A Dirac sea and thermodynamic equilibrium for the quantized three-wave interaction

Richard Beals
Department of Mathematics, Box 208283, Yale University, New Haven, Connecticut 06520-8283

David Sattinger
Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455

Eric Williams
Institute of Advanced Studies, United Nations University, Shibuya-ku Tokyo 150, Japan

(Received 3 February 1997; accepted for publication 22 August 1997)

The classical version of the three-wave interaction models the creation and destruction of waves; the quantized version models the creation and destruction of particles. The quantum three-wave interaction is described and the Bethe ansatz for the eigenfunctions is given in closed form. The Bethe equations are derived in a rigorous fashion and are shown to have a thermodynamic limit. The Dirac sea of negative energy states is obtained as the infinite density limit. Finite particle/hole excitations are determined and the asymptotic relation of energy and momentum is obtained. The Yang–Yang functional for the relative free energy of finite density excitations is constructed and is shown to be convex and bounded below. The equations of thermal equilibrium are obtained. © *1998 American Institute of Physics.* [S0022-2488(98)01801-5]

I. INTRODUCTION

The creation and destruction of particles is a fundamental process in high-energy physics. Among other things, the presence of such processes implies that the entries of the scattering matrix between states with different numbers of elementary excitations (or physical particles) can be nonzero.

Integrable systems provide solvable paradigms for various physical phenomena: e.g., for lattice statistical models, the algebraic decay of correlation functions at a critical point may be observed. The creation and destruction of particle pairs is modeled by the quantum three-wave interaction (3WI), which is obtained by quantizing a particular reduction of the $N=3$ case of the AKNS hierarchy as in Wadati and Ohkuma,[1] or by taking the continuum limit of the GL(3) lattice magnet as in Kulish and Reshetikhin.[2] The result is an interacting many-body system which does not conserve the number of particles: the interaction permits two distinct types of particles to combine to form a third, and also the reverse process.

The quantum 3WI has been considered in three forms. In one, all particles are bosons, in the ''Fermion I'' model two fermions interact to form a boson, and in the ''Fermion II'' model a boson and a fermion interact to form a fermion. In this paper we consider the Fermion I model. We find the associated first quantization problem, obtain the eigenfunctions of the Hamiltonian in closed form, and derive the Bethe equations rigorously from the requirement that the wave functions be periodic.

Because the model is first order in the derivatives, the energy is not bounded below. Nevertheless, as in Dirac's model of the electron, a stable ground state may be constructed by filling all negative energy states. One can then consider all states with reference to the filled Dirac sea of negative energy states. This approach is used in Korepin[3] and in Bergknoff and Thacker[4] to calculate the excitation spectrum of the massive Thirring model from the Bethe ansatz states.

We prove the existence of a thermodynamic limit as the length of the interval tends to infinity

0022-2488/98/39(1)/1/29/$15.00
J. Math. Phys. **39** (1), January 1998 © 1998 American Institute of Physics

with constant density. The Dirac sea is obtained in the infinite density limit. Finite particle/hole excitations from the Dirac sea correspond to the solutions of certain integral equations. Such excitations have positive relative energy, and we determine the asymptotic relation between energy and momentum. Following the treatment of a one-dimensional boson gas with delta-function interactions by Yang and Yang[5] we derive a functional for the relative free energy of finite (local) density perturbations from the Dirac sea. We prove that the functional is convex and bounded below and obtain the equation for thermal equilibrium.

Kulish and Reshetikhin[2] used the quantum inverse scattering method (also called the algebraic Bethe ansatz) to investigate GL(N) invariant transfer matrices on a one-dimensional lattice. They obtained quantized N wave models as a formal continuum limit and found explicitly the Bethe equations for the bosonic 3WI. Wadati and Ohkuma[1] also considered the bosonic case. Ohkuma and Wadati[6] treated the Fermion I and Fermion II cases and gave an algorithm for constructing the Bethe wave functions.

Ohkuma[7] investigated the Fermion I quantization of the three-wave interaction, wrote down the Bethe equations, and calculated the free energy and excitations from a ground state. However, a momentum cutoff is introduced in Ohkuma.[7] The results depend on the cutoff and do not have a finite limit when the cutoff is taken to infinity. By constructing all physical states with reference to a Dirac sea we obtain limits that are finite and are independent of any cutoff. The methods provide a basis for calculating the free energy of other integrable models whose spectrum is unbounded below.

The quantized field of the 3WI may be rewritten as a first quantization for a many-body problem. Such a reduction was carried out by Lieb and Liniger[8] for the quantization of a Bose gas–essentially the quantization of the nonlinear Schrödinger equation. They obtained a Helmholtz equation in the interior of the fundamental domain, together with boundary conditions at the interfaces associated with the interactions; see also Ref. 9. A novel feature of the 3WI is that the system of equations links functions with different numbers of variables, corresponding to the fact that individual particle numbers are not conserved.

As in Lieb and Liniger[8] we solve the differential equations in certain fundamental cones $\mathscr{C}^{m,n,p}$ and then extend the densities to all space, first by antisymmetrizing across the boundaries of $\mathscr{C}^{m,n,p}$ and then by periodicity. The differential equations for the densities force discontinuities where fermions of different types interact. However, like fermions do not interact, so to avoid discontinuities on the boundaries of the fundamental cones we impose a regularity requirement, (2.5). This requirement forces one to antisymmetrize the basic solutions of the differential equations over momenta. This in turn is fundamental to obtaining the Bethe equations, as we show by example in Sec. V.

In Secs. III and IV we sketch, in greater detail than in Refs. 1 and 6, a derivation of an algorithm for the eigenfunctions of the 3WI, and go on to obtain the explicit form of the eigenfunctions (cf. Theorem 5.1). This form generalizes the Bethe ansatz (Refs. 8 and 10–12) to systems in which the eigenvectors are superpositions of states, having different numbers of particles. In Sec. V we derive the Bethe equations from the requirement of periodicity.

In Sec. VI we reformulate the finite Bethe equations as a system of integral equations and prove existence of a certain thermodynamic limit at finite density. The Dirac sea is then obtained as an infinite density limit. In Sec. VII integral equations for the particle/hole excited states are obtained, and energy/momentum asymptotics are found. Finite density perturbations of the Dirac sea are constructed in Sec. VIII. In Sec. IX we investigate the (relative) free energy functional. The formal Euler–Lagrange equations for its minimum are the same as those proposed by Ohkuma[7] if one imposes a cutoff. Without a cutoff they have no strict solutions, but there is a natural interpretation which does yield meaningful equations for the thermal equilibrium state.

II. THE THREE-WAVE INTERACTION

The Hamiltonian of the classical three-wave interaction is

$$H=\int \sum_{j=0}^{2} ic_j \frac{\partial \Psi_j^\dagger}{\partial x} \Psi_j+\gamma(\Psi_0^\dagger \Psi_1 \Psi_2+\Psi_2^\dagger \Psi_1^\dagger \Psi_0)dx=T+\gamma(V^\dagger+V), \tag{2.1}$$

where Ψ_j are complex-valued functions and † denotes complex conjugation. The constants c_j are wave speeds, and γ is a coupling constant. Zakharov and Manakov[13] found a Lax pair for the classical system and Beals and Sattinger[14] gave a set of action-angle variables. However, integrability of the classical system will play no role in our analysis of the quantized model.

The system is formally quantized by taking the $\Psi_j^\dagger$ and Ψ_j to be creation and annihilation operators on a Fock space, † denoting Hermitian adjoints. We take $\Psi_0(x,t)$ to be a boson field operator, and Ψ_1 and Ψ_2 to be fermion fields. Thus the operators are required to satisfy the equal time commutation $([,]_-)$ or anticommutation $([,]_+)$ relations:

$$\begin{gathered}[\Psi_j(x,t),\Psi_{j'}(x',t)]_+=0, \quad [\Psi_j(x,t),\Psi_{j'}^\dagger(x',t)]_+=\delta_{jj'}\delta(x-x'),\\ [\Psi_0(x,t),\Psi_0(x',t)]_-=0, \quad [\Psi_0(x,t),\Psi_0^\dagger(x',t)]_-=\delta(x-x'),\\ [\Psi_0(x,t),\Psi_j(x',t)]_-=0, \quad [\Psi_0^\dagger(x,t),\Psi_j(x',t)]_-=0, \quad j,j'=1,2.\end{gathered} \tag{2.2}$$

The Hermitian operator T represents the kinetic energy, with distinct wave speeds c_j for each field. The interaction term $V^\dagger$ formally annihilates the two particles with wave speeds c_1 and c_2 (particles of type 1 and 2) and creates one with wave speed c_0, while V destroys a type 0 particle and creates types 1 and 2.

We assume throughout that the speeds of the fermions bracket the boson speed: $(c_1-c_0)\times(c_2-c_0)<0$. In addition, we make the simplifying assumption that

$$c_1-c_0=\alpha=c_0-c_2, \tag{2.3}$$

where α is an arbitrary constant. Without loss of generality we assume throughout this paper that $\alpha>0$.

The Hamiltonian is defined on a Hilbert space $\mathscr{H}$ consisting of the tensor product of a boson and two fermion Fock spaces. There exists a nonzero vector $|0\rangle$ in the Fock space, called the pseudovacuum, such that

$$\Psi_j(x,t)|0\rangle=0, \quad j=0,1,2.$$

Define particle number operators

$$N_j=\int dx\ \Psi_j^\dagger \Psi_j, \quad j=0,1,2.$$

The coordinates of particles of types 0, 1, and 2 are denoted by x_j, y_j, and z_j, respectively. Let

$$\Psi_0^\dagger(x)=\Psi_0^\dagger(x_1,\ldots,x_m)=\Psi_0^\dagger(x_1)\cdots\Psi_0^\dagger(x_m)$$

and define $\Psi_1^\dagger(y)$ and $\Psi_2^\dagger(z)$ similarly. Then set

$$\Psi^\dagger(x,y,z)=\Psi_0^\dagger(x)\Psi_1^\dagger(y)\Psi_2^\dagger(z), \quad x\in R^m, \quad y\in R^n, \quad z\in R^p.$$

Fundamental states in $\mathscr{H}$ are given formally by

$$\Psi^\dagger(x,y,z)|0\rangle,$$

which are eigenstates of N_0, N_1, and N_2 with eigenvalues, m, n, and p, respectively. A general state vector in this subspace may be written in the form

$$|\psi_{m,n,p}\rangle = \frac{1}{m!n!p!}\int_{R^{m,n,p}} f_{mnp}(x,y,z)\Psi^\dagger(x,y,z)|0\rangle\ dx\ dy\ dz, \tag{2.4}$$

where f_{mnp} belongs to $\mathscr{L}_2(R^{m,n,p})$ and $dx = dx_1 dx_2 \cdots dx_m$, etc. Since $\Psi^\dagger(x,y,z)$ is symmetric in the x_j and antisymmetric in the y_k and z_l, the density function f_{mnp} may be assumed to possess these same symmetries.

If a density f_{mnp} is associated with an eigenstate, then, as shown in Sec. IV, it must be discontinuous across the hyperplanes $y_k = z_l$, which we call the *interaction manifolds*. It is natural to assume that these are the only discontinuities. We therefore assume continuity across the hyperplanes $y_k = y_{k+1}$, provided $y_k \neq z_l$ for all z_l, and similarly with respect to the z coordinates. Continuity plus antisymmetry across these hyperplanes implies the following.

$$\begin{aligned} &\textit{Regularity condition}\text{: } f_{mnp} \text{ vanishes as } y_k \to y_{k\pm1} \text{ if } y_{k\pm1} \neq z_l \text{ for all } l \\ &\qquad\qquad \text{and similarly as } z_l \to z_{l\pm1} \neq y_k, \text{ all } k. \end{aligned} \tag{2.5}$$

By symmetry, we may replace the integration in (2.4) by integration over the cone $C^{m,n,p}$ given by

$$x_1 < x_2 < \cdots < x_m, \quad y_1 < y_2 < \cdots < y_n, \quad z_1 < z_2 < \cdots < z_p,$$

and write

$$|\psi_{m,n,p}\rangle = \int_{C^{m,n,p}} f_{mnp}(x,y,z)\Psi^\dagger(x,y,z)|0\rangle\ dx\ dy\ dz. \tag{2.6}$$

From now on, we write the state vectors in the form (2.6) and assume that the densities f_{mnp} are supported on the cone $C^{m,n,p}$ and satisfy the regularity condition (2.5), which implies that the antisymmetric extensions of the f_{mnp} are continuous across the collision manifolds $y_j = y_k$ and $z_j = z_k$ of like fermions.

The classical version of the Hamiltonian (2.1) is integrable and possesses an infinite number of conservation laws: see Ref. 13. Three important conserved quantities for each version are

$$N = \int 2\Psi_0^\dagger\Psi_0 + \Psi_1^\dagger\Psi_1 + \Psi_2^\dagger\Psi_2\ dx,$$

$$D = \int \Psi_1^\dagger\Psi_1 - \Psi_2^\dagger\Psi_2\ dx,$$

$$P = \int \sum_{j=0}^{2} i\Psi_j^\dagger \frac{\partial\Psi_j}{\partial x}\ dx.$$

In the quantized version operators N and D give conserved particle numbers, and P is the momentum operator. Indeed it is not difficult to show from (2.2) that N, D, and P commute with H.

III. REDUCTION TO FIRST QUANTIZATION

The quantum field problem can be rewritten as a first quantization problem for a many-body problem.

Denote the subspace of states $|\psi_{m,n,p}\rangle$ by $\mathscr{F}_{mnp}$. These are eigenstates of N and D with eigenvalues $N=2m+n+p$ and $D=n-p$, respectively, but are not eigenstates of the Hamiltonian H, since $V^\dagger$ and V create and destroy particles. However, since N and D commute with H, the quantum numbers N and D label a subspace that is invariant under H, N, and D. We denote this space by $\mathscr{F}_{N,D}$.

We now fix N and D. Within $\mathscr{F}_{N,D}$, densities can be labeled by m alone and we denote the corresponding subspace by $\mathscr{F}_m$. Note that both of $N\pm D$ must be non-negative even integers. When there are no bosons present ($m=0$), there are $(N\pm D)/2$ fermions of types 1 and 2, respectively. The maximum number of bosons occurs when the maximal number of fermions have combined, yielding $M=\min(N\pm D)/2=\frac{1}{2}(N-|D|)$.

The operators T, $V^\dagger$, and V act on the subspaces $\mathscr{F}_m$ as follows:

$$T:\mathscr{F}_m\to\mathscr{F}_m, \quad V^\dagger:\mathscr{F}_m\to\mathscr{F}_{m+1}, \quad V:\mathscr{F}_m\to\mathscr{F}_{m-1}.$$

An eigenstate of the Hamiltonian in $\mathscr{F}_{N,D}$ is a sum of vectors $|\psi_m\rangle\in\mathscr{F}_m$, and the eigenvalue equation becomes the coupled system

$$\begin{aligned}
(E-T)|\psi_0\rangle &= \gamma V|\psi_1\rangle \\
&\vdots \\
(E-T)|\psi_m\rangle &= \gamma V^\dagger|\psi_{m-1}\rangle+\gamma V|\psi_{m+1}\rangle \\
&\vdots \\
(E-T)|\psi_M\rangle &= \gamma V^\dagger|\psi_{M-1}\rangle
\end{aligned} \tag{3.1}$$

since $V|\psi_0\rangle=V^\dagger|\psi_M\rangle=0$.

Let f_m be the density corresponding to the vector $|\psi_m\rangle$. Equations (3.1) may be converted to a system of partial differential equations by calculating the action of T, $V^\dagger$, and V on the densities f_m. The calculation of the action of T is straightforward, while the calculation of V and $V^\dagger$ is somewhat more involved; details are given in Ref. 15. The action of T is

$$T|\psi_m\rangle = T\int_{\mathscr{C}^{m,n,p}} f_m(x,y,z)\Psi^\dagger(x,y,z)|0\rangle dx\ dy\ dz \tag{3.2}$$

$$= \int_{\mathscr{C}^{m,n,p}} (X_m f_m(x,y,z))\Psi^\dagger(x,y,z)|0\rangle dx\ dy\ dz, \tag{3.3}$$

where

$$X_m f = ic_0\sum_{j=0}^{m}\frac{\partial f}{\partial x_j}+ic_1\sum_{k=0}^{n}\frac{\partial f}{\partial y_k}+ic_2\sum_{l=0}^{p}\frac{\partial f}{\partial z_l}. \tag{3.4}$$

We use $V^\dagger$ to denote the action on densities as well as the action on vectors. Then

$$V^\dagger f_m = \sum_{j,k,l=1}^{m+1,n,p}(-1)^{n(m)+k+l} f_m(\hat{x}_j;\langle y,x_j,k\rangle;\langle z,x_j,l\rangle), \tag{3.5}$$

where $(x,y,z)\in\mathscr{C}^{m+1,n-1,p-1}, n(m)=(N+D)/2-m,$ and

$$\hat{x}_j=(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_m)\in\mathscr{C}^m$$

$$\langle y,x_j,k\rangle=(y_1,\ldots,y_{k-1},x_j,y_k,\ldots,y_{n-1})\in\mathscr{C}^n,\quad \text{etc.}$$

Remarks: (1) Since f_m is supported on $\mathscr{C}^{m,n,p}$, $f_m(\hat{x}_j;\langle y,x_j,k\rangle;\langle z,x_j,l\rangle)=0$ unless $x_j\in(y_{k-1},y_k)\cap(z_{l-1},z_l)$. Hence

$$(V^\dagger f_m)(x;y;z)=\sum_{j=1}^{m+1}(-1)^{n(m)+k_j+l_j}f_m(\hat{x}_j;\langle y,x_j,k_j\rangle;\langle z,x_j,l_j\rangle),\tag{3.6}$$

where k_j and l_j are uniquely determined by the constraint

$$x_j\in(y_{k_j-1},y_{k_j})\cap(z_{l_j-1},z_{l_j}).\tag{3.7}$$

(2) In (3.6) we interpret the value of f_m on the interaction manifold $x_j=y_{k_j}=z_{l_j}$ as the average of its values as $y_{k_j}\to z_{l_j}\pm$.

(3) The action of $V^\dagger$ can be thought of as gluing a type 1 fermion at y_k and a type 2 fermion at z_l to form a boson at x_j. The coordinates are relabeled to take into account the shift in numbers of variables, summed over all possible pairs. Similarly, the action of V can be thought of as the decomposition of a boson into a pair of fermions, summed over the bosons. In fact the operation of V on densities is given by

$$(Vf_m)(x,y,z)=\sum_{k,l=1}^{n+1,p+1}(-1)^{n(m)+k+l}f_m(\langle x,y_k,j_k\rangle;\hat{y}_k;\hat{z}_l)\,\delta(y_k-z_l),\tag{3.8}$$

where j_k is uniquely determined by the constraint $y_k\in(x_{j_k-1},x_{j_k})$.

We illustrate in some low-order cases. For $2m+n+p=N=2$, $f_0=f_0(y,z)$, $f_1=f_1(x)$:

$$(V^\dagger f_0)(x)=-f_0(x,x),\quad Vf_0=0,$$

$$V^\dagger f_1=0,\quad (Vf_1)(y,z)=f_1(y)\,\delta(y-z).$$

For $2m+n+p=N=4$, $f_0=f_0(y_1,y_2,z_1,z_2)$, $f_1=f_1(x,y,z)$, $f_2=f_2(x_1,x_2)$:

$$(V^\dagger f_0)(x,y,z)=f_0(x,y,x,z)-f_0(x,y,z,x)-f_0(y,x,x,z)+f_0(y,x,z,x),$$

$$(V^\dagger f_1)(x_1,x_2)=-f_1(x_1,x_2,x_2)-f_1(x_2,x_1,x_1),$$

$$V^\dagger f_2=0,\quad Vf_0=0,$$

$$(Vf_1)(y_1,y_2,z_1,z_2)=f_1(y_1,y_2,z_1)\,\delta(y_1-z_2)-f_1(y_1,y_2,z_2)\,\delta(y_1-z_1)$$

$$-f_1(y_2,y_1,z_1)\,\delta(y_2-z_2)+f_1(y_2,y_1,z_2)\,\delta(y_2-z_1),$$

$$(Vf_2)(x,y,z)=[f_2(y,x)+f_2(x,y)]\,\delta(y-z).$$

IV. THE BETHE ANSATZ

The differential equations for the densities, supplemented with the regularity condition (2.5) at the boundaries of the cones $\mathscr{C}^{m,n,p}$, describe the spectral problem in first quantization.

From the results of Sec. III the differential equations for the densities are

$$\begin{aligned} &X_0 f_0 + \gamma V f_1 = E f_0, \\ &\vdots \\ &X_m f_m + \gamma V f_{m+1} + \gamma V^\dagger f_{m-1} = E f_m \\ &\vdots \\ &X_M f_M + \gamma V^\dagger f_{M-1} = E f_M . \end{aligned} \tag{4.1}$$

Wadati and Ohkuma[1] gave an algorithm for constructing the general eigenstates. In the next two sections we find these eigenstates in closed form, in two steps. In this section we construct ''local solutions'' of Eqs. (4.1); on each of the cones $\mathscr{C}^{mnp}$ they are multiples of an exponential solution of the associated differential equation.

Because the operator V introduces delta function singularities on the interaction manifolds, solutions must have discontinuities across these manifolds. However, the local solutions violate (2.5) and thus they have additional discontinuities if extended by antisymmetry. We denote the densities of the local solutions by h_m to distinguish them from the regular densities, which will satisfy (2.5). We obtain these regular densities in Sec. V by taking linear combinations of the h_m: antisymmetrizing over the momenta.

A. Low-order cases

Two low-order cases illustrate the main features of these equations. For $N=2$, $D=0$, Eqs. (4.1) are

$$ic_1 \frac{\partial h_0}{\partial y} + ic_2 \frac{\partial h_0}{\partial z} + \gamma h_1(y)\,\delta(y-z) = E h_0(y,z), \tag{4.2a}$$

$$ic_0 \frac{\partial h_1}{\partial x} - \gamma h_0(x,x) = E h_1(x), \tag{4.2b}$$

where $h_0 = h_0(y,z)$ and $h_1 = h_1(x)$. In the first equation the interaction term vanishes off the interaction manifold $y=z$ so we take the solutions to be multiples of an exponential in each of the open sectors $y<z$ and $y>z$ of $\mathscr{C}^{0,1,1}$:

$$h_0(y,z) = \begin{cases} \phi_1^0 \exp i(\eta y - \zeta z), & y<z, \\ \phi_2^0 \exp i(\eta y - \zeta z), & y>z, \end{cases}$$

where $E = -(c_1\eta - c_2\zeta)$, and ϕ_1^0 and ϕ_2^0 are constants to be determined.

In (4.2b) we take $h_0(x,x)$ on the interaction manifold to be the average of the values of $h_0(y,z)$ from the left and right. Hence the second equation is

$$\left(ic_0 \frac{\partial}{\partial x} - E \right) h_1(x) = \gamma \frac{\phi_1^0 + \phi_2^0}{2} e^{i(\eta-\zeta)x}.$$

We seek a solution of the form $h_1(x) = Ae^{i(\eta-\zeta)x}$, so that

$$h_1(x)=g(\eta,\zeta)\,\frac{\phi_1^0+\phi_2^0}{2}\exp\, i(\eta-\zeta)x,\quad g(\eta,\zeta)=\frac{\gamma}{(c_1-c_0)\eta-(c_2-c_0)\zeta}.$$

By (2.3),

$$g(\eta,\zeta)=\frac{\gamma}{\alpha(\eta+\zeta)}:=g(\eta+\zeta).$$

We use this form of g from now on.

Finally, we determine the ratio ϕ_1^0/ϕ_2^0 by integrating (4.2a) across the interaction manifold $y=z$. We change variables, letting $d=y-z$ and $s=y+z$, and obtain

$$i(c_1-c_2)\,\frac{\partial h_0}{\partial d}+i(c_1+c_2)\,\frac{\partial h_0}{\partial s}+\gamma h_1\delta(d)=Eh_0,$$

$$h_0=h_0\left(\frac{s+d}{2},\,\frac{s-d}{2}\right),\quad h_1=h_1\left(\frac{s+d}{2}\right).$$

Integrating across the jump at $d=0$ $(y=z=x)$ we obtain

$$i(c_1-c_2)(\phi_2^0-\phi_1^0)\exp\, ix(\eta-\zeta)+\gamma g(\eta+\zeta)\,\frac{\phi_1^0+\phi_2^0}{2}\exp\, ix(\eta-\zeta)=0;$$

hence

$$\frac{\phi_2^0}{\phi_1^0}=\frac{2i(c_1-c_2)-\gamma g}{2i(c_1-c_2)+\gamma g}.$$

Using our assumption (2.3) we may write this in the simplified form

$$\phi_2^0=\theta(\eta+\zeta)\phi_1^0,\quad \theta(\lambda)=\frac{\lambda+i\omega}{\lambda-i\omega},$$

where

$$\omega=\frac{\gamma^2}{4\alpha^2}.$$

Setting $\phi_1^0=1$, we have

$$h_0(y,z)=\begin{cases}\exp\, i(\eta y-\zeta z), & y<z,\\ \theta(\eta-\zeta)\exp\, i(\eta y-\zeta z), & y>z,\end{cases}$$

and

$$h_1(x)=\tfrac{1}{2}g(\eta+\zeta)(1+\theta(\eta+\zeta))\exp\, ix(\eta-\zeta).$$

For $N=3$ and $D=1$ Eqs. (4.1) are

$$ic_1\,\frac{\partial h_0}{\partial y_1}+ic_1\,\frac{\partial h_0}{\partial y_2}+ic_2\,\frac{\partial h_0}{\partial z}-Eh_0(y_1,y_2;z)=\gamma(h_1(y_1;y_2)\,\delta(y_1-z)-h_1(y_2;y_1)\,\delta(y_2-z)),\tag{4.3a}$$

$$ic_0\frac{\partial h_1}{\partial x}+ic_1\frac{\partial h_1}{\partial y}-Eh_1(x;y)=\gamma(h_0(y,x;x)-h_0(x,y;x)), \tag{4.3b}$$

with $h_0=h_0(y_1,y_2;z)$ and $h_1=h_1(x:y)$. Arguing as above, we obtain

$$h_0(y;z)=\begin{cases}\exp\, i(\eta_1y_1+\eta_2y_2-\zeta z), & y_1<y_2<z,\\ \theta(\eta_2+\zeta)\exp\, i(\eta_1y_1+\eta_2y_2-\zeta z), & y_1<z<y_2,\\ \theta(\eta_1+\zeta)\theta(\eta_2+\zeta)\exp\, i(\eta_1y_1+\eta_2y_2-\zeta z), & z<y_1<y_2,\end{cases}$$

and

$$h_1(x;y)=\frac{1}{2}\begin{cases}(1+\theta(\eta_2+\zeta))g(\eta_2+\zeta)\exp\, i((\eta_2-\zeta)x+\eta_1y), & y<x,\\ -\theta(\eta_2+\zeta)(1+\theta(\eta_1+\zeta))g(\eta_1+\zeta)\exp\, i((\eta_1-\zeta)x+\eta_2y), & x<y.\end{cases}$$

B. The general case

The two preceding cases show that when y_k and z_l are transposed (with y_k moving to the right of z_l), h_0 is multiplied by $\theta(\eta_k+\zeta_l)$, where η_k and $-\zeta_l$ are the momenta of the corresponding fermions. Since the differential equations are local and involve only pairwise interactions, this situation extends to the general case.

We begin by introducing some notation and conventions. For fixed N and D, $\mathscr{C}^{m,n,p}$ is uniquely determined by m, so from now on we denote it by $\mathscr{C}^m$. The domain $\mathscr{C}^m$ is a union of

$$\frac{(m+n+p)!}{m!n!p!}$$

open sectors determined by the interspacing of the particles of type 0, 1, and 2. We label them by $m+n+p$-tuples $\sigma=(\epsilon_1,\epsilon_2,\ldots,\epsilon_{m+n+p})$, where $\epsilon_j=0,-1$, or $+1$ if the particle in the jth position along the real line is of type 0, 1, or 2, respectively. For example,

$$\sigma=(1,0,-1,1)\leftrightarrow\{z_1<x_1<y_1<z_2\}.$$

The boundary of $\mathscr{C}^m$ consists of the interaction manifolds $y_j=z_k$ together with the sets $y_j=y_{j+1}$ and $z_k=z_{k+1}$ at which like fermions coalesce.

A *transposition* consists of interchanging a -1 with a $+1$ to its immediate right, e.g., $(1,0,-1,1)\mapsto(1,0,1,-1)$. Each transposition corresponds to interchanging an adjacent pair of coordinates y_k and z_l; thus to every interaction manifold there is a corresponding transposition. The sectors of $\mathscr{C}^m$ are partially ordered by the number of transpositions; thus, for example, $(1,0,-1,1)<(1,0,1,-1)$. By definition, no transpositions occur between -1's and 1's separated by one or more zeroes.

Since Vh_1 is supported on the interaction manifold, $h_0=h_0(y;z)$ satisfies

$$X_0h_0=ic_1\sum_{j=1}^{n}\frac{\partial h_0}{\partial y_j}+ic_2\sum_{k=1}^{p}\frac{\partial h_0}{\partial z_k}=Eh_0 \tag{4.4}$$

in each open sector $\sigma\subset\mathscr{C}^0$.

Let

$$\sigma_0=\{y_1<\cdots<y_n<z_1<\cdots<z_p\}\sim(-1,\ldots,-1,1,\ldots,1).$$

We take

$$h_0 = \exp\, i\left(\sum_{j=1}^{n} \eta_j y_j - \sum_{k=1}^{p} \zeta_k z_k \right), \quad (y;z) \in \sigma_0,$$

with

$$E = -c_1 \sum_{j=1}^{n} \eta_j + c_2 \sum_{k=1}^{p} \zeta_k. \tag{4.5}$$

The solution in a general sector $\sigma \in \mathscr{C}^0$ is

$$h_0 = \phi_\sigma^0 \exp\, i\left(\sum_{j=1}^{n} \eta_j y_j - \sum_{k=1}^{p} \zeta_k z_k \right), \quad (y;z) \in \sigma, \tag{4.6}$$

where the constant ϕ_σ^0 is uniquely determined by σ, and is given in (4.8) below.

Lemma 4.1. Let $\sigma < \sigma'$ differ by a single transposition of fermions of types 1 and 2 at y_j, z_k with momenta η_j and $-\zeta_k$, respectively. Then

$$\phi_{\sigma'}^0 = \theta(\eta_j + \zeta_k)\, \phi_\sigma^0. \tag{4.7}$$

Remark: For $\sigma' < \sigma$ the phase factor is

$$[\theta(\eta_j + \zeta_k)]^{-1} = \theta(-\eta_j - \zeta_k).$$

Thus within a *fixed* particle number sector these conditions on the wave function are of the usual Bethe ansatz type; the ratio of phases that differ by a permutation of neighboring particles is a function of the momenta.

Proof: All the interactions are pairwise and local, so the same computations for the case $N = 2$ and $D = 0$ extend to the general case. The ratio of the constants ϕ_σ^0 and $\phi_{\sigma'}^0$, is determined entirely from the first and second equations in the hierarchy (4.1). □

Each $\sigma \in \mathscr{C}^0$ is obtained from $\sigma_0 = (-1,\ldots,-1,1,\ldots,1)$ by a unique set of transpositions (y_j, z_k), which we denote by T_σ. Thus

$$\phi_\sigma^0 = \prod_{(j,k) \in T_\sigma} \theta(\eta_j + \zeta_k). \tag{4.8}$$

Now consider the higher-order components of the density: h_m, $m \geqslant 1$. The sectors of $\mathscr{C}^m$ are characterized by $(m+n+p)$-tuples σ with m zeroes. We assign momenta to σ as follows. Regard each 0 in σ as a pair $(-1,1)$. Starting from the left, assign the -1's momenta η_1, $\eta_2,\ldots$; the $+1$'s momenta $-\zeta_1$, $-\zeta_2,\ldots$; and assign the 0's the sums of momenta $\eta_j - \zeta_k$, with the η's and ζ's taken in their normal order. For example,

$$(-1,0,1,0) \leftrightarrow (\eta_1, \eta_2 - \zeta_1, -\zeta_2, \eta_3 - \zeta_3),$$

$$(-1,1,1,0,1,0,-1) \leftrightarrow (\eta_1, -\zeta_1, -\zeta_2, \eta_2 - \zeta_3, -\zeta_4, \eta_3 - \zeta_5, \eta_4).$$

For fixed locations of the zeroes in σ the basic exponential solution is defined by

$$e_\sigma^m = \exp\, i\left(\sum_{r=1}^{m} x_r(\eta_{i_r} - \zeta_{j_r}) + \sum_{s=1}^{n} y_s \eta_{a_s} - \sum_{t=1}^{p} z_t \zeta_{b_t} \right), \tag{4.9}$$

where

$$i_r(j_r) = \#x\text{'s and } y\text{'s}(z\text{'s}) \text{ to the left of } x_r, \text{ inclusive of } x_r;$$
$$a_s(b_t) = \#x\text{'s and } y\text{'s}(z\text{'s}) \text{ to the left of } y_s, \ (z_t), \text{ inclusive of } y_s, \ (z_t). \tag{4.10}$$

Remark: From (3.7), $i_r = r + k_r - 1$, and $j_r = r + l_r - 1$.
For example,

$$e^m_\sigma(x;y;z) = \exp\, ix_1(x_1(\eta_2 - \zeta_1) + x_2(\eta_3 - \zeta_3) + y_1\eta_1 - z_1\zeta_2), \quad \sigma = (-1,0,1,0);$$
$$e^m_\sigma(x;y;z) = \exp\, i(x_1(\eta_2 - \zeta_3) + x_2(\eta_3 - \zeta_5) + y_1\eta_1 + y_2\eta_4 - z_1\zeta_1 - z_2\zeta_2 - z_3\zeta_4), \tag{4.11}$$
$$\sigma = (-1,1,1,0,1,0,-1)$$

We now describe the general case. For any $\sigma \in \mathscr{C}^m$ define g_σ by

$$g_\sigma = \prod_{r=1}^m g_{i_r j_r}, \quad g_{i_r j_r} = g(\eta_{i_r} + \zeta_{j_r}), \tag{4.12}$$

where i_r and j_r are given in (4.10). For example, for $(-1,0,1,0)$, $i_1 = 2$, $j_1 = 1$, $i_2 = 3$, $j_2 = 2$, and

$$g_\sigma = g_{21} g_{32}.$$

The components of $\mathscr{C}^m$ are determined by the location of the zeroes and the numbers and orders of $+1$'s and -1's between successive zeroes. The factor g_σ is constant on each connected component of $\mathscr{C}^m$. For $\sigma \in \mathscr{C}^m$, we define

$$\phi^m_\sigma = \phi^0_\Sigma \prod_{r=1}^m (-1)^{n(m)+k_r+l_r} \frac{1 + \theta(\eta_{i_r} + \zeta_{j_r})}{2}, \tag{4.13}$$

where $\Sigma \in \mathscr{C}^0$ is obtained from σ by replacing each 0 in σ by $(-1,1)$, and ϕ^0_Σ is given by (4.8).

Theorem 4.2: *The basic solutions of Eqs. (4.1) are given by*

$$h_m(x,y,z) = \psi^m_\sigma e^m_\sigma(x,y,z), \tag{4.14}$$

where

$$\psi^m_\sigma = \phi^m_\sigma g_\sigma = \phi^0_\Sigma \left(\frac{\gamma}{\alpha}\right)^m \prod_{r=1}^m \frac{(-1)^{n(m)+k_r+l_r}}{\eta_{i_r} + \zeta_{j_r} - i\omega}$$

and the ϕ^m_σ are given by (4.8) and (4.13); the g_σ by (4.12); and the e^m_σ by (4.9).

The proof is given in Ref. 15. In addition to Lemma 4.1, it relies on the identity

$$(-1)^{n(m)+K_r+l_r} \frac{\phi^m_{\sigma_r} + \phi^m_{\sigma'_r}}{2} = \phi^{m+1}_\sigma, \quad 1 \leqslant r \leqslant m+1,$$

where σ_r and σ'_r are obtained from σ by replacing the rth; zero by $(-1,1)$ and by $(1,-1)$, respectively. This condition insures compatibility between the phases in *different* particle number sectors; it has no analogue in the usual Bethe ansatz.

V. THE BETHE EQUATIONS

In Sec. IV we constructed basic solutions h_m of Eqs. (4.1). As noted in Sec. II we want to impose the regularity condition (2.5) that eigendensities have no unnecessary discontinuities, i.e., that their antisymmetric extensions be continuous across the collision manifold of like fermions. To obtain this property, we antisymmetrize the basic solutions over the momenta.

Permuting the momenta of the basic solution of (4.4) and (4.5) gives another solution. The general solution of (4.4) and (4.5) is a linear combination

$$e_0(J,K)=\exp\, i\left(\sum_{j=1}^{n} y_j\eta_{J_j}-\sum_{k=1}^{p} z_k\zeta_{K_k}\right),$$

where $J$and K are permutations of 1,...,n and 1,...,p, respectively. In particular,

$$f_0(y;z)=\sum_{J,K}\ \text{sgn}\ (J)\ \text{sgn}\ (K)e_0(J,K)$$

satisfies $X_0f_0=Ef_0$ in

$$\sigma_0=\{y_1<\cdots<y_n<z_1<\cdots<z_p\},$$

and satisfies the regularity condition (2.5).

The density in the other sectors is given by

$$f_0(y;z)=\sum_{J,K}\ \text{sgn}(J)\text{sgn}(K)\phi^0_\sigma(J,K)e_0(J,K), \tag{5.1}$$

where $(y;z)\in\sigma$ and

$$\phi^0_\sigma(J,K)=\prod_{(j,k)\in T_\sigma}\theta(\eta_{J_j}+\zeta_{K_k}).$$

The corresponding densities for higher m are consequently

$$f_m(x;y;z)=\sum_{J,K}\ \text{sgn}\ (J)\ \text{sgn}\ (K)\phi^m_\sigma(J,K)g_\sigma(J,K)e^m_\sigma(J,K),\quad (x;y;z)\in\sigma, \tag{5.2}$$

where $e^m_\sigma(J,K)$, $g_\sigma(J,K)$, and $\phi^m_\sigma(J,K)$ are obtained from (4.9), (4.12), and (4.13) by making the substitutions $j,\ k\mapsto J_j\ K_k$. For example, to obtain densities which vanish when $y_1=y_2\neq z$ in the case $N=3$, and $D=1$, we antisymmetric the base function over the momenta: setting

$$e_0(j,k)=\exp\, i(\eta_jy_1+\eta_ky_2-\zeta z),$$

we take

$$f_0(y_1,y_2,z)=\begin{cases} e_0(1,2)-e_0(2,1), & y_1<y_2<z,\\ \theta(\eta_2+\zeta)e_0(1,2)-\theta(\eta_1+\zeta)e_0(2,1), & y_1<z<y_2,\\ \theta(\eta_1+\zeta)\theta(\eta_2+\zeta)[e_0(1,2)-e_0(2,1)], & z<y_1<y_2.\end{cases}$$

Now $f_0(y_1,y_2,z)$ vanishes when $y_1=y_2\neq z$; but we cannot assert anything for the case $y_1<z<y_2$, since y_1 and y_2 cannot coalesce without equaling z. The corresponding density f_1 is also obtained as a superposition: setting

$$w(j,k)=g(\eta_k+\zeta)(1+\theta(\eta_k+\zeta))\exp\, i((\eta_k-\zeta)x+\eta_j y)$$

we find

$$f_1(x,y)=\begin{cases} w(1,2)-w(2,1), & y<x,\\ -\theta(\eta_2+\zeta)w(2,1)+\theta(\eta_1+\zeta)w(1,2), & x<y.\end{cases}$$

Theorem 5.1: *The antisymmetrized wave functions (5.1) and (5.2) satisfy the regularity condition (2.5) and thus determine a basis of eigenstates for the Hamiltonian (2.1).*

Proof: Consider a pair (y_k,y_{k+1}) (the proof is the same for contiguous z's) and assume there are no intervening x's or z's. Group the terms in the sum (5.1) into pairs corresponding to $J=(\ldots,j,j+1,...)$ and $J'=(\ldots,j+1,j,\ldots)$. At $y_j=y_{j+1}$ we have $e_0(J,K)=e_0(J',K)$, and the coefficient of this common value is

$$\phi^0_\sigma(J,K)-\phi^0_\sigma(J',K).$$

For any given pair (j,k) either $y_j<y_{j+1}<z_k$ or $z_k<y_j<y_{j+1}$. In the first case, neither (j,k) nor $(j+1,k)$ belong to T_σ, so neither of the factors $\theta(\eta_{J_i}-\zeta_{K_k})$ and $\theta(\eta_{J_{j+1}}-\zeta_{K_k})$ belong to the products $\phi^0_\sigma(J,K)$ and $\phi^0_\sigma(J',K)$. In the second case, both (j,k) and $(j+1,k)$ belong to T_σ, so both factors belong to both the products, and $\phi^0_\sigma(J,K)=\phi^0_\sigma(J',K)$. Therefore the coefficient of $e_0(J,K)=\epsilon_0(J',K)$ vanishes in either case, and the sum is zero. This establishes the result for f_0.

We proceed in the same way to verify that the higher-order densities f_m, $m\geqslant 1$, vanish when y_j and y_{j+1} coincide. In view of (4.16), we must show that

$$\phi^m_\sigma(J,K)g_\sigma(J,K)=\phi^m_\sigma(J',K)g_\sigma(J',K).$$

From (4.12) the constants $g_\sigma(J,K)$ depend only on the momenta associated with the bosons (the zeroes of σ), and so are unchanged under any permutation of the fermion momenta. The same applies to the product in (4.13): it depends only on the boson, not the fermion momenta. Therefore the issue for the higher-order densities reduces to the proposition $\phi^0_\Sigma(J,K)=\phi^0_\Sigma(J',K)$, which was established in the first step of this proof. □

Given the f_m defined on the cells

$$\{0\leqslant x_1<\cdots<x_m\leqslant L,\ 0\leqslant y_1<\cdots<y_n\leqslant L,\ 0\leqslant z_1<\cdots<z_p\leqslant L\},$$

we extend them to the hypercube $0\leqslant x_i,\ y_j,\ z_k\leqslant L$ by

$$h_m(x;y;z)=\text{sgn}\ (J)\ \text{sgn}\ (K)f_m(Mx,Jy,Kz), \tag{5.3}$$

where Mx, Jy, and Kz are the unique permutations of the variables x, y, and z for which

$$x_{M_1}<\cdots<x_{M_m},\quad y_{J_1}<\cdots<y_{J_n},\quad z_{K_1}<\cdots<z_{K_p}.$$

We now impose periodic boundary conditions on the wave functions h_m. These periodicity conditions lead to constraints on the momenta, known as the Bethe equations.

For example, in the simplest case $N=2$ and $D=0$ (cf. Sec. IV) we obtained

$$f_0(y,z)=\begin{cases}\exp\, i(\eta y-\zeta z), & y<z,\\ \theta(\eta+\zeta)\exp\, i(\eta y-\zeta z), & y>z.\end{cases}$$

The periodicity conditions

$$f_0(0,z)=f_0(L,z),\quad f_0(y,0)=f_0(y,L)$$

lead immediately to the Bethe equations

$$1=\theta(\eta+\zeta)e^{i\eta L},\quad e^{i\zeta L}=\theta(\eta+\zeta).$$

For the case $n=2$ and $p=1$, the conditions (5.3) lead to

$$f_0(y_1,y_2,0)=f_0(y_1,y_2,L),\quad f_0(0,y_2,z)=-f_0(y_2,L,z).$$

From the definition of f_0 in this case we obtain, respectively,

$$\theta(\eta_1+\zeta)\,\theta(\eta_2+\zeta)(\exp\, i(\eta_1y_1+\eta_2y_2)-\exp\, i(\eta_1y_2+\eta_2y_1))$$

$$=\exp\, i(\eta_1y_1+\eta_2y_2-L\zeta)-\exp\, i(\eta_1y_2+\eta_2y_1-L\zeta)$$

and (say for $y_2<z$)

$$\exp\, i(\eta_2y_2-\zeta z)-\exp\, i(\eta_1y_2-\zeta z)=-\theta(\eta_2+\zeta)\exp\, i(\eta_1y_2+\eta_2L-\zeta z)+\theta(\eta_1+\zeta)$$
$$\times\exp\, i(\eta_2y_2+\eta_1L-\zeta z).$$

From these two equations we get the Bethe equations

$$\exp\, iL\zeta=\theta(\eta_1+\zeta)\,\theta(\eta_2+\zeta),$$

$$1=\theta(\eta_1+\zeta)\exp\, i\eta_1L,\quad 1=\theta(\eta_2+\zeta)\exp\, i\eta_2L.$$

All other periodicity conditions follow from these.

Remark: The second example shows that one must antisymmetrize in momenta to achieve periodicity. For example, suppose $N=3$, $D=1$, and $0<y_2<z$. The condition $h_0(0,y_2,z)=-h_0(y_2,L,z)$ for the basic wave function obtained in Sec. IV becomes

$$h_0(0,y_2,z)=\exp\, i(\eta_2y_2-\zeta z)=-h_0(y_2,L,z)=\theta(\eta_2+\zeta)\exp\, i(\eta_1y_2+\eta_2L-\zeta z),$$

leading to the contradiction

$$\exp\, i\eta_2y_2=-\theta(\eta_2+\zeta)\exp\, i(\eta_1y_2+\eta_2L).$$

Theorem 5.2: *Let*

$$e^{i\eta_kL}=\prod_{l=1}^{p}\theta(\zeta_l+\eta_k),\quad k=1,...,n,\tag{5.4}$$

$$e^{i\zeta_lL}=\prod_{k=1}^{n}\theta(\eta_k+\zeta_l),\quad l=1,...,p.\tag{5.5}$$

Then the densities h_m defined by (5.1), (5.2), and (5.3) are L-periodic in each variable.

The equations (5.4) and (5.5) are known as the Bethe equations. The proof is given in Ref. 15.

VI. CONSTRUCTION OF A DIRAC GROUND STATE

A solution of the Bethe equations gives an eigenvector of the Hamiltonian with energy

$$E=-c_1\sum_{k=1}^{n}\eta_k+c_2\sum_{l=1}^{p}\zeta_l,\tag{6.1}$$

where n and p denote the maximum number of particles of types 1 and 2 for states in the space $\mathscr{F}_{N,D}$, i.e., $n=\frac{1}{2}(N+D)$ and $p=\frac{1}{2}(N-D)$; cf. (4.5). The energy is not bounded below. We construct a stable ground state by filling all negative energy states.

Taking logarithms in the Bethe equations 5.4 and 5.5, we get

$$\begin{aligned}\eta_k-\frac{2\pi}{L}I_k&=\sum_{l=1}^{p}\frac{1}{iL}\log\,\theta(\eta_k+\zeta_l),\quad k=1,\ldots,n,\\ \zeta_l-\frac{2\pi}{L}J_l&=\sum_{k=1}^{n}\frac{1}{iL}\log\,\theta(\eta_k+\zeta_l),\quad l=1,\ldots,p,\end{aligned}\tag{6.2}$$

where

$$\theta(\lambda)=\frac{\lambda+i\omega}{\lambda-i\omega},\quad \omega=\frac{\gamma^2}{4\alpha^2},\tag{6.3}$$

and I_k are distinct integers, as are J_l.

Taking

$$\frac{1}{i}\log\,\theta(\lambda)=\pi-2\tan^{-1}\frac{\lambda}{\omega},\tag{6.4}$$

we write the Bethe equations in the equivalent form

$$\begin{aligned}\eta_k&=\frac{2\pi}{L}I_k+\frac{2}{L}\sum_{l=1}^{p}\tan^{-1}\frac{\eta_l+\zeta_l}{\omega},\\ \zeta_l&=\frac{2\pi}{L}J_l+\frac{2}{L}\sum_{k=1}^{n}\tan^{-1}\frac{\eta_k+\zeta_l}{\omega},\end{aligned}\tag{6.5}$$

where the I_k are distinct half-integers if p is odd, distinct integers if p is even, and similarly for the J_l in relation to n. We assume for convenience that both are even and, without loss of generality, that

$$I_1<I_2<\cdots<I_n;\quad J_1<J_2<\cdots<J_p.\tag{6.6}$$

These conditions imply that there is a solution to 6.5 for which the η_k and the ζ_l are themselves strictly increasing. Indeed Eqs. (6.7) are the variational equations for the functional

$$\Sigma=\frac{1}{2}\left(\sum_{k=1}^{n}\eta_k^2+\sum_{l=1}^{p}\zeta_l^2\right)+\frac{\pi}{L}\left(\sum_{k=1}^{n}I_k\eta_k+\sum_{l=1}^{p}J_l\zeta_l\right)-\sum_{k,l=1}^{n,p}M(\eta_k+\zeta_l),$$

where

$$M(\lambda)=\int_0^{\lambda}\frac{2}{L}\tan^{-1}\frac{s}{\omega}\,ds.$$

For convenience in this section we take $\omega=1$. (This amounts to fixing a length scale, since ω has the dimension 1/length.)

Theorem 6.1: *The Bethe equations have at least one solution for each set of distinct integers I_k and distinct integers J_l.*

Proof: The term $M(\lambda)$ grows only linearly in its argument, so the quadratic terms dominate Σ as $\eta, \zeta \to \infty$. It follows that S is bounded below and tends to infinity for large values of its argument. It therefore always has a minimum. □

Uniqueness of the solutions is a subtler question. For large values of the densities, Σ is not convex, so in principle it could have multiple critical points. In the limiting cases of interest, however, we will obtain uniqueness results. In fact we use the finite case to motivate the limiting cases. Therefore we save space and smooth the exposition by stating most of the finite case results without proof and leaving mathematical details for the limiting cases.

We now turn to the construction of a stable ground state, or vacuum state. For the ground state we assume that $D=0$ (hence $n=p=N/2$) and we consider the case when the velocities of the fermions have opposite signs:

$$c_2<0<c_1. \tag{6.7}$$

Thus $\eta>0$ and $\zeta>0$ mean negative energy for the corresponding particles. *Our procedure will be to fill the available negative energy states from the top down* by choosing consecutive integers I_k, and J_l such that the solution $\{\eta_k, \zeta_l\}$ is positive *but minimally so*; we then take the infinite length and infinite density limits. The equations (6.5) are symmetric, so we take the data $I_k=J_k$ and look for a solution with $\eta_k=\zeta_k$. Thus we want

$$I_k=J_k=k-\kappa_n, \tag{6.8}$$

where κ_n is an integer to be chosen.

Theorem 6.2: *There is a largest integer $\kappa_n=O(n)$ such that (6.5) with data (6.8) has a positive solution. This solution is unique and satisfies $\eta_k=\zeta_k$, and $\eta_1<\cdots<\eta_n$. Moreover, $\eta_1\to 0$ as $n\to\infty$, $L\to\infty$ with n/L bounded above.*

We refer to the solution in Theorem 6.2 as the *cutoff ground state* (CGS). Our interest is the infinite length, finite density limit, and then the infinite density limit. For this purpose we recast (6.5) as a pair of integral equations. Assuming still that $n=p$, we let $\Delta=2\pi n/L$ and define functions on the interval $[0,\Delta)$ by

$$\begin{aligned} I(\xi)&=\frac{2\pi}{L}I_k, \quad J(\xi)=\frac{2\pi}{L}J_k, \quad \frac{2\pi(k-1)}{L}\leqslant\xi<\frac{2\pi k}{L}, \\ \eta(\xi)&=\eta_k, \quad \zeta_L(\xi)=\zeta_k, \quad \frac{2\pi(k-1)}{L}\leqslant\xi<\frac{2\pi k}{L}. \end{aligned} \tag{6.9}$$

(We hope that context will prevent confusion between the use of η and ζ to denote *functions* and the use of η as ζ as pure variables.)

Equations (6.2) with $n=p$ are equivalent to integral equations

$$\begin{aligned} \eta(\xi)&=I(\xi)+\frac{1}{\pi}\int_0^{2\pi n/L}\tan^{-1}\left(\eta(\xi)+\zeta(\xi')\right)d\xi, \\ \zeta(\xi)&=J(\xi)+\frac{1}{\pi}\int_0^{2\pi n/L}\tan^{-1}\left(\eta(\xi')+\zeta(\xi)\right)d\xi'. \end{aligned} \tag{6.10}$$

It can be shown that for every set of data I, J corresponding to sets of integers indexed as in (6.6), there is a solution (η,ζ) with positive jumps at each of the points of discontinuity $2\pi k/L$. We define approximate inverse functions accordingly:

$$\xi(\eta)=\inf\{\xi:\eta(\xi)\geqslant\eta\}; \quad \xi(\zeta)=\inf\{\xi:\zeta(\xi)\geqslant\zeta\}. \tag{6.11}$$

Remark: These functions have jumps of size $2\pi/L$, so their difference quotients with respect to η and to ζ respectively count the number of particles associated to a given momentum interval.

With the Bethe equations in the form (6.10), it is clear how to pass to the infinite length, finite density limit. To pass also to the infinite density limit we observe that (6.10) is equivalent to

$$\eta(\xi)-\eta(0)=I(\xi)-I(0)+\frac{1}{\pi}\int_0^{2\pi n/L}[\tan^{-1}(\eta(\xi)+\zeta(\xi'))-\tan^{-1}(\eta(0)+\zeta(\xi'))]\,d\xi';$$
$$\zeta(\xi)-\zeta(0)=J(\xi)-J(0)+\frac{1}{\pi}\int_0^{2\pi n/L}[\tan^{-1}(\eta(\xi')+\zeta(\xi))-\tan^{-1}\,(\eta(\xi')+\zeta(0))]\,d\xi'. \tag{6.12}$$

The data (6.8) give rise to the functions $I(\xi)=J(\xi)=\xi-c_\Delta$ in the $L\to\infty$ limit, Δ finite, and in this limit $\eta(0)=z(0)=0$. Thus in the form (6.12) we pass to the $\Delta=\infty$ limit and find the equations for the *ground state* $(\eta_0;\zeta_0=\eta_0)$:

$$\eta_0(\xi)=\xi+\frac{1}{\pi}\int_0^{\infty}[\tan^{-1}(\eta_0(\xi)+\eta_0(\xi'))-\tan^{-1}\,(\eta_0(\xi'))]\,d\xi'. \tag{6.13}$$

We obtain existence and uniqueness of the solution to (6.13) by converting it to a linear integral equation for the derivative of the inverse function:

$$\rho_0(\eta(\xi))=\frac{1}{(d\eta_0/d\xi)(\xi)}. \tag{6.14}$$

Differentiate (6.9), multiply by ρ_0, and change the variable of integration to obtain

$$\rho_0(\eta)=1-\frac{1}{\pi}\int_0^{\infty}\frac{\rho_0(\eta')d\eta'}{1+(\eta+\eta')^2}=1-K(\rho_0). \tag{6.15}$$

Remarks: (1) In view of the remark after (6.9), ρ_0 may be thought of as the *particle density function* for the ground state.

(2) Note that the kernel of the integral operator K satisfies

$$\sup_{\eta\geqslant 0}\frac{1}{\pi}\int_0^{\infty}\frac{d\eta'}{1+(\eta+\eta')^2}=\frac{1}{2}. \tag{6.16}$$

Therefore T has norm $\frac{1}{2}$ as an operator in $L^\infty(\mathbb{R}_+)$ and (6.16) has a unique solution, obtainable by successive approximations. The solution is positive and bounded.

In the next section we need both the analytic continuation $\rho_{0,t}$ of ρ to the negative axis and a dual function $\rho^*_{0,t}$ that is the analytic continuation of the solution to the same equation but with a change of sign:

$$\rho_0^*(\eta)=1+\frac{1}{\pi}\int_0^{\infty}\frac{\rho_0^*(\eta')d\eta'}{1+(\eta+\eta')^2}=1+K\rho^*_{0,t}(\eta),\quad \eta\geqslant 0. \tag{6.17}$$

In terms of these functions we define functions ξ_0 and ξ_0^*:

$$\xi_0(\eta)=\int_0^{\eta}\rho_{0,t}(s)ds,\quad \xi_0^*(\eta)=\int_0^{\eta}\rho^*_{0,t}(s)ds,\quad \eta\in\mathbb{R}. \tag{6.18}$$

Theorem 6.3: *The extension $\rho_{0,t}$ of the particle density function ρ_0 and the dual function $\rho^*_{0,t}$ satisfy the estimates*

$$0<\rho_{0,t}(\eta)<1<\rho^*_{0,t}<2, \quad \eta\in\mathbb{R};$$

$$\rho_{0,t}(\eta)=1+O(1/\eta), \quad \rho^*_{0,t}(\eta)=2+O(1/\eta) \quad \text{as} \quad \eta\to\infty; \tag{6.19}$$

$$\pi\eta\rho_{0,t}(\eta)=-1+O(\log|\eta|/|\eta|), \quad \rho^*_{0,t}=1+O(1/\eta) \quad \text{as} \quad \eta\to-\infty.$$

The functions ξ_0 and ξ_0^* have asymptotics

$$\xi_0(\eta)=\eta+O(1), \quad \xi_0^*(\eta)=2\eta+O(1) \quad \textit{as} \quad \eta\to+\infty;$$
$$\pi\xi_0(\eta)=-\log\eta+O(1), \quad \xi_0^*(\eta)=\eta+O(1) \quad \textit{as} \quad \eta\to-\infty. \tag{6.20}$$

In particular the function ξ_0 is a diffeomorphism of the line. Its inverse function η_0 satisfies the integral equation (6.13) on $\mathbb{R}_+$.

Proof: It follows from (6.16) that the solution extends analytically. The kernel is positive, so $0<\rho_0<1$. The asymptotics of ρ_0 and of ξ_0 are easily verified from (6.15). The argument for $\rho^*_{0,t}$ and ξ_0^* are similar. Finally (6.13) results from substituting (6.16) in the definition of ξ_0 and using $\rho_0 d\eta'=d\xi'$ to change the variable of integration. □

We conclude this section with the integral expressions of various quantities associated with a state described by a pair of functions η,ζ. For a finite number of particles the momentum P and energy E are

$$P=\sum_{k=1}^n \eta_k-\sum_{l=1}^n \zeta_l=\frac{L}{2\pi}\int_0^{2\pi n/L}\eta(\xi)d\xi-\frac{L}{2\pi}\int_0^{2\pi n/L}\zeta(\xi)\,d\xi,$$
$$E=-c_1\sum_{k=1}^n \eta+c_2\sum_{l=1}^n \zeta_l=-\frac{c_1L}{2\pi}\int_0^{2\pi n/L}\eta(\xi)\,d\xi+\frac{c_2L}{2\pi}\int_0^{2\pi n/L}\zeta(\xi)\,d\xi. \tag{6.21}$$

The particle number $n=p$ itself can be expressed in terms of the inverse functions:

$$n=\frac{L}{2\pi}\sup\,\xi(\eta)=\frac{L}{2\pi}\sup\,\xi(\zeta). \tag{6.22}$$

Therefore in the infinite length limit the particle densities per unit length are

$$\frac{1}{2\pi}\sup\,\xi(\eta)=\frac{1}{2\pi}\int\rho(\eta)\,d\eta, \quad \frac{1}{2\pi}\sup\,\xi(\zeta)=\frac{1}{2\pi}\int\sigma(\zeta)\,d\zeta. \tag{6.23}$$

VII. FINITE PARTICLE EXCITED STATES

In this section we construct particle-hole excitations of the ground state. These have finite positive energy with respect to the ground state in the infinite density limit, confirming the stability of the ground state.

We begin with a single particle/hole excited state for an η particle, with momentum $\eta_p<0$ for the particle and $\eta_h>0$ for the hole. Consider first a cutoff ground state with a high particle density, so that there are large momentum particles. Denote the functions for this CGS by $\eta,\zeta=\eta$. We change the CGS data by choosing k such that η_k is close to η_h and changing I_k to I_k-s, where the integer s is chosen so that I_k-s is close to $L\xi_0(\eta_p)/2\pi$. In the integral equation picture (6.10) the new data is given by functions I, $J=I$. Here $I=I$ except on an interval of length

$2\pi/L$ located near $\xi_h=\xi_0(\eta_h)$, on which it differs from I by approximately $\xi_p-\xi_h$, $\xi_p=\xi_0(\eta_p)$. This change in data changes the η solution by a bounded amount on that interval and otherwise changes both $\eta(\xi)$ and $\zeta(\xi)$ by $O(1/L)$. Therefore we write the solution $(\overline{\eta},\overline{\zeta})$ of the perturbed equation as

$$\overline{\eta}(\xi)=\eta(\xi)+e_1(\xi)/L+f(\xi),\quad \overline{\zeta}(\xi)=\zeta(\xi)+e_2(\xi)/L, \tag{7.1}$$

where e_1 vanishes on the critical interval and f is supported on the critical interval. With a similar notation for energy and momentum, the excitation momentum and energy are

$$\begin{aligned}\delta P&=\overline{P}-P=\frac{1}{2\pi}\int_0^\Delta [e_1(\xi)-e_2(\xi)]\,d\xi+f,\\ \delta E&=\overline{E}-E=\frac{1}{2\pi}\int_0^\Delta [c_1e_1(\xi)-c_2e_2(\xi)]\,d\xi+c_1f.\end{aligned} \tag{7.2}$$

Here we abuse notation and write f also for the *value* of $\overline{\eta}-\eta$ on the critical interval.

With large L the integral equations (6.10) give

$$e_1(\xi)\approx\frac{1}{\pi}\int_0^\Delta\frac{[e_1(\xi)+e_2(\xi')]\,d\xi'}{1+(\eta(\xi)+\eta(\xi'))^2}; \tag{7.3a}$$

$$f\approx\xi_p-\xi_h+\frac{1}{\pi}\int_0^\Delta[\tan^{-1}(\eta_h+f+\eta(\xi'))-\tan^{-1}(\eta_h+\eta(\xi'))]\,d\xi'; \tag{7.3b}$$

$$e_2(\xi)\approx 2[\tan^{-1}(\eta_h+f+\eta(\xi))-\tan^{-1}(\eta_h+\eta(\xi))]+\frac{1}{\pi}\int_0^\Delta\frac{[e_1(\xi')+e_2(\xi)]\,d\xi'}{1+(\eta(\xi)+\eta(\xi'))^2}. \tag{7.3c}$$

Here (7.3a) is valid outside the critical interval.

In the infinite length, infinite density limit (6.13) and (7.3b) show that $f=\eta_h-\eta_p$, so $\overline{\eta}\approx\eta_h$ on the critical interval, as desired. In this limit the remaining equations (7.3) become

$$\begin{aligned}e_1(\eta)&=\frac{1}{\pi}\int_0^\infty\frac{[e_1(\eta)+e_2(\eta')]\rho_0(\eta')\,d\eta'}{1+(\eta+\eta')^2},\\ e_2(\eta)&=\frac{1}{\pi}\int_0^\infty\frac{[e_1(\eta')+e_2(\eta)]\rho_0(\eta')d\eta'}{1+(\eta+\eta')^2}+2\tan^{-1}(\eta_p+\eta)-2\tan^{-1}(\eta_h+\eta).\end{aligned} \tag{7.4}$$

We write $\tilde{e}_j(\eta)$ for $e_j(\xi_0(\eta))\rho_0(\eta)$: The integral equation (6.15) for ρ_0 allows (7.4) to be rewritten:

$$\begin{aligned}\tilde{e}_1(\eta)&=K(\tilde{e}_2)(\eta),\\ \tilde{e}_2(\eta)&=2\tan^{-1}(\eta_p+\eta)-2\tan^{-1}(\eta_h+\eta)+K(\tilde{e}_1)(\eta).\end{aligned} \tag{7.5}$$

In the limit the excitation momentum and energy are

$$\delta P=\frac{1}{2\pi}\int_0^\infty[\tilde{e}_1(\eta)-\tilde{e}_2(\eta)]\,d\eta+(\eta_p-\eta_h), \tag{7.6}$$

$$\delta E = \frac{1}{2\pi}\int_0^\infty [-c_1\tilde{e}_1(\eta)+c_2\tilde{e}_2(\eta)]\,d\eta - c_1(\eta_p-\eta_h)$$
$$= -\alpha\left(\frac{1}{2\pi}\int_0^\infty [\tilde{e}_1(\eta)+\tilde{e}_2(\eta)]\,d\eta+\eta_p-\eta_h\right)$$
$$+c_0\left(\frac{1}{2\pi}\int_0^\infty [\tilde{\epsilon}_1(\eta)-\tilde{e}_2(\eta)]\,d\eta-\eta_p+\eta_h\right).$$

[The last part of (7.6) follows from the second part, since $2\alpha=c_1-c_2$ and $2c_0=c_1+c_2$.]

The system (7.5) and the equations (7.6) thus characterize this single-particle/hole excitation of the ground state. Again the fact that K has norm $\frac{1}{2}$ implies that (7.6) has, for each choice of η_h and η_p, a unique solution. Positivity of the kernel implies that the solution (e_1,e_2) is negative when $\eta_p-\eta_h$ is negative, so the excitation energy is positive. We can be more precise.

Theorem 7.1: *The excited state momentum and the excitation energy can be expressed in terms of the (extended) functions ξ_0, ξ_0^* of Theorem 6.3. They have the form*

$$\delta P = P_p - P_h = P(\eta_p)-P(\eta_h)=\xi_0(\eta_p)-\xi_0(\eta_h);$$
$$\delta E = E_p - E_h = E(\eta_p)-E(\eta_h)=[-\alpha\xi_0^*(\eta_p)-c_0\xi_0(\eta_p)]-[-\alpha\xi_0^*(\eta_h)-c_0\xi_0(\eta_h)]. \tag{7.7}$$

These functions have the asymptotics

$$E_p\sim 2\alpha e^{\pi|P_p|} \quad as\ \ P_p\to-\infty,$$
$$E_h\sim -c_1P_h \quad as\ \ P_h\to+\infty. \tag{7.8}$$

Proof: The inhomogeneous term for the pair (η_p,η_h) is the sum of the inhomogeneous term for the pair $(\eta_p,0)$ and the inhomogeneous term for the pair $(0,\eta_h)$ and is an odd function of its two arguments. Therefore it suffices to consider the case $\eta_p=\tilde{\eta}$ and $\eta_h=0$. We make the $\tilde{\eta}$ dependence explicit by writing $\tilde{e}_1(\eta,\tilde{\eta})$ and $\tilde{e}_2(\eta,\tilde{\eta})$. Set

$$f_1=\frac{1}{2\pi}\frac{\partial}{\partial\tilde{\eta}}[\tilde{e}_1+\tilde{e}_2],\quad f_2=\frac{1}{2\pi}\frac{\partial}{\partial\tilde{\eta}}[\tilde{e}_2-\tilde{e}_1]. \tag{7.9}$$

(Note that the $\tilde{e}_j$ vanish at $\tilde{\eta}=0$, so the sum and difference can be recovered by integrating the f_j from $\tilde{\eta}=0$.) Adding and subtracting first derivatives with respect to $\tilde{\eta}$ in (7.5), we find that as functions of η the f_j satisfy

$$f_1=Kf_1+k,\quad f_2=-Kf_2+k,\quad k(\eta,\tilde{\eta})=\frac{1}{\pi}\frac{1}{1+(\eta+\tilde{\eta})^2}. \tag{7.10}$$

Note that k is the kernel of the operator K. Let F_j denote the integral operator that has kernel f_j and denote the identity operator Id. Then (7.10) is equivalent to the operator equations

$$(Id-K)(Id+F_1)=Id,\quad (Id+K)(Id-F_2)=Id. \tag{7.11}$$

The integral $\int_0^\infty f_j(\eta,\tilde{\eta})\,d\eta$ is the same as the function $F_j(1)(\tilde{\eta})$, where 1 denotes the function that is identically 1. It follows from (7.9) and (7.10) that

$$\frac{1}{2\pi}\int_0^\infty \tilde{e}_1(\eta,\tilde{\eta})d\eta+\tilde{\eta}=\frac{1}{2}\int_0^{\tilde{\eta}}\left[\int_0^\infty (f_1-f_2)+2\right]d\eta$$
$$=\frac{1}{2}\int_0^{\tilde{\eta}}[(F_1(1)+1)+(1-F_2(1))], \tag{7.12}$$

$$\frac{1}{2\pi}\int_0^\infty \tilde{e}_2(\eta,\tilde{\eta})d\eta = \frac{1}{2}\int_0^{\tilde{\eta}}\left[\int_0^\infty (f_1+f_2)\right]d\eta$$
$$= \frac{1}{2}\int_0^{\tilde{\eta}}[F_1(1)+F_2(1)]d\eta$$
$$= \frac{1}{2}\int_0^{\tilde{\eta}}[(F_1(1)+1)-(1-F_2(1))]\,d\eta.$$

In view of (7.11), $F_1(1)+1$ is the solution of $u=Ku+1$, while $1-F_2(1)$ is the solution of $v=-Kv+1$. By (6.17) and (6.16) these functions are $\rho^*_{0,t}$ and ρ_0, respectively. It follows that

$$\frac{1}{2\pi}\int_0^\infty \tilde{e}_1(\eta,\tilde{\eta})\,d\eta+\tilde{\eta} = \frac{1}{2}[\xi_0^*(\tilde{\eta})+\xi_0(\tilde{\eta})], \tag{7.13}$$

$$\frac{1}{2\pi}\int_0^\infty \tilde{e}_2(\eta,\tilde{\eta})\,d\eta = \frac{1}{2}[\xi_0^*(\tilde{\eta})-\xi_0(\tilde{\eta})]. \tag{7.14}$$

The momentum term $P(\tilde{\eta})$ in (7.6) is the difference of the two expressions (7.13) and (7.14). Similarly, the excitation energy term $E(\tilde{\eta})$ in (7.6) is the linear combination of (7.13) and (7.14) with coefficients $-c_1$ and c_2, respectively. This proves (7.7). The asymptotics (7.8) follow from (7.7) and the asymptotics (6.20) of ξ_0 and ξ_0^*. □

Exactly similar considerations apply to a single particle/hole ζ excitation, with the roles of e_1 and e_2 interchanged. In particular, the indices 1 and 2 in (7.9) should be interchanged throughout.

Compound excitations may be treated in exactly the same way. Consider the case of particle/hole excitations in η described by $\eta_{h,i}\geqslant 0$, $\eta_{p,i}\leqslant 0$, $i=1,...,r$, and in ζ described by $\zeta_{h,j}\geqslant 0$, $\zeta_{p,j}\leqslant 0$, $j=1,...,s$. Then in the infinite density limit the Bethe equations become

$$\begin{aligned}\tilde{e}_1(\eta) &= K(\tilde{e}_2)(\eta)+2\sum_{j=1}^{s}[\tan^{-1}(\eta+\zeta_{p,j})-\tan^{-1}(\eta+\zeta_{h,j})],\\ \tilde{e}_2(\eta) &= K(\tilde{e}_1)(\eta)+2\sum_{i=1}^{r}[\tan^{-1}(\eta+\eta_{p,i})-\tan^{-1}(\eta+\eta_{h,i})].\end{aligned} \tag{7.15}$$

This state is a superposition of corresponding single-particle/hole excited states. Therefore the associated momentum and excitation energy are given by

$$\begin{aligned}\delta P &= \sum_{i=1}^{r}[\xi_0(\eta_{p,i})-\xi_0(\eta_{h,i})]+\sum_{j=1}^{s}[\xi_0(\zeta_{h,j})-\xi_0(\zeta_{p,j})],\\ \delta E &= -\alpha\left(\sum_{i=1}^{r}[\xi_0^*(\eta_{p,i})-\xi_0^*(\eta_{h,i})]+\sum_{j=1}^{s}[\xi_0^*(\zeta_{p,j})-\xi_0^*(\zeta_{h,j})]\right)\\ &\quad +c_0\left(\sum_{i=1}^{r}[\xi_0(\eta_{p,i})-\xi_0(\eta_{h,i})]-\sum_{j=1}^{s}[\xi_0(\zeta_{p,j})-\xi_0(\zeta_{h,j})]\right).\end{aligned} \tag{7.16}$$

VIII. FINITE DENSITY PERTURBATIONS OF THE GROUND STATE

In this section we study large deviations from the ground state. Again we motivate the procedure by looking briefly at the finite length, finite particle case in the form of the integral

equations (6.10). Denote by $I_0 = J_0$ the data that corresponds to the CGS data (6.8). The CGS data uses all the integers from an interval with endpoints $\pm O(n/2)$ and we are temporarily applying a low-energy cutoff that corresponds to the upper end of this interval, so the data that corresponds to an admissible perturbation of the CGS satisfies $I,\ J \leqslant I_0$. The solution can be found by iteration from the integral equations (6.10). The monotonicity properties of the mapping imply that the solution is dominated by the CGS (η_0, ζ_0):

$$\eta(\xi) \leqslant \eta_0(\xi), \quad \zeta(\xi) \leqslant \zeta_0(\xi). \tag{8.1}$$

Moreover if the perturbation is to have finite relative energy density in the limit, we expect that $I,\ J \approx I_0$ for large ξ. Note also that the jumps in the data must be at least as great as the jumps $2\pi/L$ in I_0. We pass to the infinite length, infinite (total) density limit and look for a solution of an analogue of (6.12):

$$\begin{aligned} &\eta(\xi) - I(\xi) + \frac{1}{\pi}\int_0^\infty [\tan^{-1}(\eta(\xi)+\zeta(\xi')) - \tan^{-1}(\eta_0(0)+\zeta_0(\xi'))]\ d\xi'; \\ &\zeta(\xi) - J(\xi) + \frac{1}{\pi}\int_0^\infty [\tan^{-1}(\eta(\xi')+\zeta(\xi)) - \tan^{-1}(\xi_0(\xi')+\zeta_0(\xi))]\ d\xi'. \end{aligned} \tag{8.2}$$

According to the preceding discussion the natural assumptions on the data are

$$\frac{dI}{d\xi} \geqslant 1, \quad \frac{dJ}{d\xi} \geqslant 1; \quad I(\xi) \approx J(\xi) \approx \xi, \quad \xi \gg 0. \tag{8.3}$$

It can be shown that finite particle solutions whose data is such that the formal limits are (8.2) converge to solutions of (8.2). However, it is more convenient for what follows to write the equations for the particle density functions $\rho = d\xi/d\eta$ and $\sigma = d\xi/d\zeta$:

$$\begin{aligned} \rho(\eta)\cdot I'(\eta) &= 1 - \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sigma(\zeta)d\zeta}{1+(\eta+\zeta)^2} = 1 - (K\sigma)(\eta), \\ \sigma(\zeta)\cdot J'(\zeta) &= 1 - \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\rho(\eta)d\eta}{1+(\eta+\zeta)^2} = 1 - (K\rho)(\zeta), \end{aligned} \tag{8.4}$$

where

$$I'(\eta) = \frac{dI}{d\xi}(\xi(\eta)), \quad J'(\zeta) = \frac{dJ}{d\xi}(\xi(\zeta)).$$

We have removed the upper bound on particle energies, so the integrations in (8.4) need to be taken over the whole line. Now K has norm 1 as an operator in $L^\infty(\mathbb{R})$ so solvability of (8.4) becomes an issue. We rewrite (8.4) in the form

$$\rho = \Theta(1 - K\sigma), \quad \sigma = \Psi(1 - K\rho) \tag{8.4$'$}$$

and adopt the point of view that the functions Θ and Ψ are the data. For the ground state the data are $\Psi_0 = \Theta_0$, the Heaviside function: $\Theta_0(s) = 0$, $s < 0$; $\Theta_0 = 1$, $s \geqslant 0$. More generally we can express the conditions (8.3) and the assumption that our states are perturbations of the ground state by requiring

$$0 \leqslant \Theta, \quad \Psi \leqslant 1; \qquad \lim_{|\eta| \to \infty} (\Theta - \Theta_0) = \lim_{|\zeta| \to \infty} (\Psi - \Psi_0) = 0, \tag{8.5}$$

while (8.1) gives an implicit constraint on the data (Θ, Ψ) in terms of the solution. In fact (8.1) implies the reverse relation between the inverse functions, leading to

$$\int_{-\infty}^{\eta} (\rho - \rho_0) \geqslant 0, \quad \int_{-\infty}^{\zeta} (\sigma - \sigma_0) \geqslant 0, \quad \text{all} \quad \eta, \sigma. \tag{8.6}$$

Moreover the cutoff solution (η,ζ) and the corresponding CGS (η_0, ζ_0) at any given step were defined over the same ξ interval, so the relative particle density should be zero:

$$\int_{-\infty}^{\infty} (\rho - \rho_0) = 0 = \int_{-\infty}^{\infty} (\sigma - \sigma_0). \tag{8.7}$$

We show below that (8.7) is necessary for the relative energy density to be finite.

Theorem 8.1: *Under assumption* (8.5) *the equations (8.4) have a unique solution. For the solution we have*

$$\begin{gathered} 0 \leqslant \rho \leqslant 1, \quad 0 \leqslant \sigma \leqslant 1, \\ \lim_{\eta \to \infty} \rho(\eta) = \lim_{\zeta \to \infty} \sigma(\zeta) = 1, \qquad \lim_{\eta \to -\infty} \rho(\eta) = \lim_{\zeta \to -\infty} \sigma(\zeta) = 0. \end{gathered} \tag{8.8}$$

Proof: We use Θ and Ψ also to denote the operations of multiplication by the corresponding functions. Then the equations are

$$\rho + \Theta K \sigma = \Theta, \quad \sigma + \Psi K \rho = \Psi.$$

We apply the operator $\mathrm{Id} - \Psi K$ to the first equation and $\mathrm{Id} - \Theta K$ to the second and reorganize to obtain

$$(\mathrm{Id} - \Theta K \Psi K)\rho = \Theta - \Theta K \Psi, \quad (\mathrm{Id} - \Psi K \Theta K)\sigma = \Psi - \Psi K \Theta. \tag{8.9}$$

It is enough to show that $\Theta K \Psi$ and $\Psi K \Theta$ have norm <1. The assumptions on Θ and Ψ imply easily that the integral of the kernel of either operator against either variable is bounded away from 1, so the operators have norm <1 in any L^p, $1 \leqslant p \leqslant \infty$. The asymptotics follow easily from the assumptions about Θ and Ψ, together with the trivial estimaes

$$0 \leqslant \rho \leqslant \Theta, \quad 0 \leqslant \sigma \leqslant \Psi. \qquad \square$$

One must make stronger assumptions to ensure finiteness of the integrals (8.6) and (8.7) and the relative energy density

$$E = -\frac{c_1}{2\pi} \int_{-\infty}^{\infty} (\rho - \rho_0)\, \eta \, d\eta + \frac{c_2}{2\pi} \int_{-\infty}^{\infty} (\sigma - \sigma_0)\, \zeta \, d\zeta. \tag{8.10}$$

Theorem 8.2: *The integrals (8.6) and (8.7) are finite if* $(\Theta - \Theta_0)$ *and* $(\Psi - \Psi_0)$ *are integrable.*

If $\eta(\Theta_0 - \Theta)$ *and* $\zeta(\Psi_0 - \Psi)$ *are integrable, then* E *is finite if and only if (8.7) holds. If so, then* $E \geqslant 0$.

Proof: Note that

$$\rho - \rho_0 = (\Theta - \Theta_0)(1 - K\sigma_0) - \Theta K(\sigma - \sigma_0), \tag{8.11}$$

with a similar equation for $\sigma-\sigma_0$. Analyzing this system as above, we see that integrability of $\rho-\rho_0$ and $\sigma-\sigma_0$ follows from integrability of the data in the analogue of (8.9), which follows in turn from integrability of $\Theta-\Theta_0$ and $\Psi-\Psi_0$.

Note that integrability of $\eta(\rho_0-\rho)$ at $-\infty$ follows immediately from (8.11) and integrability of $\eta(\Theta_0-\Theta)$. At $+\infty$ we observe from (8.11) that the necessary and sufficient condition is the integrability of $\eta K(\sigma-\sigma_0)$. Assume that $\zeta^2(\Psi-\Psi_0)$ is integrable at $-\infty$. Then $\zeta^2(\sigma-\sigma_0)$ is also integrable at $-\infty$ and Lebesgue's Dominated Convergence Theorem implies that

$$\lim_{\eta\to+\infty} K(\sigma-\sigma_0)=\int_{-\infty}^{\infty}(\sigma-\sigma_0)\,d\zeta. \tag{8.12}$$

This proves the necessity of the conditions (8.7).

Conversely one can deduce the integrability of $\eta K(\sigma-\sigma_0)$ and $\zeta K(\rho-\rho_0)$ from conditions (8.7) together with integrability of $\eta(\rho_0-\rho)$ and $\zeta(\sigma_0-\sigma)$. To see this, note that the mean-value zero condition allows us to replace the kernel of K with the kernel

$$\widetilde{k}(\eta,\zeta)=\frac{1}{\pi}\left[\frac{1}{1+(\eta+\zeta)^2}-\frac{1}{1+\eta^2}\right]. \tag{8.13}$$

The corresponding operator takes $L^1(\mathbb{R},(1+|z|)d\zeta)$ to $L^1(\mathbb{R},(1+|\eta|)d\eta)$, since we can multiply $\widetilde{k}$ by $(1+|\eta|)/(1+|z|)$, take the absolute value, and estimate the integral with respect to η.

Finally, to show positivity of the relative energy density we let g_1 and g_2 denote the integrals in (8.6), so that

$$\frac{dg_1}{d\eta}=\rho,\quad \frac{dg_2}{d\zeta}=\sigma,\quad g_j\geqslant 0,\quad g_j(\pm\infty)=0. \tag{8.14}$$

Integrating by parts in (8.10) we find that

$$E=\frac{c_1}{2\pi}\int_{-\infty}^{\infty}g_1(\eta)\,d\eta-\frac{c_2}{2\pi}\int_{-\infty}^{\infty}g_2(\zeta)\,d\zeta\geqslant 0.$$

□

IX. THERMODYNAMICS

We consider now the thermodynamics of the finite density perturbations of the ground state that were constructed in the preceding section. We follow, with appropriate modifications, Yang and Yang[5] and Ohkuma.[7] Consider first the case of finite L, a state with a very large number of particles, and data I, J as in (8.1). Define

$$\begin{aligned} h(s)&=s-\frac{1}{\pi}\int_0^{\Delta}[\tan^{-1}(s+\zeta(\xi'))-\tan^{-1}(\eta(0)+\zeta(\xi'))]\,d\xi',\quad s\geqslant\eta(0),\\ k(t)&=t-\frac{1}{\pi}\int_0^{\Delta}[\tan^{-1}(\eta(\xi')+t)-\tan^{-1}(\eta(\xi')+\zeta(0))]\,d\xi',\quad t\geqslant\zeta(0). \end{aligned} \tag{9.1}$$

These are strictly increasing functions with the properties that at the (discrete) values taken by the data the values are

$$h(\eta(\xi))=I(\xi),\quad k(\zeta(\xi))=J(\xi).$$

The number m of η particles from our state with momenta in a small interval $(\eta, \eta+d\eta)$ is proportional to the length of the corresponding ξ interval, i.e., if $\eta=\eta(\xi)$ and $\eta+d\eta=\eta(\xi+d\xi)$, then

$$m\approx\frac{L}{2\pi}\,d\xi.$$

The function I has jumps that are positive integer multiples of $L/2\pi$, so

$$m\leqslant n=\frac{L}{2\pi}\,(I(\xi+d\xi)-I(\xi)).$$

If data I at the m jump sites in the interval $(\xi,\xi+d\xi)$ is changed in such a way that I remains strictly increasing and the values of $LI/2\pi$ are integers, the effect on the total energy of the corresponding state will be negligible: $O((d\eta)^2)$. The possible sites are indexed by the n integers between $LI(\xi)/2\pi$ and $L(I\xi)/2\pi$. Thus the contribution to entropy from the interval is $\log\binom{n}{m}$.

In the infinite length and infinite density limits the density functions for the numbers n and m above have densities with respect to the momentum η given by $\rho=d\xi/d\eta$ and $\rho_t=dh/d\eta$, respectively. As before ρ is the particle density. We take $\rho_h=\rho_t-\rho$ to be the *hole density* and ρ_t to be the (formal) total density. Similar considerations apply to ζ particles and give particle density σ and hole density $\sigma_h=\sigma_t-\sigma$. In the notation of Sec. VIII the relations among these functions are

$$\begin{gathered}\rho=\Theta\rho_t,\quad \rho_h=(1-\Theta)\rho_t,\quad \sigma=\Psi\sigma_t,\quad \sigma_h=(1-\Psi)\sigma_t,\\ \rho_t=1-K\sigma=1-K\Psi\sigma_t,\quad \sigma_t=1-K\rho=1-K\Theta\rho_t.\end{gathered}\tag{9.2}$$

Returning to the interval $(\xi,\xi+d\xi)$, let $(\eta,\eta+d\eta)$ denote its image under h^{-1}. According to the preceding argument and Stirling's formula, the entropy density over $(\eta+d\eta)$ is approximately

$$\begin{aligned}\frac{L}{2\pi}\,[\rho_t\log\rho_t-\rho\log\rho-\rho_h\log\rho_h]&=\frac{L}{2\pi}\,[\rho\log(\rho_t/\rho)+\rho_h\log(\rho_t/\rho_h)]\\ &=-\frac{L}{2\pi}\,\rho_t[\Theta\log\Theta+(1-\Theta)\log(1-\Theta)].\end{aligned}\tag{9.3}$$

A similar formula holds for particles of the second type. Thus in the infinite length, infinite density limit the entropy per unit length is

$$\begin{aligned}S=&\frac{1}{2\pi}\int_{-\infty}^{\infty}[\rho_t\log\rho_t-\rho\log\rho-\rho_h\log\rho_h]\,d\eta\\ &+\frac{1}{2\pi}\int_{-\infty}^{\infty}[\sigma_t\log\sigma_t-\sigma\log\sigma-\sigma_h\log\sigma_h]\,d\zeta.\end{aligned}\tag{9.4}$$

The entropy density for the ground state is zero.

At a given finite temperature T the free energy density of a perturbation (ρ,σ) of the ground state relative to that of the ground state is

$$F_T(\rho,\sigma)=E-TS,\tag{9.5}$$

where the relative energy density E is given by (8.9). In order for E to be finite, we showed that the relative particle densities must vanish:

$$\int_{-\infty}^{\infty}(\rho-\rho_0)\ d\eta=0=\int_{-\infty}^{\infty}(\sigma-\sigma_0)\ d\zeta. \tag{9.6}$$

Thus the equilibrium solution should minimize F subject to the constraints (9.6).

We consider briefly the formal argument for an equilibrium. Equations (9.2) relate variations in ρ_t, ρ_h, σ_t, and σ_h to variations in the particle densities ρ and Θ:

$$\begin{aligned}\dot\rho t&=-K\dot\sigma, \quad \dot\rho_h=-\dot\rho-K\dot\sigma,\\ \dot\sigma_t&=-K\dot\rho, \quad \dot\sigma_h=-\dot\sigma-K\dot\rho.\end{aligned} \tag{9.7}$$

The constraints (9.6) require

$$\int_{-\infty}^{\infty}\dot\rho\ d\eta=0=\int_{-\infty}^{\infty}\dot\sigma\ d\zeta. \tag{9.8}$$

Taking into account (9.7) and the symmetry of the operator K, we can express the variation in the free energy as

$$\begin{aligned}\dot F=&\int_{-\infty}^{\infty}\dot\rho[-c_1\eta+T\ \log(\rho/\rho_h)+TK\ \log(\sigma_t/\sigma_h)]\ d\eta\\ &+\int_{-\infty}^{\infty}\dot\sigma[c_2\zeta+T\ \log(\sigma/\sigma_h)+TK\ \log(\rho_t/\rho_h)]\ d\zeta.\end{aligned} \tag{9.9}$$

Following Yang and Yang,[5] one can introduce functions ϵ_j:

$$e^{-\epsilon_1/T}=\rho/\rho_h, \quad e^{-\epsilon_2/T}=\sigma/\sigma_h, \tag{9.10}$$

and derive Euler–Lagrange equations in the form

$$\begin{aligned}\epsilon_1&=A_1T-c_1\eta+TK\ \log(1+e^{-e_2/T}),\\ \epsilon_2&=A_2T+c_2\zeta+TK\ \log(1+e^{-e_1/T}),\end{aligned} \tag{9.11}$$

where the A_j are the Lagrange multipliers that correspond to (9.8).

As noted in the Introduction, there are two difficulties with this argument. First, Eqs. (9.11) have no solution: taking derivatives one sees that the ϵ_j should be decreasing functions, so the functions $\log(1+e^{-e_j/T})$ are increasing and bounded above, but they are taken by K to decreasing functions and thus one needs $\epsilon_1\leqslant -c_1\eta+O(1)$ at $+\infty$. Therefore $\log(1+e^{-e_1/T})$ is non-negative and grows like η at $+\infty$, so $K\log(1+e^{-e_1/T})$ cannot be defined. Second, it is not clear that the free energy is bounded below.

The first difficulty is easily dealt with. We may assume that $\dot r$ and $\dot s$ are test functions, and (9.9) simply means that they are derivatives of test functions. Therefore the condition $\dot F=0$ simply means that the derivatives of the two bracketed terms vanish. Now K anticommutes with derivation, so the correct form of (9.11) is

$$\frac{d\epsilon_1}{d\eta}=-c_1+K\left(\Psi\ \frac{d\epsilon_2}{d\zeta}\right), \quad \frac{d\epsilon_2}{d\zeta}=c_2+K\left(\Theta\ \frac{d\epsilon_1}{d\eta}\right). \tag{9.12}$$

Combining this with

$$\Theta = \frac{\epsilon^{-\epsilon_1/T}}{1+e^{-e_1/T}}, \quad \Psi = \frac{e^{-\epsilon_2/T}}{1+e^{-e_2/T}} \tag{9.13}$$

yields a well-behaved system of equations for the thermal equilibrium particle distribution $(\rho,\sigma)=(\rho^{(T)},\sigma^{(T)})$.

To deal with the second difficulty we prove the following.

Theorem 9.1: *The free energy functional F_T is convex and bounded below by $-CT^2$, where C is a constant that depends only on the velocities c_j.*

Proof: From (9.9), the second derivative of the free energy for the variation $(\dot\rho,\dot\sigma)$ is

$$\begin{aligned} &T\int_{-\infty}^{\infty}\dot\rho\left[\frac{\dot\rho}{\rho}-\frac{\dot\rho_h}{\rho_h}+K\left(\frac{\dot\sigma_t}{\sigma_t}-\frac{\dot\sigma_h}{\sigma_h}\right)\right]d\eta+T\int_{-\infty}^{\infty}\dot\sigma\left[\frac{\dot\sigma}{\sigma}-\frac{\dot\sigma_h}{\sigma_h}+K\left(\frac{\dot\rho_t}{\rho_t}-\frac{\dot\rho_h}{\rho_h}\right)\right]d\zeta \\ &\quad=T\int_{-\infty}^{\infty}\left[\dot\rho\left(\frac{\dot\rho}{\rho}-\frac{\dot\rho_h}{\rho_h}\right)+(K\dot\sigma)\left(\frac{\dot\rho_t}{\rho_t}-\frac{\dot\rho_h}{\rho_h}\right)\right]d\eta+T\int_{-\infty}^{\infty}\left[\dot\sigma\left(\frac{\dot\sigma}{\sigma}-\frac{\dot\sigma_h}{\sigma_h}\right)+(K\dot\rho)\left(\frac{\dot\sigma_t}{\sigma_t}-\frac{\dot\sigma_h}{\sigma_h}\right)\right]dz. \end{aligned} \tag{9.14}$$

We use (9.7) and reorganize to find that the second derivative is

$$T\int_{-\infty}^{\infty}\left(\frac{\sqrt{\rho_t}}{\sqrt{\rho\rho_h}}\,\dot\rho+\frac{\sqrt{\rho}}{\sqrt{\rho_t\rho_h}}\,K\dot\sigma\right)^2 d\eta+T\int_{-\infty}^{\infty}\left(\frac{\sqrt{\sigma_t}}{\sqrt{\sigma\sigma_h}}\,\dot\sigma+\frac{\sqrt{\sigma}}{\sqrt{\sigma_t\sigma_h}}\,K\dot\rho\right)^2 d\zeta. \tag{9.15}$$

To prove boundedness we introduce the function

$$G(s)=G(1-s)=-s\,\log\,s-(1-s)\log(1-s), \quad 0\leqslant s\leqslant 1.$$

Then $G(\Theta)=G(|\Theta_0-\Theta|)$ and the entropy density can be written

$$2\pi S(\rho,\sigma)=\int_{-\infty}^{\infty}\rho_t G(\Theta)\,d\eta+\int_{-\infty}^{\infty}\sigma_t G(\Psi)\,d\zeta. \tag{9.16}$$

For the first integral we partition the line into three sets:

$$\begin{aligned} I_1&=\{\eta: G(\Theta)\leqslant|\eta(\Theta_0-\Theta)|/\delta\}, \\ I_2&=\{\eta:|\Theta_0-\Theta|\geqslant 1/2, G(\Theta)>|\eta(\Theta_0-\Theta)|/\delta\}, \\ I_3&=\{\eta:|\Theta_0-\Theta|<1/2, G(\Theta)>|\eta(\Theta_0-\Theta)|/\delta\}, \end{aligned} \tag{9.17}$$

where the constant δ is to be chosen. Note that $0\leqslant G(t)\leqslant\log 2$, so $\eta\in I_2$ implies $|\eta|<2\delta\log 2$. Also

$$-\log\,t\geqslant -2t\,\log\,t\geqslant G(t), \quad 0\leqslant t\leqslant\tfrac{1}{2},$$

so $\eta\in I_3$ implies $|\Theta_0-t|<\exp(-|\eta|/\delta)$. Now $-t\,\log\,t$ is increasing on the interval $[0,1/e]$, so we conclude that $\eta\in I_3$ implies that either $|\eta|/\delta<1$ or

$$G(\Theta)=G(|\Theta_0-\Theta|)\leqslant -2|\Theta_0-\Theta|\log|\Theta_0-\Theta|\leqslant 2\,\frac{|\eta|}{\delta}\,e^{-|\eta|/\delta}.$$

Since $1<2\log 2$ we can combine these estimates and obtain

$$\int_{-\infty}^{\infty}\rho_t G(\Theta)\ d\eta \leqslant \frac{1}{\delta}\int_{-\infty}^{\infty}\rho_t(\Theta_0-\Theta)\,\eta\ d\eta+4\delta\log 2+2\int_{-\infty}^{\infty}\frac{|\eta|}{\delta}\,e^{-|\eta|/\delta}\,d\eta$$

$$=\int_{-\infty}^{\infty}\rho_t(\Theta_0-\Theta)\,\eta\ d\eta+4\delta\log 2+4\delta. \tag{9.18}$$

We claim that

$$\int_{-\infty}^{\infty}\rho_t(\Theta_0-\Theta)\,\eta\ d\eta\leqslant\int_{-\infty}^{\infty}(\rho_0-\rho)\,\eta\ d\eta+\int_{-\infty}^{\infty}(\sigma_0-\sigma)\zeta\ d\zeta. \tag{9.19}$$

In fact (8.11) can be rewritten

$$\rho_t(\Theta_0-\Theta)=\rho-\rho_0+\Theta_0 K(\rho-\rho_0). \tag{9.20}$$

Using $\langle,\rangle$ to denote the standard inner product, we want to prove that

$$\int_{-\infty}^{\infty}\Theta_0 K(\sigma-\sigma_0)\,\eta\ d\eta=\langle K(\sigma-\sigma_0),\Theta_0\eta\rangle\leqslant\langle\sigma_0-\sigma,\zeta\rangle=-\langle Dg_2,z\rangle=\langle g_2,1\rangle, \tag{9.21}$$

where D denotes differentiation and g_2 is the function in (8.14). We use the fact that derivation anticommutes with K and integrate by parts to obtain

$$\langle K(\sigma-\sigma_0),\Theta_0\eta\rangle=\langle KDg_2,\Theta_0\eta\rangle=-\langle DKg_2,\Theta_0\eta\rangle=\langle Kg_2,D(\Theta_0\eta)\rangle=\langle g_2,K\Theta_0\rangle.$$

Now $0\leqslant K\theta_0\leqslant 1$ and $g_2\geqslant 0$, so this proves (9.19). Using (9.18), (9.19), and the analogous results for σ, we obtain

$$F_T(\rho,\sigma)\leqslant\left(c_1-\frac{2T}{\delta}\right)\int_{-\infty}^{\infty}(\rho_0-\rho)\,\eta\ d\eta+4\delta(1+\log 2)+\left(-c_2-\frac{2T}{\delta}\right)$$

$$\times\int_{-\infty}^{\infty}(\sigma_0-\sigma)\zeta\ d\zeta+4\delta(1+\log 2). \tag{9.22}$$

We take $\delta=2T/(c_1-c_2)$ and obtain $C=8(1+\log 2)/(c_1-c_2)$. □

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation Grants No. DMS-9423746 and No. DMS-9501233.

[1] M. Wadati and K. Ohkuma, ''Bethe States for the Quantum Three Wave Interaction Equation,'' J. Phys. Soc. Jpn. **53**, 1229–1237 (1984).
[2] P. P. Kulish and N. Reshetikhin, ''Diagonalization of GL(N) invariant transfer matrices and quantum N-wave system (Lee Model),'' J. Phys. A **16**, L591–L596 (1983).
[3] V. E. Korepin, ''Direct calculation of the S Matrix in the massive Thirring Model,'' Theor. Math. Phys. **41**, 168–189 (1979).
[4] H. Bergknoff and H. B. Thacker, ''Structure and solution of the massive Thirring model,'' Phys. Rev. D **19**, 3666–3681 (1979).
[5] C. N. Yang and C. P. Yang, ''Thermodynamics of a one-dimensional system of bosons with repulsive delta-function interaction,'' J. Math. Phys. **10**, 1115–1122 (1969).
[6] K. Ohkuma and M. Wadati, ''Quantum Three Wave Interaction Models,'' J. Phys. Soc. Jpn. **53**, 2899–2907 (1984).
[7] K. Ohkuma, ''Thermodynamics of the Quantum Three Wave Interaction Model,'' J. Phys. Soc. Jpn. **54**, 2817–2828 (1985).
[8] E. H. Lieb and W. Liniger, ''Exact analysis of an interacting Bose gas, I. The general solution and the ground state,'' Phys. Rev. **130**, 1605–1615 (1963).

[9] V. E. Korepin, N. M. Bogoliubov, and V. E. Izergin, *Quantum Inverse Scattering Method and Correlation Functions* (Cambridge U. P., Cambridge, 1993).
[10] H. Bethe, ''Zur Theorie der Metalle. I. Eigenwerte und Eigenfunktionen der linearen Atomkette,'' Z. Phys. **71**, 205–226 (1931).
[11] E. H. Lieb, ''Exact analysis of an interacting Bose gas, II. The excitation spectrum,'' Phys. Rev. **130**, 1616–1624 (1963).
[12] C. N. Yang and C. P. Yang, ''One-dimensional chain of anisotropic spin-spin interactions I. Proof of Bethe's hypothesis for ground state in finite system,'' Phys. Rev. **150**, 321–327 (1966).
[13] V. E. Zakharov and S. V. Manakov, ''Resonant interactions of wave packets in nonlinear media,'' JETP **18**, 243–245 (1973).
[14] R. Beals and D. H. Sattinger, ''On the Complete Integrability of Completely Integrable Systems,'' Commun. Math. Phys. **138**, 409–436 (1991).
[15] R. Beals, D. H. Sattinger, and E. Williams, ''Quantization of the three wave interaction,'' Technical Report, www.math.umn.edu/~dhs

Proof of Serban's conjecture

M. C. Bergère[a)]
CEA-Saclay, Service de Physique Théorique, F-91191 Gif sur Yvette Cedex, France

(Received 30 April 1997; accepted for publication 8 September 1997)

We prove Serban's conjecture which greatly simplifies the expression of the advanced single-particle Green function in the Calogero–Sutherland model. The importance of proving this conjecture is that it reorganizes the form factor in terms of two dimensional Coulomb gas correlators and confirms the possible existence of a bosonization procedure for this system. © *1998 American Institute of Physics.* [S0022-2488(97)02912-5]

I. INTRODUCTION

The Calogero–Sutherland model[1] attracted much attention during the last few years, due to its connection to random matrix theory, fractional statistics in one dimension, conformal field theory, etc. This model is exactly solvable and its solution is known to a large extent. The corresponding Hamiltonian

$$H=-\frac{1}{2}\sum_{i=1}^{N}\frac{d^2}{dx_i^2}+\beta(\beta-1)\frac{\pi^2}{L^2}\sum_{i<j}\frac{1}{\sin^2\left(\frac{\pi}{L}(x_i-x_j)\right)} \tag{1.1}$$

describes the motion of N particles along a circle of length L. The eigenstates of this Hamiltonian can be expressed[2] in terms of the so-called Jack polynomials in the variables $z_i=\exp(2i\pi x_i/L)$. For $\beta=0$ or 1, the particles behave like free spinless bosons or fermions, and for rational β the model describes particles with a fractional statistic (excitations from the Fermi sea with kq quasiparticles and kp quasiholes if k is a positive integer and $\beta=p/q$).

The two points correlation functions were calculated using two different methods. The first one generalizes Dyson's work on the correlations of the eigenvalues of random matrices[3] which led to explicit expressions for the static density–density correlations for $\beta=1/2,1,2$. For these values of β, Simons, Lee, and Al'tshuler in Ref. 3, using supersymmetric techniques initiated by Efetov,[3] calculated the dynamical density–density correlations; then Haldane and Zirnbauer obtained in Ref. 4, the retarded (for $\beta=2$) and the advanced (for $\beta=1/2,1,2$) single-particle Green functions.

The other method works for any rational β and uses the properties of Jack polynomials. The dynamical density–density correlations and the retarded single-particle Green functions were obtained for integer β by Lesage, Pasquier, and Serban in Ref. 5 and for rational β by Ha in Ref. 6. The correlations functions can be expressed in terms of ''form factors'' for some operator A

$$|F|^2=|\langle v_1,\ldots,v_p;w_1,\ldots,w_q|A|0\rangle|^2, \tag{1.2}$$

where the v's and the w's are the rapidities of the quasiholes and of the quasiparticles, respectively. F can be decomposed as a product

$$F=F^{(0)}\times F_\beta^{(1)}.$$

The expression $F^{(0)}$ is a purely statistical contribution given by

[a)]Membre du C.N.R.S.

0022-2488/98/39(1)/30/17/$15.00
© 1998 American Institute of Physics

$$|F^{(0)}|^2=\frac{\Pi_{i<j}|v_i-v_j|^{2/\beta}\Pi_{k<l}|w_k-w_l|^{2\beta}}{\Pi_i(v_F^2-v_i^2)^{1-1/\beta}\Pi_k(w_k^2-v_F^2)^{1-\beta}\Pi_{i,k}(v_i-w_k)^2}, \tag{1.3}$$

where v_F is half the size (in rapidity) of the Fermi sea. In the case of the advanced Green function ($A=\Psi^+$) it was found in Ref. 4 that there are two contributions: one with one quasiparticle (zero quasihole) and one with $q+1$ quasiparticles and p quasiholes (if $\beta=p/q$). For this second contribution, $F_\beta^{(1)}$ was found in Ref. 4, for $\beta=2,1,1/2$, to be

$$F_2^{(1)}=C_2\,\frac{d}{dz}\left(\frac{(z-v_1)(z-v_2)}{(z-w_1)^2}\right)_{z=w_0}, \quad F_1^{(1)}=0, \tag{1.4}$$

$$F_{1/2}^{(1)}=C_{1/2}\int_{w_0}^{\infty}\frac{dz}{\sqrt{z-w_0}}\left(\frac{z-v_1}{\sqrt{(z-w_1)(z-w_2)}}-1\right).$$

Moreover, the authors proposed the following conjecture: for any positive integer $\beta=p$,

$$F_p^{(1)}=C_p\left(\frac{d}{dz}\right)^{p-1}\left\{\frac{\Pi_{i=1}^p(z-v_i)}{(z-w_1)^p}\right\}_{z=w_0}. \tag{1.5}$$

Recently, the single particle advanced Green function have been calculated for all rational β by Serban, Lesage, and Pasquier.[7] Their method was to sum the corresponding combination of Jack polynomials over all Young tableau with r (w_0 in the thermodynamic limit) columns of length $N-1$, p legs and q arms, and to take the thermodynamic limit, that is N going to infinity. They obtained (1.3) for $F^{(0)}$, and $F_\beta^{(1)}$ as

$$\begin{aligned}F_\beta^{(1)}=&\frac{\Pi_{i=0}^q\Pi_{j=1}^p(w_i-v_j)^\beta}{\Pi_{i<j}(v_i-v_j)\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)^{2\beta-1}}\\&\times\int_{w_q}^{w_{q-1}}d\xi_{q-1}\cdots\int_{w_2}^{w_1}d\xi_1\prod_{i<j}(\xi_i-\xi_j)\prod_{i=1}^{q-1}\prod_{j=0}^{q}(\xi_i-w_j)^{\beta-1}\\&\times\partial_{v_1}\cdots\partial_{v_p}\left[\prod_{i<j}(v_i-v_j)\prod_{j=1}^{p}\left(\prod_{i=0}^{q}(w_i-v_j)^{1-\beta}\prod_{i=1}^{q-1}(\xi_i-v_j)^{-1}\right)\right].\end{aligned} \tag{1.6}$$

Then, Serban conjectured that this expression reduces to

$$F_\beta^{(1)}=(\beta-1)\,\frac{[\Gamma(\beta)]^q}{2i\pi}\oint_{C_w}dz\,\frac{\Pi_{i=1}^p(v_i-z)}{\Pi_{j=0}^q(w_j-z)^\beta}, \tag{1.7}$$

where the contour C_w surrounds the points $w_1,\ldots,w_q$.

The conjecture was proven for β integer (that is Haldane and Zirnbauer conjecture (1.5)), and the result (1.4) for $\beta=1/2$ has been obtained from (1.7), as well as the numerical equality between (1.6) and (1.7) for $\beta=1/2$ and 3/2.

In this publication, we prove Serban's conjecture, not only for $\beta=p/q$, but for a set of different β_i attached to each rapidity w_i. The importance of proving this conjecture is due to the interpretation of the result in terms of basic conformal operators. It was already pointed out by Khveshchenko[8] that the density–density correlation function and that the one particle retarded Green function could be simply re-expressed in terms of 2D Coulomb gas correlators

$$\langle V_{a_1}(z_1)\cdots V_{a_n}(z_n)\rangle=\prod_{i<j}(z_i-z_j)^{a_ia_j}\quad\text{if}\quad\sum_{i=1}^{n}a_i=0 \tag{1.8}$$

and zero otherwise. Equation (1.7) proves, in the more complicate situation of the advanced Green function, that the same property is true

$$F=F^{(0)}\times F_{\beta}^{(1)}=\frac{(\beta-1)[\Gamma(\beta)]^q}{\Pi_i(v_F^2-v_i^2)^{(\beta-1)/2\beta}\Pi_k(w_k^2-v_F^2)^{(1-\beta)/2}}$$
$$\times\frac{1}{2i\pi}\oint_{C_w}dz\left\langle V_{-\sqrt{\beta}}(z)\prod_{i=1}^{p}V_{-1/\sqrt{\beta}}(v_i)\prod_{k=0}^{q}V_{\sqrt{\beta}}(w_k)\right\rangle, \tag{1.9}$$

where a charge $\sqrt{\beta}$ is given to the quasiparticle operator, $-1/\sqrt{\beta}$ to the quasihole operator, and $-\sqrt{\beta}$ to the screening operator introduced in order to have conservation of the charge. At the present time, we do not know what has to be done with the terms coming from the Fermi sea (v_F dependent) but this might be overcome naturally if we later develop a bozonisation procedure of the Calogero–Sutherland model. Several arguments plead in favour of such a construction and specially in Ref. 9, it is shown that the Jack polynomials are the singular vectors of the Virasoro and of the W_N algebra. It is, however, surprising to see in (1.9) that the correlators V_a are not taken in complexified space-time points but in the rapidities. Let us finally mention that the relation between (1.6) and (1.7) can certainly be q-deformed if we work with the relativistic Ruijsenaars–Schneider model,[10] where the Jack polynomials are q-deformed into the Macdonald polynomials.[11]

II. PROOF OF THE CONJECTURE

Given a set of $(q+1)$ variables β_i (large enough to make all integrals convergent), and two sets of variables, $(q+1)$ variables w_i and p variables v_i, we define C as

$$C=\int_{w_q}^{w_{q-1}}d\xi_{q-1}\cdots\int_{w_2}^{w_1}d\xi_1\prod_{i<j}(\xi_i-\xi_j)\prod_{i=1}^{q-1}\prod_{j=0}^{q}|\xi_i-w_j|^{\beta_j-1}$$
$$\times\frac{\partial_{v_1}\cdots\partial_{v_p}[\Pi_{j=0}^{q}\Pi_{k=1}^{p}(w_j-v_k)^{1-\beta_j}\Pi_{i=1}^{q-1}\Pi_{k=1}^{p}(\xi_i-v_k)^{-1}\Pi_{k<l}(v_k-v_l)]}{\Pi_{j=0}^{q}\Pi_{k=1}^{p}(w_j-v_k)^{1-\beta_j}\Pi_{k<l}(v_k-v_l)}. \tag{2.1}$$

Our result is

$$C=\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}\frac{\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-1}}{\Pi_{j=0}^{q}\Pi_{k=1}^{p}(w_j-v_k)}$$
$$\times\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{k=1}^{p}(z-v_k)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}, \tag{2.2}$$

where $w_0>w_1$ and C_0 surrounds the cut $[w_0,+\infty]$; we suppose that $\Sigma_{j=0}^{q}\beta_j>p+1$ to ensure the convergence of the integral in (2.2), otherwise subtractions has to be considered.

Here, we indicate how to find the subtraction terms in the integrand (2.2) in the special case where $\Sigma_{j=1}^{q}\beta_j=r$ is an integer and when $\beta_0+r\leqslant p+1$; we proceed by analytic continuation in β_0 so that the expression (2.2) remains equal to (2.1). Indeed, if $\Sigma_{j=1}^{q}\beta_j$ is an integer, we may organize the branch cuts of $\Pi_{j=1}^{q}(z-w_j)^{\beta_j}$ so that they lie between w_1 and w_q. Then, for β_0 large enough to ensure the convergence of (2.2), the integral dz is nothing (up to a sign) but the same integral on a contour C_w which surrounds the points $w_1,\ldots,w_q$ [see Eq. (1.7)]. When β_0 decreases in such a way that $\beta_0+r\leqslant p+1$, the integral on C_w still exists and is equal to (2.1) by analytic continuation. If we wish, in that case, to come back to the contour C_0, we must make subtractions which ensure the convergence when $z\to\infty$ and which do not destroy the equality between (2.2) and (2.1). Clearly, if we subtract to

$$f_{\{\beta\}}(v_k,w_j,z)=\frac{\Pi_{k=1}^{p}(z-v_k)}{\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}$$

a finite number of terms defined as the Taylor expansion of $f_{\{\beta\}}$ around the point $v_k=w_j=w_0$ (the first subtracted term being $(z-w_0)^{p-r}$), the subtracted terms contribute nothing to the integral (2.2) on the contour C_w, and consequently, let C unchanged and equal to (2.1); however, these subtractions make the integral dz convergent at infinity and they allow the contour C_w to be transformed back into the contour C_0.

We note that if all the variables β_j are equal to p/q, we have

$$\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}=(\beta-1)[\Gamma(\beta)]^q, \tag{2.3}$$

which appears in (1.7). Also, in that case, and if $p/q\leqslant 1$, one subtraction only is needed to make the integral (2.2) convergent; the integrand in (2.2) becomes

$$\frac{1}{(w_0-z)^{\beta}}\left\{\frac{\Pi_{k=1}^{p}(z-v_k)}{\Pi_{j=1}^{q}(z-w_j)^{\beta}}-1\right\}$$

which is for instance the case for $\beta=1/2$ in (1.4).

The proof of the result (2.1–2) requires several steps of different nature; in order to have an easier reading, the main part of the proof is written in this section while four appendices are devoted to technical details.

The expression C in (2.1) is made of two parts: on the first line, the integrals $d\xi$, on the second line the derivatives ∂_v, the factorization being prevented by the $(\xi-v)^{-1}$ terms. Appendix A generalizes Cauchy's identity in order to eliminate the $(\xi-v)^{-1}$ terms. C becomes a sum of terms, each of them being factorized into $d\xi$ integrals and ∂_v derivatives without $(\xi-v)^{-1}$ terms. Appendix B generalizes a result by Dixon[12] who performs a change of variables which makes obvious the calculation of the integrals $d\xi$; the generalization consists in adding the variable w_0 to Dixon's result which generates a contour integral $\oint dz$. These two Appendices are sufficient to prove the conjecture in the case $(q-1)\geqslant p$. However, the case $(q-1)\leqslant p$ is more difficult because we have extra ∂_v derivatives to perform over terms which do not contain $(\xi-v)^{-1}$ anymore. Appendix C is devoted to the calculation and to the properties of these ∂_v derivatives. Finally, Appendix D, which also, is only needed in the case $q-1\leqslant p$, develops the properties of the integrals

$$I_P=\frac{1}{2i\pi}\oint_{C_0}dz\,\frac{P(z)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}, \tag{2.4}$$

where $P(z)$ is a polynomial in z and where C_0 surrounds the cut $[w_0,+\infty]$.

Let us mention that the proof is written for a set of generic values of the variables v_i without care to the case where several v's are equal; however, our result remain valid in that case since (2.1) has no pole singularity at $v_i=v_j$ (Appendix C (C2)).

First case: $q-1\geqslant p$:

In expression (2.1), we use the generalized Cauchy's identity (A3) to write

$$\frac{\Pi_{i<j}(\xi_i-\xi_j)\Pi_{k<l}(v_k-v_l)}{\Pi_{i=1}^{q-1}\Pi_{j=1}^{p}(\xi_i-v_j)}=\sum_{I_p}(-)^{\delta_{I_p}}\det\left|\frac{1}{\xi_{j_k}-v_j}\right|\prod_{\substack{i<j\\ i,j\notin I_p}}(\xi_i-\xi_j), \tag{2.5}$$

where $I_p=(j_1<j_2<\cdots<j_p)\subset(1,...,q-1)$ and $\delta_{I_p}=(q-1)p-\Sigma_{i=1}^{p}j_i$. Because of the absence of terms of the type (v_i-v_j) on the right hand side of (2.5), and because each term of the determinant is a product of $(\xi-v_j)^{-1}$ for different j's, the action of the derivatives ∂_{v_j} in (2.1) factorizes. We get a product of the type

$$\prod_{j=1}^{p}\frac{\partial_{v_j}[\Pi_{i=0}^{q}(w_i-v_j)^{1-\beta_i}(\xi_{a_j}-v_j)^{-1}]}{\Pi_{i=0}^{q}(w_i-v_j)^{1-\beta_i}}=\prod_{j=1}^{p}\left\{\left[\sum_{i=0}^{q}\frac{\beta_i-1}{w_i-v_j}-\frac{\partial}{\partial\xi_{a_j}}\right](\xi_{a_j}-v_j)^{-1}\right\}, \quad (2.6)$$

where the a_j's are a permutation of the j_k's in I_p. The right hand side of (2.6) can now be taken inside the integrals $d\xi$ to be integrated by parts (we assume that the corresponding Re β_i's>1 to avoid the contribution of the end-points); the expression (2.6) becomes

$$\prod_{j=1}^{p}\left[\sum_{i=0}^{q}\frac{\beta_i-1}{(w_i-v_j)(\xi_{a_j}-w_i)}\right]. \quad (2.7)$$

For any product of terms in the determinant (2.5), we have obtained a product of terms (2.7); consequently, the determinant in (2.5) can be replaced by the determinant

$$\det\left|\sum_{i=0}^{q}\frac{\beta_i-1}{(\xi_{j_k}-w_i)(w_i-v_j)}\right|=\sum_{J_p}\prod_{i\in J_p}(\beta_i-1)\det\left|\frac{1}{\xi_{j_k}-w_{l_j}}\right|\det\left|\frac{1}{w_{l_k}-v_j}\right|, \quad (2.8)$$

where $J_p=(l_1<l_2<\cdots<l_p)\subset(0,...,q)$.

The situation now, is that the determinant $|(\xi_{j_k}-v_i)^{-1}|$ in (2.5) is replaced by the determinant $|(\xi_{j_k}-w_{l_j})^{-1}|$ with the l_j's in J_p. We may now use the generalized Cauchy's identity (A3) backwards to sum over the different I_p's and we obtain, inside the sum over J_p,

$$\frac{\Pi_{i<j}(\xi_i-\xi_j)\Pi_{j<k}(w_{l_j}-w_{l_k})}{\Pi_{i=1}^{q-1}\Pi_{k=1}^{p}(\xi_i-w_{l_k})}=(-)^{\delta_{J_p}-\eta}\frac{\Pi_{i<j}(\xi_i-\xi_j)\Pi_{j<k}(w_{l_j}-w_{l_k})}{\Pi_{i=1}^{q}\Pi_{k=1}^{p}|\xi_i-w_{l_k}|}, \quad (2.9)$$

where the l_j's belong to J_p, and $\delta_{J_p}=pq-\Sigma_{k=1}^{p}l_k$ (here, $\eta=1$ if $l_1=0$ and 0 otherwise).

We proved, up to now, that

$$C=\frac{1}{\Pi_{k<l}(v_k-v_l)}\sum_{J_p}(-)^{\delta_{J_p}}\prod_{i\in J_p}(\beta_i-1)\prod_{j<k}(w_{l_j}-w_{l_k})\det\left|\frac{1}{w_{l_k}-v_j}\right|$$
$$\times\int_{w_q}^{w_{q-1}}d\xi_{q-1}\cdots\int_{w_2}^{w_1}d\xi_1\prod_{i<j}(\xi_i-\xi_j)\prod_{i=1}^{q-1}\prod_{j=0}^{q}|\xi_i-w_j|^{\beta'_j-1}, \quad (2.10)$$

where $\beta'_j=\beta_j$ if $j\notin J_p$ and $\beta'_j=\beta_j-1$ if $j\in J_p$.

We are now in a position to use Dixon's result with an extra variable w_0 (B4), (B13) and to write the second line of (2.10) as

$$\frac{\Pi_{j=0}^{q}\Gamma(\beta'_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}\prod_{0\le i<j\le q}(w_i-w_j)^{\beta'_i+\beta'_j-1}\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{1}{(w_0-z)^{\beta'_0}\Pi_{j=1}^{q}(z-w_j)^{\beta'_j}}. \quad (2.11)$$

The expression C becomes

$$C=\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}\,\frac{\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-2}}{\Pi_{k<l}(v_k-v_l)}\,\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{1}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}$$

$$\times\sum_{J_p}(-)^{\delta_{J_p}}\prod_{k=1}^{p}(z-w_{l_k})\det\left|\frac{1}{w_{l_k}-v_j}\right|\prod_{\substack{i<j\\ i,j\notin J_p}}(w_i-w_j). \tag{2.12}$$

Again, as a consequence of the generalized Cauchy's identity, we may now sum over all J_p's; from (A8) with $q+1>p$; the third line of (2.12) is nothing but

$$\prod_{k=1}^{p}(z-v_k)\,\frac{\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)\Pi_{k<l}(v_k-v_l)}{\Pi_{i=0}^{q}\Pi_{j=1}^{p}(w_i-v_j)}. \tag{2.13}$$

The fact that the indices of the w's run from 0 to q (instead of 1 to $q+1$) makes in (A8), a shift by p of δ_{J_p}; this explains why we do not have a $(-)^p$ in (2.13).

This result ends the proof for the case $(q-1)\geqslant p$.

Second case: $q-1\leqslant p$:

Again, we use the generalized Cauchy's identity in order to calculate the left hand side of (2.5), but here we have more variables v_j than variables ξ_i; using the relation (A2), we write

$$\frac{\Pi_{i<j}(\xi_i-\xi_j)\Pi_{k<l}(v_k-v_l)}{\Pi_{i=1}^{q-1}\Pi_{j=1}^{p}(\xi_i-v_j)}=\sum_{I_{q-1}}(-)^{\delta_{I_{q-1}}}\det\left|\frac{1}{\xi_i-v_{j_k}}\right|\prod_{\substack{k<l\\ k,l\notin I_{q-1}}}(v_k-v_l), \tag{2.14}$$

where $I_{q-1}=(j_1<j_2<\cdots<j_{q-1})\subset(1,...,p)$ and $\delta_{I_{q-1}}=\Sigma_{k=1}^{q-1}j_k-q+1$. Consequently, the derivatives ∂_{v_j} split into two parts depending on whether j is in I_{q-1} or not. The second line of (2.1) becomes

$$\frac{1}{\Pi_{k<l}(v_k-v_l)}\sum_{I_{q-1}}(-)^{\delta_{I_{q-1}}}\frac{\Pi_{j\notin I_{q-1}}(\partial_{v_j})[\Pi_{i=0}^{q}\Pi_{j\notin I_{q-1}}(w_i-v_j)^{1-\beta_i}\Pi_{\substack{k<l\\ k,l\notin I_{q-1}}}(v_k-v_l)]}{\Pi_{i=0}^{q}\Pi_{j\notin I_{q-1}}(w_i-v_j)^{1-\beta_i}}$$

$$\times\frac{\Pi_{j\in I_{q-1}}(\partial_{v_j})\left[\Pi_{i=0}^{q}\Pi_{j\in I_{q-1}}(w_i-v_j)^{1-\beta_i}\det\left|\frac{1}{\xi_i-v_{j_k}}\right|\right]}{\Pi_{i=0}^{q}\Pi_{j\in I_{q-1}}(w_i-v_j)^{1-\beta_i}}. \tag{2.15}$$

In the third line of (2.15), the derivatives ∂_v factorize so that this term can be calculated in exactly the same way as for the case $(q-1)\geqslant p$. The determinant $(\xi_i-v_{j_k})^{-1}$ can be replaced in the same manner by

$$\det\left|\sum_{r=0}^{q}\frac{\beta_r-1}{(\xi_i-w_r)(w_r-v_{j_k})}\right|=\sum_{J_{q-1}}\prod_{r\in J_{q-1}}(\beta_r-1)\det\left|\frac{1}{\xi_i-w_r}\right|\det\left|\frac{1}{w_r-v_{j_k}}\right|, \tag{2.16}$$

where $J_{q-1}=(l_1<l_2<\cdots<l_{q-1})\subset(0,...,q)$. We now use Cauchy's identity (A1), to write

$$\det\left|\frac{1}{\xi_i-w_r}\right|=(-)^{q(q-1)/2+\delta_{J_{q-1}}-\eta}\,\frac{\Pi_{i<j}(\xi_i-\xi_j)\Pi_{j<k}(w_{l_j}-w_{l_k})}{\Pi_{i=1}^{q-1}\Pi_{j=1}^{q-1}|\xi_i-w_{l_j}|}, \tag{2.17}$$

where $\delta_{J_{q-1}}=\Sigma_{j=1}^{q-1}l_j-q+1$ and $\eta=1$ if $l_1=0$ (and 0 otherwise).

Again, after inserting (2.17) into the integrals $d\xi$, we may integrate using Dixon's change of variables (B4), (B13) to get

$$(-)^{q(q-1)/2}\cdot\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}\prod_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-2}$$
$$\times\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{1}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}$$
$$\times\sum_{J_{q-1}}(-)^{\delta_{J_{q-1}}}\prod_{j=1}^{q-1}(z-w_{l_j})\det\left|\frac{1}{w_{l_j}-v_{j_k}}\right|\prod_{\substack{i<j\\ i,j\notin J_{q-1}}}(w_i-w_j). \tag{2.18}$$

Using a consequence of the generalized Cauchy's identity for a given set of $q-1$ variables v_{j_k}'s in I_{q-1}, the third line of (2.18) can be summed over J_{q-1}; here again, because the w's are labeled from 0 to $q+1$, $\delta_{J_{q-1}}$ has to be shifted by $q-1$ in order to apply (A8) with $q+1>q-1$. We get for the third line of (2.15)

$$(-)^{(q-1)(q-2)/2}\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}\frac{\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-1}\Pi_{\substack{k<l\\ k,l\in I_{q-1}}}(v_k-v_l)}{\Pi_{i=0}^{q}\Pi_{j\in I_{q-1}}(w_i-v_j)}$$
$$\times\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{j\in I_{q-1}}(z-v_j)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}. \tag{2.19}$$

At this point, we proved that, in the case $(q-1)\leqslant p$, the expression C may be written as

$$C=(-)^{(q-1)(q-2)/2}\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}\frac{\Pi_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-1}}{\Pi_{i=0}^{q}\Pi_{j=1}^{p}(w_i-v_j)\Pi_{k<l}(v_k-v_l)}$$
$$\times\sum_{I_{q-1}}(-)^{\delta_{I_{q-1}}}\frac{\Pi_{j\notin I_{q-1}}(\partial_{v_j})[\Pi_{i=0}^{q}\Pi_{j\notin I_{q-1}}(w_i-v_j)^{1-\beta_i}\Pi_{\substack{k<l\\ k,l\notin I_{q-1}}}(v_k-v_l)]}{\Pi_{i=0}^{q}\Pi_{j\notin I_{q-1}}(w_i-v_j)^{-\beta_i}}$$
$$\times\prod_{\substack{k<l\\ k,l\in I_{q-1}}}(v_k-v_l)\,\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{j\in I_{q-1}}(z-v_j)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}. \tag{2.20}$$

The purpose of Appendices C and D is to calculate the sum over I_{q-1} in (2.20). In (D20) we show that the second and third lines of (2.20) are equal to

$$(-)^{(q-1)(q-2)/2}\frac{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}\prod_{k<l}(v_k-v_l)$$
$$\times\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{k=1}^{p}(z-v_k)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}} \tag{2.21}$$

and this ends the proof of the conjecture in the case $(q-1)\leqslant p$.

ACKNOWLEDGMENTS

I thank D. Serban for her constant encouragement, for many discussions about the work in Ref. 7, and for her careful reading of the manuscript.

APPENDIX A

Given two sets of variables $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_m)$, if $n=m$, Cauchy's identity is

$$\frac{\Pi_{i<j}(x_i-x_j)\Pi_{k<l}(y_k-y_l)}{\Pi_{i=1}^{n}\Pi_{k=1}^{n}(x_i-y_k)}=(-)^{n(n-1)/2}\det\left|\frac{1}{x_i-y_k}\right| \tag{A1}$$

(apart from the sign, this result is relatively evident from the properties of the determinant, the homogeneity and the pole structure). Now, if $n\leqslant m$, we may generalize this relation by systematically organizing the residues of the poles; we get

$$\frac{\Pi_{i<j}(x_i-x_j)\Pi_{k<l}(y_k-y_l)}{\Pi_{i=1}^{n}\Pi_{k=1}^{m}(x_i-y_k)}=\sum_{I_n}(-)^{\delta_{I_n}}\det\left|\frac{1}{x_i-y_{j_k}}\right|\prod_{\substack{k<l\\ k,l\notin I_n}}(y_k-y_l), \tag{A2}$$

where $I_n=(j_1<j_2<\ldots<j_n)\subset(1,\ldots,m)$ and $\delta_{I_n}=\Sigma_{i=1}^{n}j_i-n$. Of course, if $n\geqslant m$, the situation is symmetric and we have

$$\frac{\Pi_{i<j}(x_i-x_j)\Pi_{k<l}(y_k-y_l)}{\Pi_{i=1}^{n}\Pi_{k=1}^{m}(x_i-y_k)}=\sum_{I_m}(-)^{\delta_{I_m}}\det\left|\frac{1}{x_{l_i}-y_k}\right|\prod_{\substack{i<j\\ i,j\notin I_m}}(x_i-x_j), \tag{A3}$$

where $I_m=(l_1<l_2<\cdots<l_m)\subset(1,\ldots,n)$ and $\delta_{I_m}=mn-\Sigma_{k=1}^{m}l_k$. We call (A2)–(A3) generalized Cauchy's identities. In (A2)–(A3), we adopt the convention that $\Pi_{\substack{i<j\\ i,j\notin I}}(z_i-z_j)=1$ if there is 0 or 1 variable $z\notin I$.

Let us note the following property: for $n<m$, if we let a given variable $x_p\to\infty$, the left-hand side of (A2) behaves as $x_p^{-(m-n+1)}\leqslant x_p^{-2}$ while the right-hand side behaves as x_p^{-1}. Consequently, the coefficient of x_p^{-1} in the right-side of (A2) is zero. This coefficient can be calculated: we define the matrix $\Delta_{I_n,p}$ such that

$$\begin{aligned}(\Delta_{I_n,p})_{ik}&=\frac{1}{x_i-y_{j_k}}\quad\text{for } i\neq p,\\ &=1\quad\text{for } i=p\end{aligned} \tag{A4}$$

then,

$$\sum_{I_n}(-)^{\delta_{I_n}}\det(\Delta_{I_n,p})\prod_{\substack{k<l\\ k,l\notin I_n}}(y_k-y_l)=0 \tag{A5}$$

We are now using (A5) to calculate

$$\Phi(z)=\sum_{I_n}(-)^{\delta_{I_n}}\prod_{k=1}^{n}(z-y_{j_k})\det\left|\frac{1}{x_i-y_{j_k}}\right|\prod_{\substack{k<l\\ k,l\notin I_n}}(y_k-y_l). \tag{A6}$$

The determinant in (A6) is a sum of terms of the type $(x_1-y_{\alpha_1})^{-1}(x_2-y_{\alpha_2})^{-1}\cdots(x_n-y_{\alpha_n})^{-1}$, where $(\alpha_1,\ldots,\alpha_n)$ is a permutation of $(j_1,\ldots,j_n)$. For such a term, we write

$$\prod_{k=1}^{n}(z-y_{j_k})=\prod_{k=1}^{n}(z-y_{\alpha_k})=\sum_{J}\prod_{k\notin J}(z-x_k)\prod_{j\in J}(x_j-y_{\alpha_j}),$$

where J is a set of $|J|$ indices $\subset(1,...n)$. Consequently, when summing over all terms of the determinant, we get

$$\prod_{k=1}^{n}(z-y_{j_k})\det\left|\frac{1}{x_i-y_{j_k}}\right|=\sum_J\prod_{k\notin J}(z-x_k)\det\left|\left(\frac{1}{x_i-y_{j_k}}\right)_J\right|,$$

where the matrices $((x_i-y_{j_k})^{-1})_J$ have $|J|$ lines of 1 corresponding to the indices $i\in J$. Their determinants are trivially zero for $|J|\geqslant 2$. Now, for $|J|=1$, the sum over I_n in (A6) gives zero because of (A5); the only non-vanishing contribution comes from $|J|=0$. We just proved that

$$\Phi(z)=\prod_{k=1}^{n}(z-x_k)\,\frac{\Pi_{i<j}(x_i-x_j)\Pi_{k<l}(y_k-y_l)}{\Pi_{i=1}^{n}\Pi_{j=1}^{m}(x_i-y_j)}. \tag{A7}$$

By symmetry, for $n>m$, we have

$$\sum_{I_m}(-)^{\delta_{I_m}}\prod_{i=1}^{m}(z-x_{I_i})\det\left|\frac{1}{x_{I_i}-y_k}\right|\prod_{\substack{i<j\\ i,j\notin I_m}}(x_i-x_j)$$

$$=\prod_{i=1}^{m}(z-y_i)\,\frac{\Pi_{i<j}(x_i-x_j)\Pi_{k<l}(y_k-y_l)}{\Pi_{i=1}^{n}\Pi_{j=1}^{m}(x_i-y_j)}, \tag{A8}$$

where, this time, we used determinants of matrices with columns of (-1)'s. Let us mention that if the indices of the x's run from 0 to $n-1$ instead of 1 to n (as it is the case in Eqs. (2.12, 2.18)), then, δ_{I_m} has to be replaced in (A8) by $\delta_{I_m}+m$.

APPENDIX B

In this Appendix, we calculate the integrals $d\xi$ which appear in the first line of (2.1). We first calculate the integrals without the variable w_0; the end of the Appendix is devoted to the introduction of w_0 generating a contour integral in the complex plane. Let us define

$$J_{q-1}=\int_{w_q}^{w_{q-1}}d\xi_{q-1}\cdots\int_{w_2}^{w_1}d\xi_1\prod_{i<j}(\xi_i-\xi_j)\prod_{i=1}^{q-1}\prod_{j=1}^{q}|\xi_i-w_j|^{\beta_j-1}. \tag{B1}$$

Clearly,

$$J_1=\frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}(w_1-w_2)^{\beta_1+\beta_2-1}.$$

A naive calculation of J_2 gives three products of hypergeometric functions, namely,

$$J_2=\prod_{i<j}(w_i-w_j)^{\beta_i+\beta_j-1}\,\frac{\Gamma(\beta_1)(\Gamma(\beta_2))^2\Gamma(\beta_3)}{\Gamma(\beta_1+\beta_2)\Gamma(\beta_2+\beta_3)}$$

$$\times\begin{bmatrix}F(\beta_3,-\beta_1;\beta_2+\beta_3;\alpha)F(1-\beta_3,\beta_1;\beta_1+\beta_2;1-\alpha)\\ +F(\beta_3,1-\beta_1;\beta_2+\beta_3;\alpha)F(-\beta_3,\beta_1;\beta_1+\beta_2;1-\alpha)\\ -F(\beta_3,1-\beta_1;\beta_2+\beta_3;\alpha)F(1-\beta_3,\beta_1;\beta_1+\beta_2;1-\alpha)\end{bmatrix}, \tag{B2}$$

where $\alpha=(w_2-w_3)/(w_1-w_3)$. The remarkable fact is that the square bracket $[\cdots]$ in (B2) is α independent and equal to

$$\frac{\Gamma(\beta_1+\beta_2)\Gamma(\beta_2+\beta_3)}{\Gamma(\beta_1+\beta_2+\beta_3)\Gamma(\beta_2)}.$$

This result is due to Elliot,[13] who generalizes Legendre's relation from the theory of elliptic integrals. A generalization of Elliot's result has been given by Dixon,[12] who proved that

$$J_{q-1}=\frac{\Pi_{j=1}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=1}^{q}\beta_j)}\prod_{i<j}(w_i-w_j)^{\beta_i+\beta_j-1} \tag{B3}$$

by performing a clever change of variables: given the function

$$f(\theta)=\prod_{j=1}^{q}(\theta-w_j)$$

we define new variables

$$x_j=(-)^{q-1}\frac{\Pi_{i=1}^{q-1}(\xi_j-w_i)}{f'(w_j)} \quad \text{for } j=1,...,q.$$

It is easy to verify that all variables x_j are $\geqslant 0$; moreover, we have

$$\sum_{j=1}^{q}x_j=1$$

so that x_q can be considered as dependant of the other x's. The jacobian of the transformation is

$$\frac{d(\xi_1,..,\xi_{q-1})}{d(x_1,...,x_{q-1})}=\frac{\Pi_{i<j}(w_i-w_j)}{\Pi_{i<j}(\xi_i-\xi_j)}.$$

Consequently,

$$J_{q-1}=\prod_{i<j}(w_i-w_j)^{\beta_i+\beta_j-1}\int_0^1\cdot\cdot\int_0^1\prod_{j=1}^{q}[dx_j\,x_j^{\beta_j-1}]\delta\left(\sum_{j=1}^{q}x_j-1\right)$$

which proves (B3).

Next, we wish to calculate (or to transform into a single contour integral in the complex plane) the integrals

$$K_{q-1}=\int_{w_q}^{w_{q-1}}d\xi_{q-1}\cdots\int_{w_2}^{w_1}d\xi_1\prod_{i<j}(\xi_i-\xi_j)\prod_{i=1}^{q-1}\prod_{j=0}^{q}|\xi_i-w_j|^{\beta_j-1}. \tag{B4}$$

In order to transform K_{q-1}, we write

$$\prod_{i=1}^{q-1}|\xi_i-w_0|=\sum_{j=1}^{q}A_j\prod_{i=1}^{q-1}|\xi_i-w_j| \tag{B5}$$

with

$$A_j=\frac{\Pi_{k\neq j}(w_0-w_k)}{\Pi_{k\neq j}|w_j-w_k|}. \tag{B6}$$

In order to calculate $|\xi_i - w_0|^{\beta_0 - 1}$, we use the following integral representation:

$$(a+b)^{\gamma} = \frac{1}{\Gamma(-\gamma)} \frac{1}{2i\pi} \int_{\sigma - i\infty}^{\sigma + i\infty} dz\ \Gamma(-z)\Gamma(z-\gamma) a^z b^{\gamma - z}, \tag{B7}$$

where Re $\gamma \leqslant \sigma \leqslant 0$. Strictly speaking, this integral representation is a priori valid for Re $\gamma < 0$. However, we may, by deformation of the contour, generalize to Re $\gamma > 0$ provided that $|\text{Im } \gamma| > 0$ in order to avoid a possible pinch of the contour and the extraction of residues. Consequently, we assume that β_0 has a small imaginary part for the demonstration and we analytically continue the result to real β_0. Now, we generalize (B7) and write

$$\prod_{i=1}^{q-1} |\xi_i - w_0|^{\beta_0 - 1} = \frac{1}{\Gamma(1-\beta_0)} \frac{1}{(2i\pi)^{q-1}} \times \int \cdots \int \prod_{j=1}^{q-1} dz_j \prod_{j=1}^{q} \left\{ \Gamma(-z_j) A_j^{z_j} \prod_{i=1}^{q-1} |\xi_i - w_j|^{z_j} \right\} \delta_{Z, \beta_0 - 1}, \tag{B8}$$

where $\delta_{Z, \beta_0 - 1}$ is a symbol which simplifies the writing and which means that

$$z_q = \beta_0 - 1 - \sum_{j=1}^{q-1} z_j .$$

We may now apply (B1, B3) in order to integrate the $d\xi$ integrals with β_j replaced by $\beta_j + z_j$ and get

$$\frac{1}{\Gamma(\Sigma_{j=0}^{q} \beta_j - 1)} \frac{1}{\Gamma(1-\beta_0)} \prod_{1 \leqslant i < j \leqslant q} (w_i - w_j)^{\beta_i + \beta_j - 1} \times \frac{1}{(2i\pi)^{q-1}} \int \cdots \int \prod_{j=1}^{q-1} dz_j \prod_{j=1}^{q} \{ \Gamma(-z_j)\Gamma(\beta_j + z_j)(w_0 - w_j)^{\beta_0 - 1 - z_j} \} \delta_{Z, \beta_0 - 1} . \tag{B9}$$

We now consider the integral

$$\frac{1}{2i\pi} \oint_{C_0} dz\ \frac{1}{(w_0 - z)^{\beta_0} \Pi_{j=1}^{q} (z - w_j)^{\beta_j}}, \tag{B10}$$

where C_0 surrounds the cut $[w_0, \infty]$. The integral (B10) can be calculated from the integral of the discontinuity along the cut $[w_0, +\infty]$

$$-\frac{\sin \pi\beta_0}{\pi} \int_0^{\infty} dx\ x^{-\beta_0} \prod_{j=1}^{q} (x + w_0 - w_j)^{-\beta_j}. \tag{B11}$$

Next, we use again the integral representation (B7), for $j = 1$ to $q-1$, in order to write (B11) as

$$-\frac{\sin \pi\beta_0}{\pi} \frac{1}{\Pi_{j=1}^{q-1} \Gamma(\beta_j)} \frac{1}{(2i\pi)^{q-1}} \int \cdots \int \prod_{j=1}^{q-1} [dz_j \Gamma(-z_j)\Gamma(\beta_j + z_j)(w_0 - w_j)^{-z_j - \beta_j}] \times \int_0^{\infty} dx\ x^{-\beta_0 + \Sigma_{j=1}^{q-1} z_j} (x + w_0 - w_q)^{-\beta_q}.$$

After integration of the dx integral, (B10) is equal to

$$-\frac{\sin \pi\beta_0}{\pi}\frac{1}{\Pi_{j=1}^{q}\Gamma(\beta_j)}\frac{1}{(2i\pi)^{q-1}}$$
$$\times\int\cdot\cdot\int\prod_{j=1}^{q-1}dz_j\prod_{j=1}^{q}[\Gamma(-z_j)\Gamma(\beta_j+z_j)(w_0-w_j)^{-z_j-\beta_j}]\delta_{Z,\beta_0-1}. \tag{B12}$$

If we compare (B12) and (B9), we just proved that

$$K_{q-1}=\frac{\Pi_{j=0}^{q}\Gamma(\beta_j)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-1)}\prod_{0\leqslant i<j\leqslant q}(w_i-w_j)^{\beta_i+\beta_j-1}\frac{(-1)}{2i\pi}\oint_{C_0}dz\,\frac{1}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}. \tag{B13}$$

This ends the calculation of K_{q-1} which represents the central part in the proof of the conjecture.

APPENDIX C

In this Appendix we calculate and we give some properties of the expression

$$A=\frac{\partial_{v_1}\cdot\cdot\partial_{v_p}[\Pi_{i=0}^{q}\Pi_{j=1}^{p}(w_i-v_j)^{1-\beta_i}\Pi_{k<l}(v_k-v_l)]}{\Pi_{i=0}^{q}\Pi_{j=1}^{p}(w_i-v_j)^{1-\beta_i}\Pi_{k<l}(v_k-v_l)}. \tag{C1}$$

In order to calculate this expression, we simply calculate the residues of the poles in the variables v_j; the first property which is clear is that there is no singularity when two v's coincide. The only way to get a pole at $v_i=v_j$ is when the derivatives ∂_{v_i} or/and ∂_{v_j} act upon the term (v_i-v_j) at the numerator of (C1). Clearly, for $f(w,v)$ symmetric in v_i and v_j,

$$\frac{\partial_{v_i}\partial_{v_j}[f(w,v)(v_i-v_j)]}{f(w,v)(v_i-v_j)}=\frac{f''_{ij}(w,v)}{f(w,v)}-\frac{f'_i(w,v)-f'_j(w,v)}{f(w,v)(v_i-v_j)} \tag{C2}$$

and the residue of the pole at $v_i=v_j$ is zero (and this property remains evidently true when performing the other derivatives). Consequently, we are going to extract the poles when $v_i=w_k$ neglecting systematically all poles at $v_i=v_j$ since their residues vanish. We write

$$A=\prod_{i=0}^{q}\frac{R_i}{w_i-v_1} \tag{C3}$$

and we find

$$R_i=(\beta_i-1)\frac{\partial_{v_2}\cdot\cdot\partial_{v_p}[f_2\cdot\cdot f_p\Pi_{j>1}(w_i-v_j)\Delta_1]}{f_2\cdot\cdot f_p\Pi_{j>1}(w_i-v_j)\Delta_1}, \tag{C4}$$

where $f_k=\Pi_{i=0}^{q}(w_i-v_k)^{1-\beta_i}$ and $\Delta_1=\Pi_{1<i<j}(v_i-v_j)$.

We now extract the poles in v_2 and write

$$A=\sum_{\{i_1,i_2\}=0}^{q}\frac{R_{i_1,i_2}}{(w_{i_1}-v_1)(w_{i_2}-v_2)}. \tag{C5}$$

We find

$$R_{i_1,i_2}=(\beta_{i_1}-1)(\beta_{i_2}-1-\delta_{i_1,i_2})\ \frac{\partial_{v_3}..\partial_{v_p}[f_3..f_p\Pi_{j>2}(w_{i_1}-v_j)(w_{i_2}-v_j)\Delta_2]}{f_3..f_p\Pi_{j>2}(w_{i_1}-v_j)(w_{i_2}-v_j)\Delta_2}, \tag{C6}$$

where $\Delta_2=\Pi_{2<i<j}(v_i-v_j)$. We can proceed and extract successively all the poles in v_i; the result can be written in the following way: given $(q+1)$ complementary (and possibly empty) sets $I_0,I_1,\ldots,I_q$ such that

$$I_r\cap I_s=\Phi,\quad \cup_{r=0}^{q}I_r=(1,..,p), \tag{C7}$$

then

$$A=\sum_{\{I_0,\ldots,I_q\}}\prod_{r=0}^{q}\left\{\frac{\Gamma(\beta_r)}{\Gamma(\beta_r-|I_r|)}\sum_{j\in I_r}\ \frac{1}{w_r-v_j}\right\}, \tag{C8}$$

where $|I_r|$ is the number of elements in I_r.

Let us write, as an example, the case where $\beta=p$ integer and where we have only two variables w_0 and w_1. Then,

$$A=(p-1)(p-1)!\sum_{I\subset(1,..,p)}C_{p-2}^{|I|-1}\prod_{i\in I}\ \frac{1}{w_0-v_i}\prod_{i\notin I}\ \frac{1}{w_1-v_i}. \tag{C9}$$

In the rest of the appendix, we show that A is the ratio of two determinants, and we develop a recurrent construction satisfied by expressions of the same type of A.

We define successively the following symbols: given the symbol $[u]=[u]_0$ attached to a given variable u, and the symbols

$$[u]_n=nu^{n-1}+u^n[u] \tag{C10}$$

then, we define the completely symmetric symbols

$$[v_1,..,v_p]=\frac{1}{\Delta}\det\begin{vmatrix}[v_1]_0 & [v_2]_0 & \cdot\cdot & [v_p]_0\\ [v_1]_1 & [v_2]_1 & \cdot\cdot & [v_p]_1\\ \cdot & \cdot & & \cdot\\ \cdot & \cdot & & \cdot\\ [v_1]_{p-1} & [v_2]_{p-1} & \cdot\cdot & [v_p]_{p-1}\end{vmatrix}, \tag{C11}$$

where $\Delta=(-)^{n(n-1)/2}\Pi_{i<j}(v_i-v_j)$ is the Vandermonde determinant. Clearly, these symbols can also be defined recursively by

$$\begin{aligned}[v,w]&=\frac{(1+v[v])[w]}{v-w}+\frac{(1+w[w])[v]}{w-v},\\ [u,v,w]&=\frac{(2u+u^2[u])[v,w]}{(u-v)(u-w)}+\text{circ.perm.},\\ &\cdots\\ [v_1,..,v_p]&=\frac{((p-1)v_1^{p-2}+v_1^{p-1}[v_1])[v_2,..,v_p]}{\Pi_{i=2}^{p}(v_1-v_i)}+\text{circ.perm.}\end{aligned} \tag{C12}$$

To illustrate this construction, we give two examples. First, let us define

$$[u]=\frac{\beta}{u}. \tag{C13}$$

Then, it is easy to construct the different symbols; using the properties of determinants, we obtain

$$[v_1,..,v_p]=\frac{\Gamma(\beta+p)}{\Gamma(\beta)}\frac{1}{v_1\ldots v_p}. \tag{C14}$$

Moreover, since

$$[v]_n=\frac{\partial_v(v^{\beta+n})}{v^{\beta}}$$

we have

$$[v_1,..,v_p]=\frac{\partial_{v_1}\cdots\partial_{v_p}[v_1^{\beta}..v_p^{\beta}\Pi_{i<j}(v_i-v_j)]}{v_1^{\beta}..v_p^{\beta}\Pi_{i<j}(v_i-v_j)}. \tag{C15}$$

As a second example, we define

$$[u]=\sum_{i=0}^{q}\frac{\beta_i-1}{w_i-u}. \tag{C16}$$

Then, from

$$[v]_n=\frac{\partial_v[v^n\Pi_{i=0}^{q}(w_i-v)^{1-\beta_i}]}{\Pi_{i=0}^{q}(w_i-v)^{1-\beta_i}} \tag{C17}$$

and the properties of determinants, we find that

$$[v_1,..,v_p]=A \tag{C18}$$

as given in (C1) and (C8).

APPENDIX D

We are interested in the structure of the integrals

$$I_{P(z)}=\frac{1}{2i\pi}\oint_{C_0}dz\,\frac{P(z)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}, \tag{D1}$$

where $P(z)$ is a polynomial and where the sum over the variables β is large enough to make the integral convergent along the cut $C_0=[w_0,+\infty]$. To simplify the writing of this Appendix, we keep abusively a z dependence on the left hand side of (D1). From

$$\frac{1}{2i\pi}\oint_{C_0}dz\,\frac{\partial}{\partial z}\left\{\frac{z^n}{(w_0-z)^{\beta_0-1}\Pi_{j=1}^{q}(z-w_j)^{\beta_j-1}}\right\}=0, \tag{D2}$$

we obtain the relation

$$-nI_{z^{n-1}\Pi_{j=0}^{q}(z-w_j)}+\sum_{j=0}^{q}(\beta_j-1)I_{z^n\Pi_{i\neq j}(z-w_i)}=0 \tag{D3}$$

In the special case where $n=0$, we have

$$\sum_{j=0}^{q}(\beta_j-1)I_{\Pi_{i\neq j}(z-w_i)}=0 \tag{D4}$$

Equation (D4) shows that the function I_{z^q} is linearly dependent of the functions I_{z^p} for $0\leqslant p<q$. Also, for $n>0$, equation (D3) shows that the functions I_{z^p} for $p>q$ can be decomposed on the basis of the q functions I_{z^p} $(0\leqslant p<q)$. We have constructed a q dimensional vectorial space E_q.

The decomposition of the functions $I_{P(z)}$ on the basis of functions I_{z^p} $(0\leqslant p<q)$ is particularly simple if the polynomial $P(z)$ is of degree $<q$; it becomes particularly difficult otherwise because of the relations (D3)–(D4). We are going to introduce other basis for the vectorial space E_q which are specially adapted when the polynomial $P(z)$ is known from its roots $P(z)=\Pi_i(z-v_i)$. Let us give ourselves q different values of the variable v say $v_1,..,v_q$; then, we define a basis of E_q as a set of q independent functions

$$I_{(j)}=I_{\Pi_{i\neq j}(z-v_i)} \quad \text{for } j=1,..,q. \tag{D5}$$

The transformation from the basis I_{z^p} to the basis (D5) can be written as

$$I_{z^p}=\sum_{j=1}^{q}\frac{v_j^p}{\Pi_{i\neq j}(v_j-v_i)}I_{(j)} \tag{D6}$$

and consequently, if $P_{q-1}(z)$ is a polynomial in z of degree at most $q-1$, then

$$I_{P_{q-1}(z)}=\sum_{j=1}^{q}\frac{P_{q-1}(v_j)}{\Pi_{i\neq j}(v_j-v_i)}I_{(j)}. \tag{D7}$$

As we already mentioned, the decomposition of $I_{z^q},I_{z^{q+1}}\cdots$ is more elaborate; let us give the example of I_{z^q}. From (D4), we have

$$D_{q+1}I_{z^q}+\sum_{j=0}^{q}(\beta_j-1)I_{\Pi_{i\neq j}(z-w_i)-z^q}=0, \tag{D8}$$

where $D_p=\Sigma_{j=0}^{q}\beta_j-p$. Since the polynomial $\Pi_{i\neq j}(z-w_i)-z^q$ is of degree $q-1$ in z, we may apply (D7); we obtain

$$I_{z^q}=\sum_{r=1}^{q}\frac{\left[v_r^q-\frac{1}{D_{q+1}}\Sigma_{j=0}^{q}(\beta_j-1)\Pi_{i\neq j}(v_r-w_i)\right]}{\Pi_{s\neq r}(v_r-v_s)}I_{(r)}. \tag{D9}$$

Given a polynomial $P_q(z)$ of degree q in z (with the coefficient of $z^q=1$), we may write

$$I_{P_q(z)}=\sum_{r=1}^{q}\frac{\left[P_q(v_r)-\frac{1}{D_{q+1}}\Sigma_{j=0}^{q}(\beta_j-1)\Pi_{i\neq j}(v_r-w_i)\right]}{\Pi_{s\neq r}(v_r-v_s)}I_{(r)}. \tag{D10}$$

If moreover, the choice of the variables $v_1,..,v_q$ which defines the basis $I_{(r)}$ coincide with the roots of the polynomial $P_q(z)$, that is $P_q(v_r)=0$ for $r=1,..,q$, then

$$I_{\Pi_{i=1}^{q}(z-v_i)}=\frac{(-)^{q+1}}{D_{q+1}}\sum_{r=1}^{q}\frac{\Pi_{i=0}^{q}(w_i-v_r)}{\Pi_{s\neq r}(v_r-v_s)}[v_r]I_{(r)}, \tag{D11}$$

where the symbol $[v]$ is defined in Appendix C (C6).

We now extend the above construction to polynomials of degree larger than q. Given a set of $p>q$ values $v_1,..,v_p$ of a variable v, we define a set of p functions as

$$\overline{I_{(r)}}=I_{\Pi_{i\neq r}(z-v_i)}. \tag{D12}$$

Clearly these p functions are necessarily dependent in the vectorial space E_q. Nevertheless, the relation (D6) remains valid

$$I_{z^j}=\sum_{r=1}^{p}\frac{v_r^j}{\Pi_{i\neq r}(v_r-v_i)}\overline{I_{(r)}}\quad\text{for } j=0,1,..,p-1. \tag{D13}$$

Now, using (D3) with $n=p-q$, and performing the same transformation (D8) to (D11), we obtain

$$I_{\Pi_{i=1}^{p}(z-v_i)}=\frac{(-)^{q+1}}{D_{p+1}}\sum_{r=1}^{p}\frac{\Pi_{j=0}^{q}(w_j-v_r)}{\Pi_{i\neq r}(v_r-v_i)}[v_r]_{p-q}\ \overline{I_{(r)}}, \tag{D14}$$

where $[v_r]_n$ is defined in Appendix C (C10).

Let us illustrate this result on the example where $p=q+1$. In that case the $q+1$ functions $\overline{I_{(r)}}$ can themselves be decomposed according to (D11) generating a set of $q(q+1)/2$ functions

$$\overline{I_{(r,s)}}=\prod_{\substack{i=1\\ i\neq r,s}}^{q+1}(z-v_i). \tag{D15}$$

When transforming (D14) in terms of $\overline{I_{(r,s)}}$, we obtain two contributions depending whether we express $\overline{I_{(r)}}$ or $\overline{I_{(s)}}$ in terms of $\overline{I_{(r,s)}}$; these two contributions construct the combination

$$\frac{[v_r]_1[v_s]_0}{v_r-v_s}+\frac{[v_s]_1[v_r]_0}{v_s-v_r}=[v_r,v_s], \tag{D16}$$

where $[v_r,v_s]$ is defined in Appendix C (C12). The expression (D14) becomes

$$I_{\Pi_{i=1}^{q+1}(z-v_i)}=\frac{(-)^{2q+2}}{D_{q+2}D_{q+1}}\sum_{r<s}\frac{\Pi_{j=0}^{q}[(w_j-v_r)(w_j-v_s)]}{\Pi_{i\neq r,s}[(v_r-v_i)(v_s-v_i)]}[v_r,v_s]\overline{I_{(r,s)}}. \tag{D17}$$

We insist again on the fact that this decomposition is not a decomposition over a basis of E_q but it is exactly the kind of decomposition we need to prove the conjecture in the case $q-1\leqslant p$.

We may generalize (D17) and we may prove by recurrence for any $p>q$, that

$$I_{\Pi_{i=1}^{p}(z-v_i)}=(-)^{(p-q+1)(q+1)}\frac{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}\sum_{J_{q-1}}\frac{\Pi_{j=0}^{q}\Pi_{k\notin J_{q-1}}(w_j-v_k)}{\Pi_{s\notin J_{q-1}}\Pi_{i\in J_{q-1}}(v_s-v_i)}$$
$$\times\sum_{r\notin J_{q-1}}\frac{[v_r]_{p-q}[\cup_{\substack{k\notin J_{q-1}\\ k\neq r}}(v_k)]}{\Pi_{\substack{i\notin J_{q-1}\\ i\neq r}}(v_r-v_i)}\overline{I_{J_{q-1}}},\tag{D18}$$

where $J_{q-1}=(j_1,..,j_{q-1})\subset(1,..,p)$ and where the notation $\overline{I_{J_{q-1}}}$ means $\overline{I_{(\cup k\notin J_{q-1})}}$. By Appendix C (C12), the sum over r in the third line of (D18) reconstructs $[\cup_{k\notin J_{q-1}}(v_k)]$.

We now rewrite (D18) in a way which is directly applicable to Sec. II Eqs. (2.20–21). We have

$$\frac{1}{\Pi_{s\notin J_{q-1}}\Pi_{i\in J_{q-1}}(v_s-v_i)}=(-)^{(2p-q)(q-1)/2+\delta_{J_{q-1}}}\frac{\Pi_{\substack{k<l\\ k,l\notin J_{q-1}}}(v_k-v_l)\Pi_{\substack{k<l\\ k,l\in J_{q-1}}}(v_k-v_l)}{\Pi_{k<l}(v_k-v_l)},\tag{D19}$$

where $\delta_{J_{q-1}}=\Sigma_{k=1}^{q-1}j_k-q+1$. From the definition of the symbol $[\cup_{k\notin J_{q-1}}(v_k)]$ as given in Appendix C (C18, C1), we may write

$$\sum_{J_{q-1}}(-)^{\delta_{J_{q-1}}}\frac{\Pi_{j\notin J_{q-1}}(\partial_{v_j})[\Pi_{i=0}^{q}\Pi_{j\notin J_{q-1}}(w_i-v_j)^{1-\beta_i}\Pi_{\substack{k<l\\ k,l\notin J_{q-1}}}(v_k-v_l)]}{\Pi_{i=0}^{q}\Pi_{j\notin J_{q-1}}(w_i-v_j)^{-\beta_i}}$$
$$\times\frac{\Pi_{\substack{k<l\\ k,l\in J_{q-1}}}(v_k-v_l)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{j\in J_{q-1}}(z-v_j)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}$$
$$=(-)^{(q-1)(q-2)/2}\frac{\Gamma(\Sigma_{j=0}^{q}\beta_j-q)}{\Gamma(\Sigma_{j=0}^{q}\beta_j-p-1)}\frac{\Pi_{k<l}(v_k-v_l)}{2i\pi}\oint_{C_0}dz\,\frac{\Pi_{j=1}^{p}(z-v_j)}{(w_0-z)^{\beta_0}\Pi_{j=1}^{q}(z-w_j)^{\beta_j}}\tag{D20}$$

which is the relation used at the end of Sec. II.

[1] F. Calogero, J. Math. Phys. **10**, 2191, 2197 (1969); B. Sutherland, *ibid.* **12**, 246 (1971); Phys. Rev. A **4**, 2019 (1971); **5**, 1372 (1972).
[2] P. J. Forrester, Nucl. Phys. B **388**, 671 (1992).
[3] F. J. Dyson, J. Math. Phys. **3**, 140,157 (1962); B. D. Simons, P. A. Lee, and B. L. Al'tshuler, Nucl. Phys. B **409**, 487 (1993); K. B. Efetov, Adv. Phys. **32**, 53 (1983).
[4] F. D. M. Haldane and M. R. Zirnbauer, Phys. Rev. Lett. **71**, 4055 (1993); M. R. Zirnbauer and F. D. M. Haldane, Phys. Rev. B **52**, 8729 (1995).
[5] F. Lesage, V. Pasquier, and D. Serban, Nucl. Phys. B **435**, 585 (1995).
[6] Z. N. C. Ha, Phys. Rev. Lett. **73**, 1574 (1994); Nucl. Phys. B **435**, 604 (1995).
[7] D. Serban, F. Lesage, and V. Pasquier, Nucl. Phys. B **466**, 499 (1996).
[8] D. V. Khveshchenko, Int. J. Mod. Phys. B **9**, 1639 (1995).
[9] K. Mimachi and Y. Yamada, Commun. Math. Phys. **174**, 447 (1995); H. Awata, Y. Matsuo, S. Odake, and J. Shiraishi, Nucl. Phys. B **449**, 347 (1995).
[10] S. N. M. Ruijsenaars and H. Schneider, Appl. Phys. **170**, 370 (1986); S. N. M. Ruijsenaars, Commun. Math. Phys. **110**, 191 (1987); H. Konno, Nucl. Phys. B **473**, 579 (1996).
[11] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed. (Clarendon, Oxford, 1995).
[12] A. L. Dixon, Proc. London Math. Soc. **3**, 206 (1905)
[13] H. Bateman, *Higher Transcendental Functions*, Vol. 1, p. 85, Eq. (13); E. B. Elliot, Messenger of Math **33**, 31 (1904).

Aharonov–Bohm effect with δ-type interaction

L. Dąbrowski
S.I.S.S.A., 34014 Trieste, via Beirut 2-4, Italy

P. Šťovíček
Department of Mathematics, Faculty of Nuclear Science, CTU, Trojanova 13, 120 00 Prague, Czech Republic

(Received 13 January 1997; accepted for publication 21 August 1997)

A quantum particle interacting with a thin solenoid and a magnetic flux is described by a five-parameter family of Hamilton operators, obtained via the method of self-adjoint extensions. One of the parameters, the value of the flux, corresponds to the Aharonov–Bohm effect; the other four parameters correspond to the strength of a singular potential barrier. The spectrum and eigenstates are computed and the scattering problem is solved. © *1998 American Institute of Physics.* [S0022-2488(98)01101-3]

I. INTRODUCTION

The purpose of this paper is to obtain and study the most general family of operators that describe the essential features of a quantum mechanical particle under the joint effect of the electromagnetic potential due to a flux ϕ together with the potential barrier supported on the infinite thin shielded solenoid.

Our initial task is to provide a class of well-defined operators corresponding to the formal differential (plus distributional) expression

$$-\sum_{\ell=1}^{3}\left(\frac{\partial}{\partial x_\ell}-A(\partial_\ell)\right)^2+\lambda\,\delta(r),$$

where x_1,x_2,x_3 are the standard coordinates in $\mathbb{R}^3$, $A=\iota(\phi/2\pi r^2)(-x_2dx_1+x_1dx_2)$ is a pure gauge potential, and $r=((x_1)^2+(x_2)^2)^{1/2}$. For that aim, we first reduce the problem to two dimensions by making use of translational symmetry with respect to the coordinate x_3. Let r, θ be, respectively, the radial and angular coordinate in $\mathbb{R}^2$, $0\leqslant r$ and $0\leqslant\theta\leqslant 2\pi$. Set $\alpha=-\phi/2\pi$. We concentrate on the case when $\alpha\notin\mathbb{Z}$ and owing to the gauge symmetry $A'=A+e^{-\iota n\theta}de^{i\nu\theta}$, $n\in\mathbb{Z}$, we can restrict ourselves to the case

$$\alpha\in\,]0,1[\,. \tag{1}$$

The method we adopt in this paper is based on self-adjoint extensions of symmetric operators. From this perspective, we try to combine two well-known cases which were already extensively discussed in the literature. The first one, with $\alpha=0$, corresponds to the so-called point interaction (cf. Ref. 1) and was studied in detail in Ref. 2. An operator in a one-parameter family is defined on a domain which is characterized by a linear relation between certain coefficients which are built up from the asymptotic behavior, as $r\to 0$, of (singular) wave functions. The starting point was a symmetric operator with a domain formed by wave functions with supports separated from the origin. The deficiency indices turned out to be (1,1).

The second case, with $\lambda=0$, corresponds to the pure Aharonov–Bohm (A–B) potential[3] and was investigated many times on different levels (see Refs. 4–8). The generalized eigenfunctions are required to belong to $H^{2,2}_{\mathrm{loc}}(\mathbb{R}^2\backslash\{0\})$ and at the origin the regular condition

$$\lim_{r\to 0_+} f(r,\theta)=0 \tag{2}$$

is imposed. As the Hamiltonian enjoys rotational symmetry, the generalized eigenfunctions are known and the scattering problem is solved explicitly.

However, it has been known already for some time that, when decomposing the Hilbert space into a direct sum as follows,

$$L^2(\mathbb{R}^2,d^2x)\simeq L^2(\mathbb{R}_+,rdr)\otimes L^2([0,2\pi],d\theta)=\oplus_{m=-\infty}^{\infty}L^2(\mathbb{R}_+,rdr)\otimes\chi_m, \tag{3}$$

where

$$\chi_m(\theta)=(2\pi)^{-1/2}e^{im\theta}, \tag{4}$$

the A–B operator decomposes correspondingly and in the channels $m=-1$ and $m=0$ the boundary condition (2) is not the most general one. But since the other boundary conditions admit wave functions that are singular at the origin, they were usually ruled out.

In our opinion, it makes good sense to consider the most general case and hence to allow even a sort of interaction between the two channels. Thus we apply to the pure A–B Hamiltonian exactly the same procedure that was used in the case of point interactions. The deficiency indices one obtains this way are (2,2) and this indicates clearly that the result is not simply a superposition of the two special cases.

II. FIVE-PARAMETER FAMILY OF HAMILTON OPERATORS AND THEIR RESOLVENTS

In order to get operators that can be consistently interpreted as describing the physical situation we are interested in, we start with the case of pure A–B effect and introduce the point interaction at $0\in\mathbb{R}^2$ in the usual way. Namely, first we consider the restriction of the self-adjoint pure A–B operator H to the space of functions with supports outside of $\{0\}$, obtaining thus a closable symmetric operator. Then we shall find all possible self-adjoint extensions of its closure $\widetilde{H}$.

The adjoint $\widetilde{H}^*$ is defined as the differential operator $-(\boldsymbol{\nabla}-A(\boldsymbol{\nabla}))^2$ on the domain

$$\psi\in\mathscr{D}(\widetilde{H}^*)\Leftrightarrow\psi\in L^2(\mathbb{R}^2)\cap H^{2,2}_{\mathrm{loc}}(\mathbb{R}^2\backslash\{0\})\quad\text{and}\quad(\boldsymbol{\nabla}-A(\boldsymbol{\nabla}))^2\psi\in L^2(\mathbb{R}^2). \tag{5}$$

On general grounds, $\widetilde{H}$ has equal deficiency indices. To find the corresponding deficiency spaces we employ the decomposition (3). Since the orthogonal projection onto $L^2(\mathbb{R}_+)\otimes\chi_m$ commutes with $\widetilde{H}$ on $\mathscr{D}(\widetilde{H})$, we can solve the eigenvalue problem

$$\widetilde{H}^*f=k^2f, \tag{6}$$

with $k=e^{\imath\pi/4}$ and $k=e^{\imath 3\pi/4}$, in each sector m of the angular momentum. Setting $f(r,\theta)=g(r)\chi_m(\theta)$, (6) becomes

$$-\left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}-\frac{(m+\alpha)^2}{r^2}\right)g=k^2g, \tag{7}$$

which, by the standard substitution $r\to kr$, leads to the Bessel equation.

Next, selecting in the two-dimensional space of solutions the one which vanishes at the infinity, we arrive at the Henkel functions

$$g(r)=H^{(1)}_{|m+\alpha|}(kr). \tag{8}$$

To ensure the integrability we still have to control the asymptotics as $r\to 0_+$. The case $\alpha=0$ is known; the L^2 solutions exist only in the sector $m=0$ and thus the the deficiency indices are (1,1). Assuming now that $0<\phi<2\pi$ and recalling the asymptotics

$$H_\nu^{(1)}(z)=\frac{2^\nu \imath}{\sin(\pi\nu)\ \Gamma(1-\nu)}\, z^{-\nu}+\frac{2^{-\nu}\imath e^{-i\pi\nu}}{\sin(\pi\nu)\ \Gamma(1+\nu)}\, z^{\nu}+O(z^{-\nu+2}), \tag{9}$$

as $z\to 0$, the integrability at 0 means that $2|m+\alpha|-1<1$, which selects precisely two angular momentum sectors: $m=-1$ and $m=0$. Thus the deficiency indices of $\widetilde{H}$ are (2,2) and the deficiency space $\mathscr{N}_\imath$ is spanned by $f_\imath^1$ and $f_\imath^2$ given by

$$f_\imath^1(r,\theta)=N_1 H^{(1)}_{1-\alpha}(kr)e^{-\imath\theta},\quad f_\imath^2(r,\theta)=N_2 H^{(1)}_{\alpha}(kr), \tag{10}$$

where $k=e^{\imath\pi/4}$ ($k=\sqrt{\imath}$ with $\mathrm{Im}\,k>0$) and the normalization constants N_1, N_2 will be determined later on.

This means that all self-adjoint extensions are in one-to-one correspondence with the elements of the unitary group $U(2)$ and are determined by boundary conditions at the origin. We treat them in detail in the next section. Thus we get, apart from α characterizing the magnetic flux, four additional parameters.

It is now straightforward to determine the domain $\mathscr{D}(\widetilde{H})$. As $\widetilde{H}=\widetilde{H}^{**}$ it holds true that

$$\psi\in\mathscr{D}(\widetilde{H})\Leftrightarrow\psi\in\mathscr{D}(\widetilde{H}^*)\quad\text{and}\quad\langle\psi,\widetilde{H}^*\varphi\rangle=\langle\widetilde{H}^*\psi,\varphi\rangle,\quad\forall\varphi\in\mathscr{N}_\imath+\mathscr{N}_{-\imath}.$$

Consequently we find that $g(r)\otimes\chi_m(\theta)\in\mathscr{D}(\widetilde{H})$, with $m=-1$, 0, if and only if $g\in L^2(\mathbb{R}_+,rdr)\cap H^{2,2}_{\mathrm{loc}}(]0,+\infty[)$, $(\partial_r^2+r^{-1}\partial_r-r^{-2}(m+\alpha)^2)g\in L^2(\mathbb{R}_+,rdr)$, and

$$\lim_{r\to 0_+} rW(g,h_\pm)=0, \tag{11}$$

where $h_\pm=H^{(1)}_{|m+\alpha|}(\sqrt{\pm\imath}r)$ (with $\mathrm{Im}\,\sqrt{\pm\imath}>0$) and the symbol $W(g,h)$ stands for the Wronskian,

$$W(g,h):=\overline{g}\partial_r h-\overline{h}\partial_r g.$$

At this point, let us make a short digression and recall a useful formula comparing resolvents of two self-adjoint extensions. It is stated in the framework of Krein's approach to the theory of self-adjoint extensions and was presented originally in Ref. 9. Let us consider a general situation when a closed symmetric operator X is given and A_0 is a self-adjoint extension of X, $X\subset A_0=A_0^*\subset X^*$. Assume that the deficiency indices of X are (d,d), $d<\infty$. Set

$$\mathscr{N}_z:=\mathrm{Ker}(X^*-z),\quad R_z^0:=(A_0-z)^{-1},\quad z\in\mathbb{C}\backslash\mathbb{R}. \tag{12}$$

The following facts are well known and easy to check. First,

$$\psi\in\mathscr{N}_w\Rightarrow\psi+(z-w)R_z^0\psi=(A_0-w)R_z^0\psi\in\mathscr{N}_z. \tag{13}$$

Second, if $U_z^0:\mathscr{N}_z\to\mathscr{N}_{\overline{z}}$, $z\in\mathbb{C}\backslash\mathbb{R}$, is the unitary mapping defining the self-adjoint extension A_0, then

$$\forall\psi\in\mathscr{N}_z,\quad U_z^0\psi=-(A_0-z)(A_0-\overline{z})^{-1}\psi. \tag{14}$$

Fix $w\in\mathbb{C}\backslash\mathbb{R}$ and a basis $\{f_w^\ell\}_\ell$ in $\mathscr{N}_w$. Set

$$f_z^\ell=f_w^\ell+(z-w)R_z^0 f_w^\ell=(A_0-w)R_z^0 f_w^\ell,\quad z\in\mathbb{C}\backslash\mathbb{R}. \tag{15}$$

Since $(A_0-w)R_z^0$ is injective and $\dim\mathscr{N}_w=\dim\mathscr{N}_z=d$, $\{f_z^1,\ldots,f_z^d\}$ is a basis in $\mathscr{N}_z$. Suppressing the superscript ℓ one can verify readily that

$$\forall z,z' \in \mathbb{C}\backslash\mathbb{R}, \quad f_z = f_{z'} + (z-z')R^0_z f_{z'}, \tag{16}$$

$$\forall z \in \mathbb{C}\backslash\mathbb{R}, \quad U^0_z f_z = -f_{\overline{z}}. \tag{17}$$

Thus, in order to reproduce the vector-valued function f_z, one can take any $z' \in \mathbb{C}\backslash\mathbb{R}$ instead of w. Furthermore, the matrix of U^0_z in the bases $\{f^{\vee}_z\}_{\vee}$, $\{f^{\vee}_{\overline{z}}\}_{\vee}$ equals -1.

To proceed further let us introduce a matrix $P(z,w)=(P^{jk}(z,w))$ of scalar products relating two spaces $\mathscr{N}^{\cdot}_z$ and $\mathscr{N}^{\cdot}_w$,

$$P^{jk}(z,w) := \langle f^j_z, f^k_w \rangle. \tag{18}$$

One finds that

$$P(z,z) = P(\overline{z,z}), \tag{19}$$

and if U is any matrix expressed in the bases $\{f^{\vee}_z\}$, $\{f^{\vee}_{\overline{z}}\}$, and representing a unitary mapping $\mathscr{N}^{\cdot}_z \to \mathscr{N}^{\cdot}_{\overline{z}}$, then

$$P(z,z) = U^* P(\overline{z,z}) U. \tag{20}$$

Furthermore, if $\mathscr{P}_z$ is the orthogonal projector onto $\mathscr{N}^{\cdot}_z$, then

$$\mathscr{P}_z = \sum_{jk} (P(z,z)^{-1})^{jk} f^j_z \langle f^k_z, \cdot \rangle. \tag{21}$$

Our primary interest is to compare A_0 with another self-adjoint extension A of X corresponding to a family of unitary mappings $U_z : \mathscr{N}^{\cdot}_z \to \mathscr{N}^{\cdot}_{\overline{z}}$. Krein's formula tells us that

$$R_z - R^0_z = (\overline{z} - z)^{-1} \mathscr{P}^*_z (U_z - U^0_z) \mathscr{P}_{\overline{z}}, \tag{22}$$

with the symbol $\mathscr{P}^*_z$ standing for the embedding of $\mathscr{N}^{\cdot}_z$ into the global Hilbert space. This means that there exists a family of $d \times d$ matrices, $M_z = (M^{jk}_z)$, such that

$$R_z - R^0_z = \sum_{jk} M^{jk}_z f^j_z \langle f^k_{\overline{z}}, \cdot \rangle. \tag{23}$$

We claim that

$$M_z - M_w = (z-w) M_z P(\overline{z},w) M_w, \tag{24}$$

$$(\overline{w} - w) M_w = (U^{-1} + 1) P(w,w)^{-1}, \tag{25}$$

where U is the matrix of $U_w : \mathscr{N}^{\cdot}_w \to \mathscr{N}^{\cdot}_{\overline{w}}$ in the bases $\{f^j_w\}$ and $\{f^j_{\overline{w}}\}$. The proof is quite straightforward and relies on Krein's formula, the first resolvent identity, and the explicit expression for $\mathscr{P}_z$. Provided M_w and M_z are invertible one can also write

$$M^{-1}_w - M^{-1}_z = (z-w) P(\overline{z},w). \tag{26}$$

What we shall need in the sequel is the particular choice of $w = -\iota$. In this case,

$$M^{-1}_z = 2\iota P(\iota,\iota)(U+1)^{-1} - (z+\iota) P(\overline{z},-\iota), \tag{27}$$

where this time U is the matrix of $U_\iota : \mathscr{N}^{\cdot}_\iota \to \mathscr{N}^{\cdot}_{-\iota}$ in the above specified bases.

Next we apply this general procedure to our problem, with $A_0 \equiv H$-the pure A–B operator. Thus, for $z \in \mathbb{C}\backslash\mathbb{R}_+$, we choose in $\mathscr{N}_z$ a particular basis which depends holomorphically on z by

$$f_z^\ell = (1+(z-\iota)R_z)f_\iota^\ell, \quad \ell = 1,2, \tag{28}$$

where R_z is the resolvent of the pure A–B operator defined by its integral kernel (Green's function)

$$G_z(r,\theta;r',\theta') = \frac{1}{2\pi}\sum_{m\in\mathbb{Z}} e^{\iota m(\theta-\theta')}\int_0^\infty (k^2-z)^{-1} J_{|m+\alpha|}(kr)J_{|m+\alpha|}(kr')k\ dk. \tag{29}$$

Recalling that

$$H_\nu^{(1)}(z) = \frac{2}{\pi\iota}\, e^{-\iota\pi\nu/2}K_\nu(-\iota z), \tag{30}$$

where

$$K_\nu(z) = \int_0^\infty e^{-z\cosh t}\cosh \nu t\ dt, \tag{31}$$

and using the formulas 8.13(3) and 8.5(12) of Ref. 10,

$$\int_0^\infty K_\mu(ay)J_\mu(xy)y\ dy = \left(\frac{x}{a}\right)^\mu (x^2+a^2)^{-1}, \tag{32}$$

$$\int_0^\infty \frac{x^\mu}{x^2+a^2}\, J_\mu(yx)x\ dx = a^\mu K_\mu(ay), \quad \mu\in\,]-1,3/2[, \quad \mathrm{Re}\ a>0, \tag{33}$$

we have

$$f_z^1(r,\theta) = N_1 e^{-\iota\pi(1-\alpha)/4}(\sqrt{z})^{1-\alpha}H_{1-\alpha}^{(1)}(\sqrt{z}r)e^{-\iota\theta},$$

$$f_z^2(r,\theta) = N_2 e^{-\iota\pi\alpha/4}(\sqrt{z})^{\alpha}H_{\alpha}^{(1)}(\sqrt{z}r), \tag{34}$$

where $z\in\mathbb{C}\backslash\mathbb{R}_+$, $\mathrm{Im}\ \sqrt{z}>0$. Particularly,

$$f_{-\iota}^1(r,\theta) = N_1 e^{\iota\pi(1-\alpha)/2}H_{1-\alpha}^{(1)}(kr)e^{-\iota\theta}, \quad f_{-\iota}^2(r,\theta) = N_2 e^{\iota\pi\alpha/2}H_\alpha^{(1)}(kr), \tag{35}$$

where $k = \sqrt{-\iota} = e^{\iota 3\pi/4}$.

A word of warning should be said here. We use the branch $(e^{\iota\varphi})^\nu = e^{\iota\varphi\nu}$, for $0\leqslant\varphi<2\pi$, which differs from the usual choice made in surveys such as Refs. 10 and 11.

Next we compute the matrix $P(z,z')$:

$$P^{\ell,m}(z,z') := \langle f_z^\ell, f_{z'}^m\rangle = P^{\ell,\ell}(z,z')\delta_{\ell,m}, \quad \ell,m = 1,2. \tag{36}$$

Since $\overline{K_\nu(z)} = K_\nu(\overline{z})$ and using the formula 6.521 of Ref. 11,

$$\int_0^\infty K_\nu(ar)K_\nu(br)r\ dr = \frac{\pi(ab)^{-\nu}(a^{2\nu}-b^{2\nu})}{2\sin(\pi\nu)(a^2-b^2)}, \tag{37}$$

we find that

$$P^{\ell,\ell'}(z,z')=N_\ell^2\,\frac{4e^{-\iota\pi\nu}}{\sin(\pi\nu)}\,\frac{(\overline{z})^\nu-(z')^\nu}{z-z'}, \tag{38}$$

where

$$\nu=\begin{cases}1-\alpha, & \text{if } \ell=1,\\ \alpha, & \text{if } \ell=2.\end{cases}$$

Making use of the identity

$$\int_0^\infty |H_\nu^{(1)}(kr)|^2 r\,dr=(\pi\,\cos(\pi\nu/2))^{-1},\quad k=\sqrt{\pm\iota},\quad \text{Im } k>0, \tag{39}$$

we choose now

$$N_1=2^{-1/2}\,\sin^{1/2}(\pi\alpha/2),\quad N_2=2^{-1/2}\,\cos^{1/2}(\pi\alpha/2). \tag{40}$$

In this case the basis $\{f_\iota^\ell\}$ (10), as well as the basis $\{f_{-\iota}^\ell\}$, is orthonormal, i.e., in the matrix notation,

$$P(\iota,\iota)=P(-\iota,-\iota)=I, \tag{41}$$

where I is the 2×2 unit matrix. Moreover, introducing the matrix

$$D=\mathrm{diag}(1-\alpha,\alpha), \tag{42}$$

we also have

$$P(\overline{z},-\iota)=(z+\iota)^{-1}\,\sin^{-1}(\pi D/2)e^{-\iota\pi D}(e^{3\iota\pi D/2}-z^D). \tag{43}$$

Here and everywhere in what follows we use the obvious notation: if a function F is well defined on the set $\{1-\alpha,\alpha\}$, then

$$F(D):=\mathrm{diag}(F(1-\alpha),F(\alpha)).$$

We conclude that all self-adjoint extensions H^U of $\widetilde{H}$ are bijectively labelled by 2×2 unitary matrices U by means of

$$\mathscr{D}(H^U):=\left\{\psi;\psi=\widetilde{\psi}+\sum_\ell\, c_\ell\psi^\ell\right\}, \tag{44}$$

where $\widetilde{\psi}\in D(\widetilde{H})$ and

$$\psi^\ell=f_\iota^\ell+\sum_m\, f_{-\iota}^m U^{m,\ell},\quad \ell=1,2,$$

and

$$H^U\psi=\widetilde{H}\widetilde{\psi}+\sum_\ell\, c_\ell\left(\iota f_\iota^\ell-\iota\sum_m\, f_{-\iota}^m U^{m,\ell}\right). \tag{45}$$

According to the above discussion, $U=-1$ corresponds to the pure A–B operator H. Moreover, diagonal matrices U describe those extensions that preserve the angular momentum (otherwise nonconserved).

III. BOUNDARY CONDITIONS

The family of operators H^U, defined so far abstractly, can be equivalently characterized as differential operators with some well-specified boundary conditions. For this purpose, we introduce four linear functionals Φ_a^n, $n=1,2$, $a=1,2$, corresponding to the two critical angular sectors and to the first two leading terms giving the asymptotic behavior of the radial part of ψ as $r\to 0$. We define

$$\begin{aligned}
\Phi_1^1(\psi) &:= \lim_{r\to 0} r^{1-\alpha}\int_0^{2\pi}\psi(r,\theta)e^{\iota\theta}\,d\theta/2\pi,\\
\Phi_2^1(\psi) &:= \lim_{r\to 0} r^{-1+\alpha}\left[\int_0^{2\pi}\psi(r,\theta)e^{\iota\theta}\,d\theta/2\pi - r^{-1+\alpha}\Phi_1^1(\psi)\right],\\
\Phi_1^2(\psi) &:= \lim_{r\to 0} r^{\alpha}\int_0^{2\pi}\psi(r,\theta)d\theta/2\pi,\\
\Phi_2^2(\psi) &:= \lim_{r\to 0} r^{-\alpha}\left[\int_0^{2\pi}\psi(r,\theta)d\theta/2\pi - r^{-\alpha}\Phi_1^2(\psi)\right].
\end{aligned} \tag{46}$$

This definition is, of course, dictated by the asymptotic behavior of Hankels functions [cf. (9)]. So for $\psi\in\mathscr{D}(H^U)$, the part of ψ which is singular or becomes singular after differentiation by ∂_τ is given by

$$(\Phi_1^1(\psi)r^{-1+\alpha}+\Phi_2^1(\psi)r^{1-\alpha})e^{-\iota\theta}+\Phi_1^2(\psi)r^{-\alpha}+\Phi_2^2(\psi)r^{\alpha}.$$

Let us first check the symmetry condition $\langle\varphi_1,H^U\varphi_2\rangle=\langle H^U\varphi_1,\varphi_2\rangle$, for $\varphi_1,\ \varphi_2\in\mathscr{D}(H^U)$. The integration by parts gives (W=Wronskian)

$$\lim_{r\to 0+}\int_0^{2\pi} rW(\varphi_1,\varphi_2)\,d\theta=0.$$

Only the singular parts of φ_1,φ_2 can contribute and thus one arrives at

$$\Phi_1(\varphi_1)^{*}D\Phi_2(\varphi_2)=\Phi_1(\varphi_2)^{*}D\Phi_2(\varphi_1),\quad \forall\varphi_1,\varphi_2\in\mathscr{D}(H^U), \tag{47}$$

where we have introduced

$$\Phi_a(\psi):=\begin{pmatrix}\Phi_a^1(\psi)\\ \Phi_a^2(\psi)\end{pmatrix},\quad a=1,2, \tag{48}$$

and D was defined in (42).

Next we apply the functionals Φ_a^n to the functions $f^{\nearrow}_{\pm\iota}$. Namely, introduce four matrices Φ_{ab}, where the label $a=1,2$ refers to the first (respectively, the second) leading coefficient, and $b=\pm$ refers to the sign in $\mathscr{N}^{\cdot}_{\pm\iota}$. They are defined by

$$(\Phi_{ab})^{n\nearrow}:=\Phi_a^n(f^{\nearrow}_{b\iota}); \tag{49}$$

so the rows of these matrices are numbered by the angular momentum n (for the sake of convenience we shifted the index by 2, $n=m+2\in\{1,2\}$) and the columns by l, corresponding to the basis in $\mathscr{N}^{\cdot}_{\pm i}$. In view of the asymptotic expansion of the functions f^l_{bi} [cf. (10) and (35)], they read explicitly

$$\Phi_{1,+}=\Phi_{1,-}=\frac{-\iota 2^{-1/2}}{\sin(\pi\alpha)}\cos^{1/2}(\pi D/2)2^{D}\Gamma(1-D)^{-1}\exp(-\iota\pi D/4),$$

$$\Phi_{2,+}=\frac{\iota 2^{-1/2}}{\sin(\pi\alpha)}\cos^{1/2}(\pi D/2)2^{-D}\Gamma(1+D)^{-1}\exp(-\iota 3\pi D/4), \tag{50}$$

$$\Phi_{2,-}=\frac{\iota 2^{-1/2}}{\sin(\pi\alpha)}\cos^{1/2}(\pi D/2)2^{-D}\Gamma(1+D)^{-1}\exp(\iota\pi D/4).$$

From the definition (44) of $\mathscr{D}(H^U)$ follows that $\psi\in L^2(\mathbb{R}^2)$ belongs to the domain of H^U if and only if

$$\begin{pmatrix}\Phi_1(\psi)\\ \Phi_2(\psi)\end{pmatrix}\in\mathrm{Ran}\begin{pmatrix}\Phi_{1,+}+\Phi_{1,-}U\\ \Phi_{2,+}+\Phi_{2,-}U\end{pmatrix}. \tag{51}$$

Inversely, assume that we are given a couple of 2×2 matrices X_1, X_2 such that $\mathrm{rank}(X_1^t,X_2^t)=2$, and consider the boundary condition

$$\begin{pmatrix}\Phi_1(\psi)\\ \Phi_2(\psi)\end{pmatrix}\in\mathrm{Ran}\begin{pmatrix}X_1\\ X_2\end{pmatrix}. \tag{52}$$

The symmetry property (47) leads to the requirement

$$X_1^* DX_2=X_2^* DX_1. \tag{53}$$

In fact, relying on the explicit form of the matrices Φ_{ab}, one can show quite straightforwardly that for any couple X_1, X_2 with the above properties there exist exactly one 2×2 unitary matrix U and $Y\in\mathrm{GL}(2,\mathbb{C})$ such that

$$X_1Y=\Phi_{1,+}+\Phi_{1,-}U,\quad X_2Y=\Phi_{2,+}+\Phi_{2,-}U. \tag{54}$$

On the contrary, if U is unitary, then $X_a=\Phi_{a,+}+\Phi_{a,-}U$, $a=1,2$, verify (53) and $\mathrm{rank}(X_1^t,X_2^t)=2$. This way we have rederived a well-known result that all self-adjoint extensions of $\widetilde{H}$ are in one-to-one correspondence with points of a real four-dimensional submanifold of the Grassmann manifold $G_2(\mathbb{C}^4)$ determined by the equation (53).

One can rewrite the boundary condition (52) in a more convenient form when making use of the biholomorphic diffeomorphism $G_2(\mathbb{C}^4)\to G_2(\mathbb{C}^4)^\sharp$. Here $G_2(\mathbb{C}^4)^\sharp$ stands for the Grassmann manifold in the space dual to $\mathbb{C}^4$. The points of $G_2(\mathbb{C}^4)^\sharp$ are represented by couples of matrices A_1, $A_2\in\mathrm{Mat}(2,\mathbb{C})$, with $\mathrm{rank}(A_1,A_2)=2$, modulo the left action of $\mathrm{GL}(2,\mathbb{C})$. The diffeomorphism is given by

$$\mathrm{Ran}\begin{pmatrix}X_1\\ X_2\end{pmatrix}\in G_2(\mathbb{C}^4)\mapsto\mathrm{Ran}\begin{pmatrix}A_1^t\\ A_2^t\end{pmatrix}\in G_2(\mathbb{C}^4),\quad\text{where } A_1X_1+A_2X_2=0. \tag{55}$$

The real submanifold of $G_2(\mathbb{C}^4)$ determined by (53) is mapped bijectively onto the real four-dimensional submanifold of $G_2(\mathbb{C}^4)^\sharp$ determined by

$$A_1D^{-1}A_2^*=A_2D^{-1}A_1^*. \tag{56}$$

We conclude that each self-adjoint extension of $\widetilde{H}$ is determined by a boundary condition of the type

$$A_1\Phi_1(\psi)+A_2\Phi_2(\psi)=0, \tag{57}$$

where A_1, $A_2 \in \mathrm{Mat}(2,\mathbb{C})$ verify (56) and $\mathrm{rank}(A_1,A_2)=2$. Two couples (A_1,A_2) and (A_1',A_2') define the same self-adjoint operator if and only if there exists $Y \in \mathrm{GL}(2,\mathbb{C})$ such that $A_1' = YA_1$, $A_2' = YA_2$.

Let us now restrict ourselves to an open dense subset of the manifold of boundary conditions (57) which we obtain by fixing $A_1 = I$ and setting $A_2 = -\Lambda$. The condition (57) means, of course, that

$$\begin{pmatrix} \Phi_1^1(\psi) \\ \Phi_1^2(\psi) \end{pmatrix} = \Lambda \begin{pmatrix} \Phi_2^1(\psi) \\ \Phi_2^2(\psi) \end{pmatrix}, \quad \forall \psi \in \mathscr{D}(H^U). \tag{58}$$

The restriction (56) then reads

$$D\Lambda = \Lambda^* D. \tag{59}$$

All matrices Λ obeying (59) can be parametrized by the aid of four real (or two real and one complex) parameters, namely

$$\Lambda = \begin{pmatrix} u & \alpha\overline{w} \\ (1-\alpha)w & v \end{pmatrix}, \quad \text{with } u,v \in \mathbb{R} \quad \text{and } w \in \mathbb{C}. \tag{60}$$

From (55) follows that one can choose $X_1 = \Lambda$, $X_2 = I$. In virtue of (54), this leads to a relation between Λ and U,

$$\Lambda = (\Phi_{1,+} + \Phi_{1,-}U)(\Phi_{2,+} + \Phi_{2,-}U)^{-1}, \tag{61}$$

provided the relevant matrix is invertible. We use the following parametrization of U,

$$U = e^{\iota\omega} \begin{pmatrix} qe^{\iota a} & -(1-q^2)^{1/2}e^{-\iota b} \\ (1-q^2)^{1/2}e^{\iota b} & qe^{-\iota a} \end{pmatrix}, \tag{62}$$

where a, b, q, $\omega \in \mathbb{R}$, $0 \leqslant q \leqslant 1$. The matrix $(\Phi_{2,+} + \Phi_{2,-}U)$ is invertible exactly when $d \neq 0$ where

$$d := \sin\omega + q\,\sin(a - \pi\alpha). \tag{63}$$

In this case, the parameters of the matrix Λ can be expressed explicitly in terms of a, b, q, ω:

$$\begin{aligned} u &= d^{-1}2^{2-2\alpha}\frac{\Gamma(2-\alpha)}{\Gamma(\alpha)}\left(\cos\left(\omega + \frac{\pi}{2}\,\alpha\right) + q\,\cos\left(a - \frac{\pi}{2}\,\alpha\right)\right), \\ \overline{w} &= d^{-1}(2(1-q^2)\sin(\pi\alpha))^{1/2}e^{-\iota(b-\pi\alpha/2-\pi/4)}, \\ w &= d^{-1}(2(1-q^2)\sin(\pi\alpha))^{1/2}e^{\iota(b-\pi\alpha/2-\pi/4)}, \\ v &= -d^{-1}2^{2\alpha}\frac{\Gamma(1+\alpha)}{\Gamma(1-\alpha)}\left(\sin\left(\omega - \frac{\pi}{2}\,\alpha\right) + q\,\sin\left(a - \frac{\pi}{2}\,\alpha\right)\right). \end{aligned} \tag{64}$$

Obviously, the parameters $u = v = 0$ and $w = 0$ correspond to the pure A–B case. Moreover, the resulting Hamiltonian preserves the angular momentum if and only if Λ is diagonal (w is responsible for the nonconservation of the angular momentum).

IV. SPECTRUM AND EIGENSPACES

In order to find the spectrum one can use Krein's formula for the resolvent $R_z^U=(H^U-z)^{-1}$,

$$R_z^U=R_z+\sum_{k,\ell} f_z^k M_z^{k\ell}\langle f_{\bar{z}}^{\ell},\cdot\rangle. \tag{65}$$

Using (41), (43), and (27) we get

$$M_z^{-1}=\sin^{-1}(\pi D/2)\ (z^D e^{-\imath\pi D}(U+1)-e^{\imath\pi D/2}U-e^{-\imath\pi D/2})(U+1)^{-1}. \tag{66}$$

Since R_z^U is a rank two perturbation of R_z, H^U and H have the same absolutely continuous spectrum, namely, $[0,\infty[$.

The discrete spectrum is determined by the equation $\det M_z^{-1}=0$, i.e.,

$$\det(p^{2D}(U+1)-e^{\imath\pi D/2}U-e^{-\imath\pi D/2})=0, \tag{67}$$

where we have introduced $p>0$ by $p^2=-z$. According to the discussion below there are no non-negative eigenvalues. If $d\neq 0$ [cf. (63)], then $(e^{\imath\pi D/2}U-e^{-\imath\pi D/2})$ is invertible and, owing to (61),

$$\begin{aligned}(U+1)(e^{\imath\pi D/2}U-e^{-\imath\pi D/2})^{-1}=&-\cos^{-1/2}(\pi D/2)2^{-D}\Gamma(1-D)e^{-\imath\pi D/4}\\ &\times\Lambda\cos^{1/2}(\pi D/2)2^{-D}\Gamma(1+D)e^{-\imath\pi D/4}.\end{aligned} \tag{68}$$

In this case, (67) is equivalent to

$$\det\left(\frac{\Gamma(1-D)}{\Gamma(1+D)}\left(\frac{p}{2}\right)^{2D}\Lambda+1\right)=0. \tag{69}$$

Rewriting (69) in terms of the parameters u, v, w [cf. (60)] we get

$$\left(\left(\frac{p}{2}\right)^{2\alpha-2}+\frac{\Gamma(\alpha)}{\Gamma(2-\alpha)}u\right)\left(\left(\frac{p}{2}\right)^{-2\alpha}+\frac{\Gamma(1-\alpha)}{\Gamma(1+\alpha)}v\right)=|w|^2. \tag{70}$$

Consider the lhs of (70) as a function $F(p)$ defined on $]0,+\infty[$. Since the rhs of (70) is always non-negative some elementary analysis gives immediately the number of solutions. The number of roots of $F(p)$ equals 0 or 1 or 2. There is no root iff $u\geqslant 0$, $v\geqslant 0$, and there are two roots iff $u<0$, $v<0$ (it may happen that the two roots coincide giving a multiple root). Denote by p_1 the smallest root of $F(p)$, if any. In the case of two roots, let p_2 be the greater one. Then clearly $F(p)$ is decreasing on $]0,p_1]$ from $+\infty$ to zero and is increasing on $[p_2,+\infty]$ from zero to the asymptotic value $uv/\alpha(1-\alpha)$.

We conclude that there are

$$\begin{aligned}&\text{two eigenvalues} && \text{if } u<0,\ v<0, \quad \text{and } \det\Lambda=uv-\alpha(1-\alpha)|w|^2>0,\\ &\text{no eigenvalue} && \text{if } u\geqslant 0,\ v\geqslant 0, \quad \text{and } \det\Lambda=uv-\alpha(1-\alpha)|w|^2\geqslant 0,\\ &&& \text{one eigenvalue} \quad \text{otherwise.}\end{aligned} \tag{71}$$

We stress once more that all eigenvalues, if any, are negative. Again, it may happen (when $w=0$) that two eigenvalues coincide, producing consequently a multiple eigenvalue.

For generic α not much can be said about what are the solutions, except the case $w=0$ when $p=2(-u\Gamma(\alpha)/\Gamma(2-\alpha))^{2-2\alpha}$ if $u<0$ and/or $p=2(-v\Gamma(1-\alpha)/\Gamma(1+\alpha))^{2\alpha}$ if $v<0$, and the case $u=v=0$ when $p=2/|w|$. An interesting particular case is $\alpha=1/2$ when we can give a complete answer about the values of the two solutions [the first case in (71)],

$$p_{\pm}=2\,\frac{u+v\pm(|w|^2+(u-v)^2)^{1/2}}{|w|^2-4uv}, \tag{72}$$

and of the solution [the third case in (71)],

$$p=2\,\frac{u+v+(|w|^2+(u-v)^2)^{1/2}}{|w|^2-4uv}. \tag{73}$$

Similar but more complicated analysis can be also performed for $\alpha=1/3$, 1/4 and partially for other fractional values of α.

As far as the eigenvectors are concerned, they have to be obtained, of course, as solutions of the differential equation (6) including the corresponding boundary conditions (58). First of all, it is clear that in the sectors of the angular momentum $m\neq -1,0$, there is a complete system of generalized (and normalized) eigenfunctions coinciding with those of the pure A–B effect,

$$(2\pi)^{-1/2}J_{|m+\alpha|}(kr)e^{\imath m\theta},\quad m=\ldots,-3,-2,1,2,\ldots,k>0. \tag{74}$$

Next we pass to the sectors $m=-1$ and $m=0$.

As far as the (true) eigenfunctions are concerned, the L^2-integrability condition at infinity restricts the eigenvalue k^2 to $k^2<0$, and picks up a unique solution, up to a multiplicative constant, in each sector $m=-1$ and $m=0$ (this means that both exponential growth and oscillatory behavior at infinity are excluded). Hence setting as before $k=ip$, with $p>0$, the eigenfunction must have the form

$$\xi H^{(1)}_{1-\alpha}(\imath pr)e^{-\imath\theta}+\eta H^{(1)}_{\alpha}(\imath pr), \tag{75}$$

where $\xi,\ \eta\in\mathbb{C}$. The boundary conditions (58) lead to the following relation between ξ and η,

$$\left(I+\left(\frac{\imath p}{2}\right)^{D}\Gamma(1-D)\Lambda\Gamma(1+D)^{-1}e^{-\imath\pi D}\left(\frac{\imath p}{2}\right)^{D}\right)\begin{pmatrix}\xi\\ \eta\end{pmatrix}=0. \tag{76}$$

Setting the determinant of this system of linear equations to zero we recover the equation on eigenvalues (70). Any nontrivial solution (ξ,η) of (76) determines an eigenfunction in accordance with (75).

As far as the generalized eigenfunctions are concerned, the eigenvalue equation admits a four-parameter solution

$$(\xi J_{-1+\alpha}(kr)+\eta J_{1-\alpha}(kr))e^{-\imath\theta}+\xi' J_{-\alpha}(kr)+\eta' J_{\alpha}(kr), \tag{77}$$

where $k>0$ and $\xi,\eta,\xi',\eta'\in\mathbb{C}$. In view of the asymptotics of $J_\nu(z)\simeq\Gamma(1+\nu)^{-1}(z/2)^{\nu}(1+O(z^2))$ and again by taking into account the boundary conditions (58) we find the relation

$$\begin{pmatrix}\xi\\ \xi'\end{pmatrix}=(k/2)^{D}\Gamma(1-D)\Lambda\Gamma(1+D)^{-1}(k/2)^{D}\begin{pmatrix}\eta\\ \eta'\end{pmatrix}. \tag{78}$$

A possible choice is

$$\begin{pmatrix} \eta \\ \eta' \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \eta \\ \eta' \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and in this way we obtain two independent solutions

$$\begin{aligned} b_1(k) &= (\xi_1 J_{-1+\alpha}(kr) + J_{1-\alpha}(kr)) e^{-\iota\theta} + \eta_1 J_{-\alpha}(kr), \\ b_2(k) &= \xi_2 J_{-1+\alpha}(kr) e^{-\iota\theta} + \eta_2 J_{-\alpha}(kr) + J_{\alpha}(kr), \end{aligned} \tag{79}$$

where

$$\begin{aligned} \begin{pmatrix} \xi_1 & \xi_2 \\ \eta_1 & \eta_2 \end{pmatrix} &= (k/2)^D \Gamma(1-D) \Lambda \Gamma(1+D)^{-1} (k/2)^D \\ &= \begin{pmatrix} u \dfrac{\Gamma(\alpha)}{\Gamma(2-\alpha)} \left(\dfrac{k}{2}\right)^{2-2\alpha} & \overline{w} \dfrac{k}{2} \\ w \dfrac{k}{2} & v \dfrac{\Gamma(1-\alpha)}{\Gamma(1+\alpha)} \left(\dfrac{k}{2}\right)^{2\alpha} \end{pmatrix}. \end{aligned} \tag{80}$$

Set

$$N(k) := \begin{pmatrix} \xi_1 & \xi_2 \\ \eta_1 & \eta_2 \end{pmatrix} + \exp(\iota\pi D). \tag{81}$$

Now we seek a pair of eigenfunctions that are complete and orthonormal in the generalized sense. In order to compute the scalar products of $b_1(k)$ and $b_2(k)$, we need to know the integrals involving the products $J_\mu(ay)J_\mu(xy)$ and $J_{-\mu}(ay)J_\mu(xy)$. Recalling the relation between the functions $H_\mu^{(1)}$ and K_μ (30) and using the limit value of (32) we have

$$\int_0^\infty H_\mu^{(1)}(ay) J_\mu(xy) y \, dy = \frac{1}{\iota\pi a} \left(\frac{x}{a}\right)^\mu \left(\frac{1}{x-a-\iota 0} - \frac{1}{x+a} \right). \tag{82}$$

Next, with the help of the distributional identity $(x-\iota 0)^{-1} = \mathscr{P}(1/x) + \iota\pi\delta(x)$, where $\mathscr{P}$ denotes the principal part, we get two identities by taking the real and imaginary parts of (82),

$$\int_0^\infty J_\mu(ay) J_\mu(xy) y \, dy = \frac{1}{a} \, \delta(x-a), \tag{83}$$

$$\int_0^\infty J_{-\mu}(ay) J_\mu(xy) y \, dy = \frac{\cos(\pi\mu)}{a} \, \delta(x-a) + \frac{\sin(\pi\mu)}{\pi a} \left(\frac{x}{a}\right)^\mu \left(\mathscr{P}\left(\frac{1}{x-a}\right) - \frac{1}{x+a} \right). \tag{84}$$

Applying (83) and (84) to the solutions (79) arranged in a row $B(k) = (b_1(k), b_2(k))$, we obtain the following 2×2 matrix of scalar products,

$$\langle B(k'), B(k) \rangle = N(k')^* N(k) \, \frac{1}{k} \, \delta(k-k'). \tag{85}$$

We observe that

$$\det N(k) = \left(\frac{uv}{\alpha(1-\alpha)} - w\overline{w} \right) \frac{k^2}{4} + \frac{u\Gamma(\alpha) e^{\iota\pi\alpha} k^{2-2\alpha}}{2^{2-2\alpha}\Gamma(2-\alpha)} - \frac{v\Gamma(1-a) e^{-\iota\pi\alpha} k^{2\alpha}}{2^{2\alpha}\Gamma(1+\alpha)} - 1 \tag{86}$$

equals minus the lhs of (69), with p being replaced by $-\imath k$, and thus, in view of our analysis of Eq. (70), $\det N(k)$ is nonvanishing for all $k\geqslant 0$. Therefore $(g_1(k),g_2(k)):=B(k)N(k)^{-1}$, being given by

$$\begin{aligned} g_1(k;r,\theta)&=\det{}^{-1}\,N(k)((\xi_1\eta_2-\xi_2\eta_1+\xi_1 e^{\imath\pi\alpha})J_{\alpha-1}(kr)e^{-\imath\theta}\\ &\quad+(\eta_2+e^{\imath\pi\alpha})J_{1-\alpha}(kr)e^{-\imath\theta}-\eta_1 J_\alpha(kr)+\eta_1 e^{\imath\pi\alpha}J_{-\alpha}(kr)),\\ g_2(k;r,\theta)&=\det{}^{-1}\,N(k)(-\xi_2 e^{-\imath\pi\alpha}J_{\alpha-1}(kr)e^{-\imath\theta}-\xi_2 J_{1-\alpha}(kr)e^{-\imath\theta}\\ &\quad+(\xi_1-e^{-\imath\pi\alpha})J_\alpha(kr)+(\xi_1\eta_2-\xi_2\eta_1-\eta_2 e^{-\imath\pi\alpha})J_{-\alpha}(kr)), \end{aligned} \tag{87}$$

where ξ_1, ξ_2, η_1, and η_2 are defined by (80), form a complete orthonormal basis of generalized eigenvectors in the subspace corresponding to the absolutely continuous spectrum in the two considered sectors,

$$\langle g_j(k'),g_\ell(k)\rangle=\frac{1}{k}\,\delta(k-k')\,\delta_{j,\ell}. \tag{88}$$

V. SCATTERING

The existence of a complete and orthonormal basis of generalized eigenvectors is sufficient to show that the wave (Møller) operators $W_\pm=\lim_{t\to\pm\infty}e^{\imath tH}e^{-\imath tH^0}$ exist and are complete. In fact, they can be exhibited explicitly as well as the scattering operator $S=(W_+)^*W_-$.

From what we have said so far it is evident that $W_\pm$ and S preserve the sectors $m\neq -1,0$, and there they are exactly the same as in the pure A–B case; in particular $S_m=e^{2\imath\delta_m}$, where $\delta_m=(|m|-|m+\alpha|)\,\pi/2$. Thus we restrict ourselves to the subspace $L^2(\mathbb{R}_+,rdr)\otimes(\mathbb{C}\chi_{-1}\oplus\mathbb{C}\chi_0)$, of remaining sectors $m=-1,0$, which is, of course, also preserved by all the relevant operators (if there is no danger of confusion we denote the restriction of an operator by the same symbol).

Using the basis $\{g_1(k),g_2(k)\}$ we define a unitary operator $\mathscr{F}$ from $L^2(\mathbb{R}_+,kdk)\otimes\mathbb{C}^2$ to $L^2(\mathbb{R}_+,rdr)\otimes(\mathbb{C}\chi_{-1}\oplus\mathbb{C}\chi_0)$ by

$$\mathscr{F}[\widetilde{\psi}]=\psi:=\sum_{j=1}^{2}\int_0^\infty g_j(k)\widetilde{\psi}_j(k)k\;dk, \tag{89}$$

with the inverse $\mathscr{F}^{-1}[\psi]=\widetilde{\psi}$ given by

$$\widetilde{\psi}_j(k):=\langle g_j(k),\psi\rangle. \tag{90}$$

The operator $\mathscr{F}$ satisfies

$$e^{-\imath tH}\mathscr{F}[\widetilde{\psi}]=\mathscr{F}[e^{-\imath tk^2}\widetilde{\psi}]. \tag{91}$$

We append the superscript '^{0}' to the relevant objects like $g_j^0(k)$ or $\mathscr{F}^0$ corresponding to the free case ($\alpha=0$, $\Lambda=0$); particularly

$$g_1^0(k;r,\theta)=J_1(kr)e^{-\imath\theta},\quad g_2^0(k;r,\theta)=J_0(kr). \tag{92}$$

Now we seek a pair of 2×2 matrices $\Omega_\pm=\Omega_\pm(k)$, generally depending on k and acting in an obvious way as a multiplication operator on $L^2(\mathbb{R}_+,kdk)\otimes\mathbb{C}^2$, so that

$$W_\pm\mathscr{F}^0=\mathscr{F}\;\Omega_\pm. \tag{93}$$

Then it follows that $S\mathscr{F}^0=\mathscr{F}^0\Omega_+^*\Omega_-$, and this means that $\Sigma:=\Omega_+^*\Omega_-$ is nothing but the scattering matrix in the momentum representation (restricted to the sectors $m=-1,0$).

We have to verify that

$$\lim_{t\to\pm\infty}\|e^{-\iota tH^0}\mathscr{F}^0[\psi]-e^{-\iota tH}\mathscr{F}[\Omega_\pm\psi]\|=0. \tag{94}$$

Due to (91), the condition (94) means that

$$\lim_{t\to\pm\infty}\left\|\sum_j\int_0^\infty e^{-\iota tk^2}\left(g_j^0(k)-\sum_\ell g_\ell(k)\Omega_\pm(k)_{\ell j}\right)\psi_j(k)k\,dk\right\|=0. \tag{95}$$

It is sufficient to prove (95) only for $\psi_j(k)$ from the dense subspace $C_0^\infty(\mathbb{R}_+)\subset L^2(\mathbb{R}_+,kdk)$. By the stationary phase method, the convergence of such an integral, as $t\to\pm\infty$, will be established provided the coefficient standing in front of the term $e^{\pm\iota kr}$ vanishes. In view of the known large x expansion

$$J_\mu(x)=(2\pi x)^{-1/2}(e^{\iota(x-\pi\mu/2-\pi/4)}+e^{-\iota(x-\pi\mu/2-\pi/4)})+\mathscr{O}(x^{-3/2}), \tag{96}$$

we obtain, by looking separately at the coefficients in front of $e^{-\iota\theta}$ and 1, that

$$\begin{aligned}\Omega_+(k)^{-1}&=\mathrm{diag}(-e^{-\iota\pi\alpha/2},e^{\iota\pi\alpha/2})\widetilde{N}(k)N(k)^{-1}=\det{}^{-1}\,N(k)\,\mathrm{diag}(-e^{-\iota\pi\alpha/2},e^{\iota\pi\alpha/2})\\&\times\begin{pmatrix}\xi_1\eta_2-\xi_2\eta_1-e^{\iota\pi\alpha}(\eta_2-\xi_1)-e^{2\iota\pi\alpha} & (e^{\iota\pi\alpha}-e^{-\iota\pi\alpha})\xi_2\\ (e^{\iota\pi\alpha}-e^{\iota\pi\alpha})\eta_1 & \xi_1\eta_2-\xi_2\eta_1-e^{-\iota\pi\alpha}(\eta_2-\xi_1)-e^{-2\iota\pi\alpha}\end{pmatrix},\end{aligned} \tag{97}$$

where

$$\widetilde{N}(k):=\begin{pmatrix}\xi_1 & \xi_2\\ \eta_1 & \eta_2\end{pmatrix}+\exp(-\iota\pi D), \tag{98}$$

and also

$$\Omega_-(k)=\mathrm{diag}(-e^{-\iota\pi\alpha/2},e^{\iota\pi\alpha/2}). \tag{99}$$

Note that

$$\widetilde{N}(k)^*\widetilde{N}(k)=N(k)^*N(k) \tag{100}$$

and so $\Omega_+(k)$ is unitary. Consequently, we can express (the entries of) Σ in terms of the parameters u,v,w [see (80)] as

$$\begin{aligned}\Sigma_{11}&=\det{}^{-1}\,N(k)\left(e^{-\iota\pi\alpha}\left(\frac{uv}{\alpha(1-\alpha)}-|w|^2\right)\frac{k^2}{4}+\frac{u\Gamma(\alpha)k^{2-2\alpha}}{2^{2-2\alpha}\Gamma(2-\alpha)}-\frac{u\Gamma(1-\alpha)k^{2\alpha}}{2^{2\alpha}\Gamma(1+\alpha)}-e^{\iota\pi\alpha}\right),\\ \Sigma_{12}&=-\det{}^{-1}\,N(k)\,\iota\,\sin(\pi\alpha)\overline{wk},\\ \Sigma_{21}&=-\det{}^{-1}\,N(k)\,\iota\,\sin(\pi\alpha)wk,\\ \Sigma_{22}&=\det{}^{-1}\,N(k)\left(e^{\iota\pi\alpha}\left(\frac{uv}{\alpha(1-\alpha)}-|w|^2\right)\frac{k^2}{4}+\frac{u\Gamma(\alpha)k^{2-2\alpha}}{2^{2-2\alpha}\Gamma(2-\alpha)}-\frac{v\Gamma(1-\alpha)k^{2\alpha}}{2^{2\alpha}\Gamma(1+\alpha)}-e^{-\iota\pi\alpha}\right),\end{aligned} \tag{101}$$

where $\det N(k)$ is given by (86). Obviously, Σ is unitary.

Let us specialize these formulas to three particular cases.

(i) If $w=0$ (conserved angular momentum), then

$$\Sigma=\mathrm{diag}\left(\frac{u\Gamma(\alpha)(k/2)^{2-2\alpha}-e^{\imath\pi\alpha}\Gamma(2-\alpha)}{e^{\imath\pi\alpha}u\Gamma(\alpha)(k/2)^{2-2\alpha}-\Gamma(2-\alpha)},\ \frac{v\Gamma(1-\alpha)(k/2)^{2\alpha}+e^{-\imath\pi\alpha}\Gamma(1+\alpha)}{e^{-\imath\pi\alpha}v\Gamma(1-\alpha)(k/2)^{2\alpha}+\Gamma(1+\alpha)}\right). \quad (102)$$

(ii) If $u=v=0$ (maximal nonconservation of angular momentum), then

$$\Sigma=Q^{-1}\times\begin{pmatrix} e^{-\imath\pi\alpha}|w|^2(k^2/4)+e^{\imath\pi\alpha} & \imath\,\sin(\pi\alpha)\overline{wk} \\ \imath\,\sin(\pi\alpha)wk & e^{\imath\pi\alpha}|w|^2(k^2/4)+e^{-\imath\pi\alpha}\end{pmatrix}, \quad (103)$$

where

$$Q=|w|^2\,\frac{k^2}{4}+1.$$

(iii) If $\alpha=1/2$, then

$$\Sigma=Q^{-1}\times\begin{pmatrix} -\imath-\imath(uv-|w|^2/4)k^2+(u-v)k & -\imath\overline{wk} \\ -\imath wk & \imath+\imath(uv-|w|^2/4)k^2+(u-v)k\end{pmatrix}, \quad (104)$$

where

$$Q=-1+(uv-|w|^2/4)k^2+\imath(u+v)k.$$

We conclude this section by giving (the kernel of) the full scattering operator in the angular representation,

$$S(k;\theta,\theta_0)=\frac{1}{2\pi}\sum_{m,n=-\infty}^{\infty}S(k;m,n)e^{\imath(m\theta-n\theta_0)}, \quad (105)$$

where

$$S(k;m,n)=\begin{cases} e^{\imath\pi(|m|-|m+\alpha|)}\delta_{mn} & \text{if } m \text{ or } n\neq -1,0, \\ \Sigma_{m+2,n+2} & \text{if } m,n=-1,0\end{cases}$$

(the shift by 2 is due to our labelling of rows and columns of Σ). The double infinite sum can be performed in the sectors $m,n\neq -1,0$ (a known result borrowed from the pure A–B effect) and yields

$$S(k;\theta,\theta_0)=\cos(\pi\alpha)\,\delta(\theta-\theta_0)+\frac{1}{2\pi}\sin(\pi\alpha)\,\frac{e^{-\imath(\theta-\theta_0)/2}}{\sin(\theta/2-\theta_0/2)}$$
$$+\frac{1}{2\pi}\sum_{m,n=-1}^{0}(\Sigma_{m+2,n+2}-e^{-(2m+1)\imath\pi\alpha}\delta_{mn})e^{\imath(m\theta-n\theta_0)}. \quad (106)$$

We recall that the differential cross section in the plane is given by $d\sigma(\theta)/d\theta=(2\pi/k)|S(k;\theta,\theta_0)|^2$.

VI. CONCLUSIONS

We have introduced and studied a five-parameter family of operators that describes a quantum mechanical particle interacting with a magnetic flux α caused by an infinite thin material solenoid.

We conjecture that such a situation is described by a (singular) δ-type potential barrier and by a electromagnetic potential concentrated along the z axis (the magnetic field vanishes in the remaining region) at least in some well-defined limit when the thickness of the solenoid$\to 0$ and its length$\to\infty$. Moreover, the limit should be largely independent of the details of approximating potentials (see, e.g., Ref. 2 for the case of point interaction and Ref. 12 for pure A–B potential), the discussion of which goes, however, beyond the scope of this paper.

One of the five parameters is just the value of the flux α and the other four correspond to the strength of a singular potential barrier (sort of a combination of Dirac δ and δ') and can be interpreted as penetrability coefficients of the shielded solenoid.

A general operator of this family corresponds to an intricate mixture between the Aharonov–Bohm effect and the point interactions which is manifested more concretely via the mixing between the angular and the radial boundary conditions. It is interesting that the result we have obtained is richer than a simple superposition of the point interaction and the pure A–B effect. For instance, for a range of parameters there are two bound states possible while for the usual interactions with a support concentrated along the z axis and symmetric under the z translations (or, equivalently, for a point interaction in two dimensions) there is always (except for the free case) exactly one bound state, and for the pure A–B effect there are no bound states at all.

In the present paper we have derived an explicit formula for the scattering matrix $S(k;\theta,\theta_0)$ depending on the five parameters. Naturally, it would be of interest to examine the differential cross-section numerically, particularly its dependance on the parameters, and to deduce some physical consequences. This is what we plan to do separately.

It is a matter of experimental measurements (interference, scattering, or in condensed matter) to establish which of the extensions in our family correspond to realizable models. On general grounds, one can distinguish some class of extensions by, e.g., symmetry requirements (conservation of the angular momentum), or postulating that there are no bound states (solenoid as a repulsive barrier).

After completing most of our work we became aware of Ref. 13 which has an overlap with some of our results, see also Ref. 14.

ACKNOWLEDGMENT

P.S. wishes to gratefully acknowledge the partial support from Grant No. 201/94/0708 of Czech GA.

[1] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, and H. Holden, *Solvable models in quantum mechanics* (Springer, New York, 1988).
[2] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, and H. Holden, ''Point interactions in two dimensions: Basic properties, approximations and applications to solid state physics,'' J. reine angew. Math. **380**, 87 (1987).
[3] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).
[4] W. C. Henneberger, ''Aharonov–Bohm scattering and the velocity operator,'' J. Math. Phys. **22**, 116 (1981).
[5] S. N. M. Ruiseenars, ''The Aharonov–Bohm Effect and Scattering Theory,'' Ann. Phys. (NY) **146**, 1 (1983).
[6] M. V. Berry, ''Exact Aharonov–Bohm wavefunction obtained by applying Dirac's magnetic phase factor,'' Eur. J. Phys. **1**, 240 (1980).
[7] F. Nicoleau, ''Matrices de diffusion pour l'opérateur de Schrödinger en présence d'un champ magnétique. Phénomene de Aharonov–Bohm,'' Ann. Inst. Henri Poincaré Phys. Theor. **61**, 329 (1994).
[8] P. Šťovíček, ''Scattering on a finite chain of vertices,'' Duke Math. J. **76**, 303 (1994).
[9] L. Dabrowski and H. Grosse, ''On nonlocal point interactions in one, two, and three dimensions,'' J. Math. Phys. **26**, 2777 (1985).
[10] A. Erdelyi (Ed.), *Tables of Integral Transforms* (Bateman Manuscript Project) (McGraw–Hill, New York, 1954), Vol. II.
[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1965).
[12] C. Magni and F. Valz-Gris, ''Can elementary quantum mechanics explain the Aharonov–Bohm effect?'' J. Math. Phys. **36**, 177 (1985).
[13] R. Adami, Tesi di Laurea, Pisa (1996) (in Italian, unpublished).
[14] R. Adami and A. Teta, ''On the Aharonov–Bohm effect,'' quant-ph/9702048.

On eigenvalues in gaps for perturbed magnetic Schrödinger operators

Rainer Hempel
Institut für Analysis, TU Braunschweig, Pockelsstr. 14, 38106 Braunschweig, Germany

Serge Z. Levendorskiĭ
Rostov Institute of National Economy, 69 B. Sadovaya, Rostov-on-Don, 344798, Russia

(Received 21 January 1997; accepted for publication 5 June 1997)

We study Schrödinger operators H_0 with a gap in the essential spectrum, perturbed by either a decreasing electric potential or a decreasing magnetic field; in both cases the strength of the perturbation is measured by a coupling constant $\lambda \geqslant 0$. Here we are mainly interested in the asymptotic behavior (as $\lambda \to \infty$) of certain counting functions for the eigenvalues that are produced by the perturbation inside the spectral gap. The case where we perturb by a potential can be handled using current technology, even if H_0 contains a fixed magnetic background. For perturbations by magnetic fields, however, we require rather strong assumptions—like exponential decay of the perturbations—to obtain a lower bound on the counting function. To gain some additional intuition, we use separation of variables in the closely related model of a Schrödinger operator with constant magnetic field in $\mathbf{R}^2$, perturbed by a rotationally symmetric magnetic field that decays at infinity.
© *1998 American Institute of Physics.* [S0022-2488(98)03501-4]

I. INTRODUCTION

(1) We consider Schrödinger operators with a spectral gap, perturbed by either a decreasing electric potential or a decreasing magnetic field. The strength of these perturbations depends on a coupling parameter λ. With λ growing, eigenvalues may move into the gap or out of the gap. Most of our results concern (lower) bounds for the number of eigenvalues that cross a fixed energy level E in the gap, as the coupling parameter increases from 0 to λ. Using the currently available technology, the first case (electric perturbations) can be handled in reasonable generality—even if the unperturbed reference operator H contains magnetic terms. In contrast, the second case (magnetic perturbations) resists virtually all attempts to adapt the methods that are so effective in the first case. It is one of the aims of the present paper to point out the differences between these two apparently similar situations.

To be more precise, we take for the unperturbed operator H a Schrödinger operator,

$$H = H(\vec{a}, V) = (-i\nabla - \vec{a}(x))^2 + V(x),$$

in $\mathbf{R}^n$, with a real electric potential $V \in L_\infty(\mathbf{R}^n)$ and a magnetic potential $\vec{a} \in C^1(\mathbf{R}^n, \mathbf{R}^n)$. Our basic assumption is that H has a gap in the essential spectrum; for instance, this is the case for many periodic potentials V, for some periodic magnetic fields, and also for combinations of electric and magnetic potentials. We also fix some energy level E inside the gap.

(2) In the first part of the paper (Sections II and III) we suppose that H_λ is defined as a relatively compact perturbation of H by an electric potential λW,

$$H_\lambda = H(\vec{a}, V) - \lambda W, \quad \lambda \in \mathbf{R},$$

where $\lambda \in \mathbf{R}$ is a coupling constant and $W \in L_\infty(\mathbf{R}^n)$ goes to zero at infinity.

As λ grows, some eigenvalue branches of the family H_λ may cross E, and it is a natural question to determine for which perturbations such crossings will occur. This question arises, for instance, in the quantum mechanical theory of impurities in insulators and semiconductors—see, e.g., the bibliography in Refs. 1, 2, 3.

In a second step, one would like to find the asymptotics of the corresponding *counting function*,

$$\widetilde{N}(E,H_\lambda) := \sum_{0<\mu<\lambda} \dim \ker(H_\mu - E), \quad \text{as} \quad \lambda \to +\infty.$$

Here the case of perturbations by non-negative or non-positive electric potentials is sufficiently well-understood[1,2,4,5–10] (see also the bibliography in Refs. 11, 12), since in this case the Birman–Schwinger principle provides a reduction to compact symmetric operators, and one can use standard variational techniques. Some of these methods can be successfully applied to divergence type operators.[13,14]

The case of W changing sign is more difficult since the corresponding Birman–Schwinger kernel is non-symmetric and does not easily yield useful spectral information. Still, an approach developed in Refs. 1, 4, 15 made it possible to investigate the existence of eigenvalue branches crossing E and to obtain *lower bounds* for the counting function for several types of decay; in Ref. 9, the ideas of Refs. 4, 15 were modified and all cases of polynomial decay were investigated. Note that *upper bounds* for the counting function are much more difficult due to possible oscillations of a single eigenvalue branch around a fixed level E (cf., however, the special results[16] for W of compact support).

So far, perturbations by a potential W with changing sign have been considered in the case of $\vec{a}=0$ only, and the first aim of our paper is to obtain similar results for perturbations of magnetic Schrödinger operators. We find that only minor changes in the scheme of Refs. 4, 15, 9 are needed, and we state and prove the corresponding results in Sections II and III. Since the magnetic case is quite similar to the non-magnetic one, we confine ourselves to the most typical cases.

(3) The second—and main—aim of the paper is to find out what the same ideas will accomplish in the case of perturbations by a magnetic potential $\vec{a}$ that decays at infinity (Section IV). Here we study magnetic Schrödinger operators of the form

$$H_\lambda = H(\lambda\vec{a},V) = (-i\nabla - \lambda\vec{a}(x))^2 + V.$$

Since a magnetic perturbation is never positive definite in the sense of quadratic forms, it is clear that the situation is difficult from the very beginning. A rather restricted class of perturbations (where the magnetic potential has compact support) was considered in Ref. 17. Here the perturbation was given by a constant field in a ball of radius R, with the support of the vector potential being contained in a ball of radius $R+1$, say. It was shown that, for R sufficiently large, cR^n eigenvalue branches cross E, as the coupling goes to ∞; here c is some positive constant. The intuition behind this result is based on the ''repulsive'' effect of a strong magnetic field, expressed in terms of a well-known quadratic form bound.[18] This bound suggests that a sufficiently strong (and not too localized) magnetic field should push eigenvalues up, so that at least some of them have to cross a given E. A further discussion of this intuition can be found in Ref. 19. In the present paper, we show that the technique of Refs. 4, 15, 9 leads to an asymptotic lower bound for the counting function $\widetilde{N}(E,H_\lambda)$, similar to the case of perturbations by an electric potential, under the—rather strong—assumption that the field and the vector potential decay rapidly, whereas one might conjecture that only some decay of the magnetic field is needed. We state the corresponding result in Theorem IV.1 and indicate which parts of the proof fail if we try to relax our assumption of rapid decay.

We still feel that the basic intuition may be correct and it is only the inadequacy of the methods used so far which prevents proving a more general result. Some support for this view comes from a closely related model problem which admits the separation of variables in polar coordinates. We present this example at the end of Section IV and show that in this case, under natural and rather general conditions, eigenvalue branches behave according to our intuition.

II. PERTURBATIONS OF MAGNETIC SCHRÖDINGER OPERATORS BY DECAYING ELECTRIC POTENTIALS: MAIN THEOREMS

We restrict our discussion to the most important and typical cases; more general results could be easily obtained by following the line of argument given for non-magnetic operators in Ref. 15 and for perturbations of magnetic operators by an electric potential (which does not change sign) in Ref. 7.

We assume that $\vec{a} \in C^1(\mathbf{R}^n, \mathbf{R}^n)$ has bounded derivatives, that $V \in L_\infty(\mathbf{R}^n)$, and that there exists a gap in the essential spectrum of H. For instance, the last condition is satisfied if either the magnetic tensor,

$$B = [b_{jk}]_{j,k=1}^n, \quad b_{jk} = \partial_j \vec{a}_k - \partial_k \vec{a}_j,$$

is constant and non-degenerate (this requires n to be even) and V is small, or $H(0,V)$ has a gap in the spectrum and the vector potential a is small (in the case of the constant B, it suffices to require that B be small).[20] There are also situations where a periodic magnetic field creates a structure of ''wells and barriers'' that lead to gaps in the essential spectrum;[19] this works for any dimension $\geqslant 2$.

If W decays slowly at infinity, we will have to assume in addition that $\rho(\cdot, H)$, the integrated density of states for H, exists:

$$\rho(\lambda, H) = \lim_{\text{vol } Q_R \to +\infty} \frac{1}{\text{vol } Q_R} N(\lambda, H(Q_R)), \tag{1}$$

where Q_R are expanding cubes, and $H(Q_R)$ is the operator of the Dirichlet problem for H in Q_R.

We suppose that $W \in L_\infty$ decays at infinity as $|x|^{-m}$ with some $m > 0$:

$$|W(x)| \leqslant C(1+|x|)^{-m}; \tag{2}$$

if $m \in (0,2]$, we assume that it stabilizes to a positively homogeneous continuous function $\widetilde{W}: \mathbf{R}^n \backslash 0 \to \mathbf{R}$ of order $-m$:

$$(W(tx) - \widetilde{W}(tx)) t^m \to 0, \text{ as } t \to +\infty, \tag{3}$$

uniformly in $x \in S_{n-1}$.

Set $N(\lambda, H-E, W) = \widetilde{N}(E, H - \lambda W)$ and $W_\pm(x) = \max\{\pm W(x), 0\}$.

Our first result deals with W decaying faster than quadratically; this is the easiest case.

II.1. Theorem: *Let (2) hold with $m > 2$. Then*

$$\liminf_{\lambda \to +\infty} N(\lambda, H-E, W) \lambda^{-n/2} \geqslant c(W_+), \tag{4}$$

where

$$c(W_+) = (2\pi)^{-n} |v_n| \int_{\mathbf{R}^n} W_+(x)^{n/2} dx, \tag{5}$$

and $|v_n|$ *is the volume of the unit ball of* $\mathbf{R}^n$.

We next discuss the borderline case where $m=2$ in (2), (3).

II.2. Theorem: *Let (2) and (3) hold with* $m=2$. *Then*

$$\liminf_{\lambda\to+\infty} N(\lambda,H-E,W)\lambda^{-n/2}(\ln\lambda)^{-1}\geqslant c_{\mathrm{int}}(\widetilde{W}_+), \tag{6}$$

where

$$c_{\mathrm{int}}(\widetilde{W}_+)=(2\pi)^{-n}\frac{1}{2}|v_n|\int_{S_{n-1}}\widetilde{W}_+(x)^{n/2}dS(x). \tag{7}$$

We finally have the case of slow decay where $m\in(0,2)$.

II.3. Theorem: *Let (2) and (3) hold with* $m\in(0,2)$, *and let a limit (1) exist for all* λ. *Then*

$$\liminf_{\lambda\to+\infty} N(\lambda,H-E,W)\lambda^{-n/m}\geqslant|c_{\mathrm{st}}^{+}(\widetilde{W}_+)-c_{\mathrm{st}}^{-}(\widetilde{W}_-)|, \tag{8}$$

where

$$c_{\mathrm{st}}^{+}(\widetilde{W}_+)=\int_{\mathbf{R}^n}dx\int_{E<t<E+\widetilde{W}_+(x)}d\rho(t,H), \tag{9}$$

$$c_{\mathrm{st}}^{-}(\widetilde{W}_-)=\int_{\mathbf{R}^n}dx\int_{E-\widetilde{W}_-(x)<t<E}d\rho(t,H). \tag{10}$$

Note that the integral in (9) converges if and only if $m\in(0,2)$.

In Ref. 10, formulas for the principal term of the asymptotics for W not changing sign are obtained. The results of Ref. 10 and formulas (4), (6), and (8), taken together, show that asymptotically,

$$N(\lambda,H-E,W)\geqslant|N(\lambda,H-E,W_+)-N(\lambda,H-E,-W_-)|.$$

III. PROOFS OF THEOREMS II.1–II.3

A. Some preliminary definitions and results

Let $\mathscr{A}$ be a quadratic form in $\mathscr{H}$, the Hilbert space, with the domain $D(\mathscr{A})$, and let $V\subset D(\mathscr{A})$ be a subspace. We assume that $\mathscr{A}$ is closable, and we set

$$\mathscr{N}(\lambda,\mathscr{A},V)=\sup\{\dim L\,|\,L\subset V,\mathscr{A}[u]<\lambda\|u\|^2_{\mathscr{H}},\forall u\in L\}.$$

For A an operator in $\mathscr{H}$ and $\mathscr{A}[u]=\langle Au,u\rangle_{\mathscr{H}}$, $D(\mathscr{A})=D(A)$, write $\mathscr{N}(\lambda,A,V)$ instead of $\mathscr{N}(\lambda,\mathscr{A},V)$.

We also set $N((a,b);A)=\dim P_{(a,b)}(A)$, where $P_{(a,b)}(A)$ is a spectral projection of A, and we set $N(\lambda;A)=N((-\infty,\lambda),A)$.

Let $\vec{a}\in C^1(\mathbf{R}^n,\mathbf{R}^n)$ with bounded derivatives, and let $U\subset\mathbf{R}^n$ be an open domain. Set

$$H^1_{\vec{a}}(\mathbf{R}^n)=\left\{u\in L_2(\mathbf{R}^n)\,|\,\|u\|_{\vec{a},1}=\left(\sum_{|\alpha|\leqslant 1}\|(D+\vec{a}(x))^{\alpha}u\|^2_{L_2}\right)^{1/2}<\infty\right\}. \tag{11}$$

This is a Hilbert space, and $C_0^{\infty}(\mathbf{R}^n)$ is a dense subspace.

Denote by $\mathring{H}^1_{\vec{a}}(U)$ the closure of $C_0^\infty(U)$ in $H^1_{\vec{a}}(\mathbf{R}^n)$. As usual, we will not distinguish in the notation between a function in $\mathring{H}^1_{\vec{a}}(U)$ and its trivial extension by 0 to all of $\mathbf{R}^n$.

It is a standard fact that $H(\vec{a},V):C_0^\infty(\mathbf{R}^n)\to L_2(\mathbf{R}^n)$ is essentially self-adjoint under the above assumptions; the closure is denoted by $H(\vec{a},V)$, too.

The proof of the following lemma is a variant of standard arguments in the theory of elliptic operators—only the part of the operator ∇ is now being played by its magnetic analogue $\nabla+i\vec{a}(x)$, and usual Sobolev spaces are being replaced by magnetic ones.

Let U be a bounded domain with the sufficiently smooth boundary, and let $H(U)$ be the operator of the Dirichlet problem for H in U.

III.1. Lemma: Let Π *be a bounded operator in* $L_2(U)$ *such that* $H(U)+\Pi$ *is invertible. Then there exists a constant* C [*depending only on the norms of the operators* Π *and* $(H(U)+\Pi)^{-1}$], *such that*

$$\|u\|_{\vec{a},1}\leqslant C\|(H(U)+\Pi)u\|_{L_2},\quad \forall u\in D(H(U)). \tag{12}$$

B. Some bounds for counting functions

Denote by B_l an open ball of radius l, centered at the origin, and let $\Omega(R,R_1)$ stand for any of the sets $Q_R, B_R, \overline{Q_R}, \overline{B_R}, \overline{B_R}\setminus B_{R_1}, Q_R\setminus\overline{Q_{R_1}}$, and $\Omega_0(R,R_1)$ for any of the sets $B_R, Q_R, B_R\setminus\overline{B_{R_1}}, Q_R\setminus\overline{Q_{R_1}}$.

The following lemma (even in a more general form) is well-known—see, e.g., Ref. 21.

III.2. Lemma: Let $s>0$ *be an integer.*

Then there exists $C=C_{s,n}$ *such that for any* $\lambda>0$ *and any* $R/2>R_1>1$,

$$\mathcal{N}(\lambda;(-\Delta)^s;C_0^\infty(\Omega_0(R,R_1)))\leqslant C\lambda^{n/2s}\ \text{meas}\ \Omega_0(R,R_1), \tag{13}$$

$$\mathcal{N}(\lambda;\|\cdot\|^2_{H^s};C_0^\infty(\Omega(R,R_1)))\leqslant C(1+\lambda^{n/2s})\text{meas}\ \Omega(R,R_1). \tag{14}$$

III.3. Lemma: Let a *and* V *be as above, so that, in particular,* V *and* ∇a *are bounded. Then there exists* C *such that for any* $\lambda>0$, *any* $R/2>R_1>1$, *and any* $x_0\in\mathbf{R}^n$,

$$\mathcal{N}(\lambda;H(\vec{a},V);C_0^\infty(x_0+\Omega_0(R,R_1)))\leqslant C(1+\lambda^{n/2})\text{meas}\ \Omega_0(R,R_1), \tag{15}$$

$$\mathcal{N}(\lambda;\|H(\vec{a},V)\cdot\|^2;C_0^\infty(\Omega(R,R_1)))\leqslant C(1+\lambda^{n/4})\text{meas}\ \Omega(R,R_1). \tag{16}$$

Proof: It is easy to see that there exist $C_j=C_j(N,V,R)$, $j=1,2$, such that

$$\langle(-\Delta)u,u\rangle_{L_2}\leqslant C_1\langle H(\vec{a},V)u,u\rangle_{L_2}+C_2\|u\|^2_{L_2},\quad \forall u\in\mathbf{C}_0^\infty(\Omega_0(R,R_1)),$$

and

$$\|u\|^2_{0,2}\leqslant C_1\|H(\vec{a},V)^2u\|_{L_2}+C_2\|u\|^2_{L_2},\quad \forall u\in\mathbf{C}_0^\infty(\Omega(R,R_1)).$$

By applying the usual variational techniques,[9,22] we see that (15) and (16) follow from (13) and (14), respectively, with the constant depending on x_0,R,R_1, though.

To reduce the proof in the general case to that of a fixed R and $x_0=0$, construct a partition of unity,

$$\sum_{k\in\mathbf{Z}^n}\chi^2_{r,k}=1,\ \text{supp}\ \chi_{r,k}\subset r\cdot k+(-r^{1/2}/4,r+r^{1/2}/4)^n, \tag{17}$$

$$|D^{\alpha}\chi_{r,k}| \leqslant C_{\alpha} r^{-|\alpha|/2}, \quad \forall \alpha, \tag{18}$$

with C_{α} independent of $r \geqslant 1/4$. It follows from (18) that

$$|\langle [H(\vec{a},V),\chi_{r,k}]\chi_{r,k}u,u\rangle| \leqslant C_1 r^{-1/2}(\langle H(\vec{a},V)u,u\rangle + C_2\|u\|^2_{L_2}), \tag{19}$$

with C_1, C_2 independent of r. Due to (17) and the locality of $H(A,V)$, (19) holds for $|\langle \Sigma_k [H(A,V),\chi_{r,k}]\chi_{r,k}u,u\rangle|$ as well, and we conclude that there exist C_1, C_2 such that uniformly in $r \geqslant 1$,

$$\langle H(\vec{a},V)u,u\rangle \geqslant (1 - C_1 r^{-1/2}) \sum_k \langle H(\vec{a},V)\chi_{r,k}u, \chi_{r,k}u\rangle - C_2 r^{-1/2}\|u\|^2_{L_2}. \tag{20}$$

Now, (20) and usual variational arguments give for $r = 2C_2^2$ the estimate

$$\mathcal{N}(\lambda; H(\vec{a},V); \mathbf{C}_0^{\infty}(x_0 + \Omega_0(R,R_1))) \leqslant \sum_{k \in I(R,R_1,r)} \mathcal{N}(C_3(1+\lambda); H(\vec{a},V); \mathbf{C}_0^{\infty}(U^+_{r,k})) \tag{21}$$

where $U^+_{r,k} = r\cdot k + (-r^{1/2}/4, r + r^{1/2}/4)^n$, and

$$I(R,R_1,r) = \{k \mid U^+_{r,k} \cap (x_0 + \Omega_0(R,R_1)) \neq \emptyset\};$$

C_3 is independent of k, of course.

$H(\vec{a},V)$ is unitarily equivalent to $H(\vec{a} - \vec{a}(rk), V_k)$, where $V_k(\cdot) = V(rk + \cdot)$, therefore

$$\mathcal{N}(C_3(1+\lambda); H(\vec{a},V); \mathbf{C}_0^{\infty}(U^+_{r,k})) = \mathcal{N}(C_3(1+\lambda); H(\vec{a} - \vec{a}(rk), V_k); \mathbf{C}_0^{\infty}(U^+_{r,0})),$$

and since V_k and $(\vec{a} - \vec{a}(rk))|_{U^+_{r,k}}$ are uniformly bounded with respect to k, we deduce from (13) an estimate,

$$\mathcal{N}(C_3(1+\lambda); H(\vec{a},V); \mathbf{C}_0^{\infty}(U^+_{r,k})) \leqslant C_4(1+\lambda^{n/2}) \text{meas } U^+_{r,k}, \tag{22}$$

with C_4 independent of k. By substituting (22) into (21), we obtain (13).

(14) is proved similarly, (12) being used. This concludes the proof of the Lemma. ■

C. A variational principle for a lower bound for $N_+(\lambda; H-E; W)$

Let E be in a gap of H and fix $a < a' < E < b' < b$ such that the interval $[a,b]$ lies in the gap. Basic for the approach developed in Ref. 1 are exponential decay estimates for the resolvent operator $(H-E)^{-1}$ of the type

$$\|\chi_R (H-E)^{-1}(1-\chi_n)\| \leqslant C e^{-\alpha(n-R)}, \quad a < E < b, \tag{23}$$

where $H = -\Delta + V$ and χ_R is the characteristic function of the ball of radius R, centered at the origin, for $n \geqslant 2R \geqslant 1$, with suitable constants C and $\alpha > 0$. We now have to work with the magnetic Sobolev spaces H^1_a and $\mathring{H}^1_a$ instead of H^1 and $\mathring{H}^1$, respectively, but it is easy to see that the ''boosting argument'' used in Ref. 1 in the proof of (23) for Schrödinger operators $-\Delta + V$ can be adapted to the magnetic case. In particular, if $U \subset \mathbf{R}^n$ is open, and $H(U)$ denotes the operator $H = H(\vec{a},V)$ acting in $L_2(U)$ with (homogeneous) Dirichlet boundary conditions, the above estimate and Lemma 2.1 show that eigenfunctions of $H(B_n)$ that are associated with

eigenvalues $\mu \in [a',b']$ are localized in a boundary region $\{n/2<|x|<n\}$, up to exponentially small errors. We may thus obtain the ''magnetic analogues'' of two basic results from Refs. 4, 15 which are formulated below.

Fix functions $\psi_l \geqslant 0$ with the properties

$$\operatorname{supp} \psi_l \subset \mathbf{R}^n \setminus B_{l/4}, \quad \psi_l|_{\mathbf{R}^n \setminus B_{l/2}} = 1,$$

$$|D^\alpha \psi_l| \leqslant C_\alpha l^{-|\alpha|}, \quad \forall \alpha,$$

uniformly in $l \geqslant 1$, and put

$$\Pi_l = P_{(a,b)}(H(B_l)), \quad c = b-a, \quad \widetilde{H}_l = H(B_l) + c\psi_l \Pi_l \psi_l .$$

III.4. Lemma: There exists $l_0>0$ such that $[a',b'] \cap \sigma(\widetilde{H}_l) = \emptyset$, for all $l>l_0$.

III.5. Theorem: *For $0<\lambda<\lambda'$, we have*

$$N(\lambda', H-E, W) \geqslant \limsup_{l \to +\infty} |N(E, \widetilde{H}_l) - N(E, \widetilde{H}_l - \lambda W)|.$$

D. Proof of Theorem II.1

Fix $s \in (1/m, 1/2)$, set $R = \lambda^s$, and construct functions $\chi_R^{\pm}$ with the following properties:

$$0 \leqslant \chi_R^{\pm} \leqslant 1, \quad (\chi_R^+)^2 + (\chi_R^-)^2 = 1, \tag{24}$$

$$\operatorname{supp} \chi_R^+ \subset \mathbf{R}^n \setminus B_{R-R^{1/2}}, \quad \operatorname{supp} \chi_R^- \subset B_R, \tag{25}$$

$$|D^\alpha \chi_R^{\pm}| \leqslant C_\alpha R^{-|\alpha|/2}, \quad \forall \alpha. \tag{26}$$

We may apply Theorem III.5 with $l \geqslant R$, and we shall obtain bounds for $N(E;\widetilde{H}_l)$ and for $N(E,\widetilde{H}_l - \lambda W)$ by using the cut-off functions $\chi_R^{\pm}$.

Due to (24) and (26) we have

$$\langle \widetilde{H}_l u, u \rangle = \sum_{j=\pm} \langle \widetilde{H}_l \chi_R^j u, \chi_R^j u \rangle + T_R(u),$$

with $T_R(u)$ admitting an estimate $|T_R(u)| \leqslant CR^{-1/2} \langle \widetilde{H}_l u, u \rangle$. Here and below we may and shall assume that $H \geqslant 1$ (hence $\widetilde{H}_l \geqslant 1$). If this is not the case we replace V and E by $V + C_0$ and $E + C_0$, respectively, where C_0 is large enough. Therefore, for some $C_1, C_2 > 0$,

$$\langle (\widetilde{H}_l - E)u, u \rangle \geqslant (1 - C_1 R^{-1/2}) \sum_{j=\pm} \langle (\widetilde{H}_l - E - C_2 R^{-1/2}) \chi_R^j u, \chi_R^j u \rangle,$$

and due to (24) and (25),

$$N(E, \widetilde{H}_l) \leqslant (1 - C_1 R^{-1/2})^{-1} N(E + C_2 R^{-1/2}, \widetilde{H}_{l,R}) + N(E + C_2 R^{-1/2}; H(B_R)), \tag{27}$$

where $\widetilde{H}_{l,R} = H(B_l \setminus \overline{B_{R_1}}) + c\psi_l \Pi_l \psi_l$ is an operator in $L_2(B_l \setminus \overline{B_{R_1}})$, and $R_1 = R - R^{1/2}$.

Similarly,

$$N(E, \widetilde{H}_l - \lambda W) \leqslant N(E, (1 - C_2 R^{-1/2}) \widetilde{H}_{l,R} - \lambda W) + N(E, (1 - C_2 R^{-1/2}) H(B_R) - \lambda W). \tag{28}$$

Estimates from below are evident:

$$N(E,\widetilde{H}_l)\geqslant N(E,\widetilde{H}_{l,R})+N(E,H(B_{R_1})), \tag{29}$$

$$N(E,\widetilde{H}_l-\lambda W)\geqslant N(E,\widetilde{H}_{l,R}-\lambda W)+N(E,H(B_{R_1})-\lambda W). \tag{30}$$

Due to our choice of R and (2) we have $|\lambda W|=o(1)$ outside B_{R_1} therefore there exists (a function) $\epsilon=\epsilon(R)\to 0$ as $R\to+\infty$ such that

$$N(E,(1-C_2R^{-1/2})\widetilde{H}_{l,R}-\lambda W)\leqslant N(E+\epsilon,\widetilde{H}_{l,R}),$$

$$N(E,\widetilde{H}_{l,R}-\lambda W)\geqslant N(E-\epsilon,\widetilde{H}_{l,R}),$$

and one easily deduces from (27)–(30) an estimate,

$$|N(E,\widetilde{H}_l)-N(E,\widetilde{H}_l-\lambda W)|\geqslant\max\{\mathscr{D}_+(\lambda,R),\mathscr{D}_-(\lambda,R)\}-\mathscr{D}_0(l,R), \tag{31}$$

where

$$\mathscr{D}_0(l,R)=N(E+\epsilon,\widetilde{H}_{l,R})-N(E-\epsilon,\widetilde{H}_{l,R}),$$

$$\mathscr{D}_\pm(\lambda,R)=\mp N(E\pm\epsilon_1,H(B_{R\pm R^{1/2}}))\pm N(E,(1\pm\epsilon_1)H(B_{R\mp R^{1/2}})-\lambda W),$$

and $\epsilon,\epsilon_1\to 0$ as $R\to+\infty$.

Construct $\phi_C^\pm$ such that $0\leqslant\phi_C^\pm\leqslant 1$,

$$(\phi_C^+)^2+(\phi_C^-)^2=1,\quad |D^\alpha\phi_C^\pm|\leqslant C_\alpha C^{-|\alpha|},\quad \forall\alpha,$$

$$\text{supp }\phi_C^-\subset B_{R-R^{1/2}+2C},\quad \phi_C^-=1\ \text{ on }\ B_{R-R^{1/2}+C},$$

with C_α independent of $C\geqslant 1$. For $u\in D(\widetilde{H}_{l,R})$, we may identify ϕ_C^+u with a function from $D(\widetilde{H}_l)$. Using this observation and the density of $\mathbf{C}_0^\infty(\overline{B_l\backslash B_{R_1}})\cap D(\widetilde{H}_{l,R})$ in $D(\widetilde{H}_{l,R})$, we obtain similarly to (27),

$$\mathscr{D}_0(l,R)\leqslant N(2\epsilon^2,(\widetilde{H}_{l,R}-E)^2)\leqslant\mathscr{N}(2\epsilon^2+\nu,\|(\widetilde{H}_l-E)\cdot\|^2,D(\widetilde{H}_l))+\mathscr{N}(2\epsilon^2+\nu,\|(H-E)\cdot\|^2,\mathbf{C}_0^\infty(B_{R_1+2C}\backslash B_{R_1})), \tag{32}$$

where $\nu\to 0$ as $C\to+\infty$. If $2\epsilon^2+\nu$ is small enough, the first term in the RHS of (32) is zero because E is in a gap of $\widetilde{H}_l$, due to Lemma III.4. Thus, it suffices to estimate the second term, and one can easily deduce from Lemma III.3 an upper bound via $C_2(C)R^{n-1}=o(R^n)$.

Thus, we may rewrite (31) as

$$|N(E,\widetilde{H}_l)-N(E,\widetilde{H}_l-\lambda W)|\geqslant\max\{\mathscr{D}_+(\lambda,R),\mathscr{D}_-(\lambda,R)\}+o(R^n). \tag{33}$$

One can compute the asymptotics of $N(E\pm\epsilon_1,H(B_{R\pm R^{1/2}}))$ and that of $N(E,(1\mp\epsilon_1)H(B_{R\pm R^{1/2}})-\lambda W)$ using the IMS-localization formula.[23] On the region where $|\lambda W|$ is not too large, we use the crude estimates provided by Lemma III.3 (and the corresponding terms can be included in the o-term of the asymptotics); the region where $|\lambda W|$ is large can be partitioned into subregions such that on each subregion, $|a-a(x^j)|$ is small with respect to $|\lambda W|^{1/2}$ (here x^j is a point from the subregion). By making an appropriate gauge transformation, we may deal with a (relatively) small magnetic potential. This means that the proof of Lemma III.3 and the

classical Weyl formula for the Laplacian now give the principal term of the asymptotics (the same as in the Weyl formula) where $\lambda W \gg 1$; where $-\lambda W \gg 1$, we see that the corresponding ''local counting function'' is zero. In the result, we obtain

$$N(E \pm \epsilon_1, H(B_{R \pm R^{1/2}})) = O(R^n), \tag{34}$$

$$N(E, (1 \mp \epsilon_1) H(B_{R \pm R^{1/2}}) - \lambda W) = c(W_+) \lambda^{n/2} + o(\lambda^{n/2}), \tag{35}$$

where $c(W_+)$ is given by (5).

Clearly, (33)–(35) yield (4). ■

E. Proof of Theorem II.2

Set $R = C_0 \lambda^{1/2}$ and argue as in the previous subsection. We obtain (31) with $\epsilon \to 0$ as $C_0 \to +\infty$. Hence, if we fix $C_0 > 0$ sufficiently large, the first term in the RHS in (32) will be zero and we will obtain (33). Once again we have (34) but due to the condition $m = 2$ (35) takes the form

$$N(E, (1 \mp \epsilon_1) H(B_{R \pm R^{1/2}}) - \lambda W) = c_{\text{int}}(\widetilde{W}_+) \lambda^{n/2} \ln \lambda + o(\lambda^{n/2} \ln \lambda), \tag{36}$$

with $c_{\text{int}}(\widetilde{W}_+)$ given by (7).

(33), (34), and (36) give (6).

For the sake of completeness, we recall why (35) and (36) are different. In both cases it can be shown that the principal term of the asymptotics is given by

$$(2\pi)^{-n} \operatorname{meas}\{(x,\xi) \mid |x| \leqslant R, |\xi|^2 - \lambda W_+(x) < 0\} = \lambda^{n/2} (2\pi)^{-n} |v_n| \int_{|x| \leqslant R} W_+^{n/2}(x) dx,$$

hence, if $m > 2$ and therefore $\int_{\mathbf{R}^n} W_+(x)^{n/2} dx$ converges, then the principal term of the asymptotics is equal to $c_+(W) \lambda^{n/2}$. If $m = 2$ then

$$\begin{aligned} \int_{|x| \leqslant R} W_+(x)^{n/2} dx &= \int_{\ln \ln \lambda \leqslant |x| \leqslant R} \widetilde{W}_+(x)^{n/2} dx (1 + o(1)) + o(\ln \lambda) \\ &= \int_{\ln \ln \lambda \leqslant |x| \leqslant C_0 \lambda^{1/2}} dr r^{-1} \int_{S_{n-1}} \widetilde{W}_+(x)^{n/2} dS(x) (1 + o(1)) + o(\ln \lambda) \\ &= \frac{1}{2} \ln \lambda \int_{S_{n-1}} \widetilde{W}_+(x)^{n/2} dS(x) (1 + o(1)), \end{aligned}$$

so that the principal term is equal to $c_{\text{int}}(\widetilde{W}_+) \lambda^{n/2} \ln \lambda$. ■

F. Proof of Theorem II.3

One modifies the proof of Theorem 1.3 in Ref. 9 in the same manner as the proofs of Theorems II.1 and II.2, above are modifications of the similar theorems for the non-magnetic case in Ref. 9. ■

IV. PERTURBATIONS BY DECAYING MAGNETIC POTENTIALS

A. Rapidly decreasing magnetic potentials

In this section, we start again from the assumption that $V \in L_\infty(\mathbf{R}^n)$ and that $H = H(0,V)$ has a gap in its (essential) spectrum. We fix E in a gap, and consider $H_\lambda := H(\lambda \vec{a}, V)$ with a magnetic

potential $a \in C^1(\mathbf{R}^n, \mathbf{R}^n)$ rapidly decaying at infinity. As before, our basic strategy is to consider an approximating problem on large balls, B_l, and to compare the number of eigenvalues below E for the perturbed problem with the corresponding number for the unperturbed problem. We expect that there should be three regions which have to be treated in a different fashion.

(1) There is a region B_R where the field produced by $\lambda \vec{a}$ eliminates all (or at least almost all) eigenvalues below E, by the Avron–Herbst–Simon bound for the quadratic forms.

(2) There is a ''transition region'' $B_{R' \setminus R}$ where the field drops in strength and is too weak to move the lowest point of the spectrum above E. Under the assumption of exponential decay, however, this transition region will be small, and we can employ a trace estimate for the magnetic Schrödinger semi-groups.

(3) Finally, in the region between the radii R' and l, the perturbation $\lambda \vec{a}$ will be so small that it can be neglected.

As it turns out, all this works if we link an exponential upper bound for a with a suitable lower bound for the strength of the field B. The rates of decay of the vector potential and the field are governed by a monotonic function $q:[0,\infty) \to \mathbf{R}_+$ such that

$$q(r) \to +\infty, \quad \text{as} \quad r \to +\infty, \tag{37}$$

and ϕ, the inverse of q, enjoys the property

$$\lim_{\varepsilon \to +0} \lim_{\lambda \to +\infty} \phi(\lambda^{1 \pm \varepsilon})/\phi(\lambda) = 1. \tag{38}$$

We assume that for any $\varepsilon > 0$ there exist positive constants $c_\varepsilon, C_\varepsilon$ such that

$$c_\varepsilon q(|x|)^{-1-\varepsilon} \leqslant b_{ij}(x), \quad \forall |x| \geqslant C_\varepsilon, \tag{39}$$

for at least one fixed pair of indices $i,j \in \{1,\ldots,n\}$, and

$$\|\vec{a}(x)\| \leqslant C_\varepsilon q(|x|)^{-1+\varepsilon}. \tag{40}$$

It is essential for our result to make sense that we require the lower bound of Eq. (39) for $|x| > C_\epsilon$ only (instead of for all $x \in \mathbf{R}^n$). In a two-dimensional situation, any (non-trivial) exponentially decreasing vector potential gives rise to a field with total magnetic flux zero and we see that the field has to change sign somewhere.

IV.1 Theorem: *Let (36)–(39) hold, and suppose that* $\rho(E, H(0,V))$, *the integrated density of states for* $H(0,V)$, *exists. Then, as* $\lambda \to +\infty$,

$$\liminf_{\lambda \to +\infty} \phi(\lambda)^{-n} \widetilde{N}(E, H(\lambda \vec{a}, V)) \geqslant |v_n| \rho(E, H(0,V)). \tag{41}$$

Example: Let $n=2$ and $\vec{a}(x,y) = (yq(r)^{-1}, -xq(r)^{-1})$, where (for large r) $q(r) = \exp(r^s)$, $s > 0$, and $r = \sqrt{x^2+y^2}$.

Then (37)–(40) hold with the indicated q and $\phi(\lambda) = (\ln \lambda)^{1/s}$.

In the case of magnetic perturbations, upper bounds for $\widetilde{N}(E, H_\lambda)$ are not just much harder to obtain, but in many cases actually impossible. In fact, there are situations with $B = d\vec{a}$ of compact support in $\mathbf{R}^2$ and non-zero flux $F = \int B$ which produce oscillating eigenvalue branches in the gap by a mechanism related to the Bohm–Aharonov effect.[24–26]

The setup for the proof of Theorem IV.1 is similar to the one for Theorem II.1. For $l \geqslant 1$, we define a function ψ_l and an operator Π_l as in Section III, and we introduce operators $\widetilde{H}_{\lambda,l} = H_\lambda(B_l) + c\psi_l \Pi_l \psi_l$, where the constant $c > 0$ is the same as in Section III.

Our proof of Theorem IV.1 is based on the following variant of Theorem III.5 which can be proved by the same arguments as in Refs. 4, 15 since the magnetic potential decays at infinity.

IV.2. Theorem: *For* $0<\lambda<\lambda'$, *we have*

$$\widetilde{N}(E,H_{\lambda'})\geqslant \limsup_{l\to+\infty}|N(E,\widetilde{H}_{l,0})-N(E,\widetilde{H}_{l,\lambda})|. \tag{42}$$

Proof of Theorem IV.1:

To estimate the functions in the RHS of (42), we fix $\varepsilon>0$ and set $R=R(\lambda,\varepsilon)=\phi(\lambda^{1+\varepsilon})$. Outside a ball $B_{\lambda^{1+\varepsilon}}$, λa is small due to (40); therefore the same arguments as in Section III allow us to deduce an estimate:

$$|N(E,\widetilde{H}_{l,\lambda})-N(E,\widetilde{H}_{l,0})|\geqslant -C\phi(\lambda^{1+\varepsilon})^{n-1/2}+N(E-\nu,H(B_R))-N(E+\nu,H_\lambda(B_R)), \tag{43}$$

where $\nu\to 0+$ as $\lambda\to+\infty$. The asymptotics of $N(E-\nu,H(B_R))$ is evident:

$$N(E-\nu,H(B_R))\sim|v_n|\rho(E,H(0,V))R^n, \quad \text{as } \lambda\to+\infty. \tag{44}$$

Suppose we have proved that

$$\lim_{\varepsilon\to+0}\ \lim_{\lambda\to+\infty}\ \phi(\lambda^{1+\varepsilon})^{-n}N(E+\nu,H_\lambda(B_R))=0. \tag{45}$$

Then we gather (42)–(45), take (38) into account, and obtain (41).

To prove (45), we use a partition of unity and the IMS-localization formula as we did in Section III, with the supports $\overline{U_j}$ of the functions involved being of diameter less than $M_1>>1$. Without loss of generality we may also assume that ∂U_j has measure 0, for all j. Fix $M>1$ and set [using the pair of indices i,j from condition (39)]

$$I_+=\{j|\lambda b_{ij}(x)\geqslant M,\ \forall x\in\overline{U_j}\},\quad I_-=\{j\notin I_+|\overline{U_j}\cap B_R\neq\emptyset\}.$$

It follows from (39) that

$$\sum_{j\in I_-}\ \text{meas } U_j\leqslant C(\text{vol } B_R-\text{vol } B_{\phi(\lambda^{1-\varepsilon})}), \tag{46}$$

where the constant is independent of M_1 and $\varepsilon>0$.

Set $U_j=\text{int }\overline{U_j}$. We have

$$N(E,H_\lambda(B_R))\leqslant\sum_{j\in I_+}N(E+\nu,H_\lambda(U_j))+\sum_{j\in I_-}N(E+\nu,H_\lambda(U_j)), \tag{47}$$

where $\nu\to 0$ as $M_1\to+\infty$. If $j\in I_+$ and M is large enough, the first eigenvalue of $H_\lambda(U_j)$ is greater than $E+\nu$—see Ref. 18. Hence the first group of summands is equal to zero and may be dropped.

By Lemma IV.3, given below, there exists a constant C such that all the terms in the second sum admit a uniform bound,

$$N(2E,H_\lambda(U_j))\leqslant C\ \text{vol } U_j. \tag{48}$$

Then, by gathering (46)–(48) and using (38), we obtain (45); this concludes the proof of Theorem IV.1. ■

The following lemma uses a trace class estimate of Simon[27] to establish the estimate claimed in Eq. (48). Note that the assumptions on $\vec{a}$ and V are far more general than our standard conditions.

IV.3. Lemma: Let $\Omega \subset \mathbf{R}^n$ be a bounded, open set such that $\partial\Omega$ has measure zero. Let $\vec{a} \in C^1(\mathbf{R}^n, \mathbf{R}^n)$, $0 \leqslant V \in L_{1,\mathrm{loc}}(\mathbf{R}^n)$ and let $H^\Omega(\vec{a},V) = (-i\nabla - \vec{a}(x))^2 + V(x)$ and $H^\Omega = H^\Omega(0,0)$ denote Schrödinger operators in $L_2(\Omega)$ with Dirichlet boundary condition on $\partial\Omega$.

Then for any $\eta \in \mathbf{R}$ there exists a constant C which is independent of Ω such that

$$N(\eta, H^\Omega(a,V)) \leqslant C \text{ meas } \Omega.$$

Proof: For all $t>0$ the semigroups obey the trace estimate,[27]

$$\mathrm{Tr}\, \exp(-tH^\Omega(a,V)) \leqslant \mathrm{Tr}\, \exp(-tH^\Omega(0,0)).$$

We take $t=1$, and denote the eigenvalues of $H^\Omega(a,V)$ by $\mu_1(a,V) \leqslant \mu_2(a,V) \leqslant \dots$.

We then have

$$\begin{aligned} N(\eta, H^\Omega(a,V)) &= \sum_{\mu_s(a,V)<\eta} 1 \leqslant \sum_{s \geqslant 1} \exp(\eta - \mu_s(a,V)) \\ &= \exp(\eta)\, \mathrm{Tr}\, \exp(-H^\Omega(a,V)) \\ &\leqslant \exp(\eta)\, \mathrm{Tr}\, \exp(-H^\Omega(0,0)). \end{aligned}$$

It is easy to modify the estimate of Ref. 27, eq. (9.4), to yield

$$\mathrm{Tr}\, e^{-H^\Omega(0,0)} \leqslant C_n \text{ meas } \Omega,$$

and the desired result follows. ■

Remark: A careful inspection of the proof of Theorem IV.1 shows that it is enough to produce an estimate,

$$N(2E; H(U_j)) \leqslant C,$$

with a constant C depending only on a bound for the *diameter* of the sets U_j. The corresponding variant of Lemma IV.3 is then slightly simpler to prove since $\mathrm{Tr}\, e^{\Delta_Q}$ can be easily estimated for a cube Q.

Remark: Finally, we wish to comment on why our method seems to require the strong assumptions made in Theorem IV.1.

The main problem stems from the incompatibility of gauge transformations with the very construction of our approximating problems involving the operators $\widetilde{H}_{R,l}$ acting in the region where the field is small. If we wish to go beyond the assumptions made in Theorem IV.1, we can no longer assume that the vector potential $\lambda\vec{a}$ is small in the entire region $B_l \setminus B_{R'}$. It would seem natural to split this region into ''cones'' by suitable cut-off functions. Due to the non-local part $\psi_l \Pi_l \psi_l$ in the operator $\widetilde{H}_{R,l}$ even this splitting is problematic. Moreover, as λ increases, any attempt at changing to a different gauge on such a cone will lead to uncontrollable fluctuations in the $\psi_l \Pi_l \psi_l$-terms.

B. A separable example

For Schrödinger operators of the type $H_\lambda = -\Delta + V - \lambda W$, $\lambda > 0$, where V is a (non-constant) periodic potential, and W a potential which decays at infinity, separation of variables seems to be impossible. There is, however, a related separable model involving purely magnetic Schrödinger operators: this model consists of a Schrödinger operator H_0 in $\mathbf{R}^2$ with a constant magnetic field, perturbed by a decreasing, radially symmetric magnetic field. As is well known, the spectrum of

H_0 is given by a discrete set of points (the *Landau levels*), each Landau level being an eigenvalue of infinite multiplicity. We may consider this as a case of a band-gap structure, where the bands are degenerated into points.

Let us now fix the notation. To define the operators H_0 and H_λ, we use the radial gauge and let for $(x,y)\in\mathbf{R}^2$,

$$\vec{a}_0(x,y):=(y,-x),\quad \vec{a}_1(x,y):=(yf(r),-xf(r)),$$

and

$$\vec{A}_\lambda(x,y):=\vec{a}_0(x,y)+\lambda\vec{a}_1(x,y),$$

where $r=\sqrt{x^2+y^2}$, and $f\in C^1([0,\infty))$ satisfies

$$rf(r)\to 0,\quad r\to\infty. \tag{49}$$

We define the Schrödinger operators,

$$H_0:=(-i\nabla-\vec{a}_0(x,y))^2,\quad H_\lambda:=(-i\nabla-\vec{A}_\lambda(x,y))^2,$$

for $\lambda\in\mathbf{R}$; the operators H_λ are essentially self-adjoint on $\mathbf{C}_0^\infty(\mathbf{R}^2)$. Following to some extent the calculations of Miller and Simon,[23,28] we wish to find out whether eigenvalues of H_λ move through the gaps of H_0, as λ increases from 0 to ∞.

We begin with a few remarks.

(1) The magnetic field associated with the perturbation $\vec{a}_1$ is easily computed to be

$$B_1(r)=2f(r)+rf'(r),$$

and the total magnetic flux through the circle $B(0,R)$, associated with $\vec{a}_1$, is $2\pi R^2 f(R)$. In particular, the flux through the plane $\mathbf{R}^2$ will be zero if $r^2f(r)\to 0$, as $r\to\infty$.

(2) Obviously, (49) implies $\vec{a}_1(x,y)\to 0$, as $r\to\infty$; it is well-known that under this condition the resolvent difference $(H_\lambda+i)^{-1}-(H_0+i)^{-1}$ is compact and thus

$$\sigma_{\mathrm{ess}}(H_\lambda)=\sigma_{\mathrm{ess}}(H_0),\quad \lambda\in\mathbf{R}. \tag{50}$$

Since $\sigma_{\mathrm{ess}}(H_0)$ consists of isolated points, it is also clear that the continuous spectrum of H_λ is empty. [As proved in Ref. 29, (50) also holds if the magnetic field decays at infinity.]

We will now try to determine whether eigenvalues of H_λ move through the gaps of H_0, as $\lambda\to\infty$. As in Ref. 23, we start from the observation that the angular momentum operator L_z commutes with H_λ, for all $\lambda\in\mathbf{R}$. The spectrum of L_z is given by $\mathbf{Z}$, and the m-th eigenspace is spanned by the funtions $u(r)e^{im\vartheta}$, with $u\in L_2(0,\infty;rdr)$. Let $\mathscr{L}_m$ denote this eigenspace. We observe that the functions $e^{im\vartheta}$ coincide with the eigenfunctions of the Laplace–Beltrami-operator $-d^2/d\vartheta^2$ on the unit circle: in fact, the eigenvalues of the Laplace–Beltrami operator are given by the numbers m^2, with $m=0,1,2,\ldots$, and the m-th eigenspace is spanned by the functions $e^{\pm im\vartheta}$.

Passing to polar coordinates, we find that $H_{\lambda,m}$, defined as the part of H_λ in $\mathscr{L}_m$, acts as

$$H_{\lambda,m}(u(r)e^{im\vartheta})=\left(-u''(r)-\frac{1}{r}u'(r)+V_{\lambda,m}(r)u(r)\right)e^{im\vartheta},$$

where $V_{\lambda,m}(r):=m^2/r^2+r^2(1+\lambda f(r))^2+2m(1+\lambda f(r))$, for $u(r)e^{im\vartheta}$ in the domain of H_λ. Hence the eigenvalues of $H_{\lambda,m}$ are the same as the eigenvalues of the operator,

$$h_{\lambda,m}u := -u'' - \frac{1}{r}u' + V_{\lambda,m}(r)u,$$

in $L_2(0,\infty;rdr)$, for all $m \in \mathbf{Z}$. By the unitary equivalence $H_\lambda \cong \oplus_{m\in\mathbf{Z}} H_{\lambda,m}$, the spectrum of H_λ coincides with the closure of $\cup_{m\in\mathbf{Z}}\sigma(h_{\lambda,m})$.

As usual, it is advantageous to pass from the weighted space $L_2(0,\infty;rdr)$ to the standard L_2-space.[30] This leads to the operators

$$\widetilde{h}_{\lambda,m}u := -u'' + \widetilde{V}_{\lambda,m}u,$$

with

$$\widetilde{V}_{\lambda,m} = \left(m^2 - \frac{1}{4}\right)\frac{1}{r^2} + r^2 + 2r^2\lambda f(r) + \lambda^2 r^2 f^2(r) + 2m + 2m\lambda f(r).$$

The operators $h_{\lambda,m}$ and $\widetilde{h}_{\lambda,m}$ are unitarily equivalent. We recall that $\widetilde{h}_{\lambda,m}$ is essentially self-adjoint on $\mathbf{C}_0^\infty(0,\infty)$ for all $0 \neq m \in \mathbf{Z}$; in the special case $m=0$, we take the Friedrichs extension starting from the form domain $\mathbf{C}_0^\infty(0,\infty)$. It is evident that $\widetilde{h}_{\lambda,m}$ has compact resolvent for all $0 \neq m \in \mathbf{Z}$ and all $\lambda \in \mathbf{R}$. In the case $m=0$ one can convince oneself that $\widetilde{h}_{0,0}$ has compact resolvent by various arguments. E.g., if $\widetilde{h}_{0,0}$ had some essential spectrum, then the operator $-\Delta + |x|^2$ in $L_2(\mathbf{R}^2)$, the harmonic oscillator in 2 dimensions, would have to possess some essential spectrum too, as may be seen by arguing backwards in our above decomposition of a spherically symmetric operator.

IV.4. Lemma: For any $m \in \mathbf{Z}$ fixed, we have

$$\inf \sigma(\widetilde{h}_{\lambda,m}) \to \infty, \quad \lambda \to \infty, \tag{51}$$

provided $f(r) > 0$, for all $r \in (0,\infty)$.

Proof: We have to show that

$$\inf\{(\widetilde{h}_{\lambda,m}\phi,\phi); \phi \in \mathbf{C}_0^\infty(0,\infty), \|\phi\| = 1\} \to \infty, \quad \lambda \to \infty. \tag{52}$$

We may henceforth restrict our attention to real-valued functions.

We first consider the case $m \neq 0$. Since

$$\|u'\|^2 - \int_0^\infty \frac{1}{4r^2}u^2 \geqslant 0,$$

it is enough to look at the potential coming from the remaining terms. Here we first note that

$$\frac{m^2}{r^2} + 2m\lambda f(r) + \lambda^2 r^2 f^2(r) = \left(\frac{m}{r} + \lambda r f(r)\right)^2 \geqslant 0,$$

and we are left with

$$\widetilde{W}_{\lambda,m}(r) := r^2 + 2r^2\lambda f(r) + 2m.$$

Clearly, for any $\varepsilon > 0$, we have

$$\widetilde{W}_{\lambda,m}(r) \to \infty, \quad \text{uniformly on } [\varepsilon,\infty), \quad \text{as } \rangle\lambda \to \infty,$$

and (52) follows.

The case $m=0$ is more delicate. If we make the assumption that (52) is not true for $m=0$, then there exists a finite number c such that $\tilde{h}_{\lambda_k,0}$ has an eigenvalue $E_k \leqslant c$, for a sequence of $\lambda_k \to \infty$. As before, we find that $\tilde{W}_{\lambda_k,0}$ tends to infinity, uniformly on any fixed interval $[\varepsilon,\infty)$. Hence the eigenfunctions ψ_k of $\tilde{h}_{\lambda_k,0}$ which are associated with the eigenvalues E_k have the property that their part in (ε,∞) tends to 0, as $k\to\infty$. Hence the sequence (ψ_k) tends weakly to 0. At the same time,

$$(\tilde{h}_{0,0}\psi_k,\psi_k) \leqslant (\tilde{h}_{\lambda_k,0}\psi_k,\psi_k) \leqslant c,$$

implying that $\sigma_{\mathrm{ess}}(\tilde{h}_{0,0}) \neq \emptyset$; this, however, is incompatible with the fact that $\tilde{h}_{0,0}$ has compact resolvent. ■

In conclusion, we have proved the following properties of our example.

IV.5. Theorem: *Suppose* $f \in C^1([0,\infty))$ *satisfies* $f(r)>0$ *on* $(0,\infty)$, *and* $rf(r)\to 0$, *as* $r\to\infty$. *Let* H_λ *be defined as above. Then*

(i) $\sigma_{\mathrm{ess}}(H_\lambda)=\sigma_{\mathrm{ess}}(H_0)$ (=*the set of Landau levels for* H_0),

(ii) *Each (non-trivial) spectral gap of* H_0 *is crossed by an infinite number of eigenvalues of* H_λ, *as* λ *increases from 0 to* ∞.

Proof: In view of eq. (50) we only have to say something about the second statement.

Each Landau level Λ_j being an eigenvalue of infinite multiplicity, it follows that there exists an infinite number of $m \in \mathbf{Z}$ such that Λ_j is an eigenvalue of $\tilde{h}_{0,m}$. By eq. (51) however, all eigenvalues of $\tilde{h}_{\lambda,m}$ will eventually move out to infinity, as we let λ increase from 0 to ∞. ■

Remark: We do not expect that the discrete eigenvalues of H_λ are monotonically increasing with λ; eventually, however, all of them move up towards $+\infty$.

ACKNOWLEDGMENTS

The second author acknowledges financial support from the Deutsche Forschungsgemeinschaft and thanks the TU Braunschweig for the warm hospitality during his visit to Braunschweig.

[1] P. A. Deift and R. Hempel, ''On the existence of eigenvalues of the Schrödinger operator $H-\lambda W$ in a gap of $\sigma(H)$,'' Commun. Math. Phys. **103**, 461–490 (1986).

[2] R. Hempel, ''On the asymptotic distribution of eigenvalue branches of the Schrödinger operator $H\pm\lambda W$ in a spectral gap of H,'' J. Reine Angew. Math. **399**, 38–59 (1989).

[3] F. Gesztezy, D. Gurarie, H. Holden, M. Klaus, L. Sadun, B. Simon, and P. Vogl, ''Trapping and cascading of eigenvalues in the large coupling limit,'' Commun. Math. Phys. **118**, 597–634 (1988).

[4] S. Alama, P. A. Deift, and R. Hempel, ''Eigenvalue branches of the Schrödinger operator $H-\lambda W$ in a gap of $\sigma(H)$,'' Commun. Math. Phys. **121**, 291–321 (1989).

[5] M. Š. Birman, ''Discrete spectrum in the gaps of a continuous one for perturbations with large coupling constant,'' in *Estimates and Asymptotics for Discrete Spectra of Integral and Differential Equations*, Advances in Soviet Mathematics (American Mathematical Society, Providence, RI, 1991), Vol. 7, pp. 57–74.

[6] M. Š. Birman, ''On the discrete spectrum in the gaps of a perturbed periodic second order operator,'' Funct. Anal. Appl. **25**, 158–161 (1991).

[7] M. Š. Birman and G. D. Raikov, ''Discrete spectrum in the gaps for perturbations of the magnetic Schrödinger operator,'' Adv. Sov. Math. **7**, 75–84 (1991).

[8] G. D. Raikov, ''Strong-electric-field eigenvalue asymptotics for the perturbed magnetic Schrödinger operator,'' Commun. Math. Phys. **155**, 415–428 (1993).

[9] S. Z. Levendorskiĭ, ''Lower bounds for the number of eigenvalue branches for the Schrödinger operator $H-\lambda W$ in a gap of H: the case of indefinite W,'' Commun. Partial Differential Equations **20**, 827–854 (1995).

[10] S. Z. Levendorskiĭ, ''The asymptotics for the number of eigenvalue branches for the magnetic Schrödinger operator $H-\lambda W$ in a gap of $\sigma(H)$,'' Math. Z. **223**, 609–625 (1996).

[11] A. Mohamed and G. D. Raikov, ''On the spectral theory of the Schrödinger operator with electromagnetic potential,'' in *Advances in Partial Differential Equations*, Pseudo-Differential Operators and Mathematical Physics (Akademie Verlag, Berlin, 1994), pp. 298–390.

[12] M. Š. Birman, ''The discrete spectrum of a periodic operator, perturbed by a decreasing potential,'' Alg. Anal. **8**, 1–18 (1996).

[13] S. Alama, M. Avellaneda, P. A. Deift, and R. Hempel, ''On the existence of eigenvalues of a divergence form operator $A+\lambda B$ in a spectral gap of A,'' Asymptotic Anal. **8**, 311–344 (1994).
[14] S. I. Boyarchenko and S. Z. Levendorskiĭ, ''An asymptotic formula for the number of eigenvalue branches of a divergence form operator $A+\lambda B$ in a spectral gap of A,'' to appear in Commun. Partial Differential Equations.
[15] R. Hempel, ''Eigenvalues in gaps and decoupling by Neumann boundary conditions,'' J. Math. Anal. Appl. **169**, 229–259 (1992).
[16] R. Hempel, ''On the asymptotic distribution of eigenvalues in gaps,'' in *Quasiclassical Methods*, IMA Volumes in Mathematics and its Applications, edited by J. Rauch and B. Simon (Springer-Verlag, New York, 1997), Vol. 95, pp. 115–124.
[17] R. Hempel and J. Laitenberger, ''Schrödinger operators with local magnetic perturbations: Existence of eigenvalues in a gap of the essential spectrum,'' in *Operator Theory: Advances and Applications* (Birkhäuser, Basel, 1994), Vol. 70, pp. 13–18.
[18] J. Avron, I. Herbst, and B. Simon, ''Schrödinger operators with singular magnetic fields, I. General Interactions,'' Duke Math. J. **45**, 847–883 (1978).
[19] R. Hempel and I. Herbst, ''Strong magnetic fields, Dirichlet boundaries, and spectral gaps,'' Commun. Math. Phys. **169**, 237–259 (1995).
[20] J. Sjöstrand, ''Microlocal analysis for the periodic magnetic Schrödinger equation and related questions,'' in *Microlocal Analysis and Applications*, Lecture Notes in Mathematics (Springer-Verlag, Berlin 1991), Vol. 1495, pp. 237–332.
[21] G. V. Rozenbljum, M. Z. Solomyak, and M. A. Shubin, ''Spectral theory of differential operators,'' *Contemporary Problems of Mathematics (Itogi Nauki i Tekhniki VINITI)* (Viniti, Moscow, 1989), Vol. 64.
[22] S. Z. Levendorskiĭ, ''Asymptotic distribution of eigenvalues of differential operators,'' (Kluwer Academic, Dordrecht, 1990).
[23] H. L. Cycon, R. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators* (Springer, New York, 1987).
[24] B. Helffer, Effet d'Aharonov-Bohm pour un état borné, Commun. Math. Phys. **119**, 315–329 (1988).
[25] I. Herbst and Sh. Nakamura, ''Schrödinger operators with strong magnetic fields: Quasi-periodicity of spectral orbits and topology,'' in preparation.
[26] R. Hempel, in preparation.
[27] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[28] K. Miller and B. Simon, ''Quantum magnetic Hamiltonians with remarkable spectral properties,'' Phys. Rev. Lett. **44**, 1705–1706 (1980).
[29] A. Iwatsuka, ''The essential spectrum of two-dimensional Schrödinger operators with perturbed constant magnetic fields,'' J. Math. Kyoto Univ. **23**, 475–480 (1983).
[30] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. II: Fourier Analysis, Self-adjointness* (Academic, New York, 1975).

On infinite-dimensional algebras of symmetries of the self-dual Yang–Mills equations

T. A. Ivanova[a)]
Bogoliubov Laboratory of Theoretical Physics, JINR,
141980 Dubna, Moscow Region, Russia

(Received 17 March 1997; accepted for publication 10 July 1997)

Infinite-dimensional algebras of hidden symmetries of the self-dual Yang–Mills equations are considered. A current-type algebra of symmetries and an affine extension of conformal symmetries introduced recently are discussed using the twistor picture. It is shown that the extended conformal symmetries of the self-dual Yang–Mills equations have a simple description in terms of Ward's twistor construction. © *1998 American Institute of Physics.* [S0022-2488(97)01112-2]

I. INTRODUCTION

The self-dual Yang–Mills (SDYM) equations in the space R^4 are manifestly invariant under the group of gauge transformations and the group of conformal transformations of the space R^4. Besides these symmetries, the SDYM equations have infinite-dimensional algebras of ''hidden symmetries''[1–6] related to gauge transformations. In Refs. 1–5 the Yang gauge[7] in which two components of gauge potential are equal to zero was used. Crane generalized the results of Refs. 1–5 on the case when a gauge is not fixed and gave the twistor interpretation of the action of symmetries on solutions of the SDYM equations.[8]

In the frames of approach of Refs. 1–4 it has been shown[9] that on the space $\mathscr{M}$ of local solutions of the SDYM equations one can also define an (infinitesimal) action of the affine Lie algebra related to conformal transformations of R^4. Later, the action of the subalgebra of this algebra was also defined on the space of solutions of the vacuum self-dual gravity equations.[10] In this paper we shall show that the new symmetry algebra described in Ref. 9 has a simple interpretation in terms of vector fields acting on transition matrices of Ward's holomorphic vector bundles over twistor space.

II. DEFINITIONS AND NOTATION

We consider the Euclidean space $R^{4,0}$ with the metric $\delta_{\mu\nu}$ and potentials A_μ of the Yang–Mills (YM) fields $F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu+[A_\mu,A_\nu]$, where $\mu,\nu,\ldots=1,...,4$. For simplicity we consider the fields A_μ and $F_{\mu\nu}$ with values in the Lie algebra $su(n)$.

The SDYM equations have the form

$$\frac{1}{2}\,\epsilon_{\mu\nu\rho\sigma}F_{\rho\sigma}=F_{\mu\nu}, \tag{1}$$

where $\epsilon_{\mu\nu\rho\sigma}$ is the completely antisymmetric tensor on $R^{4,0}$ and $\epsilon_{1234}=1$.

It is easily seen that eqs. (1) are invariant under the algebra of infinitesimal local gauge transformations:

$$\delta_\varphi A_\mu=\partial_\mu\varphi+[A_\mu,\varphi], \tag{2}$$

a)Electronic mail: ita@thsun1.jinr.dubna.su

where $\varphi(x)\in su(n)$, $x\in R^{4,0}$. It is also well-known that the SDYM equations (1) are invariant with respect to the algebra $so(5,1)$ of infinitesimal conformal transformations:

$$\delta_N A_\mu = N^\nu \partial_\nu A_\mu + A_\nu \partial_\mu N^\nu, \quad \partial_\mu := \frac{\partial}{\partial x^\mu}, \tag{3}$$

where $N=N^\nu\partial_\nu$ is any generator of the 15-parameter conformal group:

$$X_a=\delta_{ab}\eta^b_{\mu\nu}x_\mu\partial_\nu,\ \ Y_a=\delta_{ab}\overline{\eta}^b_{\mu\nu}x_\mu\partial_\nu,\ \ P_\mu=\partial_\mu,$$

$$K_\mu=\frac{1}{2}\,x_\sigma x_\sigma\partial_\mu - x_\mu D, \quad D=x_\sigma\partial_\sigma, \quad a,b,\ldots=1,2,3, \tag{4}$$

where $\{X_a\}$ and $\{Y_a\}$ generate two commuting SO(3) subgroups in SO(4), P_μ are the translation generators, K_μ are the generators of special conformal transformations and D is the dilatation generator. Here $\eta^a_{\mu\nu}=\{\epsilon^a_{bc},\mu=b,\nu=c;-\delta^a_\mu,\nu=4;-\delta^a_\nu,\mu=4\}$ is the self-dual 't Hooft tensor and $\overline{\eta}^a_{\mu\nu}=\{\epsilon^a_{bc},\mu=b,\nu=c;-\delta^a_\mu,\nu=4;\delta^a_\nu,\mu=4\}$ is the anti-self-dual 't Hooft tensor.

Let $J=(J^\nu_\mu)$ be the most general complex structure on $R^{4,0}$:[11]

$$J^\nu_\mu = s_a\overline{\eta}^a_{\mu\sigma}\delta^{\sigma\nu} \Rightarrow J^\sigma_\mu J^\nu_\sigma = -\delta^\nu_\mu, \tag{5}$$

where real numbers s_a parametrize a two-sphere S^2, $s_a s_a=1$. Using J, one can introduce (0,1) vector fields

$$\overline{V_1}=\frac{1}{2}\,(\partial_1+i\partial_2)-\frac{\lambda}{2}\,(\partial_3+i\partial_4)=\partial_{\bar y}-\lambda\partial_z,\ \ \overline{V_2}=\frac{1}{2}\,(\partial_3-i\partial_4)+\frac{\lambda}{2}\,(\partial_1-i\partial_2)=\partial_{\bar z}+\lambda\partial_y, \tag{6}$$

where $y=x_1+ix_2$, $z=x_3-ix_4$, $\overline{y}=x_1-ix_2$, $\overline{z}=x_3+ix_4$ are complex coordinates on $R^4\simeq C^2$, and $\lambda=(s_1+is_2)/(1+s_3)$ is a local complex coordinate on $S^2\simeq CP^1$.

Let S^1 denote a contour $|\lambda|=1$ in the λ-plane enclosing the origin in the counter-clockwise direction, C_+ denote all complex numbers $\lambda\in CP^1$ with $|\lambda|\leqslant 1+\alpha$, where α is any positive real number, and C_- denote all complex numbers $\lambda\in CP^1$ with $|\lambda|\geqslant 1-\alpha$ (including $\lambda=\infty$). Then C_+ and C_- form a two-set open cover of the Riemann sphere CP^1 with the intersection $C_\alpha = C_+\cap C_- = \{\lambda: 1-\alpha\leqslant|\lambda|\leqslant 1+\alpha\}$. Vector field $\partial_{\bar\lambda}:=\partial/\partial\bar\lambda$ is antiholomorphic (0,1) vector field with respect to the standard complex structure $\epsilon = id\lambda\otimes\partial_\lambda - id\lambda\otimes\partial_{\bar\lambda}$ on S^2.

Twistor space $\mathcal{Z}$ for $R^{4,0}$ is the space $\mathcal{Z}=R^{4,0}\times CP^1\simeq CP^3-CP^1\subset CP^3$ which is the space of all the constant complex structures on $R^{4,0}$.[12] This space can be covered by two coordinate patches $\mathcal{Z}=U_+\cup U_-$:

$$U_+=\{x\in R^{4,0},\lambda\in C_+\},\quad U_-=\{x\in R^{4,0},\lambda\in C_-\}, \tag{7}$$

with the intersection $U=U_+\cap U_-=\{x\in R^{4,0},\lambda\in C_\alpha=C_+\cap C_-\}$.

Twistor space $\mathcal{Z}$ is a trivial bundle π: $\mathcal{Z}\to R^{4,0}$ over $R^{4,0}$ (π:$\{x_\mu,\lambda,\bar\lambda\}\to\{x_\mu\}$) with fibre CP^1 and is a complex manifold with complex structure $\mathcal{J}=(J,\epsilon)$ on $\mathcal{Z}$. Vector fields V_1^{-},V_2^{-} from (6) and $V_3^{-}=\partial_{\bar\lambda}$ are vector fields of type (0,1) with respect to complex structure $\mathcal{J}$. Holomorphic coordinates on U_+ and U_- are $\{v^1_+=y-\lambda\bar z, v^2_+=z+\lambda\bar y, v^3_+=\lambda\}$ and $\{v^1_-=1/\lambda\ y-\bar z, v^2_-=1/\lambda\ z+\bar y, v^3_-=1/\lambda\}$. Twistor space is also a nontrivial holomorphic bundle p: $\mathcal{Z}\to CP^1$ over CP^1 with holomorphic sections (projective lines CP^1_x) parametrized by points $x\in C^4$:

$$CP^1_x=\{v^1_+=y+\lambda\tilde z,\ \ v^2_+=z+\lambda\tilde y,\ \ (y,z,\tilde y,\tilde z)\in C^4\}. \tag{8}$$

On $\mathcal{Z}$ one can introduce a real structure as a map $\tau: R^{4,0}\times S^2 \to R^{4,0}\times S^2$ defined by the formula: $\tau(x_\mu,\lambda)=(x_\mu,-1/\bar{\lambda})$, $x_\mu \in R^{4,0}$. Then, for real holomorphic sections of the bundle $\mathcal{Z}\to CP^1$ satisfying the equations $\tau(v^1_+)=\bar{v}^2_-$, $\tau(v^2_+)=-\bar{v}^1_-$ we have $\tilde{z}=-\bar{z}$, $\tilde{y}=\bar{y}$.[12,11]

III. WARD'S TWISTOR CORRESPONDENCE AND GAUGE-TYPE HIDDEN SYMMETRIES

The potentials A_μ define a connection $D:=dx^\mu(\partial_\mu+A_\mu)$ in the principal $SU(n)$-bundle $P=P(R^{4,0},SU(n))$ over $R^{4,0}$. In the standard manner we introduce the complex vector bundle $E=P\times_{SU(n)}C^n$ associated to P. Sections of this bundle are C^n-valued vector-functions depending on $x\in R^{4,0}$. Using the projection $\pi: \mathcal{Z}\to R^{4,0}$, we can pull back the bundle E with self-dual connection D to the bundle $\tilde{E}:=\pi^*E$ over $\mathcal{Z}$, and the pulled back connection $\tilde{D}:=\pi^*D$ will be flat along the fibres CP^1_x of the bundle $\mathcal{Z}\to R^{4,0}$. We can take it in the form $\tilde{D}=D+d\lambda\partial_\lambda+d\bar{\lambda}\partial_{\bar{\lambda}}$.

Let us consider local holomorphic sections ξ of the bundle $\tilde{E}$ or, in other words, local solutions of the equations $\tilde{D}^{(0,1)}_{\bar{a}}\xi=0$:

$$[(D_1+iD_2)-\lambda(D_3+iD_4)]\xi(x,\lambda,\bar{\lambda})=0, \tag{9a}$$

$$[(D_3-iD_4)+\lambda(D_1-iD_2)]\xi(x,\lambda,\bar{\lambda})=0, \tag{9b}$$

$$\partial_{\bar{\lambda}}\xi(x,\lambda,\bar{\lambda})=0, \tag{10}$$

where $\tilde{D}^{(0,1)}_{\bar{a}}$ are components of $\tilde{D}$ along (0,1) vector fields $\overline{V_a}$. We can solve eq. (10), and then eqs. (9) on $\xi(x,\lambda)$ are usually called the linear system for the SDYM equations.[13,14] It is easily seen that the compatibility conditions of the linear system (9) coincide with the SDYM equations (1).

Equations (9) have a local solution $\xi_+(x,\lambda)$ over $U_+\subset\mathcal{Z}$, a local solution $\xi_-(x,\lambda)$ over $U_-\subset\mathcal{Z}$ and $\xi_+=\xi_-$ on $U_+\cap U_-$. We can always represent $\xi_\pm$ in the form $\xi_\pm=\psi_\pm\chi_\pm$, where $SL(n,C)$-valued functions $\psi_\pm$ are holomorphic in λ on $U_\pm$, and vector-functions $\chi_\pm\in C^n$ defined on $U_\pm$ are related by

$$\chi_-=\mathcal{T}\chi_+ \tag{11}$$

on $U=U_+\cap U_-\subset\mathcal{Z}$. Here $\chi_+=\chi_+(v^a_+)$ and $\chi_-=\chi_-(v^a_-)$ are Čech fibre coordinates of the bundle $\tilde{E}$ over U_+ and U_-, and $\mathcal{T}$ is the transition matrix in the bundle $\tilde{E}$. Matrix $\mathcal{T}$ is a holomorphic $SL(n,C)$-valued function on U with non-vanishing determinant. By Ward's construction, the bundle $\tilde{E}$ is holomorphically trivial on each fibre CP^1_x, i.e., $\tilde{E}|_{CP^1_x}=CP^1_x\times C^n$.[14] This means that on CP^1_x matrix $\mathcal{T}$ can be factorized in the form

$$\mathcal{T}=\psi_-^{-1}(x,\lambda)\psi_+(x,\lambda) \tag{12}$$

for each point x from an open subset of $R^{4,0}$.

Remark: If we are given a matrix-valued function $\mathcal{T}\in G$ on $C_\alpha=C_+\cap C_-\subset CP^1$, then the Riemann–Hilbert problem is to find matrix-valued functions $\psi_+\in G$ and $\psi_-\in G$ such that ψ_+ is regular (i.e., holomorphic with non-vanishing determinant) on C_+, ψ_- is regular on C_- and $\mathcal{T}=\psi_-^{-1}\psi_+$ is regular on $C_\alpha=C_+\cap C_-$. Solution of the Riemann–Hilbert problem exists (Birkhoff's theorem) and is unique up to the left multiplication by matrix g not depending on λ: $\psi_\pm\to g\psi_\pm$ (for discussion see, e.g., Ref. 15). Thus, the splitting (12) gives a solution of a parametric Riemann–Hilbert problem (x_μ are external parameters). Besides, the matrix $\mathcal{T}$ in (12) will not change if we replace $\psi_\pm(x,\lambda)$ by $g(x)\psi_\pm(x,\lambda)$ (standard gauge transformation).

Matrix-valued functions $\psi_{\pm}$ define a trivialization of the bundle $\widetilde{E}$ over $U_{\pm}$. It follows from (9)–(12) that

$$(\partial_{\bar{y}}\psi_{+}-\lambda\partial_{z}\psi_{+})\psi_{+}^{-1}=(\partial_{\bar{y}}\psi_{-}-\lambda\partial_{z}\psi_{-})\psi_{-}^{-1}=-(A_{\bar{y}}(x)-\lambda A_{z}(x)), \tag{13a}$$

$$(\partial_{\bar{z}}\psi_{+}+\lambda\partial_{y}\psi_{+})\psi_{+}^{-1}=(\partial_{\bar{z}}\psi_{-}+\lambda\partial_{y}\psi_{-})\psi_{-}^{-1}=-(A_{\bar{z}}(x)+\lambda A_{y}(x)), \tag{13b}$$

and the potentials $\{A_{\mu}\}$ defined by (13) satisfy the SDYM equations.

For a Hermitian vector bundle when the structure group is SU(n) the real structure τ on the twistor space $\mathcal{Z}$ induces a Hermitian structure in the bundle $\widetilde{E}$.[12] Then, fields $\{A_{\mu}\}$, matrices $\psi_{\pm}$ and $\mathcal{F}$ have to satisfy the following unitarity conditions (see e.g. Ref. 8):

$$A_{y}^{\dagger}=-A_{\bar{y}},\quad A_{z}^{\dagger}=-A_{\bar{z}},\quad A_{\bar{y}}^{\dagger}=-A_{y},\quad A_{\bar{z}}^{\dagger}=-A_{z}, \tag{14a}$$

$$\psi_{+}^{\dagger}(\lambda)=\psi_{-}^{-1}\left(-\frac{1}{\bar{\lambda}}\right),\quad \psi_{-}^{\dagger}(\lambda)=\psi_{+}^{-1}\left(-\frac{1}{\bar{\lambda}}\right), \tag{14b}$$

$$\mathcal{F}^{\dagger}(\lambda)=\mathcal{F}\left(-\frac{1}{\bar{\lambda}}\right), \tag{14c}$$

where † denotes Hermitian conjugation.

From (13) one can see that gauge potentials A_{μ} do not change after transformations

$$\psi_{+}\to\psi_{+}^{equ}=\psi_{+}h_{+},\quad \psi_{-}\to\psi_{-}^{equ}=\psi_{-}h_{-}, \tag{15}$$

where h_{+} is any regular holomorphic matrix-valued function on U_{+}, and h_{-} is any regular holomorphic matrix-valued function on U_{-}. This means that the bundles with transition matrices $h_{-}^{-1}\mathcal{F}h_{+}$ and $\mathcal{F}$ are holomorphically equivalent. Notice that for the Hermitian vector bundles $\widetilde{E}$ with transition matrices $\mathcal{F}$ satisfying the ''hermiticity'' condition (14c) the equality $h_{+}^{\dagger}(\lambda)=h_{-}^{-1}(-1/\bar{\lambda})$ has to be fulfilled.

Thus, we have described a one-to-one correspondence between gauge equivalence classes of solutions to the SDYM equations on the Euclidean 4-space and equivalence classes of holomorphic vector bundles $\widetilde{E}$ over twistor space $\mathcal{Z}$, that are holomorphically trivial over each real projective line CP_{x}^{1} in $\mathcal{Z}$ (the Euclidean version of Ward's theorem[16,12,11]).

Now we will briefly remind the description of gauge-type ''hidden symmetries.'' First, we should define an infinite-dimensional complex Lie (pseudo)group H (current-type group) of holomorphic maps from $U=U_{+}\cap U_{-}\subset\mathcal{Z}$ to the group SL(n,C) with standard pointwise multiplication and the corresponding complex Lie algebra $\mathcal{H}$ of holomorphic maps from U to the Lie algebra $sl(n,C)$. We shall also consider subgroups $H_{\pm}\subset H$ of elements from H which can be extended continuously to holomorphic maps from $U_{\pm}$ to SL(n,C) and the corresponding algebras $\mathcal{H}_{\pm}\subset\mathcal{H}$.

The action of group H on transition matrix $\mathcal{F}$ preserving the hermiticity condition (14c) is given by:[8]

$$\mathcal{F}\to\widetilde{\mathcal{F}}=u(h).\mathcal{F}:=h(\lambda).\mathcal{F}(\lambda)h^{\dagger}\left(-\frac{1}{\bar{\lambda}}\right), \tag{16}$$

where $h(\lambda)\equiv h(y-\lambda\bar{z},z+\lambda\bar{y},\lambda)$, $\mathcal{F}\equiv\mathcal{F}(y-\lambda\bar{z},z+\lambda\bar{y},\lambda)$, $h^{\dagger}(-1/\bar{\lambda})\equiv h^{\dagger}(y+\bar{z}/\bar{\lambda},z-\bar{y}/\bar{\lambda},-1/\bar{\lambda})$, and in (16) we simply omitted a part of arguments. To this action the following action of the Lie algebra $\mathcal{H}$ on $\mathcal{F}$ corresponds:

$$\mathscr{F}\to\delta_\varphi\mathscr{F}=\varphi(\lambda)\mathscr{F}(\lambda)+\mathscr{F}(\lambda)\varphi^\dagger\left(-\frac{1}{\bar\lambda}\right), \tag{17}$$

where $\varphi(\lambda)\equiv\varphi(y-\lambda\bar z,z+\lambda\bar y,\lambda)\in sl(n,C)$ is any element of $\mathscr{H}$ (of course, $\varphi^\dagger(-1/\bar\lambda)$ is also an element of $\mathscr{H}$).

Consider now the following $sl(n,C)$-valued function ϕ on U

$$\phi\equiv\varphi_- -\varphi_+ :=\psi_-(\delta_\varphi\mathscr{F})\psi_+^{-1}=\psi_-\varphi(\lambda)\psi_-^{-1}+\psi_+\varphi^\dagger\left(-\frac{1}{\bar\lambda}\right)\psi_+^{-1}, \tag{18}$$

where for holomorphic ϕ expanded in Laurent series $\phi=\Sigma_{n=-\infty}^{\infty}\lambda^n\phi_n(x)$ we put

$$\varphi_+ :=\tilde\phi_0(x)-\sum_{n=1}^{\infty}\lambda^n\phi_n(x),\quad \varphi_- :=\hat\phi_0(x)+\sum_{n=-\infty}^{-1}\lambda^n\phi_n(x),\quad \hat\phi_0(x)-\tilde\phi_0(x)=\phi_0(x). \tag{19}$$

Thus, the function $\varphi_+(\lambda)\in sl(n,C)$ is holomorphic in $\lambda\in C_+\subset CP^1$, and the function $\varphi_-(\lambda)\in sl(n,C)$ is holomorphic in $\lambda\in C_-\subset CP^1$.

It follows from (13) and (18) that

$$(D_{\bar y}-\lambda D_z)\phi\equiv(\partial_{\bar y}-\lambda\partial_z)\phi+[A_{\bar y}-\lambda A_z,\phi]=0\Rightarrow(D_{\bar y}-\lambda D_z)\varphi_+=(D_{\bar y}-\lambda D_z)\varphi_-, \tag{20a}$$

$$(D_{\bar z}+\lambda D_y)\phi\equiv(\partial_{\bar z}+\lambda\partial_y)\phi+[A_{\bar z}+\lambda A_y,\phi]=0\Rightarrow(D_{\bar z}+\lambda D_y)\varphi_+=(D_{\bar z}+\lambda D_y)\varphi_-. \tag{20b}$$

The action of algebra $\mathscr{H}$ on matrix-valued functions $\psi_\pm\in\mathrm{SL}(n,C)$ and on gauge potentials $\{A_\mu\}$ is given by formulas

$$\delta_\varphi\psi_+ :=-\varphi_+\psi_+,\quad \delta_\varphi\psi_- :=-\varphi_-\psi_-, \tag{21}$$

$$\delta_\varphi A_{\bar y}-\lambda\delta_\varphi A_z :=D_{\bar y}\varphi_+-\lambda D_z\varphi_+=D_{\bar y}\varphi_--\lambda D_z\varphi_-, \tag{22a}$$

$$\delta_\varphi A_{\bar z}+\lambda\delta_\varphi A_y :=D_{\bar z}\varphi_++\lambda D_y\varphi_+=D_{\bar z}\varphi_-+\lambda D_y\varphi_-. \tag{22b}$$

It follows from (22) that

$$\delta_\varphi A_y=\oint_{S^1}\frac{d\lambda}{2\pi i\lambda^2}(D_{\bar z}\varphi_++\lambda D_y\varphi_+),\quad \delta_\varphi A_z=\oint_{S^1}\frac{d\lambda}{2\pi i\lambda^2}(D_{\bar y}\varphi_+-\lambda D_z\varphi_+),$$

$$\delta_\varphi A_{\bar y}=\oint_{S^1}\frac{d\lambda}{2\pi i\lambda}(D_{\bar y}\varphi_+-\lambda D_z\varphi_+),\quad \delta_\varphi A_{\bar z}=\oint_{S^1}\frac{d\lambda}{2\pi i\lambda}(D_{\bar z}\varphi_++\lambda D_y\varphi_+), \tag{23}$$

where $S^1=\{\lambda\in CP^1:|\lambda|=1\}$. Thus, in (17), (21) and (22) we have described the action of the current algebra $\mathscr{H}$ on the space of local solutions of the SDYM equations.

IV. HIDDEN SYMMETRIES FROM THE AFFINIZATION OF CONFORMAL ALGEBRA

In Ref. 9 to each generator N of the group SO(5,1) of conformal transformations of the Euclidean space $R^{4,0}$ Popov and Preitschopf associated an infinite number of infinitesimal symmetry transformations δ_N^n $(n=0,1,2,...)$ acting on solutions of the SDYM equations. They considered potentials A_μ with values in complex Lie algebras and did not discuss the problems of reality

for gauge fields. Considering potentials A_μ with values in su(n), we examine not only the infinitesimal symmetries from Ref. 9, which do not preserve, generally speaking, the unitarity conditions (14), but also the symmetries preserving the unitarity.

In formulas (4) we have specified a homomorphism of the Lie algebra so(5,1) of the conformal group into the Lie algebra of vector fields M, N... on $R^{4,0}$. Now we have to define the action of SO(5,1) on $\mathscr{Z}$, which preserves the holomorphicity of the bundle $\widetilde{E}\to\mathscr{Z}$. Such lift $N\to\widetilde{N}$ of vector fields on $\mathscr{Z}$ was described in Ref. 17, and the lifted generators

$$\widetilde{X}_a=X_a,\quad \widetilde{Y}_a=Y_a+2Z_a,\quad \widetilde{P}_\mu=P_\mu,\quad \widetilde{K}_\mu=K_\mu+\overline{\eta}^{\sigma}_{\sigma\mu}x_\sigma Z_a,\quad \widetilde{D}=D, \tag{24}$$

$$Z_1=\frac{1}{2}\,i(\lambda^2-1)\partial_\lambda-\frac{1}{2}\,i(\overline{\lambda^2}-1)\partial_{\overline{\lambda}},\quad Z_2=\frac{1}{2}\,(\lambda^2+1)\partial_\lambda+\frac{1}{2}\,(\overline{\lambda^2}+1)\partial_{\overline{\lambda}},\quad Z_3=i\lambda\partial_\lambda-i\overline{\lambda}\partial_{\overline{\lambda}} \tag{25}$$

are infinitesimal automorphisms of the complex structure $\mathscr{J}$ on $\mathscr{Z}$. This means, in particular, that for any generator $\widetilde{N}$ from (24) the following relations

$$[\overline{V_a},\widetilde{N}]=\overline{\alpha^b_a}(\widetilde{N})\overline{V_b} \tag{26}$$

take place for some functions $\overline{\alpha^b_a}(\widetilde{N})$ and vector fields $\overline{V_1},\overline{V_2}$ from (6) and $\overline{V_3}=\partial_{\overline{\lambda}}$.

Having generators $\widetilde{N}$ of infinitesimal holomorphic diffeomorphisms (24) of the space $\mathscr{Z}$, we consider the holomorphic vector fields $N_n=\lambda^{-n}\widetilde{N}$, $n=0,\pm1,\pm2,...$, and define the following infinitesimal transformations of transition matrix $\mathscr{F}$ of the bundle $\widetilde{E}$

$$\delta^n_N\mathscr{F}:=N_n(\mathscr{F})=\lambda^{-n}\widetilde{N}(\mathscr{F}),\quad n=0,\pm1,\pm2,\dots. \tag{27}$$

In virtue of (26) transformations (27) are holomorphic and, therefore, define the infinitesimal symmetries of the SDYM equations.

Let us introduce the $sl(n,C)$-valued function (compare with (18) from Sec. III)

$$\phi_{N_n}=\phi^-_{N_n}-\phi^+_{N_n}:=\psi_-(\delta^n_N\mathscr{F})\psi_+^{-1}=\lambda^{-n}(\widetilde{N}\psi_+)\psi_+^{-1}-\lambda^{-n}(\widetilde{N}\psi_-)\psi_-^{-1}, \tag{28}$$

$$\phi^+_{N_n}:=\widetilde{\phi}^0_{N_n}(x)-\sum_{k=1}^{\infty}\lambda^k\phi^k_{N_n}(x),\quad \phi^-_{N_n}:=\hat{\phi}^0_{N_n}(x)+\sum_{k=-\infty}^{-1}\lambda^k\phi^k_{N_n}(x),\quad \hat{\phi}^0_{N_n}-\widetilde{\phi}^0_{N_n}=\phi^0_{N_n}. \tag{29}$$

Be reminded that the splitting of the sl(n,C)-valued function $\phi_{N_n}(x,\lambda)$ into the difference of the function $\phi^-_{N_n}\in\mathrm{sl}(n,C)$ holomorphic in $\lambda\in C_-$ and of the function $\phi^+_{N_n}\in\mathrm{sl}(n,C)$ holomorphic in $\lambda\in C_+$ is a solution of the infinitesimal variant of the Riemann–Hilbert problem.

Remark: Notice that $\phi^\pm_{N_0}=-(\widetilde{N}\psi_\pm)\psi_\pm^{-1}$, but $\phi^+_{N_n}\neq-\lambda^{-n}(\widetilde{N}\psi_+)\psi_+^{-1}$, and $\phi^-_{N_n}\neq-\lambda^{-n}(\widetilde{N}\psi_-)\psi_-^{-1}$, when $n\neq0$. If we fix the gauge $\psi_+(\lambda=0)=1$, then the functions $\phi^+_{N_0}$ coincide with the functions $-\psi_{\widetilde{N}}$, $\widetilde{N}\in so(5,1)$, which are generating functions for symmetries introduced in Ref. 9.

It follows from (13), (26), and (28) that (cf. Sec. III):

$$(D_{\overline{y}}-\lambda D_z)\phi_{N_n}=0\Rightarrow(D_{\overline{y}}-\lambda D_z)\phi^+_{N_n}=(D_{\overline{y}}-\lambda D_z)\phi^-_{N_n}, \tag{30a}$$

$$(D_{\overline{z}}+\lambda D_y)\phi_{N_n}=0\Rightarrow(D_{\overline{z}}+\lambda D_y)\phi^+_{N_n}=(D_{\overline{z}}+\lambda D_y)\phi^-_{N_n}, \tag{30b}$$

$$\delta_N^n \psi_+ = -\phi_{N_n}^+ \psi_+ , \quad \delta_N^n \psi_- = -\phi_{N_n}^- \psi_- , \tag{31}$$

$$\delta_N^n A_{\bar{y}} - \lambda \delta_N^n A_z := D_{\bar{y}} \phi_{N_n}^+ - \lambda D_z \phi_{N_n}^+ = D_{\bar{y}} \phi_{N_n}^- - \lambda D_z \phi_{N_n}^- ,$$

$$\delta_N^n A_{\bar{z}} + \lambda \delta_N^n A_y := D_{\bar{z}} \phi_{N_n}^+ + \lambda D_y \phi_{N_n}^+ = D_{\bar{z}} \phi_{N_n}^- + \lambda D_y \phi_{N_n}^- , \tag{32}$$

We have

$$\delta_N^n A_y = \oint_{S^1} \frac{d\lambda}{2\pi i \lambda^2} (D_{\bar{z}} \phi_{N_n}^+ + \lambda D_y \phi_{N_n}^+), \quad \delta_N^n A_z = -\oint_{S^1} \frac{d\lambda}{2\pi i \lambda^2} (D_{\bar{y}} \phi_{N_n}^+ - \lambda D_z \phi_{N_n}^+),$$

$$\delta_N^n A_{\bar{y}} = \oint_{S^1} \frac{d\lambda}{2\pi i \lambda} (D_{\bar{y}} \phi_{N_n}^+ - \lambda D_z \phi_{N_n}^+), \quad \delta_N^n A_{\bar{z}} = \oint_{S^1} \frac{d\lambda}{2\pi i \lambda} (D_{\bar{z}} \phi_{N_n}^+ + \lambda D_y \phi_{N_n}^+). \tag{33}$$

Thus, to each generator N of the group of conformal transformations of the space $R^{4,0}$ we have corresponded an infinite number of generators δ_N^n of symmetry transformations of the SDYM equations and defined their action on the transition matrix $\mathscr{T}$, on the group-valued functions $\psi_\pm$ and on gauge potentials A_μ.

By direct calculations, it is not hard to show that

$$\delta_N^0 A_\mu = \mathscr{L}_N A_\mu \equiv N^\nu \partial_\nu A_\mu + A_\nu \partial_\mu N^\nu, \tag{34}$$

i.e. $\delta_N^0 A_\mu$ coincide with infinitesimal conformal transformation (3). Symmetries (33) with $n \geqslant 0$ are in one-to-one correspondence with the symmetries introduced in Ref. 9. To show this correspondence in more detail, let us discuss the algebraic properties of the symmetry transformations δ_N^n.

The commutator of two transformations can be easily calculated using formula (27). We have

$$[\delta_M^m, \delta_N^n] \mathscr{T} = \lambda^{-m-n} [\widetilde{M}, \widetilde{N}](\mathscr{T}) + m\lambda^{-m-n-1} \widetilde{N}^\lambda \widetilde{M}(\mathscr{T}) - n\lambda^{-m-n-1} \widetilde{M}^\lambda \widetilde{N}(\mathscr{T}), \tag{35}$$

where $\widetilde{N}^\lambda$ and $\widetilde{M}^\lambda$ are components of vector fields $\widetilde{N}$ and $\widetilde{M}$ along ∂_λ. Substituting (24) into (35), we obtain

$$[\delta_M^m, \delta_N^n] = \delta_{[M,N]}^{m+n}, \quad m,n, = 0, \pm 1, \pm 2, ... \tag{36}$$

for $M, N \in \mathscr{A} = \{P_\mu, X_a, D\}$, i.e., δ_M^m with generators M from the 8-dimensional Lie algebra $\mathscr{A}$ (see Ref. 9) form the affine Lie algebra $\mathscr{A} \otimes C[\lambda, \lambda^{-1}]$. For generators $\delta_{Y_a}^m$ we have

$$[\delta_{Y_1}^m, \delta_{Y_1}^n] = i(m-n)(\delta_{Y_1}^{m+n-1} - \delta_{Y_1}^{m+n+1}), \tag{37a}$$

$$[\delta_{Y_2}^m, \delta_{Y_2}^n] = (m-n)(\delta_{Y_2}^{m+n-1} + \delta_{Y_2}^{m+n+1}), \tag{37b}$$

$$[\delta_{Y_3}^m, \delta^n{}_{Y_3}] = 2i(m-n)\delta_{Y_3}^{m+n}, \tag{37c}$$

$$[\delta_{Y_1}^m, \delta_{Y_2}^n] = \delta_{[Y_1,Y_2]}^{m+n} + m(\delta_{Y_1}^{m+n-1} + \delta_{Y_1}^{m+n+1}) - in(\delta_{Y_2}^{m+n-1} - \delta_{Y_2}^{m+n+1}), \tag{38a}$$

$$[\delta_{Y_1}^m, \delta_{Y_3}^n] = \delta_{[Y_1,Y_3]}^{m+n} + 2im\delta_{Y_1}^{m+n} - in(\delta_{Y_3}^{m+n-1} - \delta_{Y_3}^{m+n+1}), \tag{38b}$$

$$[\delta_{Y_2}^m, \delta_{Y_3}^n] = \delta_{[Y_2,Y_3]}^{m+n} + 2im\delta_{Y_2}^{m+n} - n(\delta_{Y_3}^{m+n-1} + \delta_{Y_3}^{m+n+1}), \tag{38c}$$

$$[\delta^m_{Y_1},\delta^n_N]=\delta^{m+n}_{[Y_1,N]}-in(\delta^{m+n-1}_N-\delta^{m+n+1}_N), \tag{39a}$$

$$[\delta^m_{Y_2},\delta^n_N]=\delta^{m+n}_{[Y_2,N]}-n(\delta^{m+n-1}_N+\delta^{m+n+1}_N), \tag{39b}$$

$$[\delta^m_{Y_3},\delta^n_N]=\delta^{m+n}_{[Y_3,N]}-2in\,\delta^{m+n}_N, \tag{39c}$$

where $N\in\mathscr{A}=\{P_\mu,X_a,D\}$. Notice that the imaginary unit i can be removed from structure constants in (37)–(39) by the following change of generators $\delta^m_{Y_1}\to i\delta^m_{Y_1}$, $\delta^m_{Y_3}\to i\delta^m_{Y_3}$ (see Ref. 9). Then ''half'' of the algebra (36)–(39) with $m,n\geqslant 0$ will coincide with the one considered in Ref. 9. Let us emphasize that in (36)–(39) the numbers $m,n,=0,\pm 1,...$ are any integer numbers. It is important to understand that the change $\delta^m_{Y_1}\to i\delta^m_{Y_1}$, $\delta^m_{Y_3}\to i\delta^m_{Y_3}$ is possible only by considering complex gauge fields or by transition from the signature (4,0) to the signature (2,2).

Finally, for generators K_μ of special conformal transformations, the commutators of transformations $\delta^m_{K_\mu}$ with other transformations will be again the symmetry transformations, but with coefficients depending on the coordinates on $\mathscr{Z}$. For example,

$$[\delta^m_{P_1},\delta^n_{K_1}]=\delta^{m+n}_{[P_1,K_1]}+\frac{m}{2}\,v^2_+\delta^{m+n+1}_{P_1}-\frac{m}{2}\,v^1_+\delta^{m+n}_{P_1}, \tag{40}$$

where $v^1_+=y-\lambda\overline{z}$, $v^2_+=z+\lambda\overline{y}$ are holomorphic coordinates on $U_+\subset\mathscr{Z}$. This means that the structure constants in all the commutators with $\delta^m_{K_\mu}$ are replaced by the ''structure functions,'' that has been noticed by Popov and Preitschopf.[9] We do not write out these commutators looking similar to (40). In twistor terms, commutators like (40) appear because of the fact that the expansion in Laurent series in λ and formulas like (27), (29) etc. may be used only by the existence of the holomorphic projection $p:\mathscr{Z}\to CP^1$ leading to the distinguishing of coordinate λ. Special conformal transformations K_μ and transformations $\delta^m_{K_\mu}$ related to K_μ do not preserve the bundle $\mathscr{Z}\to CP^1$ and make invalid the formulas based on the expansion in λ. On formal level, this is connected with the fact that special conformal transformations transform λ into a function on x_μ and λ, and this leads to appearance of ''structure functions'' in formulas like (40). For correct discussion of symmetries related to K_μ, it is necessary to pass from the flat hyperKähler space $R^{4,0}$ to the four-sphere S^4 which is conformally flat (but not hyperKähler) manifold.

Be reminded that for real gauge fields with values in the Lie algebra $su(n)$ the transition functions $\mathscr{F}$ in the bundle $\widetilde{E}$ have to satisfy the condition (14c). Generally speaking, infinitesimal transformations do not (infinitesimally) preserve this condition, since

$$(\delta^n_N\mathscr{F}|_{\lambda\to -1/\overline{\lambda}})^\dagger=(-1)^n\lambda^n\widetilde{N}.\mathscr{F}(\lambda)=(-1)^n\delta^{-n}_N.\mathscr{F}(\lambda). \tag{41}$$

However, we can introduce infinitesimal symmetry transformations $\delta^n_{N_R}$ preserving reality conditions by setting

$$\delta^n_{N_R}.\mathscr{F}:=\delta^n_N.\mathscr{F}+(-1)^n\delta^{-n}_N.\mathscr{F}=(\lambda^{-n}+(-\lambda)^n)\widetilde{N}.\mathscr{F}\equiv N^R_n.\mathscr{F}. \tag{42}$$

For transformations (42), the formulas for $\phi_{N^R_n}$, $\delta^n_{N_R}\psi_\pm$ and $\delta^n_{N_R}A_\mu$ will have the same form as formulas (28)–(33) with the replacement N_n by N^R_n, and that is why we shall not write them out. As to the algebraic properties of transformations (42), then, by direct calculation, it can be shown that generators $\delta^n_{P^R_\mu}$, $\delta^n_{X^R_a}$ and $\delta^n_{D^R}$ form a closed algebra. Commutation relations analogous to (36)–(39), which we do not write out, easily follow from the definition (42).

V. CONCLUSION

Symmetries of the SDYM and the self-dual gravity equations are important in quantization of Yang–Mills model, $N=2$ strings and related models, that have been discussed, for example, in recent papers.[18,19] We think that the twistor point of view on symmetries of the SDYM equations used in this paper clarifies the geometric and algebraic sense of the so called ''hidden symmetries'' and makes them not so mystical. It would be interesting to consider the affine Lorentz symmetries[10] of the self-dual gravity equations from the twistor point of view. Notice that all the results of our paper may be generalized on the SDYM equations in the $4k$-dimensional spaces considered e.g. in Ref. 20.

ACKNOWLEDGMENTS

The author thanks the Institut für Physik, Humboldt-Universität zu Berlin for their kind hospitality, where this work was initiated, and the Fakultät für Physik, Universität Freiburg, where part of this work was done. This work was supported by the DAAD and the Heisenberg–Landau Program.

[1] M. K. Prasad, A. Sinha, and L.-L. Chau Wang, Phys. Lett. B **87**, 237 (1979); Phys. Rev. Lett. **43**, 750 (1979).
[2] K. Pohlmeyer, Commun. Math. Phys. **72**, 37 (1980).
[3] L.-L. Chau, M.-L. Ge, and Y.-S. Wu, Phys. Rev. D **25**, 1086 (1982); L.-L. Chau and Y.-S. Wu, *ibid.* **26**, 3541 (1982); L.-L. Chau, M.-L. Ge, A. Sinha, and Y.-S. Wu, Phys. Lett. B **121**, 391 (1983); L.-L. Chau, Lecture Notes Phys. **189**, 111 (1983).
[4] L. Dolan, Phys. Lett. B **113**, 387 (1982); Phys. Rep. **109**, 3 (1984).
[5] K. Ueno and Y. Nakamura, Phys. Lett. B **109**, 273 (1982); K. Takasaki, Commun. Math. Phys. **94**, 35 (1984).
[6] J. Avan, H. J. de Vega, and J. M. Maillet, Phys. Lett. B **171**, 255 (1986); J. Avan and H. J. de Vega, Int. J. Mod. Phys. A **3**, 1273 (1988).
[7] C. N. Yang, Phys. Rev. Lett. **38**, 1377 (1977).
[8] L. Crane, Commun. Math. Phys. **110**, 391 (1987).
[9] A. D. Popov and C. R. Preitschopf, Phys. Lett. B **374**, 71 (1996).
[10] A. D. Popov, M. Bordemann, and H. Römer, Phys. Lett. B **385**, 63 (1996).
[11] N. M. J. Woodhouse, Class. Quantum Grav. **2**, 257 (1985).
[12] M. F. Atiyah, N. J. Hitchin, and I. M. Singer, Proc. R. Soc. London, Ser. A **362**, 425 (1978).
[13] A. A. Belavin and V. E. Zakharov, Phys. Lett. B **73**, 53 (1978).
[14] R. S. Ward, Phys. Lett. A **61**, 81 (1977).
[15] L. J. Mason and N. M. J. Woodhouse, *Integrability, Self-Duality and Twistor Theory* (Clarendon, Oxford, 1996).
[16] M. F. Atiyah and R. S. Ward, Commun. Math. Phys. **55**, 117 (1977).
[17] M. Legaré and A. D. Popov, Phys. Lett. A **198**, 195 (1995); JETP Lett. **59**, 883 (1994).
[18] V. P. Nair and J. Schiff, Nucl. Phys. B **371**, 329 (1992); Phys. Lett. B **246**, 423 (1990); A. Losev, G. Moore, N. Nekrasov, and S. Shatashvili, Nucl. Phys. B **46** (Proc. Suppl.) (1996); S. V. Ketov, hep-th/9606142; E. J. Martinec, hep-th/9608017; D. Kutasov and E. Martinec, hep-th/9612102.
[19] T. Inami, H. Kanno, T. Ueno, and C.-S. Xiong, hep-th/9610187; D. Cangemi, hep-th/9610021; A. A. Rosly and K. G. Selivanov, hep-th/9611101; V. E. Korepin and T. Oota, hep-th/9608064.
[20] R. S. Ward, Nucl. Phys. B **236**, 381 (1984); E. Corrigan, P. Goddard, and A. Kent, Commun. Math. Phys. **100**, 1 (1985); A. Galperin, E. Ivanov, V. Ogievetsky, and E. Sokatchev, Ann. Phys. **185**, 1 (1988); A. D. Popov, Mod. Phys. Lett. A **7**, 2077 (1992); T. A. Ivanova and A. D. Popov, Theor. Math. Phys. **94**, 225 (1993).

Quantum sources and a quantum coding theorem

Christopher King[a)]
Department of Mathematics, Northeastern University, Boston, Massachusetts 02115

Andrezj Leśniewski[b)]
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138

(Received 28 March 1996; accepted for publication 16 September 1997)

We define a large class of quantum sources and prove a quantum analog of the asymptotic equipartition property. Our proof relies on using local measurements on the quantum source to obtain an associated classical source. The classical source provides an upper bound for the dimension of the relevant subspace of the quantum source, via the Shannon–McMillan noiseless coding theorem. Along the way we derive a bound for the von Neumann entropy of the quantum source in terms of the Shannon entropy of the classical source, and we provide a definition of ergodicity of the quantum source. Several explicit models of quantum sources are also presented. © *1998 American Institute of Physics.* [S0022-2488(98)00401-0]

I. INTRODUCTION

A. Context

The possibility of building a quantum computer has stimulated new interest in the quantum analog of classical information theory (see Ref. 1 for an introduction and review of current ideas). Shannon, McMillan, Khinchin, and others provided a firm foundation for the classical theory, and used the mathematics of stochastic processes to prove important theorems. In particular they obtained limits on the amount of information that can be transmitted through a channel. Although a quantum computer does not yet exist, it is reasonable to suppose that similar issues of channel capacity are relevant to its operation. So it is interesting to investigate this question, using our current understanding of quantum mechanics.

There has been much work done on this and related questions. In particular, Schumacher stated and proved a capacity result for a quantum channel in Refs. 2 and 3. Our particular interest is in the *extended* quantum signal source described by Schumacher. This is a quantum system whose state space is a (tensor) product of many copies of one fundamental state space M. The source produces a signal that is encoded by a state in M; the ensemble of possible signals is represented by a density operator ρ on M. The extended source corresponds to a sequence of such states, and has a natural interpretation as a message. The probabilistic character of the message is contained in the density operator on the tensor product of copies of M. One choice of density operator is $\rho\otimes\cdots\otimes\rho$. This corresponds to independent signals at all times, and there are no correlations between signals in the message. For this reason we call this a quantum Bernoulli source (a precise definition is provided in Sec. II). Schumacher proves his quantum noiseless coding theorem for such a source. He introduces the notion of fidelity of a quantum channel, and states his results in terms of this. On the most basic level his results show that the state space for the extended source has a relevant subspace whose dimension is determined by the von Neumann entropy of ρ. As far as the information content of the source is concerned, the rest of the state space can be ignored.

[a)]Electronic mail: king@neu.edu
[b)]Electronic mail: lesniewski@huhepl.harvard.edu

0022-2488/98/39(1)/88/14/$15.00
© 1998 American Institute of Physics

B. Results

Our main interest in this paper is to extend Schumacher's result to a large class of quantum sources. We define precisely what we mean by this in Sec. II. Roughly speaking, we consider sources that allow correlations on all time scales between signals in a message. In classical information theory the corresponding result is the Shannon–McMillan theorem, which shows that under very general assumptions the ensemble of all possible messages can be split into a relevant and an irrelevant part. The criterion for splitting is provided by the entropy of the source. The corresponding result for the quantum theory should be a splitting of the state space into a relevant subspace and an irrelevant subspace, with the von Neumann entropy as the criterion. This is precisely what Schumacher proved for the quantum Bernoulli source. We obtain such a splitting in the general case, and derive an estimate for the dimension of the relevant subspace, by computing the entropy of a classical source that is obtained by making measurements on the quantum system.

To summarize, we show that a general quantum source can be encoded by a quantum system whose dimension is smaller than the original. We give estimates for the dimension of the reduced space, based on the results of local measurements of the source. For the case of a quantum source emitting orthogonal states our estimate is tight, and for a Bernoulli source reproduces Schumacher's result. We also derive an inequality relating this experimental entropy to the true von Neumann entropy.

C. Layout

The paper is organized as follows. In sec. II we define quantum sources and recall some standard results about the construction of infinite quantum systems. We also provide some explicit examples that serve to illustrate our ideas. In Sec. III we recall the classical notion of an ergodic source, and propose a definition of ergodicity for a quantum source. As a check we prove that a quantum Bernoulli source is ergodic. In Sec. IV we recall the notion of quantum entropy. In Sec. V we pursue the quantum analog of the Shannon–McMillan theorem. To do this we use the notion of a positive operator-valued measure, and construct an associated classical source. The entropy of this associated source satisfies a lower bound involving the von Neumann entropy of the quantum source. It also determines the dimension of a relevant subspace of the quantum system, which in turn yields our quantum noiseless coding theorem.

II. QUANTUM SOURCES

A. Setup

According to the mathematical theory of information, see, e.g., Refs. 4–6, a (classical) *source* is a stochastic process. In this paper, we are concerned with discrete sources that can be described as follows. We are given a finite set $\mathscr{X}$, called the alphabet, and consider the space $\mathscr{X}^{\cdot\infty}$ of all infinite sequences $\underline{x}=\{x_n\}_{n\in Z}$, called messages. The time evolution is given by the shift T: $\mathscr{X}^{\cdot\infty}\to\mathscr{X}^{\cdot\infty}$ defined by $(T\underline{x})_n:=x_{n+1}$. The stochastic character of the process is governed by a probability measure μ on $\mathscr{X}^{\cdot\infty}$. The triple $(\mathscr{X}^{\cdot\infty},T,\mu)$ is called a source. We say that the source is *stationary*, if T preserves the measure μ.

We begin by describing informally the properties of a quantum source. In the next section we will provide a rigorous mathematical description. By analogy with the classical case, a quantum source will be a triple, consisting of *quantum messages*, the *time shift*, and a *probability distribution* for the messages. For technical reasons, it is useful to describe the space of quantum messages (which is a linear space) by the algebra of observables on it.

A *quantum source* sends a series of signals, each of which is a vector in a finite-dimensional Hilbert space $\mathscr{H}$. We assume that the source is discrete, i.e., each signal is an element of a finite set $\mathscr{S}=\{|\psi_1\rangle,\ldots,|\psi_s\rangle\}$ of normalized vectors in $\mathscr{H}$. To avoid unnecessary redundancy, we assume that the space $\mathscr{H}$ is spanned by $\mathscr{S}$, i.e. $\mathscr{H}\simeq\mathbb{C}^d$, where $d\leqslant s$. We do not assume that the elements of $\mathscr{S}$ are orthogonal to each other or even linearly independent. Indeed, this is an

important difference from the classical situation, which does not allow for forming linear superpositions of states. Denote by p_j the *a priori* probability of the state $|\psi_j\rangle$ being sent. The density matrix corresponding to the ensemble of signals $\mathscr{S}$ is then given by

$$\rho = \sum_{1 \leqslant j \leqslant s} p_j |\psi_j\rangle\langle\psi_j|. \tag{II.1}$$

As a consequence of our assumptions, $\mathrm{tr}(\rho) = 1$. Clearly, an ensemble of signals $\mathscr{S}$ and the associated *a priori* probabilities determines uniquely the density matrix ρ. On the other hand, any given density matrix corresponds to infinitely many different sets of signals. For a discussion of this point, see Ref. 7.

The observables associated with quantum signals are $d \times d$ Hermitian matrices; more formally (we will need this viewpoint shortly) they are elements of the $\mathbb{C}^*$-algebra $\mathscr{A} = \mathscr{L}(\mathscr{H})$ of linear operators on $\mathscr{H}$. Abusing slightly the language, we will refer to $\mathscr{A}$ as the algebra of observables. Using the language adopted in the operator algebra approach to quantum physics, see, e.g., Refs. 8 and 9, we define the *state* on the algebra of observables $\mathscr{A}$ associated with the density matrix ρ to be

$$\tau_1(A) := \mathrm{tr}(A\rho) = \sum_{1 \leqslant j \leqslant s} p_j \langle\psi_j|A|\psi_j\rangle. \tag{II.2}$$

B. Definition

We now propose a formal definition of a quantum source. To this end, we construct the state on the algebra of observables associated with entire (infinite) quantum messages rather than individual signals. This is technically somewhat delicate, as it involves infinite tensor products of Hilbert spaces. Let $I \subset \mathbb{Z}$ be a finite set of the form $\{M, M+1, \ldots, N-1, N\}$, where $M < N$, i.e., I is a finite collection of consecutive integers. By $\mathscr{T}$ we denote the partially ordered set of all such I's. We set $H_I := \otimes_{j \in I} \mathscr{H}_j$, where $\mathscr{H}_j = \mathscr{H}$ for all $j \in I$, and define the corresponding observable algebra $\mathsf{A}_I := \mathscr{L}(H_I) \simeq \otimes^{|I|} \mathscr{L}(\mathscr{H})$, where $|I|$ denotes the number of elements in I. For $I \subset J$, there is a natural embedding $\mathscr{L}(H_I) \hookrightarrow \mathscr{L}(H_J)$, and so we can form the union $\mathsf{A}_{\mathrm{loc}} := \cup_{I \in \mathscr{T}} \mathsf{A}_I$. The latter is a normed algebra, and we refer to its elements as *local observables*. Roughly, $\mathsf{A}_{\mathrm{loc}}$ is a collection of operators acting on the infinite tensor product $\otimes_{j \in \mathbb{Z}} \mathscr{H}_j$; every element of $\mathsf{A}_{\mathrm{loc}}$ acts as the identity on all but a finite number of factors in this product. For a local observable $A \in \mathsf{A}_{\mathrm{loc}}$, we let $\mathrm{supp}(A)$ denote its support, i.e., the smallest $I \in \mathscr{T}$ such that $A \in \mathsf{A}_I$. The norm closure A of $\mathsf{A}_{\mathrm{loc}}$ is a $\mathbb{C}^*$-algebra called the algebra of *quasilocal observables*. These concepts are borrowed from algebraic field theory and statistical mechanics, and we refer the reader to Refs. 8 and 9 for a thorough presentation.

We will use the net $\{\mathsf{A}_I\}_{I \in \mathscr{T}}$ of matrix algebras to construct the Hilbert space of states of a quantum source. Assume that we have a family $\{\Pi_I\}_{I \in \mathscr{T}}$, $\Pi_I \in \mathsf{A}_I$, satisfying the following assumptions:

1°. Each Π_I is a positive operator.

2°. If $|I| = 1$, then $\Pi_I = \rho$.

3°. Let $I \subset J$ be such that $J \setminus I \in \mathscr{T}$. Then

$$\mathrm{tr}_{H_I} \Pi_J = \Pi_I, \tag{II.3}$$

where tr_{H_I} denotes the partial trace over the factor H_I in the tensor product $H_J = H_I \otimes H_{J \setminus I}$ or $H_J = H_{J \setminus I} \otimes H_I$.

Note that the last condition implies, in particular, that $\mathrm{tr}\, \Pi_I = 1$. In other words, $\{\Pi_I\}_{I \in \mathscr{T}}$ is a *consistent* family of density matrices, and it can be thought of as a quantum mechanical counterpart of a consistent family of cylinder measures.

For each $I \in \mathscr{I}$ we define a state τ_I on A_I by

$$\tau_I(A) := \mathrm{tr}(A\Pi_I), \tag{II.4}$$

and observe that $|\tau_I(A)| \leqslant \|A\|$, uniformly in I. The consistency condition (II.3) implies that τ_I is well defined, and that the generalized limit $\tau(A) := \lim_{I \nearrow \mathbb{Z}} \tau_I(A)$ exists for all $A \in \mathsf{A}_{\mathrm{loc}}$, and satisfies $|\tau(A)| \leqslant \|A\|$. As a consequence, τ can be uniquely extended to a state on the $\mathbb{C}^*$-algebra A of quasilocal observables. We use the same symbol τ to denote this extension.

Let H, π be the GNS representation, see, e.g., Ref. 9, associated with the state τ. The Hilbert space H is the state space of the quantum source. If no confusion arises, we will write A instead of $\pi(A)$.

The additive group $\mathbb{Z}$ underlying the above construction plays the role of (discrete) time translations. Its action on the algebra of local observables is defined as follows:

$$A_I \ni A \simeq A \otimes I \to \alpha(A) := I \otimes A \simeq A \in A_{I+1}. \tag{II.5}$$

In other words, α pushes the observable to the right by one unit of time. Clearly, $\|\alpha(A)\| = \|A\|$, and so α has a unique extension to all of A, which we will denote by the same symbol. The family of automorphisms $\{\alpha^n\}_{n \in \mathbb{Z}}$ defines then a representation of $\mathbb{Z}$. The triple $(\mathsf{A}, \alpha, \tau)$ is called a *quantum source*.

We say that the quantum source is *stationary*, if the state τ is invariant under α, i.e.

$$\tau(\alpha(A)) = \tau(A), \tag{II.6}$$

for all $A \in \mathsf{A}$. From now on we will be assuming that our source is stationary. A standard result in operator algebras implies that the automorphism α is unitarily implementable on the GNS Hilbert space associated with the invariant state τ, i.e., $\alpha(A) = FAF^{-1}$ on H. We will call the unitary operator F a *quantum shift*.

Because of stationarity, we can always assume that $I \in \mathscr{I}$ is of the form $\{1,...,n\}$. To simplify the notation, we will write then Π_n instead of Π_I, and τ_n instead of τ_I.

C. Examples

The simplest example of a quantum source is a *Bernoulli source*, which we now describe. As in the classical case, a Bernoulli source produces messages that are sequences of independent signals. Accordingly, the family of density matrices $\{\Pi_I\}$ is given as follows:

$$\Pi_I = \otimes_{i \in I} \rho_i, \tag{II.7}$$

where each ρ_i equals ρ. It immediately follows that the consistency condition (II.3) is satisfied, and that the source is stationary. This is the class of sources considered by Schumacher.[2] One special feature of a Bernoulli source is that whenever $\mathrm{supp}(A) \cap \mathrm{supp}(B) = \varnothing$, we have $\tau(AB) = \tau(A)\,\tau(B)$.

There are many examples of non-Bernoulli sources. We present here a special class of stationary sources. These are all described by a signal density matrix ρ and another matrix $R \in \mathscr{L}(\mathscr{H} \otimes \mathscr{H})$ satisfying several consistency and positivity conditions. The most basic conditions are

$$R = R^*, \quad \mathrm{tr}_1((\rho \otimes I)R) = \rho, \quad \mathrm{tr}_2(R) = I. \tag{II.8}$$

Here I is the identity matrix on $\mathscr{H}$, and we introduce the notation tr_i for the partial trace over the ith factor in the n-fold tensor product $\otimes^n \mathscr{H}$. The density matrices $\{\Pi_n\}$ are then constructed recursively as follows:

$$\Pi_1 = \rho,$$

$$\Pi_{n+1} = \tfrac{1}{2}(\Pi_n \otimes I_1)(I_{n-1} \otimes R) + \tfrac{1}{2}(I_{n-1} \otimes R)(\Pi_n \otimes I_1).$$

We have denoted by I_n the identity matrix on the product $\otimes^n \mathscr{H}$. The consistency of this definition [(II.3)] is immediate. The positivity of the density matrices Π_n is a further constraint on R. We have several explicit examples for which the positivity can be proven.

In the simplest situation the matrix R satisfies the following additional conditions:

$$R \geqslant 0,$$

$$[\rho \otimes I_1, R] = 0,$$

$$[I_1 \otimes R, R \otimes I_1] = 0. \tag{II.9}$$

It follows readily that all the matrices Π_n are positive, for all $n \geqslant 1$. For example, suppose $\rho = \Sigma_{j=1}^{d} \lambda_j P_j$, where $\{\lambda_j\}$ are the eigenvalues of ρ, and $\{P_j\}$ are the corresponding orthogonal projections. Then we can take $R = \Sigma_{j=1}^{d} P_j \otimes P_j$, and all the above properties are easily seen to hold.

For our second example $\mathscr{H} = \mathbb{C}^2$, and we assume that ρ is strictly positive. Let $\{\sigma_j\}$ be the Pauli matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

By choosing a suitable basis we can write

$$\rho = \frac{1}{2} I + \frac{a}{2} \sigma_3, \tag{II.10}$$

where $|a| < 1$. Then we take the matrix R to be

$$R = I \otimes \rho + \left(\frac{a}{2} I - \frac{1}{2} \sigma_3 \right) \otimes (b\sigma_1 + c\sigma_2). \tag{II.11}$$

The consistency conditions are easily verified. Positivity of all matrices $\{\Pi_n\}$ holds for $|b|$, $|c|$ sufficiently small; the proof is given in the Appendix.

D. Reduction to classical source

The density matrix Π_n can also be written in terms of the states $\{\psi_j\}$ that span $\mathscr{S}$ [compare (II.1)] as follows:

$$\Pi_n = \sum_{1 \leqslant j_1, \ldots, j_n \leqslant s} p_{j_1, \ldots, j_n} |\psi_{j_1}\rangle\langle\psi_{j_1}| \otimes \cdots \otimes |\psi_{j_n}\rangle\langle\psi_{j_n}|.$$

The numbers $p_{j_1, \ldots, j_n}$ satisfy obvious consistency conditions. If the states $\{\psi_j\}$ are linearly independent, then they are also non-negative. Further, if the states are orthogonal, the density matrices $\{\Pi_n\}$ commute. In this case, the quantum source is equivalent to a classical source.

III. ERGODIC QUANTUM SOURCES

A. Definition

According to the individual ergodic theorem,[10] for any function $f \in L^1(\mathscr{X}^{\cdot\infty})$, the sequence of time averages

$$\langle f\rangle_N(\underline{x}) := \frac{1}{N}\sum_{0\leqslant n\leqslant N-1} f(T^n\underline{x}), \tag{III.1}$$

converges almost everywhere to a limit $\overline{f}$. The function $\overline{f}$ is invariant under T. The source $(\mathscr{X}^{\cdot\infty},T,\mu)$ is called *ergodic*, if the only functions invariant under T are constants. Consequently, for an ergodic source $\overline{f}=\int_{\mathscr{X}^{\cdot\infty}} f(\underline{x})\,d\mu(\underline{x})$. This condition is one of the several equivalent statements that could be used to define ergodicity; see, e.g., Ref. 10. There is no natural concept of convergence almost everywhere in quantum mechanics. Consequently, we adopt a notion of convergence that is natural for the algebra A. Our definition of ergodicity of a quantum source is specifically designed to suit the purposes of this paper, although it may have broader applications.

We say that a quantum source is *ergodic* if the following condition is satisfied. For any $A \in \mathsf{A}$, the time averages,

$$\langle A\rangle_N := \frac{1}{N}\sum_{0\leqslant n\leqslant N-1} F^nAF^{-n}, \tag{III.2}$$

converge in a suitable sense to the limit $\tau(A)I$, as $N\to\infty$. Specifically, we require weak convergence, i.e., for all $\phi,\ \psi\in H$,

$$\lim_{N\to\infty}(\phi,(\langle A\rangle_N-\tau(A)I)\psi)=0. \tag{III.3}$$

For $A\in\mathsf{A}$, we let $[A]$ denote the corresponding element of H. Substituting $\phi=[I]$ and $\psi=[B]$, $B\in\mathsf{A}_{\mathrm{loc}}$ in (III.3), we find that for an ergodic source,

$$\lim_{N\to\infty}\tau(\langle A\rangle_N B)=\tau(A)\tau(B), \tag{III.4}$$

for all $A,B\in\mathsf{A}_{\mathrm{loc}}$.

B. Bernoulli source

The simplest example of an ergodic source is a Bernoulli source.

Proposition III.1: A quantum Bernoulli source is ergodic.

Proof: Let $A,B,C\in\mathsf{A}_{\mathrm{loc}}$. We assert that

$$\tau(B^\dagger\langle A\rangle_N C)\to\tau(B^\dagger C)\tau(A), \tag{III.5}$$

which is equivalent to (III.3) with $\phi=[B]$, $\psi=[C]$. To prove this, we observe that there is n_0, such that

$$\tau(B^\dagger F^nAF^{-n}C)=\tau(F^nAF^{-n})\tau(B^\dagger C)=\tau(A)\tau(B^\dagger C),$$

for $n>n_0$ (this follows from the fact that the supports of A, B, and C are all finite). Consequently,

$$\tau(B^\dagger\langle A\rangle_N C)=\frac{1}{N}\sum_{0\leqslant n\leqslant n_0}\tau(B^\dagger F^nAF^{-n}C)+\frac{N-n_0-1}{N}\,\tau(A)\tau(B^\dagger C).$$

But

$$\frac{1}{N}\left|\sum_{0\leqslant n\leqslant n_0}\tau(B^\dagger F^n A F^{-n} C)\right| \leqslant \frac{n_0+1}{N}\,\|A\|\|B\|\|C\| \to 0,$$

and so $\lim_{N\to\infty}\tau(B^\dagger\langle A\rangle_N C)=\tau(A)\,\tau(B^\dagger C)$.

The remainder of the proof is a series of straightforward approximation arguments.

1°. We claim that (III.5) holds for all $A,B,C\in\mathsf{A}$. To prove this, observe that for all N,

$$\|\langle A\rangle_N\|\leqslant\|A\|. \tag{III.6}$$

For $A\in\mathsf{A}$, let $A_j\in\mathsf{A}_{\rm loc}$ be such that $\|A-A_j\|\to 0$, as $j\to\infty$. For any $B,C\in\mathsf{A}_{\rm loc}$, write

$$\begin{aligned}\tau(B^\dagger\langle A\rangle_N C-\tau(A)B^\dagger C)&=\tau(B^\dagger\langle A_j\rangle_N C-\tau(A_j)B^\dagger C)\\&\quad+\tau(B^\dagger\langle A-A_j\rangle_N C)+\tau(A_j-A)\,\tau(B^\dagger C),\end{aligned}$$

and choose j such that $\|A-A_j\|\leqslant\epsilon/(3\|B\|\|C\|)$. Now choose $N_0=N_0(j)$ such that $|\tau(B^\dagger\langle A_j\rangle_N C-\tau(A_j)B^\dagger C)|<\epsilon/3$, for all $N>N_0$. Then, using (III.6),

$$\begin{aligned}|\tau(B^\dagger\langle A\rangle_N C-\tau(A)B^\dagger C)|&<\epsilon/3+\|\langle A-A_j\rangle_N\|\|B\|\|C\|+\|A_j-A\|\|B\|\|C\|\\&<2\|A-A_j\|\|B\|+\epsilon/3\leqslant\epsilon,\end{aligned}$$

for all $N>N_0$. We have thus shown that (III.5) holds for all $A\in\mathsf{A}$ and $B,C\in\mathsf{A}_{\rm loc}$. Repeating twice almost *verbatim* the above 3ϵ argument we establish (III.5) for all $A,B,C\in\mathsf{A}$.

2°. Having established (III.3) for $\phi=[B]$, $\psi=[C]$, and $A\in\mathsf{A}$, we now show that it holds for arbitrary ϕ, $\psi\in\mathsf{H}$. Let $\phi\in\mathsf{H}$, and let B_j be such that $\|\phi-[B_j]\|<\epsilon/(2\|[C]\|\|A\|)$. Write

$$(\phi,(\langle A\rangle_N-\tau(A)I)[C])=([B_j],(\langle A\rangle_N-\tau(A)I)[C])+(\phi-[B_j],(\langle A\rangle_N-\tau(A)I)[C]),$$

and choose $N_0=N_0(j)$ so that $|([B_j],(\langle A\rangle_N-\tau(A)I)[C])|<\epsilon/2$, for all $N>N_0$. Then

$$|(\phi,(\langle A\rangle_N-\tau(A)I)[C])|<\epsilon/2+\|\phi-[B_j]\|\|[C]\|\|A\|<\epsilon.$$

Repeating this argument we establish (III.3) for all $\psi\in\mathsf{H}$. The proof of the proposition is complete.

IV. ENTROPY OF A QUANTUM SOURCE

A. Definition

In this section we construct the entropy of a quantum source. Our construction is largely standard; see, e.g., Refs. 11 and 12, but we include most of the details to make the presentation self-contained. The key mathematical input is the following lemma, known as Klein's inequality,[13] whose proof can be found in Ref. 12, as well as on p. 1122 of Ref. 11.

Lemma IV.1: Let A and B be positive trace class operators on a Hilbert space. Then

$$\mathrm{tr}(A\,\log A-A\,\log B)\geqslant\mathrm{tr}(A-B). \tag{IV.1}$$

We define the entropy associated with a sequence of n signals to be

$$H_n(\Pi):=-\mathrm{tr}_{\mathscr{H}^{\otimes n}}(\Pi_n\,\log\,\Pi_n). \tag{IV.2}$$

Substituting $A=\Pi_{m+n}$ and $B=\Pi_m\otimes\Pi_n$ in (IV.1), we obtain the following subadditivity property of $H_n(\Pi)$ (this is a special case of the well-known subadditivity of quantum mechanical entropy; see Ref. 11):

$$H_{m+n}(\Pi)\leqslant H_m(\Pi)+H_n(\Pi). \tag{IV.3}$$

A standard argument, see, e.g., Ref. 6, pp. 48–49, shows that the limit

$$h(\Pi)=\lim_{n\to\infty}\frac{1}{n}H_n(\Pi) \tag{IV.4}$$

exists. We call $h(\Pi)$ the entropy of the quantum source. Notice also that subadditivity implies the inequality

$$h(\Pi)\leqslant -\mathrm{tr}_{\mathscr{H}}(\rho\ \log\ \rho)\leqslant \log\ d. \tag{IV.5}$$

where ρ is the density matrix for the signals, and d is the dimension of the signal Hilbert space $\mathscr{H}$.

B. Examples

We can easily compute the entropy of two basic types of quantum sources introduced in Sec. II.

For a Bernoulli source, $\Pi_n=\rho\otimes\cdots\otimes\rho$, and so

$$\begin{aligned}\frac{1}{n}H_n(\Pi)&=-\frac{1}{n}\,\mathrm{tr}_{\mathscr{H}^{\otimes n}}((\otimes^n\rho)(\log\otimes^n\rho))\\&=-\frac{1}{n}\sum_{1\leqslant j\leqslant n}\mathrm{tr}_{\mathscr{H}^{\otimes n}}((\otimes^n\rho)(I\otimes\cdots\otimes\log\ \rho\otimes\cdots\otimes I))\\&=-\mathrm{tr}_{\mathscr{H}}(\rho\ \log\ \rho).\end{aligned}$$

As a result, $h(\Pi)=-\mathrm{tr}(\rho\ \log\rho)$, i.e., the entropy of a Bernoulli source is equal to the von Neumann entropy of the signal density matrix.

For the non-Bernoulli source defined in (II.9) a similar calculation yields $h(\Pi)=-\mathrm{tr}((\rho\otimes I)R\ \log R)$. We have not found a closed form expression for the entropy of the other non-Bernoulli source described in Sec. II.

V. ASYMPTOTIC EQUIPARTITION PROPERTY

A. Classical result

The classical asymptotic equipartition property, also known as the Shannon–McMillan theorem—see, e.g., Refs. 4 and 6—states that if $(\mathscr{X}^{\cdot\infty},T,\mu)$ is an ergodic source with entropy $h(\mu)$, then the sequence,

$$f_n(\underline{x}):=-\frac{1}{n}\log\ \mu(\{\underline{y}\in\mathscr{X}^{\cdot\infty}:y_1=x_1,\ldots,y_n=x_n\}), \tag{V.1}$$

converges in measure to $h(\mu)$. In other words, given $\delta,\epsilon>0$, there is n_0 such that for all $n\geqslant n_0$,

$$\mu(\{\underline{x}\in\mathscr{X}:|f_n(\underline{x})-h(\mu)|>\delta\})<\epsilon. \tag{V.2}$$

This is interpreted as saying that for any length n, there are two categories of messages sent by a source: (i) a small fraction of ''likely'' messages, each of which carries equal probability,

$$\mu(\{\underline{y}\in\mathscr{X}^{\infty}: y_1=x_1,\ldots,y_n=x_n\})\sim e^{-nh(\mu)}, \tag{V.3}$$

and (ii) the bulk of ''unlikely'' messages, whose total probability goes to zero as n goes to infinity. There are approximately $e^{nh(\mu)}$ likely messages, which is much less than the total number of messages $e^{n\log r}$ [unless $h(\mu)$ happens to equal $\log r$].

The goal of this section is to establish an analogous result for quantum sources. Our theorem generalizes Schumacher's result[2,3] to general, not necessary Bernoulli, sources.

B. POM

Let $\mathbf{A}=\{A_1,\ldots,A_r\}$, $r<\infty$ be a family of observables on $\mathscr{H}$ such that $A_j\geqslant 0$, for all j, and

$$A_1+\cdots+A_r=I. \tag{V.4}$$

Such a family is called a *positive operator-valued measure* (POM); see Ref. 7 and references therein. A POM is called *pure* if each A_J is a rank one operator. For example, any family of $d=\dim\mathscr{H}$ pairwise orthogonal projections on $\mathscr{H}$ satisfies the above conditions, and so is a pure POM. We will call the set $\mathscr{X}_{\mathbf{A}}=\{1,...,r\}$ the classical alphabet associated with the POM $\mathbf{A}$, and denote by $\mathscr{X}_{\mathbf{A}}^{\infty}$ the space of all infinite messages over the alphabet $\mathscr{X}_{\mathbf{A}}$. We can define a probability measure on $\mathscr{X}_{\mathbf{A}}^{\infty}$ associated with the quantum source $(\mathbf{A},\tau,\alpha)$. For $\{k_1,\ldots,k_n\}\in\mathscr{X}^n$, we define

$$\mu_n^{\mathbf{A}}(\{\underline{x}: x_1=k_1,\ldots,x_n=k_n\}):=\tau_n(A_{k_1}\otimes\cdots\otimes A_{k_n}). \tag{V.5}$$

This defines a consistent family of cylinder measures, as

$$\sum_{k\in\mathscr{X}_{\mathbf{A}}}\mu_{n+1}^{\mathbf{A}}(\{\underline{x}: x_1=k_1,\ldots,x_n=k_n,x_{n+1}=k\})=\tau_{n+1}\left(A_{k_1}\otimes\cdots\otimes A_{k_n}\otimes\sum_{k\in\mathscr{X}_{\mathbf{A}}}A_k\right)=\tau_{n+1}(A_{k_1}\otimes\cdots$$
$$\otimes A_{k_n}\otimes I)=\tau_n(A_{k_1}\otimes\cdots\otimes A_{k_n})=\mu_n^{\mathbf{A}}(\{\underline{x}: x_1=k_1,\ldots,x_n=k_n\}).$$

Let $\mu^{\mathbf{A}}$ denote the probability measure obtained from $\{\mu_n^{\mathbf{A}}\}$ by means of Kolmogorov's extension theorem. By T we denote the shift operator on $\mathscr{X}_{\mathbf{A}}^{\infty}$. Then the triple $(\mathscr{X}_{\mathbf{A}}^{\infty},T,\mu^{\mathbf{A}})$ forms a classical information source. We emphasize that it depends on the choice of $\mathbf{A}$. Obviously, this source is stationary.

C. Classical ergodicity from quantum

In fact, this source is ergodic if the underlying quantum source is ergodic. We state it as the following lemma.

Lemma V.1: If (A,α,τ) *is ergodic, then for any choice of* $\mathbf{A}$, *the classical source constructed above is ergodic.*

Proof: We show that $(\mathscr{X}_{\mathbf{A}}^{\infty},T,\mu^{\mathbf{A}})$ has the following property: for every $f\in L^1(\mathscr{X}^{\infty},d\mu^{\mathbf{A}})$, the sequence $\langle f\rangle_N$ converges in L^1 to $\int f\,d\mu^{\mathbf{A}}$. Using Fatou's lemma, and recalling the individual ergodic theorem, this implies that the only functions invariant under T are constants, and so the classical source is ergodic. We first observe that it is sufficient to show that if $\mathscr{C}$ is a cylinder set, and $\chi_{\mathscr{C}}$ denotes the corresponding characteristic function, then

$$\int_{\mathscr{X}_{\mathbf{A}}^{\infty}}|\langle\chi_{\mathscr{C}}\rangle_N(\underline{x})-\mu^{\mathbf{A}}(\mathscr{C})|\,d\mu^{\mathbf{A}}(\underline{x})\to 0, \tag{V.6}$$

as $N\to\infty$. A standard ''$\epsilon/3$'' argument then implies that for all $f\in L^1(\mathscr{X}^{\infty},d\mu^{\mathbf{A}})$, $\langle f\rangle_N$ converges to $\int f\,d\mu^{\mathbf{A}}$ in the L^1 norm. Furthermore, since $d\mu^{\mathbf{A}}$ is a probability measure, (V.6) will follow

from convergence of $\langle \chi_{\mathscr{C}} \rangle_N$ to $\mu^{\mathbf{A}}(\mathscr{C})$ in the L^2 norm. This in turn is implied by the following stronger result. Suppose $\mathscr{C}$ and $\mathscr{D}$ are cylinder sets, and $\chi_{\mathscr{C}}$ and $\chi_{\mathscr{D}}$ are the corresponding characteristic functions; then

$$\int_{\mathscr{X}_{\mathbf{A}}^{\infty}} \langle \chi_{\mathscr{C}} \rangle_N(\underline{x}) \chi_{\mathscr{D}}(\underline{x}) d\mu^{\mathbf{A}}(\underline{x}) \to \mu^{\mathbf{A}}(\mathscr{C}) \mu^{\mathbf{A}}(\mathscr{D}), \tag{V.7}$$

as $N \to \infty$.

In order to prove (V.7), let $\mathscr{C} = \{\underline{x} : x_1 = k_1, \ldots, x_n = k_n\}$, $\mathscr{D} = \{\underline{x} : x_{j+1} = l_1, \ldots, x_{j+m} = l_m\}$. The corresponding observables are given by

$$G(\mathscr{C}) = A_{k_1} \otimes \cdots \otimes A_{k_n}, \quad G(\mathscr{D}) = A_{l_1} \otimes \cdots \otimes A_{l_m}.$$

There is N_0 such that for all $n > N_0$, $T^n(\mathscr{C}) \cap \mathscr{D} = \varnothing$. Therefore,

$$\frac{1}{N} \int_{\mathscr{X}_{\mathbf{A}}^{\infty}} \sum_{i=N_0+1}^{N-1} \mathscr{X}_{\mathscr{C}}(T^i \underline{x}) \mathscr{X}_{\mathscr{D}}(\underline{x}) d\mu^{\mathbf{A}}(\underline{x}) = \frac{1}{N} \tau\left(\sum_{i=N_0+1}^{N-1} \alpha^i(G(\mathscr{C})) G(\mathscr{D}) \right)$$

$$= \tau(\langle G(\mathscr{C}) \rangle_N) G(\mathscr{D})) - \frac{1}{N} \tau\left(\sum_{i=0}^{N_0} \alpha^i(G(\mathscr{C})) G(\mathscr{D}) \right).$$

By our assumption of quantum ergodicity, the first term above converges to $\mu^{\mathbf{A}}(\mathscr{C}) \mu^{\mathbf{A}}(\mathscr{D})$, as $N \to \infty$. The second term is bounded as follows:

$$\left| \frac{1}{N} \tau\left(\sum_{i=0}^{N_0} \alpha^i(G(\mathscr{C})) G(\mathscr{D}) \right) \right| \leqslant \frac{N_0 + 1}{N} \prod_{p=1}^{n} \|A_{k_p}\| \prod_{q=1}^{l} \|A_{l_q}\|,$$

which converges to 0 as $N \to \infty$. Finally,

$$\frac{1}{N} \int_{\mathscr{X}_A^{\infty}} \sum_{i=0}^{N_0} \chi_{\mathscr{C}}(T^i \underline{x}) \chi_{\mathscr{D}}(\underline{x}) d\mu^A(\underline{x}) \geqslant \frac{N_0 + 1}{N} \to 0,$$

and the proof is complete.

D. Bounds for the entropy

Let $h_{\mathbf{A}}$ denote the Shannon entropy of the classical source $(\mathscr{X}_{\mathbf{A}}^{\infty}, T, \mu^{\mathbf{A}})$ (Refs. 5 and 6). In other words,

$$h_{\mathbf{A}} = -\lim_{n \to \infty} \frac{1}{n} \sum_{k_1, \ldots, k_n \in \mathscr{X}_{\mathbf{A}}} \mu_{\mathbf{A}}(\{\underline{x} : x_1 = k_1, \ldots, x_n = k_n\}) \times \log \mu_{\mathbf{A}}(\{\underline{x} : x_1 = k_1, \ldots, x_n = k_n\})$$

$$= -\lim_{n \to \infty} \frac{1}{n} \sum_{k_1, \ldots, k_n \in \mathscr{X}_{\mathbf{A}}} \tau(A_{k_1} \otimes \cdots \otimes A_{k_n}) \log \tau(A_{k_1} \otimes \cdots \otimes A_{k_n}).$$

A remarkable fact about $h_{\mathbf{A}}$ is that it can be bounded from below in terms of $h(\Pi)$, the quantum entropy of the source, and a quantity depending exclusively on the statistical properties of the signal ensemble.

Theorem V.2: *For any POM* **A** *the following inequality holds*:

$$h_{\mathbf{A}} \geqslant h(\Pi) - \sum_{1 \leqslant k \leqslant r} \operatorname{tr}(A_k \rho) \log \operatorname{tr}(A_k). \tag{V.8}$$

Remark: In particular, if **A** is a pure POM consisting of d mutually orthogonal projections, then $h_{\mathbf{A}} \geqslant h(\Pi)$.

Proof: The proof of this theorem is based on the following lemma.

Lemma V.3: Let $\mathscr{F}$ be a Hilbert space of finite dimension N, and let B be a density matrix on $\mathscr{F}$. If $A_1,\ldots,A_r$ are positive operators such that $A_1+\cdots+A_r=I$, then

$$\sum_{1 \leqslant j \leqslant r} \operatorname{tr}(A_j B) \log \operatorname{tr}(A_j B) \leqslant \operatorname{tr}(B \log B) + \sum_{1 \leqslant j \leqslant r} \operatorname{tr}(A_j B) \log \operatorname{tr}(A_j). \tag{V.9}$$

Proof: Let A_j^{kl} denote the matrix entries of A_j in an orthonormal basis consisting of eigenvectors of B. Then $\operatorname{tr}(A_j B) = \Sigma_k \, \lambda_k A_j^{kk}$, where $\lambda_1,\ldots,\lambda_N$ are the eigenvalues of B. The function $[0,1] \ni x \to f(x) := x \log x$ is convex, and so by Jensen's inequality,

$$\begin{aligned}
\sum_j f\left(\sum_k \lambda_k A_j^{kk} \right) &= \sum_j \operatorname{tr}(A_j) f\left(\sum_k \frac{A_j^{kk}}{\operatorname{tr}(A_j)} \lambda_k \right) + \sum_j \sum_k \frac{A_j^{kk}}{\operatorname{tr}(A_j)} \lambda_k f(\operatorname{tr}(A_j)) \\
&\leqslant \sum_j \sum_k A_j^{kk} f(\lambda_k) + \sum_j \operatorname{tr}(A_j B) \log \operatorname{tr}(A_j) \\
&= \sum_k f(\lambda_k) + \sum_j \operatorname{tr}(A_j B) \log \operatorname{tr}(A_j),
\end{aligned}$$

which implies (V.9).

As a consequence of this lemma, and condition 3° in the definition of a consistent family of density matrices,

$$\begin{aligned}
&\sum_{k_1,\ldots,k_n \in \mathscr{X}_{\mathbf{A}}} \operatorname{tr}((A_{k_1} \otimes \cdots \otimes A_{k_n}) \Pi_n) \log \operatorname{tr}((A_{k_1} \otimes \cdots \otimes A_{k_n} \Pi_n)) \\
&\quad \leqslant \operatorname{tr}(\Pi_n \log \Pi_n) + \sum_{k_1,\ldots,k_n \in \mathscr{X}_{\mathbf{A}}} \operatorname{tr}((A_{k_1} \otimes \cdots \otimes A_{k_n}) \Pi_n) \log \operatorname{tr}(A_{k_1} \otimes \cdots \otimes A_{k_n}) \\
&\quad = \operatorname{tr}(\Pi_n \log \Pi_n) + \sum_{1 \leqslant j \leqslant n} \sum_{k_1,\ldots,k_n \in \mathscr{X}_{\mathbf{A}}} \operatorname{tr}((A_{k_1} \otimes \cdots \otimes A_{k_n}) \Pi_n) \log \operatorname{tr}(A_{k_j}) \\
&\quad = \operatorname{tr}(\Pi_n \log \Pi_n) + n \sum_{1 \leqslant k \leqslant r} \operatorname{tr}(A_k \rho) \log \operatorname{tr}(A_k),
\end{aligned}$$

and the claim follows.

E. Main theorem

The theorem below is the main result of this section. It can be regarded as a quantum version of the Shannon–McMillan theorem.

Theorem V.4: *Let (A,τ,α) be an ergodic source, and let* **A** *be a POM, for which the operators $\{A_j\}$, $1 \leqslant j \leqslant r$, are orthogonal projections. Let $h_{\mathbf{A}}$ be the Shannon entropy of the associated classical source, and define*

$$M = \max_{1 \leqslant j \leqslant r} \operatorname{rank}(A_j), \quad m = \min_{1 \leqslant j \leqslant r} \operatorname{rank}(A_j).$$

Then, given $\delta,\epsilon>0$, *there is* n_0, *such that for all* $n\geqslant n_0$,

$$\mathscr{H}^{\otimes n}=\mathscr{S}_n\oplus\mathscr{S}_n^{\perp},$$

where $\mathscr{S}_n$ *is a subspace whose dimension satisfies*

$$\frac{\log m-\delta}{\log d}\leqslant\frac{\log\dim\mathscr{S}_n}{\log\dim\mathscr{H}^{\otimes n}}-\frac{h_{\mathbf{A}}}{\log d}\leqslant\frac{\log M+\delta}{\log d}. \tag{V.10}$$

Further, let $P_{\mathscr{S}_n}$ *be the orthogonal projection onto* $\mathscr{S}_n$. *Then for any observable* $C\in\mathscr{L}(\mathscr{H}^{\otimes n})$,

$$|\tau(CP_{\mathscr{S}_n})-\tau(C)|<\epsilon\|C\|. \tag{V.11}$$

Remark 1: We can regard $\mathscr{S}_n$ as a significant subspace of $\mathscr{H}^{\otimes n}$ in the sense that the expectation of any observable is almost completely determined by its restriction to $\mathscr{S}_n$. If the states $\{|\psi_j\rangle\}$ are orthogonal, we can take the d orthogonal projections $\{|\psi_j\rangle\langle\psi_j|\}$ for the POM. In this case the entropy $h_{\mathbf{A}}$ is equal to the von Neumann entropy, since the density operators all commute. Then there is a direct correspondence with the classical Shannon–McMillan theorem, and the quantum theory is just a restatement of the classical result.

Remark 2: In the Bernoulli case we can take the POM $\mathbf{A}$ to be d orthogonal projections onto the eigenvectors of ρ, in which case the inequality of Theorem V.2 is saturated. If ρ has a simple spectrum, this means that the Shannon entropy equals the von Neumann entropy of the quantum source. Our result then agrees with Schumacher's conclusion that the information contained in the quantum source resides in a subspace whose dimension is asymptotically $e^{nh_{\mathbf{A}}}$. In the general case we obtain only an upper bound for the dimension of the relevant subspace, and this upper bound depends on the choice of POM. For example, if each operator A_j in the POM $\mathbf{A}$ is equal to $(1/d)I$, where I is the identity, then the Shannon entropy is $h_{\mathbf{A}}=\log d$. Since this is the maximum possible entropy for a POM with d operators, we can conclude that all information about the quantum source has been lost in this measurement process. As these results show, it is advantageous to use a POM composed of orthogonal projections.

Proof: we use the POM to construct the classical source $\mathscr{X}_{\mathbf{A}}^{\infty}$ with entropy $h_{\mathbf{A}}$. Let $f_n(\underline{x})$ be the empirical entropy of a message $\underline{x}$ of length n defined in (V.1). Given $\epsilon,\delta>0$, let us define the sets

$$U_{n,\delta}=\{\underline{x}\in\mathscr{X}:|f_n(\underline{x})-h_{\mathbf{A}}|>\delta\},$$

$$L_{n,\delta}=\{\underline{x}\in\mathscr{X}:|f_n(\underline{x})-h_{\mathbf{A}}|\leqslant\delta\}.$$

By the Shannon–McMillan theorem, given $\epsilon,\delta>0$, there is n_0, such that for all $n\geqslant n_0$,

$$\sum_{\mathbf{x}\in U_{n,\delta}}\mu_{\mathbf{A}}(\{\underline{x}\})<\epsilon.$$

Since the operators $\{A_j\}$ are orthogonal projections, each tensor product $A_{k_1}\otimes\cdots\otimes A_{k_n}$ is an orthogonal projection, and hence so is the sum of these operators over the set $L_{n,\delta}$. Let $\mathscr{S}_n$ denote the range of this projection, and let $P_{\mathscr{S}_n}$ denote the orthogonal projection onto this subspace. Then for any observable C, we have

$$|\tau(CP_{\mathscr{S}_n})-\tau(C)|\leqslant\epsilon\|C\|.$$

It remains to estimate the dimension of $\mathscr{S}_n$. Since the projections $\{A_j\}$ are orthogonal, its dimension is given by

$$\dim(\mathscr{S}_n)=\sum_{\mathbf{x}\in L_{n,\delta}}\prod_{j=1}^{n}\operatorname{rank}(A_{x_j}).$$

Therefore,

$$m^n|L_{n,\delta}|\leqslant\dim(\mathscr{S}_n)\leqslant M^n|L_{n,\delta}|,$$

where $|L_{n,\delta}|$ is the size of the set $L_{n,\delta}$. The Shannon–McMillan theorem implies that

$$(1-\epsilon)e^{n(h_{\mathbf{A}}+\delta)}\geqslant|L_{n,\delta}|\geqslant e^{n(h_{\mathbf{A}}-\delta)}.$$

This leads to

$$\frac{\log m}{\log d}-\frac{\delta}{\log d}\leqslant\frac{\dim(\mathscr{S}_n)}{n\log d}-\frac{h_{\mathbf{A}}}{\log d}\leqslant\frac{\log M}{\log d}+\frac{\delta}{\log d}+\frac{\log(1-\epsilon)}{n},$$

and the result follows.

ACKNOWLEDGMENTS

Supported in part by the National Science Foundation under Grant No. DMS-9424344. We would like to thank Evelyn Wright for helpful discussions, and the referee for useful suggestions.

APPENDIX: PROOF OF POSITIVITY

We present here the proof that the R matrix (II.11) defines a family of positive density matrices, when $|b|$ and $|c|$ are sufficiently small. It is convenient to introduce the matrices

$$\omega=\frac{a}{2}\,I-\frac{1}{2}\,\sigma_3,$$

$$Q=\rho^{-1}(b\sigma_1+c\sigma_2).$$

Note that we assume $|a|<1$, so ρ^{-1} exists. Then for all $n\geqslant 2$ we define

$$S_n=I_{n-2}\otimes\omega\otimes Q.$$

It follows by direct calculation that for all $n\geqslant 1$,

$$\Pi_{n+1}=\Pi_n\otimes\rho+\tfrac{1}{2}(\Pi_n\otimes\rho)S_{n+1}+\tfrac{1}{2}S_{n+1}^{\dagger}(\Pi_n\otimes\rho).$$

In order to proceed we make the inductive assumption that $\Pi_n>0$; this implies, in particular that $(\Pi_n+u)^{-1}$ is bounded for every $u\geqslant 0$. We will prove that $(\Pi_{n+1}+u)^{-1}$ is bounded for every $u\geqslant 0$; together with the positivity of $\Pi_1=\rho$, this will establish the desired result.

Note first that $\|S_n\|\leqslant\|\omega\|\|Q\|$, and this bound is uniform in n. For convenience we define

$$w=\|\omega\|,\quad q=\|Q\|.$$

Furthermore, for any $u\geqslant 0$,

$$(\Pi_{n+1}+u)^{-1}=(\Pi_n\otimes\rho+u)^{-1}(I+\tfrac{1}{2}(\Pi_n\otimes\rho)S_{n+1}(\Pi_n\otimes\rho+u)^{-1}$$
$$+\tfrac{1}{2}S_{n+1}^{\dagger}(\Pi_n\otimes\rho)(\Pi_n\otimes\rho+u)^{-1})^{-1}.$$

Our inductive assumption implies that $(\Pi_n\otimes\rho+u)^{-1}$ is bounded for $u\geqslant 0$. Choosing $|b|$, $|c|$ sufficiently small guarantees that $\|S_n^{\dagger}\|<\epsilon$ for any $\epsilon>0$. Furthermore,

$$(\Pi_n\otimes\rho)S_{n+1}(\Pi_n\otimes\rho+u)^{-1}=([\Pi_n(I_{n-1}\otimes\omega)\Pi_n^{-1}]\otimes Q^{\dagger})(\Pi_n\otimes\rho)(\Pi_n\otimes\rho+u)^{-1}.$$

Since $\|Q^{\dagger}\|=q$ can be made arbitrarily small by choosing $|b|$, $|c|$ sufficiently small, the boundedness of $(\Pi_{n+1}+u)^{-1}$ will follow from a bound for the operator $\Pi_n(I_{n-1}\otimes\omega)\Pi_n^{-1}$, which is uniform in n. Accordingly let us define for $n\geqslant 1$,

$$A_n=\Pi_n(I_{n-1}\otimes\omega)\Pi_n^{-1}.$$

By imitating the derivation above, we obtain the recursion relation

$$A_n=(I+\tfrac{1}{2}A_{n-1}\otimes Q^{\dagger}+\tfrac{1}{2}S_n^{\dagger})(I_{n-1}\otimes\omega)(I+\tfrac{1}{2}A_{n-1}\otimes Q^{\dagger}+\tfrac{1}{2}S_n^{\dagger})^{-1}. \tag{A.1}$$

It is immediate that $\|A_1\|=\|\omega\|\leqslant 1$. We make the inductive assumption that $\|A_{n-1}\|\leqslant 1$; then (A1) implies the estimate

$$\|A_n\|\leqslant\left(1+\frac{q}{2}+\frac{wq}{2}\right)w\left(1-\frac{q}{2}-\frac{wq}{2}\right)^{-1}. \tag{A.2}$$

If we choose

$$q<2\frac{1-w}{(1+w)^2}, \tag{A.3}$$

then (A2) implies that $\|A_n\|\leqslant 1$. Hence by choosing (A3) we obtain that A_n is uniformly bounded for all n, and hence that $(\Pi_{n+1}+u)^{-1}$ is bounded for all $u\geqslant 0$. Therefore the positivity of the density matrices is proved.

[1] C. H. Bennett, ''Quantum information and computation,'' Phys. Today **48**, 24–30 (1995).
[2] B. Schumacher, ''Quantum coding,'' Phys. Rev. A **51**, 2738–2747 (1995).
[3] R. Jozsa and B. Schumacher, ''A new proof of the quantum noiseless coding theorem,'' J. Mod. Opt. **41**, 2343–2349 (1994).
[4] R. B. Ash, *Information Theory* (Dover, New York, 1990).
[5] P. Billingsley, *Ergodic Theory and Information* (Krieger, Huntington, NY, 1978).
[6] A. I. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1957).
[7] L. P. Hughston, R. Jozsa, and W. Wootters, ''A complete classification of quantum ensembles having a given density matrix,'' Phys. Lett. A **183**, 14–18 (1993).
[8] R. Haag, *Local Quantum Physics* (Springer-Verlag, Berlin, 1992).
[9] O. Bratteli and D. W. Robinson *Operator Algebras and Quantum Statistical Mechanics* (Springer-Verlag, Berlin, 1981), Vols. 1 and 2.
[10] I. P. Cornfeld, S. V. Fomin, and Ya. Sinai, *Ergodic Theory* (Springer-Verlag, Berlin, 1982).
[11] O. E. Lanford and D. W. Robinson, ''Mean entropy of states in quantum-statistical systems,'' J. Math. Phys. **9**, 1120–1125 (1968).
[12] A. Wehrl, ''General properties of entropy,'' Rev. Mod. Phys. **50**, 221–260 (1978).
[13] O. Klein, Z. Phys. **72**, 767–775 (1931).

Moving frames hierarchy and BF theory

Jyh-Hao Lee[a)]
Institute of Mathematics, Academia Sinica, Taipei 11529, Taiwan

Oktay K. Pashaev[b)]
Joint Institute for Nuclear Research, Dubna (Moscow) 141980, Russia

(Received 19 June 1997; accepted for publication 6 August 1997)

We show that the one-dimensional projection of Chern–Simons gauged nonlinear Schrödinger model is equivalent to an Abelian gauge field theory of a continuum Heisenberg spin chain. In such a theory, the matter field has geometrical meaning of coordinates in tangent plane to the spin phase space, while the $U(1)$ gauge symmetry relates to rotation in the plane. This allows us to construct the infinite hierarchy of integrable gauge theories and related magnetic models. To each of them a $U(1)$ invariant gauge fixing constraint of non-Abelian BF theory is derived. The corresponding moving frames hierarchy is obtained and the spectral parameter is interpreted as a constant-valued statistical gauge potential constrained by the 1-cocycle condition. © *1998 American Institute of Physics.*
[S0022-2488(98)01201-8]

I. INTRODUCTION

The nonlinear Schrödinger equation (NLS) in two space dimensions interacting with Abelian Chern–Simons gauge field (the Jackiw–Pi model) has attracted much attention recently due to the beautiful structure of the static limit, admitting N-soliton solution.[1] After quantization these solitons become quasiparticles with arbitrary statistics, called anyons, while the Chern–Simons gauge field is interpreted as the ''statistical'' one. The anyons have an interesting application to the planar physical phenomena as the Quantum Hall Effect.[2,3] Very recently, attempts to study the generalized statistics of many particle systems in $1+1$ dimensions and the relation with $2+1$-dimensional anyons were done.[4–6] For configurations depending of the one space direction (the *lineal* theory), the reduced theory describes $1+1$-dimensional NLS, interacting with an Abelian BF gauge field. Due to the pure gauge form it is possible to exclude this field from the gauge-invariant description. However, since in one dimension NLS is an integrable system, an interesting problem of the relation between a long-range structure of the statistical gauge field in two and integrability in one dimensions arises.

In the present paper we show that the ''trivial'' gauge field plays the crucial role for integrability of the NLS model. Namely, the homogeneous statistical gauge field is identical to the spectral parameter for the NLS linear problem. Our construction is based on a mapping of the reduced Jackiw–Pi model to non-Abelian BF gauge theory. This mapping has to appear as the gauge fixing condition corresponding to the classical Heisenberg spin model. Then, the matter field has the geometrical meaning of coordinates in tangent to the spin phase space plane, while $U(1)$ gauge symmetry relates to the rotation in the plane.

In Sec. I we consider dimensional reduction of the Jackiw–Pi model. We find a representation of the model as a $U(1)$-invariant gauge fixing condition of non-Abelian BF theory. This allows us to construct an infinite hierarchy of associated integrable models and the related gauge fixing constraints. In Sec. II we show that the model has natural interpretation as the gauge theory of a continuum Heisenberg model defined on the sphere S^2 or hyperboloid $S^{1,1}$, according to the sign

[a)]Electronic mail: leejh@ccvax.sinica.edu.tw
[b)]Electronic mail: pashaev@vxjinr.jinr.ru

0022-2488/98/39(1)/102/22/$15.00
© 1998 American Institute of Physics

of nonlinearity. Then, we derive the associated Heisenberg and moving frame hierarchies. Section III is devoted to a $U(1)$ gauge-invariant description of integrable time hierarchy in the BF theory context. We show that the B field plays the role of squared eigenfunctions for the Zakharov–Shabat linear problem. In terms of the recursion operator eigenvalue problem, the hierarchy of constraints is related to the evolution hierarchy and higher Poisson structures. In the Conclusion we discuss the chiral solitons for odd members of the hierarchy and some open problems.

II. THE JACKIW–PI MODEL IN 1+1 DIMENSIONS

The Lagrangian of a $2+1$-dimensional nonlinear Schrödinger model interacting with a Chern–Simons gauge field is[1]

$$\mathscr{L}=\frac{i}{2}\,(\overline{\psi}\,D_0\psi-\overline{D_0\psi}\,\psi)-\overline{D_1\psi}\,D_1\psi-\overline{D_2\psi}\,D_2\psi+g|\psi|^4+\frac{k}{4}\,\epsilon^{\mu\nu\lambda}A_\mu\,\partial_\nu\,A_\lambda\,, \tag{2.1}$$

where $\psi=\psi(x_1,x_2,t)$ is complex matter and $A_\mu=A_\mu(x_1,x_2,t)$, $(\mu=0,1,2)$, is a $U(1)$ Abelian (statistical) gauge field. Here, $D_\mu=\partial_\mu-(i/2)A_\mu$, denotes the covariant derivative, g and k are the self-interaction and the Chern–Simons coupling constants, respectively. The related Euler–Lagrange equations of motion are

$$i\,D_0\psi+(D_1^2+D_2^2)\psi+2g|\psi|^2\psi=0, \tag{2.2a}$$

$$\epsilon^{ij}\,\partial_i A_j=-\frac{1}{k}\,|\psi|^2, \tag{2.2b}$$

$$\partial_0 A_i-\partial_i A_0=-\frac{i}{k}\,\epsilon^{ij}(\overline{D_j\psi}\,\psi-\overline{\psi}\,D_j\psi),(i,j=1,2). \tag{2.2c}$$

In the special case when the coupling constants connected by the relation $g=1/k$, the static configurations of this model are subject to the self-dual Chern–Simons equations.[1] The last one admits the linear representation[7] and has been related to the Liouville model with N-vortex/soliton solutions.[1] However, there is no evidence that dynamics of these solitons according to Eqs. (2.2) is also integrable. Recently,[8] the integrable dynamics of the Chern–Simons solitons has been described by the Davey–Stewartson equation, being considered as the $2+1$-dimensional extension of the NLS. Furthermore, it is shown below that model (2.1) reduced to $1+1$ dimension is also integrable. The corresponding soliton dynamics subject to the NLS.

Let us consider the boundary as a rectangle on the plane. Then, the boundary dynamics of (2.1) is described by the $1+1$ reduced model. For the x_2-independent part we obtain the Lagrangian,

$$\mathscr{L}=\frac{i}{2}\,(\overline{\psi}\,D_0\psi-\overline{D_0\psi}\,\psi)-\overline{D_1\psi}\,D_1\psi-\frac{1}{4}\,B^2\overline{\psi}\psi+g|\psi|^4+\frac{k}{4}\,B\epsilon^{\mu\nu}F_{\mu\nu}, \tag{2.3}$$

where $B\equiv A_2$ is the gauge field component in the compactified direction, and $F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu$ $(\mu=0,1)$. For the vanishing matter field $\psi=0$, this Lagrangian reduces to a pure Abelian BF theory and this is why the model (2.3) can be considered as the $1+1$-dimensional BF gauged NLS. Due to the nonpropagating character, the gauge field can be eliminated from (2.3). Nevertheless, as we can see below, (2.3) is equivalent to the classical spin model and the statistical gauge potential carries an important information about the spectrum of the model. Moreover, when $\psi\neq 0$, the model (2.3) still is the BF theory, however, for a non-Abelian gauge group in a special gauge.

To proceed first we redefine gauge potentials (and the related covariant derivatives) as

$$W_0 \equiv A_0 - \frac{1}{2} B^2, \quad W_1 \equiv A_1 . \tag{2.4}$$

Then, the Lagrangian (2.3) becomes

$$\mathscr{L} = \frac{i}{2} (\overline{\psi} D_0 \psi - \overline{D_0 \psi}\, \psi) - \overline{D_1 \psi} D_1 \psi + g |\psi|^4 + \frac{k}{4} B \epsilon^{\mu\nu} F_{\mu\nu} + \Delta \mathscr{L}, \tag{2.5}$$

where

$$\Delta \mathscr{L} = - \frac{k}{6} \partial_1 B^3, \tag{2.6}$$

is the total derivative (which becomes nontrivial at the boundaries) and does not influence equations of motion. The Euler–Lagrange equations for (2.5) are given by the system

$$i\ D_0 \psi + D_1^2 \psi + 2g |\psi|^2 \psi = 0, \tag{2.7a}$$

$$\partial_0 W_1 - \partial_1 W_0 = 0, \tag{2.7b}$$

and

$$\partial_1 B = - \frac{1}{k} |\psi|^2, \tag{2.8a}$$

$$\partial_0 B = \frac{i}{k} (\overline{D_1 \psi}\, \psi - \overline{\psi} D_1 \psi), \tag{2.8b}$$

where $D_\mu = \partial_\mu - i/2 W_\mu$. Equations (2.7) are exactly the gauged NLS model (or the Heisenberg model in the tangent space form), while (2.8) are the defining relations for B (the trace of higher dimension) in terms of the charge conservation law. However, Eq. (2.8a)—the reduced Chern–Simons Gauss law—is more fundamental than (2.8b). Indeed, Eqs. (2.7) imply the charge conservation law

$$\partial_0 |\psi|^2 + i \partial_1 (\overline{D_1 \psi}\, \psi - \overline{\psi} D_1 \psi) = 0. \tag{2.9}$$

Then, Eq. (2.8b) arises from the Gauss law (2.8a) and (2.9) in a similar way to the Chern–Simons theory.[1] In other words, Eqs. (2.8) guarantee the existence of the first conservation law for (2.7). In fact, the dual current $C_\mu = \epsilon_{\mu\nu}\, \partial_\nu B$, due to (2.8), is conserved.

The above systems (2.7) and (2.8) are invariant under $U(1)$ gauge transformations,

$$\psi \to \psi e^{i\alpha}, \quad A_0 \to A_0 + 2\, \partial_0 \alpha, \quad A_1 \to A_1 + 2\, \partial_1 \alpha, \tag{2.10}$$

preserving the B field: $B \to B$. From the Lagrangian (2.5) we recognize that the B field plays the role of the Lagrange multiplier for the zero strength constraint (2.7b). Moreover, according to the one-dimensional Chern–Simons (or better to say the BF) Gauss law (2.8a), creation of a particle at a point on the line, where $|\psi|^2 \neq 0$, is accompanied with creation of the local B field gradient. This gradient represents a one-dimensional analog of the statistical magnetic field. For vanishing ψ, the B field is homogeneous, and defined up to the additive constant. Integrating (2.8a) along the x_1 line, then we obtain

$$B(+\infty)-B(-\infty)=-\frac{1}{k}\int_{-\infty}^{+\infty}|\psi|^2\,dx_1. \tag{2.11}$$

For nontrivial charged configurations (solitons) the right-hand side (rhs) does not vanish and the related B field takes the shift on the boundaries.

Actually, Eqs. (2.7) are the $1+1$-dimensional analog of the Ginzburg–Landau equations for a superconductor in the vanishing electric field, $E=\partial_0 A_1-\partial_1 A_0=0$. An important consequence of the gauge invariance (2.10) is that (2.7a) is independent of the local gauge transformations parameter $\alpha=\alpha(x,t)$. This property can be considered as a generalization of the well-known isospectrality condition for integrable systems. Indeed, the second equation (2.7) allows one to exclude the potentials A_0 and A_1 by $U(1)$ gauge transformation. Due to (2.7b) a real function $\phi(x,t)$ exists such that $A_\mu=2\,\partial_\mu\phi$. Then, we define new $\Psi=\psi e^{i\phi}$, subject to the Nonlinear Schrödinger equation,

$$i\,\partial_0\Psi+\partial_1^2\Psi+2g|\Psi|^2\Psi=0. \tag{2.12}$$

This equation admits an infinite number of conservation laws. The conservation law (2.9) in terms of Ψ is the first member of this hierarchy, having a physical meaning of the charge (number of particles) conservation.

In a traditional approach the NLS integrability follows from the Lax pair or the Zero Curvature representation with a constant spectral parameter, of which the NLS is independent (isospectrality).[9] This fact can be explained now as the consequence of a $U(1)$ gauge invariance of the gauged NLS (2.7).

The integrability of (2.7) becomes transparent if we represent the self-interaction term in a pure geometrical way. Thus, for redefined fields,

$$V_0=W_0+4g|\psi|^2,\quad V_1=W_1, \tag{2.13}$$

Eqs. (2.7) have the form

$$i\,D_0\psi+D_1^2\psi=0, \tag{2.14a}$$

$$\partial_0 V_1-\partial_1 V_0=-4g\partial_1|\psi|^2, \tag{2.14b}$$

of the linear Schrödinger equation (the quantum mechanics) in the electric field proportional to the gradient of the local particles density. The system (2.14) is equivalent to the following one:

$$D_0\psi=D_1\psi_0, \tag{2.15a}$$

$$[D_0,D_1]=2g(\overline{\psi}\psi_0-\overline{\psi_0}\psi), \tag{2.15b}$$

with the constraint,

$$\psi_0=i\,D_1\psi. \tag{2.16}$$

The above equations (2.15) define the constant curvature surface with the scalar curvature equal to coupling constant g. In fact, we can show that if ψ and ψ_0 are the zweibein fields, then the first equation (2.15a) is the torsionless condition for definition of the spin connection, identified with the Abelian gauge field V_μ. Then, the the second equation is just the constant curvature condition,

$$R=g, \tag{2.17}$$

where R is the scalar curvature, written in terms of zweibeins and the spin connection. The model (2.17) is known as the Jackiw–Teitelboim lineal gravity,[10] while (2.15) as the BF gauge theoretical formulation of it in terms of the Einstein–Cartan variables.[11] The system (2.15) appears as the Euler–Lagrange equations for the BF action (4.1), discussed in Sec. IV. Depending on the sign of g the corresponding BF theory has non-Abelian SU(2), $(g=1)$, or SU(1,1), $(g=-1)$, gauge group. Thus, the reduced Jackiw–Pi model (2.14) is defined by gauge condition (2.16) of Euclidean BF gravity characterizing a surface of the Jackiw–Teitelboim model.

The mapping (2.15), (2.16) allows us to construct an infinite hierarchy of gauge constraints, compatible with (2.16), and describing the hierarchy of corresponding equations of motion. Equation (2.15b) can be written in the form

$$\partial_0 V_1-\partial_1\left(V_0+4ig\int^x(\overline{\psi}\psi_0-\overline{\psi_0}\psi)(x')dx'\right)=0. \tag{2.18a}$$

Then, as in the above procedure from (2.7) to (2.14), but going in the opposite direction, we introduce the flat Abelian gauge field,

$$W_0=V_0+4ig\int^x(\overline{\psi}\psi_0-\overline{\psi_0}\psi)(x')dx',\quad W_1=V_1. \tag{2.18b}$$

Then, in terms of redefined covariant derivatives, (2.15) becomes

$$D_0\psi=D_1\psi_0+2g\psi\int^x(\overline{\psi}\psi_0-\overline{\psi_0}\psi)(x')dx', \tag{2.19a}$$

$$\partial_0 W_1-\partial_1 W_0=0. \tag{2.19b}$$

Of course, for nonsingular gauge configurations we can always exclude field W_μ. But, we find it is important to keep this gauge field, playing the role similar to the spectral parameter in the usual Inverse Scattering approach. Then, model (2.12) written in terms of Ψ, is explicitly invariant under the local gauge transformations (2.10) and the gauge invariance plays the role of generalized isospectrality. Below we show how far this analogy can be continued. It is convenient to write Eq. (2.19a) and its complex conjugate in the matrix form

$$\begin{pmatrix} D_0\psi \\ \overline{D_0\psi}\end{pmatrix}=\begin{pmatrix} D_1+2g\psi\int^x\overline{\psi} & -2g\psi\int^x\psi \\ -2g\overline{\psi}\int^x\overline{\psi} & \overline{D_1}+2g\overline{\psi}\int^x\psi\end{pmatrix}\begin{pmatrix}\psi_0\\ \overline{\psi_0}\end{pmatrix}, \tag{2.20}$$

Then, the constraint (2.16) can be written as

$$\begin{pmatrix}\psi_0^{(1)}\\ \overline{\psi_0^{(1)}}\end{pmatrix}=i\sigma_3\begin{pmatrix} D_1\psi \\ \overline{D_1\psi}\end{pmatrix}=\Lambda\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}, \tag{2.21}$$

where we introduce the integrodifferential operator,

$$\Lambda=i\sigma_3\begin{pmatrix} D_1+2g\psi\int^x\overline{\psi} & -2g\psi\int^x\psi \\ -2g\overline{\psi}\int^x\overline{\psi} & \overline{D_1}+2g\overline{\psi}\int^x\psi\end{pmatrix}. \tag{2.22a}$$

For a skew-symmetric linear operator (see Sec. IV) we have to consider the integral part defined by the symmetric boundaries,

$$\int^x f(y)dy=\frac{1}{2}\left(\int_{-\infty}^x f(y)dy+\int_\infty^x f(y)dy\right). \tag{2.22b}$$

The above operator Λ represents the $U(1)$ gauge covariant extension of the AKNS operator $\mathscr{L}$.[12] In fact,

$$\Lambda=\mathscr{L}+\tfrac{1}{2}W_1 I, \tag{2.23a}$$

where

$$\mathscr{L}=i\sigma_3\begin{pmatrix}\partial_1+2g\psi\int^x\overline{\psi} & -2g\psi\int^x\psi\\ -2g\overline{\psi}\overline{\int^x\psi} & \partial_1+2g\overline{\psi}\overline{\int^x\psi}\end{pmatrix}. \tag{2.23b}$$

Here I is the identity matrix and Λ is covariant under $U(1)$ gauge transformations (2.10). Then, gauged NLS (2.7) has the form

$$i\sigma_3\begin{pmatrix}D_0\psi\\ \overline{D_0\psi}\end{pmatrix}=\Lambda^2\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}. \tag{2.24}$$

This representation suggests how to create the whole hierarchy of equations associated with gauged NLS (2.24). We define the set of constraints,

$$\begin{pmatrix}\psi_0^{(n)}\\ \overline{\psi_0^{(n)}}\end{pmatrix}=\Lambda^n\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}, \tag{2.25}$$

labeled with an integer n. As mentioned above, (2.15) describes a classical motion for the BF theory, with ψ_0 playing the role of Lagrange multipliers. The arbitrariness of ψ_0 guarantees the general time reparametrization invariance of the theory, while a specific choice for ψ_0 defines the corresponding evolution. The hierarchy (2.25) can be considered as the hierarchy of gauge fixing constraints of BF theory. Then, every constraint is related to the nonlinear spin model (Heisenberg hierarchy), being just the tangent space representation for the model.

The first few constraints from (2.25) are given by

$$\begin{pmatrix}\psi_0^{(0)}\\ \overline{\psi_0^{(0)}}\end{pmatrix}=\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}, \tag{2.26}$$

$$\begin{pmatrix}\psi_0^{(1)}\\ \overline{\psi_0^{(1)}}\end{pmatrix}=i\sigma_3\begin{pmatrix}D_1\psi\\ \overline{D_1\psi}\end{pmatrix}, \tag{2.27}$$

$$\begin{pmatrix}\psi_0^{(2)}\\ \overline{\psi_0^{(2)}}\end{pmatrix}=-\begin{pmatrix}D_1^2\psi+2g|\psi|^2\psi\\ \overline{D_1^2\psi+2g|\psi|^2\psi}\end{pmatrix}, \tag{2.28}$$

$$\begin{pmatrix}\psi_0^{(3)}\\ \overline{\psi_0^{(3)}}\end{pmatrix}=-i\sigma_3\begin{pmatrix}D_1^3\psi+6g|\psi|^2D_1\psi\\ \overline{D_1^3\psi+6g|\psi|^2D_1\psi}\end{pmatrix}. \tag{2.29}$$

Thus, for every n, the corresponding $\psi_0^{(n)}$ defines the evolution equation,

$$i\sigma_3\begin{pmatrix} D_{0_n}\psi \\ \overline{D_{0_n}\psi}\end{pmatrix}=\Lambda^n\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}=\begin{pmatrix}\psi_0^{(n)}\\ \overline{\psi_0^{(n)}}\end{pmatrix}, \tag{2.30a}$$

$$[D_{0_n},D_1]=0. \tag{2.30b}$$

The first members of the hierarchy, $(n=1,2,3)$ are

$$D_{0_1}\psi-D_1\psi=0, \tag{2.31}$$

$$i\ D_{0_2}\psi+D_1^2\psi+2g|\psi|^2\psi=0, \tag{2.32}$$

$$D_{0_3}\psi+D_1^3\psi+6g|\psi|^2\ D_1\psi=0. \tag{2.33}$$

This $U(1)$ gauged hierarchy has the flat Abelian connection (2.30b). It means that every equation (2.30) in terms of the gauge-invariant variables, like (2.12), reduces to the form of the usual NLS hierarchy. Indeed, due to (2.30b) for any n there exists a real function $\alpha_n=\alpha_n(x,t_n)$, such that

$$W_1=\partial_1\alpha_n, \quad W_{0_n}=\partial_{0_n}\alpha_n. \tag{2.34}$$

Therefore, for the gauge-invariant fields $\Psi=\psi e^{i\alpha_n}$, the hierarchy (2.30) reduces to the NLS one,

$$i\sigma_3\begin{pmatrix}\partial_{0_n}\Psi\\ \overline{\partial_{0_n}\Psi}\end{pmatrix}=\mathscr{L}^n\begin{pmatrix}\Psi\\ \overline{\Psi}\end{pmatrix}=\begin{pmatrix}\Psi_0^{(n)}\\ \overline{\Psi_0^{(n)}}\end{pmatrix}. \tag{2.35}$$

However, the gauged hierarchy (2.30) contains much more information in the addition. As we show in the next section, restricted on the subclass of constant gauge potentials, Equations (2.30) provide the linear problem for any of Eqs. (2.35). Moreover, every equation of the hierarchy (2.30) can be reduced dimensionally from the related $2+1$-dimensional Chern–Simons gauged theory.

III. ABELIAN GAUGE THEORY FOR MOVING FRAMES HIERARCHY

The hierarchy of Abelian gauge theories described in the previous section by (2.30), is equivalent to the system (2.15) with the hierarchy of $U(1)$ gauge-invariant constraints (2.25). But system (2.15) is the constant curvature surface equation written in the Einstein–Cartan zweibein formalism. As we see below, every constraint (2.25) supplied to (2.15), defines an evolution equation for three-dimensional unit vector **s**, and can be considered as a nonlinear σ model on the sphere S^2 or pseudosphere $S^{1,1}$.

Let us consider the Lie group G with element g, generated by τ_i $(i=1,2,3)$, satisfying the relations

$$\tau_i\tau_j=h_{ij}+ic_{ijk}\tau_k, \tag{3.1}$$

where h_{ij} and c_{ijk} are the Killing metric and the structure constants of the algebra, respectively. We define an orthonormal trihedral set of unit vectors $\mathbf{n}_i$ in the adjoint representation,

$$(\mathbf{n}_i,\tau)=\mathbf{n}_i^k\tau_k=h_{kl}\mathbf{n}_i^k\tau^l=g\tau_ig^{-1}. \tag{3.2}$$

The Killing metric h_{ij} and structure constants $c_{ijk}=-c_{jik}$ define correspondingly the inner and the cross product between three vectors, transforming in the adjoint representation:

$$(\mathbf{n}_i,\mathbf{n}_j)=h_{ij}, \tag{3.3a}$$

$$\mathbf{n}_i\wedge\mathbf{n}_j=c_{ijk}\mathbf{n}_k. \tag{3.3b}$$

Given smooth vector fields $\mathbf{n}_i=\mathbf{n}_i(x,t)$ define at each space x and time t three vectors $(\mathbf{n}_1(x,t),\mathbf{n}_2(x,t),\mathbf{n}_3(x,t))$ forming an orthonormal basis, called the moving frame. The chiral current,

$$J_\mu=g^{-1}\,\partial_\mu g, \tag{3.4}$$

in the adjoint representation defines rotation of the moving frame by equations,

$$\partial_\mu\mathbf{n}_i=(J_\mu^{\mathrm{ad}})_{ik}\mathbf{n}_k, \tag{3.5}$$

where

$$(J_\mu^{\mathrm{ad}})_{ik}=-ic_{ijk}(J_\mu)_j=i(J_\mu)_jc_{jik}, \tag{3.6}$$

and $J_\mu=\Sigma(J_\mu)_j(1/2)\tau_j$. Matrices J_μ^{ad} have the symmetry, $(J_\mu^{\mathrm{ad}})_{ij}h_{jj}=-(J_\mu^{\mathrm{ad}})_{ji}h_{ii}$, and are antisymmetric for the SU(2) case $h_{ij}=\delta_{ij}$, $c_{ijk}=\epsilon_{ijk}$.

The current (3.4) satisfies the zero curvature equations,

$$\partial_\mu J_\nu-\partial_\nu J_\mu+[J_\mu,J_\nu]=0, \tag{3.7}$$

as compatibility conditions for the system (3.5). We decompose matrix J_μ to the diagonal and off diagonal parts, $J_\mu=J_\mu^{(0)}+J_\mu^{(1)}$, parametrized in the form[13]

$$J_\mu^{(0)}=i/4\sigma_3V_\mu,\quad J_\mu^{(1)}=\begin{pmatrix}0 & -g\overline{\psi_\mu}\\ \psi_\mu & 0\end{pmatrix}, \tag{3.8}$$

where $g=+1$ for $su(2)$, and $g=-1$, for $su(1,1)$ cases. Then, in the adjoint representation (3.6) we have the form,

$$(J_\mu)^{\mathrm{ad}}=\frac{1}{2}\begin{pmatrix}0 & V_\mu & 4g\Re(\psi_\mu)\\ -V_\mu & 0 & 4g\Im(\psi_\mu)\\ -4\Re(\psi_\mu) & -4\Im(\psi_\mu) & 0\end{pmatrix}. \tag{3.9}$$

The moving frame rotates according to equations

$$\partial_\mu\mathbf{n}_i=-1/2V_\mu\epsilon_{ij}\mathbf{n}_j-gU_{i\mu}\mathbf{s}, \tag{3.10a}$$

$$\partial_\mu\mathbf{s}=U_{i\mu}\mathbf{n}_i, \tag{3.10b}$$

where $U_\mu\equiv2(\Re(\psi_\mu),\Im(\psi_\mu))$. For vector $\mathbf{s}\equiv\mathbf{n}_3$ the constraint, $(\mathbf{s}(x,t),\mathbf{s}(x,t))=h_{33}$, is valid, where $h_{33}=1$, which means that it belongs to the two-dimensional sphere S^2 or pseudosphere $S^{1,1}$, correspondingly. Two vector fields $(\mathbf{n}_1(x),\mathbf{n}_2(x))$ at each (x,t) form a basis in the tangent plane to the corresponding manifold for $\mathbf{s}(x)$. However, by Eq. (3.10) vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ are not determined uniquely. If we choose the other pair $\mathbf{n}_1'$, $\mathbf{n}_2'$, as the rotated basis,

$$\mathbf{n}_1'=\cos\,\alpha\mathbf{n}_1-\sin\,\alpha\mathbf{n}_2,\quad \mathbf{n}_2'=\cos\,\alpha\mathbf{n}_2+\sin\,\alpha\mathbf{n}_1, \tag{3.11}$$

the related V_μ' and ψ_μ' are the $U(1)$ gauge transformed fields,

$$V'_\mu = V_\mu + 2\,\partial_\mu \alpha, \quad \psi'_\mu = e^{i\alpha}\psi_\mu . \tag{3.12a}$$

The expression (3.11) suggests that we introduce a complex basis $\mathbf{n}_+ = \mathbf{n}_1 + i\mathbf{n}_2$, $\mathbf{n}_- = \mathbf{n}_1 - i\mathbf{n}_2$, satisfying the following relations:

$$(\mathbf{n}_\pm, \mathbf{n}_\pm) = 0, \quad (\mathbf{n}_+, \mathbf{n}_-) = 2g, \tag{3.12b}$$

$$\mathbf{n}_+ \times \mathbf{s} = i\mathbf{n}_+, \mathbf{n}_- \times \mathbf{s} = -i\mathbf{n}_-, \quad \mathbf{n}_- \times \mathbf{n}_+ = 2ig\mathbf{s}. \tag{3.12c}$$

Then, we get

$$\psi_\mu = 1/2g(\partial_\mu \mathbf{s}, \mathbf{n}_+), \quad \overline{\psi_\mu} = 1/2g(\partial_\mu \mathbf{s}, \mathbf{n}_-). \tag{3.13a}$$

In terms of (3.12) the moving frame system (3.10) becomes

$$D_\mu \mathbf{n}_+ = -2g\psi_\mu \mathbf{s}, \tag{3.13b}$$

$$\partial_\mu \mathbf{s} = \psi_\mu \mathbf{n}_- + \overline{\psi_\mu}\mathbf{n}_+ , \tag{3.13c}$$

where $D_\mu \equiv \partial_\mu - i/2V_\mu$, is a $U(1)$ covariant derivative. This form is explicitly invariant under the local $U(1)$ gauge transformations,

$$\mathbf{n}_+ \to e^{i\alpha}\mathbf{n}_+ , \quad \mathbf{n}_- \to e^{-i\alpha}\mathbf{n}_- , \quad V_\mu \to V_\mu + 2\,\partial_\mu \alpha, \quad \mathbf{s} \to \mathbf{s}, \tag{3.14}$$

that are just the local rotations in tangent to the vector $\mathbf{s}$ plane. From (3.13) it follows that V_μ and ψ_μ fields subject to the system (the integrability condition),

$$D_0\psi = D_1\psi_0, \tag{3.15a}$$

$$[D_0, D_1] = 2g(\overline{\psi}\psi_0 - \overline{\psi_0}\psi), \tag{3.15b}$$

where we skip the index for the ψ_1 field. This system coincides with (2.15). Moreover, the reduced Jackiw–Pi model (2.7) corresponds to the constraint (2.16). Under this constraint, the moving frame evolution (3.13) describes

$$D_0 \mathbf{n}_+ = -2gi\, D_1\psi \mathbf{s}, \tag{3.16a}$$

$$D_1 \mathbf{n}_+ = -2g\psi \mathbf{s}, \tag{3.16b}$$

$$\partial_0 \mathbf{s} = i\, D_1\psi\, \mathbf{n}_- - i\, \overline{D_1}\,\overline{\psi}\, \mathbf{n}_+ , \tag{3.16c}$$

$$\partial_1 \mathbf{s} = \psi \mathbf{n}_- + \overline{\psi}\mathbf{n}_+ . \tag{3.16d}$$

Differentiating (3.16d),

$$\partial_1^2 \mathbf{s} = D_1\psi\, \mathbf{n}_- + \overline{D_1}\,\overline{\psi}\, \mathbf{n}_+ - 4g|q_1|^2 \mathbf{s}, \tag{3.17}$$

we immediately see that vector $\mathbf{s}$ satisfies the Landau–Lifshitz equation for the continuum isotropical Heisenberg spin chain model,

$$\partial_0 \mathbf{s} = \mathbf{s} \times \partial_1^2 \mathbf{s}, \tag{3.18}$$

where **s** belongs to the two-dimensional sphere S^2 $(g=1)$,[14] or pseudosphere $S^{1,1}$ $(g=-1)$.[15] The above results illuminate the role of the Chern–Simons statistical gauge field in the Heisenberg model. Two components, A_0, A_1, are just the gauge degrees of freedom of rotation in the tangent plane, while the $B=A_2$ component is defined by relations

$$\partial_1 B=-\frac{1}{4gk}\,(\partial_1\mathbf{s})^2, \tag{3.19a}$$

$$\partial_0 B=\frac{1}{k}\,\partial_1\mathbf{s}(\partial_1^2\mathbf{s}\wedge\mathbf{s}). \tag{3.19b}$$

The right-hand side of Eq. (3.19a) is the energy density of the model (3.18). As in Sec. II, [see Eq. (2.8a)], the local magnetic energy is always accompanied with the gradient of the B field. From Eqs. (3.19) it is evident that the total magnetic energy is a conserved quantity, defining the asymptotic jump,

$$B(+\infty)-B(-\infty)=-\frac{1}{4gk}\int_{-\infty}^{+\infty}(\partial_1\mathbf{s})^2\,dx. \tag{3.20}$$

As shown in Sec. II, the gauge constraint (2.16) is the second member of the infinite hierarchy of the gauge constraints (2.25). To show that every constraint from this hierarchy corresponds to the moving frame system and the spin model, we can proceed analogously to the case (2.18)–(2.19). After redefinition (2.18) the system (3.13) becomes

$$\begin{pmatrix} D_0\mathbf{n}_+ \\ D_0\mathbf{n}_- \end{pmatrix}=-2gs\begin{pmatrix} \psi_0 \\ \overline{\psi_0} \end{pmatrix}+2g\begin{pmatrix} \mathbf{n}_+ \\ -\mathbf{n}_- \end{pmatrix}\int^x(\overline{\psi}\psi_0-\overline{\psi_0}\psi), \tag{3.21}$$

$$\begin{pmatrix} D_1\mathbf{n}_+ \\ D_1\mathbf{n}_- \end{pmatrix}=-2gs\begin{pmatrix} \psi \\ \overline{\psi} \end{pmatrix}, \tag{3.22}$$

$$\partial_0\mathbf{s}=\psi_0\mathbf{n}_-+\overline{\psi_0}\mathbf{n}_+\,,\quad \partial_1\mathbf{s}=\psi\mathbf{n}_-+\overline{\psi}\mathbf{n}_+\,. \tag{3.23}$$

The tangent space vectors evolution (3.21) is convenient to be rewritten as

$$\begin{pmatrix} D_0\mathbf{n}_+ \\ D_0\mathbf{n}_- \end{pmatrix}=-2g\begin{pmatrix} \mathbf{s}-\mathbf{n}_+\int^x\overline{\psi} & \mathbf{n}_+\int^x\psi \\ \mathbf{n}_-\int^x\overline{\psi} & \mathbf{s}-\mathbf{n}_-\int^x\psi \end{pmatrix}\begin{pmatrix} \psi_0 \\ \overline{\psi_0} \end{pmatrix}. \tag{3.24}$$

Then, we note that compatibility condition for the system (3.22)–(3.24) is given by (2.20). Moreover, the hierarchy of the gauge fixing constraints (2.25) produces the hierarchy of the moving frame evolutions (3.24) and the related higher-order analogs of the NLS (2.30). This moving frame hierarchy generates also the higher-order analogs of the Heisenberg model (3.18). To follow this direction we need to extract information only in terms of the spin vector **s**. First we note, that the integrand of (3.21),

$$2ig(\overline{\psi}\psi_0-\overline{\psi_0}\psi)=\mathbf{s}(\partial_0\mathbf{s}\wedge\partial_1\mathbf{s}), \tag{3.25}$$

is the topological charge density for the spin configuration on the space–time world sheet, or the volume element on the spin phase space. Then, we obtain the recursion relations for the evolution equations,

$$\partial_{0_n}\mathbf{s}=\left[\mathbf{s}\wedge\partial_1-\partial_1\mathbf{s}\int^x\mathbf{s}(\partial_1\mathbf{s}\wedge\cdot)\right]\partial_{0_{n-1}}\mathbf{s}, \tag{3.26}$$

where the integer n describes the nth member of the hierarchy. The first few evolutions are given by

$$\partial_0\mathbf{s}=\partial_1\mathbf{s}, \tag{3.27}$$

$$\partial_0\mathbf{s}=\mathbf{s}\wedge\partial_1^2\mathbf{s}, \tag{3.28}$$

$$\partial_0\mathbf{s}=\mathbf{s}\wedge\mathbf{s}\wedge\partial_1^3\mathbf{s}-\tfrac{3}{2}(\partial_1\mathbf{s}\ \partial_1\mathbf{s})\partial_1\mathbf{s}. \tag{3.29}$$

The moving frame approach allows us to introduce also the traditional zero curvature description. To proceed, we need to find an appropriate reduction for (3.7)–(3.8). We note that the NLS (2.12) can be considered as a quantum mechanic, with the potential function proportional to the probability density, $U(x)=\Psi(x)\Psi(x)$. But, as is well known,[16] the Galilean boosts representation in the quantum mechanics makes use of 1-cocycle ω_1:

$$U(v)\Psi(x)=e^{-2\pi i\omega_1(x;v)}\Psi(x-vt), \tag{3.30}$$

with the following condition to satisfy:

$$\omega_1(x-ut;v)-\omega_1(x;u+v)+\omega_1(x;u)=0,(\mathrm{mod}\ Z), \tag{3.31}$$

where $U(v)$ is the unitary operator shifting the coordinates as

$$U(v)xU^{-1}(v)=x-vt. \tag{3.32}$$

Then, condition (3.31) ensures the group multiplication for $U(v)$ operators:

$$U(v)U(u)=U(u+v). \tag{3.33}$$

The quantity $\omega_1(x;v)$ is given by

$$2\pi\omega_1(x;v)=-\frac{v^2}{4}t+\frac{v}{2}x, \tag{3.34}$$

and provides the local phase transformation for the wave function. Thus, the NLS model (2.12) is invariant under the Galileo transformations,

$$x\to x'=x-vt,\quad t\to t'=t, \tag{3.35a}$$

$$\Psi\to\Psi'\ \exp\ i\left(\frac{v^2}{4}t'+\frac{v}{2}x'\right). \tag{3.35b}$$

This invariance can be interpreted in two ways. First, that the space–time transformation to a constant velocity moving frame should be accompanied with the phase transformation for the ''wave function'' Ψ. This is the 1-cocycle representation, as given above. In the second interpretation we can say that any one-parameter group ($v=\mathrm{const}$) of the phase transformations (3.35b) becomes the symmetry of the model (2.12) if we transform coordinates to the moving frame with velocity v according to (3.35a). The second interpretation has the flavor of the $U(1)$ Abelian local gauge theory with the gauge potentials transforming on the constant value. In fact, the linear in space and time phase transformations (3.35b) satisfying (3.31) generate a subgroup of the Abelian

gauge transformations, shifting the W_1 potential (2.4) on the constant velocity. This is why it is of interest to consider the restricted class of the constant gauge potentials. As shown below they are associated with the spectral parameter in the linear problem for the NLS. Therefore, in our approach the whole structure of the linear problem with the spectral parameter has the origin from the $U(1)$ gauge potential W_1, while the related phase is the linear function of space and time.

Our idea has been suggested by some known facts from the theory of superconductivity[17] and the field theory with curved momentum space.[18] For the superconductor in the ground state under the external electromagnetic action the generalized canonical momentum $\mathbf{p}=m\mathbf{v}+e/c\mathbf{A}$ vanishes. This fact provides the relation between superconducting electrons velocity and the vector potential,

$$\mathbf{v}=-e/c\mathbf{A}, \tag{3.36}$$

restricting Abelian gauge transformations to the London gauge,[17]

$$\text{div } \mathbf{A}=0. \tag{3.37}$$

In one space dimension this gauge leads to the constant gauge potential A. Furthermore, for constant potentials A_μ, transition to the new momentum,

$$p_\mu \rightarrow p_\mu - c_\mu, \tag{3.38}$$

generates the group of parallel transitions in the momentum space. For the nonrelativistic theory it is just the Galileo transformation. Finally, as was shown recently,[13] the spectral parameter arises as the constant-valued statistical gauge field of compactified $2+1$-dimensional theory. All results mentioned above suggest that the spectral parameter is deeply connected with the Abelian gauge field.

Let us consider the moving frame equations (3.16),

$$D_0\mathbf{n}_+=-2gi\ D_1\psi\ \mathbf{s}+2ig|\psi|^2\mathbf{n}_+, \tag{3.39a}$$

$$D_1\mathbf{n}_+=-2g\psi\mathbf{s}, \tag{3.39b}$$

$$\partial_0\mathbf{s}=i\ D_1\psi\ \mathbf{n}_- - i\ \overline{D_1}\overline{\psi}\mathbf{n}_+, \tag{3.39c}$$

$$\partial_1\mathbf{s}=\psi\mathbf{n}_- + \overline{\psi}\mathbf{n}_+, \tag{3.39d}$$

written in terms of the zero strength gauge potential W_μ (2.13),

$$\partial_0 W_1 - \partial_1 W_0 = 0. \tag{3.39e}$$

The general solution of the last equation is of the pure gauge form given in terms of the real function,

$$W_\mu=\partial_\mu\alpha. \tag{3.40}$$

For a restricted class of functions, linear in x and t, we put

$$W_1=2\lambda=\text{const}. \tag{3.41}$$

In this case the NLS (2.7a) becomes

$$i\ \partial_0\psi-2i\lambda\ \partial_1\psi+\partial_1^2\psi+2g|\psi|^2\psi=0, \tag{3.42}$$

where we have to choose $W_0=1/2W_1^2=2\lambda^2$. In the constant velocity frame, $v=-2\lambda$, this equation acquires the normal form (2.12)

$$i\ \partial_{0'}\psi+\partial_{1'}^2\psi+2g|\psi|^2\psi=0, \tag{3.43}$$

and any λ dependence disappears. However, it remains in the corresponding moving frame system (3.39) having the form

$$\partial_{1'}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}=\begin{pmatrix} i\lambda & 0 & -2g\psi\\ 0 & -i\lambda & -2g\overline{\psi}\\ \overline{\psi} & \psi & 0\end{pmatrix}.\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}, \tag{3.44a}$$

$$\partial_{0'}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}=\begin{pmatrix} -i\lambda^2+2ig|\psi|^2 & 0 & -2ig\ \partial_{1'}\psi+2g\lambda\psi\\ 0 & i\lambda^2-2ig|\psi|^2 & 2ig\ \partial_{1'}\overline{\psi}+2g\lambda\overline{\psi}\\ -(i\partial_{1'}\overline{\psi}+\lambda\overline{\psi}) & i\ \partial_{1'}\psi-\lambda\psi & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}. \tag{3.44b}$$

In the normal basis $\mathbf{n}_1,\mathbf{n}_2,\mathbf{s}$, we have equations

$$\partial_{1'}\begin{pmatrix}\mathbf{n}_1\\ \mathbf{n}_2\\ \mathbf{s}\end{pmatrix}=\begin{pmatrix} 0 & -\lambda & -gU_1\\ \lambda & 0 & -gU_2\\ U_1 & U_2 & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}_1\\ \mathbf{n}_2\\ \mathbf{s}\end{pmatrix} \tag{3.45a}$$

$$\partial_{0'}\begin{pmatrix}\mathbf{n}_1\\ \mathbf{n}_2\\ \mathbf{s}\end{pmatrix}=\begin{pmatrix} 0 & \lambda^2-2g|\psi|^2 & g\ \partial_{1'}U_2+g\lambda U_1\\ -\lambda^2+2g|\psi|^2 & 0 & -g\ \partial_{1'}U_1+g\lambda U_2\\ -\partial_{1'}U_2-\lambda U_1 & \partial_{1'}U_1-\lambda U_2 & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}_1\\ \mathbf{n}_2\\ \mathbf{s}\end{pmatrix}, \tag{3.45b}$$

where we defined [see Eqs. (3.10)] $U_1=2\Re(\psi)$, $U_2=2\Im(\psi)$. The first system (3.45a) in terms of the gauge-invariant fields as in (2.12) ($\lambda=0$), has the form of the Frenet equations for the curve. This suggests the idea to interpret the gauge-invariant form of (3.45) as a moving curve equation.[19] But due to the essential role of the space–time curvature that is proportional to the nonlinearity and can be negative, this analogy does not seem so useful. In opposite, we find that a more general setting—the moving frame method is going deeply in to the problem and naturally illuminates the gauge theoretical structure of it.

Note that the potentials choice,

$$W_1=2\lambda,\quad W_0=2\lambda^2, \tag{3.46}$$

which allowed us to remove the λ dependence from equations of motion (3.42), is exactly the 1-cocycle (3.35b). To reproduce the usual Lax-type representation we use the spinor representation of system (3.45), and according to (3.6) we have 2×2 Zakharov–Shabat spectral problem,

$$J_1=+\frac{i}{2}\lambda\sigma_3+\begin{pmatrix}0 & -g\overline{\psi}\\ \psi & 0\end{pmatrix}, \tag{3.47a}$$

$$J_0=i\sigma_3\left[-\frac{\lambda^2}{2}+g|\psi|^2\right]+\begin{pmatrix}0 & g(i\partial_1+\lambda)\overline{\psi}\\ (i\partial_1-\lambda)\psi & 0\end{pmatrix}, \tag{3.47b}$$

for the NLS (3.43).

To derive the hierarchy of moving frames, providing the linear problem for any of equations (2.35), we need to modify the above Galileo transformation approach (3.42)–(3.43). Let us first rewrite the hierarchy (2.30) for the constant potentials W_μ, denoting $W_1=2\lambda$:

$$
\begin{aligned}
i\partial_{0_n}\begin{pmatrix}\psi\\ -\bar{\psi}\end{pmatrix}+\frac{1}{2}\,W_{0_n}\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}&=(\mathscr{L}+\lambda I)^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}\\
&=\sum_{k=0}^{n}\binom{k}{n}\lambda^{n-k}\mathscr{L}^k\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}\\
&=\sum_{k=0}^{n}\binom{k}{n}\lambda^{n-k}\begin{pmatrix}\psi_0^{(k)}\\ \bar{\psi}_0^{(k)}\end{pmatrix}\\
&=\lambda^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}+\sum_{k=1}^{n-1}\binom{k}{n}\lambda^{n-k}\begin{pmatrix}\psi_0^{(k)}\\ \bar{\psi}_0^{(k)}\end{pmatrix}+\begin{pmatrix}\psi_0^{(n)}\\ \bar{\psi}_0^{(n)}\end{pmatrix}\\
&=\lambda^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}+\sum_{k=1}^{n-1}\binom{k}{n}\lambda^{n-k}i\sigma_3\partial_{0_k}\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}+\mathscr{L}^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix},
\end{aligned}
\tag{3.48}
$$

where $[\mathscr{L},\lambda I]=0$ is used. Choosing the gauge potentials $W_{0_n}=2\lambda^n$, and collecting all terms with the time hierarchy derivatives on the left-hand side, we get

$$
i\left\{\partial_{0_n}-\sum_{k=1}^{n-1}\binom{k}{n}\lambda^{n-k}\partial_{0_k}\right\}\begin{pmatrix}\psi\\ -\bar{\psi}\end{pmatrix}=\mathscr{L}^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}. \tag{3.49}
$$

Defining the new time hierarchy $\{t_{0_1'},t_{0_2'},\ldots,t_{0_n'},\ldots\}$, such that

$$
\partial_{0_n'}=\partial_{0_n}-\sum_{k=1}^{n-1}\binom{k}{n}\lambda^{n-k}\partial_{0_k}, \tag{3.50}
$$

we obtain for (3.49) the usual NLS hierarchy form (2.35):

$$
i\sigma_3\partial_{0_n'}\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}=\mathscr{L}^n\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix}. \tag{3.51}
$$

The matrix structure of the above time hierarchy transformation can be realized by the upper triangular matrix,

$$
\begin{pmatrix}t_1'\\ t_2'\\ t_3'\\ \vdots\\ t_n'\end{pmatrix}=\begin{pmatrix}1 & a_{12} & a_{13} & \cdots & a_{1n}\\ 0 & 1 & a_{23} & \cdots & a_{2n}\\ 0 & 0 & 1 & \cdots & a_{3n}\\ 0 & \vdots & 0 & \cdots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{pmatrix}\begin{pmatrix}t_1\\ t_2\\ t_3\\ \vdots\\ t_n\end{pmatrix}, \tag{3.52}
$$

while the time derivatives transform by the lower triangular matrix,

$$\begin{pmatrix} \partial_{0_1} \\ \partial_{0_2} \\ \partial_{0_3} \\ \vdots \\ \partial_{0_n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{12} & 1 & 0 & \cdots & 0 \\ a_{13} & a_{23} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ a_{1n} & a_{2n} & a_{3n} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \partial_{0'_1} \\ \partial_{0'_2} \\ \partial_{0'_3} \\ \vdots \\ \partial_{0'_n} \end{pmatrix}, \tag{3.53}$$

with coefficients chosen as an appropriate polynomial of the parameter λ.

By using the moving frame formulas (3.21)–(3.24) we can construct the hierarchy of moving frames, which provides also the linear problem for any member of the hierarchy (3.51). For the space part we have

$$\begin{pmatrix} D_1\mathbf{n}_+ \\ D_1\mathbf{n}_- \\ \partial_1\mathbf{s} \end{pmatrix} = \begin{pmatrix} 0 & 0 & -2g\psi \\ 0 & 0 & -2g\overline{\psi} \\ \overline{\psi} & \psi & 0 \end{pmatrix} \begin{pmatrix} \mathbf{n}_+ \\ \mathbf{n}_- \\ \mathbf{s} \end{pmatrix}, \tag{3.54}$$

in which $W_1 = 2\lambda$ gives us the space part of the moving frame system and is universal for the whole hierarchy,

$$\partial_1 \begin{pmatrix} \mathbf{n}_+ \\ \mathbf{n}_- \\ \mathbf{s} \end{pmatrix} = \begin{pmatrix} i\lambda & 0 & -2g\psi \\ 0 & -i\lambda & -2g\overline{\psi} \\ \overline{\psi} & \psi & 0 \end{pmatrix} \begin{pmatrix} \mathbf{n}_+ \\ \mathbf{n}_- \\ \mathbf{s} \end{pmatrix}. \tag{3.55}$$

For the time evolution we start from the system for the t_n time:

$$\begin{pmatrix} D_{0_n}\mathbf{n}_+ \\ D_{0_n}\mathbf{n}_- \\ \partial_{0_n}\mathbf{s} \end{pmatrix} = -2g \begin{pmatrix} \mathbf{s}-\mathbf{n}_+\int^x \overline{\psi} & \mathbf{n}_+\int^x \psi \\ s-\mathbf{n}_-\int^x \psi & \mathbf{s}-\mathbf{n}_-\int^x \psi \\ -\dfrac{1}{2g}\,\mathbf{n}_- & -\dfrac{1}{2g}\,\mathbf{n}_+ \end{pmatrix} \begin{pmatrix} \psi_0^{(n-1)} \\ \overline{\psi_0^{(n-1)}} \end{pmatrix}. \tag{3.56}$$

After substituting the value $W_{0_n} = 2\lambda^n$ and (2.25), we have the expression

$$\partial_{0_n} \begin{pmatrix} \mathbf{n}_+ \\ \mathbf{n}_- \\ \mathbf{s} \end{pmatrix} = i\lambda^n \begin{pmatrix} \mathbf{n}_+ \\ -\mathbf{n}_- \\ 0 \end{pmatrix} - 2g \begin{pmatrix} \mathbf{s}-\mathbf{n}_+\int^x \overline{\psi} & \mathbf{n}_+\int^x \psi \\ \mathbf{n}_-\int^x \overline{\psi} & \mathbf{s}-\mathbf{n}_-\int^x \psi \\ -\dfrac{1}{2g}\,\mathbf{n}_- & -\dfrac{1}{2g}\,\mathbf{n}_+ \end{pmatrix} \Lambda^{n-1} \begin{pmatrix} \psi \\ \overline{\psi} \end{pmatrix}. \tag{3.57}$$

Then, for the transformed time hierarchy using (3.50) and (3.57), we get

$$\partial_{0'_n}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}=i\lambda^n(3-2^n)\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}-2g\begin{pmatrix}\mathbf{s}-\mathbf{n}_+\int^x\overline{\psi} & \mathbf{n}_+\int^x\psi\\ \mathbf{n}_-\int^x\psi & \mathbf{s}-\mathbf{n}_-\int^x\psi\\ -\frac{1}{2g}\mathbf{n}_- & -\frac{1}{2g}\mathbf{n}_+\end{pmatrix}$$
$$\times\left\{\Lambda^{n-1}-\sum_{k=1}^{n-1}\binom{k}{n}\lambda^{n-k}\Lambda^{k-1}\right\}\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}. \tag{3.58}$$

The system (3.55), (3.58) is the moving frame system for the hierarchy (3.51). The space part is linear in the spectral parameter λ, while the time part contains the nth degree polynomial for the nth member of the hierarchy. To reproduce this dependence in the explicit form, according to (2.23a), we decompose the gauge covariant operator $\Lambda=\mathscr{L}+2I$. As a result we have the moving frame evolution in terms of the usual AKNS operator $\mathscr{L}$,

$$\partial_{0'_n}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}=i\lambda^n(3-2^n)\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}\begin{pmatrix}\mathbf{n}_+\\ \mathbf{n}_-\\ \mathbf{s}\end{pmatrix}-2g\begin{pmatrix}\mathbf{s}-\mathbf{n}_+\int^x\overline{\psi} & \mathbf{n}_+\int^x\psi\\ \mathbf{n}_-\int^x\overline{\psi} & \mathbf{s}-\mathbf{n}_-\int^x\psi\\ -\frac{1}{2g}\mathbf{n}_- & -\frac{1}{2g}\mathbf{n}_+\end{pmatrix}$$
$$\times\left\{\sum_{l=0}^{n-1}\binom{l}{n-1}-\sum_{k=1}^{n-1}\sum_{l=0}^{k-1}\binom{k}{n}\binom{l}{k-1}\right\}\lambda^{n-1-l}\mathscr{L}^l\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix}. \tag{3.59}$$

For $n=2$ these moving frame system coincides with the one (3.44) obtained by the Galileo transformations. This suggests that we should construct the hierarchy of the Galileo transformations with the properly defined time hierarchy for any of Eqs. (3.51) and the 1-cocycle conditions.

IV. BF GAUGE THEORY AND THE TIME HIERARCHY

In the present section we map the integrable hierarchy from Sec. III to the non-Abelian BF gauge theory. The BF theory is described by the action

$$S=\frac{kL}{4\pi}\int_{\Sigma_2}\mathrm{Tr}(BF), \tag{4.1}$$

where $F=\epsilon^{\mu\nu}F_{\mu\nu}$ is the gauge curvature tensor in the adjoint representation of the appropriately chosen group, while B is the world scalar Lagrange multipliers, transforming according to the coadjoint representation.[11] This model, formulated in terms of 2-form F, and the zero-form B for any 2-dimensional manifold Σ_2 defines the topological field theory.[20] For the noncompact groups SO(2,1) and ISO(1,1) it provides a gauge theoretical formulation of the Jackiw–Teitelboim lineal gravity theories.[11,21,22] This model, being the constant curvature theory, includes the world scalar Lagrangian multiplier field together with the Riemann scalar R and is a type of scalar–tensor theories or the ''dilaton'' gravity. In the gauge formulation of the gravity theory one introduces instead of the metric tensor $g_{\mu\nu}$, the Einstein–Cartan zweibeins and the spin connection viewed as independent variables. Then, the diffeomorphism invariance of the gravity theory becomes a part of the local gauge invariance. The matter ψ and the gauge field V from the Sec. II are just the zweibeins and the spin-connection fields,[23] defining the moving frame equations. The SU(2) and

SU(1,1) with the compact Abelian $U(1)$ subgroup in our case, correspond to the Euclidean gravity, with the local $O(2)$ rotations in the tangent plane, instead of the $O(1,1)$ Lorentz rotations.

To describe the time evolution, we suppose that $\Sigma_2 = R \times T$. Then the action (4.1) is

$$S = \frac{kL}{2\pi} \int \mathrm{Tr}\, B(\partial_0 J_1 - \partial_1 J_0 + [J_0, J_1]) d^2x = \frac{kL}{2\pi} \int \mathrm{Tr}(B\ \partial_0 J_1 + J_0(\partial_1 B + [J_1, B])) d^2x, \tag{4.2}$$

where we integrated the second term by parts. In terms of (3.8) parametrization and with the B field given by $B = i\phi_0\sigma_3 + g\phi\sigma_- - \overline{\phi}\sigma_+$, the action becomes

$$S = \gamma \int \left(\overline{\phi}\ \partial_0 \psi + \phi\ \partial_0 \overline{\psi} + \frac{1}{2}\ \phi_0 \partial_0 V_1 - \mathscr{H} \right) d^2x, \tag{4.3}$$

where the Hamiltonian density is given by

$$\mathscr{H} = -\{\overline{\psi_0}(D_1\phi + 2ig\phi_0\psi) + \psi_0(\overline{D_1\phi} - 2ig\phi_0\overline{\psi}) + V_0\ \tfrac{1}{2}\ (\partial_1\phi_0 + i\phi\overline{\psi} - i\overline{\phi}\psi)\}, \tag{4.4}$$

and γ is the renormalized coupling constant. Since the action is already in the first-order form we can immediately write the canonical brackets.[24]

$$\{\psi(x), \overline{\phi}(y)\} = \frac{1}{\gamma}\ \delta(x-y), \quad \{\overline{\psi}(x), \phi(y)\} = \frac{1}{\gamma}\ \delta(x-y),$$
$$\{V_1(x), \phi_0(y)\} = \frac{2}{\gamma}\ \delta(x-y). \tag{4.5}$$

The nondynamical Lagrange multipliers ψ_0, $\overline{\psi_0}$, and V_0 lead to the BF ''Gauss law'' constraints,

$$D_1\phi + 2ig\psi\phi_0 = 0, \quad \overline{D_1\phi} - 2ig\overline{\psi}\phi_0 = 0, \tag{4.6}$$

$$\partial_1\phi_0 + i\phi\overline{\psi} - i\overline{\phi}\psi = 0. \tag{4.7}$$

Due to the canonical brackets (4.5) the Gauss law constraints generate the SU(2) ($g=1$), or SU(1,1) ($g=-1$), algebra of the gauge transformations in the dynamical fields ψ, $\overline{\psi}$, and V_1. Since the Hamiltonian (4.4) is proportional to the constraints (4.6), (4.7), it is weakly vanishing, $H \approx 0$. The above properties characterize any reparametrization-invariant theory.

We can partially solve constraints (4.6), (4.7) by integration of the last one,

$$\phi_0 = -i \int^x (\overline{\psi}\phi - \psi\overline{\phi}). \tag{4.8}$$

Then, substituting to the first two (4.6), we get

$$\mathscr{M}_\psi(\phi) \equiv D_1\phi + 2g\psi \int^x (\overline{\psi}\phi - \psi\overline{\phi})(y)\, dy = 0, \tag{4.9a}$$

$$\overline{\mathscr{M}_\psi(\phi)} \equiv \overline{D_1\phi} - 2g\overline{\psi} \int^x (\overline{\psi}\phi - \psi\overline{\phi})(y)\, dy = 0, \tag{4.9b}$$

where we defined the integrodifferential operators $\mathscr{M}$. These two relations combined can be written in terms of the $U(1)$ gauge covariant Λ operator (2.22):

$$i\sigma_3\begin{pmatrix}\mathscr{M}_\psi(\phi)\\ \mathscr{M}_\psi(\phi)\end{pmatrix}=\Lambda\begin{pmatrix}\phi\\ \phi\end{pmatrix}=0. \tag{4.10}$$

This means that the constraint surface of the BF theory is defined by zero modes of the covariant Λ operator. In terms of the usual $\mathscr{L}$ operator (2.23) we have the relation

$$\mathscr{L}\begin{pmatrix}\phi\\ \phi\end{pmatrix}=-\frac{1}{2}V_1\begin{pmatrix}\phi\\ \phi\end{pmatrix}, \tag{4.11}$$

which for the constant-valued gauge potential, like in (3.41), $V_1=-2\lambda$, becomes the eigenvalue problem for the operator $\mathscr{L}$:

$$\mathscr{L}\begin{pmatrix}\phi\\ \phi\end{pmatrix}=\lambda\begin{pmatrix}\phi\\ \phi\end{pmatrix}. \tag{4.12}$$

Comparison with the theory of solitons[12] shows that our world scalars ϕ and $\overline{\phi}$ are very similar to the squared eigenfunctions for the Zhakharov–Shabat problem. Moreover, the constraints surface system (4.6), (4.7) is related to the Maxwell–Bloch equations,[25] describing the propagation of pulses through an inhomogeneously broadened and resonant medium, where ψ is the electric field envelope, ϕ is the polarization, and ϕ_0 is the excitation number density. It would be interesting to illuminate the relation between the last problem and the BF theory.

We can express the action (4.3) in terms of the bilinear form. For any two complex functions ϕ and χ vanishing at infinity, we define

$$\langle\phi,\chi\rangle=\int_{-\infty}^{\infty}(\overline{\phi}\chi+\phi\overline{\chi})dx=2\Re\int_{-\infty}^{\infty}\overline{\phi}\chi\ dx. \tag{4.13}$$

Then, the operator $\mathscr{M}_\psi(\phi)$ from (4.9) is skew symmetric with respect to this bilinear form,

$$\langle\phi,\mathscr{M}\chi\rangle=-\langle\mathscr{M}\phi,\chi\rangle, \tag{4.14}$$

if for the integral part of $\mathscr{M}$ we use the proper definition (2.22b). Another skew symmetric operator is the imaginary unit $i=\sqrt{-1}$,

$$\langle\phi,i\chi\rangle=-\langle i\phi,\chi\rangle. \tag{4.15}$$

In spite of the skew symmetry of these two operators, the product

$$\mathscr{R}=i\mathscr{M}, \tag{4.16}$$

being the recursion operator, is not Hermitian. This is due to the noncommutativity of the operators,

$$[i,\mathscr{M}](\phi)=-4ig\psi\int^{x}\overline{\phi}\psi. \tag{4.17}$$

To derive the Hamiltonian dynamics for action (4.3) we introduce the $U(1)$ gauge-invariant variables. Representing potential V_1 in a pure gauge form,

$$V_1=2\,\partial_1\alpha(x), \tag{4.18}$$

for the gauge-invariant $B(x)$ (the magnetic field analog),

$$\phi_0 = \int^x B(y)dy, \tag{4.19}$$

the Poisson bracket (4.5) results from the following canonical bracket:

$$\{\alpha(x), B(y)\} = -\frac{1}{\gamma}\,\delta(x-y). \tag{4.20}$$

Defining the $U(1)$ gauge-invariant fields,

$$\psi = \Psi e^{i\alpha}, \quad \phi = \Phi e^{i\alpha}, \tag{4.21}$$

for action (4.3) we get

$$S = \gamma \int \left[\overline{\Phi}\,\partial_0\Psi + \Phi\,\partial_0\overline{\Psi} + \left\{ \overline{\psi_0} e^{i\alpha}\left(\partial_1\Phi + 2ig\Psi \int^x B \right) + \text{c.c.} \right\} \right.$$
$$\left. + \frac{1}{2}\,(V_0 - 2\,\partial_0\alpha)\{B + i\overline{\Psi}\Phi - i\overline{\Phi}\Psi\} \right]. \tag{4.22}$$

After redefinition of the Lagrange multipliers,

$$\mathscr{V}_0 \equiv V_0 - 2\,\partial_0\alpha, \quad \overline{\Psi}_0 \equiv \overline{\psi_0} e^{i\alpha}, \quad \Psi_0 \equiv \psi_0 e^{-i\alpha}, \tag{4.23}$$

the action is written only in terms of the gauge-invariant dynamical variables Ψ, $\overline{\Psi}$, Φ, $\overline{\Phi}$:

$$S = \gamma \int \left[\overline{\Phi}\,\partial_0\Psi + \Phi\,\partial_0\overline{\Psi} + \left\{ \overline{\Psi}_0\left(\partial_1\Phi + 2ig\Psi \int^x B \right) + \text{c.c.} \right\} + \frac{1}{2}\,\mathscr{V}_0\{B + i\overline{\Psi}\Phi - i\overline{\Phi}\Psi\} \right]. \tag{4.24}$$

We can solve the Gauss law constraint explicitly,

$$B = -i(\overline{\Psi}\Phi - \Psi\overline{\Phi}). \tag{4.25}$$

Then the action is

$$S = \gamma \int (\overline{\Phi}\,\partial_0\Psi + \Phi\,\partial_0\overline{\Psi}) - H, \tag{4.26a}$$

where the Hamiltonian in terms of the bilinear form (4.13) becomes

$$H = -\gamma\langle \Psi_0, M_\Psi \Phi\rangle = \gamma\langle M_\Psi \Psi_0, \Phi\rangle. \tag{4.26b}$$

Here, the skew symmetric operator M is the noncovariant form of the $\mathscr{M}$ operator (4.9):

$$M_\Psi(\Phi) \equiv \partial_1\Phi + 2g\Psi \int^x (\overline{\Psi}\Phi - \Psi\overline{\Phi})(y)dy = 0, \tag{4.27}$$

and coincides with the one introduced by Magri.[26] The Poisson brackets for $U(1)$ gauge-invariant dynamical variables,

$$\{\Psi(x), \overline{\Phi}(y)\} = \frac{1}{\gamma}\,\delta(x-y), \quad \{\overline{\Psi}(x), \Phi(y)\} = \frac{1}{\gamma}\,\delta(x-y), \tag{4.28}$$

generate the Hamiltonian equations, given by

$$\partial_0\Psi=\{\Psi,H\}=\frac{1}{\gamma}\frac{\delta H}{\delta\overline{\Phi}}=M_\Psi(\Psi_0), \tag{4.29a}$$

$$\partial_0\overline{\Psi}=\{\overline{\Psi},H\}=-\frac{1}{\gamma}\frac{\delta H}{\delta\Phi}=\overline{M_\Psi(\Psi_0)}, \tag{4.29b}$$

or in terms of the recursion operator, $R=iM$,

$$i\ \partial_0\Psi=R_\Psi(\Psi_0),\quad -i\ \partial_0\overline{\Psi}=\overline{R_\Psi(\Psi_0)}. \tag{4.30}$$

Thus, we can see the effect of introducing the Φ field, as the canonical momentum conjugate to the dynamical field Ψ, is to eliminate the nondynamical Lagrange multipliers Ψ_0.[21] Due to the arbitrariness of these multipliers, by the recursion operator R, we can construct the whole hierarchy of the multipliers,

$$\Psi_0^{(n)}=R_\Psi(\Psi_0^{(n-1)})=\cdots=R_\Psi^n(\Psi_0^{(0)}). \tag{4.31}$$

Then, the related actions,

$$S_{n+1}=\gamma\int(\overline{\Phi}\ \partial_{0_{n+1}}\Psi+\Phi\ \partial_{0_{n+1}}\overline{\Psi})-H_{n+1}, \tag{4.32a}$$

with the Hamiltonian

$$H_{n+1}=-\gamma\langle\Psi_0^{(n)},M_\Psi\Phi\rangle=\gamma\langle M_\Psi\Psi_0^{(n)},\Phi\rangle, \tag{4.32b}$$

define the evolution hierarchy,

$$i\ \partial_{0_{n+1}}\Psi=R_\Psi^{n+1}(\Psi_0^{(0)}),\quad -i\ \partial_{0_{n+1}}\overline{\Psi}=\overline{R_\Psi^{n+1}(\Psi_0^{(0)})}. \tag{4.33}$$

By skew symmetry property of the M operator we have the relations

$$\begin{aligned}H_{n+1}&=-\gamma\langle R^{(n)}(\Psi_0^{(0)}),M(\Phi)\rangle=-\gamma\langle R^{(n-1)}(\Psi_0^{(0)}),MR(\Phi)\rangle\\&=-\gamma\langle R^{(n-2)}(\Psi_0^{(0)}),MR^2(\Phi)\rangle\\&=\cdots=-\gamma\langle\Psi_0^{(0)},MR^n(\Phi)\rangle,\end{aligned} \tag{4.34}$$

or

$$H_{n+1}=\gamma\langle iR^{(n+1)}(\Psi_0^{(0)}),\Phi\rangle=\gamma\langle i\Psi_0^{(n+1)},\Phi^{(0)}\rangle=\gamma\langle\Psi_0^{(0)},iR^{n+1}(\Phi)\rangle=\gamma\langle\Psi_0^{(0)},i\Phi^{(n+1)}\rangle, \tag{4.35}$$

where

$$\Phi^{(n+1)}=R^{n+1}(\Phi^{(0)}),\quad \Phi^{(0)}\equiv\Phi. \tag{4.36}$$

If the constraint surface is defined by the relation

$$R(\Phi^{(0)})=0, \tag{4.37}$$

then it satisfies the constraints

$$R^2(\Phi^{(0)})=R^3(\Phi^{(0)})=\cdots=0, \tag{4.38}$$

for any degree n. Moreover, the surface (4.37) defines evolution,

$$i\ \partial_{0_1}\Psi=R(\Psi^{(0)}), \tag{4.39}$$

while the higher degree surface (4.38) generates the higher evolution,

$$i\ \partial_{0_n}\Psi=R^n(\Psi^{(0)}). \tag{4.40}$$

This just confirms the evident fact that the first member of the hierarchy (4.39) completely defines the higher evolutions.

Hence, in the Hamiltonian interpretation, the canonical bracket (4.28) for the first member of hierarchy (4.39):

$$\{\Psi(x),\overline{\Phi}^{(0)}(y)\}=\frac{1}{\gamma}\,\delta(x-y),\quad \{\overline{\Psi}(x),\Phi^{(0)}(y)\}=\frac{1}{\gamma}\,\delta(x-y), \tag{4.41}$$

defines the higher Poisson brackets,

$$\{\Psi(x),\overline{\Phi}^{(n)}(y)\}=\{\Psi(x),\overline{R^n\Phi^{(0)}}(y)\}\ =\frac{1}{\gamma}\,\delta(x-y),$$

$$\{\overline{\Psi}(x),\Phi^{(n)}(y)\}=\{\overline{\Psi}(x),R^n\Phi^{(0)}(y)\}\ =\frac{1}{\gamma}\,\delta(x-y). \tag{4.42}$$

In addition, every Hamiltonian (4.35) corresponds to the higher integrals of motion for the model. Thus, the BF formulation provides a natural description of the whole integrable hierarchy.

The above procedure can be extended to the case of spectral parameter dependence, by reduction of the $U(1)$ connection to the constant value, as in Secs. II and III. The details of the calculation will be published soon.

V. CONCLUSION

The main ingredient of our consideration is the moving frame method. Historically,[27] the moving frame idea has origin from mechanics where studying the rigid body motion, one gets one-parameter frames family depending on time and completely characterizing the rigid motion. Darboux[28] and Cotton[29] has generalized the one-parameter frame to several parameters. The richness of the theory was demonstrated by Cartan[30] (*la méthode du répère mobile*), who successfully applied it to research in differential geometry. We find that this deep method is conceptually simple to study integrable models and the underlying gauge theoretical structure.

In the recent study of the dimensionally reduced Jackiw–Pi model,[4–6] to produce dynamics for the B field, an extension of the Lagrangian was given and the chiral solitons, propagating only in one direction was constructed. From our consideration it is clear that the chiral solitons can be created by the odd members of the hierarchy with the odd dispersion. This fact is valid, even for the linear approximation. Thus, for the NLS (2.32) the linear wave with the wave number k is propagating with velocity $v=k$, while the wave for $-k$ is propagating in the opposite direction with the same velocity. On the contrary, for the next member of the hierarchy, Eq. (2.33), both waves with $+k$ and $-k$ are propagating in the same direction with velocity $v=k^2$. The similar structure is valid also for solitons. This shows that the chiral solitons in our approach belong to the odd members of the NLS hierarchy.

In conclusion, as shown in the present paper, the $1+1$-dimensional projection of Chern–Simons gauge interacting theory contains the hidden hierarchy of integrable systems. The statistical gauge field, after projection, still plays the crucial role defining spectral characteristics of the system. The key point for the construction is the mapping to non-Abelian BF theory, which covers all the rich structure of an integrable system.[31–33] The commutativity of spectral flows naturally provides the hierarchy of the gauge fixing constraints and the corresponding time hierarchy. We hope that our procedure could clarify some difficult questions of the diffeomorphism invariance breaking in the theories like gravity, and the time hierarchy creation with the physical degrees of freedom from the reparametrization invariance.

ACKNOWLEDGMENTS

The authors would like to thank S. S. Roan and S. S. Lin for stimulating discussions. One of the authors (O.K.P) would like to thank Professor Fon-Che Liu and Institute of Mathematics, Academia Sinica, Taipei, for the warm hospitality. This work was supported in part by the National Science Council at Taipei, Taiwan.

[1] R. Jackiw and S. Y. Pi, Phys. Rev. Lett. **64**, 2969 (1990); Phys. Rev. D **42**, 3500 (1990).
[2] F. Wilczek, *Fractional Statistics and Anyon Superconductivity* (World Scientific, Singapore, 1990).
[3] *The Quantum Hall Effect*, edited by S. Girvin and R. Prange (Springer-Verlag, Berlin, 1990).
[4] S. J. Benetton Rabello, Phys. Rev. Lett. **76**, 4007 (1996).
[5] U. Aglietti, L. Griguolo, R. Jackiw, S.-Y. Pi, and D. Seminara, hep-th/9606141.
[6] R. Jackiw, ''A nonrelativistic chiral soliton in one dimension,'' preprint MIT-CTP N 2587, hep-th/9611185.
[7] L. Martina, O. K. Pashaev, and G. Soliani, Mod. Phys. Lett. A **8**, 3241 (1993).
[8] O. K. Pashaev, Mod. Phys. Lett. A **11**, 1713 (1996).
[9] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[10] C. Teitelboim, Phys. Lett. B **126**, 41 (1983); and in *Quantum Theory of Gravity*, edited by S. Chrsitensen (A. Hilger, Bristol, UK, 1984); R. Jackiw, *ibid.*; Nucl. Phys. B **252**, 343 (1985).
[11] R. Jackiw, Theor. Math. Phys. **92**, 979 (1992).
[12] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[13] O. K. Pashaev, J. Math. Phys. **37**, 4368 (1996).
[14] L. A. Takhtajan, Phys. Lett. A **64**, 235 (1977).
[15] O. K. Pashaev and S. A. Sergeenkov, Physica A **137**, 282 (1986).
[16] R. Jackiw, *Diverse Topics in Theoretical and Mathematical Physics* (World Scientific, Singapore, 1995).
[17] M. Tinkham, *Introduction to Superconductivity* (McGraw-Hill, New York, 1975).
[18] V. G. Kadyshevsky, in Proc. ''Problemi Kvantovoy Teorii Polya'' (Problems of Quantum Field Theory), Alushta 79 (Dubna, 1979).
[19] A. Doliwa and P. M. Santini, Phys. Lett. A **185**, 373 (1994), and references therein.
[20] M. Blau and G. Thompson, Ann. Phys. **205**, 130 (1991); E. D'Hoker, Phys. Lett. B **264**, 101 (1991).
[21] K. Isler and C. A. Trugenberger, Phys. Rev. Lett. **63**, 834 (1989).
[22] T. Fukuyama and K. Kamimura, Phys. Lett. B **160**, 259 (1985); A. Chamseddine and D. Wyler, *ibid.* **228**, 75 (1989).
[23] L. Martina, O. K. Pashaev, and G. Soliani, ''Integrable dissipative structures in the gauge theory of gravity,'' preprint Lecce University DFUL-1/02/97, 1997, Class. Quantum Grav. (in press).
[24] L. Faddeev and R. Jackiw, Phys. Rev. Lett. **60**, 1692 (1988).
[25] M. J. Ablowitz, D. J. Kaup, and A. C. Newell, J. Math. Phys. **15**, 1852 (1974).
[26] F. Magri, J. Math. Phys. **19**, 1156 (1978).
[27] S. S. Chern and W. Chen, ''Lecture notes on differential geometry'' (in Chinese, 1982).
[28] G. Darboux, *Lecons Sur la Theorie Generale des Surfaces* (Gauthier-Villars, Paris, 1987).
[29] E. Cotton, Bull. Soc. Math. Paris **33**, 42 (1905).
[30] E. Cartan, *Theorie Des Groupes Finis et Continus et la Geometrie Differentielle Traitees par la Methode du Repere Mobile* (Gauthier-Villars, Paris, 1937).
[31] J.-H. Lee and O. K. Pashaev, Int. J. Mod. Phys. A **12**, 213 (1997).
[32] O. K. Pashaev, ''Integrable models as constrained topological gauge theory,'' in *Proceedings of the 2nd Conference on Constrained Dynamics and Quantum Gravity* (in honor of Tullio Regge 65's), Santa Margerita, Italy, 1996, Nucl. Phys. Suppl. (in press).
[33] J.-H. Lee and O. K. Pashaev, ''Abelian gauge theory and integrable σ models,'' preprint MIAS 97—3, Institute of Mathematics, Academia Sinica, Taipei, June, 1997.

Two-body Dirac equations for general covariant interactions and their coupled Schrödinger-like forms

Peter Long and Horace W. Crater
The University of Tennessee Space Institute, Tullahoma, Tennessee 37388

(Received 9 July 1997; accepted for publication 28 August 1997)

We present new and useful ''external potential'' forms of the two-body Dirac equations of constraint dynamics for combined scalar, vector, pseudoscalar, pseudovector, and tensor interactions. These equations have potential applications in two-body problems for bound states in meson spectroscopy and phase shift analysis in nucleon–nucleon scattering. Toward this end, we derive their coupled Schrödinger-like forms using matrix techniques and obtain the corresponding radial equations to these forms from scalar and vector spherical harmonic decompositions. © *1998 American Institute of Physics.* [S0022-2488(98)00801-9]

I. INTRODUCTION

Dirac proposed in 1928 a single-particle wave equation for the behavior of a relativistic spin-1/2 particle.[1] The Dirac equation applies to the model of a single spin-1/2 particle in an external field. Its successful application to hydrogenic systems is dependent on the approximation that the proton is static and its magnetic moment can be ignored compared to the electron magnetic moment (i.e., there is no dipole–dipole electromagnetic interaction between the electron and proton). In general in systems such as mesons, muonium, positronium, and the deuteron, the constituent particles can have equal or near equal mass and magnetic moment: the external potential model breaks down. Theories that put both particles on an equal footing in describing two interacting spin-1/2 particles have all built on the single particle Dirac equation. Breit introduced an extension of Dirac's equation to the two-body problem in 1929.[2–4] Breit's form does not retain the manifest covariance of the original Dirac equation and relativistic corrections beyond the Coulomb potential must be handled perturbatively to avoid singularities.[5–8] Nevertheless, the Breit equation has produced accurate positronium and muonium spectra.[2–5]

The general subject matter of this work, two-body bound state wave equations, is an old problem but, judging by the very large variety of approaches,[9] one with no consensus, without yet an agreed upon formalism. Gradually over the past 15 years several authors have followed an approach[10–17] to this problem that appears to have been quite successful and shows additional promise by overcoming defects in many of the approaches listed in Ref. 9. This approach is that of Dirac's Hamiltonian constraint mechanics[18] and provides a manifestly covariant detour around many of the problems that plague the implementations and applications of the Bethe–Salpeter equation. What is remarkable about the constraint approach is that it provides a covariant three-dimensional rearrangement of the Bethe–Salpeter equation with none of the accompanying problems, including those of relative time excitations.[19] It also has all of the formal familiarity of the local nonrelativistic Schrödinger equation. The methods described in this paper derive directly from Dirac's relativistic Hamiltonian formalism (developed during the 1960s[18]) and Todorov's covariant two-body model, based on his two-body generalization of the one-body Einstein mass-shell relation ($m^2+\mathbf{p}^2=E^2$) (developed during the 1970s[9,20]). Dirac's Hamiltonian formalism is well suited for the covariant canonical quantization of relativistic two-body dynamics models such as Todorov's. Crater and Van Alstine adapted a theoretical hybrid of these two methods to a covariant description of interacting spin-1/2 particles—the coupled two-body Dirac equations.

A simplified first impression of the Crater–Van Alstine adaptation, which would not be too far off, would be to say that Dirac's constraint mechanics supplies a Hamiltonian formalism into

0022-2488/98/39(1)/124/37/$15.00
© 1998 American Institute of Physics

which Todorov's canonically conjugate variables are introduced for the purposes of calculation. Furthermore, in his quasipotential approach to the two body problem, Todorov introduced an effective ''external potential'' framework in which the potentials appear in the two-body problem, giving it an important intuitive appeal, both for ease of interpretation and practical implementation.[20] For spinless particles the Dirac constraint approach incorporates two coupled Klein–Gordon equations while for systems of two interacting spin-1/2 particles (positronium, mesons) it includes two coupled Dirac equations.

In a series of papers[10,11,13–16] Crater and Van Alstine have embedded Todorov's effective particle approach in the framework of Dirac's Hamiltonian constraint mechanics for a description of two-body systems which incorporates spin with applications to QED and quark model calculations based on QCD. Their approach yields manifestly covariant coupled equations in which interactions are introduced via minimal substitution: $p^\mu \to \pi^\mu = p^\mu - A^\mu$ for the Lorentz vector interaction and $m \to M = m + S$ for the Lorentz scalar interaction. Unlike with the Breit equation, the relativistic spin corrections need not be treated only perturbatively.[21] Since the operators are quantum mechanically well defined, one can use nonperturbative techniques (analytic as well as numerical) for the solution of bound state problems.

In recent work[14] Crater and Van Alstine have extended this approach to interactions beyond scalar and four-vector, including pseudoscalar, pseudovector, and tensor interactions in the two-body Dirac equations in the ''external potential'' framework introduced by Todorov. Their results were separate sets of coupled equations for each of these interactions acting singly, and several equations for some of the simpler interactions acting simultaneously in combination. Those authors presented a method for deriving the general case two-body Dirac equations for all interactions acting together simultaneously: however, they did not derive this important general result nor did they carry out the equally important Pauli reductions to Schrödinger-like forms.

In this paper we present the logical conclusion of results discussed in Ref. 14, deriving the general coupled equations for all interactions acting at once. We then reduce these equations to a Schrödinger-like form (its so-called Pauli-reduction) that is very useful for numerical calculations of the energy eigenstates for more realistic quark model interactions and for the difficult phase shift analysis problem of nucleon–nucleon scattering.

II. CONSTRAINT TWO-BODY DIRAC EQUATIONS

This section presents a review of previous material on constraint dynamics. One first introduces the relevant dynamical variables for the two-body Dirac equations of constraint dynamics (which we collectively call Todorov variables):

(i) relative position, $x_1 - x_2$;
(ii) total momentum, $P = p_1 + p_2$;
(iii) total c.m. ($\mathbf{P}=0$) energy, $w = \sqrt{-P^2}$;
(iv) constituent on-shell c.m. energies, $\epsilon_1 = (w^2 + m_1^2 - m_2^2)/2w$, $\epsilon_2 = (w^2 + m_2^2 - m_1^2)/2w$;
(v) relativistic reduced mass for fictitious particle of relative motion, $m_w = m_1 m_2 / w$;
(vi) c.m. energy of fictitious particle of relative motion, $\epsilon_w = (w^2 - m_1^2 - m_2^2)/2w$, and
(vii) relative momentum, $p = (\epsilon_2 p_1 - \epsilon_1 p_2)/w$. On mass shell,

$$p^2 = \epsilon_w^2 - m_w^2 = \frac{w^4 - 2(m_1^2 + m_2^2) + (m_1^2 - m_2^2)^2}{4w^2} \equiv b^2(w) = \epsilon_1^2 - m_1^2 = \epsilon_2^2 - m_2^2. \tag{2.1}$$

In terms of these variables,

$$p_1 = \epsilon_1 \hat{P} + p, \quad p_2 = \epsilon_2 \hat{P} - p, \tag{2.2}$$

where $\hat{P} = P/w$. Then one writes the Dirac equations for two free particles

$$\mathscr{S}_{10}\psi=(\theta_1\cdot p_1+m_1\theta_{51})\psi=0, \tag{2.3a}$$

$$\mathscr{S}_{20}\psi=(\theta_2\cdot p_2+m_2\theta_{52})\psi=0, \tag{2.3b}$$

where ψ is just the product of the two single-particle Dirac wave functions. These equations are compatible (they trivially commute $[\mathscr{S}_{10},\mathscr{S}_{20}]_-=0$) and become

$$\mathscr{S}_{10}\psi=(\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+m_1\theta_{51})\psi=0, \tag{2.4a}$$

$$\mathscr{S}_{20}\psi=(-\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+m_2\theta_{52})\psi=0, \tag{2.4b}$$

when expressed in terms of the Todorov variables. The ''theta'' matrices

$$\theta_i^\mu\equiv i\sqrt{\tfrac{1}{2}}\,\gamma_{5i}\gamma_i^\mu, \quad \mu=0,1,2,3, \quad i=1,2, \tag{2.5}$$

$$\theta_{5i}\equiv i\sqrt{\tfrac{1}{2}}\,\gamma_{5i}, \tag{2.6}$$

satisfy the fundamental anticommutation relations

$$[\theta_i^\mu,\theta_i^\nu]_+=-g^{\mu\nu}, \tag{2.7a}$$

$$[\theta_{5i},\theta_i^\mu]_+=0, \tag{2.7b}$$

$$[\theta_{5i},\theta_{5i}]_+=-1. \tag{2.7c}$$

(Projected ''theta'' matrices then satisfy

$$[\theta_i\cdot\hat{P},\theta_i\cdot\hat{P}]_+=1, \tag{2.8}$$

$$[\theta_i\cdot\hat{P},\theta_{i\perp}^\mu]_+=0, \tag{2.9}$$

where $\theta_{i\perp}^\mu=\theta_{i\nu}(\eta^{\mu\nu}+\hat{P}^\mu\hat{P}^\nu)$.) They are modified Dirac matrices[22] which ensure that the Dirac operators $\mathscr{S}_{10}$ and $\mathscr{S}_{20}$ are the square root operators of the corresponding mass-shell operators $-\frac{1}{2}(p_1^2+m_1^2)$ and $-\frac{1}{2}(p_2^2+m_2^2)$.

Using the Todorov variables and the above brackets, the difference

$$(\mathscr{S}_{10}^2-\mathscr{S}_{20}^2)\psi=0=-\tfrac{1}{2}(p_1^2+m_1^2-p_2^2-m_2^2)\psi \tag{2.10}$$

leads to an equation

$$P\cdot p\,\psi=\tfrac{1}{2}[w(\epsilon_1-\epsilon_2)-(m_1^2-m_2^2)]\psi=0, \tag{2.11}$$

which puts a constraint on the relative momentum (eliminating the relative energy in the c.m. frame).

We define covariant (c.m. projected) versions of the Dirac α and β matrices by

$$\beta_i=-\gamma_i\cdot\hat{P}=2\,\theta_{5i}\theta_i\cdot\hat{P}, \tag{2.12}$$

$$\alpha_i^\mu=2\,\theta_{i\perp}^\mu\,\theta_i\cdot\hat{P}, \tag{2.13}$$

and

$$\Sigma_i^\mu=\gamma_{5i}\alpha_i^\mu=2\sqrt{2}\,i\,\theta_{5i}\theta_i\cdot\hat{P}\,\theta_{\perp i}, \quad i=1,2, \tag{2.14}$$

which take on the simple form $\alpha_i^\mu=(0,\boldsymbol{\alpha}_i)$ and $\Sigma_i^\mu=(0,\boldsymbol{\Sigma}_i)$ in the center-of-mass system for which $\hat{P}=(1,\mathbf{0})$.

If we attempt to introduce scalar interactions between particles by naively making the minimal substitutions

$$m_i\rightarrow M_i=m_i+S_i\,,\quad i=1,2 \tag{2.15}$$

(as done in the one-body equation), the resulting Dirac equations

$$\mathscr{S}_1\psi=(\theta_1\cdot p+\boldsymbol{\epsilon}_1\theta_1\cdot\hat{P}+M_1\theta_{51})\psi=0, \tag{2.16a}$$

$$\mathscr{S}_2\psi=(-\theta_2\cdot p+\boldsymbol{\epsilon}_2\theta_2\cdot\hat{P}+M_2\theta_{52})\psi=0, \tag{2.16b}$$

will not be compatible because

$$[.\mathscr{S}_1,.\mathscr{S}_2]_-\psi=[\theta_1\cdot p,M_2\theta_{52}]+[M_1\theta_{51},-\theta_2\cdot p]-=-i(\partial M_1\cdot\theta_1\theta_{52}+\partial M_2\cdot\theta_2\theta_{51})\psi\neq 0 \tag{2.17}$$

(∂ is the four-gradient).

In the earlier work,[10,13] compatibility was achieved for scalar and timelike vector interactions with the help of supersymmetry arguments. For world scalar interactions, this results in the introduction of interactions by way of constituent mass potentials M_i related to one another through a relativistic ''third law''

$$\partial(M_1^2-M_2^2)=0, \tag{2.18}$$

depending on the separation variable only through the spacelike projection perpendicular to the total momentum

$$M_i=M_i(x_\perp), \tag{2.19}$$

where[18]

$$x_\perp^\mu=(\eta^{\mu\nu}+\hat{P}^\mu\hat{P}^\nu)(x_1-x_2)_\nu\,. \tag{2.20}$$

Integration of the ''third law'' condition yields

$$M_1^2-M_2^2=m_1^2-m_2^2 \tag{2.21}$$

with the hyperbolic solution

$$M_1=m_1\ \mathrm{ch}(L)+m_2\ \mathrm{sh}(L),\quad M_2=m_2\ \mathrm{ch}(L)+m_1\ \mathrm{sh}(L), \tag{2.22}$$

given in terms of a single invariant function $L=L(x_\perp)$.

Without the $x_\perp$ dependence of the potential and the relativistic ''third law'' conditions the constraints would not be compatible. While the physical importance of the $x_\perp$ dependence lies in its exclusion of the relative time from the dynamics in the c.m. frame, the ''third law'' condition relates the effective potentials each particle feels in the presence of the other in just such a way that the constraints become compatible.

The compatible spin-dependent constraints $\mathscr{S}_1$ and $\mathscr{S}_2$ below are then commuting quantum operators (the constraints are strongly compatible)

$$[.\mathscr{S}_1,.\mathscr{S}_2]=0, \tag{2.23}$$

with the resulting two-body Dirac equations for scalar interactions being

$$\mathscr{S}_1\psi=(\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+M_1\theta_{51}-i\partial L\cdot\theta_2\theta_{52}\theta_{51})\psi=0, \tag{2.24a}$$

$$\mathscr{S}_2\psi=(-\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+M_2\theta_{52}+i\partial L\cdot\theta_1\theta_{52}\theta_{51})\psi=0, \tag{2.24b}$$

where

$$\partial L=\frac{\partial M_1}{M_2}=\frac{\partial M_2}{M_1}. \tag{2.25}$$

The strong compatibility of the two constraints has been achieved by a supersymmetry which produces the extra spin-dependent recoil terms involving ∂L. These extra terms vanish, however, when one of the particles becomes infinitely massive (as seen by the parametrization $M_i=m_i+S_i$ of the scalar potential) so that one recovers the expected one-body Dirac equation in an external scalar potential. Notice the ''effective potential'' framework each particle has in the presence of the other in Eqs. (2.24). One of the primary purposes of this paper is to show we can maintain such a framework when all interactions are introduced simultaneously.

The Dirac constraint operators satisfy[10,13]

$$(\mathscr{S}_1^2-\mathscr{S}_2^2)\psi=-\tfrac{1}{2}(p_1^2+m_1^2-p_2^2-m_2^2)\psi=-P\cdot p\,\psi=0, \tag{2.26}$$

so that the relative momentum remains orthogonal to the total momentum even in the presence of scalar interactions.

One solves the compatibility problem without having to invent new supersymmetries if the scalar potential is replaced by vector, pseudoscalar, pseudovector, or tensor potentials, by first rewriting the supersymmetric or ''external potential'' form Eqs. (2.24) of the two-body Dirac equations for scalar interaction in the hyperbolic form[14] below:

$$\mathscr{S}_1\psi=(\mathrm{ch}(\Delta)\mathbf{S}_1+\mathrm{sh}(\Delta)\mathbf{S}_2)\psi=0, \tag{2.27a}$$

$$\mathscr{S}_2\psi=(\mathrm{ch}(\Delta)\mathbf{S}_2+\mathrm{sh}(\Delta)\mathbf{S}_1)\psi=0, \tag{2.27b}$$

where Δ generates the scalar potential terms in (2.24) provided that

$$\Delta=-\theta_{51}\theta_{52}L(x_\perp). \tag{2.28}$$

The operators $\mathbf{S}_1$ and $\mathbf{S}_2$ are auxiliary constraints of the form

$$\mathbf{S}_1\psi\equiv(\mathscr{S}_{10}\,\mathrm{ch}(\Delta)+\mathscr{S}_{20}\,\mathrm{sh}(\Delta))\psi=0, \tag{2.29a}$$

$$\mathbf{S}_2\psi\equiv(\mathscr{S}_{20}\,\mathrm{ch}(\Delta)+\mathscr{S}_{10}\,\mathrm{sh}(\Delta))\psi=0. \tag{2.29b}$$

Second, if one postulates that Eqs. (2.29a) and (2.29b) are valid for arbitrary Δ, then a direct consequence is that[12,14,21]

$$[P\cdot p,\Delta]\psi=0, \tag{2.30}$$

or $\Delta(x)=\Delta(x_\perp)$ and

$$P\cdot p\,\psi=0. \tag{2.31}$$

This, in turn leads to the weak compatibility of both sets of constraints:[21,14,12]

$$[\mathscr{S}_1,\mathscr{S}_2]\psi=0 \tag{2.32}$$

and

$$[\mathbf{S}_1,\mathbf{S}_2]\psi=0. \tag{2.33}$$

In our work here we shall consider Eqs. (2.27) and (2.29) for a sum of the four ''polar'' and four ''axial'' interactions. The four polar interactions[23] are

$$\Delta_L=-L\theta_{51}\theta_{52}\equiv-1\frac{L}{2}\mathscr{O}_1,\mathscr{O}_1=-\gamma_{51}\gamma_{52},\quad \text{scalar,} \tag{2.34a}$$

$$\Delta_J=J\hat{P}\cdot\theta_1\hat{P}\cdot\theta_2\equiv\mathscr{O}_2\frac{J}{2}=\beta_1\beta_2\frac{J}{2}\mathscr{O}_1,\quad \text{timelike vector,} \tag{2.34b}$$

$$\Delta_{\mathscr{G}}=\mathscr{G}\theta_{1\perp}\cdot\theta_{2\perp}\equiv\mathscr{O}_3\frac{\mathscr{G}}{2}=\gamma_{1\perp}\cdot\gamma_{2\perp}\frac{\mathscr{G}}{2}\mathscr{O}_1,\quad \text{spacelike vector,} \tag{2.34c}$$

$$\Delta_{\mathscr{F}}=4\mathscr{F}\theta_{1\perp}\cdot\theta_{2\perp}\theta_{51}\theta_{52}\hat{P}\cdot\theta_1\hat{P}\cdot\theta_2\equiv\mathscr{O}_4\frac{\mathscr{F}}{2}=\alpha_1\cdot\alpha_2\frac{\mathscr{F}}{2}\mathscr{O}_1,\quad \text{tensor (polar),} \tag{2.34d}$$

and would correspond to interactions with two polar vertices in an effective exchange model. The polar matrices are ''doubly odd,'' i.e., are odd in the number of theta matrices of both particle one and particle two. Because of this they each will anticommute with the singly odd theta matrices in $\mathscr{S}_{i0}$ of different types and the natural brackets are anticommutators. Their sum,

$$\Delta_p=\Delta_L+\Delta_J+\Delta_{\mathscr{G}}+\Delta_{\mathscr{F}}, \tag{2.35}$$

would be used in Eqs. (2.27) and (2.29) with no alteration.

The four ''axial'' interactions are[23]

$$\Delta_C=C/2\equiv\mathscr{E}_1\frac{C}{2}=-\gamma_{51}\gamma_{52}\frac{C}{2}\mathscr{O}_1,\quad \text{pseudoscalar,} \tag{2.36a}$$

$$\Delta_H=-2H\hat{P}\cdot\theta_1\hat{P}\cdot\theta_2\theta_{51}\theta_{52}\equiv-\mathscr{E}_2\frac{H}{2}=\beta_1\gamma_{51}\beta_2\gamma_{52}\frac{H}{2}\mathscr{O}_1,\quad \text{timelike pseudovector,} \tag{2.36b}$$

$$\Delta_I=-2I\theta_{1\perp}\cdot\theta_{2\perp}\theta_{51}\theta_{52}\equiv-\mathscr{E}_3\frac{I}{2}=-\gamma_{51}\gamma_{1\perp}\cdot\gamma_{52}\gamma_{2\perp}\frac{I}{2}\mathscr{O}_1,\quad \text{spacelike pseudovector,} \tag{2.36c}$$

$$\Delta_Y=-2Y\theta_{1\perp}\cdot\theta_{2\perp}\hat{P}\cdot\theta_1\hat{P}\cdot\theta_2\equiv-\mathscr{E}_4\frac{Y}{2}=-\Sigma_1\cdot\Sigma_2\frac{Y}{2}\mathscr{O}_1,\quad \text{tensor (axial).} \tag{2.36d}$$

These Δ_a's correspond to interactions with two axial vertices in an effective exchange model. The axial matrices are ''doubly even,'' i.e., are even in the number of theta matrices of both particle one and particle two. Because of this they each will commute with the singly odd theta matrices in $\mathscr{S}_{i0}$ of different types and the natural brackets to use are commutators. Their sum

$$\Delta_a=\Delta_C+\Delta_H+\Delta_I+\Delta_Y \tag{2.37}$$

would be used in Eqs. (2.27) and (2.29) but with the $\mathrm{sh}(\Delta_a)$ terms in Eq. (2.27) appearing with a negative sign instead of the plus sign as occurs for polar interactions.[14] In contrast, Eq. (2.29) has no sign changes for Δ_a.

For systems with both polar and axial interactions,[14] one uses $\Delta_p-\Delta_a$ in (2.27) and $\Delta_p+\Delta_a$ in (2.29) in place of Δ. The terms $L,\ J,\ \mathscr{G},\ \mathscr{F},\ C,\ H,\ I,\ Y$ are arbitrary invariant functions of $x_\perp$. Equation (2.29) alone is more convenient to use for the construction of Breit-like equations from the constraint formalism for general interactions. Equation (2.27) in conjunction with Eq. (2.29) is more convenient if one's aim (as is ours here) is to obtain a set of Dirac equations in an ''external potential'' form similar to that exhibited in the one-body Dirac equation, with the transformation properties one would expect for such potentials.

In subsequent sections we derive the ''external potential'' forms of the two-body Dirac equations for the sum of all eight covariant interactions. We then derive from them the necessary Schrödinger-like equations for the combined eight interactions.

III. ''EXTERNAL POTENTIAL'' FORM OF THE CONSTRAINT TWO-BODY DIRAC EQUATIONS

The compatibility of the general forms of the hyperbolic constraints (2.27a) and (2.27b) for arbitrary $\Delta(x_\perp)$ leads to a systematic method for the introduction of interactions. We find the $\mathscr{S}_1$ and $\mathscr{S}_2$ constraints (2.24) for the scalar interaction by inserting the appropriate invariant matrix Δ_L in Eqs. (2.27a) and (2.27b) and then expanding (2.29a) and (2.29b) in terms of $\Delta(x_\perp)$.

The derivation of generalizations of the form (2.24) to arbitrary combinations of the eight different invariant forms discussed above is the subject of this section.

A. Deriving the general constraints $\mathscr{S}_1$ and $\mathscr{S}_2$

For each of the individual eight interaction matrices, $\Delta_L,\Delta_J,\Delta_{\mathscr{G}},\Delta_{\mathscr{F}},\Delta_C,\Delta_H,\Delta_I,\Delta_Y$, we can derive[14] the corresponding ''external potential'' forms of the two-body Dirac equations $\mathscr{S}_1\psi=0$ and $\mathscr{S}_2\psi=0$. The basic method is same for both the axial interactions Δ_a's and polar interactions Δ_p's; the only difference arises in the choice of commutation and anticommutation brackets used in the derivation.[24]

Polar constraints from the general hyperbolic form: For each of the individual invariant polar matrices Δ_p's, Δ_L, Δ_J, $\Delta_{\mathscr{G}}$, and $\Delta_{\mathscr{F}}$ (or their sum), a generalization of Eq. (2.27) for ''external potential'' forms of the constraints $\mathscr{S}_1$ and $\mathscr{S}_2$ have the following relation to the auxiliary constraints $\mathbf{S}_1$ and $\mathbf{S}_2$:

$$\mathscr{S}_1\psi=[\mathrm{ch}(\Delta_p)\mathbf{S}_1+\mathrm{sh}(\Delta_p)\mathbf{S}_2]\psi=0, \tag{3.1a}$$

$$\mathscr{S}_2\psi=[\mathrm{ch}(\Delta_p)\mathbf{S}_2+\mathrm{sh}(\Delta_p)\mathbf{S}_1]\psi=0. \tag{3.1b}$$

Explicitly, using the generalization of these auxiliary constraints of Eq. (2.29), we find[14] that $\mathscr{S}_1$ becomes

$$\begin{aligned}\mathscr{S}_1\psi=\mathscr{S}_{10}&+\mathrm{ch}\ \Delta_p[\mathscr{S}_{10},\mathrm{ch}\ \Delta_p]_-+\mathrm{ch}\ \Delta_p[\mathscr{S}_{20},\mathrm{sh}\ \Delta_p]_+\\&+\mathrm{sh}\ \Delta_p[\mathscr{S}_{20},\mathrm{ch}\ \Delta_p]_-+\mathrm{sh}\ \Delta_p[\mathscr{S}_{10},\mathrm{sh}\ \Delta_p]_+ .\end{aligned} \tag{3.2}$$

Working out the commutation and anticommutation relations[24] in (3.2) yields[14] the generalized expression for $\mathscr{S}_1$ in the case of a polar Δ, i.e., $\Delta=\Delta_p$,

$$\begin{aligned}\mathscr{S}_1=\mathscr{S}_{10}+i\theta_2\cdot\partial\Delta_p-i\{&[\mathrm{ch}\ \Delta_p,\theta_1^\mu]_-\ \mathrm{sh}\ \Delta_p-[\mathrm{ch}\ \Delta_p,\theta_2^\mu]_-\mathrm{ch}\ \Delta_p-[\mathrm{sh}\ \Delta_p,\theta_2^\mu]_+\ \mathrm{sh}\ \Delta_p\\&+[\mathrm{sh}\ \Delta_p,\theta_1^\mu]_+\ \mathrm{ch}\ \Delta_p\}\partial_\mu\Delta_p+\{\mathrm{ch}\ \Delta_p[\theta_1^\mu,\mathrm{ch}\ \Delta_p]_--\mathrm{ch}\ \Delta_p[\theta_2^\mu,\mathrm{sh}\ \Delta_p]_+\\&-\mathrm{sh}\ \Delta_p[\theta_2^\mu,\mathrm{ch}\ \Delta_p]_-+\mathrm{sh}\ \Delta_p[\theta_1^\mu,\mathrm{sh}\ \Delta_p]_+\}p_\mu+\mathrm{ch}\ \Delta_p[\epsilon_1\theta_1\cdot\hat{P},\mathrm{ch}\ \Delta_p]_-+\mathrm{ch}\ \Delta_p[\epsilon_2\theta_2\\&\cdot\hat{P},\mathrm{sh}\ \Delta_p]_++\mathrm{sh}\ \Delta_p[\epsilon_2\theta_2\cdot\hat{P},\mathrm{ch}\ \Delta_p]_-+\mathrm{sh}\ \Delta_p[\epsilon_1\theta_1\cdot\hat{P},\mathrm{sh}\ \Delta_p]_++\mathrm{ch}\ \Delta_p[m_1\theta_{51},\mathrm{ch}\ \Delta_p]_-\\&+\mathrm{ch}\ \Delta_p[m_2\theta_{52},\mathrm{sh}\ \Delta_p]_++\mathrm{sh}\ \Delta_p[m_2\theta_{52},\mathrm{ch}\ \Delta_p]_-+\mathrm{sh}\ \Delta_p[m_1\theta_{51},\mathrm{sh}\ \Delta_p]_+\ .\end{aligned}\tag{3.3}$$

The expression for $\mathscr{S}_2$ is obtained by a similar method.

Axial constraints from the general hyperbolic form: For each of the individual invariant axial matrices Δ_a's, Δ_C, Δ_H, Δ_I, and Δ_Y (or their sum), the appropriate generalizations of Eq. (2.27) for the ''external potential'' forms $\mathscr{S}_1$ and $\mathscr{S}_2$ of the constraints have the following relations to the auxilary constraints $\mathbf{S}_1$ and $\mathbf{S}_2$ (Ref. 14):

$$\mathscr{S}_1\psi=[\mathrm{ch}\ \Delta_a\mathbf{S}_1-\mathrm{sh}\ \Delta_a\mathbf{S}_2]\psi=0,\tag{3.4a}$$

$$\mathscr{S}_2\psi=[\mathrm{ch}\ \Delta_a\mathbf{S}_2-\mathrm{sh}\ \Delta_a\mathbf{S}_1]\psi=0.\tag{3.4b}$$

The minus sign replaces the plus sign in Eqs. (3.1a) and (3.1b) so that one may take advantage of the different commutation relation of the Δ_a's and Δ_p's referred to below Eqs. (2.34) and (2.36) in conjunction with various hyperbolic identities.

The explicit calculation of $\mathscr{S}_1$ in the axial case is similar to the polar case, only the commutation relations are different due to their different symmetries. In terms of the axial invariant matrix Δ_a, $\mathscr{S}_1$ is[14]

$$\begin{aligned}\mathscr{S}_1\psi=\mathscr{S}_{10}&+\mathrm{ch}\ \Delta_a[\mathscr{S}_{10},\mathrm{ch}\ \Delta_a]_-+\mathrm{ch}\ \Delta_a[\mathscr{S}_{20},\mathrm{sh}\ \Delta_a]_-\\&-\mathrm{sh}\ \Delta_a[\mathscr{S}_{20},\mathrm{ch}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[\mathscr{S}_{10},\mathrm{sh}\ \Delta_a]_-\ .\end{aligned}\tag{3.5}$$

Working out the commutation and anticommutation relations[24] in (3.5) yields[14] the generalized expression for $\mathscr{S}_1$ in the case of an axial Δ, i.e., $\Delta=\Delta_a$,

$$\begin{aligned}\mathscr{S}_1=\mathscr{S}_{10}+i\theta_2\cdot\partial\Delta_a-i\{&[\mathrm{ch}\ \Delta_a,\theta_1^\mu]_-\mathrm{sh}\ \Delta_a-[\mathrm{ch}\ \Delta_a,\theta_2^\mu]_-\mathrm{ch}\ \Delta_a+[\mathrm{sh}\ \Delta_a,\theta_2^\mu]_-\mathrm{sh}\ \Delta_a\\&-[\mathrm{sh}\ \Delta_a,\theta_1^\mu]_-\mathrm{ch}\ \Delta_a\}\partial_\mu\Delta_a+\{\mathrm{ch}\ \Delta_a[\theta_1^\mu,\mathrm{ch}\ \Delta_a]_--\mathrm{ch}\ \Delta_a[\theta_2^\mu,\mathrm{sh}\ \Delta_a]_-\\&+\mathrm{sh}\ \Delta_a[\theta_2^\mu,\mathrm{ch}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[\theta_1^\mu,\mathrm{sh}\ \Delta_a]_-\}p_\mu+\mathrm{ch}\ \Delta_a[\epsilon_1\theta_1\cdot\hat{P},\mathrm{ch}\ \Delta_a]_-+\mathrm{ch}\ \Delta_a[\epsilon_2\theta_2\\&\cdot\hat{P},\mathrm{sh}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[\epsilon_2\theta_2\cdot\hat{P},\mathrm{ch}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[\epsilon_1\theta_1\cdot\hat{P},\mathrm{sh}\ \Delta_a]_-+\mathrm{ch}\ \Delta_a[m_1\theta_{51},\mathrm{ch}\ \Delta_a]_-\\&+\mathrm{ch}\ \Delta_a[m_2\theta_{52},\mathrm{sh}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[m_2\theta_{52},\mathrm{ch}\ \Delta_a]_--\mathrm{sh}\ \Delta_a[m_1\theta_{51},\mathrm{sh}\ \Delta_a]_-\ .\end{aligned}\tag{3.6}$$

The expression for $\mathscr{S}_2$ is obtained by a similar method.

In spite of the length of Eqs. (3.3) and (3.6), for the simple case of individual interactions (i.e., $\Delta=\Delta_L$ or $\Delta=\Delta_J$, *not* $\Delta=\Delta_L+\Delta_J+\cdots$) many of the commutators vanish, and in practice the calculations are fairly short. Equations (3.7a)–(3.14a) display the result of the calculation for the four polar interactions (Δ_p's:$\Delta_L,\Delta_J,\Delta_{\mathscr{G}},\Delta_{\mathscr{F}}$), and the four axial interactions (Δ_a's:$\Delta_C,\Delta_H,\Delta_I,\Delta_Y$) *acting independently*. As presented in Ref. 14 we obtain the following two-body Dirac equations:

1. Polar interactions—Δ_p's

a. Scalar interaction: $\Delta=\Delta_L$:

$$\mathscr{S}_1\psi=\left(\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+m_1\ \mathrm{ch}(L\mathscr{O}_1)\theta_{51}-m_2\ \mathrm{sh}(L\mathscr{O}_1)\theta_{52}-i\theta_2\cdot\frac{\partial L}{2}\mathscr{O}_1\right)\psi=0, \quad (3.7a)$$

$$\mathscr{S}_2\psi=\left(-\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+m_2\ \mathrm{ch}(L\mathscr{O}_1)\theta_{52}-m_1\ \mathrm{sh}(L\mathscr{O}_1)\theta_{51}+i\theta_1\cdot\frac{\partial L}{2}\mathscr{O}_1\right)\psi=0. \quad (3.7b)$$

b. Timelike vector interaction: $\Delta=\Delta_J$:

$$\mathscr{S}_1\psi=\left(\theta_1\cdot p+\epsilon_1\ \mathrm{ch}(J\mathscr{O}_2)\theta_1\cdot\hat{P}+\epsilon_2\ \mathrm{sh}(J\mathscr{O}_2)\theta_2\cdot\hat{P}+m_1\theta_{51}+i\theta_2\cdot\frac{\partial J}{2}\mathscr{O}_2\right)\psi=0, \quad (3.8a)$$

$$\mathscr{S}_2\psi=\left(-\theta_2\cdot p+\epsilon_2\ \mathrm{ch}(J\mathscr{O}_2)\theta_2\cdot\hat{P}+\epsilon_1\ \mathrm{sh}(J\mathscr{O}_2)\theta_1\cdot\hat{P}+m_2\theta_{52}-i\theta_1\cdot\frac{\partial J}{2}\mathscr{O}_2\right)\psi=0. \quad (3.8b)$$

c. Spacelike vector interaction: $\Delta=\Delta_{\mathscr{G}}$:

$$\mathscr{S}_1\psi=\left(\exp(\mathscr{G})\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+m_1\theta_{51}+i\ \exp(\mathscr{G})\theta_2\cdot\frac{\partial\mathscr{G}}{2}\mathscr{O}_3\right)\psi=0, \quad (3.9a)$$

$$\mathscr{S}_2\psi=\left(-\exp(\mathscr{G})\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+m_2\theta_{52}-i\ \exp(\mathscr{G})\theta_1\cdot\frac{\partial\mathscr{G}}{2}\mathscr{O}_3\right)\psi=0. \quad (3.9b)$$

d. Polar tensor interaction: $\Delta=\Delta_{\mathscr{T}}$:

$$\mathscr{S}_1\psi=\Big(\exp(\mathscr{T}\mathscr{E}_2)\theta_1\cdot p+\epsilon_1\ \mathrm{ch}(\mathscr{T}\mathscr{O}_4)\theta_1\cdot\hat{P}+\epsilon_2\ \mathrm{sh}(\mathscr{T}\mathscr{O}_4)\theta_2\cdot\hat{P}+m_1\ \mathrm{ch}(\mathscr{T}\mathscr{O}_4)\theta_{51}$$
$$+m_2\ \mathrm{sh}(\mathscr{T}\mathscr{O}_4)\theta_{52}+i\ \exp(\mathscr{T}\mathscr{E}_2)\theta_2\cdot\frac{\partial\mathscr{T}}{2}\mathscr{O}_4\Big)\psi=0, \quad (3.10a)$$

$$\mathscr{S}_2\psi=\Big(-\exp(\mathscr{T}\mathscr{E}_2)\theta_2\cdot p+\epsilon_2\ \mathrm{ch}(\mathscr{T}\mathscr{O}_4)\theta_2\cdot\hat{P}+\epsilon_1\ \mathrm{sh}(\mathscr{T}\mathscr{O}_4)\theta_1\cdot\hat{P}+m_2\ \mathrm{ch}(\mathscr{T}\mathscr{O}_4)\theta_{52}$$
$$+m_1\ \mathrm{sh}(\mathscr{T}\mathscr{O}_4)\theta_{51}-i\ \exp(\mathscr{T}\mathscr{E}_2)\theta_1\cdot\frac{\partial\mathscr{T}}{2}\mathscr{O}_4\Big)\psi=0. \quad (3.10b)$$

2. Axial interactions—Δ_a's

a. Pseudoscalar interaction: $\Delta=\Delta_C$:

$$\mathscr{S}_1\psi=\left(\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+m_1\theta_{51}+i\theta_2\cdot\frac{\partial C}{2}\mathscr{E}_1\right)\psi=0, \quad (3.11a)$$

$$\mathscr{S}_2\psi=\left(-\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+m_2\theta_{52}-i\theta_1\cdot\frac{\partial C}{2}\mathscr{E}_1\right)\psi=0. \quad (3.11b)$$

b. Timelike pseudovector interaction: $\Delta=\Delta_H$:

$$\mathscr{S}_1\psi=\Big(\theta_1\cdot p+\epsilon_1\ \mathrm{ch}(H\mathscr{E}_2)\,\theta_1\cdot\hat{P}+\epsilon_2\ \mathrm{sh}(H\mathscr{E}_2)\,\theta_2\cdot\hat{P}$$
$$+m_1\ \mathrm{ch}(H\mathscr{E}_1)\,\theta_{51}+m_2\ \mathrm{sh}(H\mathscr{E}_1)\,\theta_{52}-i\,\theta_2\cdot\frac{\partial H}{2}\,\mathscr{E}_2\Big)\psi=0, \tag{3.12a}$$

$$\mathscr{S}_2\psi=\Big(-\theta_2\cdot p+\epsilon_2\ \mathrm{ch}(H\mathscr{E}_2)\,\theta_2\cdot\hat{P}+\epsilon_1\ \mathrm{sh}(H\mathscr{E}_2)\,\theta_1\cdot\hat{P}$$
$$+m_2\ \mathrm{ch}(H\mathscr{E}_1)\,\theta_{52}+m_1\ \mathrm{sh}(H\mathscr{E}_1)\,\theta_{51}+i\,\theta_1\cdot\frac{\partial H}{2}\,\mathscr{E}_2\Big)\psi=0. \tag{3.12b}$$

c. Spacelike pseudovector interaction: $\Delta=\Delta_I$:

$$\mathscr{S}_1\psi=\Big(\exp(I\mathscr{O}_1)\,\theta_1\cdot p+\epsilon_1\theta_1\cdot\hat{P}+m_1\ \mathrm{ch}(I\mathscr{E}_3)\,\theta_{51}$$
$$+m_2\ \mathrm{sh}(I\mathscr{E}_3)\,\theta_{52}-i\ \exp(I\mathscr{O}_1)\,\theta_2\cdot\frac{\partial I}{2}\,\mathscr{E}_3\Big)\psi=0, \tag{3.13a}$$

$$\mathscr{S}_2\psi=\Big(-\exp(I\mathscr{O}_1)\,\theta_2\cdot p+\epsilon_2\theta_2\cdot\hat{P}+m_2\ \mathrm{ch}(I\mathscr{E}_3)\,\theta_{52}$$
$$+m_1\ \mathrm{sh}(I\mathscr{E}_3)\,\theta_{51}+i\exp(I\mathscr{O}_1)\,\theta_1\cdot\frac{\partial I}{2}\,\mathscr{E}_3\Big)\psi=0. \tag{3.13b}$$

d. Pseudotensor interaction: $\Delta=\Delta_Y$:

$$\mathscr{S}_1\psi=\Big(\exp(Y\mathscr{O}_2)\,\theta_1\cdot p+\epsilon_1\ \mathrm{ch}(Y\mathscr{E}_4)\,\theta_1\cdot\hat{P}+\epsilon_2\ \mathrm{sh}(Y\mathscr{E}_4)\,\theta_2\cdot\hat{P}$$
$$+m_1\theta_{51}-i\ \exp(Y\mathscr{O}_2)\,\theta_2\cdot\frac{\partial Y}{2}\,\mathscr{E}_4\Big)\psi=0, \tag{3.14a}$$

$$\mathscr{S}_2=-\exp\Big(Y\mathscr{O}_2\Big)\,\theta_2\cdot p+\epsilon_2\ \mathrm{ch}(Y\mathscr{E}_4)\,\theta_2\cdot\hat{P}+\epsilon_1\ \mathrm{sh}(Y\mathscr{E}_4)\,\theta_1\cdot\hat{P}$$
$$+m_2\theta_{52}+i\ \exp(Y\mathscr{O}_2)\,\theta_1\cdot\frac{\partial Y}{2}\,\mathscr{E}_4\psi=0. \tag{3.14b}$$

Combining polar and axial interactions: We emphasize that the preceding expressions for $\mathscr{S}_i$, (3.3) and (3.6), are also suitable for the calculation of combinations of interactions: one finds the constraints for the combined polar interaction $\Delta_p=\Delta_L+\Delta_J+\Delta_{\mathscr{G}}+\Delta_{\mathscr{F}}$ using (3.3), and the constraints for the combined axial interaction $\Delta_a=\Delta_C+\Delta_H+\Delta_I+\Delta_Y$ using (3.6). The combination of all eight interactions (four polar plus four axial) is our desired result: the equivalent of Eqs. (3.7a)–(3.10b), for the polar interactions, and Eqs. (3.11a)–(3.14b), for the axial interactions, in two compact general Dirac equations. To combine all eight interaction we need to find a general relation, like (3.3) and (3.6), that is valid for the mixed symmetries of the general invariant matrix $\Delta=\Delta_a+\Delta_p$. For this purpose it is useful to define the matrices

$$\Delta_+\equiv\Delta_p+\Delta_a, \tag{3.15}$$

and

$$\Delta_- \equiv \Delta_p - \Delta_a . \tag{3.16}$$

To find the two-body Dirac equations $\mathscr{S}_1\psi=0$ and $\mathscr{S}_2\psi=0$ for the general *compound* interaction $\Delta=\Delta_+=\Delta_p+\Delta_a$, we use as in the previous derivations for the *separate* polar and axial interactions Δ_p and Δ_a the auxilary constraints, but with $\Delta=\Delta_+=\Delta_p+\Delta_a$:

$$\mathbf{S}_1\psi=[\mathscr{S}_{10}\ \mathrm{ch}\ \Delta_+ + \mathscr{S}_{20}\ \mathrm{sh}\ \Delta_+]\psi=0, \tag{3.17a}$$

$$\mathbf{S}_2\psi=[\mathscr{S}_{20}\ \mathrm{ch}\ \Delta_+ + \mathscr{S}_{10}\ \mathrm{sh}\ \Delta_+]\psi=0. \tag{3.17b}$$

We then insert these into the constrains (with $\Delta=\Delta_-=\Delta_p-\Delta_a$)

$$\mathscr{S}_1\psi=[\mathrm{ch}\ \Delta_-\mathbf{S}_1+\mathrm{sh}\ \Delta_-\mathbf{S}_2]\psi=0, \tag{3.18a}$$

$$\mathscr{S}_2\psi=[\mathrm{ch}\ \Delta_-\mathbf{S}_2+\mathrm{sh}\ \Delta_-\mathbf{S}_1]\psi=0, \tag{3.18b}$$

and expand using the identities

$$\mathrm{ch}\ \Delta_\pm=\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a \pm \mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a , \tag{3.19a}$$

$$\mathrm{sh}\ \Delta_\pm=\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a \pm \mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a . \tag{3.19b}$$

The result[24,14] is

$$\begin{aligned}
\mathscr{S}_1=\mathscr{S}_{10}&+i\theta_2\cdot\partial\Delta_+ - i\{[\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a,\theta_1^\mu]_-\mathrm{sh}\ \Delta_+ - [\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a,\theta_2^\mu]_-\ \mathrm{ch}\ \Delta_+\\
&-[\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a,\theta_1^\mu]_+\ \mathrm{sh}\ \Delta_+ + [\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a,\theta_2^\mu]_+\ \mathrm{ch}\ \Delta_+\}\partial_\mu\Delta_+\\
&-i\{[\mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a,\theta_2^\mu]_-\ \mathrm{sh}\ \Delta_+ - [\mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a,\theta_1^\mu]_-\ \mathrm{ch}\ \Delta_+ - [\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a,\theta_2^\mu]_+\ \mathrm{sh}\ \Delta_+\\
&+[\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a,\theta_1^\mu]_+\ \mathrm{ch}\ \Delta_+\}\partial_\mu\Delta_+ + \{\mathrm{ch}\ \Delta_-[\theta_{1\perp}^\mu,\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_-\\
&-\mathrm{ch}\ \Delta_-[\theta_{2\perp}^\mu,\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+ - \mathrm{sh}\ \Delta_-[\theta_{2\perp}^\mu,\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_-\\
&+\mathrm{sh}\ \Delta_-[\theta_{1\perp}^\mu,\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+\}p_\mu + \{\mathrm{ch}\ \Delta_-[\theta_{1\perp}^\mu,\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+\\
&-\mathrm{ch}\ \Delta_-[\theta_{2\perp}^\mu,\mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_- - \mathrm{sh}\ \Delta_-[\theta_{2\perp}^\mu,\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+\\
&+\mathrm{sh}\ \Delta_-[\theta_{1\perp}^\mu,\mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_-\}p_\mu + \mathrm{ch}\ \Delta_-[\epsilon_1\theta_1\cdot\hat{P},\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_- + \mathrm{ch}\ \Delta_-[\epsilon_2\theta_2\\
&\cdot\hat{P},\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+ + \mathrm{sh}\ \Delta_-[\epsilon_2\theta_2\cdot\hat{P},\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_- + \mathrm{sh}\ \Delta_-[\epsilon_1\theta_1\cdot\hat{P},\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+\\
&+\mathrm{ch}\ \Delta_-[\epsilon_1\theta_1\cdot\hat{P},\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+ + \mathrm{ch}\ \Delta_-[\epsilon_2\theta_2\cdot\hat{P}\ \mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_- + \mathrm{sh}\ \Delta_-[\epsilon_2\theta_2\\
&\cdot\hat{P},\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+ + \mathrm{sh}\ \Delta_-[\epsilon_1\theta_1\cdot\hat{P},\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_- + \mathrm{ch}\ \Delta_-[m_1\theta_{51},\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_-\\
&+\mathrm{ch}\ \Delta_-[m_2\theta_{52},\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+ + \mathrm{sh}\ \Delta_-[m_2\theta_{52},\mathrm{ch}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_-\\
&+\mathrm{sh}\ \Delta_-[m_1\theta_{51},\mathrm{sh}\ \Delta_p\ \mathrm{ch}\ \Delta_a]_+ + \mathrm{ch}\ \Delta_-[m_1\theta_{51},\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+\\
&+\mathrm{ch}\ \Delta_-[m_2\theta_{52},\mathrm{ch}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_- + \mathrm{sh}\ \Delta_-[m_2\theta_{52},\mathrm{sh}\ \Delta_p\ \mathrm{sh}\ \Delta_a]_+\\
&+\mathrm{sh}\ \Delta_-[m_1\theta_{51},\mathrm{ch}_p\ \mathrm{ch}\ \Delta_a]_- .
\end{aligned} \tag{3.20}$$

The calculation for $\mathscr{S}_2$ is entirely similar in method and result.[14]

We can now use the above expression (3.20) (and its analog for $\mathscr{S}_2$) to calculate the general hyperbolic constraint, $\mathscr{S}_1$ (and $\mathscr{S}_2$), for all eight interactions acting in concert, i.e., for the general invariant matrix,

$$\Delta = \Delta_+ = \underbrace{\Delta_L + \Delta_J + \Delta_{\mathscr{G}} + \Delta_{\mathscr{F}}}_{\Delta_p} + \underbrace{\Delta_C + \Delta_H + \Delta_I + \Delta_Y}_{\Delta_a}, \tag{3.21}$$

where

$$\Delta_p = \Delta_L + \Delta_J + \Delta_{\mathscr{G}} + \Delta_{\mathscr{F}}, \tag{3.22}$$

$$\Delta_a = \Delta_C + \Delta_H + \Delta_I + \Delta_Y . \tag{3.23}$$

The resulting complete hyperbolic two-body Dirac equations for all eight interaction acting together are[25]

$$\begin{aligned}\mathscr{S}_1\psi = \Big\{ &\exp(\mathscr{G} + \mathscr{F}\mathscr{E}_2 + I\mathscr{O}_1 + Y\mathscr{O}_2)\Big[\theta_1\cdot p - \frac{i}{2}\,\theta_2\cdot\partial(L\mathscr{O}_1 - J\mathscr{O}_2 - \mathscr{G}\mathscr{O}_3 - \mathscr{F}\mathscr{O}_4 - C\mathscr{E}_1 + H\mathscr{E}_2 \\ &+ I\mathscr{E}_3 + Y\mathscr{E}_4)\Big] + \epsilon_1\ \mathrm{ch}(J\mathscr{O}_2 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + Y\mathscr{E}_4)\,\theta_1\cdot\hat{P} \\ &+ \epsilon_2\ \mathrm{sh}(J\mathscr{O}_2 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + Y\mathscr{E}_4)\,\theta_2\cdot\hat{P} + m_1\ \mathrm{ch}(-L\mathscr{O}_1 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + I\mathscr{E}_3)\,\theta_{51} \\ &+ m_2\ \mathrm{sh}(-L\mathscr{O}_1 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + I\mathscr{E}_3)\,\theta_{52}\Big\}\psi = 0,\end{aligned} \tag{3.24a}$$

$$\begin{aligned}\mathscr{S}_2\psi = \Big\{ &-\exp(\mathscr{G} + \mathscr{F}\mathscr{E}_2 + I\mathscr{O}_1 + Y\mathscr{O}_2)\Big[\theta_2\cdot p - \frac{i}{2}\,\theta_1\cdot\partial(L\mathscr{O}_1 - J\mathscr{O}_2 - \mathscr{G}\mathscr{O}_3 \\ &- \mathscr{F}\mathscr{O}_4 - C\mathscr{E}_1 + H\mathscr{E}_2 + I\mathscr{E}_3 + Y\mathscr{E}_4)\Big] + \epsilon_1\ \mathrm{sh}(J\mathscr{O}_2 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + Y\mathscr{E}_4)\,\theta_1\cdot\hat{P} \\ &+ \epsilon_2\ \mathrm{ch}(J\mathscr{O}_2 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + Y\mathscr{E}_4)\,\theta_2\cdot\hat{P} + m_1\ \mathrm{sh}(-L\mathscr{O}_1 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + I\mathscr{E}_3)\,\theta_{51} \\ &+ m_2\ \mathrm{ch}(-L\mathscr{O}_1 + \mathscr{F}\mathscr{O}_4 + H\mathscr{E}_2 + I\mathscr{E}_3)\,\theta_{52}\Big\}\psi = 0.\end{aligned} \tag{3.24b}$$

Note the common structures in these ''external potential'' two-body Dirac equations: coordinate-dependent kinetic term, recoil terms (not present in the one body Dirac equation), timelike or energy potential terms (including the c.m. constituent energies), and scalar potential terms (including the constituent masses).

With these results the content of the two-body Dirac equations (3.7a)–(3.14a) has been condensed into Eqs. (3.24a) and (3.24b). One can easily generate all the relations (3.7a)–(3.14b) from (3.24a) and (3.24b) by imposing the appropriate limits: for example, letting all interactions vanish in (3.24a) and (3.24b) except C reproduces the pseudoscalar constraint equations (3.11a) and (3.11b). A similar limiting procedure (i.e., letting all interaction terms vanish except the one of interest) reproduces the other constraints of Eqs. (3.7a)–(3.14a), as it must for (3.24a) and (3.24b) to be valid. In the extremely long calculation (here omitted) that leads to the above two equations the interference terms between all of the interactions are of the form given here, involving exponentials and hyperbolic functions with arguments presented as sums of the arguments for the individual interactions.

In the simplified case of combined scalar, time- and spacelike vector, and pseudoscalar interactions, we use the hyperbolic parametrization introduced earlier plus two others,

$$M_1 = m_1\ \mathrm{ch}(L) + m_2\ \mathrm{sh}(L), \tag{3.25a}$$

$$M_2 = m_2\ \mathrm{ch}(L) + m_1\ \mathrm{sh}(L), \tag{3.25b}$$

$$E_1 = \epsilon_1\ \mathrm{ch}(J) + \epsilon_2\ \mathrm{sh}(J), \tag{3.26a}$$

$$E_2 = \epsilon_2\ \mathrm{ch}(J) + \epsilon_1\ \mathrm{sh}(J), \tag{3.26b}$$

$$G = e^{\mathscr{G}}, \tag{3.27}$$

where L, J, and $\mathscr{G}$ generate scalar, timelike vector and spacelike vector interactions, respectively, while C generates pseudoscalar interactions. The resultant ''external potential'' forms of the two-body Dirac equations with

$$\Delta = \Delta_J + \Delta_L + \Delta_{\mathscr{G}} + \Delta_C \tag{3.28}$$

are

$$\mathscr{S}_1\psi = \left(G\,\theta_1\cdot p + E_1\,\theta_1\cdot\hat{P} + M_1\,\theta_{51} + i\,\frac{G}{2}\,(\theta_2\cdot\partial\mathscr{G}\mathscr{O}_3 + \theta_2\cdot\partial J\mathscr{O}_2 - \theta_2\cdot\partial L\mathscr{O}_1 + \theta_2\cdot\partial C\mathscr{E}_1)\right)\psi = 0, \tag{3.29a}$$

$$\mathscr{S}_2\psi = \left(-G\,\theta_2\cdot p + E_2\,\theta_2\cdot\hat{P} + M_2\,\theta_{52} - i\,\frac{G}{2}\,(\theta_1\cdot\partial\mathscr{G}\mathscr{O}_3 + \theta_1\cdot\partial J\mathscr{O}_2 - \theta_1\cdot\partial L\mathscr{O}_1 + \theta_1\cdot\partial C\mathscr{E}_1)\right)\psi = 0. \tag{3.29b}$$

The scalar generator produces the mass or scalar potential M_i terms, the timelike vector generator produces the energy or timelike potential E_i terms, and the spacelike vector generator produces the transverse or spacelike momentum G terms, while the pseudoscalar generator produces only spin-dependent terms. Note that the vector and scalar interactions also have additional spin-dependent recoil terms essential for compatibility. The above features are just what one would expect for interactions transforming in this way and are direct consequences of the hyperbolic parametrization of the constraints.

We point out that the above equations (without the pseudoscalar interactions) have been used successfully in quark model calculations of the meson spectra.[11,16] Any realistic model, however, must ultimately include all eight types of interactions (it being recognized, of course, that the interactions are effective or semiphenomenological).

It is of interest and importance here to characterize the contributions of each of the eight different interactions. The advantage of writing the constraints in the ''external potential'' forms given in Eqs. (3.24a) and (3.24b) is that it forces the remaining potentials from the group into forms that one can interpret as spin-dependent combinations of the simpler scalar and vector potentials. We can see this most simply by examining the groupings of the different potentials in Eqs. (3.24a) and (3.24b). Consider first the timelike pseudovector interaction defined by the function $H(x_\perp)$ and the nature of its contributions, compared with the scalar and timelike vector contributions given in Eq. (3.29). Suppose all interactions are zero except for J, L, and H. Comparison of Eqs. (3.24a) and (3.24b) with Eqs. (3.12a), (3.12b), and (3.29) show us that the timelike pseudovector interaction H contributes both to the mass potential M_i terms as well as the energy or timelike vector potential E_i terms. Furthermore, since the matrix $\mathscr{E}_2$ is spin independent, so are those contributions.

The polar–tensor interaction defined by the function $\mathscr{F}(x_\perp)$ likewise contributes both to the mass potential M_i terms as well as the energy or timelike potential E_i terms. However, unlike with $H(x_\perp)$, since the matrix $\mathscr{O}_4$ is spin independent, so are the contributions to those terms. Note also that just like the spacelike vector generator $\mathscr{G}$, $\mathscr{F}$ produces spacelike vector-dependent potential terms that are momentum dependent. These same features are shared by $I(x_\perp)$, the spacelike pseudovector interaction, and $Y(x_\perp)$, the axial–tensor interaction, with the former contributing spin-dependent terms to the mass potential and the latter contributing spin-dependent terms to the timelike vector potential. Both interactions produces transverse momentum-dependent potential terms.

Later we shall compare the above expectations with the actual contributions to the various parts of the Schrödinger-like Pauli reduction.

IV. PAULI REDUCTION

We now use the two-body Dirac equations for ψ, (3.24a) and (3.24b), to derive the Schrödinger-like eigenvalue equation from the combined interactions: $L(x_\perp)$, $J(x_\perp)$, $H(x_\perp)$, $C(x_\perp)$, $\mathscr{G}(x_\perp)$, $\mathscr{F}(x_\perp)$, $I(x_\perp)$, $Y(x_\perp)$ for future application to the quark model. The basic method we will use has some similarities to the reduction of the single particle Dirac equation to a Schrödinger-like form (Pauli reduction). First we perform the matrix multiplications in (3.24a) and (3.24b) using the matrices $\mathscr{O}_1$, $\mathscr{O}_2$, $\mathscr{O}_3$, $\mathscr{O}_4$, $\mathscr{E}_1$, $\mathscr{E}_2$, $\mathscr{E}_3$, $\mathscr{E}_4$ to write out explicitly the resulting set of simultaneous equations. This is done in analogy to the Pauli reduction of the single-particle Dirac equation. We then solve the simultaneous equation set by a process of substitution and elimination, and in the process transform from a set of coupled first-order differential equations [for ψ_1, ψ_2, ψ_3, ψ_4 of Eq. (4.1)] to a single second-order Schrödinger-like eigenvalue equation for a newly defined wave function ϕ. This places the equations in a form that one can use most efficiently in quark-model calculations.

Crater and Van Alstine have carried and applied this procedure out for combined scalar, timelike vector, and spacelike vector interactions $(L,J,\mathscr{G})$.[15,16] The result in the general case was four coupled Schrödinger-like equations coupling two spin components of the upper-upper (ψ_1) and lower-lower (ψ_4) portions of the wave function. Mourad and Sazdjian[17] have carried out a similar procedure for these three interactions but starting with Eqs. (2.29a) and (2.29b) instead of the ''external potential'' forms of Eqs. (3.24a) and (3.24b). They further developed a technique to reduce the number of coupled equations in the general case from four to two, by using different combinations of the above wave functions. Our calculations below are for all eight interactions from the ''external potential'' forms of the equations and include a procedure analogous to that introduced in Ref. 17 but applied to Eqs. (3.24a) and (3.24b).

The wave function ψ appearing in the two-body Dirac equations (3.24a) and (3.24b) is a Dirac spinor written

$$\psi=\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\\ \psi_4\end{bmatrix}, \tag{4.1}$$

where each ψ_i is itself a four-component spinor. The wave function ψ has a total of 16 components (a result of the direct product multiplication of the Dirac spinors of particle one and particle two, each of which has four components) and the matrices $\mathscr{O}_i$'s, $\mathscr{E}_i$'s are all 16×16. In terms of the gamma matrices we use the Dirac representations with block forms

$$\beta_1=\begin{pmatrix}1_8 & 0\\ 0 & -1_8\end{pmatrix},\quad \gamma_{51}=\begin{pmatrix}0 & 1_8\\ 1_8 & 0\end{pmatrix},\quad \beta_1\gamma_{51}\equiv\rho_1=\begin{pmatrix}0 & 1_8\\ -1_8 & 0\end{pmatrix}, \tag{4.2a}$$

$$\beta_2=\begin{pmatrix}\beta & 0\\ 0 & \beta\end{pmatrix},\quad \beta=\begin{pmatrix}1_4 & 0\\ 0 & -1_4\end{pmatrix}, \tag{4.2b}$$

$$\gamma_{52}=\begin{pmatrix}\gamma_5 & 0\\ 0 & \gamma_5\end{pmatrix},\quad \gamma_5=\begin{pmatrix}0 & 1_4\\ 1_4 & 0\end{pmatrix}, \tag{4.2c}$$

$$\beta_2\gamma_{52}\equiv\rho_2=\begin{pmatrix}\rho & 0\\ 0 & \rho\end{pmatrix},\quad \rho=\begin{pmatrix}0 & 1_4\\ -1_4 & 0\end{pmatrix}, \tag{4.2d}$$

$$\beta_1\gamma_{51}\gamma_{52}=\begin{pmatrix}0 & \gamma_5\\ -\gamma_5 & 0\end{pmatrix},\quad \gamma_{51}\gamma_{52}=\begin{pmatrix}0 & \gamma_5\\ \gamma_5 & 0\end{pmatrix},\quad \beta_2\gamma_{52}\gamma_{51}=\begin{pmatrix}0 & \rho\\ \rho & 0\end{pmatrix}. \tag{4.2e}$$

$$\Sigma_i=\gamma_{5i}\beta_i\gamma_{\perp i},\quad i=1,2. \tag{4.2f}$$

The last matrices are four-vector generalizations of the Pauli matrices of particles one and two. In the c.m. frame the time component is zero and the spatial components are the usual Pauli matrices for each particle. In terms of these matrices we rewrite (3.24a) and (3.24b) by multiplying the first by $\sqrt{2}i\beta_1$ and the second by $\sqrt{2}i\beta_2$ yielding

$$\begin{aligned}&[T_1(\beta_1\beta_2)+U_1(\beta_1\beta_2)\gamma_{51}\gamma_{52}]\psi\\ &\quad=\mathrm{ch}(H-\beta_1\beta_2\Sigma_1\cdot\Sigma_2\mathscr{F})(E_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)+M_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)\beta_1)\gamma_{51}\psi\\ &\qquad+\mathrm{sh}(H-\beta_1\beta_2\Sigma_1\cdot\Sigma_2\mathscr{F})(E_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)+M_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)\beta_2)\gamma_{52}\psi,\end{aligned} \tag{4.3a}$$

$$\begin{aligned}&-[T_2(\beta_1\beta_2)+U_2(\beta_1\beta_2)\gamma_{51}\gamma_{52}]\psi\\ &\quad=\mathrm{ch}(H-\beta_1\beta_2\Sigma_1\cdot\Sigma_2\mathscr{F})(E_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)+M_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)\beta_2)\gamma_{52}\psi\\ &\qquad+\mathrm{sh}(H-\beta_1\beta_2\Sigma_1\cdot\Sigma_2\mathscr{F})(E_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)+M_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)\beta_1)\gamma_{51}\psi\end{aligned} \tag{4.3b}$$

in which the kinetic and recoil terms are

$$\begin{aligned}T_1(\beta_1\beta_2)=\exp(\mathscr{G}+\mathscr{F}\beta_1\beta_2)\Big[&\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_1\cdot p-\frac{i}{2}\beta_1\beta_2(\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_2\cdot\partial(H\beta_1\beta_2-C\\ &+(\mathscr{G}\beta_1\beta_2+\mathscr{F})\Sigma_1\cdot\Sigma_2)+\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_2\cdot\partial(J\beta_1\beta_2-L+(I\beta_1\beta_2+Y)\Sigma_1\cdot\Sigma_2))\Big],\end{aligned} \tag{4.4a}$$

$$\begin{aligned}T_2(\beta_1\beta_2)=\exp(\mathscr{G}+\mathscr{F}\beta_1\beta_2)\Big[&\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_2\cdot p-\frac{i}{2}\beta_1\beta_2(\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_1\cdot\partial(H\beta_1\beta_2-C\\ &+(\mathscr{G}\beta_1\beta_2+\mathscr{F})\Sigma_1\cdot\Sigma_2)+\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_1\cdot\partial(J\beta_1\beta_2-L+(I\beta_1\beta_2+Y)\Sigma_1\cdot\Sigma_2))\Big],\end{aligned} \tag{4.4b}$$

$$U_1(\beta_1\beta_2)=\exp(\mathscr{G}+\mathscr{F}\beta_1\beta_2)\Big[\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_1\cdot p-\frac{i}{2}\,\beta_1\beta_2(\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_2\cdot\partial(J\beta_1\beta_2-L$$
$$+(I\beta_1\beta_2+Y)\Sigma_1\cdot\Sigma_2)+\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_2\cdot\partial(H\beta_1\beta_2-C+(\mathscr{G}\beta_1\beta_2+\mathscr{F})\Sigma_1\cdot\Sigma_2))\Big], \tag{4.4c}$$

$$U_2(\beta_1\beta_2)=\exp(\mathscr{G}+\mathscr{F}\beta_1\beta_2)\Big[\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_2\cdot p-\frac{i}{2}\,\beta_1\beta_2(\mathrm{ch}(I+Y\beta_1\beta_2)\Sigma_1\cdot\partial(J\beta_1\beta_2-L$$
$$+(I\beta_1\beta_2+Y)\Sigma_1\cdot\Sigma_2)+\mathrm{sh}(I+Y\beta_1\beta_2)\Sigma_1\cdot\partial(H\beta_1\beta_2-C+(\mathscr{G}\beta_1\beta_2+\mathscr{F})\Sigma_1\cdot\Sigma_2))\Big], \tag{4.4d}$$

while the generalized (spin-dependent) timelike and scalar potentials are

$$E_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)=\epsilon_1\ \mathrm{ch}(J-\beta_1\beta_2\Sigma_1\cdot\Sigma_2Y)+\epsilon_2\ \mathrm{sh}(J-\beta_1\beta_2\Sigma_1\cdot\Sigma_2Y), \tag{4.5a}$$

$$E_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)=\epsilon_2\ \mathrm{ch}(J-\beta_1\beta_2\Sigma_1\cdot\Sigma_2Y)+\epsilon_1\ \mathrm{sh}(J-\beta_1\beta_2\Sigma_1\cdot\Sigma_2Y), \tag{4.5b}$$

$$M_1(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)=m_1\ \mathrm{ch}(L+\beta_1\beta_2\Sigma_1\cdot\Sigma_2I)+m_2\ \mathrm{sh}(L+\beta_1\beta_2\Sigma_1\cdot\Sigma_2I), \tag{4.5c}$$

$$M_2(\beta_1\beta_2\Sigma_1\cdot\Sigma_2)=m_2\ \mathrm{ch}(L+\beta_1\beta_2\Sigma_1\cdot\Sigma_2I)+m_1\ \mathrm{sh}(L+\beta_1\beta_2\Sigma_1\cdot\Sigma_2I). \tag{4.5d}$$

The end result of the matrix multiplications in (4.3a) and (4.3b) is a set of eight simultaneous equations for the Dirac spinors $\psi_1,\psi_2,\psi_3,\psi_4$: a set of four from the first of the constraint two-body Dirac equations $\mathscr{S}_1\psi=0$, and another set of four simultaneous equations from the second of the constraint two-body Dirac equations $\mathscr{S}_2\psi=0$. Explicitly, in an arbitrary frame, the result of the matrix calculation produces the eight simultaneous equations ($\Sigma_i^\mu\psi\to\sigma_i^\mu\psi_{1,2,3,4}$):

$$T_1(+1)\psi_1+U_1(+1)\psi_4=\mathrm{ch}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_1(\sigma_1\cdot\sigma_2)+M_1(\sigma_1\cdot\sigma_2))\psi_3$$
$$+\mathrm{sh}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_2(\sigma_1\cdot\sigma_2)+M_2(\sigma_1\cdot\sigma_2))\psi_2, \tag{4.6a}$$

$$T_1(-1)\psi_2+U_1(-1)\psi_3=\mathrm{ch}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_1(-\sigma_1\cdot\sigma_2)+M_1(-\sigma_1\cdot\sigma_2))\psi_4$$
$$+\mathrm{sh}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_2(-\sigma_1\cdot\sigma_2)-M_2(-\sigma_1\cdot\sigma_2))\psi_1, \tag{4.6b}$$

$$T_1(-1)\psi_3+U_1(-1)\psi_2=\mathrm{ch}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_1(-\sigma_1\cdot\sigma_2)-M_1(-\sigma_1\cdot\sigma_2))\psi$$
$$+\mathrm{sh}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_2(-\sigma_1\cdot\sigma_2)+M_2(-\sigma_1\cdot\sigma_2))\psi_4, \tag{4.6c}$$

$$T_1(+1)\psi_4+U_1(+1)\psi_1=\mathrm{ch}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_1(\sigma_1\cdot\sigma_2)-M_1(\sigma_1\cdot\sigma_2))\psi_2$$
$$+\mathrm{sh}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_2(\sigma_1\cdot\sigma_2)-M_2(\sigma_1\cdot\sigma_2))\psi_3, \tag{4.6d}$$

$$-T_2(+1)\psi_1-U_2(+1)\psi_4=\mathrm{ch}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_2(\sigma_1\cdot\sigma_2)+M_2(\sigma_1\cdot\sigma_2))\psi_2$$
$$+\mathrm{sh}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_1(\sigma_1\cdot\sigma_2)+M_1(\sigma_1\cdot\sigma_2))\psi_3, \tag{4.6e}$$

$$-T_2(-1)\psi_2-U_2(-1)\psi_3=\mathrm{ch}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_2(-\sigma_1\cdot\sigma_2)-M_2(-\sigma\cdot\sigma_2))\psi_1$$
$$+\mathrm{sh}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_1(-\sigma_1\cdot\sigma_2)+M_1(-\sigma_1\cdot\sigma_2))\psi_4, \tag{4.6f}$$

$$-T_2(-1)\psi_3-U_2(-1)\psi_2=\mathrm{ch}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_2(-\sigma_1\cdot\sigma_2)+M_2(-\sigma_1\cdot\sigma_2))\psi_4 + \mathrm{sh}(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_1(-\sigma_1\cdot\sigma_2)-M_1(-\sigma_1\cdot\sigma_2))\psi_1, \tag{4.6g}$$

$$-T_2(+1)\psi_4-U_2(+1)\psi_1=\mathrm{ch}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_2(\sigma_1\cdot\sigma_2)-M_2(\sigma_1\cdot\sigma_2))\psi_3 + \mathrm{sh}(H-\sigma_1\cdot\sigma_2\mathscr{F})(E_1(\sigma_1\cdot\sigma_2)-M_1(\sigma_1\cdot\sigma_2))\psi_2. \tag{4.6h}$$

We now reduce the above set of eight equations to a second-order Schrödinger-like equation by a process of substitution and elimination using the combination of the four Dirac-spinors given below:

$$\phi_\pm\equiv\psi_1\pm\psi_4, \tag{4.7a}$$

$$\chi_\pm\equiv\psi_2\pm\psi_3. \tag{4.7b}$$

Equation (4.6a) plus Eq. (4.6d) yields

$$D_1^{++}\phi_+=E_1^{++}(\sigma_1\cdot\sigma_2)\chi_+-M_1^{+-}(\sigma_1\cdot\sigma_2)\chi_-. \tag{4.8}$$

Equation (4.6e) plus Eq. (4.6h) yields

$$-D_2^{++}\phi_+=E_2^{++}(\sigma_1\cdot\sigma_2)\chi_++M_2^{+-}(\sigma_1\cdot\sigma_2)\chi_-. \tag{4.9}$$

Equation (4.6b) plus Eq. (4.6c) yields

$$D_1^{-+}\chi_+=E_1^{-+}(\sigma_1\cdot\sigma_2)\phi_+-M_1^{-+}(\sigma_1\cdot\sigma_2)\phi_-. \tag{4.10}$$

Equation (4.6b) minus Eq. (4.6c) yields

$$D_1^{--}\chi_-=-E_1^{--}(\sigma_1\cdot\sigma_2)\phi_-+M_1^{--}(\sigma_1\cdot\sigma_2)\phi_+. \tag{4.11}$$

with the following definitions for the kinetic-recoil terms:

$$D_1^{++}\equiv T_1(+1)+U_1(+1)=e^{(\mathscr{G}+\mathscr{F}+I+Y)}\Big[\sigma_1\cdot p-\frac{i}{2}\,\sigma_2\cdot\partial[H-C+J-L + (\mathscr{G}+\mathscr{F}+I+Y)\sigma_1\cdot\sigma_2]\Big], \tag{4.12a}$$

$$D_2^{++}\equiv T_2(+1)+U_2(+1)=e^{(\mathscr{G}+\mathscr{F}+I+Y)}\Big[\sigma_2\cdot p-\frac{i}{2}\,\sigma_1\cdot\partial[H-C+J-L + (\mathscr{G}+\mathscr{F}+I+Y)\sigma_1\cdot\sigma_2]\Big], \tag{4.12b}$$

$$D_1^{-+}\equiv T_1(-1)+U_1(-1)=e^{(\mathscr{G}-\mathscr{F}+I-Y)}\Big[\sigma_1\cdot p+\frac{i}{2}\,\sigma_2\cdot\partial[-H-C-J-L + (-\mathscr{G}+\mathscr{F}-I+Y)\sigma_1\cdot\sigma_2]\Big], \tag{4.12c}$$

$$D_1^{--} \equiv T_1(-1) - U_1(-1) = e^{(\mathscr{G}-\mathscr{F}-I+Y)}\Big[\sigma_1\cdot p + \frac{i}{2}\,\sigma_2\cdot\partial[-H-C+J+L + (-\mathscr{G}+\mathscr{F}+I-Y)\sigma_1\cdot\sigma_2]\Big]. \tag{4.12d}$$

Notice how the spacelike interactions $(\mathscr{G},\mathscr{F},I,Y)$ group together to produce extra spin-dependent recoil terms and coordinate-dependent multiplicative factors to the spacelike momentum p. (We anticipated this in the last section.) Likewise, the scalar, timelike vector, pseudoscalar, and timelike pseudovector interactions (L,J,C,H) group together. The (spin-dependent) potential terms appearing are

$$E_1^{++}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H-\mathscr{F}\sigma_1\cdot\sigma_2)E_1(\sigma_1\cdot\sigma_2) + \mathrm{sh}(H-\mathscr{F}\sigma_1\cdot\sigma_2)E_2(\sigma_1\cdot\sigma_2), \tag{4.13a}$$

$$M_1^{+-}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H-\mathscr{F}\sigma_1\cdot\sigma_2)M_1(\sigma_1\cdot\sigma_2) - \mathrm{sh}(H-\mathscr{F}\sigma_1\cdot\sigma_2)M_2(\sigma_1\cdot\sigma_2), \tag{4.13b}$$

$$E_2^{++}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H-\mathscr{F}\sigma_1\cdot\sigma_2)E_2(\sigma_1\cdot\sigma_2) + \mathrm{sh}(H-\mathscr{F}\sigma_1\cdot\sigma_2)E_1(\sigma_1\cdot\sigma_2), \tag{4.13c}$$

$$M_2^{+-}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H-\mathscr{F}\sigma_1\cdot\sigma_2)M_2(\sigma_1\cdot\sigma_2) - \mathrm{sh}(H-\mathscr{F}\sigma_1\cdot\sigma_2)M_1(\sigma_1\cdot\sigma_2), \tag{4.13d}$$

$$E_1^{-+}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H+\mathscr{F}\sigma_1\cdot\sigma_2)E_1(-\sigma_1\cdot\sigma_2) + \mathrm{sh}(H+\mathscr{F}\sigma_1\cdot\sigma_2)E_2(-\sigma_1\cdot\sigma_2), \tag{4.13e}$$

$$M_1^{-+}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H+\mathscr{F}\sigma_1\cdot\sigma_2)M_1(-\sigma_1\cdot\sigma_2) + \mathrm{sh}(H+\mathscr{F}\sigma_1\cdot\sigma_2)M_2(-\sigma_1\cdot\sigma_2), \tag{4.13f}$$

$$E_1^{--}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H+\mathscr{F}\sigma_1\cdot\sigma_2)E_1(-\sigma_1\cdot\sigma_2) - \mathrm{sh}(H+\mathscr{F}\sigma_1\cdot\sigma_2)E_2(-\sigma_1\cdot\sigma_2), \tag{4.13g}$$

$$M_1^{--}(\sigma_1\cdot\sigma_2) = \mathrm{ch}(H+\mathscr{F}\sigma_1\cdot\sigma_2)M_1(-\sigma_1\cdot\sigma_2) - \mathrm{sh}(H+\mathscr{F}\sigma_1\cdot\sigma_2)M_2(-\sigma_1\cdot\sigma_2). \tag{4.13h}$$

These, together with Eqs. (4.5), display explicitly the additional contributions discussed earlier of the four interactions $H,\mathscr{F},I,Y$ to the scalar and timelike potentials generated by L and J, respectively. At this level H and $\mathscr{F}$ contribute to both while Y contributes just to the timelike and I just to the scalar potentials.

Solve Eqs. (4.8) and (4.9) for χ_+ and χ_-:

$$\chi_+ = \frac{1}{\mathscr{D}(\sigma_1\cdot\sigma_2)}\,(M_2^{+-}(\sigma_1\cdot\sigma_2)D_1^{++} - M_1^{+-}(\sigma_1\cdot\sigma_2)D_2^{++})\phi_+, \tag{4.14a}$$

$$\chi_- = -\frac{1}{\mathscr{D}(\sigma_1\cdot\sigma_2)}\,(E_2^{++}(\sigma_1\cdot\sigma_2)D_1^{++} + E_1^{++}(\sigma_1\cdot\sigma_2)D_2^{++})\phi_+, \tag{4.14b}$$

in which

$$\mathscr{D}(\sigma_1\cdot\sigma_2) \equiv E_1^{++}(\sigma_1\cdot\sigma_2)M_2^{+-}(\sigma_1\cdot\sigma_2) + E_2^{++}(\sigma_1\cdot\sigma_2)M_1^{+-}(\sigma_1\cdot\sigma_2). \tag{4.15}$$

Solve Eqs. (4.10) and (4.11) for ϕ_+:

$$E_1^{--}(\sigma_1\cdot\sigma_2)D_1^{-+}\chi_+ - M_1^{-+}(\sigma_1\cdot\sigma_2)D_1^{--}\chi_- = \mathscr{B}^2(\sigma_1\cdot\sigma_2)\phi_+, \tag{4.16}$$

in which

$$\mathscr{B}^2(\sigma_1\cdot\sigma_2)\equiv E_1^{-+}(\sigma_1\cdot\sigma_2)E_1^{--}(\sigma_1\cdot\sigma_2)-M_1^{-+}(\sigma_1\cdot\sigma_2)M_1^{--}(\sigma_1\cdot\sigma_2). \tag{4.17}$$

Next combine Eq. (4.15) with Eqs. (4.14a) and (4.14b) to yield finally our four-component equation

$$\left[E_1^{--}(\sigma_1\cdot\sigma_2)D_1^{-+}\,\frac{1}{\mathscr{D}(\sigma_1\cdot\sigma_2)}\,(M_2^{+-}(\sigma_1\cdot\sigma_2)D_1^{++}-M_1^{+-}(\sigma_1\cdot\sigma_2)D_2^{++})+M_1^{-+}(\sigma_1\cdot\sigma_2)\right.$$
$$\left.\times D_1^{--}\,\frac{1}{\mathscr{D}(\sigma_1\cdot\sigma_2)}\,(E_2^{++}(\sigma_1\cdot\sigma_2)D_1^{++}+E_1^{++}(\sigma_1\cdot\sigma_2)D_2^{++})\right]\phi_+=\mathscr{B}^2(\sigma_1\cdot\sigma_2)\phi_+ . \tag{4.18}$$

Note that, in spite of appearances, as a consequence of the relativistic ''third law'' this expression is symmetrical in particles one and two. For example, using

$$\epsilon_1^2-\epsilon_2^2=m_1^2-m_2^2 \tag{4.19a}$$

$$\text{or}\quad \epsilon_1^2-m_1^2=\epsilon_2^2-m_2^2=b^2(w), \tag{4.19b}$$

we find that since

$$\begin{aligned}\mathscr{B}^2(\sigma_1\cdot\sigma_2)&=E_1^{-+}(\sigma_1\cdot\sigma_2)E_1^{--}(\sigma_1\cdot\sigma_2)-M_1^{-+}(\sigma_1\cdot\sigma_2)M_1^{--}(\sigma_1\cdot\sigma_2)\\&=\mathrm{ch}^2(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_1(-\sigma_1\cdot\sigma_2)^2-M_1(-\sigma_1\cdot\sigma_2)^2)\\&\quad-\mathrm{sh}^2(H+\sigma_1\cdot\sigma_2\mathscr{F})(E_2(-\sigma_1\cdot\sigma_2)^2-M_2(-\sigma_1\cdot\sigma_2)^2),\end{aligned} \tag{4.20}$$

we obtain the simplification

$$\begin{aligned}\mathscr{B}^2(\sigma_1\cdot\sigma_2)&=E_1(-\sigma_1\cdot\sigma_2)^2-M_1(-\sigma_1\cdot\sigma_2)^2\\&=E_2(-\sigma_1\cdot\sigma_2)^2-M_2(-\sigma_1\cdot\sigma_2)^2\\&=b^2(w)+(\epsilon_1^2+\epsilon_2^2)\mathrm{sh}^2(J+\sigma_1\cdot\sigma_2Y)+2\epsilon_1\epsilon_2\ \mathrm{sh}(J+\sigma_1\cdot\sigma_2Y)\mathrm{ch}(J+\sigma_1\cdot\sigma_2Y)\\&\quad-(m_1^2+m_2^2)\mathrm{sh}^2(L-\sigma_1\cdot\sigma_2I)-2m_1m_2\ \mathrm{sh}(L-\sigma_1\cdot\sigma_2I)\mathrm{ch}(L-\sigma_1\cdot\sigma_2I).\end{aligned} \tag{4.21}$$

Note that this nonderivative part of our effective potential for our four-component second-order equation (4.18) is independent of both the complicated $\mathscr{F}$ and H interactions and contains the covariant eigenvalue $b^2(w)$. Its spin dependence is in accord with our earlier expectations for spacelike pseudovector and axial tensor interactions. Also, the denominator term $\mathscr{D}(\sigma_1\cdot\sigma_2)$ in Eq. (4.15) is, like the right-hand side of Eq. (4.18) above, independent of $\mathscr{F}$ and H:

$$\begin{aligned}\mathscr{D}(\sigma_1\cdot\sigma_2)&=E_1(\sigma_1\cdot\sigma_2)M_2(\sigma_1\cdot\sigma_2)+E_2(\sigma_1\cdot\sigma_2)M_1(\sigma_1\cdot\sigma_2)\\&=(\epsilon_1m_2+\epsilon_2m_1)\mathrm{ch}(J+L+(I-Y)(\sigma_1\cdot\sigma_2))\\&\quad+(\epsilon_1m_1+\epsilon_2m_2)\mathrm{sh}(J+L+(I-Y)(\sigma_1\cdot\sigma_2)).\end{aligned} \tag{4.22}$$

In light of the definitions of $D_i^{\pm\pm}$, Eq. (4.18) is a second-order Schrödinger-like eigenvalue equation for the newly defined wave function ϕ_+:

$$(p_\perp^2+\Phi(r,\sigma_1\sigma_2,w))\phi_+=b^2(w)\phi_+\,. \tag{4.23}$$

(Note that in the c.m. system $p_\perp^2=\mathbf{p}^2$.)

In the simplified case of combined scalar, time- and spacelike vector, and pseudoscalar interactions, the spin dependence on the scalar and timelike potentials disappears. Equation (4.18) becomes

$$\left[E_1D_1^{-+}\,\frac{1}{E_1M_2+E_2M_1}\,(M_2D_1^{++}-M_1D_2^{++})+M_1D_1^{--}\,\frac{1}{E_1M_2+E_2M_1}\,(E_2D_1^{++}+E_1D_2^{++})\right]\phi_+$$
$$=(E_1^2-M_1^2)\phi_+ \tag{4.24}$$

with the kinetic-recoil terms in it reducing to terms

$$D_1^{++}=e^{\mathscr{G}}\left[\sigma_1\cdot p-\frac{i}{2}\,\sigma_2\cdot\partial[-C+J-L+\mathscr{G}\sigma_1\cdot\sigma_2]\right], \tag{4.25a}$$

$$D_2^{++}=e^{\mathscr{G}}\left[\sigma_2\cdot p-\frac{i}{2}\,\sigma_1\cdot\partial[-C+J-L+\mathscr{G}\sigma_1\cdot\sigma_2]\right], \tag{4.25b}$$

$$D_1^{-+}=e^{\mathscr{G}}\left[\sigma_1\cdot p+\frac{i}{2}\,\sigma_2\cdot\partial[-C-J-L-\mathscr{G}\sigma_1\cdot\sigma_2]\right], \tag{4.25c}$$

$$D_1^{--}=e^{\mathscr{G}}\left[\sigma_1\cdot p+\frac{i}{2}\,\sigma_2\cdot\partial[-C+J+L-\mathscr{G}\sigma_1\cdot\sigma_2]\right], \tag{4.25d}$$

while the timelike and scalar potentials reduce to

$$E_1=\epsilon_1\ \mathrm{ch}(J)+\epsilon_2\ \mathrm{sh}(J), \tag{4.26a}$$

$$E_2=\epsilon_2\ \mathrm{ch}(J)+\epsilon_1\ \mathrm{sh}(J), \tag{4.26b}$$

$$M_1=m_1\ \mathrm{ch}(L)+m_2\ \mathrm{sh}(L), \tag{4.26c}$$

$$M_2=m_2\ \mathrm{ch}(L)+m_1\ \mathrm{sh}(L). \tag{4.26d}$$

Note that the scalar potentials are the same as in Eq. (2.22).

We now proceed to find the form for Eqs. (4.18) displaying all the general spin-dependent structures in $\Phi(r,\sigma_1,\sigma_2,w)$ explicitly. We proceed to the general case of all eight interactions acting simultaneously and then present the results for the simpler set of potentials. We express Eq. (4.18) explicitly in terms of its matrix (σ_1,σ_2), and operator p structure in the c.m. system $[\hat{P}=(1,\mathbf{0})]$.

We use the following abbreviations for the kinetic portions by defining the scalar function h,c,n, and vector functions **k,e,q,z,d,o**:

$$h\equiv \mathrm{e}^{(\mathscr{G}+\mathscr{F}+I+Y)},\quad \mathbf{k}\equiv\tfrac{1}{2}\nabla\ \ln(h),\quad \mathbf{z}\equiv\tfrac{1}{2}\nabla(H-C+J-L), \tag{4.27a}$$

$$c\equiv \mathrm{e}^{(\mathscr{G}-\mathscr{F}+I-Y)},\quad \mathbf{e}\equiv\tfrac{1}{2}\nabla\ \ln(c),\quad \mathbf{d}\equiv\tfrac{1}{2}\nabla(H+C+J+L), \tag{4.27b}$$

$$n\equiv \mathrm{e}^{(\mathscr{G}-\mathscr{F}-I+Y)},\quad \mathbf{q}\equiv\tfrac{1}{2}\nabla\ \ln(n),\quad o\equiv\tfrac{1}{2}\nabla(H+C-J-L). \tag{4.27c}$$

We are working in the c.m. frame [i.e., $x_\perp=(\mathbf{r},0)$], so all the interaction functions $[L(x_\perp),J(x_\perp),H(x_\perp),C(x_\perp),\mathscr{G}(x_\perp),\mathscr{F}(x_\perp),I(x_\perp),Y(x_\perp)]$ are functions of $r,F=F(r)$. Using

$$\nabla F(r)=\frac{dF(r)}{dr}\,\hat{\mathbf{r}}\equiv F'(r)\hat{\mathbf{r}},\quad \text{where } \hat{\mathbf{r}} \text{ denotes the unit vector,} \tag{4.28}$$

we can express the vector terms (**k**,**e**,**q**,**z**,**d**,**o**) explicitly in terms of the unit vector $\hat{\mathbf{r}}$. With these newly defined terms the first-order differential operators of Eqs. (4.12a)–(4.12d) become

$$D_1^{++}\equiv h[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})], \tag{4.29a}$$

$$D_2^{++}\equiv h[\boldsymbol{\sigma_2}\cdot\mathbf{p}-i\boldsymbol{\sigma_1}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})], \tag{4.29b}$$

$$D_1^{-+}\equiv c[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{d}+\mathbf{e}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})], \tag{4.29c}$$

$$D_1^{--}\equiv n[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{o}+\mathbf{q}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]. \tag{4.29d}$$

We also define the following combinations involving the inverse matrix factors in Eq. (4.18):

$$F_1(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\equiv\frac{M_2^{+-}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}{\mathscr{D}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}, \tag{4.30a}$$

$$F_2(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\equiv\frac{M_1^{+-}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}{\mathscr{D}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}, \tag{4.30b}$$

$$F_3(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\equiv\frac{E_2^{++}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}{\mathscr{D}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}, \tag{4.30c}$$

$$F_4(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\equiv\frac{E_1^{++}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}{\mathscr{D}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}. \tag{4.30d}$$

Using these definitions, Eq. (4.18) becomes

$$\begin{aligned}
&c[E_1^{--}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{d}+\mathbf{e}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]hF_1(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\phi_+\\
&\quad+n[M_1^{-+}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{o}+\mathbf{q}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]hF_3(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\phi_+\\
&\quad-c[E_1^{--}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{d}+\mathbf{e}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]hF_2(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\\
&\quad\times[\boldsymbol{\sigma_2}\cdot\mathbf{p}-i\boldsymbol{\sigma_1}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\phi_++n[M_1^{-+}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{o}+\mathbf{q}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\\
&\quad\times hF_4(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})[\boldsymbol{\sigma_2}\cdot\mathbf{p}-i\boldsymbol{\sigma_1}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\phi_+=\mathscr{B}^2(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\phi_+\,.
\end{aligned} \tag{4.31}$$

For future reference we will refer to the four sets of additive terms on the left-hand side as (4.31a)–(4.31d).

Matrix wave functions: Because of the complicated spin dependence in the general case one must find as simple approach as possible for its explicit display. Simply working through the various identities involving products of powers of $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$, although straightforward, would involve 144 terms. A simpler approach is to simply write ϕ_+ as a linear combination of total spin wave functions and replace $\sigma_1\cdot\sigma_2$ by -3 for singlet and $+1$ for triplet states, thus facilitating considerably the reduction of the matrix form of Eq. (4.31) to the desired Schrödinger-like eigenvalue form Eq. (4.23). The multiplication properties of a matrix form for the wave functions with the Pauli matrices $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$, Eq. (4.33), instead of a spinor form, make them a particularly suitable choice. For this reason we transform the four-component spinor ϕ_+ into a 2×2 matrix wave function composed of a scalar spherical harmonic part (corresponding to spin zero) and a vector spherical harmonic part (corresponding to spin one) by the redefinition (in the c.m. system)

$$\phi_+ \rightarrow \phi_+ \sigma_y \equiv \Phi \equiv \phi + \boldsymbol{\phi}\cdot\boldsymbol{\sigma}. \tag{4.32}$$

The above definition includes a transpose and left multiplication by σ_y. This leads to the properties

$$\boldsymbol{\sigma}_1 \phi_+ \rightarrow \boldsymbol{\sigma}\Phi, \quad \boldsymbol{\sigma}_2 \phi_+ \rightarrow -\Phi\boldsymbol{\sigma}. \tag{4.33}$$

This in turn leads to

$$\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2 \phi_+ \rightarrow -\boldsymbol{\sigma}\cdot\Phi\boldsymbol{\sigma} = -3\phi + \boldsymbol{\phi}\cdot\boldsymbol{\sigma}, \tag{4.34}$$

which shows the singlet and triplet nature of the two parts of the matrix wave function. In general, since an arbitrary function $F(\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2)$ can be expanded into a form

$$F(\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2) = A + B\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2, \tag{4.35}$$

we have

$$\begin{aligned} F(\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2)\phi_+ &= (A + B\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2)\phi_+ \rightarrow [A\Phi - B\boldsymbol{\sigma}\cdot\Phi\boldsymbol{\sigma}] \\ &= (A-3B)\phi + (A+B)\boldsymbol{\phi}\cdot\boldsymbol{\sigma} \\ &= F(-3)\phi + F(+1)\boldsymbol{\phi}\cdot\boldsymbol{\sigma}. \end{aligned} \tag{4.36}$$

Using this identity the right-hand side of Eq. (4.31) becomes simply

$$\mathscr{B}^2(\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2)\phi_+ \rightarrow .\mathscr{B}^2(-3)\phi + \mathscr{B}^2(+1)\boldsymbol{\phi}\cdot\boldsymbol{\sigma}. \tag{4.37}$$

In computing the four sets of terms on left-hand side of Eq. (4.31), we use the identities

$$\boldsymbol{\sigma}_1\cdot\mathbf{A}\phi_+ \rightarrow \boldsymbol{\sigma}\cdot\mathbf{A}\Phi = \mathbf{A}\cdot\boldsymbol{\phi} + \boldsymbol{\sigma}\cdot[\mathbf{A}\phi + i\mathbf{A}\times\boldsymbol{\phi}], \tag{4.38a}$$

$$\boldsymbol{\sigma}_2\cdot\mathbf{A}\phi_+ \rightarrow -\mathbf{A}\cdot\Phi\boldsymbol{\sigma} = -\mathbf{A}\cdot\boldsymbol{\phi} - \boldsymbol{\sigma}\cdot[\mathbf{A}\phi - i\mathbf{A}\times\boldsymbol{\phi}]. \tag{4.38b}$$

The calculation of Eq. (4.31) then proceeds by reducing the left-hand side to a form like the right-hand side of (4.37), $(\mathrm{R} + \mathbf{R}\cdot\boldsymbol{\sigma})$, i.e., to a form $\mathrm{L} + \mathbf{L}\cdot\boldsymbol{\sigma}$ in which we call L the ''scalar,'' and **L** the ''vector'' term, respectively. We define the left-hand side of (4.31) in these terms as

$$\begin{aligned} \Pi\Phi = \mathscr{L} &= \mathrm{L} + \mathbf{L}\cdot\boldsymbol{\sigma} \equiv \text{left-hand side}(4.31), \\ &= (4.31\mathrm{a}) + (4.31\mathrm{b}) + (4.31\mathrm{c}) + (4.31\mathrm{d}). \end{aligned} \tag{4.39}$$

The reduction of (4.31) follows by equating the like terms on the right-hand (R and **R**, which we have calculated already) and left-hand sides (L and **L**, which remain to be calculated):

$$\mathrm{L}\phi + \mathbf{L}\cdot\boldsymbol{\sigma} = \mathrm{R}\phi + \mathbf{R}\cdot\boldsymbol{\sigma} \Rightarrow \left\{ \begin{array}{l} \mathrm{L} = \mathrm{R} \\ \mathbf{L} = \mathbf{R} \end{array} \right\}. \tag{4.40}$$

Having defined the left-hand side in the terms $\mathrm{L} + \mathbf{L}\cdot\boldsymbol{\sigma}$ we now proceed to calculate L and **L** explicitly (see Appendix A).

Using the expression derived in Appendix A for L, equating the ''scalar'' terms (L=R), bringing through the operator **p** in the various portions, performing the various scalar products and vector cross products, and converting back to the variables $J, L, C, H, \mathscr{G}, \mathscr{F}, I, Y, \mathscr{D}, \mathscr{B}$, we obtain the following expression for the Schrödinger-like ''scalar'' form L=R:

$$e^{(2\mathscr{G}+4\mathscr{F}-2H)}\frac{\mathscr{D}(3)}{\mathscr{D}(1)}\left[\{\mathbf{p}^2-i[2\mathscr{G}+4\mathscr{F}-2H-\ln(\mathscr{D}(1))-J+L+3I+3Y]'\hat{\mathbf{r}}\cdot\mathbf{p}+\tfrac{1}{2}\nabla^2(H-C+J-L-3\mathscr{G}-3\mathscr{F}-3I-3Y)-\tfrac{1}{4}(3H+C+J-L-\mathscr{G}-5\mathscr{F}-3I-3Y+2\ln[\mathscr{D}(1)])'(H-C+J-L-3\mathscr{G}-3\mathscr{F}-3I-3Y)'\}\phi+e^{(2H-2\mathscr{F}-I-Y+J-L)}\frac{w(m_1-m_2)}{\mathscr{D}(1)}(J+L+I-Y)'\frac{\mathbf{L}}{r}\cdot\phi\right]$$
$$=\mathscr{B}^2(3)\phi. \tag{4.41}$$

This equation has the form

$$A\left[\left\{\mathbf{p}^2-iB'\hat{\mathbf{r}}\cdot\mathbf{p}+\frac{1}{2}\nabla^2F-\frac{1}{4}K'F'\right\}\phi+N\frac{w(m_1-m_2)}{\mathscr{D}(1)}U'\frac{\mathbf{L}}{r}\cdot\phi\right]=\mathscr{B}^2(3)\phi, \tag{4.42}$$

in which

$$A=e^{(2\mathscr{G}+4\mathscr{F}-2H)}\frac{\mathscr{D}(3)}{\mathscr{D}(1)}, \tag{4.43a}$$

$$B=2\mathscr{G}+4\mathscr{F}-2H-\ln(\mathscr{D}(1))-J+L+3I+3Y, \tag{4.43b}$$

$$F=H-C+J-L-3\mathscr{G}-3\mathscr{F}-3I-3Y, \tag{4.43c}$$

$$K=3H+C+J-L-\mathscr{G}-5\mathscr{F}-3I-3Y+2\ln(\mathscr{D}(1)), \tag{4.43d}$$

$$N=e^{(2H-2\mathscr{F}-I-Y+J-L)}, \tag{4.43e}$$

$$U=J+L+I-Y. \tag{4.43f}$$

Performing a similar operation for the ''vector'' terms ($\mathbf{L}=\mathbf{R}$) we obtain

$$E\left[\left\{\mathbf{p}^2-\mathbf{pp}\cdot+i(Q-S)'[\hat{\boldsymbol{r}}\mathbf{p}\cdot-\hat{\boldsymbol{r}}\cdot\mathbf{p}]+\left[-Q'S'-\frac{3Q'}{r}+\nabla^2Q\right]\hat{\boldsymbol{r}}\hat{\boldsymbol{r}}\cdot-\nabla^2Q+\frac{Q'}{r}+S'Q'-i\frac{S'}{r}\mathbf{L}\right\}\boldsymbol{\phi}+\frac{T}{r}\frac{(\boldsymbol{\epsilon}_1-\boldsymbol{\epsilon}_2)(m_1+m_2)}{\mathscr{D}(1)}U'\mathbf{L}\boldsymbol{\phi}\right]$$
$$+O\left[\mathbf{pp}\cdot+iQ'\mathbf{p}\hat{\boldsymbol{r}}\cdot+iR'\hat{\boldsymbol{r}}\boldsymbol{p}\cdot+\left(\nabla^2Q-\frac{2Q'}{r}-Q'R'\right)\hat{\boldsymbol{r}}\hat{\boldsymbol{r}}\cdot\right]\boldsymbol{\phi}=\mathscr{B}^2(1)\boldsymbol{\phi} \tag{4.44}$$

in which

$$O=e^{(2\mathscr{G}-2H-4\mathscr{F})}\frac{\mathscr{D}(-1)}{\mathscr{D}(-3)}, \tag{4.45a}$$

$$E=e^{(2\mathscr{G}+2H)}\frac{\mathscr{D}(-1)}{\mathscr{D}(1)}, \tag{4.45b}$$

$$Q=\tfrac{1}{2}(H-C+J-L+\mathscr{G}+I+Y), \tag{4.45c}$$

$$R=\tfrac{1}{2}(3H+C+J-L-5\mathscr{G}+7\mathscr{F}+I+Y)+\ln(\mathscr{D}(-3)), \tag{4.45d}$$

$$S=-\tfrac{1}{2}(3H+C+J-L+3\mathscr{G}-\mathscr{F}+I+Y)+\ln\,(\mathscr{D}(1)), \tag{4.45e}$$

$$T=\mathrm{e}^{(-2H+2\mathscr{F}+L-J+I+Y)}\,\frac{\mathscr{D}(-1)}{\mathscr{D}(1)}. \tag{4.45f}$$

V. REDUCTION TO RADIAL FORM

A. Spin eigenfunctions

To reduce the forms of Eqs. (4.42) and (4.44) further, making them explicit differential equations in the variable r, we define the scalar ϕ and vector $\boldsymbol{\phi}$ wave functions explicitly in terms of their spatial (r,θ,φ) and spin dependence. For spectroscopic analysis we use eigenfunctions of $\mathbf{L}^2,\mathbf{S}^2,\mathbf{J}^2$, J_z where **L**, **S**, and **J** are the orbital, spin, and total angular momentum, respectively. For mesons, the spin eigenvalue can have the value zero or one only: for $s=0$ $j=l$, while for $s=1$ either $j=l$ or $j=l\pm1$. Although the spin states are classified as singlet ($s=0$) and triplet ($s=1$), the wave functions defined below will produce two sets of coupled equations: one set couples the singlet ($s=0$, $l=j$) to the triplet ($s=1$, $l=j$) spin states, the other set couples the two triplet ($s=1$, $l=j\pm1$) spin states.

$s=0,1$, $l=j$ *spin states:* The two $j=l$ ($s=0,1$) spin states have eigenfunctions of the form

$$\Phi=\langle\mathbf{r}|nlsj\rangle=\phi+\boldsymbol{\phi}\cdot\boldsymbol{\sigma}=\frac{u_n(r)_{l0j}}{r}\,\mathscr{Y}_{jm}(\varphi,\theta)+\frac{u_n(r)_{l1j}}{r}\,\mathbf{X}_{jm}(\varphi,\theta)\cdot\boldsymbol{\sigma},$$

$$\phi=\frac{u_n(r)_{l0j}}{r}\,\mathscr{Y}_{jm}(\varphi,\theta), \tag{5.1}$$

$$\boldsymbol{\phi}\cdot\boldsymbol{\sigma}=\frac{u_n(r)_{l1j}}{r}\,\mathbf{X}_{jm}(\varphi,\theta)\cdot\boldsymbol{\sigma},$$

in which n is the radial quantum number ($n=1$ for the ground state). Here the $\mathscr{Y}_{jm}$ are the familiar scalar spherical harmonic functions and the $\mathbf{X}_{jm}$ are vector spherical harmonic functions[26,27]

$$\mathbf{X}_{jm}=\frac{\mathbf{L}}{\sqrt{j(j+1)}}\,\mathscr{Y}_{jm},$$

$$\mathbf{L}=-i\mathbf{r}\times\nabla, \tag{5.2}$$

$$j=0,1,2,...;\quad m=-j,-j+1,...0...,j-1,j.$$

For the $s=0$, $l=j$ case the wave function is

$$\phi=\frac{u(r)_{j0j}}{r}\,\mathscr{Y}_{jm}(\phi,\theta), \tag{5.3}$$

while for the $s=1$, $l=j$ case the wave function is

$$\boldsymbol{\phi}=\frac{u(r)_{j1j}}{r}\,\mathbf{X}_{jm}. \tag{5.4}$$

$s=1$, $l=j\pm1$ *spin states:* The two $l=j\pm1$ ($s=1$) spin states have eigenfunctions of the form

$$\boldsymbol{\phi}=\boldsymbol{\phi}_+ + \boldsymbol{\phi}_- = \frac{u_{j+11j}(r)}{r}\,\mathbf{Y}_{jm+} + \frac{u_{j-11j}(r)}{r}\,\mathbf{Y}_{jm-}, \tag{5.5}$$

where

$$\mathbf{Y}_{jm+} = -\sqrt{\frac{j+1}{2j+1}}\left(\hat{\mathbf{r}} - i\frac{r\mathbf{p}}{j+1}\right)\mathscr{Y}_{jm}, \tag{5.6a}$$

$$\mathbf{Y}_{jm-} = \sqrt{\frac{j}{2j+1}}\left(\hat{\mathbf{r}} + i\,\frac{r\mathbf{p}}{j}\right)\mathscr{Y}_{jm} \tag{5.6b}$$

are respectively eigenfunctions of $\mathbf{L}^2$ with eigenvalues $(j+1)(j+2)$ and $j(j-1)$.

Having defined our wave functions explicitly in terms of their angular dependence we can now carry out the indicated differential operations to produce the coupled radial eigenequations. Some of the various operators and their effect on the wave functions ϕ, $\boldsymbol{\phi}$ are listed here for use in reducing the ''scalar'' wave equation to a coupled differential form:

$$\mathbf{p}^2\phi = \frac{1}{r}\left[-\frac{d^2}{dr^2} + \frac{j(j+1)}{r^2}\right]u(r)_{j0j}\mathscr{Y}_{jm}(\phi,\theta),$$

$$\hat{\mathbf{r}}\cdot\mathbf{p}\phi = -i\frac{d}{dr}\left(\frac{u(r)_{j0j}}{r}\right)\mathscr{Y}_{jm}(\phi,\theta),$$

$$\mathbf{L}\cdot\boldsymbol{\phi} = \frac{u(r)_{l1j}}{r}\sqrt{j(j+1)}\,\mathscr{Y}_{jm}. \tag{5.7}$$

We present a complete list of the operators and their effects on the wave functions ϕ, $\boldsymbol{\phi}$, $\boldsymbol{\phi}_+$, $\boldsymbol{\phi}_-$ in Appendix B.

B. The coupled radial eigenvalue equations

The results of carrying out the differential operations on the ''scalar'' and ''vector'' equations (4.42) and (4.44), and using the relations in Appendix B produces the desired relativistic radial eigenvalue equations for all eight combined interactions:

$s=0$, $j=l$:

$$A\left[\left\{-\frac{d^2}{dr^2} + \frac{j(j+1)}{r^2} - B'\left(\frac{d}{dr} - \frac{1}{r}\right) + \frac{1}{2}\nabla^2 F - \frac{1}{4}K'F'\right\}u_{j0j} + N\,\frac{w(m_1-m_2)}{\mathscr{D}(1)}\,U'\,\frac{\sqrt{j(j+1)}}{r}\,u_{j1j}\right] = \mathscr{B}^2(-3)u_{j0j}. \tag{5.8a}$$

$s=1$, $j=l$:

$$E\left[\left\{-\frac{d^2}{dr^2} + \frac{j(j+1)}{r^2} + (S-Q)'\,\frac{d}{dr} + 2\,\frac{Q'}{r} - \nabla^2 Q + S'Q'\right\}u_{j1j} + T\,\frac{(\epsilon_1-\epsilon_2)(m_1+m_2)}{\mathscr{D}(1)}\,U'\,\frac{\sqrt{j(j+1)}}{r}\,u_{j0j}\right] = \mathscr{B}^2(1)u_{j\mathbf{1}j}. \tag{5.8b}$$

Notice that these two equations decouple either for equal masses or for the ${}^1S_0$ states.

The other two coupled equations have opposite parity from these and do not decouple except for the $^3P_0$ states. (Below we use the abbreviations $u_\pm = u_{j\pm 11j}$).

$s=1$, $j=l+1$:

$$\left\{\left[E+\frac{j}{2j+1}(O-E)\right]\left(-\frac{d^2}{dr^2}+\frac{j(j-1)}{r^2}\right)+\left[E(S-Q)'+\frac{j}{2j+1}(E(Q-S)'+O(Q+R)')\right]\frac{d}{dr}\right.$$
$$+\frac{1}{2j+1}[Q'[E(-j^2+j+2)+O(j^2-2j)]-ES'j(j+1)-OR'j^2]\frac{1}{r}$$
$$\left.+\frac{1}{2j+1}[\nabla^2 Q(-(j+1)E+Oj)+Q'(ES'(j+1)-OR'j)]\right\}u_- + \frac{\sqrt{j(j+1)}}{2j+1}$$
$$\times\left\{(E-O)\left(-\frac{d^2}{dr^2}-(2j+1)\frac{d}{dr}-\frac{j^2-1}{r^2}\right)+[E(S-Q)'-O(R+Q)']\frac{d}{dr}\right.$$
$$+(E[(2-j)Q+S(1+j)]'+O[(2-j)Q-(j+1)R]')\frac{1}{r}$$
$$\left.-(E+O)\nabla^2 Q+Q'(OR'+ES')\right\}u_+ = \mathscr{B}^2(1)u_- . \tag{5.8c}$$

$s=1$, $j=l-1$:

$$\left\{\left[E+\frac{j+1}{2j+1}(O-E)\right]\left(-\frac{d^2}{dr^2}+\frac{(j+1)(j+2)}{r^2}\right)\right.$$
$$+\left[E(S-Q)'+\frac{j+1}{2j+1}(E(Q-S)'+O(Q+R)')\right]\frac{d}{dr}$$
$$+\frac{1}{2j+1}[Q'[E(j^2+3j)-O(j^2+4j+3)]+ES'j(j+1)+OR'(j+1)^2]\frac{1}{r}$$
$$\left.+\frac{1}{2j+1}[\nabla^2 Q(-jE+O(j+1))+Q'(ES'j-OR'(j+1))]\right\}u_+$$
$$+\frac{\sqrt{j(j+1)}}{2j+1}\left\{(E-O)\left(-\frac{d^2}{dr^2}+(2j+1)\frac{d}{dr}-\frac{j(j+2)}{r^2}\right)\right.$$
$$+[E(S-Q)'-O(R+Q)']\frac{d}{dr}+(E[(3+j)Q-Sj]'+O[(3+j)Q+jR]')\frac{1}{r}$$
$$\left.-(E+O)\nabla^2 Q+Q'(OR'+ES')\right\}u_- = \mathscr{B}^2(1)u_+ . \tag{5.8d}$$

In the special case of combined scalar, time-, and spacelike vector and pseudoscalar interactions [see Eq. (4.24)] we have

$$\mathscr{B}^2(-3)=\mathscr{B}^2(1)\equiv\mathscr{B}^2=E_1^2-M_1^2=E_2^2-M_2^2$$
$$=b^2(w)+(\epsilon_1^2+\epsilon_2^2)\mathrm{sh}^2(J)+2\epsilon_1\epsilon_2\mathrm{sh}(J)\mathrm{ch}(J)$$
$$-(m_1^2+m_2^2)\mathrm{sh}^2(L)+2m_1m_2\ \mathrm{sh}(L)\mathrm{ch}(L), \tag{5.9a}$$

$$\mathscr{D}(-3)=\mathscr{D}(-1)=\mathscr{D}(1)\equiv\mathscr{D}=E_1M_2+E_2M_1$$
$$=(\epsilon_1m_2+\epsilon_2m_1)\mathrm{ch}(J+L)+(\epsilon_1m_1+\epsilon_2m_1)\mathrm{sh}(J+L), \tag{5.9b}$$

$$A=O=E=e^{2\mathscr{G}}, \tag{5.9c}$$

$$B=2\mathscr{G}-\ln(\mathscr{D})-J+L, \tag{5.9d}$$

$$F=-C+J-L-3\mathscr{G}, \tag{5.9e}$$

$$K=C+J-L-\mathscr{G}+2\ln(\mathscr{D}), \tag{5.9f}$$

$$N=\frac{1}{T}=\mathrm{e}^{(J-L)}, \tag{5.9g}$$

$$U=J+L, \tag{5.9h}$$

$$Q=\tfrac{1}{2}(-C+J-L+\mathscr{G}), \tag{5.9i}$$

$$R=\tfrac{1}{2}(C+J-L-5\mathscr{G})+\ln(\mathscr{D}), \tag{5.9j}$$

$$S=-\tfrac{1}{2}(C+J-L+3\mathscr{G})+\ln(\mathscr{D}), \tag{5.9k}$$

so that the above equations become simply

$s=0,\ j=l$:

$$\left\{-\frac{d^2}{dr^2}+\frac{j(j+1)}{r^2}-(2\mathscr{G}-\ln(\mathscr{D})-J+L)'\left(\frac{d}{dr}-\frac{1}{r}\right)+\frac{1}{2}\nabla^2(-C+J-L-3\mathscr{G})\right.$$
$$\left.-\frac{1}{4}(C+J-L-\mathscr{G}+2\ln(\mathscr{D}))'(-C+J-L-3\mathscr{G})'\right\}u_{j0j}$$
$$+\mathrm{e}^{(J-L)}\frac{w(m_1-m_2)}{\mathscr{D}}(J+L)'\frac{\sqrt{j(j+1)}}{r}u_{j1j}=\mathscr{B}^2\mathrm{e}^{-2\mathscr{G}}u_{j0j}, \tag{5.10a}$$

coupled to $s=1,\ j=l$:

$$\left\{-\frac{d^2}{dr^2}+\frac{j(j+1)}{r^2}-(2\mathscr{G}+J-L-\ln(\mathscr{D}))'\,\frac{d}{dr}+\frac{(-C+J-L+\mathscr{G})'}{r}-\frac{1}{2}\,\nabla^2(-C+J-L+\mathscr{G})\right.$$
$$\left.+\frac{1}{4}\,(2\ln(\mathscr{D})-(C+J-L+3\mathscr{G}))'(-C+J-L+\mathscr{G})'\right\}u_{j1j}$$
$$+\mathrm{e}^{(L-J)}\,\frac{(\epsilon_1-\epsilon_2)(m_1+m_2)}{\mathscr{D}}\,(J+L)'\,\frac{\sqrt{j(j+1)}}{r}\,u_{j0j}=\mathscr{B}^2e^{-2\mathscr{G}}u_{j1j}\,, \tag{5.10b}$$

and $s=1$, $j=l+1$:

$$\left\{\left(-\frac{d^2}{dr^2}+\frac{j(j-1)}{r^2}\right)+\left[\ln(\mathscr{D})-2\mathscr{G}-\frac{1}{2j+1}\,(J-L)\right]'\,\frac{d}{dr}+\left[-j\,\ln(\mathscr{D})+\frac{1}{2j+1}\,((4j^2+j\right.\right.$$
$$\left.+1)\mathscr{G}+J-L+(j-1)C)\right]'\,\frac{1}{r}+\frac{1}{4}\,(C'^2-(J-L)'^2)+\frac{1}{2j+1}\left(-\frac{1}{2}\,\nabla^2(J-L+\mathscr{G}-C)\right.$$
$$\left.\left.+\mathscr{G}'\left(\frac{2j-3}{4}\,\mathscr{G}-J+L+C\right)'+\frac{1}{2}\,\ln'(\mathscr{D})(J-L-C)'\right)\right\}u_-+\frac{\sqrt{j(j+1)}}{2j+1}\left\{-[J-L]'\,\frac{d}{dr}\right.$$
$$+[(J-L)(1-2j)+3\mathscr{G}-3C]'\,\frac{1}{r}-2\nabla^2(J-L+\mathscr{G}-C)+(J-L+\mathscr{G}-C)'$$
$$\left.\times(\ln(\mathscr{D})-2\mathscr{G})'\right\}u_+=\mathscr{B}^2e^{-2\mathscr{G}}u_-\,, \tag{5.10c}$$

coupled to $s=1$, $j=l-1$:

$$\left\{\left(-\frac{d^2}{dr^2}+\frac{(j+1)(j+2)}{r^2}\right)+\left[\ln(\mathscr{D})-2\mathscr{G}+\frac{1}{2j+1}\,(J-L)\right]'\,\frac{d}{dr}+\left[(j+1)\ln(\mathscr{D})+\frac{1}{2j+1}\right.\right.$$
$$\left.-((4j^2+7j+4)\mathscr{G}+J-L-(j+2)C)\right]'\,\frac{1}{r}+\frac{1}{4}\,(C'^2-(J-L)'^2)+\frac{1}{2j+1}\left(\frac{1}{2}\,\nabla^2(J-L+\mathscr{G}-C)\right.$$
$$\left.\left.+\mathscr{G}'\left(\frac{2j+5}{4}\,\mathscr{G}+J-L-C\right)'-\frac{1}{2}\,\ln'(\mathscr{D})(J-L-C)'\right)\right\}u_++\frac{\sqrt{j(j+1)}}{2j+1}$$
$$\times\left\{-[J-L]'\,\frac{d}{dr}+[(J-L)(2j+3)+3\mathscr{G}-3C]'\,\frac{1}{r}-2\nabla^2(J-L+\mathscr{G}-C)\right.$$
$$\left.+(J-L+\mathscr{G}-C)'\,(\ln(\mathscr{D})-2\mathscr{G})'\right\}u_-=\mathscr{B}^2e^{-2\mathscr{G}}u_+\,. \tag{5.10d}$$

Equations (5.8a)–(5.8d) and their special cases above represent the end point of our calculations.

In summary, the two-body Dirac equations of constraint dynamics (3.18) below form a covariant compatible pair of equations.

$$\mathscr{S}_1\psi=[\mathrm{ch}\;\Delta_-\mathbf{S}_1+\mathrm{sh}\;\Delta_-\mathbf{S}_2]\psi=0, \tag{5.11a}$$

$$\mathscr{S}_2\psi=[\mathrm{ch}\;\Delta_-\mathbf{S}_2+\mathrm{sh}\;\Delta_-\mathbf{S}_1]\psi=0. \tag{5.11b}$$

(The conditions for compatibility eliminate the troublesome variables of relative time and energy in the c.m. frame.) Using the forms given in Eq. (3.17) below,

$$\mathbf{S}_1\psi=[\mathscr{S}_{10}\ \mathrm{ch}\ \Delta_+ +\mathscr{S}_{20}\ \mathrm{sh}\ \Delta_+]\psi=0, \tag{5.12a}$$

$$\mathbf{S}_2\psi=[\mathscr{S}_{20}\ \mathrm{ch}\ \Delta_+ +\mathscr{S}_{10}\ \mathrm{sh}\ \Delta_+]\psi=0, \tag{5.12b}$$

for the auxiliary constraints we obtained the results given in Eqs. (3.24a) and (3.24b) for the ''external potential'' form of the two-body Dirac equations for all eight combined covariant interactions. This generalizes earlier results for the eight interactions acting separately and opens the way toward further applications in the quark model of mesons and the two-body nucleon problem. Toward that end we derived the second-order forms of the equations Eqs. (4.18) and (4.31). We achieved significant simplification in the reduction of these forms by using matrix forms of the wave function and equations, leading ultimately to the four-component Schrödinger-like forms given by the coupled equations Eqs. (4.42) and (4.44). Finally, using vector and scalar spherical harmonics we obtained the radial forms above.

APPENDIX A: CALCULATION OF L AND L

First we write

$$\mathrm{L}_\phi+\mathbf{L}\cdot\boldsymbol{\phi}\equiv\Pi\Phi, \tag{A1}$$

where

$$\Pi\Phi=\Pi_1\Phi+\Pi_2\Phi+\Pi_3\Phi+\Pi_4\Phi, \tag{A2}$$

and $\Pi_1\Phi=$(4.31a), $\Pi_2\Phi=$(4.31b), $\Pi_3\Phi=$(4.31c), and $\Pi_4\Phi=$(4.31d). Notice that *each* Π_i is a product of four matrix factors

$$\Pi_i=[\pi_{i1}][\pi_{i2}][\pi_{i3}][\pi_{i4}]. \tag{A3}$$

Explicitly Π_1 is defined

$$\Pi_1=\underbrace{cE_1^{--}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}_{[\pi_{11}]}\underbrace{[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{d}+\mathbf{e}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]}_{[\pi_{12}]}\underbrace{F_1(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})}_{[\pi_{13}]}\underbrace{[\boldsymbol{\sigma_1}\cdot\mathbf{p}-i\boldsymbol{\sigma_2}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]}_{[\pi_{14}]},$$

$$=[\pi_{11}][\pi_{12}][\pi_{13}][\pi_{14}]. \tag{A4}$$

Due to the properties (4.33), each of the $[\pi_{ij}]$ terms acts on the matrix wave function Φ to produce yet another matrix form

$$[\pi_{ij}]\Phi=\mathscr{Z}_i=\mathrm{Z}_i+\mathbf{Z_i}\cdot\boldsymbol{\sigma}. \tag{A5}$$

The multiplication of each of the four $\Pi_i\Phi$ terms in Eq. (A2) can now proceed as shown in Eq. (A6): by Eq. (A5), each factor $[\pi_{ij}]$ will act on the matrix on its right to produce yet another matrix form

$$[\pi_{i1}][\pi_{i2}][\pi_{i3}][\pi_{i4}]\Phi=[\pi_{i1}][\pi_{i2}][\pi_{i3}]\mathscr{N}_i=[\pi_{i1}][\pi_{i2}]\mathscr{P}_i=[\pi_{i1}]\mathscr{Q}_i=\mathscr{L}_i=\mathrm{L}_i+\mathbf{L}_i\cdot\boldsymbol{\sigma}. \tag{A6}$$

We use the ''script'' notation $\mathscr{N}_i, \mathscr{P}_i, \mathscr{Q}_i, \mathscr{L}_i$ to denote the matrix form, i.e.,; $\mathscr{N}_i \equiv \mathrm{N}_i + \mathbf{N}_i \cdot \boldsymbol{\sigma}$, $\mathscr{P}_i \equiv \mathrm{P}_i + \mathbf{P}_i \cdot \boldsymbol{\sigma}$ an so on. For each of the four matrix multiplications ($\Pi_i \Phi$, $i = 1,2,3,4$) of Eq. (A2) we obtain a term of the form $\mathrm{L}_i + \mathbf{L}_i \cdot \boldsymbol{\sigma}$, $i = 1,2,3,4$. For the complete calculation of Eq. (A2)

$$\begin{aligned} \mathrm{R} &\equiv \mathrm{L}_1 + \mathrm{L}_2 + \mathrm{L}_3 + \mathrm{L}_4, \\ \mathbf{R} &\equiv \mathbf{L}_1 + \mathbf{L}_2 + \mathbf{L}_3 + \mathbf{L}_4, \end{aligned} \tag{A7}$$

Inspection of Eq. (4.31) reveals that the matrix structures of the terms (4.31a) and (4.31b) are the same. In fact (4.31a) becomes (4.31b) (and vice versa) if we make the variable exchanges

$$\begin{aligned} & c \leftrightarrow n, \quad E_1^{--} \leftrightarrow M_1^{-+}, \\ & \mathbf{d} \leftrightarrow \mathbf{o}, \quad \mathbf{e} \leftrightarrow \mathbf{q}, \\ & F_1 \leftrightarrow F_3. \end{aligned} \tag{A8}$$

This simplifies the calculation of Eq. (A2): we can calculate the first term $\Pi_1 \Phi$, which is (4.31a), and then find the second term $\Pi_2 \Phi$, which is (4.31b), by the variable exchanges of Eq. (A8)

Inspection also reveals that the matrix structure of (4.31c) is the same as (4.31d): in fact (4.31c) becomes (4.31d) (and vice versa) if we make the variable exchanges

$$\begin{aligned} & -c \leftrightarrow n, \quad E_1^{--} \leftrightarrow M_1^{-+}, \\ & \mathbf{d} \leftrightarrow \mathbf{o}, \quad \mathbf{e} \leftrightarrow \mathbf{q}, \\ & F_2 \leftrightarrow F_4. \end{aligned} \tag{A9}$$

By calculating the third term $\Pi_3 \Phi$, which is (4.31c), we find the forth term $\Pi_4 \Phi$, which is (4.31d), by making the variable exchanges of Eq. (A9)

In light of the simplification produced by the variable exchanges of relations (A8) and Eq. (A9), we proceed to calculate Eq. (A2): first we calculate $\Pi_1 \Phi = \mathscr{L}_1 = \mathrm{L}_1 + \mathbf{L}_1 \cdot \boldsymbol{\sigma}$, then use it to find $\Pi_2 \Phi = \mathscr{L}_2 = \mathrm{L}_2 + \mathbf{L}_2 \cdot \boldsymbol{\sigma}$ by the variable exchanges (A8),then we calculate $\Pi_3 \Phi = \mathscr{L}_3 = \mathrm{L}_3 + \mathbf{L}_3 \cdot \boldsymbol{\sigma}$ and use it to find $\Pi_4 \Phi = \mathscr{L}_4 = \mathrm{L}_4 + \mathbf{L}_4 \cdot \boldsymbol{\sigma}$ by the variable exchanges (A9).

Calculate $\mathscr{L}_1 = \mathrm{L}_1 + \mathbf{L}_1 \cdot \boldsymbol{\sigma}$:

$$[\pi_{11}][\pi_{12}][\pi_{13}][\pi_{14}]\Phi = [\pi_{11}][\pi_{12}][\pi_{13}]\mathscr{N}_1 = [\pi_{11}][\pi_{12}]\mathscr{P}_1 = [\pi_{11}]\mathscr{Q}_1 = \mathscr{L}_1 = \mathrm{L}_1 + \mathbf{L}_1 \cdot \boldsymbol{\sigma}. \tag{A10}$$

Starting with the right-most term $\mathscr{N}_1$ [it is understood that $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ act in accordance with Eq. (4.33)],

$$[\pi_{14}]\Phi = [\boldsymbol{\sigma_1} \cdot \mathbf{p} - i\boldsymbol{\sigma_2} \cdot (\mathbf{z} + \mathbf{k}\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2)]\Phi = \mathscr{N}_1 = \mathrm{N}_1 + \mathbf{N}_1 \cdot \boldsymbol{\sigma}, \tag{A11}$$

where N_1 and $\mathbf{N}_1$ are

$$\begin{aligned} N_1 &= [\mathbf{p} + i(\mathbf{z} + \mathbf{k})] \cdot \phi, \\ \mathbf{N}_1 &= [\mathbf{p} + i(\mathbf{z} - 3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z} + \mathbf{k})] \times \phi, \end{aligned} \tag{A12}$$

$$[\pi_{13}]\mathscr{N}_1 = hF_1(\boldsymbol{\sigma_1} \cdot \boldsymbol{\sigma_2})\mathscr{N}_1 = \mathscr{P}_1 = P_1 + \mathbf{P}_1 \cdot \boldsymbol{\sigma}; \tag{A13}$$

P_1 and $\mathbf{P}_1$ are

$$\mathrm{P}_1 = hF_1(-3)N_1, \quad \mathbf{P}_1 = hF_1(+1)\mathbf{N}_1, \tag{A14}$$

$$[\pi_{12}]\mathscr{P}_1 = [\boldsymbol{\sigma_1}\cdot\mathbf{p} - i\boldsymbol{\sigma_2}\cdot(\mathbf{d}+\mathbf{e}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\mathscr{P}_1 = \mathscr{Q}_1 = \mathrm{Q}_1 + \mathbf{Q}_1\cdot\boldsymbol{\sigma}; \tag{A15}$$

and Q_1 and $\mathbf{Q}_1$ are

$$\mathrm{Q}_1 = [\mathbf{p} + i(\mathbf{d}+\mathbf{e})]\cdot\mathbf{P}_1,$$

$$\mathbf{Q}_1 = [\mathbf{p} + i(\mathbf{d}-3\mathbf{e})]P_1 + i[\mathbf{p} - i(\mathbf{d}+\mathbf{e})]\times\mathbf{P}_1, \tag{A16}$$

$$[\pi_{11}]\mathscr{Q}_1 = cE_1^{--}(\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})\mathscr{Q}_1 = \mathscr{L}_1 = \mathrm{L}_1 + \mathbf{L}_1\cdot\boldsymbol{\sigma}, \tag{A17}$$

where finally L_1 and $\mathbf{L}_1$ are

$$\mathrm{L}_1 = cE_1^{--}(-3)Q_1, \quad \mathbf{L}_1 = cE_1^{--}(+1)\mathbf{Q}_1, \tag{A18}$$

Combining the multiplicative factors, the ''scalar'' part of $\mathscr{L}_1$ is therefore

$$\mathrm{L}_1 = cE_1^{--}(-3)[\mathbf{p} + i(\mathbf{d}+\mathbf{e})]\cdot hF_1(+1)\{[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}, \tag{A19}$$

while the ''vector'' part of $\mathscr{L}_1$ is

$$\begin{aligned}\mathbf{L}_1 = cE_1^{--}(+1)\{&[\mathbf{p} + i(\mathbf{d}-3\mathbf{e})]hF_1(-3)[\mathbf{p} + i(\mathbf{z}+\mathbf{k})]\cdot\boldsymbol{\phi} + i[\mathbf{p} - i(\mathbf{d}+\mathbf{e})]\\ &\times hF_1(+1)\{[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}\}.\end{aligned} \tag{A20}$$

The ''scalar'' and ''vector'' parts of $\mathscr{L}_2$ are now easily written down by inspection when the variable exchanges of Eq. (A8) are made on L_2 and $\mathbf{L}_2$ of Eqs. (A19) and (A20). We obtain

$$\mathrm{L}_2 = nM_1^{-+}(-3)[\mathbf{p} + i(\mathbf{o}+\mathbf{q})]\cdot hF_3(+1)\{[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}, \tag{A21}$$

while the ''vector'' part of $\mathscr{L}_2$ is

$$\begin{aligned}\mathbf{L}_2 = nM_1^{-+}(+1)\{&[\mathbf{p} + i(\mathbf{o}-3\mathbf{q})]hF_3(-3)[\mathbf{p} + i(\mathbf{z}+\mathbf{k})]\cdot\boldsymbol{\phi} + i[\mathbf{p} - i(\mathbf{o}+\mathbf{q})]\\ &\times hF_3(+1)\{[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}\}.\end{aligned} \tag{A22}$$

For the scalar and vector parts of $\mathscr{L}_3$ things go very much like for $\mathscr{L}_1$ except for the first step we use Eq. (4.38b) instead of Eq. (4.38a),

$$[\pi_{34}]\Phi = [\boldsymbol{\sigma_2}\cdot\mathbf{p} - i\boldsymbol{\sigma_1}\cdot(\mathbf{z}+\mathbf{k}\boldsymbol{\sigma_1}\cdot\boldsymbol{\sigma_2})]\Phi = \mathscr{N}_3 = \mathrm{N}_3 + \mathbf{N}_3\cdot\boldsymbol{\sigma}, \tag{A23}$$

where N_3 and $\mathbf{N}_3$ are

$$\begin{aligned}\mathrm{N}_3 &= -[\mathbf{p} + i(\mathbf{z}+\mathbf{k})]\cdot\boldsymbol{\phi},\\ \mathbf{N_3} &= -[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi + i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}.\end{aligned} \tag{A24}$$

Folding this into our previous results we obtain

$$\mathrm{L}_3 = cE_1^{--}(-3)[\mathbf{p} + i(\mathbf{d}+\mathbf{e})]\cdot hF_2(+1)\{[\mathbf{p} + i(\mathbf{z}-3\mathbf{k})]\phi - i[\mathbf{p} - i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}, \tag{A25}$$

while the ''vector'' part of $\mathscr{L}_3$ is

$$\mathbf{L}_3 = cE_1^{--}(+1)\{[\mathbf{p}+i(\mathbf{d}-3\mathbf{e})]hF_2(-3)[\mathbf{p}+i(\mathbf{z}+\mathbf{k})]\cdot\boldsymbol{\phi}+i[\mathbf{p}-i(\mathbf{d}+\mathbf{e})] \times hF_2(+1)\{[\mathbf{p}+i(\mathbf{z}-3\mathbf{k})]\phi-i[\mathbf{p}-i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}\}. \tag{A26}$$

From this we determine the scalar and vector parts of $\mathscr{S}_4$ by making the variable exchanges of Eq. (A9):

$$L_4 = -nM_1^{-+}(-3)[\mathbf{p}+i(\mathbf{o}+\mathbf{q})]\cdot hF_4(+1)\{[\mathbf{p}+i(\mathbf{z}-3\mathbf{k})]\phi-i[\mathbf{p}-i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}, \tag{A27}$$

while the ''vector'' part of $\mathscr{S}_4$ is

$$\mathbf{L}_4 = -nM_1^{-+}(+1)\{[\mathbf{p}+i(\mathbf{o}-3\mathbf{q})]hF_4(-3)[\mathbf{p}+i(\mathbf{z}+\mathbf{k})]\cdot\boldsymbol{\phi}+i[\mathbf{p}-i(\mathbf{o}+\mathbf{q})]\times hF_4(+1) \times\{[\mathbf{p}+i(\mathbf{z}-3\mathbf{k})]\phi-i[\mathbf{p}-i(\mathbf{z}+\mathbf{k})]\times\boldsymbol{\phi}\}\}. \tag{A28}$$

Combining we obtain

$$\begin{aligned} L_1+L_2+L_3+L_4 = {} & h[cE_1^{--}(-3)(F_1(+1)+F_2(+1))+nM_1^{-+}(-3)(F_3(+1)-F_4(+1))]\mathbf{p}\cdot(\mathbf{p} \\ & +i(\mathbf{z}-3\mathbf{k}))\phi+i\{cE_1^{--}(-3)[h(F_1(+1)+F_2(+1))(\mathbf{d}+\mathbf{e})-\nabla(h(F_1(+1) \\ & +F_2(+1)))]+nM_1^{-+}(-3)[h(\mathbf{F}_3(+1)-F_4(+1))(\mathbf{o}+\mathbf{q})-\nabla(h(F_3(+1) \\ & -F_4(+1)))]\}\cdot(\mathbf{p}+i(\mathbf{z}-3\mathbf{k}))\phi ih[cE_1^{--}(-3)(F_1(+1)-F_2(+1))+nM_1^{-+} \\ & \times(-3)(\mathbf{F}_3(+1)+F_4(+1))]\mathbf{p}\cdot(\mathbf{p}-i((\mathbf{z}+\mathbf{k}))\times\boldsymbol{\phi}-\{cE_1^{--}(-3)[h(\mathbf{F}_1(+1) \\ & -F_2(+1))(\mathbf{d}+\mathbf{e})-\nabla(h(F_1(+1)-F_2(+1))]+nM_1^{-+}(-3)[h(F_3(+1) \\ & +F_4(+1))(\mathbf{o}+\mathbf{q})-\nabla(h(F_3(+1)+F_4(+1)))]\}\cdot(\mathbf{p}-i(\mathbf{z}+\mathbf{k}))\times\boldsymbol{\phi} \end{aligned} \tag{A29}$$

for the scalar portion.

We simplify this, leading to the result Eq. (4.42) given in the text, by using identities such as the following involving the scalar and timelike potentials:

$$E_1^{--}(-3) = \mathrm{ch}(H-3\mathscr{F})E_1(+3)-\mathrm{sh}(H-3\mathscr{F})E_2(+3), \tag{A30a}$$

$$M_1^{-+}(-3) = \mathrm{ch}(H-3\mathscr{F})M_1(+3)+\mathrm{sh}(H-3\mathscr{F})M_2(+3), \tag{A30b}$$

$$\begin{aligned} hc(F_1(+1)+F_2(+1)) &= \frac{1}{\mathscr{D}(1)}\,\mathrm{e}^{(2\mathscr{G}+2I)}\mathrm{e}^{(-H+\mathscr{F})}(M_1(1)+M_2(1)) \\ &= \mathrm{e}^{(2\mathscr{G}-H+\mathscr{F})}(M_1(3)+M_2(3)), \end{aligned} \tag{A30c}$$

$$hn(F_3(+1)-F_4(+1)) = \frac{1}{\mathscr{D}(1)}\,\mathrm{e}^{(2\mathscr{G}+2Y)}\mathrm{e}^{(-H+\mathscr{F})}(E_2(1)-E_1(1)) = \mathrm{e}^{(2\mathscr{G}-H+\mathscr{F})}(E_2(3)-E_1(3)) \tag{A30d}$$

$$\begin{aligned} \mathbf{d}+\mathbf{e}-\frac{\nabla h}{h}-\frac{\nabla(F_1(1)+F_2(1))}{F_1(1)+F_2(1)} &= \mathbf{o}+\mathbf{q}-\frac{\nabla h}{h}-\frac{\nabla(F_3(1)-F_4(1))}{F_3(1)-F_4(1)} \\ &= \frac{\hat{\mathbf{r}}}{2}(3H+C+J-L-\mathscr{G}-5\mathscr{F}-3I-3Y)'+\nabla\,\ln(\mathscr{D}(1)). \end{aligned} \tag{A30e}$$

We can simplify the vector portion below:

$$\begin{aligned}\mathbf{L_1}+\mathbf{L_2}+\mathbf{L_3}+\mathbf{L_4}=&h[cE_1^{--}(+1)(F_1(-3)+F_2(-3))+nM_1^{-+}(+1)(F_3(-3)-F_4(-3))]\mathbf{p}(\mathbf{p}\\&+i(\mathbf{z}+\mathbf{k}))\cdot\boldsymbol{\phi}+i\{cE_1^{--}(+1)[h(F_1(+1)+F_2(+1))(\mathbf{d}-\mathbf{3e})-\nabla(h(F_1(-3)\\&+F_2(-3)))]+nM_1^{-+}(+1)[h(F_3(+1)-F_4(+1))(\mathbf{o}-\mathbf{3q})-\nabla(h(F_3(-3)\\&-F_4(-3)))]\}(\mathbf{p}+i(\mathbf{z}+\mathbf{k}))\cdot\boldsymbol{\phi}ih[cE_1^{--}(+1)(F_1(+1)-F_2(+1))\\&+nM_1^{-+}(+1)(F_3(+1)+F_4(+1))]\mathbf{p}\times(\mathbf{p}+i(\mathbf{z}-\mathbf{3k}))\phi\}+\{[cE_1^{--}(+1)\\&\times[h(F_1(+1)+F_2(+1))(\mathbf{d}+\mathbf{e})+\nabla(h(F_1(+1)+F_2(+1)))]+nM_1^{-+}(+1)\\&\times[h(F_3(+1)-F_4(+1))(\mathbf{o}+\mathbf{q})+\nabla(h(F_3(+1)-F_4(+1)))]\}\\&\times(\mathbf{p}+i(\mathbf{z}-\mathbf{3k}))\phi-h[cE_1^{--}(+1)(F_1(+1)-F_2(+1))\\&+nM_1^{-+}(+1)(F_3(+1)+F_4(+1))]\mathbf{p}\times[((\mathbf{p}-i(\mathbf{z}+\mathbf{k}))\times\boldsymbol{\phi}]+i\{cE_1^{--}(+1)\\&\times[h(F_1(+1)-F_2(+1))(\mathbf{d}+\mathbf{e})+\nabla(h(F_1(+1)-F_2(+1)))]+nM_1^{-+}(+1)\\&\times[h(F_3(+1)+F_4(+1))(\mathbf{o}+\mathbf{q})+\nabla(h(F_3(+1)+F_4(+1)))]\}\\&\times(\mathbf{p}-i(\mathbf{z}+\mathbf{k}))\times\boldsymbol{\phi},\end{aligned}\tag{A31}$$

using similar identities

$$E_1^{--}(+1)=\mathrm{ch}(H+\mathscr{F})E_1(-1)-\mathrm{sh}(H+\mathscr{F})E_2(-1),\tag{A32a}$$

$$M_1^{-+}(+1)=\mathrm{ch}(H+\mathscr{F})M_1(-1)+\mathrm{sh}(H+\mathscr{F})M_2(-1),\tag{A32b}$$

$$\begin{aligned}hc(F_1(-3)+F_2(-3))&=\frac{\mathrm{e}^{(2\mathscr{G}+2I)}}{\mathscr{D}(-3)}\,\mathrm{e}^{(-H-3\mathscr{F})}(M_1(-3)+M_2(-3))\\&=\frac{e^{(2\mathscr{G}-H-3\mathscr{F})}}{\mathscr{D}(-3)}\,(M_1(-1)+M_2(1-)),\end{aligned}\tag{A32c}$$

$$\begin{aligned}hn(F_3(-3)-F_4(-3))&=\frac{\mathrm{e}^{(2\mathscr{G}+2Y)}}{\mathscr{D}(-3)}\,\mathrm{e}^{(-H-3\mathscr{F})}(E_2(-3)-E_1(-3))\\&=\frac{\mathrm{e}^{(2\mathscr{G}-H-3\mathscr{F})}}{\mathscr{D}(-3)}\,(E_2(-1)-E_1(-1)),\end{aligned}\tag{A32d}$$

$$\begin{aligned}\mathbf{d}-\mathbf{3e}-\frac{\nabla h}{h}-\frac{\nabla(F_1(-3)+F_2(-3))}{F_1(-3)+F_2(-3)}&=\mathbf{o}-\mathbf{3q}-\frac{\nabla h}{h}-\frac{\nabla(F_3(-3)-F_4(-3))}{F_3(-3)-F_4(1)}\\&=\frac{\hat{\boldsymbol{r}}}{2}\,(3H+C+J-L-5\mathscr{G}+7\mathscr{F}+I+Y)'\\&\quad+\nabla\,\ln(\mathscr{D}(-3)),\end{aligned}\tag{A32e}$$

$$hc(F_1(+1)-F_2(+1))=\frac{e^{(2\mathscr{G}+2I)}}{\mathscr{D}(+1)}\,e^{(-H+\mathscr{F})}(M_2(+1)-M_1(+1))$$
$$=\frac{e^{(2\mathscr{G}-H+\mathscr{F})}}{\mathscr{D}(+1)}\,(M_2(-1)-M_2(1-)), \tag{A32f}$$

$$hn(F_3(+1)+F_4(+1))=\frac{e^{(2\mathscr{G}+2Y)}}{\mathscr{D}(+1)}\,e^{(-H+\mathscr{F})}(E_2(+1)+E_1(+1))$$
$$=\frac{e^{(2\mathscr{G}-H+\mathscr{F})}}{\mathscr{D}(1)}\,(E_2(-1)+E_1(-1)), \tag{A32g}$$

$$\mathbf{d}+\mathbf{e}+\frac{\nabla h}{h}+\frac{\nabla(F_1(+1)-F_2(+1))}{F_1(+1)+F_2(+1)}=\mathbf{o}+\mathbf{q}+\frac{\nabla h}{h}+\frac{\nabla(F_3(+1)+F_4(+1))}{F_3(+1)+F_4(+1)}$$
$$=\frac{\hat{\boldsymbol{r}}}{2}\,(3H+C+J-L+3\mathscr{G}-\mathscr{F}+I+Y)'-\nabla\,\ln(\mathscr{D}(+1)), \tag{A32h}$$

leading to

$$e^{2\mathscr{G}-2H-4\mathscr{F}}\,\frac{\mathscr{D}(-1)}{\mathscr{D}(-3)}\Bigg[\mathbf{pp}\cdot\boldsymbol{\phi}+\frac{i}{2}\,\mathbf{p}(H-C+J-L+\mathscr{G}+\mathscr{F}+I+Y)'\hat{\boldsymbol{r}}\cdot\boldsymbol{\phi}$$
$$+\frac{i}{2}\,[3H+C+J-L-5\mathscr{G}+7\mathscr{F}+I+Y$$
$$+2\ln(\mathscr{D}(-3))]'\hat{\boldsymbol{r}}(p+(H-C+J-L+\mathscr{G}+\mathscr{F}+I+Y)'\hat{\boldsymbol{r}})\cdot\phi\Bigg]$$
$$-e^{2\mathscr{G}+2H}\,\frac{\mathscr{D}(-1)}{\mathscr{D}(1)}\left\{\mathbf{p}\times\left(\left(\mathbf{p}-\frac{i}{2}\,(H-C+J-L+\mathscr{G}+\mathscr{F}+I+Y)'\hat{r}\right)\times\phi\right)\right.$$
$$+\frac{i}{2}\,[3H+C+J-L+3\mathscr{G}-\mathscr{F}+I+Y+2\ln(\mathscr{D}(1))]'\hat{\boldsymbol{r}}$$
$$\times\left(\left(\mathbf{p}-\frac{i}{2}\,(H-C+J-L+\mathscr{G}+\mathscr{F}+I+Y)'\hat{\boldsymbol{r}}\right)\times\boldsymbol{\phi}\right)$$
$$\left.+\frac{\mathbf{L}}{r}\,e^{(2\mathscr{F}-2H+L-J+I+Y)}\,\frac{(\epsilon_1-\epsilon_2)(m_1+m_2)}{\mathscr{D}(1)}\,\phi\right\}. \tag{A33}$$

Using the following vector identities ($\mathbf{A}=A\hat{\mathbf{r}}$),

$$\mathbf{p}\times(\mathbf{p}\times\boldsymbol{\phi})=\mathbf{p}(\mathbf{p}\cdot\boldsymbol{\phi})-\mathbf{p}^2\boldsymbol{\phi}, \tag{A34a}$$

$$i\mathbf{p}\times(\mathbf{A}\times\boldsymbol{\phi})=\left[A'-\frac{A}{r}\right]\hat{\mathbf{r}}(\hat{\mathbf{r}}\cdot\boldsymbol{\phi})+\left(i\mathbf{A}\cdot\mathbf{p}-\nabla\cdot+\frac{A}{r}\right)\boldsymbol{\phi}, \tag{A34b}$$

$$\hat{\mathbf{r}}\times(\hat{\mathbf{r}}\times\boldsymbol{\phi})=\hat{\mathbf{r}}(\hat{\mathbf{r}}\cdot\boldsymbol{\phi})-\boldsymbol{\phi}, \tag{A34c}$$

$$i\mathbf{A}\times(\mathbf{p}\times\boldsymbol{\phi})=i\mathrm{A}_i\mathbf{p}\phi_i-i\mathbf{A}\cdot\mathbf{p}\boldsymbol{\phi}, \tag{A34d}$$

$$r_i \mathbf{p}\phi_i = \mathbf{r}(\mathbf{p}\cdot\boldsymbol{\phi}) + \mathbf{L}\times\boldsymbol{\phi}, \tag{A34e}$$

leads to the result given in the text [Eq. (4.44)].

APPENDIX B: VECTOR SPHERICAL HARMONICS

We list here the identities used to arrive at Eqs. (5.8a)–(5.8d). For $j=l$ we use

$$\mathbf{p}^2\boldsymbol{\phi} = -\nabla^2\boldsymbol{\phi} = \frac{1}{r}\left[-\frac{d^2}{dr^2} + \frac{j(j+1)}{r^2}\right] u(r)_{j0j}\mathscr{Y}_{jm}(\phi,\theta), \tag{B1}$$

$$i\hat{\mathbf{r}}\cdot\mathbf{p}\boldsymbol{\phi} = \hat{\mathbf{r}}\cdot\nabla\boldsymbol{\phi} = \frac{d}{dr}\left(\frac{u(r)_{j0j}}{r}\right)\mathscr{Y}_{jm}(\phi,\theta), \tag{B2}$$

$$\mathbf{L}\cdot\boldsymbol{\phi} = \frac{u(r)_{l1j}}{r}\sqrt{j(j+1)}\,\mathscr{Y}_{jm}, \tag{B3}$$

$$\mathbf{L}\boldsymbol{\phi} = \sqrt{j(j+1)}\,\frac{u(r)_{j0j}}{r}\,\mathbf{X}_{jm}, \tag{B4}$$

$$-\frac{i}{r}\,\mathbf{L}\times\boldsymbol{\phi} = \frac{u(r)_{j1j}}{r}\,\mathbf{X}_{jm}. \tag{B5}$$

For the cases $l=j\pm1$

$$\boldsymbol{\phi} = \boldsymbol{\phi}_+ + \boldsymbol{\phi}_- = \frac{u_{j+11j}(r)}{r}\,\mathbf{Y}_{jm+} + \frac{u_{j-11j}(r)}{r}\,\mathbf{Y}_{jm-}, \tag{B6}$$

where

$$\mathbf{Y}_{jm+} = -\sqrt{\frac{j+1}{2j+1}}\left(\hat{\mathbf{r}} - i\,\frac{r\mathbf{p}}{j+1}\right)\mathscr{Y}_{jm}, \tag{B7}$$

$$\mathbf{Y}_{jm-} = \sqrt{\frac{j}{2j+1}}\left(\hat{\mathbf{r}} + i\,\frac{r\mathbf{p}}{j}\right)\mathscr{Y}_{jm}. \tag{B8}$$

In the remainder we use the abbreviations $u_\pm = u_{j\pm11j}$ and $\mathbf{Y}_\pm = \mathbf{Y}_{jm\pm}$

$$-\frac{i}{r}\,\mathbf{L}\times\boldsymbol{\phi} = +\left(\frac{j+2}{r}\right)\frac{u_+}{r}\,\mathbf{Y}_+ - \left(\frac{j-1}{r}\right)\frac{u_{-1}}{r}\,\mathbf{Y}_-, \tag{B9}$$

$$\hat{\mathbf{r}}\mathscr{Y}_{jm} = \sqrt{\frac{j}{2j+1}}\,\mathbf{Y}_- - \sqrt{\frac{j+1}{2j+1}}\,\mathbf{Y}_+, \tag{B10}$$

$$\mathbf{p}\mathscr{Y}_{jm} = -\frac{i}{r}\left[(j+1)\sqrt{\frac{j}{2j+1}}\,\mathbf{Y}_- + j\sqrt{\frac{j+1}{2j+1}}\,\mathbf{Y}_+\right], \tag{B11}$$

$$\hat{\mathbf{r}}\cdot\mathbf{p}\boldsymbol{\phi}_\pm = -i\left(\frac{u'_\pm}{r} - \frac{u_\pm}{r^2}\right)\mathbf{Y}_\pm, \tag{B12}$$

$$\hat{\mathbf{r}}\cdot\boldsymbol{\phi}=-\sqrt{\frac{j+1}{2j+1}}\frac{u_+}{r}\mathscr{Y}_{jm}+\sqrt{\frac{j}{2j+1}}\frac{u_-}{r}\mathscr{Y}_{jm}, \tag{B13}$$

$$\hat{\mathbf{r}}(\hat{\mathbf{r}}\cdot\boldsymbol{\phi})=\frac{1}{2j+1}\left[\left[j\,\frac{u_-}{r}-\sqrt{j(j+1)}\,\frac{u_+}{r}\right]\mathbf{Y}_-+\left[(j+1)\,\frac{u_+}{r}-\sqrt{j(j+1)}\,\frac{u_-}{r}\right]\mathbf{Y}_+\right], \tag{B14}$$

$$\begin{aligned}\mathbf{p}(\hat{\mathbf{r}}\cdot\boldsymbol{\phi})=\frac{i}{2j+1}&\left[-j\,\frac{u_-'}{r}-j^2\,\frac{u_-}{r^2}+\left[\sqrt{j(j+1)}\left[\frac{u_+'}{r}+j\,\frac{u_+}{r^2}\right]\right]\mathbf{Y}_-+\left[-(j+1)\,\frac{u_+'}{r}+(j+1)^2\,\frac{u_+}{r^2}\right.\right.\\&\left.\left.+\sqrt{j(j+1)}\left[\frac{u_-'}{r}-(j+1)\,\frac{u_-}{r^2}\right]\right]\mathbf{Y}_+\right],\end{aligned} \tag{B15}$$

$$\mathbf{p}\cdot\boldsymbol{\phi}=\frac{i}{\sqrt{2j+1}}\left[\sqrt{j}\left[j\,\frac{u_-}{r^2}-\frac{u_-'}{r}\right]+\sqrt{j+1}\left[(j+1)\,\frac{u_+}{r^2}+\frac{u_+'}{r}\right]\right]\mathscr{Y}_{jm}, \tag{B16}$$

$$\begin{aligned}\hat{\mathbf{r}}(\mathbf{p}\cdot\boldsymbol{\phi})=\frac{i}{2j+1}&\left[j^2\,\frac{u_-}{r^2}-j\,\frac{u_-'}{r}+\left[\sqrt{j(j+1)}\left[(j+1)\,\frac{u_+}{r^2}+\frac{u_+'}{r}\right]\right]\mathbf{Y}_-\right.\\&\left.-\left[-\frac{u_-'}{r}\right]+(j+1)^2\,\frac{u_+}{r^2}+(j+1)\,\frac{u_+'}{r}+\sqrt{j(j+1)}\left[j\,\frac{u_-}{r^2}\right]\mathbf{Y}_+\right],\end{aligned} \tag{B17}$$

$$\begin{aligned}\mathbf{p}(\mathbf{p}\cdot\boldsymbol{\phi})=\frac{1}{2j+1}&\left[j\left[j(j-1)\,\frac{u_-}{r^3}-\frac{u_-''}{r}\right]+\sqrt{j(j+1)}\left[(j+1)(j-1)\,\frac{u_+}{r^3}+(2j+1)\,\frac{u_+'}{r^2}+\frac{u_+''}{r}\right]\right]\mathbf{Y}_-\\&+\frac{1}{2j+1}\left[(j+1)\left[(j+1)(j+2)\,\frac{u_+}{r^3}-\frac{u_+''}{r}\right]\right.\\&\left.+\sqrt{j(j+1)}\left[j(j+2)\,\frac{u_-}{r^3}-(2j+1)\,\frac{u_-'}{r^2}+\frac{u_-''}{r}\right]\right]\mathbf{Y}_+.\end{aligned} \tag{B18}$$

[1] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **117**, 610 (1928).
[2] G. Breit, Phys. Rev. **34**, 553 (1929).
[3] G. Breit, Phys. Rev. **36**, 383 (1930).
[4] G. Breit, Phys. Rev. **39**, 616 (1932).
[5] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One and Two Electron Atoms* (Springer, Berlin, 1957).
[6] R. W. Childers, Phys. Rev. D **26**, 2902 (1982).
[7] C. W. Wong and C. Y. Wong, Phys. Lett. B **301**, 1 (1993).
[8] C. W. Wong and C. Y. Wong, Nucl. Phys. A **562**, 598 (1993).
[9] E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951); E. E. Salpeter, Phys. Rev. **87**, 328 (1952); G. C. Wick, Phys. Rev. **96**, 1124 (1954); R. E. Cutkosky, Phys. Rev. **96**, 1135 (1954); A. A. Logunov and A. N. Tavkhelidze, Nuovo Cimento **29**, 380 (1963); A. A. Logunov *et al.*, *ibid.* **30**, 134 (1963); R. Blankenbecler and R. Sugar, Phys. Rev. **142**, 1051 (1966); I. T. Todorov, in *Properties of the Fundamental Interactions*, edited by A. Zichichi (Editrice Compositori, Bologna, 1973), Vol. 9, part C, pp. 953–979; Phys. Rev. D **3**, 2351 (1971); R. Yaes, Phys. Rev. D **3**, 3086 (1971); M. A. Stroscio, Phys. Rep. C **22**, 215 (1975); B. L. Aneva, J. I. Krapchev, and V. A. Rizov, Bulg. J. Phys **2**, 409 (1975); V. A. Rizov and I. T. Todorov, Fiz. Elem. Chastits At. Yadre **6**, 669 (1975) [translated in Sov. J. Part. Nucl. **6**, 269 (1975)]; H. J. Schnitzer, Phys. Rev. Lett. **35**, 1540 (1975); W. E. Caswell and G. P. Lepage, Phys. Rev. A **18**, 863 (1977), **18**, 810 (1978); R. Barbieri and E. Remiddi, Nucl. Phys. B **141**, 413 (1978); E. Eichten and F. Feinberg, Phys. Rev. Lett. **43**, 1205 (1979); D. Gromes, Nucl. Phys. B **131**, 80 (1977); D. P. Stanley and D. Robson, Phys. Rev. D **21**, 31 (1980); W. Krolikowski, Acta Phys. Pol. B **12**, 891 (1981); F. Gross, Phys. Rev. C **26**, 2203 (1982); S. Gupta, S. F. Radford, and W. W. Repko, Phys. Rev. D **26**, 3305 (1982); R. W. Childers, Phys. Lett. B **121**, 485 (1983); M. Arafah, R. Bhandhri, and B. Ram, Lett. Nuovo Cimento **38**, 305 (1983); D. B. Lichtenberg, W. Namgung, J. G. Wills, and E. Predazzi, Z. Phys. C **19**, 19 (1983); M. Bander, D. Silverman, B. Klima, and U. Maor, Phys. Rev. D **29**, 2038 (1984); H. Pilkuhn, J. Phys. B **17**, 381 (1985); S. Godfrey and N. Isgur, Phys. Rev. D **32**, 189 (1985); G. Bohnert, R. Decker, A. Hornberg, H. Pilkuhn, and H. G. Schlaile, Z. Phys. **2**, 23 (1986); V. B. Mandelzweig and S. J. Wallace, Phys. Lett. B **197**, 469 (1988);

E. D. Cooper and B. K. Jennings, Nucl. Phys. A **483**, 601 (1988); J. Malenfant, Phys. Rev. A **43**, 1233 (1991); J. Connell, Phys. Rev. D **43**, 1393 (1991); A. N. Mitra *et al.*, Int. J. Mod. Phys. A **7**, 121 (1992).

[10] P. Van Alstine and H. W. Crater, J. Math. Phys. **23**, 1997 (1982); H. W. Crater and P. Van Alstine, Ann. Phys. (N.Y.) **148**, 57 (1983). Supersymmetry discussions in the above references were based in part on C. A. P. Galvao and C. Teitelboim, J. Math. Phys. **21**, 1863 (1980). See also A. Barducci, R. Casalbuoni, and L. Lusanna, Nuovo Cimento A **32**, 377 (1976).

[11] H. W. Crater and P. Van Alstine, Phys. Rev. Lett. **53**, 1577 (1984); P. Van Alstine and H. W. Crater, Phys. Rev. D **34**, 1932 (1986); H. W. Crater and P. Van Alstine, Phys. Rev. D **37**, 1982 (1988).

[12] H. Sazdjian, Phys. Rev. D **33**, 3401 (1986).

[13] H. W. Crater and P. Van Alstine, Phys. Rev. D **36**, 3007 (1987).

[14] H. W. Crater and P. Van Alstine, J. Math. Phys. **31**, 1998 (1990).

[15] H. W. Crater, R. Becker, C. Y. Wong, and P. Van Alstine, Phys. Rev. D **46**, 5117 (1992).

[16] H. W. Crater and P. Van Alstine, Found. Phys. **24**, 297 (1994).

[17] J. Mourad and H. Sazdjian, J. Phys. G **21**, 267 (1995).

[18] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950); Proc. R. Soc. London, Ser. A **246**, 326 (1958); *Lectures on Quantum Mechanics* (Yeshiva University, New York, 1964). The following references detail the first applications of Dirac's constraint formalism to the two-body problem. M. Kalb and P. Van Alstine, Yale Reports, C00-3075-146 (1976), C00-3075-156 (1976); P. Van Alstine, PhD dissertation, Yale University (1976); I. T. Todorov, Dubna Joint Institute for Nuclear Research No. E2-10175, 1976; Ann. Inst. H. Poincaré A **28**, 207 (1978); L. P. Horwitz and F. Rohrlich, Phys. Rev. D **24**, 1928 (1981); F. Rohrlich, Phys. Rev. D **23**, 1305 (1981); Nucl. Phys. B **161**, 469 (1979); A. Komar, Phys. Rev. D **18**, 1881,1887 (1978); P. Droz-Vincent, Rep. Math. Phys. **8**, 79 (1975).

[19] F. Gross has noted (Ref. 9) the paradox that the extra work that must be done to keep the relative time (or relative energy) as a variable in the Bethe-Salpeter equation seems to reward one with poorer answers.

[20] I. T. Todorov, Phys. Rev. D **3**, 2351 (1971).

[21] H. W. Crater, C. W. Wong, C. Y. Wong, and P. Van Alstine, Int. J. Mod. Phys. E **5**, 589 (1996).

[22] These theta matrices have algebraic properties that permit more efficient calculation of the commutation relations appropriate to two spinning bodies permitting simplification of otherwise complicated consequences of compatibility $([\mathscr{S}_1,\mathscr{S}_2]_-\psi=0)$ when interactions are present.

[23] We omit here four additional derivative interactions of the form $\alpha_1\cdot\hat{r}\alpha_2\cdot\hat{r}[R+\beta_1\beta_2 S+\rho_1\rho_2 T+\gamma_{51}\gamma_{52}N]$ in which the invariant functions $N(x_\perp),T(x_\perp),S(x_\perp),R(x_\perp)$ describe the derivative interactions (the retardative Breit term is one such example).

[24] As pointed out in Ref. 14, the fundamental anticommutation relations (2.7) of the theta matrices lead directly through a ''pseudoclassical'' correspondence limit (Ref. 10) to a graded symplectic structure in which the θ's become two commuting sets of Grassmann variables. This space possesses a graded Poisson bracket that takes the differential form described in Ref. 10 and references contained therein. When quantized, this bracket becomes a generalized quantum bracket that is sometimes a commutator and sometimes an anticommutator, depending on the nature of its operator arguments. In terms of this bracket, all necessary commutation relations involving the quantum θ's can be carried out through operations that are isomorphic to those involving the classical brackets. The quantum bracket follows closely the classical graded structure by sorting quantum operators into even and odd classes. This accounts for the seemingly random choices of commutators and anticommutators in Eqs. (3.3), (3.6), and (3.20). For details see Ref. 14.

[25] We omit here an extremely long and detailed calculation, involving 32 brackets, each of which involves multiple combinations of the above interaction matrices. Even though the result seems plausible in hindsight (in light of the earlier individual results), we found no simple proof because of the compounding of the various interactions.

[26] J. M. Blatt and V. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952).

[27] E. L. Hill, Am. J. Phys. **22**, 211 (1954).

Unitary transformation approach to the exact solution for a class of time-dependent nonlinear Hamiltonian systems

M. Maamache
Institut de Physique, Université de Sétif, Sétif, (19000), Algeria

(Received 4 November 1996; accepted for publication 10 June 1997)

By performing unitary transformations on the time-dependent Schrödinger equation, the exact solution for a class of nonlinear Hamiltonian systems is obtained and it is shown that these time-dependent problems are related to the associated time-independent problems. In addition, the evolution operator is derived. Nonadiabatic Berry's phase is calculated on the basis of the exact solution. The theory is applied to some illustrative examples. © *1998 American Institute of Physics.*
[S0022-2488(97)02412-2]

I. INTRODUCTION

In quantum mechanics, solving the time-dependent Schrödinger equation is in general very difficult and it is very rare to be able to find an exact solution. Various methods have been used to obtain approximate solutions for such time-dependent problems. The usual methods are the adiabatic approximation, the sudden approximation, and time-dependent perturbation techniques. The existence of invariants[1,2] (constants of the motion or first integral) is one of central importance in the study of such a system, be it classical or quantum. The importance of these invariants is classical mechanics and the associated nonlinear superposition law[3,4] has been adequately discussed. In the quantum case, Lewis and Riesenfeld[5] first exploited the invariant operators to solve quantum-mechanical problems. In particular, they have derived a simple relation between eigenstates of I and solutions of the time-dependent Schrödinger equation and have applied it to the case of a quantal oscillator with time-dependent frequency. Hartley and Ray[6] have applied this technique to derive a quantum-mechanical superposition law for certain nonlinear time-dependent oscillators; an explicit form of the propagator was obtained by Dhara and Lawande.[7]

In light of the above discussion, one important question motivates our work here: how could one treat time-dependent quantum problems without making recourse to the invariant operator theory.

In this paper we answer this question from a new perspective by studying the one degree of freedom quantum time-dependent nonlinear Hamiltonian systems,

$$H=\tfrac{1}{2}\left[Z(t)p^2+Y(t)(pq+qp)+X(t)q^2+V(q,t)\right], \tag{1.1}$$

introducing a suitable transformation to make the transformed Hamiltonian time independent multiplied by a time-dependent factor. Here q,p are the quantum-mechanical operators, $V(q,t)$ is an arbitrary time-dependent potential, and $X(t)$, $Y(t)$, and $Z(t)$ are arbitrary functions of time.

Note that the class of Hamiltonians represented by (1.1) includes, for example, the so-called ''generalized harmonic oscillator'' when $V(q,t)=0$, the Caldirola–Kanai Hamiltonian when $Y(t)=V(q,t)=0$, and the case $Y(t)=0$, $Z(t)=1$ treated by Hartley and Ray[6] and Ray and Reid.[4]

For an integrable system, the Hamiltonian can be transformed into a sum of time-independent commuting operators through a series of time-dependent unitary transformations.[8] As a consequence, the solution of the time-dependent Schrödinger equation for an integrable system can be written as eigenstates of these commuting operators. In this paper, the approach that we use is based on the unitary transformation. In addition, the other advantage of our approach is that the

time evolution from an arbitrary initial state can be calculated in terms of the time-evolution operators. In Sec. II, the formalism is developed based on the unitary transformation to obtain not only the exact solution of the Schrödinger equation, but also the time-evolution operator, and we elucidate the intimate connection between the Hill–Floquet–Bloch function and the solution of the Schrödinger equation for a T-periodic Hamiltonian. The nonadiabatic Berry's phases[8–15] are determined totally by such a unitary transformation. In Sec. III, we obtain with the help of the appropriate transformations the exact solution of the system (1.1), as well as the time-evolution operator, and we find the conditions that govern the class of allowable potentials. Although the class of allowable potentials does not consist of all functions of the independent variables q and t, it is described by an arbitrary function of a single argument and includes time-dependent potentials that can be arbitrarily nonlinear in q. We obtain not only the exact solution of the Schrödinger equation but also the time-evolution operator. On the basis of the exact solution for the system represented by the Hamiltonian (1.1), a nonadiabatic Berry's phase is calculated. In Sec. IV, we give some illustrative examples. We end with a conclusion in Sec. V.

II. FORMALISM

Consider the time-dependent Schrödinger equation,

$$\imath \frac{\partial}{\partial t} |\psi(t)\rangle = H(t)|\psi(t)\rangle, \tag{2.1}$$

with $\hbar = 1$. Suppose that is a time-dependent transformation $U^{-1}(t)$,

$$|\widetilde{\psi}(t)\rangle = U^{-1}(t)|\psi(t)\rangle, \tag{2.2}$$

transforms the wave function to $|\widetilde{\psi}(t)\rangle$, which obeys the Schrödinger equation with the accordingly transformed Hamiltonian $\widetilde{H}$,

$$\imath \frac{\partial}{\partial t} |\widetilde{\psi}(t)\rangle = \widetilde{H}|\widetilde{\psi}(t)\rangle. \tag{2.3}$$

In the above equation, the new Hamiltonian operator,

$$\widetilde{H} = \left[U^{-1}H(t)U - \imath U^{-1} \frac{\partial U}{\partial t} \right], \tag{2.4}$$

is Hermitian. This requires that $U(t)$ must be unitary operator. Since the Hamiltonian of a system is time dependent, $U(t)$ results in a unitary transformation (2.2) for the wave function and a unitary transformation (2.4) for the Hamiltonian so that the form of the Schrödinger equation remains form invariant.

With the help of the appropriate unitary transformation U, we require that the solution of Schrödinger equation looks like a generalized Hill–Floquet–Bloch function, i.e., the product of a wave function satisfying a stationary equation with a time-dependent function. For this, we may make the transformed Hamiltonian $\widetilde{H}$ possessing the following form:

$$\widetilde{H}(t) = g(t)H_0, \tag{2.5}$$

where H_0 is time independent and $g(t)$ is a function of time. To get the form of the solution, let ϕ_n be the eigenstate of H_0 with eigenvalue λ_n, i.e.

$$H_0\phi_n = \lambda_n\phi_n, \quad \lambda_n = cte, \tag{2.6}$$

and suppose that the system is in the nth eigenstate of H_0 initially, at time t it evolves to the state

$$\psi_n(q,t)=\exp\left(\imath\lambda_n\int_0^t dt'\ g(t')\right)U\phi_n(q). \tag{2.7}$$

We now derive the time-evolution operator. Note that the solution (2.7) may be rewritten as

$$\psi_n(q,t)=U\ \exp\left(-\imath\int_0^t dt'\ g(t')H_0\right)\phi_n(q), \tag{2.8}$$

when $t=0$ the quantum wave function is

$$\psi_n(q,0)=U(0)\phi_n(q), \tag{2.9}$$

since

$$\psi_n(q,t)=U\ \exp\left(-\imath\int_0^t dt'\ g(t')H_0\right)U^{-1}(0)\psi_n(q,0), \tag{2.10}$$

the time-evolution operator becomes

$$\mathscr{U}(t,0)=U(t)\exp\left(-\imath\int_0^t dt'\ g(t')H_0\right)U^{-1}(0), \tag{2.11}$$

and has all properties of a unitary evolution operator,

$$\mathscr{U}^{-1}(t,0)=\mathscr{U}(0,t),\quad \mathscr{U}(t_2,t_1)=\mathscr{U}(t_2,0)\mathscr{U}^{-1}(t_1,0). \tag{2.12}$$

The time evolution of arbitrary initial states can be evaluated with the time evolution operator (2.11).

In order to give a derivation of the nonadiabatic Berry's phase, let us look at periodical Hamiltonians $H(t+T)=H(t)$ that are susceptible to describe a cyclic motion of the wave function in analogy with Bloch states in the case of spatially periodic Hamiltonian. Such spatial periodicity induces a direct sum decomposition of the spaces of states corresponding to eigenspaces of the Floquet eigenvalues of H. These eigenvalues $\{\mathscr{E}_n\}$, called ''energy bands'' (or ''quasienergies'') in solid state physics, are periodic functions in the parameters space that is the Brillouin zone.[16] Thus, an initial state $\psi_n(q,0)$ will evolve as

$$\psi_n(q,T)=U(T)\psi_n(q,0)=e^{-\imath\alpha_n(T)}\psi_n(q,0), \tag{2.13}$$

i.e., each state returns to itself after a time T, up to a phase $\alpha_n(T)=\lambda_n\int_0^T dt\,g(t)$, which is called the Hill's phase. As this is the usual condition for the so-called Floquet state (or Bloch state) we can relate the Hill's phase α_n to the quasienergies associated with Bloch wave function:

$$\alpha_n(T)=\mathscr{E}_nT. \tag{2.14}$$

The generalization of the dynamical phase is the time integral of the instantaneous expectation value of the Hamiltonian,

$$\delta_n=\int_0^T\langle\psi_n(q,t)|H(t)|\psi_n(q,t)\rangle dt=\int_0^T dt\left(\lambda_ng(t)+\langle\phi_n(q)|\imath U^{-1}\frac{\partial U}{\partial t}|\phi_n(q)\rangle\right). \tag{2.15}$$

Aharonov and Anandan[10] show that

$$\beta_n=\lambda_n\int_0^T dt\ g(t)-\delta_n=-\langle\phi_n(q)|\imath U^{-1}\frac{\partial U}{\partial t}|\phi_n(q)\rangle dt \tag{2.16}$$

is a purely geometrical property of the evolution, in the sense that it only depends on the path followed by the system in projective Hilbert space. We see that the Hill's phase can be decomposed in two terms, one of them being the nonadiabatic Berry's phase that is determined totally by the unitary transformation.

III. EXACT SOLUTION AND NONADIABATIC BERRY'S PHASE FOR NONLINEAR HAMILTONIAN SYSTEMS

Now we embark on the construction of the exact solution and the nonadiabatic Berry phase for the system described by the Hamiltonian (1.1). The key point of our analysis is to perform the three consecutive transformations:

$$|\psi_n\rangle=U_1|\psi_{1n}\rangle=\exp\left[-\frac{\imath}{2}\int_0^t Y(t')dt'(pq+qp)\right]|\psi_1\rangle, \tag{3.1}$$

$$|\psi_{1n}\rangle=U_2|\psi_{2n}\rangle=\exp[\imath C_1(t)q^2]|\psi_{2n}\rangle, \tag{3.2}$$

and

$$|\psi_{2n}\rangle=U_3|\psi_{3n}\rangle=\exp[\imath C_2(t)(pq+qp)]|\psi_{3n}\rangle, \tag{3.3}$$

where $C_1(t)$ and $C_2(t)$ are real functions of time. These coefficients are chosen in such a way that the Hamiltonian, after these transformations becomes a product of two factors, namely a time-dependent factor multiplied by a simple time-independent Hamiltonian, which represents a harmonic oscillator with a constant frequency Ω_0 with an additional time-independent potential.

The unitary operator $U_1(t)$ has the properties

$$U_1^{-1}f(p)U_1=f\left(p\ \exp\left[-\int_0^t Y(t')dt'\right]\right) \tag{3.4}$$

and

$$U_1^{-1}f(q)U_1=f\left(q\ \exp\left[\int_0^t Y(t')dt'\right]\right). \tag{3.5}$$

In a particular representation, the wave function is given by

$$\langle q|U_1|\psi_n\rangle=\exp\left[-\frac{1}{2}\int_0^t Y(t')dt'\right]\psi_n\left(q\ \exp\left[-\int_0^t Y(t')dt'\right]\right). \tag{3.6}$$

It can easily be shown that under the U_2 transformation, the coordinate and momentum operators change according to

$$U_2^{-1}pU_2=p+2C_1q, \tag{3.7}$$

$$U_2^{-1}f(q)U_2=f(q). \tag{3.8}$$

In order to solve (2.1) with the Hamiltonian specified by (1.1), we first try to remove the mixed terms in q and p in the Hamiltonian (1.1). This can be achieved by the transformation $U_1(t)$. Inserting (3.1) and (3.4)–(3.5) into (2.3), we have

$$\imath\frac{\partial}{\partial t}|\psi_{1n}(t)\rangle=\frac{1}{2}\left[\frac{1}{m(t)}p^2+k(t)q^2+V\left(q\ \exp\left[\int_0^t Y(t')dt'\right],t\right)\right]|\psi_{1n}(t)\rangle, \tag{3.9}$$

where

$$m(t)=\frac{1}{Z(t)}\exp\left[2\int_0^t Y(t')dt'\right], \tag{3.10}$$

$$k(t)=X(t)\exp\left[2\int_0^t Y(t')dt'\right]. \tag{3.11}$$

Now let us perform the transformation $U_2(t)$. Substituting (3.2) into (3.9) and using (3.7) and (3.8), we find the equation of motion for $|\psi_{2n}(t)\rangle$,

$$\begin{aligned}\imath\frac{\partial}{\partial t}|\psi_{2n}(t)\rangle=\frac{1}{2}\left[\frac{1}{m}p^2+2\frac{C_1}{m}(pq+qp)+2\left(\dot{C}_1+2\frac{C_1^2}{m}+\frac{k}{2}\right)q^2\right.\\ \left.+V\left(q\ \exp\left[\int_0^t Y(t')dt'\right],t\right)\right]|\psi_{2n}(t)\rangle.\end{aligned} \tag{3.12}$$

In order to obtain the usual time-dependent harmonic oscillator with a perturbative potential, we remove the cross term in (3.12) via the transformation $U_3(t)$. Since this transformation is formally the same as $U_1(t)$, this suggests that one takes

$$C_2(t)=-\int_0^t\frac{C_1(t')}{m(t')}dt'. \tag{3.13}$$

With this choice, (3.12) is changed under these transformation into

$$\begin{aligned}\imath\frac{\partial}{\partial t}|\psi_{3n}(t)\rangle=\frac{1}{2}\left\{\frac{1}{m}\exp\left[-4\int_0^t\frac{C_1}{m}dt'\right]p^2+2\left(\dot{C}_1+2\frac{C_1^2}{m}+\frac{k}{2}\right)\exp\left[4\int_0^t\frac{C_1}{m}dt'\right]q^2\right.\\ \left.+V\left(q\ \exp\left[\int_0^t\left(Y+2\frac{C_1}{m}\right)dt'\right],t\right)\right\}|\psi_{3n}(t)\rangle.\end{aligned} \tag{3.14}$$

We note that in Eq. (3.14), the argument of the potential V, i.e., $\exp[\int_0^t(Y+2C_1/m)dt']$, is a real function of time that we can call $\rho(t)$. Thus, we can deduce that

$$C_1(t)=\frac{m}{2}\left(\frac{\dot{\rho}}{\rho}-Y\right), \tag{3.15}$$

one can show that (3.14) becomes

$$\imath\frac{\partial}{\partial t}|\psi_{3n}(t)\rangle=\frac{1}{2}\frac{Z}{\rho^2}\left[p^2+\Omega^2q^2+\frac{\rho^2}{Z}V(\rho q,t)\right]|\psi_{3n}(t)\rangle, \tag{3.16}$$

where

$$\Omega^2=\frac{\rho^3}{Z^2}\left(\ddot{\rho}-\frac{\dot{Z}}{Z}\dot{\rho}+\left[(XZ-Y^2)+\frac{\dot{Z}}{Z}Y-\dot{Y}\right]\rho\right). \tag{3.17}$$

The central idea in this procedure is to require that the Hamiltonian governing the evolution of $|\psi_{3n}\rangle$ is the product of two parts. A time-dependent factor and a simple time-independent nonlinear Hamiltonian that represents a harmonic oscillator with a constant frequency Ω_0 with an additional time-independent potential. This condition imposes a constraint on the global time-dependent frequency and in the time-dependent potential. Let us set the global time-dependent frequency appearing in (3.17) equals to a real constant,

$$\Omega^2=\Omega_0^2, \tag{3.18}$$

which amounts to imposing a constraint on ρ. Equation (3.17) then becomes

$$\ddot{\rho}-\frac{\dot{Z}}{Z}\dot{\rho}+\left[(XZ-Y^2)+\frac{\dot{Z}}{Z}Y-\dot{Y}\right]\rho=\frac{\Omega_0^2Z^2}{\rho^3}, \tag{3.19}$$

which is the well-known auxiliary equation.[14] We now conclude our examination of (3.16) to determine the class of admissible potentials. The particular choice,

$$V(q,t)=\frac{Z}{\rho^2}W\left(\frac{q}{\rho}\right), \tag{3.20}$$

implies that the potential appearing in (3.16) is time independent, as required, where W is an arbitrary function of its argument.

It is easy to verify with the help of Eqs. (2.6)–(2.7) that one can write the solution of (3.16) as

$$\psi_{3n}(q,t)=\exp\left[-\imath\lambda_n\int_0^t\frac{Z}{\rho^2}\,dt'\right]\phi_n(q), \tag{3.21}$$

where $\phi_n(q)$ satisfies the eigenequation

$$\tfrac{1}{2}\,(p^2+\Omega_0^2q^2+W(q))\phi_n(q)=\lambda_n\phi_n(q),\quad \lambda_n=\text{const.} \tag{3.22}$$

The implication of the results is clear. The original time-dependent quantum-mechanical problem [posed through the Hamiltonian (1.1)] is completely solved if the wave function for the related (time-independent) Hamiltonian defined in Eq. (3.22) is obtained. The exact solution of the original equation (2.1) can now be found by combining the above results. We finally obtain

$$\begin{aligned}\psi_n(q,t)&=\exp\left[-\imath\lambda_n\int_0^t\frac{Z}{\rho^2}\,dt'\right]U_1U_2U_3\phi_n(q)\\&=\exp\left[-\imath\lambda_n\int_0^t\frac{Z}{\rho^2}\,dt'\right]\frac{1}{\rho^{1/2}}\exp\left[-\imath\left(\frac{Y\rho-\dot{\rho}}{Z\rho}\right)\frac{q^2}{2}\right]\phi_n\left(\frac{q}{\rho}\right).\end{aligned} \tag{3.23}$$

We now derive the time-evolution operator for the nonlinear system (1.1). The time-evolution arbitrary initial state can be evaluated with the time-evolution operator, Eq. (2.11),

$$\begin{aligned}\mathscr{U}(t,0)&=\exp\left\{\frac{\imath}{2Z}\left(\frac{\dot{\rho}}{\rho}-Y\right)q^2\right\}\exp\left\{-\frac{\imath}{2}\,\mathrm{Ln}\left(\frac{\rho}{\rho_0}\right)(pq+qp)\right\}\\&\quad\times\exp\left\{-\frac{\imath}{2}\int_0^t\frac{Z}{\rho^2}\,dt'(p^2+\Omega_0^2q^2+W(q))\right\}\exp\left\{\frac{\imath}{2Z_0}\left(\frac{\dot{\rho}_0}{\rho_0}-Y_0\right)q^2\right\},\end{aligned} \tag{3.24}$$

where ρ_0, $\dot{\rho}_0$, Z_0, and Y_0 are related to the initial values $\rho(0)$, $\dot{\rho}(0)$, $Z(0)$, and $Y(0)$.

If the parameters $X(t)$, $Y(t)$, and $Z(t)$ are periodic functions of time with period T, i.e. $(X,Y,Z)(t+T)=(X,Y,Z)(t)$, Eq. (3.19) may have periodic solutions. When the auxiliary function is periodic, i.e. $\rho(t+T)=\rho(t)$, then after one period of evolution the system returns to the initial state, except for acquiring the Hill's phase,

$$\alpha_n(T)=\lambda_n\int_0^T\frac{Z}{\rho^2}\,dt, \tag{3.25}$$

and a nonadiabatic Berry's phase is derived by use of Eq. (2.16) as

$$\begin{aligned}\beta_n(\mathscr{C})&=-\imath\int_0^T\langle\phi_n(q)|U_3^{-1}U_2^{-1}U_1^{-1}\frac{\partial}{\partial t}U_1U_2U_3|\phi_n(q)\rangle dt\\&=\langle\phi_n(q)|q^2|\phi_n(q)\rangle\oint_{\mathscr{C}}d\rho\left(\frac{\rho Y-\dot\rho}{Z}\right).\end{aligned} \tag{3.26}$$

Thus, $\beta_n(\mathscr{C})$ have a rather interesting structure being the product of two quite distinct parts, the expectation value of q^2 for the corresponding autonomous system and a circuit integral in parameter space. We note that it is always the expectation value of q^2, no matter the potential.

IV. APPLICATIONS

We now discuss applications of our formulas (3.23) and (3.26) to obtain an explicitly exact wave function and a nonadiabatic Berry's phase for some time-dependent problems. Consider first the time-dependent harmonic oscillator for which $W=0$, $Y=0$, $Z(t)=1/M(t)$, and $X(t)=M(t)\omega^2(t)$, where $M(t)$ and $\omega(t)$ are, respectively, the mass and frequency associated with the oscillator. Here Eq. (3.22) simply corresponds to a harmonic oscillator with constant frequency Ω_0;

$$\tfrac{1}{2}\,(p^2+\Omega_0^2q^2)\,\phi_n(q)=\lambda_n\phi_n(q), \tag{4.1}$$

whose solution is well known,

$$\phi_n(q)=\frac{\Omega_0^{1/4}}{(\pi^{1/2}2^n n!)^{1/2}}\,e^{-q^2/2}H_n(q\Omega_0^{1/2}) \tag{4.2}$$

and

$$\lambda_n=(n+\tfrac{1}{2})\Omega_0, \tag{4.3}$$

where H_n is the usual Hermite polynomial of order n. The solution of the Schrödinger equation for the harmonic oscillator with time-dependent mass and frequency may be written as

$$\begin{aligned}\psi_n(q,t)&=\frac{\Omega_0^{1/4}}{(\pi^{1/2}2^n n!)^{1/2}}\,e^{\imath\alpha_n(t)}U_2U_3e^{-q^2/2}H_n(q\Omega_0^{1/2})\\&=\frac{1}{(\pi^{1/2}2^n n!)^{1/2}}\left(\frac{\Omega_0}{\rho^2}\right)^{1/4}e^{\imath\alpha_n(t)}\exp\left\{\imath\,\frac{M}{2}\left(\frac{\dot\rho}{\rho}+\frac{\imath}{M\rho^2}\right)q^2\right\}H_n\left(\Omega_0^{1/2}\frac{q}{\rho}\right),\end{aligned} \tag{4.4}$$

and

$$\alpha_n(t)=-\left(n+\frac{1}{2}\right)\int_0^t\frac{\Omega_0}{M(t')\rho^2(t')}\,dt'. \tag{4.5}$$

Here the auxiliary equation (3.19) is converted in the equation

$$\ddot{\rho}+\frac{\dot{M}}{M}\,\dot{\rho}+\omega^2(t)\rho=\frac{\Omega_0^2}{M(t)\rho^3}. \tag{4.6}$$

There remains the problem of finding the nonadiabatic Berry's phase β_n that satisfies Eq. (3.26). Since the expectation value $\langle\phi_n(q)|q^2|\phi_n(q)\rangle$ refers to the harmonic oscillator, it may be replaced by $(n+1/2)$. If the parameters $M(t),\omega^2(t)$ are periodic functions of time with period T, then after one period the system acquires a nonadiabatic Berry's phase,

$$\beta_n(T)=-(n+1/2)\int_0^T\frac{M(t)\dot{\rho}^2(t)}{\Omega_0}\,dt. \tag{4.7}$$

The wave function (4.4) and the nonadiabatic Berry's phase (4.7) are in agreement with the results obtained, by using the Heisenberg picture approach, in Ref. 15.

A somewhat nontrivial example is that of a time-dependent generalized harmonic oscillator with an additional inverse quadratic potential $W=(g\rho/q)^2$,[14] where g is a constant. In this case Eq. (3.22) resembles the radial time-independent Schrödinger equation of a three-dimensional oscillator corresponding to an angular momentum $l(l+1)=g^2$,

$$\frac{1}{2}\left(p^2+\Omega_0^2q^2+\frac{g^2}{q^2}\right)\phi_n(q)=\lambda_n\phi_n(x), \tag{4.8}$$

whose solution is well known,

$$\phi_n(x)=\left[\frac{2\Omega_0^{1/2}\Gamma(n+1)}{\Gamma(n+l+3/2)}\right]^{1/2}\exp\left[-\frac{1}{2}\,x^2\right](x)^{(l+1)/4}L_n^{l+1/2}(x^2) \tag{4.9}$$

and

$$\lambda_n=2\left(n+\frac{l}{2}+\frac{3}{4}\right)\Omega_0, \tag{4.10}$$

where we have introduced the dimensionless variable $x=\Omega_0^{1/2}q$, and $L_n^{l+1/2}$ denotes the generalized Laguerre polynomials. We finally obtain the exact solution of the time-dependent singular oscillator,

$$\begin{aligned}\psi_n(x,t)&=\exp\left[-2\imath\left(n+\frac{l}{2}+\frac{3}{4}\right)\Omega_0\int_0^t\frac{Z}{\rho^2}\,dt'\right]U_1U_2U_3\phi_n(x)\\&=\exp\left[-2\imath\left(n+\frac{l}{2}+\frac{3}{4}\right)\Omega_0\int_0^t\frac{Z}{\rho^2}\,dt'\right]\psi_n(x,t),\end{aligned} \tag{4.11}$$

where

$$\psi_n(x,t)=\left[\frac{2\Omega_0^{1/2}}{\rho}\,\frac{\Gamma(n+1)}{\Gamma(n+l+3/2)}\right]^{1/2}\exp\left[-\frac{1}{2}\left(\frac{x}{\rho}\right)^2\left\{1+\imath\,\frac{Y\rho^2-\dot{\rho}\rho}{Z}\right\}\right]\left(\frac{x}{\rho}\right)^{(l+1)/4}L_n^{l+1/2}\left(\left(\frac{x}{\rho}\right)^2\right). \tag{4.12}$$

In this case the function ρ satisfies the auxiliary equation (3.19). Now let us calculate the Berry's phase for the singular oscillator from Eq. (3.26). Since the expectation value $\langle\phi_n(q)|q^2|\phi_n(q)\rangle$

refers to the radial part of a three-dimensional harmonic oscillator, it may be replaced by $2(n+l/2+3/4)$. Then after one period, the nonadiabatic Berry's phase in a cyclic evolution is given by

$$\beta_n(\mathscr{C}) = -2\left(n+\frac{c}{2}+\frac{3}{4}\right)\oint_{\mathscr{C}} \Omega_0^{-1}\left[\frac{Y\rho-\dot{\rho}}{Z}\right]d\rho, \tag{4.13}$$

where $\mathscr{C}$ is a closed circuit in the parameter space.

Note that potential $Z/\rho^2 W$ in these two examples does not depend on the auxiliary function $\rho(t)$. In general, however, the perturbative potential depends on ρ and is tailored according to the auxiliary equation for $\rho(t)$.

V. CONCLUSION

In this paper we have presented an alternative treatment for certain nonlinear time-dependent oscillators. The present treatment is based on the use of time-dependent unitary transformations. We have shown that the time-dependent Hamiltonian (1.1) is related to an associated time-independent Hamiltonian multiplied by an overall time-dependent factor. The solution of a Schrödinger equation appears as a generalized Hill–Floquet–Bloch function, i.e., the product of a wave function satisfying a stationary equation with a time-dependent function. The calculation does not presuppose the existence of an invariant operator. Moreover, we have derived an explicit time-evolution operator. We have also found the exact expression of the nonadiabatic Berry's phase in such problems. The theory is applied to the harmonic oscillator with time-dependent mass and frequency, and the time-dependent singular oscillator.

[1] H. R. Lewis, Phys. Rev. Lett. **18**, 636 (1967); J. Math. Phys. **9**, 1976 (1968).
[2] V. P. Ermakov, Univ. Izv. Kiev **20**, No. 9, 1 (1880).
[3] M. Lutzky, Phys. Lett. A **78**, 301 (1980); J. Math. Phys. **21**, 1370 (1980).
[4] J. R. Ray and L. D. Reid, Phys. Lett. A **71**, 317 (1979); **74**, 23 (1979); J. Math. Phys. **20**, 2054 (1979); **21**, 1583 (1980); **22**, 91 (1981).
[5] H. R. Lewis and W. B. Riesenfeld, J. Math. Phys. **10**, 1458 (1969).
[6] J. G. Hartley and J. R. Ray, Phys. Rev. A **24**, 2873 (1981).
[7] A. K. Dhara and S. V. Lawande, Phys. Rev. A **30**, 560 (1984).
[8] F. Li, S. J. Wang, A. Weiguny, and D. L. Lin, J. Phys. A **27**, 985 (1994).
[9] M. V. Berry, Proc. R. Soc. London, Ser. A **392**, 45 (1984).
[10] Y. Aharonov and J. Anandan, Phys. Rev. Lett. **58**, 1593 (1987).
[11] D. A. Molares, J. Phys. A **21**, L889 (1988).
[12] J. M. Cervero and J. D. Lejarreta, J. Phys. A **22**, L663 (1989).
[13] X. Gao, J. B. Xu, and T. Z. Qian, Ann. Phys. (NY) **204**, 235 (1990).
[14] M. Maamache, Phys. Rev. A **52**, 936 (1995); Ann. Phys. (NY) **254**, 1 (1997).
[15] J. Y. Ji, J. K. Kim, S. P. Kim, and K. S. Soh, Phys. Rev. A **52**, 3352 (1995).
[16] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt, Rinehart, and Winston, New York, 1976).

Abelian Chern–Simons theory. I. A topological quantum field theory

Mihaela Manoliu
Department of Mathematics, University of Texas, Austin, Texas 78712

(Received 24 March 1997; accepted for publication 29 August 1997)

We give a construction of the Abelian Chern–Simons gauge theory from the point of view of a $2+1$-dimensional topological quantum field theory. The definition of the quantum theory relies on geometric quantization ideas that have been previously explored in connection to the non-Abelian Chern–Simons theory [J. Diff. Geom. **33**, 787–902 (1991); Topology **32**, 509–529 (1993)]. We formulate the topological quantum field theory in terms of the category of extended 2- and 3-manifolds introduced in a preprint by Walker in 1991 and prove that it satisfies the axioms of unitary topological quantum field theories formulated by Atiyah [Publ. Math. Inst. Hautes Etudes Sci. Pans **68**, 175–186 (1989)]. © *1998 American Institute of Physics.* [S0022-2488(97)01812-4]

I. INTRODUCTION

A $2+1$-dimensional topological quantum field theory (TQFT) with Lagrangian the non-Abelian Chern–Simons invariant of connections was introduced by Witten several years ago.[1] It was shown that this theory generates invariants of framed oriented links in arbitrary 3-manifolds. It also leads to topological invariants of closed oriented 3-manifolds endowed with a 2-framing.[2] Witten's approach[1] to defining these invariants is based on a functional integral formulation, results from two-dimensional conformal field theory and the axioms of TQFT formulated in Ref. 3. A mathematically rigorous derivation of Witten's 3-manifold invariants is given in Refs. 4 and 5. Moreover, Walker[4] gives a complete construction of the SU(2) Chern–Simons theory, proving that it satisfies the axioms of a TQFT. A simpler model of a $2+1$-dimensional TQFT is the Chern–Simons theory with a finite gauge group treated in full detail in Ref. 6.

In line with the above mentioned developments relating to the Chern–Simons gauge theory, in this paper we present a construction of the Abelian Chern–Simons theory as a $2+1$-dimensional TQFT. Given the group $\mathbb{T}$ of complex numbers of unit modulus and an (even) integer k, called the level, we aim to associate to every closed oriented two-dimensional manifold Σ a finite-dimensional vector space $\mathscr{H}(\Sigma)$ and to every compact oriented 3-manifold X a vector Z_X in the vector space $\mathscr{H}(\partial X)$ associated to the boundary ∂X of X.

The first assignment, $\Sigma \to \mathscr{H}(\Sigma)$, results from the geometric quantization of the moduli space $\mathscr{M}_\Sigma$ of flat $\mathbb{T}$ connections on Σ, having in view that $\mathscr{M}_\Sigma$ is identified to the symplectic torus $H^1(\Sigma;\mathbb{R})/H^1(\Sigma;\mathbb{Z})$ with standard symplectic form ω_Σ. There are various models for the quantization of $\mathscr{M}_\Sigma$ according to the type of polarization one chooses on this space. The holomorphic (Kähler) quantization of $\mathscr{M}_\Sigma$ is discussed in Refs. 7–9, where it is shown that the vector spaces constructed for various choices of complex structures on $\mathscr{M}_\Sigma$ are projectively identified. In this paper we choose to quantize $\mathscr{M}_\Sigma$ by using *real* polarizations. To each rational Lagrangian subspace L in $H^1(\Sigma;\mathbb{R})$ there corresponds an invariant real polarization $\mathscr{P}_L$ of the torus $\mathscr{M}_\Sigma \cong H^1(\Sigma;\mathbb{R})/H^1(\Sigma;\mathbb{Z})$. The quantization of $\mathscr{M}_\Sigma$, at level k, in this real polarization constructs a k^g-dimensional inner product space $\mathscr{H}(\Sigma,L)$, where g denotes the genus of Σ. The quantization of symplectic tori in a real polarization is discussed in detail in Ref. 10 and the results obtained there are directly applicable to the case of the moduli space $\mathscr{M}_\Sigma$. According to Ref. 10, for any two rational Lagrangian subspaces L_1, $L_2 \subset H^1(\Sigma;\mathbb{R})$ there is a canonically defined unitary operator $F_{L_2L_1}: \mathscr{H}(\Sigma,L_1) \to \mathscr{H}(\Sigma,L_2)$. The vector spaces associated to Σ and different choices of

0022-2488/98/39(1)/170/37/$15.00
© 1998 American Institute of Physics

rational Lagrangian subspaces in $H^1(\Sigma;\mathbb{R})$ are all projectively identified, similarly to the holomorphic quantization case. The projective factor, an eighth root of unity, is expressible in terms of the Maslov–Kashiwara index of a triple of Lagrangian subspaces of $H^1(\Sigma;\mathbb{R})$. The idea of using real polarizations in quantizing the moduli space $\mathscr{M}_\Sigma$ and the method of construction of the vector space $\mathscr{H}(\Sigma,L)$ was inspired by Ref. 11, where, in the context of the non-Abelian Chern–Simons theory, the authors address the problem of quantizing the moduli space of flat SU(2) connections on a closed oriented 2-manifold in a particular real polarization of that space.

For the second assignment, $X\to Z_X$, we start from the known fact that, for X, a compact oriented 3-manifold with boundary, the image of the restriction map $H^1(X;\mathbb{R})\to H^1(\partial X;\mathbb{R})$ defines a Lagrangian subspace L_X in the symplectic vector space $(H^1(\partial X;\mathbb{R}),\omega_{\partial X})$. Then we provide a canonical construction that defines Z_X as an element of the vector space $\mathscr{H}(\partial X,L_X)$ associated to the closed 2-manifold ∂X and the rational Lagrangian subspace L_X. The construction of Z_X relies on geometric quantization ideas. Following the Chern–Simons line construction exposed in Ref. 12, we introduce a prequantum line bundle $\mathscr{L}_{\partial X}$ over the moduli space of flat connections $\mathscr{M}_{\partial X}$. The line bundle $\mathscr{L}_{\partial X}$ carries a natural connection with curvature of the symplectic form $k\omega_{\partial X}$. The rational Lagrangian subspace $L_X\subset H^1(\partial X;\mathbb{R})$ defines an invariant real polarization on $\mathscr{M}_{\partial X}$. One of the leaves of this polarization is the image Λ_X in $\mathscr{M}_{\partial X}$ of the moduli space $\mathscr{M}_X$ of flat $\mathbb{T}$ connections on X under the map $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$ determined by restricting connections on X to the boundary ∂X. We show that the leaf Λ_X is a Bohr–Sommerfeld leaf, that is, the restriction of the line bundle $\mathscr{L}_{\partial X}$ to Λ_X admits nonzero covariantly constant global sections. The vector Z_X is defined as a section of the line bundle $\mathscr{L}_{\partial X}|_{\Lambda_X}$ times a section of the bundle of half-densities on Λ_X. The former section is defined in terms of the Chern–Simons functional of flat connections on X and the latter in terms of the Reidemeister torsion invariant of the 3-manifold X. For X a closed manifold the construction gives for Z_X a complex number. The definition of Z_X, that is, finding the appropriate ingredients that should enter into this definition, was inspired in part by the results for the non-Abelian Chern–Simons theory from Ref. 11, as well as by the functional integral formulation of the closed 3-manifold SU(2) invariant given in Refs. 1 and 13.

We find that a natural way to incorporate the assignments $\Sigma\to\mathscr{H}(\Sigma)$ and $X\to Z_X$, defined as outlined above, into a TQFT is to make use of the category of *extended* 2- and 3-manifolds introduced by Walker in his treatment[4] of the non-Abelian Chern–Simons theory. In this paper an extended 2-manifold is a pair (Σ,L), with Σ a closed oriented 2-manifold and L a rational Lagrangian subspace in $H^1(\Sigma;\mathbb{R})$. An extended 3-manifold is a triple (X,L,n), with X a compact oriented 3-manifold, L a rational Lagrangian subspace in $H^1(\partial X;\mathbb{R})$, and $n\in\mathbb{Z}/8\mathbb{Z}$. Then the theory assigns to (Σ,L) the finite-dimensional inner product space $\mathscr{H}(\Sigma,L)$ and to (X,L,n) the vector $Z_{(X,L,n)}=e^{(\pi i/4)n}F_{LL_X}(Z_X)$ belonging to the vector space $\mathscr{H}(\partial X,L)$. We prove that these assignments satisfy the axioms[3] of a unitary TQFT, that is, the functoriality, orientation, disjoint union, and gluing properties.

The outline of this paper is the following. In Sec. II we begin by reviewing some fairly standard material on the moduli space of flat $\mathbb{T}$ connections. We describe properties of this moduli space for the case when connections are on two- and three-dimensional manifolds. Section III is devoted to an exposition of the properties of the classical Chern–Simons theory for the Abelian group $\mathbb{T}$ that are relevant for the sections to follow. We recall first the definitions of induced principal bundles and induced connections. Then we introduce the Chern–Simons functional of $\mathbb{T}$ connections on a 3-manifold with and without a boundary and prove a series of properties of this functional. In Sec. IV we give the construction of the prequantum line bundle $\mathscr{L}_\Sigma$ over the moduli space $\mathscr{M}_\Sigma$ of flat $\mathbb{T}$ connections on a closed 2-manifold Σ. Then we show that, if X is a 3-manifold with a boundary, the pullback of the prequantum line bundle $\mathscr{L}_{\partial X}$ over $\mathscr{M}_{\partial X}$ to $\mathscr{M}_X$ has a covariantly constant section defined in terms of the Chern–Simons functional. In Sec. V we define the quantum theory. We construct the finite-dimensional Hilbert space $H(\Sigma,L)$ associated to a closed oriented 2-manifold Σ and a rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$. Then, for a compact oriented 3-manifold X, we construct the vector Z_X belonging the Hilbert space

$\mathscr{H}(\partial X, L_X)$. Section VI contains the definition of the Abelian Chern–Simons TQFT. We introduce the notions of extended 2- and 3-manifolds, extended morphisms and gluing of extended 3-manifolds. Using the results of Sec. V, we define then a $2+1$-dimensional TQFT based on this category of extended manifolds and prove that it satisfies the required axioms. We also show that the mapping class group of an extended 2-manifold is unitarily represented on the corresponding quantum Hilbert space.

Throughout this paper, manifolds, bundles, and maps are assumed smooth.

II. THE MODULI SPACE OF FLAT $\mathbb{T}$ CONNECTIONS

A. The space of $\mathbb{T}$ connections

Let M be a smooth manifold and let $\mathscr{A}_M$ denote the space of $\mathbb{T}$ connections on M. An element Θ in $\mathscr{A}_M$ is a connection on a principal $\mathbb{T}$ bundle $P \to M$. Thus $\mathscr{A}_M$ is equal to the union

$$\mathscr{A}_M = \bigsqcup_P \mathscr{A}_P ,$$

over all principal $\mathbb{T}$ bundles P on M. For each bundle P the space $\mathscr{A}_P$ of connections on P is an affine space with vector space $2\pi i \Omega^1(M;\mathbb{R})$, where $\Omega^1(M;\mathbb{R})$ is the space of 1-forms on M (the Lie algebra of $\mathbb{T}$ is identified with $2\pi i\mathbb{R}$).

A bundle isomorphism between two principal $\mathbb{T}$ bundles P and P' over M is a $\mathbb{T}$-equivariant map $\phi: P' \to P$, which covers the identity on M. Two elements Θ and Θ' in $\mathscr{A}_M$ are *gauge equivalent* if there exists an isomorphism $\phi: P' \to P$, such that $\Theta' = \phi^* \Theta$. This defines an equivalence relation $\Theta' \sim \Theta$ on $\mathscr{A}_M$. We let $\mathscr{A}_M/\sim$ denote the space of gauge equivalence classes of $\mathbb{T}$ connections on M.

The space $\mathscr{G}_M = \mathrm{Map}(M, \mathbb{T})$ forms a group under pointwise multiplication. $\mathscr{G}_M$ is the *group of gauge transformations* on M. The space of components $\pi_0(\mathscr{G}_M)$ is isomorphic to $H^1(M;\mathbb{Z})$.[14] For any principal $\mathbb{T}$ bundle $\pi: P \to M$, the group $\mathscr{G}_M$ can be identified with the group $\mathscr{G}_P = \mathrm{Aut}(P)$ of bundle automorphisms of P. We call $\mathscr{G}_P$ the group of gauge transformations of P. For each $\mathbb{T}$ bundle P there is a right action of $\mathscr{G}_M$ on the space of connections $\mathscr{A}_P$. If $u: M \to \mathbb{T}$ is an element of $\mathscr{G}_M$ with associated bundle automorphism $\phi_u: P \to P$, the action of u on $\Theta \in \mathscr{A}_P$ is described by

$$\Theta \cdot u = \phi_u^* \Theta = \Theta + (u \cdot \pi)^* \vartheta, \tag{II.1}$$

where ϑ is the Maurer–Cartan form of $\mathbb{T}$, that is, $\vartheta/2\pi i$ generates $H^1(\mathbb{T};\mathbb{Z}) \subset H^1(\mathbb{T};\mathbb{R})$.

A principal $\mathbb{T}$ bundle $P \to M$ has flat connections if and only if $c_1(P) \in \mathrm{Tors}\, H^2(M;\mathbb{Z})$, that is, its first Chern class is a torsion class. The subspace $\mathscr{A}^f_M \subset \mathscr{A}_M$ of flat $\mathbb{T}$ connections on M is given by the union

$$\mathscr{A}^f_M = \bigsqcup_{\substack{P \\ c_1(P) \in \mathrm{Tors}\, H^2(M;\mathbb{Z})}} \mathscr{A}^f_P ,$$

where $\mathscr{A}^f_P$ is the space of connections Θ on P with curvature $F_\Theta = d\Theta = 0$. For each $\mathbb{T}$ bundle P with first Chern class torsion, the *moduli space of flat connections* on P is the quotient space $\mathscr{M}_P = \mathscr{A}^f_P/\mathscr{G}_P$. If P and P' are $\mathbb{T}$ bundles over M with $c_1(P) = c_1(P') \in \mathrm{Tors}\, H^2(M;\mathbb{Z})$, then there is a canonical isomorphism $\mathscr{M}_P \cong \mathscr{M}_{P'}$.

For each torsion class $p \in \mathrm{Tors}\, H^2(M;\mathbb{Z})$ let

$$\mathscr{A}^f_{M,p} = \bigsqcup_{\substack{P \\ c_1(P) = p}} \mathscr{A}^f_P .$$

The space $\mathcal{M}_{M,p}=\mathcal{A}^{f}_{M,p}/\sim$ is the moduli space of flat connections on principal $\mathbb{T}$ bundles over M with the first Chern class equal to p. For any principal $\mathbb{T}$ bundle $P\to M$ with $c_1(P)=p$, there is a natural isomorphism $\mathcal{M}_{M,p}\cong\mathcal{M}_P$. The moduli space $\mathcal{M}_M=\mathcal{A}^{f}_{M}/\sim$ of gauge equivalence classes of flat $\mathbb{T}$ connections on M is equal to the disjoint union

$$\mathcal{M}_M=\bigsqcup_{p\in \mathrm{Tors}\, H^2(M;\mathbb{Z})}\mathcal{M}_{M,p}.$$

Proposition II.2: Let M be a smooth manifold.
(i) There is a natural identification,

$$\mathcal{M}_M=H^1(M;\mathbb{T}).$$

(ii) $\pi_0(\mathcal{M}_M)\cong\mathrm{Tors}\, H^2(M;\mathbb{Z})$ and each connected component of $\mathcal{M}_M$ is diffeomorphic to the torus $H^1(M;\mathbb{R})/H^1(M;\mathbb{Z})$.

Proof: (i) is a consequence of a standard result in the theory of connections (Ref. 15, Chap. II) that provides the identification $\mathcal{M}_M=\Pi_\alpha\,\mathrm{Hom}(\pi_1(M_\alpha,*),\mathbb{T})$, where α indexes the connected components of M. For (ii), we consider the exact sequence of groups $0\to\mathbb{Z}\to\mathbb{R}\to\mathbb{T}\to 1$ and the induced exact cohomology sequence,

$$0\to H^1(M;\mathbb{Z})\to H^1(M;\mathbb{R})\to H^1(M;\mathbb{T})\xrightarrow{\delta}H^2(M;\mathbb{Z})\xrightarrow{i}H^2(M;\mathbb{R})\to\cdots. \tag{II.3}$$

Since $\mathrm{Im}\,\delta=\mathrm{Ker}\, i=\{p\,|\,p\in\mathrm{Tors}\, H^2(M;\mathbb{Z})\}$, for each torsion class p in $H^2(M;\mathbb{Z})$, we have $\mathcal{M}_{M,p}=\delta^{-1}(p)\cong\delta^{-1}(0)\cong H^1(M;\mathbb{R})/H^1(M;\mathbb{Z})$. □

B. The moduli space of flat $\mathbb{T}$ connections on a 2-manifold

Let Σ be a closed oriented two-dimensional manifold. We note that a principal $\mathbb{T}$ bundle $Q\to\Sigma$ has flat connections if and only if $c_1(Q)=0$, that is, the bundle Q is trivializable. For Σ the exact sequence (II.3) splits to

$$0\to H^1(\Sigma;\mathbb{Z})\to H^1(\Sigma;\mathbb{R})\to H^1(\Sigma;\mathbb{T})\to 0.$$

Thus, the moduli space of flat $\mathbb{T}$ connections on Σ is the torus

$$\mathcal{M}_\Sigma=H^1(\Sigma;\mathbb{T})\cong H^1(\Sigma,\mathbb{R})/H^1(\Sigma,\mathbb{Z}).$$

The space $H^1(\Sigma;\mathbb{R})$ is a symplectic vector space with symplectic form ω_Σ defined in de Rham cohomology by

$$\omega_\Sigma([\alpha],[\beta])=\int_\Sigma\alpha\wedge\beta,\quad [\alpha],[\beta]\in H^1(\Sigma;\mathbb{R}).$$

The integer lattice $H^1(\Sigma;\mathbb{Z})\subset H^1(\Sigma;\mathbb{R})$ is self-dual with respect to ω_Σ. The symplectic form ω_Σ on $H^1(\Sigma;\mathbb{R})$ descends to a symplectic form ω_Σ on the quotient torus $\mathcal{M}_\Sigma\cong H^1(\Sigma;\mathbb{R})/H^1(\Sigma;\mathbb{Z})$. We note that since the tangent space $T\mathcal{M}_\Sigma$ is identified with $2\pi i H^1(\Sigma;\mathbb{R})$, we have

$$\omega_\Sigma([\dot\eta],[\dot\eta'])=-\frac{1}{4\pi^2}\int_\Sigma\dot\eta\wedge\dot\eta', \tag{II.4}$$

for any $[\dot\eta],[\dot\eta']\in T\mathcal{M}_\Sigma$. Moreover, ω_Σ is normalized so that it gives $\mathcal{M}_\Sigma$ total volume equal to 1, that is,

$$\int_{\mathscr{M}_\Sigma} \frac{(\omega_\Sigma)^g}{g!} = 1,$$

where $g = 1/2\dim H^1(\Sigma;\mathbb{R})$. If $-\Sigma$ denotes the manifold Σ with the opposite orientation then by $H^1(-\Sigma;\mathbb{R})$ we understand the symplectic vector space $H^1(\Sigma;\mathbb{R})$ with symplectic form $-\omega_\Sigma$.

C. The moduli space of flat $\mathbb{T}$ connections on a 3-manifold

Let X be a compact oriented 3-manifold with boundary $\partial X \neq \varnothing$. The restriction of a connection over X to the boundary ∂X induces a map $r_x : \mathscr{M}_X \to \mathscr{M}_{\partial X}$. In view of the identifications $\mathscr{M}_X = H^1(X;\mathbb{T})$ and $\mathscr{M}_{\partial X} = H^1(\partial X;\mathbb{T})$, the moduli spaces $\mathscr{M}_X$ and $\mathscr{M}_{\partial X}$ are compact Abelian Lie groups and the restriction map r_X is a group homomorphism. We have the following.

Proposition II.5: (i) The image $\Lambda_X = \mathrm{Im}\{r_X : \mathscr{M}_X \to \mathscr{M}_{\partial X}\}$ *is a Lagrangian submanifold of the symplectic torus* $(\mathscr{M}_{\partial X}, \omega_{\partial X})$.

(ii) $r_X : \mathscr{M}_X \to \Lambda_X$ *is a principal fibre bundle with structure group a compact Abelian Lie group* H *that fits into the exact sequence*

$$1 \to \mathbb{T}^g \to H \to \mathrm{Tors}\, H^2(X;\mathbb{Z}) \to 0,$$

with $q = \dim H^1(X,\partial X;\mathbb{R}) - \dim H^0(\partial X;\mathbb{R}) + \dim H^0(X;\mathbb{R}) - \dim H^0(X,\partial X;\mathbb{R})$.

Proof: There is a long exact cohomology sequence (Ref. 16, Chap. V) associated to the pair of spaces $\partial X \subset X$, which for coefficients in $\mathbb{T}$ gives the exact sequence of homomorphisms of compact Abelian Lie groups:

$$1 \to H^0(X,\partial X;\mathbb{T}) \to H^0(X;\mathbb{T}) \to H^0(\partial X;\mathbb{T}) \to H^1(X,\partial X;\mathbb{T}) \to H^1(X;\mathbb{T}) \xrightarrow{r_X} \underset{\cap}{\Lambda_X}_{H^1(\partial X;\mathbb{T})} \to 1; \quad \text{(II.6)}$$

Λ_X is a compact subgroup and, therefore, a submanifold of $\mathscr{M}_{\partial X} = H^1(\partial X;\mathbb{T})$. At the same time the exact sequence of Abelian groups $0 \to \mathbb{Z} \to \mathbb{R} \to \mathbb{T} \to 1$ induces a long exact cohomology sequence for the relative cohomology $H^\bullet(X,\partial X)$ (Ref. 16, Chap. V). The relevant part of this sequence,

$$\cdots \to H^1(X,\partial X;\mathbb{T}) \to H^2(X,\partial X;\mathbb{Z}) \to H^2(X,\partial X;\mathbb{R}) \to \cdots,$$

implies $\pi_0(H^1(X,\partial X;\mathbb{T})) \cong \mathrm{Tors}\, H^2(X,\partial X;\mathbb{Z})$. The universal coefficient theorem (U.C.T.) (Ref. 16, Chap. V) together with Poincaré duality (P.D.) give the isomorphisms

$$\mathrm{Tors}\, H^2(X,\partial X;\mathbb{Z}) \underset{\text{(U.C.T.)}}{\cong} \mathrm{Tors}\, H_1(X,\partial X;\mathbb{Z}) \underset{\text{(P.D.)}}{\cong} \mathrm{Tors}\, H^2(X;\mathbb{Z}). \quad \text{(II.7)}$$

Thus, (ii) follows from (II.6) and (II.7). To prove (i), first recall that we identify $T\mathscr{M}_X = 2\pi i H^1(X;\mathbb{R})$ and $T\mathscr{M}_{\partial X} = 2\pi i H^1(\partial X;\mathbb{R})$. The differential $\dot{r}_X$ of r_X maps $T\mathscr{M}_X$ onto $T\Lambda_X \subset T\mathscr{M}_{\partial X}$, and we have

$$\omega_{\partial X}(\dot{r}_X[\dot{\Theta}], \dot{r}_X[\dot{\Theta}']) = -\frac{1}{4\pi^2}\int_{\partial X} \dot{r}_X(\dot{\Theta}) \wedge \hat{r}_X(\dot{\Theta}') = -\frac{1}{4\pi^2}\int_X d(\dot{\Theta}\wedge\dot{\Theta}')$$

$$= 0, \quad \text{for any } [\dot{\Theta}], [\dot{\Theta}'] \in T\mathscr{M}_X.$$

Since $\omega_{\partial X}|_{T\Lambda_X} \equiv 0$, Λ_X is an isotropic submanifold of $\mathscr{M}_{\partial X}$. The exact cohomology (homology) sequence with $\mathbb{R}$ coefficients for the pair $\partial X \subset X$ and the Poincaré duality isomorphism $H^{3-i}(X;\mathbb{R}) \cong H_i(X,\partial X;\mathbb{R})$ give the commutative diagram (Ref. 16, Chap. VI):

$$\begin{array}{ccccc} H^1(X;\mathbb{R}) & \xrightarrow{\dot{r}_X} & H^1(\partial X;\mathbb{R}) & \longrightarrow & H^2(X,\partial X;\mathbb{R}) \\ \downarrow\cong & & \downarrow\cong & & \downarrow\cong \\ H_2(X,\partial X;\mathbb{R}) & \longrightarrow & H_1(\partial X;\mathbb{R}) & \xrightarrow{j} & H_1(X;\mathbb{R}) \end{array}$$

where j and $\dot{r}_X$ are adjoint maps. Therefore dim Im $\dot{r}_X$=dim $H^1(\partial X;\mathbb{R})-$dim Ker j. Combining this with the exact sequence condition Im $\dot{r}_X$=Ker j, we find that dim Im $\dot{r}_X$ $=1/2$dim $H^1(\partial X;\mathbb{R})$. Therefore

$$L_X=\mathrm{Im}\{\dot{r}_X:H^1(X;\mathbb{R})\to H^1(\partial X;\mathbb{R})\}$$

is a Lagrangian subspace of $H^1(\partial X;\mathbb{R})$ and Λ_X a Lagrangian submanifold of $(\mathscr{M}_{\partial X},\omega_{\partial X})$. □

III. THE CHERN–SIMONS FUNCTIONAL

A. Induced principal bundles and induced connections

Consider the inclusion homomorphism $\rho:\mathbb{T}\hookrightarrow \mathrm{SU}(2)$,

$$e^{2\pi i\varphi}\mapsto\begin{pmatrix} e^{2\pi i\varphi} & 0 \\ 0 & e^{-2\pi i\varphi}\end{pmatrix},$$

from the circle group $\mathbb{T}$ into the group SU(2). The corresponding inclusion of the Lie algebra Lie $\mathbb{T}=2\pi i\mathbb{R}$ into the Lie algebra $\mathcal{S}\mathfrak{u}(2)$ is ρ_*: Lie $\mathbb{T}\hookrightarrow\mathcal{S}\mathfrak{u}(2)$,

$$\alpha\mapsto\begin{pmatrix}\alpha & 0\\ 0 & -\alpha\end{pmatrix}.$$

On the Lie algebra $\mathcal{S}\mathfrak{u}(2)$ we choose the symmetric invariant bilinear form

$$\langle\cdot,\cdot\rangle^\flat:\mathcal{S}\mathfrak{u}(2)\times\mathcal{S}\mathfrak{u}(2)\to\mathbb{R}$$

$$\times(a,b)\mapsto\langle a,b\rangle^\flat=\frac{1}{8\pi^2}\,\mathrm{Tr}(ab).$$

The bilinear form $\langle\cdot,\cdot\rangle^\flat$ is normalized so that the closed 3-form $\frac{1}{6}\langle\hat{\vartheta}\wedge[\hat{\vartheta},\hat{\vartheta}]\rangle^\flat$, with $\hat{\vartheta}$ the Maurer–Cartan form of SU(2) represents an integral cohomology class in $H^3(\mathrm{SU}(2);\mathbb{R})$. The commutator $[\cdot,\cdot]$ of two $\mathcal{S}\mathfrak{u}(2)$-valued forms is defined by taking the wedge product of forms and the natural Lie bracket in the $\mathcal{S}\mathfrak{u}(2)$ Lie algebra. The bilinear form $\langle\cdot,\cdot\rangle^\flat$ on $\mathcal{S}\mathfrak{u}(2)$ restricts to a symmetric bilinear form $\langle\cdot,\cdot\rangle$ on Lie $\mathbb{T}$:

$$\langle\cdot,\cdot\rangle:\mathrm{Lie}\ \mathbb{T}\times\mathrm{Lie}\ \mathbb{T}\to\mathbb{R}$$

$$\times(\alpha,\beta)\mapsto\langle\alpha,\beta\rangle=\langle\rho_*\alpha,\rho_*\beta\rangle^\flat=\frac{1}{4\pi^2}\,\alpha\beta.$$

As in the previous section we let ϑ denote the Maurer–Cartan form on $\mathbb{T}$ and note that $\rho^*(\hat{\vartheta})=\rho_*\vartheta$.

If M is a smooth manifold and π: $\rho\to M$ a principal $\mathbb{T}$ bundle, we can use the inclusion homomorphism ρ: $\mathbb{T}\hookrightarrow\mathrm{SU}(2)$ to extend the $\mathbb{T}$ bundle P on M to a principal SU(2) bundle $\hat{P}$ on M. The induced SU(2) bundle $\hat{P}$ is defined as the quotient $P\times_{\mathbb{T}}\mathrm{SU}(2)=(P\times\mathrm{SU}(2))/\mathbb{T}$ under the right $\mathbb{T}$ action given by $(p,a)\cdot\lambda=(p\cdot\lambda,\rho(\lambda)^{-1}a)$, for any $\lambda\epsilon\mathbb{T}$ and $(p,a)\epsilon P\times\mathrm{SU}(2)$. Let $[p,a]=(p,a)\cdot\mathbb{T}$ denote the $\mathbb{T}$ orbit through the point (p,a). We have a natural morphism of principal bundles,

$$\rho_P : P \to \hat{P},$$

$$p \mapsto [p, e],$$

covering the identity map on M. Moreover, for any connection Θ on the $\mathbb{T}$ bundle P there is an induced connection $\hat{\Theta}$ on the SU(2) bundle $\hat{P}$. The pullback of $\hat{\Theta}$ under the bundle map $\rho_P : P \to \hat{P}$ is $\rho_P^* \hat{\Theta} = \rho_* \Theta$. Similarly, the curvature forms $F_\Theta = d\Theta$ and $F_{\hat{\Theta}} = d\hat{\Theta} + \frac{1}{2}[\hat{\Theta}, \hat{\Theta}]$ are related by $\rho_P^* F_{\hat{\Theta}} = \rho_* F_\Theta$.

B. The Chern–Simons functional for a closed 3-manifold

Let X be a closed oriented 3-manifold. For any principal $\mathbb{T}$ bundle $P \to X$ there is an induced SU(2) bundle, $\hat{P} = P \times_{\mathbb{T}} \mathrm{SU}(2) \to X$, and for any $\mathbb{T}$ connection Θ on P an induced SU(2) connection $\hat{\Theta}$ on $\hat{P}$. Now, any principal SU(2) bundle over a manifold of dimension $\leqslant 3$ is trivializable. Thus $\hat{P}$ admits a global section, and we have the following.

Definition III.1: The Chern–Simons functional of a $\mathbb{T}$ connection Θ on $P \to X$ is defined by

$$S_{X,P}(\Theta) = \int_X \hat{s}^* \alpha(\hat{\Theta}) \quad (\mathrm{mod}\ 1),$$

where $\alpha(\hat{\Theta}) \in \Omega^3(\hat{P}; \mathbb{R})$ is the Chern–Simons form of the induced SU(2) connection $\hat{\Theta}$ on $\hat{P} = P \times_{\mathbb{T}} \mathrm{SU}(2)$,

$$\alpha(\hat{\Theta}) = \langle \hat{\Theta} \wedge F_{\hat{\Theta}} \rangle^\flat - \tfrac{1}{6} \langle \hat{\Theta} \wedge [\hat{\Theta}, \hat{\Theta}] \rangle^\flat,$$

and $\hat{s} : X \to \hat{P}$ is a global section of the SU(2) bundle $\hat{P}$.

The definition of $S_{X,P}(\Theta)$ does not depend on the choice of section $\hat{s} : X \to \hat{P}$. This is a consequence of the following property of the SU(2) Chern–Simons form $\alpha(\hat{\Theta})$.

Proposition III.2: If $\hat{s}, \hat{s}_1 : X \to \hat{P}$ are two sections of the SU(2) bundle $\hat{P}$ over X, then

$$\int_X \hat{s}_1^* \alpha(\hat{\Theta}) = \int_X \hat{s}^* \alpha(\hat{\Theta}) + \int_X d \langle Ad_{a^{-1}} \hat{s}^* \hat{\Theta} \wedge a^* \hat{\vartheta} \rangle^\flat - \int_X \frac{1}{6} a^* \langle \hat{\vartheta} \wedge [\hat{\vartheta}, \hat{\vartheta}] \rangle^\flat,$$

where $a : X \to \mathrm{SU}(2)$ is the map defined by $\hat{s}_1(x) = \hat{s}(x) \cdot a(x)$, for any $x \in X$.

Proof: The expression follows after simple computations from the standard relation $\hat{s}_1^* \hat{\Theta} = \mathrm{Ad}_{a^{-1}}(\hat{s} \hat{\Theta}) + a^* \hat{\vartheta}$. □

On the right-hand side of the equation in (III.2), the second integral vanishes by Stokes theorem and the assumption $\partial X = \varnothing$, while the last integral is an integer due to the normalization of the bilinear form $\langle \cdot, \cdot \rangle^\flat$ on $\mathcal{Su}(2)$. This proves that $S_{X,P}(\Theta)$ is well defined.

Remark III.3: If $P \to X$ is a trivializable bundle and $s : X \to P$ a section, then taking $\hat{s} : X \to \hat{P}$ to be $\hat{s} = \rho_P \circ s$, we have $\hat{s}^* \alpha(\hat{\Theta}) = s^* \langle \Theta \wedge F_\Theta \rangle$ and the Chern–Simons functional has as expected the expression

$$S_{X,P}(\Theta) = \int_X s^* \langle \Theta \wedge F_\Theta \rangle \quad (\mathrm{mod}\ 1).$$

Theorem III.4: *The Chern–Simons functional $S_{X,P} : \mathcal{A}_P \to \mathbb{R}/\mathbb{Z}$ defined for a closed oriented 3-manifold X and a principal $\mathbb{T}$ bundle $P \to X$ has the following properties:*

(*a*) *Functorality.*

If $\phi : P' \to P$ is a morphism of principal $\mathbb{T}$ bundles covering an orientation preserving diffeomorphism $\phi : X' \to X$ and if Θ is a connection on P, then

$$S_{X',P'}(\phi^*\Theta)=S_{X,P}(\Theta).$$

(b) Orientation

If $-X$ denotes the manifold X with the opposite orientation, then

$$S_{-X,P}(\Theta)=-S_{X,P}(\Theta).$$

(c) Disjoint union

Let X be the disjoint union $X=X_1\sqcup X_2$ and $P=P_1\sqcup P_2$ a principal $\mathbb{T}$ bundle over X. If Θ_i are connections on $P_i\to X_i$, then

$$S_{X_1\sqcup X_2,P_1\sqcup P_2}(\Theta_1\sqcup\Theta_2)=S_{X_1,P_1}(\Theta_1)+S_{X_2,P_2}(\Theta_2).$$

Proof: (a) To the principal $\mathbb{T}$ bundles P and P' there correspond the induced SU(2) bundles $\hat{P}$ and $\hat{P}'$, respectively. The morphism of $\mathbb{T}$ bundles $\phi:P'\to P$ induces a morphism of SU(2) bundles $\hat{\phi}:\hat{P}'\to\hat{P}$ covering $\phi:X'\to X$. It is defined by $\hat{\phi}([p',a])=[\phi(p'),a]$. For any section $\hat{s}':X'\to\hat{P}'$ we have a section $\hat{s}=\hat{\phi}\circ\hat{s}'\circ\overline{\phi^{-1}}:X\to\hat{P}$. Then, since $\hat{\phi}^*\hat{\Theta}$ is the SU(2) connection on $\hat{P}'$ induced from the $\mathbb{T}$ connection $\phi^*\Theta$ on P', we get

$$\begin{aligned}S_{X,P}(\Theta)&=\int_X\hat{s}^*\alpha(\hat{\Theta})\quad(\mathrm{mod}\ 1)\\&=\int_X(\overline{\phi^{-1}})^*\hat{s}'^*\hat{\phi}^*\alpha(\hat{\Theta})\quad(\mathrm{mod}\ 1)\\&=\int_{X'}\hat{s}'^*\alpha(\hat{\phi}^*\hat{\Theta})\quad(\mathrm{mod}\ 1)\\&=S_{X',P'}(\phi^*\Theta).\end{aligned}$$

(b) and (c) are straightforward. □

As a particular case of (a) of Theorem III. 4, we have the following.

Proposition III.5: The functional $S_{X,P}:\mathscr{A}_P\to\mathbb{R}/\mathbb{Z}$ for a closed 3-manifold X is invariant under the group of gauge transformations, that is,

$$S_{X,P}(\Theta\cdot u)=S_{X,P}(\Theta),$$

for any $u\in\mathscr{G}_X$. Hence $S_{X,P}$ descends to a functional on the quotient space $\mathscr{A}_P/\mathscr{G}_P$ of gauge equivalence classes of connections on P.

We also have the following property whose proof is straightforward.

Proposition III.6: The stationary points of the functional $S_{X,P}:\mathscr{A}_P\to\mathbb{R}/\mathbb{Z}$ are the flat connections, that is, $dS_{X,P}(\Theta)=0$ if and only if $F_\Theta=d\Theta=0$.

C. The Chern–Simons functional for a 3-manifold with boundary

Let X be a compact oriented 3-manifold with $\partial X\neq\varnothing$. Let $P\to X$ be a $\mathbb{T}$ bundle with first Chern class torsion. Then the $\mathbb{T}$ bundle $\partial P\to\partial X$ obtained by restricting P to the boundary ∂X is trivializable. As before, we consider the induced SU(2) bundle $\hat{P}=P\times_{\mathbb{T}}SU(2)$ on X and the bundle morphism $\rho_P:P\to\hat{P}$. For any section $s:\partial X\to\partial P$ we have a section $\rho_P\circ s:\partial X\to\partial\hat{P}$. Since the group SU(2) has $\pi_0(\mathrm{SU}(2))=\pi_1(\mathrm{SU}(2))=\pi_2(\mathrm{SU}(2))=0$, we can extend the section $\rho_P\circ s$ over ∂X to a global section $\hat{s}:X\to\hat{P}$. We make the following definition.

Definition III.7: The Chern–Simons functional of a $\mathbb{T}$ connection Θ on $P\to X$ and a section $s:\partial X\to\partial P$ is defined by

$$S_{X,P}(s,\Theta)=\int_X \hat{s}^*\alpha(\hat{\Theta}) \quad (\mathrm{mod}\ 1),$$

where $\alpha(\hat{\Theta})$ is the Chern–Simons form of the induced SU(2) connection $\hat{\Theta}$ on the bundle $\hat{P}$.

We have to check that the definition of $S_{X,P}(s,\Theta)$ does not depend on the choice of extension $\hat{s}$ of the section $\rho_P\circ s:\partial X\to\partial\hat{P}$. So let $\hat{s}_1:X\to\hat{P}$ to another extension. Then there is a map $a:X\to SU(2)$ such that $\hat{s}_1(x)=\hat{s}(x)\cdot a(x)$ and $a(x)=e$, for all $x\in\partial X$, where e denotes the identity element in SU(2). From (III.2) we get

$$\int_X \hat{s}_1^*\alpha(\hat{\Theta})-\int_X \hat{s}^*\alpha(\hat{\Theta})=\int_{\partial X}\langle \mathrm{Ad}_{a^{-1}}\hat{s}^*\hat{\Theta}\wedge a^*\hat{\vartheta}\rangle^b-\int_X \frac{1}{6}\,a^*\langle\hat{\vartheta}\wedge[\hat{\vartheta},\hat{\vartheta}]\rangle^b. \tag{III.8}$$

That $S_{X,P}(s,\Theta)$ is well defined by (III.7) follows now from the fact that, since $a|_{\partial X}=e$, the first integral on the rhs of (III.8) vanishes and the second term is an integer. To justify the last assertion we note the following.

Remark III.9: For each $a:X\to SU(2)$, let $W_{\partial X}(a)$ denote the functional,

$$W_{\partial X}(a)=-\int_X \frac{1}{6}\,a^*\langle\hat{\vartheta}\wedge[\hat{\vartheta},\hat{\vartheta}]\rangle^b.$$

Then, following the argument in Ref. 12, Sec. II, we let X' be a compact oriented 3-manifold with $\partial X'=\partial X$ and $a':X'\to\mathrm{SU}(2)$ a map such that $a'|_{\partial X'}=a|_{\partial X}$. Let $\widetilde{X}=X\cup(-X')$ be the closed 3-manifold obtained by gluing X and X' along their common boundary. The maps a and a' patch together into a map $\tilde{a}:\widetilde{X}\to\mathrm{SU}(2)$, and we have

$$W_{\partial X}(a)-W_{\partial X'}(a')=-\int_{\widetilde{X}}\frac{1}{6}\,\tilde{a}^*\langle\hat{\vartheta}\wedge[\hat{\vartheta},\hat{\vartheta}]\rangle^b\in\mathbb{Z}. \tag{III.10}$$

Since $\widetilde{X}$ is closed, the rhs is an integer due to the normalization of $\langle\cdot,\cdot\rangle^b$. Thus, $W_{\partial X}(a)$ depends modulo integers only on the restriction of $a:X\to\mathrm{SU}(2)$ to ∂X.

If $a|_{\partial X}=e$, we can choose $a':X'\to\mathrm{SU}(2)$ with $a'|_{\partial X'}=a|_{\partial X}$ to be the constant map $a'(x')=e$, for all $x'\in X'$. Then $W_{\partial X'}(a')=0$ and therefore $W_{\partial X}(a)\in\mathbb{Z}$.

Remark III.11: As in (III.3), if $P\to X$ is a trivializable $\mathbb{T}$ bundle and $s:X\to P$ a global section, then for any connection Θ on P the Chern–Simons functional $S_{X,P}(s,\Theta)$ is given by the expression

$$S_{X,P}(s,\Theta)=\int_X s^*\langle\Theta\wedge F_\Theta\rangle \quad (\mathrm{mod}\ 1).$$

The following proposition describes the dependence of the Chern–Simons functional $S_{X,P}(s,\Theta)$ on the section $s:\partial X\to\partial P$.

Proposition III.12: Let $s,s_1:\partial X\to\partial P$ be two sections of the restriction of the $\mathbb{T}$ bundle $P\to X$ to ∂X and let Θ be a connection on P. Then

$$S_{X,P}(s_1,\Theta)=S_{X,P}(s,\Theta)+\int_{\partial X}\langle s^*\Theta\wedge\lambda^*\vartheta\rangle \quad (\mathrm{mod}\ 1),$$

where $\lambda:\partial X\to\mathbb{T}$ is the map defined by $s_1(x)=s(x)\cdot\lambda(x)$, for all $x\in\partial X$.

Proof: The sections $\rho_P\circ s,\rho_P\circ s_1:\partial X\to\partial\hat{P}$ extend to sections $\hat{s},\hat{s}_1:X\to\hat{P}$. The map $a:X\to\mathrm{SU}(2)$ defined by $\hat{s}_1(x)=\hat{s}(x)\cdot a(x)$, for any $x\in X$, restricts to $a|_{\partial X}=\rho\circ\lambda$ over the boundary of X. The difference $S_{X,P}(s_1,\Theta)-S_{X,P}(s,\Theta)$ is given by an expression identical to the rhs of

the equation (III.8). The second term $W_{\partial X}(a)$ is also in this case an integer. To see that, choose in (III.10) a 3-manifold X' with $\partial X' = \partial X$, for which the map $\lambda: \partial X \to \mathbb{T}$ extends to a map $\lambda': X' \to \mathbb{T}$. Let $a' = \rho \cdot \circ \lambda': X' \to \mathrm{SU}(2)$. Then, since $\rho^* \hat{\vartheta} = \rho_* \vartheta$, we get that

$$W_{\partial X'}(a') = -\int_{X'} \frac{1}{6}\, a'^* \langle \hat{\vartheta} \wedge [\hat{\vartheta}, \hat{\vartheta}] \rangle^b = -\int_{X'} \frac{1}{6}\, \lambda'^* \langle \vartheta \wedge [\vartheta, \vartheta] \rangle = 0.$$

Thus (III.10) implies that $W_{\partial X}(a) \in \mathbb{Z}$. The first term on the rhs of (III.8) gives

$$\int_{\partial X} \langle \mathrm{Ad}_{a^{-1}} \hat{s}^* \hat{\Theta} \wedge a^* \hat{\vartheta} \rangle^b = \int_{\partial X} \langle s^* \Theta \wedge \lambda^* \vartheta \rangle,$$

which proves the proposition. □

Theorem III.13: *The Chern–Simons functional $S_{X,P}(\cdot,\cdot)$, defined for a compact oriented 3-manifold X with nonempty boundary and a principal $\mathbb{T}$-bundle $P \to X$ with first Chern class torsion, has the following properties.*

(a) Functoriality

Let $\phi: P' \to \underline{P}$ be a morphism of principal $\mathbb{T}$-bundles covering an orientation preserving diffeomorphism $\underline{\phi}: X' \to X$. If Θ is a connection on P and $s': \partial X' \to \partial P'$ a section of P' over the boundary of X', then

$$S_{X',P'}(s', \phi^* \Theta) = S_{X,P}(\partial \phi \circ s' \circ \partial \overline{\phi}^{-1}, \Theta),$$

where $\partial \phi: \partial P' \to \partial P$ and $\partial \overline{\phi}: \partial X' \to \partial X$ are restrictions to the boundary.

(b) Orientation

If $-X$ denotes the manifold X with the opposite orientation and $s: \partial X \to \partial P$, then

$$S_{-X,P}(s, \Theta) = -S_{X,P}(s, \Theta).$$

(c) Disjoint union

Let X be the disjoint union $X = X_1 \sqcup X_2$ and $P = P_1 \sqcup P_2$ a principal $\mathbb{T}$ bundle over *X with first Chern class torsion. If Θ_i are connections on $P_i \to X_i$ and $s_i: \partial X_i \to \partial P_i$ sections over the boundary, then*

$$S_{X_1 \sqcup X_2, P_1 \sqcup P_2}(s_1 \sqcup S_2, \Theta_1 \sqcup \Theta_2) = S_{X_1,P_1}(s_1, \Theta_1) + S_{X_2,P_2}(s_2, \Theta_2).$$

Proof: (a) Let $\hat{P} = P \times_{\mathbb{T}} \mathrm{SU}(2)$ and $\hat{P}' = P \times_{\mathbb{T}} \mathrm{SU}(2)$ be the induced SU(2) bundles, $\hat{\phi}: \hat{P}' \to \hat{P}$ the induced bundle map, and $\hat{\Theta}$ the induced connection on $\hat{P}$. The sections $s': \partial X' \to \partial P'$ and $s = \partial \phi \circ s' \circ \partial \phi^{-1}: \partial X \to \partial P$ determine sections $\rho_{P'} \circ \underline{s'}: \partial X' \to \partial \hat{P}'$ and $\rho_P \circ s: \partial X \to \partial \hat{P}$. Let $\hat{s}': X' \to \hat{P}'$ be an extension of $\rho_{P'} \circ s'$. Then $\hat{s} = \hat{\phi} \circ \hat{s}' \circ \phi^{-1}: X \to \hat{P}$ is an extension of $\rho_P \circ s$. Moreover, the SU(2) connection on $\hat{P}'$ induced from the $\mathbb{T}$ connection $\phi^* \Theta$ on P' is $\hat{\phi}^* \hat{\Theta}$. Applying the definition (III.7), we get

$$\begin{aligned} S_{X,P}(s, \Theta) &= \int_X \hat{s}^* \alpha(\hat{\Theta}) \quad (\mathrm{mod}\ 1) \\ &= \int_X (\overline{\phi}^{-1})^* \hat{s}'^* \hat{\phi}^* \alpha(\hat{\Theta}) \quad (\mathrm{mod}\ 1) \\ &= \int_{X'} \hat{s}'^* \alpha(\hat{\phi}^* \hat{\Theta}) \quad (\mathrm{mod}\ 1) \\ &= S_{X',P'}(s', \phi^* \Theta). \end{aligned}$$

(b) and (c) are straightforward. □

As a consequence of (III.12) and (a) of Theorem (III.13) we have the following.

Proposition III.14: The functional $S_{X,P}(\cdot,\cdot)$ changes under the group of gauge transformations as

$$S_{X,P}(s,\Theta\cdot u)=S_{X,P}(s,\Theta)+\int_{\partial X}\langle s^*\Theta\wedge u^*\vartheta\rangle \pmod 1,$$

for any $u\in\mathscr{G}_X=\mathrm{Map}(X,\mathbb{T})$.

We also note the following lemma

Lemma III.15: Let X be a compact oriented 3-manifold and $P\to X$ a $\mathbb{T}$ bundle with first Chern class torsion. If Θ and Θ' are connections on P and $s:\partial X\to\partial P$, then

$$S_{X,P}(s,\Theta')-S_{X,P}(s,\Theta)=-\int_{\partial X}s^*\langle\Theta\wedge\omega\rangle+2\int_X\langle F_\Theta\wedge\omega\rangle+\int_X\langle\omega\wedge d\omega\rangle \pmod 1,$$

where $\omega=\Theta'-\Theta$.

Proof: Let $\hat\Theta$ and $\hat\Theta'$ be the SU(2) connections on $\hat P=P\times_{\mathbb{T}}\mathrm{SU}(2)$ induced from the $\mathbb{T}$ connections Θ and Θ' on P. Let $\hat\omega=\hat\Theta'-\hat\Theta$. Then $\rho_P^*\hat\omega=\rho_*\omega$ and $[\hat\omega,\hat\omega]=0$. Routine computations give us, for the difference of the Chern–Simons forms of $\hat\Theta'$ and $\hat\Theta$, the expression

$$\alpha(\hat\Theta')-\alpha(\hat\Theta)=-d\langle\hat\Theta\wedge\hat\omega\rangle^b+2\langle F_{\hat\Theta}\wedge\hat\omega\rangle^b+\langle\hat\omega\wedge d\hat\omega\rangle^b.$$

The 3-forms $\langle\hat{F_\Theta}\wedge\hat\omega\rangle^b$ on $\hat P$ and $\langle F_\Theta\wedge\omega\rangle$ on P are both lifts to $\hat P$ and P, respectively, of the same 3-form on X. The same statement applies to the 3-forms $\langle\hat\omega\wedge d\hat\omega\rangle^b$ on $\hat P$ and $\langle\omega\wedge d\omega\rangle$ on P. Letting $\hat s:X\to\hat P$ be an extension to X of the section $\rho_P\circ s:\partial X\to\partial\hat P$, we can write

$$\int_X\hat s^*d\langle\hat\Theta\wedge\hat\omega\rangle^b=\int_{\partial X}s^*\rho_P^*\langle\hat\Theta\wedge\hat\omega\rangle^b=\int_{\partial X}s^*\langle\rho_*\Theta\wedge\rho_*\omega\rangle^b=\int_{\partial X}s^*\langle\Theta\wedge\omega\rangle.$$

Putting the above results together, we find that the difference,

$$S_{X,P}(s,\Theta')-S_{X,P}(s,\Theta)=\int_X\hat s^*\alpha(\hat\Theta')-\int_X\hat s^*\alpha(\hat\Theta),$$

is given by the expression in Lemma III.15. □

Lemma III.16: Let Θ be a connection on the $\mathbb{T}$ bundle $P\to X$ with $c_1(P)$ torsion and $s:\partial X\to\partial P$ a section over the boundary of X. If $u:X\to\mathbb{T}$ is an element of $\mathscr{G}_X$, then

$$\int_X\langle F_\Theta\wedge u^*\vartheta\rangle=\int_{\partial X}\langle s^*\Theta\wedge u^*\vartheta\rangle \pmod 1.$$

Proof: Let X' be a compact oriented 3-manifold with $\partial X'=\partial X$ and such that the map $u:\partial X\to\mathbb{T}$ extends to a map $u':X'\to\mathbb{T}$. Then glue the manifolds X and X' along their common boundary into the closed 3-manifold $\widetilde X=X\sqcup(-X')$ and extend the connection $\partial\Theta$ on ∂X to a connection Θ' over X'. Let $\widetilde\Theta=\Theta\sqcup\Theta'$ and let $\widetilde u:\widetilde X\to\mathbb{T}$ denote the map with restrictions $\widetilde u|_X=u$ and $\widetilde u|_{X'}=u'$. The integral

$$\int_{\widetilde X}\langle F_{\widetilde\Theta}\wedge\widetilde u^*\vartheta\rangle=\int_X\langle F_\Theta\wedge u^*\vartheta\rangle-\int_{X'}\langle F_{\Theta'}\wedge u'^*\vartheta\rangle\in\mathbb{Z}, \tag{III.17}$$

since $F_{\tilde{\Theta}}/2\pi i$ and $\tilde{u}^*\vartheta/2\pi i$ represent integral cohomology classes. Thus $\int_X \langle F_\Theta \wedge u^*\vartheta\rangle$ depends modulo $\mathbb{Z}$ only on the restrictions $\partial\Theta$ and $u|_{\partial X}$ to the boundary ∂X. To prove the lemma take Θ' extending $\partial\Theta$ over X' to be a connection on the trivial bundle $\pi': P' = X' \times \mathbb{T} \to X'$. Then, if $s': X' \to P'$ is a section, we have, from (III.17),

$$\begin{aligned}\int_X \langle F_\Theta \wedge u^*\vartheta\rangle &= \int_{X'} \langle F_{\Theta'} \wedge u'^*\vartheta\rangle \quad (\mathrm{mod}\ 1)\\ &= \int_{X'} s'^*\langle d\Theta' \wedge \pi'^* u'^*\vartheta\rangle \quad (\mathrm{mod}\ 1),\\ &= \int_{\partial X'} \langle s'^*\Theta' \wedge u'^*\vartheta\rangle \quad (\mathrm{mod}\ 1)\\ &= \int_{\partial X} \langle s^*\Theta \wedge u^*\vartheta\rangle \quad (\mathrm{mod}\ 1),\end{aligned}$$

where $s: \partial X \to \partial P$ is any section over the boundary. □

Proposition III.18: Let $P \to X$ be a $\mathbb{T}$ bundle with first Chern class torsion. For any section $s: \partial X \to \partial P$, the functional $S_{X,P}(s,\cdot): \mathscr{A}_P^f \to \mathbb{R}/\mathbb{Z}$ defined on the space of flat connections is invariant under the group $\mathscr{G}_X$ of gauge transformations, that is,

$$S_{X,P}(s, \Theta \cdot u) = S_{X,P}(s,\Theta),$$

for any $u \in \mathscr{G}_X$. Hence $S_{X,P}(s,\cdot)$ descends to a functional on the moduli space $\mathscr{M}_P$ of flat connections on P.

Proof: If Θ is flat, then Lemma III.16 gives $\int_{\partial X}\langle s^*\Theta \wedge u^*\vartheta\rangle = 0 \ (\mathrm{mod}\ 1)$, for any $u \in \mathscr{G}_X$. The proposition follows from (III.14) and the above result. □

IV. THE PREQUANTUM LINE BUNDLE

A. The line bundle over the moduli space of flat connections

As shown in Sec. II, to each closed oriented two-dimensional manifold Σ there corresponds the symplectic torus $(\mathscr{M}_\Sigma, \omega_\Sigma)$ of gauge equivalence classes of flat $\mathbb{T}$ connections on Σ. The moduli space $\mathscr{M}_\Sigma$ can be regarded[14] as the symplectic quotient of the infinite-dimensional affine space of connections on a trivializable $\mathbb{T}$ bundle over Σ by the action of the group of gauge transformations $\mathscr{G}_\Sigma$. As emphasized in Ref. 17 this point of view of $\mathscr{M}_\Sigma$ arises naturally in connection to the three-dimensional Chern–Simons theory.

Given a positive even integer k, we construct in this section a Hermitian line bundle $\mathscr{L}_\Sigma$ over $\mathscr{M}_\Sigma$ with a unitary connection with curvature $-2\pi i k \omega_\Sigma$. In the language of geometric quantization, such a line bundle is called a *prequantum line bundle* for the symplectic manifold $(\mathscr{M}_\Sigma, k\omega_\Sigma)$. The set of prequantum line bundles for $(\mathscr{M}_\Sigma, k\omega_\Sigma)$ is a principal homogeneous space for the cohomology group $H^1(\mathscr{M}_\Sigma; \mathbb{T})$. Since $H^1(\mathscr{M}_\Sigma; \mathbb{T})$ is nontrivial, by general geometric quantization arguments the choice of the prequantum line bundle is not unique. However, the ''Chern–Simons line construction'' from Ref. 12, Sec. II, which we are going to apply here, singles out a prequantum line bundle $\mathscr{L}_\Sigma$.

Let $Q \to \Sigma$ be a trivializable $\mathbb{T}$ bundle over the closed oriented surface Σ. To each connection η on Q we are going to associate a Hermitian line L_η by the following construction. Let $\Gamma(\Sigma; Q)$ stand for the space of sections of the line bundle $Q \to \Sigma$. Any two sections $s, s_1 \in \Gamma(\Sigma; Q)$ are

related by an element $\lambda:\Sigma\to\mathbb{T}$ of the group of gauge transformations $\mathscr{G}_\Sigma$, defined by $s_1(x)=s(x)\cdot\lambda(x)$. Thus we let L_η be the space of functions $f:\Gamma(\Sigma;Q)\to\mathbb{C}$ satisfying the relation

$$f(s_1)=f(s)c_\Sigma(s^*\eta,\lambda), \tag{IV.1}$$

where $c_\Sigma(s^*\eta,\lambda)$ is the $\mathbb{T}$-valued cocycle,

$$c_\Sigma:2\pi i\Omega^1(\Sigma;\mathbb{R})\times\mathscr{G}_\Sigma\to\mathbb{T},$$

$$c_\Sigma(\alpha,\lambda)=\exp\left(\pi ik\int_\Sigma\langle\alpha\wedge\lambda^*\vartheta\rangle\right)=\exp\left(-\pi ik\int_\Sigma\frac{\alpha}{2\pi i}\wedge\lambda^*(\vartheta/2\pi i)\right). \tag{IV.2}$$

c_Σ satisfies the cocycle identity,

$$c_\Sigma(\alpha,\lambda_1\lambda_2)=c_\Sigma(\alpha,\lambda_1)c_\Sigma(\alpha+\lambda_1^*\vartheta,\lambda_2), \tag{IV.3}$$

if and only if

$$\frac{k}{2}\int_\Sigma\lambda_1^*\left(\frac{\vartheta}{2\pi i}\right)\wedge\lambda_2^*\left(\frac{\vartheta}{2\pi i}\right)\in\mathbb{Z},$$

for any $\lambda_1,\lambda_2\in\mathscr{G}_\Sigma$. This imposes the restriction $k\in2\mathbb{Z}$. Hence, for the remainder of this paper, we are going to assume that k is a positive even integer. We note that $\dim_{\mathbb{C}}L_\eta=1$ and that a section $s:\Sigma\to Q$ determines a trivialization $L_\eta\cong\mathbb{C}$ by mapping $f\mapsto f(s)$. The complex line L_η has a Hermitian inner product determined by the standard inner product of $\mathbb{C}$: $(f_1,f_2)=\overline{f_1(s)}f_2(s)$.

As in Ref. 12, we have the property that, as η varies over the space $\mathscr{A}_Q$ of connections on Q, the lines L_η fit together smoothly into a Hermitian line bundle $\hat{\mathscr{L}}_Q\to\mathscr{A}_Q$.

Theorem IV.4: *Let Σ be a closed oriented 2-manifold. The assignment,*

$$\eta\mapsto L_{\Sigma,\eta}=L_\eta,$$

for η connections on trivializable $\mathbb{T}$ bundles over Σ, is smooth and satisfies the following.

(a) Functoriality

Let $\psi:Q'\to Q$ *be a morphism between trivializable $\mathbb{T}$ bundles covering an orientation preserving diffeomorphism $\psi:\Sigma'\to\Sigma$ and let η be a connection on Q. Then there is an induced isometry,*

$$\psi^*:L_\eta\to L_{\psi^*\eta},$$

$$f\mapsto\psi^*f,$$

where $\psi^*f\in L_{\psi^*\eta}$ is defined by $(\psi^*f)(s')=f(\psi\circ s'\circ\overline{\psi^{-1}})$, for any $s'\in\Gamma(\Sigma';Q')$.

(b) Orientation

If $-\Sigma$ denotes the manifold Σ with the opposite orientation, then there is a natural isometry,

$$L_{-\Sigma,\eta}\cong\overline{L_{\Sigma,\eta}}.$$

(c) Disjoint union

If $\Sigma=\Sigma_1\sqcup\Sigma_2$ is a disjoint union and η_i are connections on trivializable $\mathbb{T}$ bundles over Σ_i, then there is a natural isometry,

$$L_{\eta_1\sqcup\eta_2}\cong L_{\eta_1}\otimes L_{\eta_2}.$$

Proof: (a) To check that $\psi^* f$ does indeed belong to the line $L_{\psi^*\eta}$, let $\lambda':\Sigma'\to\mathbb{T}$ and let $s',s_1'\in\Gamma(\Sigma';Q')$ be sections related by $s_1'(x')=s'(x')\cdot\lambda'(x')$, for all $x'\in\Sigma'$. Then $(\psi\circ s_1'\circ\psi^{-1})(x)=(\psi\circ s'\circ\psi^{-1})(x)\cdot\lambda(x)$ at any $x\in\Sigma$, where $\lambda=\lambda'\circ\psi^{-1}:\Sigma\to\mathbb{T}$. Therefore we obtain

$$(\psi^* f)(s_1')=f(\psi\circ s_1'\circ\overline{\psi^{-1}})=f(\psi\circ s'\circ\overline{\psi^{-1}})c_\Sigma((\overline{\psi^{-1}})^* s'^*\psi^*\eta,\lambda)$$
$$=(\psi^* f)(s')c_{\Sigma'}(s'^*(\psi^*\eta),\lambda').$$

(b) follows from the fact that the $\mathbb{T}$-valued cocycle c_Σ changes to the complex conjugate expression when the orientation of Σ is reversed.

(c) η_i are connections on trivializable $\mathbb{T}$ bundles $Q_i\to\Sigma_i$. Let $s_i:\Sigma_i\to Q_i$ be some sections and $\lambda_i:\Sigma_i\to\mathbb{T}$ gauge transformations and set $s=s_1\sqcup s_2$, $\eta=\eta_1\sqcup\eta_2$, and $\lambda=\lambda_1\sqcup\lambda_2$. Then $c_\Sigma(s^*\eta,\lambda)=c_{\Sigma_1}(s_1^*\eta_1,\lambda_1)c_{\Sigma_2}(s_2^*\eta_2,\lambda_2)$, and this induces the isometry

$$L_{\eta_1}\otimes L_{\eta_2}\to L_{\eta_1\sqcup\eta_2},$$

$$f_1\otimes f_2\mapsto f,$$

where $f(s)=f_1(s_1)f_2(s_2)$, for any section $s=s_1\sqcup s_2$. □

Each section $s:\Sigma\to Q$ induces a trivialization of the line bundle $\hat{\mathscr{L}}_Q$, given by the unitary section,

$$\gamma_s:\mathscr{A}_Q\to\hat{\mathscr{L}}_Q,$$

with $\gamma_s(\eta)\in L_\eta$ defined, for each η, by $(\gamma_s(\eta))(s)=1$. We define the connection $\hat{\nabla}$ on $\hat{\mathscr{L}}_Q$ by

$$\hat{\nabla}\gamma_s=-2\pi ik\hat{\theta}_s\gamma_s, \tag{IV.5}$$

where $\hat{\theta}_s$ is the 1-form on $\mathscr{A}_Q$, given by

$$(\dot{\eta}\lrcorner\hat{\theta}_s)_\eta=-\frac{1}{2}\int_\Sigma s^*\langle\eta\wedge\dot{\eta}\rangle, \tag{IV.6}$$

for any $\eta\in\mathscr{A}_Q$ and $\dot{\eta}\in T_\eta\mathscr{A}_Q$. If $\psi\in\mathscr{G}_Q=\mathrm{Aut}(Q)$, the new section $\psi\circ s:\Sigma\to Q$ induces a new section $\gamma_{\psi\circ s}:\mathscr{A}_Q\to\mathscr{L}_Q$ related to γ_s by

$$\gamma_{\psi\circ s}(\eta)=c_\Sigma(s^*\eta,\lambda_\psi)^{-1}\gamma_s(\eta), \tag{IV.7}$$

where $\lambda_\psi:\Sigma\to\mathbb{T}$ is the map associated to ψ. It follows that the connection form $\hat{\theta}_{\psi\circ s}$ is related to $\hat{\theta}_s$ by

$$(\hat{\theta}_{\psi\circ s})_\eta=(\hat{\theta}_s)_\eta+\frac{1}{2\pi ik}\frac{dc_\Sigma(s^*\eta,\lambda_\psi)}{c_\Sigma(s^*\eta,\lambda_\psi)},$$

which implies that the definition (IV.5) of $\hat{\nabla}$ does not depend on the choice of trivialization of $\hat{\mathscr{L}}_Q\to\mathscr{A}_Q$. Since γ_s is a unitary section and $\hat{\theta}_s$ is real, the connection $\hat{\nabla}$ is compatible with the Hermitian structure on $\hat{\mathscr{L}}_Q$. The curvature of $\hat{\nabla}$ is $-2\pi ik\omega_\Sigma$, where

$$\omega_\Sigma(\dot{\eta},\dot{\eta}')=(d\hat{\theta}_s)(\dot{\eta},\dot{\eta}')=-\int_\Sigma\langle\dot{\eta}\wedge\dot{\eta}'\rangle, \tag{IV.8}$$

is the standard symplectic form on $T\mathscr{A}_Q\cong 2\pi i\Omega^1(\Sigma;\mathbb{R})$.

It follows from (a) of Theorem IV.4 that the action of the group of gauge transformations $\mathcal{G}_Q$ on $\mathcal{A}_Q$ lifts to the line bundle $\hat{\mathcal{L}}_Q$ preserving the Hermitian metric. The lift of $\mathcal{G}_Q$ to $\hat{\mathcal{L}}_Q$ preserves also the connection $\hat{\mathbf{\nabla}}$, and, as $-2\pi i k\omega_\Sigma$ is the curvature of $\hat{\mathbf{\nabla}}$, the $\mathcal{G}_Q$-action on $\mathcal{A}_Q$ preserves the symplectic form ω_Σ.

Since the action of $\mathcal{G}_Q$ on $\mathcal{A}_Q$ is symplectic and lifts to the line bundle $\pi:\hat{\mathcal{L}}_Q\to\mathcal{A}_Q$ preserving the Hermitian metric and the connection $\hat{\mathbf{\nabla}}$, there exists a moment map for this action. Applying the general construction in (Ref. 18, Sec. IV), we find that the explicit form of the moment map $\mu:\mathcal{A}_Q\to(\text{Lie }\mathcal{G}_Q)^*$, $\eta\mapsto\mu(\eta)$ is

$$(\mu(\eta))(\xi)=2k(d\xi\lrcorner\hat{\theta}_s)_\eta=-k\int_\Sigma\langle s^*\eta\wedge d\xi\rangle=-k\int_\Sigma\langle s^*\ d\eta\wedge\xi\rangle,$$

for any $\xi\in\text{Lie }\mathcal{G}_Q\cong 2\pi i\Omega^0(\Sigma;\mathbb{R})$. The preimage $\mu^{-1}(0)$, which is $\mathcal{G}_Q$ invariant, is the space $\mathcal{A}_Q^f$ of flat connections. Hence the space $\mathcal{M}_Q$ of equivalence classes of flat connections is the symplectic quotient, $\mathcal{M}_Q=\mathcal{A}_Q//\mathcal{G}_Q=\mu^{-1}(0)/\mathcal{G}_Q$. The symplectic form ω_Σ on $\mathcal{A}_Q$ pushes down to a symplectic form on $\mathcal{M}_Q$, which we continue to denote by ω_Σ.

The group $\mathcal{G}_Q$ does not act freely on $\mathcal{A}_Q$. However, the stabilizer at a point $\eta\in\mathcal{A}_Q$ is the subgroup $Z\subset\mathcal{G}_Q$ of constant gauge transformations $Z\cong\{\lambda:\Sigma\to\mathbb{T}|\lambda=\text{constant}\}\cong\mathbb{T}$ and Z acts trivially on L_η. Therefore the line bundle $\hat{\mathcal{L}}_Q\to\mathcal{A}_Q^f=\mu^{-1}(0)$ pushes down to a line bundle $\mathcal{L}_Q\to\mathcal{M}_Q$. We have the identification $\Gamma(\mathcal{M}_Q;\mathcal{L}_Q)=\Gamma(\mathcal{A}_Q^f;\hat{\mathcal{L}}_Q)^{\mathcal{G}_Q}$ of the space of sections of $\mathcal{L}_Q\to\mathcal{M}_Q$ with the space of $\mathcal{G}_Q$-invariant sections of $\hat{\mathcal{L}}_Q\to\mathcal{A}_Q^f$. The Hermitian metric and connection on $\hat{\mathcal{L}}_Q$ also push down to $\mathcal{L}_Q$.

Remark IV.9: Let us assume for the moment that in constructing the Hermitian line bundle $\hat{\mathcal{L}}_Q\to\mathcal{A}_Q$ we take the cocycle c_Σ to have, instead of (IV.2), the more general form

$$c_\Sigma(\alpha,\lambda)=\epsilon_\Sigma([\lambda])\exp\left(\pi ik\int_\Sigma\langle\alpha\wedge\lambda^*\vartheta\rangle\right),$$

where $[\lambda]$ denotes the homotopy class of the map $\lambda:\Sigma\to\mathbb{T}$. In order to satisfy the cocycle condition (IV.3), the $\mathbb{T}$-valued multiplier ϵ_Σ must have the property

$$\epsilon_\Sigma([\lambda_1][\lambda_2])=\epsilon_\Sigma([\lambda_1])\epsilon_\Sigma([\lambda_2])\exp\left(\pi ik\int_\Sigma\langle\lambda_1^*\vartheta\wedge\lambda_2^*\vartheta\rangle\right),$$

and k can be any integer. All that has been said up to this point, except (a) of Theorem IV.4, remains true. We get a Hermitian line bundle over $\mathcal{A}_Q$ with a unitary connection with curvature $-2\pi ik\omega_\Sigma$ and the action of $\mathcal{G}_Q$ lifts, preserving the metric and the connection. The restriction of this line to $\mathcal{A}_Q^f$ pushes down to a prequantum line bundle over $\mathcal{M}_Q$. The initial choice (IV.2) for c_Σ, in which ϵ_Σ is taken to be the trivial multiplier (which forces k to be even), is dictated by the requirement to satisfy also the functoriality property [see (a) of Theorem IV.4]. This can be seen immediately if we follow the proof given for (a) of Theorem IV.4. Only for this choice do we get a prequantum line bundle over the moduli space $\mathcal{M}_\Sigma$ for which there is a lift of the group of orientation preserving diffeomorphisms of Σ. The prequantum line bundle over $\mathcal{M}_\Sigma$ picked out by the above stated reasons is also closely connected to the Chern–Simons theory in three dimensions. This connection will become clear later in this section.

Given any two trivializable $\mathbb{T}$ bundles Q and Q' over Σ, let $\psi:Q'\to Q$ be a bundle isomorphism. It follows from (a) of Theorem IV.4 that the induced isomorphism between the spaces of

connections lifts to an isomorphism of line bundles,

$$\begin{array}{ccc} \hat{\mathcal{L}}_Q & \xrightarrow{\psi^*} & \hat{\mathcal{L}}_{Q'} \\ \downarrow & & \downarrow \\ \mathcal{A}_Q & \xrightarrow{\psi^*} & \mathcal{A}_{Q'} \end{array} \tag{IV.10}$$

To each gauge transformation $\phi \in \mathcal{G}_Q$ there corresponds a gauge transformation $\psi^{-1} \circ \phi \circ \psi \in \mathcal{G}_{Q'}$, so that for any connection $\eta \in \mathcal{A}_Q$ we have the commutative diagram

$$\begin{array}{ccc} L_\eta & \xrightarrow{\psi^*} & L_{\psi^*\eta} \\ {\scriptstyle \phi^*}\downarrow & & \downarrow{\scriptstyle (\psi^{-1}\circ\phi\circ\psi)^*} \\ L_{\phi^*\eta} & \xrightarrow{\psi^*} & L_{\psi^*\phi^*\eta} \end{array} \tag{IV.11}$$

of line isomorphisms defined by (a) of Theorem IV (a). The line bundle isomorphism (IV.10) pushes down to an isomorphism between the quotient line bundles by the groups of gauge transformations,

$$\begin{array}{ccc} \mathcal{L}_Q = \hat{\mathcal{L}}_Q/\mathcal{G}_Q & \longrightarrow & \mathcal{L}_{Q'} = \hat{\mathcal{L}}_{Q'}/\mathcal{G}_{Q'} \\ \downarrow & & \downarrow \\ \mathcal{A}_Q/\mathcal{G}_Q & \longrightarrow & \mathcal{A}_{Q'}/\mathcal{G}_{Q'} \end{array} \tag{IV.12}$$

It follows from (IV.11) that the above isomorphism is canonical, that is, independent of the choice of $\mathbb{T}$-bundle isomorphism $\psi: Q' \to Q$. Restricting the above considerations to the subspaces of flat connections, we obtain that, for any two trivializable $\mathbb{T}$ bundles Q and Q' over Σ, there is a canonical isomorphism of Hermitian line bundles with connections

$$\begin{array}{ccc} \mathcal{L}_Q & \longrightarrow & \mathcal{L}_{Q'} \\ \downarrow & & \downarrow \\ \mathcal{M}_Q & \longrightarrow & \mathcal{M}_{Q'} \end{array} \tag{IV.13}$$

Since $\mathcal{M}_\Sigma \cong \mathcal{M}_Q$ for any trivializable $\mathbb{T}$ bundle $Q \to \Sigma$, the above shows that we get a Hermitian line bundle,

$$\mathcal{L}_\Sigma \to \mathcal{M}_\Sigma,$$

with a unitary connection ∇ with curvature $-2\pi i k \omega_\Sigma$, where ω_Σ is the standard symplectic form on $\mathcal{M}_\Sigma$.

B. The Chern–Simons section

We now turn our attention to the three dimensional Chern–Simons theory. Let X be a compact oriented 3-manifold with nonempty boundary and $P \to X$ a principal $\mathbb{T}$ bundle with $c_1(P)$ a torsion class. If Θ is a connection on P, then Proposition III.12 shows that the function

$$e^{\pi ikS_{X,P}(\Theta)}:\Gamma(\partial X;Q)\to\mathbb{C},$$

$$s\mapsto e^{\pi ikS_{X,P}(\Theta)}(s)=e^{\pi ikS_{X,P}(s,\Theta)},$$

is an element of norm 1 in the line $L_{\partial\Theta}$ attached to the restriction of Θ to ∂X,

$$e^{\pi ikS_{X,P}(\Theta)}\in L_{\partial\Theta}. \tag{IV.14}$$

If $\partial X=\varnothing$, then $e^{\pi ikS_{X,P}(\Theta)}\in\mathbb{C}$, so we set $L_{\varnothing}=\mathbb{C}$ to account for this case too.

Theorem IV.15: *Let X be a compact oriented 3-manifold. The assignment,*

$$\Theta\to e^{\pi ikS_{X,P}(\Theta)}\in L_{\partial\Theta},$$

for Θ connections on $\mathbb{T}$-bundles $P\to X$ with $c_1(P)$ torsion, is smooth and satisfies the following.

(a) Functoriality

Let $\phi:P'\to P$ be a morphism of principal $\mathbb{T}$-bundles covering an orientation preserving diffeomorphism $\phi:X'\to X$. If Θ is a connection on P, then

$$(\partial\phi)^*e^{\pi ikS_{X,P}(\Theta)}=e^{\pi ikS_{X',P'}(\phi^*\Theta)},$$

under the induced isometry $(\partial\phi)^*:L_{\partial\Theta}\to L_{\partial\phi^*\Theta}$. In particular, if $\partial X=\varnothing$, then

$$e^{\pi ikS_{X,P}(\Theta)}=e^{\pi ikS_{X',P'}(\phi^*\Theta)}.$$

(b) Orientation

If $-X$ is the manifold X with the opposite orientation, then

$$e^{\pi ikS_{-X,P}(\Theta)}=\overline{e^{\pi ikS_{X,P}(\Theta)}}\in\overline{L_{\partial\Theta}}.$$

(c) Disjoint union

Let X be the disjoint union $X=X_1\sqcup X_2$. If Θ_i are connections on $\mathbb{T}$ bundles $P_i\to X_i$, then we have the identification

$$e^{\pi ikS_{X_1\sqcup X_2,P_1\sqcup P_2}(\Theta_1\sqcup\Theta_2)}=e^{\pi ikS_{X_1,P_1}(\Theta_1)}\otimes e^{\pi ikS_{X_2,P_2}(\Theta_2)},$$

under the isomorphism $L_{\partial\Theta_1\sqcup\partial\Theta_2}\cong L_{\partial\Theta_1}\otimes L_{\partial\Theta_2}$.

(d) Gluing

Let X be a compact oriented 3-manifold and let X^{cut} be the 3-manifold obtained by cutting X along a closed oriented two-dimensional submanifold Σ. Then $\partial X^{\mathrm{cut}}=\partial X\sqcup(-\Sigma)\sqcup\Sigma$.

Let Θ be a connection on a $\mathbb{T}$ bundle $P\to X$, with η the restriction of Θ to Σ, and let Θ^{cut} denote the induced connection on the pullback bundle $P^{\mathrm{cut}}=g^*P$ under the gluing map $g:X^{\mathrm{cut}}\to X$. If

$$\mathrm{Tr}_\eta:L_{\partial\Theta^{\mathrm{cut}}}\cong L_{\partial\Theta}\otimes\overline{L_\eta}\otimes L_\eta\to L_{\partial\Theta}$$

is the contraction map using the Hermitian inner product in $L\eta$, then

$$\mathrm{Tr}_\eta(e^{\pi ikS_{X^{\mathrm{cut}},P^{\mathrm{cut}}}(\Theta^{\mathrm{cut}})})=e^{\pi ikS_{X,P}(\Theta)}. \tag{IV.16}$$

Proof: (a) follows from (a) of Theorem IV.4 and (a) of Theorem III.13.
(b) follows from (b) of Theorem IV.4 and (b) of Theorem III.13.
(c) follows from (c) of Theorem IV.4 and (c) of Theorem III.13.

(d) Let $g:P^{\rm cut}\to P$ be the bundle map from the pullback bundle $P^{\rm cut}=g^*P$ to P, covering the gluing map $g:X^{\rm cut}\to X$. Then $\Theta^{\rm cut}=g^*\Theta$ and $\partial\Theta^{\rm cut}=\partial\Theta\sqcup\eta\sqcup\eta$. Under the isometry $L_{\partial\Theta^{\rm cut}}\to L_{\partial\Theta}\otimes\overline{L_\eta}\otimes L_\eta$, the element $e^{\pi ikS_{X^{\rm cut},P^{\rm cut}}(\Theta^{\rm cut})}\in L_{\partial\Theta^{\rm cut}}$ is mapped to

$$\frac{e^{\pi ikS_{X^{\rm cut},P^{\rm cut}}(s^{\rm cut}\Theta^{\rm cut})}}{e^{\pi ikS_{X,P}(s\Theta)}\overline{f(\sigma)}f(\sigma)}\,(e^{\pi ikS_{X,P}(\Theta)}\otimes\overline{f}\otimes f\,)\in L_{\partial\Theta}\otimes\overline{L_\eta}\otimes L_\eta,\tag{IV.17}$$

where s, σ, and $s^{\rm cut}$ are sections $s:\partial X\to\Pi$, $\sigma:\Sigma\to P|_\Sigma$ and $s^{\rm cut}=s\sqcup\sigma\sqcup\sigma:\partial X^{\rm cut}\to\partial P^{\rm cut}$. Let $\rho_P:P\to\hat P=P\times_{\mathbb{T}}{\rm SU}(2)$ and $\rho_{P^{\rm cut}}:P^{\rm cut}\to\hat P^{\rm cut}=P^{\rm cut}\times_{\mathbb{T}}{\rm SU}(2)=g^*\hat P$ be the bundle maps to the induced SU(2) bundles. There is a SU(2)-bundle morphism $\hat g:\hat P^{\rm cut}\to\hat P$ so that $\rho_P\circ g=\hat g\circ\rho_{P^{\rm cut}}$. Let $\hat s:X\to\hat P$ be the extension of $\rho_P\circ s:\partial X\to\partial\hat P$ and $\hat s^{\rm cut}:X\to\hat P^{\rm cut}$ the pullback of the section $\hat s$. Then $\hat g\circ\hat s^{\rm cut}=\hat s\circ g$ and $\hat s^{\rm cut}|_{\partial X^{\rm cut}}=\rho_{P^{\rm cut}}\circ s^{\rm cut}$. If $\hat\Theta^{\rm cut}=\hat g^*\hat\Theta$ then $\rho^*_{P^{\rm cut}}\hat\Theta^{\rm cut}=\rho_*\Theta^{\rm cut}$. Therefore

$$\begin{aligned}S_{X^{\rm cut},P^{\rm cut}}(S^{\rm cut},\Theta^{\rm cut})&=\int_{X^{\rm cut}}\hat s^{\rm cut*}\alpha(\hat\Theta^{\rm cut})\quad(\text{mod }1)\\&=\int_{X^{\rm cut}}\hat s^{\rm cut*}\alpha(\hat g^*\hat\Theta)\quad(\text{mod }1)\\&=\int_{X^{\rm cut}}g^*\hat s^*\alpha(\hat\Theta)\quad(\text{mod }1)\\&=\int_X\hat s^*\alpha(\hat\Theta)\quad(\text{mod }1)=S_{X,P}(s,\Theta).\end{aligned}$$

Taking in (IV.17) the Hermitian inner product in L_η and using the above result, we get (IV.16).□

Consider now the restriction map $r:\mathscr{A}_P\to\mathscr{A}_{\partial P}$, which sends a connection on the $\mathbb{T}$ bundle $P\to X$ to its restriction over ∂X. As previously shown, to the trivializable $\mathbb{T}$ bundle $\partial P\to\partial X$ there corresponds a Hermitian line bundle $\pi:\hat{\mathscr{L}}_{\partial P}\to\mathscr{A}_{\partial P}$ with connection $\hat\nabla$ and these pull back under r to a Hermitian line bundle $r^*\hat{\mathscr{L}}_{\partial P}\to\mathscr{A}_P$ with connection $r^*\hat\nabla$. The restriction of $r^*\hat{\mathscr{L}}_{\partial P}$ to the subspace $\mathscr{A}_P^f$ of flat connections is a *flat* line bundle, since

$$(r^*\omega_\Sigma)(\dot\Theta,\dot\Theta')=\omega_\Sigma(r_*\dot\Theta,r_*\dot\Theta')=-\int_{\partial X}\langle\partial\dot\Theta\wedge\partial\dot\Theta'\rangle=-\int_X d\langle\dot\Theta\wedge\dot\Theta'\rangle=0,$$

for any $\dot\Theta,\dot\Theta'\in T\mathscr{A}_P^f$.

As $\mathscr{G}_P\subset\mathscr{G}_{\partial P}$, the action of the group of gauge transformations $\mathscr{G}_P$ on $\mathscr{A}_P$ lifts to $r^*\hat{\mathscr{L}}_{\partial P}$. Since $r^*\hat{\mathscr{L}}_{\partial P}=\{(\Theta,f\,)\in\mathscr{A}_P\times\hat{\mathscr{L}}_{\partial P}|r(\Theta)=\pi(f\,)\}$, we see from Theorem IV.14 and Theorem IV.15 (a) that

$$\hat\sigma_P:\mathscr{A}_P\to r^*\hat{\mathscr{L}}_{\partial P},$$

$$\Theta\mapsto(\Theta,e^{\pi ikS_{X,P}(\Theta)}),\tag{IV.18}$$

is a nowhere-zero $\mathscr{G}_P$-invariant section. Moreover, we have the following.

Proposition IV.19: The section $\hat\sigma_P:\mathscr{A}_P^f\to r^*\hat{\mathscr{L}}_{\partial P}$ is covariantly constant.

Proof: Let $s:\partial X\to\partial P$ be a section. The induced section $\gamma_s:\mathscr{A}_{\partial P}\to\hat{\mathscr{L}}_{\partial P}$ pulls back to $(r^*\gamma_s)(\Theta)=(\Theta,\gamma_s(\partial\Theta)):\mathscr{A}_P\to r^*\hat{\mathscr{L}}_{\partial P}$. Since $\mathrm{e}^{\pi ikS_{X,P}(\Theta)}=\mathrm{e}^{\pi ikS_{X,P}(s,\Theta)}\gamma_s(\partial\Theta)$, we can write $\hat\sigma_P=\varphi(r^*\gamma_s)$ with $\varphi(\Theta)=\mathrm{e}^{\pi ikS_{X,P}(s,\Theta)}$. Differentiating φ with respect to $\dot\eta\in T_\Theta\mathscr{A}_P^f$, we find

$$(\dot{\eta}\varphi)(\Theta)=\frac{d}{dt}\Big|_{t=0} e^{\pi i k S_{X,P}(s,\Theta+t\dot{\eta})}=\left[-\pi i k\int_{\partial X} s^*\langle\Theta\wedge\dot{\eta}\rangle\right]\phi(\Theta).$$

On the other hand,

$$(r^*\hat{\nabla})_{\dot{\eta}}(r^*\gamma_s)=-2\pi i k(\dot{\eta}\lrcorner r^*\hat{\theta}_s)(r^*\gamma_s).$$

Combining these two results, we obtain

$$(r^*\hat{\nabla})_{\dot{\eta}}\hat{\sigma}_P=(r^*\hat{\nabla})_{\dot{\eta}}(\varphi(r^*\gamma_s))=0.$$

□

The flat line bundle $r^*\hat{\mathscr{L}}_{\partial P}\to\mathscr{A}_P^f$ together with its metric and connection push down to the quotient by $\mathscr{G}_P$ to a flat Hermitian line bundle which is identified with the pullback $r^*\mathscr{L}_{\partial P}\to\mathscr{M}_P$ of the line bundle $\mathscr{L}_{\partial P}\to\mathscr{M}_{\partial P}$ under the restriction map $r:\mathscr{M}_P\to\mathscr{M}_{\partial P}$. The $\mathscr{G}_P$-invariant and covariantly constant section $\hat{\sigma}_P:\mathscr{A}_P^f\to r^*\hat{\mathscr{L}}_{\partial P}$ corresponds to a covariantly constant section,

$$\sigma_P:\mathscr{M}_P\to r^*\mathscr{L}_{\partial P}. \tag{IV.20}$$

For any two principal $\mathbb{T}$-bundles P and P' over X with $c_1(P)=c_1(P')=p\in\mathrm{Tors}\,H^2(X;\mathbb{Z})$, let $\phi:P'\to P$ be a bundle isomorphism. This induces an isomorphism of Hermitian line bundles with connections $\phi^*:r^*\hat{\mathscr{L}}_{\partial P}\to r^*\hat{\mathscr{L}}_{\partial P'}$, such that for the corresponding Chern–Simons sections (IV.18) we have $\phi^*\hat{\sigma}_P(\Theta)=\hat{\sigma}_{P'}(\phi^*\Theta)$, for any $\Theta\in\mathscr{A}_P$. Restricting these line bundles to the subspaces of flat connections and taking the quotient by the group of gauge transformations, we obtain a canonical isomorphism of flat Hermitian lines,

$$\begin{array}{ccc} r^*\mathcal{L}_{\partial P} & \longrightarrow & r^*\mathcal{L}_{\partial P'} \\ \downarrow & & \downarrow \\ \mathcal{M}_P & \longrightarrow & \mathcal{M}_{P'} \end{array} \tag{IV.21}$$

Thus, for each $p\in\mathrm{Tors}\,H^2(X;\mathbb{Z})$, we obtain a Hermitian line bundle with a flat connection,

$$\mathscr{L}_p\to\mathscr{M}_{X,p},$$

over the connected component $\mathscr{M}_{X,p}$ of the moduli space $\mathscr{M}_X$. The line bundle $\mathscr{L}_p\to\mathscr{M}_{X,p}$ is canonically identified with the restriction to $\mathscr{M}_{X,p}\subset\mathscr{M}_X$ of the pullback of $\mathscr{L}_{\partial X}\to\mathscr{M}_{\partial X}$ under the map $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$. In view of the canonical isomorphism (IV.21), the section (IV.20) defines a nowhere-zero covariantly constant section,

$$\sigma_{X,p}:\mathscr{M}_{X,p}\to\mathscr{L}_p=r_X^*\mathscr{L}_{\partial X}|_{\mathscr{M}_{X,p}}, \tag{IV.22}$$

which we refer to as *the Chern–Simons section*.

The constructions and results of this and previous sections show that the classical Abelian Chern–Simons theory fits into the general picture described in Ref. 17, Sec. 3; that is, it has the properties axiomatized in Ref. 17, Sec. 3.3 of a classical Lagrangian field theory that formally lead, via quantization, to the axioms of a topological quantum field theory. We have in view: (i) the assignment to every two-dimensional closed oriented manifold Σ of a Hermitian line bundle $\mathscr{L}_\Sigma$ with a unitary connection ∇ over the moduli space of flat $\mathbb{T}$ connections on Σ (Sec. IV A and properties expressed in Theorem IV.4); (ii) the assignment to every three-dimensional compact oriented manifold X of the moduli space of flat $\mathbb{T}$ connections $\mathscr{M}_X$ and the map $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$

under which the line bundle $\mathscr{L}_{\partial X}$ pulls back to a line bundle $r_X^*(\mathscr{L}_{\partial X})$ with a unitary section defined by the generalized Chern–Simons section $e^{\pi i k S_{X,P}}$ (Sec. IV.B and properties expressed in Theorem IV.15).

V. THE QUANTUM THEORY

A. The Hilbert space

We consider again a closed oriented two-dimensional manifold Σ and let $g = \frac{1}{2}\dim H^1(\Sigma;\mathbb{R})$. The space of gauge equivalence classes of flat $\mathbb{T}$ connections on Σ is a symplectic $2g$-dimensional manifold $(\mathscr{M}_\Sigma, \omega_\Sigma)$. The construction in Sec. IV provides $(\mathscr{M}_\Sigma, k\omega_\Sigma)$, $k \in 2\mathbb{Z}_+$, with a Hermitian line bundle $\mathscr{L}_\Sigma$ with a unitary connection ∇ with curvature $-2\pi i k \omega_\Sigma$. According to the general geometric quantization scheme, the other data needed for the quantization of the symplectic manifold $(\mathscr{M}_\Sigma, k\omega_\Sigma)$ is a polarization.

Since $(\mathscr{M}_\Sigma, \omega_\Sigma)$ is a symplectic torus, we are going to use the results in Ref. 10 on the quantization with half-densities of symplectic tori in a real polarization. Thus, let $L \subset H^1(\Sigma;\mathbb{R})$ be a *rational* Lagrangian subspace of the symplectic vector space $(H^1(\Sigma;\mathbb{R}), \omega_\Sigma)$. The attribute rational refers to the fact that the intersection $L \cap H^1(\Sigma;\mathbb{Z})$ with the integer lattice $H^1(\Sigma;\mathbb{Z}) \subset H^1(\Sigma;\mathbb{R})$ generates L as a vector space. Under the identification of the tangent space at any point of $\mathscr{M}_\Sigma$ with $2\pi i H^1(\Sigma;\mathbb{R})$, the Lagrangian subspace L determines an invariant real polarization $\mathscr{P}_L$ of $(\mathscr{M}_\Sigma, \omega_\Sigma)$. The polarization $\mathscr{P}_L$ is the tangent bundle along the leaves of an invariant Lagrangian foliation of $\mathscr{M}_\Sigma$.

The bundle of half-densities on $\mathscr{P}_L$ is the line bundle $|\mathrm{Det}\,\mathscr{P}_L^*|^{1/2}$ over $\mathscr{M}_\Sigma$. Its restriction to any leaf Λ of $\mathscr{P}_L$ has a canonical flat connection $\nabla^{\mathscr{P}_L}$ defined as follows.[19] First, let us recall that, since the quotient map $\pi: \mathscr{M}_\Sigma \to \mathscr{M}_\Sigma/\mathscr{P}_L$ is a smooth fibration, for each point in $\mathscr{M}_\Sigma/\mathscr{P}_L$ there is a local neighborhood U such that $\mathscr{P}_L|_{\pi^{-1}(U)}$ is spanned by Hamiltonian vector fields, $[\dot{x}_1],\ldots,[\dot{x}_g]$. Then, on $\pi^{-1}(U)$, the covariant derivative $\nabla^{\mathscr{P}_L}_{[\dot{\omega}]}\mu$ of a section μ of $|\mathrm{Det}\,\mathscr{P}_L^*|^{1/2}$, along a vector field $[\dot{\omega}] \in \mathscr{P}_L|_{\pi^{-1}(U)}$, is defined by

$$(\nabla^{\mathscr{P}_L}_{[\dot{w}]}\mu)([\dot{x}_1],\ldots,[\dot{x}_g]) = [\dot{w}](\mu([\dot{x}_1],\ldots,[\dot{x}_g])).$$

Since the leaves of $\mathscr{P}_L$ are diffeomorphic to g-dimensional tori, the distribution $\mathscr{P}_L$ has a canonical density κ invariant under the Hamiltonian vector fields in $\mathscr{P}_L$ and that assigns to each integral manifold of $\mathscr{P}_L$ the volume 1. The square root of κ trivializes the line bundle $|\mathrm{Det}\,\mathscr{P}_L^*|^{1/2}$ and is a covariantly constant section of $(|\mathrm{Det}\,\mathscr{P}_L^*|^{1/2})|_\Lambda$ for each leaf Λ of $\mathscr{P}_L$.

The Hilbert space of quantization for $(\mathscr{M}_\Sigma, k\omega_\Sigma)$ will be defined in terms of sections of the line bundle $\mathscr{L}_\Sigma \otimes |\mathrm{Det}\,\mathscr{P}_L^*|^{1/2}$. The restriction of this line bundle to a leaf Λ of $\mathscr{P}_L$ has a flat connection defined by

$$\nabla_{[\dot{w}]}(\sigma \otimes \mu) = \nabla_{[\dot{w}]}\sigma \otimes \mu + \sigma \otimes \nabla^{\mathscr{P}_L}_{[\dot{w}]}\mu,$$

for any $[\dot{\omega}] \in \mathscr{P}_L|_\Lambda$ and $\sigma \otimes \mu \in \Gamma(\Lambda; \mathscr{L}_\Sigma \otimes |\mathrm{Det}\,\mathscr{P}_L^*|^{1/2})$. The union of those leaves of $\mathscr{P}_L$ for which the holonomy group of this connection is trivial defines the Bohr–Sommerfeld set $\mathscr{BS}_{\mathscr{P}_L}$ of the polarization $\mathscr{P}_L$. For each leaf Λ belonging to $\mathscr{BS}_{\mathscr{P}_L}$ we let S_Λ denote the one-dimensional vector space of covariantly constant sections of $(\mathscr{L}_\Sigma \otimes |\mathrm{Det}\,\mathscr{P}_L^*|^{1/2})|_\Lambda$. Then the Hilbert space associated to the symplectic manifold $(\mathscr{M}_\Sigma, k\omega_\Sigma)$ with real polarization $\mathscr{P}_L$ is defined to be the complex vector space,

$$\mathscr{H}(\Sigma, L) = \bigoplus_{\Lambda \subset \mathscr{BS}_{\mathscr{P}_L}} S_\Lambda,$$

with the inner product

$$\langle \sigma\otimes\mu,\sigma'\otimes\mu'\rangle=\begin{cases}0, & \text{if } \sigma\otimes\mu\in\mathscr{S}_\Lambda,\quad \sigma'\otimes\mu'\in\mathscr{S}_{\Lambda'},\quad \Lambda\neq\Lambda',\\ \displaystyle\int_\Lambda(\sigma,\sigma')\mu*\mu', & \text{if } \sigma\otimes\mu,\quad \sigma'\otimes\mu'\in\mathscr{S}_\Lambda;\end{cases}$$

(σ,σ') is the function on Λ obtained by taking the Hermitian inner product in the fiber of $\mathscr{L}_\Sigma$ and $\mu*\mu'$ the density on Λ, defined by

$$(\mu*\mu')([\dot{x}_1],\ldots,[\dot{x}_g])=\mu([\dot{x}_1],\ldots,[\dot{x}_g])\mu'([\dot{x}_1],\ldots,[\dot{x}_g]),$$

for some vector fields $[\dot{x}_1],\ldots,[\dot{x}_g]$ spanning $T\Lambda$.

Now recall that we identify $\mathscr{M}_\Sigma=2\pi iH^1(\Sigma;\mathbb{R})/2\pi iH^1(\Sigma;\mathbb{Z})$. Let $[\dot{w}_1],\ldots,[\dot{w}_g]$ be a basis for $2\pi i(L\cap H^1(\Sigma;\mathbb{Z}))$. Then, as shown in Ref. 10, Sec. 3, the Bohr–Sommerfeld leaves of $\mathscr{P}_L$ on $\mathscr{M}_\Sigma$ are determined by the requirement that their preimages in the linear symplectic space $2\pi iH^1(\Sigma;\mathbb{R})$ covering $\mathscr{M}_\Sigma$ satisfy the condition

$$e^{2\pi ik\omega_\Sigma([\dot{w}_i],[x])}=1,\quad i=1,\ldots,g. \tag{V.1}$$

For each $\mathbf{q}=(q_1,\ldots q_g)\in(\mathbb{Z}/k\mathbb{Z})^g$ the linear equations,

$$k\omega_\Sigma([\dot{w}_i],[x])=q_i\pmod k,\quad i=1,\ldots,g, \tag{V.2}$$

for $[x]\in 2\pi iH^1(\Sigma;\mathbb{R})$, define a family of parallel Lagrangian planes projecting onto a Bohr–Sommerfeld leaf Λ_q under the quotient map $2\pi iH^1(\Sigma;\mathbb{R})\to\mathscr{M}_\Sigma$. Thus, the dimension of the Hilbert space $\mathscr{H}(\Sigma,L)$ is k^g.

In conclusion, to each closed oriented surface Σ together with a rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$ we associate by the previous construction a finite-dimensional Hilbert space $\mathscr{H}(\Sigma,L)$. We recall the following results from Ref. 10.

If L_1, $L_2\subset H^1(\Sigma;\mathbb{R})$ are two rational Lagrangian subspaces, then there is a canonical unitary operator,

$$F_{L_2L_1}:\mathscr{H}(\Sigma,L_1)\to\mathscr{H}(\Sigma,L_2), \tag{V.3}$$

induced by the Blattner–Kostant–Sternberg (BKS) pairing,

$$\langle\langle .,.\rangle\rangle:\mathscr{H}(\Sigma,L_2)\times\mathscr{H}(\Sigma,L_1)\to\mathbb{C}.$$

The BKS pairing between the Hilbert spaces $\mathscr{H}(\Sigma,L_1)=\oplus_{\Lambda_1\subset\mathscr{BS}_{\mathscr{P}_{L_1}}}S_{\Lambda_1}$ and $\mathscr{H}(\Sigma,L_2)=\oplus_{\Lambda_2\subset\mathscr{BS}_{\mathscr{P}_{L_2}}}S_{\Lambda_2}$ is defined by setting

$$\langle\langle s_2\otimes\mu_2,s_1\otimes\mu_1\rangle\rangle=\int_{\Lambda_1\cap\Lambda_2}(s_2,s_1)\mu_2*\mu_1, \tag{V.4}$$

for any Bohr–Sommerfeld leaves $\Lambda_1\subset\mathscr{BS}_{\mathscr{P}_{L_1}}$, $\Lambda_2\subset\mathscr{BS}_{\mathscr{P}_{L_2}}$ and for any sections $s_1\otimes\mu_1\in S_{\Lambda_1}$, $s_2\otimes\mu_2\in S_{\Lambda_2}$. The density $\mu_2*\mu_1$ on $\Lambda_1\cap\Lambda_2$ is defined as follows.[20,10] For any point $[x]\in\Lambda_1\cap\Lambda_2$, choose a symplectic basis $([\dot{\boldsymbol{v}}_2],[\dot{\boldsymbol{w}}];[\dot{\boldsymbol{v}}_1],[\dot{\boldsymbol{t}}])$ for $(\mathscr{M}_\Sigma,k\omega_\Sigma)$, that is a basis satisfying $k\omega_\Sigma([\dot{w}_i],[\dot{t}_j])=\delta_{ij}$, $k\omega_\Sigma([\dot{v}_{2i}],[\dot{v}_{1j}])=\delta_{ij}$ and $k\omega_\Sigma([\dot{v}_{1i}],[\dot{t}_j])=k\omega_\Sigma([\dot{v}_{2i}],[\dot{t}_j])=0$, and with $[\dot{\boldsymbol{w}}]$ a basis for $T_{[x]}(\Lambda_1\cap\Lambda_2)$, $([\dot{\boldsymbol{v}}_1],[\dot{\boldsymbol{w}}])$ a basis for $T_{[x]}\Lambda_1$ and $([\dot{\boldsymbol{v}}_2],[\dot{\boldsymbol{w}}])$ a basis for $T_{[x]}\Lambda_2$. Then $\mu_2*\mu_1$ is the density on $\Lambda_1\cap\Lambda_2$ defined by

$$(\mu_2*\mu_1)([\dot{\boldsymbol{w}}])=\mu_2([\dot{\boldsymbol{v}}_2],[\dot{\boldsymbol{w}}])\mu_1([\dot{\boldsymbol{v}}_1],[\dot{\boldsymbol{w}}]).$$

The operator $F_{L_2L_1}$ is determined by the relation

$$\langle s_2\otimes\mu_2, F_{L_2L_1}(s_1\otimes\mu_1)\rangle = \langle\langle s_2\otimes\mu_2, s_1\otimes\mu_1\rangle\rangle. \tag{V.5}$$

For the unitarity proof of $F_{L_2L_1}$ we refer to Ref. 10, Sec. IV.

For any three rational Lagrangian subspaces L_1, L_2, $L_3\subset H^1(\Sigma;\mathbb{R})$, the unitary operators relating the Hilbert spaces associated to Σ and each of these Lagrangian subspaces compose transitively up to a projective factor expressible in terms of the Maslov–Kashiwara index $\tau(L_1,L_2,L_3)$:

$$F_{L_1L_3}\circ F_{L_3L_2}\circ F_{L_2L_1} = e^{(-\pi i/4)\tau(L_1,L_2,L_3)}Id. \tag{V.6}$$

For the proof of the above relation we refer to Ref. 10, Sec. VI. The definition of the index τ may be found in Refs. 21, 22, and 10.

B. The vector

We consider a compact oriented 3-manifold X with nonempty boundary ∂X. Through the construction of Sec. IV we have a prequantum line bundle $\mathscr{L}_{\partial X}$ on the symplectic space $(\mathscr{M}_{\partial X}, k\omega_{\partial X})$. The Lagrangian map $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$ determined by the restriction of connections on X to ∂X suggests a natural choice for the polarization needed in quantizing $(\mathscr{M}_{\partial X}, k\omega_{\partial X})$. That is, we take the invariant real polarization $\mathscr{P}_X$ on $\mathscr{M}_{\partial X}$ determined by the rational Lagrangian subspace $L_X\subset H^1(\partial X;\mathbb{R})$, where $L_X=\mathrm{Im}\{\dot{r}_X:H^1(X;\mathbb{R})\to H^1(\partial X;\mathbb{R})\}$. To the pair $(\partial X, L_X)$ corresponds the Hilbert space,

$$\mathscr{H}(\partial X, L_X) = \bigoplus_{\Lambda\subset\mathscr{BS}_{\mathscr{P}_X}} \mathscr{S}_\Lambda,$$

defined as the direct sum of the one-dimensional vector spaces S_Λ of parallel sections of the line bundle $\mathscr{L}_{\partial X}\otimes|\mathrm{Det}\,\mathscr{P}_X^*|^{1/2}$, supported on the leaves Λ contained in the Bohr–Sommerfeld set $\mathscr{BS}_{\mathscr{P}_X}$ of the polarization $\mathscr{P}_X$.

Our goal is to give a canonical construction of a vector Z_X in $\mathscr{H}(\partial X, L_X)$ to be associated to the 3-manifold X. First we note the following.

Proposition V.7: The Lagrangian submanifold $\Lambda_X=\mathrm{Im}\{r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}\}$ of $\mathscr{M}_{\partial X}$ is contained in the Bohr-Sommerfeld set $\mathscr{BS}_{\mathscr{P}_X}$ and is connected.

Proof: For any flat connection Θ on a $\mathbb{T}$ bundle $P\to X$ and any closed 1-form $\dot{\alpha}$ on X for which the cohomology class $[\alpha]/2\pi i$ in $H^1(X;\mathbb{R})$ is an integral, we have

$$\int_{\partial X}\langle s^*\Theta\wedge\dot{\alpha}\rangle = 0 \pmod 1,$$

for any section $s:\partial X\to\partial P$. This follows from (III.16), since $F_\Theta=0$ and since any integral class $[\dot{\alpha}]$ in $2\pi iH^1(X;\mathbb{Z})$ is the cohomology class $[u^*\vartheta]$ of the pullback of the Maurer–Cartan form ϑ of the group $\mathbb{T}$ through some map $u:X\to\mathbb{T}$. According to the definition (II.4) of the symplectic form $\omega_{\partial X}$ on $\mathscr{M}_{\partial X}=2\pi iH^1(\partial X;\mathbb{R})/2\pi iH^1(\partial X;\mathbb{Z})$, we can rewrite the above equation as

$$\omega_{\partial X}(\dot{r}_X[\dot{\alpha}],[s^*\Theta]) = 0 \pmod 1, \tag{V.8}$$

with $[s^*\Theta]\in 2\pi iH^1(\partial X;\mathbb{R})$ the cohomology class of the 1-form $s^*\Theta$ on ∂X and $\dot{r}_X[\dot{\alpha}]$ the image of $[\dot{\alpha}]$ under the restriction map $\dot{r}_X:H^1(X;\mathbb{R})\to H^1(\partial X;\mathbb{R})$.

Let $g=1/2\dim H^1(\partial X;\mathbb{R})$ and choose a basis $[\dot{w}_1],\ldots,[\dot{w}_g]$ for the subspace $2\pi i(L_X\cap H^1(\partial X;\mathbb{Z}))$ of $2\pi iH^1(\partial X;\mathbb{R})$, with $[\dot{w}_i]=\dot{r}_X[\dot{\alpha}_i]$ for some integral classes $[\dot{\alpha}_i]$

$\in 2\pi i H^1(X;\mathbb{Z})$. Given any $[\eta]\in\Lambda_X$, there exists $[\Theta]\in\mathscr{M}_X$ with Θ a flat connection on a $\mathbb{T}$ bundle $P\to X$ such that its restriction to ∂X is $\partial\Theta=\eta$. For any section $s:\partial X\to\partial P$, the element $[s^*\eta]\in 2\pi i H^1(\partial X;\mathbb{R})$ projects to $[\eta]$ under the quotient map $2\pi i H^1(\partial X;\mathbb{R})\to\mathscr{M}_{\partial X}=2\pi i H^1(\partial X;\mathbb{R})/2\pi i H^1(\partial X;\mathbb{Z})$. Then it follows from (V.8) that

$$\omega_{\partial X}([\dot{w}_i],[s^*\eta])=\omega_{\partial X}(\dot{r}_X[\dot{\alpha}_i],[s^*\Theta])=0 \pmod 1, \quad i=1,\ldots,g.$$

Now, since Λ_X is a Lagrangian submanifold of $(\mathscr{M}_{\partial X},\omega_{\partial X})$ and since $T\Lambda_X=\mathscr{P}_X|_{\Lambda_X}$, each connected component of Λ_X is a leaf of $\mathscr{P}_X$. A comparison of the above equation to the Bohr-Sommerfeld condition (V.2) shows that the leaves of the polarization $\mathscr{P}_X$ contained in Λ_X are all Bohr–Sommerfeld leaves. Moreover, it shows that Λ_X contains only one Bohr–Sommerfeld leaf, that is, Λ_X is connected. □

The above proposition suggests that we might try to define the vector Z_X in $\mathscr{H}(\partial X,L_X)$ as a parallel section of the line bundle $\mathscr{L}_{\partial X}\otimes|\mathrm{Det}\mathscr{P}_X^*|^{1/2}$ restricted to the Bohr–Sommerfeld leaf Λ_X. First we note that, since Λ_X is a connected Lagrangian submanifold of $\mathscr{M}_{\partial X}$, each connected component $\mathscr{M}_{X,p}$ of the moduli space $\mathscr{M}_X=\sqcup_{p\in \mathrm{Tors}H^2(X;\mathbb{Z})}\mathscr{M}_{X,p}$ maps *onto* Λ_X under the continuous map $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$. Now recall from Sec. IV that, for each component $\mathscr{M}_{X,p}$, the restriction $\mathscr{L}_p=r_X^*\mathscr{L}_{\partial X}|_{\mathscr{M}_{X,p}}$ of the pullback under r_X of the prequantum line bundle $\mathscr{L}_{\partial X}\to\mathscr{M}_{\partial X}$ has a nowhere-zero parallel section, the Chern–Simons section $\sigma_{X,p}$. Then, since Λ_X is a Bohr–Sommerfeld leaf and since the map $r_X:\mathscr{M}_{X,p}\to\Lambda_X$ is surjective, it follows that the section $\sigma_{X,p}$ is the pullback of a nowhere-zero covariantly constant section of $\mathscr{L}_{\partial X}|_{\Lambda_X}$, which we continue to denote by $\sigma_{X,p}$. We are thus led to define a section σ_X of $\mathscr{L}_{\partial X}|_{\Lambda_X}$ by

$$\sigma_X=\sum_{p\in \mathrm{Tors}H^2(X;\mathbb{Z})}\sigma_{X,p}. \tag{V.9}$$

σ_X is covariantly constant with respect to the connection $\boldsymbol{\nabla}$ in $\mathscr{L}_{\partial X}$.

What we need now is a covariantly constant global section μ_X of the bundle of half-densities $|\mathrm{Det}\mathscr{P}_X^*|^{1/2}$ over Λ_X. We construct such a section using the Reidemeister torsion (R-torsion) invariant T_X of the 3-manifold X. The R-torsion T_X of a compact manifold X is defined as a norm on the determinant line $|\mathrm{Det}H^\cdot(X;\mathbb{R})|$ of the cohomology of X. The definition and properties of the R-torsion may be found in Refs. 23–26.

For a compact oriented 3-manifold X with nonempty boundary the R-torsion T_X belongs to $|\mathrm{Det}H^0(X;\mathbb{R})|\otimes|\mathrm{Det}H^1(X;\mathbb{R})^*|\otimes|\mathrm{Det}H^2(X;\mathbb{R})|\otimes|\mathrm{Det}H^3(X;\mathbb{R})^*|$. By Poincaré duality $H^2(X;\mathbb{R})\cong H^1(X,\partial X,\mathbb{R})^*$. Choosing canonical trivializations $|\mathrm{Det}H^0(X;\mathbb{R})|\cong\mathbb{R}_+$ and $|\mathrm{Det}H^3(X;\mathbb{R})|\cong\mathbb{R}_+$ with respect to orthonormal bases[24,25] for $H^0(X;\mathbb{R})$ and $H^3(X;\mathbb{R})$, we can then regard T_X as a density,

$$T_X\in|\mathrm{Det}H^1(X;\mathbb{R})^*|\otimes|\mathrm{Det}H^1(X,\partial X;\mathbb{R})^*|. \tag{V.10}$$

We recall from Sec. II that $\mathscr{M}_X=H^1(X;\mathbb{T})$ and that we have an exact sequence (II.6) of compact Abelian Lie groups. The corresponding sequence between the tangent spaces at the identity of the groups in (II.6),

$$0\to H^0(X,\partial X;\mathbb{R})\to H^0(X;\mathbb{R})\to H^0(\partial X;\mathbb{R})\to H^1(X,\partial X;\mathbb{R})\to H^1(X;\mathbb{R})\to T_e\Lambda_X\to 0,$$

induces the following isomorphism between the half-densities spaces:

$$|\mathrm{Det}T_e^*\Lambda_X|^{1/2}\cong|\mathrm{Det}H^1(X;\mathbb{R})^*|^{1/2}\otimes|\mathrm{Det}H^1(X,\partial X;\mathbb{R})|^{1/2}, \tag{V.11}$$

where we used again canonical trivializations for the spaces $H^0(*;\mathbb{R})$. Let us pick an arbitrary element $w\in|\mathrm{Det}H^1(X,\partial X;\mathbb{R})^*|$. Since $H^1(X,\partial X;\mathbb{R})$ is identified with the tangent space at the

identity of the group $H^1(X,\partial X;\mathbb{T})$, w extends to an invariant density $\mathbf{w}$ on $H^1(X,\partial X;\mathbb{T})$. If $w^{-1}\in|\mathrm{Det}H^1(X,\partial X;\mathbb{R})|$ stands for the dual of w, we note that $(T_X)^{1/2}\otimes w^{-1}\in|\mathrm{Det}H^1(X;\mathbb{R})^*|^{1/2}\otimes|\mathrm{Det}H^1(X,\partial X;\mathbb{R})|^{1/2}$. Thus we can define an invariant half-density μ_X on the Abelian group Λ_X, whose value at the identity is

$$(\mu_X)_e=\left[\int_{H^1(X,\partial X;\mathbb{T})}\mathbf{w}\right](T_X)^{1/2}\otimes w^{-1}. \tag{V.12}$$

The identification between the lhs and rhs of the above equation is made via the isomorphism (V.11). As $T\Lambda_X=\mathscr{P}_X|_{\Lambda_X}$, the half-density μ_X, being Λ_X-invariant defines a covariantly constant section of the half-densities bundle $(|\mathrm{Det}\mathscr{P}H_X^*|^{1/2})|_{\Lambda_X}$. We introduce the notation

$$\mu_X=\int_{H^1(X,\partial X;\mathbb{T})}(T_X)^{1/2}, \tag{V.13}$$

with the understanding of μ_X as the invariant extension of (V.12).

Using (V.9) and (V.13), we associate then to the compact oriented 3-manifold X with boundary ∂X the vector Z_X in the Hilbert space $\mathscr{H}(\partial X,L_X)$, defined by the expression

$$Z_X=\frac{k^{m_X}}{[\#\mathrm{Tors}H^2(X;\mathbb{Z})]}\,\sigma_X\otimes\mu_X=\frac{k^{m_X}}{[\#\mathrm{Tors}H^2(X;\mathbb{Z})]}\sum_{p\in\mathrm{Tors}H^2(X;\mathbb{Z})}\sigma_{X,p}\otimes\int_{H^1(X,\partial X;\mathbb{T})}(T_X)^{1/2}, \tag{V.14}$$

where

$$m_X=\tfrac{1}{4}\,(\dim H^1(X;\mathbb{R})+\dim H^1(X,\partial X;\mathbb{R})-\dim H^0(X;\mathbb{R})-\dim H^0(X,\partial X;\mathbb{R})). \tag{V.15}$$

If X is a closed oriented 3-manifold then, according to (V.10), the square root of the R-torsion T_X is a density $(T_X)^{1/2}\in|\mathrm{Det}H^1(X;\mathbb{R})^*|$. Since $T\mathscr{M}_X\cong H^1(X;\mathbb{R})$, this defines an invariant density $(T_X)^{1/2}$ on the moduli space $\mathscr{M}_X$. From (a) of Theorem IV.15 it follows that we have a well-defined function σ_X on $\mathscr{M}_X$ whose restriction to each connected component $\mathscr{M}_{X,p}$ of $\mathscr{M}_X$ is given by $\sigma_{X,p}([\Theta])=\mathrm{e}^{\pi ikS_{X,P}(\Theta)}$, for any $\mathbb{T}$ bundle P with $c_1(P)=p$. As a result of (III.6), σ_X is constant on components. Thus, to the closed oriented 3-manifold X we associate the complex number

$$\begin{aligned}Z_X&=\frac{k^{m_X}}{[\#\mathrm{Tors}\;H^2(X;\mathbb{Z})]}\sum_{p\in\mathrm{Tors}\;H^2(X;\mathbb{Z})}\sigma_{X,p}\int_{\mathscr{M}_X}(T_X)^{1/2}\\&=k^{m_X}\sum_{p\in\mathrm{Tors}\;H^2(X;\mathbb{Z})}\int_{\mathscr{M}_{X,p}}\sigma_{X,p}(T_X)^{1/2}=k^{m_X}\int_{\mathscr{M}_X}\sigma_X(T_X)^{1/2},\end{aligned} \tag{V.16}$$

where

$$m_X=\tfrac{1}{2}\,(\dim\,H^1(X;\mathbb{R})-\dim\,H^0(X;\mathbb{R})). \tag{V.17}$$

VI. THE TOPOLOGICAL QUANTUM FIELD THEORY

The results of the previous section show that we are very close to defining a $2+1$-dimensional topological quantum field theory (TQFT) for the Lie group $\mathbb{T}$ and an even integer k. At this stage we are able to canonically associate to a closed oriented 2-manifold Σ together with a rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$ a finite-dimensional Hilbert space $\mathscr{H}(\Sigma,L)$ and to a compact oriented 3-manifold X a vector Z_X in the Hilbert space $\mathscr{H}(\partial X,L_X)$ associated to ∂X and the Lagrangian subspace $L_X\subset H^1(\partial X;\mathbb{R})$ determined by X.

A TQFT is supposed to satisfy the axioms listed in Ref. 3 that refer to functoriality, the orientation and disjoint union properties, and the gluing of manifolds along their common boundaries. In order to construct the Chern–Simons TQFT for the group $\mathbb{T}$ we are going to use the notions introduced in Ref. 4 of extended 2- and 3-manifolds, extended morphisms, and gluing of extended 3-manifolds. We give below the relevant definitions. As in Ref. 4 we use the shortened notation with the prefix ''e'' standing for ''extended.''

Definition VI.1: An e-2-manifold is a pair (Σ,L), with a Σ closed oriented two-dimensional manifold and L a rational Lagrangian subspace of $H^1(\Sigma;\mathbb{R})$.

Definition VI.2: (i) *An e-3-manifold* is a triple $(X,\ L,\ n)$, with X a compact oriented 3-manifold, L a rational Langragian subspace of $H^1(\partial X;\mathbb{R})$, and $n\in\mathbb{Z}/8\mathbb{Z}$.

(ii) The boundary of an e-3-manifold is $\partial(X,L,n)=(\partial X,L)$. A closed e-3-manifold is just a pair $(X,\ n)$ with X a closed oriented 3-manifold and $n\in\mathbb{Z}/8\mathbb{Z}$.

Definition VI.3: An extended morphism from e-2-manifold (Σ',L') to an e-2-manifold (Σ,L) is a pair (h,m) with $h:\Sigma'\to\Sigma$, an orientation preserving diffeomorphism, and $m\in\mathbb{Z}/8\mathbb{Z}$. We refer to extended morphisms between e-2-manifolds as *e-2-morphisms*.

Definition VI.4: Composition of e-2-morphism.

If $(h,m):(\Sigma',L')\to(\Sigma,L)$ and $(h',m'):(\Sigma'',L'')\to(\Sigma',L')$ are e-2-morphisms, their composition is defined to be

$$(h,m)(h',m')=(hh',m+m'+\tau(L'',h'^{*}L',(hh')^{*}L)\pmod 8),$$

where τ is the Maslov–Kashiwara index[21,22] of a triple of Lagrangian subspaces of a symplectic vector space. The e-2-morphism (id,0) acts as the identity and the inverse of an e-2-morphism (h,m) is $(h,m)^{-1}=(h^{-1},-m)$.

Definition VI.5: An extended morphism from an e-3-manifold $(X,\ L,\ n)$ to an e-3-manifold (X',L',n') is a pair (Φ,m) with $\Phi:X'\to X$ an orientation preserving diffeomorphism and $m\in\mathbb{Z}/8\mathbb{Z}$, such that

$$n'=n+m+\tau(L_{X'},L',(\partial\Phi)^{*}L)\pmod 8,$$

where $\partial\Phi:\partial X'\to\partial X$ is the induced diffeomorphism between the boundaries. We refer to extended morphisms between e-3-manifolds as *e-3-morphisms*.

To the e-3-morphism $(\Phi,m):(X',L',n')\to(X,L,n)$ there corresponds an induced e-2-morphism between the boundaries $(\partial\Phi,m):(\partial X',L')\to(\partial X,L)$.

Two closed e-3-manifolds (X,n) and (X',n') are isomorphic if there exist an orientation preserving diffeomorphism $\Phi:X'\to X$ and $n'=n$.

Definition VI.6: *Composition of e-3-morphisms.*

If $(\Phi,m):(X',L',n')\to(X,L,n)$ and $(\Phi',m'):(X'',L'',n'')\to(X',L',n')$ are e-3-morphisms, then their composition is the e-3-morphism,

$$(\Phi,m)(\Phi',m')=(\Phi\Phi',m+m'+\tau(L'',(\partial\Phi')^{*}L',(\partial\Phi\partial\Phi')^{*}L)\pmod 8).$$

The equality

$$n''=n+m+m'+\tau(L'',(\partial\Phi')^{*}L',(\partial\Phi\partial\Phi')^{*}L)+\tau(L_{X''},L'',(\partial\Phi\partial\Phi')^{*}L)\pmod 8,$$

follows from the cocycle and the symplectic invariance properties[21,22] of the index τ.

Definition VI.7: (i) If $(\Sigma_1,\ L_1)$ and $(\Sigma_2,\ L_2)$ are e-2-manifolds, their disjoint union is the e-2-manifold,

$$(\Sigma_1,L_1)\sqcup(\Sigma_2,L_2)=(\Sigma_1\sqcup\Sigma_2,L_1\oplus L_2),$$

where we have in view the identification $H^1(\Sigma_1\sqcup\Sigma_2;\mathbb{R})=H^1(\Sigma_1;\mathbb{R})\oplus H^1(\Sigma_2;\mathbb{R})$.

(ii) If (h_1,m_1) and (h_2,m_2) are e-2-morphisms, then

$$(h_1,m_1)\sqcup(h_2,m_2)=(h_1\sqcup h_2,m_1+m_2).$$

(iii) If (X_1,L_1,n_1) and (X_2,L_2,n_2) are e-3-manifolds, their disjoint union is the e-3-manifold,

$$(X_1,L_1,n_1)\sqcup(X_2,L_2,n_2)=(X_1\sqcup X_2,L_1\oplus L_2,n_1+n_2).$$

Definition VI.8: Gluing e-3-manifolds.

Let $(\widetilde{X},\widetilde{L},\widetilde{n})$ be an e-3-manifold such that $\partial(\widetilde{X},\widetilde{L},\widetilde{n})=(\partial\widetilde{X},\widetilde{L})=(Y,L)\sqcup(-\Sigma_1,L_1)\sqcup(\Sigma_2 L_2)$ and assume there exists an e-2-morphism $(h,m):(\Sigma_1,L_1)\to(\Sigma_2,L_2)$. Under these assumptions, the gluing of $(\widetilde{X},\widetilde{L},\widetilde{n})$ by (h,m) is defined to be the e-3-manifold,

$$(\widetilde{X},\widetilde{L},\widetilde{n})_{(h,m)}=(X,L,\widetilde{n}+m+\tau(\widetilde{L},L_{\widetilde{X}},L_X\oplus L_h)\quad(\mathrm{mod}\ 8)),$$

where X is the 3-manifold obtained by gluing $\widetilde{X}$ by the diffeomorphism $h:\Sigma_1\to\Sigma_2$ and where $\widetilde{L}$, $L_{\widetilde{X}}$, and $L_X\oplus L_h$ are the following Lagrangian subspaces of $H^1(\partial\widetilde{X};\mathbb{R})=H^1(Y;\mathbb{R})\oplus H^1(-\Sigma_1;\mathbb{R})\oplus H^1(\Sigma_2;\mathbb{R})$.

- $\widetilde{L}=L\oplus L_1\oplus L_2$ according to the initial assumption on $(\widetilde{X},\widetilde{L},\widetilde{n})$;
- $L_{\widetilde{X}}=\mathrm{Im}\{H^1(\widetilde{X};\mathbb{R})\to H^1(\partial\widetilde{X};\mathbb{R})\}$;
- $L_X\oplus L_h$ is the direct sum of $L_X=\mathrm{Im}\{H^1(X;\mathbb{R})\to H^1(Y;\mathbb{R})\}\subset H^1(Y;\mathbb{R})$ and of the Lagrangian subspace $L_h\subset H^1(-\Sigma_1;\mathbb{R})\oplus H^1(\Sigma_2;\mathbb{R})$ defined as the graph of the symplectic isomorphism $h^*:H^1(\Sigma_2;\mathbb{R})\to H^1(\Sigma_1;\mathbb{R})$, that is, $L_h=\{(h^*[\dot\eta],[\dot\eta])|[\dot\eta]\in H^1(\Sigma_2;\mathbb{R})\}$.

Having introduced the necessary definitions, we are ready now to construct the TQFT. For each e-2-manifold (Σ,L) we have a finite-dimensional Hilbert space $\mathscr{H}(\Sigma,L)$ by the construction of Sec. V. Now let (X,L,n) be an e-3-manifold. Then to the e-2-manifold $(\partial X,L)$ obtained as the boundary of (X,L,n) there corresponds the Hilbert space $\mathscr{H}(\partial X,L)$. On the other hand, the 3-manifold X determines a Lagrangian subspace $L_X\subset H^1(\partial X;\mathbb{R})$ and therefore an e-2-manifold $(\partial X,L_X)$ with corresponding Hilbert space $\mathscr{H}(\partial X,L_X)$. Let

$$F_{LL_X}:\mathscr{H}(\partial X,L_X)\to\mathscr{H}(\partial X,L)\tag{VI.9}$$

be the unitary isomorphism induced by the BKS pairing (V.4). Then we define the vector $Z_{(X,L,n)}$ in $\mathscr{H}(\partial X,L)$ associated to the e-3-manifold (X,L,n) by

$$Z_{(X,L,n)}=\mathrm{e}^{(\pi i/4)n}F_{LL_X}(Z_X),\tag{VI.10}$$

where Z_X is the canonical vector in $\mathscr{H}(\partial X,L_X)$ defined by the expression (V.14).

The following theorem shows that we have a unitary TQFT.

Theorem VI.11: *The assignments,*

$$e\text{-}2\text{-}manifold\ \ (\Sigma,L)\mapsto Hilbert\ \ space\ \ \mathscr{H}(\Sigma,L),$$

$$e\text{-}3\text{-}manifold\ \ (X,L,n)\mapsto vector\ \ Z_{(X,L,n)}\in\mathscr{H}(\partial X,L),$$

satisfy the following.

(a) Functoriality

To each e-2-morphism (h,m): $(\Sigma',L')\to(\Sigma,L)$ *there corresponds a unitary isomorphism,*

$$U(h,m):\mathscr{H}(\Sigma,L)\to\mathscr{H}(\Sigma',L'),$$

and these compose properly.
Let $(\Phi,m):(X',L',n')\to(X,L,n)$ *be an e-3-morphism between isomorphic e-3-manifolds and* $(\partial\Phi,m):(\partial X',L')\to(\partial X,L)$ *the induced e-2-morphism between the boundaries. Then*

$$U(\partial\Phi,m)Z_{(X,L,n)}=Z_{(X',L',n')}.$$

(b) Orientation
There is a natural isomorphism of Hilbert spaces,

$$\mathscr{H}(-\Sigma,L)\cong\overline{\mathscr{H}(\Sigma,L)}$$

and

$$Z_{(-X,L,n)}=\overline{Z_{(X,L,n)}}.$$

(c) Disjoint union
If $(\Sigma,L)=(\Sigma_1,L_1)\sqcup(\Sigma_2,L_2)$ *is a disjoint union of e-2-manifolds, then there is a natural unitary isomorphism,*

$$\mathscr{H}(\Sigma_1\sqcup\Sigma_2,L_1\oplus L_2)\cong\mathscr{H}(\Sigma_1,L_1)\otimes\mathscr{H}(\Sigma_2,L_2).$$

If $(X,L,n)=(X_1,L_1,n_1)\sqcup(X_2L_2,n_2)$ *is a disjoint union of e-3-manifolds, then*

$$Z_{(X_1\sqcup X_2,L_1\oplus L_2,n_1+n_2)}=Z_{(X_1,L_1,n_1)}\otimes Z_{(X_2,L_2,n_2)}.$$

(d) Cylinder axiom
If Σ *is a closed oriented 2-manifold,* $I=[0,1]$ *the unit interval and* $L_\Sigma\subset H^1(\Sigma;\mathbb{R})$ *a rational Lagrangian subspace, then to the e-3-manifold* $(\Sigma\times I,L_\Sigma\oplus L_\Sigma,0)$ *there corresponds*

$$Z_{(\Sigma\times I,L_\Sigma\oplus L_\Sigma,0)}=\mathrm{Id}:\mathscr{H}(\Sigma,L_\Sigma)\to\mathscr{H}(\Sigma,L_\Sigma).$$

(e) Gluing
Let X *be a compact connected oriented 3-manifold with boundary* ∂X *and let* X^{cut} *denote the manifold obtained by cutting* X *along a codimension one closed oriented submanifold* Σ*. Then* $\partial X^{\mathrm{cut}}=\partial X\sqcup(-\Sigma)\sqcup\Sigma$*. Let* $Z_{(X,L,n)}$ *be the vector associated to the e-3-manifold* (X,L,n) *and* $Z_{(X^{\mathrm{cut}},L\oplus L_\Sigma\oplus L_\Sigma,n^{\mathrm{cut}})}$ *the vector for the cut e-3-manifold* $(X^{\mathrm{cut}},L\oplus L_\Sigma\oplus L_\Sigma,n^{\mathrm{cut}})$*, with* $L_\Sigma\subset H^1(\Sigma;\mathbb{R})$ *an arbitrary rational Lagrangian subspace, and*

$$n^{\mathrm{cut}}=n-\tau(L\oplus L_\Sigma\oplus L_\Sigma,L_{X^{\mathrm{cut}}},L_X\oplus L_\Delta)\pmod 8.$$

The Lagrangian subspace $L_\Delta\subset H^1(-\Sigma;\mathbb{R})\oplus H^1(\Sigma;\mathbb{R})$ *is the diagonal. Then to the gluing of e-3-manifolds,*

$$(X^{\mathrm{cut}},L\oplus L_\Sigma\oplus L_\Sigma,n^{\mathrm{cut}})_{(\mathrm{id}_\Sigma,0)}=(X,L,n),$$

there corresponds the quantum gluing property,

$$Z_{(X,L,n)}=\mathrm{Tr}_\Sigma[Z_{(X^{\mathrm{cut}},L\oplus L_\Sigma\oplus L_\Sigma,n^{\mathrm{cut}})}],$$

where the operator Tr_Σ *is the contraction,*

$$\mathrm{Tr}_\Sigma:\mathscr{H}(\partial X^{\mathrm{cut}},L\oplus L_\Sigma\oplus L_\Sigma)\cong\mathscr{H}(\partial X,L)\otimes\overline{\mathscr{H}(\Sigma,L_\Sigma)}\otimes\mathscr{H}(\Sigma,L_\Sigma)\to\mathscr{H}(\partial X,L),$$

using the Hermitian inner product in $\mathscr{H}(\Sigma,L_\Sigma)$.

Proof: (a) *Functoriality.* The orientation preserving the diffeomorphism $h:\Sigma'\to\Sigma$ induces a map in cohomology $h^*:H^1(\Sigma;*)\to H^1(\Sigma';*)$ and therefore a symplectic diffeomorphism $h^*:\mathscr{M}_\Sigma\to\mathscr{M}_{\Sigma'}$ between the moduli spaces of flat $\mathbb{T}$ connections. Since there exist lifts of h to bundle morphisms between trivializable $\mathbb{T}$ bundles over Σ' and trivializable $\mathbb{T}$ bundles over Σ, it follows from (a) of Theorem (IV.4) and from (IV.13) that there is an induced isomorphism $h^*:\mathscr{L}_\Sigma\to\mathscr{L}_{\Sigma'}$ of prequantum line bundles covering $h^*:\mathscr{M}_\Sigma\to\mathscr{M}_{\Sigma'}$. The Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$ is mapped onto the Lagrangian subspace $h^*L\subset H^1(\Sigma';\mathbb{R})$. The map of line bundles $h^*:\mathscr{L}_\Sigma\otimes|\mathrm{Det}\,\mathscr{P}_L^*|^{1/2}\to\mathscr{L}_{\Sigma'}\otimes|\mathrm{Det}\,\mathscr{P}_{h^*L}^*|^{1/2}$ induces a map of sections, which, having in view the constructions of Sec. V, gives rise to a unitary map of Hilbert spaces $h^*:\mathscr{H}(\Sigma,L)\to\mathscr{H}(\Sigma',h^*L)$. The composition of this map with the isomorphism $F_{L',h^*L}:\mathscr{H}(\Sigma',h^*L)\to\mathscr{H}(\Sigma',L')$ arising from the BKS pairing between $\mathscr{H}(\Sigma',L')$ and $\mathscr{H}(\Sigma',h^*L)$ gives the unitary operator $F_{L',h^*L}\circ h^*:\mathscr{H}(\Sigma,L)\to\mathscr{H}(\Sigma',L')$. Thus we associate to the e-2-morphism (h,m) the unitary operator

$$U(h,m)=\mathrm{e}^{(\pi i/4)m}F_{L',h^*L}\circ h^*:\mathscr{H}(\Sigma,L)\to\mathscr{H}(\Sigma',L').$$

We note that if L and K are rational Lagrangian subspaces in $H^1(\Sigma;\mathbb{R})$, then it follows from the definition of the intertwining isomorphisms (V.3) that the following diagram commutes:

$$\begin{array}{ccc}\mathcal{H}(\Sigma,L) & \xrightarrow{F_{K,L}} & \mathcal{H}(\Sigma,K)\\ {\scriptstyle h^*}\downarrow & & \downarrow{\scriptstyle h^*}\\ \mathcal{H}(\Sigma',h^*L) & \xrightarrow{F_{h^*K,h^*L}} & \mathcal{H}(\Sigma',h^*K)\end{array}\qquad\text{(VI.12)}$$

Now let $(h',m'):(\Sigma'',L'')\to(\Sigma',L')$ be another e-2-morphism. Then, using the properties (V.6), (VI.12), and the multiplication law of e-2-morphisms, we obtain

$$U(h',m')U(h,m)=U((h,m)(h',m')).$$

This proves the first statement in (a) of Theorem VI.11.

We prove now the second statement in (a) of Theorem VI.11. The orientation preserving diffeomorphism of compact oriented 3-manifolds $\Phi:X'\to X$ induces an isomorphism $\Phi^*:\mathscr{M}_X\to\mathscr{M}_{X'}$. Its restriction $(\partial\Phi)^*:\mathscr{M}_{\partial X}\to\mathscr{M}_{\partial X'}$ lifts to an isomorphism of prequantum line bundles $(\partial\Phi)^*:\mathscr{L}_{\partial X}\to\mathscr{L}_{\partial X'}$. Moreover, we have $r_{X'}\circ\Phi^*=(\partial\Phi)^*\circ r_X$, where $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$ and $r_{X'}:\mathscr{M}_{X'}\to\mathscr{M}_{\partial X'}$ are the restriction maps to the boundaries. Therefore $(\partial\Phi)^*L_X=L_{X'}$, so there is an induced bundle isomorphism $(\partial\Phi)^*:\mathscr{L}_{\partial X}\otimes|\,\mathrm{Det}\,\mathscr{P}_X^*|^{1/2}\to\mathscr{L}_{\partial X'}\otimes|\,\mathrm{Det}\,\mathscr{P}_{X'}^*|^{1/2}$. From the construction in Sec. V of the canonical vectors Z_X and $Z_{X'}$ as sections of these line bundles, it follows that

$$(\partial\Phi)^*(Z_X)=Z_{X'},$$

under the unitary map of Hilbert spaces $(\partial\Phi)^*:\mathscr{H}(\partial X,L_X)\to\mathscr{H}(\partial X',L_{X'})$. On the other hand, to the e-2-morphism $(\partial\Phi,m):(\partial X',L')\to(\partial X,L)$ there corresponds to the unitary operator,

$$U(\partial\Phi,m)=\mathrm{e}^{(\pi i/4)m}F_{L',(\partial\Phi)^*L}\circ(\partial\Phi)^*:\mathscr{H}(\partial X,L)\to\mathscr{H}(\partial X',L').$$

Then, using the definitions (V.14), (VI.5) and the properties of (VI.12), (V.6), we find

$$\begin{aligned}U(\partial\Phi,m)Z_{(X,L,n)}&=\mathrm{e}^{(\pi i/4)m}F_{L',(\partial\Phi)^*L}\circ(\partial\Phi)^*(\mathrm{e}^{(\pi i/4)n}F_{LL_X}(Z_X))\\&=\mathrm{e}^{(\pi i/4)n'}F_{L'L_{X'}}(Z_{X'})=Z_{(X',L',n')}.\end{aligned}$$

(b) *Orientation*. Follows from (b) of Theorem (IV.4) and (b) of Theorem (IV.15).

(c) *Disjoint union*. To the disjoint union $\Sigma=\Sigma_1\sqcup\Sigma_2$ there corresponds the product of symplectic spaces,

$$(\mathscr{M}_\Sigma,\omega_\Sigma)=(\mathscr{M}_{\Sigma_1},\omega_{\Sigma_1})\times(\mathrm{M}_{\Sigma_2},\omega_{\Sigma_2})=(\mathscr{M}_{\Sigma_1}\times\mathscr{M}_{\Sigma_2},pr_{\Sigma_1}^*\omega_{\Sigma_1}+pr_{\Sigma_2}^*\omega_{\Sigma_2}),$$

where $pr_{\Sigma_i}:\Sigma_1\times\Sigma_2\to\Sigma_i$ are the natural projections. If $L_i\subset H^1(\Sigma_i;\mathbb{R})$ are rational Lagrangian subspaces, then $L_1\oplus L_2\subset H^1(\Sigma_1\sqcup\Sigma_2;\mathbb{R})=H^1(\Sigma_1;\mathbb{R})\oplus H^1(\Sigma_2;\mathbb{R})$ is again a rational Lagrangian subspace. The Bohr–Sommerfeld orbits Λ of the polarization $P_{L_1\oplus L_2}$ on $\mathscr{M}_{\Sigma_1}\times\mathscr{M}_{\Sigma_2}$ are Cartesian products $\Lambda_1\times\Lambda_2$ of Bohr–Sommerfeld orbits Λ_i for the polarizations $\mathscr{P}_{L_i}$ on $\mathscr{M}_{\Sigma_i}$, $i=1,2$. The proof of (c) follows then from (c) of Theorem (IV.4) and (c) of Theorem (IV.15) and the definitions of the Hilbert spaces associated to e-2-manifolds and of the vectors associated to e-3-manifolds.

(d) *Cylinder axiom*. Since $H^i(\Sigma\times I)\cong H^i(\Sigma)$, $i=0,1,2$, we obtain, in particular, that $\mathscr{M}_{\Sigma\times I}\cong\mathscr{M}_\Sigma$. As $\partial(\Sigma\times I)=(-\Sigma)\sqcup\Sigma$ we have $\mathscr{M}_{\partial(\Sigma\times I)}=\mathscr{M}_\Sigma\times\mathscr{M}_\Sigma$, and the restriction map

$$r_{\Sigma\times I}:\mathscr{M}_{\Sigma\times I}\cong\mathscr{M}_\Sigma\to\mathscr{M}_{\partial(\Sigma\times I)}=\mathscr{M}_\Sigma\times\mathscr{M}_\Sigma$$

is the diagonal map $[\eta],\mapsto([\eta],[\eta])$. Hence the Lagrangian subspace $L_{\Sigma\times I}=L_\Delta$, with L_Δ denoting the diagonal in $H^1(-\Sigma;\mathbb{R})\oplus H^1(\Sigma;\mathbb{R})$ and the Lagrangian submanifold $\Lambda_{\Sigma\times I}=\Delta$, with Δ the diagonal in $\mathscr{M}_\Sigma\times\mathscr{M}_\Sigma$. According to the definitions and constructions of Sec. V we find that the Chern–Simons section over Δ of the prequantum line bundle is $\sigma_{\Sigma\times I}=1$ and that the section $\mu_{\Sigma\times I}$ over Δ of the half-density bundle $|\operatorname{Det}\mathscr{P}_\Delta^*|^{1/2}$ is identified under the isomorphism $\Delta\cong\mathscr{M}_\Sigma$ with the square root $(T_\Sigma)^{1/2}$ of the R-torsion of Σ. In Ref. 27 it is proved that the density T_Σ on $\mathscr{M}_\Sigma$ coincides with the density $|(\omega_\Sigma)^g/g!|$ determined by the symplectic form ω_Σ, where $g=\frac{1}{2}\dim H^1(\Sigma;\mathbb{R})$. In view of the definition (V.14), we find therefore that the canonical vector in $\mathscr{H}(-\Sigma\sqcup\Sigma,L_\Delta)$ associated to the 3-manifold $\Sigma\times I$ is

$$Z_{\Sigma\times I}=k^{(1/4)\dim H^1(\Sigma;\mathbb{R})}\sigma_{\Sigma\times I}\otimes\mu_{\Sigma\times I}=k^{(1/4)dim\ H^1(\Sigma;\mathbb{R})}1\otimes(T_\Sigma)^{1/2}. \tag{VI.13}$$

Let us consider now the e-3-manifold $(\Sigma\times I,L_{2\Sigma,n})$, where $L_{2\Sigma}=L_\Sigma\oplus L_\Sigma$ with $L_\Sigma\subset H^1(\Sigma;\mathbb{R})$ an arbitrary rational Lagrangian subspace. To this e-3-manifold there corresponds the vector in $\mathscr{H}(-\Sigma\sqcup\Sigma,L_{2\Sigma})\cong\overline{\mathscr{H}(\Sigma,L_\Sigma)}\otimes\mathscr{H}(\Sigma,L_\Sigma)$ defined by

$$Z_{(\Sigma\times I,L_{2\Sigma},n)}=\mathrm{e}^{(\pi i/4)n}F_{L_{2\Sigma},L_\delta}(Z_{\Sigma\times I}).$$

Let us introduce a unitary basis $\{v(\Sigma,L_\Sigma)_q\}$ for the Hilbert space $\mathscr{H}(\Sigma,L_\Sigma)$. Using the definition of the unitary operator $F_{L_{2\Sigma},L_\Delta}:\mathscr{H}(-\Sigma\sqcup\Sigma,L_\Delta)\to\mathscr{H}(-\Sigma\sqcup\Sigma,L_{2\Sigma})$ in terms of the BKS pairing (V.4), we find that

$$Z_{(\Sigma\times I,L_{2\Sigma},n)}=\mathrm{e}^{(\pi i/4)n}\sum_q\overline{v(\Sigma,L_\Sigma)_{\mathbf{q}}}\otimes v(\Sigma,L_\Sigma)_{\mathbf{q}}. \tag{VI.14}$$

Thus $Z_{(\Sigma\times I,L_{2\Sigma},0)}=\mathrm{Id}\in\mathrm{Hom}[\mathscr{H}(\Sigma,L_\Sigma),\mathscr{H}(\Sigma,L_\Sigma)]$.

(e) *Gluing*. The moduli space of flat $\mathbb{T}$ connections on $\partial X^{\mathrm{cut}}=\partial X\sqcup(-\Sigma)\sqcup\Sigma$ is the symplectic manifold $(\mathscr{M}_{\partial X^{\mathrm{cut}}},\omega_{\partial X^{\mathrm{cut}}})$, with

$$\mathscr{M}_{\partial X^{\mathrm{cut}}}=\mathscr{M}_{\partial X}\times\mathscr{M}_\Sigma\times\mathscr{M}_\Sigma,$$

$$\omega_{\partial X^{\mathrm{cut}}}=pr_{\partial X}^*(\omega_{\partial X})+pr_{-\Sigma}^*(-\omega_\Sigma)+pr_\Sigma^*(\omega_\Sigma),$$

where $pr_{\partial X}$, $pr_{-\Sigma}$, and pr_{Σ} denote the projection maps from $\mathscr{M}_{\partial X}\times\mathscr{M}_{\Sigma}\times\mathscr{M}_{\Sigma}$ onto the first, second, and third factor, respectively. Let $\Delta\subset\mathscr{M}_{\Sigma}\times\mathscr{M}_{\Sigma}$ denote the diagonal. The manifold

$$C=\mathscr{M}_{\partial X}\times\Delta\subset\mathscr{M}_{\partial X^{\mathrm{cut}}},$$

is a coisotropic submanifold of $(\mathscr{M}_{\partial X^{\mathrm{cut}}},\omega_{\partial X^{\mathrm{cut}}})$. Let $TC^{\perp}$ denote the orthogonal complement with respect to $\omega_{\partial X^{\mathrm{cut}}}$ of the tangent bundle TC. The bundle $TC^{\perp}$ is the tangent bundle to an isotropic foliation $C^{\perp}$ of C. The reduced symplectic manifold $C/C^{\perp}$ is identified with $(\mathscr{M}_{\partial X},\omega_{\partial X})$ and the quotient map $C\to C/C^{\perp}$ with the natural projection $pr_{\partial X}:\mathscr{M}_{\partial X}\times\Delta\to\mathscr{M}_{\partial X}$.

With the usual notations let $\Lambda_{X^{\mathrm{cut}}}$ be the image of $\mathscr{M}_{X^{\mathrm{cut}}}$ under the restriction map $r_{X^{\mathrm{cut}}}:\mathscr{M}_{X^{\mathrm{cut}}}\to\mathscr{M}_{\partial X^{\mathrm{cut}}}$. According to (II.5), $\Lambda_{X^{\mathrm{cut}}}$ is a Lagrangian submanifold of $(\mathscr{M}_{\partial X^{\mathrm{cut}}},\omega_{\partial X^{\mathrm{cut}}})$, and we consider its reduction[28] relative to the coisotropic submanifold C. The intersection $\Lambda_{X^{\mathrm{cut}}}\cap C$ is a manifold and the tangent bundles satisfy $T(\Lambda_{X^{\mathrm{cut}}}\cap C)=T\Lambda_{X^{\mathrm{cut}}}\cap TC$. The reduction of $\Lambda_{X^{\mathrm{cut}}}$ relative to C is the image of $\Lambda_{X^{\mathrm{cut}}}\cap C$ under the quotient map $C\to C/C^{\perp}$. It is a Lagrangian submanifold of $C/C^{\perp}$ that coincides with $\Lambda_X\subset\mathscr{M}_{\partial X}$ under the identification of $C/C^{\perp}$ with $\mathscr{M}_{\partial X}$. That is,

$$\Lambda_X=pr_{\partial X}(\Lambda_{X^{\mathrm{cut}}}\cap C).$$

The kernel of the differential of $p_{\partial X}:\Lambda_{X^{\mathrm{cut}}}\cap C\to\Lambda_X$ is the vector bundle $T\Lambda_{X^{\mathrm{cut}}}\cap TC^{\perp}$.

Now let $g:X^{\mathrm{cut}}\to X$ denote the gluing map and $g^*:\mathscr{M}_X\to\mathscr{M}_{X^{\mathrm{cut}}}$ the induced map between the moduli spaces of flat $\mathbb{T}$ connections. If $r_X:\mathscr{M}_X\to\mathscr{M}_{\partial X}$ and $r_{\Sigma}:\mathscr{M}_X\to\mathscr{M}_{\Sigma}$ are the restriction maps from connections on X to connections on ∂X and Σ, respectively, then we get the commutative diagram

$$\begin{array}{ccc}\mathcal{M}_X & \xrightarrow{r_X\times r_{\Sigma}} & \mathcal{M}_{\partial X}\times\mathcal{M}_{\Sigma}\\ {\scriptstyle g^*}\downarrow & & \downarrow{\scriptstyle id\times i_{\Delta}}\\ \mathcal{M}_{X^{cut}} & \xrightarrow{r_{X^{cut}}} & \mathcal{M}_{\partial X^{cut}}=\mathcal{M}_{\partial X}\times\mathcal{M}_{\Sigma}\times\mathcal{M}_{\Sigma}\end{array}\qquad\text{(VI.15)}$$

where $i_{\Delta}:\mathscr{M}_{\Sigma}\to\mathscr{M}_{\Sigma}\times\mathscr{M}_{\Sigma}$ is the diagonal map $i_{\Delta}([\eta])=([\eta],[\eta])\in\Delta$. We note that

$$\Lambda_{X^{\mathrm{cut}}}\cap C=(r_{X^{\mathrm{cut}}}\circ g^*)(\mathscr{M}_X).\qquad\text{(VI.16)}$$

Recall that $\mathscr{M}_{X^{\mathrm{cut}}}=\bigsqcup_{p^{\mathrm{cut}}\in\mathrm{Tors}\,H^2(X^{\mathrm{cut}};\mathbb{Z})}\mathscr{M}_{X^{\mathrm{cut}},p^{\mathrm{cut}}}$ and $\mathscr{M}_X=\bigsqcup_{p\in\mathrm{Tors}\,H^2(X,\mathbb{Z})}\mathscr{M}_{X,p}$. Thus, for each p^{cut} and each connected component $[\Lambda_{X^{\mathrm{cut}}}\cap C]_c$ of $\Lambda_{X^{\mathrm{cut}}}\cap C$, there is a unique p such that

$$r_{X^{\mathrm{cut}}}^{-1}([\Lambda_{X^{\mathrm{cut}}}\cap C]_c)\cap\mathscr{M}_{X^{\mathrm{cut}},p^{\mathrm{cut}}}=g^*(\mathscr{M}_{X,p}).\qquad\text{(VI.17)}$$

Then (VI.17) and (ii) in Proposition (II.2) imply the relation

$$[\#\pi_0(\Lambda_{X^{\mathrm{cut}}}\cap C)]\cdot[\#\mathrm{Tors}\,H^2(X^{\mathrm{cut}};\mathbb{Z})]=[\#\mathrm{Tors}\,H^2(X;\mathbb{Z})].\qquad\text{(VI.18)}$$

Let us introduce the following Lagrangian subspaces of the symplectic vector space $H^1(X^{\mathrm{cut}};\mathbb{R})=H^1(X;\mathbb{R})\oplus H^1(-\Sigma;\mathbb{R})\oplus H^1(\Sigma;\mathbb{R})$:

$$\begin{aligned}L'&=L_X\oplus L_{\Sigma}\oplus L_{\Sigma},\\ L''&=L\oplus L_{\Sigma}\oplus L_{\Sigma}.\end{aligned}\qquad\text{(VI.19)}$$

To the e-3-manifold $(X^{\mathrm{cut}},L'',n^{\mathrm{cut}})$ corresponds the vector

$$Z_{(X^{\rm cut},L'',n^{\rm cut})}=\mathrm{e}^{(\pi i/4)n^{\rm cut}}F_{L''L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\in\mathscr{H}(\partial X^{\rm cut},L''),\tag{VI.20}$$

obtained as the image of the canonical vector,

$$Z_{X^{\rm cut}}=\frac{k^{m_{X^{\rm cut}}}}{[\#\mathrm{Tors}\ H^2(X^{\rm cut};\mathbb{Z})]}\,\sigma_{X^{\rm cut}}\otimes\mu_{X^{\rm cut}}\in\mathscr{H}(\partial X^{\rm cut},L_{X^{\rm cut}}),\tag{VI.21}$$

defined in Sec. V. Using the composition law (V.6), we can rewrite (VI.20) as

$$Z_{(X^{\rm cut},L'',n^{\rm cut})}=\mathrm{e}^{(\pi i/4)[n^{cut}+\tau(L_{X^{cut}},L',L'')]}F_{L''L'}\circ F_{L'L_{X^{\rm cut}}}(Z_{X^{cut}}).\tag{VI.22}$$

Let us choose unitary bases $\{v(\partial X,L_X)_{\mathbf{q}}\}$ for $\mathscr{H}(\partial X,L_X)$, $\{v(\partial X,L)_{q'}\}$ for $\mathscr{H}(\partial X,L)$, and $\{v(\Sigma,L_\Sigma)_{\ell}\}$ for $\mathscr{H}(\Sigma,L_\Sigma)$. Then we have

$$\begin{aligned}F_{L''L'}\circ F_{L'L_{X^{\rm cut}}}(Z_{X^{\rm cut}})&=F_{L''L'}\left[\sum_{\mathbf{q},\ell,\ell'}M_{\mathbf{q}\ell\ell'}v(\partial X,L_X)_{\mathbf{q}}\otimes\overline{v(\Sigma,L_\Sigma)_{\ell}}\otimes v(\Sigma,L_\Sigma)_{\ell'}\right]\\&=\sum_{\mathbf{q},\ell,\ell'}M_{\mathbf{q}\ell\ell'}[F_{LL_X}v(\partial X,L_X)_{\mathbf{q}}]\otimes\overline{v(\Sigma,L_\Sigma)_{\ell}}\otimes v(\Sigma,L_\Sigma)_{\ell'},\end{aligned}\tag{VI.23}$$

where $F_{LL_X}:\mathscr{H}(\partial X,L_X)\to\mathscr{H}(\partial X,L)$ and

$$M_{\mathbf{q}\mathbf{l}\mathbf{l}'}=\langle v(\partial X,L_X)_{\mathbf{q}}\otimes\overline{v(\Sigma,L_\Sigma)_{\mathbf{l}}}\otimes v(\Sigma,L_\Sigma)_{\mathbf{l}'},F_{L'L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut},L')}.$$

The last equality in (VI.23) follows from the definition (V.5) of the intertwining isomorphism $F_{L''L'}$ through the BKS pairing (V.4), which, in view of the special form (VI.19) of the Lagrangian subspaces L' and L'', leads to the identification $F_{L''L'}=F_{LL_X}\oplus\mathrm{Id}\oplus\mathrm{Id}$. Applying the contraction operator Tr_Σ to $Z_{(X^{\rm cut},L'',n^{\rm cut})}$ and making use of (VI.23), we get

$$\mathrm{Tr}_\Sigma[Z_{(X^{\rm cut},L'',n^{\rm cut})}]=\mathrm{e}^{(\pi i/4)[n^{\rm cut}+\tau(L_{X^{\rm cut}},L',L'')]}\sum_{\mathbf{q},\ell}M_{\mathbf{q}\ell\ell}F_{LL_X}[v(\partial X,L_X)_{\mathbf{q}}].$$

Now let $L_\Delta\subset H^1(-\Sigma;\mathbb{R})\oplus H^1(\Sigma;\mathbb{R})$ be the diagonal. It is a rational Lagrangian subspace for the symplectic form $pr_{-\Sigma}^*(-\omega_\Sigma)+pr_\Sigma^*(\omega_\Sigma)$ on $H^1(-\Sigma;\mathbb{R})\oplus H^1(\Sigma;\mathbb{R})$. We introduce the notation $L_{2\Sigma}=L_\Sigma\oplus L_\Sigma$ and define $L'_\Delta=L_X\oplus L_\Delta$. Recall that to the cylinder $\Sigma\times I$ there corresponds the canonical vector $Z_{\Sigma\times I}$ in $\mathscr{H}((-\Sigma)\sqcup\Sigma,L_\Delta)$ given by (VI.13). Associated to the e-3-manifold $(\Sigma\times I,L_{2\Sigma},0)$ we have, according to (VI.14), the vector

$$Z_{(\Sigma\times I,L_{2\Sigma},0)}=F_{L_{2\Sigma},L_\Delta}(Z_{\Sigma\times I})=\sum_{\ell}\overline{v(\Sigma,L_\Sigma)_{\ell}}\otimes v(\Sigma,L_\Sigma)_{\ell},$$

in $\mathscr{H}((-\Sigma)\sqcup\Sigma,L_{2\Sigma})$. Using this, we can write the following:

$$\begin{aligned}\sum_{\ell}M_{\mathbf{q}\ell\ell}&=\sum_{\ell}\langle v(\partial X,L_X)_{\mathbf{q}}\otimes\overline{v(\Sigma,L_\Sigma)_{\ell}}\otimes v(\Sigma,L_\Sigma)_{\ell},F_{L'L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut},L')}\\&=\langle v(\partial X,L_X)_{\mathbf{q}}\otimes Z_{(\Sigma\times I,L_{2\Sigma},0)},F_{L'L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut},L')}\\&=\langle v(\partial X,L_X)_{\mathbf{q}}\otimes F_{L_{2\Sigma},L_\Delta}(Z_{\Sigma\times I}),F_{L',L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut},L')}.\end{aligned}$$

Now, since $F_{L'L_{X^{\rm cut}}}=e^{(\pi i/4)\tau(L_{X^{\rm cut}},L'_\Delta,L')}F_{L'L'_\Delta}\circ F_{L'_\Delta,L_{X^{\rm cut}}}$, and $F_{L'L'_\Delta}=Id\otimes F_{L_{2\Sigma},L_\Delta}$ we obtain

$$\begin{aligned}\sum_{\mathbf{q}'} M_{\mathbf{q}\mathbf{q}'}&=\mathrm{e}^{(\pi i/4)\tau(L_{X^{\rm cut}},L'_\Delta,L')}\times\langle F_{L'L'_\Delta}[v(\partial X,L_X)_{\mathbf{q}}\otimes Z_{\Sigma\times I}],F_{L'L'_\Delta}\circ F_{L'_\Delta,L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut},L')}\\&=\mathrm{e}^{(\pi i/4)\tau(L_{X^{\rm cut}},L'_\Delta,L')}\langle v(\partial X,L_X)_{\mathbf{q}}\otimes Z_{\Sigma\times I},F_{L'_\Delta,L_{X^{\rm cut}}}(Z_{X^{\rm cut}})\rangle_{\mathscr{H}(\partial X^{\rm cut}L'_\Delta)}\\&=\frac{k^{m_{X^{\rm cut}}+(1/4)\dim H^1(\Sigma;\mathbb{R})}}{[\#\mathrm{Tors}\ H^2(X^{\rm cut};\mathbb{Z})]}\,\mathrm{e}^{(\pi i/4)\tau(L_{X^{\rm cut}},L'_\Delta,L')}\\&\quad\times\int_{\Lambda_{X^{\rm cut}}\cap((\Lambda_X)_{\mathbf{q}}\times\Delta)}(s(\partial X)_{\mathbf{q}}\otimes\sigma_{\Sigma\times I},\sigma_{X^{\rm cut}})(\delta_X\otimes\mu_{\Sigma\times I})*\mu_{X^{\rm cut}}.\end{aligned}$$

The last equality follows from the BKS pairing formula defining the operator $F_{L'_\Delta,L_{X^{\rm cut}}}$ and the expressions (VI.21) for $Z_{X^{\rm cut}}$ and (VI.13) for $Z_{\Sigma\times I}$. We also wrote $v(\partial X,L_X)_{\mathbf{q}}=s(\partial X)_{\mathbf{q}}\otimes\delta_X$, where δ_X is the invariant half-density on $\mathscr{P}_X$ such that $\int_{(\Lambda_X)_q}\delta_X^2=1$, for any leaf $(\Lambda_X)_{\mathbf{q}}$ of the polarization $\mathscr{P}_X$ on $\mathscr{M}_{\partial X}$. As previously shown, we have the reduction map $\Lambda_{X^{\rm cut}}\cap C\to\Lambda_X$ and, since $\Lambda_X\cap(\Lambda_X)_q\neq\varnothing$ if and only if $(\Lambda_X)_{\mathbf{q}}=\Lambda_X$, we conclude that $\Lambda_{X^{\rm cut}}\cap((\Lambda_X)_{\mathbf{q}}\times\Delta)=\varnothing$, for all leaves $(\Lambda_X)_{\mathbf{q}}\neq\Lambda_X$. Let us write

$$\mu_{X^{\rm cut}}=a_{X^{\rm cut}}\delta_{X^{\rm cut}},$$

with $\delta_{X^{\rm cut}}$ the invariant half-density on $\mathscr{P}_{X^{\rm cut}}$ satisfying $\int_{\Lambda_{X^{\rm cut}}}\delta_{X^{\rm cut}}^2=1$. We recall from (VI.13) that $\mu_{\Sigma\times I}=(T_\Sigma)^{1/2}=|(\omega_\Sigma)^g/g!|^{1/2}$ and that we have $\int_\Delta(\omega_\Sigma)^g/g!=1$, where $g=(1/2)\dim H^1(\Sigma;\mathbb{R})$. Then, making use of the results in Ref. 10, Sec. IV, we find that for each connected component $[\Lambda_{X^{\rm cut}}\cap C]_c$ of the manifold $\Lambda_{X^{\rm cut}}\cap C=\Lambda_{X^{\rm cut}}\cap(\Lambda_X\times\Delta)$ we have

$$\int_{[\Lambda_{X^{\rm cut}}\cap C]_c}(\delta_X\otimes T_\Sigma^{1/2})*\delta_{X^{\rm cut}}=k^{(1/2)[\dim(\Lambda_{X^{\rm cut}}\cap C)-(1/2)\dim H^1(\partial X^{\rm cut};\mathbb{R})]}[\#\pi_0(\Lambda_{X^{\rm cut}}\cap C)]^{-1/2}.$$

Let us introduce the notation $v(\partial X,L_X)=v(\partial X,L_X)_{\mathbf{q}}$ if $(\Lambda_X)_q=\Lambda_X$. We also note that the function $(s(\partial X)\otimes\sigma_{\Sigma\times I},\sigma_{\partial X^{\rm cut}})$ on $\Lambda_{\partial X^{\rm cut}}\cap C$ is constant on each connected component. Summarizing the previous results and using the fact that $\sigma_{X^{\rm cut}}=\Sigma_{p^{\rm cut}\in\mathrm{Tors}\,H^2(X^{\rm cut};\mathbb{Z})}\sigma_{X^{\rm cut},p^{\rm cut}}$, we can write

$$\begin{aligned}\mathrm{Tr}[Z_{(X^{\rm cut},L'',n^{\rm cut})}]&=\mathrm{e}^{(\pi i/4)[n^{\rm cut}+\tau(L_{X^{\rm cut}},L',L'')+\tau(L_{X^{\rm cut}},L'_\Delta,L')]}\\&\quad\times\frac{a_{X^{\rm cut}}}{[\#\mathrm{Tors}\ H^2(X^{\rm cut};\mathbb{Z})]}\,[\#\pi_0(\Lambda_{X^{\rm cut}}\cap C)]^{-1/2}\\&\quad\times k^{m_{X^{\rm cut}}+(1/4)\dim H^1(\Sigma;\mathbb{R})+(1/2)\dim(\Lambda_{X^{\rm cut}}\cap C)-(1/4)\dim H^1(\partial X^{\rm cut};\mathbb{R})}\\&\quad\times\left[\sum_c\sum_{p^{\rm cut}\in\mathrm{Tors}\,H^2(X^{\rm cut};\mathbb{Z})}(s(\partial X)\otimes\sigma_{\Sigma\times I},\sigma_{X^{\rm cut},p^{\rm cut}})_c\right]F_{LL_X}(v(\partial X,L_X)),\end{aligned}\tag{VI.24}$$

where $(s(\partial X)\otimes\sigma_{\Sigma\times I},\sigma_{X^{\rm cut},p^{\rm cut}})_c$ denotes the value of the function on the component $[\Lambda_{X^{\rm cut}}\cap C]_c$. Then let $[\Theta]\in\mathscr{M}_{X,p}$ with $[\eta]=[\Theta|_\Sigma]$ and $[\partial\Theta]=[\Theta|_{\partial X}]$ and let $[\Theta^{\rm cut}]=[g^*\Theta]$. Using the classical gluing formula (IV.16) and the definitions of $\sigma_{\Sigma\times I}$, $\sigma_{X,p}$, and $\sigma_{X^{\rm cut},p^{\rm cut}}$, we find that

$$(\sigma_{X,p}([\partial\Theta])\otimes\sigma_{\Sigma\times I}([\eta]\sqcup[\eta]),\sigma_{X^{\rm cut},p^{\rm cut}}([\partial\Theta]\sqcup[\eta]\sqcup[\eta]))$$

$$=(e^{\pi ikS_{X,P}(\Theta)},\mathrm{Tr}_{\eta}[e^{\pi ikS_{X^{cut},P^{cut}}(\Theta^{cut})}])=1.$$

Since both $\sigma_{X,p}$ and $s(\partial X)$ are unitary covariantly constant sections of $\mathscr{L}_{\partial X}\otimes|\mathrm{Det}\,\mathscr{P}_X^*|^{1/2}$ over Λ_X, they differ by a constant phase factor. In view of this observation of the previous relation and of the relations (VI.17)–(VI.18), the expression (VI.24) becomes

$$\begin{aligned}\mathrm{Tr}_{\Sigma}[Z_{(X^{\mathrm{cut}},L'',n^{\mathrm{cut}})}]=e^{(\pi i/4)[n^{\mathrm{cut}}+\tau(L_{X^{\mathrm{cut}}},L',L'')+\tau(L_{X^{\mathrm{cut}}},L'_{\Delta},L')]}&\\ \times\frac{a_{X^{\mathrm{cut}}}}{[\#\mathrm{Tors}\ H^2(X^{\mathrm{cut}};\mathbb{Z})]^{1/2}[\#\mathrm{Tors}\ H^2(X;\mathbb{Z})]^{1/2}}&\\ \times k^{m_{X^{\mathrm{cut}}}+(1/4)\dim H^1(\Sigma;\mathbb{R})+(1/2)\dim(\Lambda_{X^{\mathrm{cut}}}\cap C)-(1/4)\dim H^1(\partial X^{\mathrm{cut}};\mathbb{R})}&\\ \times\sum_{p\in\mathrm{Tors}\ H^2(X;\mathbb{Z})}F_{LL_X}(\sigma_{X,p}\otimes\delta_X).&\end{aligned}\tag{VI.25}$$

Let a_X denote the real constant such that $\mu_X=a_X\delta_X$. We claim that the following relations hold:

$$a^2_{X^{\mathrm{cut}}}=a^2_X\,\frac{[\#\mathrm{Tors}\ H^2(X^{\mathrm{cut}};\mathbb{Z})]}{[\#\mathrm{Tors}\ H^2(X;\mathbb{Z})]}\tag{VI.26}$$

and

$$m_{X^{\mathrm{cut}}}+\tfrac{1}{4}\dim H^1(\Sigma;\mathbb{R})+\tfrac{1}{2}\dim(\Lambda_{X^{\mathrm{cut}}}\cap C)-\tfrac{1}{4}\dim H^1(\partial X^{\mathrm{cut}};\mathbb{R})=m_X.\tag{VI.27}$$

We postpone for a moment their proof and note first that by using the definition of n^{cut}, the cocycle relation[22] for the Maslov–Kashiwara index τ and the symplectic additivity property of τ, we have

$$\begin{aligned}&n^{\mathrm{cut}}+\tau(L_{X^{\mathrm{cut}}},L',L'')+\tau(L_{X^{\mathrm{cut}}},L'_{\Delta},L')\pmod 8\\ &=n+\tau(L',L'',L'_{\Delta})\pmod 8\\ &=n+\tau(L_X,L,L_X)+\tau(L_{\Sigma}\oplus L_{\Sigma},L_{\Sigma}\oplus L_{\Sigma},L_{\Delta})\pmod 8\\ &=n\pmod 8.\end{aligned}\tag{VI.28}$$

Putting the results (VI.25)–(VI.28) together, we obtain

$$\mathrm{Tr}_{\Sigma}[Z_{(X^{\mathrm{cut}},L'',n^{\mathrm{cut}})}]=e^{(\pi i/4)n}\,\frac{k^{m_X}}{[\#\mathrm{Tors}\ H^2(X;\mathbb{Z})]}\,F_{LL_X}(\sigma_X\otimes\mu_X)=e^{(\pi i/4)n}F_{LL_X}(Z_X)=Z_{(X,L,n)},$$

which proves the gluing property (e) of theorem (VI.11).

Let us return now to the proof of the equality (VI.27). We have the following relations.

(1) The Mayer–Vietoris cohomology sequence for the spaces $X,X^{\mathrm{cut}},\Sigma$,

$$\cdots\to H^i(X;\mathbb{R})\to H^i(X^{\mathrm{cut}};\mathbb{R})\to H^i(\Sigma;\mathbb{R})\to H^{i+1}(X;\mathbb{R})\to\cdots,\tag{VI.29}$$

together with Poincaré duality, imply the relation

$$\begin{aligned}&\dim H^1(X^{\mathrm{cut}};\mathbb{R})-\dim H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R})-\dim H^0(X^{\mathrm{cut}};\mathbb{R})+\dim H^0(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R})\\ &=\dim H^1(X;\mathbb{R})-\dim H^1(X,\partial X;\mathbb{R})-\dim H^0(X;\mathbb{R})+\dim H^0(X,\partial X;\mathbb{R})+\dim H^1(\Sigma;\mathbb{R})\\ &\quad-2\dim H^0(\Sigma;\mathbb{R}).\end{aligned}\tag{VI.30}$$

(ii) The exact cohomology sequence for the pair of spaces $(\partial X \sqcup \Sigma) \subset X$ and (VI.15)–(VI.16) give

$$0 \to H^0(X;\mathbb{R}) \to H^0(\partial X \sqcup \Sigma;\mathbb{R}) \to H^1(X,\partial X \sqcup \Sigma;\mathbb{R}) \to H^1(X;\mathbb{R}) \xrightarrow{i_X^* \times (i_\Sigma^*, i_\Sigma^*)} T(\Lambda_{X^{\mathrm{cut}}} \cap C) \to 0.$$

Together with the isomorphism $H^1(X,\partial X \sqcup \Sigma;\mathbb{R}) \cong H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R})$, it implies that

$$\begin{aligned}\dim(\Lambda_{X^{\mathrm{cut}}} \cap C) = {} & \dim H^1(X;\mathbb{R}) - \dim H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R}) + \dim H^0(\partial X;\mathbb{R}) + \dim H^0(\Sigma;\mathbb{R}) \\ & - \dim H^0(X;\mathbb{R}). \end{aligned} \tag{VI.31}$$

(iii) From the exact sequence (II.6) and the fact that $\dim \Lambda_X = \frac{1}{2}\dim H^1(\partial X;\mathbb{R})$, we obtain

$$\begin{aligned}\tfrac{1}{2} \dim H^1(\partial X;\mathbb{R}) = {} & \dim H^1(X;\mathbb{R}) - \dim H^1(X,\partial X;\mathbb{R}) + \dim H^0(\partial X;\mathbb{R}) \\ & - \dim H^0(X;\mathbb{R}) + \dim H^0(X,\partial X;\mathbb{R}). \end{aligned} \tag{VI.32}$$

The claimed formula (VI.27) follows then from the relations (VI.30), (VI.31), (VI.32), and the definition (V.15).

Finally, let us prove the relation (VI.26). Referring back to the definition (V.12), we have

$$\begin{aligned}\mu_X &= \left[\int_{H^1(X,\partial X;\mathbb{T})} \mathbf{w} \right] (T_X)^{1/2} \otimes w^{-1}, \\ \mu_{X^{\mathrm{cut}}} &= \left[\int_{H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{T})} \mathbf{w}_{\mathrm{cut}} \right] (T_{X^{\mathrm{cut}}})^{1/2} \otimes w_{\mathrm{cut}}^{-1}, \end{aligned} \tag{VI.33}$$

with $w \in |\mathrm{Det}\, H^1(X,\partial X;\mathbb{R})^*|$ and $w_{\mathrm{cut}} \in |\mathrm{Det}\, H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R})^*|$. On the other hand, we introduced a_X and $a_{X^{\mathrm{cut}}}$, such that

$$\begin{aligned}\mu_X &= a_X \delta_X, \\ \mu_{X^{\mathrm{cut}}} &= \alpha_{X^{\mathrm{cut}}} \delta_{X^{\mathrm{cut}}}. \end{aligned} \tag{VI.34}$$

The R-torsion T_X of X is related to the R-torsions $T_{X^{\mathrm{cut}}}$ of X^{cut} and T_Σ by Σ by the gluing formula[26]

$$T_X = T_{X^{\mathrm{cut}}} \otimes (T_\Sigma)^{-1}, \tag{VI.35}$$

where the identification between the lhs and the rhs is made through the isomorphism of determinant lines,

$$|\mathrm{Det}\, H^\bullet(X;\mathbb{R})^*| \cong |\mathrm{Det}\, H^\bullet(X^{\mathrm{cut}};\mathbb{R})^*| \otimes |\mathrm{Det}\, H^\bullet(\Sigma;\mathbb{R})|,$$

arising from the Mayer–Vietoris sequence (VI.29). Using the exact sequence (II.6) and, due to the normalization of the density δ_X^2 on Λ_X, we see that

$$\frac{w \otimes \delta_X^2}{[\int_{H^1(X,\partial X;\mathbb{T})} \mathbf{w}]} \in |\mathrm{Det}\, H^1(X;\mathbb{R})^*|,$$

defines an invariant density on the group $H^1(X;\mathbb{T})$ that gives $H^1(X;\mathbb{T})$ volume 1. Similarly on $H^1(X^{\mathrm{cut}};\mathbb{T})$, we have the density defined by

$$\frac{w_{\mathrm{cut}}\otimes \delta^2_{X^{\mathrm{cut}}}}{[\int_{H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{T})}\mathbf{w}_{cut}]}\in |\mathrm{Det}\ H^1(X^{\mathrm{cut}};\mathbb{R})^*|.$$

The Mayer–Vietoris sequence (VI.29) and Poincaré duality induce the isomorphism of determinant lines,

$$|\mathrm{Det}\ H^1(X;\mathbb{R})^*|\cong |\mathrm{Det}\ H^1(X^{\mathrm{cut}};\mathbb{R})^*|\otimes |\mathrm{Det}\ H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{R})^*|\otimes |\mathrm{Det}\ H^1(\Sigma;\mathbb{R})|$$
$$\otimes |\mathrm{Det}\ H^1(X,\partial X;\mathbb{R})|.$$

Thus, we can define an invariant density ρ_X on the group $H^1(X;\mathbb{T})$ by setting

$$\rho_X=\frac{w_{\mathrm{cut}}\otimes \delta^2_{X^{\mathrm{cut}}}}{[\int_{H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{T})}\mathbf{w}_{\mathrm{cut}}]}\otimes\frac{w_{\mathrm{cut}}}{[\int_{H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{T})}\mathbf{w}_{\mathrm{cut}}]}\otimes T_\Sigma^{-1}\otimes\frac{w^{-1}}{[\int_{H^1(X,\partial X;\mathbb{T})}\mathbf{w}]^{-1}}.$$

From the normalization of the densities on the rhs and from the Mayer–Vietoris cohomology sequence with $\mathbb{T}$ coefficients for $X,X^{\mathrm{cut}},\Sigma$,

$$\cdots\to H^1(X;\mathbb{T})\to H^1(X^{\mathrm{cut}};\mathbb{T})\to H^1(\Sigma;\mathbb{T})\to H^2(X;\mathbb{T})\to H^2(X^{\mathrm{cut}};\mathbb{T})\to H^2(\Sigma;\mathbb{T})\to 0,$$

we see that, since $\pi_0[H^2(X;\mathbb{T})]\cong\pi_0[H^2(X^{\mathrm{cut}};\mathbb{T})]=0$, and since, as shown in the proof of Proposition (II.5), $\pi_0[H^1(X,\partial X;\mathbb{T})]\cong\mathrm{Tors}\ H^2(X;\mathbb{Z})$ and $\pi_0[H^1(X^{\mathrm{cut}},\partial X^{\mathrm{cut}};\mathbb{T})]\cong\mathrm{Tors}\ H^2(X^{\mathrm{cut}};\mathbb{Z})$, the density ρ_X gives to $H^1(X;\mathbb{T})$ volume equal to $[\#\mathrm{Tors}\ H^2(X;\mathbb{Z})]/[\#\mathrm{Tors}\ H^2(X^{\mathrm{cut}};\mathbb{Z})]$. Making use of the gluing formula (VI.35) and the relations (VI.33), (VI.34), we can write

$$\frac{w\otimes\delta^2_X}{[\int_{H^1(X,\partial X;\mathbb{T})}\mathbf{w}]}=\frac{1}{a_X^2}\,T_X\otimes\frac{w^{-1}}{[\int_{H^1(X,\partial X;\mathbb{T})}\mathbf{w}]^{-1}}=\frac{1}{a_X^2}\,T_{X^{\mathrm{cut}}}\otimes(T_\Sigma)^{-1}\otimes\frac{w^{-1}}{[\int_{H^1(X,\partial X;\mathbb{T})}\mathbf{w}]^{-1}}=\frac{a^2_{X^{\mathrm{cut}}}}{a_X^2}\,\rho_X.$$

In view of the previous observation regarding the volume of $H^1(X;\mathbb{T})$ computed with ρ_X and the fact that the density of the above equation gives $H^1(X;\mathbb{T})$ volume 1, the claimed relation (VI.26) follows. □

The mapping class group Γ_Σ of a closed oriented 2-manifold Σ is the group $\mathrm{Diff}_+(\Sigma)/\mathrm{Diff}_0(\Sigma)$ of isotopy classes of orientation preserving diffeomorphisms of Σ. The *mapping class group* $\Gamma_{(\Sigma,L)}$ of an e-2-manifold (Σ,L) is a central extension by $\mathbb{Z}/8\mathbb{Z}$ of Γ_Σ:

$$1\to\mathbb{Z}/8\mathbb{Z}\to\Gamma_{(\Sigma,L)}\to\Gamma_\Sigma\to 1.$$

As a set $\Gamma_{(\Sigma,L)}=\Gamma_\Sigma\times\mathbb{Z}/8\mathbb{Z}$ and the composition law in $\Gamma_{(\Sigma,L)}$ is, according to (VI.4), given by

$$([h],m)([h'],m')=([hh'],m+m'+\tau(L'',h'^*L',(hh')^*L)\pmod 8).$$

In the course of proving (a) of Theorem VI.11, we showed that, given a rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$, each element $[h]$ of the mapping class group Γ_Σ determines a unitary map of Hilbert spaces $h^*:\mathscr{H}(\Sigma,L)\to\mathscr{H}(\Sigma,h^*L)$. The composition of this map with the isomorphism $F_{L,h^*L}:\mathscr{H}(\Sigma,h^*L)\to\mathscr{H}(\Sigma,L)$ induced by the BKS pairing is a unitary operator $U_L([h])=F_{L,h^*L}\circ h^*$ on the Hilbert space $\mathscr{H}(\Sigma,L)$. The assignment,

$$[h]\mapsto U_L([h]),$$

defines a unitary projective representation of Γ_Σ on $\mathscr{H}(\Sigma,L)$. For the group $\Gamma_{(\Sigma,L)}$ the assignment,

$$([h],m)\mapsto U_L([h],m)=\mathrm{e}^{(\pi i/4)m}U_L([h]),$$

determines a unitary representation of this group on the Hilbert space $\mathscr{H}(\Sigma,L)$.

Since the mapping class group Γ_Σ acts on $H^1(\Sigma;\mathbb{Z})$, there is a natural homomorphism from Γ_Σ to the group $Sp(\mathscr{Z})$ of symplectic transformations of $(H^1(\Sigma;\mathbb{R}),\omega_\Sigma)$ that preserve the integer lattice $\mathscr{Z}=H^1(\Sigma;\mathbb{Z})$. According to the results in (Ref. 10, Sec. 9.1), for each rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$, there is a projective unitary representation of $Sp(\mathscr{Z})$ on the Hilbert space $\mathscr{H}(\Sigma,L)$. Let us choose an integer symplectic basis $([\dot{\mathbf{w}}];[\dot{\mathbf{w}}'])$ for $H^1(\Sigma;\mathbb{R})$, with $[\dot{w}_1],\ldots,[\dot{w}_g]$ spanning L, where $g=\frac{1}{2}\dim H^1(\Sigma;\mathbb{R})$. As shown in (Ref. 10, Sec. III), the choice of such a basis uniquely determines a unitary basis $\{v(\Sigma,L)_{\mathbf{q}}\}_{\mathbf{q}\in(\mathbb{Z}/k\mathbb{Z})^g}$ for $\mathscr{H}(\Sigma,L)$. The generators of $Sp(\mathscr{Z})$ are the elements with matrix form

$$\alpha=\begin{pmatrix} A & 0 \\ 0 & {}^tA^{-1}\end{pmatrix},\quad \beta=\begin{pmatrix} I & B \\ 0 & I\end{pmatrix},\quad \gamma=\begin{pmatrix} 0 & I \\ -I & 0\end{pmatrix},$$

with respect to the basis $([\dot{\mathbf{w}}];[\dot{\mathbf{w}}'])$, where $A\in GL(g,\mathbb{Z})$ and $B\in M(g,\mathbb{Z})$, ${}^tB=B$. The projective unitary representation of $Sp(\mathscr{Z})$ on $\mathscr{H}(\Sigma,L)$ is described by the following operators representing the generators:

$$U_L(\alpha)v(\Sigma,L)_{\mathbf{q}}=v(\Sigma,L)_{{}^tA^{-1}\mathbf{q}},$$

$$U_L(\beta)v(\Sigma,L)_{\mathbf{q}}=e^{(\pi i/k)^t\mathbf{q}B\mathbf{q}}v(\Sigma,L)_{\mathbf{q}},\tag{VI.36}$$

$$U_L(\gamma)v(\Sigma,L)_{\mathbf{q}}=k^{-q/2}\sum_{\mathbf{q}_1\in(\mathbb{Z}/k\mathbb{Z})^g} e^{(2\pi i/k)^t\mathbf{q}\mathbf{q}_1}v(\Sigma,L)_{\mathbf{q}_1}.$$

For $g=1$ we have the torus mapping class group $SL(2,\mathbb{Z})$ with standard generators the matrices,

$$S=\begin{pmatrix} 0 & 1 \\ -1 & 0\end{pmatrix}\quad\text{and}\quad T=\begin{pmatrix} 1 & 1 \\ 0 & 1\end{pmatrix},$$

subject to the relations $(ST)^3=I$, $S^4=I$. The projective unitary representation of $\mathrm{SL}(2,\mathbb{Z})$ on the Hilbert space $\mathscr{H}(T^2,L)$ associated to the torus T^2 and $L\subset H^1(T^2;\mathbb{R})$ is described by the operators:

$$U_L(T)_{qq'}=e^{(\pi i/k)q^2}\delta_{qq'},$$

$$U_L(S)_{qq'}=k^{-1/2}e^{(2\pi i/k)qq'}.$$

It coincides with the representation of the modular group $\mathbb{P}SL(2,\mathbb{Z})$ in two-dimensional rational conformal field theory.[29] The projective representation (VI.36) of $Sp(\mathscr{Z})$ on $\mathscr{H}(\Sigma,L)$ is the same projective representation as the one constructed in Refs. 9 and 30 on the vector space of theta functions at level k obtained through the quantization of the moduli space of flat $\mathbb{T}$ connections on Σ in a holomorphic polarization.

ACKNOWLEDGMENT

I would like to thank Professor Dan Freed for introducing me to the Chern–Simons theory, as well as for the helpful suggestions and advice during the development of this project.

[1] E. Witten, ''Quantum field theory and the Jones polynomial,'' Commun. Math. Phys. **121**, 351–399 (1989).
[2] M. F. Atiyah, ''On framings of 3-manifolds,'' Topology **29**, 1–7 (1990).
[3] M. F. Atiyah, ''Topological quantum field theory,'' Publ. Math. Inst. Hautes Etudes Sci. Paris **68**, 175–186 (1989).
[4] K. Walker, ''On Witten's 3-manifold invariants,'' preprint, 1991.

[5] N. Reshetikhin and V. Turaev, ''Invariants of three-manifolds via link polynomials and quantum groups,'' Invent. Math. **103**, 547–597 (1991).
[6] D. S. Freed and F. Quinn, ''Chern–Simons theory with finite gauge group,'' Commun. Math. Phys. **156**, 453–472 (1993).
[7] S. Axelrod, S. Della Pietra, and E. Witten, ''Geometric quantization of Chern–Simons gauge theory,'' J. Diff. Geom. **33**, 787–902 (1991).
[8] M. F. Atiyah, *The Geometry and Physics of Knots* (Cambridge, University Press, Cambridge, 1990).
[9] T. Gocho, ''The topological invariant of three-manifolds based on the U(1) gauge theory,'' J. Fac. Sci., Univ. Tokyo, Sect. 1 **39**, 169–184 (1992).
[10] M. Manoliu, ''Quantization of symplectic tori in a real polarization,'' J. Math. Phys. **38**, 2219–2254 (1997).
[11] L. Jeffrey and J. Weitsman, ''Half density quantization of the moduli space of flat connections and Witten's semiclassical manifold invariant,'' Topology **32**, 509–529 (1993).
[12] D. S. Freed, ''Classical Chern–Simons theory. I,'' Adv. Math. **113**, 237–303 (1995).
[13] D. S. Freed and R. Gompf, ''Computer calculations of Witten's 3-manifold invariants,'' Commun. Math. Phys. **141**, 79–117 (1991).
[14] M. F. Atiyah and R. Bott, ''The Yang–Mills equations over Riemann surfaces,'' Philos. Trans. R. Soc. London, Ser. A **308**, 523–615 (1982).
[15] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. 1.
[16] G. E. Bredon, *Topology and Geometry* (Springer-Verlag, New York, 1993).
[17] S. Axelrod, ''Geometric quantization of Chern–Simons gauge theory,'' Princeton University Ph.D. thesis, 1991.
[18] B. Kostant, ''Quantization and unitary representations,'' *Lecture Notes in Mathematics* (Springer, Berlin, 1970), Vol. 170, pp. 87–208.
[19] N. M. J. Woodhouse, *Geometric Quantization*, 2nd ed. (Clarendon, Oxford, 1992).
[20] J. Sniatycki, *Geometric Quantization and Quantum Mechanics* (Springer-Verlag, New York, 1980).
[21] M. de Gosson, ''The structure of q-symplectic geometry,'' J. Math. Pure Appl. **71**, 429–453 (1992).
[22] G. Lion and M. Vergne, ''The Weil representation, Maslov index and Theta series,'' *Progress in Mathematics*, Boston 6 (Birkhäuser, Boston, 1980).
[23] D. B. Ray and I. M. Singer, ''R-torsion and the Laplacian on Riemannian manifolds,'' Adv. Math. **7**, 145–210 (1971).
[24] D. B. Ray and I. M. Singer, ''Analytic torsion,'' Proceedings of the Symposium Pure Mathematics, American Mathematical Society, Providence, RI, 1973, Vol. 23, pp. 167–182.
[25] W. Müller, ''Analytic torsion and R-torsion of Riemannian manifolds,'' Adv. Math. **28**, 233–305 (1978).
[26] S. M. Vishik, ''Generalized Ray-Singer conjecture I. A manifold with smooth boundary,'' Commun. Math. Phys. **167**, 1–102 (1995).
[27] E. Witten, ''On quantum gauge theories in two dimensions,'' Commun. Math. Phys. **141**, 153–209 (1991).
[28] A. Weinstein, *Lectures on Symplectic Manifolds* (American Mathematical Society, Providence, RI, 1977), No. 29.
[29] E. Verlinde, ''Fusion rules and modular transformations in 2d conformal field theory,'' Nucl. Phys. B **300**, 360–376 (1988).
[30] L. Funar, ''Representations of the symplectic group and 3-manifolds,'' C. R. Acad. Sci. Paris **316**, 1067–1072 (1993).

Abelian Chern–Simons theory. II. A functional integral approach

Mihaela Manoliu[a)]
Department of Mathematics, University of Texas, Austin, Texas, 78712

(Received 24 March 1997; accepted for publication 29 August 1997)

Following Witten, [Commun. Math. Phys. **21**, 351–399 (1989)] we approach the Abelian quantum Chern–Simons (CS) gauge theory from a Feynman functional integral point of view. We show that for 3-manifolds with and without a boundary the formal functional integral definitions lead to mathematically proper expressions that agree with the results from the rigorous construction [J. Math. Phys. **39**, 170–206 (1998)] of the Abelian CS topological quantum field theory via geometric quantization. © *1998 American Institute of Physics.* [S0022-2488(98)01401-7]

I. INTRODUCTION

A rigorous construction of the CS theory for the Abelian group $\mathbb{T}$ as a topological quantum field theory satisfying Atiyah's axioms[1] is presented in Ref. 2 ($\mathbb{T}$ denotes the group of complex numbers of unit modulus). It is shown that to a closed oriented two-dimensional manifold Σ together with a rational Lagrangian subspace $L\subset H^1(\Sigma;\mathbb{R})$ one can canonically associate a quantum Hilbert space $\mathscr{H}(\Sigma,L)$ via the geometric quantization of the moduli space $\mathscr{M}_\Sigma$ of flat $\mathbb{T}$ connections on Σ. A compact oriented three-dimensional manifold X determines naturally a Lagrangian subspace $L_X\subset H^1(\partial X;\mathbb{R})$. The constructions in Ref. 2 assign canonically to X a vector Z_X in the Hilbert space $\mathscr{H}(\partial X,L_X)$ associated to the boundary ∂X and the Lagrangian subspace L_X. If the 3-manifold X is closed, then Z_X is a complex number representing a topological invariant of X.

In this paper we follow Witten's[3] approach and introduce the quantity Z_X to be associated to a 3-manifold as a Feynman functional integral (path integral). As in Ref. 3, we formally define Z_X for closed 3-manifolds as an integral over gauge equivalence classes of $\mathbb{T}$ connections on X with action the Chern–Simons functional. For the definition of the functional integral and its subsequent evaluation we rely almost entirely on the results in Refs. 4, 5. The derivation is similar to the one presented in Ref. 6 in the context of the non-Abelian version of the CS theory, which evaluates Z_X in the so-called semiclassical or stationary phase approximation limit. In the Abelian CS theory the functional integral is of a Gaussian type and the stationary phase evaluation of Z_X is the exact result. It is expressed in terms of the CS invariant of flat connections on principal $\mathbb{T}$ bundles over X with first Chern class torsion and the Ray–Singer analytic torsion invariant of X. We extend the definition of Z_X as a functional integral to the case of a 3-manifold X with nonempty boundary ∂X. In this instance the path integral is defined over a space of connections that are fixed over the boundary. Hence it is a function of the boundary connections and, in view of the results in Ref. 2, Sec. V, it can be interpreted as an element of the quantum Hilbert space $\mathscr{H}(\partial X,L_X)$.

In Sec. II we briefly review some results on integration of invariant functions and the stationary phase approximation formula for oscillatory integrals. In Sec. III and Sec. IV we discuss the definition of Z_X for compact 3-manifolds X without boundary and with boundary, respectively. We use the same notations as in Ref. 2 and we refer the reader to this reference for some definitions that we do not restate in order to avoid redundancy.

[a)]Electronic mail: miha@math.utexas.edu

II. A FINITE-DIMENSIONAL MODEL

We begin with a brief description of a finite-dimensional model that will serve as a prototype for the problem in infinite dimensions of defining Z_X as a functional integral. The main references used are Refs. 4 and 5.

Let G be a compact Lie group acting as a group of isometries of the Riemannian manifold (E,g_E). We assume that the Lie algebra Lie G of G is endowed with an invariant inner product that defines an invariant Riemannian metric on G. Then Vol G denotes the volume of G with respect to this metric. The action of G on E generates a homomorphism of Lie G into the Lie algebra of vector fields on E, which defines for each point $x\in E$ a linear map τ_x:Lie $G\to T_xE$. Let H_x denote the isotropy subgroup of G at the point $x\in E$ and Vol H_x the volume of H_x with respect to the metric induced by the invariant metric in G. We assume that the isotropy subgroups at all points $x\in E$ are conjugate to a fixed subgroup $H\subset G$. Then Vol $H_x=$Vol H, for all $x\in E$. The G-invariant Riemannian metric g_E on E induces a Riemannian metric $g_{E/G}$ on the space E/G of orbits of G in E. It is defined by setting for any $v,w\in T_{[x]}(E/G)$: $g_{E/G}(v,w)=g_E(\hat{v},\hat{w})$, where $\hat{v},\hat{w}\in T_xE$ project onto v,w and belong to the orthogonal complement of τ_x(Lie G) in T_xE. We let μ_E and $\mu_{E/G}$ be the measures on E and E/G determined by these metrics. For any G-invariant function h on E, we have[4]

$$\int_E h(x)\mu_E=\frac{\text{Vol } G}{\text{Vol } H}\int_{E/G} h(x)[\det'(\tau_x^*\tau_x)]^{1/2}\mu_{E/G}. \tag{II.1}$$

The notation det′ refers to the regularized determinant of the operator, that is, the product of all the nonzero eigenvalues. The G-invariant term (Vol G/Vol H)$[\det'(\tau_x^*\tau_x)]^{1/2}$ is the volume of the orbit through the point x.

Let us consider now a G-invariant real-valued function f on E such that the stationary points of f form a G-invariant submanifold F of E and $f(x)=\bar{f}$ for all $x\in F$. The Hessian (Hess f)$_x$ of f at the point $x\in E$ is a linear self-adjoint operator (Hess f)$_x$:$T_xE\to T_xE$ and, since f is G invariant, τ_x(Lie G)$\subset$Ker(Hess f)$_x$. We assume that $T_xF=$Ker(Hess f)$_x$ at every point $x\in F$. We also assume that the space $M=F/G$ of orbits of G in F is a manifold and we let μ_M denote the measure on M determined by the natural metric on M induced from the Riemannian metric of F. The method of the stationary phase Refs. 4 and 7 together with (II.1) give, for the asymptotic evaluation of the oscillatory integral,

$$\frac{1}{\text{Vol } G}\int_E e^{\pi i a f(x)}\mu_E \quad (a\in\mathbb{R}_+), \tag{II.2}$$

in the limit of large a, the following expression:[4,5]

$$\frac{1}{a^{(\dim E-\dim F)/2}}\frac{e^{\pi i a\bar{f}}}{\text{Vol } H}\int_M e^{(\pi i/4)\text{sgn}(\text{Hess } f)_x}\frac{|\det'(\tau_x^*\tau_x)|^{1/2}}{|\det'(\text{Hess } f)_x|^{1/2}}\mu_M. \tag{II.3}$$

sgn(Hess f)$_x$ denotes the signature of the quadratic form associated to (Hess f)$_x$. In the terminology of Refs. 4, 5 the integral (II.2) is the *partition function* of the function(al) f.

III. THE PARTITION FUNCTION FOR CLOSED MANIFOLDS

We turn now to the problem of defining Z_X as a functional integral when X is a *closed* connected oriented 3-manifold. For each torsion class $p\in\text{Tors } H^2(X;\mathbb{Z})$, we choose a principal $\mathbb{T}$ bundle π:$P\to X$ with the first Chern class $c_1(P)=p$. The space $\mathscr{A}_P$ of connections on P is an affine space with vector space $2\pi i\Omega^1(X;\mathbb{R})$. The group of gauge transformations $\mathscr{G}_P\cong\text{Map}(X,\mathbb{T})$ acts on $\mathscr{A}_P$ by

$$\Theta \cdot u = \Theta + (u \cdot \pi)^* \vartheta, \tag{III.1}$$

where ϑ is the Maurer–Cartan form of $\mathbb{T}$. The Chern–Simons functional (see Ref. 2, Def. III.1),

$$S_{X,P} : \mathscr{A}_P \to \mathbb{R}/\mathbb{Z}$$

is invariant under this action. We define Z_X by the expression

$$Z_X = \sum_{p \in \text{Tors } H^2(X;\mathbb{Z})} Z_{X,p}, \tag{III.2}$$

where $Z_{X,p}$ is the partition function of the Chern–Simons functional, that is,

$$Z_{X,p} = \int_{\mathscr{A}_P/\mathscr{G}_P} [\mathscr{D}\Theta] e^{\pi i k S_{X,P}(\Theta)}. \tag{III.3}$$

The rhs of the above equation is just a formal expression meant to suggest that, pending the existence of a ''measure'' $[\mathscr{D}\Theta]$, we integrate over the space of gauge equivalence classes of connections on P. We are going to show in the following that this functional integral can be given a precise meaning.

Let us choose a Riemannian metric on X. For each $q = 0,1,...,\dim X$, the metric determines an inner product on the space $\Omega^q(X;\mathbb{R})$ of q forms on X:

$$(\alpha, \beta) = \int_X \alpha \wedge * \beta. \tag{III.4}$$

In particular, it defines an inner product on the tangent space $T\mathscr{A}_P \cong 2\pi i \Omega^1(X;\mathbb{R})$. This makes $\mathscr{A}_P$ a Riemannian manifold and the group $\mathscr{G}_P$ acts on $\mathscr{A}_P$ by isometries. We let $\hat{\mu}$ denote the measure on $\mathscr{A}_P$ determined by this metric and μ the induced measure on the quotient space $\mathscr{A}_P/\mathscr{G}_P$. For each $\Theta \in \mathscr{A}_P$, let τ_Θ denote the differential of the map from $\mathscr{G}_P$ to $\mathscr{A}_P$, which defines by (III.1) the $\mathscr{G}_P$ action. Since Lie $\mathscr{G}_P \cong 2\pi i \Omega^0(X;\mathbb{R})$, the linear map τ_Θ is identified with the exterior differential $\tau_\Theta = d : \Omega^0(X;\mathbb{R}) \to \Omega^1(X;\mathbb{R})$. Moreover, the inner product in Lie $\mathscr{G}_P$ defined by (III.4) induces an invariant metric on the group $\mathscr{G}_P$. The isotropy subgroup of $\mathscr{G}_P$ at a point $\Theta \in \mathscr{A}_P$ is the group of constant maps from X into $\mathbb{T}$; hence, it is isomorphic to $\mathbb{T}$. We let Vol $\mathbb{T}$ denote the volume of the isotropy subgroup, computed with respect to the metric induced from that of $\mathscr{G}_P$. Its evaluation gives

$$\text{Vol } \mathbb{T} = [\text{Vol } X]^{1/2} = \left[\int_X *1 \right]^{1/2}. \tag{III.5}$$

Thus, besides the fact that we are dealing with infinite-dimensional spaces, the setting is the same as the one in Sec. II. We have the identifications $E = \mathscr{A}_P$, $G = \mathscr{G}_P$, and $f = kS_{X,P}$. The stationary points of the Chern–Simons functional $S_{X,P}$ are the flat connections; hence the critical manifold F is the subspace $\mathscr{A}_P^f = \{\Theta \in \mathscr{A}_P | F_\Theta = d\Theta = 0\}$ and the constant value of the functional $S_{X,P}$ on $\mathscr{A}_P^f$ is the Chern–Simons invariant of flat connections on P. If Θ_P is a flat connection in $\mathscr{A}_P$, any $\Theta \in \mathscr{A}_P$ can be written as $\Theta = \Theta_P + 2\pi i A$, for some $A \in \Omega^1(X;\mathbb{R})$. Using (Ref. 2, Lemma III.5), we can express

$$S_{X,P}(\Theta) = S_{X,P}(\Theta_P) - \int_X A \wedge dA \pmod 1 = S_{X,P}(\Theta_P) - (A, *dA) \pmod 1.$$

The Hessian of the functional $f=kS_{X,P}$ is the linear operator Hess $f=-k*d:\Omega^1(X;\mathbb{R})\to\Omega^1(X;\mathbb{R})$. The quotient manifold $\mathscr{M}_P=\mathscr{A}_P^f/\mathscr{G}_P$, the moduli space of flat $\mathbb{T}$ connections on P, is isomorphic to the torus $H^1(X;\mathbb{R})/H^1(X;\mathbb{Z})$. The inner product (III.4) on 1-forms determines a natural inner product on $H^1(X;\mathbb{R})$ through the Hodge–deRham isomorphism $H^1(X;\mathbb{R})\cong\mathscr{H}^1(X)$ with the space of harmonic 1-forms. Since the tangent space $T_{[\Theta]}\mathscr{M}_P\cong H^1(X;\mathbb{R})$, at any $[\Theta]\in\mathscr{M}_P$, this inner product defines a measure ν on $\mathscr{M}_P$.

In view of these observations and of the results (II.1)–(II.3) from the model of Sec. II, we *formally* write

$$Z_{X,p}=\frac{1}{\mathrm{Vol}\ \mathscr{G}_P}\int_{\mathscr{A}_P}e^{\pi ikS_{X,P}(\Theta)}\hat{\mu} \tag{III.6}$$

$$=\frac{1}{\mathrm{Vol}\ \mathbb{T}}\int_{\mathscr{A}_P/\mathscr{G}_P}e^{\pi ikS_{X,P}(\Theta)}[\det'(\tau_\Theta^*\tau_\Theta)]^{1/2}\mu \tag{III.6$'$}$$

$$=\frac{e^{\pi ikS_{X,P}(\Theta_P)}}{\mathrm{Vol}\ \mathbb{T}}\int_{\mathscr{M}_P}e^{(\pi i/4)\mathrm{sgn}(-*d)}\frac{[\det'(\tau_\Theta^*\tau_\Theta)]^{1/2}}{[\det'(-k*d)]^{1/2}}\,\nu. \tag{III.6$''$}$$

The expression (III.6$'$) will be taken as the formal definition for $Z_{X,p}$ in (III.3). The equality (III.6$''$) is the result of the fact that, since $S_{X,P}$ is a quadratic functional, the stationary phase method that gives for an oscillatory integral of the type (II.2) the asymptotic evaluation (II.3), produces in this case the exact result. We are going to show that the expression in (III.6$''$) has rigorous mathematical meaning if the determinants and signatures of the operators therein are regularized according to the definitions (A1) and (A4).

If $\Delta_q=d^*d+dd^*$ is the Laplacian on $\Omega^q(X;\mathbb{R})$ and if we let $\zeta_q=\zeta_{\Delta_q}$, then for any real number $\lambda>0$,

$$\det'(\lambda\Delta_q)=\lambda^{\zeta_q(0)}\det'(\Delta_q), \tag{III.7}$$

and[8]

$$\zeta_q(0)=-\dim\mathrm{Ker}\ \Delta_q=-\dim\ H^q(X;\mathbb{R}). \tag{III.8}$$

Now, let us make use of the result (A3) to evaluate the regularized determinant $\det'(-k*d)$. We let $S=-k*d$ acting on $\Omega^1(X;\mathbb{R})$ and $T=kd:\Omega^0(X;\mathbb{R})\to\Omega^1(X;\mathbb{R})$. Then $S^2+TT^*=k^2(d^*d+dd^*)=k^2\Delta_1$ and $T^*T=k^2d^*d=k^2\Delta_0$. Using (A3) and (III.7)–(III.8) we get

$$\det'(-k*d)=\frac{[\det'(k^2\Delta_1)]^{1/2}}{[\det'(k^2\Delta_0)]^{1/2}}=\frac{k^{-\dim H^1(X;\mathbb{R})}}{k^{-\dim H^0(X;\mathbb{R})}}\frac{[\det'\ \Delta_1]^{1/2}}{[\det'\ \Delta_0]^{1/2}}. \tag{III.9}$$

We also note that $\tau_\Theta^*\tau_\Theta=d^*d=\Delta_0$. The signature of the operator $-*d$ on $\Omega^1(X;\mathbb{R})$ is regularized via the eta invariant, that is, $\mathrm{sgn}(-*d)=\eta(-*d)$. Inserting the above results into the expression (III.6$''$), we obtain

$$Z_{X,p}=k^{m_X}e^{\pi ikS_{X,P}(\Theta_P)}e^{(\pi i/4)\,\eta(-*d)}\int_{\mathscr{M}_P}\frac{1}{[\mathrm{Vol}\ X]^{1/2}}\frac{[\det'\ \Delta_0]^{3/4}}{[\det'\ \Delta_1]^{1/4}}\,\nu, \tag{III.10}$$

where $m_X=\frac{1}{2}(\dim H^1(X;\mathbb{R})-\dim H^0(X;\mathbb{R}))$. The expression under the integral sign in the above formula is related to the Ray–Singer analytic torsion of X.

The analytic torsion of a closed Riemannian manifold was introduced in Refs. 9 and 10 as a norm on the determinant line $|\mathrm{Det}\, H^{\bullet}(X;\mathbb{R})|$ of the deRham cohomology of the manifold. For a closed 3-manifold X endowed with a Riemannian metric, the Ray–Singer analytic torsion is the density:[10]

$$T_X^a \in |\mathrm{Det}\, H^0(X;\mathbb{R})| \otimes |\mathrm{Det}\, H^1(X;\mathbb{R})^*| \otimes |\mathrm{Det}\, H^2(X;\mathbb{R})| \otimes |\mathrm{Det}\, H^3(X;\mathbb{R})^*|,$$

$$T_X^a = \delta_{|\mathrm{Det}\, H^{\bullet}(X;\mathbb{R})|} \cdot \exp\left[\frac{1}{2}\sum_{q=0}^{\dim X}(-1)^q q\,\zeta_q'(0)\right]. \tag{III.11}$$

The inner product on the space $\mathscr{H}^q(X)$ of harmonic q-forms defined by (III.4) determines an inner product on $H^q(X;\mathbb{R}) \cong \mathscr{H}^q(X)$ and, thus, a density $\delta_{|\mathrm{Det}\, H^{\bullet}(X;\mathbb{R})|}$ on $|\mathrm{Det}\, H^{\bullet}(X;\mathbb{R})|$. If $b_q = \dim \mathscr{H}^q(X)$ and $\nu_1^{(q)},\ldots,\nu_{b_q}^{(q)}$ is any orthonormal basis for $\mathscr{H}^q(X)$, then

$$\delta_{|\mathrm{Det}\, H^{\bullet}(X;\mathbb{R})|} = \bigotimes_{q=0}^{\dim X} |\nu^{(q)}|^{(-1)^q}, \tag{III.12}$$

where $\nu^{(q)} = \nu_1^{(q)} \wedge \cdots \wedge \nu_{b_q}^{(q)}$. According to the definition of regularized determinants we have $\det' \Delta_q = e^{-\zeta_q'(0)}$. Thus, T_X^a can be represented as

$$T_X^a = [(\det' \Delta_0)^0 (\det' \Delta_1)^1 (\det' \Delta_2)^{-2} (\det' \Delta_3)^3]^{1/2} \times |\nu^{(0)}| \otimes |\nu^{(1)}|^{-1} \otimes |\nu^{(2)}| \otimes |\nu^{(3)}|^{-1}. \tag{III.13}$$

The Hodge $*$-operator determines isomorphisms $\Delta_q \cong \Delta_{3-q}$ and Poincaré duality gives $H^q(X;\mathbb{R}) \cong H^{3-q}(X;\mathbb{R})^*$. Moreover, any orthonormal basis $\nu^{(0)}$ of $\mathscr{H}^0(X) \cong \mathbb{R}$ is a constant such that $|\nu^{(0)}| = [\mathrm{Vol}\, X]^{-1/2}$. Hence, the square root of the analytic torsion T_X^a can be regarded as a density $(T_X^a)^{1/2} \in |\mathrm{Det}\, H^1(X;\mathbb{R})^*|$, which is given by the expression

$$(T_X^a)^{1/2} = \frac{1}{[\mathrm{Vol}\, X]^{1/2}} \frac{[\det' \Delta_0]^{3/4}}{[\det' \Delta_1]^{1/4}}\, \nu, \tag{III.14}$$

where $\nu =: |\nu^{(1)}|^{-1} = |\nu_1^{(1)} \wedge \cdots \wedge \nu_{b_1}^{(1)}|^{-1}$ for any orthonormal basis $\nu_1^{(1)},\ldots,\nu_{b_1}^{(1)}$ of $\mathscr{H}^1(X)$. In view of the isomorphism $\mathscr{M}_P \cong H^1(X;\mathbb{R})/H^1(X;\mathbb{Z})$, the square root $(T_X^a)^{1/2}$ of the analytic torsion defines an invariant density on the moduli space of flat connections $\mathscr{M}_P$. Using (III.14), the expression (III.10) reads as

$$Z_{X,p} = k^{m_X} e^{\pi i k S_{X,P}(\Theta_P)} e^{(\pi i/4)\eta(-*d)} \int_{\mathscr{M}_P} (T_X^a)^{1/2}. \tag{III.15}$$

It is proved in Ref. 10 that the analytic torsion of a closed manifold is independent of the metric. Thus T_X^a is a manifold invariant. The only metric dependence in (III.15) comes from the phase factor $e^{(\pi i/4)\eta(-*d)}$. We recall that in the non-Abelian CS theory,[3] Witten compensates the metric dependence of the path integral by adding to the original action a counterterm, the gravitational CS action. Here, in order to make the resulting expression metric independent, *we choose* to modify our original definition (III.6) of $Z_{X,p}$ by multiplying that expression with $e^{-(\pi i/4)\eta(-*d)}$.

Starting with a heuristic functional integral formula, we have thus succeeded to associate to the closed connected 3-manifold X the topological invariant,

$$Z_X = k^{m_X} \sum_{p \in \mathrm{Tors}\, H^2(X;\mathbb{Z})} e^{\pi i k S_{X,P}(\Theta_P)} \int_{\mathscr{M}_P} (T_X^a)^{1/2}. \tag{III.16}$$

We note that the factor k^{m_X} agrees with the one appearing in Ref. 11, in the analogous expression of the closed 3-manifold invariant for the non-Abelian version of the Chern–Simons theory. A close derivation to the one given here for the factor k^{m_X}, as arising from the scaling of regularized determinants of Laplacians, is presented in Ref. 12.

IV. THE PARTITION FUNCTION FOR MANIFOLDS WITH BOUNDARY

We address now the problem of defining Z_X in the case of a compact connected oriented 3-manifold X *with boundary* ∂X. We fix a trivializable $\mathbb{T}$ bundle Q over ∂X and a section $s:\partial X\to Q$. Then, for each $p\in \mathrm{Tors}\, H^2(X;\mathbb{Z})$, we choose a $\mathbb{T}$ bundle $P\to X$ with $c_1(P)=p$ and fix a bundle isomorphism $\phi_P:\partial P\to Q$. The section $s:\partial X\to Q$ determines a section $s_P=\phi_P^{-1}\circ s:\partial X\to \partial P$. If η is a connection on Q, then we define the space

$$\mathscr{A}_P(\eta)=\{\Theta\in\mathscr{A}_P\,|\,\partial\Theta=\phi_P^*\eta\},$$

of connections on P whose restriction $\partial\Theta=\Theta|_{\partial X}$ to the boundary of X is identified with η under the given bundle isomorphism over ∂X. The space $\mathscr{A}_P(\eta)$ is an affine space with vector space $2\pi i\Omega^1_{1,\tan}(X;\mathbb{R})$, where $\Omega^1_{1,\tan}(X;\mathbb{R})$ denotes the space of 1-forms A on X whose geometrical restriction to ∂X is zero.

Let $\mathscr{G}_P(e)$ denote the subgroup in $\mathscr{G}_P$ of gauge transformations that restrict to the identity map over ∂X. The group $\mathscr{G}_P(e)$ acts on $\mathscr{A}_P(\eta)$ through (III.1) and the action is free. Given the section $s_P:\partial X\to\partial P$, the CS functional $S_{X,P}(s_P,\cdot):\mathscr{A}_P(\eta)\to\mathbb{R}/\mathbb{Z}$ (see Ref. 2, Def. III.7) is invariant under the $\mathscr{G}_P(e)$ action on $\mathscr{A}_P(\eta)$. As for the closed manifold case previously discussed, we introduce the partition function of the CS functional,

$$Z_{X,p}(\eta)=\int_{\mathscr{A}_P(\eta)/\mathscr{G}_P(e)}[\mathscr{D}\Theta]\mathrm{e}^{\pi ikS_{X,P}(s_P,\Theta)},\tag{IV.1}$$

and using similar methods aim to show in what follows that one can make sense of such a functional integral.

For this purpose let us choose a Riemannian metric on X. The inner product defined on the tangent space by the metric determines, at each point of ∂X, a normal vector to the boundary. Therefore, a differential form α on X can be decomposed into a tangential and a normal component $\alpha=\alpha_{\tan}+\alpha_{\mathrm{norm}}$ and one has $(*\alpha)_{\mathrm{norm}}=*(\alpha_{\tan})$, $(d\alpha)_{\tan}=d(\alpha_{\tan})$, $(d^*\alpha)_{\mathrm{norm}}=d^*(\alpha_{\mathrm{norm}})$. We introduce the subspaces in $\Omega^q(X;\mathbb{R})$ of q forms on X satisfying *relative* (Dirichelet) boundary conditions:[9]

$$\Omega^q_{\tan}(X;\mathbb{R})=\{\alpha\in\Omega^q(X;\mathbb{R})\,|\,\alpha_{\tan}=(d^*\alpha)_{\tan}=0\ \text{ on }\ \partial X\},$$

and *absolute* (Neumann) boundary conditions:

$$\Omega^q_{\mathrm{norm}}(X;\mathbb{R})=\{\beta\in\Omega^q(X;\mathbb{R})\,|\,\beta_{\mathrm{norm}}=(d\beta)_{\mathrm{norm}}=0\ \text{ on }\ \partial X\}.$$

Then we let $\Delta_q^{\tan}$ and Δ_q^{norm} denote the Laplace operator d^*d+dd^* acting on $\Omega^q_{\tan}(X;\mathbb{R})$ and $\Omega^q_{\mathrm{norm}}(X;\mathbb{R})$, respectively. The Hodge–deRham theory for manifolds with boundary gives the isomorphisms[8,9]

$$H^q(X,\partial X;\mathbb{R})\cong\mathscr{H}^q_{\tan}(X),$$

$$H^q(X;\mathbb{R})\cong\mathscr{H}^q_{\mathrm{norm}}(X),$$

where

$$\mathscr{H}^q_{\tan}(X)=\{\alpha\in\Omega^q(X;\mathbb{R})\,|\,\alpha_{\tan}=0 \ \text{ on } \ \partial X, d\alpha=d^*\alpha=0 \ \text{ on } \ X\},$$

$$\mathscr{H}^q_{\text{norm}}(X)=\{\beta\in\Omega^q(X;\mathbb{R})\,|\,\beta_{\text{norm}}=0 \ \text{ on } \ \partial X, d\beta=d^*\beta=0 \ \text{ on } \ X\},$$

are the spaces of harmonic forms on X with relative and absolute boundary conditions.

Let $\hat{\mu}$ be the measure on $\mathscr{A}_P(\eta)$ determined by the inner product (III.4) defined on each tangent space $T_\Theta\mathscr{A}_P(\eta)\cong\Omega^1_{1,\tan}(X;\mathbb{R})$ by the metric on X and let μ be the induced measure on the quotient space $\mathscr{A}_P(\eta)/\mathscr{G}_P(e)$. For each $\Theta\in\mathscr{A}_P(\eta)$, the $\mathscr{G}_P(e)$ action on $\mathscr{A}_P(\eta)$ determines a linear map $\tau_\Theta=d:\text{Lie}\ \mathscr{G}_P(e)\cong\Omega^0_{\tan}(X;\mathbb{R})\to T_\Theta\mathscr{A}_P(\eta)\cong\Omega^1_{1,\tan}(X;\mathbb{R})$.

The stationary points of the Chern–Simons functional $S_{X,P}(s_P,\cdot):\mathscr{A}_P(\eta)\to\mathbb{R}/\mathbb{Z}$ form the subspace of flat connections $\mathscr{A}^f_P(\eta)=\{\Theta\in\mathscr{A}_P(\eta)|F_\Theta=d\Theta=0\}$. We assume that the connection η over ∂X is flat and extends to flat connections over X. Otherwise, the spaces $\mathscr{A}^f_P(\eta)$ are empty. Let us fix an arbitrary flat connection $\Theta_P\in\mathscr{A}^f_P(\eta)$. Then any connection $\Theta\in\mathscr{A}_P(\eta)$ can be expressed as $\Theta=\Theta_P+2\pi iA$, for some $A\in\Omega^1_{1,\tan}(X;\mathbb{R})$, and we have

$$S_{X,P}(s_P,\Theta)=S_{X,P}(s_P,\Theta_P)-(A,*dA)\quad(\text{mod }1).$$

The deRham theory for manifolds with boundary gives the identification $H^1(X,\partial X;\mathbb{R})=(\text{Ker } d|_{\Omega^1_{1,\tan}(X;\mathbb{R})})/(\text{Im } d|_{\Omega^0_{\tan}(X;\mathbb{R})})$. Thus, the quotient space $\mathscr{M}_P(\eta)=\mathscr{A}^f_P(\eta)/\mathscr{G}_P(e)$ is isomorphic to the torus $H^1(X,\partial X;\mathbb{R})/H^1(X,\partial X;\mathbb{Z})$. The inner product determined on the space $H^1(X,\partial X;\mathbb{R})$ through the isomorphism with $\mathscr{H}^1_{\tan}(X)$ defines a measure ν_t on $\mathscr{M}_P(\eta)$.

As in the closed manifold case, we mimic the finite-dimensional model of Sec. II and *formally* define the partition function of the quadratic functional $S_{X,P}(s_P,\cdot)$ as

$$Z_{X,p}(\eta)=\frac{1}{\text{Vol }\mathscr{G}_P(e)}\int_{\mathscr{A}_P(\eta)}e^{\pi ikS_{X,P}(s_P,\Theta)}\hat{\mu},\tag{IV.2}$$

$$=\int_{\mathscr{A}_P(\eta)/\mathscr{G}_P(e)}e^{\pi ikS_{X,P}(s_P,\Theta)}[\det'(\tau^*_\Theta\tau_\Theta)]^{1/2}\mu,\tag{IV.2$'$}$$

$$=e^{\pi ikS_{X,P}(s_P,\Theta_P)}\int_{\mathscr{M}_P(\eta)}e^{(\pi i/4)\text{sgn}(-*d)}\frac{[\det'(\tau^*_\Theta\tau_\Theta)]^{1/2}}{[\det'(-k*d)]^{1/2}}\,\nu_t.\tag{IV.2$''$}$$

The expression (IV.2$'$) will be taken as the formal definition for $Z_{X,p}(\eta)$ in (IV.1). Taking the regularized determinants and signatures of the operators therein, the expression (IV.2$''$) has rigorous mathematical meaning.

To evaluate $\det'(-k*d)$ we use the result (A3). Let $S=k*d$ acting on the space $\Omega^1_{\tan}(X;\mathbb{R})$ and

$$\widetilde{S}=\begin{pmatrix}0 & S\\ S^* & 0\end{pmatrix}=\begin{pmatrix}0 & -k*d\\ -k*d & 0\end{pmatrix}$$

acting on $\Omega^1_{\tan}(X;\mathbb{R})\oplus\Omega^1_{\text{norm}}(X;\mathbb{R})$. We also let $\widetilde{T}$ be the operator

$$\widetilde{T}=\begin{pmatrix}kd & 0\\ 0 & kd\end{pmatrix}$$

on the space $\Omega^0_{\tan}(X;\mathbb{R})\oplus\Omega^0_{\text{norm}}(X;\mathbb{R})$. Then we find that $\widetilde{S}^2+\widetilde{T}\widetilde{T}^*=k^2(\Delta_1^{\tan}\oplus\Delta_1^{\text{norm}})$ and $\widetilde{T}^*\widetilde{T}=k^2(\Delta_0^{\tan}\oplus\Delta_0^{\text{norm}})$. According to the definition (A2), we have $\det' S=|\det'(S^*S)|^{1/2}=|\det\widetilde{S}|^{1/2}$. Using this and the relation (A3), we can write

$$\det'(-k*d)=|\det'\ \widetilde{S}|^{1/2}=\frac{|\det'(\widetilde{S}^2+\widetilde{T}\widetilde{T}^*)|^{1/4}}{|\det'(\widetilde{T}^*\widetilde{T})|^{1/4}}=\frac{|\det'(k^2\Delta_1^{\tan})\det'(k^2\Delta_1^{\mathrm{norm}})|^{1/4}}{|\det(k^2\Delta_0^{\tan})\det'(k^2\Delta_0^{\mathrm{norm}})|^{1/4}}$$

$$=\frac{k^{-(1/2)[\dim H^1(X;\mathbb{R})+\dim H^1(X,\partial X;\mathbb{R})]}}{k^{-(1/2)[\dim H^0(X;\mathbb{R})+\dim H^0(X,\partial X;\mathbb{R})]}}\frac{|\det'\ \Delta_1^{\tan}\ \det'\ \Delta_1^{\mathrm{norm}}|^{1/4}}{|\det'\ \Delta_0^{\tan}\ \det'\ \Delta_0^{\mathrm{norm}}|^{1/4}}.$$

Inserting the above result together with $\det'(\tau_\Theta^*\tau_\Theta)=\det'\ \Delta_0^{\tan}$ into (IV.2″), we obtain

$$Z_{X,p}(\eta)=k^{m_X}e^{\pi ikS_{X,p}(s_P,\Theta_P)}e^{(\pi i/4)\eta(-*d)}\int_{\mathscr{M}_P(\eta)}\frac{|\det'\ \Delta_0^{\tan}|^{5/8}|\det'\ \Delta_0^{\mathrm{norm}}|^{1/8}}{|\det'\ \Delta_1^{\tan}|^{1/8}|\det'\ \Delta_1^{\mathrm{norm}}|^{1/8}}\ \nu_t,\quad \text{(IV.3)}$$

where $m_X=(1/4)(\dim H^1(X;\mathbb{R})+\dim H^1(X,\partial X;\mathbb{R})-\dim H^0(X;\mathbb{R})-\dim H^0(X,\partial X;\mathbb{R}))$. We are going to relate the term inside the integral sign in (IV.3) to the Ray–Singer analytic torsion of X.

The same definition (III.11) of the analytic torsion as a density T_X^a on $|\mathrm{Det}\ H^\cdot(X;\mathbb{R})|$ applies when X is a Riemannian manifold with boundary.[8,13] In this case the zeta function ζ_q in (III.11) is that of the Laplace operator Δ_q^{norm} on $\Omega^q_{\mathrm{norm}}(X;\mathbb{R})$ and the density $\delta_{|\mathrm{Det}\ H^\cdot(X;\mathbb{R})|}$ on $|\mathrm{Det}\ H^\cdot(X;\mathbb{R})|$ corresponds to the natural inner product on $H^\cdot(X;\mathbb{R})\cong\mathscr{H}^\cdot_{\mathrm{norm}}(X)$. Thus, the analytic torsion of the compact connected 3-manifold X with nonempty boundary ∂X is the density,

$$T_X^a\in|\mathrm{Det}\ H^0(X;\mathbb{R})|\otimes|\mathrm{Det}\ H^1(X;\mathbb{R})^*|\otimes|\mathrm{Det}\ H^2(X;\mathbb{R})|$$

$$T_X^a=[(\det'\ \Delta_0^{\mathrm{norm}})^0(\det'\ \Delta_1^{\mathrm{norm}})^1(\det'\ \Delta_2^{\mathrm{norm}})^{-2}(\det'\ \Delta_3^{\mathrm{norm}})^3]^{1/2}|\nu^{(0)}_{\mathrm{norm}}|\otimes|\nu^{(1)}_{\mathrm{norm}}|^{-1}\otimes|\nu^{(2)}_{\mathrm{norm}}|,$$

where $\nu^{(q)}_{\mathrm{norm}}=\nu_1^{(q)}\wedge\cdots\wedge\nu_{b_q}^{(q)}$, with $b_q=\dim\mathscr{H}^q_{\mathrm{norm}}(X)$ and $\nu_1^{(q)},\ldots,\nu_{b_q}^{(q)}$ an orthonormal basis for $\mathscr{H}^q_{\mathrm{norm}}(X)$. An orthonormal basis $\nu^{(0)}_{\mathrm{norm}}$ of $\mathscr{H}^0(X)\cong\mathbb{R}$ is such that $|\nu^{(0)}_{\mathrm{norm}}|=[\mathrm{Vol}\ X]^{-1/2}$. Poincaré duality gives $H^q(X;\mathbb{R})\cong H^{3-q}(X,\partial X;\mathbb{R})^*$. The spaces $\Omega^q_{\mathrm{norm}}(X;\mathbb{R})$ and $\Omega^{3-q}_{\tan}(X;\mathbb{R})$ are isomorphic under the Hodge $*$ operator and $\det'\ \Delta_q^{\mathrm{norm}}=\det'\ \Delta_{3-q}^{\tan}$. In consequence, the square root of the analytic torsion can be regarded as a half-density $(T_X^a)^{1/2}\in|\mathrm{Det}\ H^1(X;\mathbb{R})^*|^{1/2}\otimes|\mathrm{Det}\ H^1(X,\partial X;\mathbb{R})^*|^{1/2}$, which is given by the expression

$$(T_X^a)^{1/2}=\frac{1}{[\mathrm{Vol}\ X]^{1/4}}\frac{[\det'\ \Delta_0^{\tan}]^{3/4}[\det'\ \Delta_1^{\mathrm{norm}}]^{1/4}}{[\det'\ \Delta_1^{\tan}]^{1/2}}(\nu_n)^{1/2}\otimes(\nu_t)^{1/2}.\quad \text{(IV.4)}$$

Here $\nu_n=:|\nu^{(1)}_{\mathrm{norm}}|^{-1}=|\nu_1^{(1)}\wedge\cdots\wedge\nu_{b_1}^{(1)}|^{-1}$, for any orthonormal basis $\nu_1^{(1)},\ldots,\nu_{b_1}^{(1)}$ of harmonic forms of $\mathscr{H}^1_{\mathrm{norm}}(X)$. Similarly, $\nu_t=:|\nu^{(1)}_{\tan}|^{-1}=|\widetilde{\nu}_1^{(1)}\wedge\cdots\wedge\widetilde{\nu}_{l_1}^{(1)}|^{-1}$, for any orthonormal basis $\widetilde{\nu}_1^{(1)},\ldots,\widetilde{\nu}_{l_1}^{(1)}$ of $\mathscr{H}^1_{\tan}(X)$, where $l_1=\dim\mathscr{H}^1_{\tan}(X)$. Having in view the isomorphism $\mathscr{M}_P(\eta)\cong H^1(X,\partial X;\mathbb{R})/H^1(X,\partial X;\mathbb{Z})$, we note that

$$\int_{\mathscr{M}_P(\eta)}(T_X^a)^{1/2}=:\left[\int_{\mathscr{M}_P(\eta)}\nu_t\right](T_X^a)^{1/2}\otimes(\nu_t)^{-1}\quad \text{(IV.5)}$$

is a half-density in $|\mathrm{Det}\ H^1(X;\mathbb{R})^*|^{1/2}\otimes|\mathrm{Det}\ H^1(X,\partial X;\mathbb{R})|^{1/2}$.

As in the case of a closed manifold, it was proven in Ref. 13 that the analytic torsion T_X^a of a manifold X with a boundary is independent of the metric chosen on X (the metric is assumed to be a direct product metric near ∂X). On the other hand, the partition function $Z_{X,p}(\eta)$ in (IV.3) is a metric-dependent complex number. In light of the previous observations on the analytic torsion, we can get rid of this metric dependence by multiplying $Z_{X,p}(\eta)$ by the half-density,

$$\Upsilon\in|\mathrm{Det}\ H^1(X;\mathbb{R})^*|^{1/2}\otimes|\mathrm{Det}\ H^1(X,\partial X;\mathbb{R})|^{1/2},\quad \text{(IV.6)}$$

$$\Upsilon = 2^{-\chi(\partial X)/2} e^{-(\pi i/4)\eta(-*d)} \frac{1}{[\mathrm{Vol}\ X]^{1/4}} \frac{[\det'\ \Delta_0^{\tan}]^{1/8}[\det'\ \Delta_0^{\mathrm{norm}}]^{-1/8}}{[\det'\ \Delta_1^{\tan}]^{3/8}[\det'\ \Delta_1^{\mathrm{norm}}]^{-3/8}} (\nu_n)^{1/2} \otimes (\nu_t)^{-1/2},$$

where $\chi(\partial X)$ is the Euler characteristic of ∂X. Thus, with the definition (IV.5) in view, we obtain

$$Z_{X,p}(\eta)\Upsilon = k^{m_X} e^{\pi i k S_{X,P}(s_P,\Theta_P)} \int_{\mathscr{M}_P(\eta)} 2^{-x(\partial X)/2} (T_X^a)^{1/2},$$

which is independent of the metric on X. It would be interesting to find a geometrical interpretation of the half-density factor Υ, if any exists.

Hence, to a compact connected 3-manifold X with boundary ∂X, we can associate through the path integral approach outlined above a half-density $Z_X(\eta)$ for every flat connection η over ∂X extending to flat connections over X:

$$Z_X(\eta) \in |\mathrm{Det}\ H^1(X;\mathbb{R})^*|^{1/2} \otimes |\mathrm{Det}\ H^1(X,\partial X;\mathbb{R})|^{1/2}, \tag{IV.7}$$

$$Z_X(\eta) = \sum_{p \in \mathrm{Tors}\ H^2(X;\mathbb{Z})} Z_{X,p}(\eta)\Upsilon = k^{m_X} \sum_{p \in \mathrm{Tors}\ H^2(X;\mathbb{Z})} e^{\pi i k S_{X,P}(s_P,\Theta_P)} \int_{\mathscr{M}_P(\eta)} 2^{-x(\partial X)/2} (T_X^a)^{1/2}.$$

The group $H^1(X,\partial X;\mathbb{T})$ has $\pi_0(H^1(X,\partial X;\mathbb{T})) \cong \mathrm{Tors}\ H^2(X,\partial X;\mathbb{Z}) \cong \mathrm{Tors}\ H^2(X;\mathbb{Z})$ and each connected component is isomorphic to the torus $H^1(X,\partial X;\mathbb{R})/H^1(X,\partial X;\mathbb{Z})$. Thus

$$\int_{\mathscr{M}_P(\eta)} (T_X^a)^{1/2} = \int_{H^1(X,\partial X;\mathbb{R})/H^1(X,\partial X;\mathbb{Z})} (T_X^a)^{1/2} = \frac{1}{[\#\pi_0(H^1(X,\partial X;\mathbb{T}))]} \int_{H^1(X,\partial X;\mathbb{T})} (T_X^a)^{1/2},$$

for all P, and we can write (IV.7) as

$$Z_X(\eta) = \frac{k^{m_X}}{[\#\mathrm{Tors}\ H^2(X;\mathbb{Z})]} \sum_{p \in \mathrm{Tors}\ H^2(X;\mathbb{Z})} e^{\pi i k S_{X,P}(s_P,\Theta_P)} \int_{H^1(X,\partial X;\mathbb{T})} 2^{-x(\partial X)/2} (T_X^a)^{1/2}. \tag{IV.8}$$

Ray and Singer conjectured[9] that for a closed manifold X, the analytic torsion norm T_X^a and the purely topological Reidemeister torsion norm T_X on $|\mathrm{Det}\ H^\bullet(X;\mathbb{R})|$ coincide,

$$T_X^a = T_X. \tag{IV.9}$$

A proof of the Ray–Singer conjecture and a generalization to the case of manifolds with boundary can be found in Ref. 13. It is proven that, for a compact manifold X with boundary ∂X, the two norms on $|\mathrm{Det}\ H^\bullet(X;\mathbb{R})|$, the analytic torsion norm T_X^a and the Reidemeister torsion norm T_X are related by

$$T_X^a = 2^{x(\partial X)/2} T_X. \tag{IV.10}$$

Referring to the results in (Ref. 2, Sec. V), we note that the term $\Sigma_p e^{\pi i k S_{X,P}(s_P,\Theta_P)}$ in (IV.8) gives rise to a section of the prequantum line bundle $\mathscr{L}_{\partial X}$ over the Lagrangian image $\Lambda_X \subset \mathscr{M}_{\partial X}$ of the moduli space of flat $\mathbb{T}$ connections $\mathscr{M}_X$ and that, in view of (IV.5), the half-density $\int_{H^1(X,\partial X;\mathbb{T})} (T_X^a)^{1/2}$ can be interpreted as a section of the bundle of half-densities on Λ_X. Hence, according to the definition of the quantum theory in (Ref. 2, Sec. V), the expression (IV.8) determines an element in the Hilbert space $\mathscr{H}(\partial X, L_X)$. In view of the identification (IV.10), this corresponds to the vector Z_X in $\mathscr{H}(\partial X, L_X)$ defined in (Ref. 2, Eq. V.14). It is interesting that the pure path integral (IV.3) and the half-density factor (IV.6) introduced to compensate for the metric dependence of (IV.3) combine together to produce a result that agrees with the one obtained in the

geometric quantization approach. For a closed connected oriented 3-manifold X also, the expression (III.16) for the topological invariant Z_X associated to X through the path integral approach coincides, in view of (IV.9), with the expression [Ref. 2, Eq. V.16] from the geometric quantization approach to the Abelian CS theory.

APPENDIX: REGULARIZATION OF DETERMINANTS AND SIGNATURES

• *Zeta regularization of determinants*. We follow Ref. 5. A non-negative self-adjoint operator B on a Hilbert space $\mathscr{H}$ is called regular if $\mathrm{Tr}[e^{-tB}-\Pi(B)]=\Sigma_k \alpha_k(B)t^k$ as $t\to+0$, where k runs over a finite set of non-negative numbers. $\Pi(B)$ denotes the projection operator onto the kernel of B. The zeta function $\zeta_B(s)$ of a regular operator B is defined for large Re(s) by the expression

$$\zeta_B(s)=\sum_{\lambda_j\neq 0}\lambda_j^{-s}=\frac{1}{\Gamma(s)}\int_0^\infty \mathrm{Tr}[\mathrm{e}^{-tB}-\Pi(B)]t^{s-1}\,dt,$$

where λ_j are the eigenvalues of B. The function $\zeta_B(s)$ admits a meromorphic continuation to $\mathbb{C}$ and is analytic at $s=0$. The regularized determinant $\det' B$ is defined by the expression

$$\det' B=\mathrm{e}^{-\zeta_B'(0)}. \tag{A1}$$

For an operator $B:\mathscr{H}_0\to\mathscr{H}_1$ between the Hilbert spaces $\mathscr{H}_0$ and $\mathscr{H}_1$, with adjoint operator B^* and such that B^*B is regular, the regularized determinant $\det' B$ is defined by

$$\det' B=\mathrm{e}^{-(1/2)\zeta'_{B^*B}(0)}=[\det'(B^*B)]^{1/2}. \tag{A2}$$

Let S be a self-adjoint operator an a Hilbert space $\mathscr{H}_1$. Let $T:\mathscr{H}_0\to\mathscr{H}_1$ be an operator between the Hilbert spaces $\mathscr{H}_0$ and $\mathscr{H}_1$ with adjoint $T^*:\mathscr{H}_1\to\mathscr{H}_0$ and such that $T(\mathscr{H}_0)\subset\mathrm{Ker}\,S$. The operators S^2 and T^*T are assumed regular. By definition $\det' S=(\det' S^2)^{1/2}$ and $\det' T=(\det' T^*T)^{1/2}$. Then we have[5]

$$\det' S=\frac{[\det'(S^2+TT^*)]^{1/2}}{[\det'(TT^*)]^{1/2}}=\frac{[\det'(S^2+TT^*)]^{1/2}}{[\det'(T^*T)]^{1/2}}. \tag{A3}$$

• *Regularization of signatures*. Let B be a self-adjoint operator on a Hilbert space and define the function

$$\eta_B(s)=\sum_{\lambda_j\neq 0}(\mathrm{sign}\,\lambda_j)|\lambda_j|^{-s},$$

with λ_j the eigenvalues of B. The function $\eta_B(s)$ can be meromorphically continued to $\mathbb{C}$ and has a removable singularity at $s=0$. The *eta invariant* of the operator B is defined as[4]

$$\eta(B)=\eta_B(0). \tag{A4}$$

[1] M. F. Atiyah, ''Topological quantum field theory,'' Publ. Math. Inst. Hautes Etudes Sci. Paris **68**, 175–186 (1989).
[2] M. Manoliu, ''Abelian Chern–Simons theory. I. A topological quantum field theory,'' J. Math. Phys. **39**, 170–206 (1998).
[3] E. Witten, ''Quantum field theory and the Jones polynomial,'' Commun. Math. Phys. **121**, 351–399 (1989).
[4] A. S. Schwarz, ''Instantons and fermions in the field of instanton,'' Commun. Math. Phys. **64**, 233–268 (1979).
[5] A. S. Schwarz, ''The partition function of a degenerate functional,'' Commun. Math. Phys. **67**, 1–16 (1979).
[6] M. F. Atiyah, *The Geometry and Physics of Knots* (Cambridge University Press, Cambridge, 1990).
[7] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, AMS Mathematical Surveys 14 (American Mathematical Society, Providence, RI, 1977).
[8] W. Müller, ''Analytic torsion and R-torsion of Riemannian manifolds,'' Adv. Math. **28**, 233–305 (1978).

[9] D. B. Ray and I. M. Singer, ''*R*-torsion and the Laplacian on Riemannian manifolds,'' Adv. Math. **7**, 145–210 (1971).
[10] D. B. Ray and I. M. Singer, ''Analytic torsion,'' *Proceedings of the Symposium on Pure Mathematics* 23, American Mathematical Society, 1973, pp. 167–182.
[11] D. S. Freed and R. Gompf, ''Computer calculations of Witten's 3-manifold invariants,'' Commun. Math. Phys. **141**, 79–117 (1991).
[12] D. H. Adams and S. Sen, ''Phase and scaling properties of determinants arising in topological field theories,'' Phys. Lett. B **353**, 495–500 (1995).
[13] S. M. Vishik, ''Generalized Ray–Singer conjecture. I. A manifold with smooth boundary,'' Commun. Math. Phys. **167**, 1–102 (1995).
[14] M. F. Atiyah, V. K. Patodi, and I. M. Singer, ''Spectral asymmetry and Riemannian geometry. I,'' Math. Proc. Camb. Philos. Soc. **77**, 43–69 (1975).

Infinite and finite Gleason's theorems and the logic of indeterminacy

Itamar Pitowsky[a)]
Department of Philosophy, The Hebrew University, Mount Scopus, Jerusalem 91905, Israel

(Received 20 June 1997; accepted for publication 3 September 1997)

In the first half of the paper I prove Gleason's lemma: *Every non-negative frame function on the set of rays in* $\Re^3$ *is continuous*. This is the central and most difficult part of Gleason's theorem. The proof is a reconstruction of Gleason's idea in terms of orthogonality graphs. The result is a demonstration that this theorem is actually combinatorial in nature. It depends only on a finite graph structure. In the second half of the paper I use the graph construction to obtain results about probability distributions (non-negative frame functions with weight one) on *finite* sets of rays. For example, given any two distinct nonorthogonal rays **a** and **b**, I construct a finite set of rays Γ that contains them, and has the following property: *No probability distribution on* Γ *assigns both* **a** *and* **b** *a truth value (probability zero or one) unless they are both false*. Thus the principle of indeterminacy turns into a theorem of propositional quantum logic (or partial Boolean algebras). © *1998 American Institute of Physics.* [S0022-2488(97)01612-5]

I. GLEASON'S THEOREM

Consider the n-dimensional real space $\Re^n$ equipped with the usual (Euclidean) inner product.

Definition 1: A real function f defined over the unit vectors in $\Re^n$ is called a frame function if the following two conditions hold:

(1) *For all unit vectors* $\mathbf{x}$ $f(\mathbf{x})=f(-\mathbf{x})$.
(2) *There is a constant* $C(f)$ *such that for every orthonormal basis* $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x}_n$ *in* $\Re^n$ *we have* $\Sigma_{k=1}^{n} f(\mathbf{x_k})=C(f)$. *The number* $C(f)$ *is called the weight of the frame function.*

The definition of a frame function can be extended to the n-dimensional complex space by modifying condition 1 to read: $f(e^{i\theta}\mathbf{x})=f(\mathbf{x})$ *for* $0<\theta<2\pi$, *and all unit vectors* $\mathbf{x}$. It can also be extended to include infinite-dimensional (real or complex) separable Hilbert spaces, if we allow the upper limit in the sum to be ∞. Gleason's theorem[1] provides a complete characterization of non-negative frame functions:

Theorem: *Let f be a non-negative frame function on a (real or complex) separable Hilbert space of dimension $n\geqslant 3$. Then there is a self-adjoint, non-negative definite operator W on that space such that $f(\mathbf{x})=\mathbf{x}^{\tau}W\mathbf{x}$ for all unit vectors* $\mathbf{x}$.

Gleason's theorem is a central result for the foundations of quantum mechanics. On the negative side it imposes very severe constraints on the structure of possible hidden variable theories.[2] On the positive side, it shows that the notion of ''state'' in quantum mechanics is determined by the algebraic structure of the closed subspaces of a Hilbert space. In other words, quantum ''logic'' *dictates* the probabilistic structure of quantum mechanics.

Gleason's original proof consists of three mathematically independent parts:

(1) Reduction: A proof that the general theorem follows from the special case of the three-dimensional real space $\Re^3$. This is a straightforward argument based on Hilbert space geometry.

a)Electronic mail: itamarp@vms.huji.ac.il

0022-2488/98/39(1)/218/11/$15.00
© 1998 American Institute of Physics

(2) A proof that every *continuous* non-negative frame function on $\mathfrak{R}^3$ satisfies the theorem. Here Gleason is using the continuity of the function to approximate it uniformly by linear combinations of spherical harmonics. Then, from the definition of a frame functions it follows that only functions of the desired kind could appear in the sum. This is the only truly ''analytic'' part of the original proof.
(3) Every non-negative (in fact, semibounded) frame function on $\mathfrak{R}^3$ is continuous. This part of the proof is at once the most difficult and the most elementary. Its proof requires only basic three-dimensional geometry and the definition of continuity.

The difficulty involved in the third stage of Gleason's proof and the application of ''higher mathematics'' in the second part, motivated the search for simpler demonstrations. Thus, for example, Piron[3] provided a proof of a considerably weaker version. He assumed that there exists a unit vector $\mathbf{p}$ such that $f(\mathbf{p})=C(f)=$the weight of the frame function f. Under this assumption Piron managed to avoid the ''higher mathematics'' of stage 2 and the difficulties of stage 3. Cooke *et al.*[4] succeeded in giving a full proof that avoids the difficulties in stage 2, while complicating (unnecessarily) stage 3. For example, to establish continuity, they use Tychonov's lemma, a highly nonconstructive principle.[5]

In the first part of this paper I shall prove Gleason's continuity theorem, namely: Every semibounded frame function in $\mathfrak{R}^3$ is continuous. The proof is a reconstruction of Gleason's original derivation, in terms of orthogonality graphs (which were introduced by Kochen and Specker). Thus, I shall first show (Sec. II) that certain orthogonality relations are realizable in $\mathfrak{R}^3$ (that is, certain graphs are embeddable). Moreover, we shall see that frame functions defined on the vectors *in those graphs* satisfy certain constraints. This part is purely geometrical. In Sec. III, I shall establish the continuity of the frame function by ''walking'' on the graphs, very much in the fashion of the Kochen and Specher approach. This will demonstrate that Gleason's continuity argument is not merely constructive; it is actually finitistic.[6] At its heart there are some simple combinatorial facts about finite (even if large) graphs. Using the results we can construct finite graphs on which certain finite versions of Gleason's theorem hold (Sec. IV). In particular, Theorem 6—the central result—shows the logical origin of the indeterminacy principle. This is explained in the last section.

II. TWO GEOMETRIC LEMMAS

The following is a straightforward extension of the notion of frame functions to *finite* sets of direction in space. We shall deal mainly with $\mathfrak{R}^3$, but the definition can be easily modified to accommodate real or complex spaces of any dimension.

Definition 2: Let Γ *be a finite set of unit vectors in the three-dimensional real space. A non-negative frame function on* Γ *is a function* $f:\Gamma\rightarrow\mathfrak{R}_t$ *that satisfies*

(1) $f(-\mathbf{x})=f(\mathbf{x})$ *whenever both* $\mathbf{x}$ *and* $-\mathbf{x}$ *are elements of* Γ.
(2) *There is a number* $C(f)$ *such that* $f(\mathbf{x})+f(\mathbf{y})\leqslant C(f)$ *for every orthogonal pair* $\mathbf{x},\mathbf{y}\in\Gamma$, *and* $f(\mathbf{x})+f(\mathbf{y})+f(\mathbf{z})=C(f)$ *for every orthogonal triple* $\mathbf{x},\mathbf{y},\mathbf{z}\in\Gamma$.

The proof of Gleason's theorem follows two lemmas on the embedding of graphs in the set of unit vectors in $\mathfrak{R}^3$. The graphs G_1, and G_2 below should read as follows: Each node in the graph stands for a unit vector. Two nodes in the graph are joined by an edge if, and only if, the corresponding vectors are orthogonal. I shall somewhat loosely identify each ray (that is, a one-dimensional subspace) in $\mathfrak{R}^3$ with a unit vector that spans it. The set of all unit vectors in $\mathfrak{R}^3$ is denoted by S^2. In the following all references to ''open sets'' and ''continuous functions'' correspond to the usual Euclidean topology on S^2.

The two lemmas below are geometric in nature. They follow from the metric structure of $\mathfrak{R}^3$ without any further assumption.

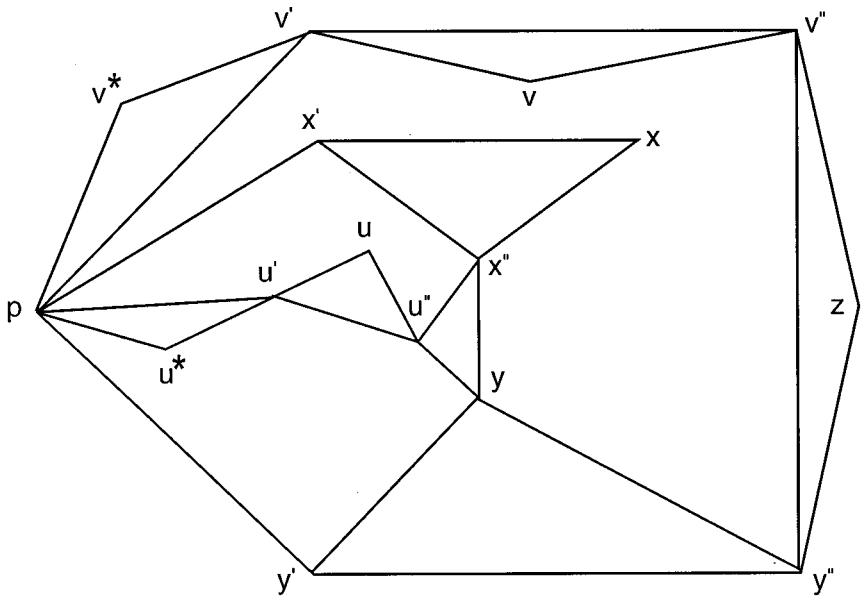

FIG. 1. The graph G_1.

Lemma 1: (a) Let $\mathbf{p}$ *and* $\mathbf{z}$ *be two unit vectors that satisfy* $\mathbf{p}\cdot\mathbf{z}\neq 0,1$. *Then there is a nonempty open set* $U\subset S^{(2)}$ *such that for each* $\mathbf{x}\in U$ *there are* $\mathbf{x}',\mathbf{x}'',\mathbf{y},\mathbf{y}',\mathbf{y}'',\mathbf{u},\mathbf{u}',\mathbf{u}'',\mathbf{v},\mathbf{v}',\mathbf{v}'',\mathbf{u}^*,\mathbf{v}^*\in S^2$, *which, together with* $\mathbf{p},\mathbf{z},\mathbf{x}$, *exhibit the orthogonality relations in the graph* G_1.

(b) Let g *be a non-negative frame function defined on the vectors in* G_1 *and satisfying*

$$g(\mathbf{x}')=g(\mathbf{u}^*),\quad g(\mathbf{y}')=g(\mathbf{v}^*). \tag{1}$$

Then

$$g(\mathbf{x})\leqslant g(\mathbf{z})+2g(\mathbf{p}). \tag{2}$$

Proof: (a) Let α be the angle between $\mathbf{p}$ and $\mathbf{z}$ and assume that $\cos\alpha>0$ (otherwise exchange $\mathbf{z}$ by $-\mathbf{z}$). Also, assume without loss of generality that $\mathbf{p}=(0,0,1)$ and $\mathbf{z}=(\cos\alpha,0,\sin\alpha)$. Consider the symmetric 3×3 matrix:

$$B=\begin{pmatrix} \sin\alpha & 0 & -\frac{1}{2}\cos\alpha \\ 0 & \sin\alpha & 0 \\ -\frac{1}{2}\cos\alpha & 0 & 0 \end{pmatrix}.$$

It has one negative eigenvalue, $\frac{1}{2}\sin\alpha-\frac{1}{2}$, corresponding to the eigenvector $\mathbf{x}_0=(\sqrt{\frac{1}{2}\sin\alpha-1/2},0,\sqrt{\frac{1}{2}\sin\alpha+\frac{1}{2}})=(\cos((\pi+2\alpha)/4),0,\sin((\pi+2\alpha)/4))$. Therefore, the quadric form $\mathbf{w}^{\mathrm{T}}B\mathbf{w}=(w_1^2+w_2^2)\sin\alpha-w_1w_3\cos\alpha$ is negative at $\mathbf{w}=\mathbf{x}_0$, and thus the set $U=\{\mathbf{w}\in S^{(2)};\mathbf{w}^{\mathrm{T}}B\mathbf{w}<\mathbf{0}\}$ is nonempty. Since the form $\mathbf{w}^{\mathrm{T}}B\mathbf{w}$ is continuous, U is opened.

Now let $\mathbf{x}\in U$, $\mathbf{x}=(x_1,x_2,x_3)$. In the following I shall not bother normalizing the vectors to have unit length. This makes no difference because we are interested in orthogonality relations only. Now put $\mathbf{x}'=(-x_2,x_1,0)$ and then $\mathbf{x}'\cdot\mathbf{p}=\mathbf{0}$. The vector $\mathbf{x}''=(x_1x_3,x_2x_3,-(x_1^2+x_2^2))$ is orthogonal to both $\mathbf{x}$ and $\mathbf{x}'$. By assumption we have $\mathbf{x}^{\mathrm{T}}B\mathbf{x}<\mathbf{0}$. Also, it is easy to see that $\mathbf{x}'^{\mathrm{T}}B\mathbf{x}'>\mathbf{0}$. Therefore, by the continuity of the quadric form $\mathbf{w}^{\mathrm{T}}B\mathbf{w}$, there is a vector $\mathbf{y}$

$=(y_1,y_2,y_3)\neq\mathbf{0}$ in the plane orthogonal to $\mathbf{x}''$, such that $\mathbf{y}^{\mathsf{T}}B\mathbf{y}=\mathbf{0}$. Put $\mathbf{y}'=(-y_2,y_1,0)$ and then $\mathbf{y}'\cdot\mathbf{p}=\mathbf{0}$. Thus, the vector $\mathbf{y}''$ which is orthogonal to both $\mathbf{y}$ and $\mathbf{y}'$ is given by $\mathbf{y}''=(y_1y_3,y_2y_3,-(y_1^2+y_2^2))$.

Now, $\mathbf{y}''\cdot\mathbf{z}=y_1y_3\cos\alpha-(y_1^2+y_2^2)\sin\alpha=-\mathbf{y}^{\mathsf{T}}B\mathbf{y}=\mathbf{0}$. To complete the proof we choose $\mathbf{v}''$ to be orthogonal to both $\mathbf{y}''$ and $\mathbf{z}$, the vector $\mathbf{v}'$ orthogonal to both $\mathbf{v}''$ and $\mathbf{p}$, and $\mathbf{v}$ orthogonal to $\mathbf{v}'$ and $\mathbf{v}''$. Similarly, we take $\mathbf{u}''$ to be orthogonal to $\mathbf{x}''$ and $\mathbf{y}$, the vector $\mathbf{u}'$ to be orthogonal to $\mathbf{u}''$ and $\mathbf{p}$ and $\mathbf{u}$ orthogonal to $\mathbf{u}'$ and $\mathbf{u}''$, the vector $\mathbf{u}^*$ orthogonal to $\mathbf{p}$ and $\mathbf{u}'$, and $\mathbf{v}^*$ orthogonal to $\mathbf{p}$ and $\mathbf{v}'$.

(b) Let g be a frame function on the vectors in G_1 that satisfy (1). We have

$$g(\mathbf{y})+g(\mathbf{y}')+g(\mathbf{y}'')=g(\mathbf{y}'')+g(\mathbf{z})+g(\mathbf{v}''),\tag{3}$$

$$g(\mathbf{x})+g(\mathbf{x}')+g(\mathbf{x}'')=g(\mathbf{x}'')+g(\mathbf{y})+g(\mathbf{u}'').\tag{4}$$

And since, in addition, g is non-negative, we have

$$g(\mathbf{v}')+g(\mathbf{v}'')\leqslant g(\mathbf{v})+g(\mathbf{v}')+g(\mathbf{v}'')=g(\mathbf{p})+g(\mathbf{v}')+g(\mathbf{v}^*),\tag{5}$$

$$g(\mathbf{u}')+g(\mathbf{u}'')\leqslant g(\mathbf{u})+g(\mathbf{u}')+g(\mathbf{u}'')=g(\mathbf{p})+g(\mathbf{u}')+g(\mathbf{u}^*).\tag{6}$$

Thus, we get $g(\mathbf{v}'')\leqslant g(\mathbf{p})+g(\mathbf{v}^*)$ and $g(\mathbf{u}'')\leqslant g(\mathbf{p})+g(\mathbf{u}^*)$. Substituting into (3) and (4) we obtain from (1)

$$g(\mathbf{x})=g(\mathbf{y})+g(\mathbf{u}'')-g(\mathbf{x}')\leqslant g(\mathbf{y})+g(\mathbf{p})+g(\mathbf{u}^*)-g(\mathbf{x}')=g(\mathbf{y})+g(\mathbf{p}),\tag{7}$$

$$g(\mathbf{y})=g(\mathbf{z})+g(\mathbf{v}'')-g(\mathbf{y}')\leqslant g(\mathbf{z})+g(\mathbf{p})+g(\mathbf{v}^*)-g(\mathbf{y}')=g(\mathbf{z})+g(\mathbf{p}).\tag{8}$$

Combining (7) and (8) we obtain $g(\mathbf{x})\leqslant g(\mathbf{z})+2g(\mathbf{p})$, as required. ■

Lemma 2: (a) Let $U\subseteq S^2$ be an open set. Then there is $\delta>0$ such that for all $\mathbf{z}_1,\mathbf{z}_2\in S^2$ that satisfy $\|\mathbf{z}_1-\mathbf{z}_2\|<\delta$ there are $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$,$\mathbf{x}_4$, $\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_7$, $\mathbf{x}_8\in U$ and $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$, $\mathbf{y}_4$, $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$, $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{w}_1$, $\mathbf{w}_2$, which together with $\mathbf{z}_1,\mathbf{z}_2$ exhibit the orthogonality relations in the graph G_2.

(b) Let f be any frame function on the vectors in G_2. Then

$$|f(\mathbf{z}_1)-f(\mathbf{z}_2)|\leqslant|f(\mathbf{x}_1)-f(\mathbf{x}_2)|+|f(\mathbf{x}_3)-f(\mathbf{x}_4)|+|f(\mathbf{x}_5)-f(\mathbf{x}_6)|+|f(\mathbf{x}_7)-f(\mathbf{x}_8)|.\tag{9}$$

Proof: (a) Let $\mathbf{x}_0\in U$ and let $\mathbf{z}_0,\mathbf{z}_0\neq\mathbf{x}_0$, be an arbitrary unit vector. Denote by $\mathbf{y}_0$ a unit vector orthogonal to both $\mathbf{x}_0$ and $\mathbf{z}_0$. Assume, without loss of generality, that $\mathbf{x}_0=(1,0,0)$ and that $\mathbf{y}_0=(0,1,0)$. Since $\mathbf{x}_0\in U$, there is $\theta>0$ such that the set of all vectors forming an angle smaller than θ with $\mathbf{x}_0$ is a subset of U. Let $\mathbf{w}_2=(-\sin(\frac{1}{2}\theta),\cos(\frac{1}{2}\theta),0)$. Then there is a neighborhood W of $\mathbf{y}_0$ with the following property: For $\mathbf{y}\in W$ let $\mathbf{x},\overline{\mathbf{x}}$ be unit vectors orthogonal to $\mathbf{y},\mathbf{w}_2$ respectively, *in the plan spanned by* $\mathbf{y}$ *and* $\mathbf{w}_2$. Then $\mathbf{x},\overline{\mathbf{x}}\in U$. (More precisely, the rays spanned by $\mathbf{x},\overline{\mathbf{x}}$ cut U, see Fig. 3.)

Similarly, there is $\phi>0$ such that the set of all unit vectors forming an angle ϕ with $\mathbf{y}_0$ is a subset of W. Let $\mathbf{w}_1$ be a vector in the plane spanned by $\mathbf{y}_0$ and $\mathbf{z}_0$ which forms an angle $\frac{1}{2}\phi$ with $\mathbf{z}_0$ and an angle $\frac{1}{2}(\pi+\phi)$ with $\mathbf{y}_0$. Then there is a neighborhood V of $\mathbf{z}_0$ with the analogous property: Let $\mathbf{z}\in V$ and $\mathbf{y},\overline{\mathbf{y}}$ be the vectors orthogonal to $\mathbf{z},\mathbf{w}_1$, respectively, in the $\mathbf{z}^{-}\mathbf{w}_1$ plane. Then the rays through $\mathbf{y},\overline{\mathbf{y}}$ cut V.

Now take any $\mathbf{z}_1,\mathbf{z}_2\in V$, and let $\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\mathbf{y}_4\in W$ be vectors such that $\mathbf{y}_1,\mathbf{y}_2$ are orthogonal to $\mathbf{z}_1,\mathbf{w}_1$, respectively, in the $\mathbf{z}_1^{-}\mathbf{w}_1$ plane, and $\mathbf{y}_3,\mathbf{y}_4$ are orthogonal to $\mathbf{w}_1,\mathbf{z}_2$, respectively, in the $\mathbf{w}_1^{-}\mathbf{z}_2$ plane while $\mathbf{v}_1,\mathbf{v}_2$ are the vectors orthogonal to the $\mathbf{z}_1^{-}\mathbf{w}_1$ and $\mathbf{w}_1^{-}\mathbf{z}_2$ planes, respectively.

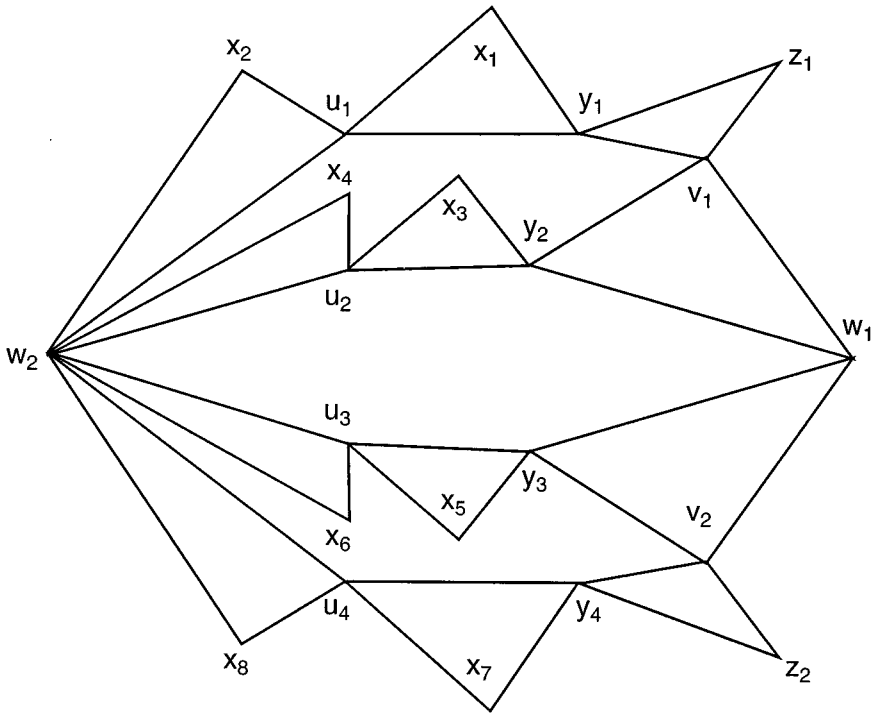

FIG. 2. The graph G_2.

In the same manner let $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_8 \in U$ be such that $\mathbf{x_{2i-1}},\mathbf{x_{2i}}$ are, respectively, orthogonal to $\mathbf{y}_i,\mathbf{w}_2$ in the $\mathbf{y_i}\bar{}\,\mathbf{w}_2$ plane, and $\mathbf{u_i}$ is a vector orthogonal to this plane ($i=1,2,3,4$). Since $\mathbf{z}_0$ is an arbitrary vector the proof of the first part is complete.

(b) Let f be a non-negative frame function on the vectors in G_2. Then

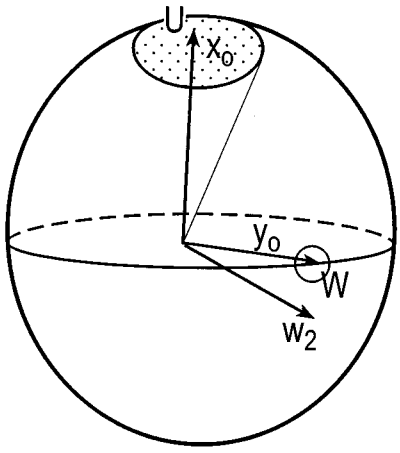

FIG. 3. The set W.

$$f(\mathbf{y}_1)+f(\mathbf{z}_1)+f(\mathbf{v}_1)=f(\mathbf{y}_2)+f(\mathbf{w}_1)+f(\mathbf{v}_1),$$
$$f(\mathbf{y}_4)+f(\mathbf{z}_2)+f(\mathbf{v}_2)=f(\mathbf{y}_3)+f(\mathbf{w}_1)+f(\mathbf{v}_2) \tag{10}$$

Subtracting the second equation (10) from the first, we get in a straightforward way

$$|f(\mathbf{z}_1)-f(\mathbf{z}_2)|\leqslant|f(\mathbf{y}_1)-f(\mathbf{y}_2)|+|f(\mathbf{y}_3)-f(\mathbf{y}_4)|. \tag{11}$$

In a similar manner we have

$$f(\mathbf{y_i})+f(\mathbf{x_{2i-1}})+f(\mathbf{u_i})=f(\mathbf{x_{2i}})+f(\mathbf{w}_2)+f(\mathbf{u_i}),\quad i=1,2,3,4, \tag{12}$$

from which it follows in the same way

$$|f(\mathbf{y}_1)-f(\mathbf{y}_2)|\leqslant|f(\mathbf{x}_1)-f(\mathbf{x}_2)|+|f(\mathbf{x}_3)-f(\mathbf{x}_4)|,$$
$$|f(\mathbf{y}_3)-f(\mathbf{y}_4)|\leqslant|f(\mathbf{x}_5)-f(\mathbf{x}_6)|+|f(\mathbf{x}_7)-f(\mathbf{x}_8)|. \tag{13}$$

Combining (11) and (13) we get inequality (9). ■

III. THE THEOREM ON CONTINUITY AND ITS COMBINATORIAL NATURE

We can now try to prove the following.

Theorem 3: *Let f be a semibounded frame-function in $\mathfrak{R}^3$. Then f is continuous.*

Proof: We can assume that f is bounded from below (otherwise, consider $-f$). By adding a constant to f it remains a frame function and we can make sure that $\inf f(x)=0$, where the infimum is taken over all unit vectors. Let $\varepsilon>0$, and choose a unit vector $\mathbf{p}$ such that $f(\mathbf{p})<\varepsilon$. Let σ be a 90° rotation around the $\mathbf{p}$ axis and define $g(\mathbf{x})=f(\mathbf{x})+f(\sigma\mathbf{x})$ for all rays $\mathbf{x}$. Then g is a frame function, $C(g)=2C(f)$. Let $m=\inf g(\mathbf{x})$, where the infimum is taken over all unit vectors $\mathbf{x}$ for which $\mathbf{x}\cdot\mathbf{p}\neq 0,1$. Choose such a vector $\mathbf{z}$ for which $g(\mathbf{z})<m+\varepsilon$. By Lemma 1, given any $\mathbf{p}$ and $\mathbf{z}$, $\mathbf{p}\cdot\mathbf{z}\neq 0,1$, there is an open set U such that for all $\mathbf{x}\in U$ the graph G_1 is realizable. Note that if $\mathbf{a}$ and $\mathbf{b}$ are any two vectors orthogonal to $\mathbf{p}$, then $g(\mathbf{a})=g(\mathbf{b})$, in particular (1) is satisfied. Therefore, by Lemma 1, also (2) is satisfied for all $\mathbf{x}\in U$. By the choice of $\mathbf{z}$ we have $g(\mathbf{x})\geqslant g(\mathbf{z})-\varepsilon$ for all $\mathbf{x}$ such that $\mathbf{x}\cdot\mathbf{z}\neq 0,1$. Therefore, combining this inequality with (2) we obtain for all $\mathbf{x},\overline{\mathbf{x}}\in U$: $g(\mathbf{x})-g(\overline{\mathbf{x}})\leqslant g(\mathbf{z})+2g(\mathbf{p})-g(\mathbf{z})+\varepsilon\leqslant 5\varepsilon$. Reversing the role of $\mathbf{x},\overline{\mathbf{x}}$ we get

$$|g(\mathbf{x})-g(\overline{\mathbf{x}})|<5\varepsilon,\quad \mathbf{x},\overline{\mathbf{x}}\in U. \tag{14}$$

Consider Lemma 2 applied to the set U and the frame function g. For any sufficiently small positive δ and all $\mathbf{z_1},\mathbf{z_2}$ such that $\|\mathbf{z}_1-\mathbf{z}_2\|<\delta$ there are $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_g}\in U$ which exhibit the orthogonality relations in the graph G_2. Apply (9) to $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_8}\in U$. We obtain from (14)

$$|g(\mathbf{z}_1)-g(\mathbf{z}_2)|<20\varepsilon,\quad \|\mathbf{z}_1-\mathbf{z}_2\|<\delta, \tag{15}$$

for all sufficiently close $\mathbf{z_2},\mathbf{z_2}$. In particular let $\mathbf{z_1}=\mathbf{p}$ and denote $U'=\{\mathbf{x};\|\mathbf{x}-\mathbf{p}\|<\delta\}$. Then all $\mathbf{x}\in U'$ satisfy $|g(\mathbf{p})-g(\mathbf{x})|<20\varepsilon$. From the definition of g we get for $\mathbf{x}\in U'$: $f(\mathbf{x})\leqslant g(\mathbf{x})\leqslant|g(\mathbf{p})-g(\mathbf{x})|+g(\mathbf{p})<20\varepsilon+2f(\mathbf{p})<22\varepsilon$. Therefore, for all $\mathbf{x_1},\mathbf{x_2}\in U'$

$$|f(\mathbf{x}_1)-f(\mathbf{x}_2)|<|f(\mathbf{x}_1)|+|f(\mathbf{x}_2)|<44\varepsilon. \tag{16}$$

Now, we take the open set U' and the frame function f and apply Lemma 2 again. We know that there is $\delta'>0$ such that for all $\mathbf{z_1},\mathbf{z_2},\|\mathbf{z_1}-\mathbf{z_2}\|<\delta'$, there are $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_8}\in U'$ which realize the graph G_2. Combining (13) and (16) we get

$$|f(\mathbf{z}_1)-f(\mathbf{z}_2)|<176\varepsilon,\quad \|\mathbf{z}_1-\mathbf{z}_2\|<\delta'. \tag{17}$$

And since $\varepsilon>0$ is arbitrary we conclude that f is continuous. ■

Despite the frequent references to ''open sets'' the above proof is, in fact, finite. Suppose that $f\geqslant 0$ is a frame function on $\Re^3$ and let $\varepsilon>0$ be a given real number. The first step is to choose vectors $\mathbf{p}$ and $\mathbf{z}$ as in the proof. This, of course, cannot be done explicitly, since f is not given explicitly. However, once $\mathbf{p}$ and $\mathbf{z}$ are fixed, the following is true:

(1) We can explicitly calculate the value of $\delta'>0$ such that if $\|\mathbf{z_1}-\mathbf{z_2}\|<\delta'$, then $|f(\mathbf{z_1})-f(\mathbf{z_2})|<176\varepsilon$.
(2) Given $\mathbf{z_1},\mathbf{z_2}$ such that $\|\mathbf{z_1}-\mathbf{z_2}\|<\delta'$ we can explicitly construct a finite (albeit large) graph and prove that $|f(\mathbf{z_1})-f(\mathbf{z_2})|<176\varepsilon$ just by ''walking'' on this graph.

As for the first point, recall that in Lemma 1 we obtained the open set $U=\{\mathbf{w}\in S^2;\mathbf{w}^{\mathrm{T}}B\mathbf{w}<0\}$ and the vector $\mathbf{x_0}=(\sqrt{\tfrac{1}{2}\sin\alpha-\tfrac{1}{2}},0,\sqrt{\tfrac{1}{2}\sin\alpha+\tfrac{1}{2}})\in U$, where $0<\alpha<1/2\pi$ is the angle between $\mathbf{p}$ and $\mathbf{z}$. From the definition of the quadric form $\mathbf{w}^{\mathrm{T}}B\mathbf{w}$ it follows that U contains the set of all unit vectors whose angle with $\mathbf{x_0}$ is less than θ, where θ is given by $\sin^2\theta=\tan^2(\pi/4-\alpha/2)$.

Now, consider Lemma 2 and the construction of the open sets W (Fig. 3) and V. A simple calculation shows that it is sufficient to take W as the set of all vectors whose angle with $\mathbf{y_0}$ is less than ϕ where $\tan^2\phi=\sin^2(\theta/2)\tan^2(\theta/2)$. This calculation repeats itself when we establish the relation between W and V. Thus V is the set of all unit vectors whose angle with $\mathbf{z_0}$ is less than ϕ' where again $\tan^2\phi'=\sin^2(\phi/2)\tan^2(\phi/2)$. In the proof of the theorem we use twice the construction of Lemma 2, first to show that if we choose $\mathbf{z_0}=\mathbf{p}$ (and therefore V to be a neighborhood of $\mathbf{p}$), then for all $\mathbf{z_1},\mathbf{z_2}\in V$ we have $|g(\mathbf{z_1})-g(\mathbf{z_2})|<20\varepsilon$. Thus we can take $\delta=2\sin\phi'$ in (15) and second we repeat the argument, this time with respect to f. Now V plays the role previously played by U. Hence, we can calculate the value of δ' in (17) in an analogous manner.

To establish the graph structure of the entire proof let us introduce some notations. Once $\mathbf{p}$, $\mathbf{z}$, and $\mathbf{x}\in U$ are chosen, the graph G_1 (Fig. 1) is determined (up to the sign of the vectors). Let us denote the graph G_1 determined by $\mathbf{p}$, $\mathbf{z}$, $\mathbf{x}$ by $G_1[\mathbf{p},\mathbf{z},\mathbf{x}]$. Similarly let $G_2[\mathbf{z_1},\mathbf{z_2},\mathbf{x_1},\mathbf{x_2},\mathbf{x_3},\mathbf{x_4},\mathbf{x_5},\mathbf{x_6},\mathbf{x_7},\mathbf{x_8}]$ denote the graph G_2 (Fig. 2) which contain these vectors. Now consider the graph G_3^*:

$$\begin{aligned}G_3^*[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2]=&G_2[\mathbf{z}_1,\mathbf{z}_2,\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_8]\cup G_2[\mathbf{p}_1,\mathbf{p},\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_8]\cup G_2[\mathbf{p}_2,\mathbf{p},\mathbf{x}_9,\mathbf{x}_{10},\ldots,\mathbf{x}_{16}]\\&\cup\cdots\cup G_2[\mathbf{p}_7,\mathbf{p},\mathbf{x}_{49},\mathbf{x}_{50},\ldots,\mathbf{x}_{56}]\cup G_2[\mathbf{p}_8,\mathbf{p},\mathbf{x}_{57},\mathbf{x}_{58},\ldots,\mathbf{x}_{64}]\cup G_1[\mathbf{x}_1,\mathbf{p},\mathbf{z}]\\&\cup G_1[\mathbf{x}_2,\mathbf{p},\mathbf{z}]\cup\cdots\cup G_1[\mathbf{x}_{64},\mathbf{p},\mathbf{z}].\end{aligned}\tag{18}$$

Now remember that σ is a 90° rotation around $\mathbf{p}$ and define G_3 by

$$G_3[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2]=G_3^*[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2]\cup\sigma G_3^*[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2]\cup\sigma^2G_3^*[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2]\cup\sigma^3G_3^*[\mathbf{p},\mathbf{z},\mathbf{z}_1,\mathbf{z}_2].\tag{19}$$

Take the graph $G_2[\mathbf{z_1},\mathbf{z_2},\mathbf{p_1},\mathbf{p_2},\mathbf{p_3},\mathbf{p_4},\mathbf{p_5},\mathbf{p_6},\mathbf{p_7},\mathbf{p_8}]$. Using (13) (with the $\mathbf{p_i}$'s instead of $\mathbf{x_i}$'s) we get

$$|f(\mathbf{z}_1)-f(\mathbf{z}_2)|\leqslant|f(\mathbf{p}_1)-f(\mathbf{p}_2)|+|f(\mathbf{p}_3)-f(\mathbf{p}_4)|+|f(\mathbf{p}_5)-f(\mathbf{p}_6)|+|f(\mathbf{p}_7)-f(\mathbf{p}_8)|.\tag{20}$$

Note that we also have

$$|f(\mathbf{p}_{2i-1})-f(\mathbf{p}_{2i})|\leqslant f(\mathbf{p}_{2i-1})+f(\mathbf{p}_{2i}),\quad i=1,2,3,4.\tag{21}$$

By construction, if $\mathbf{a}$ is a node of G_3, so is $\sigma\mathbf{a}$. Therefore, the function $g(\mathbf{w})=f(\mathbf{w})+f(\sigma\mathbf{w})$ is well defined on G_3. Take the graphs $G_2[\mathbf{p_i},\mathbf{p},\mathbf{x_{8i-7}},\mathbf{x_{8i-6}},\mathbf{x_{8i-5}},\mathbf{x_{8i-4}},\mathbf{x_{8i-3}},\mathbf{x_{8i-2}},\mathbf{x_{8i-1}},\mathbf{x_{8i}}]$, $i=1,2,\ldots,8$.

Using (13), applied to the frame function g and these graphs, we get

$$|g(\mathbf{p_i})-g(\mathbf{p})|\leqslant|g(\mathbf{x_{8i-7}})-g(\mathbf{x_{8i-6}})|+|g(\mathbf{x_{8i-5}})-g(\mathbf{x_{8i-4}})| \\ +|g(\mathbf{x_{8i-3}})-g(\mathbf{x_{8i-2}})|+|g(\mathbf{x_{8i-1}})-g(\mathbf{x_{8i}})| \tag{22}$$

for $i=1,2,...,8$. Now consider the graphs $G_1[\mathbf{p},\mathbf{z},\mathbf{x_j}]$ for $j=1,2,...,64$ and the function g. Using Eqs. (2) we conclude that $g(\mathbf{x_j})\leqslant g(\mathbf{z})+2g(\mathbf{p})$. Since, by assumption $g(\mathbf{x_j})\geqslant g(\mathbf{z})-\varepsilon$, we have

$$|g(\mathbf{x_j})-g(\mathbf{x_k})|\leqslant\varepsilon+2g(\mathbf{p}),\quad j,k=1,2,...,64. \tag{23}$$

Remembering that

$$f(\mathbf{p_i})\leqslant g(\mathbf{p_i})\leqslant|g(\mathbf{p_i})-g(\mathbf{p})|+2\,f(\mathbf{p}),\quad i=1,2,...,8, \tag{24}$$

we can combine all this to get $|f(\mathbf{z_1})-f(\mathbf{z_2})|<144f(\mathbf{p})+32\varepsilon<176\varepsilon$. Thus, the entire proof of this inequality depends only on vectors in the big graph G_3.

IV. FINITE GLEASON'S THEOREMS

It is natural to ask at this stage whether we can obtain results analogous to Gleason's theorem on smaller or even finite subsets of directions in $\mathfrak{R}^3$. The first type of finite Gleason's theorem is obtained from the infinite theorem via the compactness theorem of classical logic:

Theorem 4: *Let $A\subseteq\mathfrak{R}_t$ be a nonempty finite set of non-negative real numbers. Then there is a finite subset $\Gamma\subset S^2$ such that every frame function on Γ with range in A is constant.*

The proof is a straightforward application of the compactness theorem.[7] The compactness theorem itself is not constructive, so it does not provide a clue about the structure of the set Γ, or even about its size. Naturally, the most important case is $A=\{0,1\}$. In this case Kochen and Specker's theorem provides an actual construction of Γ.[8] The general constructive case is, to the best of my knowledge, open.

However, we can think of other finite versions of Gleason's theorem. To see that, consider first the following corollary of Gleason's theorem: *Let f be a non-negative frame function on $\mathfrak{R}^3$. If there are three linearly independent unit vectors* **p**, **q**, *and* **r**, *such that $f(\mathbf{p})=f(\mathbf{q})=f(\mathbf{r})=0$, then $f\equiv0$.* The proof follows easily from Gleason's theorem: If $\mathbf{p}^{\mathrm{T}}W\mathbf{p}=\mathbf{q}^{\mathrm{T}}W\mathbf{q}=\mathbf{r}^{\mathrm{T}}W\mathbf{r}=0$ for a basis **p**, **q**, **r**, of $\mathfrak{R}^3$, and a non-negative definite matrix W, then clearly $W\equiv0$. The following lemma is a finite version of this assertion.

Lemma 5: Let **p** *and* **z** *be two linearly independent vectors in $\mathfrak{R}^3$ and let σ be a 90° rotation around* **p**. *Then there is a finite set of unit vectors $\Gamma_1(\mathbf{p},\mathbf{z})\subset S^2$ such that $\mathbf{p},\mathbf{z},\sigma\mathbf{z}\in\Gamma_1(\mathbf{p},\mathbf{z})$, and every non-negative frame function f on $\Gamma_1(\mathbf{p},\mathbf{z})$ which satisfies $f(\mathbf{p})=f(\mathbf{z})=f(\sigma\mathbf{z})=0$ is identically zero.*

Proof: If **p** is orthogonal to **z**, then $\sigma\mathbf{z}$ is orthogonal to both, and then the theorem follows trivially for $\Gamma_1(\mathbf{p},\mathbf{z})=\{\mathbf{p},\mathbf{z},\sigma\mathbf{z}\}$. Therefore, assume that $\mathbf{p}\cdot\mathbf{z}\neq\mathbf{0}$. Consider the graph $G_3(\mathbf{p},\mathbf{z},\mathbf{z_1},\mathbf{z_2})$ in (18) and (19). In the last section we proved that there is a number $\delta>0$ such that if $\mathbf{z_1},\mathbf{z_2}\in S^2$ satisfy $\|\mathbf{z_1}-\mathbf{z_2}\|<\delta$, then we can construct a set of vectors that satisfy the orthogonality relation in G_4. Moreover, the value of δ can be explicitly calculated once **p** and **z** are given. Let f be a frame function defined on the set of vectors in the graph $G_3(\mathbf{p},\mathbf{z},\mathbf{z_1},\mathbf{z_2})$ and satisfies $f(\mathbf{p})$ $f(\mathbf{p})=f(\mathbf{z})=f(\sigma\mathbf{z})=\mathbf{0}$. Then, using inequalities (20)–(24) with $\varepsilon=0$ we conclude that $f(\mathbf{z_1})=f(\mathbf{z_2})$. Now, let **q** be a vector orthogonal to **p** in the plane spanned by **p** and **z**, and let **r** be a vector orthogonal to **p** in the plane spanned by **p** and $\sigma\mathbf{z}$. Hence, **q** and $\mathbf{r}=\sigma\mathbf{q}$ are also orthogonal to each other. Let $\mathbf{z_0}=\mathbf{p},\mathbf{z_1},\mathbf{z_2},\ldots,\mathbf{z_m}=\mathbf{q}$ be unit vectors such that $\|\mathbf{z_i}-\mathbf{z_{i-1}}\|<\delta$, $i=1,2,...,m$. Define the graph

$$G_4(\mathbf{p},\mathbf{z})=G_3(\mathbf{p},\mathbf{z},\mathbf{z_0},\mathbf{z_1})\cup G_3(\mathbf{p},\mathbf{z},\mathbf{z_1},\mathbf{z_2})\cup\cdots\cup G_3(\mathbf{p},\mathbf{z},\mathbf{z_{m-1}},\mathbf{z_m}). \tag{25}$$

and let $\Gamma_1(\mathbf{p},\mathbf{z})$ be the set of all vectors (nodes) in the graph $G_4(\mathbf{p},\mathbf{z})$. If f is a non-negative frame function on $\Gamma_1(\mathbf{p},\mathbf{z})$ which satisfies $f(\mathbf{p})=f(\mathbf{z})=f(\sigma\mathbf{z})=\mathbf{0}$, then $\mathbf{0}=f(\mathbf{p})=f(\mathbf{z_1})=f(\mathbf{z_2})=\cdots=f(\mathbf{z_{m-1}})=f(\mathbf{q})$ and $\mathbf{0}=f(\mathbf{p})=f(\sigma\mathbf{z_1})=f(\sigma\mathbf{z_2})=\cdots=f(\sigma\mathbf{z_{m-1}})=f(\mathbf{r})$. Since $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$ is an orthogonal triple the claim follows. ■

Note that the number of vectors in $G_4(\mathbf{p},\mathbf{z})$ depends on $\mathbf{p}$ and $\mathbf{z}$, since the number of copies of G_3 in (22) depends on the size of δ. It may be a large number, but the important thing is that it is finite and that G_4 is given by an explicit construction. It is possible that we can find a smaller graph that still does the job.

Let Γ be a nontrivial finite set of unit vectors in $\Re^3$. Call a non-negative frame function f on Γ *a state* if its weight is $C(f)=1$. Every non-negative frame function that is not identically zero can be normalized to be a state. Using Lemma 5 we can prove the following.

Theorem 6 (The Logical Indeterminacy Principle): *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two linearly independent nonorthogonal unit vectors in* $\Re^3$. *Then there is a finite set of unit vectors* $\Gamma(\mathbf{a},\mathbf{b})$ *such that* $\mathbf{a},\mathbf{b}\in\Gamma(\mathbf{a},\mathbf{b})$ *and for which the following statements hold: (a) There is no state* f *on* $\Gamma(\mathbf{a},\mathbf{b})$ *which satisfies* $f(\mathbf{a})=f(\mathbf{b})=1$. *(b) There is no state* f *on* $\Gamma(\mathbf{a},\mathbf{b})$ *which satisfies* $f(\mathbf{a})=1$ and $f(\mathbf{b})=0$ *(c) There is no state* f *on* $\Gamma(\mathbf{a},\mathbf{b})$ *which satisfies* $f(\mathbf{a})=0$ and $f(\mathbf{b})=1$.

Proof: We can assume that the angle between $\mathbf{a}$ and $\mathbf{b}$ is strictly between 0 and $\frac{1}{2}\pi$. (If the angle is greater than $\frac{1}{2}\pi$, substitute $-\mathbf{b}$ for $\mathbf{b}$). Let $\mathbf{p}$ be a unit vector orthogonal to $\mathbf{a}$ in the plane spanned by $\mathbf{a}$ and $\mathbf{b}$, and let $\mathbf{z}$ be a vector orthogonal to $\mathbf{b}$ in the same plane. Let σ be a 90° rotation around $\mathbf{p}$. Then $\sigma\mathbf{a}$ is orthogonal to both $\mathbf{a}$ and $\mathbf{p}$. Also, let ρ be a 90° rotation around $\mathbf{z}$. Then $\rho\mathbf{b}=\sigma\mathbf{a}$ is orthogonal to both $\mathbf{b}$ and $\mathbf{z}$. Now define

$$G_5(\mathbf{a},\mathbf{b})=G_4(\mathbf{p},\mathbf{z})\cup G_4(\mathbf{p},\mathbf{b})\cup G_4(\mathbf{z},\mathbf{a})\cup\text{the triangle}\{\mathbf{a},\mathbf{p},\sigma\mathbf{a}\}\cup\text{the triangle}\{\mathbf{b},\mathbf{z},\sigma\mathbf{a}\}$$
$$\cup\text{the triangle}\{\mathbf{a},\sigma\mathbf{b},\sigma\mathbf{z}\}\cup\text{the triangle}\{\mathbf{z},\mathbf{b},\rho\mathbf{b}\}\cup\text{the edge}\{\mathbf{b},\rho\mathbf{a}\}. \tag{26}$$

Let $\Gamma(\mathbf{a},\mathbf{b})$ be the set of all unit vectors in the graph $G_5(\mathbf{a},\mathbf{b})$. Suppose first that f is a state on $\Gamma(\mathbf{a},\mathbf{b})$ such that $f(\mathbf{a})=f(\mathbf{b})=1$. Since $\mathbf{p}$ is orthogonal to $\mathbf{a}$ we have $f(\mathbf{p})=0$. Since $\mathbf{z}$ is orthogonal to $\mathbf{b}$ we have $f(\mathbf{z})=0$. Also $\sigma\mathbf{z}$ is orthogonal to $\mathbf{a}$ (for $\sigma\mathbf{z}$ lies in the plane spanned by $\mathbf{p}$ and $\sigma\mathbf{a}$) and therefore $f(\sigma\mathbf{z})=0$. By Lemma 5 we get that f is identically zero on $\Gamma_1(\mathbf{p},\mathbf{z})$, a contradiction. Next, assume that $f(\mathbf{a})=1$ and $f(\mathbf{b})=0$. Since $\mathbf{p}$ is orthogonal to $\mathbf{a}$ we have $f(\mathbf{p})=0$. We assumed $f(\mathbf{b})=0$ and since $\sigma\mathbf{b}$ is also orthogonal to $\mathbf{a}$ we have $f(\sigma\mathbf{b})=0$. Again, by Lemma 5, f is identically zero on $\Gamma_1(\mathbf{p},\mathbf{b})$, a contradiction. Finally assume that $f(\mathbf{a})=0$ and that $f(\mathbf{b})=1$. Since $\mathbf{z}$ is orthogonal to $\mathbf{b}$ we have $f(z)=0$, we assumed $f(\mathbf{a})=0$, and since $\rho\mathbf{a}$ is also orthogonal to $\mathbf{b}$ (it lies in the plane spanned by $\mathbf{z}$ and $\rho\mathbf{b}$) we have $f(\rho\mathbf{a})=0$. Hence, by Lemma 5, f is zero on $\Gamma_1(z,\mathbf{a})$, a contradiction. ■

One can probably find realizable graphs that are simpler than G_5 and still do the job.[9] Note that (with the obvious extension of the concept of state) *the last theorem generalizes immediately to any (real or complex) Hilbert space H of finite dimension three or higher*. To see that, we simply embed $\Gamma(\mathbf{a},\mathbf{b})$ into a three-dimensional subspace $H_3\subseteq H$. Next, we take every orthogonal triple in $\Gamma(\mathbf{a},\mathbf{b})$ and add to it unit vectors from H to make it an orthonormal basis of H (we can add the same extra vectors to all orthogonal triples). Clearly all these extra vectors are orthogonal to H_3. Let $\Gamma^*(\mathbf{a},\mathbf{b})=\Gamma(\mathbf{a},\mathbf{b})\cup\{\text{the new vectors}\}$. Then $\Gamma^*(\mathbf{a},\mathbf{b})$ is finite whenever H has a finite dimension. If f is a state on $\Gamma^*(\mathbf{a},\mathbf{b})$ such that $f(\mathbf{a})=1$ or $f(\mathbf{b})=1$, then f is zero on the set of vectors orthogonal to H_3. Therefore, the restriction of such a state f to $\Gamma(\mathbf{a},\mathbf{b})$ is a state on $\Gamma(\mathbf{a},\mathbf{b})$ *as a subset of a three-dimensional space*. Thus, we can use the proof of Theorem 6.

V. THE LOGICAL ORIGINS OF THE INDETERMINACY PRINCIPLE

In the quantum logical approach[10] every unit vector in a Hilbert space (or more precisely, every one-dimensional subspace) corresponds to an atomic proposition that describes some physically possible situation. Two orthogonal vectors correspond to the description of two mutually

exclusive situations. But what about two linearly independent vectors which are not orthogonal? These correspond to ''complementary'' descriptions (to borrow Bohr's terminology).

Furthermore, in this approach quantum states correspond to probability distributions over the set of atomic propositions. A given quantum state may assign an atomic proposition probability one (''true''). Then the state is pure. A quantum state may also assign an atomic proposition probability zero (''false''). In any of these extreme cases of probability zero and one the proposition is *determined by the state*. For a physical system in a known quantum state we can find out in advance (without a measurement) which propositions are determined by the state, and what is their truth value. But there is a limit to the wealth of *a priori* knowledge that is coded into the quantum state. Consider the following:

The Indeterminacy Principle: Let **a** *and* **b** *be two distinct nonorthogonal unit vectors. Then any quantum state* f *must satisfy*

$$f(\mathbf{a}), f(\mathbf{b}) \in \{\mathbf{0}, 1\} \quad \text{only if} \quad f(\mathbf{a}) = f(\mathbf{b}) = \mathbf{0}. \tag{27}$$

In other words, *no quantum state can determine two ''complementary'' propositions unless they are both false*. In case the dimension of the space is three or higher this principle follows easily from Gleason's theorem. Furthermore, Theorem 6 shows that this is, in fact, a *logical* principle in the sense that *it is a theorem of propositional quantum logic*. To see that, let us break the proof of Theorem 6 into two parts:

(1) *Logical or combinatorial part:* We can take the graphs G_5 in (26) as a representation of abstract logical relations in classical propositional logic. In other words, the vertices represent formal atomic propositions, two propositions are joined by an edge if they are mutually exclusive, and three propositions form a triangle if they are inclusive. We can ask for the properties of probability distributions defined on a set of atomic propositions that exhibit these logical relations. Since all the inequalities and identities (1)–(13) and (20)–(24) depend only on the abstract graph structure (that is, only on orthogonality relations and not on other facts) we conclude the following.
Let f *be a real probability distribution defined on the propositions (vertices) of* G_5. *Then* $f(\mathbf{a}), f(\mathbf{b}) \in \{0,1\}$ *only if* $f(\mathbf{a}) = f(\mathbf{b}) = 0$. In this formulation **a** and **b** stand for abstract formal propositions and not their interpretation as vectors.
(2) *Geometric part:* The graph G_5 is realizable as an orthogonality graph for a set of unit vectors $\Gamma(\mathbf{a},\mathbf{b})$ in $\Re^3$.

Notice, however, that the content of the indeterminacy principle is already present in the purely combinatorial part 1. The conclusion (24) holds regardless of the content we assign to the atomic propositions.

Quantum mechanics provides an example (model) of such a set of atomic propositions. More precisely, if we take seriously the quantum-logical interpretation of subspaces and orthogonality, then we can actually construct a set of physically meaningful propositions which exhibit the logical relations in G_5.

Seen in this light Theorem 6 belongs to the logical algebraic approach to quantum theory. This approach attempts to construct as large a part of quantum theory as possible on the basis of logical principles.[11] I believe that the advantages in the present case are quite apparent. There is nothing mysterious about indeterminacy (or ''complementarity'' in the above sense). We can actually construct a finite (and classical) model of a set of atomic propositions, together with the logical relations among them, which ''force'' every probability distribution on them to satisfy (24). The only nonclassical aspect left is the identification of rays (or unit vectors) as propositions and of the orthogonality relation as mutual exclusion. The geometric fact that G_5 is embeddable does the rest.

More broadly, Theorem 6 is part of the attempt to understand the mathematical foundations of quantum mechanics. In particular, it helps to make the distinction between its physical content and mathematical artifact clear.[12]

ACKNOWLEDGMENT

This research was partially supported by the Edelstein Center for the History and Philosophy of Science at the Hebrew University.

[1] A. M. Gleason, J. Math. Mech. **6**, 885–894 (1957). The original proof is reproduced in V. S. Varadarajan, *Geometry of Quantum Mechanics*, (Springer-Verlag, Berlin, 1985), 2nd ed.

[2] J. S. Bell, Rev. Mod. Phys. **38**, 447–452 (1966). Reprinted in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge U.P., Cambridge, 1987). Gleason's theorem is also intimately related to the Kochen and Specker ''no hidden variable'' proof [S. Kochen and E. Specker, J. Math. Mech. **17**, 59–87 (1967)]. See also M. Redhead, *Incompleteness Nonlocality, and Realism* (Clarendon, Oxford, 1987). I shall discuss this connection in Secs. V and VI.

[3] C. Piron, *Foundations of Quantum Mechanics* (Addison–Wesley, Reading, MA, 1976).

[4] R. Cooke, M. Keane, and W. Moran, Math. Proc. Camb. Philos. Soc. **98**, 117–128 (1985).

[5] Thychonov's lemma asserts that an (infinite) product of compact topological spaces is compact in the product topology. The authors are taking a product of *a continuum* of spaces. Tychonov's lemma for Hausdorf spaces is only slightly weaker than the axiom of choice. See, for example, T. J. Jech, *The Axiom of Choice* (North Holland, Amsterdam, 1973).

[6] G. Hellman, [J. Philos. Logic **22**, 193–203 (1993)] demonstrated that Gleason's theorem as a whole is nonconstructive (cannot be proved by intuitionistic means alone). In order to execute the second stage in the proof we have to assume that every continuous function on $S^{(2)}$ attains its minimum. This fact cannot be proved constructively. Two points should be mentioned, however. First, there are a multitude of constructive reformulations of the minimum principle. Second, it might be possible to use the combinatorial method below to establish bound on the *modulus of continuity* of every frame function, say, prove that every frame function is Lipschitz. This will provide a fully constructive proof of Gleason's theorem as well as Theorem 4 in text.

[7] For the case where A has two elements see J. L. Bell, in *Perspectives on Quantum Reality*, edited by R. Clifton, The University of Western Ontario Series in Philosophy of Science (Kluwer, Dordrecht, 1996), pp. 227–243, in particular the corollary after Theorem 1. The same technique generalizes easily to any finite set A.

[8] In the original proof of Kochen and Specker (Ref. 2) the number of elements of Γ is 117. The present record, due to Kochen and Conway (unpublished), is 31 vectors. See A. Peres, J. Phys. A **24**, L175–L178 (1991); R. Clifton, Am. J. Phys. **61**, 443–447 (1993); J. Zimba and R. Penrose, Stud. Hist. Philos. Sci. **24**, 697–720 (1993); J. Bub, Found. Phys. **26**, 787–812 (1996); J. Bub, *Interpreting the Quantum World* (Cambridge, U.P., Cambridge, 1997).

[9] For *some* choices of **a** and **b** we can find small graphs that satisfy *some* of the relations in theorem 6. For example, Kochen and Specker's 10-element graph [figure 21 in work by Redhead (Ref. 2)] satisfies $f(1)=1\rightarrow f(2)>0$ *for every frame function*. Similarly the subgraph obtained from it by deleting vertices {2,3} satisfies $f(1)=1\rightarrow f(4)<1$. However, these graphs can be realized only for particular choices of **a** and **b**. Since we deal with frame functions (and not just zero–one truth values), it is not possible to simply ''chain'' these graphs and obtain the general result.

[10] The remarks below apply equally well to the two major versions of quantum logic: the older version of G. Birkhoff and J. von Neumann, Ann. Math. **37**, 823–843 (1936), which allows for logical operations among arbitrary propositions, and the more recent (partial Boolean algebra) version of Kochen and Specker (Ref. 2), which does not.

[11] This approach is most explicit in Varadarajan (Ref. 2). For the Parial Boolean Algebras perspective, see S. Kochen, in *Perspectives on Quantum Reality*, The University of Western Ontario Series in Philosophy of Science edited by R. Clifton (Kluwer, Dordrecht, 1996), pp. 237–243. This paper also contains a report on a remarkable result by Conway and Kochen.

[12] An analysis that clarifies this distinction in the context of Bell's inequalities is in I. Pitowsky, *Quantum Probability Quantum Logic*, Lecture Notes in Physics, Vol. 321 (Springer-Verlag, Berlin, 1989); Math. Program. **50**, 359–414 (1991); Br. J. Philos. Sci. **45**, 95–125 (1994).

On the energy spectrum of the hydrogen atom in a photon field. I

Reese T. Prosser[a)]
Dartmouth College, Hanover, New Hampshire 03755

(Received 14 May 1996; accepted for publication 24 September 1997)

In 1930 Weisskopf and Wigner gave an account, based on the Maxwell–Schrödinger equations, of the natural spectral line broadening of radiation emitted by a hydrogen atom. Their calculations were based on an approximation involving certain single-photon transitions in the perturbation series for the solutions of these equations. In Part I of this series of papers the exact expressions for both the line shift and the line broadening are obtained from the Maxwell–Dirac equations in such a way that the Weisskopf–Wigner results appear as a second order approximation. The Maxwell–Dirac Hamiltonian for the coupled fields is first shown to admit a complex analytic dilation in the energy variables. The Fredholm–Born series for the resolvent is shown to converge uniformly when certain high-energy cutoff factors are included in the interaction and the photons are given a small mass. The series is then rearranged to show that the spectrum of the modified dilated Hamiltonian, which consists of a complete set of complex eigenvalues, thresholds, and branch cuts, is only a slight perturbation of the known spectrum of the dilated Hamiltonian for the uncoupled fields. The real part of the shift of each complex eigenvalue then accounts for the spectral line shift, and the complex part accounts for the associated line broadening. Finally, the implications for the scattering matrix and the various phenomena of resonance scattering are discussed. In Part II of this series these results are shown to remain valid when the cutoff factors and the photon mass are removed. © *1998 American Institute of Physics.*
[S0022-2488(98)02201-4]

I. INTRODUCTION

The energy spectrum of the relativistic hydrogen atom, according to one of the best known calculations in all of quantum mechanics, consists of the bound state energy levels:[1]

$$\mu+\epsilon_n=\mu\left[1+\frac{e^2}{(\gamma+n_r)^2}\right]^{-1/2}<\mu, \tag{1.1}$$

together with the continuum levels:

$$\mu+\epsilon=(\mu^2+p^2)^{1/2}\geqslant\mu. \tag{1.2}$$

Here, μ and e are the electron mass and charge in atomic units in which $\hbar=c=1$, $\gamma^2=\kappa^2-e^2$, $n=n_r+|\kappa|$ is the principal quantum number, and (j,m,κ) are the angular quantum momentum numbers, of the electron.

These levels appear in each angular momentum sector as the discrete or continuum eigenvalues of the Hamiltonian operator $H_{\rm el}$ associated with the Dirac equation in the absence of radiation, and as singularities in ζ in the matrix elements $\langle g|R_{\rm el}(\zeta)|f\rangle$ of the associated resolvent operator

[a)]The author is deceased, and Part II of the planned series was never completed. The author also had no chance of reading the proofs.

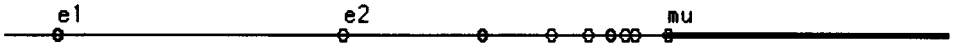

FIG. 1. Spectrum of the Hamiltonian $H_{\rm el}$ when $\theta=0$.

$R_{\rm el}(\zeta)=(H_{\rm el}-\zeta)^{-1}$. These singularities of the resolvent consist of poles at the discrete energy levels $\epsilon_n<\mu$ and a branch cut $[\mu,\infty)$ along the half-line of continuum energy levels ϵ (see Fig. 1).

If one restricts attention to a dense subset of energy states $|f\rangle$, $|g\rangle$ which admit a suitable analytic continuation $|f(\theta)\rangle$, $|g(\theta)\rangle$ into the open sector $S(\theta)=\{E:0\leqslant \arg E\leqslant\theta<\pi/2\}$ of the complex energy plane,[2–4] then the matrix elements $\langle g|H_{\rm el}|f\rangle$ made up from these states (the *analytic* matrix elements) will admit an analytic continuation into the same sector $S(\theta)$. It follows that on the Hilbert space $\mathfrak{h}_{\rm el}(\theta)$ of these analytic states the Hamiltonian $H_{\rm el}$ admits an *analytic dilation*, into the same sector $S(\theta)$. The spectrum of this analytic dilation $H_{\rm el}(\theta)$, acting on the Hilbert space $\mathfrak{h}_{\rm el}$ of analytic states $|f(\theta)\rangle$, is as shown in Fig. 2.

The presence of the radiation field changes this picture only slightly. If we adjoin a field of free photons with Hamiltonian $H_{\rm ph}$, then the spectrum of the (undilated) Hamiltonian $H_0=H_{\rm el}+H_{\rm ph}$ consists of the half-line $[\epsilon_1,\infty)$ with the eigenvalues $\epsilon_n>\epsilon_1$ embedded therein (see Fig. 3).

Moreover, the analytic matrix elements $\langle g|H_0|f\rangle$ of the Hamiltonian H_0 admit a similar analytic continuation into the same sector $S(\theta)$ of the complex energy plane. It follows that the Hamiltonian H_0 admits a similar analytic dilation $H_0(\theta)$ into the same sector $S(\theta)$, whose spectrum on the state space $\mathfrak{h}_0(\theta)$ is as shown in Fig. 4.

If we now introduce the interaction eH_1 between the electron and the radiation field, then it is generally believed that these energy levels are displaced slightly, down and to the left. Since the analytic matrix elements $\langle g|H_1|f\rangle$ of the interaction also admit an analytic continuation into the sector $S(E)$, it follows that the full Hamiltonian H also admits an analytic dilation $H(\theta)$ into this sector. The spectrum of the dilated Hamiltonian $H(\theta)$, acting on the state space $\mathfrak{h}_0(\theta)$, is thought to have the form shown in Fig. 5, in which the eigenvalues and continua are all displaced slightly to the left of, and (except for ϵ_1) slightly below their positions in Fig. 4. This displacement is a consequence of the interaction with the radiation field, and is known to be correct at least to second order in the interaction.[5,6]

It follows that the energy spectrum of the full Hamiltonian $H(\theta)$ of this system has precisely one real energy eigenvalue $\epsilon_1'<\epsilon_1$ in each angular momentum sector, associated with the ground state in that sector, and a half-line continuum $[\epsilon_1',\epsilon_1'+e^{-i\theta}\infty)$ with no embedded eigenvalues. The other discrete energy eigenvalues are now complex eigenvalues of the Hamiltonian, and they are associated with resonance scattering.

In this series of papers we propose to establish these results as rigorously as possible within the context of the quantized Maxwell radiation field interacting with the Dirac hydrogen atom in

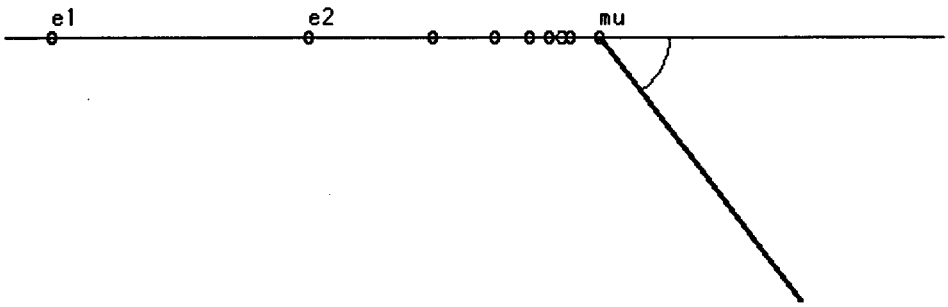

FIG. 2. Spectrum of the Hamiltonian $H_{\rm el}$ when $\theta=\pi/3$.

FIG. 3. Spectrum of the Hamiltonian H_0 when $\theta=0$.

the form described by Dirac.[6] We consider only positive-energy electron states, but we include the complete photon field, and contributions from all orders of the Born series. In order to achieve a convergent theory, however, we shall find in Part I that we must introduce cutoff factors into the interaction term at high (but not low) energies, and at large distances. Moreover, we find that we must give the photons an arbitrarily small mass $\nu>0$. The presence of a photon mass causes an essential change in the spectra of the Hamiltonian operators H_0 and H described here, in that each half-line in Figs. 4 and 5 splits into a simple eigenvalue and a bundle of half-lines, as shown in Fig. 6.

We shall see in Part II that all of the cutoff factors can be removed, and the photon mass can be removed, so that the final theory is a limiting case of a family of convergent theories for which all of these results are valid.

II. THE HAMILTONIAN

We take for the Dirac Hamiltonian $H_{\rm el}$ the form

$$H_{\rm el}=i\alpha\cdot\mathbf{grad}-\beta\mu-\frac{e^2}{r}, \tag{2.1}$$

with μ, e the electron mass and charge, and for the Maxwell Hamiltonian $H_{\rm ph}$ the form

$$H_{\rm ph}=\frac{1}{8\pi}\int[(\dot{\mathbf{A}})^2+(\mathbf{curl}\ \mathbf{A})^2+\nu^2(\mathbf{A})^2]d\mathbf{r}, \tag{2.2}$$

where ν is an arbitrarily small photon mass. As explained in Sec. VII below, we shall find it necessary to take $0<\nu\ll 1$ in order to ensure the convergence of the Fredholm–Born series. We denote by H_0 the Hamiltonian without interaction:

$$H_0=H_{\rm el}+H_{\rm ph}, \tag{2.3}$$

by eH_1 the interaction, with

$$H_1=\alpha\cdot\mathbf{A}, \tag{2.4}$$

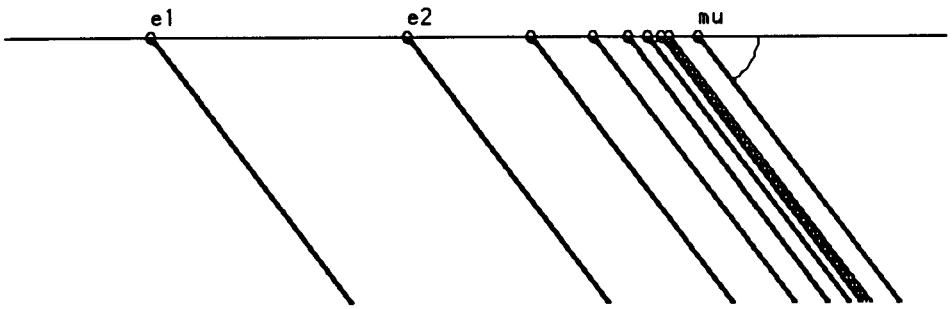

FIG. 4. Spectrum of the Hamiltonian H_0 when $\theta=\pi/3$.

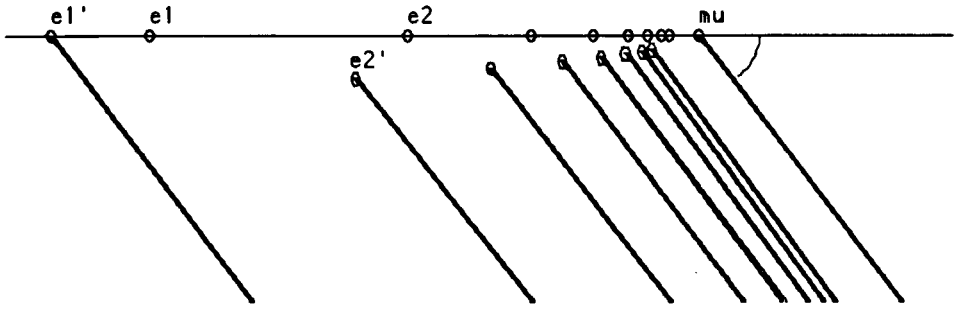

FIG. 5. Spectrum of the Hamiltonian H when $\theta=\pi/3$.

and by H the Hamiltonian with interaction:

$$H=H_0+eH_1, \tag{2.5}$$

with the coupling constant e eventually to be set equal to the electron charge e.

III. THE ELECTRON EIGENSTATES

The discrete eigenstates of the electron Hamiltonian $H_{\rm el}$ all have the form $|n,j,m,\kappa\rangle$, with n the principal quantum number and (j,m,κ) the angular quantum numbers. We shall find it convenient to suppress the dependence on (j,m,κ) and to parametrize the eigenstates by their energy ϵ_n as given by Eq. (1.1). Thus we write

$$|\epsilon_n\rangle=|n,j,m,\kappa\rangle \tag{3.1}$$

and

$$H_{\rm el}|\epsilon_n\rangle=(\mu+\epsilon_n)|\epsilon_n\rangle. \tag{3.2}$$

Similarly, the continuum eigenstates of $H_{\rm el}$ have the form $|p,j,m,\kappa\rangle$ with $p=|\mathbf{p}|$, $\mathbf{p}$ the electron momentum vector. We parametrize these states by their energy $\epsilon_p=(p^2+\mu^2)^{1/2}-\mu$ as given by Eq. (1.2), and write

$$|\epsilon\rangle=|\epsilon(p)\rangle=|p,j,m,\kappa\rangle \tag{3.3}$$

and

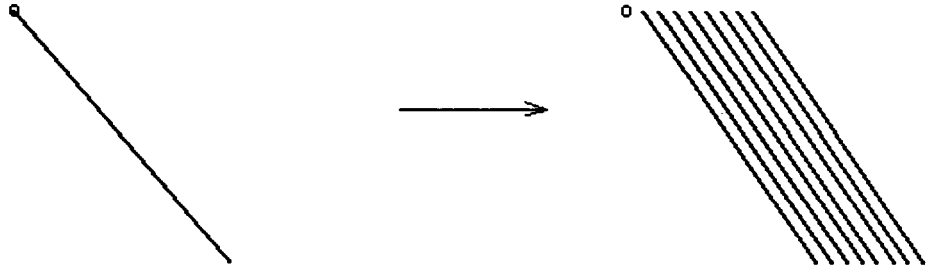

FIG. 6. Effect of replacing $\nu=0$ by $\nu>0$.

$$H_{\rm el}|\epsilon\rangle=(\mu+\epsilon)|\epsilon\rangle. \tag{3.4}$$

(For explicit descriptions of these states, see Ref. 1).

IV. THE PHOTON EIGENSTATES

The free photon states in the radiation field have the form $|k,j,m,\lambda\rangle$, with $k=|\mathbf{k}|$, $\mathbf{k}$ the photon momentum vector and (j,m,λ) the photon angular momentum quantum numbers. We parametrize these states by their energy $\omega_k=(k^2+\nu^2)^{1/2}-\nu$, and for these states we write

$$|\omega\rangle=|\omega(k)\rangle=|k,j,m,\lambda\rangle \tag{4.1}$$

and

$$H_{\rm ph}|\omega\rangle=(\nu+\omega)|\omega\rangle. \tag{4.2}$$

A typical eigenstate of the Hamiltonian $H_0=H_{\rm el}+H_{\rm ph}$, then, consists of the electron together with q photons ($q=0,1,2,...$), and has the form $|\epsilon,\omega_1,\omega_2,\ldots,\omega_q\rangle$, where ϵ takes on the discrete values ϵ_n when $\epsilon<0$ and the continuous values $\epsilon(p)$ when $\epsilon\geqslant 0$, and the ω_i take on the continuous values $\omega_i(k)$. For such states we have

$$H_0|\epsilon,\omega_1,\omega_2,\ldots,\omega_q\rangle=(\mu+q\nu+\epsilon+\omega_1+\cdots+\omega_q)|\epsilon,\omega_1,\omega_2,\ldots,\omega_q\rangle. \tag{4.3}$$

V. THE RESOLVENT $R_0(\zeta)$

Now let $R_0(\zeta)$ be the resolvent of H_0:

$$R_0(\zeta)=(H_0-\zeta)^{-1}. \tag{5.1}$$

Then for states of the form:

$$|f\rangle=\int\cdots\int f(\epsilon_n,\omega_1,\omega_2,\ldots,\omega_q)|\epsilon_n,\omega_1,\omega_2,\ldots,\omega_q\rangle d\omega_1\cdots d\omega_q, \tag{5.2}$$

we have for the matrix element $\langle g|R_0(\zeta)|f\rangle$ the expression

$$\langle g|R_0(\zeta)|f\rangle=\int\cdots\int\frac{\overline{g(\epsilon_n,\omega_1,\ldots,\omega_q)}f(\epsilon_n,\omega_1,\ldots,\omega_q)}{\mu+q\nu+\epsilon_n+\omega_1+\cdots+\omega_q-\zeta}\,d\omega_1\cdots d\omega_q. \tag{5.3}$$

It is clear that, as a function of ζ, this matrix element is analytic in the complex energy plane cut along the half-line $[\mu+\epsilon_n+q\nu,+\infty)$.

Now we choose the functions f and g in such a way that they admit analytic continuations separately in each ω_i to the half-plane $R1(\omega_i)>0$. We may take for these states the form

$$f(\epsilon_n,\omega_1,\ldots,\omega_q)=e^{-(\omega_1+\cdots+\omega_q)}p(\omega_1,\ldots,\omega_q), \tag{5.4}$$

where p is a polynomial in the ω_i. States of this form are known to be dense among all states of the form given in Eq. (5.2); we shall refer to them as *analytic states* (See Ref. 4 for details). If $|f\rangle$ and $|g\rangle$ are analytic states, then we can deform the paths of integration in Eq. (5.3) from $[0,+\infty)$ to $[0,+e^{-i\theta}\infty)$ for any θ in the interval $0<\theta<\pi/2$. Then we see that the matrix element $\langle f|R_0(\zeta)|g\rangle$ above is analytic in the complex ζ-plane cut along the half-lines $[\mu+\epsilon_n+q\nu,\mu+\epsilon_n+q\nu+e^{-i\theta}\infty)$ (see Fig. 4).

A similar argument holds for matrix elements $\langle g|R_0(\zeta)|f\rangle$ when f and g have the form

$$|f\rangle = \int \cdots \int f(\epsilon,\omega_1,\ldots,\omega_q)|\epsilon,\omega_1,\ldots,\omega_q\rangle d\epsilon\, d\omega_1\cdots d\omega_n, \tag{5.5}$$

with $\epsilon>0$, or when either one of the states f,g has the form of Eq. (5.2) and the other has the form of Eq. (5.5). Now if we choose f and g so that they admit analytic continuations separately in each ω_i to the half-plane $\mathrm{Re}(\omega_i)>0$ and in ϵ to the half-plane $\mathrm{Re}(\epsilon)>0$, then we can deform the integral paths in $\langle g|R_0(\zeta)|f\rangle$ as before, and see that the matrix element is analytic in the complex plane cut along the half-line $[\mu+q\nu,\mu+q\nu+e^{-i\theta}\infty)$, $0<\theta<\pi/2$ (see Fig. 4).

Linear combinations of such matrix elements, then, will be analytic in ζ in the complex plane cut by the half-lines $[\mu+q\nu+\epsilon_n,\mu+q\nu+\epsilon_n+e^{-i\theta}\infty)$ and $[\mu+q\nu,\mu+q\nu+e^{-i\theta}\infty)$. The points $\mu+\epsilon_n$ are poles of these matrix elements, while the points $\mu+\epsilon+q\nu$ for $q>0$ are logarithmic branch points.

VI. THE INTERACTION TERM

We write the interaction term H_1 in the angular momentum representation, using the analysis of Moses[7]

$$H_1=\alpha\cdot\mathbf{A}=\alpha\cdot(\mathbf{A}^+ + \mathbf{A}^-)=H_1^+ + H_1^-, \tag{6.1}$$

where $\mathbf{A}^+$ involves the creation of a new photon, and $\mathbf{A}^-$ the destruction of an old one. The matrix elements of $\mathbf{A}^+$ can be described as follows. If $|f\rangle$ and $|g\rangle$ are states of the form described in Eq. (5.5), with $\epsilon>0$, then

$$\langle g|\mathbf{A}^+|f\rangle=(q+1)^{1/2}\int \overline{g(\epsilon',\omega,\omega_1,\ldots,\omega_q)}F(\epsilon',\epsilon,\omega)f(\epsilon,\omega_1,\ldots,\omega_q)d\epsilon'\, d\epsilon\, d\omega\, d\omega_1\cdots d\omega_q. \tag{6.2}$$

Here, $F(\epsilon',\epsilon,\omega)$ is a function depending in a complicated way on the angular quantum numbers of the states $|\epsilon'\rangle$, $|\epsilon\rangle$ and $|\omega\rangle$, whose precise description is not important here (see Moses[7] for details). It is important, however, that the dependence of $F(\epsilon',\epsilon,\omega)$ on the energy variables can be expressed as a finite linear combination of integrals of the form

$$J(\epsilon',\epsilon,\omega)=\int_0^\infty f_1^\natural(ip'r)g_2^\natural(ipr)h^\natural(ikr)w(r)dr/r, \tag{6.3}$$

where $f^\natural$, $g^\natural$, and $h^\natural$ are the radial wave functions for the incoming and outgoing electron and the outgoing photon, which are given in terms of *confluent hypergeometric functions* $\Phi(a,c;z)$, and where w is a weight function inserted to improve the convergence of the integral. When $\epsilon>0$, the electron wave functions have the form

$$\begin{aligned} f^\natural(ipr)&=rf(r)=\frac{\Gamma(\gamma+1)}{\Gamma(2\gamma+1)}\left(\frac{\mu+\epsilon}{\epsilon}\right)^{1/2}(2ipr)^\gamma e^{ipr}\ \mathrm{Re}\ C_1\Phi(\gamma-i\nu,2\gamma+1,-2ipr),\\ g^\natural(ipr)&=rg(r)=\frac{\Gamma(\gamma+1)}{\Gamma(2\gamma+1)}\left(\frac{\mu-\epsilon}{\epsilon}\right)^{1/2}(2ipr)^\gamma e^{ipr}\ \mathrm{Im}\ C_1\Phi(\gamma-i\nu,2\gamma+1,-2ipr). \end{aligned} \tag{6.4}$$

When $\epsilon<0$, we can continue this form analytically in p from p to $\lambda=-ip$ and find

$$\begin{aligned} f^\natural(\lambda r)=rf(r)=+\frac{\Gamma(\gamma+1)}{\Gamma(2\gamma+1)}\left(\frac{\mu+\epsilon}{\epsilon}\right)^{1/2}(2\lambda)^{1/2}(2\lambda r)^\gamma e^{-\lambda r}[C_2\Phi(-n_r,2\gamma+1,2\lambda r)\\ +C_3\Phi(1-n_r,2\gamma+1,2\lambda r)], \end{aligned}$$

$$g^{\natural}(\lambda r)=rg(r)=-\frac{\Gamma(\gamma+1)}{\Gamma(2\gamma+1)}\left(\frac{\mu-\epsilon}{\epsilon}\right)^{1/2}(2\lambda)^{1/2}(2\lambda r)^{\gamma}e^{-\lambda r}[C_2\Phi(-n_r,2\gamma+1,2\lambda r) - C_3\Phi(1-n_r,2\gamma+1,2\lambda r)]. \tag{6.5}$$

The photon wave functions have the form

$$\begin{aligned} h^{\natural}(ikr)&=\frac{\Gamma(j+1)}{\Gamma(2j+2)}(2ikr)^{j+1}e^{ikr}C_4\Phi(j+1,2j+2,-2ikr), \quad \text{if } j+l_1+l_2=\text{odd},\\ &=\frac{\Gamma(j+1)}{\Gamma(2j+2)}(2ikr)^{j}e^{ikr}[C_5\Phi(j+1,2j+2,-2ikr)+C_6\Phi(j,2j+2,-2ikr)],\\ &\quad \text{if } j+l_1+l_2=\text{even}. \end{aligned} \tag{6.6}$$

Here and throughout, $j+l_1+l_2$ denote the usual angular quantum numbers, and the C_i denote normalization constants which normalize these wave functions to the energy scale.

The C_i are independent of the energy variables and bounded in the angular momentum variables. (See Ref. 1 for details.)

For the weight function $w(r)$ we assume the form

$$w(r)=e^{-\sigma\tau-\tau E}, \tag{6.7}$$

where

$$E=\mathrm{Re}(\epsilon'+\epsilon+\omega), \tag{6.8}$$

and $\sigma,\tau\ll 1$. Ultimately (in Part II), we shall want to take $\sigma,\tau\to 0$.

When the integration in J is carried out, the result can be expressed in terms of *Lauricella's hypergeometric function* $F_A^{(3)}$ (see Exton,[8] Chap. 2). When $j+l_1+l_2$ is odd, this expression takes the form

$$F^{+}(\epsilon_{n_2},\epsilon_{n_1},\omega)=\frac{(4\lambda_1\lambda_2)^{1/2}}{\Gamma(a)}\frac{\Gamma(j+1)}{\Gamma(2j+2)}x^{\gamma_1}y^{\gamma_2}z^{j+1}e^{-\tau E}C_7F_A^{(3)}(a;-n_{1r},-n_{2r},j+1;2\gamma_1+1,2\gamma_2+1,2j+2;x,y,z), \tag{6.9}$$

plus similar terms with the n_{ir} replaced with $n_{ir}-1$. Here, $a=\gamma_1+\gamma_2+j+1$, and the variables x,y,z are given by

$$x=\frac{2\lambda_1}{\sigma+\lambda_1+\lambda_2+ik}, \quad y=\frac{2\lambda_2}{\sigma+\lambda_1+\lambda_2+ik}, \quad z=\frac{2ik}{\sigma+\lambda_1+\lambda_2+ik}. \tag{6.10}$$

Note that $|x|<2$, $|y|<2$, $|z|<2$, $|x+y+z|<2$.

When $j+l_1+l_2$ is *even* this expression takes instead the form

$$F^+(\epsilon_{n_2},\epsilon_{n_1},\omega)=\frac{(4\lambda_1\lambda_2)^{1/2}}{\Gamma(a)}\frac{\Gamma(j+1)}{\Gamma(2j+2)}x^{\gamma_1}y^{\gamma_2}z^{j+1}e^{-\tau E}[C_8F_A^{(3)}(a;-n_{1r},-n_{2r},j+1;$$
$$2\gamma_1+1,2\gamma_2+1,2j+2;x,y,z)]+C_9F_A^{(3)}(a;-n_{1r},-n_{2r},j;2\gamma_1+1,2\gamma_2+1,$$
$$2j+2;x,y,z)], \tag{6.11}$$

plus similar terms with the n_{ir} replaced by $n_{ir}-1$. Here, $a=\gamma_1+\gamma_2+j$. From now on we consider in detail only the first of these cases; the second is entirely similar.

The Lauricella function $F_A^{(3)}$ is defined for $|x|+|y|+|z|<1$ by its power series:

$$F_A^{(3)}(a;b,b',b'';c,c',c'';x,y,z)=\sum_{m,n,p=0}^{\infty}\frac{(a)_{m+n+p}(b)_m(b')_n(b'')_p}{(c)_m(c')_n(c'')_p(1)_m(1)_n(1)_p}x^my^nz^p, \tag{6.12}$$

where

$$(a)_m=\frac{\Gamma(a+m)}{\Gamma(a)}. \tag{6.13}$$

In our case this can be rewritten as (see Exton,[7] Chap. 2)

$$F_A^3(a;-n_{1r},-n_{2r},j+1;2\gamma_1+1,2\gamma_2+1,2j+2;x,y,z)$$
$$=\sum_{m,n,p=0}^{\infty}\frac{(a)_{m+n+p}(-n_{1r})_m(-n_{2r})_n(j+1)_p}{(2\gamma_1+1)_m(2\gamma_2+1)_n(1)_m(1)_n(1)_p}x^my^nz^p, \tag{6.14}$$

with $a=\gamma_1+\gamma_2+j+1$. Here, the sums over m and n are finite, and the sum over p gives

$$F_A^{(3)}(a;-n_{1r},-n_{2r},j+1;2\gamma_1+1,2\gamma_2+1,2j+2;x,y,z)$$
$$=\sum_{m=0}^{n_{1r}}\sum_{n=0}^{n_{2r}}\frac{(a)_{m+n}(-n_{1r})_m(-n_{2r})_n}{(2\gamma_1+1)_m(2\gamma_2+1)_n(1)_m(1)_n}x^my^nF_A^{(1)}(a+m+n,j+1,2j$$
$$+2,z), \tag{6.15}$$

where $F_A^{(1)}(a,b,c,z)$ is *Gauss's hypergeometric function*. From *Euler's integral representation* for $F_A^{(1)}$,

$$F_A^{(1)}(a,b,c,z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1\frac{(t)^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}\,dt, \tag{6.16}$$

valid whenever $0<\mathrm{Re}(b)<\mathrm{Re}(c)$, we see that in our case the Gauss function $F_A^{(1)}$, and hence the Lauricella function $F_A^{(3)}$, as a function of z, admits an analytic continuation into the upper half z-plane. It can be shown (see Moses and Prosser[9] for details) that $F_A^{(1)}$ is bounded there by a constant times $|1-z|^{-1}$.

In our case we have

$$\tan(\pi/4-\phi/2)\leqslant|1-z|\leqslant 1 \tag{6.17}$$

when z is given by Eq. (6.10) with k replaced by $e^{-i\phi}k$. When $k^2=\omega^2+2\nu\omega$ and ω is replaced by $e^{-i\theta}\omega$, then $0\leqslant\phi<\theta<\pi/2$. It follows that the function $F(\epsilon_m,\epsilon_n,\omega)$, as a function of $\omega=(k^2+\nu^2)^{1/2}-\nu$, admits an analytic continuation into the sector

$$S(\omega)=\{e^{-i\theta}\omega:0\leqslant\theta\leqslant\theta_{\max}<\pi/2\}, \tag{6.18}$$

and is uniformly bounded there. In fact, since $|x|<2$, $|y|<2$, $|z|<2$, we have

$$|F^+(\epsilon_{n_2},\epsilon_{n_1},\omega)|\leqslant C_{10}\,\frac{(4\lambda_1\lambda_2)^{1/2}}{\tan(\pi/4-\theta/2)}\,\frac{(2)^a}{\Gamma(a)}\,\frac{\Gamma(j+1)}{\Gamma(2j+2)}\,(\lambda_1)^{\gamma_1}(\lambda_2)^{\gamma_2}(k)^{j+1}e^{-\tau k\cos\theta}. \tag{6.19}$$

A similar result holds, of course, for the case $j+l_1+l_2$ odd.

Analogous considerations hold if $\langle g|$ has the form described in Eq. (5.5). Then

$$\langle g|\mathbf{A}^+|f\rangle=(q+1)^{1/2}\int \overline{g(\epsilon',\omega_1,\ldots,\omega_q,\omega)}F^+(\epsilon',\epsilon_{n_1},\omega)f(\epsilon_{n_1},\omega_1,\ldots,\omega_q)d\epsilon'\;d\omega\;d\omega_1\ldots d\omega_q. \tag{6.20}$$

In this case the sum over m in Eq. (6.14) is finite, and the sums over n and p give (see Exton,[8] Chap. 2)

$$\begin{aligned}&F_A^{(3)}(a;-n_{1r},\gamma_2+1,j+1;2\gamma_1+1,2\gamma_2+1,2j+2;x,y,z)\\&\quad=\sum_{m=0}^{n_{1r}}\frac{(a)_m(-n_{1r})_m}{(2\gamma_1+1)_m(1)_m}\,x^mF_A^{(2)}(a+m;\gamma_2+1,j+1;2\gamma_2+1,2j+2;y,z)\end{aligned} \tag{6.21}$$

plus similar terms, where now $F_A^{(2)}(a;b,b';c,c';y,z)$ is *Appell's hypergeometric function* of two variables. From Euler's integral representation for $F_A^{(2)}$,

$$\begin{aligned}F_A^{(2)}(a;b,b';c,c';y,z)&=\frac{\Gamma(c)\Gamma(c')}{\Gamma(b)\Gamma(c-b)\Gamma(b')\Gamma(c'-b')}\\&\quad\times\int_0^1\int_0^1\frac{(s)^{b-1}(1-s)^{c-b-1}(t)^{b'-1}(1-t)^{c'-b'-1}}{(1-sy-tz)^a}\,ds\;dt,\end{aligned} \tag{6.22}$$

valid whenever $0<\mathrm{Re}(b)<\mathrm{Re}(c)$, $0<\mathrm{Re}(b')<\mathrm{Re}(c')$, we see that in our case the Appell function $F_A^{(2)}$, and hence the Lauricella function, $F_A^{(3)}$ admits an analytic continuation separately into the upper half y- and z-planes. Moreover, it can be shown (see Moses and Prosser[10]) that $F_A^{(2)}$ is bounded there by a constant times $\max\{|1-y|^{-1},|1-z|^{-1}\}$. In our case, we know that

$$\begin{aligned}|1-y|&=\frac{|\sigma+\lambda_1-i(p_2-k)e^{-i\theta}|}{|\sigma+\lambda_1+i(p_2+k)e^{-i\theta}|}>\frac{|\sigma+\lambda_1|}{|\sigma+\lambda_1+i(p_2+k)e^{-i\theta}|},\\|1-z|&=\frac{|\sigma+\lambda_1+i(p_2-k)e^{-i\theta}|}{|\sigma+\lambda_1+i(p_2+k)e^{-i\theta}|}=|1-y|.\end{aligned} \tag{6.23}$$

It follows that

$$|F^+(\epsilon',\epsilon_{n_1},\omega)|\leqslant C_{11}(2\lambda_1)^{1/2}\,\frac{1}{\Gamma(a)}\,\frac{\Gamma(\gamma_2+1)}{\Gamma(2\gamma_2+1)}\,\frac{\Gamma(j+1)}{\Gamma(2j+2)}\,|x'|^{\gamma_1}|y'|^{\gamma_2}|z'|^{j+1}e^{-\tau(\epsilon'+\omega)\cos\theta}, \tag{6.24}$$

where

$$x'=\frac{2\lambda_1}{\sigma+\lambda_1-ie^{-i\theta}q},\quad y'=\frac{2p_2}{\sigma+\lambda_1-ie^{-i\theta}q},\quad z'=\frac{2k}{\sigma+\lambda_1-ie^{-i\theta}q}, \tag{6.25}$$

with

$$q=|p_2-k|. \tag{6.26}$$

Since $|x'|<2$, $|y'|<2p_2/\sigma$, $|z'|<2k/\sigma$, it follows that the function $F(\epsilon_{p2},\epsilon_n,\omega)$ as a function of p_2 and k admits an analytic continuation into the sectors $S(p_2)$ and $S(k)$, where

$$\begin{aligned} S(p)&=\{e^{-i\theta}p:0\leqslant\theta\leqslant\theta_{\max}<\pi/2\},\\ S(k)&=\{e^{-i\theta}k:0\leqslant\theta\leqslant\theta_{\max}<\pi/2\}, \end{aligned} \tag{6.27}$$

and is uniformly bounded there. In fact, since $|x'|<2$, $|y'|<2p_2/\sigma$, $|z'|<2k/\sigma$, we have

$$\begin{aligned} |F^+(\epsilon',\epsilon_{n_1},\omega)|\leqslant C_{11}(2\lambda_1)^{1/2}\frac{(2/\sigma)^{\gamma_2+j+1}}{\Gamma(a)}\frac{\Gamma(\gamma_2+1)}{\Gamma(2\gamma_2+1)}\frac{\Gamma(j+1)}{\Gamma(2j+2)}\\ \times(\lambda_1)^{\gamma_1}(p')^{\gamma_2}(k)^{j+1}e^{-\tau(p'+k)\cos\theta}. \end{aligned} \tag{6.28}$$

A similar results holds, of course, for the case $j+l_1+l_2$ odd.

If both $\langle g|$ and $|f\rangle$ have the form in Eq. (5.5), then

$$\langle g|\mathbf{A}^+|f\rangle=\sqrt{q+1}\int\overline{g(\epsilon',\omega_1,\ldots,\omega_q,\omega)}F^+(\epsilon',\epsilon,\omega)f(\epsilon,\omega_1,\ldots,\omega_q)d\epsilon'\ d\epsilon\ d\omega\ d\omega_1\cdots d\omega_q. \tag{6.29}$$

In this case we proceed directly to Euler's integral representation for $F_A^{(3)}$:

$$\begin{aligned} &F_A^{(3)}(a;b,b',b'';c,c',c'';x,y,z)\\ &\quad=\frac{\Gamma(c)\Gamma(c')\Gamma(c'')}{\Gamma(b)\Gamma(c-b)\Gamma(b')\Gamma(c'-b')\Gamma(b'')\Gamma(c''-b'')}\\ &\qquad\times\int_0^1\int_0^1\int_0^1\frac{(r)^{b-1}(1-r)^{c-b-1}(s)^{b'-1}(1-s)^{c'-b'-1}(t)^{b''-1}(1-t)^{c''-b''-1}}{(1-rx-sy-tz)^a}\,dr\ ds\ dt, \end{aligned} \tag{6.30}$$

valid whenever $0<\mathrm{Re}(b)<\mathrm{Re}(c)$, $0<\mathrm{Re}(b')<\mathrm{Re}(c')$, $0<\mathrm{Re}(b'')<\mathrm{Re}(c'')$. From this representation we see that the Lauricella function $F_A^{(3)}$ admits an analytic continuation separately into the upper half x, y, and z planes. It can be shown that $F_A^{(3)}$ is bounded there by a constant times $\max\{|1-x|^{-1},|1-y|^{-1},|1-z|^{-1}\}$. In our case,

$$\begin{aligned} |1-x|&=\frac{|\sigma-i(p_1-p_2-k)e^{-i\theta}|}{|\sigma+i(p_1+p_2+k)e^{-i\theta}|}<1,\\ |1-y|&=\frac{|\sigma+i(p_1-p_2+k)e^{-i\theta}|}{|\sigma+i(p_1+p_2+k)e^{-i\theta}|}<1,\\ |1-z|&=\frac{|\sigma+i(p_1+p_2-k)e^{-i\theta}|}{|\sigma+i(p_1+p_2+k)e^{-i\theta}|}<1. \end{aligned} \tag{6.31}$$

This, with our weight function, leads to the estimate

$$|F^+(\epsilon',\epsilon_{p_1},\omega)|\leqslant C_{12}\frac{1}{\Gamma(a)}\frac{\Gamma(\gamma_1+1)}{\Gamma(2\gamma_2+1)}\frac{\Gamma(\gamma_2+1)}{\Gamma(2\gamma_2+1)}\frac{\Gamma(j+1)}{\Gamma(2j+2)}$$
$$\times|x'|^{\gamma_1}|y'|^{\gamma_2}|z'|^{j+1}e^{-2\tau(\epsilon'+\epsilon+\omega)\cos\theta}, \tag{6.32}$$

where

$$x'=\frac{2p_1}{\sigma-ie^{-i\theta}q},\quad y'=\frac{2p_2}{\sigma-ie^{-i\theta}q},\quad z'=\frac{2k}{\sigma-ie^{-i\theta}q}, \tag{6.33}$$

and

$$q=\min\{|-p_1+p_2+k|,|p_1-p_2+k|,|p_1+p_2-k|\}. \tag{6.34}$$

It follows that the function $F(\epsilon_{p2},\epsilon_{p1},\omega)$ as a function of p_1, p_2 and k admits an analytic continuation into the sectors $S(p_1)$, $S(p_2)$ and $S(k)$ [see Eq. (6.27)] and is uniformly bounded there. In fact, since $|x|<2p_1/\sigma$, $|y|<2p_2/\sigma$, $|z|<2k/\sigma$, we have

$$|F^+(\epsilon',\epsilon,\omega)|\leqslant C_{12}\frac{(2/\sigma)^a}{\Gamma(a)}\frac{\Gamma(\gamma_1+1)}{\Gamma(2\gamma_1+1)}\frac{\Gamma(\gamma_2+1)}{\Gamma(2\gamma_2+1)}\frac{\Gamma(j+1)}{\Gamma(2j+2)}$$
$$\times(p)^{\gamma_1}(p')^{\gamma_2}(k)^{j+1}e^{-\tau(\epsilon'+\epsilon+k)\cos\theta}. \tag{6.35}$$

We can now see that under our assumptions the matrix elements $F^+(\epsilon',\epsilon,\omega)$ are *square-summable* over the energy variables with respect to the measure

$$du=d\rho(\epsilon')d\rho(\epsilon)d\omega, \tag{6.36}$$

where we have set

$$d\rho(\epsilon)=\begin{cases}\sum_n \delta(\epsilon-\epsilon_n)d\epsilon & \text{if } \epsilon<0,\\ d\epsilon & \text{if } \epsilon\geqslant 0.\end{cases} \tag{6.37}$$

This becomes apparent if we recall that the electron quantum number κ is restricted to $|\kappa|\geqslant 1$, so that $\gamma=\sqrt{\kappa^2-e^2}$ is restricted to $\gamma\geqslant 1-e^2/2$, where $e^2=1/137\ll 1$. Hence for the bound states we can write

$$|F^+(\epsilon_{n_2},\epsilon_{n_1},\omega)|^2\leqslant C_{10}^2(\lambda_1\lambda_2)^{3-e^2}\left(\frac{2^a}{\Gamma(a)}\right)^2\left(\frac{j!k^{j+1}e^{-\tau k}}{(2j+1)!}\right)^2. \tag{6.38}$$

For the bound state $|\epsilon_n\rangle$ we have

$$\lambda_n^2=\mu^2-(\mu+\epsilon_n)^2=\mu e^2/2n^2-O(1/n^3), \tag{6.39}$$

so that

$$|F^+(\epsilon_{n_2},\epsilon_{n_1},\omega)|^2\leqslant C_{10}^2\left(\frac{\mu e^2}{n_1n_2}\right)^{3-e^2}\left(\frac{2^a}{\Gamma(a)}\right)^2\left(\frac{j!k^{j+1}e^{-\tau k}}{(2j+1)!}\right)^2. \tag{6.40}$$

It follows that this matrix element is square-integrable with respect to the energy measure du defined above. In this connection, note that

$$\sum_{n_1,n_2}\left(\frac{1}{n_1 n_2}\right)^2=\frac{\pi^4}{36}, \tag{6.41}$$

$$\int_0^\infty k^{2j+3}e^{-2\tau k\cos\theta}dk/k=\frac{(2j+2)!}{(\tau\cos\theta)^{2j+2}}, \tag{6.42}$$

and

$$\frac{(j!)^2}{2j+1!}\leqslant(2)^{-2j}. \tag{6.43}$$

Similarly, for the unbound electron states $|\epsilon\rangle$, we have

$$|F^+(\epsilon',\epsilon,\omega)|^2\leqslant C_{14}^2\left(\frac{\Gamma(\gamma_1+1)(p)^{\gamma_1}e^{-\tau p\cos\theta}}{\Gamma(2\gamma_1+1)}\right)^2\left(\frac{\Gamma(\gamma_2+1)(p')^{\gamma_2}e^{-\tau p'\cos\theta}}{\Gamma(2\gamma_2+1)}\right)^2$$
$$\times\left(\frac{\Gamma(j+1)(k)^{j+1}e^{-\tau k\cos\theta}}{\Gamma(2j+2)}\right)^2, \tag{6.44}$$

and so this matrix element is also square-integrable with respect to the measure du, thanks to the weight function $\omega(\tau)=e^{-\tau E}$.

Similar considerations hold, of course, for the matrix elements $F^+(\epsilon_{n_2},\epsilon_{p_1},\omega)$ and $F^+(\epsilon_{p_2},\epsilon_{n_1},\omega)$.

We shall use the notation

$$\|F^+\|^2=\int|F^+(\epsilon',\epsilon,\omega)|^2\,du \tag{6.45}$$

for the L^2 norm of F^+. We note that in all cases

$$\|F^+\|^2\leqslant C_{13}^2\frac{(\sigma\tau\cos\theta)^{-2a}}{\Gamma(a)^2}, \tag{6.46}$$

where $a=\gamma_1+\gamma_2+j+1$, and C_{13} is essentially independent of the angular quantum numbers.

Finally, we note that the matrix elements F^- of $\mathbf{A}^-$ are just the complex conjugates of the matrix elements F^+ of $\mathbf{A}^+$, and so yield to the same analysis with the same conclusions. In particular,

$$\|F^-\|^2=\|F^+\|^2. \tag{6.47}$$

We put

$$\|F\|^2=\sum_j(\|F_j^+\|^2+\|F_j^-\|^2), \tag{6.48}$$

where the sum is taken over the several components of the interaction.

VII. THE FREDHOLM–BORN SERIES

The *resolvent* $R(\zeta,e)=(H-\zeta)^{-1}$ of the Hamiltonian $H=H_0+eH_1$ can be expressed in terms of the resolvent $R_0(\zeta)=(H_0-\zeta)^{-1}$ of H_0 through the *resolvent equation*:

$$R(\zeta,e)=R_0(\zeta)-eR_0(\zeta)H_1R(\zeta,e). \tag{7.1}$$

If we solve this resolvent equation formally by iteration, we get the Fredholm–Born series:

$$R(\zeta,e)=\sum_{k\geqslant 0}(-e)^k(R_0(\zeta)H_1)^kR_0(\zeta)=R_0(\zeta)-eR_0(\zeta)H_1R_0(\zeta)+e^2R_0(\zeta)H_1R_0(\zeta)H_1R_0(\zeta)$$
$$-\cdots . \tag{7.2}$$

We can also write this equation in the form

$$R(\zeta,e)=R_0(\zeta)-R_0(\zeta)T_0(\zeta,e)R_0(\zeta), \tag{7.3}$$

where

$$T_0(\zeta,e)=\sum_{k\geqslant 0}(-e)^{k+1}(H_1R_0(\zeta))^kH_1=eH_1-e^2H_1R_0(\zeta)H_1+\cdots . \tag{7.4}$$

If ζ lies in the *resolvent set* $\rho(H_0)$ of H_0 (see Fig. 4), and if the operator $R_0(\zeta)H_1$ is bounded in the operator norm with bound b:

$$\|R_0(\zeta)H_1\|\leqslant b, \tag{7.5}$$

then when $eb<1$ the Fredholm–Born series converges in the operator norm to a unique solution $R(\zeta,e_0)$ of the resolvent equation.

Now the resolvent operator $R_0(\zeta)$, for ζ in $\rho(H_0)$ with $R1(\zeta)<\mu+\epsilon+\zeta_1$, is bounded on states $|f_q\rangle$ with q photons by

$$\|R_0(\zeta)\|f_q\rangle\|\leqslant\frac{1}{\delta(\zeta)}\,\||f_q\rangle\|, \tag{7.6}$$

where

$$\delta(\zeta)=\min\{|\zeta-\lambda|:\lambda\,\epsilon\sigma(H_0)\}=\operatorname{dist}\{\zeta,\sigma(H_0)\} \tag{7.7}$$

is the distance of ζ from the spectrum of H_0. It follows that

$$|\langle g_{q+1}|R_0(\zeta)H_1^+|f_q\rangle|^2\leqslant\frac{q+1}{\delta(\zeta-q\nu)^2}\,\|\langle g_{q+1}|\|^2\|F^+\|^2\||f_q\rangle\|^2. \tag{7.8}$$

We note here that, providing $\nu<\epsilon_1$,

$$\frac{q+1}{\delta(\zeta-q\nu)}=\frac{(q+1)\nu}{\nu\delta(\zeta-q\nu)}\leqslant\frac{q\nu+\epsilon_1}{\nu\delta(\zeta-q\nu)}\leqslant\begin{cases}\dfrac{\operatorname{Re}(\zeta)}{\nu\delta(\zeta)} & \text{when } q\nu<\operatorname{Re}(\zeta),\\[2ex] \dfrac{\epsilon_1}{\nu\delta(\zeta)} & \text{when } q\nu>\operatorname{Re}(\zeta).\end{cases} \tag{7.9}$$

It follows that, for all q

$$|\langle g_{q+1}|R_0(\zeta)H_1^+|f_q\rangle|^2\leqslant\frac{C_{14}}{\nu\delta(\zeta)\,\delta(\zeta-q\nu)}\,\|\langle g_{q+1}|\|^2\|F^+\|^2\||f_q\rangle\|^2, \tag{7.10}$$

where here $C_{14}=\max\{\operatorname{Re}(\zeta),\epsilon_1\}$. (Note the role of the photon mass ν in establishing this result.)

Hence

$$\|R_0(\zeta)H_1^+\|^2 \leqslant \frac{C_{14}}{\nu\delta(\zeta)^2}\,\|F^+\|^2. \tag{7.11}$$

Similarly,

$$\|R_0(\zeta)H_1^-\|^2 \leqslant \frac{C_{14}}{\nu\delta(\zeta)^2}\,\|F^-\|^2. \tag{7.12}$$

If we write

$$b(\zeta)^2 = \frac{C_{14}}{\nu\delta(\zeta)^2}\,\|F\|^2, \tag{7.13}$$

then we know that

$$\|R_0(\zeta)H_1\| \leqslant b(\zeta). \tag{7.14}$$

For fixed ζ in $\rho(H_0)$ and small values of e, or for fixed $e>0$ and $\mathrm{Re}(\zeta) \ll 0$, we have $eb(\zeta) < 1$; and for those values of e the Fredholm–Born series for the resolvent $R(\zeta)$ converges in operator norm. When $\zeta < 0$, of course, then

$$b(\zeta)^2 < \frac{C_{14}\|F\|^2}{\nu(\epsilon_1 - \zeta)^2}. \tag{7.15}$$

VIII. ANALYSIS OF THE FREDHOLM–BORN SERIES

First we observe that the angular momentum operators, including the parity and helicity operators, commute with both H_0 and H_1. This means that the state space breaks up into sectors with fixed values of the total angular momentum quantum numbers (J,M,Λ) (where $\Lambda =$ total helicity) and that these sectors are invariant under both H_0 and H_1. Moreover, since these sectors are orthogonal, the square of the norm of a typical state must be the sum of the squares of the norms of the projections of that state into each of the angular momentum sectors. A typical sector will include states of the form $|\epsilon_n, \omega_1, \ldots, \omega_q\rangle$, $q \geqslant 0$, together with all states which can be obtained from these by repeated application of the emission and absorption terms $H_1^{\pm}$ of the interaction operator. In particular, it will contain the bare electron states $|\epsilon_n\rangle$ obtained by absorption of the q photons. The angular quantum numbers of these bare electron states are already determined by the sector quantum numbers (J,M,Λ), so that these states are parametrized by their energy, as our notation suggests. We study in detail only the simplest of these sectors, for which $(J,M,\lambda) = (\frac{1}{2},0,\frac{1}{2})$, but the others can be treated in exactly the same way.

We begin our analysis be rewriting Eq. (7.2) in the form

$$R(\zeta,e) = (I - eK(\zeta) + e^2K(\zeta)^2 - \cdots)R_0(\zeta) = (I + eK(\zeta))^{-1}R_0(\zeta), \tag{8.1}$$

where now we have set

$$K(\zeta) = R_0(\zeta)H_1. \tag{8.2}$$

By the results of the previous section, we know that for $0 \leqslant \theta < \pi/2$, for $\zeta \ll 0$, and for $e \ll 1$ we have (See Eq. (7.14))

$$\|eK(\zeta,\theta)\| \leqslant eb < 1, \tag{8.3}$$

where

$$b^2=\frac{C_{14}\|F\|^2}{\nu(\epsilon_1-\zeta)^2}. \tag{8.4}$$

The Fredholm–Born series for $(I+eK(\zeta,\theta))^{-1}$ then converges in the operator norm, and we have

$$\|(I+eK)^{-1}\|\leqslant\sum_{k\geqslant 0}\|e_0K)\|^k\leqslant\frac{1}{1-eb}. \tag{8.5}$$

Moreover,

$$\begin{aligned}\|R\|&=\|(I+eK)^{-1}R_0\|\leqslant\frac{1}{1-eb}\,\frac{1}{\epsilon_1-\zeta},\\ \|R-R_0\|&=\|eR_0KR\|\leqslant\frac{eb}{1-eb}\,\frac{1}{\epsilon_1-\zeta}.\end{aligned} \tag{8.6}$$

Accordingly, we shall assume throughout this section that θ and ζ and e_0 are fixed, with $0<\theta<\pi/2$, $\zeta\leqslant 0$, and $e_0\leqslant 1$, and shall indicate explicitly the dependence of K on these variables only where relevant. Since the Fredholm–Born series converges in the operator norm, we are free to rearrange this series in any manner we please.

We now classify the states of our sector by the number q of photons present. In particular, we write $\mathfrak{h}_q$ for the space of those states with q photons present, i.e., the qth eigenspace of the photon number operator $N_{\rm ph}$. Thus

$$\mathfrak{h}_q=\{|f_q\rangle: f_q=f(\epsilon,\omega_1,\ldots,\omega_q)\}=\mathfrak{h}_{\rm el}\otimes\mathfrak{h}_{\rm ph}\otimes\cdots\otimes\mathfrak{h}_{\rm ph}\quad(q\ \text{ photons}), \tag{8.7}$$

where $\mathfrak{h}_{\rm el}$ and $\mathfrak{h}_{\rm ph}$ are the one-electron and one-photon state spaces. Then

$$\mathfrak{h}=\sum_q\ \mathfrak{h}_q=\mathfrak{h}_0+\mathfrak{h}_1+\mathfrak{h}_2+\cdots\ . \tag{8.8}$$

(We remark here that the space $\mathfrak{h}_q$ consists of states $|f_q(\epsilon,\omega_1,\ldots,\omega_q)\rangle$ which are *symmetric* in the photon energies ω_1. In rearranging the Fredholm–Born series we shall first ignore this symmetry in computing the matrix elements, and then symmetrize the results. We do this as usual by averaging over the initial states and summing over the final states for each matrix element.)

With this classification we write each operator B on $\mathfrak{h}$ as a matrix with operator entries of the form

$$B_{qp}=E_qBE_p, \tag{8.9}$$

where E_q denotes the orthogonal projection of $\mathfrak{h}$ onto $\mathfrak{h}_q$. In particular, we write

$$\begin{aligned}K_{qp}&=E_qKE_p,\\ (I+eK)^{-1}_{qp}&=E_q(I+eK)^{-1}E_p,\\ R(\zeta,e)_{qp}&=E_qR(\zeta,e)E_p=E_q(I+eK)^{-1}R_0(\zeta)E_p,\ \text{ etc.}\end{aligned} \tag{8.10}$$

We seek to determine these matrix elements $(I+eK)^{-1}_{qp}$ of the operator $(I+eK)^{-1}$.

In order to do so, we introduce the operators Q_q acting on the space $\mathfrak{h}_q$ by setting

$$Q_q=\sum_{k=2p>0}^{\infty}(-e)^{2p}\sum_{q_1,q_2,\ldots,q_{k-1}\geqslant q}K_{q,q_1}K_{q_1,q_2}\cdots K_{q_{k-1},q}, \tag{8.11}$$

where the sum here is restricted to those intermediate states with at least q photons. (Note that only the terms of even order can contribute to the sum over k.) We also introduce the raising and lowering operators $P_q^{\pm}$ by setting

$$\begin{aligned}P_q^+&=-(I+Q_q)eK_{q,q-1}^+,\\ P_q^-&=-eK_{q-1,q}^-(I+Q_q).\end{aligned} \tag{8.12}$$

We introduce the diagonal operators M_{qq} by setting inductively

$$\begin{aligned}I+M_{00}&=I+Q_0,\\ I+M_{(q+1)(q+1)}&=(I+Q_{q+1})+P_{q+1}^+(I+M_{qq})P_{q+1}^-,\end{aligned} \tag{8.13}$$

and the nondiagonal operators M_{rq} by setting

$$M_{rq}=\begin{cases}(I+M_{rr})P_{r+1}^-P_{r+2}^-\cdots P_q^- & \text{if } r<q,\\ P_r^+P_{r-1}^+\cdots P_{q+1}^+(I+M_{qq}) & \text{if } r>q.\end{cases} \tag{8.14}$$

Note that the diagonal operators $I+Q_q$ and $I+M_{qq}$, as well as the resolvent R_0, commute with the photon number operator N_{ph} and so leave the eigenspaces $\mathfrak{h}_q$ invariant.

Now, with these definitions, we can write the matrix for $(I+ek)^{-1}$ in the form

$$I+M=(I+eK)^{-1}, \tag{8.15}$$

or

$$\begin{pmatrix}I+M_{00}&M_{01}&M_{02}&\cdots\\ M_{10}&I+M_{11}&M_{12}&\cdots\\ M_{20}&M_{21}&I+M_{22}&\cdots\\ \vdots&\vdots&\vdots&\ddots\end{pmatrix}=\begin{pmatrix}I&K_{01}^-&0&\cdots\\ K_{10}^+&I&K_{12}^-&\cdots\\ 0&K_{21}^+&I&\cdots\\ \vdots&\vdots&\vdots&\ddots\end{pmatrix}^{-1}. \tag{8.16}$$

It follows that the matrix elements of the operator $(I+eK)^{-1}$, and hence of the resolvent $R=(I+eK)^{-1}R_0$, can each be expressed as a finite sum of products involving only the $K^{\pm}$ and the Q_q. In fact, if $r>q$, then

$$\begin{aligned}(I+eK)_{rq}^{-1}=(-e)^{r-q}&(I+Q_r)K^+(I+Q_{r-1})K^+\cdots K^+(I+Q_{q+1})K^+\sum_{0\leqslant k\leqslant q}(-e)^{2k}\\ &\times(I+Q_q)K^+\cdots K^+(I+Q_{q-k})K^-\cdots K^-(I+Q_q),\end{aligned} \tag{8.17}$$

while if $r\leqslant q$, then

$$\begin{aligned}(I+eK)_{rq}^{-1}=\sum_{0\leqslant k\leqslant q}&(-e)^{2k}(I+Q_r)K^+\cdots K^+(I+Q_{r-k})K^-\cdots K^-(I+Q_r)(-e)^{q-r}\\ &\times K^+(I+Q_{r+1})K^+\cdots K^+(I+Q_{q-1})K^+(I+Q_q).\end{aligned} \tag{8.18}$$

IX. ANALYSIS OF THE DIAGONAL OPERATORS Q_q

We now turn to an analysis of the diagonal operators Q_q [see Eq. (8.11)] acting on the spaces $\mathfrak{h}_q$. For this purpose we introduce

$$V_q=\sum_{k=2p>0}^{\infty}(-e)^{2p}\sum_{q_1,q_2,\ldots,q_{k-1}>q}K_{q,q_1}K_{q_1,q_2}\cdots K_{q_{k-1},q}, \tag{9.1}$$

where now the sum over the q_j is restricted to those intermediate states with at least $q+1$ photons present $(q_j>q)$. Then when $eb<1$, with $b^2=C_{14}\|F\|^2/\nu(\epsilon_1-\zeta)^2$, this series converges, and we have

$$\|V_q\|\leqslant\frac{(eb)^2}{1-(eb)^2}. \tag{9.2}$$

Moreover, we see that

$$Q_q=V_q+V_qQ_q. \tag{9.3}$$

By iteration,

$$Q_q=V_q+V_q^2+V_q^3+\cdots. \tag{9.4}$$

From the estimates above, we know that this series converges when $\|eK\|\leqslant eb<1/\sqrt{2}$, and then

$$\|I+Q_q\|\leqslant\frac{1-(eb)^2}{1-2(eb)^2},\quad I+Q_q=(I-V_q)^{-1}. \tag{9.5}$$

We also see that

$$V_q=e^2K^-(I+Q_{q+1})K^+. \tag{9.6}$$

If we combine Eq. (9.1) and Eq. (9.5), we find that

$$V_q=e^2K^-(I-V_{q+1})^{-1}K^+. \tag{9.7}$$

If we write this in the suggestive form

$$V_q=e^2\,\frac{K^-K^+}{(I-V_{q+1})}, \tag{9.8}$$

then after k iterations we get a sort of continued fraction expansion for V_q:

$$V_q=\cfrac{e^2K^-K^+}{1-\cfrac{e^2K^-K^+}{1-\cfrac{e^2K^-K^+}{\ddots\\ 1-\cfrac{e^2K^-K^+}{1-V_{q+k}}}}}. \tag{9.9}$$

X. ANALYSIS OF THE TRUNCATED RESOLVENT $F_qR(\zeta,e)F_q$

We now consider that part of the resolvent $R(\zeta,e)$ which acts on the subspace $\mathfrak{K}_q=\mathfrak{h}_0+\mathfrak{h}_1+\cdots+\mathfrak{h}_q$ of at most q photons. We let $F_q=E_0=E_1+\cdots+E_q$ be the projection on this subspace $\mathfrak{K}_q$, and consider the truncated resolvent

$$F_qR(\zeta,e)F_q=F_q(I+eK)^{-1}R_0F_q=F_q(I+eK)^{-}F_qR_0, \tag{10.1}$$

where

$$F_q(I+eK)^{-1}F_q=F_q(I+M)F_q=\begin{pmatrix} I+M_{00} & M_{01} & \cdots & M_{0q} \\ M_{10} & I+M_{11} & \cdots & M_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ M_{q0} & M_{q1} & \cdots & I+M_{qq} \end{pmatrix}. \tag{10.2}$$

This operator is related to the operator $(I+M)_q=(I+eF_qKF_q)^{-1}$ acting on the subspace $\mathfrak{K}_q$. In matrix form,

$$\begin{pmatrix} (I+M_q)_{00} & (M_q)_{01}\cdots & (M_q)_{0q} \\ (M_q)_{10} & (I+M_q)_{11}\cdots & (M_q)_{1q} \\ \vdots & \ddots & \vdots \\ (M_q)_{q1} & (M_q)_{q2}\cdots & (I+M_q)_{qq} \end{pmatrix}=\begin{pmatrix} I & K^- & 0\ldots & 0 \\ K^+ & I & K^-\ldots & 0 \\ 0 & K^+ & I\ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0\ldots & I \end{pmatrix}^{-1}. \tag{10.3}$$

If we analyze the matrix form of the operator $I+M_q=((I+eF_qKF_q)^{-1})$ in the same way as we analyzed the matrix form of the operator $I+M=(I+eK)^{-1}$, we find the same structure:

$$\begin{aligned} (I+M_q)_{00}&=(I+Q_q)_0, \\ (I+M_q)_{ss}&=(I+Q_q)_s+(P_q)s_s^+(I+M_q)_{(s-1)(s-1)})(P_q)_s^-, \\ (M_q)_{sr}&=\begin{cases} (I+M_q)_{rr}(P_q)_{r+1}^-(P_q)_{r+2}^-\cdots(P_q)_s^- & \text{if } r>s, \\ (P_q)_r^+(P_q)_{r-1}^+\cdots(P_q)_{s+1}^+(I+M_q)_{ss} & \text{if } r<s, \end{cases} \\ (I+Q_q)_s)&=(I-V_q)_s^{-1}. \end{aligned} \tag{10.4}$$

Examination of the structure for $I+M_q$ reveals that it differs from the structure for $I+M$ only in the terms $(V_q)_s$; i.e., if we replace $(V_q)_s$ by V_s throughout, then $I+M_q$ becomes $I+M$. But [see Eq. (9.9)]

$$(V_q)_s=1-\cfrac{e^2K^-K^+}{1-\cfrac{e^2K^-K^+}{1-\cfrac{e^2K^-K^+}{\ddots}}} \qquad 1-\cfrac{e^2K^-K^+}{1-(V_q)_q} \tag{10.5}$$

while

$$V_s = 1 - \cfrac{e^2 K^- K^+}{1 - \cfrac{e^2 K^- K^+}{1 - \cfrac{e^2 K^- K^+}{\ddots \\ 1 - \cfrac{e^2 K^- K^+}{1 - V_q}}}}. \tag{10.6}$$

This shows that if we replace $(V_q)_q = 0$ by V_q in the continued fraction expansion of $(V_q)_s$, we get V_s. It follows that

$$I + M_q = \begin{pmatrix} I & eK^- & 0\ldots & 0 & 0 \\ eK^+ & I & eK^- \ldots & 0 & 0 \\ 0 & eK^+ & I \ldots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0\ldots & I & eK^- \\ 0 & 0 & 0\ldots & eK^+ & I \end{pmatrix}^{-1}, \tag{10.7}$$

while

$$F_q(I+M)F_q = \begin{pmatrix} I & eK^- & 0\ldots & 0 & 0 \\ eK^+ & I & eK^- \ldots & 0 & 0 \\ 0 & eK^+ & I \ldots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0\ldots & I & eK^- \\ 0 & 0 & 0\ldots & eK^+ & I - V_q \end{pmatrix}^{-1}. \tag{10.8}$$

Since $R = (I+M)R_0$, and since

$$(I - V_q)^{-1} R_0 = ((H_0 - \zeta)(I - V_q))^{-1} = (H_0 - \zeta - U_q)^{-1}, \tag{10.9}$$

where

$$U_q = (H_0 - \zeta)V_q, \tag{10.10}$$

it follows that

$$F_q R(\zeta, e) F_q = (F_q (H_0 - \zeta - U_q) F_q)^{-1}, \tag{10.11}$$

or

$$F_q R(\zeta, e) F_q = \begin{pmatrix} H_0 - \zeta & eH_1^- & 0\ldots & 0 & 0 \\ eH_1^+ & H_0 - \zeta & eH_1^- \ldots & 0 & 0 \\ 0 & eH_1^+ & h_0 - \zeta \ldots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0\ldots & H_0 - \zeta & eH_1^- \\ 0 & 0 & 0\ldots & eH_1^+ & H_0 - \zeta - U_q \end{pmatrix}^{-1}. \tag{10.12}$$

XI. THE CONNECTED COMPONENTS OF THE OPERATOR *M*

With each contribution of the form $((eK^{\pm})^m)_{qp}$ to the matrix element M_{qp} of $I+M$ we shall associate in the usual way a Feynman diagram, with one initial electron and p initial photon lines,

and one final electron and q final photon lines. (Note that two such contributions will have the same Feynman diagram if and only if they have the same symmetrization.) We say that an external photon line is *free* if it is not involved in the interaction, i.e., if it is not connected to the rest of the diagram, and is *interacting* otherwise. We classify each contribution by how many of these external photon lines in the associated Feynman diagram are interacting, i.e., are connected to the rest of the diagram. In general, there must be at least $|p-q|$ such photon lines, and can be at most $p+q$ such lines, so we can write

$$M_{qp}=M_{qp}^{(|p-q|)}+M_{qp}^{(|p-q|+2)}+\cdots+M_{qp}^{(p+q-2)}+M_{qp}^{(p+q)}, \tag{11.1}$$

there being $p\wedge q+1$ terms in all, where $p\wedge q=p+q-|p-q|=\min\{p,q\}$. Of these terms, $M_{qp}^{(p+q-2k)}$ is the sum of all contributions having a diagram with k free photon lines.

In particular, the components of $M_{qp}^{(p+1)}$ all have diagrams with no free photon lines, and so are connected. We refer to these components as *connected components*, and to all the others as *disconnected components*.

The connected components of M_{qp} of order m in e all have kernels of the form $M_{qp}(\epsilon',(\omega')_q;\epsilon,(\omega)_p)$ which involve all the $p+q$ photon lines, and so are nonsingular. In fact, according to Eqs. (7.11)–(7.12), these kernels are all square summable, with Hilbert–Schmidt norm given by

$$\|((eK)^m)_{qp}\|^2\leqslant\|eK\|_2^{2m}\leqslant(eb)^{2m}<1. \tag{11.2}$$

It follows that they determine a Hilbert–Schmidt operator from $\mathfrak{h}_p$ to $\mathfrak{h}_q$. The same is true of the sum $C_{qp}=M_{qp}^{(p+q)}$ of all the connected components of M_{qp}, once we show that this sum converges.

For $\{q,p\}=\{0,0\}$ we have

$$C_{00}=\left(\sum_{k>0}(eK)^{2k}\right)_{00}, \tag{11.3}$$

and it follows that C_{00} has Hilbert–Schmidt norm given by

$$\|C_{00}\|\leqslant\sum_{k>0}\|eK\|_2^{2k}=\frac{(eb)^2}{1-(eb)^2},\quad\text{where}\quad b^2=\|K\|_2^2\leqslant\frac{C\|F\|_2^2}{\nu(\epsilon_1-\zeta)^2}. \tag{11.4}$$

Similarly, for $\{q,p\}\neq\{0,0\}$,

$$C_{qp}=\left(\sum_{k\geqslant 0}(eK)^{p+q+2k}\right)_{qp}, \tag{11.5}$$

and so that C_{qp} has Hilbert–Schmidt norm given by

$$\|C_{qp}\|\leqslant\sum_{k\geqslant 0}\|eK\|_2^{p+q+2k}=\frac{(eb)^{p+q}}{1-(eb)^2}, \tag{11.6}$$

It now follows that the operator C defined by these matrix elements C_{qp} acts on the Hilbert space $\mathfrak{h}$ is a well-defined Hilbert–Schmidt operator, with Hilbert–Schmidt norm given by

$$\|C\|^2=\sum_{p,q\geqslant 0}\|C_{qp}\|_2^2\leqslant\sum_{p,q\geqslant 0}\frac{(eb)^{2p+2q}}{(1-(eb)^2)^2}=\frac{1}{(1-(eb)^2)^4}. \tag{11.7}$$

Now we turn to the sum $M_{qp}^{(|p-q|+2k)}$ of disconnected components with k free photon lines. For $p=q$ we have

$$M_{qq}^{(0)}=\left(\sum_{m>0}(eK)^{2m}\right)_{qq}, \tag{11.8}$$

so that

$$\|M_{qq}^{(0)}\|\leqslant\sum_{m>0}(eb)^{2m}=\frac{(eb)^2}{1-(eb)^2}, \tag{11.9}$$

and

$$M_{qq}^{(2k)}=\left(\sum_{m\geqslant 0}(eK)^{2k+2m}\right)_{qq}, \tag{11.10}$$

so that

$$\|M_{qq}^{(2k)}\|\leqslant\sum_{m\geqslant 0}(eb)^{2k+2m}=\frac{(eb)^{2k}}{1-(eb)^2}. \tag{11.11}$$

For $p\neq q$ we have

$$M_{qp}^{(|p-q|+2k)}=\left(\sum_{m\geqslant 0}(eK)^{|p-q|+2k+2m}\right)_{qp}, \tag{11.12}$$

so that

$$\|M_{qp}^{(|p-q|+2k)}\|\leqslant\sum_{m\geqslant 0}(eb)^{|p-q|+2k+2m}=\frac{(eb)^{|p-q|+2k}}{1-(eb)^2}. \tag{11.13}$$

Since

$$M_{qp}=M_{qp}^{(|p-q|)}+M_{qp}^{(|p-q|+2)}+\ldots+M_{qp}^{(p+q)}, \tag{11.14}$$

it follows that, for $p=q$,

$$\|M_{qq}\|\leqslant\frac{(eb)^2}{1-(eb)^2}+\sum_{k>0}\frac{(eb)^{2k}}{1-(eb)^2}\leqslant\frac{2(eb)^2}{(1-(eb)2)^2}, \tag{11.15}$$

while for $p\neq q$,

$$\|M_{qp}\|\leqslant\sum_{k\geqslant 0}\frac{(eb)^{|p-q|+2k}}{1-(eb)^2}=\frac{(eb)^{|p-q|}}{(1-(eb)^2)^2}. \tag{11.16}$$

If we sum these norms over p, we get

$$\sum_p\|M_{qp}\|\leqslant\frac{2(eb)^2}{(1-(eb)^2)^2}+2\sum_{p>q}\frac{(eb)^{p-q}}{(1-(eb)^2)^2}=\frac{2(eb)^2+2(eb)}{(1-(eb)^2)^2(1-eb)}\leqslant\frac{4(eb)}{(1-eb)^3}. \tag{11.17}$$

Since this holds for all q, we see that

$$\sup_q \sum_p \|M_{qp}\| \leqslant \frac{4eb}{(1-eb)^3}. \tag{11.18}$$

In the same way, we can see that

$$\sup_p \sum_q \|M_{qp}\| \leqslant \frac{2}{(1-eb)^3}. \tag{11.19}$$

It follows that the operator M defined on the Hilbert space $\mathfrak{h}$ by the matrix elements M_{qp} is well-defined, with operator norm bounded by

$$\|M\| \leqslant \frac{4eb}{(1-eb)^3}. \tag{11.20}$$

Because the components of the matrix elements $M_{qp}^{(|p-q|+2k)}$ are disconnected when $k>0$, however, their integral kernels are singular, and so M_{qp} is not a Hilbert–Schmidt operator on $\mathfrak{h}_p$, and hence M is not a Hilbert–Schmidt operator on $\mathfrak{h}$.

XII. THE TRANSFER MAPPING *T*

At this point we recall that every connected component of the matrix element M_{qp} for $M = eK(1-eK)^{-1}$ is a product of the form $((eK^{pm})^m)_{qp}$ whose diagram is connected, and whose norm is bounded by $(eb)^m$. If we now add k free photon lines to the associated diagram, we get a disconnected diagram which describes a component of the form $((eK^{\pm})^m)_{(q+k)(p+k)}$ of the matrix element $M_{(q+k)(p+k)}$, whose norm is bounded by eb_k^m, where b_k is obtained from b by replacing ϵ by $\epsilon_1 + k\nu$ throughout:

$$b_k^2 = \frac{C_{14}\|F\|}{\nu(\epsilon_1 + k\nu - \zeta)^2} \leqslant b^2. \tag{12.1}$$

Moreover, it is clear that every disconnected component of M_{qp}^{p+q-2k} can be obtained from a unique connected component of $M_{(q-k)(p-k)}^{p+q-2k}$ simply by adding k free photon lines to the associated diagram in this way.

This observation suggests that we introduce a *transfer mapping* T, which sends each component of the matrix element M_{qp} into the corresponding component of the matrix element $M_{(q+1)(p+1)}$ obtained by adding one additional free photon line to the associated diagram. This mapping has the effect of replacing $\{p,q\}$ by $\{p+1,q+1\}$, and ϵ_1 by $\epsilon_1+\nu$ throughout in the calculation of the components of $T(M_{qp})$. In this way we assign to the matrix M a new matrix $T(M)$ with elements $T(M)_{(q+1)(p+1)} = T(M_{qp})$. If we agree that $M_{qp}=0$ if $p<0$ or $q<0$, then elements in the top row and the first column of the matrix $T(M)$ are all zero, while each of the other element $T(M)_{qp}$ is obtained from the element $M_{(q-1)(p-1)}$ immediately above and to the left in the matrix:

$$T(M) = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & T(M_{00}) & T(M_{01}) & T(M_{02}) & \cdots \\ 0 & T(M_{10}) & T(M_{11}) & T(M_{12}) & \cdots \\ 0 & T(M_{20}) & T(M_{21}) & T(M_{22}) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{12.2}$$

The discussion above shows that if the matrix M has bound $\|M\| \leqslant 4eb/(1-eb)^3$, then the matrix $T(M)$ has bound

$$\|T(M)\| \leqslant 4eb_1/(1-eb_1)^3, \quad \text{where } b_1 = \frac{\epsilon_1 - \zeta}{\epsilon_1 + \nu - \zeta}\, b < b. \tag{12.3}$$

We introduce iterates T^k of the transfer mapping T in the obvious way: $T^k(M) = T(T^{k-1}(M))$. Then $T^k M$ is obtained by adding k free photon lines to the Feynman diagram associated with each component of all the matrix elements M_{qp}, and we have

$$\|T^k(M)\| \leqslant \frac{4eb_k}{(1-eb_k)^3}, \quad \text{where } b_k = \left(\frac{\epsilon_1 + (k-1)\nu - \zeta}{\epsilon_1 + k\nu - \zeta}\right) b_{k-1} < b_{k-1}. \tag{12.4}$$

Using the transfer mapping and its iterates, we can write the matrix element M_{qp} solely in terms of the connected matrix elements C_{sr} as follows: if $p \geqslant q$, then

$$M_{qp} = M_{qp}^{(p+q)} + M_{qp}^{(p+q-2)} + \cdots + M_{qp}^{(p-q)} = C_{qp}^{(p+q)} + T(C_{(q-1)(p-1)}^{(p+q-2)}) + \cdots + T^q(C_{(q-q)(p-q)}^{(p-q)}), \tag{12.5}$$

and similarly if $p \leqslant q$. Formally, this equation tells us how to write the operators M in terms of the operators C:

$$\begin{aligned} M &= C + T(C) + T^2(C) + \cdots, \\ MR_0 &= CR_0 + T(CR_0) + T^2(CR_0) + \cdots. \end{aligned} \tag{12.6}$$

The first expression here is only formal, but the second converges in the operator norm. To see this, we estimate the norm of the right-hand side to get

$$\begin{aligned} \|MR_0\| &\leqslant \sum_{k \geqslant 0} \|T^k C R_0\| \leqslant \sum_{k \geqslant 0} \frac{4eb_k}{(1-eb_k)^3} \frac{1}{\epsilon_1 + k\nu - \zeta} \leqslant \frac{4eb}{(1-eb)^3} \sum_{k \geqslant 0} \frac{(\epsilon_1 - \zeta)^{1/2}}{(\epsilon_1 + k\nu - \zeta)^{3/2}} \\ &\leqslant \frac{4eb}{(1-eb)^3} \int_{-1}^{\infty} \frac{(\epsilon_1 - \zeta)^{1/2}) dx}{(\epsilon_1 + \nu x - \zeta)^{3/2}} \leqslant \frac{4eb}{(1-eb)^3} \frac{2(\epsilon_1 - \zeta)^{1/2}}{\nu(\epsilon_1 - \nu - \zeta)^{1/2}}, \end{aligned} \tag{12.7}$$

so that the series for MR_0 does converge to MR_0 in the operator norm.

We can rephrase this result as an operator equation:

$$MR_0 = CR_0 + T(MR_0) = CR_0 + T(M)R_0. \tag{12.8}$$

Multiplying on the right by $H_0 - \zeta$, we get

$$M = C + T(M), \tag{12.9}$$

which gives us an operator equation which must be satisfied by the operator M.

If we note that $F_0 R_0 = R_0 F_0$ and

$$R_0 = F_0 R_0 + T(R_0), \tag{12.10}$$

then we can write

$$R_0 + MR_0 = F_0 R_0 + CR_0 + T(R_0 + MR_0), \tag{12.11}$$

or

$$R = F_0 R_0 + CR_0 + T(R). \tag{12.12}$$

Similarly, if we note that

$$F_{q+1}=F_0+T(F_q),$$
$$M_{(q+1)(q+1)}=C_{(q+1)(q+1)}+T(M_{qq}), \tag{12.13}$$

then we can write

$$F_{q+1}(I+M)F_{q+1}=F_0+F_{q+1}CF_{q+1}+T(F_q(I+M)F_q), \tag{12.14}$$

which leads to a form for the truncated resolvent $R^{(q)}(\zeta)$:

$$R^{(q+1)}(\zeta)=F_{q+1}R(\zeta)F_{q+1}=F_0R_0(\zeta)+F_{q+1}C(\zeta)R_0(\zeta)F_{q+1}+T(F_qR(\zeta)F_q). \tag{12.15}$$

In the same way, we note that

$$K=F_1KF_1+T(K), \tag{12.16}$$

so that

$$I+K=F_0+F_1KF_1+T(I+K). \tag{12.17}$$

Moreover, if we write $V_q^{(2q)}$ for the connected part of V_q, then

$$V_{q+1}=V_{q+1}^{(2q+2)}+T(V_q). \tag{12.18}$$

It follows that

$$F_{q+1}(I+K-V_{q+1})F_{q+1}=F_0+F_1KF_1+V_{q+1}^{(2q+2)}+T(F_q(I+K-V_q)F_q). \tag{12.19}$$

This leads us to a form for the truncated Hamiltonian $H^{(q)}-\zeta$:

$$\begin{aligned}H^{(q+1)}-\zeta&=F_{q+1}(H_0-\zeta+eH_1-U_{q+1})F_{q+1}\\&=F_0(H_0-\zeta)F_0+F_1(eH_1)F_1-F_{q+1}U_{q+1}^{(2q+2)}F_{q+1}+T(F_q(H_0-\zeta+eH_1-U_q)F_q)\\&=F_0(H_0-\zeta)F_0+F_1(eH_1)F_1-F_{q+1}U_{q+1}^{(2q+2)}F_{q+1}T(H^{(q)}-\zeta)\\&=((I-T)(H_0-\zeta)+T(H^{(q)}-\zeta))+(F_1(eH_1)F_1-F_{q+1}U_{q+1}^{(2q+2)}F_{q+1}).\end{aligned} \tag{12.20}$$

We can write this as

$$H^{(q+1)}-\zeta=(H_1^{(q+1)}-\zeta)+(F_1(eH_1)F_1-F_{q+1}U_{q+1}^{(2q+2)}F_{q+1}), \tag{12.21}$$

where

$$H_1^{(q+1)}-\zeta=(I-T)(H_0-\zeta)+T(H^{(q)}-\zeta). \tag{12.22}$$

If we iterate this relation k times, we get

$$H^{(q+1)}-\zeta=(H_k^{(q+1)}-\zeta)+(F_k(eH_1)F_k-F_{q+1}(U_{q+1}^{(2q-2k)}+\cdots+U_{q+1}^{(2q+2)})F_{q+1}), \tag{12.23}$$

where

$$H_k^{(q+1)}-\zeta=(I-T^k)(H_0-\zeta)+T^k(H^{(q-k)}-\zeta). \tag{12.24}$$

XIII. ANALYSIS OF THE DIAGONAL OPERATORS V_q

We can write the diagonal operators V_q in the form

$$V_q = CV_q + DV_q, \tag{13.1}$$

where CV_q is the sum of the contributions from all the connected Feynman diagrams, and DV_q is the sum of the contributions from all the disconnected diagrams. We can write all contributions to the disconnected part DV_q of V_q entirely in terms of the contributions to the connected parts CV_p of V_p for $p \leqslant q$:

$$DV_q = T(CV_{q-1}) + T^2(CV_{q-2}) + \cdots + T^q(CV_0), \tag{13.2}$$

where the contributions in $T^j(CV_0)$ all have disconnected diagrams with exactly j disconnected photon lines.

Recall that CV_q denotes the sum of the terms with connected diagrams and so is Hilbert–Schmidt on $\mathfrak{h}_q$, with

$$\|CV_q\|_2 \leqslant (eb)^{2q} + (eb)^{2q+2} + \cdots = \frac{(eb)^{2q}}{1-(eb)^2}. \tag{13.3}$$

Moreover, V_q, and hence Q_q, can be expressed in terms of these operators CV_p, for $p \leqslant q$. In particular,

$$V_q = CV_q + T(CV_{q-1}) + \cdots + T^q(CV_0). \tag{13.4}$$

Now we turn to a closer analysis of the operator V_0. Since there are no contributions to $V_0 = CV_0 + DV_0$ from disconnected diagrams, we see that $DV_0 = 0$, so $V_0 = CV_0$. This means that V_0 is already a Hilbert–Schmidt operator on the space $\mathfrak{h}_0$, with

$$\|V_0\|_2 \leqslant \frac{(eb)^2}{1-(eb)^2}. \tag{13.5}$$

It follows that $Q_0 = V_0(I - V_0)^{-1}$ is also Hilbert–Schmidt, with

$$\|Q_0(\zeta, e^2\|_2 \leqslant \frac{(eb)^2}{1-2(eb)^2}. \tag{13.6}$$

We say that a Feynman diagram contains an *insertion* if it contains a proper subdiagram connected to the whole only by electron lines. We say a diagram is a *skeleton diagram* if it contains no insertion, and a nonskeleton diagram otherwise. (See Fig. 7 and Fig. 8.)

Each nonskeleton diagram can evidently be regarded as formed from a skeleton diagram by adding one or more insertions. It follows that each matrix element of the matrix $(I + eK)^{-1}$ can be regarded as formed by summing over all contributions from the different skeleton diagrams, each contribution formed by summing over all possible insertions into each of the electron lines in the skeleton. (See Fig. 9.)

Each insertion has the form

$$T^q(I + Q_0) = T^q((I - V_0)^{-1}), \tag{13.7}$$

where q here is the number of photon lines present in the skeleton at the point of insertion. If we write

$$I + Q_0' = (I + Q_0) + T(I + Q_0) + T^2(I + Q_0) + \cdots = (I - T)^{-1}(I + Q_0), \tag{13.8}$$

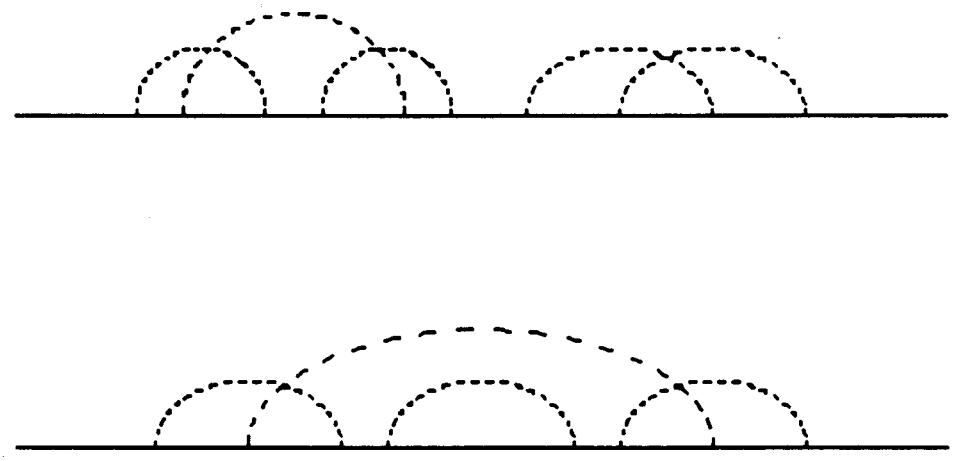

FIG. 7. Reducible vs irreducible graphs.

i.e.,

$$I+Q_0'=\begin{pmatrix} (I+Q_0) & 0 & 0 & 0 & \cdots \\ 0 & T(I+Q_0) & 0 & 0 & \cdots \\ 0 & 0 & T^2(I+Q_0) & 0 & \cdots \\ 0 & 0 & 0 & T^3(I+Q_0) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{13.9}$$

and similarly,

$$I-V_0'=(I-V_0)+T(I-V_0)+T^2(I-V_0)+\cdots=(I-T)^{-1}(I-V_0), \tag{13.10}$$

then we can write $(I+eK)^{-1}$ in the form

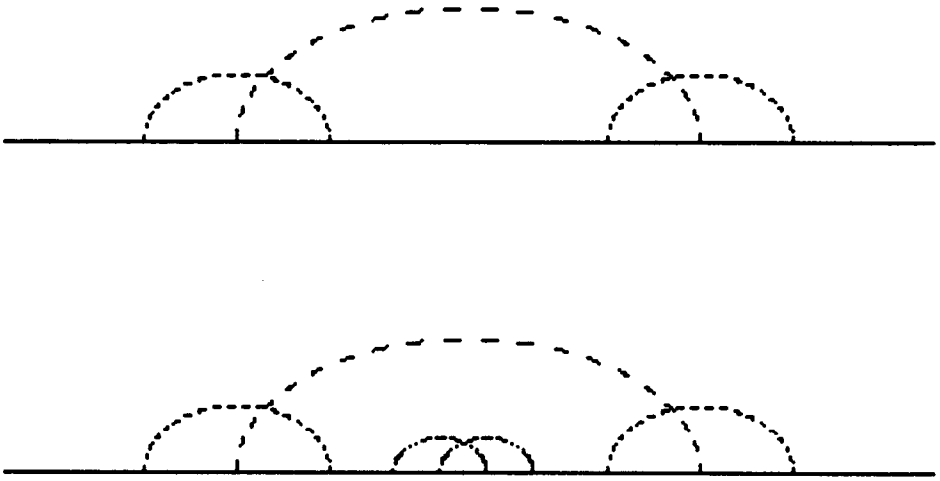

FIG. 8. Skeleton vs nonskeleton graphs.

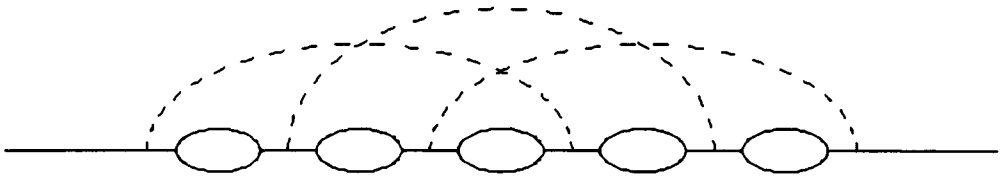

FIG. 9. Skeleton graph with insertions.

$$(I+eK)^{-1}=(I+Q_0')+\sum_{k>0}(-e)^k(((I+Q_0')K)^k(I+Q_0'))_{\mathrm{sk}}=(I-V_0')^{-1}$$

$$+\sum_{k>0}(-e)^k(((I-V_0')^{-1}K)^k(I-V_0')^{-1})_{\mathrm{sk}}, \tag{13.11}$$

where now the sums are restricted to those products associated with skeleton graphs. In this way we can express the resolvent $(I+eK)^{-1}$ in terms of the operator $(I-V_0')^{-1}$ and contributions from skeleton graphs.

In particular, this is also true of V_0 itself: V_0 can be expressed entirely in terms of $(I-V_0')^{-1}$ and contributions from skeleton graphs, in the form (see van Hove[11])

$$V_0'=\Phi(V_0'), \tag{13.12}$$

where

$$\Phi(V)=\sum_{k=2p>0}^{\infty}(-e)^k\Phi_k(V),$$

$$\Phi_k(V)=\sum_{q_1,q_2,\ldots,q_{k-1}>0}(K_{0,q_1}(I-V)^{-1}K_{q_1,q_2}(I-V)^{-1}\cdots(I-V)^{-1}K_{q_{k-1},0})_{\mathrm{sk}}. \tag{13.13}$$

To see this, we consider, for fixed $\zeta\ll 0$ in $\rho(H_0)$ and fixed $e\ll 1$, the class $\mathfrak{C}$ of bounded operators $V=V(\zeta,e^2)$ defined on $\mathfrak{h}_0$, with norm $\|V\|\leqslant 1/2$. If V is in this class, then $I-V$ is invertible, with $\|(I-V)^{-1}\|\leqslant 2$. Hence we can define the mapping Φ on the class $\mathfrak{C}$ by means of Eq. (13.13) above. If $\|eK\|\leqslant eb<1/3$, then

$$\|\Phi(V)\|\leqslant 2/5<1/2. \tag{13.14}$$

Hence if $\|eK\|\leqslant eb<1/3$, then Φ maps the class $\mathfrak{C}$ into itself. Moreover, if V and V' are any two members of this class, then we can establish the identity

$$\Phi(V)-\Phi(V')=\Phi(V)(I-V)^{-1}(V'-V)(I-V')^{-1}\Phi(V'), \tag{13.15}$$

and the bound

$$\|\Phi(V)-\Phi(V')\|\leqslant(16/25)\|V-V'\|. \tag{13.16}$$

It follows that $\Phi(V)$ is a *contraction mapping* on the class $\mathfrak{C}$. By the contraction mapping theorem, there is, for each ζ in $\rho(H_0)$ and for each e with $eb\leqslant 1/3$, a unique operator $V_0'=V_0'(\zeta,e^2)$ in $\mathfrak{C}$ for which

$$V_0'=\Phi(V_0'). \tag{13.17}$$

Moreover, we can construct V_0' recursively from the algorithm

$$V_0^{(0)}=0, \quad V_0^{(n)}=\Phi(V_0^{(n-1)}), \quad V_0^{(n)}\to V_0'. \tag{13.18}$$

Since each contribution to $V_0{}^{(n)}$ is a contribution to V_0, and since each contribution to V_0' is a contribution to some $V_0^{(n)}$, we see that $V_0^{(n)}$ must indeed converge to V_0', and that Eq. (13.12) must hold.

Now we can write the operator $R(\zeta,e)=(I+eK(\zeta))^{-1}R_0(\zeta)$ in the form

$$\begin{aligned} R(\zeta,e)&=\sum_{k\geqslant 0}(-e)^k(R_0(\zeta)H_1)^kR_0(\zeta)\\ &=\sum_{m>0}(-e)^m[((I-V_0'(\zeta e^2))^{-1}R_0(\zeta)H_1)^m(I-V_0'(\zeta,e^2))^{-1}R_0(\zeta)]_{\mathrm{sk}}. \end{aligned} \tag{13.19}$$

But

$$(I-V_0'(\zeta,^2))^{-1}R_0(\zeta)=((H_0-\zeta)(I-V_0'(\zeta,e^2)))^{-1}=(H_0-\zeta-U_0'(\zeta,e^2))^{-1}=R_0(\zeta,e^2), \tag{13.20}$$

where we have put

$$U_0'(\zeta,e^2)=(H_0-\zeta)V_0'(\zeta,e^2) \tag{13.21}$$

and

$$R_0'(\zeta,e^2)=(H_0-\zeta-U_0'(\zeta,e^2))^{-1}. \tag{13.22}$$

Hence

$$R(\zeta,e)=\sum_{m\geqslant 0}(-e)^m((R_0'(\zeta,e^2)H_1)^mR_0'(\zeta,e^2))_{\mathrm{sk}}. \tag{13.23}$$

We can also write this in the form

$$R(\zeta,e)=R_0'(\zeta,e^2)-R_0'(\zeta,e^2)T_0'(\zeta,e)R_0'(\zeta,e^2), \tag{13.24}$$

with

$$T_0'(\zeta,e)=\sum_{m\geqslant 0}(-e)^{m+1}((H_1R_0'(\zeta,e^2))^mH_1)_{\mathrm{sk}}. \tag{13.25}$$

We note that if we rewrite the Hamiltonian H in the form

$$H=(H_0-U_0'(\zeta,e^2))+(eH_1+U_0'(\zeta,e^2)), \tag{13.26}$$

then we can regard H as a *perturbation* of $(H_0-U_0'(\zeta,e^2))$ by $(eH_1+U_0'(\zeta,e^2))$. The effect of the correction $-U_0'(\zeta,e^2)$ to H_0 is to replace $R_0(\zeta)=(H_0-\zeta)^{-1}$ by $R_0'(\zeta,e^2)=(H_0-\zeta-U_0'(\zeta,e^2))^{-1}$ in the Fredholm–Born series for $R(\zeta,e)$; and the effect of the correction $+U_0'(\zeta,e^2)$ to eH_1 is to remove all terms with nonskeleton graphs from this series for $R(\zeta,e)$.

Finally, we note that $V_0(\zeta,e^2)$, and hence $U_0(\zeta,e^2)=(H_0-\zeta)V_0(\zeta,e^2)$, depends *analytically* upon ζ for ζ in $\rho(H_0)$ and analytically upon e^2 for $e^2\ll 1$. In particular,

$$V_0(\zeta,e^2)=\sum_{k>0}\frac{e^{2k}}{k!}V_0^{(k)}(\zeta), \tag{13.27}$$

where

$$V_0^{(k)}(\zeta)=(d/de^2)^k V_0(\zeta,e^2)|_{e^2=0} \tag{13.28}$$

is the sum of all products of order k in e^2 in the definition of V_0. Similarly,

$$U_0(\zeta,e^2)=\sum_{k>0}\frac{e^{2k}}{k!}U_0^{(k)}(\zeta), \tag{13.29}$$

where

$$U_0^{(k)}(\zeta)=(H_0-\zeta)V_0^{(k)}(\zeta). \tag{13.30}$$

This means that we can regard the operator $H_0-\zeta-U_0(\zeta,e^2)$ as an *analytic perturbation* of the operator $H_0-\zeta$ for fixed ζ in $\rho(H_0)$.

XIV. THE SPECTRUM OF THE FULL HAMILTONIAN $H=H(\theta)$

Now we can determine the spectrum of the dilated Hamiltonian $H(\theta)-\zeta$ for fixed dilation angle θ and complex energy ζ, with $0<\theta<\pi/2$ and $\zeta\epsilon\rho(H_0)$.

We begin with the free Hamiltonian operator $H_0-\zeta$ in matrix form:

$$H_0-\zeta=\begin{pmatrix}(H_0-\zeta)_{00} & 0 & 0 & \cdots\\ 0 & (H_0-\zeta)_{11} & 0 & \cdots\\ 0 & 0 & (H_0-\zeta)_{22} & \cdots\\ \vdots & \vdots & \vdots & \ddots\end{pmatrix}, \tag{14.1}$$

where $(H_0-\zeta)_{qq}$ acts on the space $\mathfrak{h}_q$ as multiplication by the free energy operator:

$$\begin{aligned}(H_0-\zeta)_{qq}|\epsilon,\omega_1,\ldots,\omega_q\rangle&=(H_{\rm el}+H_{\rm ph}-\zeta)_{qq}|\epsilon,\omega_1,\ldots,\omega_q\rangle\\&=(\mu+\epsilon+q\nu+\omega_1+\cdots+\omega_q-\zeta)|\epsilon,\omega_1,\ldots,\omega_q\rangle.\end{aligned} \tag{14.2}$$

It follows that the spectrum of $(H_0-\zeta)_{00}=(H_{\rm el}-\zeta)_{00}$, acting on $\mathfrak{h}_0$, consists of the discrete simple eigenvalues $\{\mu+\epsilon_m-\zeta\}$, together with the continuous half-line $L_{\mu-\zeta}$ (see Fig. 2), where

$$L_{\mu-\zeta}=\{(\mu-\zeta+e^{-1\theta}\epsilon(p)):0\leqslant p<\infty\}. \tag{14.3}$$

Similarly, the spectrum of $(H_0-\zeta)_{qq}=(H_{\rm el}+H_{\rm ph}-\zeta)_{qq}$, acting on $\mathfrak{h}_q$ for $q>0$, has no discrete eigenvalues but consists only of the continuous half-lines $L_{\mu+\epsilon_m+q\nu-\zeta}$ and $L_{\mu+q\nu-\zeta}$, where

$$L_{\mu+\epsilon_m+q\nu-\zeta}=\{(\mu+\epsilon_m+q\nu-\zeta+e^{-i\theta}(\omega_1+\cdots+\omega_q)):0\leqslant\omega_i<\infty\}, \tag{14.4}$$

and similarly for $L_{\mu+q\nu-\zeta}$. The spectrum of $H_0-\zeta$, of course, is the union of the spectra of all the $(H-\zeta)_{qq}$ (see Fig. 4 for the case $\nu=0$, and Fig. 6 for the case $\nu>0$).

If we now consider the operator $T(H_0-\zeta)$, obtained by applying the transfer mapping T to $H_0-\zeta$, we see that, since T adds one more free photon line to each Feynman diagram, $T(H_0-\zeta)$, acting on $\mathfrak{h}_{q+1}=\mathfrak{h}_q\otimes\mathfrak{h}_{\rm ph}$, has the form

$$T(H_0-\zeta)_{(q+1)(q+1)}=(H_0-\zeta)_{qq}\otimes I+I\otimes(H_{\rm ph}=(H_0-\zeta)_{(q+1)(q+1)}), \tag{14.5}$$

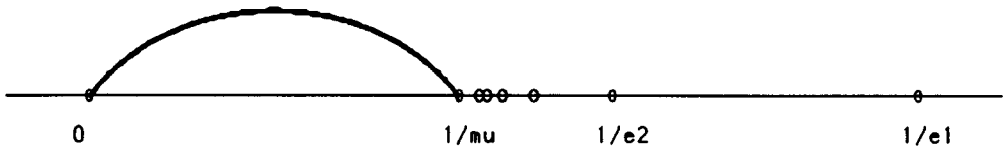

FIG. 10. Spectrum of the resolvent $R_{el}(0)=(H_{el})^{-1}$ when $\theta=\pi/3$.

so

$$T(H_0-\zeta)=\begin{pmatrix} 0 & 0 & 0 & \cdots \\ 0 & (H_0-\zeta)_{11} & 0 & \cdots \\ 0 & 0 & (H_0-\zeta)_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{14.6}$$

Hence the spectrum of $T(H_0-\zeta)_{(q+1)(q+1)}$, acting on $\mathfrak{h}_{q+1}$, has no discrete eigenvalues and consists only of the continuous half-lines $L_{\mu+\epsilon_m+(q+1)\nu-\zeta}$ and $L_{\mu+(q+1)\nu-\zeta}$. The spectrum of $T(H_0-\zeta)$ then consists of the union of all these half-lines.

The resolvent $R_0(\zeta)=(H_0-\zeta)^{-1}$ has the form

$$R_0(\zeta)=\begin{pmatrix} R_0(\zeta)_{00} & 0 & 0 & \cdots \\ 0 & R_0(\zeta)_{11} & 0 & \cdots \\ 0 & 0 & R_0(\zeta)_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{14.7}$$

where $R_0(\zeta)_{00}=(H_0-\zeta)_{00}^{-1}$. It follows by the spectral mapping theorem (see Reed and Simon,[12] Vol. IV, p. 109) that the spectrum of the resolvent $R_0(\zeta)_{00}=(H_0-\zeta)_{00}^{-1}$ for $\zeta\epsilon\rho(H_0)$ consists of the discrete eigenvalues $\{(\mu+\epsilon_m-\zeta)^{-1}\}$ together with the continuous circular arc $C_{\mu-\zeta}$, where

$$C_{\mu-\zeta}=(L_{\mu-\zeta})^{-1}=\{(\mu-\zeta+e^{-i\theta}\epsilon(p))^{-1}:0\leqslant p<\infty\}. \tag{14.8}$$

(See Fig. 10.)

Similarly, the spectrum of $(R_0(\zeta))_{qq}=(H_0-\zeta)_{qq}^{-1}$, acting on $\mathfrak{h}_q$ for $q>0$, has no discrete eigenvalues but consists of the continuous circular arcs $C_{\mu+\epsilon_m+q\nu-\zeta}$ and $C_{\mu+q\nu-\zeta}$, where

$$C_{\mu+\epsilon_m+q\nu-\zeta}=(L_{\mu+\epsilon_m+q\nu-\zeta}))^{-1}=\{(\mu+\epsilon_m+q\nu-\zeta+e^{-i\theta}(\omega_1+\cdots\omega_q))^{-1}:0\leqslant\omega_i<\infty\}, \tag{14.9}$$

and similarly for $C_{\mu+q\nu-\zeta}$. The spectrum of $R_0(\zeta)$ is then the union of the spectra of the $(R_0)_{qq}$ (see Fig. 11 for the case $\nu=0$, and Fig. 12 for the case $\nu>0$).

We note that this spectrum lies on or inside the quarter circle $S=\{z:|z|=(\epsilon_1-\zeta)^{-1}$, $\mathrm{Re}(z)\geqslant 0$, $\mathrm{Im}(z)\geqslant 0$, and so is bounded by this quarter circle.

Similarly, the operator $T(R_0(\zeta))=T((H_0-\zeta)^{-1})=(T(H_0-\zeta))^{-1}$ has the form

$$T(R_0(\zeta))=\begin{pmatrix} 0 & 0 & 0 & \cdots \\ 0 & R_0(\zeta)_{11} & 0 & \cdots \\ 0 & 0 & R_0(\zeta)_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{14.10}$$

It follows that the spectrum of $T(R_0(\zeta))$ has no eigenvalues, but consists only of the continuous circular arcs $C_{\mu+\epsilon_m+q\nu-\zeta}$ and $C_{\mu+q\nu-\zeta}$.

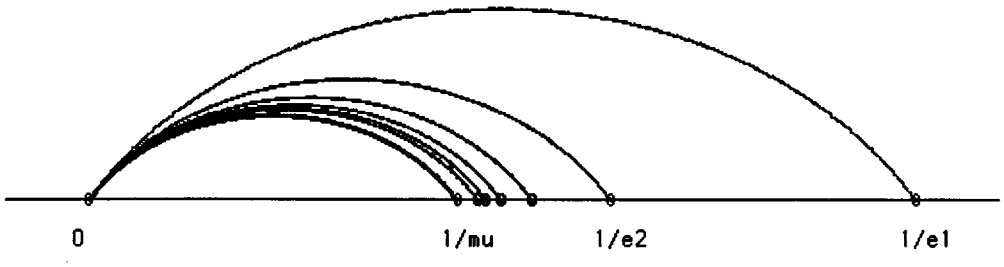

FIG. 11. Spectrum of the resolvent $R_0(0)=(H_0)^{-1}$ when $\theta=\pi/3$.

Now we consider the spectrum of the truncated resolvent operator $R(\zeta)_{00}=F_0R(\zeta)F_0$ acting on the space $\mathfrak{K}_0=\mathfrak{h}_0$:

$$R(\zeta)_{00}=F_0R(\zeta)F_0=F_0(R_0(\zeta)+M(\zeta)R_0(\zeta))F_0=R_0(\zeta)_{00}+M(\zeta)_{00}R_0(\zeta)_{00}$$
$$=R_0(\zeta)_{00}+C(\zeta)_{00}R_0(\zeta)_{00}. \tag{14.11}$$

This calculation shows that $R(\zeta)_{00}$ is a perturbation of $R_0(\zeta)_{00}$ by the compact operator $C(\zeta)_{00}R_0(\zeta)_{00}$. Standard perturbation theory for bounded operators (see Appendix A) now tells us that, for sufficiently small e, the continuous spectrum of $R_0(\zeta)_{00}$ remains unchanged, while the point spectrum may move, expand, or contract. In our case, for sufficiently small e, the continuous circular arc $C_{\mu-\zeta}$ of the spectrum of $R_0(\zeta)_{00}$ remains unchanged, while the discrete simple eigenvalues $(\mu+\epsilon_m-\zeta)^{-1}$ all migrate to nearby positions, which we choose to write as $(\mu+\epsilon_m^{(0)}-\zeta)^{-1}$ (see Fig. 13). We know (see Appendix A) that the eigenvalues $\mu+\epsilon_m^{(0)}=\mu+\epsilon_m^{(0)}\times(e^2)$ all depend analytically on e^2, with $\mu+\epsilon_m^{(0)}(0)=\mu+\epsilon_m$. Thus the spectrum of $R(\zeta)_{00}$ consists of the simple eigenvalues $(\mu+\epsilon_m^{(0)}-\zeta)^{-1}$ together with the circular arc $C_{\mu-\zeta}$. No other eigenvalues appear in the point spectrum of $R(\zeta)_{00}$, at least for sufficiently small values of e (see Appendix B). As e^2 increases, of course, these eigenvalues may disappear and other eigenvalues may appear; we are interested here only in small values of e^2.

We saw in the last section that on $\mathfrak{K}_0=\mathfrak{h}_0$ the operator $R^{(0)}(\zeta)=R(\zeta)_{00}=F_0R(\zeta)F_0$ has the form

$$R^{(0)}(\zeta)=(H^{(0)}-\zeta)^{-1}=((H_0-\zeta-U_0)_{00})^{-1}, \tag{14.12}$$

with $U_0=U_0(\zeta)=(H_0-\zeta)V_0(\zeta)$ and $H^{(0)}-\zeta=H_0-\zeta-U_0$. The spectral mapping theorem (see Ref. 12, Vol. IV, p. 109) now tells us that the spectrum of the operator $H^{(0)}-\zeta=(H_0-\zeta-U_0)_{00}$, acting on $\mathfrak{K}_0=\mathfrak{h}_0$, consists of the simple eigenvalues $\mu+\epsilon_m^{(0)}-\zeta$, together with the continuous half-line $L_{\mu-\zeta}$ (see Fig. 14).

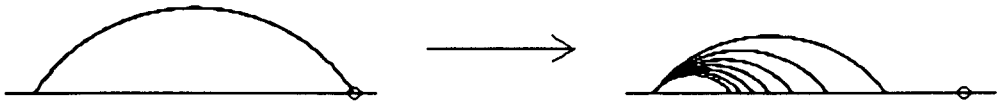

FIG. 12. Effect of replacing $\nu=0$ by $\nu>0$.

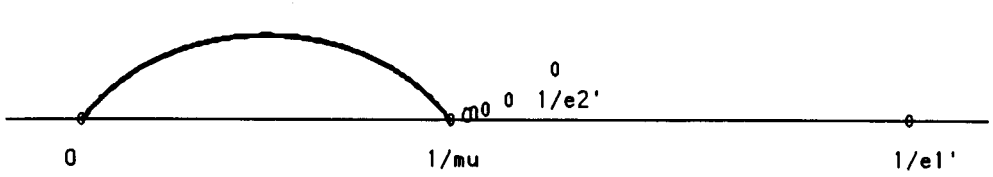

FIG. 13. Spectrum of the resolvent $R^{(0)}(0)=(H^{(0)})^{-1}$ when $\theta=\pi/3$.

From this result we see that the spectrum of the operator $(T(H_0^{(0)}-\zeta))_{11}=((H_0^{(0)}-\zeta)\otimes I+I\otimes(H_{\rm ph})_{11}$, acting on $\mathfrak{h}_1=\mathfrak{h}_0\otimes\mathfrak{h}_{\rm ph}$, consists of the sum of the spectra of $H_0^{(0)}-\zeta$ acting on $\mathfrak{h}_0$ and $H_{\rm ph}$ acting on the one-photon space $\mathfrak{h}_{\rm ph}$:

$$\sigma(T(H_0^{(0)}-\zeta))=\sigma(H_0^{(0)}-\zeta)+\sigma(H_{\rm ph}). \tag{14.13}$$

Thus the spectrum of $T(H_0^{(0)}-\zeta)_{11}$ consists of the half-lines $L_{\mu+\epsilon_m^{(0)}+\nu-\zeta}$ together with the half-line $L_{\mu+\nu-\zeta}$.

It follows that the spectrum of the operator

$$H_0^{(1)}-\zeta=(I-T)(H_0-\zeta)+T(H_0^{(0)}-\zeta)=\begin{pmatrix}(H_0-\zeta)_{00} & 0\\ 0 & T(H_0^{(0)}-\zeta)_{11}\end{pmatrix}, \tag{14.14}$$

acting on the subspace $\mathfrak{K}_1$, consists of the union of the eigenvalues $\mu+\epsilon_m-\zeta$ of $(H_0-\zeta)_{00}$ acting on $\mathfrak{h}_0$ and the half-lines $L_{\mu+\epsilon_m^{(0)}+\nu-\zeta}$ and $L_{\mu+\nu-\zeta}$ of $T(H_0^{(0)}-\zeta)_{11}$ acting on $\mathfrak{h}_1$.

By the spectral mapping theorem again, the spectrum of its inverse

$$(H_0^{(1)}-\zeta)^{-1}=\begin{pmatrix}(H_0-\zeta)_{00}^{-1} & 0\\ 0 & T(H_0^{(0)}-\zeta)_{11}^{-1}\end{pmatrix}, \tag{14.15}$$

or

$$R_0^{(1)}(\zeta)=\begin{pmatrix}R_0(\zeta)_{00} & 0\\ 0 & T(R_0^{(0)}(\zeta))_{11}\end{pmatrix}, \tag{14.16}$$

consists of the eigenvalues $(\mu+\epsilon_m-\zeta)^{-1}$, and the circular arcs $C_{\mu+\epsilon_m^{(0)}+\nu-\zeta}$ and $C_{\mu+\nu-\zeta}$.

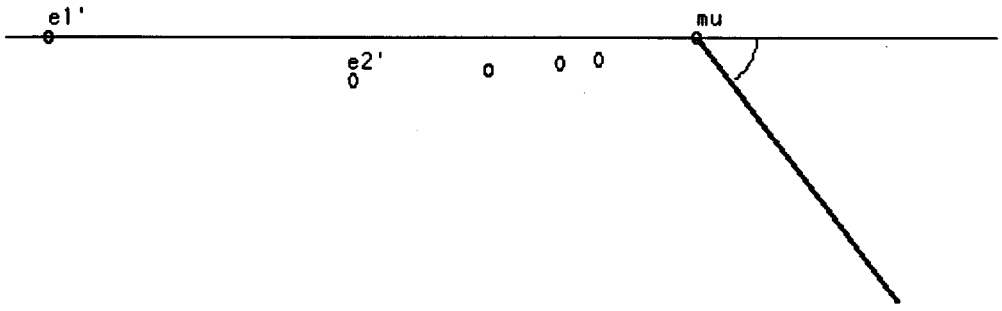

FIG. 14. Spectrum of the operator $H^{(0)}$ when $\theta=\pi/3$.

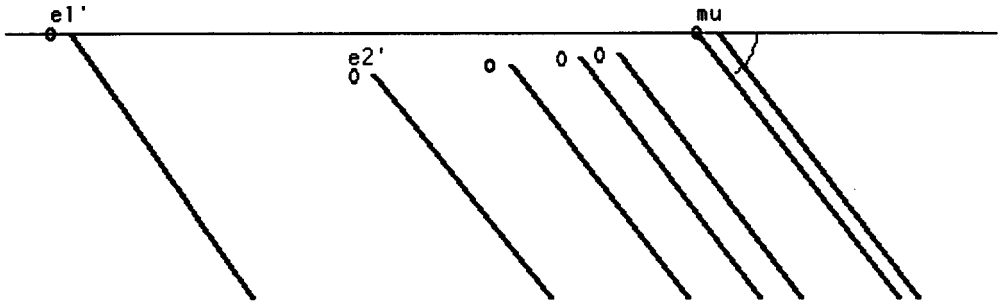

FIG. 15. Spectrum of the operator $H^{(1)}$ when $\theta=\pi/3$.

Now the truncated resolvent operator $F_1R(\zeta)F_1$, acting on $\mathfrak{K}_1$, is a compact perturbation of $(H^{(1)}-\zeta)^{-1}$:

$$\begin{aligned} F_1R(\zeta)F_1 &= F_1(R_0^{(1)}(\zeta)+C(\zeta)R_0(\zeta))F_1 \\ &= \begin{pmatrix} R_0(\zeta) & 0 \\ 0 & T(R_0^{(0)}(\zeta)) \end{pmatrix} + \begin{pmatrix} C_{00}(\zeta)R_0\zeta & C_{01}(\zeta)R_0(\zeta) \\ C_{10}(\zeta)R_0(\zeta) & C_{11}(\zeta)R_0(\zeta) \end{pmatrix}. \end{aligned} \tag{14.17}$$

It follows that the spectrum of $F_1R(\zeta)F_1$ acting on $\mathfrak{K}_1$ consists of the perturbed simple eigenvalues, which we write as $(\mu+\epsilon_m^{(1)}-\zeta)^{-1}$, together with the unperturbed circular arcs $C_{\mu+\epsilon_m^{(0)}+\nu-\zeta}$ and $C_{\mu+\nu-\zeta}$.

By the spectral mapping theorem again, the spectrum of its inverse $H^{(1)}-\zeta=(F_1R(\zeta)F_1)^{-1}$, acting on $\mathfrak{K}_1$:

$$H^{(1)}-\zeta=\begin{pmatrix} H_0-\zeta & 0 \\ 0 & T(H_0^{(0)}-\zeta) \end{pmatrix}+\begin{pmatrix} 0 & eH_1^- \\ eH_1^+ & U_1^{(2)} \end{pmatrix}, \tag{14.18}$$

consists of the eigenvalues $\mu+\epsilon_m^{(1)}-\zeta$ together with the half-lines $L_{\mu+\epsilon_m^{(0)}+\nu-\zeta}$ and $L_{\mu+\nu-\zeta}$ (see Fig. 15). Here, we write U_1 for

$$U_1=(H_0-\zeta)V_1=(H_0-\zeta)(V_1^{(2)}+V_1^{(0)})=U_1^{(2)}+U_1^{(0)}=U_1^{(2)}+T(U_0), \tag{14.19}$$

where $U_1^{(2)}=(H_0-\zeta)V_1^{(2)}$ is the connected part, and $U_1^{(0)}=T(U_0)$ is the disconnected part, of U_1.

Each of the functions $\epsilon_m^{(1)}=\epsilon_m^{(1)}(e^2)$ here is again an analytic function of e^2, with $\epsilon_m^{(1)}(0)=\epsilon_m$. They are possibly different from the functions $\epsilon_m^{(0)}(e^2)$ described above. In any case, $\mu+\epsilon_m^{(1)}$ is a singularity of the truncated resolvent operator $F_1R(\zeta)F_1$ at $\zeta=\mu+\epsilon_m^{(1)}$, just as $\mu+\epsilon_m^{(0)}$ is a singularity of the truncated resolvent operator $F_0R(\zeta)F_0$ at $\zeta=\mu+\epsilon_m^{(0)}$. But since

$$F_0R(\zeta)F_0=F_0(F_1R(\zeta)F_1)F_0, \tag{14.20}$$

we see that these singularities must agree, i.e., that $\epsilon_m^{(1)}(e^2)=\epsilon_m^{(0)}(e^2)$. Hereafter, we shall write this common value as $\epsilon_m'=\epsilon_m'(e^2)$.

Now we can proceed by induction on the photon number q. In the induction step we calculate the spectra of the relevant operators in the following order:

$$\sigma(H^{(q)}-\zeta)\to\sigma(T(H^{(q)}-\zeta))\to\sigma(H_0^{(q+1)}-\zeta)\to\sigma(R_0^{(q+1)}(\zeta))\to\sigma(R^{(q+1)}(\zeta))\to\sigma(H^{(q+1)}-\zeta). \tag{14.21}$$

Suppose we have shown that the spectrum of the operator $H^{(q)}-\zeta$, acting on the subspace $\mathfrak{K}_q$, consists of the eigenvalues $\mu+\epsilon_m^{(q)}-\zeta$ together with the half-lines $L_{\mu+\epsilon_m^{(q-j)}+j\nu-\zeta}$ and $L_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q$. Then the spectrum of the operator $T(H^{(q)}-\zeta)$, acting on the subspace $\mathfrak{K}_{(q+1)}$, consists of the eigenvalue 0, together with the half-lines $L_{\mu+\epsilon_m^{(q-j)}+(j+1)\nu-\zeta}$ and $L_{\mu+(j+1)\nu-\zeta}$ for $0\leqslant j\leqslant q$, and so the spectrum of the operator $H_0^{(q+1)}-\zeta=(I-T)(H_0-\zeta)+T(H^{(q)}-\zeta)$, acting on $\mathfrak{K}_{(q+1)}$, consists of the eigenvalues $\mu+\epsilon_m-\zeta$ together with the half-lines $L_{\mu+\epsilon_m^{(q+1-j)}+j\nu-\zeta}$ and $L_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q+1$.

By the spectral mapping theorem we know that the spectrum of its inverse $R_0^{(q+1)}(\zeta)=(H_0^{(q+1)}-\zeta)^{-1}$ consists of the eigenvalues $(\mu+\epsilon_m-\zeta)^{-1}$ together with the circular arcs $C_{\mu+\epsilon_m^{(q+1-j)}+j\nu-\zeta}$ and $C_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q+1$. Since the bounded operator $R^{(q+1)}(\zeta)=(H^{(q+1)}-\zeta)^{-1}$, which has the form

$$R^{(q+1)}(\zeta)=F_{q+1}R(\zeta)F_{q+1}=F_{q+1}(R_0(\zeta)+M(\zeta)R_0(\zeta))F_{q+1}$$
$$=R_0^{(q+1)}(\zeta)+F_{q+1}C(\zeta)R_0(\zeta)F_{q+1}, \tag{14.22}$$

is a compact perturbation of the bounded operator $R_0^{(q+1)}(\zeta)=(H_0^{(q+1)}-\zeta)^{-1}$, it follows that the spectrum of $R^{(q+1)}(\zeta)=(H^{(q+1)}-\zeta)^{-1}$ consists of the eigenvalues $(\mu+\epsilon_m^{(q+1)}-\zeta)^{-1}$ together with the circular arcs $C_{\mu+\epsilon_m^{(q+1-j)}+j\nu-\zeta}$ and $C_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q+1$. Finally, by the spectral mapping theorem again, the spectrum of the operator $H^{(q+1)}-\zeta$ acting on $\mathfrak{K}_{q+1}$ consists at least of the eigenvalues $\mu+\epsilon_m^{(q+1)}-\zeta$ together with the half-lines $L_{\mu+\epsilon_m^{(q+1-j)}+j\nu-\zeta}$ and $L_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q+1$.

This establishes the desired result, that the spectrum of $H^{(q+1)}$ acting on $\mathfrak{K}_{q+1}$ consists of the eigenvalues $\{\mu+\epsilon_m^{(q+1)}-\zeta\}$ together with the half-lines $L_{\mu+\epsilon_m^{(q+1-j)}+j\nu-\zeta}$ and $L_{\mu+j\nu-\zeta}$ for $1\leqslant j\leqslant q+1$, and thereby completes the induction argument.

Finally, we note that, as in the case $q=0$, $\mu+\epsilon_m^{(q+1)}$ is a singularity of $F_{q+1}R(\zeta)F_{q+1}$ at $\zeta=\mu+\epsilon_m^{(q+1)}$, while $\epsilon_m^{(q)}$ is a singularity of $F_qR(\zeta)F_q$ at $\zeta=\mu+\epsilon_m^{(q)}$. Since

$$F_qR(\zeta)F_q=F_q(F_{q+1}R(\zeta)F_{q+1})F_q, \tag{14.23}$$

it follows that $\epsilon_m^{(q+1)}=\epsilon_m^{(q)}$, and so, by the induction argument, $\epsilon_m^{(q+1)}=\epsilon_m'$.

Now we turn to the spectrum of the full resolvent operator $R(\zeta,e^2)$. We begin by estimating the norm of the difference between $R(\zeta,e^2)$ and the truncated resolvent $R^{(q)}(\zeta,e^2)$. Since

$$R(\zeta,e^2)-R^{(q)}(\zeta,e^2)=\sum_{r\vee s>q}R(\zeta,e^2)_{sr}, \tag{14.24}$$

where $r\vee s=\max\{r,s\}$, we have

$$\|R(\zeta,e^2)-R^{(q)}(\zeta,e^2)\|\leqslant 2\sum_{s>q}\sum_{0\leqslant r\leqslant s}\|R(\zeta,e^2)_{sr}\|\leqslant 2\sum_{s>q}\sum_{0\leqslant r\leqslant s}\|(MR_0(\zeta))_{sr}\|$$
$$\leqslant 2\sum_{s>q}\|(MR_0(\zeta))_{ss}\|+2\sum_{s>q}\sum_{0\leqslant r<s}\|(MR_0(\zeta))_{sr}\|$$
$$\leqslant\frac{4eb}{(1-eb)^3}\sum_{s>q}\frac{(\epsilon_1-\zeta)^{1/2}}{(\epsilon_1+s\nu-\zeta)^{3/2}}\leqslant\frac{4eb}{(1-eb)^3}\frac{2(\epsilon_1-\zeta)^{1/2}}{\nu(\epsilon_1+q\nu-\zeta)^{1/2}}. \tag{14.25}$$

Now we see that as $q\to\infty$, $R^{(q)}(\zeta,e^2)\to R(\zeta,e^2)$ in the operator norm.

From this result we can conclude (See appendix A) that the spectrum of the full resolvent $R(\zeta)$ consists of the union of the spectra of the truncated resolvents $R^{(q)}(\zeta)$ (see Fig. 16). It

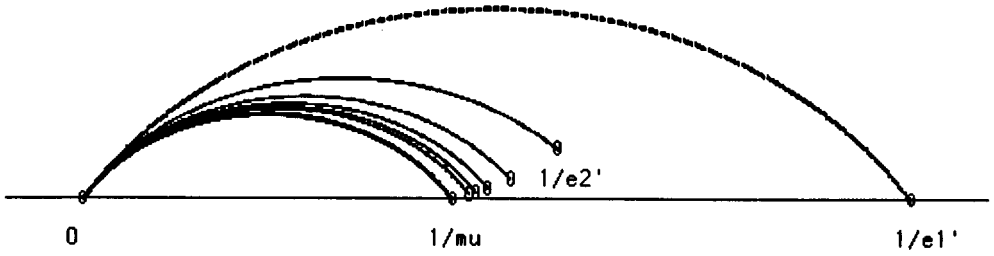

FIG. 16. Spectrum of the resolvent $R(0)=(H)^{-1}$ when $\theta=\pi/3$.

follows from the spectral mapping theorem once more that the spectrum $\sigma(H-\zeta)$ of the full Hamiltonian $H-\zeta$ consists of the union of the spectra of the partial Hamiltonians $H^{(q)}-\zeta$; i.e., it consists of the eigenvalues $\mu+\epsilon'_m-\zeta$, together with the half-lines $L_{\mu+\epsilon'_m+k\nu-\zeta}$, for $0\leqslant k<\infty$, and the half-lines $L_{\mu+k\nu-\zeta}$, for $0\leqslant k<\infty$ (see Fig. 17).

At this point we recall that the spectrum $\sigma(H)$ of the Hamiltonian H is the singular set $\mathrm{sing}(R(\zeta,e))$ of the resolvent $R(\zeta,e)$ regarded as a function of ζ. If we introduce the operators

$$H'_0=H'_0(\zeta,e^2)=(I-T)^{-1}(H_0-U_0(\zeta,e^2))$$
$$=\begin{pmatrix} H_0-U_0(\zeta,e^2) & 0 & 0 & \cdots \\ 0 & T(H_0-U_0) & 0 & \cdots \\ 0 & 0 & T^2(H_0-U_0) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{14.26}$$

and

$$R'_0(\zeta,e^2)=(H'_0(\zeta,e^2)-\zeta))^{-1}=(I-T)^{-1}(H_0-\zeta-U_0(\zeta,e^2))$$
$$=\begin{pmatrix} (H_0-U_0-\zeta)^{-1} & 0 & 0 & \cdots \\ 0 & (T(H_0-U_0-\zeta))^{-1} & 0 & \cdots \\ 0 & 0 & (T^2(H_0-U_0-\zeta))^{1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{14.27}$$

then we can conclude that the singular sets of $R'_0(\zeta,e^2)$ and $R(\zeta,e)$ as functions of ζ are the same:

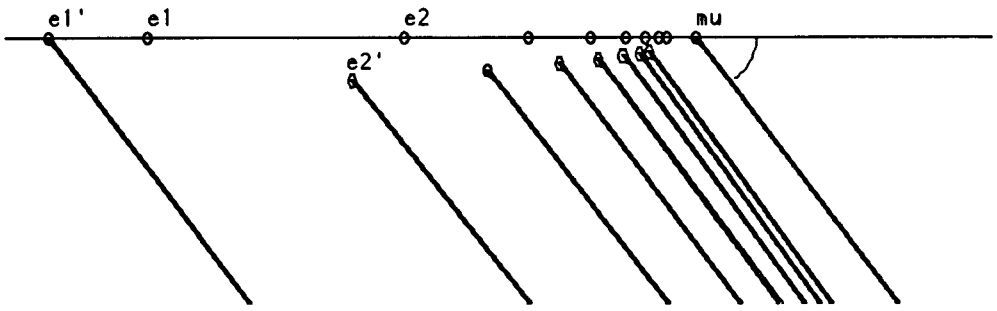

FIG. 17. Spectrum of the full Hamiltonian H when $\theta=\pi/3$.

$$\mathrm{sing}(R_0'(\zeta,e^2))=\mathrm{sing}(R(\zeta,e))=\sigma(H). \tag{14.28}$$

We shall use this fact in Sec. XV.

XV. ANALYSIS OF THE EIGENVALUES ϵ_m'

We have seen that the simple eigenvalues $\mu+\epsilon_m'=\mu+\epsilon_m'(e^2)$, together with the mass values μ and ν, completely determine, at least for small values of e^2, the spectrum of the full Hamiltonian H. In this section we examine the dependence of these simple eigenvalues on e^2. We show that, at least for small e^2, each of them is an analytic function of e^2, and we calculate the first few coefficients in the associated Taylor series expansion. In particular, we show that the leading second-order coefficient agrees with that given by Weisskopf and Wigner[5] and by Dirac,[6] as expected.

We begin by taking $e\ll 1$, and forming the projection operator

$$P(e)=\frac{-1}{2\pi i}\int_{|\mu+\epsilon_m-\eta|=r} R(\eta,e)\,d\eta, \tag{15.1}$$

where $\mu+\epsilon_m$ is the mth simple eigenvalue of H_0, and r is chosen as half the distance of $\mu+\epsilon_m$ from the rest of the spectrum of H_0. If $|\epsilon_m\rangle$ is the eigenfunction of H_0 associated with $\mu+\epsilon_m$, then

$$|\epsilon_m'\rangle=P(e)|\epsilon_m\rangle \tag{15.2}$$

is the corresponding eigenfunction of H associated with the eigenvalue $\mu+\epsilon_m'$, and

$$\mu+\epsilon_m'=\frac{\langle\epsilon_m|HP|\epsilon_m\rangle}{\langle\epsilon_m|P|\epsilon_m\rangle}=\frac{\langle\epsilon_m|(H_0+eH_1)P|\epsilon_m\rangle}{\langle\epsilon_m|P|\epsilon_m\rangle}=\mu+\epsilon_m+\frac{\langle\epsilon_m|(eH_1)P(e)|\epsilon_m\rangle}{\langle\epsilon_m|P(e)|\epsilon_m\rangle}. \tag{15.3}$$

Now

$$\begin{aligned}\langle\epsilon_m|P(e)|\epsilon_m\rangle&=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|R(\eta,e)_{00}|\epsilon_m\rangle\,d\eta\\&=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|(H_0-\eta-U_0(\eta,e^2))^{-1}|\epsilon_m\rangle\,d\eta,\end{aligned} \tag{15.4}$$

while

$$\begin{aligned}\langle\epsilon_m|(eH_1)P(e)|\epsilon_m\rangle&=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|(eH_1R(\eta,e))_{00}|\epsilon_m\rangle\,d\eta\\&=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|U_0(\eta,e^2)(H_0-\eta-U_0(\eta,e^2))^{-1}|\epsilon_m\rangle\,d\eta.\end{aligned} \tag{15.5}$$

We note here that $\langle\epsilon_m|R_0(\eta)|\epsilon_m\rangle$ is singular at $\eta=\mu+\epsilon_m$, but $\langle\epsilon_m|U_0(\eta,e^2)|\epsilon_m\rangle$ is regular there. Hence

$$\langle\epsilon_m|P(e)|\epsilon_m\rangle=\sum_{k\geqslant 0} b_k e^{2k}, \tag{15.6}$$

with

$$b_k=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|R_0(\eta)B_k(\eta)|\epsilon_m\rangle d\eta, \tag{15.7}$$

where

$$B_k(\eta)=\frac{(d/de^2)^k}{k!}\,(I-U_0(\eta,e^2)R_0(\eta))^{-1}\Big|_{e^2=0}$$

$$=\sum_{1\leqslant j\leqslant k}(\partial/\partial e_1^2+\cdots+\partial/\partial e_j^2)^k(U_0(\eta,e_1^2)R_0(\eta)\cdots U_0(\eta,e_j^2)R_0(\eta))\Big|_{e_i^2=0}. \tag{15.8}$$

In particular,

$$b_0=1,\quad b_1=0,$$

$$b_2=-\sum_{n\neq m}\frac{\langle\epsilon_m|U_0^{(1)}(\epsilon_m)|\epsilon_n\rangle\langle\epsilon_n|U_0^{(1)}(\epsilon_m)|\epsilon_m\rangle}{(\epsilon_n-\epsilon_m)^2}, \tag{15.9}$$

$$b_3=\text{etc.}$$

Similarly,

$$\langle\epsilon_m|(eH_1)P(e)|\epsilon_m\rangle=\sum_{k>0}a_ke^{2k}, \tag{15.10}$$

where

$$a_k=\frac{-1}{2\pi i}\int_{|\epsilon_m-\eta|=r}\langle\epsilon_m|B_{k+1}(\eta)|\epsilon_m\rangle d\eta, \tag{15.11}$$

with the same operator function $B_k(\eta)$. In particular,

$$a_1=-\langle\epsilon_m|U_0^{(1)}(\epsilon_m)|\epsilon_m\rangle,$$

$$a_2=-\langle\epsilon_m|U_0^{(2)}(\epsilon_m)|\epsilon_m\rangle-\sum_{n\neq m}\frac{\langle\epsilon_m|U_0^{(1)}(\epsilon_m)|\epsilon_n\rangle\langle\epsilon_n|U_0^{(1)}(\epsilon_m)|\epsilon_m\rangle}{(\epsilon_n-\epsilon_m)}, \tag{15.12}$$

$$a_3=-\langle\epsilon_m|U_0^{(3)}(\epsilon_m)|\epsilon_m\rangle-\sum_{i+j=3}\sum_{n\neq m}\frac{\langle\epsilon_m|U_0^{(i)}(\epsilon_m)|\epsilon_n\rangle\langle\epsilon_n|U_0^{(j)}(\epsilon_m)|\epsilon_m\rangle}{(\epsilon_n-\epsilon_m)}$$

$$-\sum_{n\neq m\neq p}\frac{\langle\epsilon_m|U_0^{(1)}\epsilon_n\rangle\langle\epsilon_n|U_0^{(1)}\epsilon_p\rangle\langle\epsilon_p|U_0^{(1)}\epsilon_n\rangle}{(\epsilon_n-\epsilon_m)(\epsilon_p-\epsilon_m)}-2\sum_{n\neq m}\frac{\langle\epsilon_m|U_0^{(1)}\epsilon_n\rangle\langle\epsilon_n|U_0^{(1)}\epsilon_m\rangle\langle\epsilon_n|U_0^{(1)}\epsilon_n\rangle}{(\epsilon_n-\epsilon_m)^2},$$

$$a_4=\text{etc.}$$

Combining these results, we find that

$$\epsilon_m'(e^2)=\epsilon_m+\sum_{k>0}\frac{c_ke^{2k}}{k!}, \tag{15.13}$$

with

$$c_1 = a_1 = \langle \epsilon_m | U_0^{(1)}(\epsilon_m) | \epsilon_m \rangle,$$

$$c_2 = a_2 = -\langle \epsilon_m | U_0^{(2)}(\epsilon_m) | \epsilon_m \rangle - \sum_{n \neq m} \frac{\langle \epsilon_m | U_0^{(1)}(\epsilon_m) | \epsilon_n \rangle \langle \epsilon_n | U_0^{(1)}(\epsilon_m) | \epsilon_m \rangle}{(\epsilon_n - \epsilon_m)},$$

$$\begin{aligned} c_3 = a_3 - a_1 b_2 = &-\langle \epsilon_m | U_0^{(3)}(\epsilon_m) | \epsilon_m \rangle - \sum_{i+j=3} \sum_{n \neq m} \frac{\langle \epsilon_m | U_0^{(i)}(\epsilon_m) | \epsilon_n \rangle \langle \epsilon_n | U_0^{(j)}(\epsilon_m) | \epsilon_m \rangle}{(\epsilon_n - \epsilon_m)} \\ &- \sum_{n \neq m \neq p} \frac{\langle \epsilon_m | U_0^{(1)} \epsilon_n \rangle \langle \epsilon_n | U_0^{(1)} \epsilon_p \rangle \langle \epsilon_p | U_0^{(1)} \epsilon_m \rangle}{(\epsilon_n - \epsilon_m)(\epsilon_p - \epsilon_m)} \\ &- \sum_{n \neq m} \frac{\langle \epsilon_m | U_0^{(1)} \epsilon_n \rangle \langle \epsilon_n | U_0^{(1)} \epsilon_m \rangle \langle \epsilon_m | U_0^{(1)} \epsilon_m \rangle}{(\epsilon_n - \epsilon_m)^2}, \end{aligned} \tag{15.14}$$

$$c_4 = a_4 - a_1 b_3 - a_2 b_2 = \text{etc.}$$

The leading term in this expansion for small e^2 is:

$$\epsilon'_m = \langle \epsilon_m | U_0^{(1)}(\zeta_0) | \epsilon_m \rangle = -\int \int \frac{\langle \epsilon_m | H_1 | \epsilon, \omega \rangle \langle \epsilon, \omega | H_1 | \epsilon_m \rangle}{(\mu + \epsilon + \nu + \omega) - \zeta_0} \, d\rho(\epsilon) d\omega. \tag{15.15}$$

This result agrees with that of Weisskopf and Wigner[5] and that of Dirac.[6] It is not hard to see that when $\nu < \epsilon_2 - \epsilon_1$ these leading terms, except for ϵ_1, all have negative real and imaginary parts, so that, for small e^2, the values of $\mu + \epsilon_m(e^2)$ are all displaced to the left and down (i.e., southwest in the complex energy plane) from the values of $\zeta_0 = \mu + \epsilon_m$. The exception ϵ_1 has no imaginary part, and so the ground state $\mu + \epsilon_1$ is displaced to the left but not down (i.e., west but not south).

The implicit function theorem ensures that the power series for the function $\epsilon_m(e^2)$ converges for all sufficiently small e^2, and that for such e^2 the results described above are valid. It will be possible, however, to continue the function $\zeta(e^2) = \mu + \epsilon_m(e^2)$ analytically in e^2 so long as the condition

$$eb'(\zeta) < 1 \tag{15.16}$$

is met. This condition will be violated whenever the point $\zeta(e^2)$ gets too close to the rest of the spectrum of H_0, so that $\delta'(\zeta)$ gets too small.

It can be shown, at least for small e^2, that the associated eigenvectors $|\epsilon'_m\rangle$ of the operator $H - \zeta - U_0(\zeta, e^2)$ are almost orthogonal, and that together they span the same space as do the eigenvectors $|\epsilon_m\rangle$ of the operator H_0. For these eigenvectors, of course, we have, when $\zeta = \mu + \epsilon'_m$,

$$(H_0 - U_0(\mu + \epsilon'_m, e^q)) | \epsilon'_m \rangle = (\mu + \epsilon'_m) | \epsilon'_m \rangle. \tag{15.17}$$

We note here that the expression for ϵ'_m always contains the factor $(\lambda_m)^{1+2\gamma_m}$ [see Eq. (6.5)], where

$$\lambda_m^2 = \mu^2 - \epsilon_m^2, \tag{15.18}$$

and

$$\gamma^2 = \kappa_m^2 - e^2. \tag{15.19}$$

But $\kappa_m^2 \geqslant 1$, while $e^2 = (137)^{-1} \ll 1$, so certainly $\gamma_m \geqslant 1 - e^2/2$. It follows that

$$\epsilon'_m(e^2) = O(\lambda_m^{1+2\gamma_m}) = O(1/m^{3-e^2}). \tag{15.20}$$

Since

$$\mu + \epsilon'_m(0) = \mu + \epsilon_m = \mu\left[1 - \frac{e^2}{2m^2} + O\left(\frac{e_0^4}{m^4}\right)\right], \tag{15.21}$$

we see that for small e^2 we must have

$$\mu + \epsilon'_m(\epsilon^2) = \mu\left[1 - \frac{e^2}{2m^2} + O\left(\frac{e^2}{m^{3-e^2}}\right)\right], \tag{15.22}$$

so that the relative positions and spacings of the eigenvalues ϵ_m and $\epsilon'_m(e^2)$ remain nearly the same for small e^2.

Finally, we note that each eigenvalue $\epsilon'_m(e^2)$ is independent of the dilation angle θ (see Reed and Simon,[12] Vol. IV, pp. 184 ff), so long as θ is large enough that this eigenvalue is exposed, i.e., so long as

$$-\theta < \arg(\epsilon_2(e^2) - \epsilon_1(e^2)). \tag{15.23}$$

As the dilation angle θ decreases from θ_0 to 0, however, these eigenvalues are covered in succession, beginning with $\epsilon'_2(e^2)$, by the half-line $[\mu + \epsilon'_1, \mu + \epsilon'_1 + e^{-i\theta}\infty)$, and they pass to another sheet of the complex energy plane. The corresponding eigenvectors $|\epsilon'_m(e^2)\rangle$ pass out of the state space $\mathfrak{h}_0$. As θ returns to 0, the only eigenvalue remaining on the principal sheet of the energy plane is the ground state energy eigenvalue $\epsilon'_1(e^2)$.

XVI. ANALYSIS OF THE SCATTERING MATRIX

In this section we consider the scattering of a single photon by the hydrogen atom. The measurable physical quantities for such processes are all contained in the *scattering matrix* $S(E,e)$. The scattering matrix elements are usually given by a formula of the form (see Friedrichs[13])

$$\langle E'|S(E,e)|E\rangle = \lim_{\epsilon\to 0}\ \lim_{\substack{s\to+\infty \\ t\to-\infty}} \langle E'|e^{-iH_0 s}e^{i(H+i0)(s-t)}e^{iH_0 t}|E\rangle = \delta(E'-E)\langle E'|1 - T_0(E+i0,e)|E\rangle, \tag{16.1}$$

where $T_0(\zeta,e)$ is the T-matrix introduced in Sec. VII:

$$T_0(\zeta,e) = eH_1 - e^2 H_1 R(\zeta,e) H_1 = \sum_{k\geqslant 0} e^{k+1}((-H_1 R_0(\zeta))^k H_1), \tag{16.2}$$

and where we choose $|E\rangle = |\epsilon_1, \omega_1, \ldots, \omega_p\rangle$, $|E'\rangle = |\epsilon_1, \omega'_1, \ldots, \omega'_q\rangle$ as eigenvectors of H_0. If we make this choice, then the evaluation of the limit

$$\langle E'|T_0(E+i0)|E\rangle = \lim_{\epsilon\to 0}\langle E'|T_0(E+i\epsilon)|E\rangle \tag{16.3}$$

in the above definition is plagued with multiple singularities in the expansion of the T_0-matrix elements. This behavior can be traced to the fact that, as we know, the operators H_0 and H have different spectra on $\mathfrak{h}$, so that the resolvents $R_0(\zeta)$ and $R(\zeta,e)$ have different singular sets as functions of ζ.

If we replace the Hamiltonian $H_0-\zeta$ with the modified Hamiltonian $H_0'(\zeta,e^2)$ and the resolvent $R_0(\zeta)$ by the modified resolvent $R_0'(\zeta,e^2)$ [see Eq. (13.22)] in the definition of the scattering matrix:

$$\langle E'|S'(E,e)|E\rangle=\lim_{\epsilon\to 0}\lim_{\substack{s\to+\infty\\ t\to-\infty}}\langle E'|e^{-i(H_0-U_0')s}e^{i(H+U_0'+i\epsilon)(s-t)}e^{-i(H_0-U_0')t}|E\rangle$$
$$=\delta(E'-E)\langle E'|1-T'(E+i0,e)|E\rangle, \tag{16.4}$$

where now

$$U_0'=U_0'(E,e^2),$$

$$T'(E+i0,e)=(eH_1+U_0')-(eH_1+U_0')R(E+i0,e)(eH_1+U_0')$$
$$=\sum_{k\geqslant 0}e^{k+1}((-H_1R_0'(E,e^2))^kH_1)_{\rm sk}, \tag{16.5}$$

and if we choose $|E\rangle=|\epsilon_1'(e^2),\omega_1,\ldots,\omega_p\rangle$, $|E'\rangle=|\epsilon_1'(e^2),\omega_1',\ldots,\omega_q'\rangle$ as eigenvectors of H_0-U_0, then the evaluation of the limit

$$\langle E'|T'(E+i0,e)|E\rangle=\lim_{\epsilon\to 0}\langle E'|T'(E+i\epsilon,e)|E\rangle \tag{16.6}$$

can be obtained explicitly. In fact, if we write

$$T(E+i0,e)=\sum_{k>0}e^{k+1}((-H_1R_0(E+i0,e^2))^kH_1)_{\rm sk} \tag{16.7}$$

where the sum here is over skeleton graphs only, then we shall see directly that there are no singularities at the energies $E=\mu+\epsilon_1'(e^2)+\omega$, and the terms of the sum can all be evaluated there. This follows from the fact that the operators $R(\zeta,e)$ and $R_0'(\zeta,e^2)$ have the same singular set as functions of ζ.

The evaluation of the denominators $R_0(E+i0,e^2)$ can be facilitated here by dilating all the internal energy variables ϵ'',ω'', but not the external variables ω_i,ω_j' through the dilation angle $\theta=\pi/3$, as in the previous sections. (It can be shown, by the same methods as those used in Sec. II, that the interaction terms H_1 involving both internal and external variables remain bounded by the same bounds if only the internal energy variables are dilated. See Prosser.[11])

In particular, for the scattering of a single photon of energy ω by an atom in the ground state, we have

$$|E\rangle=|\epsilon_1'(e^2),\omega\rangle,\quad |E'\rangle=|\epsilon_1'(e^2),\omega'\rangle, \tag{16.8}$$

and

$$\langle E'|S'(E,e)|E\rangle=\delta(\omega'-\omega)(1-\langle\epsilon_1'(e^2),\omega'|T'(\mu+\epsilon_1'(e^2)+\nu+\omega+i0,e)|\epsilon_1'(e^2),\omega\rangle). \tag{16.9}$$

As a function of the photon energy ω, the T'-matrix element

$$\langle\epsilon_1',\omega|T'(E,e)|\epsilon_1',\omega\rangle=\langle\epsilon_1',\omega|H_1R(E,e)H_1|\epsilon_1',\omega\rangle \tag{16.10}$$

is analytic throughout the same complex energy domain as is $R_0'(E,e)$ (see Fig. 16). In particular, it will have a singularity whenever

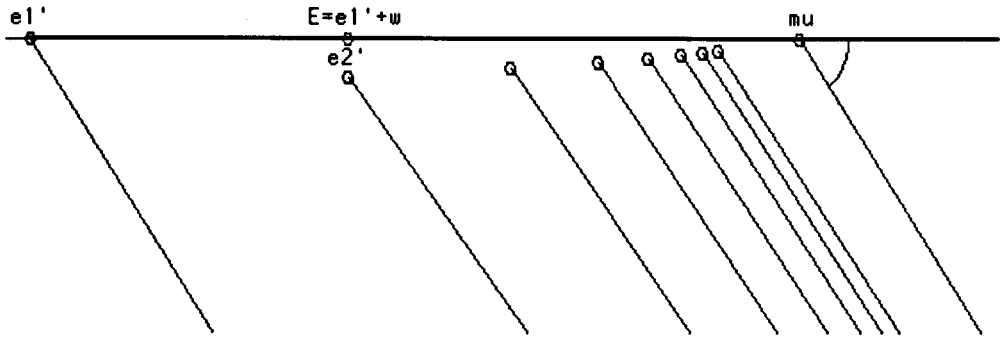

FIG. 18. Resonance values of the energy $E=\epsilon_1'+\omega$.

$$E'=\mu+\epsilon_1'+\nu+\omega=\epsilon_m', \tag{16.11}$$

i.e., whenever

$$\nu+\omega=\epsilon_m'-\epsilon_1'. \tag{16.12}$$

Near these energies E' the T-matrix element above will be large. In particular, it will be large when the photon energy ω has any of the real values (see Fig. 18)

$$\nu+\omega=\mathrm{Re}(\epsilon_m'-\epsilon_1'). \tag{16.13}$$

These values comprise the scattering resonances for this process.

We can now associate the real part $\mathrm{Re}(\epsilon_m'-\epsilon_1')$ of the energy difference $\epsilon_m'-\epsilon_1'$ with the *position* of the resonance, and the imaginary part $\mathrm{Im}(\epsilon_m'-\epsilon_1')$ with the *breadth* of the resonance, in the usual way.

We note that the presence of the branch points $\mu+\epsilon_m'+q\nu$ for $q>0$ in the spectrum of H do not cause the same sort of resonance behavior in the scattering matrix elements, because, while the resolvent $R_0'(\zeta,e^2)$ blows up as $\zeta\to\mu+\epsilon_m'$, it remains bounded as $\zeta\to\mu+\epsilon_m'+q\nu$ for all $q>0$ (see Appendix B).

A similar analysis, of course, can be made for resonance scattering processes involving more photons.

ACKNOWLEDGMENT

This work was supported by ARPA Grant No. F49620-93-1-0567.

APPENDIX A: PERTURBATION THEORY

In this section we summarize those elements of perturbation theory which are used in Secs. XIV and XV.

Essential spectra

These results rely on a theorem of Weyl:[14] If R and S are two bounded operators on a Hilbert space, with $R-S$ compact, such that the resolvent set $\rho(R)$ of R is dense and connected in the complex plane, and $\rho(S)\cap\rho(R)$ is nonempty, then the essential spectra of R and S must be the same: $\sigma_{\mathrm{ess}}(A)=\sigma_{\mathrm{ess}}(B)$ (See Reed and Simon,[13] Vol. IV, p. 111).

Now suppose A and $A(e)=A+C(e)$ are closed operators with resolvents $R(\zeta)$ and $R(\zeta,e)$. Suppose that the resolvent set $\rho(A)$ of A is dense and connected in the complex plane and includes

the half-plane $\mathrm{Re}(\zeta)\leqslant 0$. Suppose that the difference $A(e)-A=C(e)$ is relatively compact; i.e., that $R(\zeta)C(e)$ is compact for $\mathrm{Re}(\zeta)<0$. Then $R(\zeta)$ and $R(\zeta,e)$ are bounded operators with $R(\zeta,e)-R(\zeta)=R(\zeta)C(e)R(\zeta,e)$ compact, and $\zeta\epsilon p(A)\cap\rho(A(e))$ if $\zeta\ll 0$. By the spectral mapping theorem, the spectrum of $R(\zeta)$ is

$$\sigma(R(\zeta))=\{(\lambda-\zeta)^{-1}:\lambda\epsilon\sigma(A)\}, \tag{A1}$$

and similarly for the spectrum of $R(\zeta,e)$. Hence the resolvents sets of these operators are dense and connected, with nonempty intersection. By Weyl's theorem, the essential spectra of $R(\zeta)$ and $R(\zeta,e)$ must be the same. By the spectral mapping theorem again, the spectra of A and $A(e)$ must also be the same.

Point spectra

As for the point spectra of these operators, we know that under any relatively bounded analytic perturbation $A(e)=A+C(e)$ of A, with $R(\zeta)C(e)$ bounded for $\zeta\leqslant 0$ and with $C(0)=0$, each isolated simple eigenvalue ϵ will, for small values of e, move analytically to a new location $\epsilon(e)$. We shall see that for larger values of e some of these eigenvalues may disappear, and new eigenvalues may appear.

We know that, under the analytic perturbation $A(e)=A+C(e)$ with $C(0)=0$, a simple eigenvalue ϵ of A with normalized eigenfunction $|\epsilon\rangle$ will, for small e, move analytically to an eigenvalue $\epsilon(e)$ of $A(e)$ with normalized eigenfunction $|\epsilon(e)\rangle$ according to the formulas:

$$\epsilon(e)=\frac{\langle\epsilon|A(e)P(e)|\epsilon\rangle}{\langle\epsilon|P(e)|\epsilon\rangle}, \tag{A2}$$

$$|\epsilon(e)\rangle=\frac{|P(e)\epsilon\rangle}{\|P(e)|\epsilon\rangle\|}. \tag{A3}$$

Here $P(e)$ is the projection onto the eigenfunction $|\epsilon(e)\rangle$:

$$P(e)=\frac{-1}{2\pi i}\int_{|\zeta-\epsilon|=r}R(\zeta,e)d\zeta \tag{A4}$$

which is constructed from the resolvent $R(\zeta,e)$ of $A(e)$. The radius r of the integration path is chosen so that the circle $|\zeta-\epsilon|=r$ encloses only the eigenvalue ϵ, and no other point, in the spectrum of A. We also know that $R(\zeta,e)$ satisfies the resolvent equation:

$$R(\zeta,e)=R(\zeta)+R(\zeta,e)C(e)R(\zeta)=(I-R(\zeta)C(e))^{-1}R(\zeta). \tag{A5}$$

If e is so small that $\|R(\zeta)C(e)\|=b<1$ when $|\zeta-\epsilon|=r$, then

$$\|R(\zeta,e)-R(\zeta)\|\leqslant\frac{b/r}{1-b}. \tag{A6}$$

It follows from Eq. (A4) that

$$\|(P(e)-P)|\epsilon\rangle\|\leqslant\frac{b}{1-b}. \tag{A7}$$

Since

$$\langle\epsilon|P(e)|\epsilon\rangle=\langle\epsilon|P|\epsilon\rangle-\langle\epsilon|(P(e)-P)|\epsilon\rangle=1-\langle\epsilon|(P(e)-P)|\epsilon\rangle, \tag{A8}$$

we have

$$1-\frac{b}{1-b}\leqslant|\langle\epsilon|P(e)|\epsilon\rangle|\leqslant 1+\frac{b}{1-b}. \tag{A9}$$

It follows that, for any ζ with $|\zeta-\epsilon|=r$,

$$|\epsilon(e)-\epsilon|=\frac{|\langle\epsilon|(A(e)-A)P(e)|\epsilon\rangle|}{|\langle\epsilon|P(e)|\epsilon\rangle|}=\frac{|\langle\epsilon|C(e)P(e)|\epsilon\rangle|}{|\langle\epsilon|P(e)|\epsilon\rangle|}=\frac{|\langle\epsilon|(A-\zeta)(R(\zeta)C(e))P(e)|\epsilon\rangle|}{|\langle\epsilon|P(e)|\epsilon\rangle|}$$

$$\leqslant\frac{rb}{1-2b}\leqslant r \quad \text{if } b<1/3. \tag{A10}$$

The formula for the eigenfunction $\epsilon(e)$ shows that it can be expanded as a quotient if two analytic functions:

$$\epsilon(e)=\epsilon+\frac{\Sigma_{n=0}^{\infty}a_n(e)}{\Sigma_{n=0}^{\infty}b_n(e)}, \tag{A11}$$

with

$$a_n(e)=\frac{(-1)^{n+1}}{2\pi i}\int_{|\zeta-\epsilon|=r}\langle\epsilon|(A-\zeta)(R(\zeta)C(e))^{n+1}|\epsilon\rangle d\zeta, \tag{A12}$$

$$b_n(e)=\frac{(-1)^{n+1}}{2\pi i}\int_{|\zeta-\epsilon|=r}\langle\epsilon|(C(e)R(\zeta))^{n}|\epsilon\rangle d\zeta. \tag{A13}$$

These formulas lead to those in the text [see Eq. (15.3)ff.].

Comparison theorems

Now suppose that $A(e)=A+C(e)$ and $A'(e)=A+C'(e)$ are two relatively bounded analytic perturbations of A, with eigenvalues $\epsilon(e)$ and $\epsilon'(e)$ and eigenfunctions $|\epsilon(e)\rangle$ and $|\epsilon'(e)\rangle$. If we write $K(e)=R(\zeta)C(e)$ and $K'(e)=R(\zeta)C(e)$, and suppose as above that $\|K(e)\|\leqslant b$, $\|K'(e)\|\leqslant b$ with $b<1$, then

$$\epsilon(e)-\epsilon'(e)=\frac{\langle\epsilon|C(e)P(e)|\epsilon\rangle}{\langle\epsilon|P(e)|\epsilon\rangle}-\frac{\langle\epsilon|C'(e)P'(e)|\epsilon\rangle}{\langle\epsilon|P'(e)|\epsilon\rangle}$$

$$=\frac{a}{c}-\frac{a'}{c'}=\frac{ac'-a'c}{cc'}=\frac{a(c'-c)}{cc'}+\frac{(a-a')c}{cc'}. \tag{A14}$$

Now

$$c=|\langle\epsilon|P(e)|\epsilon\rangle|\geqslant 1-\frac{b}{1-b}=\frac{1-2b}{1-b}, \tag{A15}$$

and similarly for c', while

$$|c-c'|=|\langle\epsilon|(P(e)-P'(e))|\epsilon\rangle|\leqslant\|(I-K(e))^{-1}(K(e)-K'(e))(I-K'(e))^{-1}\|$$

$$\leqslant\frac{1}{(1-b)^2}\|K(e)-K'(e)\|. \tag{A16}$$

Moreover,

$$a=|\langle\epsilon|C(e)P(e)|\epsilon\rangle|\leqslant\frac{rb}{1-b},\tag{A17}$$

and similarly for a', while

$$\begin{aligned}|a-a'|&=|\langle\epsilon|(C(e)P(e)-C'(e)P'(e))|\epsilon\rangle|\\&\leqslant|\langle\epsilon|(H_0-\zeta)(K(e)P(e)-K'(e)P'(e))|\epsilon\rangle|\\&\leqslant r\|(I-K(e))^{-1}(K(e)-K'(e))(I-K'(e))^{-1}\|\leqslant\frac{r}{(1-b)^2}\|K(e)-K'(e)\|.\end{aligned}\tag{A18}$$

Hence

$$\begin{aligned}|\epsilon(e)-\epsilon'(e)|&\leqslant\frac{r(b+1)\|K(e)-K'(e)\|}{(1-b)^3}\,\frac{(1-b)^2}{(1-2b))^2}\\&\leqslant\frac{r(1+b)}{(1-2b)^3}\|K(e)-K'(e)\|\\&\leqslant 18r\|K(e)-K'(e)\|,\quad\text{if}\quad b<1/3.\end{aligned}\tag{A19}$$

Limits

Now suppose $A_n(e)=A+C_n(e)$ is a sequence of relatively bounded analytic perturbations of A, with $R(\zeta)C_n(e)\to R(\zeta)C(e)$ in operator norm as $n\to\infty$. Then $A(e)=A+C(e)$ is again a relatively bounded analytic perturbation of A. Moreover, if $\|R(\zeta)C(e)\|\leqslant b<1/3$ for all ζ with $|\zeta-\epsilon|=r$, then (by adjusting b if necessary) the same will be true for $\|R(\zeta)C(e)\|$ for all sufficiently large n. Hence, by the argument above,

$$|\epsilon(e)-\epsilon_n(e)|\leqslant 18r\|R(\zeta)C(e)-R(\zeta)C_n(e)\|.\tag{A20}$$

Moreover, this holds uniformly in e for all e such that $\|R(\zeta)C(e)\|\leqslant b<1/3\|$. We can conclude that, under these conditions, $\epsilon_n(e)\to\epsilon(e)$ as $n\to\infty$.

Disappearance of old eigenvalues

So long as e remains small, the (analytic) trajectory of the eigenvalue $\epsilon_m(e)$ starts at ϵ_m and remains disjoint the rest of the spectrum $\sigma(H)$ of H. So long as the trajectory of $\epsilon_m(e)$ remains disjoint from the rest of $\sigma(H)$, it can be continued analytically as e increases. For larger values of e, however, the trajectory of $\epsilon_m(e)$ may encounter another part of the spectrum of H, and then disappear behind that part of the spectrum onto another sheet of the complex energy plane. Or the trajectory may move off to ∞ and disappear there. The actual situation depends on the dilation angle θ as well as on the size of e.

Appearance of new eigenvalues

There is also the possibility that new eigenvalues (we call these *spurious eigenvalues*) may appear. Any spurious eigenvalue appearing under a relatively compact perturbation $\mu(e)$ must also be of finite multiplicity, and must also depend analytically on e for $e>0$. Moreover, it must either disappear or tend to a limit $\mu(0)$ as $e\to 0$. If it does not disappear before e gets to zero, then the limiting value $\mu(0)$, since it cannot lie in the resolvent set $\rho(H_0)$ of H_0, must lie somewhere

in the spectrum $\sigma(H_0)$ of H_0. It cannot lie on the interior of any branch cut, since the branch cuts are dependent on θ, while the eigenvalues are not (see Reed and Simon,[13] Vol. IV, Sec. XIII). Hence $\mu(e)$ must lie at one of the branch points $\epsilon_m + q\nu$, or it else it must lie at the branch point at ∞. We shall show in Appendix B that these spurious eigenvalues can be ruled out, at least for small values of e.

APPENDIX B: SPURIOUS EIGENVALUES

In this Appendix we show that the appearance of spurious eigenvalues in the spectrum of the operator $H^{(q)}$ can be ruled out, at least for sufficiently small values of e.

Our proof proceeds by induction. Recall that $H^{(q)} = H_0^{(q)} + F_1(eH_1)F_1 - U_q^{(c)}$ is a relatively compact perturbation of $H_0^{(q)}$. We have seen already that there are no spurious eigenvalues in the spectrum of $H^{(0)}$. Moreover, if there are no spurious eigenvalues in the spectrum of $H^{(q-1)}$, then there are no spurious eigenvalues in the spectrum of $H_0^{(q)} = (I-T)H_0 + T(H^{(q-1)})$. Any spurious eigenvalue μ of $H^{(q)}$, then, must tend, as $e \to 0$, to one of the q half-lines $L_{\epsilon_m^{(k)} + (q-k)\nu}$, $0 \leqslant k \leqslant q-1$, which make up the spectrum of $H_0^{(q)}$. Since the eigenvalues are independent of θ while the branch cuts are not, this means that μ must tend either to the branch point at $\epsilon_m^{(k)} + (q-k)\nu$ or to the branch point at ∞. These are the possibilities which we want to rule out.

We begin by recalling that

$$H^{(q)} = H_0^{(q)} + F_1(eH_1)F_1 - U_q^{(2q)} = (I-T)H_0 + T(H^{(q-1)}) + F_1(eH)F_1 - U_q^{(2q)}. \tag{B1}$$

If we iterate this relation, we find

$$\begin{aligned} H^{(q)} &= ((I-T^k)H_0 + T^k(H^{(q-k)}) + F_k(eH_1)F_k - (U_q^{2(q-k+1)} + \cdots + U_q^{(2q)}) \\ &= H_k^{(q)} + F_k(eH_1)F_k - (U_q^{2(q-k+1)} + \cdots + U_q^{(2q)}). \end{aligned} \tag{B2}$$

Thus $H^{(q)}$ is a bounded perturbation of $H_k^{(q)} = ((I-T^k)H_0 + T^k(H^{(q-k)})$, for any k, $1 \leqslant k \leqslant q$.

Now we fix k, $1 \leqslant k \leqslant q$, and we surround the branch cut $L_{\epsilon_m^{(q-k)} + k\nu}$ with a simple curve Γ_k consisting of the two half-lines $L_{\epsilon_m^{(q-k)} + k\nu - r}$ and $L_{\epsilon_m^{(q-k)} + k\nu + r}$ joined at the top by the semicircle with center at $\epsilon_m^{(q-k)} + k\nu$ and radius r. We choose r to be half the horizontal distance from the half-line $L_{\epsilon_m^{(q-k)} + k\nu}$ to the rest of the spectrum of $H_k^{(q)}$. Note that then the distance from Γ_k to $L_{\epsilon_m^{(q-k)} + k\nu}$ is just

$$\operatorname{dist}(\Gamma_k, L_{\epsilon_m^{(q-k)} + k\nu}) = r \sin\theta. \tag{B3}$$

Using the curve Γ_k, we define the projections $P_k^{(q)}$ and $P^{(q)}$ acting on the space $\mathfrak{K}_q$ as follows: For states $|f\rangle$ in the domain of $\exp(H_0)$ (and hence in the domain of $\exp(H^{(q)})$; i.e., for states of the form

$$|f\rangle = |e^{-E} g(E)\rangle = |e^{-(\epsilon + q\nu + \omega_1 + \cdots + \omega_i)\cos\theta} g(\epsilon, \omega_1, \ldots, \omega_1)\rangle, \tag{B4}$$

where $0 \leqslant i \leqslant qg$ is analytic and square integrable in the energy variables), we set

$$P_k^{(q)}|f\rangle = \frac{-1}{2\pi i} \int_{\Gamma_k} R_k^{(q)}(\eta)|f\rangle \, d\eta, \tag{B5}$$

[where $R_k^{(q)}(\eta)$ is the resolvent of $H_k^{(q)}$], and

$$P^{(q)}|f\rangle = \frac{-1}{2\pi i}\int_{\Gamma_k} R^{(q)}(\eta)|f\rangle d\eta, \tag{B6}$$

where $R^{(q)}(\eta)$ is the resolvent of $H^{(q)}$. Now

$$\|P_k^{(q)}|f\rangle\| \leqslant \left(\frac{1}{2\pi}\int_\gamma \|e^{-E}(E-\eta)^{-1}\| d\eta\right) \||g\rangle\|. \tag{B7}$$

If we write $\eta = \epsilon_m^{(q-k)} + k\nu \pm r + e^{-i\theta}\omega$, then we have

$$\|e^{-E}(E-\eta)^{-1}\| \leqslant \frac{e^{-\omega\cos\theta}}{r\sin\theta}. \tag{B8}$$

It follows that the integrals defining $P_k^{(q)}$ and $P^{(q)}$ converge strongly on all such states, and that

$$\|P_k^{(q)}|f\rangle\| \leqslant \frac{1}{\pi r\sin 2\theta}\||g\rangle\|, \tag{B9}$$

$$\|P^{(q)}|f\rangle\| \leqslant \frac{1}{\pi r\sin\theta}\frac{1}{1-eb}\||g\rangle\|, \tag{B10}$$

$$\|P^{(q)}|f\rangle - P_k^{(q)}|f\rangle\| \leqslant \frac{1}{\pi r\sin 2\theta}\frac{eb}{1-eb}\||g\rangle\|. \tag{B11}$$

The projections $P_k^{(q)}$ and $P^{(q)}$ commute with the operators $H_k^{(q)}$ and $H^{(q)}$, respectively, and project onto the subspaces $P_k^{(q)}\mathfrak{K}_q$ and $P^{(q)}\mathfrak{K}_q$ of $\mathfrak{K}_q$.

On the subspace $P_k^{(q)}\mathfrak{K}_q$, the spectrum of $H_k^{(q)}$ consists precisely of the branch cut $L_{\epsilon_m^{(q-k)}+k\nu}$. In fact, in this subspace the states are all completely disconnected, so that $P_k^{(q)} = F_k$ and $\sigma(H_k^{(q)}) = \sigma(eH_0^{(q-k)}e) + \sigma(H_{\rm ph}+\cdots+H_{\rm ph}) = L_{\epsilon_m^{(q-k)}+k\mu}$. On the subspace $P^{(q)}\mathfrak{K}_q$, the spectrum of $H^{(q)}$ consists of the same branch cut $L_{\epsilon_m^{(q-k)}+k\nu}$, together with any spurious eigenvalues μ which lie inside the curve Γ_k.

Now if $\mu = \mu(e)$ were a simple spurious eigenvalue lying inside Γ_k, and $|\mu\rangle$ were the normalized eigenfunction associated with μ, then $|\mu\rangle$ would lie in the domain of $\exp(H^{(q)})$ and hence of $\exp(H_k^{(q)})$. Hence $P^{(q)}|\mu\rangle = |\mu\rangle$ and $P_k^{(q)}|\mu\rangle = |\mu_k\rangle$ are defined by the integral formulas above, and from the estimate obtained there we have

$$\||\mu\rangle - |\mu_k\rangle\| \leqslant \frac{1}{\pi r\sin 2\theta}\frac{eb}{1-eb} = \frac{C_{15}eb}{1-eb}. \tag{B12}$$

It follows that

$$1 - \frac{C_{15}eb}{(1-eb)} \leqslant |\langle\mu_k|\mu\rangle| \leqslant 1 + \frac{C_{15}eb}{(1-eb)}. \tag{B13}$$

Now we form the operator

$$R_k^{(q)}(\mu)(H^{(q)} - H_k^{(q)}) = R_k^{(q)}(\mu)(F_k(eH_1)F_k - (U_q^{2(q-k+1)} + \cdots + U_q^{(2q)})), \tag{B14}$$

where $F_k = E_1 + \cdots + E_k$ is the projection on the space $\mathfrak{K}_k$, and we consider the inner product

$$\langle\mu_k|\mu\rangle=\langle\mu_k|R_k^{(q)}(\mu)(H_k^{(q)}-H^{(q)})|\mu\rangle=\langle\mu_k|R_k^{(q)}(\mu)(U_q^{(2q-2k)}+\cdots+U_q^{(2q)}-F_k(eH_1)F_k)|\mu\rangle. \tag{B15}$$

The left hand side here is bounded below:

$$|\langle\mu_k|\mu\rangle|\geqslant 1-\frac{C_{15}eb}{1-eb}, \tag{B16}$$

but the right-hand side is bounded above:

$$|\mathrm{RHS}|\leqslant\frac{C_{16}eb}{1-eb}. \tag{B17}$$

It follows that for e sufficiently small Eq. (B15) cannot hold; and so for e sufficiently small the spurious eigenvalue $\mu(e)$ cannot lie inside the curve Γ_k.

To get this upper bound for RHS, we first note that

$$|\langle\mu_k|R_k^{(q)}(\mu)U_q^{(2q-2i)}|\mu\rangle|\leqslant\|\langle\mu_k|\|\,\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\|\,\||\mu\rangle\|. \tag{B18}$$

But

$$\||\mu\rangle\|=1,\quad\text{and}\quad\|\langle\mu_k|\|\leqslant 1+\frac{C_{15}eb}{1-eb}. \tag{B19}$$

Moreover, we know that the components of the operator $H_k^{(q)}$ connect only the last $q-k$ of the q final photon lines (say those with energies $\omega_{k+1},\ldots,\omega_q$) in each diagram, while the components of the operator $U_q^{(2q-2i)}$ must connect the last $q-i$ of the q final photon lines. If $i<k$, then at least $k-i$ of these final photon lines (say those with energies $\omega_{i+1},\ldots,\omega_k$) must be unconnected by $H_k^{(q)}$ but connected by $U_q^{(2q-2i)}$.

If we extract from each component of $U_q^{(2q-2i)}$ the term (eH_1) which connects one of these photon lines, say the one with energy ω_k, and combine it with $R_k^{(q)}(\mu)$, we get

$$\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\|\leqslant\|P_kR_k^{(q)}(\mu)(eH_1^+)\|\left(\sum_{k=1}^{\infty}\|eK\|^{2k+1}\right)\leqslant\|P_k^{(q)}R_k^{(q)}(\mu)(eH_1^+)\|\,\frac{eb}{1-(eb)^2}$$

$$\leqslant\frac{e\|F^+\|}{|\Sigma\omega-\kappa|}\,\frac{eb}{1-(eb)^2}, \tag{B20}$$

where we write $\Sigma\omega=\omega_1+\cdots+\omega_k$ and $\mu=\epsilon_m^{(q-k)}+k\nu+e^{-i\theta}\kappa$, with $\kappa=\kappa_r+i\kappa_i$. Now $\|F^+\|$ has the form (see Sec. VI)

$$\|F^+\|\leqslant C_{16}\,\frac{j!}{(2j+1)!}\,k^{j+1}e^{-\tau\omega\cos\theta}, \tag{B21}$$

where C_{16} is another constant independent of ω and j. Since $k^2=(\omega+\nu)^2-\nu^2=\omega(\omega+2\nu)<(\omega+2\nu)^2$, we can write

$$\|F^+\|\leqslant C_{16}\,\frac{j!}{(2j+1)!}\,\omega(\omega+2\nu)^je^{-\tau\omega\cos\theta}, \tag{B22}$$

and so

$$\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\| \leqslant \frac{C_{16}j!}{(2j+1)!}\,\frac{\omega_k}{\Sigma\omega-\kappa}\,(\omega_k+2\nu)^j e^{-\tau\omega_k\cos\theta}\,\frac{(e)^2b}{1-(eb)^2}. \tag{B23}$$

Elementary calculus shows that

$$\frac{j!}{(2j+1)!}\,(\omega_k+2\nu)^j e^{-\tau\omega_k\cos\theta} \leqslant \frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j}, \tag{B24}$$

and

$$\frac{\omega_k}{|\Sigma\omega-\kappa|} \leqslant \begin{cases} \dfrac{|\kappa|}{\kappa_i} = 1/\sin\beta, & \text{if} \quad 0<|\beta|<\pi/2, \\ 1 & \text{if} \quad \pi/2\leqslant|\beta|\leqslant\pi, \end{cases} \tag{B25}$$

where

$$\beta = \arg\kappa. \tag{B26}$$

Hence

$$\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\| \leqslant \frac{C_{16}}{\sin\beta}\,\frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j}\,\frac{(e)^2b}{(1-(eb)^2)} \leqslant \frac{C_{17}(e)^2b}{1-(eb)^2}. \tag{B27}$$

We can bound

$$\|P_k^{(q)}R_k^{(q)}(\mu)F_k(eH_1)F_k\| = \|R_k^{(q)}(\mu)P_k^{(q)}F_k(eH_1)F_k\| \tag{B28}$$

in the same way. Since

$$R_k^{(q)}F_k(eH_1)F_k = E_k(eH_1)F_k = E_k(eH_1^+)E_{k-1}, \tag{B29}$$

we have, as before,

$$\|P_k^{(q)}R_k^{(q)}(\mu)F_k(eH_1)F_k\| = \|P_k^{(q)}R_k^{(q)}(\mu)(eH_1^+)\| \leqslant \frac{C_{16}e}{\sin\beta}\,\frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j} \leqslant C_{17}e. \tag{B30}$$

We note here that we can ensure that the angle $\beta=\arg\kappa=\arg(\mu-\epsilon_m^{(q-k)}-k\nu)-\theta$ is bounded away from 0, by varying θ slightly, if necessary.

We also note that in the sector with total angular momentum zero, the selection rules restrict the possible values of j for the q photons to $1\leqslant j\leqslant q$.

In the same way, we can rule out the possibility of a simple spurious eigenvalue μ which tends to the other end of the branch cut $L_{\epsilon_m^{(q-k)}+k\nu}$. We have only to rewrite Eq. (B22) in the form

$$\begin{aligned}\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\| &\leqslant \frac{j!}{(2j+1)!}\,\frac{e^{-\tau\omega_k\cos\theta}}{|\Sigma\omega-\kappa|}\,(\omega_k+2\nu)^{j+1}\,\frac{eb}{1-eb} \\ &\leqslant \frac{e^{-\tau\omega_k\cos\theta}}{|\Sigma\omega-\kappa_r-\kappa_i|}\,\frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j}\,\frac{eb}{1-eb},\end{aligned} \tag{B31}$$

and use the bound

$$\frac{e^{-\tau\omega_k\cos\theta}}{|\Sigma\omega-\kappa_r-\kappa_i|} \leqslant \frac{e^{-\tau\kappa_r\cos\theta}}{|\kappa_i|}. \tag{B32}$$

This gives the bound

$$\|P_k^{(q)}R_k^{(q)}(\mu)U_q^{(2q-2i)}\| \leqslant \frac{e^{-\tau\kappa_r\cos\theta}}{|\kappa_i|}\,\frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j}\,\frac{(e)^2 b}{1-(eb)^2}. \tag{B33}$$

Similarly,

$$\|P_k^{(q)}R_k^{(q)}(\mu)F_k(eH_1)F_k\| \leqslant \frac{e e^{-\tau\kappa_r\cos\theta}}{|\kappa_i|}\,\frac{e^{2\tau\nu\cos\theta}}{(\tau\cos\theta)^j}. \tag{B34}$$

If $\mu=\epsilon_m^{(q-r)}+k\nu+e^{-i\theta}\kappa\to\infty$ inside Γ as $e\to 0$, then $\kappa_r\to\infty$, and, as before, we can ensure that $|\kappa_i|$ is bounded away from 0 by varying θ slightly if necessary. We conclude that for e sufficiently small, (hence for κ_r sufficiently large),

$$\|P_k R_k^{(q)}(\mu)U_q^{(2q-2i)}\| \leqslant \frac{e^{-\tau\kappa_r\cos\theta}}{|\kappa_i|}\,\frac{C_{17}eb}{1-eb}, \tag{B35}$$

and hence that the spurious eigenvalue $\mu\to\infty$ cannot lie inside the curve Γ_k.

If the spurious eigenvalue μ were not simple, then in any case it would have finite multiplicity, and we could choose for $|\mu\rangle$ any normalized eigenvector of $H^{(q)}$, set $|\mu_k\rangle=P_k^{(q)}|\mu\rangle$, and proceed as before to the same conclusion.

[1] V. B. Berestetskii, E. M. Lifshitz, and L. P. Pitaevskii, *Relativistic Quantum Theory*, Part I (Pergamon, Oxford, 1973).
[2] J. Aguilar and J. M. Combes, Commun. Math. Phys. **22**, 269 (1971).
[3] E. Balslev and J. M. Combes, Commun. Math. Phys. **22**, 180 (1971).
[4] C. van Winter, J. Math. Anal. Appl. **47**, 633 (1974).
[5] V. Weisskopf and E. Wigner, Physica Acta **63**, 54 (1930).
[6] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford University Press, Oxford, 1958).
[7] H. E. Moses, Phys. Rev. A **11**, 387 (1975).
[8] H. Exton, *Multiple Hypergeometric Functions and Applications* (Horwood, Chichester, 1976).
[9] H. E. Moses and R. T. Prosser, J. Math. Phys. **33**, 1878 (1992).
[10] H. E. Moses and R. T. Prosser, J. Math. Anal. Appl. **173**, 390 (1993).
[11] L. van Hove, Proc. R. Acad. Belgium **26**, 1 (1951).
[12] M. Reed and B. Simon, *Methods of Mathematical Physics* (Academic, New York, 1978).
[13] H. Weyl, Rend. Circ. Mat. Palermo **27**, 373 (1909).

Periodic orbit theory analysis of a continuous family of quasi-circular billiards

R. W. Robinett
Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802

(Received 31 July 1997; accepted for publication 22 September 1997)

We compute the Fourier transform ($\rho(L)$) of the quantum mechanical energy level density for the problem of a particle in a two-dimensional circular infinite well (or circular billiard) as well as for several special generalizations of that geometry, namely the half-well, quarter-well, and the circular well with a thin, infinite wall along the positive x-axis (hereafter called a circular well plus baffle). The resulting peaks in plots of $|\rho(L)|^2$ versus L are compared to the lengths of the classical closed trajectories in these geometries as a simple example of the application of periodic orbit (PO) theory to a billiard or infinite well system. We then solve the Schrödinger equation for the general case of a circular well with infinite walls both along the positive x-axis and at an arbitrary angle Θ (a circular ''slice'') for which the half-well ($\Theta=\pi$), quarter-well ($\Theta=\pi/2$), and circular well plus baffle ($\Theta=2\pi$) are then all special cases. We perform a PO theory analysis of this general system and calculate $|\rho(L)|^2$ for many intermediate values of Θ to examine how the peaks in $\rho(L)$ attributed to periodic orbits change as the quasi-circular wells are continuously transformed into each other. We explicitly examine the transitions from the half-circular well to the circle plus baffle case (half-well to quarter-circle case) as Θ changes continuously from π to 2π (from π to $\pi/2$) in detail. We then discuss the general $\Theta\rightarrow 0$ limit, paying special attention to the cases where $\Theta=\pi/2^n$, as well as deriving the formulae for the lengths of closed orbits for the general case. We find that such a periodic orbit theory analysis is of great benefit in understanding and visualizing the increasingly complex pattern of closed orbits as $\Theta\rightarrow 0$. © *1998 American Institute of Physics.*
[S0022-2488(98)01301-2]

I. INTRODUCTION

There exist a wide variety of semi-classical calculational tools in quantum mechanics which help to make clear the substantial overlap between the classical and quantum mechanical descriptions of a system. Periodic orbit (PO) theory[1–5] is rapidly becoming one of the most useful semi-classical tools which can be used to make very direct connections (via a Fourier transform) between the spectrum of quantized energy levels of a quantum system (often considered as the ''smoking gun'' of wave mechanics) and the periodic orbits or closed trajectories of the corresponding classical system.

Variations of PO theory which are well-suited for spectroscopy[6] have been used to examine the structure of atoms in strong magnetic[7] and electric[8] fields and related techniques have proved useful in discussions of quantum transport.[9] Other familiar and non-trivial potentials which have proved amenable to this treatment include various harmonic-oscillator potentials[10] and especially 2- and 3-dimensional (2D and 3D) ''billiard'' systems.[1,3,5] Examples of this latter case include the exact analytic determination of the energy eigenvalues for the 1D and 3D infinite square wells,[1] the extraction of information on periodic orbits in an otherwise chaotic two-dimensional billiard system[11] as well as the general structure of the spherical and circular infinite well cases[3] and their applications as model systems to the electronic shell structure of clusters[12] and the shell structure

0022-2488/98/39(1)/278/21/$15.00
© 1998 American Institute of Physics

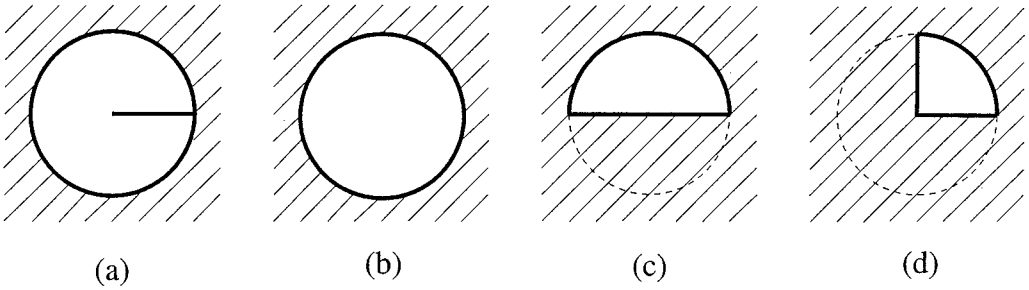

FIG. 1. Geometries for the (a) circular well plus baffle, (b) ordinary circular well, (c) half-circular well, and (d) quarter-circular well.

of heavy nuclei[13] (including the possibility of quantum tunneling[14]). A variety of two-dimensional ''billiard''-type geometries (squares, circles, circular disks) have also been studied[5,15] using these methods to investigate the semi-classical theory of orbital magnetism in the ballistic regime. Modifications of the standard formalism to include diffraction effects have also been considered.[5,16–18]

An interesting variation on the standard simple geometries is to consider billiard systems in which the billiard area or shape is allowed to vary smoothly with some parameter, call it f. One can perform a periodic orbit analysis for each value of f and then ''track'' how the quantum energy levels retain information on the shapes and lengths of the classical orbits as the potential well or billiard system is continuously deformed from one shape to another. An especially simple case, for example, is to perform a periodic orbit analysis for the rectangular box[5] as one side is made longer, starting with a square 2D infinite well potential: we will analyze just such a system (numerically) in what follows as an example.

A far more interesting and less trivial example has been recently considered by Reimann *et al.*[19] who have considered the problem of an annular infinite well (circular well of radius R with an inner circular barrier of radius fR where $0<f<1$) defined by the potential function

$$V(r)=\begin{cases}0, & \text{for } R_{min}=fR<r<R=R_{max},\\ \infty, & \text{otherwise.}\end{cases} \tag{1}$$

With the exception of two features which indicate diffraction (i.e., purely quantum or wave effects) of the particle *around* the inner annulus,[5] the remaining global features of the Fourier transform of the energy level densities are straightforwardly and accurately described by periodic orbit theory as f varies from 0 to 1.[20]

A slightly more challenging version of this problem involves the continuous deformation of ''sliced'' circular wells of varying angular sizes, Θ, as one continuously varies the opening angle Θ. In this case the potential energy function might be described via

$$V(r,\theta)=\begin{cases}0, & \text{for } 0<r<R \text{ and } 0<\theta<\Theta,\\ \infty, & \text{otherwise.}\end{cases} \tag{2}$$

Special cases are shown in Fig. 1 where the half-circular well (c) and quarter-circle well (d) can be studied using very simple symmetry arguments (as in Sec. III below) as being related to the ''ordinary'' circular well (b), which is itself not actually in the class described by Eq. (2). The other special case of a circular well plus a thin, infinite wall along the positive x-axis [case (a),

which we will call the circular well plus baffle] has not been considered before and, together with the half- and quarter-circle wells constitute special cases of a general circular ''slice'' of opening angle Θ where these special cases correspond to $\Theta = 2\pi, \pi, \pi/2$ for cases (a), (c), and (d), respectively. (We note that the circular ''slice'' systems considered here are different, in many ways much simpler, than the so-called ''wedge billiard'' system which has been extensively studied elsewhere.[21])

The purpose of this paper is then several-fold. In Sec. II, we very briefly review the main ideas of periodic orbit theory, especially as applied to 2D and 3D billiard (or infinite well) problems in very simplified form and perform simple numerical tests on a 2D rectangular billiard. Next, in Sec. III, we quickly review the periodic orbit analysis for the four related billiard systems in Fig. 1. Finally, in Sec. IV, we perform a periodic orbit theory analysis for the general circular ''slice'' case with opening angle Θ to compare how the features in the Fourier transform of the energy level spectrum ''track'' the classical closed orbits in these more general versions of the circular well. We derive a general formula for the path lengths of closed orbits in the general $\Theta \to 0$ case, paying special attention to the special cases of $\Theta = \pi/2^n$ which form ''boundaries'' at which new features begin to appear in the $\rho(L)$ spectra.

II. PERIODIC ORBIT THEORY APPLIED TO BILLIARD SYSTEMS

The main quantity of interest in periodic orbit theory is the quantum density of energy levels which is defined via

$$\rho(E) = \sum_n \delta(E - E_n), \tag{3}$$

where the summation is over all of the quantized bound state energy levels. One of the main results of periodic orbit theory[1–3,5] is that this density can be approximated semi-classically as

$$\rho(E) = \rho_0(E) + \sum_{p=1}^{\infty} \sum_{\gamma} \rho_{\gamma,p} \cos\left[p\left(\frac{S_\gamma(E)}{\hbar} - \phi_\gamma \right) \right]. \tag{4}$$

The first term, $\rho_0(E)$, is a smoothly and slowly varying function of the energy variable and will provide no information on the detailed classical orbits. The second, oscillatory term is more interesting since $S_\gamma(E)$ is the classical action corresponding to ''primitive'' (''once-around'' only) periodic orbits characterized by a label γ and $p = 1, \ldots, \infty$ runs over all possible repetitions of such trajectories. This last term arises from a path integral approach and it is thus not surprising that periodic orbits dominate since their phase contributions add coherently, period after period, while the non-periodic orbits contribute phase terms which increasingly tend to cancel over time. It is this important term which contains information on the classical closed trajectories.

Since we wish to concentrate on 2D ''billiard systems,'' we will specialize to that case immediately. For such a ''billiard,'' i.e., for a particle which is free inside an infinite well of arbitrary shape of area A and circumference L and where $\psi = 0$ on the boundary, the smooth part can be written as

$$\rho_0(E) = \frac{A}{4\pi}\left(\frac{2m}{\hbar^2} \right) - \frac{L}{8\pi}\sqrt{\frac{2m}{\hbar^2 E}} \to \frac{A}{4\pi} - \frac{L}{8\pi}\frac{1}{\sqrt{E}}, \tag{5}$$

if we set $\hbar^2/2m=1$ as is often done for simplicity. This piece may well be more familiar in three-dimensions as very similar arguments are used in the counting of the number of states in a 3D cavity typically found in explanations of black-body radiation. The systematic study of the smoothed distribution of eigenvalues for the wave equation in a bounded domain[22] and its expansion in powers of $1/k=1/\sqrt{E}$ was first carried out by Weyl and so the terms in Eq. (5) are often called the Weyl terms.

We are, of course, more interested in the oscillatory piece and for the free-particle in a 2D well the quantized energies, E_n, and ''primitive'' actions, S_γ, are given by

$$E_n=\frac{\hbar^2 k_n^2}{2m}\rightarrow k_n^2 \quad \text{and} \quad S_\gamma(E=k^2)=\hbar k L_\gamma \tag{6}$$

where the k_n are the appropriate wavenumbers and L_γ is the length of the 'primitive' periodic orbit.

Using these and Eq. (3), we then have for the equivalent density of states, but now in k-space (where we once again set $\hbar^2/2m=1$ and also ignore an unimportant overall factor of 1/2),

$$\rho(k)=\sum_n \delta(k-k_n). \tag{7}$$

The corresponding expressions for Eq. (4) in k-space are given, in turn, by

$$\rho(k)=\rho_0(k)+\sum_{p=1}^{\infty}\sum_\gamma \rho_{\gamma,p}\cos[p(kL_\gamma-\phi_\gamma)], \tag{8}$$

and the dimensions are consistent since we take, as usual, $\hbar^2/2m=1$.

We then proceed by multiplying both expressions by e^{ikL} and integrating over k to evaluate the Fourier transform of each side. For the discrete energy level term, we have

$$\rho(L)=\sum_{n=1}^{\infty}\int_{-\infty}^{+\infty}\delta(k-k_n)e^{ikL}dk=\sum_{n=1}^{\infty}e^{ik_nL}, \tag{9}$$

which can be evaluated purely by using a knowledge of the bound state energy spectrum (actually the corresponding wavenumbers, k_n, in this simplest case).

For the semi-classical oscillatory piece we have, after taking the Fourier transform,

$$\rho(L)=\sum_{p=1}^{\infty}\sum_\gamma \rho_{\gamma,p}\delta(L-pL_\gamma), \tag{10}$$

which exhibits a series of δ-function like ''spikes'' at multiples of the lengths of the ''primitive'' closed paths, i.e., at $L=pL_\gamma$. The contribution of the smooth piece, $\rho_0(k)$, yields a δ-function piece at $L=0$ which will appear in the numerical evaluation of Eq. (9), but which gives no interesting information on the periodic orbits.

Thus, if we evaluate the Fourier transform, $\rho(L)$, of the wavenumber spectrum via Eq. (9), using either theoretically calculated or experimentally determined values of the k_n for a given geometry, we should find a series of sharp peaks corresponding to the lengths of classical closed orbits. More complete analyses[5] also retain information on the prefactors ($\rho_{\gamma,p}$), but since we focus here only on the positions of the periodic orbit ''spikes'' in $\rho(L)$ and not on the ''strengths'' of the δ-function peaks, we choose to use the simple version in Eq. (9).

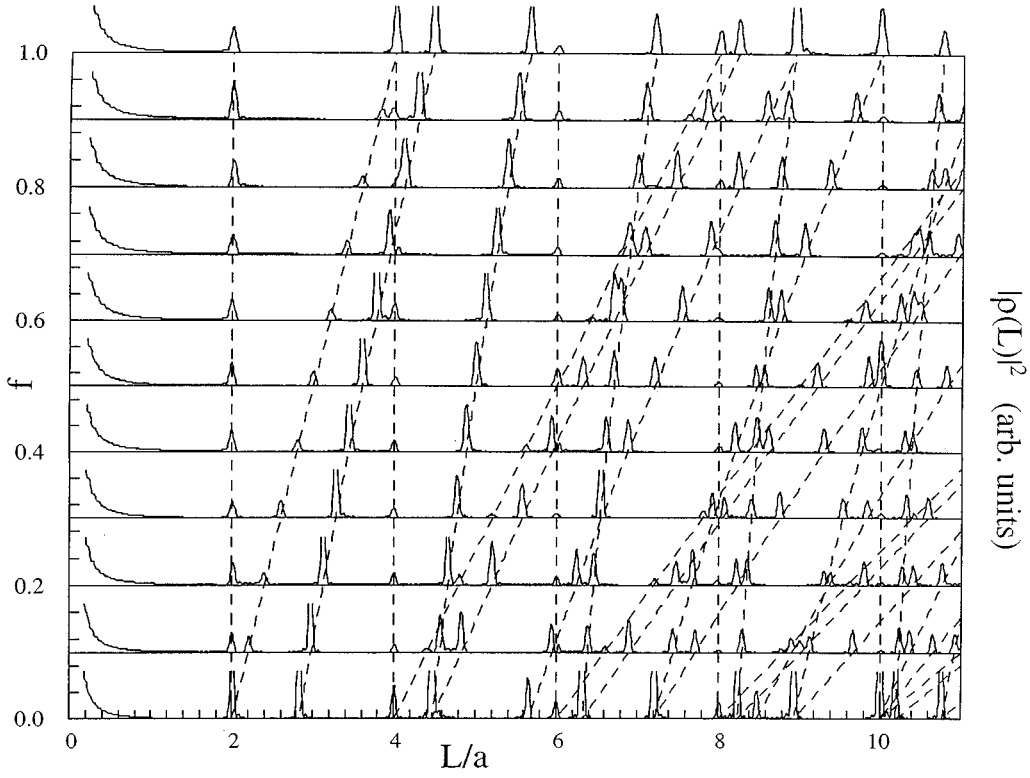

FIG. 2. Fourier transform (squared) of the wavenumber distribution, i.e., $|\rho(L)|^2$ versus L/a, for the 2D rectangular well. For each value of f shown on the left-hand axis, a ''spectrum'' corresponding to $|\rho(L)|^2$ is shown on the horizontal axis. The path length formulae of Eq. (13) are shown as dashed lines.

Because we will evaluate Eq. (9) numerically, we will necessarily use a finite number of wavevectors, namely

$$\rho_N(L)=\sum_{n=1}^{N} e^{ik_nL}, \tag{11}$$

so that the ''experimentally'' determined spikes in $\rho_N(L)$ will have a finite width. The resolution effects (due to using only a finite summation) have been discussed by Balian and Bloch.[3]

As a test of the simplified version of PO theory implied by Eq. (11), we analyze a 2D rectangular billiard with sides $L_x=a$ and $L_y=(1+f)a$ where we allow f to vary in the range $0 \leqslant f \leqslant 1$. (This system is discussed in detail in Ref. 5 where a more complete analytic approach is developed: we study it here in this more approximate formulation as a check on its numerical accuracy.) The quantized wavenumbers are trivially given by

$$k_{(i,j)}=\pi\sqrt{\frac{i^2}{L_x^2}+\frac{j^2}{L_y^2}}=\frac{\pi}{a}\sqrt{i^2+j^2/(1+f)^2}, \tag{12}$$

while the lengths of periodic orbits are equally simply given by the formula

$$L(i,j;f)=2a\sqrt{i^2+(1+f)^2j^2}. \tag{13}$$

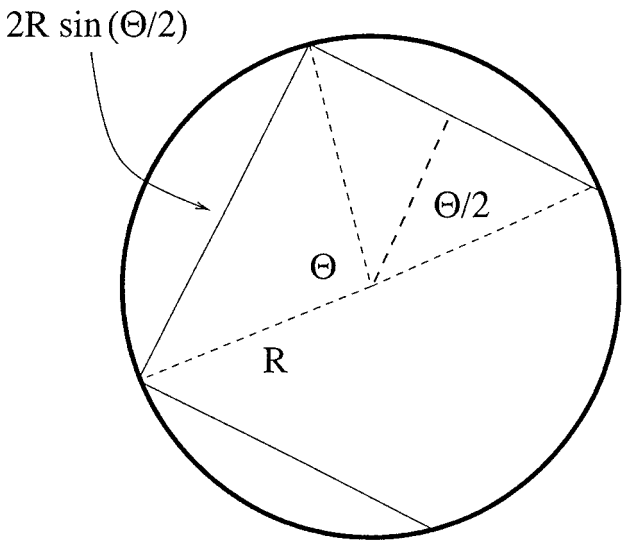

FIG. 3. Geometry leading to closed paths in a circular well. The length of each ''bounce'' is given by $2R\ \sin(\Theta/2)$ and the condition for closed orbits is $\Theta = 2\pi q/p$.

In Fig. 2 we plot the ''trajectories'' of these periodic orbit lengths (as dashed lines) as one varies f in the range (0,1) as the square is enlarged into the 1×2 rectangle. For many values of f, we then perform the numerical Fourier transform in Eq. (11) (using the lowest $N=1000$ eigenvalues) and plot the resulting $|\rho(L)|^2$ versus L/a. The locations of the peaks in $\rho(L)$ are indeed very well described globally by the intersections of the predictions of Eq. (13) with the horizontal lines of constant f. It is just this approach which we will follow when we perform a periodic orbit analysis of the general ''circular slice'' case in Sec. IV, but we first examine several special cases.

III. SPECIAL CASES

The geometry determining the structure of trajectories (closed or not) in the ordinary 2D circular well is well known[3,5,23] and is shown in Fig. 3. The distance traveled between each ''bounce'' with the circular wall is given by $L(\Theta)=2R\ \sin(\Theta/2)$. In order for the trajectory to close on itself giving rise to periodic orbits, one requires the condition that

$$p\Theta = q2\pi, \tag{14}$$

where the integer p is associated with the total number of ''bounces'' or ''hits'' on the wall and q gives the total number of ''revolutions'' made by the particle before coming back to the same point in phase space. Clearly these integers must satisfy $p \geqslant 2q$ and the lengths of periodic orbits is then given by

$$\frac{L(p,q)}{R} = 2p\ \sin\left(\frac{q\pi}{p}\right). \tag{15}$$

Repetitions of these periodic orbits are possible so that integer multiples of these values are also allowed. In fact, if p and q are relatively prime, then this labeling is unambiguous, while if $p = m\tilde{p}$ and $q = m\tilde{q}$ share a common factor, orbits labeled by $(p,q)=m(\tilde{p},\tilde{q})$ can be associated

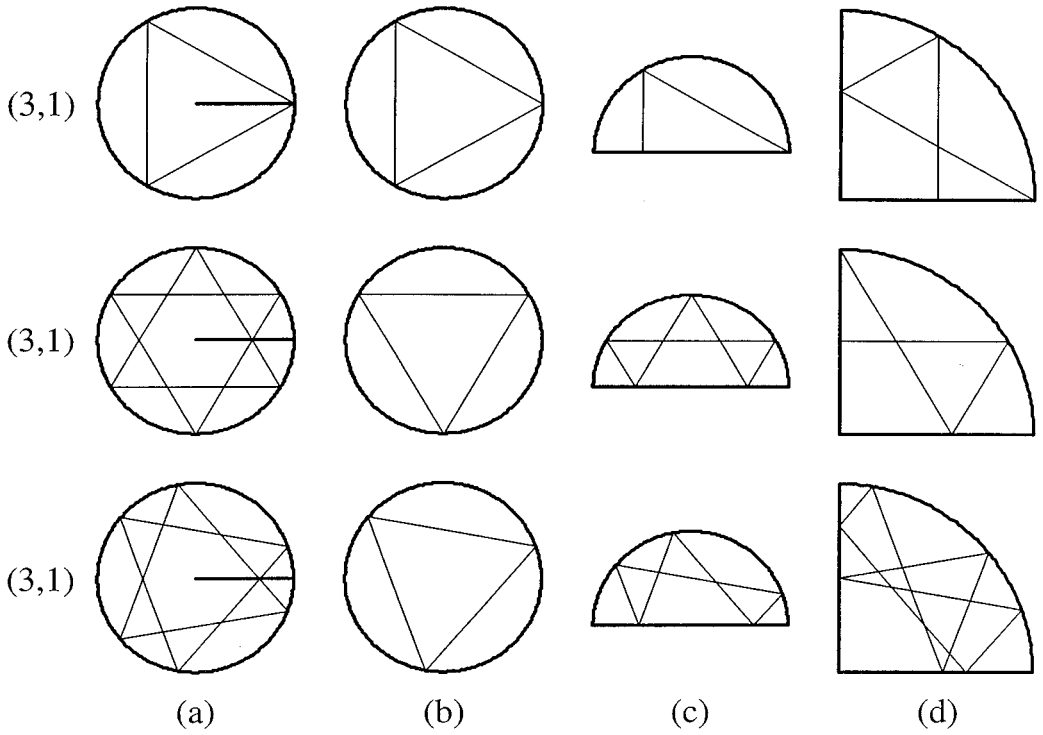

FIG. 4. Examples of possible closed orbits in each of the different wells in Fig. 1 for various values of the initial orientation of the orbit. Typical trajectories for the (a) circular well plus baffle, (b) ordinary circular well, (c) half-circular well, and (d) quarter-circular well are shown. Cases illustrating $(p,q)=(3,1)$ are shown.

with an m-fold repetition of the basic $(\tilde{p},\tilde{q})$ path; we might thus write $(6,2)=2(3,1)$ as the first repetition of the (3,1) trajectory to describe that path ''the second time around.'' For a fixed value of q, as $p\rightarrow\infty$, the path lengths approach the values

$$L(p\rightarrow\infty,q)=2\pi q, \tag{16}$$

and such paths are approaching the limit of uniform circular motion around the inner edge of the circular well. Because of this, there is a ''pile-up'' of periodic orbits as one approaches values of $L=2n\pi$ (from below) which are increasingly closely spaced as p increases. This implies that one can often not totally resolve individual (p,q) orbits near these limiting values in the $\rho(L)$ spectra.

Typical periodic orbits for the circular well for the cases $(p,q)=(3,1)$, (4,1), (5,2), (7,2), (8,3), and (2,1) are shown in Figs. 4–9, respectively, in the second column labeled (b). In each case, we show three different periodic orbits, each labeled by the same (p,q) pair, but differing only in the initial orientation of the trajectory. In each case there is obviously a continuously infinite set of trajectories for each (p,q) label.

The periodic orbits or closed trajectories in the half-circular well in column (c) of Fig. 1 are easily obtained from those of the ordinary circular well. The wall along the x-axis will give rise to specular reflections of any particle incident on it so that an allowed path (again, closed or not) in the ordinary circular well can be simply reflected in the horizontal axis and back into the upper half-plane and will result in an allowed trajectory in the half-well. Any closed path in the full well, therefore, will also be a periodic orbit in the half-well and the spectrum in $\rho(L)$ should exhibit features at the same values of L. In Figs. 4–9, we show the orbits obtained from the full well [i.e., those in column (b)] by reflection as allowed orbits in the half-well, now shown in column (c).

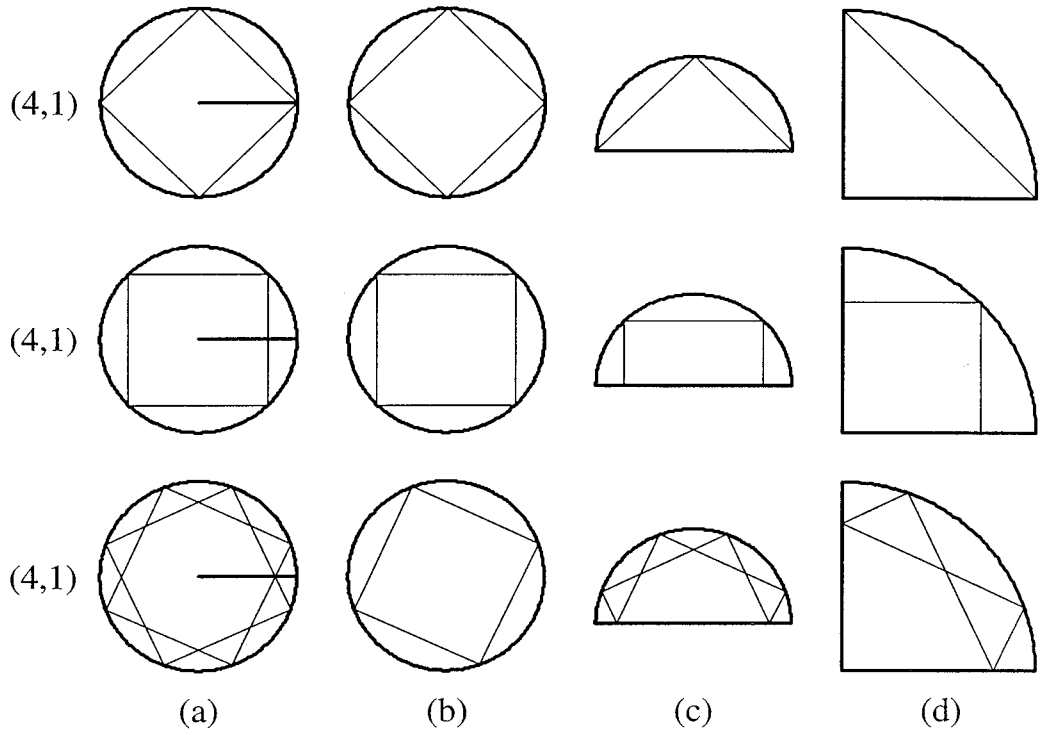

FIG. 5. The same as Fig. 4, but for $(p,q)=(4,1)$.

The one exception is the special (isolated) orbit corresponding to the $(p,q)=(2,1)$ case in Fig. 9 where the orbit through the origin closes on itself with a total path length of $L=2R$ which will lead to features in the $\rho(L)$ spectrum at integral multiples of $2R$ in addition to those in the full well at multiples of $4R$.

The periodic orbits in the quarter-well are obtained similarly by reflecting allowed orbits in the full circular well in *both* the horizontal *and* the vertical axes. Examples of such orbits are shown in Figs. 4–9 in column (d) for the same (p,q) values as used before. In this case we see that the path lengths for *even* values of p are halved as the particles return to their starting point (in phase space) in half the distance, while for *odd* values of p, the same path lengths as for the circular and half-circular wells are found.

Finally, as one can see from the examples in column (a) of Figs. 4–9, the effect of adding the baffle (the thin infinite wall along the positive x-axis) to the ordinary circular well is to double the length of all standard closed orbits as the mirroring effect of the extra wall causes all trajectories to fully retrace their paths in order to return to the same starting point in phase space. Two exceptions are illustrated in Figs. 7 and 8 for the case of *even* q values for the special orientations shown in the top row of column (a) in those two cases (compare the solid and dashed lines.) For these cases, the path lengths are unchanged from their full-well values. [We also note that the specialized geometries corresponding to the $(p,q)=(2,1)$ cases are not entirely well-defined, so that the circular well plus baffle trajectories shown in column (a) in Fig. 9 are certainly somewhat ambiguous.]

In order to compare these predictions for the path lengths of closed trajectories using the Fourier transform in Eq. (11), we require the allowed wavenumber values in these four infinite

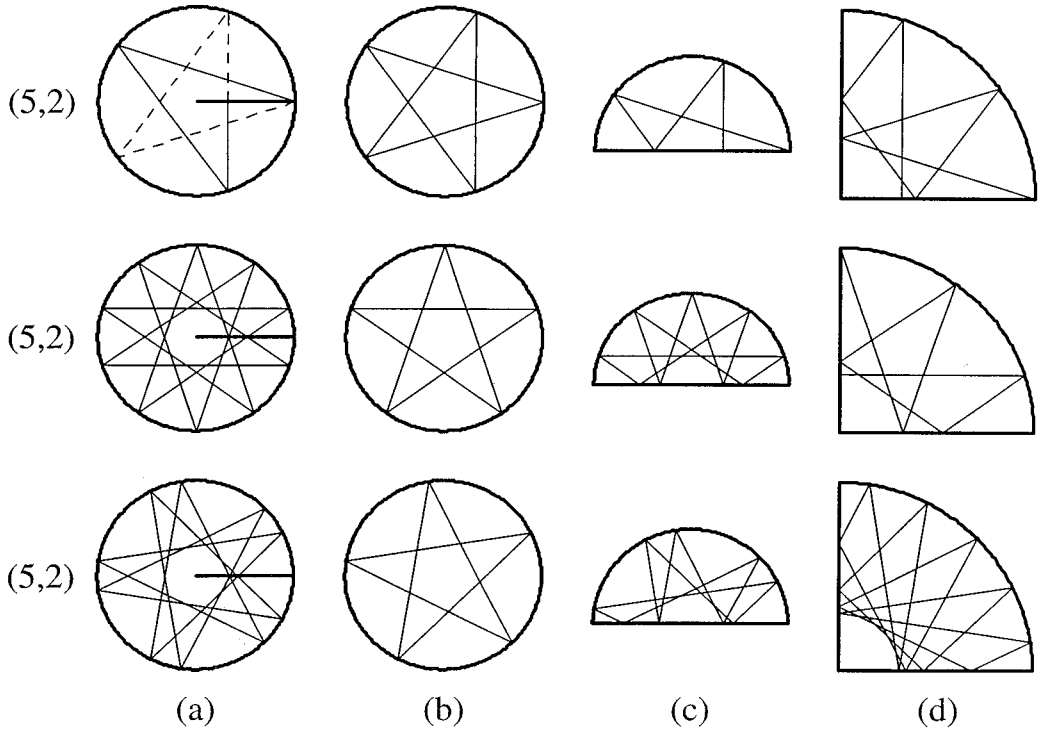

FIG. 6. The same as Fig. 4, but for $(p,q)=(5,2)$.

wells. In each case the effective (infinite well) potential function can be separated into radial and angular parts and the solutions of the Schrödinger equation (or wave equation) can be easily obtained in a product form, namely $\psi(r,\theta)=R(r)\Phi(\theta)$. For example, in the ordinary circular well case, the angular solutions are given by

$$\Phi(\theta)=\begin{cases}\sin(m\theta), & m=1,2,3,\ldots,\\ \cos(m\theta), & m=0,1,2,3,\ldots,\end{cases} \tag{17}$$

where m is an integer while the radial wavefunctions are given by (the well-behaved at the origin or regular) cylindrical Bessel functions, $R(r)=J_m(kr)$. The wavenumber spectrum is then given by $k_{n,m}=a_{n,m}/R$ where the $a_{n,m}$ denote the n-th zero of the m-th Bessel function. The spectrum is thus doubly degenerate for $m\geqslant 1$ [one zero each for the $\sin(m\theta)$ and $\cos(m\theta)$ solutions] and single-valued for the $m=0$ values.

For the half-circle case, the boundary conditions along the x-axis imply that only the $\sin(m\theta)$ solutions in Eq. (17) will be allowed, so that the spectrum in this case corresponds to one ''copy'' of the $m\geqslant 1$ spectrum of the corresponding full-well case. Similarly, the quarter-well problem, with walls along both the x- and y-axes at which the solutions must vanish, can be satisfied by those solutions in Eq. (17) of the form $\sin(m\theta)$ (which vanish along the x-axis) provided that m is an *even* integer and this limited subset of solutions then also satisfies the boundary conditions along the y-axis.

For the full-circle plus baffle case, we can still use solutions of the form $\Phi(\theta)=\sin(m\theta)$ which automatically satisfy the boundary condition along the x-axis (as $\theta\rightarrow 0$ from above) regardless of

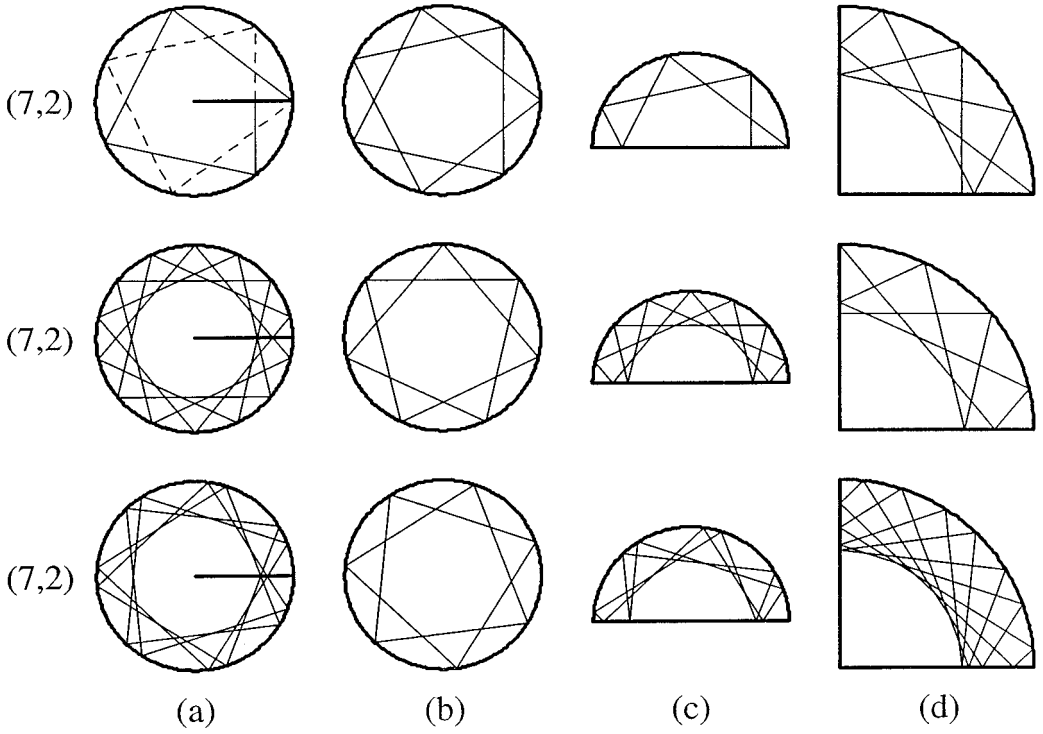

FIG. 7. The same as Fig. 4, but for $(p,q)=(7,2)$.

the value of m. In order that the solutions also vanish on the ''baffle'' as we approach the thin, infinite wall via $\theta \to 2\pi$, we also require that $\sin(m2\pi)=0$ which implies that $m=n/2$ where $n \geqslant 1$ is an integer. Thus, not only are the regular cylindrical Bessel functions still solutions (corresponding to the case where n is *even*), but now Bessel functions of half-integral order (when n is *odd*) are also required for the full solution space. These solutions (up to powers of $\sqrt{r}$) are simply the regular spherical Bessel functions which appear in the solutions of the 3D or spherical infinite well problem. The resulting wavenumber spectrum will then be determined by zeroes of both the regular *cylindrical* and *spherical* Bessel functions.

Armed with these solutions, we can numerically find a large number of the required Bessel function zeroes (we use *Mathematica*® and the FindRoot[ ] command) which we then use in the truncated Fourier transform,

$$\rho(L)=\sum_{n=1}^{N} e^{ik_nL}. \tag{18}$$

For the ordinary circular well case we use the lowest $N=3000$ wavenumbers (including each $m \geqslant 1$ value twice), while for the other three cases we use the lowest $N=2000$ values. The resulting ''spectra'' of $|\rho(L)|^2$ versus L/R are plotted in Fig. 10 for the four special cases we have considered so far, namely (a) full circle plus baffle, (b) ordinary circular well, (c) half-circle, and (d) quarter-circle cases and we label some (but not all) of the observable peaks with the appropriate (p,q) values. The overall pattern of peaks is completely consistent with the observations above and the broad shoulder seen in cases (b)–(d) near $L/R=2\pi$ is indicative of the lack of resolution mentioned earlier. The small peak at $L/R=2$ labeled $(2,1)'$ seen in the half-well case is also evident. We now turn to the more general case corresponding to arbitrary Θ to see how the

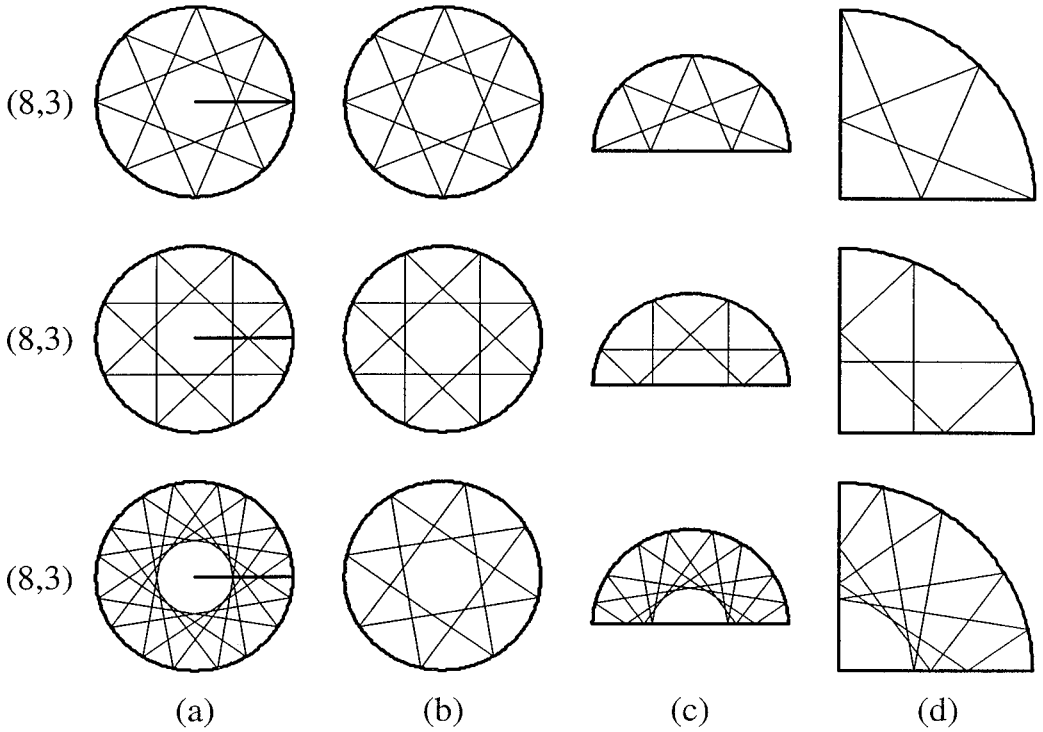

FIG. 8. The same as Fig. 4, but for $(p,q)=(8,3)$.

spectrum of the half-circle [case (c)] is changed to that of the quarter-circle of case (d) [or the circle plus baffle in case (a)] as the angle Θ is decreased (or increased) from π to $\pi/2$ (from π to 2π).

IV. MORE GENERAL CASES

The half-well, quarter-well, and circle plus baffle cases considered so far can all be considered as special cases of a general infinite well with opening angle Θ (i.e., infinite walls at $\theta=0$ and $\theta=\Theta$) as defined by the potential in Eq. (2). A convenient parameterization of the continuous change from the half-well to the circle well plus baffle case is shown in Fig. 11 (top row) where we let $\Theta=\pi(1+f)$ where $0\leqslant f\leqslant 1$. A similar description relevant for the quarter-well limit is to let $\Theta=\pi/(1+f)$ (also with $0\leqslant f\leqslant 1$) as shown in the bottom row of Fig. 11. In that case, one can continue ''shrinking'' the region beyond the quarter-well case by letting $f>1$, and we also discuss the general case below.

We begin the comparison between the classical closed orbits and the wavenumber spectrum for these more general cases by focusing first on the case where Θ increases from π. Two exemplary trajectories corresponding to $(p,q)=(4,1)$ and $(3,1)$ are shown in Fig. 12 to indicate the pattern of allowed closed trajectories. As the angle Θ is increased by a factor of $(1+f)$, the pattern of ''bounces'' discussed above remains basically unchanged except that the opening angles subtended by an individual path length are all similarly increased by the same factor. We thus expect closed orbits with lengths now given by

$$\frac{L(p,q;f)}{R}=2p\,\sin\left(\frac{q\pi(1+f)}{p}\right), \tag{19}$$

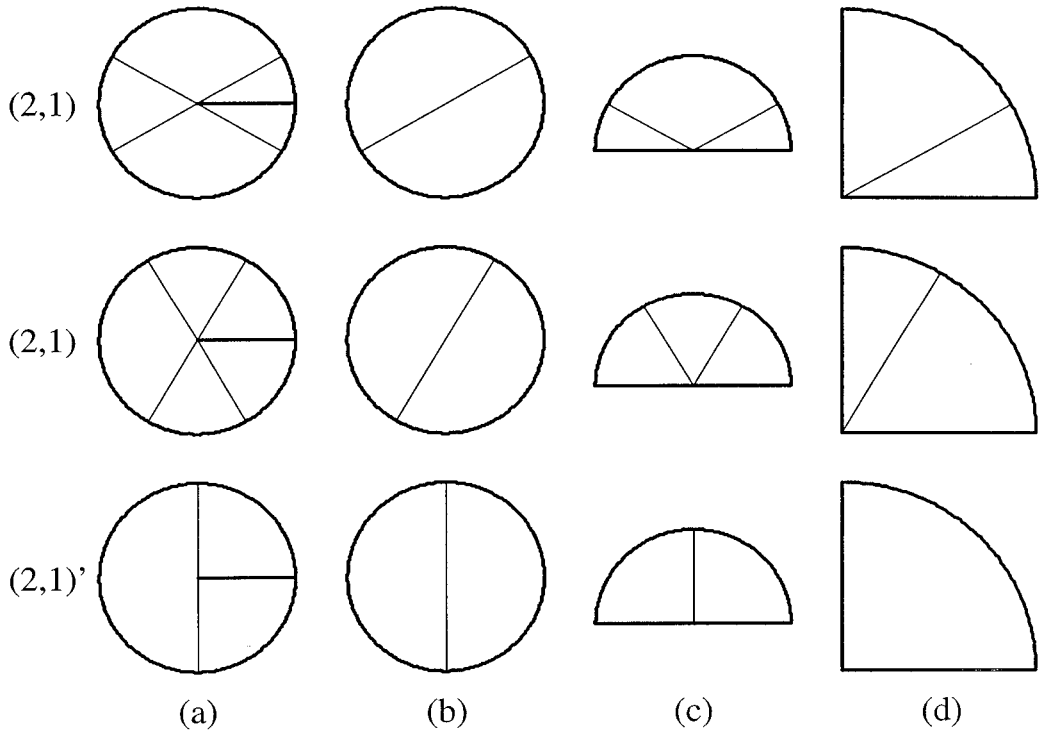

FIG. 9. The same as Fig. 4, but for $(p,q)=(2,1)$. The special case corresponding to $(2,1)'$ where the orbit length is $L=2R$ is shown in the bottom row.

as well as repetitions. This formula clearly describes the ''fate'' of the (4,1) trajectory in Fig. 12(a) as f is increased from 0 to 1. Figure 12(b), however, exhibits what can happen to other trajectories as the opening angle is increased. Beyond a certain critical value of Θ (or equivalently f), the pattern of bounces can no longer be sustained and that path should ''disappear'' from the spectrum. Operationally this occurs when $\sin(q\pi(1+f)/p)=1$ which gives a maximum value of f for certain orbits of

$$f_{max}=\frac{p}{2q}-1, \tag{20}$$

which is only relevant when $f_{max}<1$ or when $4q>p$. At that critical value, the particle makes a total of $2p$ ''bounces'' between the walls and the inner ''cusp'' and at larger angles that trajectory is no longer allowed. This phenomenon will be clearly apparent in the $\rho(L)$ spectra we discuss next.

For the quantum mechanical or wave case, we require the solutions of the Schrödinger equation in this restricted geometry. Just as discussed above, angular solutions of the form $\Phi(\theta)=\sin(m\theta)$ will satisfy the boundary conditions along the positive x-axis, while the vanishing of the wavefunction at the $\theta=\Theta$ boundary requires that $m\Theta=n\pi$ or $m=n/(1+f)$ where $n=1,2,3,\ldots$, for a given value of f. We can still evaluate the zeroes of the corresponding Bessel functions of non-integral order to obtain a long list of allowed wavenumbers. For this case, we have found the lowest-lying $N=3000$ such wavenumbers and have used them in Eq. (18) to calculate $\rho(L)$ for values of f in the range $0\leqslant f\leqslant 1$. Various such spectra are plotted in Fig. 13 (horizontal plots of $|\rho(L)|^2$ versus L/R) for different values of f. The dashed lines correspond to the predictions for classical path lengths given by Eq. (19) and the agreement between the observed peaks in the

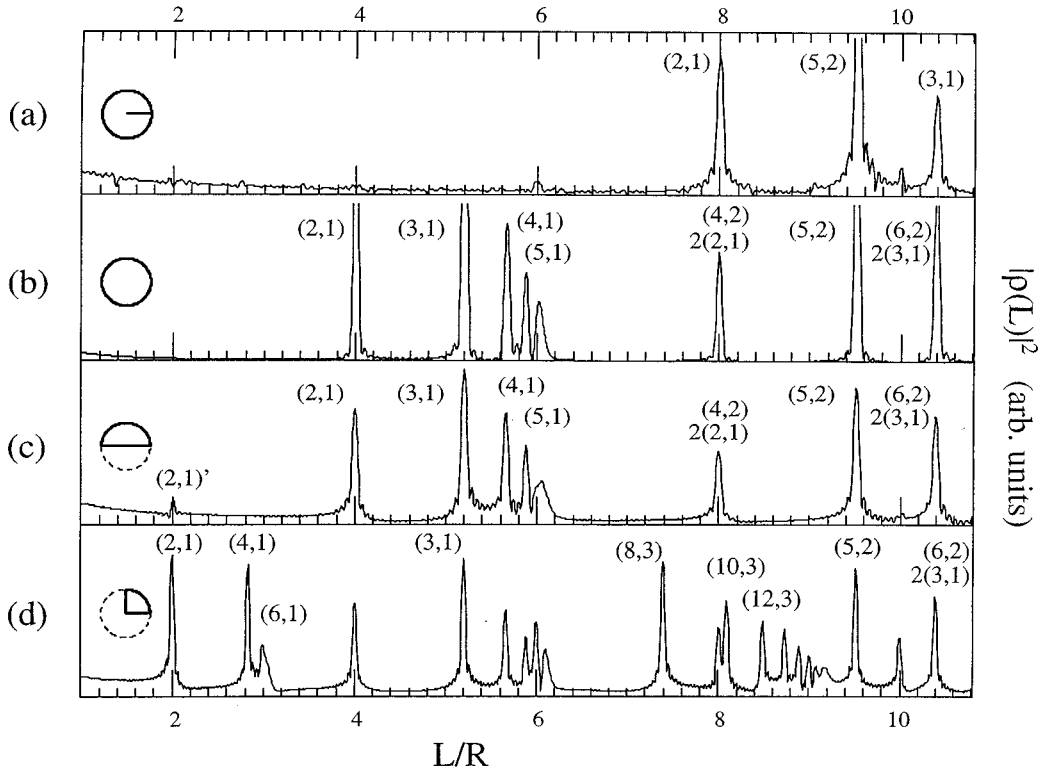

FIG. 10. Fourier transform (squared) of the wavenumber distribution, i.e., $|\rho(L)|^2$ versus L/R, for the cases of the (a) circular well plus baffle, (b) ordinary circular well, (c) half-circular well, and (d) quarter-circular well. Labels indicating the appropriate (p,q) values are shown for some (but not all) of the peaks.

spectra and the predicted locations (intersections of the dashed lines with horizontal lines of constant f) is clearly very good. As often happens, the infinite number of peaks near the limiting cases of circular motion (here just below $L/R=2\pi\rightarrow 4\pi$ as f increases) are not well resolved, especially for larger values of f. The f values at which various of the spectral features ''disappear'' from the spectrum because of the geometrical effects discussed above also is well-reproduced by the predictions of Eq. (20). Specifically, we find the values

(p,q)	f_{max}
(3,1)	1/2=0.500
(5,3)	1/4=0.250
(6,2)	1/2=0.500
(7,2)	3/4=0.750
(7,3)	1/6=0.167

which agree nicely with the spectral pattern in Fig. 13. We note that such ''bifurcations'' of orbits (the appearance or disappearance of classically allowed paths at some critical value of a parameter) also occur in the annular billiard system considered in Refs. 19 and 20: specifically, as the radius of the inner annulus is increased, certain allowed classical orbits are no longer consistent with the restricted geometry and no longer appear in the $\rho(L)$ spectrum.

We now turn our attention to the case where $\Theta=\pi/(1+f)$ is decreased which describes the transition between the half-well to quarter-well cases as f increases from 0 to 1. Just as above we expect closed path lengths (due to the decrease in Θ) now of the form

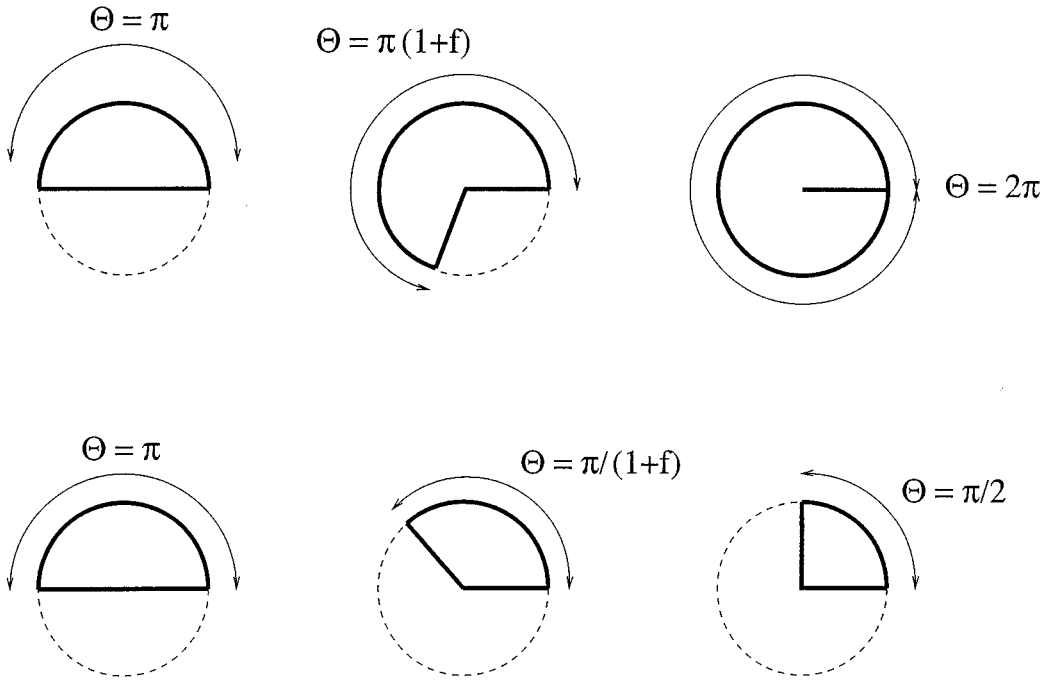

FIG. 11. Definitions of angle Θ for continuous deformation of the half-circle to the circle plus baffle [the top row where $\Theta=\pi(1+f)$] and to the quarter-circle [the bottom row where $\Theta=\pi/(1+f)$]. In each case, one has $0\leqslant f\leqslant 1$.

$$\frac{L(p,q;f)}{R}=2p\ \sin\left(\frac{q\pi}{p(1+f)}\right). \tag{21}$$

Also just as above, for a given value of f, we can now find a large number of low-lying zeroes of Bessel functions of non-integral order, $J_m(kr)$, this time labeled by $m=n(1+f)$ so that when f=1 (corresponding to the quarter-well) this reduces to even integral values as mentioned above. The resulting ''spectra'' of $|\rho(L)|^2$ versus L/R are shown in Figs. 14 and 15 for several values of f over the ranges $L/R=(2.0,8.0)$ and $L/R=(8.0,14.0)$ separately to allow for easier inspection of the more complex pattern of peaks. The predictions for the path lengths in Eq. (21) are once again shown as dashed lines. Once again, there is a large degree of overlap between the locations of the peaks and the predictions of Eq. (21). For example, the dashed ''path length trajectory'' labeled (2,1) is well described by the formula and is consistent with the illustrative example of Fig. 16 where the evolution of such a path is shown.

There are, however, a number of prominent features for the quarter-well case (i.e., along the $f=1$ horizontal line) which are not the limiting cases of the simple generalization of Eq. (15) as given by the dashed lines of Eq. (21) and these cases have been labeled by capital letters from A–K. These paths, which have been traced by solid curves, are instead described by the formulae

$$\frac{L(p,q;f)}{R}=\begin{cases}\left[2p\ \sin\left(\dfrac{2\pi q}{p(1+f)}\right)\right], & \text{for } p \text{ odd},\\ \dfrac{1}{2}\left[2p\ \sin\left(\dfrac{2\pi q}{p(1+f)}\right)\right], & \text{for } p \text{ even}.\end{cases} \tag{22}$$

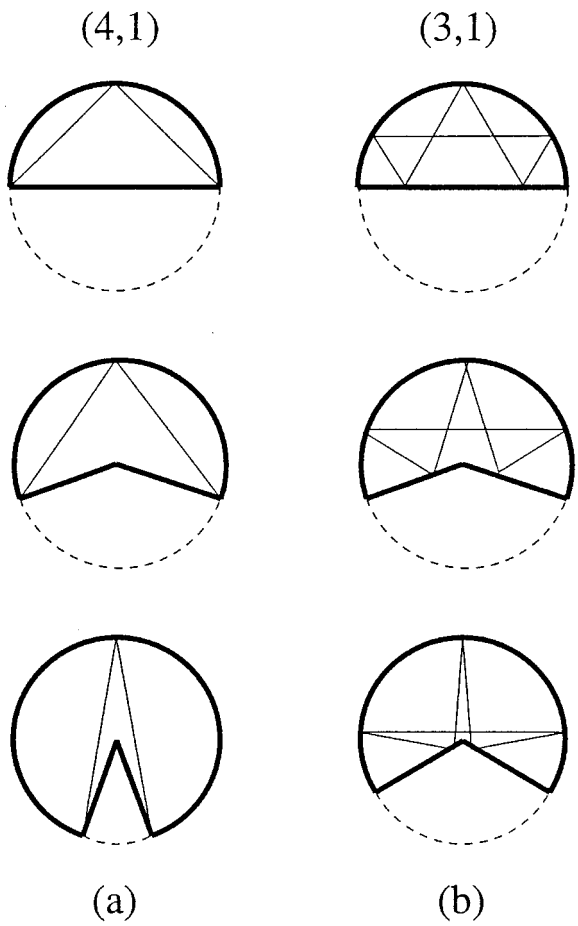

FIG. 12. Diagrams demonstrating the increasing path lengths corresponding to the (a) $(p,q)=(4,1)$ case and (b) (3,1) cases as $\Theta=\pi(1+f)$ is increased; special symmetric paths are chosen for illustrative purposes. For the (4,1) case, the formula $L(p,q;f)/R=2p\sin(\Theta(1+f)q/p)$ in Eq. (19) works for all f values in the range $0\leqslant f\leqslant 1$, while for the (3,1) case, the diagram shows the origin of the maximum value of f given by $f_{max}=p/2q-1=0.5$ corresponding to $\Theta=3\pi/2$.

The origin of these path length formulas is exemplified for a special case [namely $(p,q)=(3,1)$ and hence p odd] in Fig. 17. The evolution of a particularly symmetric (3,1) trajectory is shown in the top row, and in this case it is best to consider the angular size of the well as it is increased from $\pi/2$ to π (i.e., going from right to left). Because of the mirroring effect of the infinite wall along the x-axis, the path lengths for the upper row [where $\Theta=\pi/(1+f)$] are identical to those indicated in the bottom row as exist in the larger ''slice'' of *twice* the opening angle, namely $\Theta=2\pi/(1+f)$. Thus, familiar paths which are allowed for the quarter-well case can be transformed in direct analogy to the half-well to full circle plus baffle case; this identification explains the appearance of the additional factor of 2 in the $2\pi q/p(1+f)$ term in Eq. (22), as opposed to Eq. (19), due to mirroring.

In these cases, there is now a *minimum* value of f which is indicated before the closed orbit can be maintained in the increasingly narrow region. The effect is also illustrated in Fig. 17 for the (3,1) path and is very similar to spirit to the arguments shown in Fig. 11(b). The minimum value of f implied by requiring that $\sin(2\pi q/p(1+f))=1$ (just as above) is

$$f_{min}=\frac{4q}{p}-1, \tag{23}$$

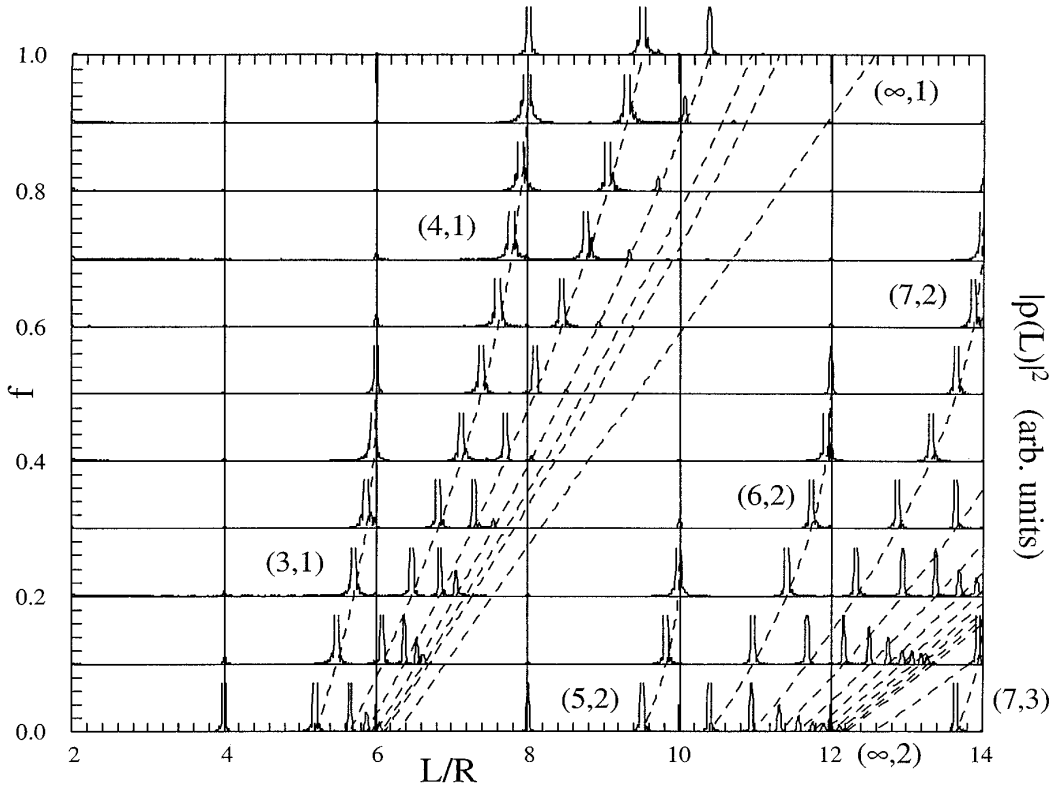

FIG. 13. Fourier transform (squared) of the wavenumber distribution, $|\rho(L)|^2$ versus L/R, for increasing values of $\Theta = \pi(1+f)$. For each value of f shown on the left-hand axis, a ''spectrum'' corresponding to $|\rho(L)|^2$ versus L/R is shown on the horizontal axis. The value $f=0\,(1)$ corresponds to the case of the half-circular (full circle plus baffle) well. The positions of the observable peaks in $\rho(L)$ are well described by the formula in Eq. (19) which are given by the dashed lines, while the values of f at which various of the dashed ''path length trajectories'' disappear are consistent with Eq. (20).

and the locations of these f values for the special ''path length trajectories'' labeled on Figs. 14 and 15 are given by

label	(p,q)	f_{min}
A	(3,1)	1/3=0.333
B	$\frac{1}{2}$(8,3)	1/2=0.500
C	$\frac{1}{2}$(10,3)	1/5=0.200
D	(5,2)	3.5=0.600
E	(6,2)=2(3,1)	1/3=0.333
F	(7,2)	1/7=0.143
G	$\frac{1}{2}$(12,5)	2/3=0.667
H	$\frac{1}{2}$(14,5)	3/7=0.429
I	$\frac{1}{2}$(16,5)	1/4=0.250
J	(7,3)	5/7=0.714
K	$\frac{1}{2}$(18,5)	1/9=0.111

and once again the f values at which the solid ''path length trajectories'' appear from the vertical lines of constant L/R given by $L/R=p$ (for even values of p) and $L/R=2p$ (for odd values of p) are very well reproduced.

With the cases considered so far, it is straightforward to generalize this analysis to the $\Theta \leqslant \pi/2$

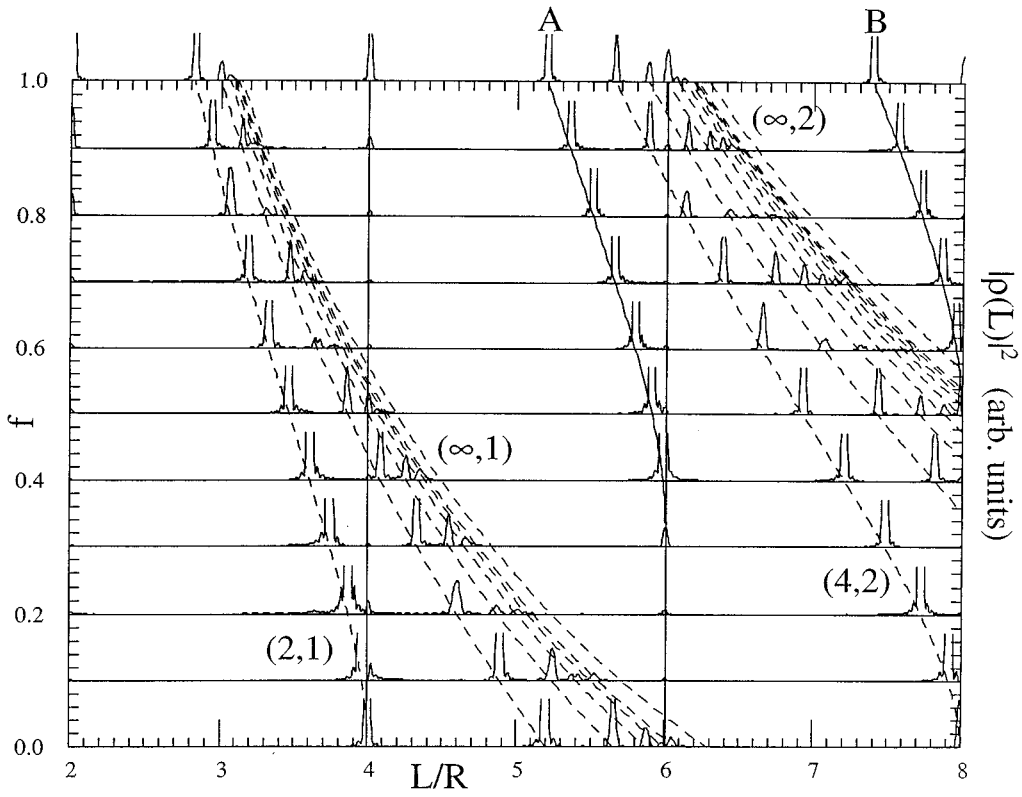

FIG. 14. The Fourier transform (squared) of the wavenumber distribution, $|\rho(L)|^2$ versus L/R, for decreasing values of $\Theta=\pi/(1+f)$. The value $f=0$ (1) corresponds to the case of the half-circular (quarter-circular) well. The range $L/R=(2.0,8.0)$ is shown. The positions of many of the observable peaks in $\rho(L)$ are well described by the formula in Eq. (21) which are given here by the dashed lines. The solid lines are described by the formulae in Eqs. (22) and the values of f at which they ''break out'' from the vertical lines corresponding to even values of L/R are all consistent with the formula $f_{min}=4q/p-1$ in Eq. (23).

region, where $f>1$. For example, in the region $\pi/4\leqslant\Theta\leqslant\pi/2$, where $3\geqslant f\geqslant 1$, the ''new'' closed orbits, i.e., those which appear to ''sprout'' from the vertical lines at even values of L/R, are given by

$$\frac{L(p,q;f)}{R}=\begin{cases}\left[4p\,\sin\left(\dfrac{4\pi q}{p(1+f)}\right)\right], & \text{for } p \text{ odd},\\ \dfrac{1}{2}\left[4p\,\sin\left(\dfrac{4\pi q}{p(1+f)}\right)\right], & \text{for } p=2m,\ m=1,2,3\ldots,\\ \dfrac{1}{4}\left[4p\,\sin\left(\dfrac{4\pi q}{p(1+f)}\right)\right], & \text{for } p=4m,\ m=1,2,3,\ldots,\end{cases} \tag{24}$$

since the values with $p=4m$ can be ''mirrored twice'' and have their path lengths cut in half a second time. In this region, the closed paths will be described by these expressions, as well as those already discussed in Eqs. (21) and (22). As an example of the increasing complexity of both the ''experimental spectra'' given by the peaks in $\rho(L)$ and the theoretical predictions of these three classes of path length formulae, in Fig. 18 we show a portion ($8\leqslant L/R\leqslant 10$) of the spectra covering the range $1\leqslant f\leqslant 3$ (corresponding to $\pi/2\geqslant\Theta\geqslant\pi/4$.) The dashed and solid curves are the

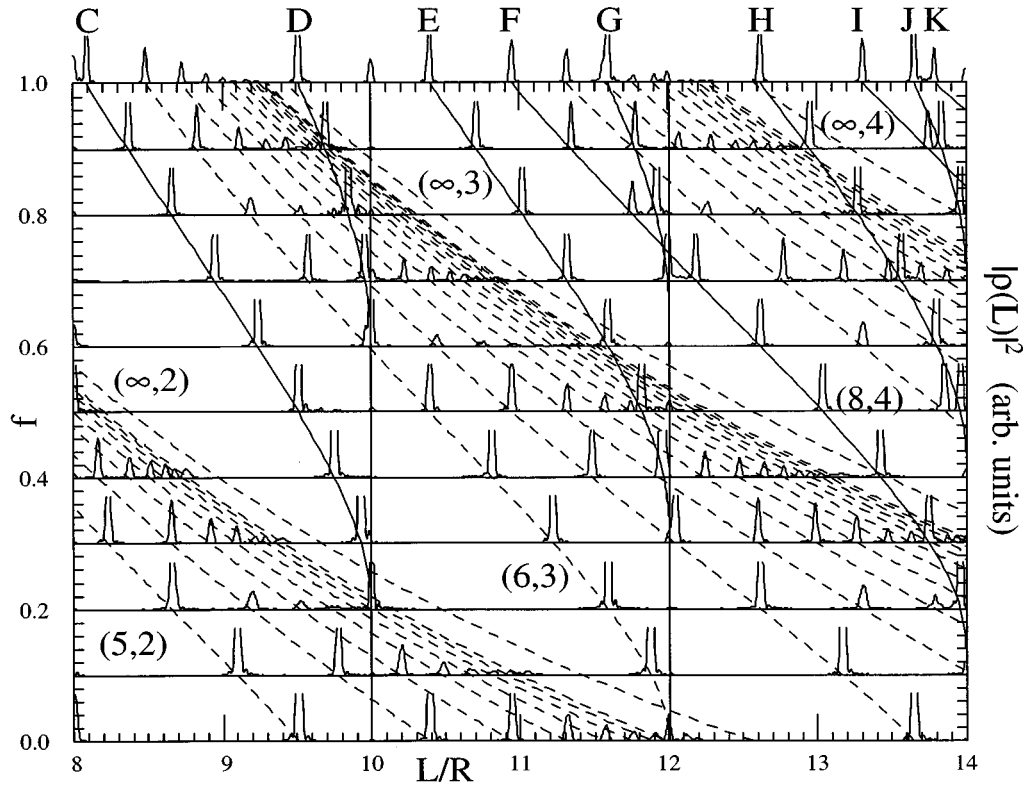

FIG. 15. The same as Fig. 14, except for the range $L/R=(8.0,14.0)$.

expressions in Eqs. (21) and (22) which were already apparent in the $\pi/2 \leqslant \Theta \leqslant \pi$ region, while the bold dashed curves correspond to the new cases described in Eq. (24) for those trajectories which first ''appear'' in the $\Theta \in (\pi/4, \pi/2)$ region. Once again, the agreement between the theoretical predictions (locations of crossings with the horizontal lines of constant f) and the locations of features in the Fourier ''spectra'' is quite good.

This pattern can then be generalized straightforwardly to the general case, namely we find that in the Θ range given by $(\pi/2^{n+1}, \pi/2^n)$ there are $n+2$ new formulae of the form in Eq. (24), namely

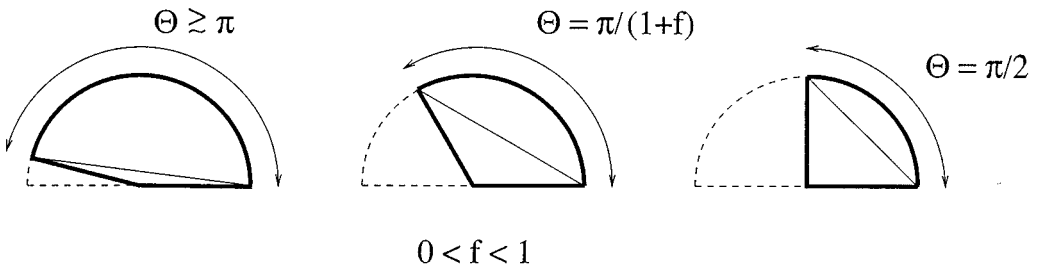

FIG. 16. The diagram illustrating the decreasing path length corresponding to the $(p,q)=(4,1)$ case as $\Theta=\pi/(1+f)$ is decreased.

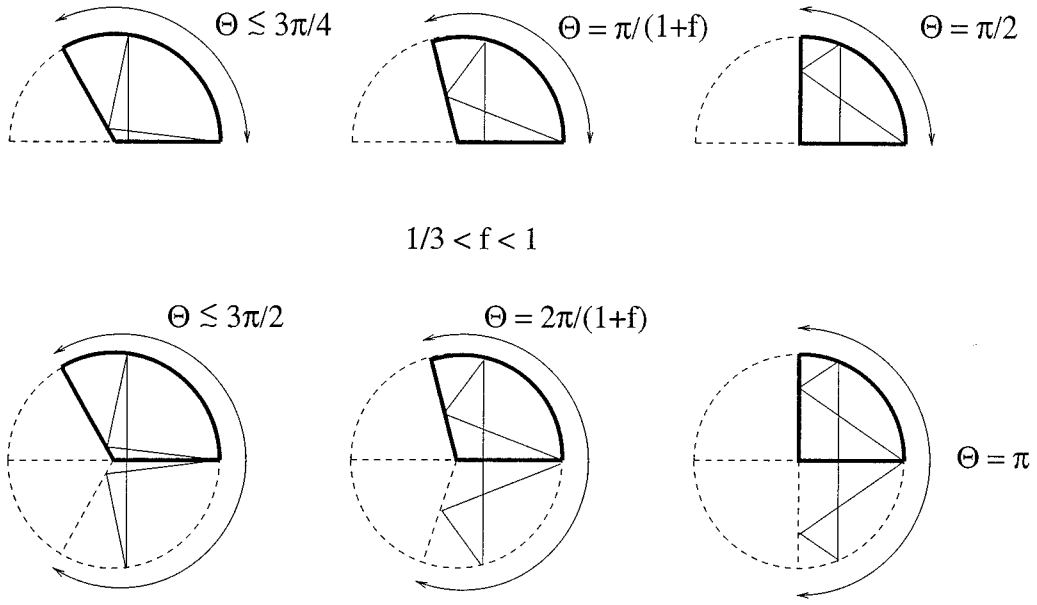

FIG. 17. The same as Fig. 16, but for the case $(p,q)=(3,1)$ illustrating the origin of the path-length formulae in Eq. (22). This figure also illustrates the origin of the minimum f value given by $f_{min}=4q/p-1=1/3$ corresponding to $\Theta=\pi/(1+f_{min})=3\pi/4$.

$$\frac{L(p,q;f)}{R}=\begin{cases}\left[2p\ \sin\left(\dfrac{2^{n+1}\pi q}{p(1+f)}\right)\right], & \text{for } p \text{ odd},\\ \dfrac{1}{2}\left[2p\ \sin\left(\dfrac{2^{n+1}\pi q}{p(1+f)}\right)\right], & \text{for } p=2m,\ m=1,2,3\ldots,\\ \dfrac{1}{4}\left[2p\ \sin\left(\dfrac{2^{n+1}\pi q}{p(1+f)}\right)\right], & \text{for } p=4m,\ m=1,2,3,\ldots,\\ \quad\vdots & \quad\vdots \qquad\qquad \vdots\\ \dfrac{1}{2^n}\left[2p\ \sin\left(\dfrac{2^{n+1}\pi q}{p(1+f)}\right)\right], & \text{for } p=2^n m,\ m=1,2,3,\ldots,\end{cases} \tag{25}$$

in addition to the path length formulae which would have already been considered for larger values of Θ. The values $\Theta=\pi/2^n$ form boundaries at which the new terms in Eq. (25) are added, with new spectral features ''appearing'' from even values of L/R as Θ is decreased from $\pi/2^n$ to $\pi/2^{n+1}$. At each value of $\Theta=\pi/2^n$, there are additional ''mirrorings'' possible for appropriate values of p.

An extension of this geometry which is also more calculationally extensive is to consider a periodic orbit theory analysis of a 3D spherical infinite well and extensions thereof. One of the original references on this subject[3] used the spherical billiard as a numerical example and it is easy to imagine extending this analysis to half-, quarter-, and eighth-spheres, as well as to 3D ''slices'' with arbitrary opening solid angle given by Ω. Many of the same symmetry arguments can be used to extract the special cases, while the general case requires zeroes of spherical Bessel functions of non-integral orders.

In conclusion, we have performed a periodic orbit theory analysis of the problem of a free particle in a two-dimensional infinite well with circular ''slice'' geometry. We have found the general pattern of periodic orbits for an arbitrary opening angle Θ and investigated

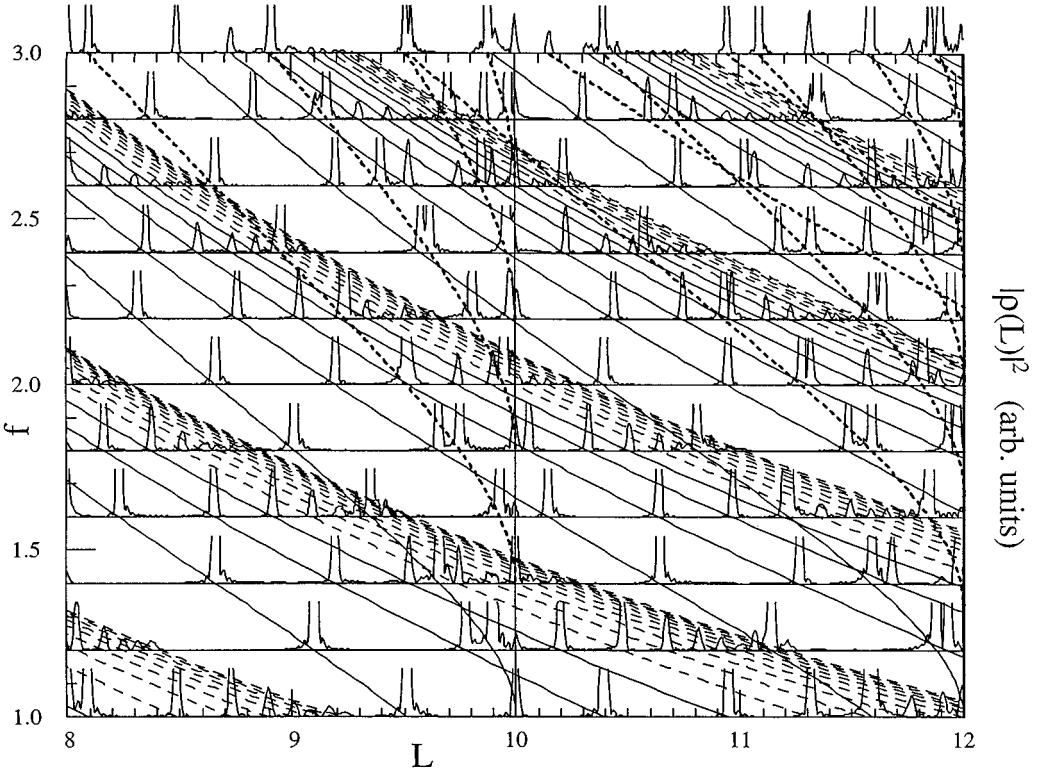

FIG. 18. The Fourier transform (squared) of the wavenumber distribution, $|\rho(L)|^2$ versus L/R, for decreasing values of $\Theta=\pi/(1+f)$ over the region $1\leqslant f\leqslant 3$ over the range $L/R=(8.0,10.0)$. The dashed (solid, bold-dashed) curves correspond to the cases described by [Eq. (21) (Eq. (22), Eq. (24)], respectively.

the appearance and disappearance of allowed orbits at the special cases $\Theta=\pi/2^n$. A simplified PO analysis, focusing on the locations of the peaks in $\rho(L)$, has found excellent global agreement with the increasingly complex pattern of allowed closed paths as Θ is decreased. No obvious signs of diffraction effects (such as seen in the annular ring billiard[19,20]) are apparent. However, a more careful analysis [one which would necessarily include the prefactor terms in Eq. (4)] of this system and the annular billiard, both of which exhibit bifurcations of orbits, should be performed to probe more carefully the structure of the features in the $\rho(L)$ spectrum near the bifurcation points to decide if a complete description requires the inclusion of diffraction effects as discussed in Refs. 5 and 16–18. For example, the diffraction effects seen most clearly in the annular billiard case (corresponding to diffraction around the inner annulus) decrease in magnitude exponentially with increasing inner radius $R_{min}=fR$. One could easily imagine small diffraction features which persist in the $\rho(L)$ spectrum beyond the values of f at which the purely classical trajectories would no longer be allowed due to such diffraction effects. Such extensions of this work are currently under study.

[1] M. C. Gutzwiller, J. Math. Phys. **11**, 1791 (1970).
[2] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, Berlin, 1990).
[3] R. Balian and C. Bloch, Ann. Phys. **69**, 76 (1972).
[4] M. V. Berry, in *Chaotic Behavior of Deterministic Systems,* Proceedings of the Les Houches Summer School, Session XXXVI, edited by G. Iooss. R. H. G. Helleman, and R. Stora (North-Holland, Amsterdam, 1983), pp. 171–272.
[5] For a recent, thorough review of many aspects of periodic orbit theory, see M. Brack and R. K. Bhaduri, *Semiclassical Physics* (Addison-Wesley, Reading, MA, 1997).

[6] M. L. Du and J. B. Delos, Phys. Rev. A **38**, 1896 (1988).
[7] D. Wintgen, Phys. Rev. Lett. **58**, 1589 (1987); A. Holle, J. Main, G. Weibusch, H. Rottke, and K. H. Welge, *ibid.* **61**, 161 (1988); H. Friedrich and D. Wintgen, Phys. Rep. **183**, 37 (1989).
[8] J. Gao and J. B. Delos, Phys. Rev. A **49**, 869 (1994); M. Courtney, H. Jiao, N. Spellmeyer, D. Kleppner, J. Gao, and J. B. Delos, Phys. Rev. Lett. **74**, 1538 (1995).
[9] K. Nakamura, *Quantum Chaos: A New Paradigm of Nonlinear Dynamics* (Cambridge University Press, Cambridge, 1993).
[10] M. Brack and S. R. Jain, Phys. Rev. A **51**, 3462 (1995).
[11] H.-J. Stöckman and J. Stein, Phys. Rev. Lett. **64**, 2215 (1990).
[12] B. Tatievski, P. Stampfli, and K. H. Bennemann, Comput. Mater. Sci. **2**, 459 (1994).
[13] V. M. Strutinsky, A. G. Magner, S. R. Ofengenden, and T. Døssing, Z. Phys. A **283**, 269 (1977); V. M. Strutinskiĭ and A. G. Magner, Sov. J. Part. Nucl. **7**, 138 (1976).
[14] O. Bohigas, D. Boosé, R. Egydio de Carvalho, and V. Marvulle, Nucl. Phys. A **560**, 197 (1993).
[15] K. Richter, D. Ullmo, and R. A. Jalabert, Phys. Rep. **276**, 1 (1996).
[16] G. Vattay, A. Wirzba, and P. E. Rosenqvist, Phys. Rev. Lett. **73**, 2304 (1994).
[17] N. Pavloff and C. Schmit, Phys. Rev. Lett. **75**, 61 (1995).
[18] H. Primack, H. Schanz, Y. Smilansky, and I. Ussishkin, Phys. Rev. Lett. **76**, 1615 (1996).
[19] S. M. Reimann, M. Brack, A. G. Magner, and M. V. N. Murthy, Surf. Rev. Lett. **3**, 19 (1996). See also an extensive discussion with references to other unpublished work in Ref. 5.
[20] R. W. Robinett, to appear in Surf. Rev. Lett.
[21] T. Szerdi and D. A. Goodings, Phys. Rev. Lett. **69**, 1640 (1992); Phys. Rev. E **48**, 3518 (1993); *ibid.* **48**, 5329 (1993).
[22] For references to the original literature, see R. Balian and C. Bloch, Ann. Phys. **60**, 401 (1970).
[23] M. V. Berry, Eur. J. Phys. **2**, 91 (1981).

New conformal gauging and the electromagnetic theory of Weyl

James T. Wheeler[a)]
Department of Physics, Utah State University, Logan, Utah 84322

(Received 7 July 1997; accepted for publication 3 September 1997)

A new eight-dimensional conformal gauging solves the auxiliary field problem and eliminates unphysical size change from Weyl's electromagnetic theory. We derive the Maurer–Cartan structure equations and find the zero curvature solutions for the conformal connection. By showing that every one-particle Hamiltonian generates the structure equations we establish a correspondence between phase space and the eight-dimensional base space, and between the action and the integral of the Weyl vector. Applying the correspondence to generic flat solutions yields the Lorentz force law, the form and gauge dependence of the electromagnetic vector potential and minimal coupling. The dynamics found for these flat solutions applies locally in generic spaces. We then provide necessary and sufficient curvature constraints for general curved eight-dimensional geometries to be in 1–1 correspondence with four-dimensional Einstein–Maxwell space–times, based on a vector space isomorphism between the extra four dimensions and the Riemannian tangent space. Despite part of the Weyl vector serving as the electromagnetic vector potential, the entire class of geometries has vanishing dilation, thereby providing a consistent unified geometric theory of gravitation and electromagnetism. In concluding, we give a concise discussion of observability of the extra dimensions. © *1998 American Institute of Physics.* [S0022-2488(98)02601-2]

I. INTRODUCTION

In 1918, following immediately on the heels of Einstein's introduction of general relativity, Weyl proposed a generalization of Riemanian space–time structure[1] to allow for local changes in the standard of length. When the gauge field required to insure this symmetry has nonvanishing curl, the lengths of vectors change as they undergo parallel transport. This nicely completes the picture of Riemann in which vectors are rotated but not changed in length. In Weyl geometry, all properties of transported vectors—magnitude *and* direction—are subject to the transport law.

Weyl identified the gauge field of local dilations with the electromagnetic potential, thereby explaining the gauge freedom of that potential and giving a geometric interpretation of electromagnetic forces. Unfortunately, the theory is inconsistent with experience because it predicts changes in the sizes of objects depending on their paths. Applied, for example, to the case of atoms moving in a background electromagnetic field, the predicted size changes would produce substantial unobserved broadening of atomic spectral lines.

The simplicity and naturalness of Weyl geometry together with the promise of a geometric understanding of electromagnetism makes the failure of Weyl's physical theory puzzling. Indeed, Weyl[1–3] and many other physicists[4–22] included proposed alternative versions of the physical interpretation in attempts to make the theory consistent. Ultimately, the work led to the current U(1) gauge theory of electromagnetism, which maintains much of the beauty of the original proposal without inconsistent predictions such as changes in atomic spectra.

In the present work, we show that a new conformal gauging avoids the standard objection[5] to this 80-year-old theory of Weyl. Since the conformal group contains the Weyl group, the spaces

[a)]Electronic-mail: jwheeler@cc.usu.edu

resulting from its gauging share certain properties with Weyl's original gauge theory of electromagnetism. However, the eight-dimensional character of the new gauging and its interpretation as a generalization of phase space eliminate the problem of unphysical size change. Before discussing these new conformal spaces, which we call *biconformal geometries*, we review some results from previous studies of the conformal group.

The conformal group is the most general set of transformations preserving ratios of infinitesimal lengths. On four-dimensional space–time the conformal group is 15 dimensional, including Lorentz transformations, translations, inverse translations, and dilations. The Lorentz transformations and translations are well known, while the dilations are simple rescalings. The inverse translations have alternatively been called special conformal transformations, conformal boosts, accelerations, or elations. We choose the term inverse translations because these transformations are pure translations at infinity. This is easily seen by translating the inverse coordinate $y^\mu = -x^\mu/x^2 \rightarrow y^\mu + a^\mu$ and then inverting again.

We will also make use of several subgroups of the conformal group. The Lorentz transformations and translations together form the Poincaré group, while the Lorentz transformations and the dilations form the homogeneous Weyl group. Adding the translations to the Weyl group gives the inhomogeneous Weyl group. The Lorentz transformations, dilations, and inverse translations also form an inhomogeneous Weyl group with the only difference being that the translations and inverse translations have opposite behavior under dilations.

Considerable attention[23–27] has been given to four-dimensional space–times which result from gauging the conformal group. In addition to the usual metric and local Lorentz structures of space–time, these models are expected, a *a priori*, to possess five gauge vectors—one for dilations and four for the inverse translations. However, it was long thought that the inverse translations could not be gauged, because the corresponding current is position dependent. From a slightly different perspective, Mansouri *et al.*[23,24] argue that making the inverse translations local is redundant with general coordinate transformations so that the corresponding gauge fields have no new effect. Nonetheless, Crispim Romão, Ferber, and Freund,[25,26] and independently Kaku, Townsend, and Van Nieuwenhuizen[27] show how to gauge the inverse translations, but find them algebraically removable from the problem. This problem is in fact generic: in any four-dimensional scale-invariant torsion-free field theory with action built quadratically from the curvatures of the conformal group these four inverse translational gauge fields may be algebraically removed from the problem.[28] The local inhomogeneous Weyl symmetry of such theories always has an effective field theory equivalent to one based on the homogeneous Weyl group, and therefore sheds no light on Weyl's theory.

By contrast, consideration of the fixed points of the various conformal transformations suggests the use of an eight-dimensional base space instead of a four-dimensional base space, corresponding to the existence of eight distinct translational generators in the Lie algebra. Requiring an eight-dimensional base space makes the gauge fields of the inverse translations act as part of the solder form on the base space, so these fields *necessarily* remain independent. Even in the case of a fully flat biconformal space this solder form contains nontrivial physical information.

Our present investigation of this new conformal gauging focuses principally on the study of flat biconformal spaces, because this is where the connection with well-known systems should be most evident and because the dynamics of flat spaces will also apply locally in generic spaces. Unlike Riemannian geometry, there is an entire nontrivial class of flat solutions. We first derive this class of solutions. Next, we show that the flat spaces possess a symplectic structure which allows their interpretation as one-particle phase spaces and provides a concise description of the corresponding Hamiltonian dynamics. Applying this Hamiltonian correspondence to the most general flat solution gives a geometric derivation of the Lorentz force law, in strong contrast to the original Weyl theory. After indicating briefly how the biconformal electromagnetic theory connects satisfactorily to U(1) gauge theory, we show how to introduce Einstein gravitation, arbitrary additional fields, and electromagnetic sources into the biconformal structure.

The layout of the paper is as follows. In Sec. II we develop biconformal structure as a conformal fiber bundle. We give particular attention to the general procedure for building connections on manifolds based on an underlying symmetry group[29–32] because it is this procedure which motivates our use of an eight-dimensional base manifold. The construction consists of two steps. First, an elementary geometry is found by taking the quotient of a given symmetry group by a subgroup containing no normal subgroup of the full group. Then, the connection of the resulting principal fiber bundle is generalized by including curvature. As an example we review the construction of general space–times by gauging the Poincaré group.[33–36] We then implement these techniques to gauge the conformal group in a new way. First, we construct an elementary geometry as the quotient, $\mathscr{C}/\mathscr{C}_0$, of the conformal group, $\mathscr{C}$, by its isotropy subgroup, $\mathscr{C}_0$. This gives a conformal Cartan connection on an eight-dimensional manifold. Next, we generalize the connection to that of a curved eight-dimensional manifold with the seven-dimensional homogeneous Weyl group as fiber by the addition of horizontal curvature two-forms to the group structure equations. The resulting eight-dimensional base manifold is called a *biconformal space* and the full fiber bundle is called the biconformal bundle.

Our first principal result is presented in Sec. III where we derive the flat solution for a frame field satisfying the zero-curvature Maurer–Cartan structure equations of biconformal space.

In Secs. IV and V we establish our next claim: the classical Hamiltonian dynamics of a single-point particle is equivalent to the specification of a seven-dimensional surface in flat biconformal space. As a consequence of the form of the structure equations, there necessarily exists a set of preferred curves satisfying the Hamiltonian equations of motion. This claim is established in two steps. First, in Sec. IV, we show that the classical Hamiltonian description of a point particle defines a class of eight-dimensional differential geometries with structure equations of manifestly biconformal type. Then in Sec. V we show that there is an embedding, unique up to local symplectic changes of basis, of the Hamiltonian system into a *flat* biconformal geometry. We also demonstrate the necessary existence of a preferred set of curves in the biconformal space satisfying the Hamiltonian equations of motion. The unique local equivalence between hypersurfaces in flat biconformal geometry and Hamiltonian systems provides a clear physical interpretation of the geometric variables of biconformal space.

The central importance of this second result is in definitively establishing the physical interpretation of the new conformal gauging. Our main goal is not to provide an alternate formulation of classical mechanics, but rather to use this embedding as a guide to subsequent physical interpretation of the elementary biconformal variables. Given the result of Sec. IV, that a classical single-particle Hamiltonian system generates a class of biconformal space–times, and its converse in Sec. V, we can conclude that for an isolated test particle in a general biconformal space the four new coordinates may be interpreted locally as the corresponding generalized particle momentum. Our identification therefore provides a physical correspondence principle for biconformal spaces. While in general spaces the extra dimensions will not necessarily represent momentum globally, we can always take the limit of a tightly confined field in a local Lorentz frame, for which the extra dimensions will permit the momentum interpretation. Indeed, when nonflat biconformal spaces are investigated in Sec. VII, we will see that the ''momentumlike'' co-solder form contains the stress-energy tensor of any gravitational sources. This generalization from momentum to the stress-energy tensor is exactly what one would anticipate in moving from a particle interpretation to a field interpretation.

For our third result, in Sec. VI, we apply the Hamiltonian correspondence of the previous sections to the general solution for flat biconformal space, to show how that flat solution predicts the following properties of the electromagnetic vector potential:

(1) The four-vector form of the potential. This is nontrivial, since it involves the reduction of the eight-dimensional one-form $\omega_0^0=\omega_{0\mu}^0(x,y)\mathbf{d}x^\mu+\omega_0^{0\mu}(x,y)\mathbf{d}y_\mu$ to a four-dimensional one-form, $\alpha_\mu(x)\mathbf{d}x^\mu$ on space–time.
(2) Its usual gauge dependence $\alpha'(x)=\alpha(x)+\mathbf{d}\phi$.

(3) Minimal coupling $p_a \to p_a - \lambda \alpha_a$.
(4) The correct equation of motion for a charged particle moving under its influence.

This section therefore provides a consistent realization of Weyl's goal[1] *of expressing electrodynamics in terms of the dilational gauge field*. Upon moving to more general biconformal spaces, the remarks of the preceding paragraph imply that in a local Lorentz frame an isolated, charged test particle will move according to the Lorentz force law.

These predictions of electromagnetic effects in biconformal space follow directly from the relationship of biconformal space to phase space and Hamiltonian dynamics, developed in Secs. IV and V. No further assumption is necessary to derive the Lorentz force law. At the same time, the biconformal theory predicts constancy of size. There is an entire class of *flat* biconformal spaces which includes sufficient freedom in the gauge fields to account for the classical electrodynamics of a charged point particle in an arbitrarily specified background electromagnetic field. The flatness condition includes vanishing dilational curvature, so none of these geometries leads to size change of any kind. In sharp contrast to Weyl geometry in which vanishing dilation implies vanishing gauge vector, the dilational gauge vector in flat biconformal space is *required to be nonzero*.

Section VI ends with a brief comparison of the biconformal model with Weyl's original gauge theory and with the standard U(1) model of electromagnetism.

In Sec. VII, we show how to introduce arbitrary gravitational and electromagnetic sources so that the results of Sec. VI remain valid in the resulting class of curved biconformal spaces. The entire class has vanishing dilation, and the electromagnetic and gravitational fields satisfy the Maxwell and Einstein equations, respectively. *This section therefore provides a consistent realization of Weyl's goal of a unified geometric theory of gravitation and electromagnetism*. While other fundamental interactions have been discovered since Weyl's time, the inclusion of both of the large scale forces in a geometric theory must be considered a step in the right direction; moreover, biconformal spaces contain additional fields which are not studied here. These additional fields might have an interpretation as the additional interactions. In any case, there is always the possibility of adding internal symmetries beyond the conformal symmetry.

The final section consists of a detailed discussion of possible consequences of regarding energy and momentum variables as four of the coordinates in an eight-dimensional space. We consider the necessary isomorphism of mathematical structure, transformation properties, and dynamical laws that must hold between momentum space variables and the biconformal co-space variables. Collisions and interactions are discussed with attention given to continuity and the proper proximity of colliding particles. Finally, we point to some experimental results which suggest a coordinatelike behavior of momentum variables.

II. THE CONSTRUCTION OF SPACE–TIMES WITH LOCAL SYMMETRIES

Before defining and interpreting biconformal spaces at the end of this section, we require some background motivating their construction. While it is customary to begin the discussion of a differential geometry with the specification of a manifold and metric pair, $(\mathcal{M},g)$, our present interest lies in developing a space–time model beginning with locally determined symmetry considerations. This allows us to construct geometries which *a priori* possess specified local symmetries. It is in part the elements of this construction which motivate our choice of an eight-dimensional base manifold for conformal gauging instead of the usual four-dimensional picture. This in turn requires us to interpret the extra four dimensions. The interpretation of the extra coordinates in the zero curvature limit is accomplished in Secs. IV–VI.

We consider symmetry groups which include the local Lorentz symmetry of space–time, seeking a procedure for developing a connection on a space–time $(\mathcal{M},g)$ from a knowledge of an experimentally determined (i.e., ''global'') symmetry group. We use the well-known techniques of Cartan and Klein,[29–32,37–39] proceeding as follows. We begin with a Lie group, $\mathcal{G}$, and a

subgroup $\mathscr{G}_0$ called the isotropy subgroup. The isotropy subgroup should contain no normal subgroup of $\mathscr{G}$ if the full group is to act effectively and transitively on the base space. Defining a projection onto the base space as the quotient $\mathscr{G}/\mathscr{G}_0$, the group manifold becomes a fiber bundle, called by Klein an *elementary geometry*. The full group will act effectively and transitively on the base space of the elementary geometry. The base space will be a manifold if and only if the fibration is regular, i.e., there exists a neighborhood of each point of any fiber which that fiber intersects only once. This regularity condition holds for the groups we consider. This base space of the elementary geometry therefore provides a manifold upon which we now generalize the connection to a Cartan connection by introducing curvature.

The generalization to a Cartan connection occurs by only requiring the action of the full group to be detectable on curves instead of globally. Specifically, we may introduce into the structure equations any curvature two-forms consistent with the resulting Bianchi identities and the following requirement. Let $\mathscr{P}(\lambda)$ be any curve in the bundle, let ω^i_j be a connection on the bundle, and let $f:\mathscr{G}\to V^{N^2}$ be a linear representation of $\mathscr{G}$ by $N\times N$ matrices. Then integrating

$$\mathbf{d}f^i_j=\omega^k_j f^i_k$$

along $\mathscr{P}(\lambda)$ yields a group transformation $f^i_j(\lambda_0)$ at each point $\mathscr{P}(\lambda_0)$. The connection ω^i_j will be a Cartan connection if the transformation $f^i_j(\lambda)$ depends only on the projection of the curve $\mathscr{P}(\lambda)$ into the base space. This condition holds if and only if the curvatures are *horizontal*, i.e., bilinear in those connection forms which vanish on the fibers. The generalization to a curved connection occurs because different curves in the base space between the same pair of points can give different group elements.

If the connection is linear, we can express this condition in terms of mappings of orthonormal frames. Consider again a curve $\mathscr{P}(\lambda)$, an initial point $\mathscr{P}(0)$ on the curve, and a frame $\mathscr{E}_a(0)$ at the point. We then can demand that there shall exist a transformation from the full group $\mathscr{G}$ mapping the initial pair $(\mathscr{P}(0),\mathscr{E}_a(0))$ to a corresponding pair $(\mathscr{P}(\lambda),\mathscr{E}_a(\lambda))$ for any point $\mathscr{P}(\lambda)$ on the curve.

A. Gauging the Poincaré group

We illustrate the method using the Poincaré group. The group may be described locally by its structure equations (equivalent to its Lie algebra), which take the form

$$\mathbf{d}\omega^a_b=\omega^c_b\wedge\omega^a_c,\quad \mathbf{d}\omega^a=\omega^b\wedge\omega^a_b, \tag{2.1}$$

where the one-forms ω^a_b,ω^a $(a,b=1,2,3,4)$ span the ten-dimensional group manifold. The only subgroup $\mathscr{G}_0$ containing no normal subgroup of $\mathscr{G}$ but which does include the Lorentz group is the Lorentz group itself, since the translations form a normal subgroup. The quotient $\mathscr{G}/\mathscr{G}_0$ leads to a four-dimensional base space with Lorentz fibers. It is easy to see that the base space is Minkowski space, and therefore the elementary geometry is the bundle of orthonormal frames over Minkowski space–time.

Continuing with the Lorentz bundle, we alter the connection to that of a curved base space by adding curvature two-forms, Ω^a_b and Ω^a, to the structure equations. Clearly there is one curvature two-form for each generator of the original Lie group. Furthermore, each curvature component is horizontal, depending only on the forms ω^a which vanish on the fibers. Functionally, the curvature components depend only on the quotient space, $\mathscr{G}/\mathscr{G}_0$. Thus,

$$\mathbf{d}\omega^a_b=\omega^c_b\wedge\omega^a_c+\Omega^a_b,\quad \mathbf{d}\omega^a=\omega^b\wedge\omega^a_b+\Omega^a, \tag{2.2}$$

or, in more familiar notation,

$$\mathbf{R}^a{}_b = \mathbf{d}\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b \quad T^a = \mathbf{d}\mathbf{e}^a + \omega^a{}_b \wedge \mathbf{e}^b, \tag{2.2$'$}$$

where $\omega^a{}_b$ is the spin connection, $\mathbf{e}^a = \omega^a$ is the vierbein, $\mathbf{R}^a{}_b$ is the curvature two-form, and $\mathbf{T}^a$ is the torsion. Notice that we will always write the metric explicitly, so we need not offset upper and lower indices. The horizontality condition, that $\mathbf{R}^a{}_b = \mathbf{R}^a{}_{bcd}\mathbf{e}^c \wedge \mathbf{e}^d$ with no terms such as $\mathbf{R}^a{}_{bcd}{}^e \mathbf{e}^c \wedge \omega^d_e$ or $\mathbf{R}^a{}_{bcd}{}^{ef}\omega^c_e \wedge \omega^d_f$, guarantees that the generator of a Poincaré transformation found by integrating the connection one-forms along any curve $\mathscr{P}(\lambda)$ will depend only on the projection of the curve into the base manifold, $\pi(\mathscr{P}(\lambda))$. Specifically, define

$$\mathbf{d}_1 \mathscr{P} = \omega^a \mathscr{E}_a, \quad \mathbf{d}_1 \mathscr{E}_a = \omega^b_a \mathscr{E}_b, \tag{2.3}$$

as the change in the point $\mathscr{P}$ and the vector frame $\mathscr{E}_b$ along any curve in the bundle, where $\mathbf{d}_1$ is the one-dimensional exterior derivative on the curve. It is convenient to introduce a full ten-dimensional frame $(\mathscr{E}_a, \mathscr{F}_a{}^b)$ at the initial point, with the frame $\mathscr{E}_a$ being the horizontal part. We choose $(\mathscr{E}_a, \mathscr{F}_a{}^b)$ to be dual to (ω^a, ω^a_b):

$$\begin{aligned} \omega^a(\mathscr{E}_b) &= \delta^a_b, \quad \omega^a(\mathscr{F}_c{}^b) = 0, \\ \omega^a_b(\mathscr{E}_c) &= 0, \quad \omega^a_b(\mathscr{F}_d{}^c) = \delta^a_d \delta^c_b. \end{aligned} \tag{2.4}$$

When Eqs. (2.3) are integrated along a curve $\mathscr{P}(\lambda)$ we find a new point-frame pair at each value of λ. Now let $\mathscr{P}(\lambda)$ be the closed perimeter of an arbitrary infinitesimal plaquette with area element S. Then explicit evaluation of the integral of $\mathbf{d}_1\mathscr{P}$ around the plaquette gives

$$\oint_{\mathscr{P}} \mathbf{d}_1 \mathscr{P} \approx \Omega^a(S)\mathscr{E}_a(0). \tag{2.5}$$

A general infinitesimal surface element S may be expanded in terms of the full frame at the initial point as

$$S = S^{ab}\mathscr{E}_a \wedge \mathscr{E}_b + S^{ab}_c \mathscr{F}^c_a \wedge \mathscr{E}_b + S^{ab}_{cd} \mathscr{F}^c_a \wedge \mathscr{F}^d_b. \tag{2.6}$$

Now taking Ω^a horizontal, $\Omega^a = \Omega^a_{bc}\omega^b \wedge \omega^c$, we use Eqs. (2.4) to evaluate

$$\begin{aligned} \Omega^a(S) &= \Omega^a_{bc}[S^{de}\omega^b(\mathscr{E}_d)\omega^c(\mathscr{E}_e) + S^{de}_f \omega^b(\mathscr{F}^f_d)\omega^c(\mathscr{E}_e) + S^{de}_{fg}\omega^b(\mathscr{F}^f_d)\omega^c(\mathscr{F}^g_e)] \\ &= \Omega^a_{bc} S^{de}\delta^b_d \delta^c_e = \Omega^a(\pi(S)), \end{aligned} \tag{2.7}$$

so that the path dependence of the point $\mathscr{P}(\lambda)$ depends only on the projection of the loop into the base manifold. A completely analogous argument holds for the integral of $\mathscr{E}_a$ and the horizontality of Ω^a_b.

Equations (2.2) or (2.2$'$) now describe a curved four-dimensional space–time with local Lorentz symmetry. The second curvature, $\Omega^a = \mathbf{T}^a$, allows the inclusion of torsion. Because the remaining Lorentz symmetry of the fibers does not mix the components of the torsion with the components of the Riemann curvature, the usual specification of general relativity, $\Omega^a = 0$, is consistent. General relativity also requires identification of the cotangent space T^* with the space spanned by the solder form $\mathbf{e}^a$. This identification is automatic here because $\mathbf{e}^a$ is required *a priori* to span the base space $\mathscr{G}/\mathscr{G}_0$. As we have carried out the construction here, the topology of the manifold is that of the quotient, R^4, but nontrivial base spaces are easily achieved by specifying

an arbitrary manifold $\mathscr{M}$ having the same (Minkowski) cotangent space and structure equations, (2.2). This substitution of $\mathscr{M}$ for $\mathscr{G}/\mathscr{G}_0$ is allowed because it alters only topological properties, while all of the structures of interest described above are local.

B. Biconformal gauging

Understanding a symmetry-based approach has significant advantages as we now turn to our examination of the conformal group. The essential point is that, in contrast to the Poincaré example, the presence of inverse translations means that there are no normal subgroups, making the choice of isotropy subgroup $\mathscr{C}_0$ nontrivial. While biconformal space is based on the choice of the homogeneous Weyl group as $\mathscr{C}_0$, it is instructive to first consider what happens if we use the inhomogeneous Weyl group instead.

As described in the Introduction, considerable attention[23–27] has been given to models which result from the choice of the 11-dimensional inhomogeneous Weyl group as $\mathscr{C}_0$. This choice is natural enough since it gives the same base manifold as in the Poincaré case. It would appear that this model simply extends the Lorentz fiber symmetry of the local Poincaré model to include dilations and inverse translations. However, in generic field theories based on this bundle the four components of the connection corresponding to the inverse translations are generically auxiliary and may be algebraically removed from the problem.[28] The resulting field theory, equivalent to one based on the homogeneous Weyl group, has had at best mixed success as a field theory. In any case, no new symmetries or additional physical fields have been gained in passing from the Weyl group to the full conformal group. By contrast, choosing $\mathscr{C}_0$ to be the *homogeneous* Weyl group from the start, we retain the full 15 degrees of freedom of the original group and acquire many new fields, with the simplest example of a biconformal space having a natural physical interpretation in terms of Hamiltonian dynamics.

We arrive at the biconformal model if we ask what property of the Lorentz transformations made them a suitable fiber symmetry for the Poincaré group. The requirement that the isotropy subgroup should have no subgroup normal in the Poincaré group ruled out the translations, which form a normal subgroup of the Poincaré group. The Lorentz symmetry is therefore the only possible fiber symmetry. However, there is another way to see that we should use the Lorentz group as the isotropy subgroup, based on fixed points. On Minkowski space, the class of translations has no fixed points, while the class of Lorentz transformations leaves the origin fixed (hence the name, ''isotropy subgroup''). Thus, we can distinguish the isotropy subgroup of the fibering from the ''translational'' symmetry of the base manifold by counting fixed points.

Returning to the conformal group, we find that when we distinguish the conformal transformations based on their fixed points, there are simply eight translations acting on *compactified* Minkowski space.[40] A special point, its null cone, and an ideal two-sphere are added at infinity to accomplish the compactification. As a result, the translations are no longer characterized by an absence of fixed points. Instead, the class of translations and the class of inverse translations each has a single fixed point (the origin and the point at infinity, respectively) while the Lorentz subgroup and the dilational subgroup leave both of these points fixed. The dilations also leave the ideal two-sphere fixed.

With these observations, we take the isotropy subgroup $\mathscr{C}_0$ to be the seven-dimensional homogeneous Weyl group, consisting of the six Lorentz transformations together with dilations. Defining a projection as the quotient $\mathscr{C}/\mathscr{C}_0$ we are led to a fiber bundle with an eight-dimensional manifold as the base space and the homogeneous Weyl group as a typical fiber. This will break *both* the translational and inverse translational symmetries when the base manifold becomes curved, but we nonetheless retain the full 15 guage field degrees of freedom.

We choose the O(4,2) representation of the conformal group for our notation,[40] with $(A,B,\ldots)=(0,1,...,5)$. Letting boldface or Greek symbols denote forms and $(a,b,\ldots)=(1,...,4)$, the O(4,2) metric η_{AB} is given by $\eta_{ab}=\mathrm{diag}\,(1,1,1,-1)$ and $\eta_{05}=\eta_{50}=1$ with all other compo-

nents vanishing. Introducing the connection one-form ω_B^A, we may express the covariant constancy of η_{AB} as

$$\mathbf{D}\eta_{AB} \equiv \mathbf{d}\eta_{AB} - \eta_{CB}\omega_A^C - \eta_{AC}\omega_B^C = 0. \tag{2.8}$$

We may break the connection form into four independent Weyl-invariant parts: the *spin connection*, ω_b^a, the *solder form*, ω_0^a, the *co-solder form*, ω_a^0, and the *Weyl vector*, ω_0^0, where the spin connection satisfies

$$\omega_b^a = -\eta_{bc}\eta^{ad}\omega_d^c \tag{2.9a}$$

and the remaining components of ω_B^A are given in terms of these by

$$\omega_0^5 = \omega_5^0 = 0, \tag{2.9b}$$

$$\omega_5^5 = -\omega_0^0, \tag{2.9c}$$

$$\omega_5^a = -\eta^{ab}\omega_b^0, \tag{2.9d}$$

$$\omega_a^5 = -\eta_{ab}\omega_0^b. \tag{2.9e}$$

These constraints reduce the number of independent fields ω_B^A to the required 15 and allow us to restrict $(A,B,\ldots) = (0,1,...,4)$ in all subsequent equations. Now it becomes clear why we write the metric explicitly instead of raising and lowering indices. The convention avoids confusion between the solder and co-solder forms. Staying with the O(4,2) index positions also indicates immediately the scaling properties of any field, since each upper index adds $+1$ to the conformal weight while each lower index adds -1.

The structure constants of the conformal Lie algebra now lead immediately to the Maurer–Cartan structure equations of the conformal group. These are simply

$$\mathbf{d}\omega_B^A = \omega_B^C \wedge \omega_C^A. \tag{2.10}$$

When broken into parts based on homogeneous Weyl transformation properties, Eq. (2.10) gives

$$\begin{aligned} \mathbf{d}\omega_b^a &= \omega_b^c \wedge \omega_c^a + \omega_b^0 \wedge \omega_0^a - \eta_{bc}\eta^{ad}\omega_d^0 \wedge \omega_0^c, \\ \mathbf{d}\omega_0^a &= \omega_0^0 \wedge \omega_0^a + \omega_0^b \wedge \omega_b^a, \\ \mathbf{d}\omega_a^0 &= \omega_a^0 \wedge \omega_0^0 + \omega_a^b \wedge \omega_b^0, \quad \mathbf{d}\omega_0^0 = \omega_0^a \wedge \omega_a^0. \end{aligned} \tag{2.11}$$

Since no finite translation can reach the point at infinity and no inverse translation can reach the origin, the space $\mathscr{C}/\mathscr{C}_0$ gives a copy of (noncompact) Minkowski space for each of the two sets of translations. The generalization to a curved base space is immediate. We have

$$\mathbf{d}\omega_b^a = \omega_b^c \wedge \omega_c^a + \omega_b^0 \wedge \omega_0^a - \eta_{bc}\eta^{ad}\omega_d^0 \wedge \omega_0^c + \Omega_b^a, \tag{2.12a}$$

$$\mathbf{d}\omega_0^a = \omega_0^0 \wedge \omega_0^a + \omega_0^b \wedge \omega_b^a + \Omega_0^a, \tag{2.12b}$$

$$\mathbf{d}\omega_a^0 = \omega_a^0 \wedge \omega_0^0 + \omega_a^b \wedge \omega_b^0 + \Omega_a^0, \tag{2.12c}$$

$$\mathbf{d}\omega_0^0 = \omega_0^a \wedge \omega_a^0 + \Omega_0^0. \tag{2.12d}$$

We will call the four types of curvature Ω^a_b, Ω^a_0, Ω^0_a, and Ω^0_0 the Riemann curvature, torsion, co-torsion, and dilational curvature, respectively. Notice that if we set ω^0_a, ω^0_0, and the corresponding curvatures to zero, we recover Eqs. (2.2) for a four-dimensional space–time with Riemannian curvature Ω^a_b and torsion Ω^a_0. If we set only $\omega^0_a=\Omega^0_a=0$, the structure equations are those of four-dimensional Weyl geometry.

We are now in a position to define a biconformal space. We state the definition for an arbitrary number of dimensions, $2d$, although our results in all subsequent sections are given using $d=4$ for concreteness.

Definition: A *biconformal space* is a principal fiber bundle (P,M,G) with conformal connection form $\omega=\{\omega^a_b,\omega^a_0,\omega^0_a,\omega^0_0\}$ $(a,b=1,...,d)$, where M is the $2d$-dimensional base manifold, G is the Weyl group of d-dimensional Minkowski space, and $\pi:\omega\rightarrow(\omega^a_0,\omega^0_a)$, where π is the canonical projection induced by $\mathscr{P}/\mathscr{G}$.

Thus, just as we can think of a space–time as a manifold and metric pair $(\mathscr{M},g)$, we can think of a biconformal space as a manifold and connection pair $(\mathscr{M},\omega)$. Note that in general relativity, the implicit demand for metric compatibility allows us to generate a unique connection, while at present there is no such theorem for biconformal spaces. We can go the other way, however, since the existence of a biconformal metric follows by combining the horizontal connection forms, (ω^a_0,ω^0_a), with the Killing metric of the conformal group.

For many purposes, the starting point of computation in a biconformal space is given by the structure equations, Eqs. (2.12). Once we specify the connection, these structure equations allow us to compute the curvatures. Conversely, given horizontal curvatures satisfying the Bianchi identities [which follow from Eqs. (2.12) by differentiation], we can integrate to find the connection. Certain combinations are also possible. For example, it is easy to see from Eqs. (2.12) that in torsion-free biconformal spaces, the entire connection and the remaining curvatures are determined by ω^a_0, ω^0_a, and ω^0_0 alone.

Horizontality requires each of the curvatures to take the form

$$\Omega^A_B=\tfrac{1}{2}\Omega^A_{Bcd}\omega^c_0\wedge\omega^d_0+\Omega^{Ac}_{Bd}\omega^d_0\wedge\omega^0_c+\tfrac{1}{2}\Omega^{Acd}_B\omega^0_c\wedge\omega^0_d. \tag{2.13}$$

Based on the interpretation of biconformal space as a generalization of phase space (Secs. IV and V below, see also Ref. 41) we will call Ω^A_{Bcd} the space–time term, Ω^{Ac}_{Bd} the cross term, and Ω^{Acd}_B the momentum term of each type of curvature. In sharp contrast to the four-dimensional gauging, in which the fiber symmetry mixes the curvatures Ω^a_b, Ω^a_0, Ω^0_a, and Ω^0_0 (see Ref. 42), our choice of the homogeneous Weyl group as the structure group not only leaves these independent, but also does not mix the space–time, cross, or momentum terms. The usefulness of this type of isolation of curvature parts is evident in general relativity, where it is consistent with the bundle structure to set the torsion to zero.

Since the base space M is a manifold, we can always introduce coordinates (x^μ,y_ν) so that the connection forms may be written as

$$\omega^A_B=\omega^A_{B\mu}(x,y)\mathbf{d}x^\mu+\omega^{A\mu}_B(x,y)\mathbf{d}y_\mu.$$

Here the capital Latin indices range over $(0,...,4)=(0,a)$ while both Greek and lower case Latin indices range from $(1,...,4)$. The Greek indices apply to general coordinates, while the lower case Latin indices are used for the Weyl invariant parts of the orthonormal O(4,2) frame introduced above. Index positions indicate conformal weight, so all metrics are written explicitly. An exception to these conventions is made in Secs. III and VI where, for the flat solution, the components of the vierbein are δ^a_μ. In these sections we simply write, for example, (x^a,y_b) in place of $(\delta^a_\mu x^\mu,\delta^\nu_b y_\nu)$.

III. FLAT BICONFORMAL SPACE

Before deriving the form of the connection of a flat biconformal space, we discuss a few of its properties.

Definition: The connection of a flat biconformal space is said to be in the *standard flat form* when it is written in the following way:

$$\omega_0^0 = \alpha_a(x)\mathbf{d}x^a - y_a\mathbf{d}x^a \equiv W_a\mathbf{d}x^a, \tag{3.1a}$$

$$\omega_0^a = \mathbf{d}x^a, \tag{3.1b}$$

$$\omega_a^0 = \mathbf{d}y_a - (\alpha_{a,b} + W_aW_b - \tfrac{1}{2}W^2\eta_{ab})\mathbf{d}x^b, \tag{3.1c}$$

$$\omega_b^a = (\eta^{ac}\eta_{bd} - \delta_d^a\delta_b^c)W_c\mathbf{d}x^d. \tag{3.1d}$$

Notice that the Weyl vector, $W_a = \alpha_a(x) - y_a$, depends on an arbitrary four-vector α_a and *also* on the additional four coordinates y_a. The presence of α_a gives the generality required for the electromagnetic vector potential, while the y_a keeps the dilational curvature zero. As a result, unlike Weyl's theory, the flat biconformal model predicts no size change. Also notice that the standard flat form is preserved by *four-dimensional* gauge transformations, $\phi(x)$, and that the gauge transformation must be associated with the undetermined vector field, $\alpha_a(x)$. Thus, the desirable properties of Weyl's original theory survive in this more general gauge theory.

The prediction of the exact form of the Weyl vector necessary for consistently modeling electromagnetism is nontrivial. In general, the dilational gauge vector of biconformal space is of the form

$$\omega_0^0 = \omega_{0\mu}^0(x,y)\mathbf{d}x^\mu + \omega_0^{0\mu}(x,y)\mathbf{d}y_\mu,$$

i.e., an eight-dimensional vector field depending on eight independent variables. Constraining the biconformal geometry to have vanishing curvatures forces $\omega_0^0 = (\alpha_\mu(x) - y_\mu)\mathbf{d}x^\mu$. This is precisely the form required to give the Lorentz force law using the independently established formulation of Hamiltonian dynamics in biconformal space. Different field strengths $\alpha_{[\mu,\nu]}$ are in 1–1 correspondence with the possible flat biconformal spaces, so that with the interpretation of α_μ as the vector potential, electromagnetic phenomena never lead to dilations.

We now turn to our first result.

Theorem: When the curvature of biconformal space vanishes, $\Omega_B^A = 0$, there exist coordinates (x^a, y_a) such that the connection takes the standard flat form. When the topology is that of R^8 the coordinates are global.

Proof: Imposing vanishing curvature, $\Omega_B^A = 0$, the equations to be solved take the form

$$\mathbf{d}\omega_b^a = \omega_b^c \wedge \omega_c^a + \omega_b^0 \wedge \omega_0^a - \eta_{bc}\eta^{ad}\omega_d^0 \wedge \omega_0^c, \tag{3.2a}$$

$$\mathbf{d}\omega_0^a = \omega_0^0 \wedge \omega_0^a + \omega_0^b \wedge \omega_b^a, \tag{3.2b}$$

$$\mathbf{d}\omega_a^0 = \omega_a^0 \wedge \omega_0^0 + \omega_a^b \wedge \omega_b^0, \tag{3.2c}$$

$$\mathbf{d}\omega_0^0 = \omega_0^a \wedge \omega_a^0. \tag{3.2d}$$

The system may be solved by making use of the involution of Eq. (3.2b).[38] This allows us to consistently set $\omega_0^a = 0$ and first solve on the subspace spanned by the remaining 11 one-forms. The initial conditions for these integral submanifolds provide a coordinate x^a such that $\omega^a = \mathbf{d}x^a$ with $x^a = \text{const}$ on each leaf. Each leaf is then a fiber bundle described by the simpler set of structure equations

$$\mathbf{d}\omega_b^a = \omega_b^c \wedge \omega_c^a , \tag{3.3a}$$

$$\mathbf{d}\omega_a^0 = \omega_a^0 \wedge \omega_0^0 + \omega_a^b \wedge \omega_b^0, \tag{3.3b}$$

$$\mathbf{d}\omega_0^0 = 0, \tag{3.3c}$$

which may be recognized as those of a flat Weyl geometry. Equation (3.3c) implies a pure-gauge form for the Weyl vector. Choosing the gauge so that the Weyl vector vanishes (on the $\omega_0^a = 0$ subspace), we are left with the structure equations for a flat Riemannian geometry. Clearly, we have the solution

$$\omega_b^a = 0, \quad \omega_a^0 = \mathbf{d}y_a , \quad \omega_0^0 = 0. \tag{3.4}$$

Next, we reintroduce the remaining four independent one-forms,

$$\omega_0^a = \mathbf{d}x^a, \tag{3.5}$$

so that (x^a, y_a) provide a set of eight independent coordinates. From the linearity of the connection one-forms in the coordinate differentials, ω_b^a, ω_a^0, and ω_0^0 will change only by terms proportional to $\mathbf{d}x^a$. We therefore may write

$$\omega_b^a = C^a{}_{bc}\mathbf{d}x^c, \tag{3.6a}$$

$$\omega_0^a = \mathbf{d}x^a, \tag{3.6b}$$

$$\omega_a^0 = \mathbf{d}y_a + B_{ab}\mathbf{d}x^b, \tag{3.6c}$$

$$\omega_0^0 = W_a \mathbf{d}x^a, \tag{3.6d}$$

where the coefficients W_a, B_{ab}, and $C^a{}_{bc}$ are functions of x^a and y_a to be found by substitution into the full structure equations, Eqs. (3.2). First we determine the form of the Weyl vector, W_a, from Eq. (3.2d). Substitution of Eqs. (3.6b)–(3.6d) yields

$$\mathbf{d}\omega_0^0 = W_{d'}{}^b \mathbf{d}y_b \wedge \mathbf{d}x^a + W_{a,b}\mathbf{d}x^b \wedge \mathbf{d}x^a = \mathbf{d}x^a \wedge (\mathbf{d}y_a + B_{ab}\mathbf{d}x^b). \tag{3.7}$$

Here a lowered comma denotes a partial derivative with respect to x^b while a raised comma denotes a partial with respect to y_b. Equating like components gives

$$W_a{}^{,b} = -\delta_a^b, \tag{3.8a}$$

$$B_{[ab]} = W_{[b,a]}. \tag{3.8b}$$

The first of these is immediately integrated to give

$$W_a = -y_a + \alpha_a(x). \tag{3.9}$$

Next we substitute Eqs. (3.6a) (3.6b), and (3.6d) into Eq. (3.2b) to find $C^a{}_{bc}$ in terms of W_a:

$$\mathbf{d}\omega_0^a = 0 = W_b \mathbf{d}x^b \wedge \mathbf{d}x^a + \mathbf{d}x^b \wedge C^a{}_{bc}\mathbf{d}x^c \tag{3.10}$$

or

$$C^a{}_{[bc]} + W_{[b}\delta_{c]}^a = 0. \tag{3.11}$$

Lowering the upper index in Eq. (3.11) and noting that $C_{abc} \equiv \eta_{ad} C^{d}{}_{bc} = -C_{bac}$, we add two even permutations and subtract the third to find

$$C^{a}{}_{bc} = -(\delta^{a}_{c}\delta^{d}_{b} - \eta^{ad}\eta_{bc}) W_{d}. \tag{3.12}$$

Finally, B_{ab} is found from Eq. (3.2a). The substitution leads to two independent equations, one from $\mathbf{d}x^{a} \wedge \mathbf{d}y_{b}$ cross-terms and the other from terms quadratic in $\mathbf{d}x^{a}$. The first is

$$(\delta^{a}_{c}\delta^{d}_{b} - \eta^{ad}\eta_{bc}) W_{d}{}^{,e} = (\eta^{ae}\eta_{bc} - \delta^{a}_{c}\delta^{e}_{b}),$$

which is identically satisfied by the form of Eq. (3.9) for W_{a}. Writing $W^{a} \equiv \eta^{ab} W_{b}$ and $W^{2} \equiv W^{a} W_{a}$, the second equation becomes

$$\begin{aligned}\delta^{a}_{d} B_{bc} - \delta^{a}_{c} B_{bd} - B^{a}_{d}\eta_{bc} + B^{a}_{c}\eta_{bd} &= (W^{a}{}_{,c} + W^{a} W_{c})\eta_{bd} - (W^{a}{}_{,d} + W^{a} W_{d})\eta_{bc} - (W_{b,c} + W_{b} W_{c})\delta^{a}_{d} \\ &\quad + (W_{b,d} + W_{b} W_{d})\delta^{a}_{c} - W^{2}(\delta^{a}_{c}\eta_{bd} - \delta^{a}_{d}\eta_{bc}).\end{aligned} \tag{3.13}$$

Contraction on the a and c indices gives an expression containing B_{ab} and its trace $B = \eta^{ab} B_{ab}$. A second trace yields $B = W^{2} - W^{a}_{,a}$. B_{ab} is then found to be

$$B_{ab} = -(W_{a,b} + W_{a} W_{b} - \tfrac{1}{2} W^{2}\eta_{ab}).$$

The antisymmetric part of this expression agrees with Eq. (3.8b). The form for B_{ab} above is now found to solve both Eq. (3.13) and the remaining structure equation, Eq. (3.2c), identically.

We conclude that in flat biconformal space there locally exist coordinates x^{a} and y_{a} such that

$$\omega^{0}_{0} = \alpha_{a}(x)\mathbf{d}x^{a} - y_{a}\mathbf{d}x^{a}, \tag{3.1a}$$

$$\omega^{a}_{0} = \mathbf{d}x^{a}, \tag{3.1b}$$

$$\omega^{0}_{a} = \mathbf{d}y_{a} - (\alpha_{a,b} + W_{a} W_{b} - \tfrac{1}{2} W^{2}\eta_{ab})\mathbf{d}x^{b}, \tag{3.1c}$$

$$\omega^{a}_{b} = -(\delta^{a}_{d}\delta^{c}_{b} - \eta^{ac}\eta_{bd}) W_{c}\mathbf{d}x^{d}. \tag{3.1d}$$

That the coordinates (x^{a}, y_{a}) are global follows immediately from the R^{8} topology of the base manifold, completing the proof.

We are now in a position to see in more detail how the biconformal structure corrects Weyl's theory. If we hold the y-coordinate constant in Eqs. (3.1), we find that Eqs. (3.1a), (3.1b), and (3.1d) are the connection forms for a four-dimensional Weyl space–time with conformally flat metric η_{ab}. The remaining expression, Eq. (3.1c), is then simply a one-form constructed from the Weyl–Ricci tensor. However, the dilational curvature of a four-dimensional Weyl geometry is given by the curl of the Weyl vector, equivalent here to the arbitrary curl of α_{a}. Thus, when viewed from Weyl's four-dimensional perspective the solution gives unphysical size change. It is only with the inclusion of the additional momentum variables proportional to y_{a} that the dilational curvature can be seen to vanish.

With the interpretation of the additional four dimensions of biconformal space as momentum variables it is clear that the actual motion of a particle is eight-dimensional. Now consider an experiment designed to detect a change in the relative size of two physical objects, for example, a pair of identical atoms. In order to see a change, a comparison must be made before and after moving the two atoms around some closed space–time loop. However, such motion necessarily involves changes in momentum as well so there is necessarily a change in y_{a} as well as x^{a}, inducing a corresponding loop in flat biconformal space. However, no loop in flat biconformal space ever encloses nonzero dilational flux, or ever results in a measurable size change. The size

change computed in the four-dimensional Weyl geometry is seen to be in error because it involves forcing a closed path while holding the momentum variables constant. Equivalently, the error occurs because the expression for the dilational curvature is incomplete.

We show in the next two sections that there is a direct correspondence between the geometric variables of biconformal geometry and the physical phase space variables of Hamiltonian dynamics. Then, in Sec. VI we apply the Hamiltonian correspondence to the general flat biconformal solution. This leads without further assumption to the Lorentz law of force and the identification of α_a with the electromagnetic vector potential.

IV. HAMILTONIAN DYNAMICS IN FLAT BICONFORMAL SPACE

Leaving biconformal geometry for the moment, we turn to a geometric approach to classical Hamiltonian dynamics. We first show that the action of a classical one-particle system may be used to generate biconformal spaces. Then, in Sec. V, we give a unique prescription for generating a *flat* biconformal space. Begin with eight-dimensional extended phase space with canonical coordinates $(x^i,t;p_i,p_4)$ where $(i,j=1,2,3)$. We assume we are given a super-Hamiltonian, $\mathscr{H}=\mathscr{H}(x^a,p_a)$.[43] Imposing the constraint $\mathscr{H}=0$ then gives p_4 as a function of the remaining seven variables, $p_4=-H(p_i,x^i,t)$, with H the usual Hamiltonian. This constraint insures that time appears as a parameter rather than an independent dynamical variable. Now consider the Hilbert form

$$\omega = H\mathbf{d}t - p_i\mathbf{d}x^i. \tag{4.1}$$

The integral of ω is the action functional. The exterior derivative of ω may always be factored:

$$\mathbf{d}\omega = \frac{\partial H}{\partial x^i}\,\mathbf{d}x^i\wedge\mathbf{d}t + \frac{\partial H}{\partial p_i}\,\mathbf{d}p_i\wedge\mathbf{d}t + \mathbf{d}x^i\wedge\mathbf{d}p_i = \left(\mathbf{d}x^i - \frac{\partial H}{\partial p_i}\,\mathbf{d}t\right)\wedge\left(\mathbf{d}p_i + \frac{\partial H}{\partial x^i}\,\mathbf{d}t\right). \tag{4.2}$$

Therefore, if we define

$$\begin{gathered}\omega^i \equiv \left(\mathbf{d}x^i - \frac{\partial H}{\partial p_i}\,\mathbf{d}t\right), \quad \omega_i \equiv \left(\mathbf{d}p_i + \frac{\partial H}{\partial x^i}\,\mathbf{d}t\right), \\ \omega^4 \equiv \mathbf{d}t, \quad \omega_4 \equiv \mathbf{d}\mathscr{H} = 0,\end{gathered} \tag{4.3}$$

then we can write

$$\mathbf{d}\omega = \omega^a\wedge\omega_a = \omega^i\wedge\omega_i\,. \tag{4.4}$$

This factoring is clearly preserved by local symplectic transformations of the six-basis (ω^i,ω_i), as well as reparametrizations of the time. Obviously these transformations include the usual canonical transformations of coordinates as a special case. One class of such allowed transformations of basis is achieved by the addition of $c_{ab}\omega^b$ to ω_a, where $c_{ab}=c_{ba}$. For the moment we take $c_{ab}=0$, but below we show the existence of a unique choice of c_{ab} that leads to a flat biconformal space.

We may also define a connection one-form, ω^a_b. This choice is uniquely determined by requiring the resulting biconformal space to be flat, but first we show that every possible choice leads to some biconformal space. Without imposing flatness any choice is possible. For this most general case, let ω^a_b be any linear combination of ω^a_0 and ω^0_a, and define the (necessarily horizontal) curvatures to be

$$\Omega^a_b \equiv \mathbf{d}\omega^a_b - \omega^c_b\wedge\omega^a_c - \omega^0_b\wedge\omega^a_0 - \eta_{bc}\,\eta^{ad}\omega^c_0\wedge\omega^0_d, \tag{4.5a}$$

$$\Omega_0^a \equiv \mathbf{d}\omega_0^a - \omega \wedge \omega_0^a - \omega_0^b \wedge \omega_b^a, \tag{4.5b}$$

$$\Omega_a^0 \equiv \mathbf{d}\omega_a^0 - \omega_a^0 \wedge \omega - \omega_a^b \wedge \omega_b^0, \tag{4.5c}$$

$$\Omega_0^0 \equiv \mathbf{d}\omega - \omega_0^a \wedge \omega_a^0 \equiv 0. \tag{4.5d}$$

The necessary presence and form of Eq. (4.4) for $\mathbf{d}\omega$, and the dependence of the curvatures on *both* ω^a and ω_a clearly show this to be a dilationally flat (i.e., $\Omega_0^0 = 0$) biconformal space. Therefore, extended phase space together with a Hamiltonian symplectic structure may be viewed as a certain kind of biconformal space. Turning this around, *we can interpret biconformal space, and therefore conformal gauge theory, as a generalization of one-particle phase space.*

It is worth pointing out that biconformal spaces, while including Hamiltonian extended phase spaces as special cases, also contain all four-dimensional pseudo-Riemannian geometries as other special cases. We therefore have a differential geometry rich enough to describe both general relativity and Hamiltonian particle dynamics. Furthermore, as we show in Sec. VII, biconformal spaces include the even larger class of all four-dimensional Weyl geometries, and this class allows the consistent geometric unification of gravity and electromagnetism.

The presence of both Hamiltonian and Riemannian structures is reassuring, since it is one aim of the study of biconformal spaces to place relativity theory and quantum theory in a common mathematical framework. While the developments here suggest only the possibility of describing, perhaps, a quantum particle in a curved background, the full picture is actually somewhat better because the symplectic two-form of the Hamiltonian structure gives a complex structure to the tangent/cotangent space and an almost complex structure to the biconformal space itself. The biconformal structure therefore gives a natural complexification of space–time in such a way that the usual real structure is immediately evident. One might hope, for example, to see some special significance for the Ashtekar connection when biconformal space is expressed in terms of the SU(2,2) conformal covering group instead of O(4,2). Whether these hopes are realized or not is the subject of current study.

We also note here that the idea of a quantum interpretation of conformal geometry agrees in some ways with earlier proposals[44–47] relating phase space, Weyl geometry, and quantum physics. These proposals, however, lack the full geometric structure of conformal gauge theory, do not demonstrate the intrinsically biconformal structure of Hamiltonian systems, and use a different inner product than that proposed in Sec. VI (see also Ref. 41).

We next show the relationship of the geometry described by Eqs. (4.5) to classical mechanics.

Since the curvatures are two-forms and because $\mathbf{d}\mathscr{H} = 0$, each of the curvatures, Ω^a and Ω_a, is necessarily at least linear in one of the six basis forms (ω^i, ω_i). Thus, six of the structure equations, Eqs. (4.5b) and (4.5c), are in involution. This involution allows us to set $\omega^i = \omega_i = 0$, thereby singling out a fibration of the bundle by one-dimensional submanifolds, i.e., the classical paths of motion. Examination of Eq. (4.3) shows that these conditions simply give Hamilton's equations of motion:

$$\mathbf{d}x^i = \frac{\partial H}{\partial p_i}\,\mathbf{d}t, \quad \mathbf{d}p_i = -\frac{\partial H}{\partial x^i}\,\mathbf{d}t. \tag{4.6}$$

The Fröbenius theorem guarantees the existence of solutions to these equations for the paths. The remaining two structure equations then reduce to

$$\mathbf{d}\omega_j^i = \mathbf{d}\omega = 0, \tag{4.7}$$

which are identically satisfied on curves.

It is of interest to further note that on the full bundle the condition that ω be exact is the Hamiltonian–Jacobi equation, since we may then write $\omega = \mathbf{d}S$. Substituting for ω and expanding $\mathbf{d}S$ gives

$$-H\mathbf{d}t + p_i \mathbf{d}x^i = \frac{\partial S}{\partial x^i}\,\mathbf{d}x^i + \frac{\partial S}{\partial p_a}\,\mathbf{d}p_a + \frac{\partial S}{\partial t}\,\mathbf{d}t, \tag{4.8}$$

so that

$$\frac{\partial S}{\partial p_i} = 0, \quad \frac{\partial S}{\partial p_4} = 0 \tag{4.9}$$

$$\frac{\partial S}{\partial x^i} = p_i, \tag{4.10}$$

and

$$H\left(\frac{\partial S}{\partial x^i}, x^i, t\right) = -\frac{\partial S}{\partial t}. \tag{4.11}$$

Therefore, since ω is the Weyl vector of the generated biconformal space, the Hamilton–Jacobi equation holds if and only if the Weyl vector is pure gauge, $\omega = \mathbf{d}S$. A gauge transformation reduces the Weyl vector to zero. Since the dilational curvature is always zero for the geometry built from a Hamiltonian system, when $\mathbf{d}\omega = 0$ we also have

$$\omega^i \wedge \omega_i = 0, \tag{4.12}$$

implying three linear dependences between these six solder forms. Together with the vanishing of $\mathbf{d}\mathscr{H}$, the Hamilton–Jacobi equation therefore specifies a four-dimensional subspace of the full biconformal space.

V. FLAT BICONFORMAL SPACE AND THE HAMILTONIAN GEOMETRY

In this section, we show how to specify a unique *flat* biconformal space for a given Hamiltonian system. This is achieved by making a judicious choice of the solder and co-solder forms.

This time we initially regard the super-Hamiltonian as an unconstrained function of all eight coordinates. The Hamiltonian, H, is again defined as the solution for $-p_4$ when $\mathscr{H}$ is zero. Thus, when $\mathscr{H}$ is unconstrained, p_4 is also variable and the Hilbert form generalizes to

$$\omega_0^0 = -p_a \mathbf{d}x^a = -p_4(\mathscr{H}, H)\mathbf{d}t - p_i \mathbf{d}x^i. \tag{5.1}$$

Then

$$\mathbf{d}\omega_0^0 = \mathbf{d}x^a \wedge \mathbf{d}p_a, \tag{5.2}$$

where we reserve the symbols ω, ω_i, and ω^i for the special case when $\mathscr{H} = 0$ and $p_4 = -H(x^i, t, p_i)$.

The solder and co-solder forms ω_0^a and ω_a^0 are now identified by comparing the expression for $\mathbf{d}\omega_0^0$ to the flat biconformal solution, Eqs. (3.1). To make Eq. (3.1a) agree with Eq. (5.1) we must have $\alpha_a(x) = 0$ and $y_a = p_a$. We can set $\alpha_a = 0$ with a gauge change as long as $\alpha_{[a,b]} = 0$. Then, with $y_a = p_a$, the remainder of the connection is fully determined to be

$$\omega_0^0 = -p_a \mathbf{d}x^a, \quad \omega_0^a = \mathbf{d}x^a,$$

$$\omega_a^0 = \mathbf{d}p_a - (p_a p_b - \tfrac{1}{2}p^2 \eta_{ab})\mathbf{d}x^b, \tag{5.3}$$

$$\omega_b^a = -(\delta_d^a \delta_b^c - \eta^{ac}\eta_{bd})p_c \mathbf{d}x^d.$$

Note that canonical changes of variables do not change ω_0^0 by more than a scale change, $\mathbf{d}\phi$, so this form of the connection is correct for any canonical variables $(x^i, p_i) \to (q^i, \pi_i)$.

We now restrict to the hypersurface $\mathscr{H} = 0$ so that $p_4 = -H(p_i, x^i, t)$. The symmetric coefficient

$$c_{ab} = p_a p_b - \tfrac{1}{2}p^2 \eta_{ab} \tag{5.4}$$

provides a symplectic change of basis with respect to the manifestly closed and nondegenerate two-form, $\mathbf{d}\omega_0^0 = \omega_0^a \wedge \omega_a^0 = \mathbf{d}x^a \wedge \mathbf{d}p_a$. The differential of ω_0^0 is still seen to factor as in Eq. (4.2) either directly by differentiation or by substitution of $p_4 = -H(p_i, x^i, t)$ into $\omega_0^a \wedge \omega_a^0$, with ω_0^a and ω_a^0 given by Eqs. (5.3). The involution of Eqs. (4.5b) and (4.5c) for ω_i and ω^i still holds, with the classical curves given by $\omega_i = \omega^i = 0$. We therefore recover the Hamiltonian system, and have shown it to lie in a unique flat biconformal space.

The resulting curves in the biconformal space are easily found by first writing the frame field in terms of ω^i and ω_i:

$$\begin{aligned}
\omega_0^0 &= \left(H - p_i \frac{\partial H}{\partial p_i}\right)\mathbf{d}t - p_i \omega^i = L(\dot{x}^i, x^i, t)\mathbf{d}t - p_i \omega^i, \\
\omega_0^i &= \omega^i + \frac{\partial H}{\partial p_i}\,\mathbf{d}t, \\
\omega_0^4 &= \mathbf{d}t, \\
\omega_i^0 &= \omega_i - (p_i p_j + \tfrac{1}{2}(H^2 - p^2)\delta_{ij})\omega^j + \left(p_i H - \frac{\partial H}{\partial x^i} - \frac{1}{2}(H^2 - p^2)\delta_{ij}\frac{\partial H}{\partial p_j}\right)\mathbf{d}t, \\
\omega_4^0 &= \left(Hp_i - \frac{\partial H}{\partial x^i}\right)\omega^i + \left(H\frac{\partial H}{\partial p_i}p_i - \frac{1}{2}(H^2 - p^2) - \frac{\partial H}{\partial x^i}\frac{\partial H}{\partial p_i}\right)\mathbf{d}t, \\
\omega_j^i &= -(\delta_l^i \delta_j^k - \delta^{ik}\eta_{jl})p_k\left(\omega^l + \frac{\partial H}{\partial p_l}\,\mathbf{d}t\right), \\
\omega_4^i &= H\omega^i - \left(\delta^{ij}p_j - H\frac{\partial H}{\partial p_i}\right)\mathbf{d}t, \\
\omega_i^4 &= -\delta_{ij}\omega_4^j.
\end{aligned} \tag{5.5}$$

The simple example of a free particle is instructive. Setting $\omega^i = \omega_i = 0$ and $H^2 = p^2 + m^2 \neq 0$, Eqs. (5.5) reduce to

$$\omega_0^0 = \frac{m^2}{H}\,\mathbf{d}t = m\,\mathbf{d}\tau, \quad \omega_0^a = \frac{\eta^{ab}p_b}{m^2}\,\omega_0^0 = u^a \mathbf{d}\tau = \mathbf{d}x^a,$$

$$\omega_a^0 = \tfrac{1}{2}p_a \omega_0^0 = \tfrac{1}{2}m p_a \mathbf{d}\tau, \quad \omega_b^a = 0,$$

where use of ω_0^0 or proper time $\mathbf{d}\tau$ in place of $\mathbf{d}t$ simplifies the expressions. The solder and co-solder forms are proportional to the displacement and momentum, respectively.

Also, we see again that the involution required to specify the classical paths necessarily exists. Since the biconformal base space is spanned by the eight forms $\mathbf{d}x^a$ and $\mathbf{d}p_a$, the only way the involution could fail is if there was an independent $\mathbf{d}t\wedge\mathbf{d}p_4$ term in the torsion or co-torsion of the $p_4=-H$ hypersurface. However, since $\mathbf{d}H$ is given *a priori* in terms of the other seven forms, this cannot happen. We conclude that the Hamiltonian dynamics of a point particle is equivalent to the specification of a hypersurface, $y_4=y_4(y_i,x^a)$, in a flat biconformal space, and the consequent necessary existence of a preferred congruence of curves in the hypersurface.

The principal significance of this result is *the unambiguous identification of the geometric quantities which arise in the gauging of the 15-dimensional conformal group with corresponding physical quantities in phase space. In particular, the extra four coordinates are identified with momenta and the integral of the Weyl vector is identified with the action*. This insight should not be construed as a replacement for the phase space description of particle mechanics. Rather, the interpretation presented here is to be regarded as the one-particle limit of a full biconformal field theory, and is intended to provide guidance for and a check on that field theory. We shall see in Sec. VII that for curved biconformal spaces, the ''momentumlike'' basis forms ω_a^0 contain the stress-energy source for the Einstein equation, further strengthening the present interpretation.

Notice that for multiparticle systems this interpretation of the variables of flat biconformal space in terms of phase space variables differs from the usual n-particle phase space of more complicated systems. In the case of multiple particles in a small region of (nearly) flat biconformal space, the particles share *both* the momentum and configuration space. Thus, while a single point of a multiparticle phase space characterizes the entire multiparticle system, the many particle biconformal model will be described by many points in the same eight-dimensional space. Nonetheless, the physical interpretation of the extra four coordinates as momenta remains valid, and each particle treated separately will (locally) obey its own set of Hamilton's equations.

The same conclusion holds for the fields in biconformal spaces—the extra dimensions will give the local field momentum. This may be seen in either of two ways. First, fields are the continuum limits of multiparticle systems and their local momentum will therefore be the limit of the particle momenta. Second, whenever a single field quantum is confined to an isolated region which is small relative to the curvature, it will have an interpretation as a single particle. The biconformal co-space must then give the momentum of that particle.

While accomplishing an interpretation of the biconformal variables, we have also circumvented problems with previous treatments of conformal gauge theory. Unlike four-dimensional conformal gaugings which always reduce to a four-dimensional Weyl geometry in which the inverse translations are auxiliary, the current approach retains the full conformal degrees of freedom. The extra four degrees of freedom are now seen to correspond to the inclusion of momentum variables in the physical description.

VI. WEYL'S THEORY IN BICONFORMAL SPACE

Now consider how the dynamics of the Hamiltonian correspondence of Secs. IV and V is modified by the presence of the vector field, α. The Weyl vector is now given by

$$\omega_0^0=-p_a\mathbf{d}x^a+\alpha_a\mathbf{d}x^a\equiv-\pi_a\mathbf{d}x^a \tag{6.1}$$

so that

$$\mathbf{d}\omega_0^0=\mathbf{d}x^a\wedge\mathbf{d}\pi_a\,. \tag{6.2}$$

Specifying $\pi_4=\pi_4(\pi_i,x^a)$ and setting $\omega^i=\omega_i=0$ again leads to Hamilton's equations, in the form

$$\mathbf{d}x^i = -\frac{\partial \pi_4}{\partial \pi_i}\,\mathbf{d}t, \tag{6.3a}$$

$$\mathbf{d}p_i = \frac{\partial \pi_4}{\partial x^i}\,\mathbf{d}t. \tag{6.3b}$$

Maintaining our previous identification $p_4 = -H = -(m^2+p^2)^{1/2}$ we find

$$\pi_4 = -(m^2+p^2)^{1/2} - \alpha_4 = -(m^2+(\pi^i+\alpha^i)(\pi_i+\alpha_i))^{1/2} - \alpha_4. \tag{6.4}$$

Hamilton's equations become

$$\mathbf{d}x^i = \frac{\pi^i+\alpha^i}{(m^2+(\pi+\alpha)^2)^{1/2}}\,\mathbf{d}t = \frac{p^i}{(m^2+p^2)^{1/2}}\,\mathbf{d}t$$

or

$$\dot{x}^i = \frac{p^i}{(m^2+p^2)^{1/2}} \tag{6.5}$$

for the position variables and

$$\mathbf{d}\pi_i = -\frac{\partial \alpha_4}{\partial x^i}\,\mathbf{d}t - \frac{\pi^j+\alpha^j}{(m^2+(\pi+\alpha)^2)^{1/2}}\,\alpha_{j,i}\mathbf{d}t = -\frac{\partial \alpha_4}{\partial x^i}\,\mathbf{d}t - \dot{x}^j\alpha_{j,i}\mathbf{d}t \tag{6.6}$$

for the momentum. The left-hand side of Eq. (6.6) expands to

$$\mathbf{d}\pi_i = \mathbf{d}p_i - \frac{\partial \alpha_i}{\partial t}\,\mathbf{d}t - \alpha_{i,j}\mathbf{d}x^j,$$

which finally leads to

$$\frac{dp_i}{dt} = \left(-\frac{\partial \alpha_4}{\partial x^i} + \frac{\partial \alpha_i}{\partial t}\right) + \dot{x}^j(\alpha_{i,j} - \alpha_{j,t}). \tag{6.7a}$$

For the time component we have

$$\begin{aligned}\frac{dp_4}{dt} &= \frac{d}{dt}(\pi_4+\alpha_4) = \frac{d}{dt}\left(-(m^2+p^2)^{1/2}\right)\\ &= -\frac{1}{(m^2+p^2)^{1/2}}\,p^i\,\frac{dp_i}{dt} = -\dot{x}^i\,\frac{dp_i}{dt} = \dot{x}^i\left(\frac{\partial \alpha_4}{\partial x^i} - \frac{\partial \alpha_i}{\partial t}\right).\end{aligned} \tag{6.7b}$$

Equations (6.7) give the Lorentz force law if we identify

$$\alpha_a = q(\phi, -A_i) = -qA_a.$$

Thus, the existence and form of the electromagnetic force on a charged particle is correctly predicted by the general solution for flat biconformal space. The presence of the vector field α_a, its gauge dependence, and its proper coupling to matter are direct consequences of the local gauge theory of scalings and of our identification of Hamiltonian dynamics as the specification of a

hypersurface in flat biconformal space. Nor does this formulation suffer the objection made to the original Weyl theory. Since we are studying precisely the flat biconformal spaces, there is no dilational curvature and no measurable size change.

In concluding this section, we comment briefly on the relationship of these considerations to the standard U(1) model of electromagnetic field theory. Biconformal space has a natural metric structure based on the Killing metric of the underlying conformal group. The Killing metric has eigenvalues ± 1 and zero signature, hence the local form

$$K_{\mathscr{A}\mathscr{B}} = \mathrm{diag}(\eta_{ab}, -\eta^{ab}), \tag{6.8}$$

where $(\mathscr{A},\mathscr{B}) = (1,...,8)$. Consequently, the proportionality constant in the identification between y_a and p_a is purely imaginary.[48] Since the vector field $\alpha_a(x)$ is real, we have

$$y_a = \lambda(ip_a + \alpha_a) \tag{6.9a}$$

$$= i\lambda(p_a - i\alpha_a) \tag{6.9b}$$

as the relationship between geometric and physical variables. Since the proportionality constant drops out of Hamilton's equations it is not measurable classically, but it must be included whenever particle paths are allowed to deviate from the classical trajectories. Thus, the real-valued four-dimensional scale invariance that preserves the form of the standard flat biconformal connection becomes a U(1) invariance when applied to the physical variables of Eq. (6.9b).

VII. COUPLING TO GRAVITY AND OTHER FIELDS

So far, we have motivated a new eight-dimensional gauging of the conformal group, found the class of flat solutions of the resulting biconformal spaces, and shown how to interpret the flat solutions as a phase space for a single particle coupled to a background electromagnetic field. In this section, we generalize these results to gravitating biconformal spaces in which the solder form satisfies the Einstein equation with arbitrary matter as source, and the vector potential identified in the previous sections satisfies the Maxwell field equations with arbitrary electromagnetic currents. The results of this section therefore comprise a unified geometric theory of gravitation and electromagnetism.

Our starting point is the full set of structure equations, Eqs. (2.12), which define the 15 curvature two-forms Ω^A_B. Each of these curvatures has the three biconformally invariant terms displayed in Eq. (2.13). In addition, we note that the two-form $\mathbf{d}\omega^0_0$ is separately biconformally invariant, and the solder and co-solder forms ω^a_0 and ω^0_a transform tensorially. As long as the corresponding Bianchi identities are satisfied, a specification of any combination of these fields may be used to invariantly determine subclasses of biconformal geometries.

To begin, we seek some general constraints to limit the number of independent fields. For classical geometries it is reasonable to assume that no classical path in phase space encloses a plaquette on which the dilation, Ω^0_0, is nonvanishing. The simplest (but by no means the only) way to guarantee this is to just set $\Omega^0_0 = 0$. We also expect that the space–time torsion will vanish in typical classical models, and again make the simplest hypothesis, that the full torsion (but *not* the co-torsion) is zero. Thus we have the general constraints

$$\Omega^0_0 = 0, \tag{7.1}$$

$$\Omega^a_0 = 0. \tag{7.2}$$

Next, we note that the vanishing of the torsion puts the solder form in involution. Assuming the resulting foliation to be regular, there exists a four-dimensional submanifold of the base space spanned by $\omega_0^a = \mathbf{e}^a$. As a final general constraint, we require the existence of a completion $\mathbf{f}_a$ to the $\mathbf{e}^a$ basis in which the space–time curvature is traceless:

$$\Omega^a_{bac} = 0 \quad [\text{in the } (\mathbf{e}^a, \mathbf{f}_a) \text{ basis}]. \tag{7.3}$$

We shall show that these three constraints are sufficient for the resulting class of geometries to take a recognizable form. There is *no* restriction of the dependence of any of the fields on the eight coordinates, (x^μ, y_ν). Notice that we now return to the use of Greek indices for general coordinates.

In addition to the general constraints above, we posit two field equations. The first provides a source for the Weyl vector, which we take here to be in the typical form for an electromagnetic current

$$*\mathbf{d}*\mathbf{d}\omega_0^0 = \mathbf{J} = J_a(x)\mathbf{e}^a. \tag{7.4a}$$

Finally, part of the co-solder form is determined by an arbitrary stress-energy tensor T_{ab} via

$$\omega_a^0 = \mathscr{T}_a + \cdots, \tag{7.4b}$$

where $\mathscr{T}_a \equiv -\frac{1}{2}(T_{ab} - \frac{1}{3}\eta_{ab}T)\mathbf{e}^b$. The consistency of these expressions and the form of the remaining part of the co-solder form are established using the general equations, Eqs. (7.1)–(7.3), together with the structure equations, Eqs. (2.12).

The central result of this section is that a biconformal space is in 1–1 correspondence with a four-dimensional Einstein–Maxwell space–time if and only if Eqs. (7.1)–(7.4) hold. The 1–1 correspondence is based on an isomorphism between the biconformal co-space at x_0^μ and the tangent space to a four-dimensional space–time at a corresponding point x_0^μ. To see roughly how this isomorphism occurs, note that the biconformal base space is spanned by the differentials $(\mathbf{d}x^\mu, \mathbf{d}y_\nu)$. Setting $x^\mu = x_0^\mu$ gives $\mathbf{d}x_0^\mu = 0$, thereby restricting to the subspace spanned by $\mathbf{d}y_\nu$. We will show that the conditions above make the biconformal space flat in the y variables, giving a four-dimensional Minkowski space with basis $\mathbf{d}y_\nu$. Then, because the basis $\mathbf{d}y_\nu$ transforms oppositely under Lorentz transformations from $\mathbf{d}x^\mu$, and because the $\mathbf{d}x^\mu$ spans the cotangent space of an orthogonal four-dimensional submanifold, there exists an isomorphism between the $\mathbf{d}y_\nu$ space and the *tangent* space of the submanifold. A more complete presentation of the techniques used and some related results are given in Ref. 49.

The equations to be solved are

$$\mathbf{d}\omega_b^a = \omega_b^c \wedge \omega_c^a + \omega_b^a \wedge \omega_0^a - \eta_{bc}\eta^{ad}\omega_d^0 \wedge \omega_0^c + \Omega_b^a, \tag{2.12a$'$}$$

$$\mathbf{d}\omega_0^a = \omega_0^0 \wedge \omega_0^a + \omega_0^b \wedge \omega_b^a, \tag{2.12b$'$}$$

$$\mathbf{d}\omega_a^0 = \omega_a^0 \wedge \omega_0^0 + \omega_a^b \wedge \omega_b^0 + \Omega_a^0, \tag{2.12c$'$}$$

$$\mathbf{d}\omega_0^0 = \omega_0^a \wedge \omega_a^0. \tag{2.12d$'$}$$

The first part of our proof follows Theorem I of Ref. 49. We begin with the Bianchi identity for Eq. (2.12b′), which follows by taking the exterior derivative and using $\mathbf{d}^2 = 0$. The result is

$$\omega_0^b \wedge \Omega_b^a = 0, \tag{7.5}$$

which in particular shows that the momentum term of the curvature vanishes

$$\Omega_b^{acd}=0. \tag{7.6}$$

Similarly, the Bianchi identity for Eq. (2.12d′) requires

$$\Omega_b^{0cd}=0 \tag{7.7}$$

for the co-torsion. Together with the vanishing of the dilation and the torsion, Eqs. (7.6) and (7.7) imply that the bundle is momentum flat, i.e., the momentum term of each curvature vanishes.

Next, we use the involution of the solder form, as noted above. The involution means that the biconformal bundle is foliated by 11-dimensional subbundles on which $\omega_0^a=0$. On this subbundle, using the momentum-flatness, the structure equations reduce to

$$\mathbf{d}\omega_b^a=\omega_b^c\wedge\omega_c^a, \tag{7.8a}$$

$$\mathbf{d}\omega_a^0=\omega_a^0\wedge\omega_0^0+\omega_a^b\wedge\omega_b^0, \tag{7.8b}$$

$$\mathbf{d}\omega_0^0=0. \tag{7.8c}$$

Equations (7.8) are just the structure equations for a flat four-dimensional Weyl geometry. Equation (7.8a) shows that we can perform a Lorentz gauge transformation on the entire bundle such that $\omega_b^a|_{\omega_0^a=0}=0$, while Eq. (7.8c) shows the existence of a scaling such that $\omega_0^0|_{\omega_0^a=0}=0$. Following these gauge changes Eq. (7.8b) takes the simple form

$$\mathbf{d}\omega_a^0=0$$

with the immediate solution

$$\omega_a^0|_{\omega_0^a=0}=f_a{}^\mu(x)\mathbf{d}y_\mu \tag{7.9}$$

where we have introduced four coordinates y_μ to span the four-dimensional base space of the submanifold. Notice that the coefficient matrix $f_a{}^\mu(x)$ must be nondegenerate.

Now, assuming the foliation to be regular, the space of leaves of the foliation is also a four-dimensional manifold spanned by ω_0^a. There exist coordinates x^μ such that

$$\omega_0^a=e_\mu{}^a(x,y)\mathbf{d}x^\mu=\mathbf{e}^a \tag{7.10}$$

so that each leaf of the foliation is given by some constant value $x^\mu=x_0^\mu$.

From the subbundle, with connection $\omega_b^a|_{x_0}=0$, $\omega_0^0|_{x_0}=0$, and $\omega_a^0|_{x_0}=\mathbf{f}_a$, we extend back to the full bundle by allowing x to vary. This can change each connection form at most by a term proportional to $\mathbf{d}x^\mu$ or equivalently $\mathbf{e}^a$, so the full connection may always be given the local form

$$\omega_b^a=C_{bc}^a\mathbf{e}^c, \tag{7.11a}$$

$$\omega_0^a=\mathbf{e}^a(x,y), \tag{7.11b}$$

$$\omega_a^0=f_a{}^\mu(x)\mathbf{d}y_\mu+B_{ab}e^b\equiv\mathbf{f}_a+\mathbf{B}_a, \tag{7.11c}$$

$$\omega_0^0=W_a\mathbf{e}^a, \tag{7.11d}$$

where W_a, B_{ab}, and C_{bc}^a are functions of all eight coordinates (x^μ,y_ν). The functional dependence of $e_\mu{}^a$, while *a priori* unspecified, is actually restricted by Eq. (2.12b′). Since ω_b^a and ω_0^0

depend only on $\mathbf{e}^a$ and not $\mathbf{f}_a$, writing out $\mathbf{de}^a$ in coordinates immediately shows that $\partial^\nu e_\mu{}^a = 0$, where we write $\partial^\nu \equiv \partial/\partial y_\nu$ and $\partial_\nu \equiv \partial/\partial x^\nu$. Therefore, we may drop the y dependence in Eq. (7.11b).

If we had made a different initial choice of the coordinates y_μ, the connection would still be of the form given by Eqs. (7.11), although W_a, B_{ab}, and C^a_{bc} would be different functions. We can therefore assume without loss of generality that the basis $(\mathbf{e}^a, \mathbf{f}_b)$ has been chosen as the one in which the space–time curvature is tracefree.

Next, after dividing B_{ab} into symmetric and antisymmetric parts,

$$B_{ab} \equiv S_{ab}(x) + F_{ab}(x) \tag{7.12}$$

with $S_{ab} = S_{(ab)}$ and $F_{ab} = F_{[ab]}$, we substitute Eq. (7.11c) into the dilation equation, Eq. (2.12d′), to find

$$\mathbf{d}\omega_0^0 = \mathbf{e}^a \wedge \mathbf{f}_a + \mathbf{F} = \mathbf{d}x^\mu \wedge f_\mu{}^\nu(x)\mathbf{d}y_\nu + \mathbf{F}, \tag{7.13}$$

where $\mathbf{F} \equiv F_{ab}\mathbf{e}^a \wedge \mathbf{e}^b = F_{\mu\nu}\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu$ and $f_\mu{}^\nu \equiv e_\mu{}^a f_a{}^\nu$. The mixed terms of Eq. (7.13) yield

$$W_{\mu'}{}^\nu = -f_\mu{}^\nu(x), \tag{7.14}$$

which is immediately integrated to give

$$\omega_0^0 \equiv W_\mu \mathbf{d}x^\mu = (-f_\mu{}^\nu(x)y_\nu + A_\mu(x))\mathbf{d}x^\mu \tag{7.15}$$

up to a gauge transformation. Here $A_\mu(x)$ is an arbitrary integration constant for the y integration.

We could continue to carry the field $f_\mu{}^\nu(x)$ through the remainder of the general constraints, but it is simpler to employ the field equation to remove it now. In a more general class of models, $f_\mu{}^\nu(x)$ provides an additional unspecified set of fields. While it is easy to conjecture that these extra fields may lead to a geometric electroweak theory, $f_\mu{}^\nu(x)$ appears on the surface to be a translational rather than a rotational Yang–Mills field, so that the required quadratic terms are missing from the field strength. The proper role of these fields is under study. For our present purpose, we simply note that the field equation

$$*\mathbf{d}*\mathbf{d}\omega_0^0 = \mathbf{J} = J_a(x)e^a \tag{7.3}$$

has no source for the part of ω_0^0 linear in y, and therefore gives

$$[*\mathbf{d}*\mathbf{d}f_\mu{}^\nu(x)]y_\nu = 0 \tag{7.16}$$

for the y-dependent part of the space–time term. This is a four-dimensional electromagnetic-type equation for the ''potential'' $f_\mu \equiv f_\mu{}^\nu(x)y_\nu$, which has the unique solution $f_\mu{}^\nu(x) = 0$ for vanishing boundary conditions at infinity. To this solution we add the particular solution $f_\nu{}^\nu = \text{const}$, since $f_\mu{}^\nu(x)$ must be invertible as noted above. Finally, a constant change of the y coordinates gives $f_\mu{}^\nu = \delta_\mu^\nu$, simplifying the co-solder form to

$$\omega_a^0 = e_a{}^\mu(x)\mathbf{d}y_\mu + B_{a\mu}(x,y)\mathbf{d}x^\mu, \tag{7.17}$$

where $e_a{}^\mu$ is the inverse to $e_\mu{}^a$.

With the solder form given by Eq. (7.17), Eq. (7.15) becomes

$$\omega_0^0 \equiv W_\mu \mathbf{d}x^\mu = (-y_\mu + A_\mu(x))\mathbf{d}x^\mu. \tag{7.18}$$

Thus, this entire class of geometries has the same ''minimal coupling'' form of the Weyl vector as we found for the flat case. Notice also that the form of Eq. (7.18) is not affected by a purely x-dependent gauge transformation, which is consistent with the interpretation of $\mathbf{A}$ as the electromagnetic vector potential.

Next, we move to Eq. (2.12b′), which becomes

$$\mathbf{d}\mathbf{e}^a = \omega_0^0 \wedge \mathbf{e}^a + \mathbf{e}^b \wedge \omega_b^a . \tag{7.19}$$

This may be uniquely solved for the connection ω_b^a. Let

$$\omega_b^a = \alpha_b^a + \beta_b^a , \tag{7.20}$$

where α_b^a is the usual metric compatible spin connection satisfying

$$\mathbf{d}\mathbf{e}^a = \mathbf{e}^b \wedge \alpha_b^a \tag{7.21}$$

and require

$$\mathbf{e}^b \wedge \beta_b^a + \omega_0^0 \wedge \mathbf{e}^a = 0. \tag{7.22}$$

Equation (7.22) is solved uniquely by

$$\beta_b^a = -W_b \mathbf{e}^a + \eta_{bc}\eta^{ad} W_d \mathbf{e}^c , \tag{7.23}$$

so we have now satisfied two of the four structure equations.

Next, we impose the tracelessness condition for the space–time term of the curvature in the $(\mathbf{e}^a, \mathbf{f}_b)$ basis:

$$\Omega^a{}_{bac} = 0. \tag{7.24}$$

Rearranging Eq. (2.12a) to solve for Ω_b^a we find

$$\Omega_b^a = \mathbf{d}\omega_b^a - \omega_b^c \wedge \omega_c^a - \omega_b^0 \wedge \omega_0^a + \eta_{bc}\eta^{ad}\omega_d^0 \wedge \omega_0^c \tag{2.12a}$$

$$= (\mathbf{d}\alpha_b^a - \alpha_b^c \wedge \alpha_c^a) + (\mathbf{d}\beta_b^a - \beta_b^c \wedge \alpha_c^a - \alpha_b^c \wedge \beta_c^a) - \beta_b^c \wedge \beta_c^a - \omega_b^0 \wedge \mathbf{e}^a + \eta_{bc}\eta^{ad}\omega_d^0 \wedge \mathbf{e}^c$$

$$\equiv \mathbf{R}_b^a(\alpha) + \mathbf{D}\beta_b^a - \beta_b^c \wedge \beta_c^a - \omega_b^0 \wedge \mathbf{e}^a + \eta_{bc}\eta^{ad}\omega_d^0 \wedge \mathbf{e}^c , \tag{7.25}$$

where we have written $\mathbf{D}$ for the covariant exterior derivative using the $\mathbf{e}^a$-compatible connection α_b^a, and $\mathbf{R}_b^a \equiv \mathbf{d}\alpha_b^a - \alpha_b^c \wedge \alpha_c^a$ is the usual Riemann curvature two-form. Substituting Eq. (7.17) for ω_a^0 and (7.23) for β_b^a, some algebra leads to

$$\Omega_b^a = \mathbf{R}_b^a + \eta_{bc}[\mathbf{D}W^c \mathbf{e}^a - \mathbf{D}W^a \mathbf{e}^c + W^2 \mathbf{e}^c \mathbf{e}^a] + \mathbf{f}_b \mathbf{e}^a - \eta_{bc}\eta^{ad}\mathbf{f}_d \mathbf{e}^c + \mathbf{B}_b \mathbf{e}^a - \eta_{bc}\eta^{ad}\mathbf{B}_d \mathbf{e}^c . \tag{7.26}$$

Finally, expand

$$\mathbf{D}W^a = -\eta^{ab}\mathbf{D}y_b + \mathbf{D}A^a . \tag{7.27}$$

While $\mathbf{D}A^a$ is independent of y_μ, we must further expand

$$\mathbf{D}y_b \equiv \mathbf{d}(e_b{}^\mu y_\mu) - \alpha_b^a e_a{}^\mu y_\mu = e_b{}^\mu \mathbf{d}y_\mu + (\mathbf{d}e_b{}^\mu - \alpha_b^a e_a{}^\mu) y_\mu . \tag{7.28}$$

The first term in Eq. (7.28) is simply $\mathbf{f}_a$. For the final term we use the Christoffel connection $\Gamma^\mu{}_{\alpha\beta}$ and the covariant constancy of $e_b{}^\mu$,

$$\mathbf{d}x^\beta D_\beta e_b{}^\mu = \mathbf{d}e_b{}^\mu - \alpha_b^a e_a{}^\mu + e_b{}^\alpha \Gamma^\mu{}_{\alpha\beta}\mathbf{d}x^\beta = 0 \tag{7.29}$$

to write

$$\mathbf{D}y_b = \mathbf{f}_b - e_b{}^\alpha \Gamma^\mu{}_{\alpha\beta}\mathbf{d}x^\beta y_\mu \equiv \mathbf{f}_b - \Gamma_{bc}\mathbf{e}^c \equiv \mathbf{f}_b - \Gamma_b\,. \tag{7.30}$$

Now substitution into Eq. (7.26) shows that the cross and momentum terms of the curvature vanish while the components of the space–time term become

$$\begin{aligned}\Omega^a_{bcd} = \mathbf{R}^a_{bcd} &- (D_c A_b + \beta_{bc} + B_{bc})\,\delta^a_d + (D_d A_b + \Gamma_{bd} + B_{bd})\,\delta^a_c \\ &+ \eta_{bf}\eta^{ag}(D_c A_g + \Gamma_{gc} + B_{gc})\,\delta^f_d - \eta_{bf}\eta^{ag}(D_d A_g + \Gamma_{gd} + B_{gd})\,\delta^f_c \\ &- W_c(W_b\delta^a_d - \eta_{bf}\eta^{ag}W_g\delta^f_d) + W_d(W_b\delta^a_c - \eta_{bf}\eta^{ag}W_g\delta^f_c) + W^2(\eta_{bc}\delta^a_d - \eta_{bd}\delta^a_c).\end{aligned} \tag{7.31}$$

Next, by setting the trace of Ω^a_{bcd} to zero, we can insure that the Einstein equation holds with arbitrary stress-energy tensor. Contracting, we find that the antisymmetric part $\Omega^c_{acb} - \Omega^c_{bca}$ is identically zero while

$$\Omega^c_{acb} = R_{ab} + 2(A_{(a;b)} + \Gamma_{ab} + S_{ab} + W_a W_b) + \eta_{ab}(A_c{}^{;c} + \Gamma_c{}^c + S_c{}^c - 2W^2) = 0 \tag{7.32}$$

relates $B_{(ab)} = S_{ab}$ to the Ricci tensor and the vector potential. If we require

$$S_{ab} \equiv \mathscr{T}_{ab} - (A_{(a;b))} + \Gamma_{ab} + W_a W_b - \tfrac{1}{2}\eta_{ab}W^2), \tag{7.33}$$

where $\mathscr{T}_{ab} \equiv -\tfrac{1}{2}(T_{ab} - \tfrac{1}{3}\eta_{ab}T)$ and T_{ab} is the electromagnetic stress-energy tensor constructed from A_b plus the stress-energy tensor from whatever other phenomenological fields one wishes to add, then Eq. (7.32) reduces to the Einstein equation with source T_{ab}. We shall see below that the choice given by Eq. (7.33) for S_{ab} greatly simplifies the expressions for both the curvature and the co-torsion.

Nothing further is required in order to satisfy the final structure equation, Eq. (2.12c). Instead, we *define* the co-torsion by Eq. (2.12c), giving

$$\Omega^0_a = \mathbf{d}\omega^0_a - \omega^b_a \wedge \omega^0_b - \omega^0_a \wedge \omega^0_0. \tag{2.12c}$$

This is easiest to evaluate if we write ω^0_a in terms of W_a wherever possible, resulting in

$$\omega^0_a = \mathscr{T}_a - \mathbf{D}W_a - W_a\mathbf{W} + \tfrac{1}{2}W^2\eta_{ab}\mathbf{e}^b. \tag{7.34}$$

Using the tracefree condition (i.e., the Einstein equation) to replace $\mathscr{T}_a$ by $\mathscr{R}_a \equiv -\tfrac{1}{2}(R_{ab} - \tfrac{1}{6}\eta_{ab}R)\mathbf{e}^b$ we find that Eq. (2.12c) takes the form

$$\Omega^0_a = \mathbf{D}\omega^0_a - \beta^b_a \wedge \omega^0_b - \omega^0_a \wedge \mathbf{W} = \mathbf{D}\omega^0_a + W_a\mathbf{d}\mathbf{W} - W^b\eta_{ac}\mathbf{e}^c\omega^0_b - \omega^0_a \wedge \mathbf{W}. \tag{7.35}$$

Substituting Eq. (7.34) for ω^0_a and using the Ricci identity $\mathbf{D}^2\omega_a = -\mathbf{R}^b_a\omega_b$ (for an arbitrary one-form ω_a) gives, after several cancellations, the surprisingly simple result

$$\Omega^0_a = \mathbf{D}\mathscr{R}_a + [\mathbf{R}^b_a - (\delta^c_a\delta^b_d - \eta^{bc}\eta_{ad})\mathscr{R}_c\mathbf{e}^d]W_b = \mathbf{D}\mathscr{R}_a + \mathbf{C}^b_a W_b\,, \tag{7.36}$$

where $\mathbf{C}^b_a$ is the Weyl curvature two-form. Notice that all derivatives of the Weyl vector have cancelled, so the resulting co-torsion also has vanishing momentum and cross terms. If we use the contracted second Bianchi identity to write $\mathbf{D}\mathscr{R}_a$ as a divergence of the Weyl curvature, we find

$$\Omega^0_a = -D_b \mathbf{C}^b_a + \mathbf{C}^b_a W_b\,, \tag{7.37}$$

so the co-torsion is simply the Weyl-covariant (not α^a_b-covariant) divergence of the Weyl curvature tensor, hence a direct measure of the deviation of the underlying space–time from conformally flat.[42]

Finally, remarkable cancellations also occur if we use Eq. (7.33) to replace $\mathbf{B}_a$ in the full curvature tensor given by Eq. (7.26). The result is simply

$$\Omega^a_b = \mathbf{C}^b_a\,. \tag{7.38}$$

The biconformal geometry is now fully specified except for the vector potential A_μ. This is fixed by the field equation

$$*\mathbf{d}*\mathbf{d}\omega^0_0 = *\mathbf{d}*\mathbf{dA} = \mathbf{J}, \tag{7.4a}$$

where $\mathbf{J}$ is the electromagnetic current.

To complete the proof of the sufficiency of Eqs. (7.1)–(7.4) for the biconformal space to correspond to an Einstein–Maxwell space–time we only need to show that the biconformal geometry is effectively four-dimensional. This happens because the extra dimensions of the biconformal base space can be identified with the tangent space of the Riemannian space–time. Such an identification works for two reasons. First, the co-vector basis $\mathbf{d}y_\mu$ has the same scale and Lorentz transformation properties as the basis vectors $\partial/\partial x^\mu$ for the tangent space, so we can identify the bases. Second, since both the co-space and the tangent space may be taken as Minkowski vector spaces (i.e., both are flat), the two complete vector spaces may be identified. Thus, while the coordinates y_μ act as momentumlike components of *cotangent* vectors $y_\mu \mathbf{d}x^\mu$, their differentials $\mathbf{d}y_\mu$ from a *tangent* basis. As discussed further in the final section, this isomorphism guarantees that the extra four dimensions do not lead to undesired new macroscopic effects.

Notice that, while the content of the biconformal space does not exceed that of an Einstein–Maxwell space–time, we cannot claim that a biconformal space satisfying Eqs. (7.1)–(7.4) is homeomophic to a four-dimensional Weyl geometry because the Weyl vector of the biconformal space retains linear y_μ dependence. It is this difference that leads to the vanishing dilation of the biconformal space.

The converse, that a four-dimensional Einstein–Maxwell space–time extends to a unique biconformal space satisfying Eqs. (7.1)–(7.4), is immediate, since, given the stress-energy tensor and electromagnetic current determining the space–time, we can find the solder form and the electromagnetic vector potential and from these directly write down the connection of the associated biconformal space as

$$\omega^0_0 = (-y_\mu + A_\mu)\mathbf{d}x^\mu, \tag{7.39a}$$

$$\omega^a_0 = \mathbf{e}^a(x), \tag{7.39b}$$

$$\omega^0_a = \mathscr{T}_a - \mathbf{D}W_a - W_a\mathbf{W} + \tfrac{1}{2}W^2\eta_{ab}\mathbf{e}^b, \tag{7.39c}$$

$$\omega^a_b = \alpha^a_b - W_b\mathbf{e}^a + \eta_{bc}\eta^{ad}W_d\mathbf{e}^c, \tag{7.39d}$$

where we invert the vector space isomorphism above to map the tangent space into the extra four dimensions of the biconformal space. The form above for the connection leads directly to the biconformal curvatures

$$\Omega^0_0 = 0, \tag{7.40a}$$

$$\Omega_0^a = 0, \tag{7.40b}$$

$$\Omega_a^0 = -D_b \mathbf{C}_a^b + \mathbf{C}_a^b W_b, \tag{7.40c}$$

$$\Omega_b^a = \mathbf{C}_b^a, \tag{7.40d}$$

and the field equations are unchanged. In terms of the natural conformal expressions $\mathscr{R}_a$, $\mathscr{T}_a$, and ω_0^0, the field equations are

$$\mathscr{R}_a = \mathscr{T}_a, \tag{7.41a}$$

$$*\mathbf{d}*\mathbf{d}\omega_0^0 = \mathbf{J}. \tag{7.41b}$$

The expressions for the curvature above make use of the tracelessness condition $\Omega_{bac}^a = 0$ to write the co-solder form in terms of the Ricci tensor instead of the stress-energy tensor. Considering the form of the curvature if we do not impose the tracelessness condition, we can interpret the Einstein equation as being that condition that reduces Ω_b^a and Ω_a^0 to the Weyl curvature and its divergence, respectively.

It should be pointed out that Eqs. (7.40) are quite remarkable for biconformal curvatures in that none of the curvatures has a cross term or momentum term. Despite the eight-dimensional formulation of the theory, all terms containing $\mathbf{d}y_\mu$ have dropped out. The space spanned by the y coordinates is therefore flat, and plays no role in the gravitational effects of this class of models.

At this point, we can only conjecture whether or not these biconformal spaces satisfy some action principle. There are certainly some action densities which are natural candidates. Specifically the curvature-quadratic eight-form $\Omega_B^A \wedge *\Omega_B^A$ is closely related to the action of the four-dimensional theories discussed in the Introduction, and clearly has the flat biconformal spaces of Sec. III as particular solutions. However, even expressing the resulting field equations is complicated, first because there are 12 Weyl-invariant parts to the curvature and second because the Bianchi identities used to reduce the field equations are correspondingly lengthy. We therefore postpone further investigation of field actions to subsequent research.

In the final section, we discuss the observability of the extra dimensions of biconformal spaces.

VIII. DISCUSSION

By placing conformal gauging on an eight-dimensional base space instead of the usual four-dimensional base space, we have overcome the long-standing problem of size change in physical models based on scale invariance. In this section, we demonstrate that our interpretation of biconformal gauge theory is consistent with experience. In particular, we examine the observable consequences of the added dimensions with regard to structure and function.

Three techniques have been used for adding extra dimensions in fundamental models of the world: (a) topological compactification of the extra dimensions, (b) construction of laws of motion or field equations that dynamically reduce the extra dimensions to a sufficiently small scale that they play no macroscopic role,[50] or (c) identification of the extra dimensions with everyday properties already associated with macroscopic matter. Technique (a) is routinely employed in Kaluza–Klein field theories, while technique (b) has been used recently[51,52] in an attempt to associate a fifth dimension with mass. The third technique, (c), was used in the development of special relativity where time came to be seen as a coordinate in a higher- (i.e., four) dimensional space rather than as a parameter of an intrinsically different type on a three-dimensional space. The local interpretation of biconformal space is of type (c). One should therefore *not* think in terms of compactification or other standard Kaluza–Klein ideas. Instead, we regard the extra four dimensions of the biconformal co-space as familiar, routinely observable macroscopic dimensions,

and ask whether a coordinate interpretation of four-momentum is consistent with experience. We demonstrate this consistency by focusing on the two necessary axes of correspondence:

Intrinsic structure: The biconformal co-space must have the same mathematical structure and transformation properties as momentum space.

Dynamical function: The biconformal dynamical laws and description of collisions or interactions must accord with experience.

We address each of these in turn, then conclude by citing positive evidence for a coordinate interpretation of momentum.

A. Intrinsic structure

For either flat or curved space–times, momentum space is the tangent space at each point of space–time. To identify the biconformal co-space with the tangent space we must check that both the Minkowski vector space structure of the tangent space and transformation properties of the tangent basis are reflected in the co-space.

The biconformal co-space is a Minkowski vector space if and only if it is flat. This flatness is obvious for the flat solutions studied in Secs. III–VI, while it is a consequence of the vanishing torsion and dilation in the curved models of Sec. VII. In either of these classes it is therefore possible to set up a vector space isomorphism between the tangent space and the co-space. We note that the same isomorphism also holds between the tangent space of a Riemannian space–time and the co-space coordinates of a *normal biconformal space*.[49] Normal spaces are defined to be torsion-free spaces in which the dilational curvature is closed, the Weyl one-form is exact, and the space–time term of the curvature tensor is trace-free. We therefore have two large, disjoint classes of biconformal geometries in which the biconformal co-space is isomorphic to the Riemannian tangent space.

There are two transformation properties of momentum variables which we must also check. First, under Lorentz transformations, p_μ transforms with the inverse to the *same* transformation as x^μ in flat spaces or $\mathbf{d}x^\mu$ in curved spaces. That is, there must not be two independent Lorentz transformations which can be applied to the space and co-space separately. This is a notable property of the eight-dimensional gauging of the four-dimensional conformal group. Because we began with the space–time conformal group, there is only a single set of local Lorentz transformations. Moreover, y_μ has the proper covariant form. The second transformation property required of momentum is that its scaling weight be -1, since (using Planck's constant) momentum has geometric units of inverse length. This corresponds correctly to the inverse length units of y_μ.

B. Dynamical function

We now consider whether Hamiltonian dynamical laws and our experience of collisions or interactions are consistent with a coordinate interpretation of momentum. There are three essential points concerning a coordinate interpretation which must correspond to our usual experience.

(1) Hamiltonian dynamical equations should describe the classical motion of particles and fields.
(2) While there is no continuity requirement on momenta, coordinates should not change discontinuously.
(3) From our experience we know that collisions only occur between particles which are nearby in their space–time coordinates, and do *not* necessarily occur when their momentum separation becomes small. We must show how this observation is consistent with the single proper biconformal separation, which depends on both position separation and on momentum separation.

For agreement on point 1, we require a match of dynamical properties between the Hamiltonian dynamics of momentum space and the dynamics of biconformal spaces. In Secs. III–VI, we

showed that the usual Hamiltonian dynamical picture is a natural property of flat biconformal spaces. Not only do the co-space coordinates act as momenta, but the structure equation

$$\mathbf{d}\omega_0^0 = \omega_0^a \omega_a^0$$

shows that $\mathbf{d}\omega_0^0$ is a manifestly closed, nondegenerate two-form, hence symplectic, and we have the usual Hamiltonian dynamical structure associated with one-particle phase space. Indeed, we demonstrate that any Hamiltonian system generates a super-Hamiltonian hypersurface in a unique flat biconformal space, and conversely that a hypersurface in a flat biconformal space gives a unique Hamiltonian system.

In the dilation-free curved geometries of Sec. VII and in normal biconformal spaces we can expect the Hamiltonian dynamical laws of flat biconformal space to hold *locally*. As long as we look at a sufficiently small neighborhood to permit a single-particle picture, the considerations of Secs. III–VI hold without modification. That fields on space–time also pose no problem may be seen from three different perspectives:

(a) The observable properties of fields are characterized by tensors built on the tangent space, so the isomorphism between the tangent space and the co-space insures that the same properties can be measured in both models.
(b) Fields and their properties are derivable as the continuum limit of many individual particles, so as long as interactions between particles are correctly predicted, the field limit may be expected to hold.
(c) Fields must admit classical, single particle limits. Thus, the highly localized field of an isolated electron moves according to a classical single particle Hamiltonian.

The net effect of these considerations is that there is no particle or field property that one might observe in a flat, dilation-free, or normal biconformal space that could not also be described in the tangent bundle of a four-dimensional Riemannian geometry. Moreover, biconformal spaces automatically predict the presence and form of electromagnetic fields and the Lorentz force law, properties which in a Riemannian space–time must be added by hand.

Now consider point 2. The nearly instantaneous change of momentum which occurs when, say, a ball bounces off a wall, seems to violate the notion of continuous motion in the momentum dimensions. However, the continuity of the actual dynamics underlying the bounce can be seen in either of two ways. Most simply, we can choose the super-Hamiltonian with a phenomenological potential representing the wall. A realistic potential will not have an actual discontinuity, so the resulting motion predicted by Hamilton's equations [or by the biconformal involution of Eqs. (4.5b) and (4.5c)] will be continuous.

A more fundamental way of seeing the continuity of momentum during the collision is to use the electrodynamic law predicted by local biconformal theory to examine the actual motion of each constituent particle of the ball in the appropriate background field. In this view, the apparent discontinuity arises because the region of biconformal space chosen for study is too large for the flat approximation to be valid. Shrinking the region to one where a single-particle interpretation is expected to hold solves the problem.

Moving to point 3, we know that in order to collide or interact strongly, the proper separation of two particles must be small. For this fact to be consistent with our experience that collisions occur whenever the space–time separation becomes small regardless of the momentum separation, is nontrivial. Indeed, recent models[51,52] in which a fifth dimension proportional to mass is added to space–time fail this test—in those models, two particles of vastly different mass will not generally collide.

In the biconformal models presented here, we find that the proper distance between colliding objects behaves according to our experience. The proper interval is given by the conformal Killing

metric of Eq. (6.8). While Eq. (6.8) applies to the orthonormal (ω_0^a, ω_a^0) basis, we easily change to the $(\mathbf{e}^a, \mathbf{f}_a)$ basis, where we find that the metric takes the form

$$g_{ab} = \begin{pmatrix} -2\mathscr{W}_{(ab)} & \delta_m{}^n \\ \delta_n{}^m & 0 \end{pmatrix},$$

where $\mathscr{W}_{ab} = \mathscr{T}_{ab} + \alpha_{a,b} + W_a W_b - \frac{1}{2} W^2 \eta_{ab}$. Therefore the squared interval ds^2 between the ball and the wall in terms of their phase space separations $(\Delta x^m, \Delta p_n)$ is

$$ds^2 = -2\mathscr{W}_{mn} \Delta x^m \Delta x^n + 2\Delta x^m \Delta p_m$$

which vanishes when the spacetime separation Δx^m vanishes and does not necessarily vanish when the momentum separation $\Delta p_m = 0$. This is exactly what is required.

In parts A and B we have shown that there is no mystery to the extra four dimensions of the biconformal co-space. Indeed, the coordinates for these extra directions are always immediately available as the energy and momentum of the system under study. Thus, if we want to probe the full biconformal geometry by ''walking off'' in the extra directions, we already know exactly how to do it. We must vary our energy or the direction of our motion. Any change of the tangent to our world line is a change in location in the momentum sector of biconformal space. Of course, because we are constrained to obey the classical laws of physics, we only experience four degrees of freedom, not eight (except as initial dynamical conditions, when we need all eight). In particular, conservation of energy and momentum constrain motions to $\mathbf{d}y_\mu = 0$ subspaces.

The new picture would be particularly convincing if there were some direct positive evidence that (perhaps under extreme conditions) the full eight-dimensional character of the world is manifest. For such evidence we must study biconformal spaces in which the fields take on a more general y dependence. Therefore, the following observations are necessarily conjectural, the investigation of the microscopic meaning of biconformal space lying clearly beyond the scope of the present paper. Nonetheless, as long as the deviation from the normal or dilation-free classes are small, we may expect the new evidence to show up as a dependence of some physical fields or parameters on momentum as well as position. We briefly discuss two physical variables which have this property: quantum mechanical wave functions, and running coupling constants.

The most evident dependence of fundamental physics on momentum is the phase space duality found in quantum mechanics. It is well understood that the wave function of a particle may be equally well represented using either momentum or position coordinates. In a very direct sense, the momentum of the particle is used as a coordinate. Representations such as the number basis for the harmonic oscillator, which lies midway between momentum and position variables, are also frequently used.

The fact that we do not independently probe both position and momentum simultaneously suggests that even at the quantum level (at least semi-classically) matter is substantially restricted to the neighborhood of some phase-space hypersurface. This might be understood in terms of biconformal space as near-normal biconformal behavior, with microscopic deviations from an involute or nearly involute subspace. The picture is consistent with the view of the path integral approach, which says that a quantum system essentially probes all phase space paths. In the biconformal picture, a path integral would simply be taken over all eight coordinates.

A second piece of evidence that momentum components act as independent coordinates is the existence of running coupling constants in quantum field theory. It is found experimentally that, at high energies, the strength of the electromagnetic and weak couplings vary with energy. Such energy-momentum dependence is easily understood if energy and momentum are coordinates, but is otherwise a somewhat nontransparent result of detailed field theoretic calculation.

Considerable progress has been made in establishing the connection between quantum systems and scale-invariant systems.[53,54] It remains to be seen to what degree specific biconformal models can be developed which will make these final observations precise.

[1] H. Weyl, Sitzung. d. Preuss. Akad. d. Wissensch., 465 (1918); *The Principle of Relativity* (Dover, New York, 1923), Chap. XI.
[2] H. Weyl, *Space-Time-Matter* (Dover, New York, 1952), 3rd Ed., Chaps. II and IV, 34+35, p. 242, et seq. Originally: H. Weyl, *Raum Zeit Materie* (J. Springer, Berlin, 1918).
[3] H. Weyl, Nature (London) **106**, 781 (1921); Math. Z. **2**, 384 (1918); Ann. Physik **54**, 117 (1918); Ann. Physik **59**, 101 (1919); Phys. Z. **21**, 649 (1920); Ann. Physik **65**, 541 (1921); Phys. Z. **22**, 473 (1921); Z. Phys. **56**, 330 (1929).
[4] W. Pauli, *Theory of Relativity*, translated by G. Field (Dover, New York, 1958), p. 192.
[5] A. Einstein, S. B. preuss. Akad. Wiss., 478 (1918), including Weyl's reply.
[6] F. London, Z. Phys. **42**, 375 (1927).
[7] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **209**, 291 (1951).
[8] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **212**, 330 (1952).
[9] W. Pauli, *Theory of Relativity*, translated by G. Field (Dover, New York, 1958), p. 192.
[10] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw–Hill, New York, 1965), p. 401.
[11] J. Ehlers, A. E. Pirani, and A. Schild, in *General Relativity*, edited by L. O'Raifeartaigh (Oxford U.P., Oxford, 1972).
[12] R. Utiyama, Prog. Theor. Phys. **50**, 2080 (1973).
[13] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **333**, 403 (1973).
[14] P. G. O. Freund, Ann. Physik **84**, 440 (1974).
[15] R. Utiyama, Prog. Theor. Phys. **53**, 565 (1975).
[16] P. G. Bergmann, *Introduction to the Theory of Relativity* (Dover, New York, 1976), Chap. XVI.
[17] K. Hayashi, M. Kasuya, and T. Shirafuji, Prog. Theor. Phys. **57**, 431 (1977).
[18] K. Hayashi and T. Kugo, Prog. Theor. Phys. **61**, 339 (1979).
[19] J. Audretsch, Phys. Rev. D **27**, 2872 (1983).
[20] J. Audretsch, F. Gähler, and N. Straumann, Commun. Math. Phys. **95**, 41 (1984).
[21] D. Ranganathan, J. Math. Phys. **28**, 2437 (1986).
[22] H. Cheng, Phys. Rev. Lett. **61**, 2182 (1988).
[23] F. Mansouri, Phys. Rev. Lett. **42**, 1021 (1979).
[24] F. Mansouri and C. Schaer, Phys. Lett. B **101**, 51 (1981).
[25] A. Ferber and P. G. O. Freund, Nucl. Phys. B **122**, 170 (1977).
[26] J. Crispim-Romão, A. Ferber, and P. G. O. Freund, Nucl. Phys. B **126**, 429 (1977).
[27] M. Kaku, P. K. Townsend, and P. Van Nieuwenhuizen, Phys. Lett. B **69**, 304 (1977).
[28] J. T. Wheeler, Phys. Rev. D **44**, 1769 (1991).
[29] É. Cartan, *La théorie des groupes finis et continus et la géométrie différentielle* (Gauthier-Villars, Paris, 1937); C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1970) and refs. therein.
[30] F. Klein, ''Erlangerprogram: Vergleichende Betrachtungen über neuere geometrischen Forschungen,'' (1872), translated in Bull. Am. Math. Soc. **2**, 215 (1893).
[31] N. Steenrod, *The Topology of Fibre Bundles* (Princeton U.P., Princeton, NJ, 1951).
[32] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1963).
[33] R. Utiyama, Phys. Rev. **101**, 1597 (1956).
[34] T. W. B. Kibble, J. Math. Phys. **2**, 212 (1961).
[35] S. W. MacDowell and F. Mansouri, Phys. Rev. Lett. **38**, 739 (1977).
[36] P. G. O. Freund, *Introduction to Supersymmetry* (Cambridge U.P., Cambridge, 1986), Chap. 21, pp. 99–105.
[37] C. Chevalley, *Theory of Lie Groups* (Princeton U.P., Princeton, NJ, 1946).
[38] Y. H. Clifton, private communication.
[39] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[40] L. P. Hughston and T. R. Hurd, Phys. Rep. **100**(5), 273 (1983).
[41] J. T. Wheeler, *Proceedings of the Seventh Marcel Grossman Meeting on General Relativity*, edited by R. T. Jantzen and G. M. Keiser (World Scientific, London, 1996), pp. 457–459.
[42] P. Szekeres, Proc. R. Soc. London, Ser. A **274**, 206 (1963).
[43] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), p. 488.
[44] E. Santamato, Phys. Rev. D **29**, 216 (1984).
[45] E. R. Caianiello, M. Gasperini, E. Predazzi, and G. Scarpetta, Phys. Lett. A **132**, 2, 82 (1988).
[46] J. T. Wheeler, Phys. Rev. D **41**, 431 (1990).
[47] E. R. Caianiello, A. Feoli, M. Gasperini, and G. Scarpetta, Int. J. Theor. Phys. **29**, 2, 131 (1990), and references therein.
[48] J. T. Wheeler, ''Why quantum mechanics is complex,'' to be published in Bull. Astron. Soc. India, Dec. 1997.
[49] J. T. Wheeler, http://xxx.lanl.gov/abs/hep-th/9706215, submitted for publication.
[50] M. Gell-Mann, Cal. Tech. field theory seminar (1985).
[51] B. Mashhoon, H. Liu, and P. Wesson, Phys. Lett. B **331**, 305 (1994).
[52] H. Liu and B. Mashhoon, Ann. Phys. **4**, 565 (1995).
[53] J. T. Wheeler, ''Quanta without quantization,'' Mod. Phys. Lett. A **12**(29), 2175–2181 (1997).
[54] J. T. Wheeler, ''String without strings,'' http://xxx.lanl.gov/abs/hep-th/9706209, submitted for publication.

On the second law of thermodynamics and contact geometry

M. Chen
Department of Mathematics, Vanier College, 821 Ste. Croix Avenue, St. Laurent, Quebec H4L 3X9, Canada

H. C. Tseng
Department of Physics, National Chung-Hsing University, Taichung, Taiwan

(Received 19 August 1997; accepted for publication 30 September 1997)

In this paper we consider the second law of thermodynamics for a dissipative system and its symmetry property in terms of contact geometry. We first show that the inaccessibility condition of Caratheodory and the assumption of semipositive definite property of the dissipative energy are equivalent to Clausius' inequality. The inaccessibility condition then gives rise to a generalized Gibbs relation (GGR). By means of the GGR a 1-form ω can be defined such that the zero of ω reproduces the GGR. Such 1-form ω has the property $\omega\wedge(d\omega)^n\neq 0$ and $\omega\wedge(d\omega)^{n+1}=0$. The integral surface of the GGR is an n-dimensional 1-graph space G (Legendre submanifold) of a 1-jet space $J^1(E_n,R)$, where E_n is the base space of $J^1(E_n,R)$ with thermodynamic variables as its coordinates. The $(2n+1)$-dimensional $J^1(E_n,R)$ equipped with the 1-form ω is also called a contact bundle K, where the intensive thermodynamic variables are considered as the contact elements to K at every x. Next we construct an isovector field X_f such that the inaccessibility condition is invariant under the contact transformations generated by X_f. Finally, suppose under some specific assumptions the dynamical equations of the thermodynamic variables x can be approximated by the flow equations of a vector field X_G on G. We can lift X_G to X_f such that the 1-graph space G as well as the inaccessibility condition are preserved under the contact transformations generated by X_f.
© *1998 American Institute of Physics.* [S0022-2488(98)03801-8]

I. INTRODUCTION

In equilibrium thermodynamics it is well recognized that Caratheodory's inaccessibility condition[1] is equivalent to the Gibbs relation where the integral surface of the Gibbs relation gives rise to the fundamental equation of equilibrium thermodynamics. In 1973 Hermann[2] suggested that classical thermodynamics could be formulated in a geometric setting in terms of a $(2n+1)$-dimensional contact manifold K with thermodynamic space E_n as the base space of K. Recently, Mrugala, Nulton, Schön, and Salamon[3] further studied the application of contact geometry to equilibrium thermodynamics pioneered by Hermann. In this paper we generalize the work of Mrugala *et al.* to irreversible thermodynamics. First we show that Caratheodory's inaccessibility condition and the assumption of semipositive definite property of dissipative energy are equivalent to Clausius' inequality.[4] The inaccessibility condition then yields the generalized Gibbs relation as well as a generalized Helmholtz potential function. We then define a 1-form ω in terms of the generalized Gibbs relation (GGR) such that the zero of ω reproduces the GGR again. The integral surface of the Pfaffian equation $\omega=0$ is an n-dimensional 1-graph space G (Legendre submanifold) of the 1-jet space $J^1(E_n,R)$ with coordinate cover (x,x^0,y), where x, x^0, y are the coordinates appearing in the GGR. In fact, x denotes the thermodynamic variables, y the intensive thermodynamic variables, and $x^0=g(x)$ is the integral of the GGR. The $(2n+1)$-dimensional 1-jet space $J^1(E_n,R)$ equipped with the 1-form ω is also called a contact bundle K whose base E_n is the thermodynamic space with coordinates x. The contact structure of the contact bundle is

specified by the 1-form ω where $\omega\wedge(d\omega)^n\neq 0$ and $\omega\wedge(d\omega)^{n+1}=0$. Let f be a C^1 function defined on K. To every f we can construct a unique vector field $X_f\in TK$ which transforms the contact ideal generated by ω into itself. The orbital equations of X_f (flows of X_f) form a 1-parameter group of transformations (contact transformations) that carry a Legendre submanifold into another Legendre submanifold. Second, we examine the conditions on f such that X_f preserves the Legendre submanifold **L**. We show that the flows of X_f lie on **L** if and only if f vanishes on **L**. As a consequence of this result, the inaccessibility condition is invariant under the contact transformations generated by such vector field X_f. Finally, suppose under some specific assumptions the dynamical equations of x can be formulated as the flows of a vector field X_G on the 1-graph space G, we can embed the flows of X_G in the flows of X_f such that G is invariant under the contact transformations generated by X_f.

Consider a system of molecules in r components contained in a region $\Omega\subset\Re^3$ with volume V where no chemical reactions take place. Let ϵ be the internal energy density, $\nu=\rho^{-1}$ the specific volume, c_i the molar fraction of species i, p the hydrostatic pressure, and $\phi_i^{(\alpha)}$ the generalized fluxes. For example, $\phi_i^{(1)}=\mathbf{J}_i$ is the mass flux vector, $\phi_i^{(2)}=\boldsymbol{\pi}_i$ is the traceless symmetric part of the pressure tensor $\overleftrightarrow{P}_i$, $\phi_i^{(3)}=\mathbf{Q}_i$ is the heat flux vector, etc. Here the subscript i refers to the ith molecular species and α represents the various kinds of fluxes. Except for the case of $\alpha=1$ and $\alpha=3$, in general $\phi_i^{(\alpha)}$ is a tensor function of order α which is related to the velocity moments given by Grad or by Eu.[5] Let $\hat{\phi}_i^{(\alpha)}=\rho^{-1}\phi_i^{(\alpha)}$. From physical point of view we may include sufficiently many fluxes for an appropriate description of irreversible processes. As in extended irreversible thermodynamics[6] we consider $\hat{\phi}_i^{(\alpha)}$ as well as the conserved thermodynamic variables ϵ, υ, c_i as a set of thermodynamic variables. For convenience we denote $x=(x^1,x^2,\ldots,x^n)=(\epsilon,\upsilon,c_1,\ldots,\hat{\phi}_r^{(k)})$. Let $\theta\in[a,b]$ be an empirical temperature in an arbitrary scale which can be employed to measure the hotness or the coldness of a nonequilibrium system at a local point $(\mathbf{r},t)$. At the moment it is difficult to discuss the relationship between θ and x. However, in Sec. II we show that θ is a monotonically increasing function of the local thermodynamic temperature T which can be considered as a C^1 function of x via the constitutive relation. Consider the fiber bundle $\Re^3\times\Re^1\times E_n$ with $(\mathbf{r},t)\in\Re^3\times\Re^1$ as the base space and E_n as the fiber, where E_n is an n-dimensional space with the fields x as local coordinates. The local state of a nonequilibrium system is then specified by x, where $x^i:\Re^3\times\Re^1\rightarrow\Re$ is assumed to be a function of x of class C^1 in Ω and satisfies the following set of dynamical equations:

$$d_t\rho=-\rho\nabla\cdot\mathbf{u}, \tag{1}$$

$$\rho d_t c_i=-\nabla\cdot\mathbf{J}_i,\quad 1\leqslant i\leqslant r, \tag{2}$$

$$\rho d_t\mathbf{u}=-\nabla\cdot\overleftrightarrow{P};\quad \overleftrightarrow{P}=\sum_a\overleftrightarrow{\pi}_a+p_b\hat{I}+p\hat{I}, \tag{3}$$

$$\rho d_t\epsilon=-\nabla\cdot\mathbf{Q}-\overleftrightarrow{P}:\nabla\mathbf{u}, \tag{4}$$

$$\rho d_t\hat{\phi}_i^{(\alpha)}=Z_i^{(\alpha)}+\Lambda_i^{(\alpha)},\quad 1\leqslant i\leqslant r,\quad 1\leqslant\alpha\leqslant k, \tag{5}$$

where $d_t=\partial_t+\mathbf{u}\cdot\nabla$ is the substantial differential operator, $\mathbf{u}$ the hydrodynamic velocity, p_b the bulk viscous pressure (the excess normal stress), $\hat{I}$ a unit second rank tensor. Furthermore, $Z_i^{(\alpha)}$ is a kinematic term depending on $\hat{\phi}_i^{(\alpha)}$ as well as the gradients of ϵ, υ, c_i, etc. In addition it contains the term $-\nabla\cdot\Psi_i^{(\alpha)}$, where $\Psi_i^{(\alpha)}$ is the flux of $\hat{\phi}_i^{(\alpha)}$.[7] On the other hand, $\Lambda_i^{(\alpha)}$ is a dissipative term which does not involve the gradients of x. In general $\Lambda_i^{(\alpha)}$ is a nonlinear function of x such that the decomposition of $Z_i^{(\alpha)}$ and $\Lambda_i^{(\alpha)}$ in (5) is unique. The detailed expressions of $Z_i^{(\alpha)}$ and $\Lambda_i^{(\alpha)}$ depend on the specific model under consideration. For example, Jou *et al.*[8] considered the following set of equations for (5):

$$\rho d_t p_b = -\tau_0^{-1}\rho\zeta\nabla\cdot\mathbf{u} + \tau_0^{-1}\rho\beta'\zeta g\nabla\cdot\mathbf{Q} - \tau_0^{-1}\rho p_b = Z^b + \Lambda^b, \tag{6a}$$

$$\rho d_t \mathbf{Q} = -\tau_1^{-1}\rho\lambda\nabla g + \tau_1^{-1}\rho\beta\lambda g^2\nabla\cdot\overleftrightarrow{\pi} + \tau_1^{-1}\rho\beta'\lambda g\nabla p_b - \tau_1^{-1}\rho\mathbf{Q} = Z^h + \Lambda^h, \tag{6b}$$

$$\rho d_t \overleftrightarrow{\pi} = -\tau_2^{-1}2\rho\eta[\nabla\mathbf{u}]^{(2)} + \tau_2^{-1}2\rho\beta\eta g[\nabla\mathbf{Q}]^{(2)} - \tau_2^{-1}\rho\overleftrightarrow{\pi} = Z^t + \Lambda^t, \tag{6c}$$

where $\Lambda^b = -\tau_0^{-1}\rho p_b$, $\Lambda^h = -\tau_1^{-1}\rho\mathbf{Q}$, $\Lambda^t = -\tau_2^{-1}\rho\overleftrightarrow{\pi}$, ζ, λ, and η are the bulk viscosity, thermal conductivity, and shear viscosity, respectively, τ_0, τ_1, τ_2 are the relaxation times for the fluxes p_b, $\mathbf{Q}$, and $\overleftrightarrow{\pi}$, respectively, β' and β are constants, and g is a positive, monotonically increasing function of θ of class C^1 on $[a,b]$. Later it will be identified with the local thermodynamic temperature T in absolute temperature scale. Furthermore, $[\nabla\mathbf{u}]^{(2)}$ represents the traceless symmetric part of $[\nabla\mathbf{u}]$.

In equilibrium thermodynamics the static pressure p is considered as a force for the volume change, and the chemical potentials μ_i as the generalized forces for the change of matter. Thus in the absence of external forces, the infinitesimal work dw is given by $dw = -p\,d\nu + \Sigma_i\mu_i dc_i$. A closer examination of (4) shows that the contribution of work due to $\Sigma_i\mu_i dc_i$ is not included in (4). Furthermore, since dissipative fluxes $\hat{\phi}_i^{(\alpha)}$ are considered as the (additional) independent variables in addition to ϵ, ν, and c_i, the classical internal energy balance Eq. (4) must be generalized in an appropriate manner so that contributions due to the dissipative fluxes are properly taken into consideration. To this end, we multiple (1) by p, (2) by $-\mu_i$ and sum over i, (5) by $X_i^{(\alpha)}$ and sum over i and α. Adding these results to (4) yields the following equation equivalent to (4):

$$\rho\left\{ d_t\epsilon + p d_t\nu - \sum_i \mu_i d_t c_i + \sum_{i,\alpha} X_i^{(\alpha)} : d_t\hat{\phi}_i^{(\alpha)} \right\}$$

$$= -\nabla\cdot\mathbf{Q} - (\overleftrightarrow{\pi} + p_b\hat{I}) : \nabla\mathbf{u} + \sum_i \mu_i\nabla\cdot\mathbf{J}_i + \sum_{i,\alpha} X_i^{(\alpha)} : [Z_i^{(\alpha)} + \Lambda_i^{(\alpha)}] = -\nabla\cdot\mathbf{Q}_c + \Sigma_d, \tag{7}$$

where

$$\mathbf{Q}_c = \mathbf{Q} - \sum_i \mu_i\mathbf{J}_i + \sum_{i,\alpha} X_i^{(\alpha)} : \Psi_i^{(\alpha)}, \tag{8}$$

$$\Sigma_d = -(\overleftrightarrow{\pi} + p_b\hat{I}) : \nabla\mathbf{u} + \sum_i \mathbf{J}_i\cdot\nabla\mu_i - \sum_{i,\alpha} \Psi_i^{(\alpha)} : (\nabla\cdot X_i^{(\alpha)}) + \sum_{i,\alpha} X_i^{(\alpha)} : [Z_i^{(\alpha)} + \nabla\cdot\Psi_i^{(\alpha)} + \Lambda_i^{(\alpha)}]. \tag{9}$$

Here '':'' denotes contraction of tensors, and $X_i^{(\alpha)}$ is the generalized potential conjugate to $\hat{\phi}_i^{(\alpha)}$. Notice that $Z_i^{(\alpha)}$ contains a term $-\nabla\cdot\Psi_i^{(\alpha)}$ which cancels the term $\nabla\cdot\Psi_i^{(\alpha)}$ in the last term of (9). In general, $X_i^{(\alpha)}$ is a nonlinear function of x. However, if $X_i^{(\alpha)}$ is linearized in powers of $\hat{\phi}_i^{(\alpha)}$, then $X_i^{(\alpha)}$ can be written as

$$X_b = -\tau_0(\rho\zeta)^{-1}p_b, \quad X_h = -\tau_1(\rho\lambda g)^{-1}\mathbf{Q}, \quad X_t = -\tau_2(\rho 2\eta)^{-1}\overleftrightarrow{\pi} \tag{10}$$

and so on, where X_b, X_h, and X_i are the conjugates of p_b, $\mathbf{Q}$, and $\overleftrightarrow{\pi}$, respectively.

Since the mass fluxes $\mathbf{J}_i$ as well as the tensorial fluxes $\Psi_i^{(\alpha)}$ can also contribute to the heat flux, $\mathbf{Q}_c$ is the net heat flux across the surface enclosing the local system centered at $\mathbf{r}$ with volume element $d\mathbf{r}$. The outward normal to the surface is assigned as the positive direction of $\mathbf{Q}_c$. On the other hand, it is well known that $\nabla\mathbf{u}$ and $\nabla\mu_i$ can be considered as the thermodynamic forces attributable to the viscous flow and diffusion, respectively. Similarly $\nabla\cdot X_i^{(\alpha)}$ can also be consid-

ered as the generalized forces attributable to the tensorial flows $\Psi_i^{(\alpha)}$. Thus the first three terms of Σ_d represent dissipative internal work. We therefore call Σ_d the rate of dissipative energy, which cannot be converted into other form of useful work. Hereafter we assume that Σ_d is semipositive definite.

In view of (7) we define the generalized infinitesimal work dw for a nonequilibrium local system by

$$dw=-p\,d\nu+\sum_i \mu_i dc_i-\sum_{i,\alpha} X_i^{(\alpha)}:d\hat{\phi}_i^{(\alpha)} \tag{11}$$

and define the local rate of change of heat per unit mass (dq/dt) by

$$\rho\,\frac{dq}{dt}=-\nabla\cdot\mathbf{Q}_c\,. \tag{12}$$

According to the traditional convention work is considered positive if it is done on the system, and negative if it is done by the system. On the other hand, heat is considered positive if it is absorbed by the system, and negative if it is liberated by the system.

In terms of (11) and (12), the balance equation for the internal energy density per unit mass by (7) can be rewritten as

$$d\epsilon=dq+dw+dq_d \tag{13}$$

with $dq_d=\rho^{-1}\Sigma_d dt$, and $dx^i=(\nabla x^i)\cdot d\mathbf{r}+(\partial_t x^i)dt$ is the substantial differential for $x^i=\epsilon,\ \nu,\ c_i$, and $\hat{\phi}_i^{(\alpha)}$.

Equation (13) is the first law of thermodynamics in local form for a dissipative system. It should be remarked that dq, dw, and dq_d in (13) cannot be interpreted as 1-forms in the vector space $\wedge^1(E_n)$ of differential forms, since they represent only the infinitesimal changes of the corresponding quantities. In other words, we consider the fiber bundle $E_n\times\wedge^1(E_n)$ with E_n as the base space and $\wedge^1(E_n)$ as the fiber. Denote $\hat{x}=(x^2,\ldots,x^n)$. Suppose p, μ_i, and $X_i^{(\alpha)}$ are C^1 functions of $\hat{x}$ and θ. For convenience we denote $\{-p,\{\mu_i\},-\{X_i^{(\alpha)}\},1\leqslant i\leqslant r,1\leqslant\alpha\leqslant k\}$ by $(w_2,w_3,\ldots,w_n)$. Define the work 1-form by

$$w=-p\,d\nu+\sum_i \mu_i dc_i-\sum_{i,\alpha} X_i^{(\alpha)}:d\hat{\phi}_i^{(\alpha)}=\sum_{j\geqslant 2} w_j(\hat{x},\theta)dx^j. \tag{14}$$

Let $\xi=d\epsilon+p\,d\nu-\Sigma_i\mu_i dc_i+\Sigma_{i,\alpha}X_i^{(\alpha)}:d\hat{\phi}_i^{(\alpha)}=dx^1-w$. We now formulate the local form of the second law as

$$\xi\wedge d\xi=0, \tag{15a}$$

$$\Sigma_d\geqslant 0;\ \ \Sigma_d=0\quad \text{at thermodynamic equilibrium.} \tag{15b}$$

In the next section we consider the global formulations of the first law and the second law.

II. GLOBAL FORMULATION OF THE FIRST LAW AND THE SECOND LAW

In order to formulate the first law and the second law in global form, we first define $(E,V,m_iN_i,\Phi_i^{(\alpha)})$ by

$$(E,V,m_iN_i,\Phi_i^{(\alpha)})=\int d\mathbf{r}\rho(\epsilon,\nu,c_i,\hat{\phi}_i^{(\alpha)}).$$

Suppose f is a function of $(\overline{r,t})$. Let $F(t)=\int d\mathbf{r}\rho f(\mathbf{r},t)$. It can be shown that

$$\frac{dF}{dt}=\int\left[\frac{\partial}{\partial t}(\rho f)+\nabla\cdot(\mathbf{u}\rho f)\right]d\mathbf{r}=\int d\mathbf{r}\rho d_t f.$$

Here $\mathbf{u}$ is the hydrodynamic velocity, $d_t=\partial_t+\mathbf{u}\cdot\nabla$, and f may be a scalar function, a vector function, or a tensor function of any order. Next, let $(p^*,\mu_i^*,(X_i^{(\alpha)})^*)$ be the global variables conjugate to $(V,N_i,\Phi_i^{(\alpha)})$. We define p^*, (μ_i^*) and $(X_i^{(\alpha)})^*$, respectively, by

$$\int d\mathbf{r}p(\rho d_t\nu)=p^*\frac{dV}{dt},$$

$$\int d\mathbf{r}\mu_i(\rho d_t c_i)=\mu_i^*\frac{dN_i}{dt},$$

$$\int d\mathbf{r}X_i^{(\alpha)}:(\rho d_t\phi_i^{(\alpha)})=(X_i^{(\alpha)})^*:\frac{d\Phi_i^{(\alpha)}}{dt},$$

where the molecular mass m_i of species i has been absorbed in μ_i^*. Denote $\hat{z}=(V,\{N_i\},\{\Phi_i^{(\alpha)}\})=(z^2,\ldots,z^n)$ and $\hat{z}^*=(p^*,\{\mu_i^*\},\{(X_i^{(\alpha)})^*\})=(z_2^*,\ldots,z_n^*)$. Then z^i's are the global thermodynamic variables, while z_i^*'s are their conjugate variables. It should be remarked that z_i^* is a function of z constructed according to phenomenological considerations.

The global formulation of the first law and the second law depend on the boundary condition of the material system. For example, if the boundary is only permeable to the mass flows, then the infinitesimal work dW can be written as

$$dW=-p^*dV+\sum_i\mu_i^*dN_i.$$

By the volume integral of (7) we obtain the first law in global form as

$$dE=dW+dQ+dQ_d, \tag{16}$$

with $dQ_d=[\int d\mathbf{r}\Sigma_d]dt$, and

$$dQ=dt\left[\int d\mathbf{r}(-\nabla\cdot\mathbf{Q}_c)\right]=dt\left[\int d\mathbf{r}\rho\frac{dq}{dt}\right],$$

$$dW=dt\left[\int d\mathbf{r}\rho\left\{-pd_t\nu+\sum_i\mu_i d_t c_i-\sum_{i,\alpha}X_i^{(\alpha)}:d_t\hat{\phi}_i^{(\alpha)}\right\}\right],$$

$$dE=dt\left[\int d\mathbf{r}\rho d_t\epsilon\right].$$

Let

$$\Omega=-p^*dV+\sum_i\mu_i^*dN_i-\sum_{i,\alpha}(X_i^{(\alpha)})^*:d\Phi_i^{(\alpha)}$$

be the work 1-form,[9] and set

$$\hat{\Omega}(t)=dE-\Omega.$$

The global form of the second law can be written as

$$\hat{\Omega}\wedge d\hat{\Omega}=0, \tag{17a}$$

$$Q_d\geqslant 0;\ \ Q_d=0 \quad \text{at thermal equilibrium.} \tag{17b}$$

Notice that $\hat{\Omega}$ is the heat 1-form if $dQ_d=0$. Thus (17a) is the inaccessibility condition of Caratheodory. In a previous paper we have shown that (17a) and (17b) in conjunction with (16) are equivalent to Kelvin's principle as well as Clausius' principle.[10]

Next we consider Clausius' inequality for cyclic processes where the material system is in contact with heat baths during the cycles. Suppose γ is a closed path in the thermodynamic space B with coordinates $(E,V,N_i,\Phi_i^{(\alpha)})$. Let T^* be the temperature of the heat reservoir and dQ the amount of heat exchanged between the system and the reservoir during an infinitesimal portion of γ. Then Clausius' inequality can be expressed as

$$\oint_\gamma ((T^*)^{-1}dQ)(\gamma)\leqslant 0. \tag{18}$$

It is well known that the path integral in (18) vanishes at thermodynamic equilibrium. Hence the maximum value of the path integral is zero. To this end we introduce a semipositive quantity $D(t)$ such that

$$\oint_\gamma ((T^*)^{-1}dQ+dD)(\gamma)=0 \tag{19}$$

subject to the constraint of the first law given by (16). Thus we rewrite (19) in terms of a Lagrange multiplier λ as the following:

$$\begin{aligned}&\oint_\gamma ((T^*)^{-1}dQ+dD+\lambda[dE-\Omega-dQ-dQ_d])(\gamma)\\ &\quad=\oint_\gamma (\lambda(dE-\Omega)+((T^*)^{-1}-\lambda)dQ+(dD-\lambda dQ_d))(\gamma)=0\end{aligned} \tag{20}$$

for all λ and γ in B.

From a physical point of view dQ, dQ_d, and dD depend on the irreversible path γ. Furthermore, $(dE-\Omega)$, dQ, and $(dD-\lambda dQ_d)$ have different physical significance. The term $(dE-\Omega)$ represents the change of internal energy and work along γ, dQ represents the change of heat between the system and its surroundings, while dQ_d represents the dissipative energy. Since (20) must be satisfied for all λ and all γ in B, we first set $\lambda=(T^*)^{-1}$. But the first term $(dE-\Omega)$ and the last term $(dD-\lambda dQ_d)$ are functionally independent. Thus (20) can be satisfied if and only if

$$\oint_\gamma (T^*)^{-1}(dE-\Omega)=0 \quad \text{and} \quad \oint_\gamma [dD-(T^*)^{-1}dQ_d]=0$$

for all γ. Consequently $dD=(T^*)^{-1}dQ_d$, or $dQ_d=T^*dD\geqslant 0$. In addition, there exists a scalar-valued function S defined on B such that

$$\hat{\Omega}=dE-\Omega=T^*dS. \tag{21}$$

Hence, by Clausius' inequality (18) we can obtain (21). Conversely, by (16), (17a), and (17b) we can obtain Clausius' inequality (18). In (21), $(T^*)^{-1}$ is an integrating factor of $\hat{\Omega}=dE-\Omega$. Equation (21) can be rewritten as the following generalized Gibbs relation:

$$T^*dS=dE+p^*dV-\sum_i \mu_i^* dN_i+\sum_{i,\alpha} (X_i^{(\alpha)})^*:d\Phi_i^{(\alpha)}. \tag{22}$$

Here S is called the generalized entropy function or calotropy function.

Proposition 1: The second law of thermodynamic formulated in (17a) and (17b) is equivalent to Clausius' inequality (18). Furthermore, there exists a generalized entropy function S such that $\hat{\Omega}=dE-\Omega=T^*dS$.

Now we consider the emergence of $(T^*)^{-1}$ as an integrating factor of $\hat{\Omega}$ via an alternative approach in terms of (17a). From geometric point of view this alternative approach is closely related to the notion of contact geometry. It is well known that (17a) is equivalent to the condition $d\hat{\Omega}=0$ for $\hat{\Omega}=0$.[11] This condition in turn is equivalent to the following condition:

$$\partial_k w_l=\partial_l w_k \quad \text{for all} \quad \theta\in[a,b],k,l\geqslant 2, \tag{23}$$

$\partial_k=(\partial/\partial z^k)$, $w=(-p^*,\{\mu_i^*\},\{-(X_i^{(\alpha)})^*\})=(w_2,\ldots,w_n)$. Hereafter we assume that the w_k's are C^1 functions of $(\hat{z},\theta)$. So far θ has been considered as a continuous parameter which does not participate in the exterior calculus in $\hat{\Omega}$ and $d\hat{\Omega}$. We now consider θ as well as $\hat{z}$ as independent variables. By (23) there exists a C^1 function $F(\hat{x},\theta)$ such that $w_j=-\partial_j F$. Let $u=-\partial_\theta F$. Then $\hat{\Omega}$ can be rewritten as

$$\hat{\Omega}=d(E-F)-ud\theta$$

and

$$d\hat{\Omega}=-du\wedge d\theta.$$

Since $\hat{\Omega}\wedge d\hat{\Omega}=-d(E-F)\wedge du\wedge d\theta=0$, there exists a C^1 function $H(u(\hat{z},\theta),\theta)$ such that $E-F=H(u(\hat{z},\theta),\theta)$. Let $H_1=\partial_u H\neq 0$, and $H_2=\partial_\theta H$. Then $\hat{\Omega}$ can be expressed as

$$\hat{\Omega}=H_1du+(H_2-u)d\theta. \tag{24}$$

Notice that (17a) is also equivalent to $d\hat{\Omega}=\eta\wedge\hat{\Omega}$ for some 1-form η. Thus

$$d\hat{\Omega}=-du\wedge d\theta=\eta\wedge\hat{\Omega}=H_1\eta\wedge du+(H_2-u)\eta\wedge d\theta, \tag{25}$$

where η, H_1 and H_2 are determined later.

Set

$$\eta=f_1(u,\theta)du+f_2(u,\theta)d\theta.$$

Then

$$d\hat{\Omega}=-du\wedge d\theta=[(H_2-u)f_1-H_1f_2]du\wedge d\theta,$$

which implies the following constraints for H_1, H_2, f_1 and f_2,

$$H_1f_2-(H_2-u)f_1=1. \tag{26}$$

Since $(\eta+\lambda\hat{\Omega})=\eta\wedge\hat{\Omega}$, we look for some function $\lambda=\lambda(u,\theta)$ such that $(\eta+\lambda\hat{\Omega})=d\mu$ is an exact 1-form. By (26) we have

$$\begin{aligned}\eta+\lambda\hat{\Omega}=d\mu&=(f_1+\lambda H_1)du+[f_2+\lambda(H_2-u)]d\theta\\&=H_1^{-1}d\theta+(f_1+\lambda H_1)[du+H_1^{-1}(H_2-u)d\theta].\end{aligned}\tag{27}$$

It is evident that $(\eta+\lambda\hat{\Omega})$ becomes an exact 1-form if $f_1+\lambda H_1=0$ with $H_1=g(\theta)$. Thus

$$d\mu=g^{-1}d\theta,$$

$$\hat{\Omega}=g(\theta)du+(H_2-u)d\theta,$$

and

$$d\hat{\Omega}=-g'(\theta)du\wedge d\theta+\left(\frac{\partial H_2}{\partial u}-1\right)du\wedge d\theta=-du\wedge d\theta.$$

Consequently $(\partial H_2/\partial u)=g'$, and $\hat{\Omega}$ becomes

$$\hat{\Omega}=gdu+(g'-1)ud\theta+c(\theta)d\theta=gud\left\{\ln(|gu|)-\int g^{-1}d\theta\right\}+db,\tag{28}$$

where $db=c(\theta)d\theta$.

Again by (17a) we can choose $c(\theta)=0$ so that $\hat{\Omega}$ can be rewritten as $\hat{\Omega}=sd\varphi$. By (27) and (28) we obtain

$$du=\eta+\lambda\hat{\Omega}=\eta+\lambda gud\ \ln(|gu|)-\lambda gu(g^{-1}d\theta)=g^{-1}d\theta.\tag{29}$$

We can now determine η and λ. Let $\lambda gu=-1$. Then $\lambda=-(gu)^{-1}$ and $\eta=u^{-1}du+g'g^{-1}d\theta$. Therefore $f_1=u^{-1}$, $f_2=g'g^{-1}$, $H_1=g$, $H_2=g'u$ and

$$\hat{\Omega}=gud\left\{\ln(|gu|)-\int g^{-1}d\theta\right\},\tag{30}$$

$$d\hat{\Omega}=(\eta+\lambda\hat{\Omega})\wedge\hat{\Omega}=g^{-1}(\theta)d\theta\wedge\hat{\Omega}.\tag{31}$$

So far the results in (30) and (31) are unique in terms of g. We can choose $g>0$ for all $\theta\in[a,b]$. Equation (31) can be solved to yield

$$\hat{\Omega}=T^*dS,\tag{32}$$

where S is a scalar-valued function defined on B, and

$$T^*=T_0^*\ \exp\left(\int g^{-1}d\theta\right),\quad T_0^*>0.$$

Notice that (32) is identical to (21). Since we have chosen $g>0$ for all $\theta\in[a,b]$, T^* can be identified as the thermodynamic temperature in the absolute temperature scale. Finally we consider the relationship between u and S. To this end we denote $F(\hat{z},\theta)=F^*(\hat{z},T^*)$, and $u^*=-\partial_{T^*}F^*(\hat{z},T^*)$. Then $u=-\partial_\theta F(\hat{z},\theta)=g^{-1}T^*u^*$, or $gu=T^*u^*$. Equation (30) then becomes

$$\hat{\Omega}=T^*dS=gud\left\{\ln(|gu|)-\int g^{-1}d\theta\right\}=T^*du^*.$$

Without loss of generality we set $u^*=S$. On the other hand, $\hat{\Omega}=d(E-F)-ud\theta$ can be rewritten as

$$\hat{\Omega}=T^*dS=dE-dF^*-SdT^*$$

which gives rise to the following result:

$$dF^*=-SdT^*+dE-T^*dS=-SdT^*-p^*dV+\sum_i \mu_i^* dN_i-\sum_{i,} (X_i^{(\alpha)}):d\Phi_i^{(\alpha)}. \tag{33}$$

In view of (33) F^* is called the generalized Helmholtz potential function. We summarize our results as the following.

Proposition 2: Suppose $\partial_k w_l=\partial_l w_k$ for all $\theta\in[a,b]$ and $k,l\geqslant 2$. Then there exists a thermodynamic temperature T^* and a scalar-valued function S defined on B such that $\hat{\Omega}=T^*dS$. Furthermore, there exists a function F^* such that dF^* is given by (33).

We make two remarks before closing the discussions of this section. First, suppose the local formulation of the second law is expressed as $\xi\wedge d\xi=0$ and $\Sigma_d\geqslant 0$. By following the same arguments as given above, we can prove the existence of a local thermodynamic temperature T and a scalar-valued function $\psi=\psi(\epsilon,v,c_i,\phi_i^{(\alpha)})$ such that $\xi=Td\psi$. Second, denote $x=(S,V,\{N_i\},\{\Phi_i^{(\alpha)}\})=(x^1,x^2,\ldots,x^n)=(x^1,\hat{x})$ and $y=(T^*,-p^*,\{\mu_i^*\},\{-(X_i^{(\alpha)})^*\})=(y_1,\ldots,y_n)=(y_1,\hat{y})$. Then the generalized Gibbs relation (22) can be rewritten as

$$dx^0=\sum_i y_i dx^i=y_i dx^i, \tag{34}$$

where $x^0=E=g(x)$ with $y_i=\partial_i g$. Let $F^*=E-T^*S=x^0-y_1x^1$. Then the 1-form for F^* given by (33) becomes

$$du^0=-x^1dy_1+\hat{y}_i d\hat{x}^i, \tag{35}$$

where $u^0=F^*=x^0-y_1x^1$. Thus the relation $u^0=x^0-y_1x^1$ defines a transformation φ which carries $(x^1,\ldots,x^n)$ to $(y_1,x^2,\ldots,x^n)$. Let $\omega=dx^0-y_idx^i$ and $\hat{\omega}=du^0+x^1dy_1-\hat{y}_id\hat{x}^i$. Then $\varphi^*\hat{\omega}=\omega$, where φ^* is the pull back of φ. The transformation φ is called the Legendre transformation. The Legendre transformation is closely related to the notion of contact geometry, which plays an important role in thermodynamics. This topic will be discussed in detail in the next section.

III. CONTACT GEOMETRY AND THERMODYNAMICS

In this section we consider the notion of contact geometry[12] in thermodynamics. Let E_n be the thermodynamic base space with coordinate cover $x=(S,V,\{N_i\},\{\Phi_i^{(\alpha)}\})=(x^1,x^2,\ldots,x^n)$ and let $G=E_n\times\mathfrak{R}$ with coordinate cover $(x,x^0)=(x,E)$. By the generalized Gibbs relation $dx^0=y_idx^i$, $y=(T^*,-p^*,\{\mu_i^*\},\{-(X_i^{(\alpha)})^*\})=(y_1,\ldots,y_n)$, x^0 is a C^1 function of x. Denote $x^0=g(x)$. Consider the $(2n+1)$-dimensional kinematic space $K=E_n\times\mathfrak{R}\times E_n'$ with coordinate cover (x,x^0,y) where x, x^0, y are independent coordinates. Define the 1-form ω in terms of the generalized Gibbs relation as

$$\omega=dx^0-y_idx^i\in\Lambda^1(K). \tag{36}$$

For every $x\in E_n$ the vector subspace of codimension 1 of T_xK defined by $\Delta_x=\{\nu\in T_xK|\langle\omega(x),\nu(x)\rangle=\omega_i\nu^i=0;\omega=\omega_idx^i,\nu=\nu^i\partial_i\}$ is called the contact hyperplane to K at x. Since $\omega\wedge(d\omega)^n\neq 0$ and $\omega\wedge(d\omega)^{n+1}=0$, the 1-form ω defines a nondegenerate hyperplane distribution $x\to\Delta_x$, where Δ_x is the kernel of ω. This distribution of hyperplanes (field of hyper-

planes) is called the contact structure of K. If λ is a nowhere vanishing function on K, then $\lambda\omega$ defines the same distribution of contact hyperplanes as ω. The differentiable manifold K equipped with the 1-form ω is called a contact manifold. By Darboux's theorem every differential 1-form defining a nondegenerate field of hyperplanes on a manifold of dimension $(2n+1)$ can be written in some local coordinate system in normal form $\omega = dx^0 - y_i dx^i$, where (x, x^0, y) are the local coordinates of K.

Consider the regular map $\psi: E_n \to G$ defined by $\psi(x) = (x, x^0)$ with $x^0 = g(x)$ and $\psi^* \mu \neq 0$, where $\mu = dx^1 \wedge dx^2 \wedge \ldots \wedge dx^n$, and ψ^* is the pull back of ψ. We extend ψ to the map $\psi: E_n \to K$ by the requirement $\psi^* \omega = 0$. Then $\psi^* \omega = \psi^*(dx^0 - y_i dx^i) = dx^0 - \psi^*(y_i) dx^i = 0$, or $\psi^*(y_i) = \partial_i g$. Thus ψ defines a global section of K. This global section can be identified as the Legendre submanifold $\mathbf{L}$ of K,

$$L = \{(x, x^0, y) \in K | x^0 = g(x), y_i = \partial_i g\}.$$

This Legendre submanifold is sometimes called the 1-graph space G.[13] Although the y_i's are independent of x and x^0 in K, however $\psi^*(y_i) = \partial_i g$, i.e., the pull back of y_i on G is $\partial_i g$.

In the contact manifold K with normal coordinates (x, x^0, y), there is a special vector field $X_E = \partial_0 = (\partial / \partial x^0)$ such that

$$i(X_E)\omega = X_E \rfloor \omega = 1, \quad i(X_E) d\omega = X_E \rfloor d\omega = 0, \tag{37}$$

where "$\rfloor$" is an inner multiplication defined by $i(X)\omega = X \rfloor \omega = \nu^i \omega_i$ for $X = \nu^i \partial_i$ and $\omega = \omega_i dx^i$. Since the tangent bundle TK can be uniquely decomposed into $TK = \ker d\omega \oplus \ker \omega$, where $\ker d\omega$ is a vector bundle generated by X_E and $\ker \omega$ is a horizontal bundle, every vector field X on TK can be uniquely decomposed as

$$X = aX_E + (X - aX_E).$$

By (37) we obtain $a = i(X)\omega$. Thus

$$X = (i(X)\omega)\partial_0 + [X - (i(X)\omega)\partial_0] = X_V + X_H.$$

We can easily verify that $i(X_V)\omega = i(X)\omega = 1$, $i(X_H)\omega = 0$, $i(X_V) d\omega = 0$, and $i(X) d\omega = i(X_H) d\omega$. Thus $X_V \in \ker d\omega$, and $X_H \in \ker \omega$. In other words, X_H lies in the contact hyperplane Δ_x, while X_V is tangent to the fiber if K is considered as a fiber bundle.

To every vector field $X \in TK$, there is a unique differentiable function f defined on K given by $f = i(X)\omega$. Since

$$i(X_H) d\omega = i(X) d\omega = -\{df - [i(X_E) df]\omega\} = -Df,$$

we have $i(X_E) Df = 0$, where

$$Df = df - [i(X_E) df]\omega$$

is called the covariant derivative of f. Let ω^b be the map; $X_H \to Df$. Consider the inverse map $\omega^\# = (\omega^b)^{-1}$. Alternatively, given Df we construct X_H such that $i(X_H) d\omega = -Df = -df + [i(X_E) df]\omega$, and $i(X_H)\omega = 0$. Thus let

$$X_H = u^0 \partial_0 + u^i \partial_i + u_i \partial^i,$$

where $\partial_0 = (\partial / \partial x^0)$, $\partial_i = (\partial / \partial x^i)$, and $\partial^i = (\partial / \partial y_i)$. Then

$$i(X_H) d\omega = u^i dy_i - u_i dx^i = -(y_i \partial_0 + \partial_i) f dx^i - (\partial^i f) dy_i.$$

Hence $u_i=(y_i\partial_0+\partial_i)f$, $u^i=-\partial^i f$ and

$$X_H=u^0\partial_0+(\partial^i f)\partial_i+[(y_i\partial_0+\partial_i)f]\partial^i.$$

On the other hand, $i(X_H)\omega=0$ yields $u^0=-y_i\partial^i f$. Therefore

$$X_H=-[y_i\partial^i f]\partial_0+y_i(\partial_0 f)\partial^i+[(\partial_i f)\partial^i-(\partial^i f)\partial_i]=-\omega^{\#}f.$$

Now that $X_V=(i(X_E)\omega)X_E=f\partial_0$, to every differentiable function f defined on K, there is a unique vector field X_f given by

$$X_f=X_V+X_H=f\partial_0-\omega^{\#}f=f\partial_0+\{-(y_i\partial^i f)\partial_0+y_i(\partial_0 f)\partial^i+[(\partial_i f)\partial^i-(\partial^i f)\partial_i]\}.$$

For convenience we rewrite X_f as

$$X_f=(f-y_i\partial^i f)\partial_0-(\partial_i f)\partial^i+[y_i(\partial_0 f)+(\partial_i f)]\partial^I=f^0\partial_0+f^i\partial_i+f_i\partial^i, \tag{38}$$

where f^0, f^i, and f_i are C^1 functions of (x,x^0,y) given by

$$f^0=f+y_i(-\partial^i f)=f+y_i f^i,$$

$$f^i=-\partial^i f,$$

$$f_i=(\partial_0 f)y_i+\partial_i f=(\partial_i+y_i\partial_0)f.$$

Thus (38) establishes a one-to-one correspondence between a C^1 function defined on K and a vector field (contact vector field) X_f. This one-to-one correspondence gives rise to an isomorphism h from the Lie algebra of vector fields onto the vector space of real-valued functions on K, whose Lie algebra structure is defined by the Jacobi bracket $[f,g]=i([X_f,X_g])\omega$ with $X_f=h^{-1}(f)$ and $X_g=h^{-1}(g)$.

In view of (38) the flows (orbital equations) of X_f can be written as

$$\frac{dx^0}{d\tau}=f^0(x,x^0,y)=f+y_i f^i,$$

$$\frac{dx^i}{d\tau}=f^i(x,x^0,y)=-\partial^i f, \tag{39}$$

$$\frac{dy_i}{d\tau}=f_i(x,x^0,y)=(y_i\partial_0+\partial_i)f.$$

These flow equations generate a 1-parameter family of maps $T_x(\tau)$: $K\to K$ with $T_x(\tau)=\exp(\tau X_f)$. Since X_f can be considered as a derivation on C^1 functions defined on K, the natural extension of X_f as a derivation on differential forms $\wedge(K)$ is the Lie derivative $L(X_f)$, where $L(X_f)\omega=i(X_f)d\omega+d(i(X_f)\omega)$. Let C be the (contact) ideal generated by ω. If $\eta\in C$, then $\eta=\lambda\omega$, $\lambda\in C^0(K)$. By (38) we have $L(X_f)\omega=(\partial_0 f)\omega=\lambda\omega$. Thus $L(X_f)C\subset C$. A vector field which carries a contact ideal onto itself is called an isovector field. Consider the composition of the regular map ψ and the 1-parameter group of maps $T_x(\tau)$, $\psi_x(\tau)=T_x(\tau)\circ\psi$. Notice that

$$\overline{\psi_x^*}(\tau)\mu=\overline{\psi^*}[(e^{\tau L(X_f)})^*\mu]=\overline{\psi^*}[\mu+\tau(L(X_f))^*\mu+\cdots]\neq 0$$

and

$$\overline{\psi_x}(\tau)\omega=\overline{\psi^*}[(e^{\tau L(X_f)})^*\omega]=\overline{\psi^*}[\omega+\tau\lambda\omega+\cdots]=0,$$

$\overline{\psi_x}$ is an embedding of $\overline{\psi}$ in a 1-parameter group of maps generated by X_f. We call $\overline{\psi_x}$ a 1-parameter group of regular maps or contact transformations generated by X_f. As $\psi_x(\tau)\omega=0$, ψ_x carries Legendre submanifold of K into another Legendre submanifold of K.

Let **L** be a Legendre submanifold of K given by the 1-graph space

$$G=\{(x,x^0,y)|x^0=g(x),y_i=\partial_i g,i=1,2,...,n\}.$$

Notice that $\{X_i=y_i\partial_0+\partial_i,\ i=1,2,...,n\}$ is a linearly independent set of vector fields such that $i(X_i)\omega=0$. Thus $X_i\in\ker\omega$. In general $X_f=f^0\partial_0+f^i\partial_i+f_i\partial^i\in\ker\omega$ if and only if $f^0-y_if^i=0$. On the other hand, X_f is tangent to G if and only if $f|_G=0$, i.e., $f^0(x,g(x),\nabla g)=(\partial_i g)f^i(x,g(x),\nabla g)$ where $\nabla g=(\partial_i g,\ldots,\partial_n g)$. By (38) and (39) we notice that

$$\begin{aligned}\left.\frac{dy_i}{d\tau}\right|_G&=(y_i\partial_0+\partial_i)f|_G=(\partial_i g)\partial_0 f^0+\partial_i f^0-(\partial_i g\partial_j g)\partial_0 f^j-(\partial_j g)\partial_i f^j\\&=(\partial_i g)(\partial_j g)\partial_0 f^i+(\partial_i\partial_j g)f^j+(\partial_j g)(\partial_i f^j)\\&\quad-(\partial_i g\partial_j g)\partial_0 f^j-(\partial_j g)\partial_i f^j\\&=(\partial_i\partial_j g)f^j=\frac{d}{d\tau}(\partial_i g).\end{aligned}$$

Hence the flows of X_f lie on G if and only if $f|_G=0$.

Proposition 3: Let f be a scalar-valued function defined on K and let **L** be a Legendre submanifold of K given by the 1-graph space G. Then **L** is invariant under the contact transformations generated by X_f if and only if $f|_G=0$.

Suppose $X_G=f^0(x,x^0)\partial_0+f^i(x,x^0)\partial_i$ is a vector field tangent to G. We can lift X_G to X_f by

$$X_f=X_G+[(y_i\partial_0+\partial_i)f]\partial^i=f^0\partial_0+f^i\partial_i+f_i\partial^i,\tag{40}$$

where $f=f^0-y_jf^i$. Comparing with (38), X_f indeed is the lift of X_G from G to TK. Furthermore, since $X_f|_G=X_G$, X_G is the projection of X_f on G. Hence,

Corollary 1: Let $X_G=f^0(x,x^0)\partial_0+f^i(x,x^0)\partial_i$ be a vector field tangent to G, such that $f^0=(\partial_i g)f^i$. Then X_f given by (40) is the lift of X_G, and X_G is the projection of X_f on G, where $f=f^0-y_if^i$. Furthermore, the 1-graph space G is invariant under the contact transformations generated by X_f.

We now apply the results obtained above to thermodynamics. Let $\hat{\Omega}=dE-\Omega$. The inaccessibility condition (17a) is equivalent to (21), which gives rise to the generalized Gibbs relation (22). Equation (22) can be recast in canonical form as $dx^0=y_idx^i$ with (y,x) as canonical coordinates. Then we define the contact 1-form $\omega=dx^0-y_idx^i$, where $\omega\wedge(d\omega)^n\neq0$ and $\omega\wedge(d\omega)^{n+1}=0$. Let K be a $(2n+1)$-dimensional manifold with coordinate cover (x,x^0,y), whose contact structure is specified by the 1-form ω. The 1-graph space $G=\{(x,x^0,y)|x^0=g(x),y_i=\partial_i g\}$ is a Legendre submanifold of K, where x^0 is the internal energy of the thermodynamic system, $x=(S,V,\{N_i\},\{\Phi_i^{(\alpha)}\})$ is the set of thermodynamic variables, and $y=(T^*,-p^*,\{\mu_i^*\},\{-(X_i^{(\alpha)})^*\})$ is a set of intensive variables conjugate to x. Alternatively we can adopt the entropy representation with $x^0=S$, $x=(E,V,\{N_i\},\{\Phi_i^{(\alpha)}\})$, and $y=((T^*)^{-1},p^*(T^*)^{-1},\{-\mu_i^*(T^*)^{-1}\},\{(X_i^{(\alpha)})^*(T^*)^{-1}\})$.

Corollary 2: The inaccessibility condition (17a) is invariant under the contact transformations generated by X_f given by (38) if $f|_G=0$. Particularly, if $X_G=f^0(x,x^0)\partial_0+f^i(x,x^0)\partial_i$ is a vector field on G, then X_G can be lifted to $X_f=X_G+[(y_i\partial_0+\partial_i)f]\partial^i$, $f=f^0-y_if^i$, such that G is invariant under the contact transformations generated by X_f.

We make some remarks in the following discussions. First, there is a one-to-one correspondence between f and X_f. In general a Legendre submanifold is carried into another Legendre submanifold under the contact transformations generated by X_f. Most of the Legendre transfor-

mations encountered in equilibrium thermodynamics are discrete transformations. It is impossible to find X_f which can generate such discrete transformations. Second, given f we can find the zero set of $f(x,x^0,y)$ and solve the equation $f(x,g(x),\nabla g(x))=0$ for $g(x)$. The solutions of g gives rise to the 1-graph space G, where G is invariant under the contact transformations generated by X_f. It should be noted that in general, the zero set of X_f is a $2n$-dimensional submanifold of X_f, whereas G is an n-dimensional Legendre submanifold of K. Furthermore the solutions of g may not be unique. Other conditions based on thermodynamic considerations must be imposed in order to obtain a unique solution of g. Third, suppose g is given. We can find a vector field $X_G=f^0(x,x^0)\partial_0+f^i(x,x^0)\partial_i$ such that $f^0=(\partial_i g)f^i$, $x^0=g(x)$. However, the choice of f^i is arbitrary. For equilibrium thermodynamics, the flows of X_G can be considered as a thermodynamic process under consideration. Once f^i has been chosen, f can be constructed by $f(x,x^0,y)=f^0(x,x^0)-y_i f^i(x,x^0)$. Finally suppose under some specific assumptions or approximations the dynamical equations of the global thermodynamic variables x^i can be expressed as $(dx^i/dt)=f^i(x,t)$. Furthermore, suppose g is also known. We can construct X_G and lift it to X_f by the method described in Corollary 1. In other words we embed the dynamical equations for x^i in the flow equations of X_f. The thermodynamic laws are then invariant under the contact transformations generated by X_f.

In the next section we consider some examples to further elucidate these remarks. The first three examples are based on equilibrium thermodynamics where $\Phi_i^{(\alpha)}=0$. The last example is associated with nonequilibrium thermodynamics.

IV. EXAMPLES

Example 1: Let $x^0=E$, $x^1=S$, $x^2=V$, $x^3=N$, $y_1=T$, $y_2=-p$, and $y_3=\mu$. The fundamental equation of thermodynamics at equilibrium can be written as $x^0=y_i x^i$, where the Gibbs relation is $dx^0=y_i dx^i$ and the Gibbs–Duhem relation is $x^i dy_i=0$. Let $f=x^0-y_i x^i=f^0-y_i f^i$. Then $f^0=x^0$ and $f^i=x^i$. By (38) and (39) the orbital equations of X_f have the solutions $x^i=x^i(0)e^{-\tau}$, $x^0=x^0(0)e^{-\tau}$, $y_i=y_i(0)$. The zero set of f is simply $x^0=y_i x^i$. If the constitutive relations are assumed for $y_i=y_i(x)$ then $x^0=g(x)=(\partial_i g)x^i=y_i(x)x^i$ is the Legendre submanifold of K with coordinate cover (x,x^0,y). Hence $G=\{(x,x^0,y)|x^0=g(x),y_i=\partial_i g\}$ is invariant under the contact transformation generated by X_f.

Example 2: Let $x^0=S$, $x^1=E$, $x^2=V$, $y_1=T^{-1}$, $y_2=pT^{-1}$. The entropy for an ideal gas can be written as

$$x^0=g(x^1,x^2)=\frac{3R}{2}\ln(x^1)+R\ln(x^2)-\left\{\frac{3R}{2}\ln(c_\nu)+R\ln(R)\right\}.$$

Thus $y_1=\partial_1 g=(3R/2)(x^1)^{-1}$ and $y_2=\partial_2 g=R(x^2)^{-1}$. In other words, we have the constitutive relations $T=(2E/3R)$, $p=V^{-1}(RT)$. Set $f^1=-x^1$ and $f^2=-x^2$. By (39) we have $x^1=x^1(0)e^{-\tau}$, $x^2=x^2(0)e^{-\tau}$, $f^0=(\partial_i g)f^i=-(5R/2)$ and $f=f^0-y_i f^i=-(5R/2)+y_i x^i$. Thus $x^0=-\frac{5}{2}R\tau+C$, $y_1=y_1(0)e^{\tau}$, and $y_2=y_2(0)e^{\tau}$, where $C=\frac{3}{2}R\ln(x^1(0)/c_\nu)+R\ln(x^2(0)/R)$. Notice that $f|_G=-\frac{5}{2}R+(\partial_1 g)x^1+(\partial_2 g)x^2=0$. Since $(y_1/y_2)=(y_1(0)/y_2(0))=p=\text{const}$, the flows of X_f describe an isobaric process on the Legendre submanifold given by the 1-graph space G.

Example 3: Let $x^0=E$, $x^1=S$, $x^2=V$, $x^3=N$, $y_1=T$, $y_2=-p$, and $y_3=\mu$.

(a) Consider $f=E-\frac{3}{2}NRT$ or $f(x,x^0,y)=x^0-(3R/2)x^3y_1=f^0-y_i x^i$. Then $f^0=x^0$, $f^1=\frac{3}{2}Rx^3$, $f^2=f^3=0$ and the vector field X_f associated with f can be constructed according to (38) as $X_f=x^0\partial_0+\frac{3}{2}Rx^3\partial_1+y_1\partial^1-\frac{3}{2}Ry_1\partial^3$. The orbital equations of X_f have the following solutions:

$$x^0=x^0(0)e^{\tau},\quad x^1=x^1(0)+\tfrac{3}{2}Rx^3(0)\tau,\quad x^2=x^2(0),\quad x^3=x^3(0),$$

$$y_1=y_1(0)e^{\tau},\quad y_2=y_2(0),\quad y_3=y_3(0)e^{\tau}-\tfrac{3}{2}Ry_1(0)\tau e^{\tau}.$$

It is important to note that the zero set of f, $x^0=\frac{3}{2}Rx^3y_1$ is a two-dimensional surface in the contact space K with coordinate cover (x,x^0,y). On the other hand it is impossible to determine a unique G by the zero set of f.

(b) Consider a system of monatomic ideal gas at equilibrium. To this end, let $f=\mu N+TS-\frac{5}{2}NRT$ or $f(x,x^0,y)=x^3y_3+x^1y_1-\frac{5}{2}Rx^3y_1=f^0-y_ix^i$. Then $f^0=0$, $f^1=-x^1+\frac{5}{2}Rx^3$, $f^2=0$, $f^3=-x^3$. By (38) we have $f_1=y_1$, $f_2=0$, $f_3=-\frac{5}{2}Ry_1+y_3$. Thus the orbital equations of X_f have the following solutions:

$$x^0=x^0(0),\quad x^1=x^1(0)e^{-\tau}+\tfrac{5}{2}Rx^3(0)\tau e^{-\tau},\quad x^2=x^2(0),\quad x^3=x^3(0)e^{-\tau},$$

$$y_1=y_1(0)e^{\tau},\quad y_2=y_2(0),\quad y_3=y_3(0)e^{\tau}-\tfrac{5}{2}Ry_1(0)\tau e^{\tau}.$$

Next we find the 1-graph space G such that $f|_G=0$. It is impossible to solve the equation $x^3\partial_3 g+x^1\partial_1 g=\frac{5}{2}Rx^3\partial_1 g$ for a unique g. Since the system under consideration is an ideal gas, by the equation of states $E=\frac{3}{2}NRT$ and $PV=NRT$, we have

$$T=\partial_1 g=\frac{2}{3x^3}\,g \quad \text{and} \quad -p=\partial_2 g=-\frac{2}{3x^2 g}.$$

Thus $g=Cx^3(x^2)^{-2/3}\exp(2x^1/3Rx^3)$. But

$$\mu=\partial_3 g=-\frac{1}{N}\,TS+\frac{5}{2}RT=\frac{5}{3}g-\frac{2}{3x^2}\,g.$$

Hence

$$E=x^0=g(x)=\alpha(x^3)^{5/3}(x^2)^{-(2/3)}\exp(2x^1/3Rx^3).$$

Set

$$E_0=x^0(0)=\alpha(x^3(0))^{5/3}(x^2(0))^{-(2/3)}\exp(2x^1(0)/3Rx^3(0)).$$

Then

$$E=\left(\frac{N}{N_0}\right)^{5/3}\left(\frac{V}{V_0}\right)^{-(2/3)}\exp\left(\frac{2}{3R}\left(\frac{S}{N}-\frac{S_0}{N_0}\right)\right)$$

or

$$S=\frac{N}{N_0}\,S_0+NR\,\ln\left\{\left(\frac{E}{E_0}\right)^{3/2}\left(\frac{V}{V_0}\right)\left(\frac{N}{N_0}\right)^{-(5/2)}\right\}$$

with $S_0=N_0R\,\ln\{E_0^{(3/2)}V_0N_0^{-(5/2)}\}$.

This example was originally proposed by Mrugala, Nulton, Schön, and Salamon.[3] Their solution is incomplete.

Example 4: In this example we consider a fluid undergoing a plane Couette flow between two parallel plates, where the lower plate is fixed and the upper plate is moving at a constant velocity u_0 in the positive x-axis.[13] The velocity vector has a gradient in the direction of the y-axis, while it has a translational symmetry in z. Thus this is a 1-dimensional problem. We assume that the fluid is incompressible and diathermic (i.e., nonheat conducting). Further we assume that both the temperature and the pressure of the system are constant. Under these assumptions, the conservative equations can be written as

$$\partial_t \rho = -\nabla \cdot (\rho \mathbf{u}) = 0,$$

$$\rho d_t \mathbf{u} = -\nabla p - \nabla \cdot \overleftrightarrow{\pi} = -\nabla \cdot \overleftrightarrow{\pi},$$

$$\rho d_t E = -\nabla \cdot \mathbf{Q} - \overleftrightarrow{P} : \nabla \mathbf{u} = 0,$$

where $\overleftrightarrow{P}$ is the stress tensor and $\overleftrightarrow{\pi}$ is the traceless symmetric part of $\overleftrightarrow{P}$. Since there is no heat flux nor viscous heating effect ($\overleftrightarrow{P}:\nabla \mathbf{u}=0$), E is constant. Consider the case where the system is near equilibrium. We can assume that $\boldsymbol{\pi}$ obeys the Maxwell equation

$$\rho d_t \hat{\pi} = -2p\gamma - (p/\eta_0)\overleftrightarrow{\pi},$$

with $\hat{\pi} = \rho^{-1}\overleftrightarrow{\pi}$, $\gamma = \frac{1}{2}\nabla \mathbf{u}$, and η_0 is the Newtonian viscosity of the fluid. By the geometry of the flow, the velocity vector $\mathbf{u}$ is simply $u_x(y)$ which is independent of x, and $\overleftrightarrow{\pi} = \pi_{xy}$. Denote $u = u_x$, $\pi = \pi_{xy}$. We have the following model equations for the plane Couette flow:

$$\partial_t u = -\rho^{-1}\partial_y \pi, \tag{41}$$

$$\partial_t \pi = -2p\gamma - p\eta_0^{-1}\pi, \tag{42}$$

with

$$\gamma = \frac{1}{2}\frac{\partial u_x}{\partial y}.$$

By eliminating π in (41) and (42) we obtain a wave equation of hyperbolic type,

$$\partial_{tt} u + 2k\partial_t u = c^2 \partial_{yy} u,$$

where $k = \frac{1}{2}(p/\eta_0)$ and $c^2 = pp^{-1}$. Next we set up the initial and boundary conditions as follows:

$$\begin{aligned} u(0,t) &= 0, \quad u(D,t) = u_0(t), \\ u(y,0) &= 0, \quad \pi(y,0) = 0. \end{aligned} \tag{43}$$

Here D is the distance between the two parallel plates. Since the fluid is viscous there is a transient period of time that $u_0(t)$ is smaller than u_0. Thus we assume $u_0(t) = u_0(1 - e^{-2kt})$. Equations (41)–(42) together with (43) now yield the following solutions for u and π,

$$u(y,t) = D^{-1}u_0 y A(t),$$

$$\pi(y,t) = \pi_0[(1+2kt)A(t) - 2kt],$$

where $A(t) = 1 - e^{-2kt}$, and $\pi_0 = -\eta_0 u_0 D^{-1}$. Notice that $u_0 D^{-1}$ is the shear rate, π_0 is the static shear stress and k^{-1} is the stress relaxation time.

For the problem of plane Couette flow there is only one thermodynamic variable π. Let X be the conjugate variable of π. As the system is near equilibrium we can approximate X by $X = -(2p)^{-1}\pi$.[14] In terms of normal coordinates (x, x^0, y) we set $\pi = x$, $X = y$, and $x^0 = g(x) = -(4p)^{-1}x^2$, with $S = T^{-1}x^0$ as the generalized entropy (calotropy) and $dS = T^{-1}dx$ as the generalized Gibbs relation.

Now (42) can be rewritten as

$$\frac{dx}{dt} = -2k\pi_0 A(t) - 2kx = f^1(x).$$

Thus we set

$$f^0(x)=\frac{dg}{dx}\ f^1(x)=kp^{-1}\pi_0 A(t)x+kp^{-1}x^2=kp^{-1}\pi_0 A(t)x-4kx^0,$$

$$f(x)=f^0(x)-yf^1(x)=-4kx^0+k\pi_0 A(t)(p^{-1}x+2y)+2kxy.$$

By (38) we have $f_1(x)=kp^{-1}\pi_0 A(t)-2ky$ and the vector field

$$X_f=f^0\partial_0+f^i\partial_i+f_i\partial^i,$$

where $\partial_0=(\partial/\partial x^0)$, $\partial_1=(\partial/\partial x)$, $\partial^1=(\partial/\partial y)$. The orbital equations of X_f are simply $(dx^0/dt)=f^0(x)$, $(dx/dt)=f^1(x)$, and $(dy/dt)=f_1(x)$. Hence we have embedded the dynamical equations of x in the flow equations of X_f. The flows of X_f form a 1-parameter group of contact transformations that preserve the Legendre submanifold $L=\{(x,x^0,y)|x^0=g(x),y=g'(x)\}$.

Finally we remark that Grmela and Jou have also applied the notion of contact geometry in their study of extended kinetic theory from a molecular point of view.[15] However their approach is different from the one presented in this paper.

[1] C. Caratheodory, Math. Ann. **67**, 355 (1905).
[2] R. Hermann, *Geometry, Physics and System* (Dekker, New York, 1973).
[3] R. Mrugala, J. D. Nulton, J. C. Schön, and P. Salamon, Rep. Math. Phys. **29**, 109 (1991).
[4] R. Clausius, Philos. Mag. **2**, 102 (1851); **24**, 201 (1862); Ann. Phys. (Leipzig) **125**, 313 (1865).
[5] (a) H. Grad, Commun. Pure Appl. Math. **2**, 331 (1949); *Handbuch der Physik*, edited by S. Flügge (Springer, Berlin, 1958), Vol. 12; (b) B. C. Eu, *Kinetic Theory and Irreversible Thermodynamics* (Wiley, New York, 1992).
[6] (a) D. Jou, N. Casas-Vasquez, and G. Lebon, *Extended Irreversible Thermodynamics* (Springer, Berlin, 1993); (b) I. Müller and T. Ruggeri, *Extended Thermodynamics* (Springer, Heidelberg, 1993).
[7] R. M. Velasco and L. S. Garcia-Colin, J. Stat. Phys. **69**, 217 (1992).
[8] See Ref. 6(a).
[9] Since dQ_d cannot be expressed in terms of x alone, it is not a 1-form in $\Lambda^1(B)$.
[10] (a) M. Chen, Phys. Rev. E **56**, 285 (1997); (b) J. Math. Phys. **38**, 3589 (1997), **38**, 5153 (1997).
[11] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Heidelberg, 1978).
[12] (a) See Ref. 11; (b) D. G. B. Edelen, *Applied Exterior Calculus* (Wiley, New York, 1985); (c) P. Liebermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Holland, 1987).
[13] See Ref. 5(b).
[14] See Ref. 5(b).
[15] M. Grmela and D. Jou, J. Math. Phys. **34**, 2290 (1993).

Information-theoretical derivation of an extended thermodynamical description of radiative systems[a)]

J. Fort[b)]
Secció de Fisica, Department d'Enginyeria Industrial, Escola Politècnica Superior, Universitat de Girona, Avda. Lluís Santaló s/n, 17071 Girona, Catalonia, Spain

J. E. Llebot
Grup de Física, Departament de Ciències Ambientals, Facultat de Ciències, Universitat de Girona, Plaça de l'Hospital 6, 17071 Girona, Catalonia, Spain

(Received 15 July 1997; accepted for publication 28 August 1997)

A radiative equation of the Cattaneo–Vernotte type is derived from information theory and the radiative transfer equation. The equation thus derived is a radiative analog of the equation that is used for the description of hyperbolic heat conduction. It is shown, without recourse to any phenomenological assumption, that radiative transfer may be included in a natural way in the framework of extended irreversible thermodynamics (EIT). © *1998 American Institute of Physics.* [S0022-2488(98)03001-1]

I. INTRODUCTION

Radiative transfer has recently become an active area of research within extended irreversible thermodynamics (EIT). However, most approaches to the subject have considered the description of purely radiation systems (photon gas),[1–3] whereas a joint thermodynamical description of radiation and matter is certainly necessary. This conclusion can be reached as follows. Because photons do not interact among themselves, it is the interaction between radiation and matter that drives an isolated radiative system towards equilibrium.[4] More generally, since in the absence of matter photons cross a small volume centered about any given point without any interaction, it is clear that the distribution function cannot be determined by the macroscopic properties at the point considered: instead, it will be determined by the properties of the emitting source. Similar arguments to the previous ones have led to the idea[5,6] that it does not seem possible to reach valid thermodynamical conclusions with regards to radiative transfer on the basis of any picture that does not take into account the absorption and emission of radiation by matter. Some phenomenological models that include matter as well as radiation have been proposed (see, e.g., Refs. 7 and 8). However, a microscopic approach should be sought in order to see to what extent the hypotheses that lie at the ground of such phenomenological models can be justified. In this paper, we make use of information theory in order to present such a microscopic approach. It will be very useful, before doing so, to briefly recall two information-theoretical results in conductive, purely matter systems. For an ideal monatomic gas, maximization of the entropy density under the constraints of fixed molecule number density n, vanishing barycentric velocity, fixed internal energy density ρu_m (ρ stands for the matter density), and fixed conductive heat flux $\mathbf{q}$ leads, keeping terms up to second-order in $\mathbf{q}$, to the following results[3] (see also Ref. 6) for the generalized Gibbs equation and for the temperature θ of extended irreversible thermodynamics

$$ds_m=\frac{1}{\theta}\,du_m-\frac{2m}{5nk_B^2T_m^3\rho}\,\mathbf{q}\,d\mathbf{q}, \tag{1}$$

[a)]Dedicated to Professor José Casas-Vázquez with occasion of his 60th anniversary and of the 25th anniversary of the first paper on Extended Irreversible Thermodynamics by his UAB group.
[b)]Electronic mail: jfort@songoku.udg.es

$$\frac{1}{\theta}=\frac{1}{T_m}\left(1+\frac{2\rho}{5n^3k_B{}^3T_m{}^3}\,\mathbf{q}\cdot\mathbf{q}\right), \tag{2}$$

where s_m is the specific entropy, m is the molecular mass (so that $\rho=mn$), k_B stands for the Boltzmann constant, T_m is the kinetic local-equilibrium temperature, which satisfies that[9,6]

$$u_m=3k_BT_m/2m, \tag{3}$$

and the definition of θ for the matter system under consideration is

$$\theta^{-1}\equiv\frac{\partial s_m}{\partial u_m}. \tag{4}$$

Equations (1) and (2) have also been derived without making use of information theory but beginning with the postulates of extended irreversible thermodynamics and making use of the results of Grad's 13-moment kinetic theory method.[9] This has provided a conceptually reasonable, sound framework for a thermodynamical description of heat conduction in further-away-from-equilibrium states than those corresponding to local-equilibrium thermodynamics. Our purpose is here to propose the information-theoretical basis of such a framework for radiative, instead of conductive, heat transfer. Therefore, in the following sections we do not consider a matter system under a conductive heat flux $\mathbf{q}$ but a radiation-matter system under a radiative heat flux, which we shall call $\mathbf{F}$.

II. INFORMATION THEORY

Assuming that the matter content of the system is a classical ideal monatomic gas, the entropy density of the system is[4]

$$\rho s=\rho s_m+\rho s_r=-k_B\int_{R^3}\frac{d^3p_m}{(2\pi\hbar)^3}\,f_m\,\ln\,f_m+2k_B\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,[(1+f_r)\ln(1+f_r)-f_r\,\ln\,f_r], \tag{5}$$

where the subindexes m and r stand for matter and radiation (photons), and s, f, and p are the corresponding entropies per unit mass, distribution functions, and momenta, respectively. Here $\hbar$ is the Planck constant h divided by 2π. We have used the same notation as in Ref. 4: for example, f_m stands for the number of molecules in d^3p_m and d^3x, divided by $d^3p_md^3x/(2\pi\hbar)^3$, with d^3x a differential of volume (in this way, $\ln f_m$ is dimensionless).

We will make use of the well-known information-theoretical approach[10,11] by maximizing (5) under the macroscopic constraints of fixed total energy density ρu, molecular number density n, and radiative energy flux $\mathbf{F}$ of the system,

$$\rho u=\rho u_m+\rho u_r=\int_{R^3}\frac{d^3p_m}{(2\pi\hbar)^3}\,\frac{p_m^2}{2m}\,f_m+2\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,p_rcf_3, \tag{6}$$

$$n=\int_{R^3}\frac{d^3p_m}{(2\pi\hbar)^3}\,f_m, \tag{7}$$

$$\mathbf{F}=2\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,p_rc\mathbf{c}f_r, \tag{8}$$

where u_m is the matter specific internal energy and u_r is the radiation energy, also per unit mass of matter. In Eq. (6) we have assumed that the barycentric velocity of the matter gas vanishes (this

allows us to concentrate on the topic that is of interest: we are neglecting heat convection in addition to heat conduction). In Eq. (8), $\mathbf{c}=c\hat{\Omega}$ is the velocity of a photon moving in a direction with unit vector $\hat{\Omega}$ and c is the speed of light *in vacuo*. Maximization of (5) under the constraints (6)–(8) finally yields for the distribution functions

$$f_m=\exp\left[-1-\lambda-\beta\,\frac{p_m^2}{2m}\right], \tag{9}$$

$$f_r=\frac{1}{\exp[\beta p_r c-\boldsymbol{\gamma}\cdot p_r c\,\mathbf{c}]-1}, \tag{10}$$

where λ, β, and $\boldsymbol{\gamma}$ are Lagrange multipliers.

III. EXTENDED IRREVERSIBLE THERMODYNAMICS OF RADIATIVE SYSTEMS

We now follow the usual information-theoretical procedure[9,12] in order to derive an extended Gibbs equation, which will be used to show that radiative transfer can be included in the framework of EIT. Since we have assumed the matter gas to be at rest, the equation of continuity implies that matter density does not change in time, i.e., the density differential vanishes, so that $dn=0$. From this fact and Eq. (5) it is not difficult to find out expressions for ds_m and ds_r, in which we make use of (9) and (10), respectively. In this way and after use of (6)–(8) we arrive at $ds_m=k_B\beta\,du_m$ and $ds_r=k_B\beta\,du_r-(k_B/\rho)\boldsymbol{\gamma}\cdot d\mathbf{F}$. Therefore

$$ds=ds_m+ds_r=k_B\beta\,du-\frac{k_B}{\rho}\,\boldsymbol{\gamma}\cdot d\mathbf{F}, \tag{11}$$

where we note that the entropy per unit mass of the radiation-matter system under consideration is a function of the *total* (radiation and matter) specific energy u and of the radiative heat flux $\mathbf{F}$. Because of this information-theoretical result, we propose the following generalization of the definition (4),

$$\theta^{-1}\equiv\frac{\partial s}{\partial u}. \tag{12}$$

From Eqs. (12) and (11) we have

$$\theta=\frac{1}{k_B\beta}. \tag{13}$$

On the other hand, it will be useful to follow Mihalas and Mihalas[13] by introducing a radiative *parameter* T_r through the following equation;

$$\rho u_r=aT_r^4, \tag{14}$$

with $a=\pi^2k_B^4/15c^3\hbar^3$ the blackbody constant.

After substitution of the matter distribution function (9) into (7) and into ρu_m, given by (6), and integration over all possible values of $\mathbf{p}_m$, it is easy to reach the result $\rho u_m=3n/2\beta$, which together with (3) implies that $T_m=1/k_B\beta$. Thus we have, taking into account Eq. (13),

$$\theta=T_m, \tag{15}$$

and this result is consistent with Eq. (2), because we are not dealing with conductive but with radiative heat transfer (i.e., we have $\mathbf{q}=0$ although $\mathbf{F}\neq 0$ in general).

We now look for a relationship between θ and T_r. We make use of the radiation distribution function (10) into (8) and into ρu_r, given by (6). Since both integrations over $\mathbf{p}_r$ are rather complicated, we assume that $\boldsymbol{\gamma}=(0,0,\gamma)$, integrate with the use of formula 3.411-1 of Ref. 14 and obtain

$$\rho u_r=\frac{a}{{k_B}^4\beta^4}\,\frac{1+\epsilon^2/3}{(1-\epsilon^2)^3}, \tag{16}$$

$$\mathbf{F}=\frac{4ca}{3{k_B}^4\beta^4}\,\frac{\boldsymbol{\epsilon}}{(1-\epsilon^2)^3}\equiv(0,0,F), \tag{17}$$

with

$$\boldsymbol{\epsilon}\equiv\frac{c}{\beta}\,\boldsymbol{\gamma}\equiv(0,0,\epsilon). \tag{18}$$

From Eqs. (14), (16), and (13) we obtain

$$\theta=T_r\left(\frac{(1-\epsilon^2)^3}{1+\epsilon^2/3}\right)^{1/4}. \tag{19}$$

This equation relates θ to T_r, but ϵ should be written in terms of macroscopic observables if we want the radiative theory to be put at the same level as the conductive one: according to the conductive relationship (2), for a given value of m (i.e., for a given matter gas) θ is a known function of T_m if the values of the macroscopic observables ρ and $\mathbf{q}$ are known. We can cope with this problem by noting that (16), (17), and (14) can be combined to yield

$$F=4caT_r{}^4\,\frac{\epsilon}{3+\epsilon^2}, \tag{20}$$

from which it follows that

$$\epsilon=\frac{2-\sqrt{4-3F^2/c^2a^2T_r{}^8}}{F/caT_r^4}. \tag{21}$$

We make use of (20) in the denominator in (19) and then apply Eq. (21). Since we will here be interested in near-equilibrium results, we assume that the value of F is low enough so that θ can be approximated by its MacLaurin expansion up to second order in F. This yields

$$\frac{1}{\theta}=\frac{1}{T_r}\left(1+\frac{15}{32c^2a^2T_r{}^8}\,F^2\right), \tag{22}$$

and this equation, which applies to radiative transfer, is analogous to the conductive equation (2). The second-order result (22) is in agreement with the expression for θ that has been obtained [Eqs. (27) and (17) in Ref. 3] by partial derivation of the radiative entropy density (which was derived on phenomenological grounds in Refs. 1 and 15) with respect to the radiative energy density. Equation (22) has also been previously derived by Casas-Vázquez and Jou for a purely radiation system [see Eq. (30) in Ref. 16], by making use of a fluctuation theory result and of the EIT phenomenology. However, there is a very important difference between these approaches and the information-theoretical one presented here: since in our definition (12) for θ we have made use of the total radiation *and matter* entropy and energy densities, the conceptual objections presented in

the first lines in this letter do not apply to our analysis. A generalized Gibbs equation can be finally derived from (11), (18), (13), and (21). Keeping in mind that we are interested in the second-order approximation for θ and s, we obtain

$$ds=\frac{1}{\theta}\,du-\frac{3}{4c^2a\rho T_r^5}\,\mathbf{F}\cdot d\mathbf{F}, \tag{23}$$

which is analogous to the conductive result (1). On the other hand, we note from (3), (15), and (22) that u_m is a function of T_r [or u_r, see Eq. (14)] and $\mathbf{F}$. Therefore, although in Eq. (23) we have taken the total specific energy $u=u_m+u_r$ and $\mathbf{F}$ as independent variables, it is clear that in this model u_r and $\mathbf{F}$ may also be chosen as independent variables. It is easy to relate du_m to du_r and $d\mathbf{F}$. Making use of the relationship thus obtained and of (23), applying (22), and neglecting higher-order terms, one finally obtains the extended Gibbs equation in the new variables,

$$ds=\left(\frac{1}{\theta}\left\{1+\frac{3k_B\rho^{1/4}}{8\,ma^{1/4}u_r^{3/4}}\right\}+\frac{315k_BF^2}{256\,m\rho^2c^2u_r^3}\right)du_r-\left(\frac{45k_B}{32mc^2\rho^2u_r^2}+\frac{3a^{1/4}}{4c^2\rho^{9/4}u_r^{3/4}}\right)\mathbf{F}\cdot d\mathbf{F}. \tag{24}$$

From this equation, (22), and (14), it is straightforward to check that ds is an exact differential, i.e.,

$$\frac{\partial^2 s}{\partial\mathbf{F}\partial u_r}=\frac{\partial^2 s}{\partial u_r\partial\mathbf{F}}, \tag{25}$$

so that the total (i.e., radiation and matter) specific entropy of the system is a state function, as it should.

IV. APPLICATION

Kremer and Müller[1] were able to deduce a radiative equation which is essentially of the Cattaneo-Vernotte relaxational type [their second Eq. (7.1)], but the corresponding relaxation time and radiative conductivity were not found out. We will now provide a microscopic derivation of that relaxational equation and find out the relaxation time and radiative conductivity explicitly.

Let us mention, for the sake of completeness, that making use of Eq. (7) it is not difficult to show that the matter distribution function (9) is locally Maxwellian, with temperature $T_m=\theta$ [see Eq. (15)]. On the other hand, making use of Eqs. (18), (21), (13), and (22), the information-theoretical radiative distribution function (10) reads, up to second order in $\mathbf{F}$,

$$f_r=f_r^{(0)}(1+\phi^{(1)}+\phi^{(2)}+\cdots), \tag{26}$$

with

$$f_r^{(0)}=\frac{1}{e^x-1}, \tag{27}$$

$$\phi^{(1)}=\frac{3x}{4}\,\frac{e^x}{e^x-1}\,\hat{\Omega}\cdot\mathbf{f}, \tag{28}$$

$$\phi^{(2)}=\tilde{\phi}^{(2)}+\tilde{\tilde{\phi}}^{(2)}, \tag{29}$$

$$\tilde{\phi}^{(2)}=-\frac{15x}{32}\,\frac{e^x}{e^x-1}\,\mathbf{f}\cdot\mathbf{f}, \tag{30}$$

$$\tilde{\tilde{\phi}}^{(2)}=\frac{9x^2}{32}\,\frac{e^x(e^x+1)}{(e^x-1)^2}\,(\hat{\Omega}\cdot\mathbf{f})^2, \tag{31}$$

where we have defined the following dimensionless quantities:

$$\mathbf{x}\equiv\frac{c}{k_B T_r}\,\mathbf{p}_r\,,\quad x\equiv|\mathbf{x}|,\quad \hat{\Omega}\equiv\frac{\mathbf{c}}{c}=\frac{\mathbf{p}_r}{p_r}=\frac{\mathbf{x}}{x},\quad \mathbf{f}\equiv\frac{\mathbf{F}}{caT_r^4}=\frac{\mathbf{F}}{cu_r\rho}, \tag{32}$$

f being the reduced flux. We have written (27)–(31) in terms of T_r, rather than in terms of $\theta=T_m$, for later convenience. The distribution function (26)–(31) is a radiative analog of the Dominguez–Jou distribution function [Eqs. (30), (33), and (34) in Ref. 3], which was also derived making use of information theory but applies to a nonrelativistic classical monatomic ideal gas under a conductive heat flux. Both in radiative and conductive systems (present paper and Ref. 3, respectively), information theory provides a simple derivation for the second-order non-equilibrium correction $\phi^{(2)}$. This correction is very difficult to find out in the framework of the kinetic theory of gases[17,18] of conductive, purely matter systems. It is worthwhile to mention that information theory is very simple mathematically but has the disadvantage that the question of what constraints should be imposed in maximizing the entropy density is an open one at present. On the other hand, it has been stressed[5] that a well-established phenomenological thermodynamics of radiative transfer does not seem to be available. This is in contrast with the situation in purely matter systems, where phenomenological thermodynamics may be a useful guide in the choice of the information-theoretical constraints (see Refs. 6, 12, 19, and, in connection with this, pages 46–47 in Ref. 9).

Our purpose here is to show that a relaxational equation follows from the radiative distribution function. In order to do so, we shall first show that although both (22) and (23) on one hand, and (26)–(31) on the other, have been obtained working up to second order in **F**, only the first nonequilibrium correction $\phi^{(1)}$ contributes to the second-order entropy corresponding to (22) and (23). Making use of Eq. (26) and of the MacLaurin expansion

$$\ln(a_0+a_1F+a_2F^2+\cdots)=\ln a_0+\frac{a_1}{a_0}F+\left(\frac{2a_2}{a_0}-\frac{a_1^2}{a_0^2}\right)\frac{F^2}{2}+O(F^3)$$

into the expression for ρs_r, given by Eq. (1), we obtain, up to second order,

$$\begin{aligned}\rho s_r=\rho s_r^{(0)}&+2k_B\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,f_r^{(0)}\,\ln\left(\frac{1}{f_r^{(0)}}+1\right)\phi^{(1)}\\&-2k_B\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,f_r^{(0)}\,\ln\left(\frac{1}{f_r^{(0)}}+1\right)\phi^{(2)}\\&-k_B\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,\frac{f_r^{(0)}}{1+f_r^{(0)}}\,\phi^{(1)^2},\end{aligned} \tag{33}$$

with

$$\rho s_r^{(0)}=2k_B\int_{R^3}\frac{d^3p_r}{(2\pi\hbar)^3}\,[(1+f_r^{(0)})\ln(1+f_r^{(0)})-f_r^{(0)}\,\ln\,f_r^{(0)}].$$

The second term on the right-hand side of (33) vanishes because (27) and (28) imply that the corresponding integrand changes sign under the transformation $\mathbf{p}_r\to-\mathbf{p}_r$. The third term, which contains $\phi^{(2)}$, is also seen to vanish after substitution of (27) and (29)–(31), application of the integral theorem (which is easy to prove)

$$\int_{R^3} d^3x\ F(x)(\mathbf{x}\cdot\mathbf{g})^2 = \frac{|\mathbf{g}|^2}{3} \int_{R^3} d^3x\ x^2 F(x)$$

[with $\mathbf{g}$ any vector and $F(x)$ an arbitrary function], and use of the integrals

$$\int_0^\infty dx\ \frac{x^4 e^x}{(e^x-1)^2} = \frac{4\pi^4}{15}, \quad \int_0^\infty dx\ \frac{x^5 e^{2x}}{(e^x-1)^3} = \frac{2\pi^4}{3} + 60\zeta(5)$$

and

$$\int_0^\infty dx\ \frac{x^5 e^x}{(e^x-1)^3} = \frac{2\pi^4}{3} - 60\zeta(5),$$

$\zeta(z)$ being the Riemann zeta function. The vanishing of the second and third terms on the right-hand side of the radiative entropy density (33) can be regarded as the radiative analogue of the (e.g., kinetic-theoretical) result[17,9] for purely matter systems

$$\int_{R^3} \frac{d^3p_m}{(2\pi\hbar)^3}\ f_m^{(0)} \phi_m^{(k)} = 0,$$

with $f_m^{(0)}$ the Maxwell–Boltzmann distribution function and $\phi_m^{(k)}$ the kth-order correction due to heat conduction and/or convection: this leads to[9]

$$\rho s_m = \rho s_m^{(0)} - \frac{k_B}{2} \int_{R^3} \frac{d^3p_m}{(2\pi\hbar)^3}\ f_m^{(0)} \phi_m^{(1)^2},$$

so that $\phi_m^{(2)}$ does not contribute to the second-order approximation to the matter entropy density in conductive and/or convective situations, just as (33) becomes

$$\rho s_r = \rho s_r^{(0)} - k_B \int_{R^3} \frac{d^3p_r}{(2\pi\hbar)^3} \frac{f_r^{(0)}}{1+f_r^{(0)}}\ \phi^{(1)^2},$$

so that $\phi^{(2)}$ does not contribute to the second-order radiation entropy density in radiative situations. We mention that this conclusion can also be reached, after rather tedious calculations, by checking that integration of the first and fourth terms on the right-hand side of (33) yields the same result for ρs_r as that obtained by substitution of (10) into ρs_r, given by (5), integration, and expansion of the result up to second order in $\mathbf{F}$. By following either of the procedures it is obtained that

$$\rho s_r = \rho s_r^{(0)} - \frac{3}{8ac^2T_r^5} F^2 + O(F^3), \tag{34}$$

with $\rho s_r^{(0)} = 4aT_r^3/3$.

In the special case of thermodynamic equilibrium ($\mathbf{F}=0$), Eq. (21) becomes $\epsilon = 0$, (15) and (19) yield $\theta = T_m = T_r$, and (26)–(31) reduce to the Planck distribution function, $f_r = f_r^{(0)}$, as they should. In the same case, (34) becomes the usual expression for the radiation entropy density in equilibrium, i.e.,[4] $\rho s_r = \rho s_r^{(0)} = 4aT_r^3/3$, and we have, instead of (23), that $ds = (1/T_r)du$. This is a Gibbs equation. Therefore, analogously to what is done in phenomenological thermodynamics of matter systems,[9] T_r may be called the local-equilibrium temperature and $s_r^{(0)} \equiv s_{rle}$ may be called the local-equilibrium radiative entropy.

We now turn our attention to the second-order theory. In order to verify its consistency, it is rather important to note that Eq. (34) can also be obtained very easily (although without finding $\rho s_r^{(0)}$ explicitly) from (23), (22), and the radiative part of the local Gibbs equation, i.e., $ds_{rle} = (1/T_r)du_r$, in the following way. From $ds_m = (1/\theta)du_m$ [see the text above Eq. (11) and Eq. (13)] and (23) we have $ds_r = (1/\theta)du_r - (3/4c^2 a\rho T_r^5)\mathbf{F}\cdot d\mathbf{F}$. By integration with the use of (22) and (14) we find the second-order result $s_r = s_{rle} - (3/8ac^2\rho T_r^5)F^2$, in agreement with Eq. (34).

The fact, derived above Eq. (34), that $\phi^{(1)}$ contributes to the second-order radiation specific entropy, whereas $\phi^{(2)}$ does not, implies that the second-order thermodynamical results (22) and (23) correspond to the first-order distribution function, i.e., $f_r \approx f_r^{(0)}(1+\phi^{(1)})$ (we stress that the same happens in EIT of purely matter systems, see pp. 83–87 in Ref. 9). From this result, we will finally be able to find an evolution equation.

It is well known that information theory on its own cannot yield evolution equations, simply because all it yields is an expression for the distribution function and an (extended) Gibbs equation. However, the information-theoretical distribution function can be[20] and has in fact been combined with dynamical methods, such as the Boltzmann equation,[21] the Liouville equation,[22] the statistical operator method,[23] or the radiative transfer equation.[24] As it is well known,[25,8], multiplication of the radiative transfer equation by $\hat{\Omega}$ and integration yields, under the grey approximation,

$$\frac{1}{c}\frac{\partial \mathbf{F}}{\partial t} + c\boldsymbol{\nabla}\cdot\mathbf{P}_r = -\sigma_a \mathbf{F}, \tag{35}$$

with σ_a the absorption coefficient, $[\boldsymbol{\nabla}\cdot\mathbf{P}_r]_i \equiv \Sigma_{k=1}^3 \partial P_{r\ ki}/\partial x_k$ and $P_{r\ ki}$ the components of the pressure tensor of radiation, namely,

$$P_{r\ ki} = \frac{2}{c}\int_{R^3} \frac{d^3 p_r}{(2\pi\hbar)^3}\, p_r c_k c_i f_r . \tag{36}$$

Substitution of $f_r \approx f_r^{(0)}(1+\phi^{(1)})$ in this expression and use of Eqs. (27), (28), and (32) shows that $\phi^{(1)}$ does not contribute to the pressure tensor. Use of formula (3.411.1) in Ref. 14 allows us to perform the remaining integration. This yields

$$P_{r\ ki} = \frac{a}{3} T_r^4 \delta_{ki}, \tag{37}$$

with $\delta_{ki} = 1$ for $k = i$ and $\delta_{ki} = 0$ for $k \neq i$. Use of (37) in Eq. (35) yields

$$\tau\frac{\partial \mathbf{F}}{\partial t} + \mathbf{F} = -\lambda \boldsymbol{\nabla} T_r, \tag{38}$$

with

$$\tau = \frac{1}{c\sigma_a}, \quad \lambda = \frac{4ca}{3\sigma_a} T_r^3. \tag{39}$$

Equation (38) is in agreement with a result previously derived in Ref. 1, but here it has been possible to introduce $\boldsymbol{\nabla} T_r$ instead of the gradient of the radiative energy density, and also to find out the explicit expressions (39) for the radiative relaxation time τ and the radiative conductivity λ. It is encouraging that in steady states Eqs. (38) and (39) become $\mathbf{F} = -(4ca/3\sigma_a)T_r^3 \boldsymbol{\nabla} T_r$, and this Fourier-type equation has been used since many years ago[26] in the theory of stellar interiors. We have seen that information theory provides a statistical-mechanical derivation of the more general results (38) and (39). Similar, less rigorous expressions have been recently advanced on

phenomenological grounds.[8] Here, information theory has also allowed us to distinguish and relate T_m to T_r [defined by Eqs. (3) and (14), respectively] through (22) and (15), i.e.,

$$\frac{1}{T_m}=\frac{1}{T_r}\left(1+\frac{15}{32c^2a^2T_r^8}F^2\right).$$

We stress that this result is valid under the same conditions as those under which (38) and (39) hold.

From the generalized Gibbs equation (23), the energy balance law, namely[27] $\rho\partial u/\partial t=\nabla\cdot\mathbf{F}$, and Eqs. (38) and (39) it is not difficult to follow the usual procedure[9] and identify the entropy production rate, for low enough values of the flux $\mathbf{F}$, as

$$\sigma^s=\frac{3\sigma_a}{4caT_r^5}\mathbf{F}\cdot\mathbf{F}, \tag{40}$$

which is semi-positive definite, in agreement with the second law. According to (39) and (40), we may write $\sigma^s=(1/\lambda T_r^2)\mathbf{F}\cdot\mathbf{F}$, which is analogous to the conductive result[9] $\sigma^s=(1/\lambda' T_m^2)\mathbf{q}\cdot\mathbf{q}$, with λ' the thermal conductivity. Equation (40) shows that in the absence of matter ($\sigma_a=0$) there is no entropy production. This corresponds to the fact that photons do not interact among themselves. Because radiative transfer in the vacuum in absolutely nonlocal, it has been previously argued[6,5] that the entropy production must arise from the interaction of radiation with matter, although the explicit formula (40) does not seem to have been derived or proposed before.

We may also note that Eq. (38) is of the Cattaneo–Vemotte type: because of conceptual motivations (see, e.g., Ref. 9), and also because Fourier's law of heat conduction, i.e., $\mathbf{q}=-\lambda'\nabla T_m$, is known to fail in very fast processes,[28] a more general law, similar to (38), for heat conduction (instead of heat radiation), namely

$$\tau'\frac{\partial\mathbf{q}}{\partial t}+\mathbf{q}=-\lambda'\nabla T_m, \tag{41}$$

with τ' the conductive relaxation time, has been considered since long ago.[29,30] Equation (41) was a fundamental starting point[31] for the development of the theory presently called extended irreversible thermodynamics (see, e.g., Refs. 9 and 20). The similarity between Eqs. (41) and (38), in itself, indicates the reasonability of including radiative transfer in the framework of EIT.

Still from another point of view, we may compare our results with previous proposals (see, e.g., Ref. 7) by saying that it seems now possible to approach the irreversible thermodynamics of radiative transfer without need, in principle, to make use of a temperature field that depends on frequency and direction. This is analogous to the fact that in nonequilibrium thermodynamics of heat conduction and convection[17,9] there is no need to introduce a temperature field that depends on the molecular speed (or energy) and direction of motion.

Just to summarize, information theory provides a nonequilibrium extension of Planck's distribution function [Eqs. (26)–(31)], a microscopic derivation of a radiative equation of the Cattaneo–Vemotte type [Eqs. (38) and (39)], and it also leads to the conclusion that the range of nonequilibrium phenomena for which EIT provides an adequate framework is enlarged with the inclusion of radiative transfer.

ACKNOWLEDGMENTS

The authors would like to thank Professor D. Jou for stimulating discussions and Professor A. S. Cukrowski for providing us with the integral theorem used under Eq. (33) and with Ref. 18. This work has been partially funded by the Comisión Interministerial de Ciencia y Tecnología of the Spanish government under grant No. CL195-1867.

[1] G. M. Kremer and I. Müller, J. Math. Phys. **33**, 2265 (1992).
[2] M. Ferrer and D. Jou, Am. J. Phys. **63**, 237 (1995), see especially the Appendix.
[3] R. Dominguez and D. Jou, Phys. Rev. E **51**, 158 (1995).
[4] L. D. Landau and E. M. Lifshitz, *Statistical Physics. Part I* (Pergamon, Oxford, 1980).
[5] C. Essex, Adv. Thermodyn. **3**, 435 (1990).
[6] R. E. Nettleton, Phys. Rev. E **53**, 1241 (1996).
[7] H. F. Albert, J. Non-Equilib. Thermodyn. **16**, 303 (1991).
[8] J. Fort and J. E. Llebot, J. Phys. A **29**, 3427 (1996).
[9] D. Jou, J. Casas-Vázquez, and G. Lebon, *Extended Irreversible Thermodynamics* (Springer, Berlin, 1996).
[10] E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957); *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz (Reidel, Dordrecht, 1983).
[11] F. Schlögl, Phys. Rep. **62**, 267 (1980).
[12] R. E. Nettleton, J. Chem. Phys. **93**, 8247 (1990).
[13] D. Mihalas and B. W. Mihalas, *Foundations of Radiation Hydrodynamics* (Oxford U.P., Oxford, 1984).
[14] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1984).
[15] A. M. Anile, S. Pennisi, and M. Sammartino, J. Math. Phys. **32**, 544 (1991).
[16] J. Casas-Vázquez and D. Jou, Phys. Rev. E **49**, 1040 (1994).
[17] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases* (Cambridge U.P., Cambridge, 1970).
[18] M. S. Shavalieff, Prikl. Mat. Mekh. **42**, 656 (1978) (in Russian).
[19] R. E. Nettleton, Phys. Rev. A **42**, 4622 (1990).
[20] R. E. Nettleton and S. Sobolev, J. Non-Equilib. Thermodyn. **20**, 205, 297 (1995); **21**, 1 (1996).
[21] J. Fort and A. S. Cukrowski, Chem. Phys. **222**, 59 (1997).
[22] B. Robertson, Phys. Rev. **144**, 15 (1966); **160**, 175 (1967).
[23] A. R. Vasconcellos, R. Luzzi, and L. S. Garcia-Colín, Phys. Rev. A **43**, 6622, 6633 (1991).
[24] J. Fort, Physica A **243**, 275 (1997).
[25] G. C. Pomraning, J. Quant. Spectrosc. Radiat. Transf. **27**, 517 (1982).
[26] A. S. Eddington, *The Internal Constitution of the Stars* (Cambridge U.P., Cambridge, 1988), p. 101.
[27] S. R. De Groot and P. Mazur, *Non-Equilibrium Thermodynamics* (Dover, New York, 1984).
[28] See, e.g., D. D. Joseph and L. Preziosi, Rev. Mod. Phys. **61**, 41 (1989); **62**, 375 (1990).
[29] C. Cattaneo, Atti Sem. Mat. Fis. Univ. Modena **3**, 33 (1948).
[30] P. Vernotte, C. R. Acad. Sci. Paris **246**, 3154 (1958).
[31] R. E. Nettleton, Phys. Fluids **3**, 216 (1960).

Conditions of Cayley's type for ellipsoidal billiard

Vladimir Dragović[a] and Milena Radnović
Mathematical institute SANU Kneza Mihaila 35, 11000 Beograd, Yugoslavia

(Received 5 March 1997; accepted for publication 15 August 1997)

The billiard system within an ellipsoid in three-dimensional space is considered. We describe periodical trajectories of that mechanical system deriving analytical condition of Cayley's type for generalized Poncelet's theorem, using methods of algebrogeometric integration. © *1998 American Institute of Physics.*
[S0022-2488(98)02001-5]

I. INTRODUCTION

The 19th century classics already understood a deep connection between the projective geometry and the theory of algebraic curves and algebraic functions. The full Poncelet's theorem is one of the clearest examples. It claims that, as well known, *if two conics are given in a plane, then some polygonal line inscribed in one and circumscribed about the other conic is closed after n steps if and only if any such a line is closed after the same number of steps*. Connection of that purely geometrical claim with the theory of elliptic functions was discovered by Jacobi, and especially Cayley. Introducing the notion of Cayley's cubic and applying interesting algebraic geometry, Cayley deduced analytical condition on two conics for the existence of a polygon with n sides inscribed in one and circumscribed about the other conic. Detailed proof of the Cayley's condition can be found in Lebesgue's book *Les Coniques*.[1]

This subject matter also gets the attention of outstanding mathematicians nowadays. Griffiths and Harris have relatively soon given, in a series of papers,[2,3] algebrogeometric proofs in a classical manner of both Poncelet's theorem and Cayley's condition, reducing the problem to the question if some point on an elliptical curve is of finite order. The curve they consider consists of pairs $(x,\xi)\in C\times D^*$, $x\in\xi$. C,D are the conics, and D^* is the dual conic—it consists of tangent lines to D. This curve differs from Cayley's cubic. Griffiths and Harris proposed an interesting generalization of Poncelet's theorem.[2]

Except geometrical, the full Poncelet's theorem has certain significance in mechanics too, especially in the theory of integrable systems. It is well known that the billiard inside an ellipse represents an integrable mechanical system.[4] The integrability is shown exactly in existence, for a given trajectory, of a conic confocal with the boundary, such that the trajectory is tangent to the conic between every two bounces.

It is quite natural to generalize the billiard system in an ellipsoid in R^n. Trajectory is then tangent to $n-1$ quadrics confocal with the boundary.[5,6] (Trajectories of the systems perturbed by potentials from Ref. 7 have the same property, too.) The corresponding generalization of Poncelet's theorem is proved in Refs. 5, 6: *if there is a closed trajectory of the billiard inside one and tangent to $n-1$ given confocal quadrics in R^n, then there are infinitely many such trajectories.*

Strong penetration of algebraic geometry in the theory of integrable systems dates from the mid-1970s, from works by Novikov, Dubrovin, Krichever, Mumford, Van Moerbeck,... . Ideas of finite zone integration are applied to the billiard system, as a system with discrete time, in papers of Veselov and Moser.[8–10]

In this paper, a generalization of Cayley's condition connected with the generalized Poncelet's theorem from Refs. 5, 6 is established. Direct application of Cayley's or Griffiths' and Harris' proof is not fruitful from our point of view. Some results and problems in this way are described

[a]Electronic mail: vladad@mi.sanu.ac.yu

in Ref. 11. So, we used some of Grifitths' and Harris' ideas, modified for hyperelliptic curves, and applied them to so-called *isospectral curves* of Veselov and Moser.[8–10] Derived conditions look like classical Cayley's formulas by their form. We should remark that, even in the planar case, the conditions obtained in this paper are not identical to those of Cayley's, although they are equivalent.

II. XYZ MODEL AND ISOSPECTRAL CURVES

In this section, following the Veselov and Moser paper,[8] the billiard will be considered as a system with discrete time, and its integration will be shown.

Let an ellipsoid in R^N be given by the equation

$$(Ax,x)=1.$$

(We can assume that A is a diagonal matrix, with all elements on the diagonal different.) The billiard inside that ellipsoid is determined by the equations

$$x_{k+1}-x_k=\mu_k y_{k+1}, \quad y_{k+1}-y_k=\nu_k A x_k,$$

with

$$\mu_k=-\frac{2(Ay_{k+1},x_k)}{(Ay_{k+1},y_{k+1})}, \quad \nu_k=-\frac{2(Ax_k,y_k)}{(Ax_k,Ax_k)}.$$

The sequence x_k represents the points of consecutive bounces of the billiard, and $y_k=(x_k-x_{k-1})/|x_k-x_{k-1}|$ are the momenta.

The Heisenberg XYZ model can be joined to the billiard system with discrete time, in the way given by Veselov and Moser in Ref. 8, which is going to be shown here.

Consider the mapping ϕ: $(x,y)\to(x',y')$,

$$x'_k=Jy_{k+1}=J(y_k+\nu_k A x_k), \quad y'_k=-J^{-1}x_k, \quad J=A^{-1/2}.$$

Notice that the dynamics of mapping ϕ contains the billiard dynamics:

$$x''_k=Jy'_{k+1}=-x_{k+1}, \quad y''_k=-J^{-1}x'_k=-y_{k+1}.$$

Define the sequence $(\overline{x_k},\overline{y_k})$ by

$$(\overline{x_0},\overline{y_0})=(x_0,y_0), \quad (\overline{x_{k+1}},\overline{y_{k+1}})=\phi(\overline{x_k},\overline{y_k}).$$

This sequence satisfies the following relations:

$$\overline{x_{k+1}}=J\overline{y_k}+\nu_k J^{-1}\overline{x_k}, \quad \overline{y_{k+1}}=-J^{-1}\overline{x_k},$$

where the multiplier v_k is such that $|\overline{y_k}|=1$, $(A\overline{x_k},\overline{x_k})=1$. These relations can be written in the following way:

$$\overline{x_{k=1}}+\overline{x_{k-1}}=\nu_k J^{-1}\overline{x_k}.$$

Let $q_k=J^{-1}\overline{x_k}$. We have

$$q_{k+1}+q_{k-1}=\nu_k J^{-1}q_k, \quad |q_k|=1.$$

These equations represent the equations of a discrete Heisenberg XYZ system.

Claim.[8] *Let* $(\overline{x_k},\overline{y_k})$ *be the sequence connected with the elliptical billiard in the way described now. Then* $q_k=J^{-1}\overline{x_k}$ *is the solution of the discrete Heisenberg system.*

Conversely, if qk *is the solution of Heisenberg system, then the sequence* $x_k=(-1)^k J_{q2k}$ *represents a trajectory of a discrete ellipsoidal billiard system.*

According to the usual scheme of algebrogeometric integration, the algebraic curve, on whose Jacobian variety the flow becomes linear, is found in Ref. 8.

The equations of the discrete XYZ model are equivalent to the isospectral deformation (see Ref. 8):

$$L_{k+1}(\lambda)=A_k(\lambda)L_k(\lambda)A_k^{-1}(\lambda),$$

where

$$L_k(\lambda)=J^2+\lambda q_{k-1}\wedge Jq_k-\lambda^2 q_{k-1}\otimes q_{k-1},$$

$$A_k(\lambda)=J-\lambda q_k\otimes q_{k-1}.$$

As all the matrices $L_k(\lambda)$ are similar to each other, they all have the same *spectral curve* Γ:

$$\det(L(\lambda)-\mu I)=0.$$

The equation of that curve is

$$\nu^2=\prod_{i=1}^{N-1}(\mu-\mu_i)\prod_{j=1}^{N}(\mu-J_j^2),$$

with $v=\lambda\Pi_{i=1}^{N-1}(\mu-\mu_i)$, $J=\mathrm{diag}(J_1,...,J_N)$, and $\mu_1,\ldots,\mu_{N-1}$ are the roots of the function

$$\phi_\mu(x,Jy)=\sum_{i=1}^{N}\frac{F_i(x,y)}{\mu-J_i^2},$$

$$F_i=x_i^2+\sum_{j\neq i}\frac{(x\wedge Jy)_{ij}^2}{J_i^2-J_j^2},\quad x=q_{k-1},\quad y=q_k.$$

Knowing the eigenvectors ψ_k of the matrices $L_k(\lambda)$, it is possible to determine uniquely the members of the sequence q_k. If D_k is the pole divisor of the function ψ_k on the curve Γ, we have[8]

$$D_{k+1}=D_k+P_\infty-P_0,$$

where the point P_∞ corresponds to the value $\mu=\infty$, and P_0 to $\mu=0$, $\lambda=(q_k,J^{-1}q_{k+1})^{-1}$.

III. CHARACTERIZATION OF PERIODICAL BILLIARD TRAJECTORIES

In this section, we are going to establish a connection between periodic billiard trajectories and points of finite order on the Jacobian variety of a corresponding hyperelliptic curve.

The next claim gives the connection between periodic sequences q_k and periodic divisors D_k.

Lemma 1: The sequence of divisors D_k *is* n *periodic if and only if the sequence* q_k *is also periodic with the period* n *or* $q_{k+n}=-q_k$ *for all* k.

Proof: If it is, for all k, $q_{k+n}=q_k$, or for all k it is $q_{k+n}=-q_k$, obviously we have $L_{k+n}=L_k$. Thus the sequence of the eigenvectors ψ_k is periodical with the period n. It follows that the sequence of the divisors is also periodic.

Suppose now $D_{k+n}=D_k$ for all k. Thus $\psi_{k+n}=C_k\psi_k$. We have (see Ref. 8)

$$\psi_{k+1}=A_k(\lambda)\psi_k.$$

Let μ_1 and μ_2 be the values of parameter μ that correspond to $\lambda=1$ on the curve Γ, and

$$\Psi_k=(\psi_k(1,\mu_1),\psi_k(1,\mu_2)).$$

From $\psi_{k+1}=A_k(\lambda)\psi_k$, we obtain $A_k(1)=\Psi_{k+1}\Psi_k^{-1}$. It follows that

$$A_k(1)=\frac{C_{k+1}}{C_k}A_{k+n}(1).$$

From the condition on A_k for all k (see Ref. 8), we get $C_k=C_{k+1}$. Thus, the sequence

$$A_k(1)=J-q_k\otimes q_{k-1}$$

is n periodic. Therefore

$$q_{k+n}=\alpha_k q_k,\quad q_{k+n-1}=\frac{1}{\alpha_k}q_{k-1}.$$

From $|q_k|=1$, we have $\alpha_k=1$ or $\alpha_k=-1$, with all α_k mutually equal, which concludes the proof. □

Since $q_{k+2n}=(-1)^n q_k$ for an n-periodic billiard, we have the following lemma.

Lemma 2: If the billiard is periodic with the period n, then the divisor sequence D_k joined to the corresponding Heisenberg XYZ system is also periodic, with period $2n$.

Proof: Let $x_{k+n}=x_k$. Join to the billiard trajectory (x_k,y_k) the corresponding flow $(\overline{x_k},\overline{y_k})$. Since

$$(\overline{x_{2k}},\overline{y_{2k}})=(-1)^k(x_k,y_k),\quad (\overline{x_{2k+1}},\overline{y_{2k+1}})=\phi(\overline{x_{2k}},\overline{y_{2k}}),$$

where ϕ is linear, we have

$$\overline{x_{k+2n}}=(-1)^n\overline{x_k}.$$

It follows that $q_{k+2n}=(-1)^n q_k$. According to Lemma 1, the divisor sequence D_k is $2n$ periodic. □

IV. THE CONDITION OF CAYLEY'S TYPE

By Lemmas 1 and 2, the problem of periodic billiard is reduced to the problem of finite-order points on the Jacobian variety of the corresponding isospectral curve. In the three-dimensional case, this curve is hyperelliptic of genus 2. The procedure similar to that given by Griffiths and Harris in Ref. 3 can be applied to hyperelliptic curves.

Let C be a hyperelliptic curve of genus 2, $J(C)$ its Jacobian variety, and

$$u:\mathrm{Div}\ C\to J(C),$$

the Abel–Jacobi map. We are interested in the points $P\in C$ with

$$nu(P)=0,$$

or, more general, in the points $P_1,\ldots,P_n\in C$ with

$$u(P_1)+\cdots+u(P_n)=0,$$

on $J(C)$. After the Abel's theorem,[12] there exists a meromorphic function f on the curve C with zeros $P_1,\dots,P_n$ and poles $Q_1,\dots,Q_n$ if and only if

$$u(P_1)+\cdots+u(P_n)-u(Q_1)-\cdots-u(Q_n)=0.$$

Thus $u(P_1)+\cdots+u(P_n)=0$ if and only if there exists a meromorphic function f with zeros $P_1,\dots,P_n$ and a pole of order n in the point P_∞. The point $u(P_\infty)$ is chosen to be neutral in $J(C)$.

Let $L(nP_\infty)$ be the vector space of meromorphic functions on C with a pole of order of at most n at the point P_∞. After the Riemann–Roch theorem,[12] we have

$$\dim L(nP_\infty)=n-1,$$

since all divisors D with $\deg D\geqslant 3$ on a hyperelliptic curve of genus 2 are nonspecial.

Let $f_1,\dots,f_{n-1}$ be a basis for the vector space $L(nP_\infty)$. The mapping

$$F:C\to P^{n-2},\quad X\mapsto[f_1(X),\dots,f_{n-1}(X)]$$

is a projective embedding whose image is a smooth algebraic curve of degree n. Hyperplane sections of this curve are zeros of functions from $L(nP_\infty)$. Thus, the following claim holds.

Lemma 3: The condition $u(P_1)+\cdots+u(P_n)=0$ *is equivalent to*

$$\operatorname{rank}\begin{bmatrix} f_1(P_1) & \cdots & f_{n-1}(P_1)\\ \cdots & \cdots & \cdots\\ f_1(P_n) & \cdots & f_{n-1}(P_n)\end{bmatrix}<n-1,$$

or, for $P_1=\cdots=P_n=P$,

$$\operatorname{rank}\begin{bmatrix} f_1(P) & \cdots & f_{n-1}(P)\\ f_1'(P) & \cdots & f_{n-1}'(P)\\ \cdots & \cdots & \cdots\\ f_1^{(n-1)}(P) & \cdots & f_{n-1}^{(n-1)}(P)\end{bmatrix}<n-1.$$

In our case C is the isospectral curve

$$\Gamma:y^2=(x-\mu_1)(x-\mu_2)(x-J_1^2)(x-J_2^2)(x-J_3^2).$$

Take point P_∞, which corresponds to $x=\infty$ to be neutral on the Jacobian variety $J(C)$. Let P_0 be one of the points that correspond to the value $x=0$. By Lemma 2, we are interested in the condition $u(2n(P_\infty-P_0))=0$, e.g., $u(2nP_0)=0$.

We can choose, similarly as in Ref. 3, the following basis for $L(2nP_\infty)$:

$$1,x,\dots,x^n,y,xy,\dots,x^{n-3}y.$$

The condition

$$\operatorname{rank}\begin{bmatrix} f_1(P_0) & \cdots & f_{2n-1}(P_0)\\ f_1'(P_0) & \cdots & f_{2n-1}'(P_0)\\ \cdots & \cdots & \cdots\\ f_1^{(2n-1)}(P_0) & \cdots & f_{2n-1}^{(2n-1)}(P_0)\end{bmatrix}<2n-1,$$

is equivalent to

$$\operatorname{rank}\begin{bmatrix} \dfrac{d^{n+1}y}{dx^{n+1}} & \cdots & \dfrac{d^{n+1}(x^{n-3}y)}{dx^{n+1}} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & & \\ \dfrac{d^{2n-1}y}{dx^{2n-1}} & \cdots & \dfrac{d^{2n-1}(x^{n-3}y)}{dx^{2n-1}} \end{bmatrix} < n-2,$$

where d^k/dx^k stands for $d^k/dx^k|_{x=0}$.

Finally, we obtained the following.

Theorem: *A condition on the three-dimensional billiard system trajectory to be n periodic is*

$$\operatorname{rank}\begin{bmatrix} A_{n+1} & A_n & \cdots & A_4 \\ A_{n+2} & A_{n+1} & \cdots & A_5 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & & & \\ A_{2n-1} & A_{2n-2} & \cdots & A_{n+2} \end{bmatrix} < n-2,$$

where

$$\sqrt{(x-\mu_1)(x-\mu_2)(x-J_1^2)(x-J_2^2)(x-J_3^2)} = A_0 + A_1 x + A_2 x^2 + \cdots .$$

By direct generalizing, it is possible to derive a corresponding condition on the billiard inside an ellipsoid in R^N.

V. DISCUSSIONS

At the end, we are going to emphasize that the condition obtained in this way is very similar to the classical Cayley's condition. But, even in the two-dimensional case, the corresponding procedure does not lead to formulas identical to Cayley's.

Let $(Cx,x)=0$ and $(Dx,x)=0$ be equations of two conics in a projective plane. Cayley's condition for the existence of a polygon with n vertices inscribed in C and circumscribed about D is (see Refs. 1, 3)

$$\begin{vmatrix} C_2 & C_3 & \cdots & C_{p+1} \\ C_3 & C_4 & \cdots & C_{p+2} \\ \cdots\cdots\cdots\cdots\cdots\cdots & & & \\ C_{p+1} & C_{p+2} & \cdots & C_{2p} \end{vmatrix} = 0, \quad n = 2p+1,$$

$$\begin{vmatrix} C_3 & C_4 & \cdots & C_{p+1} \\ C_4 & C_5 & \cdots & C_{p+2} \\ \cdots\cdots\cdots\cdots\cdots\cdots & & & \\ C_{p+1} & C_{p+2} & \cdots & C_{2p-1} \end{vmatrix} = 0, \quad n = 2p,$$

where $\sqrt{\det(tC+D)} = C_0 + C_1 t + C_2 t^2 + \cdots$.

Let $\Gamma: \nu^2 = (\mu-\mu_1)(\mu-J_1^2)(\mu-J_2^2)$ be the isospectral curve that corresponds to the XYZ model related to the billiard inside C with caustic D. Cayley's cubic $y^2 = \det(tC+D)$ and Γ are both elliptic curves, but they are obtained in a totally different way with no obvious connection.

Consider an n-periodic billiard with n even. The elliptical integral,

$$M(A,B)=\int_{\theta_A}^{\theta_b}\frac{d\theta}{\sqrt{1-k^2\cos^2\theta}}$$

is a metric on the conic C. (k is the eccentricity of the caustic ellipse D.) According to this metric, the length of the ellipse are between two bounces is always the same.[4] Since M is centrally symmetric, we have that $x_{k+p}=-x_k$, where $n=2p$.

It follows that the corresponding sequence of divisors is also n periodic for even n, and $2n$ periodic for odd n. By direct application of Ref. 3, we get the condition

$$\begin{vmatrix} B_3 & B_4 & \cdots & B_{2p+2} \\ B_4 & B_5 & \cdots & B_{2p+3} \\ \cdots & \cdots & \cdots & \cdots \\ B_{2p+2} & B_{2p+3} & \cdots & B_{4p+1} \end{vmatrix}=0, \quad n=2p+1,$$

$$\begin{vmatrix} B_3 & B_4 & \cdots & B_{p+1} \\ B_4 & B_5 & \cdots & B_{p+2} \\ \cdots & \cdots & \cdots & \cdots \\ B_{p+1} & B_{p+2} & \cdots & B_{2p-1} \end{vmatrix}=0, \quad n=2p,$$

where $\sqrt{(\mu-\mu_1)(\mu-J_1^2)(\mu-J_2^2)}=B_0+B_{1\mu}+B_2\mu^2+\cdots$.

Notice that the two matrices obtained for n odd are not of the same order. We are going to compare them in case $n=3$.

The condition on the billiard trajectory inside the conic C with caustic D to be periodic with period 3, according to Cayley's theorem,[1] is

$$C_2=0,$$

where C_2 is the coefficient with t^2 in the Maclaurin expansion for $\sqrt{\det(Ct+D)}$. According to our formula, this condition becomes

$$\begin{vmatrix} B_3 & B_4 \\ B_4 & B_5 \end{vmatrix}=0.$$

ACKNOWLEDGMENT

This research was supported by the Ministry for Science and Technology of the Republic of Serbia, Project No. 04M03.

[1] H. Lebesgue, *Les Coniques* (Gauthier-Villars, Paris, 1942), pp. 115–149.
[2] P. Griffiths and J. Harris, ''A Poncelet theorem in space,'' Comments Math Helv. **52**, 145–160 (1977).
[3] P. Griffiths and J. Harris, ''On Cayley's explicit solution to Poncelet's porism,'' Enseignement Math. **24**, 31–40 (1978).
[4] S.-J. Chang and R. Friedberg, ''Elliptical billiards and Poncelet's theorem,'' J. Math. Phys. **29**, 1537–1550 (1988).
[5] B. Crespi, S.-J. Chang, and K.-J. Shi, ''Elliptical billiards and hyperelliptic functions,'' J. Math. Phys. **34**, 2257–2289 (1993).
[6] B. Crespi, S.-J. Chang, and K.-J. Shi, ''Elliptical billiard systems and the full Poncelet's theorem in n dimensions,'' J. Math. Phys. **34**, 2242–2256 (1993).
[7] V. Dragović and B. Jovanović, ''On integrable potential perturbations of billiard system within elliopsoid,'' J. Math. Phys. **38**, 3063–3068 (1997).
[8] J. Moser and A. P. Veselov, ''Discrete versions of some classical integrable systems and factorisation of matrix polynomials,'' Commun. Math. Phys. **139**, 217–243 (1991).
[9] A. P. Veselov, ''Integrable systems with discrete time and difference operators,'' Funct. Anal. Appl. **22**, 1–13 (1988).

[10] A. P. Veselov, ''Integrable Lagrange correspondences and factorization of matrix polynomials,'' Funct. Anal. Appl. **25**, 38–49 (1991).
[11] V. Dragović and M. Radnović, ''Periodic trajectories of billiard inside quadric and algebraic curves,'' in *Proceedings of YUCTAM 97*, Vrnjačka Banja, 1997, pp. 7–12.
[12] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley-Interscience, New York, 1978).

Heisenberg chain systems from compact manifolds into S^2

You-De Wang[a)]
Math. Section, ICTP, P.O. Box 586, 34100 Trieste, Italy

(Received 22 November 1996; accepted for publication 7 July 1997)

In this paper, it is shown that the initial value problems of Heisenberg Spin Chain Systems with external magnetic field from compact Riemannian manifolds into S^2 admit global weak solutions. © *1998 American Institute of Physics.* [S0022-2488(97)01712-X]

I. INTRODUCTION

Let Ω be a bounded domain in Euclidean space R^3. Then the three-dimensional classical dynamics on Ω for the isotropic Heisenberg chain with spin density $u=(u^1,u^2,u^3)$ is described by the so-called Hamilton density

$$H=\frac{\alpha}{2}\,|\nabla u|^2-H_0 u,$$

where α is the exchange constant, u is the spin vector, and H_0 is an external magnetic field. It is a Hamiltonian dynamics appearing in the study of condensed matter physics. It is well known that the spin equation of ferromagnetic motion has the form

$$\partial_t u=\alpha u\times\Delta u+u\times H_0,$$

where $\times$ denotes the vector cross product in R^3. It is worthwhile to point out that $|u|=1$, i.e., the so-called magnetically saturated condition (see Refs. 1–3). When the external magnetic field disappears, the above system becomes

$$\partial_t u=\alpha u\times\Delta u.$$

In this paper we will pay attention to the mathematical study of the existence of weak solutions for the initial value problems of Heisenberg chain equations. Let (M,g) be a closed Riemannian manifold with metric $g=(g_{ij})$. Let Δ_M denote the Laplace–Beltrami operator of (M,g) with respect to metric g. In local coordinates $(x^1,\ldots,x^m)$ on M, the Laplace–Beltrami operator can be expressed as

$$\Delta_M=\frac{1}{\sqrt{g}}\,\frac{\partial}{\partial x^i}\left(g^{ij}\sqrt{g}\,\frac{\partial}{\partial x^j}\right),$$

where (g^{ij}) is the inverse of (g_{ij}). Let S^2 be a unit sphere of three-dimensional Euclidean space. Then we also call the following system a Heisenberg chain system (or ferromagnetic chain system):

$$\partial_t u=\alpha u\times\Delta_M u+F(x,t,u),$$

[a)]Permanent address: Institute of Mathematics, Academia Sinica, Beijing 100080, Peoples' Republic of China. Electronic mail: wyd@math03.math.ac.cn

where $u(x,t)$; $M\times R^+\to S^2$ is an unknown mapping and $F(x,t,u):M\times R^+\times S^2\to R^3$ is a smooth vector function which satisfies for any $z\in S^2$

$$F(x,t,z)\cdot z=0.$$

We will consider the following initial value problem:

$$\partial_t u=u\times\Delta_M u+F(x,t,u) \quad \text{in } M\times R^+,$$

$$u(\cdot,0)=u_0 \quad \text{on } M.$$

Definition: If for any smooth function $v\in C^\infty(M\times[0,T],R^3)$ and any T the following integral relation is always true,

$$\int_M (u(T)v(T)-u_0v(0))\,dM-\int_M u\partial_t v\,dM=-\int_M (u\times\nabla u)\nabla v\,dM+\int_M F(x,t,u)v\,dM,$$

we say that u is a global weak solution for the above initial value problem of the Heisenberg chain system.

Now let us recall some known results. In the case $M=R^1$ and $F(x,t,z)=0$, in Ref. 4 Zhou, *et al*. showed that the initial value problem of the Heisenberg chain system admits a unique global smooth solution, if the initial value mapping is smooth.

In the case M is a bounded domain with smooth boundary in R^m $(m\geqslant 2)$, in Refs. 5 and 6 Zhou and Guo established the existence of global solutions for the initial value problems of Heisenberg chain systems, but the image manifolds of solutions are three-dimensional Euclidean spaces. It seems that the target manifolds of spin vector field should be S^2 in physical problems.

Recently, in Ref. 7 the author showed that the initial value problem of the Heisenberg chain system without external magnetic field from a closed Riemannian manifold into S^2 admits a global weak solution. However, some techniques and methods employed there are not valid for the present situation. In this paper our strategy is to adopt an auxillary approximate system of the Heisenberg chain system as follows,

$$\partial_t u=\varepsilon\Delta_M u+\frac{u}{\max\{1,|u|\}}\times\Delta_M u+F\left(x,t,\frac{u}{\max\{1,|u|\}}\right),$$

and then to let $\varepsilon\to 0$ to obtain the solution of Heisenberg chain system. We can also use the methods employed here to approach the existence of weak solutions for the Heisenberg chain system with damping term, i.e., the so-called Landau–Lifchitz system. We will show it in another paper. Our main results in this paper are the following.

Theorem: Let (M,g) be a closed Riemannian manifold. Assume that for any $z\in S^2$, we always have $F(x,t,z)\cdot z=0$ and $u_0(x)\in H^1(M,S^2)$. Then the initial value problem of the Heisenberg chain system admits a global weak solution $u\in L^\infty_{\rm loc}(0,\infty;H^1(M,S^2))\cap C^{1/2}_{\rm loc}(0,\infty;L^2(M,S^2))$.

II. PROOF OF THEOREM

First, we extend smoothly $F(x,t,z)$ as a mapping from $M\times R^+\times B_1(0)$ into R^3, where $B_1(0)$ denotes a closed unit ball centered at the origin of R^3. Let $a(r):[0,1]\to[0,1]$ be a smooth function satisfying $a(r)=0$ on $[0,\delta]$ and $a(r)=1$ on $[1-\delta,1]$, where $\delta<\frac{1}{3}$ is a positive real number. Then, we define $F_e(x,t,z)=a(|z|)F(x,t,z/|z|)$ when $|z|>\delta$, and $F_e(x,t,z)=0$ when $z\in B_\delta(0)$. Obviously, we have $F_e(x,t,z)|_{S^2}=F(x,t,z)$ and $F_e(x,t,z)\cdot z=0$ for any $z\in B_1(0)$. For the sake of convenience, we still denote $F_e(x,t,z)$ by $F(x,t,z)$. Obviously, we always have

$$F(x,t,z)\cdot z=0$$

for any $(x,t,z)\in M\times R^+\times B_1(0)$. In order to establish the existence of global weak solutions for the initial value problem of the Heisenberg chain system from M into S^2, instead of the Heisenberg chain system we discuss the initial value problem of an auxillary approximate system as follows:

$$(\mathrm{IP})\begin{cases} \partial_t u=\varepsilon\Delta_M u+\dfrac{u}{\max\{1,|u|\}}\times\Delta_M u+F\left(x,t,\dfrac{u}{\max\{1,|u|\}}\right), & M\times R^+,\\ u(\cdot,0)=u_0, \quad \text{on } M. \end{cases}$$

Now we construct approximation solutions of the problem (IP) by the standard Galerkin method. Let $\omega_j(x)$ $(j=1,2,...)$ be the normalized eigenfunctions of the equation

$$\Delta_M u+\lambda u=0 \quad \text{in } M,$$

corresponding to the eigenvalues λ_j $(j=1,2,...)$. It is well known that

$$\left\{\frac{1}{\sqrt{\mathrm{vol}(M)}}=\omega_0, \quad \omega_j \;\; j=1,2,...\right\}$$

forms a normalized orthogonal basis in $H^1(M)$.

We write the approximation solutions $u^N(x,t)$ of the problem (IP) in the following form:

$$u^N(x,t)=\sum_{j=0}^{N}\alpha_{jN}(t)\omega_j(x),$$

where $\alpha_{jN}(t)$ $(j=0,1,2,...,N;\ N=1,2,...)$ are the undetermined coefficient vector functions of variable $t\in R^+$. According to the Galerkin method, the coefficient vector functions $\alpha_{jN}(t)$ $(j=0,1,2,...,N;\ N=1,2,...)$ are assumed to satisfy the following system of ordinary differential equations of first order:

$$\begin{aligned}\int_M \partial_t u^N\cdot\omega_s\,dM=\varepsilon\int_M \omega_s\cdot\Delta_M u^N\,dM+\int_M\omega_s\cdot\left(\frac{u^N}{\max\{1,|u^N|\}}\times\Delta_M u^N\right)dM\\ +\int_M\omega_s\cdot F\left(x,t,\frac{u^N}{\max\{1,|u^N|\}}\right)dM,\end{aligned}\tag{2.1}$$

$s=1,2,...,N;$ with the initial value conditions

$$\alpha_{sN}(0)=\int_M\omega_s u_0(x)dM, \quad s=0,1,2,...,N.\tag{2.2}$$

Obviously, the initial problem of the above first-order ordinary differential equations (2.1) and (2.2) can be written as

$$\frac{d}{dt}\alpha_{jN}(t)=-\varepsilon\lambda_j\alpha_{jN}(t)+f_j(x,t,\alpha_{0N},\ldots,\alpha_{NN}),$$

$$\alpha_{jN}(0)=\int_M\omega_s u_0(x)dM,$$

where $\lambda_0=0$ and

$$f_s(x,t,\alpha_{0N},\ldots,\alpha_{NN})=\int_M \omega_s\cdot\left(\frac{u^N}{\max\{1,|u^N|\}}\times\Delta_M u^N\right)dM+\int_M \omega_s\cdot F\left(x,t,\ \frac{u^N}{\max\{1,|u^N|\}}\right)dM.$$

It is easy to verify that $f_s(x,t,\alpha_{0N},\ldots,\alpha_{NN})$ $(s=0,1,...,N)$ are uniformly Lipchitz continuous. By the classical ordinary differential system theory we know that the initial problems (2.1) and (2.2) admit a unique and smooth local solution. Moreover, multiplying the two sides of (2.1) by u^N we can easily derive from (2.1) and (2.2) the following integral relation:

$$\int_M |u^N(T)|^2\ dM+\varepsilon\int_0^T\int_M |\nabla u^N|^2\ dM\ dt\leqslant \mathrm{Vol}(M). \tag{2.3}$$

And also it is easy to see from (2.1) and (2.2) that

$$\int_0^T\int_M \partial_t u^N\cdot\Delta_M u^N\ dM\ dt=\varepsilon\int_0^T\int_M |\Delta_M u^N|^2\ dM\ dt + \int_0^T\int_M F\left(x,t,\ \frac{u^N}{\max\{1,|u^N|\}}\right)\cdot\Delta_M u^N\ dM\ dt.$$

This implies

$$\varepsilon\int_M |\Delta_M u^N|^2\ dM\ dt+\frac{d}{dt}\int_M |\nabla u^N|^2\ dM\leqslant C(T)\int_M |\nabla u^N|^2\ dM+C(T), \tag{2.4}$$

since we have for $(x,t)\in M\times[0,T]$

$$\left|\nabla\left(\frac{u^N}{\max\{1,|u^N|\}}\right)\cdot\ \frac{\partial F\left(x,t,\ \frac{u^N}{\max\{1,|u^N|\}}\right)}{\partial z}\cdot\nabla\mathbf{u}^N\right|\leqslant C(T)|\nabla u^N|^2$$

and

$$\left|\nabla F\left(\cdot,t,\ \frac{u^N}{\max\{1,|u^N|\}}\right)\cdot\nabla u^N\right|\leqslant C(T)|\nabla u^N|.$$

Hence, using the Gronwall inequality we can infer from (2.4)

$$E(u^N(T))\leqslant C(T)+C(T)E(u_0),$$

where

$$E(u)=\int_M |\nabla u|^2\ dM.$$

Therefore, we conclude from the above *a priori* estimates that the ordinary differential system (2.1) and (2.2) has a unique global solution. From (2.1) and (2.2) we can also infer

$$\int_0^T\int_M |\partial_t u^N|^2\ dM\ dt = \int_0^T\int_M \partial_t u^N\cdot\left[\frac{u^N}{\max\{|u^N|,1\}}\times\Delta_M u^N\right] + \int_0^T\int_M \partial_t u^N\cdot F\left(x,t,\frac{u^N}{\max\{1,|u^N|\}}\right)dM\ dt + \varepsilon E(u^N(0))-\varepsilon E(u^N(T)). \tag{2.5}$$

Using the Holder inequality on the right of (2.5) we obtain

$$\varepsilon E(u^N(T))+\frac{1}{2}\int_0^T\int_M |\partial_t u^N|^2\ dM\ dt \leqslant \varepsilon E(u^N(0))+\int_0^T\int_M |\Delta_M u|^2\ dM\ dt + C(T).$$

Noting (2.4), we obtain from the last inequality

$$\varepsilon E(u^N(T))+\frac{1}{2}\int_0^T\int_M |\partial_t u^N|^2\ dM\ dt \leqslant \left(\varepsilon+\frac{1}{\varepsilon}\right)E(u_0)+C(T,\varepsilon,E(u_0)). \tag{2.6}$$

Therefore, when ε is fixed and $N\to\infty$, from the above *a priori* estimates we conclude that there is a function $u^\varepsilon\in W_2^{1,2}(M\times[0,T],R^3)$ for any $T\leqslant\infty$ such that

(i) $\partial_t u^N\to\partial_t u^\varepsilon$ weakly in $L^2(M\times[0,T],R^n)$,
(ii) $\Delta_M u^N\to\Delta_M u^\varepsilon$ weakly in $L^2(M\times[0,T],R^n)$,
(iii) $\nabla u^N\to\nabla u^\varepsilon$ weakly * in $L^\infty(M\times[0,T],R^n)$,
(iv) $u^N\to u^\varepsilon$ for all $t\in R^+$ and a.e. $x\in M$.

Moreover, from (2.1) and (2.2) we also conclude that for any smooth function $v\in C_\infty(M\times[0,T];R^n)$, we have

$$\int_0^T\int_M \partial_t u^\varepsilon v\ dM\ dt = \int_0^T\int_M v\cdot\left[\frac{u^\varepsilon}{\max\{|u^\varepsilon|,1\}}\times\Delta_M u^\varepsilon\right]dM\ dt + \varepsilon\int_0^T\int_M \Delta_M u^\varepsilon v\ dM\ dt+\int_0^T\int_M vF\left(x,t,\frac{u^\varepsilon}{\max\{1,|u^\varepsilon|\}}\right)dM\ dt.$$

This implies u^ε is a global weak solution of the problem (IP).

Lemma 2.1: If u^ε is a global weak solution of the problem (IP), then there holds

$$|u^\varepsilon|\leqslant 1$$

for all $t\in R^+$ and a.e. $x\in M$.

Proof: We choose $\psi=u^\varepsilon-(u^\varepsilon/|u^\varepsilon|)\min\{1,|u^\varepsilon|\}$ as a test function of the first equation of (IP). Then, multiplying the first equation of (IP) by ψ and noting $F(x,t,z)\cdot z=0$, we obtain

$$\frac{1}{2}\frac{d}{dt}\int_{|u^\varepsilon|\geqslant 1}|u^\varepsilon|^2\left(1-\frac{1}{|u^\varepsilon|}\right)dM = \frac{1}{2}\int_{|u^\varepsilon|\geqslant 1}\frac{u^\varepsilon\cdot\partial_t u^\varepsilon}{|u^\varepsilon|}\,dM - \varepsilon\int_{|u^\varepsilon|\geqslant 1}\frac{|u^\varepsilon\cdot\nabla u^\varepsilon|^2}{|u^\varepsilon|^3}\,dM-\varepsilon\int_{|u^\varepsilon|\geqslant 1}|\nabla u^\varepsilon|^2\left(1-\frac{1}{|u^\varepsilon|}\right)dM. \tag{2.7}$$

Letting $u^\varepsilon/|u^\varepsilon|$ be a test function of (IP) on set $\{|u^\varepsilon|\geqslant 1\}$, we have

$$\frac{1}{2}\int_{|u^\varepsilon|\geqslant 1}\frac{u^\varepsilon\cdot\partial_t u^\varepsilon}{|u^\varepsilon|}\,dM=\frac{\varepsilon}{2}\int_{|u^\varepsilon|\geqslant 1}\frac{|u^\varepsilon\cdot\nabla u^\varepsilon|^2}{|u^\varepsilon|^3}\,dM-\frac{\varepsilon}{2}\int_{|u^\varepsilon|\geqslant 1}|\nabla u^\varepsilon|^2\frac{1}{|u^\varepsilon|}\,dM$$
$$+\int_{|u^\varepsilon|=1}\frac{\partial u^\varepsilon}{\partial n}\cdot u^\varepsilon\,dS. \tag{2.8}$$

Here n is outward normal direction of the boundary of set $\{|u^\varepsilon|\geqslant 1\}$. Since $\partial u^\varepsilon/\partial n\cdot u^\varepsilon=\frac{1}{2}\partial|u^\varepsilon|^2/\partial n\leqslant 0$ on the boundary set $\{|u^\varepsilon|=1\}$, it is easy to see that (2.7) and (2.8) imply the following integral inequality:

$$\frac{1}{2}\frac{d}{dt}\int_{|u^\varepsilon|\geqslant 1}|u^\varepsilon|^2\left(1-\frac{1}{|u^\varepsilon|}\right)dM+\varepsilon\int_{|u^\varepsilon|\geqslant 1}|\nabla u^\varepsilon|^2\left(1-\frac{1}{|u^\varepsilon|}\right)dM\leqslant 0. \tag{2.9}$$

However, obviously this inequality implies

$$|u^\varepsilon|\leqslant 1$$

for any $t\in R^+$ and a.e. $x\in M$, since $|u^\varepsilon(x,0)|=|u_0(x)|=1$. Thus we complete the proof of the lemma.

Lemma 2.2: If u^ε is a global weak solution of (IP), then, as $\varepsilon\to 0$, there exists a function $u\in L^\infty(0,T;H^1(M,S^2))$ such that

$$u^\varepsilon\to u$$

weakly * in $L^\infty(0,T;H^1(M,S^2))$ and

$$|u(x,t)|=1$$

for all $t\in R^+$ and a.e. $x\in M$.

Proof: Obviously, we have $E(u^\varepsilon(t))\leqslant C(T)$ for $t\in[0,T]$. It implies there exists a function $u\in L^\infty(0,T;H^1(M,R^3))$ such that, neglecting a subsequence, $u^\varepsilon\to u$ weakly * in $L^\infty(0,T;H^1(M,R^3))$. Since u^ε is a global weak solution of (IP) and $F(x,t,z)\cdot z=0$, it is easy to see

$$\int_M|u^\varepsilon(T)|^2\,dM-\int_M|u^\varepsilon(0)|^2\,dM+\varepsilon\int_0^T\int_M|\nabla u^\varepsilon|^2\,dM\,dt=0.$$

Noting $|u^\varepsilon(0)|=1$, we have

$$\int_M(|u^\varepsilon(T)|^2-1)dM=-\varepsilon\int_0^T\int_M|\nabla u^\varepsilon|^2\,dM\,dt.$$

As $\varepsilon\to 0$, noting that $u^\varepsilon(\cdot,T)\to u(\cdot,T)$ strongly in $L^2(M)$ for any fixed T, we obtain

$$\int_M(|u(T)|^2-1)dM=0.$$

Yet, from Lemma 2.1 we know there holds $|u^\varepsilon(T)|\leqslant 1$ for any ε. Since $u^\varepsilon\to u$ for all $t\in[0,T]$ and a.e. $x\in M$, we also have

$$|u(T)|\leqslant 1.$$

Therefore, we conclude that there holds

$$|u(x,t)|^2=1$$

for all $t\in R^+$ and for a.e. $x\in M$, so $u(x,t)\in L^\infty(0,T;H^1(M,S^2))$.

Now we claim that $u(x,t)$ is a global weak solution of the initial value problem of Heisenberg chain system from M into S^2. In fact, since u^ε is a global weak solution of (IP) and $|u^\varepsilon(x,t)|\leqslant 1$, then for any smooth vector function $v\in C^\infty(M\times[0,T],R^n)$, we have

$$\int_0^T\int_M u^\varepsilon\partial_t v\ dM\ dt=\int_M u^\varepsilon(T)v(T)dM-\int_M u^\varepsilon(0)v(0)dM+\varepsilon\int_0^T\int_M\nabla u^\varepsilon\nabla v\ dM\ dt$$
$$+\int_0^T\int_M(u^\varepsilon\times\nabla u^\varepsilon)\nabla v\ dM\ dt-\int_0^T\int_M F(x,t,u^\varepsilon)v\ dM\ dt. \quad (2.10)$$

This implies

$$\int_0^T\int_M u^\varepsilon\partial_t v\ dM\ dt=\int_M u^\varepsilon(T)v(T)dM-\int_M u_0v(0)dM+\varepsilon\int_0^T\int_M\nabla u^\varepsilon\nabla v\ dM\ dt$$
$$+\int_0^T\int_M(u^\varepsilon\times\boldsymbol{\nabla}u^\varepsilon)\boldsymbol{\nabla}v\ dM\ dt-\int_0^T\int_M F(x,t,u^\varepsilon)v\ dM\ dt.$$

Note that $u^\varepsilon\to u$ for a.e. $x\in M$ and all $t\in[0,T]$ and for any ε

$$|F(x,t,u^\varepsilon)|\leqslant C(T).$$

So, by the well-known controlled convergence theorem we have that $F(x,t,u^\varepsilon)\to F(x,t,u)$ strongly in $L^p(M\times[0,T])$ for any $p\in[2,\infty)$. Hence,

$$\int_0^T\int_M F(x,t,u^\varepsilon)v\ dM\ dt\to\int_0^T\int_M F(x,t,u)v\ dM\ dt.$$

Then letting $\varepsilon\to 0$ in the above integral relation, we obtain

$$\int_0^T\int_M u\partial_t v\ dM\ dt=\int_M u(T)v(T)dM-\int_M u_0v(0)dM+\int_0^T\int_M(u\times\nabla u)\nabla v\ dM\ dt$$
$$-\int_0^T\int_M F(x,t,u)v\ dM\ dt.$$

Furthermore, from (2.10) we easily see there holds for any v

$$\int_0^T\int_M\partial_t u^\varepsilon v\ dM\ dt\leqslant C(\varepsilon,T)\left(\int_0^T\int_M(v^2+|\nabla v|^2)dM\ dt\right)^{1/2},$$

where $C(\varepsilon,T)\leqslant C(T)$ when ε is small enough. Hence, $\partial_t u^\varepsilon\in L^2_{\rm loc}(0,\infty;H^{-1}(M))$. Applying the interpolation formula of the Hilbert space with negative order we obtain

$$\|u^\varepsilon(\cdot,t_2)-u^\varepsilon(\cdot,t_1)\|_{L^2(M)}\leqslant C\|u^\varepsilon(\cdot,t_2)-u^\varepsilon(\cdot,t_1)\|^{1/2}_{H^1(M)}\cdot\|u^\varepsilon(\cdot,t_2)-u^\varepsilon(\cdot,t_1)\|^{1/2}_{H^{-1}(M)}$$
$$\leqslant C\ \sup_{0\leqslant t\leqslant T}\|\partial_t u^\varepsilon\|^{1/2}_{H^{-1}(M)}\cdot\sup_{0\leqslant t\leqslant T}\|u^\varepsilon\|^{1/2}_{H^1(M)}|t_2-t_1|^{1/2},$$
$$\leqslant C(T)|t_2-t_1|^{1/2}.$$

When $\varepsilon\leqslant 1$, $C(T)$ does not depend on ε. So, as $\varepsilon\to 0$, we have

$$\|u(\cdot,t_2)-u(\cdot,t_1)\|_{L^2(M)}\leqslant C(T)|t_2-t_1|^{1/2}.$$

Thus we have completed the proof of the Theorem.

If (M,g) is a compact Riemannian manifold with smooth boundary, we consider the following:

$$(\mathrm{IBP})\begin{cases}\partial_t u=u\times\Delta_M u+F(x,t,u) & \text{in } M\times R^+,\\ \dfrac{\partial u}{\partial n}=0 & \text{on } \partial M,\\ u(\cdot,0)=u_0 & \text{in } M,\end{cases}$$

where $u_0(x)\in H^1(M,S^2)$ satisfies $\partial u_0/\partial n=0$ in the sense of distribution. In the same spirit as we proved the above Theorem we can give the proof of the following

Theorem 2.1: Let (M,g) be a compact Riemannian manifold with smooth boundary ∂M. Assume that $F(x,t,z)\cdot z=0$ for any $z\in S^2$ and $u_0\in H^1(M,S^2)$ with $\partial u_0/\partial n=0$ in the sense of distribution. Then the problem (IBP) admits a global weak solution $u\in L^\infty_{\mathrm{loc}}(0,\infty;H^1(M,S^2))\cap C^{1/2}_{\mathrm{loc}}(0,\infty;L^2(M,S^2))$.

Theorem 2.2: If M is a compact manifold with $\dim(M)=1$, $F(x,t,u)$ and u_0 satisfy the same conditions as in the Theorem. Then the initial value problem of Heisenberg chain system admits a global weak solution $u\in L^\infty_{\mathrm{loc}}(0,\infty;H^1(M,S^2))\cap C^{1/2,1/4}_{\mathrm{loc}}(M\times[0,\infty))$, i.e., $u(x,0)=u_0$ in the classical sense.

Proof: In fact, since the Theorem asserts

$$\|u(x,t_1)-u(x,t_2)\|_{L^2}\leqslant C(T)|t_1-t_2|^{1/2},$$

using the Gagliardo–Nirenberg inequality we obtain

$$\|u(x,t_1)-u(x,t_2)-m\|_{L^\infty}\leqslant\|u(x,t_1)-u(x,t_2)-m\|_{L^2}^{1/2}\|u_x(x,t_1)-u_x(x,t_2)\|_{L^2}^{1/2},$$

where

$$m=\frac{1}{\mathrm{vol}(M)}\int_M(u(x,t_1)-u(x,t_2))\,dM\leqslant\left(\frac{1}{\mathrm{vol}(M)}\int_M(u(x,t_1)-u(x,t_2))^2dM\right)^{1/2}$$
$$\leqslant C(T,M)|t_1-t_2|^{1/2}.$$

Hence it is easy to see

$$\sup_{x\in M}|u(x,t_1)-u(x,t_2)|\leqslant C(T)|t_1-t_2|^{1/4},$$

for any $t_1,t_2\in[0,T]$. Noting the Sobolev embedding theorem we have

$$|u(x_1,t_1)-u(x_2,t_2)|\leqslant C(T)(|t_1-t_2|^{1/4}+|x_1-x_2|^{1/2}),$$

for any (x_1,t_1), $(x_2,t_2)\in M\times[0,T]$. Thus we complete the proof of Theorem 2.2.

ACKNOWLEDGMENTS

The author would like to thank the International Center for Theoretical Physics in Trieste for hospitality and financial support during this research and also wishes to thank the referees for their good suggestions. This research is also supported partially by the National Science Foundation.

[1] H. C. Fogedby, *Theoretical Aspects of Mainly Low Dimensional Magnetic System*, Lecture Notes Physics Vol. 131 (Springer, Berlin, 1980).
[2] K. Nakamura and T. Sasada, ''Soliton and wave trains in ferromagnets,'' Phys. Lett. A **48**, 321–322 (1974).
[3] J. Tjon and J. Wright, ''Soliton in the continuous Heisenberg spin chain,'' Phys. Rev. B **15**, 3470–3476 (1977).
[4] Y. L. Zhou, B. Guo, and S. Tan, ''Existence and uniqueness of smooth solution for system of ferromagnetic chain,'' Sci. Sinica A **34**, 257–266 (1991).
[5] Y. L. Zhou and B. Guo, ''Weak solution of system of ferromagnetic chain with several variables,'' Sci. Sinica A **27**, 1251–1266 (1987).
[6] Y. L. Zhou and B. Guo, ''The solvability of the initial value problem for the quasilinear degenerate parabolic systems,'' Proc. DD-3 Symposium (1982), pp. 713–732.
[7] Y. D. Wang, ''Ferromagnetic chain equations from closed Riemannian manifolds onto S^2,'' Int. J. Math. **6**, 93–99 (1995).

Spacetime defects: Domain walls and torsion

L. C. Garcia de Andrade[a)]
Departamento de Física Teórica - Instituto de Física - UERJ,
Rua São Fco. Xavier 524, Rio de Janeiro, RJ, Maracanã, CEP:20550-003, Brasil

(Received 20 February 1997; accepted for publication 4 September 1997)

The theory of distributions in non-Riemannian spaces is used to obtain exact static thin domain wall solutions of Einstein-Cartan equations of gravity. Curvature δ-singularities are found while Cartan torsion is given by Heaviside functions. Weitzenböck planar walls are characterized by torsion δ-singularities and zero curvature. It is shown that Weitzenböck static thin domain walls do not exist exactly as in general relativity. The global structure of Weitzenböck nonstatic torsion walls is investigated. © *1998 American Institute of Physics.*
[S0022-2488(98)00101-7]

I. INTRODUCTION

The geometry of δ-torsion singularities[1] have been extensively studied in connection with spacetime defects.[2–4] An interesting class of Weitzenböck spacetime defects representing torsion loops and line defects representing cosmic strings was recently investigated by Letelier.[5] Also, pure Riemannian defects (disclinations) with zero Cartan's torsion over lines (strings) and walls (domain walls) have been investigated by Letelier[5,6] and Wang. In their study they have considered that metric is continuous across the junction and the first derivative is represented by a Heaviside function. Thus in their approach, the δ-singularity just appears on the Riemann curvature. We shall adopt here the same procedure when dealing with Riemann-Cartan domain walls.

A Riemann-Cartan domain wall solution of the Einstein-Cartan field equation is found where the torsion 2-form is a Heaviside function and the curvature is of δ-singularity-type like in Letelier and Wang[5,6] Riemannian domain walls. The solution is obtained by assuming the symmetric energy-momentum tensor in Einstein-Cartan gravity is the same as the general relativistic thin domain wall tensor proposed by Vilenkin.[7] In the case of static Weitzenböck wall it is shown that the Vilenkin thin domain wall energy-momentum tensor does not satisfy the Einstein-Cartan field equations. Therefore, by gluing together pieces of Minkowski spaces it is possible to find solutions of Einstein-Cartan field equations as is done in Einstein field equations of general relativity.[8] Thick domain walls in Einstein-Cartan theory of gravitation can be investigated in the near future. In a certain sense this paper extends the investigation carried out by A. Wang[9,10] on plane-symmetric walls in general relativity. The important features of this paper is the fact that both torsion and curvature are necessary to guarantee that static thin domain walls exist in Einstein-Cartan theory of gravity.

II. GEOMETRY OF WEITZENBÖCK WALLS

Spacetime defects[2,5,6] appear in distinct forms. For example, they appear as torsion line defects (cosmic strings) or as torsion loops. Domain walls appear in the context of general relativity as another type of defect in spacetime. In particular, plane cosmic walls[11] have been studied in the case of plane gravitational collapse. I shall be concerned with the investigation of a type of defect which is very common in the differential geometry of fractures where torsion δ-functions are

[a)]Electronic mail: Garcia@symbcomp.uerj.br

0022-2488/98/39(1)/372/8/$15.00
© 1998 American Institute of Physics

generated from Heaviside functions. Here a static plane symmetric spacetime with zero Riemann-Cartan curvature and nonvanishing torsion, the so-called Weitzenböck spacetime, is found where one of the metric functions is a Heaviside function.

The metric is therefore a piecewise constant and torsion is a δ-singularity which changes when one passes from one plane to the other. The Weitzenböck geometry of fractures is then developed. Let us now consider the plane symmetric metric

$$ds^2=(\omega^0)^2-(\omega^1)^2-(\omega^2)^2-(\omega^3)^2, \tag{1}$$

where

$$\omega^0=e^{F/2}dt, \quad \omega^1=e^{H/2}dx, \quad \omega^2=e^{H/2}dy, \quad \omega^3=e^{G/2}dz. \tag{2}$$

The torsion 2-forms are chosen according to

$$\begin{aligned} &T^0=J^0\omega^0\wedge\omega^3, \quad T^1=J^1\omega^3\wedge\omega^1+J^2\omega^0\wedge\omega^1, \\ &T^2=J^1\omega^3\wedge\omega^2+J^2\omega^0\wedge\omega^2, \quad T^3=J^3\omega^0\wedge\omega^3. \end{aligned} \tag{3}$$

Making use of Cartan's calculus where the torsion 2-forms are given by

$$T^a=d\omega^a+\omega^a{}_b\wedge\omega^b \text{ (a,b=0,1,2,3)}, \tag{4}$$

one obtains the following 1-form connections:

$$\begin{aligned} &\omega^0{}_3=\left[J^0+e^{-G/2}\frac{F'}{2}\right]\omega^0, \quad \omega^1{}_0=\left[J^2+e^{-F/2}\frac{\dot{H}}{2}\right]\omega^0, \\ &\omega^1{}_3=\left[J^1+e^{-G/2}\frac{H'}{2}\right]\omega^1, \quad \omega^2{}_3=-\left[J^1+e^{-G/2}\frac{H'}{2}\right]\omega^2, \end{aligned} \tag{5}$$

and $J^3=(\dot{G}/2)\,e^{-F/2}$.

Making the choices

$$J^0=-e^{-G/2}\frac{F'}{2}, \quad J^1=e^{-G/2}\frac{H'}{2}, \quad J^2=e^{-F/2}\frac{\dot{H}}{2}, \tag{6}$$

one notices that for this choice of torsion functions $\omega^a_b\equiv 0$ and from the second Cartan's structure equation

$$R^a{}_b=d\omega^a{}_b+\omega^a{}_c\wedge\omega^c{}_b, \tag{7}$$

$R^a{}_b=0$ where $R^a{}_b\equiv 1/2R^a{}_{bcd}\omega^c\wedge\omega^d$ are the curvature 2-forms, thus the curvature $R^a{}_{bcd}$ vanishes.

Let us now consider the static case where $\dot{H}=\dot{G}=\dot{F}=0$.

To simplify matters one considers $J^{(0)}\equiv 0$, which implies $F'=0$ and $F=const.$ Making the choice $G=G(z)=const.$, the only surviving equation is

$$H'(z)=c_1J^1, \tag{8}$$

where c_1 is a constant. Choosing $J^1\equiv\delta(z)$, where $\delta(z)$ is Dirac δ-function, equation (8) has the following solution:

$$H(z)\equiv\theta_0(z)=\begin{cases}1, z>0\\ 0, z<0,\end{cases} \tag{9}$$

where $\theta_0(z)$ is the Heaviside function. Thus the metric is basically given by a Heaviside function and torsion is a δ -singularity.[6]

Notice that outside the defect $(z\neq 0)H(z)=0$, $F=0$, $G=const.$ and the spacetime is Minkowskian. $ds^2=dt^2-dx^2-dy^2-dz'^2$, where $dz'^2=c^2dz^2$ and since c is a constant $z'=cz$ which means a translation along the z-axis. To resume we show that a new class of fractures in Weitzenböck spacetime exists besides torsion loops and torsion lines.

From this computation it is easy to show that static thin Weitzenböck domain walls are forbidden in Einstein-Cartan theory of gravity. Thin domain walls have the following energy-momentum tensor

$$T_a{}^b=\sigma\ \mathrm{diag}(1,1,1,0)\,\delta(z). \tag{10}$$

From the Einstein-Cartan field equations

$$R_{ab}(\Gamma)-\tfrac{1}{2}g_{ab}R(\Gamma)=T_{ab} \tag{11}$$

and the Weitzenböck condition of vanishing total curvature $R_{abcd}(\Gamma)=0$, it is obvious that $\sigma=0$ and thus there is no static thin domain wall formation in Weitzenböck universe. Therefore, we are compelled to consider Riemann-Cartan domain walls. In fact, in the next section I shall present a simple solution of Einstein-Cartan field equations of gravity representing a static thin domain walls.

III. RIEMANN-CARTAN THIN DOMAIN WALLS

In this section I shall be concerned with finding a solution to Einstein-Cartan theory of gravity which describes a thin domain wall[12] given by the energy-momentum tensor equation (10).

I shall compute this solution by solving the system composed by the first and second Cartan's equations of structure given in the last section and the Einstein-Cartan field equation (11). From the 1-form connections (7) and the second Cartan structure equation, after some algebra, one obtains the components of the Riemann-Cartan nonvanishing components of curvature

$$\begin{aligned}
R^2{}_{112} &= -\frac{(H')^2}{4}e^{-F},\\
R^2{}_{332} &= -\frac{e^{-F}}{2}\left[H''+\frac{H'F'}{2}-\frac{(H')^2}{2}\right]=R^3{}_{223},\\
R^1{}_{331} &= -\frac{e^{-F}}{2}\left[H''+\frac{H'F'}{2}\right],\\
R^0{}_{330} &= -\frac{e^{-F}}{2}\left[2J^{0\prime}-\frac{F'J^0}{2}+\frac{F'^2}{2}\right],\\
R^0{}_{110} &= R^2{}_{002}=-\frac{e^{-F}}{2}\left[2J^{0\prime}+\frac{F'H'}{2}\right]
\end{aligned} \tag{12}$$

where to simplify matters we chose J^0 as the only nonvanishing component of Cartan's torsion 2-forms. This choice is compatible with Cartan's first equation of structure (4). The metric is also chosen to be static and Vilenkin hypothesis $F=G$ is also used. The Einstein tensor is defined as

$$G_{ab}=R_{ab}(\Gamma)-\tfrac{1}{2}g_{ab}R(\Gamma), \tag{13}$$

where Γ is the Riemann-Cartan connection.

The Ricci-Cartan tensor and Ricci scalar are given by

$$\begin{aligned} R_{00}&=\frac{e^{-F}}{2}\left[\frac{F'}{2}J^0-\frac{F'^2}{2}\right],\\ R_{11}&=\frac{e^{-F}}{2}\left[2J^{0\prime}+\frac{H'^2}{2}+H''+F'H'\right],\\ R_{22}&=\frac{e^{-F}}{2}[-2J^{0\prime}+H''],\\ R_{33}&=\frac{e^{-F}}{2}\left[2J^{0\prime}-\frac{F'J^0}{2}+\frac{F'^2-H'^2}{2}-H'F'\right], \end{aligned} \tag{14}$$

and

$$R\equiv R^0{}_0+R^1{}_1+R^2{}_2+R^3{}_3=\frac{e^{-F}}{2}[H''+F'H'-2J^{0\prime}+F'J^0-F'^2]. \tag{15}$$

Collecting all those terms, substitution into the Einstein-Cartan equations yields

$$\begin{aligned} G_{00} &= \frac{e^{-F}}{2}\left[2J^{0\prime}-H''-\frac{F'H'}{2}\right]=\sigma\delta(z),\\ G_{11} &= \frac{e^{-F}}{2}\left[H''-\frac{H'^2}{2}+\frac{F'H'^2}{2}\right]=-\sigma\delta(z),\\ G_{22} &= \frac{e^{-F}}{2}\left[-4J^{0\prime}+H''+\frac{H'F'}{2}+\frac{F'J^0}{2}-F'^2\right]=-\sigma\delta(z),\\ G_{33} &= \frac{e^{-F}}{2}\left[-\frac{H'^2}{2}+H''-\frac{F'H'}{2}\right]=0. \end{aligned} \tag{16}$$

This system of nonlinear ODE can be easily solved if one considers $F'=0$. With this choice Cartan torsion reduces to a Heaviside function since

$$J^{0\prime}=\frac{c\sigma}{4}\delta(z)\rightarrow J^0=\frac{c\sigma}{2}\theta_0(z), \tag{17}$$

and the metric function H″ reads

$$H''=-\frac{c\sigma}{2}\delta(z)\rightarrow H'=-\frac{c\sigma}{2}\theta_0(z). \tag{18}$$

Expression (17) means that torsion undergoes a jump across the hyperplane $z=0$ while from (18) most of the curvature components behave like a δ-singularity. For example, $R^0{}_{330}=R^0{}_{220}=R^0{}_{110}=R^1{}_{331}=R^2{}_{332}=c\sigma\delta(z)$ while $R^2{}_{112}=-(c^2/4)\sigma^2\theta_0{}^2(z)$ which is the square of the Heaviside function.

From (18) expression (12) reduces to

$$R^0{}_{330}=\tau[2\sigma\delta(z)],$$

$$R^0{}_{110}=R^2{}_{002}=\delta(z)k,$$

$$R^1{}_{331}=c[\sigma\delta(z)], \tag{19}$$

$$R^2{}_{112}=cH'^2=-\frac{c^2}{4}\sigma^2\theta_0(z),$$

$$R^2{}_{332}=c[2\sigma\delta(z)],$$

where $k\equiv 2c$. All curvature components vanish except $R^2{}_{112}$ and $R^2{}_{332}$. Since the spacetime outside z=0 plane has to be locally flat, vanishing of these components leads to $H'=0$ which yields $H=const.$ and the spacetime metric is Minkowskian. Thus our domain wall geometry is composed of two Minkowski half-spaces glued together across a torsion and curvature junction. As in the Letelier[5] paper, no attempt is made to find a field theory interpretation for the solution discussed here. It is easy to show that the Bianchi identities are identically satisfied.

This solution is similar to a solution found by Holvorcen and Letelier[12] for plane symmetric cosmic walls and gravitational collapse. To resume we have investigated two classes of solutions that represent static thin planar walls in spacetime with torsion. In one of the solutions torsion is a δ-singularity. In the other, torsion is a Heaviside function while the Riemannian curvature possess a δ-singular behavior. Notice that no physical interpretation is tried for the Weitzenböck class of solutions. Nevertheless, they may represent a class of gravitational ghosts like ghost neutrinos where the stress-energy tensor vanishes although the current is distinct from zero. Indeed, our solution is not able to represent ghost neutrinos since our axial torsion vanishes and thus no axial current of ghost neutrinos[13] is possible. Nevertheless, Letelier has found recently[2] a class of Weitzenböck spacetime defects where an axial torsion exists, being a candidate for a ghost neutrino space. Torsion and curvature δ-singularities have also been found in Jackiw[14] three dimensional planar gravity. Since domain walls in the large scale structure of the universe are either unstable or thick, the domain wall solution discussed here can only be considered as a thin domain wall approximation of a static thick domain wall in Riemann-Cartan spacetime. Work in this direction shall be addressed in the near future. A detailed investigation of the global properties of Riemann-Cartan domain walls can appear elsewhere.

IV. GLOBAL STRUCTURE OF TORSION WALLS

The investigation of maximal extension of metrics in Riemann-Cartan spacetime, like the one studied in the last section, can be extremely difficult. Therefore, to simplify matters in this section I shall address some important questions concerning nonstatic Weitzenböck torsion walls which answer many questions which appear in the more general case of domain walls in Riemann-Cartan spaces. Therefore, let us consider the metric used in Ref. 9 to investigate gravitationally repulsive domain wall solutions of Einstein's equations of general relativity. The metric is given by

$$ds^2=e^{2\nu(t,|z|)}(-dt^2+dz^2)+B(t,|z|)(dx^2+dy^2). \tag{20}$$

In the notation of Cartan's calculus, metric (20) reads

$$ds^2=-(\omega^0)^2+(\omega^1)^2+(\omega^2)^2+(\omega^3)^2, \tag{21}$$

where the basis 1-forms $\omega^a(a=0,1,2,3)$ are

$$\omega^0=e^{\nu}dt, \quad \omega^1=e^{\nu}dz, \omega^2=\sqrt{B}dx, \quad \omega^3=\sqrt{B}dy. \tag{22}$$

Choosing Cartan torsion 2-forms *a priori* by

$$T^A = J^A \omega^0 \wedge \omega^1, \tag{23}$$

where (A=0,1) and $T^i = 0$ for (i=2,3) and using Cartan's first structure equation (4), one is able to compute the following connection 1-forms:

$$\omega^0{}_1 = (J^0 + e^{-\nu}\nu_z)\omega^0, \quad \omega^2{}_0 = \frac{1}{2}\frac{B_t}{B}e^{-\nu}\omega^2,$$
$$\omega^3{}_0 = \frac{1}{2}\frac{B_t}{B}e^{-\nu}\omega^3, \quad \omega^2{}_1 = \frac{1}{2}\frac{B_z}{B}e^{-\nu}\omega^2, \quad \omega^3{}_1 = \frac{1}{2}\frac{B_z}{B}e^{-\nu}\omega^3, \tag{24}$$

and

$$J^1 = e^{-\nu}\nu_t, \tag{25}$$

the others zero. From the second Cartan's structure equation (7), one notices that the Weitzenböck condition of zero curvature can be obtained by making $\omega^a{}_b \equiv 0$. Substitution of this expression into (24) yields

$$J^0 = -e^{-\nu}\nu_z, \quad B_t = B_z = 0. \tag{26}$$

Since by definition a Weitzenböck torsion wall is one where the torsion must be equal to a δ-singularity, one should impose this condition in expression (26). This procedure can be obtained by considering the following ansatz:

$$\nu = -(t - \theta_0(z)). \tag{27}$$

Substitution of (26) into (27) yields immediately

$$J^0 = -e^{(t-\theta_0(z))}\delta(z), \quad J^1 = -e^{(t-\theta_0(z))} \tag{28}$$

for δ-singularity distributions of torsion forms. From the second equation in (26) one obtains $B = B_0 = const.$ Thus the nonstatic spacetime metric is

$$ds^2 = e^{-2(t-\theta_0(z))}(-dt^2 + dz^2) + (B_0)^2(dx^2 + dy^2). \tag{29}$$

Metric (29) can be rewritten in a conformal form

$$ds^2 = e^{-2(t-\theta_0(z))}[-dt^2 + dz^2 + (B_0)^2 e^{2(t-\theta_0(z))}(dx^2 + dy^2)]. \tag{30}$$

From the properties of the θ_0 function above, one has outside the torsion wall

$$ds^2 = \alpha e^{-2t}[-dt^2 + dz^2 + (B_0)^2 e^{2t}(dx^2 + dy^2)](z < 0), \tag{31}$$

where α is a constant.

This metric is nothing more than a metric conformal to de Sitter spacetime.

At $z > 0$ the spacetime is

$$ds^2 = e^{-2(t-1)}(-dt^2 + dz^2) + (B_0)^2(dx^2 + dy^2) \tag{32}$$

or

$$ds^2 = \alpha' e^{-2t}[-dt^2 + dz^2 + \gamma' e^{2t}(dx^2 + dy^2)](z > 0). \tag{33}$$

At t=0 the spacetimes outside the wall are flat as characterized by a spacetime defect. As $t\to\infty$ from (26) the torsion J^0 also ''explodes'' and the torsion wall suffers gravitational collapse. The solution at z=0 is obtained by gluing together the two halves of the de Sitter spaces given by the line elements (31) and (33).

Let us now consider the motion equations. In general, when one deals with spacetime with torsion test particles follow distinct paths. The first are autoparallels and the second are geodesics. Since we are just concerned with Weitzenböck spaces where the Riemann-Cartan connection vanishes autoparallel are straight lines. Thus geodesic equations read

$$\ddot{t}+\dot{t}^2+\dot{z}^2=0, \quad \ddot{z}+\dot{t}\dot{z}=0, \quad \ddot{x}=\ddot{y}=0, \tag{34}$$

where dots in (34) are the mean derivation with respect to proper time S. The geodesic of the z-coordinate is easily solved by $\dot{z}=e^{-t}$ and $\ddot{z}=-e^{-t}$. Thus considering $\dot{t}>0$ $(s>0)$, $\dot{z}>0$ and $\ddot{z}<0$ and since the speed of the test particle vanishes at $t\to\infty$ and the acceleration is opposite to the velocity, the test particle falls back to the torsion defect. The test particle is clearly bound. Notice that another solution to the geodesic motion would be $\dot{z}=-e^{-t}$ and $\ddot{z}=e^{-t}\dot{t}$, and in this case test particles are moving towards the torsion defect at $z=0$ plane. Since once more acceleration is opposite to the velocity it may be possible, in principle, that a repulsive gravitational force exists at the torsion wall and can be used to hold or at least decrease the speed of the gravitational collapse due to torsion effects.

More realistic models of nonstatic domain walls in Riemann-Cartan spacetime may be found elsewhere. Discussion of stability of cosmic strings in general relativity may be found in the paper by Gleiser and Pullin.[15] The issue of stability of domain walls in the context of Einstein-Cartan gravity may be found elsewhere as well as the study of transparency of domain walls to neutrino waves.

This is closely related to Trautman's[16] idea to use spin and torsion in the context of Einstein-Cartan theory of gravity to hold or avert gravitational collapse of stellar objects. To be pure,[5] a spacetime defect needs to be empty of matter and fields outside the defect. Therefore, to obey the rule we must force the torsion 2-form component J^1 to vanish. In this case, from expressions (25) ν_t and the metric shall not depend on a time coordinate. Thus a static torsion defect is obtained where Cartan torsion is given only by a Dirac δ-function.

From metric (29) it is easily seen that the spacetime off the wall is flat. Therefore, the metric off the wall is time independent. However, from expression (28) it is seen that the thin wall is not static. This situation occurs also in the study of thin domain walls in general relativity.

ACKNOWLEDGMENTS

I am very much in debt to Professor P. S. Letelier and Professor A. Wang for helpful discussions on the subject of this paper. Financial support from CNPq. and UERJ (FAPERJ) is gratefully acknowledged.

[1] K. P. Tod, Class. Quantum Grav. **11**, 1331 (1994).
[2] P. S. Letelier, Class. Quantum Grav. **12**, 2221 (1995).
[3] P. S. Letelier, Class. Quantum Grav. **12**, 471 (1995).
[4] R. Puntigam and H. Soleng, Volterra Distortions, Spinning Strings and Cosmic Defects, CERN preprint TH-96-26, (1996).
[5] P. S. Letelier, J. Math. Phys. **36**, 3043 (1995).
[6] P. S. Letelier and A. Wang, J. Math. Phys. **36**, 3023 (1995).
[7] A. Vilenkin, Phys. Lett. B **133**, 177 (1983).
[8] C. Köehler, Class. Quantum Grav. **12**, L11 (1995).
[9] A. Wang, J. Math. Phys. **32**, 2863 (1991).
[10] A. Wang, Phys. Rev. D **45**, 3534 (1992).
[11] P. Holvorcen and P. S. Letelier, Phys. Rev. D **49**, 6500 (1994).
[12] L. M. Widrow, Phys. Rev. D **39**, 3571 (1989).

[13] P. S. Letelier, Phys. Lett. A **79**, 283 (1980).
[14] R. Jackiw, ''Five Lectures on Planar Gravity,'' in *Proceedings of the 7th Latin American Symposium on Relativity and Gravitation-Silarg*, Cocoyoc, Mexico, 1990.
[15] R. Gleiser and J. Pullin, Class. Quantum Grav. **6**, L141 (1989).
[16] A. Trautman, Nature (London), Phys. Sci. **7**, 242 (1973).

Rotating dust solutions of Einstein's equations with 3-dimensional symmetry groups. I. Two Killing fields spanned on u^α and w^α

Andrzej Krasiński
N. Copernicus Astronomical Center and College of Science, Polish Academy of Sciences, Bartycka 18, 00 716 Warszawa, Poland

(Received 14 May 1997; accepted for publication 7 July 1997)

For a rotating dust with a 3-dimensional symmetry group all possible metric forms can be classified and, within each class, explicitly written out. This is made possible by the formalism of Plebański based on the Darboux theorem. In the resulting coordinates, the Killing vector fields (if any exist) assume a special form. Each Killing vector field may be either spanned on the fields of velocity and rotation or linearly independent of them. By considering all such cases one arrives at the classification. With respect to the structures of the groups, this is just the Bianchi classification, but with all possible orientations of the orbits taken into account. In this paper, which is part 1 of a 3-part series, all solutions are considered for which two Killing fields are spanned on velocity and rotation. The solutions of Lanczos and Gödel are identified as special cases, and their new invariant definitions are provided. In addition, a new invariant definition is given of the Ozsvath class III solution. © *1998 American Institute of Physics.* [S0022-2488(97)03112-5]

I. INTRODUCTION AND SUMMARY

The theorem of Darboux presented in Sec. II allows one to introduce invariantly defined coordinates in which the velocity field of a fluid (not assumed to be perfect) acquires a ''canonical'' form. In this paper it is further assumed that the fluid moves with zero acceleration and nonzero rotation. These assumptions result in a simplification of the metric tensor and in limitations imposed on the Killing vectors, if any exist. Within this special class of coordinates, any single Killing field may also be reduced to a ''canonical'' form, a different one in the case when it is spanned on the vector fields of velocity u^α and rotation w^α, and a different one when it is linearly independent of u^α and w^α. This gives rise to a classification of possible symmetries in rotating matter.

When there exist three linearly independent Killing fields, the classification described above gives rise to a complete classification of all possible metric forms. With respect to the algebras of the symmetry groups, this is just the Bianchi classification, but with all orientations of the orbits in the spacetime taken into account.

In every case that emerges, the commutation relations of the algebra have been solved, resulting in explicit formulae for the Killing fields, and then the Killing equations have been solved, resulting in the formulae for the metric tensors compatible with the symmetry groups considered. The degree of success in solving the Einstein equations varied very strongly from case to case. In most cases, no headway was made. In some cases the Einstein equations have been integrated either to an autonomous set of first-order equations or to a single nonlinear differential equation of second or third order. In a few cases solutions known earlier were identified in the present scheme and new invariant definitions for some of them were provided (those by Lanczos[1] and Gödel[2] will be mentioned in this paper). In three cases new solutions were found.

Since the number of cases is rather large, the results will be presented in three papers. The present paper deals with the simplest situation when two of the Killing fields are spanned on

0022-2488/98/39(1)/380/21/$15.00
© 1998 American Institute of Physics

velocity and rotation (the case of all three Killing fields being spanned on u and w is trivial—see Sec. VI).

The Darboux theorem was first applied as a tool for investigating the equations of motion and the Einstein equations by Plebański.[3] He showed that if a perfect fluid is rotating and isentropic while the particle number is conserved, then a similar consideration to the one presented here applies. The approach of Plebański was used by this author[4–8] to find a large collection of stationary, cylindrically symmetric solutions of Einstein's equations.

For a perfect fluid the assumptions of geodesic motion and nonzero rotation imply that the pressure is constant (see Ref. 9). Hence, from the point of view of thermodynamics, geodesic and rotating perfect fluids are isentropic and fall within the class considered by Plebański.[3] However, the approach based on the Darboux theorem applies to any timelike congruence that is of class C^1 and has zero acceleration and nonzero rotation. In particular, the velocity field of a rotating charged dust with zero Lorentz force, that was considered in several papers, has this property. The papers in which such solutions were discussed will be mentioned at the end of Sec. VIII; they are all within the same class of the classification introduced here.

In Sec. II the Darboux theorem is introduced. In Sec. III, the classification of first-order differential forms based on the Darboux theorem is applied to geodesic vector fields with rotation. When the vector field is the velocity field of a fluid, a class of preferred coordinates results, which shall be termed ''Plebański coordinates.'' In Sec. IV, by way of an example, the consideration of Sec. III is applied to the rotating dust solution of Stephani.[10] In Sec. V it is shown that each Killing vector field that might possibly exist in a rotating dust spacetime is determined by two functions of two variables. If the Killing field is not spanned on velocity and rotation, then the Plebański coordinates may be adapted to it so that it acquires the unique form $k^\alpha = \delta^\alpha_1$. The Gödel solution is used to illustrate the various forms of the Killing fields that may arise.

In Sec. VI, the consideration of Secs. III and V is applied to the situation when there exist three Killing vector fields. When all three of them are spanned on u^α and w^α, the result is trivial: the group becomes two-dimensional, and this case is not considered here. When two of them are spanned on u^α and w^α while the third one is not, two cases arise that correspond to different Bianchi types (II and I) of the groups. These are investigated in Secs. VII and VIII. The solutions of Lanczos[1] and of Gödel[2] emerge as special cases in both types. The Bianchi type II metrics are defined by a single third-order differential equation, the Bianchi type I metrics are determined by a set of autonomous first-order differential equations.

Finally, in Sec. IX, other invariant definitions are given: for the class III solution by Ozsvath[11] and for the solution of Gödel.[2] The former results from the following assumptions.

(1) The source in the Einstein equations is a geodesic, rotating perfect fluid.
(2) The rotation vector field is covariantly constant.

The Gödel solution, which is a subcase of this, emerges when it is assumed in addition that the shear of the fluid flow is zero.

So far, Bianchi-type solutions of Einstein's equations with a rotating source have been searched for and found by trial and error (often with nonperfect fluid sources, e.g., with heat-flow). The results of the present series of papers show that, in the case of a dust source at least, the number of allowed possibilities is limited. It is hoped that the results will direct further research toward better-defined targets.

II. THE CLASSIFICATION OF DIFFERENTIAL FORMS OF FIRST ORDER AND THE DARBOUX THEOREM

The Darboux theorem presented below exploits the simple fact that if a differential form q of first order is defined on an n-dimensional manifold M_n, then its domain is not necessarily n-dimensional. Two cases are well-known.

(1) If q is a perfect differential so that $q=df$, where f is a scalar function on M_n, then f can be chosen as one of the coordinates, and the form becomes one-dimensional.
(2) If q has an integrating factor so that $q=gdf$, where f and g are independent scalar functions on M_n, then f and g can be chosen as two of the coordinates, and then the domain of q is the two-dimensional (f,g)-surface.

The Darboux theorem summarizes all the cases that can occur. It is based on the following classification (see also Ref. 3)

Definition: Let q be a differential form of first order.
If $Q_{2l}:=dq\wedge\cdots\wedge dq$ (multiplied l times) $\neq 0$, but $q\wedge Q_{2l}=0$, then q is said to be of class $2l$.
If $Q_{2l+1}:=q\wedge Q_{2l}\neq 0$, but $dQ_{2l+1}\equiv dq\wedge Q_{2l}=0$, then q is said to be of class $(2l+1)$. □
Then the following holds.

The theorem of Darboux: The form q is of class $2l$ if and only if there exists a set of $2l$ independent functions $(\xi_1,\ldots,\xi_l,\eta_1,\ldots,\eta_l)$ such that

$$q=\eta_1 d\xi_1+\eta_2 d\xi_2+\cdots+\eta_l d\xi_l. \tag{2.1}$$

The form q is of class $(2l+1)$ if and only if there exists a set of $(2l+1)$ independent functions $(\tau,\xi_1,\ldots,\xi_l,\eta_1,\ldots,\eta_l)$ such that

$$q=d\tau+\eta_1 d\xi_1+\eta_2 d\xi_2+\cdots+\eta_l d\xi_l. \tag{2.2}$$

A proof of this theorem can be found in Ref. 12.

Evidently, the class of q cannot be larger than the dimension of the manifold on which q is defined. Hence, the Darboux theorem implies that in a four-dimensional spacetime V_4 any differential form of first order can be represented as

$$q=\sigma d\tau+\eta d\xi, \tag{2.3}$$

where σ, τ, η and ξ are scalar functions on V_4.

Any vector field u^α on V_4 defines the following differential form:

$$q_u:=u_\alpha dx^\alpha. \tag{2.4}$$

According to (2.3), in the most general case there exist scalar functions σ, τ, η and ξ such that

$$u_\alpha=\sigma\tau_{,\alpha}+\eta\xi_{,\alpha}. \tag{2.5}$$

Note that the functions in (2.5) are not uniquely defined. Since we shall not use (2.5) in the most general case, we shall determine the nonuniqueness only in the subcase that is of direct interest to us (see Sec. III).

For the most general case of (2.5), the four functions are independent in the sense that the Jacobian,

$$\frac{\partial(\sigma,\tau,\eta,\xi)}{\partial(x^0,x^1,x^2,x^3)}\neq 0. \tag{2.6}$$

Hence, they can be chosen as coordinates in the spacetime. In Refs. 4 and 7 it was shown that if u^α is the velocity field of an isentropic perfect fluid in which the particle number is conserved, then $\sigma=1/H$, where H is the enthalpy per one particle of the fluid, and further limitations on u_α follow from the particle number conservation. No other applications of (2.5) in the general case are known to this author.

III. GEODESICALLY MOVING FLUIDS

To any timelike vector field u_α normalized to unity (so that $u_\alpha u^\alpha = 1$) the formula from Refs. 13 and 14 may be applied:

$$u_{\alpha;\beta} = \dot{u}_\alpha u_\beta + \sigma_{\alpha\beta} + \omega_{\alpha\beta} + \tfrac{1}{3}\,\theta h_{\alpha\beta}, \tag{3.1}$$

which gives rise to the well-known definitions of acceleration $\dot{u}^\alpha$, expansion θ, shear $\sigma_{\alpha\beta}$ and rotation $\omega_{\alpha\beta}$. In the signature $(+---)$ used here, the projection tensor $h_{\alpha\beta}$ is

$$h_{\alpha\beta} = g_{\alpha\beta} - u_\alpha u_\beta. \tag{3.2}$$

The following properties of $\dot{u}^\alpha$, $\sigma_{\alpha\beta}$ and $\omega_{\alpha\beta}$ will be useful in further considerations:

$$\dot{u}_\alpha u^\alpha = 0, \quad \sigma_{\alpha\beta} u^\beta = \omega_{\alpha\beta} u^\beta = 0. \tag{3.3}$$

We shall assume from now on that u_α is the velocity field of a fluid and that $\dot{u}_\alpha = 0$, i.e., that the particles of the fluid move on geodesics. Then, from (2.5) we have

$$\omega_{\alpha\beta} = \tfrac{1}{2}\,(u_{\alpha,\beta} - u_{\beta,\alpha}) = \tfrac{1}{2}\,(\sigma_{,\beta}\tau_{,\alpha} - \sigma_{,\alpha}\tau_{,\beta} + \eta_{,\beta}\xi_{,\alpha} - \eta_{,\alpha}\xi_{,\beta}), \tag{3.4}$$

and from (3.3) we have

$$(u^\beta \sigma_{,\beta})\tau_{,\alpha} - (u^\beta \tau_{,\beta})\sigma_{,\alpha} + (u^\beta \eta_{,\beta})\xi_{,\alpha} - (u^\beta \xi_{,\beta})\eta_{,\alpha} = 0. \tag{3.5}$$

There are two possibilities now.

I. At least one of the four scalar products in (3.5) is nonzero. In this case (3.5) implies that at most three of the functions (σ,τ,η,ξ) are independent, and so the form (2.4) will not be of class 4.

II. All the four scalar products are zero. However, this means that the gradients of (σ,τ,ξ,η) are at every point confined to the 3-space element orthogonal to u_α, i.e., that there is a functional relation among these four functions. Again, the form (2.4) cannot be of class 4.

Hence, for a geodesically moving fluid the form (2.4) is of class at most 3, i.e., at most 3 independent functions τ, η, ξ exist such that

$$u_\alpha = \tau_{,\alpha} + \eta\xi_{,\alpha}. \tag{3.6}$$

From here on, the reasoning used in Refs. 3 and 4 applies almost unchanged. With (3.6) we have in (3.4),

$$\omega_{\alpha\beta} = \tfrac{1}{2}\,(\eta_{,\beta}\xi_{,\alpha} - \eta_{,\alpha}\xi_{,\beta}), \tag{3.7}$$

and in (3.5),

$$(u^\beta \eta_{,\beta})\xi_{,\alpha} - (u^\beta \xi_{,\beta})\eta_{,\alpha} = 0. \tag{3.8}$$

There are again two possibilities.

I. Either $(u^\beta \eta_{,\beta})$ and $(u^\beta \xi_{,\beta})$ do not vanish simultaneously, and then (3.8) implies that η and ξ are functionally related, in which case (3.6) implies that u_α is a gradient of a function, and so $\omega_{\alpha\beta} \equiv 0$.

II. Or ξ and η are not functionally related, in which case

$$u^\beta \xi_{,\beta} = u^\beta \eta_{,\beta} = 0, \tag{3.9}$$

and $\omega_{\alpha\beta} \neq 0$. We shall be interested only in the second case. The functions $\{\tau,\xi,\eta\}$ in (3.6) are determined up to the following transformations:

$$\xi=F(\xi',\eta'),\quad \eta=G(\xi',\eta'),\quad \tau=\tau'-S(\xi',\eta'), \tag{3.10}$$

where the functions F and G must obey the equation

$$F_{,\xi'}G_{,\eta'}-F_{,\eta'}G_{,\xi'}=1, \tag{3.11}$$

and then S is determined by

$$S_{,\xi'}=GF_{,\xi'}-\eta',\quad S_{,\eta'}=GF_{,\eta'}. \tag{3.12}$$

Equation (3.11) is the integrability condition of Eqs. (3.12) and it ensures that the Jacobian of the transformation (3.10) equals 1. It follows that one of the functions $\{F,G\}$ can be chosen arbitrarily, the other one is then determined by (3.11) and S is fixed up to an additive constant by (3.12). The inverse transformation to (3.10) is of exactly the same form, with the corresponding F, G and S obeying (3.11) and (3.12).

Let us now make the additional assumption that the number of particles of the fluid is conserved, i.e.,

$$(\sqrt{-g}nu^{\alpha})_{,\alpha}=0, \tag{3.13}$$

where g is the determinant of the metric tensor and n is the particle number density. This equation is a necessary and sufficient condition for the existence of a function ζ such that

$$\sqrt{-g}nu^{\alpha}=\epsilon^{\alpha\beta\gamma\delta}\xi_{,\beta}\eta_{,\gamma}\zeta_{,\delta}. \tag{3.14}$$

Note that (3.6) and (3.9) imply that

$$u^{\alpha}\tau_{,\alpha}=1, \tag{3.15}$$

and then Eq. (3.14) implies that

$$\epsilon^{\alpha\beta\gamma\delta}\tau_{,\alpha}\xi_{,\beta}\eta_{,\gamma}\zeta_{,\delta}\equiv\frac{\partial(\tau,\eta,\xi,\zeta)}{\partial(x^0,x^1,x^2,x^3)}=\sqrt{-g}n\neq 0. \tag{3.16}$$

Equation (3.14) implies also that

$$u^{\alpha}\zeta_{,\alpha}=0. \tag{3.17}$$

The function ζ is determined by (3.14) up to the transformations

$$\zeta=\zeta'+T(\xi',\eta'), \tag{3.18}$$

where T is an arbitrary function. Equation (3.16) certifies that $\{\tau,\xi,\eta,\zeta\}$ can be used as coordinates in the spacetime. If they are chosen as the $\{x^0,x^1,x^2,x^3\}=\{t,x,y,z\}$ coordinates, respectively, then Eq. (3.6) implies

$$u_0=1,\quad u_1=y,\quad u_2=u_3=0. \tag{3.19}$$

We will use these coordinates throughout the remaining part of the paper and call them ''Plebański coordinates.'' Equation (3.16) implies now that

$$g=-n^{-2}, \tag{3.20}$$

and Eq. (3.14) implies

$$u^{\alpha}=\delta^{\alpha}_0, \tag{3.21}$$

i.e., the Plebański coordinates are comoving. The rotation vector defined by

$$w^\alpha = -(1/\sqrt{-g})\,\epsilon^{\alpha\beta\gamma\delta}u_\beta u_{\gamma,\delta}, \tag{3.22}$$

assumes the form

$$w^\alpha = n\,\delta^\alpha_3. \tag{3.23}$$

Equations (3.19) and (3.21) imply that

$$g_{00}=1, \quad g_{01}=y, \quad g_{02}=g_{03}=0, \tag{3.24}$$

and also that the only nonvanishing components of the rotation tensor are

$$\omega_{12}=-\omega_{21}=1/2. \tag{3.25}$$

Note that, in contrast to Refs. 3 and 4 where barotropic perfect fluids were considered, we have not assumed anything about the form of the energy–momentum tensor so far.

If we now assume that the fluid is perfect, then we conclude from the equations of motion $T^{\alpha\beta}{}_{;\beta}=0$ that either $\omega=0$ or $p=\text{const}$ (see also Ref. 9). This means that a geodesic perfect fluid can be rotating only if it is in fact dust; the constant p can be reinterpreted as the cosmological constant. In this case, the energy–density obeys the conservation equation $(\sqrt{-g}\epsilon u^\alpha)_{,\alpha}=0$ and Eq. (3.13) need not be assumed separately. [For dust, results closely analogous to (3.19)–(3.25) were obtained by Ellis,[15] by adapting an orthonormal vector basis and a coordinate system to u^α and w^α. Of the exact solutions with nonzero rotation found by Ellis most, but not all, do belong to the collection considered in this series of papers. They will be described in paper 2.]

IV. EXAMPLE: THE STEPHANI SOLUTION[10]

The Stephani metric with $p=\text{const}\neq 0$ [Eq. (4.22) in Ref. 10] is not in fact a perfect fluid solution, as was found out while trying to construct the Plebański coordinates for it. (The error is deeply hidden and so far could not be corrected. I am grateful to H. Stephani for cooperation on this point.) Therefore, we shall consider only the dust solution, Eq. (4.8) in Ref. 10. In the original notation except for the signature, the solution is

$$ds^2=\eta_{ab}dx^a dx^b - N^2(dx^1)^2, \tag{4.1}$$

where $a,b=0,2,3$, $\eta_{ab}=\text{diag}(1,-1,-1)$ and

$$N:=\tfrac{1}{2}M\,\ln T + g_a x^a + h,$$

$$T^2:=\eta_{ab}(x^a-f^a)(x^b-f^b), \tag{4.2}$$

the functions $M(x^1)$, $f^a(x^1)$, $g_a(x^1)$, and $h(x^1)$ all being arbitrary. The velocity field and the energy–density of the dust are, respectively,

$$u_a=T_{,a}, \quad u_1=0, \tag{4.3}$$

$$\kappa\epsilon = M/(NT^2). \tag{4.4}$$

The formula for the velocity field can be written as

$$u_\alpha = T_{,\alpha} - T_{,1}x^1{}_{,\alpha}, \tag{4.5}$$

which immediately suggests the choice of the Plebański coordinates of Sec. III:

$$\tau=T, \quad \xi=x^1, \quad \eta=-T_{,1}. \tag{4.6}$$

Equation (3.14) defining ζ is here

$$MT^{-2}\eta^{ab}T_{,b}=-\epsilon^{a1cd}T_{,1c}\zeta_{,d}. \tag{4.7}$$

The following identity is useful in calculations:

$$T_{,0}{}^2-T_{,2}{}^2-T_{,3}{}^2=1. \tag{4.8}$$

Using this, one can verify that only two of the three equations (4.7) are independent. The solution of (4.7) is

$$\zeta=M[U+\lambda(T_{,1})], \tag{4.9}$$

where $\lambda(T,_1)$ is an arbitrary function, and U is defined by

$$U=\int [f^3{}_{,1}(1-X^2)-f^0{}_{,1}Y(X)+f^2{}_{,1}XY(X)]^{-1}dX, \tag{4.10}$$

the function $Y(X)$ being determined by

$$T_{,1}(1-X^2-Y^2)^{1/2}=-f^0{}_{,1}+f^2{}_{,1}X+f^3{}_{,1}Y. \tag{4.11}$$

In the integral (4.10), the coordinate x^1 and the quantity $T,_1$ are to be treated as parameters independent of X. After the integral is calculated, one should substitute for $T,_1$ from (4.11), while X and Y are to be replaced by

$$X=(x^2-f^2)/(x^0-f^0), \quad Y=(x^3-f^3)/(x^0-f^0). \tag{4.12}$$

The integral in (4.10) is expressible in terms of elementary functions, but the result is rather complicated.

As can be seen, the solution (4.1)–(4.4) becomes very complicated in the Plebański coordinates and it is unlikely that it could be found in such a form from the Einstein equations. Thus, although the coordinates are invariantly defined, they are not necessarily convenient for explicit calculations.

A collection of other solutions represented in the Plebański coordinates can be found in the extended version of Ref. 6.

V. THE KILLING VECTOR FIELDS COMPATIBLE WITH ROTATION

We shall assume that the symmetries of the spacetime (if any exist) are inherited by the source, i.e., that if the Lie derivative of the metric tensor $g_{\alpha\beta}$ along the vector field k^α is zero, $\mathcal{L}_k g_{\alpha\beta}=0$, then the velocity field and the particle number density are also invariant: $\mathcal{L}_k u^\alpha=0=\mathcal{L}_k n$. (For a pure perfect fluid source the inheritance is guaranteed.) It follows that the rotation tensor must also be invariant, $\mathcal{L}_k\omega_{\alpha\beta}=0$.

In consequence of (3.21) the equation $\mathcal{L}_k u^\alpha\equiv k^\mu u^\alpha{}_{,\mu}-u^\mu k^\alpha{}_{,\mu}=0$ implies that

$$k^\alpha{}_{,t}=0. \tag{5.1}$$

In consequence of (3.23) and of the assumption $\mathcal{L}_k n\equiv k^\alpha n_{,\alpha}=0$, the equation $\mathcal{L}_k w^\alpha=0$ implies

$$k^\alpha{}_{,z}=0. \tag{5.2}$$

The equation $\mathcal{L}_k\omega_{\alpha\beta}=0$, in consequence of (3.25) implies

$$k^1{}_{,x}+k^2{}_{,y}=0, \tag{5.3}$$

and the equation $\mathcal{L}_k u_\alpha = 0$, in consequence of (3.19) implies

$$k^0{}_{,x} = -k^2 - y k^1{}_{,x}, \quad k^0{}_{,y} = -y k^1{}_{,y}. \tag{5.4}$$

(The equations $\mathcal{L}_k u_\alpha = 0$ and $\mathcal{L}_k u^\alpha = 0$ provide independent pieces of information because the equations $\mathcal{L}_k g_{\alpha\beta} = 0$ have not been used so far.) Equation (5.3) is the integrability condition of Eqs. (5.4). The general solution of Eqs. (5.1)–(5.4) is

$$k^0 = C + \phi - y\phi_{,y}, \quad k^1 = \phi_{,y}, \quad k^2 = -\phi_{,x}, \quad k^3 = \lambda, \tag{5.5}$$

where $\phi(x,y)$ and $\lambda(x,y)$ are arbitrary functions and C is an arbitrary constant. Symmetries need not be present, in fact the Stephani[10] solution considered in Sec. IV is an example of a rotating dust solution with no symmetries. In this case $\phi = \lambda = C = 0$. However, if any symmetries are present, then the Killing vector fields must have the form (5.5).

Suppose that ϕ is not a constant, i.e., that a Killing vector field k^α exists that has a nonzero component in the x- or y-direction (in invariant terms this means that the vector field k^α is not spanned on the vector fields of velocity, u^α, and rotation, w^α). We can then, within the Plebański class defined in Sec. III, adapt the coordinates to k^α in such a way that $k^{\alpha'} = \delta_1^{\alpha'}$, i.e., so that the metric becomes independent of x'. From (3.10)–(3.12) and (3.18) the transformation functions are

$$t' = t - S(x,y), \quad x' = F(x,y), \quad y' = G(x,y), \quad z' = z + T(x,y), \tag{5.6}$$

where T is arbitrary, while F, G, and S obey

$$F_{,x}G_{,y} - F_{,y}G_{,x} = 1, \quad S_{,x} = GF_{,x} - y, \quad S_{,y} = GF_{,y}. \tag{5.7}$$

In order to lead to $k^{\alpha'} = \delta_1^{\alpha'}$ the functions F, G, and T must obey in addition

$$-(\phi + C) + GF_{,x}\phi_{,y} - GF_{,y}\phi_{,x} = 0,$$

$$F_{,x}\phi_{,y} - F_{,y}\phi_{,x} = 1, \quad G_{,x}\phi_{,y} - G_{,y}\phi_{,x} = 0, \tag{5.8}$$

$$T_{,x}\phi_{,y} - T_{,y}\phi_{,x} = -\lambda. \tag{5.9}$$

The unique solution of Eqs. (5.8) is $G = \phi + C$, which obeys (5.7) as well [in virtue of the second of (5.8)]. Equation (5.9) simply defines the accompanying T which is seen to exist always. Since ϕ was assumed nonconstant, the transformation is nonsingular (in fact its Jacobian equals just 1), and results in $\phi = y$ in the new coordinates. As already noticed, the metric becomes independent of x after the transformation. This property is preserved by the transformations (5.6), but with F, G, S, and T restricted now by

$$G = y, \quad F = x + H(y), \quad T = T(y), \quad S = \int y H_{,y} dy + A, \tag{5.10}$$

where A is an arbitrary constant and H, T are arbitrary functions. The functions given by (5.10) fulfill (5.7) identically. Note that the transformation to $k^{\alpha'} = \delta_1^{\alpha'}$ exists irrespectively of any possible functional relation among λ, ϕ, and C in (5.5); the only case when it fails is $\phi = \text{const}$.

A solution of the Einstein equations may have more than one Killing vector field of the form (5.5). In that case, the transformation (5.8)–(5.9) changes only one of them to the preferred form, the others will preserve their more complicated appearance. An example of this situation is the Gödel solution[2] transformed to the Plebański coordinates; see Refs. 4 and 5:

$$ds^2 = (dt + y dx)^2 - \tfrac{1}{2} y^2 dx^2 - (\kappa\epsilon y^2)^{-1} dy^2 - 2\kappa\epsilon^{-1} dz^2, \tag{5.11}$$

where $\kappa = 8\pi G/c^4$ and ϵ is the energy–density related to the cosmological constant by

$$\Lambda = \tfrac{1}{2}\kappa\epsilon. \tag{5.12}$$

(Note that if the cosmological constant is reinterpreted as pressure, then the resulting perfect fluid has the equation of state $\epsilon = p$. Hence, the Gödel solution may have been the first example considered in the literature of a ''stiff perfect fluid,'' now familiar from the studies of solution-generating techniques; see, e.g., Verdaguer.[16]) The symmetry group of this solution is 5-dimensional, the independent 1-parametric subgroups were given in Ref. 4. Those connected with nonconstant ϕ in (5.5) are the following three:

$$x = x' + s_1, \tag{5.13}$$

$$x = e^{-s_2}x', \quad y = e^{s_2}y', \tag{5.14}$$

$$\begin{aligned} t &= t' + (2\sqrt{2}/K)\arctan[\sqrt{2}s_3(Ky')^{-1}(1-s_3x')^{-1}], \\ x &= [K^2x'y'^2(1-s_3x') - 2s_3]/[2s_3^2 + K^2y'^2(1-s_3x')^2], \\ y &= (1-s_3x')^2y' + 2s_3^2/(K^2y'), \end{aligned} \tag{5.15}$$

where s_1, s_2, and s_3 are the group parameters and $K := (\kappa\epsilon)^{1/2}$. The Killing vectors are, respectively, $k^\alpha_{(1)} = \delta^\alpha_1$ [corresponding to $C = \lambda = 0$, $\phi = y$, the one constructed in (5.8)–(5.9)], $k^\alpha_{(2)} = -x\delta^\alpha_1 + y\delta^\alpha_2$ (corresponding to $\phi = -xy$) and $k^\alpha_{(3)} = 4(K^2y)^{-1}\delta^\alpha_0 + [x^2 - 2/(Ky)^2]\delta^\alpha_1 - 2xy\delta^\alpha_2$ [corresponding to $\phi = x^2y + 2/(K^2y)$].

VI. THE ALGEBRA OF THREE KILLING FIELDS

Suppose that three Killing vector fields exist and all three are spanned on u^α and w^α, so that $\phi = \text{const}$ in (5.5) for each of them, i.e.,

$$k^\alpha_{(i)} = C_i\delta^\alpha_0 + \lambda_i(x,y)\,\delta^\alpha_3, \quad i = 1,2,3. \tag{6.1}$$

From the Killing equations one can then easily conclude that constants α_1, α_2, and α_3 exist such that $\alpha_1 k_{(1)} + \alpha_2 k_{(2)} + \alpha_3 k_{(3)} = 0$, i.e., the symmetry group is in fact two-dimensional. Hence, no three-dimensional symmetry group with the generators (6.1) exists; for a three-dimensional group at least one of the generators must be linearly independent of u^α and w^α at every point of the spacetime region under consideration. [The algebra (6.1) corresponds to a three-dimensional group that has two-dimensional orbits, and it turns out that in the case considered the group has to be two-dimensional as well. As will follow from the whole of the present work, three-dimensional symmetry groups with two-dimensional orbits just do not exist for rotating dust.]

In Secs. VII and VIII we shall consider the situation when exactly one generator, $k^\alpha_{(1)}$, is everywhere linearly independent of u^α and w^α, while the other two, $k^\alpha_{(2)}$ and $k^\alpha_{(3)}$, are of the form (6.1). In agreement with the result of Sec. V, the Plebański coordinates can be adapted to $k^\alpha_{(1)}$ so that

$$k^\alpha_{(1)} = \delta^\alpha_1, \tag{6.2}$$

while

$$k^\alpha_{(2)} = C_2\delta^\alpha_0 + \lambda_2(x,y)\,\delta^\alpha_3, \quad k^\alpha_{(3)} = C_3\delta^\alpha_0 + \lambda_3(x,y)\,\delta^\alpha_3, \tag{6.3}$$

and the coordinate transformations preserving (6.2) and (6.3) are (5.10). Note that C_2 and C_3 cannot vanish simultaneously because otherwise the Killing equations immediately imply that either $k^\alpha_{(3)} = \text{const}\, k^\alpha_{(2)}$ (in which case the symmetry group is two-dimensional) or the metric is singular. However, with no loss of generality we can assume that

$$C_2 \neq 0 = C_3, \tag{6.4}$$

because the Killing vector fields are determined up to linear combinations among them. Hence, if initially $C_2 \neq 0 \neq C_3$, then we take $k'^\alpha_{(3)} = k^\alpha_{(3)} - (C_3/C_2)k^\alpha_{(2)}$ instead of $k^\alpha_{(3)}$ as the basis generator. If initially $C_2 = 0 \neq C_3$, then we exchange the labels ''2'' and ''3.''

We shall keep the choice (6.4) throughout Secs. VII and VIII.

VII. THE GENERATORS, THE KILLING EQUATIONS AND THE EINSTEIN EQUATIONS FOR A BIANCHI TYPE II ALGEBRA

All the cases that arise follow as limits at different stages of calculation from the generic case $\lambda_2 \neq 0$, and we shall consider it first. The commutators of the Killing vectors are

$$[k_{(1)}, k_{(2)}]^\alpha = (\lambda_{2,x}/\lambda_3)k^\alpha_{(3)}, \quad [k_{(2)}, k_{(3)}]^\alpha = 0,$$
$$[k_{(1)}, k_{(3)}]^\alpha = (\lambda_{3,x}/\lambda_3)k^\alpha_{(3)}. \tag{7.1}$$

The Killing vector fields will thus form a Lie algebra when

$$\lambda_{2,x} = b\lambda_3, \quad \lambda_{3,x} = c\lambda_3, \tag{7.2}$$

where b and c are arbitrary constants. The case $c \neq 0$ has to be considered separately. Then

$$\lambda_3 = \beta(y)e^{cx}, \quad \lambda_2 = (b/c)\beta(y)e^{cx} + \alpha(y), \tag{7.3}$$

where $\alpha(y)$ and $\beta(y)$ are arbitrary functions. However, in this case we can take $k'^\alpha_{(2)} = k^\alpha_{(2)} - (b/c)k^\alpha_{(3)}$ as the new basis generator instead of $k^\alpha_{(2)}$, and the result is equivalent to assuming $b = 0$. Hence, with $c \neq 0$, we can take $b = 0$ with no loss of generality.

The further procedure consists of the following steps.

(1) Adapt the coordinates to the Killing fields to make them as simple as possible.
(2) Solve the Killing equations for the components of the metric tensor.
(3) Simplify the metric as far as possible by coordinate transformations.
(4) Solve the Einstein equations.

The coordinate transformations in steps 1 and 3 in general lead out of the Plebański class defined in Sec. III.

This procedure will be presented in some detail below. In the present case the result from the Einstein equations is that either $c = 0$ or there is no rotation. Since we are interested in rotating solutions only, the case $c \neq 0 = b$ need not be followed further. We thus assume

$$c = 0. \tag{7.4}$$

Then

$$\lambda_3 = \beta(y), \quad \lambda_2 = b\beta(y)x + \alpha(y). \tag{7.5}$$

The algebra of the Killing vector fields is of Bianchi type II when $b \neq 0$ and of Bianchi type I when $b = 0$.

In order to simplify the Killing vectors we now transform the coordinates as follows:

$$(t',x',y')=(t,x,y), \quad z'=-(\alpha/C_2)t+z/\beta. \tag{7.6}$$

The transformation is not of the form (5.10), so the new coordinates do not belong to the Plebański class, and the forms of velocity, rotation, and the metric will no longer agree with (3.19)–(3.25). The Killing vector fields in the new coordinates become

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=\delta^\alpha_0+bx\delta^\alpha_3, \quad k^\alpha_{(3)}=\delta^\alpha_3, \tag{7.7}$$

while the velocity and rotation fields become

$$u^\alpha=\delta^\alpha_0-(\alpha/C_2)\delta^\alpha_3, \quad w^\alpha=(n/\beta)\delta^\alpha_3. \tag{7.8}$$

The transformed metric is independent of x and z, while the Killing equations for $k^\alpha_{(2)}$ imply

$$\begin{gathered}
g_{00}=1+(\alpha/C_2)^2h_{33}(y), \quad g_{01}=y+(\alpha/C_2)g_{13}, \\
g_{02}=0, \quad g_{03}=(\alpha/C_2)h_{33}(y), \\
g_{11}=h_{33}(y)(bt)^2-2h_{13}(y)bt+h_{11}(y), \\
g_{12}=-h_{23}(y)bt+h_{12}(y), \quad g_{13}=-h_{33}(y)bt+h_{13}(y), \\
g_{22}=h_{22}(y), \quad g_{23}=h_{23}(y), \quad g_{33}=h_{33}(y),
\end{gathered} \tag{7.9}$$

where $h_{ij}(y)$, $i,j=1,2,3$, are arbitrary functions of y, to be found from the Einstein equations.

The orbits of the symmetry group are now the hypersurfaces $y=\text{const}$. In order to follow the standard technique of the Bianchi-type spaces we should now carry out a coordinate transformation that preserves (7.7) and makes the y-coordinate curves orthogonal to the group orbits, so that $g'_{02}=g'_{12}=g'_{23}=0$ after the transformation. This step is not in fact necessary for solving the Einstein equations (in general it only reshuffles the unknown functions without eliminating any of them), but in the case under consideration it leads to a simplification. The transformation is

$$\begin{gathered}
t=t'+f_0(y'), \quad x=x'+f_1(y'), \quad y=f_2(y'), \\
z=z'+bf_1(y')t'+f_3(y'),
\end{gathered} \tag{7.10}$$

where $f_\alpha(y')$ must obey

$$\begin{gathered}
f_{0,y'}=-f_2f_{1,y'}+(\alpha/C_2)h_{23}f_{2,y'}, \\
(-f_2^2+h_{11}-h_{13}^2/h_{33})f_{1,y'}+(h_{12}-h_{13}h_{23}/h_{33}+\alpha f_2h_{23}/C_2)f_{2,y'}=0, \\
f_{3,y'}=-(\alpha/C_2)f_{0,y'}+(bf_0-h_{13}/h_{33})f_{1,y'}-(h_{23}/h_{33})f_{2,y'}.
\end{gathered} \tag{7.11}$$

The equations are well-defined because of the following.

I. $h_{33}\neq 0$; otherwise the rotation vector would be null, which is a physical impossibility.

II. $-f_2^2+h_{11}-h_{13}^2/h_{33}\neq 0$; otherwise the determinant of the metric tensor becomes positive; i.e., the metric acquires an unphysical signature.

Equations (7.11) are to be understood as follows. The function $f_2(y')$ can be chosen arbitrarily, therefore we choose it so that $g'_{22}=-1$ after the transformation. With $f_2(y')$ thus chosen, $f_1(y')$ is found from the second of (7.11), then $f_0(y')$ is found from the first of (7.11), and finally $f_3(y')$ is found from the third of (7.11).

After the transformation the metric becomes (primes dropped)

$$
\begin{gathered}
g_{00}=1+(\alpha/C_2+bf_1)^2h_{33},\\
g_{01}=Y(y)+(\alpha/C_2+bf_1)(-h_{33}bt+h_{13}-bf_0h_{33}),\\
g_{02}=g_{12}=g_{23}=0,\\
g_{03}=(\alpha/C_2+bf_1)h_{33},\\
g_{11}=h_{33}b^2(t+f_0)^2-2h_{13}b(t+f_0)+h_{11},\\
g_{13}=-h_{33}b(t+f_0)+h_{13},\\
g_{22}=-1,\quad g_{33}=h_{33}(y),
\end{gathered}
\tag{7.12}
$$

where $h_{ij}(y)$, $i,j=1,2,3$, $f_0(y), f_1(y)$, $Y(y)=f_2(y)$ and $\alpha(y)$ are functions to be found from the Einstein equations, and b and C_2 are arbitrary constants, $C_2\neq 0$.

For convenience in calculations we introduce the new functions $G(y)$, $A(y)$, $k_{13}(y)$, and $F(y)$ by

$$
\begin{gathered}
g_{33}=-G^2(y),\quad \alpha/C_2+bf_1=A(y),\quad h_{13}=-G^2(k_{13}+bf_0),\\
h_{11}=Y^2-F^2-k_{13}^2G^2+b^2G^2f_0^2-2bf_0k_{13}G^2.
\end{gathered}
\tag{7.13}
$$

The velocity field in the coordinates of (7.12)–(7.13) is

$$
u^\alpha=\delta^\alpha_0-A\delta^\alpha_3. \tag{7.14}
$$

Now the metric form is:

$$
ds^2=(dt+Ydx)^2-(Fdx)^2-dy^2-G^2[Adt-(bt-k_{13})dx+dz]^2. \tag{7.15}
$$

The components of the Einstein tensor will be referred to the orthonormal tetrad of forms $e^i=e^i_\alpha dx^\alpha$, $i=0,1,2,3$, uniquely implied by (7.15). Note that $e^0=u_\alpha dx^\alpha$. Hence, the Einstein equations are

$$
\begin{gathered}
G_{00}=\kappa\epsilon,\\
G_{11}=G_{22}=G_{33}=\Lambda,\quad G_{ij}=0,\quad \text{when } i\neq j,
\end{gathered}
\tag{7.16}
$$

where ϵ is the energy–density and Λ is the cosmological constant.

The equation $G_{12}=0$ implies that $bA_{,y}=0$. The case $b=0$ will be considered separately below, so we take here

$$
A=\text{const.} \tag{7.17}
$$

Then $G_{02}=0$ implies

$$
k_{13}=\text{const.} \tag{7.18}
$$

We can then carry out the coordinate transformation:

$$
z=z'-At-k_{13}x,\quad (t,x,y)=(t',x',y'), \tag{7.19}
$$

which has the same result as if

$$A = k_{13} = 0, \tag{7.20}$$

and we shall assume this from now on. The metric is still independent of x and of z, while $A = k_{13} = 0$ implies $g_{13} = 0$, i.e., the Killing vectors $k^{\alpha}_{(1)} = \delta^{\alpha}_{1}$ and $k^{\alpha}_{(3)} = \delta^{\alpha}_{3}$ are orthogonal to each other. The equation $G_{01} = 0$ then has the integral

$$Y_{,y} G/F = B = \text{const}, \tag{7.21}$$

and we can assume $B \neq 0$ because rotation would be zero with $B = 0 = Y_{,y}$.

At this point, only the diagonal components of the Einstein tensor are still nonzero, of which G_{00} just defines the energy–density, and the other three are functionally dependent (i.e., if $G_{11} = \Lambda = G_{22}$ are fulfilled, then so is $G_{33} = \Lambda$). They determine $F(y)$ and $G(y)$.

It is convenient to introduce $Y(y)$ as the new variable. The equation $G_{11} + G_{22} = 2\Lambda$ can then be written, with the help of (7.21), as

$$(F^2 G_{,Y}/G)_{,Y} = 2\Lambda G^2/B^2 - \tfrac{1}{2}, \tag{7.22}$$

and so

$$F^2 = \left(C - \frac{1}{2} Y + 2 \frac{\Lambda}{B^2} \int G^2 dY \right) G/G_{,Y}, \tag{7.23}$$

where C is a new arbitrary constant (we can assume $G_{,Y} \neq 0$ because $G_{,Y} = 0$ implies $b = 0$ from $G_{11} - G_{22} = 0$, and $b = 0$ will be considered separately). Using (7.23) in $G_{22} = \Lambda$ we obtain the following integro-differential equation that determines G:

$$-\frac{1}{4} b^2 G G_{,Y} + \frac{1}{2} (B/G)^2 \left(C - \frac{1}{2} Y + 2 \frac{\Lambda}{B^2} \int G^2 dY \right)^2 (G_{,Y}/G - G_{,YY}/G_{,Y}) = 0. \tag{7.24}$$

In the special case $\Lambda = 0$ this becomes an ordinary second-order differential equation. It is easy to get rid of the integral by transforming (7.24) appropriately and differentiating the result by Y [in this way a third-order differential equation for $G(Y)$ is obtained] or by introducing the new variable $u(Y)$ by $dY/du = 1/G^2$ [this results in a second-order equation for $G(u)$]. However, no progress toward solving (7.24) results in either case.

With the help of the equations $G_{11} = \Lambda = G_{22}$ the formula for energy–density may be simplified to

$$\kappa \epsilon = (B/G)^2 - (bG)^2 - 2\Lambda. \tag{7.25}$$

Note that the solutions considered here have a meaningful limit $b = 0$.

When $G = \text{const}$, Eqs. (7.23) and (7.24) no longer apply and one has to go back to the Einstein equations. They imply

$$G^2 = B^2/(4\Lambda) \tag{7.26}$$

(i.e., necessarily $\Lambda > 0$) and

$$F^2 = \tfrac{1}{2} Y^2 + DY + E, \tag{7.27}$$

where D and E are constants. If Y is chosen as the new coordinate in place of y, then from (7.21) and (7.26) the metric component g_{YY} is

$$g_{YY} = -(G/BF)^2 = -1/(4\Lambda F^2), \tag{7.28}$$

and the resulting metric is the Gödel solution (see Ref. 4). Note that $G=\text{const}$ is equivalent to $\epsilon=\text{const}$; see Eq. (7.25).

When $G_{,Y}\neq 0=b$, Eq. (7.24) implies $G=e^{DY+E}$, and this leads to the Lanczos solution (see Ref. 4).

These derivations of the Lanczos and Gödel solutions lead to their invariant definitions that are based on weaker assumptions than the definitions known so far. The definitions are:

(1) The source in the Einstein equations is a rotating dust.
(2) The spacetime has a 3-dimensional symmetry group.
(3) Two of the symmetry generators are spanned on the vector fields of velocity u^α and rotation w^α, while the third one is linearly independent of u^α and w^α at every point.
(4) The generators form a Bianchi type II algebra.
(5) In the solutions of the Einstein equations, the Bianchi type I limit is taken of the Bianchi type II symmetry.
(6) The Gödel solution results when the matter–density is constant; the Lanczos solution results when the density is not constant.

The generalization with respect to the earlier definition is contained in point 3: in previous derivations the two generators were assumed to be collinear with u^α and w^α, respectively, from the beginning.

VIII. THE GENERATORS, THE KILLING EQUATIONS AND THE EINSTEIN EQUATIONS FOR A BIANCHI TYPE I ALGEBRA

We shall consider the case $b=c=0$ in (7.1)–(7.2). The reasoning up to Eq. (7.16) applies also here, but (7.17) no longer follows. Instead, the equation $G_{13}=0$ can be integrated with the result

$$k_{13,y}=BF/G^3-YA_{,y}, \tag{8.1}$$

where B is an arbitrary constant; the equation $G_{01}=0$ can be integrated to

$$Y_{,y}=(C-BA)F/G, \tag{8.2}$$

where C is an arbitrary constant; and the equation $G_{03}=0$ can be integrated to

$$A_{,y}=(BY-D)/(FG^3), \tag{8.3}$$

where D is one more arbitrary constant.

At this point, only the diagonal components of the Einstein tensor survive, and $G_{00}=\kappa\epsilon-\Lambda$ just defines the energy–density. The equations $G_{11}=\Lambda=G_{22}=G_{33}$ can be written as

$$-\frac{B}{4FG}k_{13,y}+\frac{C-BA}{4FG}Y_{,y}+G_{,yy}/G-\frac{2BY-D}{4FG}A_{,y}=\Lambda, \tag{8.4}$$

$$-\frac{B}{4FG}k_{13,y}+\frac{C-BA}{4FG}Y_{,y}+\frac{F_{,y}G_{,y}}{FG}-\frac{D}{4FG}A_{,y}=\Lambda, \tag{8.5}$$

$$\frac{3B}{4FG}k_{13,y}-\frac{C-BA}{4FG}Y_{,y}+F_{,yy}/F+\frac{3D}{4FG}A_{,y}=\Lambda. \tag{8.6}$$

[In order to arrive at this form, one has to calculate B from (8.1) and replace one factor B in B^2 by the resulting expression; then replace one $Y_{,y}$ in $Y_{,y}^2$ from (8.2) and replace one $A_{,y}$ in $A_{,y}^2$ from

(8.3).] The set (8.4)–(8.6) can be integrated to a first-order set. Subtracting (8.6) from (8.4) and multiplying the result by FG we obtain an equation that is easily integrated to

$$FG_{,y}-GF_{,y}-Bk_{13}-\tfrac{1}{2}BAY+\tfrac{1}{2}CY-\tfrac{1}{2}DA=E=\text{const}. \tag{8.7}$$

Now adding (8.3) and (8.4), and multiplying the result by FG we obtain another integrable equation whose integral can be written in the form

$$FG_{,y}=\frac{1}{2}Bk_{13}+\frac{1}{2}BAY-\frac{1}{2}CY+2\Lambda\int FGdy+H_0, \tag{8.8}$$

where H_0 is an arbitrary constant. The integral can be calculated if the new variable $u(y)$ is introduced by

$$dy/du=1/(FG). \tag{8.9}$$

From (8.7) and (8.8) it follows that

$$GF_{,y}=-\frac{1}{2}Bk_{13}-\frac{1}{2}DA-E+2\Lambda\int FGdy+H_0. \tag{8.10}$$

In the set (8.4)–(8.6) there remains one equation that has still not been used. However, at this point it merely introduces a relation between the arbitrary constants, i.e., implicitly defines H_0 in terms of the other constants. This is seen as follows: substitute for $k_{13,y}$, $Y_{,y}$, $F_{,y}$, $G_{,y}$, and $A_{,y}$ from (8.1)–(8.3), (8.8), and (8.10) in (8.5), thereby obtaining an algebraic equation (i.e., one without derivatives). Differentiate it by y and eliminate the derivatives in the same way again. What results is an identity $0=0$. Hence, the left-hand side of (8.5) is identically constant in virtue of the other equations.

In terms of the variable u from (8.9), Eqs. (8.1)–(8.3), (8.8), and (8.10) form an autonomous set of first-order equations that can be investigated further by qualitative methods (see, e.g., Ref. 17). This is left as a subject for a separate study.

In analogy with the Bianchi type I spatially homogeneous (nonrotating) dust solutions (see Ref. 18) one might expect further progress by adapting the coordinates suitably (in the case considered in Ref. 18, the metric can be diagonalized). However, this author was not able to achieve any such progress.

The functions $A(y)$ and $k_{13}(y)$ have invariant meaning: they are proportional to the scalar products of the Killing vectors [see Eqs. (7.7) and (7.15) with $b=0$]:

$$A=-g_{\alpha\beta}k^{\alpha}_{(2)}k^{\beta}_{(3)}/G^2, \quad k_{13}=-g_{\alpha\beta}k^{\alpha}_{(1)}k^{\beta}_{(3)}/G^2 \tag{8.11}$$

(note that $G^2=-g_{\alpha\beta}k^{\alpha}_{(3)}k^{\beta}_{(3)}$, i.e., it is a scalar, too). Hence, $A=0$ and $k_{13}=0$ are invariant properties. Note that $A=0$ implies, through (8.3), that either $Y=\text{const}$ (in which case there is no rotation) or $B=D=0$. In the latter case, $k_{13}=\text{const}$ and the coordinate transformation $z=z'-k_{13}x$ leads to $k_{13}=0$ in the new coordinates. With $A=k_{13}=0$, the Lanczos and Gödel models result from the Einstein equations as the only solutions. Hence, one more invariant definition of these models follows, similar to the six-point definition at the end of Sec. VII. Points 1, 2, 3, and 6 remain unchanged, while points 4 and 5 are replaced by the following.

(4) The generators form a Bianchi type I algebra.

(5) From the two generators spanned on u^{α} and w^{α}, two linear combinations can be constructed that are orthogonal to each other.

Point $5'$ is equivalent to the existence of coordinates in which $A=0$.

Note that the Bianchi type I models considered in this section are more general than the Bianchi type I limit of the models from Sec. VII; those from Sec. VII had $A=k_{13}=0$ in virtue of Einstein's equations.

The assumption $k_{13}=0$ (i.e., $g_{\alpha\beta}k^{\alpha}_{(1)}k^{\beta}_{(3)}=0$) alone does not lead to any immediate progress in solving the Einstein equations.

The Lanczos solution was originally derived in Ref. 1 (an English translation, Ref. 19, is now available), and rediscovered in Ref. 20. Its limit of zero cosmological constant was rediscovered in Ref. 21 as the cylindrically symmetric subcase of a family of stationary axially symmetric solutions. Geometrical and physical properties of the Lanczos solution were discussed in Ref. 1, and, in a more modern language, also in Ref. 22 (the latter only for the case $\Lambda=0$).

Coordinate transforms of the Gödel solution were published as new solutions in Refs. 23 and 24 (concerning Ref. 23 see also Ref. 25).

A metric form that is a modest generalization of the Gödel solution (it has two unknown functions of one variable in place of Gödel's e^{x^1} and e^{2x^1}) came to be known as ''Gödel-type metric'' and became the subject of a rather large number of papers; the activity seems to have started with Ref. 26, one of the most recent appearances of it is Ref. 27. However, it was proven already in Ref. 28 that the only perfect fluid solution with this metric is the Gödel solution itself; indeed, all other ''Gödel-type solutions'' have various nonperfect fluid sources, and therefore they do not show up in the scheme considered here.

As mentioned in Sec. I, several authors considered rotating charged dust solutions under the additional assumption that the electromagnetic field $F_{\mu\nu}$ exerts no force on the charged dust particles, i.e., that $F_{\mu\nu}u^{\nu}=0$. These solutions were all derived with another, rather natural assumption: that all charges are attached to dust particles so that no currents are present apart from the one created by the dust flow. Those solutions are found in Refs. 29–36. The one in Ref. 29 has only two-dimensional symmetry, so it could not come up in this investigation. The remaining ones are stationary and cylindrically symmetric and would have shown up here, had we allowed charges and electromagnetic fields in the source. They have the following properties.

The one from Ref. 30 becomes a vacuum solution in the limit $F_{\mu\nu}=0$.

The one from Ref. 31 does not allow this limit at all.

The limit $F_{\mu\nu}=0$ of the solution from Ref. 32 is the Minkowski metric.

The Som–Raychaudhuri solution[33] reproduces the $\Lambda=0$ subcase of the Lanczos solution when $F_{\mu\nu}=0$.

The first of the six solutions by Banerjee and Banerji[34] reduces to the Gödel solution when $F_{\mu\nu}=0$. The other five behave as follows: 2 and 6 become vacuum solutions when $F_{\mu\nu}=0$, no. 5 becomes the Minkowski spacetime, no. 3 does not allow this limit at all, and no. 4 has has a two-dimensional symmetry group.

Both solutions by Mitskiévič and Tsalakou[35] are generalizations of the Gödel solution; the first one of them is in addition a generalization of the full ($\Lambda\neq 0$) Lanczos solution (In fact, the second solution has a nonzero pressure gradient that remains nonzero even after the limit $F_{\mu\nu}=0$ is taken. Another limiting transition, given in the paper, reduces the solution to Gödel's.) In the limit $\Lambda=0$, the first solution reduces to the one by Som and Raychaudhuri.[33]

The two solutions from Ref. 36 are coordinate transforms of those from Ref. 35.

Three other generalizations of the Gödel solution exist in the literature that have zero acceleration. Two were provided by Raval and Vaidya;[37] the first of them is stationary, the second expanding, both have anisotropic pressure. The third is the solution by Rebouças[38] in which the source is a free electromagnetic field (see also Ref. 39). The metric of the Rebouças solution is the same as that in the first Banerjee–Banerji solution. This coincidence was explained by Raychaudhuri and Guha Thakurta:[40] The two electromagnetic fields (one generated by a current, the other source-free) are related by a point-dependent duality rotation.

IX. ANOTHER INVARIANT DEFINITION OF THE GÖDEL AND OZSVATH CLASS III SOLUTIONS

Assumptions about invariant properties of the velocity field of matter usually lead to progress in solving the Einstein equations; the most impressive example were the shear-free normal models of Barnes,[41] where a large class of solutions resulted from the assumptions of zero shear and zero rotation in a perfect fluid source. Inspired by this, one can try to make assumptions about other vector fields characterizing fluid sources, e.g., the rotation. Indeed, it turns out that the assumption

$$w_{\alpha;\beta}=0, \tag{9.1}$$

i.e., the rotation field being covariantly constant, together with the assumption of geodesic motion of a perfect fluid source, leads uniquely to two solutions of Einstein's equations. However, both of them were obtained before by other methods. One is the Ozsvath class III metric,[11] originally identified as one of the solutions that are homogeneous in four dimensions; the other is the Gödel solution[2] which is the shear-free limit of the Ozsvath solution.

From (9.1) and from the Ricci identity $2w_{\alpha;[\beta\gamma]}=R^{\rho}{}_{\alpha\beta\gamma}w_{\rho}$ one obtains for the Ricci tensor,

$$R^{\rho}{}_{\gamma}w_{\rho}=0, \tag{9.2}$$

and then from the Einstein equations for a perfect fluid,

$$G_{\alpha\beta}+\Lambda g_{\alpha\beta}=\kappa[(\epsilon+p)u_{\alpha}u_{\beta}-pg_{\alpha\beta}], \quad \kappa=8\pi G/c^4, \tag{9.3}$$

and from $u^{\alpha}w_{\alpha}=0$ one obtains

$$\Lambda=\tfrac{1}{2}\kappa(\epsilon-p). \tag{9.4}$$

In the case $\Lambda=0$, this is the well-known ''stiff perfect fluid.'' Equation (9.4) is a necessary condition for (9.1) when the source is a perfect fluid.

As stated at the end of Sec. III, if the perfect fluid moves geodesically with rotation, then necessarily $p=\text{const}$. Equation (9.4) implies then that $\epsilon=\text{const}$, i.e., a geodesically moving and rotating perfect fluid whose rotation vector is covariantly constant must have constant matter density. Since $\Lambda=0$ may be assumed with no loss of generality (this leads only to redefining p), we shall assume this from now on. Then $\epsilon=p$ and $\epsilon+p=2p$ is a conserved quantity. Hence, we may assume

$$n=\epsilon+p=2p=\text{const} \tag{9.5}$$

in all formulae. In particular, (3.20) implies then that

$$g=\det(g_{\alpha\beta})=-(2p)^{-2}. \tag{9.6}$$

Using (9.6) and (3.23) in (9.1) we obtain, in the Plebański coordinates,

$$\tfrac{1}{2}n(g_{\alpha 3,\beta}-g_{\beta 3,\alpha}+g_{\alpha\beta,3})=0. \tag{9.7}$$

After a simple algebraic manipulation this set of equations yields the following result:

$$g_{33}=-A^2=\text{const}, \quad g_{\alpha\beta,z}=0, \quad g_{13,t}=g_{23,t}=0,$$

$$g_{23,x}-g_{13,y}=0. \tag{9.8}$$

The second equation in (9.8) means that w^α is a Killing vector, as should be expected from (9.1), (9.6), and (3.23). Equations (9.8) imply that g_{13} and g_{23} depend only on x and y, and that there exists a function $\mathscr{F}(x,y)$ such that

$$g_{23}=\mathscr{F}_{,y}, \quad g_{13}=\mathscr{F}_{,x}. \tag{9.9}$$

Since we assumed that rotation is nonzero, we know that $g_{33}=-g_{\alpha\beta}w^\alpha w^\beta/n^2\neq 0$, and so we are allowed to carry out the coordinate transformation:

$$z=z'-\mathscr{F}/g_{33}, \tag{9.10}$$

that, in virtue of (9.9), will lead to

$$g_{13}=g_{23}=0 \tag{9.11}$$

in the new coordinates. We have thus arrived at the metric form:

$$ds^2=(dt+ydx)^2-h(t,x,y)[dx+k(t,x,y)dy]^2-l(t,x,y)dy^2-A^2dz^2, \tag{9.12}$$

where h, k, and l are functions to be found from the Einstein equations and A is an arbitrary constant.

From now on, the allowed coordinate transformations are (5.6)–(5.7), but with $T=$const.

The components of the Einstein tensor will now be referred to the orthonormal tetrad implied by (9.12). The equation $G_{12}=0$ is integrated with the result

$$k_{,t}=K(x,y)l^{1/2}/h^{3/2}-1/h. \tag{9.13}$$

The equation $G_{22}=\kappa p$, with l eliminated by (9.6), is integrated with the result

$$h=[H^2(x,y)+K^2/(4\kappa p)]^{1/2}+H\,\sin[2(\kappa p)^{1/2}t+\tau(x,y)], \tag{9.14}$$

where $H(x,y)$, $k(x,y)$, and $\tau(x,y)$ are arbitrary functions. Now $G_{11}+G_{22}=2\kappa p$ imposes an additional condition on (9.13) and (9.14) that leads to $H=0$ or

$$H=(A^2-\kappa/p)^{1/2}K/(2\kappa). \tag{9.15}$$

The case $H=0$ leads to the Gödel solution (see below), so we shall consider the more general case (9.15). Then, from (9.14),

$$h=[K/(2\kappa)]\{A+(A^2-\kappa/p)^{1/2}\,\sin[2(\kappa p)^{1/2}t+\tau]\}. \tag{9.16}$$

With such h, Eq. (9.13) can be integrated with the result

$$k=[2A(\kappa p)^{1/2}h]^{-1}(A^2-\kappa/p)^{1/2}\,\cos[2(\kappa p)^{1/2}t+\tau]+L(x,y), \tag{9.17}$$

where L is a new arbitrary function. The function l is then calculated from (9.6), and an explicit solution of Einstein's equations is determined by (9.16) and (9.17).

The transformations (5.6)–(5.7) with $T=$const can now be used to simplify the metric tensor so that $L=0$. The transformation that yields this is given in Appendix A. The still allowed coordinate transformations that preserve the property $L=0$ are given by (5.6)–(5.7) with $T=$const and with the additional condition

$$KF_{,x}F_{,y}+[\kappa/(A^2pK)]G_{,x}G_{,y}=0. \tag{9.18}$$

With $L=0$, from the equations $G_{01}=G_{02}=0$ one obtains further

$$\tau_{,x}=2y(\kappa p)^{1/2}-A(p/\kappa)^{1/2}K_{,y},$$
$$\tau_{,y}=-(\kappa/p)^{1/2}K_{,x}/(AK^2). \tag{9.19}$$

The integrability condition of (9.19) is

$$(p/\kappa)^{1/2}(AK)_{,yy}+(\kappa p)^{1/2}[1/(AK)]_{,xx}-2(\kappa p)^{1/2}=0. \tag{9.20}$$

By the same method as was used in Ref. 4 it can now be shown that Eq. (9.20) is at the same time the integrability condition for such a coordinate transformation (5.6)–(5.7)–(9.18) after which (see Appendix A again)

$$K=(\kappa/A)y^2, \tag{9.21}$$

and then (9.19) implies

$$\tau=c=\text{const.} \tag{9.22}$$

The value of c can be set arbitrarily by transformations of t of the form $t=t'+\text{const}$. To match Ref. 11 one should choose

$$c=-\pi/2. \tag{9.23}$$

Finally, the functions h, k, and l in (9.12) are thus

$$h=\tfrac{1}{2}y^2\{1+[1-\kappa/(pA^2)]^{1/2}\cos[2(\kappa p)^{1/2}t]\},$$
$$k=\tfrac{1}{2}[(\kappa p)^{1/2}h]^{-1}[1-\kappa/(pA^2)]^{1/2}\sin[2(\kappa p)^{1/2}t], \tag{9.24}$$
$$l=(4p^2A^2h)^{-1}.$$

This is equivalent under a simple coordinate transformation to the Ozsvath class III solution from Ref. 11.

The velocity field $u^\alpha=\delta^\alpha_0$ for this solution has nonzero shear. The shear will vanish if and only if:

$$A^2=\kappa/p, \tag{9.25}$$

and then the Gödel solution in the form (5.11) results.

The invariant definitions of the Ozsvath class III and of the Gödel solutions given at the end of Sec. I follow from the derivation in this section.

X. CONCLUDING REMARKS

These are the main results of the paper.

1. With nonzero rotation, any Killing field, existing for a metric whose matter source inherits the symmetry, must have the form (5.5) when represented in the Plebański coordinates. When $\phi_{;\alpha}\neq 0$, the Plebański coordinates can be adapted to the Killing field so that $k^\alpha=\delta^\alpha{}_1$.

2. When two of the generators of the group are spanned on the velocity and rotation vector fields, while the third one is not, the collection of solutions with a dust source is exhausted by the following two sets.

(a) The set of Sec. VII, defined by a single differential equation (7.24), where the metric is (7.15) with $A=k_{13}=0$, F defined by (7.23) and $y(Y)$ defined by (7.21).

(b) The set of Sec. VIII, where the metric is (7.15) with $b=0$, and the metric functions are defined by an autonomous set of first-order equations (8.1)–(8.3), (8.8), and (8.10) [the integral in (8.8) and (8.10) can be calculated if the variable is changed as in (8.9)].

3. The solutions of Lanczos[1,19] and Gödel[2] are limiting cases of both sets; their invariant definitions are given at the end of Sec. VII and of Sec. VIII.

4. With no symmetries pre-assumed, if the source is a rotating geodesic perfect fluid whose rotation vector field is covariantly constant, then the solution of the Einstein equations is the homogeneous (in four dimensions) Ozsvath class III solution.[11] If shear is zero in addition, then the Gödel solution (Ref. 2; see also Ref. 4) results.

Note the modification that the results 2(a) and 2(b) introduce in theorem 3.1 of King and Ellis.[42] Those authors considered spatially homogeneous models in which the velocity field of matter was tilted (i.e., was not orthogonal) with respect to the hypersurfaces of homogeneity. Theorem 3.1 says, among other things, that there are no tilted models of type I and that tilted models of type II have zero vorticity. Evidently, this does not apply to the case where the hypersurfaces of homogeneity are timelike. The solutions of Sec. VII are of Bianchi type II, they are ''tilted'' (because the velocity field is tangent to the hypersurfaces of homogeneity), yet rotation is not zero. The solutions of Sec. VIII are tilted in the same sense, yet they are of Bianchi type I.

Other solutions that have been published earlier will be mentioned where appropriate in papers 2 and 3. A general overview of literature on related subjects will be included in paper 3.

ACKNOWLEDGMENT

The algebraic calculations for this paper were done with the help of the program Ortocartan.[43,44]

APPENDIX A: THE TRANSFORMATION TO $L=0$ IN (9.17)

A transformation of the class (5.6)–(5.7) with $T=\text{const}$ changes the functions h and k to such ones that can be cast in the form (9.16) and (9.17), respectively, with the new functions K', τ', and L' expressed through the old ones as follows:

$$K'=K(F_{,x'}+LG_{,x'})^2+\kappa G_{,x'}^2/(A^2pK), \tag{A1}$$

$$\tau'=\tau-2(\kappa p)^{1/2}S+U, \tag{A2}$$

$$L'=K'^{-1}[KF_{,x'}F_{,y'}+2KLF_{,y'}G_{,x'}+KL+KL^2G_{,x'}G_{,y'}+\kappa G_{,x'}G_{,y'}/(A^2pK)], \tag{A3}$$

where S in (A2) is the function from (5.6)–(5.7) and U is determined by

$$\cot U=2A(\kappa p)^{1/2}[2G_{,x'}(F_{,x'}+LG_{,x'})]^{-1}[-K(F_{,x'}+LG_{,x'})^2/(2\kappa)+G_{,x'}^2/(2A^2pK)]. \tag{A4}$$

Note that we are applying here (5.6)–(5.7) in reverse, i.e., with the roles of x^α and x'^α interchanged. The functions of the inverse transformation, denoted again by S, F, G, and T, still obey (5.7). For consistency of all the formulae it is convenient to choose U from the segment $(\pi, 2\pi)$. Then, the limiting cases $G_{,x'}=0$ and $F_{,x'}+LG_{,x'}=0$ are included in (A4) as the limits $U=2\pi$ and $U=\pi$, respectively (these limiting cases occur when $L_{,xx}=0$ in the original coordinates).

From (A3), the equation $L'=0$ turns out to be consistent with (5.6)–(5.7). In order to see this, one can solve (A3) and (5.7) for $F_{,x'}$ and $F_{,y'}$ and then impose the integrability condition $F_{,x'y'}-F_{,y'x'}=0$. What comes out is a well-defined (though highly nonlinear) partial differential equation of second order for G whose coefficients depend only on K and L.

The transformation preserving the property $L=0$ are (5.6)–(5.7) with (9.18), the latter easily follows from (A3). Equation (A1) with $L=0$ then shows how K is changed by such a transformation; this is useful in showing that coordinates exist in which $K_{,x}=0$ [see the remark after (9.20)]. The proof is identical as in Appendix C to Ref. 4. Note that the conclusion in Ref. 4 is weaker than it could be: the $K(x)$ [$v(t)$ in Ref. 4] is determined up to an additive constant C. Hence, by a transformation of the form $y=y'+\text{const}$ and by an appropriate choice of C one can remove the linear and the constant terms in K (resp., v) so that $K\propto y^2$ (resp., $v\propto t^2$ in Ref. 4).

[1] K. Lanczos, Z. Phys. **21**, 73 (1924).
[2] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[3] J. Plebański, *Lectures on Non-Linear Electrodynamics* (Nordita, Copenhagen, 1970), pp. 107–115 and 130–141.
[4] A. Krasiński, Acta Phys. Pol. B **5**, 411 (1974).
[5] A. Krasiński, Acta Phys. Pol. B **6**, 223 (1975).
[6] A. Krasiński, Acta Phys. Pol. B **6**, 239 (1975) (available in extended form as a preprint).
[7] A. Krasiński, J. Math. Phys. **16**, 125 (1975).
[8] A. Krasiński, Rep. Math. Phys. **14**, 225 (1978).
[9] C. B. Collins, Can. J. Phys. **64**, 191 (1986).
[10] H. Stephani, Class. Quantum Grav. **4**, 125 (1987).
[11] I. Ozsvath, J. Math. Phys. **11**, 2871 (1970).
[12] S. Sternberg, *Lectures on Differential Geometry* (Prentice–Hall, Englewood Cliffs, NJ, 1964), p. 141.
[13] G. F. R. Ellis, in ''General relativity and cosmology,'' *Proceedings of the International School of Physics ''Enrico Fermi,''* Course 47, edited by R. K. Sachs (Academic, New York, 1971), p. 104.
[14] J. Ehlers, Abhandl. Math. Naturw. Kl. Akad. Wiss. Lit. Mainz **11**, 791 (1961); English translation in Gen. Relativ. Gravit. **25**, 1225 (1993).
[15] G. F. R. Ellis, J. Math. Phys. **8**, 1171 (1967).
[16] E. Verdaguer, Phys. Rep. **229** No. 1, 1 (1993).
[17] O. J. Bogoyavlenskii, *Methods in the Qualitative Theory of Dynamical Systems in Astrophysics and Gas Dynamics* (Springer, Berlin, 1980).
[18] H. Stephani, *General Relativity*, 2nd ed. (Cambridge University Press, Cambridge, 1990), pp. 284–288.
[19] K. Lanczos, Gen. Rel. Grav. **29**, 363 (1997).
[20] J. P. Wright, J. Math. Phys. **6**, 103 (1965).
[21] W. J. van Stockum, Proc. R. Soc. Edinburgh **57**, 135 (1937).
[22] W. B. Bonnor, J. Phys. A **13**, 2121 (1980).
[23] C. Hoenselaers and C. V. Vishveshwara, Gen. Relativ. Gravit. **10**, 43 (1979).
[24] A. H. Khater and M. F. Mourad, Astrophys. Space Sci. **163**, 247 (1990).
[25] S. K. Chakraborty, Gen. Relativ. Gravit. **12**, 925 (1980).
[26] M. Novello and M. J. Rebouças, Phys. Rev. D **19**, 2850 (1979).
[27] R. X. Saibatalov, Gen. Relativ. Gravit. **27**, 697 (1995).
[28] F. Bampi and C. Zordan, Gen. Relativ. Gravit. **9**, 393 (1978).
[29] J. N. Islam, Proc. R. Soc. London, Ser. A **353**, 523 (1977).
[30] J. N. Islam, Proc. R. Soc. London, Ser. A **385**, 189 (1983).
[31] P. Wils and N. van den Bergh, Proc. R. Soc. London, Ser. A **394**, 437 (1984).
[32] A. Georgiou, Nuovo Cimento B **108**, 69 (1993).
[33] M. M. Som and A. K. Raychaudhuri, Proc. R. Soc. London, Ser. A **304**, 81 (1968).
[34] A. Banerjee and S. Banerji, J. Phys. A **1**, 188 (1968).
[35] N. V. Mitskiévič and G. A. Tsalakou, Class. Quantum Grav. **8**, 209 (1991).
[36] A. M. Upornikov, Class. Quantum Grav. **11**, 2085 (1994).
[37] H. M. Raval and P. C. Vaidya, Ann. Inst. Henri Poincaré, Sect. A **4**, 21 (1966).
[38] M. J. Rebouças, Phys. Lett. A **70**, 161 (1979).
[39] M. J. Rebouças and J. Tiomno, Phys. Rev. D **28**, 1251 (1983).
[40] A. K. Raychaudhuri and S. N. Guha Thakurta, Phys. Rev. D **22**, 802 (1980).
[41] A. Barnes, Gen. Relativ. Gravit. **4**, 105 (1973).
[42] A. R. King and G. F. R. Ellis, Commun. Math. Phys. **31**, 209 (1973).
[43] A. Krasiński, Gen. Relativ. Gravit. **25**, 165 (1993).
[44] A. Krasiński and M. Perkowski, Gen. Relativ. Gravit. **13**, 67 (1981).

Rotating dust solutions of Einstein's equations with 3-dimensional symmetry groups. II. One Killing field spanned on u^α and w^α

Andrzej Krasiński
N. Copernicus Astronomical Center and College of Science, Polish Academy of Sciences, Bartycka 18, 00 716 Warszawa, Poland

(Received 10 June 1997; accepted for publication 8 July 1997)

This is the second part of a series of 3 papers. Using the same method and the same coordinates as in part 1, rotating dust solutions of Einstein's equations are investigated that possess 3-dimensional symmetry groups, under the assumption that only one of the Killing fields is spanned on the fields of velocity u^α and rotation w^α, while the other two define vectors that are linearly independent of u^α and w^α at every point of the spacetime region under consideration. The Killing fields are found and the Killing equations solved for the components of the metric tensor in every case that arises. The Einstein equations are simplified in a few cases, three (most probably) new solutions are found, and several classes of solutions known earlier are identified in the present scheme. They include those by Ozsváth, Maitra, Ellis, King, and Vishveshwara and Winicour. © *1998 American Institute of Physics.* [S0022-2488(97)03212-X]

I. SUMMARY OF THE METHOD

This is a concise summary of results that will be used in this paper. For proofs, motivations, and references see Paper 1.[1]

Every timelike vector field u^α of unit length that has zero acceleration and nonzero rotation defines the functions $\tau(x)$, $\eta(x)$, and $\xi(x)$ such that

$$u_\alpha=\tau_{,\alpha}+\eta\xi_{,\alpha}. \tag{1.1}$$

These functions are defined up to the transformations

$$\tau=\tau'-S(\xi',\eta'), \quad \xi=F(\xi',\eta'), \quad \eta=G(\xi',\eta'), \tag{1.2}$$

where the functions F and G obey

$$F_{,\xi'}G_{,\eta'}-F_{,\eta'}G_{,\xi'}=1 \tag{1.3}$$

(this guarantees that the Jacobian of the transformation is 1), and S is determined by

$$S_{,\xi'}=GF_{,\xi'}-\eta', \quad S_{,\eta'}=GF_{,\eta'}. \tag{1.4}$$

If u^α is the velocity field of a fluid whose number of particles is conserved:

$$(\sqrt{-g}nu^\alpha)_{,\alpha}=0 \tag{1.5}$$

(where g is the determinant of the metric tensor and n is the particle number density), then one more function $\zeta(x)$ exists such that

$$\sqrt{-g}nu^\alpha=\epsilon^{\alpha\beta\gamma\delta}\xi_{,\beta}\eta_{,\gamma}\zeta_{,\delta}, \tag{1.6}$$

and it is determined up to the transformations

$$\zeta=\zeta'+T(\zeta',\eta'). \tag{1.7}$$

Note that n is not defined uniquely by (1.5). For example, if $u^\alpha=\delta^\alpha_0$ and n obeys (1.5), then $n'=nf(x,y,z)$ (where f is an arbitrary function) will also obey (1.5). This nonuniqueness allows for a greater freedom in the choice of ζ than (1.7), and the freedom will be used in some cases.

The following relations hold:

$$u^\alpha\tau_{,\alpha}=1, \quad u^\beta\xi_{,\beta}=u^\beta\eta_{,\beta}=u^\beta\zeta_{,\beta}=0,$$

$$\frac{\partial(\tau,\eta,\xi,\zeta)}{\partial(x^0,x^1,x^2,x^3)}=\sqrt{-g}n\neq 0. \tag{1.8}$$

The last of (1.8) guarantees that $\{\tau,\xi,\eta,\zeta\}$ can be chosen as coordinates; they will be called Plebański coordinates. Then, with $\{\tau,\xi,\eta,\zeta\}=\{x^0,x^1,x^2,x^3\}=\{t,x,y,z\}$,

$$u_\alpha=\delta^\alpha{}_0, \quad u_\alpha=\delta^0{}_\alpha+y\,\delta^1{}_\alpha,$$

$$g_{00}=1, \quad g_{01}=y, \quad g_{02}=g_{03}=0, \quad g=\det(g_\alpha\beta)=-n^{-2}, \tag{1.9}$$

$$w^\alpha=n\,\delta^\alpha_3, \quad \omega_{\alpha\beta}=-\omega_{\beta\alpha}=(1/2)\,\delta^1{}_\alpha\delta^2{}_\beta,$$

where w^α is the rotation vector field, and $\omega_{\alpha\beta}$ is the rotation tensor corresponding to the velocity field u_α:

$$\omega_{\alpha\beta}=\tfrac{1}{2}\,(u_{\alpha,\beta}-u_{\beta,\alpha}-\dot{u}_\alpha u_\beta+\dot{u}_\beta u_\alpha), \quad w^\alpha=-(1/\sqrt{-g})\,\epsilon^{\alpha\beta\gamma\delta}u_\beta\omega_{\gamma\delta}. \tag{1.10}$$

If $\omega_{\alpha\beta}\neq 0$ and $\dot{u}^\alpha=0$ (what is assumed throughout), then necessarily the pressure $p=\text{const}$ and κp may be interpreted as the cosmological constant ($\kappa:=8\pi G/c^4$).

If any Killing vector field exists on a manifold (on which all the assumptions specified so far are fulfilled), then, in the coordinates of (1.9), it must be of the form

$$k^\alpha=(C+\phi-y\phi_{,y})\,\delta^\alpha{}_0+\phi_{,y}\delta^\alpha{}_1-\phi_{,x}\delta^\alpha{}_2+\lambda\,\delta^\alpha{}_3, \tag{1.11}$$

where C is an arbitrary constant and $\phi(x,y)$ and $\lambda(x,y)$ are arbitrary functions of two coordinates. Whenever $\phi_{,\alpha}\neq 0$, a transformation of the class (1.2)–(1.4) can be found that leads to

$$k^\alpha=\delta^\alpha{}_1. \tag{1.12}$$

The metric then becomes independent of x, and the coordinates preserving (1.12) are determined up to the transformations

$$t'=t-\int yH_{,y}dy+A, \quad x'=x+H(y), \quad y'=y, \quad z'=z+T(y), \tag{1.13}$$

where A is an arbitrary constant and H,T are arbitrary functions.

The condition $\phi_{,\alpha}\neq 0$ that allows one to fulfill (1.12) means that the Killing vector k^α is linearly independent of the vectors u^α and w^α at every point of the spacetime region under consideration. In Paper 1, solutions of the Killing equations and of the Einstein equations were considered under the assumption that there exist three Killing vector fields on the manifold, two of which have $\phi=\text{const}$ in (1.11), while the third one has $\phi_{,\alpha}\neq 0$ and can be transformed to the form (1.12). In the present paper it is assumed that only one Killing field has $\phi=\text{const}$ (it will be $k_{(3)}$; see below), and two have $\phi_{,\alpha}\neq 0$. Only one of the two can then be transformed to the form (1.12) (it will be $k_{(1)}$); the other one ($k_{(2)}$) will preserve its general form (1.11).

In the whole paper, whenever reference is made to the Einstein tensor, the components quoted are projections of the coordinate components $G_{\alpha\beta}$ onto the orthonormal tetrad $e_i{}^\alpha$, i.e., $G_{ij} = e_i{}^\alpha e_j{}^\beta G_{\alpha\beta}$. In every case it will be self-evident how the tetrad is defined.

II. THE LIE ALGEBRA OF THE SYMMETRY GROUP

According to the assumptions made in the preceding section, there exist the following three Killing vector fields:

$$\begin{aligned} k^\alpha_{(1)} &= \delta^\alpha_1, \\ k^\alpha_{(2)} &= (C_2 + \phi - y\phi_{,y})\delta^\alpha_0 + \phi_{,y}\delta^\alpha_1 - \phi_{,x}\delta^\alpha_2 + \lambda_2(x,y)\delta^\alpha_3, \\ k^\alpha_{(3)} &= C_3\delta^\alpha_0 + \lambda_3(x,y)\delta^\alpha_3, \end{aligned} \tag{2.1}$$

where C_2 and C_3 are arbitrary constants, and ϕ, λ_2, and λ_3 are unknown functions of (x,y), to be determined from the commutation relations. The coordinates of (2.1) are determined up to (1.13).

The fields $k_{(1)}$, $k_{(2)}$, and $k_{(3)}$ will form a Lie algebra if constants $a,\ldots,j$ exist such that

$$\begin{aligned} [k_{(1)},k_{(2)}] &= ak_{(1)} + bk_{(2)} + ck_{(3)}, \\ [k_{(1)},k_{(3)}] &= dk_{(1)} + ek_{(2)} + fk_{(3)}, \\ [k_{(2)},k_{(3)}] &= gk_{(1)} + hk_{(2)} + jk_{(3)}, \end{aligned} \tag{2.2}$$

Eqs. (2.2) are equivalent to the following set:

$$\phi_{,x} - y\phi_{,xy} = b(C_2 + \phi - y\phi_{,y}) + cC_3, \tag{2.3a}$$

$$\phi_{,xy} = a + b\phi_{,y}, \tag{2.3b}$$

$$\phi_{,xx} = b\phi_{,x}, \tag{2.3c}$$

$$\lambda_{2,x} = b\lambda_2 + c\lambda_3, \tag{2.3d}$$

$$e(C_2 + \phi - y\phi_{,y}) + fC_3 = 0, \tag{2.3e}$$

$$d + e\phi_{,y} = 0, \tag{2.3f}$$

$$e\phi_{,x} = 0, \tag{2.3g}$$

$$\lambda_{3,x} = e\lambda_2 + f\lambda_3, \tag{2.3h}$$

$$h(C_2 + \phi - y\phi_{,y}) + jC_3 = 0, \tag{2.3i}$$

$$g + h\phi_{,y} = 0, \tag{2.3j}$$

$$h\phi_{,x} = 0, \tag{2.3k}$$

$$\phi_{,y}\lambda_{3,x} - \phi_{,x}\lambda_{3,y} = h\lambda_2 + j\lambda_3. \tag{2.3l}$$

A few of these equations imply alternatives that will have to be considered separately. The alternatives will be organized into a binary tree and numbered in a positional system that will enable one to quickly identify complementary cases (see Fig. 1). Of the two alternatives implied by (2.3g) we choose for the beginning, the following.

Case 1: $\phi_{,x} \neq 0$.

This implies, from Eqs. (2.3g), (2.3k), (2.3f), and (2.3j)

$$e = h = d = g = 0. \tag{2.4}$$

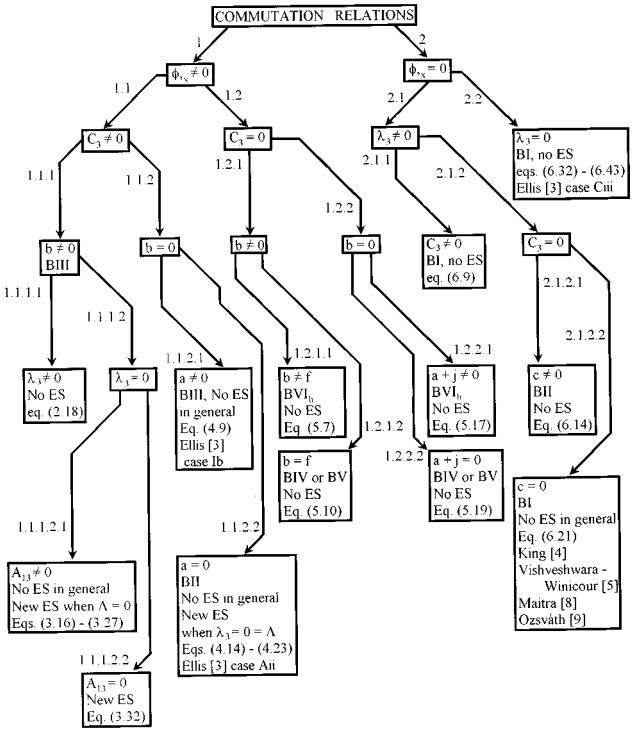

FIG. 1. The classes of metrics considered in the paper. Arrows point from more general classes to subclasses. The numbers at arrows are the case-numbers used in the text. The first entry in each rectangle is the property defining the case; all the symbols are introduced in Eqs. (2.1)–(2.3). The subsequent entries give the following information: 1. The Bianchi type of the corresponding algebra (2.2); 2. Information about the exact solutions (ES) known before or found in this paper; 3. The equation-numbers corresponding to the final result in the given case; 4. References to papers in which solutions of the given class were discussed.

Equations (2.3a)–(2.3d) survive unchanged, while the remaining ones simplify to

$$fC_3 = 0, \tag{2.5e}$$

$$\lambda_{3,x} = f\lambda_3, \tag{2.5h}$$

$$jC_3 = 0, \tag{2.5i}$$

$$\phi_{,y}\lambda_{3,x} - \phi_{,x}\lambda_{3,y} = j\lambda_3, \tag{2.5l}$$

other equations being fulfilled identically. Now we choose the following.

Case 1.1: $C_3 \neq 0$,

which implies, from (2.5),

$$f = j = 0, \tag{2.6}$$

$$\lambda_3 = \lambda_3(y), \tag{2.7h}$$

$$\phi_{,x}\lambda_{3,y} = 0. \tag{2.7l}$$

With (2.4) and (2.6), the second and third commutator in (2.2) are zero. Since in Case 1 $\phi_{,x} \neq 0$ by assumption, Eqs. (2.7h) and (2.7l) imply

$$\lambda_3 = \text{const}. \tag{2.8}$$

The solution of (2.3d) depends on whether $b=0$ or $b\neq 0$, so we follow now with case 1.1.1

Case 1.1.1: $b\neq 0$.

Then, from (2.3d),

$$\lambda_2=\beta(y)e^{bx}-c\lambda_3/b, \tag{2.9}$$

and from (2.3a)–(2.3c),

$$\phi=\alpha(y)e^{bx}-(a/b)y-C_2-(c/b)C_3, \tag{2.10}$$

where $\alpha(y)$ and $\beta(y)$ are arbitrary functions. The assumption $\phi_{,x}\neq 0$ implies $\alpha\neq 0$. The basis of the Killing vector fields can be changed to $k_{(1)}$, $k'_{(2)}=k_{(2)}+(c/b)k_{(3)}+(a/b)k_{(1)}$, and $k_{(3)}$. In the new basis,

$$k'^{\alpha}_{(2)}=e^{bx}[(\alpha-y\alpha_{,y})\delta^{\alpha}_0+\alpha_{,y}\delta^{\alpha}_1-b\alpha\delta^{\alpha}_2+\beta\delta^{\alpha}_3] \tag{2.11}$$

(i.e., by taking $k'_{(2)}$ instead of $k_{(2)}$, $c=0$ was achieved). The transformation (1.13) with $H=(1/b)\ln\alpha$, $T=(1/b)\int(\beta/\alpha)dy$ will now have the same effect as if $\beta=0$, $\alpha=1$. Finally, then, the following basis of the Killing vector fields resulted:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_1, \quad k^{\alpha}_{(2)}=e^{bx}(\delta^{\alpha}_0-b\delta^{\alpha}_2), \quad k^{\alpha}_{(3)}=C_3\delta^{\alpha}_0+\lambda_3\delta^{\alpha}_3, \tag{2.12}$$

but from now on the transformations (1.13) are allowed only with H and T being constants. The commutation relations are

$$[k_{(1)},k_{(3)}]=[k_{(2)},k_{(3)}]=0, \quad [k_{(1)},k_{(2)}]=bk_{(2)}, \tag{2.13}$$

and they correspond to Bianchi type III. In virtue of the commutator $[k_{(1)},k_{(3)}]$ being zero, coordinates can be adapted to $k_{(1)}$ and $k_{(3)}$ simultaneously so that $k^{\alpha'}_{(1)}=\delta^{\alpha'}_1$, $k^{\alpha'}_{(3)}=\delta^{\alpha'}_0$. The transformation is

$$(t',x',y')=(t,x,y), \quad z'=-\lambda_3 t+C_3 z; \tag{2.14}$$

it is nonsingular because Case 1.1 was defined by $C_3\neq 0$, but it leads out of the Plebański class of coordinates used so far. In the new coordinates (primes dropped), the metric is independent of x and t, and

$$\begin{gathered} k^{\alpha}_{(2)}=e^{bx}(\delta^{\alpha}_0-b\delta^{\alpha}_2-\lambda_3\delta^{\alpha}_3), \\ u^{\alpha}=\delta^{\alpha}_0-\lambda_3\delta^{\alpha}_3, \quad w^{\alpha}=nC_3\delta^{\alpha}_3, \\ g_{00}=1+\lambda_3{}^2 g_{33}, \quad g_{01}=y+\lambda_3 g_{13}, \\ g_{02}=\lambda_3 g_{23}, \quad g_{03}=\lambda_3 g_{33}. \end{gathered} \tag{2.15}$$

Now the Killing equations involving $k_{(2)}$ have to be solved for the components of the metric tensor. For this purpose, two cases have to be considered separately: $\lambda_3\neq 0$ and $\lambda_3=0$.

Case 1.1.1.1: $\lambda_3\neq 0$.

It is convenient to change the coordinates yet again to

$$(t',x')=(t,x), \quad Y=\lambda_3 y+bz, \quad Z=\lambda_3 y-bz. \tag{2.16}$$

The vector fields u^{α}, w^{α}, and $k^{\alpha}_{(i)}$ become (primes dropped)

$$\begin{gathered} u^{\alpha}=\delta^{\alpha}_0-b\lambda_3(\delta^{\alpha}_2-\delta^{\alpha}_3), \quad w^{\alpha}=bC_3 n(\delta^{\alpha}_2-\delta^{\alpha}_3), \\ k^{\alpha}_{(1)}=\delta^{\alpha}_1, \quad k^{\alpha}_{(3)}=\delta^{\alpha}_0, \quad k^{\alpha}_{(2)}=e^{bx}(\delta^{\alpha}_0-2b\lambda_3\delta^{\alpha}_3), \end{gathered} \tag{2.17}$$

and the metric tensor, after the transformation (2.16) and after solving the Killing equations becomes (primes dropped again)

$$
\begin{aligned}
g_{00}&=1+(\lambda_3 b)^2(g_{22}+g_{33}-2g_{23}),\\
g_{01}&=\frac{1}{2\lambda_3}\,(Y+Z)-\frac{1}{2}\,b^2\lambda_3 Y(g_{22}-g_{33})+\lambda_3 h_{13},\\
g_{02}&=b\lambda_3(g_{22}-g_{23}),\quad g_{03}=b\lambda_3(g_{23}-g_{33}),\\
g_{11}&=\left(\frac{Y}{2\lambda_3}\right)^2+\frac{1}{2\lambda_3^2}\,YZ+\left(\frac{1}{2}\,bY\right)^2(g_{22}+g_{33}+2g_{23})-\frac{b}{\lambda_3}\,h_{12}Y+h_{11},\\
g_{12}&=-\frac{1}{2}\,bY(g_{22}+g_{23})+\frac{1}{2\lambda_3}\,h_{12}+\frac{1}{2b}\,h_{13},\\
g_{13}&=-\frac{1}{2}\,bY(g_{23}+g_{33})+\frac{1}{2\lambda_3}\,h_{12}-\frac{1}{2b}\,h_{13},
\end{aligned}
\tag{2.18}
$$

where $h_{11}(Z)$, $h_{12}(Z)$, $h_{13}(Z)$, $g_{22}(Z)$, $g_{23}(Z)$, and $g_{33}(Z)$ are functions to be determined from the Einstein equations.

No progress was made with the Einstein equations in full generality in this case. Progress may be possible with additional simplifying assumptions (e.g., orthogonality relations among the Killing fields).

No solutions belonging to this case have been identified in the existing literature.

III. CASE 1.1.1.2: $\lambda_3=0$

We go back to (2.15). With $\lambda_3=0$, the coordinates of (2.15) are still in the Plebański class [the factor C_3 in w^α is allowable because n is not defined uniquely by (1.5); see the comment after (1.7)]. The formulae (2.15) still apply, with $\lambda_3=0$, and the Killing equations involving $k_{(2)}$ are solved by

$$
\begin{aligned}
g_{11}&=y^2+(by)^2 g_{22}(z)-2byh_{12}(z)+h_{11}(z),\\
g_{12}&=-byg_{22}+h_{12},\quad g_{13}=-byg_{23}(z)+h_{13}(z),
\end{aligned}
\tag{3.1}
$$

the $h_{ij}(z)$, $g_{22}(z)$, $g_{23}(z)$, and $g_{33}(z)$ being arbitrary functions. The following inequalities hold:

$$
g_{33}<0,\quad g_{22}<0,\quad g_{33}-{g_{23}}^2/g_{22}:=h_{33}<0,
\tag{3.2}
$$

for the following reasons, respectively: (i) the rotation field must be spacelike everywhere; (ii) and (iii) the determinant of the metric $g<0$ or else the metric has an unphysical signature. In virtue of (3.2) one can define the functions K_{11},K_{22},K_{33} and H_{12},H_{13},H_{23} by

$$
\begin{aligned}
{K_{11}}^2&=h_{12}^2/g_{22}+(h_{13}-h_{12}g_{23}/g_{22})^2/h_{33}-h_{11},\\
{K_{22}}^2&=-g_{22},\quad {K_{33}}^2=-h_{33},\\
H_{12}&=h_{12}/g_{22},\quad H_{23}=g_{23}/g_{22},\\
H_{13}&=h_{13}/h_{33}-h_{12}g_{23}/(g_{22}h_{33}),
\end{aligned}
\tag{3.3}
$$

the appropriate sign of ${K_{11}}^2$ being guaranteed by the signature $(+---)$ assumed throughout. [Definitions analogous to (3.3) will be introduced in several other cases in this paper. In each case, the appropriate inequalities are fulfilled for the same reasons.] With these definitions, the metric form (2.15)–(3.1) may be written as

$$ds^2=(dt+ydx)^2-(K_{11}dx)^2-\{K_{22}[(-by+H_{12})dx+dy+H_{23}dz]\}^2-[K_{33}(H_{13}dx+dz)]^2. \tag{3.4}$$

Equation (3.4) defines the orthonormal tetrad used in this section, and the tetrad is comoving, i.e., the velocity field coincides with the tetrad vector #0: $u_\alpha={e^0}_\alpha$. Without loss of generality it may be assumed that

$$K_{33}=1, \tag{3.5}$$

because this result is achieved by the coordinate transformation $z'=\int K_{33}dz$ and redefinitions of H_{13} and H_{23}. The transformation leads to a rescaling of w^α in (2.15), but, taking into account the nonuniqueness of the definition of n through (1.5), the new coordinates are effectively still in the Plebański class.

The Einstein equation $G_{01}=0$ is now integrated with the result

$$H_{23}=A_{23}K_{11}/K_{22}, \tag{3.6}$$

(A and B with indices will denote arbitrary constants). Equation (3.6) substituted into $G_{03}=0$ implies $bA_{23}/(2K_{11}K_{22})=0$, and since this is within the case 1.1.1 defined by $b\neq 0$, the result is

$$A_{23}=0=H_{23}. \tag{3.7}$$

Now $G_{02}=0$ implies

$$H_{13}=A_{13}K_{11}K_{22}, \tag{3.8}$$

and $G_{12}=0$ implies

$$H_{12,z}=B_{12}K_{11}/{K_{22}}^3. \tag{3.9}$$

Using (3.7)–(3.9) in $G_{23}=0$ one obtains $bB_{12}/(K_{11}{K_{22}}^2)=0$, and so

$$B_{12}=0, \quad H_{12}=\text{const.} \tag{3.10}$$

The coordinate transformation $t=t'-H_{12}\,x/b$, $y=y'+H_{12}/b$ has the same effect as if

$$H_{12}=0, \tag{3.11}$$

which will be assumed from now on. [The transformation preserves $k^\alpha_{(2)}, k^\alpha_{(3)}, u^\alpha, w^\alpha$ and transforms $k^\alpha_{(1)}$ into ${k'}^\alpha_{(1)}=\delta^\alpha_1+(H_{12}/b)k^\alpha_{(3)}$, i.e., in the Killing fields $H_{12}=0$ can be achieved by a change of basis in the Lie algebra.] In a further analysis two subcases have to be considered separately.

Case 1.1.1.2.1: $A_{13}\neq 0$.

It is convenient to go over to the standard Bianchi-type coordinates in which $g_{33}=-1$ while $g_{03}=g_{13}=g_{23}=0$. The following transformations do it: in the first step we transform

$$t=T-F(z)Y, \quad x=X+[\ln(bF)]/b, \quad y=bYF, \tag{3.12}$$

where $F(z)$ is determined by

$$F_{,z}/F=-bA_{13}K_{22}/[K_{11}(1+A_{13}^2K_{22}^2)] \tag{3.13}$$

(the right-hand side is finite and nonzero because of assumptions made earlier). This leads to $g_{0a}=0$. Then we transform $z=H(Z)$, where

$$\frac{dH}{dZ}=(1+A_{13}{}^2K_{22}{}^2)^{1/2} \tag{3.14}$$

and thereby achieve $g_{33}=-1$. Denoting

$$L_{11}=K_{11}\frac{dH}{dZ}, \quad L_{22}=bFK_{22}, \tag{3.15}$$

we obtain the following metric form:

$$ds^2=[dT+F(Z)(bYdX-dY)]^2-[L_{11}(Z)dX]^2-[L_{22}(Z)(-bYdX+dY)]^2-dZ^2, \tag{3.16}$$

the coordinates no longer being in the Plebański class.

The Einstein equation $G_{02}=0$ implies

$$F_{,z}=AL_{22}/L_{11}, \tag{3.17}$$

where A is an arbitrary constant. Equation (3.17) can be integrated if the new variable z' is introduced by

$$z',_Z=L_{22}/L_{11}; \tag{3.18}$$

then

$$F(z')=Az'+B, \quad B=\text{const.} \tag{3.19}$$

With z' in place of Z, the equation $G_{13}=0$ is solved by

$$L_{11}=CL_{22}\exp\left[-\frac{1}{2}\int (AF/L_{22}{}^2)dz'\right]. \tag{3.20}$$

The only tetrad components of the Einstein tensor that are still nonzero are the diagonal ones. It may be verified that $G_{11}-G_{22}\equiv 0$ in virtue of (3.17)–(3.20), while $G_{11}-G_{33}=0$ follows by differentiation of $G_{33}=\Lambda$ in virtue of (3.20). Hence, the latter is the last equation to solve, and it is

$$(L_{22}L_{22,z'})^2-\tfrac{1}{2}AFL_{22}L_{22,z'}+(b^2+A^2/4)L_{22}{}^2-\tfrac{1}{4}(bF)^2=\Lambda(L_{11}L_{22})^2. \tag{3.21}$$

Equation (3.21) is an integro-differential equation. Substituting for L_{11} from (3.20) and differentiating by z', it can be changed into an ordinary second-order differential equation that determines $L_{22}(z')$, and then L_{11} can be (in principle) calculated from (3.20). No further progress in solving the Einstein equations was achieved in the general case.

The coordinates of (3.16) are still comoving, while the rotation vector field is

$$w^\alpha=L_{11}{}^{-2}(-A\delta_1^\alpha-bAY\delta_2^\alpha+bF\delta_3^\alpha). \tag{3.22}$$

The case $A=0$ that is seen to be simpler is equivalent [via (3.17) and (3.13)] to $A_{13}=0$ or $b=0$, and these were set aside for separate consideration. The matter density is

$$\kappa\epsilon = -3\Lambda + [bF/(L_{11}L_{22})]^2 + (A/L_{11})^2. \tag{3.23}$$

When $\Lambda = 0$, an elementary solution exists. Equation (3.21) has then the integral

$$L_{22}{}^2 = \tfrac{1}{4}\,[F^2 - (A^2 + 4b^2)(z' + D)^2], \tag{3.24}$$

where D is an arbitrary constant, and L_{11} is found from (3.20) to be

$$L_{11} = CW_1^{\alpha_1} W_2^{\alpha_2}, \tag{3.25}$$

where

$$\alpha_{1,2} = \frac{1}{8b^2}\,[(A^2 + 4b^2)^{1/2} \pm A]^2, \tag{3.26}$$

and W_1, W_2 are the two factors of (3.24):

$$W_{1,2} = \tfrac{1}{2}\,[F \mp (A^2 + 4b^2)^{1/2}(z' + D)]. \tag{3.27}$$

A (modestly thorough) search through the literature by this author has not turned up any rotating dust Bianchi-type III solution with the same configuration of the Killing, velocity, and rotation vector fields as here. (The method used in the search will be disclosed in Paper 3.) Hence, (3.16)–(3.27) is tentatively proposed as a new solution.

Case 1.1.1.2.2: $A_{13} = 0$.

The invariant meaning of $A_{13} = 0$ is that the Killing field $k^{\alpha}_{(1)}$ is everywhere orthogonal to the rotation field. We go back to (3.11). The equation $G_{13} = 0$ [in the coordinates of (3.4) with later simplifications] implies

$$K_{22} = A_{22}K_{11}. \tag{3.28}$$

The equation $G_{22} + G_{33} = 2\Lambda$ is integrated with the result

$$(K_{11}K_{11,z})^2 = \Lambda K_{11}{}^4 - bK_{11}{}^2 + C, \tag{3.29}$$

where C is an arbitrary constant. Using this result in $G_{22} = \Lambda$ we obtain

$$C = 1/(4A^2). \tag{3.30}$$

The case $K_{11,z} = 0$ leads to the Gödel solution.[2] When $K_{11,z} \neq 0$, the coordinate transformation,

$$K_{11}(z) = Z, \tag{3.31}$$

leads to the following solution:

$$ds^2 = (dt + ydx)^2 - (Zdx)^2 - [AZ(-bydx + dy)]^2 - [\Lambda Z^2 - b^2 + 1/(2AZ)^2]^{-1} dZ^2. \tag{3.32}$$

The matter–density is

$$\kappa\epsilon = -3\Lambda + 1/(A^2 Z^4). \tag{3.33}$$

With such a simple solution as (3.32) it is rather improbable that it could be still unknown. Nevertheless, the remark made after Eq. (3.27) applies also here.

The subcase $\Lambda = 0$ of (3.32) is a coordinate transform of the subcase $A = 0$ of (3.16)–(3.27).

IV. CASE 1.1.2: $b=0$

We go back to (2.8). From Eqs. (2.3)–(2.8) we have

$$\phi=axy+cC_3x+\alpha(y), \quad \lambda_2=c\lambda_3x+\beta(y), \tag{4.1}$$

where α and β are unknown functions and the other symbols are constants. Applying to the resulting Killing vector fields the change of basis $k'_{(2)}=k_{(2)}-(C_2/C_3)k_{(3)}$ and the coordinate transformation (1.13) with $H=\alpha/(ay+cC_3)$, $T_{,y}=(\beta-c\lambda_3H)/(ay+cC_3)$ (the denominators are nonzero because of the assumption $\phi_{,x}\neq 0$ defining case 1), we obtain the result

$$k^\alpha_{(2)}=cC_3x\,\delta^\alpha_0+ax\,\delta^\alpha_1-(ay+cC_3)\,\delta^\alpha_2+c\lambda_3x\,\delta^\alpha_3, \tag{4.2}$$

the other Killing fields being as in (2.1). The commutation relations are

$$[k_{(1)},k_{(2)}]=ak_{(1)}+ck_{(3)}, \quad [k_{(1)},k_{(3)}]=0=[k_{(2)},k_{(3)}]. \tag{4.3}$$

The Bianchi type of this algebra is III when $a\neq 0$ and II when $a=0\neq c$ ($a=c=0$ is excluded by the assumption $\phi_{,x}\neq 0$ made in case 1).

In order to adapt the coordinates to $k_{(1)}$ and $k_{(3)}$, the following coordinate transformation is now carried out:

$$(t',x',y')=(t,x,y), \quad z'=-\lambda_3t+C_3z, \tag{4.4}$$

after which (primes dropped)

$$u^\alpha=\delta^\alpha_0-\lambda_3\delta^\alpha_3, \quad k^\alpha_{(2)}=cx\,\delta^\alpha_0+ax\,\delta^\alpha_1-(ay+c)\,\delta^\alpha_2, \tag{4.5}$$

but for solving the Killing equations, the cases $a\neq 0$ and $a=0$ have to be considered separately. We first consider the following.

Case 1.1.2.1: $a\neq 0$.

Then the Killing equations imply

$$\begin{gathered}
g_{00}=1+\lambda_3{}^2g_{33}, \quad g_{01}=y+\lambda_3g_{13}, \quad g_{02}=\lambda_3g_{23}, \quad g_{03}=\lambda_3g_{33},\\
g_{11}=-\frac{2c}{a}\,y-\left(\frac{c}{a}\right)^2+\left(\frac{c\lambda_3}{a}\right)^2g_{33}-\frac{2c\lambda_3}{a}\,(ay+c)h_{13}+(ay+c)^2h_{11},\\
g_{12}=-\frac{c\lambda_3}{a(ay+c)}\,h_{23}+h_{12}, \quad g_{13}=-\frac{c\lambda_3}{a}\,g_{33}+(ay+c)h_{13},\\
g_{22}=h_{22}/(ay+c)^2, \quad g_{23}=h_{23}/(ay+c),
\end{gathered} \tag{4.6}$$

where $g_{33}(z)$ and all $h_{ij}(z)$ are arbitrary functions. In order to introduce an orthonormal tetrad, the following new functions are defined:

$$\begin{gathered}
K_{33}{}^2=-g_{33}, \quad H_{22}=h_{22}-h_{23}{}^2/g_{33}, \quad K_{22}{}^2=-H_{22},\\
H_{12}=(h_{12}-h_{13}h_{23}/g_{33})/H_{22},\\
K_{11}{}^2=1/a^2+h_{13}{}^2/g_{33}+H_{12}{}^2H_{22}-h_{11},\\
H_{13}=h_{13}/g_{33}, \quad H_{23}=h_{23}/g_{33}.
\end{gathered} \tag{4.7}$$

After the coordinate transformation,

$$t=T+cx/a, \quad y=(e^{aY}-c)/a, \quad (x',z')=(x,z), \tag{4.8}$$

the metric form becomes

$$ds^2=(dT+a^{-1}e^{aY}dx)^2-(K_{11}e^{aY}dx)^2-[K_{22}(e^{aY}H_{12}dx+dY)]^2-[K_{33}(\lambda_3 dT+H_{13}e^{aY}dx + H_{23}dY+dz)]^2, \tag{4.9}$$

and the Killing fields become

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=ax\,\delta^\alpha_1-\delta^\alpha_2, \quad k^\alpha_{(3)}=\delta^\alpha_0. \tag{4.10}$$

No progress was made with the Einstein equations for (4.9). However, a large subset of solutions by Ellis[3] is contained here when $0=\lambda_3=H_{13}=H_{23}=H_{12}$ (the first condition means that the velocity field is a Killing field; the next two conditions mean that the Killing fields $k^\alpha_{(1)}$ and $k^\alpha_{(2)}$ are orthogonal to rotation; the invariant interpretation of the last condition is unknown). It is the case Ib of Ellis, but not in full generality. Ellis' function t, when expressed as a function of Ellis' $y(x^2)$, obeys

$$2c^2(t^2)_{,yy}+K=0, \tag{4.11}$$

where c and K are constants [it is Ellis' Eq. (4.20), appropriately transformed]. The t^2 implied by (4.11) is a polynomial of second degree. Only when the discriminant Δ of the polynomial is zero will Ellis' case Ib be a subcase of (4.9). The reason for this is that with $\Delta\neq 0$ the Ellis case Ib has a four-dimensional symmetry group acting multiply transitively on 3-dimensional orbits, and the group has no 3-dimensional subgroups. Only with $\Delta=0$ a 3-dimensional subgroup exists.

Case 1.1.2.2: $a=0$

As stated before, the Bianchi type is now II. We go back to (4.5). With $a=0\neq c$, it can be assumed that $c=1$ with no loss of generality. The velocity field and the Killing fields are

$$u^\alpha=\delta^\alpha_0-\lambda_3\delta^\alpha_3, \quad k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=x\,\delta^\alpha_0-\delta^\alpha_2, \quad k^\alpha_{(3)}=\delta^\alpha_0, \tag{4.12}$$

and the Killing equations imply

$$\begin{gathered} g_{00}=1+\lambda_3{}^2 g_{33}, \quad g_{01}=y+\lambda_3 g_{13}, \quad g_{02}=\lambda_3 g_{23}, \quad g_{03}=\lambda_3 g_{33}, \\ g_{11}=(1+\lambda_3{}^2 g_{33})y^2+2\lambda_3 y h_{13}+h_{11}, \\ g_{12}=\lambda_3 y g_{23}+h_{12}, \quad g_{13}=\lambda_3 y g_{33}+h_{13}, \end{gathered} \tag{4.13}$$

where $g_{22}(z)$, $g_{23}(z)$, $g_{33}(z)$ and the $h_{ij}(z)$ are arbitrary functions. The new functions K_{ii} and H_{ij} are defined by (4.7) with $a\to\infty$ and $h_{23}=g_{23}$. The metric is then

$$ds^2=(dt+ydx)^2-(K_{11}dx)^2-(H_{12}dx+K_{22}dy)^2-\{K_{33}[\lambda_3 dt+(\lambda_3 y+H_{13})dx+H_{23}dy+dz]\}^2. \tag{4.14}$$

Again, no progress with the Einstein equations was made at this level of generality. However, further progress was possible when

$$\lambda_3=0, \tag{4.15}$$

which means, from (4.12), that the velocity field itself becomes a Killing field. Then, by a transformation of z and a few redefinitions,

$$K_{33}=1 \tag{4.16}$$

can be achieved, and this will be assumed from now on. Some of the subcases implied by the Einstein equations turn out to be empty, so they will be mentioned only briefly. The first alternative is H_{23} being zero or nonzero. However, in both cases the Einstein equations together with coordinate transformations lead to the conclusion

$$H_{13}=0 \tag{4.17}$$

(which means $g_{\alpha\beta}k^{\alpha}_{(1)}w^{\beta}=0$). Then the equation $G_{01}=0$ is integrated with the result $H_{23}=A_{23}K_{11}K_{22}$ (the A_{ij} being arbitrary constants). Further integration of the Einstein equations must proceed separately for $A_{23}\neq 0$ and for $A_{23}=0$. However, the case $A_{23}\neq 0$ leads to a plain contradiction, and so the only case to consider is

$$H_{23}=0=A_{23} \tag{4.18}$$

(which means $g_{\alpha\beta}k^{\alpha}_{(2)}w^{\beta}=0$). The equation $G_{12}=0$ is then integrated with the result

$$H_{12,z}=H_{12}K_{22,z}/K_{22}+A_{12}K_{11}/K_{22}{}^2. \tag{4.19}$$

From $G_{11}-G_{22}$ we have

$$(K_{11}K_{22,z}-K_{22}K_{11,z})_{,z}-A_{12}{}^2K_{11}/K_{22}{}^3=0. \tag{4.20}$$

With the new variable $u(z)$ defined by

$$u_{,z}=K_{11}/K_{22}{}^3, \tag{4.21}$$

Eq. (4.20) can be integrated with the result

$$(K_{11}/K_{22})^2=-(A_{12}u)^2-2Bu+C:=W(u), \tag{4.22}$$

where B and C are arbitrary constants. With the help of (4.22), K_{11} is eliminated from the Einstein equations, and then $G_{33}=\Lambda$ remains as the only independent equation in the set. It may be written as

$$2WK_{22,u}/K_{22}=A_{12}{}^2u+B\pm(B^2+CA_{12}{}^2+1+4\Lambda WK_{22}{}^4)^{1/2}. \tag{4.23}$$

When $\Lambda=0$, this has simple elementary solutions, but several cases require separate treatment [$A_{12}\neq 0$ with $(B^2+CA_{12}{}^2)$ being positive, zero or negative, $A_{12}=0\neq B$], so (4.23) will be left as the most compact notation. This is again tentatively proposed as a new solution [the metric is defined by (4.14)–(4.19), (4.22), and (4.23)].

When $A_{12}=0$, these equations define a coordinate transform of Ellis' case Aii.[3] Then $H_{12}=B_{12}K_{22}$ from (4.19) and $B_{12}=0$ results by a coordinate transformation. The invariant meaning of the condition $A_{12}=0$ is not known to this author. With such an interpretation at hand, a new invariant definition of the Ellis case Aii would result.

V. CASE 1.2: $C_3=0$

We go back to Eqs. (2.3)–(2.5), and immediately have to consider separately the cases $b\neq 0$ and $b=0$. We take first the following.

Case 1.2.1: $b\neq 0$.

Equations (2.3a–2.3c) then imply

$$\phi=\alpha(y)e^{bx}-ay/b-C_2, \tag{5.1}$$

where α is an arbitrary function, $\alpha\neq 0$ because of the assumption $\phi_{,x}\neq 0$. Equation (2.5h) implies

$$\lambda_3=\mu(y)e^{fx}, \tag{5.2}$$

where μ is an arbitrary function; $\mu\neq 0$ because otherwise the symmetry group becomes two-dimensional. Equation (2.5l) then implies

$$\mu=B\alpha^{f/b}, \quad j=-af/b, \tag{5.3}$$

where B is an arbitrary constant. For further integration, the cases $f\neq b$ and $f=b$ have to be considered separately.

Case 1.2.1.1: $b\neq f$.

The Killing fields that result here, after a simplification of the basis by $k'_{(2)}=k_{(2)}-(a/b)k_{(1)}-[c/(f-b)]k_{(3)}$, are:

$$\begin{gathered} k^{\alpha}_{(1)}=\delta^{\alpha}{}_1, \quad k^{\alpha}_{(3)}=\alpha^{f/b}e^{fx}\delta^{\alpha}{}_3, \\ k^{\alpha}_{(2)}=e^{bx}[(\alpha-y\alpha_{,y})\delta^{\alpha}{}_0+\alpha_{,y}\delta^{\alpha}{}_1-b\alpha\delta^{\alpha}{}_2+\beta\delta^{\alpha}{}_3] \end{gathered} \tag{5.4}$$

(the basis change resulted in $a=c=0$). This algebra is of Bianchi type VI_h with the free parameter $b+f/b-f$. The Killing fields are further simplified by the coordinate transformation (1.13) with $H=b^{-1}\ln\alpha$, $T=\int(b\alpha)^{-1}\beta dy$, and by the subsequent transformation that leads out of the Plebański class:

$$t'=e^{-bx}t, \quad x'=x, \quad y'=bt+y, \quad z'=e^{-fx}z. \tag{5.5}$$

The result is (primes dropped):

$$\begin{gathered} k^{\alpha}_{(1)}=-bt\delta^{\alpha}{}_0+\delta^{\alpha}{}_1-fz\delta^{\alpha}{}_3, \quad k^{\alpha}_{(2)}=\delta^{\alpha}{}_0, \quad k^{\alpha}_{(3)}=\delta^{\alpha}{}_3, \\ u^{\alpha}=e^{-bx}\delta^{\alpha}{}_0+b\delta^{\alpha}{}_2, \quad w^{\alpha}=ne^{-fx}\delta^{\alpha}{}_3. \end{gathered} \tag{5.6}$$

The metric that results from the Killing equations is:

$$\begin{gathered} g_{00}=e^{2bx}(1+b^2h_{22}), \quad g_{01}=e^{bx}(y-bh_{12}), \quad g_{02}=-be^{bx}h_{22}, \\ g_{03}=-be^{(b+f)x}h_{23}, \quad g_{11}=h_{11}, \quad g_{12}=h_{12}, \quad g_{13}=e^{fx}h_{13}, \\ g_{22}=h_{22}, \quad g_{23}=e^{fx}h_{23}, \quad g_{33}=e^{2fx}h_{33}, \end{gathered} \tag{5.7}$$

where $h_{ij}(y), i,j=1,2,3$ are arbitrary functions. No progress with the Einstein equations was made here.

Case 1.2.1.2: $b=f$

The basis of the Killing fields is here:

$$\begin{gathered} k^{\alpha}_{(1)}=\delta^{\alpha}{}_1, \quad k^{\alpha}_{(3)}=B\alpha(y)e^{bx}\delta^{\alpha}{}_3, \\ k^{\alpha}_{(2)}=e^{bx}[(\alpha-y\alpha_{,y})\delta^{\alpha}{}_0+\alpha_{,y}\delta^{\alpha}{}_1-b\alpha\delta^{\alpha}{}_2+(cB\alpha x+\beta)\delta^{\alpha}{}_3] \end{gathered} \tag{5.8}$$

(as before, $a=0$ was achieved by a basis change). The Bianchi type is IV when $c\neq 0$ and V when $c=0$. To this basis we apply the transformation (1.13) with $H=b^{-1}\ln\alpha$, $T=\int b^{-1}(\beta e^{-bH}-cBH)dy$, and then the subsequent transformation:

$$t'=e^{-bx}t, \quad x'=x, \quad y'=bt+y, \quad z'=e^{-bx}(-ctx+z). \tag{5.9}$$

The result (i.e., the relevant vector fields and the metric that fulfills the Killing equations) is:

$$k^\alpha_{(1)}=-bt\delta^\alpha{}_0+\delta^\alpha{}_1-(ct+bz)\delta^\alpha{}_3,\quad k^\alpha_{(2)}=\delta^\alpha{}_0,\quad k^\alpha_{(3)}=\delta^\alpha{}_3,$$

$$u^\alpha=e^{-bx}\delta^\alpha{}_0+b\,\delta^\alpha{}_2-cxe^{-bx}\delta^\alpha{}_3,\quad w^\alpha=ne^{-bx}\delta^\alpha{}_3,$$

$$g_{00}=e^{2bx}(1+b^2h_{22}-2bcxh_{23}+c^2x^2h_{33}),\quad g_{01}=e^{bx}(y-bh_{12}+cxh_{13}),$$

$$g_{02}=e^{bx}(-bh_{22}+cxh_{23}),\quad g_{03}=e^{2bx}(-bh_{23}+cxh_{33}),$$

$$(g_{11},g_{12},g_{22})=(h_{11},h_{12},h_{22}),$$

$$(g_{13},g_{23})=e^{bx}(h_{13},h_{23}),\quad g_{33}=e^{2bx}h_{33},\tag{5.10}$$

where, as before, the h_{ij} are arbitrary functions of y. Also here, no progress was made with the Einstein equations.

Case 1.2.2: $b=0$.

Eqs. (2.3a)–(2.3c) and (2.3h) imply here

$$\phi=axy+\alpha(y),\quad \lambda_3=\mu(y)e^{fx},\tag{5.11}$$

where $\alpha(y)$ and $\mu(y)$ are arbitrary functions; $a\neq0$ because of the assumption defining case 1. In further integration, the cases $f\neq0$ and $f=0$ have to be considered separately. However, $f\neq0$ quickly leads to a contradiction with $a\neq0$, so the only case to consider is

$$f=0.\tag{5.12}$$

Then, from (2.3d),

$$\lambda_2=c\mu(y)x+\beta(y),\tag{5.13}$$

where β is an arbitrary function, and from (2.3l),

$$\mu=Ay^{-j/a},\tag{5.14}$$

where $A\neq0$ is an arbitrary constant. The transformation (1.13) with $H=\alpha/(ay)$, $T=\int[(\beta-cAHy^{-j/a})/(ay)]dy$ leads now to $\alpha=\beta=0$. With no loss of generality we can assume $A=1$. The coordinates will become adapted to $k_{(1)}$ and $k_{(3)}$ after the subsequent transformation:

$$z'=y^{j/a}z.\tag{5.15}$$

This leaves the metric u^α and w^α in the Plebański form (with rescaled n), while the Killing fields and the commutation relations become

$$\begin{gathered}k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=C_2\delta^\alpha_0+ax\delta^\alpha_1-ay\delta^\alpha_2+(cx-jz)\delta^\alpha_3,\quad k^\alpha_{(3)}=\delta^\alpha_3,\\ [k_{(1)},k_{(2)}]=ak_{(1)}+ck_{(3)},\quad [k_{(1)},k_{(3)}]=0,\quad [k_{(2)},k_{(3)}]=jk_{(3)}.\end{gathered}\tag{5.16}$$

In solving the Killing equations the cases $a+j\neq0$ and $a+j=0$ have to be considered separately.

Case 1.2.2.1: $a+j\neq0$.

The algebra (5.16) is then of Bianchi type VI_h [the free parameter in the standard form of the commutation relations is $(1-j/a)/(1+j/a)$]. The Killing equations lead to the following metric form:

$$
\begin{gathered}
g_{00}=1, \quad g_{01}=y, \quad g_{02}=g_{03}=0, \\
g_{11}=C^2y^{-2j/a}h_{33}-2Cy^{1-j/a}h_{13}+y^2h_{11}, \\
g_{12}=-Cy^{-1-j/a}h_{23}+h_{12}, \quad g_{13}=-Cy^{-2j/a}h_{33}+y^{1-j/a}h_{13}, \\
g_{22}=h_{22}/y^2, \quad g_{23}=y^{-1-j/a}h_{23}, \quad g_{33}=y^{-2j/a}h_{33},
\end{gathered}
\tag{5.17}
$$

where the h_{ij} are functions of the variable u defined below, and C is the constant

$$
C=c/(a+j), \quad u=e^t y^{C_2/a}. \tag{5.18}
$$

The limit $C_2=0$ of (5.17) is an allowed subcase and does not require separate treatment. With $C_2=0$ all h_{ij} become functions of t alone, and this is the first instance where a proper spatially homogeneous (necessarily tilted) Bianchi-type model appears in this scheme.

No progress in solving the Einstein equations was achieved here.

Case 1.2.2.2: $a+j=0$.

The Bianchi type is IV when $c\neq 0$ and V when $c=0$. The Killing equations lead here to the metric

$$
\begin{gathered}
g_{00}=1, \quad g_{01}=y, \quad g_{02}=g_{03}=0, \\
g_{11}=\left(\frac{c}{a}\,y\,\ln y\right)^2 h_{33}+2\,\frac{c}{a}\,y^2\,\ln y h_{13}+y^2h_{11}, \\
g_{12}=\frac{c}{a}\,y h_{23}+h_{12}, \quad g_{13}=\frac{c}{a}\,y^2\,\ln y h_{33}+y^2h_{13}, \\
g_{22}=h_{22}/y^2, \quad g_{23}=h_{23}, \quad g_{33}=y^2h_{33},
\end{gathered}
\tag{5.19}
$$

where the h_{ij} are arbitrary functions of the same $u=e^t y^{C_2/a}$ as in (5.18).

No progress with the Einstein equations was made here, either.

VI. CASE 2: $\phi_{,x}=0$

We go back to Eqs. (2.3). Now $\phi_{,y}\neq 0$ can be assumed because with $\phi_{,x}=\phi_{,y}=0$ the Killing field $k^\alpha_{(2)}$ becomes spanned on u^α and w^α, and this situation was already considered in Paper 1.[1]

If $h\neq 0$, then (2.3j) implies $\phi=-(g/h)y+A$, $A=\text{const}$, and so $k^\alpha_{(2)}=(C_2+A)\delta^\alpha_0-(g/h)\delta^\alpha_1+\lambda_2\delta^\alpha_3$. Now $k'^\alpha_{(2)}=k^\alpha_{(2)}+(g/h)k^\alpha_{(1)}$ is spanned on u^α and w^α, and so we are again in the domain of Paper 1. Hence, $h=0$, and, from (2.3j), $g=0$. In the same way it follows from (2.3b) and (2.3f) that

$$
a=b=d=e=g=h=0. \tag{6.1}
$$

From here on, the cases $\lambda_3\neq 0$ and $\lambda_3=0$ have to be considered separately.

Case 2.1: $\lambda_3\neq 0$.

Using (2.3h) with (6.1) in (2.3l) we obtain $f\phi_{,y}=j$. In the same way as above, we conclude from here that

$$
f=j=0. \tag{6.2}
$$

Now from (2.3h) and (2.3d),

$$
\lambda_3=\lambda_3(y), \quad \lambda_2=c\lambda_3 x+\beta(y), \tag{6.3}
$$

where $\lambda_3(y)$ and β are arbitrary. The only equation that remains from the set (2.3) is now $cC_3=0$, and $\phi(y)$ is an arbitrary function. We first follow

Case 2.1.1: $C_3\neq 0$.

Then

$$c=0, \tag{6.4}$$

and the algebra becomes commutative (Bianchi type I). The transformation $z'=C_3z/\lambda_3(y)$ gives the same result as if $\lambda_3=C_3$, and then

$$C_3=\lambda_3=1 \tag{6.5}$$

may be assumed without loss of generality. The Killing fields are now

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=F\delta^\alpha_0+\phi_{,y}\delta^\alpha_1+\beta(y)\delta^\alpha_3, \quad k^\alpha_{(3)}=\delta^\alpha_0+\delta^\alpha_3, \tag{6.6}$$

where

$$F:=C_2+\phi-y\phi_{,y}. \tag{6.7}$$

The Killing equations for $k^\alpha_{(1)}$ and $k^\alpha_{(3)}$ imply that the metric tensor is independent of x and that it depends on t and z only through $u:=(t-z)/2$. It may be assumed that $\phi_{,yy}\neq 0\neq F-\beta$ because with $\phi_{,yy}=0$ the Killing field $k'^\alpha_{(2)}=k^\alpha_{(2)}-\phi_{,y}k^\alpha_{(1)}$ is spanned on u^α and w^α (this is the domain of Paper 1), while $F-\beta=0$ leads, through the Killing equations, to a singular metric [$\det(g_{\alpha\beta})=0$]. Knowing this, we can adapt coordinates to all three Killing vectors by the following transformation:

$$t=Ft'+z', \quad x=\phi_{,y}t'+x', \quad y=y', \quad z=\beta t'+z'. \tag{6.8}$$

In the new coordinates, all the metric components depend only on y'. With primes dropped, the Killing, velocity, and rotation fields and the metric tensor are as follows:

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=\delta^\alpha_0, \quad k^\alpha_{(3)}=\delta^\alpha_3, \quad u^\alpha=(F-\beta)^{-1}(\delta^\alpha_0-\phi_{,y}\delta^\alpha_1-\beta\delta^\alpha_3),$$

$$w^\alpha=\{C_3n/[\lambda_3(F-\beta)]\}(\delta^\alpha_0-\phi_{,y}\delta^\alpha_1-F\delta^\alpha_3),$$

$$g_{00}=F^2-\beta^2+2y\phi_{,y}(F-\beta)+\beta^2g_{33}+2\phi_{,y}\beta g_{13}+\phi^2_{,y}g_{11}, \tag{6.9}$$

$$g_{01}=y(F-\beta)+\phi_{,y}g_{11}+\beta g_{13},$$

$$g_{02}=\phi_{,y}g_{12}+Fg_{23}, \quad g_{03}=F-\beta+\beta g_{33}+\phi_{,y}g_{13},$$

and all the $g_{ij}(y)$, $i,j=1,2,3$ are arbitrary functions.

No progress with the Einstein equations was achieved in this case.

Case 2.1.2: $C_3=0$.

This means that the Killing field $k^\alpha_{(3)}$ is collinear with the rotation vector w^α. Now we have to consider separately the cases $c\neq 0$ and $c=0$.

Case 2.1.2.1: $c\neq 0$.

The Bianchi type of the algebra (2.2) is II in this case. The transformation (1.13) with $H=-\beta/(c\lambda_3)$ leads to

$$\beta=0. \tag{6.10}$$

It may be taken for granted that $F\neq 0$ [F is still given by (6.7)] because with $F=\beta=0$ the Killing equations imply $g_{13}=g_{23}=g_{33}=0$ and so $g_{03}=0$ from (6.9), i.e., $\det(g_{\alpha\beta})=0$. The transformation $z'=z/\lambda_3$ leads to the same result as

$$\lambda_3=1 \tag{6.11}$$

(in fact, the transformation reshuffles the components g_{2j}, $j=1,2,3$ among themselves, but formally the new metric still has the Plebański form). Hence (6.10) and (6.11) will be assumed. The subsequent transformation,

$$t=Ft',\quad x=\phi_{,y}t'+x',\quad y=y',\quad z=\tfrac{1}{2}\,c\phi_{,y}t'^2+z', \tag{6.12}$$

leads to (with primes dropped)

$$\begin{gathered} k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=\delta^\alpha_0+cx\delta^\alpha_3,\quad k^\alpha_{(3)}=\delta^\alpha_3,\\ u^\alpha=F^{-1}(\delta^\alpha_0-\phi_{,y}\delta^\alpha_1-ct\phi_{,y}\delta^\alpha_3),\quad w^\alpha=(n/\lambda_3)\delta^\alpha_3. \end{gathered} \tag{6.13}$$

In the coordinates of (6.13) the Killing equations imply

$$\begin{gathered} g_{00}=(C_2+\phi)^2-(y\phi_{,y})^2+\phi^2_{,y}h_{11},\\ g_{01}=yF-ct\phi_{,y}h_{13}+\phi_{,y}h_{11},\quad g_{02}=\phi_{,y}h_{12},\quad g_{03}=\phi_{,y}h_{13},\\ g_{11}=(ct)^2g_{33}-2cth_{13}+h_{11},\quad g_{12}=-ctg_{23}+h_{12},\quad g_{13}=-ctg_{33}+h_{13}, \end{gathered} \tag{6.14}$$

where g_{22}, g_{23}, g_{33}, and the h_{ij} are arbitrary functions of y.

No progress with the Einstein equations was made.

Case 2.1.2.2: $c=0$.

We go back to Eqs. (6.1)–(6.3) with $c=C_3=0$. The algebra becomes commutative (Bianchi type I), but in contrast to Case 2.1.1 the Killing field $k^\alpha_{(3)}$ is here collinear with rotation. We have

$$\begin{gathered} \lambda_2=\beta(y),\quad \lambda_3=\lambda_3(y),\quad \phi=\phi(y),\quad k^\alpha_{(1)}=\delta^\alpha_1,\\ k^\alpha_{(2)}=F\delta^\alpha_0+\phi_{,y}\delta^\alpha_1+\beta\delta^\alpha_3,\quad k^\alpha_{(3)}=\lambda_3\delta^\alpha_3, \end{gathered} \tag{6.15}$$

where F is given by (6.7). Also here $F\neq 0$, or else we are back in the domain of Paper 1. Equation (6.11) applies here for the same reason as before, and the coordinates become adapted to the Killing fields after the transformation:

$$t=Ft',\quad x=\phi_{,y}t'+x',\quad y=y',\quad z=\beta t'+z'. \tag{6.16}$$

In the new coordinates (primes dropped)

$$\begin{gathered} k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=\delta^\alpha_0,\quad k^\alpha_{(3)}=\delta^\alpha_3,\\ u^\alpha=F^{-1}(\delta^\alpha_0-\phi_{,y}\delta^\alpha_1-\beta\delta^\alpha_3),\quad w^\alpha=(n/\lambda_3)\delta^\alpha_3, \end{gathered} \tag{6.17}$$

and the Killing equations imply

$$\begin{gathered} g_{00}=(C_2+\phi)^2-(y\phi_{,y})^2+\phi^2_{,y}g_{11}+2\beta\phi_{,y}g_{13}+\beta^2g_{33},\\ g_{01}=yF+\phi_{,y}g_{11}+\beta g_{13},\quad g_{02}=\phi_{,y}g_{12}+\beta g_{23},\quad g_{03}=\phi_{,y}g_{13}+\beta g_{33}, \end{gathered} \tag{6.18}$$

where all $g_{ij}, i,j=1,2,3$ are arbitrary functions of y.

In general, no progress was made with the Einstein equations. However, a few authors have considered this case before, and a few simple exact solutions in this class are known. Therefore, we shall introduce the standard Bianchi-type coordinates in which $g_{02}=g_{12}=g_{23}=0$, $g_{22}=-1$ in order to facilitate the comparison. The transformation to coordinates in which $g_{i2}=0$, $i=0,1,3$, is

$$t=t'+F_0(y'), \quad x=x'+F_1(y'), \quad y=F_2(y')=y', \quad z=z'+F_3(y'), \tag{6.19}$$

where the functions $F_\alpha(y')$ obey

$$\sum_{\alpha=0}^{3} g_{\alpha i}F_{\alpha,y'}=0, \quad i=0,1,3. \tag{6.20}$$

The set (6.20) may be solved for $F_{0,y'}$, $F_{1,y'}$, and $F_{3,y'}$ and then finding the F_α only requires calculation of a few integrals (which are seen to exist, although the integrands are at this point not known as explicit functions of y'). Having achieved $g_{i2}=0$ in this way, we carry out the change of variable $y'=\int(-g_{22})^{1/2}dy$ [knowing that $g_{22}<0$ because of the signature and $g_{22}=g_{22}(y)$ because of the symmetries] that leads to $g_{2'2'}=-1$. The components g_{ij} in which $i\neq 2\neq j$, the velocity and the rotation do not change after (6.19). In the new coordinates, dropping primes,

$$ds^2=[(C_2+\phi)dt+Ydx]^2-[k_{11}(\phi_{,Y}dt+dx)]^2-dy^2-\{k_{33}[(\beta+h_{13}\phi_{,Y})dt+h_{13}dx+dz]\}^2, \tag{6.21}$$

where Y is the y-coordinate of (6.18), and

$$k_{11}{}^2=y^2-g_{11}+g_{13}{}^2/g_{33}, \quad k_{33}{}^2=-g_{33}, \quad h_{13}=g_{12}/g_{33}. \tag{6.22}$$

The subcase $\beta=h_{13}=0$ of (6.21) was first considered by King,[4] and it is known in the literature as ''stationary cylindrically symmetric.'' King demonstrated that, even in this subcase, the problem is underdetermined: one function in the metric may be chosen arbitrarily [in (6.21) it is k_{33}]. The limitation $\beta=h_{13}=0$ resulted from the reflection symmetries assumed in Ref. 4, that in the coordinates of (6.21) correspond to $z\to -z$ and $(t,x)\to(-t,-x)$. King's metric ansatz can be also derived from the following assumptions, as follows from the present consideration.

1. The manifold has a 3-dimensional symmetry group whose algebra is of the Bianchi type I.
2. One Killing field $(k^\alpha_{(3)})$ is collinear with rotation, the two others are linearly independent of u^α and w^α.
3. The velocity vector field is spanned on $k^\alpha_{(1)}$ and $k^\alpha_{(2)}$ (hence $\beta=0$).
4. The Killing fields $k^\alpha_{(1)}$ and $k^\alpha_{(2)}$ are both orthogonal to $k^\alpha_{(3)}$ (hence $h_{13}=0$).

A few examples of explicit solutions of the Einstein equations are given in Ref. 4, among them the solutions of Maitra (Ref. 8; see below) and of Lanczos (Ref. 10; this one goes by the name of ''Ehlers–van Stockum'' in Ref. 4). Vishveshwara and Winicour[5] derived the same metric form as King and provided another explicit example of a solution. The solution derived by Hoenselaers and Vishveshwara[6] that was supposed to be an example of the collection of Ref. 5, turned out to be a coordinate transform of the Gödel metric; see Ref. 7.

One more example was provided by Maitra.[8] In addition to the list above, it has the following invariant property.

5. The timelike Killing field $k_{(2)}$ has unit length so that $(C_2+\phi)^2-k_{11}{}^2\phi_{,Y}{}^2=1$.

The conditions 1–5 are still insufficient to reduce (6.21) to the Maitra solution; the following coordinate-dependent relations must hold in addition:

$$(C_2+\phi)Y-k_{11}{}^2\phi_{,Y}=m, \quad Y^2-k_{11}{}^2=m^2-r^2, \tag{6.23}$$

where $r(y)$ is a new coordinate defined by

$$\ln\left(\frac{dy}{dr}\right) = -\frac{1}{4u^2}\left\{(1+u^2)^{1/2} - 1 + \frac{1}{8} - \frac{1}{4}\ln\left[\frac{1}{2}(1+u^2)^{1/2} + \frac{1}{2}\right]\right\}$$

$$u := 2r/a, \quad a = \text{const}, \tag{6.24}$$

and $m(r)$ is the function

$$m = -\tfrac{1}{2}a\{(1+u^2)^{1/2} - 1 - \ln[\tfrac{1}{2}(1+u^2)^{1/2} + \tfrac{1}{2}]\}. \tag{6.25}$$

This author was not able to interpret (6.23) in invariant terms.

The first three of the six solutions by Ozsváth[9] also belong here, and they are subcases of the class considered by King. All of Ozsváth's solutions have 4-dimensional symmetry groups whose orbits are the whole 4-dimensional manifolds. In order to place specific Ozsváth's solutions in the classification considered here, one has to identify 3-dimensional subgroups of Ozsváth's groups. Examples can be spotted by inspection in Ref. 9 in which different non-isomorphic 3-dimensional subgroups are contained in the same 4-dimensional group. Hence, the same Ozsváth's solutions should come up as limits in different classes of the present investigation. For unique and complete identification, the formulae for group generators are necessary, and these are not given for most of Ozsváth's solutions.

Case 2.2: $\lambda_3 = 0$.

We go back to Eq. (6.1). Now necessarily $C_3 \neq 0$, so it may be assumed with no loss of generality that

$$C_3 = 1, \tag{6.26}$$

and then Eqs. (6.1) together with (2.3a), (2.3e), and (2.3i) imply

$$c = f = j = 0, \tag{6.27}$$

and (2.3d) implies

$$\lambda_2 = \beta(y). \tag{6.28}$$

In this case, necessarily $\beta \neq 0$ because with $\beta = 0$ the Killing equations imply that the determinant of the metric is zero. With (6.1) and (6.26)–(6.28) the whole set (2.3) is solved and the Killing fields are

$$k^\alpha_{(1)} = \delta^\alpha_1, \quad k^\alpha_{(2)} = F\delta^\alpha_0 + \phi_{,y}\delta^\alpha_1 + \beta\delta^\alpha_3, \quad k^\alpha_{(3)} = \delta^\alpha_0, \tag{6.29}$$

where F is still given by (6.7). It may be assumed with no loss of generality that

$$C_2 = 0, \tag{6.30}$$

because this is equivalent to changing the basis to $k'^\alpha_{(2)} = k^\alpha_{(2)} - C_2 k^\alpha_{(3)}$. The Killing fields all commute to zero, so the Bianchi type is I once more, but this time with a still different position of the orbits with respect to the hydrodynamical vector fields; now the velocity field is the Killing field $k^\alpha_{(3)}$, i.e., it is tangent to the orbits.

In order to adapt the coordinates to the Killing fields we carry out the transformation

$$t' = t - (F/\beta)z, \quad x' = x - (\phi_{,y}/\beta)z, \quad y' = y, \quad z' = z/\beta. \tag{6.31}$$

The Killing fields $k^\alpha_{(1)}$ and $k^\alpha_{(3)}$ do not change, while $k^\alpha_{(2)}$, the rotation and the metric acquire the form

$$k^\alpha_{(2)}=\delta^\alpha_3, \quad w^\alpha=(n/\beta)(F\delta^\alpha_0+\phi_{,y}\delta^\alpha_1-\delta^\alpha_3),$$
$$g_{00}=1, \quad g_{01}=y, \quad g_{02}=0, \quad g_{03}=\phi, \tag{6.32}$$

other components g_{ij} are arbitrary functions of y. Similarly as it was done in (6.19)–(6.21), we can transform the metric to the Bianchi-type coordinates in which $g_{02}=g_{12}=g_{23}=0$. Then

$$ds^2=(dt+ydx+\phi dz)^2-(k_{11}dx)^2-(k_{22}dy)^2-[k_{33}(h_{13}dx+dz)]^2, \tag{6.33}$$

where k_{11}, k_{22}, k_{33}, and h_{13} are functions of y. The transformation does not change (6.32). The Einstein equation $G_{01}=0$ then implies

$$h_{13}\phi_{,y}=1-Kk_{11}k_{22}/k_{33}, \tag{6.34}$$

where K is a constant. With this, $G_{03}=0$ implies

$$\phi_{,y}k_{11}/(k_{22}k_{33})-Kh_{13}=L=\text{const}, \tag{6.35}$$

and $G_{13}=0$ implies

$$-k_{33}{}^3h_{13,y}/(k_{11}k_{22})+K\phi=M=\text{const}. \tag{6.36}$$

Equations (6.34)–(6.36) can be used to elliminate all derivatives of ϕ and h_{13} from the Einstein tensor. This is done as follows.

(1) The derivative of (6.34) is used to eliminate $h_{13}\phi_{,yy}$.
(2) Equation (6.34) is used to eliminate $(h_{13}\phi_{,y})^2$, $h_{13}\phi_{,y}$ and $h_{13}\phi_{,y}{}^2$.
(3) From (6.35), $\phi_{,yy}$ is found and eliminated from the Einstein tensor completely.
(4) From (6.36), $h_{13,yy}$ is found and eliminated from the Einstein tensor.
(5) Equations (6.36) and (6.35) are used to eliminate $h_{13,y}$ and $\phi_{,y}$.

This procedure is designed so that the result of each step applies also in the limit $h_{13}=0$. After it is completed, the Einstein tensor is diagonal.

The equation $G_{11}-G_{33}=0$ is now

$$\tfrac{1}{2}(K/k_{33})^2+k_{22}{}^{-2}[-k_{22,y}k_{33,y}/(k_{22}k_{33})+k_{33,yy}/k_{33}+k_{11,y}k_{22,y}/(k_{11}k_{22})-k_{11,yy}/k_{11}]$$
$$-\tfrac{1}{2}[(Kh_{13}+L)/k_{11}]^2-(K\phi-M)^2/k_{33}{}^4=0. \tag{6.37}$$

In this, we replace one power of $(Kh_{13}+L)$ from (6.35), one power of $(K\phi-M)$ from (6.36), multiply the result by $k_{11}k_{22}k_{33}$ and use (6.34) to eliminate $Kk_{11}k_{22}/k_{33}$ from $K^2k_{11}k_{22}/k_{33}$. The result is then integrated, and the integral is

$$(k_{11}k_{33,y}-k_{33}k_{11,y})/k_{22}-K\phi h_{13}-\tfrac{1}{2}L\phi+Mh_{13}+\tfrac{1}{2}Ky=A=\text{const}. \tag{6.38}$$

The equation $G_{11}-G_{22}=0$ is

$$-\tfrac{1}{2}(Kh_{13}+L)^2+k_{22}{}^{-2}k_{33}{}^{-1}(-k_{22,y}k_{33,y}/k_{22}+k_{33,yy}-k_{11,y}k_{33,y}/k_{11})=0. \tag{6.39}$$

In this, we eliminate one power of $(Kh_{13}+L)$ using (6.35), then use (6.34) to eliminate $h_{13}\phi_{,y}$ from the result. The equation thus obtained is integrated to

$$k_{33,y}/(k_{11}k_{22})-\frac{1}{2}L\phi-\frac{1}{2}Ky+\frac{1}{2}K^2\int(k_{11}k_{22}/k_{33})dy=B=\text{const.} \tag{6.40}$$

Now we introduce the new variable $u(y)$ and the new functions $F(y)$ and $G(y)$ by

$$u_{,y}=k_{22}k_{33}/k_{11}, \quad F=Kh_{13}+L, \quad G=K\phi-M. \tag{6.41}$$

In these variables, Eqs. (6.35) and (6.36) may be rewritten as

$$G_{,u}=KF, \quad F_{,u}=K(k_{11}{}^2/k_{33}{}^4)G. \tag{6.42}$$

These can be separated when each is differentiated by u:

$$\begin{gathered} G_{,uu}-[(Kk_{11})^2/k_{33}{}^4]G=0, \\ F_{,uu}+(4k_{33,u}/k_{33}-2k_{11,u}/k_{11})F_{,u}-[(Kk_{11})^2/k_{33}{}^4]F=0. \end{gathered} \tag{6.43}$$

Equations (6.43) determine F and G as functions of the k_{ii} (k_{22} is hidden in the definition of u). Then, (6.34) implicitly defines k_{22} as a function of k_{11} and k_{33}. Next, (6.40) defines k_{33} as a function of k_{11}, and finally (6.38) defines $k_{11}(y)$. Hence, the set is in principle solvable. The one remaining Einstein equation is $G_{22}=\Lambda$. However, at this point it may be verified that $G_{22,y}=0$ in virtue of the other equations, so $G_{22}=\Lambda$ merely defines Λ in terms of the other constants. If $\Lambda=0$, then $G_{22}=0$ imposes an algebraic relation of the constants.

Suppose that $\phi=\text{const}$. Then the coordinate transformation $t=t'-\phi z$ leads to $\phi=0$ (which means $g_{\alpha\beta}k^{\alpha}_{(2)}k^{\beta}_{(3)}=0$). Then (6.34) implies $K\neq 0$, and so (6.35) implies $h_{13}=\text{const}$. Consequently, the transformation $z=z'-h_{13}x$ leads to $h_{13}=0$ (i.e., $g_{\alpha\beta}k^{\alpha}_{(1)}k^{\beta}_{(2)}=0$). With $\phi=h_{13}=0$ the metric (6.32)–(6.33) becomes identical to the metric ansatz of Lanczos,[10] and the Lanczos solution uniquely follows. In fact, with $\phi=\text{const}$, the Killing field $k^{\alpha}_{(2)}$ becomes the second Killing field spanned on u^{α} and w^{α}, and so we land in the domain of Paper 1. Hence, the definition of the Lanczos solution given above coincides with one of those from Paper 1.

The class defined by (6.32)–(6.43) contains the case Ciii of Ellis.[3]

At this point, the whole collection of metrics considered in this paper is exhausted.

VII. CONCLUDING REMARKS

In this paper, several classes of metrics were only derived. Investigation of their properties was postponed to separate projects; at this point it is not clear which of them would deserve further pursuit in the first turn, while investigating them all would expand the paper beyond acceptable limits.

All the cases considered here have the property that one of the Killing fields ($k_{(3)}$) is spanned on the fields of velocity u^{α} and rotation w^{α} of the source, while the other two are at every point linearly independent of u^{α} and w^{α}. In every case, the explicit formulae for the Killing fields were found, and the Killing equations were solved for the Killing fields were found, and the Killing equations were solved for the components of the metric tensor. Progress with solving the Einstein equations differed from case to case. The complete listing of nonempty cases is given in Fig. 1, and the results obtained are as follows.

(1) In case 1.1.1.1 (Bianchi type III) the final result is given by Eq. (2.18). No exact solutions are known and none were derived here.
(2) In case 1.1.1.2.1 (Bianchi type III) the final results are given by (3.16)–(3.27). A (most probably) new solution was found for the subcase $\Lambda=0$, given by Eqs. (3.16), (3.18), (3.19), and (3.24)–(3.27).

(3) In case 1.1.1.2.2 (Bianchi type III) a (most probably) new solution was derived here, given by Eq. (3.32).
(4) In case 1.1.2.1 (Bianchi type III) the final result is given by Eq. (4.9). No explicit solutions are known in general, but subcases of the case Ib of Ellis[3] belong here [see the paragraph after Eq. (4.10)].
(5) In case 1.1.2.2 (Bianchi type II) the final result is given by Eqs. (4.14)–(4.23). No explicit solutions are known in general, but with $\lambda_3 = 0 = \Lambda$, a new exact solution was derived here; see the comment after Eq. (4.23). The case Aii of Ellis[3] is a subcase here.
(6) In case 1.2.1.1 (Bianchi type VI_h) the final result is (5.6)–(5.7).
(7) In case 1.2.1.2 (Bianchi types IV and V) the final result is (5.10).
(8) In case 1.2.2.1 (Bianchi type VI_h), the final result is (5.17).
(9) In case 1.2.2.2 (Bianchi types IV and V), the final result is (5.19).
(10) In case 2.1.1 (Bianchi type I), the final result is (6.9).
(11) In case 2.1.2.1 (Bianchi type II), the final result is (6.14). No explicit solutions are known in points 6–11.
(12) In case 2.1.2.2 (Bianchi type I), the final result is (6.21). No explicit solutions are known in general, but a certain subcase (known as ''stationary cylindrically symmetric'') was considered by King[4] and Vishveshwara and Winicour.[5] Examples of explicit solutions were given in Refs. 4 and 5, and also by Maitra[8] and Ozsváth;[9] see the text after Eq. (6.25).
(13) In case 2.2 (Bianchi type I), the final result is given by Eqs. (6.32)–(6.43). No explicit solutions are known.

The Bianchi types do not uniquely identify the various cases because in each case the orbit of the symmetry group has a different position with respect to the velocity and rotation fields. This is why the same Bianchi types occur in inequivalent cases.

The collection of results corresponding to all three Killing vectors being at every point linearly independent of velocity and rotation will be presented in Paper 3 (now in preparation). Paper 3 will also contain an overview of literature on solutions of Einstein's equations with rotating fluid source.

ACKNOWLEDGMENT

Calculations for this paper were done with use of the algebraic computer program Ortocartan.[11,12]

[1] A. Krasiński, J. Math. Phys. **38**, 380 (1997).
[2] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[3] G. F. R. Ellis, J. Math. Phys. **8**, 1171 (1967).
[4] A. R. King, Commun. Math. Phys. **38**, 157 (1974).
[5] C. V. Vishveshwara and J. Winicour, J. Math. Phys. **18**, 1280 (1977).
[6] C. Hoenselaers and C. V. Vishveshwara, Gen. Relativ. Gravit. **10**, 43 (1979).
[7] S. K. Chakraborty, Gen. Relativ. Gravit. **12**, 925 (1980).
[8] S. C. Maitra, J. Math. Phys. **7**, 1025 (1966).
[9] I. Ozsváth, J. Math. Phys. **6**, 590 (1965).
[10] K. Lanczos, Z. Phys. **21**, 73 (1924); English translation in Gen. Relativ. Gravit. **29**, 363 (1997).
[11] A. Krasiński, Gen. Relativ. Gravit. **25**, 165 (1993).
[12] A. Krasiński and M. Perkowski, Gen. Relativ. Gravit. **13**, 67 (1981).

Quantum space–time and classical gravity

J. Madore
Laboratoire de Physique Théorique et Hautes Energies,[a] *Université de Paris-Sud, Bât. 211, F-91405 Orsay, France*

J. Mourad
GPS, Université de Cergy-Pontoise, Site de St. Martin, F-95302 Cergy-Pontoise, France

(Received 7 July 1997; accepted for publication 9 September 1997)

A method has been recently proposed for defining an arbitrary number of differential calculi over a given noncommutative associative algebra. As an example a version of quantized space–time is considered here. It is found that there is a natural differential calculus over this version of quantized space–time using which the only possible torsion-free, metric-compatible, linear connection has zero curvature. It is then the noncummutative version of Minkowski space–time. Perturbations of this calculus are shown to give rise to nontrivial gravitational fields. © *1998 American Institute of Physics.* [S0022-2488(98)00601-X]

I. MOTIVATION AND NOTATION

Since the early days of quantum field theory physicists have been tempted to introduce some type of lattice structure on space–time to avoid the appearance of ultraviolet divergences, that is, to fix a natural value for an ultraviolet cut-off Λ. One of the disadvantages of these discrete structures is the fact that they destroy Lorentz invariance and they can hardly be considered in any respect as fundamental. It was Snyder[1] who first had the idea of using noncommuting coordinates to mimic a discrete structure in a covariant way. Since then several attempts have been made to continue this initial effort. We refer to Madore and Mourad[2] for a recent review with a historical perspective. One typically introduces four Hermitian generators q^μ of a noncommutative $*$-algebra $\mathscr{A}$ which satisfy commutation relations of the form

$$[q^\mu, q^\nu] = i\mu_P^{-2} q^{\mu\nu}. \tag{1.1}$$

The problem lies then with the interpretation of the right-hand side. One can define a succession of elements $q^{\lambda_1 \cdots \lambda_n}$ by the equations

$$[q^\lambda, q^{\mu\nu}] = i\mu_P^{-1} q^{\lambda\mu\nu}, \quad [q^\lambda, q^{\mu\nu\rho}] = i\mu_P^{-1} q^{\lambda\mu\nu\rho}, \tag{1.2}$$

and so forth. The structure of the algebra $\mathscr{A}$ is constrained by the value of these commutators. One possibility, considered by Snyder, is to choose them so as to form a representation of the Lie algebra of the de Sitter group. A second possibility, considered by Dubois-Violette and Madore,[3,4] is to choose them to form a representation of the conformal algebra. Recently Doplicher, Fredenhagen and Roberts[5,6] have argued that $q^{\mu\nu}$ should be chosen to lie in the center $\mathscr{Z}(\mathscr{A})$ of $\mathscr{A}$, that the q-tensors with more than 2 indices should vanish. We shall adopt this as a working hypothesis. There are 6 independent $q^{\mu\nu}$, which from (1.1) parameterize symplectic structures on space–time. In the commutative limit one obtains therefore a space of dimension greater than four except of course if the $q^{\mu\nu}$ are nilpotent. We shall not address here the question of the physical significance of the extra dimensions.

[a]Laboratoire associé au CNRS, URA D0063.

The idea that a noncommutative structure gives rise to an ultraviolet cut-off has been developed by several authors[6–8] since the original work of Snyder.[1,9] It is the right-hand arrow of the diagram

$$\begin{array}{ccc} \mathscr{A} & \Leftarrow & \Omega^*(\mathscr{A}) \\ \Downarrow & & \Uparrow \\ \text{Cut-off} & & \text{Gravity} \end{array}$$

The top arrow is a mathematical triviality; the $\Omega^*(\mathscr{A})$ is what gives a differential structure to the algebra. We shall define and discuss it in Section II. In Section IV we shall argue that each gravitational field is the unique ''shadow'' in the limit $\mu_P \to \infty$ of some differential structure over some non-commutative algebra. This would define the left-hand arrow of the diagram.

The composition of the three arrows is an expression of an old idea, due to Pauli and developed by Deser[10] and others,[11] that perturbative ultraviolet divergences will one day be regularized by the gravitational field. The possibility which we would like to defend is that the mechanism by which this works is through the introduction of non-commuting ''coordinates'' such as the q^μ. A hand-waving argument can be given[12] which allows one to think of the noncommutative structure of space–time as being due to quantum fluctuations of the light-cone in ordinary 4-dimensional space-time. This relies on the existence of quantum gravitational fluctuations. A purely classical argument based on the formation of black-holes has been also given.[6] In both cases the classical gravitational field is to be considered as regularizing the ultraviolet divergences through the introduction of the noncommutative structure of space–time. This can be strengthened as the conjecture that the classical gravitational field and the noncommutative nature of space–time are two aspects of the same thing. It is our purpose here to explore this relation.

In Section II we give a general prescription of how one defines differential calculi based on a set of derivations and we determine the conditions which the derivations must satisfy for the module of 1-forms to be a free module and for the module of 2-forms to be nontrivial. In Section III we associate a metric and a linear connection to each differential calculus and we argue that the association is unique. In Section IV we give the explicit calculations using a particularly simple set of derivations and we find that the associated gravitational field is trivial. In Sections V and VI we consider differential calculi which are, in a sense which we shall make precise, small perturbations of this basic differential calculus and we find that they lead to a rather simple but nontrivial gravitational field. In Section VII we consider a quotient algebra which could be considered as the noncommutative version of de Sitter space. Finally in Section VIII we discuss briefly the definition of ''gauge invariance.''

II. THE DIFFERENTIAL CALCULI

Let $\mathscr{A}$ be any unital associative $*$-algebra. Of the many differential calculi which can be constructed over $\mathscr{A}$ the largest is the differential envelope or universal differential calculus $(\Omega_u^*(\mathscr{A}), d_u)$. Every other differential calculus can be considered as a quotient of it. For the definitions we refer, for example, to the book by Connes.[13] Let $(\Omega^*(\mathscr{A}), d)$ be another differential calculus over $\mathscr{A}$. Then there exists a unique d_u-homomorphism ϕ,

$$\begin{array}{ccccccccc} \mathscr{A} & \overset{d_u}{\to} & \Omega_u^1(\mathscr{A}) & \overset{d_u}{\to} & \Omega_u^2(\mathscr{A}) & \overset{d_u}{\to} & \cdots \\ \| & & \phi_1 \downarrow & & \phi_2 \downarrow & & \\ \mathscr{A} & \overset{d}{\to} & \Omega^1(\mathscr{A}) & \overset{d}{\to} & \Omega^2(\mathscr{A}) & \overset{d}{\to} & \cdots \end{array} \tag{2.1}$$

of $\Omega_u^*(\mathscr{A})$ onto $\Omega^*(\mathscr{A})$. It is given by

$$\phi(d_u f\ ) = df. \tag{2.2}$$

The restriction ϕ_p of ϕ to each Ω_u^p is defined by

$$\phi_p(f_0 d_u f_1 \cdots d_u f_p) = f_0 df_1 \cdots df_p .$$

Consider a given algebra $\mathscr{A}$ and suppose that we know how to construct an $\mathscr{A}$-module $\Omega^1(\mathscr{A})$ and an application,

$$\mathscr{A} \overset{d}{\to} \Omega^1(\mathscr{A}). \tag{2.3}$$

Then using (2.1) there is a method of constructing $\Omega^p(\mathscr{A})$ for $p \geqslant 2$ as well as the extension of the differential. Since we know $\Omega_u^1(\mathscr{A})$ and $\Omega^1(\mathscr{A})$ we can suppose that ϕ_1 is given. It is the map such that

$$\Omega^1(\mathscr{A}) = \Omega_u^1(\mathscr{A})/\mathrm{Ker}\ \phi_1 .$$

We must construct $\Omega^2(\mathscr{A})$. The largest consistent choice would be to set

$$\Omega^2(\mathscr{A}) = \Omega_u^2(\mathscr{A})/\overline{d_u \mathrm{Ker}\ \phi_1}, \tag{2.4}$$

where

$$\overline{d_u \mathrm{Ker}\ \phi_1} = d_u \mathrm{Ker}\ \phi_1 + \Omega_u^1(\mathscr{A}) \otimes \mathrm{Ker}\ \phi_1 + \mathrm{Ker}\ \phi_1 \otimes \Omega_u^1(\mathscr{A})$$

is the bimodule generated by $d_u \mathrm{Ker}\ \phi_1$. Since

$$\Omega_u^2(\mathscr{A}) = \Omega_u^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega_u^1(\mathscr{A}),$$

we find that $\Omega^2(\mathscr{A})$ can be written also as

$$\Omega^2(\mathscr{A}) = \Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A})/\mathscr{K}, \tag{2.5}$$

with

$$\mathscr{K} = (\phi_1 \otimes \phi_1)(\overline{d_u \mathrm{Ker}\ \phi_1}) = (\phi_1 \otimes \phi_1)(d_u \mathrm{Ker}\ \phi_1). \tag{2.6}$$

It can happen that $\mathscr{K} = \Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A})$, in which case $\Omega^2(\mathscr{A}) = 0$.

Let π be the projection

$$\Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A}) \overset{\pi}{\to} \Omega^2(\mathscr{A}). \tag{2.7}$$

Then π has a right inverse ι, a map

$$\Omega^2(\mathscr{A}) \overset{\iota}{\to} \Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A}), \tag{2.8}$$

such that $\pi \circ \iota = 1$. This will allow us to identify $\Omega^2(\mathscr{A})$ as a submodule of $\Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A})$. The map ϕ_2 is defined to be the projection of $\Omega_u^2(\mathscr{A})$ onto $\Omega^2(\mathscr{A})$ so defined. From the definition of π one sees that ϕ_2 is given by

$$\phi_2 = \pi \circ (\phi_1 \otimes \phi_1). \tag{2.9}$$

The wedge product of two elements ξ and η in $\Omega^1(\mathscr{A})$ is given by $\xi\eta=\pi(\xi\otimes\eta)$. Let ξ_u be an inverse image of ξ in $\Omega^1_u(\mathscr{A})$. Then the map d from $\Omega^1(\mathscr{A})$ to $\Omega^2(\mathscr{A})$ can be written in terms of d_u as

$$d(\phi_1(\xi_u))=\phi_2(d_u\xi_u). \tag{2.10}$$

Equation (2.4) defines the largest set of 2-forms consistent with the constraints on $\Omega^1(\mathscr{A})$. The procedure can be continued by iteration to arbitrary order in p. Define for this the map ψ_p,

$$\Omega^p_u(\mathscr{A})\xrightarrow{\psi_p}\bigotimes_1^p\Omega^1(\mathscr{A}), \tag{2.11}$$

given by $\psi_p=\phi_1\otimes\cdots\otimes\phi_1$ and define[13]

$$\Omega^p(\mathscr{A})=\Omega^p_u(\mathscr{A})/(\mathrm{Ker}\ \psi_p+d_u\mathrm{Ker}\ \psi_{p-1}). \tag{2.12}$$

We have then by definition

$$\mathrm{Ker}\ \phi_p=\mathrm{Ker}\ \psi_p+d_u\mathrm{Ker}\ \psi_{p-1}. \tag{2.13}$$

For example,

$$\mathrm{Ker}\ \psi_2=\Omega^1_u(\mathscr{A})\otimes\mathrm{Ker}\ \phi_1+\mathrm{Ker}\ \phi_1\otimes\Omega^1_u(\mathscr{A}), \tag{2.14}$$

and using (2.13) we find (2.4) as a particular case. Equation (2.12) can be rewritten as

$$\Omega^p(\mathscr{A})=\bigotimes_1^p\Omega^1(\mathscr{A})/\psi_p(d_u\mathrm{Ker}\ \psi_{p-1}),$$

or in the form

$$\Omega^p(\mathscr{A})=\bigotimes_1^p\Omega^1(\mathscr{A})/\mathscr{K}_p, \tag{2.15}$$

with the $\mathscr{K}_p$ defined by the recurrence relations

$$\mathscr{K}_p=\mathscr{K}_{p-1}\otimes\Omega^1(\mathscr{A})+\Omega^1(\mathscr{A})\otimes\mathscr{K}_{p-1},\quad \mathscr{K}_2=\mathscr{K}.$$

In particular we find the expression

$$\Omega^3(\mathscr{A})=\frac{\Omega^1(\mathscr{A})\otimes\Omega^1(\mathscr{A})\otimes\Omega^1(\mathscr{A})}{\mathscr{K}\otimes\Omega^1(\mathscr{A})+\Omega^1(\mathscr{A})\otimes\mathscr{K}} \tag{2.16}$$

for the module of 3-forms.

To initiate the above construction we define the 1-forms using a set of derivations. We shall suppose that they are interior and exclude therefore the case where $\mathscr{A}$ is commutative. For each integer n let λ_i be a set of n linearly independent anti-Hermitian elements of $\mathscr{A}$ and introduce the derivations $e_i=\mathrm{ad}\lambda_i$. In general the e_i do not form a Lie algebra but they do however satisfy commutation relations as a consequence of the commutation relations of $\mathscr{A}$. We shall suppose that if an element of $\mathscr{A}$ commutes with all of the λ_i then it is in the center $\mathscr{Z}(\mathscr{A})$ of $\mathscr{A}$: g,

$$e_if=0\Rightarrow f\in\mathscr{Z}(\mathscr{A}).$$

Define the map (2.3) by

$$df(e_i)=e_i f=[\lambda_i,f]. \tag{2.17}$$

We shall suppose that there exists a set of n elements θ^i of $\Omega^1(\mathscr{A})$ such that

$$\theta^i(e_j)=\delta^i_j. \tag{2.18}$$

In the examples which we consider we shall show that the θ^i exist by explicit construction. We shall refer to the set of θ^i as a frame or Stehbein. It commutes with the elements $f\in\mathscr{A}$,

$$f\theta^i=\theta^i f. \tag{2.19}$$

The $\mathscr{A}$-bimodule $\Omega^1(\mathscr{A})$ is generated by all elements of the form fdg or of the form $(df)g$. Because of the Leibniz rule these conditions are equivalent. By definition,

$$fdg(e_i)=fe_i g,\quad (dg)f(e_i)=(e_i g)f.$$

Using the frame we can write these as

$$fdg=(fe_i g)\theta^i,\quad (dg)f=(e_i g)f\theta^i. \tag{2.20}$$

The commutation relations of the algebra constrain the relations between fdg and $(dg)f$ for all f and g. Because of the commutation relations of the algebra or, equivalently, because of the kernel of ϕ_1 in the quotient (2.4) the θ^i satisfy in general commutation relations. With the identification ι we have

$$\pi(\theta^i\otimes\theta^j)=P^{ij}{}_{kl}\theta^k\otimes\theta^l,\quad P^{ij}{}_{kl}\in\mathscr{Z}(\mathscr{A}). \tag{2.21}$$

Since π is a projection we have

$$P^{ij}{}_{mn}P^{mn}{}_{kl}=P^{ij}{}_{kl}, \tag{2.22}$$

and the product $\theta^i\theta^j$ satisfies

$$\theta^i\theta^j=P^{ij}{}_{kl}\theta^k\theta^l. \tag{2.23}$$

The module $\mathscr{K}$ is generated by the elements $(\delta^i_k\delta^j_l-P^{ij}{}_{kl})\theta^k\otimes\theta^l$. In one important case which we shall consider the θ^i anticommute. This corresponds to

$$P^{ij}{}_{kl}=\frac{1}{2}(\delta^i_k\delta^j_l-\delta^j_k\delta^i_l). \tag{2.24}$$

Define $\theta=-\lambda_i\theta^i$. Then one sees that

$$df=-[\theta,f], \tag{2.25}$$

and it follows that as a bimodule $\Omega^1(\mathscr{A})$ is generated by one element. Under the condition (2.18) the $\Omega^1(\mathscr{A})$ is free of rank n as a left or right module. It can therefore be identified with the direct sum of n copies of $\mathscr{A}$:

$$\Omega^1(\mathscr{A})=\bigoplus_1^n\mathscr{A}. \tag{2.26}$$

In this representation θ^i is given by the element of the direct sum with the unit in the ith position and zero elsewhere.

Any element θ^i_u of $\Omega^1_u(\mathscr{A})$ can be written in the form

$$\theta_u^i = \sum_\alpha f_\alpha^{(i)} \otimes g_\alpha^{(i)}, \tag{2.27}$$

with the $f_\alpha^{(i)}$ and $g_\alpha^{(i)}$ elements of $\mathscr{A}$ such that

$$\sum_\alpha f_\alpha^{(i)} g_\alpha^{(i)} = 0. \tag{2.28}$$

Let θ^i be the images of θ_u^i in $\Omega^1(\mathscr{A})$. Then the condition (2.18) can be rewritten as

$$\sum_\alpha f_\alpha^{(i)} \lambda_j g_\alpha^{(i)} = \delta_j^i. \tag{2.29}$$

The assumption that $\Omega^1(\mathscr{A})$ is free then is equivalent to the assumption that these equations have a solution for $f_\alpha^{(i)}$ and $g_\alpha^{(i)}$.

The $f_\alpha^{(i)}$ and $g_\alpha^{(i)}$ can be used to give an explicit representation of $\mathscr{K}$. Introduce the notation

$$\mathscr{A}[\theta_u^i, \mathscr{A}] = \left\{ \sum_\alpha f_\alpha [\theta_u^i, g_\alpha] \,|\, f_\alpha, g_\alpha \in \mathscr{A} \right\}.$$

Then it is easily seen that

$$\text{Ker } \phi_1 = \sum_i \mathscr{A}[\theta_u^i, \mathscr{A}]. \tag{2.30}$$

If we define then the elements ν^i of $\Omega^1(\mathscr{A}) \otimes_{\mathscr{A}} \Omega^1(\mathscr{A})$ by

$$\nu^i = \sum_\alpha f_\alpha^{(i)} \theta \otimes \theta g_\alpha^{(i)}, \tag{2.31}$$

a short calculation yields the characterization

$$\mathscr{K} = \sum_i \mathscr{A}[\nu^i, \mathscr{A}]. \tag{2.32}$$

If we define $\lambda = (\lambda_1, \ldots, \lambda_n)$ then the condition (2.29) can be written in the form

$$\mathscr{A}[\lambda, \mathscr{A}] = \bigoplus_1^n \mathscr{A}. \tag{2.33}$$

More generally let $\mathscr{B}$ be an $\mathscr{A}$-bimodule and λ an element of $\mathscr{B}$. Define a map

$$\mathscr{A} \overset{d}{\longrightarrow} \mathscr{B}, \tag{2.34}$$

by $df = [\lambda, f]$. Then we can set

$$\Omega^1(\mathscr{A}) = \mathscr{A}[\lambda, \mathscr{A}] \subset \mathscr{B}. \tag{2.35}$$

The choice $\mathscr{B} = \mathscr{A} \otimes \mathscr{A}$ with $\lambda = 1 \otimes 1$ yields the universal calculus. We see from this example that λ itself need not be an element of $\Omega^1(\mathscr{A})$.

From (2.31) we see that $\pi(\nu^i)$ can be written as

$$\pi(\nu^i)=\sum_\alpha f_\alpha^{(i)}\lambda_j\lambda_k P^{jk}{}_{lm}g_\alpha^{(i)}\theta^l\otimes\theta^m. \tag{2.36}$$

But from the identity $[\pi(\nu^i),f]=0$ we see also that

$$\pi(\nu^i)=\frac{1}{2}F^i{}_{jk}\theta^j\theta^k, \tag{2.37}$$

with $F^i{}_{jk}$ elements in $\mathcal{Z}(\mathcal{A})$ such that

$$P^{jk}{}_{lm}F^i{}_{jk}=F^i{}_{lm}\,. \tag{2.38}$$

Using (2.29) it follows that

$$\sum_\alpha f_\alpha^{(i)}K_{jk}g_\alpha^{(i)}=0,$$

if we define K_{jk} by the equation

$$2\lambda_l\lambda_m P^{lm}{}_{jk}-\lambda_i F^i{}_{jk}-K_{jk}=0. \tag{2.39}$$

The exterior derivative of θ^i is given by

$$d\theta^i=\sum_\alpha df_\alpha^{(i)}dg_\alpha^{(i)}=-[\theta,\theta^i]-\pi(\nu^i). \tag{2.40}$$

The bracket is a graded bracket. Multiplying both sides of this equation by λ_i we find the identity

$$d\theta+\theta^2=-\theta^2+\pi(\nu^i\lambda_i).$$

Using (2.39) we find that this can be written in the form

$$d\theta+\theta^2=-\frac{1}{2}K_{ij}\theta^i\theta^j, \tag{2.41}$$

and if we take the exterior derivative of (2.25) we see immediately that the coefficients K_{ij} must lie in $\mathcal{Z}(\mathcal{A})$.

The structure elements $C^i{}_{jk}$ are defined by the equation

$$d\theta^i=-\frac{1}{2}C^i{}_{jk}\theta^j\theta^k. \tag{2.42}$$

From (2.40) it follows that

$$C^i{}_{jk}=F^i{}_{jk}-2\lambda_l P^{(li)}{}_{jk}\,. \tag{2.43}$$

We have started from an integer n and a set of λ_i. The necessary and sufficient conditions for the existence of the basis θ^i are expressed in Equation (2.29). If $\Omega^2(\mathcal{A})$ is nontrivial there exists $P^{ij}{}_{kl}\neq 0$ in $\mathcal{Z}(\mathcal{A})$ such that (2.22) and (2.23) are satisfied. Conversely we could have started from elements $P^{ij}{}_{kl}$, $F^i{}_{jk}$, K_{ij} in $\mathcal{Z}(\mathcal{A})$ and looked for a solution λ_i to Equation (2.39). Define

$$C^{ij}{}_{kl}=\delta^i_k\delta^j_l-2P^{ij}{}_{kl}\,. \tag{2.44}$$

Then from (2.22) we find that

$$C^{ij}{}_{kl}C^{kl}{}_{mn}=\delta^i_m\delta^j_n.$$

From the associativity rule for the product in $\Omega^3(\mathscr{A})$ one finds that $C^{ij}{}_{kl}$ must satisfy a weak form of the Yang–Baxter equation.

Quite generally let $\mathscr{V}$ be an $\mathscr{A}$-module and π a module morphism,

$$\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}\overset{\pi}{\to}\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}, \tag{2.45}$$

with $\pi^2=\pi$. This is the algebraic generalization of the product given by (2.7). The equivalent generalization of the product in the 3-forms is a module morphism,

$$\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}\overset{\pi'}{\to}\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}, \tag{2.46}$$

with $\pi'^2=\pi'$. One has then

$$\text{Ker }\pi'=\text{Ker }\pi\otimes\mathscr{V}+\mathscr{V}\otimes\text{Ker }\pi. \tag{2.47}$$

For the product to be nontrivial we must require that $\pi'\neq 0$. Since we have

$$\pi'(\text{Ker }\pi\otimes\mathscr{V})=0,\quad \pi'(\mathscr{V}\otimes\text{Ker }\pi)=0,$$

there must exist two morphisms ϕ', ϕ'' of $\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}\otimes_{\mathscr{A}}\mathscr{V}$ into itself such that

$$\pi'=\phi'\circ\pi_{12},\quad \pi'=\phi''\circ\pi_{23}.$$

We have here used the standard convention of setting $\pi_{12}=\pi\otimes 1$ and $\pi_{23}=1\otimes\pi$. The associativity rule becomes the compatibility condition

$$\phi'\circ\pi_{12}=\phi''\circ\pi_{23}. \tag{2.48}$$

In particular if $C^{ij}{}_{kl}$ satisfies the Yang–Baxter condition,

$$C_{23}\circ C_{12}\circ C_{23}=C_{12}\circ C_{23}\circ C_{12},$$

then one can set

$$\phi'=-\frac{1}{3}(1-4\pi_{12}\circ\pi_{23}),\quad \phi''=-\frac{1}{3}(1-4\pi_{23}\circ\pi_{12}). \tag{2.49}$$

However more general solutions to (2.48) do exist.

The $F^i{}_{jk}$ must satisfy a set of modified Jacobi identities. If we choose the λ_i so that the e_i are a basis of the Lie algebra of all derivations of a matrix algebra[14] then one can choose $P^{ij}{}_{kl}=\delta^i_k\delta^j_l$. If a smaller Lie algebra[4] is chosen then $P^{ij}{}_{kl}$ is given by (2.24), $K_{ij}=0$ and the $F^i{}_{jk}$ are equal to the structure constants of the Lie algebra. An example with $K_{ij}\neq 0$ can be found.[15] Examples with $F^i{}_{jk}=0$ and $K_{ij}=0$ have also been found.[16] If $P^{ij}{}_{kl}$ is given by (2.24) with a plus instead of a minus sign, n is even and $F^i{}_{jk}=0$ then a solution to (2.39) is given by Dirac matrices. If also $K_{ij}=0$ then a solution is given by ''super-coordinates.'' In these two cases the 1-forms θ^i commute.

III. THE LINEAR CONNECTIONS

The definition of a connection as a covariant derivative was given an algebraic form in the Tata lectures by Koszul[17] and generalized to noncommutative geometry by Karoubi[18] and

Connes.[19,13] We shall use here the expressions ''connection'' and ''covariant derivative'' synonymously. A ''bimodule connection'' is a connection on a general bimodule $\mathcal{M}$, which satisfies a left and right Leibniz rule. In the particular case where $\mathcal{M}$ is the module of 1-forms we shall speak of a ''linear connection.''

Let $\mathcal{A}$ be an arbitrary algebra and $(\Omega^*(\mathcal{A}),d)$ a differential calculus over $\mathcal{A}$. One defines a left $\mathcal{A}$-connection on a left $\mathcal{A}$-module $\mathcal{H}$ as a covariant derivative,

$$\mathcal{H}\overset{D}{\to}\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\mathcal{H}, \tag{3.1}$$

which satisfies the left Leibniz rule,

$$D(f\psi)=df\otimes\psi+fD\psi, \tag{3.2}$$

for arbitrary $f\in\mathcal{A}$. This map has a natural extension,

$$\Omega^*(\mathcal{A})\otimes_{\mathcal{A}}\mathcal{H}\overset{D}{\to}\Omega^*(\mathcal{A})\otimes_{\mathcal{A}}\mathcal{H}, \tag{3.3}$$

given, for $\psi\in\mathcal{H}$ and $\alpha\in\Omega^p(\mathcal{A})$, by

$$D(\alpha\psi)=d\alpha\otimes\psi+(-1)^p\alpha D\psi.$$

The operator D^2 is necessarily left-linear.

A covariant derivative on the module $\Omega^1(\mathcal{A})$ must satisfy (3.2). But $\Omega^1(\mathcal{A})$ has also a natural structure as a right $\mathcal{A}$-module and one must be able to write a corresponding right Leibniz rule in order to construct a bilinear curvature. Quite generally let $\mathcal{M}$ be an arbitrary bimodule. Consider a covariant derivative,

$$\mathcal{M}\overset{D}{\to}\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\mathcal{M}, \tag{3.4}$$

which satisfies both a left and a right Leibniz rule. In order to define a right Leibniz rule which is consistent with the left one, it was proposed[20–22] that one introduce a generalized permutation,

$$\mathcal{M}\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})\overset{\sigma}{\to}\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\mathcal{M}.$$

The right Leibniz rule is given then as

$$D(\xi f\ )=\sigma(\xi\otimes df\ )+(D\xi)f, \tag{3.5}$$

for arbitrary $f\in\mathcal{A}$ and $\xi\in\mathcal{M}$. The purpose of the map σ is to bring the differential to the left while respecting the order of the factors. It is necessarily bilinear. Consider the case

$$\mathcal{M}=\Omega^1(\mathcal{A})$$

and let π be the projector defined by (2.7). It can be shown[20,23] that a necessary as well as sufficient condition for torsion to be right-linear is that σ satisfy the consistency condition,

$$\pi\circ(\sigma+1)=0. \tag{3.6}$$

We define a bimodule $\mathcal{A}$-connection to be the couple (D,σ). Using the fact that π is a projection one sees that the most general solution to the constraint (3.6) is given by

$$1+\sigma=(1-\pi)\circ\tau, \tag{3.7}$$

where τ is an arbitrary map,

$$\Omega^1(\mathscr{A})\otimes\Omega^1(\mathscr{A})\overset{\tau}{\to}\Omega^1(\mathscr{A})\otimes\Omega^1(\mathscr{A}). \tag{3.8}$$

If we choose $\tau=2$ then we find $\sigma=1-2\pi$ and $\sigma^2=1$. The eigenvalues of σ are then equal to ± 1.

There is at the moment no general consensus of the correct definition of the curvature of a bimodule connection. The problem is that the operator D^2 need not in general be right-linear. Nevertheless in the particular cases of interest to us here, with a module of 1-forms which is free and has a special basis such that (2.19) is satisfied, the ordinary definition of curvature is quite satisfactory.[15,16]

This general formalism can be applied in particular to the differential calculi which we have constructed in the previous section. Since $\Omega^1(\mathscr{A})$ is a free module the maps σ and τ can be defined by their action on the basis elements:

$$\sigma(\theta^i\otimes\theta^j)=S^{ij}{}_{kl}\theta^k\otimes\theta^l,\quad \tau(\theta^i\otimes\theta^j)=T^{ij}{}_{kl}\theta^k\otimes\theta^l. \tag{3.9}$$

By the sequence of identities,

$$fS^{ij}{}_{kl}\theta^k\otimes\theta^l=\sigma(f\theta^i\otimes\theta^j)=\sigma(\theta^i\otimes\theta^j f)=S^{ij}{}_{kl}f\theta^k\otimes\theta^l, \tag{3.10}$$

and the corresponding ones for $T^{ij}{}_{kl}$, we conclude that the coefficients $S^{ij}{}_{kl}$ and $T^{ij}{}_{kl}$ must lie in $\mathscr{Z}(\mathscr{A})$. From (3.7) the most general form for $S^{ij}{}_{kl}$ is

$$S^{ij}{}_{kl}=T^{ij}{}_{mn}(\delta^m_k\delta^n_l-P^{mn}{}_{kl})-\delta^i_k\delta^j_l. \tag{3.11}$$

Since $\Omega^1(\mathscr{A})$ is a free module, a covariant derivative can be defined by its action on the basis elements:

$$D\theta^i=-\omega^i{}_{jk}\theta^j\otimes\theta^k. \tag{3.12}$$

The coefficients here are elements of the algebra. The torsion 2-form is defined as usual as

$$\Theta^i=d\theta^i-\pi\circ D\theta^i.$$

There is a natural covariant derivative $D_{(0)}$ given by[15]

$$D_{(0)}\theta^i=-\theta\otimes\theta^i+\sigma(\theta^i\otimes\theta)-\nu^i, \tag{3.13}$$

which is torsion-free by (2.40). The corresponding coefficients are given by

$$\omega_{(0)}{}^i{}_{jk}=\lambda_l(S^{il}{}_{jk}-\delta^l_j\delta^i_k)+\frac{1}{2}F^i{}_{jk}. \tag{3.14}$$

The most general D is of the form

$$D=D_{(0)}+\chi, \tag{3.15}$$

where χ is an arbitrary bimodule morphism,

$$\Omega^1(\mathscr{A})\overset{\chi}{\to}\Omega^1(\mathscr{A})\otimes\Omega^1(\mathscr{A}). \tag{3.16}$$

If we write

$$\chi(\theta^i) = -\chi^i{}_{jk}\theta^j \otimes \theta^k, \tag{3.17}$$

then by an argument similar to (3.10) we conclude that

$$\chi^i{}_{jk} \in \mathcal{Z}(\mathcal{A}). \tag{3.18}$$

In general a covariant derivative is torsion-free provided the condition,

$$\omega^i{}_{lm} P^{lm}{}_{jk} = \frac{1}{2} C^i{}_{jk}, \tag{3.19}$$

is satisfied. The covariant derivative (3.15) is torsion free if and only if

$$\pi \circ \chi = 0. \tag{3.20}$$

One can define a metric by the condition

$$g(\theta^i \otimes \theta^j) = g^{ij}, \tag{3.21}$$

where the coefficients g^{ij} are elements of the algebra. To be well defined on all elements of the tensor product $\Omega^1(\mathcal{A}) \otimes_{\mathcal{A}} \Omega^1(\mathcal{A})$ the metric must be bilinear and by the sequence of identities,

$$f g^{ij} = g(f\theta^i \otimes \theta^j) = g(\theta^i \otimes \theta^j f) = g^{ij} f, \tag{3.22}$$

one concludes that the coefficients must lie in $\mathcal{Z}(\mathcal{A})$. This restriction plays an important role in the unicity argument which allows us to define the rightarrow of the diagram in Section 1. In the commutative limit the g^{ij} cannot be functions of the coordinates. In ordinary geometry an equivalence class of moving frames determines a metric and all equivalence classes correspond to the same differential calculus, the ordinary de Rham differential calculus. The choice of differential calculus does not fix the metric. In the noncommutative case on the other hand, as we have defined it, each differential calculus determines a Stehbein and thereby a metric. In the commutative limit all of the noncommutative differential calculi are either singular, if n is not equal to the classical dimension of the manifold, or have a common limit. The moving frame and the associated metric remain however as a shadow of the noncommutative structure.

The covariant derivative (3.12) is compatible with the metric if and only if [23]

$$\omega^i{}_{jk} + \omega_{kl}{}^m S^{il}{}_{jm} = 0. \tag{3.23}$$

When $F^i{}_{jk} = 0$ the condition that (3.14) be metric compatible can be written as

$$S^{im}{}_{ln} g^{np} S^{kj}{}_{mp} = g^{ik} \delta^j_l. \tag{3.24}$$

The metric we have chosen is not symmetric with respect to σ. That is

$$g^{ij} \neq S^{ij}{}_{kl} g^{kl},$$

in general. If one wishes to find a metric symmetric in the above sense then one must consider (3.24) as an equation for S and the metric and add the additional equation

$$g^{ij} = S^{ij}{}_{kl} g^{kl}. \tag{3.25}$$

The system (3.24), (3.25), if it has a solution, would yield a symmetric metric with a compatible connection.

Since we are primarily interested in the first-order effects in the commutative limit we can identify the curvature with the operator D^2. We then set as usual

$$D^2\theta^i = -\frac{1}{2}R^i{}_{jkl}\theta^k\theta^l\otimes\theta^j. \tag{3.26}$$

Since D^2 is not necessarily right-linear as an operator the last of the equivalent of the sequence of identities (3.10) is not valid and we cannot conclude that the coefficients $R^i{}_{jkl}$ necessarily lie in the center of the algebra.

IV. THE BASIC CALCULUS

In this section we shall return to the ''space–time'' algebra defined in Section I and we shall suppose, with Doplicher *et al.*, that the $q^{\mu\nu}$ lie in the center of the algebra. This permits us to suppose further that the matrix $q^{\mu\nu}$ has an inverse $q^{-1}_{\lambda\mu}$:

$$q^{-1}_{\lambda\mu}q^{\mu\nu} = \delta^\nu_\lambda.$$

We shall use this inverse to lower the indices of the generators q^μ:

$$\tilde{q}_\lambda = \mu_P^2 q^{-1}_{\lambda\mu}q^\mu.$$

A natural choice of n is $n=4$ and a natural choice of λ_μ is given by

$$\lambda_\mu = -i\tilde{q}_\mu. \tag{4.1}$$

The associated derivations defined in Section II satisfy then

$$e_\mu q^\lambda = \delta^\lambda_\mu, \tag{4.2}$$

and it follows that

$$[e_\mu, e_\nu] = 0. \tag{4.3}$$

From (4.2) it follows that

$$\theta^\lambda = dq^\lambda, \quad \theta = i\tilde{q}_\lambda dq^\lambda, \tag{4.4}$$

from which we deduce that

$$P^{\mu\nu}{}_{\rho\sigma} = \frac{1}{2}(\delta^\mu_\rho\delta^\nu_\sigma - \delta^\nu_\rho\delta^\mu_\sigma), \quad F^\lambda{}_{\mu\nu} = 0, \quad K_{\mu\nu} = i\mu_P^2 q^{-1}_{\mu\nu}. \tag{4.5}$$

One can interpret θ as a connection on a trivial bundle with the unitary elements of the algebra as a structural group.[14,24] We see from the above formula for $K_{\mu\nu}$ that $q^{\mu\nu}$ is related to the corresponding curvature. This is the noncommutative analogue of the classical result of mechanics which interprets the symplectic 2-form as the curvature of a line bundle.

From the commutation relations (2.19) one finds that the θ^λ anticommute. A possible form for σ is given therefore by

$$S^{\mu\nu}{}_{\rho\sigma} = \delta^\nu_\rho\delta^\mu_\sigma. \tag{4.6}$$

From (3.14) we see that in this case the coefficients of the connection necessarily lie in the center of the algebra. From (3.6) we see that the most general $S^{\mu\nu}{}_{\rho\sigma}$ must satisfy the constraint

$$S^{\mu\nu}{}_{[\rho\sigma]} + \delta^{\mu}_{[\rho}\delta^{\nu}_{\sigma]} = 0. \tag{4.7}$$

The most general σ is defined by a solution to Equations (3.24) and (4.7). If we restrict the $g_{\lambda\mu}$ to be the components of the Minkowski metric then the unique solution is given by (4.6). The Minkowski metric is then symmetric also with respect to σ. From (3.19) and (3.23) we see that if we require that the torsion vanish then we have

$$\omega^{\lambda}{}_{\mu\nu} = 0. \tag{4.8}$$

The space–time is therefore a noncommutative version of Minkowski space.

It is of interest to notice that Equation (4.2) defines a derivation of the algebra whatever the form of the matrix $q^{\mu\nu}$. The derivation is inner if the matrix is invertible; otherwise it is outer. Let θ^{λ} be a set of Grassmann variables. Define

$$q^{\lambda} = x^{\lambda} + \mu_P^{-1}\theta^{\lambda}, \quad q^{\mu\nu} = -2i\theta^{\mu}\theta^{\nu}. \tag{4.9}$$

Then (1.1) is satisfied and $q^{\mu\nu} \in \mathcal{Z}(\mathcal{A})$. In this case the matrix $q^{\mu\nu}$ is not invertible; it is in fact nilpotent.

Notice also that in this case the center $\mathcal{Z}(\mathcal{A})$ is nontrivial and in fact it is possible to impose that it be a smooth 6-dimensional manifold with the $q^{\mu\nu}$ as coordinates. We can, with Doplicher *et al.*, impose the conditions

$$q_{\mu\nu}q^{\mu\nu} = 0, \tag{4.10}$$

as well as

$$\epsilon_{\mu\nu\rho\sigma}q^{\mu\nu}q^{\rho\sigma} = 12,$$

which is equivalent to

$$q^{-1}_{\mu\nu} = -\frac{1}{3}\epsilon_{\mu\nu\rho\sigma}q^{\rho\sigma}. \tag{4.11}$$

The normalization is chosen for later convenience. The manifold can then be reduced to 4 dimensions. In the limit $\mu_P \to 0$ we have in fact a structure which can be regarded as a 4-dimensional manifold with a noncommutative extension à la Kaluza–Klein similar to the structures which are mentioned, for example, in Madore and Mourad.[2] In the limit $\mu_P \to \infty$ the structure can be considered to be that of an ordinary space–time with an extra 4- or 6-dimensional factor in which the Poisson structure defined by the commutator (1.1) takes its values.[25] An element of $\mathcal{Z}(\mathcal{A})$ can in no way correspond to a function on space–time in the commutative limit. Also if f is an element of $\mathcal{A}$ such that $e_{\mu}f = 0$ then we can only conclude that f is an arbitrary function of the $q^{\mu\nu}$; we cannot conclude that it is proportional to the identity. We regard the nontrivial center as something which is eventually to be eliminated, for example, by choosing $q^{\mu\nu}$ not to lie in the center. To simplify the calculations we shall suppose that the matrix of coefficients $g^{\mu\nu}$ of the metric is symmetric in the ordinary sense of the word we shall impose the condition that it be equal to the matrix of components of the ordinary Minkowski metric.

V. VARIATIONS OF THE CALCULUS

We shall now check the stability of the result (4.8) under the perturbation of the differential calculus. Introduce four arbitrary ''small'' elements f^{λ} of $\mathcal{A}$ and define

$$\tilde{f}_{\lambda} = \mu_P^2 q^{-1}_{\lambda\mu} f^{\mu}. \tag{5.1}$$

Then the elements

$$\lambda'_\mu = -i(\tilde{q}_\mu + \tilde{f}_\mu) \tag{5.2}$$

are ''near'' (4.1). In general, unless the condition (2.39) is satisfied, $\Omega^2(\mathscr{A})=0$ and the curvature will vanish. Impose the condition (2.39) and let $P^{\rho\sigma}_{(1)\mu\nu}$, $F^\lambda_{(1)\mu\nu}$ and $K_{(1)\mu\nu}$ be the first-order perturbations, respectively, of the coefficients. By simple dimensional arguments one can argue that $P^{\mu\nu}_{(1)\rho\sigma}$ must vanish. In fact it must tend to zero when the Planck mass tends to infinity but on the other hand it is without dimension and therefore cannot depend on the Planck mass. Therefore we set

$$P^{\mu\nu}_{(1)\rho\sigma}=0. \tag{5.3}$$

Using (4.5) we find that the linearization of (2.39) yields the equation

$$[\tilde{q}_\mu, \tilde{f}_\nu] - [\tilde{q}_\nu, \tilde{f}_\mu] = iF^\lambda_{(1)\mu\nu}\tilde{q}_\lambda - K_{(1)\mu\nu}. \tag{5.4}$$

Let k_μ be an arbitrary ''small'' 4-vector with the dimension of mass. Then a solution is given by

$$\tilde{f}_\lambda = k_\mu q^\mu \tilde{q}_\lambda \tag{5.5}$$

and

$$F^\lambda_{(1)\mu\nu} = k_{[\mu}\delta^\lambda_{\nu]} + 2q^{-1}_{\mu\nu}q^{\lambda\sigma}k_\sigma, \quad K_{(1)\mu\nu}=0. \tag{5.6}$$

The corresponding frame is given by

$$\theta'^\lambda = (1-k_\rho q^\rho)dq^\lambda + q^{\lambda\rho}k_\rho q^{-1}_{\mu\sigma}q^\sigma dq^\mu. \tag{5.7}$$

It will generate a new differential calculus $\Omega'^*(\mathscr{A})$ which will be in general different from $\Omega^*(\mathscr{A})$. Using (2.42) or (2.43) we find that

$$C^\lambda_{(1)\mu\nu} = F^\lambda_{(1)\mu\nu}. \tag{5.8}$$

The frame (5.7) explicitly breaks Lorentz invariance through the vector k_λ. This is quite natural since there is now present a gravitational field in the commutative limit. What is less desirable is the dependence on the extra ''coordinates'' $q^{\mu\nu}$. This dependence could perhaps be eliminated by averaging over the space of $q^{\mu\nu}$. Suppose there exists a Lorentz-invariant measure with

$$\langle q^{\mu\nu}q^{\rho\sigma}\rangle = \epsilon^{\mu\nu\rho\sigma}. \tag{5.9}$$

The normalization in (4.11) was chosen so that this equation is compatible with the condition $\langle 1\rangle = 1$. We have then

$$\langle \theta'^\lambda\rangle = dq^\lambda - q^\lambda k_\mu dq^\mu \tag{5.10}$$

and

$$\langle F^\lambda_{(1)\mu\nu}\rangle = -k_{[\mu}\delta^\lambda_{\nu]}. \tag{5.11}$$

The particular, simple form of (5.10) and (5.11) is due to the choice of normalization of the measure on the space of $q^{\mu\nu}$.

Let D' be a covariant derivative and define the coefficients $\omega'^\lambda{}_{\mu\nu}$ by the equation

$$D'\theta'^{\lambda} = -\omega'^{\lambda}{}_{\mu\nu}\theta'^{\mu}\otimes\theta'^{\nu},$$

equivalent to (3.12). Because of (4.8) we have $\omega'^{\lambda}{}_{\mu\nu} = \omega^{\lambda}_{(1)\mu\nu}$ and we can write to lowest order

$$D'\theta'^{\lambda} = -\omega^{\lambda}_{(1)\mu\nu}\theta'^{\mu}\otimes\theta'^{\nu}. \tag{5.12}$$

To extend the covariant derivative to the entire module of 1-forms we shall need the expression for the perturbed value σ' of σ. We define the coefficients $S'^{\mu\nu}{}_{\rho\sigma}$ by the equation

$$\sigma'(\theta'^{\mu}\otimes\theta'^{\nu}) = S'^{\mu\nu}{}_{\rho\sigma}\theta'^{\rho}\otimes\theta'^{\sigma},$$

equivalent to (3.12). We expand

$$S'^{\mu\nu}{}_{\rho\sigma} = \delta^{\nu}_{\rho}\delta^{\mu}_{\sigma} + S^{\mu\nu}_{(1)\rho\sigma}. \tag{5.13}$$

Each choice of $S^{\mu\nu}_{(1)\rho\sigma}$ corresponds to a definite choice of covariant derivative. As above, by simple dimensional arguments one can argue that

$$S^{\mu\nu}_{(1)\rho\sigma} = 0. \tag{5.14}$$

Therefore from (3.14) and a proper choice of χ so that (3.23) is satisfied we find that the torsion-free metric connection is given by

$$\omega^{\lambda}_{(1)\mu\nu} = \frac{1}{2}(F^{\lambda}_{(1)\mu\nu} - F_{(1)\mu\nu}{}^{\lambda} + F_{(1)\nu}{}^{\lambda}{}_{\mu}). \tag{5.15}$$

The curvature of the covariant derivative defined by this expression is essentially constant as far as ''space–time'' is concerned. From this point of view it is not particularly interesting. Using (5.11) we find that

$$\langle\omega^{\lambda}_{(1)\mu\nu}\rangle = \frac{1}{2}(k^{\lambda}g_{\mu\nu} - k_{\nu}\delta^{\lambda}_{\mu}). \tag{5.16}$$

To lowest order we find the expression

$$R^{\mu}{}_{\nu\rho\sigma} = \omega^{\mu}_{(1)\rho\tau}\omega^{\tau}_{(1)\sigma\nu} - \omega^{\mu}_{(1)\sigma\tau}\omega^{\tau}_{(1)\rho\nu} - \omega^{\mu}_{(1)\tau\nu}C^{\tau}_{(1)\rho\sigma}, \tag{5.17}$$

for the components of the curvature. It is not of particular interest to give an expression for the expectation value $\langle R^{\mu}{}_{\nu\rho\sigma}\rangle$ since it will depend critically on the probability measure. However the expectation value of the Einstein tensor must be of the form

$$\langle G_{\mu\nu}\rangle = ak_{\mu}k_{\nu} + bk^2 g_{\mu\nu}, \tag{5.18}$$

where a and b are dimensionless constants which depend on the details of the probability measure. At the present preliminary stage of the understanding of the relation between the differential calculus and the resulting curvature it is premature to consider this equation further.

VI. VARIATIONS OF THE ALGEBRA

Another way to obtain a nonvanishing gravitational field is to vary the structure of the algebra $\mathscr{A}$. We introduce 6 ''small'' elements $q^{\mu\nu}_{(1)}$ of $\mathscr{A}$ and define

$$q'^{\mu\nu} = q^{\mu\nu} + q^{\mu\nu}_{(1)}. \tag{6.1}$$

We have then

$$q'^{-1}_{\mu\nu} = q^{-1}_{\mu\nu} + q^{-1}_{\mu\rho} q^{-1}_{\nu\sigma} q^{\rho\sigma}_{(1)} . \tag{6.2}$$

Since we are here primarily interested in the effect of varying the structure of the algebra we keep the basic calculus and set

$$\lambda'_\mu = \lambda_\mu . \tag{6.3}$$

From (1.2) we find

$$[q^\lambda, q'^{\mu\nu}] = [q^\lambda, q^{\mu\nu}_{(1)}] = i\mu_P^{-1} q^{\lambda\mu\nu}_{(1)} .$$

The simplest generalization of the basic algebra is obtained by supposing that $q^{\lambda\mu\nu}_{(1)}$ lies in the center of $\mathscr{A}$. This is the extended model of Doplicher *et al.* If we impose this condition we can choose

$$q^{\mu\nu}_{(1)} = -\mu_P^{-1} q^{\lambda\mu\nu}_{(1)} \widetilde{q}_\lambda . \tag{6.4}$$

We have then

$$e'_\mu q'^\nu = e_\mu q^\nu = \delta^\nu_\mu + \mu_P^{-1} q^{-1}_{\mu\rho} q^{\nu\rho\sigma}_{(1)} \widetilde{q}_\sigma . \tag{6.5}$$

Using (4.5) we find that the linearization of (2.39) yields now the solution

$$F^\lambda_{(1)\rho\sigma} = \mu_P q^{-1}_{\rho\mu} q^{-1}_{\sigma\nu} (q^{\lambda\mu\nu}_{(1)} - q^{\mu\nu\lambda}_{(1)} + q^{\nu\mu\lambda}_{(1)}). \tag{6.6}$$

The corresponding frame is given by

$$\theta'^\lambda = dq^\lambda - \mu_P^{-1} q^{-1}_{\mu\rho} q^{\lambda\rho\sigma}_{(1)} \widetilde{q}_\sigma dq^\mu . \tag{6.7}$$

VII. de SITTER SPACE

To define a noncommutative version of a space which is not topologically trivial we shall use elementary techniques from classical geometry as well as from the quantum mechanics of constrained Hamiltonian systems. A nontrivial topological manifold V can be defined by its imbedding in a flat space of sufficiently high dimension. The algebra $\mathscr{B} = \mathscr{C}(V)$ of continuous functions on V can be identified with the algebra $\mathscr{A}$ of all continuous functions on the enveloping space modulo the ideal $\mathscr{I}$ of continuous functions which vanish on V: $\mathscr{B} = \mathscr{A}/\mathscr{I}$. As a first attempt to define a noncommutative version of de Sitter space we set $\mathscr{A}$ equal to a $*$-algebra generated by a set of elements q^i, $0 \leqslant i \leqslant 4$, which satisfy commutation relations similar to those given by (1.1):

$$[q^i, q^j] = i\mu_P^{-2} q^{ij} . \tag{7.1}$$

We introduce a Minkowski-signature metric with components g_{ij} and we define $\mathscr{I}$ to be the 2-sided ideal generated by the element

$$c_1 = g_{ij} q^i q^j - r^2 , \tag{7.2}$$

where r^2 is a real constant. Two problems present themselves immediately. The q^{ij} cannot be invertible; it would otherwise define a symplectic form on an odd-dimensional manifold. Using the commutation relations (7.1) one sees also that the 2-sided ideal generated by the element c_1 is the entire algebra and therefore $\mathscr{B} = 0$. We saw something similar to this in Section II where we noticed that the 2-sided ideal generated by the element θ is the entire algebra of forms. This second problem is connected with the fact that we are trying to define the manifold V exactly as a submanifold in spite of the fact that its ''points'' are fuzzy and only defined to within the

uncertainty μ_P^{-1}. As a solution to the first problem we add another dimension and we define V as a submanifold of a 6-dimensional space. We let then $0 \leqslant i \leqslant 5$ and we add a constraint

$$c_2 = q^5. \tag{7.3}$$

To circumvent the second problem we follow the example furnished by the quantization of constrained Hamiltonian systems. The two constraints are of second class and Dirac[26] has shown how in this case one can introduce a new bracket with respect to which they commute with each other and with the observables. One can similarly modify the commutation relations (7.1) so that c_1 commutes with the generators of $\mathscr{A}$ and thus with c_2. We introduce the components r^{ij} of an antisymmetric tensor and we set

$$[q^i, q^j]' = i\mu_P^{-2}(q^{ij} - r^{ij}). \tag{7.4}$$

From the condition $[q^i, c_2]' = 0$ we find that

$$r^{i5} = q^{i5}. \tag{7.5}$$

From the condition $[q^i, c_1]' = 0$ we find the equation

$$q_j r^{ij} + r^{ij} q_j = 2q^{ij} q_j, \tag{7.6}$$

for r^{ij}. Suppose that the element $q^{i5}q_i$ of $\mathscr{A}$ is invertible. The appropriate solution is given then by

$$r^{ij} = \frac{1}{2} q^{[ik} q^{j]5} [q_k, (q^{l5} q_l)^{-1}]_+, \tag{7.7}$$

which can also be written in the form

$$r^{ij} = q^{[ik} q^{j]5} q_k (q^{l5} q_l)^{-1} - \frac{i}{2}\, \mu_P^{-2} q^{[ik} q^{j]5} q_{kl} q^{l5} (q^{m5} q_m)^{-2}. \tag{7.8}$$

We define $\mathscr{T}$ to be the 2-sided ideal of $\mathscr{A}$ generated by the two elements c_1 and c_2. By modifying the original structure of the algebra we have obtained a nontrivial quotient algebra $\mathscr{B}$. The modification of the algebra is not ''small'' in the sense of the preceding section. We shall consider $\mathscr{B}$ as a possible noncommutative equivalent of de Sitter space. We refer to Masson[27] for some examples of quotient algebras defined without modification of the bracket.

We define $\widetilde{q}_i$ and λ_i as in Section IV. The associated derivations e_i satisfy then

$$e_i q^j = \delta_i^j - \frac{1}{2}[q_i q^{j5} - \delta_i^5 q^{jk} q_k, (q^{l5} q_l)^{-1}]_+, \tag{7.9}$$

and they no longer commute. They do not even close to form a Lie algebra:

$$[e_i, e_j] = i\mu_P^2 q_{ik}^{-1} q_{jl}^{-1} r^{kl}. \tag{7.10}$$

From the relations

$$e_i c_1 = 0, \quad e_i c_2 = 0, \tag{7.11}$$

it follows that the e_i can be considered as derivations of $\mathscr{B}$.

If we set

$$f_i{}^{jk} = \delta_i^{[j} q^{k]5} + \delta_i^5 q^{jk},$$

then in the commutative limit we can write

$$e_i = (q^{l5} q_l)^{-1} f_i{}^{jk} q_k \partial_j . \tag{7.12}$$

Since $f_i{}^{jk}$ is antisymmetric in the last 2 indices and since $f_i{}^{j5} = 0$ the e_i are vector fields on the de Sitter space. There are 6 of them and they satisfy 2 relations:

$$q_i q^{ij} e_j = 0, \quad q^{i5} e_i = 0. \tag{7.13}$$

We define a differential calculus as in Section IV. It follows from (7.11) and the construction of the differential that

$$dc_1 = 0, \quad dc_2 = 0. \tag{7.14}$$

The differential calculus can be considered then as a calculus $\Omega^*(\mathscr{B})$ over $\mathscr{B}$. It tends as it must in the commutative limit to the algebra of de Rham forms over de Sitter space. Classical de Sitter space is parallelizable but it is not obvious that $\Omega^1(\mathscr{B})$ is free as a left or right module. Because of the relations (7.13) the θ^i dual to the e_i cannot be constructed. Because of the factor $(q^{l5} q_l)^{-1}$ in the expression (7.7) the equation (2.39) has no solution for the λ_i which we have chosen.

VIII. GAUGE INVARIANCE

In (5.15) we calculated the connection associated to a perturbation of the basic differential calculus. One might have expected from the origin (5.2) of the perturbation that it would resemble rather a coordinate transformation and that the perturbed connection would vanish. The fact that this is not the case is due to the existence of the extra ''coordinates'' $q^{\mu\nu}$ which have been used to raise and lower indices. Our formalism is in fact analogous to that which can be used to describe a manifold V which is defined by its imbedding in a flat space of higher dimension so we mention this briefly. Let x^a be the coordinates of the imbedding space and y^α local coordinates of V. Then V is defined locally by functions of the form $x^a(y^\alpha)$. A variation of V to a surface V' is given by functions of the form $x'^a(y^\alpha) = x^a(y^\alpha) + h^a(y^\alpha)$. A variation of the coordinates of the imbedding space can be written in the form $x'^a = x^a + h^a(x^b)$. The most general variation of the metric on V is obtained by a variation $g'_{ab} = g_{ab} + \partial_{(a} h_{b)}$ of the components of the flat metric. In the commutative limit one can consider $(q^\lambda, q^{\mu\nu})$ as the ''coordinates'' of the imbedding space and the conditions $q^{\mu\nu} = 0$ the equations which define V.

Let V be a smooth manifold and ϕ a smooth map of V into itself. Then ϕ induces an automorphism ϕ^* of the algebra $\mathscr{C}(V)$ of smooth functions given by $\phi^* f = f \circ \phi$ and thereby a map ϕ_* of the derivations. If X is a derivation of $\mathscr{C}(V)$ then so is $\phi_* X = \phi^{*-1} \circ X \circ \phi^*$. The noncommutative equivalent of ϕ is an automorphism of the algebra $\mathscr{A}$. Consider the inner automorphism which acts on the generators by the transformation

$$q^\lambda \mapsto \mathrm{ad}\ U^{-1} q^\lambda, \quad q^{\mu\nu} \mapsto \mathrm{ad}\ U^{-1} q^{\mu\nu}. \tag{8.1}$$

Then it is obvious that $\lambda_\mu \to \mathrm{ad} U^{-1} \lambda_\mu$ and a solution to Equation (3.39) is transformed into another solution with the same values of $P^{\mu\nu}{}_{\rho\sigma}$, $F^\lambda{}_{\mu\nu}$ and $K_{\mu\nu}$. A derivation X is transformed into $X' = \mathrm{ad}\ U^{-1} \circ X \circ \mathrm{ad}\ U$ from which we deduce that

$$\theta^\alpha \mapsto \theta'^\alpha = \mathrm{ad} U \circ \theta^\alpha \circ \mathrm{ad} U^{-1}. \tag{8.2}$$

If the geometry of V is described (locally) by a moving frame θ^α then a change of moving frame is a map,

$$\theta^{\alpha} \mapsto \theta'^{\alpha} = \Lambda^{\alpha}_{\beta} \theta^{\beta}, \tag{8.3}$$

with Λ^{α}_{β} smooth functions of V. If the metric is to be left invariant then there are restrictions on the Λ^{α}_{β}:

$$\Lambda^{\alpha}_{\gamma} \Lambda^{\beta}_{\delta} g^{\gamma\delta} = g^{\alpha\beta}.$$

In the noncommutative case one could define a change of Stehbein using the same formula (8.3) but with Λ^{α}_{β} elements of the algebra $\mathscr{A}$. If $\Lambda^{\alpha}_{\beta} \in \mathscr{Z}(\mathscr{A})$ then the new Stehbein has the same status as the old; it is dual to a set of derivations. Otherwise the change is purely formal. From the left-linearity of D^2 we conclude that

$$D^2(\Lambda^{\alpha}_{\beta} \theta^{\beta}) = \Lambda^{\alpha}_{\beta} D^2 \theta^{\beta},$$

and so from (3.26) we find that

$$\Lambda^{\alpha}_{\lambda} R'^{\lambda}{}_{\beta\gamma\delta} = R^{\alpha}{}_{\lambda\mu\nu} \Lambda^{\lambda}_{\beta} \Lambda^{\mu}_{\gamma} \Lambda^{\nu}_{\delta}. \tag{8.4}$$

If we multiply this equation on both sides by $g^{\beta\gamma}$ and define the ''Ricci tensor'' as the map $\theta^{\alpha} \mapsto R^{\alpha}{}_{\delta} \theta^{\delta}$ defined by $R^{\alpha}{}_{\delta} = g^{\beta\gamma} R^{\alpha}{}_{\beta\gamma\delta}$ then we find the condition

$$\Lambda^{\alpha}_{\lambda} R'^{\lambda}{}_{\delta} = R^{\alpha}{}_{\nu} \Lambda^{\nu}_{\delta}. \tag{8.5}$$

The order of the factors is important here. There is no reason for the trace $R^{\alpha}{}_{\alpha}$ of $R^{\alpha}{}_{\delta}$ to be ''invariant'' except if $\Lambda^{\alpha}_{\beta} \in \mathscr{Z}(\mathscr{A})$. To define the analogue of the Einstein–Hilbert action one would also have to introduce a trace on the algebra such that

$$\mathrm{Tr}(R^{\alpha}{}_{\alpha}) \to \int_V R^{\alpha}{}_{\alpha},$$

in the commutative limit. There is no obvious way in which this can be done. We are therefore unable at the moment to propose a satisfactory definition of an action and indeed we are not in a position to argue that there is even a valid action principle. A discussion of this point has been made by Connes and coworkers in a series of articles[28–31] but the definition which these authors propose is valid only on the noncommutative generalizations of compact spaces with Euclidean-signature metrics. Cyclic homology groups have been proposed[19] as the appropriate generalization to noncommutative geometry of topological invariants; the appropriate definition of other, nontopological, invariants is not clear.

ACKNOWLEDGMENTS

One of the authors (JM) would like to thank S. Doplicher and K. Fredenhagen for interesting conversations. He would also like to thank A. Dimakis for a critical reading of the manuscript. This research was partially subsidized by the CEC Science project No. SCI-CT91-0729.

[1] H. S. Snyder, ''Quantized space–time,'' Phys. Rev. **71**, 38 (1947).
[2] J. Madore and J. Mourad, ''Noncommutative Kaluza–Klein theory,'' Lecture given at the 5th Hellenic School and Workshops on Elementary Particle Physics, hep-th/9601169.
[3] J. Madore 1988, ''Non-commutative geometry and the spinning particle,'' *XI Warsaw Symposium on Elementary Particle Physics*, Kazimierz, Poland, 1988 (World Scientific, 1989); ''Kaluza–Klein aspects of noncommutative geometry,'' *Proceedings of the XVII International Conference on Differential Geometric Methods in Theoretical Physics*, Chester 1988 (World Scientific, 1989).
[4] J. Madore, *An Introduction to Noncommutative Differential Geometry and Its Physical Applications* (Cambridge University Press, Cambridge, 1995).

[5] S. Doplicher, K. Fredenhagen, and J. E. Roberts, ''Spacetime quantization induced by classical gravity,'' Phys. Lett. B **331**, 39 (1994).
[6] S. Doplicher, K. Fredenhagen, and J. E. Roberts, ''The quantum structure of space–time at the Planck scale and quantum fields,'' Commun. Math. Phys. **172**, 187 (1995).
[7] A. Kempf, G. Mangano, and R. B. Mann, ''Hilbert space representation of the minimal length uncertainty relation,'' Phys. Rev. D **52**, 1108 (1995).
[8] A. Kempf and G. Mangano, ''Minimal length uncertainty relation and ultraviolet regularization,'' Phys. Rev. D **55**, 7909 (1997).
[9] H. S. Snyder, ''The electromagnetic field in quantized space–time,'' Phys. Rev. **72**, 68 (1947).
[10] S. Deser, ''General relativity and the divergence problem in quantum field theory,'' Rev. Mod. Phys. **29**, 417 (1957).
[11] C. J. Isham, A. Salam, and J. Strathdee, ''Infinity suppression in gravity-modified quantum electrodynamics,'' Phys. Rev. D **3**, 1805 (1971).
[12] J. Madore and J. Mourad, ''On the origin of Kaluza–Klein structure,'' Phys. Lett. B **359**, 43 (1995).
[13] A. Connes, *Noncommutative geometry* (Academic, New York, 1994).
[14] M. Dubois-Violette, R. Kerner, and J. Madore, ''Gauge bosons in a noncommutative geometry,'' Phys. Lett. B **217**, 485 (1989); ''Classical bosons in a noncommutative geometry,'' Class. Quantum Grav. **6**, 1709 (1989).
[15] M. Dubois-Violette, J. Madore, T. Masson, and J. Mourad, ''On curvature in noncommutative geometry,'' J. Math. Phys. **37**, 4089 (1996).
[16] A. Dimakis and J. Madore, ''Differential calculi and linear connections,'' J. Math. Phys. **37**, 4647 (1996).
[17] J. L. Koszul, ''Lectures on fibre bundles and differential geometry,'' Tata Institute of Fundamental Research, Bombay 1960.
[18] M. Karoubi, ''Connections, courbures et classes caractéristiques en *K*-théorie algébrique,'' *Current Trends in Algebraic Topology*, Part I, London, Ontario (1981).
[19] A. Connes, ''Non-commutative differential geometry,'' Publ. Inst. Hautes Etudes Sci. **62**, 257 (1986).
[20] J. Mourad, ''Linear connections in non-commutative geometry,'' Class. Quantum Grav. **12**, 965 (1995).
[21] M. Dubois-Violette and P. Michor, ''Connections on central bimodules in noncommutative differential geometry,'' J. Geom. Phys. **20**, 218 (1996).
[22] M. Dubois-Violette and T. Masson, ''On the first-order operators in bimodules,'' Lett. Math. Phys. **37**, 467 (1996).
[23] M. Dubois-Violette, J. Madore, T. Masson, and J. Mourad, ''Linear connections on the quantum plane,'' Lett. Math. Phys. **35**, 351 (1995).
[24] M. Dubois-Violette, R. Kerner, and J. Madore, ''Noncommutative differential geometry of matrix algebras,'' J. Math. Phys. **31**, 316 (1990).
[25] M. Dubois-Violette, R. Kerner, and J. Madore, ''Shadow of noncommutativity,'' J. Math. Phys. (in press).
[26] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Yeshiva University Press, New York, 1964).
[27] T. Masson, ''Submanifolds and quotient manifolds in noncommutative geometry,'' J. Math. Phys. **37**, 2484 (1996).
[28] W. Kalau and M. Walze, ''Gravity, non-commutative geometry and the Wodzicki residue,'' J. Geom. Phys. **16**, 327 (1995).
[29] T. Ackermann and J. Tolksdorf, ''A generalized Lichnerowicz formula, the Wodzicki residue, and gravity,'' J. Geom. Phys. **19**, 143 (1996).
[30] A. H. Chamseddine and A. Connes, ''Universal formula for noncommutative geometry actions: Unification of gravity and the standard model,'' Phys. Rev. Lett. **77**, 4868 (1996).
[31] J. Fröhlich, O. Grandjean, and A. Recknagel, ''Supersymmetric quantum theory, non-commutative geometry and gravitation,'' Lecture Notes, Les Houches, Summer School in Theoretical Physics, 1995, hep-th/9706132.

Conditions for regular space-times with compact spatial sections

G. Oliveira-Neto[a)]
Department of Physics, The University, Newcastle-Upon-Tyne, NE1 7RU, United Kingdom and Departamento de Campos e Particulas, Centro Brasileiro de Pesquisas Fisicas, R. Dr. Xavier Sigaud 150, Urca, CEP 22290-180, Rio de Janeiro, Brazil

(Received 15 July 1997; accepted for publication 18 August 1997)

We apply the holonomy method to study the regularity of space-times whose initially open spatial sections are made compact. We restrict our attention to a set of different expressions of the anti-de Sitter space-time in $(2+1)$-dimensions. After we have computed the holonomy matrix to this set, we explicitly derive the regularity conditions for two elements of the set. The first expression, relevant to classical cosmology, will be regular if the periods of the compactified directions are 2π. The other expression, relevant to quantum cosmology, will be regular if the cosmological constant has a certain, well defined, discrete spectrum. © *1998 American Institute of Physics.* [S0022-2488(98)01901-X]

I. INTRODUCTION

The need for different types of space-times with closed spatial sections, in classical[1] and quantum[2] models of the Universe, has produced over the last years an increasing number of works where one turns to compact, or compactify, open manifolds.[3,4]

An interesting motivation, at the classical level, for the use of space-times whose initially open spatial sections were compactified, are the works by Fagundes.[3] There, he proposed an explanation to the so-called quasar redshift controversy (see Ref. 5 for a review) by means of multiple images of a single source produced in a universe with compact hyperbolic spatial sections.

Another motivation to consider space-times with compact spatial sections comes from quantum cosmology.[2] There one works with these space-times because one does not know, in the general case, how to treat the contributions coming from boundary terms at the spatial infinity.[2] However, we may mention a few attempts to write the quantum mechanics of gravitational systems with open spatial sections.[6]

The work of Louko and Ruback is an example, in quantum cosmology, where one has to compactify the initially open, flat, spatial sections of a set of space-times in order to compute the ''no-boundary'' wave-function[7] for certain universes.

An important property of the compactification process is that it may cause the development of singularities of conical, or even more complicate nature, in the resulting space-times.[4,8] And sometimes, as in the case of the space-times contributing to the ''no-boundary'' wave-function,[7] one is not interested in the presence of singularities. It means that it would be important to determine what the conditions are for one obtaining a regular space-time with compactified spatial sections.

Recently, a technique aimed at identifying, in a systematic way, the presence of conical singularities in a given space-time has been introduced. It is the so-called holonomy method.[9]

In the present paper, we would like to apply the holonomy method to investigate the regularity of space-times with spatial sections which are closed as the result of compactifications. The use of the holonomy method implies that we will treat here only the cases where after the compactifications the space-times may have, at most, conical singularities. In particular, we would like to

[a)]Present address: Departamento de Fisica Teorica, UERJ, Rua Sao Francisco Zavier 524, Maracana, CEP 20550-013, Rio de Janeiro, Brazil; electronic mail: no4c7@symbcomp.verj.br

determine if there are conditions on the physical constants describing these space-times, when we require them to be regular, and what these conditions are.

We may divide the space-times to be studied in two sets: the first with negatively curved spatial sections and the second with flat spatial sections. Here we shall consider only space-times of the first set, leaving the others, belonging to the second set, to a future work.

As a matter of simplicity we shall restrict our attention to space-times whose spatial sections have a constant negative curvature. In order to facilitate the calculations, we shall work in $(2+1)$-dimensions and comment, in Sec. V, on the modifications needed in order to treat $(3+1)$-dimensional space-times.

The next section, Sec. II, is devoted to the introduction of the particular space-time studied here. It is the $(2+1)$-dimensional anti-de Sitter space-time with compact spatial sections. We shall modify the usual anti-de Sitter metric by the introduction of an arbitrary and constant lapse function,[10] N, which will allow us to treat a certain set of cases.

In Sec. III, after demonstrating that the singularities that may be formed by the compactifications are, at most, conical, we apply the holonomy method to the anti-de Sitter space-time, written with an unspecified lapse function. Our final result in this section is a holonomy matrix whose elements depend on N, among other quantities.

In Sec. IV, we shall derive the regularity conditions for two elements of the set mentioned above.

The first case is treated in Sec. IV A. It is the usual expression for the anti-de Sitter space-time metric, with N equal to the identity, which is relevant to classical cosmology.[3] The regularity condition, in this case, is a constraint upon the periods of the parameters introduced to describe the compactified directions.

In Sec. IV B, we treat the second case where N is a given function of few constants present in the chosen space-time. As we shall see, in the Appendix, this space-time contributes at the semi-classical level to the ''no-boundary'' wave-function of a certain universe. It means that this case is relevant to quantum cosmology. Here, the regularity condition turns out to produce a transcendental equation for the only free physical parameter of the model, the cosmological constant Λ. For certain values of the integration constants, this equation has solutions which form a discrete and finite set.

Finally, in Sec. V, we conclude by outlining the main results of the paper and commenting on generalizations of the present analysis to other $(2+1)$ and $(3+1)$-dimensional anti-de Sitter space-times.

II. (2+1)-DIMENSIONAL ANTI-DE SITTER SPACE-TIME WITH COMPACT SPATIAL SECTIONS

In $(2+1)$-dimensions the anti-de Sitter space-time has the following metric:[11]

$$ds^2 = -N^2\, dt^2 + a^2(t)(d\chi^2 + \sinh^2\chi\, d\theta^2), \tag{1}$$

where

$$0 \leqslant t \leqslant 1, \quad 0 \leqslant \chi < \infty, \quad 0 \leqslant \theta \leqslant 2\pi, \tag{2}$$

$$a(t) = \pm\alpha \sin\left(\frac{Nt}{\alpha}\right), \qquad \text{and} \quad \alpha^2 = \frac{1}{|\Lambda|}. \tag{3}$$

in Eq. (3) Λ is the negative cosmological constant.

The presence of a constant and arbitrary lapse function in the metric (1) allows us to treat a set of different expressions of the anti-de Sitter space-time, each differing from the others by the value of N. Let us call this set ADS. The physical significance of the possible choices of N can be appreciated in Ref. 10.

From the literature,[12] we know that for fixed values of t the spatial sections of the anti-de Sitter space-time (1) are two-dimensional one-sheeted hyperboloids. The separation between the origin of the coordinate system and the pole of the one-sheeted hyperboloid is given by the scale factor, $a(t)$, at that instant. It means that the spatial sections are open. Therefore, since we are interested in the case where they are closed, we must compactify them.

In what follows, with the aid of a few mathematical results we shall introduce the relevant compact spatial sections.

The one-sheeted hyperboloid mentioned above is a representation of the two-dimensional hyperbolic space H^2. In two dimensions, every closed orientable surface of genus of at least two allows a geometric structure modeled on the H^2 (see the review article Ref. 13 for more details on this result due to Poincaré, among others). Based upon this result and in order to keep our treatment as simple as possible, let us choose our closed orientable surface to be a double torus. Then, the one-sheeted hyperboloid, which describes the spatial section of the space-time (1), will be compactified into a double torus.

One way to proceed is working in a projection of the H^2 onto a plane, which results in a disc called the Poincaré Disc (P.D.).[14] The conformal transformations which takes one from the coordinates (χ,θ) to the polar coordinates (r,θ) on the P.D. are

$$r=\sqrt{\alpha^2-v_0^2}\,\tanh(\chi/2), \quad \theta \text{ is unchanged}, \tag{4}$$

where

$$v_0=\sqrt{\alpha^2-a^2(t)}. \tag{5}$$

This mapping from the H^2 to the P.D. induces a non-trivial line element on the disc, given by

$$ds_{PD}^2=4(\alpha^2-v_0^2)^2\frac{(dr^2+r^2d\theta^2)}{(\alpha^2-v_0^2-r^2)^2}. \tag{6}$$

The next step in the compactification process is the selection of a fundamental region (in general a polygon) which tessellates[15] the P.D., and may be transformed into a double torus. Then, we shall take this fundamental region and construct the double torus, or in other words, the quotient surface H^2/Γ, where Γ is a certain symmetry transformation of the H^2.

In two dimensions, the unique symmetry transformations of the H^2 capable of generating compact surfaces are the hyperbolic ones,[16] in other words, Lorentzian boosts. Our compact surface will be fabricated by the identification of the sides, through the relevant boost transformations, of the above selected polygon.

The simplest fundamental region which gives rise to a double torus is the regular octagon.[14] If we now identify the opposite sides of this polygon, with the aid of four different boost transformations (one for each relevant direction), we obtain a representation of the desired compact surface.

Therefore, the spatial sections of our space-time will be given by the time evolution of a double torus with line element (6).

III. THE HOLONOMY MATRIX

We start by writing down the line element of the space-time described by the evolution of the double torus, in Cartesian coordinates, with the aid of Eqs. (1), (2) and (6),

$$ds^2=-N^2dt^2+\frac{4(\alpha^2-v_0^2)^2(dx^2+dy^2)}{(\alpha^2-v_0^2-x^2-y^2)^2}, \tag{7}$$

where the ranges of x and y will depend on the octagon dimensions, the opposite sides of the octagon are identified, and t has the range given in Eq. (2).

Observing Eqs. (1) and (3), we notice that the anti-de Sitter space-time has a singularity at $t=0$. This singularity can, at most, be conical. One can see it by noting that for Robertson-Walker space-times all the scalars, formed out of contractions of the Riemann tensor, are proportional to appropriate powers of the curvature scalar (R). From direct calculations one finds that our space-time has an everywhere constant curvature scalar,

$$R=-6|\Lambda|. \tag{8}$$

In fact, it is known that this singularity is a coordinate singularity.[12]

The same conclusion does not directly apply to the space-time with line element (7) because at the event $t=0$ $(v_0^2=\alpha^2)$, due to the identifications mentioned after Eq. (7), a conical singularity may have formed.

Let us, then, apply the holonomy method in order to investigate the presence, or not, of a conical singularity in the space-time with line element (7), at the event $t=0$.

Following the instructions of the holonomy method,[9] we start by writing down the parallel transport equations for a fiducial vector v, along a closed loop (to be specified), in the space-time represented by Eq. (7).

The easiest way to derive them, from Eq. (7), is by working on the orthonormal basis, defined by the transformations

$$w^t=dt, \quad w^i=\frac{2(\alpha^2-v_0^2)}{\alpha^2-v_0^2-x^2-y^2}dx_i, \tag{9}$$

where $i=x,y$ and $x_i=x,y$.

The non-vanishing connection coefficient components in this basis are

$$\Gamma_{txx}=\Gamma_{tyy}=\frac{-2\dot{v}_0 v_0(x^2+y^2)}{(\alpha^2-v_0^2)(\alpha^2-v_0^2-x^2-y^2)}=-\Gamma_{xtx}=-\Gamma_{yty}, \tag{10}$$

$$\Gamma_{ijj}=\frac{-x_i}{\alpha^2-v_0^2}=-\Gamma_{iji}, \tag{11}$$

where $j=x,y$; $i\neq j$, and i and x_i vary as in Eq. (9).

The parallel transport equation for the fidutial vector, v, is given by[9]

$$\frac{dv^\alpha}{d\lambda}+v^\beta\Gamma^\alpha_{\beta\gamma}\Omega^\gamma_\delta\frac{dx^\delta}{d\lambda}=0, \tag{12}$$

where λ is an affine parameter describing the curve, the Greek indices vary over all the coordinates, and the matrix Ω relates the non-coordinate basis with the coordinate one.

In order to obtain the explicit expression of Eq. (12), we have to introduce the closed loop around which we shall parallel transport v.

We start by noting that for a given instant t there are four independent, closed directions, one for each pair of opposite, identified, sides of the octagon. Each one of the four independent close, directions gives rise to a set of closed loops, which we shall call S_i $(i=1,\ldots,4)$. All closed loops which may be constructed, taking in account the identifications, are made up of at least one element of the four S_i. Therefore, since the space-time event under investigation is in the t direction, the simplest closed loops we can choose are the elements of the four S_i at constant and different values of t. These loops will collapse to $t=0$ when we take the limit $t\to 0$.

Then, the results derived by the use of the elements of the four S_i will be enough to draw conclusions about the regularity of our space-time.

In order to apply the holonomy method to one of the four S_i, we must choose an element of this set. Let us say the closed loop C, formed by joining the middle points of each of the opposite sides. Now, we orient our axes so that the x axis coincides with C. In terms of our coordinates, the parametric equation of C, for a certain instant of t, is

$$t=\text{constant}, \quad y=0\,, \quad x=g(\lambda), \tag{13}$$

where λ is a periodic parameter varying in the range $[\lambda_0,\lambda_f]$,$\lambda_f\equiv\lambda_0+p_\lambda$, and p_λ is the period of λ, and $g(\lambda)$ is a function of λ, to be specified, which describes x in terms of λ and varies between the extreme values of x, $[-x_0(t),x_0(t)]$. The extreme values of x are the following time dependent functions from Eq. (4):

$$x_0(t)=\sqrt{\alpha^2-v_0^2}\,\tanh(\chi_0/2), \quad \theta=0,\pi, \tag{14}$$

where χ_0 is one fixed value of χ.

So, with the aid of the relevant Ω components, we may write the parallel transport equations (12) for C [Eq. (13)]:

$$\frac{dv^y}{d\lambda}=0, \tag{15}$$

$$\frac{dv^t}{d\lambda}+\frac{4\gamma f^2(df/d\lambda)}{(1-f^2)^2}v^x=0, \tag{16}$$

$$\frac{dv^x}{d\lambda}+\frac{4\gamma f^2(df/d\lambda)}{(1-f^2)^2}v^t=0, \tag{17}$$

where,

$$f(\lambda)\equiv\frac{g(\lambda)}{\sqrt{\alpha^2-v_0^2}} \quad \text{and} \quad \gamma\equiv\frac{\dot{v}_0 v_0}{\sqrt{\alpha^2-v_0^2}}. \tag{18}$$

Observing Eq. (14), we note that the main motivation for the introduction of the new parametrization function $f(\lambda)$, Eq. (18), is to restrict all the time dependence of our subsequent results to be concentrated in γ [Eq. (18)].

The solution of Eq. (15) is given by

$$v^y=v_0^y, \tag{19}$$

where v_0^y is the value of $v^y(\lambda)$ for $\lambda=\lambda_0$. The two remaining equations, (16) and (17), form a coupled system of first order differential equations. In order to solve it we shall have to introduce the explicit value of $f(\lambda)$.

One of the simplest choices for $f(\lambda)$, which gives periodic solutions for the system (16) and (17), is obtained by demanding that

$$\frac{4f^2(df/d\lambda)}{(1-f^2)^2}=\cot(\gamma\lambda), \tag{20}$$

which gives, after an integration, the following implicit equation for f as a function of λ:

$$\left(\frac{1-f}{1+f}\right)^{\gamma}\exp\left(\frac{2f\gamma}{1-f^2}\right)=A\,\sin(\gamma\lambda), \tag{21}$$

where A is an integration constant.

Now, introducing our parametrization choice, Eq. (20), in the system (16) and (17) and solving it, we find the general solutions

$$v^t(\lambda)=\frac{D}{2}\left[\frac{B^2+\sin^2(\gamma\lambda)}{B\sin(\gamma\lambda)}\right] \quad \text{and} \quad v^x(\lambda)=\frac{D}{2}\left[\frac{B^2-\sin^2(\gamma\lambda)}{B\sin(\gamma\lambda)}\right], \tag{22}$$

where B and D are integration constants to be determined by the initial conditions.

Observing the solutions (22), we note that (for non-vanishing C and D) they are singular whenever $\lambda=2n\pi/\gamma$, n being an integer. This property comes directly from our choice for the parametrizing function $f(\lambda)$ [Eq. (20)]. This singularity of the solutions does not prevent us from using them in the holonomy method, because the holonomy matrix is computed by comparing the fiducial vector v at the same point,[9] before and after the parallel transport of v around the closed loop. Therefore, we must only avoid choosing as the initial point of the loop one of the singular points of the solutions (22).

For the initial conditions,

$$v^j(\lambda=\lambda_0)=v_0^j, \tag{23}$$

where $j=t,x$, and the solutions (19) and (22), we obtain the following holonomy matrix M:

$$\begin{pmatrix} v^t(\lambda_f) \\ v^x(\lambda_f) \\ v^y(\lambda_f) \end{pmatrix} = \begin{pmatrix} M_{tt} & M_{tx} & 0 \\ M_{xt} & M_{xx} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_0^t \\ v_0^x \\ v_0^y \end{pmatrix}, \tag{24}$$

where

$$M_{tt}=M_{xx}=\frac{\sin^2(\gamma\lambda_0)+\sin^2(\gamma\lambda_f)}{2\sin(\gamma\lambda_0)\sin(\gamma\lambda_f)}\equiv M_+, \tag{25}$$

$$M_{tx}=M_{xt}=\frac{\sin^2(\gamma\lambda_0)-\sin^2(\gamma\lambda_f)}{2\sin(\gamma\lambda_0)\sin(\gamma\lambda_f)}\equiv M_-. \tag{26}$$

IV. REGULARITY CONDITIONS

The next step in the holonomy method is the determination of the limit of the holonomy matrix elements as the closed loops collapse to the event $t=0$. It means that we must take the limit of the non-trivial elements of M, given in Eqs. (25) and (26), as $t\to 0$.

The limits of M_+, Eq. (25), and M_-, Eq. (26), as $t\to 0$ are easy to determine because they depend on t only through γ. With the aid of Eqs. (5) and (18) we compute the limit of γ as

$$\lim_{t\to 0}\gamma=-N. \tag{27}$$

From Eq. (27), the desired limits of $M_\pm$ are

$$\lim_{t\to 0}M_\pm=\frac{\sin^2(N\lambda_0)\pm\sin^2(N\lambda_f)}{2\sin(N\lambda_0)\sin(N\lambda_f)}, \tag{28}$$

where the $+$ and $-$ signs on the right hand side of Eq. (28) are associated, respectively, with the limits of M_+ and M_-.

Since we would like to obtain the conditions to have regular space-times, following the holonomy method, we have to impose that the limits of M_+ and M_- are

$$\lim_{t\to 0} M_+ = 1 \quad \text{and} \quad \lim_{t\to 0} M_- = 0. \tag{29}$$

Before we impose that the limits of M_+ and M_- have the value given by Eq. (29), we must rewrite them in terms of the other free parameter (besides N) of the model. It is the period p_λ of the parameter λ which describes the closed loop C, Eq. (13).

Using the expression which relates λ_f with λ_0 and p_λ, shown just after Eq. (13), we may introduce p_λ in the limits of M_+ and M_-, Eq. (28). If, after doing that, we impose that the resulting expressions have the values given in Eq. (29), we obtain the following independent equations for generic values of λ_0:

$$\sin^2(Np_\lambda) = 0, \tag{30}$$

$$\sin(Np_\lambda)\ \cos(Np_\lambda) = 0, \tag{31}$$

$$\sin(Np_\lambda)[\cos(Np_\lambda) - 1] = 0, \tag{32}$$

$$\cos^2(Np_\lambda) - 2\cos(Np_\lambda) + 1 = 0. \tag{33}$$

The most general solution to this system of Equations (30)–(33), in the variables N and p_λ, is

$$Np_\lambda = 2\pi n, \tag{34}$$

where n is a non-zero, positive integer.

Let us now consider what conditions Eq. (34) will introduce for two elements of the ADS [Eq. (7)]. They will have definite values of the lapse function N.

A. The case relevant to classical cosmology

For this case, as mentioned in the Introduction, the lapse function is equal to the identity

$$N = 1. \tag{35}$$

It means that Eq. (34) is reduced to the following condition upon p_λ:

$$p_\lambda = 2\pi n. \tag{36}$$

From the discussion in Ref. 9, we know that the closed, or angular, direction must be defined in the range $[0,2\pi]$. Therefore, we can only accept the solution (36) for Eq. (34) such that

$$n = 1. \tag{37}$$

B. The case relevant to quantum cosmology

Here, as demonstrated in the Appendix, the lapse function has the following expression:

$$N = \alpha[\pi/2 \pm \arccos(a_1/\alpha)], \tag{38}$$

where a_1 is positive and is the value of the scale factor, $a(t)$, for $t = 1$.

In the present case we do not need to restrict n to be given by Eq. (37), because from Eq. (34) with N written above Eq. (38), we may still have the closed direction in the desired range for a generic value of n.

If we introduce N [Eq. (38)], in Eq. (34), we find the following equation for the variables p_λ and x:

$$a_1 x = \sin\left(\frac{2\pi n}{p_\lambda} x\right), \tag{39}$$

where x is related to the cosmological constant by

$$x = \sqrt{|\Lambda|} \tag{40}$$

and varies in the range $(0,1/a_1)$.

For given values of a_1 and n, Eq. (39) furnishes a relationship between p_λ and the cosmological constant. It is clear, then, that for a proper choice of p_λ one may derive from eq. (39) the allowed values of Λ such that the space-time is regular.

Suppose we adopt the same value of p_λ as in Sec. IV A, and for simplicity choose n to be given by Eq. (37). With these choices Eq. (39) reduces to

$$a_1 x = \sin x. \tag{41}$$

Equation (39) has solutions for $a_1 < 2/\pi$, and they form a discrete set. The number of solutions will depend on a_1 due to its presence in the superior limit of x.

We may write an approximate rule to the solutions of Eq. (39), if we separate them in two sets. The first set has the even order solutions and the second the odd order solutions.

Calling m the order of the solutions, we have the following approximate rules for the two sets of solutions,

$$\text{Even order: } |\Lambda|_k \approx (1-a_1)^2 (2k-1)^2 \pi^2, \quad k = 1,2,3,\ldots, \tag{42}$$

$$\text{Odd order: } \quad |\Lambda|_l \approx (1+a_1)^2 (2l\pi)^2, \quad l = 1,2,3,\ldots, \tag{43}$$

where $k = m/2$, $l = (m-1)/2$, and we have used Eq. (40). For values of n other than 1 we shall also have equations and solutions similar to Eqs. (41)–(43), respectively.

As we have discussed in Sec. III, the conditions (36), (37), (42) and (43) guarantee the regularity of the two expressions of the anti-de Sitter space-time, as investigated by using closed loops of one of the four S_i. Their regularity can only be assured if we repeat the above procedure to the other three S_i.

We start by choosing as the representative of each of the other three S_i, the closed loops which join the middle points of each remaining pairs of opposite identified sides, as we did for the S_i studied in Sec. III. In the next step, we simply rotate our (x,y) axes so that the x axis successively coincides with these three closed loops just introduced. Finally, we separately repeat the analysis from Secs. III and IV for each of the three cases.

The choice of the regular octagon as our fundamental region means that all four independent closed loops have the same length at any time. That property implies that after repeating the analysis above, we shall obtain the same regularity conditions (36), (37), (42) and (43), for each of the three independent loops above.

Therefore, we may conclude that those conditions guarantee the regularity of the two examples of different expressions of the anti-de Sitter space-time studied in this paper.

V. CONCLUSIONS

In the present paper we have applied the holonomy method to investigate the regularity of space-times whose spatial sections are closed as the result of compactifications. In particular we were interested in the conditions to obtain regular space-times.

As a matter of definiteness and simplicity we have studied a $(2+1)$-dimensional set of different expressions of the anti-de Sitter space-times (ADS). Each element of the set differs from the others by the value of a constant lapse function.

The holonomy matrix of that set was constructed in Sec. III. The regularity conditions for two particular elements of the ADS were derived in Sec. IV.

The first element of the ADS, relevant to classical cosmology,[3] will be regular only if the periods of the compactified directions are equal to 2π (Sec. IV A).

The second element of the ADS, relevant to quantum cosmology as shown in the Appendix, was studied in Sec. IV B. Here the space-time will be regular if, for given values of the initial conditions and two other free parameters, the negative cosmological constant present in the model has the discrete values of Eqs. (42) and (43).

It is important to add that the discrete spectrum for the cosmological constant has been derived just by demanding that the space-time be regular. This requirement, following Hartle and Hawking,[7] can be considered a first principle.

Our study, in the present paper, may be generalized to other cases. The most obvious generalization is to consider elements of the ADS (7), with values for N different from the two studied here.

One may also apply the above procedure to investigate the regularity of $(2+1)$-dimensional anti-de Sitter space-times whose closed spatial sections are negative curved surfaces of genus greater than two. It means that the fundamental regions are not octagons anymore. If the new polygons are regular, the analysis will be a repetition of what we did here for a greater number of independent closed loops. If, on the other hand, the polygons are not regular, we shall have to consider the regularity conditions derived for each independent closed loop, separately, and demand that they all agree.

Finally, the use of the holonomy method to examine the regularity of $(3+1)$-dimensional anti-de Sitter space-times with closed, negative curved, spatial sections follows directly from our work. The main differences are the possible uses of different projections of the H^3, and the presence of polyhedra, or complexes, as the fundamental regions.[8] It is not difficult to show that although we have the differences, mentioned above, between the $(2+1)$- and the $(3+1)$-dimensional cases, the line element (7) changes only by the straightforward introduction of an extra dimension, and the modification of the value of α.[17] Therefore, everything we did here may be generalized to this new situation.

ACKNOWLEDGMENTS

I am grateful to I. G. Moss for suggestive discussions in the course of this work. I would also like to thank J. Barrett, S. Nag, P. Parker and I. Raptis for helpful discussions or correspondences, and CAPES and FAPERJ of Brazil for the invaluable financial support.

APPENDIX: ELEMENT OF THE ADS RELEVANT TO QUANTUM COSMOLOGY

In this Appendix, we would like to show that the anti-de Sitter space-time with line element (7), lapse function (38), compact spatial sections introduced in Sec. II, and regular as assured by the conditions (42) and (43), contributes at the semi-classical level to the ''no-boundary'' wavefunction of a certain universe.

The ''no-boundary'' conditions propose that only regular Euclidean space-times, with one single spacelike boundary, should contribute to the path integral leading to the desired wave-function. It means that in the semi-classical approximation the wave-function is given by the following expression:[18]

$$\Psi_{nb}[h_{ij}^{b},\varphi_{b}]=N_{0}\sum_{k}A_{k}\exp[-I_{k}], \tag{A1}$$

where N_0 is a normalization constant, I_k is the action of the k-th solution to the Euclidean-Einstein's field equations, and the sum is over these solutions. These solutions have a unique boundary which is compact, spacelike, and has the given three-metric h_{ij}^{b} and matter or gauge fields configuration φ_b on this boundary. The prefactors A_k are given by determinants of small fluctuations about the classical solutions.

We want to compute the ''no-boundary'' wave-function of a $(2+1)$-dimensional, homogeneous, isotropic, and constant negatively curved universe. For simplicity the only source of stress-energy is a negative cosmological constant. We shall restrict our attention to the semi-classical limit.

We start by proposing the Euclidean Friedmann–Robertson–Walker (F–R–W) metric ansatz, with the compact spatial sections introduced in Sec. II . In terms of the ADM formalism,[10] it is

$$ds^{2}=N^{2}(t)dt^{2}+a^{2}(t)(d\chi^{2}+\sinh^{2}\chi d\theta^{2}), \tag{A2}$$

where $N(t)$ is the lapse function, $a(t)$ is the scale factor, t and θ vary as in Eq. (2), Sec. II, and χ has a finite superior limit due to the compactifications.

Introducing Eq. (A2) in the general expression for the Euclidean action[19], we obtain the following result,

$$I[N,a(t)]=-\mathscr{A}\int[\dot{a}^{2}(t)-N^{2}-a^{2}(t)N^{2}\Lambda]\frac{1}{N}dt, \tag{A3}$$

where $\mathscr{A}$ is a finite defined number proportional to the volume of the compact, spatial sections, and we chose the gauge where N is a constant.

The variation of I, Eq. (A3), in terms of $a(t)$ and N, will produce the Euclidean-Einstein's equations. We shall solve these equations under the ''no-boundary'' conditions, suitably adapted for minisuperspaces.[7] For our case these conditions are the following.

(1) There must be an instant when the scale factor vanishes.
(2) The solution must have the scale factor specified at the boundary.
(3) The space-time must be regular.

We obtain two solutions satisfying the three conditions above.

The space-time which is important to the present study is the Lorentzian solution,

$$N=iN_{I},\ \text{where}\ \ N_{I}=\alpha\frac{\pi}{2}\pm\arccos\left(\frac{a_{1}}{\alpha}\right); \tag{A4}$$

$$a(t)=\pm\alpha\sin\left(N_{I}\frac{t}{\alpha}\right);\quad \alpha^{2}=\frac{1}{|\Lambda|}; \tag{A5}$$

valid for $a_1<\alpha$.

The solution Equations (A4) and (A5) satisfy all the three conditions above. More specifically it satisfies condition

(1) because $a(t)$ vanishes for $t=0$;

(2) because $a(t)$ is equal to the given value a_1 at the boundary, which is characterized by $t=1$;
(3) when the conditions (42) and (43), Sec. IV B, are valid.

Therefore, we conclude that the anti-de Sitter space-time with the line element (A2), lapse function (A4), scale factor (A5), compact spatial sections given in Sec. II, and for which conditions (42) and (43), Sec. IV B, are valid, contributes to the ''no-boundary'' wave-function (A1) of this universe.

In order to complete our proof, we must identify the above solution with the anti-de Sitter space-time introduced in Sec. IV B.

We start by renaming N_I in Eqs. (A4) and (A5), calling it N. It leads us to Eq. (38), Sec. IV B. This transformation also makes the scale factor Eq. (A5) identical to the one in Eq. (3), Sec. II. Then, we apply the coordinate transformations (4), Sec. II, to the line element (A2). Finally, we write the resulting line element in Cartesian coordinates which produces Eq. (7), Sec. III, concluding our demonstration.

[1] G. F. R. Ellis, Gen. Relativ. Gravit. **2**, 7 (1971).
[2] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967).
[3] H. V. Fagundes, Ap. J. **291**, 450 (1985); Astrophys. J. **338**, 618 (1989).
[4] J. Louko and P. J. Ruback, Class. Quantum Grav. **8**, 9 (1991).
[5] G. Burbidge, Ann. (N.Y.) Acad. Sci. **375**, 123 (1981).
[6] T. Regge and C. Teitelboim, Ann. Phys. (N.Y.) **88**, 286 (1974); K. V. Kuchař, Phys. Rev. D **50**, 3961 (1994).
[7] J. B. Hartle and S. W. Hawking, Phys. Rev. D **28**, 2960 (1983).
[8] W. P. Thurston, *The Geometry and Topology of Three-Manifolds* (Princeton U. P., Princeton, NJ, 1982), pp. 3.6–3.9.
[9] G. Oliveira-Neto, J. Math. Phys. **37**, 4716 (1996).
[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973).
[11] S. Deser and R. Jackiw, Ann. Phys. (N.Y.) **153**, 405 (1984).
[12] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time* (Cambridge U.P., Cambridge, 1973), pp. 124–134.
[13] P. Scott, Bull. London Math. Soc. **15**, 401 (1983).
[14] N. L. Balazs and A. Voros, Phys. Rep. **143**, 109 (1986).
[15] C. Series, in *Dynamical Chaos*, Proceedings of a Royal Society Discussion Meeting, edited by M. V. Berry, I. C. Percival, and N. O. Weiss (Princeton U.P., Princeton, 1987), p. 171.
[16] S. Nag, *The Complex Analytic Theory of Teichmüller Spaces* (Wiley, New York, 1988); R. Benedetti and C. Petronio, *Lectures on Hyperbolic Geometry* (Universitext) (Springer-Verlag, New York, 1992), p. 65.
[17] G. Oliveira-Neto, Pre-print University of Newcastle-Upon-Tyne—U. K. NCL-93/TP-5 (1993).
[18] S. W. Hawking, in *Relativity, Groups and Topology II*, Les Houches 1983, Session XL, edited by B. S. DeWitt and R. Stora (North Holland, Amsterdam, 1984), p. 333.
[19] S. W. Hawking, in *General Relativity*, An Einstein Centenary Survey, edited by S. W. Hawking and W. Israel (Cambridge U.P., Cambridge, 1979), p. 746.

The Dirac particle in the presence of the tensor wave on the background of gravitation

G. V. Shishkin[a] and W. H. Abdel Salam
Department of Theoretical Physics, Byelorussian State University, Minsk 220080, Republic of Belarus

(Received 31 December 1996; accepted for publication 10 April 1997)

In this paper new exact solutions for the Dirac equation are presented for the case of a particle in tensor wave field on the background of gravitation. Consideration is given to different configurations introducing the tensor interaction, which result in the required separation of variables by the algebraic method and lead to the derivation of exact solutions for the Dirac equation in the corresponding field situation. At the same time, most general metrics allowing for the exact solutions are given. The cases of zero and nonzero rest mass of a particle are with treated here. © *1998 American Institute of Physics.* [S0022-2488(97)03507-X]

I. INTRODUCTION

The present paper is a direct continuation of our previous one,[1] where exact solutions of the Dirac equation in the presence of scalar, vector, or pseudoscalar wave fields on the background of gravitation have been derived, and their associated matrices have been demonstrated. An analogous problem for the case of a tensor plane wave on the background of gravitation is presented here.

Investigation of the classical Volkov problem[2,3] (an electron in the field of a plane electromagnetic wave), which was the first solution for the nonstationary Dirac equation, was continued by several authors who have contributed to the further development of this problem: Bagrov *et al.* have studied the problem on the background of static electromagnetic fields.[4–10] These authors listed all possible configurations (geometry) of statical electric and magnetic fields allowing for exact solutions of the Volkov problem, and presented these solutions. Fedorov and his colleagues[11–20] have studied an analogous problem for the particle with different spins (0,1/2,1, 3/2,2), i.e., in their works consideration has been given to wave equations by Klein–Gordon, Dirac, Duffin–Kemmer, and others in the presence of a plane electromagnetic wave. The general approach to finding appropriate exact solutions turned out to be possible due to the method of projective operators proposed by Fedorov[11,21] and the universal notation for any wave equation in the general form put forward by Fedorov.[22,23] It should be particularly emphasized that in some solutions proposed by Fedorov *et al.* allowances are made for the anomalous particle moment, i.e., for the nonminimum interaction of particle with the electric field tensor.[13,15,20] The works of Ritus *et al.*[24,25] were the first to prove the fullness and orthonormality of the Volkov solution set.

In the works of Shishkin and co-authors[26–29] the Volkov problem is generalized for the case of wave fields having different tensor structure, i.e., exact solutions for the Dirac equations are found in the presence of scalar, vector, tensor, axial, pseudoscalar waves or waves of mixed types. In these works, the problem is considered for the plane Minkovsky space–time. Derivation of the appropriate exact solutions from the use of first- or second-order ordinary differential equations with and without separation of the variables is demonstrated in Refs. 28 and 29. In Ref. 29 it is pointed out that exact solutions of the Dirac equations are possible when the equation contains up to eight field functions (waves). This work also gives the first example of exact solutions for the generalized Volkov problem on the background of gravitation. The separation of variables, wher-

[a]Author to whom all correspondence should be addressed; Minsk 220035, Drozda 8-43, Republic of Belarus.

0022-2488/98/39(1)/454/13/$15.00
© 1998 American Institute of Physics

ever required, is performed by the algebraic method proposed in Refs. 30–32 and briefly outlined in our previous work.[1]

Thus in Ref. 1 we initiated a systematic study of the Volkov problem on the background of gravitation. The present work is a continuation of the study and is devoted to the possibility of obtaining exact solutions for the Dirac equation in the presence of a tensor wave on the background of gravitation.

II. EXACT SOLUTIONS

The physical meaning associated with tensor interaction is in direct relationship to the geometry in which such an interaction is realized. For example, the external field that is introduced by the term proportional to $\gamma^x\gamma^y\mathbf{T_{xy}}$ may be associated with the nonminimum interaction of the anomalous magnetic moment with the magnetic intensity of the electromagnetic field, i.e., with the electromagnetic field tensor. Similarly, the tensor structure proportional to $\gamma^x\gamma^t\mathbf{T_{xt}}$ is associated with nonminimum interaction of the particle electric dipole with the electric intensity of the electromagnetic field. Despite the fact the experimental data[33] point to the infinitesimal value of the electric dipole moment of elementary particles, the solutions associated with $\gamma^x\gamma^t\mathbf{T_{xt}}$ are of undoubted interest to mathematicians, if only for their substantial distinction from the corresponding solutions related to $\gamma^x\gamma^y\mathbf{T_{xy}}$. Specifically, for the zero-mass particle different additional exact solutions of the Dirac equation are available. Thus there is particular interest in the following variants when they introduce the tensor interaction and differ in geometry and physical meaning:

$$\gamma^x\gamma^y\mathbf{T_{xy}}(z-t), \quad \gamma^x\gamma^z\mathbf{T_{xz}}(z-t), \quad \gamma^x\gamma^t\mathbf{T_{xt}}(z-t), \quad \gamma^z\gamma^t\mathbf{T_{zt}}(z-t).$$

Here, for the sake of definiteness, it is assumed that the tensor wave is propagating along the z axis. The remaining variants of the tensor coupling realized for the Dirac particle are reduced to those enumerated above. It should be noted that the general introduction of a tensor field presupposes the presence of all the components of a $\mathbf{T_{ik}}$ tensor. In so doing, each specific physical realization is determined by the properties of the interacting particle itself. To illustrate, in the case of an electromagnetic field the particle may or may not have the nonzero electric charge and may or may not have the magnetic or electric moment. This will enable one to determine the type of interaction, namely, minimum, nonminimum, or mixed type.

Analysis of the Dirac equation for the separation of variables reveals that separation of variables is impossible in the cases $\gamma^x\gamma^z\mathbf{T_{xz}}$ and $\gamma^x\gamma^t\mathbf{T_{xt}}$, since the resultant systems of matrix separating equations have no solutions. It seems impossible to give some recommendations concerning the exact solutions in that even in the simplest situations with a particular relationship between field functions and the space and time variables one has to use the Riccati-type or still more complex equations. However, this does not rule out computerized numerical analysis. As regards to other variants being introducing, namely; $\gamma^x\gamma^y\mathbf{T_{xy}}$ and $\gamma^z\gamma^t\mathbf{T_{zt}}$, we have succeeded in specifying the matrices allowing for the separation of variables and exact solutions, and have managed to find these solutions.

A. The specific case of $\gamma^x\gamma^y\mathbf{T_{xy}}(z-t)$

In this case, the Dirac equation for the diagonal tetrad gauge takes the form:

$$\left\{\frac{\gamma^x}{a_x^{1/2}}\left(\frac{\partial}{\partial x}-\frac{\partial_x a_x}{4a_x}\right)+\frac{\gamma^y}{a_y^{1/2}}\left(\frac{\partial}{\partial y}-\frac{\partial_y a_y}{4a_y}\right)+\frac{\gamma^z}{a_z^{1/2}}\left(\frac{\partial}{\partial z}-\frac{\partial_z a_z}{4a_z}\right)+\frac{\gamma^t}{a_t^{1/2}}\left(\frac{\partial}{\partial t}-\frac{\partial_t a_t}{4a_t}\right)+m_0\right.$$
$$\left.+2ig\gamma^x\gamma^y\mathbf{T_{xy}}(z-t)\right\}\Psi=\{H\}\Psi=0. \tag{1}$$

Here $\mathbf{T_{xy}(z-t)}$ is the tensor wave, g is a magnetic moment of the particle (including the anomalous one), and $\mathbf{a_k(x,y,z,t)}$, $\mathbf{(k=x,y,z,t)}$ are gravitational field functions determining the space–time metric:

$$ds^2=a_x(dx)^2+a_y(dy)^2+a_z(dz)^2-a_t(dt)^2, \tag{2}$$

$$\{H\}\Psi=0\Rightarrow F\{H\}\Gamma\Gamma^{-1}\Psi=0\Rightarrow(\hat{K}_{xy}+\hat{K}_{zl})\Phi, \quad \Psi=\Gamma\Phi. \tag{3}$$

Hereafter the notation will be identical to that adopted in Ref. 1.

In compliance with Refs. 30–32, separation of variables is carried out as follows:

$$\hat{K}_{xy}\hat{K}_{zt}-\hat{K}_{zy}\hat{K}_{xy}=0, \tag{4}$$

$$a_k(x,y,z,t)=v_k(x,y)w(z,t), \tag{5}$$

where $\mathbf{F}$ is a scalar factor and $\mathbf{\Gamma}$ is a matrix 4×4, which provides a mean to satisfy commutation relation (4).

Directing our attention to the explicit form of Eq. (1), we write the operators $\hat{K}_{xy}$ and $\hat{K}_{zt}$ in the most general form,

$$\hat{\mathbf{K}}_{xy}=F\left\{\frac{\gamma^x}{\mathbf{a}_x^{1/2}}\left(\frac{\partial}{\partial\mathbf{x}}-\frac{\partial_x\mathbf{a}_x}{\mathbf{4a}_x}\right)+\frac{\gamma^y}{\mathbf{a}_y^{1/2}}\left(\frac{\partial}{\partial\mathbf{y}}-\frac{\partial_y\mathbf{a}_y}{\mathbf{4a}_y}\right)+\alpha\mathbf{m}_0\right\}\Gamma, \tag{6}$$

$$\hat{K}_{zt}=F\left\{\frac{\gamma^z}{a_z^{1/2}}\left(\frac{\partial}{\partial z}-\frac{\partial_z a_z}{4a_z}\right)+\frac{\gamma^t}{a_t^{1/2}}\left(\frac{\partial}{\partial t}-\frac{\partial_t a_t}{4a_t}\right)+(1-\alpha)m_0+2ig\gamma^x\gamma^yT_{xy}(z-t)\right\}\Gamma, \tag{7}$$

where $\boldsymbol{\alpha}$ is a number equal to 1 or 0.

The commutation requirement (4) as applied to operators (6) and (7) leads to the following system of reparating equations:

$$\gamma^x\Gamma\gamma^z-\gamma^z\Gamma\gamma^x=0, \quad \gamma^y\Gamma\gamma^z-\gamma^z\Gamma\gamma^y=0,$$

$$\gamma^x\Gamma\gamma^t-\gamma^t\Gamma\gamma^x=0, \quad \gamma^y\Gamma\gamma^t-\gamma^t\Gamma\gamma^y=0,$$

$$\gamma^x\Gamma\gamma^x\gamma^y-\gamma^x\gamma^y\Gamma\gamma^x=0, \quad \gamma^y\Gamma\gamma^x\gamma^y-\gamma^x\gamma^y\Gamma\gamma^y=0, \tag{7a}$$

$$\frac{\partial}{\partial x}\frac{F}{a_z^{1/2}}=0, \quad \frac{\partial}{\partial z}\frac{F}{a_x^{1/2}}=0, \quad \frac{\partial}{\partial x}\frac{F}{a_t^{1/2}}=0, \quad \frac{\partial}{\partial t}\frac{F}{a_x^{1/2}}=0,$$

$$\frac{\partial}{\partial y}\frac{F}{a_z^{1/2}}=0, \quad \frac{\partial}{\partial z}\frac{F}{a_y^{1/2}}=0, \quad \frac{\partial}{\partial y}\frac{F}{a_t^{1/2}}=0, \quad \frac{\partial}{\partial t}\frac{F}{a_y^{1/2}}=0,$$

$$\frac{\partial}{\partial\mathbf{x}}\frac{F}{\mathbf{a}_z^{1/2}}\frac{\partial_z\mathbf{a}_z}{\mathbf{4a}_z}=\mathbf{0}, \quad \frac{\partial}{\partial\mathbf{z}}\frac{F}{\mathbf{a}_x^{1/2}}\frac{\partial_x\mathbf{a}_x}{\mathbf{4a}_x}=\mathbf{0},$$

$$\frac{\partial}{\partial\mathbf{x}}\frac{F}{\mathbf{a}_t^{1/2}}\frac{\partial_t\mathbf{a}_t}{\mathbf{4a}_t}=\mathbf{0}, \quad \frac{\partial}{\partial\mathbf{t}}\frac{F}{\mathbf{a}_x^{1/2}}\frac{\partial_x\mathbf{a}_x}{\mathbf{4a}_x}=\mathbf{0}, \tag{7b}$$

$$\frac{\partial}{\partial\mathbf{y}}\frac{F}{\mathbf{a}_z^{1/2}}\frac{\partial_z\mathbf{a}_z}{\mathbf{4a}_z}=0, \quad \frac{\partial}{\partial\mathbf{z}}\frac{F}{\mathbf{a}_y^{1/2}}\frac{\partial_x\mathbf{a}_y}{\mathbf{4a}_y}=\mathbf{0},$$

$$\frac{\partial}{\partial \mathbf{y}}\frac{F}{\mathbf{a}_t^{1/2}}\frac{\partial_t \mathbf{a}_t}{\mathbf{4a}_t}=\mathbf{0}, \quad \frac{\partial}{\partial \mathbf{t}}\frac{F}{\mathbf{a}_y^{1/2}}\frac{\partial_x \mathbf{a}_y}{\mathbf{4a}_y}=\mathbf{0},$$

$$(\gamma^x\Gamma-\Gamma\gamma^x)(\mathbf{1}-\alpha)\mathbf{m}_0=\mathbf{0}, \quad (\gamma^y\Gamma-\Gamma\gamma^y)(\mathbf{1}-\alpha)\mathbf{m}_0=\mathbf{0},$$

$$(\gamma^z\Gamma-\Gamma\gamma^z)\alpha\mathbf{m}_0=\mathbf{0}, \quad (\gamma^t\Gamma-\Gamma\gamma^t)\alpha\mathbf{m}_0=\mathbf{0},$$

$$(\gamma^x\gamma^y\Gamma-\Gamma\gamma^x\gamma^y)\alpha\mathbf{m}_0=\mathbf{0}. \tag{7c}$$

Here, Eq. (7a) forms an independent subsystem of matrix equations solvable for

$$\Gamma=\gamma^x\gamma^y. \tag{8}$$

With the use of other equations of the system (7) we get:

$$\alpha=\mathbf{1}, \quad F=1, \tag{9}$$

$$a_x=v_x(x,y), \quad \mathbf{a}_z=\mathbf{w}_z(\mathbf{z},\mathbf{t}),$$

$$a_y=v_y(x,y), \quad \mathbf{a}_t=\mathbf{w}_t(\mathbf{z},\mathbf{t}), \tag{10}$$

In this manner, for the operators $\hat{K}_{xy}$ and $\hat{K}_{zt}$ we have the following:

$$\hat{K}_{xy}=\frac{\gamma^y}{v_x^{1/2}}\left(\frac{\partial}{\partial x}-\frac{\partial_x v_x}{4v_x}\right)-\frac{\gamma^x}{v_x^{1/2}}\left(\frac{\partial}{\partial y}-\frac{\partial_y v_y}{4v_y}\right)+\gamma^x\gamma^y m_0, \tag{11}$$

$$\hat{K}_{zt}=\frac{\gamma^z\gamma^y\gamma^z}{w_z^{1/2}}\left(\frac{\partial}{\partial z}-\frac{\partial_z w_z}{4w_z}\right)+\frac{\gamma^x\gamma^y\gamma^t}{w_t^{1/2}}\left(\frac{\partial}{\partial t}-\frac{\partial_t w_t}{4w_t}\right)-2ig\mathbf{T}_{\mathbf{xy}}(z-t). \tag{12}$$

Then turning our attention to the problem in eigenfunctions and eigenvalues in accordance with

$$\hat{\mathbf{K}}_{xy}\Phi=-\hat{\mathbf{K}}_{zt}\Phi=\mathbf{k}\Phi \tag{13}$$

and introducing the light-cone variables

$$u=z-t, \quad v=z+t \tag{14}$$

for the variables **u** and **v** with the use of Eq. (12) we derive the following equation:

$$\left\{\frac{\gamma^x\gamma^y\gamma^z}{w_z^{1/2}}\left(\partial_u+\partial_v-\frac{(\partial u+\partial v)w_z}{4w_z}\right)+\frac{\gamma^x\gamma^y\gamma^t}{w_t^{1/2}}\left(-\partial_u+\partial_v-\frac{(-\partial u+\partial v)}{4w_t}\right)-2ig\mathbf{T}_{\mathbf{xy}}(u)+k\right\}\Phi(u,v)=0. \tag{15}$$

In connection with this equation the following should be noted.

(1) In order for the variables **u** and **v** to be separated (see Refs. 1 and 30–32), it is required to assume that

$$w_z=c_z(z-t)d_z(z+t), \quad w_t=c_t(z-t)d_t(z+t). \tag{16}$$

(2) In order to use the advantages of a special representation for the degenerate matrices $\gamma^z \pm \gamma^t$ (see Refs. 26–29) we assume that

$$w_z = w_t = c(z-t)d(z+t). \tag{17}$$

(3) Finally, we assume the following so that the appearance of the Riccati-type equation in the variable **v** is obviated (see Ref. 1):

$$d(z+t)=1, \quad w_z = w_t = c(z-t) = c(u), \tag{18}$$

As a result, instead of Eq. (15) we get

$$\left\{ \frac{\gamma^x \gamma^y}{c^{1/2}} (\gamma^z - \gamma^t) \frac{\partial}{\partial u} - \frac{\gamma^x \gamma^y}{c^{1/2}} (\gamma^z - \gamma^t) \frac{\partial_u c}{4c} + \frac{\gamma^x \gamma^y}{c^{1/2}} (\gamma^z + \gamma^t) \frac{\partial}{\partial v} - 2ig\mathbf{T}_{\mathbf{xy}} + k \right\} \Phi(u,v) = 0. \tag{19}$$

Further, the following representation of the Dirac matrices is used:

$$\gamma^x = \begin{pmatrix} \mathbf{0} & \sigma_3 \\ \sigma_3 & \mathbf{0} \end{pmatrix}, \quad \gamma^y = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}, \quad \gamma^z = \begin{pmatrix} \mathbf{0} & \sigma_1 \\ \sigma_1 & \mathbf{0} \end{pmatrix}, \quad \gamma^t = \mathbf{i} \begin{pmatrix} \mathbf{0} & \sigma_2 \\ \sigma_2 & \mathbf{0} \end{pmatrix}. \tag{20}$$

Here, σ_k are the corresponding Pauli matrices,

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{21}$$

For the degenerate matrices $\gamma^z \pm \gamma^+$ we have, respectively,

$$\gamma^z + \gamma^t = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \gamma^z - \gamma^t = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{22}$$

As Eq. (19) contains no field functions dependent on the variable **v**, we can assume the following:

$$\Phi \sim e^{ik_v v}. \tag{23}$$

Then Eq. (19) in the representation of Eq. (20) results in the following system of equations:

$$\begin{gathered} -\frac{2ik_v}{c^{1/2}} \Phi_2 - (2ig\mathbf{T}_{\mathbf{xy}} - k)\Phi_1 = 0, \\ \frac{2}{c^{1/2}} \left(\frac{\partial}{\partial u} - \frac{\partial_u c}{4c} \right) \Phi_1 - (2ig\mathbf{T}_{\mathbf{xy}} - k)\Phi_2 = 0, \\ \frac{2ik_v}{c^{1/2}} \Phi_4 - (2ig\mathbf{T}_{\mathbf{xy}} - k)\Phi_3 = 0, \\ -\frac{2}{c^{1/2}} \left(\frac{\partial}{\partial u} - \frac{\partial_u c}{4c} \right) \Phi_3 - (2ig\mathbf{T}_{\mathbf{xy}} - k)\Phi_4 = 0, \end{gathered} \tag{24}$$

when we have

$$\Phi_1 = \Phi_3, \quad \Phi_2 = -\Phi_4 \quad (\text{or } \Phi_1 = -\Phi_3, \quad \Phi_2 = \Phi_4). \tag{25}$$

The first pair of Eq. (24) gives the following:

$$\Phi_2 = -\frac{2ig\mathbf{T_{xy}}}{2ik_v}\, c^{1/2}\Phi_1, \tag{26}$$

and therefore in accordance with Eq. (25)

$$\left\{\frac{\partial}{\partial \mathbf{u}} - \frac{\partial_u \mathbf{c}}{\mathbf{4c}} + \frac{(\mathbf{2igT_{xy}} - \mathbf{k})}{\mathbf{4ik}_v}\, c\right\}\Phi_1 = \mathbf{0}, \tag{27}$$

$$\Phi_1 = \mathbf{c}^{1/4}\exp\left\{-\frac{\mathbf{1}}{\mathbf{4ik}_v}\int (\mathbf{2igT_{xy}} - \mathbf{k})^2 \mathbf{cdu}\right\}. \tag{28}$$

In such a manner we have established the dependence of the solution on variables **u** and **v**.

Let us now turn our attention to the operator $\hat{K}_{xy}$ [formula (11)] in order to obtain the corresponding equation:

$$\left\{\frac{\gamma^y}{v_x^{1/2}}\left(\frac{\partial}{\partial x} - \frac{\partial_x v_x}{4v_x}\right) - \frac{\gamma^x}{v_y^{1/2}}\left(\frac{\partial}{\partial y} - \frac{\partial_y v_y}{4v_y}\right) + \gamma^x\gamma^y m_0 - k\right\}\Phi(x,y) = 0, \tag{29}$$

where **x** and **y** appear symmetrically.

Separating the variables **x** and **y** according to the scheme

$$\{\mathbf{H_{x,y}}\}\Phi = \mathbf{0} \Rightarrow G\{\mathbf{H_{x,y}}\}\Lambda\Lambda^{-1}\Phi = \mathbf{0} \Rightarrow (\hat{\mathbf{K}}_\mathbf{x} + \hat{\mathbf{K}}_\mathbf{y})\Theta = \mathbf{0}, \quad \Phi = \Lambda\Theta, \tag{30}$$

$$\hat{\mathbf{K}}_\mathbf{x}\hat{\mathbf{K}}_\mathbf{y} - \hat{\mathbf{K}}_\mathbf{y}\hat{\mathbf{K}}_\mathbf{x} = \mathbf{0}, \tag{31}$$

$$v_x = \tilde{a}_x(x)\tilde{b}_x(y), \quad v_y = \tilde{a}_y(x)\tilde{b}_y(y), \tag{32}$$

and assuming

$$\hat{K}_x = \mathbf{G}\left\{\frac{\gamma^y}{\boldsymbol{\nu}_x^{1/2}}\left(\frac{\partial}{\partial x} - \frac{\partial_x \boldsymbol{\nu}_x}{\mathbf{4}\boldsymbol{\nu}_x}\right) + \gamma^x\gamma^y\mathbf{m}_0 - \alpha\mathbf{k}\right\}\Lambda, \tag{33}$$

$$\hat{K}_y = G\left\{-\frac{\gamma^x}{\nu_y^{1/2}}\left(\frac{\partial}{\partial y} - \frac{\partial_y \nu_y}{4\nu_y}\right) - (1-\alpha)k\right\}\Lambda, \tag{34}$$

we get the following system of separating equations:

$$\gamma^y\Lambda\gamma^x - \gamma^x\Lambda\gamma^y = \mathbf{0}, \quad \gamma^x\gamma^y\Lambda\gamma^x - \gamma^x\Lambda\gamma^x\gamma^y = \mathbf{0},$$

$$\partial_x\left(\frac{\mathbf{G}}{v_\mathbf{y}^{1/2}}\right) = \partial_y\left(\frac{\mathbf{G}}{v_x^{1/2}}\right) = \partial_x\left(\frac{\mathbf{G}}{v_\mathbf{y}^{1/2}}\frac{\partial_y v_y}{\mathbf{4}v_y}\right) = \partial_y\left(\frac{\mathbf{G}}{v_x^{1/2}}\frac{\partial_x v_x}{\mathbf{4}v_x}\right) = \mathbf{0}, \tag{35}$$

$$(\gamma^y\Lambda - \Lambda\gamma^y)(\mathbf{1}-\alpha)\mathbf{k} = \mathbf{0}, \quad (\gamma^x\gamma^y\Lambda - \Lambda\gamma^x\gamma^y)(\mathbf{1}-\alpha)\mathbf{k} = \mathbf{0},$$

$$(\gamma^x\Lambda - \Lambda\gamma^x)\alpha\mathbf{k} = \mathbf{0},$$

which gives

$$\Lambda = \gamma^x, \quad \alpha = \mathbf{1}, \quad \mathbf{G} = \tilde{\mathbf{a}}_y^{1/2}(\mathbf{x}), \tag{36}$$

$$v_x=\tilde{a}_x(x), \quad v_y=\tilde{a}_y(x)\tilde{b}_y(y), \tag{37}$$

$$\hat{K}_x=-\gamma^x\gamma^y\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial x}-\frac{\partial_x\tilde{a}_x}{\mathbf{4}\tilde{a}_x}\right)-\gamma^y a_y^{1/2}\mathbf{m}_0-\gamma^x\,\tilde{a}_y^{1/2}\mathbf{k}, \tag{38}$$

$$\hat{K}_y=-\frac{1}{\tilde{b}_y^{1/2}}\left(\frac{\partial}{\partial y}-\frac{\partial_y\tilde{b}_y}{4\tilde{b}_y}\right). \tag{39}$$

And going over to the problem in eigenfunctions and eigenvalues

$$\hat{K}_x\Theta=\hat{K}_y\Theta=\tilde{\mathbf{k}}\Theta. \tag{40}$$

According to Eq. (38), for the variable **x** we get

$$\left\{\gamma^x\gamma^y\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial\mathbf{x}}-\frac{\partial_x\tilde{a}_x}{\mathbf{4}\tilde{a}_x}\right)+\gamma^y a_y^{1/2}\mathbf{m}_0+\gamma^x\,\tilde{a}_y^{1/2}\mathbf{k}+\tilde{\mathbf{k}}\right\}\Theta=\mathbf{0}, \tag{41}$$

and from the representation of Eq. (20), we will get the following system;

$$\begin{aligned}
&\left\{-\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial\mathbf{x}}-\frac{\partial_x\tilde{a}_x}{\mathbf{4}\tilde{a}_x}\right)+\tilde{a}_y^{1/2}\mathbf{k}\right\}\Theta_3+(\tilde{a}_y^{1/2}\mathbf{m}_0+\tilde{\mathbf{k}})\Theta_1=\mathbf{0},\\
&\left\{\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial\mathbf{x}}-\frac{\partial_x\tilde{a}_x}{\mathbf{4}\tilde{a}_x}\right)-\tilde{a}_y^{1/2}\mathbf{k}\right\}\Theta_4+(\tilde{a}_y^{1/2}\mathbf{m}_0+\tilde{\mathbf{k}})\Theta_2=\mathbf{0},\\
&\left\{\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial x}-\frac{\partial_x\tilde{a}_x}{4\tilde{a}_x}\right)+\tilde{a}_y^{1/2}k\right\}\Theta_1-(\tilde{a}_y^{1/2}m_0-\tilde{k})\Theta_3=0,\\
&\left\{-\left(\frac{\tilde{a}_y}{\tilde{a}_x}\right)^{1/2}\left(\frac{\partial}{\partial x}-\frac{\partial_x\tilde{a}_x}{4\tilde{a}_x}\right)-\tilde{a}_y^{1/2}k\right\}\Theta_2-(\tilde{a}_y^{1/2}m_0-\tilde{k})\Theta_4=0,
\end{aligned} \tag{42}$$

when we have

$$\Theta_1=\Theta_2, \quad \Theta_3=-\Theta_4, \tag{43}$$

In this case, believing

$$\Theta_v(x)=\tilde{a}_x^{1/4}\tilde{\Theta}_v(x), \quad v=1,2,3,4, \tag{44}$$

we will get

$$\begin{aligned}
&\left(-\frac{\partial}{\partial x}+\tilde{a}_x^{1/2}k\right)\tilde{\Theta}_3+\left(\tilde{a}_x^{1/2}m_0+\left(\frac{\tilde{a}_x}{\tilde{a}_y}\right)^{1/2}\tilde{k}\right)\tilde{\Theta}_1=0,\\
&\left(\frac{\partial}{\partial x}-\tilde{a}_x^{1/2}k\right)\tilde{\Theta}_4+\left(\tilde{a}_x^{1/2}m_0+\left(\frac{\tilde{a}_x}{\tilde{a}_y}\right)^{1/2}\tilde{k}\right)\tilde{\Theta}_2=0,\\
&\left(\frac{\partial}{\partial x}+\tilde{a}_x^{1/2}k\right)\tilde{\Theta}_1-\left(\tilde{a}_x^{1/2}m_0-\left(\frac{\tilde{a}_x}{\tilde{a}_y}\right)^{1/2}\tilde{k}\right)\tilde{\Theta}_3=0,
\end{aligned} \tag{45}$$

$$\left(-\frac{\partial}{\partial x}-\widetilde{a}_x^{1/2}k\right)\widetilde{\Theta}_2-\left(\widetilde{a}_x^{1/2}m_0-\left(\frac{\widetilde{a}_x}{\widetilde{a}_y}\right)^{1/2}\widetilde{k}\right)\widetilde{\Theta}_4=0.$$

At last, in order to avoid Riccati-type equations, we have to accept

$$\widetilde{\mathbf{a}}_y=\mathbf{1}. \tag{46}$$

By introducing an integral argument

$$\xi=\int \widetilde{\mathbf{a}}_x^{1/2}\mathbf{dx}\widetilde{\Theta}_\nu(\mathbf{x})\to\widetilde{\Theta}_\nu(\xi),\quad \nu=\mathbf{1,2,3,4}, \tag{47}$$

instead of Eq. (45) we have

$$\begin{aligned}
&\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_3-(m_0+\widetilde{k})\widetilde{\Theta}_1=0,\\
&\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_4+(m_0+\widetilde{k})\widetilde{\Theta}_2=0,\\
&\left(\frac{\partial}{\partial\xi}+k\right)\widetilde{\Theta}_1-(m_0-\widetilde{k})\widetilde{\Theta}_3=0,\\
&\left(\frac{\partial}{\partial\xi}+k\right)\widetilde{\Theta}_2+(m_0-\widetilde{k})\widetilde{\Theta}_4=0,
\end{aligned} \tag{48}$$

from which

$$\widetilde{\Theta}_1=\frac{\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_3}{m_0+\widetilde{k}},\quad \widetilde{\Theta}_2=-\frac{\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_4}{m_0+\widetilde{k}}. \tag{49}$$

For $\widetilde{\Theta}_1$ and $\widetilde{\Theta}_2$ we have now the following second-order equations:

$$\left(\frac{\partial^2}{\partial\xi^2}-\mathbf{m}_0^2+\widetilde{\mathbf{k}}^2-\mathbf{k}^2\right)\widetilde{\Theta}_3=\mathbf{0},\quad \left(\frac{\partial^2}{\partial\xi^2}-\mathbf{m}_0^2+\widetilde{\mathbf{k}}^2-\mathbf{k}^2\right)\widetilde{\Theta}_2=\mathbf{0}, \tag{50}$$

and the solution may be written exactly:

$$\widetilde{\Theta}_1=\widetilde{\Theta}_2=\exp\{\pm(m_0^2+k^2-\widetilde{k}^2)^{1/2}x\}=\exp\left\{\pm(m_0^2+k^2-\widetilde{k}^2)^{1/2}\int\widetilde{a}_x^{1/2}dx\right\}. \tag{51}$$

According to the definition of the operator $\hat{K}_y$ in Eq. (39), we can get the following for the variable **y**:

$$\left\{\frac{\mathbf{1}}{\widetilde{\mathbf{b}}_y^{1/2}}\left(\frac{\partial}{\partial\mathbf{y}}-\frac{\partial_{\mathbf{y}}\mathbf{b}_{\mathbf{y}}}{\mathbf{4b_y}}\right)-\widetilde{\mathbf{k}}\right\}\Theta_\nu(\mathbf{y})=\mathbf{0},\quad \nu=\mathbf{1,2,3,4}, \tag{52}$$

from where we obtain

$$\Theta_1(\mathbf{y})=\Theta_2(\mathbf{y})=\Theta_3(\mathbf{y})=\Theta_4(\mathbf{y})=\widetilde{\mathbf{b}}_y^{1/4}\exp\left\{\int \mathbf{b}_y^{1/2}(\mathbf{y})\mathbf{dy}\right\}. \tag{53}$$

Taking into consideration exact solutions for individual variables and explicit form of the matrices $\mathbf{\Gamma}$ and Λ, we finally have the following as a solution for the Dirac equation (1):

$$\Psi=\mathbf{e}^{ik_v(z+t)}\begin{pmatrix} -\Phi_1(\mathbf{z}-\mathbf{t})\Theta_3(\mathbf{x}) \\ \Phi_2(\mathbf{z}-\mathbf{t})\Theta_3(\mathbf{x}) \\ \Phi_1(\mathbf{z}-\mathbf{t})\Theta_1(\mathbf{x}) \\ \Phi_2(\mathbf{z}-\mathbf{t})\Theta_1(\mathbf{x}) \end{pmatrix}\Theta(\mathbf{y}). \tag{54}$$

This solution is possible for the metrics of the form

$$ds^2=\widetilde{a}_x(x)(dx)^2+\widetilde{b}(y)(dy)^2+c(z-t)(dz^2-dt^2). \tag{55}$$

It should be noted that the results (54) and (55) have been obtained for the particle with a nonzero rest mass ($\mathbf{m}\neq\mathbf{0}$), but they are true also for the case $\mathbf{m}_0=\mathbf{0}$.

However, in the case of zero rest mass of particle ($\mathbf{m}_0=\mathbf{0}$) an additional more general form of the metric makes it possible to find exact solutions for Eq. (1).

Actually, instead of Eq. (5), one could obtain the following:

$$\begin{aligned} &a_x=v_x(x,y)w_x(z,t), \quad a_z=w_z(z,t), \\ &a_y=v_y(x,y)w_y(z,t), \quad a_t=w_t(z,t). \end{aligned} \tag{56}$$

Equation (56) leads to the conditions

$$\begin{aligned} &a_x=v_x(x,y)\widetilde{c}(z-t), \\ &a_y=v_y(x,y)\widetilde{c}(z-t), \\ &w_z=w_t=c(z-t). \end{aligned} \tag{57}$$

Repeating all the above calculations, we come to the solution of the form (54), where instead of Eqs. (26) and (28) we get

$$\Phi_2=-\frac{(2ig\mathbf{T}_{\mathbf{xy}}-k)}{2ik_v}\left(\frac{c}{\widetilde{c}}\right)^{1/2}\Phi_1, \tag{58}$$

$$\Phi_1=c^{1/4}\exp\left\{-\frac{1}{4ik_v}\int(2ig\mathbf{T}_{\mathbf{xy}}-k)^2\,\frac{c}{\widetilde{c}}\,du\right\} \tag{59}$$

and instead of Eqs. (49) and (50) one should assume that

$$\widetilde{\Theta}_1=\frac{1}{\widetilde{k}}\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_3, \quad \widetilde{\Theta}_2=-\frac{1}{\widetilde{k}}\left(\frac{\partial}{\partial\xi}-k\right)\widetilde{\Theta}_4, \tag{60}$$

$$\widetilde{\Theta}_1(x)=\widetilde{\Theta}_2(x)=\exp\left\{\pm(k^2-\widetilde{k}^2)^{1/2}\int \widetilde{a}_x^{1/2}dx\right\}. \tag{61}$$

Expression (53) still specifies the $\mathbf{y}$ dependence.

The most general form of the metrics allowing for the exact solutions for the case $\mathbf{m}_0=\mathbf{0}$ takes the form

$$ds^2=\widetilde{c}(z-t)\{\widetilde{a}_x(dx)^2+\widetilde{b}_y(dy)^2\}+c(z-t)\{(dz)^2-(dt)^2\}. \tag{62}$$

B. The specific case of $\gamma^z\gamma^t\mathbf{T}_{zt}(z-t)$

In this case the solution of the Dirac equation is quite similar to the previous one, except that the solution for $\mathbf{m}_0=\mathbf{0}$ here is a particular case of a more general solution for $\mathbf{m}_0\neq\mathbf{0}$ causing no differences in metrics [compare Eqs. (55) and (62)].

The Dirac equation takes the form

$$\left\{\frac{\gamma^x}{a_x^{1/2}}\left(\frac{\partial}{\partial x}-\frac{\partial_x a_x}{4a_x}\right)+\frac{\gamma^y}{a_y^{1/2}}\left(\frac{\partial}{\partial y}-\frac{\partial_y a_y}{4a_y}\right)+\frac{\gamma^z}{a_z^{1/2}}\left(\frac{\partial}{\partial z}-\frac{\partial_z a_z}{4a_z}\right)+\frac{\gamma^t}{a_t^{1/2}}\left(\frac{\partial}{\partial t}-\frac{\partial_t a_t}{4a_t}\right)\right.$$
$$\left.+m_0+2g\gamma^z\gamma^t\mathbf{T}_{zt}(z-t)\right\}\Psi=0. \tag{63}$$

Separating the variables according to Eqs. (3)–(5), we get the following:

$$\Gamma=\gamma^z\gamma^t,\quad \alpha=0,\quad F=w^{1/2}(z,t), \tag{64}$$

$$\begin{aligned} a_x&=v_x(x,y)w(z,t),\quad a_z=w_z(z,t),\\ a_y&=v_y(x,y)w(z,t),\quad a_t=w_t(z,t). \end{aligned} \tag{65}$$

In such, we get for the operators $\hat{K}_{xy}$ and $\hat{K}_{zt}$,

$$\hat{\mathbf{K}}_{xy}=\frac{\gamma^x\gamma^z\gamma^t}{v_x^{1/2}}\left(\frac{\partial}{\partial\mathbf{x}}-\frac{\partial v_x}{\mathbf{4}v_x}\right)+\frac{\gamma^y\gamma^z\gamma^t}{v_y^{1/2}}\left(\frac{\partial}{\partial\mathbf{y}}-\frac{\partial_y v_y}{\mathbf{4}v_y}\right), \tag{66}$$

$$\hat{K}_{zt}=\gamma^t\left(\frac{w}{w_z}\right)^{1/2}\left(\frac{\partial}{\partial z}-\frac{\partial_z w_z}{4w_z}\right)+\gamma^z\left(\frac{w}{w_t}\right)^{1/2}\left(\frac{\partial}{\partial t}-\frac{\partial_t w_t}{4w_t}\right)+\gamma^z\gamma^t w^{1/2}m_0+2gw^{1/2}T_{zt}(z-t). \tag{67}$$

Simplifications connected with the remarks in Sec. II A lead to

$$w=w(z-t),\quad w_z=w_t=c(z-t). \tag{68}$$

With the use of the representation (20) for the Dirac matrices we derive the following:

$$\Phi_1(z-t)=\Phi_3(z-t),\quad \Phi_2(z-t)=\Phi_4(z-t), \tag{69}$$

$$\Phi_4=\frac{w^{1/2}(m_0-2gT_{zt})-k}{2ik_\nu}\left(\frac{c}{w}\right)^{1/2}\Phi_1, \tag{70}$$

$$\Phi_1=c^{1/2}\exp\left\{\frac{1}{4ik_\nu}\int[wm_0^2-(2gw^{1/2}T_{zt}+k)^2]\frac{c}{w}\,du\right\}. \tag{71}$$

The explicit form of the operator $\hat{K}_{xy}$ (66) demonstrates that a symmetry about the substitution $\mathbf{x}\leftrightarrow\mathbf{y}$ exists between the variables $\mathbf{x}$ and $\mathbf{y}$.

When the separating of variables $\mathbf{x}$ from $\mathbf{y}$ in the operator $\hat{K}_{xy}$ is fulfilled, we have four solutions for the separating matrix, namely:

$$\Lambda=\gamma^y,\gamma^y\gamma^z\gamma^t,\gamma^x,\gamma^x\gamma^z\gamma^t. \tag{72}$$

In the process γ^y, $\gamma^y\gamma^z\gamma^t$, are equivalent to the solutions **x**, **xzt** in terms of **x**↔**y** symmetry. Because of this, it suffices to make an allowance, for example, the solutions γ^y and $\gamma^y\gamma^z\gamma^t$. Since the structure of the Dirac matrices γ^y and $\gamma^y\gamma^z\gamma^t$ is the same, specifically in the representation of Eq. (20), both matrices are the diagonal ones, systems of four equations correlated with the operators $\hat{K}_x$ and $\hat{K}_y$ will be of the same structure for both $\mathbf{\Gamma} = \gamma^y$, $\gamma^y\gamma^z\gamma^t$, and possibilities of the adequate exact solutions for the variables **x** and **y** will be associated with one and the same metric structure. Hence, we give here the results for the unique solution of Eq. (72), precisely: $\mathbf{\Gamma} = \gamma^y\gamma^z\gamma^t$.

As a result, we have the following for the operators $\hat{K}_x$ and $\hat{K}_y$:

$$\hat{\mathbf{K}}_{\mathbf{x}} = \gamma^x\gamma^y\left(\frac{\tilde{\mathbf{a}}_{\mathbf{y}}}{\tilde{\mathbf{a}}_{\mathbf{x}}}\right)^{1/2}\left(\frac{\partial}{\partial \mathbf{x}} - \frac{\partial_x \tilde{\mathbf{a}}_{\mathbf{x}}}{4\tilde{\mathbf{a}}_{\mathbf{x}}}\right) - \gamma^y\gamma^z\gamma^t \mathbf{k}\tilde{\mathbf{a}}_{\mathbf{y}}^{1/2}, \tag{73}$$

$$\hat{\mathbf{K}}_y = \frac{1}{\tilde{b}_y^{1/2}}\left(\frac{\partial}{\partial y} - \frac{\partial_y \tilde{b}_y}{4\tilde{b}_y}\right). \tag{74}$$

Here, **k** is the separation constant of **z**, **t** from **x**, **y**.

Calculations totally analogous to the previous ones give the following:

For the **x** variable:

$$\Theta_1(x) = \Theta_4(x), \quad \Theta_2(x) = \Theta_3(x),$$
$$\Theta_\nu(x) = a_\nu^{1/4}\tilde{\Theta}_\nu(x), \quad \nu = 1,2,3,4, \tag{75}$$

$$\tilde{\Theta}_1(x) = \frac{\dfrac{\partial}{\partial x}\tilde{\Theta}_3(x)}{k - \tilde{k}}, \quad \tilde{\Theta}_2(x) = \frac{\partial_x \tilde{\Theta}_4(x)}{k + \tilde{k}}, \tag{76}$$

$$\Theta_1(x) = \exp\left\{\pm(k^2 - \tilde{k}^2)\int \tilde{a}_x^{1/2} dx\right\}. \tag{77}$$

For the **y** variable:

$$\Theta_\nu(y) = \tilde{b}_y^{1/4}\exp\left\{-\tilde{k}\int \tilde{b}_y^{1/2} dy\right\}. \tag{78}$$

Here $\tilde{k}$ is constant of the separation **x** from **y**.

Considering Eqs. (64) and (72), for the solution we finally arrive to

$$\Psi = e^{ik_\nu(x+t)}\begin{pmatrix} \Phi_1(z-t)\Theta_1(x) \\ \Phi_2(z-t)\Theta_2(x) \\ -\Phi_1(z-t)\Theta_2(x) \\ -\Phi_2(z-t)\Theta_1(x) \end{pmatrix}\Theta(y). \tag{79}$$

These solutions are associated with the metric of the form

$$ds^2 = w(z-t)\{\tilde{a}_x(dx)^2 + \tilde{b}_y(dy)^2\} + c(z-t)\{(dz)^2 - (dt)^2\}. \tag{80}$$

III. CONCLUSIONS

This work as well as the previous one[1] presents a series of exact solutions for the Dirac equation in case of an external wave field in the background of gravitation. All the specific types of wave fields have been considered, namely: scalar, vector, axial, pseudoscalar,[1] and tensor fields (in this work).

All the solutions are characterized by the limit transition to the case without gravitation[26–29] and to the case of the free Dirac particle. Discussion of such limit transition is presented in Ref. 1.

Here we should also say some words about metrics associated with the exact solutions presented in the Ref. 1 and in the present article [see formulas (55), (62), (80) and also (3.52), (4.42), (7.40) in Ref. 1].

Naturally, the relevant metrics (55), (62), and (80) turned out to be the conformally plane-wave ones. By the appropriate transformation of coordinates these metrics are easily reduced to the plane-wave ones. It is natural, because the same Volkov problem proper without gravitation, is characterized by a plane symmetry, and therefore its exact solutions proved to be possible only in the Cartesian and light-cone (also ''plane'') variables. Introduction of gravitation into problems of this type should not ''impair'' the initial symmetry, otherwise, as has been demonstrated in our studies, e.g., Ref. 34, separation of variables would be impossible even in the case of the initial problem.

Note here, that plane metrics of the general relativity theory have been considered by a number of authors. Specifically, Taub[35] and McVittie[36] have demonstrated that nonpolarized plane gravitational waves are nonexistent. Bondi[37,38] has succeeded in showing that field equations permit the existence of ''plane'' wave zones having finite lengths between two regions of plane space–time. These metrics were deduced taking into account the general matric form proposed by Rosen.[39]

The metrics pointed out in the present work proved to be conformally similar in their structure and symmetry to the above-mentioned solutions.[35–39] It should be noted that a particular case of a plane nonwave gravitational metric ($d(z-t)=1$, $c(z-t)=1$), with a special selection of the functions $\tilde{a}_x(x)$ and $\tilde{b}_y(y)$ in the metrics specified by us in Eqs. (55), (62), (80) involves the degenerate case of the Kasner solution,[40] that is in agreement with a specific result of our work.[29]

ACKNOWLEDGMENTS

One of us (G.V.S.) wishes to thank Professor G. Neugebauer, Professor D. Kramer, Professor E. Schmutzer, and Professor H. Stephani, from Jena Friedrich Schiller University and Professor K. Just from the University of Arizona for critical remarks and stimulating discussion. He is also very grateful to the General Director of the BelAspher Co. Ltd., Victor Korobov, for financial support of this work.

[1] G. V. Shishkin and W. H. Abdel Salam, J. Math. Phys. (to be published).
[2] D. M. Volcov, Z. Phys. **94**, 250 (1935).
[3] D. M. Volcov, JETF **7**, 1286 (1937) (in Russian).
[4] V. G. Bagrov and D. M. Gitman, *Exact Solutions of Relativistic Wave Equations* (Kluwer, Dordrecht, 1990).
[5] V. G. Bagrov *et al.*, Izv. Vuzov, Fiz. **7**, 95 (1973) (in Russian).
[6] V. G. Bagrov *et al.*, Izv. Vuzov, Fiz. **4**, 29 (1975) (in Russian).
[7] V. G. Bagrov *et al.*, Izv. Vuzov, Fiz. **7**, 6 (1975) (in Russian).
[8] V. G. Bagrov *et al.*, Izv. Vuzov, Fiz. **8**, 73 (1975) (in Russian).
[9] V. G. Bagrov *et al.*, Izv. Vuzov, Fiz. **9**, 106 (1975) (in Russian).
[10] F. I. Fedorov, *Lorentz Group* (Science, Moscow, 1979) (in Russian).
[11] F. I. Fedorov, Rep. USSR Acad. Sci. **174**, 336 (1967) (in Russian).
[12] A. F. Radyuk and F. I. Fedorov, Rep. USSR Acad. Sci. **211**, 1091 (1973) (in Russian).
[13] F. I. Fedorov and A. F. Radyuk, Sov. J. Nucl. Phys. **20**, 509 (1974).
[14] B. V. Krilov *et al.*, Spin Particles in the Field of Plane Electromagnetic Wave, Minsk, Preprint of Physics Institute No. 113, 1976.
[15] A. F. Radyuk and F. I. Fedorov, Issue Byelorussian Acad. Sci., Phys. Math. **4**, 64 (1971) (in Russian).
[16] B. V. Krilov and F. I. Fedorov, Issue Byelorussian Acad. Sci., Phys. Math. **2**, 47 (1976) (in Russian).

[17] F. I. Fedorov *et al.*, Theor. Math. Phys. **35**, 401 (1978).
[18] F. I. Fedorov *et al.*, Acta Phys. Pol. **89**, 787 (1978).
[19] V. G. Zhulego *et al.*, Spin Particles in the Field of the Quantized Plane Electromagnetic Wave, Minsk, Preprint of Physics Institute No. 181, 1979.
[20] F. I. Fedorov, JETF **35**, 493 (1958) (in Russian).
[21] F. I. Fedorov, Rep. Byelorussian Acad. Sci. **11**, 681 (1967) (in Russian).
[22] F. I. Fedorov and A. A. Kirillov, Acta Phys. Pol. **87**, 161 (1976).
[23] V. I. Ritus, Trans. Lebedev Phys. Inst. USSR Acad. Sci. **111**, 5 (1979) (in Russian).
[24] A. I. Nikishev and V. I. Ritus, JETF **19**, 529 (1964) (in Russian).
[25] G. V. Shishkin, Nuovo Cimento B **1068**, 1137 (1991).
[26] G. V. Shishkin and M. A. Yasin, Int. J. Mod. Phys. **6**, 4831 (1991).
[27] G. V. Shishkin and M. A. Yasin, Sov. Nucl. Phys. **53**, 536 (1991).
[28] G. V. Shishkin, in *Gravitation Energy and Gravitation Waves*, edited by N. A. Chernikiv (JINR, Dubna, 1993) (in Russian).
[29] G. V. Shishkin and W. D. Cabos, J. Math. Phys. **32**, 3184 (1991).
[30] G. V. Shishkin and W. D. Cabos, J. Math. Phys. **33**, 914 (1992).
[31] G. V. Shishkin and V. M. Villalba, J. Math. Phys. **33**, 2093 (1992).
[32] Particle Properties Data booklet, Phys. Rev. D **45**, S1 (1992).
[33] G. V. Shishkin, Rep. Byelorussian Acad. Sci. **23**, 336 (1989) (in Russian).
[34] A. H. Taub, J. Math. Phys. **53**, 472 (1951).
[35] G. C. McVittie, J. Mech. Anal. **4**, 201 (1955).
[36] H. Bondi, F. A. E. Pirani, and L. Robinson, Proc. R. Soc. London, Ser. A **251**, 519 (1959).
[37] H. Bondi, Nature (London) **179**, 1072 (1957).
[38] N. Rosen, Bull. Res. Council Isr. **3**, 328 (1953).
[39] E. Kasner, Am. J. Math. **43**, 1234 (1921).

Algebraic solutions for point groups: The tetrahedral group for the group chain $T \supset C_3$

Jin-Quan Chen[a] and Peng-Dong Fan
Department of Physics, Nanjing University, Nanjing, 210008, People's Republic of China

(Received 1 April 1997; accepted for publication 16 September 1997)

A new technique, the double-induced technique, is introduced for constructing irreducible symmetry operators (ISO's) adapted to the group chain $G \supset G(s)$. Simple algebraic expressions of irreducible matrices, the ISO's and symmetry adapted functions (SAF's) are derived for the group chain $T \supset C_3$ for both single-valued and double-valued representations in a unified way. The simplicity of the results lies in the fact that the ISO's and SAF's are functions of only the quatum numbers of the group chain [the analogy of (j,m) for the group chain $SO_3 \supset SO_2$], without involving any irreducible matrix elements. The symmetries of the SAF's are disclosed for the first time. © *1998 American Institute of Physics.*
[S0022-2488(98)00201-1]

I. INTRODUCTION

It is known that the group SO_3 has analytic solutions with the primitive character $\chi(\varphi)=\sin(j+\frac{1}{2})\varphi/\sin(\varphi/2)$ and irreducible matrices $D^j_{mm'}(\alpha,\beta,\gamma)$, the well known D-functions.[1] Although any point group is a subgroup of O_3, however until present all solutions to the point groups are numerical.[2–8] It means that the primitive characters, irreducible matrices, irreducible symmetry operators (ISO's), and symmetry adapted functions (SAF's) are usually tabulated for each irrep of each group, with pages and pages of tables. It is highly desirable to seek algebraic solutions for the point groups so that the characters, irreducible matrices, ISO's and SAF's all have closed and simple formulas.

Using the symmetrized boson representation, the algebraic expressions for the single-valued (s.v.) irreducible basis have been obtained for the groups O_h and I_h.[9,10] The advantages of the symmetrized boson representation are that it allows one to get the algebraic expression of the irreducible bases for all possible cases without projection and provides a concise and intuitive picture for the vibration modes of molecules. The shortcomings of the symmetrized boson representation method are that it is applicable only to the single-valued reps and the derivation is rather complicated.

Recently, using the newly discovered factorization lemma for the semi-direct product of two Abelian groups, unified algebraic expressions are obtained for both the s.v. and double-valued (d.v.) representations of the characters, irreducible matrices, ISO's and SAF's of all dihedral groups.[11,12]

The purpose in the present series of work is to develop a new technique, called the double-induced representation technique, which can be used for obtaining algebraic solutions of any point groups for both s.v. and d.v. reps, and for the s.v. reps it is equivalent to the symmetrized boson representation. In this paper the technique is applied to the tetrahedral group T. Algebraic expressions of the ISO's of T are derived for both single-valued and double-valued reps in a unified way without using any knowledge of the irreducible matrices. The algebraic expressions of the ISO's and SAF's are functions of the quantum numbers characterizing the irreducible basis vector, without involving any irreducible matrix elements.

The outline of the paper is as follows: The double-induced rep technique is introduced in Sec.

[a]Electronic mail: jqchen@chenwang.nju.edu.cn

II and a method for uniquely determining the Euler angles of a double point group is given in Sec. III. The algebraic solutions for the group chains $T \supset C_3$ (for both s.v. and d.v.) reps, are derived in Sec. IV, while the final section is a discussion and summary.

II. THE DOUBLE-INDUCED REPRESENTATION

Since we are dealing with both the s.v. and d.v. reps in a unified way, let us begin with a brief introduction of the double point group. Consider a point group G,

$$G=\{R_i : i=1,2,\ldots,|G|\}, R_i=R_{\mathbf{n}_i}(\omega_i), \tag{1a}$$

where $R_{\mathbf{n}_i}(\omega_i)$ denotes a rotation about the axis $\mathbf{n}_i$ through angle ω_i. The corresponding double point group is

$$G^{\dagger}=\{R_i, \widetilde{R}_i : i=1,2,\ldots,|G|\}, \widetilde{R}_i=R_{\mathbf{n}_i}(\omega_i+2\pi). \tag{1b}$$

In the single-valued (s.v.) reps, or the vector reps, $\widetilde{R}_i=R_i$, while in the double-valued (d.v.), or spinor reps, $\widetilde{R}_i=-R_i$. It is convenient to introduce an index θ for specifying the type of representations,

$$\widetilde{R}_i=\theta R_i, \theta=\begin{cases} 1, & \text{for s.v. reps,} \\ -1, & \text{for d.v. reps.} \end{cases} \tag{2}$$

In deriving an algebraic expression it is vital to use quantum numbers to label irreps. It is pointed out[7] that we can use the eigenvalue λ of the so called CSCO-I or CSCO (complete set of commuting operators) of G to label irreps. The CSCO of a finite group, denoted by C, is the analogy of the set of Casimir operators of a Lie group, and the CSCO of any pure rotation point group except for the dihedral groups consists of only a single class operator.[11] Since the dihedral groups have been discussed in Ref. 12, in the following we assume that C consists of only a single class operator, say the i-th class operator C_i. For example, the CSCO of T is $C=C_{31}^{+}+C_{32}^{+}+C_{33}^{+}+C_{34}^{+}$ in the notation of Ref. 3. The correspondence between the eigenvalue λ of $C=C_i$ and the conventional irrep label (the Mulliken notation or Koster notation) is easily obtained according to the following: The eigenvalue of the class operator $C=C_i$ in the irrep Γ, denoted by λ^{Γ}, is related to the character χ_i^{Γ} as (see p. 60 in the book of Ref. 7)

$$\lambda^{\Gamma}=\frac{g_i}{h_{\Gamma}}\chi_i^{\Gamma}, \tag{3}$$

where g_i is the number of elements in the i-th class and h_{Γ} is the dimension of the irrep Γ. Using (3) from the known characters one can know immediately λ^{Γ}, and thus the correspondence is easily established between the eigenvalue λ and the the conventional irrep label Γ. For the group T, the correspondence is shown in Table I between λ and the irrep label in Koster notation Γ_1–Γ_7,[2] as well as the modified Mulliken notation A_0, A_1, A_{-1}, F for the s.v. reps and $E_{\mu}, \mu=\frac{1}{2}, \frac{3}{2}, -\frac{3}{2}$ for the d.v. reps. As will be seen the modified Mulliken notation is most suitable for obtaining algebraic solutions. Henceforth, the eigenvalue λ of the CSCO of G is referred to as the *quantum number* of the group G.

It should be pointed out that using the double-induced rep technique we can derive *ab initio* everything, the irreducible characters and matrices, the ISO's and SAF's without borrowing any thing from other work. Using the known characters is only for identifying irreps.

TABLE I. The eigenvalues λ,ρ_μ of the CSCO-II, (C,C_{3z}), of T.

	A_0 Γ_1	A_1 Γ_2	A_{-1} Γ_3	F Γ_4	$E_{1/2}$ Γ_5	$E_{3/2}$ Γ_6	$E_{-3/2}$ Γ_7
λ	4	4ϵ	$4\epsilon^*$	0	2	2ϵ	$2\epsilon^*$
μ	0	1	-1	$(1,0,-1)$	$(\frac{1}{2},-\frac{1}{2})$	$(\frac{3}{2},\frac{1}{2})$	$(\frac{3}{2},-\frac{1}{2})$
$\rho_\mu=\exp\left(-\frac{2\pi\mu i}{3}\right)$	1	ϵ	ϵ^*	$(\epsilon,1,\epsilon^*)$			

[a] $\epsilon=\exp(-2\pi i/3)$.

A central task in the group representation theory is to find the irreducible symmetry operators (ISO's). Suppose that the ISO is symmetry adapted to a canonical subgroup chain $G\supset G(s)$. The normalized $G\supset G(s)$ ISO is defined as (the usual form of the ISO is denoted as $\mathscr{P}^{(\lambda)}_{\rho\rho}=\sqrt{h_\lambda/|G|}\,\mathscr{P}^{(\lambda)}_{\rho}{}^{\overline{\rho}}$)

$$\mathscr{P}^{(\lambda)}_{\rho}{}^{\overline{\rho}}=\sqrt{\frac{h_\lambda}{|G|}}\sum_{a=1}^{|G|}D^{(\lambda)}_{\rho\rho}(R_a)^*R_a, \tag{4}$$

where h_λ is the dimension of the irrep λ, $|G|$ the order of G, R_a the a-th group element, and $D^{(\lambda)}_{\rho\rho}(R_a)$ is the irreducible matrix element of R_a in the $G\supset G(s)$ basis,

$$D^{(\lambda)}_{\rho\rho}(R_a)=\langle\psi^\lambda_\rho|R_a|\psi^\lambda_\rho\rangle.$$

Applying $\mathscr{P}^{(\lambda)}_{\rho}{}^{\overline{\rho}}$ to a normalized generic wave function, we can get all the orthonormalized irreducible basis vectors.

The intrinsic operator $\overline{R}$ of the intrinsic group $\overline{G}=\{\overline{R},\overline{S},\overline{T},\ldots\}$ is defined by[7]

$$\overline{RS}=SR,\quad \text{for any}\quad S\in G. \tag{5}$$

Notice that the set of intrinsic operators $\{\overline{R_i}^{-1},i=1,2,\cdot,|G|\}$ form what is called the right regular representation of G.[13]

If $C(s)$ is the CSCO of the subgroup $G(s)$, then $\overline{C}(s)$ is the CSCO of the intrinsic subgroup $\overline{G}(s)$, where $\overline{C}(s)$ is obtained from $C(s)$ by replacing the elements R in it by $\overline{R}$. The set of operators $(C,C(s))$ and $(C,C(s),\overline{C}(s))$ are called the CSCO-II and CSCO-III of G, which are the analogies of (J^2,J_z) and $(J^2,J_z,\overline{J_z})$ for the rotation group. According to Ref. 9 the ISO's adapted to the group chain $G\supset G(s)$ can be obtained by finding the simultaneous eigenvectors of the CSCO-III of G,

$$\begin{pmatrix} C \\ C(s) \\ \overline{C}(s)\end{pmatrix}\mathscr{P}^{(\nu)}_{\rho}{}^{\overline{\rho}}=\begin{pmatrix}\lambda\\ \rho\\ \overline{\rho}\end{pmatrix}\mathscr{P}^{(\lambda)}_{\rho}{}^{\overline{\rho}}. \tag{6}$$

The ISO $\mathscr{P}^{(\lambda)}_{\rho}{}^{\overline{\rho}}$ is a $G\supset G(s)$ basis $|\lambda\rho\rangle$ and $\overline{G}\supset\overline{G}(s)$ basis $|\lambda\,\overline{\rho}\rangle$. The $\overline{\rho}$ and ρ will be referred to as the intrinsic and external quantum numbers, respectively.

The expression for the ISO (4) can be greatly simplified by using its double-coset factored form.[14,15] Suppose that G can be decomposed into p right cosets and $q-p$ double cosets with each element occurring only once either in a right coset or a double coset,

$$G=G(s)(e+k_2+\cdots+k_p)+G(s)(k_{p+1}+\cdots+k_q)G(s), \tag{7}$$

where k_i ($k_1\equiv e$) are coset generators. Furthermore, assume that $G(s)=\{s_i:\ i=1,2,...,|G(s)|\}$ is Abelian with the un-normalized projection operator $P^{\rho}=\sum_{i=1}^{|G(s)|}D^{(\rho)}(s_i)^*s_i$. Then it is easy to show that the ISO in (4) can be expressed as

$$\mathscr{P}_{\rho}^{(\lambda)\bar{\rho}}=\sqrt{\frac{h_\lambda}{|G|}}\left[\sum_{i=1}^{p}D_{\rho\rho}^{(\lambda)}(k_i)^*P^{\rho}k_i+\sum_{i=p+1}^{q}D_{\rho\rho}^{(\lambda)}(k_i)^*P^{\rho}k_iP^{\bar{\rho}}\right]. \tag{8}$$

The ISO in the factored form sums over only the q coset generators weighted by their matrix elements.

In the numerical approach to the eigenfunction method,[7] we diagonalize the CSCO-III in the $|G|$-dimensional group space and for each set of eigenvalues $(\lambda,\rho,\bar{\rho})$ we get one eigenvector $\mathscr{P}_{\rho}^{(\nu)\rho}$ in the original form (4). A packet of programs has been written for this purpose.[16] By inputing the names of the group G and its subgroup $G(s)$, the computer will generate the characters, irreducible matrices, the Clebsch–Gordan coefficients for any crystal point group.

Now our purpose is to obtain $\mathscr{P}_{\rho}^{(\nu)\rho}$ in the factored form (8) as an algebraic function of $(\lambda,\rho,\bar{\rho})$, and we have to change our strategy. The new strategy consists of the following steps

(1) Decompose the group G into double cosets with respect to the subgroup $G(s)$,

$$G=G(s)(e+k_2+k_3+\cdots+k_q)G(s). \tag{9a}$$

In this paper we only deal with the case when $G(s)$ is a cyclic group of order n with the generator C_{nz}. In the decomposition (9a) some double cosets, say the first p double cosets, are coincident with the right or left cosets,

$$k_iG(s)=G(s)k_i,\quad i=1,2,\ldots,p, \tag{9b}$$

and Eq. (9a) reduces to (7).

(2) Construct the orthonormal common eigenvectors $\varphi_{\rho\rho}^{i}$ of $C(s)$ and $\bar{C}(s)$ by applying the projection operator of the group $G(s)$ and $\bar{G}(s)$ to the double-coset generators k_i,

$$\varphi_{\rho\rho}^{i}=\text{const}\ P^{\rho}\bar{P}^{\bar{\rho}}k_i=\text{const}\ P^{\rho}k_iP^{\bar{\rho}}, \tag{9c}$$

where P^{ρ} is the projection operator of $G(s)$ with the eigenvalue ρ of its generator C_{nz} as its irrep label, and $\bar{P}^{\rho}$ is obtained from P^{ρ} by replacing the group element R with $\bar{R}$. For deriving (9c) the defining equation (5) for the intrinsic operator $\bar{R}$ has been used. The basis vectors $\varphi_{\rho\rho}^{i}$ form what is called the *double-induced representation*, which is a representation induced from the irreps $(\rho,\bar{\rho})$ of the group $G(s)\times\bar{G}(s)$. It is seen that the basis vectors $\varphi_{\rho\rho}^{i}$ are constructed by applying the projection operators P^{ρ} and $P^{\bar{\rho}}$ to the coset generator k_i from left and right, respectively, or in a manner of speaking, by sandwiching k_i between P^{ρ} and $P^{\bar{\rho}}$.

Notice that due to (9b) and the property of the projection operator, we have

$$k_iP^{\bar{\rho}}=P^{f_i(\bar{\rho})}k_i,\varphi_{\rho\rho}^{i}=\text{const}\ \delta_{\rho,f_i(\bar{\rho})}P^{\rho}k_i,\quad i=1,2,\cdots,p, \tag{9d}$$

where $f_i(\bar{\rho})$ is a simple function of $\bar{\rho}$. It is found that for all point groups $f_i(\bar{\rho})$ is either equal to $\bar{\rho}$ [evidently $f_1(\bar{\rho})\equiv\bar{\rho}$] or $\bar{\rho}^*$. Therefore there are only pn instead of pn^2 linearly independent basis vectors $\varphi_{\rho\rho}^{i}, i=1,2,\ldots,p$ and the total number of linearly independent basis vectors in the

double induced rep is $pn+(q-p)n^2$, which is just the order of the group G. It shows that although there are qn^2 elements in the double coset decompostion (9a), no double counting problem arises here.

It is also to be noticed that due to (9d) the dimension of the double-induced rep depends on the values of ρ and $\overline{\rho}$, and the maximum dimension is equal to q, the number of double cosets of G with respect to $G(s)$, which is usually much less than the order $|G|$.

(3) Construct the representation matrix, M, of the CSCO, C, in the double-induced rep,

$$M_{ij}=\langle \varphi^{i}_{\rho\rho}|C|\varphi^{j}_{\rho\rho}\rangle. \tag{9e}$$

Notice that the matrix $M\equiv M(\rho,\overline{\rho})$ depends on ρ and $\overline{\rho}$.

(4) Diagonalize C in the double-induced rep space $L_{\rho,\overline{\rho}}=\{\varphi^{i}_{\rho\overline{\rho}}: i=1,2,\ldots,q\}$ with the $\rho,\overline{\rho}$ as parameters. The ISO's can be expressed as

$$\mathscr{P}^{(\lambda)\overline{\rho}}_{\rho}=\sum_{i=1}^{q} u_i \varphi^{i}_{\rho\overline{\rho}}. \tag{9f}$$

The coefficients u_i are determined from the following matrix equation:

$$\tilde{M}\mathbf{u}=\lambda\mathbf{u}, \tag{9g}$$

where $\mathbf{u}=\{u_1,u_2,\ldots,u_q\}_{\text{column}}$ is a column vector.

Following the above four steps we can obtain a unified and compact expression of the ISO's $\mathscr{P}^{(\lambda)\overline{\rho}}_{\rho}$ as a function of all possible values of λ, ρ, and $\overline{\rho}$. The s.v. and d.v. reps are distinguished by the possible values of quantum numbers λ, ρ and $\overline{\rho}$, and are obtained simultaneously.

(5) According to Eq. (8), the coefficients in front of the coset generator k_i in the normalized ISO's are equal to $\sqrt{h_\lambda/|G|}$ times the complex conjugate of the matrix element of k_i. In this way we can obtain an algebraic expression of the irreducible matrices of all coset generators. The irreducible matrices of other elements are easily found from those of the coset generators.

(6) For an irrep of dimension h there are h sets of ISO's distinguished by the intrinsic quantum number $\overline{\rho}$ (or the column index in the conventional language). In many cases, for example, for obtaining the SAF's, only one set of the ISO's is sufficient and the normalization is unimportant. Therefore for each group we give the explicit expression for the chosen set of unnormalized ISO's, which are termed the generalized projection operators and designated by $\mathbf{P}^{(\Gamma)}_{\rho}$.

Up to this point, the coordinate system is irrelevant. For obtaining the SAF's, we need to specify the coordinate system, and determine the Euler angles of the double-coset generators in the coordinate system. Then we apply the generalized projection operators $\mathbf{P}^{(\Gamma)}_{\rho}$ to the eigenvector $|jm\rangle$ of angular momentum, and obtain algebraic expressions for the SAF's for both s.v. and d.v. reps.

Since the normalization of the ISO's are not important except for the case of reading out the irreducible matrix elements from the ISO's as stated in step (5), we are sloppy about the numerical factor for the ISO's in the process of derivation for the sake of clarity. The normalization is easily found according to the assumption that the scalar product of two elements is $(R_a,R_b)=\delta_{ab}$ and will be inserted into the final expression of the ISO's.

III. THE CONSTRUCTION OF THE GROUP TABLE OF $T^{\dagger}$

The starting point for finding algebraic solutions to the double point group $G^{\dagger}$ is constructing its group table. The group table of a double point group $G^{\dagger}$ is usually constructed in the following way: Find the Euler angles (α,β,γ), $0\leqslant\alpha,\gamma\leqslant 2\pi$, $0\leqslant\beta\leqslant\pi$, for each element of G, assign the

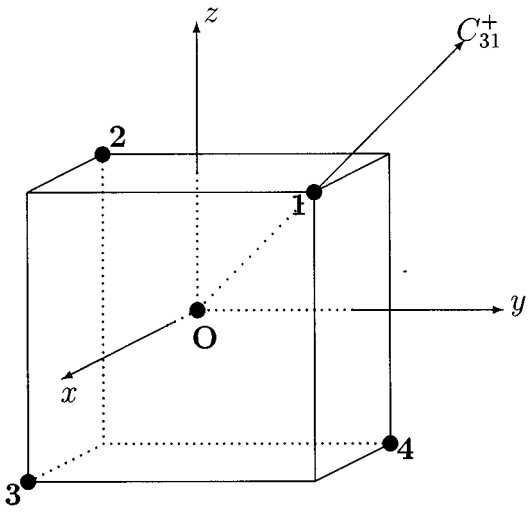

FIG. 1. The coordinate system xyz for $T \supset D_2$, with **1, 2, 3** and **4** indicating the four vertices of the tetrahedral.

matrices $\pm D^{1/2}(\alpha,\beta,\gamma)$ to the elements $(R,\widetilde{R})$, then by matrix multiplication we can form the group table of $G^{\dagger}$. However there is an ambiguity in the assignment (see p. 424 of Ref. 3): one could assign $R=D^{1/2}(\alpha,\beta,\gamma)$, $\widetilde{R}=-D^{1/2}(\alpha,\beta,\gamma)$ or $R=-D^{1/2}(\alpha,\beta,\gamma)$, $\widetilde{R}=D^{1/2}(\alpha,\beta,\gamma)$, and there is no definite rule for the assignment with the consequence that an inconsistency may arise. For example, in p. 422 of Ref. 3 the assignment is $C_{4z}=D^{1/2}(0,0,\pi/2)$, $C_{2z}=-D^{1/2}(0,0,\pi)$, which implies that $C_{4z}=\exp(-i(\pi/2)J_z)$, $C_{2z}=-\exp(-i\pi J_z)$ instead of $C_{2z}=C_{4z}^2=\exp(-i\pi J_z)$.

For avoiding this inconsistency, we enlarge the domain of the Euler angles as

$$0\leqslant\alpha\leqslant 2\pi, 0\leqslant\gamma\leqslant 4\pi, 0\leqslant\beta\leqslant\pi,$$

and determine the Euler angles for the rotation through angle ω about the axis **n** with azimuthal angles θ,φ, $R_{\mathbf{n}(\theta\varphi)}(\omega)$, uniquely by the formula[14]

$$\begin{pmatrix} e^{-(i/2)(\alpha+\gamma)}\cos\dfrac{\beta}{2}, & -e^{-(i/2)(\alpha-\gamma)}\sin\dfrac{\beta}{2} \\ e^{(i/2)(\alpha-\gamma)}\sin\dfrac{\beta}{2}, & e^{(i/2)(\alpha+\gamma)}\cos\dfrac{\beta}{2} \end{pmatrix} = \begin{pmatrix} \cos\dfrac{\omega}{2}-in_z\sin\dfrac{\omega}{2}, & -(n_y+in_x)\sin\dfrac{\omega}{2} \\ (n_y-in_x)\sin\dfrac{\omega}{2}, & \cos\dfrac{\omega}{2}+in_z\sin\dfrac{\omega}{2} \end{pmatrix}, \tag{10}$$

where

$$n_z=\cos\,\theta,\quad n_x=\sin\,\theta\,\cos\varphi,\quad n_y=\sin\,\theta\,\sin\,\varphi.$$

With the following unique assignment:

$$R_i=D^{1/2}(\alpha_i,\beta_i,\gamma_i),\quad \widetilde{R}_i=-D^{1/2}(\alpha_i,\beta_i,\gamma_i),$$

and using the D-function[17] given in the left hand side of (10) we can construct the group table for any double point group.

The coordinate system for the group T is usually chosen as in Fig. 1. The Euler angles for T and the group table of $T^{\dagger}$ are listed in Table II and Table III.

TABLE II. The Euler angles for $T^{\dagger}$ in the coordinate system of Fig. 1.

Element	α	β	γ	Element	α	β	γ
E	0	0	0	C_{2x}	0	π	π
C_{2y}	0	π	0	C_{2z}	0	0	π
C_{31}^{+}	0	$\frac{\pi}{2}$	$\frac{\pi}{2}$	C_{32}^{+}	π	$\frac{\pi}{2}$	$\frac{7\pi}{2}$
C_{33}^{+}	π	$\frac{\pi}{2}$	$\frac{5\pi}{2}$	C_{34}^{+}	0	$\frac{\pi}{2}$	$\frac{7\pi}{2}$
C_{31}^{-}	$\frac{\pi}{2}$	$\frac{\pi}{2}$	3π	C_{32}^{-}	$\frac{3\pi}{2}$	$\frac{\pi}{2}$	2π
C_{33}^{-}	$\frac{\pi}{2}$	$\frac{\pi}{2}$	0	C_{34}^{-}	$\frac{3\pi}{2}$	$\frac{\pi}{2}$	3π

Remark 1: It is found that the new group table has the desired feature that the elements $R,S,T,\ldots,$ belong to the same class and $\widetilde{R},\widetilde{S},\widetilde{T},\ldots,$ belong to a different class, instead of a mix up of tilded and untilded elements in the same class.

Remark 2: The Euler angles in Table II are different from those given in p. 54 by Bradley and Cracknell[3] (notice that the definition of the Euler angles α and γ in Ref. 3 is opposite to what is used here).

In later developments we need to use different quantization axes for different subgroup chains and it is important to know the relation between the Euler angles of the same operator in two different coordinates systems. Suppose that the new coordinate system $x'y'z'$ is obtained from the original one xyz by a rotation $R(\phi,\theta,\zeta)$, i.e., $xyz \xrightarrow{R(\phi,\theta,\zeta)} x'y'z'$, then the Euler angles (α',β',γ') of an operator in the new system is determined by

$$D^{1/2}(\alpha',\beta',\gamma')=D^{1/2}(-\zeta,-\theta,-\phi)D^{1/2}(\alpha,\beta,\gamma)D^{1/2}(\phi,\theta,\zeta). \tag{11}$$

In passing we note that through out the paper we will use the Condon–Shortley definition for $|jm\rangle$. When $j=l$, l being an integer,

$$|lm\rangle \equiv Y_{lm},$$

TABLE III. The group table of $T^{\dagger}$.

E	1	2	3	4	5	6	7	8	9	10	11	12
C_{2x}	2	$\widetilde{1}$	4	$\widetilde{3}$	$\widetilde{8}$	7	$\widetilde{6}$	5	10	$\widetilde{9}$	12	$\widetilde{11}$
C_{2y}	3	$\widetilde{4}$	$\widetilde{1}$	2	$\widetilde{6}$	5	8	$\widetilde{7}$	11	$\widetilde{12}$	$\widetilde{9}$	10
C_{2z}	4	3	$\widetilde{2}$	$\widetilde{1}$	$\widetilde{7}$	$\widetilde{8}$	5	6	12	11	$\widetilde{10}$	$\widetilde{9}$
C_{31}^{+}	5	$\widetilde{6}$	$\widetilde{7}$	$\widetilde{8}$	$\widetilde{9}$	12	10	11	1	3	4	2
C_{32}^{+}	6	5	8	$\widetilde{7}$	11	$\widetilde{10}$	12	9	$\widetilde{3}$	1	$\widetilde{2}$	4
C_{33}^{+}	7	$\widetilde{8}$	5	6	12	9	$\widetilde{11}$	10	$\widetilde{4}$	2	1	$\widetilde{3}$
C_{34}^{+}	8	7	$\widetilde{6}$	5	10	11	9	$\widetilde{12}$	$\widetilde{2}$	$\widetilde{4}$	3	1
C_{31}^{-}	9	12	10	11	1	$\widetilde{2}$	$\widetilde{3}$	$\widetilde{4}$	$\widetilde{5}$	7	8	6
C_{32}^{-}	10	$\widetilde{11}$	$\widetilde{9}$	12	2	1	$\widetilde{4}$	3	8	$\widetilde{6}$	5	7
C_{33}^{-}	11	10	$\widetilde{12}$	$\widetilde{9}$	3	4	1	$\widetilde{2}$	6	8	$\widetilde{7}$	5
C_{34}^{-}	12	$\widetilde{9}$	11	$\widetilde{10}$	4	$\widetilde{3}$	2	1	7	5	6	$\widetilde{8}$

where the spherical harmonic Y_{lm} is defined in accordance with Rose.[17] The following properties of the D-function[17] are useful for deriving the SAF's and their symmetries,

$$C_{2x}|jm\rangle=(-)^{-j}|j-m\rangle,\quad C_{2y}|jm\rangle=(-)^{j-m}|j-m\rangle,\qquad C_{2z}|jm\rangle=(-)^{m}|jm\rangle, \tag{12a}$$

$$d^j_{mm'}(\beta)=(-1)^{m-m'}d^j_{m',m}(\beta)=(-1)^{m-m'}d^j_{-m-m'}(\beta),\qquad d^j_{mm'}\left(\frac{\pi}{2}\right)=(-1)^{j-m'}d^j_{-m',m}\left(\frac{\pi}{2}\right)=(-1)^{j-m'}d^j_{-m,m'}\left(\frac{\pi}{2}\right), \tag{12b}$$

where $d^j_{mm'}(\beta)=\langle jm|e^{-i\beta J_y}|jm'\rangle$.

IV. THE GROUP CHAIN $T\supset C_3$

The CSCO of T is

$$C=C_{31}^{+}+C_{32}^{+}+C_{33}^{+}+C_{34}^{+}=\mathbf{5}+\mathbf{6}+\mathbf{7}+\mathbf{8}, \tag{13}$$

where and in the following we use a bold face integer to denote the group element listed in Table III, e.g., **5** represents the element $R_5=C_{31}^{+}$.

From the known characters of the group T and Eq. (3) one can know immediately that C has seven distinct eigenvalues $(4,4\epsilon,4\epsilon^*,0)$, and $(2,2\epsilon,2\epsilon^*)$, $\epsilon=e^{-2\pi i/3}$, corresponding to the seven irreps; in our notation they are the s.v. reps (A_0,A_1,A_{-1},F) and the d.v. reps $(E_{1/2},E_{3/2},E_{-3/2})$; see Table I.

A. The C_3 double-induced representation

It is known that C_3 is a canonical subgroup of T for both the s.v. and d.v. reps and its CSCO can be chosen as $C(s)=C_{31}^{+}$. Choosing the axis of C_{31}^{+} as the quantization axis, the operator C_{31}^{+} can be written as

$$C_{31}^{+}=\exp\left(-\frac{2\pi J_z i}{3}\right). \tag{14}$$

Its eigenvalue is denoted by $\rho\equiv\rho_\mu$,

$$\rho=\rho_\mu=\exp\left(-\frac{2\pi\mu i}{3}\right),\mu=\begin{cases}0,\pm1, & \text{for s. v. reps},\\ \pm\ \frac{1}{2},\ \frac{3}{2}, & \text{for d. v. reps.}\end{cases} \tag{15}$$

As in Ref. 5 we use μ as the label of the irreps of C_3. Notice that the irreps $\frac{3}{2}$ and $-\frac{3}{2}$ of C_3 are equivalent, $-\frac{3}{2}\cong\frac{3}{2}$, and

$$C_{31}^{3+k}=\theta C_{31}^{k},\quad (\rho)^{3+k}=\theta\rho^{k}, \tag{16}$$

where θ is the index introduced in (2).

The CSCO-II of T is (C,C_{31}^{+}), and each irreducible basis vectors of T can be uniquely labeled by the eigenvalues of the CSCO-II, as shown in Table I. For the d.v. reps, it is more convenient to use the value μ instead of the value $\rho_\mu=\exp(-2\pi\mu i/3)$ as the component label.

We begin with the construction of the eigenvectors $\varphi^{i}_{\rho\rho}$—of C_{31}^{+} and $\overline{C_{31}^{+}}$ in the group space. The projection operator of C_3 is defined as

$$P^{\mu}\equiv P^{\rho_\mu}=(P^{\mu})^{\dagger}=\frac{1}{2}\sum_{k=0}^{5}(\rho_{\mu}^{*})^{k}(C_{31})^{k}. \tag{17}$$

Regarding P^{ρ} as a vector in the group space, we have

$$(P^{\rho})^{*}=P^{\rho *},\quad (P^{\mu})^{*}=P^{-\mu},\quad P^{3/2}=P^{-3/2}. \tag{18}$$

Due to (16), for both s.v. and d.v. reps we have

$$P^{\rho}=P^{\mu}=e+\rho C_{31}^{-}+\rho^{*}C_{31}^{+}=\mathbf{1}+\rho\mathbf{9}+\rho^{*}\mathbf{5}. \tag{19}$$

The double coset decomposition of the group T with respect to C_3 is

$$T=C_3(e+C_{2z})C_3. \tag{20}$$

Notice that any member in a coset can be chosen as the coset generator, but once chosen, it should be kept fixed. Here we choose e and C_{2z} as the coset generators.

According to (9c), sandwiching the two coset generators (e,C_{2z}) between P^{ρ} and $P^{\overline{\rho}}$, and inserting normalization coefficients we get two common eigenvectors of C_{31}^{+} and $\overline{C_{31}^{+}}$,

$$\varphi^{1}_{\rho,\overline{\rho}}=\frac{1}{3\sqrt{3}}P^{\rho}eP^{\overline{\rho}},\quad \varphi^{2}_{\rho,\overline{\rho}}=\tfrac{1}{3}P^{\rho}C_{2z}P^{\overline{\rho}}. \tag{21a}$$

Using $P^{\rho}P^{\rho'}=3\delta_{\rho,\rho'}P^{\rho}$ it is easy to show that

$$\varphi^{1}_{\rho\overline{\rho}}=\delta_{\rho\overline{\rho}}\varphi^{1}_{\rho},\quad \varphi^{1}_{\rho}=\frac{1}{\sqrt{3}}P^{\rho},$$

$$\varphi^{2}_{\rho\overline{\rho}}=\tfrac{1}{3}P^{\rho}C_{2z}P^{\overline{\rho}}=\frac{1}{\sqrt{3}}[\rho^{*}\overline{\rho}\,\varphi^{2}_{\rho}+\rho\overline{\rho^{*}}\varphi^{3}_{\rho}+\varphi^{4}_{\rho}], \tag{21b}$$

where φ^{i}_{ρ} are eigenvectors of C_{3z},

$$\varphi^{1}_{\rho}=\frac{1}{\sqrt{3}}P^{\rho},\quad \varphi^{2}_{\rho}=\varphi^{1}_{\rho}C_{2x},\quad \varphi^{3}_{\rho}=\varphi^{1}_{\rho}C_{2y},\quad \varphi^{4}_{\rho}=\varphi^{1}_{\rho}C_{2z}. \tag{21c}$$

The two basis vectors $\varphi^{1}_{\rho\overline{\rho}}$, $\varphi^{2}_{\rho\overline{\rho}}$—form the double-induced rep. Recall what was mentioned in Sec. II that there are only $12=1\times3+1\times3^{2}$ linearly independent basis vectors in (21b).

Using the group table, the two common eigenvectors of C_{31}^{+} and $\overline{C_{31}^{+}}$ can be written explicitly,

$$\varphi^{1}_{\rho}=\frac{1}{\sqrt{3}}[\mathbf{1}+\rho\mathbf{9}+\rho^{*}\mathbf{5}],$$

$$\varphi^{2}_{\rho\overline{\rho}}=\tfrac{1}{3}[(\mathbf{4}+\rho\mathbf{11}+\theta\rho^{*}\mathbf{8})+\overline{\rho}(\mathbf{12}+\rho\mathbf{6}+\rho^{*}\mathbf{2})+\theta\overline{\rho^{*}}(\mathbf{7}+\theta\rho\mathbf{3}+\rho^{*}\mathbf{10})]. \tag{22}$$

For later convenience, one element in each of the basis vectors of the double-induced rep is chosen as its representative. Although any element in a basis vector can be chosen as its representative, the simplest choice here is to choose the element **1** and **4** as the representatives of φ^{1} and φ^{2}.

The task of finding the eigenvectors of the CSCO-III in the 12-dimensional group space of T is reduced to finding the eigenvector of the CSCO in the two-dimensional double-induced rep space $L_{\rho\bar{\rho}}=\{\varphi^i_{\rho\bar{\rho}}: i=1,2\}$. Since $\rho\equiv\rho_\mu=\exp(-2\pi\mu i/3)$, the basis vectors $\varphi^i_{\rho\bar{\rho}}$ can be written as $\varphi^i_{\mu\bar{\mu}}$, $\mu,\bar{\mu}=0,\pm\frac{1}{2},\pm 1,\frac{3}{2}$.

The next step is to find the representation matrix M of the operator C in the double-induced representation. A key point for finding the matrix elements M_{ji} in $C\varphi^i_{\rho\bar{\rho}}=\Sigma_j M_{ji}\varphi^j_{\rho\bar{\rho}}$, is that we only need to find the coefficients in front of the representatives of $\varphi^1_{\rho\bar{\rho}}$ and $\varphi^2_{\rho\bar{\rho}}$, i.e., in front of **1** and **4**. This can be accomplished by the following procedure: (1). Highlight the elements **1** and **4** in rows 5–8 (which correspond to the elements $C^+_{31}-C^+_{34}$) of the group table of $T^\dagger$. (2) Collecting the highlighted numbers in each column, we obtain the following table:

R_i	**9**	**10**	**11**	**12**
CR_i	**1** $+\theta$**4**	**1**$+\theta$**4**	**1**+**4**	**1** +**4**

The table shows, e.g., $CR_9=R_1+\theta R_4+\cdots$. The terms involving elements other than R_1 and R_4 are ignored. Using the above table and (22) we have

$$C\varphi^1_\rho=\frac{1}{\sqrt{3}}[\rho(\mathbf{1}+\theta\mathbf{4})]=\rho\varphi^1_\rho+\sqrt{3}\,\theta\rho\varphi^2_{\rho\bar{\rho}},$$
$$C\varphi^2_{\rho\bar{\rho}}=\tfrac{1}{3}[\rho(\mathbf{1}+\mathbf{4})+\bar{\rho}(\mathbf{1}+\mathbf{4})+\theta(\bar{\rho}\rho)^*(\mathbf{1}+\theta\mathbf{4})]=\sqrt{3}\rho\varphi^1_\rho\delta_{\rho\bar{\rho}}+(\rho+\bar{\rho}+\rho^*\bar{\rho}^*)\varphi^2_{\rho\bar{\rho}}, \tag{23}$$

where $\rho^{*2}=\theta\rho$ has been used.

From (23) we obtain the transpose of the rep matrix M of C,

$$\tilde{M}=\begin{pmatrix}\rho\delta_{\rho\bar{\rho}}, & \sqrt{3}\rho\delta_{\rho\bar{\rho}}\\ \sqrt{3}\theta\rho\delta_{\rho\bar{\rho}}, & \rho+\bar{\rho}+(\rho\bar{\rho})^*\end{pmatrix}. \tag{24}$$

The diagonalization of the matrix $\tilde{M}$ in (24) is carried out separately for two cases.

Case (i) $\rho\neq\bar{\rho}$

When $\rho\neq\rho$, or $\bar{\mu}\neq\mu$, M shrinks into one-dimensional and there is only one non-vanishing vector, $\varphi^2_{\mu\bar{\mu}}$, which is necessarily an eigenvector of the operator C with the eigenvalue $\lambda=M_{22}=\rho_\mu+\rho_{\bar{\mu}}+(\rho_\mu\rho_{\bar{\mu}})^*$,

$$\lambda=\begin{cases}0,\\ 2,\\ 2\epsilon, & \text{for}(\mu,\bar{\mu})\ \text{or}(\bar{\mu},\mu)=\begin{cases}(1,-1),(0,1),(0,-1),\\ (\frac{1}{2}-\frac{1}{2}),\\ (\frac{3}{2},\frac{1}{2}),\\ (\frac{3}{2},-\frac{1}{2}),\end{cases}\\ 2\epsilon^*,\end{cases} \tag{25a}$$

where

$$\epsilon=\rho_1=\exp\left(-\frac{2\pi i}{3}\right). \tag{25b}$$

Consulting Table I we know that the four eigenvalues $0,2,2\epsilon,2\epsilon^*$ of λ in (25a) correspond to the four irreps $F,E_{1/2},E_{3/2}$, and $E_{-3/2}$, respectively. Hence we have the basis vectors

$$\mathscr{P}_{\mu}^{(F)\overline{\mu}}=\varphi^{2}_{\mu\overline{\mu}},\quad \mu\neq\overline{\mu}=0,\pm1, \tag{26a}$$

$$\mathscr{P}_{\mu}^{(E_\nu)\overline{\mu}}=\varphi^{2}_{\mu\overline{\mu}},\quad (\mu,\overline{\mu})\ \text{or}(\overline{\mu},\mu)=\begin{cases}(\frac{1}{2},-\frac{1}{2}),\\ (\frac{3}{2},\frac{1}{2}),\\ (\frac{3}{2},-\frac{1}{2}),\end{cases}\quad \text{for}\quad \nu=\begin{cases}\frac{1}{2}\\ \frac{3}{2}\\ -\frac{3}{2}\end{cases}. \tag{26b}$$

Case (ii) $\overline{\rho}=\rho$

For $\rho=\overline{\rho}$, the matrix $\widetilde{M}$ is simplified as

$$\widetilde{M}=\rho_\mu\begin{pmatrix}1, & \sqrt{3}\\ \theta\sqrt{3}, & \theta+2\end{pmatrix}. \tag{27}$$

The eigenvector is

$$\mathscr{P}_{\mu}^{(\lambda)\mu}=N_{\lambda\mu}\left[\varphi^{1}_{\mu}-\frac{1}{\sqrt{3}}(1-\lambda\rho_{-\mu})\varphi^{2}_{\mu\mu}\right]. \tag{28}$$

For each μ, the matrix $\widetilde{M}$ in (27) has two eigenvalues, which are easily found as

$$\lambda=(0,4\rho_\mu),\quad \mu=0,\pm1,\quad \text{for s.v.}, \tag{29a}$$

$$\lambda=2\rho_{\mu\mp1/2}=\begin{cases}(2,2\epsilon)=(\lambda^{E_{1/2}},\lambda^{E_{3/2}}), & \mu=\frac{1}{2},\\ (2\epsilon^*,2)=(\lambda^{E_{-3/2}},\lambda^{E_{1/2}}), & \mu=-\frac{1}{2},\\ (2\epsilon,2\epsilon^*)=(\lambda^{E_{3/2}},\lambda^{E_{-3/2}}), & \mu=\frac{3}{2}.\end{cases} \tag{29b}$$

The eigenvalue $\lambda=4\rho_\mu$ corresponds to the one-dimensional irreps $A_\mu,\mu=0,\pm1$.

B. The $T\supset C_3$ ISO's

Substituting the eigenvalues (29) into (28), we can get the ISO's.

(1) The $T\supset C_3$ s.v. reps:

$$\begin{cases}\mathscr{P}^{(A_\mu)}=\frac{1}{2}(\varphi^{1}_{\mu}+\sqrt{3}\varphi^{2}_{\mu\mu}),\\ \mathscr{P}^{(F)\mu}_{\mu}=\frac{1}{2}(\sqrt{3}\varphi^{1}_{\mu}-\varphi^{2}_{\mu\mu}),\end{cases}\quad \mu=0,\pm1. \tag{30}$$

From (21), (30), and (26a) we get all the $T\supset C_3$ s.v. ISO's

$$\begin{cases} \mathscr{P}^{(A_\mu)} = \dfrac{1}{\sqrt{12}}[P^\mu + P^\mu C_{2z} P^\mu], \\ \mathscr{P}_\mu^{(F)\mu} = \frac{1}{2}[P^\mu - \frac{1}{3} P^\mu C_{2z} P^\mu], \quad \mu = 1,0,-1, \\ \mathscr{P}_\mu^{(F)\overline{\mu}} = \frac{1}{3} P^\mu C_{2z} P^{\overline{\mu}}, \quad \overline{\mu} \neq \mu. \end{cases} \tag{31}$$

Using the fact that

$$P^{\rho_\mu} C_{2z} P^{\rho_\mu} = P^{\rho_\mu}(C_{2x} + C_{2y} + C_{2z}), \tag{32}$$

$\mathscr{P}^{A_\mu}$ can be rewritten as

$$\mathscr{P}^{A_\mu} = \frac{1}{\sqrt{12}} P^{\rho_\mu} \wp^{++}, \tag{33}$$

with $\wp^{++} = 1 + C_{2x} + C_{2y} + C_{2z}$.

(2) The $T \supset C_3$ d.v. reps

For d.v. reps, (28) can be simplified as

$$\mathscr{P}_\mu^{(\lambda)\mu} = \frac{1}{\sqrt{2}}(\varphi_\mu^1 \mp i \varphi_{\mu\mu}^2), \lambda = 2\rho_{\mu \pm 1/2}, \mu = \tfrac{3}{2}, \pm \tfrac{1}{2}. \tag{34}$$

Combining it with (26b) we get the $T \supset C_3$ ISO's,

$$\begin{cases} \mathscr{P}_\mu^{(E_{1/2})\mu} = \dfrac{1}{\sqrt{6}}\left(P^\mu + \dfrac{(-1)^\mu}{\sqrt{3}} P^\mu C_{2z} P^\mu \right), \\ \mathscr{P}_\mu^{(E_{1/2})\overline{\mu}} = -\dfrac{i}{3} P^\mu C_{2z} P^{\overline{\mu}}, \overline{\mu} \neq \mu = \pm \tfrac{1}{2}, \end{cases} \tag{35a}$$

$$\begin{cases} \mathscr{P}_\mu^{(E_{3/2})\mu} = \dfrac{1}{\sqrt{6}}\left(P^\mu - \dfrac{(-1)^\mu}{\sqrt{3}} P^\mu C_{2z} P^\mu \right), \\ \mathscr{P}_\mu^{(E_{3/2})\overline{\mu}} = -\dfrac{i}{3} P^\mu C_{2z} P^{\overline{\mu}}, \overline{\mu} \neq \mu = \tfrac{3}{2}, \tfrac{1}{2}, \end{cases} \tag{35b}$$

where the phase factor $-i$ is chosen for convenience. The ISO's for $E_{-3/2}$ are obtained from the complex conjugation [for a definition see (18)] of the ISO's of $E_{3/2}$ as

$$\mathscr{P}_{-\mu}^{(E_{-3/2})-\overline{\mu}} = (\mathscr{P}_\mu^{(E_{3/2})\overline{\mu}})^*. \tag{36}$$

Notice that $C_{2z}^{-1} = \theta C_{2z}$ and the ISO's obey the condition

$$(\mathscr{P}_\mu^{(E_\nu)\overline{\mu}})^\dagger = \mathscr{P}_{\overline{\mu}}^{(E_\nu)\mu}.$$

C. The $T \supset C_3$ irreducible matrices

The operator C_{31}^{+} is diagonal and its representation matrix is readily known. The representation matrices of the coset generator C_{2z} are obtained in the following way: by comparing (31) and (35) with (8) we know that the coefficients of $P^{\mu}C_{2z}P^{\overline{\mu}}$ in (31) and (35) are equal to $\sqrt{h_\lambda/|G|}D^{\Gamma}_{\mu\overline{\mu}}(C_{2z})^*$. Consequently from (31) we immediately get $T \supset C_3$ s.v. matrix elements of C_{31}^{+} and C_{2z}, $D^{F}_{\mu\mu'}(C_{31}^{+}) = \delta_{\mu\mu'}\rho_\mu$, $D^{F}_{\mu\mu'}(C_{2z}) = \frac{2}{3} - \delta_{\mu\mu'}$, or in the matrix form

$$D^{F}(C_{31}^{+}) = \begin{pmatrix} \rho_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \rho_{-1} \end{pmatrix}, D^{F}(C_{2z}) = \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix}. \tag{37}$$

Similarly from (35) we get the $T \supset C_3$ d.v. irreducible matrices:

$$D^{E_{1/2}}(C_{31}^{+}) = \begin{pmatrix} \rho_{1/2} & 0 \\ 0 & \rho_{-1/2} \end{pmatrix}, D^{E_{3/2}}(C_{31}^{+}) = \begin{pmatrix} \rho_{3/2} & 0 \\ 0 & \rho_{1/2} \end{pmatrix}, \tag{38a}$$

$$D^{E_{1/2}}(C_{2z}) = D^{E_{3/2}}(C_{2z}) = i\begin{pmatrix} -\sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\ \sqrt{\frac{2}{3}} & \sqrt{\frac{1}{3}} \end{pmatrix}. \tag{38b}$$

The irreducible matrices of the irreps $E_{-3/2}$ and $E_{3/2}$ are complex conjugate to each other,

$$D^{E_{-3/2}}(R_a) = D^{E_{3/2}}(R_a)^*.$$

The matrices of other elements are easily obtained by

$$D^{(\lambda)}_{\mu\overline{\mu}}(C_{31}^{j}C_{2z}C_{31}^{k}) = \rho_\mu^{j} D^{(\lambda)}_{\mu\overline{\mu}}(C_{2z})\rho_{\overline{\mu}}^{k} = \epsilon^{\mu j + \overline{\mu k}} D^{(\lambda)}_{\mu\overline{\mu}}(C_{2z}), \tag{39}$$

where $C_{31}^{j} \equiv (C_{31}^{+})^{j}$.

Using the group table of $T^{\dagger}$ we can obtain the pair of index (j,k) for each element of $T^{\dagger}$, as shown below:

(j,k)	(00)	(01)	(02)	(10)	(11)	(12)	(20)	(21)	(22)
$C_{31}^{j}C_{2z}C_{31}^{k}$	C_{2z}	$\tilde{C}_{33}^{+}$	$\tilde{C}_{34}^{-}$	$\tilde{C}_{34}^{+}$	$\tilde{C}_{32}^{-}$	$\tilde{C}_{2x}$	$\tilde{C}_{33}^{-}$	$\tilde{C}_{2y}$	C_{32}^{+}

Equation (39) shows that the irreducible matrix elements of all elements in the same double coset differ from that of the coset generator only by a phase factor $\epsilon^{\mu j + \overline{\mu k}}$. Equations (37)–(39) are the algebraic expressions for the irreducible matrices of $T^{\dagger}$.

D. The $T \supset C_3$ generalized projection operators

To obtain the SAF's, the normalization of the ISO's is unimportant and for each irrep we only need one set of the ISO's. We collect these chosen unnormalized ISO's below, and call them the generalized projection operators.

1. $T \supset C_3$ s.v. generalized projection operators

Taking $\overline{\mu}=0$ for the irrep F from (31) we have

$$\begin{cases} \mathbf{P}^{A}{}_{\mu}=P^{\mu}+P^{\mu}C_{2z}P^{\mu}, \\ \\ \mathbf{P}^{F}_{\mu}=3P^{\mu}\delta_{\mu 0}-(-1)^{\mu}P^{\mu}C_{2z}P^{\overline{\mu=0}}, \end{cases} \qquad \mu=1,0,-1. \tag{40a}$$

2. $T \supset C_3$ d.v. generalized projection operators

For the irrep E_{μ}, there are two sets of the ISO's distinguished by $\overline{\mu}=(\frac{1}{2},-\frac{1}{2}),(\frac{1}{2},\frac{3}{2})$, and $(-\frac{1}{2},-\frac{3}{2})$ for the irreps $E_{1/2}, E_{3/2}$, and $E_{-3/2}$, respectively. Taking $\overline{\mu}=\frac{1}{2},\frac{3}{2}$, and $-\frac{3}{2}$ for the irreps $E_{1/2}$, $E_{3/2}$, and $E_{-3/2}$, respectively, from (35) we have:

$$\mathbf{P}^{E_{\nu}}_{\mu}=\sqrt{3}P^{\mu}\delta_{\mu\nu}+(-1)^{\mu-\frac{1}{2}+|\nu|}P^{\mu}C_{2z}P^{\nu},$$

$$\nu=\tfrac{1}{2},\ \tfrac{3}{2},\ \mu=\nu,\ \nu-1;\nu=-\tfrac{3}{2},\ \mu=\nu,\ \nu+1. \tag{40b}$$

Equation (40) is the algebraic expression for the generalized projection operator, and they are all we need for constructing the s.v. and d.v. SAF's of $T^{\dagger}$. Compared to the double-coset factored form (8), it represents a great simplification in that it is merely a function of only two quantum numbers ν and μ instead of all the irreducible matrix elements of the coset generators.

E. The $T \supset C_3$ SAF's

Up to now the coordinate system has not been specified explicitly. For obtaining the SAF's we need to introduce the coordinate system. The usual coordinate system for the group T is shown in Fig. 1, which is suitable for the group chain $T \supset D_2$. For the group chain $T \supset C_3$, we choose the coordinate system $x'y'z'$ as shown in Fig. 2, with z' as the axis of C_{31}^{+}. The coordinate system $x'y'z'$ is obtained from the system xyz in Fig. 1 by a rotation $R(\omega)$,

$$xyz \overset{R(\omega)}{\rightarrow} x'y'z', R(\omega)\equiv R\left(0,\ \cos^{-1}\frac{1}{\sqrt{3}},\ \frac{\pi}{4}\right). \tag{41}$$

Notice that $\cos^{-1}(1/\sqrt{3})$ is the angle between the z axis and the axis of C_{31}^{+} in Fig. 1. Using (11) and Table II we obtain the Euler angles in the system $x'y'z'$ for the coset generator C_{2z}, $C_{2z}=R(\pi,\cos^{-1}(-\frac{1}{3}),0)$.

The projection operator method with the projection operator in its original form (4) is quite tedious, but it becomes very efficient with the algebraic form in (40). For example, applying (40a) to $|jm\rangle$ and using

$$P^{\mu}|j\overline{m}\rangle=3\,\delta_{\mu,\overline{m}(\mathrm{mod}\ 3)}|j\overline{m}\rangle, C_{2z}|j\overline{m}\rangle=\sum_{m}\ e^{-im\pi}d^{j}_{m\overline{m}}(\beta_0)|jm\rangle, \beta_0=\cos^{-1}(-\tfrac{1}{3}), \tag{42}$$

we immediately obtain the $T \supset C_3$ s.v. SAF:

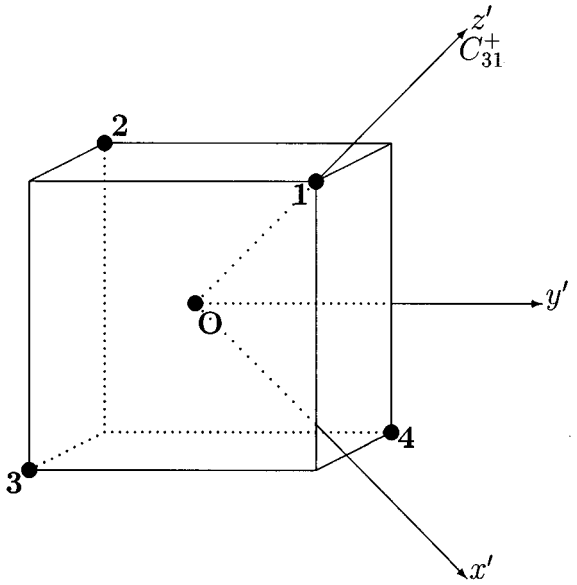

FIG. 2. The coordinate system $x'y'z'$ for $T\supset C_3$, $xyz \overset{R(\omega)}{\rightarrow} x'y'z'$, $R(\omega)\equiv R(0,\cos^{-1}(1/\sqrt{3}),\ \pi/4)$.

1. One-dimensional s.v. irreps

$$\psi^{(A_\mu)\overline{m}}=Y_{l\overline{m}}+3\sum_{m=\mu}^{(3)}(-1)^m d^l_{m\overline{m}}(\beta_0)Y_{lm},$$
$$\overline{m}=\mu(\ \mathrm{mod}\ 3),\quad \mu=0,\pm 1, \tag{43a}$$

where the normalization factor is suppressed for simplicity and the summation symbol stands for

$$\sum_{m=\mu}^{(n)}\equiv\sum_{m=\mu,\mu+n,\mu-n,\ldots}.$$

2. Three-dimensional s.v. irreps

$$\psi_\mu^{(F)\overline{m}}=Y_{l\overline{m}}\delta_{\mu 0}-(-1)^\mu\sum_{m=\mu}^{(3)}(-1)^m d^l_{m\overline{m}}(\beta_0)Y_{lm},$$
$$\overline{m}=0(\mathrm{mod}\ 3),\mu=0,\pm 1. \tag{43b}$$

It is easily found that

$$\begin{cases}\langle\psi^{(A_\mu)\overline{m}}|lm\rangle=(-1)^{m+\overline{m}}\langle\psi^{(A_{-\mu})-\overline{m}}|l-m\rangle,\\ \langle\psi_\mu^{(F)\overline{m}}|lm\rangle=(-1)^{m+\overline{m}}\langle\psi_{-\mu}^{(F)-\overline{m}}|l-m\rangle,\end{cases}\qquad \mu=0,\pm 1, \tag{44a}$$

where the SAF's ψ_μ^Γ are, of course, supposed to have been normalized.

Therefore for multiplicity free case the SAF's in (43) for A_0 and F have the symmetries

$$\langle \psi^{A_0}|lm\rangle=(-1)^m\langle \psi^{A_0}|l-m\rangle,$$
$$\langle \psi^{F}_{\mu}|lm\rangle=(-1)^m\langle \psi^{F}_{-\mu}|l-m\rangle, \quad \mu=0,\pm 1. \tag{44b}$$

However, for the non-multiplicity free case, the above symmetries no longer hold. Since the intrinsic quantum number is irrelevant, we can use the following as the new SAF's:

$$\psi'^{(\Gamma)\overline{m}}_{\mu}=\frac{1}{\sqrt{2}}(\psi^{(\Gamma)\overline{m}}_{\mu}\pm\psi^{(\Gamma)-\overline{m}}_{\mu}), \quad \Gamma=A_0,F, \tag{45}$$

where the ''+'' and ''−'' signs are taken for $\overline{m}>0$ and $\overline{m}<0$, respectively. Using (43) we get the new SAF's for $T\supset C_3$,

$$\psi'^{(A_0)\overline{m}}=Y_{l\overline{m}}\pm Y_{l-\overline{m}}+3\sum_{m=0}^{(3)}(-1)^m[d^l_{m\overline{m}}(\beta_0)\pm d^l_{m-\overline{m}}(\beta_0)]Y_{lm}, \tag{46a}$$

$$\psi'^{(F)\overline{m}}_{\mu}=(Y_{l\overline{m}}\pm Y_{l-\overline{m}})\delta_{\mu,0}-(-1)^{\mu}\sum_{m=\mu}^{(3)}(-)^m[d^l_{m\overline{m}}(\beta_0)\pm d^l_{m-\overline{m}}(\beta_0)]Y_{lm}, \tag{46b}$$

where $\overline{m}=0 \pmod 3$, $\mu=0,\pm 1$. The SAF's for $A_{\pm 1}$ are still given by (43a), i.e.,

$$\psi'^{(A_\mu)\overline{m}}=Y_{l\overline{m}}+3\sum_{m=\mu}^{(3)}(-1)^m d^l_{m\overline{m}}(\beta_0)Y_{lm}, \overline{m}=\mu(\mathrm{mod}\ 3), \mu=\pm 1. \tag{46c}$$

The SAF's ψ' have the desired symmetries:

$$\langle \psi'^{(A_0)\overline{m}}|lm\rangle=(-1)^{m+\overline{m}}\,\mathrm{sign}(\overline{m})\langle \psi'^{(A_0)\overline{m}}|l-m\rangle,$$
$$\langle \psi'^{(A_\mu)\overline{m}}|lm\rangle=(-1)^{m+\overline{m}}\langle \psi'^{(A_{-\mu})-\overline{m}}|l-m\rangle, \mu=\pm 1, \tag{47}$$
$$\langle \psi'^{(F)\overline{m}}_{\mu}|lm\rangle=(-1)^{m+\overline{m}}\,\mathrm{sign}(\overline{m})\langle \psi'^{(F)\overline{m}}_{-\mu}|l-m\rangle, \mu=0,\pm 1.$$

Notice that the index $\overline{m}$ in (46) serves as the multiplicity label. In general the SAF's $\psi^{(\Gamma)\overline{m}}_{\mu}$ with different $\overline{m}$ may be neither orthogonal nor linearly independent. With these algebraic expressions, it is easy to find the exact numerical expression of the SAF's $\psi^{(\Gamma)\overline{m}}_{\mu}$ for any l with the help of some software, say the Maple software. The multiplicity or frequency τ^l_Γ for the occurrence of the irrep Γ of T in the subduced rep $l\downarrow\Gamma$ is calculated first and used as a control parameter for the calculation: If the multiplicity τ^l_Γ is equal to one, then for obtaining the SAF's from the expressions (46) one only needs to let $\overline{m}$ take the lowest possible value, i.e., zero for the irrep A_0 or F, and ± 1 and for the irrep $A_{\pm 1}$. Otherwise one should let $\overline{m}$ take τ^l_Γ different values: $0,3,-3,6,-6,...$, for A_0 or F, $1,-2,4,-5,\ldots$, for A_1 and $-1,2,-4,5,\ldots$, for A_{-1}. The linear independence of the SAF's $\psi^{(\Gamma)\overline{m}}_{\mu}$ with different $\overline{m}$ is checked in each step. No orthogonal procedure is included for keeping the simplicity of the algebraic expression of the SAF's. The $T\supset C_3$ SAF's for $j=1$–3, and a part for $j=6,8$ are listed in Table IV.

Applying (40b) to $|jm\rangle$ we obtain the $T\supset C_3$ d.v. SAF:

TABLE IV. The $T \supset C_3$ s.v. SAF's. The superscript is the multiplicity label. For $l=6$ and 8 only the SAF's with multiplicity are listed.

$\overline{m}=0$ $|F\,-1\rangle = -|1\,-1\rangle$

$|F0\rangle = |10\rangle$

$|F1\rangle = |11\rangle$

$\overline{m}=-1 \quad |A_{-1}\rangle = \frac{\sqrt{2\cdot 3}}{3}|2\,-1\rangle - \frac{\sqrt{3}}{3}|22\rangle$

$\overline{m}=1 \quad |A_1\rangle = \frac{\sqrt{3}}{3}|2\,-2\rangle + \frac{\sqrt{2\cdot 3}}{3}|21\rangle$

$\overline{m}=0 \quad |F\,-1\rangle = \frac{\sqrt{3}}{3}|2\,-1\rangle + \frac{\sqrt{2\cdot 3}}{3}|22\rangle$

$|F0\rangle = |20\rangle$

$|F1\rangle = \frac{\sqrt{2\cdot 3}}{3}|2\,-2\rangle - \frac{\sqrt{3}}{3}|21\rangle$

$\overline{m}=0$

$|A_0\rangle = -\frac{\sqrt{2}}{3}|3\,-3\rangle + \frac{\sqrt{5}}{3}|30\rangle + \frac{\sqrt{2}}{3}|33\rangle$

$\overline{m}=0$

$|F\,-1\rangle^1 = \frac{\sqrt{2\cdot 3}}{6}|3\,-1\rangle - \frac{\sqrt{2\cdot 3\cdot 5}}{6}|32\rangle$

$|F0\rangle^1 = \frac{\sqrt{2\cdot 5}}{6}|3\,-3\rangle + \frac{2}{3}|30\rangle - \frac{\sqrt{2\cdot 5}}{6}|33\rangle$

$|F1\rangle^1 = -\frac{\sqrt{2\cdot 3\cdot 5}}{6}|3\,-2\rangle - \frac{\sqrt{2\cdot 3}}{6}|31\rangle$

$\overline{m}=3$

$|F\,-1\rangle^2 = -\frac{2\sqrt{3\cdot 5\cdot 7}}{21}|3\,-1\rangle + \frac{\sqrt{3\cdot 7}}{21}|32\rangle$

$|F0\rangle^2 = \frac{2\sqrt{7}}{21}|3\,-3\rangle - \frac{\sqrt{2\cdot 5\cdot 7}}{21}|30\rangle + \frac{\sqrt{7}}{3}|33\rangle$

$|F1\rangle^2 = \frac{4\sqrt{3\cdot 7}}{21}|3\,-2\rangle - \frac{\sqrt{3\cdot 5\cdot 7}}{21}|31\rangle$

$\overline{m}=0$

$|A_0\rangle^1 = \frac{\sqrt{2\cdot 3\cdot 7\cdot 11}}{54}|6\,-6\rangle - \frac{\sqrt{3\cdot 5\cdot 7}}{27}|6\,-3\rangle + \frac{4\sqrt{2}}{9}|60\rangle + \frac{\sqrt{3\cdot 5\cdot 7}}{27}|63\rangle + \frac{\sqrt{2\cdot 3\cdot 7\cdot 11}}{54}|66\rangle$

$\overline{m}=3$

$|A_0\rangle^2 = \frac{\sqrt{2\cdot 3\cdot 5}}{18}|6\,-6\rangle + \frac{\sqrt{3\cdot 11}}{9}|6\,-3\rangle + \frac{\sqrt{3\cdot 11}}{9}|63\rangle - \frac{\sqrt{2\cdot 3\cdot 5}}{18}|66\rangle$

$\overline{m}=-2$

$|A_1\rangle^1 = \frac{\sqrt{2\cdot 3\cdot 7\cdot 11\cdot 13}}{324}|8\,-8\rangle + \frac{\sqrt{3\cdot 5\cdot 11\cdot 13}}{81}|8\,-5\rangle + \frac{16\sqrt{3}}{81}|8\,-2\rangle - \frac{\sqrt{3\cdot 5\cdot 7}}{162}|81\rangle$
$+ \frac{11\sqrt{2\cdot 3\cdot 5\cdot 11}}{324}|84\rangle - \frac{\sqrt{3\cdot 7\cdot 11\cdot 13}}{162}|87\rangle$

$\overline{m}=1$

$|A_1\rangle^2 = \frac{16\sqrt{3\cdot 5\cdot 11\cdot 13}}{81\sqrt{487}}|8\,-8\rangle - \frac{8\sqrt{2\cdot 3\cdot 7\cdot 11\cdot 13}}{81\sqrt{487}}|8\,-5\rangle - \frac{4\sqrt{2\cdot 3\cdot 5\cdot 7}}{81\sqrt{487}}|8\,-2\rangle + \frac{487\sqrt{2\cdot 3}}{81\sqrt{487}}|81\rangle$
$+ \frac{37\sqrt{3\cdot 7\cdot 11}}{81\sqrt{487}}|84\rangle + \frac{11\sqrt{2\cdot 3\cdot 5\cdot 11\cdot 13}}{81\sqrt{487}}|87\rangle$

TABLE IV. *(Continued).*

$\overline{m}=-1$

$$|A_{-1}\rangle^1=\frac{11\sqrt{2\cdot3\cdot5\cdot11\cdot13}}{81\sqrt{487}}|8\ -7\rangle-\frac{37\sqrt{3\cdot7\cdot11}}{81\sqrt{487}}|8\ -4\rangle+\frac{487\sqrt{2\cdot3}}{81\sqrt{487}}|8\ -1\rangle+\frac{4\sqrt{2\cdot3\cdot5\cdot7}}{81\sqrt{487}}|82\rangle$$
$$-\frac{8\sqrt{2\cdot3\cdot7\cdot11\cdot13}}{81\sqrt{487}}|85\rangle-\frac{16\sqrt{3\cdot5\cdot11\cdot13}}{81\sqrt{487}}|88\rangle$$

$\overline{m}=2$

$$|A_{-1}\rangle^2=\frac{\sqrt{3\cdot7\cdot11\cdot13}}{162}|8\ -7\rangle+\frac{11\sqrt{2\cdot3\cdot5\cdot11}}{324}|8\ -4\rangle+\frac{\sqrt{3\cdot5\cdot7}}{162}|8\ -1\rangle+\frac{16\sqrt{3}}{81}|82\rangle$$
$$-\frac{\sqrt{3\cdot5\cdot11\cdot13}}{81}|85\rangle+\frac{\sqrt{2\cdot3\cdot7\cdot11\cdot13}}{324}|88\rangle$$

3. Two-dimensional d.v. irreps

$$\psi_{\mu}^{(E_\nu)\overline{m}}=\frac{1}{\sqrt{3}}\delta_{\mu\nu}|j\overline{m}\rangle+(-1)^{|\nu|-1/2}\sum_{m=\mu}^{(3)}(-1)^{m-\mu}d^{j}_{m\overline{m}}(\beta_0)|jm\rangle,\overline{m}=\nu(\mathrm{mod}\ 3),$$
$$\nu=\tfrac{1}{2},\ \tfrac{3}{2},\ \mu=\nu,\ \nu-1;\nu=-\tfrac{3}{2},\ \mu=\nu,\ \nu+1. \tag{48}$$

Using (12b) it can be shown that the SAF's for $E_{\pm3/2}$ have the symmetry

$$\langle\psi_{\mu}^{(E_{3/2})\overline{m}}|\,jm\rangle=(-1)^{m-\overline{m}}\langle\psi_{-\mu}^{(E_{-3/2})-\overline{m}}|\,j-m\rangle,\mu=\tfrac{1}{2},\tfrac{3}{2}, \tag{49}$$

while from the numerical values of (48) it is known that the SAF's for $E_{1/2}$ in the multiplicity free case have the symmetry

$$\langle\psi_{1/2}^{(E_{1/2})}\,|\,jm\rangle=\pm\langle\psi_{-1/2}^{(E_{1/2})}\,|j-m\rangle,$$

however for the non-multiplicity free case they do not have such a symmetry. The lack of symmetry for the SAF $\psi^{(E_{1/2})}$ is due to the lack of symmetry of the ISO's in (35a). Since the intrinsic quantum number $\overline{\mu}$ is irrelevant for obtaining the SAF's, we can choose the following as the generalized projection operator:

$$\mathbf{P}'^{(E_{1/2})}_{\mu}=\mathscr{P}^{(E_{1/2})(1/2)}_{\mu}+\mathscr{P}^{(E_{1/2})-1/2}_{\mu}. \tag{50a}$$

Applying (50a) to the trial state $(|j\overline{m}\rangle+|j-\overline{m}\rangle)$, we obtain the new SAF's for the irrep $E_{1/2}$,

$$\psi'^{(E_{1/2})\overline{m}}_{1/2}=|j\overline{m}\rangle+\sqrt{3}\sum_{m=1/2}^{(3)}(-)^{m-1/2}[d^{j}_{m\overline{m}}(\beta_0)-\sqrt{2}d^{j}_{m-\overline{m}}(\beta_0)]|jm\rangle,$$
$$\psi'^{(E_{1/2})\overline{m}}_{-1/2}=|j-\overline{m}\rangle+\sqrt{3}\sum_{m=-1/2}^{(3)}(-)^{m+1/2}[d^{j}_{m-\overline{m}}(\beta_0)+\sqrt{2}d^{j}_{m\overline{m}}(\beta_0)]|jm\rangle, \tag{50b}$$

where $\overline{m}=\frac{1}{2}$ (mod 3). The SAF's in (50b) have the desired symmetry:

$$\langle \psi'^{(E_{1/2})\overline{m}}_{\mu} | jm\rangle = (-)^{m-\overline{m}} \langle \psi'^{(E_{1/2})\overline{m}}_{-\mu} | j-m\rangle. \tag{50c}$$

Hence we see that the symmetries of the SAF's for the non-multiplicity free case depend on the choice of the intrinsic quantum number and the trial state to be projected.

Using (48) for $E_{\pm 3/2}$ and (50b) for $E_{1/2}$, we can find the exact numerical expression of the SAF's $\psi^{(\Gamma)\overline{m}}_{\mu}$ for any j. The SAF's for $j=\frac{1}{2}-\frac{9}{2}$ are listed in Table V. It should be mentioned that the symmetrization of the SAF's as indicated in (47) and (50c) leads to simpler numerical results.

V. DISCUSSIONS

A. The left-, right-, and double-induced representations

There are three possible ways to construct an induced representation of a group G from the ISO's P^{ρ} of its subgroup $G(s)$. We first deompose G into left-, right-, and double-cosets as

$$\begin{aligned} G &= (e + \not{c}_2 + \not{c}_3 + \cdots) G(s) \\ &= G(s)(e + r_2 + r_3 + \cdots) \\ &= G(s)(e + k_2 + k_3 + \cdots) G(s). \end{aligned} \tag{51}$$

Then the three sets of the basis vectors $\{\not{c}_i P^{\rho}\}$, $\{P^{\rho} r_i\}$, and $\{P^{\rho} k_i P^{\overline{\rho}}\}$ form the left-, right-, and double-induced representations of G, respectively. The conventionally called induced representation[3] is the left-induced representation. If $G(s)$ is an invariant subgroup of G, then the double-induced representation is reduced to the left- or the right-induced representation, which are equivalent to one another. Therefore, the double-induced representation is a generalization of the conventional induced representation.

B. The induction method for finding irreps

Since the induction technique is the pillar for finding irreps of point groups in traditionally group rep theory, it is worth showing the difference between the technique proposed in this paper and the usual induction technique. To show the difference, in this subsection we first sketch the induction method for finding irreps of a group G from irreps of its maximal invariant subgroup.[3] Recall that an invariant subgroup of G is called maximal if it is not a subgroup of another invariant subgroup of G. Assume that H is a maximal invariant subgroup of G and C_q is a cyclic group of prime order q,

$$G = H \wedge C_q, C_q = \{s^n : n = 0,1,\ldots,q-1\}.$$

We can factorize G into left cosets in the form of

$$G = (e + s + s^2 + \cdots + s^{q-1}) H.$$

Suppose that $D^{(\mu,0)} \equiv D^{(\mu)}$ is an irrep of H of dimension $|\mu|$ with basis vectors $\psi^{(\mu)}_{0m}$,

$$h \psi^{(\mu)}_{0m} = \sum_{m'} D^{(\mu)}_{m'm}(h) \psi^{(\mu)}_{0m'}.$$

Define

$$\psi^{(\mu)}_{nm} = s^n \psi^{(\mu)}_{0m}, \quad n = 0,1,\ldots,q-1,$$

which span the induced rep of G. For given n, the basis vectors $\psi^{(\mu)}_{nm}$ carry the representation $D^{(\mu,n)}$ of H,

TABLE V. The $T \supset C_3$d.v. SAF's. The superscript is the multiplicity label. For $j=\frac{9}{2}$ only the SAf's with multiplicity are listed.

$\overline{m}=\frac{1}{2}$

$|E\,\frac{1}{2}\ \ \frac{1}{2}\rangle=|\frac{1}{2}\,\frac{1}{2}\rangle$

$|E\,\frac{1}{2}\ -\frac{1}{2}\rangle=|\frac{1}{2}\ -\frac{1}{2}\rangle$

$\overline{m}=\frac{3}{2}$

$|E\,\frac{3}{2}\ \ \frac{1}{2}\rangle=-|\frac{3}{2}\ \frac{1}{2}\rangle$

$|E\,\frac{3}{2}\ \ \frac{3}{2}\rangle=\frac{\sqrt{2\cdot 3}}{3}|\frac{3}{2}\ -\frac{3}{2}\rangle+\frac{\sqrt{3}}{3}|\frac{3}{2}\,\frac{3}{2}\rangle$

$\overline{m}=-\frac{5}{2}$

$|E\,\frac{1}{2}\ \ \frac{1}{2}\rangle=-\frac{2}{3}|\frac{5}{2}\ -\frac{5}{2}\rangle+\frac{\sqrt{5}}{3}\ |\frac{5}{2}\,\frac{1}{2}\rangle$

$|E\,\frac{1}{2}\ -\frac{1}{2}\rangle=-\frac{\sqrt{5}}{3}\ |\frac{5}{2}-\frac{1}{2}\rangle-\frac{2}{3}\ |\frac{5}{2}\,\frac{5}{2}\rangle$

$\overline{m}=\frac{3}{2}$

$|E\,\frac{3}{2}\ \ \frac{1}{2}\rangle=\frac{\sqrt{5}}{3}|\frac{5}{2}-\frac{5}{2}\rangle+\frac{2}{3}\ |\frac{5}{2}\,\frac{1}{2}\rangle$

$|E\,\frac{3}{2}\ \ \frac{3}{2}\rangle=\frac{1}{3}\ |\frac{5}{2}-\frac{3}{2}\rangle+\frac{2\sqrt{2}}{3}\ |\frac{5}{2}\,\frac{3}{2}\rangle$

$\overline{m}=-\frac{3}{2}$

$|E_{-\frac{3}{2}}\ -\frac{1}{2}\rangle=|\frac{3}{2}-\frac{1}{2}\rangle$

$|E_{-\frac{3}{2}}\ -\frac{3}{2}\rangle=\frac{\sqrt{3}}{3}|\frac{3}{2}\ -\frac{3}{2}\rangle-\frac{\sqrt{2\cdot 3}}{3}|\frac{3}{2}\,\frac{3}{2}\rangle$

$\overline{m}=-\frac{3}{2}$

$|E_{-\frac{3}{2}}\ -\frac{1}{2}\rangle=-\frac{2}{3}\,|\frac{5}{2}-\frac{1}{2}\rangle+\frac{\sqrt{5}}{3}\ |\frac{5}{2}\,\frac{5}{2}\rangle$

$|E_{-\frac{3}{2}}\ -\frac{3}{2}\rangle=\frac{2\sqrt{2}}{3}\ |\frac{5}{2}-\frac{3}{2}\rangle-\frac{1}{3}\,|\frac{5}{2}\,\frac{3}{2}\rangle$

$\overline{m}=\frac{1}{2}$

$|E\,\frac{1}{2}\ \ \frac{1}{2}\rangle^1=-\frac{\sqrt{2}}{6}|\frac{7}{2}-\frac{5}{2}\rangle+\frac{\sqrt{5}}{3}|\frac{7}{2}\,\frac{1}{2}\rangle+\frac{\sqrt{2\cdot 7}}{6}|\frac{7}{2}\,\frac{7}{2}\rangle$

$|E\,\frac{1}{2}\ -\frac{1}{2}\rangle^1=-\frac{\sqrt{2\cdot 7}}{6}|\frac{7}{2}-\frac{7}{2}\rangle+\frac{\sqrt{5}}{3}|\frac{7}{2}-\frac{1}{2}\rangle+\frac{\sqrt{2}}{6}|\frac{7}{2}\,\frac{5}{2}\rangle$

$\overline{m}=\frac{7}{2}$

$|E\,\frac{1}{2}\ \ \frac{1}{2}\rangle^2=-\frac{\sqrt{2\cdot 3\cdot 5\cdot 7}}{18}|\frac{7}{2}-\frac{5}{2}\rangle-\frac{\sqrt{3\cdot 7}}{9}|\frac{7}{2}\,\frac{1}{2}\rangle+\frac{\sqrt{2\cdot 3\cdot 5}}{18}|\frac{7}{2}\,\frac{7}{2}\rangle$

$|E\,\frac{1}{2}\ -\frac{1}{2}\rangle^2=\frac{\sqrt{2\cdot 3\cdot 5}}{18}|\frac{7}{2}-\frac{7}{2}\rangle+\frac{\sqrt{3\cdot 7}}{9}|\frac{7}{2}-\frac{1}{2}\rangle-\frac{\sqrt{2\cdot 3\cdot 5\cdot 7}}{18}|\frac{7}{2}\,\frac{5}{2}\rangle$

$\overline{m}=\frac{3}{2}$

$|E\,\frac{3}{2}\ \ \frac{1}{2}\rangle=\frac{2\sqrt{2\cdot 3}}{9}|\frac{7}{2}-\frac{5}{2}\rangle-\frac{\sqrt{3\cdot 5}}{9}|\frac{7}{2}\,\frac{1}{2}\rangle+\frac{\sqrt{2\cdot 3\cdot 7}}{9}|\frac{7}{2}\,\frac{7}{2}\rangle$

$|E\,\frac{3}{2}\ \ \frac{3}{2}\rangle=-\frac{2\sqrt{2}}{3}|\frac{7}{2}\ -\frac{3}{2}\rangle+\frac{1}{3}|\frac{7}{2}\,\frac{3}{2}\rangle$

$\overline{m}=-\frac{3}{2}$

$|E_{-\frac{3}{2}}\ -\frac{1}{2}\rangle=\frac{\sqrt{2\cdot 3\cdot 7}}{9}|\frac{7}{2}-\frac{7}{2}\rangle+\frac{\sqrt{3\cdot 5}}{9}|\frac{7}{2}-\frac{1}{2}\rangle+\frac{2\sqrt{2\cdot 3}}{9}|\frac{7}{2}\,\frac{5}{2}\rangle$

$|E_{-\frac{3}{2}}\ -\frac{3}{2}\rangle=\frac{1}{3}|\frac{7}{2}-\frac{3}{2}\rangle+\frac{2\sqrt{2}}{3}|\frac{7}{2}\,\frac{3}{2}\rangle$

$\overline{m}=-\frac{3}{2}$

$|E\,\frac{3}{2}\ \ \frac{1}{2}\rangle^1=\frac{\sqrt{7\cdot 11}}{11}|\frac{9}{2}-\frac{5}{2}\rangle-\frac{2\sqrt{11}}{11}|\frac{9}{2}\,\frac{1}{2}\rangle$

$|E\,\frac{3}{2}\ \ \frac{3}{2}\rangle^1=-\frac{2\sqrt{2\cdot 7\cdot 11}}{99}|\frac{9}{2}-\frac{9}{2}\rangle+\frac{\sqrt{3\cdot 11}}{9}|\frac{9}{2}-\frac{3}{2}\rangle-\frac{2\sqrt{2\cdot 3\cdot 11}}{99}|\frac{9}{2}\,\frac{3}{2}\rangle+\frac{8\sqrt{7\cdot 11}}{99}|\frac{9}{2}\,\frac{9}{2}\rangle$

$\overline{m}=\frac{3}{2}$

$|E\,\frac{3}{2}\ \ \frac{1}{2}\rangle^2=-\frac{\sqrt{2\cdot 7}}{6}|\frac{9}{2}-\frac{5}{2}\rangle-\frac{\sqrt{2}}{3}|\frac{9}{2}\,\frac{1}{2}\rangle-\frac{\sqrt{2\cdot 7}}{6}|\frac{9}{2}\,\frac{7}{2}\rangle$

$|E\,\frac{3}{2}\ \ \frac{3}{2}\rangle^2=-\frac{2\sqrt{7}}{9}|\frac{9}{2}-\frac{9}{2}\rangle-\frac{\sqrt{2\cdot 3}}{18}|\frac{9}{2}-\frac{3}{2}\rangle+\frac{4\sqrt{3}}{9}|\frac{9}{2}\,\frac{3}{2}\rangle-\frac{\sqrt{2\cdot 7}}{18}|\frac{9}{2}\,\frac{9}{2}\rangle$

$\overline{m}=-\frac{3}{2}$

$|E_{-\frac{3}{2}}\ -\frac{1}{2}\rangle^2=-\frac{\sqrt{2\cdot 7}}{6}|\frac{9}{2}-\frac{7}{2}\rangle+\frac{\sqrt{2}}{3}|\frac{9}{2}-\frac{1}{2}\rangle-\frac{\sqrt{2\cdot 7}}{6}|\frac{9}{2}\,\frac{5}{2}\rangle$

$|E_{-\frac{3}{2}}\ -\frac{3}{2}\rangle^2=\frac{\sqrt{2\cdot 7}}{18}|\frac{9}{2}-\frac{9}{2}\rangle+\frac{4\sqrt{3}}{9}|\frac{9}{2}-\frac{3}{2}\rangle+\frac{\sqrt{2\cdot 3}}{18}|\frac{9}{2}\,\frac{3}{2}\rangle-\frac{2\sqrt{7}}{9}|\frac{9}{2}\,\frac{9}{2}\rangle$

TABLE V. *(Continued).*

$\overline{m}=\frac{3}{2}$

$$|E_{-\frac{3}{2}}\ -\tfrac{1}{2}\rangle^1=-\frac{2\sqrt{11}}{11}|\tfrac{9}{2}-\tfrac{1}{2}\rangle-\frac{\sqrt{7\cdot 11}}{11}|\tfrac{9}{2}\tfrac{5}{2}\rangle$$

$$|E_{-\frac{3}{2}}\ -\tfrac{3}{2}\rangle^1=\frac{8\sqrt{7\cdot 11}}{99}|\tfrac{9}{2}-\tfrac{9}{2}\rangle+\frac{2\sqrt{2\cdot 3\cdot 11}}{99}|\tfrac{9}{2}-\tfrac{3}{2}\rangle+\frac{\sqrt{3\cdot 11}}{9}|\tfrac{9}{2}\tfrac{3}{2}\rangle+\frac{2\sqrt{2\cdot 7\cdot 11}}{99}|\tfrac{9}{2}\tfrac{9}{2}\rangle$$

$$h\psi_{nm}^{(\mu)}=hs_n\psi_{0m}^{(\mu)}=s_nh'\psi_{0m}^{(\mu)}=s^n\sum_{m'}D_{m'm}^{(\mu)}(h')\psi_{0m'}^{(\mu)}=\sum_{m'}D_{m'm}^{(\mu,n)}(h)\psi_{nm'}^{(\mu)}, \tag{52}$$

with

$$D^{(\mu,n)}(h)=D^{(\mu)}(s^{-n}hs^n),\quad n=0,1,\ldots,q-1.$$

Case (i)

If the representations $D^{(\mu,n)}(H),n=0,1,\ldots,q-1$, are mutually inequivalent, then the induced rep $\mu\uparrow G$ is irreducible. The typical example is $D^{A_{\pm 1}}(T)\uparrow O=D^E$.

Case (ii)

If the representations $D^{(\mu,n)}(H),n=0,1,\ldots,q-1$, are mutually equivalent, then the induced rep $\mu\uparrow G$ is reducible. It can be reduced into q mutually inequivalent irreps $\mathscr{D}^{(\mu_n)},n=0,1,\ldots(q-1)$, of G each of dimension $|\mu|$, with the irreducible matrices,

$$\mathscr{D}^{(\mu_n)}(h)=D^{(\mu)}(h),\quad \mathscr{D}^{(\mu_n)}(s)=\exp(2\pi ni/q)M,$$

while the matrix M is decided from the following matrix equations:

$$D^{(\mu)}(s^{-1}hs)=M^{-1}D^{(\mu)}(h)M, \tag{53a}$$

$$M^n=I, \tag{53b}$$

where h in (53a) runs over all generators of H and I is a unit matrix.

The typical example is $D^{A_0}(T)\uparrow O=D^{A_1}\oplus D^{A_2},D^F(T)\uparrow O=D^{F_1}\oplus D^{F_2}$.

C. Comparison between the two induction techniques

The shortcomings of the induction method are as follows. (1) The theory is applicable only to the case when H is a maximal invariant subgroup of G. (2) The procedure is laborious. We need to solve a set of matrix equations (53) to get the matrix M. (3) Only numerical solutions for the irreps of G are found.

Folland proposed an improved induction method,[18] which is also applicable only to the case when $G(s)$ is an invariant subgroup of G.

Some special features of the double-induced technique are the following.

1. The ISO's are obtained by seeking the eigenvectors of the CSCO of G in the double-induced rep in a unified way regardless of the induced representations are reducible or irreducible, regardless they are s.v. or d.v., and regardless the subgroup is invariant or non-invariant.

2. All are derived *ab initio*. Only for identifying the irreps obtained by the eigenfunction method, do we need to consult the character table.

3. *Algebraic* instead of numerical solutions for the ISO's and irreducible matrices are found.

4. The present approach allows a high level of flexibility in choosing both the subgroup $G(s)$ according to the physical problem under consideration and the intrinsic subgroup $\overline{G(s')}$ according to the symmetrization requirement for the SAF's in the non-multiplicity case.

5. With the algebraic expression for the SAF's we are able for the first time to derive many symmetries of the SAF's, which are otherwise hidden in the numerical tables.

6. The algebraic expressions are functions of only the quantum numbers of the chosen group chain, while the ISO's in the coset factored form are functions of all the matrix elements of the coset generators.

Damhus *et al.*[19] gave the $T \supset C_3$ for $j=0, \frac{1}{2}, 1, \frac{3}{2}, 2$. Our results differ from Ref. 19 in phases.

Summary: The algebraic expression of the generalized projection operator greatly simplifies the projection operator method, enabling us to obtain a simple algebraic expression for the SAF's and thereby make the extensive tabulation of the SAF's obsolete.

The group chain $T \supset D_2 \supset C_2$ can be treated similarly and will be reported elsewhere. The application of this approach to the octahedral group and the icosahedral group as well as the relation between the double-induced representation and the symmetrized boson representation will be addressed in subsequent papers.[20]

ACKNOWLEDGMENTS

This work was supported in part by the Fundamental Research Fund of State Science and Technology Committee of China and the Research Fund of the National Educational Committee of China.

[1] E. P. Wigner, *Group Theory and Its Applications to Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959), Chaps. 24 and 26.
[2] G. F. Koster, J. O. Dimmock, R. G. Wheeler, and H. Statz, *Properties of the Thirty-Two Point Groups* (MIT Press, Cambridge, MA, 1963).
[3] C. J. Bradley and A. P. Cracknell, *Symmetry in Solids* (Oxford University Press, London, 1972).
[4] R. McWeeny, *Symmetry. An Introduction to Group Theory and Its Applications* (Pergamon, Oxford, 1963).
[5] P. H. Butler, *Point Group Symmetry Applications: Methods and Table* (Plenum, New York, 1981).
[6] J. S. Griffith, *The Irreducible Tensor Method for Molecular Symmetry Groups* (Prentice–Hall, London, 1962); *Theory of Transition Metal Ions* (Cambridge University Press, Cambridge, England, 1961).
[7] J. Q. Chen, *Group Representation Theory for Physicists* (World Scientific, Singapore, 1989); J. Q. Chen, M. J. Gao, and G. Q. Ma, Rev. Mod. Phys. **57**, 211 (1985).
[8] W. G. Harter, *Principle of Symmetry, Dynamics and Spectroscopy* (Wiley, New York, 1993).
[9] J. Q. Chen, A. Klein, and J. L. Pin, J. Math. Phys. **37**, 2400 (1996).
[10] J. Q. Chen and J. L. Pin, J. Math. Phys. **38**, 387 (1997).
[11] J. Q. Chen, ''Factorization of irreducible symmetry operators for semi-direct product groups,'' submitted to J. Math. Phys.
[12] J. Q. Chen and P. D. Fang, ''Algebraic solutions for dihedral groups D_n, C_{nv}, D_{nd}, and D_{nh} with arbitrary n,'' 1997 preprint.
[13] A. J. Coleman, Induced and subduced representations in *Group Theory and its Applications*, edited by E. M. Lobel (Academic, New York, 1968), Vol. I, p. 57.
[14] G. McLellian, J. Chem. Phys. **345**, 1350 (1961).
[15] Q. Zhang and X. Z. Li, Int. J. Quantum Chem. **29**, 293 (1986).
[16] J. L. Ping, Q. R. Zheng, B. Q. Chen, and J. Q. Chen, Comput. Phys. Commun. **52**, 355 (1989).
[17] M. E. Rose, *Elemental Theory of Angular Momentum* (Wiley, New York, 1957).
[18] N. O. Folland, J. Math. Phys. **18**, 31 (1977).
[19] T. Damhus, S. E. Harnung, and C. E. Schaffer, Theor. Chim. Acta **65**, 405 (1984).
[20] J. Q. Chen and P. D. Fan, ''The algebraic solution of point groups: The octahedral group, 1997 preprint; ''The algebraic solution of point groups: The icosahedral group,'' 1997 preprint.

q-deformed star products and Moyal brackets

Ömer F. Dayi[a)]
TÜBITAK—Research Institute for Basic Sciences,
P.O. Box 6, 81220 Çengelköy—Istanbul, Turkey

(Received 8 November 1996; accepted for publication 15 September 1997)

The standard and antistandard ordered operators acting on two-dimensional q-deformed phase space are shown to satisfy algebras which can be called $q-W_\infty$. q-star products and q-Moyal brackets corresponding to these algebras are constructed. Some applications like defining q-classical mechanics and q-path integrals are discussed. © *1998 American Institute of Physics.*
[S0022-2488(98)00901-3]

I. INTRODUCTION

A way of visualizing quantum mechanics as $\hbar$-deformation of the classical case is to utilize symbols of operators, star products and Moyal brackets.[1] If we denote the symbol map by $\mathcal{S}_O^{(\hbar)}$, to an operator $\hat{f}$ there corresponds a c-number object $f_O^{(\hbar)}$,

$$S_O^{(\hbar)}(\hat{f})=f_O^{(\hbar)},$$

where the subscript O denotes the operator ordering adopted. Noncommutativity of quantum mechanics is taken into account in terms of the star product defined such that

$$\mathcal{S}_O^{(\hbar)}(\hat{f}\hat{g})=\mathcal{S}_O^{(\hbar)}(\hat{f})\star_O^{(\hbar)}\mathcal{S}_O^{(\hbar)}(\hat{g})$$

and the symbol corresponding to the commutator (divided by $\hbar$) of the operators $\hat{f}$ and $\hat{g}$ is the Moyal bracket (or $\star^\hbar$-bracket):

$$\{f_O^{(\hbar)}(p,x),g_O^{(\hbar)}(p,x)\}_M\equiv\frac{1}{i\hbar}\,(f_O^{(\hbar)}(p,x)\star_O^\hbar g_O^{(\hbar)}(p,x)-g_O^{(\hbar)}(p,x)\star_O^\hbar f_O^{(\hbar)}(p,x)), \tag{1}$$

where p and x are classical phase space variables.

Recently $\star$-products and Moyal brackets are used in some diverse context such as constructing a representation of noncommuting forms[2] and they appeared in the formulation of two-dimensional area-preserving diffeomorphisms.[3] The latter is closely related to the quantum Hall effect which also exhibits a q-deformed structure where q is related to filling fraction.[4] Hence, understanding q-deformations in terms of symbols of the q-deformed operators may shed light on some aspects of the quantum Hall effect. Although the quantum Hall effect is a many-body problem, it is based on a field theory which is given in a two-dimensional phase space when it is restricted to the lowest Landau level. Hence, understanding q-deformed star products and Moyal brackets in a two-dimensional phase space is the first step in this direction.

One of the essential properties of the ordinary star product is its associativity. In this case it is known that the unique deformation of the Poisson bracket is the Moyal bracket (1).[5] On the other hand it was shown that it is impossible to deform Heisenberg dynamics by keeping the algebra of the observables associative.[6] Hence, to obtain a nontrivial deformation of the quantum mechanics we should sacrifice associativity of the star product. In fact, q-deformed algebras are not associative, so that a q-deformed star product leading to them should also be nonassociative.

a)Electronic mail: dayi@mam.gov.tr

Here, after deriving the algebras of the standard (or XP) and the antistandard (or PX) ordered operators q-deformed star products and Moyal brackets corresponding to them are constructed and the obstacles in generalizing Weyl ordering to the q-deformed operators are emphasized.

Once a q-Moyal bracket is obtained an immediate application is to define q-classical mechanics in terms of q-Poisson brackets. This and some other applications like classical as well as quantum mechanical properties of q-canonical transformations and a definition of q-path integrals on general grounds are briefly discussed.

II. ORDINARY SYMBOLS AND STAR PRODUCTS

In the two-dimensional phase space given in terms of the ordinary ($\hbar$-deformed) canonical operators

$$[P_\hbar,\ X_\hbar]=i\hbar, \tag{2}$$

an operator $\mathscr{O}$ can be written as

$$\mathscr{O}=\sum_{(m,n)>0}\mathscr{O}_{m,n}g(P_\hbar^m,X_\hbar^n), \tag{3}$$

where $g(P_\hbar^m,X_\hbar^n)$ are some functions of $P_\hbar^m$, $X_\hbar^n$ and $\mathscr{O}_{m,n}$ are some constant coefficients depending on the operator ordering scheme adopted. Hence, it is sufficient to deal with the monomials $g(P_\hbar^m,X_\hbar^n)$ as far as the algebraic properties of the operators are concerned. We are interested in the following three different ordering schemes which constitute complete basis.

Standard (or XP) ordering: In this scheme one deals with the monomials $X_\hbar^nP_\hbar^m$. They satisfy the Lie algebra[7]

$$[X_\hbar^nP_\hbar^m,\ X_\hbar^kP_\hbar^l]=\sum_{r=1}^{\infty}(i\hbar)^r r!\left\{\binom{k}{r}\binom{m}{r}-\binom{n}{r}\binom{l}{r}\right\}X_\hbar^{n+k-r}P_\hbar^{l+m-r}. \tag{4}$$

Antistandard (or PX) ordering: the suitable monomials $P_\hbar^nX_\hbar^m$, can be shown to satisfy[7]

$$[P_\hbar^nX_\hbar^m,\ P_\hbar^kX_\hbar^l]=\sum_{r=1}^{\infty}(-i\hbar)^r r!\left\{\binom{k}{r}\binom{m}{r}-\binom{n}{r}\binom{l}{r}\right\}P_\hbar^{n+k-r}X_\hbar^{l+m-r}. \tag{5}$$

The algebras (4) and (5) are called W_∞.[3]

Weyl ordering: The monomials

$$T_{m,n}^{(\hbar)}=\exp\left(P_\hbar\frac{\partial}{\partial\mathscr{P}}+X_\hbar\frac{\partial}{\partial\mathscr{Q}}\right)\mathscr{P}^m\mathscr{Q}^n|_{\mathscr{P}=|\mathscr{Q}=0}, \tag{6}$$

where $\mathscr{P}$ and $\mathscr{Q}$ are c-number variables, constitute a complete basis, and satisfy the Lie algebra[8,9]

$$[T_{m,n}^{(\hbar)},\ T_{k,l}^{(\hbar)}]=\sum_{a=0}^{b}(i\hbar)^{2a+l}B_{mnkl}^aT_{m+k-2a-l,n+l-2a-1}^{(\hbar)}, \tag{7}$$

where $b=\min\{[(m+k-1)/2,(n+l-1)/2]\}$. $B_{00kl}=B_{mn00}\equiv 0$ and for the other values of the indices

$$B_{mnkl}^a\equiv\sum_{c=0}^{2a+1}\frac{(-1)^c m!n!k!l!}{(2a+1-b)!b!(m+c-2a-1)!(n-c)!(k-c)!(l+c-2a-1)!}. \tag{8}$$

For our purposes it is sufficient to deal with the symbols of the monomials in each ordering scheme, which are given by

$$\mathscr{S}_S^{(\hbar)}[X_h^m P_h^n]=\mathscr{S}_A^{(\hbar)}[P_\hbar^m X_\hbar^n]=\mathscr{S}_W^{(h)}[T_{m,n}^{(h)}]=p^m x^n, \tag{9}$$

where p and x are c-number variables.

Related star products are[1,7]

$$\star_S^h \equiv \exp(i\hbar \overleftarrow{\partial}_p \vec{\partial}_x), \tag{10}$$

$$\star_A^\hbar \equiv \exp(-ih \overleftarrow{\partial}_x \vec{\partial}_p), \tag{11}$$

$$\star_W^\hbar \equiv \exp\left[\frac{-i\hbar}{2}(\overleftarrow{\partial}_x \vec{\partial}_p - \overleftarrow{\partial}_p \vec{\partial}_x)\right]. \tag{12}$$

We used $\partial/\partial x \equiv \partial_x$; $\partial/\partial p \equiv \partial_p$. One of them can be utilized in the Moyal bracket (1) in terms of the related symbols. Observe that $\star^\hbar$-products are associative,[1] so that the algebraic properties of the commutators are preserved.

III. CLASSICAL ($\hbar=0$) q-DEFORMED SYMBOLS, STAR PRODUCTS, AND MOYAL BRACKETS

Classical ($\hbar=0$) q-deformed canonical operators are defined as

$$P_q X_q - q X_q P_q = 0. \tag{13}$$

Now, the monomials of the different ordering schemes are equivalent up to an overall q-dependent factor. Thus it is sufficient to consider the algebra

$$q^{nk} T_{m,n}^{(q)} T_{k,l}^{(q)} - q^{ml} T_{k,l}^{(q)} T_{m,n}^{(q)} = 0, \tag{14}$$

where $T_{m,n}^{(q)} \equiv P_q^m X_q^n$. They give a complete basis in P_q, X_q space and the basis operators in another ordering scheme are equal to $T_{m,n}^{(q)}$ up to an overall q-dependent constant.

Symbols of the operators $T_{m,n}^{(q)}$ are

$$\mathscr{S}^{(q)} T_{m,n}^{(q)} = p^m x^n. \tag{15}$$

It is possible to define associative $\star^q$-products

$$\star_S^q \equiv \exp(\nu \overleftarrow{\partial}_p p x \vec{\partial}_x), \tag{16}$$

$$\star_A^q \equiv \exp(-\nu \overleftarrow{\partial}_x x p \vec{\partial}_p), \tag{17}$$

$$\star_W^q \equiv \exp\left[\frac{-\nu}{2}(\overleftarrow{\partial}_x x p \vec{\partial}_p - \overleftarrow{\partial}_p p x \vec{\partial}_x)\right], \tag{18}$$

where $\nu \equiv \ln q$ and the subscripts denote the resemblance to the ordinary star products (10)–(12). In fact, (16)–(18) can be obtained from (10)–(12) by a canonical transformation in accord with the fact that there exists a unique deformation of the Poisson bracket if the star product is associative.

The symbols defined in (15) satisfy

$$q^{nk} p^m x^n \star^q p^k x^l - q^{ml} p^k x^l \star^q p^m x^n = 0, \tag{19}$$

for any $\star^q$-product (16)–(18).

IV. q-SYMBOLS, q-STAR! q-SYMBOLS, q-STAR PRODUCTS AND q-MOYAL q-MOYAL BRACKETS

q-deformed ($\hbar \neq 0$) canonical variables satisfy

$$PX - qXP = i\hbar. \tag{20}$$

To reproduce (14) in $\hbar = 0$ limit we define the q-commutator as

$$[t(P^m,X^n),\ t(P^k,X^l)]_q \equiv q^{nk} t(P^m,X^n) t(P^k,X^l) - q^{ml} t(P^k,X^l) t(P^m,X^n), \tag{21}$$

where $t(P^m,X^n)$ is a function of P^m and X^n, depending on the operator ordering scheme adopted. Observe that the weights of the q-commutator (21) change according to the operators which are considered.

As one can easily see, the standard or the antistandard ordered monomials form a complete basis in the q-phase space given by P and X.

By explicit calculations q-algebra satisfied by the standard ordered monomials $X^n P^m$ can be derived:

$$[X^n P^m,\ X^k P^l]_q = \sum_{r=1}^{\infty} (i\hbar)^r [r]! \left\{ q^{(k-r)(n-r)+ml} \begin{bmatrix} k \\ r \end{bmatrix} \begin{bmatrix} m \\ r \end{bmatrix} - q^{(m-r)(l-r)+nk} \begin{bmatrix} n \\ r \end{bmatrix} \begin{bmatrix} l \\ r \end{bmatrix} \right\} X^{n+k-r} P^{l+m-r}, \tag{22}$$

where we used the definitions of the q-factorial

$$[n]! \equiv \left(\frac{1-q^n}{1-q} \right)! \equiv [1][2]\cdots[n-1][n],$$

and the q-binomial coefficient

$$\begin{bmatrix} n \\ r \end{bmatrix} \equiv \frac{[n]!}{[n-r]![r]!}.$$

Similarly we see that in the antistandard ordering scheme the monomials $P^m X^n$ satisfy the q-algebra

$$[P^n X^m,\ P^k X^l]_q = \sum_{r=1}^{\infty} (-i\hbar)^r q^{r(r-1)/2} [r]! \left\{ \begin{bmatrix} k \\ r \end{bmatrix} \begin{bmatrix} m \\ r \end{bmatrix} - \begin{bmatrix} n \\ r \end{bmatrix} \begin{bmatrix} l \\ r \end{bmatrix} \right\} P^{n+k-r} X^{m+l-r}. \tag{23}$$

The q-deformed algebras (22) and (23) can be called q-W_∞. For other definitions of q-W_∞ algebra see Ref. 10.

Symbol maps for the standard and antistandard orderings are defined like the ordinary case:

$$\mathscr{S}_S(X^m P^n) = \mathscr{S}_A(P^m X^n) = p^m x^n. \tag{24}$$

As emphasized before, the ordinary star products should be associative so that the Moyal bracket satisfies an identity corresponding to the Jacobi identity satisfied by the ordinary operators. However, now the situation is altered drastically: we do not any more deal with the commutators which are not aware of their entries but with the q-commutators (21) which change according to their entries, i.e., the underlying algebraic structure is nonassociative. Thus the associativity condition cannot be preserved and we can obtain a nontrivial deformation of the Poisson bracket other than $\hbar$-deformation.

In terms of the q-derivative

$$D_z f(z) \equiv \frac{f(z)-f(qz)}{(1-q)z}, \tag{25}$$

we can construct q-star products for the standard and antistandard orderings as

$$\star_S \equiv \sum_{r=0}^{\infty} \frac{(i\hbar)^r}{[r]!} \overleftarrow{D}_p^r \exp(\nu \overleftarrow{\partial}_p p x \vec{\partial}_x) \vec{D}_x^r, \tag{26}$$

$$\star_A \equiv \sum_{s=0}^{\infty} (-\nu \overleftarrow{\partial}_x x)^s \sum_{r=0}^{\infty} \frac{(-i\hbar)^r q^{r(r-1)/2}}{[r]!} \overleftarrow{D}_x^r \vec{D}_p^r (p \vec{\partial}_p)^s. \tag{27}$$

One can see that if the q-Moyal bracket is defined as

$$\{p^m x^n, p^k x^l\}_{q-M} \equiv \frac{1}{i\hbar} (q^{nk} p^m x^n \star p^k x^l - q^{ml} p^k x^l \star p_m x^n), \tag{28}$$

by using the symbols (24) and the q-star products (26)–(27), q-Moyal algebras corresponding to (22) and (23) are obtained. In Ref. 11 another q-star product is defined by using the coherent states maps, however it does not lead to the algebras which we deal within (22) and (23).

Generalization of the Weyl ordering (6) to the q-phase space is not obvious: a term of a monomial can be generalized by assuming that it is weighted with a factor q^γ, where γ is a number depending on the term under consideration. To emphasize the difficulties related to this ordering procedure, let us suppose that there exist operators $T_{m,n}$ leading to the ordinary Weyl ordered operators in the $q=1$ limit, satisfying

$$[T_{m,n},\ T_{k,l}]_q = \sum_{r,s=0}^{A,B} \mathscr{C}_{mnkl}^{rs}(\hbar,q) T_{r,s}, \tag{29}$$

where for obtaining the correct classical limit $\mathscr{C}$ should satisfy

$$\mathscr{C}_{mnkl}^{(K)rs}(0,q) = 0. \tag{30}$$

An operator algebra is proposed in Ref. 9 as a generalization of (7) by replacing the factorial terms with the q-factorials:

$$[T_{m,n}^{(GF)}, T_{k.l}^{(GF)}]_q = \sum_{a=0}^{b} \sum_{c=0}^{2a+1} \frac{(-1)^c [m]![n]![k]![l]!}{[2a+1-b]![b]![m+c-2a-1]![n-c]![k-c]![l+c-2a-1]!} \times (i\hbar)^{2a+1} T_{m+k-2a-1,n+l-2a-1}^{(GF)}, \tag{31}$$

where b is given as in (7). The symbol map independent of the definition of the q-Weyl ordered operators is

$$S_W T_{m,n}^{(GF)} = p^m x^n, \tag{32}$$

so that the $\star$-product reproducing (31), in terms of the q-Moyal bracket (28), is

$$\star^W \equiv \sum_{M=0}^{\infty} (-\nu/2)^M \sum_{L=0}^{M} \frac{(-1)^L}{(M-L)!L!} (\overleftarrow{\partial}_x x)^{M-L} (\overleftarrow{\partial}_p p)^L \sum_{\alpha=0}^{\infty} (-i\hbar/2)^\alpha$$
$$\times \sum_{\beta=0}^{\alpha} \frac{(-1)^\beta}{[\alpha-\beta]![\beta]!} \overleftarrow{D}_x^{\alpha-\beta} \overleftarrow{D}_p^{\beta} \vec{D}_p^{\alpha-\beta} \vec{D}_x^{\beta} (p\vec{\partial}_p)^{M-L} (x\vec{\partial}_x)^L. \tag{33}$$

Obviously, independent of how the generalization is done, we have

$$T_{m,0} = P^m, \quad T_{0,m} = X^m. \tag{34}$$

Hence, by studying the q-commutators $[P^m, X^m]_q$ one can try to reach to the other q-weighted monomials: e.g.,

$$P^2X^2 - q^4X^2P^2 = i\hbar[2](PX + q^2XP) \tag{35}$$

suggests that $T_{1,1} \approx (PX + q^2XP)$. Similarly, explicit calculation gives

$$P^2X^3 - q^6X^3P^2 = [2](PX^2 + q^4X^2P + q^2XPX), \tag{36}$$

yielding $T_{1,2} \approx (PX^2 + q^4X^2P + q^2XPX)$. However, the q-commutator of this operator with P,

$$P(PX^2 + q^4X^2P + q^2XPX) - q^2(PX^2 + q^4X^2P + q^2XPX)P = i\hbar(1 + q + q^2)(PX + q^3XP), \tag{37}$$

suggests $T_{1,1} \approx (PX + q^3XP)$, which is in contradiction with the one suggested by (35).

In fact, having obstacles in defining q-Weyl ordering is not surprising. If one considers monomials reproducing the Weyl ordered ones in the $q=1$ limit with a definite q weight, they cannot constitute a complete basis. For having a complete basis one should have monomials with all possible q weights.

V. APPLICATIONS

In terms of the q-star products and the related q-Moyal brackets (26)–(28), one can study quantum as well as classical dynamics on general grounds.

If we deal with the $\star s$-product, q-classical dynamics can be given in terms of the q-Poisson bracket defined for the observables $f(p,x) = \Sigma_i f_i(p,x)$ and $g(p,x) = \Sigma_j g_j(p,x)$ where f_i and g_j are monomials in p, x, as

$$\{f(x,p), g(x,p)\}_{q-P} \equiv \lim_{\hbar \to 0} \{f(x,p), g(x,p)\}_{q-M}$$
$$= \sum_{i,j} [q^{\alpha(f_i, g_j)} (D_p f_i) \exp(\nu \overleftarrow{\partial}_p p x \overleftarrow{\partial}_x)(D_x g_j)$$
$$- q^{\alpha(g_j, f_i)} (D_p g_j) \exp(\nu \overleftarrow{\partial}_p p x \overleftarrow{\partial}_x)(D_x f_i)], \tag{38}$$

where $\alpha(p^m x^n, p^k x^l) = nk$. Thus, if $H = \Sigma_k H_k$ is the classical Hamiltonian where H_k are monomials, the equation of motion of the observable $f(p,x)$ is

$$\tau_q(f) = \{H, f\}_{q-P}, \tag{39}$$

where $\lim_{q \to 1} \tau_q = d/dt$. Now, observe that in general

$$\{H, fg\}_{q-P} \neq f\{H, g\}_{q-P} + g\{H, f\}_{q-P}, \tag{40}$$

which leads to

$$\tau_q(fg) \neq \tau(f)g + f\tau(g). \tag{41}$$

Obviously, there are some exceptions like f or g is constant. One may think that $\star s$ of f and g should be considered on the left-hand side of (41). However, in the limit $\hbar \to 0$ the q-star product will yield $\star_S^{(q)}$ (16), and $f \star_S^{(q)} g = q^{\alpha(f,g)} fg$, so that there is not any difference.

In the ordinary classical mechanics canonical transformations leave the basic Poisson brackets invariant. One of these is the point transformation defined as

$$u = f(x); \quad p_u = (\partial_x f(x))^{-1} p, \tag{42}$$

where f is an invertible function. For the q-classical mechanics point transformation can be defined as

$$u = f(x); \quad p_u = (D_x f(x))^{-1} p. \tag{43}$$

Now, in terms of the q-Poisson bracket (38) one can observe that

$$\{u, p_u\}_{q-P} = -q^{\alpha(u,p_u)j} \quad \{p_u, u\}_{q-P} = 1, \tag{44}$$

instead of the ones satisfied by p,x:

$$\{x,p\}_{q-P} = -q; \quad \{p,x\}_{q-P} = 1. \tag{45}$$

In (44) $\alpha(u,p_u) = x\partial_x \log f(x)$, which is a number. Let us have an example where $f(x) = \sqrt{x}$:

$$u = \sqrt{x}; \quad p_u = \left[\frac{1}{2}\right]^{-1} \sqrt{x} p, \tag{46}$$

so that

$$\{u, p_u\}_{q-P} = -q^{1/2}; \quad \{p_u, u\}_{q-P} = 1. \tag{47}$$

In fact, the transformation (46) was studied in Ref. 12 and it was found that it is a q-canonical transformation if the phase space operators satisfy

$$\hat{p}\hat{x} - q\hat{x}\hat{p} = i\hbar, \tag{48}$$

$$\hat{p}_u \hat{u} - \sqrt{q}\hat{u}\hat{p}_u = i\hbar, \tag{49}$$

which are consistent with (45) and (47).

When we deal with q-quantum mechanics in the Heisenberg picture, time evolution of an observable f is given by

$$\tau(f) = \{H,f\}_{q-M} = \frac{1}{i\hbar} \sum_{i,j} (q^{\alpha(H_i, f_j)} H_i \star f_j - q^{\alpha(f_j, H_i)} f_j \star H_i). \tag{50}$$

Here $\star$ indicates one of the q-star products (26) and (27).

In the Schrödinger picture time evolution of a time-dependent state vector ψ is given by

$$i\hbar\, \tau(\psi(t)) = \hat{H}\psi(t), \tag{51}$$

where $\hat{H}$ is the q-Hamiltonian operator, i.e., $\mathscr{S}(\hat{H}) = H$.

Contrary to the ordinary quantum mechanics, in the q-deformed case the relation between the Schrödinger picture (51) and the Heisenberg picture (50) is not clear. The unique common feature

of the deformed and nondeformed cases is the fact that in both of the cases symbols of the monomials are the same, namely, $p^m x^n$. The difference lies in the definition of the related star products. Thus we may still assume that the symbol of the evolution operator $\hat{U}(t)$ is

$$U(t)=\mathscr{S}(\hat{U}(t))=e^{(it/h)H}. \tag{52}$$

Then we can adopt the definition of the path integral of the ordinary time evolution given in terms of the star products and symbols in Ref. 13 to define the q-path integral as

$$G(t)=\lim_{N\to\infty} U\left(\frac{t}{N}\right)\star\cdots\star U\left(\frac{t}{N}\right). \tag{53}$$

When the canonical transformation (42) is performed there will be some q-quantum corrections in the transformed Hamiltonian. The kinetic term including the q-quantum corrections can be studied in terms of the $\star s$-product similar to the ordinary case[7] by defining the transformed kinetic term as

$$\tilde{H}_0=(D_x f)^{-1/2}\star_s p\star_s (D_x f)^{-2}\star_s D_x f\star_s p\star_s (D_x f)^{-1/2}. \tag{54}$$

Because of the nonassociativity of the q-star product we should specify in which order the multiplications will be performed in (54). However, we do not possess a general procedure.

As it is briefly illustrated, q-star products and q-Moyal brackets are very useful in studying several aspects of q- deformations on general grounds. However, the relations to the other formulations of q-dynamics (e.g., see Ref. 14) and q- path integral definitions[15] should be studied.

ACKNOWLEDGMENTS

A part of this work is done in II. Institut für Theoretische Physik, Uni Hamburg with the support of A. von Humboldt Foundation. I thank them for their kind hospitality. I am also thankful to M. Reuter for the discussions which were crucial for this work.

[1] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. (New York) **111**, 61 (1978); **111**, 111 (1978).
[2] M. Reuter, Int. J. Mod. Phys. A **11**, 1253 (1996).
[3] D. B. Fairlie and C. K. Zachos, Phys. Lett. B **224**, 101 (1989); I. Bakas, Phys. Lett. B **228**, 57 (1989); C. N. Pope, L. J. Romans, and X. Shen, Phys. Lett. B **236**, 173 (1990); J. Hoppe and P. Schaller, Phys. Lett. B **237**, 407 (1990).
[4] S. Iso, D. Karabali, and B. Sakita, Phys. Lett. B **296**, 143 (1992); A. Capelli, C. A. Trugenberger, and G. R. Zemba, Nucl. Phys. B **396**, 465 (1993); A. Capelli, C. A. Trugenberger, and G. R. Zemba, Nucl. Phys. B **398**, 531 (1993); I. I. Kogan, Int. J. Mod. Phys. A **9**, 3887 (1994); M. Flohr and R. Varnhagen, J. Phys. A **27**, 3999 (1994).
[5] W. Arveson, Commun. Math. Phys. **89**, 77 (1983); P. Fletcher, Phys. Lett. B **248**, 323 (1990).
[6] M. Pillin, Commun. Math. Phys. **180**, 23 (1996).
[7] G. V. Dunne, J. Phys. A **21**, 2321 (1988).
[8] C. M. Bender and G. V. Dunne, Phys. Rev. D **40**, 3504 (1989).
[9] I. M. Gelfand and D. B. Fairlie, Commun. Math. Phys. **136**, 487 (1991).
[10] C-Z. Zha, J. Math. Phys. **35**, 517 (1994); E. H. Al Kinani and M. Zakkari, Phys. Lett. B **357**, 105 (1995).
[11] A. Odzijewicz, ''Quantum algebras and q-special functions related to coherent states maps of the disc,'' Warsaw University preprint IFT-18/95.
[12] Ö. F. Dayi and I. H. Duru, J. Phys. A **28**, 2395 (1995).
[13] F. A. Berezin, Usp. Fiz. Nauk **132**, 497 (1980).
[14] A. Chodos and D. G. Caldi, J. Phys. A **24**, 5505 (1991).
[15] L. Baulieu and E. G. Floratos, Phys. Lett. B **258**, 171 (1991); M. Chaichian and A. P. Demichev, Phys. Lett. B **320**, 273 (1994).

Solutions of 1+1 relativistic field models

A. Degasperis
Dipartimento di Fisica, Università di Roma ''La Sapienza,'' P.le Aldo Moro 2, 00185 Roma, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy

S. De Lillo
Dipartimento di Fisica, Università di Perugia, Via Pascolini, 06100 Perugia, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, Italy

F. Tinebra
Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy

(Received 19 November 1996; accepted for publication 15 July 1997)

A class of classical relativistic nonlinear field models has been recently introduced by two of us [A. Degasperis and F. Tinebra, J. Math. Phys. **34**, 2950 (1993)]; they are solvable by an explicit mapping to free field equations. Here we show the behavior of particular solutions of these models for both scalar and spinor fields. In the real spectrum we find brachion and tachion localized solutions, and we show the large-time behavior of wave packets, while in the complex spectrum we prove that the asymptotic states may contain solitons, antisolitons, and point-like charged particles (''pointons''). We report on several processes such as soliton–antisoliton pair creation and annihilation, and pointon decay in two solitons. © *1998 American Institute of Physics.* [S0022-2488(98)02701-7]

I. INTRODUCTION

A new class of relativistic field equations has been recently introduced by two of us.[1] Two simple instances of them are the following nonlinear versions of the Klein–Gordon equation:

$$\Box Q+[m^2-2\epsilon^{\mu\nu}\ \mathrm{Im}(Q^*_{,\mu}Q_{,\nu})-g^{\mu\nu}\ \mathrm{Im}(Q^*Q_{,\mu})\mathrm{Im}(Q^*Q_{,\nu})]Q=0, \tag{1.1}$$

and of the Dirac equation

$$\gamma^\mu\phi_{,\mu}+(im-\epsilon_{\mu\nu}\gamma^\mu\overline{\phi}\gamma^\nu\phi)\phi=0. \tag{1.2}$$

While being invariant with respect to the Lorentz group of coordinate transformations, these PDE's possess the nice property of being solvable by a transformation of the field variable which maps them into their linearized form. The price we pay to buy this virtue is that they are constrained to live in a 1+1 dimensional space–time. For motivations and further observations, together with reference to the literature, the reader is referred to Ref. 1, where the method of construction, based on charge conservation, of our relativistic models is detailed (for charge conservation combined with gauge transformation in a nonrelativistic context, see Ref. 2). In fact, in Ref. 1 the attention is mainly devoted to deriving the general class of solvable PDE's, for both scalar and spinor fields, together with systems of coupled equations, and only marginal results are given on general properties and solutions.

The purpose of this paper is to present special classes of solutions whose explicit expression obtains by mapping simple linear combinations of exponential solutions of the free field. According to whether these exponentials are everywhere bounded (real spectrum) or unbounded (complex spectrum), the corresponding solutions of our nonlinear model show different behaviors. The resulting phenomenology is better described by the charge density rather than by the field itself. Thus, in the real part of spectrum, we prove the existence of brachion and tachion-type solutions

with power-localization in the spatial coordinate, and we show that nonlinear wavepackets have a boomerang behavior. As for the complex spectrum, both scalar and spinor solitons show up with a rich phenomenology, including pair-production and annihilation. Of course, in contrast with the scalar case, where one can use terms borrowed from elementary particle physics, in the spinor case this is not possible since classical spinor variables commute rather than anticommute, this leading to the fact that all states have positive charge.

The derivation of our results is elementary, though actual computations are rather lenghty. Our presentation is not always aimed to generality, and we content ourselves with pointing out behaviors of simple solutions. Throughout the paper, we present results in relativistically invariant form. To this aim, for the benefit of the reader, we introduce here our notation, and, for future reference, few useful formulas.

Space–time coordinates are equivalently denoted

$$z=(x^0,x^1)=(t,x)=x^\mu, \tag{1.3}$$

according to convenience. The light speed is set equal to unity, $c=1$, and greek indices take the values 0 and 1, while two equal indices are meant to be summed over these values. Lorentz transformations

$$x^\mu \to x'^\mu = L^\mu_\nu x^\nu \tag{1.4}$$

are defined by the 2×2 matrix L^μ_ν, which may be given the following equivalent expressions:

$$L^\mu_\nu = \gamma \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}, \quad \gamma \equiv (1-\beta^2)^{-1/2}, \quad -1<\beta<1, \tag{1.5a}$$

$$L^\mu_\nu = \begin{pmatrix} \cosh\,\alpha & \sinh\,\alpha \\ \sinh\,\alpha & \cosh\,\alpha \end{pmatrix}, \quad -\infty<\alpha<+\infty, \tag{1.5b}$$

where $\beta=\mathrm{tgh}\,\alpha$; in this second notation, the group property is more transparent, as it reads $L(\alpha_1)L(\alpha_2)=L(\alpha_1+\alpha_2)$, and the Lorentz matrix takes the exponential expression

$$L^\mu_\nu = L(\alpha) = \exp(\alpha\sigma_1), \tag{1.6}$$

σ_1 being the Pauli matrix $\sigma_1=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$. The metric tensor

$$g_{\mu\nu} = \mathrm{diag}(1,-1) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{1.7}$$

and the skew-symmetric Ricci tensor

$$\epsilon_{\mu\nu} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{1.8}$$

are left unchanged by a Lorentz transformation, so that, with any pair of vectors a^μ and b^μ, the following expressions:

$$g_{\mu\nu}a^\mu b^\nu = a^0b^0 - a^1b^1, \quad \epsilon_{\mu\nu}a^\mu b^\nu = a^0b^1 - a^1b^0, \tag{1.9}$$

are Lorentz-invariant. We also note that

$$g^{\mu\nu}=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},\quad \epsilon^{\mu\nu}=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix},\quad \epsilon_{\mu\alpha}\epsilon_{\nu\beta}g^{\alpha\beta}=-g_{\mu\nu}. \tag{1.10}$$

With two given vectors a^μ and b^μ there can be associated, at most, three independent invariants, so that the following four invariants:

$$I_a\equiv g_{\mu\nu}a^\mu a^\nu,\quad I_b\equiv g_{\mu\nu}b^\mu b^\nu,\quad I\equiv g_{\mu\nu}a^\mu b^\nu,\quad \tilde{I}\equiv\epsilon_{\mu\nu}a^\mu b^\nu, \tag{1.11}$$

are not independent, namely,

$$I^2=I_aI_b+\tilde{I}^2. \tag{1.12}$$

The basic ingredient of the solutions of linear differential equations is the exponential, $\exp(g_{\mu\nu}u^\mu x^\nu)$, where u^μ is a complex vector with fixed norm,

$$g_{\mu\nu}u^\mu u^\nu=m^2, \tag{1.13}$$

where the mass m is a given (positive) real number. This complex vector u^μ is parametrized in the following way:

$$u^\mu=\eta\ \cos\ \theta k^\mu+i\tau\ \sin\ \theta\epsilon^{\mu\nu}k_\nu, \tag{1.14}$$

where the scalars η and τ are sign parameters,

$$\eta=\pm 1,\quad \tau=\pm 1,$$

and are referred to as, respectively, ''charge'' and ''chirality'' (see below). The ''spectral angle'' parameter θ is a scalar to be taken in the interval

$$0\leqslant\theta\leqslant\frac{\pi}{2}, \tag{1.15}$$

and the real vector k^μ is defined as

$$k^\mu=(\omega,k),\quad \omega=+\sqrt{m^2+k^2}, \tag{1.16}$$

the wave number k being a real parameter, $-\infty<k<+\infty$, the associated frequency ω being positive definite, $m\leqslant\omega$. An alternative parametrization of the real vector k^μ may be

$$k^\mu=m(\cosh\ \alpha,\sinh\ \alpha),\quad -\infty<\alpha<+\infty, \tag{1.17}$$

namely $\omega=m\ \cosh\alpha$ and $k=m\ \sinh\alpha$. The expression (1.14), together with (1.16), justifies the introduction of the following notation:

$$E(k,\eta,\theta,\tau)\equiv\exp(-iu^\mu x_\mu). \tag{1.18}$$

For future reference, we note that, if u_n^μ and u_m^μ are any two vectors satisfying condition (1.13), then (an asterisk stands for complex conjugation) the vector $u_n^{\mu *}-u_m^\mu$ is orthogonal to the vector $u_n^{\mu *}+u_m^\mu$, and, therefore,

$$\epsilon^{\mu\nu}(u_{n\nu}^*+u_{m\nu})=[\lambda_{nm}/(m^2-\mu_{nm})](u_n^{\mu *}-u_m^\mu), \tag{1.19a}$$

$$\epsilon^{\mu\nu}(u_{n\nu}^*-u_{m\nu})=-[\lambda_{nm}/(m^2+\mu_{nm})](u_n^{\mu *}+u_m^\mu), \tag{1.19b}$$

where we have introduced the two invariants

$$\mu_{nm} \equiv g_{\mu\nu} u_n^{\mu *} u_m^{\nu}, \quad \lambda_{nm} \equiv \epsilon_{\mu\nu} u_n^{\mu *} u_m^{\nu}, \tag{1.20}$$

which, because of (1.12), satisfy the relation

$$\mu_{nm}^2 = m^4 + \lambda_{nm}^2. \tag{1.21}$$

As for spinors, our notation is the following. The Dirac vector $\gamma^\mu = (\gamma^0, \gamma^1)$ is

$$\gamma^0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma^1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{1.22}$$

which satisfies the usual condition $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$. If φ is the spinor

$$\varphi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \tag{1.23}$$

then the corresponding bar-conjugated spinor is defined as

$$\overline{\varphi} = \varphi^+ \gamma^0 = (\varphi_1^* \quad -\varphi_2^*). \tag{1.24}$$

The transformation of spinors that corresponds to the Lorentz transformation (1.4) reads

$$\varphi \to \varphi' = \wedge \varphi, \tag{1.25}$$

where the 2×2 matrix $\wedge$ takes the expression

$$\wedge = (1 + \gamma + \beta\gamma\sigma_1)/\sqrt{2(1+\gamma)} = \exp\left(\frac{\alpha}{2}\,\sigma_1\right), \tag{1.26}$$

α, β, and γ being the same as in (1.5) and (1.6). Thus, if φ and ψ are two spinors, then $\overline{\varphi\psi}$ and $\overline{\varphi}\gamma^\mu\psi$ are, respectively, scalar and vector with respect to Lorentz transformations. As basic spinors, we use the following expressions:

$$\mathscr{E}(k, \eta, \theta, \tau) = E(k, \eta, \theta, \tau)\chi(k, \eta, \theta, \tau), \tag{1.27}$$

where the exponential function $E(k, \eta, \theta, \tau)$ is defined by (1.18), (1.14), and (1.16), and the constant spinor $\chi(k, \eta, \theta, \tau)$ has the expression

$$\chi(k, \eta, \theta, \tau) \equiv \frac{1}{2}\sqrt{m\cos\theta}[1 + \eta + (1-\eta)\gamma^1]\exp\left[\frac{1}{2}(\eta\alpha + i\tau\theta)\sigma_1\right]\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{1.28}$$

with α such that $k = m \sinh\alpha$, and with the normalization

$$\overline{\chi}\chi = \eta m \cos^2\theta. \tag{1.29}$$

The spinor (1.28) obtains by applying the Lorentz transformation (1.25), with (1.26), to the spinor $\chi(0, \eta, \theta, \tau)$ in the rest reference frame (i.e., for $k = 0$).

For future reference, we report here, without proof, few properties of the basic spinor (1.28). To this aim, we first introduce the following simplifying notation:

$$\chi_n = \chi(k_n, \eta_n, \theta_n, \tau_n), \quad n = 1,2,..., \tag{1.30}$$

$$\overline{\chi_n}\chi_m \equiv S_{nm}\,, \quad \overline{\chi_n}\gamma^\mu\chi_m \equiv V^\mu_{nm}\,. \tag{1.31}$$

Since the vector V^μ_{nm} satisfies the equation

$$\epsilon_{\mu\nu}V^\mu_{nm}(u^{\nu *}_n + u^\nu_m) = 0, \tag{1.32}$$

its expression has to be of the form

$$V^\mu_{nm} = \Gamma_{nm}(u^{\mu *}_n + u^\mu_m). \tag{1.33}$$

Moreover, we find that

$$\Gamma_{nm} = [m/(m^2+\mu_{nm})]S_{nm}\,, \tag{1.34}$$

where the two invariants μ_{nm} and S_{nm} are defined by (1.20) and, respectively, (1.31). The explicit expression of S_{nm} follows from its definition, together with (1.28), and turns out to read

$$S_{nm} = \eta_n[\cos\,\theta_n\,\cos\,\theta_m(m^2+\mu_{nm})/2]^{1/2}, \quad \eta_n\eta_m = 1, \tag{1.35a}$$

$$S_{nm} = \eta_n\lambda_{nm}[\cos\,\theta_n\,\cos\,\theta_m/2(m^2-\mu_{nm})]^{1/2}, \quad \eta_n\eta_m = -1. \tag{1.35b}$$

We note that the space component of the complex wave vector u^μ, see (1.14), can be any complex number, and plays the role of the spectral variable. Bounded free field solutions obtain only if the spectral variable is real ($\theta=0$), where sums of exponentials provide periodic or quasi-periodic solutions, and integrals over k yield wave packet solutions. Complex u^μ solutions are exponentially unbounded; however, they are still of interest as they may be mapped into bounded solutions of the nonlinear field equations. Because of this, we discuss below the two cases separately, namely, after a short presentation of the models given in Sec. II, in Sec. III we deal with the real spectrum and in Sec. IV with the complex spectrum. Section V is devoted to conclusions.

II. MODELS AND LINEARIZATION

Here we shortly review the linearization method, and discuss few general properties of the mapping which takes solutions of the linear (free-field) equation into solutions of the nonlinear equation. Let us consider first the complex scalar field $q(z)$ which satisfies the Klein–Gordon equation (m is a real constant),

$$\Box q + m^2 q = 0, \tag{2.1}$$

and, consequently, the conservation law,

$$j^\mu{}_{,\mu} = 0, \tag{2.2}$$

for the current

$$j_\mu = \mathrm{Im}(q q^*_{,\mu}). \tag{2.3}$$

In the following method, it is convenient to introduce, together with the field $q(z)$, an associated real scalar field $v(z)$ with the property of satisfying the field equation

$$v_{,\mu} = \epsilon_{\mu\nu}j^\nu, \tag{2.4}$$

which is certainly integrable because of the conservation law (2.2). Thus, from now on, the pair of fields $q(z)$ and $v(z)$ are the dynamical variables of the model.

Consider now the mapping $q \to Q$, $v \to V$ as defined by the formulas

$$Q=q\ \exp[f(v)],\quad V=g(v), \tag{2.5}$$

with the condition of being invertible, namely

$$q=Q\ \exp[-F(V)],\quad v=G(V). \tag{2.6}$$

At this point, f and g are arbitrary real functions of a real variable (except for the obvious requirement that the operations below be meaingful), while F and G obtain via the invertibility condition,

$$F(g(v))=f(v),\quad G(g(v))=v. \tag{2.7}$$

We note that, because of the reality of the function f, the transformation (2.5), just in contrast with respect to the local SU(1) gauge transformation, changes the amplitude of the field value q, but not its phase. Moreover, its dependence on the real field v, together with the field equation (2.4), guarantees that also the current associated with the new field Q,

$$J_\mu=\mathrm{Im}(QQ^*_{,\mu}), \tag{2.8}$$

is conserved, i.e.,

$$J^\mu_{,\mu}=0. \tag{2.9}$$

This property of our mapping, which is basic in the present approach, directly follows from the transformation rule

$$J_\mu=j_\mu\ \exp[2f(v)] \tag{2.10}$$

and Eq. (2.4).

Let us look now at the equations satisfied by the new fields Q and V. For the latter one, the equation reads

$$V_{,\mu}=H(V)\epsilon_{\mu\nu}J^\nu,\ \ H(V)\equiv\exp[-2F(V)]/G'(V), \tag{2.11}$$

and it is evident that, with no loss of generality (it is just a matter of a change of the variable V), one can ask that $H(V)=1$, so that V formally satisfies the same equation as v, that is

$$V_{,\mu}=\epsilon_{\mu\nu}J^\nu. \tag{2.12}$$

On the contrary, the equation for Q is no more linear and comes out to be

$$\Box Q+[m^2+(F''-F'^2)J_\alpha J^\alpha+2F'\epsilon^{\alpha\beta}\ \mathrm{Im}(Q^*_{,\alpha}Q_{,\beta})]Q=0. \tag{2.13}$$

We first observe that with the choice $H(V)=1$, see (2.11), $F(V)$ and $G(V)$ are related to each other by

$$G(V)=\Gamma+\int_0^V dV'\ \exp[-2F(V')], \tag{2.14}$$

where Γ is an integration constant. As a consequence of the inversion formulas (2.7), also the function $g(v)$ is no more arbitrary, as it is given by the expression

$$g(v)=\gamma+\int_0^v dv'\ \exp[2f(v')], \tag{2.15}$$

γ being an arbitrary constant. Therefore, the mapping $q \to Q$ is characterized by a function $f(v)$ and a real parameter γ, while $F(V)$ and Γ are consequently determined by the inversion condition (2.7).

Before proceeding further, it is worth pointing out that the class of mappings $M(\gamma,f):(q,v)\to(Q,V)$ under consideration, can be restricted by using an equivalence relation according to the following definition. Two mappings $M(\gamma,f)$ and $M(\tilde{\gamma},\tilde{f})$ are equivalent to each other if, for any solution (q,v) of (2.1) and (2.4) there exists another solution of the same equations, say $(\tilde{q},\tilde{v})$, such that the corresponding transformed fields are the same: $M(\gamma,f):(q,v)\to(Q,V)$ and $M(\tilde{\gamma},\tilde{f}):(\tilde{q},\tilde{v})\to(Q,V)$. This means that the two corresponding nonlinear equations, say (2.13) with F and $\tilde{F}$, have the same manifold of solutions and, therefore, do not deserve separate analysis. In particular, it is easily proved that, if $u:(q,v)\to(\tilde{q},\tilde{v})$ is a transformation which leaves the free field Eqs. (2.1) and (2.4) invariant, and if the composed map $M(\gamma,f)\star u$ is of the same type, i.e., $M(\gamma,f)\star u = M(\tilde{\gamma},\tilde{f})$, then $M(\gamma,f)$ and $M(\tilde{\gamma},\tilde{f})$ are equivalent. A simple application of this result is the following. Consider the two parameter transformation $u:(q,v)\to(\tilde{q},\tilde{v})$ with

$$\tilde{q}=aq, \quad \tilde{v}=a^2v+b, \tag{2.16}$$

where a and b are real parameters; then, given a mapping $M(\gamma,f):q\to Q$, the composed mapping $M(\tilde{\gamma},\tilde{f})=M(\gamma,f)\star u$ is readly found to correspond to

$$\tilde{f}(v)=f(a^2v+b)+\ln a, \quad \tilde{\gamma}=\gamma+\int_0^b dv\ \exp[2f(v)]. \tag{2.17}$$

Therefore, we conclude that, by choosing b and a so that

$$\int_0^b dv\ \exp[2f(v)]=-\gamma, \tag{2.18}$$

and

$$a=\exp[-f(b)], \tag{2.19}$$

within the class of equivalence of any mapping $M(\gamma,f)$, there exists a mapping $M(\tilde{\gamma},\tilde{f})$ such that $\tilde{\gamma}=0$ and $\tilde{f}(0)=0$. This observation gives us the possibility to consider, without loss of generality, only mappings of the form $M(0,f(v))$, with the additional property $f(0)=0$. It is also easy to see that the inverse mapping, characterized by the function $F(V)$ and the paramter Γ, satisfies the same condition, say $\Gamma=F(0)=0$. In conclusion, here and below we confine our attention to invertible mappings $q\leftrightarrow Q$, $v\leftrightarrow V$ of the form (2.5) and (2.6) with

$$g(v)=\int_0^v dv'\ \exp[2f(v')], \quad G(V)=\int_0^V dV'\ \exp[-2F(V')], \tag{2.20}$$

which are characterized by one real function $f(v)$ (or, equivalently, $F(V)$, see (2.7)) with the property

$$f(0)=0 \quad (\text{or } F(0)=0). \tag{2.21}$$

Once $f(v)$ (and, therefore, $F(V)$) is fixed, the nonlinear field equations of interest are (2.13) and (2.12); we note that the nonlinearity in the complex scalar field Q in (2.13) is only cubic and quintic for any choice of the function $F(V)$.

It is obvious that our mapping, see (2.5), (2.6), and (2.20), translates whatever property of the free field equations into the corresponding one for the nonlinear Eqs. (2.12) and (2.13). Instances of interesting properties are the transformations which map a solution (Q,V) into another solution $(\widetilde{Q},\widetilde{V})$, and the nonlinear superposition principle. The formal way to construct the relevant formulas is simple. Suppose $M:(q,v)\to(Q,V)$ is the mapping, and $u:(q,v)\to(\widetilde{q},\widetilde{v})$ is a transformation which leaves the free field equations invariant, then the transformation $U:(Q,V)\to(\widetilde{Q},\widetilde{V})$, with $U=M\star u\star M^{-1}$, maps solutions of the nonlinear field equations into solutions of the same equations. For example the transformation u (2.16) yields the transformation U given by the explicit expressions,

$$\widetilde{Q}=aQ\ \exp[f(a^2G(V)+b)-F(V)],\quad \widetilde{V}=g(a^2G(V)+b). \tag{2.22}$$

As for the nonlinear superposition principle, the way to proceed is similar; suppose (Q_j,V_j), $j=1,2,...,N$, are N solutions of the nonlinear field equations, and $q_j=Q_j\ \exp[-F(V_j)]$ the corresponding solutions of the K-G equation (2.1), then let q be the linear superposition, $q=\Sigma_j\ a_jq_j$, where the a_j's are arbitrary complex numbers and v is the corresponding solution of (2.4), with (2.3), then the new solution (Q,V) given by the nonlinear superposition of the (Q_j,V_j)'s reads

$$Q=\sum_j\ a_jQ_j\ \exp[F(V)-F(V_j)],\quad V=g(v). \tag{2.23}$$

Deriving a more explicit expression of this formula is a simple task, and is left to the interested reader.

Models for spinor fields can be similarly constructed.[1] Since the formulas for the linearizing transformation are the same as before, we confine ourselves to report below the basic equations. The free field $\varphi(z)=\binom{\varphi_2}{\varphi_1}$ satisfies the Dirac equation (see Sec. I for the notation)

$$\gamma^\mu\varphi_{,\mu}+im\varphi=0, \tag{2.24}$$

while the corresponding current

$$j^\mu=\overline{\varphi}\gamma^\mu\varphi,\quad \overline{\varphi}\equiv\varphi^\dagger\gamma^0 \tag{2.25}$$

satisfies the conservation law (2.2).

Again, it is convenient to introduce the associated scalar field $v(z)$, which satisfies the field equation (2.4). Then, for a given function $f(v)$, with $f(0)=0$, we associate to the free field variables φ and v the new fields ϕ and V through the transformation formulas

$$\phi=\exp[f(v)]\varphi,\quad V=g(v), \tag{2.26}$$

while the inverse transformation reads (compare with (2.5) and (2.6))

$$\varphi=\exp[-F(V)]\phi,\quad v=G(V). \tag{2.27}$$

As before, the four functions f, F, g, and G are related to each other by Eqs. (2.20), together with the condition (2.21), and the new scalar field V satisfies again Eq. (2.12), with $J^\mu=\overline{\phi}\gamma^\mu\phi$, or more explicitly,

$$V_{,\mu}=\epsilon_{\mu\nu}\overline{\phi}\gamma^\nu\phi. \tag{2.28}$$

The field equation for the new spinor ϕ is nonlinear, and takes the expression

$$\gamma^{\mu}\phi_{,\mu}+[im-F'(V)\epsilon_{\mu\nu}\gamma^{\mu}\overline{\phi}\gamma^{\nu}\phi]\phi=0, \tag{2.29}$$

with a cubic nonlinearity in the spinor field. Of course, a two-parameter family of transformations which map a given solution (ϕ,V) of the field equations (2.29) and (2.28) into a new solution $(\tilde{\phi},\tilde{V})$ of the same equations can be found in exactly the same way as for the scalar case, and it looks as in (2.22), namely

$$\tilde{\phi}=a\phi\ \exp[f(a^2G(V)+b)-F(V)], \quad \tilde{V}=g(a^2G(V)+b). \tag{2.30}$$

The same holds for the nonlinear superposition formula of solutions of (2.28) and (2.29), whose derivation does not need to be repeated here.

The generalization of this approach to systems of coupled equations for scalar and spinor fields has been given in Ref. 1. Here, we shortly report on few examples of particular field models, which correspond to simple choices of the function $f(v)$ [or, equivalently, $F(V)$ or $g(v)$ or $G(V)$] which defines the linearizing transformation. First, we observe that a preliminary requirement is that all the four functions f, g, F, and G be explicitly expressed in terms of elementary functions. Second, apart from obvious simplicity criteria, we ask that the mapping $q\to Q$ be either nonsingular or have a simple singularity structure. This point is particularly relevant in the context of the scalar field, as contrasted with the nonrelativistic case, because the scalar function v, which enters in the transformations (2.5) and (2.26), can assume any real value. In fact, modulo an additive constant of integration, v is the (space) integral of the charge density, $v_x=-j_0$, which is definite positive in the nonrelativistic case, $v_x=|\psi|^2>0$, while in the relativistic case $v_x=\mathrm{Im}(q^*q_t)$ may take any real value (see also below). However, in the case of the spinor field, the charge density $j_0=|\varphi_1|^2+|\varphi_2|^2$ is positive and the discussion of the problem of avoiding singularities of the nonlinear field ϕ is simpler than in the scalar one. In order to clearly show the singularities in the variable v, it is convenient to introduce the real function

$$D(v)=\exp[-2f(v)]=1/g'(v) \tag{2.31}$$

so that the transformation of the charge density reads [see (2.10)]

$$J_0=j_0/D(v);$$

in this notation, the singularities of J_0 may arise only from the zeros of the function $D(v)$ whenever j_0 is nonsingular. We also point out that, because of (2.21), $D(v)$ should be chosen so as to satisfy the condition

$$D(0)=1. \tag{2.32}$$

Example A: $D(v)$ has only one zero,

$$D(v)=(1+av)^{\alpha/a}, \quad a\alpha>0, \tag{2.33}$$

$$f(v)=-\frac{\alpha}{2a}\ln(1+av), \quad g(v)=\frac{1}{a-\alpha}[(1+av)^{1-\alpha/a}-1], \tag{2.33a}$$

$$F(V)=-\frac{\alpha}{2(a-\alpha)}\ln[1+(a-\alpha)V], \quad G(V)=\frac{1}{a}\{[1+(a-\alpha)V]^{a/(a-\alpha)}-1\}, \tag{2.33b}$$

$$F'(V)=-\frac{\alpha}{2}[1+(a-\alpha)V]^{-1}, \quad F''(V)-(F'(V))^2=\frac{\alpha}{4}(2a-3\alpha)[1+(a-\alpha)V]^{-2}. \tag{2.33c}$$

Example A_1: this is the $a \to 0$ limit of Example A, $D(v)$ has no zero for finite v,

$$D(v) = \exp(\alpha v), \tag{2.34}$$

$$f(v) = -\frac{\alpha}{2} v, \quad g(v) = -\frac{1}{\alpha} [\exp(-\alpha v) - 1], \tag{2.34a}$$

$$F(V) = \frac{1}{2} \ln(1 - \alpha V), \quad G(V) = -\frac{1}{\alpha \ln(1 - \alpha V)}, \tag{2.34b}$$

$$F'(V) = -\frac{\alpha}{2} (1 - \alpha V)^{-1}, \quad F''(V) - (F'(V))^2 = -\frac{3}{4} \alpha^2 (1 - \alpha V)^{-2}. \tag{2.34c}$$

Example A_2: this is the $a \to \alpha$ limit of Example A,

$$D(v) = 1 + av, \tag{2.35}$$

$$f(v) = -\frac{1}{2} \ln(1 + av), \quad g(v) = \frac{1}{a} \ln(1 + av), \tag{2.35a}$$

$$F(V) = -\frac{a}{2} V, \quad G(V) = \frac{1}{a} [\exp(aV) - 1], \tag{2.35b}$$

$$F'(V) = -\frac{a}{2}, \quad F''(V) - (F'(V))^2 = -\frac{a^2}{4}. \tag{2.35c}$$

Example B_1: $D(v)$ has two complex zeros,

$$D(v) = (a^2 v^2 + 1)^{1/2}, \quad a > 0, \tag{2.36}$$

$$f(v) = -\frac{1}{4} \ln(a^2 v^2 + 1), \quad g(v) = \frac{1}{a} \ln \sqrt{1 + a^2 v^2} + av), \tag{2.36a}$$

$$F(V) = -\frac{1}{2} \ln[\cosh(aV)], \quad G(V) = \frac{1}{a} [\sinh(aV)], \tag{2.36b}$$

$$F'(V) = -\frac{a}{2} \operatorname{tgh}(aV), \quad F''(V) - (F'(V))^2 = -\frac{a^2}{4} [2 - \operatorname{tgh}^2(aV)]. \tag{2.36c}$$

Example B_2: $D(v)$ has two complex zeros,

$$D(v) = (a^2 v^2 + 1), \tag{2.37}$$

$$f(v) = -\frac{1}{2} \ln(a^2 v^2 + 1), \quad g(v) = \frac{1}{a} \operatorname{arctg}(av), \tag{2.37a}$$

$$F(V) = \ln \cos(aV), \quad G(V) = \frac{1}{a} \operatorname{tg}(aV), \tag{2.37b}$$

$$F'(V) = -\operatorname{atg}(aV), \quad F''(V) - (F'(V))^2 = -a^2 [1 + 2 \operatorname{tg}^2(aV)]. \tag{2.37c}$$

Example B_3: $D(v)$ has two complex zeros,

$$D(v)=(a^2v^2+1)^{3/2}, \tag{2.38}$$

$$f(v)=-\frac{3}{4}\ln(a^2v^2+1), \quad g(v)=\frac{v}{\sqrt{a^2v^2+1}}, \tag{2.38a}$$

$$F(V)=\frac{3}{4}\ln(1-a^2V^2), \quad G(V)=\frac{V}{\sqrt{1-a^2V^2}}, \tag{2.38b}$$

$$F'(V)=-\frac{3}{2}a^2V/(1-a^2V^2), \quad F''(V)-(F'(V))^2=-\frac{3a^2}{2}\left(1+\frac{5}{2}a^2V^2\right)\Big/(1-a^2V^2)^2. \tag{2.38c}$$

It should be noted that only in the models A and A_2, $D(v)$ vanishes on the real axis, i.e., for $v=-1/a$, and that only with the choice A_2 the field equations (2.13) in the scalar case, and (2.29) in the spinor case, are not coupled to the scalar field V; these equations coincide with (1.1) and (1.2) for $a=-2$ i.e., $F'(V)=1$).

III. SPECIAL SOLUTIONS-REAL SPECTRUM

In this section we describe few properties of particular solutions of the nonlinear field equations introduced in the previous section. The main emphasis here is on our capability to write down explicit expressions of solutions of nonlinear PDEs by transforming explicit expressions of the linear free field, whose building blocks are, of course, the elementary exponential solutions. Before proceeding to this task, however, we briefly look at the initial value problem associated with our field equations.

For the free scalar field, this is the problem of computing the solution $q(x,t)$ of the K-G equation (2.1), and the solution $v(x,t)$ of the auxiliary equation (2.4), which satisfy the conditions

$$q(x,0)=q_0(x), \quad q_t(x,0)=q_1(x), \quad x\epsilon\Re, \tag{3.1a}$$

$$v(x,0) \xrightarrow[x\to-\infty]{} l; \tag{3.1b}$$

both choices of the initial time $t=0$, and of the point $x=-\infty$ in (3.1b), are made for the sake of convenience. If $q_0(x)$ and q_1 are assumed to be in $L^2(\Re)$, then the solution $q(x,t)$ has the Fourier integral representation

$$q(x,t)=\int_{-\infty}^{+\infty}\frac{dk}{\sqrt{2\pi}}\{q+(k)\exp[i(kx-\omega t)]+q-(k)\exp[i(kx+\omega t)]\}, \tag{3.2}$$

where the complex amplitudes $q\pm(k)$ are easily expressed in terms of the Fourier transforms of the initial values (3.1a). Once the spectral integral (3.2) is performed, the potential function $v(x,t)$ obtains by quadrature, namely,

$$v(x,t)=\int_{-\infty}^{x}dx'\ \mathrm{Im}[q^*(x',t)q_t(x',t)]+l. \tag{3.3}$$

The point we wish to make here is that the similar initial value problem for the transformed nonlinear equation (2.13) cannot, in general, be mapped into the problem (3.1) with L^2 initial values (3.1a). Indeed, consider the problem of finding the solutions $Q(x,t)$ and $V(x,t)$ of, respectively, (2.13) and (2.12) with the initial values

$$Q(x,0)=Q_0(x), \quad Q_t(x,0)=Q_1(x), \quad x\epsilon\mathfrak{R} \tag{3.4a}$$

in $L^2(\mathfrak{R})$, and the condition

$$V(x,0) \xrightarrow[x\to-\infty]{} L. \tag{3.4b}$$

Then, these conditions can be mapped in conditions for the corresponding free field q according to the transformation (2.6), namely,

$$\begin{aligned} q_0(x)&=Q_0(x)\exp[-F(V_0(x))], \\ q_1(x)&=[Q_1(x)-F'(V_0(x))V_1(x)Q_0(x)]\exp[-F(V_0(x))], \end{aligned} \tag{3.5}$$

where we have set, for notational convenience,

$$\begin{aligned} V_0(x)&=V(x,0)=L+\int_{-\infty}^{x} dx'\ \mathrm{Im}[Q_0^*(x')Q_1(x')], \\ V_1(x)&=V_t(x,0)=\mathrm{Im}[Q_0^*(x)Q_{0x}(x)]. \end{aligned} \tag{3.6}$$

As it will be shown below, even if the initial values Q_0 and Q_1 are L^2 functions, the transformation factor $\exp[-F(V_0)]$, see (3.5), may well be singular as $|x|\to\infty$ so that the spectral representation (3.2) of the corresponding free field is not valid. This property of integrable nonlinear equations is well-known to occurr in both cases of S-integrability and C-integrability.[1] The obvious reason for this property is that there may exist solutions of the free field q whose spectral representation is given by an integral in the spectral variable k which runs, in contrast with the Fourier case (3.2), off the real axis, while yielding, via the mapping (2.5), L^2 solutions Q. A detailed discussion of this problem has been given in Ref. 3 for the Eckhaus equation, which is the nonrelativistic analog of the scalar equation (1.1), and it is not repeated here. Similar considerations apply also to the spinor field, namely, to Eq. (2.29).

At any rate, the method of solution of initial value problems is outside the scope of this paper, and, in fact, it requires separate consideration of each of the field models described in the previous section.

We now turn our attention to the simpler task of constructing solutions of the nonlinear equations by transforming known solutions of the free field. Moreover, in the following, the properties of a given nonlinear field will be described by considering the associated charge density, both in the scalar and spinor case, rather than the field solution itself. This is particularly justified whenever we show plots of x-profiles. If the integrals defining the total charge c of the free field,

$$c=\int_{-\infty}^{+\infty} dx j_0(x,t) \tag{3.7}$$

and the total charge C of the nonlinear field,

$$C=\int_{-\infty}^{+\infty} dx J_0(x,t), \tag{3.8}$$

exist, then the relevant formulas follow from (2.4) and (2.12), and read

$$c=l-r, \quad C=L-R=g(l)-g(r), \tag{3.9}$$

where we use the asymptotic conditions (3.1b) and (3.4b), together with the notation

$$v(x,t) \xrightarrow[x\to+\infty]{} r, \quad V(x,t) \xrightarrow[x\to+\infty]{} R, \tag{3.10}$$

which defines the constants r and R, and the relation $V=g(v)$ (see (2.5) or (2.26)). In this respect, we also note that it may well happen that the free field charge c is infinite, while the corresponding charge C is finite (see below).

Because of linearity, the free field solutions are combinations of the elementary solutions (1.18), which, on the real spectrum ($\theta=0$), read

$$E(k,\eta)=\exp(-i\eta k^{\mu}x_{\mu})=\exp[i\eta(kx-\omega t)], \quad \eta=\pm 1, \tag{3.11}$$

for the scalar field equation (2.1), and (1.27) (with $\theta=0$)

$$\epsilon(k,\eta)=E(k,\eta)\chi(k,\eta), \quad \eta=\pm 1, \tag{3.12a}$$

$$\chi(k,\eta)=\frac{1}{\sqrt{8}}[(1+\eta)+(1-\eta)\gamma^{1}]\begin{pmatrix}\sqrt{\omega+m}\\ \eta k/\sqrt{\omega+m}\end{pmatrix}, \quad \eta=\pm 1, \tag{3.12b}$$

for the spinor field equation (2.4); here k^{μ} is the timelike vector (1.16).

A. Brachions and tachions

Let us consider first the scalar field. The simplest solutions of the K-G equation (2.1) obtain as linear combinations of a finite number of exponentials of the form (3.11). However, the corresponding nonlinear field solutions may be singular as a consequence of the particular mapping (2.5) one considers. Therefore, for this class of solutions, we limit our consideration to the models B_2 and B_3 we have previously introduced (see (2.37) and (2.38)), since these are the only ones to possess regular and finite-charge solutions. Moreover, in order to show the main features of these solutions, it is sufficient to consider the sum of two exponentials, namely,

$$q(z)=a_1E(k_1,\eta_1)+a_2E(k_2,\eta_2), \tag{3.13}$$

because the extension to the generic sum of N exponentials does not take into play any novel feature. The scalar field $v(z)$ corresponding to (3.13) obtains by integrating the differential equation (2.4), where the current j^{μ} is given by (2.3), i.e.,

$$j^{\mu}(z)=k^{\mu}+(\eta_1k_1^{\mu}+\eta_2k_2^{\mu})\mathrm{Re}[a_1a_2^{*}E(k_1,\eta_1)E^{*}(k_2,\eta_2)], \tag{3.14}$$

and it is easily found to have the expression

$$v(z)=\epsilon_{\mu\nu}x^{\mu}k^{\nu}+\sigma[(\mu+m^2)/(\mu-m^2)]^{1/2}\,\mathrm{Im}[a_1a_2^{*}E(k_1;\eta_1)E^{*}(k_2,\eta_2)]+\overline{v,} \tag{3.15}$$

where we have introduced the vector

$$k^{\mu}=\eta_1|a_1|^2k_1^{\mu}+\eta_2|a_2|^2k_2^{\mu}, \tag{3.16}$$

the relativistic invariant

$$\mu=\eta_1\eta_2g_{\mu\nu}k_1^{\mu}k_2^{\nu} \tag{3.17}$$

and the sign $\sigma, \sigma^2 = 1$, according to the formulas

$$\lambda = \eta_1 \eta_2 \epsilon_{\mu\nu} k_1^\mu k_2^\nu, \quad \sigma = \text{sign } \lambda, \tag{3.18}$$

while $\overline{v}$ is a real integration constant. The derivation of expression (3.15) comes from the observation that, for any two real vectors k_1^μ and k_2^μ on the same branch of the mass-shell, see (1.16), the relation $\mu^2 = \lambda^2 + m^4$ holds together with the identity

$$\epsilon_{\mu\nu}(\eta_1 k_1^\nu + \eta_2 k_2^\nu) = \sigma[(\mu + m^2)/(\mu - m^2)]^{1/2}(\eta_2 k_{2\mu} - \eta_1 k_{1\mu}), \tag{3.19}$$

which is a special case of (1.19a) (i.e., for $\theta_1 = \theta_2 = 0$). The nonlinear equations are (2.12) and (2.13), with the specifications (2.37c) for the model B_2 and (2.38c) for the model B_3. The corresponding expressions for the nonlinear fields Q and V explicitly read

$$Q = q/(1 + a^2 v^2)^{1/2}, \quad V = \frac{1}{a} \operatorname{arctg}(av), \tag{3.20}$$

for B_2, and

$$Q = q/(1 + a^2 v^2)^{3/4}, \quad V = v/(1 + a^2 v^2)^{1/2}, \tag{3.21}$$

for B_3, while the current J^μ (see (2.8)) is given by the general formula

$$J^\mu = j^\mu / D(v), \tag{3.22}$$

see (2.10) and (2.31), with the specifications (2.37) and (2.38) for, respectively, B_2 and B_3.

Let us consider first the simpler case in which only one exponential is present in (3.13), namely let us set $a_1 = A$, $a_2 = 0$, $k_1^\mu = k^\mu$, $\eta_1 = \eta$. Moreover, since k^μ is time-like, by performing a Lorentz transformation we can set $k = 0$, i.e., $k^\mu = (m, 0)$, without loss of generality. In this case, also the currents j^μ and J^μ are time-like, and the charge density of the nonlinear field is

$$J^0 = \eta m |A|^2 / D(v), \tag{3.23}$$

with

$$v(x,t) = -\eta m |A|^2 x + \overline{v}. \tag{3.24}$$

The corresponding charge readily obtains from the general formula (3.9) (with $l = \eta\infty$ and $r = -\eta\infty$), and reads

$$C = \eta C_0, \tag{3.25}$$

where the elementary charge C_0 takes the values

$$C_0 = \frac{\pi}{a} \quad \text{for } B_2, \quad C_0 = \frac{2}{a} \quad \text{for } B_3. \tag{3.26}$$

Since the current J^μ is conserved, and it vanishes as $x \to \pm\infty$ for any fixed value of t and in any frame of reference, the charge C is relativistically invariant. We conclude, therefore, that this solution of the nonlinear field models B_2 and B_3 describes a (algebraically) localized object, whose (''quantized'') charge may have two apposite values that, remarkably enough, do not depend on the amplitude A and are given only in terms of the structure constant a of the model.

We consider next the generic solution corresponding to the free field (3.13). In this case the norm of the vector k^μ, see (3.16), may take any value, according to the value of the invariant μ given by (3.17). In fact, since

$$k^\mu k_\mu = m^2(\alpha_1^2+\alpha_2^2)+2\mu\alpha_1\alpha_2, \quad \alpha_j \equiv |a_j|^2, \tag{3.27}$$

and

$$\mu = \eta_1\eta_2(\omega_1\omega_2 - k_1k_2), \tag{3.28}$$

it is convenient to separate the case in which the two exponentials in (3.13) are in the same branch, namely $\eta_1\eta_2=1$, from that in which they are in different branches, namely $\eta_1\eta_2=-1$.

In the first case, let us set $\eta_1=\eta_2=\eta$; then

$$\mu = \omega_1\omega_2 - k_1k_2 \geqslant m^2 \tag{3.29}$$

and the vector k^μ is time-like for any choice of k_1, k_2, a_1, and a_2. Therefore, without loss of generality, we can choose a frame of reference in which $\alpha_1k_1+\alpha_2k_2=0$, or, equivalently, $k^\mu = \eta(\alpha_1\omega_1+\alpha_2\omega_2,0)$ so that the scalar field v takes the expression

$$v(x,t) = -\eta(\alpha_1\omega_1+\alpha_2\omega_2)x + \sigma[\alpha_1\alpha_2(\mu+m^2)/(\mu-m^2)]^{1/2}\cos\,\theta(x,t)+\overline{v}, \tag{3.30}$$

where the explicit expression of the phase $\theta(x,t)$ can be easily derived from (3.15). Since also in this case we have that $l=\eta\infty$ and $r=-\eta\infty$, for the same argument as before, the corresponding charge is again relativistically invariant and has the same value (3.25). Moreover, as it is easily verified, there are values of the parameters k_1, k_2, a_1, a_2, and η such that the currents j^μ, and, therefore, J^μ, are time-like in the whole space-time, or time-like in some regions and space-like in other regions (but, for no choice of the parameters, space–like in the whole space-time). Our conclusion on this case is that this solution is again algebraically confined in space, with a parameter-independent charge which may take the two values (3.25). Adding one exponential to the previous free field q has, therefore, given only some internal motion through the time-dependence of the phase function θ that appears in the expression (3.30) (yielding a sort of breather). A plot of the corresponding charge density J^0 (see (3.14) and (3.22)) for the model B_2 is provided in Fig. 1.

In the second case, i.e. for $\eta_1=-\eta_2=\eta$, since

$$\mu = -(\omega_1\omega_2 - k_1k_2) \geqslant -m^2, \tag{3.31}$$

the vector k^μ (see (3.16)) is time-like if

$$-m^2\left(\frac{\alpha_1^2+\alpha_2^2}{2\alpha_1\alpha_2}\right) < \mu, \tag{3.32}$$

and is space-like if

$$\mu < -m^2\left(\frac{\alpha_1^2+\alpha_2^2}{2\alpha_1\alpha_2}\right). \tag{3.33}$$

If the parameters a_1, a_2, k_1, and k_2 are chosen so as to satisfy (3.32), then the vector $\epsilon_{\mu\nu}\,k^\nu$, which is orthogonal to the time-like vector k^μ, is, of course, space-like, and the corresponding nonlinear field solution has the same properties as in the previous case; there exist a frame of reference in which $k^\mu=\eta(\alpha_1\omega_1-\alpha_2\omega_2,0)$, the total charge is invariant and takes two values according to the formula

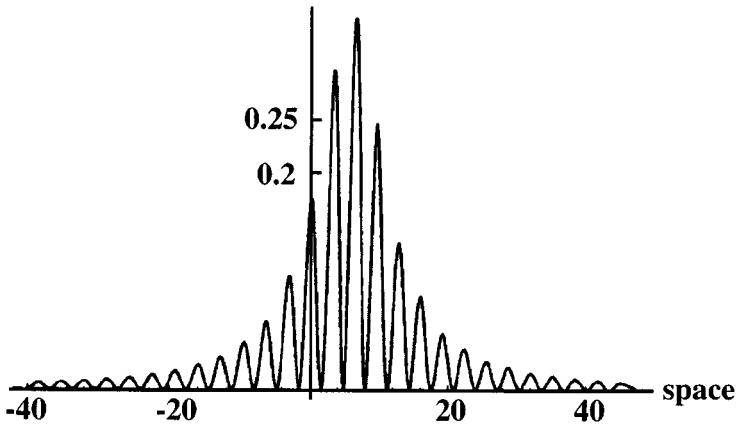

FIG. 1. Brachion profile. The charge density $J^0(x,t)$ is plotted against the space coordinate x at $t=10$ for the scalar model B_2 with $a=1$, $m=1$. The parameters are $\eta_1=\eta_2=1$, $\tau_1=-\tau_2=1$, $\theta_1=\theta_2=0$, $\beta_1=-\beta_2=7/10$, $a_1=a_2=0.1$, $\overline{v}=0$ and the charge takes the elementary value $C_0=\pi$.

$$C=\eta C_0\ \mathrm{sign}(\alpha_1\omega_1-\alpha_2\omega_2). \tag{3.34}$$

The only difference, with respect to the previous case, is that the sign of the charge depends on the relative amplitude $|a_1|^2/|a_2|^2$ and on the spectral variable, say k_1, as explicitely shown in (3.34). Again, for any value of the invariant μ and amplitudes $|a_1|$ and $|a_2|$, there are space–time regions (which may invade the whole space–time) where the current J^μ (and, of course, j^μ, see (3.22)) is time-like.

Let us finally choose k_1 and k_2 so as to satisfy the inequality (3.33). Now the vector k^μ is space-like while $\epsilon_{\mu\nu}k^\nu$ is time-like, and this implies that, in a generic frame of reference, the localized profile of $J^0(x,t)$ moves with the velocity

$$\beta=(\alpha_1k_1-\alpha_2k_2)/(\alpha_1\omega_1-\alpha_2\omega_2), \tag{3.35}$$

whose modulus is larger than the speed of light,

$$|\beta|>1. \tag{3.36}$$

We may call this solution a ''tachion'' and the previous solutions (with velocity smaller than the light-speed) ''brachions,'' just to emphasize the difference between these two classes of solutions. Tachions behave very differently with respect to ordinary particles. In the present example, there is a frame of reference in which the tachion moves with infinite speed, this being the case when (see (3.35)),

$$\alpha_1\omega_1=\alpha_2\omega_2. \tag{3.37}$$

Moreover, in this coordinate system the tachion is completely delocalized since the function $v(x,t)$ becomes periodic in the spatial variable x,

$$v(x,t)=\eta(\alpha_1k_1-\alpha_2k_2)t+\sqrt{\alpha_1\alpha_2}[(m^2+\mu)/\lambda]\cos\ \theta(x,t)+\overline{v}. \tag{3.38}$$

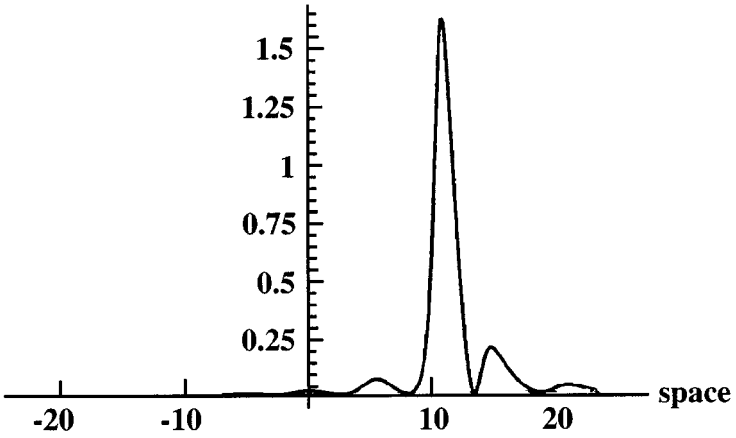

FIG. 2. Tachion profile. The charge density $J^0(x,t)$ is plotted against the space coordinate x at $t=4$ for the scalar model B_2 with $a=1$, $m=1$. The parameters are $\eta_1=-\eta_2=1$, $\tau_1=-\tau_2=1$, $\theta_1=\theta_2=0$, $\beta_1=\sqrt{3}/2$, $\beta_2=-1/2$, $a_1=a_2=1$, $\overline{v}=0$. The absolute value of the charge is $C_0=\pi$ and the tachion velocity is $\beta=2.732$.

Because of this property, the corresponding current does not asymptotically vanish as $x\to\pm\infty$ in all frames of reference, and, therefore, the total charge is no more relativistically invariant. Indeed, while the charge C takes only the two elementary values $\pm C_0$, according to the general formula (3.34), its sign depends on the particular frame of reference since the quantity $\alpha_1\omega_1-\alpha_2\omega_2$ may change its sign by a Lorentz transformation. In fact, the jump of the charge sign obtains by a Lorentz transformation that changes the sign of the tachion velocity (compare (3.35) with (3.34)). The profile of a tachion is plotted in Fig. 2.

This analysis can be easily extended to solutions originated by a sum of N, rather than just two (see (3.3)), exponentials. The following general conclusion can be drawn by looking at the vector $k^\mu\equiv\Sigma\,\eta_n|a_n|^2k_n^\mu$; the corresponding nonlinear field solutions are localized brachions if the vector k^μ is time-like, and tachions if, instead, k^μ is space-like. For the tachion solutions, the charge density profile is localized, with the exception of the frame of reference in which it moves infinitely fast, where it is quasiperiodic in the coordinate x. The absolute value of the charge of all these solutions is always $|C_0|$, independently of N and their associated parameters a_n, k_n, η_n, while the charge itself is Lorentz invariant only for brachions. We finally observe that the associated scalar field $V(x,t)=g(v)$, for the models under consideration B_2 and B_3, has the shape of a modulated kink whose jump $V(-\infty,t)-V(+\infty,t)$ is the total charge C, see (3.9).

Nonlinear spinor fields corresponding to sums of elementary solutions (3.12) can be similarly constructed. We now briefly report on this class of solutions by pointing out the main difference with respect to the scalar case. This stems from the fact that the charge density j^0, and therefore J^0, is positive definite for any solution $\varphi=\binom{\varphi_2}{\varphi_1}$ of the Dirac equation (2.24), namely,

$$j^0=|\varphi_1|^2+|\varphi_2|^2, \tag{3.39}$$

and, therefore, all finite charge solutions of spinor field equations have positive charge. The point here is that our spinor field is a classical commuting variable and, in contrast with the quantized theory of anticommuting fields, there is no spin-statistics relation.

Let us start again with the free field solution [see (3.12) and (3.11)]

$$\varphi(z)=a_1E(k_1,\eta_1)\chi(k_1,\eta_1)+a_2E(k_2,\eta_2)\chi(k_2,\eta_2), \tag{3.40}$$

whose corresponding current (see (2.25)) has the expression

$$j^\mu(z)=k^\mu+2V^\mu\ \mathrm{Re}[a_1a_2^*E(k_1,\eta_1)E^*(k_2,\eta_2)]. \tag{3.41}$$

Here the vector k^μ takes the expression

$$k^\mu=|a_1|^2k_1^\mu+|a_2|^2k_2^\mu; \tag{3.42}$$

as it follows from the fact that the spinor $\chi(k,\eta)$ has been defined by (3.12) so as to satisfy the normalization condition (see Sec. I),

$$\overline{\chi(}k,\eta)\gamma^\mu\chi(k,\eta)=k^\mu=(\omega,k). \tag{3.43}$$

Moreover, the vector V^μ in (3.41), that is

$$V^\mu=\overline{\chi(}k_1,\eta_1)\gamma^\mu\chi(k_2,\eta_2), \tag{3.44}$$

has the following expression [as a special case of formula (1.33)],

$$V^\mu=m[2(m^2+\mu)]^{-1/2}(k_1^\mu+k_2^\mu),\quad \text{if}\ \ \eta_1=\eta_2=\eta, \tag{3.45a}$$

$$V^\mu=(m/\lambda)[(m^2-\mu)/2]^{1/2}(k_2^\mu-k_1^\mu),\quad \text{if}\ \ \eta_1=-\eta_2=\eta, \tag{3.45b}$$

where the relativistic invariants μ and λ are defined by (3.17) and (3.18).

Finally, integrating Eq. (2.4), with (3.41), yields the scalar field

$$v(z)=\epsilon_{\mu\nu}\,x^\mu k^\nu+\eta(m/\lambda)[2(m^2+\mu)]^{1/2}\ \mathrm{Im}[a_1a_2^*E(k_1,\eta_1)E^*(k_2,\eta_2)]+\overline{v},\quad \text{if}\ \ \eta_1=\eta_2=\eta, \tag{3.46a}$$

$$v(z)=\epsilon_{\mu\nu}\,x^\mu k^\nu-\eta m[2/(m^2-\mu)]^{1/2}\ \mathrm{Im}[a_1a_2^*E(k_1,\eta_1)E^*(k_2,\eta_2)]+\overline{v},\quad \text{if}\ \ \eta_1=-\eta_2=\eta, \tag{3.46b}$$

where $\overline{v}$ is an integration constant.

Similarly to the scalar case, finite charge solutions, corresponding to the free field (3.40), are obtained for the nonlinear models B_2 and B_3 only. Therefore we consider the nonlinear equations (2.28) and (2.29), with (2.37c) for B_2 and (2.38c) for B_3, whose solutions are explicitly given by the transformation

$$\phi=\varphi/(1+a^2v^2)^{1/2},\quad V=\frac{1}{a}\,\mathrm{arctg}(av), \tag{3.47}$$

and, respectively,

$$\phi=\varphi/(1+a^2v^2)^{3/4},\quad V=v/(1+a^2v^2)^{1/2}, \tag{3.48}$$

while the current J^μ is provided by (3.22), with (2.37) for B_2 and (2.38) for B_3.

The properties of these solutions are rather simple, since the vector k^μ, in contrast with the scalar field models, is timelike for any choice of the parameters a_1, a_2, k_1, k_2, η_1, and η_2. As a consequence, there exists a frame of reference in which

$$k^\mu=(|a_1|^2\omega_1+|a_2|^2\omega_2,0), \tag{3.49}$$

and the spinor field ϕ models a localized object at rest.

This is easily shown by the expression of the charge density $J^0=j^0/D(v)$, where the field v is of the form

$$v(x,t)=-x(|a_1|^2\omega_1+|a_2|^2\omega_2)+\alpha\ \cos\ \theta(x,t)+\overline{v}, \tag{3.50}$$

the expressions of the constant α, and of the phase function $\theta(x,t)$, being easily derived from formulas (3.46). Thus we conclude that these solutions describe an elementary spinorial object (namely no break-up process occurs during the time evolution) which travels no faster than light (brachions) with the same positive elementary, relativistically invariant, charge C_0 given by (3.26) (the localized x-profile of the charge density looks as in Fig. 1). Indeed, these same conclusions hold for solutions corresponding to a sum of any finite number of elementary spinors (3.12), these being localized brachions with the parameter-independent elementary charge C_0.

B. Wave packets

Here we consider the nonlinear field solutions which obtain by transforming the wave packet solutions of the free field. The following analysis is aimed to provide few asymptotic properties of such solutions as $t\to\pm\infty$. In particular, we will focus our attention on the asymptotic expression of the velocity of the center of charge, and on the leading term of the charge density $J^0(x,t)$ in the direction $x=\beta t$, with $-1\leqslant\beta\leqslant 1$. Since the computational technique, which is based on the Fourier analysis of the free field solutions, is standard, the derivation of the final expression is not detailed.

Let us consider first the scalar field. The free field wave packet is given by the integral (3.2) in terms of two complex amplitudes $q_+(k)$ and $q_-(k)$, which we assume to be rapidly decreasing functions of the real variable k. These amplitudes are, respectively, related to the positive and negative charge components of the field, as shown by the integral expression of the total charge c, see (3.7), that is

$$c=\int_{-\infty}^{+\infty}dk\ \omega(k)[|q_+(k)|^2-|q_-(k)|^2]. \tag{3.51}$$

The behavior of the free field is well-known to be simple. In fact, the center of charge, with coordinate $x_c(t)$, defined by the formula

$$cx_c(t)=\int_{-\infty}^{+\infty}dx\ xj^0(x,t), \tag{3.52}$$

moves with the velocity $\dot{x}_c(t)\equiv s(t)$, whose expression is provided by the conservation law (2.2), and reads

$$cs(t)=\int_{-\infty}^{+\infty}dx\ j^1(x,t). \tag{3.53}$$

The right-hand side integral can be easily related to the Fourier amplitudes, so that the center of charge velocity takes the alternative expression

$$cs(t)=\int_{-\infty}^{+\infty}dk\ k\{(|q_+(k)|^2+|q_-(k)|^2)+2\ \mathrm{Re}[q_+^*(k)q_-(k)\exp(2i\omega t)]\}. \tag{3.54}$$

It then follows that the free field wave packet asymptotically moves with the same velocity s_a,

$$s_a=\frac{1}{c}\int_{-\infty}^{+\infty} dk\ k[|q_+(k)|^2+|q_-(k)|^2], \tag{3.55}$$

both in the past ($t\to-\infty$) and in the future ($t\to+\infty$). This is not so for the corresponding nonlinear field wave packet, which may indeed display a boomerang behavior as its asymptotic velocities $S_{\text{in}}(t=-\infty)$ and $S_{\text{out}}(t=+\infty)$ have apposite sign. In order to show this effect, let us give first the asymptotic behaviors of the current $j^\mu(x,t)$, and of the associated scalar field $v(x,t)$, as $t\to\pm\infty$ in the direction $x=\beta t$, with $-1<\beta<1$. These can be derived from the well-known leading term of the free field, that is

$$q(\beta t,t)\xrightarrow[t\to\pm\infty]{} (m|t|)^{-1/2} q^{\text{out}}_{\text{in}}(\beta,t), \tag{3.56a}$$

$$q^{\text{out}}_{\text{in}}(\beta,t)=m\gamma^{3/2}[q_+(m\beta\gamma)\exp(-i\alpha_\pm)+q_-(-m\beta\gamma)\exp(i\alpha_\pm)], \tag{3.56b}$$

$$\alpha_\pm=mt/\gamma\pm\frac{\pi}{4},\quad \gamma=(1-\beta^2)^{-1/2}. \tag{3.56c}$$

In fact, inserting these expressions into the definition (2.3), we obtain the leading term of the current j^μ, namely,

$$j^\mu(\beta t,t)\xrightarrow[t\to\pm\infty]{} |t|^{-1} j^\mu_a(\beta), \tag{3.57a}$$

$$j^\mu_a(\beta)=m^2\gamma^4[|q_+(m\beta\gamma)|^2-|q_-(-m\beta\gamma)|^2](1,\beta), \tag{3.57b}$$

in the usual notation $j^\mu=(j^0,j^1)$. As for the scalar field $v(x,t)$, its expression in terms of the Fourier amplitudes of the free field obtains by inserting (3.2) in (3.3); we leave out the simple, and tedious, steps of this computation and display only the following asymptotic formula

$$v(\beta t,t)\xrightarrow[t\to\pm\infty]{} v^{\text{out}}_{\text{in}}(\beta), \tag{3.58a}$$

$$v^{\text{out}}_{\text{in}}(\beta)=\mp\frac{1}{2}\int_{-\infty}^{+\infty} dk\ \omega\left[\operatorname{sign}\left(\beta-\frac{k}{\omega}\right)|q_+(k)|^2-\operatorname{sign}\left(\beta+\frac{k}{\omega}\right)|q_-(k)|^2\right]+\overline{v}. \tag{3.58b}$$

For the sake of completeness, we note that the integration constant $\overline{v}$, which appears in this expression, is related to the asymptotic constants l and r, see (3.3) and (3.10), by the equations

$$l=\tfrac{1}{2}c+\overline{v},\quad r=-\tfrac{1}{2}c+\overline{v}. \tag{3.59}$$

Therefore, the total charge C, see (3.8), of the nonlinear field can be conveniently computed via (3.9), namely,

$$C=g(\tfrac{1}{2}c+\overline{v})-g(-\tfrac{1}{2}c+\overline{v}). \tag{3.60}$$

We are now in the position to derive the asymptotic behaviors of both the current $J^\mu(\beta t,t)$ of the nonlinear field in the direction $x=\beta t$, and the velocity of the center of charge, whose position $X_c(t)$ is, of course,

$$X_c(t)=\frac{1}{C}\int_{-\infty}^{+\infty} dx\ xJ^0(x,t). \tag{3.61}$$

In fact, the general transformation law (3.22), together with (3.57) and (3.58), imply that

$$J^{\mu}(\beta t,t) \xrightarrow[t\to\pm\infty]{} |t|^{-1} J^{\mu}_{\substack{\text{out}\\ \text{in}}}(\beta), \tag{3.62a}$$

$$J^{\mu}_{\substack{\text{out}\\ \text{in}}}(\beta) = j^{\mu}_{a}(\beta)/D(v^{\text{out}}_{\text{in}}), \tag{3.62b}$$

whereas the asymptotic values of the nonlinear wave packet velocity $S(t)\equiv \dot{X}_c(t)$, whose expression is (compare with (3.53))

$$CS(t) = \int_{-\infty}^{+\infty} dx J^1(x,t), \tag{3.63}$$

can be derived as follows. We rewrite this last equation as

$$CS(t) = \int_{-\infty}^{-|t|} dx J^1(x,t) + \int_{|t|}^{+\infty} dx J^1(x,t) + |t| \int_{-1}^{1} d\beta J^1(\beta t,t), \tag{3.64}$$

and take the limits as $t\to\pm\infty$; then, using (3.62) yields the formula

$$S(t) \xrightarrow[t\to\pm\infty]{} S^{\text{out}}_{\text{in}} = \frac{1}{C} \int_{-1}^{1} d\beta j^1_a(\beta)/D(v^{\text{out}}_{\text{in}}). \tag{3.65a}$$

For computational purpose, we deem it convenient to rewrite some of the expressions given above in the following alternative forms:

$$j^{\mu}_{a}(\beta) = \frac{\omega^3}{m^2} [|q_+(k)|^2 - |q_-(-k)|^2](\omega,k), \tag{3.57c}$$

$$v^{\text{out}}_{\text{in}}(\beta) = \mp\frac{1}{2} \left(\int_{-\infty}^{k} dk' - \int_{k}^{+\infty} dk' \right) \omega(k')[|q_+(k')|^2 + |q_-(-k')|^2] + \overline{v}, \tag{3.58c}$$

$$S^{\text{out}}_{\text{in}} = \frac{1}{C} \int_{-\infty}^{+\infty} dk k[|q_+(k)|^2/D(v^{\text{out}}_{\text{in}}(\beta)) + |q_-(k)|^2/D(v^{\text{out}}_{\text{in}}(-\beta))], \tag{3.65b}$$

with the obvious relation between the variables β and k

$$k = m\gamma\beta, \quad \omega(k) = m\gamma, \quad \beta = k/\omega. \tag{3.66}$$

As already pointed out, we note also here that even a well-behaved free field wave packet q, see (3.2), may yield a singular nonlinear field Q. The singularities correspond, of course, to the zeros of the function $D(v)$, see (3.22). For the sake of simplicity, we avoid this problem by considering only those models whose corresponding function $D(v)$ has no finite real zero; these are the models A_1, see (2.34), B_1, see (2.36), B_2, see (2.37) and B_3, see (2.38). For these models, to any free field wave packet there corresponds a nonsingular, finite charge, solution of the nonlinear field equations. Let us now show the boomerang effect. Consider a free field solution in the frame of reference in which the velocity of its center of charge asymptotically vanishes as $t\to\pm\infty$. This is the case if we assume that

$$|q_{\pm}(-k)|^2 = |q_{\pm}(k)|^2, \tag{3.67}$$

as this implies that $s_a=0$, see (3.55). It is then convenient to rewrite the expression (3.58c) in the following form

$$v_{\text{in}}^{\text{out}}=\overline{v\mp}\,\omega(k), \tag{3.68}$$

with, by taking into account (3.67),

$$w(k)=\frac{1}{2}\left(\int_{-\infty}^{k}dk'-\int_{k}^{+\infty}dk'\right)\omega(k')[|q_+(k')|^2+|q_-(k')|^2]. \tag{3.69}$$

We note now that the parity property (3.67) implies that the function $w(k)$ be odd,

$$w(-k)=-w(k); \tag{3.70}$$

it then follows that, by using (3.65b), the asymptotic velocities S_{out} and S_{in} of the nonlinear wave packet turn out to take the expressions

$$S_{\text{in}}^{\text{out}}=\frac{1}{C}\int_{-\infty}^{+\infty}dk\;k[|q_+(k)|^2/D(\overline{v\mp}w(k))+|q_-(k)|^2/D(\overline{v\pm}w(k))], \tag{3.71}$$

with the consequence, because of (3.67) and (3.70), that

$$S_{\text{out}}=-S_{\text{in}}, \tag{3.72}$$

and this proves that the nonlinear field wave packet behaves like a boomerang.

A similar analysis can be carried on for the spinor field. Here below we just report the relevant formulas. Again, two complex rapidly decreasing Fourier amplitudes, $\varphi_+(k)$ and $\varphi_-(k)$, completely characterize the free spinor field (see (3.12)),

$$\varphi(x,t)=\int_{-\infty}^{+\infty}\frac{dk}{\sqrt{2\pi}}[\varphi_+(k)\mathscr{E}(k,+)+\varphi_-(k)\mathscr{E}(k,-)], \tag{3.73}$$

whose charge c, see (3.7) and (3.39), is the positive quantity

$$c=\int_{-\infty}^{+\infty}dk\;\omega(k)[|\varphi_+(k)|^2+|\varphi_-(k)|^2]. \tag{3.74}$$

Of course, the center of charge coordinate x_c and velocity s are still given by (3.52) and, respectively, (3.53), while the Fourier integral representation of the velocity now reads

$$cs(t)=\int_{-\infty}^{+\infty}dk\{k[|\varphi_+(k)|^2-|\varphi_-(k)|^2]+2m\;\text{Re}[\varphi_+^*(k)\varphi_-(k)\exp(2i\omega t)]\}. \tag{3.75}$$

In order to derive this expression we have taken into account the definition (3.12) of the spinor $\chi(k,\eta)$, together with Eq. (3.44), (3.45), and (3.19). In the limit as $t\to\pm\infty$, this velocity takes the asymptotic value s_a,

$$s_a=\frac{1}{c}\int_{-\infty}^{+\infty}dk\;k[|\varphi_+(k)|^2-|\varphi_-(k)|^2], \tag{3.76}$$

which is the same in the past and in the future. Similarly to the scalar field, the behavior of the nonlinear wave packet is quite different. In order to obtain the asymptotic values S_{out} and S_{in} of the

velocity of the center of charge $S(t)$, as defined by (3.63), as $t\to+\infty$ and $t\to-\infty$, respectively, it is convenient to derive first the asymptotic behavior of the free-field current $j^\mu(\beta t,t)$,

$$j^\mu(\beta t,t)\xrightarrow[t\to\pm\infty]{}|t|^{-1}[j_a^\mu(\beta)+j_\pm^\mu(\beta,t)], \tag{3.77}$$

where

$$j_a^\mu(\beta)=m^2\gamma^4[|\varphi_+(m\gamma\beta)|^2+|\varphi_-(-m\gamma\beta)|^2](1,\beta), \tag{3.78a}$$

$$j_\pm^\mu(\beta,t)=2m^2\gamma^4\ \mathrm{Re}[\varphi_+^*(m\gamma\beta)\varphi_-(-m\gamma\beta)\exp(2i\alpha_\pm)](\beta,1), \tag{3.78b}$$

with $\alpha_\pm(\beta,t)$ as defined by (3.54c). These formulas follow from the asymptotic behavior,

$$\varphi(\beta t,t)\xrightarrow[t\to\pm\infty]{}(m|t|)^{-1/2}\varphi_{\text{in}}^{\text{out}}(\beta,t), \tag{3.79a}$$

$$\varphi_{\text{in}}^{\text{out}}(\beta,t)=m\gamma^{3/2}[\varphi_+(m\gamma\beta)\chi(m\gamma\beta_1,+)\exp(-i\alpha_\pm)+\varphi_-(-m\gamma\beta)\chi(-m\gamma\beta,-)\exp(i\alpha_\pm)], \tag{3.79b}$$

together with the expression (2.25) of the current. Next we obtain the asymptotic behavior of the scalar field $v(x,t)$, namely the functions $v_{\text{out}}(\beta)$ and $v_{\text{in}}(\beta)$ as defined by (3.58a). The lines of computation are the same as in the scalar case, and are not repeated. The resulting expressions are

$$v_{\text{in}}^{\text{out}}(\beta)=\mp\frac{1}{2}\int_{-\infty}^{+\infty}dk\ \omega\left[\operatorname{sign}\left(\beta-\frac{k}{\omega}\right)|\varphi_+(k)|^2+\operatorname{sign}\left(\beta+\frac{k}{\omega}\right)|\varphi_-(k)|^2\right]+\overline{v}, \tag{3.80a}$$

or, equivalently (see (3.66)),

$$v_{\text{in}}^{\text{out}}(\beta)=\mp\frac{1}{2}\left(\int_{-\infty}^{k}dk'-\int_{k}^{+\infty}dk'\right)\omega(k')[|\varphi_+(k')|^2-|\varphi_-(-k')|^2]+\overline{v}. \tag{3.80b}$$

The expression of the total charge C, see (3.8), is, of course, given again by Eq. (3.60), and also Eq. (3.62), (3.63), and (3.64) are still holding. It is also easy to ascertain the validity of the asymptotic formula (3.65) where now the expressions (3.58a) and (3.80) have to be considered, so as to yield the integral representation

$$S_{\text{in}}^{\text{out}}=\frac{1}{C}\int_{-\infty}^{+\infty}dk k[|\varphi_+(k)|^2/D(v_{\text{in}}^{\text{out}}(\beta))-|\varphi_-(k)|^2/D(v_{\text{in}}^{\text{out}}(-\beta))]. \tag{3.81}$$

The observation about the boomerang effect we have already made for the scalar field applies without any change also to the present case and its proof does not need to be repeated.

IV. SPECIAL SOLUTIONS-COMPLEX SPECTRUM

We consider now those nonlinear field solutions which correspond to linear combinations of finitely many exponential solutions of the free field, where the wave vector u^μ, see (1.18), is complex. According to our parametrization (1.14) of u^μ, this clearly amounts to requiring that the spectral angle θ be nonvanishing, $\theta\neq0$, with the further implication that, in contrast with the solutions discussed in the previous section, also the sign parameter $\tau=\pm1$ comes into play (see (1.14)).

Let us look first at the scalar field models. In this case, the generic free field solution is

$$q=\sum_{n=1}^{N} A_n E_n, \tag{4.1}$$

where N is a positive integer, A_n is an arbitrary complex number and, in the notation (1.18) for exponentials, E_n shortly stands for the expression,

$$E_n \equiv E(k_n, \eta_n, \theta_n \tau_n), \quad \theta_n \neq 0, \quad n=1,2,...,N. \tag{4.2}$$

We note that the generic exponential $E(k,\eta,\theta,\tau)=\exp(-iu_\mu x^\mu)$ is not bounded, since, as shown by the explicit expression

$$E(k,\eta,\theta,\tau)=\exp[i\eta\ \cos\ \theta(kx-\omega t)]\exp[-\tau\ \sin\ \theta(\omega x-kt)], \tag{4.3}$$

it diverges to the left (i.e., as $x\to-\infty$) if $\tau=+1$ and to the right (i.e., as $x\to+\infty$) if $\tau=-1$. Just because of this property, it seems appropriate to refer to τ as ''chirality.''

Since we prefer to deal with the current j^μ, see (2.3), rather than with the field q, we introduce the N-dimensional vector, and matrix, notation. Thus, the current associated with the scalar field solution (4.1) takes the following expression:

$$j^\mu=(\mathbf{E},\mathbf{G}^\mu\mathbf{E}), \tag{4.4}$$

where $\mathbf{E}$ is the vector,

$$\{\mathbf{E}\}_n=A_n E_n, \tag{4.5}$$

$\mathbf{G}^\mu$ is the Hermitian matrix whose entries are

$$\{\mathbf{G}^\mu\}_{nm}=\tfrac{1}{2}(u_n^{\mu *}+u_m^\mu), \tag{4.6}$$

and the scalar product is defined as

$$(\mathbf{A},\mathbf{B})=\sum_{n=1}^{N} A_n^* B_n. \tag{4.7}$$

In this notation, and by taking into account the relation (1.19a), the differential equation (2.4) is easily integrated to yield the auxiliary scalar field

$$v=(\mathbf{E},\mathbf{M}\mathbf{E})+\overline{v}, \tag{4.8}$$

where $\overline{v}$ is an integration constant and the entries of the Hermitian matrix $\mathbf{M}$ are given in terms of the invariants μ_{nm} and λ_{nm}, see (1.20),

$$\{\mathbf{M}\}_{nm}=M_{nm}=\frac{1}{2i}\,\lambda_{nm}/(m^2-\mu_{nm}). \tag{4.9}$$

These formulas are the basic tools for the construction of the nonlinear field solutions. Indeed our task here is to discuss the properties of the current (2.8) of the nonlinear field corresponding to (4.1) via the transformation (2.5), namely of the explicit expression (see (3.22) and (4.4)),

$$J^\mu=(\mathbf{E},\mathbf{G}^\mu\mathbf{E})/D(v), \tag{4.10}$$

where v is given by (4.8) and (4.9). Moreover, in order to guarantee that generic solutions be singularity free, in this section we confine our attention to model B_2, see (2.37), so that the expression (4.10) specializes to

$$J^{\mu}=(\mathbf{E},\mathbf{G}^{\mu}\mathbf{E})/[1+a^2((\mathbf{E},\mathbf{M}\mathbf{E})+\overline{v})^2], \tag{4.11}$$

with (4.6) and (4.9).

Let us consider first the total charge C corresponding to the current (4.11); this is given by the general formula (3.9), that presently reads (see (2.37a)),

$$C=\frac{1}{a}\,[\operatorname{arctg}(al)-\operatorname{arctg}(ar)], \tag{4.12}$$

where l and r are, respectively, the asymptotic values $v(-\infty,t)$ and $v(+\infty,t)$ of $v(x,t)$ (see (3.3) and (3.10)). Since these asymptotic values depend on the chirality of the exponentials E_n, it is convenient to specify, without any loss of generality, that the first p exponentials have positive chirality, while the remaining $N-p$ chiralities are negative, i.e.,

$$\tau_1,\tau_2,\ldots,\tau_p>0,\quad \tau_{p+1},\tau_{p+2},\ldots,\tau_N<0, \tag{4.13}$$

where, of course, $p=0$ if all chiralities are negative, and, vice versa, $p=N$ if they are all positive. Moreover, it is expedient to label the exponentials E_n so as to order their rate of growth within each chirality group in the following way:

$$\omega_1 \sin\theta_1<\cdots<\omega_p \sin\theta_p,\quad \omega_{p+1}\sin\theta_{p+1}>\cdots>\omega_N\sin\theta_N. \tag{4.14}$$

These assumptions, together with the expression

$$M_{nn}=\tfrac{1}{2}\tau_n\eta_n \operatorname{cotg}\theta_n, \tag{4.15}$$

which follows from (4.9), (1.20), and (1.14), imply the asymptotic values

$$l=\eta_p\infty,\quad r=-\eta_{p+1}\infty \quad \text{if } 1\leqslant p\leqslant N-1, \tag{4.16a}$$

$$l=\overline{v},\quad r=-\eta_1\infty \quad \text{if } p=0, \tag{4.16b}$$

$$l=\eta_N\infty,\quad r=\overline{v} \quad \text{if } p=N. \tag{4.16c}$$

The implication of the expression (4.12) of the charge C is therefore that, if the state contains both chiralities, the charge is either vanishing or takes the elementary values $\pm C_0=\pm\pi/a$ (see (3.26)) according to the formula,

$$C=\frac{\pi}{2a}\,(\eta_p+\eta_{p+1}),\quad 1\leqslant p\leqslant N-1, \tag{4.17a}$$

while if the chiralities are all positive, or all negative, the total charge can take any value between $-C_0$ and $+C_0$, i.e., $-C_0<C<+C_0$, depending on the integration constant $\overline{v}$, the explicit expressions being [the arctg function is evaluated in the interval $(-\pi/2),(\pi/2)$]

$$C=\frac{\pi}{2a}\,\eta_1+\frac{1}{a}\operatorname{arctg}(a\overline{v}),\quad \text{if } p=0, \tag{4.17b}$$

$$C=\frac{\pi}{2a}\,\eta_N-\frac{1}{a}\operatorname{arctg}(a\overline{v}),\quad \text{if } p=N. \tag{4.17c}$$

Here, and in the following, we omit to consider states with degeneracies such as, for instance, $\omega_p \sin\theta_p=\omega_{p-1}\sin\theta_{p-1}$, and we strictly stick to the condition (4.14).

Let us proceed now to the analysis of the x-profile of the current J^μ, and to its asymptotic evolution at very large time both in the past and in the future. Since the expression (4.11) of the current is a rational function of exponentials of the form (4.3), it is easily seen that, at any fixed time t, J^μ has an exponentially localized profile in the space coordinate x, which, however complicate it may look, shows (see the orderings (4.13) and (4.14), together with (4.6) and (4.15)) the exponential tails

$$J^\mu \to \frac{4}{a^2}\ \eta_{p+1}\ \sin\ \theta_{p+1} \mathrm{tg}\theta_{p+1} k^\mu_{p+1}\ \exp[-2\ \sin\ \theta_{p+1}(\omega_{p+1}x - k_{p+1}t)], \quad x \to +\infty, \tag{4.18a}$$

$$J^\mu \to \frac{4}{a^2}\ \eta_p\ \sin\ \theta_p \mathrm{tg}\theta_p k^\mu_p\ \exp[2\ \sin\ \theta_p(\omega_p x - k_p t)], \quad x \to -\infty. \tag{4.18b}$$

These formulas hold if $1 \leqslant p \leqslant N-1$. In the case $p=0$ (all chiralities are negative, $\tau_n = -1$), the index p in (4.18b) should be replaced with N, while, if $p=N$ (all chiralities are positive, $\tau_n = +1$), the index p in (4.18a) should be formally vanished.

More interesting is the dynamical problem of the time evolution of this profile, and, particularly, the large time behavior and the breaking up processes. Our approach is standard, and we look at the asymptotic y-profile of the current $J^\mu(y+\beta t, t)$, as $t \to \pm\infty$, in a frame of reference moving with velocity β. To this aim, it is convenient to introduce the characteristic velocities

$$\beta_n = k_n/\omega_n, \quad n = 1,...,N, \tag{4.19}$$

so that the modulus of the exponentials (4.2), according to (4.3), depend on the new coordinate y and velocity β as follows:

$$|E_n| = \exp(-\tau_n \omega_n\ \sin\ \theta_n y)\exp[-\tau_n \omega_n\ \sin\ \theta_n(\beta - \beta_n)t]; \tag{4.20}$$

therefore, we conclude that, if β do not coincide with any of the characteristic velocities β_n, $n=1,2,...,N$, then the current $J^\mu(y+\beta t, t)$ [see (4.11)] vanishes as $t \to \pm\infty$ for almost all values of y. This implies that no smooth profile, with traveling speed β, shows up at large time. On the other hand, if β equals one of the velocities β_n, see (4.19), which we assume to be all different from each other for the sake of simplicity, i.e.,

$$\beta_n \neq \beta_m \quad \text{if } n \neq m, \tag{4.21}$$

then the limit state

$$J^\mu_\pm(y) = \lim_{t \to \pm\infty} J^\mu(y+\beta t, t) \tag{4.22}$$

does not vanish if β is such that either the condition

$$\tau_n(\beta - \beta_n) \geqslant 0, \quad n = 1,2,...,N \tag{4.23}$$

holds, as for the $t \to +\infty$ limit, or the condition

$$\tau_n(\beta - \beta_n) \leqslant 0, \quad n = 1,2,...,N, \tag{4.24}$$

is satisfied so as to guarantee the existence of the $t \to -\infty$ limit. In order to investigate on the existence of a velocity β which satisfies condition (4.23) or condition (4.24), we conveniently split the set $\{\beta_n\}$ of the characteristic velocities into two subsets B_+ and B_- according to their corresponding chiralities,

$$B_+ = \{\beta_n : \tau_n = +1\}, \quad B_- = \{\beta_n : \tau_n = -1\}, \tag{4.25}$$

to say that $\beta_1, \ldots, \beta_p \in B_+$ and $\beta_{p+1}, \ldots, \beta_N \in B_-$, as it follows from our ordering (4.13). In fact, it can be readily proved that, for the existence of a solution β of condition (4.23) or condition (4.24), it is necessary and sufficient, that the two sets B_+ and B_- be disjoined in the following sense. Let $\beta^{\pm}_{\max}$ and $\beta^{\pm}_{\min}$ be defined as

$$\beta^{\pm}_{\max} = \mathrm{Max}\{\beta_n\}, \quad \beta^{\pm}_{\min} = \min\{\beta_n\}, \quad \beta_n \epsilon B_{\pm}, \tag{4.26}$$

then either

$$\beta^{+}_{\max} < \beta^{-}_{\min} \quad \text{or} \quad \beta^{-}_{\max} < \beta^{+}_{\min}. \tag{4.27}$$

If this condition is not met by the set of velocities β_n, then no asymptotic profile $J^{\mu}_{\pm}(y)$, see (4.22), exists as an ordinary function of y, though an asymptotic state may appear as a Dirac distribution (see also below). In the following, we discuss only those solutions J^{μ} for which the condition (4.27) for the existence of asymptotic profiles is satisfied.

Consider first the case in which both chiralities enter in the solution, namely, $1 \leqslant p \leqslant N-1$. Assume also that, of the two possibilities (4.27), the condition

$$\beta^{+}_{\max} < \beta^{-}_{\min} \tag{4.28}$$

is satisfied. Then, while the inequality (4.24) has no solution β, both $\beta = \beta^{+}_{\max}$ and $\beta = \beta^{-}_{\min}$ satisfy the inequality (4.23). Moreover, to simplify our discussion, we can change the frame of reference by a Lorentz transformation so as to have

$$\beta^{+}_{\max} = -\overline{\beta}, \quad \beta^{-}_{\min} = \overline{\beta}, \quad \overline{\beta} > 0. \tag{4.29}$$

According to our previous considerations, we conclude that, in the present case, no y-profile asymptotically shows up at $t \to -\infty$, while, as $t \to +\infty$, two profiles separate from each other traveling with opposite velocities $\overline{\beta}$ and $-\overline{\beta}$. Since these two profiles clearly obtain from the general expression (4.11) when all exponentials, but one, disappear, at this point it is convenient to write down the expression of the current J^{μ} for $N=1$, namely the 1-soliton profile. By dropping the unnecessary index, and using the notation (1.14) and the formula (4.15), the one-soliton expression of the current (4.11) reads

$$J^{\mu}(k, \eta, \theta, \tau) = \eta \frac{\sin\,\theta}{a} k^{\mu} / (\sqrt{1 + a^2 \overline{v}^2}\, \cosh\, \xi + \eta \tau a \overline{v}), \tag{4.30}$$

where $k^{\mu} = (\omega, \kappa)$, $\omega = \sqrt{m^2 + k^2}$,

$$\xi \equiv 2\tau \sin\,\theta \epsilon_{\mu\nu} k^{\mu} (x^{\nu} - \overline{x^{\nu}}) = 2\tau \sin\,\theta\, [\omega(x - \overline{x}) - k(t - \overline{t})] = 2\tau\omega \sin\,\theta\, [x - \overline{x} - \beta(t - \overline{t})], \tag{4.31}$$

and

$$\exp(2\tau \sin\,\theta \epsilon_{\mu\nu} k^{\mu} \overline{x^{\nu}}) = \frac{a}{2} \mathrm{cotg}\,\theta |A|^2 / \sqrt{1 + a^2 \overline{v}^2}. \tag{4.32}$$

It should be noted that the profile of solitons with the same spectral parameters k and θ and integration constant $\overline{v}$ may still look different because of the relative sign $\eta\tau$ of charge and chirality, as shown by the expression of the denominator.

The total charge C of the soliton is given by the general expression (4.17), which presently reads

$$C=\frac{\pi}{2a}\,\eta-\frac{\tau}{a}\,\mathrm{arctg}(a\overline{v}), \tag{4.33}$$

while we note that, because of the arbitrariness of the parameter $\overline{v}$, the amplitude and size of the soliton are independent from each other.

Going back to our asymptotic analysis, if $1\leqslant p\leqslant N-1$, and the condition (4.28) is satisfied, see (4.29), then, as $t\to+\infty$, the nonlinear field solution breaks up into two one-soliton states,

$$J^{\mu}\to J^{\mu}(-\overline{k},\eta_{+},\theta_{+},1)+J^{\mu}(\overline{k},\eta_{-},\theta_{-},-1),\quad t\to+\infty, \tag{4.34}$$

where (see (4.19))

$$\overline{k}=m\overline{\beta}/\sqrt{1-\overline{\beta}^{2}}, \tag{4.35}$$

and η_{+}, $\theta_{+}(\eta_{-},\theta_{-})$ are the charge and spectral angle associated with the particular exponential corresponding to $-k(k)$ (note that the chiralities are, by construction, $\tau_{+}=+1$ and $\tau_{-}=-1$).

A peculiar outcome of this analysis is that the charge C does not necessarily coincide with the sum of the charges of the asymptotic one-soliton states into which the N-exponential solution (4.11) breaks. Indeed, while the total charge in the case we are considering is given by (4.17a), the sum of the charges of the two one-soliton states which appear at $t\to+\infty$, see (4.34) and (4.33), is

$$C_{+\infty}=\frac{\pi}{2a}\,(\eta_{+}+\eta_{-}), \tag{4.36}$$

$\eta+\,+\eta-$ being not necessarily equal to $\eta_{p}+\eta_{p+1}$ for $N\geqslant 3$, and on the other hand, no solitons states exist in the limit as $t\to-\infty$. This property, which is obviously due to the fact that the two limit operations as $|x|\to\infty$ and $|t|\to\infty$ do not commute with each other, does not contradict the charge conservation law as it is shown below while considering particularly simple solutions. Here we limit ourselves to noticing that a picture of the evolution of the charge density $J^{0}(x,t)$ would look like the process of creation of a couple of solitons.

The evolution is quite different if we assume instead that all chiralities are equal, say positive, so that $p=N$ (see (4.13)), while still keeping the assumption (4.14). In fact, in this case, both inequalities (4.23) and (4.24) have one, and only one, solution, namely $\beta=\beta^{+}_{\max}$ satisfies (4.23) and $\beta=\beta^{+}_{\min}$ satisfies (4.24). It is convenient to take again advantage of a Lorentz transformation to set, without any loss of generality,

$$\beta^{+}_{\min}=-\overline{\beta},\quad \beta^{+}_{\max}=\overline{\beta}>0, \tag{4.37}$$

and to use the same notation as before (and note that now the chiralities are positive, $\tau_{+}=\tau_{-}=1$). Therefore, we conclude that the two limits

$$J^{\mu}_{\pm}(y)=\lim_{t\to\pm\infty}J^{\mu}(y\pm\overline{\beta}t,t) \tag{4.38}$$

exist, and are the one-soliton solution (see (4.30))

$$J^{\mu}_{\pm}=J^{\mu}(\pm\overline{k},\eta_{\pm},\theta_{\pm},1), \tag{4.39}$$

with charge

$$C_{\pm}=\frac{\pi}{2a}\,\eta_{\pm}-\frac{1}{a}\,\mathrm{arctg}(a\overline{v}). \tag{4.40}$$

Again, we note that the charge $C_{\pm}$ of the asymptotic states may not equal the charge of the nonlinear field solution, which, according to (4.17c), reads

$$C=\frac{\pi}{2a}\,\eta_N-\frac{1}{a}\,\mathrm{arctg}(a\overline{v}). \tag{4.41}$$

Indeed, this happens if η_+ and/or η_- do not coincide with η_N, and the way the charge conservation law is not broken is discussed below. It should be also pointed out that, because of the opposite sign of the velocities of the in and out-soliton states, the nonlinear field solution pictures an object which moves like a boomerang. The dynamics in the alternative case in which all chiralities are negative, say $p=0$ (see (4.13)), looks very much the same (i.e., it describes a boomerang-type object) and does not need to be detailed.

The remaining case, in which the inequality

$$\beta^{-}_{\max}<\beta^{+}_{\min} \tag{4.42}$$

holds, instead of (4.28), is easily seen to correspond to a situation in which time is reversed. Indeed, in this case, only the condition (4.24), which implies the existence of asymptotic solitons as $t\to-\infty$, has solutions, namely $\beta=\beta^{-}_{\max}$ and $\beta=\beta^{+}_{\min}$, while the opposite condition (4.23), for the existence of solitons at $t=+\infty$, has no solution. The resulting process can be roughly described as the annihilation of two solitons. Since the analysis of this nonlinear field solution is similar to that of the creation process described above, we deem it unnecessary to discuss it any further. It is more instructive to detail the behavior of a few particularly simple solutions.

Let us consider first the zero-charge $N=2$ solution which describes the soliton–antisoliton annihilation (solitons and antisolitons are defined by the expression (4.30) with $\eta=1$ and, respectively, $\eta=-1$, so that their charge densities are positive and, respectively, negative). This is the case with $p=1$, i.e., $\tau_1=1$ and $\tau_2=-1$, and $\eta_1+\eta_2=0$ so as to have $C=0$ (see (4.17a)). The annihilation process is implied by the condition (4.42), which is satisfied with the choice

$$\beta_2=-\overline{\beta},\quad \beta_1=\overline{\beta}>0; \tag{4.43}$$

the evolution is shown in Fig. 3. The plot in Fig. 4 shows instead the soliton–antisoliton creation [with, of course, $\beta_1=-\beta$ and $\beta_2=\beta>0$, see condition (4.28)].

More illuminating, with respect to the asymptotic behavior, is the nonvanishing charge $N=2$ solution, with opposite chiralities, i.e., $p=1$. Just to be definite, let us set $\eta_1=\eta_2=1$ so that the total charge takes the elementary value, $C=C_0=(\pi/a)$ [see (4.17a)], and let us take the velocities β_1 and β_2 as in (4.43), so that the two solitons have a head-on collision and disappear since no soliton profile is found as $t\to+\infty$. Let us choose, for simplicity, the following values of the parameters: $\theta_1=\theta_2=\theta$, $A_1=A_2=1$, and note that the profiles of the two incoming solitons are different because $\eta_1\tau_1=1$ and $\eta_2\tau_2=-1$ [see (4.30)]. Then, let us compute the corresponding charge density $J^0(x,t)$, see (4.11) with (4.6) and (4.9),

$$J^0=\frac{2\overline{\omega}\cos\,\overline{\theta}(\cosh\,\xi+\cos\,\varphi)\exp(2\overline{k}\,\sin\,\overline{\theta t})}{\left\{1+a^2\left[\overline{v}+\left(\cot\mathrm{g}\,\overline{\theta}\,\sinh\,\xi+\frac{\overline{\omega}}{k}\sin\,\varphi\right)\exp(2\overline{k}\,\sin\,\overline{\theta t})\right]^2\right\}}, \tag{4.44a}$$

where we have set $(\overline{k}=\overline{\beta}\,\overline{\omega})$,

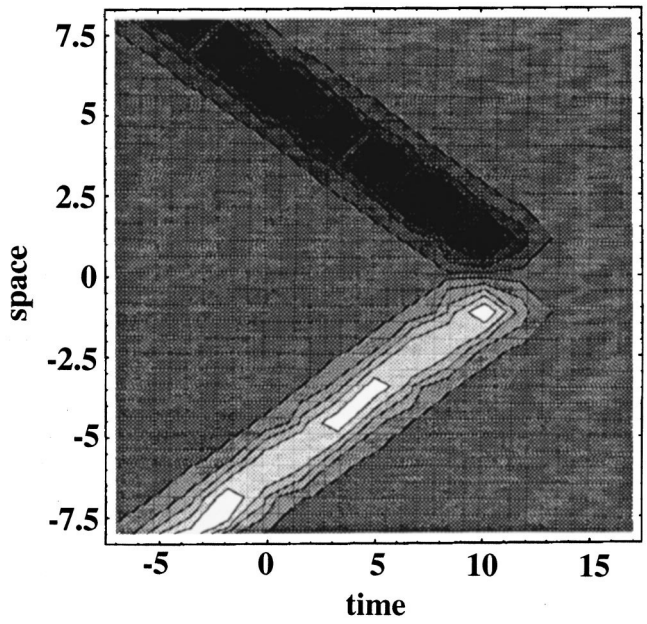

FIG. 3. Soliton–antisoliton annihilation. Contour plot of the charge density $J^0(x,t)$ for the scalar model B_2 with $a=1$, $m=1$. The parameters are $\eta_1=-\eta_2=1$, $\tau_1=-\tau_2=1$, $\theta_1=\theta_2=\pi/6$, $\beta_1=-\beta_2=2/3$, $a_1=a_2=0.8$, $\overline{v}=0$. The soliton and antisoliton have opposite charge, as shown by the grey level.

$$\xi=2\overline{\omega}\sin\overline{\theta x},\quad \varphi=2\overline{k}\cos\overline{\theta x}. \tag{4.44b}$$

It is now easy to derive from this explicit expression the asymptotic limit of $J^0(x,t)$ as $t\to+\infty$, where no soliton exists. In fact, since the function $\operatorname{cotg}\theta\sinh\xi+(\overline{\omega}/\overline{k})\sin\varphi$ in the denominator of the right-head side of (4.44a) is monotonically increasing with just one (simple) zero at $x=0$, it follows by direct computation that this limit exists as a Dirac distribution, and it is found to be

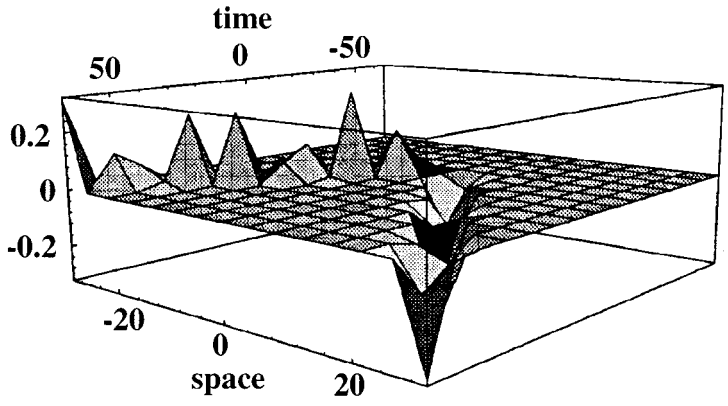

FIG. 4. Soliton–antisoliton creation. Three-dimensional plot of the charge density $J^0(x,t)$ for the scalar model B_2 with $a=1$, $m=1$. The parameters are $\eta_1=-\eta_2=-1$, $\tau_1=-\tau_2=1$, $\theta_1=\theta_2=\pi/4$, $\beta_1=-\beta_2=-1/3$, $a_1=a_2=10$, $\overline{v}=0$.

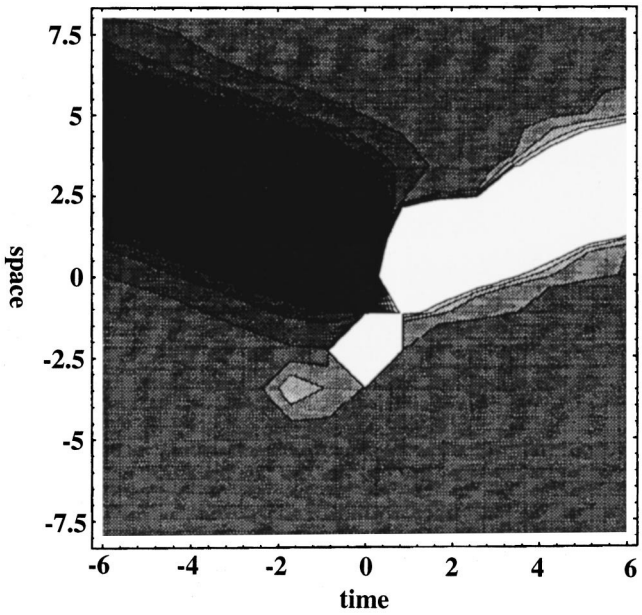

FIG. 5. Pointon absorption. Contour plot of the charge density $J^0(x,t)$ for the scalar model B_2 with $a=1$, $m=1$. Since the pointon tends to a Dirac distribution as $t\to-\infty$, its track is not visible back far in the past because of low numerical resolution. The parameters are $\eta_1=-\eta_2=1$, $\tau_1=\tau_2=1$, $\theta_1=\pi/6$, $\theta_2=\pi/3$, $\beta_1=-\beta_2=1/2$, $a_1=a_2=2$, $\overline{v}=0$. The charge of the incoming soliton is $-\pi/2$, that of the outgoing soliton is $\pi/2$, and the pointon charge is π.

$$\lim_{t\to+\infty} J^0(x,t)=\frac{\pi}{a}\,\delta(x)=C_0\delta(x). \tag{4.45}$$

This result shows that, after the collision of the two solitons, an elementary charged particle is left at rest in the origin with the remarkable property of becoming point-like as $t\to+\infty$. This mechanism of creation (annihilation) of charged point-like particles, whenever the in and out-soliton states have different charges, is quite general, and we have explicity shown it in this simple case just to avoid complicated expressions.

There are two different kinds of elementary objects in this field model, which show themselves as asymptotic states; these are the soliton, and the point-like particle, which we call ''pointon'' to stress their point-size structure. The current J^μ associated with a pointon is simply a Dirac distribution,

$$J^\mu(x,t)=\eta C_0 v^\mu \delta(\epsilon_{\alpha\beta}v^\alpha x^\beta), \tag{4.46}$$

where v^μ is its velocity vector, namely, $v^\mu=(\gamma,\beta\gamma)$ with $-1<\beta<1$ and $v^\mu\nu_\mu=1$. Of course, $\eta=1$ for pointons and $\eta=-1$ for antipointons. The process we have described above is therefore the formation of a standing pointon at the origin as a result of a two-soliton collision. The opposite process, in which two solitons appears as $t\to+\infty$, while a pointon was at the origin at $t=-\infty$, pictures instead the decay of a pointon in two solitons.

Although we leave to the interested reader the task to explore more complex solutions of our nonlinear field model, we conclude from our findings that the complex spectrum solutions break up at large $|t|$ into systems which generically may contain solitons and antisolitons together with pointons and antipointons. Two examples are shown in Figs. 5 and 6.

A quite similar analysis of complex-spectrum solutions of the spinor field models goes as well along the same lines. Since only few obvious modifications of the scalar field results are required,

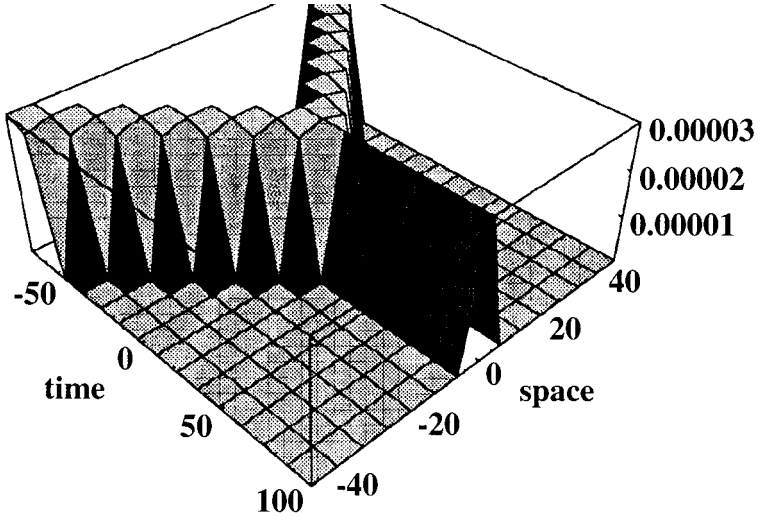

FIG. 6. Pointon creation. Three-dimensional plot of the charge density $J^0(x,t)$ for the scalar model B_2 with $a=1$, $m=1$. The charge of the two incoming solitons is $\pi/2$, and the pointon charge is π. The parameters are $\eta_1=\eta_2=1$, $\tau_1=-\tau_2=1$, $\theta_1=\theta_2=\pi/4$, $\beta_1=-\beta_2=1/2$, $a_1=a_2=0.03$, $\overline{v}=0$.

our discussion of the spinor field given below is very terse with emphasis only on the main changes. Again we consider only solutions of the model B_2, see (2.37), for which the expression of the current J^μ is still given by (4.11) with the following specifications. The spinor free field to start with is

$$\varphi=\sum_{n=1}^{N} A_n \mathscr{E}_n, \tag{4.47}$$

where A_n is an arbitrary complex number, and the elementary spinor $\mathscr{E}_n$ stands, according to our notation (1.27) with (1.18), for

$$\mathscr{E}_n=\mathscr{E}(k_n,\eta_n,\theta_n,\tau_n)=E_n\chi_n, \tag{4.48}$$

where the exponential E_n is (4.2) and χ_n is the spinor (1.28), namely,

$$\chi_n=\chi(k_n,\eta_n,\theta_n,\tau_n). \tag{4.49}$$

The free field current j^μ is easily computed, and it can be given again the expression (4.4) in vector notation, where, however, the $N\times N$ matrix $\mathbf{G}^\mu$ takes now the following form:

$$\{\mathbf{G}^\mu\}_{nm}=\Gamma_{nm}(u_n^{\mu *}+u_m^\mu), \tag{4.50}$$

where the scalar Γ_{nm} comes from the general formula (1.33), with (1.31), and its expression obtains by combining (1.34) with (1.35),

$$\Gamma_{nm}=m\,\eta_n[\cos\,\theta_n\,\cos\,\theta_m/2(m^2+\mu_{nm})]^{1/2}, \quad \eta_n\eta_m=1, \tag{4.51a}$$

$$\Gamma_{nm}=m\,\eta_n[\lambda_{nm}/(m^2+\mu_{nm})][\cos\,\theta_n\,\cos\,\theta_m/2(m^2-\mu_{nm})]^{1/2}, \quad \eta_n\eta_m=-1. \tag{4.51b}$$

Again, the relation (1.19a) leads to the integration of the differential equation (2.4), whose solution has the same expression as (4.8), where the $N\times N$ matrix $\mathbf{M}$ in the present case reads

$$\{\mathbf{M}\}_{nm}=M_{nm}=\frac{1}{i}\,\Gamma_{nm}\lambda_{nm}/(m^2-\mu_{nm}). \tag{4.52}$$

Having established these basic formulas, the following analysis of the solution (4.11) parallels that one discussed for the scalar case. Indeed, all arguments and reasonings are exactly the same, the only difference being due to the introduction of the factor Γ_{nm} in (4.50) and (4.52) (note that replacing Γ_{nm} with 1/2 reproduces the scalar field results).

The total charge is again provided by the expression (4.12); in order to evaluate the asymptotic values l and r we assume the labeling (4.13) of the chiralities τ_n, with the ordering (4.14), and we use the expression

$$M_{nn}=\tfrac{1}{2}\tau_n\ \mathrm{cotg}\ \theta_n, \tag{4.53}$$

which follows from (4.52), the previous formula (4.15), and the equality $\Gamma_{nn}=\eta_n/2$. The resulting limit values are

$$l=+\infty,\quad r=-\infty,\quad \text{if } 1\leqslant p\leqslant N-1, \tag{4.54a}$$

$$l=\overline{v},\quad r=-\infty,\quad \text{if } p=0, \tag{4.54b}$$

$$l=+\infty,\quad r=\overline{v},\quad \text{if } p=N. \tag{4.54c}$$

As a consequence of (4.12), we find that the charge of the spinor field solution is

$$C=C_0=\frac{\pi}{a},\quad 1\leqslant p\leqslant N-1 \tag{4.55a}$$

if both positive and negative chiralities are present, while in the opposite two cases

$$C=\frac{\pi}{2a}+\frac{1}{a}\,\mathrm{arctg}(a\overline{v}),\quad p=0, \tag{4.55b}$$

$$C=\frac{\pi}{2a}-\frac{1}{a}\,\mathrm{arctg}(a\overline{v}),\quad p=N. \tag{4.55c}$$

As expected from the positiveness of the charge density J^0, the total charge (4.55) is always positive. The exponential tails of the profile at fixed t, and at the leading order, are

$$J^\mu\to\frac{4}{a^2}\sin\theta_{p+1}tg\,\theta_{p+1}k^\mu_{p+1}\exp[-2\sin\theta_{p+1}(\omega_{p+1}x-k_{p+1}t)],\quad x\to+\infty, \tag{4.56a}$$

$$J^\mu\to\frac{4}{a^2}\sin\theta_p tg\,\theta_p k^\mu_p\exp[2\sin\theta_p(\omega_p x-k_p t)],\quad x\to-\infty. \tag{4.56b}$$

These expressions are valid if $1\leqslant p\leqslant N-1$. In the case $p=0$ all chiralities are negative and the index p in (4.56b) should be replaced by N; instead, if $p=N$, the asymptotic expression (4.56a) holds if one formally sets $p=0$.

The investigation of the evolution of the N-exponential solution (4.47) through the large-time asymptotic behavior of the corresponding current J^μ is the same as in the scalar case, and it is not repeated. Again, we limit our analysis to those solutions with characteristic velocities β_n which satisfy the important condition (4.27). In these cases, the asymptotic states may contain solitons

and pointons, as shown for scalar fields. As a difference with respect to scalar solitons, we note that the spinor soliton profile do not depend on the sign parameter η, its expression being

$$J^{\mu}(k,\theta,\tau)=\frac{\sin\,\theta}{a}\,k^{\mu}/(\sqrt{1+a^2\overline{v^2}}\,\cosh\,\xi+\tau a\overline{v}), \tag{4.57}$$

where ξ is specified by (4.31) and (4.32); this, of course, implies that also the charge does not depend on η,

$$C=\frac{\pi}{2a}-\frac{\tau}{a}\,\mathrm{arctg}(a\overline{v}). \tag{4.58}$$

As already pointed out, the fact that the charge has only one sign (positive) is a consequence of the commutativity of our spinor field variables, as contrasted with the theory of quantized fields.

Therefore, in the spinor field model, since only positively charged states exist, the elementary objects, which show up at large time are solitons and pointons. With the exception of those processes corresponding to solutions with vanishing charge, such as soliton–antisoliton creation and annihilation, all other reactions occurred in the scalar case such as pointon decay and formation, do also exist for spinor field solutions, and can be easily investigated by analogy.

V. CONCLUSIONS

The field models we have introduced are C-integrable, $U(1)$ invariant and relativistically invariant, and the conserved Noether current is local. The special solutions we have explicitly constructed are sufficiently representative of the main features of the models. Some of the field equations we have explicitly displayed have been chosen because their discrete spectrum solutions are free of singularities. In particular, the solutions of the model B_2 [see (2.37)] for the scalar fields $Q(x,t)$ and $V(x,t)$,

$$\Box Q+\{m^2-a^2[1+2tg^2(aV)]J_{\alpha}\,J^{\alpha}+2a\;\;\mathrm{tg}(aV)\,\epsilon^{\alpha\beta}\,\mathrm{Im}(Q_{,\alpha}Q^{*}_{,\beta})\}Q=0, \tag{5.1a}$$

$$V_{,\mu}=\epsilon_{\mu\nu}J^{\nu},\quad J_{\mu}\equiv\mathrm{Im}(QQ^{*}_{,\mu}), \tag{5.1b}$$

have been discussed in detail. In fact, each model can be given both a scalar version and a spinor version. Thus, for instance the model B_2 for the spinor field ϕ and the scalar field V is obtained by replacing Eq. (5.1) with

$$\gamma^{\mu}\phi_{,\mu}+[im+a\;\;\mathrm{tg}(aV)\,\epsilon_{\mu\nu}\gamma^{\mu}\overline{\phi}\gamma^{\nu}\phi]\phi=0, \tag{5.2a}$$

$$V_{,\mu}=\epsilon_{\mu\nu}\overline{\phi}\gamma^{\nu}\phi. \tag{5.2b}$$

The most remarkable properties of these models appear when looking at nondispersive localized solutions. Their total charge is independent on amplitude and phase, and its absolute value C_0 is characteristic of the model; for instance, for (5.1) and (5.2), $C_0=\pi/a$. Moreover, in the scalar case, the large time asymptotic analysis in the solitonic sector shows that, in addition to smooth, exponentially localized, profiles, there may appear point-like states described by Dirac distributions (pointons and antipointons). The charge carried by pointons (antipointons) turns out to be the elementary charge $C_0(-C_0)$ of the model. The dynamics described by these solutions is rather rich, and we have displayed simple instances of processes such as pair creation and annihilation, as well as pointon absorption and emission. In this respect, we find it remarkable that in our models, smooth initial conditions for the fields evolve, in the limit as $t\to\infty$, to asymptotic states containing point-like particles, whereas these objects are usually considered as foreign to field theory.

We finally observe that the class of nonlinear field models considered here, is still integratable when a parametric forcing is introduced. Indeed, in this case the nonlinear equations are mapped into linear equations with a potential term. It would therefore be of interest to introduce, both in the scalar and in the spinor case, some particular form of external potential, and to analyze the asymptotic properties of the solutions. We expect that, in analogy with the nonrelativistic case,[4] such properties should be strongly influenced by the parametric forcing. We leave this matter to future investigation.

[1] A. Degasperis and F. Tinebra, J. Math. Phys. **34**, 2950 (1993).
[2] A. Kundu, Physica D **25**, 399 (1987).
[3] F. Calogero and S. De Lillo, Inverse Probl. **3**, 633 (1987).
[4] M. J. Ablowitz and S. De Lillo, Nonlinearity **7**, 1143 (1994).

Dirac operator on a disk with global boundary conditions

H. Falomir, R. E. Gamboa Saraví, and E. M. Santangelo
Departamento de Física, Facultad de Ciencias Exactas,
Universidad Nacional de La Plata, Argentina

(Received 20 October 1996; accepted for publication 10 September 1997)

We compute the functional determinant for a Dirac operator in the presence of an Abelian gauge field on a bidimensional disk, under global boundary conditions of the type introduced by Atiyah–Patodi–Singer. We also discuss the connection between our result and the index theorem. © *1998 American Institute of Physics.* [S0022-2488(97)03312-4]

I. INTRODUCTION

The wide application of functional determinants in Quantum and Statistical Physics is by now a well known fact. Typically, one is faced with the necessity of defining a regularized determinant for elliptic differential operators, among which the Dirac first order one plays a central role. An interesting related problem is the modification of physical quantities due to the presence of boundaries. The study of boundary effects has lately received much attention,[1–9] since it is of importance in many different situations, like effective models for strong interactions, quantum cosmology and application of QFT to statistical systems, among others.

In previous work,[10,11] we have studied elliptic Dirac boundary problems in the case of local boundary conditions. In particular, we have developed for this case a scheme for evaluating determinants from the knowledge of the associated Green's function, based on Seeley's theory of complex powers.[12]

Another type of boundary conditions extensively studied in the literature are global ones, of the type introduced by Atiyah, Patodi and Singer (APS)[13] in connection with the index theorem for manifolds with boundaries (see, Ref. 14, for a review.) Other motivations for considering these global (or spectral) conditions are their consistency with charge conjugation and chiral invariance, and the presence of topological obstructions for the chiral Dirac operator under local boundary conditions (although this restriction no longer holds when considering the whole Dirac operator[10]).

In this paper we present the complete evaluation of the determinant of the Dirac operator on a disk, in the presence of an axially symmetric Abelian flux, under spectral boundary conditions (see Refs. 15, 16 and references therein for related work), chosen so that they make self adjoint the operator under consideration.

In Section II we establish our conventions and set the problem to be considered.

In Section III we study zero modes and obtain the projector on the null space of the operator.

Section IV is devoted to the evaluation of the Green's function of the problem in the subspace orthogonal to such null space. The knowledge of this Green's function is necessary in the evaluation of the determinant, which is done in Section V. This evaluation is performed in two steps. The first one involves the consideration of the quotient of determinants with and without gauge field, under a fixed boundary condition. For this situation, we grow the field along a continuous path through a parameter α, and perform a point splitting regularization of the α-derivative of the logarithm of the quotient of determinants, written in terms of the Green's function.

The second step amounts to the evaluation, via ζ-function, of the quotient of the free operators under different global boundary conditions, which is possible thanks to the explicit determination of their eigenfunctions and corresponding eigenvalues.

The use of a different regularization in each step is enforced for the following reasons: In the first place, a construction of complex powers is not available for the case of spectral boundary conditions (see Ref. 18). As regards the Green's function approach, it cannot be applied to the

0022-2488/98/39(1)/532/13/$15.00
© 1998 American Institute of Physics

second step, because there is no smooth path connecting both boundary conditions.

In Section VI we discuss the application of the APS index theorem to our example, with emphasis on the effect of the boundary isomorphism σ appearing in the Dirac operator.

Finally, Section VII contains our conclusions.

II. SETTING OF THE PROBLEM

We will evaluate the determinant of the operator $D=i\not{\partial}+\not{A}$ acting on functions defined on a two dimensional disk of radius R, under APS boundary conditions.

We take A_μ to be an Abelian gauge field. As it is well known, it can be written as $A_\mu=\epsilon_{\mu\nu}\partial_\nu\phi+\partial_\mu\eta$ ($\epsilon_{01}=-\epsilon_{10}=1$). We set $\eta\equiv 0$, thus choosing the Lorentz gauge. For ϕ we take a smooth (real) bounded function $\phi=\phi(r)$; then $A_r=0$ and $A_\theta(r)=-\partial_r\phi(r)=-\phi'(r)$. We call

$$\kappa=\frac{\Phi}{2\pi}=\frac{1}{2\pi}\oint_{r=R}A_\theta R\,d\theta=-R\phi'(R). \tag{1}$$

Our convention for two dimensional Dirac matrices is

$$\gamma_0=\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad \gamma_1=\begin{pmatrix}0&-i\\i&0\end{pmatrix},\quad \gamma_5=\begin{pmatrix}1&0\\0&-1\end{pmatrix}, \tag{2}$$

which satisfy

$$\gamma_\mu\gamma_\nu=\delta_{\mu\nu}I+i\epsilon_{\mu\nu}\gamma_5. \tag{3}$$

Therefore, the free Dirac operator can be written as

$$i\not{\partial}=i(\gamma_0\partial_0+\gamma_1\partial_1)=2i\begin{pmatrix}0&\dfrac{\partial}{\partial X}\\[2ex]\dfrac{\partial}{\partial X^*}&0\end{pmatrix}, \tag{4}$$

where $X=x_0+ix_1$ and $X^*=x_0-ix_1$ or, in polar coordinates

$$i\not{\partial}=i\left(\gamma_r\partial_r+\frac{1}{r}\gamma_\theta\partial_\theta\right), \tag{5}$$

with

$$\gamma_r=e^{-i\gamma_5\theta}\gamma_0=\begin{pmatrix}0&e^{-i\theta}\\e^{i\theta}&0\end{pmatrix},\quad \gamma_\theta=e^{-i\gamma_5\theta}\gamma_1=\begin{pmatrix}0&-ie^{-i\theta}\\ie^{i\theta}&0\end{pmatrix}. \tag{6}$$

With these conventions, the full Dirac operator can be written as

$$D=e^{-\gamma_5\phi(r)}\,i\not{\partial}e^{-\gamma_5\phi(r)}=\begin{pmatrix}0&\sigma^{-1}(\partial_r+B)\\\sigma(-\partial_r+B)&0\end{pmatrix}, \tag{7}$$

with

$$\sigma=-ie^{i\theta}, \tag{8}$$

and

$$B(r)=-\frac{i}{r}\partial_\theta+\partial_r\phi(r). \tag{9}$$

At the boundary, the self adjoint operator $B(r)$ becomes

$$B\equiv B(R)=-\frac{1}{R}[i\partial_\theta+\kappa], \tag{10}$$

with eigenvectors $e^{in\theta}$ and eigenvalues $(1/R)(n-\kappa)$. We take the radial variable to be conveniently adimensionalized through multiplication by a fixed constant with dimensions of mass.

We will consider the action of the differential operator D on the space of functions satisfying global boundary conditions, characterized by

$$(\mathscr{P}_{\geqslant}\quad \sigma(1-\mathscr{P}_{\geqslant})\sigma^*)\begin{pmatrix}\varphi(R,\theta)\\ \chi(R,\theta)\end{pmatrix}=0, \tag{11}$$

with

$$\mathscr{P}_{\geqslant}=\frac{1}{2\pi}\sum_{n\geqslant k+1} e^{in\theta}\langle e^{in\theta},\cdot\rangle, \tag{12}$$

where k is the integer such that $k<\kappa\leqslant k+1$. Notice that

$$\sigma(1-\mathscr{P}_{\geqslant})\sigma^*=\sigma\,\mathscr{P}_{<}\,\sigma^*=\frac{1}{2\pi}\sum_{n\leqslant k+1} e^{in\theta}\langle e^{in\theta},\cdot\rangle=\mathscr{P}_{\leqslant} \tag{13}$$

and the operator so defined, $(D)_\kappa$, turns out to be self adjoint. The presence of σ as in (11) has a relevant consequence on the boundary conditions for the lower components. As we will discuss later, this also reflects on the form of the index theorem for the present situation.

Our aim is to compute the quotient of the determinants of the operators $(D)_\kappa$ and $(i\not\partial)_{\kappa=0}$. Since the global boundary conditions in Eq. (11) depend on the flux Φ as a step function, there is no continuous family connecting both operators. So, we will proceed in two steps:

$$(D)_\kappa\rightarrow(i\not\partial)_\kappa\rightarrow(i\not\partial)_{\kappa=0}. \tag{14}$$

In the first step, where there is no change of boundary conditions, we can grow the gauge field by varying α from 0 to 1 in

$$D_\alpha=i\not\partial+\alpha\not{A}=e^{-\alpha\gamma_5\phi(r)}i\not\partial e^{-\alpha\gamma_5\phi(r)},\ \text{ with }\ 0\leqslant\alpha\leqslant1, \tag{15}$$

thus going smoothly from the free to the full Dirac operator. The explicit knowledge of the Green's function will allow us to perform the calculation of this step, where we will use a gauge invariant point splitting regularization of the α-derivative of the determinant. The second step will be achieved by using a ζ-function regularization, after explicitly computing the spectra.

But this is not the whole story: As we are going to see, these global boundary conditions give rise to the presence of zero modes, which must be taken into account.

III. ZERO MODES

We will here show that the operator $(D_\alpha)_\kappa$ has $|k+1|$ zero modes.

From (7), we get

$$D_\alpha\psi=0\Leftrightarrow\not\partial e^{-\alpha\gamma_5\phi(r)}\psi=0, \tag{16}$$

or, equivalently

$$\begin{pmatrix} 0 & e^{-i\theta}\left(\partial_r - \frac{i}{r}\partial_\theta\right) \\ e^{i\theta}\left(\partial_r + \frac{i}{r}\partial_\theta\right) & 0 \end{pmatrix} \begin{pmatrix} e^{-\alpha\phi(r)} & 0 \\ 0 & e^{\alpha\phi(r)} \end{pmatrix} \begin{pmatrix} \varphi(r,\theta) \\ \chi(r,\theta) \end{pmatrix} = 0. \tag{17}$$

Now, we introduce the expansions

$$\varphi(r,\theta) = \sum_{n=-\infty}^{\infty} \varphi_n(r) e^{in\theta}, \quad \chi(r,\theta) = \sum_{n=-\infty}^{\infty} \chi_n(r) e^{in\theta}. \tag{18}$$

The solutions are thus given by

$$\varphi_n(r) = a_n r^n e^{\alpha\phi(r)}, \quad \chi_n(r) = b_n r^{-n} e^{-\alpha\phi(r)}, \tag{19}$$

where the coefficients a_n and b_n are to be determined from the regularity requirement at the origin and the boundary conditions at $r=R$, Eq. (11). Thus

$$\begin{aligned} \varphi(r,\theta) &= e^{\alpha\phi(r)} \sum_{n=0}^{k} a_n r^n e^{in\theta}, \quad \text{for } k \geqslant 0, \\ \chi(r,\theta) &= e^{-\alpha\phi(r)} \sum_{n=0}^{-k-2} b'_n r^n e^{-in\theta}, \quad \text{for } k < -1. \end{aligned} \tag{20}$$

Then, there are $|k+1|$ zero modes which, once normalized, are given by

$$\begin{aligned} &\frac{e^{\alpha\phi(r)}}{\sqrt{2\pi q_n(R;\alpha)}} \begin{pmatrix} X^n \\ 0 \end{pmatrix}, \quad \text{for } 0 \leqslant n \leqslant k, \quad \text{if } k \geqslant 0; \\ &\frac{e^{-\alpha\phi(r)}}{\sqrt{2\pi p_n(R;\alpha)}} \begin{pmatrix} 0 \\ X^{*n} \end{pmatrix}, \quad \text{for } 0 \leqslant n \leqslant -k-2, \quad \text{if } k < -1. \end{aligned} \tag{21}$$

Here, the normalization factors are

$$q_n(u;\alpha) = \int_0^u e^{2\alpha\phi(r)} r^{2n+1} dr, \quad p_n(u;\alpha) = \int_0^u e^{-2\alpha\phi(r)} r^{2n+1} dr. \tag{22}$$

Notice that, for $k=-1$ (in particular, when $\Phi=0$), there is no zero mode.

So, the kernel of the orthogonal projector on $\mathrm{Ker}(D_\alpha)_\kappa$ is

$$\begin{aligned} P_\alpha(x,y) &= \sum_{n=0}^{k} \frac{e^{\alpha[\phi(|X|)+\phi(|Y|)]}}{2\pi q_n(R;\alpha)} \begin{pmatrix} (XY^*)^n & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{if } k \geqslant 0, \\ P_\alpha(x,y) &= \sum_{n=0}^{-k-2} \frac{e^{-\alpha[\phi(|X|)+\phi(|Y|)]}}{2\pi p_n(R;\alpha)} \begin{pmatrix} 0 & 0 \\ 0 & (X^*Y)^n \end{pmatrix}, \quad \text{if } k < -1. \end{aligned} \tag{23}$$

Since P_α is an orthogonal projector, $P_\alpha^2 = P_\alpha$, we have

$$(\partial_\alpha P_\alpha)(1-P_\alpha) = P_\alpha(\partial_\alpha P_\alpha). \tag{24}$$

Being $(D_\alpha+P_\alpha)_\kappa$ invertible, we can define

$$\mathrm{Det}'(D_\alpha)_\kappa \equiv \mathrm{Det}(D_\alpha+P_\alpha)_\kappa, \tag{25}$$

and write

$$\frac{\mathrm{Det}'(D)_\kappa}{\mathrm{Det}(i\partial\!\!\!/)_{\kappa=0}} = \frac{\mathrm{Det}(D+P_1)_\kappa}{\mathrm{Det}(i\partial\!\!\!/+P_0)_\kappa}\,\frac{\mathrm{Det}(i\partial\!\!\!/+P_0)_\kappa}{\mathrm{Det}(i\partial\!\!\!/)_{\kappa=0}}. \tag{26}$$

As mentioned above, we can compute

$$\frac{\partial}{\partial\alpha}[\ln\,\mathrm{Det}(D_\alpha+P_\alpha)_\kappa] = \mathrm{Tr}[(\mathcal{A}+\partial_\alpha P_\alpha)G_\alpha], \tag{27}$$

after determining the Green's function $G_\alpha(x,y)$ satisfying

$$(D_\alpha+P_\alpha)G_\alpha(x,y)=\delta(x,y), \quad (\mathscr{P}_{\geqslant} \quad \mathscr{P}_{\leqslant})G_\alpha(x,y)\big|_{r=R}=0, \tag{28}$$

and applying a regularization prescription to the trace in (27). Then, by integrating in α from 0 to 1 we will get the first quotient in the r.h.s. of (26).

The computation of this Green's function will be the subject of the next section.

IV. CALCULATION OF THE GREEN'S FUNCTION

Notice that, since $(D_\alpha)_\kappa$ is self adjoint, we can write

$$G_\alpha=(1-P_\alpha)\mathscr{G}_\alpha(1-P_\alpha)+P_\alpha, \tag{29}$$

where $\mathscr{G}_\alpha(x,y)$ is the kernel of the right-inverse of $(D_\alpha)_\kappa$ on the orthogonal complement of $Ker(D_\alpha)_\kappa$. From Eq. (7), it is easy to see that $\mathscr{G}_\alpha(x,y)$ can be obtained by solving the differential equation

$$i\partial\!\!\!/\begin{pmatrix}\varphi\\ \chi\end{pmatrix}=\begin{pmatrix}f\\ g\end{pmatrix}, \tag{30}$$

for $\binom{\chi}{\varphi}$ satisfying

$$(\mathscr{P}_{\geqslant} \quad \mathscr{P}_{\leqslant})e^{\alpha\gamma_5\phi(R)}\begin{pmatrix}\varphi\\ \chi\end{pmatrix}=0. \tag{31}$$

Let us start by considering functions f and g identically vanishing in a neighbourhood of $r=0$. We expand

$$\begin{aligned}\varphi(r,\theta)&=\sum\ \varphi_n(r)e^{in\theta}, \quad \chi(r,\theta)=\sum\ \chi_n(r)e^{in\theta}\\ f(r,\theta)&=\sum\ f_n(r)e^{in\theta}, \quad g(r,\theta)=\sum\ g_n(r)e^{in\theta}.\end{aligned} \tag{32}$$

Equation (30) together with the boundary conditions at $r=R$ (Eq. (31)) and at $r=0$ (regularity), univocally determine the φ_n's (χ_n's) in terms of the f_n's (g_n's), except for $|k+1|$ of them, which correspond to the zero modes of the system. We chose the particular solution

$$\varphi(r,\theta)=i\sum_{n\geqslant k+1}e^{in\theta}r^n\int_r^R dr'(r')^{-n}g_{n+1}(r')-i\sum_{n\leqslant k}e^{in\theta}r^n\int_0^r dr'(r')^{-n}g_{n+1}(r')$$
$$=\frac{i}{2\pi}\int_r^R dr'\int_0^{2\pi}d\theta'\sum_{n\geqslant k+1}e^{in(\theta-\theta')}\left(\frac{r}{r'}\right)^n e^{-i\theta'}g(r',\theta')$$
$$-\frac{i}{2\pi}\int_0^r dr'\int_0^{2\pi}d\theta'\sum_{n\leqslant k}e^{in(\theta-\theta')}\left(\frac{r}{r'}\right)^n e^{-i\theta'}g(r',\theta'),\tag{33}$$

and

$$\chi(r,\theta)=\frac{i}{2\pi}\int_r^R dr'\int_0^{2\pi}d\theta'\sum_{n\leqslant k+1}e^{in(\theta-\theta')}\left(\frac{r'}{r}\right)^n e^{i\theta'}f(r',\theta')$$
$$-\frac{i}{2\pi}\int_0^r dr'\int_0^{2\pi}d\theta'\sum_{n\geqslant k+2}e^{in(\theta-\theta')}\left(\frac{r'}{r}\right)^n e^{i\theta'}f(r',\theta').\tag{34}$$

Therefore, calling

$$\mathscr{G}^{(0)}(x,y)=\frac{1}{2\pi i}\begin{pmatrix}0 & \dfrac{1}{X-Y}\left(\dfrac{X}{Y}\right)^{k+1}\\ \dfrac{1}{X^*-Y^*}\left(\dfrac{Y^*}{X^*}\right)^{k+1} & 0\end{pmatrix},\tag{35}$$

we have

$$\begin{pmatrix}\varphi\\ \chi\end{pmatrix}(x)=\int d^2y\,\mathscr{G}^{(0)}(x,y)\begin{pmatrix}f\\ g\end{pmatrix}(y).\tag{36}$$

The particular solution in (33) and (34) makes $\mathscr{G}^{(0)}(x,y)$ Hermitian (although this will prove to be not essential).

Notice that $\binom{\varphi}{\chi}$ so constructed, satisfying Eq.(31), is neither in $[\mathrm{Ker}((D_\alpha)_\kappa e^{\alpha\gamma_5\phi})]^\perp$, nor in L_2 for arbitrary $\binom{f}{g}$ (as expected from the existence of zero modes).

Restricting $\binom{\varphi}{\chi}$ to $[\mathrm{Ker}((D_\alpha)_\kappa e^{\alpha\gamma_5\phi})]^\perp$ and $\binom{f}{g}$ to $[\mathrm{Ker}((D_\alpha)_\kappa e^{-\alpha\gamma_5\phi})]^\perp$ corresponds to the composition

$$\mathscr{G}_\alpha=(1-P_\alpha)e^{\alpha\gamma_5\phi}\mathscr{G}^{(0)}e^{\alpha\gamma_5\phi}(1-P_\alpha),\tag{37}$$

which is an integral operator with locally integrable kernel (an anti-diagonal matrix whose only singularity is at $Y=X$). It is worth to stress that, due to the presence of the projector $(1-P_\alpha)$, $\mathscr{G}_\alpha$ is independent of the particular choice made in (33) and (34).

The addition of P_α [Eq. (23)] leads, for $k\geqslant 0$, to

$$G_\alpha(x,y)$$
$$=\frac{1}{2\pi i}\begin{pmatrix}ie^{\alpha[\phi(|X|)+\phi(|Y|)]}\displaystyle\sum_{n=0}^{k}\frac{(XY^*)^n}{q_n(R;\alpha)} & e^{\alpha[\phi(|X|)-\phi(|Y|)]}\left[\dfrac{1}{X-Y}+\dfrac{1}{Y}\displaystyle\sum_{n=0}^{k}\left(\frac{X}{Y}\right)^n\frac{q_n(|Y|;\alpha)}{q_n(R;\alpha)}\right]\\ e^{-\alpha[\phi(|X|)-\phi(|Y|)]}\left[\dfrac{1}{X^*-Y^*}-\dfrac{1}{X^*}\displaystyle\sum_{n=0}^{k}\left(\frac{Y^*}{X^*}\right)^n\frac{q_n(|X|;\alpha)}{q_n(R;\alpha)}\right] & 0\end{pmatrix},\tag{38}$$

which is the inverse of $(D_\alpha+P_\alpha)_\kappa$, as can be easily verified by direct inspection. For the $k<-1$ case, we get

$$G_\alpha(x,y) = \frac{1}{2\pi i}\begin{pmatrix} 0 & e^{\alpha[\phi(|X|)-\phi(|Y|)]}\left[\frac{1}{X-Y}-\frac{1}{X}\sum_{n=0}^{-k-2}\left(\frac{Y}{X}\right)^n\frac{p_n(|X|;\alpha)}{p_n(R;\alpha)}\right] \\ e^{-\alpha[\phi(|X|)-\phi(|Y|)]}\left[\frac{1}{X^*-Y^*}+\frac{1}{Y^*}\sum_{n=0}^{-k-2}\left(\frac{X^*}{Y^*}\right)^n\frac{p_n(|Y|;\alpha)}{p_n(R;\alpha)}\right] & ie^{-\alpha[\phi(|X|)+\phi(|Y|)]}\sum_{n=0}^{-k-2}\frac{(X^*Y)^n}{p_n(R;\alpha)} \end{pmatrix}. \tag{39}$$

V. EVALUATION OF THE DETERMINANT

Now we have all the necessary elements to compute the first quotient in the r.h.s. of Eq. (26). In fact, from Eqs. (29) and (24)

$$\mathrm{Tr}[(\partial_\alpha P_\alpha)G_\alpha]=0. \tag{40}$$

So, Eq. (27) becomes

$$\frac{\partial}{\partial\alpha}[\ln\,\mathrm{Det}(D_\alpha+P_\alpha)_\kappa]=\mathrm{Tr}[\not{A}G_\alpha]. \tag{41}$$

As usual, the kernel of the operator inside the trace is singular at the diagonal, so we must introduce a regularization. We will employ a point-splitting one where, following Schwinger,[17] we will introduce a phase factor in order to preserve gauge invariance. For $k\geqslant 0$, from (38), we take

$$\begin{aligned}\mathrm{Tr}[\not{A}G_\alpha] &\equiv \mathrm{Reg.Part}\left\{\int_{r<R} d^2x\ \mathrm{tr}[\not{A}(x)G_\alpha(x,x+\epsilon)e^{i\alpha(\epsilon_r A_r(x)+\epsilon_\theta A_\theta(x))}]\right\}_{\epsilon\to 0}\\ &= -\frac{1}{2\pi}\mathrm{Reg.Part}\int_{r<R} d^2x\ \phi'(r)\left\{e^{-i\theta}e^{\alpha\phi'(r)[\epsilon_r-i\epsilon_\theta]}\left[\frac{e^{i\theta}}{\epsilon_r-i\epsilon_\theta}+\frac{1}{X^*}\sum_{n=0}^{k}\frac{q_n(|X|;\alpha)}{q_n(R;\alpha)}\right.\right.\\ &\quad\left.+O(\epsilon)\right]+e^{i\theta}e^{-\alpha\phi'(r)[\epsilon_r+i\epsilon_\theta]}\left[-\frac{e^{-i\theta}}{\epsilon_r+i\epsilon_\theta}+\frac{1}{X}\sum_{n=0}^{k}\frac{q_n(|X|;\alpha)}{q_n(R;\alpha)}+O(\epsilon)\right]\Bigg\}_{\epsilon\to 0}\\ &= \mathrm{Reg.Part}\left\{-\left[\frac{2i\epsilon_\theta}{\epsilon_r^2+\epsilon_\theta^2}\right]\int_0^R dr\ r\phi'(r)-\frac{\alpha}{\pi}\int_{r<R} d^2x\phi'^2-2(k+1)\phi(R)\right.\\ &\quad\left.+\sum_{n=0}^{k}\frac{\partial}{\partial\alpha}\ln[q_n(R;\alpha)]\right\}_{\epsilon\to 0}.\end{aligned} \tag{42}$$

The regular part at $\epsilon=0$ drops out the first term and, integrating in α from 0 to 1, we get

$$\ln\left[\frac{\mathrm{Det}(D+P_1)_\kappa}{\mathrm{Det}(i\not{\partial}+P_0)_\kappa}\right]=-\frac{1}{2\pi}\int_{r<R} d^2x\phi'^2-2(k+1)\phi(R)+\sum_{n=0}^{k}\ln\left[2(n+1)\frac{q_n(R;1)}{R^{2(n+1)}}\right]. \tag{43}$$

When there are no zero modes ($k+1=0$) only the first term in the r.h.s. survives. For $k<-1$, from (40), one gets a similar expression where the sum runs from 0 to $-k-2=|k+1|-1$, and the norms q_n are replaced by p_n, defined in Eq. (22). This result can also be obtained by performing the change $\phi\to-\phi$ and $k+1\to-(k+1)$. Moreover, as expected from (7), this expression is invariant under the shift $\phi(r)\to\phi(r)+\epsilon$, for any constant ϵ.

In the following, we will obtain the second quotient of determinants in Eq. (26) by computing explicitly the spectra of the free Dirac operators and using a ζ-function regularization. The equation to be solved is

$$i\not{\partial}\psi=\lambda\psi, \tag{44}$$

or, explicitly

$$i\begin{pmatrix} 0 & e^{-i\theta}\left(\partial_r-\frac{i}{r}\partial_\theta\right) \\ e^{i\theta}\left(\partial_r+\frac{i}{r}\partial_\theta\right) & 0 \end{pmatrix}\begin{pmatrix}\varphi(r,\theta)\\ \chi(r,\theta)\end{pmatrix}=\lambda\begin{pmatrix}\varphi(r,\theta)\\ \chi(r,\theta)\end{pmatrix}. \tag{45}$$

Now, by expanding

$$\varphi(r,\theta)=\sum_{n=-\infty}^{\infty}\varphi_n(r)e^{in\theta},\quad \chi(r,\theta)=\sum_{n=-\infty}^{\infty}\chi_n(r)e^{in\theta}, \tag{46}$$

we get

$$\begin{aligned} i\varphi_n'(z)-i\frac{n}{z}\varphi_n(z)&=\frac{\lambda}{|\lambda|}\chi_{n+1}(z),\\ i\chi_{n+1}'(z)+i\frac{(n+1)}{z}\chi_{n+1}(z)&=\frac{\lambda}{|\lambda|}\varphi_n(z), \end{aligned} \tag{47}$$

where $z=|\lambda|r$. So, $\varphi_n(z)$ satisfies

$$\varphi_n''(z)+\frac{1}{z}\varphi_n'(z)+\left(1-\frac{n^2}{z^2}\right)\varphi_n(z)=0. \tag{48}$$

Therefore, the general solution of Eq. (46) is

$$\psi(r,\theta)=\sum_{n=-\infty}^{\infty}\psi_n\begin{pmatrix} J_n(|\lambda|r)e^{in\theta}\\ -i\frac{|\lambda|}{\lambda}J_{n+1}(|\lambda|r)e^{i(n+1)\theta}\end{pmatrix}. \tag{49}$$

Now, by imposing the global boundary conditions given in Eq. (11), it follows that: If $n\geqslant k+1$, the upper component must vanish at the boundary, which implies $\lambda=\pm j_{n,l}/R$ ($j_{n,l}$ is the l-th zero of $J_n(z)$). Analogously, if $n\leqslant k$, it is the lower component which must vanish, and so $\lambda=\pm j_{n+1,l}/R$. Therefore, the eigenvalues are

$$\lambda_{n,l}=\pm j_{n,l}/R,\quad \text{for } n=0,\pm1,\pm2,\ \ldots \quad \text{and } l=1,2,\ \ldots\ . \tag{50}$$

Notice that $j_{-n,l}=j_{n,l}$, and that, for $n=k+1$ the eigenvalues appear twice, once for an eigenfunction with vanishing upper component at the boundary, and once for another one with vanishing lower component.

From these eigenvalues we can construct the ζ-function

$$\zeta_{(i\not{\partial}+P_0)_\kappa}(s)=|k+1|+(1+e^{-i\pi s})\left\{\sum_{n=-\infty}^{n=\infty}\sum_{l=1}^{\infty}\left(\frac{j_{n,l}}{R}\right)^{-s}+\sum_{l=1}^{\infty}\left(\frac{j_{|k+1|,l}}{R}\right)^{-s}\right\}. \tag{51}$$

The first term, $|k+1|$, is the multiplicity of the 0-eigenvalue of $(i\partial\!\!\!/)_\kappa$. It is also interesting to note that the double sum in the r.h.s. (which is independent of k) corresponds to the ζ-function of the Laplacian on a disk with Dirichlet boundary conditions, thus being analytic at $s=0$.[12] The second sum, to be explicitly computed below, is regular at $s=0$. Then $\zeta_{(i\partial\!\!\!/+P_0)_\kappa}(s)$ is regular at the origin. As far us we know, the regularity of the ζ at the origin for non local boundary conditions has not been established in general.[18]

From (52), we can write

$$\begin{aligned}\ln\left[\frac{\mathrm{Det}(i\partial\!\!\!/+P_0)_\kappa}{\mathrm{Det}(i\partial\!\!\!/)_{\kappa=0}}\right] &= -\frac{d}{ds}[\zeta_{(i\partial\!\!\!/+P_0)_\kappa}(s)-\zeta_{(i\partial\!\!\!/)_{\kappa=0}}(s)]_{s=0}\\ &= -\frac{d}{ds}\left[(1+e^{-i\pi s})R^s\left\{\sum_{l=1}^{\infty}(j_{|k+1|,l})^{-s}-\sum_{l=1}^{\infty}(j_{0,l})^{-s}\right\}\right]_{s=0}\\ &= -2\left[f'_{|k+1|}(0)-f'_0(0)+\left(\ln R-\frac{i\pi}{2}\right)[f_{|k+1|}(0)-f_0(0)]\right],\end{aligned} \tag{52}$$

where the function f is defined as

$$f_\nu(s)\equiv\sum_{l=1}^{\infty}(j_{\nu,l})^{-s}. \tag{53}$$

Taking into account the asymptotic expansion for the zeros of Bessel functions[19]

$$(j_{\nu,l})^{-s}-(l\pi)^{-s}+s\left(\frac{2\nu-1}{4}\right)\pi(l\pi)^{-s-1}\sim sO(l^{-s-2}), \tag{54}$$

we can write

$$f_\nu(s)=\sum_{l=1}^{\infty}\left[(j_{\nu,l})^{-s}-(l\pi)^{-s}+s\left(\frac{2\nu-1}{4}\right)\pi(l\pi)^{-s-1}\right]+\pi^{-s}\left[\zeta(s)-s\left(\frac{2\nu-1}{4}\right)\zeta(s+1)\right], \tag{55}$$

where $\zeta(s)$ is Riemann's ζ-function. By (55), the first series converges for Re $s>-1$, and equals 0 at $s=0$. Thus

$$f_\nu(0)=-\frac{\nu}{2}-\frac{1}{4}, \tag{56}$$

and

$$f'_\nu(0)=-\frac{1}{2}\ln 2+\left(\frac{2\nu-1}{4}\right)(\ln\pi-\gamma)-\sum_{l=1}^{\infty}\ln\left[\frac{j_{\nu,l}}{l\pi}e^{-([2\nu-1/4l])}\right], \tag{57}$$

where γ is Euler's constant. By replacing Eqs. (57) and (58) into Eq. (53), we get

$$\ln\left[\frac{\mathrm{Det}(i\partial\!\!\!/+P_0)_\kappa}{\mathrm{Det}(i\partial\!\!\!/)_{\kappa=0}}\right]=-|k+1|\left[\frac{i\pi}{2}-\gamma-\ln\left(\frac{R}{\pi}\right)\right]+2\sum_{l=1}^{\infty}\ln\left[\frac{j_{|k+1|,l}}{j_{0,l}}e^{-([|k+1|/2l])}\right]. \tag{58}$$

Finally, adding (44) and (59), and taking into account that we have used a gauge invariant procedure, we obtain for the quotient in the l.h.s. of (26)

$$\ln\left[\frac{\mathrm{Det}(D+P_1)_\kappa}{\mathrm{Det}(i\partial\!\!\!/)_{\kappa=0}}\right]=-\frac{1}{2\pi}\int_{|x|<R}\int_{|y|<R}d^2x\ d^2y A_\mu(x)[\delta_{\mu\nu}\delta(x-y)-\partial_\mu\partial_\nu\partial^{-2}(x-y)]A_\nu(y)$$
$$-2(k+1)\phi(R)+\sum_{n=0}^{k}\ln\left[2(n+1)\frac{q_n(R;1)}{R^{2(n+1)}}\right]-|k+1|\left[\frac{i\pi}{2}-\gamma-\ln\left(\frac{R}{\pi}\right)\right]$$
$$+2\sum_{l=1}^{\infty}\ln\left[\frac{j_{|k+1|,l}}{j_{0,l}}e^{-(|k+1|/2l)}\right]. \tag{59}$$

The first term is the integral on the disk of the same expression appearing in the well-known result for the boundaryless case.[20] Note that this final result is also invariant under the transformation: $\phi\to\phi+\epsilon$ for any constant ϵ.

Now, a comment is in order concerning global axial transformations and their relationship to zero modes. Under such a transformation, with constant ϵ,

$$e^{-\gamma_5\epsilon}(i\partial\!\!\!/)_{\kappa=0}e^{-\gamma_5\epsilon}=(i\partial\!\!\!/)_{\kappa=0} \tag{60}$$

is invariant. Moreover

$$e^{-\gamma_5\epsilon}(D+P_1)_\kappa e^{-\gamma_5\epsilon}=(D+e^{-\gamma_5\epsilon}P_1e^{-\gamma_5\epsilon})_\kappa, \tag{61}$$

while its inverse has the structure

$$G^{(\epsilon)}(x,y)=(1-P_1)\mathcal{G}(x,y)(1-P_1)+e^{\gamma_5\epsilon}P_1\ e^{\gamma_5\epsilon}. \tag{62}$$

Therefore, since γ_5 leaves Ker $(D)_\kappa$ invariant,

$$\frac{\partial}{\partial\epsilon}\ln\left[\frac{\mathrm{Det}(e^{-\gamma_5\epsilon}(D+P_1)_\kappa e^{-\gamma_5\epsilon})}{\mathrm{Det}(e^{-\gamma_5\epsilon}(i\partial\!\!\!/)_{\kappa=0}e^{-\gamma_5\epsilon})}\right]=-\mathrm{Tr}[e^{-\gamma_5\epsilon}\{\gamma_5,P_1\}e^{-\gamma_5\epsilon}G^{(\epsilon)}]$$
$$=-2\ \mathrm{Tr}[\gamma_5P_1]=-2(N_+-N_-), \tag{63}$$

where $N_{+(-)}$ is the number of positive(negative) chirality zero modes.

It can be verified that our strategy leads to the correct result for the index of $(D)_\kappa$. By following the same procedure that lead to Eq. (60), we can compute the quotient of determinants in the l.h.s of (64). In fact, using Eq. (63) instead of Eq. (29), the only difference appears in the first term of the r.h.s. of (52), where a factor $e^{\pm 2\epsilon s}$ arises. Thus, after performing the ϵ-derivative

$$N_+-N_-=k+1, \tag{64}$$

which agrees with our previous result for the number of zero modes.

For the sake of completeness, in the next section we will show how the presence of the σ's in (7) and (11) affects the index theorem.

VI. RELATION TO THE INDEX THEOREM

Here we will reobtain the index by following the Atiyah–Patodi–Singer approach.[13]

The Dirac operator in polar coordinates can also be written as[15,16]

$$D=\frac{1}{\sqrt{r}}\begin{pmatrix}0 & e^{-i\theta/2}(\partial_r+B)e^{-i\theta/2}\\ e^{i\theta/2}(-\partial_r+B)e^{i\theta/2} & 0\end{pmatrix}\sqrt{r}, \tag{65}$$

subject to the boundary conditions in Eq. (11). It is to be stressed that the presence of the factors $e^{\pm i\theta/2}$ coming from the σ's will produce a departure from the index formula in Ref. 13.

To be be consistent with (11) we call $\mathscr{D}$ the differential operator $(-\partial_r+B)$ acting on functions

$$f(r,\theta)=\sum_{n=-\infty}^{\infty} f_n(r)e^{i(n+1/2)\theta} \tag{66}$$

satisfying

$$\mathscr{P}_{\geqslant}[e^{-i\theta/2}f(R,\theta)]=0. \tag{67}$$

Its adjoint, $\mathscr{D}^+$, is the differential operator (∂_r+B) defined on the space of functions

$$g(r,\theta)=\sum_{n=-\infty}^{\infty} g_n(r)e^{i(n-1/2)\theta} \tag{68}$$

satisfying

$$\mathscr{P}_{\leqslant}[e^{i\theta/2}g(R,\theta)]=0. \tag{69}$$

Then $D^+D=(1/\sqrt{r})\,e^{-i\theta/2}\,\mathscr{D}^+\mathscr{D}\,e^{i\theta/2}\sqrt{r}$, where $\mathscr{D}^+\mathscr{D}=-\partial_r^2+B^2$ acting on functions satisfying

$$\begin{cases}\mathscr{P}_{\geqslant}[e^{-i\theta/2}f(R,\theta)]=0,\\ \mathscr{P}_{\leqslant}[e^{i\theta/2}\mathscr{D}f(R,\theta)]=0,\end{cases} \tag{70}$$

or, equivalently,

$$\begin{cases} f_n(R)=0, & \text{for } n\geqslant k+1\\ -f_n'(R)+\dfrac{n+1/2-\kappa}{R}f_n(R)=0, & \text{for } n\leqslant k.\end{cases} \tag{71}$$

Similarly, $\mathscr{D}\mathscr{D}^+=-\partial_r^2+B^2$ acting on functions $g(r,\theta)$ satisfying

$$\begin{cases} g_n(R)=0, & \text{for } n\leqslant k+1\\ g_n'(R)+\dfrac{n+1/2-\kappa}{R}g_n(R)=0, & \text{for } n\geqslant k+2.\end{cases} \tag{72}$$

Now, we have for the heat kernels

$$\mathrm{Tr}\{e^{-tD^+D}\}-\mathrm{Tr}\{e^{-tDD^+}\}=\mathrm{Tr}\{e^{-t\mathscr{D}^+\mathscr{D}}\}-\mathrm{Tr}\{e^{-t\mathscr{D}\mathscr{D}^+}\}. \tag{73}$$

When written in this fashion, the boundary contribution to the r.h.s. in (73), $K(t)$, can be easily constructed from Eqs. (2.16) and (2.17) in Ref. 13, thus getting

$$K(t)=\frac{1}{2}\,\mathrm{erfc}\left(-\left(k+\frac{1}{2}-\kappa\right)\sqrt{t}\,\right)-\sum_{n\neq k}\frac{\mathrm{sig}(n+\frac{1}{2}-\kappa)}{2}\,\mathrm{erfc}\left(\left|\,n+\frac{1}{2}-\kappa\right|\,\sqrt{t}\,\right). \tag{74}$$

Taking into account that

$$\lim_{t\to\infty} K(t) = m(\kappa) = \begin{cases} 0 & \text{for } k+\frac{1}{2} < \kappa \leqslant k+1 \\ \frac{1}{2} & \text{for } \kappa = k+\frac{1}{2}, \\ 1 & \text{for } k < \kappa < k+\frac{1}{2} \end{cases} \tag{75}$$

we obtain

$$\int_0^\infty dt\ t^{s-1}\{K(t)-m(\kappa)\} = \frac{-\Gamma(s+1/2)}{2s\sqrt{\pi}}\,\eta_{(B+\,1/2R)}(2s), \tag{76}$$

where

$$\eta_{(B+\,1/2R)}(s) = R^s \sum_{n\neq\kappa-\,1/2} \mathrm{sig}\left(n+\frac{1}{2}-\kappa\right)\left|n+\frac{1}{2}-\kappa\right|^{-s}. \tag{77}$$

In particular

$$\eta_{(B+\,1/2R)}(0) = \begin{cases} 2(\kappa-k-1) & \text{for } k+\frac{1}{2} < \kappa \leqslant k+1 \\ 0 \quad \text{for } \kappa = k+\frac{1}{2} \\ 2(\kappa-k) & \text{for } k < \kappa < k+\frac{1}{2} \end{cases}. \tag{78}$$

Finally, taking into account that, for $k<\kappa\leqslant k+1$,

$$m(\kappa) - \tfrac{1}{2}\eta_{(B+\,1/2R)}(0) = k+1-\kappa = \tfrac{1}{2}[1-h(B)-\eta_{(B)}(0)], \tag{79}$$

where $h(B)=\dim\ker(B)$, and from the asymptotic expansion[13] of the heat kernels in (74), we have

$$\text{index } D = \kappa + \frac{[1-h(B)-\eta_{(B)}(0)]}{2} = k+1, \tag{80}$$

in agreement with (64). The first term in the intermediate expression is the well known contribution from the bulk.[20] The second one is the boundary contribution of APS, shifted by 1/2. This correction, due to the presence of the factor σ in (7), has already been obtained in[15] with slightly different spectral boundary conditions.

VII. CONCLUSIONS

In this paper we have achieved the complete evaluation of the determinant of the Dirac operator on a disk, in the presence of an axially symmetric flux, under global boundary conditions of the type intoduced by Atiyah, Patodi, and Singer.[13] To this end, we have proceeded in two steps: In the first place, we have grown the gauge field while keeping the boundary condition fixed. This calculation was possible thanks to the exact knowledge of the zero modes and the Green's function (in the complement of the null space.) Here, a gauge invariant point splitting regularization was employed. In references[10,11] we developed, for the case of local boundary conditions, a regularization scheme based on Seeley's complex powers.[12] Its application to the present problem leads to the same result as the point splitting we used here, even though its relation to the ζ-function is not guaranteed. In fact, as far as we know, the construction of complex powers for elliptic boundary problems with global boundary conditions is still under study.[18]

In the second step, we have explicitly obtained the eigenvalues of $(i\partial\!\!\!/ + P_0)_\kappa$. We have shown that the corresponding ζ-function is regular at the origin and we have evaluated the quotient of the free Dirac operators for two different global boundary conditions.

We have verified that our complete result is in agreement with the APS index theorem, adapted to our example.

ACKNOWLEDGMENT

We thank María Amelia Muschietti and Jorge Solomin for valuable discussions. This work was partially supported by CONICET, Argentina.

[1] M. Bordag, B. Geyer, K. Kirsten, and E. Elizalde, Commun. Math. Phys. **179**, 215 (1996).
[2] J. S. Dowker and J. S. Apps, Int. J. Mod. Phys. D **5**, 799 (1996).
[3] E. Elizalde, M. Lygren, and D. V. Vasilevich, J. Math. Phys. **37**, 3105 (1996).
[4] Klaus Kirsten and Guido Cognola, Class. Quantum Grav. **13**, 633 (1996).
[5] Peter D. D'Eath and Giampiero Esposito, Phys. Rev. D **43**, 3234 (1991).
[6] Peter D. D'Eath and Giampiero Esposito, Phys. Rev. D **44**, 1713 (1991).
[7] A. Wipf and S. Durr, Nucl. Phys. B **443**, 201 (1995).
[8] M. De Francia, H. Falomir, and E. M. Santangelo, Phys. Rev. D **45**, 2129 (1992).
[9] M. De Francia, Phys. Rev. D **50**, 2908 (1994).
[10] H. Falomir, R. E. Gamboa Saraví, M. A. Muschietti, E. M. Santangelo, and J. E. Solomin, J. Math. Phys. **37**, 5805 (1996).
[11] H. Falomir, R. E. Gamboa Saraví, M. A. Muschietti, E. M. Santangelo, and J. E. Solomin, On the relation between determinants and Green functions of Dirac operators with local boundary conditions, Bulletin des Sciences Mathématiques (in press).
[12] R. T. Seeley, Am. J. Math. **91**, 889 (1969); **91**, 963 (1969).
[13] M. F. Atiyah, V. K. Patodi, and I. M. Singer, Math. Proc. Camb. Philos. Soc. **77**, 43 (1975).
[14] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[15] A. V. Mishchenko and Yu. A. Sitenko, Ann. Phys. (N.Y.) **218**, 199 (1992).
[16] E. F. Moreno, Phys. Rev. D **48**, 921 (1993).
[17] J. Schwinger, Phys. Rev. **82**, 664 (1951); **128**, 2425 (1962).
[18] G. Grubb and R. T. Seeley, C. R. Acad. Sci., Ser. I: Math. **317**, 1123 (1993); Zeta and eta fuctions for Atiyah–Patodi–Singer operators, Copenh. Univ., Math. Dept, Preprint Ser. (1993), No. 11; Weakly parametric pseudodifferentials operators and Atiyah–Patodi–Singer boundary problems, Copenh. Univ., Math. Dept, Preprint Ser. (1993), No. 15.
[19] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980).
[20] See, for example, R. E. Gamboa Saraví, M. A. Muschietti, F. A. Schaposnik, and J. E. Solomin, Ann. Phys. (N.Y.) **157**, 360 (1984).

On pairs of difference operators satisfying $[D, X] = \mathrm{Id}$

Andrzej Z. Górski[a)]
Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208

Jacek Szmigielski[b)]
Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5E6, Canada

(Received 4 April 1997; accepted for publication 17 September 1997)

Dedicated to the memory of Gerda Dorfmeister.

Different finite difference replacements for the derivative are analyzed in the context of the Heisenberg commutation relation. The type of the finite difference operator is shown to be tied to whether one can naturally consider D and X to be self-adjoint and skew self-adjoint or whether they have to be viewed as creation and annihilation operators. The first class, generalizing the central difference scheme, is shown to give unitary equivalent representations. For the second case we construct a large class of examples, generalizing previously known difference operator realizations of $[D, X] = \mathrm{Id}$. © *1998 American Institute of Physics.* [S0022-2488(98)01001-9]

I. INTRODUCTION

The idea of discrete physics and discrete space–time is a very old one. To the best of our knowledge, the oldest references in a physical journal are Refs. 1 and 2, and it has been reconsidered many times in the past (see Ref. 3 for an extensive bibliographical review on the subject). The basic motivations for all those efforts were: (a) the presence of ultraviolet infinities in the standard quantum field theory, and (b) understanding the origin of the fundamental length (mass) scale in the Einstein's general relativity theory.

In recent years there has been a growing number of attempts to find an underlying discrete structure of space–time. At the same time interesting discrete structures have surprisingly emerged within originally continuous models. The most interesting examples can be found in string theory,[4] Roger Penrose's spin network calculus,[5] the loop representation of quantum gravity[6] (see also Ref. 7), the Bekenstein black hole entropy problem,[8–12] or the recent approach advocated by G. t' Hooft.[13–15] Also, it is interesting to mention in this context the old work by Aharonov *et al.*,[16] where the modular variables for the coordinate and momentum operators were introduced to describe an infinite slit experiment with magnetic field.

Increasing popularity of discrete models in theoretical physics has been stimulated by several successful attempts to discretize continuous models for technical reasons. This has advanced our understanding of discrete techniques. Here, the best examples are the lattice QCD[17] and the Regge calculus in classical gravity.[18]

On the other hand, there have been several attempts to incorporate discrete space–time at the *kinematical* level and to investigate its physical consequences.[19,20] This line of research can be also associated with the q-deformations of the Poincare group.[21] Another interesting approach has been recently proposed by Mazur,[12,22] who has given additional physical arguments in favor of space–time discreteness.

[a)]The Kościuszko Foundation Fellow. Permanent address: Institute of Nuclear Physics, Radzikowskiego 152, 31–342 Kraków, Poland. Electronic mail: gorski@alf.ifj.edu.pl
[b)]Electronic mail: szmigiel@math.usask.ca

We assume a discrete coordinate space (see also Ref. 23) and investigate its consequences for the Heisenberg commutation relations. This approach is purely kinematical, as the Heisenberg relations are, and does not depend on the details of the underlying dynamics. Such ''unusual'' realizations of the Heisenberg algebra have also been found in 2D gravity.[24] These representations are not unitarily equivalent to the Schrödinger representation. In particular, they cannot be exponentiated to the Weyl form of the canonical commutation relation.

To motivate our discussion we first consider the standard quantum mechanical representation of the momentum operator:

$$p_\mu = -i\hbar \frac{\partial}{\partial x^\mu}. \tag{1.1}$$

Hence, the coordinate discretization can be implemented through a derivative discretization:

$$\frac{d}{dx} \to D_{\Delta x}, \tag{1.2}$$

where $D_{\Delta x}$ is a discretized derivative (difference) operator and Δx is the discretization parameter (usually a fixed coordinate spacing). Even though we will call $D_{\Delta x}$ a discretized derivative it should be noted at this point that initially this is just some difference operator acting, for example, on the space of smooth functions on $\mathbb{R}$. However, such operators can be naturally restricted to act on the space of functions on $\Delta x \times \mathbb{Z}$. It is in this sense that we talk about a discrete derivative.

We will consider a wide class of possible derivative discretization schemes of the form

$$D_{\Delta x}(M+N+1) \equiv D_{\Delta x} \equiv \sum_{k=-M}^{+N} \alpha_k E_{\Delta x}^k, \tag{1.3}$$

where α_k are real constant coefficients and M and N are integer indices corresponding to the lowest and highest nonzero terms (hence, $\alpha_{-M} \cdot \alpha_N \neq 0$). We take the coefficients α_k to be constant as one would like to have the same definition of the discrete derivative at all points in the coordinate space. The expression will be called n-point for $n=M+N+1$. For convenience, let us define the shift operator $E_{\Delta x}^n$ as

$$E_{\Delta x}^n \, f(x) \equiv E^n f(x) = f(x+n\Delta x). \tag{1.4}$$

One can easily check that this operator fulfills the following commutation relations:

$$\begin{aligned} [E^m, E^n] &= 0, \\ [E^m, x] &= m\Delta x E^m, \\ [E^m, xE^n] &= m\Delta x E^{m+n}, \\ [E^m, E^n x] &= m\Delta x E^{m+n}. \end{aligned} \tag{1.5}$$

Several authors have used a purely imaginary Δx and α_k. This case is not addressed in the present paper and our methods are not directly applicable to it.

In addition, the formal continuum limit reads

$$\lim_{\Delta x \to 0} E_{\Delta x}^n = 1. \tag{1.6}$$

In the simplest case, $N=M=1$ the difference operator can be expressed by the Heine bracket $[\cdot]_q$ defined for the dimensionless variables $P.q$ as (see Appendix A for its basic properties)

$$H(P,q)\equiv[P]_q=\frac{q^P-q^{-P}}{q-q^{-1}}, \tag{1.7}$$

where $q=e^i$ and $P=-i(\partial/\partial x)$. This notation is widely used in the context of q-deformations and, in particular, to define the q-deformed Poincaré group[21] (see also Ref. 22).

This paper is organized as follows. In Sec. II we investigate the basic properties of the difference operator (1.3). Section III is devoted to a construction of the Heisenberg algebra for the operator (1.3) and its conjugate operator $X_{\Delta x}$ when the latter is a bounded operator (finite series). In Sec. IV we consider the case of unbounded operators. In Sec. V we introduce the momentum formulation of our problem and within this representation we complete our proofs in the unbounded case in Sec. VI. Section VII contains the summary of our results.

II. THE DISCRETE DERIVATIVE

In this section we shall investigate the general properties of the discretized derivative (1.3) and the corresponding momentum operator. We would like to be able to talk about the convergence of discretized operators to the ordinary ones when the parameter Δx tends to zero and whenever we can make sense of this limit we will call it *the classical limit*. In particular, we stipulate that for analytic functions

$$D_{\Delta x}\, f(x)\to\frac{d}{dx}\, f(x), \quad \Delta x\to 0, \tag{2.1}$$

holds pointwise in x. Consequently, by writing (2.1) in terms of the translation operator $\epsilon^{k\Delta x}d/dx$ we obtain

$$D_{\Delta x}\, f(x)\equiv\sum_{k=-M}^{+N}\alpha_k\, e^{k\Delta x d/dx}f(x). \tag{2.2}$$

Applying the Taylor expansion to (1.3) at $x=x_i$, we have

$$D_{\Delta x}\, f(x_i)=f(x_i)\sum_{k=-M}^{+N}\alpha_k+\frac{1}{1!}\frac{df(x_i)}{dx}\sum_{k=-M}^{+N}\alpha_k\, k\Delta x+\cdots+\frac{1}{n!}\, f^{(n)}(x_i)\sum_{k=-M}^{+N}\alpha_k k^n\Delta x^n+\cdots . \tag{2.3}$$

From (2.3), we then get the following two conditions for the coefficients α_k to fulfill (2.1):

$$\sum_{k=-M}^{+N}\alpha_k=0, \quad \sum_{k=-M}^{+N}k\alpha_k=\frac{1}{\Delta x}. \tag{2.4}$$

We may also sharpen the notion of a classical limit by specifying the accuracy of the discretization scheme. For example, the coefficients α_k can be chosen to give the best fit to the ordinary derivative by requiring that all terms in (2.3) vanish up to the order Δx^{M+N}. As a result, in addition to (2.4), the following set of algebraic equations for the coefficients must be satisfied:

$$\sum_{k=-M}^{k=+N}k^n\alpha_k=0 \quad \text{for } n=2,3,...,(M+N), \tag{2.5}$$

and we get instead of (2.3) the following approximation for the derivative:

TABLE I. Values of the lowest optimal α coefficients.

$N=M$	α_1	α_2	α_3	α_4	α_5	α_6
1	1/2					
2	2/3	−1/12				
3	3/4	−3/20	1/60			
4	4/5	−1/5	4/105	−1/280		
5	5/6	−5/21	5/84	−5/504	1/1260	
6	6/7	−15/56	5/63	−1/56	1/385	−1/5544

$$D_{\Delta x}f(x_i)=f'(x_i)+\frac{\Delta x^{M+N+1}}{(M+N+1)!}\,f^{(M+N+1)}(x_i)\sum_{k=-M}^{+M}\alpha_k k^{M+N+1}+o(\Delta x^{M+N+1}). \tag{2.6}$$

Equations (2.4) and (2.5) comprise the set of $M+N+1$ linear equations for $M+N+1$ coefficients α_k and they can be solved explicitly. The corresponding determinant is the *Vandermonde* determinant and its value is $\det\|A\|=1!\cdot 2!\cdot\cdots\cdot(2N)!\neq 0$. Consequently, there exists a unique solution. Its form is given by the following:

Theorem 1: *The discrete derivative of the form (1.3) gives the best fit to the ordinary derivative, i.e., it has expansion (2.3) with the first nonvanishing coefficient for* Δx^{M+N+1}*, if and only if the coefficients* α_k *are of the following form*:

$$\begin{aligned}\alpha_0&=\frac{-1}{\Delta x}\left[\sum_{k=1}^{N}\frac{1}{k}-\sum_{k=1}^{M}\frac{1}{k}\right],\\ \alpha_k&=\frac{(-1)^{k+1}}{\Delta x}\,\frac{M!N!}{k(M+k)!(N-k)!}.\end{aligned} \tag{2.7}$$

Higher-order terms in the expansion (2.3) cannot vanish, as the system would become overdetermined.

In the symmetric case, i.e., when $M=N$, Theorem 1 implies the symmetry between forward and backward terms. Indeed, in this case the coefficients (2.7) have the following simple properties:

$$\alpha_k(\Delta x)=-\alpha_{-k}(\Delta x)=\alpha_{-k}(-\Delta x). \tag{2.8}$$

Moreover, for a fixed k,

$$\lim_{N\to\infty}\alpha_k(\Delta x)=(-1)^{k+1}\,\frac{1}{(\Delta x)k},\quad k\neq 0. \tag{2.9}$$

The meaning of (2.9) is explained in Appendix B. Moreover, Eq. (2.8) implies the symmetry of the discrete derivative:

$$D_{\Delta x}=+D_{-\Delta x}. \tag{2.10}$$

The operator $D_{\Delta x}$ satisfying (2.10) we shall call **symmetric** (in Δx). Later on we shall limit ourselves to this case as it preserves the parity symmetry. Discretization schemes and corresponding coefficients satisfying, in addition to (2.4), (2.5) and symmetric we shall call **optimal**. In fact, in the special symmetric case $N=1$, (2.8) implies that the derivative is optimal and we obtain the well-known central difference scheme. The numerical values of the optimal α-coefficients for the lowest $(2N+1)$-point schemes are given in Table I.

In analogy to (2.2), the optimal derivative can be rewritten as

$$D_{\Delta x}f(x)\equiv 2\sum_{k=1}^{N}\alpha_k \sinh\left(k\Delta x\,\frac{d}{dx}\right)f(x). \tag{2.11}$$

In particular, for the optimal three-point (=central difference) discretization scheme we have

$$D_{\Delta x}f(x)=\frac{1}{\Delta x}\sinh\left(\Delta x\,\frac{d}{dx}\right)f(x)=\frac{1}{i\Delta x}\sin\left(i\Delta x\,\frac{d}{dx}\right)f(x). \tag{2.12}$$

The representation (2.12) is exactly the derivative represented by the Heine symbol (1.7).[21,22]

Originally, Heisenberg introduced his matrix mechanics by postulating that mathematical operations in classical equations of motion should be reinterpreted (that led to the correspondence principle). In this spirit Dirac postulated that the ''quantum differentiation'' must satisfy the additivity and the Leibniz rule. From this he obtained the fundamental (Heisenberg) commutation relations for coordinates and momenta.[25] We would like to proceed in an analogous way. However, in our case not all general rules for derivatives are satisfied by (1.3), subject to (2.4). The basic properties of the ordinary derivative are listed below along with some comments on how those properties change if one uses $D_{\Delta x}$ instead of d/dx:

(1) Additivity. Ordinary and discrete derivative are both linear:

$$D_{\Delta x}[f(x)+g(x)]=D_{\Delta x}f(x)+D_{\Delta x}g(x).$$

(2) Leibniz rule. The rule for differentiation of the product is not satisfied exactly by (1.3). The deviation is of order $O(\Delta x)$ and for the optimal scheme of order $O(\Delta x^{2N})$.[26]

(3) Derivative of the composite function. The same result as for the Leibniz rule holds for the derivative of a composite function: $D_{\Delta x}[f\circ g(x)]=D_{\Delta x}f(g(x))\times D_{\Delta x}g(x)+O(\Delta x^{2N})$.

(4) Derivative of monomials. Discrete derivative of the monomial x^n is identical to the ordinary derivative if the discretization scheme is optimal and the exponent $n\leqslant 2N$. In general, assuming (2.4) and (2.8), we have

$$D_{\Delta x}(N)x^n=nx^{n-1}+\sum_k\sum_{i=2}^{n}\alpha_k k^i\binom{n}{k}\Delta x^i x^{n-i}. \tag{2.13}$$

(5) Heisenberg commutation relation. This will be the subject of the following sections. We shall present a construction of the discrete counterpart of the Schrödinger couple x and d/dx satisfying a discrete analog of the continuous Heisenberg commutation relation: $[d/dx, x]f(x)=f(x)$. Naively, that is, by replacing only d/dx with $D_{\Delta x}$ and using (1.3), we get

$$[D_{\Delta x}, x]f(x)=\Delta x\sum_k k\,\alpha_k\, f(x+k\Delta x). \tag{2.14}$$

For the optimal discretization we will construct a position operator called X which together with $D_{\Delta x}$ forms a conjugate pair, that is the relation $[D_{\Delta x}, X]=I$ holds on a dense domain in a Hilbert space.

(6) Hermiticity. In the symmetric case we have a natural Hermitian conjugation induced from the relation $(E_{\Delta x}^n)^\dagger=E_{\Delta x}^{-n}$. The operator $D_{\Delta x}$, like the operator d/dx, is formally anti-Hermitian if $M=N$ and $\alpha_k(\Delta x)=-\alpha_{-k}(\Delta x)$ and we have

$$\int (D_{\Delta x}f(x))g(x)\,dx=-\int f(x)(D_{\Delta x}g(x))\,dx. \tag{2.15}$$

A more detailed analysis will be given in the subsequent sections.

III. LIE ALGEBRAIC DISCRETIZATION

The notion of the *Lie algebraic discretization* has been introduced in Ref. 27. It can be described as a special case of representation theory of the enveloping algebra of the Heisenberg algebra in which at least one of the generators is a difference operator. As a result, some linear difference equations have a representation theoretic meaning similar to the D-module interpretation of linear differential equations [see Ref. 28, in particular Sec. VI]. The most important application seems to be to describe quasi-polynomial solutions to a certain class of difference equations. On the other hand the whole procedure has certain appeal of *déjà vu* in the context of basic quantum mechanics and from this point of view one can regard the authors' concept of the Lie algebraic discretization as a quantization procedure.

Following Ref. 27, we start with (1.1) and we look for a conjugate operator $X\equiv X_{\Delta x}$ that satisfies

$$[D_{\Delta x}, X]=1, \tag{3.1}$$

with the additional condition ('classical limit')

$$\lim_{\Delta x\to 0} X_{\Delta x}=x. \tag{3.2}$$

In Ref. 27, only the special cases of the forward and backward *Euler schemes* that are not symmetric under reflections (2.10) have been discussed. This can be improved in two ways. First, we can start with the more general discretization scheme (1.3). Second, we can require that $D_{\Delta x}$ be formally skew-Hermitian whereas $X_{\Delta x}$ be formally Hermitian relative to some $\star$ anti-involution to ensure a natural physical interpretation.

In Ref. 27 it has been shown that for derivatives in the form

$$D^{+}_{\Delta x}f(x)\equiv\frac{f(x+\Delta x)-f(x)}{\Delta x},\quad D^{-}_{\Delta x}f(x)\equiv\frac{f(x)-f(x-\Delta x)}{\Delta x}, \tag{3.3}$$

we get the Heisenberg algebra for D^{+} and $x(1-\Delta x D^{-})$:

$$[D^{+}, x(1-\Delta x D^{-})]=1. \tag{3.4}$$

However, as has been mentioned in Sec. II, we are not satisfied with the Euler forms of the discrete derivative (3.3), as these schemes are not optimal and symmetric with respect to reflections.[29] In particular, this form of the discrete derivative does not seem to be skew-Hermitian with respect to any natural Hermitian structure. Also, the central difference scheme (and optimal, in general) was shown to enable a consistent formulation of the action principle and the Euler–Lagrange equations in the classical theory.[30] Another possibility would be that the pair $D_{\Delta x}, X$ is a pair of conjugate operators (annihilation and creation operators). But here again, no natural Hermitian structure on the Hilbert space of states seems to induce that.

First, we take up a question of generalizing the results of Ref. 27 to a wider class of discrete operators. As the starting point we generalize the form of the coordinate operator used in Ref. 27, stipulating the following Ansatz:

$$X\equiv\sum_{k=-N}^{M}\beta_k\, x\, E^{k}, \tag{3.5}$$

subject to the condition (continuous limit)

$$\sum_{k=-N}^{M} \beta_k = 1. \tag{3.6}$$

In fact, one can add to (3.5) two additional terms without changing the conclusions. Thus X might be taken to be of the form

$$X \equiv \sum_{k=-N}^{M} \beta_k \, x \, E^k + \sum_{k=-N}^{M} \beta_k' E^k x + \sum_{k=-N}^{M} \beta_k'' E^k. \tag{3.7}$$

The last term commutes with the derivative $D_{\Delta x}$ and does not change commutation relation (3.1). It changes, however, condition (3.6), replacing it with

$$\sum_{k=-N}^{M} (\beta_k + \beta_k') = 1, \quad \sum_{k=-N}^{M} \beta_k'' = 0. \tag{3.8}$$

The second term on the right-hand side of (3.7), due to (1.5), has the same commutation relation as (3.5), leading to the substitution

$$\beta_k \to \beta_k + \beta_k' . \tag{3.9}$$

From now on, we restrict our attention to (3.5). From (3.1) we get the following conditions for the β's:

$$\begin{aligned} &\sum_{k=-N}^{M} \beta_k = 1, \\ &\Delta x \sum_{\substack{k=-M \\ -N \leqslant n-k \leqslant M}}^{N} \alpha_k \, \beta_{n-k} \, k = \delta_{n0}, \end{aligned} \tag{3.10}$$

where n goes from $-(M+N)$ to $+(M+N)$. Here, (3.10) is a set of $(2M+2N+2)$ equations for $(2M+2N+2)$ coefficients α_k, β_k plus two additional equations (2.4). Hence, the system is formally overdetermined. Let us start with the second equation above (set of $2N+2M+1$ equations) and rewrite it as two separate sets, each of size $M+N+1$, corresponding to $n \leqslant 0$ and $0 \leqslant n$, respectively. For simplicity we introduce matrix notation in the $(M+N+1)$-dimensional space of coefficients. Let us define vectors

$$\beta = \begin{bmatrix} \beta_{-N} \\ \beta_{-N+1} \\ \vdots \\ \beta_M \end{bmatrix}, \quad b = \begin{bmatrix} 1/\Delta x \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3.11}$$

and two $(N+M+1) \times (N+M+1)$ matrices:

$$A_-=\begin{bmatrix} N\alpha_N & \dots & \dots & -M\alpha_{-M} \\ (N-1)\alpha_{N-1} & \dots & -M\alpha_{-M} & 0 \\ \dots & \dots & \dots & \dots \\ -M\alpha_{-M} & 0 & \dots & 0 \end{bmatrix},$$
$$A_+=\begin{bmatrix} N\alpha_N & \dots & \dots & -M\alpha_{-M} \\ 0 & N\alpha_N & \dots & (-M+1)\alpha_{-M+1} \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & N\alpha_N \end{bmatrix}. \tag{3.12}$$

Now, our equations can be rewritten as two sets of equations for the coefficients β_k in the following simple form:

$$A_-\ \beta=b, \quad A_+\ \beta=b, \tag{3.13}$$

where the first equations in both sets are identical. The determinants of the systems (3.13) are

$$\det\|A_-\|=(-M\alpha_{-M})^{M+N+1}, \quad \det\|A_+\|=(N^{M+N+1}\alpha_N)^{M+N+1}, \tag{3.14}$$

and (3.13) can be solved explicitly. Indeed, there are three possibilities to consider: (1) Both determinants are nonzero (i.e., $M\cdot N\neq 0$); (2) one of the determinants is zero, then either $M=0$ or $N=0$; or (3) both determinants are zero ($M=N=0$). In the last case the derivative expansion (1.3) consists of one nonzero term ($\alpha_0 E_0$) only and the second condition of (2.4) cannot be fulfilled. In the first case, using Cramer's rule, we get two different solutions to (3.13), hence the system is inconsistent. The only nontrivial possibility is to have the second case. Let us assume that $\alpha_N\neq 0$. Then from the second equation of (3.13) we get the following (unique) solution for β's:

$$\beta_{-N}=\frac{1}{N\alpha_N\Delta x}, \quad \beta_{-N+1}=\cdots=\beta_M=0, \tag{3.15}$$

and the first equation of the set (3.10) implies: $\beta_{-N}=1$ and $\alpha_N=1/(N\Delta x)$. Now, consistency of the first equation of (3.13) implies that the only nonzero α's can be α_0 and α_N. Finally, Eqs. (2.4) are satisfied as well if $\alpha_0=-\alpha_N$. An identical solution can be found for $\alpha_{-M}\neq 0$. Hence, we have the following.

Theorem 2: *For the discretized derivative operator (1.3) satisfying (2.4) and the coordinate operator in the form (3.5) with condition (3.2) to satisfy Heisenberg commutation relation (3.1) the coefficients α_k, β_k must be either in the forward Euler form:*

$$\alpha_k\equiv\alpha_k^F=-\frac{1}{N\Delta x}\ \delta_{k0}+\frac{1}{N\Delta x}\ \delta_{N,k}, \quad \beta_k\equiv\beta_k^F=+\delta_{-N,k}, \tag{3.16}$$

or in the backward Euler form:

$$\alpha_k\equiv\alpha_k^B=\frac{1}{M\Delta x}\ \delta_{k0}-\frac{1}{M\Delta x}\ \delta_{-M,k}, \quad \beta_k\equiv\beta_k^B=+\delta_{-M,k}. \tag{3.17}$$

Thus, for X of the form (3.5) the only way to satisfy (3.1) is to take $D_{\Delta x}$ as the Euler derivatives (3.3). In this case the coordinate operators are

$$X^{B,F}=x(1-\Delta x D_{N\Delta x}^{\mp})=xE_{\Delta x}^{\pm N}. \tag{3.18}$$

This proves that the Turbiner–Smirnov discretization scheme is unique under conditions (3.1), (3.2), and (3.5). Hence, an optimal derivative with the discretized coordinate operator of the form (3.5) cannot satisfy the Heisenberg commutation relation. In fact, this is impossible for any symmetric discretization scheme with finite series in (3.5).

IV. TRANSCENDENTAL OPERATORS X

In the previous section we have limited our considerations to the coordinate operators (3.5) in the form of a finite series. However, there is still a possibility of having the Heisenberg algebra with the discrete derivative of the form (1.3) and an infinite series for X. We shall investigate the algebraic aspects of this problem in this section and a complete, rigorous solution of the problem will be given in Sec. VI. We should mention that the operator X is always unbounded and eventually special care needs to be exercised when dealing with any expressions involving X. In this section we study only formal aspects of our problem, leaving aside important questions like the domain of the definition of X or in what sense we understand expressions below involving infinite sums. We will come back to those questions in Secs. V and VI.

In the present section we postulate that our coordinate operator has the following form:

$$X=x\sum_{k=-\infty}^{\infty}\beta_m E_{\Delta x}^m, \tag{4.1}$$

and we have an infinite set of linear algebraic equations (3.10) for $n=0,\pm 1,\pm 2,\pm 3,\ldots$. In addition, the continuous limit condition (3.6) is tentatively taken to mean

$$\sum_{k=-\infty}^{+\infty}\beta_k=1. \tag{4.2}$$

Eventually, we will have to regularize this condition since in many examples a direct evaluation of this limit does not make sense.

We define infinite-dimensional matrices and vectors analogous to (3.11) and (3.12):

$$\beta_+=\begin{bmatrix}\beta_{-N}\\ \beta_{-N+1}\\ \beta_{-N+2}\\ \vdots\end{bmatrix}\quad \beta_-=\begin{bmatrix}\beta_M\\ \beta_{M-1}\\ \beta_{M-2}\\ \vdots\end{bmatrix},\quad b=\begin{bmatrix}1/\Delta x\\ 0\\ 0\\ \vdots\end{bmatrix}, \tag{4.3}$$

$$A_+=\begin{bmatrix}N\alpha_N & \cdots & -M\alpha_{-M} & 0 & \cdots & \cdots\\ 0 & N\alpha_N & \cdots & -M\alpha_{-M} & 0 & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\end{bmatrix},$$
$$A_-=\begin{bmatrix}-M\alpha_{-M} & \cdots & N\alpha_N & 0 & \cdots & \cdots\\ 0 & -M\alpha_{-M} & \cdots & N\alpha_N & 0 & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\end{bmatrix}. \tag{4.4}$$

Formally, we get, instead of (3.13), the following two sets of equations:

$$A_+\beta_+=b,\quad A_-\beta_-=b. \tag{4.5}$$

In both sets the first equation is identical and relates at most $(M+N+1)$ coefficients β_m. Each next equation involves one more (higher-order) coefficient that can be explicitly computed.

Hence, our system has $(M+N)$ free parameters. In the special case, when $M\cdot N=0$, the first or last element in each row in (4.4) vanishes and we have $(M+N-1)$ free parameters. This implies for the kernel of operators A_+, A_-

$$D\equiv \dim \ker A_+ = \dim \ker A_- = M+N-\delta_{MN,0}. \tag{4.6}$$

Hence, we have proven the following.

Theorem 3: *For operator X in the form of infinite series (4.1) there is an* $(M+N-\delta_{MN,0})$*-parameter family of formal solutions to the Heisenberg relation (3.1).*

Remark: We do not claim at this point that any of the formal solutions satisfy (4.2).

In the case $M\cdot N\neq 0$, Eqs. (4.5) imply one algebraic condition for the coefficients $\{\beta_{-N},\ldots,\beta_M\}$:

$$\sum_{k=-M}^{+N} k\ \alpha_k\ \beta_{-k}=\frac{1}{\Delta x}, \tag{4.7}$$

and the higher coefficients β_k $(k>M$ or $k<-N)$ can be uniquely determined from

$$\beta_{M+i}=\frac{-1}{M\alpha_{-M}}\sum_{k=-M+1}^{+N} k\ \alpha_k\ \beta_{i-k},$$

$$\beta_{-(N+i)}=\frac{-1}{N\alpha_N}\sum_{k=-M}^{N-1} k\ \alpha_k\ \beta_{-(i+k)}, \tag{4.8}$$

where $i=1,2,\ldots$ with additional condition (4.2).

In the remainder of this section we shall assume that the coefficients α_k are optimal and the coefficients β_k have the symmetry

$$\beta_k=\beta_{-k}, \quad -N\leqslant k\leqslant N. \tag{4.9}$$

This implies the symmetry of the discrete coordinate operator with respect to the Δx reflections:

$$X_{\Delta x}=X_{-\Delta x}, \tag{4.10}$$

in analogy to the discrete derivative operator (2.10).

Indeed, now we have $N+1$ β-parameters $\{\beta_0,\beta_1,\beta_2,\ldots,\beta_N\}$ subject to

$$\sum_{k=1}^{N} k\ \alpha_k\ \beta_k=\frac{1}{2\Delta x}. \tag{4.11}$$

The higher-order coefficients (β_k for $|k|>N$) are uniquely determined from

$$\beta_{N+i}=\frac{-1}{N\alpha_N}\sum_{k=1}^{N} k\ \alpha_k\ \beta_{i-k}-\frac{1}{N\alpha_N}\sum_{k=1}^{N-1} k\ \alpha_k\ \beta_{i+k}, \quad i=1,2,\ldots. \tag{4.12}$$

Using induction on i and (4.7), we obtain the claim

$$\beta_{-(N+i)}=\beta_{N+i}. \tag{4.13}$$

We remark that the normalization condition (4.2) can be rewritten as

$$2\sum_{k=1}^{+\infty}\beta_k = 1-\beta_0, \tag{4.14}$$

i.e., we are left with $(N-1)$ free parameters β. All considerations in this section were rather formal as we deal with infinite series whose convergence is unclear. In fact, as we shall see below the series might diverge in the usual sense. We discuss in detail a concrete example to illustrate this point.

In the case $N=1$, (4.12) implies the following condition for the β-coefficients:

$$\beta_{i+2}=\beta_i, \quad i=0,1,2,\ldots . \tag{4.15}$$

Hence, we have two free parameters: β_0 and β_1. The latter can be determined from (4.11), giving

$$\beta_1=+1. \tag{4.16}$$

This provides an exact formula for all odd coefficients in the following compact form:

$$\begin{aligned}\beta_{2k+1}&=(-1)^k, \quad k=0,1,2,\ldots,\\ \beta_{-(2k+1)}&=\beta_{2k+1},\end{aligned} \tag{4.17}$$

and for even coefficients we have

$$\beta_{2k}=(-1)^k\beta_0. \tag{4.18}$$

It is clear from (4.17) that our series is indeed divergent and we must redefine the classical limit to make sense of the summation in (4.1). There is no unique way of doing that. We will say more about this in the next section.

At the moment, we are still left with one free parameter, β_0, and one equation, the normalization condition (4.14). As a special case we can put

$$\beta_0=0. \tag{4.19}$$

Hence, a formal solution for the discretized coordinate operator, X, in the $N=1$ case is (4.1) with β-coefficients given by

$$\begin{aligned}\beta_{2k}&=0 \quad \forall k\\ \beta_{2k+1}&=(-1)^k, \quad k=0,1,2,\ldots,\\ \beta_{-(2k+1)}&=\beta_{2k+1}, \quad k=0,1,2,\ldots .\end{aligned} \tag{4.20}$$

Now, this solution can be compared with the solution found by Frappat and Sciarino.[31] Their solution, after restricting it to one dimension, reads

$$X=x\,\frac{1}{E+E^{-1}}+\frac{1}{E+E^{-1}}\,x. \tag{4.21}$$

There is a formal analogy between their solution and (4.20). To see that analogy one expands the first term in (4.21) at $E=0$ and the second one at $E^{-1}=0$. There is, however, an important difference between the approach of the present paper and that of Ref. 31. In their approach, the discretization parameter Δx is purely imaginary. As a consequence, the shift operator E is formally Hermitian and positive definite, while the momentum operator corresponding to our $D_{\Delta x}$ is unbounded. Moreover, since the shift is in the imaginary direction it is hard to see what this has

to do with a discrete space–time. By contrast, we are treating the shift as a unitary operator, thus retaining the real discretization parameter, which, in turn, ensures that the momentum is bounded.

V. MOMENTUM FORMULATION

In this section we reformulate some of the results obtained earlier as well as clarify some of the points left obscure in our discussion of Eq. (4.21).

The basic idea is to realize the algebra of shift operators as an algebra of multiplication operators on the Hilbert space of square integrable functions on the circle. One can also view this realization as the momentum space counterpart of the content of the preceding sections. Throughout this section $\Delta x=1$.

At the core of the idea of discretization is that in all operations one proceeds by performing a sequence of discrete steps of size, in this section, 1. This calls for bringing the Fourier transform into the picture. To this end we will first reinterpret (1.5). Our setup is as follows. We consider the unit circle,

$$S^1=\{z\in\mathbb{C}:\ |z|=1\}, \tag{5.1}$$

and the Hilbert space $H=L^2(S^1)$. We recall

$$L^2(S^1)=\left\{ f=\sum_{-\infty}^{\infty} a_n z^n\ :\ \sum_{-\infty}^{\infty} |a_n|^2<\infty \right\}. \tag{5.2}$$

We use the following parametrization of the circle S^1: $z=\exp i\theta$, $\theta\in\mathbb{R}$ (mod 2π). Thus a function $f\in L^2(S^1)$ can be written as

$$f(e^{i\theta})=\sum_{n\in\mathbb{Z}} a_n\, e^{in\theta}, \tag{5.3}$$

and a_n are the Fourier coefficients of f. Since the shift operator E defined in (1.4) corresponds to the elementary step of unit size we postulate that $E=T_z$, where T_z is the operator of multiplication by z. We will say more about multiplication operators towards the end of this section. Till further notice we also suppress writing T. Thus we set

$$E^n=z^n, \quad x=-z\,\frac{d}{dz}, \quad n\in\mathbb{Z}. \tag{5.4}$$

It is routine to check that (1.5) holds. Furthermore we define

$$D(z)\ :=\sum_{k=-M}^{N} \alpha_k\, z^k. \tag{5.5}$$

In other words the difference operator becomes now a multiplication by a Laurent polynomial in z. We might observe that the operator x above is one of the generators of the Virasoro algebra (with the trivial central term), called L_0. Now we reinterpret Eqs. (2.4) and (2.5). We see that (2.4) goes into

$$D(z=1)=0, \quad (L_0 D)(z=1)=-1. \tag{5.6}$$

Similarly, (2.5) goes into

$$(L_0^n D)(z=1)=0 \quad \text{for } n=2,3,\ldots,(M+N). \tag{5.7}$$

Now, it is immediate that the coordinate operator X should be sought [see (3.4)] in the form

$$X=-z\sum_{n=-\infty}^{\infty}\beta_n\frac{d}{dz}z^n, \tag{5.8}$$

where

$$\sum_{n=-\infty}^{\infty}\beta_n=1. \tag{5.9}$$

Let us now, at least formally, introduce the function $h(z)=\Sigma_{-\infty}^{\infty}\beta_n z^n$. From the above condition we get that $h(1)=1$. Moreover X can be written

$$X(z)=L_0h(z)=-zh(z)\frac{d}{dz}-zh'(z). \tag{5.10}$$

As was already explained earlier, we can in fact ignore the second part and look for X in the form

$$X(z)=-f(z)\frac{d}{dz},\quad f(1)=1. \tag{5.11}$$

Thus X is formally an element of the Virasoro algebra. We want that

$$[D(z), X(z)]=\left[f(z)\frac{d}{dz},D(z)\right]=1, \tag{5.12}$$

where $D(z)$ is a function (Laurent polynomial) defined previously. We therefore obtain that

$$f(z)D'(z)=1, \tag{5.13}$$

which formally implies

$$f(z)=\frac{1}{D'(z)}. \tag{5.14}$$

We note that the condition $f(1)=1$ is automatically satisfied as a consequence of the second condition imposed on D. However, it is important to underscore the fact that the latter expression can only be understood formally at this stage of our analysis. Yet, there are some special cases when we do not expect to have any difficulties. For example, if $D(z)=A+Bz$, that is $M=0$, $N=1$. Then we impose the following conditions:

$$D(1)=0,\quad D'(1)=1, \tag{5.15}$$

which gives $D(z)=z-1$ and, consequently, $f(z)=1$. Thus we get the pair $D(z)=z-1$, $X(z)=-d/dz$, which clearly satisfies (5.12). Using (5.10) we get that in this case $h(z)=1/z$. Hence $\beta_{-1}=1$ and $\beta_i=0$, $i\neq -1$. Going back to the difference operator realization we get that $D=E-I$, $X=xE^{-1}$, which is the pair found in Ref. 27 for $\Delta x=1$. We can contrast this example with that of the symmetric discretization scheme for which $D(z)=\frac{1}{2}(z-z^{-1})$. In that case $f(z)=2z^2/(1+z^2)$, which is not bounded on the unit circle, and, consequently, $f(z)$ is not a multiplier [that is the multiplication by f is not a bounded operator on $L^2(S^1)$]. This example is explained in detail in the next section.

However, there are many cases for which f is a multiplication operator on the circle. We present now a general construction of X for which f is a multiplier. One can view this construction as a generalization of Ref. 27.

First, we review a few elementary facts from the theory of doubly infinite Toeplitz operators.[32] Initially, we can consider a vector space of sequences $a=\{a_n\}_{n\in\mathbb{Z}}$ such that $\Sigma_{n\in\mathbb{Z}}|a_n|^2<\infty$. This space, called ℓ^2, is isomorphic to $L^2(S^1)$, the fact well known from the theory of Fourier series. On ℓ^2 we introduce a linear operator T_c,

$$a_m\to\sum_{n\in\mathbb{Z}} c_{m-n}a_n, \tag{5.16}$$

where $c=\{c_n\}_{n\in\mathbb{Z}}$ is a sequence of numbers. It is known that T_c is a bounded operator on ℓ^2 iff c is a sequence of the Fourier coefficients of an essentially bounded function on S^1. We recall that $\phi(\theta)$ is an essentially bounded function on S^1 if it is bounded almost everywhere there. The isomorphism between ℓ^2 and L^2 allows one to study the Toeplitz operator T_c as acting on $L^2(S^1)$. We will say that $\phi(\theta)$ is a symbol of T_c if under this isomorphism T_c acts on $L^2(S^1)$ as

$$(T_c f\,)(\theta)=\phi(\theta)f(\theta),\quad f\in L^2(S^1). \tag{5.17}$$

Because of this relation we often write T_ϕ to denote the Toeplitz operator with the symbol ϕ. We will need in this section one more result from the theory of Toeplitz operators. The result below gives a complete description, in terms of symbols, of invertible Toeplitz operators.

Lemma 1: T_ϕ is invertible if and only if $1/\phi$ is essentially bounded. If this holds, $T_\phi^{-1}=T_{1/\phi}$.

Now we apply this lemma to our problem. Let us first consider a Laurent polynomial $g(z)$ such that $\mathrm{Res}_{z=0}\, g=0$ and g is nowhere zero on S^1. We can normalize such a g to satisfy $g(1)=1$. Now we define

$$D(e^{i\theta}):=i\int_0^\theta e^{it}g(e^{it})dt, \tag{5.18}$$

where we used $dz=ie^{i\theta}d\theta$. Observe that since there is no z^{-1} in g, $D(z)$ is again a Laurent polynomial satisfying (5.6). Now we set $f(z)=1/g(z)$. We see that $f(z)$ is bounded on S^1, so, by the lemma above, T_f is then the bounded inverse of T_g. Now we define the operator $X(z)=-T_f\partial/\partial z$: $C^\infty(S^1)\to C^\infty(S^1)$. By exactly the same computation as the one leading up to (5.5) we obtain that on $C^\infty\subset L^2(S^1)$, $X(z)$ and $D(z)$ so defined satisfy the Heisenberg commutation relation as well as (5.6).

Our original setup has been formulated in terms of the ℓ^2 space. Now we would like to return to it. The map which maps back $L^2(S^1)$ to ℓ^2 is given by the Fourier series method: $f\to\{\hat{f}(n)\}_{n\in z}$, where $\hat{f}(n)$ is the nth Fourier coefficient of f. We recall that using (5.10) we can describe the coefficients β_n as

$$\beta_n=\hat{f}(n+1),\quad n\in\mathbb{Z}. \tag{5.19}$$

Example: $g(z)=a+bz^{-2}$, $|a/b|\neq 1$. The condition $g(1)=1$ gives $a+b=1$. Thus $D(z)=az+(1-2a)+(a-1)z^{-1}$, $|a/(1-a)|\neq 1$. We have two cases to consider depending on whether $|a/(1-a)|<1$ (case 1) or $|a/(1-a)|>1$ (case 2).

Case 1: $f(z)=z^2\{1/[az^2+(1-a)]\}$.

We observe that f is analytic inside the unit circle S^1 and thus its Fourier expansion coincides with its Taylor expansion around $z=0$. We obtain

$$\hat{f}(n)=\begin{cases}\dfrac{(-1)^{(n-2)/2}}{1-a}\left(\dfrac{a}{1-a}\right)^{(n-2)/2}, & \text{if } n\geqslant 2 \text{ and } n \text{ is even;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.20}$$

Thus

$$\beta_n=\begin{cases}\dfrac{(-1)^{(n-1)/2}}{1-a}\left(\dfrac{a}{1-a}\right)^{(n-1)/2}, & \text{if } n\geqslant 1 \text{ and } n \text{ is odd;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.21}$$

Case 2: $f(z)=\{1/[a+(1-a)z^{-2}]\}$, $|a/(1-a)|>1$.

We observe that f is analytic outside of the unit circle S^1 and thus its Fourier expansion coincides with its Taylor expansion around $z=\infty$. We obtain

$$\hat{f}(n)=\begin{cases}\dfrac{(-1)^{n/2}}{a}\left(\dfrac{1-a}{a}\right)^{(-n)/2}, & \text{if } n\leqslant 0 \text{ and } n \text{ is even;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.22}$$

Consequently,

$$\beta_n=\begin{cases}\dfrac{(-1)^{(n+1)/2}}{a}\left(\dfrac{1-a}{a}\right)^{(-n-1)/2}, & \text{if } n\leqslant -1 \text{ and } n \text{ is odd;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.23}$$

This example illustrates very well the fact that the optimal discretization is special even from the point of view of the Heisenberg commutation relations. Indeed the case excluded from the example is $a=\frac{1}{2}$ which is precisely the optimal case. We can get, however, some information about this case making a approach $\frac{1}{2}$. This allows us to discuss the formal solution (4.20). One way of interpreting this solution relies on a regularization of D which essentially amounts to moving the zeros of D' off the unit circle. We now present a simple example of such a regularization. First we set $a=\frac{1}{2}-\epsilon$, $\epsilon>0$. We observe that $|a/(1-a)|<1$, so, we are dealing with the first case above. Directly from (5.21) we obtain

$$\beta_n(\epsilon)=\begin{cases}\dfrac{(-1)^{(n-1)/2}}{1/2-\epsilon}\left(\dfrac{1/2-\epsilon}{1/2+\epsilon}\right)^{(n-1)/2}, & \text{if } n\geqslant 1 \text{ and } n \text{ is odd;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.24}$$

It is interesting to note that

$$\beta_n(0)=\begin{cases}2(-1)^{(n-1)/2}, & \text{if } n\geqslant 1 \text{ and } n \text{ is odd;}\\ 0, \quad \text{otherwise.}\end{cases} \tag{5.25}$$

This is a formal solution to (4.7) and (4.8), with no symmetry condition imposed on β's. It has exactly the same status as the symmetric solution (4.20). Thus we see that the formal solutions of Sec. V can be justified through an appropriate regularization. In addition we have to properly interpret (4.2) (classical limit). It makes sense to consider this expression for $\epsilon\neq 0$. Then, as it is easy to check,

$$\sum_{k=-\infty}^{\infty}\beta_k(\epsilon)=1. \tag{5.26}$$

The same expression does not make sense for $\epsilon=0$, clearly indicating that one cannot interchange $\lim_{\epsilon-0}$ with Σ.

The final item we would like to discuss in this section is the question of Hermiticity or rather lack thereof. In our formulation the pair D and X may not necessarily consist of (even formally) Hermitian or adjoint elements. This is plain, for example, if one looks at the pair $D=E-1$ and $X=xE^{-1}$ or even more general pairs found above. One could, however, make this pair adjoint to each other, thus turning D and X into a pair consisting of a creation and an annihilation operator, by constructing a proper representation space. One possibility would be to consider a Fock representation, which carries an essentially unique Hermitian form. Thus, for the last example, the completion of the space span $\{e_n=X^n 1,\ n\geqslant 0\}$ with respect to the inner product $\langle e_m,e_n\rangle = \delta_{m,n},n!$ gives us a Fock space on which D acts as an annihilation operator and X acts as a creation operator, respectively, and both act as difference operators. Another possibility is to consider a coherent representation, i.e., such that is generated from the vector v: $Dv=\lambda v$, $\lambda \in \mathbb{C}$, whose special case is the Fock representation obtained for $\lambda=0$. In the case of $D=E-1$ the vacuum state v is a quasiperiodic function satisfying $v(x+1)=(\lambda+1)v(x)$. One obtains the representation space as the span $\{\epsilon_n=X^n v,\ n\geqslant 0\}$ and the unique Hermitian form is defined by $\langle v,v\rangle=1$ and $\langle e_m,e_n\rangle=\langle v,D^m X^n v\rangle$, where in the last expression one moves X to the left and D to the right making use of (5.12), the formula $\langle u,X^n w\rangle=\langle D^n u,w\rangle$, and $Dv=\lambda v$.

In the next section we study the optimal discretization for which we show that one can construct pairs of Hermitian (self-adjoint) operators D and X.

VI. MOMENTUM FORMULATION APPROACH TO OPTIMAL DISCRETIZATION

Now we take on the case of the optimal discretization. Thus $M=N$ throughout this section. Also, in this section we are interested in self-adjoint pairs X, P relative to the standard inner product on $L^2(S^1)$. For convenience we multiply D from previous sections by $1/i$ and denote the resulting function by D_N.

The problem of classifying the pairs of operators satisfying Heisenberg commutation relations under an additional assumption that one of the operators be bounded was considered in Ref. 33. Since our difference operators are bounded, we are studying a special case of that classification, included in Ref. 33, in particular in Theorem 8.5 therein. The approach below is direct, however, and can be used as an introduction to a more encompassing treatment of Ref. 33 (see also Refs. 34 and 35). Moreover, we have a different physical motivation. Our interest lies in concrete bounded operators explicitly given by difference operators, reflecting the underlying assumption that the coordinate space is discrete. We recall some basic definitions from Ref. 33. Let H denote a separable Hilbert space over $\mathbb{C}$.

Definition: Let P,Q be operators in H, and Ω a dense subspace of H. We call (P,Q) a conjugate pair on Ω iff

(K1) *P and Q are symmetric, $\Omega\subset D(P)\cap D(Q)$,*

(K2) *$P\Omega\subset\Omega$, $Q\Omega\subset\Omega$,*

(K3) *$P=\overline{P/\Omega}$, $Q=\overline{Q/\Omega}$,*

(K4) *Q is bounded,*

(K5) *P has a self-adjoint extension.*

In this section we essentially show that our X and D_N form a conjugate pair in the above sense.

We start with the following.

Lemma 2: Let $D_N(z)$ be optimal. Then $D_N'(z)$ has two simple roots on the circle S^1.

Proof: Since $z=e^{i\theta}$, $0\leqslant\theta<2\pi$, we want to show that $(d/d\theta)D_N\equiv D_N'(\theta)$ has exactly two zeros for $\theta\in[0.2\pi)$.

To derive $D_N'(\theta)$ we first compute $D_N(\theta)$. We find that

$$D_N(\theta)=2\sum_{k=1}^{N}\frac{(-1)^{k+1}}{k}\frac{(N!)^2}{(N+k)!(N-k)!}\sin k\theta.$$

Hence,

$$D_N'(\theta)=2\sum_{k=1}^{N}(-1)^{k+1}\frac{(N!)^2}{(N+k)!(N-k)!}\cos k\theta$$

or, in terms of z,

$$D_N'(\theta)=\sum_{\substack{k=-N\\k\neq 0}}^{N}\frac{(-1)^{k+1}(N!)^2}{(N+k)!(N-k)!}z^k=\frac{(-1)^{N+1}(N!)^2}{(2N)!}z^{-N}\sum_{\substack{k=-N\\k\neq 0}}^{N}(-1)^k\binom{2N}{N}z^k$$
$$=\frac{(-1)^{N+1}(N!)^2}{(2N)!}z^{-N}\left[(1-z)^{2N}-(-1)^N\binom{2N}{N}z^N\right].$$

Thus it suffices to show that the polynomial $(1-z)^{2N}-(-1)^N\binom{2N}{N}z^N$ has exactly two zeros for $z\in S^1$. Indeed, we obtain $[(1-z)^2/z]^N=(-1)^N\binom{2N}{N}$ or $\sin^2\theta/2=\frac{1}{4}\sqrt[N]{\binom{2N}{N}}$ This equation has two solutions, provided $\frac{1}{4}\sqrt[N]{\binom{2N}{N}}<1$ and $N\geqslant 1$.

Indeed, for $N=1$ we get $\frac{1}{4}\cdot 2<1$ and there are two roots, $\theta=\pi/2,\frac{3}{2}\pi$. Now, we proceed by induction. First we observe that the above inequality is equivalent to $\binom{2N}{N}<4^N$. We already proved it for $N=1$. We note that

$$\binom{2N+2}{N+1}=\binom{2N}{N}\frac{(2N+1)(2N+2)}{(N+1)^2}=\binom{2N}{N}\frac{2(2N+1)}{(N+1)},$$

and we use the induction hypothesis to prove the claim.

To prove that the zeros are simple, we show that $d/dz[(1-z)^{2N}-(-1)^N\binom{2N}{N}z^N]$ is not zero there. Using that $(1-z)^{2N}-(-1)^N\binom{2N}{N}z^N=0$, we obtain that this derivative can be zero if and only if $2z^2-2z-1=0$. Since this equation has two real zeros not equal to ± 1 we conclude that the zeros of $D_N'(\theta)$ are simple. ■

Now, we turn to the problem of the localization of those zeros. We will show that both zeros move towards $\theta=\pi$ as $N\to\infty$. First we prove three technical lemmas.

Lemma 3:

$$\frac{\left(1+1/2N\right)}{\left(1+1/N\right)}<\frac{1+\dfrac{1}{2(N+1)}}{1+1/(N+1)},\quad N\geqslant 1.$$

Proof: Consider

$$\left(1+\frac{1}{2N}\right)\left(1+\frac{1}{N+1}\right)=\frac{2N^2+5N+2}{2N(N+1)}\quad\text{and}\quad\left(1+\frac{1}{N}\right)\left(1+\frac{1}{2(N+1)}\right)=\frac{2N^2+5N+3}{2N(N+1)}.$$

Hence,

$$\left(1+\frac{1}{2N}\right)\left(1+\frac{1}{N+1}\right)<\left(1+\frac{1}{N}\right)\left(1+\frac{1}{2(N+1)}\right),$$

which implies the claim. ■

TABLE II. Roots of D'_N.

N	1	10	100	2500	10 000
root 1	$\frac{1}{2}\pi$	2.32	2.80	3.06	3.10
root 2	$\frac{3}{2}\pi$	3.96	3.48	3.27	3.19

The next lemma improves on the estimate used in the course of the proof of Lemma 2.
Lemma 4:

$$\binom{2N}{N}<4^N\left[\frac{(1+1/2N)}{(1+1/N)}\right]^N, \quad N\geqslant 1.$$

Proof: The proof goes by induction on N. The only nontrivial step is to observe that

$$\binom{2(N+1)}{N+1}=\binom{2N}{N}\frac{4(1+1/2N)}{(1+1/N)},$$

which, by the induction hypotheses, implies

$$\binom{2(N+1)}{N+1}<4^{N+1}\left[\frac{\left(1+1/2N\right)}{\left(1+1/N\right)}\right]^{N+1}.$$

Now, the claim follows from Lemma 3. ■

Finally, we have the following.
Lemma 5: The sequence $a_N=1/4\sqrt[N]{\binom{2N}{N}}$, $N=1,2,...,$ *is increasing and* $\lim_{N-\infty} a_N=1$.
Proof: By Lemma 4, $a_N<(1+1/2N)/(1+1/N)$. Hence.

$$4\binom{2N}{N}a_N<4\binom{2N}{N}\frac{(1+1/2N)}{(1+1/N)}.$$

The left-hand side equals $\binom{2N}{N}^{(N+1)/N}$, from which we get $\binom{2N}{N}^{(N+1)/N}<\binom{2(N+1)}{(N+1)}$, which implies that $a_N<a_{N+1}$. To compute $\lim_{N\to\infty} a_N$ we observe that $\lim_{N\to\infty}\exp(1/N)[\ln(2N)!-2\ln(N)!]$ can be computed using the Stirling formula:

$$\Gamma(x)=e^{-x}x^{x-1/2}(2\pi)^{1/2}\left[1+\frac{1}{12x}+\cdots\right],$$

valid for large x. Thus

$$\lim_{N\to\infty} a_N=\frac{1}{4}\lim_{N\to\infty}\exp\frac{1}{N}\left[\ln(2N+1)^{2N+1/2}-2\ln(N+1)^{N+1/2}\right]=\frac{1}{4}\lim_{N\to\infty} e^{2\ln((2N+1)(N+1))}=1,$$

thus completing the proof. ■

To find the roots of $D'_N(\theta)$ we solve $\sin^2(\theta_N/2)=a_N$. By Lemma 5 we see that θ_N's are approaching π as N increases. Table II illustrates this phenomenon numerically.

Let us denote by θ_N^+, and θ_N^- the zeros of $D'_N(\theta), \theta_N^{\pm}\in[0,2\pi]$. Then on the intervals $I_N^1=(\theta_N^- -2\pi,\theta_N^+)$, and $I_N^2=(\theta_N^+,\theta_N^-)$, $D_N(\theta)$ is monotone and thus $D_N(\theta)$ has an inverse which

we denote by ϕ_N^1 and ϕ_N^2, respectively. Now, we observe that the Hilbert space $H=L^2(S^1)$ $\simeq L^2([\theta_N^- - 2\pi, \theta_N^-])$ admits the direct sum decomposition: $H=L^2(I_N^1)\oplus L^2(I_N^2)$. Moreover, we have

$$L^2(I_N^i)\simeq L^2(D_N(\mathrm{I}_N)), \quad \tau_i:\ \psi(\theta)\to|(D_N'\circ\phi_N^i)(y)|^{-1/2}\psi(\phi_N^i(y)),$$

where $y=D_N(\theta)$ and $i=1,2$. Now we show that τ is a unitary isomorphism.

Lemma 6: τ_i is a unitary isomorphism, $i=1,2$.

Proof:

$$\langle\psi_1|\psi_2\rangle=\int_{I_N^i}\overline{\psi_1(\theta)}\psi_2(\theta)d\theta=\int_{D_N(I_N^i)}\overline{\psi_1(\phi_N^i(y))}\psi_2(\phi_N^i(y))\left|\frac{dy}{d\theta}\right|^{-1}dy.$$

Observe, however, that $|dy/d\theta|=|(D_N'\circ\phi_N^i)(y)|$. Hence,

$$\langle\psi_1|\psi_2\rangle_{L^2(I_N^i)}=\int_{D_N(I_N^i)}(\overline{\tau_i\psi_1})(y)(\tau_i\psi_2)(y)\ dy=\langle\tau_i\psi_1|\tau_i\psi_2\rangle_{L^2(D_N(I_N^i))},$$

which ends the proof. ■

We will need an explicit form of the inverse of τ_i. A simple computation yields

$$\tau_i^{-1}:\ \phi\to|D_N'(\theta)|^{1/2}\phi(D_N(\theta)), \quad \theta\in I_N^i, \quad \phi\in L^2(D_N(I_N^i)).$$

Now, we want to define the skew-Hermitian part of the operator $[1/D_N'(\theta)](d/d\theta)[1/D_N'(\theta)]$ on $L^2(D_N)$. It is easier, however, to consider first its push-forward under τ:

$$\begin{aligned}
&\frac{1}{2}\,\tau_i\circ\left[\frac{1}{D_N'(\theta)}\frac{d}{d\theta}+\frac{d}{d\theta}\frac{1}{D_N'(\theta)}\right]\circ\tau_i^{-1}\phi(y)\\
&\quad=\frac{1}{2}\,\tau_i\circ\left[\frac{1}{D_N'(\theta)}\frac{d}{d\theta}+\frac{d}{d\theta}\frac{1}{D_N'(\theta)}\right](\mathrm{sign}(D_N'(\theta))D_N'(\theta))^{1/2}\phi(D_N(\theta))\\
&\quad=\frac{1}{2}\,\tau_i\circ\left[\frac{1}{D_N'(\theta)}\,\phi(D_N(\theta))+2\frac{\mathrm{sign}\,D_N'(\theta)}{(\mathrm{sign}(D_N'(\theta))D_N'(\theta))^{1/2}}\,\partial_\theta\phi(D_N(\theta))\right]\\
&\quad=\tau_i\circ\frac{\mathrm{sign}\,D_N'(\theta)}{(\mathrm{sign}(D_N'(\theta))D_N'(\theta))^{1/2}}\,\partial_\theta\phi(D_N(\theta))\\
&\quad=|D_N'\circ\varphi_N^i(y)|^{-1/2}\frac{(\mathrm{sign}(D_N'(\theta))\circ\varphi_N^i)(y)}{|D_N'(\theta)\circ\varphi_N^i(y)|^{1/2}}\frac{\partial}{\partial\varphi_N^i(y)}\,\phi(y)=\frac{1}{D_N'\circ\varphi_N^i(y)}\frac{\partial}{\partial\varphi_N^i(y)}\,\phi(y).
\end{aligned}$$

Since $\partial/\partial y=[\partial\varphi_N^i(y)/\partial y][\partial/\partial\varphi_N^i(y)]$ we need to compute $\partial\varphi_N^i(y)/\partial y$. To this end we simply observe that $(\varphi_N^i\circ D_N)(\theta)=\theta$. Hence,

$$\frac{d\varphi_N^i}{dy}=\frac{1}{D_N'(\theta)}=\frac{1}{(D_N'\circ\varphi_N^i)(y)}.$$

This implies

$$\left(\tau_1\circ\frac{1}{2}\left[\frac{1}{D_N'(\theta)}\frac{d}{d\theta}+\frac{d}{d\theta}\frac{1}{D_N'(\theta)}\right]\circ\tau_1^{-1}\right)(y)=\frac{\partial}{\partial y}, \quad y\in D_N(I_N^i).$$

We thus have the following important theorem.

Theorem 4: *Let*

$$X_N=\frac{i}{2}\left[\frac{1}{D_N'(\theta)}\frac{d}{d\theta}+\frac{d}{d\theta}\frac{1}{D_N'(\theta)}\right]$$

be defined on a dense set $S\in L^2(S^1)$. Then X_N is unitary equivalent to $i\partial/\partial y|_{L^2(D_N(I_N^1))} \oplus i\partial/\partial y|_{L^2(D_N(I_N^2))}$, *defined on the dense space* $\tau_1(S\cap L^2(I_N^1))+\tau_2(S\cap L^2(I_N^2))$.

We can prove a similar statement for $D_N(\theta)$. We have: $\tau_i\circ D_N(\theta)\circ\tau_i^{-1}\phi(y)=\tau_1 \circ D_N(\theta)|D_N'(\theta)|^{1/2}\phi(D_N(\theta))=|D_N'\circ\varphi^i(y)|^{-1/2}D_N\circ\varphi^i(y)|D_N'\circ\varphi^i(y)|^{1/2}\phi(D_N\circ\varphi^i)(y)=y\phi(y), \quad y \in D_N(I_N^i)$. Hence we have the following.

Theorem 5: *Let* P_N *be the operator of multiplication by the function* D_N. *Then* P_N *is unitary equivalent to* $y|_{L^2(D_N(I_N^1))}\oplus y|_{L^2(D_N(I_N^2))}$.

Corollary: The pair $Q=D_N$ and $P=-X_N$ is a conjugate pair on Ω, where $\Omega=\tau^{-1}(S_1) \oplus\tau^{-1}(S_2)$, $S_i=\{f\in L^2(D_N(I_N^i))$, f is absolutely continuous, $f=0$ on the boundary of $D_N\}$, $i=1,2$.

Remark: Observe that X_N is closed and symmetric on Ω. We discuss its self-adjoint extensions below.

First, however, we determine the spectrum of D_N or to be more precise T_{D_N}. By the well-known theorem on Toeplitz operators,[32] Spec $(T_{D_N})=$ essential range of $D_N=$ range of D_N, where the latter is a consequence of continuity. Because X_N decomposes into a direct sum of operators of the type id/dy on the finite intervals $D_N(I_N^i)$ we present a brief description of basic features of such operators. For convenience we take a finite interval $[-1,1]$. The following discussion is based on Ref. 36 and all the details can be found there. The operator id/dy is known to have a one-parameter family of self-adjoint extensions. More precisely, let us set $id/dy=T_\alpha$, defined on $D(T_\alpha)=\{\varphi:\varphi$ is absolutely continuous and $\varphi(-1)=\alpha\varphi(1)\}$, where $\alpha\in\mathbb{C}$, $|\alpha|=1$.

To interpret α we solve the eigenvalue problem: $i(d/dy)\psi=\lambda\psi$ with $\psi(-1)=\alpha\psi(1) =e^{i\varphi_0}\psi(1)$, $\alpha=e^{i\varphi_0}$, $\varphi_0\in[0,2\pi)$. Thus we obtain; $\psi(y)=e^{i\lambda y}$ and $e^{-i\lambda}=e^{i\varphi_0}e^{i\lambda}$ from which it follows that $2\lambda+\varphi_0=2\pi n$, $n\in\mathbb{Z}$. Hence, $\lambda=\lambda_n=n\pi-\varphi_0/2$, $n\in Z$. The appearance of π is accidental and can easily be removed by rescaling T_α. The spectrum of T_α is a lattice of physical positions. We would like to point out that there is no canonical way of identifying this lattice with the lattice we have started with. It seems to be compelling in fact to consider as a real physical space the lattice of eigenvalues of T_α and the corresponding Hilbert space $L^2(D_N(I_N^1))$, for example. Thus we would, so to speak, avoid the ''spectrum doubling'' problem by decree. On the other hand, this would have to be interpreted quantum mechanically as saying that the process of quantization amounts in this case to finding an irreducible representation of the Heisenberg commutation relations subject to the condition that the spectrum of the position operator be discrete. This representation would then have a relation to the calculus of difference operators as described above, yet the physical Hilbert space would be only ''half'' of the Hilbert space that naturally carries the action of difference operators.

To close this section we would like to mention that the optimal discretization has a very nice property that $\lim_{N\to\infty} D_N(\theta)=\theta$. This fact is proven in Appendix B. It is only in this case that we get an irreducible representation of the Heisenberg commutation relations in the case of the optimal discretization.

VII. SUMMARY AND CONCLUSIONS

In this paper we have investigated a wide class of discretization schemes for the derivative and/or momentum operator defined by (1.3). The optimal subclass of these schemes has been distinguished. These schemes give the best fit to the continuous operators and are symmetric with respect to the space reflections. The particular form of those schemes has been determined in Sec.

II. It has been shown that the solution found in Ref. 27 is the only solution in the form of finite series for the operator X. We also have determined that the commutation relation for the annihilation–creation-type of operators can be satisfied for a large class of difference operators, thus extending the result of Ref. 27, if one admits X in a form of an infinite series in E.

Our analysis in Sec. VI has shown that upon quantization the classically different optimal discretization schemes are all unitary equivalent. As a result, the number of points used to ''delocalize'' d/dx is unessential. For a fixed number of these points, say $2N+1$, the representation of the canonical commutation relations is reducible. Furthermore, by the result of Appendix B its irreducible components are unitary equivalent to the case with an infinite ($N\to\infty$) number of points.

Further extensions of this work can be done in several directions. First, one can investigate the case of an imaginary shift, as has been mentioned in Sec. I. Second, more general ansatz for the P and X operators can be analyzed. In particular, we suggest that the representation used in Ref. 24 can be included into our scheme by considering operator X being a nonlinear function of x. Finally, it would be interesting to study simple quantum mechanical systems within our approach.

It was brought to our attention that the so-called umbral calculus[37] rests on the same principle as our work. Umbral calculus is specifically designed to deal with representations of the canonical commutation relations on spaces of polynomials. Even though we do not make this assumption we recognize the evident similarity between the umbral calculus and our work. This deserves further study (for a very readable introduction to umbral calculus see also Ref. 38 where some physical aspects of umbral calculus are discussed).

ACKNOWLEDGMENTS

We are greatly indebted to P. O. Mazur for his insightful comments and constant encouragement. Also, we thank A. Herdegen for many illuminating suggestions regarding the content of Sec. VI. We also would like to thank the referee for pointing out to us the relevance of umbral calculus to the present work. The authors gratefully acknowledge the support of the State Committee for Scientific Research Republic Poland (KBN) 2 P03B 140 10 and the Natural Sciences and Engineering Research Council of Canada. AZG would like to thank the Kościuszko Foundation (NY) for the financial support and the Department of Physics, University of South Carolina, for hospitality.

APPENDIX A: HEINE SYMBOLS

The Heine symbol defined by (1.7) exhibit many interesting features. The deformation parameter $q\geqslant 0$ is usually bound to the (complex) region: $|q|\leqslant 1$ and it is parametrized exponentially by a (dimensionless) physical parameter l:

$$q=\exp\left(\frac{il}{2}\right). \tag{A1}$$

The above parametrization allows us to rewrite the Heine symbol (1.7) in the following widely used form:

$$[P]_q\equiv\frac{\sin(lP/2)}{\sin(l/2)}=\frac{\sinh(ilP/2)}{\sinh(il/2)}. \tag{A2}$$

Function $H(P,q)$ is antisymmetric in P:

$$H(P,q)=-H(-P,q). \tag{A3}$$

Also, it has the following limits for fixed q:

$$\lim_{P\to 0} H(P,q)=0,$$

$$\lim_{P\to 1} H(P,q)=1, \tag{A4}$$

$$\lim_{P\to\infty} H(P,q)=\infty, \quad q\neq 0,1.$$

As a function of q the function $H(P,q)$ is invariant under the transformation

$$q\to q'\equiv\frac{1}{q}. \tag{A5}$$

For $P=1$ it is constant: $H(1,q)=1$ and for $P<1$ $(P>1)$ it has maximum (minimum) at $q=1$ equal to P. The point $P=q=1$ is singular and the limits $\lim_{q\to 1}$ and $\lim_{P\to 1}$ do not commute. Also, the limit $\lim_{q\to 0}$ is singular—it is equal to 0 for $P\in[0,1)$, equal to 1 for $P=1$, and undefined for $P>1$.

APPENDIX B: A REMARK ABOUT D_N

Let us consider the function $f(\theta)=\theta$, $-\pi<\theta<\pi$. This function has the Fourier expansion

$$\theta=\sum_{n=1}^{\infty} b_n \sin n\theta,$$

where

$$b_n=\frac{2(-1)^n}{n}.$$

Let us recall that

$$D_N(\theta)=i\sum_{n=1}^{N} \frac{2(-1)^{n+1}(N!)^2}{n(N+n)!(N-n)!}\sin n\theta.$$

The main idea now is to compare the Nth partial sum S_N of $if(\theta)$ with the above sum. Clearly,

$$S_N(\theta)=i\sum_{n=1}^{N} \frac{2(-1)^{n+1}}{n}\sin n\theta.$$

The two sums look alike; the difference is the factor $(N!)^2/(N-n)!(N+n)!$.

We claim that $D_N(\theta)$ is an approximation to $i\theta$, i.e., $\lim_{N\to\infty} D_N(\theta)=i\theta$ in $(-\pi,\pi)$. This approximation is a result of a specific method of summation of the Fourier series, similar to the Fejer method of arithmetic means. Let us briefly review this subject. We refer the reader to Ref. 39 for more details. Given an infinite matrix $a=[a_{ij}:i,j=0,1,...]$ and a sequence $\{S_n:n=0,1,...\}$ we define a new sequence σ_n, $n=0,1,...$, by

$$\sigma_N=\sum_{n=0}^{\infty} a_{Nn}S_n.$$

We require that a satisfy the following conditions:

(i) $\lim_{N\to\infty} a_{Nn}=0$.
(ii) $B_N:=\Sigma_{n=0}^{\infty}|a_{Nn}|$ exists for every $N=0,1,...$, and the set $\{B_N:N=0,1,...\}$ is bounded.
(iii) $\lim_{N\to\infty}\Sigma_{n=0}^{\infty}a_{Nn}=1$.

If the σ_N have a limit s, then we say that the sequence of partial sums S_N is *a-summable* to the limit s. The most important fact for us is the following:

Theorem B1:[39] *If a satisfies (i), (ii), and (iii) and if S_N tends to a finite limit s, then* $\lim_{N\to\infty}\sigma_N=s$.

We claim that D_N are precisely the linear means σ_N for an appropriate choice of a. We define

$$a_{Nn}=\begin{cases}\dfrac{(2n+1)(N!)^2}{(N-n)!(N+n+1)!}, & \text{if } 0<n\leqslant N\\ 0, & \text{otherwise}\end{cases}.$$

It is easy to check that a so defined satisfy (i) and (ii). To see that (iii) is satisfied as well we observe that $a_N n$ satisfy

$$\sum_{m=n}^{N} a_{Nm}=\frac{(N!)^2}{(N-n)!(N+n)!},\quad 0\leqslant n\leqslant N.$$

In particular, for $n=1$ we obtain

$$\sum_{n=1}^{N} a_{Nn}=\frac{(N!)^2}{(N-1)!(N+1)!}=\frac{N}{N+1}.$$

By taking the limit $N\to\infty$ we obtain (iii). To prove our claim we need to show that $D_N=\sigma_N$. This follows from (B1). Finally, applying Theorem B1 we conclude that $\lim_{N\to\infty} D_N(\theta)=i\theta$.

[1] V. Ambarcumian and D. Ivanienko, Z. Phys. **64**, 563 (1930).
[2] The first questions about the world geometry were already posed by Riemann in the XIXth century: G. F. B. Riemann, in *Über die Hypothesen, welche der Geometrie zugrunde liegen*, edited by H. Weyl (Springer-Verlag, Berlin, 1919).
[3] P. Gibbs, ''The Small Scale Structure of Space-Time,'' preprint PEG-06-95 (1995), hep-th/9506171.
[4] I. Klebanov and L. Suskind, Nucl. Phys. B **309**, 175 (1988).
[5] R. Penrose, in *Quantum Theory and Beyond*, edited by E. A. Bastin (Cambridge U. P. Cambridge, 1971).
[6] G. Horowitz, talk at the Conference on Quantum Gravity, Durham (1994).
[7] C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995).
[8] J. D. Bekenstein, Lett. Nuovo Cimento **11**, 467 (1974).
[9] V. F. Mukhanov, JETP Lett. **44**, 63 (1986).
[10] T. Jacobson, Phys. Rev. D **44**, 1731 (1991).
[11] J. D. Bekenstein and V. F. Mukhanov, Phys. Lett. B **360**, 7 (1995).
[12] P. O. Mazur, Acta Phys. Pol. B **27**, 1849 (1996).
[13] G. 't Hooft, Nucl. Phys. B **342**, 471 (1990).
[14] G. 't Hooft, Class. Quantum Grav. **10**, 1023 (1993).
[15] G. 't Hooft, Class. Quantum Grav. **10**, 1653 (1993).
[16] Y. Aharonov, H. Pendelton, and A. Petersen, Int. J. Theor. Phys. **2**, 213 (1969).
[17] K. Wilson, Phys. Rev. D **10**, 2445 (1974).
[18] T. Regge, Nuovo Cimento **19**, 558 (1961).
[19] L. Bombelli, J. Lee, D. Meyer, and R. D. Sorkin, Phys. Rev. Lett. **59**, 521 (1987).
[20] L. J. Garay, Int. J. Mod. Phys. A **10**, 145 (1995).
[21] J. Lukierski, A. Nowicki, and H. Ruegg, Phys. Lett. B **293**, 344 (1992).
[22] P. O. Mazur, Acta Phys. Pol. B **26**, 1685 (1995).
[23] P. O. Mazur, Phys. Rev. Lett. **57**, 929 (1986); **59**, 2380 (1987).
[24] M. R. Douglas, Phys. Lett. B **238**, 176 (1990).
[25] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **109**, 642,Sec.3 (1926).
[26] There exists, however, a way of restoring the Leibniz rule by modyfying the product: Y. Bouguenaya and D. B. Fairlie, J. Phys. A **19**, 1049 (1986) and R. S. Ward, Phys. Lett. A **165**, 325 (1992). We are grateful to Prof. D. B. Fairlie for bringing these references to our attention.

[27] Yu. Smirnov and A. Turbiner, Mod. Phys. Lett. A **10**, 1795 (1995).
[28] S. C. Coutinho, *A Primer of Algebraic D-modules*, *LMSST 33* (Cambridge U.P., Cambridge, 1995).
[29] It can be shown that for classical dynamical systems the optimal schemes are closer to continuous derivative with respect to the stability analysis and they preserve the Hamiltonian structure of the phase space, as well (A. Z. Górski and J. Szmigielski, in preparation).
[30] E. A. B. Cole, Nuovo Cimento A **66**, 645 (1970).
[31] L. Frappat and A. Sciarrino, Phys. Lett. B **347**, 28 (1995).
[32] H. Widom, in *Studies in Real and Complex Analysis*, Vol. 3, edited by I. H. Hirschman, Jr., Prentice-Hall, Inc., Englewood Cliffs, NJ, 1965.
[33] G. Dorfmeister and J. Dorfmeister, J. Funct. Anal. **57**, 301 (1984).
[34] D. Schmüdgen, J. Funct. Anal. **50**, 8 (1983).
[35] D. Schmüdgen, Publ. Res. Inst. Math. Sci., Kyoto Univ. **19**, 601 (1983).
[36] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vols. 1 and 2.
[37] G-C. Rota, *Finite Operator Calculus* (Academic, New York, 1975).
[38] A. Dimakis and F. Mueller-Hoissen, J. Phys. A **29**, 6861 (1996).
[39] A. Zygmunt, *Trigonometric Series* (Cambridge U. P., Cambridge, 1959).

Lattice functions, wavelet aliasing, and SO(3) mappings of orthonormal filters

Sarah John
Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211

(Received 2 June 1997; accepted for publication 15 September 1997)

A formulation of multiresolution in terms of a family of dyadic lattices $\{S_j; j \in Z\}$ and filter matrices $M_j \subset \mathrm{U}(2) \subset \mathrm{GL}(2,C)$ illuminates the role of aliasing in wavelets and provides exact relations between scaling and wavelet filters. By showing the $\{DN; N \in Z_+\}$ collection of compactly supported, orthonormal wavelet filters to be strictly $\mathrm{SU}(2) \subset \mathrm{U}(2)$, its representation in the Euler angles of the rotation group SO(3) establishes several new results: a 1:1 mapping of the $\{DN\}$ filters onto a set of orbits on the SO(3) manifold; an equivalence of $D\infty$ to the Shannon filter; and a simple new proof for a criterion ruling out pathologically scaled nonorthonormal filters. © *1998 American Institute of Physics.* [S0022-2488(98)00301-6]

I. INTRODUCTION

The theory of wavelets, developed in recent years,[1–12] has found widespread applications in several areas in science and technology. Among their useful features are the phase space structure, which allows simultaneous windowing in the spatial and frequency domains, and their scale transformations, which allows microscopic focusing as well as zooming to examine spectral details both at the local as well as global scales. Wavelets, therefore, are more versatile than the fixed scale Wigner or Gabor phase space transforms, or the strictly global Fourier transforms.

Wavelets are either continuous or discrete. Continuous wavelets have a high degree of redundancy when spanning $L^2(R)$, thereby allowing discretization without loss of information. The discretized wavelets constitute frames or unconditional (Riesz) basis for $L^2(R)$ and is the basis for multiresolution analysis introduced by Mallat.[2,3] A particular collection of compactly supported orthonormal wavelets $\{DN; N \in Z_+\}$ first constructed by Daubechies,[1] which include the Haar and Daubechies wavelets, has proven useful in wavelet applications, its members mainly characterized by their finite interval of support and the order of differentiability. Another distinction between members of this collection relates to aliasing, a phenomenon associated with undersampling that would be particularly relevant in the selection of an optimal wavelet for any given application. For instance, in zoom, aliasing artifacts can seriously degrade the quality of reconstruction at any given scale, and in this case the Shannon wavelet would be perfect with unaliased subspaces but has the shortcoming of unbounded support. Again, in ill-posed inverse problems,[13] with noise, bandlimiting, and other signal distortions, a more rigorous information theoretic analysis is needed to determine optimized wavelets.[14,15]

This paper emphasizes the role of aliasing in multiresolution.[2] A key development is the use of the lattice or ''comb'' function $S_0(x)$, consisting of a train of Dirac deltas, to represent a unit lattice in the multiresolution subspace V_0, together with all the scaled lattices $\{S_j; j \in Z\}$ obtained by dyadic scale transformations acting on S_0. The algebra relating S_j at different scales is derived and the structure of 2×2 filter matrices $M_j \subset \mathrm{GL}(2,C)$ determined from two simultaneous linear equations representing aliased information.[16] Here $\mathrm{GL}(2,C)$ is the group of invertible 2×2 matrices. This approach may be compared to the linear phase filter theory[10,11] that by contrast relates only to unitary filter matrices, $M_j \subset \mathrm{U}(2)$, whereas the present formalism is more general, starting with $M_j \subset \mathrm{GL}(2,C)$, and eventually restricting to unitary matrices, $M_j \subset \mathrm{U}(2) \subset \mathrm{GL}(2,C)$ for multiresolution analysis. Related approaches using sampling may be found in the texts by Kaiser[7] and Holschneider.[9] Some well-known results, see for example the excellent texts by Chui,[5]

Daubechies,[6] and the more recent book by Hernandez and Weiss,[12] are rederived to illustrate the useful role of the lattice functions S_j, while other results are considered in greater detail, for instance, distinct representations of unitary matrices are written down explicitly to show how exact relations between scaling and wavelet filters may be obtained.

Several new results are also derived. It is shown that the relevant group for the $\{DN\}$ family is SU(2)⊂U(2). As SU(2) has a 2:1 homomorphism into the rotation group SO(3), a representation of the SU(2) group elements in the three Euler angles provides a framework for mapping the $\{DN\}$ filters onto a collection of closed orbits on the SO(3) manifold. The filters can then be viewed as being related to each other via rotations in three-dimensional space. This is distinct from the usual one-parameter representation of the filters that can only be representative of the U(1) group and not of the full structure of the non-Abelian SU(2) Lie group requiring three continuous parameters. The Euler angle representation also leads to a proof relating $D\infty$ to the Shannon filter and yields a simple new proof for a criterion ruling out pathologically scaled nonorthonormal filters.[17–19]

The paper is divided into the following sections. Section II briefly reviews continuous and discrete wavelets, and multiresolution.

In Sec. III a family of dyadic lattices $\{S_j\}$ and its algebra are derived and their role in multiresolution established. A derivation of the dilation equation is included to illustrate the useful role of lattice functions S_j. Although sampling is implicitly incorporated in all earlier works, the explicit use of lattice functions is new here.

In Sec. IV some well-known concepts on sampling and aliasing are discussed and a technique for the removal of aliasing provides the structure of 2×2 filter matrices $M_j\subset \mathrm{GL}(2,C)$. With $M_j\subset \mathrm{U}(2)$, exact relations between multiresolution filter elements are deduced from the general structure of unitary matrices. Further, a representation of the SU(2) group elements in the Euler angles leads to a heuristic derivation of the Haar filter.

In Sec. V the relevant group for the $\{DN\}$ filters is shown to be SU(2). Its representation in the Euler angles provides a 1:1 mapping of the filters onto a collection of orbits on the SO(3) manifold, and leads to a proof relating $D\infty$ to the Shannon filter.

Section VI discusses pathologically scaled filters with a criterion to eliminate them and an elementary new proof for it is given using the Euler angles.

Finally, Sec. VII gives a brief summary and outlook.

II. REVIEW

A brief review of continuous and discrete wavelet transforms as well as multiresolution is included for completeness and to establish notations and conventions.

A. Continuous wavelets

A continuous wavelet transform $W_f(a,b)$ of a function $f\in L^2(R^1)\cap L^1(R^1)$ is defined as a cross correlation between f and the conjugate of a continuous wavelet family

$$\left\{ \psi_{ab}:\psi_{ab}(x)=a^{-1/2}\psi\left(\frac{x-b}{a}\right);\ a\in R^1_+,\ b\in R^1 \right\} \tag{2.1}$$

and is given by

$$W_f(a,b)=\int_{-\infty}^{\infty} dx\ f(x)a^{-1/2}\overline{\psi\left(\frac{x-b}{a}\right)} \tag{2.2}$$

with Fourier transform,

$$\hat{W}_f(a,\omega)=a^{1/2}\hat{f}(\omega)\overline{\hat{\psi}(a\omega)}. \tag{2.3}$$

Their inverse transforms are given by the Grossmann–Morlet formulas,

$$f(x)=\frac{1}{C_\psi}\int_0^\infty da\ a^{-2}\int_{-\infty}^{\infty} db\ W_f(a,b)a^{-1/2}\psi\left(\frac{x-b}{a}\right) \tag{2.4}$$

and

$$\hat{f}(\omega)=\frac{1}{C_\psi}\int_0^\infty da\ a^{-2}\hat{W}_f(a,\omega)a^{1/2}\hat{\psi}(a\omega), \tag{2.5}$$

respectively. Here, the constant

$$C_\psi=\int_0^\infty (d\xi/\xi)|\hat{\psi}(\xi)|^2 \tag{2.6}$$

with the requirement

$$0<C_\psi<\infty \tag{2.7}$$

sets an admissibility condition on ψ, that ψ must have mean zero,

$$\int_{-\infty}^{\infty} dx\ \psi(x)=\hat{\psi}(0)=0, \tag{2.8}$$

indicating oscillations in the function.

Two completeness relations are implicit in the inverse transforms. From Eq. (2.5) one obtains

$$1=\frac{1}{C_\psi}\int_0^\infty da\ a^{-1}|\hat{\psi}(a\omega)|^2 \tag{2.9}$$

and from Eq. (2.4) the Grossmann–Morlet reproducing kernel $K(x,x')$ defined via

$$f(x)=\int_{-\infty}^{\infty} dx'\ K(x,x')f(x') \tag{2.10}$$

and given by

$$K(x,x')=\frac{1}{C_\psi}\int_0^\infty da\ a^{-3}\int_{-\infty}^{\infty} db\,\psi\left(\frac{x-b}{a}\right)\overline{\psi\left(\frac{x'-b}{a}\right)} \tag{2.11}$$

obtains a completeness relation in Hilbert space,

$$I=\int_0^\infty da\ a^{-2}\int_{-\infty}^{\infty} db|\psi_{ab}\rangle\langle\psi_{ab}|, \tag{2.12}$$

as can be seen in an expansion for $f(x)$ in the form

$$f(x)=\langle x|f\rangle=\int_0^\infty da\ a^{-2}\int_{-\infty}^\infty db\int_{-\infty}^\infty dx'\langle x|\psi_{ab}\rangle\langle\psi_{ab}|x'\rangle\langle x'|f\rangle,$$

where $\langle x|\psi_{ab}\rangle=\psi_{ab}(x)$ as given by Eq. (2.1).

B. Discrete wavelets

A continuous wavelet family, expressed in two continuous variables, has high redundancy when spanning a one-dimensional function space $L^2(R^1)$ and this in turn allows a discretization of the variables without loss of information. A discrete wavelet family, such as a dyadic family, given by

$$\{\psi_{jk}:\psi_{jk}(x)=2^{j/2}\psi(2^jx-k);\ j,k\in Z\},\tag{2.13}$$

constitutes a frame, or more restrictively, an unconditional (Riesz) basis for $L^2(R^1)$, depending on the wavelet ψ. Frames allow linear dependence among its elements implying redundancy, whereas Riesz bases constitute a subset of frames with linearly independent elements and no redundancy.

The Riesz coefficient R_{jk} for $f\in L^2(R^1)$ is defined as a cross correlation, or equivalently as an inner product $\langle\psi_{jk}|f\rangle$ given by

$$R_{jk}=\int_{-\infty}^\infty dx'\ 2^{j/2}\overline{\psi(2^jx'-k)}f(x').\tag{2.14}$$

A simple inversion formula is obtained in terms of its dual family $\{\tilde{\psi}_{jk}:j,k\in Z\}$, and is given by

$$f(x)=\sum_{j,k}\ R_{jk}\tilde{\psi}_{jk}.\tag{2.15}$$

For an orthonormal family, $\tilde{\psi}=\psi$ and one may deduce a discrete reproducing kernel $K_d(x,x')$ from Eqs. (2.14) and (2.15) by writing

$$f(x)=\int_{-\infty}^\infty dx'\ K_d(x,x')f(x')\tag{2.16}$$

with

$$K_d(x,x')=\sum_{j,k}\ 2^j\psi(2^jx-k)\overline{\psi(2^jx'-k)},\tag{2.17}$$

which gives a completeness relation in Hilbert space of the form

$$\sum_{j,k}\ |\psi_{jk}\rangle\langle\psi_{jk}|=I.\tag{2.18}$$

To generate multiresolution, on the other hand, the discrete *wavelet* coefficient C_{jk} must be defined as a convolution with an orthonormal wavelet family $\{\psi_j:\psi_j(x)=2^{j/2}\psi(2^jx);j\in Z\}$ followed by sampling on a scaled lattice. With convolution defined by

$$C_j(x)=\int_{-\infty}^\infty dx'\ f(x')2^{j/2}\psi(2^j(x-x')),$$

sampling on the scaled lattice gives

$$C_{jk}=\int_{-\infty}^{\infty} dx'\ f(x')2^{j/2}\psi(k-2^j x'). \tag{2.19}$$

This definition is necessary for multiresolution, as made clear in the derivation of the dilation equation (see Sec. III E). The wavelet coefficients may also be viewed as Riesz coefficients in the wavelet basis $\{\psi^{-}{}_{jk}:\psi^{-}{}_{jk}=2^{j/2}\psi(k-2^j x);j,k\in Z\}$, where the negative sign in the superscript indicates an inversion $2^j x-k\rightarrow-(2^j x-k)$ compared to the basis given by Eq. (2.13).

C. Multiresolution

A discrete wavelet family generates multiresolution,[2,3] which leads to an approximation theory for analyzing a function $f\in L^2(R^1)$ at distinct scales. The underlying concept in multiresolution is a nested sequence of closed subspaces $\{V_j:V_j\subset V_{j+1};j\in Z\}$, which has the following properties for dyadic scaling:

(i) $f(2x)\in V_{j+1}\leftrightarrow f(x)\in V_j$,
(ii) $\cup_{j\in Z}V_j$ is dense in $L^2(R^1)$, $\cap_{j\in Z}V_j=\{0\}$,
(iii) $V_{j+1}=V_j\oplus W_j$, $V_j\perp W_j$, and $L^2(R^1)=\oplus_{j\in Z}W_j$,
(iv) there exist related pairs of functions $\phi,\psi\in V_0$ such that $\{\phi_{jk}:\phi_{jk}=2^{j/2}\phi(2^j x-k);k\in Z\}$ forms an orthonormal basis for V_j, $\{\psi_{jk}:\psi_{jk}=2^{j/2}\psi(2^j x-k);k\in Z\}$ an orthonormal basis for W_j, and $\{\psi_{jk};j,k\in Z\}$ an orthonormal basis for $L^2(R^1)$. ϕ is the scaling function and ψ is the wavelet.

The spaces ν_j, w_j of $2^{j+1}\pi$ periodic functions, with $\nu_j\oplus w_j=\nu_{j+1}$ are defined by

$$\nu_j=\{\hat{m}_j;\forall\hat{m}_j(\omega)\in L^2[-2^j\pi,2^j\pi]\}, \tag{2.20}$$

$$w_j=\{\hat{m}_j;\forall\hat{m}_j(\omega)\in L^2[-2^{j+1}\pi,-2^j\pi;2^j\pi,2^{j+1}\pi]\}, \tag{2.21}$$

where w_j obviously is a translation of ν_j space and

$$\hat{m}_j(\omega)=\sum_k m_k e^{-i2^{-j}\omega k},\quad m_k\in C. \tag{2.22}$$

A necessary condition for multiresolution is the dilation equation,

$$2^{1/2}\hat{\phi}_j(2\omega)=\hat{m}_j(\omega)\hat{\phi}_j(\omega), \tag{2.23}$$

where $\hat{m}_j(\omega)\in\nu_j$. For $j=0$ and with $\hat{m}'(\omega)=2^{-1/2}\hat{m}_0(\omega)$, iteration ''constructs'' $\hat{\phi}=\hat{\phi}_0$,

$$\hat{\phi}(\omega)=\Pi_{n=1}^{\infty}\hat{m}'(2^{-n}\omega). \tag{2.24}$$

The constraints placed on $\hat{m}=\hat{m}_0$ to obtain an orthonormal basis for V_0, $\{\phi_{0k};k\in Z\}$, is the unitarity relation

$$|\hat{m}(\omega)|^2+|\hat{m}(\omega-\pi)|^2=2 \tag{2.25}$$

with $\hat{m}(0)=2^{1/2}$ implying that only a subset of ν_j qualify as multiresolution filters.

Multiresolution subspaces V_j and W_j consisting of $f_j\in L^2$ are defined by

$$V_j=\{\hat{f}_j:\hat{f}_j(\omega)=\hat{m}_j(\omega)\hat{\phi}_j(\omega);\quad\forall\hat{m}_j\in\nu_j\}, \tag{2.26}$$

$$W_j=\{\hat{f}_j:\hat{f}_j(\omega)=\hat{m}_j(\omega)\hat{\psi}_j(\omega);\quad\forall\hat{m}_j\in\nu_j\}, \tag{2.27}$$

which basically implies an expansion in the orthonormal basis of the subspace. For example,

$$f_j(x)=\sum_k m_k \phi_{jk}(x), \quad f_j \in V_j . \tag{2.28}$$

III. LATTICE FUNCTIONS AND WAVELET THEORY

In this section a family of dyadically scaled lattice functions $\{S_j ; j \in Z\}$ is introduced and their algebra derived. Following a well-known method, these functions derive expressions for a sequence $\{f_n\}$ as a continuous variable in x. Their useful role in multiresolution analysis is exemplified by a derivation of the dilation equation. As the δ function plays a pivotal role in the formulation, a few of its properties are now delineated.

A. The δ function and its properties

The Dirac δ function is defined by

$$\int_{-\infty}^{\infty} \delta(x)\ dx=1, \tag{3.1}$$

with

$$\delta(x)=0, \quad x \neq 0. \tag{3.2}$$

Strictly, this definition precludes δ from the L^2 space, as it must assume an infinite value at $x=0$ and therefore is to be considered as a ''generalized'' function, the definition also providing the following rule:

$$\delta(a^j x)=|a|^{-j}\delta(x). \tag{3.3}$$

A family of scaled deltas $\{\delta_j(x); j \in Z\}$ is now introduced, with

$$\delta_j(x)=a^{j/2}\delta(a^j x), \quad a \in R^1_+,$$

which on applying the rule [Eq. (3.3)] may be written as

$$\delta_j(x)=a^{-j/2}\delta(x). \tag{3.4}$$

These scaled deltas are to be viewed as existing in a scaled space. This allows one to define unit L^p norm, $1 \leqslant p < \infty$, by associating a factor $a^{j/2}$ for each delta in the measure, and is given by

$$\int_{-\infty}^{\infty} |\delta_j(x)|^p a^{jp/2}\ dx=\int_{-\infty}^{\infty} |\delta(x)|^p\ dx=1.$$

Then it follows that

$$\int_{-\infty}^{\infty} \delta_j(\cdot-m_1)\,\delta_j(\cdot-m_2)\cdots\delta_j(\cdot-m_p)a^{jp/2}\ dx=\delta_{m_1,m_2\cdots m_p}, \tag{3.5}$$

where $\delta_j(\cdot-m_1)=a^{j/2}\delta(a^j x-m_1)$, etc.

One very useful property of the delta function is the linear functional

$$\int_{-\infty}^{\infty} f(x')\,\delta(x-x')dx'=f(x), \tag{3.6}$$

which picks out the value of f at a given value x.

An integral representation for the delta function given by

$$\delta(x)=(2\pi)^{-1}\int_{-\infty}^{\infty}e^{ikx}\,dk \tag{3.7}$$

is useful in proving results in Sec. III B.

B. Lattice functions: S_j

A unit lattice is represented by the comb or lattice function S_0 consisting of a train of Dirac deltas

$$S_0(x)=\sum_n\ \delta(x-n) \tag{3.8}$$

with Fourier transform,

$$\hat{S}_0(\omega)=\sum_k\ e^{-ik\omega}=2\pi\sum_n\ \delta(\omega-2\pi n), \tag{3.9}$$

where the second equality follows from the Poisson summation formula written in the form

$$\sum_k\ F(k)=\sum_n\ \left\{\int_{-\infty}^{\infty}e^{2\pi inx}\ F(x)\ dx\right\}, \tag{3.10}$$

with $F(x)=e^{-ix\omega}$, and an application of Eq. (3.7).

A family of dyadically scaled lattices $\{S_j;j\in Z\}$ is now introduced, with

$$S_j(x)=2^{j/2}\ S_0(2^jx)=2^{j/2}\sum_n\ \delta(2^jx-n)=\sum_n\ \delta_j(x,n), \tag{3.11}$$

where

$$\delta_j(x,n)=2^{j/2}\delta(2^jx-n). \tag{3.12}$$

Their Fourier transforms are given by

$$\hat{S}_j(\omega)=2^{-j/2}\hat{S}_0(2^{-j}\omega)=2^{-j/2}2\pi\sum_n\ \delta(2^{-j}\omega-2\pi n). \tag{3.13}$$

Applying Eq. (3.3) to the preceding equations obtains the following forms:

$$S_j(x)=2^{-j/2}\sum_n\ \delta(x-2^{-j}n), \tag{3.14}$$

$$\hat{S}_j(\omega)=2^{j/2}\ 2\pi\sum_n\ \delta(\omega-2^{j+1}\pi n). \tag{3.15}$$

The following properties of the lattice functions are easily proved,

$$S_j^2(x)=S_j(x), \tag{3.16}$$

$$S_j(x)S_{j-1}(x)=S_{j-1}(x). \tag{3.17}$$

A relation between S_j and S_{j+1}, deduced from the defining equation (3.11) is given by

$$S_{j+1}(x)=2^{1/2}S_j(2x), \tag{3.18}$$

with Fourier transform

$$\hat{S}_{j+1}(\omega)=2^{-1/2}\hat{S}_j(\omega/2). \tag{3.19}$$

Similarly, Eq. (3.15) obtains a relation between S_j and S_{j-1} as follows:

$$\begin{aligned}\hat{S}_{j-1}(\omega)&=2^{(j-1)/2}2\pi\sum_n \delta(\omega-2^j\pi n)\\ &=2^{(j-1)/2}2\pi\sum_{n\ \text{even}}[\delta(\omega-2^j\pi n)+\delta(\omega-2^j\pi(n+1))]\\ &=2^{-1/2}[S_j(\omega)+S_j(\omega-2^j\pi)]\end{aligned} \tag{3.20}$$

with inverse Fourier transform given by

$$S_{j-1}(x)=2^{-1/2}S_j(x)[1+e^{i2^j\pi x}], \tag{3.21}$$

which is easily verified for $j=0$ when

$$S_{-1}(x)=2^{-1/2}S_0(x)[1+e^{i\pi x}]=2^{1/2}\sum_{n\ \text{even}}\delta(x-n)=2^{1/2}\sum_n \delta(x-2n). \tag{3.22}$$

C. Sampling

Sampling constitutes a mapping of $f\in L^2(R^1)$ onto a sequence $\{f_n\}$ with the map invertible for Shannon or sufficient sampling. The following proposition extends to dyadic lattices a well-known representation of a sampled sequence $\{f_n\}$ written as a generalized function, one advantage being that conversion formulas such as $f*g\leftrightarrow\hat{f}\hat{g}$ valid for continuous functions can now be readily adapted.

Note: In the following $f_{S_j}\in\nu_j$, the space of $2^{j+1}\pi$ periodic functions, and $f\in L^2$.

Proposition 3.1: A sequence $\{f_k=f(2^{-j}k)\}$ obtained by sampling $f\in L^2$ on a lattice S_j has representation

$$f_{S_j}(x)=f(x)S_j(x) \tag{3.23}$$

with Fourier transform represented in either form

$$\hat{f}_{S_j}(\omega)=2^j\sum_n \hat{f}(\omega-2^{j+1}\pi n) \tag{3.24}$$

or

$$\hat{f}_{S_j}(\omega)=\sum_n f(2^{-j}n)e^{-i2^{-j}n\omega}. \tag{3.25}$$

Proof: Equation (3.23) provides the required sampled sequence $\{f_k=f(2^{-j}k)\}$ since

$$f_{S_j}(x)=f(x)\sum_n 2^{j/2}\delta(2^j x-n)$$

and (including a factor 2^j to the measure for the two deltas in the integrand)

$$\begin{aligned} f_{S_j}(k)&=\int_{-\infty}^{\infty} f_{S_j}(x)\,\delta_j(x,k)2^j\,dx \\ &=\sum_n 2^{2j}\int_{-\infty}^{\infty} f(x)\,\delta(2^j x-n)\,\delta(2^j x-k)dx \\ &=\sum_n f(2^{-j}n)\,\delta_{k,n} \\ &=f(2^{-j}k). \end{aligned}$$

One method for obtaining the Fourier transform of Eq. (3.23) is by applying the convolution formula $f*g\leftrightarrow\hat{f}\hat{g}$,

$$\begin{aligned} \hat{f}_{S_j}(\omega)&=\hat{f}(\omega)*\hat{S}_j(\omega) \\ &=2^j\,2\pi\sum_n\int_{-\infty}^{\infty}(2\pi)^{-1}\,d\omega'\,\hat{f}(\omega')\,\delta(\omega-\omega'-2^{j+1}\pi n) \\ &=2^j\sum_n\hat{f}(\omega-2^{j+1}\pi n). \end{aligned}$$

Note the use of phase-space measure $(2\pi)^{-1}dx\,d\omega$, which implies $d\omega\rightarrow(2\pi)^{-1}\,d\omega$. Alternatively, by performing a direct Fourier transform,

$$\hat{f}_{S_j}(\omega)=\int_{-\infty}^{\infty}f(x)S_j(x)e^{-i\omega x}2^{j/2}\,dx=\sum_n f(2^{-j}n)e^{-i2^{-j}n\omega}.$$

The equivalence of the two forms can also be established directly by the Poisson summation formula. □

D. Multiresolution and lattice functions

The space ν_j of all $2^{j+1}\pi$ periodic functions is equivalent to the space of sampled sequences of f on S_j, $\forall f\in L^2$, that is,

$$\nu_j\equiv\{\hat{f}_{S_j}:\hat{f}_{S_j}(\omega)=\hat{f}(\omega)*\hat{S}_j(\omega);\ \ \forall f\in L^2\}\equiv\{\hat{m}_j;\forall\hat{m}_j(\omega)\in L^2[-2^j\pi,2^j\pi]\},\tag{3.26}$$

and therefore V_j as specified by Eq. (2.26) can be expressed in terms of S_j in the form

$$V_j=\{\hat{f}_j:\hat{f}_j(\omega)=\hat{f}(\omega)*\hat{S}_j(\omega)\,\hat{\phi}_j(\omega);\ \ \forall\hat{f}\in L^2\}.\tag{3.27}$$

The multiresolution subspaces for the Shannon wavelet have a particularly simple structure with $\hat{\phi}_j(\omega)=\chi_j(\omega)$, where the characteristic function χ_j is defined by

$$\chi_j(\omega)=\begin{cases}1, & 0\leqslant|\omega|<\Omega_j\\ 0, & \text{otherwise,}\end{cases}\tag{3.28}$$

and where $\Omega_j = 2^j \pi$ is the Nyquist bandwidth of S_j. In this case,

$$V_j = \{\hat{f}_j : \hat{f}_j(\omega) = \hat{f}_{S_j}(\omega)\chi_j(\omega);\ \ \forall \hat{f} \in L^2\} \tag{3.29}$$

and the orthogonal wavelet subspaces W_j, $L^2 = \oplus_{j \in Z} W_j$ form a union of disjoint regions in the frequency domain. All multiresolution properties are then immediately apparent since Ω_j spans V_j exactly, and consequently,

(i) $\Omega_j < \Omega_{j+1} \rightarrow V_j \subset V_{j+1}$,
(ii) $\hat{f}_j(\omega) \in V_j \leftrightarrow \hat{f}_j(2\omega) \in V_{j-1}$,
(iii) $\Omega_{j \rightarrow \infty} = \infty \rightarrow \cup_{j \in Z} V_j = V_{j \rightarrow \infty} = \infty$ is dense in $L^2(R_1)$,
(iv) $\Omega_{j \rightarrow -\infty} = 0 \rightarrow \cap_{j \in Z} V_j = V_{j \rightarrow -\infty} = \{0\}$,
(v) $\chi_j \in V_j$ generates an orthonormal basis for V_j.

E. Dilation equation using S_j

Wavelet theory formulated in terms of a sequence of sampled values $\{f_n\}$ has the drawback that mathematical operations such as convolution and sampling are not expressed symbolically, and consequently many of the underlying concepts are not evident per se in the equations. By using lattice functions S_j to represent $\{f_n\}$ as a distribution, the expressions attain a more obvious interpretation while also simplifying mathematical operations in symbolic manipulations. As an illustration, a derivation of the dilation equation,

$$\hat{\phi}_j(2\omega) = 2^{-1/2} \hat{m}_{S_j}(\omega) \hat{\phi}_j(\omega) \tag{3.30}$$

follows incorporating the lattice functions S_j.

There are two distinct methods for obtaining the scaling and wavelet coefficients. For instance, the scaling coefficients D_{jk} may be obtained either via a direct convolved projection of $f \in L^2$ into ν_j or alternatively via successive multiresolution decomposition in the form $\nu_{j+1} \rightarrow \nu_j \oplus w_j$. The first method is the true wavelet analysis, where f is convolved with $\{\phi_j(x) = 2^{j/2}\phi(2^j x); j \in Z\}$ followed by sampling on $\{S_j; j \in Z\}$, obtaining

$$D_{S_j}(x) = (f * \phi_j(x)) S_j(x) \quad \in \nu_j = L^2[-2^j\pi, 2^j\pi], \tag{3.31}$$

$$D_{S_{j-1}}(x) = (f * \phi_{j-1}(x)) S_{j-1}(x) \quad \in \nu_{j-1} = L^2[-2^{j-1}\pi, 2^{j-1}\pi], \tag{3.32}$$

with the scaling coefficients given by $\{D_{jk}\} \equiv \{D_{S_j}(k); k \in Z\}$.

The second method is the multiresolution analysis, with a successive projection from $\nu_j \rightarrow \nu_{j-1}$. This is achieved by convolving sequence $\{D_{jk}; k \in Z\}$ with masking coefficients $\{m_k; k \in Z\}$, $\hat{m}_{S_j}(\omega) \in \nu_j$, followed by 2:1 downsampling, obtaining

$$D_{S_{j-1}}(x) = (m_{S_j}(x) * D_{S_j}(x)) S_{j-1}(x) \in \nu_{j-1}, \tag{3.33}$$

where on substituting for $D_{S_j}(x)$ and $D_{S_{j-1}}(x)$ from Eqs. (3.31) and (3.32)

$$f * \phi_{j-1}(x) S_{j-1}(x) = m_{S_j}(x) * f * \phi_j(x) S_j(x) S_{j-1}(x) = m_{S_j}(x) * f * \phi_j(x) S_{j-1}(x) \tag{3.34}$$

and therefore

$$f * \phi_{j-1}(x) = m_{S_j}(x) * f * \phi_j(x) \tag{3.35}$$

with Fourier transform

$$\hat{f}(\omega)\hat{\phi}_{j-1}(\omega)=\hat{m}_{S_j}(\omega)\hat{f}(\omega)\hat{\phi}_j(\omega). \tag{3.36}$$

On substituting $\hat{\phi}_{j-1}(\omega)=2^{1/2}\hat{\phi}_j(2\omega)$ one obtains the dilation equation

$$\hat{\phi}_j(2\omega)=2^{-1/2}\hat{m}_{S_j}(\omega)\hat{\phi}_j(\omega), \quad \hat{m}_{S_j}(\omega)\in\nu_j \tag{3.37}$$

with inverse Fourier transform

$$\phi_j(x/2)=2^{1/2}\sum_n m_n\phi_j(x-2^{-j}n). \tag{3.38}$$

The wavelet coefficients C_{jk} are similarly obtained, with the ''wavelet'' projection from L^2 given by

$$C_{S_{j-1}}(x)=(f*\psi_{j-1}(x))S_{j-1}(x)\in w_{j-1} \tag{3.39}$$

and ''multiresolution'' from $\nu_j\rightarrow w_{j-1}$ with masking sequence $\{m^1{}_k:k\in Z\}$, $\hat{m}^1{}_{S_j}(\omega)\in w_j$, given by

$$C_{S_{j-1}}(x)=(m^1{}_{S_j}(x)*D_{S_j}(x))S_{j-1}(x) \quad \in w_{j-1}, \tag{3.40}$$

obtaining

$$\hat{\psi}_j(2\omega)=2^{-1/2}\hat{m}^1{}_{S_j}(\omega)\hat{\phi}_j(\omega), \quad \hat{m}^1{}_{S_j}(\omega)\in w_j. \tag{3.41}$$

IV. WAVELET ALIASING

Multiresolution analysis, first introduced by Mallat[2,3] for two-dimensional image processing, has the desirable feature of reconstructing images in increasing resolution as higher scales are incorporated. Some of these added details arise from the unfolding of aliased frequencies that were present in a blurred form at lower scales. An important distinction between various multiresolutions obtained from distinct wavelet families relates to the amount of aliasing present in the wavelet coefficients. In this section some well-known concepts of sampling and aliasing are discussed and a general formalism for removal of aliasing forms the framework for understanding how multiresolution subspaces arise, the approach also clearly revealing the way aliasing gets incorporated into wavelet coefficients. Also, exact relations between wavelet and scaling filters are obtained and a representation of the filters in the Euler angles provides a heuristic derivation of the Haar filter.

A. Aliasing

For a given lattice S_j with sampling interval $\Delta_j=2^{-j}$, the Nyquist bandwidth $\Omega_j=2\pi(2\Delta_j)^{-1}=2^j\pi$ corresponds to the maximum frequency that can be accurately sampled on that lattice. Therefore, a function $f\in L^2$ bandlimited to $[-\Omega,\Omega]$ when sampled on a lattice S_j with bandwidth $\Omega_j<\Omega$ results in a summing of frequencies beyond Ω_j with those within the passband $[-\Omega_j,\Omega_j]$ giving rise to the phenomenon of aliasing. A familiar example is the Moire patterns seen in a rotating fan, where a low frequency rotation appears superimposed on the actual high frequency rotation due to the insufficient sampling rate of the eye. Aliasing is more easily analyzed in the frequency domain where the Fourier transform of a sampled function is given by

$$\hat{f}_{S_j}(\omega)=2^j\sum_n \hat{f}(\omega-2\Omega_j n), \quad \Omega_j=2^j\pi \tag{4.1}$$

and on multiplying both sides by the characteristic filter $\chi_j(\omega)$ [given by Eq. (3.28)] one obtains the passband frequencies given by

$$\hat{g}(\omega)=\hat{f}_{S_j}(\omega)\chi_j(\omega)=2^j\sum_n \hat{f}(\omega-2^{j+1}\pi n)\chi_j(\omega). \tag{4.2}$$

One particular reconstruction for f from its sampled values is obtained by taking the inverse Fourier transform of the first equality obtaining

$$g(x)=f_{S_j}(x)*(\mathscr{F}^{-1}\chi_j)(x)=f_{S_j}(x)*2^j\ \mathrm{sinc}(2^j\pi x), \tag{4.3}$$

where $\mathrm{sinc}(x)=\sin(x)/x$. In the case of either sufficient or oversampling, when $\Omega_j\geqslant\Omega$, there are no overlap of the translates in Eq. (4.2) into the passband region, and the filter χ picks out only one term, namely, $\hat{g}(\omega)=2^j\hat{f}(\omega)$, with inverse Fourier transform

$$g(x)=2^jf(x)=\sum_n \int_{-\infty}^{\infty} f(x')2^{-j/2}\delta(x'-2^{-j}n)2^j\ \mathrm{sinc}(2^j\pi(x-x'))2^{j/2}\,dx',$$

obtaining the famous *Whittacker–Shannon–Kotel'nikov* interpolation formula for exact reconstruction from sampled values $f(\Delta_j n)$,

$$f(x)=\sum_n f(\Delta_j n)\mathrm{sinc}(\Omega_j(x-\Delta_j n)). \tag{4.4}$$

For insufficient sampling, when $\Omega_j<\Omega$, the translates in the summation in Eq. (4.2) overlap producing aliasing (Fig. 1). This precludes reconstruction of $f(x)$ from a single sampled sequence $\{f_n\}$. In signal processing, techniques for removal of aliasing, that unfold both high and low frequency components from subsampled data, such as the filter bank and subband coding techniques, rely on several distinct subsampled sequences obtained by phase filtering and processed simultaneously. In image processing, distinct subsampled sets may be obtained by varying the optical transfer function, effected by changing the width of the camera lens.[16]

Wavelet aliasing also has its origin in undersampling as a wavelet coefficient C_{jk} is obtained by convolving $f\in L^2$ with wavelet ψ_j followed by sampling on S_j, given by

$$C_j(x)=f(x)*\psi_j(x)S_j(x)\in w_j, \tag{4.5}$$

with Fourier transform

$$\hat{C}_j(\omega)=\hat{f}\hat{\psi}_j*\hat{S}_j(\omega)=2^j\sum_n \hat{f}(\omega-2^{j+1}\pi n)\hat{\psi}_j(\omega-2^{j+1}\pi n). \tag{4.6}$$

Now if the bandwidth of the convolved function $\hat{f}\hat{\psi}_j$ exceeds the passband Ω_j of the sampling lattice, adjacent, possibly several translates overlap producing aliased coefficients $C_{jk}=C_j(k)$.

In dyadic multiresolution, C_{jk} is obtained via successive 2:1 downsampling, with aliasing occurring at most from one adjacent translate at each scale, thereby enabling the much simpler algorithm used in practice for successive reconstruction and removal of aliasing.

B. Removal of aliasing

As an approximation theory, multiresolution pertains to a reconstruction sequence $\{f^j;j\in Z\}\rightarrow f$, obtained by incorporating successively higher scales $\nu_{j-1}\oplus w_{j-1}\rightarrow\nu_j$, which improves sharpness and fidelity by incorporating higher frequencies while simultaneously unfolding aliased frequencies. The formulation for the removal of aliasing starts with a more general class of

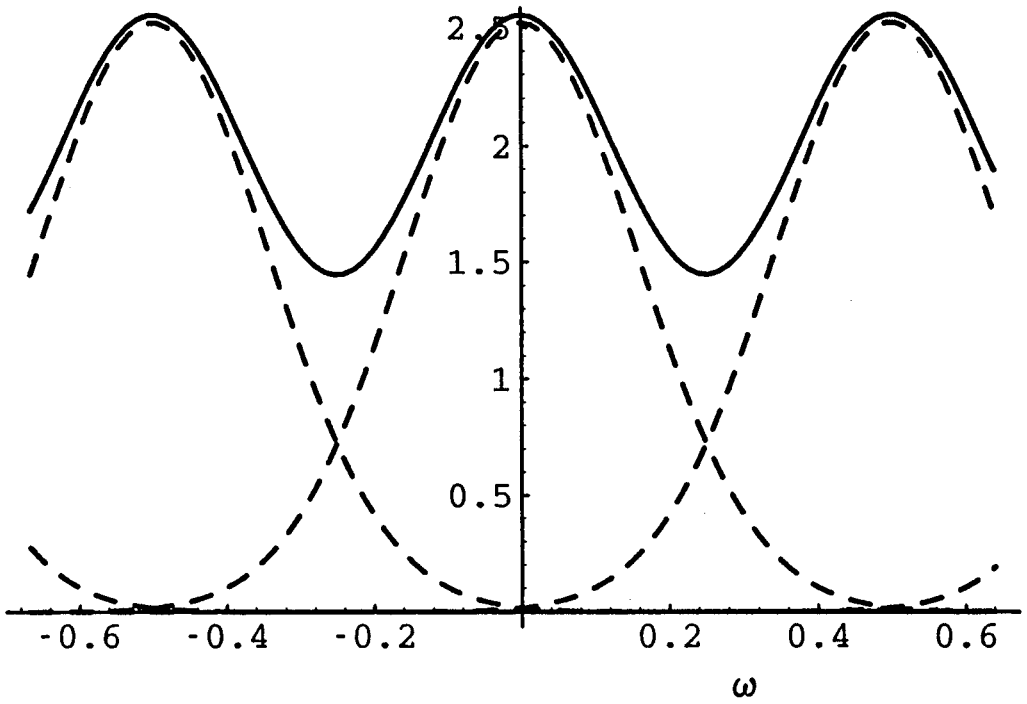

FIG. 1. Aliasing due to undersampling. The figure plots Eq. (4.1) for $j=-1$. The dotted lines are the translated functions $\frac{1}{2}\hat{f}(\omega-\pi n)$ and the solid line is their sum, representing the sampled function. Aliasing arises from the overlapping regions of the translates (ω is in units of 2π).

masking coefficients $\{m_k\}\in C$ than indicated for multiresolution, and subsequently restricted to obey orthonormality relations. The following proposition indicates that in dyadic multiresolution, at each 2:1 downsampling, aliasing accumulates from overlap of a single nearest translate.

Proposition 4.1: Aliasing in 2:1 *downsampling of a sequence* $\{f_k=f(2^{-j}k);k\in Z\}$ *arises from overlap of a single nearest translate, with*

$$\hat{f}_{S_{j-1}}(\omega)=2^{-1/2}[\hat{f}_{S_j}(\omega)+\hat{f}_{S_j}(\omega-2^j\pi)]. \tag{4.7}$$

Proof: From Proposition (3.1) the sequence has representation $f_{S_j}(x)$ with 2:1 downsampling given by

$$f_{S_{j-1}}(x)=f_{S_j}(x)S_{j-1}(x) \tag{4.8}$$

with Fourier transform

$$\begin{aligned}\hat{f}_{S_{j-1}}(\omega)&=\hat{f}_{S_j}(\omega)*\hat{S}_{j-1}(\omega)\\&=\hat{f}_{S_j}(\omega)*2^{-1/2}[\hat{S}_j(\omega)+\hat{S}_j(\omega-2^j\pi)]\\&=2^{-1/2}[\hat{f}_{S_j}(\omega)+\hat{f}_{S_j}(\omega-2^j\pi)],\end{aligned} \tag{4.9}$$

where the second equality follows from Eq. (3.20) and the last from the equivalence

$$f_{S_j}(x)S_j(x)\equiv f_{S_j}(x). \tag{4.10}$$

□

Evidently, the original sequence cannot be recovered directly from the downsampled sequence. However, consider the following. Generate a sequence $\{g^0{}_k; k \in Z\}$ by convolving $\{f_k; k \in Z\}$ with a masking sequence $\{m^0{}_k; k \in Z\}$ followed by a 2:1 downsampling, obtaining the sequence $\{g^0{}_k; k \in 2Z\}$. Then from Proposition (4.1)

$$\hat{g}^0{}_{S_{j-1}}(\omega) = 2^{-1/2}[\hat{g}^0{}_{S_j}(\omega) + \hat{g}^0{}_{S_j}(\omega - 2^j\pi)], \tag{4.11}$$

where

$$\hat{g}^0{}_{S_j}(\omega) = \hat{m}^0{}_{S_j}(\omega)\hat{f}_{S_j}(\omega). \tag{4.12}$$

This suggests that to recover f_k a second equation may be set up with a different masking sequence $\{m^1{}_k\}$, with the two equations written out explicitly in the form

$$\hat{g}^0{}_{S_{j-1}}(\omega) = 2^{-1/2}[\hat{m}^0{}_{S_j}(\omega)\hat{f}_{S_j}(\omega) + \hat{m}^0{}_{S_j}(\omega - 2^j\pi)\hat{f}_{S_j}(\omega - 2^j\pi)], \tag{4.13}$$

$$\hat{g}^1{}_{S_{j-1}}(\omega) = 2^{-1/2}[\hat{m}^1{}_{S_j}(\omega)\hat{f}_{S_j}(\omega) + \hat{m}^1{}_{S_j}(\omega - 2^j\pi)\hat{f}_{S_j}(\omega - 2^j\pi)]. \tag{4.14}$$

Solving for $\hat{f}_{S_j}$, one obtains

$$\mathbf{F}_j(\omega) = M_j^{-1}(\omega)\mathbf{G}_{j-1}(\omega), \tag{4.15}$$

where

$$\mathbf{F}_j(\omega) = \begin{pmatrix} \hat{f}_{S_j}(\omega) \\ \hat{f}_{S_j}(\omega - 2^j\pi) \end{pmatrix}, \quad \mathbf{G}_{j-1}(\omega) = \begin{pmatrix} \hat{g}^0{}_{S_{j-1}}(\omega) \\ \hat{g}^1{}_{S_{j-1}}(\omega) \end{pmatrix} \tag{4.16}$$

and

$$M_j(\omega) = 2^{-1/2}\begin{pmatrix} \hat{m}^0{}_{S_j}(\omega) & \hat{m}^0{}_{S_j}(\omega - 2^j\pi) \\ \hat{m}^1{}_{S_j}(\omega) & \hat{m}^1{}_{S_j}(\omega - 2^j\pi) \end{pmatrix}, \tag{4.17}$$

the only requirement for a solution being that M_j be invertible, that is, $M_j \in \mathrm{GL}(2,C)$, the general linear group of 2×2 invertible matrices. The components of $\mathbf{F}_j$ are defined in orthogonal intervals within the passband of S_j, the upper component $\hat{f}_{S_j}(\omega)$ in $0 \leqslant |\omega| < \Omega_j/2$ and the lower component $\hat{f}_{S_j}(\omega - 2^j\pi)$ being the same as $\hat{f}_{S_j}(\omega)$ in the interval $\Omega_j/2 \leqslant |\omega| < \Omega_j$. The components of $\mathbf{G}_{j-1}$ are defined on S_{j-1}, each representing distinct aliased information. A solution in the interval $0 \leqslant |\omega| \leqslant \Omega_j/2$ unfolds the aliasing obtaining $\hat{f}_{S_j}(\omega)$ in $0 \leqslant |\omega| < \Omega_j$.

C. Unitary groups and multiresolution

Most of the results in this section are implicit in earlier works. In general, the only requirement on $M_j \in \mathrm{GL}(2,C)$ is that $\det M_j \neq 0$ in the interval $0 \leqslant |\omega| \leqslant \Omega_j/2$ for a solution to exist. A further restriction of M_j to the unitary group, $M_j \subset \mathrm{U}(2)$ has $M_j{}^\dagger M_j = I$ and

$$\mathbf{G}_{j-1}{}^\dagger \mathbf{G}_{j-1} = (M_j\mathbf{F}_j)^\dagger M_j\mathbf{F}_j = \mathbf{F}_j{}^\dagger \mathbf{F}_j \tag{4.18}$$

preserves norm and orthogonality. This corresponds to multiresolution and one may now identify the components of $\mathbf{F}_j$ to be specified in ν_j and that of $\mathbf{G}_{j-1}$ in orthonormal subspaces $\nu_{j-1} \oplus w_{j-1}$. The sequences $\{m^0{}_k\}, \{m^1{}_k\}$ are now the scaling and wavelet masking coefficients and the components of $\mathbf{G}_{j-1}$ the scaling and wavelet coefficients.

Note: To simplify notation only the $j=0$ case will be considered in the rest of the paper and the subscript S_j will be dropped.

The next proposition derives exact relations between scaling and wavelet filters from the general structure of U(2) matrices.

Proposition 4.2: Let $M_U \in \mathrm{U}(2)$ *and* $M_S \in \mathrm{SU}(2)$ *with* $M_U = e^{i\theta} M_S$, $\theta \in R$, *and where* U(2) *and* SU(2) *are the unitary and the special unitary groups. Then the elements* $\hat{m}^0(\omega), \hat{m}^1(\omega)$ *of* $M_j \in \mathrm{U}(2)$ *are uniquely related as*

$$\hat{m}^0(\omega) = e^{i2\theta} \hat{m}^{1*}(\omega - \pi), \tag{4.19}$$

$$\hat{m}^1(\omega) = -e^{i2\theta} \hat{m}^{0*}(\omega - \pi), \tag{4.20}$$

obtaining the unitarity relation

$$2^{-1}(|\hat{m}^0(\omega)|^2 + |\hat{m}^0(\omega - \pi)|^2) = 1. \tag{4.21}$$

Proof: A general representation for $M_S \in \mathrm{SU}(2)$ is given by

$$M_S = \begin{pmatrix} \xi & \eta \\ -\eta^* & \xi^* \end{pmatrix}, \tag{4.22}$$

where $\eta, \xi \in C$. The requirement $\det M_S = |\eta|^2 + |\xi|^2 = 1$ reduces the number of independent variables to three. A subset of the unitary group $M_U \subset \mathrm{U}(2)$ that represent M_j may be parametrized in ω and equated to M_j,

$$e^{i\theta(\omega)} \begin{pmatrix} \xi(\omega) & \eta(\omega) \\ -\eta^*(\omega) & \xi^*(\omega) \end{pmatrix} = 2^{-1/2} \begin{pmatrix} \hat{m}^0(\omega) & \hat{m}^0(\omega - \pi) \\ \hat{m}^1(\omega) & \hat{m}^1(\omega - \pi) \end{pmatrix}. \tag{4.23}$$

Comparing ξ, ξ^* with corresponding elements on the right-hand side proves Eq. (4.19), and a similar comparison of η, η^* proves Eq. (4.20). These relations together with $\det M_S = 1$ obtain the unitarity relation. □

Corollary: For $M_S \in M_j \subset \mathrm{SU}(2)$, η, ξ are related as

$$\eta(\omega) = \xi(\omega - \pi), \tag{4.24}$$

obtained by comparing the top rows of Eq. (4.23).

D. Euler angle representation of $M_j \subset$ SU(2)

A useful representation of $M_S \in \mathrm{SU}(2)$ is in the three Euler angles of rigid body rotations given by

$$M_S{}^E = \begin{pmatrix} e^{i(\alpha+\gamma)/2} \cos(\beta/2) & ie^{-i(\alpha-\gamma)/2} \sin(\beta/2) \\ ie^{-i(\alpha-\gamma)/2} \sin(\beta/2) & e^{-i(\alpha+\gamma)/2} \cos(\beta/2) \end{pmatrix} \tag{4.25}$$

and exhibits 4π periodicity in the Euler angles. This corresponds to a 2:1 homomorphism with the rotation group SO(3), with two group elements $\pm M_S \in \mathrm{SU}(2)$ mapping onto a single group element in SO(3). Substituting into the left-hand side of Eq. (4.23) and also including an ω dependency for the Euler angles $(\alpha(\omega), \beta(\omega), \gamma(\omega))$ to indicate the subset of SU(2) matrices representing M_j yields the following relations:

$$2^{-1/2} \hat{m}^0(\omega) = e^{i(\alpha(\omega)+\gamma(\omega))/2} \cos(\beta(\omega)/2), \tag{4.26}$$

$$2^{-1/2}\hat{m}^1(\omega)=e^{-i(\alpha(\omega)-\gamma(\omega)-\pi)/2}\sin(\beta(\omega)/2). \tag{4.27}$$

This form permits a mapping of the SU(2) filters onto orbits on the SO(3) manifold. The SO(3) manifold consists of all points within a sphere of radius π,[20] with each point mapping to a specific Euler set (α, β, γ). The frequency dependent Euler set $(\alpha(\omega), \beta(\omega), \gamma(\omega))$ representing a particular filter therefore traces a continuous orbit on the SO(3) group manifold as a function of ω.

In the following proposition, relations between Euler angles at ω and $\omega-\pi$ are given, that obtains a heuristic derivation of the Haar filter.

Proposition 4.3: The Euler angles (α,β,γ) *at* ω *and* $\omega-\pi$ *are related as*

$$\alpha(\omega-\pi)=\alpha(\omega)+\pi,\quad \beta(\omega-\pi)=\beta(\omega)-\pi,$$
$$\gamma(\omega-\pi)=-\gamma(\omega). \tag{4.28}$$

Proof: Since $\hat{m}^0(\omega-\pi)=-\hat{m}^{1*}(\omega)$ from Eq. (4.20), then from Eqs. (4.26) and (4.27),

$$e^{i(\alpha(\omega-\pi)+\gamma(\omega-\pi))/2}\cos(\beta(\omega-\pi)/2)=e^{i(\alpha(\omega)-\gamma(\omega)+\pi))/2}\sin(\beta(\omega)/2) \tag{4.29}$$

and the results are deduced by comparing the Euler angles on both sides. □

E. Haar filter

One particular guess for the Euler angles that checks with Eqs. (4.28) is

$$\alpha(\omega)=-\omega,\quad \beta(\omega)=\omega,\quad \gamma(\omega)=0, \tag{4.30}$$

and results in the Haar filter matrix M_H [substituting in Eq. (4.25)],

$$M_H=\begin{pmatrix} e^{-i\omega/2}\cos(\omega/2) & ie^{-i\omega/2}\sin(\omega/2) \\ ie^{i\omega/2}\sin(\omega/2) & e^{i\omega/2}\cos(\omega/2)\end{pmatrix}. \tag{4.31}$$

Therefore one may infer

$$\hat{m}^0(\omega)=2^{1/2}e^{-i\omega/2}\cos(\omega/2)=2^{-1/2}(1+e^{-i\omega}) \tag{4.32}$$

and the masking coefficients for $\hat{m}^0(\omega)=\Sigma_n m_n e^{-i\omega n}$ are read off as $\{m^0{}_k\}=\{2^{-1/2},2^{-1/2}\}$. There also exists a second possibility $(\alpha(\omega),\beta(\omega),\gamma(\omega))\equiv(\omega+2\pi,\omega,0)$ and this corresponds to the other 2:1 map giving $-M_H$ and is typical for 2:1 homomorphism.

V. ORTHONORMAL FILTERS

It is shown that the underlying group structure for the $\{DN\}$ collection of orthonormal filters is SU(2). Using the Euler angle representation for SU(2) the filters are then expressed in terms of the three Euler angles. Also a proof relating $D\infty$ to Shannon filter is established.

A. SU(2) structure of $\{DN\}$ filters

A systematic construction of orthonormal scaling filters assumes the following unitarity and admissibility relations:

$$|\hat{m}^0(\omega)|^2+|\hat{m}^0(\omega-\pi)|^2=2, \tag{5.1}$$

$$\hat{m}^0(0)=2^{1/2}. \tag{5.2}$$

The pioneering work by Daubechies[1] paved the way for several construction procedures, and the version by Strichartz[21] will be followed here. In this section it will be shown that the relevant group structure for the $\{DN\}$ filters is SU(2)⊂U(2).

The Strichartz method consists in the binomial expansion of

$$[\cos^2(\omega/2)+\sin^2(\omega/2)]^{2N-1}=1 \tag{5.3}$$

written in the form of the unitarity relation [Eq. (5.1)]

$$Q_1(\omega)+Q_2(\omega)=1, \tag{5.4}$$

where

$$x=\cos(\omega/2). \tag{5.5}$$

Q_1, Q_2 are given by

$$Q_1(\omega)=2^{-1}|\hat{m}^0(\omega)|^2=x^{2N}\sum_{k=0}^{N-1}\binom{2N-1}{k}x^{2N-2-2k}(1-x^2)^k, \tag{5.6}$$

$$Q_2(\omega)=Q_1(\omega-\pi)=2^{-1}|\hat{m}^0(\omega-\pi)|^2=(1-x^2)^N\sum_{k=0}^{N-1}\binom{2N-1}{k}(1-x^2)^{N-1-k}x^{2k}. \tag{5.7}$$

Further, using the Riesz lemma, Q_1 and Q_2 are expressed as polynomials in $Z=e^{-i\omega}$,[1]

$$Q_1(\omega)=|x^N e^{i\omega(N-1)/2}R_{N-1}(Z)|^2, \tag{5.8}$$

$$Q_2(\omega)=|(1-x^2)^{N/2}e^{i(\omega-\pi)(N-1)/2}R_{N-1}(-Z)|^2, \tag{5.9}$$

with

$$R_{N-1}(Z)=\sum_{n=0}^{N-1}b_n e^{-i\omega n}, \quad b_n\in R^1, \tag{5.10}$$

and

$$\sum_{n=0}^{N-1}b_n=1. \tag{5.11}$$

The unitarity relation therefore takes the form

$$x^{2N}|e^{i\omega(N-1)/2}R_{N-1}(Z)|^2+(1-x^2)^N|e^{i(\omega-\pi)(N-1)/2}R_{N-1}(-Z)|^2=1 \tag{5.12}$$

and this clearly indicates that the ''regularity'' term x^{2N} cannot be factored out from Q_1 (or from $\hat{m}_0$), while still maintaining the unitarity relation, thereby implying regularity to be an intrinsic feature of multiresolution.

Substituting for $x=\cos(\omega/2)$ in Eqs. (5.8) and (5.9), Q_1 and Q_2 take the form

$$Q_1(\omega)=|e^{i\omega(2N-1)/2}P_{2N-1}(Z)|^2, \tag{5.13}$$

$$Q_2(\omega)=|e^{i(\omega-\pi)(2N-1)/2}P_{2N-1}(-Z)|^2, \tag{5.14}$$

where

$$P_{2N-1}(Z)=2^{-N}(1+e^{-i\omega})^N\sum_{n=0}^{N-1} b_n e^{-i\omega n},\quad b_n\in R^1 \tag{5.15}$$

is a $(2N-1)$th polynomial in Z. From Eqs. (5.6), (5.13), and (4.23) relations between the filter, the binomial expansion series, and the variables of unitary matrices is obtained,

$$2^{-1/2}\hat{m}^0(\omega)=e^{i\omega(2N-1)/2}P_{2N-1}(Z)=e^{i\theta(\omega)}\xi(\omega) \tag{5.16}$$

and on comparing the second equality,

$$\theta(\omega)=\omega(2N-1)/2,\quad \xi(\omega)=P_{2N-1}(Z). \tag{5.17}$$

The filter $\hat{m}^0(\omega)$ represents an orthonormal scaling filter and its companion wavelet filter is obtained from $\hat{m}^1(\omega)=-e^{i2\theta}\hat{m}^{0*}(\omega-\pi)$, yielding

$$2^{-1/2}\hat{m}^1(\omega)=-e^{i(\omega+\pi)(2N-1)/2}P^*_{2N-1}(-Z). \tag{5.18}$$

Together, these filters define a unitary filter matrix,

$$M_j=e^{i\omega(2N-1)/2}\begin{pmatrix} P_{2N-1}(Z) & ie^{-i\pi N}P_{2N-1}(-Z)\\ ie^{i\pi N}P^*_{2N-1}(-Z) & P^*_{2N-1}(Z)\end{pmatrix}. \tag{5.19}$$

A comparison with the unitary matrix in the form $M_U=e^{i\theta}M_S$ obtains $\theta=\omega(2N-1)/2\in R^1$ and M_S given by the matrix in Eq. (5.19), with $\xi(\omega)=P_{2N-1}(Z)$ and $\eta(\omega)=ie^{-i\pi N}P_{2N-1}(-Z)$.

By definition the $\{DN\}$ filters are

$$2^{-1/2}\hat{m}^0(\omega)=P_{2N-1}(Z), \tag{5.20}$$

$$2^{-1/2}\hat{m}^1(\omega)=-P^*_{2N-1}(-Z), \tag{5.21}$$

where the wavelet filter is obtained from $\hat{m}^1(\omega)=-\hat{m}^{0*}(\omega-\pi)$ and compares with M_S in Eq. (5.19) except for a constant phase factor. This shows that a conventional $\{DN\}$ filter has only an underlying SU(2) group structure. These filters can therefore be expressed entirely in the Euler angles.

B. Euler angle representation of $\{DN\}$ filters

SU(2) is 2:1 homomorphic to the rotation group SO(3) [there are two group elements, $\pm M_S\in \mathrm{SU}(2)$ that map to a single element in SO(3)]. A parametrization of SU(2) in terms of the rotation angles, in particular, the three Euler angles, provides a graphic portrayal of the $\{DN\}$ filters as a collection of orbits on the SO(3) manifold, with the members related via rotations in three-dimensional space. The transformation into the Euler representation may be accomplished by writing

$$P(Z)=e^{i\phi_1(\omega)/2}|P(Z)|, \tag{5.22}$$

$$P(-Z)=e^{i\phi_2(\omega)/2}|P(-Z)|, \tag{5.23}$$

and therefore Eqs. (5.20) and (5.21) become

$$2^{-1/2}\hat{m}^0(\omega)=P_{2N-1}(Z)=e^{i\phi_1(\omega)/2}|P_{2N-1}(Z)|, \tag{5.24}$$

$$2^{-1/2}\hat{m}^1(\omega)=-P^*_{2N-1}(-Z)=e^{-i(\phi_2(\omega)-2\pi)/2}|P_{2N-1}(-Z)|. \tag{5.25}$$

On comparing with Eqs. (4.26) and (4.27)

$$e^{i(\alpha(\omega)+\gamma(\omega))/2}\cos(\beta(\omega)/2)=e^{i\phi_1(\omega)/2}|P_{2N-1}(Z)|, \tag{5.26}$$

$$e^{-i(\alpha(\omega)-\gamma(\omega)-\pi)/2}\sin(\beta(\omega)/2)=e^{-i(\phi_2(\omega)-2\pi)/2}|P_{2N-1}(-Z)|. \tag{5.27}$$

One may then deduce from these equations,

$$\phi_1(\omega)=\alpha(\omega)+\gamma(\omega), \tag{5.28}$$

$$\phi_2(\omega)=\alpha(\omega)-\gamma(\omega)+\pi, \tag{5.29}$$

$$\cos(\beta(\omega)/2)=|P_{2N-1}(Z)|, \tag{5.30}$$

and on solving for α, γ, we obtain

$$\alpha(\omega)=(\phi_1(\omega)+\phi_2(\omega)-\pi)/2, \tag{5.31}$$

$$\gamma(\omega)=(\phi_1(\omega)-\phi_2(\omega)+\pi)/2. \tag{5.32}$$

The transformation is particularly simple for $N=1$ when Eq. (5.15) gives

$$P_1(Z)=2^{-1}(1+e^{-i\omega})=e^{-i\omega/2}\cos(\omega/2), \tag{5.33}$$

$$P_1(-Z)=2^{-1}(1-e^{-i\omega})=e^{-i(\omega-\pi)/2}\sin(\omega/2) \tag{5.34}$$

and comparing with Eqs. (5.22) and (5.23) one obtains $\phi_1(\omega)=-\omega$, $\phi_2(\omega)=-\omega+\pi$ and $|P_1(Z)|=\cos(\omega/2)$. Substituting in Eqs. (5.30)–(5.32),

$$\alpha(\omega)=-\omega,\quad \beta(\omega)=\omega,\quad \gamma(\omega)=0, \tag{5.35}$$

to be compared with Eq. (4.30) obtained heuristically for the Haar filter.

C. Shannon filter: $D\infty$

The $\{DN\}$ collection is a denumerable set of compactly supported orthonormal filters with increasing support. At one extreme is the Haar filter $(D1)$ with minimal support, while in the $N\to\infty$ limit, as will be shown in this section, $D\infty$ corresponds to the Shannon filter with noncompact support but having well-defined bandlimiting properties. An indication toward this limiting trend can be seen in Fig. 2, a plot for $|\hat{m}^0(\omega)|$ vs ω for $N=1,2,10$; the figure also revealing a qualitative measure for aliasing, dependent on the width of the filter, indicating maximum aliasing for Haar filter, and totally aliasing free for Shannon filter. In the following proposition it is shown that the Shannon filter corresponds to $D\infty$.

Proposition 5.1: $D\infty$ is equivalent to the Shannon filter as defined by

$$|\hat{m}^0(\omega)|=\begin{cases}1, & 0\leqslant|\omega|<\dfrac{\pi}{2}\\[2ex] 0, & \dfrac{\pi}{2}\leqslant|\omega|<\pi.\end{cases} \tag{5.36}$$

Proof: Since from Eqs. (5.24) and (5.30)

$$2^{-1/2}|\hat{m}^0(\omega)|=|P_{2N-1}(Z)|=|\cos(\beta(\omega)/2)|, \tag{5.37}$$

the following equivalent statement holds true for $N\to\infty$,

$$\beta(\omega)\to 0,\quad 0\leqslant|\omega|<\pi/2, \tag{5.38}$$

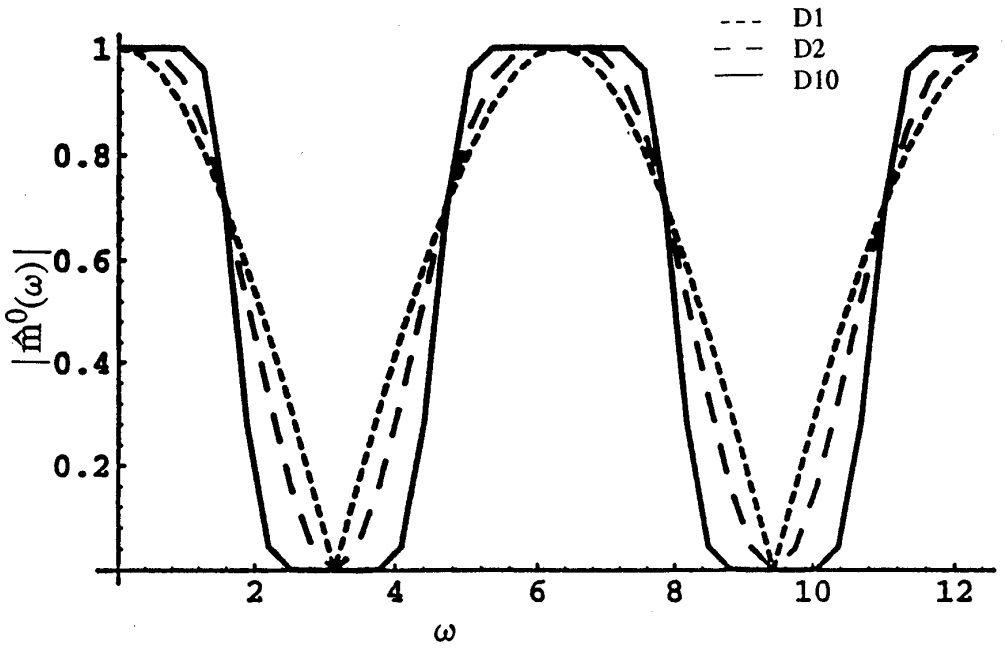

FIG. 2. The 2π periodic filter $|\hat{m}^0(\omega)|$ vs ω for $D1$, $D2$, $D10$. The first zero is at $\omega=\pi$. In 2:1 downsampling this figure is overlayed with one shifted by π, producing aliasing from overlapping regions that vanishes in the limiting case $D\infty$.

$$\beta(\omega)\to\pi, \quad \pi/2\leqslant|\omega|<\pi \tag{5.39}$$

and is proved using the De Moivre–Laplace theorem. The theorem states that the binomial function $f_M(k)$ defined by

$$f_M(k)=\binom{M}{k}q^k p^{M-k}, \tag{5.40}$$

p,q positive, $p+q=1$, $(M,k\in Z_+)$, approaches a normal or Gaussian distribution $g_M(k)$ in the limit $Mpq\gg 1$, centered at mean Mq and width $(Mpq)^{1/2}$, that is,

$$f_M(k)\to g_M(k), \quad Mpq\gg 1, \tag{5.41}$$

where

$$g_M(k)=(2\pi Mpq)^{-1/2}e^{-(k-Mq)^2/(2Mpq)}. \tag{5.42}$$

To prove Eq. (5.38), let

$$p=\cos^2(\omega/2), \quad q=\sin^2(\omega/2), \tag{5.43}$$

then for $0\leqslant|\omega|<\pi/2$, one has $1\leqslant p<\frac{1}{2}$ and $0\leqslant q<\frac{1}{2}$. From Eqs. (5.6) and (5.37),

$$\cos^2(\beta(\omega)/2)=Q_1(\omega)=\sum_{k=0}^{N-1}\binom{2N-1}{k}p^{2N-1-k}q^k=\sum_{k=0}^{N-1}f_M(k), \tag{5.44}$$

where $M=2N-1$. In the limit $N\to\infty$, $Mpq\gg 1$, the stated theorem has $f_M(k)\to g_M(k)$ and also replacing the sum by an integration over k, Eq. (5.44) becomes

$$I=\cos^2(\beta(\omega)/2)\approx\int_0^N g_M(k)dk. \tag{5.45}$$

Now for $0\leqslant q<\frac{1}{2}$ one has $0\ll Mq\ll N$, showing the upper limit to be well beyond the mean of the Gaussian $g_M(k)$. Similarly with the width $(Mpq)^{1/2}\ll Mq$, one may conclude that the significant contribution from the integrand lies well within the limits of integration [a rough estimate requires $M^{1/2}(\frac{1}{2}-q)>3(pq)^{1/2}$], whence it follows $I\approx 1$ and therefore in the limit $N\to\infty$,

$$\cos(\beta(\omega)/2)\to 1 \tag{5.46}$$

or

$$\beta(\omega)\to 0,\quad 0\leqslant|\omega|<\pi/2 \tag{5.47}$$

proves Eq. (5.38).

A similar argument holds for Eq. (5.39) with $\pi/2\leqslant|\omega|<\pi$ and $\frac{1}{2}\leqslant p<0$, $\frac{1}{2}\leqslant q<1$. Now $Mq\gg N$ and with $(Mpq)^{1/2}\ll Mq-N$ the integrand is well outside the integration limits and hence $I\approx 0$ and in the limit $N\to\infty$,

$$\beta(\omega)\to\pi,\quad \pi/2\leqslant|\omega|<\pi \tag{5.48}$$

completes the proof. □

VI. SCALED FILTERS

Filters satisfying both unitarity and admissibility relations do not necessarily produce orthonormal basis.[17–19] Dlin and Madych[19] gave nontrivial examples of such filters. An additional sufficiency relation to ensure orthonormality of the scaling function ϕ is stated as

$$|\hat{m}^0(\omega)|\neq 0,\quad 0\leqslant|\omega|\leqslant\pi/3 \tag{6.1}$$

and follows from a corollary to Cohen's theorem[17] (see also in Ref. 6, Corollary 6.3.2). Lawton's criterion,[18] requires nondegeneracy on the eigenvalue 1 of a $(2K-1)\times(2K-1)$ matrix $\mathbf{A}$ defined by

$$A_{lk}=\sum_{n=0}^{K} m(n)\overline{m}(k-2l+n),\quad -K+1\leqslant l,\ k\leqslant K-1,$$

where $m(n)=0$ if $n<0$ or $n>K$. That both Cohen's and Lawton's conditions are equivalent (see Ref. 6) allows one to focus on Eq. (6.1). A rather elegant new proof for this criterion is given here in the Euler angle representation, but first, the cause for lack of sufficiency of the unitarity relation is examined in the next two propositions.

Proposition 6.1: Let

$$\hat{m}^0(\omega)=\sum_n m(n)e^{-i\omega n} \tag{6.2}$$

represent an orthonormal filter for V_0 with periodicity 2π. Then their scaled versions

$$\hat{m}^0{}_N(\omega)=\sum_n m(n)e^{-i\omega nN}, \quad N \text{ odd integer} \tag{6.3}$$

also satisfy the unitarity relation but with a reduced periodicity $2\pi/N$ *that preclude them from generating an orthonormal basis for* V_0.

Proof: The unitarity relation gives

$$|\hat{m}^0(N\omega)|^2+|\hat{m}^0(N\omega-\pi)|^2=2$$

and since $\hat{m}^0(\omega)$ is 2π periodic in ω, for N odd one may write,

$$|\hat{m}^0(N\omega)|^2+|\hat{m}^0(N(\omega-\pi))|^2=2. \tag{6.4}$$

Scaled filters $\hat{m}^0{}_N(\omega)$ are defined by (note that this is not dyadic scaling, but will be termed pathological scaling)

$$\hat{m}^0{}_N(\omega)=\hat{m}^0(N\omega)=\sum_n m(n)e^{-i\omega nN}=\sum_n m_N(n)e^{-i\omega n} \tag{6.5}$$

with the new coefficients defined by $m_N(Nn)=m(n)$, $n\in Z$, and 0 otherwise. Substituting in Eq. (6.4) it is seen that the scaled filters satisfy the unitarity relation (for N even the equation does not reduce to this form),

$$|\hat{m}^0{}_N(\omega)|^2+|\hat{m}^0{}_N(\omega-\pi)|^2=2 \tag{6.6}$$

but with a reduced periodicity $2\pi/N$ that preclude them from being an orthonormal filter for V_0. □

Next it is shown that the orthonormality condition

$$\sum_k |\hat{\phi}(\omega-2\pi k)|^2=1 \tag{6.7}$$

derives a series of dyadically scaled relations that are not necessarily unitarity relations.

Proposition 6.2: If $\{\phi(x-n);n\in Z\}$ *is an orthonormal basis for* V_0 *such that* $\langle\phi(0),\phi(n)\rangle=\delta_{n,0}$ *and* $\hat{m}^0(\omega)$ *an orthonormal filter for* V_0, *then the relations*

$$\sum_k |\hat{\phi}(\omega-2^{j+1}\pi k)|^2=2^{-j}, \quad \forall j\in Z \tag{6.8}$$

are satisfied, and they derive the relation

$$|\hat{m}_j^0(\omega)|^2+|\hat{m}_j^0(\omega-\pi)|^2=2, \quad j\geqslant 0 \tag{6.9}$$

with $\hat{m}_j^0(\omega)$ *denoting dyadic scaling. Since for* $j\neq 0$ *these do not represent unitarity relations, they cannot be considered as a sufficient condition on an orthonormal filter.*

Proof: To prove Eq. (6.8) define

$$|\hat{\phi}(2^j(\omega'-2\pi k))|^2=\hat{f}(\omega'-2\pi k), \tag{6.10}$$

where $\omega'=2^{-j}\omega$. Then, since

$$\mathscr{F}^{-1}[\hat{\phi}(2^j(\omega'-2\pi k))]=2^{-j}e^{-i2\pi kx}\phi(2^{-j}x), \tag{6.11}$$

the inverse Fourier transform of Eq. (6.10) is obtained as

$$f(n)=2^{-2j}e^{-i2\pi kn}\int_{-\infty}^{\infty}\phi(2^{-j}x')\phi^*(2^{-j}(n-x'))dx'=2^{-j}e^{-i2\pi kn}\delta_{n,0}, \tag{6.12}$$

where the last equality follows from the orthonormality $\langle\phi(0),\phi(n)\rangle=\delta_{n,0}$. Application of the Poisson summation formula,

$$\sum_k \hat{f}(\omega'-2k\pi)=\sum_n f(n)e^{i\omega' n}=2^{-j}, \tag{6.13}$$

proves Eq. (6.8).

Next, as in the standard method, using the dilation equation to write

$$\hat{\phi}(2\omega-2^{j+1}\pi k)=\hat{m}^0(\omega-2^j\pi k)\hat{\phi}(\omega-2^j\pi k)$$

and expanding the following summation in even and odd integers,

$$2^{-j}=\sum_k |\hat{\phi}(2\omega-2^{j+1}\pi k)|^2=2^{-1}(|\hat{m}^0(\omega)|^2+|\hat{m}^0(\omega-2^j\pi)|^2)\sum_k |\hat{\phi}(\omega-2^{j+1}\pi k)|^2, \tag{6.14}$$

obtains the following relations for $j\leqslant 0$:

$$|\hat{m}^0(\omega)|^2+|\hat{m}^0(\omega-2^j\pi)|^2=2,$$

which implies Eq. (6.9) using the transformation $\omega\rightarrow 2^j\omega$. For $j\neq 0$ these are not unitarity relations, the requirement then being

$$|\hat{m}^0{}_j(\omega)|^2+|\hat{m}^0{}_j(\omega-2^j\pi)|^2=2. \tag{6.15}$$

□

Both foregoing propositions indicate the possibility of obtaining scaled versions of a true orthonormal filter if only the unitarity relation and the condition on a scaling filter $\hat{m}_0(0)=2^{1/2}$ are employed for construction. Both emphasize exact 2π periodicity for orthonormal filters. For dyadic scaling, the $j>0$ filters are eliminated by the admissibility condition on wavelets. For pathological scaling, a well-known sufficiency relation must be satisfied. The following proposition gives the necessary and sufficient conditions on orthonormal filters.

Proposition 6.3: The necessary and sufficient conditions on a filter $\hat{m}^0(\omega)$ *to represent an orthonormal filter for* V_0 *are*:

(i) *the unitarity relation*: $|\hat{m}^0(\omega)|^2+|\hat{m}^0(\omega-\pi)|^2=2$,
(ii) *the scaling filter condition*: $\hat{m}^0(0)=2^{1/2}$,
(iii) *the wavelet admissibility condition*: $\hat{m}^0(\pi)=0$,
(iv) *a sufficiency relation eliminating pathological scaling*: $\hat{m}^0(\omega)\neq 0;\ 0\leqslant|\omega|\leqslant\pi/3$.

Proof: To prove the sufficiency relation, first it will be shown that for a valid filter $\hat{m}^0(\omega)$ there can be no zeros in $0\leqslant|\omega|<\pi$ and next, for the lowest order scaling at $N=3$, the first zero appears at $\omega=\pi/3$. The first condition is easily established in the Euler angle representation, where

$$2^{-1/2}|\hat{m}^0(\omega)|=|\cos(\beta(\omega)/2)| \tag{6.16}$$

evidently has no zero in the interval $0\leqslant|\beta(\omega)|<\pi$ because the right-hand side does not. A valid filter has a 1:1 mapping between ω and $\beta(\omega)$ since $|\cos(\beta(\omega)/2)|$ is 2π periodic in β, and $|\hat{m}^0(\omega)|$, 2π periodic in ω, giving the simultaneous intervals

$$0\leqslant|\beta(\omega)|<\pi,\quad 0\leqslant|\omega|<\pi \tag{6.17}$$

and so it may be concluded that there is no zero for $\hat{m}^0(\omega)$ as well in $0 \leqslant |\omega| < \pi$. The first zero appears at $\omega = \pi$ as can also be seen from

$$\hat{m}^0(\pi) = \sum_{n \text{ even}} m(n) - \sum_{n \text{ odd}} m(n) = -\hat{m}^{1*}(0) = 0, \tag{6.18}$$

where the last equality follows from the required admissibility condition on a wavelet filter. This less restrictive interval $[-\pi, \pi]$ was first given by Mallat (see also Fig. 2).

Next, for the $N=3$ scaled version given by Eq. (6.5) one has at $\omega = \pi/3$,

$$\hat{m}^0{}_3(\pi/3) = \hat{m}^0(3(\pi/3)) = 0, \tag{6.19}$$

showing the first zero to be at $\omega = \pi/3$ for the lowest scaling. The nonzero interval for β remains invariant for the cosine function, and one has

$$0 \leqslant |\beta(\omega)| < \pi, \quad 0 \leqslant |\omega| < \pi/3, \tag{6.20}$$

implying multiple mapping from $\beta(\omega)$ to ω in the interval $0 \leqslant |\omega| < \pi$. □

VII. SUMMARY AND OUTLOOK

This paper formulates wavelet theory in terms of a family of dyadically scaled lattice functions S_j and its algebra. A set of linear equations representing convolution and sampling determines the structure of a 2×2 filter matrix $M_j \subset \mathrm{GL}(2,C)$ defined via

$$\mathbf{G}_{j-1} = M_j \mathbf{F}_j,$$

where $\mathbf{F}_j, \mathbf{G}_{j-1}$ are vectors with components defined on S_j and S_{j-1} lattices, respectively. When $M_j \subset \mathrm{U}(2)$, the projections from S_j to S_{j-1} preserve norm and orthogonality and correspond to the multiresolution decomposition $\nu_j \rightarrow \nu_{j-1} \oplus w_{j-1}$, the components of M_j then constituting wavelet and scaling filters. This allows exact relations between the filters to be deduced from the general structure of a unitarity matrix.

The underlying group for the $\{DN\}$ filters is shown to be SU(2), with its representation in the Euler angles revealing several interesting properties. These include, a mapping to a set of orbits on the SO(3) manifold with members related to each other by three-dimensional rotations; an equivalence of $D\infty$ to Shannon filter; and a simple new proof for a sufficiency relation ruling out nonorthonormal pathological scaling. The amount of aliasing in multiresolution filters is delineated qualitatively as shown in Fig. 2, with the Haar filter producing maximum aliasing, while the Shannon filter is completely aliasing free.

The formulation for dyadic scaling can be extended directly to (N)th order scaled lattices resulting in $N \times N$ filter matrix $M_j \subset \mathrm{U}(N)$, M_j now defining N orthonormal filters and a multiresolution decomposition $\nu_{j+1} \rightarrow \nu_j \oplus w_j^1 \oplus \cdots \oplus w_j^{N-1}$. As in dyadic scaling it should be possible to relate the various filters, but the complexity increases with increasing N.

ACKNOWLEDGMENTS

This work was supported by AFOSR/NM research Grant No. F49620-95-1-0060, and Missouri University Department of Physics & Astronomy. This support is acknowledged with appreciation. The author also thanks Professor B. DeFacio for encouragement of this work.

[1] I. Daubechies, ''Orthonormal bases of compactly supported wavelets,'' Commun. Pure Appl. Math. **41**, 909–996 (1988).
[2] S. G. Mallat, ''Multiresolution approximations and wavelet orthonormal bases of $L^2(R^1)$,'' Trans. Am. Math. Soc. **315**, 69–88 (1989).

[3] S. G. Mallat, ''A theory for multiresolution signal decomposition: The wavelet representation,'' IEEE Trans. Acoust. Speech Signal Process. **37**, 2091–2110 (1989).
[4] Y. Meyer, *Wavelets and Operators* (Cambridge University Press, Cambridge, 1992); and *Wavelets, Algorithms and Applications* (SIAM, Philadelphia, 1993).
[5] C. K. Chui, *An Introduction to Wavelets* (Academic, San Diego, 1992), Vol. 1.
[6] I. Daubechies, *Ten Lectures in Wavelets* (SIAM, Philadelphia, 1992).
[7] G. Kaiser, *A Friendly Guide to Wavelets* (Birkhauser, Boston, 1994).
[8] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software* (A. K. Peters, Wellesley, MA, 1994).
[9] M. Holschneider, *Wavelets, An Analysis Tool* (Clarendon, Oxford, 1995).
[10] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding* (Prentice–Hall, Englewood Cliffs, NJ, 1995).
[11] G. Strang and T. Nguyen, *Wavelets and Filter Banks* (Cambridge University Press, Wellesley, MA, 1996).
[12] E. Hernandez and G. Weiss, *A First Course on Wavelets* (Chemical Rubber, Boca Raton, FL, 1996).
[13] O. Brander and B. DeFacio, ''The role of filters and the singular value decomposition for the inverse Born approximation,'' Inverse Probl. **2**, 375–393 (1986).
[14] C. E. Shannon, ''A mathematical theory of communication,'' Bell Syst. Tech. J. **27**, 379–423 (1948).
[15] F. O. Huck, C. L. Fales, J. A. McCormick, and S. A. Park, ''Image gathering system design for information and fidelity,'' J. Opt. Soc. Am. A **5**, 285–299 (1988).
[16] S. John, ''Recovery of aliased signals for high resolution digital imaging,'' Proc. SPIE **1385**, 39–47 (1990).
[17] A. Cohen, ''Ondelletes, analyses multiresolutions et filtres miroir en quadrature,'' Ann. Inst. Henri Poincaré, Anal. non lineare **7**, 439–459 (1990).
[18] W. M. Lawton, ''Necessary and sufficient conditions for constructing orthonormal wavelet bases,'' J. Math. Phys. **32**, 57–60 (1991).
[19] F. Dlin and W. R. Madych, ''Scaling functions and sequences associated with orthonormal wavelets,'' Houston J. Math. **21**, 129–148 (1995).
[20] L. I. Schiff, *Quantum Mechanics* (McGraw–Hill, New York, 1968).
[21] R. S. Strichartz, ''Construction of orthonormal wavelets,'' in *Wavelets, Mathematics and Applications*, edited by J. J. Benedetto and M. W. Frazier (Chemical Rubber, Boco Raton, FL, 1994), pp. 23–50.

Spectrum of the Dirac operator on $Gr_2(\mathbb{C}^{m+2})$

Jean-Louis Milhorat[a)]
Département de Mathématiques, Université de Nantes,
UMR CNRS 6629, 44322 Nantes Cedex 03, France

(Received 1 July 1997; accepted for publication 23 July 1997)

The spectrum of the Dirac operator, acting on the quaternion-Kähler spin symmetric space $Gr_2(\mathbb{C}^{m+2})$, is explicitly computed by harmonic analysis methods: in particular ''branching rules'' for irreducible representations of the Lie group $\mathrm{SU}(m+2)$ and its subgroup $\mathrm{S}(\mathrm{U}(m)\times\mathrm{U}(2))$, are given. © *1998 American Institute of Physics.* [S0022-2488(98)02401-3]

I. INTRODUCTION

In the study of the relations between the spectrum of the Dirac operator and the geometry of a compact spin manifold, it is natural to consider symmetric spaces corresponding to the geometry. Indeed, for these manifolds, the spectrum of the Dirac operator can be (theoretically) computed by classical harmonic analysis methods.

In this paper, we compute the Dirac spectrum of $Gr_2(\mathbb{C}^{m+2})$. This is motivated by results of Refs. 1 and 2 about the determination of a lower bound for the square of the first eigenvalue of the Dirac operator, acting on a compact spin quaternion-Kähler manifold with positive scalar curvature. Symmetric spaces with this geometry are known: the only compact quaternion-Kähler symmetric spaces are, besides five exceptional cases, the quaternionic projective space $\mathbb{H}P^m$, the Grassmannians $Gr_2(\mathbb{C}^{m+2})$ and $\widetilde{Gr_4}(\mathbb{R}^{m+4})$.[3] In fact, these spaces are the only known examples of compact quaternion-Kähler manifolds, see Ref. 4. Furthermore, the quaternionic projective space is spin, but the last two Grassmannians are spin only for m even.[5] It is then interesting to study the Dirac operator on these three families of Grassmannians. The spectrum of the Dirac operator on $\mathbb{H}P^m$ was given in Ref. 6. Here, we consider the case of the Grassmannian $Gr_2(\mathbb{C}^{m+2})=G/K$, with $G=\mathrm{SU}(m+2)$ and $K=\mathrm{S}(\mathrm{U}(m)\times\mathrm{U}(2))$.

For symmetric spaces, the spectrum of the Dirac operator D is totally determined by the spectrum of its square. Eigenvalues of D^2 are related to irreducible G-modules V_G, whose decomposition under the action of K, has an irreducible K-module occurring in the corresponding decomposition of the spinor space.[7–10] They have the following form[10]:

$$\lambda = c(V_G) + \frac{S}{8},$$

where $c(V_G)$ is the eigenvalue of the Casimir operator of V_G.

In the following, we give the ''branching rules'' for $G=\mathrm{SU}(m+2)$ and $K=\mathrm{S}(\mathrm{U}(m)\times\mathrm{U}(2))$ (see Proposition 2 below), the decomposition into irreducible K-modules of the spinor space (Proposition 3), which enables us to compute the spectrum:

Proposition 1: *On the compact spin Riemannian symmetric space* $Gr_2(\mathbb{C}^{m+2}) \simeq \mathrm{SU}(m+2)/\mathrm{S}(\mathrm{U}(m)\times\mathrm{U}(2))$, *m even, endowed with the metric induced from the Killing form sign-changed, the spectrum of the square of the Dirac operator is*

(i) If $m=2$,

$$Spec(D^2, Gr_2(\mathbb{C}^4)) = \{\tfrac{1}{8}\lambda_{k,l}\,;\,\tfrac{1}{8}\mu_{k,l}\},$$

[a)]Electronic mail: milhorat@math.univ-nantes.fr

0022-2488/98/39(1)/594/16/$15.00
© 1998 American Institute of Physics

with

$$\lambda_{k,l}=(2k+l+4)^2+(l+2)^2, \quad k,l\in\mathbb{N},$$

$$\mu_{k,l}=(2k+l+2)^2+(l+1)^2-1, \quad k,l\in\mathbb{N},(k,l)\neq(0,0).$$

(ii) If $m\geqslant 4$,

$$Spec(D^2,Gr_2(\mathbb{C}^{m+2}))=\left\{\begin{array}{lll}\frac{1}{2(m+2)}\lambda_{k,l,n,p,q}; & \frac{1}{2(m+2)}\mu_{k,l,n,p,q}; & \frac{1}{2(m+2)}\nu_{k,l,p,q};\\ \frac{1}{2(m+2)}\eta_{k,l,p,q}; & \frac{1}{2(m+2)}\theta_{k,l,p,q}; & \frac{1}{2(m+2)}\zeta_{k,l,p,q}\end{array}\right\},$$

with

$$\begin{aligned}\lambda_{k,l,n,p,q}&=(2k+l+n+m)^2+(l+1)^2+(n-q)^2+2(p+q)(k+l+n)\\&\quad+(p+q)(m-q-2)+pq-5,\\&\quad k,l,n,p,q\in\mathbb{N},s.t.0\leqslant p\leqslant m/2-2,0\leqslant q\leqslant m/2-2,0\leqslant n\leqslant q \quad\text{and}\\&\quad\left\{\begin{array}{l}k\geqslant 1 \quad if\ p\leqslant m/2-3,\\(k,l)\neq(0,0) \quad if\ p=m/2-2 \quad\text{and}\ 1\leqslant n\leqslant q,\\(k,l)\neq(0,0) \quad\text{and}\ (k,l)\neq(0,1) \quad if\ p=q=m/2-2 \quad\text{and}\ n=0,\end{array}\right.\\\mu_{k,l,n,p,q}&=\left(2k+l+n+\frac{3m}{2}-p\right)^2+(l+1)^2+n^2+2(p-q)\left(k+l+\frac{m}{2}\right)\\&\quad+p(m-(p+4))-\frac{m^2}{4}+2m-9,\\&\quad k,l,n,p,q\in\mathbb{N},s.t.\ 0\leqslant p\leqslant m/2-2,0\leqslant q\leqslant p,0\leqslant n\leqslant p-q,\\\nu_{k,l,p,q}&=\left(2k+l+\frac{3m}{2}-p+1\right)^2+(l+1)^2+2(p-q)\left(k+l+\frac{m}{2}+1\right)\\&\quad+p(m-(p+2))-\frac{m^2}{4}+m-2,\\&\quad k,l,p,q\in\mathbb{N},s.t.0\leqslant p\leqslant m/2-1,0\leqslant q\leqslant p,\\\eta_{k,l,p,q}&=\left(2k+l+\frac{3m}{2}-p+1\right)^2+(l+2)^2+2(p-q)\left(k+l+\frac{m}{2}+2\right)\\&\quad+p(m-(p+2))-\frac{m^2}{4}+m-1, \quad k,l,p,q\in\mathbb{N},s.t.\ 0\leqslant p\leqslant m/2-2,0\leqslant q\leqslant p,\\\theta_{k,l,p,q}&=(2k+l+m+2(p+1))(2k+l+m+2)+2q\left(k+l+\frac{m}{2}+p+2\right)+(l+2)^2,\\&k,l,p,q\in\mathbb{N},s.t.0\leqslant p\leqslant m/2-1,0\leqslant q\leqslant m/2+p-1,k\geqslant 1\ \ if\ q\leqslant m/2+p-2,\end{aligned}$$

$$\zeta_{k,l,p,q}=(2k+l+m+2(p+1))(2k+l+m)+2q\left(k+l+\frac{m}{2}+p+1\right)+(l+1)^2-1,$$
$$k,l,p,q\in\mathbb{N}, s.t.\ 0\leqslant p\leqslant m/2-2, 0\leqslant q\leqslant m/2+p,$$
$$k\geqslant 1 \ \ if\ \ q\leqslant m/2+p-1.$$

II. THE GRASSMANNIAN $Gr_2(\mathbb{C}^{m+2})$

We *denote* by $M_{p,q}(\mathbb{C})$ the space of $p\times q$ matrices with complex coefficients.
Let $G=\mathrm{SU}(m+2)=\{A\in M_{m+2}(\mathbb{C}); {}^t\overline{A}A=I_{m+2}; \det A=1\}$ and

$$K=\mathrm{S}(\mathrm{U}(m)\times\mathrm{U}(2))=\left\{A=\begin{pmatrix}B & 0\\ 0 & C\end{pmatrix}; \begin{cases}B\in U(m),\\ C\in U(2),\end{cases} \det A=1\right\}.$$

The quotient space G/K, identified with the Grassmannian $Gr_2(\mathbb{C}^{m+2})$, is a compact simply connected irreducible symmetric space of ''type I.''[11] The symmetric space structure is given by the involutive automorphism $\sigma: G\to G$, $A\to SAS^{-1}$, where $S=\left(\begin{smallmatrix}I_m & 0\\ 0 & -I_2\end{smallmatrix}\right)$, since $K=\{A\in G;\sigma(A)=A\}$. Let $\mathfrak{G}$ (resp. $\mathfrak{K}$) be the Lie algebra of G (respective of K) and let $\mathfrak{G}=\mathfrak{K}\oplus\mathfrak{P}$ be the Lie algebras decomposition induced by the involution σ_*. One has

$$\mathfrak{P}=\{X\in\mathfrak{G};\sigma_*(X)=-X\}=\left\{X_U=\begin{pmatrix}0 & U\\ -{}^t\overline{U} & 0\end{pmatrix}; U\in M_{m,2}(\mathbb{C})\right\}.$$

Restricted to K, the adjoint representation Ad_G of G acts on $\mathfrak{P}\simeq M_{m,2}(\mathbb{C})$ by

$$Ad_G\left(\begin{pmatrix}B & 0\\ 0 & C\end{pmatrix}\right)\cdot U=BU{}^t\overline{C}. \tag{1}$$

Since this representation is irreducible, all G-invariant metrics on G/K are homothetic. We consider the G-invariant metric g induced by the restriction to $\mathfrak{P}$ of the sign-changed Killing form of $\mathfrak{G}$. The symmetric space G/K is Einstein[4] and, with this metric, its scalar curvature is $S=2m$.

Let $(1,i,j,k)$ be the standard basis of $\mathbb{H}$. We identify $\mathbb{H}$ with $\mathbb{C}\oplus j\mathbb{C}$ and $\mathbb{H}^m$ with $\mathbb{C}^m\oplus j\mathbb{C}^m$ ($\mathbb{H}$ acts on the right). In the following, we identify $M_{m,2}(\mathbb{C})\simeq\mathfrak{P}$ with $\mathbb{H}^m\simeq\mathbb{R}^{4m}$ by

$$U=(\xi_1,\xi_2)\in M_{m,2}(\mathbb{C})\to\xi=\xi_1+j\overline{\xi_2}\in\mathbb{C}^m\oplus j\mathbb{C}^m, \tag{2}$$

and *denote* by α the homomorphism $K\to\mathrm{SO}(4m)$; $k\to Ad(k)_{|\mathfrak{P}}$. The group $\alpha(K)$ is a subgroup of $\mathrm{Sp}(m)\mathrm{Sp}(1)$ so the $4m$-dimensional manifold $(G/K,g)$ has a quaternion-Kähler structure.[12,13]

Finally let us recall that the Killing form of $\mathfrak{G}$ is given by $B(X,Y)=2(m+2)\Re[\mathrm{Tr}(XY)]$, so restriction of B to $\mathfrak{P}$ is

$$B(X_U,X_V)=-4(m+2)\Re(\xi_U,\xi_V), \tag{3}$$

where ξ_U (respectively ξ_V) are the elements of $\mathbb{H}^m$ corresponding to X_U (X_V) by the isomorphism given in (2), and where $(,)$ denotes the usual symplectic product on $\mathbb{H}^m$.

III. SPIN STRUCTURE AND DIRAC OPERATOR ON $Gr_2(\mathbb{C}^{m+2})$

The symmetric space $(G/K,g)$ has a spin structure (necessarily unique because it is simply connected) if there is a lift to $Spin(4m)$ of the homomorphism $\alpha: K\to\mathrm{SO}(4m)$, $k\to Ad(k)_{|\mathfrak{P}}$, (see Sec. II), hence if m is even.[5] Henceforward, except in Sec. IV, m is assumed to be even. The

lift of α is *denoted* by $\tilde{\alpha}$. The complex Clifford algebra $\mathbb{C}\ell_{4m}$, admits, up to equivalence, a unique irreducible representation $\rho:\mathbb{C}\ell_{4m}\to Hom_{\mathbb{C}}(\Sigma,\Sigma)$. The induced representation $Spin(4m)\to Hom_{\mathbb{C}}(\Sigma,\Sigma)$ will be also denoted by ρ.

For the definition and properties of the Dirac operator D on a spin symmetric space see Refs. 7–10. Here, all we need to know is the following: the spectrum of the Dirac operator D is $Spec(D,Gr_2(\mathbb{C}^{m+2}))=\{\pm\sqrt{\lambda};\lambda\in Spec(D^2,Gr_2(\mathbb{C}^{m+2}))\}$, with[10]

$$Spec(D^2,Gr_2(\mathbb{C}^{m+2}))=\left\{c_{[\alpha]}+\frac{S}{8};[\alpha]\in[G],m([\alpha]_{|K},\Sigma_i)\geqslant 1\right\}, \tag{4}$$

where (i) $[G]$ is the set of equivalence classes of unitary irreducible representations of G; (ii) $\Sigma=\oplus_{i=1}^{n}\Sigma_i$, the decomposition of Σ into irreducible K-modules under the action of K by the representation $\rho\circ\tilde{\alpha}$; (iii) $m([\alpha]_{|K},\Sigma_i)$ the multiplicity of the K-module Σ_i in the decomposition into irreducible K-modules of a G-module representing $[\alpha]$; (iv) $c_{[\alpha]}$ the eigenvalue of the Casimir operator of the representation $[\alpha]$.

Hence we are lead to state the ''branching rules'' giving the decomposition of an irreducible G-module into irreducible K-modules, give the decomposition (ii), determine the $[\alpha]$'s such that $m([\alpha]_{|K},\Sigma_i)\geqslant 1$, compute the corresponding eigenvalues with the help of Freudenthal's formula.

IV. BRANCHING RULES OF SU($m+2$) UNDER THE ACTION OF S(U(m)×U(2))

For results on representation theory, we refer to Refs. 14–17.

Let $T=\{$ diag. matrix $(e^{i\beta_1},\ldots,e^{i\beta_{m+2}});\beta_1,\ldots,\beta_{m+2}\in\mathbb{R}\}$, the usual maximal torus of U$(m+2)$. Its Lie algebra is $\mathfrak{T}=\{$ diag. matrix $(i\beta_1,\ldots,i\beta_{m+2});\beta_1,\ldots\beta_{m+2}\in\mathbb{R}\}$. We *denote* by $(y_1,\ldots,y_{m+2})$ the basis of the dual $\mathfrak{T}^*$ given by

$$y_k[\text{ diag. matrix }(i\beta_1,\ldots,i\beta_{m+2})]=\beta_k. \tag{5}$$

Let $T'=T\cap \mathrm{SU}(m+2)$. This is a maximal common torus of G and K. Its Lie algebra is $\mathfrak{T}'=\{$ diag. matrix $(i\beta_1,\ldots,i\beta_{m+2});\beta_1,\ldots\beta_{m+2}\in\mathbb{R},\Sigma_{k=1}^{m+2}\beta_k=0\}$. We *denote* by $(x_1,\ldots,x_{m+1})$ the basis of the dual $\mathfrak{T}'^*$ given by

$$x_k[\text{ diag. matrix }(i\beta_1,\ldots,i\beta_{m+2})]=\beta_k. \tag{6}$$

A weight $\mu\in i\mathfrak{T}'^*$ such that $\mu=\Sigma_{k=1}^{m+1}\lambda_k\hat{x}_k$, in the basis $(\hat{x}_k\equiv ix_k)_{k=1,\ldots,m+1}$, is *denoted* by

$$\mu=(\lambda_1,\ldots,\lambda_{m+1}). \tag{7}$$

We can also express μ in the basis $(\hat{y}_k\equiv iy_k)_{k=1,\ldots,m+2}$, of $i\mathfrak{T}^*$: $\mu=\Sigma_{k=1}^{m+2}\mu_k\hat{y}_k$; $\Sigma_{k=1}^{m+2}\mu_k=0$. In this case we *denote* μ by

$$\mu=[\mu_1,\ldots,\mu_{m+2}]. \tag{8}$$

The correspondence between the two expressions (7), (8) is

$$\lambda_i=\mu_i-\mu_{m+2}\Leftrightarrow\begin{cases}\mu_i=\lambda_i+\lambda_0\\ \mu_{m+2}=\lambda_0:=-\dfrac{1}{m+2}\displaystyle\sum_{i=1}^{m+1}\lambda_i,\end{cases}\qquad i=1,\ldots,m+1. \tag{9}$$

In the following, we consider the invariant scalar product on $\mathfrak{T}'$ given by $\langle X,Y\rangle = -\Re[\mathrm{Tr}(XY)] = -1/2(m+2)B(X,Y), \forall X,Y \in \mathfrak{T}'$, where B is the Killing form of $\mathfrak{G}$ (see Sec. II). We also *denote* by $\langle,\rangle$ its extension to $i\mathfrak{T}'^*$. Then for $\mu=[\mu_1,\ldots,\mu_{m+2}]$ and $\mu' = [\mu'_1,\ldots,\mu'_{m+2}] \in i\mathfrak{T}'^*$, we get

$$\langle \mu,\mu'\rangle = \sum_{i=1}^{m+2} \mu_i \mu'_i . \tag{10}$$

A. Determination of [*G*]

As G is simply connected, every class in $[G]$ corresponds to a dominant weight and conversely.

Lemma 1: *A weight of* G, $\beta=(k_1,\ldots,k_{m+1})=[p_1,\ldots,p_{m+2}]$ *is dominant iff*

$$\begin{cases} k_i \in \mathbb{N}, i=1,\ldots,m+1, \\ k_1 \geqslant k_2 \geqslant \cdots \geqslant k_{m+1}, \end{cases} \Leftrightarrow \begin{cases} (p_i - p_{i+1}) \in \mathbb{N}, \\ i=1,\ldots,m+1 . \end{cases}$$

Proof: We choose as positive roots $\hat{x}_i - \hat{x}_j, 1 \leqslant i < j \leqslant m+1$; $\hat{x}_1+\hat{x}_2+\cdots+2\,\hat{x}_i+\cdots+\hat{x}_{m+1}, i=1,\ldots,m+1$, such that the simple roots are $\theta_k = \hat{x}_k - \hat{x}_{k+1}; 1 \leqslant k \leqslant m$; $\theta_{m+1} = \hat{x}_1+\hat{x}_2+\cdots+2\,\hat{x}_{m+1}$. The conditions for β to be a dominant weight are $2\langle\beta,\theta_k\rangle/\langle\theta_k,\theta_k\rangle \in \mathbb{N}$, $k=1,\ldots,m+1$. Hence the result. ■

The Weyl group W_G of G is generated by the reflections across $\theta_k^{\perp}$. Hence it acts on $i\mathfrak{T}'^*$ by

$$\mu=[\mu_1,\ldots,\mu_{m+2}] \rightarrow [\mu_{\sigma(1)},\ldots,\mu_{\sigma(m+2)}], \sigma \in \mathfrak{S}_{m+2}. \tag{11}$$

B. Determination of [K]

We choose as positive roots for K, $\hat{x}_i - \hat{x}_j, 1 \leqslant i < j \leqslant m$; $\hat{x}_1+\hat{x}_2+\cdots+2\,\hat{x}_{m+1}$, such that the simple roots are

$$\theta'_k = \hat{x}_k - \hat{x}_{k+1}; 1 \leqslant k \leqslant m-1, \theta'_m = \hat{x}_1+\hat{x}_2+\cdots+2\,\hat{x}_{m+1}. \tag{12}$$

The Weyl group W_K of K is generated by the reflections across $\theta'^{\perp}_k$. Hence it acts on $i\mathfrak{T}'^*$ by

$$\mu=[\mu_1,\ldots,\mu_{m+2}] \rightarrow [\mu_{\sigma(1)},\ldots,\mu_{\sigma(m)},\mu_{s(m+1)},\mu_{s(m+2)}], \quad \begin{cases} \sigma \in \mathfrak{S}_m, \\ s \in \mathfrak{S}_2. \end{cases} \tag{13}$$

Conditions for a weight $\beta \in \mathfrak{T}'^*$ to be dominant are $2\langle\beta,\theta'_k\rangle/\langle\theta'_k,\theta'_k\rangle \in \mathbb{N}$, $k=1,\ldots,m$, so we get

Lemma 2: *A weight of* K, $\beta=(l_1,\ldots,l_{m+1})=[q_1,\ldots,q_{m+2}]$ *is dominant iff*

$$\begin{cases} (l_i - l_{i+1}) \in \mathbb{N}, i=1,\ldots,m-1, \\ l_{m+1} \in \mathbb{N}, \end{cases} \Leftrightarrow \begin{cases} (q_i - q_{i+1}) \in \mathbb{N}, \\ i=1,\ldots,m-1\ ,\ i=m+1. \end{cases}$$

Every class $[\alpha]$ in $[K]$ is related to a dominant weight of K, but the converse is not true as K is not semi-simple. Indeed, dominant weights induce possibly multivalued representations, so we determine those giving a single-valued representation.

Lemma 3: *A dominant weight of* K, $\beta=(l_1,\ldots,l_{m+1})=[q_1,\ldots,q_{m+2}]$ *induces a single-valued representation iff*

$$l_i \in \mathbb{Z} \Leftrightarrow q_i - q_{m+2} \in \mathbb{Z}, i=1,\ldots,m-1.$$

Such weights are *called* analytically dominant weights.

Proof: The center of the Lie algebra $\mathfrak{K}$ is $\mathfrak{C}=\mathbb{R}$. diag. matrix$(i,\ldots,i,-im/2,-im/2)$. As $\mathfrak{K}$ is of compact type, one has the decomposition $\mathfrak{K}=\mathfrak{C}\oplus\mathfrak{C}^{\perp}$, where $\mathfrak{C}^{\perp}\simeq\mathfrak{su}(m)\oplus\mathfrak{su}(2)$ is semi-simple. This decomposition induces $\mathfrak{T}'=\mathfrak{C}\oplus\mathfrak{T}''$, where $\mathfrak{T}''=\mathfrak{C}^{\perp}\cap\mathfrak{T}'$, and

$$i\mathfrak{T}'^{*}=i\mathfrak{C}^{*}\oplus i\mathfrak{T}''^{*}. \tag{14}$$

Let $\beta=(l_1,\ldots,l_{m+1})=[q_1,\ldots,q_{m+2}]$, a dominant weight of K. By (14), one has the decomposition $\beta=\gamma+\delta$, with $\gamma=n(1,1,\ldots,1,0)\in i\mathfrak{C}^{*}$, $n=1/m\Sigma_{i=1}^{m}l_i-1/2l_{m+1}$, and $\delta=(l_1',\ldots,l_{m+1}')\in i\mathfrak{T}''^{*}$,

$$\begin{cases} l_i'=l_i-n, & i=1\ldots,m,\\ l_{m+1}'=l_{m+1}. \end{cases}$$

Let

$$H=\left\{A=\begin{pmatrix} B & 0\\ 0 & C\end{pmatrix};B\in \mathrm{SU}(m);C\in \mathrm{SU}(2)\right\}.$$

This is a compact simply connected subgroup of K with Lie algebra $\mathfrak{C}^{\perp}$. The weight δ is dominant for H: indeed, relative to the maximal torus $T''=T'\cap H$, positive roots can be chosen such that the simple roots are restrictions of the θ_k' (12), hence conditions for a weight in $i\mathfrak{T}''^{*}$ to be dominant are the same as in lemma 2. Since H is simply connected, there exists, up to equivalence, a unique irreducible representation of H corresponding to the weight δ. Let $\rho:H\to Gl_{\mathbb{C}}(V)$, such a representation.

Now, for any

$$A=\begin{pmatrix} B & 0\\ 0 & C\end{pmatrix}\in K,$$

let $B'=(\det B)^{-1/m}B\in \mathrm{SU}(m)$ and $C'=(\det C)^{-1/2}C\in \mathrm{SU}(2)$. Of course, B' is defined up to an m-root of the unity, and C' is defined up to ± 1. Let $\tilde{\rho}$ the possibly multivalued representation $K\to Gl_{\mathbb{C}}(V)$,

$$\tilde{\rho}(A)=(\det B)^{n}\rho\left(\begin{pmatrix} B' & 0\\ 0 & C'\end{pmatrix}\right).$$

We claim that this is an irreducible (possibly multivalued) representation of K with dominant weight β:

Irreducibility comes from the fact that the restriction of $\tilde{\rho}$ to H is the (single-valued) irreducible representation ρ of H seen before. Furthermore, because $\tilde{\rho}_{*}[X=\text{ diag. matrix}(i\beta_1,\ldots,i\beta_{m+2})]\cdot v=\beta(X)v$, β is the dominant weight of $\tilde{\rho}$. Finally, as $\det B=e^{i\theta}$, where θ is defined up to $2\pi\mathbb{Z}$, a necessary and sufficient condition for $\tilde{\rho}$ to be single valued is $(e^{2k\pi i})^{n}\rho[\text{diag. matrix}(\exp(-2k\pi i/m),\ldots,\exp(-2k\pi i/m),e^{k\pi i},e^{k\pi i})]\cdot v=v,\forall k\in\mathbb{Z}$. This is equivalent to: $l_m\in\mathbb{Z}$, hence to: $l_i\in\mathbb{Z},i=1,\ldots,m-1$, because of the dominance conditions of lemma 2. ■

C. Branching rules of *G* under the action of *K*

We state the branching rules, following very closely the computation methods (based on Weyl's formula for characters) of Ref. 18.

Proposition 2: *Let $\beta_G=[p_1,\ldots,p_{m+2}]$ the dominant weight of an irreducible G-*module V_G *(see Lemma 1). The decomposition of V_G into irreducible K-modules contains the K-*module V_K *of dominant weight $\beta_K=[q_1,\ldots,q_{m+2}]$, (see lemmas 2, 3), if and only if the following three conditions are satisfied:*

(i) $(q_{m+2}-p_{m+2})\in\mathbb{Z}$,

(ii) $p_{i+2}\leqslant q_i\leqslant p_i$, $1\leqslant i\leqslant m$,

(iii) in the (finite) power series expansion in X of $(X-X^{-1})^{-m}\Pi_{i=1}^{m+1}(X^{r_i+1}-X^{-(r_i+1)})$, *where*

$$\begin{cases} r_1=p_1-\max(q_1,p_2), \\ r_i=\min(q_{i-1},p_i)-\max(q_i,p_{i+1}), \quad 2\leqslant i\leqslant m, \\ r_{m+1}=\min(q_m,p_{m+1})-p_{m+2}, \end{cases}$$

the coefficient of $X^{(q_{m+1}-q_{m+2}+1)}$ does not vanish. Moreover, the value of this coefficient is the multiplicity of the K-module V_K.

Remark 1: *Note that by (i) and the conditions of Lemmas 1, 2, 3, the r_i in (iii) are integers. Furthermore, by (ii) these integers are non-negative. Hence, the coefficient of X^k, $k>0$, in the expansion of condition (iii) is a non-negative integer. Moreover, the coefficient of X^{-k} is equal to the opposite coefficient of X^k.*

Proof. Let $\beta_G=(k_1,\ldots,k_{m+1})=[p_1,\ldots,p_{m+2}]$ be the dominant weight of an irreducible G-module V_G. By Weyl's formula, the character of the corresponding representation is $\chi_G(\beta_G)=J_G(\beta_G+\delta_G)/J_G(\delta_G)$, where $J_G(\beta)=\Sigma_{w\in W_G}(\det w)e^{w\cdot\beta}$, and where δ_G *is* the half-sum of the positive roots of G.

The corresponding notions for K are *denoted* by χ_K, J_K, δ_K respectively.

Let $S(\beta_G)$ the finite set of dominant weights of K, counted with their multiplicity, corresponding to the irreducible K-modules occurring in the decomposition of V_G. One has

$$\chi_G(\beta_G)=\sum_{\beta_K\in S(\beta_G)}\chi_K(\beta_K). \tag{15}$$

By Weyl's formula, (15) is equivalent to

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\sum_{\beta_K\in S(\beta_G)}J_K(\beta_K+\delta_K). \tag{16}$$

The proof consists in dividing $J_G(\beta_G+\delta_G)$ by $J_G(\delta_G)/J_K(\delta_K)$, and trying to set the result in the form $\Sigma J_K(\beta_K+\delta_K)$, where β_K are analytically dominant weights of K. Since in this case the $J_K(\beta_K+\delta_K)$ are linearly independent, we thus obtain the decomposition (16).

Let Φ_G^+ (respectively Φ_K^+) be the set of positive roots of G (respective of K). It's a general result that $J_G(\delta_G)=\Pi_{\alpha\in\Phi_G^+}(e^{\alpha/2}-e^{-(\alpha/2)})$, $J_K(\delta_K)=\Pi_{\alpha\in\Phi_K^+}(e^{\alpha/2}-e^{-(\alpha/2)})$. Since $\Phi_K^+\subset\Phi_G^+$, we get $J_K(\delta_K)/J_G(\delta_G)=\Pi_{\alpha\in\Phi_G^+\setminus\Phi_K^+}(e^{\alpha/2}-e^{-(\alpha/2)})^{-1}$. One has $\Phi_G^+\setminus\Phi_K^+=\{\hat{x}_i-\hat{x}_{m+1},\hat{x}_i-\hat{x}_{m+2},\ 1\leqslant i\leqslant m\}$, where $\hat{x}_{m+2}$ *denotes* the restriction of $\hat{y}_{m+2}$ to $i\mathfrak{T}'^*$. Thus we get

$$\frac{J_K(\delta_K)}{J_G(\delta_G)}=\prod_{i=1}^{m}e^{(\hat{z}_i+\hat{z}_i')/2}(e^{\hat{z}_i}-1)^{-1}(e^{\hat{z}_i'}-1)^{-1}, \tag{17}$$

with the following *notations*: $\hat{z}_i:=\hat{x}_i-\hat{x}_{m+1}$, $\hat{z}_i':=\hat{x}_i-\hat{x}_{m+2}$. Since

$$\delta_G=\sum_{i=1}^{m+1}(m+2-i)\hat{x}_i=\frac{1}{2}\sum_{i=1}^{m+2}(m+3-2i)\hat{y}_i, \tag{18}$$

we have $\beta_G+\delta_G=(k'_1,\ldots,k'_{m+1})=[p'_1,\ldots,p'_{m+2}]$, where $k'_i:=k_i+(m+2-i)$, $i=1,\ldots,m+1$. Hence, by the description of the Weyl group W_G (see (11)), we get

$$J_G(\beta_G+\delta_G)=\det[e^{p'_i\hat{x}_j}]_{1:m+2}. \tag{19}$$

Here and in the following we *denote* by $[a_{ij}]_{p:q}$ a square matrix whose suffixes i,j range from p to q.

The half-sum of the positive roots δ_K is $\delta_K=\frac{1}{2}(\Sigma_{i=1}^{m}(m+2-2i)\hat{x}_i+2\,\hat{x}_{m+1})$. If $\beta_K=(l_1,\ldots,l_{m+1})=[q_1,\ldots,q_{m+2}]$ is an analytically dominant weight of K, then $\beta_K+\delta_K=(l'_1,\ldots,l'_{m+1})=[q'_1,\ldots,q'_{m+2}]$, where $l'_i:=l_i+1/2\,(m+2-2i)$, $i=1,\ldots,m$, $l'_{m+1}:=l_{m+1}+1$. By the description of the Weyl group W_K (13), we get $J_K(\beta_K+\delta_K)=\det[e^{q'_i\hat{x}_j}]_{1:m}\det[e^{q'_i\hat{x}_j}]_{m+1:m+2}$. Since $l'_{m+1}=q'_{m+1}-q'_{m+2}$, we have $\det[e^{q'_i\hat{x}_j}]_{m+1:m+2}=e^{q'_{m+2}(\hat{x}_{m+1}+\hat{x}_{m+2})}(e^{l'_{m+1}\hat{x}_{m+1}}-e^{l'_{m+1}\hat{x}_{m+2}})$, hence since $\hat{x}_{m+1}+\hat{x}_{m+2}=-\Sigma_{i=1}^{m}\hat{x}_i$, and $l'_i=q'_i-q'_{m+2}$, $i=1,\ldots,m$, we get

$$J_K(\beta_K+\delta_K)=\mu(l'_{m+1})\det[e^{l'_i\hat{x}_j}]_{1:m}. \tag{20}$$

Here and in the following, we *set* $\mu(k)=e^{k\hat{x}_{m+1}}-e^{k\hat{x}_{m+2}}$. In (19), multiplying each column C_i by $e^{-p'_{m+2}\hat{x}_i}$, we get, since $\Sigma_{i=1}^{m+2}\hat{x}_i=0$,

$$J_G(\beta_G+\delta_G)=\det[e^{k'_i\hat{x}_j}]_{1:m+2}\quad\text{with}\quad k'_{m+2}=0\ .$$

Subtracting the $(m+2)$th line multiplied by $e^{k'_i\hat{x}_{m+2}}$ from the ith line, $i=1,\ldots,m+1$, then developing relatively to the last column we get

$$J_G(\beta_G+\delta_G)=\det[e^{k'_i\hat{x}_j}-e^{k'_i\hat{x}_{m+2}}]_{1:m+1}.$$

Subtracting the $(i+1)$th line multiplied by $\mu(k'_i)/\mu(k'_{i+1})$, from the ith line, $i=1,\ldots,m$, then developing relatively to the last column we get

$$J_G(\beta_G+\delta_G)=\left(\prod_{i=1}^{m-1}\mu(k'_{i+1})^{-1}\right)\det[a_{ij}]_{1:m},$$

where

$$a_{ij}=\mu(k'_{i+1})(e^{k'_i\hat{x}_j}-e^{k'_i\hat{x}_{m+2}})-\mu(k'_i)(e^{k'_{i+1}\hat{x}_j}-e^{k'_{i+1}\hat{x}_{m+2}}).$$

Because of the expression of $J_K(\delta_K)/J_G(\delta_G)$ (17), we divide a_{ij} by $(e^{\hat{z}_j}-1)(e^{\hat{z}'_j}-1)$, the result is $\mu(1)^{-1}\Sigma_{h\in\mathbb{N}}P_i(h)e^{h\hat{x}_j}$, where

$$P_i(h)=\begin{cases}-e^{(k'_{i+1}+1)(\hat{x}_{m+1}+\hat{x}_{m+2})}\mu(k'_i-k'_{i-1})\mu(-(h+1)) & \text{if } 0\leqslant h<k'_{i+1},\\ e^{(\hat{x}_{m+1}+\hat{x}_{m+2})}\mu(k'_{i+1})\mu(k'_i-(h+1)) & \text{if } k'_{i+1}\leqslant h<k'_i-1\ ,\\ 0 \quad otherwise.\end{cases}$$

Since we have $\hat{z}_i+\hat{z}'_i=2\,\hat{x}_i-(\hat{x}_{m+1}+\hat{x}_{m+2})$, we get

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\frac{e^{-m/2\,(\hat{x}_{m+1}+\hat{x}_{m+2})}}{\mu(1)^m(\Pi_{i=2}^m\mu(k_i'))}\det\left[\sum_{h\in\mathbb{N}}P_i(h)e^{(h+1)\hat{x}_j}\right]_{1:m},$$

$$=\frac{e^{m/2\,(\hat{x}_{m+1}+\hat{x}_{m+2})}}{\mu(1)^m(\Pi_{i=2}^m\mu(k_i'))}\det\left[\sum_{h\in\mathbb{N}}Q_i(h)e^{h\hat{x}_j}\right]_{1:m},$$

where

$$Q_i(h)=\begin{cases}-e^{k_{i+1}'(\hat{x}_{m+1}+\hat{x}_{m+2})}\mu(k_i'-k_{i-1}')\mu(-h) & if\ 1\leqslant h\leqslant k_{i+1}',\\ \mu(k_{i+1}')\mu(k_i'-h) & if\ k_{i+1}'<h<k_i',\\ 0 & otherwise.\end{cases}$$

By properties of the determinant, we can write

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\frac{e^{m/2(\hat{x}_{m+1}+\hat{x}_{m+2})}}{\mu(1)^m(\Pi_{i=2}^m\mu(k_i'))}\sum_{h_i'\in\mathbb{N}}\left(\prod_{i=1}^m Q_i(h_i')\right)\det[e^{h_i'\hat{x}_j}]_{1:m}. \tag{21}$$

Since $\det[e^{h_i'\hat{x}_j}]_{1:m}=0$, if $h_i'=h_j'$, $i\neq j$, the h_i' in the sum (21) can be assumed all distinct. Then, always by the properties of the determinant, we can write

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\frac{e^{m/2(\hat{x}_{m+1}+\hat{x}_{m+2})}}{\mu(1)^m(\Pi_{i=2}^m\mu(k_i'))}\Sigma\det[Q_i(h_j')]_{1:m}\det[e^{h_i'\hat{x}_j}]_{1:m},$$

where the sum is taken over all of the m-tuples $(h_1',\ldots,h_m')$ of integers satisfying

$$h_1'>h_2'>\cdots>h_m'>0\ . \tag{22}$$

If such an $(h_1',\ldots,h_m')$ does not satisfy

$$k_{i+2}'<h_i'<k_i',\ i=1,\ldots,m, \tag{23}$$

then $\det[Q_i(h_j')]_{1:m}$ vanishes. Indeed, if $h_m'\geqslant k_m'$, then $h_i'\geqslant k_m'$, so $Q_m(h_i')=0$ and all the entries of the last line are 0. If $h_j'\geqslant k_j'$, for some $j=1,\ldots,m-1$, then $\det[Q_i(h_j')]_{1:m}=\det[Q_i(h_j')]_{1:j}\det[Q_i(h_j')]_{j+1:m}$, with $\det[Q_i(h_j')]_{1:j}=0$, because all the entries of the last line are 0. If $h_j'\leqslant k_{j+2}'$, for some $j=1,\ldots,m-1$, then we transform $\det[Q_i(h_j')]_{1:m}$ by subtracting the jth column, multiplied by $\mu(-h_k')/\mu(-h_j')$, from the kth column, $k=j+1,\ldots,m$. The result is $\det[Q_i(h_j')]_{1:m}=\det[Q_i(h_j')]_{1:j}\det[b_{ij}]_{j+1:m}$, where all the entries of the first line of $[b_{ij}]_{j+1:m}$ are 0. Hence, we get

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\frac{e^{m/2(\hat{x}_{m+1}+\hat{x}_{m+2})}}{\mu(1)^m(\Pi_{i=2}^m\mu(k_i'))}\sum{}'\det[Q_i(h_j')]_{1:m}\det[e^{h_i'\hat{x}_j}]_{1:m}, \tag{24}$$

where Σ' means that the sum is taken over all of the m-tuples $(h_1',\ldots,h_m')$ satisfying (22) and (23).

Assuming that the h_i' satisfy (22) and (23), we compute $\det[Q_i(h_j')]_{1:m}$ by subtracting the $(j-1)$th column, multiplied by $\mu(-h_j')/\mu(-h_{j-1}')$, from the jth column, $j=m,m-1,\ldots,2$. The resulting matrix $[c_{ij}]_{1:m}$ is a tridiagonal matrix such that $c_{j,j+1}c_{j+1,j}=0$, $j=1,\ldots,m-1$, hence its determinant is the product of the diagonal elements. So we get

$$\det[Q_i(h'_j)]_{1:m}=Q_1(h'_1)\prod_{i=2}^{m}\left(Q_i(h'_i)-\frac{\mu(-h'_i)}{\mu(-h'_{i-1})}Q_i(h'_{i-1})\right). \tag{25}$$

Now every $\mu(k'_i)$ factors through c_{ii}: $c_{ii}=\mu(k'_i)e^{(k'_{i+1}-s'_{i-1})(\hat{x}_{m+1}+\hat{x}_{m+2})}\mu(r'_i)\mu(-s'_i)/\mu(-s'_{i-1})$, with

$$r'_i:=\min(h'_{i-1},k'_i)-\max(h'_i,k'_{i+1}),\quad i=1,\ldots,m,\quad h'_0\equiv k'_1,$$
$$s'_i:=\min(h'_i,k'_{i+1}),\quad i=0,\ldots,m.$$

Hence (25) gives

$$\det[Q_i(h'_j)]_{1:m}=\left(\prod_{i=1}^{m}\mu(k'_i)e^{(k'_{i+1}-s'_{i-1})(\hat{x}_{m+1}+\hat{x}_{m+2})}\mu(r_i')\right)\frac{\mu(-s'_m)}{\mu(-s'_0)}.$$

Now, since $\mu(-k)=-\mu(k)e^{-k(\hat{x}_{m+1}+\hat{x}_{m+2})}$, $\forall k\in\mathbb{Z}$, we obtain, *setting* $r'_{m+1}:=s'_m$, $\det[Q_i(h'_j)]_{1:m}=e^{P(\hat{x}_{m+1}+\hat{x}_{m+2})}(\Pi_{i=2}^{m}\mu(k'_i))(\Pi_{i=1}^{m+1}\mu(r'_i))$, where $P:=\Sigma_{i=0}^{m}(k'_{i+1}-s'_i)$. So (24) gives

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\sum{}'e^{(P+m/2)(\hat{x}_{m+1}+\hat{x}_{m+2})}\frac{\Pi_{i=1}^{m+1}\mu(r'_i)}{\mu(1)^m}\det[e^{h'_i\hat{x}_j}]_{1:m}. \tag{26}$$

With (16) and (20) in mind, we write the second part of equality (26)

$$\sum{}'\sum_{l'_{m+1}\in\mathbb{N}^*}e^{(P+m/2)(\hat{x}_{m+1}+\hat{x}_{m+2})}\frac{\Pi_{i=1}^{m+1}\mu(r'_i)}{\mu(1)^m\mu(l'_{m+1})}\mu(l'_{m+1})\det[e^{h'_i\hat{x}_j}]_{1:m}. \tag{27}$$

Let $X:=\exp((\hat{x}_{m+1}-\hat{x}_{m+2})/2)$. One has $\mu(k)=\exp(k(\hat{x}_{m+1}+\hat{x}_{m+2})/2)(X^k-X^{-k})$, $\forall k\in\mathbb{Z}$. Then the sum in (27) is equal to

$$\sum{}'\sum_{l'_{m+1}\in\mathbb{N}^*}\frac{\Pi_{i=1}^{m+1}(X^{r'_i}-X^{-r'_i})}{(X-X^{-1})^m(X^{l'_{m+1}}-X^{-l'_{m+1}})}\mu(l'_{m+1})\det[e^{l_i'\hat{x}_j}]_{1:m},$$

where

$$l'_i:=h'_i-M/2,\quad M:=\left(\sum_{i=1}^{m+1}k'_i-\sum_{i=1}^{m}h'_i-l'_{m+1}\right). \tag{28}$$

Now, we look for conditions on l'_{m+1}, such that $\mu(l'_{m+1})\det[e^{l_i'\hat{x}_j}]_{1:m}=J_K(\beta_K+\delta_K)$, where β_K is an analytically dominant weight of K, (see (20)). Let

$$\begin{cases} l_i:=l'_i-m/2+i-1, & i=1,\ldots,m,\\ l_{m+1}:=l'_{m+1}-1\ . \end{cases} \tag{29}$$

We have $l_{m+1}\in\mathbb{N}$ and by (22), $l_i-l_{i+1}\in\mathbb{N}$, $i=1,\ldots,m-1$, hence $\beta_K=(l_1,\ldots,l_{m+1})$ is a dominant weight for K (see lemma 2). Furthermore it is an analytically dominant weight iff

$$(M+m)/2\in\mathbb{Z}, \tag{30}$$

From (28) and (29), we get $M=2/(m+2)\,\Sigma_{i=1}^{m+1}(k_i-l_i)+m$, so the condition (30) is equivalent to

$$h_0:=\frac{1}{m+2}\sum_{i=1}^{m+1}(k_i-l_i)\in\mathbb{Z}. \tag{31}$$

Let $h_i:=l_i+h_0$, $i=1,\ldots,m$, then (23) gives

$$\begin{cases} k_{i+2}\leqslant h_i\leqslant k_i, & i=1,\ldots,m-1,\\ 0\leqslant h_m\leqslant k_m. \end{cases} \tag{32}$$

Furthermore, we have

$$\begin{cases} r_1'=k_1-\max(h_1,k_2)+1,\\ r_i'=\min(h_{i-1},k_i)-\max(h_i,k_{i+1})+1, & i=2,\ldots,m,\\ r_{m+1}'=\min(h_m,k_{m+1})+1. \end{cases} \tag{33}$$

Now, since each analytically dominant weight $\beta_K=(l_1,\ldots,l_{m+1})$ satisfying conditions (31), (32), corresponds by (28), (29) to a unique $(m+1)$-tuple $(h_1',\ldots,h_m',l_{m+1}')$ of positive integers satisfying conditions (22), (23) and (30), we can write (26) as

$$J_G(\beta_G+\delta_G)\frac{J_K(\delta_K)}{J_G(\delta_G)}=\sum\frac{\Pi_{i=1}^{m+1}(X^{r_i'}-X^{-r_i'})}{(X-X^{-1})^m(X^{(l_{m+1}+1)}-X^{-(l_{m+1}+1)})}J_K(\beta_K+\delta_K), \tag{34}$$

where the sum is taken over all the analytically dominant weights $\beta_K=(l_1,\ldots,l_{m+1})$ satisfying (31), (32), and the following condition:

In (34), the coefficient of $X^{(l_{m+1}+1)}$ in the power series expansion in X of $(X-X^{-1})^m\Pi_{i=1}^{m+1}(X^{r_i'}-X^{-r_i'})$, does not vanish.

This gives the decomposition (16). Now it's easy to see from (9), that for a weight $\beta_K=(l_1,\ldots,l_{m+1})=[q_1,\ldots,q_{m+1}]$ conditions (31), (32), and the last condition obtained are respectively equivalent to conditions (i), (ii), (iii) of the Proposition 2. Hence the result. ■

V. DECOMPOSITION OF THE SPINOR SPACE Σ UNDER THE ACTION OF K.

Notations are those of Secs. II and III. We henceforward assume that m is even. We consider the same spinor representation as in Ref. 19:

Let $(X_k)_{1\leqslant k\leqslant 4m}$ be the orthonormal basis of $\mathfrak{P}\simeq M_{m,2}(\mathbb{C})$ given by

$$\begin{cases} X_{4k-3}=\dfrac{1}{2\sqrt{m+2}}E_{k,1}\\ X_{4k-2}=-i\dfrac{1}{2\sqrt{m+2}}E_{k,1}\\ X_{4k-1}=-\dfrac{1}{2\sqrt{m+2}}E_{k,2}\\ X_{4k}=-i\dfrac{1}{2\sqrt{m+2}}E_{k,2} \end{cases},\quad 1\leqslant k\leqslant m,$$

where E_{ij} *denotes* the usual basis of $M_{m,2}(\mathbb{C})$. Under the identification $\mathfrak{P}\simeq\mathbb{H}^m$ given in (2), this basis corresponds to the orthonormal basis $\mathscr{B}=(e_k, Ie_k, Je_k, Ke_k)$, $1\leqslant k\leqslant m$, of $\mathbb{R}^{4m}$ used in Ref. 19. The following vectors

$$\begin{cases} x_k=\frac{1}{2}(e_k+iIe_k), & \overline{x_k}=\frac{1}{2}(e_k-iIe_k), \\ y_k=\frac{1}{2}(Je_k+iKe_k), & \overline{y_k}=\frac{1}{2}(Je_k-iKe_k), \end{cases}$$

form a Witt basis of $\mathbb{C}^{4m}$: one has $(x_k,x_l)=(\overline{x_k},\overline{x_l})=(y_k,y_l)=(\overline{y_k},\overline{y_l})=0$,$(x_k,\overline{x_l})=(y_k,\overline{y_l})=\frac{1}{2}\delta_{kl}$, where $(,)$ is the complex bilinear extension to $\mathbb{C}^{4m}$ of the usual scalar product of $\mathbb{R}^{4m}$.

Let $\Phi=\overline{x_1}\cdot\overline{x_2}\cdots\overline{x_m}\cdot\overline{y_1}\cdot\overline{y_2}\cdots\overline{y_m}\in\mathbb{C}l_{4m}$. (Clifford multiplication is *denoted* by a centered dot). The vector space $\Sigma=\mathbb{C}l_{4m}\cdot\Phi$, is of complex dimension 4^m. Indeed a basis is given by

$$\varphi_{IJ}=x_{i_1}\cdot x_{i_2}\cdots x_{i_p}\cdot y_{j_1}\cdot y_{j_2}\cdots y_{j_q}\cdot\Phi,$$

where

$$\begin{cases} I=(i_1,\ldots,i_p), & 0\leqslant i_1<\cdots<i_p\leqslant m, \\ J=(j_1,\ldots,j_q), & 0\leqslant j_1<\cdots<j_q\leqslant m, \end{cases}$$

with the convention $x_0=\overline{x_0}=1$.

The spin representation is then defined by

$$\rho:\mathbb{C}\!\not{l}_{4m}\to Hom_{\mathbb{C}}(\Sigma,\Sigma),\varphi\to(\rho(\varphi):\psi\cdot\Phi\to\varphi\cdot\psi\cdot\Phi).$$

Let $X=\text{diag. matrix}(i\beta_1,\ldots,i\beta_{m+1},i\beta_{m+2})\in\mathfrak{T}'$, $\Sigma_{k=1}^{m+2}\beta_k=0$. By the results of the end of Sec. II, we have

$$\tilde{\alpha}_*(X)=\frac{1}{2}\sum_{k=1}^{m}(\beta_{m+1}-\beta_k)e_k\cdot Ie_k+\frac{1}{2}\sum_{k=1}^{m}(\beta_k-\beta_{m+2})Je_k\cdot Ke_k.$$

Hence from the relations

$$\begin{cases} e_k\cdot Ie_k\cdot\varphi_{IJ}=i\varphi_{IJ} & if \quad k\notin I, \\ e_k\cdot Ie_k\cdot\varphi_{IJ}=-i\varphi_{IJ} & if \quad k\in I, \\ Je_k\cdot Ke_k\cdot\varphi_{IJ}=i\varphi_{IJ} & if \quad k\notin J, \\ Je_k\cdot Ke_k\cdot\varphi_{IJ}=-i\varphi_{IJ} & if \quad k\in J, \end{cases}$$

we get

$$(\rho_*\circ\tilde{\alpha}_*)(X)\cdot\varphi_{IJ}=i\left(\sum_{k=1}^{p}\beta_{i_k}-\sum_{k=1}^{q}\beta_{j_k}+\left(\frac{m}{2}-p\right)\beta_{m+1}+\left(q-\frac{m}{2}\right)\beta_{m+2}\right)\varphi_{IJ}. \tag{35}$$

So, under the action of the maximal torus T', the spinor space splits into the direct sum of the $\mathbb{C}\varphi_{IJ}$, with corresponding weights $[\varepsilon_1,\ldots,\varepsilon_m,(m/2-p),(q-m/2)]$, $\varepsilon_k=-1,0,$ or 1. Hence from lemma 2, we deduce

Proposition 3: *The spinor space* Σ *splits into the direct sum of* $[(m+1)(m/2+1)]$ *irreducible K-modules* $\Sigma_{p,q}$, $0\leqslant p+q\leqslant m$, *with dominant weight*

$$\beta_{p,q}=[\underbrace{1,\ldots,1}_{p},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q},(m/2-p),(q-m/2)].$$

Furthermore the multiplicity is 1.

Proof: By Weyl's dimension formula, one has

$$n_{p,q}:=\dim\Sigma_{p,q}=\frac{(m-(p+q)+1)^2}{m+1}\binom{m+1}{p}\binom{m+1}{q},$$

and a proof by induction shows that $\Sigma_{0\leqslant p+q\leqslant m}n_{p,q}=4^m=\dim\Sigma$, so, the multiplicity is 1. ■

VI. SPECTRUM OF D^2

A. Determination of the $[\alpha]$'s in $[G]$ such that $m([\alpha]_{|K},\Sigma_{p,q})\geqslant 1$

From Propositions 2 and 3, we get, after some tedious computations:

Proposition 4: *Let* $\beta_G=[p_1,\ldots,p_{m+2}]$ *the dominant weight of an irreducible G-module* V_G. *Then the decomposition of* V_G *into irreducible K-modules contains a K-module isomorphic to* $\Sigma_{p,q}$ *if and only if* β_G *is of the following form*:

(i) If $m=2$,

$$\beta_G=[k+l,k,-k,-(k+l)],\quad k,l\in\mathbb{N},\quad (k,l)\neq(0,0),$$
$$\beta_G=[k+l+2,k,-(k+1),-(k+l+1)],\quad k,l\in\mathbb{N},$$
$$\beta_G=[k+l+1,k+1,-k,-(k+l+2)],\quad k,l\in\mathbb{N}.$$

(ii) If $m\geqslant 4$,

$$\begin{cases}\beta_G=[k+l+n+m/2-(p'+2),\underbrace{k,1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q},-(k+n),-(k+l+m/2-(q'+2))],\\ \quad 0\leqslant p'\leqslant m/2-2,\ 0\leqslant q'\leqslant m/2-2,\ k,l,n\in\mathbb{N},\ s.t.0\leqslant n\leqslant m/2-(q'+2)\quad and\\ \quad k\geqslant 1\ \ if\ p'\geqslant 1\ ,\\ \quad (k,l)\neq(0,0)\ \ if\ p'=0\quad and\ \ 1\leqslant n\leqslant m/2-(q'+2),\\ \quad (k,l)\neq(0,0)\quad and\ \ (k,l)\neq(0,1)\ \ if\ p'=q'=n=0,\\ \beta_G=[k+l+m/2-(p'+2),k+n,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-k,-(k+l+n+m/2-(q'+2))],\\ \quad 0\leqslant p'\leqslant m/2-2,\ 0\leqslant q'\leqslant m/2-2,\ k,l,n\in\mathbb{N},\ s.t.0\leqslant n\leqslant m/2-(p'+2)\quad and\\ \quad k\geqslant 1\ \ if\ q'\geqslant 1,\\ \quad (k,l)\neq(0,0)\ \ if\ q'=0\quad and\ \ 1\leqslant n\leqslant m/2-(p'+2),\\ \quad (k,l)\neq(0,0)\quad and\ \ (k,l)\neq(0,1)\ \ if\ p'=q'=n=0,\end{cases}\tag{36}$$

$$\begin{cases}\beta_G=[k+l+n+1,k+1,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+n-m/2+p'+3),-(k+l+m/2-(q'+1))],\\ \quad m/2-2\leqslant p'\leqslant m-4,\ 0\leqslant q'\leqslant m-(p'+4),\ k,l,n\in\mathbb{N},\ s.t.0\leqslant n\leqslant m-(p'+q'+4),\\ \beta_G=[k+l+m/2-(p'+1),k+n-m/2+p'+3,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+1),-(k+l+n+1)],\\ \quad m/2-2\leqslant q'\leqslant m-4,\ 0\leqslant p'\leqslant m-(q'+4),\ k,l,n\in\mathbb{N},\ s.t.0\leqslant n\leqslant m-(p'+q'+4),\end{cases}\tag{37}$$

$$\begin{cases} \beta_G=[k+l+1,k+1,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k-m/2+p'+2),-(k+l+m/2-q')], \\ \quad m/2-1\leqslant p'\leqslant m-2,\ 0\leqslant q'\leqslant m-(p'+2),\ \ k,l\in\mathbb{N}, \\ \beta_G=[k+l+m/2-p',k-m/2+q'+2,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+1),-(k+l+1)], \\ \quad m/2-1\leqslant q'\leqslant m-2,\ 0\leqslant p'\leqslant m-(q'+2),\ \ k,l\in\mathbb{N}, \end{cases} \tag{38}$$

$$\begin{cases} \beta_G=[k+l+1,k+1,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k-m/2+p'+1),-(k+l+m/2-q'+1)], \\ \quad m/2\leqslant p'\leqslant m-2,\ 0\leqslant q'\leqslant m-(p'+2),\ \ k,l\in\mathbb{N}, \\ \beta_G=[k+l+m/2-p'+1,k-m/2+q'+1,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+1),-(k+l+1)], \\ \quad m/2\leqslant q'\leqslant m-2,\ 0\leqslant p'\leqslant m-(q'+2),\ \ k,l\in\mathbb{N}, \end{cases} \tag{39}$$

$$\begin{cases} \beta_G=[k+l+m/2-p',k+m/2-p',\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-k,-(k+l+m-(p'+q'))], \\ \quad 0\leqslant p'\leqslant m/2-1,\ 0\leqslant q'\leqslant m-(p'+2),\ \ k,l\in\mathbb{N},\ \ s.t.k\geqslant 1\ \ if\ q'\geqslant 1\ , \\ \beta_G=[k+l+m-(p'+q'),k,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+m/2-q'),-(k+l+m/2-q')], \\ \quad 0\leqslant q'\leqslant m/2-1,\ 0\leqslant p'\leqslant m-(q'+2),\ \ k,l\in\mathbb{N},\ \ s.t.k\geqslant 1\ \ if\ p'\geqslant 1, \end{cases} \tag{40}$$

$$\begin{cases} \beta_G=[k+l+m/2-(p'+1),k+m/2-(p'+1),\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-k,-(k+l+m-(p'+q'+2))], \\ \quad 0\leqslant p'\leqslant m/2-2,\ 0\leqslant q'\leqslant m-(p'+2),\ \ k,l\in\mathbb{N},\ \ s.t.k\geqslant 1\ \ if\ q'\geqslant 1, \\ \beta_G=[k+l+m-(p'+q'+2),k,\underbrace{1,\ldots,1}_{p'},0,\ldots,0,\underbrace{-1,\ldots,-1}_{q'},-(k+m/2-(q'+1)),-(k+l+m/2-(q'+1))], \\ \quad 0\leqslant q'\leqslant m/2-2,\ 0\leqslant p'\leqslant m-(q'+2),\ \ k,l\in\mathbb{N},\ \ s.t.k\geqslant 1\ \ if\ p'\geqslant 1, \end{cases} \tag{41}$$

B. Computation of the eigenvalues

From (4) and since $S=2^m$ (see Sec. II), eigenvalues of D^2 are obtained from the list of dominant weights occurring in Proposition 4 by the formula

$$\lambda(\beta_G)=c(\beta_G)+\frac{m}{4},$$

where $c(\beta_G)$ is the eigenvalue of the Casimir operator of the representation corresponding to β_G. By Freudenthal's formula we have

$$c(\beta_G)=\frac{1}{2(m+2)}\langle\beta_G,\beta_G+2\delta_G\rangle, \tag{42}$$

where $\langle,\rangle$ is the scalar product (10), and δ_G the half-sum of the positive roots of G, see (18). Note that if $\beta_G=[p_1,\ldots,p_{m+2}]$ is a dominant weight of G, so is $\beta'_G=[-p_{m+2},-p_{m+1},\ldots,-p_1]$, and it is easily verified that $\langle\beta_G,\beta_G+2\delta_G\rangle=\langle\beta'_G,\beta'_G+2\delta_G\rangle$. Hence the two types of dominant weights occurring in (36), (41) respectively, give the same eigenvalue. Computation of (42) for each of these cases leads to Proposition 1. In case $m=2$, we recover, up to a homothetic change of the metric, a result of Strese on the Dirac spectrum of the Grassmannian $Gr_4(\mathbb{R}^6)\simeq Gr_2(\mathbb{C}^4)$.[20]

Because of the expression of the eigenvalues of D^2, it is easy to see that the lowest one is $\mu_{0,1}=3/2$, if $m=2$, and $\lambda_{1,0,0,0,0}=\mu_{0,0,0,m/2-2,m/2-2}=m(m+4)/(2(m+2))$, if $m\geqslant 4$. Hence the square of the first eigenvalue of the Dirac operator is $\lambda^2=m(m+4)/(2(m+2))$. So the ratio $\nu:=\lambda^2/S$ of the square of the first eigenvalue by the scalar curvature is

$$\nu=\frac{m+4}{4\,(m+2)}>\nu_0=\frac{m+3}{4\,(m+2)}$$

where ν_0 is the corresponding ratio for the quaternionic projective space. Thus the conjecture of Ref. 1 is verified for the Grassmannian $Gr_2(\mathbb{C}^{m+2})$.

Furthermore, the following characterization of the first eigenvalue can be given:

Let $\Sigma(G/K)$ the spinor bundle of $Gr_2(\mathbb{C}^{m+2})=G/K$. From the results of Ref. 2, the space $\Sigma^\lambda(G/K)=\{\psi\in\Gamma(\Sigma(G/K));D^2\psi=\lambda^2\psi\}$, splits into the orthogonal decomposition

$$\Sigma^\lambda(G/K)=\bigoplus_{k=0}^{m}\Sigma^\lambda_k(G/K), \tag{43}$$

where $\Sigma^\lambda_k(G/K)=\{\psi\in\Sigma^\lambda(G/K);\Omega\cdot\psi=(6m-4k(k+2))\psi\}$, Ω being the fundamental 4-form given by the quaternion-Kähler structure of G/K.

Proposition 5: *If λ is the first eigenvalue of the Dirac operator, then there exists a non trivial spinor field in $\Sigma^\lambda_0(G/K)$.*

Proof: From Propositions 1 and 4, the lowest eigenvalue corresponds to the weight $\beta_G=[\underbrace{1,\ldots,1}_{m/2},0,0,\underbrace{-1,\ldots,-1}_{m/2}]$. From Proposition 2, the (up to equivalence) G-module V_G corresponding to β_G contains in its decomposition under the action of K, a K-module V_K isomorphic to $\Sigma_{m/2,m/2}$. Indeed, the dominant weight of $\Sigma_{m/2,m/2}$ is $\beta_{m/2,m/2}=[\underbrace{1,\ldots,1}_{m/2},\underbrace{-1,\ldots,-1}_{m/2},0,0]$. Conditions (i) and (ii) are easily verified. The expansion of condition (iii) equals $(X-X^{-1})$, so the condition (iii) is also verified. Furthermore the multiplicity of V_K is 1.

From (35), the vector of Σ, $\varphi=x_1\cdot x_2\cdots x_{m/2}\cdot y_1\cdot y_2\cdots y_{m/2}\cdot\Phi$, is a weight-vector for the weight $\beta_{m/2,m/2}$.

Now, from the general description of the spectrum of the Dirac operator on a symmetric space (see Refs. 5, 7, 9, 10), a spinor field $\psi\in\Sigma^\lambda(G/K)$ can be constructed as follows:

Let $\hat\rho$ be the representation $G\to Hom_{\mathbb{C}}(V_G,V_G)$. The condition that V_G contains an irreducible K-module isomorphic to $\Sigma_{m/2,m/2}$ is equivalent to the existence of a K-equivariant homomorphism $L:V_G\to\Sigma$, such that $L_{|V_K}$ is the isomorphism $V_K\simeq\Sigma_{m/2,m/2}$. The spinor bundle of G/K is isomorphic to $G\times_K\Sigma\to G/K$, $[g,\phi]\to[g]$, where K acts on the left on Σ by the representation $\rho\circ\widetilde{\alpha}$ of Sec. III. Then, the spinor field ψ defined by $\psi:[g]\to[g,(L\circ\hat\rho(g^{-1}))\cdot v]$, with $L(v)=\varphi$, belongs to $\Sigma^\lambda(G/K)$.

Under the decomposition (43), the spinor field ψ splits into $\psi=\Sigma_{k=0}^m\psi_k$. Now, from results of Ref. 19, the action of the fundamental 4-form Ω on ψ is

$$(\Omega\cdot\psi)([g])=[g,\mathbf{\Omega}\cdot(L\circ\hat\rho(g^{-1}))(v)],$$

where $\mathbf{\Omega}=\Sigma_{c.p.I,J,K}\mathbf{\Omega}_I\cdot\mathbf{\Omega}_I-6m$, with $\mathbf{\Omega}_I=\Sigma_{i=1}^{m}(e_i\cdot Ie_i+Je_i\cdot Ke_i)$, (e_i,Ie_i,Je_i,Ke_i), being the orthonormal basis of $\mathbb{R}^{4m}$ introduced in Sec. V.

If e denotes the neutral element in G, we get $(\Omega\cdot\psi)([e])=[e,\mathbf{\Omega}\cdot\varphi]$. From the result of page 331 in Ref. 19 we have $\mathbf{\Omega}\cdot\varphi=6m\varphi$. Hence $\psi_0([e])=\psi([e])\neq 0$, and ψ_0 is a non-trivial vector of $\Sigma_0^{\lambda}(G/K)$. Hence the result. ■

ACKNOWLEDGMENT

The author would like to thank Oussama Hijazi for useful comments on this paper.

[1] O. Hijazi and J.-L. Milhorat, J. Math. Pures Appl. **74**, 387 (1995).
[2] O. Hijazi and J.-L. Milhorat, Ann. Global An. Geom. **15**, 117 (1997).
[3] J. A. Wolf, J. Math. Mech. **14**, 1033 (1965).
[4] A. Besse, *Einstein Manifolds* (Springer-Verlag, Berlin, 1987).
[5] M. Cahen and S. Gutt, Simon Stevin **62**, 209 (1988).
[6] J.-L. Milhorat, C. R. Acad. Sci. Paris **314**, 69 (1992).
[7] C. Bär, *Das Spectrum von Dirac-Operator* (Bonner Mathematische, Schifren, 1990), p. 217.
[8] M. Cahen, A. Franc, and S. Gutt, Lett. Math. Phys. **18**, 165 (1989).
[9] R. Parthasarathy, Ann. Math. **69**, 1 (1972).
[10] S. Sulanke, *Berechnung des Spektrums des Quadrates des Dirac-operators auf der Sphäre* (Humbolt-Universität, Berlin, 1979).
[11] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces*, 2nd ed. (Academic, New York, 1978).
[12] S. Salamon, Invent. Math. **67**, 143 (1982).
[13] C. LeBrun and S. Salamon, Invent. Math. **118**, 109 (1994).
[14] J. F. Adams, *Lectures on Lie groups* (Benjamin, New York, 1967).
[15] J. Dieudonné, *Eléments d'Analyse* (Cahiers Sci. 38, Gauthier-Villars, Paris, 1975), Vol. 5.
[16] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer-Verlag, Berlin, 1972).
[17] D. P. Zelobenko, *Compact Lie groups and Their Representations*, Translations of Math. Monographs, Vol. 40 (American Mathematical Society, Providence, RI, 1973).
[18] C. Tsukamoto, Osaka J. Math. **18**, 407 (1981).
[19] O. Hijazi and J.-L. Milhorat, J. Geom. Phys. **15**, 320 (1995).
[20] H. Strese, Math. Nachr. **98**, 53 (1980).

Representation properties, Racah sum rule, and Biedenharn–Elliott identity for $U_q(\mathrm{osp}(1|2))$

Pierre Minnaert and Marek Mozrzymas[a)]
Centre de Physique Théorique et Modélisation,[b)] *Université Bordeaux I, France*[c)]

(Received 11 July 1997; accepted for publication 5 August 1997)

It is shown that the universal R matrix in the tensor product of two irreducible representation spaces of the quantum superalgebra $U_q(\mathrm{osp}(1|2))$ can be expressed by Clebsch–Gordan coefficients. The Racah sum rule satisfied by $U_q(\mathrm{osp}(1|2))$ Racah coefficients and $6-j$ symbols is derived from the properties of the universal R matrix in the tensor product of three representation spaces. Considering the tensor product of four irreducible representations, it is shown that Biedenharn–Elliott identity holds for $U_q(\mathrm{osp}(1|2))$ Racah coefficients and $6-j$ symbols. A recursion relation for $U_q(\mathrm{osp}(1|2))$ $6-j$ symbols is derived from the Biedenharn–Elliott identity. © *1998 American Institute of Physics.* [S0022-2488(97)03011-9]

I. INTRODUCTION

In three previous papers, Refs. 1–3, we have studied the properties of irreducible representations of the quantum superalgebra $U_q(\mathrm{osp}(1|2))$. It was shown that it is possible to construct Racah–Wigner calculus for this quantum superalgebra, in a completely similar way as in the classical algebra su(2)[4] and the quantum algebra $U_q(\mathrm{su}(2))$[5–7] cases. In this paper, in order to complete our study of the representations of this simple quantum superalgebra, we shall consider the properties of its universal R matrix in the tensor product of irreducible representations.

The universal R matrix for $U_q(\mathrm{osp}(1|2))$ has been derived and its basic properties have been exhibited in Ref. 8. It was also considered in the construction of vertex model solutions of the graded Yang–Baxter equation in Ref. 9. In the present paper we study the properties of the universal R matrix in the tensor product of two irreducible $U_q(\mathrm{osp}(1|2))$ representations. We give explicit formulas for the matrix representation of the universal R matrix in different bases of the tensor product of representations. In particular, we show that matrix elements of the universal R matrix in the reduced basis of the tensor product of two representations can be expressed with $U_q(\mathrm{osp}(1|2))$ Clebsch–Gordan coefficients (denoted sq-CGc). Analytic expressions for sq-CGc of any tensor product of irreducible representations are known,[1,10,11] therefore the matrix representation of the universal R matrix can be calculated for any tensor product.

In Ref. 2, we have defined and studied properties of Racah coefficients (denoted $sqRc$) and $6-j$ symbols (denoted $sq6-j$) for the quantum superalgebra $U_q(\mathrm{osp}(1|2))$. In particular, it was shown that, as in the classical case, $sq6-j$ symbols satisfy not only the usual tetrahedral symmetry but present also an additional symmetry of the Regge type. Another property satisfied by $sqRc$ and $sq6-j$ symbols is the pseudo-orthogonality relation. This relation can be considered an algebraic relation satisfied by $sqRc$ and $sq6-j$ symbols. It is known that in the classical Racah–Wigner calculus for su(2) or $U_q(\mathrm{su}(2))$, Racah coefficients and $6-j$ symbols satisfy, besides orthogonality relations, other algebraic relations, namely the Racah sum rule and the Biedenharn–Elliott identity. These algebraic identities have been extended to the corresponding features of the superalgebra osp(1|2), Ref. 12. In this paper we extend the Racah sum rule and the Biedenharn–Elliott identity for $sqRc$ and $sq6-j$ symbols. In both cases, the structure of these relations is

[a)]On leave of absence from Institute of Theoretical Physics, University of Wroclaw, Poland.
[b)]Unité Associée au CNRS, URA 1537.
[c)]Postal address: 19, rue du Solarium, 33174 Gradignan, Cedex.

0022-2488/98/39(1)/610/15/$15.00
© 1998 American Institute of Physics

completely similar to the corresponding classical ones, the only difference concerns the phases that are more complicated in the case of $U_q(\mathrm{osp}(1|2))$. As in the classical cases, we also derive from the Biedenharn–Elliott identity a three-term recurrence relation between $sq6-j$ symbols.

This paper is organized in the following way: in Sec. II we study the properties of the universal R matrix in the tensor product of irreducible representations of $U_q(\mathrm{osp}(1|2))$. In Sec. III, considering the tensor product of three $U_q(\mathrm{osp}(1|2))$ representations, we derive the Racah sum rule for $sqRc$ and $sq6-j$ symbols and we derive a simple algebraic identity following from the sum rule. Finally, in Sec. IV, we prove the Biedenharn–Elliott identity for $sq6-j$ symbols and the three-term recurrence relation.

II. PROPERTIES OF THE UNIVERSAL R MATRIX IN THE IRREDUCIBLE REPRESENTATIONS OF THE QUANTUM SUPERALGEBRA $U_q(\mathrm{osp}(1|2))$

A. The $U_q(\mathrm{osp}(1|2))$ universal R matrix

The quantum superalgebra $U_q(\mathrm{osp}(1|2))$ is generated by four elements: $\mathbf{1}$, H (even) and $v_\pm$ (odd) with the following (anti)commutation relations:

$$[H,v_\pm]=\pm\frac{1}{2}v_\pm, \quad [v_+,v_-]_+=-\frac{\sinh(\eta H)}{\sinh(2\eta)}, \tag{2.1}$$

where the deformation parameter η is real and $q=e^{-\eta/2}$ (we choose $\eta>0$ so that $q<1$). The quantum superalgebra $U_q(\mathrm{osp}(1|2))$ is a Hopf algebra with the following coproduct Δ^q and antipode S:

$$\Delta^q(v_\pm)=v_\pm\otimes q^H+q^{-H}\otimes v_\pm, \tag{2.2}$$

$$\Delta^q(H)=H\otimes\mathbf{1}+\mathbf{1}\otimes H, \quad \Delta^q(\mathbf{1})=\mathbf{1}\otimes\mathbf{1}, \tag{2.3}$$

$$S(H)=-H, \quad S(v_\pm)=-q^{\pm 1/2}v_\pm. \tag{2.4}$$

One can consider another coalgebra structure and antipode defined by

$$\Delta'^q(v_\pm)=v_\pm\otimes q^{-H}+q^H\otimes v_\pm, \tag{2.5}$$

$$S'(H)=-H, \quad S'(v_\pm)=-q^{\mp 1/2}v_\pm, \tag{2.6}$$

which, together with relations (2.1), define another Hopf algebra structure that we denote $U'_q(\mathrm{osp}(1|2))$. Both Hopf algebra structures are related by

$$U'_q(\mathrm{osp}(1|2))=U_{q^{-1}}(\mathrm{osp}(1|2)). \tag{2.7}$$

It is known[8] that a canonical element $R^q\in U_q(\mathrm{osp}(1|2))\otimes U_q(\mathrm{osp}(1|2))$ exists, called the universal R matrix, that defines a similarity relation between Δ'^q and Δ^q:

$$R^q\Delta^q=\Delta'^qR^q. \tag{2.8}$$

This universal R matrix is

$$R^q=q^{4H\otimes H}\sum(-1)^k q^{k(k-1)/4}\frac{(1+q^{-1})^k}{[k]!\,\gamma^k}(q^Hv_+)^k\otimes(q^Hv_-)^k, \tag{2.9}$$

and it satisfies the following relations:

$$R^{q13}R^{q23}=(\Delta\otimes id)R^q\equiv R^{q12,3}, \tag{2.10}$$

$$R^{q13}R^{q12}=(id\otimes\Delta)R^q\equiv R^{q1,23}, \tag{2.11}$$

where the indices $i,j=1,2,3$, $i\neq j$, in R^{qij} show the embeding of $R^q=\Sigma r_n\otimes r'_n$ into the tensor product $U_q(\mathrm{osp}(1|2))\otimes U_q(\mathrm{osp}(1|2))\otimes U_q(\mathrm{osp}(1|2))$. The above relations imply the Yang–Baxter equation for the universal R matrix,

$$R^{q12}R^{q13}R^{q23}=R^{q23}R^{q13}R^{q12}. \tag{2.12}$$

In the following we will consider the properties of the universal R matrix in the tensor product of irreducible representations of the quantum superalgebra $U_q(\mathrm{osp}(1|2))$.

B. The $U_q(\mathrm{osp}(1|2))$ finite-dimensional representations

A representation of a quantum superalgebra $U_q(\mathrm{osp}(1|2))$ in a finite-dimensional graded space V is a homomorphism $T:U_q(\mathrm{osp}(1|2))\to L(V,V)$ of the associative graded algebra $U_q(\mathrm{osp}(1|2))$ into the associative graded algebra $L(V,V)$ of linear operators in V, such that

$$[T(H),T(v_\pm)]=\pm\frac{1}{2}\,T(v_\pm),\quad [T(v_+),T(v_-)]_+=-\frac{\sinh(\eta T(H))}{\sinh(2\eta)}. \tag{2.13}$$

Let us recall the main results concerning these representations, Ref. 1. Any finite-dimensional grade star representation of $U_q(\mathrm{osp}(1|2))$ is characterized by four parameters: the highest weight l (a non-negative integer), the parity λ of the highest weight vector in the representation space, and the signature parameters $\varphi,\psi=0,1$ of the Hermitian form in the representation space V. The parity λ and the signature φ define the class $\epsilon=0,1$ of the grade star representation by the relation $\epsilon=\lambda+\varphi+1 \pmod 2$. The irreducible representation space of highest weight $l, V=V^l(\lambda)$ is a graded vector space of dimension $2l+1$ with basis $e^{lq}_m(\lambda)$, where $-l<m<l$, the parameter $\lambda=0,1$ is the parity of the highest weight vector $e^l_l(\lambda)$, and $\deg(e^l_m(\lambda))=l-m+\lambda \pmod 2$. The vectors $e^{lq}_m(\lambda)$ are constructed from the vector $e^l_l(\lambda)$ in a standard way so they depend on q via the normalization factor. The vectors $e^{lq}_m(\lambda)$ are pseudo-orthogonal with respect to a Hermitian form in the representation space, denoted $(\,,\,)$, and their normalization is determined by the signature parameters φ,ψ:

$$(e^{lq}_m(\lambda),e^{lq}_{m'}(\lambda))=(-1)^{\varphi(l-m)+\psi}\delta_{mm'}. \tag{2.14}$$

The operators $T(v_\pm)$ and $T(H)$ act on the basis $e^{lq}_m(\lambda)$ in the following way:

$$T(H)e^{lq}_m(\lambda)=\frac{m}{2}\,e^{lq}_m(\lambda),$$

$$T(v_+)e^{lq}_m(\lambda)=(-1)^{l-m}\sqrt{[l-m][l+m+1]\gamma}\;e^{lq}_{m+1}(\lambda),$$

$$T(v_-)e^{lq}_m(\lambda)=\sqrt{[l+m][l-m+1]\gamma}\;e^{lq}_{m-1}(\lambda), \tag{2.15}$$

where the symbol $[n]$ is the graded quantum symbol defined by

$$[n]=\frac{q^{-n/2}-(-1)^n q^{n/2}}{q^{-1/2}+q^{1/2}}, \tag{2.16}$$

and $\gamma=\cosh(\eta/4)/\sinh(2\eta)$. Note that although the symbol $[n]$ is not invariant under the exchange $q\leftrightarrow q^{-1}$, the relations (2.15) are invariant with respect to this substitution. Furthermore, it follows from the explicit formulas for $e^{lq}_m(\lambda)$, Ref. 11, that

$$e_m^{lq}(\lambda)=e_m^{lq^{-1}}(\lambda), \tag{2.17}$$

thus both quantum algebras $U_q(osp(1|2))$ and $U_{q^{-1}}(osp(1|2))$ define the same basis in the representation space $V^l(\lambda)$. In the next section, we shall see that this is no longer true when we consider tensor product of representation spaces.

C. Tensor product of representations and the $\mathscr{T}$ matrix

In the following, the representation T of class ϵ that acts in the representation space $V^l(\lambda)$ is denoted by T^l. The generators $v_\pm$ and H are represented in the tensor product of two irreducible representations of the same class, $T^{l_1}\otimes T^{l_2}$ by the operators

$$v_\pm^{\otimes}(1,2)=(T^{l_1}\otimes T^{l_2})(\Delta(v_\pm))=T^{l_1}(v_\pm)\otimes q^{T^{l_2}(H)}+q^{-T^{l_1}(H)}\otimes T^{l_2}(v_\pm), \tag{2.18}$$

$$H^{\otimes}(1,2)=(T^{l_1}\otimes T^{l_2})(\Delta(H))=T^{l_1}(H)\otimes T^{l_2}(\mathbf{1})+T^{l_1}(\mathbf{1})\otimes T^{l_2}(H). \tag{2.19}$$

The tensor product $T^{l_1}\otimes T^{l_2}$ is simply reducible. The standard basis $e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2)$ and reduced basis $e_m^{lq}(l_1,l_2,\lambda)$ of the representation space $V^{l_1}(\lambda_1)\otimes V^{l_2}(\lambda_2)$ are related by Clebsch–Gordan coefficients in the following way:

$$e_m^{lq}(l_1,l_2,\lambda)=\sum_{m_1,m_2}(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_q\, e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2), \tag{2.20}$$

or equivalently,

$$\sum_{l,m}(-1)^{(l-m)L}(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_q e_m^{lq}(l_1,l_2,\lambda)=(-1)^{(l_1-m_1)(l_2-m_2)}e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2), \tag{2.21}$$

where $m_1+m_2=m$, $L=l_1+l_2+l$, and l is an integer satisfying the conditions

$$|l_1-l_2|\leqslant l\leqslant l_1+l_2,\quad \lambda=L+\lambda_1+\lambda_2(\mathrm{mod}\ 2). \tag{2.22}$$

Both bases $e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2)$ and $e_m^{lq}(l_1,l_2,\lambda)$ are orthogonal but not positive definite, i.e., we have

$$(e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2),e_{m_1'}^{l_1q}(\lambda_1)\otimes e_{m_2'}^{l_2q}(\lambda_2))=(-1)^{(l_1-m_1+\lambda_1)(l_2-m_2+\lambda_2)}\delta_{m_1m_1'}\delta_{m_2m_2'}, \tag{2.23}$$

$$(e_m^{lq}(l_1,l_2,\lambda),e_{m'}^{l'q}(l_1,l_2,\lambda))=\delta_{ll'}\delta_{mm'}(-1)^{\varphi(l-m)+\psi}, \tag{2.24}$$

where

$$\varphi=L+\lambda_1+\varphi_2(\mathrm{mod}\ 2),\quad \psi=(L+\lambda_2)\lambda_1+\varphi_2L+\psi_1+\psi_2(\mathrm{mod}\ 2). \tag{2.25}$$

We may consider also another reduced basis built with the basis $e_m^{lq^{-1}}(\lambda)$:

$$e_m^{lq^{-1}}(l_1,l_2,\lambda)=\sum_{m_1,m_2}(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_{q^{-1}}e_{m_1}^{l_1q^{-1}}(\lambda_1)\otimes e_{m_2}^{l_2q^{-1}}(\lambda_2), \tag{2.26}$$

$$=\sum_{m_1,m_2}(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_{q^{-1}}e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2), \tag{2.27}$$

where the second line is deduced from Eq. (2.17). This basis is normalized in the same way (2.24) as the basis $e_m^{lq}(l_1,l_2,\lambda)$, but the two reduced bases are different because we have, in general,

$$(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_{q^{-1}}\neq(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_q. \tag{2.28}$$

Thus, there exists an automorphism $\mathscr{T}^q(l_1,l_2)$ of the space $V^{l_1}(\lambda_1)\otimes V^{l_2}(\lambda_2)$ that relates both bases:

$$\mathscr{T}^q(l_1,l_2)(e_m^{lq}(l_1,l_2,\lambda))=e_m^{lq^{-1}}(l_1,l_2,\lambda).$$

Using the orthogonality relations (2.20), (2.21), one can show that this operator has the following matrix form in the basis $e_m^{lq}(l_1,l_2,\lambda)$:

$$[\mathscr{T}^q(l_1,l_2)]_{l'm',lm}=(-1)^{(l_1+l_2+l')(l'-m')}\sum_{m_1m_2}(-1)^{(l_1-m_1)(l_2-m_2)}\times(l_1m_1\lambda_1,l_2m_2\lambda_2|l'm'\lambda')_q(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_{q^{-1}}. \tag{2.29}$$

From this expression it follows that $[\mathscr{T}^q(l_1,l_2)]^{-1}=\mathscr{T}^{q^{-1}}(l_1,l_2)$. Note also that the matrix $\mathscr{T}^q(l_1,l_2)$ is diagonal in the indices m',m but it is not diagonal in the indices l',l. The symmetries of Clebsch–Gordan coefficients imply the following properties for the matrix elements $[\mathscr{T}^q(l_1,l_2)]_{l'm',lm}$:

$$[\mathscr{T}^q(l_1,l_2)]_{lm,l'm'}=(-1)^{(l_1+l_2+l'+l+m)(l'+l)}[\mathscr{T}^{q^{-1}}(l_1,l_2)]_{l'm',lm}, \tag{2.30}$$

$$[\mathscr{T}^{q^{-1}}(l_1,l_2)]_{l'm',lm}=(-1)^{m(l'+l)}(-1)^{(1/2)l(l-1)+(1/2)l'(l'-1)}[\mathscr{T}^q(l_1,l_2)]_{l'-m',l-m}, \tag{2.31}$$

$$[\mathscr{T}^q(l_2,l_1)]_{l'm',lm}=(-1)^{(l_1+l_2+\lambda_1+\lambda_2)(l'+l)}(-1)^{(1/2)l(l-1)+(1/2)l'(l'-1)}[\mathscr{T}^{q^{-1}}(l_1,l_2)]_{l'm',lm} \tag{2.32}$$

D. The permutation operator τ and the universal R matrix

We shall now consider the tensor product of the previous irreducible representations in reverse order, namely $T^{l_2}\otimes T^{l_1}$. For this we introduce the permutation operator τ acting in any graded tensor product space as follows:

$$\tau(a\otimes b)=(-1)^{\deg(a)\deg(b)}(b\otimes a). \tag{2.33}$$

Therefore, for the quantum superalgebra $U_q(\mathrm{osp}(1|2))$ we have

$$\tau\Delta^q(v_\pm)=\Delta^{q^{-1}}(v_\pm)=R^q\Delta^q(v_\pm)(R^q)^{-1}, \tag{2.34}$$

$$\tau\Delta^q(H)=\Delta^{q^{-1}}(H)=R^q\Delta^q(H)(R^q)^{-1}, \tag{2.35}$$

which means that the action of the operator τ for the coproduct of generators of $U_q(\mathrm{osp}(1|2))$ coïncides with the action of the universal R matrix. The action of the operator τ in the tensor product of representations $T^{l_1}\otimes T^{l_2}$ reverse the order of the tensor product, and we have

$$\tau(e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2))=(-1)^{(l_1-m_1+\lambda_1)(l_2-m_2+\lambda_2)}e_{m_2}^{l_2q}(\lambda_2)\otimes e_{m_1}^{l_1q}(\lambda_1), \tag{2.36}$$

and

$$\tau v_{\pm}^{\otimes q}(l_1,l_2)\tau^{-1}=v_{\pm}^{\otimes q^{-1}}(l_2,l_1), \tag{2.37}$$

$$\tau H^{\otimes q}(l_1,l_2)\tau^{-1}=H^{\otimes q^{-1}}(l_2,l_1). \tag{2.38}$$

Using the following symmetry property of Clebsch–Gordan coefficients,[1]

$$\begin{aligned}(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_q=&(-1)^{(l_1-m_1+\lambda_1)(l_2-m_2+\lambda_2)}(-1)^{(l_1+l_2+l)(\lambda_1+\lambda_2)+\lambda_1\lambda_2}\\&\times(-1)^{(1/2)(l_1+l_2-l)(l_1+l_2-l+1)}(l_2m_2\lambda_2,l_1m_1\lambda_1|lm\lambda)_{q^{-1}},\end{aligned} \tag{2.39}$$

one can check that the action of τ on the reduced basis reads as

$$\tau(e_m^{lq}(l_1,l_2,\lambda))=(-1)^{(l_1+l_2+l+\lambda_1)(l_1+l_2+l+\lambda_2)}(-1)^{(1/2)(l_1+l_2-l)(l_1+l_2-l-1)}e_m^{lq^{-1}}(l_2,l_1,\lambda). \tag{2.40}$$

Thus the operator τ that relates the bases $e_m^{lq}(l_1,l_2,\lambda)$ and $e_m^{lq^{-1}}(l_2,l_1,\lambda)$ is a representation isomorphism:

$$\tau:(T^{l_1}\otimes T^{l_2})(\mathrm{U}_q(\mathrm{osp}(1|2)))\to(T^{l_2}\otimes T^{l_1})(\mathrm{U}_{q^{-1}}(\mathrm{osp}(1|2))). \tag{2.41}$$

The operator R^q in the representation $T^{l_1}\otimes T^{l_2}$ will be denoted $R^q(l_1,l_2)\equiv R^q(1,2)$. In such a representation its action is

$$R^q(1,2)v_{\pm}^{\otimes q}(l_1,l_2)R^q(1,2)^{-1}=v_{\pm}^{\otimes q^{-1}}(l_1,l_2), \tag{2.42}$$

$$R^q(1,2)H^{\otimes q}(l_1,l_2)R^q(1,2)^{-1}=H^{\otimes q^{-1}}(l_1,l_2). \tag{2.43}$$

Since $R^q(1,2)$ does not exchange the two spaces, in this case the action of the operator τ does not coïncide with the action of $R^q(1,2)$.

1. Matrix elements of $R^q(1,2)$ in the tensor product basis

Using relations (2.15) one can show that, in the standard basis $e_{m_1}^{l_1q}(\lambda_1)\otimes e_{m_2}^{l_2q}(\lambda_2)$, the operator $R^q(1,2)$ has the following matrix form:

$$\begin{aligned}[\mathbb{R}^q(1,2)]_{m_1m_2,n_1n_2}=&(-1)^{k(\lambda_1+1)+(1/2)k(k-1)}q^{(1/4)k(3k+1)}q^{(1/2)k(n_1-n_2)+(n_1+k)(n_2-k)}\\&\times\frac{(1+q^{-1})^k}{[k]!}\left(\frac{[l_1-n_1]![l_1+n_1+k]![l_2+n_2]![l_2-n_2+k]!}{[l_1+n_1]![l_1-n_1-k]![l_2-n_2]![l_2+n_2-k]!}\right)^{1/2}\\&\times\delta_{m_1,n_1+k}\delta_{m_2,n_2-k}.\end{aligned} \tag{2.44}$$

For the simplest case when $l_1=l_2=1$, $R^q(1,2)_{m_1m_2,n_1n_2}$ has the following form:

$$[\mathscr{R}^q(1,1)]_{m_1m_2,n_1n_2}=\begin{pmatrix} q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & a & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & q^{-1} & 0 & b & 0 & e & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & c & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & d & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q \end{pmatrix}, \tag{2.45}$$

where

$$a=d=(-1)^{\lambda_1}(q-q^{-1}), \quad c=b=q^{-1/2}a, \quad e=(-1)^{\lambda_1}(1+q^{-1})a. \tag{2.46}$$

When $\lambda_1=0$, this matrix coïncides with the corresponding matrix given in Ref. 8.

2. Matrix elements of $R^q(1,2)$ and $R^{q^{-1}}(1,2)$ in the reduced bases

Relations (2.42), (2.43) suggest that the vector $e_m^{lq^{-1}}(l_1,l_2,\lambda)$ is proportional to the vector $R^q(1,2)e_m^{lq}(l_1,l_2,\lambda)$. In the Appendix we prove that indeed we have

$$R^q(1,2)e_m^{lq}(l_1,l_2,\lambda)=q^{(1/2)(l(l+1)-l_1(l_1+1)-l_2(l_2+1))}e_m^{lq^{-1}}(l_1,l_2,\lambda). \tag{2.47}$$

From this relation it follows that

$$[R^q(1,2)]^{-1}=R^{q^{-1}}(1,2). \tag{2.48}$$

Furthermore, formula (2.47) allows one to calculate the matrix elements of the operators $R^q(1,2)$ and $R^{q^{-1}}(1,2)$ in the reduced bases $e_m^{lq}(l_1,l_2,\lambda)$ and $e_m^{lq^{-1}}(l_1,l_2,\lambda)$. Let us introduce the following notation: $\mathbb{R}_e^q(\mathbb{R}_e^{q^{-1}})$ is the matrix of the operator $R^q(1,2)(R^{q^{-1}}(1,2))$ in the basis $e_m^{lq}(l_1,l_2,\lambda)$, $\mathbb{R}_{\underline{e}}^q(\mathbb{R}_{\underline{e}}^{q^{-1}})$ is the matrix of the operator $R^q(1,2)(R^{q^{-1}}(1,2))$ in the basis $e_m^{lq^{-1}}(l_1,l_2,\lambda)$, and let $\mathbb{D}^q$ be a diagonal matrix of the form

$$[\mathbb{D}^q]_{lm,l'm'}=q^{(1/2)(l(l+1)-l_1(l_1+1)-l_2(l_2+1))}\delta_{l,l'}\delta_{m,m'}. \tag{2.49}$$

Using relation (2.47) and the orthogonality relations (2.20), (2.21) for sq-CGc, one can derive the following matrix relations:

$$\mathbb{R}_e^q=\mathscr{T}^q\mathbb{D}^q, \quad \mathbb{R}_e^{q^{-1}}=\mathbb{D}^{q^{-1}}\mathscr{T}^{q^{-1}}, \tag{2.50}$$

$$\mathbb{R}_{\underline{e}}^q=\mathbb{D}^q\mathscr{T}^q, \quad \mathbb{R}_{\underline{e}}^{q^{-1}}=\mathscr{T}^{q^{-1}}\mathbb{D}^{q^{-1}}, \tag{2.51}$$

where $\mathscr{T}^q$ is the matrix given by the formula (2.29). In particular, for the first relation we have

$$[\mathbb{R}_e^q]_{l'm',lm}=q^{(1/2)(l(l+1)-l_1(l_1+1)-l_2(l_2+1))}(-1)^{(l_1+l_2+l')(l'-m')}\sum_{m_1m_2}(-1)^{(l_1-m_1)(l_2-m_2)}$$
$$\times(l_1m_1\lambda_1,l_2m_2\lambda_2|l'm'\lambda')_q(l_1m_1\lambda_1,l_2m_2\lambda_2|lm\lambda)_{q^{-1}}, \tag{2.52}$$

i.e., the matrix elements of the operator $R^q(1,2)$ can be expressed in terms of sq-CGc. The values of the sq-CGc for an arbitrary tensor product of irreducible representations can be calculated from the analytical expression given in Refs. 2, 10, and 11, therefore the matrix elements of $R^q(1,2)$ in any tensor product of irreducible representations can be deduced from formula (2.52). For example, calculating the values of corresponding sq-CGc for $l_1=l_2=1$, we get from formula (2.52),

$$[\mathbb{R}^q_e]_{l'm',lm}=\begin{pmatrix} q^{-2}a & 0 & 0 & 0 & 0 & 0 & qc & 0 & 0 \\ 0 & q^{-1}b & 0 & 0 & 0 & qd & 0 & 0 & 0 \\ 0 & 0 & q^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q^{-1}b & 0 & 0 & 0 & -qd & 0 \\ 0 & 0 & 0 & 0 & q & 0 & 0 & 0 & 0 \\ 0 & -q^{-1}d & 0 & 0 & 0 & qb & 0 & 0 & 0 \\ -q^{-2}c & 0 & 0 & 0 & 0 & 0 & qa & 0 & 0 \\ 0 & 0 & 0 & q^{-1}d & 0 & 0 & 0 & qb & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q \end{pmatrix}, \tag{2.53}$$

where

$$a=\frac{q}{q^2-q+1}, \quad b=\frac{2q}{q^2+1}, \quad c=\frac{q^3-q^2+q-1}{\sqrt{q^2+1}(q^2-q+1)}, \quad d=(-1)^{\lambda_1}\frac{q^2-1}{q^2+1}. \tag{2.54}$$

From expressions (2.42), (2.43), and (2.47), it follows that the operator $R^q(1,2)$ yields an isomorphism of two representations:

$$R^q(1,2):(T^{l_1}\otimes T^{l_2})(U_q(osp(1|2)))\rightarrow(T^{l_1}\otimes T^{l_2})(U_{q^{-1}}(osp(1|2))). \tag{2.55}$$

E. The operator $\mathscr{R}^q$

Finally, in the following we will use the operator $\mathscr{R}^q$ that is a composition of the operators τ and R^q:

$$\mathscr{R}^q=\tau\circ R^q. \tag{2.56}$$

Properties of the operators τ and R^q, in particular, the Yang–Baxter equation (2.12) satisfied by R^q, implies that $\mathscr{R}^q$ satisfies the following relation:

$$(\mathscr{R}^q\otimes id)(id\otimes\mathscr{R}^q)(\mathscr{R}^q\otimes id)=(id\otimes\mathscr{R}^q)(\mathscr{R}^q\otimes id)(id\otimes\mathscr{R}^q). \tag{2.57}$$

Further, from relations (2.10) we get the relations

$$(\mathscr{R}^q\otimes id)(id\otimes\mathscr{R}^q)=\tau(12,3)R^{q}{}_{12,3}\equiv\mathscr{R}^{q}{}^{12,3}, \tag{2.58}$$

$$(id\otimes\mathscr{R}^q)(\mathscr{R}^q\otimes id)=\tau(1,23)R^{q}{}_{1,23}\equiv\mathscr{R}^{q}{}^{1,23}, \tag{2.59}$$

where

$$\tau(12,3)=(\tau\otimes id)(id\otimes\tau), \quad \tau(1,23)=(id\otimes\tau)(\tau\otimes id). \tag{2.60}$$

In the tensor product $T^{l_1}\otimes T^{l_2}$ the operator $\mathscr{R}^q$ is represented by the operator

$$\mathscr{R}^q(1,2)=\tau\circ R^q(1,2),$$

and formulas (2.58), (2.59) take the form

$$(\mathscr{R}^q(1,3)\otimes id(2))(id(1)\otimes\mathscr{R}^q(2,3))=\mathscr{R}^{q12,3}(12,3), \tag{2.61}$$

$$(id(2)\otimes\mathscr{R}^q(1,3))(\mathscr{R}^q(1,2)\otimes id(3))=\mathscr{R}^{q1,23}(1,23). \tag{2.62}$$

Using relations (2.40), (2.47) we derive immediately the action of $\mathscr{R}^q(1,2)$ on the reduced basis,

$$\begin{aligned}\mathscr{R}^q(1,2)e^{lq}_m(l_1,l_2,\lambda)=&(-1)^{(l_1+l_2+l+\lambda_1)(l_1+l_2+l+\lambda_2)}\\&\times(-1)^{(1/2)(l_1+l_2-l)(l_1+l_2-l-1)}\\&\times q^{(1/2)[l(l+1)-l_1(l_1+1)-l_2(l_2+1)]}e^{lq}_m(l_2,l_1,\lambda),\end{aligned} \tag{2.63}$$

and its action on the operators $v^{\otimes q}_{\pm}(l_1,l_2)$, $H^{\otimes q}(l_1,l_2)$ reads as

$$\mathscr{R}^q(1,2)v^{\otimes q}_{\pm}(l_1,l_2)\mathscr{R}^q(1,2)^{-1}=v^{\otimes q}_{\pm}(l_2,l_1), \tag{2.64}$$

$$\mathscr{R}^q(1,2)H^{\otimes q}(l_1,l_2)\mathscr{R}^q(1,2)^{-1}=H^{\otimes q}(l_2,l_1), \tag{2.65}$$

where

$$\mathscr{R}^q(1,2)^{-1}=R^{q^{-1}}(1,2)\circ\tau. \tag{2.66}$$

The operator $\mathscr{R}^q(1,2)$ defines then an isomorphism of representations,

$$\mathscr{R}^q(1,2):(T^{l_1}\otimes T^{l_2})(U_q(osp(1|2)))\rightarrow(T^{l_2}\otimes T^{l_1})(U_q(osp(1|2))). \tag{2.67}$$

The universal R matrix is the appropriate tool for obtaining numerous relations between Racah coefficients and $sq6\text{-}j$ symbols. In the following section we shall first use it to derive the Racah sum rule for $sq6\text{-}j$ symbols.

III. RACAH SUM RULE FOR THE sq-6j SYMBOLS

In order to derive the Racah sum rule, we will consider the action of the operator $\mathscr{R}^q$ on the reduced bases of the tensor product of three irreducible representations of the same class $T^{l_1}\otimes T^{l_2}\otimes T^{l_3}$. The operators $H,v_{\pm}$ are represented in this tensor product by

$$v^{\otimes}_{\pm}(1,2,3)=(T^{l_1}\otimes T^{l_2}\otimes T^{l_3})((\Delta\otimes id)\Delta(v_{\pm})), \tag{3.1}$$

$$H^{\otimes}(1,2,3)=(T^{l_1}\otimes T^{l_2}\otimes T^{l_3})((\Delta\otimes id)\Delta(H)). \tag{3.2}$$

The reduction of the tensor product $V^{l_1}(\lambda_1)\otimes V^{l_2}(\lambda_2)\otimes V^{l_3}(\lambda_3)$ of representation spaces can be done, as in the classical case, in two different schemes:

$$T^l\subset((T^{l_1}\otimes T^{l_2})\otimes T^{l_3}),\quad T^l\subset(T^{l_1}\otimes(T^{l_2}\otimes T^{l_3})). \tag{3.3}$$

The bases $e^{lq}_m(l_1,l_{23},\lambda)$ and $e^{lq}_m(l_{12},l_3,\lambda)$ corresponding to these reduction schemes are orthogonal and normalized in the following way:

$$(e^{lq}_m(l_{12},l_3,\lambda),e^{l'q}_{m'}(l'_{12},l_3,\lambda))=(-1)^{\varphi_{12,3}(l-m)+\psi_{12,3}}\delta_{ll'}\delta_{mm'}\delta_{l_{12}l'_{12}}\equiv g_{ll_{12}m,l'l'_{12}m'}, \tag{3.4}$$

$$(e^{lq}_m(l_1,l_{23},\lambda),e^{l'q}_{m'}(l_1,l'_{23},\lambda))=(-1)^{\varphi_{1,23}(l-m)+\psi_{1,23}}\delta_{ll'}\delta_{mm'}\delta_{l_{23}l'_{23}}\equiv g_{ll_{23}m,l'l'_{23}m'}, \tag{3.5}$$

where

$$\varphi_{1,23}=\varphi_{12,3}=\mathscr{S}+\lambda_1+\lambda_2+\varphi_3(\mathrm{mod}\ 2), \tag{3.6}$$

$$\psi_{12,3}=(l_1+l_2+l_{12})(\mathscr{S}+1)+(\lambda_1+\lambda_3+\varphi_2)\mathscr{S}+\sum_{i<j}\lambda_i\lambda_j+\sum_{i=1}^{3}\psi_i\ (\mathrm{mod}\ 2), \tag{3.7}$$

$$\psi_{1,23}=(l_2+l_3+l_{23})(\mathscr{S}+1)+(\lambda_1+\lambda_3+\varphi_2)\mathscr{S}+\sum_{i<j}\lambda_i\lambda_j+\sum_{i=1}^{3}\psi_i\ (\mathrm{mod}\ 2), \tag{3.8}$$

with

$$\mathscr{S}=l_1+l_2+l_3+l,\quad \lambda=\sum_{i=1}^{3}(\lambda_i+l_i)+l. \tag{3.9}$$

The sqRacah coefficients $U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)$ of the quantum superalgebra $\mathrm{U}_q(\mathrm{osp}(1|2))$ are defined, in the standard way, as the coefficients that relate reduced bases in two different reduction schemes,[2]

$$e_m^{lq}(l_{12},l_3,\lambda)=\sum_{l_{23}}(-1)^{(l_2+l_3+l_{23})(\mathscr{S}+1)}U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)e_m^{lq}(l_1,l_{23},\lambda), \tag{3.10}$$

or equivalently,

$$e_m^{lq}(l_1,l_{23},\lambda)=\sum_{l_{12}}(-1)^{(l_1+l_2+l_{12})(\mathscr{S}+1)}U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)e_m^{lq}(l_{12},l_3,\lambda). \tag{3.11}$$

We have the following relation between $sqRc$ and $sq6-j$ symbols:[2]

$$\begin{aligned}U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)=&(-1)^{(\lambda_2+1)(l_2+l+l_{23}+l_{12})}(-1)^{(l_1+l_2+l_3)(l+1)}(-1)^{l_{12}(l_1+l_{23}+l)}\\&\times(-1)^{l_{23}(l_{12}+l_3+l)}(-1)^{\mathscr{S}(\mathscr{S}-1)/2}\sqrt{[2l_{12}+1][2l_{23}+1]}\\&\times\begin{Bmatrix}l_1 & l_2 & l_{12}\\ l_3 & l & l_{23}\end{Bmatrix}_q^s.\end{aligned} \tag{3.12}$$

Let us consider the matrix $W_{l'm',lm}$ of the operator $id(1)\otimes\mathscr{R}^q(3,2)$ in the bases $e_m^{lq}(l_{12},l_3,\lambda)$ and $e_m^{lq}(l_{13},l_2,\lambda)$,

$$W_{l'm',lm}=g_{l'l'_{12}m',l''l''_{12}m''}(e_{m''}^{l''q}(l''_{12},l_3,\lambda),id(1)\otimes\mathscr{R}^q(3,2)e_m^{lq}(l_{13},l_2,\lambda)) \tag{3.13}$$

Using relations (2.20), (3.10), and (2.63), one obtains

$$\begin{aligned}&id(1)\otimes\mathscr{R}^q(3,2)e_m^{lq}(l_{13},l_2,\lambda)\\&=\sum_{l_{12},l_{23}}(-1)^{(l_1+l_3+l_{12}+l_{23})(\mathscr{S}+1)}(-1)^{(l_2+l_3+l_{23}+\lambda_2)(l_2+l_3+l_{23}+\lambda_3)}\\&\quad\times(-1)^{(1/2)(l_2+l_3-l_{23})(l_2+l_3-l_{23}-1)}q^{(1/2)[l_{23}(l_{23}+1)-l_2(l_2+1)-l_3(l_3+1)]}\\&\quad\times U^s(l_1,l_3,l_2,l,l_{13},l_{23},q)U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)e_m^{lq}(l_{12},l_3,\lambda)\end{aligned} \tag{3.14}$$

Therefore, the matrix $W_{l'm',lm}$ has the form

$$W_{l'm',lm}=\sum_{l_{23}}(-1)^{(l_1+l_3+l_{12}+l_{23})(\mathscr{S}+1)}(-1)^{(l_2+l_3+l_{23}+\lambda_2)(l_2+l_3+l_{23}+\lambda_3)}$$
$$\times(-1)^{(1/2)(l_2+l_3-l_{23})(l_2+l_3-l_{23}-1)}q^{(1/2)[l_{23}(l_{23}+1)-l_2(l_2+1)-l_3(l_3+1)]}$$
$$\times U^s(l_1,l_3,l_2,l,l_{13},l_{23},q)U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)\delta_{l',l}\delta_{m',m}. \tag{3.15}$$

It is diagonal and it does not depend on m.

On the other hand, we have

$$id(1)\otimes\mathscr{R}^q(3,2)e^{lq}_m(l_{13},l_2,\lambda)=([\mathscr{R}^q(1,2)]^{-1}\otimes id(3))(\mathscr{R}^q(1,2)\otimes id(3))$$
$$\times(id(1)\otimes\mathscr{R}^q(3,2))e^{lq}_m(l_{13},l_2,\lambda), \tag{3.16}$$
$$=(-1)^{(l_2+l_{13}+l+\lambda_2)(l_2+l_{13}+l+\lambda_{13})}$$
$$\times(-1)^{(1/2)(l_2+l_{13}-l)(l_2+l_{13}-l-1)}q^{(1/2)[l(l+1)-l_2(l_2+1)-l_{13}(l_{13}+1)]}$$
$$\times([\mathscr{R}^q(1,2)]^{-1}\otimes id(3))e^{lq}_m(l_2,l_{13},\lambda), \tag{3.17}$$

where we have used relations (2.61), (2.63). Applying further relations (2.20), (3.10), and (2.63), we find that

$$([\mathscr{R}^q(1,2)]^{-1}\otimes id(3))e^{lq}_m(l_2,l_{13},\lambda)$$
$$=\sum_{l_{12}}(-1)^{(l_1+l_2+l_{12})(\mathscr{S}+1)}(-1)^{(l_1+l_2+l_{12}+\lambda_1)(l_1+l_2+l_{12}+\lambda_2)}$$
$$\times(-1)^{(1/2)(l_1+l_2-l_{12})(l_1+l_2-l_{12}-1)}q^{-(1/2)[l_{12}(l_{12}+1)-l_1(l_1+1)-l_2(l_2+1)]}$$
$$\times U^s(l_2,l_1,l_3,l,l_{12},l_{13},q)e^{lq}_m(l_{12},l_3,\lambda). \tag{3.18}$$

Substitution of this relation into (3.16) and of the resulting formula into relation (3.13) yields another expression for the matrix $W_{l'm',lm}$,

$$W_{l'm',lm}=(-1)^{(l_2+l_{23}+l+\lambda_2)(l_2+l_{13}+l+\lambda_{13})}(-1)^{(l_1+l_2+l_{12}+\lambda_1)(l_1+l_2+l_{12}+\lambda_2)}(-1)^{(l_1+l_2+l_{12})(\mathscr{S}+1)}$$
$$\times(-1)^{(1/2)(l_2+l_{13}-l)(l_2+l_{13}-l-1)}(-1)^{(1/2)(l_1+l_2-l_{12})(l_1+l_2-l_{12}-1)}$$
$$\times q^{(1/2)[l(l+1)-l_2(l_2+1)-l_{13}(l_{13}+1)]}q^{-(1/2)[l_{12}(l_{12}+1)-l_1(l_1+1)-l_2(l_2+1)]}$$
$$\times U^s(l_2,l_1,l_3,l,l_{12},l_{13},q)\delta_{l',l}\delta_{m',m}. \tag{3.19}$$

Comparison of this expression with relation (3.15) furnishes the Racah sum rule for $U_q(\mathrm{osp}(1|2))$ Racah coefficients:

$$\sum_{l_{23}}(-1)^{(l_2+l_3+l_{23})(\mathscr{S}+1)}(-1)^{(l_2+l_3+l_{23}+\lambda_2)(l_2+l_3+l_{23}+\lambda_3)}$$
$$\times(-1)^{(1/2)(l_2+l_3-l_{23})(l_2+l_3-l_{23}-1)}q^{(1/2)[l_{23}(l_{23}+1)-l_2(l_2+1)-l_3(l_3+1)]}$$
$$\times U^s(l_1,l_3,l_2,l,l_{13},l_{23},q)U^s(l_1,l_2,l_3,l,l_{12},l_{23},q)$$
$$=(-1)^{(l_2+l_{23}+l+\lambda_2)(\mathscr{S}+\lambda_1+\lambda_3)}(-1)^{(l_1+l_2+l_{12}+\lambda_1)(l_1+l_2+l_{12}+\lambda_2)}(-1)^{(1/2)(l_2+l_{13}-l)(l_2+l_{13}-l-1)}$$
$$\times(-1)^{(1/2)(l_1+l_2-l_{12})(l_1+l_2-l_{12}-1)}q^{(1/2)[l(l+1)+l_1(l_1+1)-l_{12}(l_{12}+1)-l_{13}(l_{13}+1)]}$$
$$\times U^s(l_2,l_1,l_3,l,l_{12},l_{13},q). \tag{3.20}$$

Substituting in this expression relation (3.12) that relates sqRacah coefficients and $sq6-j$ symbols, we derive the Racah sum rule satisfied by the latter:

$$\sum_{l_{23}} (-1)^{(l_1+l_2+l_{13}+l_{23})(\mathscr{L}+1)+\mathscr{L}l}(-1)^{(l_2+l_3+l_{23})l_{23}+(l_1+l_2+l_{12})l_{12}+(l_1+l_3+l_{13})l_{13}}(-1)^{\mathscr{L}(\mathscr{L}+1)/2}$$
$$\times(-1)^{(1/2)(l_2+l_3-l_{23})(l_2+l_3-l_{23}-1)}(-1)^{(1/2)(l_1+l_2-l_{12})(l_1+l_2-l_{12}-1)}$$
$$\times(-1)^{(1/2)(l_2+l_{13}-l)(l_2+l_{13}-l-1)}q^{(1/2)[l_{23}(l_{23}+1)+l_{12}(l_{12}+1)+l_{13}(l_{13}+1)]}[2l_{23}+1]$$
$$\times\begin{Bmatrix} l_1 & l_2 & l_{12} \\ l_3 & l & l_{23} \end{Bmatrix}_q^s \begin{Bmatrix} l_1 & l_3 & l_{13} \\ l_2 & l & l_{23} \end{Bmatrix}_q^s = q^{(1/2)[l(l+1)+l_1(l_1+1)+l_2(l_2+1)+l_3(l_3+1)]} \begin{Bmatrix} l_2 & l_1 & l_{12} \\ l_3 & l & l_{13} \end{Bmatrix}_q^s . \tag{3.21}$$

Using the orthogonality relation for the $sq6-j$ symbols,

$$\sum_{l_{12}} (-1)^{(l_1+l_2+l_{12})(\mathscr{L}+1)}[2l_{12}+1][2l_{23}+1] \begin{Bmatrix} l_1 & l_2 & l_{12} \\ l_3 & l & l_{23} \end{Bmatrix}_q^s \begin{Bmatrix} l_1 & l_2 & l_{12} \\ l_3 & l & l'_{23} \end{Bmatrix}_q^s$$
$$= (-1)^{(l_2+l_3+l_{23})(\mathscr{L}+1)} \delta_{l_{23},l'_{23}}, \tag{3.22}$$

one can rewrite the Racah sum rule (3.21) in the following form:

$$\sum_{l_{13}} (-1)^{\Theta} q^{-(1/2)[l_{23}(l_{23}+1)+l_{12}(l_{12}+1)+l_{13}(l_{13}+1)]}[2l_{13}+1] \begin{Bmatrix} l_2 & l_1 & l_{12} \\ l_3 & l & l_{13} \end{Bmatrix}_q^s \begin{Bmatrix} l_1 & l_3 & l_{13} \\ l_2 & l & l_{23} \end{Bmatrix}_q^s$$
$$= q^{-(1/2)[l(l+1)+l_1(l_1+1)+l_2(l_2+1)+l_3(l_3+1)]} \begin{Bmatrix} l_1 & l_2 & l_{12} \\ l_3 & l & l_{23} \end{Bmatrix}_q^s , \tag{3.23}$$

where the phase argument Θ is

$$\Theta = \tfrac{1}{2}(l_2+l_3-l_{23})(l_2+l_3-l_{23}-1)+\tfrac{1}{2}(l_1+l_2-l_{12})(l_1+l_2-l_{12}-1)$$
$$+\tfrac{1}{2}(l_2+l_{13}-l)(l_2+l_{13}-l-1)+\tfrac{1}{2}\mathscr{L}(\mathscr{L}+1)+\mathscr{L}l+(l_2+l_3+l_{23})l_{23}$$
$$+(l_1+l_2+l_{12})l_{12}+(l_1+l_3+l_{13})l_{13}. \tag{3.24}$$

In the limit $q\to 1$ this formula becomes the Racah sum rule for the $6-j$ symbols for the superalgebra osp(1|2) and it takes the form given in Ref. 12.

If $l_{12}=0$ and $l_1=l_2$, $l_3=l$, then the Racah sum rule (3.23) leads to the summation formula

$$\sum_{l_{13}} (-1)^{(l_1+l_3)(l_{13}+l'_{13})+l'_{13}} q^{-(1/2)[l_{13}(l_{13}+1)+l'_{13}(l'_{13}+1)]}[2l_{13}+1] \begin{Bmatrix} l_1 & l_3 & l_{13} \\ l_1 & l_3 & l'_{13} \end{Bmatrix}_q^s$$
$$= q^{-(1/2)[l_1(l_1+1)+l_3(l_3+1)]}. \tag{3.25}$$

IV. BIEDENHARN–ELLIOTT IDENTITY FOR THE QUANTUM SUPERALGEBRA $U_q(\mathrm{osp}(1|2))$

The tensor product $T^{l_1}\otimes T^{l_2}\otimes T^{l_3}\otimes T^{l_4}$ of four irreducible representations of the same class can be reduced into irreducible representations according to one of the following five reduction schemes:

$$[(T^{l_1}\otimes T^{l_2})\otimes T^{l_3}]\otimes T^{l_4}, \tag{4.1}$$

$$[T^{l_1}\otimes (T^{l_2}\otimes T^{l_3})]\otimes T^{l_4}, \tag{4.2}$$

$$T^{l_1}\otimes [(T^{l_2}\otimes T^{l_3})\otimes T^{l_4}], \tag{4.3}$$

$$T^{l_1}\otimes [T^{l_2}\otimes (T^{l_3}\otimes T^{l_4})], \tag{4.4}$$

$$(T^{l_1}\otimes T^{l_2})\otimes (T^{l_3}\otimes T^{l_4}). \tag{4.5}$$

The Biedenharn–Elliott identity is concerned with the relation between the schemes (4.1) and (4.4). One can go from one to the other either by the chain

$$(4.1)\rightarrow(4.2)\rightarrow(4.3)\rightarrow(4.4), \tag{4.6}$$

or by the chain

$$(4.1)\rightarrow(4.5)\rightarrow(4.4), \tag{4.7}$$

each step involving sqRacah coefficients. Since the overall matrices that connect the two reduction schemes, (4.1) and (4.4), must be independent of the chain, there must exist a relation between combinations of products of three sqRacah coefficients for the first chain and combinations of products of two sqRacah coefficients for the second chain. This is precisely the Biedenharn–Elliott identity. Writing down the equality of corresponding matrix elements, the Biedenharn–Elliott identity for sqRacah coefficients reads as

$$\begin{aligned}
&\sum_{l_{23}} (-1)^{(l_2+l_3+l_{23})(l_2+l_3+l_4+l_{234}+1)}(-1)^{(l_1+l_2+l_{12})(l_1+l_2+l_3+l_{123}+1)}\\
&\quad\times(-1)^{(l_1+l_{23}+l_{123})(l_1+l_{23}+l_4+l+1)}U^s(l_1,l_2,l_3,l_{123},l_{12},l_{23},q)U^s(l_1,l_{23},l_4,l,l_{123},l_{234},q)\\
&\quad\times U^s(l_2,l_3,l_4,l_{234},l_{23},l_{34},q)\\
&\quad=(-1)^{(l_1+l_2+l_{12})(l_1+l_2+l_{34}+l+1)}(-1)^{(l_{12}+l_3+l_{123})(l_{12}+l_3+l_4+l+1)}\\
&\qquad\times U^s(l_{12},l_3,l_4,l,l_{123},l_{34},q)U^s(l_1,l_2,l_{34},l,l_{12},l_{234},q).
\end{aligned} \tag{4.8}$$

For $sq6$-j symbols it is rewritten as

$$\begin{aligned}
&\sum_{l_{13}} (-1)^{\Theta}[2l_{23}+1]\begin{Bmatrix} l_1 & l_2 & l_{12}\\ l_3 & l_{123} & l_{23}\end{Bmatrix}_q^s\begin{Bmatrix} l_1 & l_{23} & l_{123}\\ l_4 & l & l_{234}\end{Bmatrix}_q^s\begin{Bmatrix} l_2 & l_3 & l_{23}\\ l_4 & l_{234} & l_{34}\end{Bmatrix}_q^s\\
&\quad=\begin{Bmatrix} l_{12} & l_3 & l_{123}\\ l_4 & l & l_{34}\end{Bmatrix}_q^s\begin{Bmatrix} l_1 & l_2 & l_{12}\\ l_{34} & l & l_{234}\end{Bmatrix}_q^s,
\end{aligned} \tag{4.9}$$

where the phase argument Θ is given by

$$\begin{aligned}
\Theta=&\frac{1}{2}\left(\sum_{all} l_i\right)\left(\sum_{all} l_i-1\right)+l_{23}(l+l_{12}+l_{34})+(l_4+l_{234})(l_3+l+l_{123}+l_{34})\\
&+(l_3+l_{12}+l_{123})(l_3+l_2+l+l_{12}+l_{34})+(l_1+l_2+l_{12})(l_1+l_3+l_4+l_{34}+l+l_{234})\\
&+l_2+l_4+l+l_{34}+l_{123}+l_{234},
\end{aligned} \tag{4.10}$$

where $\Sigma_{all}\, l_i$ means summation over all 10 arguments appearing in $sq6\text{-}j$ symbols. Similarly, as in the cases of $su(2)$ and $u_q(su(2))$ the three-term recurrence relation for $sq6\text{-}j$ symbols follows from the Biedenharn–Elliott identity (4.9). Putting $l=1$ into this identity and using the known analytical expressions of the $sq6\text{-}j$ symbols with one argument equal to 1 (see the table in Ref. 3), we derive a recurrence relation that can be written, after a change in the notation,

$$A\begin{Bmatrix} a & b & e \\ c & d & f-1 \end{Bmatrix}_q^s + B\begin{Bmatrix} a & b & e \\ c & d & f \end{Bmatrix}_q^s + C\begin{Bmatrix} a & b & e \\ c & d & f+1 \end{Bmatrix}_q^s = 0, \tag{4.11}$$

where

$$A=(-1)^{a+d+f+1}[2][2f+2]([a+d+f+1][a-d+f][a+d-f+1][-a+d+f][b+c+f+1] \times[b-c+f][b+c-f+1][-b+c+f])^{1/2}, \tag{4.12}$$

$$B=(-1)^{a+c+f+1}[2f+1]([a-d+f][a+d-f]+(-1)^{a+b+f}[-a+d+f][a+d+f+2]) \times([b+c-f][b-c+f]+(-1)^{b+c+f}[-b+c+f][b+c+f+2])+(-1)^{\Psi}[2f][2f+1] \times[2f+2]([e+c-d][e-c+d]+(-1)^{e+c+d}[-e+c+d][e+c+d+2]), \tag{4.13}$$

$$C=(-1)^{b+c+f}[2][2f]([a+d+f+2][a-d+f+1][a+d-f][-a+d+f+1] \times[b+c+f+2][b-c+f+1][b+c-f][-b+c+f+1])^{1/2}, \tag{4.14}$$

and

$$\Psi=(a+b+e)(a+b+c+d)+(d+e+c)(b+d+e+f)+(a+d+f)(a+c+e+f)+c+d. \tag{4.15}$$

APPENDIX: ACTION OF THE UNIVERSAL *R* MATRIX ON THE REDUCED BASIS $e_m^{lq}(l_1,l_2,\lambda)$

In this appendix we prove the relation (2.47). For this, we recall that the vector $e_m^{lq}(l_1,l_2,\lambda)$ can be expressed in the form (Ref. 1)

$$e_m^{lq}(l_1,l_2,\lambda)=\frac{1}{(N(l,m,q))^{1/2}}\,P_{ml}^{lq\otimes}(1,2)e_{l_1}^{l_1q}(\lambda_1)\otimes e_{l-l_1}^{l_2q}(\lambda_2), \tag{A1}$$

where

$$P_{ml}^{lq\otimes}(1,2)=\left(\frac{[l+m]!}{[2l]![l-m]!}\,\gamma^{-(l-m)}\right)^{1/2} v_-^{\otimes q}(l_1,l_2)^{l-m}P^{lq\otimes}(1,2), \tag{A2}$$

$$P^{lq\otimes}(1,2)=(T^{l_1}\otimes T^{l_2})(\Delta(P^{lq}))=\sum_{r=0}^{\infty} c_r(l)(v_-^{\otimes}(1,2))^r(v_+^{\otimes}(1,2))^r, \tag{A3}$$

and P^{lq} is the projection operator on the highest weight vector of weight l. The coefficients $c_r(l)$ are of the form

$$c_r(l)=\frac{[2l+1]!}{[2l+r+1]![r]!\,\gamma^r}. \tag{A4}$$

The value of the normalization factor $N(l,m,q)$ is given by the formula

$$N(l,m,q)=q^{(1/2)(l_1+l_2-l)(l_1+l_2-l-1)}\,\frac{[2l+1]![2l_1]!}{[l_1+l_2+l]![l_1+l_2+l+1]!}. \tag{A5}$$

From relations (2.42) and (A2), (A3) it follows that

$$R^q(1,2)P^{lq\otimes}_{ml}(1,2)R^{q^{-1}}(1,2)=P^{lq^{-1}\otimes}_{ml}(1,2),$$

and one deduces successively the relations

$$R^q(1,2)e^{lq}_m(l_1,l_2,\lambda)=(N(l,m,q))^{-1/2}R^q(1,2)P^{lq\otimes}_{lm}(1,2)R^{q^{-1}}(1,2)R^q(1,2)e^{l_1q}_{l_1}(\lambda_1)\otimes e^{l_2q}_{l-l_1}(\lambda_2), \tag{A6}$$

$$=(N(l,m,q))^{-1/2}q^{l_1(l-l_1)}P^{lq^{-1}\otimes}_{ml}(1,2)e^{l_1q}_{l_1}(\lambda_1)\otimes e^{l_2q}_{l-l_1}(\lambda_2), \tag{A7}$$

$$=\left(\frac{N(l,m,q^{-1})}{N(l,m,q)}\right)^{1/2}q^{l_1(l-l_1)}e^{lq^{-1}}_m(l_1,l_2,\lambda). \tag{A8}$$

The norm ratio $N(l,m,q^{-1})/N(l,m,q)$ is calculated from the definition (A5), and one gets finally the action of the universal R-matrix on the reduced basis $e^{lq}_m(l_1,l_2,\lambda)$:

$$R^q(l_1,l_2)e^{lq}_m(l_1,l_2,\lambda)=q^{(1/2)[l(l+1)-l_1(l_1+1)-l_2(l_2+1)]}e^{lq^{-1}}_m(l_1,l_2,\lambda). \tag{A9}$$

ACKNOWLEDGMENT

One of the authors (M.M.) is grateful to the Laboratoire de Physique Théorique for its warm hospitality.

[1] P. Minnaert and M. Mozrzymas, J. Phys. A **28**, 669 (1995).
[2] P. Minnaert and M. Mozrzymas, J. Math. Phys. **36**, 907 (1995).
[3] P. Minnaert and M. Mozrzymas, J. Math. Phys. **38**, 2676 (1997).
[4] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison-Wesley, London, 1981).
[5] A. N. Kirillov and N. Yu. Reshetikhin, preprint E-9-88, Leningrad Branch (V. A. Steklov Mathematics Institute, Leningrad, 1988).
[6] M. Nomura, J. Math. Phys. **30**, 2397 (1989).
[7] Yu. F. Smirnov, V. N. Tolstoy, and Yu. I. Kharitonov, Sov. J. Nucl. Phys. **53**, 1068 (1991).
[8] P. P. Kulish and N. Yu. Reshetikhin, Lett. Math. Phys. **18**, 143 (1989).
[9] H. Saleur, Nucl. Phys. B **336**, 363 (1990).
[10] P. P. Kulish, LOMI preprint published in Zapiski Nautchnovo Seminaria, LOMI, 1990.
[11] P. Minnaert and M. Mozrzymas, J. Math. Phys. **35**, 3132 (1994).
[12] P. Minnaert and S. Toshev, J. Math. Phys. **35**, 5057 (1994).

Unobstructedness of Calabi–Yau orbi-Kleinfolds

Z. Ran[a)]
Department of Mathematics, University of California, Riverside, California 92521

(Received 29 August 1997; accepted for publication 4 September 1997)

We show that Calabi–Yau spaces with certain types of hypersurface-quotient singularities have unobstructed deformations. This applies in particular to all Calabi–Yau orbifolds nonsingular in codimension 2.
© *1998 American Institute of Physics.* [S0022-2488(98)02501-8]

I. INTRODUCTION

In recent years Calabi–Yau manifolds have attracted a great deal of attention, motivated both by their role in classification theory and by their connections with physics, in particular the phenomenon of mirror symmetry. Both the classification theory and physics viewpoints suggest looking more generally at a suitable class of *singular* Calabi–Yau spaces, e.g., most known mirror symmetry constructions involve forming a quotient under a finite group action, thus leading naturally to Calabi–Yau *orbifolds* and families of such.

Now a basic result about Calabi–Yau manifolds X is the theorem of Bogomolov–Tian–Todorov asserting that X has unobstructed deformations. Some extensions of this theorem to the case of singular X have been considered in Refs. 1–4. However,these extensions do not cover, e.g., the case of orbifolds. Thus one is naturally led to ask, as did Morrison, whether unobstructedness holds for Calabi–Yau orbifolds, say nonsingular in codimension 2. Our purpose here is to answer this question affirmatively and, in fact, to prove a rather more general statement which at the same time generalizes the reuslt of Ref. 2, allowing (local) quotients of Kleinfolds, whence the term ''orbi-Kleinfolds.''

The proof is particularly easy in the orbifold case, requiring little more than a simple combination of (either form of) the ''dual unobstructedness criterion'' of Ref. 5 with the ideas surrounding Schlessinger's rigidity theorem.[6] The more general orbi-Kleinfold case requires more work, which we feel is justified by the interest of the result.

II. SOME DEFINITIONS

We begin by recalling that a *Kleinian* singularity is an isolated simple hypersurface singularity. These are classified by Arnold into types $A–D–E$ (cf. Ref. 7, p. 132). We generalize this notion as follows.

Definition 1: An isolated hypersurface singularity (X,p) is said to be weakly Kleinian *provided.*

(i) *X is cDV (i.e., a general surface section through p is a DuVal or Kleinian singularity*);
(ii) *X admits a* $\mathbb{C}^*$ *action with positive weights*;
(iii) *for a resolution* $\widetilde{X}\to X$*, all components of the exceptional locus are smooth divisors E with* $H^{p,q}(E)=0$ *for* $p\neq q$.

Remark: Note that (i) implies that X is a rational singularity, hence $H^{p,0}(E)=0$, $p>0$. In particular, in dimension 3, (i) implies (iii).

Next recall from Ref. 2 that an isolated singularity (X,p) is said to be *good* if for a resolution

a)Electronic mail: ziv@ucrmath.ucr.edu

$\tilde{X}\to X$ and $\hat{\tilde{X}}$ the formal neighborhood of the exceptional locus, the map induced by exterior derivative

$$d:H^i(\Omega^j_{\tilde{X}})\to H^i(\Omega^{j+1}_{\tilde{X}})$$

is injective in the range $i,j>0$, $i\neq j$, $i+j<\dim X$. All threefold rational singularities, and A_1 singularities in all dimensions are good. Presumably all Kleinian singularities are good, but this is unproven. The unobstructedness result of Ref. 2 as stated allowed good Kleinian singularities, but it was remarked by Namikawa in Ref. 4 that the proof covers the good weakly Kleinian case as well.

Definition 2: An analytic variety X is said to be a (weak) orbi-Kleinfold *if X is locally of the form V/G, where V has good (weak) Kleinian singularities and G is a finite group acting on V which is ''small'' in that for all* $g\in G$, $p\in V^g$, *the induced action of g on the Zariski tangent* T_pV *has no eigenspace of codimension exactly 1.*

Remarks:

(1) All orbifolds are orbi-Kleinfolds (cf. Ref. 8).
(2) In dimension 3, all terminal singularities are, by Mori's classification (cf. Ref. 9) (cyclic) quotients of isolated cDV points, hence ''almost'' weakly orbi-Kleinian; thus the class of weak orbi-Kleinfolds is rather more general than that of orbifolds.
(3) All orbi-Kleinfolds are Cohen–Macaulay, because a finite quotient of a CM variety is CM [cf. Eagon–Hochster, ''Cohen–Macaulay rings, invariant theory, and the generic perfection of determinantal loci,'' Am J. Math (1971); indeed this follows easily by observing that CM is equivalent to an arbitrary system of parameters being regular, i.e., having trivial Koszul homology, and the fact that Koszul homology commutes with the averaging (or Reynolds) operator; I am indebted to M. Hochster for this remark].
(4) If $X=V/G$ is a local weak orbi-Kleinfold, then V admits a G-equivariant resolution $\tilde{V}$, yielding an orbifold $X_{\text{orb}}=V/G$ with a birational morphism to X. This construction clearly globalizes, yielding an ''orbifold resolution'' $X_{\text{orb}}\to X$ for any weak orbi-Kleinfold, which is an isomorphism off the (discrete) nonorbifold locus of X.

Definition 3: A Calabi–Yau (weak) orbi-Kleinfold *is a compact irreducible (weak) orbi-Kleinfold X such that*

(i) *X admits a resolution of singularities by a Kähler manifold*;
(ii) *the dualizing sheaf* $\omega_X\cong\mathscr{O}_X$:

III. THEOREM AND PROOF

Our main result is the following:

Theorem 4: Any Calabi–Yau weak orbi-Kleinfold nonsingular in codimension 2 has unobstructed deformations.

Remark: This generalizes Theorem 1 of Ref. 2. In dimension 3, unobstructedness of Calabi–Yau spaces with terminal singularities has been proven by Namikawa,[4] but not all weak orbi-Kleinfolds are terminal, nor conversely.

Proof of theorem: Let X be a Calabi–Yau weak orbi-Kleinfold of dimension n, $\Theta=\mathscr{H}om(\Omega^1_X,\mathscr{O}_X)$ its tangent sheaf, and $j:X_{\text{reg}}\to X$ the inclusion of the open subset of regular points. In view of our depth hypothesis, results of Schlessinger,[6] as exposed in Ref. 10, Secs. I 9–10, show that we have an isomorphism of deformation functors

$$\text{Def}(X)\xrightarrow{\sim}\text{Def}(X_{\text{reg}}).$$

In particular the first-order deformation group

$$T^1(X)=\mathrm{Ext}^1(\Omega^1_X,\mathscr{O}_X)\cong H^1(X_{\mathrm{reg}},\Theta),$$

and, more importantly for us obstructions to deforming X may be taken in the group $H^2(X_{\mathrm{reg}},\Theta)$. The following is in essence due to Schlessinger (and is valid more generally for quotients, nonsingular in codimension 2, of isolated hypersurface singularities).

Lemma 5: (i) $R^1j_*\Theta_{X_{\mathrm{reg}}}$ is supported in the nonorbifold locus of X; (ii) $R^2j_*\Theta_{X_{\mathrm{reg}}}=0$.

Proof: We work locally on a small open set $U=V/G\subset X$, where V is a local hypersurface with one (possibly) singular point p. Put

$$V_0=V-\{p\},$$

$$V_{00}=\text{subset of } V_0 \text{ where } G \text{ acts freely.}$$

Note that $V-V_{00}$ has codimension $\geqslant 3$ in X and $U_{\mathrm{reg}}=V_{00}/G$. As in Ref. 10, we have

$$H^i(V_{\mathrm{reg}},\Theta)=H^i(V_{00},\Theta_{V_{00}})^G=H^i(V_0,\Theta_{V0})^G,\quad i=1,2,$$

and $H^i(V_0,\Theta_{V_0})$ coincides with $T^i_{V,p}$ and in particular vanishes if p is regular or $i=2$. □

Now put

$$\widetilde{\Omega}_X=j_*\Omega_{X_{\mathrm{reg}}}.$$

This is a complex of reflexive sheaves on X and by Lemma 2 we have

$$H^2(X_{\mathrm{reg}},\Theta_{X_{\mathrm{reg}}})=H^2(X,j_*\Theta_{X_{\mathrm{reg}}})\hookrightarrow\mathrm{Ext}^2(\widetilde{\Omega}^1_X,\mathscr{O}_X),$$

hence the latter group is an obstruction group for deformations of X. By Grothendieck duality, $\mathrm{Ext}^2(\widetilde{\Omega}^1_X,\mathscr{O}_X)=\mathrm{Ext}^2(\widetilde{\Omega}^1_X,\omega_X)$ is dual to $H^{n-2}(\widetilde{\Omega}^1_X)$. Now the following result generalizes simultaneously results of Steenbrink[8] and Ref. 2, Proposition 4.

Proposition 6: Let X be a compact weak orbi-Kleinfold of dimension $n\geqslant 2$, $\pi_0:X_{orb}\to X$ an orbifold resolution $\pi_1:\widetilde{X}\to X_{orb}$ a resolution of singularities with $\widetilde{X}$ Kähler. Then

(i) $\widetilde{\Omega}^{\cdot}_X=\pi_{0*}\widetilde{\Omega}^{\cdot}_{X_{\mathrm{orb}}}=(\pi_0\circ\pi_1)_*\Omega^{\cdot}_{\widetilde{X}}$;

(ii) *$\widetilde{\Omega}^{\cdot}_X$ is a resolution of the constant sheaf $\mathbb{C}_X$;*

(iii) *the Hodge–De Rham spectral sequence*

$$E_1^{p,q}=H^q(X,\widetilde{\Omega}^p)\Rightarrow H^{p,q}(X,\mathbb{C})$$

degenerates at E_1 in degrees $\leqslant n-1$.

The argument that Proposition 6 implies Theorem 4, based on the ''T^2-injecting criterion'' (Theorem 1.1, (ii) of Ref. 5) is identical to the corresponding argument in Ref. 2. In the case of orbifolds, Proposition 6 is due to Steenbrink[8] (and consequently the reader only interested in orbifolds may stop reading here). Our proof of the Proposition combines ideas from Refs. 8 and 2.

Proof of Proposition 6: (i) The assertion is local, and at orbifold points has been proven by Steenbrink (Ref. 8), Lemma 1.11). For the nonorbifold points, consider

$$U=V/G\subset X,$$

$$\omega\in\Gamma(U,\widetilde{\Omega}^i_X)=\Gamma(U_{\mathrm{reg}},\Omega^i_{\mathrm{reg}}).$$

By Ref. 8, Lemma 1.8, ω lifts to $\tilde{\omega}\in\Gamma(V_{\text{reg}},\Omega^i_{V_{\text{reg}}})$. We may assume our desingularization of X is locally over U of the form $\tilde{V}/G$ where $\tilde{V}\to V$ is some desingularization (which will also blow up the nonfree locus of G, in addition to sing V). By Ref. 2, Proposition 4, (i), $\tilde{\omega}$ extends holomorphically to $\tilde{V}$ and being G-invariant descends to $\tilde{V}/G$, as required. (ii) This is a consequence of Ref. 8, Lemma 1.8, Ref. 2, Proposition 4, and exactness of the functor of G-invariants. (iii) We prove the vanishing of the differentials d_1^{ij}, $i+j<n$, the case of d_r^{ij}, $r\geqslant 2$, being similar. Denote by $\tilde{\Omega}^j_X\subset\tilde{\Omega}^j_X$ the subsheaf of closed forms and similarly for $\tilde{\Omega}^j_{X_{\text{orb}}},\tilde{\Omega}^j_{\tilde{X}}$. Consider the Leray spectral sequence

$$E_2^{p,q}=H^p(X,R^q\pi_{0*}\tilde{\Omega}^j_{X_{\text{orb}}})\Rightarrow H^i(X_{\text{orb}},\tilde{\Omega}^j_{X_{\text{orb}}})$$

and the anlogous one for $\tilde{\Omega}^j_{X_{\text{orb}}}$. Using that π_0 is an isomorphism off the finite nonorbifold locus of X, we get a diagram

$$H^0(R^{i-1}\pi_{0*}\tilde{\Omega}^j_{X_{\text{orb}}})\to H^0(R^{i-1}\pi_{0*}\tilde{\Omega}^{j+1}_{X_{\text{orb}}})\overset{\beta_{ij}}{\to}H^0(R^{i-1}\pi_{0*}\tilde{\Omega}^{j+1})$$

$$(*)H^i(\overset{\downarrow}{\tilde{\Omega}^j_X})\to H^i(\overset{\downarrow}{\hat{\Omega}^{j+1}_X})\to H^i(\overset{\downarrow}{\tilde{\Omega}^{j+1}_X})$$

$$H^i(\overset{\downarrow}{\tilde{\Omega}^j_{X_{\text{orb}}}})\overset{\alpha_{ij}}{\to}H^i(\overset{\downarrow}{\hat{\Omega}^{j+1}_{X_{\text{orb}}}})\to H^i(\overset{\downarrow}{\tilde{\Omega}^{j+1}_{X_{\text{orb}}}})L$$

in which the composite of the middle-row arrow coincides with d_1^{ij}. by Steenbrink,[8] we have

$$H^i(\tilde{\Omega}^{j+1}_{X_{\text{orb}}})=\mathbb{H}^i(F^{j+1}\tilde{\Omega}^{\cdot}_{X_{\text{orb}}})=F^{j+1}\mathbb{H}^{i+j+1}(\tilde{\Omega}^{\cdot}_{X_{\text{orb}}}),$$

where F denotes the Hodge (or stupid) filtration, and it follows easily that α_{ij} vanishes. To prove $d_1^{ij}=0$ for $i+j<n$, $j\neq i-2$, it will suffice, by $(*)$, to prove the vanishing of β_{ij} in this range. For this we use our goodness hypothesis. Let $\hat{X}_{\text{orb}}$ be the formal neighborhood of the exceptional locus of π_0 and similarly for $\hat{\tilde{X}}$. Then we have a diagram

$$H^{i-1}(\hat{X}_{\text{orb}},\hat{\Omega}^{j+1}_{X_{\text{orb}}})\overset{\hat{\beta}_{ij}}{\to}H^{i-1}(\hat{X}_{\text{orb}},\Omega^{j+1}_{\text{orB}})$$

$$(**)H^{i-1}(\hat{\tilde{X}},\hat{\Omega}^{j+1}_{\tilde{X}})\overset{\tilde{\beta}_{ij}}{\to}H^{i-1}(\hat{X},\Omega^{j+1}_{X}),$$

where $\hat{\beta}_{ij},\tilde{\beta}_{ij}$ are the obvious maps. By goodness, $\tilde{\beta}_{ij}$ vanishes whenever $i-1\neq j+1$, because the supposedly injective exterior-derivative map vanishes on its image. If we can prove the vertical maps π^* in $(**)$ are injective, it then follows that $\hat{\beta}_{ij}$, hence β_{ij}, vanishes. But the injectivity of π^* follows from Steenbrink's duality argument (Ref. 8, Proof of Theorem 1.12), applied on the formal scheme $\hat{X}_{\text{orb}}$.

Finally, it remains to prove the vanishing of $d_1^{i,i-2}$, $2\leqslant i\leqslant(n+1)/2$. This is done just as in Ref. 2, comparing $H^{\cdot}(X,\mathbb{C})$ with $H^{\cdot}(X_{\text{orb}},\mathbb{C})$ and using the fact that the component of the exceptional locus of π_0 are quotients of manifolds with only (p,p) cohomology, hence by Steenbrink's theory are themselves orbifolds with all their cohomology of type (p,p), hence contribute via Gysin only (p,p) classes to $H^{\cdot}(X_{\text{orb}})$. □

Remark: After this was written, the author became aware of Fujiki's paper,[11] which contains numerous results on symplectic orbifolds and their deformations, but not including unobstructedness, which Fujiki includes as an hypothesis in some statements. In particular, in his Theorem 4.8, p. 116, the hypothesis ''$\mathscr{S}$ is smooth'' holds automatically.

ACKNOWLEDGMENTS

We are grateful to Professor D. R. Morrison for suggesting the problem of unobstructedness of Calabi–Yau orbifolds.

This work was supported in part by NSF DMS 9202050.

[1] Y. Kawamata, ''Unobstructed deformations—a remark on a paper by Z. Ran'' (preprint).
[2] Z. Ran, ''Deformations of Calabi–Yau Kleinfolds'' in ''Essays on Mirror Symmetry,'' edited by S. T. Yau (International, Hong Kong, 1992), pp. 451–457.
[3] G. Tian, ''Smoothing 3-folds with trivial canonical bundle and ordinary double points'' (preprint), pp. 458–479.
[4] Y. Namikawa, ''On deformations of Calabi–Yau 3-folds with terminal singularities'' (preprint).
[5] Z. Ran, ''Hodge theory and deformations of maps,'' Compositio Mathematica **97**, 309–328 (1995).
[6] M. Schlessinger, ''Rigidity of quotient singularities,'' Invent. Math. **14**, (1971).
[7] E. J. N. Looijenga, *Isolated singular points on complete intersections* (Cambridge University Press, Cambridge, 1984).
[8] J. H. M. Steenbrink, ''Mixed Hodge structure on the vanishing cohomology,'' in *Real and complex singularities* (Oslo, 1976), pp. 513–536.
[9] M. Reid, ''A young person's guide to canonical singularities,'' in Proc. Symp. Pure Math, AMS.
[10] M. Artin, *Deformations of Singularities* (Tata Institute, Bombay, 1976).
[11] A. Fujiki, ''On primitively symplectic compact Kähler *V*-manifolds of dimension four,'' in *Classification of algebraic and analytic manifolds*, edited by K. Ueno (Birkhauser, Boston, 1983), pp. 71–250.

Algebraic expressions for $O(N)\supset O(N-1)$ reduction factors for the three-rowed irreducible representations

Hong-zhou Sun[a)]
Department of Physics, Tsinghua University, Beijing 100084, People's Republic of China and Institute of Theoretical Physics, Academia Sinica, Beijing 100080, People's Republic of China

Dong Ruan
Department of Physics, Tsinghua University, Beijing 100084, People's Republic of China

(Received 1 October 1996; accepted for publication 4 September 1997)

Using the concept of irreducible tensor basis and the properties of irreducible tensor operator, we give all $O(N)\supset O(N-1)$ reduction factors for $(m_1m_2m_30\cdot 0)\otimes(100\cdot 0)$. © *1998 American Institute of Physics.* [S0022-2488(97)00712-3]

I. INTRODUCTION

Reduction factors are very useful in many physical problems. For example, in algebraic models of atomic[1,2] and nuclear[3–5] physics, in order to classify wave-functions and to gain the analytic formulae for the electric and magnetic transition rates, one needs to know the algebraic expressions for reduction factors (reduced Wigner coefficients).

$O(N)\supset O(N-1)$ reduction factors have been studied by many authors. For $N=3$, they are the well-known Clebsch–Gordon coefficients of $O(3)$.[6,7] $O(4)\supset O(3)$ reduction factors in the $O(4)\supset O(3)\supset O(2)$ basis were first obtained by Biedenharn.[8] For some special reduction factors of $O(N)\supset O(N-1)$, Gavrilik[9] considered the case of coupling $l_1\otimes l_2$ to (L_1L_2) with all three symmetric irreducible representations (i.e., with $L_2=0$). Later Ališauskas *et al.*[10–12] derived the algebraic expressions for the coupling $l_1\otimes l_2$ to (L_1L_2) by use of the substitution group technique of $SP_4(O_5)$ and Pan *et al.*[13] discussed the case of the coupling $(w_1w_2)\otimes(w_30)$ to $(w_1'w_1')$ with $w_1+w_2+w_3=w_1'+w_2'$ in terms of $\mathrm{SU}(N)\supset\mathrm{SU}(N-1)\supset O(N-1)$. However, these results are very complicated and since they only considered the less than two-rowed irreducible representations for $O(N)$ and $O(N-1)$, from these results, one can not obtain all $O(6)\supset O(5)$ reduction factors for $(m_{16}m_{26}m_{36})\otimes(1\,0\,0)$, with $m_{36}\neq 0$, which are very important in the $O(6)$ limit case in the interacting boson model (IBM).[3,4]

In the 1960's an effective method—irreducible tensor basis method, was put forward by Baird *et al.*[14] and Sun.[15] This method relies on the concept of an irreducible tensor basis, that is, the generators of group G are composed of the generators of its subgroup G' and the irreducible tensor operators of G'. So it can lead to the recurrent relations for the irreducible representations and for the reduction factors easily and naturally, which can reduce the amount of calculation largely. The method has been applied widely.[14–23] In this paper, we will use the irreducible tensor basis method to calculate the reduction factors of $O(N)\supset O(N-1)$ for the three-rowed irreducible representations of $O(N)$ and $O(N-1)$.

This paper is arranged as follows. In Section II, the basic results of $O(3)$ are reviewed. In Section III, the irreducible tensor basis of $O(4)$ is constructed. In Section IV, the irreducible representations of $O(4)$ are derived. In Section V, the reduction factors of $O(4)\supset O(3)$ are evaluated. In Section VI, further calculations are discussed and the reduction factors of $O(N)\supset O(N-1)$ for the three-rowed irreducible representations of $O(N)$ and $O(N-1)$ are obtained. A brief conclusion is given in the final section.

[a)]Electronic mail: long@ibm320h.phy.pku.edu.cn

0022-2488/98/39(1)/630/20/$15.00
© 1998 American Institute of Physics

II. BASIC RESULTS OF $O(3)$

In fact, the irreducible tensor basis method is very parallel to the treatment of the quantum theory of angular momentum, where the groups of interest are $O(3)$ and $SU(2)$.[6,7] In the framework of the irreducible tensor basis method, some important concepts and tools in the theory of angular momentum such as irreducible tensor operator and the Wigner–Eckart theorem, which are very useful for evaluating the matrix elements of physical quantities and the Clebsch–Gordon coefficients of $O(3)$, will be extended and then be used to discuss the groups $O(N)$ $(N>3)$. To this end, we first have to construct the new basis for $O(N)$ $(N\geqslant 3)$—the irreducible tensor basis, from their old generators I_{pq}, then can apply the (generalized) Wigner–Eckart theorem to evaluating the irreducible representations of $O(N)$ and $O(N)\supset O(N-1)$ reduction factors. In order to see clearly how to construct the irreducible tensor basis for the orthogonal groups, and how to utilize the irreducible tensor basis method, let us begin with the simplest case—$O(3)$.

It is well known that the generators $I_{pq}(p>q)$ of $O(N)$ satisfy the commutation relations[24,16]

$$[I_{pq},I_{rs}]=\delta_{ps}I_{qr}+\delta_{qr}I_{ps}-\delta_{pr}I_{qs}-\delta_{qs}I_{pr},$$

$$p,q,r,s=1,2,...,N, \tag{1}$$

where $I_{pq}=E_{pq}-E_{qp}$ and E_{pq} is a $N\times N$ matrix with only one nonzero element 1 in the pth row and qth column position,

$$E_{pq}=\begin{bmatrix} 0 & \cdot & 0 & 0 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & 0 & 0 & \cdot & 0 \\ 0 & \cdot & 0 & 1 & 0 & \cdot & 0 \\ 0 & \cdot & 0 & 0 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & 0 & 0 & \cdot & 0 \end{bmatrix} \leftarrow p\text{th row}.$$

$$\uparrow$$

$$q\text{th column}$$

When $N=3$, Eq. (1) gives the commutation relations satisfied by the generators $\{I_{32},I_{31},I_{21}\}$ of $O(3)$.

First, it is convenient to replace the three linearly independent generators $\{I_{32},I_{31},I_{21}\}$ of $O(3)$ by their linear combinations. This can be done by defining[16]

$$L_0=iI_{21},\quad L_{\pm 1}=-\frac{1}{\sqrt{2}}(I_{31}\pm iI_{32}). \tag{2}$$

They satisfy

$$[L_0,L_{\pm 1}]=\pm L_{\pm 1},\quad [L_{+1},L_{-1}]=L_0. \tag{3}$$

The above equations are the usual definition of angular momentum operator **L** in the quantum mechanics,[6,7] and L_0 and $L_{\pm 1}$ are the three spherically harmonic components of **L**.

TABLE I. The Clebsch–Gordon coefficients of $O(3)\left[\begin{smallmatrix} l & 1 \\ m' & q \end{smallmatrix}\middle| \begin{smallmatrix} l' \\ m \end{smallmatrix}\right]$.

m'	$q \backslash l'$	$l+1$	l	$l-1$
$m-1$	1	$\sqrt{\frac{(l+m)(l+m+1)}{(2l+1)(2l+2)}}$	$-\sqrt{\frac{(l+m)(l-m+1)}{2l(2l+1)}}$	$-\sqrt{\frac{(l-m)(l-m+1)}{2l(l+1)}}$
m	0	$\sqrt{\frac{(l-m+1)(l+m+1)}{(l+1)(2l+1)}}$	$\frac{m}{\sqrt{l(l+1)}}$	$\sqrt{\frac{(l-m)(l+m)}{l(2l+1)}}$
$m+1$	-1	$\sqrt{\frac{(l-m)(l-m+1)}{(2l+1)(2l+2)}}$	$\sqrt{\frac{(l-m)(l+m+1)}{2l(l+1)}}$	$-\sqrt{\frac{(l+m)(l+m+1)}{2l(2l+1)}}$

Take the Gel'fand–Tsetlin basis $\left|\begin{smallmatrix} l \\ m \end{smallmatrix}\right]$ as the basis vectors carrying the irreducible representations of $O(3)$, where l labels the irreducible representations of $O(3)$ while m labels those of $O(2)$. (In the quantum theory of angular momentum, $|lm\rangle \equiv \left|\begin{smallmatrix} l \\ m \end{smallmatrix}\right]$ are the common eigenvectors of $\{\mathbf{L}^2, L_0\}$.) As the components of l for distinguishing the individual basis vectors, m must obey the constraint $-l \leqslant m \leqslant l$. Thus, by means of the standard treatment of quantum theory of angular momentum,[6,7] we have immediately the irreducible representations of $O(3)$,

$$\begin{aligned} &\left[\begin{matrix} l \\ m' \end{matrix}\middle| L_0 \middle| \begin{matrix} l \\ m \end{matrix}\right] = m\,\delta(m'm), \\ &\left[\begin{matrix} l \\ m' \end{matrix}\middle| L_{\pm 1} \middle| \begin{matrix} l \\ m \end{matrix}\right] = \mp \sqrt{\frac{1}{2}(l \mp m)(l \pm m + 1)}\,\delta(m' m \pm 1), \end{aligned} \tag{4}$$

and the Clebsch–Gordon coefficients of $O(3)$ (the case of $l \times 1$ considered in this paper are given in Table I).

Here, we have chosen the phase conventions for the Clebsch–Gordon coefficients (Table I) of $O(3)$,

$$\begin{aligned} &\left[\begin{matrix} l & 1 \\ l & 0 \end{matrix}\middle| \begin{matrix} l' \\ l \end{matrix}\right] > 0, \\ &\left[\begin{matrix} l & 1 \\ m & q \end{matrix}\middle| \begin{matrix} l' \\ m' \end{matrix}\right] = \left[\begin{matrix} 1 & l \\ q & m \end{matrix}\middle| \begin{matrix} l' \\ m' \end{matrix}\right]. \end{aligned} \tag{5}$$

Note that in later calculations of the reduction factors of $O(N)\supset O(N-1)(N>3)$, we will choose the phase conventions consistent with Eq. (5).

Finally, in order to construct the irreducible tensor basis of $O(4)$ in the next section, let us review the usual definition of an irreducible tensor operator of $O(3)$, employed by Racah[25] initially. The set $\mathbf{T}^{(3)}(l)$ is called the irreducible tensor operator of rank l of $O(3)$, if its linearly independent components $T^{(3)}\left(\begin{smallmatrix} l \\ m \end{smallmatrix}\right)$ satisfy

$$\begin{aligned} &\left[L_i,\ T^{(3)}\begin{pmatrix} l \\ m \end{pmatrix}\right] = \sum_{m'} T^{(3)}\begin{pmatrix} l \\ m' \end{pmatrix}\left[\begin{matrix} l \\ m' \end{matrix}\middle| L_i \middle| \begin{matrix} l \\ m \end{matrix}\right], \\ &i = 0, \pm 1; \quad m, m' = -l, -l+1, \ldots, l-1, l, \end{aligned} \tag{6}$$

where $[{}^{l}_{m'}|L_i|^{l}_{m}]$ are the irreducible representation matrices of $O(3)$. Equation (6) shows that under the operations of $O(3)$ the irreducible tensor operator $\mathbf{T}^{(3)}(l)$ transforms according to the irreducible representations l of $O(3)$. The number of components $T^{(3)}({}^{l}_{m})$ is equal to the dimension $2l+1$ of the irreducible representations l of $O(3)$.

The definition of the irreducible tensor operator of $O(3)$ has been generalized to $O(N)(N>3)$ (see Appendix A).

III. THE IRREDUCIBLE TENSOR BASIS OF $O(4)$

In order to apply the irreducible tensor basis method to $O(4)$, we first have to construct the irreducible tensor basis of $O(4)$ by combining its old generators $\{I_{43},I_{42},I_{41},I_{32},I_{31},I_{21}\}$. Observe that some of them, $\{I_{32},I_{31},I_{21}\}$, in fact are the generators of the subgroup $O(3)$ of $O(4)$, which have been combined to the angular momentum operators $\mathbf{L}$ [see Eq. (2)] of $O(3)$ in the last section. Then the remaining three generators, $\{I_{43},I_{42},I_{41}\}$, can be conveniently grouped to display their transformation properties under the subgroup $O(3)$. Defining

$$\begin{aligned} T^{(3)}\begin{pmatrix}1\\0\end{pmatrix}&\equiv iI_{43},\\ T^{(3)}\begin{pmatrix}1\\+1\end{pmatrix}&\equiv\sqrt{\frac{1}{2}}\,(I_{42}-iI_{41}),\\ T^{(3)}\begin{pmatrix}1\\-1\end{pmatrix}&\equiv\sqrt{\frac{1}{2}}\,(I_{42}+iI_{41}),\end{aligned} \tag{7}$$

it is checked easily that they satisfy Eq. (6). Hence, $T^{(3)}({}^{1}_{q})$ $(q=-1,0,1)$, as three components form an irreducible tensor operator $\mathbf{T}^{(3)}(1)$ of rank 1 of $O(3)$.

Hence, the generators of $O(4)$ are composed of the generators $\mathbf{L}$ of the subgroup $O(3)$ and the irreducible tensor operator $\mathbf{T}^{(3)}(1)$ of rank 1 of $O(3)$. Note that since we seek unitary representations, in Eq. (7), we have so defined the irreducible tensor operator $\mathbf{T}^{(3)}(1)$ that

$$T^{(3)\dagger}\begin{pmatrix}1\\q\end{pmatrix}=(-)^{q}T^{(3)}\begin{pmatrix}1\\-q\end{pmatrix}. \tag{8}$$

Then the commutation relations satisfied by the generators of $O(4)$ are the following.

(1) The commutation relations of $O(3)$ [Eq. (3)].
(2) The commutation relations between the generators $\mathbf{L}$ of $O(3)$ and the irreducible tensor operator $\mathbf{T}^{(3)}(1)$ of $O(3)$ [Eq. (6)].
(3) The commutation relations between the components of $\mathbf{T}^{(3)}(1)$. Using Eq. (1), we get immediately

$$\begin{aligned}\left[T^{(3)}\begin{pmatrix}1\\0\end{pmatrix},\quad T^{(3)}\begin{pmatrix}1\\\pm1\end{pmatrix}\right]&=\pm L_{\pm1},\\ \left[T^{(3)}\begin{pmatrix}1\\+1\end{pmatrix},\quad T^{(3)}\begin{pmatrix}1\\-1\end{pmatrix}\right]&=-L_0.\end{aligned} \tag{9}$$

We have also found from the commutation relations of $O(4)$ that, in the irreducible tensor basis of $O(4)$, Cartan's subalgebra of $O(4)$ is

$$H_1=L_0, \quad H_2=T^{(3)}\begin{pmatrix}1\\0\end{pmatrix}. \tag{10}$$

IV. THE IRREDUCIBLE REPRESENTATIONS OF $O(4)$

Take the Gel'fand–Tsetlin basis,[24]

$$\begin{vmatrix} v_1 & v_2 \\ l & \\ m & \end{vmatrix}, \tag{11}$$

as the basis vectors carrying the irreducible representations (v_1v_2) of $O(4)$. All the entries in the Gel'fand pattern (11) must obey the following betweenness relations:

$$v_1\geqslant l\geqslant|v_2|\geqslant 0, \quad l\geqslant|m|\geqslant 0. \tag{12}$$

Hence, evaluating the irreducible representations of $O(4)$ is to evaluate the matrix elements of all generators of $O(4)$ in the Gel'fand–Tsetlin basis (11). Fortunately, owing to the fact that the irreducible representations of the subgroup $O(3)$ in the Gel'fand–Tsetlin basis have been given in the last section, now we need only to evaluate the matrix elements of the remaining generators, the irreducible tensor operator $\mathbf{T}^{(3)}(1)$, of $O(4)$. Furthermore, by means of the Wigner–Eckart theorem,

$$\begin{bmatrix} v_1 & v_2 \\ l' & \\ m' & \end{bmatrix} T^{(3)}\begin{pmatrix}1\\q\end{pmatrix} \begin{bmatrix} v_1 & v_2 \\ l & \\ m & \end{bmatrix} = \begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix} T^{(3)}(1) \begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix} \begin{bmatrix} l & 1 & l' \\ m & q & m' \end{bmatrix}, \tag{13}$$

we can simplify the problem of calculating the matrix representations of $\mathbf{T}^{(3)}(1)$ into that of evaluating the $T^{(3)}(1)$ reduced matrix elements,

$$\begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix} T^{(3)}(1) \begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}, \quad l'=l,\ l\pm 1. \tag{14}$$

We begin with calculating the matrix elements of Eq. (8): it follows that

$$\begin{bmatrix} v_1 & v_2 \\ l' & \\ m' & \end{bmatrix} T^{(3)\dagger}\begin{pmatrix}1\\q\end{pmatrix} \begin{bmatrix} v_1 & v_2 \\ l & \\ m & \end{bmatrix} = (-)^q \begin{bmatrix} v_1 & v_2 \\ l' & \\ m' & \end{bmatrix} T^{(3)}\begin{pmatrix}1\\-q\end{pmatrix} \begin{bmatrix} v_1 & v_2 \\ l & \\ m & \end{bmatrix}. \tag{15}$$

We square Eq. (15), apply the Wigner–Eckart theorem, and then run over l', l, m', m, and q using the orthogonality of the Clebsch–Gordon coefficients, to obtain the symmetry relation satisfied by the $T^{(3)}(1)$ reduced matrix elements,

$$\begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix} T^{(3)}(1) \begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix} = \sqrt{\frac{2l+1}{2l'+1}} \begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix} T^{(3)}(1) \begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix}. \tag{16}$$

Here, we have chosen the phase conventions for the $T^{(3)}(1)$ reduced matrix elements,

$$\begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\geqslant 0, \quad l'\neq l. \tag{17}$$

In later calculations of the reduced matrix elements of the irreducible tensor operators $\mathbf{T}^{(N)}\times(10\cdot 0)$ of rank $(10\cdot 0)$ of $O(N)$ $(N>3)$, we will choose the phase conventions consistent with Eq. (17).

Next calculate the matrix elements of a commutation relation,

$$\left[T^{(3)}\begin{pmatrix} 1 \\ +1 \end{pmatrix}, \quad T^{(3)}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right]=-L_0. \tag{18}$$

By virtue of the Wigner–Eckart theorem, together with the Clebsch–Gordon coefficients (Table I) of $O(3)$ and the symmetry relation Eq. (16), it is straightforward to obtain the following two equations satisfied by the $T^{(3)}(1)$ reduced matrix elements:

$$\begin{bmatrix} v_1 & v_2 \\ l-1 & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l-1 & \end{bmatrix}=\sqrt{\frac{l+1}{l-1}}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}, \tag{19}$$

and

$$\begin{aligned}\frac{2l-1}{2l+1}\begin{bmatrix} v_1 & v_2 \\ l-1 & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}^2 &= l-\frac{1}{l+1}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}^2 \\ &\quad +\frac{1}{l+1}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l+1 & \end{bmatrix}^2.\end{aligned} \tag{20}$$

We observe that Eq. (19) is a recurrent equation and Eq. (20) is an inhomogeneous recurrent equation.

Now evaluate the matrix elements of $T^{(3)}(\begin{smallmatrix}1\\0\end{smallmatrix})$ in the highest weight basis vector. It follows, using Eq. (10) and the highest weight theorem,[2,16] that

$$\begin{bmatrix} v_1 & v_2 \\ \hat{l} & \\ \hat{m} & \end{bmatrix}\left| T^{(3)}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|\begin{bmatrix} v_1 & v_2 \\ \hat{l} & \\ \hat{m} & \end{bmatrix}=\begin{bmatrix} v_1 & v_2 \\ v_1 & \\ v_1 & \end{bmatrix}\left| T^{(3)}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|\begin{bmatrix} v_1 & v_2 \\ v_1 & \\ v_1 & \end{bmatrix}=v_2, \tag{21}$$

where $\hat{l}$ stands for the highest weight representation (i.e., the allowed maximum of l) of $O(3)$ contained in (v_1v_2) of $O(4)$, so $\hat{l}=v_1$. $\hat{m}$ has the similar definition. Application of the Wigner–Eckart theorem and the Clebsch–Gordon coefficients (Table I) of $O(3)$ to the above equation gives rise to

$$\begin{bmatrix} v_1 & v_2 \\ \hat{l} & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ \hat{l} & \end{bmatrix}=v_2\sqrt{\frac{v_1+1}{v_1}}. \tag{22}$$

Then inserting it into the recurrent equation (19) and utilizing Eq. (19) repeatedly we can obtain, in light of mathematical induction,

$$\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\left\| T^{(3)}(1) \right\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}=\frac{(v_1+1)v_2}{\sqrt{l(l+1)}}. \tag{23}$$

Inserting the above equation into Eq. (20), and from the highest weight state, utilizing Eq. (20) repeatedly, we can obtain, by mathematical induction,

$$\begin{bmatrix} v_1 & v_2 \\ l-1 & \end{bmatrix} \left\| T^{(3)}(1) \right\| \begin{matrix} v_1 & v_2 \\ l & \end{matrix} = \sqrt{\frac{(v_1+l+1)(v_1-l+1)(v_2+l)(l-v_2)}{l(2l-1)}}. \tag{24}$$

By virtue of the symmetry relation (16), and replacing l by $l+1$, we then get

$$\begin{bmatrix} v_1 & v_2 \\ l+1 & \end{bmatrix} \left\| T^{(3)}(1) \right\| \begin{matrix} v_1 & v_2 \\ l & \end{matrix} = \sqrt{\frac{(v_1+l+2)(v_1-l)(v_2+l+1)(l-v_2+1)}{(l+1)(2l+3)}}. \tag{25}$$

Equation (23), Eq. (24), and Eq. (25) give all the reduced matrix elements of $\mathbf{T}^{(3)}(1)$. Thus we obtain the irreducible representations of $O(4)$.

V. $O(4)\supset O(3)$ REDUCTION FACTORS

In terms of the concept of coupling coefficients, the coupled Gel'fand–Tsetlin basis of $O(4)$ constructed from two Gel'fand–Tsetlin bases distinguished by a single prime and two primes may be written as

$$\begin{aligned} \begin{vmatrix} v_1 & v_2 \\ l & \\ m & \end{vmatrix} &= \Sigma \begin{vmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \\ m' & & m'' & \end{vmatrix} \begin{bmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \\ m' & & m'' & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \\ m & \end{vmatrix} \\ &= \Sigma \begin{vmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \\ m' & & m'' & \end{vmatrix} \begin{bmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix} \begin{bmatrix} l' & l'' \\ m' & m'' \end{vmatrix} \begin{matrix} l \\ m \end{matrix}. \end{aligned} \tag{26}$$

where

$$\begin{bmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \\ m' & & m'' & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \\ m & \end{vmatrix}$$

are the coupling coefficients of $O(4)$. In order to gain the second equation, we have used Racah's factorization lemma,[25,2] which decompose the coupling coefficients of $O(4)$ into the product of the Clebsch–Gordon coefficients of the subgroup $O(3)$ and factor

$$\begin{bmatrix} v_1' & v_2' & v_1'' & v_2'' \\ l' & & l'' & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix}.$$

This factor is referred to as the $O(4)\supset O(3)$ reduction factor, whose value depends only on the irreducible representations of $O(4)$ and $O(3)$.

Consider a coupled irreducible tensor of rank 1 of $O(3)$,

$$T^{(3)}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = T^{(3)\prime}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + T^{(3)\prime\prime}\begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{27}$$

It is worthy of reminding the reader that all the three irreducible tensor operators in Eq. (27) are the generators of $O(4)$. Now calculate the matrix element of Eq. (27) in the coupled Gel'fand–Tsetlin basis (26) with $(v_1'' \; v_2'')=(1\;0)$, $l''\equiv\beta$ and $m''=0$, where $T^{(3)\prime}(\begin{smallmatrix}1\\0\end{smallmatrix})$ and $T^{(3)\prime\prime}(\begin{smallmatrix}1\\0\end{smallmatrix})$ act on the Gel'fand–Tsetlin basis of a single prime and two primes, respectively. It follows, in light of the Wigner–Eckart theorem, that we can obtain the following equation satisfied by $O(4)\supset O(3)$ reduction factors,

$$\sum_{\tilde{l}} C_1(\tilde{l})\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ \tilde{l}' & & \beta & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ \tilde{l} & \end{bmatrix}=\sum_{\tilde{l}} C_2(\tilde{l})\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ \tilde{l} & & \beta & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ \tilde{l} & \end{bmatrix}$$
$$+\sum_{\tilde{\beta}} C_3(\tilde{\beta})\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & \tilde{\beta} & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}, \tag{28}$$

where the three known coefficients are

$$C_1(\tilde{l})=\begin{bmatrix} \tilde{l}' & \beta & \tilde{l} \\ m & 0 & m \end{bmatrix}\begin{bmatrix} l & 1 & \tilde{l} \\ m & 0 & m \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ \tilde{l} & \end{bmatrix}\Big\| T^{(3)}(1)\Big\|\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix},$$
$$C_2(\tilde{l})=\begin{bmatrix} \tilde{l} & \beta & \tilde{l} \\ m & 0 & m \end{bmatrix}\begin{bmatrix} \tilde{l} & 1 & \tilde{l}' \\ m & 0 & m \end{bmatrix}\begin{bmatrix} v_1' & v_2' \\ l' & \end{bmatrix}\Big\| T^{(3)}(1)\Big\|\begin{bmatrix} v_1' & v_2' \\ \tilde{l} & \end{bmatrix}, \tag{29}$$
$$C_3(\tilde{\beta})=\begin{bmatrix} l' & \tilde{\beta} & l \\ m & 0 & m \end{bmatrix}\begin{bmatrix} \tilde{\beta} & 1 & \beta \\ 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ \beta & \end{bmatrix}\Big\| T^{(3)}(1)\Big\|\begin{bmatrix} 1 & 0 \\ \tilde{\beta} & \end{bmatrix}.$$

Therefore, inserting the needed Clebsch–Gordon coefficients (Table I) of $O(3)$ and the irreducible representations [Eqs. (23)–(25)] of $O(4)$ obtained previously into Eq. (29), we can solve Eq. (28) to obtain, by mathematical induction, all $O(4)\supset O(3)$ reduction factors,

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & q & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}, \quad q=0,1. \tag{30}$$

Using the method of highest weight decomposition or Schur function method,[2,16] we can get the selection rules for $O(4)\supset O(3)$ reduction factors, Eq. (30), that is, for a fixed $(v_1\ v_2)$, $(v_1'v_2')$, in Eq. (30) is allowed to take $(v_1\pm 1 v_2)$ and $(v_1 v_2\pm 1)$ only, and similarly, for a fixed l, l' in Eq. (30) is allowed to take l only for $q=0$, and to take l, $l\pm 1$ only for $q=1$. In order to be consistent with Eq. (5) here, we have adopted the following phase conventions for $O(4)\supset O(3)$ reduction factors:

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ \hat{l} & & 0 & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ \hat{l} & \end{bmatrix}>0,$$
$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & q & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}=\begin{bmatrix} 1 & 0 & v_1' & v_2' \\ q & & l' & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}. \tag{31}$$

The algebraic expressions for all $O(4)\supset O(3)$ reduction factors for $(v_1 v_2)\otimes(10)$, as a special case of $N=4$, are given in Table III.

In fact, in the process of solving Eq. (28), we have reduced the amount of computation due to the following equation:

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & 1 & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}=\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & 0 & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\Big\| T^{(3)}(1)\Big\|\begin{bmatrix} v_1 & v_2 \\ l' & \end{bmatrix}$$
$$-\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l & & 0 & \end{bmatrix}\begin{bmatrix} v_1 & v_2 \\ l & \end{bmatrix}\begin{bmatrix} v_1' & v_2' \\ l & \end{bmatrix}\Big\| T^{(3)}(1)\Big\|\begin{bmatrix} v_1' & v_2' \\ l' & \end{bmatrix}, \tag{32}$$

which can be obtained easily by taking $\beta=0$ in Eq. (28). Equation (32) gives a relation between the reduction factors

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & 1 & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix}$$

and the simpler ones

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l & & 0 & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix}.$$

Since

$$\begin{bmatrix} v_1' & v_2' \\ l & \end{bmatrix} \left\| T^{(3)}(1) \right\| \begin{bmatrix} v_1' & v_2' \\ l' & \end{bmatrix}$$

have been gained in the last section, we need merely to calculate the simpler reduction factors

$$\begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l & & 0 & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix}.$$

By means of Eq. (32), we can also obtain the symmetry relation satisfied by the $O(4)\supset O(3)$ reduction factors,

$$\begin{bmatrix} v_1 & v_2 & 1 & 0 \\ l & & q & \end{bmatrix} \begin{vmatrix} v_1' & v_2' \\ l' & \end{vmatrix} = (-1)^q \sqrt{\frac{d(v_1'v_2')d(l)}{d(v_1v_2)d(l')}} \begin{bmatrix} v_1' & v_2' & 1 & 0 \\ l' & & q & \end{bmatrix} \begin{vmatrix} v_1 & v_2 \\ l & \end{vmatrix}, \tag{33}$$

where $d(v_1v_2)$ and $d(l)$ stand for the dimensions of the irreducible representations (v_1v_2) of $O(4)$ and l of $O(3)$, respectively, namely,

$$\begin{aligned} d(v_1v_2)&=(v_1-v_2+1)(v_1+v_2+1), \\ d(l)&=2l+1. \end{aligned} \tag{34}$$

The symmetry relation, Eq. (33), can further reduce the amount of calculation required to construct tables of $O(4)\supset O(3)$ reduction factors greatly.

VI. FURTHER CALCULATIONS

From Section II to Section V, we have observed that the irreducible tensor basis method is a kind of recurrent method. That is, for the case in this paper, in terms of the Clebsch–Gordon coefficients of $O(3)$ for $l\otimes 1$, we first gain the irreducible representations (v_1v_2) of $O(4)$; then using the obtained irreducible representations of $O(4)$, we obtain all $O(4)\supset O(3)$ reduction factors for $(v_1v_2)\otimes(1\ 0)$. This method may go further to $O(5), O(6), \ldots$. We have calculated up to $O(6)$, and obtained the irreducible representations of $O(6)$ and the $O(6)\supset O(5)$ reduction factors, however, for economical pages, the procedure is no longer repeated here. Some important concepts and tools needed in the further evaluation of the irreducible representations and the reduction factors, such as the definition of an irreducible tensor operator, the generalized Wigner–Eckart theorem, and the explicit construction of the irreducible tensor bases of $O(5)$ and $O(6)$, are given in Appendices A, B, and C, respectively.

From the calculated results of $O(6)$, We have also found that there exists a very simple relation depending on rank N for both irreducible representations and reduction factors, therefore,

TABLE II. The Irreducible representations of $O(N)$.

n_1'	n_2'	n_3'	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{bmatrix} \Big\| T^{(N-1)} \Big\| \begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix}$
n_1+1	n_2	n_3	$\left\{\frac{(n_1-5+N)(m_1-n_1)(m_1+n_1-2+N)(n_1-m_2+1)(n_1+m_2-3+N)(n_1-m_3+2)(n_1+m_3-4+N)}{(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	$\left\{\frac{(n_1+2)(m_1-n_1+1)(m_1+n_1-3+N)(n_1-m_2)(n_1+m_2-4+N)(n_1-m_3+1)(n_1+m_3-5+N)}{(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	$\left\{\frac{(n_2-6+N)(m_1-n_2+1)(m_1+n_2-3+N)(m_2-n_2)(m_2+n_2-4+N)(n_2-m_3+1)(n_2+m_3-5+N)}{(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	$\left\{\frac{(n_2+1)(m_1-n_2+2)(m_1+n_2-4+N)(m_2-n_2+1)(m_2+n_2-5+N)(n_2-m_3)(n_2+m_3-6+N)}{(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	$\left\{\frac{(n_3-7+N)(m_1-n_3+2)(m_1+n_3-4+N)(m_2-n_3+1)(m_2+n_3-5+N)(m_3-n_3)(m_3+n_3-6+N)}{(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	$\left\{\frac{n_3(m_1-n_3+3)(m_1+n_3-5+N)(m_2-n_3+2)(m_2+n_3-6+N)(m_3-n_3+1)(m_3+n_3-7+N)}{(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

we can extend easily the irreducible representations of $O(6)$ for $(m_{16}m_{26}m_{36})$ and the reduction factors of $O(6)\supset O(5)$ for $(m_{16}m_{26}m_{36})\otimes(1\,0\,0)$ to the case of $O(N)$ but for the three-rowed irreducible representations. Thus we can gain by mathematical induction the irreducible representations of $O(N)$ for $(m_1m_2m_30\cdot0)$,

$$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{bmatrix} \Big\| T^{(N-1)} \Big\| \begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix},$$

and all $O(N)\supset O(N-1)$ reduction factors for $(m_1m_2m_30\cdot0)\otimes(1\,0\cdot0)$,

$$\left[\begin{array}{cccccc|cccccc} m_1 & m_2 & m_3 & 0 & \cdot & 0 & 1 & m_1' & m_2' & m_3' & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & & q & n_1 & n_2 & n_3 & 0 & \cdot & \end{array}\right], \quad q=0,1,$$

where $(m_1m_2m_30\cdot0)$ labels the three-rowed irreducible representations of $O(N)$ while $(n_1n_2n_30\cdot)$ labels the three-rowed irreducible representations of $O(N-1)$, and $(1)\equiv(1\,0\cdot0)$, $(q)\equiv(q0\cdot)$.

Of course, we can obtain directly the selection rules for the irreducible representations of $O(N)$ and $O(N)\supset O(N-1)$ reduction factors by use of the method of highest weight decomposition of Shur function method.[2,16] The selection rules for the irreducible representations of $O(N)$ are the following: for a fixed $(n_1n_2n_30\cdot0)$, $(n_1'n_2'n_3'0\cdot0)$ is allowed to take $(n_1\pm1n_2n_30\cdot0)$, $(n_1n_2\pm1n_30\cdot0)$ and $(n_1n_2n_3\pm10\cdot0)$. The selection rules for $O(N)\supset O(N-1)$ reduction factors are the following: for a fixed $(m_1m_2m_30\cdot0)$, $(m_1'm_2'm_3'0\cdot0)$ is allowed to take $(m_1\pm1m_2m_30\cdot0)$, $(m_1m_2\pm1m_30\cdot0)$ and $(m_1m_2m_3\pm10\cdot0)$, and for a fixed $(n_1n_2n_30\cdot0)$, $(n_1'n_2'n_3'0\cdot0)$ is allowed to take $(n_1n_2n_30\cdot0)$ for $q=0$, and to take $(n_1\pm1n_2n_30\cdot0)$, $(n_1n_2\pm1n_30\cdot0)$ and $(n_1n_2n_3\pm10\cdot0)$ for $q=1$.

Similarly, the irreducible representations of $O(N)$ for $(m_1m_2m_30\cdot0)$ and $O(N)\supset O(N-1)$ reduction factors for $(m_1m_2m_30\cdot0)\otimes(1\,0\cdot0)$ have similar phase conventions as the irreducible representations $(m_{16}m_{26}m_{36})$ of $O(6)$ and $O(6)\supset O(5)$ reduction factors for $(m_{16}m_{26}m_{36})\otimes(1\,0\,0)$, respectively. In fact, all of them are consistent with the phase conventions of $O(3)$.

TABLE III. The reduction factors of $O(N)\supset O(N-1)$.

(a) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 & 1 \\ n_1' & n_2' & n_3' & 0 & \cdot & & q \end{bmatrix} \begin{vmatrix} m_1+1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{vmatrix}$
n_1	n_2	n_3	0	$\left\{\frac{(m_1-n_1+1)(m_1+n_1-2+N)(m_1-n_2+2)(m_1+n_2-3+N)(m_1-n_3+3)(m_1+n_3-4+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	1	$\left\{\frac{(m_1-n_1+1)(m_1-n_2+2)(m_1+n_2-3+N)(m_1-n_3+3)(m_1+n_3-4+N)(n_1-5+N)(m_1-n_1)(n_1-m_2+1)(n_1+m_2-3+N)(n_1-m_3+2)(n_1+m_3-4+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$\left\{\frac{(m_1+n_1-2+N)(m_1-n_2+2)(m_1+n_2-3+N)(m_1-n_3+3)(m_1+n_3-4+N)(n_1+2)(m_1+n_1-3+N)(n_1-m_2)(n_1+m_2-4+N)(n_1-m_3+1)(n_1+m_3-5+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$\left\{\frac{(m_1-n_1+1)(m_1+n_1-2+N)(m_1-n_2+2)(m_1-n_3+3)(m_1+n_3-4+N)(n_2-6+N)(m_1-n_2+1)(m_2-n_2)(m_2+n_2-4+N)(n_2-m_3+1)(n_2+m_3-5+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$\left\{\frac{(m_1-n_1+1)(m_1+n_1-2+N)(m_1+n_2-3+N)(m_1-n_3+3)(m_1+n_3-4+N)(n_2+1)(m_1+n_2-4+N)(m_2-n_2+1)(m_2+n_2-5+N)(n_2-m_3)(n_2+m_3-6+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$\left\{\frac{(m_1-n_1+1)(m_1+n_1-2+N)(m_1-n_2+2)(m_1+n_2-3+N)(m_1-n_3+3)(n_3-7+N)(m_1-n_3+2)(m_2-n_3+1)(m_2+n_3-5+N)(m_3-n_3)(m_3+n_3-6+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$\left\{\frac{(m_1-n_1+1)(m_1+n_1-2+N)(m_1-n_2+2)(m_1+n_2-3+N)(m_1+n_3-4+N)n_3(m_1+n_3-5+N)(m_2-n_3+2)(m_2+n_3-6+N)(m_3-n_3+1)(m_3+n_3-7+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4-N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

TABLE III. (*Continued.*)

(b) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 & 1 \\ n_1' & n_2' & n_3' & 0 & \cdot & & q \end{bmatrix} \begin{vmatrix} m_1-1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{vmatrix}$
n_1	n_2	n_3	0	$\left\{\frac{(m_1-n_1)(m_1+n_1-3+N)(m_1-n_2+1)(m_1+n_2-4+N)(m_1-n_3+2)(m_1+n_3-5+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	0	$-\left\{\frac{(m_1+n_1-3+N)(m_1-n_2+1)(m_1+n_2-4+N)(m_1-n_3+2)(m_1+n_3-5+N)(n_1-5+N)(m_1+n_1-2+N)(n_1-m_2+1)(n_1+m_2-3+N)(n_1-m_3+2)(n_1+m_3-4+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$-\left\{\frac{(m_1-n_1)(m_1-n_2+1)(m_1+n_2-4+N)(m_1-n_3+2)(m_1+n_3-5+N)(n_1+2)(m_1-n_1+1)(n_1-m_2)(n_1+m_2-4+N)(n_1-m_3+1)(n_1+m_3-5+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$-\left\{\frac{(m_1-n_1)(m_1+n_1-3+N)(m_1+n_2-4+N)(m_1-n_3+2)(m_1+n_3-5+N)(n_2-6+N)(m_1+n_2-3+N)(m_2-n_2)(m_2+n_2-4+N)(n_2-m_3+1)(n_2+m_3-5+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$-\left\{\frac{(m_1-n_1)(m_1+n_1-3+N)(m_1-n_2+1)(m_1-n_3+2)(m_1+n_3-5+N)(n_2+1)(m_1-n_2+2)(m_2-n_2+1)(m_2+n_2-5+N)(n_2-m_3)(n_2+m_3-6+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$-\left\{\frac{(m_1-n_1)(m_1+n_1-3+N)(m_1-n_2+1)(m_1+n_2-4+N)(m_1+n_3-5+N)(n_3-7+N)(m_1+n_3-4+N)(m_2-n_3+1)(m_2+n_3-5+N)(m_3-n_3)(m_3+n_3-6+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$-\left\{\frac{(m_1-n_1)(m_1+n_1-3+N)(m_1-n_2+1)(m_1+n_2-4+N)(m_1-n_3+2)n_3(m_1-n_3+3)(m_2-n_3+2)(m_2+n_3-6+N)(m_3-n_3+1)(m_3+n_3-7+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

TABLE III. (*Continued.*)

(c) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix} 1 \Big\vert \begin{matrix} m_1 & m_2+1 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{matrix} \Big]$
n_1	n_2	n_3	0	$\left\{\frac{(n_1-m_2)(n_1+m_2-3+N)(m_2-n_2+1)(m_2+n_2-4+N)(m_2-n_3+2)(m_2+n_3-5+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	1	$\left\{\frac{(n_1-m_2)(m_2-n_2+1)(m_2+n_2-4+N)(m_2-n_3+2)(m_2+n_3-5+N)(n_1-5+N)(m_1-n_1)(m_1+n_1-2+N)(n_1-m_2+1)(n_1-m_3+2)(n_1+m_3-4+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$-\left\{\frac{(n_1+m_2-3+N)(m_2-n_2+1)(m_2+n_2-4+N)(m_2-n_3+2)(m_2+n_3-5+N)(n_1+2)(m_1-n_1+1)(m_1+n_1-3+N)(n_1+m_2-4+N)(n_1-m_3+1)(n_1+m_3-5+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$\left\{\frac{(n_1-m_2)(n_1+m_2-3+N)(m_2-n_2+1)(m_2-n_3+2)(m_2+n_3-5+N)(n_2-6+N)(m_1-n_2+1)(m_1+n_2-3+N)(m_2-n_2)(n_2-m_3+1)(n_2+m_3-5+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$\left\{\frac{(n_1-m_2)(n_1+m_2-3+N)(m_2+n_2-4+N)(m_2-n_3+2)(m_2+n_3-5+N)(n_2+1)(m_1-n_2+2)(m_1+n_2-4+N)(m_2+n_2-5+N)(n_2-m_3)(n_2+m_3-6+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$\left\{\frac{(n_1-m_2)(n_1+m_2-3+N)(m_2-n_2+1)(m_2+n_2-4+N)(m_2-n_3+2)(n_3-7+N)(m_1-n_3+2)(m_1+n_3-4+N)(m_2-n_3+1)(m_3-n_3)(m_3+n_3-6+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$\left\{\frac{(n_1-m_2)(n_1+m_2-3+N)(m_2-n_2+1)(m_2+n_2-4+N)(m_2+n_3-5+N)n_3(m_1-n_3+3)(m_1+n_3-5+N)(m_2+n_3-6+N)(m_3-n_3+1)(m_3+n_3-7+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

TABLE III. (*Continued.*)

(d) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix} 1 \left\vert \begin{matrix} m_1 & m_2-1 & m_3 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{matrix} \right. q \Big]$
n_1	n_2	n_3	0	$\left\{\frac{(n_1-m_2+1)(n_1+m_2-4+N)(m_2-n_2)(m_2+n_2-5+N)(m_2-n_3+1)(m_2+n_3-6+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	1	$\left\{\frac{(n_1+m_2-4+N)(m_2-n_2)(m_2+n_2-5+N)(m_2-n_3+1)(m_2+n_3-6+N)(n_1-5+N)(m_1-n_1)(m_1+n_1-2+N)(n_1+m_2-3+N)(n_1-m_3+2)(n_1+m_3-4+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$-\left\{\frac{(n_1-m_2+1)(m_2-n_2)(m_2+n_2-5+N)(m_2-n_3+1)(m_2+n_3-6+N)(n_1+2)(m_1-n_1+1)(m_1+n_1-3+N)(n_1-m_2)(n_1-m_3+1)(n_1+m_3-5+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$-\left\{\frac{(n_1-m_2+1)(n_1+m_2-4+N)(m_2+n_2-5+N)(m_2-n_3+1)(m_2+n_3-6+N)(n_2-6+N)(m_1-n_2+1)(m_1+n_2-3+N)(m_2+n_2-4+N)(n_2-m_3+1)(n_2+m_3-5+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$-\left\{\frac{(n_1-m_2+1)(n_1+m_2-4+N)(m_2-n_2)(m_2-n_3+1)(m_2+n_3-6+N)(n_2+1)(m_1-n_2+2)(m_1+n_2-4+N)(m_2-n_2+1)(n_2-m_3)(n_2+m_3-6+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$-\left\{\frac{(n_1-m_2+1)(n_1+m_2-4+N)(m_2-n_2)(m_2+n_2-5+N)(m_2+n_3-6+N)(n_3-7+N)(m_1-n_3+2)(m_1+n_3-4+N)(m_2+n_3-5+N)(m_3-n_3)(m_3+n_3-6+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$-\left\{\frac{(n_1-m_2+1)(n_1+m_2-4+N)(m_2-n_2)(m_2+n_2-5+N)(m_2-n_3+1)n_3(m_1-n_3+3)(m_1+n_3-5+N)(m_2-n_3+2)(m_3-n_3+1)(m_3+n_3-7+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

TABLE III. (*Continued.*)

(e) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix} 1 \left\vert \begin{matrix} m_1 & m_2 & m_3+1 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{matrix} \right] q$
n_1	n_2	n_3	0	$\left\{\frac{(n_1-m_3+1)(n_1+m_3-4+N)(n_2-m_3)(n_2+m_3-5+N)(m_3-n_3+1)(m_3+n_3-6+N)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	1	$\left\{\frac{(n_1-m_3+1)(n_2-m_3)(n_2+m_3-5+N)(m_3-n_3+1)(m_3+n_3-6+N)(n_1-5+N)(m_1-n_1)(m_1+n_1-2+N)(n_1-m_2+1)(n_1+m_2-3+N)(n_1-m_3+2)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$-\left\{\frac{(n_1+m_3-4+N)(n_2-m_3)(n_2+m_3-5+N)(m_3-n_3+1)(m_3+n_3-6+N)(n_1+2)(m_1-n_1+1)(m_1+n_1-3+N)(n_1-m_2)(n_1+m_2-4+N)(n_1+m_3-5+N)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$\left\{\frac{(n_1-m_3+1)(n_1+m_3-4+N)(n_2-m_3)(m_3-n_3+1)(m_3+n_3-6+N)(n_2-6+N)(m_1-n_2+1)(m_1+n_2-3+N)(m_2-n_2)(m_2+n_2-4+N)(n_2-m_3+1)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$-\left\{\frac{(n_1-m_3+1)(n_1+m_3-4+N)(n_2+m_3-5+N)(m_3-n_3+1)(m_3+n_3-6+N)(n_2+1)(m_1-n_2+2)(m_1+n_2-4+N)(m_3-n_2+1)(m_2+n_2-5+N)(n_2+m_3-6+N)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$\left\{\frac{(n_1-m_2+1)(n_1+m_3-4+N)(n_2-m_3)(n_2+m_3-5+N)(m_3-n_3+1)(n_3-7+N)(m_1-n_3+2)(m_1+n_3-4+N)(m_2-n_3+1)(m_2+n_3-5+N)(m_3-n_3)}{(m_3+1)(2m_3-6+N)(m_1+m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$\left\{\frac{(n_1-m_3+1)(n_1+m_3-4+N)(n_2-m_3)(n_2+m_3-5+N)(m_3+n_3-6+N)n_3(m_1-n_3+3)(m_1+n_3-5+N)(m_2-n_3+2)(m_2+n_3-6+N)(m_3+n_3-7+N)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

TABLE III. (*Continued.*)

(f) n_1'	n_2'	n_3'	q	$\begin{bmatrix} m_1 & m_2 & m_3 & 0 & \cdot & 0 \\ n_1' & n_2' & n_3' & 0 & \cdot & \end{bmatrix} 1 \Big\vert \begin{matrix} m_1 & m_2 & m_3-1 & 0 & \cdot & 0 \\ n_1 & n_2 & n_3 & 0 & \cdot & \end{matrix} \Big] q$
n_1	n_2	n_3	0	$\left\{\frac{(n_1-m_3+2)(n_1+m_3-5+N)(n_2-m_3+1)(n_2+m_3-6+N)(m_3-n_3)(m_3+n_3-7+N)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}\right\}^{1/2}$
n_1+1	n_2	n_3	1	$\left\{\frac{(n_1+m_3-5+N)(n_2-m_3+1)(n_2+m_3-6+N)(m_3-n_3)(m_3+n_3-7+N)(n_1-5+N)(m_1-n_1)(m_1+n_1-2+N)(n_1-m_2+1)(n_1+m_2-3+N)(n_1+m_3-4+N)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1-1	n_2	n_3	1	$-\left\{\frac{(n_1+m_3+2)(n_2-m_3+1)(n_2+m_3-6+N)(m_3-n_3)(m_3+n_3-7+N)(n_1+2)(m_1-n_1+1)(m_1+n_1-3+N)(n_1+m_2)(n_1+m_2-4+N)(n_1-m_3+1)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_1-3+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_1-n_3+2)(n_1+n_3-5+N)}\right\}^{1/2}$
n_1	n_2+1	n_3	1	$\left\{\frac{(n_1-m_3+2)(n_1+m_3-5+N)(n_2+m_3-6+N)(m_3-n_3)(m_3+n_3-7+N)(n_2-6+N)(m_1-n_2+1)(m_1+n_2-3+N)(m_2-n_2)(m_2+n_2-4+N)(n_2+m_3-5+N)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2-1	n_3	1	$-\left\{\frac{(n_1-m_3+2)(n_1+m_3-5+N)(m_2-m_3+1)(m_3-n_3)(m_3+n_3-7+N)(n_2+1)(m_1-n_2+2)(m_1+n_2-4+N)(m_2-n_2+1)(m_2+n_2-5+N)(n_2-m_3)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_2-5+N)(n_1-n_2+1)(n_1+n_2-4+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3+1	1	$-\left\{\frac{(n_1-m_3+2)(n_1+m_3-5+N)(n_2-m_3+1)(n_2+m_3-6+N)(m_3+n_3-7+N)(n_3-7+N)(m_1-n_3+2)(m_1+n_3-4+N)(m_2-n_3+1)(m_2+n_3-5+N)(m_3+n_3-6+N)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$
n_1	n_2	n_3-1	1	$-\left\{\frac{(n_1-m_3+2)(n_1+m_3-5+N)(n_2-m_3+1)(n_2+m_3-6+N)(m_3-n_3)n_3(m_1-n_3+3)(m_1+n_3-5+N)(m_2-n_3+2)(m_2+n_3-6+N)(m_3-n_3+1)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2n_3-7+N)(n_1-n_3+2)(n_1+n_3-5+N)(n_2-n_3+1)(n_2+n_3-6+N)}\right\}^{1/2}$

The algebraic expressions for the irreducible representations of $O(N)$ for $(m_1m_2m_30\cdot 0)$ are given in Table II, and the algebraic expressions for $O(N)\supset O(N-1)$ reduction factors for $(m_1m_2m_30\cdot 0)\otimes(1\,0\cdot 0)$ are given in Table III. The orthonormality conditions satisfied by reduction factors have been checked by virtue of software ''MATHEMATICA.''

VII. CONCLUSION

Using the irreducible tensor basis method, we obtain, by a recurrent approach, the algebraic expressions for the irreducible representations of $O(N)$ for $(m_1m_2m_30\cdot 0)$ and all reduction factors of $O(N)\supset O(N-1)$ for $(m_1m_2m_30\cdot 0)\otimes(1\,0\cdot 0)$. These results are elegant and practical. The irreducible spinor representations of $O(N)$ and corresponding reduction factors of $O(N)\supset O(N-1)$ will be given elsewhere.

Compared with other methods, the irreducible tensor basis method in this paper is simple and distinct. This method can not only be generalized to the more general cases of $O(N)$ but also be used to discuss the classical Lie algebras systematically. This is being studied.

ACKNOWLEDGMENTS

The authors would like to thank Professor Q. Z. Han and Professor G. L. Long for helpful discussions. The project was supported by the Chinese National Natural Science Foundation.

APPENDIX A: THE IRREDUCIBLE TENSOR OPERATOR OF $O(N)$

We denote the generators of $O(N)$ by $X_k^{(N)}$ $(k=1,2,...,N(N-1)/2)$. The set $\mathbf{T}^{(N)}(\Gamma_N)$ is said to be an irreducible tensor operator of rank Γ_N of $O(N)$, if under the operations of $O(N)$ its linearly independent components $T^{(N)}\binom{\Gamma_N}{\gamma_N}$ transform according to the irreducible representations Γ_N of $O(N)$, that is,

$$\left[X_k^{(N)},\ T^{(N)}\begin{pmatrix}\Gamma_N\\ \gamma_N\end{pmatrix}\right]=\sum_{\gamma_N'} T^{(N)}\begin{pmatrix}\Gamma_N\\ \gamma_N'\end{pmatrix}\begin{bmatrix}\Gamma_N\\ \gamma_N'\end{bmatrix} X_k^{(N)}\Bigg|\begin{matrix}\Gamma_N\\ \gamma_N\end{matrix}\Bigg],$$

where

$$\begin{bmatrix}\Gamma_N\\ \gamma_N'\end{bmatrix} X_k^{(N)}\Bigg|\begin{matrix}\Gamma_N\\ \gamma_N\end{matrix}\Bigg]$$

is the representation matrix of $O(N)$ and γ_N stands for the components of Γ_N. It can also be shown that the number of components of $\mathbf{T}^{(N)}(\Gamma_N)$ is equal to the dimension $d\Gamma_N$ of the irreducible representation Γ_N of $O(N)$.[14,2,16]

APPENDIX B: THE GENERALIZED WIGNER–ECKART THEOREM

For the typical case in this paper, we have to evaluate the matrix elements of the irreducible tensor operator $\mathbf{T}^{(N-1)}(\widetilde{\Gamma}_{N-1})$ of rank $\widetilde{\Gamma}_{N-1}$ of $O(N-1)$ in the Gel'fand–Tsetlin basis of $O(N)$.[24] Application of the generalized Wigner–Eckart theorem to the matrix elements of $\mathbf{T}^{(N-1)}\times(\widetilde{\Gamma}_{N-1})$, whose components have been conveniently labeled by the Gel'fand pattern of $O(N-1)$, gives rise to

$$\left[\begin{array}{c|c|c} \Gamma_N & \widetilde{\Gamma}_{N-1} & \Gamma_N \\ \Gamma_{N-1} & \widetilde{\Gamma}_{N-2} & \Gamma'_{N-1} \\ \Gamma_{N-2} & \cdot & \Gamma'_{N-2} \\ \cdot & \widetilde{\Gamma}_2 & \cdot \\ \Gamma_2 & & \Gamma'_2 \end{array}\right] T^{(N-1)} = \left[\begin{array}{c|c|c} \Gamma_N & \widetilde{\Gamma}_{N-1} & \Gamma_{N-1} \\ \Gamma_{N-1} & \widetilde{\Gamma}_{N-2} & \Gamma_{N-2} \\ \cdot & \cdot & \cdot \\ \Gamma'_2 & \widetilde{\Gamma}_2 & \Gamma_2 \end{array}\right],$$

where Γ_N stands for the irreducible representations of $O(N)$, Γ_{N-1} stands for those of $O(N-1)$, and so forth. Here

$$\left[\begin{array}{c} \Gamma_N \\ \Gamma_{N-1} \end{array}\right\| T^{(N-1)}(\widetilde{\Gamma}_{N-1}) \left\|\begin{array}{c} \Gamma_N \\ \Gamma'_{N-1} \end{array}\right]$$

are the $T^{(N-1)}(\widetilde{\Gamma}_{N-1})$ reduced matrix elements and

$$\left[\begin{array}{c|c|c} \Gamma'_{N-1} & \widetilde{\Gamma}_{N-1} & \Gamma_{N-1} \\ \Gamma'_{N-2} & \widetilde{\Gamma}_{N-2} & \Gamma_{N-2} \\ \cdot & \cdot & \cdot \\ \Gamma'_2 & \widetilde{\Gamma}_2 & \Gamma_2 \end{array}\right]$$

are the coupling coefficients of $O(N-1)$. Using Racah's factorization lemma,[25,2] the coupling coefficients of $O(N-1)$ can be decomposed into a product of $N-3$ reduction factors of $O(i)\supset O(i-1)$ $(i=3,4,...,N-1)$.

APPENDIX C: THE IRREDUCIBLE TENSOR BASIS OF $O(5)$ AND $O(6)$

1. The irreducible tensor basis of $O(5)$

The irreducible tensor basis of $O(5)$ includes the following two parts: (1) The generators $X_p^{(4)}$ $(p=1,2,...,6)$ of $O(4)$; (2) a set of irreducible tensors of rank (1 0) of $O(4)$, $\mathbf{T}^{(4)}(1\,0)$, with components

$$T^{(4)}(\beta_{-1}) = -i\sqrt{\tfrac{1}{2}}\,(I_{52}+iI_{51}), \quad T^{(4)}(\beta_{+1}) = -i\sqrt{\tfrac{1}{2}}\,(I_{52}-iI_{51}),$$

$$T^{(4)}(\beta_3) = I_{53}, \quad T^{(4)}(\beta_4) = iI_{54},$$

where β_i $(i=-1,+1,3,4)$ are the Gel'fand patterns labeling the basis vectors carrying the basic representations (1 0) of $O(4)$,

$$\beta_{-1} = \begin{array}{ccc} 1 & & 0 \\ & 1 & \\ -1 & & \end{array}, \quad \beta_{+1} = \begin{array}{ccc} 1 & & 0 \\ & 1 & \\ & 1 & \end{array},$$

$$\beta_3 = \begin{array}{ccc} 1 & & 0 \\ & 1 & \\ 0 & & \end{array}, \quad \beta_4 = \begin{array}{ccc} 1 & & 0 \\ & 0 & \\ & 0 & \end{array}.$$

Thus, the commutation relations of the generators of $O(5)$ are the following.

(1) The commutation relations of the generators $X_p^{(4)}$ of $O(4)$.
(2) The commutation relations between $X_p^{(4)}$ and $\mathbf{T}^{(4)}(1\,0)$. According to the definition of irreducible tensor of $O(4)$ they are

$$[X_p^{(4)},\ T^{(4)}(\beta_i)]=\sum_j T^{(4)}(\beta_j)[\beta_j|X_p^{(4)}|\beta_i].$$

The basic representation matrix, $[\beta_j|X_p^{(4)}|\beta_i]$, of $O(4)$ have been obtained in Section IV.

(3) The commutation relations between the components of $\mathbf{T}^{(4)}(1\ 0)$. Using Eq. (1), we can obtain immediately

$$[T^{(4)}(\beta_{-1}),\ T^{(4)}(\beta_{+1})]=-L_0,\quad [T^{(4)}(\beta_{-1}),\ T^{(4)}(\beta_3)]=-L_{-1},$$
$$[T^{(4)}(\beta_{-1}),\ T^{(4)}(\beta_4)]=T^{(3)}_{-1};\quad [T^{(4)}(\beta_{+1}),\ T^{(4)}(\beta_3)]=L_{+1},$$
$$[T^{(4)}(\beta_{+1}),\ T^{(4)}(\beta_4)]=T^{(3)}_{+1};\quad [T^{(4)}(\beta_3),\ T^{(4)}(\beta_4)]=T^{(3)}_0.$$

2. The irreducible tensor basis of $O(6)$

The irreducible tensor basis of $O(6)$ includes the following two parts: (1) The generators $X_q^{(5)}$ $(q=1,2,...,10)$ of $O(5)$; (2) a set of irreducible tensors of rank $(1\ 0)$ of $O(5)$, $\mathbf{T}^{(5)}(1\ 0)$, with components

$$T^{(5)}(\gamma_{-1})=-\sqrt{\tfrac{1}{2}}\,(I_{62}+iI_{61}),\quad T^{(5)}(\gamma_{+1})=-\sqrt{\tfrac{1}{2}}\,(I_{62}-iI_{61}),$$

$$T^{(5)}(\gamma_3)=-iI_{63},\quad T^{(5)}(\gamma_4)=I_{64},\quad T^{(5)}(\gamma_5)=iI_{65},$$

where γ_i $(i=-1,+1,3,4,5)$ are the Gel'fand patterns labeling the basis vectors carrying the basic representation $(1\ 0)$ of $O(5)$, namely,

$$\gamma_{-1}=\begin{matrix}1&0\\1&0\\1&\\-1&\end{matrix},\quad \gamma_{+1}=\begin{matrix}1&0\\1&0\\1&\\1&\end{matrix},\quad \gamma_3=\begin{matrix}1&0\\1&0\\1&\\0&\end{matrix},$$

$$\gamma_4=\begin{matrix}1&0\\1&0\\0&\\0&\end{matrix},\quad \gamma_5=\begin{matrix}1&0\\0&0\\0&\\0&\end{matrix}.$$

Thus, the commutation relations of the generators of $O(6)$ are the following.

(1) The commutation relations of the generators $X_q^{(5)}$ of $O(5)$.

(2) The commutation relations between $X_q^{(5)}$ and $\mathbf{T}^{(5)}(1\ 0)$. According to the definition of the irreducible tensor of $O(5)$, they are

$$[X_q^{(5)},\ T^{(5)}(\gamma_i)]=\sum_j T^{(5)}(\gamma_j)[\gamma_j|X_q^{(5)}|\gamma_i].$$

(3) The commutation relations between the components of $\mathbf{T}^{(5)}(1\ 0)$. By direct calculation, we have

$$[T^{(5)}(\gamma_{-1}),\ T^{(5)}(\gamma_{+1})]=L_0,\quad [T^{(5)}(\gamma_{-1}),\ T^{(5)}(\gamma_3)]=-L_{-1},$$
$$[T^{(5)}(\gamma_{-1}),\ T^{(5)}(\gamma_4)]=-T^{(3)}_{-1},\quad [T^{(5)}(\gamma_{-1}),\ T^{(5)}(\gamma_5)]=T^{(4)}(\beta_{-1});$$
$$[T^{(5)}(\gamma_{+1}),\ T^{(5)}(\gamma_3)]=L_{+1},\quad [T^{(5)}(\gamma_{+1}),\ T^{(5)}(\gamma_4)]=-T^{(3)}_{+1},$$
$$[T^{(5)}(\gamma_{+1}),\ T^{(5)}(\gamma_5)]=T^{(4)}(\beta_{+1});\quad [T^{(5)}(\gamma_3),\ T^{(5)}(\gamma_4)]=-T^{(3)}_0$$
$$[T^{(5)}(\gamma_3),\ T^{(5)}(\gamma_5)]=T^{(4)}(\beta_3),\quad [T^{(5)}(\gamma_4),\ T^{(5)}(\gamma_5)]=T^{(4)}(\beta_4).$$

[1] B. R. Judd, *Operator Techniques in Atomic Spectroscopy* (McGraw–Hill, New York, 1963).
[2] B. G. Wybourne, *Classical Groups for Physicists* (John Wiley, New York, 1974).

[3] A. Arima and F. Iachello, Ann. Phys. (N.Y.) **123**, 468 (1979).
[4] P. Van Isacker, A. Frank, and H. Z. Sun, Ann. Phys. (N.Y.) **157**, 183 (1984).
[5] G. L. Long, J. Phys. A **28**, 64 (1995).
[6] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton University Press, New York, 1957).
[7] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison–Wesley, Reading MA, 1981).
[8] L. C. Biedenharn, J. Math. Phys. **2**, 433 (1961).
[9] A. M. Gavrilik, Kiev Institute of Theoretical Physics, preprint ITP-73-155P, 1973 (in Russian); also see preprint ITP-77-26E, 1973 (in English).
[10] E. Z. Norvaisas and S. J. Ališauskas, Liet. Fiz. Rinkinys **14**, 443 (1974).
[11] S. J. Ališauskas, Part. Nuclei **14**, 1336 (1983); Sov. J. Part. Nucl. **14**, 563 (1983).
[12] S. J. Ališauskas, J. Phys. A **20**, 35 (1987).
[13] FPan and Y. F. Cao, J. Math. Phys. **29**, 2384 (1988).
[14] G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963); **5**, 1723,1730 (1964); **6**, 1845 (1965).
[15] H. Z. Sun, Sci. Sinica **14**, 840 (1965).
[16] Q. Z. Han and H. Z. Sun, *Group Theory* (Peking University Press, Beijing, 1987).
[17] H. Z. Sun, Phys. E. Fort. Phys. Nucl. **4**, 73,137,265 (1980).
[18] H. Z. Sun and Q. Z. Han, Sci. Sinica **24**, 914 (1981).
[19] Q. Z. Han, Nuovo Cimento A **64**, 391 (1981).
[20] Q. Z. Han, H. Z. Sun, M. Zhang, and D. H. Feng, J. Math. Phys. **26**, 1822 (1985).
[21] Z. R. Yu, O. Scholten, and H. Z. Sun, J. Math. Phys. **27**, 442 (1986).
[22] Q. Z. Han, E. G. Zhao, and H. Z. Sun, J. Phys. A **22**, 1495 (1989).
[23] H. Z. Sun and D. Ruan, preprint AS-ITP-96-14, 1996 [including some practical tables for $O(N)\supset O(N-1)$ reduction factors].
[24] I. M. Gel'fand, R. A. Minlos, and Z. Ya. Shapiro, *Representations of the Rotation and Lorentz Groups and Their Applications* (Pergamon Press, New York, 1963).
[25] G. Racah, Group Theory and Spectroscopy, Lectures at the Institute for Advanced Study (Princeton University Press, New York, 1951).

Automorphisms of $sl(2)$ and dynamical r-matrices

A. V. Tsiganov[a)]
Department of Mathematical and Computational Physics, Institute of Physics, St. Petersburg University, 198 904, St. Petersburg, Russia

(Received 11 December 1996; accepted for publication 16 September 1997)

Two outer automorphisms of the space of infinite-dimensional representations of $sl(2)$ algebra are considered. The similar constructions for the loop algebras and Yangians are presented. The corresponding linear and quadratic R-brackets include the dynamical r-matrices. © *1998 American Institute of Physics.* [S0022-2488(98)00701-4]

I. INTRODUCTION

We are concerned with the representation theory of loop algebras and Yangians and with the method of separation of variables in framework of the inverse scattering method. Our aim is to apply the representation theory of $sl(2)$ and $sl(N)$ algebras into the r-matrix formalism.

Let us start with the following Levi-Civita theorem:[1] if the Hamilton–Jacobi equation associated with Hamiltonian,

$$H=g^{ij}(x_1,\ldots,x_n)p_ip_j+h^j(x_1,\ldots,x_n)p_j+U(x_1,\ldots,x_n),\quad \{p_i,x_j\}=\delta_{ij}, \tag{I.1}$$

can be integrated by separation of variables then the equation is integrated with $U=0$ (I.1), i.e., in the absence of a force, in the same system of coordinates.

This paper is devoted to the similar problem. Consider a classical Hamiltonian system completely integrable on a $2n$-dimensional symplectic manifold D. It means that the system possesses n independent integrals $I_1,\ldots,I_n$ in the involution,

$$\{I_i(x,p),I_j(x,p)\}=0,\quad i,j=1,\ldots,n,$$

where $\{x_j,p_j\}_{j=1}^n$ is some coordinate system in D.

Introduce the following mapping:

$$I_j\to I'_j=\sum_i^n a_{ij}(x,p)\cdot[I_i+v_i(x,p)], \tag{I.2}$$

here a_{ij} and v_j are certain functions on D. In contrast with the Calogero and Toda models[2] we shall consider special mappings (I.2), which preserve the phase space of initial integrable model.

Under the suitable conditions for functions a_{ij} and v_j, the mapping (I.2) could define a new integrable system on D with independent integrals $I'_1,\ldots,I'_n$ in the involution

$$\{I'_i(x,p),I'_j(x,p)\}=0,\quad i,j=1,\ldots,n,$$

and could preserve the property of separability in the same coordinate system.

Remember that systems discovered by Stäckel (see Ref. 2) have the similar properties. Their integrals are

a)Electronic mail: tsiganov@mph.niif.spb.su

0022-2488/98/39(1)/650/15/$15.00
© 1998 American Institute of Physics

$$I_j=\sum_{i=1}^{n} a_{ij}(x_1,\ldots,x_n)[p_i^2+v_i(x_i)], \tag{I.3}$$

with the special functions $a_{ij}(x_1,\ldots,x_n)$ and the arbitrary functions $v_i(x_i)$.

We try to make a first step to understand the algebraic roots of constraints on the functions a_{ij} and v_j, which guarantee the properties of integrability and separability for mapping (I.2) in the framework of the inverse scattering method.

II. OUTER AUTOMORPHISMS OF REPRESENTATIONS OF $SL(2)$

Let W be an infinite-dimensional representation of the Lie algebra $sl(2)$ in linear space V defined in the Cartan–Weyl basis $\{s_3,s_\pm\}\in \mathrm{End}(V)$ equipped with the natural bracket,

$$[s_3,s_\pm]=\pm s_\pm, \quad [s_+,s_-]=2s_3, \tag{II.4}$$

and the single Casimir operator,

$$\mathbf{\Delta}=s_3^2+\frac{1}{2}(s_+s_-+s_-s_+). \tag{II.5}$$

If operator s_+ is invertible in $\mathrm{End}(V)$ then the mapping

$$\begin{aligned} &s_3\to s_3'=s_3, \quad s_+\to s_+'=s_+,\\ &s_-\to s_-'=s_-+fs_+^{-1}, \quad f\in\mathbb{C}, \end{aligned} \tag{II.6}$$

is an outer automorphism of the space of infinite-dimensional representations of $sl(2)$ in V. The mapping (II.6) shifts the spectrum of $\mathbf{\Delta}$ on the parameter f,

$$\Delta\to\Delta'=\Delta+f. \tag{II.7}$$

Let us call the mapping (II.6) an additive automorphism.

Assuming in addition the value of Casimir operator $\mathbf{\Delta}$ is equal to zero $\Delta=0$ in W, then the mapping

$$\begin{aligned} &s_j\to s_j'=s_j\cdot(1-fs_+^{-1}), \quad j=\pm,3, \quad f\in\mathbb{C},\\ &\mathbf{\Delta}\to\mathbf{\Delta}'=\mathbf{\Delta}\cdot(1-fs_+^{-1})^2, \quad \Delta=\Delta'=0, \end{aligned} \tag{II.8}$$

is another outer automorphism of representation W. Let us call the mapping (II.8) a multiplicative automorphism.

These special infinite-dimensional representations W could be obtained by using the previous additive automorphism (II.6). Notice that the similar representations with $\Delta=0$ are well known in the quantum conformal field theory.[3]

We can suppose that additive (II.6) and multiplicative (II.8) automorphisms define the one-parametric realizations $W(f)$ of $sl(2)$.

For instance, the realization of $sl(2)$ with one free parameter f in the classical mechanics is given by

$$s_3=\frac{xp}{2}, \quad s_+=\frac{x^2}{2}, \quad s_-=-\frac{p^2}{2}+\frac{f}{x^2}, \quad \Delta'=f, \tag{II.9}$$

where (x,p) is a pair of canonical coordinate and momenta with the classical Poisson bracket $\{p,x\}=1$. Similar realization of generators s_j as the differential operators in quantum case is

$$s_3=x\partial_x-l, \quad s_+=x, \quad s_-=x\partial_x^2-2l\partial_x+\frac{f}{x}, \quad \Delta'=l(l+1)+f. \tag{II.10}$$

For the quantum Calogero system the more sophisticated representation of $sl(2)$ is equal to

$$s_3=\frac{1}{4}\sum_{k=1}^{N}(x_kD_k+D_kx_k), \quad s_+=2\sum_{k=1}^{N}x_k^2, \quad s_-=-2\sum_{k=1}^{N}D_k^2,$$

where x_k are coordinates and D_k are the corresponding Dunkl operators.[4]

Using the inner automorphisms of $sl(2)$ the one parametric mappings (II.6) and (II.8) can be generalized. Let W be an infinite-dimensional representation of $sl(2)$ in V defined by operators $s_j\in\mathrm{End}(V)$ with the standard bracket and the single quadratic Casimir operator $\mathbf{\Delta}$,

$$[s_i,s_j]=\varepsilon_{ijk}s_k, \quad \mathbf{\Delta}=s_1^2+s_2^2+s_3^2. \tag{II.11}$$

Introduce the general linear operator $b\in\mathrm{End}(V)$ as

$$b=\alpha_1s_1+\alpha_2s_2+\alpha_3s_3, \quad \alpha_j\in\mathbb{C}, \tag{II.12}$$

where α_j are three arbitrary parameters. If b is invertible operator, then the mapping

$$s_j\to s_j'=s_j+\alpha_jb^{-1} \tag{II.13}$$

is an outer automorphism of the space of infinite-dimensional representations of $sl(2)$ in V, if the parameters α_j lie on the cone

$$\alpha_1^2+\alpha_2^2+\alpha_3^2=0. \tag{II.14}$$

For this mapping the spectrum of the Casimir operator $\mathbf{\Delta}$ (II.11) is additively shifted on the parameter $f=\Sigma\alpha_j$,

$$\Delta\to\Delta'=\Delta+f=\Delta+\sum_{j=1}^{3}\alpha_j. \tag{II.15}$$

If the value of the Casimir operator $\mathbf{\Delta}$ is equal to zero $\Delta=0$ and the general linear operator $b\in\mathrm{End}(V)$ (II.12) is invertible, then the mapping,

$$s_j\to s_j'=s_j(1-b^{-1}), \quad \mathbf{\Delta}\to\mathbf{\Delta}'=\mathbf{\Delta}\cdot(1-b^{-1})^2, \tag{II.16}$$

is an outer multiplicative automorphism of representation W.

Automorphisms (II.13) and (II.16) define the general two- and three-parametric realizations $W(\alpha_j)$ of $sl(2)$.

For the $sl(N)$ algebra the similar additive automorphism shifted the highest central element of N-th order can be taken in the form

$$s_{ij}\to s_{ij}'=s_{ij}+\frac{A_{ij}^{(m-1)}}{B^{(m)}}, \quad s_{ij}\in\mathrm{End}(V), \quad m=N(N-1)/2. \tag{II.17}$$

In operators s_{ij} functions $B^{(m)}$ and $A_{ij}^{(m-1)}$ are certain polynomials of degrees m and $m-1$. Here we shall not go into details of this problem, one example related to $sl(3)$ will be presented in Section IV.

Motivated by realization (II.9) we present one application of automorphism (II.6) in the theory of integrable systems. Consider a classical Hamiltonian system completely integrable on the phase space $D=\mathbb{R}^{2n}$ with the natural Hamiltonian,

$$H=\sum_{j=1}^{n} p_j^2+V(x_1,\ldots,x_n).$$

Let the phase space be identified completely or partially with the m coadjoint orbits in $sl(2)^*$ as (II.9). Then the mapping,

$$H\to H'=H+\sum_{j=1}^{m}\frac{f_j}{x_j^2},\quad f_j\in\mathbb{R}, \tag{II.18}$$

preserves the properties of integrability and separability. The list of such systems can be found in Ref. 2.

Our main aim is to develop similar constructions for the loop algebras and Yangians. Some examples of the integrable systems related with this approach have been considered in Refs. 5–7.

III. DYNAMICAL *R*-MATRICES ASSOCIATED TO $\widetilde{SL(2)}$

All details of the general r-matrix scheme can be found in a review[8] and in references therein. We recall briefly necessary elements of r-matrix scheme for loop algebra $\mathfrak{g}=\widetilde{sl(2)}$ to assume the standard identification of the dual spaces.

The loop algebra $\mathfrak{g}=\widetilde{sl(2)}$ consists of Laurent polynomials $s_j(\lambda)=\Sigma_k s_j\lambda^k$, $j=1,2,3$ of a spectral parameter λ with coefficients in $\mathfrak{a}=sl(2)$ and commutator $[s_i\lambda^l,s_j\lambda^m]=\varepsilon_{ijk}s_k\lambda^{l+m}$. The standard R-bracket associated with loop algebra $\mathfrak{g}$ is defined by the following decomposition of $\mathfrak{g}$ into a linear sum of two subalgebras:

$$\mathfrak{g}=\mathfrak{g}_+\dot{+}\mathfrak{g}_-,\quad \mathfrak{g}_+=\oplus_{i\geqslant 0}\mathfrak{a}\lambda^i,\quad \mathfrak{g}_-=\oplus_{i<0}\mathfrak{a}\lambda^i,\quad R=P_+-P_-. \tag{III.1}$$

Here R is a standard r-matrix and $P_\pm$ means the projection operators onto $\mathfrak{g}_\pm$ parallel $\mathfrak{g}_\mp$.

The Lax equation may be presented in the form

$$\frac{dL(\lambda)}{dt}=-\mathrm{ad}^*_{\mathfrak{g}}M\cdot L,\quad M=\frac{1}{2}R(dP(L)), \tag{III.2}$$

where the Lax matrix $L(\lambda)$ belongs to $\mathfrak{g}^*$, P is an ad*-invariant polynomial on $\mathfrak{a}^*$. For an algebra $\widetilde{sl(2)}$ polynomial $P(L)$ is a function of the unique invariant polynomial $\Delta(\lambda)=\Sigma_{j=1}^{3}s_j^2(\lambda)$. Let $P(L)=\phi(\lambda)\Delta(\lambda)$, where ϕ is a function of λ. The integrals of motion I_k related to flow (III.2) are

$$I_k(L)=\mathrm{Res}_{\lambda=0}(\phi_k(\lambda)\Delta(\lambda)), \tag{III.3}$$

where $\phi_k(\lambda)$ are various functions of spectral parameter defining a complete set of integrals of motion.[8]

The Lax matrix $L(\lambda)$ (III.2) is defined on the whole infinite-dimensional phase space $\mathfrak{g}^*$. It is well known that the standard R-bracket associated with (III.1) has also a large collection of finite-dimensional Poisson (ad^*_R-invariant) subspaces,[8]

$$\mathscr{L}_{M,N}=\oplus_{j=-M}^{N}\mathfrak{a}^{*}\lambda^{j}, \quad \text{provided } M\geqslant 0;\ N\geqslant 1, \tag{III.4}$$

and, as a rule, the concrete physical systems are related to the restrictions of the flow (III.2) to certain low-dimensional Poisson submanifolds $\mathscr{L}_{M,N}$ (III.4).

The r-matrix scheme is extended easily to the twisted subalgebras of loop algebra $\mathfrak{g}$ and corresponding r-matrices have rational, trigonometric and elliptic dependence on the spectral parameter. We shall work with a tensor form of the Lax equations and R-bracket on $\widetilde{sl(2)}$, that allows us to consider all the r-matrices simultaneously. In addition, for brevity, we shall consider the $\widetilde{sl(2,\mathbb{R})}$ only.

The Lax matrix $L(\lambda)$ (III.2) associated to the Hamiltonian (III.3) is given by

$$L(\lambda)=\sum_{k=1}^{3} s_k(\lambda)\sigma_k, \quad \Delta(\lambda)=\frac{1}{2}\,\mathrm{tr}L^{2}(\lambda)=-\det L(\lambda), \tag{III.5}$$

where σ_j are Pauli matrices. The R-bracket related to decomposition (III.1) take the following form:

$$\{\overset{1}{L}(\lambda),\overset{2}{L}(\mu)\}=[r_{12}(\lambda,\mu),\overset{1}{L}(\lambda)]-[r_{21}(\lambda,\mu),\overset{2}{L}(\mu)], \tag{III.6}$$

where the standard notations are introduced:

$$\begin{gathered}\overset{1}{L}(\lambda)=L(\lambda)\otimes I, \quad \overset{2}{L}(\mu)=I\otimes L(\mu),\\ r_{12}(\lambda,\mu)=\sum_{k=1}^{3} w_k(\lambda,\mu)\cdot\sigma_k\otimes\sigma_k, \quad r_{21}(\lambda,\mu)=\Pi r_{12}(\mu,\lambda)\Pi.\end{gathered} \tag{III.7}$$

Here Π is the permutation operator of auxiliary spaces and $w_k(\lambda,\mu)$ are certain functions of spectral parameters only. Their explicit dependence on λ,μ is not important for the moment.

Returning now to the additive automorphism of representations of $sl(2)$ (II.13) we introduce the related mappings on the loop algebras $\mathfrak{g}=\widetilde{sl(2,\mathbb{R}}$ or $\mathfrak{g}=\oplus^{n}\widetilde{sl(2,\mathbb{R}}$.

The entries of the Lax matrix are Laurent polynomials with coefficients in $sl(2)$. Let us consider the mapping transforming these coefficients. If parameters a_j lie on the cone (II.14), then the first mapping similar to the automorphism (II.13),

$$s_j(\lambda)=\sum_k s_j\lambda^{k}\to s_j'(\lambda)=\sum_k [s_j+\alpha_j b^{-1}]\lambda^{k}, \quad b=\sum_{j=1}^{3}\alpha_j s_j, \tag{III.8}$$

is a canonical change of variables on $sl(2)$. This mapping is defined for the infinite-dimensional representations of $sl(2)$ and mapping (III.8) is a Poisson map with respect to the first natural Lie–Poisson bracket and to the second linear Lie–Poisson bracket associated with the standard R-matrix (III.1) on $\widetilde{sl(2)}$. For the certain Lax representation $L(\lambda)$ mapping (III.8) transforms the integrals of motion I_k similar to (II.18).

Let us introduce the general linear operator $b(\lambda)\in\widetilde{sl(2)}$,

$$b(\lambda)=\sum_{j=1}^{3}\alpha_j(\lambda)s_j(\lambda), \tag{III.9}$$

where $\alpha_j(\lambda)$ are functions of spectral parameter and of central elements on $\widetilde{sl(2)}$. If functions $\alpha_j(\lambda)$ lie on the cone,

$$\alpha_1^2(\lambda)+\alpha_2^2(\lambda)+\alpha_3^2(\lambda)=0, \tag{III.10}$$

then the second mapping similar to the automorphism (II.13),

$$s_j(\lambda)\to s_j'(\lambda)=s_j(\lambda)+\alpha_j(\lambda)b^{-1}(\lambda), \tag{III.11}$$

is a Poisson map with respect to the first natural Lie–Poisson bracket. For the mapping (III.11) the Lax matrix (III.5), the invariant polynomial $\Delta(\lambda)$ and integrals of motion I_k (III.3) are additively shifted,

$$L(\lambda)\to L'(\lambda)=L(\lambda)+\sum_{k=1}^{3}\alpha_k\sigma_k\cdot b^{-1}(\lambda), \tag{III.12}$$

$$\Delta(\lambda)\to\Delta'(\lambda)=\Delta(\lambda)+V(\lambda)=\Delta(\lambda)+\sum_{k=1}^{3}\alpha_k(\lambda), \tag{III.13}$$

$$I_k\to I_k'=I_k+U_k,\quad U_k=\mathrm{Res}_{\lambda=0}(\phi_k(\lambda)V(\lambda))\in\mathbb{C}. \tag{III.14}$$

Restriction (III.10) guarantees that initial integrals of motion I_k (III.3) transform by some constants $U_k\in\mathbb{C}$ (I.2). Hence, two Lax matrices $L(\lambda)$ and $L'(\lambda)$ [the original and the image of mapping (III.11)] correspond to the same integrable system.

The entries of the initial matrix $L(\lambda)$ belong to $\mathfrak{g}^*$. The entries of the matrix $L'(\lambda)$ are the Laurent polynomials of spectral parameter λ with coefficients from the universal enveloping algebra of $\mathfrak{a}^*$. So, to begin with the element of the loop algebra we construct special element of the universal enveloping algebra $U(\mathfrak{g}^*)$ defined by the following algebra homomorphism:

$$U(\mathfrak{g}^*)\to\mathbb{C}[\lambda,\lambda^{-1}]\otimes U(\mathfrak{a}^*).$$

Nevertheless, the second Poisson bracket $\{L'(\lambda),L'(\mu)\}$ can be directly calculated, because of all the necessary Poisson brackets between $s_j(\lambda)$ and $b(\lambda)$ are preassigned by (III.6).

Further we consider a special class of the linear dynamical R-bracket related to $\widetilde{sl(2)}$ and defined by the following second restriction on coefficients $\alpha_j(\lambda)$:

$$\{b(\lambda),b(\mu)\}=g(\lambda,\mu)b(\lambda)-g(\mu,\lambda)b(\mu),\quad \lambda,\mu\in\mathbb{C}, \tag{III.15}$$

or, that is equivalent,

$$w_j(\lambda,\mu)\alpha_j(\mu)\alpha_i(\lambda)-w_i(\lambda,\mu)\alpha_i(\mu)\alpha_j(\lambda)=g(\lambda,\mu)\cdot\alpha_k(\lambda), \tag{III.16}$$

where (j,i,k) are cyclic permutations of indices (1,2,3) and the scale function $g(\lambda,\mu)$ depends on the spectral parameters only. This restriction is closely related with the separation of variables method. It guarantees that all zeroes of $b(\lambda)$ are mutually commuting. Below we assume that both conditions (III.10) and (III.15) are always fulfilled for the mapping (III.11).

In this case we get the Lax matrix $L'(\lambda)$; (III.12) obeys the linear R-bracket (III.6), where constant r_{ij}-matrices substituted by r'_{ij}-matrices depending on dynamical variables,

$$r_{12}(\lambda,\mu)\to r'_{12}=r_{12}+\sum_{i,j=1}^{3}\alpha_{ij}(\lambda,\mu)\sigma_i\otimes\sigma_j, \tag{III.17}$$

with coefficients α_{ij} being

$$\alpha_{ij}(\lambda,\mu)=g(\lambda,\mu)\,\frac{\alpha_j(\mu)\alpha_i(\mu)w_i(\lambda,\mu)}{b^2(\mu)}-g(\mu,\lambda)\,\frac{\alpha_i(\lambda)\alpha_j(\lambda)w_j(\lambda,\mu)}{b^2(\lambda)}. \qquad \text{(III.18)}$$

For the proof see Ref. 9.

Dynamical matrices $r'_{jk}(\lambda,\mu)$ (III.17) obey the classical dynamical Yang–Baxter equation,

$$[r'_{12}(\lambda,\mu),r'_{13}(\lambda,\nu)]+[r'_{12}(\lambda,\mu),r'_{23}(\mu,\nu)]+[r'_{32}(\nu,\mu),r'_{13}(\lambda,\nu)]+[L'_2(\mu),r'_{13}(\lambda,\nu)]$$
$$-[L'_3(\nu),r'_{12}(\lambda,\mu)]+[X_{123}(\lambda,\mu,\nu),L'_2(\mu)-L'_3(\nu)]=0, \qquad \text{(III.19)}$$

where an explicit expression of the tensor $X(\lambda,\mu,\nu)$ is not important for the moment (see Refs. 5, 9, and 10). This equation is the image of a standard 2-cocycle in the universal enveloping algebra of the loop algebra, which consists of Laurent polynomials of spectral parameter λ with coefficients from the corresponding universal enveloping algebra of the Lie algebra.

We can suppose that mapping (III.11) defines a family of Lax matrices $L'(\lambda)$ and dynamical r-matrices (III.17) with a fixed set of parameters $\alpha_j(\lambda)$ for certain integrable system. Thus, the Lax matrix $L(\lambda)$ (III.2), the constant r-matrices (III.7) and the standard classical Yang–Baxter equation on $\widetilde{sl(2)}$ are the limits of the dynamical ones at $\alpha_j(\lambda)=0$.

The mapping (III.11) allows us to construct an infinite set of the Lax matrices associated to the loop algebra $\widetilde{sl(2)}$, which correspond to different integrable systems. According to Ref. 9 we introduce an infinite set of mappings,

$$s_j(\lambda)\to s'_j(\lambda)=s_j(\lambda)+[\alpha_j(\lambda)b^{-1}(\lambda)]_{M_jN_j}, \qquad \text{(III.20)}$$

where $[z]_{MN}$ means the restriction of z onto the certain Poisson subspace $\mathscr{S}_{MN}$ (III.4) of the standard R-bracket (III.1).

As an example, we can use the linear combinations of the following Laurent and Fourier projections:

$$[z]_{MN}=\left[\sum_{k=-\infty}^{+\infty} z_k\lambda^k\right]_{MN}\equiv\sum_{k=-M}^{N} z_k\lambda^k,$$
$$[z]_{MN}=\left[\sum_{k=-\infty}^{+\infty} z_k\exp(k\cdot\lambda)\right]_{MN}\equiv\sum_{k=-M}^{N} z_k\exp(k\cdot\lambda), \qquad \text{(III.21)}$$

if the corresponding r-matrices have rational or trigonometric dependence on a spectral parameter.[9]

Generally speaking, we can not describe now all the possible restrictions of the mapping (III.11) to certain low-dimensional submanifolds, which allow us to define the second Poisson bracket on $\widetilde{sl(2)}$. However, for a fairly large class of properly defined projections (III.21) the mappings (III.20) are the Poisson maps with respect to the second R-bracket and the corresponding dynamical r-matrices obey the dynamical Yang–Baxter equation (III.19).

For instance, an application of projections (III.21) yields dynamical r-matrices,

$$r_{12}(\lambda,\mu)\to r'_{12}=r_{12}+\sum_{i,j=1}^{3}\alpha_{ij}(\lambda,\mu)\sigma_i\otimes\sigma_j, \qquad \text{(III.22)}$$

with the following coefficients:

$$\alpha_{ij}(\lambda,\mu)=g(\lambda,\mu)w_i(\lambda,\mu)\left[\frac{\alpha_j(\mu)\alpha_i(\mu)}{b^2(\mu)}\right]_{MN}-g(\mu,\lambda)w_j(\lambda,\mu)\left[\frac{\alpha_i(\lambda)\alpha_j(\lambda)}{b^2(\lambda)}\right]_{MN}. \tag{III.23}$$

The essential feature of the restricting mappings (III.20) is, in comparison with the mapping (III.11), that the invariant polynomial $\Delta(\lambda)$ and all integrals of motion I_k are shifted now on the items depending on dynamical variables,

$$\begin{aligned}\Delta(\lambda)&\to\Delta_{MN}(\lambda)=\Delta(\lambda)+V_{MN}(s_j,\alpha_j,\lambda),\\ I_k&\to I_k'=I_k+\mathrm{Res}_{\lambda=0}(\phi_k(\lambda)V_{MN}(s_j,\alpha_j,\lambda)),\end{aligned} \tag{III.24}$$

with V_{MN} being

$$V_{MN}(s_j,\alpha_j,\lambda)=\sum_{k=1}^{3}(2s_k(\lambda)+[a_j(\lambda)b^{-1}(\lambda)]_{MN})[a_j(\lambda)b^{-1}(\lambda)]_{MN}. \tag{III.25}$$

Hence, images of integrals of motion I_k' of the mapping (III.20) are functionally different from the original ones I_k.

The second feature of the mapping (III.20) is that it can be also applied to the finite-dimensional representations of $sl(2)$. Consider the multipole Lax matrices related to the rational r-matrix.[8] Let $\mathfrak{g}=\oplus^n\widetilde{sl(2,\mathbb{R}}$ and elements of mapping (III.11) are

$$\alpha_1=1,\quad \alpha_2=i,\quad \alpha_3=0,\quad b(\lambda)=\sum_{k=1}^{n}\frac{s_+^{(k)}}{\lambda-e_k},\quad e_j\neq e_k\in\mathbb{R}. \tag{III.26}$$

The Taylor projection of $b^{-1}(\lambda)$ is determined by the following recurrence relations:[5]

$$[b^{-1}(\lambda)]_{MN}|_{M=0}=\sum_{k=0}^{N}\mathscr{T}_k\lambda_{N-k},\quad \mathscr{T}_k=\sum_{i=1}^{n}\left(s_+^{(i)}\sum_{j=0}^{k-1}\mathscr{T}_{k-1-j}e_i^j\right),\quad \mathscr{T}_0=1. \tag{III.27}$$

The Taylor projection (III.27) is the well-defined polynomial of the nilpotent operators $s_+^{(k)}$ without the negative powers, which can also be used for finite-dimensional representations of $sl(2)$.

Once again, the mappings (III.20) are not just isomorphisms of the Lax matrices $L(\lambda)$ and $L'(\lambda)$, in contrast with mapping (III.11), but they also preserve the second Poisson R-bracket together with all the commuting integrals of motion I_k', i.e., preserve the properties of integrability and separability.

Thus, the mappings (III.20) play the role of a dressing procedure allowing one to construct the Lax matrices $L'_{MN}(\lambda)$ for an infinite set of new integrable systems starting from the single known Lax matrix $L(\lambda)$ associated to one integrable model. Many well-known integrable mechanical systems, such as Henon–Heiles, Garnier, Neumann (and many others) can be embedded into the proposed scheme.[5,6,9,10]

Now we shall briefly discuss the second multiplicative automorphism (II.16). Let us start with some Lax matrix $L(\lambda)\in\widetilde{sl(2)}^*$. This matrix relates to the integrable system with the Lax representation (III.2) and with the following integrals of motion (III.3):

$$I_k=\mathrm{Res}_{\lambda=0}(\phi_k(\lambda)\Delta(\lambda)).$$

Recall that the multiplicative automorphism (II.16) of $sl(2)$ transforms the Casimir operator by the rule

$$\mathbf{\Delta}\to\mathbf{\Delta}'=\mathbf{\Delta}\cdot\varphi(b)=\mathbf{\Delta}\cdot(1-b^{-1})^2.$$

Motivated by this transformation we consider the new set of integrals I'_k [see (I.2)] defined by

$$I'_k = \mathrm{Res}_{\lambda=0}(\varphi_k(b,\lambda)\cdot\Delta(\lambda)), \tag{III.28}$$

where $\varphi_k(b,\lambda)$ are certain functions of $b(\lambda)$ (III.9) and of the spectral parameter λ. The involution conditions,

$$\{I'_i, I'_j\}=0, \quad i,j=1,\ldots,n,$$

yield the set of equations in $\varphi_k(b,\lambda)$,

$$\mathrm{Res}_{\lambda\mu=0}\{\Delta(\lambda),b(\mu)\}\varphi_j(\lambda)\varphi_k(\mu)\left(\frac{\partial \ln \varphi_j}{\partial b}(\lambda)\Delta(\lambda)-\frac{\partial \ln \varphi_k}{\partial b}(\mu)\Delta(\mu)\right)=0.$$

New integrals of motion I'_j could be functionally different from the original ones I_j.

In the quantum case the Poisson brackets should be replaced by the standard commutator relations on $sl(2)$ and the non-dynamical linear R-matrix bracket (III.6) becomes the following commutator relations:

$$[\overset{1}{L}(\lambda),\overset{2}{L}(\mu)]=[\mathfrak{r}_{12}(\lambda,\mu),\overset{1}{L}(\lambda)]-[\mathfrak{r}_{21}(\lambda,\mu),\overset{2}{L}(\mu)], \quad \mathfrak{r}(\lambda,\mu)=-i\hbar r(\lambda,\mu). \tag{III.29}$$

The theory of the general quantum linear or quadratic R-bracket with dynamical r-matrices is not well developed yet, but a nice feature of the presented special Poisson mappings (III.11) and (III.20) is that it admits a natural quantization. All the necessary commutator relations between operators $s_j(\lambda)$ and the linear invertible operator $b(\lambda)$ (III.9) are preassigned by (III.29) and, therefore, the introduction of the quantum dynamical r-matrices (III.17) is a straightforward calculation in this particular case. A similar calculation yields the direct quantum counterpart of the dynamical Yang–Baxter equation.

IV. DYNAMICAL *R*-MATRICES AND SEPARATION OF VARIABLES

The presented algebraic constructions are intimately connected with the method of separation of variables in classical mechanics. We shall use the technique developed by Sklyanin in the framework of an r-matrix formalism and based on the application of the Baker–Akhiezer function.[11] Recall, that the Baker–Akhiezer function $\Psi(\lambda)$ is the eigenvector of the Lax matrix,

$$L(\lambda)\Psi(\lambda)=z(\lambda)\Psi(\lambda),$$

corresponding to the eigenvalue $z(\lambda)$, which has certain analyticity properties. Since an eigenvector is defined up to a scalar factor, to exclude the ambiguity in the definition of $\Psi(\lambda)$ one has to fix a normalization of $\Psi(\lambda)$ imposing a linear constraint,

$$\sum_{j=1}^{N}\beta_j(\lambda)\Psi_j(\lambda)=1.$$

The main purpose of this Section is to find a correspondence between the outer automorphisms of infinite-dimensional representations of $sl(2)$ and the normalization of the Baker–Akhiezer function.

According to the Sklyanin recipe the variables of separation (x_j,p_j) are defined by the poles of the properly normalized Baker–Akhiezer function $\Psi(\lambda)$ and the corresponding eigenvalues $z(\lambda)$ of the Lax matrix.[11] For the Lax matrix $L(\lambda)\in\widetilde{sl(2)}^*$ the poles x_j of $\Psi(\lambda)$ are zeroes of the following function:

$$B(\lambda,\beta_1,\beta_2)=\beta_1^2(\lambda)s_+(\lambda)-\beta_2^2(\lambda)s_-(\lambda)-2\beta_1(\lambda)\beta_2(\lambda)s_3(\lambda). \tag{IV.1}$$

Introduce functions $\alpha_j(\lambda)$ as

$$\alpha_1(\lambda)=[(\beta_1^2-\beta_2^2)f](\lambda),\quad \alpha_2(\lambda)=i[(\beta_1^2+\beta_2^2)f](\lambda),\quad \alpha_3=-2[\beta_1\beta_2 f](\lambda), \tag{IV.2}$$

where $f(\lambda)$ is a certain common function of a spectral parameter only. The coefficients $\alpha_j(\lambda)$ lie on the cone (III.10) and define the linear operator $b(\lambda,\alpha_j)=B(\lambda,\beta_1,\beta_2)$ (III.9) for the mappings (III.11) and (III.20). The second restriction (III.15) on $\alpha_j(\lambda)$ guarantees that all zeroes x_j of $b(\lambda)$ are mutually commuting $\{x_j,x_k\}=0$.

For the linear R-bracket the eigenvalues $z(\lambda)$ of the Lax matrix $L(\lambda)$ corresponding to zeroes x_j are canonically conjugated momenta p_j. The pairs of variables (x_j,p_j) lie on the spectral curve,

$$W(p_j,x_j)=0,\quad W(z,\lambda)=\det(z-L(\lambda))=z^2-\Delta(\lambda), \tag{IV.3}$$

which fits exactly to separated equations.[11] The mappings (III.20) change the separated equations (IV.3) as

$$p_j^2-\Delta(x_j,I_k)=0\to p_j^2-\Delta'(x_j,I_k')=0, \tag{IV.4}$$

where $\Delta'(\lambda)$ and new integrals of motion I_k' are given by (III.24). Thus, by taking the single Lax matrix $L(\lambda)$ associated to some integrable system, which can be integrated by separation of variables in coordinates $\{x_j,p_j\}$, we get an infinite set of completely integrable systems determined by the mappings (III.20), which are separable in the same variables $\{x_j,p_j\}$.

The linear operator $b(\lambda)$ (III.9) is a symmetric function of its zeroes $\{x_j\}_{j=1}^n$ and dynamical r-matrices (III.17) depends of the spectral parameters and only half of the dynamical variables $\{x_j\}_{j=1}^n$. Moreover, if normalization $\beta(\lambda)$ of $\Psi(\lambda)$ is given by an arbitrary constant numeric vector, then three functions $\alpha_j(\lambda)$ (IV.2) differ by certain numeric constants β_j,

$$\alpha_j(\lambda)=\beta_j f(\lambda),\quad \beta_j\in\mathbb{C},\quad \sum \beta_j^2=0. \tag{IV.5}$$

Then by (IV.5) the invariant polynomial $\Delta(\lambda)$ and all the integrals of motion I_k (III.3) are shifted on items said to be potentials depending on separated coordinates x_j (I.2):

$$\begin{aligned}\Delta'&=\Delta+V_{MN}(x_1,\ldots,x_n)=\Delta+2b(x_j,\lambda)\left[\frac{f(\lambda)}{b(x_j,\lambda)}\right]_{MN},\\ I_k'&=I_k+U_k(x_1,\ldots,x_n)=I_k+\mathrm{Res}_{\lambda=0}(\phi_k(\lambda)V_{MN}(x_1,\ldots,x_n)).\end{aligned} \tag{IV.6}$$

Thus, all integrable systems related to the mappings (III.20) with the constant normalization of the corresponding Baker–Akhiezer function obey to the Levi-Civita theorem[1] (I.1).

Integrals of motion I_j' (III.28) associated to the multiplicative automorphism (II.16) could be different from original ones on certain functions $a_{ij}(x_1,\ldots,x_n)$ of only half of the dynamical variables $\{x_j\}_{j=1}^n$ [see (I.2) and the Stäckel systems (I.3)]. Here functions a_{ij} are defined by (III.28) for any concrete representation.

The method of separation of variables for the Lax matrices $L(\lambda)\in\widetilde{sl(N)}^*$ was studied in Ref. 11. The separated coordinates are obtained as zeroes of the certain polynomial $B(\lambda)$ of degree $N(N-1)/2$ in components of $L(\lambda)$. The spectral curve $W(z,\lambda)=\det(z-L(\lambda))$ is a nonhyperellip-

tic algebraic curve for $N>2$ and the Levi-Civita theorem can not be applied directly to this case. Still the polynomial $B(\lambda)$ depending on separated coordinates can be used to construct a counterpart of the mappings (III.11) and (III.20) for $\widetilde{sl(N)}$.

Consider, for instance, the loop algebra $\widetilde{sl(3)}$. Let the Lax matrix $L(\lambda)\in\widetilde{sl(3)}^*$ obey the standard linear R-bracket (III.6) with the rational r-matrix $r_{12}=(\lambda-\mu)^{-1}\Pi$, where Π is the permutation operator: $\Pi x\otimes y=y\otimes x, x,y\in\mathbb{C}^3$. The entries $s_{ij}(\lambda)$ of the Lax matrix $L(\lambda)$ are constructed in variables s_{ij} $(i,j=1,2,3.\ \Sigma s_{ii}=0)$ with the following standard Poisson brackets:

$$\{s_{ij},s_{km}\}=s_{im}\delta_{jk}-s_{kj}\delta_{im},$$

which define the natural Lie–Poisson bracket on $\widetilde{sl(3)}^*$. According to Ref. 11, the simplest choice of normalization $\beta(\lambda)$ of $\Psi(\lambda)$ is

$$\beta_1=\beta_2=0,\quad \beta_3=1, \tag{IV.7}$$

then the polynomial $B(\lambda)$ is given by

$$B(\lambda)=s_{32}(\lambda)u_{13}(\lambda)-s_{31}(\lambda)u_{23}(\lambda), \tag{IV.8}$$

where $u_{ij}(\lambda)$ is (ij)-cofactor of the determinant of $L(\lambda)$. The polynomial $B(\lambda)$ (IV.8) allows us to introduce the following mapping:

$$\begin{aligned} &s_{ij}\to s'_{ij}=s_{ij},\quad (ij)\neq(13),(23),\\ &s_{13}\to s'_{13}=s_{13}+s_{32}f(\lambda)B^{-1}(\lambda),\\ &s_{23}\to s'_{23}=s_{23}-s_{31}f(\lambda)B^{-1}(\lambda), \end{aligned} \tag{IV.9}$$

as an analog of the mapping (III.11) with the normalization (IV.7). This mapping leaves fixed the spectral invariants τ_1 and τ_2 and shifts the third invariant polynomial τ_3,

$$\begin{aligned} &\tau_1(\lambda)=\mathrm{tr}L(\lambda)=0\to\tau'_1(\lambda)=\tau_1(\lambda)=0,\\ &\tau_2(\lambda)=\mathrm{tr}L^2(\lambda)\to\tau'_2(\lambda)=\tau_2(\lambda),\\ &\tau_3(\lambda)=\det L(\lambda)\to\tau'_3(\lambda)=\tau_3(\lambda)+f(\lambda). \end{aligned}$$

The general Lax matrix $L'(\lambda)$ defined by (IV.9) obeys the linear R-bracket with the dynamical r-matrices.

So, we can apply the known polynomials $B(\lambda)$ and the method of separation of variables to construct the analogs of mappings (III.11) and (III.20) for the loop algebra $\widetilde{sl(N)}$. It would be interesting to find the general counterparts of outer automorphisms (II.13) and (II.16) for infinite-dimensional representations of $sl(N)$ and to consider the inverse problem of choosing the correct normalization of Baker–Akhiezer function by using these automorphisms. A similar, but more difficult problem arises for the simple Lie algebras other than $sl(N)$.

V. DYNAMICAL *R*-MATRICES AND QUADRATIC *R*-BRACKET

Now we consider analogs of the additive and multiplicative automorphisms of $sl(2)$ (II.6)–(II.8) for the quadratic R-bracket in classical mechanics.

Introduce the formal particular mapping,

$$T(u)=\begin{pmatrix} A & B \\ C & D \end{pmatrix}(u)\to T'(u)=\begin{pmatrix} A+fD^{-1} & B \\ C & D \end{pmatrix}(u). \tag{V.1}$$

If the initial matrix $T(u)$ obeys the standard quadratic R-bracket,

$$\{\overset{1}{T}(u),\overset{2}{T}(v)\}=[r(u-v),\overset{1}{T}(u)\overset{2}{T}(v)], \quad \text{with } r=(u-v)^{-1}\Pi, \tag{V.2}$$

then the image $T'(u)$ of mapping (V.1) obeys the dynamical quadratic R-bracket,

$$\{\overset{1}{T'}(u),\overset{2}{T'}(v)\}=[r(u-v),\overset{1}{T'}(u)\overset{2}{T'}(v)]+\overset{2}{T'}(v)s_{21}\overset{1}{T'}(u)-\overset{1}{T'}(u)s_{12}\overset{2}{T'}(v). \tag{V.3}$$

Here dynamical matrices s_{jk} are given by

$$s_{12}(u,v)=\frac{1}{u-v}\left[\frac{f(u)}{D^2(u)}-\frac{f(v)}{D^2(v)}\right]\cdot\sigma_-\otimes\sigma_+,$$
$$s_{21}(u,v)=\Pi s_{12}(v,u)\Pi. \tag{V.4}$$

In contrast to the linear case, the mapping (V.1) related to additive automorphism (II.6) changes the form of the R-bracket. However, if the functions $D(u)$ and $f(u)$ (V.1) are independent of spectral parameter u, then the dynamical matrices s_{jk} in (V.3) go to zero.[6] Moreover, in this case, we can use the more general mapping [see (V.1)],

$$A(u)\to A'(u)=A(u)+g(D),$$

where $g(D)$ is an arbitrary function on entry D. This mapping changes the standard R-bracket (V.2) to the dynamical bracket (V.3), but it preserves the property of integrability,

$$\{\mathrm{tr}T'(u),\mathrm{tr}T'(v)\}=0.$$

We have to emphasize that additive and multiplicative automorphisms of $sl(2)$ (II.6)–(II.8) give rise to the integrable systems associated to the two root systems BC_n and D_n, respectively.[6] Consider the special solutions,

$$U_A(u)=\begin{pmatrix} A(u) & B(u) \\ C(u) & -A(-u) \end{pmatrix}, \tag{V.5}$$

of the classical reflection equation,

$$\{\overset{1}{U}(u),\overset{2}{U}(v)\}=[r(u-v),\overset{1}{U}(u)\overset{2}{U}(v)]+\overset{1}{U}(u)r(u+v)\overset{2}{U}(v)-\overset{2}{U}(v)r(u+v)\overset{1}{U}(u). \tag{V.6}$$

with the rational $r=(u-v)^{-1}\Pi$ r-matrix. An application of the additive and multiplicative automorphisms of $sl(2)$ (II.6)–(II.8) is more tricky in the quadratic case.

If the entry $B(u)$ is independent of spectral parameter u, i.e. $\partial B/\partial u=0$, we can construct the new solution of the reflection equation (V.6),

$$U_C=U_A+\begin{pmatrix} 0 & 0 \\ \gamma B^{-1} & 0 \end{pmatrix}, \quad \gamma\in\mathbb{R}. \tag{V.7}$$

Assuming in addition that entry $A(u)$ is a linear function of u, i.e. $A(u)=ua_1+a_2$, let us introduce the second bounary matrix,

$$U_{BC}=U_C+\begin{pmatrix} \dfrac{\alpha}{u}+\beta & 0 \\ \left[\dfrac{\alpha}{u}(A(u)-A(-u))+\beta(A(u)+A(-u))\right]B^{-1} & \dfrac{\alpha}{u}-\beta \end{pmatrix},$$

$$\alpha,\beta,\in\mathbb{R}, \tag{V.8}$$

which is a new solution of the reflection equation (V.6). These solutions U_C (V.7) and U_{BC} (V.8) could be associated with the additive automorphism (II.6).

If the central element $\Delta=\det U_A(u)=0$ is equal to zero and $\partial B/\partial u=0$ the third boundary matrix,

$$U_D=\begin{pmatrix} A(u)-A(-u)B^{-1} & B\cdot(1-B^{-1})^2 \\ C(u) & -A(-u)+A(u)B^{-1} \end{pmatrix}, \tag{V.9}$$

is a solution of the reflection equation (V.6). This solution could be associated with the multiplicative automorphism (II.8).

The reflection equation (V.6) has an important covariance property: if $T(u)$ and $U(u)$ satisfy the relations (V.2) and (V.6), then the matrix

$$U'(u)=T(u)K(u)T^{-1}(-u), \tag{V.10}$$

is also a solution of the reflection equation provided that the Poisson brackets between the entries of initial matrices are equal to zero $\{K_{ij}(u),T_{km}(u)\}=0$. Using this property the special solution $U_A(u)$ (V.5) is given by

$$U_A(u)=T(u)\sigma_{\pm}T^{-1}(-u), \quad \Delta=\det U_A(u)=0, \tag{V.11}$$

$$\sigma_+=\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_-=\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

If the entry $T_{12}(u)$ $(T_{21}(u))$ is independent of the spectral parameter u, the entry U_{12} (U_{21}) is independent too. The list of the models with such matrices $T(u)$ includes the Heizenberg magnets, the nonlinear Schrödinger equation, the Toda lattices and many others.

Of course, taking another initial constant solution of the reflection equation in (V.11) the boundary matrices U_{BC} can be obtained by the usual rule[6] (V.10). However, the boundary matrix $U_D(u)$ can not be expanded on simplest factors.

As an example, here we consider the Toda lattices. The initial boundary matrix $U_A(u)$ is given by

$$U_A(u)=\begin{pmatrix} (u-p)\exp(q) & \exp(2q) \\ u^2-p^2 & (u+p)\exp(q) \end{pmatrix}, \tag{V.12}$$

where (q,p) is a pair of canonically conjugate variables $\{p,q\}=1$. According to Ref. 6 matrices U_A, U_{BC} and U_D correspond to the Toda lattices associated with the Lie algebras of the A_n, $B_n(\beta=\gamma=0)$, $C_n(\alpha=\beta=0)$ and D_n series, respectively. Among the Hamiltonians, in comparison with (I.2), there are

$$H_A=\frac{1}{2}\sum_{j=1}^{n} p_j^2+\sum_{j=1}^{n-1} \exp(x_{j+1}-x_j),$$

$$H_{BC}=H_A+\gamma \exp(-2x_1)+(2\alpha+2\beta p_1)\exp(-x_1),$$

$$H_D=H_A+\exp(-x_1-x_2).$$

In classical and quantum mechanics the boundary matrices (V.8) and (V.9) have been used for the relativistic Toda lattices and for the Heisenberg XXX and XXZ models in Ref. 6.

The more complicated dynamical quadratic R-bracket has been introduced in Ref. 7 for the Neumann top, Kowalewski top and Toda lattice associated to the Lie algebra G_2.

VI. CONCLUSIONS

The outer automorphisms of an infinite-dimensional representation of $sl(2)$ gives the possibility of constructing new Lax matrices. The corresponding linear and quadratic R-brackets include the dynamical r-matrices, which obey the dynamical Yang–Baxter equations.

For the loop algebras as a second step we applied the certain projection of a general Lax matrix onto the low-dimensional subspaces, which preserve the R-bracket. Thus, the set of the Lax matrices associated to the different integrable systems can be obtained.

Another natural way to understand the origin of dynamical r-matrices is to consider the reduction procedure. In this approach we start with some integrable system on the rich initial phase space supplied with a symplectic action of a symmetry group. By using a relatively simple invariant Hamiltonian and factorizing the corresponding dynamics by the symmetry group we get a smaller phase space with a nontrivial dynamic. Then the dynamical r-matrices describe the Poisson–Dirac brackets on reduced phase space. Of course, the automorphisms of the Lie algebras play an important role in this approach too.

In a proposed scheme we also begin with some integrable system and then construct another integrable system by using the automorphisms of Lie algebras. However, in contrast with Calogero models, the initial phase space is the same space for all considered systems. To apply the outer automorphisms of Lie algebras we construct Poisson mappings from one special representation of the loop algebra to special representations of the corresponding universal enveloping algebra. So, we preserve the phase space, but transform representations of algebras in the same linear space.

The similar dynamical deformations of quadratic R-bracket has been applied for the construction of Lax matrix for integrable systems associated to the root systems BC_n and D_n.

Among the known and possible examples there is a wide class of the Stäckel systems; the integrable extensions of the classical tops—Euler top, Manakov and Steklov tops, Lagrange and Goryachev–Chaplygin top; generalizations of the Heisenberg and Gaudin magnets.

The following problems, however, remain open. We have not an exhaustive description of outer automorphisms of infinite-dimensional representations of simple Lie algebras and of the all admissible low-dimensional submanifolds, which allow a well-defined restriction to the corresponding linear classical and quantum R-brackets.

ACKNOWLEDGMENTS

We would like to thank I. V. Komarov for reading the manuscript and making valuable comments. This research has been partially supported by RFBR Grant No. 96-0100537.

[1] T. Levi-Civita, Math. Ann. **59**, 383 (1904).
[2] A. Perelomov, *Integrable System of Classical Mechanics and Lie Algebras* (Birkhauser, Basel, 1991).
[3] L. N. Lipatov, JETP Lett. **59**, 596 (1994).
[4] C. Dunkl, Trans. Am. Math. Soc. **311**, 167 (1989).
[5] J. Eilbeck, V. Enolskii, V. Kuznetsov, and A. Tsiganov, J. Phys. A **27**, 567 (1994).

[6] A. Tsiganov, J. Phys. A **27**, 6759 (1994).
[7] A. Tsiganov, J. Math. Phys. **38**, 196 (1997).
[8] A. Reyman and M. Semenov-Tian Shansky, in *Dynamical Systems VII*, edited by V. Arnold and S. Novikov (Springer, Berlin, 1993).
[9] P. Kulish, S. Rauch-Wojciechowski, and A. Tsiganov, J. Math. Phys. **37**, 3365 (1996).
[10] Y. Zeng and J. Hietarinta, J. Phys. A **29**, 5241 (1996).
[11] E. Sklyanin, Prog. Theor. Phys. Suppl. **118**, 35 (1995).

Demazure characters and WZW fusion rules

Mark A. Walton[a)]
Department of Physics, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada

(Received 2 July 1997; accepted for publication 15 August 1997)

The Demazure character formula is applied to the Verlinde formula for the fusion rules of Wess–Zumino–Witten (WZW) models. We follow Littelmann's derivation of a generalized Littlewood–Richardson rule that computes tensor products for simple Lie algebras. A combinatorial rule for WZW fusions does not result, however. Only a modified version of the Littlewood–Richardson rule is obtained that computes an (old) upper bound on the fusion coefficients of A_r WZW models. We argue that this is because the characters of simple Lie algebras appear in this treatment, instead of the corresponding affine characters. The Bruhat order on the affine Weyl group must be implicated in any combinatorial rule for WZW fusions; the Bruhat order on subgroups of this group (such as the finite Weyl group) does not suffice. © *1998 American Institute of Physics.* [S0022-2488(97)02612-1]

I. INTRODUCTION

The fusion coefficients $N_{l,m}^n \in \mathbf{Z}_{\geqslant 0}$ of a rational conformal field theory (RCFT) may be defined by[1]

$$\frac{S_{l,p}}{S_{0,p}}\frac{S_{m,p}}{S_{0,p}} = \sum_{n\in\Phi} N_{l,m}^n \frac{S_{n,p}}{S_{0,p}}, \quad \forall p \in \Phi. \tag{1}$$

Here, the indices $l,m,n,p,\ldots \in \Phi$ label the primary fields of the RCFT, with the index 0 specifying the identity field. $S_{i,j}$ denotes an element of the matrix describing the transformation of the characters of the primary fields under the change of torus modulus $\tau \to -1/\tau$.

We will restrict attention to those RCFTs realizing an affine Kac–Moody algebra $X_{r,k}$ that is the central extension at fixed level k of the loop algebra of the simple Lie algebra X_r of rank r. These are often called Wess–Zumino–Witten (WZW) models.[2]

If the chiral algebra is not extended, the primary fields of such theories are in one-to-one correspondence with the integrable highest-weight representations of $X_{r,k}$. The set of highest weights of these affine representations can therefore be used to label the primary fields. Alternatively, we can label the primary fields using the following set of dominant weights of X_r:

$$P_{\geqslant}(X_{r,k}) := \left\{ \lambda = \sum_{i=1}^{r} \lambda_i \omega^i \;\middle|\; \lambda_i \in \mathbf{Z}_{\geqslant 0},\ \sum_{i=1}^{r} \lambda_i a_i^\vee \leqslant k \right\}. \tag{2}$$

ω^i denotes the ith fundamental weight of X_r, and $a_i^\vee$ is the corresponding comark, so that $1 + \Sigma_{i=1}^r a_i^\vee = h^\vee$, the dual Coxeter number. $F(X_r)$ will stand for the set of fundamental weights of X_r. The WZW fusion rules may be written as

$$\frac{S_{\lambda,\zeta}}{S_{0,\zeta}}\frac{S_{\mu,\zeta}}{S_{0,\zeta}} = \sum_{\nu\in P_{\geqslant}(X_{r,k})} N_{\lambda,\mu}^{\nu} \frac{S_{\nu,\zeta}}{S_{0,\zeta}}, \quad \forall \zeta \in P_{\geqslant}(X_{r,k}). \tag{3}$$

[a)]Supported in part by NSERC. Electronic mail: walton@hg.uleth.ca

The ratios of matrix elements of S appearing in (3) are intimately related to the characters of representations of X_r.[3] Let $R(\lambda)$ be the integrable representation of X_r with highest weight λ, and let $P(\lambda)$ denote the set of its weights. If

$$P(X_r) := \left\{ \lambda = \sum_{i=1}^{r} \lambda_i \omega^i \mid \lambda_i \in \mathbf{Z} \right\} \tag{4}$$

is the set of integral weights of X_r, we have $P(\lambda) \subset P(X_r)$. Let e^λ denote the formal exponential of the weight λ, with the property $e^\lambda e^\mu = e^{\lambda+\mu}$. The formal character of $R(\lambda)$ may be written as

$$\mathrm{ch}_\lambda := \sum_{\sigma \in P(\lambda)} \mathrm{mult}_\lambda(\sigma) e^\sigma, \tag{5}$$

where $\mathrm{mult}_\lambda(\sigma)$ is the multiplicity of the weight σ in the representation $R(\lambda)$, and $\rho = \sum_{i=1}^{r} \omega^i$ is the Weyl vector.

Define

$$e^\sigma(\zeta) := \exp\left[\frac{-2\pi i}{k+h^\vee} \, \sigma \cdot (\zeta + \rho) \right]. \tag{6}$$

Kac and Peterson showed that

$$\frac{S_{\lambda,\zeta}}{S_{0,\zeta}} = \mathrm{ch}_\lambda(\zeta) \quad (\lambda, \zeta \in P_{\geqslant}(X_{r,k})). \tag{7}$$

This result allows us to relate the WZW fusion coefficients $N_{\lambda,\mu}^{\nu}$ to the coefficients appearing in the decomposition of a tensor product of two representations of X_r. Let

$$P_{\geqslant}(X_r) := \left\{ \lambda = \sum_{i=1}^{r} \lambda_i \omega^i \middle| \lambda_i \in \mathbf{Z}_{\geqslant 0} \right\} \tag{8}$$

be the set of dominant weights of X_r. Then, if the tensor product coefficients $T_{\lambda,\mu}^{\nu}$ are defined by

$$R(\lambda) \otimes R(\mu) = \bigoplus_{\nu \in P_{\geqslant}(X_r)} T_{\lambda,\mu}^{\nu} R(\nu), \tag{9}$$

the corresponding characters obey

$$\mathrm{ch}_\lambda \ \mathrm{ch}_\mu = \sum_{\nu \in P_{\geqslant}(X_r)} T_{\lambda,\mu}^{\nu} \ \mathrm{ch}_\nu . \tag{10}$$

The Weyl character formula is

$$\mathrm{ch}_\lambda = \frac{\Sigma_{w \in W} (\det w) e^{w(\lambda+\rho)}}{\Sigma_{w \in W} (\det w) e^{w\rho}}, \tag{11}$$

where W is the Weyl group of X_r. Therefore,

$$\mathrm{ch}_\lambda(\zeta) = \frac{\Sigma_{w \in W} (\det w) \zeta^{w(\lambda+\rho)}}{\Sigma_{w \in W} (\det w) \zeta^{w\rho}}, \tag{12}$$

where we have defined

$$\zeta^{\sigma} := e^{\sigma}(\zeta), \tag{13}$$

for notational convenience. The Weyl formula makes manifest the Weyl symmetry of the characters:

$$\mathrm{ch}_{\lambda}(\zeta) = (\det w)\mathrm{ch}_{w.\lambda}(\zeta) = \mathrm{ch}_{\lambda}(w.\zeta), \quad \forall w \in W(\zeta \in P_{\mathbf{R}}(X_r)), \tag{14}$$

where $w.\lambda := w(\lambda+\rho)-\rho$ denotes the shifted action of the Weyl group element w, and $P_{\mathbf{R}}(X_r) := \{\lambda = \Sigma_{i=1}^{r}\lambda_i\omega^i | \lambda_i \in \mathbf{R}\}$. When $\zeta \in P(X_r)$, this symmetry can be extended to include the elements of the Weyl group of $X_{r,k}$, the so-called affine Weyl group $\hat{W}$:

$$\mathrm{ch}_{\lambda}(\zeta) = (\det w)\mathrm{ch}_{w.\lambda}(\zeta) = \mathrm{ch}_{\lambda}(w.\zeta), \quad \forall w \in \hat{W} \quad (\zeta \in P(X_r)). \tag{15}$$

Let $S(X_r)$ signify the set of simple roots of X_r. W is generated by the primitive reflections r_{α},

$$r_{\alpha}\lambda := \lambda - (\lambda \cdot \alpha^{\vee})\alpha, \quad \alpha \in S(X_r), \tag{16}$$

where $\alpha^{\vee}$ is the simple coroot dual to the simple root α. $\hat{W}$ has one additional generator r_{α_0} corresponding to the zeroth affine simple root $\alpha_0 = \delta - \theta$.[4] Here θ is the highest root of X_r, and δ is the imaginary root of the affine algebra that yields $\delta \cdot \lambda = k$ for any level k weight λ (and $\delta \cdot \beta = 0$ for any root β). The action of r_{α_0} in (15) is therefore

$$r_{\alpha_0}.\lambda = r_{\theta}.\lambda + (k+h^{\vee})(\theta - \delta). \tag{17}$$

Since ζ corresponds to a level k weight, and ρ to a level $h^{\vee}$ one, we have $\zeta^{\delta} = 1$. Then the identity,

$$\zeta^{r_{\alpha_0}.\lambda} = \zeta^{r_{\theta}.\lambda + (k+h^{\vee})\theta} = \zeta^{r_{\theta}.\lambda}, \tag{18}$$

leads to the full affine Weyl invariance of (15).

Comparing Eqs. (3) and (10), taking into account the Kac–Peterson relation[3] and the affine Weyl symmetry (15), one finds[5,4,6,7]

$$N_{\lambda,\mu}^{\nu} = \sum_{w \in \hat{W}} (\det w) T_{\lambda,\mu}^{w.\nu}. \tag{19}$$

Using the well-known connection between multiplicities and tensor product coefficients,

$$T_{\lambda,\mu}^{\nu} = \sum_{w \in W} (\det w)\mathrm{mult}_{\mu}(w.\nu - \lambda), \tag{20}$$

we can also write

$$N_{\lambda,\mu}^{\nu} = \sum_{w \in \hat{W}} (\det w)\mathrm{mult}_{\mu}(w.\nu - \lambda). \tag{21}$$

These formulas make it clear that the WZW fusion coefficients are integers, while it is not evident from the form of the matrix S in (3). They do not, however, confirm that the fusion coefficients are non-negative integers, as must be.

That the tensor product coefficients are non-negative is made manifest by rules for their computation that reduce to *counting* certain objects, usually Young tableaux (see Ref. 8, for

example). The most famous is the Littlewood–Richardson rule that computes the tensor product coefficients of the simple Lie algebra A_r. Such rules do not involve overcounting and cancellations, and will be called combinatorial rules.

Algorithms like the Littlewood–Richardson rule involve Young tableaux. The action of the affine Weyl group on the Young tableaux for classical simple Lie algebras can be simply described. So, (19) allows the combinatorial methods for computing tensor products to be adapted to the computation of WZW fusion rules.[9] The adaptation is not combinatorial, however; the factor $\det w = \pm 1$ in (19) means the adapted methods involve overcounting and cancellations.

Can one find a purely combinatorial method for the computation of WZW fusion rules? To the best of our knowledge, no such rule is known. (Of course, for sufficiently large level k, when the fusion coefficient coincides with the corresponding tensor product coefficient, we can use the combinatorial rules for tensor products.) Such a rule, if it exists, should find wide application, much as the original Littlewood–Richardson rule has. It could, for example, settle the conjecture of a *threshold level* for WZW fusions.[10,11] More importantly, however, it would provide a constructive mathematical proof that fusion coefficients are non-negative integers, something that is obvious and necessary from a physical point of view. It may also deepen our understanding of WZW fusions; perhaps it will provide a definition of WZW fusion expressed solely in terms of standard (integrable, highest weight, positive integer level) representations, analogous to that for simple Lie algebra tensor products.

We may try to use (19) as a guide to a combinatorial rule. The problem with (19), however, is the factor $\det w = \pm 1$, inherited from the Weyl character formula (11). So, perhaps a non-negative character formula can lead to a combinatorial rule. The Demazure character formula[12,13] is such a non-negative formula, and Littelmann[14] has derived the Littlewood–Richardson rule and generalizations from it. In this work, we adapt his methods to the computation of WZW fusion rules, in an attempt to find a purely combinatorial rule. Although our results are only valid for $X_r = A_r$, we use universal language appropriate to all simple Lie algebras (and their untwisted affine analogs) whenever possible. This reflects our hope that a universal result, valid for all $X_{r,k}$, will eventually be found.

As far as we know, this work is the first direct application of Demazure character formulas to WZW fusions. In light of Littelmann's results, this is a natural attempt at finding a combinatorial rule for WZW fusions. Unfortunately, no such rule is obtained. Our effort does make it clear why, however. We hope it will therefore help as a stepping stone to a combinatorial rule for WZW fusions.

II. DEMAZURE CHARACTER FORMULAS

Introduce the primitive Demazure operators D_α:

$$D_\alpha(e^\mu) := \frac{e^\mu - e^{r_\alpha \cdot \mu}}{1 - e^{-\alpha}}, \tag{22}$$

for all $\alpha \in S(X_r)$. These linear operators are associated with the primitive reflections of the Weyl group W of X_r, and are defined for all $\mu \in P(X_r)$. More explicitly,

$$D_\alpha(e^\mu) = \begin{cases} e^\mu + \cdots + e^{\mu - (\mu\cdot\alpha^\vee)\alpha}, & \mu\cdot\alpha^\vee \geqslant 0; \\ 0, \quad \mu\cdot\alpha^\vee = -1; \\ -e^{\mu+\alpha} - \cdots - e^{\mu + (-\mu\cdot\alpha^\vee - 1)\alpha}, & \mu\cdot\alpha^\vee \leqslant -2. \end{cases} \tag{23}$$

A Demazure operator can also be associated to every element of the Weyl group of X_r. Suppose $w = r_{\beta_1} r_{\beta_2} \cdots r_{\beta_l}$, with all $\beta_i \in S(X_r)$, is a decomposition of $w \in W$. Such a decomposition is said to have length l. For fixed $w \in W$, any decomposition of minimum length is called a reduced

decomposition. The length $l(w)$ of any reduced decomposition of $w \in W$ is known as the length of w. Now, if $w = r_{\beta_1} r_{\beta_2} \cdots r_{\beta_l}$ is a reduced decomposition of $w \in W$, then we define

$$D_w := D_{\beta_1} \circ D_{\beta_2} \circ \cdots \circ D_{\beta_l}. \tag{24}$$

It makes sense to label these operators with the Weyl group element w, since they do not depend on which reduced decomposition is used. It is important, however, that a *reduced* decomposition be used in the definition of D_w; we have $r_\alpha^2 = \mathrm{id}$, but

$$D_\alpha(D_\alpha(e^\mu)) = D_\alpha(e^\mu). \tag{25}$$

This last relation follows from the more general one:

$$D_\alpha(e^\lambda D_\alpha(e^\mu)) = D_\alpha(e^\lambda) D_\alpha(e^\mu). \tag{26}$$

The Weyl group of any simple Lie algebra contains a unique element of maximum length. If w_L symbolizes the longest element of the Weyl group W, we have the following Demazure character formula:

$$\mathrm{ch}_\lambda = D_{w_L}(e^\lambda). \tag{27}$$

It is also useful to introduce other primitive Demazure operators related to the D_α:

$$d_\alpha(e^\mu) := (D_\alpha - id)(e^\mu) = \frac{e^{\mu-\alpha} - e^{r_\alpha \cdot \mu}}{1 - e^{-\alpha}}, \tag{28}$$

for all $\alpha \in S(X_r)$. More explicitly,

$$d_\alpha(e^\mu) = \begin{cases} e^{\mu-\alpha} + \cdots + e^{\mu-(\mu\cdot\alpha^\vee)\alpha}, & \mu \cdot \alpha^\vee \geqslant 1; \\ 0, \quad \mu \cdot \alpha^\vee = 0; \\ -e^\mu - \cdots - e^{\mu-(\mu\cdot\alpha^\vee+1)\alpha}, & \mu \cdot \alpha^\vee \leqslant -1. \end{cases} \tag{29}$$

These operators can also be generalized to d_w, for each $w \in W \backslash id$, in a way similar to that defining the D_w. For later convenience, we also define $d_{id}(e^\lambda) := e^\lambda$.

To write other Demazure character formulas, we need the Bruhat partial order on the elements of the Weyl group W. Let r_α be the Weyl reflection across the hyperplane in weight space normal to the root α, i.e. $r_\alpha \lambda = \lambda - (\lambda \cdot \alpha^\vee)\alpha$. The length $l(w)$ of a Weyl group element w is defined as the number of primitive reflections in a reduced decomposition of the element. For $v_1, v_2 \in W$, we write $v_1 \leftarrow v_2$ if and only if both $v_1 = r_\alpha v_2$ for some positive root α, and $l(v_1) = l(v_2) + 1$. Then for $w, \widetilde{w} \in W$, the Bruhat order is defined by $w > \widetilde{w}$ if $w \leftarrow v_1 \leftarrow v_2 \leftarrow \cdots \leftarrow v_n \leftarrow \widetilde{w}$, for some $v_1, v_2, \ldots, v_n \in W$.

Another formula for the character of the representation $R(\lambda)$ is

$$\mathrm{ch}_\lambda = \sum_{v \in W} d_v(e^\lambda). \tag{30}$$

A generalization is

$$D_w(e^\lambda) = \sum_{v \leq w} d_v(e^\lambda). \tag{31}$$

The formula

$$D_\alpha(d_{r_\alpha v}(e^\mu))=d_{r_\alpha v}(e^\mu)+d_v(e^\mu), \quad \text{if } r_\alpha v\prec v, \quad \alpha\in S(X_r), \tag{32}$$

makes explicit the relation between the two types of Demazure operators. Using (28) this translates into

$$d_v(e^\mu)=d_\alpha\circ d_{r_\alpha v}(e^\mu), \quad \text{if } r_\alpha v\prec v, \quad \alpha\in S(X_r). \tag{33}$$

III. TABLEAUX AND THE WEYL GROUP

In order to relate the Demazure formulas to the Littlewood–Richardson rule, the Weyl group must be related to Young tableaux and standard tableaux. This is done by recognizing certain sequences of elements of W as the appropriate generalizations of standard tableaux. The so-called minimal defining chains (MDCs) of Weyl group elements are in one-to-one correspondence with certain vectors of fixed weight that form a basis of the irreducible representation $R(\lambda)$ of X_r.[15,16] For the algebra A_r, it is well known that such vectors are in one-to-one correspondence with standard tableaux. MDCs are what work for all simple Lie algebras X_r.

Suppose μ is the highest weight of an integrable representation $R(\mu)$ of X_r. Such a weight can be expressed as a sum of fundamental weights $\mu=\nu^1+\nu^2+\cdots+\nu^n$, where $\nu^j\in F(X_r)$, $1\leqslant j\leqslant n$. If we fix an order $>$ on the elements of $F(X_r)$, so that we can also impose $\nu^i\geqslant\nu^{i+1}$ on the sum, the expression is unique. Now consider a sequence $c=(w_1,w_2,\ldots,w_n)$ of n Weyl group elements $w_i\in W$. The sequence, or chain, is a defining chain if it respects the Bruhat order: $w_1\preceq w_2\preceq\cdots\preceq w_n$. A sequence of weights of fundamental representations corresponding to each defining chain and weight μ is given by

$$P_\mu(c):=(w_1\nu^1,w_2\nu^2,\ldots,w_n\nu^n). \tag{34}$$

Different defining chains, however, lead to the same weight sequence. A unique *minimal defining chain* (MDC) can be associated with the weight sequence: $c=(w_1,\ldots,w_n)$ is a minimal defining chain if for any other defining chain $\tilde{c}=(\tilde{w}_1,\ldots,\tilde{w}_n)$ satisfying $P_\mu(c)=P_\mu(\tilde{c})$, we have $w_j\preceq\tilde{w}_j$, for all $1\leqslant j\leqslant n$.

Define the μ-weight $p_\mu[c]$ of the MDC c as

$$p_\mu[c]=p_\mu[(w_1,\ldots,w_n)]:=w_1\nu^1+\cdots+w_n\nu^n. \tag{35}$$

Let $\mathscr{C}_\mu[\sigma]$ be the set of MDCs with μ-weight σ, and let $\mathscr{C}_\mu(w)$ denote the set of MDCs with the last element equal to $w\in W$. The weights of a representation $R(\mu)$ and those of MDCs are related:

$$\mathrm{mult}_\mu(\sigma)=|\mathscr{C}_\mu[\sigma]|, \tag{36}$$

so that

$$\mathrm{ch}_\mu=\sum_{\sigma\in P(\mu)}\sum_{c\in\mathscr{C}_\mu[\sigma]}e^{p_\mu[c]}. \tag{37}$$

Also relevant here is the connection with Demazure operators:

$$d_w(e^\lambda)=\sum_{c\in\mathscr{C}_\lambda(w)}e^{p_\lambda[c]}. \tag{38}$$

In the case $X_r=A_r$, the usual standard tableaux are recovered as follows. Let $\{e_i|i=1,\ldots,r+1;\ e_i\cdot e_j=\delta_{i,j}\}$ be an orthonormal basis of $\mathbf{R}^{r+1}$. The fundamental weights can be chosen to be $\omega^i=e_1+e_2+\cdots+e_i-i\psi/(r+1)$, with $\psi=\Sigma_{i=1}^{r+1}e_i$.[17] The Weyl group of A_r acts as the group of permutations on the e_i. Therefore the weights $w_j\nu^j$ $[w_j\in W,\ \nu^j\in F(A_r)]$ in the weight sequence

TABLE I. Minimal defining chains and corresponding standard tableaux for the A_2 representation of highest weight $2\omega^1+\omega^2=:(2,1)$.

MDC	weight sequence	tableau	weight
(id, id, id)	$((0,1),(1,0),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&1\\\hline 2\\\cline{1-1}\end{array}$	$(2,1)$
(id, id, r_1)	$((0,1),(1,0),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&2\\\hline 2\\\cline{1-1}\end{array}$	$(0,2)$
(id, r_1, r_1)	$((0,1),(-1,1),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&2&2\\\hline 2\\\cline{1-1}\end{array}$	$(-2,3)$
(r_2, r_2, r_2)	$((1,-1),(1,0),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&1\\\hline 3\\\cline{1-1}\end{array}$	$(3,-1)$
(id, id, r_2r_1)	$((0,1),(1,0),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&3\\\hline 2\\\cline{1-1}\end{array}$	$(1,0)$
(r_2, r_2, r_1r_2)	$((1,-1),(1,0),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&2\\\hline 3\\\cline{1-1}\end{array}$	$(1,0)$
(r_2, r_1r_2, r_1r_2)	$((1,-1),(-1,1),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&2&2\\\hline 3\\\cline{1-1}\end{array}$	$(-1,1)$
(id, r_1, r_2r_1)	$((0,1),(-1,1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&2&3\\\hline 2\\\cline{1-1}\end{array}$	$(-1,1)$
(r_1r_2, r_1r_2, r_1r_2)	$((-1,0),(-1,1),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&2\\\hline 3\\\cline{1-1}\end{array}$	$(-3,2)$
(r_2, r_2, r_2r_1)	$((1,-1),(1,0),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&1&3\\\hline 3\\\cline{1-1}\end{array}$	$(2,-2)$
$(r_2, r_1r_2, r_2r_1r_2)$	$((1,-1),(-1,1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&2&3\\\hline 3\\\cline{1-1}\end{array}$	$(0,-1)$
(id, r_2r_1, r_2r_1)	$((0,1),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&3&3\\\hline 2\\\cline{1-1}\end{array}$	$(0,-1)$
$(r_1r_2, r_1r_2, r_2r_1r_2)$	$((-1,0),(-1,1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&3\\\hline 3\\\cline{1-1}\end{array}$	$(-2,0)$
(r_2, r_2r_1, r_2r_1)	$((1,-1),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 1&3&3\\\hline 3\\\cline{1-1}\end{array}$	$(1,-3)$
$(r_1r_2, r_2r_1r_2, r_2r_1r_2)$	$((-1,0),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&3&3\\\hline 3\\\cline{1-1}\end{array}$	$(-1,-2)$

$P_\mu(c)=(w_1\nu^1, w_2\nu^2, \ldots, w_n\nu^n)$ of a MDC c can be associated with a column of boxes, with numbers from 1 to $r+1$ in each, increasing down the column. A box $\boxed{2}$, say, corresponds to a summand e_2 in the expression for the weight, modulo multiples of ψ. If the fundamental weights of A_r are ordered $\omega^1<\omega^2<\cdots<\omega^r$, then if the columns of boxes corresponding to the weights in the weight sequence of a MDC are assembled into a tableaux, the columns will have heights that do not increase from left to right. Furthermore, the Bruhat order imposed on a MDC ensures that the numbers in the rows of the tableau also do not increase from left to right. But these are precisely the defining properties of a standard tableau for A_r. (These are sometimes called semi-standard tableaux, when standard tableaux designate those with numbers increasing in both rows and columns.) [The further *minimal* property of the MDCs is necessary for (36) and (38).]

The MDCs, corresponding weight sequences, and standard tableaux associated with vectors in the A_2 representation $R(2,1)$ of highest weight $2\omega^1+\omega^2=:(2,1)$ are given as an example in Table I. There we use the shorthand notation $r_{\alpha_i}=:r_i$.

If a standard tableau corresponds to a weight sequence $P_\mu(c)$ of the MDC c, then we say that it has shape μ. The shape of a standard tableau is determined by the configuration of its boxes, and not by the numbers therein. If the numbers are removed from a standard tableau of shape μ, a Young tableau of shape μ results. Young tableaux also enter into the Littlewood–Richardson rule.

It will be useful to define subchains obtained from other chains by keeping the last i elements on the right, and also by keeping the first $n-i$ elements on the left. If $c=(w_1, w_2, \ldots, w_n)$ is a defining chain, we define

$$R_i(c):=(w_{n-i+1}, w_{n-i+2}, \ldots, w_n) \quad \text{and} \quad L_i(c):=(w_1, w_2, \ldots, w_{n-i}). \tag{39}$$

If $\mu=\nu^1+\nu^2+\cdots+\nu^n\in P_{\geqslant}(X_{\tau r})$, we also define

$$R_i(\mu):=\nu^{n-i+1}+\nu^{n-i+2}+\cdots+\nu^n \quad \text{and} \quad L_i(\mu):=\nu^1+\nu^2+\cdots+\nu^{n-i}. \tag{40}$$

IV. LITTELMANN'S GENERALIZATION OF THE CLASSICAL LITTLEWOOD–RICHARDSON RULE

From (10) and the Weyl character formula (11), we have

$$\sum_{w\in W}(\det w)e^{w(\lambda+\rho)}\ \mathrm{ch}_\mu=\sum_{\nu\in P_{\geqslant}(X_r)}T^{\nu}_{\lambda,\mu}\sum_{w\in W}(\det w)e^{w(\nu+\rho)}. \tag{41}$$

Defining the linear operator Θ by

$$\Theta(e^\lambda):=\sum_{w\in W}(\det w)e^{w(\lambda+\rho)}, \tag{42}$$

this can be rewritten as

$$\Theta(e^\lambda\ \mathrm{ch}_\mu)=\Theta\left(\sum_{\nu\in P_{\geqslant}(X_r)}T^{\nu}_{\lambda,\mu}e^\nu\right). \tag{43}$$

What Littelmann showed was that the left-hand side simplifies because it contains terms of the form

$$\Theta(e^\kappa d_v(e^\gamma))=0, \quad \text{for } r_\alpha v\prec v\in W, \quad \kappa\cdot\alpha^\vee=0, \quad \alpha\in S(X_r). \tag{44}$$

To see why these contributions vanish, first note that by the definition (42), we get

$$\Theta(e^{w\cdot\gamma})=(\det w)\Theta(e^\gamma), \quad \forall w\in W, \tag{45}$$

so that

$$\Theta(D_\alpha(e^\gamma))=\Theta(e^\gamma), \quad \forall\alpha\in S(X_r). \tag{46}$$

Equation (26) above then gives

$$\Theta(e^\kappa D_\alpha(e^\gamma))=\Theta(D_\alpha(e^\kappa D_\alpha(e^\gamma)))=\Theta(D_\alpha(e^\kappa)D_\alpha(e^\gamma))=\Theta(D_\alpha(e^\kappa)e^\gamma), \tag{47}$$

for any simple root $\alpha\in S(X_r)$. This implies

$$\Theta(e^\kappa d_\alpha(e^\gamma))=\Theta(d_\alpha(e^\kappa)e^\gamma), \tag{48}$$

by the definition (28) of the operators d_α. From (33) we then get

$$\Theta(e^\kappa d_v(e^\gamma))=\Theta(d_\alpha(e^\kappa)d_{r_\alpha v}(e^\gamma)), \quad \text{for } r_\alpha v\prec v. \tag{49}$$

But since $d_\alpha(e^\kappa)=0$ for $\kappa\cdot\alpha^\vee=0$, (44) is proved.

The vanishing contributions to the tensor product $R(\lambda)\otimes R(\mu)$ encoded in (44) can be identified easily using the MDCs. From (37), the left-hand side of (43) contains expressions of the form

$$\Theta(e^{\lambda+p_\mu[c]})=\Theta(e^{[\lambda+p_{R_i(\mu)}[R_i(c)]]+p_{L_i(\mu)}[L_i(c)]}), \tag{50}$$

where $1\leqslant i\leqslant n$. Suppose that for the particular MDC under consideration,

$$\lambda + p_{R_i(\mu)}[R_i(c)] \in P_{\geqslant}(X_r), \quad \text{for } i = 1,...,l < n;$$

$$\lambda + p_{R_{l+1}[\mu]}(R_{l+1}(c)) \notin P_{\geqslant}(X_r). \tag{51}$$

For $X_r = A_r$, we have $w\nu \cdot \alpha^\vee \in \{-1,0,1\}$, for all $\alpha \in S(X_r)$ and $\nu \in F(X_r)$. With $c = (w_1,\ldots,w_n)$ and $P_\mu(c) = (w_1\nu^1,\ldots,w_n\nu^n)$, (51) implies that $(\lambda + p_{R_l(\mu)}[R_l(c)]) \cdot \alpha^\vee = 0$ and $(\lambda + p_{R_{l+1}(\mu)} \times [R_{l+1}(c)]) \cdot \alpha^\vee = -1$, for some simple root $\alpha \in S(X_r)$. This means that $(w_{n-l}\nu^{n-l}) \cdot \alpha^\vee = -1$ and that $r_\alpha w_{n-l} \prec w_{n-l}$.

By (38) then, the expression in (50) is contained in

$$\Theta(e^{\lambda + p_{R_l(\mu)}[R_l(c)]} d_{w_{n-l}}(e^{\mu - R_l(\mu)})), \tag{52}$$

which vanishes by (44). The Bruhat order obeyed by the elements of minimal defining chains, and their connection with Demazure characters, ensures that all terms in (52) of the form (50) obeying (51) are present on the left-hand side of (43), if one is.

A minimal defining chain c is called λ dominant if

$$\lambda + p_{R_i(\mu)}[R_i(c)] \in P_{\geqslant}(X_r), \quad \text{for all } 0 \leqslant i \leqslant n. \tag{53}$$

So, only the λ-dominant MDCs can contribute to the left-hand side of (43). Define $\mathscr{C}_{\lambda,\mu}[\sigma]$ to be the set of λ-dominant MDCs of μ weight σ. Then, once the terms of the form (52) are eliminated from the left-hand side of (43), we are left with

$$\Theta\left(\sum_{\nu \in P_{\geqslant}(X_r)} \dim\, \mathscr{C}_{\lambda,\mu}[\nu - \lambda] e^\nu \right) = \Theta\left(\sum_{\nu \in P_{\geqslant}(X_r)} T^\nu_{\lambda,\mu} e^\nu \right). \tag{54}$$

But, it is easy to see that

$$\Theta\left(\sum_{\mu \in P_{\geqslant}(X_r)} a_\mu e^\mu \right) = \Theta\left(\sum_{\nu \in P_{\geqslant}(X_r)} b_\nu e^\nu \right) \Rightarrow a_\mu = b_\mu, \quad \forall \mu \in P_{\geqslant}(X_r). \tag{55}$$

Littelmann's generalization of the Littlewood–Richardson rule then follows:

$$\boxed{T^\nu_{\lambda,\mu} = |\mathcal{C}_{\lambda,\mu}[\nu - \lambda]|} \ . \tag{56}$$

In the case $X_r = A_r$, the classical Littlewood–Richardson rule is recovered. Let a standard tableau be called λ dominant if the corresponding MDC is λ dominant. To find the tensor product coefficient $T^\nu_{\lambda,\mu}$, one simply counts the number of λ-dominant standard tableaux of shape μ and weight $\nu - \lambda$. These can be found by taking each standard tableau of shape μ and weight $\nu - \lambda$ and adding the weights of its columns, from right to left, to λ. If the addition of the weight of one of the columns results in a nondominant weight, no contribution results. If this does not happen, the standard tableau contributes 1 to $T^\nu_{\lambda,\mu}$.

The rule is most easily implemented by adding the columns directly to the Young tableau of shape λ, to form a mixed tableau. The shape of the mixed tableau is similar to the shape of a Young tableau. One takes the columns of the standard tableau, and any box $\boxed{3}$, e.g., to the third row of the Young tableau, etc. If the addition in this manner of any column results in a mixed tableau of nondominant shape, the standard tableau does not contribute.

For example, consider the decomposition of the tensor product of A_2 representations: $R(2,0)\otimes R(2,1)=R(0,3)\oplus R(1,1)\oplus R(2,2)\oplus R(3,0)\oplus R(4,1)$, where $R(\lambda_1,\lambda_2)$ means $R(\lambda)$ with $\lambda=\lambda_1\omega^1+\lambda_2\omega^2$. We have $T^{(3,0)}_{(2,0),(2,1)}=1$ while there are two standard tableaux of shape (2,1) and of the appropriate weight $(3,0)-(2,0)=(1,0)$:

$$\begin{array}{|c|c|c|}\hline 1&1&2\\\hline 3\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&1&3\\\hline 2\\\cline{1-1}\end{array}\ . \tag{57}$$

Only the first standard tableau contributes because by adding its columns, from right to left, to the Young tableau $\square\square$ of shape (2,0), we obtain the following sequence of mixed tableaux:

$$\begin{array}{|c|c|}\hline \ &\ \\\hline 2\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline \ &\ &1\\\hline 2\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|c|}\hline \ &\ &1&1\\\hline 2\\\cline{1-1}3\\\cline{1-1}\end{array}\ , \tag{58}$$

each of which has a dominant shape. In attempting to do the same for the second standard tableau, we immediately encounter

$$\begin{array}{|c|c|}\hline \ &\ \\\hline 3\\\cline{1-1}\end{array}\ . \tag{59}$$

This mixed tableau is clearly not dominant, and so does not contribute to the tensor product.

Only five tableaux from Table I contribute to the tensor product $R(2,0)\otimes R(2,1)$. The rest must result in terms of the form (44). For example, the five tableaux,

$$\begin{array}{|c|c|c|}\hline 1&1&3\\\hline 2\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&2&3\\\hline 2\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&1&3\\\hline 3\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&3&3\\\hline 2\\\cline{1-1}\end{array}\ ,\quad \begin{array}{|c|c|c|}\hline 1&3&3\\\hline 3\\\cline{1-1}\end{array}\ , \tag{60}$$

furnish together such a form, with $\kappa=(2,0)$, $\gamma=(2,1)$, and $v=r_2r_1$.

The form of the Littlewood–Richardson rule just described differs from the original, in which the only numbered tableau that enters is

$$\begin{array}{|c|c|c|}\hline 1&1&1\\\hline 2\\\cline{1-1}\end{array}\ . \tag{61}$$

But the translation is simple, and is described explicitly in Ref. 18. The version involving standard tableaux is the most convenient, since it can be directly related to vectors in representations of X_r.

V. MODIFIED LITTLEWOOD–RICHARDSON RULES FOR FUSIONS

In this section we will follow Littelmann's derivation of a generalized Littlewood–Richardson rule,[14] as outlined in the previous section, and attempt to find a combinatorial rule for WZW fusions.

From (3) and (7), we can follow steps in the previous section to arrive at

$$\Theta(e^{\lambda}\ \mathrm{ch}_{\mu})(\sigma)=\Theta\left(\sum_{\nu\in P_{\geqslant}(X_{r,k})} N^{\nu}_{\lambda,\mu}e^{\nu}\right)(\sigma), \tag{62}$$

where λ, μ, σ are elements of $P_{\geqslant}(X_{r,k})$.

From (19) it is clear that the Weyl group $\hat{W}$ of $X_{r,k}$ must be implicated. Using the action (17) of the $\hat{W}$ generator r_{α_0} that is additional to the generators of W, it is simple to show that

$$\Theta(e^{w\cdot\gamma})(\sigma)=(\det w)\Theta(e^{\gamma})(\sigma), \quad \forall w\in\hat{W}, \tag{63}$$

when σ, $\gamma\in P(X_r)$. But another ingredient in Littelmann's derivation of a generalized Littlewood–Richardson rule is the Demazure character formulas for $\mathrm{ch}_\mu(\sigma)$ that involve the Bruhat order on W. We know of no way to write a similar formula for $\mathrm{ch}_\mu(\sigma)$ involving the Bruhat order on the full affine Weyl group $\hat{W}$. Formulas involving the Bruhat order on subgroups of $\hat{W}$ *can* be written, however, as we now demonstrate.

In order to use the affine Weyl symmetry (63), the weights involved must have level k: $\delta\cdot\gamma=k$ is required. If two such weights are added, however, we get level $2k$. Let ω^0 be the zeroth affine fundamental weight, so that $\delta\cdot\omega^0=1$. If ξ, ϕ have level k, i.e. $\delta\cdot\phi=\delta\cdot\xi=k$, define $\bar{\phi}:=\phi-k\omega^0$. Then $\delta\cdot(\xi+\bar{\phi})=k$. Notice that $\sigma^{\omega^0}=1$, so that we are able to interpret the weight μ on the left-hand side of (62) above as $\bar{\mu}$, in order to exploit the affine symmetry (63).

With $w=r_{\alpha_0}$ in (63), it is easy to show that

$$\Theta(d_{\alpha_0}(e^{\gamma}))(\sigma)=0, \tag{64}$$

for $\delta\cdot\gamma=k$, with $\gamma\in P(X_r)$, if d_{α_0} is defined by (28) with $\alpha=\alpha_0$. From this,

$$\Theta(e^{\gamma}d_{\alpha_0}(e^{\bar{\phi}}))(\sigma)=0, \quad \text{if } \gamma\cdot\alpha_0=k-\theta\cdot\gamma=0 \tag{65}$$

follows, since $\gamma+r_{\alpha_0}\cdot\bar{\phi}=r_{\alpha_0}\cdot(\gamma+\bar{\phi})$ in that case. Finally then, if we define Demazure operators d_u for any $u\in\hat{W}$ by the reduced decomposition of u, we find

$$\Theta(e^{\gamma}d_u(e^{\bar{\phi}}))(\sigma)=0, \quad \text{if } k-\theta\cdot\gamma=0 \quad \text{and} \quad r_{\alpha_0}u\prec u, \tag{66}$$

where now $\prec$ indicates the Bruhat order on $\hat{W}$.

For $X_r=A_r$, let W^m denote the subgroup of $\hat{W}$ generated by all the affine primitive reflections except r_{α_m}:

$$W^m=\langle r_{\alpha_i} \mid i\in\{0,1,2,\ldots,r\}\setminus m\rangle. \tag{67}$$

So all $W^m\cong W$, and $W^0=W$. Denote by $x\in W$ the Weyl group element that permutes the simple roots of A_r and $-\theta$ in a cyclic manner:

$$x(-\theta,\alpha_1,\alpha_2,\ldots,\alpha_r)=(\alpha_1,\alpha_2,\ldots,\alpha_r,-\theta). \tag{68}$$

Using these definitions we can write a modified Demazure character formula,

$$\boxed{\mathrm{ch}_{\bar{\mu}}(\sigma) \;=\; \sum_{v\in W^m} d_v(e^{x^m\bar{\mu}})(\sigma)} \quad , \tag{69}$$

involving the subgroup W^m of the affine Weyl group $\hat{W}$.

Defining chains and corresponding tableaux can also be associated to the affine subgroups W^m. The highest weight μ of an integrable representation $R(\mu)$ can be expressed uniquely as a sum $\mu = x^m\nu^1 + x^m\nu^2 + \cdots + x^m\nu^n$, where $\nu^j \in F(A_r)$, and the order $\nu^i \geqslant \nu^{i+1}$ has been fixed. A sequence $c=(w_1,w_2,\ldots,w_n)$ of elements of W^m can be associated with such a weight. If $w_j \preceq w_{j+1}$, the chain is a defining chain. If it is also the minimal defining chain corresponding to the following sequence of weights:

$$P^m_\mu(c) := (w_1x^m\nu^1, w_2x^m\nu^2, \ldots, w_nx^m\nu^n), \tag{70}$$

we will call it an m-MDC.

Define the μ weight $p_{\mu,m}[c]$ of the m-MDC c as

$$p_{\mu,m}[c] = p_{\mu,m}[(w_1,w_2,\ldots,w_n)] := w_1x^m\nu^1 + \cdots + w_nx^m\nu^n. \tag{71}$$

Let $\mathscr{C}^m_\mu[\phi]$ be the set of m-MDCs with μ weight ϕ, and let $\mathscr{C}^m_\mu(w)$ denote the set of m-MDCs with the last element equal to $w \in W^m$. The weights of a representation $R(\mu)$ and those of m-MDCs are related:

$$\mathrm{mult}_\mu(\phi) = |\mathscr{C}^m_\mu[\phi]|, \tag{72}$$

so that

$$\mathrm{ch}_{\overline{\mu}}(\sigma) = \sum_{\phi \in P(\mu)} \sum_{c \in \mathscr{C}^m_\mu[\phi]} e^{p_{\mu,m}[c]}(\sigma). \tag{73}$$

The relevant connection with Demazure operators is given by

$$d_w(e^\lambda)(\sigma) = \sum_{c \in \mathscr{C}^m_\lambda(w)} e^{p_{\lambda,m}[c]}. \tag{74}$$

The previous two equations are valid for any $\sigma \in P(A_r)$, and the previous three are consequences of the Weyl invariance of characters.

A minor variation of the usual standard tableaux associated to $X_r = A_r$ can encode the m-MDCs. In terms of the orthonormal basis $\{e_i\}$ of $\mathbf{R}^{r+1}$ related to the usual standard tableaux, the action of $x \in W$ is very simple:

$$x(e_1,e_2,\ldots,e_{r+1}) = (e_2,e_3,\ldots,e_{r+1},e_1). \tag{75}$$

Using the same ordering of fundamental weights: $\omega^r > \omega^{r-1} > \cdots > \omega^1$, we can define m-standard tableaux as the numbered tableaux of shape μ with numbers appearing in the order

$$(m+1, m+2, \ldots, r+1, 1, 2, \ldots, m), \tag{76}$$

from left to right in its rows, and from top to bottom with no repetitions in its columns. [Other tableaux, with orderings that are noncyclic permutations of (76), can also be defined. Their relation to fusion coefficients, however, is not clear.] Notice that 0-standard tableaux are the usual standard tableaux.

As an example, consider again the A_2 representation of highest weight $2\omega^1 + \omega^2 = (2,1)$. For A_2, $x = r_{\alpha_1}r_{\alpha_2}$, and the corresponding 1-MDCs, and 1-standard tableaux are shown in Table II.

The derivation of a modified Littlewood–Richardson rule from (62), (66), (73), and (74) follows the derivation of the classical Littlewood–Richardson rule outlined in the previous section. To state the result, we define c, an m-MDC of length n, to be (λ,k) dominant if

TABLE II. Minimal defining chains and corresponding 1-standard tableaux for the A_2 representation of highest weight $2\omega^1+\omega^2=:(2,1)$, using the affine Weyl subgroup $W^1=\langle r_0, r_2\rangle$.

1-MDC	weight sequence	tableau	weight
(id, id, id)	$((-1,0),(-1,1),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&2\\\hline 3\\\cline{1-1}\end{array}$	$(-3,2)$
(r_0, r_0, r_0)	$((0,1),(-1,1),(-1,1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&2\\\hline 1\\\cline{1-1}\end{array}$	$(-2,3)$
(id, id, r_2)	$((-1,0),(-1,1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&3\\\hline 3\\\cline{1-1}\end{array}$	$(-2,0)$
(id, r_2, r_2)	$((-1,0),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&3&3\\\hline 3\\\cline{1-1}\end{array}$	$(-1,-2)$
(r_0, r_0, r_2r_0)	$((0,1),(-1,1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&3\\\hline 1\\\cline{1-1}\end{array}$	$(-1,1)$
(id, id, r_0r_2)	$((-1,0),(-1,1),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&1\\\hline 3\\\cline{1-1}\end{array}$	$(-1,1)$
(id, r_2, r_0r_2)	$((-1,0),(0,-1),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&3&1\\\hline 3\\\cline{1-1}\end{array}$	$(0,-1)$
(r_0, r_2r_0, r_2r_0)	$((0,1),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 2&3&3\\\hline 1\\\cline{1-1}\end{array}$	$(0,-1)$
(r_2r_0, r_2r_0, r_2r_0)	$((1,-1),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 3&3&3\\\hline 1\\\cline{1-1}\end{array}$	$(1,-3)$
(r_0, r_0, r_0r_2)	$((0,1),(-1,1),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&2&1\\\hline 1\\\cline{1-1}\end{array}$	$(0,2)$
$(r_0, r_2r_0, r_0r_2r_0)$	$((0,1),(0,-1),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&3&1\\\hline 1\\\cline{1-1}\end{array}$	$(1,0)$
(id, r_0r_2, r_0r_2)	$((-1,0),(1,0),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&1&1\\\hline 3\\\cline{1-1}\end{array}$	$(1,0)$
$(r_2r_0, r_2r_0, r_0r_2r_0)$	$((1,-1),(0,-1),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 3&3&1\\\hline 1\\\cline{1-1}\end{array}$	$(2,-2)$
(r_0, r_0r_2, r_0r_2)	$((0,1),(1,0),(1,0))$	$\begin{array}{\|c\|c\|c\|}\hline 2&1&1\\\hline 1\\\cline{1-1}\end{array}$	$(2,1)$
$(r_2r_0, r_0r_2r_0, r_0r_2r_0)$	$((1,-1),(0,-1),(0,-1))$	$\begin{array}{\|c\|c\|c\|}\hline 3&1&1\\\hline 1\\\cline{1-1}\end{array}$	$(3,-1)$

$$\lambda + p_{R_i(\mu)}[R_i(c)] \in P_{\geqslant}(X_{r,k}), \quad \text{for all } 0 \leqslant i \leqslant n. \tag{77}$$

Let $\mathscr{C}^{m,k}_{\lambda,\mu}[\sigma]$ denote the set of (λ,k)-dominant m-MDCs of μ weight σ. The modified Littlewood–Richardson rule for $A_{r,k}$ fusions is then

$$\boxed{N^{\nu}_{\lambda,\mu} \;\leq\; |\mathcal{C}^{m,k}_{\lambda,\mu}[\nu-\lambda]| \;\; \forall\; 0 \leq m \leq r\ .} \tag{78}$$

Let an m-standard tableau be called (λ,k) dominant if the corresponding m-MDC is (λ,k) dominant. To treat the fusion coefficient $N^{\nu}_{\lambda,\mu}$ for $A_{r,k}$, one simply counts the number of (λ,k)-dominant m-standard tableaux of shape μ and weight $\nu-\lambda$. These can be found by taking each m-standard tableau of shape μ and weight $\nu-\lambda$ and adding the weights of its columns, from right to left, to λ. If the addition of the weight of one of the columns results in a nondominant weight $\notin P_{\geqslant}(A_{r,k})$, no contribution results. If this does not happen, the m-standard tableau contributes 1 to the upper bound on $N^{\nu}_{\lambda,\mu}$.

The modified rule is easily implemented by adding the columns directly to the Young tableau of shape λ, to form a mixed tableau. If the addition in this manner of any column results in a mixed tableau of shape $\notin P_{\geqslant}(A_{r,k})$, the m-standard tableau does not contribute. As an example, consider the $A_{2,k=3}$ fusion rule: $R(2,0)\otimes_3 R(2,1)=R(1,1)\oplus R(0,3)$. Choosing $m=1$ as in Table

II, we can explain $N^{(3,0)}_{(2,0),(2,1)}=0$ (for $k=3$). From Table II we read that the two 1-standard tableaux of shape (2,1) and of the appropriate weight $(3,0)-(2,0)=(1,0)$ are

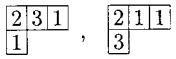

(79)

Neither of these tableaux contribute. The first does not, because by adding its columns, from right to left, to the Young tableau $\square\square$ of shape (2,0), we obtain the following sequence of mixed tableaux:

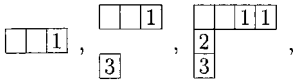

(80)

the second of which has shape $\notin P_{\geqslant}(A_{2,3})$. For the second standard tableau, we find

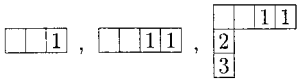

(81)

Again, the second mixed tableau has a shape $\notin P_{\geqslant}(A_{2,3})$ this time because it has a number of columns (of less than three boxes) greater than $k=3$. In a similar fashion, the complete fusion rule is found correctly, at level 3 and all higher levels.

If we compare the noncombinatorial expressions for tensor product coefficients (21), and for fusion coefficients (20), it is clear that the modified Littlewood–Richardson rule (78) calculates the right-hand side of

$$N^{\nu}_{\lambda,\mu} \leqslant T^{(m)\nu}_{\lambda,\mu} := \sum_{w\in W^m} (\det w)\,\mathrm{mult}_{\mu}(w.\nu-\lambda). \tag{82}$$

We conjecture that

$$\boxed{N^{\nu}_{\lambda,\mu} \;=\; T^{(m)\nu}_{\lambda,\mu} \quad \text{when} \quad \mathrm{mult}_{\mu}(r_{\alpha_m}.\nu-\lambda)=0} \quad . \tag{83}$$

It is not hard to find an example when the bound (78) is not saturated. Consider the $A_{4,4}$ fusion $R(1,1,1,1)\otimes_4 R(1,1,1,1)$, with decomposition containing four copies of $R(1,1,1,1)$, so that $N^{(1,1,1,1)}_{(1,1,1,1),(1,1,1,1)}=4$. The modified Littlewood–Richardson rule gives $T^{(m)(1,1,1,1)}_{(1,1,1,1),(1,1,1,1)}=16, 6, 5, 5, 6,$ for $m=0,1,2,3,4$, respectively.

The $A_{4,4}$ fusion coefficient $N^{(1,1,1,1)}_{(1,1,1,1),(1,1,1,1)}=4$ *can* be recovered from the Littlewood–Richardson rule, however, if level-rank duality[19] is used. But this fix is not general. If the level is increased by one, it fails: the $A_{4,5}$ fusion coefficient $N^{(1,1,1,1)}_{(1,1,1,1),(1,1,1,1)}=14$ cannot be recovered using (78) and level-rank duality. The problem is deeper than that, as we discuss in the next section.

We now show that the modified Littlewood–Richardson rule (78) gives a known upper bound on WZW A_r fusion coefficients. Let J denote the diagram outer automorphism of $P_{\geqslant}(A_{r,k})$, with action

$$J\lambda = J\left(\sum_{i=1}^{r} \lambda_i \omega^i\right) = \left(k - \sum_{i=1}^{r} \lambda_i\right)\omega^1 + \sum_{i=2}^{r} \lambda_{i-1}\omega^i. \tag{84}$$

What we will show is that

$$T^{\nu}_{\lambda,\mu} = T^{(0)\nu}_{\lambda,\mu} = T^{(m)J^m\nu}_{J^m\lambda,\mu}. \tag{85}$$

From (20) we can write

$$T^{(m)J^m\nu}_{J^m\lambda,\mu} = \sum_{w\in W^m} (\det w)\,\mathrm{mult}_{\mu}(J^m((J^{-m}wJ^m)(\nu+\rho)-(\lambda+\rho))), \tag{86}$$

using $J^m\rho=\rho$. But for $w\in W^m$, we have $J^{-m}wJ^m\in W$. Furthermore, if σ is any weight satisfying $\sigma\cdot\delta=0$, then $J^m\sigma=x^m\sigma$. The Weyl invariance of the multiplicities then establishes (85).

The bound computed by (78) is just (82), or

$$N^{\nu}_{\lambda,\mu} \leqslant T^{J^{-m}\nu}_{J^{-m}\lambda,\mu}. \tag{87}$$

But it is well known that

$$N^{\nu}_{\lambda,\mu} = N^{J^{-m-n}\nu}_{J^{-m}\lambda,J^{-n}\mu}, \tag{88}$$

for all m,n. And with $N^{\nu}_{\lambda,\mu}\leqslant T^{\nu}_{\lambda,\mu}$, we obtain

$$N^{\nu}_{\lambda,\mu} \leqslant T^{J^{-m-n}\nu}_{J^{-m}\lambda,J^{-n}\mu}. \tag{89}$$

This last result is even stronger than the upper bound (82). It was originally conjectured to be saturated for all triples $\lambda,\mu,\nu\in P_{\geqslant}(A_{r,k})$, for some choice of m and n.[6]

VI. DISCUSSION

The classical Littlewood–Richardson rule for A_r tensor products, and generalizations for other simple Lie algebras [see (56)], can be derived from Demazure character formulas, as Littelmann demonstrated.[14] Here we wrote Demazure character formulas (69) for ratios of elements of the modular S matrices of affine Kac–Moody algebras. These were then used in the Verlinde formula in an attempt to derive a combinatorial rule for WZW A_r fusion rules.

The resulting modified Littlewood–Richardson rule (78), however, only provides an upper bound on the fusion coefficients. The basic reason is the use of the subgroups W^m of the affine Weyl group in the Demazure character formula (69), instead of the full affine Weyl group and its Bruhat order. The cancellations in (20) can be ordered systematically using the Bruhat order on the finite Weyl group W, so that a combinatorial rule results from (30) and (38). The affine Bruhat order is needed, however, to order all the cancellations in (21) in a similar way. But we were unable to write a character formula for the relevant quantities (7) that involves the Bruhat order on the full affine Weyl group.

Perhaps such a character formula can be written. On the other hand, a different derivation of the Littlewood–Richardson rule suggests that it may be difficult. Let $R(\mu;\phi)$ denote the subspace of the representation $R(\mu)$ consisting of vectors of weight ϕ. Then (see Ref. 20, Thm. 4, Sec. 78, for example)

$$T^{\nu}_{\lambda,\mu} = \dim\{V\in R(\mu;\nu-\lambda) \mid E(-\alpha_i)^{\nu_i+1}V=0, \quad \forall i\in\{1,2,..,r\}\}, \tag{90}$$

where $E(-\alpha_i)$ is the generator of X_r in the Cartan–Weyl basis that decreases the weight of a vector by α_i. So-called ''good'' bases for the representations of X_r exist[21] that result in combinatorial rules when used in (90). In particular, such a good basis for $X_r = A_r$ can be labelled by the standard tableaux (or, equivalently, by Gelfand–Tsetlin patterns[22]), so that the classical Littlewood–Richardson rule results. The modified Littlewood–Richardson rule (78) corresponds to

$$T^{(m)\nu}_{\lambda,\mu} = \dim\{V \in R(\mu;\nu-\lambda) \mid E(-\alpha_i)^{\nu_i+1} V = 0, \quad \forall i \in \{0,1,2,...,r\}\backslash m\}, \tag{91}$$

where we identify $E(-\alpha_0) := E(+\theta)$. A Weyl-transformed version of the good basis of Ref. 22 exists, and the rule involving the m-standard tableaux results. However, a natural generalization of the last two relations is

$$N^{\nu}_{\lambda,\mu} = \dim\{V \in R(\mu;\nu-\lambda) \mid E(-\alpha_i)^{\nu_i+1} V = 0, \quad \forall i \in \{0,1,2,...,r\}\}. \tag{92}$$

This equation was conjectured in Ref. 11, and was motivated by the depth rule of Ref. 2. But no basis of representations of X_r exists that is good with respect to (92),[11] as can be verified explicitly for $X_r = A_r$ using the results of Ref. 22.

It seems then that a different approach is required, one relating fusions more directly to affine representations. This would allow the use of good bases for affine representations, or formulas for their characters involving the affine Bruhat order. In this context, we should mention the more recent work of Littelmann: in Ref. 16 the generalized Littlewood–Richardson rule of (78) was further generalized to a rule for tensor products of all symmetrizable Kac–Moody algebras, and the language of minimal defining chains was upgraded to one of certain paths.

ACKNOWLEDGMENTS

Thanks go to C. Cummins for conversations and for informing us of Ref. 14, T. Gannon for helpful conversations and a critical reading of the manuscript, P. Mathieu for many helpful conversations, and V. Kac for correspondence pointing out the problem of finding a combinatorial rule for WZW fusions (this was also mentioned in Ref. 7).

[1] E. Verlinde, ''Fusion rules and modular transformations in 2D conformal field theory,'' Nucl. Phys. B **300(FS22)**, 360–376 (1988).
[2] D. Gepner and E. Witten, ''Strings on group manifolds,'' Nucl. Phys. B **278**, 493–549 (1986).
[3] V. G. Kac and D. Peterson, ''Infinite-dimensional Lie algebras, theta functions and modular forms,'' Adv. Math. **53**, 125–264 (1984).
[4] V. G. Kac, *Infinite dimensional Lie algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).
[5] M. A. Walton, ''Algorithm for WZW fusion rules: A proof,'' Phys. Lett. B **241**, 365–368 (1990); ''Fusion rules in Wess–Zumino–Witten models,'' Nucl. Phys. B **340**, 777–790 (1990).
[6] P. Furlan, A. Ganchev, and V. Petkova, ''Quantum groups and fusion rules multiplicities,'' Nucl. Phys. B **343**, 205–227 (1990).
[7] F. M. Goodman and H. Wenzl, ''Littlewood–Richardson coefficients for Hecke algebras at roots of unity,'' Adv. Math. **82**, 244–265 (1990).
[8] G. R. E. Black, R. C. King, and B. G. Wybourne, ''Kronecker products for compact semisimple Lie groups,'' J. Phys. A: Math. Gen. **16**, 1555–1589 (1983).
[9] C. J. Cummins and R. C. King, ''WZW fusion rules for the classical Lie algebras,'' Can. J. Phys. **72**, 342–344 (1994).
[10] C. J. Cummins, P. Mathieu, and M. A. Walton, ''Generating functions for WZNW fusion rules,'' Phys. Lett. B **254**, 386–390 (1991).
[11] A. N. Kirillov, P. Mathieu, D. Sénéchal, and M. A. Walton, ''Can fusion coefficients be calculated from the depth rule?,'' Nucl. Phys. B **391**, 651–674 (1993).
[12] M. Demazure, ''Désingularisation des variétés de Schubert généralisées,'' Ann. Ecole Norm. Sup. **7**, 53–88 (1974); ''Une nouvelle formule des caractères,'' Bull. Sci. Math. 2 Ser. 2 **98**, 163–172 (1974).
[13] S. Kumar, ''Demazure character formula in arbitrary Kac–Moody setting,'' Invent. Math. **89**, 395–423 (1987).
[14] P. Littelmann, ''A generalization of the Littlewood–Richardson rule,'' J. Alg. **130**, 328–368 (1990).
[15] V. Lakshmibai and C. S. Seshadri, ''Standard monomial theory,'' in *Proceedings of the Hyderabad Conference on Algebraic Groups* (Manoj Prakashan, Madras, 1991), pp. 279–323.

[16] P. Littelmann, ''A Littlewood–Richardson rule for symmetrizable Kac–Moody algebras,'' Invent. Math. **116**, 329–346 (1994).
[17] N. Bourbaki, *Groupes et Algèbres de Lie* (Hermann, Paris, 1968), Chaps. IV–VI.
[18] J. Weyman, ''Pieri's formulas for classical groups,'' Contemp. Math. **88**, 177–184 (1989).
[19] A. Kuniba and T. Nakanishi, ''Level-rank duality in fusion RSOS models,'' in *Modern Quantum Field Theory*, Proceedings of the International Colloquium on Modern Quantum Field Theory, Bombay, 1990, edited by S. Das *et al.* (World Scientific, New York, 1991).
[20] D. P. Zelobenko, *Compact Lie Groups and their Representations* (American Mathematical Society, Providence, RI, 1973).
[21] O. Mathieu, ''Good bases for G-modules,'' Geom. Ded. **36**, 51–66 (1990).
[22] I. M. Gelfand and A. V. Zelevinsky, ''Multiplicities and proper bases for $gl(n)$,'' *Group Theoretical Methods in Physics*, Proceedings of the 3rd seminar, Yurmala (Nauka, Moscow, 1985), Vol. 2, pp. 22–31.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Diffeomorphism invariant integrable field theories and hypersurface motions in Riemannian manifolds

Martin Bordemann[a)]
Fakultät für Physik, Universität Freiburg, Hermann-Herder-Str. 3, 79104 Freiburg i. Br., Federal Republic of Germany

Jens Hoppe[b),c)]
Institut für Theoretische Physik, ETH Hönggerberg, CH 8093 Zürich, Switzerland

(Received 2 January 1995; accepted for publication 3 September 1997)

We discuss hypersurface motions in Riemannian manifolds whose normal velocity is a function of the induced hypersurface volume element and derive a second-order partial differential equation for the corresponding time function $\tau(x)$ at which the hypersurface passes the point x. Equivalently, these motions may be described in a Hamiltonian formulation as the singlet sector of certain diffeomorphism-invariant field theories. At least in some (infinite class of) cases, which could be viewed as a large-volume limit of Euclidean M-branes moving in an arbitrary $M+1$-dimensional Riemannian manifold, the models are integrable: In the time-function formulation the equation becomes linear [with $\tau(x)$ a harmonic function on the embedding Riemannian manifold]. We explicitly compute solutions to the large volume limit of Euclidean membrane dynamics in $\mathbb{R}^3$ by methods used in electrostatics and point out an additional gradient flow structure in $\mathbb{R}^n$. In the Hamiltonian formulation we discover infinitely many hierarchies of integrable, multidimensional, N-component theories possessing infinitely many diffeomorphism invariant, Poisson commuting, conserved charges. © *1998 American Institute of Physics.* [S0022-2488(97)00912-2]

I. INTRODUCTION

In physics, and mathematics, surface motions are usually considered independent of the parametrization, the most prominent example perhaps being the ''flow by mean curvature,'' which has been of intense mathematical interest[1] as well as of physical importance.[2] When studying the dynamics of relativistic membranes, on the other hand, one encounters[3] in a partially gauge-fixed formulation, a very interesting equation (describing the time evolution of a two-dimensional surface in $\mathbb{R}^3$) that is parametrization *dependent:* this means that different parametrizations of the initial surface Σ_0 generically lead to geometrically different shapes Σ_t at later times, t. In the case at hand, there is a residual symmetry group of area-preserving diffeomorphisms, but any reparametrization whose Jacobi determinant is different from 1 will generically change the motion. Alternatively, the equation (though first order in time) may be viewed as requiring the initial unparametrized shape Σ_0 *and* the initial unparametrized normal velocity field on Σ_0 as initial conditions (the parametrization of Σ_0, and of Σ_t, then follows, up to area-preserving diffeomorphisms of Σ_0—that do not change the geometry of the motion—from the equation). Moreover, it was shown in Ref. 4 that this parametrization-dependent equation also arises when studying a diffeomorphism-invariant field theory and restricting to the zero levels of certain constraints [such as Eq. (34) in Sec. IV of this paper]. In Ref. 5, on the other hand, some surface motions in $\mathbb{R}^3$ were shown to be best described by a (second-order partial differential) equation for $\tau(\mathbf{x})$, the time at which Σ_t contains $\mathbf{x}$.

In this paper, we significantly generalize this situation (to higher dimensions, curved embedding spaces, and more general dynamics) by considering a rather general class of parametrization-

a)Electronic mail: mbor@phyq1.physik.uni-freiburg.de
b)Heisenberg Fellow. On leave of absence from Karlsruhe University.
c)Electronic mail: hoppe@itp.phys.ethz.ch

 © 1998 American Institute of Physics

dependent hypersurface motions in Riemannian manifolds that are first order in time, viz. Eq. (1) in Sec. II. For these type of hypersurface motions we shall derive in Sec. II a second-order partial differential equation for the time-function $\tau(x)$, which in some cases (that may be viewed as large-volume limits of Euclidean M-branes or, equivalently, as highly generalized, multilinear, field theoretic extensions of Nahm's matrix equations[6]) turns out to simply be $\Delta\tau=0$, where Δ is the Laplacian of the embedding Riemannian manifold in which the motion takes place. The fact that the evolution (1) is not purely geometric, but requires the initial surface *and* its initial normal velocity (which is also reflected in the fact that the equation for the time function is second order) can be understood from a Hamiltonian point of view: we shall give an equivalent diffeomorphism-invariant (constrained) Hamiltonian formulation of these evolutions in Sec. IV. As an example, we consider the particular case of three-dimensional Euclidean space in Sec. III: the solution of the Laplace equation for the time function becomes easy to handle thanks to the formal equivalence with the potential equation of electrostatics: we compute explicit solutions of these surface motions determined by singular one-dimensional membranes at $t=-\infty$ and $t=+\infty$ (corresponding to negatively and positively charged composed loops of wire, respectively) between which two-dimensional surfaces develop in time.

Finally, and perhaps most importantly, we are led to an infinite set of hierarchies of multidimensional integrable diffeomorphism-invariant N-component field theories for which we are able to provide a complete set of Poisson-commuting, diffeomorphism-invariant, conserved charges (Sec. IV).

II. SOME HYPERSURFACE MOTIONS IN RIEMANNIAN MANIFOLDS AND THEIR TIME FUNCTIONS

Let Σ be an orientable compact connected manifold of dimension M with a fixed volume form ρ. We shall denote coordinates on Σ by $\varphi^r, 1\leqslant r\leqslant M$. Let $(\mathcal{N},\zeta)$ be an orientable Riemannian manifold of dimension $N=M+1$ on which we shall denote coordinates by $x^i, 1\leqslant i\leqslant M+1$ and whose volume form induced by ζ will be denoted by ω_ζ. Consider the set of smooth immersions $x:\Sigma\to\mathcal{N}$, $\mathrm{imm}(\Sigma,\mathcal{N})$. For each immersion x there is a unique outward normal field $n[x]:\Sigma\to T\mathcal{N}$ (depending on x and its first derivatives) attached to the parametrized hypersurface $x:\Sigma\to\mathcal{N}$. Moreover, let $x^*\zeta$ denote the Riemannian metric on Σ obtained by pulling back ζ to Σ by x and denote by $\omega_{x^*\zeta}$ the corresponding volume form on Σ. It follows that there is a unique smooth positive function $\sqrt{g[x]}/\rho:\Sigma\to\mathbb{R}$ depending on x, its first derivatives, and the reference volume form ρ such that $\omega_{x^*\zeta}=(\sqrt{g[x]}/\rho)\rho$. Finally, let α be a positive smooth function on (a suitable open interval of) the positive real line with a nowhere vanishing derivative. Consider now the hypersurface motion $(-\epsilon,\epsilon)\to\mathrm{imm}(\Sigma,\mathcal{N}):t\mapsto x(t)$ in $\mathcal{N}$ described by the following differential equation:

$$\dot{x}:=\frac{\partial x}{\partial t}=\alpha(\sqrt{g[x]}/\rho)n[x]. \tag{1}$$

In local coordinates this equation looks as follows: writing

$$g_{rs}=\frac{\partial x^i}{\partial\varphi^r}\,\frac{\partial x^j}{\partial\varphi^s}\,\zeta_{ij}(x), \tag{2}$$

for the induced metric $x^*\zeta$, we get

$$\dot{x}^i=\alpha(\sqrt{\det(g_{rs})}/\rho)\,\frac{1}{M!\sqrt{\det(g_{rs})}}\,\zeta^{ij}(x)\sqrt{\det(\zeta_{kl})}(x)\,\epsilon_{ji_1\cdots i_M}\epsilon^{r_1\cdots r_M}\,\frac{\partial x^{i_1}}{\partial\varphi^{r_1}}\cdots\frac{\partial x^{i_M}}{\partial\varphi^{r_M}}. \tag{3}$$

The following choices of α are of particular interest:

$$\text{Lorentzian } M\text{-brane:}\quad \dot{x}=\sqrt{1-(\sqrt{g[x]}/\rho)^2}\,n[x], \tag{4}$$

$$\text{Euclidean } M\text{-brane:}\quad \dot{x}=\sqrt{(\sqrt{g[x]}/\rho)^2-1}\,n[x], \tag{5}$$

and the large-volume limit of (5),

$$\dot{x} = (\sqrt{g[x]}/\rho) n[x]. \tag{6}$$

The small-volume limit of (4),

$$\dot{x} = n[x] \tag{7}$$

(which may be called the *Optical model*) is not contained in this class (since the prefactor α of the normal is constant). Nevertheless, it can be solved directly since (7) is easily seen to imply the following equation of free motion:

$$\frac{\partial^2 x^i}{\partial t^2} + \Gamma^i_{jk}(x) \frac{\partial x^j}{\partial t} \frac{\partial x^k}{\partial t} = 0 \tag{8}$$

(where Γ^i_{jk} denote the Christoffel symbols of ζ), which can be solved as soon as the geodesic flow of $(\mathcal{N},\zeta)$ is explicitly known (e.g., for flat $\mathbb{R}^N$, the N-sphere, etc.). Some of the following results will also remain true for (7).

As in Ref. 5 we are trying to derive a differential equation for the time function.

Theorem II. 1: *Let Σ, ρ, $\mathcal{N}$, ζ and α be defined as above.*

(1) Suppose that there is a positive real number ϵ and a solution x of the above equation (1) such that the map $x:(-\epsilon,\epsilon)\times\Sigma\to\mathcal{N}:(t,\varphi)\mapsto x(t,\varphi)$ is a diffeomorphism onto its image, $\mathcal{N}_\epsilon$, which is an open neighborhood of the initial hypersurface $x(0,\Sigma)$.

Then the first component of the inverse map $x^{-1}:\mathcal{N}_\epsilon\to(-\epsilon,\epsilon)\times\Sigma$ that we shall call the time function *τ satisfies the following equations where $\nabla\tau := \zeta^{\#}(d\tau)$ ($\nabla^i\tau = \zeta^{ij}\,\partial_j\tau$) denotes the gradient of τ, $|v| := \sqrt{\zeta(v,v)} = \sqrt{\zeta_{ij}v^i v^j}$ for any tangent vector v to $\mathcal{N}$, and the symbol ";" stands for the covariant derivative,*

$$n[x] = \frac{\nabla\tau}{|\nabla\tau|}(x), \tag{9}$$

$$|\nabla\tau|(x) = \frac{1}{\alpha(\sqrt{g[x]}/\rho)}, \tag{10}$$

$$0 = |\nabla\tau|^2 \zeta^{ij}\tau_{;ij} + (\tilde{\beta}(|\nabla\tau|) - 1)\,\tau_{;ij}\,\nabla\tau^i\,\nabla\tau^j, \tag{11}$$

where $\tilde{\beta}$ is defined by

$$\tilde{\beta}(z) := -z\,\frac{\partial}{\partial z}\log(\alpha^{-1}(1/z)). \tag{12}$$

Equations (9) and (10) *remain true for positive α, whose derivative may vanish.*

(2) Conversely, suppose that the smooth function $\tau:\mathcal{N}\to\mathbb{R}$ is a solution of the second-order partial differential equation (11) and that there is a positive real number ϵ such that all its level surfaces $\tau = c$ for $c\in(-\epsilon,\epsilon)$ are diffeomorphic to Σ.

Then there is a parametrization of the zero level surface Σ_0 of τ, $x_0:\Sigma\to\Sigma_0$ and a positive real number r, such that Eq. (10) is satisfied for x_0 and $r\rho$. Moreover, let Φ_t denote the flow of the vector field $X := \nabla\tau/|\nabla\tau|^2$, i.e. the solution of the ordinary differential equation $\partial\Phi_t(x)/\partial t = X(\Phi_t(x))$ with initial condition $\Phi_0(x) = x$. Then for each parametrization x_0 of the zero level surface that satisfies the above condition (10) the map $x:(-\epsilon,\epsilon)\times\Sigma\to\mathcal{N}$, defined by

$$x(t,\varphi) := \Phi_t(x_0(\varphi)), \tag{13}$$

satisfies Eq. (1) (with ρ replaced by $r\rho$) with initial condition x_0.

Proof: 1. It is obvious that the gradient of τ will be orthogonal on the surfaces of constant time whence the gradient of τ is proportional to the surface normal $n[x]$ that gives (9). Using the inverse function theorem

$$\delta^i_j = \frac{\partial x^i}{\partial t}\frac{\partial \tau}{\partial x^j} + \frac{\partial x^i}{\partial \varphi^r}\frac{\partial \varphi^r}{\partial x^j},$$

contracting with $\zeta_{ik} n[x]^k$, and using Eq. (1) and the definition of the surface normal we get the second equation (10). This leads to

$$\dot{x} = \frac{\boldsymbol{\nabla}\tau}{|\boldsymbol{\nabla}\tau|^2}(x). \tag{14}$$

In order to obtain a differential equation for the time function τ in the case where the derivative of the function α never vanishes, we differentiate Eq. (10) with respect to time and use Eq. (14) to get rid of $\dot{x}$. Thereby the time derivative of the left-hand side of (10) can solely be expressed by second covariant derivatives of the time function. To do the same for the right-hand side we use the equation

$$\dot{\sqrt{g[x]}} = \tfrac{1}{2}\sqrt{g[x]}\, g^{rs}\dot{g}_{rs},$$

obtain by the chain rule,

$$\partial_r \dot{x}^i = \partial_j(\boldsymbol{\nabla}\tau^{\,i}/|\boldsymbol{\nabla}\tau|^2)\partial_r x^j,$$

observe that the orthogonal projection π onto the tangent spaces of the hypersurfaces can be expressed in two ways by

$$g^{rs}\ \partial_r x^i\ \partial_s x^j = \pi^{ij}(x) = \zeta^{ij}(x) - n[x]^i n[x]^j$$

(see e.g., Ref. 3), and replace the normal vectors by the normalized gradients of the time function (9), which gives the second-order equation (11).

2. Denote by Σ_0 the level surface $\tau = 0$ and by $i:\Sigma_0 \to \mathcal{M}$ the canonical inclusion. Let $\omega_{i^*\zeta}$ denote the volume form on Σ_0 induced by the metric ζ. Consider the modified volume form,

$$\frac{1}{\alpha^{-1}(1/|\boldsymbol{\nabla}\tau|(i))}\ \omega_{i^*\zeta},$$

on Σ_0. Let r be the unique positive real number such that the two following integrals are equal:

$$\int_\Sigma r\rho = \int_{\Sigma_0} \frac{1}{\alpha^{-1}(1/|\boldsymbol{\nabla}\tau|(i))}\ \omega_{i^*\zeta}.$$

By Moser's lemma (see Ref. 7) there is a diffeomorphism $x_0:\Sigma \to \Sigma_0$ such that the two volume forms are diffeomorphic, i.e.,

$$r\rho = x_0^*\left(\frac{1}{\alpha^{-1}(1/|\boldsymbol{\nabla}\tau|(i))}\ \omega_{i^*\zeta}\right) = \frac{1}{\alpha^{-1}(1/|\boldsymbol{\nabla}\tau|(x_0))}\ \omega_{x_0^*\zeta},$$

whence Eq. (10) is satisfied for $t = 0$ and $r\rho$.

Observe now that for all points $y \in \Sigma_0$ we have

$$\tau(\Phi_t(y)) = t,$$

which can be seen by differentiating the left-hand side with respect to t and showing it to be equal to 1. When we insert $y = x_0(\varphi)$ in this equation and differentiate with respect to φ^r we see that the

gradient of the time function is always orthogonal to the surfaces $\Phi_t(x_0(\Sigma))$. Hence, $\nabla\tau(x(t,\varphi))/|\nabla\tau(x(t,\varphi))|$ is equal to the hypersurface normal $n[x_t](\varphi)$ whence we have the differential equation

$$\dot{x}(t,\varphi)=\frac{\nabla\tau(x(t,\varphi))}{|\nabla\tau(x(t,\varphi))|^2}=\frac{1}{|\nabla\tau(x(t,\varphi))|}\,n[x_t](\varphi).$$

We have already shown above that the function (for fixed $\varphi\in\Sigma$)

$$f(t):=\alpha(\sqrt{g_{rs}(t,\varphi)}/(r\rho(\varphi)))|\nabla\tau(x(t,\varphi))|$$

equals 1 for $t=0$ by construction of x_0. Upon differentiating f with respect to t, eliminating $\dot{x}$ by the equation above, and using the second-order equation (11), we get the first-order time-dependent differential equation for f:

$$\dot{f}(t)=\frac{\tau_{;ij}\nabla\tau^i\,\nabla\tau^j\widetilde{\beta}(|\nabla\tau|)}{|\nabla\tau|^4}(x(t,\varphi))\left(\frac{1}{\widetilde{\beta}(|\nabla\tau|)(x(t,\varphi))}-\frac{1}{\widetilde{\beta}(|\nabla\tau|(x(t,\varphi))/f(t))}\right)f(t).$$

Since the right-hand side smoothly depends on f in a neighborhood of $f=1$ the obvious constant solution $f(t)=1$ is unique. This proves Eq. (10) for all t. Q.E.D.

In other words, the open neighborhood $\mathcal{N}_\epsilon$ will be foliated by the level sets of τ, which are nothing else but the surfaces of constant time. Equations (9) and (14) have also been considered in Refs. 8 and 9 and only use the fact that $\dot{x}$ is proportional to the unit normal.

Note that for nonconstant α the above surface motion is *not* reparametrization invariant, which is here reflected in the fact that the time function satisfies a second-order partial differential equation whose normal derivative at the zero level surface encodes the ''volume part'' of the parametrization, i.e., the induced volume form compared to a reference volume. On the other hand, for the Optical model (7) the function α is equal to unity: this model of surface motion is invariant under reparametrization, and the differential equation for the time function is first order, viz. Eq. (10) for $\alpha=1$, which is the eikonal equation $|\nabla\tau|^2=1$ known in optics.

The second-order equations for all the time functions dealt with in Theorem II.1 can be derived from a Lagrangian: define the real-valued smooth function F as any solution of the first-order differential equation:

$$F'(z)=\frac{C}{\alpha^{-1}(1/z)}, \tag{15}$$

where C is an arbitrary nonzero real number. Define the Lagrangian $\mathcal{L}$ for the time function τ as:

$$\mathcal{L}(\partial\tau):=\sqrt{\det(\zeta_{ij})}F(|\nabla\tau|). \tag{16}$$

A straightforward calculation shows that the Euler Lagrange equations produce (11) for all F solving (15).

For the model (6) we can choose the Lagrangian density $\mathcal{L}(\partial\tau)$ equal to $\sqrt{\det(\zeta_{ij})}1/2|\nabla\tau|^2$, and get the Laplace equation

$$\Delta\tau=0, \tag{17}$$

where Δ is the Laplacian operator $[1/\sqrt{\det(\zeta_{ij})}]\partial_i(\sqrt{\det(\zeta_{ij})}\zeta^{ij}\partial_j$. This means that—despite its complicated nonlinear appeal—the model (6) is integrable at least for all those Riemannian manifolds for which the Laplace equation can be solved in a sufficiently explicit way, e.g., for $\mathbb{R}^n$, the sphere S^n, and hyperbolic space H^n. We shall discuss the surface motion in $\mathbb{R}^3$ (for which a linearizability was already noted in Ref. 10 without using the time function) according to its harmonic time function in more detail in the next section.

In numerical relativity theory the foliation of a given space–time into spacelike hypersurfaces, which are level surfaces of a (Lorentz) harmonic time function, has been dealt with by Bona and Massó in Ref. 11. This method of ''harmonic slicing'' avoids the occurrence of space–time singularities on these surfaces.

For the Lorentzian membrane we can choose the Lagrangian equal to the function $\sqrt{\det(\zeta_{ij})}\sqrt{|\nabla\tau|^2-1}$ and get the Euler–Lagrange equations,

$$(1-|\nabla\tau|^2)\Delta\tau+\tau_{;ij}\,\nabla\tau^{\,i}\,\nabla\tau^{\,j}=0, \tag{18}$$

whereas for the Euclidean membrane we can choose $\mathscr{L}(\partial\tau)$ equal to the function $\sqrt{\det(\zeta_{ij})}\sqrt{1+|\nabla\tau|^2}$ and get the equation

$$(1+|\nabla\tau|^2)\Delta\tau-\tau_{;ij}\,\nabla\tau^{\,i}\,\nabla\tau^{\,j}=0. \tag{19}$$

Note that the global description of surface motion by parametrization may drastically differ from the description by the level sets of a time function: although being equivalent for short time intervals (and some technical assumptions, see Theorem (II.1) the former allows the membrane to be at the same position at different times, but fixes the topological type, whereas the latter allows for varying topological type, but every point in space is contained in at most one level surface. In both pictures one has to admit violations of the regularity of the mappings (i.e., points where the Jacobians cease to have maximal rank) in order to produce interesting situations.

III. HARMONIC TIME FLOW IN THREE-DIMENSIONAL EUCLIDEAN SPACE AND ELECTROSTATICS

In this section we should like to discuss some solutions of the model (6) in $\mathbb{R}^3$ in terms of its harmonic time function. A comparison with the Euclidean membrane model (5) shows that (6) is an approximation of the Euclidean membrane model in the regime where the induced spatial membrane area element $\sqrt{\det(g_{rs})}$ is large compared to a reference area element ρ.

The Laplace equation (17) for flat $\mathbb{R}^3$ can be treated by methods known in electrostatics (see, e.g., Ref. 12). For example, consider a piece of wire of length L that is uniformally charged with line charge $\lambda=q/L$ and that is lying along the z axis between $-p:=L/2$ and p: its electrostatic potential τ is given by

$$\tau(\mathbf{x})=\lambda\,\log\left(\frac{\sqrt{x^2+y^2+(z+p)^2}+z+p}{\sqrt{x^2+y^2+(z-p)^2}+z-p}\right). \tag{20}$$

This function is harmonic outside the interval $[-p,p]$ along the z axis on which it becomes $\pm\infty$ according to the sign of λ and tends to zero for large distances from the origin. It is known that all the equipotential surfaces of this function are axially symmetric ellipsoids whose focal points are the end points of the piece of wire. In case λ is negative we get the—at least in electrostatics—unusual picture of the equipotential surfaces as surfaces of constant time: for $\tau=-\infty$ it is a singular line that blows up into bigger and bigger ellipsoids, and at $\tau=0$ the level surface is the sphere at spatial infinity.

This situation can be generalized a bit: let

$$\mathbf{x}_i^-,\mathbf{x}_j^+:S^1\to\mathbb{R}^3,\quad 1\leqslant i\leqslant m,\quad 1\leqslant j\leqslant n \tag{21}$$

be closed curves in $\mathbb{R}^3$, which are either points or immersed with only a finite number of self-intersection points. Suppose furthermore that any two different curves do not intersect. Moreover, let $\lambda_i^-,\lambda_j^+:S^1\to\mathbb{R}$ be positive smooth functions where $1\leqslant i\leqslant m$, $1\leqslant j\leqslant n$. On $\mathbb{R}^3$ minus all the images of the closed curves, $\mathbf{x}_1^-(S^1),\ldots,\mathbf{x}_m^-(S^1),\mathbf{x}_1^+(S^1),\ldots,\mathbf{x}_n^+(S^1)$ we can define the following harmonic function τ:

$$\tau(\mathbf{x}):=-\sum_{i=1}^{m}\int_0^{2\pi}d\varphi\,\frac{\lambda_i^-(\varphi)}{|\mathbf{x}-\mathbf{x}_i^-(\varphi)|}+\sum_{j=1}^{n}\int_0^{2\pi}d\varphi\,\frac{\lambda_j^+(\varphi)}{|\mathbf{x}-\mathbf{x}_j^+(\varphi)|}. \tag{22}$$

(It is harmonic because $\mathbf{x}\mapsto 1/|\mathbf{x}-\mathbf{y}|$ is harmonic for all $\mathbf{x}\neq\mathbf{y}$.) From the point of view of electrostatics, this a finite collection of closed curves carrying either positive or negative line charges. Due to the presence of the denominators in the above formula for τ it is clear that there is a negative real number $-t_0$ of a very large absolute value such that the equipotential surface $\tau=-t_0$ breaks up into m connected surfaces located each in the vicinity of one of the ''incoming curves'' $x_i^-(S^1)$. The following intuitive argument shows that each connected piece is diffeomorphic to a Riemann surface of genus equal to the number of loops that are generated by the self-intersections. In other words: the level surface piece near such a curve looks like a two-sphere if the curve is just a point, like a two-torus if the curve is immersed with no self-intersection, like a surface of genus two if the curve looks like the numeral 8, and so forth. Indeed, since $x_i^-(S^1)$ is compact we can cover it by a finite number of balls of diameter L where L can be chosen arbitrarily small. Now suppose that the map x_i^- is well behaved enough so that we can approximate it by a sequence of straight pieces of wire, each carrying a constant line charge equal to the mean value of λ_i^- over the piece of the curve. Formula (20) gives an approximate solution of the potential τ. From this formula one can infer that if the distance of the point $\mathbf{x}$ to the curve is very small compared to L, the potential is approximately proportional to the natural logarithm of the distance. Hence, for ''very early'' values of τ, the connected components of the equipotential surface are approximated by surfaces of constant distance from a given curve that visibly have the required topology. One has the same argumentation for very large or ''very late'' values of τ: here the connected components of the equipotential surfaces are certain Riemann surfaces centered around the ''outgoing curves'' $x_j^+(S^1)$.

We can interpret this geometrical picture as a theory for propagating membranes with a prescribed change of topology: if the time equals $-\infty$ one has m incoming membranes shrunk to one-dimensional objects whose homotopy and initial ''shape'' are encoded in the curves $\mathbf{x}_i^-$ with line charge functions λ_i^-. These m membranes will grow bigger and possibly melt together (which can be interpreted as an interaction) and for large positive time values the surface of equal time decomposes again into n connected components that will eventually center around the outgoing curves $\mathbf{x}_j^+$ with the final ''shape'' described by λ_j^+ and will be equal to these curves for $t=+\infty$.

In order to get more general solutions one has to solve the Dirichlet problem for the potential τ outside m conducting Riemann surfaces held at value $-t_0$ and n conducting Riemann surfaces held at value $+t_0$ for a positive real number t_0 whose solution, in principle, exists and is unique.

In order to get explicit solutions for the model (6) in higher-dimensionsional Euclidean space $\mathbb{R}^{M+1}$ we can easily generalize the above construction (22): we have to replace the curves $\mathbf{x}_i^{\mp}(\varphi)$ and by (at most) $M-1$-dimensional surfaces $\mathbf{x}_i^{\mp}(\varphi^1,\ldots,\varphi^{M-1})$ equipped with higher-dimensional ''line charge'' functions and the line integral over $1/|\mathbf{x}-\mathbf{x}_i^-(\varphi)|$ by a surface integral (of codimension two) over $1/|\mathbf{x}-\mathbf{x}_i^-(\varphi^1,\ldots,\varphi^{M-1})|^{M-1}$. More generally, one has to explicitly solve the higher-dimensional analog of the above-mentioned Dirichlet problem whose solution in principle exists and is unique.

Finally, one should point out that (6) may also be written as

$$\dot{x}^i=\int_\Sigma d^M\widetilde{\varphi}\; H^{ij}(\varphi,\widetilde{\varphi})\;\frac{\delta W}{\delta x^j(\widetilde{\varphi})}\,[x], \tag{23}$$

with

$$H^{ij}(\varphi,\widetilde{\varphi}):=\frac{\delta^{ij}\delta^M(\varphi,\widetilde{\varphi})}{\rho(\varphi)}, \tag{24}$$

$$W[x]:=\frac{1}{M+1}\int_\Sigma d^M\varphi\; x^i(\varphi)\epsilon_{ii_1\cdots i_M}\epsilon^{r_1\cdots r_M}\frac{1}{M!}\,\partial_{r_1}x^{i_1}(\varphi)\cdots\partial_{r_M}x^{i_M}(\varphi), \tag{25}$$

in $\mathbb{R}^n$, i.e., are gradient flows with respect to the volume functional W and the metric $H^{ij}(\varphi,\widetilde{\varphi})$.

IV. HAMILTONIAN FORMULATION AND CONSERVED CHARGES

Consider now a Hamiltonian field theory with fields (x,p) from Σ into the cotangent bundle $T^*\mathscr{N}$ of $\mathscr{N}$, where the conjugate momentum field p is supposed to be densitized.[13] Let the Hamiltonian be of the form

$$H(x,p) := \int_\Sigma d^M\varphi \sqrt{g}[x] h\left(\frac{|p|}{\sqrt{g}[x]}\right), \tag{26}$$

where h is a smooth function of one variable, $\sqrt{g} := \sqrt{g}[x] := \sqrt{\det g_{rs}}$ is defined as in (2), $p^2 := p_i p_j \zeta^{ij}(x), |p| := \sqrt{p^2}$, and u will be short for $|p|/\sqrt{g}$.[14]

From the canonical Poisson structure,

$$\{F,G\} = \int_\Sigma d^M\varphi \left(\frac{\delta F}{\delta x^i(\varphi)} \frac{\delta G}{\delta p_i(\varphi)} - \frac{\delta G}{\delta x^i(\varphi)} \frac{\delta F}{\delta p_i(\varphi)}\right), \tag{27}$$

we obtain Hamiltonian equations of motion,

$$\frac{\partial x^i}{\partial t} = \zeta^{ij}(x) h' \frac{p_j}{|p|}, \tag{28}$$

$$\frac{\partial p_i}{\partial t} = \zeta_{ij}(x)\partial_r((h-uh')\sqrt{g}g^{rs}\,\partial_s x^j) + \zeta_{ij}(x)(h-uh') \times \sqrt{g}g^{rs}\,\partial_r x^k\,\partial_s x^l\,\Gamma^j_{kl}(x) - \frac{1}{2}\,\partial_i\zeta^{jk}(x)h'\frac{p_j p_k}{|p|}. \tag{29}$$

Defining

$$C_r(x,p) := p_i\,\partial_r x^i \tag{30}$$

(the infinitesimal generators of diffeomorphisms), it is easy to check that (28) and (29) imply

$$\frac{\partial C_r}{\partial t} = 0 \tag{31}$$

(H is invariant under diffeomorphisms ϕ: $\Sigma \to \Sigma$), as well as

$$\frac{\partial}{\partial t}(\sqrt{g}h) = \partial_r\left(\left(\frac{hh'}{u} - h'^2\right) g^{rs} C_s(x,p)\right). \tag{32}$$

For solutions $(x(t),p(t))$ of the equations of motion satisfying $C_r(x(0),p(0)) = 0$ the rhs of Eq. (32) vanishes identically [for all t due to (31)], hence (for such solutions)

$$\sqrt{g}[x(t)] h\left(\frac{|p|(t)}{\sqrt{g}[x(t)]}\right) =: \rho, \tag{33}$$

with ρ some *time-independent* density on Σ. Inverting (33) to obtain $|p|/\sqrt{g}$ as a function of $\sqrt{g}/\rho$, and noting that for $N-M=1$ the condition

$$C_r(x,p) = p_i\,\partial_r x^i = 0, \quad 1 \leqslant r \leqslant M, \tag{34}$$

for x an immersion means that $\zeta^{ij}p_j$ must lie in the direction normal to the hypersurface defined by $\varphi \mapsto x(t,\varphi)$, one sees that (28) can be rewritten in the form mentioned in Sec. II [see (1)],

$$\frac{\partial x}{\partial t} = \alpha\left(\frac{\sqrt{g}}{\rho}\right) n, \tag{35}$$

with the functions α and h being related via

$$\alpha(z)z^2\partial_z\left(h^{-1}\left(\frac{1}{z}\right)\right)=-1. \tag{36}$$

In particular, for $h(z)=\sqrt{2z}$ we get $\alpha(z)=z$ (6) and

$$H[x,p]=\int_\Sigma d^M\varphi(4g[x]p^2)^{1/4}. \tag{37}$$

For $h(z)=\sqrt{1+z^2}$, which corresponds to $\alpha(z)=\sqrt{1-z^2}$, we get the Hamiltonian,

$$H[x,p]=\int_\Sigma d^M\varphi\sqrt{p^2+g[x]}, \tag{38}$$

of the relativistic M-brane in $M+1$-dimensional Minkowski space. Here $h(z)=z$, implying $\alpha(z)=1$, gives the Hamiltonian

$$H[x,p]=\int_\Sigma d^M\varphi\sqrt{p^2}, \tag{39}$$

of the Optical model (7).

Let us for a moment concentrate on (37): Motivated by the fact that the equation for the time function is linear [see (17)] and the observation,[5] made for $N=3$, that a Lax pair formulation of $\partial\mathbf{x}/\partial t=\sqrt{g}\mathbf{n}$ implies the time independence of $\int Q(x(t,\varphi^1,\varphi^2)d\varphi^1\,d\varphi^2$ for harmonic polynomials $Q(x^1,x^2,x^3)$, one is led to the conjecture that for all $N=M+1$, ρ, and $Q(x)$ any harmonic function of x,

$$\mathsf{Q}[x]:=\int_\Sigma d^M\varphi\ \rho Q(x), \tag{40}$$

will be independent of time if the hypersurface $\Sigma_t=x_t(\Sigma)$ evolves according to (6). The proof is a trivial application of Gauss' Theorem: assuming that for all t the moving surface Σ_t is a boundary of some open set V_t of $\mathscr{N}$ and writing dS_t^i for the surface element, we get

$$\frac{\partial\mathsf{Q}[x(t)]}{\partial t}=\int_\Sigma d^M\varphi\ \rho\ \partial_iQ(x(t))\ \frac{\partial x^i}{\partial t}=\int_\Sigma d^M\varphi\sqrt{g}[x(t)]n^i[x(t)]\partial_iQ(x(t))=\int_{\Sigma_t}dS_t^i\ \partial_iQ(x(t))$$

$$=\int_{V_t}d^Nx\sqrt{\det(\zeta_{kl})}\partial_j(\zeta^{ji}\ \partial_iQ)=\int_{V_t}d^Nx\sqrt{\det(\zeta_{kl})}\Delta Q=0.$$

In the Hamiltonian formulation (37) the nondynamical density $\rho(\varphi)$ is simply replaced by the Hamiltonian density

$$\mathscr{H}[x,p](\varphi):=\sqrt{g}[x](\varphi)h\left(\frac{|p|(\varphi)}{\sqrt{g}[x](\varphi)}\right), \tag{41}$$

i.e., $\mathscr{H}[x,p]=(g[x]p^2)^{1/4}$ in the case of (37). Indeed, for any harmonic Q the following functionals on phase space,

$$\mathsf{Q}[x,p]:=\int_\Sigma d^M\varphi(gp^2)^{1/4}Q(x), \tag{42}$$

will be time independent for solutions $(x(t),p(t))$ satisfying (28), (29), and the constraint (34).

Do the quantities (42) Poisson commute (on the reduced space)? The answer is ''yes'' (in fact, they form an infinite, presumably complete set of Poisson commuting, reparametrization invariant charges, thus providing the solution of a nontrivial diffeomorphism-invariant field theory)—but let

us first make some observations valid for all theories of the general form (26), i.e. arbitrary h, N, M, and ζ: With the canonical Poisson structure (27) one has the following Poisson brackets, where $f,\widetilde{f}$ are arbitrary smooth real-valued functions on Σ and X,Y are arbitrary vector fields on Σ:

$$\left\{\int_\Sigma d^M\varphi\ X^r C_r, \int_\Sigma d^M\varphi\ Y^s C_s\right\} = \int_\Sigma d^M\varphi [X,Y]^r C_r, \tag{43}$$

$$\left\{\int_\Sigma d^M\varphi\ f\mathscr{H}, \int_\Sigma d^M\varphi\ \widetilde{f}\mathscr{H}\right\} = \int_\Sigma d^M\varphi\left(\frac{hh'}{u}-h'^2\right)\times g^{rs}(f\ \partial_s\widetilde{f}-\widetilde{f}\ \partial_s f)C_r, \tag{44}$$

$$\left\{\int_\Sigma d^M\varphi\ f\mathscr{H}, \int_\Sigma d^M\varphi\ Y^s C_s\right\} = -\int_\Sigma d^M\varphi\ Y^s\ \partial_s f\ \mathscr{H}, \tag{45}$$

where $[X,Y]^r$ denotes the Lie bracket $\partial_s Y^r\ X^s-\partial_s X^r\ Y^s$ of X and Y.

While (43), together with (31), imply that, in principle, the Hamiltonian theory (26) may be reduced to a Hamiltonian theory on a diffeomorphism-invariant phase space (involving only $N-M$ fields and $N-M$ momenta) one should note that on the reduced phase space the density $\mathscr{H}$ (41) can no longer be defined, unless integrated over particular functions. This explains that although $\dot{\mathscr{H}}=0$ (32) for solutions of (28) and (29), satisfying the constraint (31), the Dirac bracket of $\mathscr{H}$ with H [see, e.g., Ref. 4 for the case (38)] would not give zero, due to (45). In particular, $\mathscr{H}$ is *not* a conserved quantity in the Hamiltonian theory.

The charges Q (42), however, *are* conserved on the reduced phase space: $\mathsf{Q}[x,p]=\int_\Sigma d^M\ \varphi\ \mathscr{H}Q(x)$ commutes with C_r,

$$\left\{\int_\Sigma d^M\varphi\ \mathscr{H}Q(x), \int_\Sigma d^M\varphi\ Y^r C_r\right\} = -\int_\Sigma d^M\varphi\ Y^r\ \partial_r Q(x)\mathscr{H} + \int_\Sigma d^M\varphi\ \mathscr{H}\ \partial_i Q(x) Y^r\ \partial_r x^i = 0 \tag{46}$$

[irrespective of Q, h, and $N-M$; just use (45) with $\widetilde{f}=Q(x)$ for the first term on the rhs of (46)], as well as with H (provided $C_r=0$, $N=M+1$, and $hh'=\text{const}$). Moreover,

$$\begin{aligned}
&\left\{\int_\Sigma d^M\varphi\ \mathscr{H}Q(x), \int_\Sigma d^M\varphi\ \mathscr{H}\widetilde{Q}(x)\right\}\\
&\quad=\int_\Sigma d^M\varphi\left(\frac{hh'}{u}-h'^2\right)g^{rs}(Q(x)\partial_s\widetilde{Q}(x)-\widetilde{Q}(x)\partial_s Q(x))C_r\\
&\qquad-\int_\Sigma d^M\varphi\sqrt{g}[x]h'h\ \frac{p_i}{|p|}\ \zeta^{ij}(x)(Q(x)\partial_j\widetilde{Q}(x)-\widetilde{Q}(x)\partial_j Q(x))\\
&\quad=\text{const}\int_\Sigma d^M\varphi\sqrt{g}[x]n^j[x](Q(x)\partial_j\widetilde{Q}(x)-\widetilde{Q}(x)\partial_j Q(x))\\
&\quad=\text{const}\int_{V_t} d^N x\sqrt{\zeta kl}\,\partial_i(\zeta^{ij}(Q\partial_j\widetilde{Q}-\widetilde{Q}\partial_j Q))=0,
\end{aligned} \tag{47}$$

where again we have assumed that Σ_t is the boundary of V_t. In (47), the first equality is general [using (44), with $f=Q(x)$, $\widetilde{f}=\widetilde{Q}(x)$, for the first term], the second equality holds if $C_r=0$ ($1\leqslant r\leqslant M$) and $N=M+1$ (so that $\zeta^{ij}p_j$ is proportional to the surface normal), and the last equality holds if Q and $\widetilde{Q}$ are harmonic functions (it actually holds as long as $\widetilde{Q}\ \Delta Q-Q\ \Delta\widetilde{Q}=0$, which is a much weaker requirement).

The preceding observation has far-reaching consequences. It not only proves the Poisson commutativity of the conserved charges Q given in (42) hence the integrability of (37); the freedom of choosing $Q(x)$ in $\mathbb{R}^3$, e.g. as a superposition $\Sigma_{l=0,|m|\leqslant l}^{\infty}\ q_{lm}Y_{lm}(\mathbf{x})$ with $Y_{lm}(\mathbf{x})$ being the solid spherical harmonics, neatly matches the degrees of freedom to be expected from the one

single field that is left over after the Hamiltonian reduction] but—confirming that integrable theories come in hierarchies—shows that *any* of the charges (42) with $\Delta Q=0$ may be considered as the Hamiltonian of another diffeomorphism-invariant field theory [possessing (42) as commuting conserved charges].

Due to

$$\int_\Sigma d^M\varphi\ Q(x)(4g[x]p^2)^{1/4}=\int_\Sigma d^M\varphi(4\widetilde{g}[x]\widetilde{p}^2)^{1/4}=:\widetilde{H}[x,p], \tag{48}$$

corresponding to $\widetilde{Q}(x)=1$ for a Riemannian manifold $(\mathscr{N},\widetilde{\zeta})$ with a conformally equivalent metric,

$$\widetilde{\zeta}_{ij}(x):=(Q(x))^{4/(N-2)}\zeta_{ij}(x) \tag{49}$$

[implying that $\sqrt{\widetilde{g}}=(Q(x))^{2(N-1)/N-2)}\sqrt{g}$, $\widetilde{p}^{\,2}=\widetilde{p}_i\widetilde{p}_j\widetilde{\zeta}^{ij}=p_ip_j\zeta^{ij}Q^{-4/(N-2)}$; $N\neq 2$; the very special case $N=2$ can be dealt with separately], and the fact that under the conformal change (49) the scalar curvature changes according to

$$\widetilde{R}=Q^{-4/(N-2)}\left(R-4\frac{N-1}{N-2}\frac{\Delta Q}{Q}\right), \tag{50}$$

$$=Q^{-4/(N-2)}R, \tag{51}$$

one sees that each hierarchy (of integrable Hamiltonian systems, resp. harmonic time functions) consists of hypersurface motions in Riemannian manifolds having conformally equivalent metrics and satisfying

$$\widetilde{R}\widetilde{\zeta}_{ij}=R\zeta_{ij}\,. \tag{52}$$

This intriguing observation should have many important consequences. Taking, e.g., $\mathscr{N}=\mathbb{R}^3$ and $\zeta_{ij}=\delta_{ij}$, the above suggests that the solutions of the linear, but nontrivial, equation,

$$\boldsymbol{\nabla}\cdot(Q^2(\mathbf{x})\boldsymbol{\nabla}\tau(\mathbf{x}))=0, \tag{53}$$

resp., the corresponding nonlinear equations,

$$\frac{\partial\mathbf{x}}{\partial t}=Q^2(\mathbf{x})\ \frac{\sqrt{\det(\partial_r\mathbf{x}\cdot\partial_s\mathbf{x})}}{\rho}\ \mathbf{n}, \tag{54}$$

are related to

$$\nabla^2\tau(\mathbf{x})=0, \tag{55}$$

resp.,

$$\frac{\partial\mathbf{x}}{\partial t}=\frac{\sqrt{\det(\partial_r\mathbf{x}\cdot\partial_s\mathbf{x})}}{\rho}\ \mathbf{n}, \tag{56}$$

via some generalized Bäcklund transformation for all $Q(\mathbf{x})$ satisfying

$$\nabla^2Q(\mathbf{x})=0. \tag{57}$$

ACKNOWLEDGMENTS

The authors would like to thank J. Fröhlich, D. Giulini, K. Happle, and M. Struwe for valuable discussions, C. Kiefer for pointing out Ref. 11, and the Physics Department of the ETH Zürich for friendly hospitality.

[1] K. A. Brakke, ''The motion of a Surface by its mean curvature,'' *Mathematical Notes* (Princeton University Press, Princeton NJ, 1978); G. Huisken, ''Flow by mean curvature of convex surfaces into spheres,'' J. Diff. Geom. **20**, 237 (1984); L. Evans and J. Spruck, ''Motion of level sets by mean curvature,'' *ibid.* **33**, 635 (1991); Y. G. Chen, Y. Giga, and S. Goto, ''Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations,'' *ibid.* **33**, 749 (1991).
[2] P. Pelcé, *Dynamics of Curved Fronts* (Academic, New York, 1988); S. Osher and J. A. Sethian, J. Comput. Phys. **79**, 12 (1988).
[3] M. Bordemann and J. Hoppe, ''The dynamics of relativistic membranes II: Nonlinear waves and covariantly reduced membrane equations,'' Phys. Lett. B **325**, 359–365 (1994).
[4] J. Hoppe, ''Canonical 3+1 description of relativistic membranes,'' Yukawa Institute preprint YITP/K-1079.
[5] J. Hoppe, ''Surface motions and fluid dynamics,'' Phys. Lett. B **335**, 41–44 (1994).
[6] N. Hitchin, ''On the construction of Monopoles,'' Commun. Math. Phys. **89**, 145–190 (1989).
[7] J. Moser, ''On the volume element on a manifold,'' Trans. Am. Math. Soc. **120**, 286–294 (1965).
[8] K. Nakayama, J. Hoppe, and M. Wadati, ''On the level-set formulation of geometrical models,'' J. Phys. Soc. Jpn. **64**, 403–407 (1995).
[9] K. Nakayama and M. Wadati, ''Reaction-diffusion system in a curved space and the KPZ equation,'' J. Phys. Soc. Jpn. **64**, 1501–1505 (1995).
[10] E. G. Floratos and G. K. Leontaris, Phys. Lett. B **223**, 153 (1989); R. S. Ward, ''Linearization of the $SU(\infty)$ Nahm equations,'' *ibid.* **234**, 81–84 (1990).
[11] C. Bona and J. Massó, ''Harmonic synchronisations of spacetime,'' Phys. Rev. D **38**, 2419–2422 (1988).
[12] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1975).
[13] Although we shall always be working in canonical x^i and p_i coordinates (a bundle chart) on $T^*\mathscr{N}$ this has a global meaning: consider the one-dimensional real vector bundle of contravariant densities, $|\Lambda|T\Sigma$, over Σ [its smooth sections can be regarded as smooth real-valued functions $\hat{\rho}$ on the bundle $L(\Sigma)$ of linear frames of Σ that are equivariant under the structure group $GL(n,\mathbb{R})$ in the sense $\hat{\rho}(vg)=|\det(g)|^{-1}\hat{\rho}(v)$ for all $v\in L(\Sigma)$ and $g\in GL(n,\mathbb{R})$], with bundle projection $|\Lambda|\tau_\Sigma:|\Lambda|T\Sigma\to\Sigma$. Let $\tau^*_{\mathscr{N}}:T^*\mathscr{N}\to\mathscr{N}$ denote the canonical bundle projection. The pair (x,p) can now be regarded as a smooth vector bundle homomorphism $|\Lambda|T\Sigma\to T^*\mathscr{N}$ over the map $x:\Sigma\to\mathscr{N}$, i.e. $x\circ|\Lambda|\tau_\Sigma=\tau^*_{\mathscr{N}}\circ(x,p)$. Alternatively, since the bundle of densities is trivializable we may choose a fixed covariant nonvanishing positive reference density ρ_0 on Σ (a section in the bundle dual to $|\Lambda|T\Sigma$ and that induces an integral over Σ) and write $(x,p)=(x,p')\rho_0$, where (x,p') is an arbitrary smooth map $\Sigma\to T^*\mathscr{N}$ such that $x=\tau^*_{\mathscr{N}}\circ(x,p')$. In this fashion we can establish a (noncanonical) one-one-correspondence between the set of ''densitized maps'' as defined above and all smooth maps $\Sigma\to T^*\mathscr{N}$.
[14] In order to avoid clumsy notation we shall henceforth suppress the arguments $\varphi\in\Sigma$ in the integrals, the field x and its derivatives in $\sqrt{g}[x]$ and $n[x]$, and the argument u of h and h' from time to time.

New integrable generalizations of Calogero–Moser quantum problem

O. Chalykh[a] and M. Feigin[a]
Department of Mathematics and Mechanics, Moscow State University, Moscow 119899, Russia

A. Veselov[b]
Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom and Landau Institute for Theoretical Physics, Kosygina 2, Moscow 117940, Russia

(Received 21 July 1997; accepted for publication 9 September 1997)

A one-parameter deformation of Calogero–Moser quantum problem is introduced. It is shown that corresponding Schrödinger operator is integrable for any value of the parameter and algebraically integrable in case of integer value. © *1998 American Institute of Physics.* [S0022-2488(98)00902-5]

I. INTRODUCTION

Let's consider the following Schrödinger operator describing the one-dimensional $(N+1)$-body problem:

$$L=\sum_{j=1}^{N+1}\frac{\hat{p}_j^2}{2m_j}+m(m+1)\sum_{i<j}^{N+1}\frac{m_i m_j}{(q_i-q_j)^2}, \tag{1}$$

where all but one of the ''masses'' are equal, $m_1=\cdots=m_N=1$, $m_{N+1}=m^{-1}$, m is some real parameter, $\hat{p}_j=i(\partial/\partial q_j)$ $(j=1,\ldots,N+1)$.

When $m=1$ one has a particular case of the Calogero operator[1] considered in Ref. 2.

The aim of the present paper is to show that such a deformation of the Calogero problem is still integrable in the quantum (but not in the classical!) case.

Recall that a Schrödinger operator in $\mathbf{R}^n$ is called *integrable* if there are n commuting differential operators $L_1=L,L_2,\ldots,L_n$ with constant and independent highest symbols $P_i(k)$. If there exists at least one more commuting operator L_{N+1} with a constant highest symbol $P_{n+1}(k)$ such that it takes the different values at the solutions of the algebraic system $P_i(k)=c_i$ $(i=1,\ldots,n)$, then L is called *algebraically integrable* (see Refs. 3 and 4).

Our main result is the following:

Theorem: *The Schrödinger operator (1) is integrable for an arbitrary m and algebraically integrable for any integer m.*

We prove also that the same is true for the corresponding trigonometric version

$$L=\sum_{j=1}^{N+1}\frac{\hat{p}_j^2}{2m_j}+m(m+1)\sum_{i<j}^{N+1}\frac{m_i m_j\omega^2}{\sin^2\omega(q_i-q_j)}, \tag{2}$$

$m_1=\ldots=m_N=1$, $m_{N+1}=m^{-1}$, which is the deformation of the Sutherland operator.[5]

We conjecture that the same is true also for the elliptic version

$$L=\sum_{j=1}^{N+1}\frac{\hat{p}_j^2}{2m_j}+m(m+1)\sum_{i<j}^{N+1}m_i m_j\wp(q_i-q_j), \tag{3}$$

$m_1=\ldots=m_N=1$, $m_{N+1}=m^{-1}$.

[a]Electronic mail: chalykh@nw.math.msu.su, mfeigin@dnttm.ru
[b]Electronic mail: A.P.Veselov@lboro.ac.uk

 © 1998 American Institute of Physics

It is known to be true for $m=1$ (integrability,[6] algebraic integrability[7,8] for $N=2$ and Refs. 9, 10, and 11 for any N). For $N=2$ and any m the integrability of (3) was proven in Ref. 8.

In the coordinates $x_1=q_1, \ldots, x_N=q_N, x_{N+1}=(1/\sqrt{m})\,q_{N+1}$ the operators (1), (2), (3) have the form

$$L=-\Delta+\sum_{i<j}^{N} 2m(m+1)u(x_i-x_j)+\sum_{i=1}^{N} 2(m+1)u(x_i-\sqrt{m}x_{N+1}) \tag{4}$$

with $u(x)=x^{-2}, \omega^2 \sin^{-2}\omega x$ or $\wp(x)$, $\Delta=\partial_1^2+\ldots+\partial_{N+1}^2, \partial_j=\partial/\partial x_j$.

Olshanetsky and Perelomov[6] proposed the generalizations of the Calogero–Moser system of the form

$$L=-\Delta+\sum_{\alpha\in\mathfrak{R}_+} m_\alpha(m_\alpha+1)(\alpha,\alpha)u((\alpha,x)), \tag{5}$$

related to any root system $\mathfrak{R}$. Our operator (4) is related to the set $\mathbf{A}_N(m)$ which consists of the vectors $\alpha=e_i-e_j$ $(i<j\leqslant N)$ with multiplicities $m_\alpha=m$ and $\alpha=e_i-\sqrt{m}e_{N+1}$ $(i=1,\ldots,N)$ with $m_\alpha=1$. We may consider this set as a deformation of the root system $\mathbf{A}_N$.

It turns out that such a deformation exists also for the root system $\mathbf{B}_2$. It depends on two parameters l,m and consists of the vectors $\alpha_1=\sqrt{2m+1}e_1$ with multiplicity l, $\alpha_2=\sqrt{2l+1}e_2$ with multiplicity m and two vectors $\alpha_{3,4}=\frac{1}{2}(\alpha_1\pm\alpha_2)$ with multiplicity 1. We denote this set as $\mathbf{B}_2(l,m)$. The corresponding Schrödinger operator has the form:

$$\begin{aligned} L=&-\Delta+l(l+1)(2m+1)u(\sqrt{2m+1}x_1)+m(m+1)(2l+1)u(\sqrt{2l+1}x_2)+(l+m+1)\\ &\times u(\tfrac{1}{2}\sqrt{2m+1}x_1+\tfrac{1}{2}\sqrt{2l+1}x_2)+(l+m+1)u(\tfrac{1}{2}\sqrt{2m+1}x_1-\tfrac{1}{2}\sqrt{2l+1}x_2). \end{aligned} \tag{6}$$

We prove that this operator is integrable in rational $u=x^{-2}$ and trigonometric $u=\omega^2\sin^{-2}\omega x$ cases.

All these deformations appeared in Ref. 12 within a construction of the multidimensional Baker–Akhiezer function related to a configuration of hyperplanes proposed in Refs. 2 and 4.

In the rational case we can add the quadratic term Ω^2x^2 to the potential without destroying integrability:

$$\begin{aligned} L=&\hat{p}_1^2+\ldots+\hat{p}_N^2+m\hat{p}_{N+1}^2+2m(m+1)\sum_{i<j}^{N}\frac{1}{(q_i-q_j)^2}+2(m+1)\\ &\times\sum_{i=1}^{N}\frac{1}{(q_i-q_{N+1})^2}+\Omega^2\left(q_1^2+\ldots+q_N^2+\frac{1}{m}q_{N+1}^2\right). \end{aligned}$$

II. INTERTWINING OPERATORS AND INTEGRABILITY OF THE DEFORMED CALOGERO OPERATOR

We start with the standard Calogero operator $\mathscr{L}$ in $\mathbf{R}^N$:

$$\mathscr{L}=-\Delta+\sum_{i<j}^{N} 2m(m+1)(x_i-x_j)^{-2}. \tag{7}$$

It is well known that the operator $\mathscr{L}$ is integrable for any m. Moreover, for the symmetric polynomials $s_i(k)=k_1^i+\cdots+k_N^i$ $(i=1,\ldots,N)$ there exist commuting differential operators $\mathscr{L}_i$ with the highest symbols $s_i(k)$.[13,6] In Ref. 4 it was proven that for any *integer m* the operator (7) is *algebraically integrable*.

We will prove now that the same is true for the deformed Calogero operator,

$$L=-\Delta+\sum_{i<j}^{N}\frac{2m(m+1)}{(x_i-x_j)^2}+\sum_{i=1}^{N}\frac{2(m+1)}{(x_i-\sqrt{m}x_{N+1})^2}. \tag{8}$$

The idea of the proof is similar to the one we used in Ref. 2: We present a differential operator D which intertwines the operator (8) and

$$\hat{\mathscr{L}} = -\partial_{N+1}^2 + \mathscr{L}. \tag{9}$$

Let's introduce a set of differential operators $D_n^{i_1,\dots,i_k}$ by the following recursive formulas:

$$D_n = 1,\ D_n^i = \partial_i - \sqrt{m}\partial_n - (m+1)(x_i - \sqrt{m}x_n)^{-1},$$

$$D_n^{i_1,\dots,i_k} = D_n^{i_1,\dots,i_{k-1}} \circ (\partial_{i_k} - \sqrt{m}\partial_n) - (m+1)(x_{i_k} - \sqrt{m}x_n)^{-1} D_n^{i_1,\dots,i_{k-1}} \tag{10}$$

$$+ \sum_{s=1}^{k-1} \frac{m(m+1)}{(x_{i_s} - x_{i_k})^2} D_n^{i_1,\dots,\hat{i}_s,\dots,i_{k-1}}.$$

Here $\partial_i = \partial/\partial x_i$, $i_1, \dots, i_k$ are different natural indices less than n and the symbol $\hat{i}_s$ means that the index i_s must be omitted.

Theorem 1: *The operator $D = D_{N+1}^{1,\dots,N}$ defined by the formulas (10) intertwines the operator (8) and Calogero operator $\hat{\mathscr{L}}$*:

$$L \circ D = D \circ \hat{\mathscr{L}}. \tag{11}$$

The theorem can be proved by repeating the arguments of Ref. 14 where the case $m = 1$ was considered.

Corollary: For any operator $\hat{A}$ commuting with Calogero operator $\hat{\mathscr{L}}$ the operator $A = D \circ \hat{A} \circ D^*$ commutes with L where D^* means the operator formally adjoint to D.

Indeed, since L and $\hat{\mathscr{L}}$ are self-adjoint, it follows from (11) that $D^* \circ L = \hat{\mathscr{L}} \circ D^*$ and therefore

$$AL = D\hat{A}D^*L = D\hat{A}\hat{\mathscr{L}}D^* = D\hat{\mathscr{L}}\hat{A}D^* = LD\hat{A}D^* = LA.$$

Now taking integrals for $\hat{\mathscr{L}}$ as $\hat{A}$ we obtain independent integrals for L. The same arguments as in Ref. 15 show that these integrals commute with each other.

Taking the symmetric integrals $\hat{A} = \mathscr{L}_i$ $(i = 1, \dots, N)$ and $\hat{A} = \partial_{N+1}$ we obtain $N+1$ commuting integrals and therefore the integrability for the deformed Calogero system (8). Using the extra integrals of the system (7) found in Ref. 4 for integer m we construct in such a way the extra integrals for the system (8) to prove its algebraic integrability.

In the trigonometric case (4) the intertwining operator D is determined by the recursion formulas (cf. Ref. 14):

$$D_n = 1,\ D_n^i = \partial_i - \sqrt{m}\partial_n - (m+1)\,\omega \cot \omega(x_i - \sqrt{m}x_n),$$

$$D_n^{i_1,\dots,i_k} = D_n^{i_1,\dots,i_{k-1}} \circ (\partial_{i_k} - \sqrt{m}\partial_n) - (m+1)\,\omega \cot \omega(x_{i_k} - \sqrt{m}x_n) D_n^{i_1,\dots,i_{k-1}}$$

$$+ \sum_{s=1}^{k-1} \frac{m(m+1)\omega^2}{\sin^2 \omega(x_{i_s} - x_{i_k})} D_n^{i_1,\dots,\hat{i}_s,\dots,i_{k-1}},$$

$$D = D_{N+1}^{1,\dots,N}.$$

It satisfies the relation

$$L \circ D = D \circ \hat{\mathscr{L}},$$

where

$$L=-\Delta+\sum_{i<j}^{N}\ 2m(m+1)\frac{\omega^2}{\sin^2\ \omega(x_i-x_j)}+\sum_{i=1}^{N}\ 2(m+1)\frac{\omega^2}{\sin^2\ \omega(x_i-\sqrt{m}x_{N+1})}, \tag{12}$$

and

$$\hat{\mathscr{L}}=-(\partial_1^2+\ \ldots\ +\partial_{N+1}^2)+\sum_{i<j}^{N}\ 2m(m+1)\frac{\omega^2}{\sin^2\ \omega(x_i-x_j)} \tag{13}$$

is the Sutherland operator[5].

Since the Sutherland system is integrable for any m (see Ref. 6) and algebraically integrable for integer m[4] we can use the above argument to prove the same property for the deformed Sutherland operator (12).

Theorem 2: *The deformed Calogero–Sutherland operators (8), (12) are integrable for any value of the parameter m and algebraically integrable when m is an integer.*

Notice that the existence of the intertwining operator D allows one to find the solutions of the corresponding Schrödinger equation $L\psi=\lambda\psi$ as $\psi=D\psi_0$ where ψ_0 is the solution of the equation $\hat{\mathscr{L}}\psi_0=\lambda\psi_0$. In particular, for integer m we have the eigenfunction for L of the form

$$\psi=P(k,x)e^{(k,x)}, \tag{14}$$

where $P(k,x)$ is some polynomial on k with the highest terms

$$P(k)=\prod_{i<j}^{N}\ (k_i-k_j)^m\prod_{i=1}^{N}\ (k_i-\sqrt{m}k_{N+1})$$

(cf. Refs. 4 and 12). This function satisfies certain properties in k-variables (multidimensional Baker–Akhiezer function, see Ref. 12).

The intertwining operator exists also for the Schrödinger operator $L+\Omega^2x^2=L_\Omega$:

$$L_\Omega=-\Delta+\sum_{i<j}^{N}\frac{2m(m+1)}{(x_i-x_j)^2}+\sum_{i=1}^{N}\frac{2(m+1)}{(x_i-\sqrt{m}x_{N+1})^2}+\Omega^2(x_1^2+\cdots+x_{N+1}^2). \tag{15}$$

It is defined by the following recursive formulas:

$$D_n=1,\ D_n^i=\partial_i-\sqrt{m}\partial_n-\left(\frac{m+1}{x_i-\sqrt{m}x_n}+\Omega(x_i-\sqrt{m}x_n)\right),$$

$$D_n^{i_1,\ldots,i_k}=D_n^{i_1,\ldots,i_{k-1}}\circ(\partial_{i_k}-\sqrt{m}\partial_n)-\left(\frac{m+1}{x_{i_k}-\sqrt{m}x_n}+\Omega(x_{i_k}-\sqrt{m}x_n)\right)D_n^{i_1,\ldots,i_{k-1}}$$

$$+\sum_{s=1}^{k-1}\left(\frac{m(m+1)}{(x_{i_s}-x_{i_k})^2}-m\Omega\right)D_n^{i_1,\ldots,\hat{i}_s,\ldots,i_{k-1}}.$$

The following result can be proven similar to Ref. 14.

Theorem 3: *For* $D_\Omega=D_{N+1}^{1,\ldots,N}$ *we have*:

$$D_\Omega\circ(\mathscr{L}_\Omega+2N\Omega)=L_\Omega\circ D_\Omega, \tag{16}$$

where

$$\mathscr{L}_\Omega=-\Delta+\sum_{i<j}^{N}\frac{2m(m+1)}{(x_i-x_j)^2}+\Omega^2(x_1^2+\cdots+x_{N+1}^2). \tag{17}$$

Corollary: L_Ω *is integrable for any* m.

Again, the integrals can be obtained as

$$L_i^{\Omega}=D_{\Omega}\mathscr{L}_i D_{\Omega}^{*}$$

where $\mathscr{L}_i$ $(i=1,\ldots,N+1)$ are the integrals for the Calogero operator (17).

We should note that not all the integrals have such a form. For example, we can prove using the results of Ref. 12 the following

Theorem 4: *For the deformed Calogero operator (8) there exist commuting integrals $L_s(m)$ with the highest symbols*

$$p_s(k)=k_1^s+\ldots+k_N^s+m^{\frac{s-2}{2}}k_{N+1}^s.$$

When $m=1$ these integrals coincide with the standard integrals L_s of the Calogero system,[13] so $L_s(m)$ can be considered as their deformation.

III. THREE-PARTICLE CASE

We consider now the deformed three-particle Sutherland operator (2):

$$L=-\partial_1^2-\partial_2^2-m\partial_3^2+2m(m+1)u(q_1-q_2)+2(m+1)u(q_1-q_3)+2(m+1)u(q_2-q_3), \tag{18}$$

where $u(x)=\omega^2\ \sin^{-2}\ \omega x$.

In this case the intertwining operator D has the following form:

$$\begin{aligned}D=&(\partial_1-m\partial_3)(\partial_2-m\partial_3)-(m+1)\omega\ \cot\ \omega(q_1-q_3)(\partial_2-m\partial_3)\\&-(m+1)\omega\ \cot\ \omega(q_2-q_3)(\partial_1-m\partial_3)+(m+1)^2\omega^2\ \cot\ \omega(q_1-q_3)\cot\ \omega(q_2-q_3)\\&+m(m+1)\omega^2\ \sin^{-2}\ \omega(q_1-q_2).\end{aligned}$$

We have $L\circ D=D\circ\mathscr{L}$ with

$$\mathscr{L}=-\partial_1^2-\partial_2^2-m\partial_3^2+2m(m+1)\omega^2\ \sin^{-2}\ \omega(q_1-q_2).$$

When m is an integer we can intertwine $\mathscr{L}$ with $-\Delta=-\partial_1^2-\partial_2^2-m\partial_3^2$ by the following operator S:

$$\mathscr{L}\circ S=S\circ(-\Delta),$$

$$\begin{aligned}S=&(\partial_1-\partial_2-2m\omega\ \cot\ \omega(q_1-q_2))\circ(\partial_1-\partial_2-2(m-1)\omega\ \cot\ \omega(q_1-q_2))\ldots\circ(\partial_1-\partial_2\\&-2\omega\ \cot\ \omega(q_1-q_2)).\end{aligned}$$

It follows from this that the operator $D\circ S$ intertwines L and $-\Delta$:

$$L\circ(D\circ S)=D\circ\mathscr{L}\circ S=(D\circ S)\circ(-\Delta).$$

The quantum integrals for L can be written explicitly:

$$L_1=\partial_1+\partial_2+\partial_3,\ L_2=L,$$

$$L_3=\partial_1\partial_2\partial_3+\frac{1-m}{2}(\partial_1+\partial_2)\partial_3^2+\frac{(1-m)(1-2m)}{6}\partial_3^3$$
$$+(m+1)u(q_2-q_3)\partial_1+(m+1)u(q_1-q_3)\partial_2+m(m+1)u(q_1-q_2)\partial_3$$
$$+\frac{1-m^2}{2}(u(q_1-q_3)\circ\partial_3+\partial_3\circ u(q_1-q_3))$$
$$+\frac{1-m^2}{2}(u(q_2-q_3)\circ\partial_3+\partial_3\circ u(q_2-q_3)),$$

where $u(x)=\omega^2\ \sin^{-2}\omega x$.

While $\omega\to 0$ we obtain the analogous formulas for the deformed Calogero operator. It can be checked that the above formulas work also for the elliptic version of the operator (18) with $u(x)=\wp(x)$. This provides the integrability of the deformed elliptic Calogero–Moser system of three particles for the quantum case.

IV. DEFORMATION OF B_2-CASE

Let's consider the following Schrödinger operator:

$$L=-(2m+1)\frac{\partial^2}{\partial x_1^2}-(2l+1)\frac{\partial^2}{\partial x_2^2}+\frac{l(l+1)(2m+1)}{\sin^2 x_1}+\frac{m(m+1)(2l+1)}{\sin^2 x_2}$$
$$+\frac{(l+m+1)}{\sin^2\frac{1}{2}(x_1+x_2)}+\frac{(l+m+1)}{\sin^2\frac{1}{2}(x_1-x_2)}, \tag{19}$$

which is a trigonometric version of (6) after the change of variables $x_1\to\sqrt{2m+1}x_1, x_2\to\sqrt{2l+1}x_2$. When $l=m$ it coincides with the special case of B_2-generalization of the Calogero–Sutherland system. It turned out that the integrability takes place also for any l and m. We prove this by constructing operator D intertwining L and

$$\mathscr{L}=-(2m+1)\frac{\partial^2}{\partial x_1^2}-(2l+1)\frac{\partial^2}{\partial x_2^2}+\frac{l(l+1)(2m+1)}{\sin^2 x_1}+\frac{m(m+1)(2l+1)}{\sin^2 x_2}. \tag{20}$$

The explicit form of the intertwining operator is the following:

$$D=(2m+1)^2\partial_1^2-(2l+1)^2\partial_2^2$$
$$-(m+l+1)\cot\tfrac{1}{2}(x_1+x_2)((2m+1)\partial_1-(2l+1)\partial_2)$$
$$-(m+l+1)\cot\tfrac{1}{2}(x_1-x_2)((2m+1)\partial_1+(2l+1)\partial_2)$$
$$+(m+l+1)^2\cot\tfrac{1}{2}(x_1+x_2)\cot\tfrac{1}{2}(x_1-x_2)$$
$$-\frac{l(l+1)(2m+1)^2}{\sin^2 x_1}+\frac{m(m+1)(2l+1)^2}{\sin^2 x_2}, \tag{21}$$

where ∂_i denotes $\partial/\partial x_i$.

Theorem 5: *The operator D given by formula (21) intertwines the Schrödinger operators (19) and (20):*

$$D\circ\mathscr{L}=L\circ D.$$

Corollary: The operator (19) is integrable for any l,m and algebraically integrable for integer l and m.

The integrals have the form $I_k=D\circ\mathscr{L}_k\circ D^*$ $(k=1,2)$,

$$\mathscr{L}_1=-\partial_1^2+\frac{l(l+1)}{\sin^2 x_1},\quad \mathscr{L}_2=-\partial_2^2+\frac{m(m+1)}{\sin^2 x_2},\quad \mathscr{L}_1+\mathscr{L}_2=\mathscr{L}.$$

For integer l,m the operator (19) can be intertwined with $\mathscr{L}_0=-(2m+1)\partial_1^2-(2l+1)\partial_2^2$ by means of the following operator $S=D\circ S_1\circ S_2$, $L\circ S=S\circ\mathscr{L}_0$ where

$$S_1=(\partial_1-l\ \cot x_1)\circ(\partial_1-(l-1)\cot x_1)\circ\cdots\circ(\partial_1-\cot x_1),$$

$$S_2=(\partial_2-m\ \cot x_2)\circ(\partial_2-(m-1)\cot x_2)\circ\cdots\circ(\partial_2-\cot x_2).$$

In this case we obtain the extra integrals A_1,A_2 by the same method, $A_k=S\circ\partial_k\circ S^*$, which supply the algebraic integrability for L.

In the rational limit we have the operator

$$L=-(2m+1)\partial_1^2-(2l+1)\partial_2^2+\frac{l(l+1)(2m+1)}{x_1^2}$$

$$+\frac{m(m+1)(2l+1)}{x_2^2}+\frac{4(l+m+1)}{(x_1+x_2)^2}+\frac{4(l+m+1)}{(x_1-x_2)^2}.$$

V. CONCLUDING REMARKS

The natural question is what happens with our system in the classical limit. To obtain such a limit we multiply the operator (4) by $\hbar^2$ and replace $i\hbar\,\partial/\partial x_k$ by p_k. We have the following result:

$$\sum_{k=1}^{N+1}p_k^2+\sum_{i<j}^{N}2\hbar^2m(m+1)u(x_i-x_j)+\sum_{i=1}^{N}2\hbar^2(m+1)u(x_i-\sqrt{m}x_{N+1}).$$

Then in the limit $\hbar\to 0$, $m\to\infty, m\hbar\to\mu$ we obtain formally the Hamiltonian of the classical N-body Calogero system with one additional non-interacting particle:

$$H=\sum_{k=1}^{N+1}p_k^2+\sum_{i<j}^{N}2\mu^2u(x_i-x_j).$$

Notice that the formal classical analogue of our problem, i.e., the Hamiltonian system with the Hamiltonian

$$\sum_{k=1}^{N+1}\frac{p_k^2}{2m_k}+\gamma\sum_{i<j}^{N}m_im_j(q_i-q_j)^{-2}$$

seems to be integrable only if all the masses are equal. At least, for the three-particle case with the Hamiltonian

$$H=p_1^2+p_2^2+mp_3^2+\gamma\left((q_1-q_2)^{-2}+\frac{1}{m}(q_1-q_3)^{-2}+\frac{1}{m}(q_2-q_3)^{-2}\right)$$

and $m\neq 1$ one can show that any integral I which is a polynomial of the third order in momenta can be expressed via H and the total momentum $p_1+p_2+p_3$. So only for the case $m=1$ we have independent integral of the third order.

It is interesting to note also that in the limit $\omega\to\infty, m+1\to 0, \omega(m+1)\to$const we formally have the following operator

$$L=-\Delta+\sum_{k<l}^{N}\mu\,\delta(x_k-x_l)-\sum_{k=1}^{N}\mu\,\delta(x_k-ix_{N+1}),$$

or, in the coordinates $q_1=x_1,\ldots,q_N=x_N,q_{N+1}=ix_{N+1}$,

$$L=-\partial_1^2-\ldots-\partial_N^2+\partial_{N+1}^2+\sum_{k<l}^{N}\mu\,\delta(q_k-q_l)-\sum_{k=1}^{N}\mu\,\delta(q_k-q_{N+1}).$$

Another remark is about the relation of our system (1) to Huygens' principle. It can be shown that our deformations of the Calogero operator correspond to new examples of Huygensian equations in Minkowsky' space similar to the Coxeter case considered in Ref. 16. More precisely, the equation

$$(\Box_{M,1}+u(x))\phi=0$$

where $\Box_{M,1}=\partial_t^2-\partial_1^2-\ldots-\partial_M^2$ is the D'Alembert operator in $\mathbf{R}^{M,1}$ and the potential $u=u(x_1,\ldots,x_{N+1})$ has the form

$$u(x)=\sum_{i<j}^{N}\frac{2m(m+1)}{(x_i-x_j)^2}+\sum_{i=1}^{N}\frac{2(m+1)}{(x_i-\sqrt{m}x_{N+1})^2}$$

and satisfies Huygens' principle (in the Hadamard sense) if the parameter m is any integer and the dimension M is any odd number greater than $mN(N-1)+2N+2$. A similar result is true for the deformed B_2-system.

The last remark is about a difference version of our system. Such a version of the Calogero–Moser quantum problem has been discovered by Ruijsenaars (see, e.g., Ref. 17). In the rational case Ruijsenaars's system can be described as a dual system in the ''bispectral sense'' to the trigonometric CMS problem (Sutherland system). One can show (see Ref. 18), that the same relation is valid for our generalizations of the Sutherland system (12) and the following difference operator, which can be considered as a deformation of Ruijsenaars's operator:

$$\mathscr{D}(m)=\sum_{i=1}^{N}\left(1-\frac{\omega}{k_i-\sqrt{m}k_{N+1}+\gamma}\right)\prod_{j\neq i}^{N}\left(1-\frac{m\omega}{k_i-k_j}\right)T_i^{\omega}+m^{-1}\prod_{i=1}^{N}\left(1+\frac{m\omega}{k_i-\sqrt{m}k_{N+1}+\gamma}\right)T_{N+1}^{\sqrt{m}\omega}. \tag{22}$$

Here $\gamma=(1-m)\omega/2$ and $T_i^{\omega}\phi(k_1,\ldots,k_i,\ldots,k_{N+1})=\phi(k_1,\ldots,k_i+\omega,\ldots,k_{N+1})$ is the ω-shift in the ith direction. When $m=1$ $\gamma=0$ and $\mathscr{D}(m)$ coincides with one of the Ruijsenaars operators.[17]

ACKNOWLEDGMENTS

This work was partially supported by the Russian Fund for Fundamental Research (Grant No. 96-01-01404) and INTAS (Grant No. 93-166-ext). One of the authors (O.Ch.) is grateful to the Department of Mathematical Sciences, Loughborough University for the hospitality during his two-month visit and to the Royal Society for the support of this visit. We would like to thank Yuri Berest, John Gibbons and Vladimir Inozemtzev for useful remarks and stimulating discussions.

[1] F. Calogero, ''Solution of the one-dimensional n-body problem with quadratic and/or inversely quadratic pair potential,'' J. Math. Phys. **12**, 419–436 (1971).
[2] O. A. Chalykh and A. P. Veselov, ''Commutative rings of partial differential operators and Lie alrebras,'' Commun. Math. Phys. **126**, 597–611 (1990).
[3] I. M. Krichever, ''Methods of algebraic geometry in the theory of nonlinear equations,'' Russ. Math. Surveys **32**, 180–208 (1977).
[4] A. P. Veselov, K. L. Styrkas, and O. A. Chalykh, ''Algebraic integrability for Schrödinger equation and finite reflection groups,'' Theor. Math. Phys. **94**, 253–275 (1993).
[5] B. Sutherland, ''Exact results for a quantum many-body problem in one dimension,'' Phys. Rev. A **4**, 1372–1390 (1971).
[6] M. A. Olshanetsky and A. M. Perelomov, ''Quantum integrable systems related to Lie algebras,'' Phys. Rep. **94**, 313–404 (1983).
[7] M. Schmidt and A. Veselov, ''Elliptic Calogero–Moser quantum problem and deformations of algebraic surfaces,'' preprint of TU, Berlin (1996).
[8] L. Khodarinova ''On quantum elliptic Calogero-Moser problem,'' Vestnik MGU (to be published).
[9] G. Felder and A. Varchenko, ''Integral representations of solutions of the elliptic Knizhnik–Zamolodchikov–Bernard equation,'' Int. Math. Res. Notices **5**, 221–233 (1995).
[10] G. Felder and A. Varchenko, ''Three formulas for eigenfunctions of integrable Schrödinger operators,'' preprint of UNC (1995).

[11] A. Braverman, P. Etingof, and D. Gaitsgory ''Quantum integrable systems and differential Galois theory,'' preprint (1996).
[12] A. P. Veselov, M. V. Feigin, and O. A. Chalykh, ''New integrable deformations of quantum Calogero–Moser problem,'' Usp. Mat. Nauk **51**, 185–186 (1996).
[13] F. Calogero, C. Marchioro, and O. Ragnisco, ''Exact solution of the classical and quantal one-dimensional many-body problems with the two-body potential $V_a(x)=g^2a^2/sh^2ax$,'' Lett. Nuovo Cimento **13**, 383–387 (1975).
[14] O. A. Chalykh and A. P. Veselov, ''Integrability in the theory of Schrödinger operator and harmonic analysis,'' Commun. Math. Phys. **152**, 29–40 (1993).
[15] O. A. Chalykh, ''Additional integrals for the generalized Calogero–Moser quantum problem,'' Teor. Mat. Fiz. **109**, 28–33 (1996).
[16] Yu. Yu. Berest and A. P. Veselov, ''Hadamard's problem and Coxeter groups: new examples of Huygens' equation,'' Funct. Anal. Appl. **28**, 3–15 (1994).
[17] S. N. M. Ruijsenaars, ''Systems of Calogero–Moser type,'' in *Proceedings of the 1994 Banff summer school 'Particles and fields'*, edited by G. Semenoff and L.Vinet (to be published).
[18] O. A. Chalykh, M. V. Feigin, and A. P. Veselov, ''Multidimensional Baker-Akhiezer functions and Huygens' Principle,'' (unpublished).

Augmented temporal logic formalism for histories-based generalized quantum mechanics

Tulsi Dass[a] and Yogesh Joglekar[b]
Department of Physics, Indian Institute of Technology, Kanpur 208016, India

(Received 22 May 1997; accepted for publication 24 October 1997)

An axiomatic development of dynamics of systems in the framework of histories is given which contains the history versions of classical and traditional quantum mechanics as special cases. We consider theories which admit a quasitemporal structure (a generalization of the concept of time using partial semigroups) and whose ''single time'' propositions have the mathematical structure of a logic; isomorphism of logics at different ''instants of time'' is not assumed. The concept of directed partial semigroup is introduced to incorporate the concept of direction of flow of time in quasitemporal theories. Starting with a few simple axioms, the space of history propositions is explicitly constructed and shown to be an orthoalgebra (as envisaged in the scheme of Isham and Linden [J. Math. Phys. **35**, 5452–5476 (1994)]); its subspace consisting of history filters (''homogeneous histories'') is a meet semilattice. The partial semigroups employed allow semi-infinite irreducible decompositions. © *1998 American Institute of Physics.*
[S0022-2488(98)01402-9]

I. INTRODUCTION

In the traditional formulation of quantum mechanics, all measurements/observations refer to external (classical) observers; such a formalism obviously cannot serve as a basis for quantum cosmology. The consistent histories approach to quantum mechanics[1–7] is an outcome of efforts to evolve a formalism for quantum mechanics applicable to closed systems, in particular, the Universe.

The first paper to employ objects analogous to what are now called histories was probably the one by Houtappel, Van Dam, and Wigner[8] (HVW) who employed, in their treatment of geometric invariance principles, conditional probabilities $\Pi(B|A)$ where $A=(\alpha,r_\alpha;\beta,r_\beta;\ldots;\varepsilon,r_\varepsilon)$ denotes a set of measurements α, β,..., ε (at times $t_\alpha,t_\beta,\ldots,t_\varepsilon$) with outcomes $r_\alpha,\ldots,r_\varepsilon$ and similarly $B=(\zeta,r_\zeta;\ldots;\nu,r_\nu)$; the quantity $\Pi(B|A)$ denotes the probability of the realization of the sequence B, given A. [We have changed HVW's notation $\Pi(A|B)$ to $\Pi(B|A)$ to bring it in line with standard usage in probability theory.] The ordering of the times $t_\alpha,\ldots,t_\varepsilon$, $t_\zeta,\ldots,t_\nu$ is arbitrary. The objects $\Pi(B|A)$ are quite general and can be employed in classical as well as quantum mechanics. The measurements refer to external observers.

The objects presently used in histories-based quantum theory are closely related to the Π functions. By a history one typically means a sequence of measurements/records for a system in given initial state. For a system S in the state represented by the density operator $\rho(t_0)$, the joint probability of the sequence of measurement outcomes

$$(P_1,t_1;P_2,t_2;\ldots;P_n,t_n);t_1<t_2<\ldots<t_n, \tag{1.1}$$

where $t_1>t_0$ and $P_1,\ldots,P_n$ are appropriate Schödinger picture projection operators, is

$$\mathrm{Prob}(P_1,t_1;\ldots;P_n,t_n|\rho(t_0))=\mathrm{Tr}(C_\alpha\rho(t_0)C_\alpha^\dagger). \tag{1.2}$$

Here α denoted the history (1.1) and

[a]Electronic mail: tulsi@iitk.ernet.in
[b]Present address: Physics Department, University of Indiana, Bloomington, IN 47405; electronic mail: yojoglek@indiana.edu

 © 1998 American Institute of Physics

$$C_\alpha = P_n U(t_n, t_{n-1}) \cdots P_2 U(t_2, t_1) P_1 U(t_1, t_0), \tag{1.3}$$

where $U(t', t'') = \exp[-iH(t'-t'')/\hbar]$ is the unitary evolution operator. The main inputs in this derivation are (i) unitary evolution of the undisturbed system during the open time intervals (t_j, t_{j+1}) $(j=0,1,...,n-1)$, (ii) reduction of state vector at times $t_1, t_2, \ldots, t_{n-1}$, and (iii) transition probability formula employed at the final time $t=t_n$.

In the interpretive scheme of Griffith and Omnes,[2,3] the Hilbert space-based mathematical formalism is retained, the reduction postulate is discarded and (1.2) is proposed to be interpreted as the probability for the history (1.1) for the closed system S (no external observers) with the further stipulation that situations involving measurements by external observers are to be dealt with by treating the closed system (system+observer) as above.

Some restrictions need to be imposed on admissible histories if the formula (1.2) is to be used as a classical probability. This is because the classical addition of probabilities cannot hold in the presence of quantum mechanical interference. The desired ''consistency condition'' (or ''decoherence condition'') between two histories α and β [with $\beta = (P'_1, t'_1; \ldots; P'_m, t'_m)$, say] turns out to be

$$\mathrm{Re}[d(\alpha, \beta)] = 0, \tag{1.4}$$

where the so-called decoherence functional $d(\alpha, \beta)$ is given by

$$d(\alpha, \beta) = \mathrm{Tr}[C_\alpha \rho(t_0) C_\beta^+]. \tag{1.5}$$

Note that

$$\mathrm{Prob}(\alpha) = d(\alpha, \alpha). \tag{1.6}$$

Gell-Mann and Hartle[4–7] (GH) have pursued a program for evolving a framework for the quantum dynamics of the Universe as a whole. Within that framework, they aim at understanding the origin of the quasiclassical domain of familiar experience taking as basic inputs the action of elementary particles and the initial condition of the Universe. In the ''generalized quantum mechanics'' (of a closed system) evolved by them, the basic ingredients are a set of fine-grained histories [usually taken to be the paths $q^i(t)$ in configuration space], allowed coarse grainings, and a space of decoherence functional $d(\alpha, \beta)$ (complex-valued functionals defined on pairs of histories in coarse-grained sets) for which four conditions [hermiticity, positivity, additivity, and normalization; see Eqs. (3.62a)–(3.62d)], which are satisfied by the $d(.,.)$ of Eq. (1.5), are *postulated*. The decoherence functionals serve, as above, to define decoherence conditions and probabilities of decoherent histories. No reference is made to an underlying Hilbert space.

Isham[9] (see also Isham and Linden[10]) has proposed that the proper mathematical framework for the axiomatic development of a GH type scheme is that of temporal quantum logic, a multitime generalization of traditional quantum logic.[11–15] In his scheme a partial semigroup (for definition, see Sec. II) structure was introduced in a pair of spaces [a space $\mathscr{T}$ of temporal supports (replacing the space of time sequences $t_1, \ldots, t_n$ employed in the traditional description of histories) and a space $\mathscr{U}$ of history filters (homogeneous histories); our notations are slightly different from those of Refs. 9 and 10] to generalize the concept of time and to describe the composition of histories in appropriate temporal relation to form larger histories. A homomorphism σ of $\mathscr{U}$ onto $\mathscr{T}$ associates, with an element α of $\mathscr{U}$, the element $\sigma(\alpha)$ of $\mathscr{T}$ representing the ''sequence of times'' at which the propositions of the history α are specified. The space $\mathscr{U}$ was assumed to have, besides the partial semigroup structure, the structure of a meet semilattice[16] with the operations of partial order $\leqslant$ (coarse graining) and a meet operation $\wedge$ (simultaneous realization of two histories). To accommodate the operation of negation of a history, the space $\mathscr{U}$ was proposed to be embedded in a larger space Ω (denoted as $\mathscr{U}.\mathscr{P}$ in Refs. 9 and 10) of history propositions (''inhomogeneous'' histories); the structure of an orthocomplemented lattice was envisaged (but not derived) for Ω. The decoherence functionals (objects associating complex numbers with pairs of elements of Ω) were assumed to have the same properties as in the GH scheme. A formalism of this type has the generality required to accommodate in it history versions of field theories on general space times (which may not admit foliation in terms of spacelike surfaces).

In Ref. 10, a more general scheme was proposed in which the basic objects were the space Ω of history propositions and the space $\mathscr{D}$ of decoherence functionals. The space Ω (whose elements will be denoted as $\underline{\alpha}$, $\underline{\beta}$,...) was postulated to have the structure of an orthoalgebra,[17] a structure incorporating partial order$\leqslant$(coarse graining), disjointness$\perp$(mutual exclusion), and disjoint joint operation $\oplus$ ($\alpha\oplus\beta$, defined only when $\alpha\perp\beta$, is to mean ''α or β'') with the stipulation that $\alpha\leqslant\beta$ if and only if there exists a (unique) $\gamma\varepsilon\,\Omega$ such that $\beta=\alpha\oplus\gamma$. It has two distinguished elements 0 (the null history proposition which is impossible to realize) and 1 (the unit history proposition which is sure to be realized); one has $0\leqslant\alpha\leqslant 1$ for all $\alpha\varepsilon\,\Omega$. For every $\alpha\varepsilon\,\Omega$, the relation $\alpha\leqslant 1$ implies the existence of a unique element$\neg\alpha$ (the negation of α) of Ω such that $\alpha\oplus\neg\alpha=1$. History theories with quasitemporal structure (i.e., those covered by an Isham type scheme) constitute a subclass of theories covered by this more general scheme.

While histories, in principle, contain information about the structural features and dynamics of a system, the Isham–Linden formalism is not adequately equipped to bring it out in an autonomous framework. In the present work we present a more detailed formalism for quasitemporal theories with a view to fulfill this need. Our framework is defined by three axioms A_1–A_3. The first axiom A_1 introduces a quasitemporal structure in terms of a triple $(\mathscr{U},\mathscr{T},\sigma)$ as in the Isham scheme. The second is a causality axiom which forbids ''closed time loops.'' Our main new axiom is A_3 in which the spaces $\mathscr{U}_\tau=\sigma^{-1}(\tau)$ for nuclear elements in $\mathscr{T}$ (''points of time''), the spaces of ''single time propositions,'' are assumed to have the structure of a logic as defined in Ref. 14, a natural assumption to make in view of the fact that this structure was arrived at (in the 1960s) by very general considerations. (In any case, it is worth studying history theories with this structure assumed for the $\mathscr{U}_\tau$'s and generalizing the concept of time as described above.) We, however, do not assume isomorphism of $\mathscr{U}_\tau$'s for different τ's. This generality, we feel, must be kept in order that the formalism is applicable to systems whose structural properties undergo qualitative change with temporal evolution. Using these three axioms, the space Ω of history propositions (''inhomogeneous histories'') is explicitly constructed and shown to be an orthoalgebra as envisaged in the Isham–Linden scheme. The subspace $\widetilde{\mathscr{U}}$ of Ω consisting of ''homogeneous histories'' (obtained from $\mathscr{U}$ by removing some redundancies) is shown to be a meet-semilattice.

An important subclass of the theories covered by these axioms is the one in which $\mathscr{U}_\tau=\mathscr{P}(\mathscr{H}_\tau)$, the lattice of projection operators in a separable Hilbert space $\mathscr{H}_\tau$, the Hilbert spaces $\mathscr{H}_\tau$ for different τ's being not necessarily naturally isomorphic. We shall refer to these theories as Hilbert space-based theories. Traditional quantum mechanics (in history version) obviously belongs to this class of theories.

The space $\mathscr{D}$ of decoherence functionals in both the GH and Isham–Linden schemes is not properly integrated with the basic framework; it is, for example, not made clear as to what it is that distinguishes different elements of $\mathscr{D}$ and when is a particular element of $\mathscr{D}$ relevant. In traditional quantum mechanics, a decoherence functional is determined by the evolution operator and the initial state [see Eq. (1.5)]. In the Isham–Linden formalism there is no explicitly defined concept of state or of evolution; information about both of these is contained in the decoherence functionals. It is, however, not clear how to bring out and use this information.

In our formalism, the $\mathscr{U}_\tau$'s being logics, both the concepts can be introduced in the conventional manner. Our axiom A_4 introduces temporal evolution in terms of an appropriate class of mappings between the $\mathscr{U}_\tau$'s. The last axiom A_5 associates a decoherence function with a given law of temporal evolution and a given initial state. The formalism would be complete if an explicit expression for the decoherence functional [analogous to that in Eq. (1.5) in the case of traditional quantum mechanics] is given. We are able to do it only for the Hilbert space-based theories. The decoherence functional proposed for these theories [Eq. (3.77)] is an adaptation from Eq. (4.19) of Ref. 10; it is, expectedly, a straightforward generalization of the expression in Eq. (1.5) for traditional quantum mechanics.

An innovation introduced in the present work is that, in our axiomatic scheme, a general history is allowed to have countably semi-infinite temporal support; this means that (homogeneous) histories are generally taken to involve single ''time'' propositions at ''times'' $\tau_1,\tau_2,\ldots$ or $\ldots\tau_{-1},\tau_0$. Similarly, an inhomogeneous history is generally taken to include a countable infinity of mutually exclusive homogeneous histories. This, we believe, should be helpful in addressing questions relating to, for example, the long time behavior of the Universe (under suitable hypotheses) besides facilitating analytical work in the present setting.

The rest of the paper is organized as follows. Section II is devoted to the treatment of partial semigroups. Here, apart from recalling a few definitions from Ref. 9, an improved treatment of the so-called unit and absorbing elements in partial semigroups is given, some new definitions are introduced and a useful lemma is proved. In the last section an axiomatic development of our formalism is presented. (We call it the augmented temporal logic formalism.) After developing the formalism in the form indicated above in Secs. III A–III D, Hilbert space-based theories are treated in Sec. III E and (history version of) classical mechanics in Sec. III F.

II. PARTIAL SEMIGROUPS

In the treatment of histories in traditional quantum theory, the time sequences $(t_1,\ldots,t_n)$ are employed only for book keeping; the properties of time t as a real variable are not used. It was Isham's[9] observation that the relevant mathematical structure employed in the book keeping of histories is that of a partial semigroup.

We briefly recall a few definitions. A semigroup is a set with an associative composition rule (product) defined on it. A partial semigroup (psg) is a set on which a similar composition rule is defined but not necessarily for all pairs of elements. Denoting the composition in a psg $\mathscr{K}$ (with elements $s,t,u,\ldots$) by $\circ$ we have

$$(s\circ t)\circ u = s\circ(t\circ u) \tag{2.1}$$

whenever the two sides are defined.

A homomorphism of a psg $\mathscr{K}$ into another psg $\mathscr{K}'$ is a mapping $\sigma:\mathscr{K}\to\mathscr{K}'$ such that, for all $s,t\varepsilon\mathscr{K}$ with $s\circ t$ defined, we have $\sigma(s)\circ\sigma(t)$ defined and

$$\sigma(s\circ t)=\sigma(s)\circ\sigma(t). \tag{2.2}$$

If σ is bijective, then it is called an isomorphism,

There are two psg's (we call them $\mathscr{K}_1$ and $\mathscr{K}_2$) implicitly involved in the description of histories of the form (1.1). We have

$$\mathscr{K}_1=\{\text{ finite ordered subsets of } R,\ \text{the real line}\}.$$

A typical element of $\mathscr{K}_1$ is

$$t=\{t_1,t_2,\ldots,t_n\};\quad t_1<t_2<\cdots<t_n. \tag{2.3}$$

If $s=\{s_1,s_2,\ldots,s_m\}\ \varepsilon\mathscr{K}_1$ such that $s_m<t_1$, we have $s\circ t$ defined and

$$s\circ t=\{s_1,s_2,\ldots,s_m,t_1,\ldots,t_n\}=s\cup t. \tag{2.4}$$

It is useful to adopt the convention

$$\{t_1\}\circ\{t_1\}=\{t_1\}. \tag{2.5}$$

With this convention we have $s\circ t$ defined for $s_m\leqslant t_1$.

The other psg $\mathscr{K}_2$ has, as its elements, histories of the form (1.1). With

$$\begin{aligned}\alpha&=\{(P_1,s_1),\ (P_2,s_2),\ldots,(P_m,s_m)\};\ s_1<s_2<\cdots<s_m,\\ \beta&=\{(P'_1,t_1),\ (P'_2,t_2),\ldots,(P'_n,t_n)\};\ t_1<t_2<\cdots<t_n,\end{aligned} \tag{2.6}$$

and $s_m<t_1$, we have $\alpha\circ\beta$ defined and given by

$$\alpha\circ\beta=\{(P_1,s_1),\ (P_2,s_2),\ldots,(P_m,s_m),\ (P'_1,t_1),\ldots,(P'_n,t_n)\}. \tag{2.7}$$

There is a homomorphism σ from $\mathscr{K}_2$ onto $\mathscr{K}_1$ given by

$$\sigma(\alpha)=s,\ \ \sigma(\beta)=t,\ \ \sigma(\alpha\circ\beta)=s\circ t. \tag{2.8}$$

The triple $(\mathscr{K}_2,\mathscr{K}_1,\sigma)$ defines what Isham[9] calls a quasitemporal structure (a pair of psg's with a homomorphism from one onto the other).

A prototype of psg's employed in our work later is an ''elder brother'' of $\mathscr{K}_1$, namely

$\mathscr{K}_3$ = {ordered subsets of R which are at most countably semi-infinite}.

Typical elements of $\mathscr{K}_3$ are of the form

$$s=\{s_1,s_2,\ldots,s_m\} \quad \text{(finite)},$$

$$s=\{s_1,s_2,\ldots,\} \quad \text{(right infinite)},$$

$$s=\{\ldots,s_{-1},s_0\} \quad \text{(left infinite)}.$$

The product $s\circ t$ is defined if at least one of s and t is finite, s is right finite, t is left finite, and the largest entry in s is not greater than the smallest entry in t; then $s\circ t=s\cup t$.

A unit element in a psg $\mathscr{K}$ is an element e such that

$$e\circ s=s\circ e=s \quad \text{for all } s\varepsilon\mathscr{K}. \tag{2.9}$$

A psg may or may not have a unit element; when it has, it is unique. [*Proof:* $e'=e'\circ e=e$.]

The empty set [when formally considered as an element of the psg in question with composition rule analogous to that for e in Eq. (2.9)] serves as the unit element in $\mathscr{K}_1$ and $\mathscr{K}_3$ above. In $\mathscr{K}_2$, there is no unit element. However, if we introduce an equivalence relation in $\mathscr{K}_2$ such that histories differing by any number of entries of form (p_j,t_j) where $P_j=I$, the identity operator, are equivalent then, denoting the equivalence class of α as $[\alpha]$, the set $\tilde{\mathscr{K}}_2$ of equivalence classes in $\mathscr{K}_2$ has a psg structure defined by

$$[\alpha]\circ[\beta]=[\alpha\circ\beta]. \tag{2.10}$$

All histories having only identity operators in their entries from an equivalence class which serves as the unit element in $\tilde{\mathscr{K}}_2$.

An absorbing element in a psg $\mathscr{K}$ is an element a such that

$$a\circ s=s\circ a=a \quad \text{for all } s\varepsilon\mathscr{K}. \tag{2.11}$$

Such an element, if it exists, is unique. [*Proof:* $a'=a'\circ a=a$.]

The psg's $\mathscr{K}_1$ and $\mathscr{K}_3$ have no absorbing elements. The psg's $\mathscr{K}_2$ and $\tilde{\mathscr{K}}_2$ also have no absorbing elements but, if we allow the use of the null projection operator in defining histories and strengthen the equivalence relation defining $\tilde{\mathscr{K}}_2$ by the stipulation that all histories having $P_j=0$ as one or more entries are to be treated as equivalent, then the resulting equivalence class will serve as the absorbing element in $\tilde{\mathscr{K}}_2$.

In this connection we note here that, in the so-called HPO (history projection operator) version[9] of histories in standard quantum theory (see the Appendix for details), the space of homogeneous projection operators of the form (A4) forms a psg isomorphic to the final $\tilde{\mathscr{K}}_2$ above with the unit operator $\tilde{I}$ and the null operator $\tilde{0}$ on the tensor product space $\tilde{\mathscr{H}}$ serving, respectively, as the unit and absorbing elements.

Let $\mathscr{K}$ be a psg with unit and absorbing elements e, a as above and σ a homomorpism of $\mathscr{K}$ into $\mathscr{K}'$. We must have, for all $s\in\mathscr{K}$,

$$\sigma(e)\circ\sigma(s)=\sigma(s)\circ\sigma(e)=\sigma(e\circ s)=\sigma(s),$$

$$\sigma(a)\circ\sigma(s)=\sigma(s)\circ\sigma(a)=\sigma(a\circ s)=\sigma(a).$$

It follows that

(i) if $\mathscr{K}$ has a unit element e and/or an absorbing element a and σ is onto, then $\mathscr{K}'$ must correspondingly have a unit element e' and/or an absorbing element a' and we must have

$$\sigma(e)=e', \quad \sigma(a)=a'. \tag{2.12}$$

(ii) More generally, $\sigma(e)$ and $\sigma(a)$ act, respectively, as unit and absorbing elements in the homomorphic image $\sigma(\mathscr{K})$ of $\mathscr{K}$ in $\mathscr{K}'$.

In a psg, elements other than the unit and the absorbing element will be called typical.

A psg $\mathscr{K}$ will be called directed if, for any two different typical elements s, t in $\mathscr{K}$, if $s\circ t$ is defined, then $t\circ s$ is not defined. Note that, in this definition, the composition $s\circ s$ is allowed. All the psg's considered above are directed. This property appears to be appropriate for defining the direction of flow of time in theories with quasitemporal structure.

A useful concept relating to partial semigroups which serves to introduce the concept of ''point of time'' in the framework of histories-based theories is that of irreducible decompositions of typical elements of a psg. This means expressing a typical element of a psg as a composition of ''nuclear'' elements which cannot be expressed as nontrivial compositions. For example, general elements of $\mathscr{K}_1$, $\mathscr{K}_3$, and $\tilde{\mathscr{K}}_2$ have irreducible decompositions of the form [we use the HPO notation (A4) for $\tilde{\mathscr{K}}_2$]

$$\mathscr{K}_1: \quad \{t_1,t_2,\ldots,t_n\}=\{t_1\}\circ\{t_2\}\circ\cdots\circ\{t_n\}, \tag{2.13a}$$

$$\mathscr{K}_3: \quad \{t_1,t_2,\ldots\}=\{t_1\}\circ\{t_2\}\circ\cdots, \tag{2.13b}$$

$$\tilde{\mathscr{K}}_2: \quad \overline{\alpha}=\alpha_{t_1}\circ\alpha_{t_2}\circ\cdots\circ\alpha_{t_n}. \tag{2.13c}$$

In $\mathscr{K}_1$ and $\mathscr{K}_3$, the nuclear elements correspond to points of time and, in $\tilde{\mathscr{K}}_2$, to single time ''history'' propositions. The set of nuclear elements of a psg $\mathscr{K}$ will be denoted as $\mathscr{N}(\mathscr{K})$. For all the psg's introduced above, clearly $\mathscr{N}(\mathscr{K})=R$.

The term ''nontrivial'' employed in the definition of a nuclear element refers to the possibility of trivial compositions of the form $s\circ e=e\circ s=s$ or of the form (2.5).

Not all psg's admit irreducible decompositions. For example analogs of $\mathscr{K}_1$ and $\mathscr{K}_3$ with intervals as their elements do not admit such decompositions.

We shall employ psg's of the type of $\mathscr{K}_3$ above which admit semi-infinite irreducible decompositions and in which conventions of the type of Eq. (2.5) are incorporated. We shall call such psg's special. To formalize it, we shall call a psg $\mathscr{K}$ (which may or may not contain unit and/or absorbing elements) special if there exists a subset $\mathscr{N}(\mathscr{K})$ of typical elements of $\mathscr{K}$ (whose elements will be called nuclear elements of $\mathscr{K}$) such that

(i) for every $s\varepsilon\mathscr{N}(\mathscr{K})$, we have $s\circ s=s$;
(ii) every non-nuclear typical element s of $\mathscr{K}$ admits a unique irreducible decomposition into a composition of elements of $\mathscr{N}(\mathscr{K})$ which is at most semi-infinite, i.e.,

$$s=s_1\circ\cdots\circ s_n \quad \text{or} \quad s=s_1\circ s_2\circ\cdots \quad \text{or} \quad s=\cdots\circ s_{-1}\circ s_0,$$

where $s_j\varepsilon\mathscr{N}(\mathscr{K})$. The uniqueness of these decompositions must be understood as modulo trivial compositions of the form $s_k\circ s_k=s_k$.

A subset A (which is at most countable and whose elements are listed in some definite order) of a psg $\mathscr{K}$ will be called orientable if there exists a bijection π of A such that, in $\pi(A)$, the composition of every pair of consecutive elements is defined. An ordering of elements of A permitting such a composition [such as that given by $\pi(A)$ above] will be called an orientation of A. In a semigroup every countable subset is orientable and every ordering of such a subset gives an orientation. It is clear from the example of semigroups that, in a psg, orientations of orientable subsets need not be unique.

One might expect that, in a directed psg, the orientations of orientable subsets are unique. This need not be true. As a counterexample, consider the three element directed psg $\mathscr{K}=\{a,b,c\}$ with $a\circ b$, $b\circ c$, and $c\circ a$ defined. One can have $a\circ b\circ c$, $b\circ c\circ a$, and $c\circ a\circ b$ as possible orientations of $\mathscr{K}$.

The following lemma will be used in Sec. III B.

Lemma (2.1): Let $\mathscr{K}$ be a special psg which does not admit elements expressible as nontrivial closed chains of the form $s\circ\cdots\circ s$ for any $s\varepsilon\mathscr{N}(\mathscr{K})$. [Such a psg is directed and its orientable

subsets have unique orientations; see propositions (3.2) and (3.3) below.]. Let $E_1, E_2,\ldots,E_n$ be orientable subsets of $\mathcal{N}(\mathcal{K})$ such that $E_i \cup E_j$ is orientable for every pair (i,j). Then $E = \cup_i E_i$ is orientable.

Proof: Let $E_i = \{s_k^{(i)}, k \varepsilon A_i\}, \quad i = 1,...,n$, where A_i are some countable index sets. To show that E is orientable, we must show that we can construct, by composition of all elements of E, a single element, say s_E, of $\mathcal{K}$.

Let us pick up some element of E, say $s_{i_1}^{(1)}$. We must show that, starting with $s_{i_1}^{(1)}$, one can go on adjoining element of E on the right and/or on the left (to construct an element of $\mathcal{K}$) untill all members of E are used up. An element s in E will be called a potential successor of $s_{i_1}^{(1)}$ if there exists an orientable subset of E containing both $s_{i_1}^{(1)}$ and s such that its (unique) orientation is given by an ordering of the form $\{\ldots, s_{i_1}^{(1)}, \ldots, s, \ldots\}$; similarly we define a potential predecessor. Since $E_i \cup E_j$ is orientable for every pair (i,j), every element in E other than $s_{i_1}^{(1)}$ is either a potential successor or a potential predecessor of $s_{i_1}^{(1)}$ (but not both because $\mathcal{K}$ does not admit nontrivial closed chains).

It is adequate to show how the collection F of potential successors of $s_{i_1}^{(1)}$, if nonempty, is to be used up. For this, it is adequate to show that, in F (assumed nonempty), there is an ''earliest'' element s^1 (by this we mean that every other element in F is a potential successor of s^1).

In every E_i the collection F_i of potential successors of $s_{i_1}^{(1)}$ is either empty or has an earliest element. In E_1, it is $s_{i_1+1}^{(1)}$ (if it exists); in $E_j (j \neq 1)$, it is the first element of E_j on the right of $s_{i_1}^{(1)}$ in the ordering showing the orientation of $E_1 \cup E_j$. The possible successors of $s_{i_1}^{(1)}$ (in the construction of s_E) are, therefore, in the (unique) set

$$[F] = \{s_{i_1+1}^{(1)}, s_{j_1}^{(2)}, \ldots, s_{k_1}^{(n)}\}, \text{ say}$$

where some of the entries on the right may be vacuous. By assumption, $[F]$ is nonempty. If it has only one element, it is the unique successor of $s_{i_1}^{(1)}$. If there are more than one, note that every pair in $[F]$ has a definite orientation which implies that $[F]$ has an earliest element. This element is the successor of $s_{i_1}^{(1)}$. Similarly for the predecessor. Q.E.D.

III. THE AUGMENTED TEMPORAL LOGIC FORMALISM

In this section we shall present an axiomatic development of the dynamics of (closed) systems in the framework of histories. This, as we shall see, contains the history versions of classical mechanics and traditional quantum mechanics as special cases.

A. The basic ingredients

We shall denote our axioms as $A_1,\ldots,A_5$. In this section we shall take up the first three of them which define the basic structure.

1. A_1 (quasitemporal structure axiom)

Associated with every dynamical system is a *history system* which, by definition, is a triple $(\mathcal{U}, \mathcal{T}, \sigma)$ where $\mathcal{U}$ (called ''the space of history filters'') and $\mathcal{T}$ (called ''the space of temporal supports'') are special psg's and σ is a homomorphism of $\mathcal{U}$ onto $\mathcal{T}$ such that

$$\sigma[\mathcal{N}(\mathcal{U})] = \mathcal{N}(\mathcal{T}). \tag{3.1}$$

Elements of $\mathcal{U}$ (to be called history filters) will be denoted as $\alpha, \beta, \ldots$ and those of $\mathcal{T}$ (to be called temporal supports) as $\tau, \tau', \ldots$. If $\alpha \circ \beta$ and $\tau \circ \tau'$ are defined, we shall write $\alpha \lhd \beta$ (which means α precedes β) and $\tau \lhd \tau'$ (τ precedes τ'). We shall often refer to ''temporally ordered'' sequences of the form $\alpha \lhd \beta \lhd \ldots \lhd \gamma$ and $\tau \lhd \tau' \lhd \ldots \lhd \tau''$ as chains.

2. A_2 (causality axiom)

If $\alpha, \beta, ..., \gamma \varepsilon \mathcal{N}(\mathcal{U})$ are such that

$$\alpha \lhd \beta \lhd \ldots \lhd \gamma \quad \text{with} \quad \sigma(\alpha) = \sigma(\gamma) \tag{3.2}$$

then we must have

$$\alpha = \beta = \cdots = \gamma. \tag{3.3}$$

In essence, this axiom forbids histories corresponding to ''closed time loops.'' This is one of the weakest formulations of causality.

We collect, in the first proposition below, a few auxiliary results (easily derived from the above two axioms).

Proposition (3.1): (a) If $\tau, \tau', \ldots, \tau'' \varepsilon \mathcal{N}(\mathcal{T})$ are such that

$$\tau \lhd \tau' \lhd \cdots \lhd \tau'' \lhd \tau$$

then we must have

$$\tau = \tau' = \cdots = \tau''.$$

(b) If τ', $\tau'' \varepsilon \mathcal{N}(\mathcal{T})$ are such that (i) and/or (ii) below holds:

(i) $\tau' \lhd \tau$ for any $\tau \varepsilon \mathcal{N}(\mathcal{T})$ if and only if $\tau'' \lhd \tau$,
(ii) $\tau \lhd \tau'$ for any $\tau \varepsilon \mathcal{N}(\mathcal{T})$ if and only if $\tau \lhd \tau''$, then we must have $\tau' = \tau''$

(c) Analogue of (b) for $\mathcal{N}(\mathcal{U})$ also holds.
(d) If $\tau \varepsilon \mathcal{T}$ is non-nuclear, then $\tau \circ \tau$ is not allowed.
(e) If $\alpha \varepsilon \mathcal{U}$ is non-nuclear then $\alpha \circ \alpha$ is not allowed.

Proposition (3.2): The partial semigroups $\mathcal{T}$ and $\mathcal{U}$ are directed.

Proof: We prove it only for $\mathcal{T}$. Suppose it is not directed. Then there are typical elements $\tau, \tau' \varepsilon \mathcal{T}$, $\tau \neq \tau'$ such that $\tau \lhd \tau'$ and $\tau' \lhd \tau$. It τ, τ' are both nuclear then the chain $\tau \lhd \tau' \lhd \tau$ gives $\tau = \tau'$ (contradiction). If one or both of τ, τ' are non-nuclear, $\tau \circ \tau'$ and $\tau' \circ \tau$ both can be defined only if both have finite irreducible decompositions, say

$$\tau = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_m, \quad \tau' = \tau_1' \circ \tau_2' \circ \cdots \circ \tau_n' (n \geq 1, m > 1).$$

The element $\tau \circ \tau' \circ \tau$ contains, in its irreducible decomposition, the element

$$\tau_1 \circ \cdots \circ \tau_m \circ \tau_1' \circ \cdots \circ \tau_n' \circ \tau_1$$

giving $\tau_1 = \tau_2 = \cdots = \tau_m = \tau_1' = \tau_2' = \cdots = \tau_n'$ which, in turn, given $\tau' = \tau = \tau_1 \varepsilon \mathcal{N}(\mathcal{T})$ (contradiction). Q.E.D.

Proposition (3.3): Orientations of orientable subsets of $\mathcal{N}(\mathcal{T})$ and $\mathcal{N}(\mathcal{U})$ are unique.

Proof: We prove it for $\mathcal{N}(\mathcal{T})$. Suppose the proposition is not true. Then there exists an orientable subset $A = \{\tau_1, \tau_2, \ldots\}$ of $\mathcal{N}(\mathcal{T})$ having (at least) two different orientations. This implies that, for some pair (j,k), $j \neq k$, we have τ_j appearing earlier than τ_k in one orientation and later in the other so that we can form the two chains

$$\tau_j \lhd \cdots \lhd \tau_k \quad \text{and} \quad \tau_k \lhd \cdots \lhd \tau_j$$

giving, in turn, the chain

$$\tau_j \lhd \cdots \lhd \tau_k \lhd \cdots \lhd \tau_j$$

which implies $\tau_j = \tau_k$ (contradiction). Q.E.D.

We next introduce an axiom relating to the spaces $\mathcal{U}_\tau = \sigma^{-1}(\tau)$ $[\tau \varepsilon \mathcal{N}(\mathcal{T})]$ of ''single-time propositions;'' we give them the same structure as one has for the analogous objects in the

traditional proposition calculus for quantum mechanics. Formulations of proposition calculus by different authors differ in some details; we find the one of Varadarajan[14] best suited for our needs.

3. A_3 (logic structure axiom)

Every $\mathscr{U}_\tau$ (for τ nuclear) has the structure of a *logic* as defined in Ref. 14. This means that the following operations are defined in $\mathscr{U}_\tau$ with the indicated interpretations. [Some (harmless) redundancy in the following description can be ignored.]

(a) Partial order ($\leqslant$). The relation $\alpha \leqslant \beta$ for $\alpha, \beta \in \mathscr{U}_\tau$ should mean that the proposition α implies β or α is finer than β.

(b) In each $\mathscr{U}_\tau$, there is a unique null proposition 0_τ (the absurd proposition which can never be true) and a unique unit proposition 1_τ (the trivial proposition which is always true) such that

$$0_\tau \leqslant \alpha \leqslant 1_\tau \quad \text{for all} \quad \alpha \in \mathscr{U}_\tau. \tag{3.4}$$

(c) For each subset $A=\{\alpha_1, \alpha_2, \ldots\}$ of $\mathscr{U}_\tau$ which is at most countable, the greatest lower bound (glb) of A denoted as $\bigwedge_i \alpha_i$ (it is the coarsest proposition finer than all propositions in A) and the least upper bound (lub) of A denoted as $\bigvee_i \alpha_i$ (it is the finest proposition coarser than all propositions in A) exist in $\mathscr{U}_\tau$. For $\alpha, \beta \in \mathscr{U}_\tau$ their glb (or ''meet'') $\alpha \wedge \beta$ stands for ''α and β'' and lub (or ''join'') $\alpha \vee \beta$ for ''α or β.''

(d) Negation ($\neg$) or orthocomplementation. It is a bijection $\neg: \mathscr{U}_\tau \to \mathscr{U}_\tau$ such that

$$\text{(i)} \quad \neg(\neg\alpha) = \alpha \quad \text{for all} \quad \alpha \in \mathscr{U}_\tau, \tag{3.5a}$$

$$\text{(ii)} \quad \alpha \leqslant \beta \quad \text{if and only if} \quad \neg\beta \leqslant \neg\alpha, \tag{3.5b}$$

$$\text{(iii)} \quad \alpha \wedge (\neg\alpha) = 0_\tau, \quad \text{for all} \quad \alpha \in \mathscr{U}_\tau, \tag{3.5c}$$

$$\text{(iv)} \quad \alpha \vee (\neg\alpha) = 1_\tau \quad \text{for all} \quad \alpha \in \mathscr{U}_\tau. \tag{3.5d}$$

The proposition $\neg\alpha$ is to be interpreted as the negation of α, i.e., α is realized if and only if $\neg\alpha$ is not realized. We have

$$\neg(0_\tau) = 1_\tau, \quad \neg(1_\tau) = 0_\tau, \tag{3.6}$$

(e) If $\alpha, \beta \in \mathscr{U}_\tau$, then $\alpha \leqslant \beta$ if and only if there exists an element γ in $\mathscr{U}_\tau$ such that

$$\gamma \leqslant (\neg\alpha) \quad \text{and} \quad \gamma \vee \alpha = \beta. \tag{3.7}$$

In fact,[11] γ is unique and given by

$$\gamma = (\neg\alpha) \wedge \beta; \tag{3.8}$$

it is often denoted as $\beta - \alpha$. Denoting the disjoint join operation in (3.7) by $\oplus$, we have

$$\beta = \alpha \oplus [(\neg\alpha) \wedge \beta]. \tag{3.9}$$

We do not assume isomorphism of $\mathscr{U}_\tau$'s (as logics) for different τ's; this is in contrast to the situation in the proposition calculus for traditional quantum mechanics where one assumes the logics at different instants of time to be isomorphic.

Given $\alpha, \beta \in \mathscr{U}_\tau$, we shall say that $\alpha \perp \beta$ if $\alpha \leqslant \neg\beta$ (which is equivalent to $\beta \leqslant \neg\alpha$). This is a disjointness relation implying that if α is realized, then β is ruled out and vice versa. It implies the relation $\alpha \wedge \beta = 0_\tau$ (simultaneous realization of α and β is impossible).

Recalling Eq. (3.1) and the definition of $\mathscr{U}_\tau$, we have

$$\mathscr{N}(\mathscr{U}) = \bigcup_{\tau \in \mathscr{N}(\mathscr{T})} \mathscr{U}_\tau. \tag{3.10}$$

For a typical $\alpha \in \mathscr{U}$, we define supp(α) (called the temporal support of α) as the unique collection of elements of $\mathscr{N}(\mathscr{T})$ appearing in the irreducible decomposition of $\sigma(\alpha)$. Clearly supp(α) is an

orientable subset of $\mathcal{N}(\mathcal{T})$ with the unique orientation provided by the unique irreducible decomposition of $\sigma(\alpha)$. For any $\tau\varepsilon\,\mathrm{supp}(\alpha)$, the nuclear element in the irreducible decomposition of α projecting to τ under σ will be denoted as α_τ.

An $\alpha\varepsilon\mathcal{U}$ will be called

(i) a null history filter if $\alpha_\tau = 0_\tau$ for at least one $\tau\,\varepsilon\,\mathrm{supp}(\alpha)$;
(ii) a unit (or trivial) history filter if $\alpha_\tau = 1_\tau$ for all $\tau\,\varepsilon\,\mathrm{supp}(\alpha)$.

All the null history filters are obviously physically equivalent (they represent absurd histories) as are all the unit history filters. It is desirable to remove the redundancy implied by the presence of more than one of each type. We do this by constructing from $\mathcal{U}$, a psg $\tilde{\mathcal{U}}$ imitating the construction of $\tilde{\mathcal{K}}_2$ from $\mathcal{K}_2$ in Sec. II. In $\tilde{\mathcal{U}}$ all the null history filters form a single null element $\tilde{0}$ and all the unit history filters from a single unit element $\tilde{1}$. A general element $\tilde{\alpha}$ of $\tilde{\mathcal{U}}$ is an equivalence class of elements $\{\alpha\}$ of $\mathcal{U}$ differing from each other by superfluous 1_τ's in their irreducible decompositions. The irreducible decomposition of a general element of $\mathcal{U}$ defines, in an obvious manner, the same for the corresponding element of $\tilde{\mathcal{U}}$. It is now easily seen that $\tilde{\mathcal{U}}$ is a directed special psg with unit element $\tilde{1}$ and absorbing element $\tilde{0}$.

We now construct a psg $\tilde{\mathcal{T}}$ which is in the same relation to $\tilde{\mathcal{U}}$ as $\mathcal{T}$ is to $\mathcal{U}$. The typical elements of $\tilde{\mathcal{T}}$ and the homomorphism $\tilde{\sigma}:\tilde{\mathcal{U}}\to\tilde{\mathcal{T}}$ (restricted to typical elements) can be obtained in a single step by defining $\tilde{\tau}=\tilde{\sigma}(\tilde{\alpha})$ to be the reduced object obtained from $\tau=\sigma(\alpha)$ (for any representative α of $\tilde{\alpha}$) by deleting the images under σ of the redundant 1_τ's in the irreducible decomposition of α. We augment $\tilde{\mathcal{T}}$ by defining its unit element $\tilde{e}$ and absorbing element $\tilde{a}$ as

$$\tilde{e}=\phi(\text{the empty set}), \tag{3.11a}$$

$$\tilde{a}=\mathcal{N}(\mathcal{T}) \tag{3.11b}$$

and augmenting the psg structure of $\tilde{\mathcal{T}}$ by the relations

$$\tilde{e}\circ\tilde{\tau}=\tilde{\tau}\circ\tilde{e}=\tilde{\tau}\quad\text{for all}\quad\tilde{\tau}\varepsilon\tilde{\mathcal{T}}, \tag{3.12a}$$

$$\tilde{a}\circ\tilde{\tau}=\tilde{\tau}\circ\tilde{a}=\tilde{a}\quad\text{for all}\quad\tilde{\tau}\varepsilon\tilde{\mathcal{T}}. \tag{3.12b}$$

The homomorphism $\tilde{\sigma}$ can now be extended to include the relations

$$\tilde{\sigma}(\tilde{1})=\tilde{e}, \tag{3.13a}$$

$$\tilde{\sigma}(\tilde{0})=\tilde{a}. \tag{3.13b}$$

Defining, for any $\tilde{\alpha}\varepsilon\tilde{\mathcal{U}}$ $\mathrm{supp}(\tilde{\alpha})=\tilde{\sigma}(\tilde{\alpha})$ [considered as a subset of $\mathcal{N}(\tilde{\mathcal{T}})$], we have

$$\mathrm{supp}(\tilde{1})=\phi, \tag{3.14a}$$

$$\mathrm{supp}(\tilde{0})=\mathcal{N}(\tilde{\mathcal{T}}). \tag{3.14b}$$

Intuitively Eq. (3.14a) is easily understood from the definitions of $\tilde{1}$ and $\mathrm{supp}(\tilde{\alpha})$. Equation (3.14b) may be understood as formalizing the position that, given a representative α of some $\tilde{\alpha}\neq\tilde{0}$, a representative β of $\tilde{0}$ can always be chosen such that $\mathrm{supp}\,\beta\supset\mathrm{supp}\,\alpha$. Whenever any of the equations (3.11b), (3.12b), (3.13b), and (3.14b) creates problems of consistency and/or interpretation, one must go back to the description of the relevant situation in terms of the null history filters in $\mathcal{U}$.

Temporal supports of typical elements of $\tilde{\mathcal{U}}$ are orientable subsets of $\mathcal{N}(\tilde{\mathcal{T}})$; however, since $\tilde{0}$ is only a formal construction, there is no compelling need to assume the whole of $\mathcal{N}(\tilde{\mathcal{T}})$ $[=\mathrm{supp}(\tilde{0})]$ to be orientable. Indeed, as we shall see below, the formalism can be consistently developed without this assumption.

We next consider some operations/relations in $\tilde{\mathcal{U}}$ which the rich structure on $\mathcal{U}_\tau$'s permits us to define.

a. *Partial Order* ($\preccurlyeq$). For $\tilde{\alpha},\tilde{\beta}\in\tilde{\mathscr{U}}$, we say that $\tilde{\alpha}\preccurlyeq\tilde{\beta}$ (to mean that $\tilde{\alpha}$ is finer than $\tilde{\beta}$) if supp $\tilde{\beta}\subset$supp $\tilde{\alpha}$ and (if supp $\tilde{\beta}$ is nonempty) there exist representatives α and β of $\tilde{\alpha}$ and $\tilde{\beta}$ in $\mathscr{U}$ (with supp $\beta\subset$supp α) such that $\alpha_\tau\preccurlyeq\beta_\tau$ for all $\tau\in$ supp β.

Proposition (3.4): The relation $\preccurlyeq$ defined above is a partial order.

Proof: Clearly $\tilde{\alpha}\preccurlyeq\tilde{\alpha}$. Transitivity is easily seen from the definition and transitivity of the partial order on $\mathscr{U}_\tau$'s. Now, if $\tilde{\alpha}\preccurlyeq\tilde{\beta}$ and $\tilde{\beta}\preccurlyeq\tilde{\alpha}$, we have supp $\tilde{\alpha}=$supp $\tilde{\beta}$ and (if these supports are nonempty) $\alpha_\tau\preccurlyeq\beta_\tau$ and $\beta_\tau\preccurlyeq\alpha_\tau$ for each τ implying $\tilde{\alpha}=\tilde{\beta}$. When supp $\tilde{\alpha}=$supp $\tilde{\beta}=\varphi$, we trivially have $\tilde{\alpha}=\tilde{\beta}(=\tilde{1})$. Q. E. D.

It immediately follows from the above definition that

$$\tilde{0}\preccurlyeq\tilde{\alpha}\preccurlyeq\tilde{1}\quad\text{for all}\quad\tilde{\alpha}\in\tilde{\mathscr{U}}. \tag{3.15}$$

b. *Disjointness* ($\perp$). For $\tilde{\alpha},\tilde{\beta}\in\tilde{\mathscr{U}}$ we say that $\tilde{\alpha}\perp\tilde{\beta}$ (i.e., $\tilde{\alpha}$ and $\tilde{\beta}$ are disjoint which is to mean that if one is realized, the other is ruled out) if they have representatives α and β in $\mathscr{U}$ such that $s_{\alpha\beta}\equiv(\text{supp }\alpha)\cap(\text{supp }\beta)$ is nonempty and $\alpha_\tau\perp\beta_\tau$ for at least one $\tau\in s_{\alpha\beta}$.

This definition has the following desirable implications:

(i) The null element $\tilde{0}$ is disjoint with all $\tilde{\alpha}$ in $\tilde{\mathscr{U}}$.

(ii) The only element of $\tilde{\mathscr{U}}$ disjoint with $\tilde{1}$ is $\tilde{0}$.

c. *Disjoint Joint Operation* ($\oplus$). This operation is defined between two typical elements $\tilde{\alpha}$ and $\tilde{\beta}$ of $\tilde{\mathscr{U}}$ if

(i) supp $\tilde{\alpha}=$supp $\tilde{\beta}$,

(ii) $\tilde{\alpha}_{\tau_0}\perp\tilde{\beta}_{\tau_0}$ for one element $\tau_0\in\text{supp}(\tilde{\alpha})$,

(iii) $\tilde{\alpha}_\tau=\tilde{\beta}_\tau$ for $\tau\in\text{supp}(\tilde{\alpha})-\{\tau_0\}$.

Given these conditions (which imply $\tilde{\alpha}\perp\tilde{\beta}$; this special case of disjointness will be referred to as weak disjointness), we define $\tilde{\alpha}\oplus\tilde{\beta}$ as follows:

$$\text{supp}(\tilde{\alpha}\oplus\tilde{\beta})=\text{supp }\tilde{\alpha}=\text{supp }\tilde{\beta}$$

$$(\tilde{\alpha}\oplus\tilde{\beta})_\tau=\begin{cases}\tilde{\alpha}_{\tau_0}\vee\tilde{\beta}_{\tau_0} & \text{for}\quad\tau=\tau_0\\ \tilde{\alpha}_\tau(=\tilde{\beta}_\tau) & \text{for}\quad\tau\in\text{supp}(\tilde{\alpha})-\{\tau_0\}\end{cases}. \tag{3.16}$$

It is easily verified that

$$\text{(i)}\quad\tilde{\alpha}\oplus\tilde{\beta}=\tilde{\beta}\oplus\tilde{\alpha}, \tag{3.17a}$$

$$\text{(ii)}\quad(\tilde{\alpha}\oplus\tilde{\beta})\oplus\tilde{\gamma}=\tilde{\alpha}\oplus(\tilde{\beta}\oplus\tilde{\gamma}). \tag{3.17b}$$

d. *Meet Operation* ($\wedge$). For general elements $\tilde{\alpha}$, $\tilde{\beta}\in\tilde{\mathscr{U}}$, we define $\tilde{\alpha}\wedge\tilde{\beta}$ as follows:

(i) If $\text{supp}(\tilde{\alpha})\cup\text{supp}(\tilde{\beta})$ is not orientable, we put $\tilde{\alpha}\wedge\tilde{\beta}=\tilde{0}$.

(ii) If supp $\tilde{\alpha}\cup$supp $\tilde{\beta}$ is orientable, we define

$$\text{supp}(\tilde{\alpha}\wedge\tilde{\beta})=\text{supp}(\tilde{\alpha})\cup\text{supp}(\tilde{\beta}), \tag{3.18a}$$

$$(\tilde{\alpha}\wedge\tilde{\beta})_\tau=\begin{cases}\tilde{\alpha}_\tau & \text{for}\quad\tau\in\text{supp }\tilde{\alpha}-\text{supp }\tilde{\beta}\\ \tilde{\beta}_\tau & \text{for}\quad\tau\in\text{supp }\tilde{\beta}-\text{supp }\tilde{\alpha}\\ \tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau & \text{for}\quad\tau\in\text{supp }\tilde{\alpha}\cap\text{supp }\tilde{\beta}\end{cases}. \tag{3.18b}$$

The history filter $\tilde{\alpha}\wedge\tilde{\beta}$ is to mean "$\tilde{\alpha}$ and $\tilde{\beta}$" (i.e., it is realized if and only if both $\tilde{\alpha}$ and $\tilde{\beta}$ are realized).

Proposition (3.5): The following relations hold in $\tilde{\mathscr{U}}$ ($\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$ denote general elements of $\mathscr{U}$):

$$\text{(i)}\quad\tilde{\alpha}\wedge\tilde{\alpha}=\tilde{\alpha}, \tag{3.19a}$$

(ii) $\tilde{\alpha}\wedge\tilde{\beta}=\tilde{\beta}\wedge\tilde{\alpha}$, (3.19b)

(iii) $\tilde{\alpha}\wedge\tilde{1}=\tilde{\alpha}$, (3.20a)

(iv) $\tilde{\alpha}\wedge\tilde{0}=\tilde{0}$, (3.20b)

(v) $\tilde{\alpha}\perp\tilde{\beta}$ implies $\tilde{\alpha}\wedge\tilde{\beta}=\tilde{0}$. [(iv) is a special case of this.]

(vi) When $\tilde{\alpha}\circ\tilde{\beta}$ is defined, we have

$$\tilde{\alpha}\wedge\tilde{\beta}=\tilde{\alpha}\circ\tilde{\beta}. \tag{3.21}$$

(vii) $\tilde{\alpha}\wedge\tilde{\beta}\leqslant\tilde{\alpha}$, (3.22)

(viii) $\tilde{\alpha}\leqslant\tilde{\alpha}\wedge\tilde{\beta}$ if and only if $\tilde{\alpha}\leqslant\tilde{\beta}$. This, with (vii), implies that $\tilde{\alpha}=\tilde{\alpha}\wedge\tilde{\beta}$ if and only if $\tilde{\alpha}\leqslant\tilde{\beta}$.

(ix) If $\tilde{\gamma}\leqslant\tilde{\alpha}$ and $\tilde{\gamma}\leqslant\tilde{\beta}$, then $\tilde{\gamma}\leqslant(\tilde{\alpha}\wedge\tilde{\beta})$. (3.23)

(x) If $\tilde{\alpha}\perp\tilde{\beta}$, then $(\tilde{\alpha}\wedge\tilde{\gamma})\perp(\tilde{\beta}\wedge\tilde{\gamma})$ for all $\tilde{\gamma}\in\tilde{\mathscr{U}}$.

Proof: Items (i)–(iv) follow immediately from the definition given above. Item (v) which, in view of item (iv) and the result that $\tilde{\alpha}\perp\tilde{1}$ only if $\tilde{\alpha}=\tilde{0}$, need be proved only for generic $\tilde{\alpha}$ and $\tilde{\beta}$, follows easily from the definitions of $\perp$ and $\wedge$.

Item (vi) follows immediately by noting that, when $\tilde{\alpha}\circ\tilde{\beta}$ is defined, we have supp $\tilde{\alpha}\cap$ supp $\tilde{\beta}=\phi$.

Item (vii): When $\tilde{\alpha}$ and/or $\tilde{\beta}$ equals $\tilde{0}$ or $\tilde{1}$, it is easily seen to follow from items (iii), (iv), (i), and Eq. (3.15). For generic $\tilde{\alpha}$ and $\tilde{\beta}$, the required support condition is trivially satisfied and the required relation follows from the definition (3.18) and the relation $\tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau\leqslant\tilde{\alpha}_\tau$.

Item (viii): Cases when $\tilde{\alpha}$ and/or $\tilde{\beta}$ equals $\tilde{0}$ or $\tilde{1}$ are easily handled. We, therefore, take $\tilde{\alpha}$ and $\tilde{\beta}$ to be generic.

Given $\tilde{\alpha}\leqslant\tilde{\alpha}\wedge\tilde{\beta}$ we have

$$\text{supp } \tilde{\alpha}\cup\text{supp } \tilde{\beta}=\text{supp}(\tilde{\alpha}\wedge\tilde{\beta})\subset\text{supp } \tilde{\alpha} \tag{A}$$

which implies

$$\text{supp } \tilde{\beta}\subset\text{supp } \tilde{\alpha}; \tag{B}$$

moreover,

$$\tilde{\alpha}_\tau\leqslant(\tilde{\alpha}\wedge\tilde{\beta})_\tau \quad \text{for all } \tau\in\text{supp } \tilde{\alpha}\cup\text{supp } \tilde{\beta}. \tag{C}$$

Now

$$(\tilde{\alpha}\wedge\tilde{\beta})_\tau=\tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau \quad \text{for } \tau\in\text{supp } \tilde{\alpha}\cap\text{supp } \tilde{\beta}=\text{supp } \tilde{\beta}. \tag{D}$$

Equations (C) and (D) give

$$\tilde{\alpha}_\tau\leqslant\tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau \quad \text{for all } \tau\in\text{supp } \tilde{\beta}. \tag{E}$$

But $\tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau\leqslant\tilde{\beta}_\tau$. Hence (E) gives

$$\tilde{\alpha}_\tau\leqslant\tilde{\beta}_\tau \quad \text{for all } \tau\in\text{supp } \tilde{\beta}. \tag{F}$$

The relations (B) and (F) imply $\tilde{\alpha}\leqslant\tilde{\beta}$.

Conversely, given $\tilde{\alpha}\leqslant\tilde{\beta}$, we have (B) and (F). The relation (B) implies (A) whereas (F), along with the trivial condition $\tilde{\alpha}_\tau\leqslant\tilde{\alpha}_\tau$ implies (E) which, in turn, implies, with Eq. (3.18) and (B), the relation (C). The relations (A) and (C) then imply $\tilde{\alpha}\leqslant\tilde{\alpha}\wedge\tilde{\beta}$.

Item (ix): The case $\tilde{\gamma}=\tilde{0}$ is trivial; $\tilde{\gamma}=\tilde{1}$ is possible only if $\tilde{\alpha}=\tilde{\beta}=\tilde{1}$ which, with (iii) above, reduces the required relation to triviality. We can, therefore, take $\tilde{\gamma}$ to be generic. The cases when $\tilde{\alpha}$ and/or $\tilde{\beta}$ equals $\tilde{0}$ or $\tilde{1}$ are trivial. We can, therefore, take $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$ to be generic.

The given conditions imply the following relations:

$$\text{supp}\ \tilde{\alpha}\subset\text{supp}\ \tilde{\gamma},\quad \tilde{\gamma}_\tau\leqslant\tilde{\alpha}_\tau\quad\text{for all}\quad\tau\in\text{supp}\ \tilde{\alpha},$$

$$\text{supp}\ \tilde{\beta}\subset\text{supp}\ \tilde{\gamma},\quad \tilde{\gamma}_\tau\leqslant\tilde{\beta}_\tau\quad\text{for all}\quad\tau\in\text{supp}\ \tilde{\beta}.$$

This implies that $\text{supp}\ \tilde{\alpha}\cup\text{supp}\ \tilde{\beta}\subset\text{supp}\ \tilde{\gamma}$ and

$$\tilde{\gamma}_\tau\leqslant\tilde{\alpha}_\tau\wedge\tilde{\beta}_\tau\quad\text{for all}\quad\tau\in\text{supp}\ \tilde{\alpha}\cap\text{supp}\ \tilde{\beta}.$$

These results, along with the given conditions and Eq. (3.18) imply $\tilde{\gamma}\leqslant\tilde{\alpha}\wedge\tilde{\beta}$.

Item (x): The given condition $\tilde{\alpha}\perp\tilde{\beta}$ implies that

$$\tilde{S}_{\alpha\beta}\equiv(\text{supp}\ \tilde{\alpha})\cap(\text{supp}\ \tilde{\beta})\neq\phi\quad\text{and}\quad\tilde{\alpha}_{\tau_0}\perp\tilde{\beta}_{\tau_0}\quad\text{for some}\quad\tau_0\in\tilde{S}_{\alpha\beta}.$$

Clearly, $\tau_0\in\text{supp}(\tilde{\alpha}\wedge\tilde{\gamma})\cap\text{supp}(\tilde{\beta}\wedge\tilde{\gamma})$.

Now, if $\tau_0\notin\text{supp}\ \tilde{\gamma}$ we have $(\tilde{\alpha}\wedge\tilde{\gamma})_{\tau_0}=\tilde{\alpha}_{\tau_0}$ and $(\tilde{\beta}\wedge\tilde{\gamma})_{\tau_0}=\tilde{\beta}_{\tau_0}$ which are disjoint. If $\tau_0\in\text{supp}\ \tilde{\gamma}$, we have $(\tilde{\alpha}\wedge\tilde{\gamma})_{\tau_0}=\tilde{\alpha}_{\tau_0}\wedge\tilde{\gamma}_{\tau_0}$ and $(\tilde{\beta}\wedge\tilde{\gamma})_{\tau_0}=\tilde{\beta}_{\tau_0}\wedge\tilde{\gamma}_{\tau_0}$. We have

$$\tilde{\alpha}_{\tau_0}\wedge\tilde{\gamma}_{\tau_0}\leqslant\tilde{\alpha}_{\tau_0}\leqslant(\neg\tilde{\beta}_{\tau_0})\leqslant(\neg\tilde{\beta}_{\tau_0})\vee(\neg\tilde{\gamma}_{\tau_0})=\neg(\tilde{\beta}_{\tau_0}\wedge\tilde{\gamma}_{\tau_0})$$

proving the desired disjointness. Q.E.D.

This proposition gives relations between various operations in $\tilde{\mathscr{U}}$. In particular, the collection $(\tilde{\mathscr{U}},\leqslant,\wedge,\tilde{0},\tilde{1})$ can now be seen to be a meet-semilattice. [For a well-organized treatment of lattices and semilattices see, for example, Ref. 16.] As mentioned in Sec. I, a meet-semilattice structure was envisaged for the space of history filters in the scheme of Isham[9] for quasitemporal theories. It is the triple $(\tilde{\mathscr{U}},\tilde{\mathscr{T}},\tilde{\sigma})$ [and not the original $(\mathscr{U},\mathscr{T},\sigma)$] which corresponds to the triple in Ref. 9.

B. The space $\Omega=\Omega(\tilde{\mathscr{U}},\tilde{\mathscr{T}})$

As was noted in Ref. 9, the space $\mathscr{U}$ (or its refined version $\tilde{\mathscr{U}}$) of history filters (''homogeneous histories'') is not adequate for a histories-based formalism of (generalized) quantum mechanics. An important ingredient which is present in $\mathscr{U}_\tau$'s (families of ''single time'' propositions) but absent in $\mathscr{U}$ or $\tilde{\mathscr{U}}$ is the concept of negation. To incorporate this, we shall now embed $\tilde{\mathscr{U}}$ in a larger space Ω of history propositions (''inhomogeneous histories''). To motivate the following constructions, it is useful to point out here that the spaces $\tilde{\mathscr{U}}$ and Ω are the analogues of, respectively, the spaces of homogeneous and inhomogeneous projection operators in the HPO space $\tilde{\mathscr{H}}$ (see the Appendix).

The elements of Ω (to be called history propositions), will be taken to be of the form

$$\underline{\alpha}=\{\tilde{\alpha}^{(1)},\tilde{\alpha}^{(2)},\ldots\},\tag{3.24}$$

where the entries on the right are at most a countable collection of elements of $\tilde{\mathscr{U}}$ such that

(i) the order of the entries on the right is immaterial;
(ii) $\tilde{\alpha}^{(j)}\perp\tilde{\alpha}^{(k)}$ for $j\neq k$;
(iii) inclusion of any number of $\tilde{0}$'s inside the braces leaves $\underline{\alpha}$ unchanged;
(iv) the union of temporal supports of every finite subcollection of non-null $\tilde{\alpha}^{(j)}$'s is orientable;

(v) if $\tilde{\alpha}$ and $\tilde{\beta}$ are weakly disjoint (as defined in Sec. III A) with $\tilde{\alpha}\oplus\tilde{\beta}=\tilde{\gamma}$, then

$$\{\tilde{\alpha},\tilde{\beta},...\}=\{\tilde{\gamma},...\}, \tag{3.25}$$

where entries represented by dots are identical on the two sides.

For consistency, we must check that, given $\tilde{\delta}\in\tilde{\mathscr{U}}$ in the collection represented by ... in Eq. (3.25), we have $\tilde{\delta}\perp\tilde{\gamma}$ if and only if $\tilde{\delta}\perp\tilde{\alpha}$ and $\tilde{\delta}\perp\tilde{\beta}$. To verify this, let τ_0 be the point of weak disjointness of $\tilde{\alpha}$ and $\tilde{\beta}$ and τ' a point of disjointness of $\tilde{\gamma}$ and $\tilde{\delta}$. Since $\tilde{\alpha}_\tau=\tilde{\beta}_\tau=\tilde{\gamma}_\tau$ for all $\tau\neq\tau_0$, the required consistency condition is trivially satisfied if $\tau'\neq\tau_0$. We need, therefore, only consider the case when the disjointness of $\tilde{\gamma}$ with $\tilde{\delta}$ holds only at $\tau'=\tau_0$. We have

$$\tilde{\alpha}_{\tau_0}\perp\tilde{\beta}_{\tau_0},\ \ \tilde{\gamma}_{\tau_0}=\tilde{\alpha}_{\tau_0}\vee\tilde{\beta}_{\tau_0},\ \ \neg\tilde{\gamma}_{\tau_0}=(\neg\tilde{\alpha}_{\tau_0})\wedge(\neg\tilde{\beta}_{\tau_0}).$$

Now (i) given $\tilde{\gamma}_{\tau_0}\perp\tilde{\delta}_{\tau_0}$ we have $\tilde{\delta}_{\tau_0}\leqslant\neg\tilde{\gamma}_{\tau_0}\leqslant\neg\tilde{\alpha}_{\tau_0}$ implying that $\tilde{\delta}\perp\tilde{\alpha}$ and similarly $\tilde{\delta}\perp\tilde{\beta}$. (ii) Given $\tilde{\delta}_{\tau_0}\perp\tilde{\alpha}_{\tau_0}$ and $\tilde{\delta}_{\tau_0}\perp\tilde{\beta}_{\tau_0}$, we have $\tilde{\delta}_{\tau_0}\leqslant\neg\tilde{\alpha}_{\tau_0}$ and $\tilde{\delta}_{\tau_0}\leqslant\neg\tilde{\beta}_{\tau_0}$ which gives

$$\tilde{\delta}_{\tau_0}\leqslant(\neg\tilde{\alpha}_{\tau_0})\wedge\neg\tilde{\beta}_{\tau_0}=\neg(\tilde{\alpha}_{\tau_0}\wedge\tilde{\beta}_{\tau_0})=\neg\tilde{\gamma}_{\tau_0}$$

implying $\tilde{\delta}\perp\tilde{\gamma}$. This completes the verification.

The object $\underline{\alpha}$ of Eq. (3.24) represents a history proposition whose realization means realization of any of the mutually exclusive $\tilde{\alpha}^{(j)}$'s [$\tilde{\alpha}^{(1)}$ or $\tilde{\alpha}^{(2)}$ or $\tilde{\alpha}^{(3)}$ or...]; it is analogous to a sum of mutually orthogonal homogeneous projection operators in the HPO space $\tilde{\mathscr{H}}$. [See Eq. (3.33) below.]

The mapping $\tilde{\alpha}\to\{\tilde{\alpha}\}$ defines an embedding of $\tilde{\mathscr{U}}$ in Ω.

An intuitively appealing representation of $\underline{\alpha}$, equivalent to that in (3.24), is as a formal disjoint join:

$$\underline{\alpha}=\tilde{\alpha}^{(1)}\oplus\tilde{\alpha}^{(2)}\oplus\cdots \tag{3.26}$$

[with stipulations analogous to those in (i)–(v) above]. Equation (3.25) then appears more natural in the form

$$\tilde{\alpha}\oplus\tilde{\beta}\oplus\cdots=\tilde{\gamma}\oplus\cdots. \tag{3.27}$$

We shall, in fact, introduce a $\oplus$ operation in Ω below and a representation of the form (3.26) will appear [see Eq. (3.33)] with $\tilde{\alpha}^{(j)}$'s replaced by $\{\tilde{\alpha}^{(j)}\}$'s and $\oplus$ as the genuine disjoint join operation between the entries. The representation (3.24), however, has its own conveniences and we shall retain it.

We define the null element $\underline{0}$ and the unit element $\underline{1}$ in Ω as

$$\underline{0}=\{\overline{0}\},\quad \underline{1}=\{\overline{1}\}. \tag{3.28}$$

Elements of Ω other than these two will be called generic.

It must be noted that representation of the form (3.24) for an $\underline{\alpha}\in\Omega$ is not unique, even after disposing off the redundant $\tilde{0}'$s and the weakly disjoint pairs. For example,

$$\underline{\alpha}=\{\neg\alpha_{\tau_1}\circ\alpha_{\tau_2},\ \ \alpha_{\tau_1}\circ\neg\alpha_{\tau_2},\ \ \neg\alpha_{\tau_1}\circ\neg\alpha_{\tau_2}\}$$

can be represented as $\{\neg\alpha_{\tau_1},\alpha_{\tau_1}\circ\neg\alpha_{\tau_2}\}$ and also as $\{\neg\alpha_{\tau_1}\circ\alpha_{\tau_2},\neg\alpha_{\tau_2}\}$. This point must be taken care of to ensure that the definitions of various operations/relations in Ω are unambiguous.

We now consider operations and relations in Ω. The interpretation of these objects must be obvious from the discussion so far.

a. *Partial Order* ($\leqslant$). Given $\underline{\alpha},\underline{\beta}\in\Omega$, we say that $\underline{\alpha}\leqslant\underline{\beta}$ if they admit representations $\underline{\alpha}=\{\tilde{\alpha}^{(j)};j\in A\}$, $\underline{\beta}=\{\tilde{\beta}^{(k)};k\in B\}$ where A and B are suitable index sets, such that, for every $j\in A$, there is a $k\in B$ such that $\tilde{\alpha}^{(j)}\leqslant\tilde{\beta}^{(k)}$.

Recalling Eq. (3.15), we have

$$\underline{0} \leqslant \underline{\alpha} \leqslant \underline{1} \quad \text{for all} \quad \underline{\alpha} \in \Omega. \tag{3.29}$$

Before introducing other operations in Ω, it is useful to introduce the concept of support-compatibility. We shall say that $\underline{\alpha}, \underline{\beta} \in \Omega$ are support-compatible if

(i) at least one of them is $\underline{0}$ or $\underline{1}$, or
(ii) both are generic satisfying the condition that, for any representation $\underline{\alpha} = \{\tilde{\alpha}^{(i)}; i \in A\}$, $\underline{\beta} = \{\tilde{\beta}^{(j)}; j \in B\}$, the joint collection $\{\tilde{\alpha}^{(i)}, \tilde{\beta}^{(j)}; i \in A, j \in B\}$ is such that the union of temporal supports of any finite subset is orientable.

b. *Disjointness* ($\perp$). Given $\underline{\alpha} = \{\tilde{\alpha}^{(1)}, \tilde{\alpha}^{(2)}, \ldots\}$ and $\underline{\beta} = \{\tilde{\beta}^{(1)}, \tilde{\beta}^{(2)}, \ldots\}$ we shall say that $\underline{\alpha} \perp \underline{\beta}$ if they are support-compatible and if $\tilde{\alpha}^{(i)} \perp \tilde{\alpha}^{(j)}$ for all i, j. In view of the discussion after Eq. (3.25), this definition suffers no ambiguity due to nonuniqueness of representation of elements of Ω.

It immediately follows from this definition that

$$\text{(i)} \quad \underline{0} \perp \underline{\alpha} \quad \text{for all} \quad \underline{\alpha} \ \text{in} \ \Omega, \tag{3.30a}$$

$$\text{(ii)} \quad \underline{\alpha} \perp \underline{1} \quad \text{if and only if} \quad \underline{\alpha} = \underline{0}. \tag{3.30b}$$

c. Disjoint Join Operation ($\oplus$) Given two disjoint elements $\underline{\alpha} = \{\tilde{\alpha}^{(i)}; i \in A\}$ and $\underline{\beta} = \{\tilde{\beta}^{(j)}; j \in B\}$ in Ω we define

$$\underline{\alpha} \oplus \underline{\beta} = \{\tilde{\alpha}^{(i)}, \tilde{\beta}^{(j)}, i \in A, j \in B\}. \tag{3.31}$$

Lemma (2.1) permits the construction of disjoint sums of a countable number of pairwise disjoint elements of Ω (note that only the union of temporal supports of any finite subset of the joint collection of history filters needs to be orientable).

It is easily verified that

$$\underline{\alpha} \oplus \underline{\beta} = \underline{\beta} \oplus \underline{\alpha}$$

$$(\underline{\alpha} \oplus \underline{\beta}) \oplus \underline{\gamma} = \underline{\alpha} \oplus (\underline{\beta} \oplus \underline{\gamma})$$

$$\underline{\alpha} \oplus \underline{0} = \underline{\alpha} \quad \text{for all} \quad \underline{\alpha} \in \Omega. \tag{3.32}$$

The $\underline{\alpha}$ of (3.24) can now be represented as

$$\underline{\alpha} = \{\tilde{\alpha}^{(1)}\} \oplus \{\tilde{\alpha}^{(2)}\} \oplus \ldots . \tag{3.33}$$

This representation makes clear the correspondence between Ω and the family $\mathscr{P}'$ of inhomogeneous projectors in the HPO space $\tilde{\mathscr{H}}$ expressible as countable orthogonal sums of homogeneous projectors.

The $\oplus$ operations in $\tilde{\mathscr{H}}$ and Ω are compatible. Indeed, if $\tilde{\alpha}$ and $\tilde{\beta}$ are weakly disjoint with $\tilde{\alpha} \oplus \tilde{\beta} = \tilde{\gamma}$ we have

$$\{\tilde{\alpha}\} \oplus \{\tilde{\beta}\} = \{\tilde{\alpha}, \tilde{\beta}\} = \{\tilde{\alpha} \oplus \tilde{\beta}\} = \{\tilde{\gamma}\}.$$

An (at most countable) set of mutually disjoint history propositions $\underline{\alpha}, \underline{\beta}, \ldots$ will be called *complete* (or *exhaustive*) if

$$\underline{\alpha} \oplus \underline{\beta} \oplus \cdots = \underline{1}. \tag{3.34}$$

Given $\underline{\alpha}, \underline{\beta} \in \Omega$ which are mutually disjoint, it is easily seen from the definitions given above that

$$\underline{\alpha} \leqslant (\underline{\alpha} \oplus \underline{\beta}). \tag{3.35}$$

Proposition (3.6): If $\underline{\alpha}$, $\underline{\beta}$, $\underline{\gamma}$ are elements of Ω, then

(i) $\underline{\alpha} \oplus \underline{\beta} \leqslant \underline{\alpha} \oplus \underline{\gamma}$ implies $\underline{\beta} \leqslant \underline{\gamma}$;

(ii) $\underline{\alpha}\oplus\underline{\beta}=\underline{\alpha}\oplus\underline{\gamma}$ implies $\underline{\beta}=\underline{\gamma}$.

Proof: (i) Let $\underline{\alpha}=\{\tilde{\alpha}^{(i)};i\in A\}$, $\underline{\beta}=\{\tilde{\beta}^{(j)};j\in B\}$, and $\underline{\gamma}=\{\tilde{\gamma}^{(k)};k\in C\}$. We have $\underline{\alpha}\oplus\underline{\beta}=\{\tilde{\alpha}^{(i)},\tilde{\beta}^{(j)};i\in A,j\in B\}$ and $\underline{\alpha}\oplus\underline{\gamma}=\{\tilde{\alpha}^{(i)},\tilde{\gamma}^{(k)};i\in A,k\in C\}$. The given condition $\underline{\alpha}\oplus\underline{\beta}\leqslant\underline{\alpha}\oplus\underline{\gamma}$ implies that, for every $\tilde{\beta}^{(j)}$, there exists either an $\tilde{\alpha}^{(i)}$ such that $\tilde{\beta}^{(j)}\leqslant\tilde{\alpha}^{(i)}$ or a $\tilde{\gamma}^{(k)}$ such that $\tilde{\beta}^{(j)}\leqslant\tilde{\gamma}^{(k)}$. The first possibility is ruled out because $\tilde{\alpha}^{(i)}\perp\tilde{\beta}^{(j)}$ for all $i\in A$, $j\in B$. The second condition implies $\underline{\beta}\leqslant\underline{\gamma}$.
(ii) The given condition $\underline{\alpha}\oplus\underline{\beta}=\underline{\alpha}\oplus\underline{\gamma}$ implies

$$\underline{\alpha}\oplus\underline{\beta}\leqslant\underline{\alpha}\oplus\underline{\gamma}\quad\text{and}\quad\underline{\alpha}\oplus\underline{\gamma}\leqslant\underline{\alpha}\oplus\underline{\beta}$$

which implies, respectively, $\underline{\beta}\leqslant\underline{\gamma}$ and $\underline{\gamma}\leqslant\underline{\beta}$ giving $\underline{\beta}=\underline{\gamma}$. Q.E.D.

We now move towards introducing a negation operation in Ω. The following theorem will play a crucial role.

Theorem (3.7): Given $\underline{\alpha},\underline{\beta}\in\Omega$, we have $\underline{\alpha}\leqslant\underline{\beta}$ if and only if there exists a $\underline{\gamma}\in\Omega$ such that $\underline{\beta}=\underline{\alpha}\oplus\underline{\gamma}$; the element $\underline{\gamma}$ is unique.

As the ''if'' part of the theorem is already implied by Eq. (3.35), we have only to consider the ''only if'' part.

In the main part of the proof below, we shall ignore possible complications arising from support considerations. Appropriate remarks relating to this will be made at the end of the proof.

For $\underline{\alpha}=\underline{0}$, the desired result follows trivially from the first of the order relations in (3.29) and Eq. (3.32). If $\underline{\alpha}=\underline{1}$, the second relation in (3.29) implies that $\underline{\alpha}\leqslant\underline{\beta}$ if and only if $\underline{\beta}=\underline{1}$; the desired result now holds trivially with $\underline{\gamma}=\underline{0}$. We therefore need consider only the case when $\underline{\alpha}$ is a generic element of Ω represented as in Eq. (3.24). In this case $\underline{\beta}$ is either generic or equals $\underline{1}$.

Let

$$\underline{\alpha}=\{\tilde{\alpha}^{(1)}\}\oplus\{\tilde{\alpha}^{(2)}\}\oplus\cdots\quad\underline{\beta}=\{\tilde{\beta}^{(1)}\}\oplus\{\tilde{\beta}^{(2)}\}\oplus\cdots.$$

When $\underline{\beta}=\underline{1}$, we have, say, $\tilde{\beta}^{(1)}=\tilde{1}$ and $\tilde{\beta}^{(k)}=\tilde{0}$ for $k>1$. As a representative of $\tilde{\beta}^{(1)}$, we may choose any $\beta\in\mathscr{U}$ with supp(β), any orientable subset of $\mathscr{N}(\mathscr{U})$, and $\beta_\tau=1_\tau$ for all $\tau\in$ supp(β).

The relation $\underline{\alpha}\leqslant\underline{\beta}$ means that, for every $\tilde{\alpha}^{(j)}$, there is a $\tilde{\beta}^{(k)}$ such that

$$\tilde{\alpha}^{(j)}\leqslant\tilde{\beta}^{(k)}. \tag{3.36}$$

In general, there may be more than one $\tilde{\alpha}^{(j)}$'s satisfying (3.36) with the same $\tilde{\beta}^{(k)}$. There may be some $\tilde{\beta}^{(k)}$'s not involved in any relation of the form (3.36); these will be called ''spectators'' and the remaining ones ''active.'' The spectator $\tilde{\beta}^{(k)}$ clearly go as ''summands'' in $\underline{\gamma}$.

To prove the theorem, it is clearly adequate to consider a general active $\tilde{\beta}^{(k)}$ such that

$$\{\tilde{\alpha}^{(p)}\}\leqslant\{\tilde{\beta}^{(k)}\}\quad\text{for } p=j,j',j'',\ldots \tag{3.37}$$

and show that

$$\{\tilde{\beta}^{(k)}\}=[\{\tilde{\alpha}^{(j)}\}\oplus\{\tilde{\alpha}^{(j')}\}\oplus\cdots]\oplus\rho(k), \tag{3.38}$$

where the entries in the parentheses are in one-to-one correspondence with the values of p in (3.37). To prove (3.38), it is, in turn, adequate to prove that

$$\{\tilde{\beta}^{(k)}\}=\{\tilde{\alpha}^{(j)}\}\oplus\underline{\lambda}^{(\cdots)} \tag{3.39}$$

with

$$\underline{\lambda}^{(\cdots)}=\{\tilde{\alpha}^{(q)}\}\oplus\cdots\quad\text{for } q=j',j'',\ldots\ . \tag{3.40}$$

When $\underline{\beta}$ is generic, the relation (3.37) implies that supp($\tilde{\beta}^{(k)}$)$\subset$supp($\tilde{\alpha}^{(p)}$) and

$$\tilde{\alpha}_\tau^{(p)} \leqslant \tilde{\beta}_\tau^{(k)} \quad \text{for all} \quad \tau \in \text{supp}(\tilde{\beta}^{(k)}). \tag{3.41}$$

Equation (3.9) now gives

$$\tilde{\beta}_\tau^{(k)} = \tilde{\alpha}_\tau^{(p)} \oplus \tilde{\in}_\tau^{(k,p)} \quad \text{for all} \quad \tau \in \text{supp}\ \tilde{\beta}^{(k)}. \tag{3.42}$$

Now, given an element $\mu \in \mathscr{U}$ and an orientable subset A of $\mathscr{N}(\mathscr{T})$ containing supp(μ), we can always extend μ to another element $\mu' \in \mathscr{U}$ with supp(μ')$=A$ by taking

$$\mu'_\tau = \begin{cases} \mu_\tau & \text{for} \quad \tau \in \text{supp}(\mu) \\ 1_\tau & \text{for} \quad \tau \in A\text{-supp}(\mu) \end{cases}. \tag{3.43}$$

We shall call such an extension trivial (for the obvious reason that the corresponding element of $\tilde{\mathscr{U}}$ is left unchanged by such an extension).

Using such a trivial extension, if necessary, we can choose representatives $\alpha^{(p)}$ and $\beta^{(k)}$ of $\tilde{\alpha}^{(p)}$ and $\tilde{\beta}^{(k)}$, respectively, such that supp($\alpha^{(p)}$)$=$supp($\beta^{(k)}$). Equation (3.42) now gets replaced by

$$\beta_\tau^{(k)} = \alpha_\tau^{(p)} \oplus \in_\tau^{(k,p)} \quad \text{for all} \quad \tau \in \text{supp}(\alpha^{(p)}). \tag{3.44}$$

When $\underline{\beta} = \underline{1}$ we have, with $\beta \in \mathscr{U}$ introduced as above, the analogue of Eq. (3.44) directly:

$$\beta_\tau (= 1_\tau) = \alpha_\tau^{(p)} \oplus \neg \alpha_\tau^{(p)} \quad \text{for all} \quad \tau \in \text{supp}(\alpha^{(p)}). \tag{3.45}$$

Now consider Eq. (3.44) for $p=j$ and q where q is any of the values $j',j'',\ldots$ of p in (3.37). Since $\alpha^{(j)} \perp \alpha^{(q)}$, we have

$$\alpha_{\tau_0}^{(q)} \leqslant \in_{\tau_0}^{(k,j)} \quad \text{for some} \quad \tau_0 \in (\text{supp}\ \alpha^{(j)}) \cap (\text{supp}\ \alpha^{(q)}). \tag{3.46}$$

Appealing to Eq. (3.9) once again, we have

$$\in_{\tau_0}^{(k,j)} = \alpha_{\tau_0}^{(q)} \oplus \cdots.$$

Equation (3.44) now gives

$$\beta_{\tau_0}^{(k)} = \alpha_{\tau_0}^{(j)} \oplus \alpha_{\tau_0}^{(q)} \oplus \cdots\ . \tag{3.47}$$

Let

$$\beta^{(k)} = \lambda' \circ \beta_{\tau_0}^{(k)} \circ \lambda'', \tag{3.48}$$

where λ' and λ'' are elements of $\mathscr{U}$ representing entries in the irreducible decomposition of $\beta^{(k)}$ ''before'' and ''after'' τ_0, respectively. From Eqs. (3.48) and (3.47), we have

$$\{\beta^{(k)}\} = \{\lambda' \circ \alpha_{\tau_0}^{(j)} \circ \lambda''\} \oplus \{\lambda' \circ \alpha_{\tau_0}^{(q)} \circ \lambda''\} \oplus \cdots\ . \tag{3.49}$$

Recalling Eq. (3.44) (with $p=j,q$) we have

$$\{\lambda' \circ \alpha_{\tau_0}^{(j)} \circ \lambda''\} = \{\alpha^{(j)}\} \oplus \cdots \tag{3.50a}$$

and

$$\{\lambda' \circ \alpha_{\tau_0}^{(q)} \circ \lambda''\} = \{\alpha^{(q)}\} \oplus \cdots\ . \tag{3.50b}$$

Equations (3.49), (3.50a), and (3.50b) together complete the proof of Eqs. (3.39) and (3.40) and hence that of the existence of the promised $\underline{\gamma}$ modulo possible support problems (arising from support requirements in the definition of elements of Ω).

All support problems are taken care of by the following two observations:

(i) Since $\underline{\alpha} \in \Omega$, the union of supports of any finite subcollection of $\tilde{\alpha}^{(j)}$'s is given as orientable.

(ii) The right-hand sides of all equations in the proof involving possibly infinite $\oplus$ operations are such that the union of supports of any finite number of ''summands'' equals the union of supports of a finite number of $\tilde{\alpha}^{(j)}$'s.

The uniqueness of $\underline{\gamma}$ follows immediately from proposition 3.6 (ii). Q.E.D.

Corollary (3.8): For every $\underline{\alpha} \in \Omega$, there is a unique $\underline{\gamma} \in \Omega$ (which is $\perp$ to $\underline{\alpha}$) such that $\underline{\alpha} \oplus \underline{\gamma} = \underline{1}$.

Proof: Take $\underline{\beta} = \underline{1}$ in *theorem (3.7).* Q.E.D.

We denote the history proposition $\underline{\gamma}$ given by *corollary (3.8)* as $\neg \underline{\alpha}$ and call it the negation of $\underline{\alpha}$. It is the unique history proposition which is realized if and only if $\underline{\alpha}$ is not realized.

Clearly, the mapping $\neg : \Omega \to \Omega$ taking $\underline{\alpha}$ to $\neg \underline{\alpha}$ is a bijection satisfying the conditions (i) $\neg(\neg \underline{\alpha}) = \underline{\alpha}$, (ii) $\underline{\alpha} \oplus \neg \underline{\alpha} = \underline{1}$, and (iii) $\neg \underline{0} = \underline{1}$; $\neg \underline{1} = \underline{0}$. These are the traditional defining properties of negation (or orthocomplementation).

Proposition (3.9): Given $\underline{\alpha}$, $\underline{\beta} \in \Omega$ which are support-compatible, we have $\underline{\alpha} \perp \underline{\beta}$ if and only if $\underline{\alpha} \leqslant \neg \underline{\beta}$.

Proof: (i) Given $\underline{\alpha} \perp \underline{\beta}$. Let $\underline{\gamma} = \underline{\alpha} \oplus \underline{\beta}$. We have

$$\underline{1} = \underline{\gamma} \oplus \neg \underline{\gamma} = \underline{\alpha} \oplus \underline{\beta} \oplus (\neg \underline{\gamma}) = \underline{\beta} \oplus \neg \underline{\beta}$$

giving

$$\underline{\alpha} \oplus (\neg \underline{\gamma}) = \neg \underline{\beta}$$

which implies, by Eq. (3.35), $\underline{\alpha} \leqslant (\neg \underline{\beta})$.

(ii) Given $\underline{\alpha} \leqslant \neg \underline{\beta}$. We have $\neg \underline{\beta} = \underline{\alpha} \oplus \underline{\delta}$, say. Now $(\neg \underline{\beta}) \perp \underline{\beta}$ implies $\underline{\alpha} \perp \underline{\beta}$ (recall the definition of $\oplus$ in Ω). Q.E.D.

It is clear from the developments in this section that Ω is an orthoalgebra as envisaged in the scheme of Isham and Linden.[10]

C. States and observables; temporal evolution

As the propositions in each $\mathscr{U}_\tau$ have the structure of a logic, we can define observables, states, and probability distributions of observables as in the traditional proposition calculus. Of course, since $\mathscr{U}_\tau$'s for different τ's are generally not isomorphic, a similar nonisomorphism of spaces of observables and states at different τ's will hold in the present formalism.

A state at ''time'' $\tau \in \mathscr{N}(\mathscr{T})$ is a kind of generalized probability on $\mathscr{U}_\tau$; more precisely, it is a map $p_\tau : \mathscr{U}_\tau \to R$ such that

(i) $0 \leqslant p_\tau(\alpha) \leqslant 1$ for each $\alpha \in \mathscr{U}_\tau$;
(ii) $p_\tau(0_\tau) = 0$, $p_\tau(1_\tau) = 1$;
(iii) given a sequence $\alpha_1, \alpha_2, \ldots$ in $\mathscr{U}_\tau$ which is pairwise disjoint (i.e., $\alpha_i \leqslant \neg \alpha_j$, for $i \neq j$), we have

$$p_\tau(\vee_i \alpha_i) = \sum_i p_\tau(\alpha_i). \tag{3.51}$$

We shall denote the collection of states at ''time'' τ by $\mathscr{S}(\tau)$.

An observable at ''time'' τ is a map $A_\tau : B(R) \to \mathscr{U}_\tau$ [where $B(R)$ is the σ-algebra of Borel sets in R, the smallest σ-algebra containing all the open intervals] such that

(i) $A_\tau(\phi) = 0_\tau$, $A_\tau(R) = 1_\tau$;
(ii) given disjoint sets E, F in $B(R)$, we have $A_\tau(E)$ and $A_\tau(F)$ disjoint;
(iii) if $E_1, E_2, \ldots$ is at most a countable collection of pairwise disjoint sets in $B(R)$, we have

$$A_\tau(\cup_k E_k) = \vee_k A_\tau(E_k). \tag{3.52}$$

We shall denote the collection of observables at ''time'' τ by $\mathcal{O}(\tau)$.

The probability distribution of the observable A_τ in the state p_τ is given by the measure $\mu^{(A_\tau,p_\tau)}$ on the measurable space $(R,B(R))$ given by

$$\mu^{(A_\tau,p_\tau)}(E)=p_\tau(A_\tau(E)) \quad \text{for all } E\in B(R). \tag{3.53}$$

It is worth remarking here that the ''time'' dependence of our states and observables corresponds neither to the Schrödinger picture nor Heisenberg picture (our states and observables both have τ dependence). This is because we are not operating in the framework of a fixed logic; we have different (generally nonisomorphic) logics at different τ's. When all the $\mathcal{U}_\tau$'s are isomorphic, we can recover the (analogue of) Heisenberg picture by referring all the states p_τ's in various $\mathcal{S}(\tau)$'s to a single $\mathcal{S}(\tau)$, say $\mathcal{S}(\tau_0)$; similarly, a Schrödinger type picture is realized by referring all the observables to $\mathcal{O}(\tau_0)$.

It appears natural to assume that the temporal evolution of the system is given by homomorphisms (structure preserving mappings) of logics; we do this in our next axiom.

A_4 (temporal evolution axiom)

The temporal evolution of the system is given by a family of mappings $V(\tau',\tau):\mathcal{U}_\tau\to\mathcal{U}_{\tau'}$ [defined for each pair $\tau,\ \tau'\in\mathcal{N}(\mathcal{T})$ such that $\tau\lhd\tau'$] such that

(i) $V(\tau',\tau)$ is onto;
(ii) $V(\tau',\tau)(0_\tau)=0_{\tau'}$, $V(\tau',\tau)(1_\tau)=1_{\tau'}$;
(iii) given $\alpha^{(1)},\ \alpha^{(2)},\ldots\in\mathcal{U}_\tau$ we have

$$V(\tau',\tau)\left(\bigvee_j \alpha^{(j)}\right)=\bigvee_j V(\tau',\tau)(\alpha^{(j)}),$$

$$V(\tau',\tau)\left(\bigwedge_j \alpha^{(j)}\right)=\bigwedge_j V(\tau',\tau)(\alpha^{(j)});$$

(iv) $V(\tau',\tau)(\neg\alpha)=\neg[V(\tau',\tau)(\alpha)]$ for all $\alpha\in\mathcal{U}_\tau$;
(v) if $\tau\lhd\tau'$, $\tau'\lhd\tau''$, and $\tau\lhd\tau''$, we have

$$V(\tau'',\tau')\cdot V(\tau',\tau)=V(\tau'',\tau),$$

where the dot between the V's indicates composition of mappings. (Note: These requirements have not been stated in minimal terms; there is some redundancy.) It is easily seen by employing Eq. (3.9) and (iii) and (iv) above that the evolution mappings $V(.,.)$ preserve partial order, i.e., given $\alpha\leqslant\beta$ in $\mathcal{U}_\tau$, we have $V(\tau',\tau)(\alpha)\leqslant V(\tau',\tau)(\beta)$ in $\mathcal{U}_\tau{}'$.

A comment is in order about condition (i) above. It reflects the view that everything that is in the system at a ''later time'' must have evolved from something in the system (supposedly closed) at an ''earlier time.'' An evolution of a closed system (through structure preserving mappings) which is not onto does not appear to make physical sense.

What about the property of $V(\tau',\tau)$ being injective (one-one)? In fact, structure preserving mappings of logics are generally taken to be injective.[14] Let us say that an evolution is nondegenerate if the evolution mapping is injective; otherwise it is degenerate. For a nondegenerate evolution, the mapping $V(\tau',\tau)$ will be bijective and therefore an isomorphism of the logics $\mathcal{U}_\tau$ and $\mathcal{U}_{\tau'}$. It follows that, if the generality of nonisomorphic $\mathcal{U}_\tau$'s is to be retained, the possibility of degenerate evolution must be admitted.

In situations where the $\mathcal{U}_\tau$'s are mutually isomorphic, we shall generally take the evolution maps $V(\tau',\tau)$ to be bijective.

The evolution described by $V(\tau',\tau)$ induces a mapping from $\mathcal{O}(\tau)$ to $\mathcal{O}(\tau')$ which maps $A_\tau\in\mathcal{O}(\tau)$ to $A_{\tau'}\in\mathcal{O}(\tau')$ given by

$$A_{\tau'}(E)=V(\tau',\tau)(A_\tau(E)) \quad \text{for all } E\in B(R). \tag{3.54}$$

If the mapping $V(\tau',\tau)$ is invertible, we also have a mapping on states taking $p_\tau\in\mathcal{S}(\tau)$ to $p_{\tau'}\in\mathcal{S}(\tau')$ given by

$$p_{\tau'}(\beta)=p_{\tau}(V(\tau',\tau)^{-1}(\beta)) \quad \text{for all} \quad \beta\in\mathscr{O}_{\tau'}. \tag{3.55}$$

The probability distribution $\mu^{(A_\tau,p_\tau)}$ of A_τ in the state p_τ is mapped to the probability distribution $\mu^{(A_{\tau'},p_{\tau'})}$ of $A_{\tau'}$ in the state $p_{\tau'}$ given by, for every $E\in B(R)$,

$$\mu^{(A_{\tau'},p_{\tau'})}(E)=p_{\tau'}(A_{\tau'},(E))=p_{\tau'}[V(\tau',\tau)(A_\tau(E))]=p_\tau(A_\tau(E))=\mu^{(A_\tau,p_\tau},(E). \tag{3.56}$$

The invariance of the probability measure μ reflected in Eq. (3.56) is essentially a consequence of the ''picture'' we are employing [see the remark after Eq. (3.53) above]. To see this we consider the corresponding situation in traditional quantum mechanics. Let $\hat{A}$ be a self-adjoint operator and E_λ its spectral family. In any state Ψ, we have

$$(\Psi,\hat{A}\Psi)=\int_R \lambda\, d(\Psi,E_\lambda\Psi). \tag{3.57}$$

The probability distribution $\mu^{(\hat{A},\Psi)}$ is given by

$$\mu^{(\hat{A},\Psi)}(E)=\int_E d(\Psi,E_\lambda\Psi) \quad \text{for all} \quad E\in B(R). \tag{3.58}$$

Now consider time dependence of states and observables (in our composite picture) given by the unitary evolution operator $U(t',t)$:

$$\Psi(t')=U(t',t)\Psi(t),$$

$$\hat{A}(t')=U(t',t)\hat{A}(t)U(t',t)^{-1}. \tag{3.59}$$

This gives

$$(\Psi(t'),\hat{A}(t')\Psi(t'))=(\Psi(t),\hat{A}(t)\Psi(t)) \tag{3.60}$$

which implies

$$\mu^{(\hat{A}_{t'},\Psi(t'))}=\mu^{(\hat{A}_t,\Psi(t))} \tag{3.61}$$

as claimed.

D. Decoherence functionals

Our fifth axiom relates to decoherence functionals. It has two parts.

A_5 (a) (Decoherence functionals)

Given a state $p_0\in\mathscr{S}(\tau_0)$ for some $\tau_0\in\mathscr{N}(\mathscr{T})$ and a law of evolution $V(\tau',\tau)$, we have a decoherence functional $d=d_{p_0,V}$ which is a mapping $d:\Omega\times\Omega\to C$ such that

(i) $d(\underline{\alpha},\underline{\beta})^*=d(\underline{\beta},\underline{\alpha})$ (hermiticity). (3.62a)

(ii) $d(\underline{\alpha},\underline{\alpha})\geqslant 0$ (positivity). (3.62b)

(iii) If $\underline{\alpha}$, $\underline{\beta}$, ... is at most a countable collection of pairwise disjoint of elements of Ω, we have

$$d(\underline{\alpha}\oplus\underline{\beta}\oplus\ldots,\underline{\gamma})=d(\underline{\alpha},\underline{\gamma})+d(\underline{\beta},\underline{\gamma})+\cdots \quad \text{(countable additivity)}. \tag{3.62c}$$

(iv) $d(\underline{1},\underline{1})=1$ (normalization). (3.62d)

The space of decoherence functionals will be denoted as $\mathscr{D}$.

A complete set $\mathscr{C}$ of history propositions is said to be (weakly) decoherent (or consistent) with respect to a decoherence functional d if

$$\text{Re } d(\underline{\alpha},\underline{\beta})=0, \quad \underline{\alpha},\underline{\beta}\in\mathscr{C}, \quad \underline{\alpha}\neq\underline{\beta}. \tag{3.63}$$

A_5 (b) (Probability interpretation)

The probability that a history $\underline{\alpha}$ in a complete set $\mathscr{C}$ which is decoherent with respect to a decoherence functional d is realized is given by

$$P(\underline{\alpha})=d(\underline{\alpha},\underline{\alpha}). \tag{3.64}$$

For $\underline{\alpha},\underline{\beta}\in\mathscr{C}$, we have the classical probability sum rule:

$$P(\underline{\alpha}\oplus\underline{\beta})=d(\underline{\alpha}\oplus\underline{\beta},\underline{\alpha}\oplus\underline{\beta})=d(\underline{\alpha},\underline{\alpha})+d(\underline{\beta},\underline{\beta})+2\,\text{Re}\,d(\underline{\alpha},\underline{\beta})=P(\underline{\alpha})+P(\underline{\beta}). \tag{3.65}$$

Finally

$$1=d(\underline{1},\underline{1})=d(\underline{\alpha}\oplus\underline{\beta}\oplus,\ldots,\underline{\alpha}\oplus\underline{\beta}\oplus\cdots)=\sum_{\underline{\alpha}\in\mathscr{C}}P(\underline{\alpha}). \tag{3.66}$$

We shall give an explicit construction of $d_{p_0,V}$ in a subclass of theories in Sec. III E below.

E. Hilbert space-based theories

An important subclass of the class of theories covered by the formalism presented above are those in which, corresponding to each $\tau\in\mathscr{N}(\mathscr{T})$, we have a separable Hilbert space $\mathscr{H}_\tau$ (there being generally no natural isomorphism between $\mathscr{H}_\tau$'s for different τ's) and $\mathscr{U}_\tau=\mathscr{P}(\mathscr{H}_\tau)$, the lattice of projection operators in $\mathscr{H}_\tau$. Here $\mathscr{T}$ is a general space of temporal supports satisfying the conditions of axioms A_1 and A_2. The object $\mathscr{U}$ in such a theory is a psg whose elements α have irreducible decompositions of the form

$$\alpha=[\cdots\alpha_{\tau_1}\circ\alpha_{\tau_2}\circ\alpha_{\tau_3}\circ\cdots], \tag{3.67}$$

where

$$\cdots\lhd\tau_1\lhd\tau_2\lhd\tau_3\lhd\cdots \tag{3.68}$$

and the bracket in Eq. (3.67) is to serve as a reminder that the decompositions can be at most semi-infinite.

In a straightforward extension of the HPO formalism given in the Appendix, elements of $\mathscr{U}$ (more appropriately, of $\tilde{\mathscr{U}}$) are homogeneous projection operators in the Hilbert space

$$\tilde{\mathscr{H}}=\bigotimes_{\tau\in N(\mathscr{T})}{}^{w}\,\mathscr{H}_\tau. \tag{3.69}$$

Corresponding to α of Eq. (3.67), we have a projection operator $\overline{\alpha}$ in $\tilde{\mathscr{H}}$ given by [compare Eq. (A4)]

$$\overline{\alpha}=[\cdots\oplus\alpha_{\tau_1}\oplus\alpha_{\tau_2}\oplus\alpha_{\tau_3}\oplus\cdots]. \tag{3.70}$$

Objects like $\overline{\alpha}$ of Eq. (3.70), in which the number of nontrivial α_{τ_i}'s is not restricted to be finite, should be understood as elements of the space

$$\mathscr{B}=\bigotimes_{\tau\in N(\mathscr{T})}\mathscr{B}(\mathscr{H}_\tau). \tag{3.71}$$

Here $\mathscr{B}(\mathscr{H}_\tau)$ denotes the algebra of bounded linear operators on $\mathscr{H}_\tau$. The construction[9] of $\mathscr{B}$ is similar to that of $\tilde{\mathscr{H}}$. One first defines a space $\mathscr{A}$ (analogous to the spaced $\mathscr{F}$ in the Appendix) whose general element A is a map from $\mathscr{N}(\mathscr{T})$ to the Cartesian product

$$X=\prod_{\tau\in N(\mathscr{T})}\mathscr{B}(\mathscr{H}_\tau) \tag{3.72}$$

such that $A(\tau)=I_\tau$ (the identity operator on $\mathscr{H}_\tau$) for all but a finite set of values of τ. Clearly $\mathscr{A}$ is a subset of X [identifying $A\in\mathscr{A}$ with the set of values $A(\tau)$]. Algebraic operations on $\mathscr{A}$ are defined in the obvious manner:

$$(aA+bB)(\tau)=aA(\tau)+bB(\tau);\quad AB(\tau)=A(\tau)B(\tau)\quad \text{for all } a,b,\in C,\qquad \text{and}\quad \tau\in N(\mathscr{T}).$$

Weak operator topologies on $\mathscr{B}(\mathscr{H}_\tau)$'s define the corresponding product topology π_W on X. The space $\mathscr{B}$ is defined as the closure of $\mathscr{A}$ in the topological space (X,π_W).

The process of association of homogeneous projection operators in $\mathscr{B}$ with elements of $\mathscr{U}$ takes care of the formality of going from the triple $(\mathscr{U},\mathscr{T},\sigma)$ to $(\tilde{\mathscr{U}},\tilde{\mathscr{T}},\tilde{\sigma})$. This is because the superfluous I_τ's get automatically ignored in passing from α to $\underline{\alpha}$ above and the presence of a 0_τ anywhere in α makes $\underline{\alpha}=0$.

The operations in $\tilde{\mathscr{U}}$ described in Sec. III A get translated into those on homogeneous projection operators in $\tilde{\mathscr{H}}$. Partial order ($\leqslant$) translates to the partial order among projectors. Disjointness gets translated to orthogonality of projectors. The operation $\oplus$ becomes orthogonal sum. Finally, the meet operation is easily seen to correspond to the meet operation on projectors (which is equivalent to that on the subspaces of the Hilbert space $\tilde{\mathscr{H}}$).

Elements of Ω are inhomogeneous projection operators in $\tilde{\mathscr{H}}$ expressible as countable sums of mutually orthogonal homogeneous projection operators. Statements in the previous paragraph apply, with due alteration of details, also to the operations in Ω described in Sec. III B. The negation operation $\neg$ translates into subtraction from the unit operator $\tilde{I}$.

For each $\tau\in\mathscr{N}(\mathscr{T})$, the state space $\mathscr{S}(\tau)$ is the family of density operators in $\mathscr{H}_\tau$ and the space $\mathscr{O}(\tau)$ of observables is the family of projectors of the form [see Eq. (3.58)]

$$p^{(\hat{A})}(E)=\int_E dE_\lambda, \tag{3.73}$$

where $\hat{A}$ is a self-adjoint operator in $\mathscr{H}_\tau$ and E is a Borel set in R. The maps $P^{(\cdot)}$: $B(R)\to\mathscr{P}(\mathscr{H}_\tau)$ are in one-to-one correspondence with self-adjoint operators in $\mathscr{H}_\tau$.

The history version of traditional quantum mechanics presented in Sec. I is clearly a special case of these theories with $\mathscr{T}=\mathscr{K}_1$ of Sec. II [which implies $\mathscr{N}(\mathscr{T})=R$, the traditional time] and, for each $\tau\in R$, $\mathscr{H}_\tau$ an isomorphic copy of the quantum mechanical Hilbert space $\mathscr{H}$.

The decoherence functional (1.5) can be easily extended to finite histories in general Hilbert space-based theories. The one constructed below [Eq. (3.77)] is essentially a generalization of Eq. (4.19) of Ref. 10 to the case of general temporal support space $\mathscr{T}$. Let

$$\alpha=\alpha_{\tau_1}\circ\cdots\circ\alpha_{\tau_n},\quad \beta=\beta_{\tau_1'}\circ\cdots\circ\beta_{\tau_m'}$$

be two histories with the same initial state $\rho(\tau_0)$. It is understood here that $\tau_0\lhd\tau_1$ and $\tau_0\lhd\tau_1'$. For every pair τ, $\tau'\in\mathscr{N}(\mathscr{T})$ with $\tau\lhd\tau'$, let $K(\tau',\tau)$: $\mathscr{H}_\tau\to\mathscr{H}_{\tau'}$ be the linear evolution maps satisfying the condition

$$K(\tau'',\tau')K(\tau',\tau)=K(\tau'',\tau) \tag{3.74}$$

for all triples τ, τ', τ'' such that $\tau\lhd\tau'\lhd\tau''$ and $\tau\lhd\tau''$. Let us assume that there exists $\tau_f\in\mathscr{N}(\mathscr{T})$ such that

$$\tau_n\lhd\tau_f\qquad\text{and}\qquad\tau_m'\lhd\tau_f. \tag{3.75}$$

[The existence of such a τ_f (which is trivial when $\mathscr{N}(\mathscr{T})=R$) is not guaranteed by our axioms and is a new assumption.] Let [compare Eq. (1.3)]

$$C_\alpha'=K(\tau_f,\tau_n)\alpha_{\tau_n}K(\tau_n,t_{n-1})\cdots\alpha_{\tau_2}K(\tau_2,\tau_1)\alpha_{\tau_1}K(\tau_1,\tau_0) \tag{3.76}$$

with a similar expression for C_β'. We now define

$$d(\alpha,\beta)=N\ \mathrm{Tr}[C'_\alpha \rho(\tau_0) C'^{+}_\beta] \tag{3.77a}$$

with

$$N^{-1}=\mathrm{Tr}[A\rho(\tau_0)B^{+}], \tag{3.77b}$$

where A and B are operators obtained from C'_α and C'_β, respectively, by putting each of the α_{τ_i} and $\alpha_{\tau'_i}$ equal to the unit operator. It is implicitly assumed here that the right-hand side of Eq. (3.77b) does not vanish. If the temporal relations among the temporal supports of α and β are such that compositions of the form (3.74) are permitted for all successive pairs, we have $A=K(\tau_f,\tau_0)$ and $B=K(\tau_f,\tau_0)^{+}$. Now, if the maps $K(.,.)$ satisfy the condition

$$K(\tau',\tau)^{\dagger}K(\tau',\tau)=I_\tau, \tag{3.78}$$

we have $N=1$.

To be consistent, we must connect the maps $K(\tau',\tau)$ to the evolution maps $V(\tau',\tau)$ postulated in the axiom A_4. In the present case the latter are maps $V(\tau',\tau):\mathscr{P}(\mathscr{H}_\tau)\to\mathscr{P}(\mathscr{H}_{\tau'})$ satisfying the conditions of axiom A_4 (i.e., they are onto and logic homomorphisms). To have this evolution implemented by maps $K(\tau',\tau):\mathscr{H}_\tau\to\mathscr{H}_{\tau'}$ employed above, it appears necessary to assume that $V(.,.)$'s are injective (one-one) as well. In this case they are logic isomorphisms; Wigner's theorem[14] then implies the existence of unitary or antiunitary maps $K(\tau',\tau):\mathscr{H}_\tau\to\mathscr{H}_{\tau'}$ such that, for any $P\in\mathscr{P}(\mathscr{H}_\tau)$,

$$V(\tau',\tau)(P)=K(\tau',\tau)PK(\tau',\tau)^{+}. \tag{3.79}$$

Equation (3.74) shows that $K(\tau'',\tau)$ is unitary when $K(\tau',\tau)$ and $K(\tau'',\tau')$ are both unitary or both antiunitary. The simplest possibility, therefore, is that all $K(\tau',\tau)$ are unitary.

Conversely, if linear maps $K(\tau',\tau)$ are given and are used to define $V(\tau',\tau)$ through Eq. (3.79), we must have the relation (3.78) if the condition $P^2=P$ is to be preserved. The condition that the map $V(\tau',\tau)$ maps the identity operator I_τ to $I_{\tau'}$ demands that

$$K(\tau',\tau)K(\tau',\tau)^{\dagger}=I_{\tau'}, \tag{3.80}$$

implying that $K(.,.)$ be unitary.

The subclass of theories considered in this section is general enough to include history version of Hilbert space-based quantum field theory on general (i.e., not necessarily globally hyperbolic) space times.

F. Classical mechanics

As promised earlier, we shall now present the history version of classical mechanics in the framework of theories described in Secs. III A–III D.

We consider classical mechanics of a system with phase space Γ. For simplicity we take $\mathscr{T}=\mathscr{K}_1$ of Sec. II which gives $\mathscr{N}(\mathscr{T})=R$, the traditional space for the flow of time. For each $t\in R$, the space $\mathscr{U}_t$ of ''single time propositions'' at time t is the collection $B(\Gamma)$ of all Borel subsets of Γ which is a Boolean logic. A general element of $\mathscr{U}$ is a collection

$$\alpha=\{\alpha(\tau_1),\ldots,\alpha(t_n);t_i\in R,t_1<\cdots<t_n;\alpha(t_i)\in\mathscr{U}_{t_i}\} \tag{3.81a}$$

which is represented, in the notation of Sec. III A, as

$$\alpha=\alpha(t_1)\circ\alpha(t_2)\circ\cdots\circ\alpha(t_n). \tag{3.81b}$$

The transition from $\mathscr{U}$ to $\tilde{\mathscr{U}}$ is made by the stipulation that the null history $\tilde{0}$ and the unit history $\tilde{1}$ are to be taken as single elements of $\tilde{\mathscr{U}}$ [corresponding, respectively, to situations where $\alpha(t)=\phi$ for at least one $t\in\mathrm{supp}(\alpha)$ and $\alpha(t)=\Gamma$ for all $t\in\mathrm{supp}(\alpha)$] and all other elements of $\tilde{\mathscr{U}}$ are $\alpha(t)$'s of the form (3.81) [now called $\tilde{\alpha}(t)$] such that $\tilde{\alpha}(t)\neq\phi$ or Γ for any $t\in\mathrm{supp}(\tilde{\alpha})$.

Given $\tilde{\alpha}$, $\tilde{\beta} \in \tilde{\mathscr{U}}$ with supp $\tilde{\alpha}=A$ and supp $\tilde{\beta}=B$, the various operations in $\tilde{\mathscr{U}}$ are to be interpreted as follows:

(i) $\tilde{\alpha} \leqslant \tilde{\beta}$ if $B \subset A$ and $\tilde{\alpha}(t) \subset \tilde{\beta}(t)$ for all $t \in B$.
(ii) $\tilde{\alpha} \perp \tilde{\beta}$ if $A \cap B \neq \phi$ and $\tilde{\alpha}(t) \cap \tilde{\beta}(t) = \phi$ or at least one $t \in A \cap B$.
(iii) Given $\tilde{\alpha}$ and $\tilde{\beta}$ weakly disjoint [i.e., $\tilde{\alpha}, \tilde{\beta}$ are disjoint, $\tilde{\alpha}(t_0) \cap \tilde{\beta}(t_0) = \phi$, where $t_0 \in A=B$ and $\tilde{\alpha}(t)=\tilde{\beta}(t)$ for $t \in A-\{t_0\}$] we have $\tilde{\alpha} \oplus \tilde{\beta} \in \tilde{\mathscr{U}}$ given by

$$(\tilde{\alpha} \oplus \tilde{\beta})(t) = \begin{cases} \tilde{\alpha}(t_0) \cup \tilde{\beta}(t_0) & \text{for } t=t_0 \\ \tilde{\alpha}(t) (=\tilde{\beta}(t)) & \text{for } t \in A-\{t_0\} \end{cases}. \tag{3.82}$$

(iv) $\tilde{\alpha} \wedge \tilde{\beta} = \tilde{0}$ if $S \equiv \text{supp } \tilde{\alpha} \cup \text{supp } \tilde{\beta}$ is not orientable. When S is orientable, we have [see Eq. (3.18)]

$$(\tilde{\alpha} \wedge \tilde{\beta})(t) = \begin{cases} \tilde{\alpha}(t) & \text{for } t \in A-B \\ \tilde{\beta}(t) & \text{for } t \in B-A, \\ \tilde{\alpha}(t) \subset \tilde{\beta}(t) & \text{for } t \in A \cap B. \end{cases} \tag{3.83}$$

Elements of Ω are, of course, at most countable collections of mutually disjoint elements of $\tilde{\mathscr{U}}$. The operations $\leqslant$, $\perp$, and $\oplus$ in Ω have a straightforward description which is left to the reader. We only describe $\neg \underline{\alpha}$ for $\underline{\alpha}=\{\tilde{\alpha}^{(1)}, \tilde{\alpha}^{(2)}, \ldots\}$.

First we define $\neg\{\tilde{\alpha}\}$ for any $\tilde{\alpha} \in \tilde{\mathscr{U}}$. Let $\text{supp}(\tilde{\alpha})=A=\{t_1, \ldots, t_n\}$ and let $h(\tilde{\alpha})$ be the collection of all elements of $\tilde{\mathscr{U}}$ of the form $\tilde{\alpha}'(t_1) \circ \cdots \circ \tilde{\alpha}'(t_n)$ where $\tilde{\alpha}'(t_i)=\tilde{\alpha}(t_1)$ or $\Gamma - \tilde{\alpha}(t_1)$ $[=\neg\tilde{\alpha}(t_1)$ in the notation of Sec. III A]. The collection $h(\tilde{\alpha})$ includes $\tilde{\alpha}$ itself and history filters disjoint with $\tilde{\alpha}$; it is an element of Ω. In fact

$$h(\tilde{\alpha}) = \underline{1}; \tag{3.84}$$

this is because $h(\tilde{\alpha})$ contains all the possible 2^n history propositions corresponding to the possible choices of the system point being in the subset $\tilde{\alpha}(t_i)$ or $\Gamma - \tilde{\alpha}(t_i)$ at time t_i for $i=1,\ldots,n$. It therefore represents a history proposition which is sure to be realized. We now have

$$\neg\{\tilde{\alpha}\} = h(\tilde{\alpha}) - \{\tilde{\alpha}\}. \tag{3.85}$$

Now let $\underline{\alpha}=\{\tilde{\alpha}^{(1)}, \tilde{\alpha}^{(2)}, \ldots\}$. Put

$$h(\underline{\alpha}) = \bigcup_i h(\tilde{\alpha}^{(i)}), \tag{3.86}$$

where the index i labels the constituents $\tilde{\alpha}^{(i)}$ of $\underline{\alpha}$. By the argument given above, we have again

$$h(\underline{\alpha}) = \underline{1} \tag{3.87}$$

and finally

$$\neg\underline{\alpha} = h(\underline{\alpha}) - \underline{\alpha}. \tag{3.88}$$

States at any time are probability measures on the measurable space $(\Gamma, B(\Gamma))$. Observables at any time t are measurable maps $A_t : B(R) \to B(\Gamma)$. For $E \in B(R)$, $A_t(E)$ represents the subset of the phase space corresponding to the observable (represented by the mapping) A_t having values in E. The probability that, at time t, the observable A_t has values in E, the state being p_t, is

$$P\{A_t \in E | p_t\} = p_t(A_t(E)). \tag{3.89}$$

Temporal evolution, according to the scheme of Sec. III C, is given, for $t<t'$, by [identifying $\mathscr{U}_t = \mathscr{U}_{t'} = B(\Gamma)$] a bijective measurable map $V(t',t)$: $B(\Gamma) \to B(\Gamma)$. This, of course, includes the traditional Hamiltonian evolution as a special case.

Given an initial state p_{t_0} and the family of temporal evolution maps $V(.,.)$, we construct the decoherence functional $d=d_{p_0,V}$ taking a clue from the D_{cl} of Ref. 6 [Eq. IV. 5.3] constructed for Hamiltonian evolution with a given Hamiltonian. We do it for two histories of the form $\alpha=(\alpha_{t_1},\ldots,\alpha_{t_n})$ and $\beta=(\beta_{t_1},\ldots,\beta_{t_n})$ such that $t_0<t_1<\cdots<t_n$. (The general case can be reduced to this by trivially extending the temporal supports of α and β to the union of their temporal supports.) We have [writing $V_{t't}$ for $V(t',t)$]

$$d(\alpha,\beta)=\int_\Gamma dp_{t_0}(\xi_0)K^{(\alpha,\beta)}(\xi_0), \tag{3.90a}$$

where [using the notation $\xi(t;\xi_0)=V_{tt_0}(\xi_0)$]

$$K^{(\alpha,\beta)}(\xi_0)=\begin{cases}1 & \text{if, for every } t_j(j=1,\ldots,n),\xi(t_j;\xi_0)\in\alpha_{t_j}\cap\beta_{t_j}\\ 0 & \text{otherwise}\end{cases}. \tag{3.90b}$$

Explicitly

$$K^{(\alpha,\beta)}(\xi_0)=\int_{\alpha_{t_1}}d\xi_1\int_{\beta_{t_1}}d\eta_1\cdots\int_{\alpha_{t_n}}d\xi_n\int_{\beta_{t_n}}d\eta_n\,\delta(\xi_1-\eta_1)\cdots\delta(\xi_n-\eta_n)\cdot$$

$$\delta[\xi_1-V_{t_1t_0}(\xi_0)]\cdots\delta[\xi_n-V_{t_nt_{n-1}}(\xi_{n-1})]. \tag{3.91}$$

Equation (3.90b) shows that $K^{(\alpha,\beta)}$ is the characteristic function of the subset $E_{\alpha\beta}$ of Γ consisting of those points ξ_0 for which $\xi(t_j;\xi_0)$ lies in $\alpha_{t_j}\cap\beta_{t_j}$ for each $j=1,\ldots,n$. We can, therefore, write

$$d(\alpha,\beta)=p_{t_0}(E_{\alpha\beta}). \tag{3.92}$$

In most practical situations, the probability measure p_{t_0} admits a density and we have $dp_{t_0}(\xi_0)=\rho_{t_0}(\xi_0)d\xi_0$. The δ-functions $\delta(\xi_i-\eta_i)$ in Eq. (3.91) ensure that $d(\alpha,\beta)=0$ for $\alpha\neq\beta$ (in any complete set of histories) showing that the decoherence condition (3.63) is trivially satisfied. Finally, the probability $P(\alpha)=d(\alpha,\alpha)$ that the history α is realized, given by Eq. (3.90), is an expected result.

APPENDIX: HPO (HISTORY PROJECTION OPERATOR) VERSION[9] OF HISTORIES IN STANDARD QUANTUM THEORY

Taking the view that histories of a system are (multitime) propositions about the system, it is desirable to have them represented, if possible, by projection operators in a suitable Hilbert space. To construct this space, consider the Cartesian product of the (naturally isomorphic) copies $\mathscr{H}_t$ $(t\in R)$ of the quantum mechanical Hilbert space $\mathscr{H}$ of the system:

$$\mathscr{V}=\Pi_{t\in R}\mathscr{H}_t. \tag{A1}$$

An element v of the $\mathscr{V}$ is a function $v:R\to\mathscr{H}$ taking $t\in R$ to $v_t\in\mathscr{H}_t\equiv\mathscr{H}$. Let w be a fixed vector in $\mathscr{V}$ whose ''component vectors'' w_t are unit vectors: $\|w_t\|_{\mathscr{H}_t}=1$. Let $\mathscr{F}$ be the subspace of $\mathscr{V}$ consisting of vectors v such that $v_t=w_t$ for all but a finite set of t values. The scalar product $(,)$ on $\mathscr{H}$ induces the following scalar product on $\mathscr{F}$:

$$(v',v)_{\mathscr{F}}=\Pi_{t\in R}(v'_t,v_t). \tag{A2}$$

The completion $\tilde{\mathscr{H}}$ of this inner product space is the desired Hilbert space; it is the infinite tensor product of the $\mathscr{H}_t$'s denoted as

$$\tilde{\mathscr{H}}=\underset{t\in R}{\otimes^w}\mathscr{H}_t. \tag{A3}$$

General linear operators on $\tilde{\mathscr{H}}$ are linear combinations of operators of the form $A=A_{t_1}\otimes\cdots\otimes A_{t_n}$, where A_{t_i} is a linear operator on $\mathscr{H}_{t_i}$ (it should be considered as a tensor product $\otimes_t A_t$ of operators such that $A_t=I_t$, the identity operator on $\mathscr{H}_t$ for all but a finite number of t's) or limits of such linear combinations. Given a set of projection operators $(\alpha_{t_1},\ldots,\alpha_{t_n})$ constituting a history α, the operator

$$\overline{\alpha}=\alpha_{t_1}\otimes\cdots\otimes\alpha_{t_n} \tag{A4}$$

is easily seen to be a projection operator in $\tilde{\mathscr{H}}$; such projection operators are referred to as homogeneous projection operators. Given a set $\alpha^{(1)},\ldots,\alpha^{(k)}$ of homogeneous projection operators in $\tilde{\mathscr{H}}$ which are mutually orthogonal (i.e., $\alpha^{(i)}\,\alpha^{(j)}=0$ for $i\neq j$), the sum $\alpha^{(1)}+\cdots+\alpha^{(k)}$ is also an (inhomogeneous) projection operator. In the HPO formalism, histories are represented by such projection operators in $\tilde{\mathscr{H}}$. One refers to histories as homogeneous or inhomogeneous depending on whether they are represented by homogeneous or inhomogeneous projection operators. If $\overline{\alpha}$ and β are mutually orthogonal homogeneous projection operators representing the histories (more precisely, history propositions) α and β, respectively, then the projection operator $\overline{\alpha}+\beta$ represents the history proposition which means that either α or β is realized.

Inclusion of inhomogeneous histories facilitates the introduction of the logical operation of negation of a history proposition. Given α, $\overline{\alpha}$ as above, the projection operator representing the negation of the history α is the inhomogeneous projection operator $\tilde{I}-\overline{\alpha}$ where $\tilde{I}$ is the identity operator on $\tilde{\mathscr{H}}$.

[1] R. B. Griffiths, J. Stat. Phys. **36**, 219 (1984); Am. J. Phys. **55**, 11 (1986).
[2] R. Omnes, J. Stat. Phys. **53**, 893 (1988); **53**, 933 (1988); **53**, 957 (1988); **57**, 357 (1989); Ann. Phys. (N.Y.) **201**, 354 (1990).
[3] R. Omnes, Rev. Mod. Phys. **64**, 339 (1992).
[4] M. Gell-Mann and J. Hartle, in *Complexity, Entropy and the Physics of Information, SFI Studies in the Science of Complexity*, edited by W. Zurek (Addison-Wesley, Reading, MA, 1990), Vol. VIII, pp. 425–458.
[5] J. Hartle, in *Quantum Cosmology and Baby Universes*, edited by S. Coleman, J. Hartle, T. Piran, and S. Weinberg (World Scientific, Singapore, 1991).
[6] J. B. Hartle, in *Gravitation and Quantization*, Proceedings of the 1992 Les Houches Summer School, edited by B. Julia and J. Zinn-Justin, Les Houches Summer School Proceedings, Vol. LVII (North-Holland, Amsterdam, 1995).
[7] M. Gell-Mann and J. Hartle, Phys. Rev. D **47**, 3345 (1993).
[8] R. M. F. Houtappel, H. Van Dam, and E. P. Wigner, Rev. Mod. Phys. **37**, 595 (1965).
[9] C. J. Isham, J. Math. Phys. **35**, 2157 (1994).
[10] C. J. Isham and N. Linden, J. Math. Phys. **35**, 5452 (1994).
[11] G. Birkhoff and J. von Neumann, Ann. Math. **37**, 823 (1936).
[12] G. W. Mackey, *Mathematical Foundations of Quantum Mechanics* (Benjamin, New York, 1963).
[13] J. M. Jauch, *Foundation of Quantum Mechanics* (Addison-Wesley, London, 1973).
[14] V. S. Varadarajan, *Geometry of Quantum Theory*, 2nd ed. (Springer-Verlag, New York, 1985).
[15] E. G. Beltrametti and G. Cassinelli, *The Logic of Quantum Mechanics* (Addison-Wesley, Reading, MA, 1981).
[16] G. Gratzer, *General Lattice Theory* (Birkhäuser Verlag, Basel, 1978).
[17] D. J. Foulis, R. J. Greechie, and G. T. Ruttimann, Int. J. Theor. Phys. **31**, 789 (1992).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Shadow of noncommutativity

Michel Dubois-Violette[a] and John Madore[b]
Laboratoire de Physique Théorique et Hautes Energies,[c] Université Paris XI, Bâtiment 211 91 405 Orsay Cedex, France

Richard Kerner[d]
Laboratoire de Physique Théorique,[e] Gravitation et Cosmologie Relativistes, Tour 22/12-4ème étage-Boîte 142, 4, place Jussieu-75005 Paris, France

(Received 28 July 1997; accepted for publication 7 October 1997)

We analyze the structure of the $\kappa=0$ limit of a family of algebras $\mathscr{A}_\kappa$ describing noncommutative versions of space–time, with κ a parameter of noncommutativity. Assuming the Poincaré covariance of the $\kappa=0$ limit, we show that, besides the algebra of functions on Minkowski space, $\mathscr{A}_0$ must contain a nontrivial extra factor $\mathscr{A}_0^I$ which is Lorentz covariant and which does not commute with the functions whenever it is not commutative. We give a general description of the possibilities and analyze some representative examples. © *1998 American Institute of Physics.* [S0022-2488(98)03201-0]

I. INTRODUCTION

In this paper we wish to give a heuristic analysis of the structure of a noncommutative space–time, as might arise from the interplay of quantum theory with gravity, in the limit where the space–time is the usual commutative Minkowski space, that is in the limit when the Planck length tends to zero (when gravity is switched off). We shall show that in this limit one necessarily captures a nontrivial extra factor in the limit algebra besides the algebra $\mathscr{C}(M)$ of functions on Minkowski space M. In view of what is known of particle interactions, it is natural to expect that this extra factor has something to do with gauge theory. We have been aware of this conclusion for some time; it was indeed one of our main motivations in studying gauge theory over the noncommutative algebra of matrix-valued functions,[1] the model for the extra factor being then described by the matrix algebra. We shall see, however, that a matrix algebra is too small to be the extra factor and that furthermore a noncommutative extra factor cannot commute with the algebra $\mathscr{C}(M)$.

We recall briefly the arguments which suggest that the interplay between quantum theory and gravitation leads to a noncommutative space–time. There is first an old semi-classical argument which is recalled Ref. 2, showing that localization loses its meaning at distances of the order of the Planck length λ_p. The argument is that, because of quantum theory, in order to localize an event to within $\Delta x^\mu \sim a$, one needs to transfer an amount of energy of order $1/a$ and that then, in view of general relativity, if a is too small, say $a<\lambda_p$, the energy would create a ''black hole.'' In fact this semi-classical argument can be made more precise[2] and leads to limitations of the form $\Delta x^0 \Sigma \Delta x^k \gtrsim \lambda_p^2$ and $\Sigma_{k<\ell} \Delta x^k \Delta x^\ell \gtrsim \lambda_p^2$ in order to avoid the above phenomena. This suggests that the x^μ do not commute, that the space–time is noncommutative. It is worth noticing here that this is not the only conclusion. For instance, one can argue that since in this argument the x^μ have a length scale, then the metric must enter somewhere and that the above uncertainty relations can be consequences of the quantization of the metric instead of a ''quantization'' of space–time itself.

There is, however, a second argument of a different nature. In classical general relativity, one solves locally Einstein's equations and one extends maximally the solutions. Doing this one obtains possibly a space–time carrying a nontrivial topology. This extension is not just a math-

[a]Electronic mail: flad@qcd.th.u-psud.fr
[b]Electronic mail: madore@qcd.th.u-psud.fr
[c]Laboratoire associé au Centre National de la Recherche Scientifique—URA D0063.
[d]Electronic mail: rk@ccr.jussieu.fr
[e]Laboratoire associé au Centre National de la Recherche Scientifique—URA D0769.

 © 1998 American Institute of Physics

ematical artifact but is physically relevant. Consider, for instance, the analog of the classical self-energies of point charges. More generally, this has to be taken into account in order to use the old physical idea[3] that gravitation acts as an ultraviolet ''regularization'' because of the fact that it is attractive and has also its own energy-momentum density source. A problem arises when one quantizes the gravitational field. The topology of space–time must then be sensitive to the states of the field. This suggests again that the functions on space–time should be replaced by elements of a noncommutative *-algebra acting on the same states as the quantized gravitation field.

In short there are several arguments suggesting that the space–time becomes noncommutative on the scale of Planck length. Although this is certainly not the only possibility, it is worth studying the consequences which follow from such an assumption when the gravitational interaction is switched off. It is the aim of the next section to analyze the ''shadow'' of such a noncommutativity in the limit when one recovers the Poincaré-covariant physical theory in the usual Minkowski space. This analysis is sharpened in Sec. III where it is pointed out that the limit has a structure of crossed product and the Hochschild two-cocycle associated with the deformation is identified as a group cocycle of the dual of the group of space–time translations. Then, in Sec. IV, we shall discuss in this context various recent proposals and, in particular, the generalizations of gauge theory using noncommutative differential calculi.[1,4,5]

II. THE COMMUTATIVE SPACE–TIME LIMIT

Let $\mathscr{A}_\kappa$ be a one-parameter family of associative algebras. Here we think of κ as being the gravitational constant λ_p^2 and we think of $\mathscr{A}_\kappa$ as being a noncommutative version of the functions on space–time. Technically, we use the framework of formal deformation theory of associative algebras.[6,7] This means that we assume that, as vector spaces, all the $\mathscr{A}_\kappa$ coincide with a fixed vector space E and that, for $f,g\in E$, one can expand their product $(fg)_\kappa$ in $\mathscr{A}_\kappa$ as

$$(fg)_\kappa = fg + \kappa c(f,g) + o(\kappa^2), \tag{1}$$

where $fg=(fg)_0$ is the product in $\mathscr{A}_0$. We also assume that there is a distinguished element, $1\in E$, which is the unit for each algebra $\mathscr{A}_\kappa$. The bilinear map $(f,g)\mapsto c(f,g)$ is then a normalized Hochschild two-cocycle of $\mathscr{A}_0$ with values in $\mathscr{A}_0$.[6] In terms of commutators, (1) yields

$$[f,\ g]_\kappa = [f,\ g] - i\kappa\{f,g\} + o(\kappa^2), \tag{2}$$

where the bracket $(f,g)\mapsto\{f,g\}$ is defined by

$$\{f,g\} = i(c(f,g) - c(g,f)) \tag{3}$$

and where $[f,\ g]_\kappa$ and $[f,\ g]$ denote the commutator in $\mathscr{A}_\kappa$ and in $\mathscr{A}_0$, respectively. The first order in κ of the identity $[h,\ (fg)_\kappa]_\kappa = ([h,\ f]_\kappa g)_\kappa + (f[h,\ g]_\kappa)_\kappa$ yields the condition

$$i([h,\ c(f,g)] - c([h,\ f],g) - c(f,[h,\ g])) = f\{h,g\} - \{h,fg\} + \{h,f\}g. \tag{4}$$

This implies that, if h is an element of the center $Z(\mathscr{A}_0)$ of $\mathscr{A}_0$, then the endomorphism δ_h of $\mathscr{A}_0$ defined by $\delta_h(f)=\{h,f\}$ is a derivation of $\mathscr{A}_0$. The center $Z(\mathscr{A}_0)$ of $\mathscr{A}_0$ is stable under the derivations of $\mathscr{A}_0$ and therefore $Z(\mathscr{A}_0)$ is stable under the bracket (3). On the other hand, if f, g, and h are elements of $Z(\mathscr{A}_0)$, then the lowest nontrivial order in κ (i.e., the second order) of $[[f,\ g]_\kappa,\ h]_\kappa = [[f,\ h]_\kappa,\ g]_\kappa + [f,\ [g,h]_\kappa]_\kappa$ yields the condition

$$\{\{f,g\},h\} = \{\{f,h\},g\} + \{f,\{g,h\}\}. \tag{5}$$

So one can summarize the above discussion by the following proposition:

Proposition 1: The center $Z(\mathscr{A}_0)$ of $\mathscr{A}_0$ is a (commutative) Poisson algebra with Poisson bracket given by (3) and one defines a linear mapping $z\mapsto\delta_z$ of $Z(\mathscr{A}_0)$ into the Lie algebra Der$(\mathscr{A}_0)$ of all derivations of $\mathscr{A}_0$ by setting $\delta_z(f)=\{z,f\}$ for $z\in Z(\mathscr{A}_0)$ and $f\in\mathscr{A}_0$ which satisfies $\delta_{zz'}=z\delta_{z'}+z'\delta_z$ for $z,z'\in Z(\mathscr{A}_0)$.

This result is known; it is the first part of Proposition 1.2 of Ref. 8 (see also Ref. 9).

We wish to represent the noncommutative analog of real functions by Hermitian elements. This leads us to add the following reality condition to the above general structure. We assume that

the $\mathscr{A}_\kappa$ are complex $*$-algebras such that all the underlying complex involutive vector spaces coincide. Thus the vector space E must be a complex vector space equipped with an antilinear involution, $f\mapsto f^*$, and a distinguished Hermitian element $\mathbb{1}=\mathbb{1}^*$. The parameter κ is real (consistently) and one has $(fg)^*_\kappa=(g^*f^*)_\kappa$ and $(f\mathbb{1})_\kappa=(\mathbb{1}f)_\kappa=f$. It follows that the normalized cocycle c satisfies $c(f,g)^*=c(g^*,f^*)$, which implies that the bracket (3) is real, i.e., $\{f,g\}^*=\{f^*,g^*\}$. Therefore, *the set* $Z_{\mathbb{R}}(\mathscr{A}_0)$ *of all Hermitian elements of the center* $Z(\mathscr{A}_0)$ *of* $\mathscr{A}_0$ *is a real (commutative) Poisson algebra and the map* $z\mapsto\delta_z$ *induces a real linear mapping of* $Z_{\mathbb{R}}(\mathscr{A}_0)$ *into the real Lie algebra* $\mathrm{Der}_{\mathbb{R}}(\mathscr{A}_0)$ *of all Hermitian derivations of* $\mathscr{A}_0$.

We now return to our specific problem. The $\mathscr{A}_\kappa$ are noncommutative versions of the algebra of functions on space–time and we wish to recover Poincaré-invariant physics on Minkowski space M in the limit $\kappa=0$. We thus assume that the Poincaré group $\mathfrak{P}$ acts by $*$-automorphisms $(\Lambda,a)\to\alpha_{(\Lambda,a)}$ on $\mathscr{A}_0$, that $\mathscr{A}_0$ contains as a $*$-subalgebra the (commutative) algebra $\mathscr{C}(M)$ of (smooth) functions on Minkowski space, and that the action of $\mathfrak{P}$ on $\mathscr{A}_0$ induces its usual action on $\mathscr{C}(M)$ associated with the corresponding transformations of M. We now argue that $\mathscr{C}(M)$ cannot be the whole $\mathscr{A}_0$. In fact, assume that $\mathscr{C}(M)$ is equal to $\mathscr{A}_0$. Then, in view of Proposition 1, there is a Poisson bracket on M. This Poisson bracket is nontrivial since we have assumed that the $\mathscr{A}_\kappa$ are noncommutative. On the other hand, there does not exist a nontrivial Poincaré-invariant Poisson bracket on M. Indeed let $(f,g)\mapsto\{f,g\}$ be such a Poisson bracket. Then $\{f,g\}=\omega^{\mu\nu}(\partial_\mu f)(\partial_\nu g)$ where $\omega^{\mu\nu}=\{x^\mu,x^\nu\}$ must be an antisymmetric tensor field on M which is constant by translational invariance and Lorentz invariant. However rotational invariance already implies that the three-vector $E^i=\omega^{0i}$ and $B^i=\Sigma_{j,k=1}^3\epsilon_{ijk}\omega^{jk}$, $i\in\{1,2,3\}$, vanish which implies $\omega^{\mu\nu}=0$ and therefore $\{f,g\}=0\forall f$, $g\in\mathscr{C}(M)$. It seems unreasonable to us that the Poincaré invariance be broken at the first order in κ. In fact, at this order, we expect a Poincaré-invariant theory involving a spin-2 field linearly coupled to the other fields. Once this hypothesis is accepted, it follows that the inclusion $\mathscr{C}(M)\subset\mathscr{A}_0$ must be a strict one; the $\kappa=0$ limit $\mathscr{A}_0$ of the $\mathscr{A}_\kappa$ must contain an extra factor. It follows also that the normalized two-cocycle c of $\mathscr{A}_0$ defined by (1) is Poincaré invariant, so one has the condition

$$\alpha_{(\Lambda,a)}(c(f,g))=c(\alpha_{(\Lambda,a)}(f),\alpha_{(\Lambda,a)}(g)),\quad\forall(\Lambda,a)\in\mathfrak{B},\quad\forall f,g\in\mathscr{A}_0,\tag{6}$$

which implies the invariance of the bracket (3).

Let $x^\mu\in\mathscr{C}(M)$ be Minkowskian coordinates. Then the algebra $\mathscr{C}(M)$ is generated by the x^ν and the action of the Poincaré group on it is given by

$$\alpha_{(\Lambda,a)}x^\mu=\Lambda^{-1\mu}_\nu(x^\nu-a^\nu\mathbb{1}).\tag{7}$$

By choosing an origin, one can identify $\mathscr{C}(M)$ with the Hopf algebra of functions on the group of translations. Since $\mathscr{C}(M)$ is a subalgebra of $\mathscr{A}_0$, the latter is (in particular) a bimodule over $\mathscr{C}(M)$. Furthermore, by restricting attention to the action of translations, $\mathscr{A}_0$ is in fact a bicovariant bimodule over the algebra of functions on the group of translations.[10,11] By standard arguments,[10,11] $\mathscr{A}_0$ is isomorphic as left $\mathscr{C}(M)$-module to $\mathscr{C}(M)\otimes\mathscr{A}_0^I$ where $\mathscr{A}_0^I$ denotes the subalgebra of translationally invariant elements of $\mathscr{A}_0$: $\mathscr{A}_0^I=\{f\in\mathscr{A}_0|\,\alpha_{(1,a)}f=f,\forall a\}$. In fact $\mathscr{A}_0$ is isomorphic to $\mathscr{C}(M)\otimes\mathscr{A}_0^I$ as $(\mathscr{C}(M),\mathscr{A}_0^I)$-bimodule. Thus in order to recover the complete structure of algebra of $\mathscr{A}_0$ in the representation $\mathscr{C}(M)\otimes\mathscr{A}_0^I$, it is sufficient to describe the right multiplication by elements of $\mathscr{C}(M)$ of elements of $\mathscr{A}_0^I$. For each Minkowskian coordinate x^μ, it follows from (7) that $\mathscr{A}_0^I$ is stable under the derivation $f\mapsto\mathrm{ad}(x^\mu)(f)=[x^\mu,f]$ and therefore one has $fx^\mu=x^\mu f-\mathrm{ad}(x^\mu)(f)$, $\forall f\in\mathscr{A}_0^I$, which can be written in the representation $\mathscr{A}_0=\mathscr{C}(M)\otimes\mathscr{A}_0^I$:

$$fx^\mu=x^\mu\otimes f-\mathbb{1}\otimes\mathrm{ad}(x^\mu)(f),\quad\forall f\in\mathscr{A}_0^I.\tag{8}$$

From this one deduces the right multiplication by elements of $\mathscr{C}(M)$, i.e., the right $\mathscr{C}(M)$-module structure of $\mathscr{C}(M)\otimes\mathscr{A}_0^I$, which is in fact isomorphic to $\mathscr{A}_0^I\otimes\mathscr{C}(M)$ as right $\mathscr{C}(M)$-module. In the following, we shall denote by X^μ the four commuting derivations of $\mathscr{A}_0^I$ induced by the $\mathrm{ad}(x^\mu)$. The algebra $\mathscr{A}_0^I$ is invariant under the automorphisms $\alpha_{(\Lambda,0)}$. We shall denote by $\alpha^I:\Lambda\mapsto^I_\Lambda$, the corresponding homomorphism of the Lorentz group into the group Aut $(\mathscr{A}_0^I)$ of all $*$-automorphisms of $\mathscr{A}_0^I$.

One can thus summarize the above discussion by the following presentation of $\mathscr{A}_0$. One starts with a unital $*$-algebra $\mathscr{A}_0^I$ equipped with four commuting anti-Hermitian derivations X^μ and an action $\Lambda \mapsto \alpha_\Lambda^I$ of the Lorentz group by automorphisms of $\mathscr{A}_0^I$ such that

$$\alpha_\Lambda^I \circ X^\mu = \Lambda_\nu^{-1\mu} X^\nu \circ \alpha_\Lambda^I \tag{9}$$

and $\mathscr{A}_0$ is ''generated'' as unital $*$-algebra by $\mathscr{A}_0^I$ and four Hermitian elements x^μ with relations $x^\mu x^\nu = x^\nu x^\mu$ and $x^\mu f = f x^\mu + X^\mu(f)$, $f \in \mathscr{A}_0^I$. We put quotes on the word generated in the above sentence because we do not pay attention here to functional analysis aspects, for instance appropriate completions, etc. The Poincaré group acts on $\mathscr{A}_0$ by the action $\alpha_{(\Lambda,a)}$ on $\mathscr{C}(M)$ defined by (7) and by $\alpha_{(\Lambda,a)}(f) = \alpha_\Lambda^I(f)$ for $f \in \mathscr{A}_0^I$.

We have assumed that the bracket (3) is not identically zero on $\mathscr{C}(M)$. This implies that the $c^{\mu\nu} = c(x^\mu, x^\nu)$ do not all vanish. On the other hand, it follows from (7) that $c^{\mu\nu}$ are elements of $\mathscr{A}_0^I$ and (6) implies then that one has

$$\alpha_\Lambda^I(c^{\mu\nu}) = \Lambda_\alpha^{-1\mu} \Lambda_\beta^{-1\nu} c^{\alpha\beta}. \tag{10}$$

Thus the homomorphism $\alpha^I : \mathfrak{L} \to \mathrm{Aut}(\mathscr{A}_0^I)$ of the Lorentz group $\mathfrak{L}$ into the group of $*$-automorphisms of $\mathscr{A}_0^I$ *is never trivial*. This places severe restrictions on the structure of the extra factor $\mathscr{A}_0^I$. For instance, $\mathscr{A}_0^I$ cannot be a finite-dimensional matrix algebra because on such an algebra all automorphisms are inner and, on the other hand, it is known that the Lorentz group has no nontrivial finite-dimensional unitary representation. The same argument shows that, if $\mathscr{A}_0^I$ admits an injective (eventually unbounded) $*$-representation in a Hilbert space, then $\mathscr{A}_0^I$ must be infinite dimensional.

III. CROSSED-PRODUCT STRUCTURE OF $\mathscr{A}_0$

There is another useful presentation of $\mathscr{A}_0$ which is a sort of exponentiation of the previous one and which can be used for a functional-analytic development of the framework. Let $\mathfrak{T}^*$ be the space $\mathbb{R}^4$ considered as the dual of the group $\mathfrak{T}$ of translations of the Minkowski space M. Instead of taking the x^μ as generators of $\mathscr{C}(M)$ we now choose the exponentials $u(k) = \exp(ik_\mu x^\mu)$ for $k \in \mathfrak{T}^*$. One has

$$u(0) = \mathbb{1}, \quad u(k)u(\ell) = u(k+\ell), \quad u(k)^* = u(-k) = u(k)^{-1}, \tag{11}$$

and (7) now yields with $(\Lambda k)_\mu = k_\nu \Lambda_\mu^{-1\nu}$,

$$\alpha_{(\Lambda,a)} u(k) = u(\Lambda k) \exp(-i(\Lambda k)_\mu a^\mu). \tag{12}$$

If $f \in \mathscr{A}_0^I$, then $\tau_k(f) = u(k) f u(-k)$ is also in $\mathscr{A}_0^I$ and $k \mapsto \tau_k$ is a homomorphism of the additive group $\mathfrak{T}^*$ into the group of $*$-automorphisms of $\mathscr{A}_0^I$. This homomorphism is in fact the exponential version of the derivations X^μ, $(\tau_k = \exp(ik_\mu X^\mu))$, and (9) is replaced by

$$\alpha_\Lambda^I \circ \tau_k = \tau_{\Lambda k} \circ \alpha_\Lambda^I. \tag{13}$$

The $*$-algebra $\mathscr{A}_0$ is generated by the $*$-algebra $\mathscr{A}_0^I$ and the $u(k)$, $k \in \mathfrak{T}^*$ with the relations (11) and the relation

$$u(k) f = \tau_k(f) u(k), \quad \forall f \in \mathscr{A}_0^I \quad \text{and} \quad \forall k \in \mathfrak{T}^*. \tag{14}$$

The action of the Poincaré group on $\mathscr{A}_0$ is given by (12) and by

$$\alpha_{(\Lambda,a)}(f) = \alpha_\Lambda^I(f), \quad \forall f \in \mathscr{A}_0^I. \tag{15}$$

The consistency is ensured by the relation (13). We can summarize the above discussion by the following proposition:

Proposition 2: The $*$-subalgebra $\mathscr{A}_0^I$ of translationally invariant elements of $\mathscr{A}_0$ is equipped with an action τ of the dual group $\mathfrak{T}^*$ of the group $\mathfrak{T}$ of translations by $*$-automorphisms and $\mathscr{A}_0$

is isomorphic to the crossed product of $\mathscr{A}_0^I$ with $\mathfrak{T}^*$ for τ. Furthermore, $\mathscr{A}_0^I$ is equipped with an action α^I of the Lorentz group by $*$-automorphisms which is connected with τ by (13) and the action of the Poincaré group on $\mathscr{A}_0$ is given by (12) and (15).

In this proposition, the crossed product is taken in the category of unital $*$-algebras. However, one can accommodate various functional analytic generalizations by working in the appropriate categories of topological $*$-algebras. This partly depends what one has in mind for the algebra $\mathscr{C}(M)$ of functions on space–time [the linear decomposition of the elements of $\mathscr{C}(M)$ over the $u(k)$ being the Fourier transformation].

Let us now consider the description of the cocycle c in this framework. The restriction of c to $\mathscr{C}(M)\times\mathscr{C}(M)$ is an $\mathscr{A}_0$-valued normalized Hochschild two-cocycle on $\mathscr{C}(M)$ which is described by the $c(u(k),u(l))$, $k,l\in\mathfrak{T}^*$. The invariance and the reality conditions yield

$$\alpha_{(\Lambda,a)}c(u(k),u(\ell))=c(u(\Lambda k),u(\Lambda\ell))\exp(-i(\Lambda(k+\ell))_\mu a^\mu), \tag{16}$$

$$c(u(k),u(\ell))^*=c(u(-\ell),u(-k)). \tag{17}$$

Define $\gamma(k,\ell)\in\mathscr{A}_0$ for $k,\ell\in\mathfrak{T}^*$ by

$$c(u(k),u(\ell))=\gamma(k,\ell)u(k+\ell). \tag{18}$$

It follows from (16) (with $\Lambda=I$) that the $\gamma(k,\ell)$ are translationally invariant, that is $\gamma(k,\ell)\in\mathscr{A}_0^I$, $\forall k,\ell\in\mathfrak{T}^*$. The reality condition (17) becomes

$$\gamma(k,\ell)^*=\tau_{k+\ell}\gamma(-\ell,-k) \tag{19}$$

and (16) yields the action

$$\alpha_\Lambda^I\gamma(k,\ell)=\gamma(\Lambda k,\Lambda\ell). \tag{20}$$

On the other hand the cocycle relation on $\mathscr{C}(M)$

$$u(k)c(u(\ell),u(m))-c(u(k)u(\ell),u(m))+c(u(k),u(\ell)u(m))-c(u(k),u(\ell))u(m)=0.$$

is equivalent to the relation

$$\tau_k\gamma(\ell,m)-\gamma(k+\ell,m)+\gamma(k,\ell+m)-\gamma(k,\ell)=0, \tag{21}$$

and, since $\gamma(k,0)=\gamma(0,k)=0$ follows from $c(\mathbb{1},x)=c(x,\mathbb{1})=0$, one has the following result:

Lemma 1: The mapping $\gamma:\mathfrak{T}^*\times\mathfrak{T}^*\to\mathscr{A}_0^I$ defined by (18) is an $\mathscr{A}_0^I$-valued normalized two-cocycle of the group $\mathfrak{T}^*$ for τ which is real, in the sense of (19), and Lorentz invariant, in the sense of (20).

Thus starting from $(\mathscr{A}_0^I,\tau,\alpha^I)$, one reconstructs $\mathscr{A}_0$ with its Poincaré automorphism by using Proposition 2 and one reconstructs the restriction to $\mathscr{C}(M)\times\mathscr{C}(M)$ of the Hochschild cocycle c from the $\mathscr{A}_0^I$-valued group cocycle γ as in Lemma 1 by formula (18). The missing items are the values of c on $\mathscr{A}_0^I\times\mathscr{C}(M)$ and on $\mathscr{A}_0^I\times\mathscr{A}_0^I$. Indeed let f and g be arbitrary elements of $\mathscr{A}_0^I$ and k,ℓ be in $\mathfrak{T}^*$. Then the cocycle relation implies

$$\begin{aligned}c(fu(k),u(\ell)g)=&(f\gamma(k,\ell)\tau_{k+\ell}(g)+c(f,\tau_{k+\ell}(g)))u(k+\ell)\\&+c(f\tau_{k+\ell}(g),u(k+\ell))-fu(k)c(u(\ell),g)-c(f,u(k))u(\ell)g\\&+fc(u(k+\ell),g)-fc(\tau_{k+\ell}(g),u(k+\ell)).\end{aligned} \tag{22}$$

This relation expresses the cocycle c on $\mathscr{A}_0\times\mathscr{A}_0$ in terms of its restrictions to $\mathscr{C}(M)\times\mathscr{C}(M)$, to $\mathscr{A}_0^I\times\mathscr{A}_0^I$, to $\mathscr{A}_0^I\times\mathscr{C}(M)$, and to $\mathscr{C}(M)\times\mathscr{A}_0^I$. By taking into account the reality condition (17), the restriction of c to $\mathscr{C}(M)\times\mathscr{A}_0^I$ can be expressed in terms of its restriction to $\mathscr{A}_0^I\times\mathscr{C}(M)$. Thus c is known whenever its restrictions to $\mathscr{C}(M)\times\mathscr{C}(M)$, to $\mathscr{A}_0^I\times\mathscr{A}_0^I$, and to $\mathscr{A}_0^I\times\mathscr{C}(M)$ are known. It follows from (6) that the restriction c^I of c to $\mathscr{A}_0^I\times\mathscr{A}_0^I$ is $\mathscr{A}_0^I$-valued and, therefore, c^I is a normalized ($\mathscr{A}_0^I$-valued) two-cocycle of $\mathscr{A}_0^I$ which satisfies the invariance condition

$$\alpha^I_\Lambda c^I(f,g)=c^I(\alpha^I_\Lambda f,\alpha^I_\Lambda g), \tag{23}$$

and the reality condition

$$c^I(f,g)^*=c^I(g^*,f^*), \tag{24}$$

for any $f,\ g\in\mathscr{A}^I_0$ and $\Lambda\in\mathfrak{L}$.

We define $\lambda(f,k)$ for $f\in\mathscr{A}^I_0$ and $k\in\mathfrak{T}^*$ by

$$c(f,u(k))-c(u(k),\tau_{-k}(f))=\lambda(f,k)u(k). \tag{25}$$

It follows from (12) that $\lambda(f,k)$ belongs to $\mathscr{A}^I_0$ and, therefore, λ is a map of $\mathscr{A}^I_0\times\mathfrak{T}^*$ into $\mathscr{A}^I_0$, linear in the first factor, which satisfies the Lorentz-invariance condition

$$\alpha^I_\Lambda\lambda(f,k)=\lambda(\alpha^I_\Lambda f,\Lambda k), \tag{26}$$

the normalization conditions

$$\lambda(f,0)=0,\quad \lambda(\mathbb{1},k)=0, \tag{27}$$

and the reality condition

$$\lambda(f,k)^*=-\tau_k(\lambda(\tau_{-k}(f^*),-k)). \tag{28}$$

The cocycle relation for c implies that one has

$$f\gamma(k,\ell)-\gamma(k,\ell)f=\tau_k(\lambda(\tau_{-k}(f),\ell))-\lambda(f,k+\ell)+\lambda(f,k) \tag{29}$$

and

$$c^I(f,g)-\tau_k(c^I(\tau_{-k}(f),\tau_{-k}(g)))=f\lambda(g,k)-\lambda(fg,k)+\lambda(f,k)g \tag{30}$$

for any $k,l\in\mathfrak{T}^*$ and $f,g\in\mathscr{A}^I_0$. Conversely if c^I, γ, λ as above are such that c^I is a Hochschild two-cocycle on $\mathscr{A}^I_0$ and such that (21), (29), and (30) are satisfied, then c is a two-cocycle on $\mathscr{A}_0$, i.e., the cocycle relation for c is equivalent to the cocycle relation for c^I and the relations (21), (29), and (30).

We now describe another way to present these relations. We first extend the action τ of $\mathfrak{T}^*$ on $\mathscr{A}^I_0$ into an action on the $\mathscr{A}^I_0$-valued Hochschild cochains on $\mathscr{A}^I_0$ by setting $\tau_k(\omega)(f_1,\ldots,f_n)=\tau_k(\omega(\tau_{-k}(f_1),\ldots,\tau_{-k}(f_n)))$ for $k\in\mathfrak{T}^*$, $f_i\in\mathscr{A}^I_0$, where $\omega\in C^n(\mathscr{A}^I_0,\mathscr{A}^I_0)$ is an $\mathscr{A}^I_0$-valued Hochschild n-cochain on $\mathscr{A}^I_0$. Let $C^{r,s}$ denote the space of $C^r(\mathscr{A}^I_0,\mathscr{A}^I_0)$-valued s-cochains on the group $\mathfrak{T}^*$. One has two differentials on the direct sum $C=\oplus C^{r,s}$. The first one is the composition with the Hochschild differential of $C(\mathscr{A}^I_0,\mathscr{A}^I_0)$ which will be denoted by $\delta^{(1,0)}$. The second one is the group differential of $\mathfrak{T}^*$ for its action τ on $C^r(\mathscr{A}^I_0,\mathscr{A}^I_0)$ which will be denoted by $(-1)^r\delta^{(0,1)}$. One has $\delta^{(1,0)}(C^{r,s})\subset C^{r+1,s}$ $\delta^{(0,1)}(C^{r,s})\subset C^{r,s+1}$ and $\delta^{(1,0)}\delta^{(0,1)}+\delta^{(0,1)}\delta^{(1,0)}=0$. Therefore $\delta=\delta^{(1,0)}+\delta^{(0,1)}$ is again a differential of degree one for the total degree $r+s$. One can restrict all differentials to the space of normalized cochains (i.e., normalized group cochains with values in normalized Hochschild cochains). One has $c^I\in C^{2,0}$, $\lambda\in C^{1,1}$, and $\gamma\in C^{0,2}$. Furthermore, the cocycle condition for c^I and the relations (21), (29), and (30) reduce to $(\delta^{(1,0)}+\delta^{(0,1)})(c^I+\lambda+\gamma)=0$. Indeed the (3,0) part of this relation $\delta^{(1,0)}c^I=0$ is the cocycle relation for c^I, the (2,1) part $\delta^{(0,1)}c^I+\delta^{(1,0)}\lambda=0$ is relation (30), and (1,2) part $\delta^{(0,1)}\lambda+\delta^{(1,0)}\gamma=0$ is relation (29), and the (0,3) part $\delta^{(0,1)}\gamma=0$ is the group cocycle relation (21).

Proposition 3: The cocycle condition for c is equivalent to $(\delta^{(1,0)}+\delta^{(0,1)})(c^I+\lambda+\gamma)=0$, i.e., to the cocycle condition for $c^I+\lambda+\gamma$ in C. Moreover, the addition of the Hochschild coboundary of a translationally invariant one-cochain to c is equivalent to the addition to $c^I+\lambda+\gamma$ of a term $(\delta^{(1,0)}+\delta^{(0,1)})(\beta^{(1,0)}+\beta^{(0,1)})$ with $\beta^{(1,0)}\in C^{(1,0)}$ and $\beta^{(0,1)}\in C^{(0,1)}$.

Indeed, if one adds to c the coboundary of b with b such that $\alpha_{(1,a)}(b(f))=b(\alpha_{(1,a)}(f))$ for $a\in\mathfrak{T}$ and $f\in\mathscr{A}_0$, then $\beta^{(1,0)}$ is the restriction of b to $\mathscr{A}^I_0$ and $\beta^{(0,1)}$ is given by $\beta^{(0,1)}(k)=b(u(k))u(-k)$ for $k\in\mathfrak{T}^*$; the invariance of b implies that $\beta^{(1,0)}$ and $\beta^{(0,1)}$ are $\mathscr{A}^I_0$ valued.

The above result is a particular case of results of Ref. 12 on the Hochschild cohomology of crossed products. Here the simplification comes from the fact that we are only interested on cochains of $\mathscr{A}_0$ which are invariant by translations (and in fact by Poincaré transformations).

It is worth noticing here that the data c^I, λ, and γ do not determine c completely. However, by using the formula (22) one can show that c is determined by c^I, λ, and γ to within the coboundary of a translationally invariant one-cochain b of $\mathscr{A}_0$ which is such that the corresponding $\beta^{(1,0)}$ and $\beta^{(0,1)}$ as above vanish identically.

For $\kappa \neq 0$, the algebra $\mathscr{A}_\kappa$ is generated by the noncommutative version of the functions on space–time. This implies that $\mathscr{A}_0$ is generated by the algebra $\mathscr{C}(M)$ of functions on space–time and the iterated applications of the cocycle c on $\mathscr{C}(M)$. More precisely, we define an increasing filtration F^n of $\mathscr{A}_0$ by unital *-subalgebras $F^n(\mathscr{A}_0)$ by setting $F^0(\mathscr{A}_0)=\mathscr{C}(M)$ and $F^{n+1}(\mathscr{A}_0)$ $=\{$*the subalgebra of $\mathscr{A}_0$ generated by $F^n(\mathscr{A}_0)$ and by $c(F^r(\mathscr{A}_0), F^s(\mathscr{A}_0)\}$* for $r+s=n(\in\mathbb{N})$. Then, our assumption means that one has $\mathscr{A}_0=\cup_n F^n(\mathscr{A}_0)=F^\infty(\mathscr{A}_0)$. Correspondingly one has an increasing filtration of $\mathscr{A}_0^I$ by unital *-subalgebras $F^n(\mathscr{A}_0^I)$ $=F^n(\mathscr{A}_0)\cap\mathscr{A}_0^I$ which is such that $F^0(\mathscr{A}_0^I)=\mathbb{C}\!\!1$, $\mathscr{A}_0^I=\cup F^n(\mathscr{A}_0^I)=F^\infty(\mathscr{A}_0^I)$ and which is Lorentz invariant, i.e., $\alpha_\Lambda^I F^n(\mathscr{A}_0^I)\subset F^n(\mathscr{A}_0^I)$ for $\Lambda\in\mathscr{L}$.

IV. DISCUSSION

The simplest cases correspond to the situation where $\mathscr{C}(M)$ is in the center of $\mathscr{A}_0$, i.e., where τ is trivial. In this case, one can assume that, up to a coboundary, c is antisymmetric on $\mathscr{C}(M)$ and in fact on $Z(\mathscr{A}_0)$ [this corresponds to a weak regularity assumption on the commutative algebra $Z(\mathscr{A}_0)$]. It then follows from the discussion in the end of last section that one has $\mathscr{A}_0$ $=Z(\mathscr{A}_0)$, i.e., that $\mathscr{A}_0$ is a commutative Poisson algebra and that the $\mathscr{A}_\kappa$ are obtained by ''quantization'' of $\mathscr{A}_0$. In this case $\mathscr{A}_0$ is, as algebra, the tensor product $\mathscr{C}(M)\otimes\mathscr{A}_0^I$. Since the $\{x^\mu,x^\nu\}$ are elements of $\mathscr{A}_0^I$ and since the Lorentz group acts there by automorphisms, $\mathscr{A}_0^I$ must contain as subalgebra an algebra of functions on a union of Lorentz orbits of antisymmetric two-tensors. The $\{x^\mu,x^\nu\}$ are then coordinate functions on this manifold of antisymmetric tensors and the orbits occurring are labeled by the pairs (α,β) of real numbers such that α is in the spectrum of $g_{\mu\mu'}g_{\nu\nu'}\{x^\mu,x^\nu\}\{x^{\mu'},x^{\nu'}\}$ and β is in the spectrum of $\varepsilon_{\alpha\beta\gamma\delta}\{x^\alpha,x^\beta\}\{x^\gamma,x^\delta\}$. It is natural to require time reversal and parity be defined and therefore to assume that whenever one has the orbit (α,β), one also has the orbit $(\alpha,-\beta)$. When furthermore one has $\{x^\lambda,\{x^\mu,x^\nu\}\}$ $=0(\forall\lambda,\mu,\nu)$, then $\mathscr{A}_0^I$ is just such an algebra of functions on a union of orbits of antisymmetric tensors. This is precisely the case for the algebra $\mathscr{A}_0$ which is the $\kappa=0$ limit of the model of Doplicher, Fredenhagen, and Robert[2] where the orbits there are (0,1) and $(0,-1)$. It is not very difficult to construct examples with $\{x^\lambda,\{x^\mu,x^\nu\}\}\neq 0$.

As pointed out above, in order to have a noncommutative algebra $\mathscr{A}_0$, it is necessary (and obviously sufficient) that the algebra $\mathscr{C}(M)$ of functions on space–time not be included in the center of $\mathscr{A}_0$, which means that τ is nontrivial. Since by assumption the cocycle γ defined by (18) is nontrivial, the simplest cases with τ nontrivial are the cases where λ and c^I vanish. In such a case, it follows from (29) that the image of γ is in the center of $\mathscr{A}_0^I$ and that $F^{n+1}(\mathscr{A}_0)$ $=F^1(\mathscr{A}_0)$, $\forall n\in\mathbb{N}$. Therefore, in view of the discussion at the end of last section, $\mathscr{A}_0$ $=F^1(\mathscr{A}_0)$ and $\mathscr{A}_0^I$ is the commutative algebra generated by the $\gamma(k,\ell)$, (or, equivalently, by the $\{x^\mu,x^\nu\}$). Thus in such a case $\mathscr{A}_0$ is the crossed product of the commutative algebra $\mathscr{A}_0^I$ with $\mathfrak{T}^*$ for τ. An example of this situation is provided by the $\kappa=0$ limit of an example elaborated by Doplicher and Fredenhagen described in Section 2 of Ref. 13. It is worth noticing here that in the $\kappa=0$ limits of examples of Refs. 2 and 13 the orbits of antisymmetric two-tensors occurring are (0,1) and $(0,-1)$ (recall that in these examples $\mathscr{A}_0^I$ is commutative). This is connected with the fact that these authors construct $\mathscr{A}_\kappa$ in such a way that physically motivated space–time uncertainty relations are implemented.

Generically τ is nontrivial and λ and c^I do not vanish. A simple example of this kind can be easily found. For $\kappa\in\mathbb{R}$, let $\mathscr{A}_\kappa$ be the unital *-algebra generated by Hermitian elements x^μ, L^μ, and $I^{\mu\nu}$ $(\mu,\nu=0,1,2,3)$ satisfying the relations

$$[x^\mu,x^\nu]=i\kappa I^{\mu\nu},$$

$$[x^\lambda,I^{\mu\nu}]=i(g^{\lambda\nu}L^\mu-g^{\lambda\mu}L^\nu), \tag{31}$$

$$[x^\mu, L^\nu] = i\kappa I^{\mu\nu},$$

$$[I^{\lambda\rho}, I^{\mu\nu}] = i(g^{\lambda\nu} I^{\mu\rho} - g^{\rho\nu} I^{\mu\lambda} + g^{\rho\mu} I^{\nu\lambda} - g^{\lambda\mu} I^{\nu\rho}),$$

$$[L^\lambda, I^{\mu\nu}] = i(g^{\lambda\nu} L^\mu - g^{\lambda\mu} L^\nu), \tag{32}$$

$$[L^\mu, L^\nu] = i\kappa I^{\mu\nu},$$

where $g^{\mu\nu}$ denotes the Minkowskian metric with $g^{00} = -1$ and $g^{11} = g^{22} = g^{33} = 1$. It follows from these relations that for $\kappa \neq 0$, $\mathscr{A}_\kappa$ is generated by the x^μ. This implies that $\mathscr{A}_0$ satisfies the property $\mathscr{A}_0 = \cup_n F^n(\mathscr{A}_0)$ of the end of last section. In fact here one has again $F^{n+1}(\mathscr{A}_0) = F^1(\mathscr{A}_0)$, $\forall n \in \mathbb{N}$. Furthermore there is (for any κ) an action of the Poincaré group $\mathfrak{B}$ by *-automorphisms $(\Lambda, a) \mapsto \alpha_{(\Lambda,a)}$ on $\mathscr{A}_\kappa$ given by

$$\alpha_{(\Lambda,a)} x^\mu = \Lambda^{-1\mu}_{\ \ \nu}(x^\nu - a^\nu), \quad \alpha_{(\Lambda,a)} L^\mu = \Lambda^{-1\mu}_{\ \ \nu} L^\nu, \quad \alpha_{(\Lambda,a)} I^{\mu\nu} = \Lambda^{-1\mu}_{\ \ \alpha} \Lambda^{-1\nu}_{\ \ \beta} I^{\alpha\beta}.$$

The commutation relations (32) between the $I^{\mu\nu}$ and the L^λ are the relations of the Lie algebra of SO(4,1) if $\kappa > 0$, of SO(3,2) if $\kappa < 0$, and of the Poincaré group if $\kappa = 0$. It follows that the $I^{\mu\nu}$ and the L^λ generate the corresponding enveloping algebra. The $x^\mu - L^\mu$ are in the center $Z(\mathscr{A}_\kappa)$ of $\mathscr{A}_\kappa$. Therefore the algebra $\mathscr{A}_\kappa$ is the tensor product of the commutative algebra generated by the $x^\mu - L^\mu$ ($\mu = 0,1,2,3$) with the above enveloping algebra. In fact the center $Z(\mathscr{A}_\kappa)$ of $\mathscr{A}_\kappa$ is generated by the $x^\mu - L^\mu$ and the two casimirs $\mathfrak{c}_2$ and $\mathfrak{c}_4$ given by

$$\mathfrak{c}_2 = \kappa g_{\lambda\mu} g_{\rho\nu} I^{\lambda\rho} I^{\mu\nu} + 2 g_{\alpha\beta} L^\alpha L^\beta, \tag{33}$$

$$\mathfrak{c}_4 = g^{\rho\rho'} \varepsilon_{\rho\lambda\mu\nu} L^\lambda I^{\mu\nu} \varepsilon_{\rho'\lambda'\mu'\nu'} L^{\lambda'} I^{\mu'\nu'}, \tag{34}$$

where $\varepsilon_{\rho\lambda\mu\nu}$ is the completely antisymmetric tensor with $\varepsilon_{0123} = 1$. Therefore $\mathscr{A}_0$ is the tensor product of the commutative algebra generated by the $x^\mu - L^\mu$ with the enveloping algebra of the Poincaré Lie algebra generated by the $I^{\mu\nu}$ and the L^λ. Notice, however, that the $I^{\mu\nu}$ and the L^λ have nothing to do with the physical Poincaré group $\mathfrak{B}$ acting on $\mathscr{A}_0$ by the automorphisms $\alpha_{(\Lambda,a)}$; in particular the L^λ have the dimension of a length. The algebra $\mathscr{A}_0^I$ is simply the subalgebra generated by the $I^{\mu\nu}$ and the L^λ ($\simeq$ enveloping algebra of a Poincaré Lie algebra). The τ_k are given by $\tau_k(I^{\mu\nu}) = I^{\mu\nu} + k^\mu L^\nu - k^\nu L^\mu$ and $\tau_k(L^\lambda) = L^\lambda$. One can compute the cocycle c, e.g., one has $c(x^\mu, x^\nu) = -\frac{1}{2} I^{\mu\nu}$. Finally, allowing exponentials, the cocycle γ is given by

$$\gamma(k, \ell) = -\frac{i}{2}\left(k_\mu \ell_\nu I^{\mu\nu} - \left(\frac{2}{3} k_\rho \ell^\rho + \frac{1}{3} \ell_\rho \ell^\rho\right) k_\lambda L^\lambda + \left(\frac{1}{3} k_\rho \ell^\rho + \frac{2}{3} k_\rho k^\rho\right) \ell_\lambda L^\lambda\right),$$

and one has $\lambda(L^\mu, k) = -k_\nu (I^{\mu\nu} + \frac{1}{2}(k^\mu L^\nu - k^\nu L^\mu))$ and $c^I(L^\mu, L^\nu) = -\frac{1}{2} I^{\mu\nu}$ and so forth. Since $\mathfrak{c}_2$ and $\mathfrak{c}_4$ are in the center of $\mathscr{A}_\kappa$ and since they are translationally invariant it is natural to fix them and to add the corresponding relations in the definition of the $\mathscr{A}_\kappa$. Since the element $\mathfrak{c}_2$ has the dimension of a length squared, there are three natural ways to fix it: (i) $\mathfrak{c}_2 = \kappa$, (ii) $\mathfrak{c}_2 = -\kappa$, and (iii) $\mathfrak{c}_2 = 0$. Any of these choices leads to $g_{\mu\nu} L^\mu L^\nu = 0$ in $\mathscr{A}_0$ (in fact in $\mathscr{A}_0^I$). Remembering that $\mathscr{A}_0^I$ has the structure of the enveloping algebra of a Poincaré Lie algebra, the latter condition is the analog of a zero mass condition (although here it has a different meaning as we pointed out). Thus if one assumes furthermore that $\mathfrak{c}_4$ is fixed in such a way that the representations occurring are ''zero mass'' representations with strictly positive ''spin,'' one finds again a characteristic two-sheeted structure for $\mathscr{A}_0^I$ (by allowing the two helicities). The origin of the frequent occurrence of this two-sheeted structure in this context is obviously due to the fact that the full Lorentz group is not connected.

In all the above examples the $\mathscr{A}_\kappa$ are Poincaré covariant. This is not needed; only $\mathscr{A}_0$ must have this property. In fact one can easily produce an example where only $\mathscr{A}_0$ is Poincaré covariant by deforming the universal enveloping algebras occurring in the previous example for $\kappa \neq 0$ (in the sense used in the theory of quantum groups).

If one wishes to establish a connection between the extra factor $\mathscr{A}_0^I$ here and the one occurring in recent noncommutative versions of gauge theory,[1,4,5] one should remember that, according

to our analysis, $\mathcal{A}_0^I$ dimensional and that it can be noncommutative only if some of its elements do not commute with the functions on space–time. Concerning the first point, one could expect that, by some contractibility argument, a finite-dimensional approximation can be found. However the second point remains. This suggests that it is worth trying to enlarge the setting of the noncommutative models of gauge theory.

ACKNOWLEDGMENTS

It is a pleasure to thank Ivan Todorov for stimulating discussions, Victor Nistor for enlightening correspondence, and Sunggoo Cho for pointing out to us a mistake in the previous version of the paper.

[1] M. Dubois-Violette, R. Kerner, and J. Madore, ''Gauge bosons in a noncommutative geometry,'' Phys. Lett. B **217**, 485 (1989); ''Classical bosons in a noncommutative geometry,'' Class. Quantum Grav. **6**, 1709 (1989); ''Noncommutative differential geometry of matrix algebras,'' J. Math. Phys. **31**, 316 (1990); ''Noncommutative differential geometry and new models of gauge theory,'' J. Math. Phys. **31**, 323 (1990).
[2] S. Doplicher, K. Fredenhagen, and J. E. Roberts, ''The quantum structure of spacetime at the Planck scale and quantum fields,'' Commun. Math. Phys. **172**, 187 (1995).
[3] S. Deser, ''General relativity and the divergence problem in quantum field theory,'' Rev. Mod. Phys. **29**, 417 (1957).
[4] A. Connes and J. Lott, ''Particles models and non commutative geometry,'' Nucl. Phys. B, Proc. Suppl. **11**, 19 (1990); A. Connes, *Non-commutative geometry* (Academic, New York, 1994).
[5] R. Coquereaux, ''Higgs fields and superconnections,'' in *Differential Geometric Methods in Theoretical Physics*, Rapallo 1990, edited by C. Bartocci, U. Bruzzo, and R. Cianci, *Lecture Notes in Physics Vol. 375* (Springer Verlag, New York, 1991); R. Coquereaux, R. Häuβling, and F. Scheck, ''Algebraic connections on parallel universe,'' Int. J. Mod. Phys. A **10**, 89–98 (1995).
[6] M. Gernstenhaber, ''On the deformation of rings and algebras,'' Ann. Math. **79**, 59 (1964).
[7] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, ''Deformation theory and quantization. I. Deformations of symplectic structures, Ann. Phys. (N.Y.) **111**, 61 (1978); ''II. Physical applications,'' Ann. Phys. (N.Y.) **111**, 111 (1978).
[8] N. Reshetikhin, A. Voronov, and A. Weinstein, ''Semiquantum geometry,'' Preprint Juin 1996.
[9] C. De Concini, V. G. Kac, and C. Procesi, ''Some quantum analogues of solvable Lie groups,'' *Geometry and Analysis*, Tata Institute of Fundamental Research, hep-th/9308138.
[10] S. L. Woronowicz, ''Differential calculus on compact matrix pseudogroups (quantum groups),'' Commun. Math. Phys. **122**, 125 (1989).
[11] C. Bertrand, ''Calcul différentiel non commutatif sur une algèbre de Hopf. Un outil essentiel, la représentation τ'', Bull. Sci. Math. **119**, 49 (1995).
[12] V. Nistor, ''Group cohomology and the cyclic cohomology of crossed products,'' Invent. Math. **99**, 411 (1990).
[13] S. Doplicher, ''Quantum spacetime,''Ann. Inst. Henri Poincaré **64**, 543 (1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

The role of a form of vector potential — normalization of the antisymmetric gauge

Wojciech Florek[a] and Stanisław Wałcerz[b]
A. Mickiewicz University, Institute of Physics, Computational Physics Division, ul. Umultowska 85, 61–614 Poznań, Poland

(Received 23 July 1997; accepted for publication 14 October 1997)

Results obtained for the antisymmetric gauge $\mathbf{A}=[Hy,-Hx]/2$ by Brown and Zak are compared with those based on pure group-theoretical considerations and corresponding to the Landau gauge $\mathbf{A}=[0,Hx]$. Imposing the periodic boundary conditions one has to be very careful since the first gauge leads to a factor system which is not normalized. A period N introduced in Brown's and Zak's papers should be considered as a *magnetic* one, whereas the *crystal* period is in fact $2N$. The ''normalization'' procedure proposed here shows the equivalence of Brown's, Zak's, and other approaches. It also indicates the importance of the concept of *magnetic cells*. Moreover, it is shown that factor systems (of projective representations and central extensions) are gauge-dependent, whereas a commutator of two magnetic translations is gauge-independent. This result indicates that a *form* of the vector potential (a gauge) is also important in physical investigations. © *1998 American Institute of Physics.* [S0022-2488(98)01202-X]

I. INTRODUCTION

The discovery of the quantum Hall effect[1,2] led to remarkable interest in two-dimensional electron systems subjected to a magnetic field.[3] Since 1980 authors working in different fields — from applied to mathematical physics — have considered related problems and many new features have been observed and discussed.[4] One of the most interesting questions is the dynamic of two-dimensional electrons in a periodic potential and an external magnetic field.[5] The first results, in the tight binding approximation, were presented by Peierls,[6] shortly after Landau's[7] discovery of the quantization of electron states in a magnetic field. A new impact was due to Brown[8] and Zak[9,10] who independently introduced magnetic translation operators in two different, but equivalent, ways. Both approaches were based on group-theoretical considerations and led to the broadening of the Landau levels and quantization of a magnetic field.[8,11] Although more than thirty years have passed, their papers are still considered as fundamental ones.[5] Brown and Zak proved that the problem considered is in fact two-dimensional and their investigations confirmed the importance of projective representations and central extensions in quantum physics.[12] On the other hand, Zak's and Brown's results were not gauge-independent — only a completely antisymmetric vector potential was considered by both authors. An attempt to consider gauge-equivalent vector potentials leads to some ambiguities and misconceptions if it is not done carefully. A bit simpler and more clear results can be obtained from pure group-theoretical considerations. For example, Divakaran and Rajagopal did not consider gauges at all and they worked with central extensions and projective representations only.[13] However, pure mathematical description may not provide us with an intuitive image of the physical phenomena. Moreover, many experiments and theories indicate the importance of vector potential,[14] so it is necessary to include gauges and potentials in considerations.

The aim of this paper is to show sources of misconceptions, ambiguities, and unexpected gauge-dependence of the problem. In particular, factor systems of projective representations and central extensions introduced by Brown and Zak have been carefully checked and compared with those obtained from pure group-theoretical considerations.[13,15,16] It occurs that they can be con-

[a]Electronic mail: florek@phys.amu.edu.pl
[b]Electronic mail: stanw@phys.amu.edu.pl

 © 1998 American Institute of Physics

sidered as standard but they are not normalized.[12,17] This last fact is the main source of differences between Brown's and Zak's approaches. Moreover, it indicates points at which a form of the vector potential is important, i.e., the points at which the problem is not gauge-independent.

In this paper we propose a procedure of ''normalization'' of those factor systems, which enables us to identify irreps introduced by Brown and Zak. A comparison of these irreps with those obtained for central extensions of finite translation groups leads to a concept of the so-called *magnetic cells*[10] and shows that Brown and Zak considered in fact finite lattices with a period $2N$ not N.

For the sake of clarity, the following simplifications arising from the quoted papers are assumed. Position ($\mathbf{r}$, $\mathbf{R}$), momentum ($\mathbf{p}$), and vector potential ($\mathbf{A}$) are considered to be two-dimensional vectors. Note that $\mathbf{r}=(x,y)$ is any vector of $\mathbb{R}^2$, whereas $\mathbf{R}=(X,Y)\in\mathbb{Z}^2$ denotes a vector of a square lattice with $\mathbf{a}_1=\hat{\mathbf{x}}$ and $\mathbf{a}_2=\hat{\mathbf{y}}$, so the area of the elementary cell is equal to 1. The magnetic field is perpendicular to the x-y plane and $\mathbf{H}=H\hat{\mathbf{z}}$. The periodic boundary conditions are imposed on representations of $\mathbb{Z}^2$ and the periods are equal, i.e., $N_1=N_2=N$; the finite translation group and its representations can be considered equivalently.

The paper is organized as follows. In Sec. II the most fundamental formulas of Brown's and Zak's papers are recalled and equivalence of their approaches are indicated. Basic properties of projective representations are briefly presented, too. The role of factor systems is briefly discussed in Sec. III. The next section is devoted to determination of the equivalence of different approaches. From the physical point of view it is done by introducing the concept of magnetic cells. The results obtained are discussed in Sec. V.

II. DIFFERENT DESCRIPTIONS OF MAGNETIC TRANSLATION GROUPS

From the algebraic point of view there are two equivalent descriptions of the magnetic translation operators. Brown[8] investigated a *projective* representation of the translation group T then imposed the magnetically periodic boundary conditions on it. On the other hand, Zak[9] introduced a closed set of noncommuting operators which, in fact, form a covering group T' of T so its standard (vector) representations are *projective* representations of T.[17] The finiteness of these representations was again achieved by imposing the periodic boundary conditions. These two approaches are related by a formula which follows from the induction procedure if one constructs representations of the covering group. Since T' is a central extension of T by the group $\mathrm{U}(1)$ (or its subgroup referred to hereafter as a group of factors and denoted F)[15,16] then its (vector) representations can be written as

$$\Xi[\alpha,\mathbf{R}]=\Gamma(\alpha)D(\mathbf{R}), \tag{1}$$

where $\alpha\in F\subset\mathrm{U}(1)$, $\mathbf{R}\in T$, Γ is a vector representation of F and D is a *projective* representation of T. A factor system $m(\mathbf{R},\mathbf{R}')$ of this representations is determined by the relation[17]

$$D(\mathbf{R})D(\mathbf{R}')=m(\mathbf{R},\mathbf{R}')D(\mathbf{R}+\mathbf{R}'), \tag{2}$$

whereas the multiplication rule for T' reads

$$[\alpha,\mathbf{R}][\alpha',\mathbf{R}']=[\alpha\alpha'\mu(\mathbf{R},\mathbf{R}'),\mathbf{R}+\mathbf{R}'] \tag{3}$$

with $\mu(\mathbf{R},\mathbf{R}')$ being a factor system of a central extension. These factor systems are related to each other by the formula

$$m(\mathbf{R},\mathbf{R}')=\Gamma[\mu(\mathbf{R},\mathbf{R}')]. \tag{4}$$

This relation establishes the equivalence of both approaches. Moreover, both authors assumed the antisymmetric vector potential (gauge) $\mathbf{A}=(\mathbf{H}\times\mathbf{r})/2=[Hy,-Hx]/2$ and were not able to generalize their considerations to other gauges, in particular their approaches did not include the Landau gauge. On the other hand, their results and some conclusions are different in some points which will be discussed here[18] and compared with the results obtained for the Landau gauge.

All considerations and formulas given above are also valid for a finite group T_N and its (finite-dimensional) representations. In addition, we can apply to this case a version of the Burnside theorem which reads that nonequivalent irreducible projective representations of T_N with the same factor system $m(\mathbf{R},\mathbf{R}')$ satisfy the following condition[17]

$$\sum_j [{}^jD]^2 = |T_N| = N^2, \tag{5}$$

where j labels nonequivalent representations (there is no expression for a number of these representations) and $[{}^jD]$ denotes the dimension of ${}^jD$. Since F is an Abelian group then it has $|F|$ irreducible nonequivalent representations and each of them determines different (nonequivalent) factor system $m(\mathbf{R},\mathbf{R}')$ according to (4). It follows from (5) that irreducible representations of T'_N determined by (1) satisfy the Burnside theorem.

Brown[8] defined a magnetic translation operator as

$$\hat{T}(\mathbf{R}) = \exp[-i\mathbf{R}\cdot(\mathbf{p} - e\mathbf{A}/c)/\hbar], \tag{6}$$

where $\mathbf{p}$ is the kinetic momentum and $\mathbf{A}$ is the vector potential such that $\nabla\times\mathbf{A} = \mathbf{H}$. These operators form a projective representation of T with a factor system[8]

$$m(\mathbf{R},\mathbf{R}') = \exp[-\pi i(\mathbf{R}\times\mathbf{R}')\cdot\mathbf{H}/\varphi_0], \tag{7}$$

where $\varphi_0 = ch/e$. Brown showed that one can impose the periodic boundary conditions $\hat{T}(N\mathbf{a}_j)\psi = \psi$ if (notice simplifications assumed in this paper — in fact, H denotes hereafter the magnetic flux through one primitive cell)

$$H = \frac{l}{N}\varphi_0, \tag{8}$$

where l is mutually prime with N, i.e., $\gcd(l,N) = 1$. Hence a factor system of a finite projective representation ${}^lD$ for H satisfying (8) is given as

$$m_l([X,Y],[X',Y']) = \exp[-\pi i l(XY' - YX')/N]. \tag{9}$$

Brown showed that there is the unique (up to equivalence) irreducible projective representation with a dimension N and matrix elements[8]

$${}^l_1D_{jk}[X,Y] = \exp\left[\pi i \frac{lX}{N}(Y + 2j)\right]\delta_{j,k-Y}, \tag{10}$$

where $j,k = 0,1,\ldots,N-1$ and $\delta_{j,k}$ is calculated modulo N (Brown labeled rows and columns by $j,k = 1,2,\ldots,N$).

Zak[9] considered a covering group of the translation group consisting of operators

$$\tau(\mathbf{R}|\mathbf{R}_1,\ldots,\mathbf{R}_j) = \hat{T}(\mathbf{R})\exp[2\pi i\phi(\mathbf{R}_1,\ldots,\mathbf{R}_n)/\varphi_0], \tag{11}$$

where $\Sigma_i\mathbf{R}_i = \mathbf{R}$ and $\phi(\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_j)$ is the flux of the magnetic field through a polygon enclosed by a loop consisting of the vectors $\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_j,-\mathbf{R}$. The periodicity condition was the same as (8) for even N but for odd N Zak proved that the condition

$$H = 2\frac{l}{N}\varphi_0 \tag{12}$$

should be satisfied. This condition implies that for even N the number of different factors is $2N$, whereas it equals N for odd N.[9,10] The result obtained by Zak agreed with Azbel's considerations[19] who showed that wave functions had to be periodic functions of H with the period $2\varphi_0$. It is worthwhile noting that Azbel also worked with the antisymmetric gauge. Zak did not introduce a factor system in an explicit way (it was not necessary in his constructions of representations) but

it can be easily found by considering multiplication of coset representatives $\tau(\mathbf{R}|\mathbf{R})$,[17,20] which simply are equal to $\hat{T}(\mathbf{R})$ [see (11)]. Therefore the factor system is also given by (9), but now it is the factor system of the covering group being a central extension so it should be denoted as $\mu_l(\mathbf{R},\mathbf{R}')$. Zak[10] also showed that matrix elements of an irreducible N-dimensional representation should be [only the coset representatives $\tau(\mathbf{R},\mathbf{R})$ are taken into account here][21]

$$ {}^{l}_{2}D_{jk}[\tau([X,Y]|[X,Y])]=\exp\left[2\pi i\frac{lX}{bN}(Y+2k)\right]\delta_{j,k+Y}, \tag{13}$$

where $b=1,2$ for N odd and even, respectively. It is obvious that this representation corresponds to the irreducible representation $\Gamma(\alpha)=\alpha$ of the factor group F, so $m_l(\mathbf{R},\mathbf{R}')=\mu_l(\mathbf{R},\mathbf{R}')$. According to (8) and (12) changes of H are related to changes of l but they were interpreted in different ways. In Brown's considerations H determines a factor system of projective representations in a direct way — different values of H satisfying (8) lead to nonequivalent projective representations. On the contrary, Zak considered different (nonequivalent) central extensions of T with factor systems μ_l. However, Zak assumed that only the representations $\Gamma(\alpha)=\alpha$ were physical whereas the others were rejected as nonphysical in further considerations.[10,22] It means, according to Zak, that for a central extension with a factor system μ_l one has to find projective representations ${}^{l}D$ with a factor system $m_l=\Gamma(\mu_l)=\mu_l$. The same result can be obtained while considering only the factor system μ_1 and next all irreducible representations Γ_l of F such that $\Gamma_l(\mu_1)=m_l$. Thus all representations necessary in physical applications, considered by Zak as representations of different although isomorphic groups, can be obtained by use of ''nonphysical'' representations Γ_l with $l>1$. Nevertheless, it seems that the representations introduced by Brown (10) could be used in Zak's approach to construct (vector) representations of T' (finite or not) according to (1). A comparison of (10) and (13) shows that for odd N Brown and Zak used different representations. However, for even N $(b=2)$ we have

$$ {}^{l}_{1}D_{jk}[X,-Y]={}^{l}_{2}D_{jk}[\tau([X,Y]|[X,Y])], \tag{14}$$

where the sign ''$-$'' originates from a different choice of the sign of e assumed by Zak[18] (in Zak's approach eigenvectors of $D[1,0]$ are permuted by $D[0,1]$ in the opposite direction than that assumed in Brown's definition).

The third approach is based on pure group-theoretical considerations and consists in determination of all possible central extensions of a finite group $\mathbb{Z}_N^2$ (in general $\mathbb{Z}_{N_1}\otimes\mathbb{Z}_{N_2}$) by an infinite (U(1)) or finite ($C_N=\{\alpha\in\mathbb{C}|\alpha^N=1\}$) group of factors F.[13,15,16] It was shown, by means of the Mac Lane method, that all nonequivalent factor systems corresponding to finite magnetic translation groups can be written as[15,16]

$$\mu_k([X,Y],[X',Y'])=\exp(2\pi ikYX'/N) \tag{15}$$

with $k=0,1,\ldots,N-1$. Some important facts have to be mentioned:

This formula resembles the Landau gauge $\mathbf{A}=[0,Hx]$; recently it has been shown that this convergence is not accidental.[23]

The fraction k/N can be interpreted as H/φ_0,[15,16] so the resulting numbers constitute a periodic function of H with the period φ_0 in agreement with Brown's result but contrary to Zak's and Azbel's results.

In both Brown's and Zak's approaches a (group-theoretical) commutator of two magnetic translations, corresponding to vectors $[X,Y]$ and $[X',Y']$, is equal to

$$c([X,Y],[X',Y'])=\exp[-2\pi i(XY'-YX')H/\varphi_0] \tag{16}$$

and it is the same as obtained in the cited papers[15,16] if H does not satisfy (12) but (8).

There are no additional conditions imposed on k and on N [i.e., the results are valid for both odd and even N and for $\gcd(k,N)>1$]. Taking into account only parameters $k=l$ mutually prime with N it can be shown that N-dimensional irreducible projective representations of T_N (or ''physical'' vector representations of the extension of T_N by C_N) have the following matrix elements

$$ {}^{l}_{3}D_{jk}[X,Y]=\exp\left(2\pi i\frac{l}{N}Xj\right)\delta_{j,k-Y}. \tag{17}$$

Representations with $\gcd(k,N)>1$ were briefly discussed elsewhere,[22] but the difference between odd and even N was not considered there.

III. PROJECTIVE REPRESENTATIONS — STANDARD AND NORMALIZED FACTOR SYSTEMS

To compare different descriptions of the magnetic translation groups we have to discuss not only projective representations themselves but also their factor systems. To begin with we recall now some definitions related to factor systems and their properties.[17,20,24,25] As one can see factor systems appear in the definition of a projective representation (2) and in the multiplication rule for a central extension of groups (3). Factor systems m_l are determined directly (as in Brown's approach) or via factor systems μ_l for central extensions by means of the ''physical'' representations $\Gamma(\alpha)=\alpha$ and the formula (4).

A factor system $m:T\times T\to\mathbb{C}$ has to satisfy the following condition[20,25]

$$m(\mathbf{R},\mathbf{R}')m(\mathbf{R}+\mathbf{R}',\mathbf{R}'')=m(\mathbf{R}',\mathbf{R}'')m(\mathbf{R},\mathbf{R}+\mathbf{R}'') \tag{18}$$

for all $\mathbf{R},\mathbf{R}',\mathbf{R}''$. A *trivial* factor system $t(\mathbf{R},\mathbf{R}')$ is determined by any mapping $f:T\to\mathbb{C}$ according to

$$t(\mathbf{R},\mathbf{R}')=f(\mathbf{R})f(\mathbf{R}')/f(\mathbf{R}+\mathbf{R}'). \tag{19}$$

Since T is Abelian then each trivial factor system is symmetric, i.e., $t(\mathbf{R},\mathbf{R}')=t(\mathbf{R}',\mathbf{R})$. If

$$m'(\mathbf{R},\mathbf{R}')=t(\mathbf{R},\mathbf{R}')m(\mathbf{R},\mathbf{R}') \tag{20}$$

then factor systems m and m' are called *equivalent*. Notice, however, that the projective representations determined by equivalent factor systems are *nonequivalent*.[17] Since all factor systems for a given T form an Abelian group Φ and a set of trivial factor systems Θ is its normal subgroup then elements of the factor group $M=\Phi/\Theta$ (known as the *Schur multiplicator*) correspond to representatives of classes of an equivalent factor system. A factor system is called *standard* if it satisfies

$$m(\mathbf{R},\mathbf{0})=m(\mathbf{0},\mathbf{R})=1,\quad \forall\mathbf{R}. \tag{21}$$

A factor system of an N-dimensional projective representation is *normalized* if

$$m(\mathbf{R},\mathbf{R}')\in C_N,\quad \forall\mathbf{R},\mathbf{R}' \tag{22}$$

(i.e., each factor is the Nth root of 1).

It is well known that the Schur multiplicator of $\mathbb{Z}_N^2$ is C_N,[26] so the factor system (15) is normalized and standard since $m_k([0,0],[X,Y])=1$. Moreover, it is periodic with respect to Y and X' — the period is equal to N. In particular we have $m_k([N,0],[X,Y])=m_k([X,Y],[N,N])=1$, etc. On the other hand, the factor system (9) of N-dimensional representations (10) or (13) is not normalized because some of factors do not belong to C_N but to C_{2N} instead. It also means that this system is standard, because $m_l([0,0],[X,Y])=1$, but it appears that N does not serve as a period because, for example, $m_l([0,N],[1,0])=\exp(\pi il)=(-1)^l$. This fact stirs up a conflict between the conditions obtained by Brown and those obtained by Zak and, moreover, leads to difficulties in studying magnetic translations for the antisymmetric gauge. Of course, one may work with factor systems (and, hence, representations) which are neither standard nor normalized, but such considerations have to be carried very carefully and results obtained have to be carefully interpreted, too.[17] Brown and Zak did not check normalization of their factor systems and this led to ambiguity of their results [cf. (8) and (12)].

At first let us notice that Brown took into account one requirement only, namely[8]

$$\hat{T}(N\mathbf{a}_j)\hat{T}(\mathbf{R})=\hat{T}(\mathbf{R})\hat{T}(N\mathbf{a}_j), \tag{23}$$

i.e., that $\hat{T}(N\mathbf{a}_j)$ would commute with any other operator. On the other hand, Zak demanded in addition that[9]

$$m(N\mathbf{a}_j, N\mathbf{a}_k) = 1\,, \tag{24}$$

i.e., that ${}^lD(N\mathbf{a}_j)$ should behave as a constant factor. As follows from (9)

$$m_l([N,0],[0,N]) = (-1)^{lN},$$

so for odd N the magnetic field H has to be twice as high as in (8). Note that representations (17), corresponding to the Landau gauge, satisfy, for both odd and even N, the following stronger condition

$${}^l_3D(N\mathbf{a}_j)\,{}^l_3D(\mathbf{R}) = {}^l_3D(\mathbf{R}+N\mathbf{a}_j) = {}^l_3D(\mathbf{R}), \tag{25}$$

i.e., both $N\mathbf{a}_j$ and $\hat{T}(N\mathbf{a}_j)$ are equal to the unit element in the translation group T_N and in the group of magnetic translation operators, respectively. The condition (12) (for odd N) removes these problems but, however, leads to another question why odd and even N should be considered separately while both cases can be evidently treated as one with the use of the Landau gauge.

While considering restrictions imposed on H by the periodic boundary conditions with the Landau gauge, i.e., the standard factor system (15), one can see that the condition $H = k\varphi_0/N$ is sufficient. So, it seems that Brown's approach is well-supported. Moreover, it should be noted that Zak weakened his requirements and later on he considered only Brown's condition.[27] To enlighten the problem we have to check whether the factor systems (9) and (15) are equivalent or not. At first note that the group-theoretical commutator of operators of any projective representations (of an Abelian group T) is equal to

$$c(\mathbf{R},\mathbf{R}') = D(\mathbf{R})D(\mathbf{R}')[D(\mathbf{R}')D(\mathbf{R})]^{-1} = \frac{m(\mathbf{R},\mathbf{R}')}{m(\mathbf{R}',\mathbf{R})}, \tag{26}$$

then it is the same for all equivalent factor systems. Since the equivalence of factor systems means the equivalence of vector potentials (gauges),[2,23] then the above commutator does not depend on **A** but rather on **H**, and in this sense this commutator only (not a factor system) has the physical meaning — if $D(\mathbf{R})$ represents a lattice translation in the presence of a magnetic field then the commutator corresponds to a loop determined by vectors $\mathbf{R},\mathbf{R}',-\mathbf{R},-\mathbf{R}'$ and its value depends on the flux through a nonprimitive, in general, cell determined by these vectors. So, Brown's requirement (23) leading to the condition (8) was based on a reasonable assumption. However, the factor system considered was not normalized which led to a disagreement with Zak's results.

IV. EQUIVALENCE OF FACTOR SYSTEMS AND REPRESENTATIONS

Let us consider a mapping $f_w[X,Y] = \exp(2\pi i w XY)$, $w \in \mathbb{R}$, which determines the following trivial factor system

$$t_w([X,Y],[X',Y']) = \exp[-2\pi i w(XY' + YX')]. \tag{27}$$

The factors obtained belong to C_N, i.e., t is standard and normalized, if $w = j/N$. For example for $j = k$ the factor system (15) is transformed to

$$\overline{\mu_k}([X,Y],[X',Y']) = (\mu_k \circ t_{k/N})([X,Y],[X',Y']) = \exp(-2\pi \mathrm{i} k XY'/N), \tag{28}$$

which corresponds to another form of the Landau gauge $\overline{\mathbf{A}} = [-Hy,0]$. It is important that if t is not normalized then a new factor system $m' = tm$ is not normalized, too. This is, however, the case which leads to the factor system (9) determined by Brown and Zak — one has to put $w = k/2N$. This, and the previous discussion on the commutator, proves that the stronger condition introduced by Zak following from (24) is superfluous. It can be easily shown for odd N, since for l mutually

prime with odd N also $\gcd(2l,N)=1$ (the mapping $l\mapsto 2l$ is an automorphism of $\mathbb{Z}_N$ which changes the order of elements only). Therefore in the formulas obtained by Brown, (8)–(10), one can replace $l<N$ in the following way

$$l=\begin{cases} 0, \text{ for } l=0 \\ 2k=2l', \text{ for even } l\neq 0 \\ 2k-1=2(k+N')=2l', \text{ for odd } l, \end{cases} \tag{29}$$

where $N'=(N-1)/2$, $k=1,2,\ldots,N'$, and $l'=1,2,\ldots,2N'=N-1$. In this way a relation similar to (14) is obtained

$${}^{l}_{1}D_{jk}[X,-Y]={}^{l'}_{2}D_{jk}[\tau([X,Y]|[X,Y])], \tag{30}$$

where l and l' are interrelated by (29). In the same way one can transform the factor system (15) into (9). If $\gcd(l,N)=1$ then l is replaced by l', so

$$m_l([X,Y],[X',Y'])=\exp[2\pi i(2l')YX'/N] \tag{31}$$

and next w is taken to be l'/N. The factor system obtained

$$t_w m_l([X,Y],[X',Y'])=\exp[\pi i l(YX'-XY')/N] \tag{32}$$

is exactly the same as (9). In a sense, we have performed a ''normalization'' of the factor system used by Brown and Zak. In other words, the projective representations (10) do not satisfy the condition (24) for odd N since they are not given in *normalized* form. Such a form can be obtained by substitution $l\to 2l'$ which leads to the condition (12) determined by Zak.

Anyway, this way of normalization is not possible in the case of even N, since in general $\gcd(l,N)\neq\gcd(2l,N)$. However, we can use a hint given by Zak, who did not exploit it in full. At the end of his paper[9] Zak noticed that a finite magnetic translation group contains N^3 elements[21] for odd N whereas for even N the number of elements is two times bigger. It suggests that a group considered by him was, in fact, an extension of T_N by C_{2N} — the factors obtained were not normalized since they did not belong to the multiplicator of T_N.

In a previous paper[15] it was shown that central extensions of T_N by C_{2N} which correspond to magnetic translation groups have factor systems

$${}_2m_k([X,Y],[X',Y'])=\exp\left(\frac{\pi i}{N}2kYX'\right) \tag{33}$$

with $k=0,1,\ldots,N-1$. A mapping f_w assigns to each $[X,Y]$ an element of C_{2N} so it is well defined for $w=k/2N$, i.e.,

$$f_w[X,Y]=\exp\left(\frac{\pi i}{N}kXY\right).$$

Note that the product kXY is calculated modulo $2N$ and, therefore, f_w is a multivalued function: in $\mathbb{Z}_N$ numbers X and $X+N$ represent the same element, whereas

$$f_w[X+N,Y]=(-1)^{kY}\exp\left(\frac{\pi i}{N}kXY\right)$$

is not equal to $f_w[X,Y]$, in a general case. To calculate a trivial factor system t_w according to (19) one has to determine $f_w(\mathbf{R}+\mathbf{R}')$. Let us assume, at this moment only formally, that a sum of vectors in this formula will *not* be calculated modulo N. Then

$$t_w([X,Y],[X',Y'])=\exp\left[-\frac{\pi i}{N}k(XY'+X'Y)\right] \tag{34}$$

and

$$_2m_k t_w([X,Y],[X',Y'])=\exp\left[\frac{\pi i}{N}k(YX'-XY')\right],$$

which coincides with (9). It means that in order to obtain the above results we have to treat $[X,Y]$ as an element of T_{2N} rather than that of T_N. So, in fact, we have considered a larger, $2N\times 2N$, lattice although a parameter labeling nonequivalent central extensions has been taken to be equal $2k$. Even for $\gcd(k,N)=1$ we have $\gcd(2k,2N)=2$, so the condition accompanying (8) that l is mutually prime with the period, is not fulfilled in this case. This leads to the magnetic periodicity with a period N though the lattice (crystal) period is $2N$. This problem was briefly discussed by Brown[8] and Zak[10] and its solution is possible through a concept of *magnetic cells* — if $\gcd(l,N)=\lambda>1$ [in (8) and (12)] then the $N\times N$ lattice can be decomposed into $(N/\lambda)\times(N/\lambda)$ magnetically periodic sublattices, which form a $\lambda\times\lambda$ lattice of magnetic cells. In the case considered $\lambda=2$ and the $(2N)\times(2N)$ lattice is decomposed into four $N\times N$ sublattices consisting of points $[X,Y]$, $[X+N,Y]$, $[X,Y+N]$, and $[X+N,Y+N]$, respectively, where $X,Y\in\mathbb{Z}_N$. According to (17) and (33) an irreducible projective representation of T_{2N} should be N-dimensional in this case. It follows from the Burnside's theorem that there are *four* such representations and they can be chosen as

$$_3^lD_{jk}^{\kappa_x,\kappa_y}[X,Y]=(-1)^{\kappa_x\epsilon_x+\kappa_y\epsilon_y}\exp\left(2\pi i\frac{l}{N}Xj\right)\delta_{j,k-Y}, \tag{35}$$

where $\kappa_x,\kappa_y=0,1$ and ϵ_x (ϵ_y) is equal to 0 for $X<N$ ($Y<N$) and to 1 otherwise. Therefore the representations considered by Brown and Zak (14) are equivalent to $_3^lD^{0,0}$. However, the latter is clearly periodic with the period N and satisfies the condition (25). Since the trivial factor system (34) is not normalized in C_N then the representations (14) satisfy the condition (24) and do not satisfy (25).

V. DISCUSSION AND FINAL REMARKS

Summarizing the above discussion on odd and even periods N we can state that the antisymmetric gauge $\mathbf{A}=[Hy,-Hx]/2$, considered by Brown and Zak, corresponds in fact to the crystal period $2N$ and the magnetic period N. If one, like Brown and Zak, does not take this fact into account then results obtained can lead to erroneous conclusions. For example, in this way the additional condition (12) was derived. Investigations of the standard and normalized factor system (15), corresponding to the Landau gauge, have clearly indicated points at which the magnetic translation groups are ''gauge-dependent'' and how one can ''normalize'' factor systems and representations.

The magnetically periodic boundary conditions of projective representations, when the Landau gauge is assumed, can be invoked if

$$H=\frac{l}{N}\varphi_0, \tag{36}$$

where $l=0,1,\ldots,N-1$. [If $\gcd(l,N)=\lambda$ then the magnetic period is equal to N/λ, whereas the crystal period is still N.] The factor system (15) [and also the representation (17) and the physical properties] is a periodic function of H with a period φ_0. Hence the different magnetic response of the considered system can be observed only for N values of $H=l\varphi_0/N$.

If N is an odd integer then $\gcd(2l,N)=\gcd(l,N)$ so the magnetic periodicity is the same in both cases and l in (36) can be replaced by $2l'$, which is equivalent to Zak's condition (12). However, the successively counted values of HN/φ_0 have to be arranged in a different order: $0,\ N+1,\ 2,\ N+3,\ldots,2N-2,\ N-1$. If these values were arranged in the increasing order, i.e., $0,\ 2,\ldots,N-1,\ N+1,\ldots,2N-2$, it might suggest that the condition (12) has to be taken into

account and that the magnetic period is $2\varphi_0$. The only way to settle this problem is by investigation of a system described by a Hamiltonian with a nonperiodic part, e.g., the paramagnetic term.

The case of even N has quite a different nature. As was shown above, the factor systems and representations considered by Brown and Zak describe a lattice with the crystal period $2N$ and the magnetic period equal N. The condition (36) yields $H=l\varphi_0/(2N)$, with $l=0,1,\ldots,2N-1$, but to achieve the magnetic period N only even values of $l=2l'$ are considered, so $H=l'\varphi_0/N$. The representation (10) [see also (14)] is one of four nonequivalent irreducible representations which can be determined in this case. It can be easily seen that Brown's considerations for odd N can also be interpreted in this way [since the decomposition of a $(2N)\times(2N)$ lattice into four $N\times N$ lattices does not depend on the parity of N]. It means, in particular, that the antisymmetric gauge for $N=2$ can be introduced only if one considers the 4×4 lattice with $H=0$ (a trivial case) or $H=\varphi_0/2$.

In this work the descriptions of the magnetic translation group proposed by Brown and Zak were compared with the results obtained by means of the Mac Lane method.[15] The first authors assumed the antisymmetric gauge, whereas the Mac Lane method led to the Landau gauge. Due to a factor $\frac{1}{2}$ in the antisymmetric gauge some problems arise when one introduces the magnetically periodic boundary conditions. More careful considerations put forward by Zak gave the additional condition (12) for an odd period N. However, the condition (36) obtained for the Landau gauge resembles the Brown's condition (8) and does not depend on the parity of N. This condition was obtained from the group-theoretical considerations leading to the factor system (15). In the next step k/N was interpreted as H/φ_0. At first sight it can be interpreted as any value proportional to H, e.g., as $2H/\varphi_0$. However, the first choice is confirmed by the value of group-theoretical commutator, which does not depend on the gauge or, in the other words, is identical for all equivalent factor systems.

Let us also remind that in this work H in fact denotes the magnetic flux through one primitive cell. Therefore, according to (8) or (36), the total flux through the $N\times N$ lattice is equal to

$$\Phi=lN\varphi_0, \tag{37}$$

i.e., to an integer multiplicity of the fluxon. To introduce the antisymmetric gauge one has to consider a $(2N)\times(2N)$ lattice and even $l=2l'$. Hence the total flux equals $\Phi=4l'N\varphi_0$, so the flux through one $N\times N$ magnetic cell is equal to $l'N\varphi_0$, which is consistent with the previous value (37), and the flux trough one primitive cell is equal to $H_a=l'\varphi_0/N$. On the other hand, the flux through one primitive cell of the $(2N)\times(2N)$ lattice (assuming the Landau gauge) is $H_L=l\varphi_0/(2N)$, so in general it is a half of H_a. Therefore we can set up the certain procedure: For a given magnetic field H_a the antisymmetric gauge can be introduced if the magnetically periodic boundary conditions admit two times smaller H_L. For the sake of illustration let us consider the $(2N)\times(2N)$ lattice and $H=\varphi_0/N$. Then, from (16), one obtains a commutator corresponding to the primitive vectors $[1,0]$ and $[0,1]$ as

$$c([1,0],[0,1])=\exp(-2\pi i/N).$$

The formula (9) gives the following values of the corresponding factors [see (26)]

$${}_1m([1,0],[0,1])=\exp(-\pi i/N)$$

and

$${}_1m([0,1],[1,0])=\exp(\pi i/N),$$

whereas (15) leads to

$${}_2m([1,0],[0,1])=1$$

and

$${}_2m([0,1],[1,0])=\exp(2\pi i/N).$$

So, the flux through the primitive cell, corresponding to the commutator and independent of the gauge, was decomposed in two different ways into fluxes through ''primitive'' triangles. However, the first decomposition (related to the antisymmetric gauge) is not possible for the minimal flux $H=\varphi_0/(2N)$. If considering any other trivial factor system (27) determined by the parameter $w\in\mathbb{R}$ one can obtain many other decompositions of the commutator into factors. It can be viewed as the decomposition of the flux through the primitive cell into fluxes through the ''lower'' and ''upper'' triangle. In particular, the other Landau gauge, corresponding to the factor system (28), changes roles of these triangles since one obtains

$$\overline{m}([1,0],[0,1])=\exp(2\pi i/N)$$

and

$$\overline{m}([0,1],[1,0])=1\,.$$

Despite the fact that the physical properties are gauge-independent we have noticed that the form of the vector potential **A** has a certain importance in the mathematical description of a system. One has to be especially careful considering projective representations or extensions of groups, since some equivalent factor systems are neither standard nor normalized. However, it may occur that in certain applications or in other descriptions of the same problem it is more convenient to use such a form of **A**.

[1] K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980); For recent reviews see, e.g., *The Quantum Hall Effect,* edited by R. E. Prange and S. M. Grivin (Springer, New York, 1990); *Quantum Hall Effect,* edited by M. Stone (World Scientific, Singapore, 1992); M. Janssen, O. Viehweger, U. Fastenrath, and J. Hajdu, *Introduction to the Theory of the Integer Quantum Hall Effect* (VCH, Weinheim, 1994).
[2] B. Huckestein, Rev. Mod. Phys. **67**, 357 (1995).
[3] Two-dimensional electron systems are reviewed in T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).
[4] See, e.g., recently published papers and the references therein: P. Kleinert and V. V. Bryskin, Phys. Rev. B **55**, 1469 (1997); V. A. Geyler and V. A. Margulis, *ibid.* **55**, 2543 (1997); Q. W. Shi and K. Y. Szeto, *ibid.* **55**, 4558 (1997).
[5] A. Kol and N. Read, Phys. Rev. B **48**, 8890 (1993); O. Steffens, M. Suhrke, and P. Rotter, *ibid.* **55**, 4486 (1997); E. I. Rashba, L. E. Zhukov, and A. L. Efros, *ibid.* **55**, 5306 (1997); K. Ishikawa, N. Maeda, T. Ochiai, and H. Suzuki, cond-mat/9704023.
[6] R. Peierls, Z. Phys. **80**, 763 (1933).
[7] L. Landau, Z. Phys. **64**, 629 (1930).
[8] E. Brown, Bull. Am. Phys. Soc. **8**, 256 (1963); Phys. Rev. **133**, A1038 (1964).
[9] J. Zak, Phys. Rev. **134**, A1602 (1964).
[10] J. Zak, Phys. Rev. **134**, A1607 (1964).
[11] J. Zak, Phys. Rev. **136**, A776 (1964).
[12] More detailed discussion on central extensions was presented in P. P. Divakaran, A. K. Rajagopal, Physica C **176**, 457 (1991); P. P. Divakaran, Rev. Math. Phys. **6**, 167 (1994); M. S. Raghanathan, *ibid.* **6**, 207 (1994).
[13] P. P. Divakaran and A. K. Rajagopal, Int. J. Mod. Phys. B **9**, 261 (1995).
[14] The most spectacular and the best known is the Aharonov–Bohm effect [Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959); Y. Aharonov and J. Anandan, Phys. Rev. Lett. **58**, 1593 (1987)] but some new ideas have been presented lately by F. Ghaboussi, quant-ph/9702054, cond-mat/9701128, cond-mat/9703080, cond-mat/9705108.
[15] W. Florek, Rep. Math. Phys. **34**, 81 (1994).
[16] W. Florek, Rep. Math. Phys. **38**, 235 (1996).
[17] S. L. Altmann, *Induced Representations in Crystal and Molecules* (Academic, London, 1977).
[18] In this work it is assumed, as in Brown's paper, that the charge of an electron is $-e$, i.e., $e>0$. Zak used $e<0$, so some of his formulas differ in sign from Brown's ones, but in fact are identical. However, we use **H** to denote the magnetic field, as Zak did, while Brown used **B**.
[19] M. Ya. Azbel, Zh. Eksp. Teor. Fiz. **44**, 980 (1963) [Sov. Phys. JETP **17**, 665 (1963)].
[20] A. O. Barut and R. Rączka, *Theory of Group Representations and Applications* (PWN-Polish Scientific, Warsaw, 1977).
[21] Since a two-dimensional lattice is considered then a term m_3n_3/N is omitted in (13) as compared with Zak's formula and the finite magnetic translation group consisting of bN^3 elements is considered [b is defined in the text below (13)].
[22] W. Florek, Phys. Rev. B **55**, 1449 (1997).
[23] W. Florek, Acta Phys. Pol. A **92**, 399 (1997).
[24] C. W. Curtis and I. Reiner, *Representation Theory of Finite Groups and Associative Algebras* (Interscience, New York, 1962).
[25] A. G. Kurosh, *Group Theory* (Chelsea, New York, 1960).
[26] A. Babakhanian, *Cohomological Methods in Group Theory* (Marcel Dekker, New York, 1972).
[27] J. Zak, Phys. Rev. **136**, A1647 (1964).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Renormalized contact potential in two dimensions

R. J. Henderson and S. G. Rajeev
Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627

(Received 14 April 1997; accepted for publication 22 September 1997)

We obtain for the attractive Dirac δ-function potential in two-dimensional quantum mechanics a renormalized formulation that avoids reference to a cutoff and running coupling constant. Dimensional transmutation is carried out before attempting to solve the system, and leads to an interesting eigenvalue problem in $N-2$ degrees of freedom (in the center of momentum frame) when there are N particles. The effective Hamiltonian for $N-2$ particles has a nonlocal attractive interaction, and the Schrodinger equation becomes an eigenvalue problem for the logarithm of this Hamiltonian. The three-body case is examined in detail, and in this case a variational estimate of the ground-state energy is given. © *1998 American Institute of Physics.* [S0022-2488(98)01702-2]

I. INTRODUCTION

Scale invariance, and the ultraviolet divergences for which it is responsible, is an essential feature of the quantum field theories, QCD and electroweak, that comprise the standard model of elementary particle interactions. The divergences that are obtained in the process of quantizing the classical field theories on which the quantum dynamics are based can be removed via the mathematical procedure of renormalization. In condensed matter physics renormalization techniques are used to obtain mathematical models of physical phenomena (e.g., phase transitions) which, despite the presence of a characteristic scale, provided by a lattice spacing, are scale independent in nature. In contrast, renormalization in the context of elementary particle interactions is necessary to make the fundamental theory physically sensible.

That an awkward renormalization procedure is necessary to permit the quantum field theory paradigm to successfully describe elementary particle interactions might be viewed as evidence that quantum field theory is not the proper framework for this problem. Some other exotic and finite theory might be more conceptually accurate and less mathematically cumbersome. On the other hand, one may take Wilson's point of view[1] on the role of renormalization in particle physics: a renormalizable quantum field theory may be viewed as an *effective model* which approximates for low enough energies (or long enough distances) a more fundamental and comprehensive theory. In addition, then, to seeking a more fundamental theory underlying renormalizable quantum field theories, we may aspire to a deeper understanding of the very successful effective theories we have.

Toward this end we take the point of view that renormalizable interactions might be given a *finite formulation*,[2] which avoids the need for renormalization altogether, without the necessity of discarding the framework of quantum field theory. The *renormalized interactions* that would be part of this formulation could not be scale invariant; the scaling symmetry would be broken explicitly at the outset, and not through the renormalization procedure. Other properties of the erstwhile ''fundamental'' interactions might be modified as well.

This approach to finding a finite, effective theory of particle interactions would be tantamount to a reordering of the conventional analysis. It would require renormalizing the theory completely *before* embarking on efforts to solve it, rather than renormalizing in parallel with the finding of solutions. No nonphysical cutoffs or running parameters would appear in the formulation or solutions of the renormalized theory. Rather, such a theory could be formulated as a well-posed mathematical problem: a set of differential equations with appropriate boundary conditions, for example. If this point of view is the correct one, then the reason we cannot write down such a formulation of QCD or electroweak theory is not because they are only low-energy effective theories (though they may be), but rather because we have not yet achieved a deep enough

 © 1998 American Institute of Physics

understanding of these theories to write them down in the simplest way.

Although we believe finding a finite formulation of renormalizable interactions to be a worthwhile goal, attacking this problem directly, in say QCD, appears too formidable at this time. We have in the past, though, found examples of simpler, asymptotically free renormalizable theories that lend themselves to this approach. In Ref. 3 we renormalized the large-N limit of the $1+1$-dimensional non-Abelian Thirring (or Gross–Neveu) model before finding some exact solutions. The scale-breaking renormalized interaction manifested itself as a restriction on the domain of the Hamiltonian operator. A similar role for the Hamiltonian domain was found in Ref. 4 in the study of a quantum mechanical system of two particles in two dimensions attracted by a Dirac δ-function potential. The same system was also examined in the path integral picture, where it was found that the renormalized interaction appeared as a subtle modification to the Wiener measure.

In this paper we continue with our investigation of quantum mechanical particles interacting through an attractive Dirac δ-function potential in two dimensions. We first work out the case of three particles in detail, and eventually extend our ideas to the N-body case.

II. REVIEW OF THE TWO-BODY PROBLEM

Before embarking in the next section on the three-body problem, let us recall the renormalized formulation, in the Hamiltonian picture, of the two-body problem. In the two-body case, after separating out the center of mass coordinate, the original Schrodinger equation in configuration space is

$$-2\Delta\Psi_\lambda(\overline{x})-g\,\delta^2(\overline{x})\Psi_\lambda(\overline{x})=\lambda\Psi_\lambda(\overline{x}), \tag{1}$$

where Δ is the two-dimensional Laplacian, and g is a positive, dimensionless coupling constant. We have taken the masses of the two particles to be $m_1=m_2=\frac{1}{2}$ and chosen units such that $\hbar=1$. In momentum space the Schrodinger equation reads

$$2p^2\Psi_\lambda(\overline{p})-\frac{g}{(2\pi)^2}\int d^2p\,\Psi_\lambda(\overline{p})=\lambda\Psi_\lambda(\overline{p}). \tag{2}$$

This eigenvalue problem, however, is nonphysical since, due to scale invariance, the presence of even a single negative energy solution implies a continuum of negative energy states extending down to minus infinity. On the other hand, if one attempts to restrict the domain of the Hamiltonian to the positive energy sector only, one finds that a complete set of eigenstates cannot be found: there will be no zero angular momentum eigenstate.[2] The Hamiltonian cannot be self-adjoint under such a choice of domain. (The situation is analogous to the presence of solutions to the Dirac equation of unbounded negative energy, and the impossibility of ignoring or ''throwing away'' these states.)

This system's illness can be cured via renormalization. First one regularizes the system by introducing a momentum cutoff (upper bound). Then the coupling constant, g, is required to depend on the cutoff in such a way that the ground-state energy of the regularized system remains finite as the cutoff is removed, i.e., taken to infinity, This procedure removes g from the problem, replacing it with a parameter having dimensions of energy, a trade sometimes called ''dimensional transmutation.'' This new parameter is arbitrary, characterizes the strength of the interaction, and can be taken to be the ground-state energy of the two-body system.

The customary way to obtain renormalized solutions to the problem (see Refs. 5–9), then, is to solve the regularized system first, and then take the limits of solutions (wave functions, scattering amplitudes, etc.) as the cutoff is taken to infinity. In Ref. 4 we showed that an equivalent, but simpler, formulation is given by the following two equations:

$$2p^2\Psi_\lambda(\overline{p})-\lim_{p\to\infty}2p^2\Psi_\lambda(\overline{p})=\lambda\Psi_\lambda(\overline{p}) \tag{3}$$

and

$$\int d^2p \left(\Psi_\lambda(\overline{p}) - \frac{\eta_{\Psi_\lambda}}{2p^2+\mu^2} \right) = 0, \tag{4}$$

where $\eta_\lambda \equiv \lim_{p\to\infty} 2p^2\Psi_\lambda(\overline{p})$.

These two equations give an example of what we mean by a *finite formulation* of a renormalizable theory. The theory at this stage has been renormalized, and can be treated as a well-posed mathematical problem. The first equation is recognized as a renormalized version of the Schrodinger equation. The interaction has become the term $-\eta_{\Psi_\lambda} = -\lim_{p\to\infty} 2p^2\Psi(\overline{p})$ on the left-hand side. This term is the *renormalized interaction.* Wave functions for which η_{Ψ_λ} is zero do not take part in the interaction. Interestingly, there is no adjustable parameter in the Hamiltonian operator appearing in this equation. In fact, the appropriate Hamiltonian in configuration space is the Laplacian with no interaction term at all: the deviation from the free theory is contained entirely in the specification of the Laplacian's domain.

The second equation specifies the domain of the Hamiltonian. In configuration space it is a local condition that implies that wave functions with zero angular momentum diverge logarithmically at the origin; in configuration space, then, the interaction appears as this boundary condition on wave functions. The parameter μ^2 has dimensions of energy, and can be picked arbitrarily. For any choice of $\mu^2 > 0$, the Hamiltonian will be self-adjoint (and in fact is a self-adjoint extension of the free Hamiltonian[10]). Choosing μ^2 corresponds to selecting the strength of the attractive interaction and it turns out that $-\mu^2$ is the ground-state energy, and that the free theory corresponds to letting $\mu^2 = 0$.

The above pair of equations can be solved completely to yield all the energy eigenstates in the two-body case.[2,4] Although we cannot analytically solve the N-body, or even the three-body, system, we can formulate the problem in a similar way, finding the analogs of Eqs. (3) and (4), which can then be solved approximately.

III. RENORMALIZING THE THREE-BODY PROBLEM

The original three-body Schrodinger equation in configuration space is, when $m_1 = m_2 = m_3 = 1/2$ and $\hbar = 1$,

$$(-\Delta_1 - \Delta_2 - \Delta_3)\Psi_\lambda - g(\delta^2(\overline{x_1} - \overline{x_2}) + \delta^2(\overline{x_2} - \overline{x_3}) + \delta^2(\overline{x_3} - \overline{x_1}))\Psi_\lambda = \lambda\Psi_\lambda, \tag{5}$$

where Δ_i is the two-dimensional Laplacian in coordinate $\overline{x_i}$, $\Psi_\lambda = \Psi_\lambda(\overline{x_1}, \overline{x_2}, \overline{x_3})$, and g is the positive, dimensionless coupling constant. In momentum space this becomes

$$(k_1^2 + k_2^2 + k_3^2)\Psi_\lambda - \frac{g}{(2\pi)^2}\left(\int d^2k_{12}\, \Psi_\lambda + \int d^2k_{23}\, \Psi_\lambda + \int d^2k_{31}\, \Psi_\lambda \right) = \lambda\Psi_\lambda, \tag{6}$$

where $\Psi_\lambda = \Psi_\lambda(\overline{k_1}, \overline{k_2}, \overline{k_3})$ and $\overline{k_{ij}} = \frac{1}{2}(\overline{k_i} - \overline{k_j})$.

As in the two-body problem, we work in momentum space, in the center of momentum frame where $K \equiv k_1 + k_2 + k_3 = 0$. Defining $\overline{p_1} \equiv \frac{2}{3}(\frac{1}{2}(k_2 + k_3) - k_1)$ (see Ref. 11 for a discussion of coordinate choices in the general three-body problem), the Schrodinger equation in this frame is

$$\left(2k_{23}^2 + \frac{3}{2}\, p_1^2 \right)\Psi_\lambda - \frac{g}{(2\pi)^2}\left(\int d^2k_{12}\, \Psi_\lambda + \int d^2k_{23}\Psi_\lambda + \int d^2k_{31}\, \Psi_\lambda \right) = \lambda\Psi_\lambda, \tag{7}$$

where we choose the two independent coordinates to be $(\overline{k_{23}}, \overline{p_1})$, such that $\Psi_\lambda = \Psi_\lambda(\overline{k_{23}}, \overline{p_1})$, in which case k_{12} and k_{31} stand for $-\frac{1}{2}k_{23} - \frac{3}{4}\overline{p_1}$ and $-\frac{1}{2}k_{23} + \frac{3}{4}\overline{p_1}$, respectively. We could just as well have chosen $(k_{31}, \overline{p_2})$ or $(k_{12}, \overline{p_3})$ and would have found the same Schrodinger equation, as $2k_{23}^2 + \frac{3}{2}p_1^2 = 2k_{31}^2 + \frac{3}{2}p_2^2 = 2k_{12}^2 + \frac{3}{2}p_3^2$ since the three particles are identical. From now on we will use this symmetry to write this combination simply as $2k^2 + \frac{3}{2}p^2$. The particle singled out by the choice of coordinates [e.g., particle ''1'' with the choice $(\overline{k_{23}}, \overline{p_1})$] is sometimes called the ''spectator'' particle. Then the other two may be referred to as the ''interacting pair,'' even though in reality all three particles interact with each other.

For the same reasons that (2) was found to be nonphysical, (6) also represents an ill-posed physical problem. In this case if one tries to restrict the Hamiltonian domain to positive energy states, one will find that states with zero angular momentum in the $\overline{k_{23}}$ variable are missing.[2] A renormalization of the system is again necessary. By carrying out the renormalization before, and independent of, solving the system, we can cast the problem in a form analogous to (3) and (4) above.

The first step in our renormalization program is to regularize the system by reconsidering the problem in an artificial momentum space wherein the momenta are bounded above by a cutoff. From Wilson's point of view, this cutoff represents some scale beyond which we are not entitled to apply our low-energy effective theory. Here it is convenient to impose this condition by requiring $2k^2+\frac{3}{2}p^2<\Lambda^2$. We must also allow the coupling constant, g, to depend on Λ. By doing this we will be able to specify that g change with Λ in such a way that the physics predicted by the effective theory we derive is independent of the cutoff. The resulting Schrodinger equation is

$$\left(2k^2+\frac{3}{2}\,p^2-\lambda\right)\Psi_\lambda^\Lambda=\frac{g(\Lambda/\nu)}{(2\pi)^2}\sum_i\ \int^\Lambda d^2k_i\ \Psi_\lambda^\Lambda, \tag{8}$$

where ν is an arbitrary parameter with dimensions of momentum, $\int^\Lambda$ indicates the restriction $2k^2+\frac{3}{2}p^2<\Lambda^2$, the sum has three terms, and from now on k_1, k_2, and k_3 will correspond to, and be used interchangeably with, k_{23}, k_{31}, and k_{12}, respectively.

The presence of Λ breaks scale invariance, such that there is no longer the instability problem preventing us from allowing negative energy states. Also, if we define $f_{\lambda,i}^\Lambda(\overline{p_i})\equiv[g(\Lambda/\nu)/(2\pi)^2]\int^\Lambda d^2k_i\ \Psi_\lambda^\Lambda$, then (8) can be written

$$\left(2k^2+\frac{3}{2}\,p^2-\lambda\right)\Psi_\lambda^\Lambda=\sum_i\ \ f_{\lambda,i}^\Lambda(\overline{p_i}). \tag{9}$$

If we decide to consider Ψ_λ^Λ to be a function of $\overline{k_{23}}$ and $\overline{p_1}$, then it is also useful to note

$$\overline{p_2}\equiv\tfrac{2}{3}(\tfrac{1}{2}(\overline{k_3}+\overline{k_1})-\overline{k_2})=-\overline{k_{23}}-\tfrac{1}{2}\overline{p_1} \tag{10}$$

and

$$\overline{p_3}\equiv\tfrac{2}{3}(\tfrac{1}{2}(\overline{k_1}+\overline{k_2})-\overline{k_3})=\overline{k_{23}}-\tfrac{1}{2}\overline{p_1}. \tag{11}$$

It is left to determine what $g(\Lambda/\nu)$ must be in order that this regularized system will have a well-defined limit as we let $\Lambda\to\infty$. By integrating (8) over any of the k_i it is not difficult to show[2] that $g(\Lambda/\nu)$ must satisfy, for large enough Λ/ν,

$$\frac{(2\pi)^2}{g(\Lambda/\nu)}=\int_{k<\Lambda}\frac{d^2k}{2k^2+\nu^2}, \tag{12}$$

where we keep in mind that ν^2 is arbitrary. Equation (12) implies that $g(\Lambda/\nu)\to 0$ as we let $\Lambda\to\infty$, indicating that the interaction we consider is *asymptotically free*. This expression for g is, with ν^2 replaced by μ^2, precisely what was found during the renormalization of the two-body system.[4] This is a result we should expect. In fact it would be reasonable to fix $g(\Lambda/\nu)$ using the two-body result rather than to rederive it in the three-body case. This correspondence allows us to identify $-\nu^2=-\mu^2$ as the binding energy of the two particle bound state, i.e., the ground-state energy of the two-body system. From now on we will use $-\mu^2$ to denote this quantity.

By taking the $k_{23}\to\infty$ limit of (8) we can make the identification

$$\lim_{k_{23}\to\infty}2k_{23}^2\Psi_\lambda^\Lambda=\frac{g(\Lambda/\mu)}{(2\pi)^2}\int^\Lambda d^2k_{23}\ \Psi_\lambda^\Lambda=f_{\lambda,1}^\Lambda(\overline{p_1}). \tag{13}$$

This equation has counterparts with $\overline{k_{23}}$ replaced by $\overline{k_{31}}$ or $\overline{k_{12}}$, such that (8) could also be written

$$\left(2k^2+\frac{3}{2}\,p^2-\lambda\right)\,\Psi_\lambda^\Lambda=\sum_i\,\lim_{k_i\to\infty}\,2k_i^2\Psi_\lambda^\Lambda. \tag{14}$$

This equation does not contain g, and will remain valid as we let $\Lambda\to\infty$. In this limit, (14) becomes the three-body renormalized Schrodinger equation, comparable to (3) in the two-body case. Again, no coupling parameter appears in the renormalized Hamiltonian. The parameter characterizing the interaction strength, which we continue to take to be the two-body binding energy, will appear in the domain (boundary condition) equation.

We can find the domain equation by combining (12) and (13). Together, as $\Lambda\to\infty$, these equations impose the following conditions on wave functions:

$$\int\,d^2k_i\left(\Psi_\lambda(\overline{k_i},\overline{p_i})-\frac{f_{\lambda,i}(\overline{p_i})}{2k_i^2+\mu^2}\right)=0, \tag{15}$$

where $f_\lambda(\overline{p_i})=\lim_{k_i\to\infty}2k_i^2\Psi_\lambda$. One way of understanding the choice of $g(\Lambda/\nu)=g(\Lambda/\mu)$ in (12) is that g must have the form which ensures that (13) will continue to hold true as we let $\Lambda\to\infty$.

Equation (15) is in fact three equations, one each for $i=1,2,3$. They are integrals over momentum space, but in configuration space they are local conditions dictating that the wave functions either go to zero or diverge logarithmically, at a rate determined by μ, whenever the coordinates of two particles are made to coincide. That the wave function can blow up under these circumstances is unusual, but not proscribed since the singularities are square integrable.

We thus find that the three-body analogs of the two-body equations (4) and (3) are (15) and the three-body renormalized Schrodinger equation:

$$\left(2k^2+\frac{3}{2}\,p^2-\lambda\right)\Psi_\lambda=\sum_i\,f_{\lambda,i}(\overline{p_i}). \tag{16}$$

The equations (15) and (16) comprise the renormalized, finite formulation of the three-body problem we seek. The form of these equations indicates that the unknown parts of the wave functions are essentially their limits as the relative momenta of particles are taken to infinity, i.e., the functions $f_{\lambda,i}$, $i=1,2,3$. We will require the particles to satisfy Bose statistics, in which case Ψ_λ must be symmetric under permutations of particle indices. Then the $f_{\lambda,i}$ must all be the same function, which we will call f_λ. f_λ is a function of one momentum variable rather than two, and so the eigenvalue problem has already been simplified. The analogous quantity, $\eta_{\Psi_\lambda}\equiv\lim_{p\to\infty}2p^2\,\Psi_\lambda(\overline{p})$, in the two-body case was simply a constant which could be fixed by normalization, so that the eigenvalue problem could be trivially solved. In the N-body case we will find an eigenvalue problem in $N-2$ degrees of freedom in the center of momentum frame.

Although the crux of our problem in solving the system is finding the functions f_λ for all eigenvalues λ, (15) and (16) contain not only f_λ but the wave function Ψ_λ as well. It would be convenient to eliminate Ψ_λ from the problem altogether, and obtain the equation satisfied by f_λ alone. This we can do by employing the Lippmann–Schwinger formulation[12] of the Schrodinger equation. The Lippmann–Schwinger approach is conventionally reserved for scattering states, but here we will find it a convenient starting point in both the scattering and bound state sectors of the theory.

The Lippmann–Schwinger equation for our three-body system can be written down by inspection using (16). It is

$$\Psi_\lambda=g(\overline{p_1},\overline{k_{23}})\,\delta\left(2k^2+\frac{3}{2}\,p^2-\lambda\right)+\frac{\Sigma_i f_\lambda^\pm(\overline{p_i})}{2k^2+\frac{3}{2}p^2-\lambda\mp i\epsilon}. \tag{17}$$

We choose for concreteness to write (17) in the basis $(\overline{k_{23}},\overline{p_1})$. $g(\overline{p_1},\overline{k_{23}})$ is arbitrary, excepting the fact that it must be symmetric under permutations of particle indices.

The energy $\epsilon>0$ in (17) is infinitesimally small. It must appear when $\lambda>0$ to make the division by the singular operator $2k^2+\frac{3}{2}p^2-\lambda$ well defined. The choice of sign, $\pm$ in the denominator corresponds to a choice of boundary conditions: the upper sign implying outgoing scattered

waves (which is ordinarily the physical case), and the lower sign meaning converging scattered waves. Eventually ϵ can be taken to zero, but will serve to regulate otherwise divergent integrals in intermediate steps. We add the label $\pm$ to $f_\lambda^\pm$ to signify this choice of boundary conditions.

The first term in (17) represents the kernel of the operator $2k^2+\frac{3}{2}p^2-\lambda$, which appears on the left-hand side of (16). It is a solution to the *free* Schrodinger equation. In the case $\lambda>0$ it represents the unscattered portion of the wave, and in configuration space gives the wave function its asymptotic behavior as particles become infinitely separated. When $\lambda<0$ this term disappears. Negative energy states are comprised of only the second term, and their form is specified by (17) up to the unknown function f_λ. All information contained in (16) is also in (17).

At this point we must part ways with the usual Lippmann–Schwinger analysis. No potential appears in (17), and we cannot use the self-consistency of this equation alone to solve for the wave function. In place of the usual term involving the potential stands the unknown function $f_\lambda^\pm$. The only further information we can glean about $f_\lambda^\pm$ from (17) comes from taking the limit of this equation as $k_{23}\to\infty$. Since $f_\lambda^\pm(\overline{p_1})=\lim_{k_{23}\to\infty} 2k_{23}^2\Psi_\lambda$, taking this limit and using (10) and (11) gives

$$\lim_{k_{23}\to\infty}\left(f_\lambda(\overline{k_{23}})+f_\lambda(-\overline{k_{23}})\right)=0. \tag{18}$$

However, the bosonic symmetry requires that $f_\lambda(-\overline{k_{23}})=f_\lambda(\overline{k_{23}})$, so that (18) becomes the simple asymptotic condition:

$$\lim_{k_{23}\to\infty} f_\lambda(\overline{k_{23}})=0. \tag{19}$$

To gain more information about f_λ, we need to use (15). This once more highlights the fact that, in this system, the interaction between particles is encoded in the boundary conditions on wave functions rather than in a conventional potential energy term. Inserting (17) into (15) and performing integrations where possible yields the following implicit equation for f_λ:

$$\left(\ln\left(\frac{1}{\mu^2}\left|\lambda-\frac{3}{2}p^2\right|\right)\mp i\pi\Theta\left(\lambda-\frac{3}{2}p^2\right)\right)f_\lambda^\pm(\overline{p})-\frac{2}{\pi}\int d^2k\,\frac{f_\lambda^\pm(\overline{k})}{[k^2+p^2+\overline{p\cdot k}-(\lambda/2)\mp(i\epsilon/2)]}$$
$$=\frac{4}{\pi^2}\int d^2k\; g(\overline{p,k})\,\delta\left(2k^2-\frac{3}{2}p^2-\lambda\right). \tag{20}$$

This ungainly looking equation contains all the information about wave functions and energy eigenvalues we need. Imposing the requirement that $f_\lambda^\pm$ has asymptotic behavior such that the integral in (20) converges will ensure (18) holds true. When $\lambda>0$, this equation is an inhomogeneous linear equation in $f_\lambda^\pm$. Operating on $f_\lambda^\pm$ on the left-hand side of (20) is a symmetric integral operator, which must be inverted to get $f_\lambda^\pm$ for any given g. When $\lambda<0$, the inhomogeneous term proportional to g vanishes identically, leaving an eigenvalue problem for $f_\lambda^\pm$ and λ. In the following section we partially solve the problem of finding $f_\lambda^\pm$ for positive and negative energies, and show that the negative energy sector yields an explicit form for a renormalized, nonlocal effective Hamiltonian.

IV. TOWARD SOLVING THE THREE-BODY PROBLEM

All solutions to the *free* Schrodinger equation have positive energy and can be categorized by the following choice of basis:

$$\Psi_\lambda^{\text{Free}}=g(\overline{p_1},\overline{k_{23}})\,\delta(2k^2+\tfrac{3}{2}p^2-\lambda), \tag{21}$$

where

$$g(\overline{p_1},\overline{k_{23}})=\sum_\Pi\;\delta^2\left(\overline{k_{23}}-\frac{1}{2}\left(\overline{\kappa_{\Pi(2)}}-\overline{\kappa_{\Pi(3)}}\right)\right)\delta^2\left(\overline{p_1}-\left(\overline{\kappa_{\Pi(2)}}+\overline{\kappa_{\Pi(3)}}\right)\right). \tag{22}$$

The constant vectors $\overline{\kappa_1}$, $\overline{\kappa_2}$, and $\overline{\kappa_3}$ are the momenta of the three particles and label the states. There are no restrictions on them other than they represent the particles in the center of momentum frame wherein $\overline{\kappa_1}+\overline{\kappa_1}+\overline{\kappa_3}=0$. The energy is $\lambda=\kappa_1^2+\kappa_2^2+\kappa_3^2$. The sum in (22) is over the six permutations of (1,2,3), and gives g the bosonic symmetry we require. For our purposes it is convenient to absorb the k_{23} dependence into p_1 using the delta function in (21) and to diagonalize the angular momentum in the angle, θ_{23}, of k_{23}. In this case g becomes

$$g_n(\overline{p_1},\theta_{23})=e^{in\theta_{23}}\sum_{\Pi} e^{-in\theta_{\Pi(2),\Pi(3)}}\delta\left(k_{23}^2-\frac{1}{4}(\overline{\kappa_{\Pi(2)}}-\overline{\kappa_{\Pi(3)}})^2\right)\delta^2(\overline{p_1}-(\overline{\kappa_{\Pi(2)}}+\overline{\kappa_{\Pi(3)}})), \tag{23}$$

where n is an integer, and $\theta_{i,j}$ denotes the angle of $\frac{1}{2}(\overline{\kappa_i}-\overline{\kappa_j})$.

Recall that in the two-body problem only those states carrying zero angular momentum participated in the interaction. The wave functions with nonzero angular momentum were just the free ones. We find a similar situation in the three-body case. The free wave functions

$$\Psi_{n,\lambda}=g_n(\overline{p_1},\theta_{23})\,\delta(2k^2+\tfrac{3}{2}p^2-\lambda) \tag{24}$$

for $n\neq 0$ are the positive energy solutions we seek in the sector of the Hilbert space where the angular momenta of any pair of particles is nonzero. Clearly (24) solves (16) with $f_\lambda=0$, for any n. For $n=0$, due to the integration over θ_{23}, these wave functions also satisfy (15) for $f_\lambda=0$. Note that the right-hand side of (20) becomes zero in this case. This is consistent with the linear operator on the left-hand side being nonsingular.

Free wave functions with quantum number n not equal to zero do not scatter. Nontrivial, i.e., interacting, wave functions can therefore be taken to have the form (17), with $g(\overline{p_1},\overline{k_{23}})\to g_0(\overline{p_1})$. g_0 is the symmetrized $[g_0(\overline{p_1})=g_0(\overline{p_2})=g_0(\overline{p_3})]$ function, for $n=0$, given in (23). It is independent of θ_{23}, so the integral on the right-hand side of (20) is trivial, and this equation becomes

$$\left(\ln\left(\frac{1}{\mu^2}\left|\lambda-\frac{3p^2}{2}\right|\right)\mp i\pi\Theta\left(\lambda-\frac{3}{2}p^2\right)\right)f_\lambda^{\mp}(\overline{p})-\frac{2}{\pi}\int d^2k\,\frac{f_\lambda^{\pm}(\overline{k})}{[k^2+p^2+\overline{p\cdot k}-(\lambda/2)\mp(i\epsilon/2)]}$$
$$=\frac{2}{\pi}\Theta\left(\lambda-\frac{3}{2}p^2\right)g_0(\overline{p}). \tag{25}$$

Positive energy scattering solutions $f_\lambda^{\pm}$ exist for all $\lambda>0$ and all g_0. Note that the operator acting on $f_\lambda^{\pm}$ in (25) is rotationally invariant. One can let $f_\lambda^{\pm}(\overline{p})\sim e^{il\theta_p}$, where the integer l is the total angular momentum of the three-particle system, and solve each angular momentum sector independently. It is also worth noting that a rotation in the angle θ_{23} is *not* a symmetry of the interacting Hamiltonian; although scattered wave functions have a free part independent of this angle, the scattered components of these states will in general depend on θ_{23}. Equation (25) is as far as we will take our analysis in the positive energy sector.

For the consideration of negative energy states, let us take $\lambda=-\eta^2$. In this case, (17) gives the form of the wave functions as

$$\Psi_\lambda=\frac{\Sigma_i f_\lambda(\overline{p_i})}{2k^2+\frac{3}{2}p^2+\eta^2}. \tag{26}$$

The infinitesimal parameter ϵ can be set to zero in this case. When $\lambda=-\eta^2$, (25) becomes an eigenvalue problem for f_λ and λ:

$$\ln\left(\frac{1}{\mu^2}\left(\eta^2+\frac{3}{2}p^2\right)\right)f_\lambda(\overline{p})-\frac{2}{\pi}\int d^2k\,\frac{f_\lambda(\overline{k})}{[k^2+p^2+\overline{p\cdot k}+(\eta^2/2)]}=0. \tag{27}$$

For a separable attractive potential [i.e., one having the form $\langle\overline{k}|V|\overline{k'}\rangle=-v(k)v(k')$] the Faddeev integral equation for the T-matrix,[13] written as an equation for the wave function, can be

reduced to an equation in just one variable. Bruch and Tjon[14] have shown that, in the limit in which such a potential is made to have zero range in real space, this equation is precisely the eigenvalue equation we have found in (27).

Our eigenvalue problem can be brought to a more conventional form by writing it in terms of dimensionless variables. We let $\overline{p}=\eta\overline{x}$, $\overline{k}=\eta\overline{y}$, and $f_\lambda(\overline{p})=u_\lambda(\overline{x})$. Then (27) becomes

$$\ln\left(1+\frac{3}{2}x^2\right)u_\lambda(\overline{x})-\frac{2}{\pi}\int d^2y\,\frac{u_\lambda(\overline{y})}{(y^2+x^2+\overline{x\cdot y}+\frac{1}{2})}=\ln\left(\frac{\mu^2}{\eta^2}\right)u_\lambda(\overline{x}). \tag{28}$$

Denoting the operator on the left-hand side of (28) by W, we can rewrite that equation as

$$Wu_\lambda=\ln\left(\frac{\mu^2}{\eta^2}\right)u_\lambda\,. \tag{29}$$

We find, therefore, a linear eigenvalue problem for the eigenfunction u_λ and the eigenvalue $\ln(\mu^2/\eta^2)$. Note that W is a symmetric, nonlocal operator. Its eigenfunctions, u_λ, give us the undetermined part [see (26)] of the wave function, Ψ_λ.

Since the energy is $-\eta^2$, we can identify the operator

$$H\equiv-\mu^2e^{-W} \tag{30}$$

as the Hamiltonian, with one degree of freedom (in addition to the center of mass) effectively integrated out, of the renormalized system in the negative energy sector. We thus find that the logarithm of the effective Hamiltonian of the spectator particle, in momentum space, is an integral operator. The form of the integral operator W reveals that the renormalized, effective ''spectator Hamiltonian'' includes an attractive potential which is nonlocal. It is important to realize, however, that the nonlocality of this Hamiltonian arises because we have effectively integrated out one degree of freedom, a simplification we could make because the actual interaction is of zero range. A positive attribute of the formulation in terms of this one-particle Hamiltonian is that the boundary conditions, and therefore the interactions, are built in, and do not come in through some supplementary condition on wave functions.

The operator W consists of two terms: an unbounded positive kinetic-energy-like operator [a multiplication by $\ln(\frac{3}{2}x^2+1)$ in momentum space] and an interaction term [the integral operator in (28)]. In light of numerical and variational evidence we conjecture, but have not proved, that the interaction part of W is bounded below. Certainly the Hamiltonian including both the ''kinetic'' and ''potential'' energy terms is bounded below. For a proof of this see Ref. 14.

If the effective potential energy term alone is in fact bounded below, it indicates that there are a finite number of normalizable bound states, as well as a continuum of negative energy scattering states wherein a bound state of two particles scatters from the third particle, and the kinetic energy of scattering is less in magnitude than the two-particle binding energy, $-\mu^2$. These states will have counterparts in the positive energy sector for which the kinetic energy dominates. Also in the positive energy sector will be scattering states of three unbound fundamental particles. In fact Bruch and Tjon[14] have carried out a numerical study of (28) and concluded that there are just two three-particle bound states, a result consistent with the above picture.

The expression for H in (30), in combination with the explicit form of W, is one of our key results. It gives us a rare glimpse at an explicit renormalized Hamiltonian, which in most theories can at best be asymptotically approached order by order in perturbation theory. In this system the entire interaction appears in the domain of the renormalized Hamiltonian. The fact that the interaction is of zero range allows us to obtain a compact form for the eigenvalue problem by effectively integrating out one degree of freedom. That the simplest way to write the resulting Hamiltonian is in terms of its logarithm is interesting, but should not surprise us in light of the prevalence of logarithmic dependences in asymptotically free theories. We will find in the N-body case, as well, that the Hamiltonian is best written in terms of its logarithm. Before doing this, however, we will analyze the three-body case a little further, finding in the next section an approximation to the ground state and ground state energy.

V. APPROXIMATING THE THREE-BODY GROUND STATE

Our system of three particles is invariant under rotations, so that we may use this symmetry to simultaneously diagonalize the Hamiltonian and the total angular momentum. Letting the integer l be the angular momentum quantum number, the form of negative energy wave functions becomes

$$\Psi_{\lambda,l}=\frac{\Sigma_i e^{il\theta_i} f_{\lambda,l}(p_i)}{2k^2+\frac{3}{2}p^2+\eta^2}, \tag{31}$$

where θ_i is the angle of $\overline{p_i}$.

Letting $u_{\lambda,l}(x)=f_{\lambda,l}(p)$, the eigenvalue equation, (28), can be decomposed into separate angular momentum sectors, giving us

$$\ln\left(\frac{3}{2}x+1\right)u_{\lambda,l}(x)-2\int_0^\infty \frac{dy}{(xy)^{l/2}}\frac{(\sqrt{(x+y+\frac{1}{2})^2-xy}-(x+y+\frac{1}{2}))^l}{\sqrt{(x+y+\frac{1}{2})^2-xy}}u_{\lambda,l}(y)=\ln\left(\frac{\mu^2}{\eta^2}\right)u_{\lambda,l}(x). \tag{32}$$

The ground state will correspond to the (or one of the) $l=0$ negative energy state(s). Letting $-\gamma^2$ be the ground state energy, the form of this state, which we will call Ψ_0, is

$$\Psi_0=\frac{\Sigma_i f_0(p_i)}{2k^2+\frac{3}{2}p^2+\gamma^2}, \tag{33}$$

with $f_0(p)=u_0(x)$ being the eigenstate corresponding to the lowest eigenvalue $\ln(\mu^2/\eta^2)$ solving the equation

$$\ln\left(\frac{3}{2}x+1\right)u(x)-2\int_0^\infty dy\,\frac{u(y)}{\sqrt{(x+y+\frac{1}{2})^2-xy}}=\ln\left(\frac{\mu^2}{\eta^2}\right)u(x). \tag{34}$$

By definition u_0 solves this equation for the value $\eta^2=\gamma^2$. Equivalently we can regard u_0 as the function which minimizes the quadratic functional:

$$Q[u]\equiv\int_0^\infty \ln\left(\frac{3}{2}x+1\right)u^2(x)-2\int_0^\infty dx\int_0^\infty dy\,\frac{u(x)u(y)}{\sqrt{(x+y+\frac{1}{2})^2-xy}} \tag{35}$$

subject to the constraint $\int_0^\infty dx\ u^2(x)=1$.

One way of obtaining an approximation to the ground state wave function is to numerically solve a discretized version of (34). That is, we can diagonalize the matrix version of the operator W, W_{ij}, obtained by replacing the continuous variables $0<x,\ y<\infty$ by discrete ones $0<i\Delta$, $j\Delta<N\Delta$ for integers i, j, and N and positive Δ. The task then is to diagonalize W_{ij}. Doing this amounts to making a discrete approximation to the regularized problem within which there is a high momentum cutoff. For large $N\Delta$ the unique positive eigenvector should give a useful picture of the ground state of the renormalized problem.

Using this method to obtain a numerical estimate of the ground state wave function shows that u_0 has a long, power-law-type tail. In fact, $u_0(x)$ is quite well approximated by a function of the form $\sqrt{b}/(b+x)$. This rational function is normalized such that the integral of its square on the positive real line is unity, and therefore can be taken as a variational ansatz for u_0. Inserting this form in the quadratic functional $Q[u]$ and minimizing with respect to the parameter b yields the upper bound on the ground state energy:

$$\lambda_0=-\gamma^2<-\mu^2e^{2.6}\approx-13.5\mu^2. \tag{36}$$

Recalling that $-\mu^2$ has the meaning of the two-body ground state energy, we see that the three-body energy is significantly less, which is physically reasonable. Also, this bound is consistent with the estimate $\lambda_0 \approx -16.1\mu^2$ found by Bruch and Tjon using a numerical diagonalization of the Hamiltonian. In addition they find one other bound state, with spectator particle angular momentum $l=1$, having energy $\approx -1.25\mu^2$.

VI. THE *N*-BODY PROBLEM

The original Schrodinger equation of the N-body problem in momentum space is

$$\sum_{i=1}^{N} k_i^2 \Psi_\lambda - \frac{g}{(2\pi)^2} \sum_{i<j} \int d^2 k_{ij}\, \Psi_\lambda = \lambda \Psi_\lambda , \tag{37}$$

where we again take $\hbar = 1$ and the particle masses to be $m_i = 1/2$. $\overline{k_{ij}} = \frac{1}{2}(\overline{k_i} - \overline{k_j})$ are, as in the three-body case, the relative momenta of pairs of particles. As before the system is ill defined due to scale invariance. We regularize the problem with a momentum cutoff Λ such that $\Sigma_i k_i^2 < \Lambda^2$, and define for each pair of particles the functions

$$f^{\Lambda}_{\lambda,ij} = \frac{g(\Lambda/\mu)}{(2\pi)^2} \int^{\Lambda} d^2 k_{ij}\, \Psi_\lambda . \tag{38}$$

$g(\Lambda/\mu)$ is as given in (12), and the superscript Λ on the integral indicates the restriction $\Sigma_i k_i^2 < \Lambda^2$. Recall that $-\mu^2$ is the two-body ground state energy.

As Λ is taken to ∞ in (38), $g(\Lambda/\mu)$ is driven to zero, and picks out only the logarithmically divergent part of the integral over the relative momenta of particles. The renormalized Schrodinger equation then becomes

$$\left(\sum_{i=1}^{N} k_i^2 - \lambda \right) \Psi_\lambda = \sum_{i<j} f_{\lambda,ij} . \tag{39}$$

The functions $f_{\lambda,ij}$ are the limits as $\Lambda \to \infty$ of the $f^{\Lambda}_{\lambda,ij}$, and in this limit (38) and (12) ensure that $f_{\lambda,ij} = \lim_{k_{ij}\to\infty} 2k_{ij}^2 \Psi_\lambda$. $f_{\lambda,ij}$ is independent, then, of k_{ij}, and if we continue to consider the particles to be bosons, then the necessary permutation symmetry will imply that only one of these functions need be specified. The others will follow by permutation of particle indices.

In a similar manner as in the two- and three-body problems, we can show that the renormalized Schrodinger equation, (39), must be supplemented by the domain equations:

$$\int d^2 k_{ij} \left(\Psi_\lambda - \frac{f_{\lambda,ij}}{2k_{ij}^2 + \mu^2} \right) = 0. \tag{40}$$

There is no difficulty in using the Lippmann–Schwinger form of (39) in combination with (40) to derive an equation like (20) for the N-body problem. The full equation is not very illuminating, however, so we do not reproduce it here. Rather we confine ourselves to the negative energy sector and again let $\lambda = -\eta^2$. The wave functions then have the form

$$\Psi_\lambda = \frac{\Sigma_{i<j} f_{\lambda,ij}}{\Sigma_{i=1}^{N} k_i^2 + \eta^2} . \tag{41}$$

Inserting this form into (40), taking, for example, the integration to be over the relative momentum k_{12}, and transforming to dimensionless variables $\overline{x_i} = k_i/\eta$, yields the eigenvalue equation:

$$\ln\left(1 + \frac{1}{2}(\overline{x_1} + \overline{x_2})^2 + \sum_{i=3}^{N} x_i^2 \right) u_{\lambda,ij} - \frac{2}{\pi} \int d^2 x_{12} \frac{\Sigma_{\substack{i<j \\ (i,j)\neq(1,2)}} u_{\lambda,ij}}{\Sigma_{i=1}^{N} x_i^2 + 1} = \ln\left(\frac{\mu^2}{\eta^2} \right) u_{\lambda,ij} , \tag{42}$$

where $u_{\lambda,ij}(\overline{x_1},\ldots,\overline{x_N})=f_{\lambda,ij}(\overline{k_1},\ldots,\overline{k_N})$. One must keep in mind that $u_{\lambda,ij}$ is independent of $\frac{1}{2}(\overline{x_i}-\overline{x_j})$, so that in the center of the momentum frame this function depends on $N-2$ two-dimensional vectors. The linear operator on the left-hand side of (42) is the N-body version of the operator W. We may, as in the three-body system, identify the renormalized spectator Hamiltonian in the negative energy sector as the exponential of this operator, i.e., $H=-\mu^2 e^{-W}$.

We conclude that the N-body problem can be given a renormalized formulation that is in principle no more complicated than for the three-body problem. Of course it is much more difficult to solve, or even to estimate its ground state and ground state energy. Our most intriguing result, however, that the simplest renormalized formulation is in terms of the logarithm of the Hamiltonian which is an integral operator in momentum space, continues to hold in the negative energy sector of the N-body problem for arbitrary N.

We have found, therefore, for all N, a renormalized, i.e., *finite*, formulation of the system of N particles interacting through an attractive Dirac δ-function potential.

An undesirable feature of our formulation is the schism between the treatment of positive and negative energy states. States for all positive values of energy exist, and are found by inverting a linear integral operator in momentum space. Negative energy states, in contrast, come as solutions to an eigenvalue problem. This is the price we pay for combining Eqs. (16) and (15) into a formulation of the problem in which the boundary condition on wave functions is built in, and not a supplementary condition.

In a sense, however, all the essential features of the system are included in the negative energy sector. In this sector are contained normalizable bound states, scattering among these composite particles, as well as scattering of the composite particles with the elementary excitations of the theory. It is possible, in fact, that as we allow $N\to\infty$, keeping the ground state energy finite, that it is only this sector which evolves into the resulting field theory. The field theory found this way should admit a finite formulation along the lines presented here, and is a natural direction for future research.

ACKNOWLEDGMENTS

This work was supported in part by funds provided by the U.S. Department of Energy under Grant No. DE-FG02-91ER40685.

[1] K. G. Wilson, Phys. Rev. D **2**, 1438 (1970); Phys. Rev. B **4**, 3174 and 3184 (1971); Phys. Rep. C **12C**, 75 (1974); Rev. Mod. Phys. **55**, 583 (1983).
[2] R. J. Henderson, Ph.D. Dissertation, University of Rochester, Rochester, NY (1997).
[3] R. J. Henderson and S. G. Rajeev, Int. J. Mod. Phys. A **10**, 3765 (1995).
[4] R. J. Henderson and S. G. Rajeev, J. Math. Phys. **38**, 2171 (1997).
[5] C. Thorn, Phys. Rev. D **19**, 639 (1979).
[6] K. Huang, *Quarks, Leptons and Gauge Fields* (World Scientific, Singapore 1982).
[7] K. S. Gupta and S. G. Rajeev, Phys. Rev. D **48**, 5940 (1993).
[8] C. Manuel and R. Tarrach, Phys. Lett. B **328**, 113 (1994).
[9] G. Amelino-Camelia, Phys. Lett. B **326**, 282 (1994); Phys. Rev. D **51**, 2000 (1995).
[10] S. Albeverio, F. Gesztesy, R. Hoegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics* (Springer-Verlag, New York, 1988).
[11] L. D. Faddeev, *Mathematical Aspects of the Three-body Problem in the Quantum Scattering Theory* (Israel Program for Scientific Translations, Jerusalem, 1965).
[12] B. A. Lippmann and J. Schwinger, Phys. Rev. **79**, 469 (1950).
[13] R. D. Amato and J. V. Noble, Phys. Rev. D **5**, 1992 (1972).
[14] L. W. Bruch and J. A. Tjon, Phys. Rev. A **19**, 425 (1979).

Splitting in large dimension and infrared estimates. II. Moment inequalities

B. Helffer
UA 760 du CNRS, Département de mathématiques,
Bât 425, F-91405 Orsay Cédex, France

(Received 28 March 1997; accepted for publication 4 September 1997)

This is the continuation of notes written for the NATO-ASI conference in Il Ciocco (September 96) consisting of the analysis of the links between estimating the splitting between the two first eigenvalues for the Schrödinger operator H and the proof of infrared estimates for quantities attached to Gaussian-type measures. These notes were mainly reporting on the ''old'' contributions of Dyson, Fröhlich, Glimm, Jaffe, Lieb, Simon, and Spencer (in the 1970s) in connection with more recent contributions of Pastur, Khoruzhenko, Barbulyak, and Kondrat'ev which treat in general more sophisticated models. Here we concentrate on the simplest model related to field theory and extend the results of Barbulyak and Kondrat'ev by mixing ideas coming from Pastur and Khozurenko related to the use of Bogolyubov's inequality with classical inequalities due to Ginibre, Lebowitz, Sokal, and others, or, in the case when the temperature T is zero, by applying rather elementary estimates on Schrödinger operators, in order to find lower bounds for second-order moments attached to the measure $\phi \mapsto \mathrm{Tr}\phi \exp -\beta H/\mathrm{Tr} \exp -\beta H$ with $\beta = 1/T$. This question was ''left to the reader'' in lectures given by J. Fröhlich in 1976 [Acta Phys. Austriaca, Suppl. XV, 133–269 (1976)], but we think that it is worthwhile to do this ''homework'' carefully. © *1998 American Institute of Physics.* [S0022-2488(98)03002-3]

I. INTRODUCTION

Our aim is to understand in the large dimension limit $L \to +\infty$ the splitting between the two lowest eigenvalues of the following Schrödinger operator

$$H_\Lambda^{\mathrm{per}} := -h^2\Delta + \sum_{j \in \Lambda} v(x_j) + \frac{\mathscr{J}}{2} \sum_{j \sim k} |x_j - x_k|^2, \tag{I.1}$$

where

(i) Λ is a square periodic lattice $\Lambda := \Lambda(L) = (\mathbb{Z}/L\mathbb{Z})^d$ identified (up to translation) with $\{0,...,L-1\}^d$ in $\mathbb{Z}^d$,
(ii) $j \sim k$ means j nearest neighbor of k in Λ (seen as a periodic lattice),
(iii) v is a one-particle even potential on $\mathbb{R}^n$ defining as typical cases either a single well or a symmetric double well,
(iv) $\mathscr{J}$ is the intensity of the interaction.

As is well known, the first eigenvalue of the Schrödinger operator is simple, so the splitting is always strictly positive. Its behavior with respect to L as $L \to +\infty$ depends actually heavily on the nature of v and on the size of $\mathscr{J}$, which is always assumed positive.

The parameter $\mathscr{J}$ satisfies the condition

$$\mathscr{J} > 0. \tag{I.2}$$

The case $\mathscr{J} = 0$, which could be considered for comparison, corresponds to the case without interaction and the analysis in this case is immediately reduced to the semi-classical analysis of the so-called one-particle Hamiltonian on $L^2(\mathbb{R}^n)$,

 © 1998 American Institute of Physics

$$H_0 := -h^2\Delta_y + v(y). \tag{I.3}$$

This model can also be written as

$$H_\Lambda^{\mathrm{per}} = -h^2\Delta + \sum_{\ell \in \Lambda} \tilde{v}(x_\ell) - \mathscr{J}\sum_{i\sim j} x_i \cdot x_j\,, \tag{I.4}$$

with

$$\tilde{v}(y) = v(y) + \mathscr{J}dy^2 \quad \forall y \in \mathbb{R}^n. \tag{I.5}$$

This sometimes leads (see Ref. 1) to another notion of one-particle operator:

$$\tilde{H}_0 := -h^2\Delta_y + \tilde{v}(y). \tag{I.6}$$

In this article we will in particular discuss the relevance of these various one-particle Hamiltonians in the problem of the phase transition.

For a one-dimensional lattice ($d=1$), this model was analyzed in Ref. 2 by extension of the study of another model analyzed in collaboration with J. Sjöstrand.[3] It was proved that

$$\lambda_2(L;h) - \lambda_1(L;h) \leqslant C \exp -\left(\frac{\epsilon L}{h}\right), \tag{I.7}$$

for C, ϵ independent of L and for $h<h_0$, h_0 independent of L.

But we met an additional and unfortunate condition

$$L \leqslant Ch^{-N_0}, \tag{I.8}$$

without knowing if it was a technical or deep condition.

The main object of the NATO-ASI notes was to analyze another approach based on the so-called infrared estimates which were developed by many authors in the late 70's including Fröhlich, Simon, and Spencer,[4] Dyson, Lieb, and Simon,[5] and Glimm and Jaffe,[6] and more recently by Pastur and Khozurenko[7] and Barbulyak and Kondrat'ev.[1] In particular, we show how this infrared approach also gives information about the splitting for our initial questions. It gave actually a rather complete answer for all the lattices of dimension strictly greater than 1. In this case, the result does not seem to be related to tunneling properties. In the case when the lattice is of dimension 1, we obtained only a partial result under the condition

$$|\Lambda| \leqslant \exp \frac{\varepsilon}{h}, \tag{I.9}$$

for some suitable $\varepsilon>0$. All these results were obtained relatively easily as extensions of these contributions. Because, particularly in Ref. 1 semi-classical analysis is involved, we tried to be more precise than in the original paper, using our more precise knowledge of the tunneling.[8] On the other hand we progressively got the impression that some assumptions were actually irrelevant in the problem and our main goal is to present here rather optimal results, although limited by the bounds of this approach (see Ref. 9 for discussions).

Let us now recall the main results obtained through the infrared estimates. Following Ref. 1, we now consider the operator $\exp -\beta H_\Lambda^{\mathrm{per}}$ and the associated

$$\left\langle \left| \frac{1}{|\Lambda|} \sum_{k\in\Lambda} x_k \right|^2 \right\rangle_{\beta,\Lambda} := \frac{\mathrm{Tr}\,(|(1/|\Lambda|)\Sigma_{k\in\Lambda} x_k|^2 \exp(-\beta H_\Lambda^{\mathrm{per}}))}{\mathrm{Tr}\,(\exp(-\beta H_\Lambda^{\mathrm{per}}))}. \tag{I.10}$$

Physically the strictly positive parameter β corresponds to the inverse of the temperature.

We now introduce the so-called parameter of long-range order

$$P(\beta) = \lim_{|\Lambda|\to\infty} P_\Lambda(\beta), \tag{I.11}$$

with

$$P_\Lambda(\beta)=\left\langle \left| \frac{1}{|\Lambda|}\sum_{k\in\Lambda} x_k \right|^2 \right\rangle_{\beta,\Lambda}. \tag{I.12}$$

The presence of the long-range order, i.e., the strict positivity of $P(\beta)$, will serve as a test for phase transition (cf. Refs. 5 and 7).

When the limit of $\langle\cdot\rangle_{\beta,\Lambda}$ as $|\Lambda|\to+\infty$ exists, we shall denote it by $\langle\cdots\rangle_\beta$. Let

$$E(p)=\sum_{i=1}^{d}(1-\cos p^{(i)}), \tag{I.13}$$

where Λ^* is the dual lattice

$$\Lambda^*=\{p=(p^{(1)},\ldots,p^{(d)})|p^{(i)}=2\pi k^{(i)}/m,\ \ 0\leqslant k^{(i)}\leqslant m-1;\ \ 1\leqslant i\leqslant d\}. \tag{I.14}$$

The method of infrared estimates gives (cf., for example, Ref. 5, (52), p. 368, and also Theorem 3.2 and Theorem 5.1 in this article) the following lower bound.

Theorem I.1: *If $d\geqslant 3$ and under the following assumption on v,*

$$v(x)\geqslant ax^2+b,\quad \text{with}\ \ a>0, \tag{I.15}$$

we have

$$P(\beta)\geqslant\langle x_k^2\rangle_\beta-\frac{n}{2}\cdot\frac{1}{(2\pi)^d}\int_{]-\pi,\pi[^d}\left(\frac{h^2}{\mathscr{F}E(p)}\right)^{1/2}\coth\left[(h^2\beta^2\mathscr{F}E(p))^{1/2}\right]dp. \tag{I.16}$$

This lower bound is deduced by a limiting argument (thermodynamic limit) from the actually more useful inequality (for our questions relative to the splitting), which is relative to the finite lattice case, and which is given by the following theorem essentially due to Ref. 5.

Theorem I.2: *Under the assumption (I.15) on v, we have, for any $k\in\Lambda$, the following universal estimates:*

$$P_\Lambda(\beta)\geqslant\langle x_k^2\rangle_{\beta,\Lambda}-\frac{n}{2}\cdot\frac{1}{|\Lambda|}\sum_{p\in\Lambda^*\setminus\{0\}}\left(\frac{h^2}{\mathscr{F}E(p)}\right)^{1/2}\coth[(\beta^2h^2\mathscr{F}E(p))^{1/2}]. \tag{I.17}$$

Pastur and Khozurenko[7] and more recently Barbulyak and Kondrat'ev[1] try (in particular) to analyze how one can control the quantity $\langle x_k^2\rangle_{\beta,\Lambda}$ from below in order to deduce existence for long-range order. The second authors use semi-classical analysis for the one-particle Hamiltonian operator $\widetilde{H}_0$ and standard inequalities (GKS) reducing the problem to this case. The first authors use for a rather specific model the Bogolyubov inequality. Their result is no more semi-classical and explicit conditions for long-range order are given.

Extending preliminary results announced in Ref. 10, we push these results in various directions:

(i) very weak assumptions in the limit case $\beta=+\infty$, without using (GKS),
(ii) stronger but still rather weak assumptions for the case β large, but using standard moment inequalities.

We obtain, for example, the following theorem.

Theorem I.3: *Let us assume $d\geqslant 2$. If the potential v satisfies (I.15),*

$$v(x)=v(-x), \tag{I.18}$$

$$v(0)>\inf v, \tag{I.19}$$

then there exists h_0 such that we have, for $0<h<h_0$,

$$\lim_{|\Lambda|\to+\infty}(\lambda_2^{\Lambda}(h)-\lambda_1^{\Lambda}(h))=0, \tag{I.20}$$

where λ_1^{Λ} *and* λ_2^{Λ} *are the two lowest eigenvalues of* $H_{\Lambda}^{\mathrm{per}}$.

The direct application of the result by Barbulyak and Kondrat'ev would have given, in the limit $\beta\to+\infty$, the conditions $d\geqslant 3$, $n=1$, and the conditions on $\tilde{v}$ that

$$\tilde{v}(y)\geqslant ay^2+b, \quad \text{with } a>\mathscr{J}d, \tag{I.21}$$

and that

$$\text{for some } q_0>0, \text{ the function } \tilde{v} \text{ attains its strong global minimum at the points } \pm q_0. \tag{I.22}$$

On the other hand, the conditions given by the authors are much stronger [but lead to a stronger result (treatment of the case β large) which will be analyzed later].

In any case, what seems still open is the complete analysis of the tunneling with control of the dimension and we recall from Ref. 9 the important point that the Peierls approch will give additional information, particularly for $d=1$, which cannot be analyzed by this method. In reference to these enlightening lectures, what we are doing here is essentially what was left to the reader in these notes (Ref. 9, p. 237). We recently received a preprint of Albeverio, Kondratiev, and Rebenko[11] based on the Peierls argument.

This article is organized as follows.

(i) In Sec. II, we recall the relations between the splitting and some notions related to the absence of long-range order. We recall also the general strategy of the infrared estimates.
(ii) In Sec. III, we recall a family of estimates extending in various directions the FKG estimates: GKS, FKG, GHS, Lebowitz, and Gaussian domination estimates. These estimates will play an important role in the treatment of the case of β large but are not used in the treatment of $\beta=+\infty$.
(iii) Section IV is devoted to the analysis of lower bounds for the second-order moments in the case $\beta=+\infty$. This is a combination of standard semi-classical analysis with control with respect to the dimension (harmonic approximation) in combination with *a priori* estimates. Although this would be very interesting to analyze, no study of the tunneling is involved and only a rather weak localization of the first eigenfunction.
(iv) In Sec. V, we extend the results proposed by Barbulyak and Kondrat'ev by following the same strategy but instead of reducing, through Ginibre-type inequalities GKS, to a one-particle problem, we take advantage of a reduction to a p-particle problem.
(v) Section VI is mainly reporting on the partially alternative strategy followed by Pastur and Khozurenko and based on the use of the Bogolyubov inequality. This approach is not semi-classical and gives more explicit estimates for specific models.
(vi) In Sec. VIII we propose extensions of the model treated by Pastur and Khozurenko by implementation of the results recalled in Sec. III. This leads to less explicit results but permits, for example, the analysis of the dependence on $\mathscr{J}$ of the transition of phase.
(vii) The last section presents some open questions.

II. CONNECTION WITH THE SPLITTING

We recall here rather standard relations connecting the splitting between the two first eigenvalues of the Schrödinger operator with various quantities introduced in the preceding section. They are quite general and true for any Schrödinger operator

$$H_V:=-h^2\Delta+V,$$

where V is C^{∞} semi-bounded, tending to ∞ as $|x|\to+\infty$ and satisfying

$$V(-x)=V(x).$$

We have the following lemma.

Lemma II.1:

$$\int \left| \frac{1}{|\Lambda|} \sum_i x_i \right|^2 \phi_i^\Lambda(x^\Lambda)^2 dx^\Lambda \leqslant \frac{n}{|\Lambda|} (\lambda_2^\Lambda - \lambda_1^\Lambda)^{-1}. \tag{II.1}$$

This is indeed just the minimax principle saying that

$$(\lambda_2^\Lambda - \lambda_1^\Lambda) = \inf_{\substack{f \in C_0^\infty \\ \int f(x^\Lambda)\phi_1^\Lambda(x^\Lambda)^2\, dx^\Lambda = 0}} \frac{\langle |\nabla f|^2 \rangle}{\langle |f|^2 \rangle}, \tag{II.2}$$

where $\langle \cdot \rangle$ is computed for the measure $\phi_1^\Lambda(x^\Lambda)^2 dx^\Lambda$:

$$\langle \cdot \rangle = \langle \cdot \rangle_{\infty,\Lambda} \tag{II.3}$$

(remembering that it corresponds to the case $\beta = +\infty$). Here dx^Λ is the Lebesgue measure on $(\mathbb{R}^n)^{|\Lambda|}$.

This is then applied to the function

$$f^{(j)}(x^\Lambda) = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} x_i^{(j)}, \tag{II.4}$$

for $j = 1,...,n$. The condition of symmetry on V (which is a consequence of the symmetry of v for our specific example) gives the orthogonality of $f^{(j)}\Phi_1^\Lambda$ to Φ_1^Λ.

In the case when $x_i \in \mathbb{R}^n$ with $n > 1$, we have to introduce the functions $f^{(j)}(x^\Lambda) = (1/|\Lambda|)\Sigma_{i \in \Lambda} x_i^{(j)}$ (for $j = 1,...,n$).

This means that if we can prove by other means the property that

$$\left[\int \left| \frac{1}{|\Lambda|} \sum_i x_i \right|^2 \phi_1^\Lambda(x^\Lambda)^2 dx^\Lambda \right] \geqslant \rho > 0, \tag{II.5}$$

with ρ independent of Λ, then we get that

$$\lim_{|\Lambda| \to +\infty} |\lambda_2^\Lambda - \lambda_1^\Lambda| = 0. \tag{II.6}$$

A lower bound for $P_\Lambda(+\infty)$ gives an upper bound for the splitting, but Theorem I.2 gives the starting point for finding this lower bound and this will then permit us to prove Theorem I.3. We shall mainly follow the proof of Barbulyak and Kondrat'ev but with a small change.

In Ref. 1, the authors indeed take first the limit $|\Lambda| \to +\infty$ and then the limit $\beta \to +\infty$. We shall proceed for the application to the splitting in the inverse order.

Proof of Theorem I.3, modulo a lower bound of $\langle x_k^2 \rangle_{\beta,\Lambda}$: We assume in this part that we have proved the existence of a constant $\delta(\beta,h) > 0$ such that for any Λ (with sometimes a restriction $L \geqslant L_0$)

$$\langle x_k^2 \rangle_{\beta,\Lambda} \geqslant \delta(\beta,h). \tag{II.7}$$

We shall sometimes get in the next sections a control of $\delta(\beta,h)$, with β and h, of the following type.

(i) In the case $\beta = +\infty$,
(As1) there exists $\delta_0 > 0$ and $h_0 > 0$ such that for $h \in]0,h_0]$
$$\delta(\infty,h) \geqslant \delta_0.$$

(i) In the case β large,
(As2) there exists $\delta_0 > 0$, $h_0 > 0$ such that for $h \in]0,h_0]$ there exists $\beta_0(h)$ such that, for all $\beta \geqslant \beta_0(h)$,

$$\delta(\beta,h)\geqslant\delta_0,$$

or

(As3) there exists $\delta_0>0$, $h_0>0$, and β_0 such that for $h\in]0,h_0]$ and $\beta\geqslant\beta_0$,

$$\delta(\beta,h)\geqslant\delta_0.$$

What is common to all these estimates is that they are independent of Λ with possibly the restriction L large enough. This last restriction is usually not important because we shall always take the limit $L\to+\infty$ at the end.

We start from (I.17)

$$P_\Lambda(\beta)\geqslant\langle x_k^2\rangle_{\beta,\Lambda}-\frac{nh}{2}\cdot\frac{1}{|\Lambda|}\sum_{p\in\Lambda^*\setminus\{0\}}\left(\frac{1}{\mathscr{J}E(p)}\right)^{1/2}\coth\left[(h^2\beta^2\mathscr{J}E(p))^{1/2}\right]. \tag{II.8}$$

This leads, under assumption (II.7) to

$$P(\beta)\geqslant\delta(\beta,h)-\frac{nh}{2\mathscr{J}^{1/2}}\cdot\frac{1}{(2\pi)^d}\int_{]-\pi,\pi[^d}E(p)^{-1/2}\coth\left[(h^2\beta^2\mathscr{J}E(p))^{1/2}\right]\,dp. \tag{II.9}$$

The discussion of the case β large is then related to the analysis of $\delta(\beta,h)$ for β large and its comparison with the quantity

$$\frac{nh}{2\mathscr{J}^{1/2}}\cdot\frac{1}{(2\pi)^d}\int_{]-\pi,\pi[^d}E(p)^{-1/2}\coth\left[(h^2\beta^2\mathscr{J}E(p))^{1/2}\right]\,dp.$$

The condition $d\geqslant3$ appears naturally in this case to keep the integral appearing on the rhs of (II.9) convergent.

Taking the limit $\beta\to+\infty$ in (II.8), we first obtain

$$\lim_{\beta\to+\infty}P_\Lambda(\beta)\geqslant\langle x_k^2\rangle_{+\infty,\Lambda}-\frac{nh}{2}\cdot\frac{1}{|\Lambda|}\sum_{p\in\Lambda^*\setminus\{0\}}\left(\frac{1}{\mathscr{J}E(p)}\right)^{1/2}. \tag{II.10}$$

Now, as $|\Lambda|\to+\infty$ and if $d\geqslant2$, the right-hand side is estimated from below, under the assumption (As1) by

$$P(+\infty)\geqslant\delta_0-\frac{nh}{2}\,\mathscr{J}^{-1/2}I_d, \tag{II.11}$$

where we recall that

$$I_d=\frac{1}{(2\pi)^d}\int_{]-\pi,\pi[^d}E(p)^{-1/2}\,dp. \tag{II.12}$$

We now observe that I_d is finite for $d\geqslant2$. Then $P(\infty)$ becomes strictly positive when $h\leqslant h_0$ and

$$\delta_0-\frac{nh}{2}\,\mathscr{J}^{-1/2}I_d>0. \tag{II.13}$$

In some cases h_0 will be explicitly computed (e.g., in the cases considered by Ref. 7) and we can in this case discuss the influence of $\mathscr{J}$ for fixed h, in particular (II.13) is satisfied for $\mathscr{J}$ large enough. In other cases (e.g., in the cases considered by Ref. 1), h_0 is not explicited by the (semi-classical) analysis and we can only say that (II.13) is satisfied for h small enough.

In particular we obtain the proof of Theorem I.3 modulo the verification of (As1).

The case $d=1$. Another interesting point is to analyze the case $d=1$. The sum

$$\frac{1}{|\Lambda|}\sum_{p\in\Lambda^*\setminus\{0\}}\left(\frac{1}{\mathscr{J}E(p)}\right)^{1/2}$$

is divergent as $|\Lambda| \to +\infty$, but this divergence is controlled in $\ln(|\Lambda|)$. This leads to the statement that the splitting remains in $\mathcal{O}(1/|\Lambda|)$ for $h \leqslant h_0$ and $|\Lambda| \leqslant \exp T/h$, where T is explicitly computable.

This condition is much weaker than the condition (I.8) given in Ref. 3. As $h \to 0$, the estimate still works, under assumption (As1), if

$$\ln |\Lambda| \leqslant C \frac{\delta_0}{h} [2\mathscr{J}]^{1/2}, \tag{II.14}$$

where C is a universal constant. We note also that the phenomenon is effectively related to $\mathscr{J} \neq 0$.

To summarize, we prove in this case that the following holds.

Theorem II.2: *Let $d=1$ and v satisfy (I.15), (I.18), and (I.19). Then there exists a constant C such that*

$$\lambda_2^\Lambda - \lambda_1^\Lambda \leqslant \frac{C}{|\Lambda|} \frac{1}{(1 - Ch \log|\Lambda|)_+}. \tag{II.15}$$

As mentioned in the Introduction and as was communicated to us by J. Fröhlich, this condition on $|\Lambda|$ is due to the method (when $n=1$). An approach using the Peierls trick could be more effective in this case[9] (see also Refs. 12 and 13).

III. GINIBRE- OR LEBOWITZ-TYPE INEQUALITIES

Ginibre-type inequalities are well presented in Refs. 14 and 15. They correspond to extensions of the so-called FKG inequalities[16] (see Ref. 17 for a more recent review) and permit the estimate of $\langle x_k^2 \rangle_{\beta,\Lambda}$ from below by the average $\langle x_k^2 \rangle_{\beta,\hat{H}}$ taken over the measure corresponding to the ''formal'' decoupled Hamiltonian

$$\hat{H} = -h^2 \sum_{k \in \mathbb{Z}^d} \frac{\partial^2}{\partial x_k^2} + \sum_{k \in \mathbb{Z}^d} \tilde{v}(x_k) \equiv \sum_{k \in \mathbb{Z}^d} \tilde{H}_k, \tag{III.1}$$

with separate variables.

Less formally, we introduce

$$\hat{H}^\Lambda = \sum_{k \in \Lambda} \tilde{H}_k \tag{III.2}$$

and we obtain[1] the following.

Proposition III.1: Under the assumptions (I.15), (I.18), and (I.21)–(I.22) on the potential $\tilde{v}$ and $\mathscr{J}$, the following inequality is true, for any $k \in \Lambda$:

$$\langle x_k^2 \rangle_{\beta,\Lambda} \geqslant \langle x_k^2 \rangle_{\beta,\Lambda,\tilde{H}^\Lambda}. \tag{III.3}$$

Let us describe briefly the version of the Ginibre inequalities which is needed here. This is related to the control of the sign of the correlations attached to the more general measure

$$Z^{-1} \exp \sum_{ij} J_{ij} x_i x_j \prod_{j=1}^{m} d\nu_j(x_j). \tag{III.4}$$

We look at the partial derivative of $\langle (x_k)^2 \rangle$ with respect to J_{ij}. Here $\langle \cdot \rangle$ is taken with respect to the measure (III.4).

This gives

$$\partial_{J_{ij}} \langle (x_k)^2 \rangle = \langle x_i x_j (x_k)^2 \rangle - \langle (x_k)^2 \rangle \cdot \langle x_i x_j \rangle.$$

The right-hand side appears as the pair correlation of the two functions $f = x_i x_j$ and $g = x_k^2$.

We shall deduce from the Ginibre's inequalities that this expression is positive:

$$\partial_{J_{ij}}\langle (x_k)^2\rangle \geqslant 0. \tag{III.5}$$

Let us briefly recall some elements of this theory (cf. Ref. 15, pp. 119–124, for a nice exposition). We recall the following.

Theorem III.2: *Let $\mathscr{F}_1$ be the set of functions on $\mathbb{R}$ which are non-negative and monotone on $[0,+\infty)$ and either even or odd. Let $\mathscr{F}_n$ be the functions on $\mathbb{R}^m$ of the form $f_1(x_1)\cdots f_n(x_n)$ with $f_i \in \mathscr{F}_1$. Let $d\mu$ be a probability measure of the form (III.4) where $J_{ij}\geqslant 0$ and each $d\nu_j$ has the form* $\exp(f_j(x))d\lambda_j(x)$ *with $f_j \in \mathscr{F}_1$ and $d\lambda_j$ even. Then*

$$\begin{aligned} &\text{(GKS1)} \quad \langle f\rangle \geqslant 0, \\ &\text{(GKS2)} \quad \langle fg\rangle \geqslant \langle f\rangle \cdot \langle g\rangle, \end{aligned} \tag{III.6}$$

for all $f,g \in \mathscr{F}_{m+1}$.

The reader can find a proof of the theorem in Ref. 15.

The measure in consideration is attached to the restriction to the diagonal of the distribution kernel of $\exp -\beta H_\Lambda^{\text{per}}$, where $\mathscr{J}$ becomes a parameter in $[0, \mathscr{J}_0]$ and has not exactly the structure which is introduced above. However, using the Trotter product formula which describes the kernel of $\exp -\beta H_\Lambda^{\text{per}}$ as the limit in a weak sense of $(\exp -[\beta V(x)/N]\exp(\beta\Delta/N))^N$, we observe that the measure is a limit of measures satisfying the assumptions of the theorem for $\mathscr{J}\geqslant 0$. We just have to control (cf. Ref. 15, pp. 120 and 122) the limit procedure in order to obtain the result.

Application: The Hamiltonian $\hat{H}^\Lambda$ describes a system of noninteracting particles. Consequently, we obtain immediately, for $k \in \Lambda$,

$$\langle x_k^2\rangle_{\beta,\Lambda;\tilde{H}} = \langle x_k^2\rangle_{\beta,\tilde{H}_k} = \langle x_0^2\rangle_{\beta,\tilde{H}_0}, \tag{III.7}$$

for all $k \in \mathbb{Z}^d$.

Here we recall that $\tilde{H}_k$ is the one-particle Hamiltonian at $k \in \mathbb{Z}^d$.

We consequently have obtained the following.

Proposition III.3: Under the previous assumptions on $\tilde{v}$ and $\mathscr{J}$ (in particular $n=1$), we have

$$\langle x_k^2\rangle_{\beta,\Lambda} \geqslant \langle x_0^2\rangle_{\beta,H_0}. \tag{III.8}$$

In particular we obtain the condition (As2) mentioned in the previous section by observing that it is sufficient for treating the case β large enough to analyze the case $\beta = +\infty$ for $\langle x_0^2\rangle_{\beta,H_0}$ and to use weak estimates for one-particle Hamiltonians. This will be generalized in one of the next sections.

Remark III.4: When $n>1$, we can no longer apply this technique directly. The case when we have a rotational symmetry can probably be treated by taking polar coordinates. The results in this case are mentioned in Ref. 6. Pastur and Khoruzenko proceeded differently in the case of the model $v(x)=(1-|x|^2)^2$. We shall discuss this point later.

We now present two other inequalities which may be important in order to compare moments relative to various Hamiltonians. We discover these inequalities in the nice paper by A. Sokal,[18] who refers to Refs. 19 and 20. The name of Lebowitz[21] is in reference to the particular case of spin systems.

Theorem III.5 (GS and Lebowitz inequalities): *Let $d\mu(x)$ be a probability measure on $\mathbb{R}^m$ of the type*

$$d\mu(x) = Z^{-1}\exp\left[\frac{1}{2}\sum_{i,j} J_{ij}x_ix_j + \sum_i h_ix_i\right]\prod_{i=1}^m d\nu_i(x_i) \tag{III.9}$$

with

$$J_{ij} = J_{ji} \geqslant 0, \quad h_i \geqslant 0, \tag{III.10}$$

where $d\nu_i$ is for each i a measure on $\mathbb{R}$ of the type

$$d\nu_i(y)=\exp\,-\widetilde{v}_i(y)\;dy, \tag{III.11}$$

where $\widetilde{v}_i$ *is an even differentiable function such that*

$$\widetilde{v}'\ \ is\ \ convex\ \ on\ \ ]0,+\infty[. \tag{III.12}$$

Then we have

$$(\mathrm{GS})\ \langle x_ix_jx_k\rangle\leqslant\langle x_ix_j\rangle\langle x_k\rangle++\langle x_ix_k\rangle\langle x_j\rangle+\langle x_jx_k\rangle\langle x_i\rangle-2\langle x_i\rangle\langle x_j\rangle\langle x_k\rangle, \tag{III.13}$$

and, when $h=(h_1,\ldots,h_m)=0$,

$$(\mathrm{Leb})\ \langle x_ix_jx_kx_\ell\rangle\leqslant\langle x_ix_j\rangle\langle x_kx_\ell\rangle+\langle x_ix_\ell\rangle\langle x_jx_k\rangle+\langle x_ix_k\rangle\langle x_jx_\ell\rangle. \tag{III.14}$$

We shall also use the following extension mentioned by Sokal who refers to Refs. 22–25.

Theorem III.6 (Gaussian domination inequalities): *Under the assumptions of validity of Theorem III.5, we have, with* $h=(h_1,\ldots,h_m)=0$, *for any multindex* $\alpha=(\alpha_1,\ldots,\alpha_p)$ *with* p *even and* $\alpha_j\in\{1,...,m\}$ *and with* $x^\alpha=\Pi_{j=1}^p x_{\alpha j}$,

$$\langle x^\alpha\rangle\leqslant\sum_{\text{pairings}}\prod\langle x_kx_\ell\rangle, \tag{III.15}$$

where the sum is over all the pairs giving a decomposition of α.

IV. SEMI-CLASSICAL ANALYSIS OF THE MANY-PARTICLE PROBLEM. THE CASE $\beta=+\infty$

We have seen in Sec. III that we can estimate from below, using the Ginibre inequalities, the quantity $\langle x_j^2\rangle_{\beta,\Lambda}$ (for $j\in\Lambda$) by the corresponding quantity $\langle y^2\rangle_{\beta,H_{\widetilde{v}}}$, attached to one particle potential $\widetilde{v}$. We can then first analyze the limit case $\beta=+\infty$ by semi-classical analysis and prove that for β large enough an analogous estimate from below remains true.

This was the approach developed by Barbulyak and Kondrat'ev[1] that we further developed in Ref. 10.

This result appears to be only semi-satisfactory because it does not seem so natural to impose that $\widetilde{v}$ creates a double well and one is more likely waiting for the weaker assumption that v creates a double well. Of course this is not important when $\mathscr{J}$ is sufficiently small, but it appears more important as $\mathscr{J}$ is large.

Our alternative idea is to treat the case $\beta=+\infty$ by analyzing directly the quantity $\langle x_j^2\rangle_{\infty,\Lambda}$, without reducing to a one-particle Hamiltonian. If successful, the advantage is that we do not need any assumption related to the application of statistical mechanics like the Ginibre inequalities.

This is of course not sufficient for having the result for large β independently of Λ but is sufficient for the analysis of the splitting. The proposition is the following.

Proposition IV.1: Let us assume that v *is* C^∞, *positive, satisfying*

$$\lim\inf_{|x|\to+\infty}v(x)>0, \tag{IV.1}$$

such that there exists $q_0>0$, *with*

$$v(x)>0,\ \forall x\ \ s.t.\ \ |x|<q_0, \tag{IV.2}$$

and the property that there exists $y_0\in\mathbb{R}^n$ *such that* $|y_0|=q_0$ *and*

$$v(y_0)=0. \tag{IV.3}$$

Then, for any $\epsilon>0$, *there exists* $h_0(\epsilon)>0$ *such that, for* $h\in]0,h_0(\epsilon)]$, $\forall j\in\Lambda$, *we have*

$$\langle x_j^2\rangle_{\infty,\Lambda}\geqslant(q_0-\epsilon)^2. \tag{IV.4}$$

This contains the double-well case. The even character of v is not assumed in this proposition, but note that this assumption is quite important in other parts of the problem in order, for example, to have $\langle x_j \rangle_{\infty,\Lambda} = 0$.

The argument is actually the conjunction of the existence of a minimum of v outside the origin and of the property that $v > 0$ in a neighborhood of 0.

Let us explain for simplification the proof in the case $n = 1$ and when v is even.

Let us denote by Φ_1 the first normalized eigenfunction of H_Λ^{per}. We have to estimate from below $\int x_j^2 \Phi_1^2 \, dx$. It is trivial that, for any ϵ such that $q_0 > \epsilon > 0$, we have

$$\int x_j^2 \Phi_1^2 \, dx \geqslant (q_0 - \epsilon)^2 \int_{x \in \mathbb{R}^{|\Lambda|}; |x_j| \geqslant q_0 - \epsilon} \Phi_1^2 \, dx. \tag{IV.5}$$

We rewrite this inequality in the form

$$\int x_j^2 \Phi_1^2 \, dx \geqslant (q_0 - \epsilon)^2 - (q_0 - \epsilon)^2 \int_{x \in \mathbb{R}^{|\Lambda|}; |x_j| \leqslant q_0 - \epsilon} \Phi_1^2 \, dx. \tag{IV.6}$$

We are then reduced to the control of $\int_{x \in \mathbb{R}^{|\Lambda|}; |x_j| \leqslant q_0 - \epsilon} \Phi_1^2 \, dx$. For this estimate, we come back to the equation satisfied by Φ_1 and obtain, also using the invariance by translation of H_Λ^{per},

$$\int v(x_j) \Phi_1^2 \, dx \leqslant \lambda_1^\Lambda / |\Lambda|. \tag{IV.7}$$

Let $m_\epsilon = \inf_{y \in [-q_0 + \epsilon, q_0 - \epsilon]} v(y)$. This quantity is strictly positive by assumption. We obtain in particular

$$\int_{|x_j| \leqslant q_0 - \epsilon} \Phi_1^2 \, dx \leqslant \frac{1}{m_\epsilon} \cdot \frac{\lambda_1^\Lambda}{|\Lambda|}. \tag{IV.8}$$

We have to control $\lambda_1^\Lambda / |\Lambda|$ independently of $|\Lambda|$. This point is the consequence of the rather standard proof of the existence of the thermodynamic limit (see, for example Ref. 2), but we need more, that is, the control with dependence on h. We can add to v the potential $(y^2 - q_0^2)^2$ in order to get a potential with two nondegenerate wells at $\pm q_0$. Using the minimax principle and the harmonic approximation, the following lemma is proved in Ref. 2 and we state it here in a very weak form.

Lemma IV.2: Let us assume that v has nondegenerate minima q_i with $v(q_i) = 0$. Then there exists a constant $C >$ and h_0 such that for $h \in]0, h_0]$ we have

$$\frac{\lambda_1^\Lambda}{|\Lambda|} \leqslant Ch. \tag{IV.9}$$

Proof of Theorem I.3: As explained in Sec. II, this is an immediate consequence of the infrared estimates once we have proven the existence of $\delta_0 > 0$ such that $\langle x_j^2 \rangle_{\beta,\Lambda} \geqslant \delta_0$. We can first assume that $\inf v = 0$. Under the assumption that $v(0) > 0$ and (I.15), it is clear that there exists y_0 such that $v(y_0) = 0$ and such that the conditions of Proposition IV.1 are satisfied.

Remark IV.3: The proof also gives a ''weak'' control of the convergence of the type; for any $\epsilon > 0$, there exists $h_0(\epsilon)$ such that, for $h \in]0, h_0(\epsilon)]$,

$$\lambda_2^\Lambda(h) - \lambda_1^\Lambda(h) \leqslant n \, \frac{1 + \epsilon}{q_0^2} \cdot \frac{1}{|\Lambda|}.$$

Remark IV.4: Let us emphasize that the assumptions are much weaker than the assumptions given by Barbulyak and Kondrat'ev, but their proof is still valid for β large. This is not the case for our proof. We shall see in the next section how we can analyze this case.

Let us also emphasize that we have not used any result of Sec. III.

Remark IV.5: Similar to the case treated in Sec. II, we can have, in the case $d=1$, some estimate like (II.15) under the weaker assumptions of Theorem I.3, (I.15), (I.18), and (I.19).

V. THE CASE β FINITE, LARGE ENOUGH

We want to treat the case when β is finite by improving the ideas of Barbulyak and Kondrat'ev. Our theorem will be the following

Theorem V.1: *Let us assume that v is a C^∞ potential on $\mathbb{R}$ with two nondegenerate minima at $\pm q_0$. Then, for all $\varepsilon>0$, there exists $h(\varepsilon)>0$ such that, for all $h\in]0,h(\varepsilon)]$, there exists $\beta_1(h)$ such that, for $|\Lambda|=L^d$, large enough, for any $k\in\Lambda$,*

$$\langle x_k^2\rangle_{\beta,\Lambda}\geqslant(1-\varepsilon)q_0^2. \tag{V.1}$$

This proof is modeled on the proof sketched by Barbulyak and Kondrat'ev.[1] These authors reduce, through the Ginibre inequalities, to the study of a one-particle problem attached to the potential $\tilde{v}$. The condition $n=1$ appears only in the application of these inequalities.

For simplicity, we explain the case when $d=1$. In the general case, this will correspond to the idea that Λ may be decomposed in smaller boxes.

Our unique idea is to replace the reduction to a one-particle problem by a reduction to a p-particle problem with p large enough but then fixed in the thermodynamic limit.

Let us divide L by p. The number p will be chosen large enough later. We start from

$$L=pL_0+\ell.$$

This leads to a natural decomposition of $\{1,...,L\}$ in L_0+1 intervals I_k $(k=1,...,L_0+1)$ with $I_k=\{(k-1)p,\ldots,kp-1\}$ for $k=1,...,L_0$. Following Ref. 1, we destroy the terms $\mathscr{J}x_jx_k$ when j and k do not belong to the same interval. We recall that these interaction terms appear only for j and k nearest neighbors in $\mathbb{Z}/L\mathbb{Z}$.

Using the Ginibre inequalities, we obtain

$$\langle x_k^2\rangle_{\beta,\Lambda}=\frac{1}{L}\sum_{k=1}^{L}\langle x_k^2\rangle_{\beta,\Lambda}\geqslant\left(1-\frac{\ell}{L}\right)\left(\frac{1}{p}\sum_{q=1}^{p}\langle x_q^2\rangle_{\beta,H^{(p)}}\right). \tag{V.2}$$

Here the last mean value is associated with the operator $\exp-\beta H^{(p)}$ and $H^{(p)}$ is the p-particle operator

$$H^{(p)}=\sum_{j=1}^{p}v(x_j)+\frac{\mathscr{J}}{2}\sum_{j=1}^{p-1}|x_j-x_{j+1}|^2+\frac{\mathscr{J}}{2}(x_1^2+x_p^2). \tag{V.3}$$

When $p=1$, we recover the one-particle Hamiltonian $H^{(1)}=\widetilde{H}_0$ introduced by Ref. 1. So, what we have written above is just the extension of what was written in the case $p=1$ by these authors.

In the case when $p=1$, the inequalities obtained from the semi-classical analysis of $H^{(1)}$ were rather weak, particularly when $\mathscr{J}$ was large, and lead to (I.21) and (I.22). Our aim is to eliminate this condition by playing with p large enough.

We now implement the ideas appearing in Sec. IV but for $H^{(p)}$. The new point is that we have destroyed the invariance by circular permutation of the model and we have to measure the effect of this modification.

The first problem is that we have now for the first eigenvalue $\lambda_1^{(p)}(h)$ of $H^{(p)}$ a weaker estimate than for the first eigenvalue of $H^\Lambda(h)$.

Lemma V.2: There exists C and h_0 such that, for any $h\in]0,h_0]$ and any p,

$$\lambda_1^{(p)}(h)\leqslant Cph+C. \tag{V.4}$$

This gives in particular that $\lambda_1^{(p)}(h)/p$ will be small for large p and small h. Following our previous analysis, we now observe that

$$\sum_{j=1}^{p} \int_{\mathbb{R}^p} v(x_j)(\varphi_1^{(p)}(x))^2 \, dx^{(p)} \leqslant \lambda_1^{(p)}(h). \tag{V.5}$$

The difference with the previous argument is that we can no longer study independently the different terms of the sum on the left-hand side of (IV.6). This does not appear to be a big problem.

As for (IV.6), we can write

$$\sum_j \int x_j^2(\varphi_1^{(p)}(x))^2 \, dx^{(p)} \geqslant p(q_0-\epsilon)^2 - q_0^2 \sum_j \int_{|x_j|\leqslant q_0-\epsilon} (\varphi_1^{(p)}(x))^2 \, dx^{(p)}. \tag{V.6}$$

We then obtain

$$\sum_j \int x_j^2(\varphi_1^{(p)}(x))^2 \, dx^{(p)} \geqslant p(q_0-\epsilon)^2 - \frac{q_0^2}{m_\epsilon} \cdot \sum_j \int_{|x_j|\leqslant q_0-\epsilon} v(x_j)(\varphi_1^{(p)}(x))^2 \, dx^{(p)}, \tag{V.7}$$

hence

$$\sum_j \int x_j^2 |\varphi_1^{(p)}(x)|^2 \, dx^{(p)} \geqslant p(q_0-\epsilon)^2 - \frac{q_0^2}{m_\epsilon} \cdot \sum_j \int v(x_j)|\varphi_1^{(p)}(x)|^2 \, dx^{(p)},$$

$$\geqslant p(q_0-\epsilon)^2 - \frac{q_0^2}{m_\epsilon} \cdot \lambda_1^{(p)}(h). \tag{V.8}$$

Finally, using Lemma V.2, we prove

$$\sum_j \int x_j^2 |\varphi_1^{(p)}(x)|^2 \, dx^{(p)} \geqslant p(q_0-\epsilon)^2 - \widetilde{C}(ph+1). \tag{V.9}$$

This is equivalent, dividing by p, to

$$\frac{1}{p}\left(\sum_j \int x_j^2 |\varphi_1^{(p)}(x)|^2 \, dx^{(p)} \right) \geqslant (q_0-\epsilon)^2 - \widetilde{C}h - \widetilde{C}\,\frac{1}{p}. \tag{V.10}$$

So we have proved the following lemma.

Lemma V.3: For any $\epsilon>0$, there exists h_ϵ and p_ϵ such that, for $h \in]0, h_\epsilon]$ and $p \geqslant p_\epsilon$,

$$\frac{1}{p}\left(\sum_{j=1}^{p} \int x_j^2 |\varphi_1^{(p)}|^2 \, dx^{(p)} \right) \geqslant (q_0-\epsilon)^2. \tag{V.11}$$

Using (V.2), this gives Theorem V.1. The argument for comparing the situation $\beta = +\infty$ and the situation β large is now possible, once p is fixed.

VI. ABOUT THE BOGOLYUBOV INEQUALITY

Let us first recall some of the results obtained by Pastur and Khozurenko. We shall then analyze the technique which is involved in order to prepare the improvements of the next section. These authors discussed the case when

$$v(x) = -\frac{a}{2}|x|^2 + \frac{b}{4}|x|^4 \tag{VI.1}$$

with $a>0$ and $b>0$.

They prove the following theorem.

Theorem VI.1: *If $d \geqslant 3$ and if*

$$\mathscr{J} > h^2(n+2)^2 b^2 I_d^2/4a^2, \tag{VI.2}$$

then there exists a temperature β_0^{-1} *such that, for* $\beta > \beta_0$, *the corresponding* $P(\beta)$ *is strictly positive.*

The control of $\langle x_k^2 \rangle_{\beta,\Lambda}$ is obtained by using the Bogolyubov inequality. We recall that this inequality (see Ref. 26, Lemma 5.5.1) states the following.

Lemma VI.2: Under suitable assumptions, as discussed in detail in Ref. 27, on the various domains of the possibly not bounded operator C *and of the self-adjoint operator* H, *we have the following inequality:*

$$\langle [[C^*, H], C] \rangle_{\beta,\Lambda} \geqslant 0. \tag{VI.3}$$

In our context, this inequality is applied with

$$C = \frac{1}{i} \sum_{j \in \Lambda} \partial_{x_j}. \tag{VI.4}$$

This leads to by considering the limit $\beta \to +\infty$ before to take the thermodynamic limit, the following result for the splitting between the two first eigenvalues as $|\Lambda|$ tends to ∞.

Theorem VI.3: *If* $d \geqslant 2$, *then, for* v *defined by (VI.1) and* $h_0 = 2a\mathscr{J}^{1/2}/(n+2)bI_d$, *we have, for* $h \in]0, h_0[$,

$$\lim_{|\Lambda| \to +\infty} |\lambda_2^\Lambda - \lambda_1^\Lambda| = 0. \tag{VI.5}$$

For this theorem, we need only the Bogolyubov inequality in the limiting case $\beta = +\infty$. In this case and when $H = -\Delta + V$ is a Schrödinger operator, the inequality (VI.3) becomes simply

$$\int ([[C^*, H], C]\Phi)(x)\Phi(x)\, dx \geqslant 0, \tag{VI.6}$$

where Φ is the first normalized positive eigenfunction of H. This is easily and directly obtained by the minimax principle. In particular, if C is given by (VI.4), we obtain the following.

Lemma VI.4:

$$\sum_{jk} \int (\partial^2 V/\partial x_j \partial x_k)(x)\Phi(x)^2\, dx \geqslant 0. \tag{VI.7}$$

When $V(x)$ has the form (we take for simplification $n = 1$)

$$V(x) = \sum_j v(x_j) + J \sum_{i \sim j} |x_i - x_j|^2,$$

we simply obtain

$$\sum_j \int v''(x_j)\Phi(x)^2\, dx \geqslant 0. \tag{VI.8}$$

By invariance of Φ in this case, we obtain the following.

Proposition VI.5: For any j,

$$\int v''(x_j)\Phi(x)^2\, dx \geqslant 0. \tag{VI.9}$$

More generally, for any β, *we have the property*

$$\langle v''(x_j) \rangle_{\beta,\Lambda} \geqslant 0. \tag{VI.10}$$

In the more specific case when $v(x)=(b/4)x^4-(a/2)x^2$, we obtain

$$\int x_j^2\Phi(x)^2\,dx\geqslant\frac{a}{3b}. \tag{VI.11}$$

The proof through the infrared estimates is then easy using the arguments presented in Sec. II.

VII. EXTENSIONS OF PASTUR AND KHOZURENKO: UNIVERSAL LOWER BOUNDS

A. Introduction

If we compare (VI.11) with the semi-classical lower bound obtained by Barbulyak and Kondrat'ev, we note that when $n=1$ and $a>2\mathscr{J}d$, the minimum of $\widetilde{v}$ is for $q=q_0$ with $q_0^2=(a-2\mathscr{J}d)/b$. This leads asymptotically (as $h\to 0$) to (VI.11) with $a/3b$ replaced by $(a-2\mathscr{J}d)/b-\varepsilon$ with $\varepsilon>0$. The improvement we gave in Sec. IV leads asymptotically (as $h\to 0$) to a lower bound by $a/b-\varepsilon$ with $\varepsilon>0$, which seems optimal in the semi-classical regime.

The main goal of this section is to extend the results proposed by Pastur and Khozurenko.

Theorem VII.1: *Let us assume that $n=1$ and that v satisfies (I.15), (I.18), and the condition*

$$v''(x)\leqslant-\gamma_0+\gamma_1x^2, \tag{VII.1}$$

for some real γ_0,γ_1 with $\gamma_0>0$. Then, for any $h>0$,

$$\frac{\gamma_0}{\gamma_1}\leqslant\langle x_j^2\rangle_{\beta,\Lambda}. \tag{VII.2}$$

In many cases, we can take $\gamma_0=-v''(0)$. The point 0 corresponds, for example, to the top between the two wells.

We note also that this approach does not make use of the Ginibre inequalities. This leaves possible extensions for the case $n>1$ in the case when Hess $v(0)$ has a strictly negative eigenvalue which seems generic for a double-well problem and also true in the radial case considered by Ref. 7.

We observe also that the left-hand side in (VII.2) is independent of h.

B. Proof of the theorem

The Bogolyubov inequality was giving

$$\langle v''(x_j)\rangle_{\beta,\Lambda}\geqslant 0. \tag{VII.3}$$

The theorem follows immediately from (VII.1).

Because the left-hand side of (VII.2) is independent of β and Λ, the arguments developed in Sec. II work in the same way: long-range order for h small enough or $\mathscr{J}$ large enough. Moreover, the control of ''h small enough'' is explicit.

We recall that in the semi-classical approach developed in the spirit of Barbulyak and Kondratev ''h small enough'' is implicit.

As already mentioned, the condition (VII.2) can be verified when

$$v''(0)<0, \tag{VII.4}$$

which is a condition of concavity at 0, and when the condition at ∞

$$v''(x)\leqslant Cx^2,\quad\forall x>C, \tag{VII.5}$$

is satisfied. The first one seems natural in the context. The second one is apparently more technical, at least in the semi-classical context, as we shall see in the next subsection for an example.

C. The case $n=1$, $v(x)=bx^{2k}-ax^2$

In this case we have

$$v''(x)=2k(2k-1)bx^{2k-2}-2a, \tag{VII.6}$$

with $a>0$. We then obtain the inequality, for any $j\in\Lambda$,

$$\frac{a}{k(2k-1)b}\leqslant\langle x_j^{(2k-2)}\rangle_{\beta,\Lambda}, \tag{VII.7}$$

as a consequence of the Bogolyubov inequality.

For $k=2$, this was enough for having a universal lower bound for $\langle x_j^2\rangle_{\beta,\Lambda}$, which was decisive for the use of the infrared estimate.

The problem is then to treat the case $k>2$. The Gaussian domination inequalities give a rather satisfactory answer. They give indeed the inequality

$$\langle x_j^{(2k-2)}\rangle_{\beta,\Lambda}\leqslant C(k)\langle x_j^2\rangle^{k-1}, \tag{VII.8}$$

where $C(k)$ is a universal constant counting the number of pairings. Newman[23] gives

$$C(k)=\frac{(2k-2)!}{2^{k-1}(k-1)!}. \tag{VII.9}$$

Consequently, a lower bound for $\langle x_j^2\rangle_{\beta,\Lambda}$ (which has to be uniform with respect to Λ, β, and j) will be a consequence of an upper bound for $\langle x_j^{2k}\rangle_{\beta,\Lambda}$. We note here that everything is actually independent of j.

The first idea could be to analyze first the case when $\beta=+\infty$. This leads to some results which are not basically better than the results obtained in Sec. IV. We just mention that these results are obtained by combination of (VI.9) and the inequality

$$\langle v(x_j)\rangle_{\infty,\Lambda}\leqslant\frac{\lambda_1^{\Lambda}}{|\Lambda|}. \tag{VII.10}$$

In the case β large, the arguments of the preceding example work if we have control of $\langle x_j^{2k}\rangle_{\beta,\Lambda}$ by a constant (using in this case a Hölder inequality) or if we can more directly have a control of $\langle x_j^{2k-2}\rangle_{\beta,\Lambda}$ by a power of $\langle x_j^2\rangle_{\beta,\Lambda}$. In the next section, we shall exploit this idea in combination with the Ginibre's inequalities.

D. Using Lebowitz inequality or Gaussian domination inequalities

As already said, the general principle of the Lebowitz inequalities is to control any correlation by the pair correlations under rather weak assumptions. Our strategy will be to modify the initial potential v in order to get a new potential v_1 satisfying the assumptions of Theorem III.6. This will lead to the following improvement of Theorem VII.1.

Theorem VII.2: *Let us assume that* $n=1$ and *that* v *satisfies (1.15), (1.18), and the condition*

$$v''(0)<0. \tag{VII.11}$$

Let us also assume the existence of ℓ *and* C *such that*

$$|v'''(x)|\leqslant Cx^{\ell}\quad\forall x>C. \tag{VII.12}$$

Then, there exists $\rho>0$ *such that, for any* $h>0$, $\beta>0$ *and* Λ, *we have*

$$\rho\leqslant\langle x_j^2\rangle_{\beta,\Lambda}. \tag{VII.13}$$

Outside this unfortunate condition $n=1$ due to the fact that we use the Ginibre inequality, this theorem is rather satisfactory. The computation of ρ is not very explicit but can be done on examples.

Proof of Theorem VII.2: The main new step is the following lemma.

Lemma VII.3: Under the assumption of Theorem VII.2, there exists $k\in\mathbb{N}$, $\alpha_0>0$, *and* $\alpha_1>0$ *such that*

$$v_1(x)=v(x)+\alpha_0 x^4+\alpha_1 x^{2k} \tag{VII.14}$$

satisfies

$$v_1'(x) \text{ is convex on } \mathbb{R}^+. \tag{VII.15}$$

We observe that v_1 satisfies (I.15), (I.18), and for suitable $\gamma_0>0$ and $\gamma_1>0$ the estimate

$$v_1''(x)\leqslant -\gamma_0+\gamma_1 x^{2k}. \tag{VII.16}$$

The key remark is that the Ginibre inequalities give the estimate

$$\langle x_j^2\rangle_{\beta,\Lambda,v}\geqslant\langle x_j^2\rangle_{\beta,\Lambda,v_1}. \tag{VII.17}$$

Here $\langle x_j^2\rangle_{\beta,\Lambda,v}$(resp. $\langle x_j^2\rangle_{\beta,\Lambda,v_1}$) means that we consider H_Λ^{per} with the one-particle potential v (resp. v_1).

We are now reduced to the problem of finding a lower bound for $\langle x_j^2\rangle_{\beta,\Lambda,v_1}$. With the properties of convexity of v_1' and the estimate (VII.17), this case is easily treated by combination of the Bogolyubov inequality and of the Gaussian domination inequality.

VIII. CONCLUSION

In this article, we have given explicit conditions on the potential v under which a good lower bound can be found for the second-order moment.

When $\beta=+\infty$, the results are not far from the optimal one and no restriction in the dimension n is involved. Nevertheless, restrictions on d appear when applying these inequalities in the infrared estimate strategy for proving transition of phase.

When we want to extend this study to the case β large, we meet immediately the condition $n=1$ [which could probably be extended to the case $n>1$ but with invariance by SO(n)] and this is due to the fact that we use these monotonicity arguments related to various extensions of FKG inequalities. The importance of the two restrictions on d and n is sometimes technical, sometimes perhaps deeper, and has surely to be investigated more carefully. The analysis of the splitting given here is rather rough due to our weak control of the localization of the first eigenfunction in the thermodynamic limit. This also has to be analyzed in the future in order to see the effects of the tunneling between the wells in the spirit of our previous works with J. Sjöstrand.[3,28]

ACKNOWLEDGMENTS

Preliminary results were announced at the conference in Ascona (June 1996) and the NATO conference in Castel Vecchio (September 1996). We thank J. Fröhlich for a motivating discussion in Ascona.

We thank also the European Union which partially supported this research through the TMR Programme of the European Commission–Network Postdoctoral training program in partial differential equations and application in quantum mechanics.

[1] V. S. Barbulyak and Y. Kondrat'ev, ''The quasiclassical limit for the Schrödinger operator and phase transitions in quantum statistical physics,'' Funct. Anal. Appl. **26(2)**, 61–64 (1992).

[2] B. Helffer, ''Recent results and open problems on Schrödinger operators, Laplace integrals and transfer operators in large dimension,'' *Advances in PDE*, in *Schrödinger operators, Markov semigroups, Wavelet analysis, operator algebras*, edited by M. Demuth, E. Schrohe, B. W. Schultze, and J. Sjöstrand, (Akademie Verlag, Berlin, 1996), pp. 11–162.

[3] B. Helffer and J. Sjöstrand, ''Semiclassical expansions of the thermodynamic limit for a Schrödinger equation II,'' Helv. Phys. Acta **65**, 748–765 (1992) and Erratum **67**, 1–3 (1994).

[4] J. Fröhlich, B. Simon, and T. Spencer, ''Infrared bounds, phase transitions, and continuous symmetry breaking,'' Commun. Math. Phys. **50**, 79–85 (1976).
[5] F. J. Dyson, E. H. Lieb, and B. Simon, ''Phase transitions in quantum spins systems with isotropic and nonisotropic interactions,'' J. Stat. Phys. **18**(4), 335–383 (1978).
[6] J. Glimm and A. Jaffe, *Quantum physics (a functional integral point of view)* (Springer Verlag, New York, 1987), 2nd ed.
[7] L. A. Pastur and B. A. Khoruzhenko, ''Phase transitions in quantum models of rotators and ferroelectrics,'' Teor. Mat. Fiz. **73**(1), 111–124 (1987).
[8] B. Helffer and J. Sjöstrand, ''Multiple wells in the semiclassical limit,'' Commun. PDE **9**(4), 337–408 (1984).
[9] J. Fröhlich, ''Phase transitions, Goldstone bosons and topological superselection rules,'' Acta Phys. Austriaca Suppl. **XV**, 133–269 (1976).
[10] B. Helffer, ''Splitting in large dimension and infrared estimates,'' in *Microlocal Analysis and Spectral Theory*, edited by L. Rodino, Proceedings of the NATO-ASI conference (Kluwer, Dordrecht, 1997), pp. 307–348.
[11] S. Albeverio, A. Yu Kondratiev, and A. L. Rebenko, ''Peierls argument and long-range order behaviour of quantum lattice systems with unbounded spins,'' Preprint (1996).
[12] J. Glimm, A. Jaffe, and T. Spencer, ''Phase transitions for ϕ_2^4 quantum fields,'' Commun. Math. Phys. **45**, 203–216 (1975).
[13] A. L. Rebenko, ''Peierls argument and long range behavior of quantum lattice systems with unbounded spins,'' Lecture in Ascona (June 1996).
[14] B. Simon, *The $P(\phi)_2$ Euclidean quantum field theory*, Princeton Series in Physics (Princeton U.P., Princeton, NJ, 1974).
[15] B. Simon, *Functional integration and quantum physics*, Pure and Applied Mathematics, No. 86 (Academic, New York, 1979).
[16] C. Fortuin, P. Kastelyn, and J. Ginibre, ''Correlation inequalities on some partially ordered sets,'' Commun. Math. Phys. **22**, 89–103 (1971).
[17] E. Laroche, ''Inégalités de corrélation sur $\{-1,+1\}^n$ et dans $\mathbb{R}^n$,'' Ann. Inst. Henri Poincaré: Phys. Theor. **29**(4), 531–567 (1993).
[18] A. D. Sokal, ''Mean-field bounds and correlation inequalities,'' J. Stat. Phys. **28**(3), 431–439 (1982).
[19] R. S. Ellis, J. L. Monroe, and C. M. Newman, ''The GHS and other correlations inequalities for a class of even ferromagnets,'' Commun. Math. Phys. **46**, 167–182 (1976).
[20] R. S. Ellis and C. M. Newman, ''Necessary and sufficient conditions for the GHS inequality with applications to analysis and probability,'' Trans. AMS **237**, 83–99 (1978).
[21] J. Lebowitz, ''GHS and other inequalities,'' Commun. Math. Phys. **35**, 87–92 (1974).
[22] D. Brydges, J. Fröhlich, and T. Spencer, ''The random walk representation of classical spin systems and correlation inequalities,'' Commun. Math. Phys. **83**, 123–150 (1978).
[23] C. M. Newman, ''Gaussian correlation inequalities for ferromagnets,'' Z. Wahrscheinlichkeitstheor. Verwandte Geb. **33**, 75–93 (1975).
[24] G. S. Sylvester, ''Representations and inequalities for Ising model Ursell functions,'' Commun. Math. Phys. **42**, 209–220 (1975).
[25] J. Bricmont, ''The gaussian inequalities for multicomponent rotators,'' J. Stat. Phys. **17**, 289–300 (1977).
[26] D. Ruelle, *Statistical mechanics*, Math. Physics monograph series (Benjamin, New York, 1969).
[27] M. Bouziane and P. A. Martin, ''Bogolyubov inequality for unbounded operators and the Bose gas,'' J. Math. Phys. **17**, 1848–1851 (1976).
[28] B. Helffer and J. Sjöstrand, ''Semiclassical expansions of the thermodynamic limit for a Schrödinger equation,'' Astérisque **210**, 135–181 (1992).

Unitary implementations of one-parameter squeezing groups

Reinhard Honegger and Alfred Rieckers
Institut für Theoretische Physik, Universität Tübingen D-72076 Tübingen, Germany

(Received 27 May 1997; accepted for publication 24 September 1997)

For the case of infinitely many photon (Boson) modes we investigate the unitary implementability of a class of symplectic one-parameter groups (more exactly, of the associated groups of Bogoliubov automorphisms on the CCR algebra) in the Fock representation and in representations of the CCR algebra, which are symplectically related and inequivalent to Fock. Furthermore, the existence of the associated (squeezing) quadratic Hamiltonians is discussed. Finally, applications in the theory of quantization in QED and in the Luttinger model are pointed out. © *1998 American Institute of Physics.* [S0022-2488(98)01701-0]

I. INTRODUCTION

In the theoretical descriptions of squeezing processes, quadratic Hamiltonians of the form $(1/2)\Sigma_{k=1}^{N}(\zeta_k a_k^* a_k^* + \overline{\zeta_k} a_k a_k)$ with finite mode number $N<\infty$ are commonly used in the interaction picture.[1] For finitely many modes, however, by the Stone-von Neumann uniqueness theorem (e.g., Ref. 2, Corollary 5.2.12, and Refs. 3, 4) the only relevant representation of the CCR is the Fock representation. Thus, for $N<\infty$ it suffices to treat the above Hamiltonian in Fock space, only.

For $N=\infty$, the above quadratic Hamiltonian has a meaning as an operator in the Fock representation, if and only if the squeezing parameters $\zeta_k \in \mathbb{C}$ (the classical pumping fields) are square summable, that is $\Sigma_{k=1}^{\infty}|\zeta_k|^2<\infty$, a result that is derived in Ref. 5. However, for infinitely many photon modes, $N=\infty$, there exist various inequivalent representations of the CCR, which from the physical point of view describe different macroscopic aspects (sectors) of the considered Boson system. Hence, it seems interesting to investigate quadratic Hamiltonians also in representations, which are inequivalent to the Fock representation.

For the infinite mode case we continue here and supplement our discussions from Refs. 5 and 6 by the investigation of quadratic Hamiltonians of the more general type,

$$H=\sum_{k=1}^{\infty}\left[\frac{1}{2}(\zeta_k a_k^* a_k^* + \overline{\zeta_k} a_k a_k) + c_k a_k^* a_k\right], \tag{1.1}$$

where $\zeta_k \in \mathbb{C}$ (with the complex conjugates $\overline{\zeta_k}$) and $c_k \in \mathbb{R}$. It is assumed that the complete set of orthonormalized eigenmodes u_k is indexed by $k \in \mathbb{N}$, such that for the annihilation operators holds $a_k \equiv a(u_k)$, etc. We further assume that the modes $\{u_k | k \in \mathbb{N}\}$ constitute a basis for the complex one-photon test function space E the completion of which is the Hilbert space $\mathscr{H}$. The unique C^*-Weyl algebra $\mathscr{W}(E)$, which is simply called the *CCR-algebra*, is considered as the field algebra of our photon system. For each $k \in \mathbb{N}$ we define the mode e_k by the phase shift $e_k := \exp\{(i/2)\arg(\zeta_k)\}u_k$. By the (complex) linearity of the (smeared) creation operators, $E \ni f \mapsto a^*(f)$, and the (complex-) antilinearity of the annihilation operators, $E \ni f \mapsto a(f)$, our Hamiltonian (1.1) then takes the form

$$H=\sum_{k=1}^{\infty}\left[\frac{s_k}{2}(a^*(e_k)^2 + a(e_k)^2) + c_k a^*(e_k)a(e_k)\right], \tag{1.2}$$

where we have set $s_k := |\zeta_k|, \forall k \in \mathbb{N}$.

In general, one treats squeezing processes, Bogoliubov transformations, renormalizations, and diagonalizations on the level of the second quantization, only. We deal here first with the one-Boson features, which we lift in a second step to the level of the second quantization. Proceeding

 © 1998 American Institute of Physics

this way, especially for our infinite mode case, demonstrates that many problems of the above types are already solvable on the one-particle level, which is easier to handle.

In fact, with the Hamiltonian (1.2), resp. (1.1), there are associated two self-adjoint (complex-) linear one-photon operators on the one-photon Hilbert space $\mathscr{H}$,

$$C:=\sum_{k=1}^{\infty} c_k|e_k\rangle\langle e_k|, \quad S:=\sum_{k=1}^{\infty} s_k|e_k\rangle\langle e_k|, \tag{1.3}$$

and the (complex-) antilinear involution J satisfying

$$J_f=\sum_{k=1}^{\infty}\langle f|e_k\rangle e_k, \quad \forall f\in\mathscr{H} \tag{1.4}$$

(the scalar product $\langle\cdot|\cdot\rangle$ on $\mathscr{H}$ is assumed to be linear in the right factor). Hence C, S, and J are mutually commuting operators on $\mathscr{H}$ having the phase-transformed orthonormalized photon modes e_k, $k\in\mathbb{N}$, as eigenvectors. In a sense described in Ref. 6, the Hamiltonian H is a formal second quantization of $C-JS$.

We first regard quite general one-parameter groups of the type $T_t=\exp\{it(C-JS)\}$, $t\in\mathbb{R}$, on the infinite-dimensional one-photon Hilbert space $\mathscr{H}$, resp. on E. Each real-linear operator T_t turns out to be a symplectic transformation on $\mathscr{H}$, that is

$$\mathrm{Im}\langle T_t f|T_t g\rangle=\mathrm{Im}\langle f|g\rangle, \quad \forall f,g\in\mathscr{H}.$$

On the second quantized level in the algebraic sense for every time $t\in\mathbb{R}$ there exists a unique *-automorphism α_{T_t} on the CCR-algebra $\mathscr{W}(E)$—a Bogoliubov transformation—such that

$$\alpha_{T_t}(W(f))=W(T_t f), \quad \forall f\in E.$$

Here $W(f)=\exp\{i2^{-1/2}(a^*(f)+a(f))\}\in\mathscr{W}(E), f\in E$, denote the Weyl operators, which generate the CCR-algebra $\mathscr{W}(E)$.

Let us assume a representation Π of the CCR-algebra $\mathscr{W}(E)$ on the representation Hilbert space $\mathscr{K}_\Pi$. Then, from purely algebraic calculations it follows, that α_{T_t} is *formally* implemented by $\exp\{itH\}$, that is, for every $t\in\mathbb{R}$ one has

$$\Pi(\alpha_{T_t}(A))=\exp\{itH\}\Pi(A)\exp\{-itH\}, \quad \forall A\in\mathscr{W}(E). \tag{1.5}$$

We use the notion *formally*, since the existence of the unitaries $\exp\{itH\}$ is not ensured and the operator series in (1.2) may not converge to a well-defined self-adjoint operator on $\mathscr{K}_\Pi$. However, in some cases there may exist a well-defined self-adjoint operator $\widetilde{H}$, which unitarily implements the group $T_t, t\in\mathbb{R}$, in Π in the sense of (1.5). In this case $\widetilde{H}$ is a renormalization of the (possibly not existing) Hamiltonian H from (1.2).

In the present work we investigate the existence of unitary implementations (1.5) for one-parameter symplectic groups of the above type $T_t=\exp\{it(C-JS)\}$ (more exactly, of the associated group of Bogoliubov automorphisms α_{T_t} on the CCR-algebra) in the Fock representation Π_F, and in representations that are symplectically related to Fock. The T_t are not necessarily bounded operators on E. For the symplectically related representations $\Pi_T:=\Pi_F\circ\alpha_T$ of the CCR-algebra, we restrict ourselves to Bogoliubov transformations α_T associated with symplectic one-photon transformations of the form $T=\exp\{JR\}$, where the self-adjoint (complex-) linear R commutes with S, C, and J, and thus has the spectral representation,

$$R=\sum_{k=1}^{\infty} r_k|e_k\rangle\langle e_k|, \tag{1.6}$$

with some $r_k\in\mathbb{R}$. The specialization to discrete operators C, S, and R is justified for many applications in physics (cf. also Sec. V). Even if C would have a continuous part of the spectrum, it could be made discrete by an arbitrarily small Hilbert–Schmidt perturbation. As we shall demonstrate, the Hilbert space representations Π_F and Π_T of $\mathscr{W}(E)$ are unitarily equivalent, if

and only if R is Hilbert–Schmidt. Our main interest, however, concerns the case where R is not Hilbert–Schmidt, since then Π_F and Π_T are *disjoint* representations, and thus they physically belong to *different macroscopic* sectors of the infinite Boson (photon) system (see Theorem 2.3 and Proposition 3.3 below).

For $s_k=0,\forall k\in\mathbb{N}$, that is $S=0$, we are in the situation of Warchall's work,[7] who completely analyzed the implementability of a unitary (and thus symplectic) one-parameter group $T_t=\exp\{itC\}$ with a positive self-adjoint C in representations symplectically related to Fock. Here, with our discrete C one gets more insight into the basic structure of Warchall's result revealing the role of an *essential infrared part* of the one-photon Hamiltonian C (cf. Theorem 2.7 below).

In the Secs. III B and IV C we generalize Warchall's result to true one-parameter squeezing groups (i.e., with nontrivial S) and give as a main result of our work necessary and sufficient criteria for their unitary implementability.

We investigate in Sec. IV the existence of Hamiltonians of the type (1.2), resp. of renormalizations of H for the above representations Π_F and Π_T. For the Fock representation it is demonstrated that for the class of quadratic Hamiltonians (1.2) one obtains more involved results than the Hilbert–Schmidt property for S given in Ref. 6, allowing even some unbounded squeezing parts S. Especially, in Theorem 4.10, it is elaborated a renormalization concept, which treats the infrared and the ultraviolet part of a generating Hamiltonian in a different manner. Also, the diagonalizability of the quadratic Hamiltonian H, resp. of the group T_t, by symplectic transformations $T=\exp\{JR\}$ of the above kind is analyzed. In Theorems 4.12 and 4.13 we take up and work out an old idea of Segal[8] (cf. also Ref. 9), that a positive diagonal form of an implementing (renormalized) Hamiltonian is not only of practical but also of fundamental conceptual importance, providing a particle interpretation of the quantized system. We formulate necessary and sufficient conditions for the diagonalizability, discuss the equivalence class of this distinguished representation, and demonstrate that positivity is a logical consequence of the diagonalization requirement.

In Sec. V we apply our results to two typical problems of Boson field theory. First, we clarify the connection between the Dirac- and the harmonic oscillator-quantization schemes in nonrelativistic QED and discuss the corresponding field Hamiltonians. Then we study the Bosonic version of the Luttinger Hamiltonian and point out how a considerable generalization of the treated class of interactions for the Luttinger model may be achieved, including strictly local potentials.

As indicated above, there occur on the one-particle level complex-linear, complex-antilinear, and real-linear operators. If not stated otherwise, we always mean with the expressions ''operator'' and ''linear operator'' complex-linear mappings. On the level of the second quantization (C^*-algebraic level, and operators on Fock space), however, we are concerned with complex-linear operators, only.

II. IMPLEMENTATIONS OF SYMPLECTIC TRANSFORMATIONS

Let us first introduce our notation and collect some preliminary results.

Throughout the present section let E be an arbitrary separable complex pre-Hilbert space with (right-linear) scalar product $\langle\cdot|\cdot\rangle$ and norm completion $\mathscr{H}$. In physical applications E is interpreted as the one-Boson test function space of a Boson system. For scalar Bosons E is a (dense) subspace of $\mathsf{L}^2(\Lambda)$, whereas $E\subseteq\mathsf{L}^2(\Lambda)\otimes\mathbb{C}^{2s+1}$ for Bosons with the spin s.[10,2] Here $\Lambda\subseteq\mathbb{R}^3$ denotes the quantization volume in position space. For photons the quantization procedure in the Coulomb gauge leads to a test function space E, as described in Sec. V A.

Since E and $\mathscr{H}$ are complex vector spaces, each real-linear operator $B:E\to\mathscr{H}$ decomposes uniquely into its (complex-) linear part B_l and its (complex-) antilinear part B_a, that is, $B=B_l+B_a$ with (the multiplication with the complex i commutes with B, if and only if B is complex-linear),

$$B_l:=\tfrac{1}{2}(B-iBi)\qquad\text{and}\qquad B_a:=\tfrac{1}{2}(B+iBi). \tag{2.1}$$

A. Symplectic transformations

The imaginary part of the scalar product is a nondegenerate symplectic form on E. A mapping $T:E\to E$ is called a *symplectic transformation* on E, if it is real linear, with range $T(E)=E$, and fulfills

$$\mathrm{Im}\langle Tf|T_g\rangle = \mathrm{Im}\langle f|g\rangle, \quad \forall f,g \in E. \tag{2.2}$$

The set of all symplectic transformations on E is denoted by $\mathscr{T}(E)$. The relation (2.2) implies the injectivity of T, and thus each symplectic transformation T acts bijectively on E. The inverse mapping $T^{-1}: E \to E$ is a symplectic transformation on E, too. Hence $\mathscr{T}(E)$ forms a group (with respect to the multiplication).

As a transformation in $\mathscr{H}$, a symplectic $\mathscr{T} \in \mathscr{T}(E)$ is a densely defined, real-linear, operator with domain E. The following results are shown in Ref. 6.

Lemma 2.1: Let $\mathscr{T} \in \mathscr{T}(E)$. Then on $\mathscr{H}$ there exist a self-adjoint (linear) operator R, a (linear) unitary U, and an antilinear involution J commuting with R (that is, $J = J^* = J^{-1}$ and $J\phi(R) = \phi(R)^* J$ for every bounded measurable function $\phi: \mathbb{R} \to \mathbb{C}$), so that

$$T_l \subseteq U \cosh(R), \quad T_a \subseteq UJ \sinh(R),$$
$$(T^{-1})_l \subseteq \cosh(R) U^*, \quad (T^{-1})_a \subseteq -J \sinh(R) U^*,$$

and all these operators leave E invariant [T_l is the linear part and T_a the antilinear part of T according to (2.1)]. This decomposition is unique, if $R \geqslant 0$.

Moreover, the following assertions hold:

(a) *Let $T \in \mathscr{T}(E)$. If one of the mappings R, T, T_l, resp., T_a is bounded, then all are bounded.*
(b) *If $T \in \mathscr{T}(E)$ is bounded, then T extends by continuity to an element of $\mathscr{T}(\mathscr{H})$.*
(c) *If T is symplectic on the whole of $\mathscr{H}$, that is $T \in \mathscr{T}(\mathscr{H})$, then T is bounded.*
Thus, $\{T \in \mathscr{T}(E) | T \text{ is bounded}\} \equiv \{T \in \mathscr{T}(\mathscr{H}) | T(E) = E\}$ is a subgroup of $\mathscr{T}(\mathscr{H})$.

A symplectic one-parameter group is a family $\{T_t | t \in \mathbb{R}\} \subset \mathscr{T}(E)$, which satisfies the group properties $T_0 = \mathbf{1}$ and $T_s T_t = T_{s+t}$, $\forall s,t \in \mathbb{R}$. In the literature the notion of strong continuity of (semi-) groups is treated on *complete* spaces.[2] Since symplectic transformations are real linear, the above lemma suggests we take the (complete) real Hilbert space $\mathscr{H}_r$, which coincides with $\mathscr{H}$, but is equipped with the real scalar product $\mathrm{Re}\,\langle \cdot | \cdot \rangle$, when considering strong continuity for symplectic one-parameter groups. That is, a *strongly continuous* symplectic group $\{T_t | t \in \mathbb{R}\}$ consists of symplectic transformations $T_t \in \mathscr{T}(\mathscr{H})$ [each $T_t \in \mathscr{T}(E)$ has to be bounded and thus extends continuously to the whole of $\mathscr{H}$] and fulfills the continuity $\lim_{t \to 0} \|T_t f - f\| = 0$ for every $f \in \mathscr{H}_r \equiv \mathscr{H}$.

In general, a strongly continuous symplectic group $T_t \in \mathscr{T}(\mathscr{H})$, $t \in \mathbb{R}$, has the growth property $\|T_t\| \leqslant c \exp\{\beta|t|\}$, $\forall t \in \mathbb{R}$, for some $\beta \geqslant 0$ and $c \geqslant 1$; cf., e.g., Ref. 2, Proposition 3.1.3. If there is no $\beta \geqslant 0$ such that c may be chosen equal to unity, then by Ref. 6 the antilinear part G_a of the real-linear generator G of $T_t = \exp\{tG\}$, has to be unbounded, provided that G_a exists as a densely defined (closable) operator.

In Ref. 6 the generator G of a strongly continuous symplectic group $T_t = \exp\{tG\} \in \mathscr{T}(\mathscr{H})$, $t \in \mathbb{R}$, with the growth property $\|T_t\| \leqslant \exp\{\beta|t|\}$, $\forall t \in \mathbb{R}$ for some $\beta \geqslant 0$ is determined to be of the form $G = iC + D$ with a self-adjoint linear (possibly unbounded) operator C on $\mathscr{H}$ and an antilinear self-adjoint bounded operator D on $\mathscr{H}$. Conversely, for a self-adjoint linear C the group $u_t := \exp\{itC\}$, $t \in \mathbb{R}$, regarded as a strongly continuous unitary group on the real Hilbert space $\mathscr{H}_r$, may be perturbed with an antilinear bounded D,[2]

$$\exp\{t(iC+D)\} = \sum_{n=0}^{\infty} \int_{t_n=0}^{t} dt_n \cdots \int_{t_1=0}^{t_2} dt_1 \; u_{t_1} D u_{t_2 - t_1} D \cdots u_{t_n - t_{n-1}} D u_{t - t_n}, \tag{2.3}$$

where the norm convergence is ensured by the boundedness of D. Then the series (2.3) gives a strongly continuous symplectic group in $\mathscr{T}(\mathscr{H})$ with the growth estimates $\|\exp\{t(iC+D)\}\| \leqslant \exp\{\|D\||t|\}$ for all $t \in \mathbb{R}$. Since the u_t are linear and the perturbation D is antilinear, we have that the linear part of the real-linear $\exp\{t(iC+D)\}$ is given by summing over the even powers in the expansion series (2.3), and the antilinear part by summing over the odd ones.

B. Unitary implementations

The C^*-algebra for the Boson system with test function space E is the Weyl algebra $\mathscr{W}(E)$ over E—the CCR-algebra—, which is generated by the unitary Weyl operators $W(f)$, $f\in E$, satisfying Ref. 2, Theorem 5.2.8.

$$W(f)W(g)=\exp\{-i/2\,\mathrm{Im}\langle f|g\rangle\}W(f+g),\quad W(f)^*=W(-f),\quad \forall f,\quad g\in E. \tag{2.4}$$

The self-adjoint field operators for the regular representation Π of $\mathscr{W}(E)$ are given by

$$\Phi_\Pi(f):=-i\,\frac{d}{dt}\,\Pi(W(tf))|_{t=0},\quad f\in E. \tag{2.5}$$

The functional $E\ni f\mapsto\Phi_\Pi(f)$ is real linear. The smeared annihilation and creation operators,

$$a_\Pi(f):=2^{-1/2}(\Phi_\Pi(f)+i\Phi_\Pi(if)),\quad a_\Pi^*(f):=2^{-1/2}(\Phi_\Pi(f)-i\Phi_\Pi(if)), \tag{2.6}$$

respectively, fulfill the CCR $[a_\Pi(f),a_\Pi^*(g)]=\langle f|g\rangle\mathbb{1}$, $\forall f,\,g\in E$. It is $a_\Pi(f)^*=a_\Pi^*(f)$, the map $E\ni f\mapsto_\Pi(f)$ is antilinear and $E\ni f\mapsto a_\Pi^*(f)$ is linear (e.g., Ref. 2, Subsection 5.2.3).

For each $T\in\mathscr{T}(E)$ there exists a (unique) *-automorphism α_T on $\mathscr{W}(E)$ with

$$\alpha_T(W(f))=W(Tf),\quad \forall f\in E, \tag{2.7}$$

which is called the *Bogoliubov transformation associated with* T. Thus, the mapping $\mathscr{T}(E)\ni T\mapsto\alpha_T$ is a representation of the group $\mathscr{T}(E)$ as *-automorphisms on $\mathscr{W}(E)$.

Definition 2.2 (implementation): Consider a representation Π *of the CCR-algebra* $\mathscr{W}(E)$ *on the representation Hilbert space* $\mathscr{K}_\Pi$.

(a) *The Bogoliubov transformation* α_T *on* $\mathscr{W}(E)$, *or briefly the associated symplectic* $T\in\mathscr{T}(E)$, *is denoted to be (unitarily) implementable in* Π, *if there exists a unitary operator* U *(depending on the symplectic* T*) acting on* $\mathscr{K}_\Pi$, *such that*

$$\Pi(\alpha_T(A))=U\Pi(A)U^*,\quad \forall A\in\mathscr{W}(E). \tag{2.8}$$

The set $\mathscr{T}_\Pi(E)$ *of those symplectic transformations* $T\in\mathscr{T}(E)$, *that are implementable in* Π, *constitutes a subgroup of* $\mathscr{T}(E)$. *But, in general,* $\mathscr{T}_\Pi(E)\neq\mathscr{T}(E)$.

(b) *The one-parameter group* $\{T_t|t\in\mathbb{R}\}$ *of symplectic transformations* $T_t\in\mathscr{T}_\Pi(E)$, $\forall t\in\mathbb{R}$ *is denoted to be continuously implementable in* Π, *if the associated implementing unitaries* U_t, $t\in\mathbb{R}$, *on* $\mathscr{K}_\Pi$ *may be chosen as a strongly continuous one-parameter group, that is,* $U_t=\exp\{itH_\Pi\}$ *with some implementing self-adjoint operator* H_Π *on* $\mathscr{K}_\Pi$, *such that*

$$\Pi(\alpha_{T_t}(A))=\exp\{itH_\Pi\}\Pi(A)\exp\{-itH_\Pi\},\quad \forall A\in\mathscr{W}(E),\quad \forall t\in\mathbb{R}. \tag{2.9}$$

The implementing U and H_Π from (2.8), resp. (2.9), are, in general, nonunique. However, if the representation Π is irreducible, then U is unique up to a phase, and H_Π is unique up to a constant.

Let us remark, that $T_t\in\mathscr{T}_\Pi(E)$ for all $t\in\mathbb{R}$ with the implementing unitaries U_t, $t\in\mathbb{R}$, from (2.8) does not imply the continuous implementability of the symplectic group $\{T_t|t\in\mathbb{R}\}$. In general, the unitaries U_t, $t\in\mathbb{R}$ are not even choosable to constitute a group.

C. Representations symplectically related to Fock

In the present work we exclusively are interested in the Fock representation Π_F, and in representations $\Pi_T:=\Pi_F\circ\alpha_T$ of $\mathscr{W}(E)$, where $T\in\mathscr{T}(E)$. Obviously, for $T=\mathbb{1}$ we have the Fock representation $\Pi_1=\Pi_F$. The representation Π_T is called *symplectically related* to the Fock representation.[7] We briefly write $\mathscr{T}_F(E)$ and $\mathscr{T}_T(E)$, for the groups consisting of those symplectic $\widetilde{T}\in\mathscr{T}(E)$, which are implementable in Π_F, resp., in Π_T.

By construction the representation Hilbert space for the symplectically related representation $\Pi_T=\Pi_F\circ\alpha_T$ of $\mathscr{W}(E)$, $T\in\mathscr{T}(E)$, is the same as for the Fock representation Π_F, namely the Bose–Fock space $F_+(\mathscr{H})$, and we stick to this realization in the following for convenience.

The field, annihilation, and creation operators associated with Π_F are denoted by an index F, that is, $\Phi_F(f)$, $a_F(f)$, resp., $a_F^*(f)$, $f\in E$, and those associated with Π_T by an index T, that is $\Phi_T(f)$, $a_T(f)$, resp. $a_T^*(f)$, $f\in E$, according to Eqs. (2.5) and (2.6). All these operators act on the Fock space $F_+(\mathscr{H})$, and it holds that

$$\Phi_T(f)=\Phi_F(Tf),\quad a_T(f)=a_F(T_lf)+a_F^*(T_af),\quad \forall f\in E. \tag{2.10}$$

It is well known that the Fock representation Π_F is the GNS representation (cf. e.g., Ref. 2, Definition 2.3.18) of the vacuum state ω_F on $\mathscr{W}(E)$, which is uniquely given by

$$\langle\omega_F;W(f)\rangle=\exp\{-\tfrac{1}{4}\|f\|^2\}=\langle\Omega_F|\Pi_F(W(f))\Omega_F\rangle,\quad \forall f\in E,$$

where $\Omega_F\in F_+(\mathscr{H})$ is the Fock vacuum vector. Thus $(\Pi_T,F_+(\mathscr{H}),\Omega_F)$ is (unitarily equivalent to) the GNS triple of the Bogoliubov transformed vacuum state:

$$\omega_T:=\omega_F\circ\alpha_T,$$

the so-called squeezed vacuum.[1,11,12,5] We see that the GNS triple for ω_T is realized with the Fock space $F_+(\mathscr{H})$ as a GNS Hilbert space. The Fock representation Π_F is irreducible,[2] and thus Π_T is so, too. Hence, the associated von Neumann algebra $\Pi_T(\mathscr{W}(E))''$ is equal to $\mathscr{B}(F_+(\mathscr{H}))$, the algebra of all bounded operators on Fock space.

If $T\in\mathscr{T}(E)$ is bounded, then it extends continuously to $T\in\mathscr{T}(\mathscr{H})$ by Lemma 2.1, and thus then one may extend Π_T to the enlarged Weyl algebra $\mathscr{W}(\mathscr{H})\supseteq\mathscr{W}(E)$, since $\mathscr{H}\ni f\mapsto\Pi_T(W(f))=\Pi_F(W(Tf))$ is continuous in the strong operator topology by Ref. 2, Proposition 5.2.4. For the general theory concerning continuous enlargements of testfunction spaces and representations, we refer to Ref. 13.

Let us introduce the folium $\mathscr{F}_T$ consisting of the Π_T-normal states on $\mathscr{W}(E)$ (cf., e.g., Ref. 2, Definition 2.4.25, Ref. 14, Section 1, 15, 16), that is, for each $\omega\in\mathscr{F}_T$ there is a density operator $\rho_{\omega,T}$ on Fock space such that

$$\langle\omega;A\rangle=\mathrm{tr}[\rho_{\omega,T}\Pi_T(A)],\quad \forall A\in\mathscr{W}(E),\quad \omega\in\mathscr{F}_T. \tag{2.11}$$

Since Π_T is irreducible, Eq. (2.11) defines a one-to-one correspondence between $\mathscr{F}_T$ and the density operators on Fock space. For $T=\mathbf{1}$ we obtain the folium $\mathscr{F}_1\equiv\mathscr{F}_F$ of the Fock-normal states on $\mathscr{W}(E)$.

In the case of $T\in\mathscr{T}_F(E)$ there is, by Definition 2.2(a), a unitary U_T on Fock space (unique up to a phase), such that

$$\Pi_T(A)=\Pi_F(\alpha_T(A))=U_T\Pi_F(A)U_T^*,\quad \forall A\in\mathscr{W}(E). \tag{2.12}$$

This relation expresses the unitary equivalence of Π_F and Π_T. Moreover, for $T\in\mathscr{T}_F(E)$ the Bogoliubov transformed field operators (2.10) are obtained with the unitary transformation U_T,

$$\Phi_T(f)=\Phi_F(Tf)=U_T\Phi_F(f)U_T^*,\quad \forall f\in E. \tag{2.13}$$

Theorem 2.3: *Let $T\in\mathscr{T}(E)$. Then the following holds:*
(a) The following conditions are equivalent:

(i) $T\in\mathscr{T}_F(E)$;
(ii) *the closure of the antilinear part T_a of T is Hilbert–Schmidt[17] on $\mathscr{H}$;*
(iii) *there is a normalized $\psi_T\in F_+(\mathscr{H})$ with $a_T(f)\psi_T=0\,\forall f\in E$ (where then $\psi_T=U_T\Omega_F$ up to a phase);*
(iv) $\omega_T\in\mathscr{F}_F$ *(or equivalently,* $\omega_F\in\mathscr{F}_T$);
(v) $\mathscr{F}_T=\mathscr{F}_F$;
(vi) Π_T *and* Π_F *are quasiequivalent (and thus unitarily equivalent by their irreducibility);*

(vii) *there is an *-automorphism* γ_T *on* $\mathcal{B}(F_+(\mathcal{H}))$ *such that* $\gamma_T(\Pi_F(A))=\Pi_T(A)$, $\forall A \in \mathcal{W}(E)$ *(which then is given by* U_T *according to Eq. (2.12)).*

(b) The following conditions are equivalent:

(i) $T \notin \mathcal{T}_F(E)$;
(ii) $\mathcal{T}_T \cap \mathcal{T}_F = \emptyset$;
(iii) Π_T *and* Π_F *are disjoint representations of* $\mathcal{W}(E)$ *(e.g., Ref. 2, Subsection 4.2.2);*
(iv) *the (algebraic) transition probability (cf. Ref. 18) between the vacuum* $\omega_F \in \mathcal{T}_F$ *and every other state* $\omega \in \mathcal{T}_T$ *is zero.*

Proof: (a)(i)⇔(a)(ii) is proven in Ref. 19 (cf. also Refs. 20, 21, Sec. IX, and 22), and is generalized in Ref. 6. (a)(i)⇔(a)(iii)⇔(a)(iv) is also shown in Ref. 6. (a)(iv)⇔(a)(v) follows from the fact that $\mathcal{T}_T$ (resp., $\mathcal{T}_F$) is the smallest folium in the state space of $\mathcal{W}(E)$ containing ω_T (resp. ω_F). The equivalences (a)(v)⇔(a)(vi)⇔(a)(vii) are standard results for quasiequivalence, e.g. Ref. 2, Subsection 2.4.4, applied to our specific situation. (b)(i)⇔(b)(ii)⇔(b)(ii) is the negation of (a), since Π_T and Π_F are the (irreducible) GNS representations of the pure states ω_T, resp. ω_F. (b)(iii)⇔(b)(iv) is proved in Ref. 18, Proposition 2.8. ■

We see that the antilinear part T_a being Hilbert–Schmidt induces a lot of further properties. For the use of the folium $\mathcal{T}_T$ for a generalized Schrödinger dynamics we refer to Ref. 15.

Let us remark, that for $T \in \mathcal{T}_F(E)$ we have T_a to be Hilbert–Schmidt by the above theorem, which yields T_a and thus T to be bounded. Consequently, T extends continuously to an element of $\mathcal{T}_F(\mathcal{H})$ by Lemma 2.1. That is, $\mathcal{T}_F(E)$ is the subgroup of $\mathcal{T}_F(\mathcal{H})$ consisting of those $T \in \mathcal{T}_F(\mathcal{H})$ with $T(E)=E$.

D. Implementations in the Fock representation

By Theorem 2.3 the symplectic $T \in \mathcal{T}(\mathcal{H})$ is implementable in the Fock representation Π_F, if and only if its antilinear part T_a is Hilbert–Schmidt on $\mathcal{H}$. Let us turn to the implementability of symplectic *one-parameter groups* of the type (2.3) and communicate a result proven in Ref. 6.

Theorem 2.4: *For linear* $C=C^*$ *and antilinear Hilbert–Schmidt* $D=D^*$ *on* $\mathcal{H}$*, it holds that* $\exp\{t(iC+D)\} \in \mathcal{T}_F(\mathcal{H}) \forall t \in \mathbb{R}$ *(cf. Section II A). Moreover, the symplectic one-parameter group* $\{\exp\{t(iC+D)\} | t \in \mathbb{R}\}$ *is continuously implementable in* Π_F.

E. Implementations in symplectically related representations

As discussed above, the representation Hilbert space of the symplectically related representation $\Pi_T=\Pi_F \circ \alpha_T$ of $\mathcal{W}(E)$ is the Fock space. Thus, the implementability in Π_T leads back to the implementability in Π_F, but for the transformed symplectic transformation:

Lemma 2.5: For $T \in \mathcal{T}(E)$ *consider the symplectically related representation* $\Pi_T=\Pi_F \circ \alpha_T$. *Then for each* $\tilde{T} \in \mathcal{T}(E)$ *it holds:* $\tilde{T} \in \mathcal{T}_T(E)$, *if and only if* $T\tilde{T}T^{-1} \in \mathcal{T}_F(E)$.

Furthermore, the associated implementing unitaries U, *for* $\tilde{T}$ *in* Π_T *and for* $T\tilde{T}T^{-1}$ *in* Π_F *(both acting on Fock space), agree up to a phase. That is, for all* $f \in E$ *we have, according to the equations (2.7) and (2.8):*

$$\Pi_T(W(\tilde{T}f))=\Pi_T(\alpha_{\tilde{T}}(W(f)))=U\Pi_T(W(f))U^*,$$

$$\Pi_F(W(T\tilde{T}T^{-1}f))=\Pi_F(\alpha_{T\tilde{T}T^{-1}}(W(f)))=U\Pi_F(W(f))U^*.$$

Proof: With Eq. (2.8) and $\Pi_F \circ \alpha_T=\Pi_T$, we have $U\Pi_T(W(f))U^*=(\Pi_T \circ \alpha_{\tilde{T}})(W(f))=\Pi_T(W(\tilde{T}f))$, $\forall f \in E$, if and only if $U\Pi_F(W(Tf))U^*=\Pi_F(W(T\tilde{T}f)) \forall f \in E$. Since T acts bijectively on E, we may put $f=T^{-1}g$, and thus the last equation is equivalent to $U\Pi_F(W(g))U^*=\Pi_F(W(T\tilde{T}T^{-1}g))=(\Pi_F \circ \alpha_{T\tilde{T}T^{-1}})(W(g))$, $\forall g \in E$. Hence the unitary implementability of $\tilde{T}$ in Π_T is equivalent to the unitary implementability of $T\tilde{T}T^{-1}$ in Π_F, and the associated implementing unitaries agree up to a phase since Π_F and Π_T are irreducible. ■

Obviously, this Lemma immediately extends to continuously implementable *one-parameter groups* of symplectic transformations: For $T \in \mathcal{T}(E)$ consider the symplectically related representation $\Pi_T=\Pi_F \circ \alpha_T$ of $\mathcal{W}(E)$. Then the symplectic group $T_t \in \mathcal{T}(E)$, $t \in \mathbb{R}$, is continuously imple-

mentable in Π_T, if and only if the transformed symplectic group $\hat{T}_t := TT_tT^{-1} \in \mathscr{T}(E)$, $t \in \mathbb{R}$, is continuously implementable in Π_F. For the implementing strongly continuous unitary group $U_t = \exp\{itH_{\mathrm{im}}\}$, $t \in \mathbb{R}$, with some self-adjoint H_{im} on Fock space (which is unique up to a constant), it follows according to Eq. (2.9) that

$$\begin{aligned} \Pi_T(W(T_t f)) &= \Pi_T(\alpha_{T_t}(W(f))) = \exp\{itH_{\mathrm{im}}\}\Pi_T(W(f))\exp\{-itH_{\mathrm{im}}\}, \\ \Pi_F(W(TT_tT^{-1}f)) &= \Pi_F(\alpha_{TT_tT^{-1}}(W(f))) = \exp\{itH_{\mathrm{im}}\}\Pi_F(W(f))\exp\{-itH_{\mathrm{im}}\}. \end{aligned} \tag{2.14}$$

For the norm-complete test function space $E = \mathscr{H}$, Warchall investigated in Ref. 7 the implementability in Π_T of a unitary (and thus symplectic) one-parameter group $T_t = \exp\{itC\}$, $t \in \mathbb{R}$, on $\mathscr{H}$ with a self-adjoint linear generator $C \geqslant 0$. The unitary group $\{\exp\{itC\} | t \in \mathbb{R}\}$ is continuously implementable in the Fock representation by Segal's second quantized $\Gamma(\exp\{itC\}) = \exp\{it d\Gamma(C)\}$, cf., e.g., Ref. 2, Subsection 5.2.1 or Ref. 3, Sec. X.7.

Definition 2.6 (infrared part): Since the neighborhood of the zero frequency determines the infrared behavior of a Hamiltonian, we call here the self-adjoint operator C on $\mathscr{H}$ to have an essential infrared part, if the dimension of $P_C([-a,a])\mathscr{H}$ is infinite for every $a > 0$, where $P_C(\cdot)$ is the spectral projection valued measure for C.

If $P_C([-\varepsilon,\varepsilon])\mathscr{H}$ is finite for some $\varepsilon > 0$, then C has an inessential infrared part (which, e.g., is valid for strictly positive C, i.e., $C \geqslant c\mathbb{1}$ for some $c > 0$).

Observe that C has an essential infrared part, if and only if zero is an element of the essential spectrum of the self-adjoint linear operator C, Ref. 3, Subsection VII.2. Using Definition 2.6 we are able to cast a main result of Ref. 7 into the following concise form.

Theorem 2.7 (Warchall): *Let $C \geqslant 0$ be a self-adjoint operator on $\mathscr{H}$. We denote by $\mathscr{C}_{\mathrm{im}}(C)$ the class of those $T \in \mathscr{T}(\mathscr{H})$, for which $\{\exp\{itC\} | t \in \mathbb{R}\}$ is continuously implementable in Π_T. (Each $T \in \mathscr{T}(\mathscr{H})$ is bounded by Lemma 2.1.) The following holds*:

(a) *If C has an inessential infrared part, then $\mathscr{C}_{\mathrm{im}}(C) = \mathscr{T}_F(\mathscr{H})$.*
(b) *Let C have an essential infrared part. Then the inclusion $\mathscr{T}_F(\mathscr{H}) \subset \mathscr{C}_{\mathrm{im}}(C)$ is proper. However, one has in general $\mathscr{C}_{\mathrm{im}}(C) \neq \mathscr{T}(\mathscr{H})$.*

Thus the group $\{\exp\{itC\} | t \in \mathbb{R}\}$ is continuously implementable in some symplectically related representations Π_T inequivalent (disjoint) to the Fock representation Π_F (cf. Theorem 2.3), if and only if C has an essential infrared part.

In the Secs. III B and IV C, we generalize Warchall's result in a certain sense.

By definition, $\mathscr{T}_F(E)$ is the group of those symplectic transformations T on E for which T is unitarily implementable in the Fock representation Π_F, which immediately implies the following result concerning the implementability in Π_F.

Lemma 2.8: Suppose $T \in \mathscr{T}_F(E)$ in the situation of the Lemma 2.5. Because of the group property of $\mathscr{T}_F(E)$ we have the following equivalence; $\tilde{T} \in \mathscr{T}(E)$ is implementable in Π_F, if and only if $T\tilde{T}T^{-1}$ is implementable in Π_F, or in other words, $\tilde{T}$ is implementable in Π_F, if and only if $\tilde{T}$ is implementable in Π_T. Assume in this case that U_T, $U_{\tilde{T}}$, and $U_{T\tilde{T}T^{-1}}$ are the unitaries implementing T, $\tilde{T}$, resp., $T\tilde{T}T^{-1}$ in Π_F, then $U \equiv U_{T\tilde{T}T^{-1}} = U_T U_{\tilde{T}} U_T^*$ (up to a phase; the implementing U is from Lemma 2.5), which also implements $\tilde{T}$ in Π_T.

Furthermore, the symplectic group $T_t \in \mathscr{T}(E)$, $t \in \mathbb{R}$, is continuously implementable in Π_F, if and only if the transformed symplectic group $\hat{T}_t := TT_tT^{-1} \in \mathscr{T}(E)$, $t \in \mathbb{R}$, is continuously implementable in Π_F. If H and $\hat{H}$ are the implementing self-adjoint operators for T_t, resp., $\hat{T}_t$ in Π_F, then $H_{\mathrm{im}} \equiv \hat{H} = U_T H U_T^*$ [up to an additive constant; the implementing H_{im} is from the equation (2.14)].

III. SYMPLECTICALLY TRANSFORMED GROUPS

Let $\mathscr{H}$ be an infinite-dimensional separable complex Hilbert space with fixed ONB $\{e_k | k \in \mathbb{N}\}$. Let C, S, and R be possibly unbounded, self-adjoint, linear operators on $\mathscr{H}$ having the e_k as eigenvectors:

$$Ce_k = c_k e_k, \quad Se_k = s_k e_k, \quad \mathrm{Re}_k = r_k e_k, \quad \text{with } c_k, s_k, r_k \in \mathbb{R}, \quad \forall k \in \mathbb{N}. \tag{3.1}$$

Finally, let J be the antilinear involution on $\mathscr{H}$ satisfying $Je_k = e_k$, $\forall k \in \mathbb{N}$. Obviously, C, S, R, and J mutually commute [cf. the Eqs. (1.3), (1.4), and (1.6), but in contrast to the Introduction here also $s_k < 0$ is allowed].

For convenience we choose from now on the one-particle test function space E as the space of all finite (complex-) linear combinations of our ONB $\{e_k | k \in \mathbb{N}\}$. The exponentials,

$$T_t := \exp\{it(C - JS)\} \in \mathscr{T}(E), \quad t \in \mathbb{R}, \tag{3.2}$$

constitute a well-defined one-parameter group of symplectic transformations on E, which follows from the application of Sec. II A [cf. Eq. (2.3) with $\mathrm{D} = -iJS$) to each invariant subspace $T_t(\mathbb{C}e_k) = \mathbb{C}e_k$, $k \in \mathbb{R}$. Another group of this kind is introduced by a *symplectic shift* as follows:

$$\hat{T}_t := TT_tT^{-1} \in \mathscr{T}(E), \quad t \in \mathbb{R}, \tag{3.3}$$

where

$$T := \exp\{JR\} = \cosh(R) + J\sinh(R) \in \mathscr{T}(E). \tag{3.4}$$

Because C, S, R, J, and thus $T_t, T, \hat{T}_t$ decompose according to $\mathscr{H} = \oplus_{k=1}^{\infty} \mathbb{C}e_k$, it holds that $\lim_{t\to 0} \|T_t f - f\| = 0$ and $\lim_{t\to 0} \|\hat{T}_t f - f\| = 0$ for every $f \in E$. Since E is not complete, these continuity relations do not imply the strong continuity of both groups on the complete $\mathscr{H}$, as described in Sec. II A. In general, T_t, resp. $\hat{T}_t$, and T may be unbounded symplectic transformations on E (by Lemma 2.1, since E is not complete).

We stick to the notation (3.1)–(3.4) throughout the present and the following sections.

A. The generator of the shifted group

The following result ensures that shifts with symplectic transformations of the form (3.4) leave the class (3.2) of symplectic groups invariant, that is, $\hat{T}_t$ is of the type (3.2), too.

Lemma 3.1: The generator of the symplectic group $\{\hat{T}_t | t \in \mathbb{R}\} \subset \mathscr{T}(E)$ is given by $i(\hat{C} - J\hat{S})$, that is, $\hat{T}_t = \exp\{it(\hat{C} - J\hat{S})\}$, $\forall t \in \mathbb{R}$, with the self-adjoint linear operators on $\mathscr{H}$,

$$\hat{C} := \cosh(2R)C + \sinh(2R)S, \quad \hat{S} := \cosh(2R)S + \sinh(2R)C, \tag{3.5}$$

and the e_k, $k \in \mathbb{N}$, are the eigenvectors for $\hat{C}$ and $\hat{S}$ with the associated eigenvalues,

$$\hat{c}_k := \cosh(2r_k)c_k + \sinh(2r_k)s_k, \quad s_k := \cosh(2r_k)s_k + \sinh(2r_k)c_k. \tag{3.6}$$

Conversely, C and S are obtained from $\hat{C}$ and $\hat{S}$ by

$$C = \cosh(2R)\hat{C} - \sinh(2R)\hat{S}, \quad S = \cosh(2R)\hat{S} - \sinh(2R)\hat{C}.$$

Proof: It follows from (3.3) that $\hat{T}_t = \exp\{t\hat{G}\}$, and thus its generator $\hat{G}$ leaves each $\mathbb{C}e_k$, $k \in \mathbb{N}$, invariant, E being a complex-linear space we may decompose $\hat{G}$ into its linear and antilinear part, $\hat{G}_l =: i\hat{C}$ and $\hat{G}_a =: -iJ\hat{S}$, respectively. Now $Ti(C - JS)T^{-1} = i(\hat{C} - J\hat{S})$, the decomposition (3.4), and $T^{-1} = \exp\{-JR\} = \cosh(R) - J\sinh(R)$, and the formulas

$$\cosh(x)^2 = \sinh(x)^2 + 1, \quad 2\sinh(x)^2 = \cosh(2x) = -1, \quad 2\cosh(x)\sinh(x) = \sinh(2x) \tag{3.7}$$

imply the remaining statements. ■

We now study the growth estimations of the groups T_t and $\hat{T}_t$ in the case where they constitute strongly continuous one-parameter groups on the whole of the real Hilbert space $\mathscr{H}_r$ ($\mathscr{H}$ equipped with the real scalar product $\mathrm{Re}\langle \cdot | \cdot \rangle$), that is, $\|T_t\| < \infty$ for all $t \in \mathbb{R}$ and thus each T_t extends continuously to the whole of $\mathscr{H}$ by Lemma 2.1, and $\lim_{t\to 0} \|T_t f - f\| = 0$ for all $f \in \mathscr{H}$, and analogous relations for $\hat{T}_t$ (see Sec. II A).

Theorem 3.2 (growth property): *If S is bounded, then the group $\{T_t|t\in\mathbb{R}\}$ is strongly continuous on $\mathscr{H}_r$ with $\|T_t\|\leqslant\exp\{\|S\||t|\}$ for all $t\in\mathbb{R}$.*

Let S and R both be bounded. Then $\|T\|=\exp\{\|R\|\}=\|T^{-1}\|$, and the symplectic one-parameter group $\{\hat{T}_t|t\in\mathbb{R}\}$ is strongly continuous on $\mathscr{H}_r$ with $\|\hat{T}_t\|\leqslant\exp\{2\|R\|\}\exp\{\|S\|t\}$ for all $t\in\mathbb{R}$ (also for unbounded $\hat{S}$). But $\hat{T}_t$ has the growth estimation $\|\hat{T}_t\|\leqslant\exp\{\beta|t|\}$, $\forall t\in\mathbb{R}$ with some $\beta\geqslant 0$, if and only if $\hat{S}$ is bounded (in which case β may be chosen as $\|\hat{S}\|$).

Proof: For bounded T (which is equivalent with R being bounded by Lemma 2.1) it follows from the construction (3.3) that the shifted group $\hat{T}_t$ is strongly continuous on $\mathscr{H}_r$, if and only if T_t is strongly continuous on $\mathscr{H}_r$. Now our assertions are consequences of the results stated in Sec. II A. ■

From (3.5) it is seen that for unbounded C but bounded S and bounded R, it may occur that $\hat{S}$ is unbounded, too [e.g., if $C\geqslant 0$ and $R\geqslant\gamma\mathbb{1}$ for some $\gamma>0$, then $\hat{S}\geqslant\sinh(2\gamma)C$]. The above Theorem then implies the growth estimation $\|\hat{T}_t\|\leqslant c\ \exp\{\beta|t|\}$, $\forall t\in\mathbb{R}$ with some $c>1$ and $\beta\geqslant 0$, and one cannot find a $\beta\geqslant 0$ such that c may be chosen as $c=1$. Thus, we have here some examples of strongly continuous groups (not only semigroups), where it is not possible to put $c=1$, supplementing the rare instances in the literature; cf. Ref. 23, p. 14.

B. Unitary implementations

Let us first consider the symplectic shift $T=\exp\{JR\}$ from Eq. (3.4). By Theorem 2.3 the condition $T\in\mathscr{T}_F(E)$ is equivalent to the unitary equivalence of the representations Π_F and Π_T (which then physically are not distinguishable). This contrasts the case $T\notin\mathscr{T}_F(E)$, where Π_F and Π_T are inequivalent (disjoint) representations of $\mathscr{W}(E)$, and thus describe the macroscopically different sectors $\mathscr{F}_F$ and $\mathscr{F}_T$ (disjoint folia) of the Boson system.

Proposition 3.3: For our $T=\exp\{JR\}\in\mathscr{T}(E)$ we have the following equivalences:

(i) *R is Hilbert–Schmidt on $\mathscr{H}$ (especially, R is bounded)*;
(ii) $\sinh(R)$ *is Hilbert–Schmidt on $\mathscr{H}$*;
(iii) $\tanh(R)$ *is Hilbert–Schmidt on $\mathscr{H}$*;
(iv) *$T=\exp\{JR\}\in\mathscr{T}_F(E)$ (especially, T is bounded).*

Proof: From Theorem 2.3 and Eq. (3.4), it follows that $T\in\mathscr{T}_F(E)$, if and only if $T_a=J\sinh(R)$ is Hilbert–Schmidt, or equivalently, if and only if $\sinh(R)$ is Hilbert–Schmidt (since J is an antiunitary). Let us first assume $\sinh(R)$ to be Hilbert–Schmidt. Then from $\sinh(x)=\Sigma_{n\,\mathrm{odd}}x^n/n!\geqslant x$ for $x\geqslant 0$, one gets

$$\|\sinh(R)\|_{\mathrm{HS}}^2=\sum_{k=1}^{\infty}\ \sinh(r_k)^2=\sum_{k=1}^{\infty}\ \sinh(|r_k|)^2\geqslant\sum_{k=1}^{\infty}\ |r_k|^2,$$

which implies R to be Hilbert–Schmidt, too. The converse relation, that R Hilbert–Schmidt implies $\sinh(R)$ to be Hilbert–Schmidt, follows from the power series for $\sinh(R)$, respectively, from Theorem 2.4. Now observe that the Hilbert–Schmidt operators on $\mathscr{H}$ constitute an ideal in the algebra of all bounded operators on $\mathscr{H}$. But R bounded implies $\cosh(R)$ and $\cosh(R)^{-1}$ to be bounded. Then the equivalence of (ii) and (iii) follows from $\cosh(R)\tanh(R)=\sinh(R)$, and from the fact that $\sinh(R)$, resp., $\tanh(R)$ Hilbert–Schmidt yields R bounded ($\Sigma_k\tanh(r_k)^2<\infty$ implies $\sup_k|\tanh(r_k)|<1$, which yields $\sup_k|r_k|<\infty$, and thus R is bounded). ■

If S is Hilbert–Schmidt, then Theorem 2.4 implies $T_t\in\mathscr{T}_F(E)$, $\forall t\in\mathbb{R}$. Beside this result, we have another case, which allows S not to be Hilbert–Schmidt.

Proposition 3.4: Let $SC^{-1}|_{\ker(C)^\perp}+S|_{\ker(C)}$ be Hilbert–Schmidt on $\mathscr{H}$ (here $\ker(C)^\perp$ is the (complex) orthogonal complement of the kernel $\ker(C)$ of C in $\mathscr{H}$), that is $\Sigma_{k,c_k\neq 0}(s_k^2/c_k^2)+\Sigma_{k,c_k=0}s_k^2<\infty$. Then we have $T_t\in\mathscr{T}_F(E)$, $\forall t\in\mathbb{R}$.

Proof: For $t\in\mathbb{R}$ consider the continuous functions $f_t:\mathbb{R}\to\mathbb{C}$ with $f_t(x)=(\exp\{-i2tx\}-1)/x$ for $x\neq 0$ and $f_t(0)=-i2t$. Since $J\exp\{i\tau C\}=\exp\{-i\tau C\}J$, $\forall\tau\in\mathbb{R}$ and $[S,C]=0$ we have for each $t\in\mathbb{R}$,

$$F_t := \int_0^t \exp\{i\tau C\}(-iJS)\exp\{-i\tau C\}\mathrm{d}\tau = \frac{1}{2} JSf_t(C).$$

Now the assumption exhibits $t \in \mathbb{R} \mapsto \|F_t\|_{\mathrm{HS}}$ to be continuous. Finally, use the expansion series for $(T_t)_a$ [summing over the odd numbers in the Dyson series (2.3)] to show that $(T_t)_a$ is Hilbert–Schmidt. ■

Now we turn to the continuous implementability of the symplectic *one-parameter groups* T_t and $\hat{T}_t$, $t \in \mathbb{R}$. By Lemma 2.5, $\{T_t | t \in \mathbb{R}\}$ is continuously implementable in the symplectically related representation $\Pi_T = \Pi_F \circ \alpha_T$, if and only if $\{\hat{T}_t | t \in \mathbb{R}\}$ is continuously implementable in the Fock representation Π_F.

First, let us consider the case $T \in \mathscr{T}_F(E)$, where Π_F and Π_T are unitarily equivalent.

Proposition 3.5: Let S and R be Hilbert–Schmidt. Then $\{T_t | t \in \mathbb{R}\}$ *and* $\{\hat{T}_t | t \in \mathbb{R}\}$ *are continuously implementable in the Fock representation* Π_F.

Proof: Consequence of Lemma 2.8, Proposition 3.3, and Theorem 2.4. ■

Example 3.6: Even in the situation of the above proposition, where S and R both are Hilbert–Schmidt (and thus bounded), the operator C may be chosen such that $\hat{S}$ *is unbounded: Let,* e.g., $r_k > 0 \forall k \in \mathbb{N}$, *and choose C with* $c_k \geqslant k/\sinh(2r_k)$, $\forall k \in \mathbb{R}$. *Then (3.5) yields* $\hat{S}$ *to be unbounded (since* $\hat{s}_k \geqslant k$, $\forall k \in \mathbb{N}$*), and it is nevertheless* $\hat{T}_t \in \mathscr{T}_F(E)$, $\forall t \in \mathbb{R}$.

Suppose S to be Hilbert–Schmidt. Then $\{T_t | t \in \mathbb{R}\}$ is continuously implementable in Π_F by Theorem 2.4. Let us now construct some $T \notin \mathscr{T}_F(E)$ so that $\{T_t | t \in \mathbb{R}\}$ also is continuously implementable in Π_T. We arrive at the mentioned generalization of Warchall's result[7] (cf. Theorem 2.7). Observe that in our context the spectral projections $P_C(\mathscr{B})$ of C ($\mathscr{B} \subseteq \mathbb{R}$ denotes an arbitrary Borel subset) and the operators S, R, and J mutually commute.

Theorem 3.7: *Let us write* $R \in \mathscr{R}(C)$, *whenever our operators R* and *C from (3.1) fulfill the condition that for some* $a > 0$ *the operators* $\tanh(R) C P_C([-a,a])$ *and* $R P_C(\mathbb{R} \backslash [-a,a])$ *both are Hilbert–Schmidt on* $\mathscr{H}$. The following holds.

(I) If R is bounded and S is Hilbert–Schmidt, then for $T = \exp\{JR\}$ *it is equivalent:*

(i) $T_t = \exp\{it(C - JS)\} \in \mathscr{T}_T(E)$ *for all* $t \in \tau$ *for an arbitrary open interval* $\tau \subseteq \mathbb{R}$ *containing the origin,*

(ii) $R \in \mathscr{R}(C)$,

(iii) *the symplectic group* $\{T_t | t \in \mathbb{R}\}$ *is continuously implementable in* Π_T *(which is equivalent to the continuous implementability of* $\{T_t | t \in \mathbb{R}\}$ *in* Π_F*).*

(II) For the condition $\mathscr{R}(C)$ *the following assertions are valid:*

(a) *If R is Hilbert–Schmidt (and thus bounded), then* $R \in \mathscr{R}(C)$.

(b) *Let C have an inessential infrared part. Then* $R \in \mathscr{R}(C)$, *if and only if R is Hilbert–Schmidt.*

(c) *There exist some bounded and some unbounded* $R \in \mathscr{R}(C)$, *which are not Hilbert–Schmidt, if and only if C has an essential infrared part.*

(d) $\mathscr{R}(C)$ *consists of all bounded and all unbounded operators R of the form (3.1), if and only if C is Hilbert–Schmidt on* $\mathscr{H}$.

Proof: HS_r denotes the real-linear Hilbert–Schmidt operators on $\mathscr{H}_r$, and HS the (complex-) linear Hilbert–Schmidt operators on $\mathscr{H}$. It is $\mathrm{HS} \subset \mathrm{HS}_r$. In the proof we use the ideal property of HS (resp. HS) in the algebra of bounded operators on $\mathscr{H}$ (resp., $\mathscr{H}_r$). (I)(iii)⇒(I)(i) is immediate. Let us prove (I)(ii)⇒(I)(iii): We decompose the Fock representation and the Fock space into a tensor product, $\Pi_F = \Pi_{F,1} \otimes \Pi_{F,2}$ on $F_+(\mathscr{H}) = F_+(\mathscr{H}_1) \otimes F_+(\mathscr{H}_2)$, where $\mathscr{H}_1 := P_C \times([-a,a])\mathscr{H}$ and $\mathscr{H}_2 := P_C(\mathbb{R} \backslash [-a,a])\mathscr{H}$. Restricting first to $\mathscr{H}_2$, we have that $S|_{\mathscr{H}_2} \in \mathrm{HS}$ and $R|_{\mathscr{H}_2} \in \mathrm{HS}$. Thus, it follows from Proposition 3.5 that the restricted group $\hat{T}_t|_{\mathscr{H}_2}$ is continuously implementable in $\Pi_{\mathrm{F},2}$. Now we restrict to $\mathscr{H}_1$. If $\hat{S}|_{\mathscr{H}_1} \in \mathrm{HS}$, then the restricted group $\hat{T}_t|_{\mathscr{H}_1}$ is continuously implementable in $\Pi_{F,1}$ by Theorem 2.4. Forming the direct sum $\hat{T}_t$

$=\hat{T}_t|_{\mathscr{H}_1}\oplus\hat{T}_t|_{\mathscr{H}_2}$ yields the result. But $\cosh(R)$ being bounded implies $2\cosh(R)^2\tanh(R)CP_C$ $2\cosh(R)^2\tanh(R)CP_C([-a,a])=\sinh(2R)CP_C([-a,a])\in\mathrm{HS}$, and thus Eq. (3.5) gives $\hat{S}|_{\mathscr{H}_1}$ $\in\mathrm{HS}$.

The proof of (I)(i)⇒(I)(ii) picks up and generalizes some features of Warchall's original proof. For later purposes let us generalize the situation to an arbitrary $X\in\mathscr{T}(E)$. Let $X_l\subseteq W\cosh(Z)$ and $X_a\subseteq WI\sinh(Z)$ be the decomposition of X stated in Lemma 2.1 with the unitary W, the antilinear involution I commuting with the self-adjoint operator Z. Then

$$\begin{aligned}\cosh(Z)^{-1}W^*(XT_tX^{-1})_aW\cosh(Z)^{-1}&=\cosh(Z)^{-1}W^*\{X_a(T_t)_l(X^{-1})_l+X_l(T_t)_a(X^{-1})_l\\&\quad+X_l(T_t)_l(X^{-1})_a+X_a(T_t)_a(X^{-1})_a\}W\cosh(Z)^{-1}\\&=[I\tanh(Z),T_t]-[I\tanh(Z),(T_t)_a]+(T_t)_a\\&\quad-I\tanh(Z)(T_t)_aL\tanh(Z).\end{aligned}$$

But by assumption it is $S\in\mathrm{HS}$, that is, $T_t\in\mathscr{T}_F(E)$ or equivalently $(T_t)_a\in\mathrm{HS}_r$ by Theorem 2.3. Hence, the Dyson expansion (2.3) with $D=-iJS\in\mathrm{HS}_r$ yields the decomposition $T_t=\exp\{itC\}$ $+B_t$ with some $B_t\in\mathrm{HS}_r$. Since $\cosh(Z)^{-1}$ and $\tanh(Z)$ are bounded (also for unbounded Z), we thus have

$$(XT_tX^{-1})_a\in\mathrm{HS}_r\Leftrightarrow[I\tanh(Z),T_t]\in\mathrm{HS}_r\Leftrightarrow[I\tanh(Z),\exp\{itC\}]\in\mathrm{HS}_r.$$

Now for $a\in\mathbb{R}$ let $P_a^+:=P_C([a,\infty[)$ and $P_a^-:=P_C(]-\infty,-a])$, and assume $XT_tX^{-1}\in\mathscr{T}_F(E)$. Since $[P_a^\pm,\exp\{itC\}]=0$, it is $[P_a^\pm I\tanh(Z)P_0^\pm,\exp\{itC\}]=P_a^\pm[I\tanh(Z),\exp\{itC\}]P_0^\pm\in\mathrm{HS}_r$. From arguments similar to the part I⇒III (of the proof) from Ref. 7, Theorem 2, we obtain

$$P_a^-I\tanh(Z)P_0^-\in\mathrm{HS},\quad P_a^+I\tanh(Z)P_0^+\in\mathrm{HS}\quad\forall a>0. \tag{3.8}$$

Here in our concrete situation, however, we have $X=T$ with $W=1$, $I=J$, and $Z=R$, where C, J, and R commute. Thus, (3.8) implies $\tanh(R)P_C(\mathbb{R}\backslash[-a,a])\in\mathrm{HS}$, which with Proposition 3.3 yields $RP_C(\mathbb{R}\backslash[-a,a])\in\mathrm{HS}$ for all $a>0$. Also with the same arguments as in Ref. 7, Theorem 2 one finally gets $\tanh(R)CP_C([-a,a])\in\mathrm{HS}$ for all $a>0$.

Part (II) is easily checked; also see the examples below. ■

The above theorem includes Warchall's result[7] applied to our context: For $S=0$, and $C\geqslant 0$ one has $T=\exp\{JR\}\in\mathscr{C}_{\mathrm{im}}(C)$ [where the class $\mathscr{C}_{\mathrm{im}}(C)\subseteq\mathscr{T}(\mathscr{H})$ is defined in Theorem 2.7], if and only if $R\in\mathscr{R}(C)$ and R is bounded (in Ref. 7 they are considered bounded symplectic transformations T on the whole of $\mathscr{H}$; cf. Lemma 2.1).

From Proposition 3.3 and Theorem 2.3 one immediately obtains the following consequence of Theorem 3.7.

Corollary 3.8: Let S be Hilbert–Schmidt. Then the group $\{T_t|t\in\mathbb{R}\}$ is continuously implementable in some symplectically related representations Π_T, $T=\exp\{JR\}$, inequivalent, resp. disjoint to the Fock representation Π_F, if and only if C has an essential infrared part; these Π_T are characterized in terms of the operator R as follows:

(a) *$R\in\mathscr{R}(C)$ is bounded but not Hilbert–Schmidt.*
(b) *For unbounded R, however, the condition $R\in\mathscr{R}(C)$ is necessary but not sufficient for the continuous implementability of $\{T_t|t\in\mathbb{R}\}$ in Π_T. For example, a sufficient condition is the Hilbert–Schmidt property of $\hat{S}P_C([-a,a])=(\cosh(2R)S+\sinh(2R)C)P_C([-a,a])$ for some $a>0$.*

Finally, let us give examples for the cases $R\in\mathscr{R}(C)$ and $R\notin\mathscr{R}(C)$.

Example 3.9: Choose a subset $N\subseteq\mathbb{N}$ with $\Sigma_{k\in N}c_k^2<\infty$, and define R by $r_k=0$ for $k\notin N$ and $r_k\in\mathbb{R}$ arbitrary for $k\in N$. Then one observes the following:

(a) It holds *$RP_C(\mathbb{R}\backslash[-a,a])=0$ for each $a>\sup_{k\in N}|c_k|$ and $\tanh(R)C$ is Hilbert–Schmidt, and thus $R\in\mathscr{R}(C)$.*
(b) *Only if the cardinality of N is infinite, it is possible to choose R not to be Hilbert–Schmidt*

[that is, $T=\exp\{JR\}\notin\mathscr{T}_F(E)$], which makes the representation Π_T disjoint to the Fock representation. But an infinite N is selectable, if and only if C has an essential infrared part.

(c) *For C with an essential infrared part select an infinite N, and let, e.g., $|r_k|\geqslant d$ for all $k\in N$ with some $d>0$. Then R is not Hilbert–Schmidt, but $R\in\mathscr{R}(C)$. Here R may be chosen bounded as well as unbounded.*

Example 3.10: Let S be Hilbert–Schmidt and suppose C to have an essential infrared part. Hence for (each) $q>1$ there exists an infinite subset $N\subseteq\mathbb{N}$ with $\Sigma_{k\in N}q^k c_k^2<\infty$ and $\Sigma_{k\in N}q^k s_k^2<\infty$. Define an unbounded R by $r_k=0$ for $k\notin N$ and $\cosh(2r_k)\leqslant q^k$ for $k\in N$. Then $\cosh(2R)S$ and $\sinh(2R)C$ both are Hilbert–Schmidt, and thus $\hat{\mathrm{S}}$ is so, too. Consequently, $\{T_t|t\in\mathbb{R}\}$ is continuously implementable in Π_T by Corollary 3.8 (b).

Example 3.11: Suppose C not to be Hilbert–Schmidt. Define a (bounded or unbounded) R with $|r_k|\geqslant d$ for all $k\in\mathbb{N}$ and some $d>0$. Then $R\notin\mathscr{R}(C)$ (which forces Π_T to be disjoint to Π_F), and thus $\{T_t|t\in\mathbb{R}\}$ is not (continuously) implementable in Π_T.

IV. QUADRATIC HAMILTONIANS IN REPRESENTATION SPACES

Let us first consider a representation Π of $\mathscr{W}(E)$ with the representation Hilbert space $\mathscr{K}_\Pi$. With the annihilation and creation operators of the representation Π from Eq. (2.6), $a_\Pi(f)$, resp. $a_\Pi^*(f)$, $f\in E$, one may introduce the formal Hamiltonian,

$$H_{C,S}^{\Pi}:=\sum_{k=1}^{\infty}\left[\frac{s_k}{2}\,(a_\Pi^*(e_k)^2+a_\Pi(e_k)^2)+c_k a_\Pi^*(e_k)a_\Pi(e_k)\right], \tag{4.1}$$

on the representation Hilbert space $\mathscr{K}_\Pi$. It is not clear whether the operator series in (4.1) converges at all, and thus, whether $H_{C,S}^{\Pi}$ is indeed a well-defined self-adjoint operator on $\mathscr{K}_\Pi$. However, from purely algebraic calculations it follows that $\exp\{itH_{C,S}^{\Pi}\}$ formally implements our symplectic group $\{T_t=\exp\{it(C-JS)\}|t\in\mathbb{R}\}\subset\mathscr{T}(E)$ in the representation Π. If the Hamiltonian $H_{C,S}^{\Pi}$ has no mathematical meaning as a self-adjoint operator on $\mathscr{K}_\Pi$, then nevertheless a well-defined self-adjoint operator H_Π may exist on $\mathscr{K}_\Pi$, which (continuously) implements $\{T_t|t\in\mathbb{R}\}$ in Π according to Definition 2.2 (b). Then H_Π is a *renormalized* Hamiltonian of the system, which differs from the formal (nonexisting) $H_{C,S}^{\Pi}$ by an additive diverging renormalization term.

Here we investigate Hamiltonians of the type (4.1) in the Fock representation $\Pi\equiv\Pi_F$ and in the symplectically related representation $\Pi\equiv\Pi_T=\Pi_F\circ\alpha_T$, where $T=\exp\{JR\}$ with R from (3.1). Thus we have here analoguously to (4.1) the two formal Hamiltonians,

$$H_{C,S}^{F}:=\sum_{k=1}^{\infty}\left[\frac{s_k}{2}\,(a_F^*(e_k)^2+a_F(e_k)^2)+c_k a_F^*(e_k)a_F(e_k)\right], \tag{4.2}$$

$$H_{C,S}^{T}:=\sum_{k=1}^{\infty}\left[\frac{s_k}{2}\,(a_T^*(e_k)^2+a_T(e_k)^2)+c_k a_T^*(e_k)a_T(e_k)\right], \tag{4.3}$$

on Fock space. Especially $H_{C,S}^{T}$ coincides with $H_{C,S}^{F}$ for $T=\mathbb{1}$, resp., $R=0$. We also discuss the formal Hamiltonian $H_{\hat{C},\hat{S}}^{F}$, which formally implements the shifted group $\{\hat{T}_t=\exp\{it(\hat{C}-J\hat{S})\}|t\in\mathbb{R}\}$ in the Fock representation Π_F. It is obtained from (4.2) by replacing C and S by $\hat{\mathrm{C}}$ and $\hat{\mathrm{S}}$, that is, by replacing the associated eigenvalues s_k and c_k by the eigenvalues $\hat{s}_k$ and $\hat{c}_k$ of $\hat{\mathrm{S}}$ and $\hat{\mathrm{C}}$ from (3.6), respectively.

The formal Hamiltonian $H_{C,S}^{T}$ should be approximated (in some sense) by the sequence

$$H_{CP_n,SP_n}^{T}=\sum_{k=1}^{n}\left[\frac{s_k}{2}\,(a_T^*(e_k)^2+a_T(e_k)^2)+c_k a_T^*(e_k)a_T(e_k)\right],\quad n\in\mathbb{N}, \tag{4.4}$$

where the orthogonal projections $P_n:=\Sigma_{k=1}^{n}|e_k\rangle\langle e_k|$, $n\in\mathbb{N}$, commute with all the operators C, $\hat{\mathrm{C}}$, S, $\hat{\mathrm{S}}$, R, and J. These *local* Hamiltonians H_{CP_n,SP_n}^{T}, $n\in\mathbb{N}$, are well-defined self-adjoint operators on Fock space, which follows from Eq. (4.6) and Theorem 4.3 below.

The Bogoliubov transformed annihilation operators are given on Fock space by

$$a_T(e_k)=\cosh(r_k)a_F(e_k)+\sinh(r_k)a_F^*(e_k),\quad k\in\mathbb{N}, \tag{4.5}$$

which follows from (2.10) and the fact that by (3.4) $T_l=\cosh(R)$ and $T_a=J\sinh(R)$ for our symplectic $T=\exp\{JR\}\in\mathscr{T}(E)$.

Observation 4.1: For every $k\in\mathbb{N}$ it holds on Fock space:

$$\left[\frac{s_k}{2}\,(a_T^*(e_k)^2+a_T(e_k)^2)+c_k a_T^*(e_k)a_T(e_k)\right]=\left[\frac{\hat{s}_k}{2}\,(a_F^*(e_k)^2+a_F(e_k)^2)+\hat{c}_k a_F^*(e_k)a_F(e_k)\right]$$
$$+\tfrac{1}{2}\,(\hat{c}_k-c_k)\mathbb{1}.$$

Proof: Insert (4.5) on the left side. Then the CCR $[a_F(e_k),a_F^*(e_l)]=\delta_{k,l}\mathbb{1}$, Eq. (3.6) and the formulas (3.7) yield the result. ■

The above lemma implies immediately the following connection for the local Hamiltonians:

$$H^T_{CP_n,SP_n}=H^F_{CP_n,\hat{S}P_n}+\kappa_n\mathbb{1},\quad \forall n\in\mathbb{N}, \tag{4.6}$$

where κ_n is given by

$$\kappa_n:=\frac{1}{2}\,\mathrm{tr}[(\hat{C}-C)P_n]=\frac{1}{2}\sum_{k=1}^{n}\,(\hat{c}_k-c_k), \tag{4.7}$$

with the relation [following from (3.5)]

$$\hat{C}-C=2\,\sinh(R)^2C+\sinh(2R)S. \tag{4.8}$$

Observation 4.2: *The convergence of the series $\kappa:=\lim_{n\to\infty}\kappa_n=\frac{1}{2}\Sigma_{k=1}^{\infty}(\hat{c}_k-c_k)$ does not mean that $\hat{C}-C$ is of trace class on $\mathscr{H}$. But $\hat{C}-C$ being of trace class is equivalent to the* absolute *convergence of the series, in which case $\kappa=\frac{1}{2}\,\mathrm{tr}[\hat{C}-C]$.*

A. Hamiltonians in the Fock representation

Let us first investigate the limit of the local Hamiltonians $H^F_{CP_{n,SPn}}$, $n\in\mathbb{N}$.

Theorem 4.3: *The following assertions are valid.*

(a) *Let S be Hilbert–Schmidt. Then the operator series in (4.2), respectively, the limit*

$$\lim_{n\to\infty}\,H^F_{CP_n,SP_n}=H^F_{C,S},$$

converge in the strong resolvent sense to the self-adjoint $H^F_{C,S}$, and they also converge pointwise [with respect to the norm of $F_+(\mathscr{H})$] on the dense set L_C of the finite particle number vectors over the domain of C of one-particle wave functions. L_C is a core for $H^F_{C,S}$. Especially the vacuum vector Ω_F is contained in the domain of $H^F_{C,S}$, and it holds that $\|H^F_{C,S}\Omega_F\|=1/\sqrt{2}\|S\|_{HS}$. Moreover, the unitary group $\exp\{itH^F_{C,S}\}$, $t\in\mathbb{R}$, *implements the symplectic group $\{T_t|t\in\mathbb{R}\}$ in the Fock representation Π_F.*

(b) *Suppose the operator series (4.2) to converge in the strong resolvent sense to a selfadjoint operator $H^F_{C,S}$ on Fock space. Then we have the following equivalences:*

(i) The vacuum vector Ω_F is contained in the domain of $H^F_{C,S}$;
(ii) S is Hilbert–Schmidt on $\mathscr{H}$.

Proof: (a) By Ref. 6, $\lim_{n\to\infty}H^F_{CP_nSP_n}\psi=H^F_{C,S}\psi$ converges in Fock space for the vectors ψ from the core L_C. Thus, by Ref. 24, Theorem 9.16 the operator series in (4.2), respectively, the limit $\lim_{n\to\infty}H^F_{CP_n,SP_n}=H^F_{C,S}$, converge in the strong resolvent sense.

(b) By Ref. 6, Theorem 3.7, $\exp\{itH^F_{CP_n,SP_n}\}$ implements $T_{t,n}:=\exp\{it(C-JS)P_n\}$ in Π_F. But $\lim_{n\to\infty}H^F_{CP_n,SP_n}=H^F_{C,S}$ in the resolvent sense is equivalent to $\lim_{n\to\infty}\exp\{itH^F_{CP_n,SP_n}\}=\exp\{itH^F_{C,S}\}$ in the strong operator topology. Consequently, $\exp\{itH^F_{C,S}\}$ implements T_t

$=\lim_{n\to\infty} T_{t,n}$ in Π_F, which yields $T_t \in \mathscr{T}_F(E)$. Now Ref. 6, Theorem 3.8 implies that the condition $\Omega_F \in \mathscr{D}(H^F_{C,S})$ is equivalent to the Hilbert–Schmidt property of S. ■

Observation 4.4: If S is Hilbert–Schmidt, then one has $H^F_{C,S}=H^F_{C,0}+H^F_{O,S}$, *where* $H^F_{C,O} \equiv d\Gamma(C)$ *is the second quantization of C, and* $\exp\{itH^F_{C,S}\}$ *is obtained by a Dyson expansion with the ''free'' evolution* $\exp\{itH^F_{C,0}\}$ *and the perturbation* $H^F_{0,S}$.[6] *However, if S is not a Hilbert–Schmidt operator, then the expression* $H^F_{0,S}$ *has no meaning as an operator on Fock space, which is shown in Ref. 5.*

Subsequently, in Corollary 4.6(b) we derive that for $C\neq 0$ *also some operators S are allowed, which are not Hilbert–Schmidt (and possibly unbounded), so that the self-adjoint limiting operator* $H^F_{C,S}$ *of the form (4.2) exists in the strong resolvent sense. But then the vacuum vector* Ω_F *is not contained in the domain of definition of* $H^F_{C,S}$, *and the additive decomposition* $H^F_{C,S}=H^F_{C,0}+H^F_{0,S}$ *is meaningless.*

From now on in the present section *let us suppose S and R to be Hilbert–Schmidt.* Then $T=\exp\{JR\}\in\mathscr{T}_F(E)$ by Proposition 3.3 and by Theorem 4.3 we have for all $t\in\mathbb{R}$ that $T, T_t, T_t \in \mathscr{T}_F(E)$ since $\mathscr{T}_F(E)$ is a group, and T_t is (continuously) implemented by $\exp\{itH^F_{C,S}\}$ in the Fock representation Π_F. By Theorem 2.3 the condition $T\in\mathscr{T}_F(E)$ is equivalent to the unitary equivalence of the representations Π_F and Π_T, where the associated implementing unitary U_T is defined in Eq. (2.12).

Proposition 4.5: Let R be Hilbert–Schmidt on $\mathscr{H}$. *Then the symplectic group* $\{\exp\{tJR\}|t\in\mathbb{R}\}\subset\mathscr{T}(E)$ *is continuously implemented by* $\exp\{itQ_R\}$ *in the Fock representation* Π_F *with the self-adjoint operator,*

$$Q_R := \sum_{k=1}^{\infty} \frac{r_k}{2}\,(a^*_F(e^{i\pi/r}e_k)^2 + a_F(e^{i\pi/4}e_k)^2) = i\sum_{k=1}^{\infty}\frac{r_k}{2}\,(a^*_F(e_k)^2 - a_F(e_k)^2),$$

on Fock space, where the operator series converges in the strong resolvent sense, and it also converges pointwise on the finite particle number vectors L. L *is a core for* Q_R.

Especially the implementing unitary U_T *for our symplectic* $T=\exp\{JR\}\in\mathscr{T}_F(E)$ *is given (up to a phase) by* $U_T=\exp\{iQ_R\}$.

Proof: One has $\exp\{tJR\}=\exp\{-it\hat{J}R\}$, where the antilinear involution $\hat{J}:=iJ$ on $\mathscr{H}$ satisfies $\widetilde{J}(e^{i\pi/4}e_k)=e^{i\pi/4}e_k, \forall k\in\mathbb{N}$. The involution $\widetilde{J}$ commutes with our $R=\Sigma_{k=1}^{\infty} r_k|e_k\rangle\langle e_k| = \Sigma_{k=1}^{\infty} r_k|e^{i\pi/4}e_k\rangle\langle e^{i\pi/4}e_k|$. Thus, the result follows from Ref. 6, or in a way similar to the proof of Theorem 4.3(a). ■

Here the Bogoliubov transforms from Eq. (4.5) are obtained with the unitary transformation U_T, that is, $a_T(e_k)=U_T a_F(e_k) U_T^* = \cosh(r_k)a_F(e_k)+\sinh(r_k)a^*_F(e_k)$, which follow from (2.13). Thus (4.6) takes the form

$$H^T_{CP_n,SP_n} = U_T H^F_{CP_n,SP_n} U_T^* = H^F_{CP_n,\hat{S}P_n} + \kappa_n \mathbb{1}, \quad \forall n\in\mathbb{N}, \tag{4.9}$$

with κ_n from (4.7). Performing the limit $n\to\infty$ in (4.9), we gain with standard results on the strong resolvent convergence (cf., e.g., Ref. 3, Sec. VIII.), with Lemma 2.8, and with Theorem 4.3 the following result.

Corollary 4.6: Let S and R both be Hilbert–Schmidt. Then it holds on Fock space.

(a) *The self-adjoint limiting operator,*

$$H^T_{C,S} = U_T H^F_{C,S} U_T^* = \lim_{n\to\infty} H^T_{CP_n,SP_n}$$

exists, where the limit and the series converge in the strong resolvent sense, and they also converge pointwise on $U_T(L_C)$ *with respect to the norm of* $F_+(\mathscr{H})$. $U_T(L_C)$ *is a core for* $H^T_{C,S}$, *and* $\exp\{itH^T_{C,S}\}$ *implements* $\hat{T}_t$ *in* Π_F *and* T_t *in* Π_T.

(b) *If* $\lim_{n\to\infty}\kappa_n=\kappa\in\mathbb{R}$ *exists, then the self-adjoint operator*

$$H^F_{\hat{C},\hat{S}} = \lim_{n\to\infty} H^F_{\hat{C}P_n,\hat{S}P_n}$$

exists, where the limit and the series converge in the strong resolvent sense, and they also

converge pointwise on $U_T(L_C)$ with respect to the norm of $F_+(\mathscr{H})$. Moreover, $U_T(L_C)$ is a core for $H^F_{\hat{C},\hat{S}}$, and we have $H^F_{\hat{C},\hat{S}}=H^T_{C,S}-\kappa\mathbf{1}$. Thus, both $\exp\{itH^F_{\hat{C},\hat{S}}\}$ *and* $\exp\{itH^T_{C,S}\}$ *implement $\hat{T}_t$ in Π_F and T_t in Π_T.*

(c) *If $\lim_{n\to\infty}\kappa_n$ does not exist, then $H^F_{\hat{C},\hat{S}}$ has no meaning as a self-adjoint operator on Fock space, and the limit and the operator series in (b) do not converge in the weak (and thus strong) resolvent sense.*

Thus, for S and R being Hilbert–Schmidt, the implementing self-adjoint operator $H^T_{C,S}$ for T_t in Π_T has the form (4.3). However, for the implementation of the shifted group $\hat{T}_t$ in the Fock representation Π_F, the associated formal Hamiltonian $H^F_{\hat{C},\hat{S}}$ of the desired form (4.2) may not exist. Hence, the self-adjoint $H^T_{C,S}$ is a *renormalization* of $H^F_{\hat{C},\hat{S}}$, where the difference is given by the additive (possibly diverging) renormalization term $\kappa=\lim_{n\to\infty}\kappa_n$ in the sense of

$$H^T_{C,S}=U_T H^F_{C,S}U_T^*=\lim_{n\to\infty}(H^F_{\hat{C}P_n,\hat{S}P_n}+\kappa_n\mathbf{1}).$$

Example 4.7: Let S and R both be Hilbert–Schmidt.

(a) We use (4.8) to construct situations in which $\kappa=\frac{1}{2}\,\mathrm{tr}[\hat{C}-C]$ exists.

• Assume $\hat{S}$ to be Hilbert-Schmidt [which, e.g., is a consequence of (3.5) for bounded C]. Then $\hat{C}-C$ is of trace class on $\mathscr{H}$.

• Define R and C by $\sinh(r_k):=k^{-3}$ *and $c_k:=k^4$ for all $k\in\mathbb{N}$. Then R is Hilbert–Schmidt and $\hat{C}-C$ is of trace class. If in addition $S\geqslant 0$, then $\hat{S}\geqslant\mathbb{1}$ is unbounded (since $\hat{s}_k\geqslant k$, $\forall k\in\mathbb{N}$). Nevertheless, the quadratic self-adjoint Hamiltonian $H^F_{\hat{C},\hat{S}}$ exists and differs from $H^T_{C,S}$ by the finite $\kappa\mathbb{1}$.*

(b) Let us give an example where κ does not exist: Let $r_k>0$ and $S\geqslant 0$, and define C by $c_k\geqslant\sinh(r_k)^{-2}$, then $\hat{C}-C\geqslant 2\mathbb{1}$ by (4.8), and thus $\kappa=\infty$.

Proof: We prove the first part of (a): The Hilbert–Schmidt operators constitute an ideal in the bounded operators on $\mathscr{H}$. Thus, from the boundedness of $\cosh(R)^{-1}$ and $\cosh(R)$ (since R is bounded), from (3.5) and the formula $2\cosh(x)\sinh(x)=\sinh(2x)$, we obtain

$$\sinh(R)C=\tfrac{1}{2}\,(\hat{S}-\cosh(2R)S)\cosh(R)^{-1},$$

to be Hilbert–Schmidt (also for unbounded C). Now observe that the product of two Hilbert–Schmidt operators gives a trace class operator, and the result follows from (4.8) and the fact that $\sinh(2R)$ and $\sinh(R)$ are Hilbert–Schmidt by Proposition 3.3. The other parts follow by direct computation. ■

B. Hamiltonians in symplectically related (non-Fock) representations

Here we allow R to not be Hilbert–Schmidt, which by Proposition 3.3 is equivalent to $T=\exp\{JR\}\notin\mathscr{T}_F(E)$. Then the symplectically related representation Π_T is disjoint to the Fock representation Π_F, and the above implementing unitary U_T does not exist.

Performing the limit $n\to\infty$ in (4.6), we obtain with Theorem 4.3 (a) applied to $\hat{C}$ and $\hat{S}$ and with the analogous arguments as in the previous section the following result.

Proposition: 4.8: Let $\hat{S}$ be Hilbert–Schmidt. Then the following holds.

(a) *The self-adjoint limiting operator,*

$$H^F_{\hat{C},\hat{S}}=\lim_{n\to\infty}H^F_{\hat{C}P_n,\hat{S}P_n},$$

exists in the strong resolvent sense. The unitary group $\exp\{itH^F_{\hat{C},\hat{S}}\}$ *continuously implements the symplectic one-parameter group $\{\hat{T}_t|t\in\mathbb{R}\}$ in Π_F and the group $\{T_t|t\in\mathbb{R}\}$ in the representation Π_T (which is disjoint to Π_F for R not Hilbert–Schmidt).*

(b) *If* $\lim_{n\to\infty}\kappa_n=\kappa\in\mathbb{R}$ *exists, then performing the limit* $n\to\infty$ *in (4.6) gives the Hamiltonian* $H^T_{C,S}=H^F_{\hat{C},\hat{S}}+\kappa\mathbb{1}$ *from (4.3) as a well-defined self-adjoint operator on Fock space, and its series converges in the strong resolvent sence. Furthermore, both* $\exp\{itH^T_{C,S}\}$ *and* $\exp\{itH^F_{\hat{C},\hat{S}}\}$ *continuously implement* $\hat{T}_t$ *in* Π_F *and* T_t *in* Π_T.

(c) *If* $\lim_{n\to\infty}\kappa_n$ *does not exist, then* $H^T_{C,S}$ *has no meaning as a self-adjoint operator on Fock space.*

The self-adjoint operator $H^F_{\hat{C},\hat{S}}$ *is a renormalization of the formal Hamiltonian* $H^T_{C,S}$ *differring therefrom by the (possibly diverging) renormalization term* κ *in the sense of*

$$H^F_{\hat{C},\hat{S}}=\lim_{n\to\infty}(H^T_{CP_n,SP_n}-\kappa_n\mathbb{1}). \tag{4.10}$$

Example 4.9 (continuation of Example 3.9): Let S *be of trace class (thus* S *is Hilbert–Schmidt, and* T_t *is implementable in* Π_F *by* $\exp\{itH^F_{C,S}\}$*), and choose a subset* $N\subseteq\mathbb{N}$ *with* $\Sigma_{k\in N}|c_k|<\infty$*, and define* R *by* $r_k=0$ *for* $k\notin N$ *and* $r_k\in\mathbb{R}$ *arbitrary for* $k\in N$*, but with* $\sup_k|r_k|<\infty$*, such that* R *is bounded. Then the following holds.*

(a) *The equations (4.8) and (3.5) imply* $\hat{C}-C$ *and* $\hat{S}$ *to be of trace class, which ensures* $\hat{S}$ *to be Hilbert–Schmidt and the existence of* $\kappa=\frac{1}{2}\,\mathrm{tr}[\hat{C}-C]$.

(b) *Only if the cardinality of* N *is infinite, it is possible to choose* R *not to be Hilbert–Schmidt, which makes the representation* Π_T *disjoint to the Fock representation. But an infinite* N *is selectable, if and only if* C *has an essential infrared part.*

Collecting the results of Observation 4.1, Corollary 4.6 (a), and Proposition 4.8 (a), we get a mixed implementing Hamiltonian in the case of Theorem 3.7 (I). From its proof [especially from part (I)(ii)⇒(I)(iii)] we obtain the following form of a more involved renormalization.

Theorem 4.10 (continuation of Theorem 3.7): *Let* S *be Hilbert–Schmidt, and suppose* $R\in\mathscr{R}(C)$ *such that* $\hat{S}P_C([-a,a])$ *is Hilbert–Schmidt for an* $a>0$ [*this automatically follows for each bounded* $R\in\mathscr{R}(C)$]. *Then it holds for each* $b>0$: $\hat{S}P_C([-b,b])$ *is Hilbert–Schmidt and the well-defined self-adjoint operator,*

$$H_b:=H^F_{\hat{C}P_C([-b,b]),\hat{S}P_C([-b,b])}+H^T_{CP_C(\mathbb{R}\setminus[-b,b]),SP_C(\mathbb{R}\setminus[-b,b])},$$

implements the symplectic one-parameter group $\{\hat{T}_t|t\in\mathbb{R}\}$ *in* Π_F *and the group* $\{T_t|t\in\mathbb{R}\}$ *in the non-Fock representation* Π_T. *Thus* H_b *is a renormalization of the (formal) expressions* $H^F_{\hat{C},\hat{S}}$*, resp.* $H^T_{C,S}$*, in the sense of*

$$\begin{aligned}H_b&=\lim_{n\to\infty}(H^F_{\hat{C}P_n,\hat{S}P_n}+\tfrac{1}{2}\,\mathrm{tr}[(\hat{C}-C)P_C(\mathbb{R}\setminus[-b,b])P_n]\mathbb{1})\\&=\lim_{n\to\infty}(H^T_{CP_n,SP_n}-\tfrac{1}{2}\,\mathrm{tr}[(\hat{C}-C)P_C([-b,b])P_n]\mathbb{1}).\end{aligned}$$

Furthermore, if $(\hat{C}-C)P_C(\mathbb{R}\setminus[-b,b])$ *is of trace class, then* $H^F_{\hat{C},\hat{S}}$ *exits as a well-defined self-adjoint operator on Fock space, and if* $(\hat{C}-C)P_C([-b,b])$ *is of trace class, then* $H^T_{C,S}$ *exists as a well-defined self-adjoint operator on Fock space [in which case the series from (4.2), resp. (4.3), converge in the strong resolvent sense].*

C. Diagonalization

In the literature diagonalizations mostly are performed in the Fock representation, cf., e.g., Ref. 25 and 20. Thus, in terms of our notion a Fock representation diagonalization of the *existing* quadratic operator $H^F_{C,S}$ with (possibly unbounded) $S\neq 0$ means to find a Hilbert–Schmidt operator R, such that $\hat{S}=0$ and $\lim_{n\to\infty}\kappa_n=\kappa\in\mathbb{R}$ exists. Then all terms in Eq. (4.9) exist in the limit $n\to\infty$, where U_T is the implementing unitary for $T=\exp\{JR\}\in\mathscr{T}_F(E)$ by the Propositions 3.3 and 4.5, and the self-adjoint operator,

$$H^F_{C,0}=H^T_{C,S}-\kappa\mathbb{1}=U_T H^F_{C,S}U_T^*-\kappa\mathbb{1}, \tag{4.11}$$

is referred to as the (Fock) *diagonalization* of $H^F_{C,S}$.

However, here we do not restrict ourselves to the Fock representation and also allow arbitrary unbounded operators S, R, and C of the type (3.1) and nonexisting renormalization terms κ. The following definition characterizes *diagonalization* only on the one-particle level and does not use the notion of a continuous implementation of the symplectic one-parameter group $\{T_t|t\in\mathbb{R}\}$.

Definition 4.11 (diagonalization): The symplectic group $T_t=\exp\{it(C-JS)\}\in\mathscr{T}(E)$, $t\in\mathbb{R}$, is called diagonalizable by $T=\exp\{JR\}\in\mathscr{T}(E)$, if $\hat{S}=0$. Then the shifted (unitary) group $TT_tT^{-1}=\hat{T}_t=\exp\{it\hat{C}\}$ is denoted the diagonalization of T_t [cf. Eq. (3.3) and Lemma 3.1].

Theorem 4.12: *We have the following equivalences:*

(i) $\{T_t=\exp\{it(C-JS)\}|t\in\mathbb{R}\}$ is diagonalizable;

(ii) $|s_k|<|c_k|$ for every $k\in\mathbb{N}$ with $s_k\neq 0$.

If the conditions (i) and (ii) are satisfied, then the eigenvalues r_k, $k\in\mathbb{N}$, for the diagonalizing $T=\exp\{JR\}\in\mathscr{T}(E)$ are given by $c_k\tanh(2r_k)=-s_k$; thus r_k is unique for $c_k\neq 0$ and arbitrary for $c_k=0$.

Furthermore, if (i) and (ii) are valid, then the following holds.

(a) *$\hat{C}=\sqrt{C^2-S^2}\geqslant 0$ [observe that $C^2\geqslant S^2$ by (ii)], which yields $H^F_{\hat{C},0}\geqslant 0$.*

(b) *R is bounded, if and only if $\sup\{|s_k|/|c_k||k,c_k\neq 0\}<1$ (provided the eigenvalues r_k are chosen for those $k\in\mathbb{N}$ with $c_k=0$ such that $\sup\{|r_k||k,c_k=0\}<\infty$).*

(c) *R is Hilbert–Schmidt (which is equivalent to $T=\exp\{JR\}\in\mathscr{T}_F(E)$, resp., to the existence of the above implementing unitary U_T,), if and only if $\Sigma_{k,c_k\neq 0}s_k^2/c_k^2<\infty$ (provided the eigenvalues r_k are chosen for those $k\in\mathbb{N}$ with $c_k=0$, such that $\Sigma_{k,c_k=0}r_k^2<\infty$).*

(d) *$\hat{C}-C$ is of trace class, if and only if $\Sigma_{k,c_k\neq 0}s_k^2/|c_k|<\infty$, in which case κ exists (cf. Observation 4.2) with $\kappa=\frac{1}{2}\,\mathrm{tr}[\hat{C}-C]=-1/2\Sigma_{k,c_k\neq 0}c_k(1-\sqrt{1-s_k^2/c_k^2})$.*

Proof: By (3.5), resp. (3.6), diagonalization condition $\hat{S}=0$ is equivalent to $s_k=-\tanh(2r_k)c_k$, $\forall k\in\mathbb{N}$. Since $|\tanh(x)|<1$, $\forall x\in\mathbb{R}$, it is only possible to realize $s_k=-\tanh(2r_k)c_k$ with an $r_k\in\mathbb{R}$ if (ii) is fulfilled.

If $c_k=0=s_k$, then (3.6) gives $\hat{c}_k=0$. But for $c_k\neq 0$ we obtain with the formulas $\tanh(2r_k)=-s_k/c_k$, $\cosh(x)^2-\sinh(x)^2=1$ and $1/\cosh(x)=\sqrt{1-\tanh(x)^2}$, $\forall x\in\mathbb{R}$,

$$\begin{aligned}\hat{c}_k=\cosh(2r_k)c_k+\sinh(2r_k)s_k&=c_k\left(\cosh(2r_k)+\sinh(2r_k)\,\frac{s_k}{c_k}\right)\\&=c_k\left(\cosh(2r_k)-\frac{\sinh(2r_k)^2}{\cosh(2r_k)}\right)=\frac{c_k}{\cosh(2r_k)}\\&=c_k\,\sqrt{1-\tanh(2r_k)^2}=\sqrt{c_k^2-s_k^2},\end{aligned}$$

which implies the operator relation $\hat{C}=\sqrt{C^2-S^2}\geqslant 0$. Thus $\hat{c}_k-c_k=-c_k(1-\sqrt{1-s_k^2/c_k^2})$. Now observe that $0\leqslant(1-\sqrt{1-s_k^2/c_k^2})\leqslant 1$ and that $1\leqslant(1+\sqrt{1-s_k^2/c_k^2})\leqslant 2$, which implies

$$\left(1-\sqrt{1-\frac{s_k^2}{c_k^2}}\right)\leqslant\left(1-\sqrt{1-\frac{s_k^2}{c_k^2}}\right)\left(1+\sqrt{1-\frac{s_k^2}{c_k^2}}\right)=\frac{s_k^2}{c_k^2}\leqslant 2\left(1-\sqrt{1-\frac{s_k^2}{c_k^2}}\right).$$

Thus, we obtain the estimation

$$\mathrm{tr}[|\hat{C}-C|]=\sum_{k=1}^{\infty}|\hat{c}_k-c_k|=\sum_{k,c_k\neq 0}|c_k|\left(1-\frac{s_k^2}{c_k^2}\right)\leqslant\sum_{k,c_k\neq 0}\frac{s_k^2}{|c_k|}\leqslant 2\,\mathrm{tr}[\hat{C}-C|],$$

which gives (d). Part (c) follows from Proposition 3.3. ∎

Since $\hat{S}=0$ is Hilbert–Schmidt we may apply Proposition 4.8, and we obtain from Eq. (4.10) the following generalization of the Fock diagonalization (4.11), which is appropriate for many

models (cf., e.g., Sec. V), and that gives another generalization of Warchall's result in Theorem 2.7. Observe that, in contrast to Theorem 2.7, here E is a proper dense subspace of $\mathscr{H}$, and thus by Lemma 2.1 there occur also unbounded symplectic transformations $V \in \mathscr{T}(E)$.

Theorem 4.13 (particle representation of squeezed Bosons): *Let be given the dynamical symplectic group* $\{T_t = \exp\{it(C-JS)\} | t \in \mathbb{R}\} \subset \mathscr{T}(E)$ *for squeezed Bosons with possibly unbounded* S *and* C.

(i) *The diagonalization condition* $|s_k| < |c_k|$ *for all* $k \in \mathbb{N}$ *with* $s_k \neq 0$ *is necessary and sufficient that there exists a Fock symplectically related representation* Π_T *of the CCR-algebra* $\mathscr{W}(E)$, *in which the one-particle dynamics* T_t *is continuously implemented in terms of a (renormalized) diagonal self-adjoint operator,*

$$H^F_{\hat{C},0} = \lim_{n\to\infty} (H^T_{CP_n, SP_n} - \kappa_n \mathbf{1}) = \sum_{k=1}^{\infty} \hat{c}_k a^*_F(e_k) a_F(e_k). \tag{4.12}$$

$T = \exp\{JR\} \in \mathscr{T}(E)$*—and thus* Π_T*—is unique up to the modes* k *with vanishing* c_k, *and* R *is not necessarily bounded. This holds also in the case where the implementing* U_T *does not exist [i.e.,* $T \notin \mathscr{T}_F(E)$*] or the limit* $\lim_{n\to\infty} \kappa_n = \kappa$ *diverges (which then implies that* $H^T_{C,S}$ *does not exist). The positivity of the second quantized* $H^F_{\hat{C},0} \equiv d\Gamma(\hat{C})$ *follows automatically [cf. part (a) of Theorem 4.12], and thus allows for a particle interpretation (cf. Ref. 8).*

(ii) *Suppose the diagonalization condition* $|s_k| < |c_k|$ *for every* $k \in \mathbb{N}$ *with* $s_k \neq 0$ *to be valid. Let us denote by* $\mathscr{C}_{\mathrm{im}}(C,S)$ *the class of those* $V \in \mathscr{T}(E)$ *[not necessarily of the form (3.4)], for which* $\{T_t | t \in \mathbb{R}\}$ *is continuously implementable in* Π_V. *Then it holds for the diagonalizing* $T = \exp\{JR\}$:

(a) *If* $\hat{C} = \sqrt{C^2 - S^2}$ *has an inessential infrared part, then* $\mathscr{C}_{\mathrm{im}}(C,S) = \mathscr{T}_T(E)$. *That is, each representation* Π_V, $V \in \mathscr{T}(E)$, *in which* $\{T_t | t \in \mathbb{R}\}$ *is continuously implementable, is unitarily equivalent to the diagonal representation* Π_T.

(b) *If* $\hat{C} = \sqrt{C^2 - S^2}$ *has an essential infrared part, then the inclusion* $\mathscr{T}_T(E) \subset \mathscr{C}_{\mathrm{im}}(C,S)$ *is proper. However, one has, in general,* $\mathscr{C}_{\mathrm{im}}(C,S) \neq \mathscr{T}(E)$. *Observe that an essential infrared part of* C *implies always an essential infrared part for* $\hat{C} = \sqrt{C^2 - S^2}$. *However, if* C *has an inessential infrared part, then both an essential, resp. an inessential, infrared part may occur for* $\hat{C}$.

Proof: Since $\mathscr{T}(E)$ is a group, there exists for each $V \in \mathscr{T}(E)$ a unique $X \in \mathscr{T}(E)$ such that $V = XT$, resp., $X = VT^{-1}$. Let us first reduce the implementability of T_t in Π_V to the implementability of $\hat{T}_t$ in Π_X. If $V \in \mathscr{C}_{\mathrm{im}}(C,S)$ with the implementing strongly continuous unitary group $\{U_t | t \in \mathbb{R}\}$, then $\Pi_X(W(TT_t f)) = \Pi_V(W(T_t f)) = U_t \Pi_V(W(f)) U_t^* = U_t \Pi_X(W(Tf)) U_t^*$ for all $f \in E$. Since T acts bijectively on E, we may put $f = T^{-1}g$, and obtain $U_t \Pi_X(W(g)) U_t^* = \Pi_X(W(TT_t T^{-1} g))$ for all $g \in E$. Hence, T_t is continuously implementable in Π_V, if and only if the diagonal $TT_t T^{-1} = \hat{T}_t = \exp\{it\hat{C}\}$ is continuously implementable in Π_X. Now we reduce the unitary equivalence of Π_V and Π_T to the unitary equivalence of Π_X and Π_V. Let $X \in \mathscr{T}_F(E)$, that is, $\Pi_X(A) = U_X \Pi_F(A) U_X^*$ for all $A \in \mathscr{W}(E)$ with some unitary U_X on Fock space; see (2.12). Then $\Pi_X(W(f)) = U_X \Pi_T(W(T^{-1} f)) U_X^*$. Replacing $h = T^{-1} f$ yields $\Pi_V(W(h)) = \Pi_X(W(Th)) = U_X \Pi_T(W(h)) U_X^*$ for all $h \in E$. Hence, $X \in \mathscr{T}_F(E)$, if and only if $V = XT \in \mathscr{T}_T(E)$. Summarizing, we have shown the following relations:

$$\mathscr{T}_T(E) = \mathscr{T}_F(E) T \subseteq \mathscr{C}_{\mathrm{im}}(C,S) = \mathscr{C}_{\mathrm{im}}(\hat{C},0) T.$$

By Lemma 2.1 there exist in the hypothetic case $E = \mathscr{H}$ only bounded symplectic transformations, and thus the result immediately would arise from Warchall's Theorem 2.7. But for our proper dense $E \subset \mathscr{H}$ here we know that $\{X \in \mathscr{T}(E) | X \text{ is bounded}\} \equiv \{X \in \mathscr{T}(\mathscr{H}) | X(E) = E\}$ is a proper subgroup of $\mathscr{T}(\mathscr{H})$. For proving (II)(b) the reader may apply the Examples 3.9 and 3.10 to $\hat{C}$ to construct various $X = \exp\{JR_0\} \in \mathscr{T}(E)$ of the type (3.4) with self-adjoint, bounded (or possibly unbounded) operators R_0 having the e_k as eigenvectors but not being Hilbert–Schmidt, such that $R_0 \in \mathscr{R}(\hat{C})$ [and such that $\sinh(2R_0)\hat{C} P_{\hat{C}}([-a,a]) \in \mathrm{HS}$ for an $a > 0$ in the case of unbounded R_0]. Then, by Corollary 3.8 $\hat{T}_t = \exp\{it\hat{C}\}$ is continuously implementable in Π_X, or equivalently T_t in Π_{XT}. For proving part (II)(a) we replace T_t by $\hat{T}_t$ in the proof of part (I)(i)$\Rightarrow$(I)(ii) in Theorem 3.7, then we obtain from Eq. (3.8) that $P_{\hat{C}}([a,\infty[) \tanh(Z) \in \mathrm{HS}$ for all $a > 0$ [since $P_0^+ = \mathbf{1}$ for $\hat{C}$

$\geqslant 0$ and $I=I^{-1}$ commutes with $\tanh(Z)$]. The existence of an inessential infrared part for $\hat{C}$ is equivalent to the existence of an $a>0$ for which $P_{\hat{C}}([0,a[)\mathscr{H}$ is finite dimensional. Thus $\tanh(Z)\in \mathrm{HS}$, which yields $Z\in \mathrm{HS}$, and thus $X=W\exp\{IZ\}\in\mathscr{T}_F(E)$, which may be shown similarly to Proposition 3.3. ■

In order to study the diagonalization procedure in detail let us consider the two series $\Sigma_{k,c_k\neq 0}(s_k^2/c_k^2)$ and $\Sigma_{k,c_k\neq 0}(s_k^2/|c_k|)$ of the parts (c), resp. (d), of Theorem 4.12. The reader may easily construct examples for C and S with $|s_k|<|c_k|$ for $s_k\neq 0$ (diagonalization condition), such that none, resp. only one, resp. both, power series converge. However, there is some redundancy in the case of an inessential infrared part for C.

Proposition 4.14: If C has an inessential infrared part, then the condition $\Sigma_{k,c_k\neq 0}(s_k^2/|c_k|)<\infty$ implies $\Sigma_{k,c_k\neq 0}(s_k^2/c_k^2)<\infty$. This means that $\hat{C}-C$ being of trace class (i.e., the existence of κ) implies $T=\exp\{JR\}\in\mathscr{T}_F(E)$ and thus the existence of the implementing unitary U_T, provided that $\Sigma_{k,c_k=0}r_k^2<\infty$.

Proof: The condition of an inessential infrared part yields $c:=\inf\{|c_k||c_k\neq 0\}>0$, and thus

$$\infty>\sum_{k,c_k\neq 0}\frac{s_k^2}{|c_k|}=\sum_{k,c_k\neq 0}|c_k|\,\frac{s_k^2}{c_k^2}\geqslant c\sum_{k,c_k\neq 0}\frac{s_k^2}{c^2},$$

which gives the result. ■

V. APPLICATIONS

A. Connection between the Dirac- and the harmonic oscillator-quantization in QED

Let us consider a specific Boson system, namely, the photon field. We assume $\mathscr{H}$ to be the Hilbert space $\mathbf{L}^2_{\mathrm{div}}(\Lambda,\mathbb{C}^3)$ of $\mathbb{C}^3$-valued, divergence-free (i.e., $\boldsymbol{\nabla}\cdot f=0$ in the distributional sense) functions on the quantization volume $\Lambda\subseteq\mathbb{R}^3$ in position space, where Λ is equipped with suitable boundary conditions for the component-wise acting Laplacian Δ. The antilinear involution J on $\mathscr{H}$ is chosen to be the component-wise complex conjugation, i.e., $Jf(x)=\overline{f(x)}$, $\forall x\in\Lambda$, $f\in\mathscr{H}$.

A dense complex testfunction space $E\subseteq\mathscr{H}$ is appropriate for QED in the Coulomb gauge, where the conjugate fields are the vector potential $\mathbf{A}$ and $\boldsymbol{\Pi}=-\epsilon_0^{-1}\mathbf{E}^T$ (ϵ_0 is the dielectric constant of the vacuum) with the transversal electric field $\mathbf{E}^T$ (cf., e.g., Ref. 26). As usual, we rescale the physical constants ϵ_0, the velocity of light, and $\hbar$ to unity. After smearing,[27] we form the fundamental field expressions

$$\Phi_D(f)=\mathbf{A}(\mathrm{Re}(f))+\boldsymbol{\Pi}(\mathrm{Im}(f)),\quad f\in E, \tag{5.1}$$

which, according to the analogy with $\Sigma_{j=1}^n(q_j\,\mathrm{Re}(c_j)+p_j\,\mathrm{Im}(c_j))$ for finitely many degrees of freedom, formally have to satisfy the canonical commutation relations CCR,

$$[\Phi_D(f),\Phi_D(g)]\subseteq i\,\mathrm{Im}\langle f|g\rangle\mathbb{1},\quad f,g\in E. \tag{5.2}$$

The mathematical realization of the CCR in Eq. (5.2), however, requires a complex Hilbert space on which the field expressions $\Phi_D(f)$, $f\in E$, are well-defined self-adjoint operators. Thus, one commonly turns over to the Weyl form of the CCR (2.4), which defines also here the C^*-algebra $\mathscr{W}(E)$ of observables for the photon field. Then the CCR (5.2) are recovered in every regular representation Π of $\mathscr{W}(E)$, where the field operators $\Phi_D(f)\equiv\Phi_\Pi(f)$, $f\in E$, are realized according to Eq. (2.5) as well-defined self-adjoint operators on the representation Hilbert space $\mathscr{K}_\Pi$. This procedure is a straightforward generalization of Dirac's quantization description [as indicated by the index D in $\Phi_D(f)$] and is in coincidence with Ref. 28, $D_{III}1(9)$, if one makes explicit the projection onto the transversal components in the scalar product (cf. also Refs. 29–31).

For simplicity, we assume Λ to be bounded and its boundary conditions such that $\langle f|-\Delta g\rangle=\langle\boldsymbol{\nabla}\times f|\boldsymbol{\nabla}\times g\rangle$ for $f,g\in E$. Since Δ and J commute, there exists an orthonormal base (ONB) $\{e_k|k\in\mathbb{N}\}$ of $\mathscr{H}$ consisting of common eigenvectors, $-\Delta e_k=\omega_k^2 e_k$, where the increasing eigenfrequencies $0<\omega_1<\omega_2<\cdots$ accumulate at infinity depending linearly on the absolute momentum values, and $Je_k=e_k$, i.e., $e_k(x)\in\mathbb{R}^3$, $\forall x\in\Lambda$. Then the classical Hamiltonian function,

$$H_{\text{class}}=\frac{1}{2}\int_{\Lambda}[\mathbf{\Pi}(x)^2+(\nabla\times\mathbf{A}(x))^2]d^3x,$$

leads—after replacing the classical functions $\mathbf{A}(x)$ and $\mathbf{\Pi}(x)$, $x\in\Lambda$, by the associated field operators from Eq. (5.1)[32]—to the (formal) quantized Hamiltonian of the free field with respect to the Dirac quantization,

$$\begin{aligned}H&=\frac{1}{2}\sum_{k=1}^{\infty}[\Phi_D(ie_k)^2+\omega_k^2\Phi_D(e_k)^2]\\&=\sum_{k=1}^{\infty}\left[\frac{\omega_k^2-1}{4}(a_D^*(e_k)^2+a_D(e_k)^2)+\frac{\omega_k^2+1}{2}a_D^*(e_k)a_D(e_k)+\frac{\omega_k^2+1}{4}\mathbb{1}\right]\end{aligned}\tag{5.3}$$

(for more details we refer to Ref. 29). The annihilation and creation operators, $a_D(f)$ and $a_D^*(f)$, $f\in E$ are defined as in (2.6). Let us emphasize, that up to now the representation Π of $\mathscr{W}(E)$ is not specified, in which the field operators $\Phi_D(f)$, resp. the Hamiltonian H, are considered to be realized.

The more popular quantization procedure uses field operators, which lead to the diagonalized oscillator form of the field Hamiltonian. This Hamiltonian is obtained by means of the Bogoliubov transformation $(\alpha_T)^{-1}=\alpha_{T^{-1}}$ of $\mathscr{W}(E)$ associated with the symplectic $T=\exp\{JR\}=\cosh(R)+J\sinh(R)\in\mathscr{T}(E)$, where the self-adjoint operator R acting on $\mathscr{H}$ is defined by

$$R:=\ln((-\Delta)^{-1/4}).$$

Observation 5.1: R has a purely discrete spectrum: $Re_k=r_ke_k$, where $r_k:=\ln(\omega_k^{-1/2})$ for all $k\in\mathbb{N}$, which implies $Te_k=\omega_k^{-1/2}e_k$ and $T(ie_k)=\omega_k^{1/2}(ie_k)$, $\forall k\in\mathbb{N}$. Moreover, it holds that $Tf=(-\Delta)^{-1/4}\,\mathrm{Re}(f)+i(-\Delta)^{1/4}\,\mathrm{Im}(f)$ for all $f\in E$ (for more details to this decomposition we refer to Ref. 12, Subsection 3.1).

Assuming $\Phi_D(f)$ to be realized in the representation Π with Hilbert space $\mathscr{K}_\Pi$, the transformed field operators $\Phi_{\text{osc}}(f)$ are then given by

$$\Phi_{\text{osc}}(Tf)=\Phi_D(f),\quad \forall f\in E,\tag{5.4}$$

resp. $\Phi_{\text{osc}}(f)=\Phi_D(T^{-1}f)$, $\forall f\in E$, where the $\Phi_{\text{osc}}(f)$, $f\in E$, are the field operators for the representation $\Pi\circ(\alpha_T)^{-1}$ of $\mathscr{W}(E)$ (also having $\mathscr{K}_\Pi$ as representation Hilbert space). With Observation 5.1 we obtain for every $k\in\mathbb{N}$ the relations $\Phi_D(e_k)=\omega_k^{-1/2}\Phi_{\text{osc}}(e_k)$ and $\Phi_D(ie_k)=\omega_k^{1/2}\Phi_{\text{osc}}(ie_k)$. Insertion into (5.3) gives

$$H=\frac{1}{2}\sum_{k=1}^{\infty}w_k[\Phi_{\text{osc}}(ie_k)^2+\Phi_{\text{osc}}(e_k)^2]=\sum_{k=1}^{\infty}w_k\left[a_{\text{osc}}^*(e_k)a_{\text{osc}}(e_k)+\frac{1}{2}\mathbb{1}\right]\tag{5.5}$$

[the annihilation and creation operators, $a_{\text{osc}}(f)$ and $a_{\text{osc}}^*(f)$, $f\in E$, are defined in the standard manner (2.6)]. Thus, the Bogoliubov transformation (5.4) formally diagonalizes indeed the Hamiltonian H.

In order to give (5.5) a rigorous meaning as a self-adjoint operator, H is usually interpreted in the (oscillator-) Fock representation Π_F of $\mathscr{W}(E)$. That is, one makes the identification $\Pi_F\equiv\Pi\circ(\alpha_T)^{-1}$ and the $\Phi_{\text{osc}}(f)\equiv\Phi_F(f)$, $f\in E$, are identified with the Fock field operators. If one drops the vacuum self-energy $\frac{1}{2}\sum_{k=1}^{\infty}\omega_k$ one finds, in fact, that H reduces to

$$H^F_{\sqrt{-\Delta},0}=\sum_{k=1}^{\infty}w_ka_F^*(e_k)a_F(e_k)\geqslant 0,\tag{5.6}$$

which is a well-defined self-adjoint operator on Fock space, and which is just the second quantization of the so-called one-photon Hamiltonian $\sqrt{-\Delta}$ (for the second quantization, see, e.g., Refs. 33, Ref. 2, Sec. 5.2.1, or Ref. 3, Sec. X.7).

Since we have interpreted Eq. (5.5) in the Fock representation Π_F, Eq. (5.3) has to be interpreted in the symplectically related representation $\Pi_T=\Pi_F\circ\alpha_T\equiv\Pi$, that is, $\Phi_D(f)\equiv\Phi_T(f)$, $f\in E$. Hence H from (5.3) agrees [up to the diverging constant $\frac{1}{4}\Sigma_{k=1}^{\infty}(\omega_k^2+1)$] with the formal Hamiltonian $H_{C,S}^T$ defined in (4.3), where the self-adjoint one-photon operators C and S on $\mathscr{H}$ have the modes e_k as eigenvectors,

$$Ce_k=c_ke_k,\quad Se_k=s_ke_k,\quad \text{where}\quad c_k:=\frac{w_k^2+1}{2},\quad s_k:=\frac{w_k^2-1}{2},\quad \forall k\in\mathrm{N}. \tag{5.7}$$

However, $H_{C,S}^T$ differs from the diagonal $H^F_{\sqrt{-\Delta},0}$ by the diverging renormalization term $\kappa=1/4\Sigma_{k=1}^{\infty}(2\omega_k-\omega_k^2-1)$ according to Eq. (4.12). Thus, the formal Hamiltonian $H_{C,S}^T$ does not exist as a self-adjoint operator.

Let us turn to the one-photon level. With the operators C,S,R, and the involution J we are in the situation of Sec. III. As mentioned at the beginning of Sec. IV, the expression $\exp\{itH_{C,S}^T\}$ formally implements the symplectic group $T_t:=\exp\{it(C-JS)\}\in\mathscr{T}(E)$, $t\in\mathbb{R}$, in the symplectically related representation Π_T. (The one-photon test function space E has to be chosen so that the T_t and $T=\exp\{JR\}$ leave E invariant, e.g., as the finite (complex-) linear combinations of the ONB elements.) However, since R is not Hilbert–Schmidt, the Bogoliubov transformation α_T, resp. its inverse, is *not* unitarily implementable in the Fock representation, and thus the symplectically related representation Π_T is *disjoint* to the Fock representation Π_F (see Theorem 2.3 and Proposition 3.3).

For the transformed symplectic group $\hat{T}_t=TT_tT^{-1}=\exp\{it(\hat{C}-J\hat{S})\}\in\mathscr{T}(E)$ from (3.3), we obtain by the equations (3.5) that $\hat{S}=0$ and that $C=\sqrt{-\Delta}$ is the one-photon Hamiltonian, i.e., $\hat{c}_k=\omega_k$, $\forall k\in\mathbb{N}$. This means that our $T=\exp\{JR\}$ diagonalizes T_t in the sense of Definition 4.11 [observe that (5.7) implies $|s_k|<|c_k|$, $\forall k\in\mathbb{N}$, and thus the Theorems 4.12 and 4.13 are applicable] with the diagonalization

$$\hat{T}_t=\exp\{it\hat{C}\}=\exp\{it\sqrt{-\Delta}\},\quad t\in\mathbb{R},$$

which is the one-photon dynamics on the one-photon test function space. From Proposition 4.8 we know that the diagonal self-adjoint operator $H^F_{\sqrt{-\Delta},0}$ from (5.6), resp. the unitary group $\exp\{itH^F_{\sqrt{-\Delta},0}\}$, (continuously) implements the symplectic group $\hat{T}_t$ in the Fock representation Π_F and T_t in the symplectically related representation Π_T.

Summing up our somewhat complicated analysis of the usual quantization procedure, we want to emphasize the following points: The fundamental quantization strategy of Dirac leads to a unique C^*-Weyl algebra $\mathscr{W}(E)$, and in combination with a kind of correspondence principle to a formal squeezing Hamiltonian (5.3). In order to find a well-defined dynamics, which comes as close to the formal procedure as possible, one should first extract from (5.3) the one-photon dynamics $T_t=\exp\{it(C-JS)\}$ and introduce the corresponding quasifree automorphism dynamics on $\mathscr{W}(E)$. Since the diagonalization condition is fulfilled, one has a particle representation for the quantized electrodynamic field system. By Theorem 4.13 (II) (a) this representation is unique up to unitary equivalence in a finite cavity, for which $\hat{C}=\sqrt{-\Delta}$ has an inessential infrared part. The infinite volume limit apparently has an essential infrared part, but transcends the present formalism.

B. The Luttinger Hamiltonian

The expression (1.1) is not the most general form of a quadratic Hamiltonian. However, the form (1.1) comprises also some quadratic Hamiltonians, which on-the first look seem to be more general than (1.1). Consider the Hamiltonian,

$$H=\sum_{m=1}^{\infty}\left[(\eta_m a^*(u_{1,m})a^*(u_{2,m})+\overline{\eta_m}a(u_{1,m})a(u_{2,m}))+p_m(a^*(u_{1,m})a(u_{1,m})+a^*(u_{2,m})a(u_{2,m}))\right], \tag{5.8}$$

where $\eta_m\in\mathbb{C}$ and $p_m\in\mathbb{R}$. The indices 1 and 2 distinguish two kinds of modes $u_{1,m}$ and $u_{2,m}$, $m\in\mathbb{N}$, which all are normalized and mutually orthogonal (in quantum optical squeezing processes

1 may stand for the signal modes and 2 for the idler modes). Here for each $m \in \mathbb{N}$ the modes $u_{1,m}$ and $u_{2,m}$ are assumed to have the same eigenfrequencies p_m. Such quadratic Hamiltonians are brought into the form (1.2) by superposing both kinds of modes: The combined orthonormalized modes e_k, $k \in \mathbb{N}$, are chosen as

$$e_{2m-1} := \frac{\alpha_m}{\sqrt{2}}(u_{1,m}+u_{2,m}), \quad e_{2m} := \frac{\alpha_m}{\sqrt{2}}(u_{1,m}-u_{2,m}), \quad \forall m \in \mathbb{N},$$

where $\alpha_m := \exp\{(i/2)\arg(\eta_m)\}$ (cf. the phase transformation from u_k to e_k in the Introduction; also Refs. 5 and 6). From the linearity of $f \mapsto a^*(f)$, one obtains the transformation $\eta_m a^*(u_{1,m})a^*(u_{2,m}) = (|\eta_m|/2)(a^*(e_{2m-1})^2 - a^*(e_{2m})^2)$, which finally leads to

$$H = \sum_{k=1}^{\infty} \left[\frac{s_k}{2}(a^*(e_k)^2 + a(e_k)^2) + c_k a^*(e_k)a(e_k) \right], \tag{5.9}$$

where $s_{2m-1} = |\eta_m|$, $s_{2m} = -|\eta_m|$, and $c_{2m-1} = c_{2m} = p_m$ for all $m \in \mathbb{N}$. Such a situation occurs in the popular Luttinger model.

The Luttinger model (cf., e.g., Ref. 34 for more physical references, and Refs. 35, 36 for more mathematical references) deals with the thermodynamics and statistical mechanics of a self-interacting, one-dimensional Fermion field enclosed in a one-dimensional periodic box of length $0 < L < \infty$. It is directly connected with the Thirring model in field theory and the Tomonaga model for massive, spinless Fermions. In the literature on solid-state physics and on field theory there also exist various generalizations of the Luttinger model (cf. Refs. 34–36, and references therein).

The attraction of the Luttinger model arises from its complete solubility. This is insofar important since certain infrared divergences make an approach based on conventional Fermion many-body perturbation theory useless. The solution in effect resums all these divergences. Hence the Luttinger model attracted much attention and has been used for checking the validity of approximations and testing new ideas.

The feature that allows the solution of the Luttinger model is its exact linear Fermion dispersion. Then by means of the so-called bosonization, the interacting (Fermionic) Hamiltonian K_L can be transformed into a quadratic expression H_L in Boson collective modes and the charge operators $Q_\pm$ for the Fermions moving in the plus and minus direction, that is, $K_L = H_L + (\pi/L)(Q_+^2 + Q_-^2)$, where H_L is of type (5.8) with $p_m = (2\pi/L)m$ and $\eta_m = -p_m w_m$ for all $m \in \mathbb{N}$, and the $w_m \in \mathbb{C}$ are the Fourier coefficients of the 4-Fermion potential.[36] More exactly, the bosonization leads to an orthonormalized set of modes u_m of unspecified nature, where first $m = \pm 1, \pm 2, \pm 3, \ldots$. Then the form (5.8) for H_L arises when identifying the above two kinds of Bosons by $u_{1,m} \equiv u_m$ and $u_{2,m} \equiv u_{-m}$, where now $m = 1,2,3,\ldots$. The Fourier coefficients w_m are assumed to be real and to satisfy the following three conditions:[36]

$$(L-1)\ |w_m| < 1, \quad \forall m \in \mathbb{N}; \quad (L-2)\ \sum_{m=1}^{\infty} |w_m|^2 < \infty; \quad (L-3)\ \sum_{m=1}^{\infty} m|w_m|^2 < \infty.$$

As mentioned above, H_L may be transformed into the form (5.9), where now $c_{2m-1} = c_{2m} = p_m = (2\pi/L)m$ and $s_{2m-1} = -s_{2m} = p_m|w_m|$ for all $m \in \mathbb{N}$. Applying the diagonalization Theorems 4.12 and 4.13 for the complex one-Boson test function space E, which is (finitely) spanned by $\{e_k | k \in \mathbb{N}\}$, resp. by $\{u_m | m = \pm 1, \pm 2, \ldots\}$, we find the following results.

- $(L-1)$ is equivalent to the diagonalizability of H_L [in our generalized sense, where in Eq. (4.12) the operator $H_{C,S}^T$ is identified with H_L).
- $(L-2)$ is equivalent to the implementability of the diagonalizing symplectic $T = \exp\{JR\}$ in the Fock representation Π_F of $\mathscr{W}(E)$ with the implementing unitary U_T (which in turn is equivalent to R being Hilbert–Schmidt).
- $(L-3)$ is equivalent to the existence of the renormalization constant $\kappa \in \mathbb{R}$.

That $(L-3)$ implies $(L-2)$ is immediate and reveals an application of Proposition 4.14. Thus, whenever $(L-1)$ to $(L-3)$ are valid, the diagonalization of H_L may be treated according to Eq. (4.11) in the Fock representation. Of course, for the Luttinger Hamiltonian this Fock diagonalization [that is, by assuming the conditions $(L-1)$ to $(L-3)$] has been done in the literature, but on the level of the Bosonic operators (the level of the second quantization), only.

It has been shown in the literature that the commuting charges $Q_\pm$ both have the discrete spectrum $\mathbb{Z}$. Since the $Q_\pm$ also commute with the Bosonic annihilation and creation operators, and thus with H_L, each joint spectral subspace of the $Q_\pm$ reduces the CCR system. This suggests the realization of the Luttinger dynamics as a tensor product: Let us realize the dynamics arising from the part $(\pi/L)(Q_+^2+Q_-^2)$ of K_L as a group of *-automorphisms $\beta_t, t\in\mathbb{R}$, on an associated commutative C^*-algebra $\mathscr{A}$. As mentioned at the beginning of Sec. IV, the formal Luttinger–Boson Hamiltonian H_L in its canonical form (5.9) leads to the symplectic group $T_t=\exp\{it(C-JS)\}\in\mathscr{T}(E)$. Consequently, the dynamics arising from H_L is given by the group of Bogoliubov *-automorphisms $\alpha_{T_t}, t\in\mathbb{R}$, on the CCR-algebra $\mathscr{W}(E)$. The combination of both gives the Luttinger dynamics in terms of the tensor product,

$$\{\alpha_{T_t}\otimes\beta_t | t\in\mathbb{R}\}, \quad \text{on} \quad \mathscr{W}(E)\otimes\mathscr{A}. \tag{5.10}$$

Thus we arrived at a purely C^*-algebraic formulation of the Luttinger dynamics, which is independent of any Hilbert space representation and is also independent of the existence of the formal Hamiltonian H_L as a self-adjoint operator on any representation Hilbert space. Obviously, the diagonalization of T_t (if it exists) takes place only on the first factor of the tensor product, and it is given according to Definition 4.11 and Theorem 4.12 on the one-Boson level, only. This illustrates the one-Boson aspects accompanying the dynamics and the diagonalization of the Luttinger Hamiltonian.

The insight we gain from our foregoing analysis is the possibility of a considerable generalization of the Luttinger model. Since the C^*-algebraic formulation of the Luttinger dynamics in (5.10) is well defined for arbitrary Fourier coefficients $w_m\in\mathbb{C}$, $m\in\mathbb{N}$, it is not necessary that the symplectic Luttinger-one-Boson dynamics $T_t=\exp\{it(C-JS)\}$ is diagonalizable at all. Thus, the Luttinger model is treatable with very arbitrary 4-Fermion potentials. A well-defined, in general, renormalized Hamiltonian is found in each representation, where the dynamics (5.10) is continuously implementable.

However, if in order to have a particle interpretation a diagonalization is desired (in our generalized sense), then the complex Fourier coefficients w_m only have to fulfill the necessary and sufficient condition $(L-1)$ for the diagonalizability of $\{T_t | t\in\mathbb{R}\}$. Then the diagonalized Hamiltonian may be well defined in a representation, which is disjoint to the original representation.

As is demonstrated in Ref. 37 the connection to the Fermion algebra is mathematically subtle. Assuming this Boson–Fermion correspondence being solved, our foregoing analysis allows us to consider the model on the C^*-algebraic level independently of any Hilbert space representation (and thus also in every implementing representation inequivalent to Fock) and to incorporate very arbitrary 4-Fermion potentials in the theory of the Luttinger model, e.g., our class of interactions would cover purely local 4-Fermion theories.

ACKNOWLEDGMENTS

This work has been supported by the Deutsche Forschungsgemeinschaft.

[1] H. P. Yuen, ''Generalized coherent states and the statistics of two-photon-lasers,'' Phys. Lett. A **51**, 1–2 (1975); H. P. Yuen, ''Two-photon coherent states of the radiation field,'' Phys. Rev. A **13**, 2226–2243 (1976); D. F. Walls, ''Squeezed states of light,'' Nature (London) **306**, 141–146 (1983); R. Loudon and P. L. Knight, ''Squeezed light,'' J. Mod. Opt. **34**, 709–759 (1987); X. Ma and W. Rhodes, ''Multimode squeeze operators and squeezed states,'' Phys. Rev. A **41**, 4625–4631 (1990).

[2] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I, II* (Springer, Berlin, 1979, 1981).

[3] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I, II* (Academic, New York, 1980, 1975).

[4] E. Prugovečki, *Quantum Mechanics in Hilbert Spaces* (Academic, New York, 1971).

[5] R. Honegger and A. Rieckers, ''Squeezing operations in Fock space and beyond,'' Physica A **242**, 423–438 (1997).

[6] R. Honegger and A. Rieckers, ''Squeezing Bogoliubov transformations on the infinite mode CCR-algebra,'' J. Math. Phys. **37**, 4292–4309 (1996).

[7] H. A. Warchall, ''Implementation of automorphism groups in certain representations of the canonical commutation relations,'' J. Math. Phys. **23**, 2221–2228 (1982).
[8] I. E. Segal, ''Foundations of the theory of dynamical systems of infinitely many degrees of freedom,'' Part I in Mat. Fys. Medd. Dansk. Vid. Selsk. **31**, 1–39 (1959); Part II in Can. J. Math. **13**, 1–18 (1961); I. E. Segal, ''Mathematical characterization of the physical vacuum for a linear Bose–Einstein field,'' Illinois J. Math. **6**, 500–523 (1962).
[9] M. Weinless, ''Existence and uniqueness of the vacuum for linear quantized fields,'' J. Funct. Anal. **4**, 350–379 (1969).
[10] H. Araki and E. J. Woods, ''Representations of the canonical commutation relations describing a nonrelativistic infinite free Bose gas,'' J. Math. Phys. **4**, 637–662 (1963).
[11] R. Honegger and A. Rieckers, ''Squeezing of optical states on the CCR-algebra,'' to appear in Publ. RIMS Kyoto Univ.
[12] R. Honegger and A. Rieckers, ''Squeezed variances of smeared Boson fields,'' Helv. Phys. Acta **70**, 507–541 (1997).
[13] R. Honegger, ''On the continuous extension of states on the CCR-algebra,'' Lett. Math. Phys. **42**, 11–25 (1997).
[14] R. Honegger, ''Global cocycle dynamics for infinite mean field quantum systems interacting with the boson gas,'' J. Math. Phys. **37**, 263–282 (1996).
[15] R. Honegger, ''Enlarged testfunction spaces for the global free Folia dynamics on the CCR-algebra,'' to appear in J. Math. Phys.
[16] R. Haag, R. V. Kadison, and D. Kastler, ''Nets of C^*-algebras and classification of states,'' Commun. Math. Phys. **16**, 81–104 (1970); G. L. Sewell, ''States and dynamics of infinitely extended physical systems,'' *ibid.* **33**, 43–51 (1973); R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras I, II* (Academic, New York, 1983, 1986).
[17] A bounded linear (resp. antilinear) operator D on the Hilbert space $\mathscr{H}$ is called Hilbert–Schmidt, if its Hilbert–Schmidt norm $\|D\|_{\mathrm{HS}} := (\Sigma_{e\in\mathscr{O}}\|De\|^2)^{1/2}$ is finite, where $\|\cdot\|_{\mathrm{HS}}$ does not depend on the specific ONB $\mathscr{O}$ chosen for $\mathscr{H}$ (ONB means orthonormal base); cf., e.g., Ref. 3, Sec. VI.6.
[18] T. Gerisch, A. Rieckers, and S. Zanzinger, ''Operator algebraic transition probabilities and disjointness,'' Tübingen, preprint (1997).
[19] D. Shale, ''Linear symmetries of free Boson fields,'' Trans. Am. Math. Soc. **103**, 149–167 (1962).
[20] F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966).
[21] D. Petz, ''An invitation to the algebra of canonical commutation relations,'' in *Leuven Notes in Mathematical and Theoretical Physics* (Leuven University, Leuven, 1990), Vol. 2.
[22] A. van Daele and A. Verbeure, ''Unitary equivalence of Fock representations on the Weyl algebra,'' Commun. Math. Phys. **20**, 268–278 (1971).
[23] E. B. Davies, *One-Parameter Semigroups* (Academic, New York, 1980).
[24] J. Weidmann, *Linear Operators in Hilbert Spaces* (Springer, Berlin, 1980).
[25] H. Araki, ''On the diagonalization of a bilinear Hamiltonian by a Bogoliubov transformation,'' Publ. RIMS Kyoto Univ. Ser. A **4**, 387–412 (1968).
[26] R. Loudon, *The Quantum Theory of Light* (Clarendon, Oxford, 1979).
[27] The smearing of the classical observable represented by the divergence-free vector function $\mathbf{G}(x)\in\mathbb{R}^3$ is given by $\mathbf{G}(g)=\int_\Lambda \mathbf{G}(x)\cdot g(x)d^3x=\langle \mathbf{G}|g\rangle$ for every real test function $g\in E$.
[28] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Photons & Atoms, Introduction to QED* (Wiley, New York, 1989).
[29] R. Honegger, ''Globale Quantentheorie der Strahlung,'' Doctoral thesis, Tübingen, 1991.
[30] H. Haken, *Light 1* (North-Holland, Amsterdam, 1981).
[31] H. M. Nussenzveig, *Introduction to Quantum Optics* (Gordon and Breach, London, 1973).
[32] It is $H_{\mathrm{class}}=1/2(\|\Pi\|^2+\|\nabla\times\mathbf{A}\|^2)$. The decomposition with the ONB $\{e_k|k\in\mathbb{N}\}$ yields $\|\mathbf{\Pi}\|^2=\Sigma_k\mathbf{\Pi}(e_k)^2$ and $\|\nabla\times\mathbf{A}\|^2=\langle\mathbf{A}|-\Delta\mathbf{A}\rangle=\Sigma_k\omega_k^2\mathbf{A}(e_k)^2$, which gives $H_{\mathrm{class}}=1/2\Sigma_k[\mathbf{\Pi}(e_k)^2+\omega_k^2\mathbf{A}(e_k)^2]$. Now the replacement of the classical smeared field observables $\mathbf{\Pi}(e_k)$ and $\mathbf{A}(e_k)$ by the quantized field expressions leads to (5.3). Hereby Eq. (5.1) has been used, taking into account that the eigenmode functions e_k are real valued, $Je_k=e_k$, respectively, $e_k(x)\in\mathbb{R}^3$, $\forall x\in\Lambda$.
[33] J. Cook, ''Mathematics of the second quantization,'' Trans. Am. Math. Soc. **74**, 222–245 (1953).
[34] F. D. M. Haldane, '''Luttinger liquid theory' of one-dimensional quantum fields: I. Properties of the Luttinger model and their extension to the general 1D interacting spinless Fermi gas,'' J. Phys. C **14**, 2585–2609 (1981); C. B. Dover, ''Properties of the Luttinger model,'' Ann. Phys. (N.Y.) **50**, 500–533 (1968).
[35] E. H. Lieb and D. C. Mattis, *Mathematical Physics in One Dimension* (Academic, New York, 1966); D. A. Uhlenbrock, ''Fermions and associated bosons of one-dimensional model,'' Commun. Math. Phys. **4**, 64–76 (1967).
[36] R. Heidenreich, R. Seiler, and D. A. Uhlenbrock, ''The Luttinger model,'' J. Stat. Phys. **22**, 27–57 (1980).
[37] A. L. Carey and K. C. Hannabuss, ''Temperature states on loop groups, theta functions and the Luttinger model,'' J. Funct. Anal. **75**, 128–160 (1987).

Quantum theory of the real and the complexified projective line

Frank Leitenberger[a)]
Institut für Theoretische Physik, Universität Wien, Boltzmanngasse 5, 1090 Wien, Austria

(Received 28 October 1996; accepted for publication 6 October 1997)

Quantum deformations of sets of points of the real and the complexified projective line are constructed. These deformations depend on the deformation parameter q and certain further parameters λ_{ij}. The deformations for which the subspace of polynomials of degree three has a basis of ordered monomials are selected. We show that the corresponding algebras of three points have ''polynomiality.'' Invariant elements which turn out to be cross ratios in the classical limit are defined. For the special case $|\lambda_{ij}|=1$ a quantum cross ratio with properties similar to the classical case is presented. As an application a quantum version of the real Euclidean distance is given. © *1998 American Institute of Physics.*
[S0022-2488(98)00402-2]

I. INTRODUCTION

Noncommutative geometry can be viewed as the quantum deformation of an increasing set of mathematical objects and theories. In this article we introduce a quantized version of the projective geometry in one dimension.

First we consider the algebra of coordinates of points t_i, $i \in I$, of the real line. All possible Poisson brackets between these variables (which are covariant with respect to the action of the standard Poisson structure of the Möbius group) are parametrized by a large set of parameters λ_{ij}, $i,j \in I$, which obey certain second-order polynomial conditions. These quasi-classical structures give a good insight into the quantized situation.

In Sec. III we replace the algebra of coordinates t_i by a $U_q(\mathrm{sl}(2,\mathbb{R}))$-module algebra of noncommutative coordinates v_i. We show that (up to reordering the points) third degree polynomials have a basis of ordered polynomials if a condition in analogy to the Jacobi identity is satisfied. There is one exceptional structure, which has no quasi-classical limit.

For algebras of three points we give a proof that the subspaces of homogeneous polynomials of higher degrees have also the classical dimension. The condition (which plays the role of the Jacobi identity) takes the form of the Einstein addition law for velocities, where the light velocity corresponds to the quantization parameter q. We consider also the complexified case.

In Sec. IV we consider invariant elements C_{ijkl} of four points, which turn out to be cross ratios in the classical limit. Invariant elements occurred first in Ref. 1 in a similar situation for the quantum sphere. Our class of invariant elements contain the elements of the type of Ref. 1 as a subclass. However, these elements do not satisfy a reality condition (*-invariance) and have no simple behavior with respect to permutations of the four points. We show that for parameters $|\lambda_{ij}|=1$ certain rational functions $f(C_{ijkl},q)$ have the classical properties of the cross ratio.

As an application we define a quantization of the Euclidean distance on the real line.

One can say that we quantize products of one-dimensional spaces equipped with general $\mathrm{SL}(2,\mathbb{R})$ covariant Poisson brackets. We discuss algebraic aspects with topological aspects left open.

This article is the first part of a program of the author to develop a quantum deformation of certain parts of the geometry and analysis of the real line (cross ratios, invariant theory of binary forms, elliptic and Abelian differentials, etc.).

We used the computer algebra program Mathematica 3.0 forming computations in III.C and III.E.

[a)]Electronic mail: leitenbe@doppler.thp.univie.ac.at

 © 1998 American Institute of Physics

II. THE QUASI-CLASSICAL LIMIT

Consider the one-dimensional real projective line $\mathbb{P}$, the action of the Möbius group G on it, and an arbitrary set of variable points $P_i \in \mathbb{P}$, $i \in I$, on it. Let t be a coordinate on $\mathbb{P}$ and t_i the coordinates of the points P_i and let the Möbius group $G \cong \mathrm{SL}(2,\mathbb{R})/\{\pm I\}$ act on the points P_i according to

$$t_i \to \frac{at_i + b}{ct_i + d}.$$

Endow G with the structure of a Poisson group:

$$\{a,b\} = ab, \quad \{a,c\} = ac, \quad \{a,d\} = 2bc,$$

$$\{b,c\} = 0, \quad \{b,d\} = bd, \quad \{c,d\} = cd.$$

Let I be an arbitrary set. The general G-covariant (cf. Ref. 2) pre-Poisson bracket on the algebra of polynomials of the variables t_i, $i \in I$, is given by

$$\{t_j, t_i\} = t_j^2 - t_i^2 - \lambda_{ij}(t_i - t_j)^2 \tag{1}$$

with arbitrary coefficients $\lambda_{ij} = -\lambda_{ji}$.

Proof: The G-covariant pre-Poisson brackets are given as sums of a special G-covariant bracket and an arbitrary G-invariant bracket (cf. Ref. 2). The general G-invariant bracket is given by

$$\{t_j, t_i\} = \lambda_{ij}(t_i - t_j)^2.$$

Therefore it is sufficient to check the G-covariance for one of the above structures by explicit calculation. ■

A calculation yields that this pre-Poisson bracket becomes a Poisson bracket (i.e., it satisfies the Jacobi identity) if and only if

$$\lambda_{ij}\lambda_{jk} - \lambda_{ij}\lambda_{ik} - \lambda_{ik}\lambda_{jk} = -1 \tag{2}$$

for three different arbitrary $i,j,k \in I$. The relation is invariant with respect to a permutation of the indices.

Instead of giving the general solution of this equation we prefer to describe some basic examples.

Example 1: Let I be a linear ordered set. I may have arbitrary cardinality. Set $\lambda_{ij} = 1$ for $i < j$, $i,j \in I$. We obtain

$$\{t_j, t_i\} = 2t_i(t_j - t_i), \quad i < j.$$

Conversely, let $|\lambda_{ij}| = 1$ for arbitrary different $i,j \in I$. Then we can define a linear order relation on I. We set $i < j$, if $\lambda_{ij} = 1$ (the transitivity law of the order relation corresponds to the Jacobi identity).

Example 2: Let $I = \mathbb{Z}$ be the set of integers, $(\alpha_k)_{k \in \mathbb{Z}}$ be a sequence of positive integers, and let

$$\lambda_{ij} = \coth\left(\sum_{k=i}^{j-1} \alpha_k\right), \quad \alpha_k > 0, \quad i < j.$$

The Poisson property follows from the addition formula

$$\coth(\alpha + \beta) = \frac{1 + \coth(\alpha)\coth(\beta)}{\coth(\alpha) + \coth(\beta)}.$$

We have

$$\{t_j,t_i\}=t_j^2-t_i^2-\coth\left(\sum_{k=i}^{j-1}\alpha_k\right)(t_i-t_j)^2,\quad i<j.$$

Note that $|\lambda_{ij}|>1$ because of $|\coth(\alpha)|>1$ for $\alpha\neq 0$. We obtain again Example 1 in the limit $\alpha_k\to\infty$.

Example 3: In order to get examples with $|\lambda_{ij}|<1$ for certain $i,j\in I$, let $I=\mathbb{Z}$ and

$$\lambda_{ij}=\coth\left(\sum_{k=i}^{j-1}\left(\alpha_k+\frac{i\pi}{2}\right)\right),\quad \alpha_k>0.$$

Then we have $|\lambda_{ij}|<1$ for $i-j$ odd and $|\lambda_{ij}|>1$ for $i-j$ even.

Example 4: Now we give an example with $\lambda_{ij}=0$ for certain $i,j\in I$. We remark that if i is fixed, there is only one different index j with $\lambda_{ij}=0$ [because of Eq. (2)]. Let $I=\mathbb{Z}\times\mathbb{Z}_2$ and $z_i:=v_{(i,0)}$ and $z_{\overline{i}}^*:=v_{(i,1)}$ be the corresponding variables. One can check Eq. (2) for $\lambda_{i,\overline{i}}=0$ and $\lambda_{ij}=\lambda_{\overline{i}\,\overline{j}}=\lambda_{i\,\overline{j}}=\lambda_{\overline{i}j}=1$ for $i<j$. The Poisson bracket is given by

$$\{z_{\overline{i}}^*,z_i\}=(z_{\overline{i}}^{*2}-z_i^2),$$

$$\{z_j,z_i\}=2z_i(z_j-z_i),\quad i<j,$$

$$\{z_{\overline{j}}^*,z_{\overline{i}}^*\}=2z_{\overline{i}}^*(z_{\overline{j}}^*-z_{\overline{i}}^*),\quad i<j,$$

$$\{z_j,z_{\overline{i}}^*\}=2z_{\overline{i}}^*(z_j-z_{\overline{i}}^*),\quad i<j,$$

$$\{z_{\overline{j}}^*,z_i\}=2z_i(z_{\overline{j}}^*-z_i),\quad i<j.$$

Because z_i and z_i^* transform according to

$$z_i\to\frac{az_i+b}{cz_i+d},\quad z_i^*\to\frac{az_i^*+b}{cz_i^*+d},$$

we identify z_i and z_i^* with complex and conjugate complex coordinates of the complexified projective line.

Example 5: Now we explain how to put structures together to a new one. Let I' be an ordered set. Consider variables $t_i^{(j)}$, $i\in I^{(j)}$, $j\in I'$ (i.e., $I=\cup_{j\in I'}I^{(j)}$). Suppose that for fixed j the variables $t_i^{(j)}$, $i\in I^{(j)}$, carry an arbitrary Poisson structure. We combine these structures according to

$$\{t_{i'}^{(j')},t_i^{(j)}\}=2t_i^{(j)}(t_{i'}^{(j')}-t_i^{(j)}),\quad j<j'.$$

The verification uses the fact that Eq. (2) is satisfied, if two of the numbers λ_{ij}, λ_{ik}, and λ_{jk} are equal to 1.

Example 6: An example with complex λ_{ij} is given by $I=\{1,2,3,4\}$ and $\lambda_{12}=\lambda_{23}=\lambda_{34}=i$, $\lambda_{14}=-i$, and $\lambda_{13}=\lambda_{24}=0$.

III. QUANTIZATION OF POINT SETS OF THE REAL AND COMPLEX PROJECTIVE LINE

A. A class of $U_q(sl_2,\mathbb{R})$-module algebras

Let $U_q(sl_2,\mathbb{R})$, $|q|=1$, $q\neq\pm 1,\pm i$, be the unital *-algebra, given by generators E, F, K, K^{-1} and relations

$$KK^{-1}=K^{-1}K=I,\quad KE=qEK,\quad KF=q^{-1}FK,$$

$$[E,F]=\frac{K^2-K^{-2}}{q-q^{-1}},$$

and the involution

$$E^*=F, \quad F^*=E, \quad K^*=K^{-1}.$$

For an arbitrary set I consider the unital complex algebra A which is freely generated by the elements x_i, $i\in I$.

An action of $\mathrm{U}_q(\mathrm{sl}(2,\mathbb{R}))$ on A is given, if we set

$$E(1)=F(1)=0, \quad K(1)=1, \quad E(x_i)=1, \quad F(x_i)=-x_i^2, \quad K(x_i)=q^{-1}x_i, \tag{3}$$

and require for $x,y\in A$

$$K(xy)=K(x)K(y), \tag{4}$$

$$E(xy)=E(x)K(y)+K^{-1}(x)E(y), \tag{5}$$

$$F(xy)=F(x)K(y)+K^{-1}(x)F(y). \tag{6}$$

A proof can be given along the lines of Ref. 3, where a similar module of two variables occurs.

Further we consider ideals I of A, which are generated by all (nonvanishing) elements of the form

$$X_{ij}:=a_{ij}x_ix_j+b_{ij}x_i^2+c_{ij}x_j^2+d_{ij}x_jx_i, \tag{7}$$

where $a_{ij}=d_{ji}$ and $b_{ij}=c_{ji}$ (i.e., the elements are invariant with respect to a change of i and j). The action of $\mathrm{U}_q(\mathrm{sl}(2,\mathbb{R}))$ on A induces an action on the factor algebra A/I if and only if I is $\mathrm{U}_q(\mathrm{sl}(2,\mathbb{R}))$ invariant. We denote the image of the quotient map of x_i by v_i.

Proposition 1: The ideal I is $\mathrm{U}_q(\mathrm{sl}(2,\mathbb{R}))$ *invariant if and only if*

$$a_{ij}q^2+b_{ij}(1+q^2)+d_{ij}=0$$

and

$$a_{ij}+c_{ij}(1+q^2)+d_{ij}q^2=0. \tag{8}$$

The ideal I is generated by the elements

$$I_{ij}^{\lambda}=[x_j,x_i]-\frac{q^2-1}{q^2+1}\,(x_j^2-x_i^2-\lambda_{ij}(x_i-x_j)^2),$$

with

$$\lambda_{ij}=\frac{q^2+1}{q^2-1}\,\frac{a_{ij}+d_{ij}}{a_{ij}-d_{ij}}, \quad \lambda_{ij}=-\lambda_{ji}, \quad i,j\in I,$$

for $a_{ij}\neq d_{ij}$ and

$$I_{ij}^{\infty}=(x_i-x_j)^2,$$

for $a_{ij}=d_{ij}$.

Proof: First we consider the case $a_{ij}\neq d_{ij}$. Using (3)–(6), we have

$$\begin{aligned}E(X_{ij}):&=E(a_{ij}x_ix_j+b_{ij}x_i^2+c_{ij}x_j^2+d_{ij}x_jx_i)\\&=\left(a_{ij}q+b_{ij}\left(q+\frac{1}{q}\right)+d_{ij}\,\frac{1}{q}\right)x_i+\left(a_{ij}\,\frac{1}{q}+c_{ij}\left(q+\frac{1}{q}\right)+d_{ij}q\right)x_j.\end{aligned}$$

Because there are no linear relations between the generators, the last expression is contained in I, if and only if it is zero. Therefore the equations (8) are necessary.

Now we show that the equations (8) are also sufficient. We have to show that $K(X_{ij})$, $F(X_{ij}) \in I$. We have

$$K(X_{ij}) := q^{-2} X_{ij} \in I.$$

Further

$$F(X_{ij}) := -\frac{a_{ij}}{q} x_i^2 x_j - a_{ij} q x_i x_j^2 - \frac{b_{ij}}{q} x_i^3 - b_{ij} q x_i^3 - \frac{c_{ij}}{q} x_j^3 - c_{ij} q x_j^3 - \frac{d_{ij}}{q} x_j^2 x_i - d_{ij} q x_i x_j^2.$$

Now let

$$\alpha := \frac{1+q^2}{q} \frac{a_{ij}}{d_{ij} - a_{ij}}, \quad \beta := \frac{1+q^2}{q} \frac{d_{ij}}{a_{ij} - d_{ij}}.$$

Because of (8) we have

$$\alpha(a_{ij} + b_{ij}) = -\frac{a_{ij}}{q}, \quad \alpha(a_{ij} + c_{ij}) = -a_{ij} q, \quad \beta(c_{ij} + d_{ij}) = -\frac{d_{ij}}{q},$$

$$\beta(b_{ij} + d_{ij}) = -d_{ij} q, \quad (\alpha + \beta) = -\left(q + \frac{1}{q} \right), \quad (d_{ij}\alpha + a_{ij}\beta) = 0.$$

Therefore

$$\begin{aligned} F(X_{ij}) &= \alpha(a_{ij} + b_{ij}) x_i^2 x_j + \alpha(a_{ij} + c_{ij}) x_i x_j^2 + \beta(c_{ij} + d_{ij}) x_j^2 x_i + \beta(b_{ij} + d_{ij}) x_j x_i^2 + (\alpha + \beta) b_{ij} x_i^3 \\ &\quad + (\alpha + \beta) c_{ij} x_j^3 + (d_{ij}\alpha + a_{ij}\beta) x_i x_j x_i + (d_{ij}\alpha + a_{ij}\beta) x_j x_i x_j \\ &= \alpha(x_i X_{ij} + X_{ij} x_j) + \beta(x_j X_{ij} + X_{ij} x_i) \in I, \end{aligned}$$

i.e., the equations (8) are sufficient.

Further we confirm the second statement of the proposition. We show that the generators X_{ij} and I_{ij}^{λ} differ only by a constant nonzero factor:

$$\begin{aligned} I_{ij}^{\lambda} &= \left(-1 - \lambda_{ij} \frac{q^2-1}{q^2+1} \right) x_i x_j + (1 + \lambda_{ij}) \frac{q^2-1}{q^2+1} x_i^2 + (-1 + \lambda_{ij}) \frac{q^2-1}{q^2+1} x_j^2 + \left(1 - \lambda_{ij} \frac{q^2-1}{q^2+1} \right) x_j x_i \\ &= \frac{-2}{a_{ij} - d_{ij}} \left(a_{ij} x_i x_j + \frac{-a_{ij} q^2 - d_{ij}}{1+q^2} x_i^2 + \frac{-a_{ij} - d_{ij} q^2}{1+q^2} x_j^2 + d_{ij} x_j x_i \right) \\ &= \frac{-2}{a_{ij} - d_{ij}} (a_{ij} x_i x_j + b_{ij} x_i^2 + c_{ij} x_j^2 + d_{ij} x_j x_i). \end{aligned}$$

The second statement of the proposition follows.

Finally, let $a_{ij} = d_{ij} \neq 0$. We obtain $b_{ij} = c_{ij} = -a_{ij} = -d_{ij}$. The proposition follows.

We exclude the case I_{ij}^{∞} from the following considerations. Proposition 1 admits the following Definition.

Definition 1: Let I be an arbitrary set. A'_{λ}, $\lambda = (\lambda_{ij}; i, j \in I)$ is the left $U_q(sl_2, \mathbb{R})$-module algebra given by generators v_i, $i \in I$, and relations

$$[v_j, v_i] = \frac{q^2-1}{q^2+1} (v_j^2 - v_i^2 - \lambda_{ij} (v_i - v_j)^2). \tag{9}$$

Let λ_{ij} be fixed. In the quasi-classical limit $q \to 1$ the bracket (9) becomes the pre-Poisson bracket (1). By Proposition 1 the modules A'_{λ} are the most general $U_q(sl(2, \mathbb{R}))$-modules given by generators and quadratic relations.

In order to define cross ratios (see below) we need also elements of the form $(\Sigma \alpha_i v_i)^{-1}$.

Definition 2: We obtain the extension A_λ of A'_λ if we adjoin the inverses of the nonvanishing finite sums $(\Sigma \alpha_i v_i)^{-1}$ to A_λ and require that

$$\left(\sum \alpha_i v_i\right)^{-1}\left(\sum \alpha_i v_i\right)=\left(\sum \alpha_i v_i\right)\left(\sum \alpha_i v_i\right)^{-1}=1. \tag{10}$$

The $U_q(sl_2,\mathbb{R})$-module structure extends by the requirements (2)–(4) to all $x,y\in A_\lambda$.

B. Polynomials of degree three

There is a principle about a correspondence between the Jacobi identity for Poisson brackets and the Poincaré–Birkhoff–Witt (PBW) theorem for the quantized algebras. We will prove the statement of the PBW theorem only for polynomials of degree three. We obtain that condition (2) for the Jacobi identity characterizes also the classical behavior of third-order polynomials [with the exception of two special values for $\lambda_{i,j}$ and one structure where (2) is not satisfied]. The main part of the proof is Lemma 2.

Remark: Let b_{ij}, $c_{ij}\neq 0$ (i.e., $|\lambda_{ij}|\neq 1$). Then on the right side in relation (12) occurs a monomial which is lexicographically smaller and a monomial which is lexicographically greater than the monomial on the left side. Therefore we can not use the Bergman theory (cf. Ref. 4) for a proof of the PBW theorem.

Let $i<j$ and suppose $d_{ij}\neq 0$ [or the equivalent condition $\lambda_{i,j}\neq(q^2+1)/(q^2-1)$]. It follows

$$v_j v_i := \alpha_{ij} v_i v_j+\beta_{ij} v_i^2+\gamma_{ij} v_j^2 \tag{11}$$

with

$$\alpha_{ij}:\frac{-a_{ij}}{d_{ij}}=\frac{-(q^2+1)-\lambda_{ij}(q^2-1)}{\lambda_{ij}(q^2-1)-(q^2+1)},$$

$$\beta_{ij}:=\frac{-b_{ij}}{d_{ij}}=\frac{1-q^2\alpha_{ij}}{1+q^2}=\frac{(q^2-1)(1+\lambda_{ij})}{\lambda_{ij}(q^2-1)-(q^2+1)},$$

$$\gamma_{ij}:=\frac{-c_{ij}}{d_{ij}}=\frac{q^2-\alpha_{ij}}{1+q^2}=\frac{(q^2-1)(-1+\lambda_{ij})}{\lambda_{ij}(q^2-1)-(q^2+1)}$$

[cf. formulas (7)–(9)].

Lemma 1: (i) *Let* $i<j$, $\lambda_{i,j}\neq(q^2+1)/(q^2-1)$, and v_i and v_j *be two of the generators of* A'_λ. *Then*

$$(1-\beta_{ij}\gamma_{ij})v_j^2 v_i=(\beta_{ij}^2(1+\alpha_{ij}))v_i^3+(\alpha_{ij}\beta_{ij}(1+\alpha_{ij}))v_i^2 v_j+(\alpha_{ij}^2+\alpha_{ij}\beta_{ij}\gamma_{ij})v_i v_j^2 + (\gamma_{ij}(1+\alpha_{ij}))v_j^3.$$

We have $1-\beta_{ij}\gamma_{ij}\neq 0$ *for the coefficient on the left side if and only if* $\lambda_{i,j}\neq(q^4+1)/(q^4-1)$.

(ii) *Let* $1-\beta_{ij}\gamma_{ij}\neq 0$. *Then we can reduce every monomial of degree three into a sum of ordered monomials.*

Proof: (i) We have

$$\begin{aligned} v_j v_j v_i &= \alpha_{ij} v_j v_i v_j+\beta_{ij} v_j v_i^2+\gamma_{ij} v_j^3 \\ &= \alpha_{ij} v_j v_i v_j+(\beta_{ij}\alpha_{ij} v_i v_j v_i+\beta_{ij}\beta_{ij} v_i^3+\gamma_{ij}\beta_{ij} v_j^2 v_i)+\gamma_{ij} v_j^3 \\ &= (\alpha_{ij}\alpha_{ij} v_i v_j^2+\alpha_{ij}\beta_{ij} v_i^2 v_j+\alpha_{ij}\gamma_{ij} v_j^3)+(\beta_{ij}\alpha_{ij}\alpha_{ij} v_i^2 v_j \\ &\quad +\beta_{ij}\alpha_{ij}\beta_{ij} v_i^3+\beta_{ij}\alpha_{ij}\gamma_{ij} v_i v_j^2)+\beta_{ij}\beta_{ij} v_i^3+\gamma_{ij}\beta_{ij} v_j^2 v_i+\gamma_{ij} v_j^3. \end{aligned}$$

The identity follows. Further

$$(1-\beta_{ij}\gamma_{ij})=1-\frac{(q^2-1)(1+\lambda_{ij})}{(\lambda_{ij}(q^2-1)-(q^2+1))}\frac{(q^2-1)(-1+\lambda_{ij})}{(\lambda_{ij}(q^2-1)-(q^2+1))}=\frac{2(q^4+1)-2\lambda_{ij}(q^4-1)}{(\lambda_{ij}(q^2-1)-(q^2+1))^2}.$$

Therefore $1-\beta_{ij}\gamma_{ij}=0$ if and only if $\lambda_{i,j}=(q^4+1)/(q^4-1)$.

(ii) Using (11) we can replace every nonordered monomial by a sum of lexicographically smaller monomials or monomials of the type $v_jv_jv_i$. Because of Lemma 1(i) we can reduce $v_jv_jv_i$, $i<j$, to ordered monomials. Therefore we can reduce every polynomial of degree three with finite steps to a sum of ordered polynomials. ■

Lemma 2: (i) *Let $i<j<k$ be three indices, x_i,x_j,x_k, be three of the generators of A, $X_{ij}\in I$ [cf. (7)], $V\subset I$ be the subspace of all elements x_aX_{ab}, x_bX_{ab}, $X_{ab}x_a$, $X_{ab}x_b$, and*

$$X_{ijk}:=-a_{jk}a_{ik}X_{ij}x_k-a_{jk}x_jX_{ik}-X_{jk}x_i+x_kX_{ij}+a_{ij}X_{ik}x_j+a_{ik}a_{ij}x_iX_{jk}\in I.$$

Then we have

$$\begin{aligned}X_{ijk}=&\frac{(\lambda_{ij}\lambda_{jk}-\lambda_{ij}\lambda_{ik}-\lambda_{ik}\lambda_{jk}+1)(q^2-1)^2(q^2+1)}{k_{ij}k_{ik}k_{jk}}\\&\times(p_1x_i^3+p_2x_j^3+p_3x_k^3+p_4x_i^2x_j+p_5x_ix_j^2+p_6x_i^2x_k\\&+p_7x_ix_k^2+p_8x_j^2x_k+p_9x_jx_k^2)\mathrm{mod}(V)\end{aligned}\tag{12}$$

with

$$p_1:=\frac{4(\lambda_{ij}-\lambda_{ik})q^6(q^2-1)}{l_{ij}l_{ik}},\quad p_2:=\frac{-2(q^2-1)K}{l_{ij}l_{jk}},$$

$$p_3:=\frac{4(\lambda_{ik}-\lambda_{jk})(q^2-1)}{l_{ik}l_{jk}},$$

$$p_4:=\frac{-2m_{ij}q^2}{l_{ij}},\quad p_5:=\frac{2m_{ij}}{l_{ij}},\quad p_6:=\frac{2m_{ik}q^2}{l_{ik}},$$

$$p_7:=\frac{-2m_{ik}}{l_{ik}},\quad p_8:=\frac{-2m_{jk}q^2}{l_{jk}},\quad p_9:=\frac{2m_{jk}}{l_{jk}},$$

and

$$k_{ij}:=1+q^2+\lambda_{ij}(1-q^2),\quad l_{ij}:=1+q^4+\lambda_{ij}(1-q^4),$$

$$m_{ij}:=1+q^2+\lambda_{ij}(-1+q^2),$$

$$\begin{aligned}K:=&-(1+q^2+q^4+q^6)+\lambda_{12}(1-q^2+q^4+q^6)+\lambda_{23}(-1-q^2+q^4-q^6)\\&+\lambda_{12}\lambda_{23}(1-q^2-q^4+q^6).\end{aligned}$$

The right side of (12) vanishes if and only if $\lambda_{ij}\lambda_{jk}-\lambda_{ij}\lambda_{ik}-\lambda_{ik}\lambda_{jk}=-1$ or $\lambda_{ij}=(1+q^2)/(1-q^2)$, $\forall i<j$.

Proof: One can check that in X_{ijk} all terms $x_ax_bx_c$ with three different a,b,c cancel. The remaining monomials have only two different indices. Because we used in the reduction process of $v_av_bv_b$, $v_av_av_b$, and $v_av_bv_a$ in the proof of Lemma 1.1 only reduction rules which correspond to x_aX_{ab}, x_bX_{ab}, $X_{ab}x_a$, and $X_{ab}x_b$, we can replace the remaining monomials of type $x_ax_bx_b$, $x_ax_ax_b$, and $x_ax_bx_a$ by sums of ordered polynomials mod (V). This was done, using a computer calculation.

The right side of (12) vanishes if and only if (2) is satisfied or $p_1,\ldots,p_9=0$. From $p_1=p_3=p_4=0$ follows $\lambda_{ij}=\lambda_{ik}=\lambda_{jk}=(1+q^2)/(1-q^2)$. It follows $p_5,\ldots,p_9=0$ and $p_2=0$ because of

$$(1-q^2)^2K=-(1-q^2)^2(1+q^2+q^4+q^6)+(1-q^2)(1+q^2)(-2q^2+2q^4)$$
$$+(1+q^2)^2(1-q^2-q^4+q^6)=0,$$

i.e., $p_1,\ldots,p_9=0$ if and only if $\lambda_{ij}=(1+q^2)/(1-q^2)$. ■

Theorem 1: *Let A'_λ be the algebra of n points. We can represent every polynomial of degree less than three as a unique sum of ordered polynomials v_i, v_iv_j, and $v_iv_jv_k$, $i\leqslant j\leqslant k$, if and only if every three parameters $\lambda_{ij},\lambda_{ik},\lambda_{jk}, i<j<k$ satisfy the conditions (2):*

$$\lambda_{ij}\lambda_{jk}-\lambda_{ij}\lambda_{ik}-\lambda_{ik}\lambda_{jk}=-1$$

and the additional conditions $\lambda_{ij}\neq(q^2+1)/(q^2-1),(q^4+1)/(q^4-1)$, $i<j$, or if

$$\lambda_{ij}=\lambda_{ik}=\lambda_{jk}=\frac{1+q^2}{1-q^2}\,.$$

Proof: (i) Let $\lambda_{ij}=(q^2+1)/(q^2-1)$ (i.e., $d_{ij}=0$). Because of (7) we obtain the linear dependence of ordered monomials $a_{ij}v_iv_j+b_{ij}v_i^2+c_{ij}v_j^2=0$.

Let $\lambda_{ij}=(q^4+1)/(q^4-1)$. Because of Lemma 1 we obtain a linear dependence of ordered monomials.

Finally let i,j,k be three indices, such that (2) is not satisfied and let $\lambda_{ij}\neq(1+q^2)/(1-q^2)$ for $i<j$. If we apply the quotient map $A\to A/I$ to (12) we obtain a linear dependence between ordered monomials.

Therefore the conditions of Theorem 1 are necessary.

(ii) Now let the conditions of the Theorem be satisfied. Because the relations (9) are homogeneous, A'_λ is the direct sum of the subspaces of homogeneous polynomials.

Because of $\lambda_{ij}\neq(q^2+1)/(q^2-1)$ (i.e., $d_{ij}\neq0$) we can represent every polynomial of degree $\leqslant2$ as a sum of ordered monomials.

It remains to consider monomials of degree three. Because of $\lambda_{ij}\neq(q^4+1)/(q^4-1)$ and Lemma 1(ii) we can reduce every monomial of degree three in finite steps to a sum of ordered monomials.

We show that the ordered monomials are linear independent. The space of third degree polynomials in the free algebra A is n^3-dimensional. The Ideal I is generated by the $2n\binom{n}{2}$ elements x_iX_{jk} and $X_{jk}x_i$, $j<k$. Because (2) is satisfied for $i<j<k$ or $\lambda_{ij}=(1+q^2)/(1-q^2)$ for $i<j$, the left side of (12) in Lemma 2 vanishes mod (V). Therefore we have $\binom{n}{3}$ relations between the generators x_iX_{jk} and $X_{jk}x_i$ of I. These relations are independent, because elements x_iX_{jk} and $X_{jk}x_i$ with three fixed different indices occur only in one of the relations. Therefore $\dim I\leqslant2n\binom{n}{2}-\binom{n}{3}$. It follows that the dimension of the space of third degree polynomials in the factor algebra $A'_\lambda=A/I$ is greater than or equal $n^3-2n\binom{n}{2}+\binom{n}{3}=n^2+\binom{n}{3}$. We have shown above that this space is spanned by ordered polynomials, i.e., the dimension of this space space is less than or equal $n+2\binom{n}{2}+\binom{n}{3}=n^2+\binom{n}{3}$. Therefore the ordered monomials are linear independent. ■

Remark: If (2) is satisfied and $\lambda_{ij}=(q^2+1)/(q^2-1)$ or $(q^4+1)/(q^4-1)$ for certain $i<j$ one can change the order of the indices in order to satisfy the conditions of Theorem 1.

Remark: We expect an analogue of Theorem 1 for polynomials of higher degrees. One difficulty is to classify the additional conditions. For example, for fourth-order polynomials one can derive the additional condition $\beta_{ij}\gamma_{ij}(1+\alpha_{ij})\neq1$.

Remark: If we introduce the new parameters $c:=i(q^2+1)/(q^2-1)$ and $\phi_{ij}:=c/\lambda_{ij}$, $(\lambda_{ij}\neq0)$ we have

$$[v_j,v_i]=\frac{i}{c}\,(v_j^2-v_i^2)-\frac{i}{\phi_{ij}}\,(v_i-v_j)^2.$$

For condition (2) we get the form

$$\phi_{ik}=\frac{\phi_{ij}+\phi_{jk}}{1+\phi_{ij}\phi_{jk}/c^2}$$

of the Einstein addition theorem for velocities.

In subsections C and D we give some examples.

C. Points of the real projective line

Consider A_λ with real λ_{ij} and endow A_λ with the trivial involution $v_i^*=v_i$. The relations (9) are obviously $*$-invariant. Therefore the involution on A induces an involution on A'_λ and A_λ.

In view of Theorem 1 and the Remark after Theorem 1 we give the following definition.

Definition 3: The algebra $A_\lambda^{\mathbb{R}}$ of coordinates of noncommutative points v_i of the real line is the $U_q(\mathrm{sl}(2,\mathbb{R}))$ module A_λ, if A_λ carries the involution $*$ and A'_λ has the property (2) or is the exceptional structure of Theorem 1.

Example 7: Let I be an arbitrary linear-ordered set and let $\lambda_{ij}=1$ for $i<j$ (cf. Example 1). It follows

$$[v_j,v_i]=(q^2-1)v_i(v_j-v_i),\quad i<j.$$

We denote this structure by $A_1^{\mathbb{R}}$.

Example 8: Let I be the set $\mathbb{Z}$ of integers and $\lambda_{ij}=\coth(\Sigma_{k=i}^{j-1}\alpha_k)$, $\alpha_k>0$, $i<j$ (cf. Example 2). It follows

$$[v_j,v_i]=\frac{q^2-1}{q^2+1}\left(v_j^2-v_i^2-\coth\left(\sum_{k=i}^{j-1}\alpha_k\right)(v_i-v_j)^2\right),\quad i<j.$$

Example 9: Let I be the set of integers $\lambda_{ij}=(1+q^2)/(1-q^2)$, $i<j$ (cf. Theorem 1). It follows

$$v_jv_i=\frac{v_i^2+q^2v_j^2}{1+q^2}.$$

We denote this structure by $A_E^{\mathbb{R}}$. Because of $\lambda_{ij}\to\infty$ for $q\to1$ this structure has no quasi-classical limit.

D. Points of the complexified projective line

Let I' be an arbitrary set and $I=I'\times\mathbb{Z}_2$. We use the notations $w_i:=v_{(i,0)}$ and $w_{\overline{i}}^*:=v_{(i,1)}$, $i\in I'$. Let w_i and $w_{\overline{i}}^*$, $i\in I'$, be connected by the involution $*$ (i.e., $w_{\overline{i}}^{**}=w_i$). Then the relations (9) get the form

$$[w_j,w_i]=\frac{q^2-1}{q^2+1}\,(w_j^2-w_i^2-\lambda_{ij}(w_i-w_j)^2),$$

$$[w_j,w_{\overline{i}}^*]=\frac{q^2-1}{q^2+1}\,(w_j^2-w_{\overline{i}}^{*2}-\lambda_{\overline{ij}}(w_{\overline{i}}^*-w_j)^2),$$

$$[w_{\overline{j}}^*,w_{\overline{i}}^*]=\frac{q^2-1}{q^2+1}\,(w_{\overline{j}}^{*2}-w_{\overline{i}}^{*2}-\lambda_{\overline{ij}}(w_{\overline{i}}^*-w_{\overline{j}}^*)^2).$$

Lemma 3: The above relations are $*$-invariant (and therefore induce an involution on A'_λ), if and only if $\overline{\lambda_{ij}}=\lambda_{\overline{ij}}$ and $\overline{\lambda_{\overline{ij}}}=\lambda_{\overline{ij}}$.

We denote the algebra generated by elements w_i and w_i^* and the above relations with $\overline{\lambda_{ij}}=\lambda_{\overline{ij}}$ and $\overline{\lambda_{\overline{ij}}}=\lambda_{\overline{ij}}$ by A_λ^c.

In order to check (2) for A_λ^c it is sufficient to check $\lambda_{\alpha\beta}\lambda_{\beta\gamma}-\lambda_{\alpha\beta}\lambda_{\alpha\gamma}-\lambda_{\alpha\gamma}\lambda_{\beta\gamma}=-1$ for $(\alpha,\beta,\gamma)=(i,j,k),(i,j,k),(i,i,j)$, $\forall i<j<k$. [For real coefficients one has to check the case $(\alpha,\beta,\gamma)=(i,j,k)$ and $\lambda_{\overline{ij}}=\lambda_{ij},\lambda_{i\overline{j}}=\lambda_{\overline{i}j}=\lambda_{ij}^{-1}$ (i.e., $\lambda_{i\overline{i}}=0$).]

Definition 4: By the algebra $A_\lambda^{\mathbb{C}}$ of quantized complex points w_i of the complexified projective line we denote the $U_q(\mathrm{sl}_2,\mathbb{R})$ module A_λ^c, if A_λ^c satisfies (2) and A_λ^c carries the involution $*$.

We consider the elements w_i and $w_{\bar{i}}^*$, respectively, as quantized complex and quantized complex conjugated points, respectively.

Example 10: Let I' be a linear ordered set, $I=I'\times\mathbb{Z}_2$, and let λ_{ij} be defined as in Example 4. We obtain

$$[w_{\bar{i}}^*,w_i]=\frac{q^2-1}{q^2+1}\,(w_{\bar{i}}^{*2}-w_i^2),$$

$$[w_j,w_i]=(q^2-1)w_i(w_j-w_i),\quad i<j,$$

$$[w_{\bar{j}}^*,w_{\bar{i}}^*]=(q^2-1)w_{\bar{i}}^*(w_{\bar{j}}^*-w_{\bar{i}}^*),\quad i<j,$$

$$[w_j,w_{\bar{i}}^*]=(q^2-1)w_{\bar{i}}^*(w_j-w_{\bar{i}}^*),\quad i<j,$$

$$[w_{\bar{j}}^*,w_i]=(q^2-1)w_i(w_{\bar{j}}^*-w_i),\quad i<j.$$

We denote this structure by $A_1^{\mathbb{C}}$

Example 11: Let $I'=\mathbb{Z}$, $I=I'\times\mathbb{Z}_2$, and $\lambda_{ij}=\coth(\Sigma_{k=i}^{j-1}\alpha_k)$, $\alpha_k>0$, $i<j$. It follows

$$[w_{\bar{i}}^*,w_i]=\frac{q^2-1}{q^2+1}\,(w_{\bar{i}}^{2*}-w_i^2),$$

$$[w_j,w_i]=\frac{q^2-1}{q^2+1}\left(w_j^2-w_i^2-\coth\left(\sum_{k=i}^{j-1}\alpha_k\right)(w_i-w_j)^2\right),\quad i<j,$$

$$[w_j,w_{\bar{i}}^*]=\frac{q^2-1}{q^2+1}\left(w_j^2-w_{\bar{i}}^{*2}-\tanh\left(\sum_{k=i}^{j-1}\alpha_k\right)(w_{\bar{i}}^*-w_j)^2\right),\quad i<j,$$

$$[w_{\bar{j}}^*,w_i]=\frac{q^2-1}{q^2+1}\left(w_{\bar{j}}^{*2}-w_i^2-\tanh\left(\sum_{k=i}^{j-1}\alpha_k\right)(w_i-w_{\bar{j}}^*)^2\right),\quad i<j,$$

$$[w_{\bar{j}}^*,w_{\bar{i}}^*]=\frac{q^2-1}{q^2+1}\left(w_{\bar{j}}^{*2}-w_{\bar{i}}^{*2}-\coth\left(\sum_{k=i}^{j-1}\alpha_k\right)(w_{\bar{i}}^*-w_{\bar{j}}^*)^2\right),\quad i<j.$$

Example 10 arises again in the limit $\alpha_k\to\infty$.

Example 12: An example with complex coefficients is given by $I=\{1,2,\overline{1,2}\}$ and $\lambda_{12}=\lambda_{2\bar{2}}=\lambda_{\bar{2}\bar{1}}=\lambda_{\bar{1}\bar{1}}=i$ and $\lambda_{1\bar{2}}=\lambda_{\bar{1}\bar{2}}=O$ (cf. Example 6).

E. Algebras of three points

We will say that $A_\lambda{}'$ is polynomial, if the subspaces of homogeneous polynomials have the classical dimensions.

From Theorem 1 and the Remark after Theorem 1 follows that the subspaces of homogeneous polynomials of degree $\leqslant 3$ of the algebras $A_\lambda{}'$ which satisfy (2) and of the exceptional algebra $A_E^{\mathbb{R}}$ have the classical dimensions.

We expect that these algebras are polynomial. We give a proof for algebras of three points.

Lemma 4: Let $A_\lambda{}'$ be the algebra with three generators v_1,v_2,v_3 and let (2) be satisfied. $A_\lambda{}'$ is equivalent to the algebra with generators u_1,u_2,u_3 and relations

$$u_2u_1=\frac{1}{q^2}\,u_1u_2,$$

$$u_3u_1=\frac{1}{q^2}\,u_1u_3+\left(\frac{1}{q^2}-1\right)(\lambda_{1,2}-1)(\lambda_{2,3}+1)u_2^2, \tag{13}$$

$$u_3u_2=\frac{1}{q^2}\,u_2u_3.$$

The equivalence is given by a linear transformation between the generators u_i and v_i.

Proof: Consider the transformation

$$\begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix}=\begin{pmatrix} \lambda_{1,2}-1 & -(\lambda_{1,2}+\lambda_{2,3}) & \lambda_{2,3}+1\\ -(\lambda_{1,2}+1) & \lambda_{1,2}+\lambda_{2,3} & -(\lambda_{2,3}-1)\\ (\lambda_{1,2}+1)^2(\lambda_{2,3}+1) & -(\lambda_{1,2}+\lambda_{2,3})(\lambda_{1,2}-1)(\lambda_{2,3}+1) & (\lambda_{2,3}-1)^2(\lambda_{1,2}-1)\end{pmatrix}\times\begin{pmatrix} v_1\\ v_2\\ v_3\end{pmatrix}. \tag{14}$$

We denote the matrix by (k_{ij}). For the determinant one obtains

$$\det((k_{ij}))=-8(\lambda_{1,2}+\lambda_{2,3})^2.$$

[If $\det((k_{ij}))=0$, there exists a permutation of the three indices such that $\det((k_{ij}))\neq 0$.]

Let

$$\begin{aligned} A_{12}&:=2u_1u_2-2q^2u_2u_1,\\ A_{13}&:=2u_1u_3-2q^2u_3u_1+(1-q^2)(\lambda_{1,2}-1)(\lambda_{2,3}+1)u_2^2,\\ A_{23}&:=2u_2u_3-2q^2u_3u_2.\end{aligned} \tag{15}$$

Because of (13) we have $A_{12}=A_{13}=A_{23}=0$. We form the vanishing expressions

$$\begin{aligned} B_{12}&:=(k_{33}A_{12}-k_{23}A_{13}+k_{13}A_{23})/\det(K),\\ B_{13}&:=(-k_{32}A_{12}+k_{22}A_{13}-k_{12}A_{23})/\det(K),\\ B_{23}&:=(k_{31}A_{12}-k_{21}A_{13}+k_{11}A_{23})/\det(K).\end{aligned} \tag{16}$$

If we insert (15) and (14) in B_{12} by a computer calculation, we obtain

$$B_{12}=(1-q^2)(\lambda_{12}+1)v_1^2+(1-q^2)(\lambda_{12}-1)v_2^2+(q^2(\lambda_{12}+1)+(-\lambda_{12}+1))v_1v_2+(q^2(\lambda_{12}-1)+(-\lambda_{12}-1))v_2v_1.$$

We transform this expression in

$$B_{12}=-(1+q^2)(v_2v_1-v_1v_2)+(1-q^2)(v_1^2-v_2^2+\lambda_{12}(v_1-v_2)^2).$$

Similarly we obtain

$$B_{13}=-(1+q^2)(v_3v_1-v_1v_3)+(1-q^2)\left(v_1^2-v_3^2+\frac{1+\lambda_{12}\lambda_{23}}{\lambda_{12}+\lambda_{23}}\,(v_1-v_3)^2\right),$$

$$B_{23}=-(1+q^2)(v_3v_2-v_2v_3)+(1-q^2)(v_2^2-v_3^2+\lambda_{13}(v_2-v_3)^2).$$

Because B_{12}, B_{13}, and B_{23} vanish, the formulas (9) follow. ■

Proposition 2: Let (2) be satisfied. Then $A_\lambda{}'$ is polynomial.

Proof: The transformation between the v_i and u_i does not change the dimensions of the subspaces of homogeneous polynomials. We have to show the polynomiality of the algebra (13).

By the Diamond lemma (cf. Ref. 4) this algebra has the PBW property (and is therefore ''polynomial'') if and only if the ''overlap''

$$(u_3u_2)u_1-u_3(u_2u_1)$$

gives zero when it is reduced by means of (13) to a linear combination of ordered monomials (cf. Ref. 4 for details). Let $a:=1/q^2$ and $\epsilon:=(1/q^2-1)(\lambda_{1,2}-1)(\lambda_{2,3}+1)$. We obtain

$$\begin{aligned}(u_3u_2)u_1-u_3(u_2u_1)&=au_2(u_3u_1)-a(u_3u_1)u_2\\&=a^2(u_2u_1)u_3+a\epsilon u_2^3-a^2u_1(u_3u_2)-a\epsilon u_2^3\\&=a^3u_1u_2u_3-a^3u_1u_2u_3=0.\end{aligned}$$ ■

Finally we consider the exceptional structure $A_E^{\mathbb{R}}$. It is given by

$$v_jv_i=\frac{v_i^2+q^2v_j^2}{1+q^2},\quad i<j \tag{17}$$

(cf. Example 9).

Proposition 3: The algebra $A_E^{\mathbb{R}}$ is polynomial.

Proof: Consider the transformation

$$\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}q^2&q^2&q^2\\1&q^2&q^2\\1&1&q^2\end{pmatrix}\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}. \tag{18}$$

We denote the matrix by (k_{ij}). We obtain for the determinant

$$\det((k_{ij}))=q^2(q^2-1)^2\neq 0.$$

Let

$$\begin{aligned}A_{12}&:=-(1+q^2)v_2v_1+v_1^2+q^2v_2^2,\\A_{13}&:=-(1+q^2)v_3v_1+v_1^2+q^2v_3^2,\\A_{23}&:=-(1+q^2)v_3v_2+v_2^2+q^2v_3^2.\end{aligned} \tag{19}$$

Because of (17) we have $A_{12}=A_{13}=A_{23}=0$. We form the vanishing expressions B_{12}, B_{13}, and B_{23} [cf. (16)] and insert (19) and (14). A calculation yields

$$\begin{aligned}B_{12}&=-(1+q^2)u_2u_1=0,\\B_{13}&=u_1u_2+u_2u_1+u_1u_3-q^2u_3u_1=0,\\B_{23}&=-u_1u_2-u_2u_1+u_2u_3-q^2u_3u_2=0.\end{aligned}$$

Therefore the algebra has generators u_1, u_2, and u_3 and relations

$$\begin{aligned}u_2u_1&:=0,\\u_3u_1&:=\frac{1}{q^2}(u_1u_3+u_1u_2),\\u_3u_2&:=\frac{1}{q^2}(u_2u_3-uu_2).\end{aligned} \tag{20}$$

The monomials on the left side are lexicographically greater than the monomials on the right side. Therefore it is sufficient to show that the ''overlap''

$$(u_3u_2)u_1-u_3(u_2u_1)$$

gives zero when it is reduced by means of (20) to a linear combination of ordered monomials (cf. Proof of Proposition 2). We have

$$(u_3u_2)u_1-u_3(u_2u_1)=\frac{1}{q^2}\,(u_2(u_3u_1)-u_1(u_2u_1))=\frac{1}{q^4}\,((u_2u_1)u_3+(u_2u_1)u_2)=0. \qquad \blacksquare$$

Remark: There is a close relation between the PBW property and the Jacobi identity. Consider an algebra with generators x, y, and z and relations $[x, y]=a$, $[z, x]=b$, and $[y, z]=c$, where a, b, and c are certain polynomials of x, y, and z. We obtain for zyx the two reductions

$$(zy)x=xyz-cx+yb-az$$

and

$$z(yx)=xyz-za+by-xc.$$

Therefore the vanishing of the ''overlap'' of the Diamond lemma has the form of the Jacobi identity, in the sense that

$$(zy)x-z(yx)=(xc-cx)+(yb-by)+(za-az)=0.$$

One can further show that the Jacobi identity for the three generators v_1, v_2, and v_3 with the commutator (9) is equivalent to the PBW property (cf. section 4.4 of Ref. 4).

F. Projective coordinates

As the algebra of *noncommutative projective coordinates* B_μ *for* A_λ we denote the $U_q(sl(2,\mathbb{R}))$-module algebra generated by elements x_i, y_i, y_i^{-1}, $(\sum \alpha_i x_i y_i^{-1})^{-1}$, $i, j \in I$, and relations

$$y_i y_i^{-1}=y_i^{-1}y_i=1,$$

$$\left(\sum\ \alpha_i x_i y_i^{-1}\right)^{-1}\left(\sum\ \alpha_i x_i y_i^{-1}\right)=\left(\sum\ \alpha_i x_i y_i^{-1}\right)\left(\sum\ \alpha_i x_i y_i^{-1}\right)^{-1}=1,$$

and

$$x_i y_i=q y_i x_i\,, \tag{21}$$

$$x_j x_i=\mu_{ij}^{(1)} x_i x_j\,, \tag{22}$$

$$y_j y_i=\mu_{ij}^{(1)} y_i x_j\,, \tag{23}$$

$$x_j y_i=\mu_{ij}^{(2)} y_i x_j+\left(\mu_{ij}^{(1)}-\frac{1}{q}\,\mu_{ij}^{(2)}\right)x_i y_j\,, \tag{24}$$

$$y_j x_i=\mu_{ij}^{(2)} x_i y_j+(\mu_{ij}^{(1)}-q\mu_{ij}^{(2)})y_i x_j \tag{25}$$

with

$$\mu_{ji}^{(1)}=\frac{1}{\mu_{ij}^{(1)}} \quad \text{and} \quad \mu_{ji}^{(2)}=1\Big/\left[\mu_{ij}^{(2)}\left(\frac{\mu_{ij}^{(1)}}{\mu_{ij}^{(2)}}\left(q+\frac{1}{q}-\frac{\mu_{ij}^{(1)}}{\mu_{ij}^{(2)}}\right)\right)\right].$$

The $U_q(sl(2,\mathbb{R}))$-module structure on B_μ is given by

$$E(x_i)=q^{1/2}y_i, \quad E(y_i)=0,$$

$$F(x_i)=0, \quad F(y_i)=q^{-1/2}x_i,$$

$$K(x_i)=q^{-1/2}x_i, \quad K(y_i)=q^{1/2}y_i.$$

and Eqs. (4)–(6).

Lemma 5: The elements $v_i:=q^{1/2}x_iy_i^{-1}$, $i\in I$, *satisfy (9) with*

$$\lambda_{ij}=\frac{q^2-2q(\mu_{ij}^{(1)}/\mu_{ij}^{(2)})+1}{1-q^2} \quad \text{or} \quad \mu_{ij}:=\frac{\mu_{ij}^{(1)}}{\mu_{ij}^{(2)}}=\frac{1+q^2+\lambda_{ij}(q^2-1)}{2q},$$

and $\mathrm{U}_q(\mathrm{sl}(2,\mathbb{R}))$ *acts on* v_i *according to (3).*

This can be checked by an explicit calculation. Therefore we can map A_λ onto a submodule of B_μ. We call x_i and y_i *noncommutative coordinates* of v_i.

Remark: One can show that the algebra of projective coordinates is *polynomial* if and only if $\mu_{ij}^{(1)}=q^{\pm 1}\mu_{ij}^{(2)}$, i.e., $\lambda_{ij}=\pm 1$. Therefore for $\lambda_{ij}\neq\pm 1$ the polynomiality of the algebra A'_λ does not yield the polynomiality of the algebra of projective coordinates.

IV. QUANTUM CROSS RATIOS

A. Invariant elements

In this section we investigate a class of invariant elements in A_λ.

Definition 5: We say that $x\in A_\lambda$ (or B_μ) is *invariant* if

$$E(x)=0, \quad F(x)=0, \quad K(x)=x.$$

Following Ref. 1 we define

$$(ij):=q^{-1/2}x_iy_j-q^{1/2}y_ix_j \quad \text{and} \quad [ij]:=v_i-v_j.$$

One can check that (ij) is invariant and we have

$$(ij)=y_i[ij]y_j. \tag{26}$$

Now we use the invariants $(ij)\in B_\mu$ to combine invariants of A_λ. We define

$$C_{ijkl}:=(il)(kl)^{-1}(kj)(ij)^{-1}.$$

Proposition 3: (i) $C_{ijkl}\in B_\mu$, $(i,j,k,l\in I)$ *is invariant.*

(ii) *We can represent* C_{ijkl} *by elements* v_i (i.e., $C_{ijkl}\in A_\lambda$) *and we have the formula*

$$\begin{aligned}C_{ijkl}:=&(q+q^{-1}-\mu_{il})(v_i-v_l)((q+q^{-1}-\mu_{ik})v_k-(q+q^{-1}-\mu_{il})v_l\\&+(\mu_{ik}-\mu_{il})v_i)^{-1}((q+q^{-1}-\mu_{ik})v_k-(q+q^{-1}-\mu_{ij})v_j\\&+(\mu_{ik}-\mu_{ij})v_i)(q+q^{-1}-\mu_{ij})^{-1}(v_i-v_j)^{-1}.\end{aligned} \tag{27}$$

Proof: The element $(ij)^{-1}$ and products of invariant elements are invariant because of (4)–(6). The assertion follows.

(ii) From (21)–(25) follow the commutation rules

$$y_iv_i=q^{-1}v_iy_i, \quad y_iv_j=((q+q^{-1}-\mu_{ij})v_j+(\mu_{ij}-q)v_i)y_i. \tag{28}$$

Further, because of (26) we have

$$C_{ijkl}=y_i[il][kl]^{-1}[kj][ij]^{-1}y_i^{-1}. \tag{29}$$

The assertion follows, if we move y_i to the right applying (28). ■

Proposition 4: In analogy to the classical case we have

$$C_{ilkj}=\frac{1}{C_{ijkl}},\quad C_{ijlk}=1-C_{ijkl},\quad C_{ikjl}=\frac{C_{ijkl}}{C_{ijkl}-1},$$

$$C_{iklj}=\frac{1}{1-C_{ijkl}},\quad C_{iljk}=1-\frac{1}{C_{ijkl}}.$$

Proof: One checks the equations $C_{ilkj}C_{ijkl}=1$ and $C_{ijlk}+C_{ijkl}=1$ using (29) and the rule $[ik]=[ij]+[jk]$. The five identities can combined from them. ■

Remark: In the general case, the elements C_{ijkl} are not $*$-invariant and the identities $C_{ijkl}=C_{lkji}=C_{klij}=C_{jilk}$ are in general not valid (see below).

B. The cross ratio

We restrict the consideration to the structure $A_1^{\mathbb{R}}$ of Example 7, i.e., $\lambda_{ij}=1$. [If $|\lambda_{ij}|=1$, we can reorder the indices such that $\lambda_{ij}=1$ (cf. Example 1).]

We can choose the algebra of projective coordinates with $\mu_{ij}^{(1)}=q^2$ and $\mu_{ij}^{(2)}=q$ (cf. Lemma 5), i.e.,

$$x_iy_i=qy_ix_i,\quad x_jx_i=q^2x_ix_j,\quad y_jy_i=q^2y_iy_j,$$

$$x_jy_i=qy_ix_j+(q^2-1)x_iy_j,\quad y_jx_i=qx_iy_j. \tag{30}$$

Lemma 6: Let $i<j<k<l$ be four ordered indices. Then we have

$$(kl)(ij)=q^6(ij)(kl),\quad (jk)(il)=(il)(jk),$$

$$(jl)(ik)=q^4(ik)(jl)+(q^6-q^4)(ij)(kl),$$

$$(jk)(ij)=q^4(ij)(jk),\quad (ik)(ij)=q^2(ij)(ik),$$

$$(jk)(ik)=q^2(ik)(jk),\quad (ij)^*=(ij).$$

Proof: We prove the fourth rule. Let $i<j<k$. From (30) follows

$$\begin{aligned}x_k(x_iy_j-qy_ix_j)&=q^2x_ix_ky_j-q^2y_ix_kx_j-(q^3-q)x_iy_kx_j\\&=q^3x_iy_jx_k+(q^4-q^2)x_ix_jy_k-q^4y_ix_jx_k\\&\quad-(q^4-q^2)x_ix_jy_k=q^3(x_iy_j-qy_ix_j)x_k,\end{aligned}$$

i.e., $x_k(ij)=q^3(ij)x_k$. Similarly, one derives $y_k(ij)=q^3(ij)y_k$, $x_j(ij)=q(ij)x_j$, and $y_j(ij)=q(ij)y_j$. The identity $(jk)(ij)=q^4(ij)(jk)$ follows. The proof of the remaining rules is similar. ■

In a similar situation some of the invariant elements C_{ijkl} were first considered in Ref. 1. Let $i<j<k<l$ be four elements of I in increasing order. With $C:=C_{ijkl}$ we derive using Proposition 4 and Lemma 6

$$C_{ijkl}=C_{klij}=C,\quad C_{jilk}=C_{lkji}=C^*=q^2C,$$

$$C_{ilkj}=C_{kjil}=\frac{1}{C},\quad C_{jkli}=C_{lijk}=C^*_{ilkj}=\frac{1}{q^2C}=\frac{c_{ilkj}}{q^2},$$

$$C_{ijlk}=C_{klji}=1-C,\quad C_{jikl}=C_{lkij}=C^*_{ijlk}=1-q^2C=q^2C_{ijlk}+(1-q^2),$$

$$C_{ikjl}=C_{kilj}=\frac{C}{C-1},\quad C_{jlik}=C_{ljki}=C^*_{ikjl}=\frac{q^2C}{q^2C-1}=\frac{q^2C_{ikjl}}{1+(q^2-1)C_{ikjl}},$$

$$C_{iklj}=C_{kijl}=\frac{1}{1-C},\quad C_{jlki}=C_{ljik}=C^*_{iklj}=\frac{1}{1-q^2C}=\frac{C_{iklj}}{q^2+(1-q^2)C_{iklj}},$$

$$C_{iljk}=C_{kjli}=1-\frac{1}{C},\quad C_{jkil}=C_{likj}=C^{*}_{iljk}=1-\frac{1}{q^2C}=\frac{1}{q^2}\,C_{iljk}+\left(1-\frac{1}{q^2}\right).$$

The Remark after Proposition 4 indicates that C_{ijkl} is not a good candidate of a quantum cross ratio. Now we make the following observation:

Lemma 7: Let i, j, k, l be four indices in arbitrary order and let $C^{*}_{ijkl}=:f(C_{ijkl},q^2)$, where f is a certain rational function. Then $f(C_{ijkl},q)$ is a $*$-invariant element of $A^{\mathbb{R}}_{\lambda}$.

This can be checked with the above formulas. Now we give the following definition.

Definition 6: We say that the $*$-invariant element $f(C_{ijkl},q)$ is the *quantum cross ratio* of the quadruple (v_i,v_j,v_k,v_l). We use the notation C_{ijkl}.

The elements C_{ijkl} have the desired properties. They tend to the classical cross ratio for $q\to 1$ and for $i<j<k<l$ we obtain

$$\mathscr{C}_{ijkl}=qC=q[il][kl]^{-1}[kj][ij]^{-1},$$

$$\mathscr{C}_{ilkj}=\frac{1}{C_{ijkl}}=\frac{1}{qC}=\frac{1}{q}\,C_{ilkj},$$

$$\mathscr{C}_{ijlk}=1-C_{ijkl}=1-qC=qC_{ijlk}+(1-q),$$

$$\mathscr{C}_{ikjl}=\frac{\mathscr{C}_{ijkl}}{\mathscr{C}_{ijkl}-1}=\frac{qC}{qC-1}=\frac{qC_{ikjl}}{1+(q-1)C_{ikjl}},$$

$$\mathscr{C}_{iklj}=\frac{1}{1-\mathscr{C}_{ijkl}}=\frac{1}{1-qC}=\frac{C_{iklj}}{q+(1-q)C_{iklj}},$$

$$\mathscr{C}_{iljk}=1-\frac{1}{\mathscr{C}_{ijkl}}=1-\frac{1}{qC}=\frac{1}{q}\,C_{iljk}+\left(1-\frac{1}{q}\right),$$

and

$$\mathscr{C}_{ijkl}=\mathscr{C}_{lkji}=\mathscr{C}_{klij}=\mathscr{C}_{jilk}.$$

C. The Euclidean distance

In the classical case a Euclidean distance on $P\mathbb{R}$ can be defined, if we fix the three points t_0 (zero-point), t_1 (one-point), and t_∞ (the point at infinity). The distance $d(t_j,t_i)$ of two points t_j, t_i is given by

$$d(t_j,t_i):=C_{0,1,\infty,j}-C_{0,1,\infty,i}.$$

This observation leads us to a notion of a noncommutative Euclidean distance. Fix three different generators v_0, v_1, and v_∞ of $A^{\mathbb{R}}_1$. Then the *noncommutative distance* of the generators v_i and v_j is given by the $*$-invariant element

$$D_{v_0,v_1,v_\infty}(v_j,v_i):=\mathscr{C}_{0,1,\infty,j}-\mathscr{C}_{0,1,\infty,i}.$$

[1] C.-S. Chu, P.-M. Ho, and B. Zumino, ''The Braided Quantum 2-Sphere,'' preprint q-alg 9507013, UCB-PTH-95/25.
[2] J.-H. Lu and A. Weinstein, J. Diff. Geom. **31**, 501 (1990).
[3] S. Klimek and A. Lesniewski, J. Funct. Anal. **115**, 1 (1993).
[4] G. M. Bergman, Adv. Math. **29**, 178 (1978).

Particles in singular magnetic field

Władysław Marcinek[a)]
Institute of Theoretical Physics, University of Wrocław,
Pl. Maksa Borna 9, 50-204 Wrocław, Poland

(Received 29 August 1997; accepted for publication 10 September 1997)

An algebraic formalism for description of quantum states of charged particle with spin moving in two-dimensional space under the influence of a singular magnetic field is developed in terms of graded algebras. The fundamental assumption is that the particle is transformed into a composite system which consists of quasiparticles, quasiholes, and magnetic fluxes. Such a system is endowed with generalized statistics determined by a grading group and a commutation factor on it. Composite systems corresponding to the quantum Hall effect and the electronic magnetotransport anomaly are described. The Fock space representation is also given. © *1998 American Institute of Physics.* [S0022-2488(98)01902-1]

I. INTRODUCTION

In the last few years charged particles moving under the influence of a strong and singular magnetic field have been studied from a few different points of view by many authors. New and very specific effects in two-dimensional systems of particles have been discovered (see Refs. 1–6, for example). A fundamental quantity for the description of these new effects is the so-called *filling factor* υ which is defined as the ratio p/N of the number of particles p by the number of magnetic fluxes N. *A magnetic flux* is a magnetic field concentrated completely on a vertical line carrying an elementary flux $\Phi_0 := h/e$. For integer-valued υ there is the well-known *integer quanlum Hall effect* (IQHE). For fractional-valued υ with odd denominator there is the famous *fractional quantum Hall effect* (FQHE).[1] It is a very interesting and fascinating fact that for a fractional-valued filling factor with an even denominator the corresponding effect is completely different from the above-mentioned quantum Hall effect. This is the so-called *electronic magnetotransport anomaly* (the $\frac{1}{2}$-anomaly). In this case the magnetic field is completely compensated such that particles move as though in the absence of the magnetic field! In this way we obtain the following scheme:

$$\upsilon = \begin{cases} 1,2,\ldots & \text{IQHE}, \\ \frac{1}{2}, \frac{1}{4}, \ldots & \text{the } \frac{1}{2} \text{ anomaly}, \\ \frac{1}{3}, \frac{1}{5}, \ldots & \text{FQHE}. \end{cases} \tag{1}$$

At present there is no satisfactory theory for unified description of all these effects. The most popular theory for the quantum Hall effect is the Chern–Simons–Landau–Ginzburg (CSLG) theory (see for example, Ref. 7 and references therein). In this theory electrons are described as bosons carrying an odd number of flux quanta. Such a composite system is said to be *a composite boson*. The $\frac{1}{2}$-anomaly can be explained in a similar way. In this case electrons can be considered as a degenerate system of new fermions in the absence of the magnetic field. Such particles are called *composite fermions*.[4,5,8] They can be imagined as electrons carrying an even number of magnetic fluxes.

An interesting concept for a composite system which consists of a particle with charge e and a singular magnetic field concentrated completely on a vertical line has been introduced by Wilczek[9] (see also Ref. 10). The charge is moving around the singular line in the flat region in which there is no influence of the magnetic field. Every rotation yields a certain phase factor q. Observe that the space on which the particle moving is $\mathscr{M} \equiv \mathbb{R}^2 \backslash \{s\}$, where $s \in \mathbb{R}^2$ is the point of

[a)]Electronic mail: wmar@ift.uni.wroc.pl

 © 1998 American Institute of Physics

intersection of the magnetic line with the plane $\mathbb{R}^2$. The factor q can be in general a complex number of the following form: $q:=\exp(i\varphi)$, where $0\leqslant\varphi<2\pi$ is the so-called statistics parameter. This means that the statistics of the particle are determined by the value of φ. For $\varphi=0$ we have boson, for $\varphi=\pm\pi$—fermion. For arbitrary $\varphi\in[0,2\pi)$ we obtain anyons. Note that the name ''anyons'' is sometimes reserved for the case when q is the mth root of unity (see, for example, the papers of Majid,[11,12] where anyons are given in terms of certain Z_m-graded structures). The description of the statistics corresponding to composite fermions or bosons is a problem. A few remarks on this problem have been given by Jacak *et al.* in Ref. 13. According to their suggestion, the value of the statistics parameter need not be restricted to the interval $[0,2\pi)$, but can be an arbitrary real number. If $\varphi=(2k+1)\pi$, $k=1,2,...$, for example, then one can expect to obtain nothing else but the usual fermion statistics. However, according to Jacak, *et al.*, in this case we obtain a new kind of statistics, just the statistics of composite fermions. The case of $\varphi=2k\pi$, $k=1,2,...$, corresponds to composite bosons. An algebraic formalism for a particle system with generalized statistics has been considered by the author in Refs. 14–17 (see also Refs. 18–21). It is interesting that in this algebraic approach all commutation relations for particles equipped with arbitrary statistics can be described as a representation of the so-called quantum Weyl algebra $\mathscr{W}$ (or Wick algebra).[22–25] A similar approach has also been considered by other author (see Refs. 24, and 26–30). In this attempt the creation and anihilation operators act on a certain quadratic algebra $\mathscr{A}$. The creation operators act as the multiplication in $\mathscr{A}$ and the anihilation ones act as a noncommutative contraction (noncommutative partial derivatives). The algebra $\mathscr{A}$ plays the role of noncommutative Fock space. The applications for particles in a singular magnetic field have been described in Refs. 20 and 21. As a result the fact that at $\upsilon=\frac{1}{2}$ the system is degenerate and at $\upsilon=1/N$ for N odd the Landau level is fractionally filled has been obtained in an obvious way.

In this paper an algebraic formalism for a system of charged particles with spin $\frac{1}{2}$ in singular magnetic field is developed. We restrict our attention to the system with the filling factor $\upsilon=1/N$ for simplicity. The generalization for arbitrary filling factor is possible but it needs more complicated notation. The fundamental assumption of the paper is that the particle is transformed into a composite system which consists of quasiparticles, quasiholes, and magnetic fluxes. The transformation is a result of interactions of charged particle with spin with magnetic field. If N is even, then the system is identified as composite fermions, and if N is odd, then we obtain composite bosons. It is obvious that the statistics of our composite systems cannot be described by the scalar statistics parameter $q=\exp(i\varphi)$. We use here the concept of the commutation factor[31] in order to describe the statistics. The paper is organized as follows. The fundamental concept of the paper based on Refs. 19–21 is given in Sec. II. In Sec. III the quantum Weyl algebra is described in detail. It is shown that it is in fact a Γ-graded generalized (color) Lie algebra.[31] The representation of the quantum Weyl algebra is considered in Sec. IV. The Fock space of quantum states for composite fermions and composite bosons is described in Sec. V. The notion of commutation factors is briefly summarized in the Appendix. Note that the algebraic approach to the flat particle system in the singular magnetic field is not complete theory, but it can be further developed. One can write down the Hamiltonian and study the corresponding field theory. One can use the formalism described by Liguori and Mintchev[32] for such study. We hope that the possibility for the unified description of the quantum Hall effect and the electronic magnetotransport anomaly based on the algebraic formalism presented in this paper are clear.

II. FUNDAMENTAL ASSUMPTIONS

Let us consider a system of charged particles with spin $\frac{1}{2}$ moving on a flat space $\mathbb{R}^2$ under the influence of a singular perpendicular magnetic field. The magnetic field is sufficiently strong for the polarization effects of the spin. The *singularily* of the magnetic field means here that the field is completely concentrated into fluxes. We assume that there are on average N magnetic fluxes per particle. This means that the filling factor is $\upsilon=1/N$. Our fundamental assumption is that the problem of many particles interacting with the singular magnetic field can be reduced to the study of a system consisting of just one particle and N magnetic fluxes. In this way we can restrict our attention to the study of such system. It is known that quantum states of charged particles in the magnetic field are described by Landau levels. Here is an additional degeneracy of such states connected with the singular nature of the field. The structure of such degeneration is determined

by the change of phase of the particle. Let us study the degeneration in more details. We assume that the particle is in the lowest Landau level. The ''effective'' space for the particle in a singular magnetic field is $\mathscr{M}\equiv\mathbb{R}^2\backslash\{s_1,\ldots,s_N\}$, where s_i is the point of intersection of the ith magnetic flux with the plane. Let us denote by $P_m\mathscr{M}$ the space of all homotopy classes of paths which starts at m_0 and end at arbitrary point $m\in\mathscr{M}$. It is known that the union $\cup_m P_m\mathscr{M}=P\mathscr{M}$ is a covering space $P=(P\mathscr{M},\pi,\mathscr{M})$. The projection $\pi:P\mathscr{M}\to\mathscr{M}$ is given by

$$\pi(\xi)=m \quad \text{iff} \quad \xi\in P_m\mathscr{M}. \tag{2}$$

The homotopy class $P_{m_0}\mathscr{M}$ of all paths which start at m_0 and end at the same point m_0 (i.e. a loop space) can be naturally endowed with a group structure. This group is known as the fundamental group of $\mathscr{M}$ at m_0 and is denoted by $\pi_1(\mathscr{M},m_0)$. Observe that two groups $\pi_1(\mathscr{M},m_0)$ and $\pi_1(\mathscr{M},m)$ at two arbitrary points m_0 and m, respectively, are isomorphic. Hence we can introduce the fundamental group $\pi_1(\mathscr{M})$ for the whole space $\mathscr{M}$ in an obvious way. Note that in our case the fundamental group $\pi_1(\mathscr{M})$ for the space $\mathscr{M}\equiv\mathbb{R}^2\backslash\{S_1,\ldots,S_N\}$ is a free group generated by $\tilde{\sigma}_i$, $(i=1,...,N)$ and is denoted by $G_N\equiv G_N(\tilde{\sigma}_1,\ldots,\tilde{\sigma}_N)$, where $\tilde{\sigma}_i$ is the homotopy class of all paths which encloses the point m_i and none of the remaining points m_j for $j\neq i$, (see Ref. 33 for example in more detail). Next we assume that for every class of loops the phase change of particle under consideration is fixed. We introduce an equivalence relation in the fundamental group G_N. Two homotopy classes of loops in G_N are equivalent if and only if the phase changes corresponding to this classes are equal. The equivalence class of $\tilde{\sigma}_i$ with respect to the above relation is denoted by σ_i. Let Γ be a quotient group of the fundamental group G_N by the above relation. It is obvious that the phase change corresponding to $\tilde{\sigma}_i\tilde{\sigma}_j$ and $\tilde{\sigma}_j\tilde{\sigma}_i$ is the same for all $i\neq j$. Hence we have the relation $\sigma_i\sigma_j=\sigma_j\sigma_i$ in the quotient group Γ. Observe that generators σ_i of Γ can be identified with rotations around the axis at the point s_i. This means that the group Γ can be identified with the group $Z\oplus\cdots\oplus Z$ which contains N copies of the group of integers. Every rotation around the axis at s_i yields the same phase factor q_i. Two rotations around two different axis s_i and s_j are independent and they yield the phase factor c_{ij}. It is natural to assume that $c_{ij}c_{ji}=1$ and $c_{ii}\equiv q_i$. It is obvious that there is a commutation factor $c:\Gamma\times\Gamma\to\mathbb{C}\backslash\{0\}$ on the group Γ (see the Appendix) such that $c(\sigma_i,\sigma_j)\equiv c_{ij}$. The commutation factor c generalizes the scalar statistics parameter $q=\exp(i\varphi)$, where φ is an arbitrary real number. It follows from the above considerations that the structure of degenerations of Landau levels can be described by the commutation factor c on Γ. Structures of different kinds of degenerations correspond to nonequivalent classes of commutation factors on the group Γ. This means that we can use the concept of color statistics and related Γ-graded structures in order to obtain the algebraic formalism for the particle in the singular magnetic field.[17] It is interesting that in certain particular cases (see the Appendix) the grading group can be reduced to the group $\Gamma=Z_n^N\equiv Z_n\oplus\cdots\oplus Z_n$ for $n>2$ or to $Z_2^N\equiv Z_2\oplus\cdots\oplus Z_2$.

It is interesting that the structure of degeneration of the Landau levels can be described as quantum states of new particles equipped with generalized statistics. We assume that the particle under consideration is transformed into a composite system which consists of quasiparticles, quasiholes, and magnetic fluxes. Such a composite system is said to be *a composite parlticle or generalized quon.* The transformation of the particle into the composite system is a result of interactions of charged particles with spin with magnetic field. Note that if the particle is coupled with a certain flux in such a way that the magnetic field of this flux is canceled, then we say that we have *a quasiparticle.* In this case the particle is said to be *bound* to the flux. In the opposite case, i.e., when the magnetic field is not canceled, we say we have quasiholes. A *quasihole* is a ''free'' flux which behaves like a particle endowed with fermion statistics. Hence the points of intersection of N magnetic fluxes with the plane must be a set of N different points $s_1,\ldots,s_N$. The number of quasiparticles and quasiholes is equal to the number of magnetic fluxes. This means that the filling factor for the transformed system is $\nu'=(m+n)/N\equiv 1$, where m is the number of quasiparticles, and n is the number of quasiholes. The ''effective'' magnetic field is

$$B_{\text{eff}}:=B-m\Phi_0, \tag{3}$$

where B is the external magnetic field. Observe that if $B=N\Phi_0$ and the number of quasiparticles is equal to N, then the magnetic field is completely canceled! A quasiparticle is in fact the charged

particle bound to a single magnetic flux. The particles bound to two different magnetic fluxes are understood as a system of two different quasiparticles. Obviously the particle can not be bound to two fluxes at the same point! Quasiparticles as components of certain composite particles also have their own statistics. These statistics are determined by the commutation factor c on the group Γ. For an electron in a singular magnetic field it is natural to assume that we have $c_{ii} \equiv \exp(i\varphi)$, where $\varphi = \pi(N+1)$, i.e., $c_{ii} \equiv -(-1)^N$ and $c_{ij} \equiv \exp(i\pi N) \equiv (-1)^N$ for $i \neq j$. The reason that in the formula for c_{ii} we have that φ is equal to the number of magnetic fluxes N plus 1 is the spin of electrons. The above expressions for the factor c mean that two different quasiparticles for N even commute and for N odd anticommute. This also means that the composite boson can contain only one quasiparticle. The number of quasiparticles can be equal to N for composite fermions only! In other words, our quasiparticles are examples of particles equipped with the so-called color statistics, i.e., the parastatistics in Green's representation.[17] This means that one can use the concept of color statistics and related structures with the Abelian grading group in order to obtain an algebraic formalism for particles in a singular magnetic field.

III. ON QUANTUM WEYL ALGEBRAS

Let us denote by Θ^i the quantum quasiparticle state coupled with the flux at s_i. The corresponding conjugate state is denoted by Θ_j^*. It is natural to assume that there is a finite-dimensional vector space E equipped with a basis Θ^i, $i = 1,...,N = \dim E$. The complex conjugate space E^* is endowed with a basis Θ_i^* such that we have that $\langle \Theta_i^* | \Theta^j \rangle = \delta_i^j$ for $i,j = 1,...,N$. Let us give the notion of quantum Weyl algebras $\mathscr{W} \equiv \mathscr{W}_{\Gamma,c}(N)$ corresponding to an arbitrary grading group Γ and a commutation factor c on it.

Definition: A (Γ,c)-quantum Weyl algebra is a quotient algebra

$$\mathscr{W} \equiv \mathscr{W}_{\Gamma,c}(N) := T(E \oplus E^*)/I_{\Gamma,c}, \tag{4}$$

where $I_{\Gamma,c}$ is an $*$-ideal in the tensor algebra T(E⊕E) generated by the following elements*:

$$\begin{gathered} \Theta_i^* \otimes \Theta^j - c_{ij}\Theta^j \otimes \Theta_i^* - \delta_i^j \mathbf{1}, \quad \text{for all } i,j, \\ \Theta^i \otimes \Theta^j - c_{ij}\Theta^j \otimes \Theta^i, \quad \Theta_j^* \oplus \Theta_i^* - c_{ij}\Theta_i^* \otimes \Theta_i^*, \quad \text{for } i \neq j, \\ (\Theta^i)^2, \quad (\Theta_j^*)^2 \quad \text{if } c_{ii} = -1 \end{gathered} \tag{5}$$

(no sum), for i,j=1,...,N, where c_{ij} are coefficient of certain commutation factors on the grading group Γ.

The above definition means that $\mathscr{W}_{\Gamma,c}(N)$ is a $*$-algebra generated by Θ^i and Θ_j^* $(i,j = 1,...,N)$ subject to relations

$$\begin{gathered} \Theta_i^* \Theta^i = 1 + q_i \Theta^i \Theta_i^* \quad \text{for } i = j, \quad \Theta_i^* \Theta^j = c_{ji}\Theta^j \Theta_i^* \quad \text{for } i \neq j, \\ \Theta^i \Theta^j = c_{ij}\Theta^j \Theta^i \quad \text{for all } i,j, \quad \Theta_j^* \Theta_i^* = c_{ij}\Theta_i^* \Theta_j^* \quad \text{for all } i,j, \\ (\Theta^i)^2 = 0, \quad (\Theta_j^*)^2 = 0, \quad \text{if } c_{ii} = -1, \end{gathered} \tag{6}$$

where $q_i := c_{ii} : i = 1,...,N$ are diagonal elements of c (note that the same symbols for elements in the tensor algebra $T(E \oplus E^*)$ and for generators of the algebra $\mathscr{W}_{\Gamma,c}(N)$ have been used for simplicity). The algebra $\mathscr{W}_{\Gamma,c}(N)$ should be also denoted by $\mathscr{W}_{\Gamma,c}\langle \Theta^1,\ldots,\Theta^N,\Theta_i^*,\ldots,\Theta_N^* \rangle$.

Theorem: *The quantum Weyl algebra $\mathscr{W} \equiv \mathscr{W}(\Gamma,c)$ is a Γ-graded c-Lie algebra.*

Proof: We introduce the Γ-gradation of the algebra $\mathscr{W}$ as follows: for generators of the algebra we define that grade $(\Theta^i) \equiv |\Theta^i| := \sigma_i$, and grade $(\Theta_i^*) \equiv |\Theta_i^*| := -\sigma_i$, where $\{\sigma_i\}_{i=1}^N$ is a set of generators of Γ. All monomials in Θ^i and Θ_j^* modulo generating relations are homogeneous elements of $\mathscr{W}$. We use the formula $|XY| = |X| + |Y|$ for the extension of gradation for arbitrary homogeneous elements of the algebra. The generalized bracket $[.,.]_c : \mathscr{W} \otimes \mathscr{W} \to \mathscr{W}$ is defined by the formula

$$[.,.]_c X \otimes Y \equiv [X, Y]_c := XY - c(|X|,|Y|)YX, \tag{7}$$

where X,Y are arbitrary homogeneous elements of the algebra $\mathscr{W}$. We can calculate the bracket as follows: for generators we obtain the relation

$$[\Theta_i^*,\ \Theta^j]_c := \Theta_i^*\Theta_j - c(|\Theta_i^*|,|\Theta^j|)\Theta^j\Theta_i^* = \delta_i^j, \tag{8}$$

where $c(|\Theta_i^*|,|\Theta^j|)=c(-\sigma_i,\sigma_j)=c_{ji}$ and the relations (6) have been used. In a similar way we obtain the formula

$$\begin{aligned}[\Theta^i,\Theta_j^*]_c &:= \Theta^i\Theta_j^* - c(|\Theta^i|,|\Theta_j^*|)\Theta_j^*\Theta^i \\ &= -c(|\Theta^i|,|\Theta_j^*|)(\Theta_i^*\Theta_j - c(|\Theta_i^*|,|\Theta^j|)\Theta^j\Theta_i^*) \\ &= -c_{ij}\delta_i^j. \end{aligned} \tag{9}$$

For the extension of our generalized bracket to the higher-order homogeneous elements of the algebra we can use the following formulas:

$$\begin{aligned}[\Theta_i^*,\ \Theta^jX]_c &= c_{ij}\Theta^j[\Theta_i^*,\ X] + \delta_i^jX, \\ [0_i^*X,\ \Theta^j]_c &= \Theta_i^*[X,\ \Theta^j] - c_{ji}c(|X|,\sigma_j)\,\delta_i^jX, \end{aligned} \tag{10}$$

for arbitrary homogeneous X in $\mathscr{W}$. One can prove these formulas by induction. One can see that the generalized bracket is c-anticommutative,

$$[X,\ Y]_c = -c(|X|,|Y|)[Y,\ X]_c, \tag{11}$$

for homogeneous X and Y in $\mathscr{W}$. Finally, it is not difficult to calculate the following generalized Jacobi identity.

$$[X,\ [Y,\ Z]_c]_c = [[X,\ Y]_c,\ Z]_c + c(|X|,|Y|)[Y,\ [X,\ Z]_c]_c. \tag{12}$$

It is interesting that for arbitrary generalized Lie algebra $L\equiv L_{\Gamma,c}$ there is a corresponding Lie algebra or superalgebra $\mathrm{s}(L)$.[31,34] This algebra is called the superization of L and we have the following crossed product:

$$L := \mathrm{s}(L)\#C_{\Gamma,b}(N), \tag{13}$$

where $C_{\Gamma,b}(N)$ is Γ-graded b-commutative $*$-bialgebra. This means that we have the relation $gh=b(|g|,|h|)hg$ for all homogeneous $g,h\in C_{\Gamma,b}(N)$ of grade $|g|$ and $|h|$, respectively. The factor b is given by the formula (A6). The bialgebra $C_{\Gamma,b}(N)$ as a unital, associative algebra is generated by e^i,e_j^* and the following relations:

$$\begin{aligned} &e^ie^j = b_{ij}e^je^i, \quad e_i^*e_j^* = b_{ij}e_j^*e_i^*, \quad \text{for } i\neq j, \\ &e_i^*e^j = e^je_i^* = \delta_i^j, \quad \text{for all } i,j \end{aligned} \tag{14}$$

(no sum). The gradation is given in the standard way. This means that $|e^i|=\sigma_i$ and $|e_i^*| = -\sigma_i,\sigma_i$ are generators of the group Γ. The $*$-operation and the comultiplication by the formulas

$$(e^i)^* := e_i^*, \quad (e_i^*)^* := e^i, \tag{15}$$

i.e., $(e^i)^*\equiv(e^i)^{-1}$ and

$$\Delta(e^i) := e^i\otimes 1 + 1\otimes e^i, \tag{16}$$

respectively. The extension to the whole bialgebra is obvious. For the tensor product we have here the relation

$$(g\otimes h)(k\otimes l) = b(|h|,|k|)gk\otimes hl \tag{17}$$

for homogeneous $g,h,k,l \in C_{\Gamma,b}(N)$. Observe that the bialgebra $C_{\Gamma,b}(N)$ is not a usual Hopf algebra, but a Γ-graded generalized Hopf algebra (see Ref. 35) and a $*$-bialgebra.[36] The relation (13) means that L is a subalgebra of the tensor product $\mathrm{s}(L)\otimes C_{\Gamma,b}(N)$ spanned by elements of the form $X=x\otimes g$ for homogeneous $X\in L$, $x\in \mathrm{s}(L)$, $g\in C_{\Gamma,b}(N)$ of the same grade, i.e., $|X|=|x|=|g|=\alpha\in\Gamma$. The algebra $\mathrm{s}(L)$ is also Γ graded, but this gradation can be reduced to the group Γ/Γ_0. The reduction is given by the quotient map $\pi:\Gamma\to(\Gamma/\Gamma_0)$. In this way we obtain that the gradation of the algebra $\mathrm{s}(L)$ can be given by the group Z_2 or is trivial. Let us describe the superization $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ of the quantum Weyl algebra $\mathscr{W}_{\Gamma,c}(N)\equiv\mathscr{W}_{\Gamma,c}\langle\Theta^1,\ldots,\Theta^N,\Theta_1^*,\ldots,\Theta_N^*\rangle$ in terms of generators. The algebra $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ is generated by x^i and x_j^* $(i,j=1,\ldots,N)$ and relations

$$\begin{aligned} x_i^*x_i &= 1+q_ix_ix_i^* \quad \text{for } i=j,\\ x_i^*x_j &= c'_{ji}x_jx_i^* \quad \text{for } i\neq j,\\ x^ix^j &= c'_{ij}x^jx^i \quad \text{for all } i,j, \end{aligned} \tag{18}$$

where c' is given by the formula (A7). This means that $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ is also a quantum Weyl algebra, $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))\equiv\mathscr{W}_{\pi(\Gamma),c'}$. The crossed product of algebras $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ and $C_{\Gamma,b}(N)$ is a subalgebra of the tensor product $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))\otimes C_{\Gamma,b}(N)$ generated by elements

$$\Theta^i:=x^i\otimes e^i \quad \Theta_i^*:=x_i^*\otimes e_i^* \quad (\text{no sum}). \tag{19}$$

Note that the above expressions for Θ^i and Θ_j^* are not unique. There is a freedom of choosing the generators of the bialgebra $C_{\Gamma,b}(N)$ described by the orthogonal group $\mathrm{O}(N)$.[37] Let us consider a few simple examples.

Example 1: Assume that $\Gamma=Z\otimes\cdots\otimes Z$ (N-summands) and the factor c is given by the relation (A9), where $\Omega_{ij}=1$ for $i\neq j$. Hence the commutation rules (6) can be given in the following form:

$$\begin{aligned} \Theta_i^*\Theta^i-q_i\Theta^i\Theta_i^* &= 1 \quad \text{for } i=j,\\ \Theta_i^*\Theta^j-\omega^{-1}\Theta^j\Theta_i^* &= 0 \quad \text{for } i<j,\\ \Theta^i\Theta^j-\omega\Theta^j\Theta^i &= 0 \quad \text{for } i<j. \end{aligned} \tag{20}$$

Example 2: Let us consider the corresponding case for $\Gamma=Z_2\otimes\cdots\otimes Z_2$, in more detail. If we substitute the commutation factor c of the form (A13) into the formulas (6), then we obtain the following relations:

$$\begin{aligned} \Theta_i^*\Theta^i+(-1)^N\Theta^i\Theta_i^* &= 1 \quad \text{for } i=j,\\ \Theta_i^*\Theta^j-(-1)^{\Omega_{ij}}\Theta^j\Theta_i^* &= 0 \quad \text{for } i<j,\\ \Theta^i\Theta^j-(-1)^{\Omega_{ij}}\Theta^j\Theta^i &= 0 \quad \text{for } i<j. \end{aligned} \tag{21}$$

It is interesting that in this case the bialgebra $C_{\Gamma,b}(N)$ as algebra is reduced to the standard Clifford algebra C_N. In fact we have $(e^i)^*\equiv e^i$ and we obtain the well-known relations $e_ie_j+e_je_i=2\delta_{ij}$ for $i,j=1,\ldots,N$.

Example 3: If N is even, we obtain the following relations for composite fermions:

$$\begin{aligned} \Theta_i^*\Theta^i-\Theta^i\Theta_i^* &= 1 \quad \text{for } i=j,\\ \Theta_i^*\Theta^j+\Theta^j\Theta_i^* &= 0 \quad \text{for } i<j,\\ \Theta^i\Theta^j+\Theta^j\Theta^i &= 0 \quad \text{for } i<j, \end{aligned} \tag{22}$$

The quantum Weyl algebra generated by these relations is denoted by $\mathscr{W}_{cf}(N)$. We can see that the algebra $s(\mathscr{W}_{cf}(N))$ is generated by the usual canonical commutation relations for the system of N bosons.

Example 4: If N is odd, then we obtain relations for composite bosons:

$$\begin{aligned} &\Theta_i^*\Theta^i+\Theta^i\Theta_i^*=1 \quad \text{for } i=j,\\ &\Theta_i^*\Theta^j-\Theta^j\Theta_i^*=0 \quad \text{for } i<j,\\ &\Theta^i\Theta^j-\Theta^j\Theta^i=0 \quad \text{for } i<j, \end{aligned} \tag{23}$$

The corresponding quantum Weyl algebra is denoted by $\mathscr{W}_{cb}(N)$. In this case the algebra $s(\mathscr{W}_{cb}(N))$ is generated by the usual canonical anticommutation relations for the system of N fermions.

Example 5: Let C_N be the Clifford algebra. If we substitute $\Theta^i \equiv \Theta^* \equiv \frac{1}{2}e_i$, and

$$c_{ij}=\begin{cases} +1 & \text{for } i=j,\\ -1 & \text{for } i\neq j, \end{cases}$$

then we obtain a particular example of the quantum Weyl algebra denoted by $C_\Gamma(N)$.

IV. REPRESENTATION OF THE QUANTUM WEYL ALGEBRA

In this section we are going to construct a representation of the quantum Weyl algebra $\mathscr{W} \equiv \mathscr{W}_{\Gamma,c}(N)$ on the c-symmetric algebra $\mathscr{A}$. An algebra defined as the quotient $\mathscr{A} \equiv \mathscr{A}_c(E) := TE/I_c$ where I_c is an ideal in TE generated by the following relations,

$$\begin{aligned} &\Theta^i\Theta^j-c_{ij}\Theta^j\Theta^i \quad \text{for } i\neq j,\\ &(\Theta^i)^2 \quad \text{if } c_{ii}=-1, \end{aligned} \tag{24}$$

is said to be c-symmetric algebra over E.[14] The algebra $\mathscr{A} \equiv \mathscr{A}_c(E)$ can also be denoted by $\mathscr{A}_c\langle\Theta^1,\ldots,\Theta^N\rangle$, sometimes. It is easy to see that $\mathscr{A}_c(E)$ is a Γ-graded c-commulative algebra.

Theorem: *Let $\mathscr{W}$ be a Clifford–Weyl algebra and $\mathscr{A}$ be a c-symmetric algebra. Then there is a representation a: $\mathscr{W} \to End(\mathscr{A})$ of the quantum Weyl algebra $\mathscr{W} \equiv \mathscr{W}_{\Gamma,c}(N)$ on the Γ-graded c-commutative algebra $\mathscr{A}$.*

Proof: We must prove that there is a well-defined homogeneous homomorphism $a: \mathscr{W} \to \mathrm{End}(\mathscr{A})$ which transforms the algebra $\mathscr{W}$ as generalized Lie algebra into the generalized Lie algebra $\mathrm{End}(\mathscr{A})$ of linear endomorphisms of $\mathscr{A}$. This means that we have the relation

$$a_{[X,\,Y]_c}f=[a_X,\,a_Y]_c f \tag{25}$$

for every homogeneous $X, Y \in \mathscr{W}$ and $f \in \mathscr{A}$. The bracket on the right-hand side of the above formula is defined by the relation

$$[a_X,\,a_Y]_c := a_X a_Y - c(|X|,|Y|)a_Y a_X. \tag{26}$$

For generators Θ^i and $\Theta_j^*(i,j=1,\ldots,N)$ we define

$$a_{\Theta^i} f := m(\Theta^i \otimes f) \equiv \Theta^i f, \quad a_{\Theta_i^*}: f \equiv ev_k(\Theta_i^* \otimes f), \tag{27}$$

for every $f \in \mathscr{A}^k$, where $ev_k: E^* \otimes \mathscr{A}^{\otimes k} \to \mathscr{A}^{\otimes k-1}$ are a set of linear mappings defined by the following formulas:

$$\begin{aligned} &ev_1(\Theta_i^* \otimes \Theta^j) := \langle \Theta_i^* | \Theta^j \rangle = \delta_i^j,\\ &ev_k(\Theta_i^* \otimes \Theta^j f) := ev_1(\Theta_i^* \otimes \Theta^j) f + c_{ij}\Theta^j ev_{k-1}(\Theta_i^* \otimes f), \end{aligned} \tag{28}$$

for $f \in \mathscr{A}^k$. These mappings are said to be *a right evaluation*. Observe that we have the relation

$$ev_{k-1}(\Theta_i^*\otimes ev_k(\Theta_j^*\otimes f))-c_{ij}ev_{k-1}(\Theta_j^*\otimes(ev_k(\Theta_i^*\otimes f))=0 \tag{29}$$

on $E^*\otimes E^*\otimes\mathscr{A}^k$. We have for example,

$$ev_2(\Theta_i^*\otimes\Theta^k\Theta^1)=\delta_i^k\Theta^l+c_{ik}\delta_{il}\Theta^k. \tag{30}$$

We can see that the evaluation $ev:=\{ev_k:k=1,...\}$ can be well defined in a consistent way on the whole algebra $\mathscr{A}$ (see Refs. 18, 22, and 23). We can also see that the definition (27) can be extended in a consistent way for an arbitrary homogeneous element of the algebra $\mathscr{W}$. Obseve that we have here the following lemma: □

Lemma: We have on the algebra $\mathscr{A}$ the following commutation relations for the representation a of algebra $\mathscr{W}_{\Gamma,c}(N)$ on $\mathscr{A}$:

$$[a_{\Theta_i^*},\ a_{\Theta^j}]_c=\delta_i^j\mathbf{1},\quad [a_{\Theta^i},\ a_{\Theta^j}]_c=0,\quad [a_{\Theta_i^*},\ a_{\Theta_j^*}]_c=0. \tag{31}$$

Proof: Using the relations (27) for the first relation (31) we obtain

$$\begin{aligned}([a_{\Theta_i^*},\ a_{\Theta^j}]_c)f&=(a_{\Theta_i^*},\ a_{\Theta^j}-c_{ij}a_{\Theta_j^*}\ a_{\Theta^i})f\\&=[ev_{l+1}(\Theta_i^*\otimes\Theta^j f)-c_{ij}\Theta^j ev_l(\Theta_i^*\otimes f)]\\&=ev_1(\Theta_i^*\otimes\Theta^j)f=\delta_{ij}f,\end{aligned} \tag{32}$$

where $f\in\mathscr{A}^l$ and the relation (28) has been used. The second relation (31) follows immediately from the c-commutativity of the algebra $\mathscr{A}$. The last relation follows from the equation (29).

Let $a:\mathscr{W}_{\Gamma,c}(N)\to\mathrm{End}(\mathscr{A})$ be a Fock representation of $\mathscr{W}_{\Gamma,c}(N)$ on $\mathscr{A}$. Then there is a corresponding Fock representation $\tilde{a}:\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))\to\mathrm{End}(\mathrm{s}(\mathscr{A}_c))$ of the algebra $\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ on $\mathrm{s}(\mathscr{A})$, where $\mathrm{s}(\mathscr{A})$ is a commutative (or super, commutative) algebra, the superization of $\mathscr{A}$. For the representation a we have the following decomposition:

$$\begin{aligned}a_{\Theta^i}&\equiv a_{x^i}\otimes e^i:=\tilde{a}_{\Theta^i}\otimes L_{e^i},\\a_{\Theta_i^*}&\equiv a_{x_i^*}\otimes_{e_i^*}:=\tilde{a}_{\Theta_i^*}\otimes L_{e_i^*}\end{aligned} \tag{33}$$

where $x^i,\ x_i^*\in\mathrm{s}(\mathscr{W}_{\Gamma,c}(N))$ are given by (19), $e^i,\ e_i^*\in C_{\Gamma,b}(N)$, and $L:C_{\Gamma,b}(N)\to\mathrm{End}(C_{\Gamma,b}(N))$ is the left regular representation of $C_{\Gamma,b}(N)$ on itself, i.e., $L_g h=gh$ for arbitrary $g,h\in C_{\Gamma,b}(N)$. Obviously the above decomposition is not unique.

V. FOCK SPACE FOR COMPOSITE FERMION AND BOSSON

Let us describe the Fock space for composite fermion and composite boson in detail. The ground state vector $|0\rangle$ is defined as usual, i.e. $a_{\Theta_i^*}|0\rangle=0$. We use here the notation

$$a_{\Theta^i}|0\rangle\equiv a_i^+|0\rangle\equiv\Theta^i \tag{34}$$

and

$$a_a^+|0\rangle\equiv(a_1^+)^{a^1}\cdots(a_N^+)^{a^N}|0\rangle\equiv(\Theta^1)^{a^1}\cdots(\Theta^N)^{a^N}. \tag{35}$$

We use the similar notation for the operators $\tilde{a}_{x^i}$ and L_{e^i}. From relations (33) we obtain

$$a_\alpha^+|0\rangle=\tilde{a}_\alpha^+\otimes L_\alpha|0\rangle=(\Theta^1)^{\alpha^1}\cdots(\Theta^N)^{\alpha^N}\otimes(e^1)^{\alpha^1}\cdots(e^N)^{\alpha^N}. \tag{36}$$

Let us consider an example of representation corresponding to an electron in a singular magnetic field. We assume that the electron is represented by one Grassmann variable Θ, i.e., $\Theta^2=0$. For this representation we introduce here the following state vectors.

$$a_i^+|0\rangle := \begin{pmatrix} 1 \\ \vdots \\ \Theta \\ \vdots \\ 1 \end{pmatrix}, \tag{37}$$

where Θ is on the ith row. In this way an arbitrary state vector can be given in the following form,

$$a_\alpha^+|0\rangle := \begin{pmatrix} \Theta^{\alpha^1} \\ \vdots \\ \Theta^{\alpha^N} \end{pmatrix}, \tag{38}$$

for $\alpha=(\alpha^1,\ldots,\alpha^N)$, $\alpha^i=0$ or 1 for $i=1,...,N$. Now let us study this representation in more detail. Observe that for even N we obtain

$$\Theta^i\Theta^j=\Theta^j\Theta^i, \quad (\Theta^i)^2=0. \tag{39}$$

We have

$$\Theta^i\Theta^j=\begin{pmatrix} 1 \\ \vdots \\ \Theta \\ \vdots \\ \Theta \\ \vdots \\ 1 \end{pmatrix}=\Theta^j\Theta^i \tag{40}$$

and

$$(\Theta^i)^2=\begin{pmatrix} 1 \\ \vdots \\ \Theta^2 \\ \vdots \\ 1 \end{pmatrix}=0. \tag{41}$$

Let us consider the case of $N=2$ in more detail. In this case we have the following states:

$$\Theta^1=\begin{pmatrix} \Theta \\ 1 \end{pmatrix}, \quad \Theta^2=\begin{pmatrix} 1 \\ \Theta \end{pmatrix}, \tag{42}$$

and

$$\Theta^1\Theta^2=\begin{pmatrix} \Theta \\ \Theta \end{pmatrix}. \tag{43}$$

The filling factor for all these states (42) and (43) is $v=\frac{1}{2}$. These two states (42) contain quasihotes. Observe that the state (43) describes an electron bond to two fluxes which is transformed into a system containing two quasiparticles and two fluxes. This system is nonlocal due to the anticommutativity of fluxes. Note that the state (43) is the unique state corresponding to the filling factor $v=\frac{1}{2}$ without quasiholes. Observe that this state is completely filled by quasiparticles. This means that for $B=2\Phi_0$ we obtain that $B_{\text{eff}}=0$. In this way the state (43) corresponds to the $\frac{1}{2}$-anomaly.

For odd N we obtain

$$\Theta^i\Theta^j=-\Theta^j\Theta^i \quad \text{for } i\neq j. \tag{44}$$

For $i=j$ we obtain the identity $\Theta^i\Theta^i=\Theta^i\Theta^i$. In this case the space of states can also be represented by the variable Θ such that we have

$$\Theta^i=\begin{pmatrix}1\\ \vdots\\ \Theta\\ \vdots\\ 1\end{pmatrix}, \tag{45}$$

where Θ is on the ith row. We have

$$\Theta^i\Theta^j=\begin{pmatrix}1\\ \vdots\\ \Theta\\ \vdots\\ \Theta\\ \vdots\\ 1\end{pmatrix}=-\Theta^j\Theta^i. \tag{46}$$

This means that

$$\Theta^i\Theta^j=0 \tag{47}$$

for all $i\neq j$. Observe that the quantum state x^ix^j $(i\neq j)$ corresponding to particle coupled to magnetic fluxes at two different points disappear,

$$\Theta^i\Theta^j=x^ix^j\otimes e^ie^j=0, \tag{48}$$

and the state describing the particle coupled to a few fluxes at the same point is also impossible. In fact we have

$$(\Theta^i)^2=(x_i)^2\otimes(e_i)^2=(x_i)^2\otimes 1=(x_i)^2. \tag{49}$$

This means that the state $(\Theta^i)^2$ is not equipped with a flux. Let us consider as an example the case of $N=3$, i.e., the filling factor is $v=\frac{1}{3}$. In this case we have the states

$$\Theta^1=\begin{pmatrix}\Theta\\ 1\\ 1\end{pmatrix},\quad \Theta^2=\begin{pmatrix}1\\ \Theta\\ 1\end{pmatrix},\quad \Theta^3=\begin{pmatrix}1\\ 1\\ \Theta\end{pmatrix}, \tag{50}$$

which contain two quasiholes. Observe that the following states,

$$\Theta^1\Theta^2,\quad \Theta^1\Theta^3,\quad \Theta^2\Theta^3, \tag{51}$$

which contain one quasihole, and the state

$$\Theta^1\Theta^2\Theta^3, \tag{52}$$

which does not contain quasiholes, are impossible. Hence in this case the single quasiparticle states with two quasiholes are possible! This also means that the lowest Landau level cannot be completely filled by quasiparticles but fractionally filled only! This is just the FQHE.

We can see that there is a well-defined scalar product on $\mathscr{A}$ (Ref. 25),

$$\langle\Theta^\sigma|\Theta^\tau\rangle_{q,b}=\Sigma_{\pi\in S_n}\chi_n(\pi)\langle\Theta^{i_1}|\Theta^{\pi(j_1)}\rangle\cdots\langle\Theta^{i_n}|\Theta^{\pi(j_n)}\rangle, \tag{53}$$

where $\Theta^\sigma,\ \Theta^\tau\in\mathscr{A}^n$, $\Theta^\sigma=\Theta^{i_1}\cdots\Theta^{i_n}$, $\Theta^\tau=\Theta^{j_1}\cdots\Theta^{j_n}$, and

$$\chi(\pi):=\Pi_{(i,j)\in J(\pi)}c_{ij}=\Pi_{(i,j)\in J(\pi)}q_i^{n_i}b_{ij}, \tag{54}$$

$J(\pi):=\{(i,j):1\leqslant i\leqslant j\leqslant n\}$, $\pi(i)>\pi(j)$, and n_i is the number of elements of the set $K(\pi):=\{(i,j)\in J(\pi):i=j\}$. It follows immediately from the considerations of Bozejko and Speicher[38] that the corresponding scalar product is positive definite.

ACKNOWLEDGMENTS

The paper is partially supported by KBN, Grant No. 2P03B 130,12 and by the Katholischer Akademischer Auslander Dienst, Bonn, Germany. The author would like to thank J. Lukierski for discussion, R. Gielerak for some remarks, and A. Borowiec for other help.

APPENDIX:

Let us briefly recall the concept of commutation factors on an arbitrary Abelian group Γ.[31] A mapping $c:\Gamma\times\Gamma\to\mathbb{C}$ such that

$$c(\alpha+\beta,\gamma)=c(\alpha,\gamma)c(\alpha,\gamma),\quad c(\alpha,\beta+\gamma)=c(\alpha,\beta)c(\beta\gamma), \tag{A1}$$

is said to be a bicharacter on Γ. If in addition we have the relation

$$c(\alpha,\beta)c(\beta,\alpha)=1, \tag{A2}$$

then c is said to be a commutation factor on Γ. We use here the following notation $c(\sigma_i,\sigma_i)\equiv c_{ii}\equiv q_i$ and $c(\sigma_i,\sigma_j)=c_{ij}$ for $i\neq j$, where $q_i=+1$ or -1 is said to be a parity of c, $c_{ij}\in\mathbb{C}\setminus\{0\}$ are deformation parameters such that $c_{ij}c_{ji}=1$, $\alpha=\Sigma_i\alpha^i\sigma_i$, $\beta=\Sigma_j\alpha^j\sigma_j$, and σ_i are generators of Γ. Note that we have in general the following formula for the commutation factor c on the group Γ:

$$c(\alpha,\beta)=\Pi_{i,j}(c_{ij})^{\alpha^i\beta^j}=\Pi_i(c_{ij})^{\alpha^i\beta^i}\Pi_{i<j}(c_{ij})^{\alpha^i\beta^j-\alpha^j\beta^i}. \tag{A3}$$

It is obvious that the set $\Gamma_0:=\{\alpha\in\Gamma:c(\alpha,\alpha)=1\}$ is a subgroup of Γ of index at most 2. This means that the quotient Γ/Γ_0 is isomorphic to the group Z_2 or is trivial. The commutation factor c can be given in the following form,

$$c(\alpha,\beta)=c'(\alpha,\beta)b(\alpha,\beta), \tag{A4}$$

where b and c' are two new commutation factors on Γ. The factor b is defined by the relations

$$b(\sigma_i,\sigma_j)\equiv b_{ij}:=\begin{cases}+1 & \text{for } i=j,\\ -c_{ij} & \text{for } i\neq j \text{ if } q_i=q_j=-1,\\ c_{ij} & \text{in the remaining cases.}\end{cases} \tag{A5}$$

It follows immediately from the formula (A3) that

$$b(\alpha,\beta)=\Pi_{i<j}(b_{ij})^{\alpha^i\beta^j-\alpha^j\beta^i}. \tag{A6}$$

The factor c' is defined by the formula $c'_{ii}\equiv c_{ii}\equiv q_i$ and

$$c'(\sigma_i,\sigma_j)\equiv c'_{ij}=\begin{cases}-1 & \text{if } q_i=q_j=-1,\\ +1 & \text{otherwise,}\end{cases} \tag{A7}$$

for $i\neq j$. It is interesting that the factor c' can be reduced to the group $\pi(\Gamma)\equiv\Gamma/\Gamma_0$, where $\pi:\Gamma\to\Gamma/\Gamma_0$ is the quotient map. We can see that there is the relation

$$c'(\alpha,\beta)=(-1)^{\pi(\alpha)\pi(\beta)}. \tag{A8}$$

For $\Gamma=Z^N\equiv Z\oplus\cdots\oplus Z$ (N-summands) we have

$$q_i=(-1)^{S_i},\quad b_{ij}=\omega^{\Omega_{ij}}, \tag{A9}$$

where $S_i=0,1$, and Ω_{ij} are elements of a skew-symmetric integer-valued matrix, and $\omega\neq -1$ is some complex parameter.[39] If $\omega=\exp(2\pi i/n)$, $n\geqslant 3$, and $S_i\equiv 0$, then the grading group $G\equiv Z^N$ can be reduced to $\Gamma=Z_n^N\equiv Z_n\oplus\cdots\oplus Z_n$. If $\omega=-1$, then the grading group Γ can be reduced to the group

$$G\equiv Z_2^N:=Z_2\oplus\cdots\oplus Z_2 \quad (N \ \ summands). \tag{A10}$$

It follows from our considerations in Sec. II that for the charged particle in the singular magnetic field we have $S_i=N+1 \pmod 2$, and

$$\Omega_{ij}:=\begin{cases} 0 & \text{for } i=j, \\ 1 & \text{for } i\neq j. \end{cases} \tag{A11}$$

This means that

$$b_{ij}:=\begin{cases} +1 & \text{for } i=j, \\ -1 & \text{for } i\neq j, \end{cases} \tag{A12}$$

and the factor c is given by the formula

$$c_{ij}=-(-1)^N(-1)^{\Omega_{ij}}. \tag{A13}$$

In such case if the number N of fluxes is even (i.e., for composite fermions) we obtain

$$c_{ij}=-(-1)^{\Omega_{ij}}=\begin{cases} -1 & \text{for } i=j, \\ +1 & \text{for } i\neq j. \end{cases} \tag{A14}$$

If N is odd (composite bosons), then we obtain

$$c_{ij}=(-1)^{\Omega_{ij}}:=\left\{\begin{matrix} +1 & \text{for } i=j, \\ -1 & \text{for } i\neq j. \end{matrix}\right\}. \tag{A15}$$

[1] D. C. Tsui, H. L. Stormer, and A. C. Gossard, Phys. Rev. Lett. **48**, 1559 (1982).
[2] Z. F. Ezawa and H. Hotta, Phys. Rev. B **46**, 7765 (1992).
[3] B. I. Halperin, P. A. Lee, and N. Read, Phys. Rev. B **46**, 7312 (1993).
[4] J. K. Jain, Phys. Rev. Lett. **63**, 199 (1989); Phys. Rev. B **40**, 8079 (1989); **41**, 7653 (1990).
[5] R. R. Du, H. L. Stormer, D. C. Tsui, A. S. Yeh, L. N. Pfeiffer, and K. W. West, Phys. Rev. Lett. **73**, 3274 (1994).
[6] A. Arai, J. Math. Phys. **36**, 2569 (1995).
[7] V. Kalmeyer and S-C. Zhang, Phys. Rev. B **46**, 9889 (1992).
[8] V. J. Goldman, B. Su, and J. K. Jain, Phys. Rev. Lett. **72**, 2065 (1994).
[9] F. Wilczek, Phys. Rev. Lett. **49**, 957 (1982).
[10] F. Wilczek and A. Zee, Phys. Rev. Lett. **51**, 2250 (1983).
[11] S. Majid, ''Anyonic quantum groups,'' in *Spinors, Twistors and Clifford Algebras*, edited by Z. Oziewicz *et al.* (Kluwer Academic, Dordrecht, 1993), p. 327.
[12] S. Majid, Czech. J. Phys. **44**, 1073 (1994).
[13] L. Jacak, P. Sitko, and K. Wieczorek, ''Anyons and composite fermions,'' Scientific Papers of the Institute of Physics of the Technical University of Wroclaw No. 31, Monographs No. 20, Wroclaw (1995).
[14] W. Marcinek, J. Math. Phys. **33**, 1631 (1992).
[15] W. Marcinek, Rep. Math. Phys. **34**, 325 (1994).
[16] W. Marcinek, Rep. Math. Phys. **33**, 117 (1993).
[17] W. Marcinek, Int. J. Mod. Phys. A **10**, 1465 (1995).
[18] W. Marcinek, ''On Commutation Relations for Quons,'' q-alg/9512015 (in preparation).
[19] W. Marcinek, ''On quantum Weyl algebras and generalized quons,'' in *Proceedings of the Symposium: Quantum Groups and Quantum Spaces*, Warsaw, Poland, November 20–29, 1995, edited by R. Budzynski, W. Pusz, and S. Zakrzewski (Banach Center, Warsaw, 1997).
[20] W. Marcinek, ''Topology and quantization,'' in *Proceeding of the IVth International School on Theoretical Physics, Symmetry and Structural Properties*, Zajaczkowo k. Poznania, August 29–September 4 1996, Poland, ed. by T. Lulek, W. Florek, and B. Lulek (World Scientific, Singapore, 1997), pp. 415–424.
[21] W. Marcinek, ''On algebraic model of composite fermions and bosons,'' *Proceedings of the IXth Max Born Symposium*, Karpacz, Poland, September 25–28, 1996, in *New Symmetries in the Theories of Fundamental Interactions*, ed. by J. Lukierski and M. Morymas (PWN, Warszawa, 1997), pp. 307–312.
[22] W. Marcinek and R. Ralowski, ''Particle operators from braided geometry,'' in *Quantum Groups, Formalism and Applications*, XXX Karpacz Winter School in Theoretical Physics, 1994, edited by J. Lukierski *et al.* (1995), pp. 149–154.

[23] W. Marcinek and Robert Rałowski, J. Math. Phys. **36**, 2803 (1995).
[24] P. E. T. Jorgensen, L. M. Schmith, and R. F. Werner, J. Funct. Anal. **134**, 33 (1995).
[25] R. Ralowski, ''On deformations of commutation relation algebras,'' Phys. Lett. A. **30**, 3235 (1997).
[26] R. Scipioni, Phys. Lett. B **327**, 56 (1994).
[27] Y. Ting and W. Zhao-Yan, Sci. China A **37**, 1472 (1994).
[28] S. Meljanac and A. Perica, Mod. Phys. Lett. A **9**, 3293 (1994).
[29] M. Pillin, Commun. Math. Phys. **180**, 23 (1996).
[30] A. K. Mishra and G. Rajasekaran, J. Math. Phys. **38**, 466 (1997).
[31] M. Scheunert, J. Math. Phys. **20**, 712 (1979).
[32] A. Liguori and M. Mintchew, Commun. Math. Phys. **169**, 635 (1995).
[33] A. P. Balachandram, G. Marmo, B. S. Skagerstam, and A. Steam, *Classical Topology and Quantum States* (World Scientific, Singapore, 1991).
[34] W. Marcinek, Rep. Math. Phys. **27**, 385 (1989).
[35] W. Marcinek, Rep. Math. Phys. **32**, 1 (1993).
[36] W. Marcinek, ''On *-Algebras and Noncommutative Geometry,'' Preprint KL-TH 95/19 (1995).
[37] W. Marcinek, Europhys. Lett. **18**, 133 (1992).
[38] M. Bożejko, R. Speicher, Math. Ann. **300**, 97 (1994).
[39] Z. Oziewicz, ''Lie algebras for arbitrary grading group,'' in *Differential Geometry and Its Applications*, edited by J. Janyska and D. Krupka (World Scientific, Singapore, 1990), pp. 148–154.

Vacuum polarization and the geometric phase: Gauge invariance

Jouko Mickelsson
Department of Theoretical Physics, Royal Institute of Technology, S-10044 Stockholm, Sweden

(Received 13 May 1997; accepted for publication 17 September 1997)

A nonperturbative approach to the vacuum polarization for quantized fermions in external vector potentials is discussed. It is shown that by a suitable choice of counterterms the vacuum polarization phase is both gauge and renormalization independent, within a large class of nonperturbative renormalizations. © *1998 American Institute of Physics.* [S0022-2488(98)02602-4]

I. INTRODUCTION

For Dirac fermions in (non-second quantized) external Yang–Mills fields the one-particle scattering matrix S is a well-defined unitary operator, and, moreover, it has a canonical second quantization $\hat{S}$ operating in the fermionic Fock space;[1] for twice differentiable vector potentials one only needs to assume certain falloff properties at spatial and time infinity.

On the other hand, if one tries to expand the quantum scattering matrix in ordinary Dyson–Feynman perturbation series one meets the well-known vacuum polarization divergences, which must be taken care by suitable (infinite) subtractions. Actually, it is only the phase of the vacuum expectation value $\langle 0|\hat{S}|0\rangle$ that is diverging; the absolute value is uniquely defined and finite by canonical quantization. The crucial point is that when passing from S to $\hat{S}$ using the rules of canonical Hamiltonian quantization the phase remains ill defined. It is exactly this quantity that is diverging in perturbation theory.

One of the basic principles of relativistic quantum field theory is locality. The fields are supposed to either commute (bosons) or anticommute (fermions) at space-like separations. This is the case at least in four or higher space–time dimensions; in lower dimensions there are more alternatives. For this reason one prefers field theories in which the action is a local differential polynomial of the fields. On the other hand, the requirement that the action is a function of the fields and their derivatives up to a finite degree at the same space–time point is by no means a necessary condition for the locality in the above sense. In this paper we continue the work on a nonperturbative renormalization, which breaks the locality in the narrow sense (the renormalized action contains field derivatives up to infinite degree) but nevertheless is local in the framework of Wightman axioms.

This work was initiated in Ref. 2 and further generalized in Ref. 3. In the previous paper[3] the question of gauge and renormalization independence was left open. We fill the gap in the present paper in the case of infinitesimal gauge and renormalization variations.

II. THE RENORMALIZATION

We shall study massless Dirac fermions coupled to a gauge potential A in Minkowski space. The potential is a smooth 1-form $A_\mu(x)dx^\mu$ in space–time with values in the Lie algebra $\mathbf{g}$ of a compact gauge group G. The elements of $\mathbf{g}$ are represented by Hermitian (according to physics literature convention) matrices in the complex vector space $\mathbb{C}^N$. The free Dirac operator is then $i\gamma^\mu\partial_\mu$. The metric is $x^\mu x_\mu = x_0^2 - x_1^2 - \cdots - x_d^2$. The Dirac gamma matrices satisfy $\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 2g_{\mu\nu}$, γ_0 is Hermitian, and γ_k is anti-Hermitian for $k \neq 0$. We fix a Hermitian matrix Γ such that $\Gamma^2 = 1$ and $\Gamma\gamma_\mu = -\gamma_\mu\Gamma$. The chiral projectors are $P_\pm = \frac{1}{2}(\Gamma \pm 1)$.

The Dirac Hamiltonian in background gauge field A is

$$D_A = -\gamma^0\gamma^k(i\partial_k + A_k) - A_0.$$

 © 1998 American Institute of Physics

We shall assume that $A(x)$ and its derivatives vanish faster than $|x|^{-d/2}$ when $|x|\to\infty$.

The one-particle scattering operator S is defined as the limit of the time evolution operator in the interaction picture, $U_I(t,-t)\to S$, as $t\to\infty$. The time evolution in the Schrödinger picture is defined by

$$i\ \partial_t U(t,t_0)=h(t)U(t,t_0),\quad \text{with}\ \ U(t_0,t_0)=1, \tag{1}$$

and in the interaction picture by

$$i\ \partial_t U_I(t,t_0)=V_I(t)U_I(t,t_0),\quad \text{with}\ \ U_I(t_0,t_0)=1. \tag{2}$$

The interaction is $V_I(t)=e^{ith_0}V(t)e^{-ith_0}$, the total Hamiltonian being $h(t)=D_A=h_0+V(t)$ with $h_0=D_0=-i\gamma_0\gamma^k\partial_k$. The quantum divergences are related to the fact that when $V=-\gamma^0\gamma^k A_k(\mathbf{x},t)-A_0(\mathbf{x},t)$ is the interaction with a Yang–Mills potential, then the quantization of $\hat{U}_I(t,-\infty)$ for intermediate times $t<\infty$ is not well defined. Let $\epsilon=h_0/|h_0|$ be the sign of the free Hamiltonian. In general, a one-particle operator K has a canonical quantization in the Fock space of free fermions iff $[\epsilon,K]$ is Hilbert–Schmidt. For a review of this and other properties of representations of CAR algebra, see Ref. 4. Generally, the Hilbert–Schmidt property is not satisfied when $K=V(t)$.

Let us recall the basic facts about (free) Fock space quantization of the canonical anticommutation relations. To each vector $u\in H$ in the one-particle Hilbert space one associates a creation operator $a^*(u)$ and an annihilation operator $a(u)$, the latter depending antilinearly on the parameter and the former linearly, with the only nonzero anticommutation relations,

$$a^*(u)a(v)+a(v)a^*(u)=\langle v,u\rangle,$$

where $\langle v,u\rangle$ is the inner product in H. Let $H=H_+\oplus H_-$ be splitting to a pair of closed infinite-dimensional subspaces. In our discussion below H_+ will be the spectral subspace corresponding to non-negative energies with respect to the free Dirac hamiltonian D_0. There is a unique (up to equivalence) irreducible representation of the CAR algebra in a Hilbert space $\mathscr{F}$ such that

$$a^*(u)|0\rangle=0=a(v)|0\rangle,\quad \text{for}\ \ u\in H_-\quad\text{and}\quad v\in H_+,$$

where $|0>\in\mathscr{F}$ is a normalized vector, the vacuum.

Let a_i^*, a_j, with $i,j\in\mathbb{Z}$, be a complete set of creation and annihilation operators such that the index $i\geqslant 0$ corresponds to non-negative energies and $i<0$ to negative energies,

$$a_i^*a_j+a_ja_i^*=\delta_{ij}.$$

The normal ordering is defined by $a_i^*a_j:=a_i^*a_j$, except when $i=j<0$, and then $a_i^*a_j:=-a_ja_i^*$. If X,Y is a pair of one-particle operators then the canonical quantizations,

$$\hat{X}=\sum\ X_{ij}:a_i^*a_j:,$$

satisfy the algebra

$$[\hat{X},\hat{Y}]=\widehat{[X,Y]}+\omega(X,Y),$$

and the quantum operators are defined such that $[\hat{X},a_i^*]=\Sigma X_{ji}a_j^*$. The 2-cocycle ω is defined by[5]

$$\omega(X,Y)=\tfrac{1}{4}\ \mathrm{tr}\ \epsilon[\epsilon,X][\epsilon,Y]. \tag{3}$$

Here we meet again the Hilbert–Schmidt condition: The 2-cocycle is defined only for bounded operators X such that $[\epsilon,X]$ is Hilbert–Schmidt.

The starting point for our discussion here is the following renormalization that makes U_I quantizable.[2,3] For each (smooth) potential $A=A_0\ dt+A_k\ dx^k$ one defines a unitary operator T_A in the one-particle space with the following property. Define $U_{\mathrm{ren}}(t,t_0)=T_A(t)U_I(t,t_0)T_A(t_0)^{-1}$.

Then (1) $[\epsilon, U_{\rm ren}(t,t_0)]$ is Hilbert–Schmidt, and (2) $T_A(t)\to 1$ as $t\to\pm\infty$. The last property quarantees that the renormalization does not affect the scattering matrix S whereas the first condition guarantees that the renormalized time evolution is quantizable in the free Fock space.

The Hilbert–Schmidt condition on operators $[\epsilon, g]$ defines the restricted unitary group U_1 of unitary operators g.[6] The second quantization of elements in U_1 defines a central extension $\hat{U}_1$, as discussed in detail in Ref. 6. Now we have a smooth path of operators $g(t)=U_{\rm ren}(t)=U_{\rm ren}(t,-\infty)$ in U_1 with the initial condition $g(-\infty)=1$ and $g(+\infty)=S$. The central extension of U_1 defines a natural connection in the circle bundle $\hat{U}_1\to U_1$. The curvature of this connection has a simple formula: It is simply given as $\mathrm{curv}(X,Y)=\omega(X,Y)$, where X,Y are tangent vectors at $g\in U_1$, identified through left translation on the group as elements of the Lie algebra of U_1. The phase of $\hat{S}$ is defined *through parallel transport*, with respect to the connection above.

Our definition of phase of $\hat{S}$ is *causal*: A scattering process in an external field A followed by the scattering in A' defines the same phase as the scattering for $A''=A\cup A'$, when both A,A' have finite nonoverlapping support in time and A'' denotes a field $A''(t)=A(t)$ for $t<t_0$ and $A''(t)=A'(t)$ for $t>t_0$; here t_0 separates the supports of A,A'. The operator T_A is not uniquely defined. It is more convenient to define the transformation first in the Schrödinger picture (1). A simple formula that works is (here $\not{A}=\gamma^0\gamma^k A_k$ and $\not{E}=\partial_t\not{A}-\not{\partial}A_0+[A_0,\not{A}]$)

$$T_A=\exp\left(\frac{1}{4}\left[\frac{1}{D_0},\not{A}\right]-\frac{1}{8}\left[\frac{1}{D_0}\not{A}\frac{1}{D_0},\not{A}\right]-\frac{i}{4}\frac{1}{D_0}\not{E}\frac{1}{D_0}\right), \tag{4}$$

where it is understood that the singularity at the zero modes of D_0 is taken care of by an infrared regularization, for example, $1/D_0\to D_0/(D_0^2+\alpha^2)$ for some nonzero real number α. In the interaction picture one uses the operator $\exp(ith_0)T_A\exp(-ith_0)$. Note that this choice commutes with the chiral projection operators $P_\pm$.

For the proof of validity of the choice (4) it is convenient to use the symbol calculus for pseudodifferential operators. A PSDO B acting on vector-valued functions in $\mathbb{R}^d$ is represented by a smooth matrix-valued function $b(x,p)$ of coordinates x_i and momenta p_i. The operators that we need have an asymptotic expansion,

$$b(x,p)\sim b_k+b_{k-1}+b_{k-2}+\cdots,$$

where each b_j is a smooth function of x,p for $|p|>1$, homogeneous of degree j in the momenta. The asymptotic expansion for the product of two operators A,B is obtained from the formula

$$a*b=\sum\frac{(-i)^{|m|}}{m!}\frac{\partial^m a}{\partial x^m}\frac{\partial^m b}{\partial p^m},$$

where the multi-index notation $m=(m_1,m_2,\ldots,m_d)$ is used; $|m|=m_1+m_2+\cdots m_d$, $m!=m_1!m_2!\cdots m_d!$, etc. In particular, the degree of the highest term in the product is the sum of degrees of the factors. Each differentiation in momentum space lowers the degree of the operator by one.

If A is an operator with symbol a such that $a(x,p)\to 0$ faster than $|x|^{-d}$ as $|x|\to\infty$ and $\deg a<-d$ then A is a trace-class operator and

$$\mathrm{tr}\,A=\frac{1}{(2\pi)^d}\int \mathrm{tr}\,a(x,p)dx\,dp.$$

We shall now assume that the components A_μ of the vector potential satisfy the above boundary condition at $|x|\to\infty$.

The time evolution equation for $U_{\rm ren}(t,t_0)=T_A(t)U(t,t_0)T_A(t_0)^{-1}$ in the Schrödinger picture is

$$i\,\partial_t U_{\rm ren}(t,t_0)=(h_0+W(t))U_{\rm ren}(t,t_0),$$

with

$$W(t)=(i\ \partial_t T_A)T_A^{-1}+T_A(h_0+V(t))T_A^{-1}-h_0. \tag{5}$$

Expanding the exponential and arranging terms according to powers of the inverse of momentum (i.e., of D_0,) one gets $[\epsilon,W]=R_1+R_2+\cdots$, where the dots denote terms that behave explicitly as $|D_0|^k$ with $k\leqslant -2$ for high momenta, and

$$R_1=\frac{1}{2}\frac{D_0}{|D_0|}\mathcal{A}-\frac{1}{2}\mathcal{A}\frac{D_0}{|D_0|}+\frac{1}{4}|D_0|\mathcal{A}\frac{1}{D_0}-\frac{1}{4}D_0\mathcal{A}\frac{1}{|D_0|}+\frac{1}{4}\frac{1}{|D_0|}\mathcal{A}D_0-\frac{1}{4}\frac{1}{D_0}\mathcal{A}|D_0|,$$

and the second term, which is quadratic in $\mathcal{A}$, is

$$R_2=\frac{1}{4}\frac{1}{D_0}[|D_0|,\mathcal{A}^2]\frac{1}{D_0}+\frac{1}{4}\left[\frac{1}{|D_0|},\mathcal{A}^2\right]+\frac{1}{8}D_0\left[\mathcal{A}\frac{1}{D_0}\mathcal{A},\frac{1}{|D_0|}\right]+\frac{1}{8}\left[\mathcal{A}\frac{1}{D_0}\mathcal{A},\frac{1}{|D_0|}\right]D_0$$
$$+\frac{1}{8}\frac{1}{D_0}\left[\mathcal{A}\frac{1}{D_0}\mathcal{A},|D_0|\right]+\frac{1}{8}\left[\frac{1}{D_0}\mathcal{A}\frac{1}{D_0},|D_0|\right]\frac{1}{D_0}.$$

Since the commutator $[|D_0|^k,\mathcal{A}]$ is of order $|D_0|^{k-1}$ in momenta, all terms in R_2 are actually of order -2 or less. In order to estimate R_1 we write it in equivalent form,

$$R_1=\frac{1}{4}\frac{D_0}{|D_0|}\left[[\mathcal{A},|D_0|],\frac{1}{|D_0|}\right]-\frac{1}{4}\left[[\mathcal{A},|D_0|],\frac{1}{|D_0|}\right]\frac{D_0}{|D_0|},$$

and observe both terms are of order -2 for the same reason as in the case of R_2. We have disregarded all the low-order terms because in three space dimensions the condition that a PSDO is Hilbert–Schmidt (which was required for canonical quantization) is precisely the requirement that the operator vanishes for high momenta faster than $|p|=|D_0|$ raised to power $-3/2$. Thus, after our renormalization (=conjugation by the time-dependent unitary operator T_A) the gauge interaction can be lifted to a finite operator in the fermionic Fock space.

The total renormalized Hamiltonian is now a sum of the free (unbounded) self-adjoint Hamiltonian and the bounded self-adjoint interaction. By the Kato–Rellich theorem[7] the total Hamiltonian is self-adjoint, and according to Stone's theorem it defines the time evolution as a strongly continuous unitary one-parameter group in the Fock space.

This method can be extended to other interactions as well, under very general conditions.[3] It has been also used to derive the chiral anomaly in the Hamiltonian framework.[2,3]

The curvature formula (3) is *nonlocal*. The computation of the trace involves space derivatives up to arbitrary high power because of the Green's function $1/D_0$ in the renormalization prescription. However, the curvature is equivalent in cohomology to a local formula. We use the residue formula,[8]

$$\text{Res } B=\frac{1}{(2\pi)^d}\int_{|p|=1}\text{tr } b_{-d}(x,p)d^dx\ d^{d-1}p.$$

It behaves like trace, $\text{Res}[A,B]=0$. Let $\mathbf{g}$ be the Lie algebra of bounded pseudodifferential operators satisfying the conditions (1) that the degree of the commutator $[\epsilon,X]$ is strictly less than $-d/2$ (here the space dimension $d=3$), (2) the symbol of X and its derivatives decrease faster than $|x|^{-d/2}$ as $|x|\to\infty$. We have then the following.

Theorem 1: *Define the following 2-cocycle on* $\mathbf{g}$*:*

$$\omega_{\text{loc}}(X,Y)=\text{Res }\epsilon[\log|p|,X]Y, \tag{6}$$

where $|p|$*, the length of three momentum, is the symbol of* $|D_0|$*. Then the difference* $\omega-\omega_{\text{loc}}$ *is a coboundary.*

For a proof see Ref. 9 (and Ref. 2 for the special case of renormalized current operators). Actually, in Ref. 9 it was assumed that the manifold is compact. This was needed for the convergence of the integral over x in the trace formula for PSDOs. Our boundary condition does the same job.

Without the sign ϵ this formula would give the Radul cocycle, which is a 2-cocycle on the algebra of *all* PSDOs on a compact manifold.[10] In quantum field theory it is important to keep the sign because this is the cocycle arising from normal ordering. The equivalence of the two cocycles means simply that if $\widetilde{X}=\hat{X}+\lambda(X)$ for a suitable complex linear form on the algebra of one-particle operators then

$$[\widetilde{X},\widetilde{Y}]=\widetilde{[X,Y]}+\omega_{\mathrm{loc}}(X,Y). \tag{7}$$

The great advantage of the residue formula is that it only depends on the term in the appropriate PSDO that is precisely of asymptotic degree $-d$ (here -3) when expanded asymptotically in powers of $1/|p|$. In operator products each derivative in configuration space is associated with a derivative in momentum space. Since a differentiation in momentum space decreases the homogeneous degree of the operator, it follows that a calculation of the residue can involve derivatives with respect to x_i's only up to a finite order.

III. GAUGE INDEPENDENCE

The prescription for the vacuum polarization phase in Sec. II is not gauge independent. Furthermore, if we change the renormalization operator T_A the phase will be changed, too. Of course, in the case of chiral fermions the gauge dependence is to be expected because of the chiral anomaly; this was discussed in Ref. 3.

In the case of Dirac fermions we expect on topological grounds that one should be able to construct counterterms in the Hamiltonian such that the gauge variance of the counterterms exactly balances the gauge variation of the phase, so that the total phase would be gauge invariant. We shall confirm this explicitly below in the case of infinitesimal variations. At the same time we shall also get rid of the renormalization-dependent part of the phase.

Suppose that we have constructed another renormalization operator T'_A that commutes with the chiral projection operators and with γ_0, such that each term in the asymptotic expansion,

$$T'_A=1+t_{-1}(A)+t_{-2}(A)+\cdots,$$

in homogeneous terms t_k of degree k in momenta, is a local differential polynomial in the components of the vector potential A. The condition that the renormalization commutes with γ_0 means that the action of T'_A on the left chiral sector is equivalent to the action on the right-hand spinors.

The change of phase of the quantum scattering is given by a parallel transport around a loop of time evolution operators obtained by first traveling the path $g(t)$ from the point $g(-\infty)=1$ to $g(\infty)=S$ defined by the renormalized time evolution due to T_A and then following backward in time the time evolution defined by the renormalization T'_A. But the logarithm of the parallel transport phase is equal to the integral of the curvature over a surface bounded by the loop.

In the following we shall use the local curvature formula (6). This has the advantage that the parallel transport around a closed loop is given by a space integral of a local differential expression in the vector potential. This is in the spirit of local quantum field theory: A change in the renormalization corresponds to a local counterterm in the Lagrangian.

Since all our operators, the (renormalized) Hamiltonian, the renormalization operator, and gauge transformations, are composed of a pair of operators (K_L,K_R) acting on the two chiral sectors, we can split the curvature into chiral constituents as follows. The left and right spinors are indentified through the unitary transformation γ_0 so that we can think of K_L and K_R as acting in the same space. Thus we may define $K_\pm=K_L\pm K_R$. Since the sign ϵ is odd, $\epsilon_L=-\epsilon_R$, we can write

$$2\omega(X,Y)=\omega_L(X_-,Y_+)+\omega_L(X_+,Y_-).$$

Here $l=\log|p|$, $|p|$ is the symbol of $|h_0|$, and $\omega_L(X;Y)=\mathrm{Res}\ \epsilon_L[l,X_L]Y_L$. We want to show that this is equivalent to a 2-form ω_- such that $\omega_-(X,Y)=0$ if either $X_-=0$ or $Y_-=0$. We denote $\mathrm{Res}'\ X=\mathrm{Res}\ \epsilon_L X$.

Let GL_1 denote the group of all invertible bounded PSDOs g in H such that the degree of $[\epsilon,g]$ is strictly less than $-d/2$. The Lie algebra of GL_1 is then equal to $\mathbf{g}$. Let $GL_{1,c}$ be the subgroup of GL_1 consisting of the chirally diagonal operators $f=(f_L,f_R)$.

The time evolution operator in Schrödinger picture does not have a good asymptotic expansion in inverse powers of momenta and therefore it is not immediately clear that one can define the Wodzicki residue for a product involving $U(t)$ as a factor. For example, the free time evolution $\exp(it\not{\partial})$ contains arbitrary high powers of momenta and therefore it is not a classical PSDO. However, the time evolution $U_I(t)=U_I(t,-\infty)$ in the interaction picture has a well-behaved symbol. The leading terms (in powers of momenta) are obtained from the Dyson expansion in which the individual terms are composed of products of the Green's function of the Dirac operator and the interaction Hamiltonian. This gives directly a classical asymptotic expansion for the PSDO $U_I(t)$.

In the following we deal with the ''Maurer–Cartan'' form $\theta=[l,f]f^{-1}$. This is a classical PSDO when $f=U_I(t)$. In the case of the Schrödinger picture it is replaced by $\exp(ith_0)\theta\times\exp(-ith_0)$. The computation of residues for this type of operators involves oscillatory integrals. These can be handled by Guillemin's method for Fourier integral operators.[11] For this reason the curvature formula extends to the bigger group GL_F, which is the smallest group containing GL_1, and the free time evolution operators $\exp(ith_0)$.

Lemma: The restriction of the curvature ω_{loc} to the chirally diagonal subgroup $GL_{F,c}\subset GL_F$ is equivalent in de Rham cohomology to the 2-form

$$\omega'(X,Y)=\omega_-(X_-,Y_-)=-\mathrm{Res}'[l,X_-]Y_-+\mathrm{Res}'[l,f_L]f_L^{-1}[X_-,Y_-].$$

Proof: Define the 1-form,

$$\phi(f;X)=\mathrm{Res}'[l,f_L]f_L^{-1}X_-,$$

on the group $GL_{F,c}$. The exterior derivative of this is computed through

$$(d\phi)(X,Y)=\mathscr{L}_X\phi(f;Y)-\mathscr{L}_Y\phi(f;X)-\phi(f;[X,Y]),$$

using the definition of the left action, $\mathscr{L}_XF(f)=(\partial/\partial t)F(e^{-tX}f)|_{t=0}$. The result is

$$(d\phi)(X,Y)=-\mathrm{Res}'[l,X_-]Y_--\mathrm{Res}'[l,X_L]Y_L+\mathrm{Res}'[l,X_R]Y_R+\mathrm{Res}'[l,f_L]f_L^{-1}[X_-,Y_-].$$

Adding the right-hand side to $\omega(X,Y)$ gives $\omega_-(X,Y)$.

We denote by $q(R)$ the canonical second quantization of an operator R and $W=W(A,T)$ is the renormalized interaction (5).

Theorem 2: *The modified quantum Hamiltonian,*

$$q'(h_0+W)=q(h_0+W)+\phi(U(A;t);h_0+W),$$

has the property that the phase of the quantum scattering matrix is invariant under infinitesimal gauge transformations. Moreover, the phase is invariant with respect to all infinitesimal variations X of T_A such that (1) X is a PSDO in the Lie algebra of U_1, (2) X is chirally even, i.e., it commutes with γ_0 and Γ.

Proof: The second term on the right-hand side means that we compute the quantum phase using a modified connection, corresponding to adding to the old connection (with curvature ω_{loc}) the 1-form ϕ. In general, a variation of the parallel transport phase defined by a connection θ on any manifold is given by

$$\mathscr{L}_X\int\theta(\gamma'(t))dt=\int\Omega_{g(t)}(\gamma'(t),X)dt,$$

where Ω is the curvature form and $X(t)$ is a vector field along the curve $\gamma(t)$. Applying this to the case when $\gamma(t)=U(t)$ [the tangent vector $\gamma'(t)$ is replaced by the element $h_0+W=iU'(t)U(t)^{-1}$], and X is an infinitesimal gauge transformation, shows that at time t the infinitesimal gauge variation of the phase is

$$\mathscr{L}_X(\text{phase}) = \omega'(h_0 + W, X). \tag{8}$$

But now $X_- = 0$ and therefore this expression vanishes and thus also the total gauge variation of the quantum phase as an integral with respect to t of (8).

Remark: We started from the standard minimal coupling of fermions to external gauge fields. We performed a renormalization by the field-dependent unitary operator T_A. The new interaction Hamiltonian, although nonlocal, is still unitarily equivalent to the original local interaction. There is a very natural generalization of our discussion in the spirit of noncommutative geometry.[12] The original interaction $\gamma^\mu A_\mu$ could be replaced by any bounded pseudodifferential operator without affecting our results (or even by a larger class of operators; see Ref. 3). This would lead to a Hamiltonian counterpart of the noncommutative Yang–Mills action functional. It would be an intermediate step between the ordinary Yang–Mills theory and the abstract universal Yang–Mills theory proposed by Rajeev.[13]

[1] J. Palmer, ''Scattering automorphisms of the Dirac field,'' J. Math. Anal. Appl. **64**, 189–215 (1978); S. N. M. Ruijsenaars, ''Charged particles in external fields; I, classical theory,'' J. Math. Phys. **18**, 720–737 (1976); ''Gauge invariance and implementability of the *S*-operator for spin-0 and spin-1/2 particles in time-dependent classical fields,'' J. Funct. Anal. **33**, 47–57 (1979).

[2] J. Mickelsson, ''Wodzicki residue and anomalies of current algebras,'' *Integrable Systems and Strings*, Springer Lecture Notes in Physics, edited by A. Alekseev, A. Hietamäki, K. Huitu, A. Morozov, and A. Niemi, 1994, Vol. 436, pp. 123–135; ''Regularization of current algebra,'' *Constraint Theory and Quantization Methods*, edited by F. Colomo, L. Lusanna, and G. Marmo (World Scientific, Singapore, 1994), pp. 72–79; ''Renormalization of current algebra,'' *Generalized Symmetries in Physics*, edited by H.-D. Doebner, V. K. Dobrev, and A. G. Ushveridze (World Scientific, Singapore, 1994), pp. 328–337.

[3] E. Langmann and J. Mickelsson, ''Scattering matrix in external field problems,'' J. Math. Phys. **37**, 3933–3953 (1996).

[4] H. Araki, ''Bogoliubov automorphisms and Fock representations of canonical anticommutation relations,'' in *Contemporary Mathematics* (American Mathematical Society, Providence, 1987), Vol. 62, pp. 23–141.

[5] L.-E. Lundberg, ''Quasi-free 'second quantization','' Commun. Math. Phys. **50**, 103–112 (1976).

[6] A. Pressley and G. Segal, *Loop Groups* (Clarendon, Oxford, 1986).

[7] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II* (Academic, London, 1975), p. 162.

[8] M. Wodzicki, ''Noncommutative residue,'' in *K-Theory, Arithmetic and Geometry*, Lecture Notes in Mathematics, edited by Yu. I. Manin (Springer-Verlag, Berlin, 1987), Vol. 1289, pp. 320–399; ''Local invariants of spectral asymmetry,'' Invt. Math. **75**, 143–177 (1984).

[9] M. Cederwall, G. Ferretti, B. E. W. Nilsson, and A. Westerberg, ''Schwinger terms and cohomology of pseudodifferential operators,'' Commun. Math. Phys. **175**, 203–220 (1996).

[10] A. O. Radul, ''Lie algebra of differential operators, their central extensions, and $\mathscr{W}$-algebras,'' Funct. Anal. Appl. **25**, 33–49 (1991).

[11] V. Guillemin, ''Residue traces for certain algebras of Fourier integral operators,'' J. Funct. Anal. **115**, 391–417 (1993); ''Gauged Lagrangian distributions,'' Adv. Math. **102**, 184–201 (1993).

[12] A. Connes, *Noncommutative Geometry* (Academic, London, 1994).

[13] S. G. Rajeev, ''Universal gauge theory,'' Phys. Rev. D **42**, 2779–2791 (1990).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Ladder operators for isospectral oscillators

S. Seshadri, V. Balakrishnan, and S. Lakshmibala
Department of Physics, Indian Institute of Technology, Madras 600 036, India

(Received 10 June 1997; accepted for publication 9 October 1997)

We present, for the isospectral family of oscillator Hamiltonians, a systematic procedure for constructing raising and lowering operators satisfying any prescribed ''distorted'' Heisenberg algebra (including the q-generalization). This is done by means of an operator transformation implemented by a shift operator. The latter is obtained by solving an appropriate partial isometry condition in the Hilbert space. Formal representations of the nonlocal operators concerned are given in terms of pseudo-differential operators. Using the new annihilation operators, new classes of coherent states are constructed for isospectral oscillator Hamiltonians. The corresponding Fock–Bargmann representations are also considered, with specific reference to the order of the entire function family in each case. © *1998 American Institute of Physics.* [S0022-2488(98)03301-5]

I. INTRODUCTION

Generalized coherent states of different kinds have been defined and investigated for some years now, particularly in quantum optics and quantum measurement theory.[1–6] Recently, two different kinds of coherent states (CS) have been constructed[7,8] in relation to a one-parameter (λ) family of Hamiltonians $\tilde{H}$ that are isospectral companions of the standard oscillator Hamiltonian $H=a^{\dagger}a$. (The latter corresponds to the limit $|\lambda|\to\infty$.) For each finite value of λ, the eigenstates of $\tilde{H}$ consist of a zero-energy state $|\theta_0\rangle$ and a set of states $\{|\theta_n\rangle\}$ with eigenvalues $n=1,2,...$, the *latter* being obtained from the oscillator states $|n\rangle$ ($n=0,1,...$) by a nonunitary transformation [see Eq. (2.5) below]. Fernandez *et al.*[7] have constructed coherent states that are eigenstates of an annihilation operator associated with $\{|\theta_n\rangle, n\geqslant 1\}$. However, these states are not coherent states in the group-theoretic or Perelomov sense,[9] since they are not obtained by the action of a unitary ''displacement'' operator on the base state $|\theta_1\rangle$. On the other hand, Kumar and Khare[8] have shown that $\tilde{H}$ itself is essentially unitarily equivalent to H. This fact enables them to obtain coherent states simply by a unitary transformation of the standard oscillator CS. These states are thus coherent states in the group-theoretic sense in addition to being annihilation operator eigenstates. As may be anticipated, they reduce to the standard oscillator CS in the limit $|\lambda|\to\infty$. Subsequently, Fernandez *et al.*[10] have generalized their earlier work[7] by introducing a real parameter w as follows: the commutator of the lowering and raising operators in the $\{|\theta_n\rangle, n\geqslant 0\}$ basis is a diagonal operator given by diag $(0,w,1,1,1,...)$. This has been termed a ''distorted'' Heisenberg algebra. In the special case $w=1$, the CS constructed are also CS in the group-theoretic sense.

A number of interesting problems arise, which we address in this paper: the generalization of the algebra to encompass an arbitrary diagonal operator as the basic commutator, including the corresponding q-generalization; the development of a systematic procedure to find raising and lowering operators in this case; and the construction of the corresponding CS (which are *not* unitarily equivalent to the standard oscillator CS). Our strategy is to find a suitable operator transformation that takes the original ladder operators to the required new ones. The transformation is implemented by means of a *shift operator* that is constructed so as to satisfy appropriate partial isometry conditions dictated by the algebra imposed. We also derive representations for the new raising and lowering operators in terms of pseudo-differential operators, to enable one to compare their structure explicitly with their original counterparts. Finally, we examine the Fock–Bargmann (or entire analytic function) representation of states based on the CS obtained in various cases, with particular reference to the order of the entire function families concerned.

 © 1998 American Institute of Physics

II. NOTATION AND REVIEW

For ready reference, we recapitulate briefly the relevant parts of Refs. 7, 8, and 10. Given $H=a^\dagger a$ with $[a,a^\dagger]=1$, we have $H|n\rangle=n|n\rangle$ $(n=0,1,...)$. In the position representation, $a=2^{-1/2}(x+d)$ and $a^\dagger=2^{-1/2}(x-d)$, where $d=d/dx$. Operators b and $b^\dagger$ are then defined, given in the position representation by

$$b=2^{-1/2}(x+d+\phi(x)), \quad b^\dagger=2^{-1/2}(x-d+\phi(x)), \tag{2.1}$$

and the condition $bb^\dagger=aa^\dagger$ imposed. This implies that the function $\phi(x)$ satisfies the Riccati equation

$$\phi'+2x\phi+\phi^2=0, \tag{2.2}$$

where $\phi'=d\phi/dx$. Equation (2.2) has the one-parameter family of solutions

$$\phi_\lambda(x)=\frac{e^{-x^2}}{(\lambda+\int_0^x dy e^{-y^2})}, \tag{2.3}$$

where λ is a real number such that $|\lambda|>\sqrt{\pi}/2$ (to avoid a singular ϕ_λ). Now consider the Hamiltonian (actually, a family of Hamiltonians, parametrized by λ)

$$\tilde{H}=b^\dagger b. \tag{2.4}$$

In order to keep the notation simple, the λ dependence of b, $b^\dagger$, and $\tilde{H}$ has not been indicated explicitly. Although $bb^\dagger=aa^\dagger$ (that is, the supersymmetric partners of H and $\tilde{H}$ are identical to each other), $b^\dagger b\neq a^\dagger a$. It is straightforward to check that the states

$$|\theta_n\rangle=n^{-1/2}b^\dagger|n-1\rangle (n=1,2,...) \tag{2.5}$$

are eigenstates of the Hamiltonian $\tilde{H}$ with eigenvalues $n=1,2,...$. The *ground state* $|\theta_0\rangle$ of $\tilde{H}$ is obtained by requiring that

$$b|\theta_0\rangle=0. \tag{2.6}$$

The eigenstates $\{|\theta_n\rangle, n=0,1,...\}$ form a complete orthonormal set of states in the Hilbert space $\mathscr{H}$ spanned by the Fock basis $\{|n\rangle\}$. It will be useful to write $\mathscr{H}_0=\text{span}|0\rangle$, $\mathscr{H}_1=\text{span}\{|n\rangle, n\geq 1\}$, $\tilde{\mathscr{H}}_0=\text{span}|\theta_0\rangle$, and $\tilde{\mathscr{H}}_1=\text{span}\{|\theta_n\rangle, n\geq 1\}$, so that $\mathscr{H}=\mathscr{H}_0+\mathscr{H}_1=\tilde{\mathscr{H}}_0+\tilde{\mathscr{H}}_1$. (For easy identification, we shall use a tilde for spaces and operators pertaining to the isospectral case and drop the tilde for their counterparts relating to the standard oscillator.) It is evident that H and $\tilde{H}$ are isospectral, and that $\tilde{H}\to H$ in the limit $|\lambda|\to\infty$. However, since $b^\dagger b\neq a^\dagger a$, the commutator $[b, b^\dagger]\neq 1$. Thus $b^\dagger$ and b are *not* the appropriate raising and lowering operators for $\tilde{H}$. Fernandez *et al.*[7] consider the lowering operator (in their notation)

$$A=b^\dagger a b, \tag{2.7}$$

which annihilates not only $|\theta_0\rangle$ but also $|\theta_1\rangle$. One finds

$$A|\theta_0\rangle=0, \quad A|\theta_n\rangle=(n-1)n^{1/2}|\theta_{n-1}\rangle. \tag{2.8}$$

Similarly,

$$A^\dagger|\theta_0\rangle=0, \quad A^\dagger|\theta_n\rangle=n(n+1)^{1/2}|\theta_{n+1}\rangle. \tag{2.9}$$

These relations help us understand how the Hilbert space $\mathscr{H}$ breaks up in a natural fashion into the direct sum $\tilde{H}_0+\tilde{H}_1$. As A is an annihilation operator in $\tilde{H}_1$, its eigenstates, satisfying $A|z;\theta_1\rangle=z|z;\theta_1\rangle (z\in\mathscr{C})$, may be regarded as coherent states. They are given by[7]

$$|z;\theta_1\rangle = (\text{const}) \sum_{n=0}^{\infty} \frac{z^n}{n!\sqrt{(n+1)!}} |\theta_{n+1}\rangle. \tag{2.10}$$

However, since $[A, A^\dagger] \neq 1$ even in the subspace $\tilde{H}_1$, the state $|z;\theta_1\rangle$ cannot be obtained by a unitary transformation of the form $\exp(zA^\dagger - \overline{z}A)$ acting on the ground state $|\theta_1\rangle$. Therefore $|z;\theta_1\rangle$ is not a CS in the group-theoretic sense, nor can *generalized* coherent states be built up from it by the standard procedure—namely, by using $|z:\theta_1\rangle$ as the ground state of an appropriate Hamiltonian obtained by a unitary transformation of $\tilde{H}$.

In contrast to the foregoing, Kumar and Khare[8] have shown that b and $b^\dagger$ can be written in the form

$$b = aU^\dagger, \quad b^\dagger = Ua^\dagger, \tag{2.11}$$

where U is unitary ($UU^\dagger = U^\dagger U = 1$) in the full space $\mathscr{H}$, and formally determined the matrix elements of U in the basis $\{|n\rangle\}$. Therefore $bb^\dagger = aa^\dagger$, as required. Further, $\tilde{H} = b^\dagger b = UHU^\dagger$, so that $\tilde{H}$ and H are actually unitarily equivalent. Writing $\tilde{a} = UaU^\dagger$, one has $\tilde{H} = \tilde{a}^\dagger \tilde{a}$ where $[\tilde{a}, \tilde{a}^\dagger] = 1$. The eigenstates $|\theta_n\rangle$ of $\tilde{H}$ are just the unitary transforms of those of H, i.e.,

$$|\theta_n\rangle = U|n\rangle \quad \text{and} \quad U^\dagger|\theta_n\rangle = |n\rangle, \quad n = 0,1,\ldots. \tag{2.12}$$

The operator $\tilde{a}$ only annihilates the vacuum $|\theta_0\rangle$. The state

$$|\alpha;\theta_0\rangle = \exp\left(-\frac{|\alpha|^2}{2}\right) \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |\theta_n\rangle \quad (\alpha \in \mathscr{C}) \tag{2.13}$$

is both an annihilation operator eigenstate and a CS in the Perelomov sense. It is readily verified that (i) $|\alpha;\theta_0\rangle = U|\alpha\rangle$ where $|\alpha\rangle$ is the standard oscillator CS; (ii) $|\alpha;\theta_0\rangle$ is obtainable by a unitary displacement operator acting on $|\theta_0\rangle$; and (iii) $|\alpha;\theta_0\rangle \to |\alpha\rangle$ in the limit $|\lambda| \to \infty$. Generalized CS may be constructed from $|\alpha;\theta_0\rangle$ in essentially the same manner as in the case of $|\alpha\rangle$. Viewed in this manner, the isospectral case is simply a unitary transformation of the standard oscillator problem, and nothing new is gained by considering it in preference to the latter.

In order to have ladder operators in $\tilde{\mathscr{H}}_1$ whose commutator is equal to unity, and also to generalize this algebra to some extent, Fernandez *et al.* have subsequently defined[10] a new lowering operator C_w (in their notation) according to

$$C_w = b^\dagger f(H) ab. \tag{2.14}$$

Here, the form of $f(H)$ is deduced from the requirement that the commutator $[C_w, C_w^\dagger]$ be equal to diag $(w,1,1,\ldots)$ in $\tilde{\mathscr{H}}_1$, where w is any positive number. (The operators C_w and $C_w^\dagger$ continue to annihilate the subspace $\tilde{\mathscr{H}}_0$). It is found that

$$f(H) = \frac{1}{1+H}\left(\frac{w+H}{2+H}\right)^{1/2}. \tag{2.15}$$

It is evident that diag $(w,1,1,\ldots)$ is not the most general diagonal form possible for the basic commutator, although it can be shown to be the only consistent one once the specific form given by Eq. (2.14) is assumed for the lowering operator. As we shall see, a procedure of wider applicability is to look for an operator transformation on $\tilde{a}$ of the type $\tilde{S}^\dagger \tilde{a} \tilde{S}$. In the next section, we develop a systematic procedure to find the transformation $\tilde{S}$, and from it the ladder operators *for any prescribed diagonal form of their commutator*, including the case of the q-generalization of the isospectral oscillator.

III. SHIFT OPERATORS AND OPERATOR TRANSFORMATIONS

We wish to implement transformations on the ladder operators $\tilde{a}$ and $\tilde{a}^\dagger$ so as to obtain a lowering operator

$$\tilde{a}_1 = \tilde{S}^\dagger \tilde{a} \tilde{S} \tag{3.1}$$

and a Hermitian conjugate raising operator

$$\tilde{a}_1^\dagger = \tilde{S}^\dagger \tilde{a}^\dagger \tilde{S} \tag{3.2}$$

with the following properties: (i) both operators must annihilate $|\theta_0\rangle$ (so that $\tilde{\mathscr{H}}_0$ is the kernel of the transformation) and (ii) their commutator is a prescribed diagonal operator in $\tilde{\mathscr{H}}_1$. For the sake of clarity, we solve the problem first in the standard Fock basis, and then use the operator U to transform to the isospectral case.

Accordingly, we first seek to eliminate the subspace $\mathscr{H}_0$ from $\mathscr{H}$, and to find conjugate lowering and raising operators in $\mathscr{H}_1$ given by

$$a_1 = S^\dagger a S, \quad a_1^\dagger = S^\dagger a^\dagger S, \tag{3.3}$$

whose commutator is equal to *unity* throughout the latter space. For this we need to find an operator S in $\mathscr{H}$ that satisfies the *partial isometry*

$$SS^\dagger = 1, \quad S^\dagger S = 1 - |0\rangle\langle 0|. \tag{3.4}$$

It is easy to see that this is solved by the *shift* operator

$$S = \sum_{n=0}^{\infty} |n\rangle\langle n+1|. \tag{3.5}$$

Equations (3.3) then yield

$$a_1 = \sum_{n=1}^{\infty} n^{1/2} |n\rangle\langle n+1|, \quad a_1^\dagger = \sum_{n=1}^{\infty} n^{1/2} |n+1\rangle\langle n|. \tag{3.6}$$

Therefore $[a_1,\ a_1^\dagger] = 1 - |0\rangle\langle 0| = \mathrm{diag}\,(0,1,1,\ldots)$, as required. It is easily verified that

$$a_1|0\rangle = a_1|1\rangle = 0, \quad a_1|n\rangle = (n-1)^{1/2}|n-1\rangle, \quad n \geqslant 1, \tag{3.7}$$

and

$$a_1^\dagger|0\rangle = 0, \quad a_1^\dagger|n\rangle = n^{1/2}|n+1\rangle, \quad n \geqslant 1. \tag{3.8}$$

Further, we note that the operator S has the closed form $(1+H)^{-1/2}a$ in $\mathscr{H}$. Hence we may also write

$$a_1 = a^\dagger (1+H)^{-1/2} a (1+H)^{-1/2} a, \quad a_1^\dagger = a^\dagger (1+H)^{-1/2} a^\dagger (1+H)^{-1/2} a. \tag{3.9}$$

Next, consider the *general* ''distorted'' Heisenberg algebra in which the commutator is specified (in the Fock basis) to be

$$[a_1,\ a_1^\dagger] = \mathrm{diag}(0, w_1, w_2, \ldots) = \sum_{n=1}^{\infty} w_n |n\rangle\langle n|, \tag{3.10}$$

where $\{w_n\}$ are real positive constants. Now S is found by generalizing the partial isometry condition, Eq. (3.4), to read

$$SS^\dagger = \sum_{n=0}^{\infty} c_n |n\rangle\langle n|, \quad S^\dagger S = \sum_{n=0}^{\infty} c_n |n+1\rangle\langle n+1|, \tag{3.11}$$

where the coefficients c_n are to be determined. The deviation of the c_n's from unity is a measure of the extent to which the shift operator deviates from the exact partial isometry specified in Eqs. (3.4). Solving for S and $S^\dagger$ in the form

$$S=\sum_{n=0}^{\infty} c_n^{1/2}|n\rangle\langle n+1|, \quad S^\dagger=\sum_{n=0}^{\infty} c_n^{1/2}|n+1\rangle\langle n| \tag{3.12}$$

and using Eqs. (3.3) and (3.10), we obtain the recursion relations

$$c_0c_1=w_1, \quad (n+1)c_nc_{n+1}-nc_nc_{n-1}=w_{n+1}, \quad n=1,2,3,\ldots . \tag{3.13}$$

We may choose $c_0=1$ without loss of generality, and solve the recursion relations to obtain

$$c_n=\frac{(n-1)!!}{n!!}\,\frac{(w_1+w_2+\cdots+w_n)!!}{(w_1+w_2+\cdots+w_{n-1})!!}. \tag{3.14}$$

Here the notation $(w_1+w_2+\ldots+w_n)!!$ stands for the product

$$(w_1+w_2+\cdots+w_n)\times(w_1+w_2+\cdots+w_{n-2})\cdots .$$

Substituting the values for c_n given above in Eq. (3.12) yields S and $S^\dagger$. The corresponding expressions for a_1 and $a_1^\dagger$ are then obtained from Eqs. (3.3), and their commutator is guaranteed to satisfy Eq. (3.10).

We are ready now to go over to the isospectral case. As the corresponding shift and ladder operators are merely the unitary transformations

$$\tilde{S}=USU^\dagger, \quad \tilde{a}_1=Ua_1U^\dagger, \quad \tilde{a}_1^\dagger=Ua_1^\dagger U^\dagger, \tag{3.15}$$

we find

$$\tilde{S}=\sum_{n=0}^{\infty} c_n^{1/2}|\theta_n\rangle\langle\theta_{n+1}|, \tag{3.16}$$

where $c_0=1$ as already stated, and c_n is given by Eq. (3.14). We thus obtain the solutions

$$\tilde{a}_1=\sum_{n=1}^{\infty} (w_1+w_2+\cdots+w_n)^{1/2}|\theta_n\rangle\langle\theta_{n+1}|, \tag{3.17}$$

$$\tilde{a}_1^\dagger=\sum_{n=1}^{\infty} (w_1+w_2+\cdots+w_n)^{1/2}|\theta_{n+1}\rangle\langle\theta_n| \tag{3.18}$$

for the ladder operators in $\tilde{\mathscr{H}}$. It is easily verified from these expressions that $[\tilde{a}_1, \tilde{a}_1^\dagger]$ is diag $(0,w_1,w_2,\cdots)$ in the *transformed* basis, i.e.,

$$[\tilde{a}_1,\tilde{a}_1^\dagger]=\sum_{n=1}^{\infty} w_n|\theta_n\rangle\langle\theta_n|, \tag{3.19}$$

as required [cf. Eq. (3.10)]. The action of $\tilde{a}_1$ and $\tilde{a}_1^\dagger$ on $|\theta_n\rangle$ is given by

$$\tilde{a}_1|\theta_0\rangle=\tilde{a}_1|\theta_1\rangle=0, \quad \tilde{a}_1|\theta_n\rangle=(w_1+w_2+\cdots+w_{n-1})^{1/2}|\theta_{n-1}\rangle, \quad n\geqslant 2, \tag{3.20}$$

and

$$\tilde{a}_1^\dagger|\theta_0\rangle=0, \quad \tilde{a}_1^\dagger|\theta_n\rangle=(w_1+w_2+\cdots+w_n)^{1/2}|\theta_{n+1}\rangle, \quad n\geqslant 1. \tag{3.21}$$

For ready reference, the various lowering operators involved and the transformations connecting them are indicated schematically in Fig. 1.

In order to see how our general expressions simplify in particular cases, it is convenient to use Eq. (2.5) and the usual representation of a in the Fock basis to express Eq. (3.17) in the form

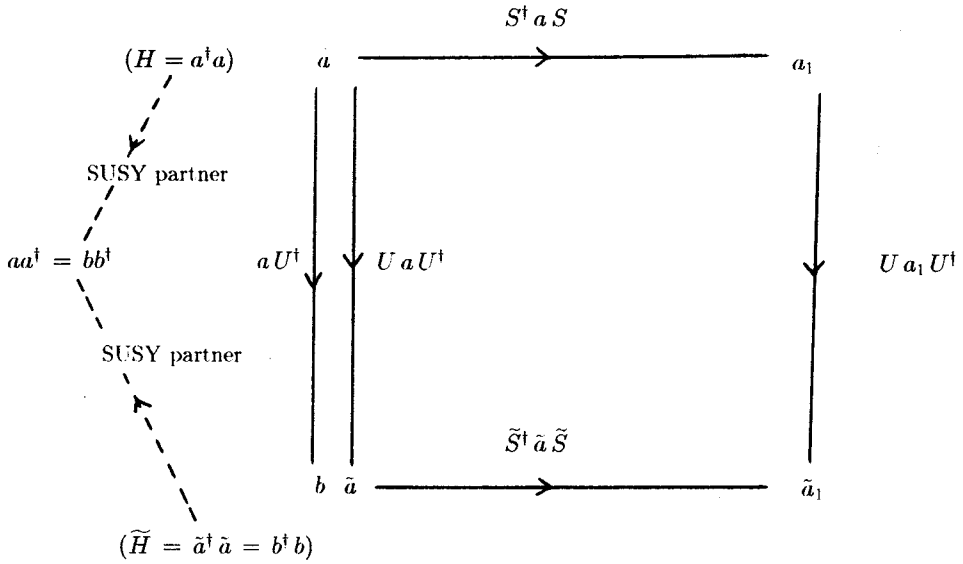

FIG. 1. The various transformed versions of the lowering operator a. S is the general shift operator given by Eq. (3.12) while U is the unitary transformation connecting the Fock basis $\{|n\rangle\}$ to the isospectral basis $\{|\theta_n\rangle\}$. $\tilde{S}$ stands for $USU^\dagger$.

$$\tilde{a}_1 = b^\dagger \sum_{n=0}^{\infty} \left(\frac{W_{n+1}}{(n+1)(n+2)} \right)^{1/2} |n\rangle\langle n+1|b, \tag{3.22}$$

or in the equivalent form

$$\tilde{a}_1 = b^\dagger (1+H)^{-1/2} \left\{ \sum_{n=0}^{\infty} W_{n+1}^{1/2} |n\rangle\langle n+1| \right\} (1+H)^{-1/2} b, \tag{3.23}$$

where we have written

$$W_n = w_1 + w_2 + \cdots + w_n \tag{3.24}$$

for brevity. It is clear from the above that, whenever $W_{n+1}^{1/2}$ can be written as some function of n, say $F(n)$, we obtain the closed-form expression

$$\tilde{a}_1 = b^\dagger (1+H)^{-1} f(H) a (1+H)^{-1/2} b. \tag{3.25}$$

If, further, $F(n)$ is a product of the form $(n+1)^{1/2} G(n) G(n+1)$, we obtain the *symmetric* form

$$\tilde{a}_1 = b^\dagger (1+H)^{-1/2} G(H) a G(H) (1+H)^{-1/2} b. \tag{3.26}$$

Corresponding expressions obtain, of course, for $a_1^\dagger$ as well. A remark is in order here on the square-root operator occurring in these expressions. As $(1+H)^{-1}$ is a bounded positive definite operator in $\mathscr{H}$, it has a unique positive square root in that space, by virtue of the square-root lemma.[11] Following the procedure of Ref. 12, this can be defined rigorously in terms of the resolvent operator $(\xi + 1 + H)^{-1}$ as

$$(1+H)^{-1/2} = \frac{1}{\pi} \int_0^\infty d\xi \, \xi^{-1/2} (1+\xi+H)^{-1}. \tag{3.27}$$

We are now ready to read off various special cases of the foregoing.

(i) First of all, if we set $w_n = w$ for *all* $n \geq 1$, the operator in curly brackets in Eq. (3.23) above is simply $w^{1/2} a$, and we get the symmetric expression

$$\tilde{a}_1 = w^{1/2} b^\dagger (1+H)^{-1/2} a (1+H)^{-1/2} b. \tag{3.28}$$

(ii) Next, we note that the ''distorted'' Heisenberg algebra introduced in Ref. 10 corresponds to setting $w_1=w$ and $w_n=1$ for $n\geqslant 2$, so that $W_{n+1}=(w+n)$. Therefore $\tilde{a}_1$ reduces, in this special case, to a closed-form expression— namely, the operator C_w defined in Eqs. (2.14) and (2.15).

(iii) Again, if we choose $w_n=n$ itself—i.e., we have the algebra

$$[\tilde{a}_1,\ \tilde{a}_1^\dagger]=UHU^\dagger=\tilde{H}=b^\dagger b, \tag{3.29}$$

we obtain

$$\tilde{a}_1=2^{-1/2}b^\dagger(1+H)^{-1/2}ab. \tag{3.30}$$

(iv) On the other hand, if we set $w_1=w$ and $w_n=0$ for $n\geqslant 2$, we obtain

$$\tilde{a}_1=w^{1/2}b^\dagger(1+H)^{-1}a(1+H)^{-1/2}b. \tag{3.31}$$

(v) Finally, the q-generalization[13] of the isospectral case is immediately found by setting $w_n=q^n$: this implies that

$$[\tilde{a}_1,\ \tilde{a}_1^\dagger]=q^{\tilde{H}}, \tag{3.32}$$

which is a transformed, equivalent, version of a q-commutation relation between a conjugate pair of ladder operators. We obtain in this case the result

$$\tilde{a}_1=q^{1/2}b^\dagger(1+H)^{-1}\left(\frac{1-q^{1+H}}{1-q}\right)^{1/2}a(1+H)^{-1/2}b. \tag{3.33}$$

It is readily seen that Eq. (3.33) reduces to Eq. (3.28) (with $w=1$) in the limit $q\to 1$.

IV. REPRESENTATION IN TERMS OF PSEUDO-DIFFERENTIAL OPERATORS

The expressions obtained in the preceding section involve various nonlocal operators, for which it is appropriate to exhibit suitable explicit representations. Kernels in the position basis for the integral operators represented by $\tilde{a}_1$ and $\tilde{a}_1^\dagger$ may of course be written down at once from Eqs. (3.17) and (3.18) in terms of the wave function $\tilde{\psi}_n(x)=\langle x|\,\theta_n\rangle$ and its complex conjugate. *Formal* representations for the operators concerned in a form that is an extension of the standard ones $a=2^{-1/2}(x+d)$, $a^\dagger=2^{-1/2}(x-d)$, may be given in terms of pseudo-differential operators,[14,15] i.e., operators of the form

$$\sum_{n=0}^{N}u_n(x)d^n+\sum_{n=0}^{\infty}u_{-n}(x)d^{-n},$$

where N is a finite positive integer (and $d\equiv d/dx$). Here d^{-1} is the antiderivative, defined recursively through

$$d^{-1}(f\cdot)=\sum_{n=0}^{\infty}(-1)^n f^{(n)}d^{-1-n}(\cdot), \tag{4.1}$$

where $f^{(n)}$ is the nth derivative of $f(x)$. This leads to the basic relationship

$$[d^{-r},\ f]=\sum_{n=1}^{\infty}(-1)^n\binom{n+r-1}{n-1}f^{(n)}d^{-n-r}. \tag{4.2}$$

With the help of the foregoing we can find expansions for the operators $(d^2+u(x))^{1/2}$ and $(d^2+u(x))^{-1/2}$ in powers of d^{-1}. In particular, we find

$$(1+H)^{-1/2}=-2^{1/2}i(d^2-x^2-1)^{-1/2}=-2^{1/2}i\{d^{-1}+\tfrac{1}{2}(1+x^2)d^{-3}-\tfrac{3}{2}xd^{-4}+\cdots\}. \tag{4.3}$$

Using these results, formal expansions for $\tilde{a}_1$ and $\tilde{a}_1^\dagger$ may be derived in those cases discussed in Sec. III in which closed-form expressions obtain for these operators. In particular, corresponding to the algebra[10] $[\tilde{a}_1, \tilde{a}_1^\dagger] = \mathrm{diag}\,(w,1,1,...)$ in $\tilde{\mathscr{H}}_1$, for which $\tilde{a}_1$ is the operator C_w given by Eqs. (2.14) and (2.15), we obtain the following expansions [quoted up to $\mathscr{O}(d^{-2})$ for brevity]:

$$2^{1/2}\tilde{a}_1 = x + d - (w - 2 - \phi_\lambda')d^{-1} + [x(2-w) + \phi_\lambda\phi_\lambda' + x\phi_\lambda' + 2\phi_\lambda]d^{-2} + \cdots \tag{4.4}$$

and

$$2^{1/2}\tilde{a}_1^\dagger = x - d + (w - 2 - \phi_\lambda')d^{-1} + [x(2-w) - x\phi_\lambda' - \phi_\lambda\phi_\lambda']d^{-2} + \cdots . \tag{4.5}$$

Further, we find

$$\tilde{a}_1\tilde{a}_1^\dagger = \tfrac{1}{2}(-d^2 + x^2 + 2w - 3) - \phi_\lambda', \tag{4.6}$$

while in the space $\mathrm{span}\{|\theta_n\rangle, n \geqslant 2\}$ we also have

$$\tilde{a}_1^\dagger\tilde{a}_1 = \tfrac{1}{2}(-d^2 + x^2 + 2w - 5) - \phi_\lambda'. \tag{4.7}$$

When $w = 1$, Eq. (4.7) is valid in all of $\tilde{\mathscr{H}}_1$.

Finally, we remark that in the limit $|\lambda| \to \infty$, $\phi_\lambda(x)$ and its derivatives vanish, $U \to 1$, $b \to a$, $\tilde{\mathscr{H}}_1 \to \mathscr{H}_1$, $\tilde{a}_1 \to a_1$, etc., in all the foregoing expressions.

V. COHERENT STATES

With the operators $\tilde{a}_1$, $\tilde{a}_1^\dagger$ in $\tilde{\mathscr{H}}_1$ and their commutation relation at hand [Eqs. (3.17)–(3.19)], we may construct coherent states (CS) in the usual manner, i.e., as eigenstates of $\tilde{a}_1$ built upon the base state $|\theta_1\rangle$. We find, for every $\zeta \in \mathscr{C}$,

$$\tilde{a}_1|\zeta;\theta_1\rangle = \zeta|\zeta;\theta_1\rangle, \tag{5.1}$$

where the normalized eigenstate is given by

$$|\zeta;\theta_1\rangle = h^{-1/2}(|\zeta|^2)\sum_{n=0}^{\infty} d_n^{1/2}\zeta^n|\theta_{n+1}\rangle \tag{5.2}$$

with

$$d_0 = 1, \quad d_n = (W_1 \cdots W_n)^{-1}. \tag{5.3}$$

The normalization factor in Eq. (5.2) is defined by

$$h(|\zeta|^2) = \sum_{n=0}^{\infty} d_n|\zeta|^{2n}. \tag{5.4}$$

A question of interest is the kind of Fock–Bargmann or entire function representation of states[9,16] in $\tilde{\mathscr{H}}_1$ associated with the CS constructed above—specifically, the *order*[17] of the class of entire functions involved. A state $|\Psi\rangle$ of finite norm is represented by the analytic function

$$\Psi(\zeta) = \sum_{n=0}^{\infty} d_n^{1/2}\langle\theta_{n+1}|\Psi\rangle\zeta^n. \tag{5.5}$$

As the state $|\zeta;\theta_1\rangle$ in Eq. (5.2) is normalized to unity, we have $|\Psi(\zeta)| \leqslant \|\Psi\||h|^{1/2}$. Hence the asymptotic behavior of $\Psi(\zeta)$ is decided by that of $h^{1/2}$ which, in turn, is controlled by the growth properties of the sequence $\{d_n\}$. In particular, the *order* of the class of entire functions to which $\Psi(\zeta)$ belongs is given by

$$\rho = \lim_{n\to\infty} \frac{2n \log n}{\log(W_1 \cdots W_n)}. \tag{5.6}$$

In the special cases (i)–(v) considered at the end of Sec. III, this leads to the following results: In case (i), $(W_1 \ldots W_n) = w^n n!$, while in case (ii)[10] it is $\Gamma(n+w)/\Gamma(n)$. In both cases it follows easily from Eq. (5.6) that $\rho = 2$, i.e., the order is exactly the same as it is for the Fock–Bargmann representation corresponding to the standard oscillator CS. In case (iii), $(W_1 \ldots W_n) = 2^{-n} n!(n+1)!$, leading to $\rho = 1$. All these cases are covered by the following generalization: if the coefficient w_n of the algebra $[\tilde{a}_1, \tilde{a}_1^\dagger] = \mathrm{diag}(w_1, w_2, \ldots)$ has the leading asymptotic behavior n^ν for large n, with $\nu > -1$, then $(W_1 \cdots W_n) \sim n^{n(1+\nu)}$, which leads to the result

$$\rho = \frac{2}{1+\nu} \tag{5.7}$$

for the order of the corresponding entire function representation of $\tilde{\mathscr{H}}_1$. Cases (i) and (ii) correspond to $\nu = 0$, while $\nu = 1$ in case (iii). In case (iv), $h = w(w - |\zeta|^2)^{-1}$, so that the representation is not in terms of entire functions at all.

For the q-generalization, case (v), $(W_1 \cdots W_n)$ is a ''q-factorial'' given by

$$q(q+q^2)\cdots(q+q^2+\cdots+q^n) = q^n (q-1)^{(1-n)} (q^2-1)\cdots(q^n-1). \tag{5.8}$$

When $q < 1$, it is readily established that the series defining h has a finite radius of convergence ($|\zeta| = q^{1/2}$), as in case (iv). When $q = 1$, of course, we recover case (i) or (ii) with $w = 1$, so that $\rho = 2$. On the other hand, when $q > 1$, $(W_1 \cdots W_n)$ has the leading asymptotic behavior $q^{n^2/2}$, which implies that $\rho = 0$.

The CS considered above have been constructed as annihilation operator eigenstates. In the special case $w_n = 1$ for all $n \geqslant 1$, they are also CS in the group-theoretic sense. Each coefficient c_n in Eq. (3.14) is now equal to unity, and the shift operator $\tilde{S}$ obeys the exact partial isometry condition

$$\tilde{S}\,\tilde{S}^\dagger = 1, \quad \tilde{S}^\dagger \tilde{S} = 1 - |\theta_0\rangle\langle\theta_0|. \tag{5.9}$$

We then have $[\tilde{a}_1, \tilde{a}_1^\dagger] = 1$ in $\tilde{\mathscr{H}}_1$, and the Hamiltonian $\tilde{H}_1 = \tilde{a}_1^\dagger \tilde{a}_1$ with eigenstates $|\theta_n\rangle$ and eigenvalues $(n-1)$, $n = 1,2,\ldots$, is the exact analog in $\tilde{\mathscr{H}}_1$ of the Hamiltonian $H = a^\dagger a$ in $\mathscr{H}$. The transformation from $(a, a^\dagger)$ to $(\tilde{a}_1, \tilde{a}_1^\dagger)$ is now a true operator map, in the sense that any function $\phi(a, a^\dagger)$ transforms as

$$\phi(a, a^\dagger) \to \phi(\tilde{a}_1, \tilde{a}_1^\dagger) = \tilde{S}^\dagger \phi(a, a^\dagger) \tilde{S}. \tag{5.10}$$

The coherent states

$$|\zeta; \theta_1\rangle = \exp\left(-\frac{|\zeta|^2}{2}\right) \sum_{n=0}^{\infty} \frac{\zeta^n}{\sqrt{n!}} |\theta_{n+1}\rangle, \quad \zeta \in \mathscr{C}, \tag{5.11}$$

are normalized eigenstates of the annihilation operator $\tilde{a}_1$. They are *also* CS in the Perelomov sense, for we have $|\zeta; \theta_1\rangle = D(\zeta, \bar{\zeta})|\theta_1\rangle$, where the displacement operator

$$D(\zeta, \bar{\zeta}) = \exp(\zeta \tilde{a}_1^\dagger - \bar{\zeta} \tilde{a}_1) \tag{5.12}$$

is unitary in $\tilde{\mathscr{H}}_1$. Moreover, as in the case of the standard oscillator CS, the state $|\zeta; \theta_1\rangle$ is the ground state of the displaced Hamiltonian $D\tilde{H}_1 D^\dagger$. The excited states $|\zeta; \theta_n\rangle$ $(n \geqslant 2)$ of this Hamiltonian, i.e., the generalized coherent states corresponding to $|\zeta; \theta_1\rangle$, are obtained *either* by operating on $|\zeta; \theta_1\rangle$ by the transformed raising operator $(D\tilde{a}_1^\dagger D^\dagger)^{n-1}$, *or* by displacing the corresponding base state $|\theta_n\rangle$:

$$|\zeta; \theta_n\rangle = (D\tilde{a}_1^\dagger D^\dagger)^{n-1} |\zeta; \theta_1\rangle = D|\theta_n\rangle, \quad n = 2,3,\ldots. \tag{5.13}$$

Thus, these coherent states and generalized CS have in $\tilde{\mathscr{H}}_1$ the same standing as their standard oscillator counterparts have in $\mathscr{H}$. They therefore constitute the closest analogs of the latter for the isospectral family of oscillators in $\tilde{\mathscr{H}}_1$, while remaining a distinctly new (unitarily inequivalent) class of coherent states. However, the generalized algebra of Eq. (3.19) leads, as we have seen, to an even wider variety of possible coherent states with interesting properties.

ACKNOWLEDGMENTS

We are grateful to Suresh Govindarajan for helpful discussions on pseudodifferential operators. S.S. acknowledges financial support from the Council of Scientific and Industrial Research, India, in the form of a Junior Research Fellowship.

[1] C. M. Caves and B. L. Schumaker, Phys. Rev. A **31**, 3068, 3093 (1985).
[2] H. P. Yuen, Phys. Rev. A **13**, 2226 (1976).
[3] G. S. Agarwal, J. Opt. Soc. Am. B **5**, 1940 (1988).
[4] C. L. Mehta, A. K. Roy, and G. M. Saxena, Phys. Rev. A **46**, 1565 (1992).
[5] P. Shanta, S. Chaturvedi, V. Srinivasan, G. S. Agarwal, and C. L. Mehta, Phys. Rev. Lett. **72**, 1447 (1994).
[6] S. Seshadri, S. Lakshmibala, and V. Balakrishnan, Phys. Rev. A **55**, 869 (1997).
[7] D. J. Fernandez, V. Hussin, and L. M. Nieto, J. Phys. A **27**, 3547 (1994).
[8] M. S. Kumar and A. Khare, Phys. Lett. A **217**, 73 (1996).
[9] A. Perelomov, *Generalized Coherent States and Their Applications* (Springer-Verlag, Berlin, 1986).
[10] D. J. Fernandez, L. M. Nieto, and O. Rosas-Ortiz, J. Phys. A **28**, 2693 (1995); O. Rosas-Ortiz, J. Phys. A **29**, 3281 (1996).
[11] M. Simon and B. Reed, *Methods of Modern Mathematical Physics* (Academic, New York, 1972), Vol. I.
[12] I. M. Gel'fand and L. A. Dikii, Funkts. Anal. Prilozhen. **10**, 13 (1976) [Funct. Anal. Appl. **10**, 259 (1976)].
[13] S. Chaturvedi and V. Srinivasan, Phys. Rev. A **44**, 8020 (1991).
[14] I. M. Gel'fand and L. A. Dikii, Usp. Mat. Nauk **30**, 67 (1975). [Russ. Math. Surv. **30**, 77 (1975)].
[15] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge U.P., Cambridge, 1990).
[16] V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961).
[17] E. C. Titchmarsh, *The Theory of Functions* (Oxford U.P., London, 1962).

Algebraic characterization of vector supersymmetry in topological field theories

S. P. Sorella
UERJ, Universidade do Estado do Rio de Janeiro, Departamento de Física Teórica, Instituto de Física, UERJ, Rua São Francisco Xavier, 524, 20550-013, Maracanã, Rio de Janeiro, Brazil

L. C. Q. Vilar, O. S. Ventura, and C. A. G. Sasaki
CBPF, Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290-180 Urca, Rio de Janeiro, Brazil

(Received 5 February 1997; accepted for publication 23 September 1997)

A cohomological characterization of a class of linearly broken Ward identities is provided. The examples of the topological vector supersymmetry and of the Landau ghost equation are discussed. The existence of these linearly broken Ward identities lies in the BRST exactness of suitable antifield dependent cocycles with negative ghost number. This algebraic set up turns out to be related to the cohomological reformulation of the Noether theorem given by G. Barnich, F. Brandt, and M. Henneaux, Commun. Math. Phys. **174**, 57, 93 (1995); M. Henneaux, J. Pure Appl. Algebra **100**, 3 (1995). © *1998 American Institute of Physics.*
[S0022-2488(98)01501-1]

I. INTRODUCTION

The topological field theories TFT (Ref. 1) are known to be characterized, besides their BRST symmetry, by the so-called *vector supersymmetry*;[2–5] an additional invariance with many properties.

The first property is that the generators of the vector susy carry a Lorentz index and, together with the BRST operator, give rise to an algebra of the Wess–Zumino-type which, closing on-shell on the space–time translations, allows for a supersymmetric interpretation of the TFT.[2–5]

The second feature is that, in the case of the TFT of the Schwartz-type, the vector susy is present only after the introduction of all the ghost fields needed in order to quantize the model. It is therefore an invariance of the fully quantized action but not of its classical part.[4] This property is unavoidably related to the choice of the gauge fixing term, meaning that the vector susy can exist only for certain values of the gauge parameters present in the gauge fixing condition. As an example, let us mention that in the case of the three-dimensional Chern–Simons model the vector susy turns out to select, among the class of the linear covariant gauge fixings, the Landau gauge.[2,3]

For the TFT of the cohomological type the use of the equivariant approach[6] allows us to obtain the vector susy through the introduction of the ghost fields corresponding to the so called shift symmetry.[1] In this case the vector susy can be related to the twisted version of extended supersymmetry. For instance, in the case of the Witten's four-dimensional Euclidean topological Yang–Mills[7] in flat space–time the vector susy can be obtained by twisting the standard $N=2$ super Yang–Mills.[8,9] After the twist, the fermionic spinor generators of $N=2$ become a self-dual two form, a vector and a scalar. The vector susy is identified here with the twisted vector generators.

The third aspect is that after the introduction of the antifields (or BRST external sources), the algebra between the vector susy Ward operator and the BRST operator closes off-shell on the space–time translations, without making use of the equations of motion. Moreover the vector susy loses now the property of being an exact invariance of the fully quantized action. Rather, it yields a broken Ward identity.[3–5] It is a remarkable feature, however, that the corresponding breaking term is a classical breaking, i.e., a breaking which is purely linear in the quantum fields. It is known that this kind of breaking does not get renormalized by the quantum corrections and does not spoil the usefulness of the corresponding Ward identity.[10] The latter turns out to be free from

 © 1998 American Institute of Physics

anomalies and has been shown to play a crucial role in order to establish the ultraviolet finiteness of the TFT.[3–5]

Let us also mention that the vector susy Ward operator can be introduced on a more abstract geometrical level,[11–15] independently from the fact of being a (linearly broken) symmetry of the action. In this case the vector Ward operator turns out to be useful in order to solve the BRST consistency condition for anomalies and invariant counterterms. In addition, it allows us to encode the relevant informations (BRST transformations, cohomology classes, etc.) into a unique equation which takes the form of a generalized zero curvature condition.[16–18]

These properties make the vector susy interesting and motivate further research about its origin. This is particularly needed for the non-Abelian TFT of the Schwartz-type which are not manifestly related to an extended supersymmetric algebra. (In the case of the free Abelian BF models in two and three dimensions the vector susy can be obtained by means of a suitable twist of extended supersymmetry for matter multiplets.[19]) In this case the vector susy has been introduced essentially by hand[2–4] and later on has been related to the existence of a conserved current obtained from the BRST exactness of the corresponding energy-momentum tensor.[20,10]

Moreover, a general set up accounting for all the features of the vector susy has not yet been completely worked out. This is the aim of the present work. We shall be able to provide an algebraic cohomological characterization of the existence of the vector susy and of the related linear classical breaking term. It will be also seen that it is precisely the requirement of linearity in the quantum fields which selects a particular set of gauge parameters, clarifying thus the relationship between the vector susy and the gauge fixing condition.

We shall prove, in particular, that the vector susy is deeply related to a BRST antifield dependent cocycle with ghost number -1. When this cocycle is trivial the vector susy Ward identity can be established and turns out to be necessarily accompanied by a breaking term linear in the fields. On the other hand, when this cocycle is nontrivial, the vector susy cannot be implemented and we are left with an example of an antifield dependent cohomology class.

The algebraic framework presented here is related to the cohomological reformulation of the Noether theorem given by Ref. 21. We shall observe indeed that the aforementioned vector cocycle can be related to a set of currents among which one can identify the BRST invariant energy-momentum tensor, meaning that this vector cocycle is related to the Poincaré transformations.[21] Recently, it has been considered in connection with the problem of collecting additional global invariances into a unique generalized Slavnov–Taylor identity.[22]

The paper is organized as follows. In Sec. II we introduce the general algebraic setup; we present the relevant properties of the vector antifield cocycle and we discuss its relation with the BRST invariant energy-momentum tensor. In Sec. III the pure Yang–Mills model is considered. We shall check that in this case the vector cocycle is nontrivial. Section IV is devoted to a detailed analysis of several examples of TFT. In these cases the BRST triviality of the vector cocycle will be at the origin of the existence of the vector susy Ward identities. Finally, in Sec. V we shall present a second example of a linearly broken Ward identity whose contact terms can be obtained through an exact antifield cocycle carrying a group index. The corresponding Ward identity, known as the *Landau ghost equation*,[23] turns out to be related to the rigid gauge invariance.

II. GENERAL NOTATIONS AND THE VECTOR COCYCLE

In order to present the general algebraic set up let us begin by fixing the notations. We shall work in a flat D-dimensional space–time equipped with a set of fields generically denoted by $\{\varphi^i\}$, i labeling the different kinds of fields needed in order to properly quantize the model; gauge fields, ghosts, ghosts for ghosts, etc. Following the standard quantization procedure we introduce for each field φ^i of ghost number $\mathcal{N}^{\cdot}_{\varphi^i}$ and dimension d_{φ^i}, the corresponding antifield φ^{i*} with ghost number $-(1+\mathcal{N}^{\cdot}_{\varphi^i})$ and dimension $(D-d_{\varphi^i})$. We shall also assume that the set of fields $\{\varphi^i\}$ do not explicitly contains the antighosts and their corresponding Lagrange multipliers which, being grouped in BRST doublets, do not contribute to the BRST cohomology.[24] Accordingly, some of the antifields φ^{i*} have to be understood as shifted antifields[10] which already take into account the antighosts.

In order to define the model we have now to introduce the classical gauge fixed *reduced* action[10] $\Sigma(\varphi,\varphi^*)$. (The name *reduced* action is used to denote the part of the complete gauge fixed

action which does not depend on the Lagrange multipliers and which depends on the antighosts only through the shifted antifields. As it will be discussed in the next sections, the dependence of the complete action on the Lagrange multipliers as well as on the gauge parameters present in the gauge fixing condition will be determined by requiring that the breaking term corresponding to the topological vector susy Ward identity is at most linear in the quantum fields.) This is done by requiring Σ to be power counting renormalizable and to be the solution of the classical homogeneous Slavnov–Taylor (or master equation) identity[25]

$$\int d^D x \, \frac{\delta \Sigma}{\delta \varphi^i} \, \frac{\delta \Sigma}{\delta \varphi^{i*}} = \frac{1}{2} \, \mathscr{B}_\Sigma \Sigma = 0, \tag{2.1}$$

where $\mathscr{B}_\Sigma$ denotes the nilpotent linearized Slavnov–Taylor operator,

$$\mathscr{B}_\Sigma = \int d^D x \left(\frac{\delta \Sigma}{\delta \varphi^i} \, \frac{\delta}{\delta \varphi^{i*}} + \frac{\delta \Sigma}{\delta \varphi^{i*}} \, \frac{\delta}{\delta \varphi^i} \right), \tag{2.2}$$

$$\mathscr{B}_\Sigma \mathscr{B}_\Sigma = 0, \tag{2.3}$$

As usual, the classical action $\Sigma(\varphi,\varphi^*)$ will be assumed to be invariant under the space–time translations, i.e.,

$$\mathscr{P}_\mu \Sigma = \int d^D x \left(\partial_\mu \varphi^i \, \frac{\delta \Sigma}{\delta \varphi^i} + \partial_\mu \varphi^{i*} \, \frac{\delta \Sigma}{\delta \varphi^{i*}} \right) = 0, \tag{2.4}$$

The classical Slavnov–Taylor identity (2.1) and the translation invariance (2.4) will be taken therefore as the basic starting points for the characterization of the classical action $\Sigma(\varphi,\varphi^*)$ and for the algebraic analysis which will be carried out in the next sections. We remark here that the requirement of the translation invariance, as expressed by Eq. (2.4), does not imply any further restrictions on Σ than those that are tacitly assumed in any local field theory.

Let us now introduce the following integrated local vector polynomial, linear in the antifields φ^{i*}, of ghost number -1 and dimension $(D+1)$,

$$\Omega_\nu^{-1} = \int d^D x [\, \omega_\nu^{-1}]_{D+1} \equiv \int d^D x (-1)^{(1+\mathscr{N}_{\varphi^i})} \varphi^i \partial_\nu \varphi^{i*}. \tag{2.5}$$

It is easily proven that the above expression is $\mathscr{B}_\Sigma$-invariant. In fact, one has

$$\mathscr{B}_\Sigma \Omega_\nu^{-1} = \int d^D x \left(\partial_\nu \varphi^i \, \frac{\delta \Sigma}{\delta \varphi^i} + \partial_\nu \varphi^{i*} \, \frac{\delta \Sigma}{\delta \varphi^{i*}} \right) = \mathscr{P}_\nu \Sigma = 0, \tag{2.6}$$

due to the translation invariance of the classical action $\Sigma(\varphi,\varphi^*)$. Having proven the invariance of Ω_ν^{-1} we have now to establish if it eventually identifies a cohomology class of the operator $\mathscr{B}_\Sigma$. We have the following two possibilities, namely,

(1) Ω_ν^{-1} is $\mathscr{B}_\Sigma$-exact, i.e.,

$$\Omega_\nu^{-1} = \mathscr{B}_\Sigma \Xi_\nu^{-2}, \tag{2.7}$$

Ξ_ν^{-2} being an integrated local polynomial of ghost number -2 and dimension $(D+1)$.
(2) Ω_ν^{-1} belongs to the integrated cohomology of $\mathscr{B}_\Sigma$,

$$\Omega_\nu^{-1} \neq \mathscr{B}_\Sigma \Xi_\nu^{-2}. \tag{2.8}$$

The detailed analysis of the two equations (2.7), (2.8) will be the main subject of the next sections. Let us limit here to underline that while the first possibility (2.7) turns out to be a feature

of the TFT, the second one (2.8) is typical of a Yang–Mills theory and, more generally, of any model possessing an invariant energy-momentum tensor $T_{\mu\nu}$ which cannot be written as a pure $\mathscr{B}_\Sigma$-variation, i.e.,

$$\mathscr{B}_\Sigma T_{\mu\nu}=0, \quad T_{\mu\nu}\neq\mathscr{B}_\Sigma\Lambda_{\mu\nu}. \tag{2.9}$$

In this latter case Ω_ν^{-1} provides an example of a nontrivial antifield dependent cocycle of the operator $\mathscr{B}_\Sigma$, giving thus an explicit realization of the general results of Ref. 21.

Concerning now the first equation (2.7), it is worthwhile to recall that one of the properties of the TFT is precisely that of having a nonphysical energy momentum tensor.[1,7] This means that the energy momentum tensor $T^{\rm top}_{\mu\nu}$ of a TFT can be written as a pure $\mathscr{B}_\Sigma$-variation,

$$T^{\rm top}_{\mu\nu}=\mathscr{B}_\Sigma\Lambda_{\mu\nu}, \tag{2.10}$$

for some local $\Lambda_{\mu\nu}$, yielding thus a more direct indication of the BRST exactness of the vector cocycle Ω_ν^{-1}.

For a better understanding of the relationship between Ω_ν^{-1} and the energy-momentum tensor, let us write down the descent equations[10,21,24] corresponding to the integrated invariance condition (2.6),

$$\begin{aligned}
&\mathscr{B}_\Sigma[\,\omega_\nu^{-1}]_{D+1}=\partial^{\mu_1}[\,\omega^0_{\mu_1\nu}]_D,\\
&\mathscr{B}_\Sigma[\,\omega^0_{\mu_1\nu}]_D=\partial^{\mu_2}[\,\omega^1_{[\mu_1\mu_2]\nu}]_{D-1},\\
&\mathscr{B}_\Sigma[\,\omega^1_{[\mu_1\mu_2]\nu}]_{D-1}=\partial^{\mu_3}[\,\omega^2_{[\mu_1\mu_2\mu_3]\nu}]_{D-2},\\
&\cdots,\\
&\cdots,\\
&\mathscr{B}_\Sigma[\,\omega^{D-2}_{[\mu_1\mu_2\ldots\mu_{D-1}]\nu}]_2=\partial^{\mu_D}[\,\omega^{D-1}_{[\mu_1\mu_2\ldots\mu_{D-1}\mu_D]\nu}]_1,\\
&\mathscr{B}_\Sigma[\,\omega^{D-1}_{[\mu_1\mu_2\ldots\mu_{D-1}\mu_D]\nu}]_1=0,
\end{aligned} \tag{2.11}$$

where the $[\,\omega^{j-1}_{[\mu_1\mu_2\ldots\mu_j]\nu}]_{D-j+1}$ with $(j=0,\ldots,D)$ are local currents of ghost number $(j-1)$, antisymmetric in the lower indices $(\mu_1,\mu_2,\ldots,\mu_j)$, and of dimension $(D-j+1)$. The usefulness of working with the system (2.11) is due to the fact that these equations relate the local cocycle $[\,\omega_\nu^{-1}]_{D+1}$ with currents of lower dimension which can provide a more easy and transparent interpretation of the physical meaning of Ω_ν^{-1}. We observe indeed that the cocycle $[\,\omega^0_{\mu_1\nu}]_D$ entering the second equation of the system (2.11) has the same quantum numbers of the energy momentum tensor, being of dimension D, of ghost number 0, and possessing two Lorentz indices. From the general results on the BRST cohomology[24] it follows that, independently from the fact that the cohomology of the operator $\mathscr{B}_\Sigma$ is empty in the sectors with dimension lower than D, the existence of a nontrivial local cohomology in the sector of dimension D and with two free Lorentz indices necessarily implies the nontriviality of the upper level of dimension $D+1$. It becomes apparent thus that the nontriviality of the BRST invariant energy-momentum tensor is deeply related to the existence of a nontrivial integrated cohomology in the sector of dimension $D+1$ and negative ghost-number.

In fact, as we shall see explicitly in the case of the Yang–Mills theory, the current $[\,\omega^0_{\mu_1\nu}]_D$ is the improved BRST invariant energy-momentum tensor. On the other hand, the use of the descent equations (2.11) gives us a simple way to establish the triviality of the vector cocycle Ω_ν^{-1} in the case of the TFT. This is actually due to the fact that the field content of the topological models gives rise to a BRST cohomology which is empty in the various sectors of the descent equations (2.11), implying that the corresponding energy momentum tensor is BRST trivial, according to Eq. (2.10). Moreover, as we shall see in detail later on, the right-hand side of Eq. (2.7) expressing the

BRST triviality of Ω_ν^{-1} will provide the contact terms of the vector susy Ward identity of the topological theories. In this latter case we shall also check that the left-hand side of Eq. (2.7) reduces to a classical linear breaking, showing then that the topological vector susy Ward identity is always linearly broken.

III. PURE YANG–MILLS THEORY

As a first application of the general algebraic set up discussed in the previous section, let us prove here that in the case of pure Yang–Mills the vector cocycle Ω_ν^{-1} cannot be written as an exact $\mathscr{B}_\Sigma$-term. Let us begin by considering the complete fully quantized gauge fixed Yang–Mills action which, choosing a Feynman gauge, reads

$$\mathscr{S}=\int d^4x\left(-\frac{1}{4g^2}F^a_{\mu\nu}F^{a\mu\nu}+b^a\partial A^a+\frac{\alpha}{2}b^ab^a+\partial^\mu\overline{c}^a(D_\mu c)^a\right)+\int d^4x\hat{A}^{a*}_\mu(D^\mu c)^a$$
$$-\tfrac{1}{2}f_{abc}C^{*a}c^bc^c, \tag{3.1}$$

where $(c,\overline{c},b)$ are the Faddeev–Popov ghost, the antighost and the Lagrange multiplier fields and

$$(D_\mu c)^a=\partial_\mu c^a+f^a_{bc}A^b_\mu c^c \tag{3.2}$$

is the covariant derivative with f_{abc} the totally antisymmetric structure constants of a compact semisimple Lie group G. The two antifields $(\hat{A}^*,C^*)$ are introduced in order to properly define the nonlinear BRST transformations of the gauge field A and of the ghost c. The dimensions and the ghost numbers of all the fields and antifields are assigned as in the following diagram:

	A	$\hat{A}^*$	c	C^*	$\bar{c}$	b
dim.	1	3	0	4	2	2
gh-numb.	0	-1	1	-2	-1	0

(3.3)

As it is well known,[10] the complete action $\mathscr{S}$ is characterized by the classical Slavnov–Taylor identity expressing the invariance of (3.1) under the BRST transformations

$$\int d^4x\left(\frac{\delta\mathscr{S}}{\delta A^a_\mu}\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}+\frac{\delta\mathscr{S}}{\delta c^a}\frac{\delta\mathscr{S}}{\delta C^{*a}}+b^a\frac{\delta\mathscr{S}}{\delta\overline{c}^a}\right)=0, \tag{3.4}$$

and by the linear gauge fixing condition[10]

$$\frac{\delta\mathscr{S}}{\delta b^a}=\partial A^a+\alpha b^a. \tag{3.5}$$

A further identity, the antighost equation[10]

$$\frac{\delta\mathscr{S}}{\delta\overline{c}^a}+\partial^\mu\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}=0, \tag{3.6}$$

follows from the gauge fixing condition (3.5) and from the Slavnov–Taylor identity (3.4), implying that the antighost $\overline{c}$ and the antifield $\hat{A}^*$ can enter only through the shifted antifield $A^{*a\mu}$ of ghost number -1 and dimension 3,

$$A^{*a\mu}=\hat{A}^{*a\mu}+\partial^\mu\overline{c}^a. \tag{3.7}$$

Introducing now the reduced Yang–Mills action $\Sigma(A,A^*,c,C^*)$ defined by the identities (3.5) and (3.6),

$$\mathscr{S}=\Sigma+\int d^4x\left(b^a\partial A^a+\frac{\alpha}{2}\,b^a b^a\right), \tag{3.8}$$

i.e.,

$$\Sigma=\int d^4x\left(-\frac{1}{4g^2}\,F^a_{\mu\nu}F^{a\mu\nu}+A^{*a\mu}(D_\mu c)^a-\frac{1}{2}\,f_{abc}C^{*a}c^b c^c\right), \tag{3.9}$$

it is easily verified that Σ obeys the homogeneous Slavnov–Taylor identity,

$$\int d^4x\left(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta\Sigma}{\delta A^{*a\mu}}+\frac{\delta\Sigma}{\delta c^a}\frac{\delta\Sigma}{\delta C^{*a}}\right)=\frac{1}{2}\,\mathscr{B}_\Sigma\Sigma=0, \tag{3.10}$$

with

$$\mathscr{B}_\Sigma=\int d^4x\left(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta}{\delta A^{*a\mu}}+\frac{\delta\Sigma}{\delta A^{*a\mu}}\frac{\delta}{\delta A^{a\mu}}+\frac{\delta\Sigma}{\delta c^a}\frac{\delta}{\delta C^{*a}}+\frac{\delta\Sigma}{\delta C^{*a}}\frac{\delta}{\delta c^a}\right), \tag{3.11}$$

and

$$\mathscr{B}_\Sigma\mathscr{B}_\Sigma=0. \tag{3.12}$$

For the vector cocycle Ω^{-1}_ν of Eq. (2.5) we have now

$$\Omega^{-1}_\nu=\int d^4x[\,\omega^{-1}_\nu]_5\equiv\int d^4x(c^a\partial_\nu C^{*a}-A^a_\mu\partial_\nu A^{*a\mu}). \tag{3.13}$$

Expression (3.13) turns out to be $\mathscr{B}_\Sigma$-invariant, according to Eq. (2.6),

$$\mathscr{B}_\Sigma\Omega^{-1}_\nu=\mathscr{P}_\nu\Sigma=0. \tag{3.14}$$

We are now ready to show that in the case of Yang–Mills theory the vector cocycle Ω^{-1}_ν of Eq. (3.13) is nontrivial. Let us proceed by assuming the converse. We suppose then that Ω^{-1}_ν can be written as an exact $\mathscr{B}_\Sigma$-term,

$$\Omega^{-1}_\nu=\mathscr{B}_\Sigma\Xi^{-2}_\nu, \tag{3.15}$$

for some integrated local polynomial Ξ^{-2}_ν of ghost number -2 and dimension 5. From Table I it follows that the most general form for Ξ^{-2}_ν is given by

$$\Xi^{-2}_\nu=\zeta\int d^4x C^{*a}A^a_\nu, \tag{3.16}$$

where ζ is an arbitrary parameter which has to be fixed by the exactness condition (3.15). Thus, from Eq. (3.15), we should have the algebraic equality

$$\int d^4x(c^a\partial_\nu C^{*a}-A^a_\mu\partial_\nu A^{*a\mu})=\zeta\int d^4x(C^{*a}\partial_\nu c^a+A^a_\nu(D_\mu A^{*\mu})^a). \tag{3.17}$$

However it is almost immediate to check that the above equation has no solution for ζ, showing then that Ω^{-1}_ν cannot be written as a pure $\mathscr{B}_\Sigma$-variation,

$$\Omega^{-1}_\nu\neq\mathscr{B}_\Sigma\Xi^{-2}_\nu. \tag{3.18}$$

Therefore in the case of Yang–Mills theory the vector cocycle Ω^{-1}_ν identifies a cohomology class of the operator $\mathscr{B}_\Sigma$ in the sector of the integrated local polynomials of ghost number -1 and dimension 5, providing thus an explicit example of an antifield dependent cohomology class of $\mathscr{B}_\Sigma$.

For a better understanding of the nontriviality of the vector cocycle Ω_ν^{-1} let us discuss now its relationship with the energy-momentum tensor. To this purpose we analyze the descent equations corresponding to the local polynomial $[\omega_\nu^{-1}]_5$,

$$\begin{aligned}
\mathscr{B}_\Sigma[\omega_\nu^{-1}]_5 &= \partial^{\mu_1}[\omega^0_{\mu_1\nu}]_4,\\
\mathscr{B}_\Sigma[\omega^0_{\mu_1\nu}]_4 &= \partial^{\mu_2}[\omega^1_{[\mu_1\mu_2]\nu}]_3,\\
\mathscr{B}_\Sigma[\omega^1_{[\mu_1\mu_2]\nu}]_3 &= \partial^{\mu_3}[\omega^2_{[\mu_1\mu_2\mu_3]\nu}]_2,\\
\mathscr{B}_\Sigma[\omega^2_{[\mu_1\mu_2\mu_3]\nu}]_2 &= \partial^{\mu_4}[\omega^3_{[\mu_1\mu_2\mu_3\mu_4]\nu}]_1,\\
\mathscr{B}_\Sigma[\omega^3_{[\mu_1\mu_2\mu_3\mu_4]\nu}]_1 &= 0.
\end{aligned} \tag{3.19}$$

After some straightforward algebraic manipulations, for the local currents $(\omega^{j-1}_{[\mu_1\ldots\mu_j]\nu},\ j=1,\ldots,4)$ one finds

$$[\omega^0_{\mu_1\nu}]_4 = \frac{1}{g^2}\left(F_{\mu_1\sigma}F^{a\sigma}_\nu - \frac{1}{4}\,g_{\mu_1\nu}F^a_{\rho\sigma}F^{a\rho\sigma}\right) + \mathscr{B}_\Sigma(A^{*a}_{\mu_1}A^a_\nu), \tag{3.20}$$

$$[\omega^1_{[\mu_1\mu_2]\nu}]_3 = \epsilon_{\mu_1\mu_2}{}^{\mu_3\mu_4}\partial_{\mu_3}[\tilde{\omega}^1_{\mu_4\nu}]_2, \tag{3.21}$$

$$[\omega^2_{[\mu_1\mu_2\mu_3]\nu}]_2 = \epsilon_{\mu_1\mu_2\mu_3\mu_4}.\mathscr{B}_\Sigma[\tilde{\omega}^1_{\mu_4\nu}]_2 + \epsilon^{\mu_4}_{\mu_1\mu_2\mu_3}\partial_{\mu_4}[\tilde{\omega}^2_\nu]_1, \tag{3.22}$$

$$[\omega^3_{[\mu_1\mu_2\mu_3\mu_4]\nu}]_1 = \epsilon_{\mu_1\mu_2\mu_3\mu_4}.\mathscr{B}_\Sigma[\tilde{\omega}^2_\nu]_1, \tag{3.23}$$

with $[\tilde{\omega}^1_{\mu_4\nu}]_2$ and $[\tilde{\omega}^2_\nu]_1$ local arbitrary polynomials in the fields and their derivatives. From the above expressions we observe that while the last three currents are trivial solutions of the descent equations (3.19), the first one, $[\omega^0_{\mu_1\nu}]_4$, yields the familiar expression of the BRST invariant improved energy-momentum tensor which, as it is well known, belongs to the cohomology of $\mathscr{B}_\Sigma$.[24] One sees thus that, as already remarked in Sec I, the existence of a nontrivial invariant energy-momentum tensor is at the origin of the nontriviality of the vector cocycle Ω_ν^{-1}. Let us conclude this section by remarking that the exact term $\mathscr{B}_\Sigma(A^{*a}_{\mu_1}A^a_\nu)$ which naturally appears in the right-hand side of Eq. (3.20) is needed in order to ensure the off-shell conservation of the improved Yang–Mills energy-momentum tensor $T^{\rm YM}_{\mu\nu} = F^a_{\mu\sigma}F^{a\sigma}_\nu - \frac{1}{4}g_{\mu\nu}F^a_{\rho\sigma}F^{a\rho\sigma}$. Of course, the same conclusions hold if the pure Yang–Mills action is supplemented with the introduction of matter fields.

IV. THE CASE OF THE TOPOLOGICAL FIELD THEORIES

Having established the nontriviality of Ω_ν^{-1} for the Yang–Mills case, we analyze now the vector cocycle (2.5) in the context of the TFT. In this case, it turns out that Ω_ν^{-1} is always BRST exact,

$$\Omega_\nu^{-1} = \mathscr{B}_\Sigma \Xi_\nu^{-2}, \tag{4.1}$$

for some integrated local field polynomial Ξ_ν^{-2} with ghost number -2. The reason for the exactness of Ω_ν^{-1} in the TFT relies on their field content which, as proven by several authors,[26–29,4,5,10] does not allow for cohomology classes with free Lorentz indices, as the ones needed in order to have nontrivial solutions of the descent Eqs. (2.11). In fact, in the case of the TFT of the cohomological type,[30–32] the topological shift symmetry implies the emptiness of the BRST cohomology[1,5,26–28] and therefore the triviality of the vector cocycle Ω_ν^{-1}. (Let us remark here that, as discussed by Refs. 26–29, the relevant cohomology for the topological theories of the Witten's type is the so-called equivariant cohomology. The latter is the restriction of the BRST, cohomology to gauge invariant local field polynomials which do not depend on the Faddeev–

Popov ghost field c. Contrary to the BRST cohomology, the equivariant cohomology is not empty and turns out to provide a consistent definition of the Witten's observables.) Concerning now the Schwartz-type models, it can be proven that their field content allows for nontrivial cohomology classes.[3,4,10] Moreover, as observed in Refs. 16 and 17, these models can be formulated in a pure geometrical way due to the fact that all the fields (gauge fields, ghosts, ghosts for ghosts, antifields, etc.) can be viewed as being the components of a generalized gauge connection obeying a zero curvature condition. This structure, called complete ladder structure, implies that all the BRST cohomology classes can be identified with invariant polynomials in the undifferentiated dimensionless scalar ghosts present in the models.[3,4,10] From this result it follows that the cohomology relevant for the solutions of Eqs. (2.11) vanishes, meaning thus that also in the case of the Schwartz-type TFT the vector cocycle Ω_ν^{-1} is trivial.

Although Eq. (4.1) could seem empty of any interesting information because of its BRST exactness, it turns out that its content is far from being irrelevant. In fact the right-hand side of Eq. (4.1), due to the form (2.2) of the operator $\mathscr{B}_\Sigma$, is easily seen to provide the contact terms of a Ward identity. This identity, due to the presence of Ω_ν^{-1} in the left-hand side of Eq. (4.1), cannot express an exact invariance of the theory. Instead, Eq. (4.1) will have the meaning of a broken Ward identity. Furthermore, it is not difficult to convince oneself that this breaking term is linear in the quantum fields. Therefore it does not get renormalized by the radiative corrections and does not spoil the usefulness of the corresponding Ward identity. As one can easily understand, the existence of this linear breaking term is a direct consequence of the fact that the general expression of Ω_ν^{-1} given in Eq. (2.5) is linear in the fields, apart from possible quadratic terms coming from the fact that some of the antifields φ^{i*} have been shifted in order to account for the antighosts. We shall check, however, that these quadratic terms turn out to be handled in a very simple way, being reinterpreted as pure contact terms by means of an appropriate choice of the gauge parameters present in the gauge fixing. This means that the corresponding Ward identity is consistent with the Quantum Action Principles[33] only for certain classes of gauge fixings. This feature will be shown later on in the cases of the vector susy for three-dimensional Chern–Simons and of the Landau ghost equation.

Equation (4.1) has thus the meaning of a purely algebraic cohomological characterization of a Ward identity. Indeed, once one is able to prove that the cocycle Ω_ν^{-1} is trivial, one knows automatically that Eq. (4.1) has to be necessarily satisfied for some local polynomial Ξ_ν^{-2}. The strategy to be followed is now straightforward. One writes down the most general expression for Ξ_ν^{-2} compatible with the dimension of the space–time and with the field content of the model under consideration. This expression will, in general, depend on a set of free constant coefficients which will be fixed by the validity of Eq. (4.1). Recalling then that the functional form of the operator $\mathscr{B}_\Sigma$ [see Eq. (2.2)] depends on the reduced action Σ, it becomes apparent that the coefficients of Ξ_ν^{-2} are determined by the form of the classical action Σ. This means that whenever Ω_ν^{-1} is trivial, Eq. (4.1) tells us that the classical action Σ obeys a further Ward identity whose contact terms are precisely given by Eq. (4.1). This provides a simple mechanism for an algebraic characterization of a linearly broken symmetry of the action. Of course, this additional Ward identity will have the same quantum numbers of the vector cocycle Ω_ν^{-1}. It will carry a Lorentz index and it will have a negative ghost number. This vector identity, always present in the TFT, is known as the topological vector supersymmetry.[2–5]

Let us spend now a few words on the possibility of generalizing this algebraic mechanism in order to find other kinds of unknown global symmetries. What seems to emerge from the previous analysis is that the antifield cocycles which are relevant for the existence of a *class of linearly broken* Ward identities are those which are linear in the fields $\{\varphi^i\}$ and in the antifields $\{\varphi^{i*}\}$, and that are BRST exact. It is this last property which, when cast in the form of Eq. (4.1), allows us to identify the contact terms of the corresponding Ward identity. It is worthwhile to remark that an equation of the kind of (4.1) implies that the related Ward identity is always broken, due to the existence of a nonvanishing term in left-hand side of (4.1). However, being this term linear in the quantum fields, the resulting breaking term is classical. In Sec. V, we will discuss another example of a linearly broken Ward identity related to the rigid gauge invariance, whose contact terms can be characterized in a cohomological way by means of an equation of the type of (4.1).

Let us first focus on the discussion of the vector susy Ward identity in the case of the TFT. We shall present in detail the examples of the three-dimensional Chern–Simons, of the four-

dimensional BF model and of the so-called two-dimensional b-c and b-c β-γ ghost systems.

A. The three-dimensional Chern–Simons model

Using the same notations of the Sec. III, for the complete three-dimensional Chern–Simons action[1] quantized in the Landau gauge we have

$$\mathscr{S}=\mathscr{S}_{cs}+\int d^3x\left(b^a\partial A^a+\partial^\mu\overline{c}^a(D_\mu c)^a+\hat{A}^*_\mu(D^\mu c)^a-\frac{1}{2}f_{abc}C^{*a}c^bc^c\right), \tag{4.2}$$

where $\mathscr{S}_{cs}$ is given by

$$\mathscr{S}_{cs}=-\frac{k}{4\pi}\int d^3x\,\epsilon^{\mu\nu\rho}\left(A^a_\mu\partial_\nu A^a_\rho+\frac{1}{3}f_{abc}A^a_\mu A^b_\nu A^c_\rho\right), \tag{4.3}$$

k identifying the inverse of the coupling constant. The fields and the antifields have now the following quantum numbers:

	A	$\hat{A}^*$	c	C^*	$\bar{c}$	b
dim.	1	2	0	3	1	1
gh-numb.	0	-1	1	-2	-1	0

(4.4)

For the classical Slavnov–Taylor identity one has

$$\int d^3x\left(\frac{\delta\mathscr{S}}{\delta A^a_\mu}\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}+\frac{\delta\mathscr{S}}{\delta c^a}\frac{\delta\mathscr{S}}{\delta C^{*a}}+b^a\frac{\delta\mathscr{S}}{\delta\overline{c}^a}\right)=0. \tag{4.5}$$

Making use of the Landau gauge condition

$$\frac{\delta\mathscr{S}}{\delta b^a}=\partial A^a, \tag{4.6}$$

and of the antighost equation

$$\frac{\delta\mathscr{S}}{\delta\overline{c}^a}+\partial^\mu\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}=0, \tag{4.7}$$

for the reduced Chern–Simons action $\Sigma(A,A^*,c,C^*)$,

$$\mathscr{S}=\Sigma+\int d^3x\,b^a\partial A^a, \tag{4.8}$$

we obtain

$$\Sigma=\mathscr{S}_{cs}+\int d^3x\left(A^{a*}_\mu(D^\mu c)^a-\frac{1}{2}f_{abc}C^{*a}c^bc^c\right), \tag{4.9}$$

where, as usual, A^{a*}_μ is the shifted antifield,

$$A^{*a\mu}=\hat{A}^{*a\mu}+\partial^\mu\overline{c}^a. \tag{4.10}$$

Finally, taking into account Eqs. (4.6), (4.7) and the expression (4.9), the Slavnov–Taylor identity (4.5) becomes

$$\int d^3x\left(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta\Sigma}{\delta A^{*a\mu}}+\frac{\delta\Sigma}{\delta c^a}\frac{\delta\Sigma}{\delta C^{*a}}\right)=\frac{1}{2}\,\mathscr{B}_\Sigma\Sigma=0, \tag{4.11}$$

with

$$\mathscr{B}_\Sigma=\int d^3x\left(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta}{\delta A^{*a\mu}}+\frac{\delta\Sigma}{\delta A^{*a\mu}}\frac{\delta}{\delta A^{a\mu}}+\frac{\delta\Sigma}{\delta c^a}\frac{\delta}{\delta C^{*a}}+\frac{\delta\Sigma}{\delta C^{*a}}\frac{\delta}{\delta c^a}\right), \tag{4.12}$$

and

$$\mathscr{B}_\Sigma\mathscr{B}_\Sigma=0. \tag{4.13}$$

Repeating the same procedure done in the case of the Yang–Mills theory, for the vector cocycle Ω^{-1}_ν we can write

$$\Omega^{-1}_\nu=\int d^3x[\,\omega^{-1}_\nu]_4\equiv\int d^3x(c^a\partial_\nu C^{*a}-A^a_\mu\partial_\nu A^{*a\mu}), \tag{4.14}$$

and

$$\mathscr{B}_\Sigma\Omega^{-1}_\nu=\mathscr{P}_\nu\Sigma=0. \tag{4.15}$$

In this case, as already mentioned, Ω^{-1}_ν turns out to be exact

$$\Omega^{-1}_\nu=\mathscr{B}_\Sigma\Xi^{-2}_\nu, \tag{4.16}$$

where Ξ^{-2}_ν is an integrated local polynomial of dimension 4 and ghost number -2. From Table II it follows that the most general expression for Ξ^{-2}_ν is given by

$$\Xi^{-2}_\nu=\int d^3x\left(\sigma C^{*a}A^a_\nu+\frac{\tau}{2}\,\epsilon_{\nu\sigma\tau}A^{*a\sigma}A^{*a\tau}\right), \tag{4.17}$$

where σ and τ are arbitrary parameters. Using now the expressions of the linearized Slavnov–Taylor operator (4.12) and of the reduced Chern–Simons action (4.9), for the right-hand side of Eq. (4.16) we have

$$\mathscr{B}_\Sigma\Xi^{-2}_\nu=\int d^3x\left(\sigma A^a_\nu\frac{\delta\Sigma}{\delta c^a}+\sigma C^{*a}\frac{\delta\Sigma}{\delta A^{*a\nu}}-\tau\epsilon_{\nu\sigma\tau}A^{*a\sigma}\frac{\delta\Sigma}{\delta A^a_\tau}\right). \tag{4.18}$$

Comparing then both sides of Eq. (4.16), for the coefficients σ and τ we get

$$\sigma=-1,\quad \tau=\frac{2\pi}{k}, \tag{4.19}$$

so that Eq. (4.16) can be rewritten as

$$\int d^3x\left(A^a_\nu\frac{\delta\Sigma}{\delta c^a}+C^{*a}\frac{\delta\Sigma}{\delta A^{*a\nu}}+\frac{2\pi}{k}\,\epsilon_{\nu\sigma\tau}A^{*a\sigma}\frac{\delta\Sigma}{\delta A^a_\tau}+c^a\partial_\nu C^{*a}-A^a_\mu\partial_\nu A^{*a\mu}\right)=0. \tag{4.20}$$

Finally, moving from the reduced action Σ to the complete action $\mathscr{S}$ of Eq. (4.2) and making use of the gauge condition (4.6), the identity (4.20) can be cast in the form of a linearly broken Ward identity, namely,

$$\mathscr{W}_\nu\mathscr{S}=\Delta^{\rm cl}_\nu, \tag{4.21}$$

with

$$\mathscr{W}_\nu = \int d^3x \left(A^a_\nu \frac{\delta}{\delta c^a} + C^{*a} \frac{\delta}{\delta \hat{A}^{*a\nu}} + \frac{2\pi}{k} \epsilon_{\nu\sigma\tau} (\hat{A}^{*a\sigma} + \partial^\sigma \overline{c^a}) \frac{\delta}{\delta A^a_\tau} + \partial_\nu \overline{c^a} \frac{\delta}{\delta b^a} \right), \quad (4.22)$$

and

$$\Delta^{\rm cl}_\nu = \int d^3x \left(C^{*a} \partial_\nu c^a - \hat{A}^{*a\mu} \partial_\nu A^a_\mu - \frac{2\pi}{k} \epsilon_{\nu\sigma\tau} \hat{A}^{*a\sigma} \partial^\tau b^a \right). \quad (4.23)$$

Equation (4.21) is easily recognized to be the well known topological vector susy Ward identity of the three-dimensional Chern–Simons theory.[2,3] Let us conclude this section by observing that in the case in which instead of a Landau gauge fixing condition we had adopted a Feynman gauge [see Eq. (3.5)], the right-hand side of Eq. (4.21) would have been modified by the additional term

$$\alpha \int d^3x b^a \partial_\nu \overline{c^a}, \quad (4.24)$$

which, being quadratic in the quantum fields, would have spoiled the usefulness of the identity (4.21). We see therefore that the requirement that the breaking term $\Delta^{\rm cl}_\nu$ is at most linear in the quantum fields leads to the vanishing of the gauge parameter, $\alpha=0$, selecting thus the Landau gauge.

B. The four-dimensional BF model

As the second example, we shall present the four-dimensional BF model[1] whose classical invariant action is given by

$$\mathscr{S}_{\rm BF} = -\frac{1}{4} \int d^4x \epsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} B^a_{\rho\sigma}. \quad (4.25)$$

The quantization of this model requires the BV procedure[25] due to the presence of ghosts for ghosts. We shall limit here only to report the final result, reminding the reader the numerous references[1] for the technical details. Using the same notations of Ref. 4, in order to gauge fix the invariant action (4.25) we introduce a set of Lagrange multipliers $(b^a, h^a_\mu, \omega^a, \lambda^a)$, a set of anti-ghosts $(\overline{c^a}, \overline{\xi^a_\mu}, \overline{\phi^a}, e^a)$, and a triple of ghosts (c^a, ξ^a_μ, ϕ^a). For the gauge fixing action S_{gf} we have

$$\begin{aligned} \mathscr{S}_{gf} = \int d^4x (& b^a \partial A^a - \partial^\mu \overline{c^a} (D_\mu c)^a + h^a_\nu \partial_\mu B^{a\mu\nu} + \omega^a \partial \xi^a + h^{a\mu} \partial_\mu e^a + \omega^a \lambda^a - \partial^\mu \overline{\phi^a} ((D_\mu \phi)^a \\ & + f_{abc} c^b \xi^c_\mu) + \tfrac{1}{2} f_{abc} \epsilon^{\mu\nu\rho\sigma} (\partial_\mu \overline{\xi^a_\nu}) (\partial_\rho \overline{\xi^b_\sigma}) \phi^c - \partial^\mu \overline{\xi^a}{}^\nu ((D_\mu \xi_\nu)^a - (D_\nu \xi_\mu)^a + f_{abc} B^b_{\mu\nu} c^c) \\ & - \lambda^a \partial \overline{\xi^a}). \end{aligned} \quad (4.26)$$

The quantum numbers of all the fields and ghosts are assigned as follows:

	A	B	c	ξ	ϕ	$\bar{c}$	$\bar{\xi}$	$\bar{\phi}$	e	b	h	ω	λ
dim.	1	2	0	1	0	2	1	2	2	2	1	2	2
gh-numb.	0	0	1	1	2	−1	−1	−2	0	0	0	−1	1

(4.27)

Introducing a set of antifields $(\hat{A}^{*a\mu}, \hat{B}^{*a\mu\nu}, C^{*a}, \phi^{*a}, \hat{\xi}^{*a\mu})$ associated, respectively, to the fields $(A^{a\mu}, B^{a\mu\nu}, c^a, \phi^a, \xi^{a\mu})$,

$$\mathscr{S}_{\rm ext}=\int d^4x\Big(\frac{1}{2}\,\hat{B}^{*a\mu\nu}((D_\nu\xi_\mu)^a-(D_\mu\xi_\nu)^a-f_{abc}B^b_{\mu\nu}c^c+f_{abc}\epsilon_{\mu\nu\rho\sigma}\partial^\rho\overline{\xi^{b\sigma}}\phi^c)+\hat{\xi}^{*a\mu}((D_\mu\phi)^a$$
$$+f_{abc}c^b\xi^c_\mu)+\frac{1}{8}\,f_{abc}\epsilon_{\mu\nu\rho\sigma}\hat{B}^{*a\mu\nu}\hat{B}^{*b\rho\sigma}\phi^c-\hat{A}^{*a\mu}(D_\mu c)^a+\frac{1}{2}\,f_{abc}C^{*a}c^bc^c$$
$$+f_{abc}\phi^{*a}c^b\phi^c\Big), \tag{4.28}$$

	$\hat{A}^*$	$\hat{B}^*$	C^*	ϕ^*	$\hat{\xi}^*$
dim.	3	2	4	4	3
gh-numb.	−1	−1	−2	−3	−2

(4.29)

we have that the complete action $\mathscr{S}$

$$\mathscr{S}=\mathscr{S}_{\rm BF}+\mathscr{S}_{gf}+\mathscr{S}_{\rm ext}, \tag{4.30}$$

obeys the following Slavnov–Taylor identity:

$$\int d^4x\Big(\frac{\delta\mathscr{S}}{\delta A^a_\mu}\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}+\frac{\delta\mathscr{S}}{\delta c^a}\frac{\delta\mathscr{S}}{\delta C^{*a}}+\frac{1}{2}\frac{\delta\mathscr{S}}{\delta B^a_{\mu\nu}}\frac{\delta\mathscr{S}}{\delta\hat{B}^{*a\mu\nu}}+\frac{\delta\mathscr{S}}{\delta\phi^a}\frac{\delta\mathscr{S}}{\delta\phi^{*a}}+\frac{\delta\mathscr{S}}{\delta\xi^a_\mu}\frac{\delta\mathscr{S}}{\delta\hat{\xi}^{*a\mu}}+h^a_\mu\frac{\delta\mathscr{S}}{\delta\overline{\xi^a_\mu}}$$
$$+b^a\frac{\delta\mathscr{S}}{\delta\overline{c^a}}+\omega^a\frac{\delta\mathscr{S}}{\delta\overline{\phi^a}}+\lambda^a\frac{\delta\mathscr{S}}{\delta e^a}\Big)=0. \tag{4.31}$$

In order to define the reduced action for the BF model let us write down the gauge-fixing conditions,

$$\frac{\delta\mathscr{S}}{\delta b^a}=\partial A^a,\quad\frac{\delta\mathscr{S}}{\delta h^{a\mu}}=\partial_\mu e^a+\partial^\nu B^a_{\nu\mu},$$
$$\frac{\delta\mathscr{S}}{\delta\omega^a}=\lambda^a+\partial\xi^a,\quad\frac{\delta\mathscr{S}}{\delta\lambda^a}=-\partial\overline{\xi^a}-\omega^a. \tag{4.32}$$

Commuting the above conditions with the Slavnov–Taylor identity (4.31) we get the antighost equations (see Tables III and IV),

$$\frac{\delta\mathscr{S}}{\delta\overline{c^a}}+\partial^\mu\frac{\delta\mathscr{S}}{\delta\hat{A}^{*a\mu}}=0,\quad\frac{\delta\mathscr{S}}{\delta\overline{\phi^a}}-\partial^\mu\frac{\delta\mathscr{S}}{\delta\hat{\xi}^{*a\mu}}=0,$$
$$\frac{\delta\mathscr{S}}{\delta e^a}=-\partial h^a,\quad\frac{\delta\mathscr{S}}{\delta\overline{\xi^{a\nu}}}+\partial^\mu\frac{\delta\mathscr{S}}{\delta\hat{B}^{*a\mu\nu}}=-\partial_\nu\lambda^a. \tag{4.33}$$

Introducing the following shifted antifields:

$$A^{*a\mu}=\hat{A}^{*a\mu}+\partial^\mu\overline{c^a},\quad\xi^{*a\mu}=\hat{\xi}^{*a\mu}-\partial^\mu\overline{\phi^a},$$
$$B^{*a\mu\nu}=\hat{B}^{*a\mu\nu}+(\partial^\mu\overline{\xi^{a\nu}}-\partial^\nu\overline{\xi^{a\mu}}), \tag{4.34}$$

for the reduced BF action Σ we get

$$\mathscr{S}=\Sigma+\int d^4x(b^a\partial A^a+h^a_\nu\partial_\mu B^{a\mu\nu}+\omega^a\partial\xi^a+h^{a\mu}\partial_\mu e^a+\omega^a\lambda^a+-\lambda^a\partial\overline{\xi^a}), \tag{4.35}$$

and

$$\Sigma = \int d^4x \Big(-\frac{1}{4}\,\epsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} B^a_{\rho\sigma} + \frac{1}{2}\, B^{*a\mu\nu}((D_\nu \xi_\mu)^a - (D_\mu \xi_\nu)^a - f_{abc} B^b_{\mu\nu} c^c + f_{abc}\epsilon_{\mu\nu\rho\sigma}\partial^\rho \overline{\xi^{b\sigma}} \phi^c)$$
$$+ \xi^{*a\mu}((D_\mu \phi)^a + f_{abc} c^b \xi^c_\mu) + \frac{1}{8}\, f_{abc}\epsilon_{\mu\nu\rho\sigma} B^{*a\mu\nu} B^{*b\rho\sigma} \phi^c - A^{*a\mu}(D_\mu c)^a + \frac{1}{2}\, f_{abc} C^{*a} c^b c^c$$
$$+ f_{abc} \phi^{*a} c^b \phi^c \Big). \tag{4.36}$$

As usual, the reduced action Σ obeys the homogeneous Slavnov–Taylor identity

$$\frac{1}{2}\,\mathscr{B}_\Sigma \Sigma = 0 = \int d^4x \Big(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta\Sigma}{\delta A^{*a\mu}} + \frac{\delta\Sigma}{\delta c^a}\frac{\delta\Sigma}{\delta C^{*a}} + \frac{1}{2}\,\frac{\delta\Sigma}{\delta B^a_{\mu\nu}}\frac{\delta\Sigma}{\delta B^{*a\mu\nu}} + \frac{\delta\Sigma}{\delta \phi^a}\frac{\delta\Sigma}{\delta \phi^{*a}} + \frac{\delta\Sigma}{\delta \xi^a_\mu}\frac{\delta\Sigma}{\delta \xi^{*a\mu}} \Big), \tag{4.37}$$

with

$$\mathscr{B}_\Sigma = \int d^4x \Big(\frac{\delta\Sigma}{\delta A^a_\mu}\frac{\delta}{\delta A^{*a\mu}} + \frac{\delta\Sigma}{\delta A^{*a\mu}}\frac{\delta}{\delta A^a_\mu} + \frac{\delta\Sigma}{\delta c^a}\frac{\delta}{\delta C^{*a}} + \frac{\delta\Sigma}{\delta C^{*a}}\frac{\delta}{\delta c^a} + \frac{1}{2}\,\frac{\delta\Sigma}{\delta B^a_{\mu\nu}}\frac{\delta}{\delta B^{*a\mu\nu}}$$
$$+ \frac{1}{2}\,\frac{\delta\Sigma}{\delta B^{*a\mu\nu}}\frac{\delta}{\delta B^a_{\mu\nu}} + \frac{\delta\Sigma}{\delta \phi^a}\frac{\delta}{\delta \phi^{*a}} + \frac{\delta\Sigma}{\delta \phi^{*a}}\frac{\delta}{\delta \phi^a} + \frac{\delta\Sigma}{\delta \xi^a_\mu}\frac{\delta}{\delta \xi^{*a\mu}} + \frac{\delta\Sigma}{\delta \xi^{*a\mu}}\frac{\delta}{\delta \xi^a_\mu} \Big), \tag{4.38}$$

and

$$\mathscr{B}_\Sigma \mathscr{B}_\Sigma = 0. \tag{4.39}$$

Let us turn now to the invariant vector cocycle (2.5), which in the present case takes the form

$$\Omega^{-1}_\nu \equiv \int d^4x \Big(c^a \partial_\nu C^{*a} - A^a_\mu \partial_\nu A^{*a\mu} + \xi^a_\mu \partial_\nu \xi^{*a\mu} - \phi^a \partial_\nu \phi^{*a} - \frac{1}{2}\, B^a_{\mu\tau} \partial_\nu B^{*a\mu\tau} \Big), \tag{4.40}$$

and

$$\mathscr{B}_\Sigma \Omega^{-1}_\nu = \mathscr{P}_\nu \Sigma = 0. \tag{4.41}$$

As in the case of the Chern–Simons model, the vanishing of the cohomology of the operator $\mathscr{B}_\Sigma$ (Refs. 4 and 10) in the sector of the integrated local polynomials with ghost number -1 and with a free Lorentz index implies that Ω^{-1}_ν is an exact cocycle,

$$\Omega^{-1}_\nu = \mathscr{B}_\Sigma \Xi^{-2}_\nu, \tag{4.42}$$

for some local integrated polynomial Ξ^{-2}_ν of dimension 5 and ghost number -2. In fact, repeating the same procedure done in the previous example, Ξ^{-2}_ν is easily found to be

$$\Xi^{-2}_\nu = \int d^4x \Big(C^{*a} A^a_\nu + \frac{1}{2}\, \epsilon_{\sigma\tau\mu\nu} A^{*a\sigma} B^{*a\tau\mu} - \phi^{*a} \xi^a_\nu - \xi^{*a\mu} B^a_{\mu\nu} \Big). \tag{4.43}$$

Finally, moving from the reduced action (4.36) to the complete one (4.30), we get the linearly broken vector Ward identity,

$$\mathscr{W}_\nu \mathscr{S} = \Delta^{\mathrm{cl}}_\nu, \tag{4.44}$$

with

$$\mathscr{W}^{*}_{\nu}=\int d^4x\Big(\frac{1}{2}\,\epsilon_{\sigma\tau\mu\nu}(\hat{B}^{*a\tau\mu}+\partial^{\tau}\overline{\xi}^{a\mu}-\partial^{\mu}\overline{\xi}^{a\tau})\,\frac{\delta}{\delta A^{a}_{\sigma}}+A^{a}_{\nu}\,\frac{\delta}{\delta c^{a}}-\partial_{\nu}\overline{c}^{a}\,\frac{\delta}{\delta b^{a}}-\frac{1}{2}\,\epsilon_{\sigma\tau\mu\nu}(\hat{A}^{*a\sigma}$$
$$+\partial^{\sigma}\overline{c}^{a})\,\frac{\delta}{\delta B^{a}_{\tau\mu}}-B^{a}_{\mu\nu}\,\frac{\delta}{\delta\xi^{a}_{\mu}}-\xi^{a}_{\nu}\,\frac{\delta}{\delta\phi^{a}}+\overline{\phi}^{a}\,\frac{\delta}{\delta\overline{\xi}^{a\mu}}-\partial_{\nu}\overline{\phi}^{a}\,\frac{\delta}{\delta\omega^{a}}+\partial_{\nu}e^{a}\,\frac{\delta}{\delta\lambda^{a}}$$
$$-(\omega^{a}\delta^{\tau}_{\nu}+\partial_{\nu}\overline{\xi}^{a\tau})\,\frac{\delta}{\delta h^{a\tau}}+C^{*a}\,\frac{\delta}{\delta\hat{A}^{*a\nu}}+\phi^{*a}\,\frac{\delta}{\delta\hat{\xi}^{*a\nu}}-\hat{\xi}^{*a\mu}\,\frac{\delta}{\delta\hat{B}^{*a\mu\nu}}\Big),\tag{4.45}$$

and the classical breaking $\Delta^{\mathrm{cl}}_{\nu}$ given by

$$\Delta^{\mathrm{cl}}_{\nu}=\int d^4x\Big(\hat{A}^{*a\mu}\partial_{\nu}A^{a}_{\mu}-C^{*a}\partial_{\nu}c^{a}-\hat{\xi}^{*a\mu}\partial_{\nu}\xi^{a}_{\mu}+\phi^{*a}\partial_{\nu}\phi^{a}+\frac{1}{2}\,\hat{B}^{*a\tau\mu}\partial_{\nu}B^{a}_{\tau\mu}$$
$$-\frac{1}{2}\,\epsilon_{\sigma\tau\mu\nu}\hat{B}^{*a\tau\mu}\partial^{\sigma}b^{a}+\frac{1}{2}\,\epsilon_{\sigma\tau\mu\nu}\hat{A}^{*a\sigma}\partial^{\tau}h^{a\mu}\Big).\tag{4.46}$$

Equation (4.44) is recognized to be the well known topological vector susy Ward identity of the four-dimensional BF systems.[4] Let us conclude by remarking that the above construction can be easily generalized to the BF systems in higher space–time dimensions, reproducing the results of Refs. 4 and 10. In particular, it is not difficult to check that the requirement that the breaking term (4.46) is at most linear in the quantum fields completely fixes the gauge parameters which would be left free by the Slavnov–Taylor identity.[4]

C. The *b*-*c* ghost system

We present here, as a two-dimensional example of a topological model, the so-called b-c ghost system whose action reads

$$\mathscr{S}_{bc}=\int dz d\overline{z}\; b\,\overline{\partial} c,\tag{4.47}$$

where the fields $b=b_{zz}$ and $c=c^{z}$ are anticommuting and carry, respectively, ghost number -1 and $+1$. The action (4.47) is recognized to be the ghost part of the quantized bosonic string[1,34] and is invariant under the following nonlinear BRST transformations:

$$sc=c\partial c,\quad sb=-(\partial b)c-2b\partial c.\tag{4.48}$$

In particular, the expression in the right-hand side of the BRST transformation of the field b is the component T_{zz} of the energy-momentum tensor corresponding to the action (4.47). This property allows for the topological interpretation of the b-c ghost system.

Transformations (4.48) being nonlinear, one needs to introduce two antifields $(b^{*}=b^{z*}_{\overline{z}},c^{*}=c^{*}_{zz\overline{z}})$ of ghost number 0 and -2,

$$\mathscr{S}_{\mathrm{ext}}=\int dz d\overline{z}(c^{*}c\partial c+b^{*}(c\partial b-2b\partial c)).\tag{4.49}$$

The complete action

$$\mathscr{S}=\mathscr{S}_{bc}+\mathscr{S}_{\mathrm{ext}},\tag{4.50}$$

obeys thus the Slavnov–Taylor identity

$$\int dz d\overline{z}\Big(\frac{\delta\mathscr{S}}{\delta b}\,\frac{\delta\mathscr{S}}{\delta b^{*}}+\frac{\delta\mathscr{S}}{\delta c}\,\frac{\delta\mathscr{S}}{\delta c^{*}}\Big)=\frac{1}{2}\,\mathscr{B}_{\mathscr{S}}\mathscr{S}=0,\tag{4.51}$$

$\mathscr{B}_{\mathscr{S}}$ denoting the linearized operator

$$\mathscr{B}_{\mathscr{S}}=\int dzd\overline{z}\left(\frac{\delta\mathscr{S}}{\delta b}\frac{\delta}{\delta b^*}+\frac{\delta\mathscr{S}}{\delta b^*}\frac{\delta}{\delta b}+\frac{\delta\mathscr{S}}{\delta c}\frac{\delta}{\delta c^*}+\frac{\delta\mathscr{S}}{\delta c^*}\frac{\delta\mathscr{S}}{\delta c}\right), \tag{4.52}$$

and

$$\mathscr{B}_{\mathscr{S}}\mathscr{B}_{\mathscr{S}}=0. \tag{4.53}$$

	c	b	c^*	b^*
dim	0	1	1	0
$gh-numb$	1	−1	−2	0

(4.53)

Concerning now the vector cocycle (2.5), here written in components, we have

$$\begin{aligned}\Omega_z^{-1}&=\int dzd\overline{z}(b\partial b^*+c\partial c^*),\\ \Omega_{\overline{z}}^{-1}&=\int dzd\overline{z}(b\partial\overline{b}^*+c\partial\overline{c}^*),\end{aligned} \tag{4.55}$$

and

$$\mathscr{B}_{\mathscr{S}}\Omega_z^{-1}=\mathscr{P}_z\mathscr{S}=0,\quad \mathscr{B}_{\mathscr{S}}\Omega_{\overline{z}}^{-1}=\mathscr{P}_{\overline{z}}\mathscr{S}=0. \tag{4.56}$$

As done before, we look for the solution of the BRST exact conditions

$$\Omega_z^{-1}=\mathscr{B}_{\mathscr{S}}\Xi_z^{-2},\quad \Omega_{\overline{z}}^{-1}=\mathscr{B}_{\mathscr{S}}\Xi_{\overline{z}}^{-2}, \tag{4.57}$$

for some local integrated polynomials $(\Xi_z^{-2},\Xi_{\overline{z}}^{-2})$ of ghost number -2 and dimension 1 (see Table 5). After some trivial algebraic manipulations one easily finds

$$\Xi_z^{-2}=-\int dzd\overline{z}\,c^*,\quad \Xi_{\overline{z}}^{-2}=-\int dzd\overline{z}\,c^*b^*. \tag{4.58}$$

Turning then Eqs. (4.57) into contact terms by means of the linearized operator $\mathscr{B}_{\mathscr{S}}$ of Eq. (4.52), we get the two linearly broken Ward identities

$$\int dzd\overline{z}\frac{\delta\mathscr{S}}{\delta c}=-\int dzd\overline{z}(b\partial b^*+c\partial c^*), \tag{4.59}$$

and

$$\int dzd\overline{z}\left(b^*\frac{\delta\mathscr{S}}{\delta c}+c^*\frac{\delta\mathscr{S}}{\delta b}\right)=-\int dzd\overline{z}(b\partial\overline{b}^*+c\partial\overline{c}^*), \tag{4.60}$$

which are nothing but the topological vector susy Ward identities of the b-c ghost system.[13,17,10]

D. The b-c β-γ ghost system

It is known that the b-c system can be generalized to include the so-called β-γ ghosts,

$$\mathscr{S}_{bc\beta\gamma}=\int dzd\overline{z}\,(b\partial\overline{c}+\beta\partial\overline{\gamma}), \tag{4.61}$$

where $\beta=\beta_{zzz}$ and $\gamma=\gamma^{zz}$ are anticommuting fields of ghost number -1 and 1, respectively. The action (4.61) is the simplest example of a b-c β-γ system and can be easily generalized to the case of ghosts of higher conformal weight.[35] The BRST transformations read now

$$sc = c\partial c, \quad s\gamma = c\partial\gamma - 2(\partial c)\gamma,$$

$$sb = -(\partial b)c - 2b\partial c - 2(\partial\beta)\gamma - 3\beta\partial\gamma, \tag{4.62}$$

$$s\beta = -(\partial\beta)c - 3\beta\partial c.$$

Introducing external fields $(c^*, b^*, \beta^*, \gamma^*)$,

$$\mathscr{S}_{\text{ext}} = \int dz d\overline{z}\,(c^* sc + b^* sb + \beta^* s\beta + \gamma^* s\gamma), \tag{4.63}$$

for the Slavnov–Taylor identity we get

$$\int dz d\overline{z}\left(\frac{\delta\mathscr{S}}{\delta b}\frac{\delta\mathscr{S}}{\delta b^*} + \frac{\delta\mathscr{S}}{\delta c}\frac{\delta\mathscr{S}}{\delta c^*} + \frac{\delta\mathscr{S}}{\delta\gamma}\frac{\delta\mathscr{S}}{\delta\gamma^*} + \frac{\delta\mathscr{S}}{\delta\beta}\frac{\delta\mathscr{S}}{\delta\beta^*}\right) = 0, \tag{4.64}$$

with

$$\mathscr{S} = \mathscr{S}_{bc\beta\gamma} + \mathscr{S}_{\text{ext}}. \tag{4.65}$$

Finally, following the same procedure of the previous example, the linearly broken Ward identities of Eqs. (4.59), (4.60) generalize now to

$$\int dz d\overline{z}\,\frac{\delta\mathscr{S}}{\delta c} = -\int dz d\overline{z}\,(b\partial b^* + c\partial c^* + \gamma\partial\gamma^* + \beta\partial\beta^*), \tag{4.66}$$

and

$$\int dz d\overline{z}\left(b^*\frac{\delta\mathscr{S}}{\delta c} + c^*\frac{\delta\mathscr{S}}{\delta b} + \beta^*\frac{\delta\mathscr{S}}{\delta\gamma} + \gamma^*\frac{\delta\mathscr{S}}{\delta\beta}\right) = \overline{\Delta}^{\text{cl}}, \tag{4.67}$$

with

$$\overline{\Delta}^{\text{cl}} = -\int dz d\overline{z}\,(b\partial\overline{b}^* + c\partial\overline{c}^* + \gamma\partial\overline{\gamma}^* + \beta\partial\overline{\beta}^*). \tag{4.68}$$

V. THE CASE OF THE RIGID GAUGE INVARIANCE: THE LANDAU GHOST EQUATION

In this last section we shall analyse another interesting example of a linearly broken Ward identity whose contact terms can be characterized in the same purely algebraic cohomological way of the topological vector susy. This Ward identity is not related to a specific gauge model, being present in all the cases in which the rigid gauge invariance is an exact symmetry of the action. In order to present its derivation in a detailed way we shall consider the four dimensional pure Yang–Mills theory of the Sec. III, the generalization to other models being straightforward.

The rigid gauge invariance is an exact symmetry of the Yang–Mills action (3.9), expressing the simple fact that all the fields and antifields belong to the adjoint representation of the gauge group,

$$\mathscr{R}_a^{\text{rig}}\Sigma = \int d^4x f_{abc}\left(A^b_\mu\frac{\delta\Sigma}{\delta A^c_\mu} + A^{*b}_\mu\frac{\delta\Sigma}{\delta A^{*c}_\mu} + c^b\frac{\delta\Sigma}{\delta c^c} + C^{*b}\frac{\delta\Sigma}{\delta C^{*c}}\right) = 0. \tag{5.1}$$

Let us consider now the following integrated local polynomial, linear in the (shifted) antifields (A^*_μ, C^*), of ghost number -1, dimension four, and possessing a group index in the adjoint representation

$$\Omega_a^{-1} = \int d^4x f_{abc}(A^{b\mu}A^{*c}_\mu - c^b C^{*c}) \tag{5.2}$$

As in the case of the vector cocycle of Eq. (2.5), the above expression turns out to be the BRST invariant. In fact we have

$$\mathscr{B}_\Sigma \Omega_a^{-1} = \mathscr{R}_a^{\rm rig}\Sigma = 0, \tag{5.3}$$

due to the rigid invariance (5.1). The operator $\mathscr{B}_\Sigma$ appearing in the above equation is the linearized Slavnov–Taylor operator defined in Eq. (3.11). However, unlike the vector cocycle Ω_ν^{-1}, the cocycle Ω_a^{-1} is always $\mathscr{B}_\Sigma$-exact, due to a very well known result of the BRST cohomology stating that there are no nontrivial cohomology classes with one group index.[24] Therefore we may write

$$\Omega_a^{-1} = \mathscr{B}_\Sigma \Xi_a^{-2}, \tag{5.4}$$

for some local integrated polynomial Ξ_a^{-2} of ghost number -2 and dimension four. From Table I it follows that the most general form for Ξ_a^{-2} can be written as

$$\Xi_a^{-2} = \zeta \int d^4x C_a^*, \tag{5.5}$$

ζ being an arbitrary parameter. Using the expression of the operator $\mathscr{B}_\Sigma$ of Eq. (3.11) and of the reduced Yang–Mills action Σ given in Eq. (3.9), the condition (5.4) becomes

$$\int d^4x f_{abc}(A^{b\mu}A_\mu^{*c} - c^b C^{*c}) = \zeta \int d^4x \, \frac{\delta \Sigma}{\delta c^a} = \zeta \int d^4x f_{abc}(A^{b\mu}A_\mu^{*c} - c^b C^{*c}), \tag{5.6}$$

which gives $\zeta = 1$. Moving now from the reduced action Σ to the complete Yang–Mills action $\mathscr{S}$ (3.9) and making use of the gauge condition (3.5) and of the definition (3.7), Eq. (5.6) takes the form

$$\int d^4x \left(\frac{\delta \mathscr{S}}{\delta c^a} - f_{abc}\overline{c}^b \frac{\delta \mathscr{S}}{\delta b^c} \right) = \int d^4x f_{abc}(A^{b\mu}\hat{A}_\mu^{*c} - c^b C^{*c}) + \alpha \int d^4x f_{abc} b^b \overline{c}^c. \tag{5.7}$$

We observe that the left-hand side of this equation, besides a pure linear breaking, contains a term which is quadratic in the quantum fields, i.e., $\alpha f_{abc} b^b \overline{c}^c$. This term, being subject to renormalization, would have been defined as an insertion, spoiling then the usefulness of Eq. (5.7). Therefore we see that the requirement that the breaking term is at most linear in the quantum fields implies again the vanishing of the gauge parameter, $\alpha = 0$. Setting then $\alpha = 0$ in the gauge condition (3.5), we obtain the linearly broken identity

$$\int d^4x \left(\frac{\delta \mathscr{S}}{\delta c^a} - f_{abc}\overline{c}^b \frac{\delta \mathscr{S}}{\delta b^c} \right) = \int d^4x f_{abc}(A^{b\mu}\hat{A}_\mu^{*c} - c^b C^{*c}), \tag{5.8}$$

which is recognized to be the so-called *ghost equation* Ward identity,[23] always present in the Landau gauge. Let us conclude by recalling that the ghost equation (5.8), although valid only in the Landau gauge, turns out to be a very useful tool in order to study the ultraviolet finiteness properties of a large class of gauge invariant local field polynomials.[36] These gauge invariant polynomials can be promoted at the quantum level to local insertions whose anomalous dimensions are independent from the gauge fixing parameter α. Therefore they can be studied without loss of generality in the Landau gauge. In particular, the ghost equation (5.8) allows us to prove the vanishing of the anomalous dimensions of the invariant ghost monomials $\mathrm{tr}(c^{2n+1})$, which are deeply related to the gauge anomalies and to the generalized Chern–Simons terms.[10] (We have used here the matrix notation $c = c^a T_a$, T_a being the generators of the gauge group, $[T_a, T_b] = i f_{ab}{}^c T_c$.) This result is of great importance in order to prove the Adler–Bardeen nonrenormalization theorem for the gauge and the $U(1)$ axial anomalies.[36,37,10] We mention, finally, that the ghost equation (5.8) has been proven to be renormalizable.[10,23] Recently, it has been extended to the case of $N = 1$ supersymmetric gauge theories in superspace.[38]

VI. CONCLUSION

A purely algebraic characterization of the topological vector susy and of the Landau ghost equation Ward identities has been provided. These Ward identities, always linearly broken, are obtained by exploiting the BRST exactness condition of suitable antifield cocycles with ghost number -1. Applications to other kinds of linearly broken Ward identities as well as to other TFT and to superspace supersymmetric models are under investigation.

ACKNOWLEDGMENT

The Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq-Brazil is gratefully acknowledged for the financial support.

[1] D. Birmingham, M. Blau, M. Rakowski, and G. Thompson, Phys. Rep. **209**, 129 (1991).
[2] D. Birmingham, M. Rakowski, and G. Thompson, Nucl. Phys. B **329**, 83 (1990); D. Birmingham and M. Rakowski, Mod. Phys. Lett. A **4**, 1753 (1989).
[3] F. Delduc, F. Gieres, and S. P. Sorella, Phys. Lett. B **225**, 367 (1989); F. Delduc, C. Lucchesi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **346**, 313 (1990).
[4] E. Guadagnini, N. Maggiore, and S. P. Sorella, Phys. Lett. B **255**, 65 (1991); N. Maggiore and S. P. Sorella, Int. J. Mod. Phys. A **8**, 929 (1993); C. Lucchesi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **395**, 325 (1993).
[5] D. Birmingham and M. Rakowski, Phys. Lett. B **269**, 103 (1991); **275**, 289 (1992); **289**, 271 (1992); A. Brandhuber, O. Moritsch, M. W. de Oliveira, O. Piguet, and M. Schweda, Nucl. Phys. B **341**, 173 (1994).
[6] D. Birmingham, M. Blau, and G. Thompson, Int. J. Mod. Phys. A **5**, 24, 4721 (1990).
[7] E. Witten, Commun. Math. Phys. **117**, 353 (1988).
[8] D. Birmingham and M. Rakowski, Phys. Lett. B **272**, 217 (1991); **273**, 74 (1991).
[9] M. Mariño, ''The Geometry of Supersymmetric Gauge Theories in Four Dimension,'' US-Ft-2/97, hep-th/9701128; F. Fucito, A. Tanzini, L. C. Q. Vilar, O. S. Ventura, C. A. G. Sasaki, and S. P. Sorella, ''Algebraic Renormalization: Perturbative twisted considerations on topological Yang–Mills theory and on $N=2$ supersymmetric gauge theories, CBPF-NF-043/97, hep-th/9707209.
[10] O. Piguet and S. P. Sorella, *Algebraic Renormalization*, Monographs Series (Springer, Berlin, 1995), Vol. m28.
[11] S. P. Sorella, Commun. Math. Phys. **157**, 231 (1993); S. P. Sorella and L. Tataru, Phys. Lett. B **324**, 351 (1994).
[12] M. Werneck de Oliveira and S. P. Sorella, Int. J. Mod. Phys. A **9**, 2979 (1994); O. Moritsch, M. Schweda, and S. P. Sorella, Class. Quantum Grav. **11**, 1225 (1994); O. Moritsch and M. Schweda, Helv. Phys. Acta **67**, 289 (1994); P. A. Blaga, O. Moritsch, M. Schweda, T. Sommer, L. Tataru, H. Zerrouki, Phys. Rev. D **51**, 2792 (1995); O. Moritsch, M. Schweda, T. Sommer, L. Tataru, H. Zerrouki, ''BRST Cohomology of Yang–Mills Gauge Fields in the Presence of Gravity in Ashtekar Variables,'' hep-th/9409081; S. Emery, O. Moritsch, M. Schweda, T. Sommer, and H. Zerrouki, Helv. Phys. Acta **68**, 167 (1995); O. Piguet, ''On the Role of Vector Supersymmetry in Topological Field Theory,'' UGVA-DPT 1995/02-880, hep-th/9502033.
[13] M. Werneck de Oliveira, M. Schweda, and S. P. Sorella, Phys. Lett. B **315**, 93 (1993); L. Tataru and I. V. Vancea, Int. J. Mod. Phys. **11**, 375 (1996).
[14] A. Boresch, M. Schweda, and S. P. Sorella, Phys. Lett. B **328**, 36 (1994).
[15] M. Carvalho, L. C. Queiroz Vilar, and S. P. Sorella, Int. J. Mod. Phys. A **10**, 27,3877 (1995).
[16] S. Axelrod and I. M. Singer, *Chern-Simons Perturbation Theory, XXth Conference on Differential Geometry Methods in Physics*, June 1991, hep-th/9110056; S. Axelrod and I. M. Singer, J. Diff. Geom. **39**, 173 (1994).
[17] M. Carvalho, L. C. Queiroz Vilar, C. A. G. Sasaki, and S. P. Sorella, J. Math. Phys. **37**, 5310,5325 (1996).
[18] L. C. Queiroz Vilar, C. A. G. Sasaki, and S. P. Sorella, ''Superspace Descent Equations and Zero Curvature Formalism of the Four Dimensional N=1 Supersymmetric Yang–Mills Theories,'' CBPF-NF-049/96, hep-th 9610220, Int. J. Mod. Phys. A (submitted).
[19] R. Brooks, J. G. Demers, and C. Lucchesi, Nucl. Phys. B **415**, 353 (1994).
[20] N. Maggiore and S. P. Sorella, Nucl. Phys. B **377**, 236 (1992).
[21] G. Barnich, F. Brandt, and M. Henneaux, Commun. Math. Phys. **174**, 57,93 (1995); M. Henneaux, J. Pure Appl. Algebra **100**, 3 (1995).
[22] F. Brandt, M. Henneaux, and A. Wilch, Phys. Lett. B **387**, 320 (1996); F. Brandt, M. Henneaux, and A. Wilch, ''Ward Identities for Rigid Symmetries of Higher Order,'' hep-th 9611056.
[23] A. Blasi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **356**, 154 (1991).
[24] F. Brandt, N. Dragon, and M. Kreuzer, Phys. Lett. B **231**, 263 (1989); Nucl. Phys. B **332**, 224,250 (1990;) M. Dubois-Violette, M. Henneaux, M. Talon, and C. M. Viallet Phys. Lett. B **289**, 361 (1992).
[25] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. B **102**, 27 (1981); Phys. Rev. D **28**, 2567 (1983).
[26] S. Ouvry, R. Stora, and P. Van Baal, Phys. Lett. B **220**, 159 (1989); A. Blasi and R. Collina, *ibid.* **222**, 159 (1989); S. P. Sorella, *ibid.* **228**, 159 (1989)
[27] J. Kalkman, Commun. Math. Phys. **153**, 447 (1993); R. Stora, *Equivariant Cohomology and Topological Theories*, in *BRS Symmetry*, edited by M. Abe, N. Nakanishi, and Iojima (Universal Academy Press, Tokyo, Japan, 1996); R. Stora, F. Thuillier, and J. C. Wallet, *Algebraic Structure of Cohomological Field Theory Models and Equivariant Cohomology*, lectures given at the First Caribbean Spring School of Mathematics and Theoretical Physics, edited by R. Coquereaux, M. Dubois-Violette, and P. Flad (World Scientific, Singapore, 1995); R. Stora, ''Exercise in Equivariant Cohomology,'' ENSLAPP-A-619/96, hep-th/9611114; ''De la fixation de jauge consideree comme un des beaux arts et de la symetrie de Slavnov qui s'ensuit,'' ENSLAPP-A-620/96, hep-th/9611115; ''Exercises in Equivariant Cohomology and Topological Theories,'' hep-th/9611116.

[28] C. Becchi, R. Collina, and C. Imbimbo Phys. Lett. B **322**, 79 (1994); C. Becchi and C. Imbimbo, Nucl. Phys. B **462**, 571 (1996); C. Becchi, C. Imbimbo, and S. Giusto, ''Gauge dependence in topological gauge theories,'' GEF-Th/96-20, hep-th/9611113.
[29] F. Delduc, N. Maggiore, O. Piguet, and S. Wolf, ''Note on Constrained Cohomology,'' UGVA-DPT 1996/05-925, hep-th/9605158, Phys. Lett. B (to be published).
[30] L. Baulieu and I. M. Singer, Nucl. Phys. B **15**, 12 (1988); Commun. Math. Phys. **135**, 253 (1991).
[31] J. M. F. Labastida and M. Pernici, Phys. Lett. B **212**, 56 (1988).
[32] E. Witten, Commun. Math. Phys. **118**, 411 (1988).
[33] J. H. Lowenstein, Phys. Rev. D **4**, 2281 (1971); Commun. Math. Phys. **24**, 1 (1971); Y. M. P. Lam, Phys. Rev. D **6**, 2145 (1972); D**7**, 2943 (1973); T. E. Clark and J. H. Lowenstein, Nucl. Phys. B **113**, 109 (1976).
[34] M. Green, J. Schwarz, and E. Witten, *Superstring Theory* (Cambridge, University Press, Cambridge, 1987). Vol. I and II.
[35] S. Lazzarini and R. Stora, *Ward Identities for Lagrangian Conformal Models*, in *Knots, Topology and Quantum Field Theory*, 13th ed., Johns Hopkins Workshop, edited by L. Lusanna (World Scientific, Singapore, 1989).
[36] O. Piguet and S. P. Sorella, Nucl. Phys. B **381**, 373 (1992).
[37] O. Piguet and S. P. Sorella, Nucl. Phys. B **395**, 661 (1993).
[38] O. Piguet and S. P. Sorella, Phys. Lett. B **371**, 238 (1996).

Quantization of a particle in a background Yang–Mills field

Yihren Wu
Department of Mathematics, Hofstra University, Hempstead, New York 11550

(Received 27 June 1997; accepted for publication 29 October 1997)

Two classes of observables defined on the phase space of a particle are quantized, and the effects of the Yang–Mills field are discussed in the context of geometric quantization. © *1998 American Institute of Physics.* [S0022-2488(98)02202-6]

I. INTRODUCTION

Let Q be a Riemannian manifold considered as the configuration space of a particle, the purpose of this paper is to discuss the quantization of the observables on the phase space T^*Q of this particle when it is moving under the influence of a background Yang–Mills field, the Yang–Mills potential is a connection α on a principal bundle N over Q.

The free G action on N can be lifted to a Hamiltonian G action on T^*N with an equivariant moment map $J:T^*N\to\mathfrak{g}^*$. Let $\mu\epsilon\mathfrak{g}^*$, and denote by $\mathcal{O}_\mu$ the coadjoint orbit through μ. Then $J^{-1}(\mathcal{O}_\mu)/G$ has a canonical symplectic structure given by the Marsden–Weinstein reduction.[1] This reduced phase space is the appropriate phase space of a particle in a background Yang–Mills field α of charge μ.[2]

We will denote by $\mathcal{Q}(X)$ the quantization of the phase space X, suppressing in our notation the choices of polarizations and prequantization line bundles, etc., via the standard procedure of geometric quantization.[3,4] Suppose we choose the vertical polarization on T^*N so that the quantization $\mathcal{Q}(T^*N)$ of T^*N gives $L^2(N)$. Moreover, suppose the coadjoint orbit $\mathcal{O}_\mu$ is integral, so that the quantization of this coadjoint orbit gives an irreducible representation space $\mathcal{H}_\mu$ of G.[5,6] A theorem of Guillemin–Sternberg[7] (also see Ref. 8) then suggests that the quantization of $J^{-1}(\mathcal{O}_\mu)/G$ is given by $\mathrm{Hom}_G(\mathcal{H}_\mu, L^2(N))$, the space of G-equivariant linear maps from $\mathcal{H}_\mu$ to $L^2(N)$. And this result holds independent of whether there is a Yang–Mills field present in the background. Thus, when some technical assumptions are made so that the procedure of geometric quantization can be carried out smoothly, the Yang–Mills field plays no role in the quantization of the phase space $J^{-1}(\mathcal{O}_\mu)/G$.

We will discuss the effect of the Yang–Mills field in quantizing observables that are lifted from functions on T^*Q. We will show that the resulting quantum operators are expressed in terms of the covariant derivatives, which is defined by the connection α. In particular, we will show that the quantum operators for f of the form $\frac{1}{2}\|p\|^2+V(q)$ are expressed in terms of the covariant Laplace operator and the Ricci curvature. This is obtained by a standard Blattner–Kostant–Sternberg (BKS) pairing approach.[9] Our results are in agreement with those of Landsman,[10] who arrived at the conclusion via deformation quantization.

Outline of this paper is as follows; In Sec. II we give a detailed exposition of our problem in order to standardize the notations used throughout this paper. As a prelude to our result, we note that if the gauge group is Abelian, we will recover the Dirac quantization of a charge particle in the presense of an electromagnetic field. In Sec. III we introduce local coordinates to facilitate our calculation, and state some results concerning the Hamiltonian vector fields for our observables. We follow closely the treatment of[8] on the polarization chosen for the phase space in Sec. IV. Our results when f is polarization preserving are given in Sec. V, and the quantization of $\frac{1}{2}\|p\|^2+V(q)$ using BKS pairing appears in Sec. VI.

II. PRELIMINARY DISCUSSIONS

Let N be a principal G bundle over Q where G is compact, with Lie algebra $\mathfrak{g}$, and the group action is on the right. We define two functions $R_g:N\to N$ and $\hat{n}:G\to N$, where

 © 1998 American Institute of Physics

$$R_g(n)=\hat{n}(g)=ng, \tag{1}$$

and we denote by F_* the Jacobian of F. A connection is a linear map $\alpha(n):T_nN\to\mathfrak{g}$ for each $n\in N$, satisfying
(i) $\mathrm{Ad}_{g^{-1}}\alpha(n)=\alpha(ng)R_{g*}:T_nN\to\mathfrak{g}$,
(ii) $\alpha(n)\hat{n}_*=\mathrm{Id}:\mathfrak{g}\to\mathfrak{g}$.

The free G action on N can be lifted to a Hamiltonian G action on T^*N with moment map $J:T^*N\to\mathfrak{g}^*$ given by $J(\xi,n)=\hat{n}^*\cdot\xi$.

Let $N^{\#}$ be the pullback bundle over T^*Q: Explicitly,

$$N^{\#}=\{(p,n)|p\in T_q^*Q,\ \text{where}\ \ q=\pi(n)\}.$$

Define a diffeomorphism,

$$\begin{gathered}\chi:N^{\#}\times\mathfrak{g}^*\to T^*N,\\ (p,n,\mu)\to(\xi,n)\quad \text{where}\ \ \xi=\pi^*(p)+\langle\mu,\alpha\rangle\in T_n^*N.\end{gathered} \tag{2}$$

This map in turn induces an α-dependent projection $\pi_\alpha:T^*N\to T^*Q$, and their corresponding symplectic forms are related by

$$\Omega_{T^*N}=\pi_\alpha^*\Omega_{T^*Q}+\mathrm{d}\langle\mu,\alpha\rangle. \tag{3}$$

One shows that the moment map is simply the projection $N^{\#}\times\mathfrak{g}^*\to\mathfrak{g}^*$, i.e., $\chi^{-1}(J^{-1}(\mu))=N^{\#}\times\{\mu\}$. Thus, for each μ, the G action on $N^{\#}$ induces a G action on $J^{-1}(\mu)$. This action is noncanonical ($R_g^*\Omega_{T^*N}\neq\Omega_{T^*N}$) in general:

$$(\xi,n)\to(\xi_g,ng),\quad \text{where}\ \ \xi_g=R_{g^{-1}}^*\xi+R_{g^{-1}}^*[(\mathrm{Ad}_{g^{-1}}^*-\mathrm{Id})\mu]\alpha(n).$$

However, this action coincides with the canonical G action when restricted to the isotropy subgroup,

$$H=\{g\in G|\mathrm{Ad}_g^*\mu=\mu\},$$

of μ. (There $\mathrm{Ad}_{g^{-1}}^*\mu=\mu$ and $\xi_g=R_{g^{-1}}^*\xi$.) The relevant phase space becomes $J^{-1}(\mathcal{O}_\mu)/G=J^{-1}(\mu)/H=N^{\#}/H\times\{\mu\}$.

Given an observable on the phase space of the particle $f:T^*Q\to\mathbf{R}$, by the projection π_α, the pullback map, which we continue to denote by f,

$$f=\pi_\alpha^*f:T^*N\simeq N^{\#}\times\mathfrak{g}^*\to\mathbf{R}, \tag{4}$$

is invariant with respect to both the canonical G action and the noncanonical one. In particular, f is independent on the charge variables $\mu\in\mathfrak{g}^*$.

We assume that $\mathcal{O}_\mu$ is integral, so that $\mathcal{Q}(\mathcal{O}_\mu)=\mathcal{H}_\mu$ is an irreducible representation space of G induced by $\rho_\mu:H\to U(1)$. Choose an orthonormal basis $\{\phi_i\}$ for $\mathcal{H}_\mu$, where ϕ_i is a holomorphic function on the Kähler manifold G/H.[5] Then $\Psi\in\mathcal{Q}(J^{-1}(\mu)/H)=\mathrm{Hom}_G(\mathcal{H}_\mu,L^2(N))$ is determined by $\Psi(\phi_i)=\Psi_i\in L^2(N)$. Using orthonormality of ϕ_i and G equivariance of Ψ, we write

$$\Psi=\sum\ \Psi_i\phi_i:N\times G/H\to N\times_G G/H\to\mathbf{C},$$

which is uniquely determined by $\psi=\Psi(-,eH):N\to\mathbf{C}$ with the condition $\psi(nh)=\rho_\mu(h^{-1})\psi(n)$ for all $h\in H$. So we see that the Yang–Mills potential plays no role in quantizing the relevant phase space, it simply picks up the multiplicity of the charge sector $\mathcal{O}_\mu$ in $L^2(N)$ (cf. Refs. 7 and 8).

If we are to quantize an observable that is a pullback of f on the phase space of the particle T^*Q, the connection α plays an important role. As an illustration, suppose the charge $\mu\in\mathfrak{g}^*$ is G

invariant, then $H=G$. This will be the case if, for instance, G is Abelian. Under this condition, $J^{-1}(\mu)/H$ is diffeomorphic to T^*Q via the projection π_α, and the canonical symplectic form on the reduction space pushes forward onto T^*Q. So $J^{-1}(\mu)/H$ is symplectomorphic to T^*Q if T^*Q is equipped with the ''effective'' symplectic form $\Omega_{\text{eff}}=\Omega_{T^*Q}+\langle\mu,\Omega_{\mathbf{\nabla}}\rangle$, where $\Omega_{\mathbf{\nabla}}$ is a two-form on Q that pulls back to the curvature form $d\alpha$ on N.[11] It is with respect to this effective symplectic form that the quantization procedure must be carried out. Since the adjustment is a two-form on Q, quantization of T^*Q using the vertical polarization still gives $L^2(Q)$. However, Hamiltonian vector fields $\mathscr{H}_{f_i}$ associated with observables f_i and the Poisson bracket are defined in terms of the effective form

$$\Omega_{\text{eff}}(\mathscr{H}_{f_i},-)=-df_i,\quad \{f_1,f_2\}=\Omega_{\text{eff}}(\mathscr{H}_{f_1},\mathscr{H}_{f_2}).$$

The quantization of observables must preserve the Poisson bracket,

$$[\mathscr{Q}(f_1),\mathscr{Q}(f_2)]=i\hbar\mathscr{Q}(\{f_1,f_2\}).$$

When we carry out the geometric quantization with respect to Ω_{eff}, the result is the Dirac quantization with α as the vector potential associated with an electromagnetic field (cf. Ref. 3).

III. THE HAMILTONIAN VECTOR FIELDS

Let us first introduce local canonical coordinates (ξ_i,n_i) on T^*N so that $\Omega_{T^*N}=\mathrm{d}\xi_i\ \mathrm{d}n_i$, similarly (p_a,q_a) on T^*Q with $\Omega_{T^*Q}=\mathrm{d}p_a\ \mathrm{d}q_a$, and let $\alpha=A_{si}\ \mathrm{d}n_i$, where $i=1\cdots$ dim N, $a=1\cdots$ dim Q, $s=1\cdots$ dim G, and repeated indices are summed. For each $n\in N$, let us denote the horizontal lift $T_qQ\to T_nN$ by the matrix $M_{i\sigma}(n)$, $i=1\cdots$ dim N, $\sigma=1\cdots$ dim Q. We have

$$\frac{\partial q_a}{\partial n_i}M_{ib}=\delta_{ab}, \tag{5}$$

and the covariant derivative is the horizontal lift of $\partial/\partial q_a$:

$$\mathscr{D}_a=M_{ia}\frac{\partial}{\partial n_i}. \tag{6}$$

In these coordinates, the canonical one-form and the symplectic two-form of T^*N can be calculated using (3),

$$\begin{aligned}\chi^*\xi_i\ \mathrm{d}n_i&=\left(p_a\frac{\partial q_a}{\partial n_i}+\mu_sA_{si}\right)\mathrm{d}n_i,\\ \chi^*\Omega_{T^*N}&=\frac{\partial q_a}{\partial n_i}\mathrm{d}p_a\ \mathrm{d}n_i+\mu_s\frac{\partial A_{si}}{\partial n_j}\mathrm{d}n_j\ \mathrm{d}n_i+A_{si}\ \mathrm{d}\mu_s\ \mathrm{d}n_i.\end{aligned} \tag{7}$$

Let $f:N^{\#}\times\mathfrak{g}^*\to\mathbf{R}$ be a pullback function from T^*Q, and let $\mathscr{H}_f$ be its Hamitonian vector field,

$$\mathscr{H}_f=B_a\frac{\partial}{\partial p_a}+C_i\frac{\partial}{\partial n_i}+U_s\frac{\partial}{\partial\mu_s}.$$

Using $\chi^*\Omega_{T^*N}(\mathscr{H}_f,-)=-\mathrm{d}f$, we get

$$\begin{aligned}\frac{\partial f}{\partial q_a}\frac{\partial q_a}{\partial n_i}&=-B_a\frac{\partial q_a}{\partial n_i}+\mu_s\left(\frac{\partial A_{sj}}{\partial n_i}-\frac{\partial A_{si}}{\partial n_j}\right)C_j-A_{si}U_s,\\ \frac{\partial f}{\partial p_a}&=\frac{\partial q_a}{\partial n_i}C_i,\end{aligned} \tag{8}$$

$$\frac{\partial f}{\partial \mu_s} = A_{si} C_i .$$

Since f is invariant with respect to the canonical G action, $\mathscr{H}_f$ is tangent to the subspace $J^{-1}(\mu) \simeq N^{\#} \times \{\mu\}$, thus $U_s = 0$. As remarked after (4), f is independent of μ, thus $A_{si} C_i = 0$, which implies $\mathscr{H}_f$ is horizontal. Moreover, letting $B_a = -\partial f / \partial q_a + E_a$, we have

$$\mathscr{H}_f = \left[-\frac{\partial f}{\partial q_a} \frac{\partial}{\partial p_a} + C_i \frac{\partial}{\partial n_i} \right] + E_a \frac{\partial}{\partial p_a}.$$

The terms in the bracket is the horizontal lift of the Hamiltonian vector field of f with respect to the usual symplectic form Ω_{T^*Q} on T^*Q. E_a satisfies

$$E_a \frac{\partial q_a}{\partial n_i} = \mu_s \left(\frac{\partial A_{sj}}{\partial n_i} - \frac{\partial A_{si}}{\partial n_j} \right) C_j .$$

So we summarize the properties of $\mathscr{H}_f$ needed for our purpose.

Proposition 1: If $f: T^*N \to \mathbf{R}$ is a pullback of a function on T^*Q, then for all $\mu \in \mathfrak{g}$ the Hamiltonian vector field $\mathscr{H}_f$ is a vector field on the subspace $J^{-1}(\mu)$, and as the total space of a principal bundle over T^*Q with connection α, $\mathscr{H}_f$ is horizontal. If differs from the horizontal lift of the standard Hamiltonian vector field on T^*Q by a field in the vertical direction. With respect to the local coordinates chosen, we have explicitly

$$\mathscr{H}_f = \left[-\frac{\partial f}{\partial q_a} \frac{\partial}{\partial p_a} + M_{ia} \frac{\partial f}{\partial p_a} \frac{\partial}{\partial n_i} \right] + \mu_s M_{ia} M_{jb} \left(\frac{\partial A_{sj}}{\partial n_i} - \frac{\partial A_{si}}{\partial n_j} \right) \frac{\partial f}{\partial p_b} \frac{\partial}{\partial p_a}. \tag{9}$$

Furthermore, we have

$$\langle \chi^* \xi_i \, dn_i, \mathscr{H}_f \rangle = p_a \frac{\partial f}{\partial p_a}. \tag{10}$$

IV. POLARIZATION

We first state some well-known results concerning the quantization of integral coadjoint orbits $\mathscr{O}_\mu$. Let $\mathfrak{h}$ be the Lie algebra of H, we say that $\mathscr{O}_\mu$ is integral if the map

$$v \in \mathfrak{h} \to 2\pi i \langle v, \mu \rangle \tag{11}$$

is the derivative of a global character, i.e., there is a group homomorpism $\rho_\mu : H \to U(1)$ such that $\rho_{\mu *}$ is the map given in (11). A version of the Borel–Weil theorem, due to Kirillov[5] and Kostant[6] asserts that there is a one-to-one correspondence between the integral orbits of G and its unitary irreducible representations, and these representations can be constructed by the method of geometric quantization applied to the coadjoint orbit $\mathscr{O}_\mu$, which we will briefly explain.

Let $\mathscr{L}$ be the prequantization line bundle $G \times_{\rho_\mu} \mathbf{C}$ over $\mathscr{O}_\mu \simeq G/H$ with the connection induced by the map ρ_μ. It is known that $\mathscr{O}_\mu$ is a Kähler manifold with complex coordinates with respect to which the G action is holomorphic. There is a standard G–equivariant polarization quantizing with respect to the line bundle and this polarization gives $\mathscr{H}_\mu$ whose elements are holomorphic functions on $\mathscr{O}_\mu$. The polarization, known as the positive Kähler polarization, is given by left translation of a set of $v_k \in \mathfrak{g} \otimes \mathbf{C}$, the complexification of $\mathfrak{g}$, so that the polarization is generated by

$$V_k(\mathrm{Ad}_g \, \mu) = g_* v_k \in T_{\mathrm{Ad}_g \, \mu} \mathscr{O}_\mu . \tag{12}$$

As a polarization, v_k thought of as vectors in $T_\mu \mathscr{O}_\mu$ satisfies

$$\Omega_{\mathscr{O}_\mu}(v_h, v_k) = 0, \tag{13}$$

where $\Omega_{\mathcal{O}_\mu}$ is the canonical symplectic form on $\mathcal{O}_\mu$. The specifics of the choices of v_k will not be important in what follows. It is worth mentioning that V_k is contained in the vertical polarization on T^*N.

Consider the complex distribution on $N^\# \times \mathcal{O}_\mu$ generated by $\{\partial/\partial p_a, V_k\}$; it is G equivariant and thus projects onto $N^\# \times_G \mathcal{O}_\mu \simeq N^\#/H \times \{\mu\}$. One checks that the image is a polarization, which we denote by $\mathcal{P}$.[8]

It is easy to represent $\mathcal{P}$ in local coordinates on $N^\#$; $\partial/\partial p_a$ are vector fields on $N^\#$, and V_k corresponds to $\hat{n}_* v_k$ as complex vector fields on N, $\hat{n}_*$ as in (1). This is so since the assignment $(p,n,\mathrm{Ad}_g\mu) \mapsto (p,ng,\mu)$ defines the projection $N^\# \times \mathcal{O}_\mu \to N^\# \times_G \mathcal{O}_\mu \simeq N^\#/H \times \{\mu\}$. Thus, the vector field generated by G action on $\mathcal{O}_\mu$ translates to the vector field generated by G action on N.

The Hilbert space structure on $L^2(N)$ is given by integration with respect to the measure $\mathrm{d}n = \mathrm{d}\sigma \sqrt{\det(g)}\mathrm{d}q$, where $\mathrm{d}\sigma$ is a Haar measure on G that we transfer to a measure on the fiber in the projection $N \to Q$ and g is the metric on Q. Using the half-form bundle formalism the wave functions are of the form $\psi(n)\sqrt{\mathrm{d}n}$. It is clear that the of the Haar measure will play no role in our consideration as the polarization and all Hamiltonians in question are G invariant. To keep the half-form bundle formalism to a minimum, we may identify the wave functions as $\psi(n)\det g^{1/4}$. We will determine explicitly the differential operators corresponding to f, so that

$$\psi(n)\det g^{1/4} \mapsto [\mathcal{Q}(f)\psi(n)]\det g^{1/4}. \tag{14}$$

In quantizing f that is linear in the momentum variables, the $\det g^{1/4}$ term will give rise to the covariant divergence, and for $f = \frac{1}{2}\|p\|^2 + V(q)$, it results in the Ricci curvature. The appearance of the Ricci curvature is also reported in Ref. 4.

V. POLARIZATION PRESERVING CASE

If $f: T^*Q \to \mathbf{R}$ is linear in p, $f = K_a(q)p_a$, then one easily checks, using (9), that $\exp t\mathcal{H}_{f*}\mathcal{P} = \mathcal{P}$. In fact, we have the following.

Proposition 2:

$$\left[\mathcal{H}_f, \frac{\partial}{\partial p_a}\right] = \frac{\partial K_a}{\partial q_b}\frac{\partial}{\partial p_b}, \tag{15}$$

$$[\mathcal{H}_f, V_k] \in \mathrm{span}\left\{\frac{\partial}{\partial p_a}\right\}, \tag{16}$$

where the brackets refer to the Lie algebra bracket on vector fields.

Proof: Equation (15) is by direct computation. We have

$$\mathcal{H}_f = \left(E_b - p_c\frac{\partial K_c}{\partial q_b}\right)\frac{\partial}{\partial p_b} + K_b M_{ib}\frac{\partial}{\partial n_i}, \tag{17}$$

where E_b is independent of p_a, and (15) results.

To show (16), we first realize from Proposition 1 that $\mathcal{H}_f = W_1 + W_2$, where W_1 is the horizontal lift of a vector field on T^*Q, thus $[W_1, V_k] = 0$ as V_k is generated by the group action on N. W_2 is of the form $F_a(\partial/\partial p_a)$. Since V_k is independent of p, $[W_2, V_k] = V_k(F_a)(\partial/\partial p_a)$, where $V_k(F_a)$ refers to applying the vector field as a differential operator to the coefficient function F_a. □

The importance of (16) is that $[\mathcal{H}_f, V_k]$ is a combination of vectors fields in $\mathcal{P}$ that does not involve the V_h vector fields. According to (7.12) of Ref. 3, the quantization of f is then given by

$$\mathscr{Q}(f)\psi = -\frac{i\hbar}{\det g^{1/4}}\left[\mathscr{H}_f(\psi(n)\det g^{1/4}) + \frac{1}{2}\sum_{a=1}^{\dim Q}\frac{\partial K_a}{\partial q_a}(\psi(n)\det g^{1/4})\right]$$

$$= -i\hbar\left[\mathscr{H}_f(\psi) + \frac{1}{2}\left(\frac{1}{\sqrt{\det g}}\mathscr{H}_f\left(\sqrt{\det g}\right) + \sum_{a=1}^{\dim Q}\frac{\partial K_a}{\partial q_a}\right)\psi\right]. \tag{18}$$

The Hamiltonian vector field $\mathscr{H}_f$ projects to a vector field $V^{\#}$ on N, which is the horizontal lift to the projection of $\mathscr{H}_f$ onto Q, where $V = K_b(\partial/\partial q_b)$. Note that $\mathscr{H}_f(\sqrt{\det g}) = V(\sqrt{\det g})$. The divergence of the vector field V on Q is defined[12] through the relation

$$\mathrm{d}*V = \mathrm{div}\ V\sqrt{\det g}\,\mathrm{d}q.$$

The covariant divergence on N is defined as the divergence of the horizontal lift $V^{\#}$. We have

$$\mathrm{div}\ V = \frac{1}{\sqrt{\det g}}V\left(\sqrt{\det g}\right) + \sum_{a=1}^{\dim Q}\frac{\partial K_a}{\partial q_a}\ .$$

.

Using (17), (18), and the fact that ψ is independent of p, we have

Proposition 3: $\mathscr{Q}(f)\psi = -i\hbar(K_a\mathscr{D}_a + \frac{1}{2}\mathrm{div}\ V)\psi$, where $\mathscr{D}$ is the covariant derivative with respect to the connection α.

Since $\det g$ is a function of n through q, the divergence and the covariant divergence are the same.

VI. BKS PAIRING CASE

Let $\mathscr{P}$ and $\mathscr{P}'$ be transversal polarizations on T^*N, and denote their associated quantum spaces by $\mathscr{Q}$ and $\mathscr{Q}'$. The BKS pairing gives rise to a map $B:\mathscr{Q}' \to \mathscr{Q}$ such that

$$\langle B(\psi), \phi\rangle = \int_{T^*N}\psi\overline{\phi}(\det\omega)^{1/2}\,\mathrm{d}l,$$

where $\mathrm{d}l$ is the Liouville form $\mathrm{d}\xi_1\cdots\mathrm{d}\xi_n\,\mathrm{d}n_1\cdots\mathrm{d}n_n$. Since the volume form on N is $\mathrm{d}n = \sqrt{\det g}\mathrm{d}n_1\cdots\mathrm{d}n_n$, we have

$$B\psi(n) = \frac{1}{\sqrt{\det g}}\int_{T^*_nN}\psi(\det\omega)^{1/2}\,\mathrm{d}\xi_1\cdots\mathrm{d}\xi_n. \tag{19}$$

Let $f = \frac{1}{2}g^{ab}p_ap_b + V(q)$, where g^{ab} is the inverse of the metric g_{ab}. From (9), we have

$$\mathscr{H}_f = -\left(\frac{\partial V}{\partial q_a} + \frac{1}{2}\frac{\partial g^{bc}}{\partial q_a}p_bp_a\right)\frac{\partial}{\partial p_a} + M_{ia}g^{ab}p_b\frac{\partial}{\partial n_i} + \mu_sM_{ia}M_{jb}\left(\frac{\partial A_{sj}}{\partial n_i} - \frac{\partial A_{si}}{\partial n_j}\right)g^{bc}p_c\frac{\partial}{\partial p_a},$$

and note the linear dependence of the coefficients on the p variables. We denote by $\mathscr{P}_t = \exp t\mathscr{H}_{f*}\mathscr{P}$, here the two polarizations $\mathscr{P}$ and $\mathscr{P}_t$ do not intersect transversely. We claim the following.

Proposition 4: Vector fields generated by the group action are in $\mathscr{P}\cap\mathscr{P}_t$.

Proof: It suffices to show that

$$\Omega\left(\exp t\mathscr{H}_{f*}V_k, \frac{\partial}{\partial p_a}\right) = 0 \tag{20}$$

and

$$\Omega(\exp t\mathscr{H}_{f*}V_k, V_h) = 0, \tag{21}$$

with Ω as in (7). Let $\overline{p(}p,n,t)$ and $\overline{n(}p,n,t)$ denote the flow generated by $\exp t\mathscr{H}_f$ at $(p,n) \in N^{\#}$ with μ fixed, $(\overline{p,n,}\mu)=\exp t\mathscr{H}_f(p,n,\mu)$. Then $\exp t\mathscr{H}_{f*}V_k=V_k(\overline{p_a})(\partial/\partial p_a)+V_k(\overline{n_i}) \times(\partial/\partial n_i)$. So $\Omega(\exp t\mathscr{H}_{f*}V_k,\partial/\partial p_a)=(\partial q_a/\partial n_i)V_k(\overline{n_i})$. Recall that $V_k=\hat{n}_* v_k$ is vertical and $\partial q_a/\partial n_i$ is the Jacobian of the projection $\pi:N\to Q$. Then $\pi_*\hat{n}_*=0$ implies that (20) holds.

Equation (21) follows from general principle; since f is G-invariant (with respect to the noncanonical G action), V_h is equivariant with respect to the flow: $\exp t\mathscr{H}_{f*}V_h(p,n) =V_h(\exp t\mathscr{H}_f(p,n))$. Since the flow of a Hamiltonian vector field preserves the symplectic form, we have

$$\Omega(\exp t\mathscr{H}_{f*}V_k(p,n),V_h(\exp t\mathscr{H}_f(p,n)))=\Omega(V_k,V_h)=\Omega_{\mathscr{O}_\mu}(v_h,v_k)=0,$$

where v_h and v_k belong to a polarization on $\mathscr{O}_\mu$ to begin with (13). □

This being the case, quantization of f via BKS pairing involves integrating only over the p variables, i.e., the fiber coordinates of the projection $\Pi:N^{\#}\to N$. According to (7.20) of Ref. 3, together with the similarity transform (14) adjustment and the adjustment in the BKS pairing described in (19),

$$\mathscr{Q}(f)\psi(n)=\frac{1}{\det g^{1/2}}\frac{1}{\det g^{1/4}}\, i\hbar\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0}\Psi_t(n), \tag{22}$$

where

$$\Psi_t(n)=(i\hbar)^{-\dim Q/2}\int_{\Pi^{-1}(n)}[\det\omega_{ab}]^{1/2}\exp(i\hbar^{-1}L)\Psi(p,n,t)\mathrm{d}p, \tag{23}$$

$$\Psi(p,n,t)=\psi(\overline{n(}p,n,t))\times[\det g(\overline{n(}p,n,t))]^{1/4}, \tag{24}$$

$$\omega_{ab}=\Omega\left(\frac{\partial}{\partial p_a},\exp t\mathscr{H}_{f*}\frac{\partial}{\partial p_b}\right), \tag{25}$$

$$L=t\left(\frac{1}{2}\|p\|^2+V(q)\right)-2\int_0^t V(\overline{n(}p,n,s))\mathrm{d}s. \tag{26}$$

The manipulation follows closely that of Sniatycki.[3] Making the substitution $x_a=tp_a$, we have results analogous to (7.26) and (7.27) of Ref. 3.

Proposition 5:

$$\lim_{t\to 0+} t^{-\dim Q/2}\exp\left(\frac{i}{\hbar}\frac{\|x\|^2}{2t}\right)=(2\pi\hbar)^{\dim Q/2}e^{\pi i\,\mathrm{sgn}(g)/4}\sqrt{\det g}\,\delta(x), \tag{27}$$

$$\frac{\partial}{\partial t}t^{-\dim Q/2}\exp\left(\frac{i}{\hbar}\frac{\|x\|^2}{2t}\right)=\frac{i\hbar}{2}g_{ab}\frac{\partial^2}{\partial x_a\,\partial x_b}t^{-\dim Q/2}\exp\left(\frac{i}{\hbar}\frac{\|x\|^2}{2t}\right). \tag{28}$$

Proof: The first equation follows from the method of the stationary phase (cf. Ref. 13), which for n-dimensional space reads as

$$\int_{\mathbf{R}^n}\alpha(y)e^{ik\phi(y)}\,\mathrm{d}y=\left(\frac{2\pi}{k}\right)^{n/2}\sum_{y|\mathrm{d}\phi(y)=0}e^{\pi i\,\mathrm{sgn}\,H(y)/4}\frac{e^{ik\phi(y)}\alpha(y)}{\sqrt{|\det H(y)|}}+O(k^{-n/2-1}). \tag{29}$$

The $1/t$ factor in (27) plays the role of the large parameter k. H is the Hessian of ϕ, which in our case is $g^{\mu\nu}$; thus $\det H=\det g^{-1}$ and sgn is the signature of the metric. The only stationary point in (27) is $x=0$, thus the right-hand side of (27) has a (dim Q-dimensional) delta function at $x=0$.

The second equation is a straightforward computation. We need the fact that $g_{\alpha\mu}g^{\mu b}=\delta_{ab}$, and $\Sigma_a\Sigma_b g_{ab}g^{ab}=\Sigma_a\delta_{aa}=\dim Q$ in the course of the computation. □

One checks that ω_{ab} in (25) is

$$\frac{\partial q_a}{\partial n_i}\,\frac{\partial \overline{n_i}}{\partial p_b} = t g^{ab} + \text{higher-order terms in } t,$$

using (30) below. Then $[\det \omega_{ab}]^{1/2}\, \mathrm{d}p \sim t^{-\dim Q/2}\, \mathrm{d}x$, providing us with the needed factor to apply the results of Proposition 5. Thus, in determining $\mathrm{d}/\mathrm{d}t\big|_{t=0}\Psi_t(n)$, we need only to consider terms involving t, tp_a, and $t^2 p_a p_{\underline{b}}$ while ignoring terms of the form $t^2 p_a$ and all higher-order terms. Using (9), the expansion of $\overline{n}$, expressed in the t and x_a variables up to the relevant terms, are

$$\widetilde{n}_i(n,x) = n_i + M_{ia} g^{\mu a} x_\mu + \frac{1}{2}\left[M_{jb}\left(\frac{\partial M_{ia}}{\partial n_j}\, g^{\nu b} g^{\mu a} + M_{ia}\, \frac{\partial g^{\mu a}}{\partial q_b}\, g^{\nu b} \right) - \frac{1}{2}\, M_{ia}\, \frac{\partial g^{\mu\nu}}{\partial q_b}\, g^{ba} \right] x_\mu x_\nu, \tag{30}$$

which is independent of t. And $\Psi_t(n)$ in (23) is reduced to

$$\Psi_t(n) = (i\hbar)^{-\dim Q/2} \int_{\Pi^{-1}(n)} t^{-\dim Q/2}\, \exp\left(\frac{i}{\hbar}\, \frac{\|x\|^2}{2t} \right) \Phi(n,x,t)\, \mathrm{d}x,$$

where

$$\Phi(n,x,t) = \exp(-i\hbar^{-1} t V(q)) \times \psi(\overline{n}(n,x)) \det g(\overline{n}(n,x))^{1/4}. \tag{31}$$

By applying Proposition 5, integration by parts yields

$$\mathscr{Q}(f)\psi(n) = \frac{(-2\pi i)^{\dim Q/2} e^{\pi i\, \mathrm{sgn}/4}}{\det g^{1/4}} \left[-\frac{\hbar^2}{2}\, g_{ab}\, \frac{\partial^2 \Phi}{\partial x_a\, \partial x_b}\bigg|_{x,t=0} + i\hbar\, \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{x=t=0} \Phi \right]. \tag{32}$$

The $\sqrt{\det H}$ term that appears in the method of stationary phase formula results in a $g^{1/2}$ factor (27) that cancels with the $\det g^{1/2}$ on the right-hand side of (22).

Since n is fixed, we can choose a normal coordinate system[12] around $q = \pi(n)$ so that $\partial g^{\mu\nu}/\partial q_a = 0$ for all μ, ν, and q_a when evaluated at n (i.e., at $x = t = 0$). A direct computation shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{x=t=0} \Phi(n,x,t) = -i\hbar^{-1} V(q)\psi(n) \det g^{1/4}, \tag{33}$$

$$g_{ab}\, \frac{\partial^2}{\partial x_a\, \partial x_b}\bigg|_{x=t=0} \Phi(n,x,t) = \det g^{1/4} \bigg[g^{\mu\nu} M_{j\nu}\, \frac{\partial M_{i\mu}}{\partial n_j}\, \frac{\partial \psi}{\partial n_i} + g^{\mu\nu} M_{i\mu} M_{j\nu}\, \frac{\partial^2 \psi}{\partial n_i\, \partial n_j} + \frac{1}{4}\, g^{\mu\nu} g^{ab}\, \frac{\partial^2 g_{\mu\nu}}{\partial q_a\, \partial q_b} \bigg]. \tag{34}$$

Here we have made repeated use of the identity (Ref. 12 p. 302)

$$\frac{1}{\det g}\, \frac{\partial \det g}{\partial q_b} = \frac{1}{\det g}\, \frac{\partial \det g}{\partial g_{\mu\nu}}\, \frac{\partial g_{\mu\nu}}{\partial q_b} = g^{\mu\nu}\, \frac{\partial g_{\mu\nu}}{\partial q_b}.$$

In normal coordinates, the covariant Laplace operator reduces to

$$\Delta_\alpha \psi = \frac{1}{\det g^{1/2}}\, \mathscr{D}_\mu (\det g^{1/2} g^{\mu\nu}\, \mathscr{D}_\nu \psi) = g^{\mu\nu} M_{i\mu} M_{j\nu}\, \frac{\partial^2 \psi}{\partial n_i\, \partial n_j} + g^{\mu\nu} M_{j\nu}\, \frac{\partial M_{i\mu}}{\partial n_j}, \tag{35}$$

and the Ricci curvature becomes

$$R = g^{ik}(\partial_k \Gamma^j_{ji} - \partial_j \Gamma^j_{ki} + \Gamma^j_{km}\Gamma^m_{ji} - \Gamma^j_{jm}\Gamma^m_{ki}) = (g^{ik} g^{\mu\nu} - g^{i\nu} g^{k\mu})\partial_i\, \partial_k g_{\mu\nu} = \tfrac{3}{2} g^{ik} g^{\mu\nu} \partial_i\, \partial_k g_{\mu\nu}. \tag{36}$$

Here Γ^k_{ij} are the Christoffel symbols, and the identity

$$\Gamma^m_{ij,k}+\Gamma^m_{jk,i}+\Gamma^m_{ki,j}=0$$

is used to show $g^{i\nu}g^{k\mu}\partial_i\,\partial_k g_{\mu\nu}=-\frac{1}{2}g^{ik}g^{\mu\nu}\partial_i\,\partial_k g_{\mu\nu}$. We must caution the readers that these expressions only hold in normal coordinates. However, by combining (32)–(36) we can express our final result in a coordinate invariant form.

Proposition 6. Quantization of $f=\frac{1}{2}\|p\|2+V(q)$ gives

$$\mathscr{Q}(f)\psi(n)=(-2\pi i)^{\dim Q/2}e^{\pi i\,\mathrm{sgn}(g)/4}\left(-\frac{\hbar^2}{2}\left[\Delta_\alpha+\frac{1}{6}R\right]+V(q)\right)\psi(n).$$

We conclude with a final remark. The Yang–Mills field is defined[14] as the curvature $\mathscr{D}\alpha$ of the Yang–Mills potential α, whereas the contribution of this connection in the local expression of the symplectic form is $d\alpha$. They are related by

$$\mathscr{D}\alpha(v,w)=\mathrm{d}\,\alpha(\mathrm{hor}\,v,\mathrm{hor}\,w),$$

where hor denotes the horizontal projection. Since the vector fields of concern are all horizontal, the effect of $\mathrm{d}\alpha$ is equivalent to the curvature.

ACKNOWLEDGMENT

We wish to thank N. P. Landsman for helpful comments on the preliminary version of this work.

[1] J. Marsden and A. Weinstein, Rep. Math. Phys. **5**, 121 (1974).
[2] A. Weinstein, Lett. Math. Phys. **2**, 417 (1978).
[3] J. Śniatycki, *Geometric Quantization and Quantum Mechanics* (Springer-Verlag, New York, 1980).
[4] N. M. J. Woodhouse, *Geometric Quantization*, 2nd ed. (Clarendon, Oxford, 1992).
[5] A. A. Kirillov, *Elements of the Theory of Representation* (Springer-Verlag, Berlin, 1976).
[6] B. Kostant, ''Quantization and unitary representation,'' in *Modern Analysis and Applications*, Lecture Notes in Mathematics, 1970, Vol. 170, pp. 87–207.
[7] V. W. Guillemin and S. Sternberg, Invent. Math. **67**, 515 (1982).
[8] M. A. Robson, J. Geom. Phys. **19**, 207 (1996).
[9] R. J. Blattner, Proc. Symp. Pure Math. **26**, 87 (1973).
[10] N. P. Landsman, J. Geom. Phys. **12**, 93 (1993).
[11] M. Kummer, Indiana Univ. Math. J. **30**, 281 (1981).
[12] Y. Choquet-Bruhat, C. de Witt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1977).
[13] V. W. Guillemin and S. Sternberg, *Geometric Asymptotics* (American Mathematical Society, Providence, 1977).
[14] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Hamiltonian systems on *k*-cosymplectic manifolds

Manuel de León
Instituto de Matemáticas y Física Fundamental, Consejo Superior de Investigaciones Científicas, Serrano 123, 28006 Madrid, Spain

Eugenio Merino
Escola Politécnica Superior de Enxeñería, Universidade da Coruña, Mendizábal s.n., 15403 Ferrol, A Coruña, Spain

José A. Oubiña
Departamento de Xeometría e Topoloxía, Facultade de Matemáticas, Universidade de Santiago de Compostela, Campus Sur, Santiago, Spain

Paulo R. Rodrigues
Departamento de Geometria, Instituto de Matemática, Universidade Federal Fluminense, 24020-005-Niteroi, Brazil

Modesto R. Salgado
Departamento de Xeometría e Topoloxía, Facultade de Matemáticas, Universidade de Santiago de Compostela, Campus Sur, Santiago, Spain

(Received 7 April 1997; accepted for publication 12 September 1997)

The Hamiltonian framework on symplectic and cosymplectic manifolds is extended in order to consider classical field theories. To do this, the notion of *k*-cosymplectic manifold is introduced, and a suitable Hamiltonian formalism is developed so that the field equations for scalar and vector Hamiltonian functions are derived. © *1998 American Institute of Physics.* [S0022-2488(98)01102-5]

I. INTRODUCTION

There are many attempts in recent years to develop a Hamiltonian formalism for classical field theories. In the 1960s the so-called multisymplectic formalism was developed by the Tulczyjew school in Warsaw (see Refs. 1–4) and independently by García and Pérez-Rendón[5,6] and Goldschmidt and Sternberg.[7] This approach was revisited by Martin,[8,9] Gotay,[10,11] and more recently by Cantrijn *et al.*,[12–14] de León *et al.*,[15,16] Sardanashvily,[17,18] Kanatchikov,[19] and Echevarría *et al.*[20] (also see Refs. 21 and 22).

At the same time, other approaches were suggested, with different names but defining the same geometrical object. Thus, Gunther introduced the so-called polysymplectic structures,[23] Awane[24,25] the *k*-symplectic manifolds, also discussed by Puta[26] and Norris (see Ref. 27 and the references therein), and de León *et al.*[28,29] studied the so-called *p*-almost cotangent structures. Along this paper, we shall adopt the name of a (almost) *k*-symplectic structure. The original idea in all these papers was to abstract the geometrical ingredients of the *k*-cotangent bundle $(T^1_k)^*Q$ of a manifold Q, say the manifold of 1-jets with target $0 \in \mathbf{R}^k$ of all the mappings from Q into $\mathbf{R}^k$. Since $(T^1_k)^*Q$ may be canonically identified with the Whitney sum of k copies of T^*Q, we can transport the canonical symplectic structure on it to $(T^1_k)^*Q$. It should be noticed that the vertical distribution has to be incorporated in the game. This structure can be also derived by using the solder form of the linear frame bundle of Q (see Ref. 27). We also remark that in Ref. 29 it was presented an integrability theorem that provides Darboux coordinates as in the symplectic case.

Given a *k*-symplectic manifold, we obtain a Hamiltonian system by introducing a Hamiltonian function. Thus, the corresponding field equations can be derived[23] as the integral sections of a *k*-vector field. Moreover, if we consider vector-valued Hamiltonian functions, we obtain a Hamiltonian vector field, and in this case, the Hamiltonian has to be of polynomial form.[24,25]

The purpose of this paper is to extend the study to the case of field equations involving the independent parameters in an explicit way; in other words, the Hamiltonian function is of the form $H(t^i,x^\alpha,x^i_\alpha)$, where $1 \leqslant i \leqslant k$, $1 \leqslant \alpha \leqslant n$. To define a convenient geometrical structure, we extract the geometric ingredients of the so-called stable cotangent bundle of k^1-covelocities $\mathbf{R}^k$

 © 1998 American Institute of Physics

$\times(T^1_k)^*Q$ of a manifold Q. Thus, a family $(\eta_i,\omega_i,V;\ 1\leqslant i\leqslant k)$, where each η_i is a 1-form, each ω_i is a 2-form, and V is an nk-dimensional distribution on a $k(n+1)+n$-dimensional manifold M satisfying some compatibility relations, will be called an almost k-cosymplectic structure on M, and M an almost k-cosymplectic manifold. After identifying such a structure as a kind of G-structure on M, an integrability theorem is proved and Darboux coordinates are introduced on a k-cosymplectic manifold (Sec. II). This result shows that a k-cosymplectic manifold is locally isomorphic to a stable cotangent bundle of k^1-covelocities; in Sec. III we prove that, under some global conditions, there is also a global isomorphism. In Sec. IV, we introduce the dynamics; in fact, the geometry of a k-cosymplectic manifold allows us to write the field equations in a global way. An example is discussed. The case of vector-valued Hamiltonian functions is studied in Sec. V. Finally, we compare our geometrical approach with the so-called multisymplectic formalism (Sec. VI). In a forthcoming paper we shall examine the corresponding Euler–Lagrange formalism.

II. ALMOST *k*-COSYMPLECTIC STRUCTURES

Let $\pi:\mathbf{R}^k\times Q\to Q$ be a trivial fibered manifold and denote by $J^1\pi$ the manifold of 1-jets of local sections of π, $J^1\pi$ will be called the stable cotangent bundle of k^1-covelocities of the manifold Q. $J^1\pi\to Q$ is a vector bundle over Q with standard fiber $\mathbf{R}^k\times\mathbf{R}^{nk}$, where $\dim Q=n$. We have a canonical identification $J^1\pi\cong\mathbf{R}^k\times(T^1_k)^*Q$ in such a way that we can introduce local coordinates on $\mathbf{R}^k\times(T^1_k)^*Q$ as follows. If (x^α) are local coordinates on Q, then $(t^i,x^\alpha,x^i_\alpha)$, $1\leqslant i\leqslant k$, $1\leqslant\alpha\leqslant n$, is a local coordinate system on $\mathbf{R}^k\times(T^1_k)^*Q$. The reason for the above notation is the following. For $k=1$ we have $J^1\pi\cong\mathbf{R}\times T^*Q$, and (t,q^α,p_α) stands for the canonical coordinates on it.

Define a family $((\eta_0)_i,(\omega_0)_i,V_0)$ of 1-forms $(\eta_0)_i$, 2-forms $(\omega_0)_i$ and a distribution V_0 given by

$$(\eta_0)_i=dt^i,\quad (\omega_0)_i=dx^\alpha\wedge dx^i_\alpha,\quad V_0=\left\langle\frac{\partial}{\partial x^1_\alpha},\ldots,\frac{\partial}{\partial x^k_\alpha}\right\rangle.$$

A simple inspection in local coordinates shows that the following relations hold:

(i) $(\eta_0)_1\wedge\cdots\wedge(\eta_0)_k\neq 0$,
(ii) $\dim(\ker(\omega_0)_1\cap\cdots\cap\ker(\omega_0)_k)=k$,
(iii) $\ker(\eta_0)_1\cap\cdots\cap\ker(\eta_0)_k\cap\ker(\omega_0)_1\cap\cdots\cap\ker(\omega_0)_k=\{0\}$,
(iv) $(\eta_0)_{i|V0}=0$, $(\omega_0)_{i|V0\times V0}=0$ $(1\leqslant i\leqslant k)$.

Inspired in the above geometrical model, we introduce the following definition.

Definition II.1: Let M be a differentiable manifold of dimension $k(n+1)+n$. A family $(\eta_i,\omega_i,V;\ 1\leqslant i\leqslant k)$ where each η_i is a 1-form, each ω_i is a 2-form, and V is an nk-dimensional distribution on M, such that

(i) $\eta_1\wedge\cdots\wedge\eta_k\neq 0$,
(ii) $\dim(\ker\omega_1\cap\cdots\cap\ker\omega_k)=k$,
(iii) $\ker\eta_1\cap\cdots\cap\ker\eta_k\cap\ker\omega_1\cap\cdots\cap\ker\omega_k=\{0\}$,
(iv) $\eta_{i|V}=0$, $\omega_{i|V\times V}=0$ $(1\leqslant i\leqslant k)$,

will be called an *almost k-cosymplectic structure*, and the manifold M an *almost k-cosymplectic manifold.*

In particular, if $k=1$, then $\dim M=2n+1$, and (η,ω) is an almost cosymplectic structure on M,[30,31] and (η,ω,V) an almost stable cotangent structure on M.[32] From the conditions of Definition II.1 we deduce that there exist k vector fields $\xi_1,\ldots,\xi_k$ on M satisfying

$$\iota_{\xi_i}\eta_j=\delta_{ij},\quad \iota_{\xi_i}\omega_j=0, \tag{1}$$

with $1\leqslant i,\ j\leqslant k$, and that will be called the Reeb vector fields associated to the almost k-cosymplectic structure. Our intention is to get integrability conditions that provide Darboux coordinates for such a kind of geometrical structure. To go further in this direction, we first shall

describe an almost k-cosymplectic structure as a type of G-structure. Let us recall that a G-structure on an m-dimensional manifold M is a G-reduction of the linear frame bundle of M.[33] Here G a Lie subgroup of $Gl(m,\mathbf{R})$.

Define k distributions $V_1,\ldots,V_k$ on M by

$$V_i=\cap_{j\neq i}\ \ker\ \omega_j\cap\ker\ \eta_1\cap\cdots\cap\ker\ \eta_k.$$

Then

$$V_i\subset V,\quad \dim\ V_i=n,\quad V_i\cap(+_{j\neq i}V_j)=\{0\},$$

for all $1\leqslant i\leqslant k$, and therefore we have

$$V=V_1\oplus\cdots\oplus V_k.$$

Since (1) we can choose a linear frame at a point $x\in M$ of the form $\{\xi_1,\ldots,\xi_k,Y_{k+1},\ldots,Y_{k(n+1)+n}\}$, such that

$$\langle Y_{k+1},\ldots,Y_{k(n+1)+n}\rangle=\ker\ \eta_1\cap\cdots\cap\ker\ \eta_k,$$

$$\langle Y_{k+in+1},\ldots,Y_{k+in+n}\rangle=V_i,$$

for all $1\leqslant i\leqslant k$. Moreover, we have $\langle\xi_1,\ldots,\xi_k\rangle=\ker\ \omega_1\cap\cdots\cap\ker\ \omega_k$. With respect to this linear frame, each ω_i is written as

$$\omega_i=\begin{pmatrix}0&0&\cdots&0&\cdots&0\\0&P_i&\cdots&-Q_i^t&\cdots&0\\\vdots&\vdots&&\vdots&&\vdots\\0&Q_i&\cdots&0&\cdots&0\\\vdots&\vdots&&\vdots&&\vdots\\0&0&\cdots&0&\cdots&0\end{pmatrix},$$

with $P_i=-P_i^t$ and $\det Q_i\neq 0$. We construct a new frame $\{\xi_1,\ldots,\xi_k,X_{k+1},\ldots,X_{k(n+1)+n}\}$ at x given by the formula

$$\begin{pmatrix}\xi_1\\\vdots\\\xi_k\\X_{k+1}\\\vdots\\X_{k(n+1)+n}\end{pmatrix}=(\xi_1,\ldots,\xi_k,Y_{k+1},\ldots,Y_{k(n+1)+n})\begin{pmatrix}I_k&0&0&\cdots&0\\0&A&0&\cdots&0\\0&B_1&C_1&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&B_k&0&\cdots&C_k\end{pmatrix},$$

where

$$A=Q_k^{-1},\quad C_k=I_n,\quad B_k=\tfrac{1}{2}\,(Q_k^{-1})^tP_kQ_k^{-1},$$

$$C_i=(Q_i^{-1})^tQ_k^t,\quad B_i=\tfrac{1}{2}\,(Q_i^{-1})^tP_iQ_k^{-1},$$

for all $1\leqslant i\leqslant k$. With respect to this new frame (called adapted to the k-cosymplectic structure) the forms $\omega_1,\ldots,\omega_k$, are expressed by the matrices $(\omega_1)_0,\ldots,(\omega_k)_0$ given by

$$(\omega_1)_0=\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -I_n & \cdots & 0 \\ 0 & I_n & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},\ldots,(\omega_k)_0=\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -I_n \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & I_n & 0 & \cdots & 0 \end{pmatrix}.$$

Each pair of adapted frames is related by a change with associated matrix

$$\begin{pmatrix} I_k & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & B_1 & C & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & B_k & 0 & \cdots & C \end{pmatrix},$$

with $B_i^tA=A^tB_i$ and $C=(A^{-1})^t$ for all $1\leqslant i\leqslant k$. The group G of such matrices is a Lie subgroup of $Gl(k(n+1)+n,\mathbf{R})$. Now we obtain a G-structure on M by collecting all the adapted linear frames at all the points of M. Conversely, let us suppose a given G-structure $\mathbf{B}$ on M. Then we may define an almost k-cosymplectic structure $(\eta_i,\omega_i,V;1\leqslant i\leqslant k)$ on M, as follows. If $\{e_1,\ldots,e_{k(n+1)+n}\}$ is the canonical basis of $\mathbf{R}^{k(n+1)+n}$ and $\{e^1,\ldots,e^{k(n+1)+n}\}$ is the corresponding dual basis, we set $(\eta_i)_x(X)=e^i(u^{-1}(X))$, $(\omega_i)_x(X,Y)=(u^{-1}(X)(\omega_i)_0(u^{-1}(Y))^t$, and

$$V_x=u(\langle e_{k+n+1},\ldots,e_{k(n+1)+n}\rangle),$$

where $x\in M$, $X,Y\in T_xM$, and $u\in B$ is an adapted linear frame,

$$u:\mathbf{R}^{k(n+1)+n}\to T_xM.$$

Summing up, we have proved the following.

Proposition II.2: A manifold M of dimension $k(n+1)+n$ admits an almost k-cosymplectic structure if and only if the structure group of its tangent bundle is reducible to the group G of matrices of the form

$$\begin{pmatrix} I_k & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & B_1 & C & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & B_k & 0 & \cdots & C \end{pmatrix},$$

with $B_i^tA=A^tB_i$ and $C=(A^{-1})^t$ for all $1\leqslant i\leqslant k$.

Definition II.3: Let M be a $k(n+1)+n$-dimensional manifold with an almost k-cosymplectic structure (η_i,ω_i,V). We say that (η_i,ω_i,V) is integrable if the corresponding G-structure is integrable, and in such a case it is called *k-cosymplectic.*

A G structure on a manifold M is said to be integrable[33] if around each point of M there exist local coordinates (z^u) such that the local frame $\{\partial/\partial z^u\}$ is adapted to the G-structure, that is, it takes values in the reduced G-bundle.

Therefore, an almost k-cosymplectic structure (η_i,ω_i,V) on a manifold M is integrable if around each point of M there exist local coordinates $(s^i,x^\alpha,x^i_\alpha;1\leqslant i\leqslant k,1\leqslant\alpha\leqslant n)$ such that

$$\eta_i=ds^i,\quad \omega_i=dx^\alpha\wedge dx^i_\alpha,\quad V=\left\langle\frac{\partial}{\partial x^1_\alpha},\ldots,\frac{\partial}{\partial x^k_\alpha}\right\rangle_{\alpha=1,\ldots,n}. \tag{2}$$

Such coordinate functions will be called *Darboux* or *canonical coordinates*.

The following theorem characterizes the integrability of a k-cosymplectic structure.

Theorem II.4: *An almost k-cosymplectic structure* (η_i,ω_i,V) *on a manifold* M *is integrable if and only if the following conditions are satisfied:*

$$d\eta_i=0, \quad d\omega_i=0, \quad [V,V]\subset V \quad (1\leqslant i\leqslant k). \tag{3}$$

Proof: The integrability of a G structure obviously implies (3). Conversely, let us suppose that (3) holds. Define $K_x=\ker\ \eta_1(x)\cap\cdots\cap\ker\ \eta_k(x)$, for each $x\in M$. We have

$$T_xM=\langle\xi_1\rangle_x\oplus\cdots\oplus\langle\xi_k\rangle_x\oplus K_x,$$

and the distributions

$$\langle\xi_1\rangle,\langle\xi_2\rangle,\ldots,\langle\xi_k\rangle,K_x,$$

are all of them integrable. Moreover, the sum of each pair of these distributions is also integrable and $TM=\langle\xi_1\rangle\oplus\cdots\oplus\langle\xi_k\rangle\oplus K$. From Lemma 1.3 of Ref. 34 we deduce that for every point $x\in M$ there exists a local coordinate system $(\overline{t_i},\overline{x^\alpha},\overline{x^i_\alpha};1\leqslant i\leqslant k,1\leqslant\alpha\leqslant n)$ in a neighborhood U of x, such that

$$\langle\xi_i\rangle=\left\langle\frac{\partial}{\partial t_i}\right\rangle, \quad K=\left\langle\frac{\partial}{\partial\overline{x^\alpha}},\frac{\partial}{\partial\overline{x^i_\alpha}}\right\rangle_{\substack{1\leqslant\alpha\leqslant n\\ 1\leqslant i\leqslant k}}.$$

On $\overline{U}$, for each fixed $1\leqslant i\leqslant k$, we have

$$\xi_i=f_i\,\frac{\partial}{\partial\overline{t_i}},$$

and, since $\eta_j(\xi_i)=\delta_{ij}$, we deduce that

$$\eta_i=\frac{1}{f_i}\,d\overline{t_i}=g_i\,d\overline{t_i}.$$

Since the 1-forms $\eta_1,\ldots,\eta_k$ are closed, we get

$$0=d\eta_i=\frac{\partial g_i}{\partial\overline{t_j}}\,d\overline{t_j}\wedge d\overline{t_i}+\frac{\partial g_i}{\partial\overline{x^\alpha}}\,d\overline{x^\alpha}\wedge d\overline{t_i}+\frac{\partial g_i}{\partial\overline{x^\alpha_j}}\,d\overline{x^\alpha_j}\wedge d\overline{t_i},$$

which implies

$$\frac{\partial g_i}{\partial\overline{x^\alpha}}=0=\frac{\partial g_i}{\partial\overline{x^\alpha_j}},$$

for all $1\leqslant i,\ j\leqslant k,\ 1\leqslant\alpha\leqslant n$, and

$$\frac{\partial g_i}{\partial\overline{t_j}}=0,$$

for all $1\leqslant i,\ j\leqslant k,\ i\neq j$. Then $g_i=g_i(\overline{t_i})$ and we can define k functions $h^i=h^i(\overline{t_i})$ where each h^i is a primitive function of g_i. Define a new coordinate system $(\widetilde{t}_i,\widetilde{x}^\alpha,\widetilde{x}^i_\alpha)$ on a neighborhood $\widetilde{U}$ of x by

$$\widetilde{t}_i=h^i(\overline{t_i}), \quad \widetilde{x}^\alpha=\overline{x^\alpha}, \quad \widetilde{x}^i_\alpha=\overline{x^i_\alpha},$$

for $1\leqslant i\leqslant k$ and $1\leqslant\alpha\leqslant n$. With respect to these coordinates, we write

$$\eta_i = d\widetilde{t}_i, \quad \xi_i = \frac{\partial}{\partial \widetilde{t}_i}, \quad K = \left\langle \frac{\partial}{\partial \widetilde{x}^\alpha}, \frac{\partial}{\partial \widetilde{x}^i_\alpha} \right\rangle_{\substack{1\leqslant\alpha\leqslant n \\ 1\leqslant i\leqslant k}}.$$

Since K is an integrable distribution there exists an integral submanifold W of K through x such that $(W,\omega_1,\ldots,\omega_k,V)$ is a k-symplectic manifold (see Refs. 24, 25, 28, and 29). Then there exists a coordinate neighborhood (U',x^α,x^i_α) with $1\leqslant i\leqslant k$, $1\leqslant\alpha\leqslant n$, in W such that

$$\omega_{i_{|U'}} = dx^\alpha \wedge dx^i_\alpha, \quad V_{|U'} = \left\langle \frac{\partial}{\partial x^1_\alpha}, \ldots, \frac{\partial}{\partial x^k_\alpha} \right\rangle_{1\leqslant\alpha\leqslant n},$$

for all $1\leqslant i\leqslant k$. If the coordinate functions $\widetilde{t}_i$ of $\widetilde{U}$ are defined in $(-\epsilon_i,\epsilon_i)$, for each $1\leqslant i\leqslant k$, we take the following coordinate neighborhood of x in M:

$$(\mathbf{U} = (-\epsilon_1,\epsilon_1)\times\cdots\times(-\epsilon_k,\epsilon_k)\times U', s^i, x^\alpha, x^i_\alpha),$$

with $s^i = \widetilde{t}_i$, and then we have

$$\begin{aligned}\omega_i = {} & (A_i)^j_r\, ds^j\wedge ds^r + (B_i)^j_\alpha\, ds^j\wedge dx^\alpha + (C_i)^j_{\alpha,s} ds^j\wedge dx^r_\alpha \\ & + (D_i)^\alpha_{\beta,j}\, dx^\alpha\wedge dx^j_\beta + (E_i)^\alpha_\beta\, dx^\alpha\wedge dx^\beta + (F_i)^{\alpha,j}_{\beta,s}\, dx^j_\alpha\wedge dx^r_\beta.\end{aligned}$$

Since $\iota_{\xi_j}\omega_i = 0$, we deduce that

$$(A_i)^j_r = (B_i)^j_\alpha = (C_i)^j_{\alpha,s} = 0,$$

for all $1\leqslant i,j,\ s\leqslant k$, $1\leqslant\alpha\leqslant n$, and from $d\omega_i = 0$ we obtain

$$\frac{\partial (E_i)^\alpha_\beta}{\partial s^j} = \frac{\partial (F_i)^{\alpha,j}_{\beta,s}}{\partial s^l} = \frac{\partial (D_i)^\alpha_{\beta,j}}{\partial s^r} = 0.$$

Therefore

$$(E_i)^\alpha_\beta(s^j,x^\sigma,x^j_\sigma) = (E_i)^\alpha_\beta(0,x^\sigma,x^j_\sigma) = 0,$$

$$(F_i)^{\alpha,j}_{\beta,s}(s^l,x^\sigma,x^l_\sigma) = (F_i)^{\alpha,j}_{\beta,s}(0,x^\sigma,x^l_\sigma) = 0,$$

$$(D_i)^\alpha_{\beta,j}(s^r,x^\sigma,x^r_\sigma) = (D_i)^\alpha_{\beta,j}(0,x^\sigma,x^r_\sigma) = \delta_{\alpha\beta}\delta_{ij},$$

from which we get that

$$\omega_i = dx^\alpha\wedge dx^i_\alpha.$$

Moreover, for each $x\in U$, we have

$$V(x)\subset T_xU',$$

and then

$$V = \left\langle \frac{\partial}{\partial x^1_\alpha}, \ldots, \frac{\partial}{\partial x^k_\alpha} \right\rangle_{1\leqslant\alpha\leqslant n},$$

which completes the proof of the theorem. ■

Remark II.5: When $k=1$, the condition $[\xi,V]\subset V$ is a necessary condition for the integrability of the almost stable cotangent structure (η,ω,V), and it is an independent condition of the other ones (see Ref. 32). For $k>1$, we have the following.

Proposition II.6: If M is an almost k-cosymplectic manifold with $k>1$, and $\xi_1,\ldots,\xi_k$ are the Reeb vector fields associated to the almost k-cosymplectic structure, then

$$(d\eta_j=0, d\omega_j=0, \quad \forall 1\leqslant j\leqslant k)\Rightarrow [\xi_i, V]\subset V,$$

for all $1\leqslant i\leqslant k$.

III. *k*-COSYMPLECTIC MANIFOLDS THAT ARE STABLE *k*-COTANGENT BUNDLES

Let $(M, \eta_i, \omega_i, V; 1\leqslant i\leqslant k)$ be a k-cosymplectic manifold. For each $1\leqslant i\leqslant k$, we consider the distributions

$$V_i=\ker\ \eta_1\cap\cdots\cap\ker\ \eta_k\cap_{j\neq i}\ker\ w_j,$$

$$K_i=\ker\ \eta_1\cap\cdots\cap\ker\ \eta_k\cap\ker\ \omega_i=\oplus_{j\neq i}V_j,$$

which are all of them involutive and hence define foliations on M. If $\xi_1,\ldots,\xi_k$ are the Reeb vector fields, then we put

$$W=\langle\xi_1,\ldots,\xi_k\rangle\oplus V,$$

$$W_i=\langle\xi_1,\ldots,\hat{\xi}_i,\ldots,\xi_k\rangle\oplus K_i,$$

where $\hat{}$ over an element means that this element is omitted. Again, W and W_i are integrable distributions of dimension $k(n+1)$ and $(n+1)(k-1)$, respectively.

Definition III.1: The k-cosymplectic structure (η_i, ω_i, V) defines a fibration on M when the following conditions are satisfied.

(i) The leaf space $M_i=M/W_i$ defined by the involutive distribution W_i is a quotient manifold of M, and the canonical projection $p_i: M\to M_i$ is a fibration for each $1\leqslant i\leqslant k$.

(ii) The leaf space $Q=M/W$ defined by the involutive distribution W is a quotient manifold of M and the canonical projection $p: M\to Q$ is a fibration.

In such a case, we have the following commutative diagram:

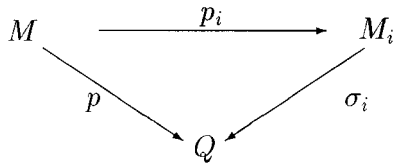

where σ_i is the induced projection. Moreover, we can choose a suitable coordinate system $(s^i, x^\alpha, x^i_\alpha)$, such that

$$p_i(s^1,\ldots,s^k,x^\alpha,x^1_\alpha\ldots,x^k_\alpha)=(s^i,x^\alpha,x^i_\alpha),$$

$$p(s^1,\ldots,s^k,x^\alpha,x^1_\alpha,\ldots,x^k_\alpha)=(x^\alpha),$$

$$\sigma_i(s^i,x^\alpha,x^i_\alpha)=(x^\alpha).$$

On each M_i we define a cosymplectic structure $(\overline{\eta_i}, \overline{\omega_i})$ by putting

$$\overline{\eta_i}(X)=\eta_i(\overline{X}), \quad \overline{\omega_i}(X,Y)=\omega_i(\overline{X,Y}),$$

with $\overline{X,Y}\in TM$, such that $(p_i)_*(\overline{X})=X$ and $(p_i)_*(\overline{Y})=Y$. Moreover, each V_i defines a distribution V_i on M_i such that $(\overline{\eta_i}, \overline{\omega_i}, \overline{V_i})$ is a stable almost cotangent structure on M.[32]

Definition III.2: Let $y\in M$ be such that $p(y)=x$ and $p_i(y)=y_i$ for each $1\leqslant i\leqslant k$. If $\beta\in T^*_xQ$, the i-vertical lift of β to T_yM is the unique vector $\beta^{(i)}\in T_yM$, such that $\beta^{(i)}\in V_i(y)$ and

$$\iota_{T(p_i)\beta^{(i)}}\overline{\eta_i}=0, \quad \iota_{T(p_i)\beta^{(i)}}\overline{w_i}=(\sigma_i)^*\beta.$$

In a similar way to the case of k-symplectic and stable almost cotangent structures (see Ref. 29), we have the following results.

Proposition III.3: Given the 1-forms β, γ on Q, we have

$$[\beta^{(i)},\gamma^{(j)}]=0, \quad [\beta^{(i)},\xi_j]=0,$$

for all $1\leqslant i,\ j\leqslant k$.

Proposition III.4: If ∇ is a symmetric connectic adapted to the k-cosymplectic structure on M that define a fibration, then ∇ induces, by restriction, a flat connection on each leaf of the fibrations $p_i:M\to M_i$.

It should be noticed that such an adapted symmetric connection always exists from the general theory of G structures (see Ref. 33). Now we can prove the main theorem of this section.

Theorem III.5: *Let $(M,\eta_1,\ldots,\eta_k,\omega_1,\ldots,\omega_k,V)$ be a k-cosymplectic manifold that defines a fibration $p:M\to Q$. Let ∇ be a symmetric connection adapted to the structure. If the flat connection induced by ∇ is geodesically complete, each fiber of p is connected and simply connected and each leaf of $\langle\xi_i\rangle\oplus V_i$ is connected for all $1\leqslant i\leqslant k$, then M is an affine bundle modeled in the vector bundle $\mathbf{R}^k\times(T^1_k)^*Q$. Therefore, M admits a structure of vector bundle on Q isomorphic to $\mathbf{R}^k\times(T^1_k)^*Q$.*

Proof: We need only to define a mapping,

$$\rho:M\times_Q(\mathbf{R}^k\times(T^1_k)^*Q)\to M,$$

such that, for each $z\in Q$,

$$\rho_z:p^{-1}(z)\times(\tau^k)^{-1}(z)\to p^{-1}(z),$$

is a free and transitive action. For any $\beta\in T^*Q$ and any $t\in\mathbf{R}^k$, we can define k vector fields $\hat{\beta}^{t^k}_i=t^i\xi_i+\beta^{(i)}$. Each $(\hat{\beta}^{t^k}_i)$ is a geodesic vector field and then its flow $\exp t\hat{\beta}^{t^i}_i$ is global. For any $(t^1,\ldots,t^k)\in\mathbf{R}^k$, $(\beta_1,\ldots,\beta_k)\in(T^1_k)^*Q$, and $y\in p^{-1}(z)$, we then define

$$\rho_z(y,t^1,\ldots,t^k,\beta_1,\ldots,\beta_k)=(\exp\hat{\beta}^{t^k}_k)\cdots(\exp\hat{\beta}^{t^2}_2)(\exp\hat{\beta}^{t^1}_1)(y). \qquad \blacksquare$$

Therefore, M is an affine bundle modeled over $\mathbf{R}^k\times(T^1_k)^*Q$, and if $s:Q\to M$ is a global section of p we can define an isomorphism $F_s:M\to\mathbf{R}^k\times(T^1_k)^*Q$ by identifying s with the zero section in M.

Proposition III.6: For any global section s of p, the diffeomorphism $F_s:M\to\mathbf{R}^k\times(T^1_k)^*Q$ defined above, satisfies

$$\eta_i=F_s^*((\eta_0)_i+(\tau^k)^*\beta_i), \quad \omega_i=F_s^*((\omega_0)_i+(\tau^k)^*\phi_i),$$

where $(\eta_0)_i$, $(\omega_0)_i$ are the canonical forms on $\mathbf{R}^k\times(T^1_k)^*Q$.

Proof: The result follows from the formulas:

$$F_*(\alpha^{(i)})=\alpha^{(i)0}, \quad F_*(\xi_i)=(\xi_0)_i,$$

where $\alpha^{(i)}$ and $\alpha^{(i)_0}$ are the (i) lifts of α to M and $\mathbf{R}^k\times(T^1_k)^*Q$, respectively, and ξ_i, $(\xi_0)_i$ are the Reeb vector fields on M and $\mathbf{R}^k\times(T^1_k)^*Q$, respectively. $\blacksquare$

Definition III.7: A k-cosymplectic structure satisfying all the hypotheses of Theorem III.5 will be called *regular*.

Two regular k-cosymplectic structures $(M,p,Q,\eta_i\omega_i,V)$ and $(\overline{M},\overline{p},Q,\overline{\eta_i},\overline{\omega_i},\overline{V})$ over the same manifold Q are equivalent if there exists a bundle isomorphism $F:M\to M$ over the identity on Q, such that

$$F^*\overline{\eta_i}-\eta_i=p^*(df_i), \quad F^*\overline{\omega_i}-\omega_i=p^*(d\beta_i),$$

for all $1\leqslant i\leqslant k$, where f_i is a function on Q and β_i is a 1-form on Q. Then $F^*\overline{\eta_i}-\eta_i$ and $F^*\overline{\omega_i}-\omega_i$ are cohomologous to zero for all $1\leqslant i\leqslant k$.

Proposition III.8: There is a 1–1 correspondence between the set of equivalence classes of regular k-cosymplectic structures over Q and the elements of

$$(H^1(Q,\mathbf{R})\times\cdots\times H^1(Q,\mathbf{R}))\oplus(H^2(Q,\mathbf{R})\times\cdots\times H^2(Q,\mathbf{R})),$$

where $H^1(Q,R)$ and $H^2(Q,R)$ are, respectively, the first and the second de Rham cohomology groups of Q.

The above results are a natural extension of those obtained by Thompson and Schwardmann.[35,36]

An alternative procedure can be developed by using the following theorem due to Nagano.[37]

Theorem III.9: *Suppose that there exists a vector field C on a manifold M satisfying the following conditions.*

(i) *C generates a global one-parameter transformation group on M.*
(ii) *For each point $x\in M$ there exists a unique* $\lim_{t\to-\infty}(\exp tC)(x)$, where $\exp tC$ *denotes the flow of C.*
(iii) *The characteristic operator $(A_C)_x$ associated to C satisfies $((A_C)_x)^2=(A_C)_x$ for each singular point x of C.*
(iv) *The set of S of the singular points of C is a submanifold of M such that codimension S* $=\operatorname{rank}(A_C)_x$, *for all $x\in S$.*

Then there exists a unique vector bundle structure on M such that C is the canonical vector field.

Theorem III.10: *Let M be a $k(n+1)+n$-dimensional manifold endowed with a k-cosymplectic structure $(\eta_i,\omega_i,V;1\leqslant i\leqslant k)$ such that the 1-forms η_i and the 2-forms ω_i are globally exact, i.e., there exist k functions f_i and k 1-forms α_i such that*

$$\eta_i=df_i,\quad \omega_i=d\alpha_i,$$

for all $1\leqslant i\leqslant k$, and with $(\alpha_i)_{]V}=0$ and $\alpha_i(\xi_j)=0$. Let $\overline{C_1},\ldots,\overline{C_k}$ be the vector fields on M, defined by

$$\iota_{\overline{C_i}}\eta_j=0,\quad \iota_{\overline{C_i}}\omega_j=\delta_{ij}\alpha_j,$$

for all $1\leqslant i,j\leqslant k$, and define

$$C_i=f_i\xi_i+\overline{C_i},$$

for $1\leqslant i\leqslant k$, where $\xi_1,\ldots,\xi_k$ are the Reeb vector fields. Then

$$\iota_{C_i}\eta_j=\delta_{ij}f_j,\quad \iota_{C_i}\omega_j=\delta_{ij}\alpha_j.$$

If the vector fields $C_1,\ldots,C_k$ satisfy (i)–(ii) *then there exists a unique vector bundle structure on M that is isomorphic to the stable bundle of k^1-covelocities $\mathbf{R}^k\times(T^1_k)^*S$ of the singular submanifold S of $C=C_1+\cdots+C_k$. Moreover, this isomorphism transports the canonical k-cosymplectic structure and the canonical vector field of $\mathbf{R}^k\times(T^1_k)^*S$ to (η_i,ω_i,V) and C, respectively.*

Proof: The vector field $C=C_1+\cdots+C_k$ satisfies the conditions (i)–(iv) in Theorem III.9, and then there exists a vector bundle isomorphism ϕ over the manifold S of the singular points of C:

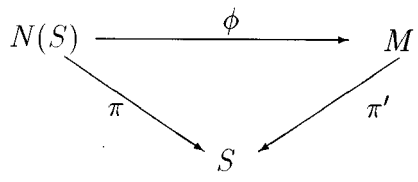

where π is the canonical projection and π' is the projection induced from π via ϕ. Then, for each

$x \in S$, we can transport the k-cosymplectic structure on T_xM to a k-cosymplectic structure $(\eta_i(x), \omega_i(x), M_x)$ on the vector space $T_xM \oplus M_x$, with $M_x = \pi^{-1}(x)$, by using the linear isomorphism:

$$1 + \phi_x : T_xM = T_xS \oplus N_xS \to T_xS \oplus M_x .$$

In fact, if we put $f_i(x) = \phi_x(\xi_i(x))$ and $W_i(x) = \phi_x(V_i(x))$, we have

$$M_x = \langle f_1(x), \ldots, f_k(x) \rangle \oplus W_1(x) \oplus \cdots \oplus W_k(x),$$

and we define a linear isomorphism,

$$G_x : M_x \to \mathbf{R}^k \times (T_x^*S \oplus \cdots \oplus T_x^*S),$$

by

$$G_x(f_i, 0, \ldots, 0) = (e_i 0, \ldots, 0),$$

$$G_x(0, 0, \ldots, v_i, \ldots, 0) = (0, 0, \ldots, \iota_{v_i}\omega_i, \ldots, 0),$$

where $\{e_i, 1 \leqslant i \leqslant k\}$ is the canonical basis of $\mathbf{R}^k$ and $v_i \in W_i(x)$. We finally get a vector bundle isomorphism:

$$G : M = \cup_{x \in S} M_x \to \mathbf{R}^k \times (T_k^1)^*S = \cup_{x \in S} (\mathbf{R}^k \times (T_x^*S \oplus \cdots \oplus T_x^*S)).$$

■

IV. HAMILTONIAN SYSTEMS ON *k*-COSYMPLECTIC MANIFOLDS

In order to introduce dynamics on a k-cosymplectic manifold we need to consider a Hamiltonian function defined on it. The dynamics will be given by k-vector fields, thus we first recall this notion, which is a natural extension of the notion of vector fields. Let M be an arbitrary manifold and

$$\tau^k : T_k^1 M \to M,$$

its tangent bundle of k^1-velocities. Let us recall that $T_k^1 M$ is the manifold of 1-jets at $0 \in \mathbf{R}^k$ of mappings from R^k into M.

Definition IV.1: A section $s : M \to T_k^1 M$ of the projection τ^k will be called a *k-vector field* on M.

Since $T_k^1 M$ may be canonically identified with the Whitney sum of k copies of TM, say

$$T_k^1 M \equiv TM \oplus \cdots \oplus TM,$$

we deduce that a k-vector field s defines k vector fields $X_1, \ldots, X_k$ on M by projecting s onto every factor. From now on, we shall identify s with the k-tuple $(X_1 \ldots, X_k)$.

Definition IV.2: An application $\sigma : U \subset \mathbf{R}^k \to M$ defined on some open neighborhood of $0 \in \mathbf{R}^k$ will be called an *integral section* of a k-vector field $(X_1, \ldots, X_k)$ passing through a point $x \in M$ if and only if

$$\sigma(0) = x, \quad \sigma_*(t)\left(\frac{\partial}{\partial t^i}\right) = X_i(\sigma(t)) \quad \text{for all} \quad t \in U.$$

We say that a k-vector field $(X_1, \ldots, X_k)$ on M is *integrable* if there is an integral section passing through each point of M.

Let us remark that if σ is an integral section of a k-vector field $(X_1, \ldots, X_k)$, then each curve on M defined by $\sigma_i = \sigma \circ J_i$, where $J_i : \mathbf{R} \to \mathbf{R}^k$ is the natural inclusion $J_i(t) = (0, \ldots, t, \ldots, 0)$, is an integral curve of the vector field X_i on M, with $1 \leqslant i \leqslant k$.

Now, assume that M is a k-cosymplectic manifold with k-cosymplectic structure $(\eta_1, \ldots, \eta_k, \omega_1, \ldots, \omega_k, V)$. We define two vector bundle morphisms $\Omega^\flat$ and $\Omega^\sharp$ as follows:

$$\Omega^\flat : TM \to (T^1_k)^* M$$

$$X \to \Omega^\flat(X) = (\iota_X \omega_1 + \eta_1(X)\,\eta_1, \ldots, \iota_X \omega_k + \eta_k(X)\,\eta_k)$$

and

$$\Omega^\sharp : T^1_k M \to T^* M$$

$$(X_1, \ldots, X_k) \to \Omega^\sharp(X_1, \ldots, X_k),$$

such that

$$\Omega^\sharp(X_1, \ldots, X_k)(Y) = \operatorname{trace}((\Omega^\flat(X_j))_i(Y)) = \sum_{i=1}^{k} (\Omega^\flat(X_i))_i(Y) = \sum_{i=1}^{k} (\omega_i(X_i, Y) + \eta_i(X_i)\,\eta_i(Y)),$$

for all $Y \in TM$. The above morphisms induce two morphism of $C^\infty(M)$ modules between the corresponding spaces of sections.

Remark IV.3: If $k=1$ then $\Omega^\flat = \Omega^\sharp$ is defined from TM onto T^*M, and it is the morphism $\chi_{\eta,\omega}$ defined on the cosymplectic manifold (M, η, ω) by (see Refs. 30 and 31)

$$\chi_{\eta,\omega}(X) = \iota_X \omega + \eta(X)\,\eta.$$

If $(s^i, x^\alpha, x^i_\alpha;\ 1 \leqslant i \leqslant k,\ 1 \leqslant \alpha \leqslant n)$ are Darboux coordinates for the k-cosymplectic structure (η_i, ω_i, V), and if the k-vector field $(X_1, \ldots, X_k)$ is expressed with respect to this system by

$$X_i = (X_i)^j \frac{\partial}{\partial s^j} + (X_i)^\alpha \frac{\partial}{\partial x^\alpha} + (X_i)^j_\alpha \frac{\partial}{\partial x^i_\alpha},$$

then

$$\Omega^\sharp(X_1, \ldots, X_k) = \sum_i (X_i)^i\, ds^i - \sum_{i,\alpha} (X_i)^i_\alpha\, dx^\alpha + \sum_{i,\alpha} (X_i)^\alpha\, dx^i_\alpha. \tag{4}$$

Let $H : M \to \mathbf{R}$ be a function on M. If $(X_1, \ldots, X_k)$ is a k-vector field on M, the equations

$$\eta_i(X_j) = \delta_{ij}, \quad \forall i, j,$$

$$\Omega^\sharp(X_1, \ldots, X_k) = dH + \sum_{i=1}^{k} (1 - \xi_i(H))\,\eta_i, \tag{5}$$

would imply

$$(X_i)^j = \delta_{ij} \quad \frac{\partial H}{\partial x^\alpha} = -\sum_{i=1}^{k} (X_i)^i_\alpha, \quad \frac{\partial H}{\partial x^i_\alpha} = (X_i)^\alpha. \tag{6}$$

From these local conditions we can define, in a neighborhood of each point $x \in M$, a k-vector field that satisfies (6). For example, we can put

$$(X_i)^j = \delta_{ij}, \quad (X_1)^1_\alpha = \frac{\partial H}{\partial x^\alpha}, \quad (X_i)^j_\alpha = 0, \quad \text{for } i \neq 1 \neq j, \quad (X_i)^\alpha = \frac{\partial H}{\partial x^i_\alpha}.$$

Now one can construct a global k-vector field, which is a solution of (6), by using a partition of the unity.

Remark IV.4: Equations (6) have not, in general, a unique solution. In fact, if we denote by $\mathscr{M}_k(C^\infty(M))$ the space of matrices of order k whose entries are functions on M, and we define the map

$$\eta^\sharp : T^1_k M \to \mathscr{M}_k(C^\infty(M))$$

$$(X_1,\ldots,X_k) \to (\eta_i(X_j)),$$

the solutions of (6) are given by $(X_1,\ldots,X_k)+(\ker \Omega^\sharp \cap \ker\ \eta^\sharp)$, where $(X_1,\ldots,X_k)$ is a particular solution.

Definition IV.5: Any k-vector field $(X_1,\ldots,X_k)$ on M, such that

$$\eta_i(X_j)=\delta_{ij},$$

$$\Omega^\sharp(X_1,\ldots,X_k)=dH+\sum_{i=1}^{k}(1-\xi_i(H))\eta_i,$$

for all $1\leqslant i,j\leqslant k$, will be called an *evolution k-vector field* on M associated with the Hamiltonian function H.

Let $(X_1,\ldots,X_k)$ be an evolution k-vector field associated to H, and assume that it is integrable. Let

$$\sigma : \mathbf{R}^k \to M$$

$$(t^i) \to (\sigma^j(t^i), \alpha^\alpha(t^i), \sigma^j_\alpha(t^i)),$$

be an integral section of $(X_1,\ldots,X_k)$; then, we have

$$\frac{\partial \sigma^j}{\partial t^i}=\delta_{ij}, \quad \frac{\partial \sigma^\alpha}{\partial t^i}=(X_i)^\alpha, \quad \frac{\partial \sigma^j_\alpha}{\partial t^i}=(X_i)^j_\alpha,$$

for all $1\leqslant i,\ j\leqslant k$ and $1\leqslant\alpha\leqslant n$. Therefore, Eqs. (6) give

$$\frac{\partial H}{\partial x^\alpha}=-\sum_{i=1}^{k}\frac{\partial \sigma^i_\alpha}{\partial t^i},$$

$$\frac{\partial H}{\partial x^i_\alpha}=\frac{\partial \sigma^\alpha}{\partial t^i},$$

with $1\leqslant i\leqslant k$ and $1\leqslant\alpha\leqslant n$, which are the field equations for H.

Remark IV.6: Let $(X_1,\ldots,X_k)$ be an evolution k-vector field. Since $\eta_i(X_j)=\delta_{ij}$, it follows that the vector fields $X_1,\ldots,X_k$ on M are linearly independent.

Example IV.7: We shall use the above formalism to obtain an intrinsic version for the electrostatic equations. Consider $\mathbf{R}^3$ with a metric g with components g_{ij}. Let $\phi:\mathbf{R}^3\to\mathbf{R}$ be the electric potential and $p=(p_1,p_2,p_3)$: $\mathbf{R}^3\to\mathbf{R}^3$ the electric field. Denote by (t^1,t^2,t^3) the standard coordinates on $\mathbf{R}^3$, and set $\sqrt{g}=\sqrt{\det g_{ij}}$. By $r(t)$ we denote the scalar function that gives the density of electric charge on $\mathbf{R}^3$. Consider on $M=\mathbf{R}^3\times(T^1_3)^*\mathbf{R}$ the canonical 3-cosymplectic structure $(\eta_i,\omega_i,V;1\leqslant i\leqslant 3)$. We denote by $(t^1,t^2,t^3,x,x^1,x^2,x^3)$ the local coordinates on $\mathbf{R}^3\times(T^1_3)^*\mathbf{R}$ induced by the standard coordinate (x) on $\mathbf{R}$, and define a Hamiltonian function,

$$H:\mathbf{R}^3\times(T^1_3)^*\mathbf{R}\to\mathbf{R}$$

$$(t^1,t^2,t^3,x,x^1,x^2,x^3)\to\sqrt{g}\left(4\pi r(t^1,t^2,t^3)x+\frac{1}{2g}\,g_{ij}x^ix^j\right).$$

Consider the equations

$$\eta i_i(X_j)=\delta_{ij} \quad (1\leqslant i,j\leqslant 3),$$

$$\Omega^\sharp(X_1,X_2,X_3)=dH+\sum_{i=1}^{3}(1-\xi_i(H))\eta_i, \tag{7}$$

where ξ_1, ξ_2, ξ_3 are the Reeb vector fields and (X_1,X_2,X_3) is a 3-vector field on $\mathbf{R}^3\times(T_3^1)^*\mathbf{R}$. Let $\sigma:\mathbf{R}^3\to\mathbf{R}^3\times(T_3^1)^*\mathbf{R}$, $\sigma(t)=(t,\psi(t),\psi^1(t),\psi^2(t),\psi^3(t))$ be an integral section of an evolution 3-vector field that is a solution of (7). Then we obtain

$$4\pi r(t^1,t^2,t^3)\sqrt{g}=-\left(\frac{\partial\psi^1}{\partial t^1}+\frac{\partial\psi^2}{\partial t^2}+\frac{\partial\psi^3}{\partial t^3}\right),$$

$$\frac{1}{\sqrt{g}}\left(\sum_{i=1}^{3}g_{ij}\psi^j\right)=\frac{\partial\psi}{\partial t^i},$$

which are the electrostatic equations, and then the components $\psi(t)$ and $(\psi^1(t),\psi^2(t),\psi^3(t))$ of σ are, respectively, the electric potential ϕ and electric field $p=(p_1,p_2,p_3)$ on $\mathbf{R}^3$.

Actually, in most of the real electrostatic problems, one has to deal with finite regions and with the boundary conditions established in the limits of those regions. To do that the Green identities have to be used.

Thus, the suitable boundary conditions to the Poisson (Laplace) equations to present a unique well-behaved solution inside a finite region are modeled. Physically, we know that the specification of the potential on a closed surface defines a unique boundary problem (Dirichlet boundary conditions). The specification of the electric field at any point of the closed surface, i.e., the specification of the distribution charge on the surface (Neumann boundary conditions), also defines a unique problem. Those propositions can be shown by taking into account the first Green identity (see Ref. 38 and the references therein).

V. R^k-VALUED HAMILTONIANS

We can also consider Hamiltonian functions $\mathbf{R}^k$-valued on a k-cosymplectic manifold M.

In Darboux coordinates we have

$$H=(H^1(s^i,x^\alpha,x^i_\alpha),\ldots,H^k(s^i,x^\alpha,x^i_\alpha)),$$

or equivalently,

$$H=H^ie_i,$$

where $\{e_1,\ldots,e_k\}$ denotes the standard basis in $\mathbf{R}^k$.

Definition V.1: If M is a k-cosymplectic manifold, a vector field X on M such that

$$\eta_i(X)=1, \quad \text{for all } i=1,...,k,$$

$$\Omega^b(X)=dH+\left[\sum_{i=1}^{k}((1-\xi_i(H^i))\eta_i)e_i\right], \tag{8}$$

where $\xi_1,\ldots,\xi_k$ are the Reeb vector fields associated with the k-cosymplectic structure, will be called a *k-Hamiltonian system* associated to H.

In Darboux coordinates, we have

$$X=\frac{\partial}{\partial s^1}+\cdots+\frac{\partial}{\partial s^k}+X^\alpha\frac{\partial}{\partial x^\alpha}+X^i_\alpha\frac{\partial}{\partial x^i_\alpha},$$

where

$$\frac{\partial H^i}{\partial s^j}=0, \quad \text{if } i\neq j, \quad \frac{\partial H^i}{\partial x^\alpha}=-X^i_\alpha, \quad \frac{\partial H^i}{\partial x^j_\alpha}=\delta_{ij}X^\alpha.$$

If

$$\sigma:(-\epsilon,\epsilon)\subset\mathbf{R}\to M$$

$$t\to(\sigma^i(t),\sigma^\alpha(t),\sigma^i_\alpha(t))$$

is an integral curve of X, then we obtain the Hamilton equations associated to the $\mathbf{R}^k$-valued Hamiltonian H:

$$\frac{\partial H^i}{\partial x^\alpha}=-\frac{d\sigma^i_\alpha}{dt},$$

$$\frac{\partial H^i}{\partial x^j_\alpha}=\delta_{ij}\frac{d\sigma^\alpha}{dt}.$$

From these equations and using the same argument as in Refs. 24 and 25, we can prove the following result.

Proposition V.2: An admissible Hamiltonian function $H=(H^1,\ldots,H^k)$ is necessarily of the form

$$H^i=\sum_{\alpha=1}^{n}\varphi_\alpha(s^i,x_1,\ldots,x_n)x^i_\alpha+\psi^{\,i}(s^i,x_1,\ldots,x_n),$$

for $1\leqslant i\leqslant k$.

Given an admissible Hamiltonian function $H=(H^1,\ldots,H^k)$ on M, we can define the local vector field,

$$X=\frac{\partial}{\partial s^1}+\cdots+\frac{\partial}{\partial s^k}+\frac{\partial H^i}{\partial x^i_\alpha}\frac{\partial}{\partial x^\alpha}-\frac{\partial H^i}{\partial x^\alpha}\frac{\partial}{\partial x^i_\alpha},$$

which satisfies (8). By using a partition of the unity, it is possible to define a unique global vector field X on M satisfying (8).

An application. Let $L:\mathbf{R}\times TQ\to\mathbf{R}$ be a hyperregular Lagrangian function, that is, the Legendre transformation $\mathrm{Leg}:\mathbf{R}\times TQ\to\mathbf{R}\times T^*Q$ defined by L is a diffeomorphism. Denote by $H=E_L\circ\mathrm{Leg}^{-1}$ the Hamiltonian energy, where $E_L=C(L)-L$, C being the infinitesimal generator of the dilations on TQ. We construct a 2-cosymplectic structure on the manifold $M=\mathbf{R}^2\times(TQ\oplus T^*Q)$ as follows.

First of all, let ω_Q denote the canonical symplectic form on T^*Q and $\omega_L=\mathrm{Leg}^*\,\omega_Q$ the Poincaré–Cartan two-form on TQ.[39] If (q^α) are coordinates on Q we denote by (q^α,v^α) and (q^α,p_α) the induced coordinates on TQ and T^*Q, respectively. So, we have coordinates $(t^1,t^2,q^\alpha,v^\alpha,p_\alpha)$ on M. Denote by τ: $M\to\mathbf{R}\times TQ$ and $\pi:M\to\mathbf{R}\times T^*Q$ the canonical projections, locally given by

$$\tau(t^1,t^2,q^\alpha,v^\alpha,p_\alpha)=(t^1,q^\alpha,v^\alpha),$$

$$\pi(t^1,t^2,q^\alpha,v^\alpha,p_\alpha)=(t^2,q^\alpha,p_\alpha).$$

If we put

$$\eta_1=dt^1, \quad \eta_2=dt^2, \quad \omega_1=\tau^*\omega_L, \quad \omega_2=\pi^*\omega_Q,$$

we deduce that $(\eta_1,\eta_2,\omega_1,\omega_2,V)$ is a 2-cosymplectic structure on M where $V=\mathrm{Ker}\,T\rho$, $\rho:M\to Q$ being the canonical projection.

Consider now a Hamiltonian function,

$$H:M\to\mathbf{R}^2,$$

given by

$$H(t^1,t^2,X_q,\alpha_q)=(\tau^*E_L,\langle X_q,\alpha_q\rangle-\tau^*L).$$

In local coordinates we obtain

$$H(t^1,t^2,q^\alpha,v^\alpha,p_\alpha)=\left(v^\alpha\frac{\partial L}{\partial v^\alpha}-L,q^\alpha p_\alpha-L\right).$$

Let now X be a vector field on M, which is Hamiltonian for H. After a direct computation, we deduce

$$X=\frac{\partial}{\partial t^1}+\frac{\partial}{\partial t^2}+v^\alpha\frac{\partial}{\partial q^\alpha}+X^1_\alpha\frac{\partial}{\partial v^\alpha}-\frac{\partial H}{\partial q^\alpha}\frac{\partial}{\partial p_\alpha},$$

where

$$\frac{\partial^2 L}{\partial t^1\,\partial v^\beta}+v^\alpha\frac{\partial^2 L}{\partial q^\alpha\,\partial v^\beta}+X^1_\alpha\frac{\partial^2 L}{\partial v^\alpha\,\partial v^\beta}-\frac{\partial L}{\partial q^\beta}=0.$$

If $\sigma(t)=(\sigma_1(t),\sigma_2(t),\sigma^\alpha(t),\sigma^1_\alpha(t),\sigma^2_\alpha(t))$ is an integral curve of X, we have

$$\frac{d\sigma^\alpha}{dt}=v^\alpha,\quad \frac{d\sigma^1_\alpha}{dt}=X^1_\alpha,\quad \frac{d\sigma^2_\alpha}{dt}=-\frac{\partial H}{\partial q^\alpha}.$$

Therefore, the projections of σ onto Q satisfy simultaneously the Euler–Lagrange and Hamilton equations.

VI. MULTISYMPLECTIC FORMALISM ON JET BUNDLES

In this section we obtain the Hamiltonian formalism above as a particular case of the multisymplectic formalism on jet bundles.[12]

Let $\pi:E\to M$ be a fibered manifold. We denote by Λ^k_0E and Λ^k_1E the following vector bundles on E:

$$\Lambda^k_0E=\{\sigma\in\Lambda^kE/\iota_v\sigma=0,\quad \forall_v\in V\pi\},$$

$$\Lambda^k_1E=\{\sigma\in\Lambda^kE/\iota_{v_1}\iota_{v_2}\sigma=0,\quad \forall v_1,v_2\in V\pi\},$$

where $V\pi$ is the vertical bundle of π. Here, Λ^rE denotes the space of r-forms on the manifold E, that is,

$$\Lambda^rE=\cup_{z\in E}\Lambda^r(T^*_zE).$$

If we take fibered coordinates (s^i,x^α), then each element in Λ^k_0E can be expressed as $x(s^i,x^\alpha)d^ks$, and each element of Λ^k_1E as $x(s^i,x^\alpha)d^ks+x^i_\alpha(s^i,x^\alpha)dx^\alpha\wedge d^{k-1}s_i$, where $d^ks=ds^1\wedge\cdots\wedge ds^k$ and $d^{k-1}s^i=\iota(\partial/\partial s^i)\,d^ks$. Thus, we can introduce local coordinates (s^i,x^α,x) and $(s^i,x^\alpha,x,x^i_\alpha)$ on Λ^k_0E and Λ^k_1E, respectively.

We denote by $\mathscr{M}\pi=\Lambda^k_1E$. Since Λ^k_0E is a vector subundle of $\mathscr{M}\pi$, then $\mathscr{M}\pi/\Lambda^k_0E=J^1\pi^*$ is a quotient manifold of dimension $(k+1)n+k$ with local coordinates $(s^i,x^\alpha,x^i_\alpha)$.

Moreover, we have the following canonical projections:

$$\mu: \mathscr{M}\pi \to J^1\pi^*,$$

$$\pi^*_{1,0}: J^1\pi^* \to E,$$

$$\pi^*_1: J^1\pi^* \to M.$$

On $\mathscr{M}\pi$ there is a canonical k-form Θ, which is defined as follows:

$$\Theta_\omega(X_1,\ldots,X_k) = \omega(\nu_* X_1,\ldots,\nu_* X_k),$$

for all $X_i \in T_\omega \mathscr{M}\pi$ and $\omega \in \mathscr{M}\pi$, where $\nu: \mathscr{M}\pi \to E$ is the canonical projection.

Definition VI.1: The $(k+1)$-form on $\mathscr{M}\pi$, defined by

$$\Omega = d\Theta,$$

is called the *multisymplectic structure* on $\mathscr{M}\pi$.

Definition VI.2: A global section $h: J^1\pi^* \to \mathscr{M}\pi$ of μ is called a *Hamiltonian.* Given a Hamiltonian h on $J^1\pi^*$, we can define a $(k+1)$-form on $J^1\pi^*$ given by

$$\Omega_h = -h^*\Omega.$$

Definition VI.3: A section $\gamma: M \to J^1\pi^*$ of π^*_1 satisfies the Hamilton equations associated to a Hamiltonian h if and only if

$$\gamma^*(\iota_X \Omega_h) = 0,$$

for all vector field X on $J^1\pi^*$.

Put

$$h(s^i, x^\alpha, x^i_\alpha) = (s^i, x^\alpha, x = -H(s^i, x^\alpha, x^i_\alpha), x^i_\alpha),$$

with $H: J^1\pi^* \to \mathbf{R}$.

Then a section $\gamma(s^i) = (s^i, \gamma^\alpha(s^i), \gamma^j_\alpha(s^i))$ as in Definition VI.3 satisfies the following system of differential equations:

$$\frac{\partial H}{\partial x^\alpha} = -\sum_{i=1}^{k} \frac{\partial \gamma^i_\alpha}{\partial s^i},$$

$$\frac{\partial H}{\partial x^i_\alpha} = \frac{\partial \gamma^\alpha}{\partial s^i},$$

which are the field equations for H (see Sec. IV).

Next, we shall consider the particular case where $\pi: E \to M$ is the trivial bundle $(\mathbf{R}^k \times Q, \pi, \mathbf{R}^k)$.

In this case we get

$$\mathscr{M}\pi = \mathbf{R}^k \times \mathbf{R} \times (T^1_k)^* Q, \quad J^1\pi^* = \mathbf{R}^k \times (T^1_k)^* Q.$$

Consider the canonical k-cosymplectic structure $((\eta_0)_i, (\omega_0)_i, V_0)$ on $J^1\pi^* = \mathbf{R}^k \times (T^1_k)^* Q$ described in Sec. II. Notice that $(\omega_0)_i = -d(\lambda_0)_i$, where $(\lambda_0)_i = x^i_\alpha \, dx^\alpha$, and $(t^i, x^\alpha, x^i_\alpha)$ are Darboux coordinates (see Sec. II).

If $h: J^1\pi^* = \mathbf{R}^k \times (T^1_k)^* Q \to \mathscr{M} = \mathbf{R}^k \times \mathbf{R} \times (T^1_k)^* Q$ is a Hamiltonian on $J^1\pi^*$ such that

$$h(t^i, x^\alpha, x_\alpha) = (t^i, -H(t^i, x^\alpha x^i_\alpha), x^\alpha, x^\alpha_i),$$

then we obtain the following.

Lemma VI.4: We have

$$\Theta_h = -H(\eta_0)_1 \wedge \cdots \wedge (\eta_0)_k + \sum_{i=1}^{k} (\lambda_0)_i \wedge \iota_{\xi_i}((\eta_0)_1 \wedge \cdots \wedge (\eta_0)_k),$$

where $(\xi_0)_i = \partial/\partial t^i (1 \leqslant i \leqslant k)$, *are the Reeb vector fields on* $\mathbf{R}^k \times (T_k^1)^*Q$.

Thus, we deduce the following.

Proposition VI.5: A section ψ *of the canonical projection* $\pi^*: \mathbf{R}^k \times (T_k^1)^*Q \to \mathbf{R}^k$ *satisfies the Hamilton equations in Definition VI.3 if and only if* ψ *is an integral section of an evolution k-vector field on* $\mathbf{R}^k \times (T_k^1)^*Q$ *associated to* H.

ACKNOWLEDGMENTS

This work has been partially supported through grants DGICYT (Spain) (Project No. PB94-0106) and XUNTA DE GALICIA (Projects No. XUGA13101A94 and No. XUGA13101A96). MdL also acknowledges the support by FAPERJ-Proc. E-26/170.236/96, RJ, Brazil. PR wishes to thank the hospitality of the Departamento de Xeometría e Topoloxía, Universidade de Santiago de Compostela, where part of this work was prepared. We also acknowledge the referee for his valuable comments and remarks.

[1] J. Śniatycki, ''On the geometric structure of classical field theory in Lagrangian formulation,'' Proc. Cambridge Philos. Soc. **68**, 475–483 (1970).
[2] J. Kijowski, ''A finite-dimensional canonical formalism in the classical field theory,'' Commun. Math. Phys. **30**, 99–128 (1973).
[3] J. Kijowski and W. Szczyrba, ''Multisymplectic manifolds and the geometric construction of the Poisson bracket in field theory,'' in *Géométrie Symplectique et Physique Mathématique* (CNRS, Paris, 1975), pp. 347–378.
[4] J. Kijowski and W. M. Tulczyjew, *A Symplectic Framework for Field Theories*, Lecture Notes in Physics (Springer-Verlag, Berlin, 1979), p. 107.
[5] P. L. García and A. Pérez-Rendón, ''Symplectic approach to the theory of quantized fields. I,'' Commun. Math. Phys. **13**, 24–44 (1969).
[6] P. L. García and A. Pérez-Rendón, ''Symplectic approach to the theory of quantized fields. II,'' Arch. Ration. Mech. Anal. **43**, 101–124 (1971).
[7] H. Goldschmidt and S. Sternberg, ''The Hamilton–Cartan formalism in the calculus of variations,'' Ann. Inst. Fourier (Grenoble) **23**, 203–267 (1975).
[8] G. Martin, ''Dynamical structures for *k*-vector fields,'' Int. J. Theor. Phys. **27**, 571–585 (1988).
[9] G. Martin, ''A Darboux theorem for multi-symplectic manifolds,'' Lett. Math. Phys. **16**, 133–138 (1988).
[10] M. J. Gotay, ''An exterior differential systems approach to the Cartan form,'' in *Symplectic Geometry and Mathematical Physics*, Progress in Mathematics (Birkhauser, Boston, 1991), Vol. 99, pp. 160–188.
[11] M. J. Gotay, ''A multisymplectic framework for classical field theory and the calculus of variations I: Covariant Hamiltonian formalism,'' in *Mechanics, Analysis and Geometry: 200 Years after Lagrange*, North-Holland Delta Series (North-Holland, Amsterdam, 1991), pp. 203–235.
[12] J. F. Cariñena, M. Crampin, and L. A. Ibort, ''On the multisymplectic formalism for first order field theories,'' Diff. Geom. Appl. **1**, 345–374 (1991).
[13] F. Cantrijn, A. Ibort, and M. de León, ''On the geometry of multisymplectic manifolds,'' preprint IMAFF, 1996.
[14] F. Cantrijn, A. Ibort, and M. de León, ''Hamiltonian structures on multisymplectic manifolds,'' to appear in *Proceedings of the Workshop on Geometry and Physics on the occasion of the 65th birthday of W. Tulczyjew*, IIASS, Vietri sul Mare (Italy), 29 September–3 October 1996.
[15] M. de León, J. C. Marrero, and J. Marín-Solano, ''A geometrical approach to Classical Field Theories: A constraint algorithm for singular theories,'' in *New Developments in Differential Geometry (Debrecen, 1994)*, Mathematical Applications (Kluwer Academic, Dordrecht, 1996), Vol. 350, pp. 291–312.
[16] M. de León, J. C. Marrero, and J. Marín-Solano, ''Ehresmann connections in classical field theories,'' in *Proceedings of the III Fall Workshop: Differential Geometry and its Applications*, Granada, 19–20 September 1994, Anales de Física, Monografías (CIEMAT, Madrid, 1995), Vol. 2, pp. 73–89.
[17] G. Sardanashvily, *Gauge Theory in Jet Manifolds* (Hadronic, Palm Harbor, 1993).
[18] G. Sardanashvily, *Generalized Hamiltonian Formalism for Field Theory: Constrained Systems* (World Scientific, Singapore, 1995).
[19] I. V. Kanatchikov, ''Basic structures of the covariant canonical formalism for fields based on the De Donder–Weyl theory,'' PITHA 94/47.
[20] A. Echevarría-Enríquez, M. C. Muñoz-Lecanda, and N. Román-Roy, ''Geometry of Lagrangian first-order classical field theories,'' Fortschr. Phys. **44**, 235–280 (1996).
[21] M. de León and P. R. Rodrigues, *Generalized Classical Mechanics and Field Theory*, North-Holland Mathematics Studies (North-Holland, Amsterdam, 1985), p. 112.
[22] E. Binz, J. Śniatycki, and H. Fisher, *Geometry of Classical Fields*, North-Holland Mathematics Studies (North-Holland, Amsterdam, 1989), p. 154.
[23] Ch. Günther, ''The polysymplectic Hamiltonian formalism in field theory and calculus of variations I: The local case,'' J. Diff. Geom. **25**, 23–53 (1987).
[24] A. Awane, ''Structures *k*-symplectiques,'' Doctoral dissertation, Mulhouse, 1992.
[25] A. Awane, *k*-symplectic structures, J. Math. Phys. **33**, 4046–4052 (1992).

[26] M. Puta, ''Some remarks on the k-symplectic manifolds,'' Tensor, N.S. **47**, 109–115 (1988).
[27] L. K. Norris, ''Symplectic geometry on T^*M derived from n-symplectic geometry on LM,'' J. Geom. Phys. **13**, 51–78 (1994).
[28] M. de León, I. Méndez, and M. Salgado, ''p-almost cotangent structures,'' Boll. Unione Mat. Italiana **7-A**, 97–107 (1993).
[29] M. de León, I. Méndez, and M. Salgado, ''Regular p-almost cotangent structures,'' J. Korean Math. Soc. **25**, 273–287 (1988).
[30] C. Albert, ''Le théoreme de réduction de Marsden–Weinstein en géometrie cosymplectique et de contact,'' J. Geom. Phys. **6**, 627–649 (1994).
[31] F. Cantrijn, M. de León, and E. A. Lacomba, ''Gradient vector fields on cosymplectic manifolds,'' J. Phys. A **25**, 175–188 (1992).
[32] M. de León, E. Merino, J. A. Oubiña, and M. Salgado, ''Stable almost cotangent structures,'' to appear in Boll. Unione Mat. Italiana.
[33] A. Fujimoto, ''Theory of G-structures,'' Publications of the Study Group of Geometry, Tokyo, 1972, Vol. 1.
[34] E. T. Kobayashi, ''A remark on the Nijenhuis tensor,'' Pac. J. Math. **12**, 963–977 (1962).
[35] G. Thompson, ''Integrable almost cotangent structures and Legendrian bundles,'' Math. Proc. Cambridge Philos. Soc. **101**, 61–78 (1987).
[36] G. Thompson and U. Schwardmann, ''Almost tangent and cotangent structures in the large,'' Trans. Am. Math. Soc. **327**, 313–328 (1991).
[37] T. Nagano, ''1-forms with the exterior derivate of maximal rank,'' J. Diff. Geom. **2**, 253–264 (1968).
[38] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).
[39] M. de León and P. R. Rodrigues, *Methods of Differential Geometry in Analytical Mechanics*, North-Holland Mathematics Series 152 (North-Holland, Amsterdam, 1989).

On the derivation of boundary integral equations for scattering by an infinite two-dimensional rough surface

J. A. DeSanto
Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, Colorado 80401-1887

P. A. Martin
Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom

(Received 9 June 1997; accepted for publication 17 September 1997)

A plane acoustic wave is incident upon an infinite, rough, impenetrable surface S. The aim is to find the scattered field by deriving a boundary integral equation over S, using Green's theorem and the free-space Green's function. This requires careful consideration of certain integrals over a large hemisphere of radius r; it is known that these integrals vanish as $r\to\infty$ if the scattered field satisfies the Sommerfeld radiation condition, but that is not the case here—reflected plane waves must be present. It is shown that the well-known Helmholtz integral equation is not valid in all circumstances. For example, it is not valid when the scattered field includes plane waves propagating away from S along the axis of the hemisphere. In particular, it is not valid for the simplest possible problem of a plane wave at normal incidence to an infinite flat plane. Some suggestions for modified integral equations are discussed. © *1998 American Institute of Physics.* [S0022-2488(98)01302-4]

I. INTRODUCTION

A bounded three-dimensional obstacle with a smooth surface S is surrounded by a compressible fluid. A plane time-harmonic sound wave is incident on the obstacle; the problem is to calculate the scattered field u. In order to have a well-posed boundary-value problem (with existence and uniqueness), one imposes the Sommerfeld radiation condition,

$$r\left(\frac{\partial u}{\partial r}-iku\right)\to 0 \quad \text{as } r\to\infty, \tag{1}$$

uniformly in all directions. Here r is a spherical polar coordinate, k is the wave number, and the time-dependence is $e^{-i\omega t}$. Physically, the radiation condition ensures that the scattered waves propagate outwards, away from the obstacle.

A well-known method for solving the above problem is to derive a boundary integral equation for the boundary values of u on S. In the derivation, Green's theorem is applied to u and a fundamental solution G, in the region bounded internally by S and externally by C_r, a large sphere of radius r. It turns out that the radiation condition implies that the integral

$$I(u;C_r)\equiv\int_{C_r}\left(u\frac{\partial G}{\partial r}-G\frac{\partial u}{\partial r}\right)ds\to 0 \quad \text{as } r\to\infty, \tag{2}$$

and so only boundary integrals over S remain. For more information, see Colton and Kress.[1]

Assume that S is a sound-hard surface (Neumann condition). Then, the method described above leads to the following boundary integral equation:

$$u(p)-\int_S u(q)\frac{\partial G}{\partial n_q}(p,q)ds_q=\int_S\frac{\partial u_{\text{inc}}}{\partial n_q}G(p,q)ds_q, \quad p\in S, \tag{3}$$

 © 1998 American Institute of Physics

here, u_{inc} is the given incident wave. One can also derive an equation for the boundary values of the total field $u_{\text{tot}}=u_{\text{inc}}+u$; this boundary integral equation is

$$u_{\text{tot}}(p)-\int_S u_{\text{tot}}(q)\frac{\partial G}{\partial n_q}(p,q)\,ds_q=2u_{\text{inc}}(p), \quad p\in S. \tag{4}$$

We shall refer to Eqs. (3) and (4) as *standard Helmholtz integral equations*. Similar equations can be derived for sound-soft surfaces ($u_{\text{tot}}=0$ on S).

Suppose now that the obstacle is *unbounded*. The prototypical problem is scattering (reflection) of a plane wave by an infinite flat plane, S. As is well known, the incident wave is reflected specularly as a single propagating plane wave. More generally, if S is an infinite rough surface, an incident plane wave will be scattered into a spectrum of plane waves. For such problems, the Sommerfeld radiation condition is definitely inappropriate as it is not satisfied by a plane wave. Nevertheless, it is customary to proceed, *assuming* that the scattered field can be represented in terms of plane waves, at least at some distance from S. Typically, this requires the discarding of an integral such as (2), but with the large sphere C_r replaced by a large *hemisphere* H_r. This paper began as an attempt to justify this step.

In a previous paper,[2] we derived boundary integral equations of Helmholtz type for *one-dimensional* rough surfaces. We found that the standard Helmholtz integral equations are valid, except that the right-hand side of Eq. (4) must be replaced by $u_{\text{inc}}(p)$ for grazing incidence.

It is perhaps surprising that an analysis of this kind has not been given before: most authors have been content to write down an integral equation such as Eq. (4), prior to extensive numerical computations. However, it turns out that the necessary analysis for scattering by a *two-dimensional* rough surface is not straightforward and, moreover, it yields some surprises. For example, the simplest problem, namely reflection of a plane wave at *normal* incidence upon a *flat* surface, leads to divergent integrals: the standard Helmholtz integral equation (3) is not valid for this problem.

The paper is organized as follows. Section II is devoted to formulating the problem, with some background on angular-spectrum representations and integral representations (using G). Green's theorem is applied inside a volume whose closed boundary is made up of three pieces: the large hemisphere H_r; a large circular piece, S_r, of the rough surface; and a cylindrical surface T_r, joining S_r and H_r. Estimation of integrals over H_r is carried out in Secs. III–V. Thus the method of stationary phase for multiple integrals and an expansion method are used in Secs. III and IV, respectively, but only for a single plane wave. Results for $I(u;H_r)$ are obtained in Sec. V. The contribution from integrating over T_r is considered in Sec. VI. Unlike in two dimensions (one-dimensional rough surface), this contribution may not be negligible; it is evaluated under additional, but reasonable, *a priori* assumptions on the form of the scattered field near S. This is a weakness of the present analysis. Finally, boundary integral equations of the Helmholtz type are derived in Sec. VII. Further work is needed to tighten up the analysis and to investigate the numerical consequences.

II. FORMULATION

Consider the scattering of a plane wave by a two-dimensional rough surface, S, described by

$$z=s(x,y), \quad -\infty<x<\infty, \quad -\infty<y<\infty$$

with $-h<s(x,y)\leqslant 0$ and some constant $h\geqslant 0$. The acoustic medium occupies $z>s$ and, for definiteness, we assume that S is a smooth, sound-hard surface. Thus we can write the total field as

$$u_{\text{tot}}=u_{\text{inc}}+u,$$

where u is the scattered field. The incident plane wave is

$$u_{\text{inc}}(r,\theta,\phi)=\exp\{i\mathbf{k}_i\cdot\mathbf{x}\}, \quad 0\leqslant\theta_i\leqslant\tfrac{1}{2}\pi, \tag{5}$$

where $\mathbf{k}_i = k(\sin\theta_i, 0, -\cos\theta_i)$, θ_i is the angle of incidence (it is the angle between the direction of propagation and the negative z-axis),

$$\mathbf{x} = r\hat{\mathbf{x}} = r(\sin\theta\,\cos\phi, \sin\theta\,\sin\phi, \cos\theta),$$

and (r, θ, ϕ) are spherical polar coordinates: $x = r\,\sin\theta\cos\phi$, $y = r\,\sin\theta\sin\phi$ and $z = r\cos\theta$. All the fields $u_{\rm tot}$, $u_{\rm inc}$, and u satisfy the Helmholtz equation,

$$(\nabla^2 + k^2)u = 0$$

for $z > s$. The boundary condition is

$$\frac{\partial u_{\rm tot}}{\partial n} = 0 \quad \text{on } S, \tag{6}$$

where $\partial/\partial n$ denotes normal differentiation *out* of the acoustic medium.

A. Reflection by a flat surface

It is instructive to consider the very simple problem of reflection by a flat surface, so that $s = 0$. The textbook solution for the scattered field is

$$u(r, \theta, \phi) = \exp\{i\mathbf{k}_{\rm s}\cdot\mathbf{x}\} \quad \text{for } 0 \leqslant \theta_i < \tfrac{1}{2}\pi, \tag{7}$$

where $\mathbf{k}_{\rm s} = k(\sin\theta_i, 0, \cos\theta_i)$. When $\theta_i = \frac{1}{2}\pi$ (''grazing incidence''), we have $u \equiv 0$: the incident wave satisfies the boundary condition on S.

So, for $0 \leqslant \theta_i < \frac{1}{2}\pi$,

$$u_{\rm tot} = 2\,e^{ikx\sin\theta_i}\cos(kz\cos\theta_i)$$

solves the problem. But consider

$$u'_{\rm tot} \equiv u_{\rm tot} + u_g \tag{8}$$

with

$$u_g = V(\beta)e^{ik(x\cos\beta + y\sin\beta)},$$

where β and $V(\beta)$ are arbitrary, with $-\pi < \beta \leqslant \pi$. $u'_{\rm tot}$ also ''solves'' the problem, in that it satisfies the Helmholtz equation and the boundary condition. Of course, we disallow this second solution, unless $V \equiv 0$: but why? The answer is because of the radiation condition (which we have yet to specify). For example, take $\beta = 0$ and $V(0) = 1$, so that $u_g = e^{ikx}$; this gives an ''outgoing'' grazing wave at $x = +\infty$ but it is an ''incoming'' grazing wave at $x = -\infty$, we must therefore exclude it. Indeed, we must exclude *all* contributions u_g, for any β and V.

A similar condition is imposed on the two-dimensional problem.[3] However, the three-dimensional problem has another feature, for we could consider replacing u_g in Eq. (8) by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} V(\beta)e^{ik(x\cos\beta + y\sin\beta)}d\beta,$$

where V is a continuous function; but, as u_g has been excluded, we must also exclude all linear combinations of such plane grazing waves. In particular, by taking $V(\beta) = (-i)^n e^{in\beta}$, we see that we must exclude the cylindrical standing waves

$$J_n(kR)e^{in\phi}, \tag{9}$$

where J_ν is a Bessel function,[4] $R = r\sin\theta$, and (R, ϕ, z) are cylindrical polar coordinates of the point at $\mathbf{x}$. On the other hand, the exact scattered field, given by Eq. (7), when evaluated on any plane $z =$ constant, has an azimuthal Fourier component proportional to

$$J_n(k_i R)e^{in\phi}, \tag{10}$$

where $k_i = k\sin\theta_i < k$. Thus if one wants to formulate a radiation condition, mathematically, it must be such that fields (9) are excluded but fields (10) are permitted.

This discussion suggests that the specification of a *mathematical* radiation condition for the present class of problems (plane-wave scattering by an infinite two-dimensional rough surface) will not be straightforward. However, the *physical* purpose of a radiation condition is clear: it is to exclude all ''incoming'' waves apart from the incident wave. We shall return to radiation conditions in Sec. II B.

B. Angular-spectrum representations

For any rough surface S, the scattered field in the half-space $z>0$ may be written using an angular-spectrum representation,

$$\begin{aligned} u(x,y,z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(\mu,\nu)e^{ik(\mu x+\nu y+mz)}\frac{d\mu\ d\nu}{m(\kappa)} \\ &= \int_0^{\pi/2}\int_{-\pi}^{\pi} A(\alpha,\beta)v(r,\theta,\phi;\alpha,\beta)d\alpha\ d\beta + \text{evanescent terms}. \end{aligned} \tag{11}$$

Here F is the *spectral amplitude*, $A(\alpha,\beta)=F(\sin\alpha\cos\beta,\ \sin\alpha\sin\beta)$, $\kappa=\sqrt{\mu^2+\nu^2}$, and

$$m(\kappa)=\begin{cases}\sqrt{1-\kappa^2}, & 0\leqslant\kappa<1,\\ i\sqrt{\kappa^2-1}, & \kappa>1,\end{cases}$$

the function v is defined by

$$v(r,\theta,\phi;\alpha,\beta)=\exp\{i\mathbf{k}\cdot\mathbf{x}\},\quad 0\leqslant\alpha\leqslant\frac{1}{2}\pi,\quad |\beta|\leqslant\pi, \tag{12}$$

where

$$\mathbf{k}=k\hat{\mathbf{k}}=k(\sin\alpha\ \cos\beta,\sin\alpha\ \sin\beta,\cos\alpha).$$

The integrals are superpositions of plane waves; they are propagating, homogeneous plane waves when $0\leqslant\kappa<1$, and they are evanescent, inhomogeneous plane waves when $\kappa>1$. In Eq. (11), we see the propagating plane waves explicitly: they propagate in the direction of $\hat{\mathbf{k}}$, with an (unknown) complex amplitude, $A(\alpha,\beta)$; the ''evanescent terms'' decay exponentially with z. For more information on angular-spectrum representations, see Clemmow[5] and DeSanto and Martin.[6]

In general, the spectral amplitude must be considered as a generalized function. Thus it is convenient to extract a continuous component from F, writing the scattered field as

$$u=u_{\rm pr}+u_{\rm ev}+u_{\rm con}, \tag{13}$$

where

$$u_{\rm pr}(r,\theta,\phi)=\sum_{n=0}^{N} A_n v(r,\theta,\phi;\alpha_n,\beta_n),$$

$$u_{\rm ev}(r,\theta,\phi)=\sum_{m=1}^{M} B_m w(r,\theta,\phi;\mu_m,\nu_m), \tag{14}$$

$$u_{\rm con}(x,y,z)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathscr{C}(\mu,\nu)\ e^{ik(\mu x+\nu y+mz)}\frac{d\mu\ d\nu}{m(\kappa)}, \tag{15}$$

$$w(r,\theta,\phi;\mu,\nu)=\exp\{ikr\sin\theta\,[\mu\cos\phi+\nu\sin\phi]-kr\cos\theta\,\sqrt{\kappa^2-1}\},\tag{16}$$

and $\kappa=\sqrt{\mu^2+\nu^2}>1$ in Eq. (16). The first term in Eq. (13) is a sum of propagating waves; the coefficients A_n and the angles, α_n and β_n, are unknown in general. The second term in Eq. (13) is a sum of evanescent waves; B_m, μ_m, and ν_m are unknown in general. The third term in Eq. (13) is a continuous spectrum of plane waves; the unknown function $\mathscr{C}$ is continuous. See Sec. V for further comments.

Let us now return to the radiation condition. Having chosen an origin O, arbitrarily, we consider a large hemisphere H_r, with radius r and center O. We then require that all propagating plane-wave components $v(r,\theta,\phi;\alpha_n,\beta_n)$ in u propagate *outwards* through H_r, away from O. This is almost built into the decomposition (13): we have to be careful with grazing waves [$\alpha_n=\frac{1}{2}\pi$; see the discussion following Eq. (8)]. A simple way to impose our radiation condition is to split the half-space $z>0$ and the hemisphere H_r into four parts. Thus with

$$H_r^m=\left\{(r,\theta,\phi):0\leqslant\theta\leqslant\frac{1}{2}\pi,\frac{1}{2}(m-3)\pi\leqslant\phi<\frac{1}{2}(m-2)\pi\right\},\quad m=1,2,3,4,$$

being the surfaces of four octants of a sphere, we require the following conditions for the regions specified:

$$\begin{aligned}
&m=1:\quad \text{in } x<0,\ y\leqslant 0,\quad \text{use } -\pi\leqslant\beta_n<-\tfrac{1}{2}\pi,\\
&m=2:\quad \text{in } x\geqslant 0,\ y<0,\quad \text{use } -\tfrac{1}{2}\pi\leqslant\beta_n<0,\\
&m=3:\quad \text{in } x>0,\ y\geqslant 0,\quad \text{use } 0\leqslant\beta_n<\tfrac{1}{2}\pi,\\
&m=4:\quad \text{in } x\leqslant 0,\ y>0,\quad \text{use } \tfrac{1}{2}\pi\leqslant\beta_n<\pi.
\end{aligned}\tag{17}$$

This partitioning makes it easy to ensure that only plane waves propagating out through H_r^m are included. This is the form of radiation condition used to derive boundary integral equations.

C. Boundary integral equations

One way to determine the scattered field is to derive a boundary integral equation over the rough surface S. An appropriate fundamental solution is the free-space Green's function

$$G(P,Q)=G(\mathbf{y},\mathbf{x})=\frac{-1}{2\pi}\,\frac{\exp\{ik|\mathbf{x}-\mathbf{y}|\}}{|\mathbf{x}-\mathbf{y}|},$$

where $\mathbf{x}$ and $\mathbf{y}$ are the position vectors of Q and P, respectively, with respect to the origin O. Apply Green's theorem to u and G in the region D_r with boundary $\partial D_r=H_r\cup S_r\cup T_r$, where H_r is a large hemisphere H_r of radius r and center O,

$$S_r=\{(x,y,z):z=s(x,y),\ 0\leqslant x^2+y^2<r^2\}$$

is a truncated rough surface, and

$$T_r=\{(x,y,z):x^2+y^2=r^2,\ s(x,y)\leqslant z\leqslant 0\}\tag{18}$$

is the surface of a truncated circular cylinder joining H_r and S_r. The result is

$$2u(P)=\int_{\partial D_r}\left\{u(q)\,\frac{\partial G}{\partial n_q}(P,q)-\frac{\partial u}{\partial n}\,G(P,q)\right\}dS_q,$$

where $P\in D_r$, $q\in\partial D_r$, and $\partial/\partial n_q$ denotes normal differentiation at q. Use of the boundary condition (6) yields

$$2u(P)=\int_{S_r}\left\{u(q)\frac{\partial G}{\partial n_q}(P,q)+\frac{\partial u_{\rm inc}}{\partial n}\,G(P,q)\right\}\,dS_q+I(u;H_r)+I(u;T_r), \tag{19}$$

where

$$I(u;\mathscr{S})=\int_{\mathscr{S}}\left\{u(q)\,\frac{\partial G}{\partial n_q}(P,q)-\frac{\partial u}{\partial n}\,G(P,q)\right\}\,dS_q$$

and normal differentiation is taken in a direction away from the origin [so that $\partial/\partial n=\partial/\partial r$ on H_r, consistent with Eq. (2)].

The next step is to estimate $I(u;H_r)$ and $I(u;T_r)$ for large r. Before estimating $I(u;H_r)$, using Eq. (13), we consider a single propagating plane-wave component in Eq. (13). Thus we shall evaluate $I(v;H_r)$ as $r\to\infty$, where v is defined by Eq. (12). In fact, as $v(r,\theta,\phi;\alpha,\beta)=v(r,\theta,\phi-\beta;\alpha,0)$, we can assume without loss of generality that $\beta=0$; we write

$$v(r,\theta,\phi;\alpha,0)=v(r,\theta,\phi;\alpha)=v(\mathbf{x};\alpha),\quad 0\leqslant\alpha\leqslant\frac{1}{2}\pi.$$

Indeed, we shall evaluate the limit using two different methods; these are the method of stationary phase (Sec. III) and an expansion method (Sec. IV). The reasons for this twofold evaluation are (i) the results are surprising, and (ii) the expansion method is natural but it is complicated and it leads to some subtle nonuniform behavior. We shall discuss the evaluation of $I(u;H_r)$ itself for large r in Sec. V. We then consider the contribution from $I(u;T_r)$ to Eq. (19) in Sec. VI. Finally, we will derive boundary integral equations from Eq. (19) in Sec. VII.

III. THE METHOD OF STATIONARY PHASE

We use the method of stationary phase[7] to estimate $I(v;H_r)$. It turns out that there are three cases, depending on the angle α.

We are interested in large values of $r=|\mathbf{x}|$, for fixed $\mathbf{y}$ and k. We have

$$G(P,q)\simeq(B/r)\,\exp\{ik(r-\mathbf{y}\cdot\hat{\mathbf{x}})\}, \tag{20}$$

where $B=-\frac{1}{2}/\pi$. Hence for large r,

$$v\,\frac{\partial G}{\partial r}-G\,\frac{\partial v}{\partial r}\simeq ik\frac{B}{r}\,(1-\hat{\mathbf{k}}\cdot\hat{\mathbf{x}})\,\exp\{ikr(1+\hat{\mathbf{k}}\cdot\hat{\mathbf{x}})\}\,\exp\{-ik\mathbf{y}\cdot\hat{\mathbf{x}}\}$$

and then

$$I(v;H_r)\simeq iB\,e^{ikr}\,L(kr),$$

where

$$L(\lambda)=\lambda\int_{\mathscr{D}}g(\theta,\phi)\,e^{i\lambda F(\theta,\phi)}\,d\theta\,d\phi, \tag{21}$$

$$g(\theta,\phi)=(1-\hat{\mathbf{k}}\cdot\hat{\mathbf{x}})\,\exp\{-ik\mathbf{y}\cdot\hat{\mathbf{x}}\}\,\sin\theta, \tag{22}$$

$$F(\theta,\phi)=\hat{\mathbf{k}}\cdot\hat{\mathbf{x}}\quad\text{and}\quad\mathscr{D}=\{(\theta,\phi):0\leqslant\theta\leqslant\theta_0<\pi,-\pi\leqslant\phi\leqslant\pi\}$$

is the rectangular domain of integration; this corresponds to integrating over a spherical cap subtending an angle of $2\theta_0$ at the origin, for the hemisphere H_r, we set $\theta_0=\frac{1}{2}\pi$.

We now examine three cases in turn. These are $\alpha=0$, $0<\alpha<\frac{1}{2}\pi$, and $\alpha=\frac{1}{2}\pi$.

TABLE I. Stationary-phase points for $L(\lambda)$.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
θ	0	0	π	π	α	$\pi-\alpha$	$\pi-\alpha$
ϕ	$\frac{1}{2}\pi$	$-\frac{1}{2}\pi$	$\frac{1}{2}\pi$	$-\frac{1}{2}\pi$	0	π	$-\pi$

A. The method of stationary phase: $\alpha=0$

Consider the special case $\alpha=0$, corresponding to a plane wave v propagating along the z-axis. Then we have $F(\theta,\phi)=\cos\theta$, which means that the only dependence on ϕ is through $\mathbf{y}\cdot\hat{\mathbf{x}}$:

$$L(\lambda)=\lambda\int_0^{\theta_0}(1-\cos\theta)\ b(\theta;\mathbf{y})\ e^{i\lambda\cos\theta}\sin\theta\ d\theta, \tag{23}$$

where

$$b(\theta;\mathbf{y})=\int_{-\pi}^{\pi}\exp\{-ik\mathbf{y}\cdot\hat{\mathbf{x}}\}\ d\phi.$$

Introducing spherical polar coordinates for $\mathbf{y}$,

$$\mathbf{y}=\rho\hat{\mathbf{y}}=\rho(\sin\Theta\ \cos\Phi,\sin\Theta\ \sin\Phi,\cos\Theta), \tag{24}$$

we can integrate over ϕ to obtain

$$b(\theta;\mathbf{y})=2\pi\ e^{-ik\rho\cos\theta\ \cos\Theta}\ J_0(k\rho\sin\theta\ \sin\Theta).$$

Now, the integral (23) can be estimated for large λ using the (one-dimensional) method of stationary phase. The only stationary-phase point is at $\theta=0$; as the integrand vanishes at $\theta=0$, we deduce that $L(\lambda)=O(1)$ as $\lambda\to\infty$. In fact, an integration by parts shows that

$$L(\lambda)\sim i(1-\cos\theta_0)\ b(\theta_0;\mathbf{y})\ e^{i\lambda\cos\theta_0}\quad\text{as }\lambda\to\infty,$$

in particular, for the hemisphere ($\theta_0=\frac{1}{2}\pi$), we obtain

$$I(v;H_r)=e^{ikr}\ J_0(k\rho\sin\Theta)+O((kr)^{-1/2})\quad\text{as } kr\to\infty,\quad\text{for } \alpha=0. \tag{25}$$

B. The method of stationary phase: $0<\alpha<\frac{1}{2}\pi$

Return to the integral $L(\lambda)$, defined by Eq. (21). We estimate $L(\lambda)$ for large λ, using the method of stationary phase for two-dimensional integrals.[8] Thus we look for stationary-phase points $\mathbf{c}=(\theta,\phi)\in\mathscr{D}$ at which grad $F=\mathbf{0}$; such points may be in the interior of $\mathscr{D}$ or on the boundary, $\partial\mathscr{D}$. Each $\mathbf{c}$ contributes a term to $L(\lambda)$ proportional to

$$g(\mathbf{c})\ e^{i\lambda F(\mathbf{c})},$$

the next term being $O(\lambda^{-1})$.[9] In general, $\partial\mathscr{D}$ contributes terms of $O(\lambda^{-1/2})$, whereas corners of $\partial\mathscr{D}$ contribute terms of $O(\lambda^{-1})$; all these contributions are smaller than those from stationary-phase points.

We have

$$F(\theta,\phi)=\sin\theta\ \cos\phi\ \sin\alpha+\cos\theta\ \cos\alpha.$$

Elementary calculations show that grad $F=\mathbf{0}$ at seven points $\mathbf{c}_j$ ($j=1,2,\ldots,7$) in the range $0\leqslant\theta\leqslant\pi$, $|\phi|\leqslant\pi$ (which is larger than $\mathscr{D}$); see Table I. Substituting into Eq. (22), we see that

$$g(\mathbf{c}_1)=g(\mathbf{c}_2)=g(\mathbf{c}_3)=g(\mathbf{c}_4)=g(\mathbf{c}_5)=0$$

and

$$g(\mathbf{c}_6)=g(\mathbf{c}_7)=2\sin\alpha\ \exp\{i\mathbf{k}\cdot\mathbf{y}\}.$$

So, if we integrate over a region $\mathscr{D}$ that does *not* include $\mathbf{c}_6$ and $\mathbf{c}_7$, then $L(\lambda)=o(1)$ as $\lambda\to\infty$. This will be the case if

$$\alpha<\theta_0<\pi-\alpha,$$

in particular, for the hemisphere, we obtain

$$I(v;H_r)=O((kr)^{-1/2}) \quad \text{as } kr\to\infty, \quad \text{for } 0<\alpha<\tfrac{1}{2}\pi.$$

C. The method of stationary phase: $\alpha=\frac{1}{2}\pi$

For the hemisphere, with $\alpha=\theta_0=\frac{1}{2}\pi$, we find that $\mathbf{c}_6$ and $\mathbf{c}_7$ are on $\partial\mathscr{D}$. Their contribution can be found;[10] the result is

$$I(v;H_r)=2\exp\{i\mathbf{k}\cdot\mathbf{y}\}+O((kr)^{-1/2}) \quad \text{as } kr\to\infty, \quad \text{for } \alpha=\tfrac{1}{2}\pi.$$

D. Summary

The axis of symmetry of the hemisphere H_r is the z-axis. We have considered plane waves v propagating out of the hemisphere, at an angle α to the z-axis. We have seen that $I(v;H_r)\to 0$ as $r\to\infty$, for $0<\alpha<\frac{1}{2}\pi$. For $\alpha=\frac{1}{2}\pi$ (''grazing waves,'' with respect to the plane $z=0$), $I(v;H_r)\to 2\exp\{i\mathbf{k}\cdot\mathbf{y}\}$, a finite quantity, as $r\to\infty$. For $\alpha=0$ (''normal waves,'' with respect to $z=0$), $I(v;H_r)\sim e^{ikr}\,J_0(k\rho\sin\Theta)$, which means that $I(v;H_r)$ does not have a limit (in this case) as $r\to\infty$. This unpleasant result can be verified directly when $\rho=0$; in this special case, we have

$$\begin{aligned}L(\lambda)&=2\pi\lambda\int_0^{\pi/2}(1-\cos\theta)\ e^{i\lambda\cos\theta}\sin\theta\ d\theta\\&=2\pi i\{1-i\lambda^{-1}(1-e^{i\lambda})\}\end{aligned}$$

exactly, after an integration by parts. In fact, in this special case, we can evaluate I exactly, without using the approximation (20); the result is

$$I(v;H_r)=e^{ikr}, \quad \rho=\alpha=0. \tag{26}$$

As in the two-dimensional case,[2] it is possible to derive a uniform approximation for α near $\frac{1}{2}\pi$.[11] The situation for small α is more complicated.[12] We shall not pursue these nonuniformities here.

IV. AN EXPANSION METHOD

We shall evaluate the integral over the hemisphere, $I(v;H_r)$, using suitable expansions of v and G in spherical polar coordinates. Thus

$$v(\mathbf{x};\alpha)=\exp\{i\mathbf{k}\cdot\mathbf{x}\}=\sum_{n=0}^{\infty}(2n+1)\ i^n\ j_n(kr)\ P_n(\cos\theta_1),$$

where $j_n(x)=(\frac{1}{2}\pi/x)^{1/2}\,J_{n+1/2}(x)$ is a spherical Bessel function,

$$\cos\theta_1=\hat{\mathbf{k}}\cdot\hat{\mathbf{x}}=\sin\theta\ \sin\alpha\ \cos\phi+\cos\theta\ \cos\alpha$$

and P_n is a Legendre polynomial. Similarly,

$$G(P,q)=G(\mathbf{y},\mathbf{x})=\frac{-ik}{2\pi}\sum_{n=0}^{\infty}(2n+1)\ j_n(k\rho)\ h_n(kr)\ P_n(\cos\theta_2)$$

for $r>\rho$, where

$$\cos\theta_2=\hat{\mathbf{y}}\cdot\hat{\mathbf{x}}=\sin\theta\ \sin\Theta\ \cos(\phi-\Phi)+\cos\theta\ \cos\Theta,$$

$h_n(x)\equiv h_n^{(1)}(x)=(\frac{1}{2}\pi/x)^{1/2}\ H_{n+1/2}^{(1)}(x)$ is a spherical Hankel function and $H_\nu^{(1)}$ is a Hankel function. Hence

$$I(v;H_r)=\sum_{l=0}^{\infty}\sum_{s=0}^{\infty}(2l+1)(2s+1)\ i^s\ j_l(k\rho)\ W_{ls}(kr)\ A_{ls},$$

where

$$W_{ls}(w)=-iw^2\{j_s(w)h_l'(w)-j_s'(w)h_l(w)\} \tag{27}$$

and

$$A_{ls}=\frac{1}{2\pi}\int_{-\pi}^{\pi}\int_0^{\pi/2}P_l(\cos\theta_2)\ P_s(\cos\theta_1)\ \sin\theta\ d\theta\ d\phi.$$

Note that W_{ll} is essentially a Wronskian, given by $W_{ll}=1$.

Let us evaluate A_{ls}. The addition theorem for Legendre polynomials gives

$$P_s(\cos\theta_1)=\sum_{m=0}^{s}\epsilon_m c_s^m P_s^m(\cos\theta)\ P_s^m(\cos\alpha)\ \cos m\phi,$$

$$P_l(\cos\theta_2)=\sum_{n=0}^{l}\epsilon_n c_l^n P_l^n(\cos\theta)\ P_l^n(\cos\Theta)\ \cos n(\phi-\Phi),$$

where P_n^m is an associated Legendre function, $\epsilon_0=1$, $\epsilon_m=2$ for $m\geqslant 1$, and

$$c_s^m=[(s-m)!]/[(s+m)!].$$

Hence, integrating over ϕ, we obtain

$$A_{ls}=\sum_{m=0}^{s}\epsilon_m c_l^m c_s^m P_l^m(\cos\Theta)\ P_s^m(\cos\alpha)\ B_{ls}^m\cos m\Phi, \tag{28}$$

where

$$B_{ls}^m=\int_0^{\pi/2}P_l^m(\cos\theta)\ P_s^m(\cos\theta)\ \sin\theta\ d\theta=\int_0^1 P_l^m(\mu)\ P_s^m(\mu)\ d\mu.$$

[Actually, the upper limit on the summation in Eq. (28) should be $\min\{l,s\}$, but this is of no consequence as $P_n^m\equiv 0$ for $m>n$.] B_{ls}^m has been evaluated by Hulme.[13] It turns out that

$$B_{2l+m,2s+m+1}^m=\frac{2^{2m}(-1)^{l+s+1}\Gamma(l+m+\frac{1}{2})\Gamma(s+m+\frac{3}{2})}{\pi(2l-2s-1)(l+s+m+1)\ l!\ s!}, \tag{29}$$

$$B_{2l+m+1,2s+m}^m=\frac{2^{2m}(-1)^{l+s+1}\Gamma(l+m+\frac{3}{2})\Gamma(s+m+\frac{1}{2})}{\pi(2s-2l-1)(l+s+m+1)\ l!\ s!}, \tag{30}$$

$B_{ll}^m=[(2l+1)\ c_l^m]^{-1}$ and $B_{ls}^m=0$ if $|l-s|$ is a positive even integer.

Now, returning to I, we see that it can be expressed in the form

$$I(v;H_r)=\sum_{l=0}^{\infty}\sum_{s=0}^{\infty}\sum_{m=0}^{s}\epsilon_m Q_{ls}^m\cos m\Phi,$$

where

$$Q_{ls}^m=(2l+1)(2s+1)\ i^s\ j_l(k\rho)\ W_{ls}(kr)\ c_l^m c_s^m\ B_{ls}^m\ P_l^m(\cos\Theta)\ P_s^m(\cos\alpha).$$

Interchanging the summations over s and m, we obtain

$$I(v;H_r)=\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{s=m}^{\infty}\epsilon_m Q_{ls}^m\cos m\Phi=\sum_{m=0}^{\infty}\epsilon_m\left\{\sum_{l=0}^{\infty}\sum_{s=m}^{\infty}Q_{ls}^m\right\}\cos m\Phi.$$

But $Q_{ls}^m\equiv 0$ for $m>l$, whence

$$I(v;H_r)=\sum_{m=0}^{\infty}\epsilon_m\Lambda_m\cos m\Phi,$$

where

$$\Lambda_m=\sum_{l=0}^{\infty}\sum_{s=0}^{\infty}Q_{l+m,s+m}^m.$$

This shows that I can be expressed as a Fourier series in the azimuthal coordinate of the observation point **y**, which is as expected.

A. A special case: $\rho=0$ ($\mathbf{y}=0$)

The expression for Λ_m is complicated. To gain some insight into its evaluation, we start with the special case of $\rho=0$. Then, only Q_{0s}^0 is not identically zero, whence

$$\begin{aligned}I(v;H_r)&=\sum_{s=0}^{\infty}Q_{0s}^0\\&=1+i\sum_{n=0}^{\infty}(4n+3)(-1)^n\ W_{0,2n+1}(kr)\ B_{0,2n+1}^0\ P_{2n+1}(\cos\alpha),\end{aligned}$$

where we have used $B_{00}^0=W_{00}=1$ and $B_{0,2n}^0=0$; also

$$B_{0,2n+1}^0=\int_0^1 P_{2n+1}(\mu)\ d\mu=\frac{(-1)^n\ \Gamma(n+\frac{1}{2})}{2\sqrt{\pi}\ (n+1)!}.$$

Given the definition (27), we define

$$S(\lambda;\alpha)=\sum_{n=0}^{\infty}(4n+3)(-1)^n\ B_{0,2n+1}^0\ j_{2n+1}(\lambda)\ P_{2n+1}(\cos\alpha)\tag{31}$$

so that

$$I(v;H_r)=1+\lambda\ e^{i\lambda}\{S+iS'+i\lambda^{-1}S\},\tag{32}$$

where we have simplified using $h_0(\lambda)=e^{i\lambda}/(i\lambda)$.

Proceeding formally, we substitute the known asymptotic approximation

$$j_m(\lambda)\sim\lambda^{-1}\sin\left(\lambda-\frac{1}{2}m\pi\right)\quad\text{as }\lambda\to\infty, \tag{33}$$

into Eq. (31), giving $S(\lambda;\alpha)\simeq -C(\alpha)\ \lambda^{-1}\cos\lambda$, where

$$C(\alpha)=\sum_{n=0}^{\infty}(4n+3)\ B^0_{0,2n+1}\ P_{2n+1}(\cos\alpha).$$

Hence, from Eq. (32),

$$I(v;H_r)\simeq 1-C(\alpha)\quad\text{as } kr\to\infty.$$

But $C(\alpha)$ is a known Fourier–Legendre expansion:

$$C(\alpha)=\begin{cases}1, & 0<\alpha<\frac{1}{2}\pi,\\ -1, & \frac{1}{2}\pi<\alpha<\pi,\\ 0, & \alpha=0,\ \frac{1}{2}\pi,\pi,\end{cases}$$

and is defined by periodicity for other values of α. Hence we deduce that

$$I(v;H_r)=o(1)\quad\text{as } kr\to\infty,\quad\text{for } \rho=0\ \text{ and }\ 0<\alpha<\tfrac{1}{2}\pi,$$

which is correct, but

$$I(v;H_r)=1+o(1)\quad\text{as } kr\to\infty,\quad\text{for } \rho=0\text{ and }\alpha=0,\tfrac{1}{2}\pi,$$

which is incorrect. This last result follows from $C(0)=0$, giving $S(\lambda;0)=o(\lambda^{-1})$ as $\lambda\to\infty$. In fact, we have

$$S(\lambda;0)=\sum_{n=0}^{\infty}\frac{(4n+3)\Gamma(n+\frac{1}{2})}{2\sqrt{\pi}\ (n+1)!}\ j_{2n+1}(\lambda)=\frac{1-\cos\lambda}{\lambda}, \tag{34}$$

exactly. This can be shown either by using a formula due to Gegenbauer,[14] namely

$$\left(\frac{1}{2}\lambda\right)^{\gamma}=\sum_{n=0}^{\infty}\frac{(2n+\gamma)\ \Gamma(n+\gamma)}{n!}\ J_{2n+\gamma}(\lambda)$$

with $\gamma=-\frac{1}{2}$, or by an application of the Mellin transform with respect to λ (the resulting series can be summed using the known formula for $F(a,b;c;1)$ where F is the Gauss hypergeometric function; the sum can be inverted using the Mellin convolution theorem).

The formal calculation above shows that we cannot expect to obtain the correct result if we simply replace the spherical Bessel functions by their large-argument asymptotic approximations, at least in the special cases of grazing ($\alpha=\frac{1}{2}\pi$) and normal ($\alpha=0$) plane waves. We examine the latter case next.

B. Another special case: $\alpha=0$

When $\alpha=0$, only those terms with $m=0$ contribute, giving

$$I(v;H_r)=\sum_{l=0}^{\infty}\sum_{s=0}^{\infty}Q^0_{ls}=\sum_{l=0}^{\infty}(2l+1)\ i^l\ b_l(kr)\ j_l(k\rho)\ P_l(\cos\Theta),$$

where

$$b_l(\lambda)=\sum_{s=0}^{\infty}(2s+1)\,i^{s-l}\,W_{ls}(\lambda)\,B^0_{ls}$$
$$=1+\lambda^2\{h_l'(\lambda)\,S_l(\lambda)-h_l(\lambda)\,S_l'(\lambda)\},\tag{35}$$

$$S_l(\lambda)=\sum_{\substack{s=0\\ s\neq l}}^{\infty}(2s+1)(-i)^{l-s+1}\,B^0_{ls}\,j_s(\lambda),$$

and we have used Eq. (27). It remains to evaluate $S_l(\lambda)$, and thence $b_l(\lambda)$ for large λ. Before doing this, let us consider the expected final answer, namely Eq. (25). From another formula due to Gegenbauer,[15] we have

$$\int_0^{\pi}J_0(w\sin\theta)\,P_n(\cos\theta)\,\sin\theta\,d\theta=\begin{cases}2\,(-1)^{n/2}\,P_n(0)\,j_n(w), & n \text{ even},\\ 0, & n \text{ odd},\end{cases}$$

whence

$$J_0(w\sin\theta)=\sum_{n=0}^{\infty}(4n+1)(-1)^n\,P_{2n}(0)\,j_{2n}(w)\,P_{2n}(\cos\theta).$$

Thus we are expecting to find that

$$b_{2l}(\lambda)=e^{i\lambda}\,P_{2l}(0)=e^{i\lambda}\,\frac{(-1)^l\,\Gamma(l+\frac{1}{2})}{\sqrt{\pi}\,l!}\quad\text{and}\quad b_{2l+1}=0.\tag{36}$$

Now, consider S_{2l+1}. From Eq. (30) with $m=0$, we have

$$S_{2l+1}(\lambda)=\frac{\Gamma(l+\frac{3}{2})}{2\pi\,l!}\sum_{n=0}^{\infty}\frac{(4n+1)\,\Gamma(n+\frac{1}{2})}{(n+l+1)(n-l-\frac{1}{2})n!}j_{2n}(\lambda).\tag{37}$$

Using

$$[\Gamma(n+a)]/[\Gamma(n+b)]\sim n^{a-b}\quad\text{as } n\to\infty,\tag{38}$$

we see that, for large n, the terms in the series (37) behave like $n^{-3/2}\,j_{2n}(\lambda)$. Given that $|J_\nu(\lambda)|\leqslant 1$ for all real λ and for all positive ν,[16] we deduce that the series (37) is absolutely and uniformly convergent. Thus we can replace $j_{2n}(\lambda)$ by its large-argument asymptotic approximation (33) to obtain

$$S_{2l+1}(\lambda)\sim\mathscr{A}_l\lambda^{-1}\sin\lambda\quad\text{as }\lambda\to\infty,$$

where

$$\mathscr{A}_l=\frac{\Gamma(l+\frac{3}{2})}{\pi\,l!}\sum_{n=0}^{\infty}\frac{(2n+\frac{1}{2})\,\Gamma(n+\frac{1}{2})}{(n+l+1)(n-l-\frac{1}{2})}\,\frac{(-1)^n}{n!}$$

is a real constant. Substituting into Eq. (35), using $h_n(\lambda)\sim\lambda^{-1}(-i)^{n+1}\,e^{i\lambda}$ as $\lambda\to\infty$, we find that

$$b_{2l+1}(\lambda)=1+(-1)^l\mathscr{A}_l+o(1)\quad\text{as }\lambda\to\infty.\tag{39}$$

Finally, we evaluate $\mathscr{A}_l$ using a contour-integral method. Splitting into partial fractions, the series can be written as

$$\sum_{n=0}^{\infty} \frac{\Gamma(n+\frac{1}{2})}{n+l+1} \frac{(-1)^n}{n!} + \sum_{n=0}^{\infty} \frac{\Gamma(n+\frac{1}{2})}{n-l-\frac{1}{2}} \frac{(-1)^n}{n!}. \tag{40}$$

Then, consider integrating the function $\Gamma(\frac{1}{2}-z)\ \Gamma(z)/(l-z+1)$ around a large circular contour in the complex z-plane. $\Gamma(z)$ has simple poles at $z=-n$, $n=0,1,2,\ldots$, with residue $(-1)^n/n!$. Thus the residues at the poles of $\Gamma(z)$ and $\Gamma(\frac{1}{2}-z)$ give rise to the first and second sums, respectively, in Eq. (40). The residue from the pole at $z=l+1$ is

$$-\Gamma(-\tfrac{1}{2}-l)\ \Gamma(l+1) = (l+\tfrac{1}{2})^{-1} l!\ \Gamma(\tfrac{1}{2}-l).$$

Combining Eq. (38) with the formula

$$\Gamma(\tfrac{1}{2}+z)\ \Gamma(\tfrac{1}{2}-z) = \pi \sec \pi z \tag{41}$$

shows that the contribution from integrating around the large circular contour vanishes as the circle expands to infinity. Hence the calculus of residues gives

$$\frac{\pi l!}{\Gamma(l+\frac{3}{2})} \mathscr{A}_l + \frac{l!\ \Gamma(\frac{1}{2}-l)}{l+\frac{1}{2}} = 0$$

whence $\mathscr{A}_l = (-1)^{l+1}$, using Eq. (41). Hence Eq. (39) gives $b_{2l+1}(\lambda) \to 0$ as $\lambda \to \infty$, as predicted by Eq. (36).

Next, consider S_{2l}. From Eq. (29) with $m=0$, we have

$$S_{2l}(\lambda) = \frac{\Gamma(l+\frac{1}{2})}{2\pi\ l!} \sum_{n=0}^{\infty} \frac{(4n+3)\ \Gamma(n+\frac{3}{2})}{(n+l+1)(n-l+\frac{1}{2})\ n!}\, j_{2n+1}(\lambda). \tag{42}$$

The terms in this series decay like $n^{-1/2}\, j_{2n+1}(\lambda)$ as $n \to \infty$, which is not fast enough to guarantee uniform convergence. However, we have

$$S_0(\lambda) = S(\lambda;0) = \lambda^{-1}(1-\cos\lambda)$$

from Eq. (34), so that

$$S_{2l}(\lambda) = \frac{\Gamma(l+\frac{1}{2})}{\sqrt{\pi}\ l!}\, S_0(\lambda) + T_{2l}(\lambda),$$

where

$$T_{2l}(\lambda) = \frac{\Gamma(l+\frac{1}{2})}{2\pi\ l!} \sum_{n=0}^{\infty} \frac{(4n+3)\ \Gamma(n+\frac{3}{2})}{n!} \left\{ \frac{1}{(n+l+1)(n-l+\frac{1}{2})} - \frac{1}{(n+1)(n+\frac{1}{2})} \right\} j_{2n+1}(\lambda),$$

this series is absolutely and uniformly convergent, whence $T_{2l}(\lambda) \sim \mathscr{B}_l \lambda^{-1} \cos\lambda$ as $\lambda \to \infty$, where

$$\mathscr{B}_l = \frac{\Gamma(l+\frac{1}{2})}{\pi\ l!} \sum_{n=0}^{\infty} \Gamma(n+\tfrac{3}{2}) \left\{ \frac{1}{n+1} + \frac{1}{n+\frac{1}{2}} - \frac{1}{n+l+1} - \frac{1}{n-l+\frac{1}{2}} \right\} \frac{(-1)^n}{n!}$$

is a real constant. Substituting back into Eq. (35), we find that

$$b_{2l}(\lambda) = \frac{(-1)^l\ \Gamma(l+\frac{1}{2})}{\sqrt{\pi}\ l!}\, e^{i\lambda} + 1 - \frac{(-1)^l\ \Gamma(l+\frac{1}{2})}{\sqrt{\pi}\ l!} + (-1)^l \mathscr{B}_l + o(1) \tag{43}$$

as $\lambda \to \infty$. The first term on the right-hand side is in accord with the prediction (36). The next three terms sum to zero, as expected. This can be shown by evaluating $\mathscr{B}_l$ using a contour-integral method (as for $\mathscr{A}_l$): use

$$\Gamma\left(\frac{3}{2}-z\right)\ \Gamma(z)\{(1-z)^{-1}-(l-z+1)^{-1}\}.$$

This completes the proof for $\alpha=0$ using an expansion method. It demonstrates that the expansion method is much more complicated to use (and more subtle, due to nonuniform convergence of one of the component series) than an approach based on the method of stationary phase.

V. ASYMPTOTIC BEHAVIOR OF $I(u;H_r)$

When a plane wave is reflected by a rough surface S, we can use the angular-spectrum representation (13) for the reflected field in $z>0$. Thus we have

$$I(u;H_r)=I(u_{\rm pr};H_r)+I(u_{\rm ev};H_r)+I(u_{\rm con};H_r).$$

For $I(u_{\rm ev};H_r)$, with $u_{\rm ev}$ defined by Eq. (14), we have

$$I(w;H_r)\simeq iB\lambda\ e^{i\lambda}\int_{\mathscr{D}} g(\theta,\phi)\ e^{i\lambda F(\theta,\phi)}\ d\theta\ d\phi,$$

where w is defined by Eq. (16), $\lambda=kr$, $B=-\frac{1}{2}/\pi$,

$$F(\theta,\phi)=[\mu\cos\phi+\nu\sin\phi]\sin\theta+i\cos\theta\ \sqrt{\kappa^2-1},$$

$$g(\theta,\phi)=[1-F(\theta,\phi)]\ \exp\{-ik\mathbf{y}\cdot\hat{\mathbf{x}}\}\ \sin\theta,$$

and $\mathscr{D}$ is defined below Eq. (22) with $\theta_0=\frac{1}{2}\pi$. As $\kappa>1$, grad $F\neq\mathbf{0}$ in $\mathscr{D}$, whence the vector

$$\mathbf{f}(\theta,\phi)=g(\text{grad } F)/|\text{grad } F|^2$$

is well defined. Then, use of the identity

$$ge^{i\lambda F}=(i\lambda)^{-1}\{\text{div }(e^{i\lambda F}\mathbf{f})-e^{i\lambda F}\text{div }\mathbf{f}\},$$

together with the divergence theorem in $\mathscr{D}$, shows that $I(w;H_r)=o(1)$ as $kr\to\infty$.[17] Hence from Eq. (14),

$$I(u_{\rm ev};H_r)\to 0 \quad \text{as } r\to\infty.$$

As $\mathscr{C}$ in Eq. (15) is continuous, we know that $u_{\rm con}$ satisfies the Sommerfeld radiation condition.[18] It follows that

$$I(u_{\rm con};H_r)\to 0 \quad \text{as } r\to\infty.$$

Finally, consider $I(u_{\rm pr};H_r)$. If $0\leqslant\alpha_n<\frac{1}{2}\pi$, the results of the previous sections are immediately applicable. For a typical grazing-wave component v, propagating at some angle β, our radiation condition implies that we consider $I(v;H_r^m)$, with m chosen according to Eq. (17). The stationary-phase analysis of Sec. III shows that $I(v;H_r^m)\to 0$ as $r\to\infty$. For the other three values of m, v will either be incoming through the corresponding H_r^m, and so its contribution must be discarded; or the contribution is negligible, using the method of stationary phase again.

Thus, in summary, let us extract the normal waves ($\alpha=0$) from $u_{\rm pr}$, and write

$$u_{\rm pr}(r,\theta,\phi)=A_0\ e^{ikr\cos\theta}+\text{other propagating waves with } 0<\alpha<\tfrac{1}{2}\pi, \tag{44}$$

where the coefficient A_0 is unknown. Then

$$I(u;H_r)\sim A_0\ e^{ikr}\ \mathscr{U}_0(P) \quad \text{as } r\to\infty, \tag{45}$$

where

$$\mathscr{U}_0(P)\equiv\mathscr{U}_0(\rho,\Theta,\Phi)=J_0(k\rho\sin\Theta). \tag{46}$$

VI. ASYMPTOTIC BEHAVIOR OF $I(u;T_r)$

The truncated cylindrical surface T_r is defined by Eq. (18). A point $q\in T_r$, with position vector $\mathbf{x}$, has cylindrical polar coordinates (r,ϕ,z). Then, for large r,

$$|\mathbf{x}-\mathbf{y}|\simeq r-\rho\sin\Theta\ \cos(\phi-\Phi),$$

where the fixed point P has position vector $\mathbf{y}$ and spherical polar coordinates (ρ,Θ,Φ) defined by Eq. (24). Hence as $\partial/\partial n=\partial/\partial r$ on T_r,

$$u\frac{\partial G}{\partial n}-G\frac{\partial u}{\partial n}\simeq\frac{B}{r}\ e^{ikr}\left(iku-\frac{\partial u}{\partial r}\right)\exp\{-ik\rho\sin\Theta\ \cos(\phi-\Phi)\},$$

where $B=-\frac{1}{2}/\pi$. Let

$$\mathscr{E}(r,\phi)=\int_s^0\left(\frac{\partial u}{\partial r}-iku\right)\ dz,$$

where the lower limit is $s(r\cos\phi,r\sin\phi)$. Then

$$I(u;T_r)\sim\frac{1}{2\pi}\ e^{ikr}\int_{-\pi}^{\pi}\mathscr{E}(r,\phi)\ \exp\{-ik\rho\sin\Theta\ \cos(\phi-\Phi)\}\ d\phi \quad \text{as } r\to\infty.$$

When is it true that $I(u;T_r)\to 0$ as $r\to\infty$? One sufficient condition is that

$$s\to 0 \ \text{ as } \ r\to\infty, \quad \text{for all } \phi,$$

this means that the rough surface approaches the flat plane $z=0$ at large distances, in all directions.

Another sufficient condition is that

$$\frac{\partial u}{\partial r}-iku=o(1) \ \text{ as } r\to\infty, \quad \text{for all } \phi \ \text{ and } \ s\leqslant z\leqslant 0,$$

so that u satisfies a form of radiation condition in directions parallel to $z=0$. In particular, this will be the case if u comprises outgoing cylindrical or spherical waves.

Next, consider a typical plane wave v, defined by Eq. (12); thus let

$$v=e^{ikr\sin\alpha\cos\phi}\ e^{ikz\cos\alpha} \ \text{ with } \ 0<\alpha\leqslant\tfrac{1}{2}\pi.$$

Then, a Fourier expansion in ϕ leads to Fourier components involving $J_n(kr\sin\alpha)$ and $J_n'(kr\sin\alpha)$. As $J_n(w)=O(w^{-1/2})$ as $w\to\infty$, we deduce that $I(v;T_r)\to 0$ as $r\to\infty$. This result includes grazing waves, but not normal waves for which $\alpha=0$. When $\alpha=0$, $v=e^{ikz}$ whence $\partial v/\partial r=0$ and the integration over z in $\mathscr{E}$ is trivial. The result is

$$\begin{aligned}I(v;T_r)&\sim\frac{1}{2\pi}\ e^{ikr}\int_{-\pi}^{\pi}(e^{iks}-1)\ \exp\{-ik\rho\sin\Theta\ \cos(\phi-\Phi)\}\ d\phi\\&=\frac{1}{2\pi}\ e^{ikr}\int_{-\pi}^{\pi}e^{iks}\ \exp\{-ik\rho\sin\Theta\ \cos(\phi-\Phi)\}\ d\phi-e^{ikr}\ J_0(k\rho\sin\Theta) \ \text{ as } r\to\infty.\end{aligned}$$

Thus, in summary, suppose that for $s\leqslant z\leqslant 0$ and large r,

$$u = A_0\, e^{ikz} + \tilde{u},$$

where the coefficient A_0 is the same as in Eq. (44) and $\tilde{u}$ is such that $I(\tilde{u};T_r)\to 0$ as $r\to\infty$. Then

$$I(u;T_r)\sim A_0\, e^{ikr}\{\mathscr{U}_s(r;P)-\mathscr{U}_0(P)\} \quad \text{as } r\to\infty, \tag{47}$$

where $\mathscr{U}_0$ is defined by Eq. (46) and

$$\mathscr{U}_s(r;P)\equiv\mathscr{U}_s(r;\rho,\Theta,\Phi)=\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{iks}\,\exp\{-ik\rho\sin\Theta\,\cos(\phi-\Phi)\}\,d\phi; \tag{48}$$

the dependence on r comes from the exponent $s=s(r\cos\phi, r\sin\phi)$.

VII. BOUNDARY INTEGRAL EQUATIONS

In Sec. II, we used Green's theorem to obtain the integral representation

$$2u(P)=\int_{S_r}\left\{u(q)\,\frac{\partial G}{\partial n_q}(P,q)+\frac{\partial u_{\rm inc}}{\partial n}\,G(P,q)\right\}\,dS_q+I(u;H_r)+I(u;T_r), \tag{49}$$

when $P\in D_r$, the region bounded by the hemisphere H_r, the truncated rough surface S_r, and the truncated circular cylinder T_r. Note that the left-hand side of this equation does not depend on r, so that the right-hand side of the equation must have a limit as $r\to\infty$. Before taking this limit, it is instructive to consider a very simple example.

A. An example

Consider a plane wave *normally* incident upon a *flat* surface at $z=-h$. Thus

$$u_{\rm inc}=e^{-ikz} \quad \text{and } u=A_0\, e^{ikz} \quad \text{with } A_0=e^{2ikh}.$$

This is the exact solution. Let us see how this solution is reconstructed by the representation (49). For simplicity, we take P at the origin; this will permit all the integrals to be evaluated exactly (without any asymptotic approximations). Thus, from Eq. (26), we have

$$I(u;H_r)=A_0\, e^{ikr}. \tag{50}$$

On S_r, we have $u=A_0\, e^{-ikh}$, $\partial u/\partial n=-iku$,

$$G=BR_\sigma^{-1}e^{ikR_\sigma},\quad \partial G/\partial n=hR_\sigma^{-2}(ik-R_\sigma^{-1})G$$

and $R_\sigma=\sqrt{\sigma^2+h^2}$, whence the integral over S_r is

$$\begin{aligned} I(u;S_r)&=2\pi A_0 B\, e^{-ikh}\int_0^r e^{ikR_\sigma}\{ik+hR_\sigma^{-1}(ik-R_\sigma^{-1})\}\frac{\sigma d\sigma}{R_\sigma}\\ &=-A_0\, e^{-ikh}\int_h^R e^{ikt}\{ik+ht^{-1}(ik-t^{-1})\}dt\\ &=-A_0\, e^{-ikh}\,[e^{ikt}+ht^{-1}\,e^{ikt}]_{t=h}^{R}\\ &=2A_0-A_0\left(1+\frac{h}{R}\right)e^{ik(R-h)}, \end{aligned} \tag{51}$$

where $R=\sqrt{r^2+h^2}$.

On T_r, we have $u=A_0\, e^{ikz}$ and $\partial u/\partial n=\partial u/\partial r=0$, whence

$$I(u;T_r)=2\pi A_0 Br\int_{-h}^{0} e^{ikz}\frac{\partial}{\partial r}\left(\frac{e^{ikR_z}}{R_z}\right)dz$$

$$=-A_0\,[(1-zR_z^{-1})\,e^{ik(R_z+z)}]_{z=-h}^{0}$$

$$=-A_0\,e^{ikr}+A_0\left(1+\frac{h}{R}\right)e^{ik(R-h)}, \tag{52}$$

where $R_z=\sqrt{r^2+z^2}$.

Adding Eqs. (50)–(52), we see that their sum is exactly $2A_0$, which is $2u(P)$ evaluated at the origin. Note that, as $r\to\infty$,

$$I(u;S_r)\sim 2A_0-A_0\,e^{ik(r-h)},$$

$$I(u;T_r)\sim -A_0\,e^{ikr}+A_0\,e^{ik(r-h)},$$

and $I(u;H_r)$ does not simplify further. Thus the boundary integral over the truncated rough surface does not have a limit as $r\to\infty$. Moreover, the integral over the truncated cylinder T_r does not have a limit as $r\to\infty$, and it is not negligible. This is a genuine three-dimensional effect, which is not seen in the two-dimensional case.[2]

B. Taking the limit: A new finite-part integral

Letting $r\to\infty$ in Eq. (49), we obtain

$$2u(P)=\oint_S\left\{u(q)\,\frac{\partial G}{\partial n_q}(P,q)+\frac{\partial u_{\rm inc}}{\partial n}\,G(P,q)\right\}dS_q,\quad P\in D_\infty, \tag{53}$$

where D_∞ is the unbounded region $z>s$,

$$\oint_S\cdots dS=\lim_{r\to\infty}\left\{\int_{S_r}\cdots dS+A_0\,e^{ikr}\,\mathscr{U}_s\right\} \tag{54}$$

and $\mathscr{U}_s$ is defined by Eq. (48). Note that the terms involving $\mathscr{U}_0$ in Eqs. (45) and (47) cancel.

The definition (54) is a nonstandard form of finite-part integral. It reduces to the standard definition of a principal-value integral at infinity if the coefficient A_0 vanishes; in other words, the standard definition is only appropriate if the scattered field does *not* include a discrete plane wave propagating up the z-axis. For a general rough surface, we do not know *a priori* whether $A_0=0$ or not.

We can express A_0 in terms of the values of $u(r,\theta,\phi)$ on $z=0$ ($\theta=\frac{1}{2}\pi$). Thus write

$$u\left(r,\frac{1}{2}\pi,\phi\right)=\sum_{n=-\infty}^{\infty}u_n(r)\,e^{in\phi}.$$

Hence, as a plane wave propagating along the z-axis is constant on planes of constant z, we have

$$A_0=\lim_{r\to\infty}u_0(r)=\frac{1}{2\pi}\lim_{r\to\infty}\int_{-\pi}^{\pi}u\left(r,\frac{1}{2}\pi,\phi\right)\,d\phi.$$

However, this formula is not useful, as we do not know u on $z=0$. What we would like is a similar formula, involving the boundary values of u on S.

Let us combine the identity

$$A_0=\frac{2}{r^2}\int_0^r u_0(\sigma)\;\sigma\,d\sigma+\frac{2}{r^2}\int_0^r[A_0-u_0(\sigma)]\sigma\,d\sigma$$

with an assumption that $u_0(r)=A_0+O(r^{-\delta})$ as $r\to\infty$, where $\delta>0$. Then, the second integral is $o(1)$ as $r\to\infty$, whence

$$A_0=\lim_{r\to\infty}\frac{2}{r^2}\int_0^r u_0(\sigma)\ \sigma\ d\sigma$$

$$=\frac{1}{\pi}\lim_{r\to\infty}\frac{1}{r^2}\int_{\mathscr{D}_r} u\left(\sigma,\frac{1}{2}\pi,\phi\right)\ \sigma\ d\sigma\ d\phi, \tag{55}$$

where $\mathscr{D}_r$ is the circular disk on $z=0$ with $x^2+y^2<r^2$. This is another formula for A_0 in terms of u on $z=0$. However, this formula involves u everywhere on $\mathscr{D}_r$, so we can relate it to an integral over S_r, using Green's theorem.

Consider the truncated circular cylinder V_r with axis along the z-axis and radius r, between $z=0$ and $z=s$. Thus one end of V_r is S_r and the other is the circular disk $\mathscr{D}_r$. Apply Green's theorem in V_r to u and $\sin kz$, giving

$$k\int_{\mathscr{D}_r} u\ dS+\int_{S_r}\left\{u\ \frac{\partial}{\partial n}(\sin kz)-\frac{\partial u}{\partial n}\sin kz\right\}\ dS-\int_{T_r}\frac{\partial u}{\partial n}\sin kz\ dS=0,$$

where T_r, defined by Eq. (18), is the curved part of V_r on which $x^2+y^2=r^2$. The third integral is $O(r)$ as $r\to\infty$, as it is bounded by $2\pi rM\max(s)$, where M is a bound on $|\partial u/\partial r|$. Hence, comparing with Eq. (55), and making use of Eq. (6), we see that

$$A_0=\lim_{r\to\infty}\mathscr{A}(r)\quad\text{where}\quad\mathscr{A}(r)=\frac{-1}{\pi kr^2}\int_{S_r}\left\{u\ \frac{\partial}{\partial n}(\sin kz)+\frac{\partial u_{\rm inc}}{\partial n}\sin kz\right\}\ dS, \tag{56}$$

which is a formula for A_0 in terms of the boundary values of u.

For the example in Sec. VII A, one can check that Eq. (56) does produce the correct value for A_0.

C. A boundary integral equation for the scattered field

Letting $P\to p\in S$ in Eq. (53) gives

$$u(p)=\oint_S\left\{u(q)\ \frac{\partial G}{\partial n_q}(p,q)+\frac{\partial u_{\rm inc}}{\partial n}\ G(p,q)\right\}\ dS_q,\quad p\in S. \tag{57}$$

This would be a boundary integral equation for u on S if we knew A_0. A_0 is defined by the formula (56), which requires $u(q)$ for $q\in S$; but we can only find $u(q)$ by solving Eq. (57).

Two possible ways to proceed are as follows. First, one could *assume* that $A_0=0$. In that case, the integral in Eq. (57) becomes an ordinary principal-value integral at infinity. However, the integral will diverge if it turns out that $A_0\neq 0$; see the example in Sec. VII A.

Second, from Eq. (56), we could replace the constant A_0 by $\mathscr{A}(r)$, where r is the radius of the truncated rough surface. This gives

$$u(p)=\int_{S_r}\left\{u(q)\ \frac{\partial\widetilde{G}}{\partial n_q}(r;p,q)+\frac{\partial u_{\rm inc}}{\partial n}\ \widetilde{G}(r;p,q)\right\}dS_q,\quad p\in S_r, \tag{58}$$

where

$$\widetilde{G}(r;P,Q)=G(P,Q)-(\pi kr^2)^{-1}\ e^{ikr}\ \mathscr{U}_s(r;P)\ \sin kz, \tag{59}$$

$Q=(x,y,z)$ and $\mathscr{U}_s$ is defined by Eq. (48). Thus we have a plausible boundary integral equation for u on S_r. Note that, in any numerical treatment, the rough surface would have to be truncated to S_r, so our ''plausible'' integral equation seems to be optimal in some sense. It remains to carry out numerical experiments.

VIII. CONCLUSION

In this paper we have considered the three-dimensional problem of the reflection of a plane wave by an infinite two-dimensional rough surface, defined by $z=s(x,y)$ with $-h\leqslant s\leqslant 0$. We have shown that the derivation of a boundary integral equation for this problem, akin to the Helmholtz integral equation for scattering by a bounded obstacle, is by no means straightforward. In particular, if the scattered field includes a plane wave propagating along the z-axis away from the rough surface (''normal waves''), then the usual Helmholtz integral equation is not valid: the boundary integral diverges. We have offered a modified integral equation which reduces to the standard Helmholtz integral equation when normal waves are absent.

The situation just described is unsatisfactory, even though the mathematical difficulty may be overcome. Indeed, this difficulty is due entirely to the unphysical problem posed at the outset: *plane-wave* reflection by an *infinite* rough surface. Clearly, we can realize neither a plane wave nor an infinite rough surface. Moreover, the mathematical difficulty disappears if we consider either point-source insonification or a finite patch of roughness on an otherwise flat surface.[2]

Several papers have been written in which a Helmholtz integral equation was used to provide ''exact'' or ''benchmark'' numerical solutions for plane-wave reflection by an infinite one-dimensional rough surface, the purpose being to validate various approximate theories.[19] We have shown here that the same integral equation cannot be used for two-dimensional surfaces, in general, unless one changes the problem, as suggested in the previous paragraph.

ACKNOWLEDGMENTS

This work was begun while P.A.M. was on sabbatical leave at the Colorado School of Mines. He is grateful to the Department of Mathematical and Computer Sciences for its hospitality, and acknowledges receipt of a Fulbright Scholarship Grant.

[1] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory* (Wiley, New York, 1983).
[2] J. A. DeSanto and P. A. Martin, ''On the derivation of boundary integral equations for scattering by an infinite one-dimensional rough surface,'' J. Acoust. Soc. Am. **102**, 67–77 (1997).
[3] Sec. I A of Ref. 2.
[4] G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, Cambridge, 1944), 2nd ed.
[5] P. C. Clemmow, *The Plane Wave Spectrum Representation of Electromagnetic Fields* (Pergamon, Oxford, 1966).
[6] J. A. DeSanto and P. A. Martin, ''On angular-spectrum representations for scattering by infinite rough surfaces,'' Wave Motion **24**, 421–433 (1996).
[7] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals* (Dover, New York, 1986).
[8] Sec. 8.4 of Ref. 7.
[9] Eqs. 42, 44, and 46 in Sec. 8.4 of Ref. 7.
[10] Eq. (8.4.46) of Ref. 7.
[11] Sec. 9.6 of Ref. 7.
[12] See Exercise 9.15(b) in Ref. 7.
[13] A. Hulme, ''The wave forces acting on a floating hemisphere undergoing forced periodic oscillations,'' J. Fluid Mech. **121**, 443–463 (1982).
[14] Sec. 5.2 of Ref. 4.
[15] Eq. 2 on p. 379 of Ref. 4, with $\nu=\frac{1}{2}$ and $\psi=\frac{1}{2}\pi$.
[16] See p. 406 of Ref. 4.
[17] See Sec. 8.2 of Ref. 7 for details.
[18] G. C. Sherman, J. J. Stamnes, and É. Lalor, ''Asymptotic approximations to angular-spectrum representations,'' J. Math. Phys. **17**, 760–776 (1976).
[19] See Ref. 2 for a discussion and references.

Integrable structure of the new Calogero models

V. Karimipour
Institute for Studies in Theoretical Physics and Mathematics, P.O. Box 19395-5531, Tehran, Iran
and Department of Physics, Sharif University of Technology, PO Box 11365-9161, Tehran, Iran

(Received 2 January 1997; accepted for publication 12 September 1997)

We show that the integrability of the dynamical system recently proposed by Calogero and characterized by the Hamiltonian $H=\Sigma_{j,k}^{N} p_j p_k\{\lambda+\mu\ \cos[\nu(q_j-q_k)]\}$ is due to a simple algebraic structure. It is shown that this system is canonically equivalent to the classical Heisenberg magnet with long-range interactions. Our method shows clearly how these types of systems can be generalized and provides the mechanism of integrability of a large class of similar systems. © *1998 American Institute of Physics.* [S0022-2488(98)03101-6]

I. INTRODUCTION

It has recently been shown[1] that the dynamical system characterized by the Hamiltonian

$$H=\sum_{j,k=1}^{N} p_j p_k\{\lambda+\mu\ \cos[\nu(q_j-q_k)]\} \tag{1}$$

and the standard Poisson brackets

$$\{q_i,q_j\}=\{p_i,p_j\}=0,\quad \{q_i,p_j\}=\delta_{ij}, \tag{2}$$

is completely integrable. Before going on we perform a canonical transformation $q\to q/\nu$, $p\to\nu p$, and set the parameter ν equal to unity. In Ref. 1 Calogero showed that this system has the following properties:

(1) The quantities

$$c_{jk}=p_j p_k\{1-\cos(q_j-q_k)\} \tag{3}$$

are constants of the motion.

(2) There are N independent constants of the motion in involution with each other and H is a function of these integrals. These are the total momentum $P=\Sigma_{i=1}^{N} p_i$ and

$$h_m=\sum_{j,k=1}^{m} c_{j,k},\quad m=2,3,...,N,$$

$$\{h_m,h_{m'}\}=\{h_m,P\}=0. \tag{4}$$

(3) The quantities C and S defined by

$$C=\sum_{j=1}^{N} p_j\cos q_j,\quad S=\sum_{j=1}^{N} p_j\sin q_j \tag{5}$$

are constants of the motion if $\lambda+\mu=0$ or if $P=0$, otherwise they evolve simply in time.

(4) The following extra relations hold:

$$\{C,P\}=-S,\quad \{S,P\}=C,\quad \{C,S\}=-P, \tag{6}$$

 © 1998 American Institute of Physics

$$\{C,H\}=-2(\lambda+\mu)PS, \quad \{S,H\}=2(\lambda+\mu)PC. \tag{7}$$

The last two relations verify the statement made in part (3) above.

(5) The initial value problem for this Hamiltonian system was solved in explicit form.

The starting point of Calogero and his main line of reasoning is to demand that a Hamiltonian of the general form

$$H=\sum_{j,k=1}^{N} p_j p_k f(q_j-q_k) \tag{8}$$

has constants of the motion of the following form:

$$c_{jk}=p_j p_k g(q_j-q_k), \tag{9}$$

where f and g are functions to be specified. By this requirement he arrives at a functional equation for f and g, one solution of which leads to the Hamiltonian (1) and the conserved quantities (3). However, the mutual Poisson brackets of these integrals of motion are complicated:[1]

$$\begin{aligned}\{c_{jk},c_{j'k'}\}=&\delta_{jj'}p_j p_k p_k'\{\sin(q_j-q_k)+\sin(q_k-q_k')+\sin(q_{k'}-q_j)\}\\&+(j\leftrightarrow k)+(j'\leftrightarrow k')+(j\leftrightarrow k,j'\leftrightarrow k').\end{aligned} \tag{10}$$

The clever guess of Calogero is that the quantities given in (4) are the required integrals of motion which are in involution, hence the integrability of the system. He also demonstrated that the equations of motion can be derived from a Lax pair.

After all these calculations one is tempted to ask the following natural questions: Is there any algebraic structure behind the integrability of this system? Is the integrability of this system related somehow to the existence of the classical Yang–Baxter matrix or to some bi-Hamiltonian structure? Can one construct more general systems?

It is the aim of this paper to answer the above questions. We will find that the integrability of these systems is due to a very simple algebraic and geometrical structure which is related to the long-range interactions and the factorizability of the Hamiltonian. This structure is completely different from the ones which are encountered in the study of systems with local interactions.

II. THE ALGEBRAIC STRUCTURE

Let us define the variables

$$x_j=p_j\cos q_j, \quad y_j=p_j\sin q_j, \quad z_j=ip_j. \tag{11}$$

In the following we also use the notation $x^1=x$, $x^2=y$, and $x^3=z$. The Hamiltonian can now be written as

$$H=\sum_{j,k=1}^{N} -\lambda z_j z_k+\mu(x_j x_k+y_j y_k), \tag{12}$$

where the new variables satisfy the following su(2) Poisson bracket relations:

$$\{x_i^a,x_j^b\}=i\epsilon^{abc}x_j^c\delta_{ij}. \tag{13}$$

Remark: We use the complex number i only for notational convenience in later manipulations. In fact, the Poisson bracket between the real dynamical variables x_i, y_i, and $z'_j = p_j$ is related to the su(1,1) algebra with the Casimir $C = x^2 + y^2 - z'^2$.
Now define the variables

$$X^a_m = x^a_1 + x^a_2 + x^a_3 + \cdots + x^a_m . \tag{14}$$

It is obvious that for each m these sets of variables satisfy the same relations among themselves as in (13) and form a copy of su(2) algebra, and, furthermore, since the smaller copies of the algebra are embedded in the larger copies, we have

$$\{X^a_i , X^b_j\} = i\epsilon^{abc} X^c_{(i,j)} , \tag{15}$$

where (i,j) is meant to denote the minimum of i and j, i.e.,

$$\{X^a_2 , X^b_2\} = i\epsilon^{abc} X^c_2 , \quad \{X^a_2 , X^b_3\} = i\epsilon^{abc} X^c_2 .$$

Defining for each copy, say the mth one, the Casimir function

$$C_m = \sum_{a=1}^{3} X^a_m X^a_m , \tag{16}$$

we obtain

$$\{C_i , X^b_j\} = 2i\epsilon^{abc} X^a_i X^c_{(i,j)} , \tag{17}$$

$$\{C_i , C_j\} = 4i\epsilon^{abc} X^a_i X^b_j X^c_{(i,j)} . \tag{18}$$

We now note that in the last formula the indices i and j are not dummy variables; however, the index (i,j) is either equal to i or to j. In any case, the tensor which is contracted with ϵ^{abc} is symmetric with respect to the interchange of two of the indices (a,c) or (b,c), hence the right-hand side identically vanishes:

$$\{C_i , C_j\} = 0.$$

It is interesting to note that although the Casimir of one copy does not commute with the generators of another copy as seen from (17), the Casimirs of different copies commute among themselves. The Hamiltonian can now be written as

$$H = -\lambda Z^2_N + \mu(X^2_N + Y^2_N) = -(\lambda + \mu) Z^2_N + \mu C_N . \tag{19}$$

Using (17), (14), and (15) the following relations can also be verified directly:

$$\{C_i , X^b_N\} = 0, \quad b = 1,2,3, \quad \{H, Z_N\} = 0, \tag{20}$$

$$\{X_N , H\} = 2i(\lambda + \mu) Z_N Y_N , \quad \{Y_N , H\} = -2i(\lambda + \mu) Z_N X_N . \tag{21}$$

We see very clearly the essence of integrability of the system and have been able to avoid the intermediate constants of the motion c_{jk} with their complicated Poisson brackets and directly reach the integrals of motion which are in involution.

Proposition: The Casimir functions C_i, $i = 2,3,...,N$, and $P \equiv -iZ_N$ are N involutive integrals of motion. Note that C_1 (even if it were not zero) is a constant which characterizes the phase space of the system and is not to be counted as an integral of the motion. The Casimir functions C_m modulo a minus sign are exactly equal to the integrals h_m defined in Ref. 1.

In fact we have

$$C_m=\sum_{a=1}^{3} X_m^a X_m^a=\sum_{a=1}^{3}\sum_{j,k=1}^{m} x_j^a x_k^a \tag{22}$$

$$=\sum_{j,k=1}^{m}(x_j x_k+y_j y_k+z_j z_k)=-\sum_{j,k=1}^{m} p_j p_k\{1-\cos(q_j-q_k)\}. \tag{23}$$

From (11) and (13) we readily find the algebraic meaning of all the quantities introduced in Ref. 1 and summarized in the Introduction:

$$Z_N=iP, \quad C_m=-h_m, \quad X_N=C, \quad Y_N=S. \tag{24}$$

All the algebraic and Poisson bracket relations between these quantities found in Ref. 1, follow from the above identification.

Remark: The Casimir functions are all functionally independent, since for any pair of integers m and n, with $m<n$, C_n depends on dynamical variables not present in C_m.

III. THE INITIAL VALUE PROBLEM

Integrability of a system does not always imply that the system can be solved in closed form. However, the system (1) has the additional virtue that its initial value problem is also solved in closed form.[1] The aim of this section is to show algebraically why this is possible.

In terms of the new variables one can easily write the equations of motions. From (17) we have

$$\{C_N, X_j^b\}=2i\epsilon^{abc}X_N^a X_j^c, \tag{25}$$

or

$$\{C_N, X_j\}=2i(Z_N Y_j-Y_N Z_j), \tag{26}$$

$$\{C_N, Y_j\}=2i(X_N Z_j-Z_N X_j), \tag{27}$$

$$\{C_N, X_j\}=2i(Y_N X_j-X_N Y_j), \tag{28}$$

from which we can obtain the equations of motion,

$$\begin{pmatrix} \frac{d}{dt}X_j \\ \frac{d}{dt}Y_j \\ \frac{d}{dt}Z_j \end{pmatrix}=2i\begin{pmatrix} 0 & \lambda Z_N & \mu Y_N \\ -\lambda Z_N & 0 & -\mu X_N \\ -\mu Y_N & \mu X_N & 0 \end{pmatrix}\begin{pmatrix} X_j \\ Y_j \\ Z_j \end{pmatrix}, \tag{29}$$

which is of the form

$$\frac{d}{dt}\Psi_j=A\Psi_j. \tag{30}$$

The time dependence of the matrix A is itself determined from

$$\frac{d}{dt}X_N=\{X_N,H\}=2i(\lambda+\mu)Z_NY_N, \tag{31}$$

$$\frac{d}{dt}Y_N=\{Y_N,H\}=-2i(\lambda+\mu)Z_NX_N, \tag{32}$$

$$\frac{d}{dt}Z_N=\{Z_N,H\}=0, \tag{33}$$

or simply

$$\frac{d}{dt}\mathbf{S}_N=i\omega J_3\mathbf{S}_N, \tag{34}$$

where we have set $\mathbf{S}_N=(X_N,Y_N,Z_N)$ and J_1, J_2, and J_3 are generators of the rotation group: $(J_i)_{kl}=-i\epsilon_{ikl}$. In any case Z_N is constant and the dynamics of X_N and Y_N is given by

$$X_N(t)=X_N(0)\cos\,h\omega t+iY_N(0)\sin\,h\omega t,\quad Y_N(t)=Y_N(0)\cos\,h\omega t-iX_N(0)\sin\,h\omega t,$$

where $\omega=2(\lambda+\mu)Z_N$.

Note that from the above equations $X_N(t)^2+Y_N(t)^2=X_N(0)^2+Y_N(0)^2$. Once the time dependence of A is determined, the time dependence of Ψ_j will be determined from (30):

$$\Psi_j(t)=T\,\exp\int_0^t A(t')dt'\;\Psi_j(0). \tag{35}$$

However, one can go beyond this formal solution. We note that A can be written as follows:

$$A=-\omega J_3+2\mu(\mathbf{S}_N\cdot\mathbf{J}). \tag{36}$$

Knowing the dynamics of $\mathbf{S}_N$ from (34), we have

$$\mathbf{S}_N(t)\cdot\mathbf{J}=U(t)^{-1}\mathbf{S}_N(0)\cdot\mathbf{J}U(t), \tag{37}$$

where

$$U(t)=e^{\omega tJ_3}. \tag{38}$$

Therefore the dynamics of A is just a simple rotation around the z axis:

$$A(t)=e^{-\omega tJ_3}A(0)e^{\omega tJ_3}\equiv U^{-1}(t)A(0)U(t). \tag{39}$$

Multiplying both sides of this equation by $U(t)$ and defining $\Phi_i(t)=U(t)\Psi_i(t)$ we find

$$\frac{d}{dt}\Phi_i(t)=(\omega J_3+A(0))\Phi_i(t)=(2\mu\mathbf{S}_N(0)\cdot\mathbf{J})\Phi_i(t). \tag{40}$$

This is a simple evolution equation governed by a constant matrix the solution of which is simply

$$\Phi_i(t)=e^{2\mu\mathbf{S}_N(0)\cdot\mathbf{J}t}\Phi_i(0) \tag{41}$$

or

$$\Psi_i(t)=e^{-\omega tJ_3}e^{2\mu\mathbf{S}_N(0)\cdot\mathbf{J}t}\Psi_i(0). \tag{42}$$

Note that from Eq. (14) the vectors

$$\psi_j = \begin{pmatrix} x_j \\ y_j \\ z_j \end{pmatrix} = \Psi_j - \Psi_{j-1}$$

have the same equation of motion as above. We have thus arrived at a closed form solution of the initial value problem for Ψ_i and thus for ψ_i. Our analysis shows clearly why the initial value problem in Ref. 1 can be solved in closed form. Compared with the solution of Ref. 1, our solution has the merit that the initial data $\psi_j(0)$ appear explicitly in the solution, and no intermediate parameters (e.g., the parameters t_j in eqs. 2.19 and 2.20 of Ref. 1) are involved. Obtaining the evolution of the original variables q_i and p_i is now a matter of a change of coordinates, which I leave for the interested reader. I personally think, when viewed as a system of interacting classical spins (see example 2 in Sec. V), this solution is more interesting physically.

IV. GENERALIZATIONS

We can now generalize our construction and include the models of Calogero as a special case. Let g be a simple Lie algebra of ank r with generators e^a, $a=1,2,...,\dim g$, and relations

$$[e^a,\ e^b] = C^{ab}{}_c e^c. \tag{43}$$

Also, let g^* be its dual with basis e_a. It is well known[2] that the Lie structure on g induces a Poisson structure on $C^\infty(g^*)$,

$$\{x^a, x^b\} = C^{ab}{}_c x^c, \tag{44}$$

where the x^a, $a=1,2,...,\dim g = \dim g^*$ are the local coordinates in g^*.

In general the Poisson bracket is degenerate. To make it nondegenerate one restricts it to the submanifolds of $C^\infty(g^*)$ naturally defined by setting the values of the Casimir functions equal to constants. These submanifolds (symplectic leaves) are always even dimensional and the Poisson bracket becomes nondegenerate on them. By Darboux theorem one can then define local canonical coordinates and momenta on these submanifolds [the analogue of Eq. (11) above].

One can now define the variables X^a_m exactly as in (14). All the formalism of Sec. II can be followed exactly except that there are r Casimir functions involved labeled by

$$C^\alpha_m, \quad \alpha = 1,2,...,r, \quad m = 2,3,...,N,$$

with

$$\{C^\alpha_m, C^\beta_n\} = 0. \tag{45}$$

Furthermore one has

$$\{C^\alpha_m, h^\beta_N\} = 0, \tag{46}$$

where h^β_N correspond to the Cartan subalgebra elements of the Nth copy of g. Obviously any Hamiltonian of the general form

$$H = H(C^1_2, \ldots, C^r_N, h^1_N, \ldots, h^r_N) \tag{47}$$

defines an integrable system which is a generalization of the one introduced in Ref. 1.

The simplest of such Hamiltonians reads

$$H=(\lambda+\mu)\cdot h_N+\mu C_N. \tag{48}$$

Here

$$\lambda\cdot h_N\equiv\sum_\alpha \lambda_\alpha h_N^\alpha$$

and

$$C_N=\sum_{a,b} g_{ab}X_N^a X_N^b$$

with g_{ab} being the Cartan metric of the Lie algebra. The dynamical equations of motion for this system can be solved exactly along the same lines outlines in Sec. III.

Example 1: Consider again the algebra su(1,1) with relations

$$\{x,y\}=-z',\quad \{y,z'\}=x,\quad \{z',x\}=y. \tag{49}$$

The symplectic leaves are defined by

$$x^2+y^2-z'^2=c,$$

where c is a constant. These submanifolds are of two completely different geometries. For $c=0$ they are double cones with a singularity at the apex. This is the leaf chosen by Calogero with the canonical coordinates (11). The other leaves where $c\neq 0$ are hyperboloids with the following canonical coordinates:

$$z'=p,\quad x=\sqrt{p^2+c}\,\cos q,\quad y=\sqrt{p^2+c}\,\sin q. \tag{50}$$

Hence a generalization of the model of Ref. 1 is defined by

$$H=\sum_{j,k=1}^{N}\lambda p_j p_k+\mu\{\sqrt{p_j^2+c}\sqrt{p_k^2+c}\,\cos[(q_j-q_k)]\} \tag{51}$$

with the integrals of motion given by

$$h_m=\sum_{j,k=1}^{m} p_j p_k-\{\sqrt{p_j^2+c}\sqrt{p_k^2+c}\,\cos[(q_j-q_k)]\}$$

and

$$P=\sum_{i=1}^{N} p_i.$$

Example 2: Classical Heisenberg magnet with long-range interactions. We replace the su(1,1) algebra with su(2). The Casimirs are now equal to $C=x^2+y^2+z^2$ and the symplectic leaves are two-dimensional spheres. This is the classical Heisenberg Magnet describing a system of classical spins interacting with each other.

All the results we have obtained are readily adopted to this case simply by the following replacement: $\mathbf{x}_i^a\to i\mathbf{S}_i^a$, $a=1,2,3$.

The Hamiltonian (12) now turns into

$$H=\sum_{j,k=1}^{N}\lambda\mathbf{S}_j^z\mathbf{S}_k^z-\mu(\mathbf{S}_j^x\mathbf{S}_k^x+\mathbf{S}_j^y\mathbf{S}_k^y). \tag{52}$$

For $\lambda+\mu=0$ we have the isotropic magnet with su(2) symmetry.

The physical interpretation of the solution [Eqs. (34) and (42)] is now clear. In the isotropic case where $\omega=0$ the total magnetization $\mathbf{S}_N$ is constant and all the other spins simply rotate around the total spin with equal frequency $2\mu|\mathbf{S}_N|$. In the unisotropic case $\omega\neq 0$ the total spin rotates around the z axis with frequency ω and all the other spins rotate around the instantenous position of the total spin with the same frequency as in the isotropic case.

Example 3: Classical Heisenberg magnet with more general interactions. One can also define the Hamiltonian to be

$$H=\lambda_1 C_1+\lambda_2 C_2+\cdots,\ \lambda_N C_N+BZ_N, \tag{53}$$

which can be rewritten as

$$H=\sum_{i,j}\ \eta_{ij}\mathbf{S}_i\cdot\mathbf{S}_j+B\sum_i\ \mathbf{S}_i^z, \tag{54}$$

where

$$\eta_{ii}=\sum_{r=i}^{N}\lambda_r, \quad \eta_{ij}=2\sum_{r=\max(i,j)}^{N}\lambda_r.$$

Here the coupling parameters between different spins are not equal and with a suitable choice of the parameters λ_i's may represent a system of more physical interest than the Hamiltonian (53). The last term in the Hamiltonian represents the interaction of the spins with the magnetic field.

Remark: In Refs. 3 and 4, systems of very similar form to (1), where the momenta p_j come under the square root, have been shown to be exactly solvable. The integrable structure of these models is completely different[5] and should not be confused with that of the present model.

[1] F. Calogero, Phys. Lett. A **201**, 306–310 (1995).
[2] L. D. Faddeev, *Integrable Models in* $1+1$ *Dimensional Quantum Fields Theory*, Les Houches Lectures 1982 (Elsevier, Amsterdam, 1984).
[3] F. Calogero, J. Math. Phys. **36**, 4832 (1995).
[4] F. Calogero and J. F. van Diejen, Phys. Lett. A **205**, 143 (1995).
[5] V. Karimipour, ''A solvable Hamiltonian system, integrability and action-angle variables,'' J. Math. Phys. **38**, 1577–1582 (1997).

Quantum critical fluctuations in a ferroelectric model: Quasi-average approach

Patrice Joussot[a] and Valentin A. Zagrebnov[b]
Centre de Physique Théorique, CNRS-Luminy, Case 907, 13288 Marseille cedex 09, France

(Received 30 October 1996; accepted for publication 10 October 1997)

We show that critical fluctuations in a quantum ferroelectric model depend not only on the range of harmonic interactions but also on infinitesimal perturbations breaking symmetry of the model. This means that algebra of fluctuation operators is degenerated for (unique) critical equilibrium state. © *1998 American Institute of Physics.* [S0022-2488(98)00502-7]

I. INTRODUCTION

The idea to construct *pure equilibrium states* (phases) by infinitesimal perturbations of Hamiltonians comes back to the Bogoliubov ''quasi-averages.'' [1] It was generalized later to include *mixed states*.[2] For a lattice system defined by a Hamiltonian H_Λ (Λ is a finite subset of d-dimensional hypercubic lattice $\mathbb{Z}^d$ of the volume $V=|\Lambda|$), it can be formulated as follows:

(i) Let $\{\tau_l(-)\}_{l\in\mathbb{Z}^d}$ be space-translation automorphisms and $\{B_l=\tau_l(B)\}_{l\in\mathbb{Z}^d}$, where B is a local self-adjoint operator in the site $l=0$. Define a perturbed Hamiltonian

$$H_\Lambda(h)\equiv H_\Lambda-\sum_{l\in\Lambda} hB_l, \quad h\in\mathbb{R}^1. \tag{1.1}$$

If for $h\neq 0$ the perturbation breaks a symmetry G of H_Λ, then the limiting state for $\Lambda\to\mathbb{Z}^d$ (the Bogoliubov's quasi-average),

$$\langle-\rangle^*=\lim_{h\to 0}\lim_{\Lambda}\langle-\rangle_{\Lambda,h}, \tag{1.2}$$

picks out pure states for which the symmetry G can be broken at low temperatures (see, e.g., Ref. 3, Ch. 6).

(ii) If the external field has the form

$$h=\hat{h}/V^\alpha, \tag{1.3}$$

then an obvious generalization of (1.2) either coincides with pure states (for $\alpha<\alpha_c$) or gives a family of mixed states enumerated by $\hat{h}$ and $\alpha\geqslant\alpha_c$ (Ref. 2).

As is proven in Refs. 4 and 5 for the case of a quantum model of ferroelectrics (structural phase transitions), the algebra of quantum fluctuations depends on the parameter α *in pures states (1.2) with broken symmetry*.

The aim of the present paper is to scrutinize this algebra on the critical line (i.e., for a *unique* equilibrium state) by explicit calculation of its dependence on perturbations (1.3). As far as the corresponding critical equilibrium state exhibits long-range correlations, one might to assume that *critical fluctuations* have to be sensitive with respect to these kind of perturbations. On the other hand, they also have to be sensitive to the decay of direct interaction between particles: in the model[4–6] it corresponds to the decay of harmonic-force matrix elements $\{\phi_{l,l'}\}_{l,l'\in\mathbb{Z}^d}$ given by

[a]DEA Sciences des Matériaux, Université de la Méditerranée (Aix-Marseille II), Centre de Recherche sur les Mécanismes de la Croissance Cristalline. Electronic mail: joussot@cpt.univ-mrs.fr
[b]Centre de Physique Théorique and Département de Physique, Université de la Méditerranée (Aix-Marseille II). Electronic mail: zagrebnov@cpt.univ-mrs.fr

 © 1998 American Institute of Physics

$$\phi_{l,l'} \sim |l-l'|^{-(d+\sigma)} \quad \text{for } |l-l'| \to \infty. \tag{1.4}$$

If $\sigma \geqslant 2$, one classifies (1.4) as a *short-range* interaction while the case $0<\sigma<2$ corresponds to *long-range* interactions because the corresponding lattice Fourier transforms of $\phi_{l,l'}$ have the following asymptotics for $q \to 0$:

$$\widetilde{\phi}(q) \sim \begin{cases} s^{\sigma} q^{\sigma} + o(q^{\sigma}), & 0<\sigma<2, \\ s^2 q^2 + o(q^2), & \sigma \geqslant 2, \end{cases} \tag{1.5}$$

with a constant $s>0$.

Let A be a local operator. Then the corresponding fluctuation operator for a limiting translation-invariant state $\langle - \rangle$ is defined as the limit of the finite-volume fluctuation operator:

$$F^{\delta_A}(A) = \lim_{\Lambda} F_{\Lambda}{}^{\delta_A}(A) = \lim_{\Lambda} \frac{1}{V^{1/2+\delta_A}} \sum_{l \in \Lambda} (\tau_l(A) - \langle A_l \rangle_{\Lambda,h}),$$

where the exponent δ_A is defined via the noncommutative central limit theorem and corresponds to the condition that the variance $\lim_{\lambda} \langle (F_{\Lambda}{}^{\delta_A}(A))^2 \rangle$ is nontrivial.[4–6] In fact, dependence of the fluctuation operator exponent δ_A on σ is studied on the critical line for $h \equiv 0$ in Ref. 6. Therefore, our aim is to find dependence of exponent δ_A on the parameter α in the presence of the long-range interaction (1.4). Note that $\delta_A = \delta_A(\alpha, \sigma)$ is directly related to the critical exponent η describing decay of the two-point correlation function on the critical line: $\eta = 2 - 2d\,\delta_A$.

The paper is organized as follows. In the next section (using notations from Refs. 4–6) we briefly recall our anisotropic quantum ferroelectric model. Detailed analysis of fluctuations on the critical line as well as the main result, calculations of exponents $\delta_Q(\alpha, \sigma)$ and $\delta_P(\alpha, \sigma)$ for displacement and momentum fluctuations, are collected in Sec. III. Concluding remarks are left to Sec. IV.

II. MODEL AND PHASE TRANSITION

Let $\mathbb{Z}^d$ be a d-dimensional hypercubic lattice. With each lattice site $l \in \mathbb{Z}^d$ occupied by a particle of mass m, we associate position and momentum operators $Q_l \in \mathbb{R}^1$ and $P_l = (\hbar/i) \times (\partial/\partial Q_l)$ corresponding to anisotropic one-dimensional particle displacements from this site. Let Λ be a finite rectangular subset of $\mathbb{Z}^d$ with periodic boundary conditions and Λ^* be the lattice dual to Λ. The local Hamiltonian of the model[4–6] is given by

$$H_{\Lambda}(h) = \sum_{l \in \Lambda} \frac{P_l^2}{2m} + \frac{1}{4} \sum_{l,l' \in \Lambda} \phi_{l,l'} (Q_l - Q_{l'})^2 + \frac{a}{2} \sum_{l \in \Lambda} Q_l^2 + VW\left(\frac{1}{V} \sum_{l \in \Lambda} Q_l^2 \right) - h \sum_{l \in \Lambda} Q_l. \tag{2.1}$$

The second term in (2.1) represents a harmonic potential between particles, whereas the last term corresponds to external field breaking the $\{Q_l \to -Q_l\}_{l \in \mathbb{Z}^d}$ symmetry. The fourth term in (2.1) represents a one-site lattice anharmonic potential in the *spherical approximation*[4] known as the concept of *self-consistent phonons*.[7] The model (2.1) can be solved by the approximation Hamiltonian method.[4] The corresponding approximating Hamiltonian has the form

$$H_{\Lambda}(c,h) = \sum_{l \in \Lambda} \frac{P_l^2}{2m} + \frac{1}{4} \sum_{l,l' \in \Lambda} \phi_{l,l'} (Q_l - Q_{l'})^2 + \frac{a}{2} \sum_{l \in \Lambda} Q_l^2$$

$$+ W'(c) \sum_{l \in \Lambda} Q_l^2 - h \sum_{l \in \Lambda} Q_l + V(W(c) - cW'(c)).$$

Then the free-energy density of the system (2.1) is equal to

$$f(T,h) = \lim_{\Lambda} f_{\Lambda}[H_{\Lambda}(h)] = \lim_{\Lambda} \sup_{c} f_{\Lambda}[H_{\Lambda}(c,h)],$$

where

$$\sup_c f_\Lambda[H_\Lambda(c,h)] = f_\Lambda[H_\Lambda(c_{\Lambda,h}(T,\lambda))]$$

$$= \frac{1}{\beta V}\sum_{q\in\Lambda^*} \ln\left[2\sinh\frac{\beta\lambda\Omega_q(c_{\Lambda,h}(T,\lambda))}{2}\right] - \frac{1}{2}\frac{h^2}{\Delta(c_{\Lambda,h}(T,\lambda))}$$

$$+ [W(c_{\Lambda,h}(T,\lambda)) - c_{\Lambda,h}(T,\lambda)W'(c_{\Lambda,h}(T,\lambda))]. \tag{2.2}$$

Here $c_{\Lambda,h}(T,\lambda)>0$ is a solution of the spherical self-consistency equation

$$c = \frac{h^2}{\Delta^2(c)} + \frac{1}{V}\sum_{q\in\Lambda^*}\frac{\lambda}{2\Omega_q(c)}\coth\frac{\beta\lambda}{2}\,\Omega_q(c), \tag{2.3}$$

with

$$\Omega_q^2(c) = \Delta(c) + \omega_q^2, \tag{2.4}$$

$$\Delta(c) = a + 2W'(c), \tag{2.5}$$

$$\omega_q^2 = \tilde{\phi}(0) - \tilde{\phi}(q), \tag{2.6}$$

$$\tilde{\phi}(q) = \sum_{l\in\Lambda}\phi_{l,0}\exp(-iql), \tag{2.7}$$

$$\lambda = \frac{\hbar}{\sqrt{m}}, \tag{2.8}$$

$$\beta = (kT)^{-1}, \tag{2.9}$$

where $\{\Omega_q(c)\}_{q\in\Lambda^*}$ are frequencies of the approximating harmonic Hamiltonian $H_\Lambda(c,h)$, $\Delta(c)$ is a gap in the spectrum of (self-consistent) phonons, and λ is the quantum parameter of the model.

Let $a>0$ and $W:\mathbb{R}^1_+\to\mathbb{R}^1_+$ be a monotonous decreasing function with $W''(c)\geqslant w>0$ and $\Delta(c=0)<0$. Then by (2.5) and (2.6) one gets that stability domain $D=[c^*,\infty)$ of the approximating Hamiltonian $H_\Lambda(c,h)$ corresponds to

$$c^* = \inf\{c: c\geqslant 0, \Delta(c)\geqslant 0\} > 0, \tag{2.10}$$

and $\Delta(c^*)=0$.

To study the phase diagram of the model (2.1) we have to pass in (2.3) to the thermodynamic limit $\Lambda\to\mathbb{Z}^d$. Then limiting solutions satisfy

$$c_h(T,\lambda) = \rho_h(T,\lambda) + I_d(c_h(T,\lambda),T,\lambda), \tag{2.11}$$

where

$$\rho_h(T,\lambda) = \lim_\Lambda\left\{\frac{1}{V}\frac{\lambda}{2\sqrt{\Delta(c_{\Lambda,h}(T,\lambda))}}\coth\frac{\beta\lambda}{2}\sqrt{\Delta(c_{\Lambda,h}(T,\lambda))} + \frac{h^2}{\Delta^2(c_{\Lambda,h}(T,\lambda))}\right\} \tag{2.12}$$

and

$$I_d(c,T,\lambda) = \frac{\lambda}{(2\pi)^d}\int_{q\in B_d} d^dq\,\frac{1}{2\Omega_q}\coth\frac{\beta\lambda}{2}\,\Omega_q(c).$$

Here, $B_d=\{q\in\mathbb{R}^d; |q|\leqslant\pi\}$ is the first Brillouin zone.

From (2.11) one gets that for $h=0$ and $T=0$, there is λ_c such that $c^*\leqslant I_d(c^*,0,\lambda)$ for $\lambda\geqslant\lambda_c$. Then the line of critical temperatures $T_c=T_c(\lambda)$ which separates the phase diagram (see Fig. 1) is defined by the condition

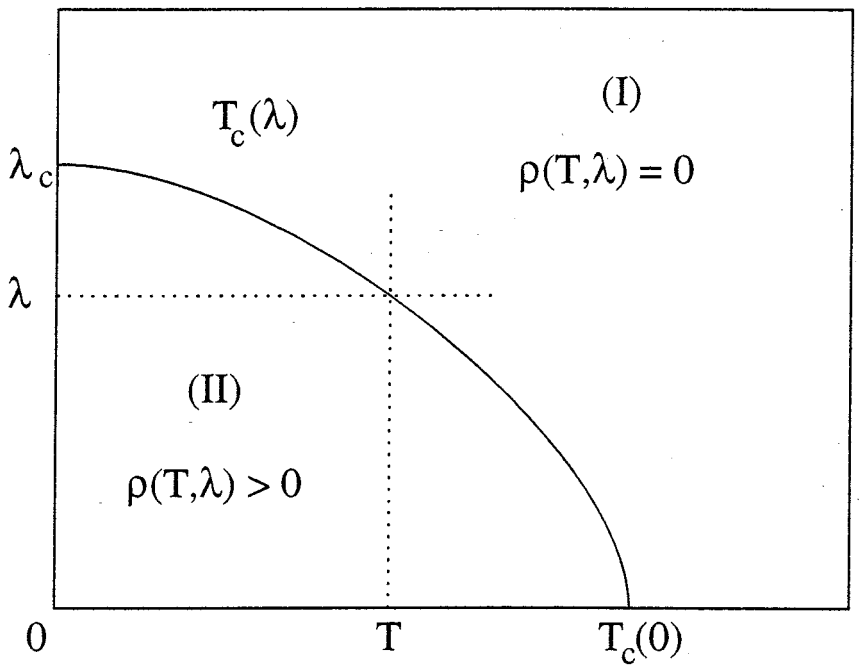

FIG. 1. Phase diagram of the model (2.1).

$$c^* = I_d(c^*, T_c(\lambda), \lambda), \quad \lambda \leqslant \lambda_c . \tag{2.13}$$

For a fixed $\lambda < \lambda_c$ we observe a well-known temperature-driven phase transition at $T_c(\lambda) > 0$ with an order parameter related to $\rho \equiv \rho_{h=0}$, while for a fixed $T < T_c(0)$ one gets a phase transition driven by the quantum parameter (2.8).

Let $\eta_{\Lambda,h}(\cdot) \equiv \langle \cdot \rangle_{H_{\Lambda,h}}$ be finite-volume equilibrium state corresponding to (2.1). Note that the $\{Q_l \to -Q_l\}_{l \in \mathbb{Z}^d}$ symmetry of the Hamiltonian (2.1) for $h = 0$ implies

$$\lim_{\Lambda} \langle Q_l \rangle_{H_{\Lambda,h=0}} = 0.$$

Now let $h \neq 0$. Then for the limiting state one obtains

$$\eta_h(Q_l) = \lim_{\Lambda} \eta_{\Lambda,h}(Q_l) = \frac{h}{\Delta(c_h)}. \tag{2.14}$$

Above the critical line $T_c(\lambda) < T$ (disordered phase) Eqs. (2.3) and (2.11) give that $\lim_{h \to 0} c_h(T,\lambda) = c(T,\lambda) > c^*$. Therefore, $\Delta(c_h) > 0$. Hence, by (2.12) $\rho_{h=0}(T,\lambda) = 0$ and by (2.14) one obtains

$$\lim_{h \to 0} \eta_h(Q_l) = 0.$$

In ordered phase, i.e., for $T < T_c(\lambda)$, by (2.3) and (2.11) we get $\lim_{h \to 0} c_h(T,\lambda) = c^*$. Then

$$\rho(T,\lambda) = \lim_{h \to 0} \rho_h(T,\lambda) = c^* - I_d(c^*, T, \lambda) = \lim_{h \to 0} \frac{h^2}{\Delta^2(c_h)} > 0, \tag{2.15}$$

and Eqs. (2.14) and (2.15) yield

$$\eta_{\pm}(Q_l) = \lim_{h \to \pm 0} \eta_h(Q_l) = \pm \sqrt{\rho(T,\lambda)} \neq 0. \tag{2.16}$$

Therefore, using the Bogoliubov quasi-averages, we obtain two extremal translation invariant equilibrium states η_+ and η_-.

Consider now fluctuation operators corresponding to particle position and momentum displacements defined by

$$F^{\delta_Q}(Q)=\lim_{\Lambda}\frac{1}{V^{1/2+\delta_Q}}\sum_{i\in\Lambda}(Q_i-\eta_{\Lambda,h}(Q_i)), \tag{2.17}$$

and

$$F^{\delta_P}(P)=\lim_{\Lambda}\frac{1}{V^{1/2+\delta_P}}\sum_{i\in\Lambda}(P_i-\eta_{\Lambda,h}(P_i)), \tag{2.18}$$

where exponents δ_Q and δ_P are chosen so that variances of $F^{\delta_Q(Q)}$ and $F^{\delta_P(P)}$ in the state η_h are nontrivial.

Calculation of variances of (2.17) and (2.18) gives

$$\lim_{\Lambda}\eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_Q}}\sum_{i\in\Lambda}(Q_i-\eta_{\Lambda,h}(Q_i))\right\}^2\right)$$

$$=\lim_{\Lambda}\frac{1}{V^{2\delta_Q}}\frac{\lambda}{2\sqrt{\Delta(c_{\Lambda,h}(T,\lambda))}}\coth\frac{\beta\lambda}{2}\sqrt{\Delta(c_{\Lambda,h}(T,\lambda))} \tag{2.19}$$

and

$$\lim_{\Lambda}\eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_P}}\sum_{i\in\Lambda}(P_i-\eta_{\Lambda,h}(P_i))\right\}^2\right)$$

$$=\lim_{\Lambda}\frac{1}{V^{2\delta_P}}\frac{\lambda m\sqrt{\Delta(c_{\Lambda,h}(T,\lambda))}}{2}\coth\frac{\beta\lambda}{2}\sqrt{\delta(c_{\Lambda,h}(T,\lambda))}. \tag{2.20}$$

By Eqs. (2.3) and (2.11), above the critical line for any h one has $\lim_\Lambda \Delta(c_{\Lambda,h}(T,\lambda))>0$. Consequently, (2.19) and (2.20) imply that $\delta_Q=\delta_P=0$, i.e., both momentum and displacement fluctuation operators are normal. Below the critical line, quasi-averages controled by the external field $h=\hat{h}/V^\alpha$ show (for $T>0$ and finite range interactions $\sigma\geqslant 2$) that the momentum fluctuation operator is normal while the position fluctuation operator is abnormal with the exponent δ_Q which depends on α.[4] Of course, the domain $\rho(T,\lambda)\neq 0$ is a particular one because there we have more than one pure equilibrium state. That is why a challenge problem is to study whether exponents δ_Q and δ_P depend on the parameter $\alpha>0$ for the *long-range* interaction $0<\sigma<2$ exactly on the critical line where the equilibrium state is *unique*.

III. FLUCTUATIONS ON THE CRITICAL LINE

Since by (2.13) and (2.15) one has $\rho(T_c(\lambda),\lambda)=0$, on the critical line Eq. (2.12) gets the following form:

$$\lim_{\Lambda}\left\{\frac{1}{V}\frac{\lambda}{2\sqrt{\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))}}\coth\frac{\lambda\beta_c(\lambda)}{2}\sqrt{\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))}+\frac{\hat{h}^2}{V^{2\alpha}\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))^2}\right\}=0, \tag{3.1}$$

where by $h=\hat{h}/V^\alpha$ and (2.11)–(2.13) we have $\lim_\Lambda c_{\Lambda,h}(T_c(\lambda),\lambda)=c^*$. Hence

$$\lim_{\Lambda}\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))=0. \tag{3.2}$$

Now we have to distinguish two cases:

(a) $T_c(\lambda)>0$, then (3.1) reduces to

$$\lim_{\Lambda}\left\{\frac{1}{V\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))\beta_c(\lambda)}+\frac{\hat{h}^2}{V^{2\alpha}\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))^2}\right\}=0; \tag{3.3}$$

(b) $T_c(\lambda_c)=0$, then (3.1) takes the form

$$\lim_{\Lambda}\left\{\frac{\lambda_c}{2V\sqrt{\Delta(c_{\Lambda,h}(0,\lambda_c))}}+\frac{\hat{h}^2}{V^{2\alpha}\Delta(c_{\Lambda,h}(0,\lambda_c))^2}\right\}=0. \tag{3.4}$$

These cases both imply that, for $V\to\infty$, the gap Δ (3.2) has the asymptotic behavior $\Delta\simeq V^{-\gamma}$: with $0<\gamma<\min(1,\alpha)$ [see (3.3)] or $0<\gamma<\min(2,\alpha)$ [see (3.4)].

It is Eq. (2.3) that is the key allowing one to obtain these asymptotics in explicit form. To make it evident, we rewrite (2.3) as

$$\begin{aligned}&(c_{\Lambda,h}-c^*)+[c^*-I_d(c_{\Lambda,h},T_c,\lambda)]+\left[I_d(c_{\Lambda,h},T_c,\lambda)\right.\\&\left.-\sum_{q\in\Lambda^*,q\neq 0}\frac{\lambda}{2\Omega_q(c_{\Lambda,h})}\coth\frac{\beta_c\lambda\Omega_q(c_{\Lambda,h})}{2}\right]\\&=\left(\frac{\hat{h}}{V^{\alpha}\Delta}\right)^2+\frac{1}{V}\frac{\lambda}{2\sqrt{\Delta}}\coth\left(\frac{\beta_c\lambda\sqrt{\Delta}}{2}\right).\end{aligned} \tag{3.5}$$

Lemma 3.1: If (T,λ) belongs to the critical line $(T_c(\lambda),\lambda)$ and $T_c(\lambda)>0$, then the asymptotic volume behavior of the gap $\Delta(c_{\Delta,h}(T_c(\lambda),\lambda))$ for $h=\hat{h}/V^{\alpha}$ is defined by

$$\gamma=\begin{cases}\text{if } d>2\sigma:\\ \frac{2}{3}\alpha \quad \text{for } \alpha<\frac{3}{4}=\alpha_c,\\ \frac{1}{2} \quad \text{for } \alpha\geqslant\frac{3}{4},\\ \text{if } d=2\sigma:\\ \frac{2}{3}\alpha+0 \quad \text{for } \alpha<\frac{3}{4}=\alpha_c,\\ \frac{1}{2}+0 \quad \text{for } \alpha\geqslant\frac{3}{4},\\ \text{if } \sigma<d<2\sigma:\\ \alpha\,\frac{\sigma}{d+\sigma} \quad \text{for } \alpha<\frac{1}{2}+\frac{\sigma}{2d}=\alpha_c,\\ \frac{\sigma}{d} \quad \text{for } \alpha\geqslant\frac{1}{2}+\frac{\sigma}{2d}.\end{cases}$$

Proof: Since $T_c(\lambda)>0$, the right-hand side of (3.5) has asymptotics (3.3) or

$$O\left\{\left(\frac{\hat{h}}{V^{\alpha}\Delta}\right)^2+\frac{1}{V\Delta}\right\}. \tag{3.6}$$

We define α_c by the condition $O[(1/V^{\alpha_c}\Delta)^2]=O(1/V\Delta)$. Then for the asymptotics of (3.6) one obtains

$$O\left(\left(\frac{\hat{h}}{V^{\alpha_c}\Delta}\right)^2+\frac{1}{V\Delta}\right)=O\left(\frac{1}{V\Delta}\right), \tag{3.7}$$

i.e., for $\alpha\geqslant\alpha_c$, the gap Δ has the same asymptotic behavior as for $h=0$[6].

Now we have to consider three regimes for the harmonic-force matrix element decay (1.4):

(a) Let $\sigma<d<2\sigma$. Then for $h=0$ Eq. (3.5) gives[6] the exponent $\gamma=\sigma/d$, i.e., $\Delta\simeq V^{-\sigma/d}$, and (3.7) yields $\alpha_c=\sigma/2d+1/2$.
(b) Let $d=2\sigma$. In this limiting case one has $\gamma=1/2+0$, so we obtain $\alpha_c=3/4$.
(c) Let $d>2\sigma$. Then (3.5) provides[6] $\gamma=1/2$. Hence $\alpha_c=3/4$.

We consider first the case $\alpha\geqslant\alpha_c$. Asymptotics in (3.5) are the same as for $\hat{h}=0$, i.e., we reproduce the results of Ref. 6: $\gamma=\sigma/d$, if $\sigma<d<2\sigma$; $\gamma=1/2+0$, if $d=2\sigma$; and $\gamma=1/2$, if $d>2\sigma$. Then we consider $\alpha<\alpha_c$ for the three cases (a), (b), and (c) correspondingly. Now the asymptotics of the first term in (3.6) will dominate.

(a) Let $\sigma<d<2\sigma$. Then (3.5) has the following asymptotic form[6]:

$$O[(\Delta+\Delta^{d/\sigma-1}+V^{\sigma/d-1})V^{2\alpha}\Delta^2]=O(1).$$

By the condition $\sigma<d<2\sigma$, one obtains

$$\gamma=2\alpha\,\frac{\sigma}{d+\sigma}. \tag{3.8}$$

(b) Let $d=2\sigma$. Then asymptotic form of (3.5) is

$$O[(\Delta+\Delta\ln(\Delta)+V^{1/2})V^{2\alpha}\Delta^2]=O(1).$$

Therefore, we obtain

$$\gamma=\frac{2\alpha}{3}+0. \tag{3.9}$$

(c) Let $d>2\sigma$. Then the asymptotic form of (3.5) is

$$O[(\Delta=V^{\sigma/d-1})V^{2\alpha}\Delta^2]=O(1).$$

By $d>2\sigma$, one gets that

$$\gamma=\frac{2\alpha}{3}. \tag{3.10}$$

The lemma is proven.

Proposition 3.2: If (T,λ) belongs to the critical line $(T_c(\lambda),\lambda)$ and $T_c(\lambda)>0$, then the algebra of fluctuation operators is Abelian. The momentum fluctuation operator $F^{\delta_P}(P)$ is normal ($\delta_P=0$) while the position fluctuation operator $F^{\delta_Q}(Q)$ is abnormal with critical exponent given by

$$\delta_Q=\begin{cases}
\text{if } d>2\sigma:\\
\frac{1}{3}\alpha \quad \text{for } \alpha<\frac{3}{4}=\alpha_c,\\
\frac{1}{4} \quad \text{for } \alpha\geqslant\frac{3}{4},\\
\text{if } d=2\sigma:\\
\frac{1}{3}\alpha+0 \quad \text{for } \alpha<\frac{3}{4}=\alpha_c,\\
\frac{1}{4}+0 \quad \text{for } \alpha\geqslant\frac{3}{4},\\
\text{if } \sigma<d<2\sigma:\\
\alpha\,\frac{\sigma}{d+\sigma} \quad \text{for } \alpha<\frac{1}{2}+\frac{\sigma}{2d}=\alpha_c,\\
\frac{\sigma}{2d} \quad \text{for } \alpha\geqslant\frac{1}{2}+\frac{\sigma}{2d}.
\end{cases}$$

Proof: To check that the algebra of fluctuation operators generated by F^{δ_Q} and F^{δ_F} is Abelian it is enough to calculate limit of the commutator:

$$\lim_{\Lambda}[F_{\Lambda}^{\delta_P},F_{\Lambda}^{\delta_Q}]=\lim_{\Lambda}\frac{1}{|\Lambda|^{1+\delta_P+\delta_Q}}\sum_{l,l'\in\Lambda}[P_l,Q_{l'}]. \tag{3.11}$$

The second part of the Proposition results from Eqs. (2.19) and (2.20). On the critical line and for $h=\hat{h}/V^{\alpha}$ they get the form

$$\lim_{\Lambda}\eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_Q}}\sum_{i\in\Lambda}(Q_i-\eta_{\Lambda,h}(Q_i))\right\}^2\right)=\lim_{\Lambda}\frac{1}{V^{2\delta_Q}}\frac{kT_c(\lambda)}{\Delta(c_{\Lambda,h}(T_c(\lambda),\lambda))}. \tag{3.12}$$

and

$$\lim_{\Lambda}\eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_P}}\sum_{i\in\Lambda}(P_i-\eta_{\Lambda,h}(P_i))\right\}^2\right)=\lim_{\Lambda}\frac{1}{V^{2\delta_P}}mkT_c(\lambda). \tag{3.13}$$

So, the variance (3.12) is nontrivial if and only if $\delta_Q=\gamma/2>0$, while (3.13) is nontrivial if and only if $\delta_P=0$. Hence, the limit of commutator (3.11) is equal to zero. One gets explicit values of δ_Q by Lemma (3.1).

Remark that if we put $\sigma=2$, then Proposition 3.2 reproduces results for short-range interactions $\sigma\geqslant 2$ (see Ref. 5).

Lemma 3.3: If (T,λ) coincides with the point $(0,\lambda_c,)$ then the asymptotic volume behavior of the gap $\Delta(c_{\Lambda,h}(0,\lambda_c),\lambda_c)$ is given by

$$\gamma=\begin{cases}\text{if } d>\frac{3\sigma}{2}:\\ \frac{2}{3}\alpha \quad \text{for } \alpha<1=\alpha_c,\\ \frac{2}{3} \quad \text{for } \alpha\geqslant 1,\\ \\ \text{if } d=\frac{3\sigma}{2}:\\ \frac{2}{3}\alpha+0 \quad \text{for } \alpha<1=\alpha_c,\\ \frac{1}{2}+0 \quad \text{for } \alpha\geqslant 1,\\ \\ \text{if } \frac{\sigma}{2}<d<\frac{3\sigma}{2}:\\ \alpha\ \frac{\sigma}{2d+3\sigma} \quad \text{for } \alpha<\frac{1}{2}+\frac{3\sigma}{4d}=\alpha_c,\\ \frac{\sigma}{d} \quad \text{for } \alpha\geqslant\frac{1}{2}+\frac{3\sigma}{4d}.\end{cases}$$

Proof: At this particular point on the critical line the right-hand side of (3.5) gets the following asymptotic form:

$$O\left[\left(\frac{\hat{h}}{V^{\alpha}\Delta}\right)^2+\frac{1}{V\Delta^{1/2}}\right]. \tag{3.14}$$

Similar to Lemma 3.1 we define α_c in such a way that $O[(1/V^{\alpha_c}\Delta)^2]=O(1/V\Delta^{1/2})$ and we distinguish three cases of the interaction decay: (a) $\sigma/2<d<3\sigma/2$, (b) $d=3\sigma/2$, and (c) $d>3\sigma/2$.

If $\alpha\geqslant\alpha_c$, then the gap Δ has the same asymptotic behavior as for $\hat{h}=0$[6]:

(a) For $\sigma/2<d<3\sigma/2$, it gives $O(\Delta)=O(V^{-\sigma/d})$. So by (3.14) we get $\alpha_c=3\sigma/4d+\frac{1}{2}$.
(b) If $d=3\sigma/2$, then $O(\Delta)=O(V^{-(2/3+0)})$, so we obtain $\alpha_c=1$.

(c) Let $d>3\sigma/2$. Then one gets $O(\Delta)=O(V^{-2/3})$, so $\alpha_c=1$.

Therefore, if $\alpha\geqslant\alpha_c$, then the last term in (3.14) dominates. Hence, γ has same values as for $\alpha=\alpha_c$.

Now we pass to $\alpha<\alpha_c$. Then for (a) $\sigma/2<d<3\sigma/2$, the asymptotic form of (3.5) is[6]

$$O([(\Delta+\Delta^{d/\sigma-1/2}+V^{\sigma/2d-1})])=O\left(\frac{1}{V^{2\alpha}\Delta^2}\right),$$

which yields $\gamma=4\alpha\sigma/(2d+3\sigma)$. For (b) $d=3\sigma/2$, one gets similarly to (3.9) that $\gamma=2\alpha/3+0$. Finally, for (c) $d>3\sigma/2$ the asymptotic form of (3.5) becomes

$$O((\Delta+V^{\sigma/2d-1}))=O\left(\frac{1}{V^{2\alpha}\Delta^2}\right)$$

and we find $\gamma=2\alpha/3$.

Proposition 3.4: If (T,λ) coincides with the point $(0,\lambda_c)$, then algebra of fluctuation operators is non-Abelian: the position fluctuation operator $F^{\delta_Q}(Q)$ is abnormal ($\delta_Q>0$), while the momentum fluctuation operator $F^{\delta_P}(P)$ is supernormal with $\delta_P=-\delta_Q$ and

$$\delta_Q=\begin{cases}\text{if } d>\frac{3\sigma}{2}:\\ \frac{1}{6}\,\alpha \text{ for } \alpha<1=\alpha_c,\\ \frac{1}{6} \text{ for } \alpha\geqslant 1,\\ \\ \text{if } d=\frac{3\sigma}{2}:\\ \frac{1}{6}\,\alpha+0 \text{ for } \alpha<1=\alpha_c,\\ \frac{1}{6}+0 \text{ for } \alpha\geqslant 1,\\ \\ \text{if } \frac{\sigma}{2}<d<\frac{3\sigma}{2}:\\ \alpha\,\frac{\sigma}{2d+3\sigma} \text{ for } \alpha<\frac{1}{2}+\frac{3\sigma}{4d}=\alpha_c,\\ \frac{\sigma}{4d} \text{ for } \alpha\geqslant\frac{1}{2}+\frac{3\sigma}{4d}.\end{cases}$$

Proof: If $T=0$ and $\lambda=\lambda_c$, variances (2.19) and (2.20) get the form

$$\lim_{\Lambda}\ \eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_Q}}\sum_{i\in\Lambda}(Q_i-\eta_{\Lambda,h}(Q_i))\right\}^2\right)=\lim_{\Lambda}\frac{1}{V^{2\delta_Q}}\frac{\lambda_c}{\sqrt{\Delta(c_{\Lambda,h}(0,\lambda_c))}}$$

and

$$\lim_{\Lambda}\ \eta_{\Lambda,h}\left(\left\{\frac{1}{V^{1/2+\delta_P}}\sum_{i\in\Lambda}(P_i-\eta_{\Lambda,h}(P_i))\right\}^2\right)=\lim_{\Lambda}\frac{1}{V^{2\delta_P}}\frac{\lambda_c m}{2}\sqrt{\Delta(c_{\Lambda,h}(0,\lambda_c))},$$

which implies $\delta_Q=-\delta_P$. For the rest one just has to apply Lemma 3.3 with $\delta_Q=\gamma/4$. Since $\delta_Q+\delta_P=0$, the non-Abelian nature of the algebra of fluctuation operators follows from the nontrivial commutator (3.11).

IV. CONCLUSION

Using a particular example of exactly soluble ferroelectric model we show that critical fluctuations depend on parameters σ and α. The first of them measures the power decay of the

harmonic-force matrix elements with the distance between sites of the lattice, while the second one controls the rate of the ''switching off'' of the infinitesimal external field breaking the Hamiltonian symmetry related to a phase transition.

The sensitivity with respect to σ is perfectly understandable: the long-range interaction is obviously involved in the formation of the long-range correlations in the critical equilibrium state. On the other hand, dependence of the exponents δ_P and δ_Q on α is less evident and is unexpected on the critical line where the limiting Gibbs state is unique. This means that the quasi-average perturbation $\sim \hat{h}/V^{\alpha}$, $\alpha>0$, is able to change fluctuations in the case of unique equilibrium state if the latter is critical. Note that there is no trace of perturbation $\sim \hat{h}/V^{\alpha}$, $\alpha>0$, above the critical line, i.e., for a unique noncritical state, while it appears again below the critical line even in pure states.[4]

The dependence of δ_P and δ_Q on α on the critical line means that such macro-observables as order parameters or fluctuations may have nontrivial structure in the unique critical equilibrium state. It seems that for the first time this kind of phenomenon has been observed in Ref. 8 in the example of a nontrivial structure of magnetization and its fluctuations in the ground state of the one-dimensional Ising model.

ACKNOWLEDGMENTS

One of us (VAZ) benefited greatly from discussions with André Verbeure, Charles-Edouard Pfister, and Senya Shlosman on problems discussed in this paper. We thank André Verbeure for directing our attention to Ref. 8. We are indebted to the referee for numerous constructive critical remarks.

[1] N. N. Bogoliubov, *Lectures on Quantum Statistics* (Gordon and Breach, New York, 1970), Vol. 2.
[2] N. Angelescu and V. A. Zagrebnov, J. Stat. Phys. **41**, 323 (1985).
[3] D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969).
[4] A. Verbeure and V. A. Zagrebnov, J. Stat. Phys. **69**, 329 (1992).
[5] A. Verbeure and V. A. Zagrebnov, J. Stat. Phys. **79**, 377 (1995).
[6] A. Car and V. A. Zagrebnov, Physica A **212**, 398 (1994).
[7] A. D. Bruce and K. A. Cowley, *Structural Phase Transition* (Taylor & Francis, London, 1981).
[8] R. H. Schonman and N. P. Tanaka, J. Stat. Phys. **61**, 241 (1990).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Colored noise and a characteristic level crossing problem

M. M. Kłosek
Department of Mathematical Sciences, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201

P. S. Hagan
Bank Paribas, New York, New York 10019

(Received 4 November 1996; accepted for publication 12 September 1997)

In this paper we discuss the activation of a dynamical system of one spatial variable forced by an external Gaussian stationary colored noise. The model is an analog of the Smoluchowski equation when external fluctuations are present. We assume that the underlying deterministic dynamics are derived from a potential forming a well. We enlarge the state space by including the noise variable, and then use half-range expansion techniques, singular perturbations, and matched asymptotics to solve the boundary value problem for the steady-state probability density function, in the asymptotic limit of small correlation time and over the entire regime of noise strength. We find a uniformly valid expression for the activation rate to the top of the potential barrier (a characteristic point for the unforced dynamical system). We show the important effects of the boundary behavior of the process on its activation rate. If the intensity of the noise is small compared to the autocorrelation of the noise, then the effective activation energy is modified by the noise autocorrelation. We consider an example of the activation process in a double well potential, and compare our uniformly valid results with results reported in the literature. © *1998 American Institute of Physics.* [S0022-2488(98)00501-5]

I. INTRODUCTION

Over the years activation phenomena in noisy dynamical systems have been studied extensively. There are two commonly distinguished classes of problems. The first is when the noisy fluctuations and the underlying deterministic dynamics are related by the fluctuation–dissipation principle (i.e., detailed balance holds).[1] The second is the class of open systems, in which deterministic dynamics are acted on by external fluctuations.

In this paper we discuss the activation of a one-dimensional spatial variable, $x(t)$, by an external Markovian Gaussian colored noise, $y(t)$, with nonzero correlation time. The underlying deterministic dynamics are assumed to occur in a potential well and the activation of $x(t)$ to the top of the potential barrier is the subject of this paper.

The deterministic dynamics for one spatial variable and in one degree of freedom can be viewed as the high damping limit of Newtonian dynamics—the analog of the Smoluchowski approximation to the Langevin equation.[1] The model has been derived independently a number of times with the spatial variable having specific physical interpretations. For example, one spatial variable has been used to describe dynamics along the reaction coordinate in a liquid medium;[1] models of rotational diffusion lead to one-dimensional dynamics for the angular coordinate in the presence of external fluctuations;[2] reconfiguration of proteins between two states under external temperature changes can be described by one-dimensional dynamics;[3] the time evolution of the phase of a laser obeys a first-order differential equation;[4] to name but a few.

We model the external fluctuations by Gaussian colored noise: the Ornstein–Uhlenbeck process.[5] This noise carries a finite energy and thus is a more realistic description of physical fluctuations than Gaussian white noise. In the limit of small correlation time the colored noise approaches the white noise. The white noise case is much easier to handle analytically than the colored noise case, and for that reason white noise has been used extensively to describe the fluctuation phenomena. However, colored noise may provide a more accurate description of the noisy phenomena; for example, in the context of laser-induced transitions,[4,6,7] as a model for

 © 1998 American Institute of Physics

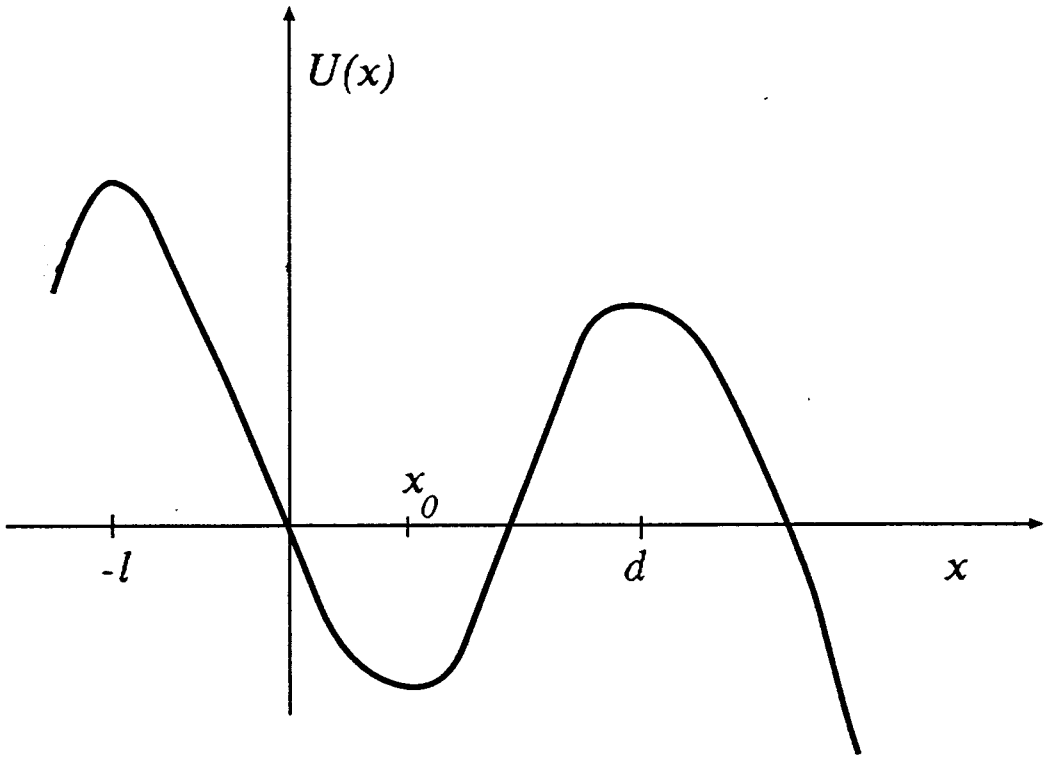

FIG. 1. A typical potential $U(x)$.

deviations from the thermal equilibrium by including nonzero correlation times,[8] with regard to modeling of biological membranes,[9] and many others.

The state variable $x(t)$ acted on by colored noise $y(t)$ is neither a diffusion nor a Markov process, due to the nonzero correlation time of $y(t)$. The phenomena of the activation of a non-Markovian process $x(t)$ and its numerical characterization require careful description. The concept of the reaction rate is well understood in the framework of diffusion processes, which are constructed from dynamical systems acted on by weak Gaussian white noise. If the underlying deterministic dynamics exhibit in the phase space (say) two domains of attraction of stable equilibria separated by the deterministic separatrix, then we say that activation occurs if a trajectory that originates at a stable equilibrium reaches an arbitrarily small neighborhood of the other stable equilibrium; that is, if a well-to-well transition occurs. In the weak noise limit, the transition rate between the stable equilibria is half the rate of a transition from a stable equilibrium to the deterministic separatrix. The activation rate to the separatrix can be calculated as the inverse of the mean first passage time for the trajectories to reach the separatrix from the stable state. Equivalently, the activation rate is inversely proportional to the first principal eigenvalue of the Fokker–Planck operator with absorbing boundary conditions on the separatrix.[10] In particular, in the case of one-dimensional dynamics in a potential well, as shown in Fig. 1, acted on by a weak Gaussian white noise, the domain of attraction of the stable equilibrium at x_0 is the interval $(-l,d)$, and the points $-l$ and d form the zero-dimensional separatrix.

The formalism described above is not directly applicable to the case of non-Markovian processes. Here we define the activation for the process $x(t)$ in the case when the underlying deterministic dynamics exhibit a well, as shown in Fig. 1. We say that the state variable x is activated if its coordinate changes from x_0 to d or $-l$, so that a trajectory reaches the top of the barrier. This definition corresponds, for example, to a deviation of the polarization angle;[2] reconfiguration of proteins;[3] phase fluctuations of laser light;[4] the orbiting model for ion–molecule reactions;[11] and others. Moreover, it is a straightforward generalization of activation by Gaussian white noise, as discussed above. We refer to the deviation of trajectories from their initial (stable) position to the barrier top as the characteristic level crossing problem.

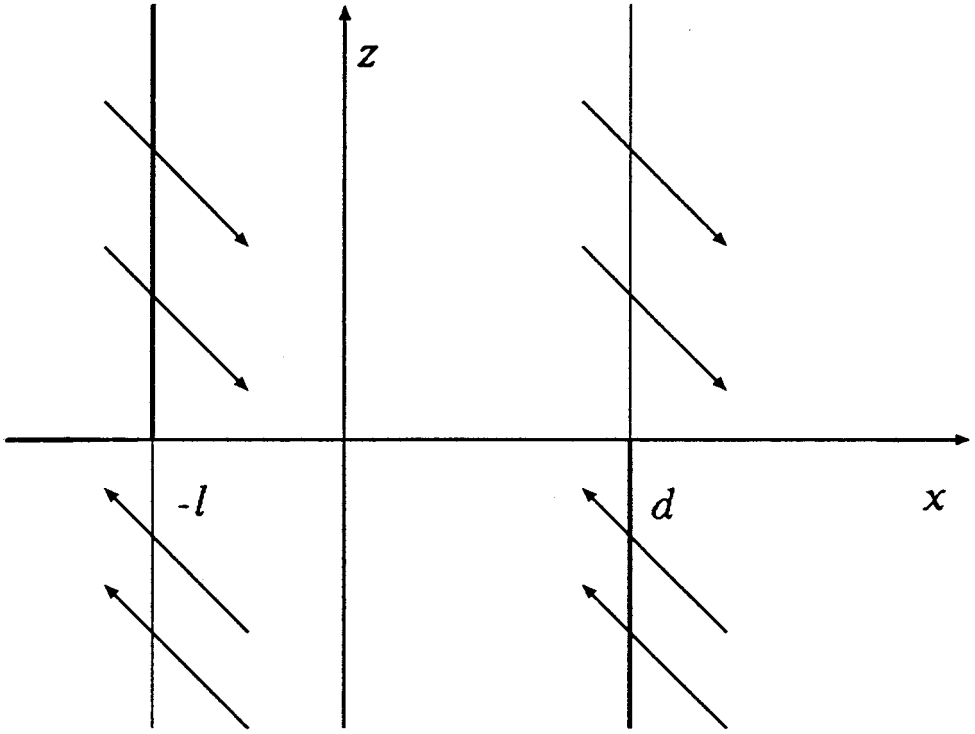

FIG. 2. The strip $D\equiv(-l,d)\times\mathscr{R}$. Arrows indicate the direction of the deterministic flow. The absorbing boundary conditions for the forward Fokker–Planck operator can only be prescribed on the parts of the boundaries where the flow is inwardly directed.

Next, we define the measure of the activation from the stable equilibrium for a non-Markovian process $x(t)$. In the case considered here, the difficulty of dealing with non-Markovian processes can be avoided by enlarging the state space by the noise variable $y(t)$ so that the two-dimensional process $(x(t),y(t))$ becomes Markov. Since the noise takes any value in $\mathscr{R}$, then the state space for the process $(x(t),y(t))$ is a strip $D=(-l,d)\times\mathscr{R}$, as shown in Fig. 2. We use the formalism of Markovian processes to calculate the mean first passage time of this two-dimensional diffusion to either level at $x=-l$, or d, and we take it to be the measure of the activation. A different (and nonequivalent) way to define the activation rate for the non-Markovian process $x(t)$ can be based on the observation that the process $(x(t),y(t))$ is a two-dimensional diffusion process acted on by Gaussian white noise, and the mean first passage time to the deterministic separatrix for the two-dimensional dynamics is a measure of the activation rate, as has been calculated elsewhere.[12] If we assume that trajectories originate with a random value of the noise (e.g., taken according to the stationary distribution), then it is possible for the transition to have occurred before the state variable equals the coordinate of the barrier top—see Fig. 3. Such an approach does not correspond to the class of problems in which the activation is determined solely by the deviation of the variable $x(t)$ from the equilibrium (cf. the above examples). In the enlarged phase plane, this latter class of problems corresponds to the situation when trajectories of the process $(x(t),y(t))$ reach a line $x=$const, with any value of the noise.

To summarize, in this paper we define the activation of the process $x(t)$ to be the activation of the Markov process $(x(t),y(t))$ to a prescribed level $x=$const. Our measure of the activation is the mean first passage time (mfpt) of the trajectories of the Markov process $(x(t),y(t))$ to $x=$const. This definition is relevant to situations in which the measure of transition depends only on the value of x and not on the value of the noise. We are able to solve the boundary value problem for the mean first passage time (mfpt) thanks to the newly developed analytical technique of half-range expansions.[13–15]

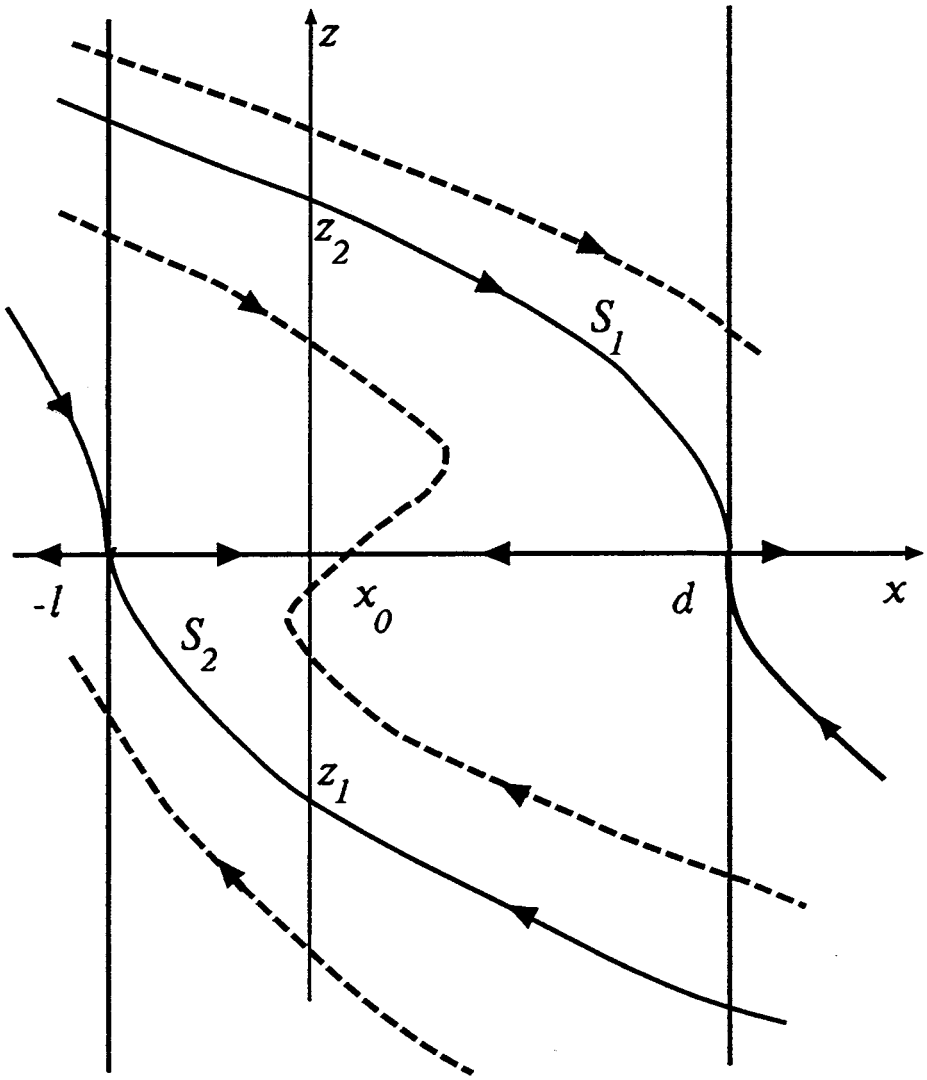

FIG. 3. The curve $S_1 \cup S_2$ (marked with a solid line) is the deterministic separatrix: the boundary of the domain of attraction in the two-dimensional phase space of the stable equilibrium at $(x_0,0)$. If the initial state of the noise, z_0, falls outside of the interval (z_1,z_2), then the initial state (x_0,z_0), of the two-dimensional system is outside of the domain of attraction of the stable equilibrium. According to the definition of the activation as the transition to the deterministic separatrix, this would mean that the system has been already activated, before the state variable x reached the level at either $-l$ or d. According to the definition of activation used in this work x must reach $-l$ or d, with any value of z.

II. MATHEMATICAL PRELIMINARIES AND RESULTS

We study activation of a system in which the deterministic dynamics has one degree of freedom,

$$\dot{x}(t) = -U'(x) + y(t), \tag{2.1}$$

and is acted on by a stationary Ornstein–Uhlenbeck process $y(t)$. The process $y(t)$ is a solution to the stochastic differential equation,[5]

$$\dot{y}(t) = -\frac{1}{\epsilon^2}\, y(t) + \frac{\sqrt{2\sigma}}{\epsilon^2}\, \dot{w}(t), \tag{2.2}$$

with the initial probability density $y(0) \sim \mathcal{N}(0,\sigma/\epsilon^2)$. Here $\dot{w}(t)$ denotes Gaussian white noise. It follows that the average $\langle y(t) \rangle = 0$, hence the autocorrelation of the process $y(t)$ is given by

$$\langle (y(t_1)-\langle y(t_1)\rangle)(\mathbf{y}(t_2)-\langle y(t_2)\rangle)\rangle = \langle y(t_1)y(t_2)\rangle = \frac{\sigma}{\epsilon^2}\, e^{-|t_1-t_2|/\epsilon^2}. \tag{2.3}$$

The parameters σ and ϵ are selected so ϵ^{-2} measures the width of the power spectrum of the noise and σ measures the noise intensity. In the limit as $\epsilon \to 0$, the colored noise $y(t)$ approaches, in distribution, Gaussian white noise of intensity σ; that is, $y(t) \to \sqrt{2\sigma}\dot{w}(t)$. We consider the process $x(t)$ on the interval $[-l,d]$ assuming that the deterministic potential has local extrema at $x=-l,d$, that is, $U'(-l)=U'(d)=0$, so the boundaries are characteristic.

The case of dynamics (2.1) in a bistable potential has played a special role in the literature.[16–18] Even in this ''simple'' case, the analysis of such a random system poses several analytical difficulties. For one, the stochastic process of the state variable $x(t)$ is non-Markovian. On the other hand, the stochastic process in the enlarged state space obtained by including the noise variable is Markovian. However, the (quasi-)stationary probability density function (pdf) of this Markovian process cannot be found explicitly for arbitrary deterministic dynamics. Therefore it is natural to pose questions about how a small but nonzero correlation time of the stochastic forcing term affects the activation of the system. The statistics for the case of zero correlation time forcing term, that is, the Gaussian standard white noise, are known. The problem of finding the asymptotic correction to the activation due to color of the forcing noise has led to a prolific discussion in the literature.[16–18] The discussion evolved about the fact that several methods that were proposed to solve the problem led to different results. One reason for some confusion has been the fact that the enlarged Markovian process depends not on one but on two perturbation parameters: the autocorrelation (color) ϵ^2 and the strength (intensity) σ of the noise. We will show below that the asymptotic limits of weak and nearly white noise are not interchangeable.

In this paper we present uniformly valid results on the activation of the process (2.1). We conduct a systematic asymptotic analysis, which allows us to keep track of the interplay of the parameters ϵ and σ. Specifically, we discuss activation by colored noise over the entire range of its intensity; that is, we do not take the strength σ to be necessarily small compared to the color ϵ, as is customarily assumed in the literature. We find that three distinct regimes need to be considered. The first regime occurs when the correlation time ϵ^2 is asymptotically smaller than the noise strength σ; the noise strength σ itself may be $O(1)$ or smaller in this regime, but $\epsilon^2 \ll \sigma$. We find that in this regime the color of the noise effects the activation rate algebraically. A distinguished regime occurs when $\sigma = O(\epsilon^2)$. In this regime the leading-order effects change from algebraic in color of the noise to the exponential in the noise intensity. Finally, we consider the third regime $\sigma \ll \epsilon^2 \ll 1$. In all three regimes a systematic treatment of the boundary behavior is essential to obtain the correct activation rate from the calculations of the mfpt.

Mathematically, the problem of calculating the mfpt can be formulated in terms of a boundary value problem for either the forward Fokker–Planck operator, $\mathscr{L}_{x,y}$,

$$\frac{\partial p}{\partial t} = \mathscr{L}_{x,y}p \equiv -\frac{\partial}{\partial x}\left[(-U'(x)+y)p\right] + \frac{\partial}{\partial y}\left(\frac{1}{\epsilon^2}\, yp\right) + \frac{\sigma}{\epsilon^4}\frac{\partial^2 p}{\partial y^2}, \tag{2.4}$$

or its adjoint, the backward Kolmogorov operator $\mathscr{L}^{*}_{x,y}$. The boundary conditions for the operator $\mathscr{L}_{x,y}$ can only be prescribed on parts of the boundary where the deterministic flow is inwardly directed.[19] In the present case this implies that the boundary conditions can be prescribed on half-lines: $\{x=-l, y>0\}$ and $\{x=d, y<0\}$ (cf. Fig. 2). The analytical solution to such a boundary value problem became possible only recently with the development of the closed form half-range expansion technique[13,14] and its generalization.[15]

In the limit as $\epsilon \to 0$ the process $y(t)$ approaches in distribution the standard Gaussian white noise of intensity σ. Hence the generator of the process $(x(t),y(t))$, $\mathscr{L}_{x,y}$, approaches the generator, $\mathscr{L}_x$, of a one-dimensional diffusion process of the dynamics $x(t)$ driven by the Gaussian white noise of intensity σ. Several authors[16–18,20] attempted to derive a diffusion approximation to the non-Markovian process $x(t)$, in the asymptotic limit of $\epsilon \ll 1$, by constructing a generator whose coefficients depend on the parameter ϵ, and that reduces to $\mathscr{L}_x$ when $\epsilon = 0$. This approach has been criticized since the meaning of the approximation of the non-Markov process by a Markov process is not well defined.[12] Moreover, these procedures have been applied to the free space operator $\mathscr{L}_{x,y}$ without reference to boundary conditions. We show that the statistics of the

non-Markovian process $(x(t),y(t))$, for example the mfpt, are the same as those of a diffusion process whose generator can be derived from the generator of $\mathscr{L}_{x,y}$, but very specific boundary conditions must be supplied for this diffusion process. That is, it is not enough to construct the free space generator $\mathscr{L}_x$, but rather the asymptotic analysis must be conducted on the boundary value problem for the generator $\mathscr{L}_{x,y}$, which in turn leads to a boundary value problem for the generator $\mathscr{L}_x$.

In this paper we solve the boundary value problem for the forward operator $\mathscr{L}_{x,y}$ with absorbing boundary conditions on half-lines $\{x=-l,y>0\}$ and $\{x=d,y<0\}$. We conduct a perturbation analysis, using boundary layer and half-range expansion techniques. Then we employ the solution to calculate various statistics of the process $(x(t),y(t))$ such as the mfpt and the distribution of exit points. Our results are uniformly valid for $\epsilon\ll 1$ and $\sigma=O(1)$ or smaller. We show that the mfpt of the process $(x(t),y(t))$ to a prescribed level $x=\text{const}$ depends on the mfpt of a one-dimensional diffusion process, $X(t)$, with absorbing boundary conditions imposed at different spatial locations than the boundary conditions for the process $(x(t),y(t))$ were prescribed. The parameters of this one-dimensional diffusion process depend on the color of the noise ϵ and the strength of the noise σ. The locations of the effective boundaries are extrapolated beyond the original interval in proportion to the color and the square root of the intensity of the noise. If $\epsilon\ll 1$ and σ is arbitrary, then the mfpt depends algebraically on the color of the noise. The effects of the color are seen in the effective diffusion coefficient, $D(x)$, the effective potential, $\Psi(x)$, the extrapolated boundaries, and the boundary layer solutions at the absorbing boundaries. If $\sigma\ll 1$, the mfpt is exponentially large in σ. The effects of the boundary layer analysis are of order $O(\epsilon^2)$; thus, they are transcendentally small compared to the other effects. The asymptotic limits of $\epsilon\ll 1$ and $\sigma\ll 1$ are not interchangeable. The expression for the mfpt depends on the relative size of the parameters. If $\epsilon\ll\sqrt{\sigma}$ then, to leading order, corrections to the mfpt are independent of the color of the noise; that is, the white noise approximation holds. In the distinguished limit regime of $\epsilon=O(\sqrt{\sigma})$, the effects due to the color and the intensity of the noise give terms of the same order. If $\sigma\ll\epsilon^2$ then the color of the noise gives the leading-order corrections both in the preexponential factor and in the exponent. This novel result shows that if the noise intensity is sufficiently small compared to the noise autocorrelation, then the activation energy (in addition to the preexponential factor) is changed due to the color of the noise. We call this a non-Markovian effect. This phenomena is similar to the energy gap observed in some ion–molecule reactions.[21]

To summarize, we construct the effective diffusion process $X(t)$, which is the solution to the Itô stochastic differential equation,

$$\dot{X}(t)=-\frac{\Psi'(X)}{D(X)}+\sqrt{\frac{2\sigma}{D(X)}}\,\dot{w}, \tag{2.5}$$

where the effective potential $\Psi(x)$ and the effective diffusion coefficient $D(x)$ are given by

$$\Psi(x)=U(x)+\frac{\epsilon^2}{2}\,(U'(x))^2-\frac{\epsilon^4}{2}\int_{x_0}^{x}U'^2(s)U'''(s)ds+O(\epsilon^6), \tag{2.6}$$

$$D(x)=1+\epsilon^2U''(x)+\epsilon^4\left[\tfrac{1}{2}\sigma U^{iv}(x)+\tfrac{1}{4}U''^2(x)-U'(x)U'''(x)\right]+O(\epsilon^6), \tag{2.7}$$

respectively. The process $X(t)$ satisfies absorbing boundary conditions at the points

$$x_R^*=d+\epsilon\sqrt{\sigma}\,|\zeta(1/2)|+O(\epsilon^3), \tag{2.8}$$

$$x_L^*=-l-\epsilon\sqrt{\sigma}\,|\zeta(1/2)|+O(\epsilon^3), \tag{2.9}$$

where ζ denotes the Riemann zeta function. Statistics of the process $x(t)$ can be approximated by statistics of the process $X(t)$ with an $O(\epsilon^3)$ error. In the entire range of the parameters the exit probabilities are the same for the two processes—compare Eqs. (5.18) and (5.19). If $\sigma=O(1)$, then the mfpt of the process $x(t)$ equals the mfpt for the process $X(t)$ plus an $O(\epsilon^2)$ correction due to the boundary layer solutions—compare Eq. (5.17). If $\sigma\ll 1$, then the mfpt of the two processes are equal within a transcendentally small error. Therefore the process $X(t)$ can be

interpreted as the Markovian approximation to the non-Markovian process $x(t)$ in the asymptotic limit of $\sigma \ll 1$. It is necessary to retain an $O(\epsilon^4)$ term in the expansion of Ψ in Eq. (2.6) in order to construct an expression for the mfpt uniformly valid for $\sigma \ll 1$ and $\epsilon \ll 1$. In particular, this order of the expansion is required to handle the distinguished limit regime of the parameters—see Eq. (6.5) and the discussion of Sec. VI.

The double well potential has played a special role in studying activation by colored noise. It has been established that for such a potential the leading-order correction to the mfpt due to the color of the noise appears in the preexponential factor at the $O(\epsilon^2)$ order, and it is proportional to 3/2. Our systematic perturbation analysis shows that if terms of $O(\epsilon^2)$ are to be kept in the preexponential factor, then additional terms must also be included in the formula for the mfpt. In Eq. (7.2) we illustrate the uniform formula (6.5) for the case of the double well potential. A term of $O(\epsilon)$ due to the extrapolated boundaries must be kept in the preexponential factor. More importantly, other terms due to the strength and the color of the noise must be retained to keep the asymptotic accuracy in the formula for the mfpt. In particular, Eq. (6.7) shows that if $\sigma \ll \epsilon^2$, then an exponential term proportional to ϵ^4/σ must be retained. These results are new—compare Refs. 16, 17, 18, and 22.

Our results give new insight to activation phenomena by driven external fluctuations. These results are only possible due to recent developments of rigorous mathematical techniques for solving explicitly the boundary layer problems when data is given only on half the boundaries to the domain.[13–15]

III. FORMULATION OF THE PROBLEM

We study the activation of the process $x(t)$ [Eq. (2.1)] due to the noise $y(t)$ [Eq. (2.2)], to boundaries of the interval $[-l,d]$, given that the process originates at $x_0 \in (-l,d)$, $O(1)$ away from the boundaries $x=-l(d)$. We assume that the deterministic potential $U(x)$ has a local minimum at x_0, and local maxima at $-l$ and d, such that $U'(-l)=0$, $U'(d)=0$ [the case of noncharacteristic boundaries $U'(-l)\neq 0$, $U'(d)\neq 0$ will be considered elsewhere].[23] Without any loss of generality we assume that $U'(x)\neq 0$ for $x \in (-l,x_0)\cup(x_0,d)$—see Fig. 1.

We rewrite

$$y(t)=\frac{\sqrt{\sigma}}{\epsilon}\, z(t), \tag{3.1}$$

and we observe that the pair $(x(t),z(t))$ is a Markov process that satisfies the system of stochastic differential equations,

$$\begin{aligned} \dot{x}(t) &= -U'(x)+\frac{\sqrt{\sigma}}{\epsilon}\, z, \\ \dot{z}(t) &= -\frac{1}{\epsilon^2}\, z+\sqrt{\frac{2}{\epsilon^2}}\, \dot{w}(t). \end{aligned} \tag{3.2}$$

The transition probability density function (tpdf), of the process (x,z), $p(x,z,t|x_0,z_0)$, satisfies the forward Fokker–Planck equation,[24]

$$\frac{\partial p}{\partial t}=\mathscr{L}_{x,z}p \equiv -\frac{\partial}{\partial x}\left[\left(-U'(x)+\frac{\sqrt{\sigma}}{\epsilon}\, z\right)p\right]+\frac{\partial}{\partial z}\left(\frac{1}{\epsilon^2}\, zp\right)+\frac{1}{\epsilon^2}\,\frac{\partial^2 p}{\partial z^2}. \tag{3.3}$$

We consider the process (x,z) on the strip $D \equiv (-l,d)\times\mathscr{R}$. We assume that all trajectories of the process (x,z) originate at (x_0,z_0), where the coordinate z_0 is selected according to the stationary noise density $r(z)=(1/\sqrt{2\pi})\exp\{-z^2/2\}$:

$$\lim_{t\to 0} p(x,z,t|x_0,z_0)=\delta(x-x_0)r(z_0)\,\delta(z-z_0). \tag{3.4}$$

The fact that no other trajectories enter the domain D implies absorbing boundary conditions on the left and on the right, respectively,

$$p(-l,z,t|x_0,z_0)=0, \quad \text{for} \quad z>0,$$

$$p(d,z,t|x_0,z_0)=0, \quad \text{for} \quad z<0; \tag{3.5}$$

see Fig. 2. We solve the problem (3.3–3.5) and we calculate the mean first passage time (mfpt) of the process (x,z) to the boundaries $x=-l$ and $x=d$. First we introduce the function $\hat{P}(x,z|x_0)$,

$$\hat{P}(x,z|x_0)=\int_{-\infty}^{+\infty}\int_{0}^{+\infty} p(x,z,t|x_0,z_0)dt\ dz_0. \tag{3.6}$$

The function $\hat{P}$ can be interpreted as the time spent by a trajectory at (x,z) given it originated at x_0, prior to absorption.[25] We seek the solution to Eq. (3.3) in the form

$$\overline{P}(x,z|x_0)=\frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}Q(x,z|x_0). \tag{3.7}$$

The function Q satisfies the boundary value problem

$$L_{x,z}Q\equiv Q_{zz}-zQ_z+\epsilon^2[U'(x)Q]_x-\epsilon\sqrt{\sigma}zQ_x=-\epsilon^2\delta(x-x_0), \tag{3.8}$$

$$Q(-l,z|x_0)=0, \quad \text{for } z>0, \tag{3.9}$$

$$Q(d,z|x_0)=0, \quad \text{for } z<0. \tag{3.10}$$

In Sec. IV we solve the problem (3.8)–(3.10) in the asymptotic limit of $\epsilon\ll 1$. We do not assume that the strength of the noise σ is small compared to other parameters of the problem. The problem (3.8)–(3.10) with boundary conditions prescribed only on the parts of the boundaries where the deterministic flow is inward is a typical case to analyze with the half-range expansion technique.[13,15]

After we calculate the solution Q we use it in Sec. IV to compute the following: (1) the mfpt of a trajectory to the level $x=-l,d$, $E\tau(x_0)$, given that the trajectory started at x_0,

$$E\tau=\int\int_{D}\frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}Q(x,z|x_0)dx\ dz; \tag{3.11}$$

(2) the probability density that the trajectory exits D with the value of the noise z,

$$p(x(\tau)=d,z(\tau)\in(z,z+dz)|x(0)=x_0)=\frac{\sqrt{\sigma}}{\epsilon}\,\frac{z}{\sqrt{2\pi}}\,e^{-z^2/2}Q(d,z|x_0)dz, \tag{3.12}$$

$$p(x(\tau)=-l,z(\tau)\in(z,z+dz)|x(0)=x_0)=-\frac{\sqrt{\sigma}}{\epsilon}\,\frac{z}{\sqrt{2\pi}}\,e^{-z^2/2}Q(-l,z|x_0)dz; \tag{3.13}$$

(3) the probability that the trajectory exits D on the left, p^{LEFT}, and the probability that the trajectory exits D on the right p^{RIGHT},

$$p^{\mathrm{LEFT}}=-\frac{\sqrt{\sigma}}{\epsilon}\int_{-\infty}^{0}\frac{z}{\sqrt{2\pi}}\,e^{-z^2/2}Q(-l,z|x_0)dz, \tag{3.14}$$

$$p^{\mathrm{RIGHT}}=\frac{\sqrt{\sigma}}{\epsilon}\int_{0}^{+\infty}\frac{z}{\sqrt{2\pi}}\,e^{-z^2/2}Q(d,z|x_0)dz. \tag{3.15}$$

The factor $\sqrt{\sigma}/\epsilon$ in Eqs. (3.12)–(3.15) results from the scaling (3.1).

In Sec. V we calculate the mfpt of the trajectories of the process (x,y) to the absorbing boundaries in the asymptotic limit of $\epsilon \ll 1$ and $\sigma \ll 1$.

IV. ANALYSIS

A. Outer solution

Away from the source and the boundaries, at a distance greater than $O(\epsilon)$, the function Q satisfies the outer problem,

$$L_{x,z} Q^{\mathrm{OUT}}(x,z) = 0. \tag{4.1}$$

As shown in Appendix A, in the limit of $\epsilon \ll \sqrt{\sigma}$, the solution to Eq. (4.1) can be expanded in an asymptotic series in powers of ϵ. In the limit of $\sigma \ll 1$, the (quasi-)stationary solution to Eq. (4.1) has a WKB structure; that is, it is exponential in $1/\sigma$. The transition between the two asymptotic behaviors of Q^{OUT} occurs for $\epsilon = O(\sqrt{\sigma})$, which is where the validity of the expansion in powers of ϵ breaks down. In Appendix A we also calculate the solution to Eq. (4.1) in the distinguished limit $\epsilon = O(\sqrt{\sigma})$—see Eqs. (A25) and (A28). We use it to construct the solution to Eq. (4.1), which is uniformly valid for $\epsilon \ll 1$, $\sigma = O(1)$ or smaller. We write

$$Q^{\mathrm{OUT}}(x,z) = \begin{cases} C_1 Q_1^{\mathrm{OUT}}(x,z) + D_1 Q_2^{\mathrm{OUT}}(x,z), & -l < x < x_0 \\ C_2 Q_1^{\mathrm{OUT}}(x,z) + D_2 Q_2^{\mathrm{OUT}}(x,z), & x_0 < x < d \end{cases}, \tag{4.2}$$

where

$$Q_1^{\mathrm{OUT}}(x,z) = \exp\left\{ -\frac{\Psi(x)}{\sigma} \right\} K(x,z), \tag{4.3}$$

$$Q_2^{\mathrm{OUT}}(x,z) = Q_1^{\mathrm{OUT}}(x,z) \int_{x_0}^{x} e^{\Psi(s)/\sigma}\, ds + M(x,z), \tag{4.4}$$

and where functions $\Psi(x)$, $K(x,z)$, and $M(x,z)$ have asymptotic expansions

$$\Psi(x) = U(x) + \frac{\epsilon^2}{2} (U'(x))^2 - \frac{\epsilon^4}{2} \int_{x_0}^{x} U'^2(s) U'''(s) ds + O(\epsilon^6), \tag{4.5}$$

$$K(x,z) = \left[1 + \frac{3}{2} \epsilon^2 U''(x) + O(\epsilon^3) \right] \exp\left\{ -\frac{1}{2} (1 + \epsilon^2 U''(x)) \left(z - \frac{\epsilon}{\sqrt{\sigma}} U'(x) \right)^2 + \frac{1}{2} z^2 + O(\epsilon^3/\sqrt{\sigma}) \right\}, \tag{4.6}$$

$$M(x,z) = \frac{\sigma}{U'(x)} \left[1 - \exp\left\{ \frac{\epsilon}{\sqrt{\sigma}} z U'(x) - \frac{\epsilon^2}{2\sigma} U'^2(x) \right\} \right] + O(\sigma^2). \tag{4.7}$$

We observe that Q^{OUT} is a well-defined function, even at points where $U'(x)$ vanishes. Here C_1, C_2, D_1, and D_2 are arbitrary constants, which are determined in the course of further analysis. We note that if $\epsilon \ll \sqrt{\sigma}$ and we expand the functions Ψ, K, and M in powers of ϵ, then we recover the expansion (A8). For $\sigma \ll 1$, after rearranging terms, Q_1^{OUT} agrees with the WKB expansion (A17), while for $\epsilon = O(\sqrt{\sigma})$ it equals the distinguished limit expansion given by Eqs. (A25) and (A28).

B. Analysis near the source

We introduce a local variable $\tilde{x} \equiv (x - x_0)/\epsilon\sqrt{\sigma}$. Since the point x_0 is a minimum of the potential, we expand $U(x)$ about x_0 as

$$U(x)=U(x_0)+\tfrac{1}{2}U''(x_0)(x-x_0)^2+\cdots . \tag{4.8}$$

We denote by Q^S the solution to Eq. (3.8) written in terms of $\tilde{x}$. We have

$$Q^S_{zz}-zQ^S_z-zQ^S_{\tilde{x}}+O(\epsilon^3)=0, \quad \text{for } \tilde{x}\neq 0, \tag{4.9}$$

$$Q^S(\tilde{x}=0^+,z)-Q^S(\tilde{x}=0^-,z)=\frac{\epsilon}{\sqrt{\sigma}}\frac{1}{z}. \tag{4.10}$$

Moreover Q^S must match Q^{OUT} as $\tilde{x}\to-\infty,+\infty$, respectively.

According to the calculations in Appendix B, we expand the right-hand side of Eq. (4.10) in terms of eigenfunctions $\{V_0,\tilde{V}_0,V_n^{\pm}\}$ as shown in Eqs. (B8) and (B10). Then the solution to Eq. (49) that satisfies the jump condition (4.10) is given by

$$\begin{aligned} Q^{S+}(\tilde{x},z)&=a_0+\tilde{a}_0(\tilde{x}-z)+\frac{\epsilon}{\sqrt{\sigma}}\sum_{n=1}^{\infty} I_n^- e^{-\sqrt{n}\tilde{x}}V_n^-(z), \quad \text{if } \tilde{x}>0,\\ Q^{S-}(\tilde{x},z)&=b_0+\tilde{b}_0(\tilde{x}-z)-\frac{\epsilon}{\sqrt{\sigma}}\sum_{n=1}^{\infty} I_n^+ e^{\sqrt{n}\tilde{x}}V_n^+(z), \quad \text{if } \tilde{x}<0, \end{aligned} \tag{4.11}$$

where $a_0=b_0$, $\tilde{b}_0-\tilde{a}_0=\epsilon/\sqrt{\sigma}$. The matching conditions between Q^{S+} and Q^{OUT} give $a_0=C_2\exp\{-U(x_0)/\sigma\}$, $\tilde{a}_0=\epsilon\sqrt{\sigma}D_2$, and the matching conditions between Q^{S-} and Q^{OUT} give $b_0=C_1\exp\{-U(x_0)/\sigma\}$, $\tilde{b}_0=\epsilon\sqrt{\sigma}D_1$, so

$$C_1=C_2, \quad D_2-D_1=1/\sigma. \tag{4.12}$$

C. Boundary layer analysis

In this section we construct the boundary layer solutions at $x=-l$ and at $x=d$, under the assumption $U'(-l)=U'(d)=0$. First we introduce a local variable $\xi\equiv(x+l)/\epsilon\sqrt{\sigma}$, and rewrite Eq. (3.8) for the local solution Q^L in terms of ξ and z as

$$\begin{aligned} &Q^L_{zz}-zQ^L_z-zQ^L_{\xi}+O(\epsilon^2)=0,\\ &Q^L(\xi=0,z)=0, \quad \text{for } z>0. \end{aligned} \tag{4.13}$$

Moreover, we require that Q^L matches the outer solution Q^{OUT} as $\xi\to\infty$, that is

$$Q^L\sim\left[D_1\int_{-l}^{x_0}e^{\Psi(y)/\sigma}\,dy+C_1\right]e^{-\Psi(-l)/\sigma}+D_1\epsilon\sqrt{\sigma}(\xi-z), \quad \text{as } \eta\to\infty. \tag{4.14}$$

According to the results of Refs. 13 and 14 quoted in Appendix B, the solution to Eq. (4.13) is given by

$$Q^L(\xi,z)=A^L\left[\alpha+\xi-z+\sum_{n=1}^{\infty}\beta_n V_n^-(z)e^{-\sqrt{n}\xi}\right]. \tag{4.15}$$

We recall that $\alpha\equiv|\zeta(1/2)|$ and the coefficients β_n and the eigenfunctions V_n are defined in Eq. (B13) and Eq. (B4), respectively. The constant A^L is to be determined from the matching condition (4.14).

Similarly, analysis about $x=d$ in the local variable $\eta\equiv(d-x)/\epsilon\sqrt{\sigma}$ leads to the boundary value problem for Q^R, given by

$$Q^R_{zz} - zQ^R_z + zQ^R_\eta = 0,$$

$$Q^R(\eta = 0, z) = 0, \qquad \text{for} \quad z < 0, \tag{4.16}$$

$$Q^R \sim \left[D_2 \int_{x_0}^{d} e^{\Psi(y)/\sigma}\, dy + C_2 \right] e^{-\Psi(d)/\sigma} + D_2 \epsilon \sqrt{\sigma}(\eta - z), \quad \text{as} \ \eta \to \infty.$$

We write the solution to Eq. (4.16) as

$$Q^R(\eta, z) = A^R \left[\alpha + \eta + z + \sum_{n=1}^{\infty} (-1)^n \beta_n V_n^+(z) e^{-\sqrt{n}\eta} \right], \tag{4.17}$$

where A^R is an arbitrary constant to be determined from matching conditions. Matching between Q^R and Q^{OUT} gives

$$\begin{aligned} A^R \alpha e^{\Psi(d)/\sigma} &= D_2 \int_{x_0}^{d} e^{\Psi(y)/\sigma}\, dy + C_2, \\ A^R &= -D_2 \epsilon \sqrt{\sigma}, \end{aligned} \tag{4.18}$$

while the matching between Q^L and Q^{OUT} gives

$$\begin{aligned} A^L \alpha e^{\Psi(-l)/\sigma} &= -D_1 \int_{-l}^{x_0} e^{\Psi(y)/\sigma}\, dy + C_1, \\ A^L &= D_1 \epsilon \sqrt{\sigma}. \end{aligned} \tag{4.19}$$

Combining Eqs. (4.12), (4.18), and (4.19) we determine all constants

$$\begin{aligned} D_2 &= -\frac{1}{\sigma} \frac{K(-l)}{K(-l) + K(d)}, \\ D_1 &= \frac{1}{\sigma} \frac{K(d)}{K(-l) + K(d)}, \\ C_1 = C_2 &= \frac{1}{\sigma} \frac{K(d)K(-l)}{K(-l) + K(d)}, \end{aligned} \tag{4.20}$$

where

$$K(-l) \equiv \int_{-l - \epsilon\sqrt{\sigma}\alpha}^{x_0} e^{\Psi(x)/\sigma}\, dx + O(\epsilon^3), \tag{4.21}$$

$$K(d) \equiv \int_{x_0}^{d + \epsilon\sqrt{\sigma}\alpha} e^{\Psi(x)/\sigma}\, dx + O(\epsilon^3). \tag{4.22}$$

V. SOLUTIONS

A. Uniform solution

We write the uniform solution to the problem (3.8) in the form

$$Q^{\mathrm{UNI}}(x,z) = Q^{\mathrm{OUT}}(x,z) + Q^{\mathrm{IN}(-l)}(x,z) + Q^{\mathrm{IN}(d)}(x,z) + Q^{\mathrm{IN}(S)}(x,z), \tag{5.1}$$

where $Q^{\mathrm{IN}(-l)}$, $Q^{\mathrm{IN}(d)}$ and $Q^{\mathrm{IN}(S)}$ denote the contributions due to the boundary layer on the left, on the right, and at the location of the source, respectively. They are given by

$$Q^{\mathrm{IN}(-l)}(x,z)=\frac{\epsilon}{\sqrt{\sigma}}\frac{K(d)}{K(-l)+K(d)}\sum_{n=1}^{\infty}\beta_n e^{-\sqrt{n}(x+l)/\epsilon\sqrt{\sigma}}V_n^-(z), \tag{5.2}$$

$$Q^{\mathrm{IN}(d)}(x,z)=\frac{\epsilon}{\sqrt{\sigma}}\frac{K(-l)}{K(-l)+K(d)}\sum_{n=1}^{\infty}\beta_n e^{-\sqrt{n}(d-x)/\epsilon\sqrt{\sigma}}V_n^-(-z), \tag{5.3}$$

$$Q^{\mathrm{IN}(S)}(x,z)=\begin{cases}\dfrac{\epsilon}{\sqrt{\sigma}}\displaystyle\sum_{n=1}^{\infty}I_n^-V_n^-(z)e^{-\sqrt{n}(x-x_0)/\epsilon\sqrt{\sigma}}, & \text{if } x>x_0,\\ -\dfrac{\epsilon}{\sqrt{\sigma}}\displaystyle\sum_{n=1}^{\infty}I_n^+V_n^+(z)e^{\sqrt{n}(x-x_0)/\epsilon\sqrt{\sigma}}, & \text{if } x<x_0,\end{cases} \tag{5.4}$$

with the coefficients β_n given by Eq. (B13) and the eigenfunctions $V_n^{\pm}$ given by Eq. (B4).

B. Effective diffusion

Next, we discuss the marginal steady-state pdf corresponding to the outer solution, that is, $P_{\mathrm{mar}}^{\mathrm{OUT}}\equiv\langle Q^{\mathrm{OUT}}(x,z),1\rangle$. Here $\langle\cdot,\cdot\rangle$ denotes the inner product defined in Appendix B in Eq. (B3) as $\langle f,g\rangle=1/\sqrt{2\pi}\int_{-\infty}^{+\infty}f(z)g(z)\exp(-z^2/2)dz$. We obtain

$$P_{\mathrm{mar}}^{\mathrm{OUT}}(x)=[1+\epsilon^2U''(x)]e^{-\Psi(x)/\sigma}\begin{cases}D_1\displaystyle\int_{-l-\epsilon\sqrt{\sigma}\alpha}^{x}e^{\Psi(y)/\sigma}\,dy, & x<x_0,\\ -D_2\displaystyle\int_{d+\epsilon\sqrt{\sigma}\alpha}^{x}e^{\Psi(y)/\sigma}\,dy, & x>x_0.\end{cases} \tag{5.5}$$

We observe that $P_{\mathrm{mar}}^{\mathrm{OUT}}$ does not vanish at $x=-l,d$, but rather it vanishes at the points extrapolated beyond the interval $(-l,d)$ whose coordinates are given by

$$x_R^*=d+\epsilon\sqrt{\sigma}\alpha+O(\epsilon^3), \tag{5.6}$$

$$x_L^*=-l-\epsilon\sqrt{\sigma}\alpha+O(\epsilon^3). \tag{5.7}$$

$P_{\mathrm{mar}}^{\mathrm{OUT}}$ can be interpreted as the steady-state pdf of a diffusion process X, which is the solution to the Itô stochastic differential equation,

$$\dot{X}(t)=-\frac{\Psi'(X)}{D(X)}+\sqrt{\frac{2\sigma}{D(X)}}\,\dot{w}, \tag{5.8}$$

where $\Psi(X)$ is given by Eq. (4.5) and $D(X)\equiv1+\epsilon^2U''(x)$. The process $X(t)$ has absorbing boundaries at x_R^* and x_L^*, with a trajectory originating at $x=x_0$ with probability one. In other words, the steady-state pdf, $P(X,t)$ of the process $X(t)$ satisfies the problem

$$\frac{\partial P(X,t)}{\partial t}=\mathscr{L}_XP(X,t)\equiv\frac{\partial}{\partial X}\left(\frac{\Psi'(X)}{D(X)}\,P(X,t)\right)+\sigma\,\frac{\partial^2}{\partial X^2}\left(\frac{1}{D(X)}\,P(X,t)\right), \tag{5.9}$$

$$\lim_{t\to0}P(X,t|x_0)=\delta(X-x_0), \tag{5.10}$$

$$P(X,t|x_L^*)=0,\quad P(X,t|x_R^*)=0. \tag{5.11}$$

The process $X(t)$ can be interpreted as a Markovian approximation of a non-Markovian process $x(t)$. We refer to $X(t)$ as the effective diffusion. The meaning and the validity of this approximation is discussed here and in the subsequent sections. We solved the boundary value problem for the generator, $\mathscr{L}_{x,z}$, of the diffusion process $(x(t),z(t))$ with absorbing boundary conditions—Eqs. (3.3)–(3.5). The projection of this two-dimensional problem on a one-dimensional space is

given by the generator $\mathscr{L}_X$ of a diffusion process (5.8) with absorbing boundary conditions at the extrapolated points (5.6) and (5.7). The error of this approximation is of $O(\epsilon^3)$. The effective potential Ψ and the effective diffusion coefficient $D(X)$ depend on the potential $U(x)$ and on the small parameter ϵ. The dependence of the effective diffusion on ϵ has been observed before.[16–18,12] In contrast to previous work[16,12,22] we keep the expansion of Ψ up to $O(\epsilon^4)$ in order to derive properties of the processes $(x(t),z(t))$ and $X(t)$, which are uniformly valid in ϵ and σ. This asymptotic order of the expansion of Ψ plays an essential role in understanding how the intensity and the color of the noise affect the activation of the process $x(t)$. In particular, it is important in the case when the two parameters are small, and as they change their relative magnitudes through the distinguished limit regime. The absorbing boundary conditions for the process X are prescribed at the points extrapolated beyond the original spatial location of the boundaries for the process $x(t)$. The size of the extrapolation depends on both ϵ and σ, but it is independent of the potential $U(x)$. This is due to the fact that the boundaries are characteristic, and to leading order the process $(x(t),z(t))$ in the boundary layer regime behaves like the integral of the Ornstein–Uhlenbeck process. Hence the extrapolated boundaries are the same as found in Refs. 13 and 14.

C. The mfpt

Next we use $Q^{\mathrm{UNI}}(x,z)$ to calculate the mfpt to the boundary of the strip D. According to the formula (3.11) we have

$$\begin{aligned} E\tau &= \int_{-l}^{d} \langle Q^{\mathrm{UNI}}(x,z),1\rangle dx = \int_{-l}^{d} \langle Q^{\mathrm{OUT}}(x,z),1\rangle dx + \int_{-l}^{d} \langle Q^{\mathrm{IN}(s)}(x,z),1\rangle dx \\ &\quad + \int_{-l}^{d} [\langle Q^{\mathrm{IN}(-l)}(x,z),1\rangle + \langle Q^{\mathrm{IN}(d)}(x,z),1\rangle] dx \\ &\equiv E\tau^{\mathrm{OUT}} + E\tau^{S} + E\tau^{\mathrm{BL}}. \end{aligned} \tag{5.12}$$

Here $E\tau^{\mathrm{OUT}}$, $E\tau^{S}$, and $E\tau^{\mathrm{BL}}$ denote the contribution to the mfpt due to the outer solution, the solution in the neighborhood of the source, and the absorbing boundary layer solution, respectively. We write

$$E\tau^{\mathrm{OUT}} = \int_{-l}^{d} P_{\mathrm{mar}}^{\mathrm{OUT}}(x)dx = \int_{x_L^*}^{x_R^*} R_{\mathrm{mar}}^{\mathrm{OUT}}(x)dx - \epsilon^2 \frac{\alpha^2}{2} + O(\epsilon^3) \equiv E\tau_{ExB}^{\mathrm{OUT}} - \epsilon^2 \frac{\alpha^2}{2} + O(\epsilon^3), \tag{5.13}$$

where $E\tau_{ExB}^{\mathrm{OUT}}$ denotes the mfpt of the diffusion process (5.8) to the extrapolated boundaries (5.6) and (5.7). The contribution to the mfpt in Eq. (5.12) due to the boundary layers is given by

$$E\tau^{\mathrm{BL}} = \int_{-l}^{d} \langle Q^{\mathrm{IN}(-l)},1\rangle dx + \int_{-l}^{d} \langle Q^{\mathrm{IN}(d)},1\rangle dx = \mu\epsilon^2, \tag{5.14}$$

where the coefficient μ is calculated numerically as the sum of the series

$$\mu \equiv \sum_{n=1}^{\infty} \beta_n \frac{1}{\sqrt{n}} \langle V_n^-(z),1\rangle = -0.206\,184\ldots . \tag{5.15}$$

The contribution to the mfpt due to the boundary layer at the location of the source is given by

$$E\tau^{S} = \int_{-l}^{d} \langle Q^{\mathrm{IN}(S)},1\rangle dx = \epsilon^2 \left[\sum_{n=1}^{-\infty} \frac{1}{\sqrt{n}} I_n^- \langle V_n^-(z),1\rangle - \frac{1}{\sqrt{n}} I_n^+ \langle V_n^+(z),1\rangle \right] = \epsilon^2. \tag{5.16}$$

The last identity follows from Eqs. (B16) and (B20). Combining Eqs. (5.13), (5.14), and (5.16), we write

$$E\tau = E\tau_{ExB}^{\mathrm{OUT}} + \epsilon^2\left(-\frac{\alpha^2}{2}+\mu+1\right)+O(\epsilon^3). \tag{5.17}$$

Equation (5.17) shows the dependence of the statistics of the marginal process $x(t)$ on the color and the intensity of the noise. First, one of the contributions to the mfpt of the process $X(t)$ comes from the statistics of the process $X(t)$ defined in Eq. (5.8). The diffusion coefficient, $D(X)$, and the effective potential, Ψ, depend on the color of the noise. Second, the absorbing boundary conditions for the process $X(t)$ are extrapolated beyond the location of absorbing boundaries of the process $x(t)$ by a distance proportional to the color of the noise. Third, the correction $\epsilon^2 \times(-\alpha^2/2+\mu+1)$ is due to the boundary layers at the absorbing boundaries and at the location of the source. The effects both of the color of the noise and the potential on the mfpt can be seen in $E\tau_{ExB}^{\mathrm{OUT}}$. We study asymptotic properties of $E\tau_{ExB}^{\mathrm{OUT}}$ as ϵ and σ vary in Sec. VI.

We also use Q^{UNI} to calculate the probabilities of exiting on the left and right. According to Eqs. (3.14) and (3.15), we have

$$p^{\mathrm{LEFT}} = \frac{\sqrt{\sigma}}{\epsilon}A^L = \frac{K(d)}{K(-l)+K(d)} = \frac{\int_{x_0}^{x_R^*} e^{\Psi(x)/\sigma}\,dx}{\int_{x_L^*}^{x_R^*} e^{\Psi(x)/\sigma}\,dx}+O(\epsilon^3), \tag{5.18}$$

$$p^{\mathrm{RIGHT}} = \frac{\sqrt{\sigma}}{\epsilon}A^R = \frac{K(-l)}{K(-l)+K(d)} = \frac{\int_{x_L^*}^{x_0} e^{\Psi(x)/\sigma}\,dx}{\int_{x_L^*}^{x_R^*} e^{\Psi(x)/\sigma}\,dx}+O(\epsilon^3). \tag{5.19}$$

The probabilities p^{LEFT} and p^{RIGHT} are equal to the exit probabilities on the left and right for the process $X(t)$.

VI. ASYMPTOTIC LIMIT OF WEAK NEARLY WHITE COLORED NOISE

The dependence of the mfpt on the two parameters is complicated, as can be seen from the simple considerations below. As an example, we assume that $\epsilon \ll 1$, and write

$$E\tau_{ExB}^{\mathrm{OUT}} = E\tau^0 + \epsilon E\tau^1 + \epsilon^2 E\tau^2 + O(\epsilon^3). \tag{6.1}$$

Here $E\tau^0$ denotes the mfpt of the diffusion process,

$$\dot{x} = -U'(x) + \sqrt{2\sigma}\dot{w}, \tag{6.2}$$

to absorption at $x=-l,d$. That is, the leading term corresponds to the approximation of the colored process by the white noise. The term $E\tau^1$ can be interpreted as the effect of the extrapolated boundaries, and so on. Formulas for $E\tau^i$, $i=0,1,...$, are rather complicated and one could try to simplify them by considering them in the standard asymptotic limit of small noise strength, $\sigma \ll 1$. Assuming that $U(-l)>U(d)$, we would obtain

$$E\tau^0 = \frac{\pi}{\sqrt{-U''(d)U''(x_0)}}\, e^{[U(d)-U(x_0)]/\sigma}[1+O(\sqrt{\sigma})]. \tag{6.3}$$

The last step of this analysis poses several problems. We derived the uniform solution to Eq. (3.8)–(3.10) and hence we calculated the formula for the mfpt, Eq. (5.12), valid in the asymptotic limit of $\epsilon \ll 1$. In particular, our results are valid for $\epsilon \ll \sqrt{\sigma}$. Therefore the term of $O(\sqrt{\sigma})$ in Eq. (6.3) is asymptotically greater than the first-order correction, $\epsilon E\tau^1$, which contains effects of the color of the noise. Therefore, disregarding the term of $O(\sqrt{\sigma})$ while keeping terms of $O(\epsilon)$ and $O(\epsilon^2)$ is incorrect. In order to obtain a correct asymptotic formula for the mfpt in the limit of $\sigma \ll 1$, one should retain as many terms in the expansion (6.3) as necessary for the order of magnitude of the disregarded terms to be asymptotically smaller than ϵ.

In order to fully understand the dependence of the mfpt on the two parameters we construct the formula for the mfpt that is uniformly valid in the asymptotic limit of $\epsilon \ll 1$ and $\sigma \ll 1$. If $\sigma \ll 1$ then the mfpt is exponentially large in this parameter, and according to Eq. (5.17) with a transcendentally small error we have

$$E\tau = E\tau_{ExB}^{\mathrm{OUT}} = \int_{x_L^*}^{x_R^*} P_{mar}^{\mathrm{OUT}}(x)dx = D_1 \int_{x_L^*}^{x_0} D(x) e^{-\Psi(x)/\sigma} \int_{x_L^*}^{x} e^{\Psi(y)/\sigma}\, dy\, dx - D_2 \int_{x_0}^{x_R^*} D(x) e^{-\Psi(x)/\sigma} \int_{x}^{x_R^*} e^{\Psi(y)/\sigma}\, dy\, dx. \tag{6.4}$$

We evaluate the Laplace-type integrals in Eq. (6.4) for $\sigma \ll 1$. First we note that the asymptotic expansion of the function Ψ in Eq. (4.5) was derived including terms up to $O(\epsilon^4)$. We assume that $\epsilon^6 \ll \sigma \ll 1$. (In order to obtain the answer for smaller values of σ, additional terms in the asymptotic expansion of Ψ should be kept.) Without any loss of generality we assume that $U(-l) > U(d)$, and we obtain

$$\begin{aligned} E\tau^{\mathrm{OUT}} &= \frac{\pi}{2\sqrt{-U''(x_0)U''(d)}} \exp\left\{ \frac{U(d)-U(x_0)}{\sigma} - \frac{\epsilon^4}{2\sigma} \int_{x_0}^{d} U'^2(y) U'''(y) dy \right\} \\ &\times \left\{ 1 + \sqrt{\frac{2}{\pi}} \sqrt{-U''(d)} \left(\epsilon\alpha - \sqrt{\sigma}\, \frac{1}{3} \frac{U'''(d)}{3U''^2(d)} \right) + \sigma \left[\frac{1}{8} \frac{U^{iv}(d)}{U''^2(d)} - \frac{5}{24} \frac{U'''^2(d)}{U''^3(d)} \right] - \epsilon^2 \frac{U''(d)}{2} \right\} \\ &\times \left\{ 1 - \sqrt{\frac{2}{\pi}} \sqrt{\sigma}\, \frac{1}{3} \frac{U'''(x_0)}{U''^{3/2}(x_0)} \right) + \sigma \left[-\frac{1}{8} \frac{U^{iv}(x_0)}{U''^2(x_0)} + \frac{5}{24} \frac{U'''^2(x_0)}{U''^3(x_0)} \right] + \epsilon^2 \frac{U''(x_0)}{2} \right\} \\ &\times \{1 + O(\epsilon^4, \epsilon^2\sqrt{\sigma}, \sigma^{3/2})\}. \end{aligned} \tag{6.5}$$

Equation (6.5) is uniformly valid for $\epsilon \ll 1$, $\epsilon^6 \ll \sigma \ll 1$. It shows that, depending on the asymptotic values of ϵ and σ, the main contributions to the formula (6.5) result from different effects. The terms proportional to powers of σ in the preexponential factor in Eq. (6.5) stem from the evaluation of the Laplace-type integrals. The term proportional to ϵ is the effect of the extrapolated boundaries, while the term proportional to ϵ^2 comes from the effective diffusion coefficient. The term in the exponent proportional to ϵ^4/σ is due to the effective potential. If $\epsilon \ll \sigma \ll 1$ then the leading-order terms in the preexponential term of Eq. (6.5) are due to the Laplace-type integrals; the effects of the color of the noise show up only as lower-order contributions. In this case the white noise approximation is valid with an error $O(\sqrt{\sigma})$. If $\epsilon = O(\sigma)$ then the order of magnitude of the term due to the extrapolated boundaries is the same as the order of magnitude of the second-order correction in the Laplace-type integral. So that if σ decreases to be $O(\epsilon)$, the effects of the color of the noise begin to show in Eq. (6.5). As σ decreases to the distinguished limit case, $\epsilon = O(\sqrt{\sigma})$, the effects of the color of the noise (through the effective diffusion coefficient and the effective potential, and the extrapolated boundaries) and the noise intensity (through the Laplace-type integrals) are of the same order and all of them must be retained in Eq. (6.5). If $\epsilon^6 \ll \sigma \ll \epsilon^4$ then the color of the noise gives the main order contribution. In the case of $\sigma \ll \epsilon^2$ the contribution in Eq. (6.5) due to the effective potential (the term proportional to ϵ^4/σ) shows that the color of the noise changes the activation energy. We refer to this phenomena as the non-Markovian effects on the activation rate.

If the strength of the noise is asymptotically larger than the color of the noise, then the noise strength gives the main contribution to the mfpt. In the distinguished limit regime, both parameters give main order contributions to the mfpt. If the noise strength is asymptotically smaller than the color of the noise, then the expression for the mfpt is changed due to the color of the noise, both in the preexponential factor and in the exponent. Thus, due to the color of the noise the effective activation energy is changed by a term of $O(\epsilon^4)$.

VII. EXAMPLE: THE DOUBLE WELL POTENTIAL

We illustrate the discussion of the preceding section with an example of the double well potential $U(x)=x^4/4-x^2/2$. We take $x_0=-1$ and $d=0$. The mfpt from x_0 to d due to the white noise is, to leading order, given by—compare Eq. (6.3),

$$E\tau_{\mathrm{WN}}=\frac{\pi}{2\sqrt{2}}\,e^{1/4\sigma}. \tag{7.1}$$

Then the formula (6.5) reduces to

If $\epsilon\ll\sigma\ll 1$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\sqrt{\sigma}\sqrt{\frac{1}{\pi}}+\frac{3}{2}\,\sigma+O(\epsilon,\sigma^{3/2})\right\}.$$

If $\sigma=O(\epsilon)$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\epsilon\alpha\sqrt{\frac{2}{\pi}}+\sqrt{\sigma}\sqrt{\frac{1}{\pi}}+\frac{3}{2}\,\sigma+O(\epsilon\sqrt{\sigma},\epsilon^2,\sigma^{3/2})\right\}.$$

If $\epsilon^2\ll\sigma\ll\epsilon$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\epsilon\alpha\sqrt{\frac{2}{\pi}}+\sqrt{\sigma}\sqrt{\frac{1}{\pi}}+\frac{3}{2}\,\sigma+\epsilon\sqrt{\sigma}\,\frac{\sqrt{2}}{\pi}+O(\epsilon\sqrt{\sigma},\epsilon^2,\sigma^{3/2})\right\}.$$

If $\sigma=\epsilon^2$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\epsilon\sqrt{\frac{2}{\pi}}\left(\alpha+\frac{\sqrt{2}}{2}\right)+\epsilon^2\left(\frac{3}{2}+\frac{\sqrt{2}}{\pi}\right)+O(\epsilon^3)\right\}.$$

If $\epsilon^4\leqslant\sigma\leqslant\epsilon^2$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\epsilon\sqrt{\frac{2}{\pi}}\,\alpha+\sqrt{\sigma}\,\frac{1}{\sqrt{\pi}}+\frac{3}{2}\,\epsilon^2+O(\epsilon^3,\sigma^{3/2},\epsilon\sqrt{\sigma})\right\}.e^{\epsilon^4/(8\sigma)}.$$

If $\epsilon^6\leqslant\sigma\leqslant\epsilon^4$ then

$$E\tau=E\tau_{\mathrm{WN}}\left\{1+\epsilon\sqrt{\frac{2}{\pi}}\,\alpha+\frac{3}{2}\,\epsilon^2+O(\sigma^{1/2})\right\}\cdot e^{\epsilon^4/(8\sigma)}. \tag{7.2}$$

The ''famous'' factor $3\epsilon^2/2$ does not show in the preexponential factor until $\sigma=O(\epsilon^2)$ or smaller. If $\epsilon^6\ll\sigma\ll\epsilon^4$ then no term proportional to a power of σ shows up to leading order in the preexponential factor. The term proportional to ϵ is the result of the extrapolated boundaries or equivalently the boundary layer analysis at the absorbing boundary location of the original problem. The factor $3\epsilon^2/2$ is due to the effective diffusion coefficient. However, if it is to be included in the formula for the mfpt, the exponential term with the exponent proportional to ϵ^4/σ must also be included in this expression. It is necessary in order for the mfpt expression to be asymptotically correct. Therefore, if the term of $3\epsilon^2/2$ order is to be included, other terms must be kept as well. Equation (6.7) gives the asymptotically exact expression for the mfpt to of the trajectories to the top of the potential barrier in the double well potential. The last two cases in Eq. (6.7) illustrate the effective change in the activation energy due to the color of the noise for sufficiently small noise intensity. In this case the effective activation energy increases from 1/4 to $1/4+\epsilon^4/8$.

VIII. SUMMARY

In this paper we analyze the activation of one-dimensional dynamics $x(t)$ by stationary Gaussian colored noise. We define the activation as the variation of the spatial variable $x(t)$ from its

initial state to a prescribed level $x=\text{const}$. In the examples considered here we took the level to be the coordinate of the barrier top, so that the boundary is characteristic. A different way of defining the activation is related to the transitions up to the deterministic separatrix in the two-dimensional state space $(x(t),y(t))$. These two types of activation problems are distinct, and they apply to different physical situations. Mathematically, the two problems lead to boundary value problems for the same Fokker–Planck operator, but with boundary conditions applied on different curves. In the latter case the conditions are prescribed on a separatrix. In contrast, for the level crossing problem, we must impose boundary conditions at (half-)lines $x=\text{const}$ Consequently, the activation rates differ between the two cases. In the past, some authors have defined the activation problem of the non-Markovian process $x(t)$ as the level crossing problem, but then tried to calculate the well-to-well transition. This has led to some confusion in the field—see Refs. 16–18 and the discussions therein. In the Introduction we clarify the distinction between these two problems. Another confusing topic in the literature has been the use of Markovian approximations to the evolution of the process $x(t)$. We show that the evolution of the process $x(t)$ cannot be treated separately from its boundary behavior.

Our analysis extends and corrects the analysis of Ref. 22. We analyze the level crossing problem for colored noise whose autocorrelation is taken to be small compared to $O(1)$ quantities like principal frequencies, the distance of the initial point to the level, etc. However, the correlation time can be large compared to the noise strength, when the latter is taken to be asymptotically small. This interplay of these small parameters and a systematic perturbation analysis lead to new insights into non-Markovian activation phenomena.

We are able to construct the outer solution to the problem (3.8)–(3.10), which is uniformly valid in the entire considered range of the parameters. Next, we match the outer solution to the half-range boundary layer solution. Consequently we obtain in Eq. (5.1) the uniform solution to the problem (3.8)–(3.10). These calculations allow the construction of the outer effective diffusion operator in Eqs. (5.9), which depends on the color of the noise at $O(\epsilon^2)$. An analog to this result is the Smoluchowski approximation to the Längevin equation since we construct a one-dimensional approximation $X(t)$ to the two-dimensional process $(x(t),y(t))$. The outer equation is equipped with the absorbing conditions at the extrapolated boundaries (5.6) and (5.7). The location of the new boundaries follows from the boundary layer analysis in Sec. IV. The validity of the approximation (5.9) to the original problem depends on the relative size of the color and intensity of the noise. If the intensity of the noise σ is of $O(1)$, while $\epsilon\ll 1$, then the outer problem alone (5.9)–(5.11) does not give the correct approximation to the original problem (3.3)–(3.5). From Eq. (5.17) we see that the boundary layer functions give a correction of $O(\epsilon^2)$ to the mfpt. This correction is of the same asymptotic order as the corrections to the outer diffusion operator due to the color of the noise and so cannot be ignored. This effect of the boundaries has not been derived before. If σ is asymptotically small compared to $O(1)$ quantities of the original problem, then Eqs. (5.9)–(5.11) provide a valid approximation to the original problem.

In Sec. VI we use Eq. (5.9)–(5.11) to calculate the mfpt in the asymptotic limit of $\sigma\ll 1$. We observe that the effects of the color of the noise on the activation rate change from algebraic to exponential as the intensity of the noise decreases through the distinguished limit of $\sigma=O(\epsilon)$. If $\sigma\ll\epsilon\ll 1$, then the color of the noise changes the activation energy, and thus modifies the Arrhenius factor. We call these new results non-Markovian effects on the activation rate.

We note that the half-range expansion technique of Refs. 13 and 14 outlined here, and its extensions in Ref. 15, are applicable to a larger class of diffusion-type transport problems. The example of the colored noise activation problem shows that the correct analytical treatment of the boundary conditions affects coefficients and parameters of such transport problems in a nontrivial way.

APPENDIX A: OUTER SOLUTION

In this appendix we calculate the solution to the outer problem (4.1),

$$Q_{zz}(x,z)-zQ_z(x,z)+\epsilon^2[U'(x)Q(x,z)]_x-\epsilon\sqrt{\sigma}zQ_x(x,z)=0, \tag{A1}$$

in the asymptotic limit $\epsilon\ll 1$ with σ being of $O(1)$ or smaller. First for $\epsilon\ll\sqrt{\sigma}$ we seek Q in the form of the asymptotic series,

$$Q(x,z)=\sum_{n=0}^{\infty}\frac{(-\epsilon\sqrt{\sigma})^n}{n!}Q^n(x)\mathrm{He}_n(z), \tag{A2}$$

where $\mathrm{He}_n(z)$ denotes the Hermite polynomials.[26] Substituting the series (A2) into Eq. (A1) we obtain the recursive system of equations for Q^n,

$$\sigma Q^1_x+[U'(x)Q^0]_x=0,$$

$$Q^n=Q^{n-1}_x+\frac{\epsilon^2}{n}[\sigma Q^{n+1}+U'(x)Q^n]_x, \quad n=1,2,3,\ldots. \tag{A3}$$

Next we solve the system (A3) by seeking Q^n in the form of the asymptotic series,

$$Q^n(x)=Q^n_0(x)+\epsilon^2Q^n_1(x)+\epsilon^4Q^n_2(x)+\cdots, \quad n=0,1,2,\ldots. \tag{A4}$$

We determine $Q^0(x)$ with an $O(\epsilon^6)$ error by solving the system of equations,

$$\hat{L}Q^0_0\equiv\sigma[Q^0_0(x)]_{xx}[U'(x)Q^1_0]_x=0,$$

$$\hat{L}Q^0_1=-\sigma[\sigma Q^2_0+U'(x)Q^1_0]_{xx}, \tag{A5}$$

$$\hat{L}Q^0_2=-\sigma[\sigma Q^2_1+U'(x)Q^1_1]_{xx},$$

where

$$Q^1_0(x)=[Q^0_0(x)]_x \quad Q^2_0(x)=[Q^0_0(x)]_{xx},$$

$$Q^1_1(x)=[Q^0_1(x)-U''(x)Q^0_0(x)]_x, \tag{A6}$$

$$Q^2_1(x)=[Q^0_1(x)]_{xx}-[\tfrac{3}{2}U'''(x)Q^0_0(x)+2U''(x)(Q^0_0(x))_x]_x.$$

In particular,

$$Q^0_0(x)=\left[D_1\int_{x_0}^{x}e^{U(t)/\sigma}\,dt+C_1\right]e^{-U(x)/\sigma}. \tag{A7}$$

Combining terms up to $O(\epsilon^2)$ we observe that Q^{OUT} can be written in the form

$$Q^{\mathrm{OUT}}(x,z)=C_1Q^{\mathrm{OUT}}_1(x,z)+D_1Q^{\mathrm{OUT}}_2(x,z), \tag{A8}$$

where

$$Q^{\mathrm{OUT}}_1(x,z)=\left[1+\frac{3}{2}\epsilon^2U''(x)+O(\epsilon^3)\right]\exp\left\{\frac{-\psi(x)}{\sigma}+\frac{z^2}{2}-\frac{\epsilon^2U''(x)z^2}{2}-\frac{1}{2}\left(\frac{\epsilon U'(x)}{\sqrt{\sigma}}-z\right)^2\right\}, \tag{A9}$$

$$Q^{\mathrm{OUT}}_2(x,z)=Q^{\mathrm{OUT}}_1(x,z)\int_{x_0}^{x}e^{\psi(s)/\sigma}\,ds-\epsilon\sqrt{\sigma}z-\frac{1}{2}\epsilon^2z^2U'(x)+\frac{1}{2}\epsilon^2U'(x)+O(\epsilon^3), \tag{A10}$$

with

$$\psi(x)\equiv U(x)+\tfrac{1}{2}\epsilon^2U'^2(x). \tag{A11}$$

Next we calculate the (quasi-)stationary pdf of the marginal process $x(t)$. We have

$$P^{\mathrm{mar}}(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}Q(x,z)e^{z^2/2}dz=Q^0(x)=Q_0^0(x)+\epsilon^2Q_1^0(x)+\epsilon^4Q_2^0(x)+O(\epsilon^6). \quad \text{(A12)}$$

Taking $Q^0(x)=\exp(-U(x)/\sigma)$ in Eq. (A12), we obtain

$$\begin{aligned}P^{\mathrm{mar}}(x)&=e^{-U(x)/\sigma}\left[1+\epsilon^2\left(U''(x)-\frac{1}{2\sigma}U'^2(x)\right)+\epsilon^4\left[\frac{\sigma}{2}U^{iv}(x)+\frac{1}{4}U''^2(x)-3U'''(x)U'(x)\right.\right.\\&\left.\left.+\frac{1}{2\sigma}\int_{x_0}^{x}U'^2(y)U'''(y)dy+\frac{1}{8\sigma^2}U'^4(x)\right]+O(\epsilon^6)\right]\\&=\{1+\epsilon^2U''(x)+O(\epsilon^4)]\}\\&\times\exp\left\{-\frac{1}{\sigma}\left[U(x)+\frac{\epsilon^2}{2}U'^2(x)-\frac{\epsilon^4}{2}\int_{x_0}^{x}U'^2(y)U'''(y)dy+O(\epsilon^6)\right]\right\}. \quad \text{(A13)}\end{aligned}$$

Next we outline the calculations of the (quasi-)stationary pdf in the asymptotic limit $\sqrt{\sigma}\ll\epsilon\ll1$. First we rescale the process $y(t)$ in Eq. (2.1) as—compare Ref. 12,

$$y(t)=\frac{1}{\epsilon}v(t), \quad \text{(A14)}$$

and we write the system for the Markov process (x,v) in the form equivalent to Eq. (3.2) as

$$\begin{aligned}\dot{x}(t)&=-U'(x)+\frac{1}{\epsilon}v,\\\dot{v}(t)&=-\frac{1}{\epsilon^2}v+\sqrt{\frac{2\sigma}{\epsilon^2}}\dot{w}(t).\end{aligned} \quad \text{(A15)}$$

The (quasi-)stationary pdf $p(x,v)$ satisfies the forward Fokker–Planck operator for Eq. (A15)

$$\mathscr{L}_{x,v}p\equiv-\frac{\partial}{\partial x}\left[\left(-U'(x)+\frac{1}{\epsilon}v\right)p\right]+\frac{\partial}{\partial v}\left(\frac{1}{\epsilon^2}vp\right)+\frac{\sigma}{\epsilon^2}\frac{\partial^2p}{\partial v^2}=0. \quad \text{(A16)}$$

We seek the solution to Eq. (A16) in the WKB form,

$$p(x,v)=\mathscr{K}(x,v)e^{-v^2/2\sigma}e^{-\Phi(x,v)/\sigma}. \quad \text{(A17)}$$

Upon substituting Eq. (A17) into Eq. (A16) we find the eikonal equation for Φ,

$$\Phi_v^2+v\Phi_v-[\epsilon^2U'(x)-\epsilon v]\Phi_x=0, \quad \text{(A18)}$$

and the transport equation for $\mathscr{K}$,

$$-\Phi_{vv}.\mathscr{K}-2\Phi_v.\mathscr{K}_v-v.\mathscr{K}_v+\epsilon^2U''(x).\mathscr{K}+[\epsilon^2U'(x)-\epsilon v].\mathscr{K}_x=0. \quad \text{(A19)}$$

We are unable to solve the system (A18)–(A19) exactly, but we determine its solution in the asymptotic series in powers of ϵ (cf. Ref. 12). We seek Φ in the form of the asymptotic series,

$$\Phi(x,v)=\phi^0(x,v)+\epsilon\phi^1(x,v)+\epsilon^2\phi^2(x,v)+\cdots. \quad \text{(A20)}$$

Upon substituting the series (A20) into Eq. (A18), and equating to zero coefficients of separate powers of ϵ we obtain a recursive system of equations for ϕ^i. We solve this system requiring Φ to be nonsingular in v. We obtain

$$\phi^0(x,v)=U(x),$$

$$\phi^1(x,v)=-vU'(x),$$

$$\phi^2(x,v)=\frac{v^2}{2}U''(x)+U'^2(x), \tag{A21}$$

$$\phi^3(x,v)=-\frac{v^3}{6}U'''(x)-vU'(x)U''(x),$$

$$\phi^4(x,v)=\frac{v^4}{24}U^{iv}(x)+\frac{v^2}{4}U'(x)U'''(x)+\int_{x_0}^{x}U'(y)U''^2(y)dy,$$

and so on. Then we seek the solution to the transport equation (A19) for $\mathscr{K}\equiv\mathscr{K}(x,v)$,

$$-\phi_{vv}\mathscr{K}-2\Phi_v\mathscr{K}_v-v\mathscr{K}_v+\epsilon^2U''(x)\mathscr{K}+[\epsilon^2U'(x)-\epsilon v]\mathscr{K}_x=0, \tag{A22}$$

in the asymptotic series of powers of ϵ. Using the expansion (A21), we obtain

$$\mathscr{K}(x,v)=1+\tfrac{3}{2}\epsilon^2U''(x)+O(\epsilon^3). \tag{A23}$$

Next we calculate the (quasi-)stationary pdf of the marginal process $x(t)$ in the asymptotic limit $\sqrt{\sigma}\ll\epsilon\ll1$. Using the fact that $\int U'U''^2\,dx=U'^2U''/2-\int U'^2U'''\,dy/2$, we have

$$\begin{aligned}P^{\mathrm{mar}}(x)&=\frac{1}{\sqrt{2\pi\sigma}}\int_{-\infty}^{\infty}\mathscr{K}(x,v)e^{-v^2/2\sigma}e^{-\Phi(x,v)/\sigma}\,dv\\&=[1+\epsilon^2U''(x)]\\&\quad\times\exp\left\{-\frac{1}{\sigma}\left[U(x)+\frac{\epsilon^2}{2}U'^2(x)-\frac{\epsilon^4}{2}\int_{x_0}^{x}U'^2(y)U'''(y)dy\right]\right\}(1+O(\epsilon^3)).\end{aligned} \tag{A24}$$

We retain the $O(\epsilon^4)$ term in the exponent while disregarding $O(\epsilon^3)$ terms in the preexponential factor in order to construct the uniform approximation to the pdf and the mfpt, as discussed in Sec. III.

Next, we construct the solution to Eq. (4.1) in the distinguished limit $\epsilon=O(\sqrt{\sigma})$. To this end we seek Q^{OUT} in the form

$$Q^{\mathrm{OUT}}=\exp\left\{-\frac{1}{\sigma}U(x)-\frac{\epsilon^2}{\sigma}U'^2(x)+\frac{\epsilon}{\sqrt{\sigma}}zU'(x)\right\}K(x,z). \tag{A25}$$

Substituting the expansion (A25) into Eq. (4.1), then changing variables $(x,z)\to(x,v\equiv2\epsilon U'(x)/\sqrt{\sigma}-z)$ and writing $\sqrt{\sigma}=l\epsilon$ we obtain the equation for K,

$$K_{vv}-vK_v+\epsilon^2\left\{\frac{2}{l}U'[lv-U']K_v+l\left[v-\frac{U'}{l}\right]K_x+\left[U''+\frac{1}{l}U'U''v-v^2U''\right]K\right\}=0. \tag{A26}$$

We seek the solution Eq. (A26) in the form of the asymptotic series,

$$K=K_0+\epsilon^2K_1+\epsilon^4K_2+\cdots, \tag{A27}$$

to obtain

$$K(x,z)=1+\epsilon^2\left[\frac{3}{2}\,U''-\frac{z^2}{2}\,U''+\frac{z}{l}\,U'U''-\frac{1}{2}\,\frac{U'^2U''}{l^2}-\frac{1}{2l^2}\int_{x_0}^{x}U'^2(y)U'''(y)dy\right]+O(\epsilon^4). \tag{A28}$$

APPENDIX B: BOUNDARY LAYER SOLUTION

The local analysis about the location of the source and about the location of the absorbing characteristic boundaries leads to a problem for the differential operator given by

$$Q_{zz}-zQ_z-zQ_x=0. \tag{B1}$$

Separation of variables in Eq. (B1) leads to the eigenvalue problem for the operator L given by

$$LV\equiv V''-zV'=\lambda zV. \tag{B2}$$

We observe that L is self-adjoint in the inner product,

$$\langle V,W\rangle=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-z^2/2}V(z)W(z)dz. \tag{B3}$$

Following the analysis, of Refs. 13 and 14 we find that $\lambda=0$ is a double eigenvalue of L with the eigenfunction $V_0=1$, and the generalized eigenfunction $\tilde{V}_0=z$. All other eigenvalues and their corresponding eigenfunctions of L are given by

$$\lambda_n^{\pm}=\pm\sqrt{n}\leftrightarrow V_n^{\pm}(z)=e^{z^2/4}e^{-(z\pm2\sqrt{n})^2/4}\,\mathrm{He}_n(z\pm2\sqrt{n}), \tag{B4}$$

where $\mathrm{He}_n(x)$ denotes the Hermite polynomial.[26] The eigenfunctions $\{V_0,\tilde{V}_0,V_n^{\pm}\}$ are orthogonal in the inner product (B3) with the weight function $w(z)=z$, that is,

$$\langle zV_n^{\pm},V_m^{\pm}\rangle=\langle zV_0,\tilde{V}_0\rangle=\langle zV_n^{\pm},V_0\rangle=0,\quad \text{for } n\neq m, \tag{B5}$$

$$\langle zV_n^{\pm},\tilde{V}_0\rangle=0,\quad \langle zV_n^{\pm},V_m^{\mp}\rangle=0, \tag{B6}$$

and the normalization constants are

$$\langle zV_n^{\pm},V_n^{\pm}\rangle=\mp2\sqrt{n}n!. \tag{B7}$$

The set $\{V_0,\tilde{V}_0,V_n^{\pm},\ n=1,2,3,...\}$ is a complete set of eigenfunctions of the operator L.[27,28] The expansion of a function $f(z)$ in terms of this basis can be written as

$$f(z)=I_0V_0(z)+\tilde{I}_0\tilde{V}_0(z)+\sum_{n=1}^{\infty}I_n^{+}V_n^{+}(z)+\sum_{n=1}^{\infty}I_n^{-}V_n^{-}(z), \tag{B8}$$

where

$$I_n^{\pm}=\frac{\langle zf(z),V_n^{\pm}(z)\rangle}{\langle zV_n^{\pm}(z),V_n^{\pm}(z)\rangle},\quad I_0=\frac{\langle z^2f(z),1\rangle}{\langle z^2,1\rangle},\quad \tilde{I}_0=\frac{\langle zf(z),1\rangle}{\langle z^2,1\rangle}. \tag{B9}$$

In particular if $f(z)=1/z$ then the coefficients (B9) become

$$I_0=0,\quad \tilde{I}_0=1,\quad I_n^{+}=-\frac{e^{-n/2}n^{n/2}}{2\sqrt{n}n!},\quad I_n^{-}=-(-1)^nI_n^{+}. \tag{B10}$$

According to the analysis of Refs. 13, 14, the solution to Eq. (B1), which satisfies the absorbing boundary conditions,

$$Q(x=0,z)=0,\quad \text{for}\quad z>0, \tag{B11}$$

and does not grow faster than linearly as $x\to\infty$ is given by

$$Q(x,z)=A\left[\alpha+x-z+\sum_{n=1}^{\infty}\beta_n V_n^-(z)e^{-\sqrt{n}x}\right]. \tag{B12}$$

Here $\alpha\equiv|\zeta(1/2)|$, where ζ denotes the Riemann zeta function, and

$$\beta_n=-(-1)^n\frac{N(\sqrt{n})}{2n!\sqrt{n}} \tag{B13}$$

and

$$N(s)=\prod_{k=1}^{\infty}\left(1+\frac{s}{\sqrt{k}}\right)^{-2s(\sqrt{k}-\sqrt{k-1})}\left(\frac{k+1}{k}\right)^{s^2/2}. \tag{B14}$$

We evaluate the following sum numerically [compare Eq. (5.14)]

$$\sum_{n=1}^{\infty}\frac{1}{\sqrt{n}}\beta_n\langle V_n^-(z),1\rangle=-\sum\nolimits_{n=1}^{\infty}\frac{N(\sqrt{n})n^{n/2}e^{-n/2}}{2nn!}\equiv\mu=-0.206\,184\ldots. \tag{B15}$$

We also need to evaluate the following sum [see Eq. (5.16)]:

$$\mathscr{S}_1\equiv\sum_{n=1}^{\infty}\frac{1}{\sqrt{n}}[I_n^-\langle V_n^-(z),1\rangle-I_n^+\langle V_n^+(z),1\rangle]=\sum_{n=1}^{\infty}\frac{\langle 1,V_n^-\rangle^2}{\sqrt{n}\langle zV_n^-,V_n^-\rangle}-\sum_{n=1}^{\infty}\frac{\langle 1,V_n^+\rangle^2}{\sqrt{n}\langle zV_n^+,V_n^+\rangle}. \tag{B16}$$

First, we observe that as a result of Eq. (B6) we have

$$\langle 1,V_n^-\rangle=\langle 1-z\tilde{V}_0,V_n^-\rangle=\frac{1}{2}\langle Lz^2,V_n^-\rangle=-\frac{\sqrt{n}}{2}\langle z^2,zV_n^-\rangle. \tag{B17}$$

Similarly,

$$\langle V_n^+,1\rangle=\frac{\sqrt{n}}{2}\langle z^2,zV_n^+\rangle. \tag{B18}$$

We observe that if f and g are any square integrable functions in the inner product (B3) then

$$\langle zf,g\rangle=\langle z^2,f\rangle\langle z,g\rangle+\langle z,f\rangle\langle z^2,g\rangle+\sum_{n=1}^{\infty}\frac{\langle zf,V_n^+\rangle\langle zg,V_n^+}{\langle zV_n^+,V_n^+\rangle}+\sum_{n=1}^{\infty}\frac{\langle zf,V_n^-\rangle\langle zg,V_n^-}{\langle zV_n^-,V_n^-\rangle}. \tag{B19}$$

Hence, combining Eqs. (B8), (B17), (B18), and (B19) we have

$$\mathscr{S}_1=-\frac{1}{2}\left[\sum_{n=1}^{\infty}\frac{\langle z^2,zV_n^-\rangle}{\langle zV_n^-,V_n^-\rangle}\langle 1,V_n^-\rangle+\sum_{n=1}^{\infty}\frac{\langle z^2,zV_n^+\rangle}{\langle zV_n^+,V_n^+\rangle}\langle 1,V_n^+\rangle\right]=\left\langle\frac{1}{2}(3-z^2),1\right\rangle=1. \tag{B20}$$

[1] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and Natural Sciences*, 2nd ed. (Springer-Verlag, New York, 1985).
[2] M. V. Evans, Adv. Chem. Phys. **83**, 361 (1992).
[3] A. Ikegami, Adv. Chem. Phys. **46**, 363 (1981).
[4] R. G. K. Habiger, H. Risken, M. James, F. Moss, and W. Schleich, Phys. Rev. A **41**, 3950 (1990).
[5] G. E. Uhlenbeck and L. S. Ornstein, Phys. Rev. **36**, 823 (1930); in *Selected Papers on Noise and Stochastic Processes* (Dover, New York, 1954).
[6] R. Short, R. Mandel, and R. Roy, Phys. Rev. Lett. **49**, 647 (1982).
[7] S. N. Dixit and P. S. Sahini, Phys. Rev. Lett. **50**, 1273 (1983).
[8] U. Zürcher, J. Chem. Phys. **103**, 4829 (1995).

[9] H. C. Tuckwell, J. Theor. Biol. **127**, 427 (1987).
[10] T. Naeh, M. M. Kłosek, B. J. Matkowsky, and Z. Schuss, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **50**, 595 (1990).
[11] H. Nakamura, Adv. Chem. Phys. **82**, 243 (1992).
[12] M. M. Kłosek, B. J. Matkowsky, and Z. Schuss, Phys. Rev. A **38**, 2606 (1988); SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **48**, 425 (1988).
[13] P. S. Hagan, C. R. Doering, and C. D. Levermore, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **49**, 1480 (1989).
[14] P. S. Hagan, C. R. Doering, and C. D. Levermore, J. Stat. Phys. **54**, 1321 (1989).
[15] P. S. Hagan, M. M. Kłosek, in preparation.
[16] K. Lindenberg and B. J. West, *The Nonequilibrium Statistical Dynamics of Open and Closed Systems* (VCH City, 1990), and references therein.
[17] M. Dykman and K. Lindenberg, in *Contemporary Problems in Statistical Physics* (SIAM, Philadelphia, 1994), and references therein.
[18] P. Hänggi and P. Jung, Adv. Chem. Phys. **89**, 239 (1995).
[19] G. Fichera, *Boundary Value Problems in Differential Equations* (University of Wisconsin Press, Madison, WI, 1960).
[20] L. Cao, D. Wu, and X. Luo, Phys. Rev. A **50**, 4244 (1994).
[21] J. Troe, Adv. Chem. Phys. **82**, 485 (1992).
[22] C. R. Doering, P. S. Hagan, and C. D. Levermore, Phys. Rev. Lett. **59**, 2129 (1987).
[23] M. M. Kłosek and P. S. Hagan, submitted for publication.
[24] L. Arnold, *Stochastic Differential Equations, Theory and Applications* (Wiley, New York, 1974).
[25] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes* (Academic, New York, 1981).
[26] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, D.C., 1964).
[27] R. Beals and V. Protopopescu, J. Stat. Phys. **32**, 565 (1983).
[28] V. Protopopescu, J. Phys. A **20**, L1239 (1987).

Variational principles for nonlinear dynamical systems

Vicenç Méndez
Grup de Física, Departament de Ciències Ambientals, Facultat de Ciències, Universitat de Girona, C/. Albareda 3-5, 1701 Girona, Catalonia, Spain

(Received 26 February 1997; accepted for publication 8 September 1997)

A variational method for Hamiltonian systems is analyzed. Two different variational characterization for the frequency of nonlinear oscillations is also supplied for non-Hamiltonian systems. © *1998 American Institute of Physics.* [S0022-2488(98)02702-9]

I. INTRODUCTION

The dynamical study of nonlinear oscillatory systems is often reduced to numerical calculus or to some approximated analytical techniques. Most of them are based in perturbation methods[1] as for instance the Poincaré–Lindsted method, the averaging (Krylov–Bogoliuvov–Mitropolsky) method, the Shohat expansion, or the multiscale method, to mention a few. They provide approximate solutions as well as the relationship between the frequency of the nonlinear oscillations and a small parameter (physically identified for the system, as the amplitude of the oscillations). Other approximate techniques consist in the qualitative analysis on the phase space by linearizing the differential equation around the fixed points. Periodic or exponential behavior of the system may be predicted after studying the stability of the fixed points.

Many different physical systems reduce to one-dimensional nonlinear ODE of second order, which may be studied by using the approximate techniques mentioned above.[2,3] An important property of them is that the frequency depends explicitly on the amplitude of the oscillations. As we have mentioned, the perturbative techniques yield to an approximate solution in power series of the amplitude only.

Recently, Benguria and Depassier (BD for short)[4] have made use of a variational principle in order to solve the bifurcation problem for second-order nonlinear Hamiltonian systems. They also apply the principle to obtain simple approximate closed formulas for the frequency of large-amplitude oscillations.

In this work we extend this method to other situations. Basically, the method consists in transforming the differential equation $\ddot{u}+u=N(u)$, which may be written in the phase space variables (p,u), with $p=-u'$ (where the prime means differentiation with respect to the variable $\tau=2\omega t/\pi$) in an integral equation multiplying the differential equation by an auxiliary function $g(u)$, and integrating. The variational method provide a lower bound for an integral functional. From the Euler–Lagrange equations for the functional, one may calculate the auxiliary function $g(u)$ at a first approximation by taking $p(u)$ from the linear case $(N=0)$.

First of all, we generalize this transformation for Hamiltonian systems. We multiply the differential equation by $p^n g$ and apply the variational method. From this generalization we obtain the first interesting result: the variational characterization depends strongly on the exponent n. In this sense, we obtain a better fit for $n=-1$ than for $n=0$ (the BD's case).

However, as the systems are Hamiltonian, it is possible to find a relationship between p and u in the nonlinear situation. Using it, instead of its linear approach, one can find a new relationship for g being the variational result for the frequency, always equal to the exact solution. This is shown in three specific and physically interesting cases.

Second, as non-Hamiltonian systems are very often found in practice and arise in many physical systems, we introduce in this work a special analysis for them. Following the same ideas employed for Hamiltonian systems, one obtains an explicit dependence of the variational characterization on $p(u)$. For non-Hamiltonian systems it is not possible to obtain a relationship between p and u, at least exactly. In consequence, we adopt two ways to proceed. From one hand, we apply twice the variational method and find two coupled Euler–Lagrange equations, one of them corresponding to the same functional that appears in Hamiltonian systems and the other one to the

 © 1998 American Institute of Physics

new one that contains $p(u)$. The consequent variational characterization does not depend now on $p(u)$. This method also apply for large amplitude. On the other hand, we find for a class of non-Hamiltonian systems in which the above method does not hold, a perturbative solution to $p(u)$ for small nonlinearities. As a special example we treat the Van der Pol oscillator.

II. VARIATIONAL PRINCIPLES FOR HAMILTONIAN SYSTEMS

In this section we generalize the BD's method and we show, by using the Duffing oscillator as an example, the generality of the principle. Starting from a nonlinear Hamiltonian dynamical system of the form

$$\ddot{u}+u=N(u), \quad \text{with } u(0)=a, \quad \dot{u}(0)=0, \tag{1}$$

this becomes

$$u''+\lambda u=N(u,\lambda), \tag{2}$$

where the prime symbol stands for the differentiation respect to the angular variable $\tau=2\omega t/\pi$, ω being the angular frequency of the nonlinear system (1) and $\lambda=(\pi/2\omega)^2$. We restrict our study to $u\in(0,a)$, where a is the initial amplitude of movement, so $u(\tau=1)=0$. Defining the variable $p=-u'>0$ we reduce the order of (2), and it is written as

$$p\,\frac{dp}{du}+\lambda u=N(u,\lambda). \tag{3}$$

Multiplying (3) by $p^n g(u)$, where $g(u)$ is an auxiliary function to be determined such that $g(0)=0$, and integrating, we obtain

$$\lambda\int_0^a up^n g(u)du=\int_0^a\left(p^n N(u,\lambda)g(u)+\frac{1}{n+2}\,p^{n+2}g'(u)\right)du. \tag{4}$$

Notice that for $n=0$ we recover the Benguria and Depassier case. We consider now, for a fixed g and $n>-2$, the functional

$$J_g[v]=\frac{(-1)^n}{n+2}\int_0^a (v')^{n+2}g'(v)dv=\frac{(-1)^{n+1}}{n+2}\int_0^1 (v')^{n+3}g'(v)d\tau, \tag{5}$$

defined for v such that it satisfies $v(0)=a$, $v(1)=0$, and $v'<0$ in $\tau\in(0,1)$. Defining $\phi(v,v')=(-1)^{n+1}(v')^{n+3}g'(v)/(n+2)$, the Euler–Lagrange equations for J_g may be integrated once to obtain $\phi-v'\,\partial\phi/\partial v'=\text{const}$, provided that ϕ does not depend explicitly on τ. From the Euler–Lagrange equation we obtain

$$(-1)^{n+1}(\tilde{v}')^{n+3}g'(\tilde{v})=K>0, \tag{6}$$

where $\tilde{v}$ is the value of v such that for a fixed g, $J_g\geqslant 0$ has a unique minimum. The auxiliary function g may be obtained by integrating (6). So, once g is determined, we get

$$J_g[v]\geqslant J_g[\tilde{v}]=\frac{K(a)}{n+2}, \tag{7}$$

where $K(a)$ is determined through the boundary condition on $\tilde{v}$. So, we have from (4), (5), and (7),

$$\lambda\geqslant\frac{\int_0^a N(u,\lambda)g(u)du+\dfrac{K}{n+2}}{\int_0^a up^n g(u)du}. \tag{8}$$

This constitutes the result of our first variational principle applied to nonlinear Hamiltonian oscillators. Notice the strong dependence of the variational principle on the exponent n, which we will show later in a specific example. For values of n lower than -2 this variational principle does not apply and for the other values the upper bound for the frequency depends explicitly on n.

If we take $n=0$ in (8), we recover the Benguria and Depassier situation,[4] where (8) reduces to

$$\lambda \geqslant \frac{\int_0^a N(u,\lambda)g(u)du + \frac{K}{2}}{\int_0^a ug(u)du}, \tag{9}$$

and (6) is $(u')^3 g'(u) = -K$. Taking $n=-1$, (8) becomes

$$\lambda \geqslant \frac{\int_0^a \frac{N(u,\lambda)g(u)}{p(u)}\,du + K}{\int_0^a \frac{ug(u)}{p(u)}\,du}, \tag{10}$$

and the Euler–Lagrange equation reduces to $(u')^2 g'(u) = K$.

Following Ref. 5, we can also derive, from another variational principle, a new restriction on the characteristic parameters of the system. Taking $n=0$ in (4), we may define the function

$$\psi(p) = \lambda f(u)g + \tfrac{1}{2}\, p^2 g'(u),$$

where $N(\lambda,u) = \lambda f(u)$. This function has, for each value of u, a minimum at $p_{\min}=0$, provided that $g'(u)>0$. So $\psi(p_{\min}) = \lambda f(u)g \leqslant \psi(p)$, and finally

$$\int_0^a ug(u)du \geqslant \int_0^a f(u)g(u)du. \tag{11}$$

This constitutes the second variational principle for Hamiltonian systems. Let us now apply both principles on the Duffing oscillator and the nonlinear pendulum.

A. The Duffing oscillator

It is known that the exact solution for the Duffing oscillator,[2]

$$\ddot{u} + u + \delta u^3 = 0, \quad \text{with } u(t=0)=a, \quad \dot{u}(t=0)=0, \tag{12}$$

is given by

$$u(t,\delta) = a\ \ \mathrm{cn}(t\sqrt{1+\delta a^2}; k), \tag{13}$$

where cn is the Jacobi elliptic function and $k^2 \equiv \delta a^2/2(1+\delta a^2)$. The exact expression for the frequency is then

$$\omega_{\mathrm{ex}} = \frac{\pi}{2}\sqrt{1+\delta a^2}\,\frac{1}{F\left(\frac{\pi}{2},k\right)}, \tag{14}$$

where $F(k)$ is called the complete elliptic integral of the first order. Taking $n=0$ in (8) we recover the BD situation,[4] which yields

$$\omega_{n=0} \leqslant \sqrt{1+\frac{3}{4}\,\delta a^2}. \tag{15}$$

Let us show in this example and in the following one the dependence of the frequency on the value of n taken in (8). First, we analyze the linear case, for the Duffing oscillator, and compare the results obtained for $n=0$ and for $n=-1$. In this case, (10) leads to

$$\lambda \geqslant \frac{K}{\int_0^a \frac{ug(u)}{p(u)}\,du}, \tag{16}$$

K and $g(u)$ must be calculated from

$$(u')^2 g'(u) = K. \tag{17}$$

In the linear case ($N=0$), corresponding to the linear oscillator, one has $(u')^2=\lambda(a^2-u^2)$. Replacing this in (17), we obtain

$$g(u) = \ln\left(\frac{a+u}{a-u}\right), \tag{18}$$

as the expression for the auxiliary function. Solving (17) with $u(0)=a$, $u(1)=0$, one obtains $K=a\pi^2/2$. On the other hand, defining $x \equiv u/a$, one has, from (18),

$$\int_0^a \frac{ug(u)}{p(u)}\,du = \frac{a}{\sqrt{\lambda}} \int_0^1 \ln\left(\frac{1+x}{1-x}\right) \frac{x\,dx}{\sqrt{1-x^2}} = \frac{a\pi}{\sqrt{\lambda}}.$$

Finally, from (16) we find $\lambda \geqslant (\pi/2)^2$, which is the same result obtained by BD.[4] For the nonlinear case ($N \neq 0$), we find a different result, as we show in turn. The main difference between the cases $n=-1$ and $n=0$ is that, to solve the latter, one does not need the relationship between p and u, while for $n=-1$ this is necessary. For the Duffing oscillator, we find, by making use of the Hamiltonian as a constant of motion, that

$$p = \sqrt{\lambda a^2\left(1+\frac{\delta a^2}{2}\right)}\sqrt{(1-x^2)(sx^2+1)}, \tag{19}$$

where $s \equiv \delta a^2/(2+\delta a^2)$. From (10) and defining the following integrals:

$$I_1(s) = \int_0^1 \ln\left(\frac{1+x}{1-x}\right) \frac{x\,dx}{\sqrt{(1-x^2)(sx^2+1)}},$$

$$I_1(s) = \int_0^1 \ln\left(\frac{1+x}{1-x}\right) \frac{x^3\,dx}{\sqrt{(1-x^2)(sx^2+1)}},$$

we obtain

$$\lambda \geqslant \left(\frac{\pi^2}{2}\right)\left(1+\frac{\delta a^2}{2}\right) \frac{1}{(I_1+\delta a^2\, I_2)^2}. \tag{20}$$

The integrals I_1 and I_2 may be calculated, after some integrations by parts, in terms of the complete elliptic integral of the second kind, and we find finally for the frequency that

$$\omega_{n=-1} \leqslant \frac{2}{\pi}\sqrt{1+\delta a^2} E\left(\frac{\pi}{2}, \sqrt{\frac{s}{s+1}}\right). \tag{21}$$

In Fig. 1 we represent the exact value ω_{ex} given by (14), the BD solution $\omega_{n=0}$ given by (15) in solid lines, and the new approach $\omega_{n=-1}$ given by (21) in dashed lines. As one can see, the

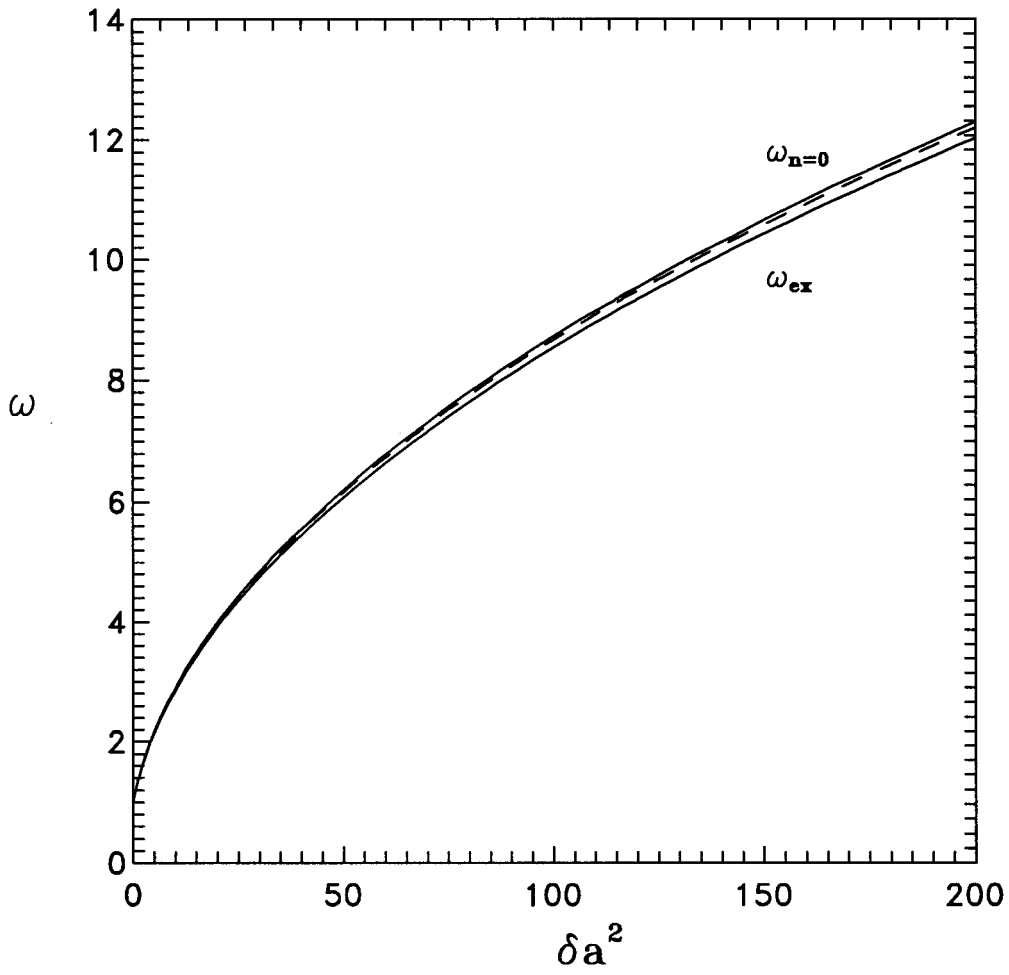

FIG. 1. The frequency of the Duffing oscillator obtained by the BD method ($\omega_{n=0}$, solid lines), the new variational method ($\omega_{n=-1}$, dashed lines) versus the numerical exact solution (ω_{ex}, solid line).

latter is a slightly better result than the BD solution. So, we have shown the dependence of the frequency on the value of n. Applying now the second variational method (11) to the Duffing equation, and taking the BD method we find the following restriction:

$$\frac{-4}{3a^2} \leqslant \delta.$$

For $\delta > 0$ the second variational method does not add any additional information, but for $\delta < 0$ this may be understood[3] as a bound for δ, that is, $\delta \leqslant 4/3a^2$. For $\delta < 0$ we have that

$$\omega_{n=0} \leqslant \sqrt{1 - \frac{3}{4}\,\delta a^2}.$$

The frequency has real value if $\delta \leqslant 4/3a^2$, which coincides with the result obtained by using the second variational method.

B. The nonlinear pendulum

We study with some detail another very known nonlinear dynamical system namely the nonlinear pendulum. First of all we show that the $n = -1$ case may lead us to the exact solution. Later, we apply the BD approach to this solution. As it is widely known, the equation for the nonlinear pendulum is given by

$$\ddot{u}+\frac{g}{l}\sin(u)=0, \quad \text{with } u(t=0)=a, \quad \dot{u}(t=0)=0. \tag{22}$$

Defining again $\tau=2\omega t/\pi$ and

$$\lambda\equiv\left(\frac{\pi}{2}\right)^2\frac{g}{l}, \tag{23}$$

and taking into account the conditions $u(t=0)=a$ and $u(t=\pi/2\omega)=0$, (22) is written as

$$u''+\lambda\ \sin(u)=0, \quad \text{with } u(\tau=0)=a, \quad u'(\tau=1)=0.$$

The exact solution[2] is given by

$$\omega_{\text{ex}}=\frac{\pi}{2}\ \sqrt{\frac{g}{l}}\ \frac{1}{F\left(\frac{\pi}{2},\sin\left(\frac{\alpha}{2}\right)\right)}. \tag{24}$$

Let us now to apply the $n=-1$ approach. The variational method leads to

$$\lambda\geqslant\frac{K}{\int_0^a\frac{g(u)}{p(u)}\sin(u)du}, \tag{25}$$

with

$$(u')^2g'=K. \tag{26}$$

As (22) may be integrated once, we get

$$p(u)=\sqrt{2\lambda[\cos(u)-\cos(a)]}. \tag{27}$$

So, omitting overall multiplicative constants we identify

$$g'(u)=\frac{1}{\cos(u)-\cos(a)},$$

and integrating, we find for the auxiliary function

$$g(u)=\frac{1}{\sin(a)}\ln\left[\frac{\sin(a)\tan\left(\frac{u}{2}\right)+1-\cos(a)}{\sin(a)\tan\left(\frac{u}{2}\right)-1+\cos(a)}\right]. \tag{28}$$

The value of K is found to be

$$K=2F^2\left(\frac{\pi}{2},\sin\left(\frac{a}{2}\right)\right), \tag{29}$$

from (26). Calculating the integral in (25), we obtain

$$\int_0^a\frac{g(u)}{p(u)}\sin(u)du=\frac{2}{\sqrt{\lambda}}F\left(\frac{\pi}{2},\sin\left(\frac{a}{2}\right)\right),$$

after some integrations by parts. Finally, from (23), (25), and (29), one gets

$$\omega_{n=-1} \leqslant \frac{\pi}{2} \sqrt{\frac{g}{l}} \frac{1}{F\left(\frac{\pi}{2}, \sin\left(\frac{a}{2}\right)\right)},$$

which coincides within the exact result. If we had made some approximations, as, for instance, take for $g(u)$ or $p(u)$ the same value as for the linear oscillator, the final result had not been equal to the exact one. We show this now for $n=0$. Equation (8) writes in the BD case,

$$\lambda \geqslant \left(\frac{\pi}{2}\right)^2 + \frac{4}{\pi a^3} \int_0^a \frac{uN(u,\lambda)du}{\sqrt{1-u^2/a^2}}, \tag{30}$$

where we take $N(u,\lambda)=\lambda(u-\sin(u))$. Calculating the integral in (30) and from the definition of λ, one obtains

$$\omega_{n=0} \leqslant \sqrt{\frac{g}{l}} \sqrt{\frac{2}{a} J_1(a)}, \tag{31}$$

where $J_1(a)$ is the Bessel function of the first kind. Because $J_1(a)$ alternates the sign, we must restrict the solution given by (31) to those values of a such that $J_1(a)$ is positive. For small values of a this condition requires $a \lesssim 3$. Precisely this condition may be recovered from the second variational principle. Taking $f(u)=u-\sin(u)$ and (28) we find from (11) $J_1(a) \geqslant 0$. In Fig. 2 we plot the exact solution and the variational solution for the BD's method.

C. Systems of the form $\ddot{u}+\alpha u^n=0$

We consider now kinds of dynamical systems, which, in their linear approximation ($N=0$) do not reduce to linear oscillators. These systems may also be treated as the previous ones. First we apply the variational methods and we compare the solution with the exact one. We make, for mathematical simplicity the analysis by using the particular case[6] $n=3$. It is known that this system describes periodic oscillations.[2] The system may be integrated once by using the same initial conditions as in previous examples. Hence, we get, after using the independent variable τ,

$$p(u) = \sqrt{\frac{\alpha\lambda}{2}} \sqrt{a^4-u^4}, \tag{32}$$

and this one may be integrated once again to get

$$\omega_{\mathrm{ex}} = \frac{\pi}{2} a\sqrt{\alpha} \frac{1}{F\left(\frac{\pi}{2}, \frac{1}{\sqrt{2}}\right)}. \tag{33}$$

Applying the first variational method (10) (with $n=-1$) one obtains

$$\alpha\lambda \geqslant \frac{K}{\int_0^a \frac{u^3 g(u)}{p(u)} du} \tag{34}$$

where $(u')^2 g' = K$. So, the auxiliary function $g(u)$ is given by

$$g(u) = \frac{1}{2} \ln\left(\frac{a+u}{a-u}\right) + \arctan\left(\frac{u}{a}\right), \tag{35}$$

and K by

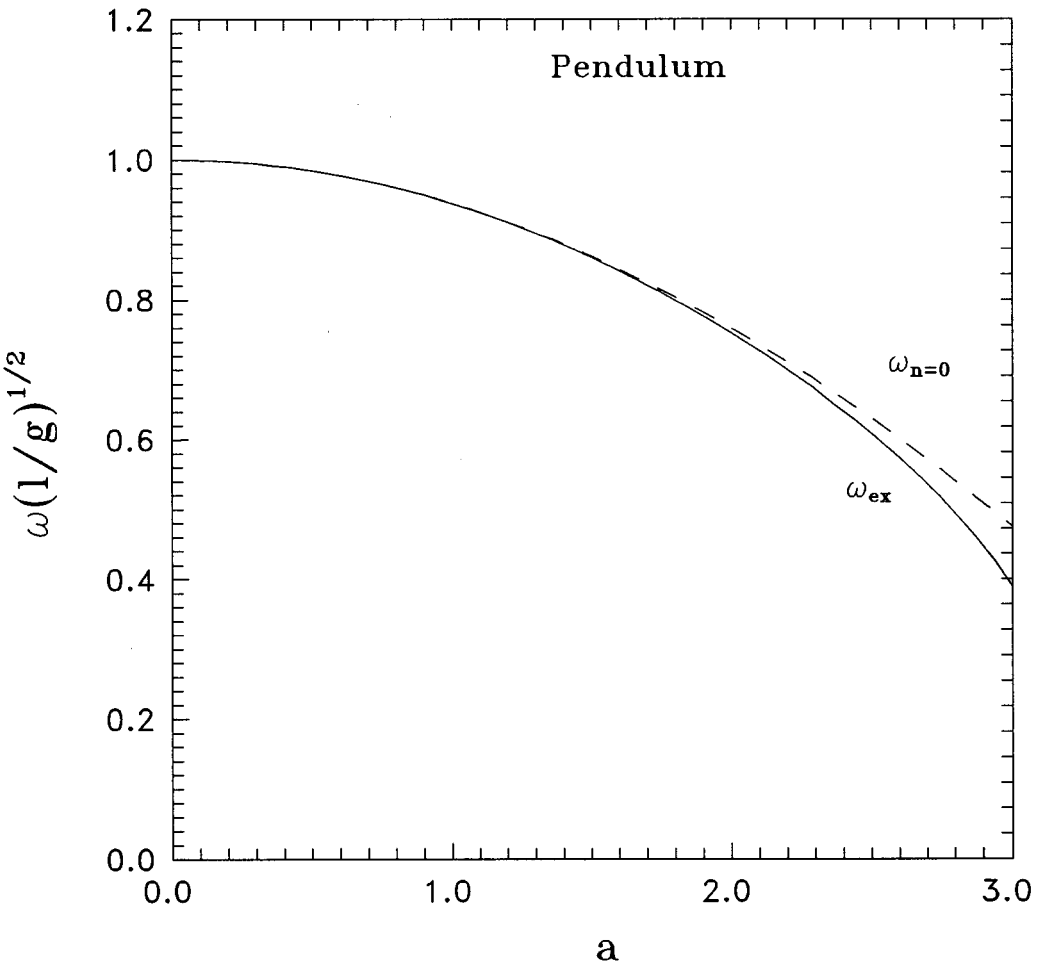

FIG. 2. The frequency of the nonlinear pendulum obtained by the BD method ($\omega_{n=0}$, dashed lines), versus the numerical exact solution (ω_{ex}, solid line).

$$K(a)=aF^2\left(\frac{\pi}{2},\frac{1}{\sqrt{2}}\right).$$

The integral involved in (34) is given by

$$\int_0^a \frac{u^3 g(u)}{p(u)}\,du=a^2\sqrt{\frac{2}{\alpha\lambda}}\int_0^1 \frac{x^3\,dx}{\sqrt{1-x^4}}\left[\frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)+\arctan(x)\right]=\frac{a^2}{\sqrt{\alpha\lambda}}F\left(\frac{\pi}{2},\frac{1}{\sqrt{2}}\right).$$

Using the previous calculations, we may write (34) in terms of the frequency by

$$\omega_{n=-1}\leqslant\frac{\pi}{2}\,a\sqrt{\alpha}\,\frac{1}{F\left(\frac{\pi}{2},\frac{1}{\sqrt{2}}\right)},$$

where the equality gives the exact result (33). On the other hand, we may also apply the BD's method to this case. Taking in (2) $N=\lambda(u-u^3)$ and $n=0$ is (8) we find $K=(\pi/2)^3a^3$ and the auxiliary function is given by

$$g(u)=\frac{u}{\sqrt{1-u^2/a^2}}. \tag{36}$$

We finally find for the frequency

$$\omega_{n=0}\leqslant\frac{a}{2}\sqrt{3\alpha}. \tag{37}$$

Calculating numerically $\omega_{n=-1}(=\omega_{ex})$ we find $\omega_{\mathrm{ex}}=0.847a\sqrt{\alpha}$ and $\omega_{n=0}\leqslant 0.866a\sqrt{\alpha}$. We may observe the good agreement between the exact and the BD's solutions.

III. DYNAMICAL SYSTEMS DEPENDING ON THE FIRST DERIVATIVE

We extend here the variational principles found in the first section to nonlinear dynamical systems that contains explicitly the term $\dot{u}$ and powers of it. Some of them are Hamiltonian systems, that is, they have a first integral, and others are dissipative, such as the Van der Pol's equation. We focus our attention to systems of the form

$$\ddot{u}+\epsilon f(u)\dot{u}^n+u=0. \tag{38}$$

This equation may be written as

$$p\,\frac{dp}{du}+\lambda u+\epsilon\lambda^{(2-n)/2}f(u)p^n=0. \tag{39}$$

By using the change of variables $\tau=2\omega t/\pi$ and the definition of λ, we may arrive to

$$\lambda\geqslant\frac{\epsilon(-1)^{n+1}\lambda^{(2-n)/2}\int_0^a f(u)g(u)p^n du+\frac{K}{2}}{\int_0^a ug(u)du}, \tag{40}$$

with $(u')^3g'(u)=-K$. Notice that in the variational method given by (40) we must know the explicit expression of $p(u)$ and this is only possible if the system admits a first integration. In general, one must use a new approach to (40). In this sense, we derive two new ways to proceed; one of them consists in applying the variational method on the functional of the numerator of (40) but this method only holds if $\int(u)>0$ for any $u\in(0,a)$, the other one consists in making a perturbative expansion on the phase space in order to find an approximate solution for $p(u)$. Let us illustrate the first way.

From (40) we define the functional

$$J_g^*[v]=(-1)^{n+1}\int_0^a f(v)g(v)p^n\,dv=\int_0^1 f(v)g(v)(v')^{n+1}\,d\tau, \tag{41}$$

where v is such that satisfies $v(0)=a$, $v(1)=0$, and $v'<0$ in $\tau\in(0,1)$. Defining $\phi(v,v')=f(v)g(v)(v')^{n+1}$, we get from the Euler–Lagrange equations,

$$nf(\tilde{v})g(\tilde{v})(\tilde{v}')^{n+1}=K^* \tag{42}$$

where $\tilde{v}$ is the value of v such that for a fixed g, $J_g\geqslant 0$ has a unique minimum. Notice that $f(\tilde{v})\geqslant 0$ is required for any $\tilde{v}\in(0,a)$, and this is not possible for the Van der Pol's equation. So, once g is determined from $(\tilde{v}')^3g'(\tilde{v})=-K$, we get

$$J_g[v]\geqslant J_g[\tilde{v}]=\frac{K^*(a)}{n}, \tag{43}$$

and $K^*(a)$ may be calculated from (42). Thus, from (40), we get the variational restriction

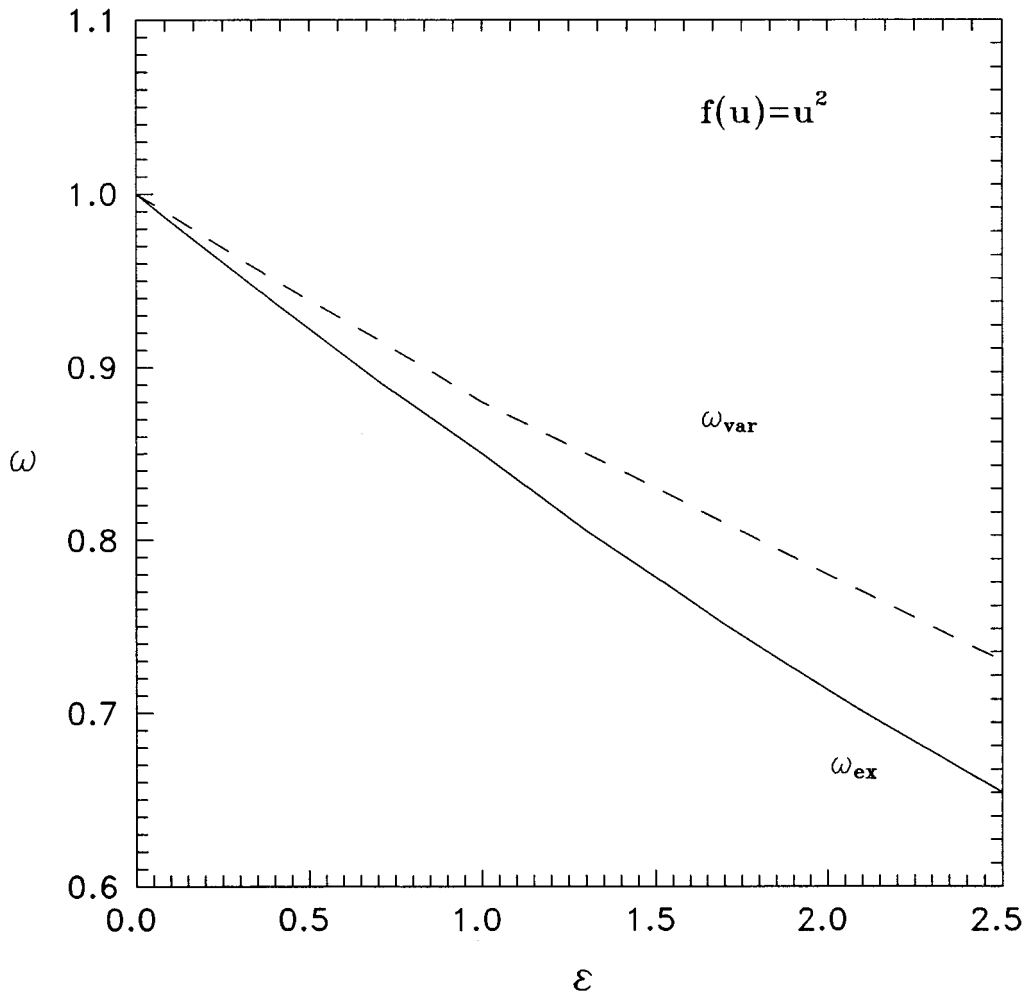

FIG. 3. The frequency of a non-Hamiltonian oscillator given in (46) obtained by the variational method (ω_{var}, dashed lines), versus the numerical exact solution (ω_{ex}, solid line).

$$\lambda \geqslant \frac{\epsilon \lambda^{(2-n)/2} \dfrac{K^*}{n} + \dfrac{K}{2}}{\int_0^a u g(u) du}. \tag{44}$$

Taking the auxiliary function g given in the linear case $\epsilon = 0$ [Eq. (36)], Eq. (44) becomes

$$\lambda \geqslant \left(\frac{\pi}{2}\right)^2 + \frac{4\epsilon}{\pi a^3} \lambda^{(2-n)/2} \frac{K^*}{n}. \tag{45}$$

Let us to apply this variational method to the specific case

$$\ddot{u} + \epsilon u^2 \dot{u} + u = 0, \quad \text{with} \quad u(t=0) = a, \quad \dot{u}(t=0) = 0. \tag{46}$$

In this case, $f(u) = u^2$ and $n = 1$. From (42) we obtain $K^* = \pi^2 a^5/32$, and from (45),

$$\omega_{var} \leqslant \frac{-b + \sqrt{b^2+4}}{2}, \quad \text{with} \quad b = \frac{a^2 \epsilon}{4}. \tag{47}$$

In Fig. 3 we plot the exact numerical solution for the frequency versus its variational solution (47). We develop now the second variational method for the specific situation in which $f < 0$ for

some $u \in (a,0)$. Starting from (40) the problem is to find an expression for $p(u)$. As a first approximation, we may take the expression for p given by the linear case. In this approach we obtain

$$\lambda \geqslant \left(\frac{\pi}{2}\right)^2 + \frac{4\epsilon(-1)^{n+1}}{\pi a^2}\lambda \int_0^a u(a^2-u^2)^{(n-1)/2} f(u)du. \tag{48}$$

A better approach may be obtained by the perturbative solution in the phase space. Solving perturbatively (39) for small ϵ, we get

$$p(u) = p_0(u,\lambda) + \epsilon p_1(u,\lambda) + O(\epsilon^2),$$

where

$$p_0(u,\lambda) = \sqrt{\lambda(a^2-u^2)},$$

$$p_1(u,\lambda) = -\frac{\lambda^{(5-n)/2}}{\sqrt{a^2-u^2}} \int_a^u (a^2-z^2)^2 f(z)dz.$$

So, the variational method yields

$$\lambda \geqslant \left(\frac{\pi}{2}\right)^2 - \frac{4\epsilon}{\pi a^2} \int_0^a \frac{u f(u)(p_0+\epsilon p_1)^n}{\sqrt{a^2-u^2}}\, du. \tag{49}$$

Both approaches are applied as an illustration to the Van der Pol oscillator.

Taking $n=1$ and $f(u)=u^2-1$ in (38), we recover the well-known differential for the Van der Pol oscillator,

$$\ddot{u} + \epsilon(u^2-1)\dot{u} + u = 0, \tag{50}$$

with the initial conditions $u(t=0)=a$, $\dot{u}(t=0)=0$. It is known that (50) presents a limit cycle for $a=2$. Thus, we study the periodic behavior, and the frequency, in particular, of (50) near of the limit cycle given by $a=2$. In the first approximation (assuming g and p are given by the linear case, $\epsilon=0$), given by (48) we find for the frequency

$$\omega_1 \leqslant \sqrt{1-\frac{2\epsilon}{\pi}}. \tag{51}$$

Using the second approximation, that is, p given up to first order in ϵ, we obtain

$$p_1(u) = -\sqrt{\lambda}\,\frac{u}{4}(4-u^2),$$

and finally

$$\omega_2 \leqslant \sqrt{1-\frac{2\epsilon}{\pi}+\frac{\epsilon^2}{4}}. \tag{52}$$

We may solve numerically the Van der Pol's equation for different ϵ and obtain ω_{ex}. We plot in Fig. 4 the exact numerical solution and the ω_2 solution for the frequency.

IV. CONCLUSIONS

A variational principle applied to Hamiltonian systems has been developed by Benguria and Depassier.[4] In this paper we extend their method along the following new lines:

(i) The initial transformation, for Hamiltonian systems, which transforms the differential equation in an integral equation, is generalized by introducing an exponent n. The final variational

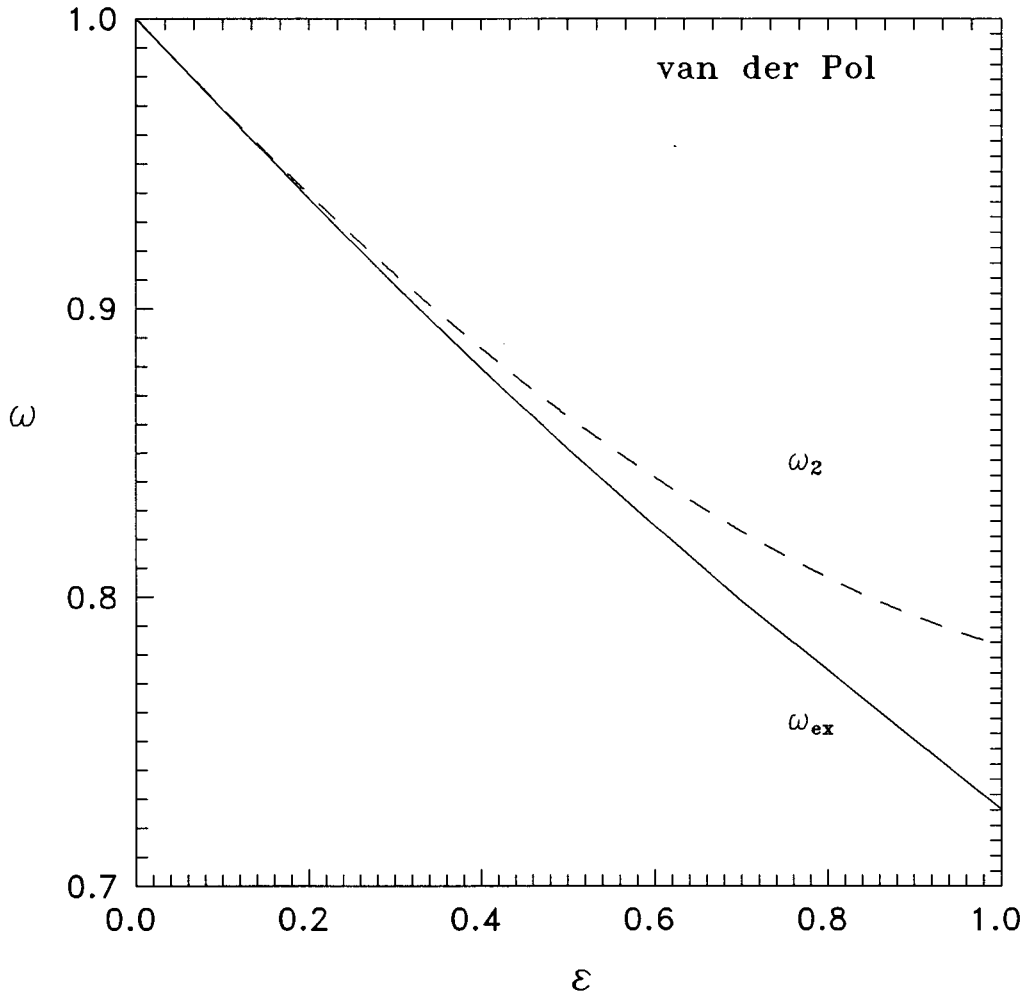

FIG. 4. The frequency of the Van der Pol oscillator obtained by the variational method (ω_2, dashed lines), versus the numerical exact solution (ω_{ex}, solid line).

restriction that relates the frequency with the amplitude, depends strongly on n. The variational method proposed by BD only holds for $n>-2$. A second variational method is also applied to Hamiltonian systems and it supplies new constraints between the characteristic parameters involved in the system. We have specified the results to some selected systems as the Duffing oscillator, the nonlinear pendulum, and systems of the form $\ddot{u}+\alpha u^n=0$. Better fits those obtained by the method of BD, to the exact solution for the frequency have been discovered.

(ii) The variational principle is also applied to systems with an explicit dependence on the first time derivative. Some of them are Hamiltonian, and they may be exactly characterized. Others are non-Hamiltonian and we may proceed in two different ways. First, we have proposed a new variational method that is applied to two different functionals in order to avoid the explicit dependence on $p(u)$. This method does not hold for systems with $f<0$ for some $u\in(0,a)$. So, we develop an approximate variational method for them. The results are not as good as those obtained for Hamiltonian systems, but they may be understood as upper bounds on the frequency. This approximation consists in finding a perturbative solution for p and the results fit very well for weak nonlinearities.

ACKNOWLEDGMENTS

I dedicate this paper to Professor José Casas-Vázquez on his 60th birthday. I also would like to thank Professor David Jou for helpful and stimulating discussions. I acknowledge the support of the program Formación de Personal Investigador under Grant No. CLI95-1867 and the Spanish Ministry of Education and Science under Grant No. PB94-0718.

[1] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics* (Springer-Verlag, NY, 1981).
[2] R. E. Mickens, *An Introduction to Nonlinear Oscillators* (Cambridge University Press, Cambridge, 1981).
[3] R. Grimshaw, *Nonlinear Ordinary Differential Equations*, Applied Mathematical Texts and Engineering Science texts (Blackwell Scientific Publications, Nottingham, Great Britain, 1990).
[4] R. D. Benguria and M. C. Depassier, Phys. Rev. Lett. **77**, 2847 (1996).
[5] R. D. Benguria and M. C. Depassier, Phys. Rev. Lett. **77**, 1171 (1996).
[6] P. Vainshtein, G. Ziskind, and M. Fichman, Phys. Rev. Lett. **78**, 551 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Distortion of a phase space under the Darboux transformation

Boris F. Samsonov[a)]
Tomsk State University, 36 Lenin Avenue, 634050 Tomsk, Russia

(Received 21 April 1997; accepted for publication 23 September 1997)

The Darboux transformation operator technique is applied to construct exactly solvable anharmonic singular oscillator potentials and to study their coherent states. Classical system corresponding to a transformed quantum system is constructed with the help of the coherent states technique. It is shown that at classical level the Darboux transformation may be treated as a transformation of Kähler potential, which leads to a distortion of the initial phase space. © *1998 American Institute of Physics.* [S0022-2488(98)01601-6]

I. INTRODUCTION

In quantum mechanics there exists a beautiful recipe to construct a family of exactly solvable Hamiltonians starting from an initial Hamiltonian h_0 for which the solutions of the Schrödinger equation $h_0\psi_E=E\psi_E$ are known. Since this method is very simple in applications and possesses an intrinsic charm, it attracts a thorough attention of several generations of research. A central point of this theory is a notion of a transformation operator introduced and investigated by Delsart and Lions.[1] According to their approach, every operator L intertwining the initial Hamiltonian h_0 and transformed Hamiltonian h_1,

$$Lh_0=h_1L, \tag{1}$$

is a *transformation operator* (originally called a transmutation operator[1]). It is obvious that L transforms solutions ψ_E ($\psi_E\notin\ker L$) of the initial Schrödinger equation onto solutions $\varphi_E=L\psi_E$ of a *transformed Schrödinger equation* $h_1\varphi_E=E\varphi_E$, corresponding to the same eigenvalue E. In the case of a first-order differential operator L and a one-dimensional Hamiltonian h_0, the relation (1) leads to[2] the transformation proposed by Moutard[3] and studied by Darboux[4] (also see Ref. 5). If L is a differential operator of order N, it is called[2] *N-order Darboux transformation operator*. In this paper we will restrict ourselves by first-order transformations.

The Darboux transformation operator technique is a very useful tool in obtaining and investigating new exactly solvable quantum problems. It enters many fields of physics and mathematics. In soliton theory, the Darboux transformation permits one to construct and investigate properties of multisoliton solutions to nonlinear evolution equations (see Ref. 6 and references therein). In quantum mechanics this technique takes the form of a factorization method. Using this method, Schrödinger[7] was able to solve the hydrogen atom problem in a pure algebraic manner. Subsequently, Infeld and Hall[8] generalized the Schrödinger method and obtained a wide class of solvable potentials. In the inverse quantum scattering problem,[9] this transformation is used to construct potentials with given properties and to obtain appropriate solutions to the Schrödinger equation. New developments in this field have recently been presented in Ref. 10. In the context of supersymmetric quantum mechanics (for reviews see Ref. 11), the Hamiltonian h_0 and its superpartner h_1 form a matrix-valued Hamiltonian (so-called super-Hamiltonian) $\mathscr{H}$. Darboux transformation operator L and its Hermitian conjugate L^+ participate in the construction of supercharge operators, which, together with $\mathscr{H}$, form a superalgebra. Initially the supersymmetric quantum mechanics served as a toy model to illustrate the problem of supersymmetry breaking in quantum field theories. Recently, new features in supersymmetry breakdown in quantum mechanics caused by

a)Electronic mail: samsonov@phys.tsu.tomsk.su

 © 1998 American Institute of Physics

higher-order Darboux transformations have been discovered.[12] Besides quantum field theories, the same ideas find applications in many others fields of theoretical physics such as nuclear, atomic, solid state, and statistical physics.[11]

In spite of the long period of application of the Darboux transformation operator method to quantum mechanical problems, the question of correspondence between the transformed quantum system and a classical system has not been properly discussed in the literature. In a recent paper[13] a classical mechanics counterpart of the transformed system has been described. It consists in addition of a total time derivative of a purely imaginary function to the Lagrangian of a classical system. This transformation does not affect the equations of motion. Canonical quantization of the new classical system gives the transformed quantum system.

The concept of coherent states plays an important role in several fields of physics and mathematics (see, for example, Refs. 14 and 15). In this paper we exploit the property of coherent states to establish the correspondence between the quantum mechanics and the classical mechanics. For any quantum system we can construct using the coherent states technique,[14,16] a classical system possessing the following remarkable property. The geometric quantization[17] of the resulting system gives a holomorphic representation of the initial quantum system. By these means we may establish a correspondence between quantum and classical systems. Recently it has been noted[18] that an application of the Darboux transformation operator to the coherent states of the initial quantum system gives a set of states that may be treated as coherent states of the transformed system. This observation opens the door to establishing a correspondence between transformed quantum and classical systems and obtaining a classical counterpart of the Darboux transformation.

In this paper we realize the above program for a system with the Hamiltonian,

$$h_0=-d^2/dx^2+x^2/4+b/x^2, \quad b>0, \quad x\in[0,\infty). \tag{2}$$

We will not discuss the significance of the spectral problem for the Hamiltonian (2) (more precisely, for its closure in the Hilbert space H of square integrable on interval $[0,\infty)$ functions with the zero boundary condition at the origin). We only notice the fact that this Hamiltonian appears in different physical problems such as the central potential problem, the Calogero problem,[19] fractional statistics, and anyons,[20] two-dimensional QCD,[21] and others.

We obtain exactly solvable partners for the Hamiltonian (2). Next, we construct the coherent states for the transformed *a lá* Darboux Hamiltonians. These states, as well as the coherent states for the input system (initial coherent states), are labeled by the points of a unite disk on the complex z plane. We consider two different Kähler structures on the unit disk. The first one is the well-known Poincaré model of the Lobachevsky plane and the second is the Darboux transformation of the first. The expression for the transformed Kähler potential is obtained. We also notice that in contrast to the initial manifold the transformed one has a varying curvature. The integration measure realizing the identity decomposition for the transformed coherent states is calculated. The holomorphic representation for the discrete basis set as well as for the coherent states, lowering and raising operators and Darboux transformation operators is constructed. It is established that at the classical level the Darboux transformation may be considered as such a transformation of the Kähler potential and Hamilton function that preserves the equations of motion.

II. COHERENT STATES OF SINGULAR OSCILLATOR

In this section we briefly review the well-known[14] properties of the coherent states for the Hamiltonian (2) that we will need further.

The dynamical symmetry algebra for the quantum system with the Hamiltonian (2) is su(1,1). In coordinate representation, the generators of this algebra are expressed in terms of the harmonic oscillator annihilation a and creation a^+ operators,

$$k_0=\tfrac{1}{2}h_0, \quad k_+=\tfrac{1}{2}[(a^+)^2-b/x^2], \quad k_-=\tfrac{1}{2}[a^2-b/x^2], \quad a=d/dx+x/2, \quad a^+=-d/dx+x/2,$$

and satisfy the standard commutation relations,

$$[k_0,\ k_\pm]=\pm k_\pm, \quad [k_-,\ k_+]=2k_0.$$

Since we consider an irreducible representation the corresponding su(1,1) Casimir operator takes the constant value $\mathscr{C}=1/2[k_+k_-+k_-k_+]-k_0^2=3/16-b/4=k(1-k)$. The value of $k=1/2+(1/4)\sqrt{1+4b}$ defines the ground state (vacuum) energy $E_0=2k$. The vacuum state $|0\rangle$ is defined by the equations $k_-|0\rangle=0$ and $k_0|0\rangle=k|0\rangle$. Other discrete basis eigenfunctions $|n\rangle$ of the Hamiltonian h_0 are defined by means of the raising operator k_+,

$$k_+|n\rangle=-\sqrt{(n+1)(n+2k)}|n+1\rangle,$$

$$|n\rangle=(-1)^n\sqrt{\frac{\Gamma(2k)}{n!\Gamma(2k+n)}}\,(k_+)^n|0\rangle,\quad k_0|n\rangle=(k+n)|n\rangle,$$

and in the coordinate representation they have the form

$$\psi_n(x)=\langle x|n\rangle=[n!2^{1-2k}\Gamma^{-1}(n+2k)]^{1/2}x^{2k-1/2}\exp(-x^2/4)L_n^{2k-1}(x^2/2), \tag{3}$$

where $L_n^{\alpha}(z)$ are the Laguerre polynomials.

The coherent states $|z\rangle$ can be obtained by applying the unitary displacement operator D_z to the vacuum state,

$$|z\rangle=D_z|0\rangle=e^{zk_+}\exp[\ln(1-z\overline{z})k_0]e^{-\overline{z}k_-}|0\rangle=N_{0z}\sum_{n=0}^{\infty}a_nz^n|n\rangle,$$

$$N_{0z}=(1-z\overline{z})^k,\quad a_n=(-1)^n\sqrt{\frac{\Gamma(2k+n)}{n!\Gamma(2k)}},\quad |z|<1.$$

In the coordinate representation they look like

$$\psi_z(x)=2^{1/2-k}\Gamma^{-1/2}(2k)(1-z)^{-2k}(1-z\overline{z})^kx^{2k-1/2}\exp[-(1/4)(1-z)^{-1}(1+z)x^2],\quad |z|<1. \tag{4}$$

The linear manifold spanned by the vectors $\{|z\rangle\}$ forms an everywhere dense set in the Hilbert space H with the following identity decomposition:

$$\int_{|z|<1}|z\rangle\langle z|d\mu(z)=1,\quad d\mu(z)=\frac{2k-1}{\pi}\,(1-z\overline{z})^{-2}\,dz\,d\overline{z}. \tag{5}$$

The system $\{|z\rangle\}$ is overcomplete. The completeness of different subsystems of this system was studied in detail in Ref. 22.

The Fourier coefficients c_n of a vector $|\psi\rangle\in H$ with respect to the basis $\{|n\rangle\}$ define a holomorphic representation $\psi(z)$ of the vector $|\psi\rangle$ in the space of functions that are holomorphic in the unit disk,

$$\langle\overline{z}|\psi\rangle=N_{0z}\psi(z),\quad \psi(z)=\sum_{n=0}^{\infty}a_nc_nz^n.$$

The functions $\psi_n(z)=a_nz^n$ realize the holomorphic representation of the basis vectors $|n\rangle$. For the holomorphic representation of the coherent state vectors we have

$$\zeta^{(0)}(z)=N_{0z}^{-1}\langle\overline{z}|\zeta\rangle=(1-\zeta\overline{\zeta})^k(1-\zeta z)^{-2k}.$$

Let $\mathscr{L}$ be the lineal of entire analytic functions $\psi(z)$ defined in the unit disk, such that

$$\int_{|z|<1}|\psi(z)|^2(1-|z|^2)^{2k}\,d\mu(z)<\infty.$$

If now we define an inner product in the lineal $\mathscr{L}$,

$$\langle \psi_1(z)|\psi_2(z)\rangle \equiv \int_{|z|<1} e^{-f^{(0)}} \overline{\psi_1}(z)\psi_2(z)d\mu(z) = \langle \psi_1|\psi_2\rangle, \tag{6}$$

where $\langle \psi_1|\psi_2\rangle$ is the inner product in the Hilbert space H then the completion of $\mathscr{L}$ with respect to this inner product gives the Hilbert space $\mathscr{H}$. The function $f^{(0)}=\ln|\langle 0|z\rangle|^{-2}=\ln(1-z\overline{z})^{-2k}$ is the Kähler potential in the unit disk. The corresponding metric has the form

$$ds^2 = g\ dz\ d\overline{z}, \quad g=\frac{\partial^2 f^{(0)}}{\partial z\ \partial \overline{z}}=\frac{2k}{(1-z\overline{z})^2}.$$

As usually the imaginary part of this metric defines a symplectic 2-form $\omega=-ig\ dz\wedge d\overline{z}$ and consequently a Poisson bracket $\{H_1,H_2\}$ of the functions H_1 and H_2 defined on the unit disk. The unit disk represents now a phase space of a constant Gauss curvature $\mathscr{K}^{(0)}=-2/k$. Making use of the covariant Berezin symbols,[23] we establish a one-to-one correspondence between the classical observables and the holomorphic representation of the quantum mechanical operators. For example, the following classical observables:

$$K_0=\langle \overline{z}|k_0|\overline{z}\rangle = k\,\frac{1+z\overline{z}}{1-z\overline{z}},$$

$$K_-=\langle \overline{z}|k_-|\overline{z}\rangle=\frac{2k\overline{z}}{1-z\overline{z}}, \quad K_+=\langle \overline{z}|k_+|\overline{z}\rangle=\frac{2kz}{1-z\overline{z}},$$

correspond to the generators of su(1,1) algebra of quantum operators. They form the basis of the algebra isomorphic to su(1,1) with respect to the introduced Poisson bracket.

If we choose K_0 as a classical Hamilton function, then the evolution of the system is described by the Hamiltonian equations,

$$\dot{z}=\{z,K_0\}=-iz, \quad \dot{\overline{z}}=\{\overline{z},K_0\}=i\overline{z},$$

where the overdot stands for the time derivative.

Geometric quantization[17] of the obtained classical system leads to the holomorphic representation of the quantum system described above. (In our case this is a quantization on the Lobachevsky plane,[14] which really is a Berezin[24] quantization on a Kähler manifold.) The value $h=(2k)^{-1}$ plays the role of the Plank constant.[14]

To conclude this section we present the explicit expressions for the Bergman kernel, which takes off the integration in the $\mathscr{L}$ space,

$$\delta^{(0)}(z,\overline{z'})=\sum_{n=0}^{\infty}\psi_n(z)\overline{\psi_n}(z')=(1-z\overline{z'})^{-2k},$$

and for the su(1,1) generators

$$k_0(z)=z\,\frac{d}{dz}+k, \quad k_+(z)=z^2\,\frac{d}{dz}+2kz, \quad k_-(z)=\frac{d}{dz},$$

in holomorphic representation.

III. DARBOUX TRANSFORMATION OF COHERENT STATES

In a simplest case[5,2] the Darboux transformation operator is the first-order differential operator of the form

$$L=-L_0(x)+d/dx=-u'(x)/u(x)+d/dx, \tag{7}$$

where the prime denotes the derivative with respect to x. When acting on the solutions $\psi_n(x)$ of the initial Schrödinger equation,

$$h_0\psi_n(x)=E_n\psi_n(x), \quad E_n=2n+2k,$$

it transforms them into the solutions of another Schrödinger equation,

$$\varphi_n(x)=N_nL\psi_n(x), \quad h_1\varphi_n(x)=E_n\varphi_n(x), \tag{8}$$

with the same eigenvalues E_n. The factor N_n is introduced to ensure that the states φ_n are normalized to unity. New exactly solvable Hamiltonian $h_1=h_0+A(x)$ is defined by the potential difference $A=A(x)=-2(\ln u)''$. The function $u=u(x)$, called a *transformation function*,[2] is a solution to the initial Schrödinger equation,

$$h_0u(x)=\alpha u(x), \tag{9}$$

with $\alpha\leqslant E_0$.

For the sake of simplicity we will restrict ourselves to the case when $\alpha<E_0$, $u(x)\neq 0$, $\forall x\in(0,\infty)$ and $1/u(x)$ is not a square integrable function in the interval $[0,\infty)$. In this case $u\notin H$ and the set $\{|\varphi_n\rangle\}$ constitutes a complete basis in the Hilbert space H, provided the initial system $\{|\psi_n\rangle\}$ is complete.[2] In terms of the supersymmetric quantum mechanics this case corresponds to a broken supersymmetry.

Let us now choose the following solution of the initial equation (9):

$$u=u_p(x)=x^{2k-1/2}L_p^{2k-1}(y)\exp(x^2/4), \quad y=-x^2/2,$$

$$\alpha=-2(k+p), \quad p=0,1,2,...,$$

as the transformation function. It is worth stressing that $L_p^{2k-1}(y)\neq 0$, $\forall x\neq 0$ when $y=-x^2/2$ (<0) and, hence, the function $u_p(x)$ is nodeless for $x>0$. The function $L_0(x)=L_{0p}(x)$ has in this case the form

$$L_{0p}(x)=\frac{1-4k}{2x}-\frac{x}{2}-\frac{xL_{p-1}^{2k}(y)}{L_p^{2k-1}(y)},$$

and for the potential difference we obtain

$$A=A_p(x)=-1+\frac{4k-1}{x^2}-2\frac{x^2L_{p-2}^{2k+1}(y)+L_{p-1}^{2k}(y)}{L_p^{2k-1}(y)}+2x^2\left[\frac{L_{p-1}^{2k}(y)}{L_p^{2k-1}(y)}\right]^2, \quad y=-x^2/2.$$

It is worth noticing that because of the condition $L_p^{2k-1}(y)\neq 0$ for $x\neq 0$, the potential difference $A_p(x)$ has a single pole at $x=0$. The asymptotic behavior of this function near the point $x=0$ is as follows: $A_p(x)\to -1-p/k+(4k-1)/x^2$. In the limit as $x\to\infty$ this function tends to -1. For $b\neq 0$ the behavior of the transformed potentials $V_p(x)=V_0(x)+A_p(x)$ is nearly the same as the behavior of the initial potential $V_0(x)=x^2/4+b/x^2$. There is only a small difference in the region of the minimum. When $b=0$ the initial potential is regular at $x=0$, whereas all $V_p(x)$ are singular: $V_p(x)\to -1-4p/3+2/x^2$. As an illustration, we give the explicit expression for two first potential differences:

$$A_0(x)=-1+\frac{4k-1}{x^2}, \quad A_1(x)=A_0(x)+\frac{4}{4k+x^2}-\frac{32k}{(4k+x^2)^2}.$$

Thus, the transformed Hamiltonians $h_1=h_1^{(p)}=-d^2/dx^2+V_p(x)$ are bounded from below and are essentially self-adjoint operators defined on the dense set in the Hilbert space H. All these Hamiltonians have only a discrete spectrum. The position of the energy levels is the same as before transformation: $E_n=2n+2k$, $n=0,1,2,...$.

The operator $L^+=-L_0(x)-d/dx$ realizes the transformation in the inverse direction,

$$|\psi_n\rangle=N_nL^+|\varphi_n\rangle,$$

and together with L participates in the following factorization:

$$L^+L=h_0-\alpha, \quad LL^+=h_1-\alpha. \tag{10}$$

We note that in our case operators L and L^+ are well defined $\forall\psi\in H$ and are conjugated to each other with respect to the inner product in the space H. This observation, together with the factorization properties (10), is very useful in calculating integrals involving transformed functions. In particular, for the normalization constant N_n we readily find $N_n^{-2}=\langle\varphi_n|h_1-\alpha|\varphi_n\rangle=E_n-\alpha=2p+4k+2n$.

Now, by following the method of Ref. 18 we find the transformed coherent states,

$$\varphi_z(x)=N_{1z}L\psi_z(x)=N\sum_{n=0}^{\infty} b_n z^n\varphi_n,$$

$$N=N_{0z}N_{1z}/N_0, \quad b_n=a_nN_0/N_n.$$

From the factorization properties (10) one readily derives the value of the normalization constant

$$N_{1z}^{-2}=\frac{4k+2p-2pz\overline{z}}{1-z\overline{z}}.$$

Using the identity operator,

$$\sum_{n=0}^{\infty}|\varphi_n\rangle\langle\varphi_n|=1, \tag{11}$$

one can calculate the integration measure $\nu(z)$, which realizes the following identity decomposition:

$$\int_{|z|<1}|\varphi_z\rangle\langle\varphi_z|d\nu(z)=1. \tag{12}$$

For this purpose we pass from the complex variables z and $\overline{z}$ to the polar ones $z=\sqrt{x}\exp i\phi$ and look for the measure in the form $d\nu=1/2h(x)dx\,d\phi$. The integration over the angle variable ϕ in (12) is trivial. Then, making use of the formula (11), we obtain the equation for the function $h(x)$:

$$\pi\int_0^1 dx\ h(x)\ \frac{x^n(1-x)^{2k+1}}{2k+p-px}=\frac{n!\Gamma(2k)}{\Gamma(n+2k)(n+2k+p)}.$$

So, we arrive at a moments problem well known in mathematics.[25] In our case the solution to this problem can readily be obtained with the help of the integral representation of the beta function,

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}=\int_0^1 x^{a-1}(1-x)^{b-1}\,dx,$$

and the following identity:

$$\frac{\Gamma(n+1)\Gamma(2k)}{(n+2k+p)\Gamma(n+2k)}=\sum_{j=0}^{p}C_p^j\,\frac{2k-1}{2k+p-j-1}\,B(n+j-1,2k+p-j),$$

where C_p^j are the binomial coefficients. It is not difficult to prove this identity by induction. The final expression for the function $h(x)$ is

$$h(x)=\frac{2k-1}{\pi}\,(2k+p-px)\sum_{j=0}^{p}C_p^j\,\frac{x^j(1-x)^{p-j-2}}{2k+p-j-1}.$$

Once we know the measure $\nu(z)$ we can construct the holomorphic representation $\varphi(z)$ in the unit disk of a vector $|\varphi\rangle \in H$, which is defined by its Fourier coefficients c_n with respect to the base $\{|\varphi_n\rangle\}$,

$$\langle \varphi_{\overline{z}}|\varphi\rangle = N\varphi(z), \quad \varphi(z) = \sum_{n=0}^{\infty} b_n c_n z^n.$$

It follows from these relations the holomorphic representation for the transformed discrete basis vectors $\varphi_n(z) = b_n z^n$ and for the transformed coherent states,

$$\zeta^{(1)}(z) = N^{-1}\langle \varphi_{\overline{z}}|\varphi_\zeta\rangle = \frac{2k+p+p\zeta z}{\sqrt{2k+p}} \frac{(1-\zeta\overline{\zeta})^{k+1/2}}{(1-\zeta\overline{z})^{2k+1}} (2k+p-p\zeta\overline{\zeta})^{-1/2}.$$

With the help of the transformation operators L and L^+, we construct the lowering and raising operators $p_\pm = Lk_\pm L^+$ for the transformed discrete basis functions (8),

$$p_+|\varphi_n\rangle = -(N_n N_{n+1})^{-1}[(n+1)(n+2k)]^{1/2}|\varphi_{n+1}\rangle,$$

$$p_-|\varphi_n\rangle = -(N_n N_{n-1})^{-1}[n(n+2k-1)]^{1/2}|\varphi_{n-1}\rangle.$$

Together with the operator $p_0 = 1/2h_1$ they form a nonlinear algebra,

$$[p_0, p_\pm] = \pm p_\pm, \quad [p_-, p_+] = 2[2k(1-k) - p_0\alpha + 4p_0^2](2p_0 - \alpha).$$

It is interesting that the operators $p_\pm$ are of the fourth degree in the derivative in the coordinate representation while in the holomorphic one they are of the second degree,

$$p_0(z) = z\frac{d}{dz} + k, \quad p_-(z) = 2z\frac{d^2}{dz^2} + 2(2k+p)\frac{d}{dz},$$

$$p_+(z) = 2z^3\frac{d^2}{dz^2} + 2z(2k+p+2)\left(z\frac{d}{dz} + 2k\right).$$

Since the operators L and L^+ realize the correspondence between the basis sets $\{|\psi_n\rangle\}$ and $\{|\varphi_n\rangle\}$ we can find their holomorphic representation,

$$L(z) = \sqrt{\frac{2}{2k+p}}\left(z\frac{d}{dz} + 2k + p\right), \quad L^+(z) = \sqrt{2(2k+p)}.$$

By assuming that the inner product of two vectors $|\varphi_1\rangle$ and $|\varphi_2\rangle$ is independent on the representation used and taking into account the identity decomposition (12), we can equip the lineal $\mathscr{L}$ with a new inner product,

$$\langle \varphi_1|\varphi_2\rangle = \langle \varphi_1(z)|\varphi_2(z)\rangle = \int_{|z|<1} \overline{\varphi_1}(z)\varphi_2(z)e^{-f^{(1)}}\, d\nu(z). \tag{13}$$

It is natural to declare the function

$$f^{(1)} = f^{(1)}(z,\overline{z}) = \ln|\langle \varphi_0|\varphi_z\rangle|^{-2} = f^{(0)} + \ln\frac{\langle \psi_z|h_0 - \alpha|\psi_z\rangle}{E_0 - \alpha},$$

a new Kähler potential. It then follows the expression for a new Hermit metric:

$$g^{(1)} = f^{(1)}_{z\overline{z}} = \frac{2k+1}{(1-z\overline{z})^2} - \frac{p(2k+p)}{(2k+p-pz\overline{z})^2}.$$

The unit disk represents now a phase space with the varying Gauss curvature $\mathscr{K}^{(1)}=-(2/g)\times(\partial^2/\partial z\partial\overline{z})\ln g^{(1)}$. The explicit expression for $\mathscr{K}^{(1)}$ is rather complicated so it is omitted here. We will note only that $\mathscr{K}^{(1)}\to 0$ when $k\to\infty$.

The function

$$\delta^{(1)}(z,\overline{z'})=\sum_{n=0}^{\infty}\varphi_n(z)\overline{\varphi_n}(z')=(1-z\overline{z'})^{-2k-1}\,\frac{2k+p-pz\overline{z}}{2k+p}$$

plays the role of the transformed Bergman kernel that takes off the integration with respect to the new inner product (13).

By making use of the covariant Berezin symbols, we can establish correspondence between the quantum mechanical operators and the classical observables that are the functions of a point in the unit disk. The function

$$H_1=\langle\varphi_z^-|h_1|\varphi_z^-\rangle=2k+\frac{4kz\overline{z}(2k+p+1-pz\overline{z})}{(1-z\overline{z})(2k+p-pz\overline{z})}$$

corresponds to the transformed Hamiltonian operator, and therefore it is the transformed Hamilton function. Similarly, the functions P_+ and P_-,

$$P_+=\overline{P_-}=\langle\varphi_z^-|p_+|\varphi_{\overline{z}}^-\rangle=\frac{4k\overline{z}}{2k+p-pz\overline{z}}\left[\frac{(2k+1)(2k+2)}{(1-z\overline{z})^2}+(p-1)\,\frac{2k+1}{1-z\overline{z}}+p(p-1)\right],$$

correspond to the lowering and raising operators.

The change of the Kähler potential induces the change of the symplectic form that remains closed because of its dimension. As a consequence, we have another Poisson bracket in the unit disk. The functions $P_0=1/2H_1$ and $P_\pm$ form now neither a closed Lie algebra nor a polynomial one with respect to the new Poisson bracket, but it is not difficult to check the following relations:

$$\{P_0,P_\pm\}=\pm iP_\pm,\quad \{P_0,z\}=iz,\quad \{P_0,\overline{z}\}=-i\overline{z}.$$

Two latter relations mean that the equation of trajectories for the transformed classical system in the transformed phase space remains unchanged under the Darboux transformation.

IV. CONCLUSION

Using the coherent states technique we have constructed a classical counterpart of the transformed *a lá* Darboux quantum singular oscillator. It has been shown that the Darboux transformation for this system translates into the transformation of the Kähler potential. New Kähler potential has a supplementary summand of the form $\ln(H_0-\alpha)$, where H_0 is the initial Hamilton function corresponding to the quantum Hamiltonian operator for the singular oscillator. This leads to the distortion of the initial phase space. However, the distortion is consistent with the change of the Hamilton function in such a manner that the curves of a constant energy (classical trajectories) in the transformed phase space remain unchanged. A similar observation has recently been discussed in a Lagrangian formalism and for the canonical method of quantization.[13]

ACKNOWLEDGMENT

This research has been partially supported by the Russian Foundation for Basic Research Grant No. 97-02-16279.

[1] J. Delsart, Comptes. Rend. Acad. Sci. Paris. **206**, 178 (1938); Colloque Int. Nancy. 29 (1956); J. L. Lions, Colloques Int. Nancy. 125 (1956); J. Delsart and J. L. Lions, Comment Math. Helv. **32**, 113 (1957).
[2] V. G. Bagrov and B. F. Samsonov, Fiz. Elem. Chastits At. Yadra **28**, 951 (1997); Theor. Math. Phys. **104**, 1051 (1995).
[3] T. F. Moutard, J. École Polytechnique **XLV**, 1 (1878).
[4] G. Darboux, Comptes Rend. Acad. Sci. Paris. **94**, 1343 (1882); **94**, 1456 (1882); *Leçons sur la Théorie Générale des Surfaces et les Application Géométriques du Calcul Infinitésimale. Deuxiem Partie* (Gautier-Villar et Fils, Paris, 1889).
[5] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1926).
[6] V. Matveev and M. Salle, *Darboux Transformations and Solitons* (Springer, Berlin, 1991).

[7] E. Schrödinger, Proc. R. Irish. Acad A **46**, 9 (1940); **47**, 53,183 (1941).
[8] T. E. Infeld, Phys. Rev. **59**, 737 (1941); H. Hull and T. E. Infeld, *ibid.* **74**, 905 (1948); T. E. Infeld and H. Hull, Rev. Mod. Phys. **53**, 21 (1951).
[9] L. D. Faddev, Usp. Mat. Nauk **14**, 57 (1959); B. N. Zakhariev and A. A. Suzko, *Direct and Inverse Problems* (Springer-Verlag, Heidelberg, 1990).
[10] A. A. Suzko, Int. J. Mod. Phys. A **12**, 277 (1997); B. F. Samsonov, J. Phys. A **28**, 6989 (1995).
[11] G. Junker, *Supersymmetric Methods in Quantum and Statistical Physics* (Springer, Berlin 1996); F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **251**, 267 (1995); V. A. Kostelecky and D. K. Campbell, *Supersymmetry in Physics* (North-Holland, Amsterdam, 1985).
[12] B. F. Samsonov, Mod. Phys. Lett. A **11**, 1563 (1996).
[13] A. A. Neto and E. D. Filho, Mod. Phys. Lett. A **12**, 899 (1997).
[14] A. M. Perelomov, *Generalized Coherent States and Their Applications* (Springer, Berlin, 1986).
[15] I. A. Malkin and V. I. Man'ko, *Dynamical Symmetries and Coherent States of Quantum Systems* (Nauka, Moscow, 1979); J. R. Klauder and B. S. Skagerstam, *Coherent States: Applications in Physics and Mathematical Physics* (World Scientific, Singapore, 1985).
[16] J. R. Klauder, J. Math. Phys. **4**, 1055,1058 (1963); **5**, 177 (1964).
[17] B. Kostant, Lect. Notes Math. **170**, 87 (1970); A. Kirillov, *Eléments de la Théorie des Représentation* (Mir, Moscou, 1974); N. M. J. Woodhouse, *Geometric Quantization* (Oxford University Press, 1991); J. R. Klauder, Geometric quantization from a coherent state viewpoint, e-print quant-ph/9510008, 1995.
[18] V. G. Bagrov and B. F. Samsonov, Zh. Eksp. Teor. Fiz. **109**, 1105 (1996); J. Phys. A **29**, 1011 (1996).
[19] F. Calogero, J. Math. Phys. **10**, 2191,2197 (1969); **12**, 419 (1971).
[20] J. M. Leinaas and J. Mirheim, Phys. Rev. B **37**, 9286 (1988); A. P. Polychronacos, Nucl. Phys. B **324**, 597 (1989); F. D. M. Haldane, Phys. Rev. Lett. **67**, 937 (1991).
[21] J. A. Minahan and A. P. Polychronacos, Phys. Lett. B **326**, 288 (1994).
[22] A. M. Perelomov, Funct. Anal. Appl. **7**, 57 (1973).
[23] F. A. Berezin, Izv. Akad. Nauk SSSR Ser. Mat. **37**, 1134 (1972).
[24] F. A. Berezin, Dokl. Akad. Nauk SSSR Ser. Mat. **211**, 1263 (1973); Izv. Akad. Nauk SSSR Ser. Mat. **38**, 1116 (1974); **39**, 363 (1975).
[25] N. I. Akhiezer, *Classical Moments Problem* (Phys.-Math. Literature State Publishing House, Moscow, 1961).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

The reduced self-dual Yang–Mills equation, binary and infinitesimal Darboux transformations

N. V. Ustinov[a)]
Department of Theoretical Physics, Kaliningrad State University, Al. Nevsky street 14, 236041, Kaliningrad, Russia

(Received 12 May 1997; accepted for publication 4 September 1997)

The self-dual Yang–Mills equation in the (2+1)-dimensional space–time is considered. Binary Darboux transformation of a new kind is applied to obtain the infinite hierarchy of solutions expressed through ones of corresponding Lax pairs on an initial solution of the equation studied. Sufficient conditions are given for the solutions built to take values in the class of unimodular positive-definite matrices. A new infinitesimal Darboux transformation is introduced and a particular solution of the linearization of the self-dual Yang–Mills equation is given. Two hierarchies of infinitesimal symmetries depending on a solution of the nonlinear equation are extracted. © *1998 American Institute of Physics.* [S0022-2488(97)00612-9]

I. INTRODUCTION

The self-dual Yang–Mills equation (SDYM) has been thoroughly investigated by various authors. In particular, it has been found that many famous nonlinear integrable equations can be deduced from SDYM.[1–3] An extension of this relationship to the higher-dimensional space–time seems to become an attractive problem in the theory of integrable equations.[4,5] Nevertheless, it is worth paying more attention to some classical results concerning SDYM.

In the case of gauge group SU(n), SDYM is equivalent to an equation for the unimodular positive-definite matrix-valued function. It is well known that in the space–time with signature (2,2) this equation is reduced to the evolution equation

$$(J_t J^{-1})_t = (J_z J^{-1})_{\bar{z}}, \tag{1}$$

assuming matrix-valued function J being independent of one of the coordinates.[6] This equation, which is the relativistic-invariant three-dimensional reduction of the full SDYM, will be considered hereafter.

The Cauchy problem of Eq. (1) for rapidly decreasing initial values was solved by means of the inverse scattering transformation method (IST) based on the Riemann conjugation problem.[6] The reductions on scattering data, which guarantee the unimodularity and hermiticity of matrix J, were given. Soliton solutions corresponding to a purely discrete part of the scattering data[7] were considered. The explicit formula of the one-soliton solution was presented and the solutions finite in all directions were obtained. However, there are some difficulties in using the IST formalism. The problem of recovering a positive-definite solution of Eq. (1) from the scattering data has not yet been discussed. A recurrent formula of the multisoliton solution contains no contributions of single soliton components in a uniform way. Moreover, a modification of IST presented in the work[8] led to the wrong expression of a two-soliton solution.

In this paper we apply the Darboux transformation (DT) technique[9] to overcome the problems noted above. In Sec. II we recall the representations of the equation considered as the compatibility condition of linear systems of differential equations (Lax pairs) and introduce an automorphism[10] determined by the reduction constraint on matrix J. A theorem establishing binary DT (BDT) is presented. It should be stressed that the BDT introduced here differs from the known one.[9] This transformation is a particular case of the sequence of elementary DTs[11] and gives the usual BDT as a result of an appropriate number of iterations. (It was noted in Ref. 12 that

a)Electronic mail: ustinov@theor.phys.ksu.kern.ru

 © 1998 American Institute of Physics

the usual BDT makes sense after substituting the vector solutions for matrix ones. However, a factorization of such a BDT was not considered.) Sufficient conditions covariant with respect to the BDT are given to construct the solutions of Eq. (1) belonging to the class of unimodular positive-definite matrices. The infinite hierarchies of solutions of Eq. (1) and corresponding ones of Lax pairs are obtained in Sec III. The scattering data of the solutions built are shown to differ by the discrete part only. To put it another way, the solution of Eq. (1) with the given spectral data is generated by BDT iterations starting from one having the same continuous part. A basic property underlying the DT technique allows one to express the formula of successive iterations of BDT through the Lax pair solutions corresponding to an initial solution of the reduced SDYM equation. The contributions of one-soliton solutions are included in this formula uniformly. The DT technique is developed in Sec. IV to obtain a particular solution for the linearization of Eq. (1). An infinitesimal DT is introduced and two hierarchies of infinitesimal symmetries, which contain explicitly matrix J, are extracted.

II. BINARY DARBOUX TRANSFORMATION

Equation (1) arises from the compatibility condition of a system of linear equations (Lax pair),[6]

$$\begin{aligned} \Psi_t &= -\lambda \Psi_{\bar{z}} + U\Psi, \\ \Psi_z &= -\lambda \Psi_t + V\Psi. \end{aligned} \tag{2}$$

Here λ is the spectral parameter ($\lambda \in C$), $z = x_1 + ix_2$, $\Psi \equiv \Psi(t, z, \bar{z}, \lambda)$ is the nondegenerate $n*n$ matrix, currents U and V are defined by the solution of Eq. (1),

$$\begin{aligned} U &= J_t J^{-1}, \\ V &= J_z J^{-1}. \end{aligned} \tag{3}$$

Let us introduce for the sake of convenience the dual Lax pair with spectral parameter æ,

$$\begin{aligned} \Xi_t &= -æ\Xi_{\bar{z}} - \Xi U, \\ \Xi_z &= -æ\Xi_t - \Xi V, \end{aligned} \tag{4}$$

on the matrix-valued function $\Xi \equiv \Xi(t, z, \bar{z}, æ)$. Spaces of the matrix solutions of direct and dual pairs (2), (4) are connected:

$$\Xi(t, z, \bar{z}, \lambda) = \Phi(\bar{z}\lambda t + \lambda^2 z)\Psi(t, z, \bar{z}, \lambda)^{-1},$$

where $\Phi(\bar{z}\lambda t + \lambda^2 z)$ is a solution of Eq. (2) on the zero background.

Matrix-column and matrix-row solutions of Lax pairs will be called *vector solutions* in the following. A vector solution of the direct pair with spectral parameter λ and one of the dual pair with spectral parameter æ will be, respectively, designated ψ and ξ, omitting a dependence on the arguments for brevity.

It should be noted that Eq. (1) may be represented as a set of first-order equations,

$$\begin{aligned} U_t &= V_{\bar{z}}, \\ U_z - V_t &= [V, U]. \end{aligned} \tag{5}$$

The second equation in (5) is nothing but the compatibility condition of system (3) according to a statement,

$$J\Phi(\bar{z}) = \Psi(t, z, \bar{z}, \lambda)|_{\lambda = 0}. \tag{6}$$

This relation will be used in obtaining Darboux transformation formulas.

As it was mentioned in the previous section, the additional restrictions are imposed on solutions of Eq. (1) in the gauge group SU(n) case, namely,

$$\det J = 1, \tag{7a}$$

$$J^+ = J, \tag{7b}$$

$$\varphi^+ J \varphi > 0 \quad (\varphi = (\varphi_1, \ldots \varphi_n)^T, \quad \varphi^+ \varphi \neq 0). \tag{7c}$$

If J is a $2*2$ matrix satisfying these constraints then Eq. (1) has a local real Lagrangian and a corresponding Hamiltonian to be positively defined.[13]

Condition (7b) defines an automorphism[10] in the spaces of Lax pair solutions,

$$\Psi(\lambda) \to J[\Psi(\overline{\lambda^{-1}})^+]^{-1} \in \{\Psi(\lambda)\},$$

$$\Xi(æ) \to [\Xi(\overline{æ^{-1}})^+]^{-1} J^{-1} \in \{\Xi(æ)\},$$

and gives a connection between the spaces of vector solutions: matrix-row $\psi^+ J^{-1}$ is a solution of the dual pair with spectral parameter $\overline{\lambda^{-1}}$. One can consequently put

$$\xi = \psi^+ J^{-1}, \tag{8}$$

implying $æ = \overline{\lambda^{-1}}$. Lax pair solutions satisfying condition (8) will be called *connected through automorphism*. It is obvious that an existence of the automorphism leads to the restrictions on the scattering data. In particular, the locations of zeros and poles of the Riemann conjugate problem are related. A solution of a Lax pair having the zero at point μ on the spectral parameter plane has also the pole at point $\overline{\mu^{-1}}$, and vice versa.

The usual DT[9] and also the elementary DTs (EDTs)[11,14] may be used to obtain the new solutions of Eq. (1) together with the corresponding solutions of Lax pairs. The method of keeping the reduction constraints under DT iterations[11,14,15] allows us to restrict our consideration on binary DT (BDT)[11] (its relation with EDTs will be noted below).

Theorem: *Let φ and χ be the vector solutions of direct and dual Lax pairs* (2,4) *with spectral parameters μ and ν, respectively ($\mu \neq \nu$). Then the matrix functions $\Psi[1]$ and $\Xi[1]$, defined as*

$$\Psi[1] = \left(E - \frac{\mu - \nu}{\lambda - \nu} P \right) \Psi, \tag{9a}$$

$$\Xi[1] = \Xi \left(E - \frac{\mu - \nu}{\mu - æ} P \right), \tag{9b}$$

where

$$P = \frac{\varphi \chi}{(\chi, \varphi)}, \qquad (\chi, \varphi) = \chi \varphi \equiv \sum_{i=1}^{n} \chi_i \varphi_i, \tag{10}$$

are solutions of Lax pairs with new currents,

$$U[1] = U - (\mu - \nu) P_{\overline{z}}, \qquad V[1] = V - (\mu - \nu) P_t. \tag{11}$$

The corresponding solution of Eq. (1) is given by the formula

$$J[1]=\alpha\left(E+\frac{\mu-\nu}{\nu}P\right)J \tag{12}$$

(α is a constant).

Proof: Substituting quantities (9a), (9b), and (11) into the systems

$$\begin{aligned}\Psi[1]_t&=-\lambda\Psi[1]_{\bar{z}}+U[1]\Psi[1],\\ \Psi[1]_z&=-\lambda\Psi[1]_t+V[1]\Psi[1],\end{aligned} \tag{13a}$$

and

$$\begin{aligned}\Xi[1]_t&=-æ\Xi[1]_{\bar{z}}-\Xi[1]U[1],\\ \Xi[1]_z&=-æ\Xi[1]_t-\Xi[1]V[1],\end{aligned} \tag{13b}$$

one obtains a set of differential equations on matrix P,

$$\begin{aligned}P_t&=-\tfrac{1}{2}(\mu+\nu)P_{\bar{z}}+[P,\tfrac{1}{2}(\mu-\nu)P_{\bar{z}}-U],\\ P_z&=-\tfrac{1}{2}(\mu+\nu)P_t+[P,\tfrac{1}{2}(\mu-\nu)P_t-V].\end{aligned} \tag{14}$$

The left-hand side of Eq. (10) solves it. Therefore, systems (13) are compatible and new currents (11) satisfy system (5). Expression (12) for new solution $J[1]$ follows from its relation with the transformed solution of the direct Lax pair (13a):

$$J[1]\Phi(\overline{z})=\Psi[1]|_{\lambda=0.} \tag{15}$$

■

The formulas (9)–(12) constitute binary DT. In the sequel we shall state that BDT is performed *with* vector solutions φ and χ. The definitions of transformed vector solutions $\psi[1]$ and $\xi[1]$ follow from Eqs. (9), replacing the matrix solutions by corresponding vector ones. Transformation $\{\psi,\xi\}\rightarrow\{\psi[1],\xi[1]\}$ has the next property: $\psi[1]=0$ if $\lambda=\mu$ and $\psi=\varphi$, $\xi[1]=0$ if $æ=\nu$ and $\xi=\chi$. An existence of the kernel is a common feature of any DT and underlies this technique. The statement may be proven, allowing for that BDT is a particular case of successive carrying out of two EDTs (namely, one of the EDTs of the direct Lax pair and a proper one of the dual pair[11]). Such a representation of BDT also yields a relation,

$$\det\Psi[1]=\frac{\lambda-\mu}{\lambda-\nu}\det\Psi, \tag{16}$$

displaying a presence of additional zero and pole in the discrete part of the scattering data[7] of the transformed solution $J[1]$. The vector solutions φ and χ complete a description of the new scattering data and define an evolution of them. The successive iterations of BDT give rise to a solution of Eq. (1) with the arbitrary discrete part of the scattering data.

Consider the constraints on BDT allowing us to build a solution of SDYM satisfying conditions (7). It is seen from Eqs. (15) and (16) that a unimodular solution is obtained, putting in Eq. (12)

$$\alpha=(\nu/\mu)^{1/n}.$$

A Hermiticity of Eq. (1) solutions imposes the restrictions on scattering data. Therefore, one has to choose the vector solutions φ and χ connected through the automorphism for the transformed solution to inherit this reduction. It is remarkable that relation (8) occurs then to be covariant (within an unimportant scalar factor) with respect to such a BDT:

$$\xi[1]=|\mu|^{-2/n}\psi[1]^{+}J[1]^{-1},\quad\text{if}\quad\xi=\psi^{+}J^{-1}.$$

Combining these constraints on a BDT (9)–(12), the expression for the new unimodular Hermitian solution of Eq. (1) becomes

$$J[1]=|\mu|^{-2/n}\left(J+(|\mu|^2-1)\,\frac{\varphi\varphi^+}{(\varphi^+,J^{-1}\varphi)}\right). \tag{17}$$

The last formula may be written down in the form

$$J[1]=AJ^{-1}A,$$

where

$$A=|\mu|^{-1/n}\left(J+(-1\pm|\mu|)\,\frac{\varphi\varphi^+}{(\varphi^+,J^{-1}\varphi)}\right).$$

It shows obviously that, assuming an initial solution to satisfy condition (7c), we generate the new positive-definite solution of Eq. (1) by keeping the reduction (7b). Therefore, the problem of recovering a positive-definite solution of SDYM is reduced to the case of the scattering data having the continuous part only.

III. ITERATIONS OF BINARY DARBOUX TRANSFORMATION

Let $\varphi^{(i)}$ and $\chi^{(i)}$ $(i=1,\ \ldots\ ,N)$ be the vector solutions of direct and dual Lax pairs (2), (4) with spectral parameters $\mu^{(i)}$ and $\nu^{(i)}$. For the sake of simplicity we suggest that the values of the spectral parameters are mutually different. The matrix-column and matrix-row functions,

$$\varphi^{(i)}[1]=\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\mu^{(i)}-\nu^{(1)}}\,P[1]\right)\varphi^{(i)},$$

$$\chi^{(i)}[1]=\chi^{(i)}\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\mu^{(1)}-\nu^{(i)}}\,P[1]\right),$$

$$\left(P[1]=\frac{\varphi^{(1)}\chi^{(1)}}{(\chi^{(1)},\varphi^{(1)})}\right),$$

are the vector solutions of the transformed direct and dual pairs with spectral parameters $\mu^{(i)}$ and $\nu^{(i)}$ with the new currents. The corresponding solution of Eq. (1) is obtained by performing the BDT (9)–(12) with solutions $\varphi^{(1)}$ and $\chi^{(1)}$. Carrying out successively the iterations of BDT with transformed vector solutions $\varphi^{(i)}[i-1]$ and $\chi^{(i)}[i-1]$ $(i=2,\ \ldots\ ,N)$, defined as follows:

$$\varphi^{(i)}[i-1]=\left(E-\frac{\mu^{(i-1)}-\nu^{(i-1)}}{\mu^{(i)}-\nu^{(i-1)}}\,P[i-1]\right)\cdots\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\mu^{(i)}-\nu^{(1)}}\,P[1]\right)\varphi^{(i)},$$

$$\chi^{(i)}[i-1]=\chi^{(i)}\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\mu^{(1)}-\nu^{(i)}}\,P[1]\right)\cdots\left(E-\frac{\mu^{(i-1)}-\nu^{(i-1)}}{\mu^{(i-1)}-\nu^{(i)}}\,P[i-1]\right),$$

where

$$P[i]=\frac{\varphi^{(i)}[i-1]\chi^{(i)}[i-1]}{(\chi^{(i)}[i-1],\varphi^{(i)}[i-1])},$$

produces the recurrent formulas for the transformed matrix solutions,

$$\Psi[N]=\left(E-\frac{\mu^{(N)}-\nu^{(N)}}{\lambda-\nu^{(N)}}\,P[N]\right)\cdots\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\lambda-\nu^{(1)}}\,P[1]\right)\Psi,$$

$$\Xi[N]=\Xi\left(E-\frac{\mu^{(1)}-\nu^{(1)}}{\mu^{(1)}-æ}\,P[1]\right)\cdots\left(E-\frac{\mu^{(N)}-\nu^{(N)}}{\mu^{(N)}-æ}\,P[N]\right).$$

Using the definitions of the transformed vector solutions and, as a consequence, the identities similar to the next one,

$$\frac{(\chi^{(i)}[1],\varphi^{(i)}[1])}{\mu^{(i)}-\nu^{(i)}}\,\frac{(\chi^{(1)},\varphi^{(1)})}{\mu^{(1)}-\nu^{(1)}}=\begin{vmatrix}\dfrac{(\chi^{(i)},\varphi^{(i)})}{\mu^{(i)}-\nu^{(i)}} & \dfrac{(\chi^{(1)},\varphi^{(i)})}{\mu^{(i)}-\nu^{(1)}}\\ \dfrac{(\chi^{(i)},\varphi^{(1)})}{\mu^{(1)}-\nu^{(i)}} & \dfrac{(\chi^{(1)},\varphi^{(1)})}{\mu^{(1)}-\nu^{(1)}}\end{vmatrix}, \tag{18}$$

allows us to express the projectors $P[i]$ in terms of the initial Lax pair solutions. Taking then into account that $\psi[N]=0$ if $\lambda=\mu^{(i)}$, $\psi=\varphi^{(i)}$, and $\xi[N]=0$ if $æ=\nu^{(i)}$, $\xi=\chi^{(i)}$, one leads the Nth iteration formulas to the form

$$\Psi[N]=\left(E-\sum_{i,k=1}^{N}(\Delta[N]^{-1})_{ik}\,\frac{\varphi^{(i)}\chi^{(k)}}{\lambda-\nu^{(k)}}\right)\Psi, \tag{19a}$$

$$\Xi[N]=\Xi\left(E-\sum_{i,k=1}^{N}(\Delta[N]^{-1})_{ik}\,\frac{\varphi^{(i)}\chi^{(k)}}{\mu^{(i)}-æ}\right), \tag{19b}$$

$$\Delta[N]_{ik}=\frac{(\chi^{(i)},\varphi^{(k)})}{\mu^{(k)}-\nu^{(i)}},$$

$$(i,k=1,\ \ldots\ ,N).$$

The relation of $\Psi[N]$ with new solution $J[N]$ of Eq. (1) yields

$$J[N]=\alpha[N]\left(E+\sum_{i,k=1}^{N}(\Delta[N]^{-1})_{ik}\,\frac{\varphi^{(i)}\chi^{(k)}}{\nu^{(k)}}\right)J. \tag{20}$$

So, we rewrite the recurrent formula in a more convenient manner, displaying obviously an equal contribution of each soliton component. Roughly speaking, these solutions of SDYM describe N-soliton solutions developing on an initial background.

Now we shall set out to construct the new solutions satisfying reductions (7). It was proven in the previous section that the condition keeping reduction (7b) is sufficient to inherit reduction (7c) by the transformed solution of SDYM and the connection through the automorphism (8) is covariant with respect to such a BDT. These properties impose the next conditions under BDT iterations for expression (20) to ensure the unimodular positive-definite solution of Eq. (1),

$$\nu^{(i)}=(\overline{\mu^{(i)}})^{-1}, \tag{21a}$$

$$\chi^{(i)}=\varphi^{(i)+}J^{-1}, \tag{21b}$$

$$\alpha[N]=\prod_{i=1}^{N}|\mu^{(i)}|^{-2/n}. \tag{21c}$$

This way the DT technique solves the problem of recovering the positive-definite solutions of Eq. (1) for the discrete part of scattering data. A problem remaining is to build the solutions taking values in the class of unimodular positive-definite matrices for the case of purely continuous scattering data. It should be noted that matrix J in Eq. (20) is an arbitrary positive-definite solution of SDYM. It may be, for example, the self-similar solution satisfying an equation for principal chiral field on the plane.[6]

The BDT iterations formulas (19)–(20) with restrictions (21) imposed allow us to produce a great variety of positive-definite solutions of Eq. (1). It was supposed here that the spectral

parameters of vector solutions forming the BDT are different. Then the solution built possesses, in the general case, the additional pairs of zeros and poles in the discrete part of scattering data and corresponds to the many-soliton solution ''added'' on an initial background. The case of some spectral parameters values to be equal is treated as a limit by assuming the analytical dependence of the vector solutions on spectral parameters. This way, the solutions of Eq. (1), which are not imbedded into the usual IST scheme, arise.

Consider, for example, the second iteration of BDT keeping reductions (7). Let $\mu^{(1)}=\mu$, $\mu^{(2)}=\overline{\mu^{-1}}+\epsilon$, and $(\varphi^{(1)+}J^{-1},\varphi^{(2)})=O(\epsilon)$. Taking the limit $\epsilon\to 0$, the coefficients of matrix $J[2]$ become

$$J[2]_{ik}=\begin{vmatrix} J_{ik} & -(\varphi^{(1)+}J^{-1},J^{(k)}) & -(\varphi^{(2)+}J^{-1},J^{(k)}) \\ \varphi_i^{(1)} & \dfrac{(\varphi^{(1)+}J^{-1},\varphi^{(1)})}{|\mu|^{(2)}-1} & \mu^{-1}(A^{+}J^{-1},\varphi^{(1)}) \\ \varphi_i^{(2)} & \bar{\mu}^{-1}(\varphi^{(1)+}J^{-1},A) & \dfrac{|\mu|^2}{1-|\mu|^2}(\varphi^{(2)+}J^{-1},\varphi^{(2)}) \end{vmatrix} \Bigg/ \begin{vmatrix} \dfrac{(\varphi^{(1)+}J^{-1},\varphi^{(1)})}{|\mu|^2-1} & \mu^{-1}(A^{+}J^{-1},\varphi^{(1)}) \\ \bar{\mu}^{-1}(\varphi^{(1)+}J^{-1},A) & \dfrac{|\mu|^2}{1-|\mu|^2}(\varphi^{(2)+}J^{-1},\varphi^{(2)}) \end{vmatrix}, \tag{22}$$

where

$$A=\left.\frac{\partial\varphi^{(2)}}{\partial\mu^{(2)}}\right|_{\mu^{(2)}=\overline{\mu^{-1}}},\quad J=(J^{(1)},\ \ldots\ ,J^{(n)}).$$

As it is seen from Eqs. (16) and (21a), the solution constructed has no additional zeros and poles in the scattering data. A zoo of solutions finite in all directions is generated by expressions (22) starting from the zero background.

IV. INFINITESIMAL DARBOUX TRANSFORMATION

In this section we obtain the explicit expressions for infinitesimal symmetries of Eq. (1). A functional is said to be the infinitesimal symmetry if it satisfies the linearized equation for given nonlinear equation. The infinitesimal transformation $J\to J+\epsilon\ \delta J$ (ϵ is an infinitesimal parameter) gives the linearization of Eq. (1), namely,

$$(J(J^{-1}\ \delta J)_t J^{-1})_t=(J(J^{-1}\ \delta J)_z J^{-1})_{\bar{z}}. \tag{23}$$

The particular solution of Eq. (23) is derived from the BDT (9)–(12). Let $\nu=\mu+\epsilon$, $(\chi,\varphi)=1+O(\epsilon)$ and proceed to the limit $\epsilon\to 0$. From Eq. (12) one has

$$\delta J=-\frac{\varphi\chi}{\mu}J. \tag{24a}$$

The corresponding infinitesimal transformations of the solutions and the currents of Lax pairs (2), (4) $\{\Psi,\Xi,U,V\}\to\{\Psi+\epsilon\ \delta\Psi,\Xi+\epsilon\ \delta\Xi,U+\epsilon\ \delta U,V+\epsilon\ \delta V\}$ are produced by the BDT formulas also,

$$\delta\Psi=\frac{\varphi\chi}{\lambda-\mu}\Psi,$$

$$\delta\Xi=\Xi\,\frac{\varphi\chi}{\mu-æ},$$
$$\delta U\equiv(\delta J)_t J^{-1}-U\;\delta J\;J^{-1}=(\varphi\chi)_{\overline{z}},\tag{24b}$$
$$\delta V\equiv(\delta J)_z J^{-1}-V\;\delta J\;J^{-1}=(\varphi\chi)_t.$$

We name these formulas the infinitesimal Darboux Transformation (IDT). Having yielded a solution of the linear equation, the expression for symmetry (24a) is generalized,

$$\delta J=\int\Psi(\lambda)\Xi(\lambda)\,\frac{d\lambda}{\lambda}\,J.\tag{25a}$$

This representation of infinitesimal symmetry may be used in the study of the stability of SDYM solutions. Infinitesimal transformations $\delta\Psi(\mu)$ and $\delta\Xi(\mu)$ of Lax pair solutions with spectral parameter μ, which are associated with symmetry (25a), are given by the expressions

$$\delta\Psi(\mu)=\int\frac{\Psi(\lambda)\Xi(\lambda)}{\lambda-\mu}\,d\lambda\;\Psi(\mu),$$
$$\delta\Xi(\mu)=\Xi(\mu)\int\frac{\Psi(\lambda)\Xi(\lambda)}{\mu-\lambda}\,d\lambda.\tag{25b}$$

For a Hermitian solution of Eq. (1), one obtains Hermitian symmetry by adding into Eq. (25a) the integral with solutions connected through the automorphism,

$$\delta J=\int(\lambda^{-1}\Psi(\lambda)\Xi(\lambda)J+\overline{\lambda^{-1}}J\Xi(\lambda)^{+}\Psi(\lambda)^{+})da,$$

where the solutions of Lax pairs depend on the real parameter a through their dependence on spectral parameter $\lambda\equiv\lambda(a)$.

Hierarchies of infinitesimal symmetries expressed in terms of the Eq. (1) solution result from expanding the solutions of Lax pairs in a power series on the spectral parameter. Consider the expansions in the neighborhoods of two points on spectral parameter plane $\lambda=0$ and $\lambda=\infty$. Let $A^{(i)}$, $B^{(i)}$, $C^{(i)}$, $D^{(i)}$ $(i=0,1,...)$ be the matrices of functions of the space–time coordinates. Substitution in Eqs. (2) and (4) of formal Taylor series,

$$\Psi(\lambda)=\sum_{i=0}^{\infty}\lambda^{i}A^{(i)},$$
$$\Xi(æ)=\sum_{i=0}^{\infty}æ^{i}B^{(i)},$$

for the first case and

$$\Psi(\lambda)=\sum_{i=0}^{\infty}\lambda^{-i}C^{(i)},$$
$$\Xi(æ)=\sum_{i=0}^{\infty}æ^{-i}D^{(i)},$$

for the second one gives

$$A^{(0)}=-J\Phi^{(0)}(\overline{z}),$$

$$A^{(i)}=-J\left(\int (J^{-1}A_{\overline{z}}^{(i-1)}\,dt+J^{-1}A_t^{(i-1)}\,dz)+\Phi^{(i)}(\overline{z})\right),$$

$$B^{(0)}=-F^{(0)}(\overline{z})J^{-1},$$

$$B^{(i)}=-\left(\int (B_{\overline{z}}^{(i-1)}J\,dt+B_t^{(i-1)}J\,dz)+F^{(i)}(\overline{z})\right)J^{-1},$$

$$C^{(0)}=G^{(0)}(z),$$

$$C^{(i)}=-\int (J(J^{-1}C^{(i-1)})_z\,dt+J(J^{-1}C^{(i-1)})_t\,d\overline{z})+G^{(i)}(z),$$

$$D^{(0)}=H^{(0)}(z),$$

$$D^{(i)}=-\int ((D^{(i-1)}J)_zJ^{-1}\,dt+(D^{(i-1)}J)_tJ^{-1}\,d\overline{z})+H^{(i)}(z).$$

Inserting these expansions in Eq. (25a) generates two infinite hierarchies of symmetries $\delta J^{(i)}$ ($i=0,\ 1,\ ...$). The first members are

$$\delta J^{(0)}=JT(\overline{z}),$$

$$\delta J^{(1)}=J\left[T(\overline{z}),\int (J^{-1}J_{\overline{z}}dt+J^{-1}J_t\,dz)\right]-JT'(\overline{z})t$$

and

$$\delta J^{(0)}=S(z)J,$$

$$\delta J^{(1)}=\left[S(z),\int (J_zJ^{-1}\,dt+J_tJ^{-1}\,d\overline{z})\right]J+S'(z)Jt,$$

where $T(\overline{z})=\Phi^{(0)}(\overline{z})F^{(0)}(\overline{z})$, $S(z)=G^{(0)}(z)H^{(0)}(z)$.

A solution of the linearized equation and a hierarchy of infinitesimal symmetries may be obtained by means of the BDT for other integrable nonlinear equations. The approach suggested in this section resembles the infinitesimal version of the ''dressing'' method.[16] An expansion on spectral parameter powers of the bilinear product of Lax pairs solutions was also used in the work of Ref. 17 to derive a hierarchy of symmetries of the KP equation.

V. CONCLUSION

The technique developed may be applied to the equation for the principal chiral field on the plane, which coincides with the self-similar reduction of Eq. (1),[6] as well as to SDYM in $4k$-dimensional space–time.[4] Moreover, expressions (20) and (25) give new solutions of generalized self-dual Yang–Mills flows[5] and of the linearization of them by supposing that the vector functions $\varphi^{(i)}$, $\chi^{(i)}$ and matrix ones $\Psi(\lambda)$, $\Xi(\lambda)$ are the solutions of corresponding direct and dual Lax pairs. Different reduction constraints on a solution of SDYM are treated in the similar manner, employing the associated automorphism.[11] An analysis of the solutions for these equations, their stability, and the scattering properties of the Eq. (1) solutions, which are not described by the usual IST formalism, are the subjects of the future study.

ACKNOWLEDGMENT

The author is grateful to M. A. Salle, who pointed out the work in Ref. 6 and suggested applying the Darboux transformation technique.

[1] R. S. Ward, Philos. Trans. R. Soc. London, Ser. A **315**, 451 (1985).

[2] L. J. Mason and G. A. J. Sparling, Phys. Lett. A **137**, 29 (1989).
[3] M. J. Ablowitz, S. Chakrayarty, and L. A. Takhtajan, Commun. Math. Phys. **159**, 289 (1993).
[4] K. Takasaki, ''Integrable systems in gauge theory, Kähler geometry and super KP hierarchy—symmetries and algebraic point of view,'' in *Proceedings of the International Congress of Mathematics*, Kyoto, 1990, Vol. II, pp. 1205–1214.
[5] C. H. Gu, Acta Appl. Math. **39**, 349 (1995).
[6] S. V. Manakov and V. E. Zakharov, Lett. Math. Phys. **5**, 247 (1981).
[7] S. P. Novikov, S. V. Manakov, L. P. Pitaevsky, and V. E. Zakharov, *Theory of Solitons: The Inverse Scattering Method* (Consultants Bureau, New York, 1984).
[8] J. Villarroel, Inverse Probl. **5**, 1157 (1989).
[9] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer-Verlag, Berlin, 1991).
[10] A. V. Mikhailov, Physica D **3**, 73 (1983).
[11] N. V. Ustinov, Ph.D. thesis, St. Petersburg University, 1995.
[12] W. Oevel and W. Schief, Rev. Math. Phys. **6**, 1301 (1994).
[13] K. Pohlmeyer, Commun. Math. Phys. **74**, 111 (1980).
[14] S. B. Leble and N. V. Ustinov, J. Phys. A **26**, 5007 (1993); ''Deep reductions for matrix systems, invariant forms, and elementary Darboux transforms,'' in *Nonlinear Evolution Equations and Dynamical Systems*, edited by V. Makhan'kov *et al.* (World Scientific, Singapore, 1993) pp. 34–41.
[15] S. B. Leble and N. V. Ustinov, Inverse Probl. **10**, 617 (1994).
[16] A. Yu. Orlov and E. I. Schulman, Lett. Math. Phys. **12**, 171 (1986).
[17] J. Matsukidaira, J. Satsuma, and W. Strampp, J. Math. Phys. **31**, 1426 (1990).

Nonlinear constraints and soliton solutions of 1+2-dimensional three-wave equation

Zixiang Zhou
Institute of Mathematics, Fudan University, Shanghai 200433, China

(Received 27 January 1997; accepted for publication 7 July 1997)

Using the nonlinear constraint method, the explicit expressions of localized solitons of the 1+2-dimensional three-wave equation are obtained. After the lth Darboux transformation, each component of the solution has, at most, l^2 peaks when the parameters for each single Darboax transformation are large enough. This gives a corresponding asymptotic property for the DSI equation, although all the peaks of each component move in the same velocity here. © *1998 American Institute of Physics.* [S0022-2488(98)03701-3]

I. INTRODUCTION

The 1+2-dimensional three-wave equation describes the nonlinear interaction of three waves and is widely investigated.[1] The inverse scattering transformation has been developed systematically[2–6] and there are many other considerations for this equation.[7,8]

Since the work of Ref. 9 on the 1+1-dimensional problem, the method of nonlinear constraint has been discussed in various papers. In 1+2 dimensions, this method is first applied to the KP equation by Refs. 10 and 11. Generally, the nonlinear constraint in 1+2 dimensions is a method to reduce the original problem to a new 1+2-dimensional problem that can be dealt with by a 1+1-dimensional method. For that new problem, the Darboux transformation has been well developed.[12,13] Using the nonlinear constraint and Darboux transformation methods, Ref. 14 proved that for any equation in the hyperbolic $su(n)$ AKNS system, which includes the three-wave equation, there are localized solitons (i.e., the solitons that decay at infinity in all directions).

If a certain condition [condition (16) of Remark 1] is satisfied so that different solitons have different velocities, the asymptotic solution as $t\to\pm\infty$ has, at most, l^2 peaks if the solution is given by the lth Darboux transformation from the zero solution. This is true for the DSI equation. However, for the three-wave equation, that condition never holds, and all the solitons in each component move in the same velocity. Therefore, the asymptotic solution is still very complicated (not split up into several single solitons). In this paper, we show that when certain parameters are large enough, these peaks can also be separated. This is true, not only for the asymptotic solution (under a weaker condition), but also for the solution at finite time t (under a stronger condition).

In Sec. II of this paper we describe the Lax pair of the three-wave equation, whose nonlinear constraint and the Darboux transformation. The main conclusion has been obtained in Ref. 14. In Sec. III the behavior of a single-soliton solution is discussed in detail. In Sec. IV, the asymptotic behavior of the multisoliton solutions is obtained when the parameters of each single Darboux transformation (not time t) are large enough. This gives a corresponding property for the asymptotic solutions of the DSI equation, although all the solitons move in the same velocity here. Some figures describing the behaviors and interactions of the ''two-soliton'' solutions are shown.

II. NONLINEAR CONSTRAINT AND DARBOUX TRANSFORMATION

The 1+2-dimensional three-wave equation is

$$\begin{aligned} f_{1,t}&=\alpha_1 f_{1,y}+\beta_1 f_{1,x}+(\alpha_3-\alpha_2)f_2\overline{f_3},\\ f_{2,t}&=\alpha_2 f_{2,y}+\beta_2 f_{2,x}+(\alpha_1-\alpha_3)f_1 f_3,\\ f_{3,t}&=\alpha_3 f_{3,y}+\beta_3 f_{3,x}+(\alpha_2-\alpha_1)\overline{f_1}f_2, \end{aligned} \tag{1}$$

 © 1998 American Institute of Physics

where α_i, β_i are real constants.

It is known that this equation has a Lax pair

$$\begin{aligned}\Psi_y &= J\Psi_x + U(x,y,t)\Psi,\\ \Psi_t &= K\Psi_x + V(x,y,t)\Psi,\end{aligned} \tag{2}$$

where

$$J=\begin{pmatrix} J_1 & 0 & 0\\ 0 & J_2 & 0\\ 0 & 0 & J_3\end{pmatrix},\quad K=\begin{pmatrix} K_1 & 0 & 0\\ 0 & K_2 & 0\\ 0 & 0 & K_3\end{pmatrix},\quad U=(u_{ij})=\begin{pmatrix} 0 & f_1 & f_2\\ -\overline{f_1} & 0 & f_3\\ -\overline{f_2} & -\overline{f_3} & 0\end{pmatrix},$$

J_i, K_i are real constants with $J_i \neq J_j$ for $i \neq j$, and V is a 3×3 matrix whose diagonal entries are zero. Moreover, we suppose that

$$\det\begin{pmatrix} 1 & J_1 & K_1\\ 1 & J_2 & K_2\\ 1 & J_3 & K_3\end{pmatrix} \neq 0. \tag{3}$$

The integrability conditions of (2) are

$$[J,V]=[K,U],\quad U_t - V_y + JV_x - KU_x + [U,V]=0. \tag{4}$$

Eliminating V, we get an equation of U, which is just (1) with $\alpha_1=(K_1-K_2)/(J_1-J_2)$, $\alpha_2=(K_1-K_3)/(J_1-J_3)$, $\alpha_3=(K_2-K_3)/(J_2-J_3)$, $\beta_i=K_i-J_i\alpha_i$ $(i=1,2,3)$.

There is another linear system related with (2), which is

$$\Phi_x=\begin{pmatrix} i\lambda I & iP\\ iP^* & 0\end{pmatrix}\Phi,\quad \Phi_y=\begin{pmatrix} i\lambda J+U & iJP\\ iP^*J & 0\end{pmatrix}\Phi,\quad \Phi_t=\begin{pmatrix} i\lambda K+V & iKP\\ iP^*K & 0\end{pmatrix}\Phi, \tag{5}$$

where $P(x,y,t)$ is a 3×3 matrix. The integrability conditions of (5) give

$$P_y = JP_x + UP,\quad P_t = KP_x + VP, \tag{6}$$

$$U_x=[J,PP^*], \tag{7}$$

$$U_t - V_y + [U,V] - JPP^*K + KPP^*J = 0. \tag{8}$$

Clearly the system (6) is the same as (2), except that Ψ is replaced by P. If (7) is considered, (8) is just the three-wave equation (4). Moreover, (7) gives a nonlinear constraint between U and P [P is a solution of the Lax pair (2)]. Therefore, we relate the $1+2$-dimensional problem (2) with the problem (5) by the nonlinear constraint (7). For (5), the standard method in $1+1$ dimensions can be applied.

In Ref. 15, this method was applied for 3×1 matrix P. In that case, nonlocalized solutions of the $1+2$-dimensional three-wave equation are obtained. However, localized solutions can only be obtained by choosing P as a $3\times s$ $(s\geqslant 3)$ matrix with rank 3.

If we can solve (6)–(8), the explicit solutions (f_i) of (1) are obtained, although they are not all solutions of (1).

Reference 14 showed that the equations (1) have localized soliton solutions that tend to zero exponentially at infinity. These solutions can be obtained by a Darboux transformation for (5), provided that the parameters in the Darboux transformation are chosen properly. More precisely, the Darboux transformation can be constructed as follows from the zero seed solution.

Take l complex numbers $\lambda_1,\ldots,\lambda_l$ such that $\lambda_j \neq \lambda_k$ $(j\neq k)$ and $\lambda_j \neq \overline{\lambda_k}$ for all $1\leqslant j,k\leqslant l$. Let

$$H_j=\begin{pmatrix} e^{\sqrt{-1}\lambda_j(x+Jy+Kt)}\\ C^{(j)}\end{pmatrix}, \tag{9}$$

where $C^{(j)}$'s are 3×3 nondegenerate constant matrices. Let $\Sigma=(\Sigma_{jk})_{1\leqslant j,k\leqslant l}$ be a block matrix given by

$$\Sigma_{jk}=\frac{H_j^* H_k}{\lambda_k-\overline{\lambda}_j}. \tag{10}$$

Then

$$G(\lambda)=\prod_{j=1}^{l}(\lambda-\overline{\lambda}_j)\left(1-\sum_{j,k=1}^{l}\frac{H_j}{\lambda-\lambda_k}(\Sigma^{-1})_{jk}H_k^*\right) \tag{11}$$

is a Darboux matrix of degree l for (5) with $U=P=0$. That is, for any solution Φ of (5) with $U=P=0$, $\widetilde{\Phi}=G\Phi$ satisfies (5) for certain U and P. It is known that the action of the Darboux matrix of degree l is equivalent to the successive actions of l Darboux matrices of degree one.

From the standard calculation of a Darboux transformation, we have

$$\begin{aligned} u_{\alpha\beta}&=2\sqrt{-1}\sum_{j,k=1}^{l}(h_j(\Sigma^{-1})_{jk}h_k^*)_{\alpha\beta}\\ &=2\sqrt{-1}\sum_{j,k=1}^{l}(\Sigma^{-1})_{\alpha\beta}e^{\sqrt{-1}(\lambda_j-\overline{\lambda}_k)x+\sqrt{-1}(\lambda_j J_\alpha-\overline{\lambda}_k J_\beta)y+\sqrt{-1}(\lambda_j K_\alpha-\overline{\lambda}_k K_\beta)t}. \end{aligned} \tag{12}$$

Reference 14 proved that when $C^{(j)}$'s are nondegenerate, the solution (12) is localized, and it decays at infinity exponentially. Moreover, the peaks in $u_{\alpha\beta}$ move in the velocity (v_x,v_y) given by

$$v_x=-\frac{J_\alpha K_\beta-J_\beta K_\alpha}{J_\alpha-J_\beta},\quad v_y=-\frac{K_\alpha-K_\beta}{J_\alpha-J_\beta}. \tag{13}$$

For a three-wave equation, these velocities are independent of λ. Therefore, all the solitons move in the same velocity for given J_α, K_α.

Remark 1: For a more general linear system [it comes from the hyperbolic $su(N)$ AKNS system),

$$\begin{gathered} \Phi_x=\begin{pmatrix} i\lambda I & iP\\ iP^* & 0\end{pmatrix}\Phi,\quad \Phi_y=\begin{pmatrix} i\lambda J+U & iJP\\ iP^*J & 0\end{pmatrix}\Phi,\\ \Phi_t=\begin{pmatrix} W(x,y,t,\lambda) & X(x,y,t,\lambda)\\ -X^*(x,y,t,\lambda) & Z(x,y,t,\lambda)\end{pmatrix}\Phi, \end{gathered} \tag{14}$$

the multisoliton solutions are taken from $U=P=0$, $X=0$, $Z=0$,

$$W(x,y,t,\lambda)=\begin{pmatrix} i\beta_1(\lambda) & & \\ & \ddots & \\ & & i\beta_N(\lambda)\end{pmatrix}. \tag{15}$$

In that case, Proposition 4 of Ref. 14 shows that if

$$\det\begin{pmatrix} 1 & J_i & \beta_{iI}(\lambda_\alpha)/\lambda_{\alpha I}\\ 1 & J_j & \beta_{jI}(\lambda_\beta)/\lambda_{\beta I}\\ 1 & J_k & \beta_{kI}(\lambda_\gamma)/\lambda_{\gamma I}\end{pmatrix}\neq 0 \tag{16}$$

for any mutually different pairs (i,α), (j,β), and (k,γ), then as $t\to\infty$, each entry $U_{ij}^{[l]}$ of the solution $U^{[l]}$ given by the lth Darboux transformations splits up into at most l^2 peaks. Here z_I is the imaginary part of a complex number z. Actually, (16) implies that different solitons have different velocities.

For a three-wave equation, $\beta_i(\lambda)=K_i$, which is independent of λ. (16) is always violated by taking pairs (i,α), (i,β), (i,γ) for different α,β,γ. Hence Proposition 4 of Ref. 14 is no longer valid. From (13), the formula for the velocities for the solitons, the solitons will not split as $t\to\infty$.

Remark 2: Since $u_{\beta\alpha}^*=-u_{\alpha\beta}$, we will always consider $u_{\alpha\beta}$ $(\alpha<\beta)$ from now on.

III. SINGLE-SOLITON SOLUTIONS

Take $l=1$,

$$H=\begin{pmatrix} e^{\sqrt{-1}\lambda_0(x+Jy+Kt)} & \\ & C \end{pmatrix}, \tag{17}$$

where $C=(C_{jk})$ is a 3×3 nondegenerate constant matrix, then

$$\Sigma=\frac{1}{2b\sqrt{-1}}\,(e^{-2b(\zeta+J_\mu\eta)-2b\pi_\mu t}\delta_{\mu\nu}+g_{\mu\nu})_{3\times3}, \tag{18}$$

with

$$\begin{gathered} x=v_x t+\xi,\quad y=v_y t+\eta,\\ \pi_\mu=v_x+J_\mu v_y+K_\mu,\quad a=\mathrm{Re}(\lambda_0),\quad b=\mathrm{Im}(\lambda_0),\quad g_{\mu\nu}=C_{\sigma\mu}^* C_{\sigma\nu}. \end{gathered} \tag{19}$$

Since the solitons moves in a fixed velocity for each component $u_{\alpha\beta}$, we consider the solution in a moving frame with velocity (v_x,v_y). The asymptotic behavior of the soliton is clearly seen in this frame.

When (v_x,v_y) takes the value in (13), $\pi_\alpha=\pi_\beta=0$. Without loss of generality, suppose $\pi_2=\pi_3=0$, then $\pi_1\neq0$ by considering (3).

Let $(h_{\mu\nu})=(g_{\mu\nu})^{-1}$, $g=(g_{\mu\nu})$, $h=(h_{\mu\nu})$, then by direct calculation from (12), we get the new solution of (1) as follows:

$$\begin{aligned} u_{12}&=-\frac{4b}{\Delta}\,(h_{12}\sqrt{\det g}\,e^{b\xi+\theta_3}-g_{12}\sqrt{\det h}\,e^{-b\xi-\theta_3})e^{\sqrt{-1}a(J_1-J_2)\eta+\sqrt{-1}a\pi_1 t},\\ u_{13}&=-\frac{4b}{\Delta}\,(h_{13}\sqrt{\det g}\,e^{b\xi+\theta_2}-g_{13}\sqrt{\det h}\,e^{-b\xi-\theta_2})e^{\sqrt{-1}a(J_1-J_3)\eta+\sqrt{-1}a\pi_1 t},\\ u_{23}&=-\frac{4b}{\Delta}\,(h_{23}e^{s(t)}-g_{23}e^{-s(t)})e^{\sqrt{-1}a(J_2-J_3)\eta}, \end{aligned} \tag{20}$$

where

$$\begin{aligned} \Delta&=e^{s(t)}(e^{\rho}+g_{11}\det h e^{-\rho}+h_{22}e^{\theta_3-\theta_2}+h_{33}e^{\theta_2-\theta_3})\\ &\quad+e^{-s(t)}(e^{-\rho}+h_{11}\det g e^{\rho}+g_{22}e^{\theta_2-\theta_3}+g_{33}e^{\theta_3-\theta_2}),\\ s(t)&=b(\xi+J_1\eta+\pi_1 t)+\tfrac{1}{2}\ln(\det g),\\ \rho&=b(2\xi+J_2\eta+J_3\eta),\quad \theta_j=bJ_j\eta\quad(j=1,2,3). \end{aligned} \tag{21}$$

Whenever $b\neq0$, u_{23} tends to a single localized peak as $t\to\pm\infty$, and both u_{12}, u_{13} tend to zero exponentially. More precisely, when $b\pi_1 t\to+\infty$ (ξ, η bounded),

$$u_{12}\sim 0, \quad u_{13}\sim 0,$$
$$u_{23}\sim \frac{-4bh_{23}e^{\sqrt{-1}a(J_2-J_3)\eta}}{e^{\rho}+g_{11}\det he^{-\rho}+h_{22}e^{\theta_3-\theta_2}+h_{33}e^{\theta_2-\theta_3}}, \tag{22}$$

and when $b\pi_1 t\to-\infty$ (ξ, η bounded),

$$u_{12}\sim 0, \quad u_{13}\sim 0,$$
$$u_{23}\sim \frac{4bg_{23}e^{\sqrt{-1}a(J_2-J_3)\eta}}{e^{-\rho}+h_{11}\det ge^{\rho}+g_{22}e^{\theta_2-\theta_3}+g_{33}e^{\theta_3-\theta_2}}. \tag{23}$$

The asymptotic solutions of u_{23} for $t\to-\infty$ and $t\to+\infty$ differ in the amplitudes as well as the phases. That is because there is an interaction of u_{12}, u_{13} and u_{23} around $t=0$. The total energy $\int(|u_{12}|^2+|u_{13}|^2+|u_{23}|^2)dx\,dy$ still conserves.

At an extreme case, let $h_{23}=0$ and $g_{23}\neq 0$ [this is always possible since the only constraints on $(g_{\mu\nu})$ are Hermitian and positivity], then a u_{23} wave disappears as $t\to+\infty$ although it exists as $t\to-\infty$.

On the other hand, if $(g_{\mu\nu})$ is unitary, then $h_{\mu\nu}=\overline{g_{\nu\mu}}$ and the asymptotic behavior of $|u_{23}|$ as $t\to-\infty$ and $t\to+\infty$ are the same.

The asymptotic behavior of u_{12} is similar when $\pi_1=\pi_2=0$, so is that of u_{13} when $\pi_1=\pi_3=0$.

IV. MULTISOLITON SOLUTIONS

Since the velocities of solitons do not depend on the choice of λ, Proposition 4 of Ref. 14 that is explained in Remark 1 fails. However, we can still see several peaks after twice Darboux transformations, although all these peaks move in the same velocity.

For $l>1$, we have obtained the solution $u^{[l]}_{\alpha\beta}$ given by (12). Now that the limit $t\to\pm\infty$ cannot separate these peaks, we consider the asymptotic behavior of the solutions when $C\to\infty$ in some way. This describes the separation of the peaks when the phase of each single soliton is large enough.

It is well known that H_j and $H_j\cdot\mathrm{diag}(\gamma_1^{(j)},\gamma_2^{(j)},\gamma_3^{(j)})$ ($\gamma_1^{(j)},\gamma_2^{(j)},\gamma_3^{(j)}$ are constants) give the same solution $u^{[l]}_{\alpha\beta}$ [see also (10) and (11)]. Therefore, we can rewrite H_j in (9) as

$$H_j=\begin{pmatrix} e^{\sqrt{-1}\lambda_j(x+Jy+Kt)-QL^{(j)}} & \\ & D^{(j)}\end{pmatrix}e^{QL^{(j)}}, \tag{24}$$

where $L^{(j)}=\mathrm{diag}(L_1^{(j)},L_2^{(j)},L_3^{(j)})$ are arbitrary real constant diagonal matrices, Q is a real number, and $D^{(j)}=C^{(j)}e^{-QL^{(j)}}$.

The discussion in Secs. IV and VI of Ref. 14 can be used directly here. Let $x=\xi_0+Qw_\xi$, $y=\eta_0+Q\omega_\eta$ where w_ξ, w_η are real constants (''velocity'' with respect to Q). Let

$$\begin{aligned}\sqrt{-1}\phi_{i\alpha}&\equiv\sqrt{-1}\lambda_i(x+J_\alpha y+K_\alpha t)-QL_\alpha^{(i)}\\&=\sqrt{-1}\lambda_i(\xi_0+J_\alpha\eta_0+K_\alpha t)+(\sqrt{-1}\lambda_i(w_\xi+J_\alpha w_\eta)-L_\alpha^{(i)})Q.\end{aligned}$$

Suppose

$$\det\begin{pmatrix}1 & J_\alpha & L_\alpha^{(i)}/\lambda_{iI}\\ 1 & J_\beta & L_\beta^{(j)}/\lambda_{jI}\\ 1 & J_\gamma & L_\gamma^{(k)}/\lambda_{kI}\end{pmatrix}\neq 0 \tag{25}$$

for mutually different pairs (α,i), (β,j), (γ,k). Here λ_{iI} is the imaginary part of λ_α.

Now fix $D^{(j)}$, $L^{(j)}$, and let $Q\to+\infty$ or $-\infty$. Note that

$$\mathrm{Re}(\sqrt{-1}\lambda_i(w_\xi+J_\alpha w_\eta)-L_\alpha^{(i)})=-\lambda_{iI}(w_\xi+J_\alpha w_\eta)-L_\alpha^{(i)}.$$

Therefore, Im $\phi_{i\alpha}$ keeps finite if and only if $\lambda_{iI}(w_\xi+J_\alpha w_\eta)+L_\alpha^{(i)}=0$.

The discussion in Secs. IV and VI of Ref. 14 implies the following.

Proposition 1: Suppose the matrix H_j in (9), (10), (11) can be written as

$$H_j=\begin{pmatrix} e^{QA^{(j)}+B^{(j)}} \\ D^{(j)} \end{pmatrix}$$

where Q is a parameter, $A^{(j)}=diag(A_1^{(j)},\ldots,A_N^{(j)}),B^{(j)}=diag(B_1^{(j)},\ldots,B_N^{(j)})$ are $N\times N$ diagonal matrices independent of $Q,D^{(j)}$ are $N\times N$ matrices independent of Q. Let $u_{\alpha\beta}^{[l]}$ be the solution given by l Darboux transformations. Then the following occurs.

(1) If there is at most one pair (α,i) such that $A_\alpha^{(i)}=0$, then $u_{\mu\nu}^{[l]}\to 0$ for all $\mu<\nu$ when $Q\to+\infty$ or $-\infty$.

(2) If there are exactly two pairs $(\alpha,i)\neq(\beta,j)$ $(\alpha<\beta)$ such that $A_\alpha^{(i)}=A_\beta^{(j)}=0$, then, when $Q\to+\infty$ or $-\infty$, $u_{\mu\nu}^{[l]}\to 0$ for all $\mu<\nu$ except $\mu=\alpha$, $\nu=\beta$.

For the H_j's in (24), Proposition 1 implies the following.

(1) If (w_ξ,w_η) is chosen so that there is at most one pair (α,i), such that $\lambda_{iI}(w_\xi+J_\alpha w_\eta)+L_\alpha^{(i)}=0$, then $u_{\mu\nu}^{[l]}\to 0$ for all $\mu<\nu$ when $Q\to+\infty$ or $-\infty$.

(2) If (w_ξ,w_η) is chosen so that there are exactly two pairs $(\alpha,i)\neq(\beta,j)(\alpha<\beta)$ such that

$$\begin{aligned}\lambda_{iI}(w_\xi+J_\alpha w_\eta)+L_\alpha^{(i)}=0,\\ \lambda_{jI}(w_\xi+J_\beta w_\eta)+L_\beta^{(j)}=0,\end{aligned}\tag{26}$$

i.e.,

$$w_\xi=-\frac{L_\alpha^{(i)}/\lambda_{iI}-L_\beta^{(j)}/\lambda_{jI}}{J_\alpha-J_\beta},\quad w_\eta=-\frac{J_\alpha L_\beta^{(j)}/\lambda_{jI}-J_\beta L_\alpha^{(i)}/\lambda_{iI}}{J_\alpha-J_\beta},\tag{27}$$

then, when $Q\to+\infty$ or $-\infty$, $u_{\mu\nu}^{[l]}\to 0$ for all $\mu<\nu$ except $\mu=\alpha$, $\nu=\beta$.

(3) If (25) holds for mutually different pairs (α,i), (β,j), (γ,k), then for any (w_ξ,w_η), there are at most two pairs (α,i), (β,j) such that (26) holds.

Therefore, for any α, β, unless (w_ξ,w_η) takes the value in (27), $u_{\alpha\beta}^{[l]}\to 0$. Since i,j in (27) can be $1,\ldots,l$, respectively, $u_{\alpha\beta}^{[l]}$ has at most l^2 peaks. Thus, we have the following.

Theorem 1: *Let λ_j $(j=1,\ldots,l)$ be complex numbers such that $\lambda_j\neq\lambda_k(j\neq k),\lambda_j\neq\overline{\lambda_k}$ for all $1\leqslant j,k\leqslant l$. Let $L^{(j)}$'s be 3×3 constant diagonal matrices such that (25) holds, $D^{(j)}$'s be 3×3 nondegenerate constant matrices. Let $C^{(j)}=D^{(j)}e^{QL^{(j)}}$ and construct H_j as in (9). Then as $Q\to+\infty$ or $-\infty$, each component $u_{\alpha\beta}^{[l]}$ of the solution derived by the Darboux transformation of degree l has at most l^2 peaks for fixed time t.*

Remark 3: Generally, when (26) holds, $u_{\alpha\beta}^{[l]}$ tends to a nonzero function unless its coefficient tends to zero. Hence, usually there are exact l^2 peaks for each $u_{\alpha\beta}^{[l]}$ when Q is large enough. Of course, if a certain coefficient vanishes (e.g., $h_{23}=0$ or $g_{23}=0$ for u_{23} in a single-soliton case), the corresponding peak disappears.

When we only consider the asymptotic behavior of $u_{\alpha\beta}^{[l]}$, a condition weaker than (25) is enough to separate the peaks as $Q\to\infty$. This is because as $t\to\infty$, the interaction between $u_{\alpha\beta}^{[l]}$ and any other $u_{\mu\nu}^{[l]}$ vanishes.

As in Sec. III consider u_{23} and choose the velocity such that $\pi_2=\pi_3=0$, then $\pi_1\neq 0$. Take the coordinates ξ, η given by (19). Since $b_j\equiv\lambda_{jI}\neq 0$, $b_j\pi_1 t\to\infty$ as $t\to\infty$. Therefore, the limit Darboux matrix can be constructed by

$$H_j(-\infty)=\begin{pmatrix} 0 & 0 & 0 \\ 0 & e^{\sqrt{-1}\lambda_j(\xi+J_2\eta)-QL_2^{(j)}} & 0 \\ 0 & 0 & e^{\sqrt{-1}\lambda_j(\xi+J_3\eta)-QL_3^{(j)}} \\ D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix}$$

if $b\pi_1 t\to-\infty$, or

$$H_j(+\infty)=\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{\sqrt{-1}\lambda_j(\xi+J_2\eta)-QL_2^{(j)}} & 0 \\ 0 & 0 & e^{\sqrt{-1}\lambda_j(\xi+J_3\eta)-QL_3^{(j)}} \\ 0 & D_{12} & D_{13} \\ 0 & D_{22} & D_{23} \\ 0 & D_{32} & D_{33} \end{pmatrix}$$

if $b\pi_1 t\to+\infty$. Now let $\xi=\xi_0+Qw_\xi$, $\eta=\eta_0+Qw_\eta$. Clearly, if $(H_j(-\infty))_{(1,1)}$ is replaced by $e^{-|Q|}$ and $(H_j(+\infty))_{(1,1)}$ is replaced by $e^{|Q|}$, the limit Darboux transformation is the same when $Q\to\infty$. Then, the coefficients of Q in the power of $(H_j(\pm\infty))_{(1,1)}$ are always nonzero. Therefore, according to Proposition 1, we only need to guarantee that (25) holds for $(\alpha,i)\neq(\beta,j)$ with $\alpha,\beta\in\{2,3\}$ so that $u_{\alpha\beta}^{[I]}(-\infty)$, the asymptotic solution for $t\to-\infty$, does not tend to zero, and $u_{\mu\nu}^{[I]}(-\infty)\to0$ for any $(\mu,\nu)\neq(\alpha,\beta)$. This implies that we only need that (25) holds for all mutually different pairs (α,i), (β,j), (γ,k) with $\alpha,\beta,\gamma\in\{2,3\}$.

Since α, β, γ can only take two values and J_α's are mutually different, (25) is equivalent to

$$L_\alpha^{(i)}/\lambda_{iI}\neq L_\alpha^{(j)}/\lambda_{jI}\quad(\alpha=2,3,\quad i\neq j).$$

The same argument can be applied to $u_{23}^{[I]}(+\infty)$, the asymptotic solution for $t\to+\infty$.

Clearly, there are corresponding conclusions for $u_{12}^{[I]}$, $u_{13}^{[I]}$. Therefore, we have the following.

Theorem 2: *Given two integers α, β with $1\leqslant\alpha<\beta\leqslant3$. Let λ_j $(j=1,...,l)$ be as in Theorem 1, $D^{(j)}$'s be 3×3 nondegenerate constant matrices. Let $L^{(j)}$'s be 3×3 constant diagonal matrices such that $L_\alpha^{(i)}/\lambda_{iI}\neq L_\alpha^{(j)}/\lambda_{jI}$, $L_\beta^{(i)}/\lambda_{iI}\neq L_\beta^{(j)}/\lambda_{jI}$ for $i\neq j$. Let $C^{(j)}=D^{(j)}e^{QL^{(j)}}$ and construct H_j as in (9). Then as $Q\to+\infty$ or $-\infty$, the asymptotic solutions $u_{\alpha\beta}^{[I]}(-\infty)$ and $u_{\alpha\beta}^{[I]}(+\infty)$ have at most l^2 peaks respectively.*

Remark 4: Both Theorem 1 and Theorem 2 are valid for the N-wave equation, except that 3 is replaced by N. Here we discuss three-wave equation for its simplicity and importance.

Remark 5: Return to the case of single-soliton solutions. In (20), (21), the phase of the soliton is characterized by

$$\phi_1=\ln(e^{s^{(j)}(t)}+e^{-s^{(j)}(t)}h_{11}^{(j)}\det g^{(j)})-\ln(e^{-s^{(j)}(t)}+e^{s^{(j)}(t)}g_{11}^{(j)}\det h^{(j)}),$$

$$\phi_2=\ln(e^{s^{(j)}(t)}h_{22}^{(j)}+e^{-s^{(j)}(t)}g_{33}^{(j)})-\ln(e^{-s^{(j)}(t)}g_{22}^{(j)}+e^{s^{(j)}(t)}h_{33}^{(j)}).$$

If we let $L_{\alpha\alpha}^{(j)}=g_{\alpha\alpha}^{(j)}/Q$, then, when $L^{(j)}$ is fixed, $Q\to+\infty$ implies $\phi_1\to+\infty$ and ϕ_2 is bounded.

Here we show some figures that describe the behavior of the solitons given by a Darboux transformation of degree 2. Figures 1, 3, 5, and 7 correspond to time $t=-16$, 0, 12, and 16, respectively. As discussed above, these solitons have at most four peaks. Here the vertical axis is $|u|^{1/2}\equiv(|u_{12}|^2+|u_{13}|^2+|u_{23}|)^{1/4}$. Therefore, all three components are shown in one figure. The corresponding parameters are

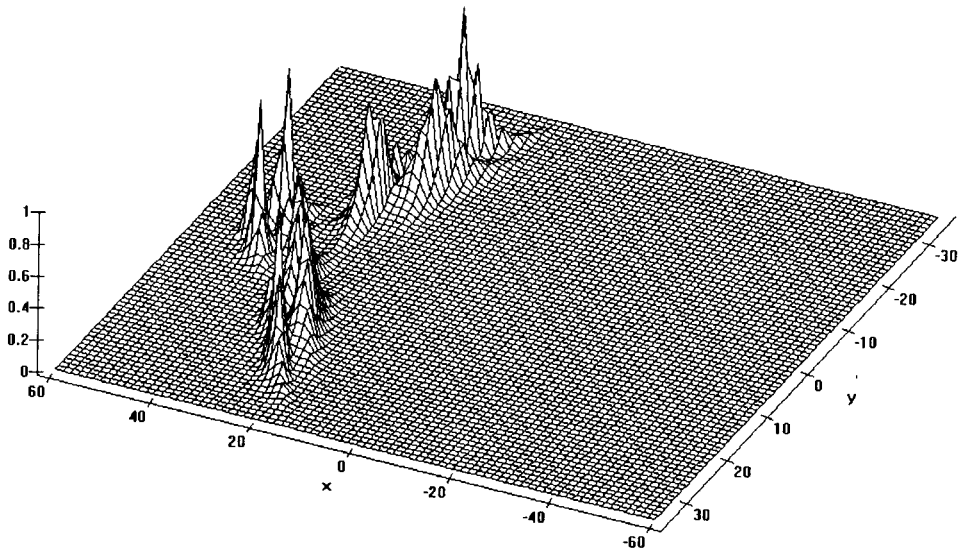
1
0.8
0.6
0.4
0.2
0
60
40
20
0
-20
-40
-60
x
-30
-20
-10
0
10
20
30
y

FIG. 1

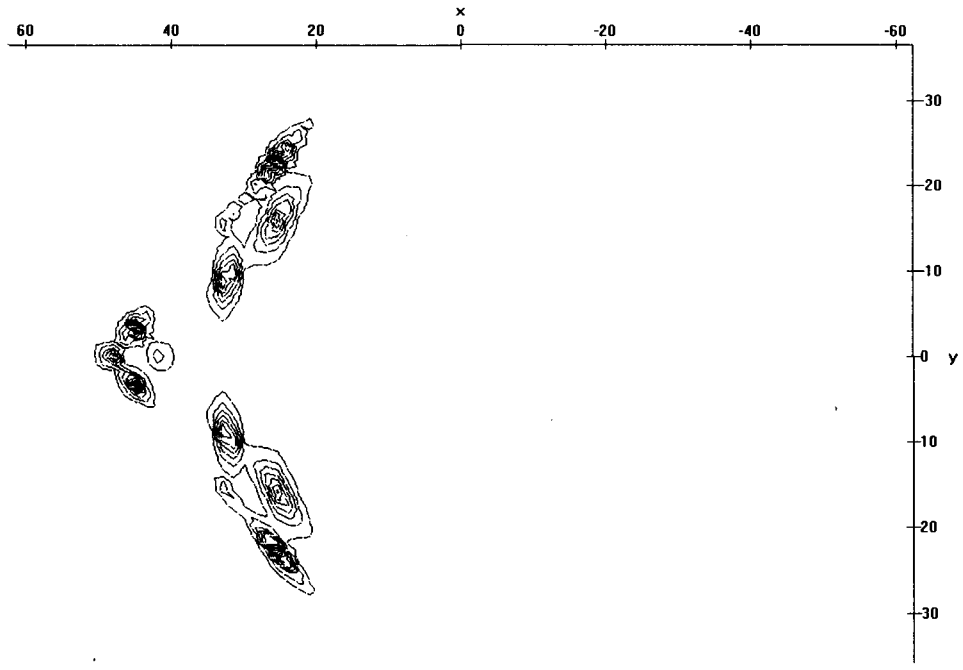
x
60
40
20
0
-20
-40
-60
30
20
10
0
y
10
20
30

FIG. 2

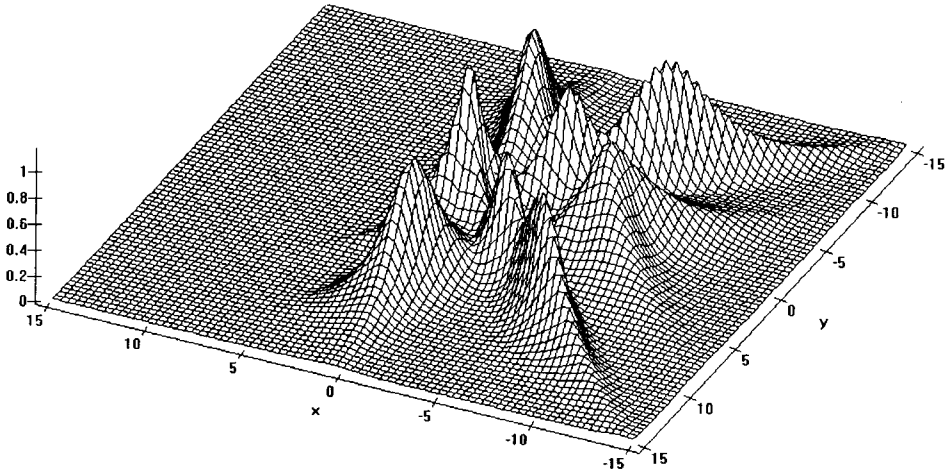
1
0.8
0.6
0.4
0.2
0
15
10
5
0
-5
-10
-15
x
-15
-10
-5
0
5
10
15
y

FIG. 3

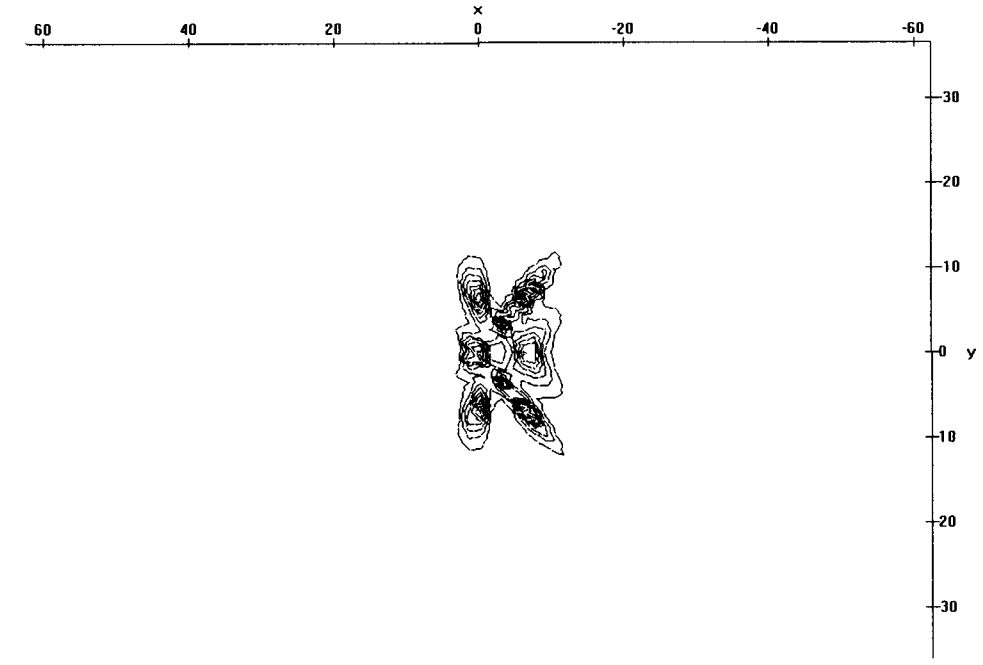
x
60
40
20
0
-20
-40
-60
30
20
10
0
-10
-20
-30
y

FIG. 4

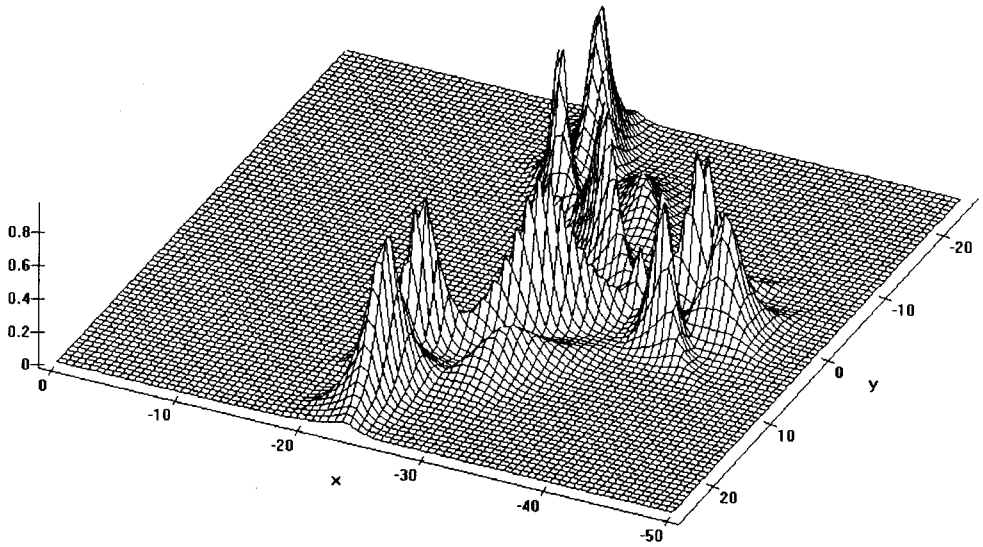
0.8
0.6
0.4
0.2
0
0
-10
-20
x
-30
-40
-50
-20
-10
0
y
10
20

FIG. 5

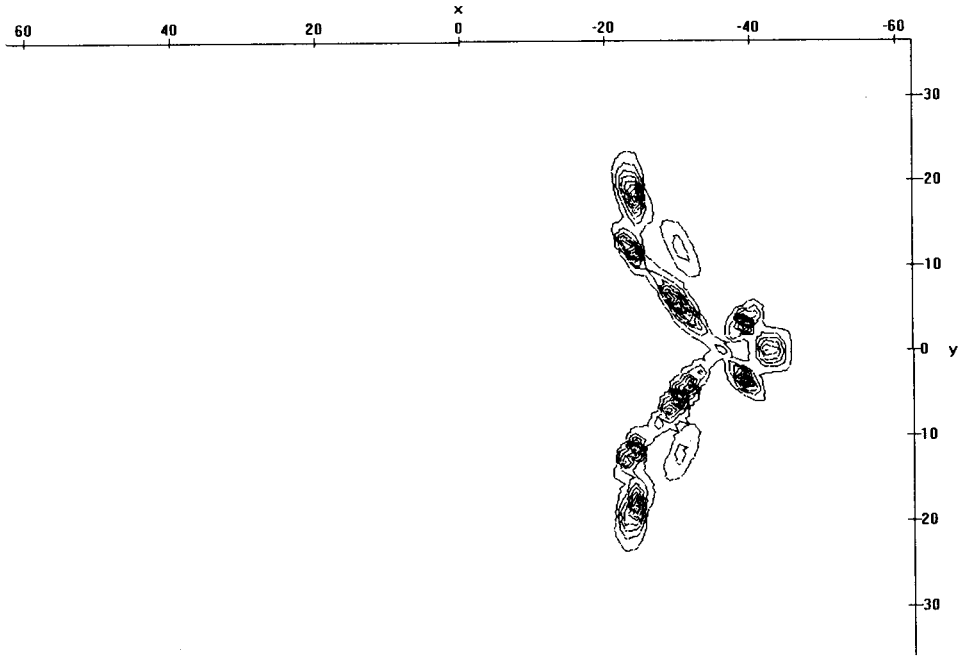
x
60
40
20
0
-20
-40
-60
30
20
10
0
y
10
20
30

FIG. 6

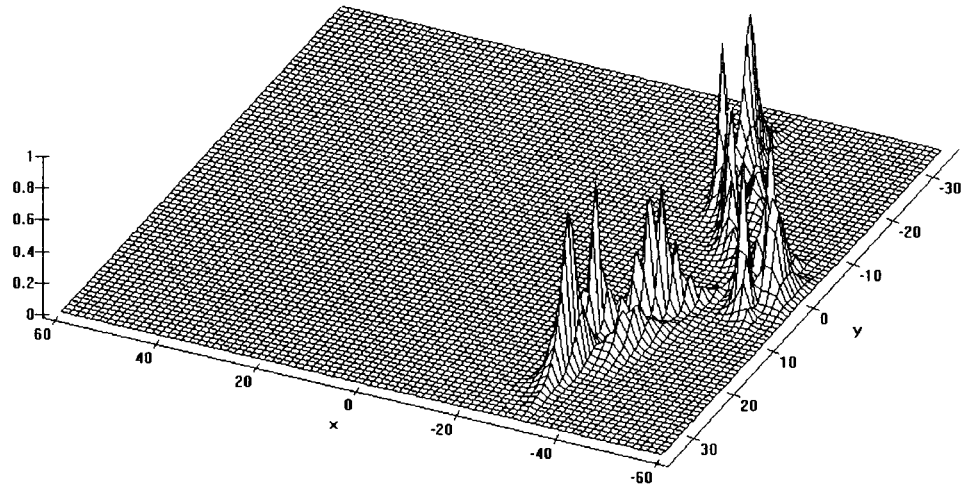
1
0.8
0.6
0.4
0.2
0
60
40
20
0
-20
-40
-60
x
-30
-20
-10
0
y
10
20
30

FIG. 7

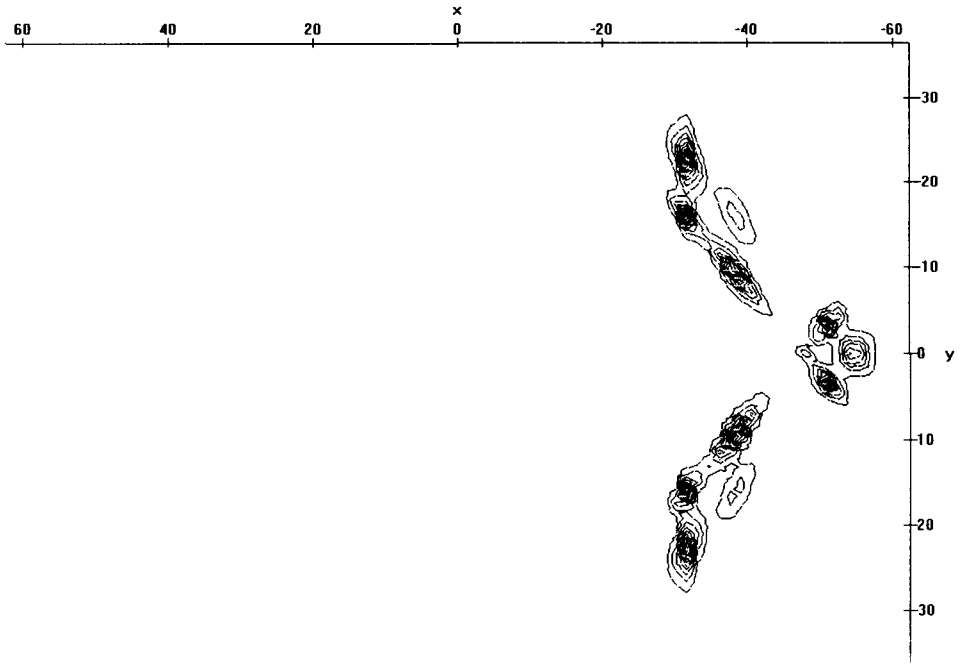
x
60
40
20
0
-20
-40
-60
30
20
10
0
y
10
20
30

FIG. 8

$$J=\begin{pmatrix}1&0&0\\0&0&0\\0&0&-1\end{pmatrix},$$

$$K=\begin{pmatrix}3&0&0\\0&2&0\\0&0&3\end{pmatrix},$$

$$\lambda_1=1-2\sqrt{-1},\quad \lambda_2=2+\sqrt{-1},$$

$$C_2=\begin{pmatrix}1&0&1\\1&1&0\\0&1&1\end{pmatrix},\quad C_1=256C_2.$$

Figures 2, 4, 6, and 8 are the corresponding contour plot of Figs. 1, 3, 5, and 7. Note that in Fig. 3 and Fig. 5, the scales of the coordinates are different from the others so that the peaks can be seen more clearly.

ACKNOWLEDGMENTS

This work was supported by Chinese National Research Project ''Nonlinear Science,'' Chinese National Science Foundation for Youth, and the Fok Ying-Tung Education Foundation of China. The author would like to thank Professor Chaohao Gu for his encouragement.

[1] V. E. Zakharov and A. B. Shabat, ''A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem,'' Funct. Anal. Appl. **8**, 226–235 (1974).
[2] A. S. Fokas and L. Y. Sung, ''On the solvability of the N-wave, Davey–Stewartson and Kadomtsev–Petviashvili equations,'' Inverse Probl. **8**, 673–708 (1992).
[3] A. S. Fokas and V. E. Zakharov, ''The dressing method and nonlocal Riemann–Hilbert problems,'' J. Nonlinear Sci. **2**, 109–134 (1992).
[4] A. I. Nachman and M. J. Ablowitz, ''Multidimensional inverse scattering for first-order systems,'' Stud. Appl. Math. **71**, 251–262 (1984).
[5] F. Pempinelli, ''New features of soliton dynamics in 2+1 dimensions,'' Acta Appl. Math. **39**, 445 (1995).
[6] L. Y. Sung and A. S. Fokas, ''Inverse problem for N×N hyperbolic systems on the plane and the N-wave interactions,'' Commun. Pure Appl. Math. **44**, 535–571 (1991).
[7] M. Boiti, F. Pempinelli, and P. C. Sabatier, ''First and second order nonlinear evolution equations from an inverse spectral problem,'' Inverse Probl. **9**, 1–37 (1993).
[8] Y. Cheng, ''Symmetries and hierarchies of equations for the (2+1)-dimensional N-wave interaction,'' Physica D **34**, 277–288 (1989).
[9] C. W. Cao, ''Nonlinearization of Lax equations of AKNS system,'' Chinese Sci. A, 701–707 (1989).
[10] Y. Cheng and Y. S. Li, ''The constraint of the KP equation and its special solutions,'' Phys. Lett. A **157**, 22 (1991).
[11] B. Konopelchenko, J. Sidorenko, and W. Strampp, ''(1+1)-dimensional integrable systems as symmetry constraints of (2+1) dimensional systems,'' Phys. Lett. A **157**, 17 (1991).
[12] C. H. Gu, ''On the interaction of solitons for a class of integrable systems in the space–time $\mathbf{R}^{n+1}$,'' Lett. Math. Phys. **26**, 199–209 (1992).
[13] C. H. Gu and Z. X. Zhou, ''On Darboux transformations for soliton equations in high dimensional space-time,'' Lett. Math. Phys. **32**, 1–10 (1994).
[14] Z. X. Zhou, ''Soliton solutions for some equations in 1+2 dimensional hyperbolic $su(N)$ AKNS system,'' Inverse Probl. **12**, 89–109 (1996).
[15] Z. X. Zhou, ''Explicit solutions of N-wave equation in 1+2 dimensions,'' Phys. Lett. A **168**, 370–374 (1992).

Classification of multidimensional inflationary models

S. Mignemi
Dipartimento di Matematica, Università di Cagliari,
viale Merello 92, I-09123 Cagliari, Italy

H.-J. Schmidt
Universität Potsdam, Institut für Mathematik,
D-14415 Potsdam, PF 601553, Am Neuen Palais 10, Germany

(Received 12 May 1997; accepted for publication 6 August 1997)

We define under which circumstances two multiwarped product spacetimes can be considered equivalent and then we classify the spaces of constant curvature in the Euclidean and Lorentzian signature. For dimension $D=2$, we get essentially twelve representations, for $D=3$ exactly eighteen. More generally, for every even D, $5D+2$ cases exist, whereas for every odd D, $5D+3$ cases exist. For every D, exactly one half of them has the Euclidean signature. Our definition is well suited for the discussion of multidimensional cosmological models, and our results give a simple algorithm to decide whether a given metric represents the inflationary de Sitter spacetime (in unusual coordinates) or not. © *1998 American Institute of Physics.* [S0022-2488(98)02301-9]

I. INTRODUCTION

Recently, a lot of papers have dealt with multidimensional cosmological models, cf., e.g. Refs. 1–3 and references cited there. In these works, the ansatz for the metric is a generalized warped product of several spaces, with the warping functions depending on the time t only, i.e.,

$$ds^2=dt^2-\sum_{k=1}^{n} a_k^2(t)d\sigma_k^2, \tag{1.1}$$

where each $d\sigma_k^2$ represents a Riemannian space of dimension $d_k\geqslant 1$. In the usual interpretation, one requires $d_1=3$ and $d\sigma_1^2$ to be the physical 3-space, whereas all other spaces $d\sigma_k^2$, $k\geqslant 2$ are internal spaces. However, our approach is not restricted to this interpretation.

One of the often discussed questions in this context is the appearance of an inflationary phase of cosmic evolution. Its geometry is usually represented by a spacetime of constant negative curvature, called de Sitter spacetime. It is well-known how to represent the de Sitter spacetime of arbitrary dimension in the form (1.1), but up to now there does not exist a classification of the metrics of the general form (1.1) to decide which of them corresponds to spaces of constant curvature.

To elucidate what this means, let us consider the metrics (1.2) and (1.3),

$$ds^2=dt^2-\sinh^2t dx^2-\cosh^2t(dy^2+\sin^2y dz^2) \tag{1.2}$$

[see, e.g., Ref. 4, Eq. (3.18)]. The metric (1.2) is of the type (1.1) and is obviously a cosmological model of Kantowski–Sachs type. The metric (1.2) is known to represent a vacuum solution of Einstein's equations with a positive cosmological term. In agreement with the cosmological no-hair theorem it is asymptotically de Sitter. In the coordinates used, the metric (1.2) seems to define an anisotropic model, however (cf. Ref. 4) it is nothing but a piece of the isotropic de Sitter spacetime in unusual coordinates, and moreover, the coordinate transformation necessary to bring it to the classical form is quite involved.

A related metric is (cf. Ref. 5, Eq. (4.6)),

$$ds^2=dt^2+\sin^2t dx^2+\cos^2t dy^2, \tag{1.3}$$

 © 1998 American Institute of Physics

which is the metric of the standard 3-sphere in unusual coordinates.

It is the purpose of the present paper to give a complete classification of all the metrics of type (1.1) which represent spaces of constant curvature.

The paper is organized as follows: Section II introduces the generalized warped products and defines the normal form of the metric (1.1). Section III gives the results for the dimension $D\leqslant 3$ and Section IV covers the remaining cases $D\geqslant 4$. In Sec. V, we consider a slightly different form of the metric, where the a_k can additionally depend on one of the spatial coordinates, but the $d\sigma_k^2$ have dimension 1. Section VI discusses the results.

II. GENERALIZED WARPED PRODUCTS

Usually, a warped product is the conformally transformed Cartesian product between two Riemannian spaces, where the conformal factor, called the warping function, depends on the coordinates of one of the factors only. If there are more than two factor spaces, then there exist several possible generalizations of that notion. We will restrict ourselves to the following types of generalized warped products: all $a_k>0$ and

$$ds^2=dt^2\pm\sum_{k=1}^{n} a_k^2(t)d\sigma_k^2 \tag{2.1}$$

and each $d\sigma_k^2$ is a Riemannian space of dimension $d_k\geqslant 1$. Thus we consider the Euclidean (upper sign) and the Lorentzian (lower sign) signatures of the metric. In this way we are able to cover both Eqs. (1.2) and (1.3). Obviously, the metric (2.1) represents a usual warped product for $n=1$. The dimension of ds^2 is $D=1+\Sigma d_k$. We want to find out under which circumstances the metric (2.1) represents a space of constant curvature locally (so we do not discuss whether zeroes of the functions $a_k(t)$ give rise to coordinate singularities and whether our metrics are geodesically complete).

To get a uniquely defined classification, we start by explaining under which circumstances two metrics of type (2.1) will be considered to be the same:

(A) Any constant $c_k>0$ can be shifted between $a_k^2(t)$ and $d\sigma_k^2$.
(B) ds^2 can be multiplied by a positive constant, or in other words: t and every $a_k(t)$ can be multiplied by the same positive factor.
(C) Linear transformations of t will be allowed.
(D) Permutations of the indices k are allowed.
(E) Coordinate transformations involving only the coordinates of one of the spaces $d\sigma_k^2$ are allowed.
(F) If two warping functions coincide, $a_j(t)=a_k(t)$, $(j\neq k)$, then $a_j^2(t)d\sigma_j^2+a_k^2(t)d\sigma_k^2$ can be replaced by $a_j^2(t)(d\sigma_j^2+d\sigma_k^2)$.

Comment: (B) means that we do not distinguish between homothetically equivalent spaces, or, in other words, a change of the length unit will not be considered essential. (F) means that the number n of terms in the sum of Eq. (2.1) is not fixed. However, we can define the normal form of the representation (2.1) as follows: whenever possible (if necessary by the help of introducing suitable constants c_k according to (A), the replacement (F) shall be done (of course, this procedure terminates uniquely). Equivalently we may say: the normal form of the representation is that one where the number n is minimal. Furthermore it holds that the metric (2.1) is in normal form if the quotients $a_j(t)/a_k(t)$ are constant for $k=j$ only.

For technical reasons, we will sometimes discuss spaces in non-normal form. However, if the number n of factors is explicitly mentioned, we always refer this to the normal form and, without loss of generality, we may restrict to metrics (2.1) in normal form.

III. CLASSIFICATION FOR DIMENSION $D\leqslant 3$

$D=1$ implies $n=0$ and $ds^2=dt^2$ represents flat space. So we have just one representation for $D=1$.

A. Dimension $D=2$

Let $D=2$, then $d\sigma_1^2=dx^2$ is flat. (Spaces of constant curvature in $D=2$ have been discussed, for example, in Ref. 6.)

We write $a(t)$ instead of $a_1(t)$ and get

$$ds^2=dt^2\pm a^2(t)dx^2, \tag{3.1}$$

which is already in normal form. This space is of constant curvature iff the curvature scalar $R=-2\ddot{a}/a$ is constant, where a dot denotes d/dt. Both signatures can be considered simultaneously, because they are related by the imaginary coordinate transformation $x\to ix$. (Here we give the derivation for the known case $D=2$ in detail in order to show how our formulation works, we do not repeat the analogous arguments in the other cases.) According to (B), we may restrict to the three cases $R=0$, $R=2$ and $R=-2$. For $R=0$, $a(t)$ must be linear in t, and according to (A) and (C), the following representations are singled out:

$$ds^2=dt^2\pm dx^2 \tag{3.2}$$

and

$$ds^2=dt^2\pm t^2dx^2. \tag{3.3}$$

Clearly, these are the Cartesian and polar coordinates respectively for the flat space. For $R=2$, $\ddot{a}+a=0$ has to be solved. According to (A) and (C), only one representation, the standard metric of the 2-sphere, appears:

$$ds^2=dt^2\pm\sin^2 t dx^2. \tag{3.4}$$

For $R=-2$, however, three different representations appear for each of the two signatures: (By condition (C) the solution e^{-2t} can be excluded.)

$$ds^2=dt^2\pm e^{2t}dx^2 \tag{3.5}$$

further

$$ds^2=dt^2\pm\sinh^2 t dx^2 \tag{3.6}$$

and

$$ds^2=dt^2\pm\cosh^2 t dx^2. \tag{3.7}$$

An explicit coordinate transformation between the metrics (3.5) and (3.6) is deduced in the appendix of Ref. 4. Metric (3.4) and (3.6) are related by multiplying s, t and x by i.

Thus we have exactly twelve different representations of a space of constant curvature for $D=2$. Six of them have $R<0$ and only two have $R>0$.

B. Dimension $D=3$: First part

Let $D=3$. The space is of constant curvature iff the three eigenvalues of the Ricci tensor coincide. Let us start recalling the known cases. The Cartesian product of flat spaces is flat. So we can get from Eqs. (3.2), (3.3) for the $R=0$ cases:

$$n=1,\quad ds^2=dt^2\pm(dx^2+dy^2) \tag{3.8}$$

and

$$n=2,\quad ds^2=dt^2\pm(t^2dx^2+dy^2) \tag{3.9}$$

For $R>0$ we know the examples

$$n=1,\quad ds^2=dt^2+\sin^2 t(dx^2+\sin^2 x dy^2) \tag{3.10}$$

and

$$n=2, \quad ds^2=dt^2\pm(\sin^2 t dx^2+\cos^2 t dy^2). \tag{3.11}$$

Here, (3.10) is the standard three-sphere and (3.11) is the metric (1.3).[5] Replacing x by ix, we get from the metric (3.10) the Lorentzian signature with $R>0$:

$$n=1, \quad ds^2=dt^2-\sin^2 t(dx^2+\sinh^2 x dy^2). \tag{3.12}$$

For $R<0$ we get analogously, as in Section III A:

$$n=1, \quad ds^2=dt^2\pm e^{2t}(dx^2+dy^2) \tag{3.13}$$

and

$$n=2, \quad ds^2=dt^2\pm(\sinh^2 t dx^2+\cosh^2 t dy^2) \tag{3.14}$$

(cf. Ref. 5, Eq. (3.9)). So, we have again twelve different cases. However, before checking their completeness in Section III D, we have to discuss the case $n=1$ in the next section.

C. The case $n=1$ in arbitrary dimension

Let $D\geqslant 3$ be arbitrary and be $n=1$ in Eq. (2.1). So the index 1 may be omitted and $d=D-1$. We write $d\sigma^2=h_{\alpha\beta}dx^\alpha dx^\beta$, where $h_{\alpha\beta}$ depends on the spatial coordinates only, $\alpha,\beta=1,\ldots,d$. For $ds^2=dt^2\pm a^2(t)d\sigma^2$ we get $R_{00}=-d\ddot{a}/a$. Indices $j,k,\ldots$ take the D values $0,1,\ldots,d$. If ds^2 is a space of constant curvature, then it must be an Einstein space, i.e. Ricci tensor and metric tensor differ by a constant factor only. So we may assume, without loss of generality due to condition (B), that

$$R_{jk}=\lambda d g_{jk} \quad \text{with } \lambda\in\{-1,1,0\} \tag{3.15}$$

Now, $g_{00}=1$ and we set $\lambda=-\ddot{a}/a$. So we are in a case similar to that of Sec. III A. Consequently, only the same six functions $a(t)$ may appear: $a=1$ and $a=t$ for $R=0$, $a=\sin t$ for $R>0$, and $a=e^t$, $a=\sinh t$, $a=\cosh t$ for $R<0$.

Let us now apply the $\alpha\beta$-component of Eq. (3.15), and let $P_{\alpha\beta}$ denote the Ricci tensor of $d\sigma^2$. We get:

$$P_{\alpha\beta}=R_{\alpha\beta}\pm h_{\alpha\beta}[a\ddot{a}+(d-1)\dot{a}^2]. \tag{3.16}$$

From (3.15) and (3.16) we get

$$P_{\alpha\beta}=\Lambda h_{\alpha\beta}, \quad \Lambda=\pm(d-1)(\dot{a}^2+\lambda a^2). \tag{3.17}$$

It is easy to see that Λ is a constant. Concluding, for $n=1$, $d\sigma^2$ must be an Einstein space of constant curvature scalar.

D. Dimension $D=3$: Second part

Now we are ready to complete the discussion for $D=3$. For $n=1$, we have $d=2$, and by Sec. III C, $d\sigma^2$ has constant curvature scalar. This means that $d\sigma^2$ is of constant curvature here, and (3.17) reads $\dot{a}^2+\lambda a^2=\pm\Lambda$. $\lambda=0$, $a=1$ gives the metric (3.8). $\lambda=0$, $a=t$, gives $\Lambda=\pm 1$, namely

$$n=1, \quad ds^2=dt^2+t^2(dx^2+\sin^2 x dy^2) \tag{3.18}$$

which is flat space in spherical coordinates, and

$$n=1, \quad ds^2=dt^2-t^2(dx^2+\sinh^2 x dy^2), \tag{3.19}$$

$\lambda=1$, $a=\sin\ t$ gives $\Lambda=\pm 1$, i.e., the metric (3.10) and (3.12). $\lambda=-1$, $a=e^t$ yields $\Lambda=0$, i.e., the metric (3.13). $\lambda=-1$, $a=\sinh t$ yields $\Lambda=\pm 1$, i.e.,

$$n=1, \quad ds^2=dt^2+\sinh^2\ t(dx^2+\sin^2\ xdy^2) \tag{3.20}$$

and

$$n=1, \quad ds^2=dt^2-\sinh^2\ t(dx^2+\sinh^2\ xdy^2). \tag{3.21}$$

Finally, $\lambda=-1$, $a=\cosh t$ yields $\Lambda=\mp 1$, i.e.,

$$n=1, \quad ds^2=dt^2+\cosh^2\ t(dx^2+\sinh^2\ xdy^2) \tag{3.22}$$

and

$$n=1, \quad ds^2=dt^2-\cosh^2\ t(dx^2+\sin^2\ xdy^2). \tag{3.23}$$

This completes the discussion for $n=1$, and the six metrics (3.18)–(3.23) together with the twelve metrics of Section III B form already eighteen different representations of a space of constant curvature in $D=3$. It remains to check completeness for the case $n=2$. So we have to use the ansatz

$$ds^2=dt^2\pm(a^2(t)\ dx^2+b^2(t)\ dy^2) \tag{3.24}$$

and have to ensure that the three eigenvalues of the Ricci tensor of (3.24) coincide. This is equivalent to

$$\frac{\ddot{a}}{a}=\frac{\ddot{b}}{b}=\frac{\dot{a}}{a}\frac{\dot{b}}{b}. \tag{3.25}$$

A detailed investigation of Eq. (3.25) shows that all solutions are already covered by the eighteen cases mentioned before.

Let us summarize the results for $D=3$: six representations with $n=2$ and twelve representations with $n=1$ form a complete classification. Six of these are flat, four have positive curvature and the remaining eight have $R<0$.

IV. CLASSIFICATION FOR DIMENSION $D\geqslant 4$

From now on, we put $D\geqslant 4$. It holds: ds^2 is of constant curvature iff it is an Einstein space with vanishing Weyl tensor. In Sec. IV A we review the known cases, and in Sec. IV B we give a general account of conformal flatness.

A. Dimension $D\geqslant 4$: First part

Let $\delta_{\alpha\beta}$ be the Kronecker tensor. The generalization of Eqs. (3.8), (3.9) with $R=0$ reads

$$n=1, \quad ds^2=dt^2\pm\delta_{\alpha\beta}dx^\alpha dx^\beta \tag{4.1}$$

and

$$n=2, \quad ds^2=dt^2\pm(t^2dx^2+\delta_{\alpha\beta}dx^\alpha dx^\beta). \tag{4.2}$$

Generalizing (3.18), (3.19) with $R=0$ yields

$$n=1, \quad ds^2=dt^2+t^2d\Omega_k^2 \tag{4.3}$$

and

$$n=1, \quad ds^2=dt^2-t^2d\overline{\Omega_k^2} \tag{4.4}$$

where $k=D-1$ and $d\Omega_k^2$ is the metric of the standard sphere of dimension k, while $d\overline{\Omega_k^2}$ denotes the corresponding space of constant negative curvature.

For $R>0$ we get from (3.10) and (3.12)

$$n=1, \quad ds^2=dt^2+\sin^2 t d\Omega_k^2 \tag{4.5}$$

and

$$n=1, \quad ds^2=dt^2-\sin^2 t d\overline{\Omega_k^2}, \tag{4.6}$$

Equation (4.6) is the anti-de Sitter spacetime written as an open Friedman model for $D=4$.

For $R<0$, we get from (3.13)

$$n=1, \quad ds^2=dt^2\pm e^{2t}\delta_{\alpha\beta}dx^\alpha dx^\beta \tag{4.7}$$

and from (3.20)–(3.23)

$$n=1, \quad ds^2=dt^2+\sinh^2 t d\Omega_k^2 \tag{4.8}$$

and

$$n=1, \quad ds^2=dt^2-\sinh^2 t d\overline{\Omega_k^2}. \tag{4.9}$$

Further,

$$n=1, \quad ds^2=dt^2+\cosh^2 t d\overline{\Omega_k^2} \tag{4.10}$$

and

$$n=1, \quad ds^2=dt^2-\cosh^2 t d\Omega_k^2. \tag{4.11}$$

The metric (4.11) represents the de Sitter spacetime as a closed Friedman model for $D=4$. So, for every $D\geqslant 4$, we have found twelve representations with $n=1$, namely Eqs. (4.1) and (4.3)–(4.11).

A further generalization of (4.3)–(4.4) is possible with $R=0$ and $2\leqslant k\leqslant D-2$,

$$n=2, \quad ds^2=dt^2+t^2 d\Omega_k^2+\delta_{\alpha\beta}dx^\alpha dx^\beta \tag{4.12}$$

and

$$n=2, \quad ds^2=dt^2-t^2 d\overline{\Omega_k^2}-\delta_{\alpha\beta}dx^\alpha dx^\beta \tag{4.13}$$

Together with (4.2) this yields $2(D-2)$ representations with $R=0$, $n=2$.

For $D=4$, the following representations are known, cf. Eq. (1.2) above and Ref. 4 for details.

$R>0$,

$$n=2, \quad ds^2=dt^2+\sin^2 t dx^2+\cos^2 t d\Omega_2^2 \tag{4.14}$$

and

$$n=2, \quad ds^2=dt^2-\sin^2 t dx^2-\cos^2 t d\overline{\Omega_2^2}. \tag{4.15}$$

$R<0$,

$$n=2, \quad ds^2=dt^2+\sinh^2 t dx^2+\cosh^2 t d\overline{\Omega_2^2} \tag{4.16}$$

and

$$n=2, \quad ds^2=dt^2-\sinh^2 t dx^2-\cosh^2 t d\Omega_2^2. \tag{4.17}$$

Further,

$$n=2, \quad ds^2=dt^2+\cosh^2 t dx^2+\sinh^2 t d\Omega_2^2 \tag{4.18}$$

and

$$n=2, \quad ds^2=dt^2-\cosh^2 t dx^2-\sinh^2 t d\overline{\Omega}_2^2. \tag{4.19}$$

So, for $D=4$, we have already 10 representations with $n=2$. By the way, the metric (4.17) is the same as (1.2).

B. Conformal flatness

Spaces of constant curvature are conformally flat, and for $D\geqslant 4$, conformal flatness is equivalent to the vanishing of the conformally invariant Weyl tensor C^i_{jkl}, defined as

$$C_{ijkl}=R_{ijkl}-\frac{1}{D-2}(R_{ik}g_{jl}+R_{jl}g_{ik}-R_{jk}g_{il}-R_{il}g_{jk})+\frac{R}{(D-1)(D-2)}(g_{ik}g_{jl}-g_{jk}g_{il}). \tag{4.20}$$

Now, let $n=1$, i.e., $ds^2=dt^2\pm a^2(t)d\sigma^2$. According to Sec. III C, $d\sigma^2$ must be an Einstein space. But ds^2 is conformally flat, hence $d\hat{s}^2=ds^2/a^2(t)$ is conformally flat too. After a time-rescaling, we set

$$d\hat{s}^2=d\hat{t}^2\pm d\sigma^2.$$

Inserting this metric into Eq. (4.20), applying $\hat{C}_{ijkl}=0$ and the fact that $d\sigma^2$ is an Einstein space, gives as a result that $d\sigma^2$ must be a space of constant curvature. This proves that our classification for the case $n=1$ given in Sec. IV A is already complete.

Let $n\geqslant 2$ in the following. We want to show that every $d\sigma_k^2$ is a space of constant curvature. For $d_k=1$, this is trivial. For any $d_k\geqslant 2$ we denote d_k by d and $D-d$ by m. Because of $n\geqslant 2$ we have $m\geqslant 2$. Let $d\sigma_k^2=h_{\alpha\beta}(x^\alpha)dx^\alpha dx^\beta$ with $\alpha,\beta=1,\ldots,d$, and we write

$$ds^2=\pm a_k^2(t)d\sigma_k^2+d\Omega^2,$$

where $d\Omega^2$ is simply a description of the rest. Let $d\hat{s}^2=\pm ds^2/a_k^2(t)$, then we may write

$$d\hat{s}^2=h_{\alpha\beta}(x^\alpha)dx^\alpha dx^\beta+h_{AB}(x^A)dx^A dx^B,$$

where $A,B=0,d+1,\ldots,D-1$ (h_{AB} has unspecified signature). $d\sigma_k^2$ has curvature scalar $R_{(1)}$ and $d\hat{\Omega}^2=h_{AB}dx^A dx^B$ curvature scalar $R_{(2)}$. Therefore, the metric $d\hat{s}^2$ has curvature scalar $\hat{R}=R_{(1)}+R_{(2)}$. Then we apply (4.20) and the fact that $C_{\alpha A\beta B}$ vanishes. This gives

$$R_{\alpha\beta}h_{AB}+R_{AB}h_{\alpha\beta}=\frac{\hat{R}}{D-1}h_{\alpha\beta}h_{AB}. \tag{4.21}$$

We calculate the trace of this equation by multiplying with $h^{\alpha\beta}h^{AB}$. Observing that $d\geqslant 2$ and $m\geqslant 2$ this leads to

$$\frac{R_{(1)}}{d(d-1)}+\frac{R_{(2)}}{m(m-1)}=0. \tag{4.22}$$

$R_{(1)}$ depends on x^α only, $R_{(2)}$ on x^A only, consequently both of them are constant. By the way, for $m=d=D/2$ we get additionally $\hat{R}=0$. Knowing the constancy of $R_{(1)}$, we multiply Eq. (4.21) by h^{AB} and get the result that $d\sigma_k^2$ is an Einstein space with constant curvature scalar. In the last step we apply $C_{\alpha\beta\gamma\delta}=0$ with Eq. (4.20) to show that $d\sigma_k^2$ is a space of constant curvature.

C. Dimension $D \geqslant 4$: Second part

According to the previous results, it remains to discuss the cases with

$$n \geqslant 2, \quad ds^2 = dt^2 \pm \sum_{k=1}^{n} a_k^2(t) d\sigma_k^2, \tag{4.23}$$

where each of the spaces $d\sigma_k^2$ represents a space of constant curvature, and the quotients $a_k(t)/a_j(t)$ are constant for $k=j$ only. Let us start with the simplest case: every $d_k=1$, i.e., $n=D-1$, and we have to check the positive signature case only

$$n \geqslant 3, \quad ds^2 = dt^2 + \sum_{k=1}^{n} a_k^2(t)(dx^k)^2. \tag{4.24}$$

We define $\alpha_k = \dot{a}_k/a_k$, $H = \Sigma \alpha_k$, $H_2 = \Sigma \alpha_k^2$, and $\beta_k = \alpha_k - H/n$. Consequently, $\Sigma \beta_k = 0$. The equation $R_{AB} = \lambda g_{AB}$, λ = const, $A,B = 0,\ldots,n$ yields

$$\lambda + \dot{H} + H_2 = 0, \tag{4.25}$$

and

$$\lambda + \dot{\alpha}_k + \alpha_k H = 0. \tag{4.26}$$

Summing up the n equations (4.26), we get

$$n\lambda + \dot{H} + H^2 = 0. \tag{4.27}$$

It holds: $a_k(t)/a_j(t)$ is constant iff $\beta_k = \beta_j$.

Let $B = \Sigma \beta_k^2$, then it holds: the metric (4.23) has $n \geqslant 2$ in the normal form iff $B \neq 0$.

Proof: $B=0$ iff $\beta_k = 0$ for every k. But equality of all the β_k implies $\beta_k = 0$ because of $\Sigma \beta_k = 0$. q.e.d.

Inserting $H_2 = B + H^2/n$ into Eq. (4.25) we get

$$n\lambda + n\dot{H} + nB + H^2 = 0. \tag{4.28}$$

Equation (4.26) can be written as

$$n\lambda + n\dot{\beta}_k + \dot{H} + n\beta_k H + H^2 = 0. \tag{4.29}$$

(Of course, summing up the n equations (4.29) we recover (4.27).) The difference between (4.27) and (4.28) leads to

$$\dot{H} = -\frac{n\ B}{n-1}. \tag{4.30}$$

Because of $B>0$ we have $\dot{H}<0$.

The difference between (4.29) and (4.27) reads

$$\dot{\beta}_k + \beta_k H = 0. \tag{4.31}$$

This implies $\dot{B} = -2HB$ and with (4.30)

$$\left(\frac{\dot{B}}{B}\right) = \frac{2n}{n-1} B. \tag{4.32}$$

For $\lambda = 0$, we get from (4.27) $H = 1/t$, $(t>0)$, i.e., with (4.30),

$$B=\frac{n-1}{nt^2}. \tag{4.33}$$

With $H=1/t$, we can solve Eq. (4.31), obtaining $\beta_k=p_k/t$, $p_k=$const. Together with Eq. (4.33), we finally get

$$\sum p_k=0, \quad \sum p_k^2=\frac{n-1}{n}.$$

Consequently, $\alpha_k=\beta_k+1/(nt)=q_k/t$, with

$$\sum q_k=\sum q_k^2=1. \tag{4.34}$$

One can integrate these equations to $a_k(t)=t^{q_k}$ and Eqs. (4.34) give the defining condition for the d-dimensional Kasner solution.

Now we have to check under which conditions the Kasner solution

$$ds^2=dt^2+\sum_{k=1}^{n} t^{2q_k}(dx^k)^2 \tag{4.35}$$

is flat. (By the way, $n>1$ and (4.34) already imply $B>0$.) However, this requires the calculation of the Weyl tensor.

D. Calculating the Weyl tensor

Now we take the metric (4.24), assume that it is already an Einstein space, and calculate the Weyl tensor with the notations used in Sec. IV C. In this section the sum conventions will not be used. It turns out that the non-trivial components of the Weyl tensor read

$$0=C^{0k}_{0k}=-\dot{\alpha}_k-\alpha_k^2+c_1 \tag{4.36}$$

and

$$0=C^{jk}_{jk}=-\alpha_j\alpha_k+c_2, \quad j\neq k, \tag{4.37}$$

where c_1 and c_2 are constants which vanish iff ds^2 has vanishing curvature scalar. Now we are prepared to continue our discussion.

E. Dimension $D\geqslant 4$: Third part

Let $R=0$, i.e., $c_2=0$ in Eq. (4.37). This means that at most one of the functions α_j may be different from zero. If all of them are zero, we get Eq. (4.1), if one of them is non-zero, then we get (4.2), so no new representations are found here.

Let $R\neq 0$, i.e., $c_2\neq 0$ in Eq. (4.37). Let j,k,l be three different indices (which is possible because $n\geqslant 3$). Then (4.37) implies $c_2=\alpha_l\alpha_j=\alpha_l\alpha_k\neq 0$, hence $\alpha_l\neq 0$, and consequently $\alpha_j=\alpha_k$. Because of the arbitrariness in the choice of j and k, we conclude that all the functions α_k coincide. So we have the case $n=1$ in the normal form which has already been discussed above. This concludes the discussion for the case that all the spaces $d\sigma_k^2$ are flat. By condition (D) (see Sec. II), we may now assume that $d\sigma_1^2$ has non-vanishing curvature and consequently $d_1\geqslant 2$. So we have to check under which circumstances (again, the index 1 will be omitted)

$$ds^2=\pm a^2(t)d\sigma^2+dt^2\pm\sum_{k=2}^{n} a_k^2(t)d\sigma_k^2 \tag{4.38}$$

represents a space of constant curvature. If $a(t)$ is constant, this is impossible. Hence, $\alpha=\dot{a}/a\neq 0$. So, we have again the same situation as in Sec. IV B and we may apply Eq. (4.22) as follows: let

$$d\overline{s^2}=\frac{1}{a^2(t)}\left[dt^2\pm\sum_{k=2}^{n}a_k^2(t)d\sigma_k^2\right], \tag{4.39}$$

then $d\overline{s^2}$ must be a space of non-vanishing constant curvature.

Let $D=4$. Then necessarily $n=2$, and $d\sigma_k^2$ is one-dimensional, hence a flat space and $d\sigma^2$ has to be a plane of constant curvature. The results are easily calculated and are given by Eqs. (4.12)–(4.19), where we have to insert $k=2$ and $D=4$ in Eqs. (4.12) and (4.13).

Higher dimensions can be discussed by induction over D: Eq. (4.39) has the same structure (up to a redefinition of t), and therefore we can deduce all possible metrics, from which we can calculate the corresponding metrics (4.38) of dimension at least higher by two.

However, this is quite involved, so we prefer a different approach. A more careful inspection of Eq. (4.38) leads to the observation that Eq. (4.37) remains valid. Therefore, the discussion at the beginning of Sec. IV E still holds. Thus we can restrict to the case $n=2$ in the following. So we have to check: $R=0$,

$$ds^2=dt^2\pm(a^2(t)d\sigma^2+\delta_{\alpha\beta}dx^\alpha dx^\beta) \tag{4.40}$$

with $\dot{a}=0$ and $d\sigma^2$ not flat.

$R\neq 0$,

$$ds^2=dt^2\pm(a^2(t)d\sigma^2+b^2(t)d\sigma_2^2) \tag{4.41}$$

with $\dot{a}\neq 0$, $\dot{b}\neq 0$, and $d\sigma^2$ not flat.

For the case $R=0$, we get $a(t)=t$ and the metrics (4.12), (4.13). For the case $R\neq 0$, we get the generalization of (4.14)–(4.19) to higher dimensions, which read

$$n=2,\quad ds^2=dt^2+\sin^2 t d\Omega_j^2+\cos^2 t d\Omega_k^2 \tag{4.42}$$

and

$$n=2,\quad ds^2=dt^2-\sin^2 t d\overline{\Omega_j^2}-\cos^2 t d\overline{\Omega_k^2} \tag{4.43}$$

for $R>0$ and

$$n=2,\quad ds^2=dt^2+\sinh^2 t d\Omega_j^2+\cosh^2 t d\overline{\Omega_k^2} \tag{4.44}$$

and

$$n=2,\quad ds^2=dt^2-\cosh^2 t d\Omega_j^2-\sinh^2 t d\overline{\Omega_k^2} \tag{4.45}$$

for $R<0$, with $j+k=D-1$. In Eq. (4.42), only $2\leqslant j\leqslant k$ give new cases, these are $(D-4)/2$ for even D and $(D-3)/2$ for odd D. The same result appears for Eq (4.43). In Eq. (4.44), all $j,\ k\geqslant 2$ give new cases, $D-4$ in sum, as does Eq. (4.45).

V. A MORE GENERAL ANSATZ

In this section, we consider a slightly different ansatz for the metric of a space of constant curvature, which is suitable for a direct determination of its general form from the solution of a simple system of differential equations. It is well-known, in fact, that a maximally symmetric space is uniquely characterized by the property that

$$R_{ijkl}=\lambda(g_{ik}g_{jl}-g_{il}g_{jk}) \tag{5.1}$$

with a constant λ. Adopting an appropriate ansatz for the metric, (5.1) gives a set of differential equations which determine the conditions under which the space is of constant curvature.

We take a metric of the form

$$ds^2=dt^2\pm\left[a^2(t)d\rho^2+\sum_{i=1}^{D-2}b_i^2(\rho,t)(dx^i)^2\right], \tag{5.2}$$

where the metric functions depend on t and one ''spatial'' coordinate ρ. A more general form of the ansatz would require that $a=a(t,\rho)$, but our choice is necessary in order to obtain a tractable set of differential equations from (5.1). Anyway, our choice is not restrictive since the metric (5.2) can be seen as a multi-warped product of a two-dimensional space, spanned by the coordinates t and ρ, with a $(D-2)$-dimensional space, and is well known that the metric of a two-dimensional space of constant curvature can always be put in the form $dt^2\pm a^2(t)d\rho^2$ by a suitable choice of coordinates.

For simplicity, in this section we consider the case of Lorentzian signature. The Euclidean case is easily obtained following the same procedure. Inserting the ansatz (5.2) into (5.1), one obtains the following equations

$$\ddot{a}=\lambda a, \tag{5.3}$$

$$\ddot{b}_i=\lambda b_i, \tag{5.4}$$

$$\dot{a}\dot{b}_i-\frac{b_i''}{a}=\lambda a b_i, \tag{5.5}$$

$$\dot{b}_i\dot{b}_j-\frac{b_i'b_j'}{a^2}=\lambda b_i b_j, \tag{5.6}$$

$$a\dot{b}_i'-\dot{a}b_i'=0, \tag{5.7}$$

where a dot denotes a derivative with respect to t and a prime a derivative with respect to ρ. Equation (5.7) implies that

$$\frac{d}{dt}\left(\frac{b_i'}{a}\right)=0. \tag{5.8}$$

Hence, either $b_i'=0$ or $b_i=a(t)g_i(\rho)$. We can now pass to discuss the solutions of (5.3)–(5.7). In the following, we tacitly assume that the transformations of Section II can be applied to change the form of the solutions to their normal form.

A. $\lambda>0$

If $\lambda>0$, without loss of generality we can take $\lambda=1$ by condition (B). Equation (5.3) then implies $a=Ae^t+Be^{-t}$ with A and B integration constants. Moreover, if $b_i'=0$, Eq. (5.4) implies that $b_i=C_ie^t+D_ie^{-t}$, with C_i and D_i integration constants. Equations (5.5) and (5.6) then read

$$AD_i=-BC_i, \quad C_iD_j=-C_jD_i. \tag{5.9}$$

Only two distinct classes of solutions are available: if either $A=C_i=0$ or $B=D_i=0$, one has for any dimension D a solution of the general form (4.7). Further non-trivial solutions are available only in $D=3$, when $D_1=-BC_1/A$, or in $D=2$ and, modulo the transformations of Sec. II they reduce to the form (3.14) in $D=3$ or (3.6), (3.7) in $D=2$.

If $b_i'\neq0$, instead, one has

$$b_i=A(t)g_i(\rho)=(Ae^t+Be^{-t})g_i(\rho),$$

which substituted in (5.5), (5.6) gives

$$g_i''=-4ABg_i, \quad g_i'g_j'=4ABg_ig_j. \tag{5.10}$$

One must distinguish the case $AB>0$ from $AB<0$. If $AB>0$, $a(t)$ can be reduced to $\cosh t$, while, after a rescaling of ρ, $g_i=F_i\sin\rho+G_i\cos\rho$ and (5.10) implies $F_iF_j+G_iG_j=0$. A solu-

tion to the last equation is available only for $D\leqslant 4$, and can be put in the form (here and in the following, lower dimensional solutions can be obtained by suppressing some of the coordinates $x,y,\ldots$)

$$ds^2=dt^2-\cosh^2 t(d\rho^2+\sin^2 \rho dx^2+\cos^2 \rho dy^2), \tag{5.11}$$

which is of the type (4.11).

If $AB<0$, instead, $a(t)$ can be reduced to $\sinh t$, while $g_i=F_ie^{\rho}+G_ie^{-\rho}$ and (5.10) implies $F_iG_j+G_iF_j=0$. A solution to this equation valid in any dimensions, is given by $F_i=0$ or $G_i=0$ and can be written

$$ds^2=dt^2-\sinh^2 t\left(d\rho^2+e^{2\rho}\sum dx_k^2\right). \tag{5.12}$$

Another solution is available when $D\leqslant 4$, if $F_2G_1=-F_1G_2$, which can be reduced to the form

$$ds^2=dt^2-\sinh^2 t(d\rho^2+\sinh^2 \rho dx^2+\cosh^2 \rho dy^2). \tag{5.13}$$

Both solutions (5.11) and (5.12) are special cases of (4.9). A last possibility remains to be investigated, namely the case in which some of the b_i are independent of ρ and some are not. It is easy to see that in this case, Eq. (5.6) is a consequence of (5.5) and hence one gets two new five-dimensional solutions, and one in arbitrary dimensions, by combining (5.11)–(5.13) with (3.14)

$$ds^2=dt^2-\sinh^2 t\left(d\rho^2+e^{2\rho}\sum dx_k^2\right)-\cosh^2 tdy^2, \tag{5.14}$$

$$ds^2=dt^2-\sinh^2 t(d\rho^2+\sinh^2\rho dx^2+\cosh^2 \rho dy^2)-\cosh^2 tdz^2, \tag{5.15}$$

$$ds^2=dt^2-\cosh^2 t(d\rho^2+\sin^2 \rho dx^2+\cos^2 \rho dy^2)-\sinh^2 tdz^2. \tag{5.16}$$

The solutions (5.14)–(5.16) are all special cases of (4.45).

B. $\lambda<0$

If $\lambda<0$, (5.3) yields $a=A\sin t+B\cos t$, with A and B integration constants. If $b_i'=0$, Eq. (5.4) is solved by $b_i=C_i\sin t+D_i\cos t$ and Eqs. (5.5) and (5.6) imply

$$AC_i=-BD_i, \quad C_iC_j=-D_iD_j. \tag{5.17}$$

These relations admit solutions only for $D=3$, for example $B=C_1=0$ or $A=D_1=0$. All these solutions are equivalent, via the transformations of Section II, to (3.11).

If $b_i'\neq 0$, instead, one has

$$b_i=A(t)g_i(\rho)=(A\sin t+B\cos t)g_i(\rho)$$

which substituted in (5.5), (5.6) give respectively,

$$g_i''=-(A^2+B^2)g_i, \quad g_i'g_j'=-(A^2+B^2)g_ig_j. \tag{5.18}$$

These equations have solution $g_i=F_ie^{\rho}+G_ie^{-\rho}$, with $F_iG_j+G_iF_j=0$. Solutions to this equation exist up to four dimensions and take the form

$$ds^2=dt^2-\sin^2 t(d\rho^2+\sinh^2 \rho dx^2+\cosh^2 \rho dy^2) \tag{5.19}$$

and hence are of the type (4.6).

Finally, in the case in which some of the b_i are independent of ρ and some are not, it is easy to see that Eq. (5.10) is always satisfied and hence one gets one new five-dimensional solution, which again is a special case of (4.43)

$$ds^2=dt^2-\sin^2 t(d\rho^2+\sinh^2 \rho dx^2+\cosh^2 \rho dy^2)+\cos^2 tdz^2. \tag{5.20}$$

C. $\lambda = 0$

If $\lambda = 0$, a and b must be linear in t. Proceeding as before, one obtains the following possibilities: if $a = t$, b_i is constant for any i or can have the form $b_1 = t \sinh \rho$, $b_2 = t \cosh \rho$. One then recovers the solutions (4.2), or

$$ds^2 = dt^2 - t^2(d\rho^2 + \sinh \rho^2 dx^2 + \cosh^2 \rho dy^2), \tag{5.21}$$

which is equivalent to (3.19). If a is constant, one can have $b_1 = \rho$, $b_i = \text{const}$, $i \neq 1$ and hence (4.13).

To conclude this section, we observe that no new solutions are introduced by the ansatz (5.2). However, we did not formally define which is the generalization of condition (F), and so in this section, the classification is not such a strict one as in Sec. IV. This fact may lead to the conjecture that even assuming a more general dependence of the metric on the ''spatial'' coordinates, no new forms can be obtained for the solutions, in addition to those listed in the previous sections.

VI. DISCUSSION

Now we have finished the classification of the spaces of constant curvature, and it is useful to sum up potential applications of this. First, consider any type of cosmic no-hair theorem stating that under certain circumstances every solution of the Einstein field equation is asymptotically a space of constant curvature. To find out regions where this theorem applies it is necessary to know which spacetimes of a given class, e.g., the multidimensional warped products we considered here, is exactly a (generalized) de Sitter spacetime. We found unexpected forms on the one hand, and on the other hand, it came out that the Cartesian product between a four-dimensional de Sitter spacetime and one or more static internal spaces is never a space of constant curvature.

Example: metric (4.19) is a metric for a cosmological model of Bianchi type III. Usually, Bianchi type III consists of anisotropic models only, and so (4.19) may be used to test the range of validity of the no-hair theorem within the class of Bianchi type III models, which seems not to have been done up to now.

This list may help identifying ''new solutions'' of Einstein's field equations with or without cosmological term in an arbitrary number of dimensions. The most remarkable result of our paper seems to be that no metric with more than two factors is possible. If n is the number of factors, we have found for every $D \geqslant 2$ twelve representations with $n = 1$ and $5D - 10$ representations with $n = 2$ if D is even and $5D - 9$ representations with $n = 2$ if D is odd.

ACKNOWLEDGMENTS

H.J.S. thanks the University of Cagliari for hospitality and the DFG for financial support.

[1] U. Kasper, M. Rainer, and A. Zhuk, Gen. Relativ. Gravit. **29**, 1123 (1997).
[2] V. D. Ivashchuk and V. N. Melnikov, Grav. Cosmol. **2**, 211 (1996).
[3] U. Bleyer, D.-E. Liebscher, H.-J. Schmidt and A. Zhuk, Wiss. Z. Univ. Jena **39**, 20 (1991).
[4] H.-J. Schmidt, Fortschr. Phys. **41**, 179 (1993).
[5] J. Bičak and J. Griffiths, Ann. Phys. (N.Y.) **252**, 180 (1996).
[6] M. Cadoni and S. Mignemi, Mod. Phys. Lett. A **10**, 367 (1995).

Fixed points due to gravitational lenses

A. O. Petters
Department of Mathematics, Princeton University, Princeton, New Jersey, 08544

F. J. Wicklin
The Geometry Center and School of Mathematics, University of Minnesota, Minneapolis, Minnesota, 55454

(Received 7 May 1997; accepted for publication 10 October 1997)

A fixed point of a gravitational lensing map represents those positions from which a pointlike light source has a lensed image that, despite gravitational lensing, corresponds to the original position of the source. In this paper we study fixed points of lensing maps due to a generic gravitational lens with applications to nonsingular isolated lenses, and to point-mass lenses with continuous matter and shear. Counting formulas and bounds on the number of fixed points are determined. The results include an odd-number fixed-point theorem for nonsingular isolated deflectors. Information on the positions of fixed points are found for the case of point masses on a lens plane with and without shear. The methods of the paper are based on Morse theory, complex variables, and resultants. © *1998 American Institute of Physics.* [S0022-2488(98)02002-7]

I. INTRODUCTION

Consider a light source at y in R^2 undergoing gravitational lensing by a matter distribution with Newtonian potential ψ. Associated with ψ is a planar lensing map defined by $\boldsymbol{\eta}(x)=x-\text{grad}\,\psi(x)$. The lensed images of a light source at y are solutions x of the lens equation: $\boldsymbol{\eta}(x)=y$. Due to light deflection, a lensed image of a light source does not represent the actual position or magnification of the source. We are interested in studying those positions from which a light source has a lensed image that, in spite of a deflector, coincides with the original position of the light source. They are ''fixed points'' due to the lens. Formally, a *fixed point* of $\boldsymbol{\eta}$ is a point x_0 such that $\boldsymbol{\eta}(x_0)=x_0$. Note that by construction, fixed points of $\boldsymbol{\eta}$ are critical points of ψ. To the best of our knowledge, these points have not been studied before in gravitational lensing. In addition, given the fundamental role of this notion in mathematics, fixed-point lensed images warrant investigation, at least on mathematical grounds. Fixed points are vital to the field of discrete dynamical systems, and the planar mappings arising in gravitational lensing may be analyzed using geometric techniques from this field.

We will show that fixed points not only exist, but obtain counting formulas and lower bounds, as well as some results on their positions. Our methods employ the Morse-theoretic and Resultant techniques in the recent papers of Petters[1,2] and Petters and Witt.[3] Some of these results were announced at the Eighth Marcel Grossman Meeting on General Relativity.

The outline of our paper is as follows. In Sec. II we present counting formulas and lower bounds on the number of fixed points due to a ''generic'' gravitational lens. The remainder of the paper then deals with applications to nonsingular isolated lenses (Sec. III), point masses (Sec. IV), and point masses with continuous matter and shear (Sec. V). Section V also includes upper bounds on the number of fixed points.

II. FIXED POINTS DUE TO A GENERAL LENS

Some basic terminologies. Let $L=R^2-A$, where A is a finite collection of points in R^2, and denote the Laplacian on R^2 by Δ. A *gravitational lens potential with infinite singularities A* is a smooth map $\psi: L\to R$ such that $\Delta\psi(x)\geqslant 0$ for all $x\in L$, and $\psi(x)\to-\infty$ as $x\to a$ for each $a\in A$. We call ψ *isolated* if $\psi(x)\to+\infty$ as $|x|\to\infty$. Also, ψ is *nonsingular* when $A=\varnothing$; otherwise,

 © 1998 American Institute of Physics

ψ is *singular*. The *(single-plane) lensing map* induced by ψ is the map $\boldsymbol{\eta}: L \to S$ defined by $\boldsymbol{\eta}(x) = x - \boldsymbol{\alpha}(x)$, where $S = R^2$ and $\alpha = \text{grad}\, \psi$. We call L and S, resp., the *lens* and *light source plane* of $\boldsymbol{\eta}$, while $\boldsymbol{\alpha}$ is the *bending-angle vector field* of ψ.

A lensed image of a light source at $y \in S$ is a solution x of the lens equation $y = \boldsymbol{\eta}(x)$. If x_0 is a fixed point of $\boldsymbol{\eta}$, then x_0 is a lensed image of a light source at $y = x_0$. We can also assign a critical point type to lensed images. In fact, the lensed images of a light source at y are critical points of the *time-delay function* $T_y: L \to R$ induced at y by ψ:

$$T_y(x) = \frac{|x-y|^2}{2} - \psi(x).$$

We have grad $T_y(x) = 0$ if and only if $\boldsymbol{\eta}(x) = y$. The *magnification* of a lensed image x_0 of y is defined to be $1/|\det[\text{Jac}\ \boldsymbol{\eta}](x_0)| = 1/|\det[\text{Hess}\ T_y](x_0)|$. If the magnification of x_0 is 1 and a tiny disc is centered about y, then the resulting extended lensed image about x_0 has approximately the same area as the disc.

Morse Boundary Conditions B. It is practically unfeasible to study the fixed points of lensing maps due to arbitrary gravitational lens potentials ψ. However, if ψ satisfies certain ''generic'' properties, then we can obtain a lower bound and counting formula for the number of fixed points of the induced lensing map $\boldsymbol{\eta}$. First, we shall assume that ψ is nondegenerate. In fact, for almost all a in R^2 with $|a|$ sufficiently small, the linear perturbation,

$$\psi^a(x) = \psi(x) - a \cdot x,$$

of ψ is nondegenerate gravitational lens potential on L.[4] As $|a| \to 0$, the potential ψ^a becomes an arbitrarily small linear perturbation of ψ on any bounded open subset of L. Second, let D be a *regular domain* in L, i.e., a compact subset of L whose boundary ∂D is a disjoint union of smooth simple closed curves. Since the critical points of a nondegenerate ψ are isolated, we can choose D such that ∂D contains no critical point of ψ. *Denote by* D_- *the union of all open subsets of* ∂D *where the bending-angle vector field of* ψ *points into the interior of* D. If ψ is nondegenerate, then for almost all points $a \in R^2$, the critical points in D of a linear perturbation ψ^a tend uniformly and bijectively with preservation of critical point type to those of ψ in D as $|a| \to 0$. Moreover, for $|a|$ sufficiently small, the restricted map $\psi^a|_{\partial D}$ is nondegenerate. Proofs of these results appear in Ref. 5. We shall then define ψ to obey *Morse boundary conditions B* relative to a regular domain D in L if all critical points of ψ lying in D are nondegenerate, no critical point of ψ lies on ∂D, and the restriction $\psi|_{\partial D}$ is nondegenerate. This property is ''generic'' in the sense that it is satisfied by an arbitrarily small linear perturbation of ψ in any bounded open subset containing D.

Notation:

(1) $F(\boldsymbol{\eta}; D)$ = set of fixed points of $\boldsymbol{\eta}$ in D. $|F(\boldsymbol{\eta}; D)|$ = total number of fixed points of $\boldsymbol{\eta}$ in D.
(2) $M_{\min}(\psi; D)$ = total number of minima critical points of ψ in D. $M_{\text{sad}}(\psi; D)$ = total number of saddle critical points of ψ in D. Note that ψ has no maximum critical point [since $\Delta\psi(x) \geqslant 0$ for all $x \in L$].
(3) $M'_{\min}(\psi|_{D_-})$ = total number of minima critical points of $\psi|_{D_-}$. $M'_{\max}(\psi|_{D_-})$ = total number of maxima critical points of $\psi|_{D_-}$.

Since a regular domain represents a portion of the sky where observations are made, it suffices to consider regular domains that are *closed g-hole discs* (i.e., a manifold that is diffeomorphic to a closed disc from which g disjoint open discs are excised). Think of the holes as places where a singularity of the gravitational lens is located.

Theorem 1: *Let* $\boldsymbol{\eta}: L \to S$ *be a lensing map induced by a gravitational lens potential* ψ *with infinite singularities A. Suppose that* ψ *obeys Morse boundary conditions B relative to a closed g-hole disc D, where each hole of D has exactly one element of A. (Note that the number of singularities of* ψ *may exceed g.) Then the fixed points of* $\boldsymbol{\eta}$ *in D satisfy*

$$\begin{aligned} |F(\boldsymbol{\eta}; D)| &= 2M_{\min}(\psi; D) + g - 1 + [M'_{\min}(\psi|_{D_-}) - M'_{\max}(\psi|_{D_-})] \\ &= 2M_{\text{sad}}(\psi; D) - g + 1 - [M'_{\min}(\psi|_{D_-}) - M'_{\max}(\psi|_{D_-})]. \end{aligned}$$

Proof: This theorem immediately follows from application of Morse inequalities B. (An introduction to these powerful inequalities, with detailed proofs, appears in a work by Morse and Cairns.[6]) First, since ψ is nondegenerate with no critical point on ∂D, compactness of D implies that ψ has finitely many critical points in D. Consequently, $|F(\boldsymbol{\eta};D)| = M_{\min}(\psi;D) + M_{\text{sad}}(\psi;D)$, which is finite. Second, because ψ also satisfies Morse boundary conditions B relative to D, Morse inequalities B yield

$$M_{\min}(\psi;D) \geqslant B_0(D) - M'_{\min}(\psi|_{D_-}),$$

$$M_{\text{sad}}(\psi;D) - M_{\min}(\psi;D) \geqslant [B_1(D) - M'_{\max}(\psi|_{D_-})] - [B_0(D) - M'_{\min}(\psi|_{D_-})],$$

$$M_{\max}(\psi;D) - M_{\text{sad}}(\psi;D) + M_{\min}(\psi;D) = B_2(D) - [B_1(D) - M'_{\max}(\psi|_{D_-})] + [B_0(D) - M'_{\min}(\psi|_{D_-})],$$

where $B_i(D)$ is the ith Betti number of D. The Betti numbers of a closed g-hole disc are $B_0(D)=1$, $B_1(D)=g$, and $B_2(D)=0$ (e.g., Ref. 7). Since $M_{\max}(\psi;D)=0$ and $|F(\boldsymbol{\eta};D)| = M_{\min}(\psi;D)+M_{\text{sad}}(\psi;D)$, the theorem follows. q.e.d.

In the sequel, we will usually abbreviate $F(\boldsymbol{\eta};D)$ by $F(\boldsymbol{\eta})$, $M_{\min}(\psi;D)$ by $M_{\min}(\psi)$, etc. We will also define $\psi_- = -\psi$ and $\boldsymbol{\alpha}_- = -\boldsymbol{\alpha}$.

III. APPLICATIONS: NONSINGULAR ISOLATED LENSES

Theorem 2: *Let ψ be a nonsingular isolated nondegenerate gravitational lens potential. Suppose that ψ obeys Morse boundary conditions B. Then fixed points of the lensing map $\boldsymbol{\eta}$ induced by ψ exist, and satisfy the following*:

(1) $1 \leqslant |F(\boldsymbol{\eta})| < \infty$.
(2) *Odd-number fixed-point theorem:* $|F(\boldsymbol{\eta})| = 2M_{\min}(\psi) - 1 = 2M_{\text{sad}}(\psi) + 1$.

Proof: The fixed points of $\boldsymbol{\eta}$ coincide with the critical points of ψ_-. Since ψ is isolated, all critical points of ψ_- lie in the interior of some closed disc D. By choosing the radius of D sufficiently large, we can also assume that $\boldsymbol{\alpha}_-$ is inward pointing on ∂D (so $\boldsymbol{\alpha}$ is outward pointing on ∂D). Consequently, $|F(\boldsymbol{\eta})| < \infty$ and D_- is empty. The theorem is now a consequence of Theorem 1. q.e.d.

Without explicitly knowing the gravitational lens potential modeling a particular smooth nonsingular isolated mass distribution, Theorem 2 enables us to conclude that the lens produces an *odd* number of fixed points and, hence, at least one.

Example (elliptic potential): An elliptical potential with ellipticity ϵ and finite core x_c is of the form

$$\psi_{\text{ep}}(x) = C\sqrt{x_{\text{cr}}^2 + (1-\epsilon)u^2 + (1+\epsilon)v^2},$$

where $x=(u,v)$, $C>0$, $x_{\text{cr}}>0$, and $0 \leqslant \epsilon < 1$. The bending-angle vector field is

$$\boldsymbol{\alpha}_{\text{ep}}(x) = C^2((1-\epsilon)u, (1+\epsilon)v)/\psi_{\text{ep}}(x).$$

There is only one fixed point, namely, the origin.

Although Theorem 2 is stated for isolated potentials, the proof of the theorem shows that we can also determine an odd-number fixed-point theorem for any nonsingular potential for which $M'_{\min}(\psi|_{D_-}) - M'_{\max}(\psi|_{D_-})$ is an even number.

Example (perturbed Plummer potential): A perturbed Plummer potential with shear γ at infinity is of the form

$$\psi_{\text{ppl}}(x) = \frac{\kappa_0}{2} \ln(1+|x|^2) - \frac{\gamma}{2}(u^2 - v^2),$$

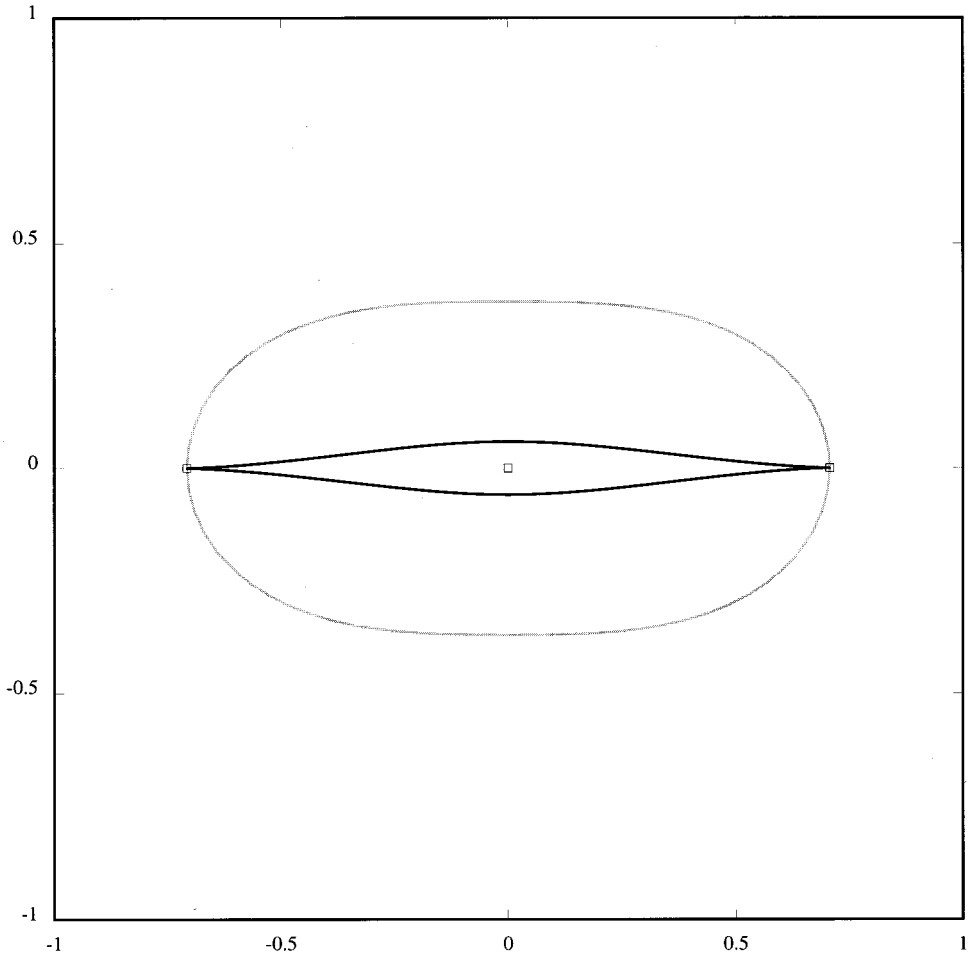

FIG. 1. A superposition of the critical curves (grey) and caustics (black) for the perturbed Plummer model, $\kappa_0=0.75$, $\gamma=0.5$. There are three fixed points (marked by boxes), one at the origin and one at each cusp $(\pm 1/\sqrt{2}, 0)$.

where κ_0 and γ are positive. This potential is not isolated. Employing polar coordinates (R,θ), the potential is approximated for large values of R by $\tilde{\psi}(R,\theta)=-\gamma R^2\cos 2\theta$. On a large disc of radius R, D_- is approximated by the intervals $-\pi/4<\theta<-\pi/4$ and $3\pi/4<\theta<5\pi/4$. Hence $\tilde{\psi}|_{D_-}$ has two critical points, both minima. Hence $M'_{\min}(\psi|_{D_-})-M'_{\max}(\psi|_{D_-})=2-0=2$, and so there will be an odd number of fixed points.

In fact, for this potential we can explicitly compute the fixed points. Since the associated bending-angle vector field is

$$\boldsymbol{\alpha}_{\text{ppl}}(x)=\left(\left(\frac{\kappa_0}{1+|x|^2}-\gamma\right)u,\left(\frac{\kappa_0}{1+|x|^2}+\gamma\right)v\right),$$

there is always a real fixed point at $x=(0,0)$. If $\kappa_0>\gamma$, two more fixed points occur at $x=(\pm\sqrt{\kappa_0/\gamma-1},0)$. Note that for $\kappa_0<\gamma$, the origin is a saddle; for $\kappa_0>\gamma$, the origin is a minimum, and the other two fixed points are saddle points. Figure 1 depicts the locations of these fixed points. Notice the lensing of a disc light source centered at the fixed point in Fig. 2.

In the (typical) case that no explicit formula for the location of the fixed points exists, iterative methods may be used to numerically approximate the fixed points (e.g., Ref. 8) These methods typically use some sort of Newton iteration and require the use of an initial guess (or ''seed'') to start the iteration; these methods are not guaranteed to find all of the fixed points that exist, even if the algorithm is run using thousands of initial seeds. It is imperative, therefore, to have estimates

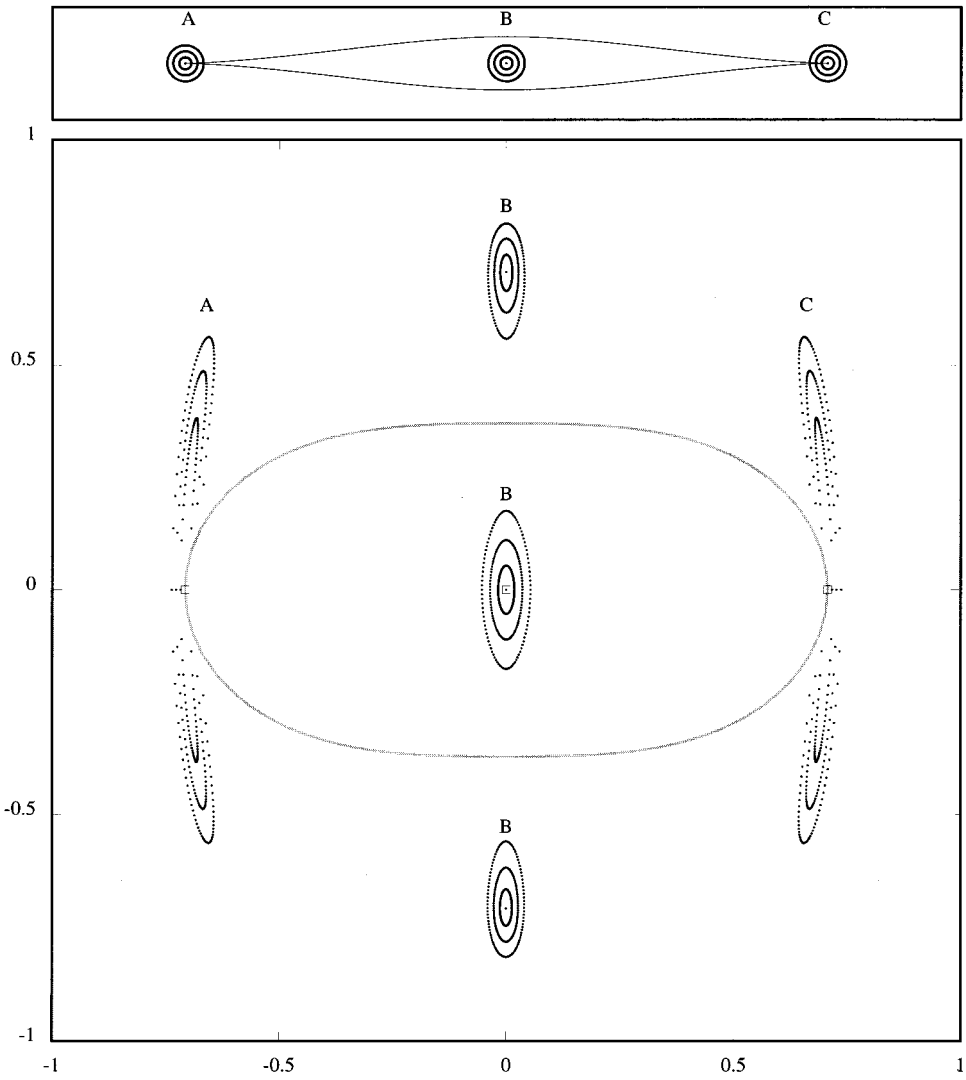

FIG. 2. (Upper) An extended light source is placed at each fixed point, A, B, and C, in the perturbed Plummer model. (Lower) Images of the extended sources. The fixed point at the origin has three images; the fixed point at each cusp is arclike.

and bounds on the number and location of fixed points, so that we may be more confident in the results of our numerics. All of the theorems in this paper provide estimates of this sort, and therefore provide important practical information to computational mathematicians and physicists.

The reader may wonder about the likelihood of a fixed point located on a critical curve. These situations appear in Figs. 1 and 2, but, in general, these are not generic occurrences. If ξ is a fixed point, then grad $\psi(\xi)=0$, but this condition is independent of the condition that ξ is on a critical curve: $\det[d\boldsymbol{\eta}](\xi)=\det[I-\mathrm{Hess}\,(\psi)](\xi)=0$. These two conditions mean that the condition that a fixed point is on a critical curve is a codimension-1 phenomenon. For example, evaluating $\det[I-\mathrm{Hess}\,(\psi)]$ for the perturbed Plummer model at the fixed point gives a determinant of $(1+2\gamma(1-\gamma/\kappa_0))(1-2\gamma)$ and this quantity only vanishes along the curve $\kappa_0=2\gamma^2/(1+2\gamma)$ in the two-dimensional parameter space.

It is even more unlikely that a fixed point on a critical curve is mapped onto a cusp point by $\boldsymbol{\eta}$. The point ξ is a cusp point if the critical curve is nonsingular at ξ, and the tangent vector to the critical curve is contained in the kernel of $d\boldsymbol{\eta}$. In symbols, if T is a tangent vector to the critical curve at ξ, then $d\boldsymbol{\eta}(\xi)\cdot T=0$. (To be certain that ξ is mapped to a ''mathematical cusp,'' and not a higher-order singularity, we should also check that the family of tangent vectors near ξ pass

transversally through the kernel of $d\boldsymbol{\eta}$, but we will not go into these details here.) For example, since the Plummer potential is invariant under the Z_2-symmetry $(u,v)\to(u,-v)$, we know that critical curves passing through the line $v=0$ must have vertical tangent vectors. The Jacobian of the lensing map (evaluated at the fixed point) sends the vector $[0,1]$ to the vector $[0,\ 1-2\gamma]$, and so we see that $d\boldsymbol{\eta}(\xi)$ sends a tangent vector to the zero vector only when $\gamma=\frac{1}{2}$. Of course, for extended light sources, there is an open set of parameter values for which the geometry of Fig. 2 would be observable, so the physical realization of this figure will not be as rare as is indicated by a mathematical analysis of the point-source case.

IV. APPLICATIONS: POINT MASSES WITH CONTINUOUS MATTER AND SHEAR

A. Preliminaries

The negative of the gravitational lens potential due to g point masses m_l at positions $\boldsymbol{\xi}_l$ on a plane lens with constant-density continuous matter κ_c (in units of the critical density) and non-normalized shear $\gamma\geqslant 0$ (at infinity) is given by

$$\psi_-(x)=\frac{(\gamma-\kappa_c)}{2}\,u^2-\frac{(\gamma+\kappa_c)}{2}\,v^2-\sum_{l=1}^{g} m_l \ln|x-\boldsymbol{\xi}_l|, \tag{1}$$

where $x=(u,v)$. The fixed points due to ψ are solutions x of the equation $\boldsymbol{\alpha}_-(x)=0$, where

$$\boldsymbol{\alpha}_-(x)=((-\kappa_c+\gamma)u,(-\kappa_c-\gamma)v)-\sum_{l=1}^{g} m_l\,\frac{x-\boldsymbol{\xi}_l}{|x-\boldsymbol{\xi}_l|^2}. \tag{2}$$

The lensing map $\boldsymbol{\eta}:L\to S$ induced by ψ is given by $\boldsymbol{\eta}(x)=x+\boldsymbol{\alpha}_-(x)$, where $L=R^2-\{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_g\}$ and $S=R^2$. Explicitly,

$$\boldsymbol{\eta}(x)=\begin{pmatrix}1-\kappa_c+\gamma & 0\\ 0 & 1-\kappa_c-\gamma\end{pmatrix}x-\sum_{l=1}^{g} m_l\,\frac{x-\boldsymbol{\xi}_l}{|x-\boldsymbol{\xi}_l|^2}. \tag{3}$$

Now, since $\psi_-(x)\to+\infty$ as $x\to\boldsymbol{\xi}_l$, where $l=1,\ldots,g$, there is an open disc about $\boldsymbol{\xi}_l$ that contains no critical point of ψ_-. In addition, if $\kappa_c=\gamma$, then $\psi_-(x)\to-\infty$ as $|x|\to\infty$. Moreover, if $\kappa_c\neq\gamma$, then $|\boldsymbol{\alpha}_-(x)|\to+\infty$ as $|x|\to\infty$. Consequently, all critical points lie in a bounded region. Let D be a closed g-hole disc, where each hole of D has exactly one singularity of ψ. By choosing the circumradius of the inner-boundary curves and inradius of the outer-boundary curve of D sufficiently small and large, resp., we can assume that all critical points of ψ_- lie in the interior of D. Hence, *if ψ_- is nondegenerate, then it has finitely many critical points, that is, $|F(\boldsymbol{\eta})|=M_{\min}(\psi)+M_{\mathrm{sad}}(\psi)<\infty$, where $\boldsymbol{\eta}$ is the lensing map induced by ψ.* Note that $M_{\max}=0$ since $\Delta\psi=2\kappa_c\geqslant 0$.

Though Theorem 1 yields counting formulas and lowers bounds on $|F(\boldsymbol{\eta})|$, it provides no upper bound. However, by introducing complex variables,[9,10] we shall determine upper bounds on $|F(\boldsymbol{\eta})|$ for point masses on a lens plane with shear and continuous matter.

Express x and $\boldsymbol{\xi}_l$ in complex form: $x=z=u+iv$ and $\boldsymbol{\xi}_l=\xi_l^1+i\xi_l^2$. By Eq. (2), we must investigate the zeros of

$$\boldsymbol{\alpha}_-(z,\overline{z})=-\kappa_c z+\gamma\overline{z}+\sum_{l=1}^{g}\frac{m_l}{\boldsymbol{\xi}_l-\overline{z}}, \tag{4}$$

which is equivalent to finding the simultaneous roots of

$$p(z,\overline{z})=-\kappa_c z\prod_{l=1}^{g}(\overline{\boldsymbol{\xi}_l}-\overline{z})+\gamma\overline{z}\prod_{l=1}^{g}(\overline{\boldsymbol{\xi}_l}-\overline{z})+\sum_{l=1}^{g} m_l\prod_{\substack{r=1\\ r\neq l}}^{g}(\overline{\boldsymbol{\xi}_r}-\overline{z}) \tag{5}$$

and

$$\overline{p(z,\overline{z})}=\gamma z\prod_{l=1}^{g}(\xi_l-z)-\kappa_c\overline{z}\prod_{l=1}^{g}(\xi_l-z)+\sum_{l=1}^{g}m_l\prod_{\substack{r=1\\r\neq l}}^{g}(\xi_r-z). \tag{6}$$

Thus, if x is a fixed point of η*, then* $(z,\overline{z})$ *is a root of both (5) and (6). We can then translate the study of fixed points into that of the common zeros of* $p(z,\overline{z})$ *and* $\overline{p(z,\overline{z})}$*. Applications of the theory of resultants will then yield upper bounds on* $|F(\eta)|$*.*

B. Fixed points due to point masses with $\kappa_c=0$ and $\gamma\geqslant 0$

Throughout this subsection, assume that $\gamma\geqslant\kappa_c=0$.

Notation:

(1) $M_{\text{tot}}=m_1+\cdots+m_g$.
(2) $\hat{M}_l=M_{\text{tot}}-m_l$.

Theorem 3: *Let* ψ *be a nondegenerate gravitational lens potential due to g point masses on a lens plane with shear* $\gamma\geqslant 0$ *and no continuous matter* $(\kappa_c=0)$*. Then the fixed points of the lensing map* η *induced by* ψ *satisfy the following*:

(1) *The set* $F(\eta)$ *consists of all roots of the equation* $\overline{p(z)}=0$*, where*

$$\overline{p(z)}=\gamma z\prod_{l=1}^{g}(\xi_l-z)+\sum_{l=1}^{g}m_l\prod_{\substack{r=1\\r\neq l}}^{g}(\xi_r-z) \tag{7}$$

and has only simple zeros. Consequently,

$$|F(\eta)|=M_{\text{sad}}(\psi)=\begin{cases}g+1, & \text{if } \gamma>0,\\ g-1, & \text{if } \gamma=0.\end{cases} \tag{8}$$

(2) *The fixed points* $y_1,\ldots,y_{n(\gamma)}$ *of* η *obey*

$$y_1+\cdots+y_{n(\gamma)}=\begin{cases}\dfrac{\xi_1+\cdots+\xi_g}{\gamma}, & \text{if } \gamma>0,\\ \dfrac{\hat{M}_1\xi_1+\cdots+\hat{M}_g\xi_g}{M_{\text{tot}}}, & \text{if } \gamma=0,\end{cases} \tag{9}$$

where $n(\gamma)=g+1$ *if* $\gamma>0$*, and* $n(\gamma)=g-1$ *if* $\gamma=0$*. Moreover, no fixed point of* η *has magnification 1.*

Proof: (1) We have $|F(\eta)|=M_{\text{sad}}(\psi)+M_{\min}(\psi)$. But $\Delta\psi=2\kappa_c=0$. Hence, $M_{\min}(\psi)=0$. Since $\kappa_c=0$, Eqs. (5) and (6) are equivalent to the single-variable equation $\overline{p(z)}=0$, where $\overline{p}$ is given by Eq. (7). In addition,

$$\frac{d\overline{\alpha}_-}{dz}(z)=\frac{p'(z)\Pi_{l=1}^{g}(\xi_l-z)-p(z)[\Pi_{l=1}^{g}(\xi_l-z)]'}{\Pi_{l=1}^{g}(\xi_l-z)^2}$$

and $\det[\text{Hess}\,\psi](y)=|(d\overline{\alpha}_-/dz)(y)|^2\neq 0$, where $\overline{\alpha}_-$ is the conjugate of $\overline{\alpha}_-$ and $'$ denotes differentiation relative to z. Hence, all zeros of $p(z)$ are simple. The remainder of part (1) now follows from the degree of $\overline{p(z)}$.

(2) We want to expand the products in Eq. (6). Recall that the kth elementary symmetric function in $t_1,\ldots,t_n$ is

$$s_k=\sum_{1\leqslant j_1<\cdots<j_k\leqslant n}\xi_{j_1}^{l}\cdots\xi_{j_k}^{l},$$

where $1\leqslant k\leqslant n$ (e.g., $s_1=t_1+\cdots+t_n$ and $s_n=t_1\cdots t_n$). Let

$$\xi_j^l=\begin{cases}\xi_j, & j<l\\ \xi_{j+1}, & j\geqslant l,\end{cases}$$

where $1\leqslant j\leqslant g-1$ and $1\leqslant l\leqslant g$. Denote the kth elementary symmetric function in $\xi_1^l,\ldots,\xi_{g-1}^l$ by ξ_k^l (e.g., $\xi_j^1=\xi_{j+1}$ and $\xi_j^g=\xi_j$). Note that $s_1^l=\xi_1^l+\cdots+\xi_{g-1}^l=\xi_1+\cdots+\hat{\xi}_l+\cdots+\xi_g^l$, where $\hat{\xi}_l$ is omitted, and $s_{g-1}^l=\xi_1^l\cdots\xi_{g-1}^l$. Also, $s_1^1=\xi_2+\cdots+\xi_{g-1}$ and $s_{g-1}^g=\xi_1\cdots\xi_{g-1}$.

We expand Eq. (7) as follows:

$$\begin{aligned}\overline{p(z)}&=\gamma z\prod_{l=1}^{g}(\xi_l-z)+(-1)^{g-1}\sum_{l=1}^{g}m_l\prod_{j=1}^{g-1}(z-\xi_j^l)=(-1)^g z(z^g-s_1z^{g-1}+\cdots+(-1)^{g-1}s_g)\\&\quad+(-1)^{g-1}\left\{M_{\mathrm{tot}}z^{g-1}-\sum_{l=1}^{g}m_l(\xi_1+\cdots+\hat{\xi}_l+\cdots+\xi_g)z^{g-2}+\text{lower-order terms}\right\}\\&=(-1)^g(z^{g+1}-s_1z^g+\cdots+(-1)^{g-1}s_gz)\\&\quad+(-1)^{g-1}\left\{M_{\mathrm{tot}}z^{g-1}-\left(\sum_{l=1}^{g}\hat{M}_l\xi_l\right)z^{g-2}+\text{lower-order terms}\right\}.\end{aligned}$$

Since the sum of the zeros of a complex polynomial $a_nz^n+a_{n-1}z^{n-1}+\cdots+a_0$ is $-a_{n-1}/a_n$, the first claim in part (2) of the Theorem follows. For the second claim, observe that the determinant of the Jacobian of $\boldsymbol{\eta}$ is given by

$$\det J(z,\overline{z})=\frac{\partial w}{\partial z}\frac{\overline{\partial w}}{\partial z}-\frac{\partial w}{\partial\overline{z}}\frac{\overline{\partial w}}{\partial\overline{z}}=1-\left|\frac{d\overline{\boldsymbol{\alpha}}_-}{dz}(z)\right|^2.$$

Hence, a fixed-point lensed image z_0 has magnification 1 if and only if $|(d\overline{\boldsymbol{\alpha}}_-/dz)(z_0)|=0$. Nondegeneracy of ψ rules out this possiblity. q.e.d.

Theorem 3 shows that a one point-mass lens has no fixed points, while a point-mass with shear generically produces two fixed points, as shown in Figs. 3 and 4. In the nongeneric case (i.e., the gravitational lens potential is degenerate), the latter has one fixed point of multiplicity two; it is a minimum lensed image with magnification 1. A nondegenerate binary lens has a fixed point at position

$$y=\frac{m_1\xi_2+m_2\xi_1}{m_1+m_2}.$$

If the positions ξ_1 and ξ_2 of the masses m_1 and m_2 are switched, then y is the *center of mass* of the resulting system. For a ternary lens system, the fixed points can also be located theoretically. In fact, note that zeros of Eq. (7) are solutions of

$$\sum_{l=1}^{g}\frac{m_l}{z-\xi_l}=0.$$

The zeros of this equation for $g=3$ was studied by Siebeck as early as 1864. He established (see Ref. 11) that the zeros are the foci of the ellipse which meets the line segment between ξ_j and ξ_k at the point ζ_{jk}, where $|\zeta_{jk}-\xi_j|=(m_j/m_k)|\zeta_{jk}-\xi_k|$ (see Fig. 3). Siebeck also covered the case for general g, where the fixed points become the foci of a certain algebraic curve. Part (2) of Theorem 3 further shows that the sum of the positions of the fixed points (counting with multiplicity) of a g-point mass lens, where $g\geqslant 2$, is the center of mass of the lens system if m_l is replaced by $\hat{M}_l$.

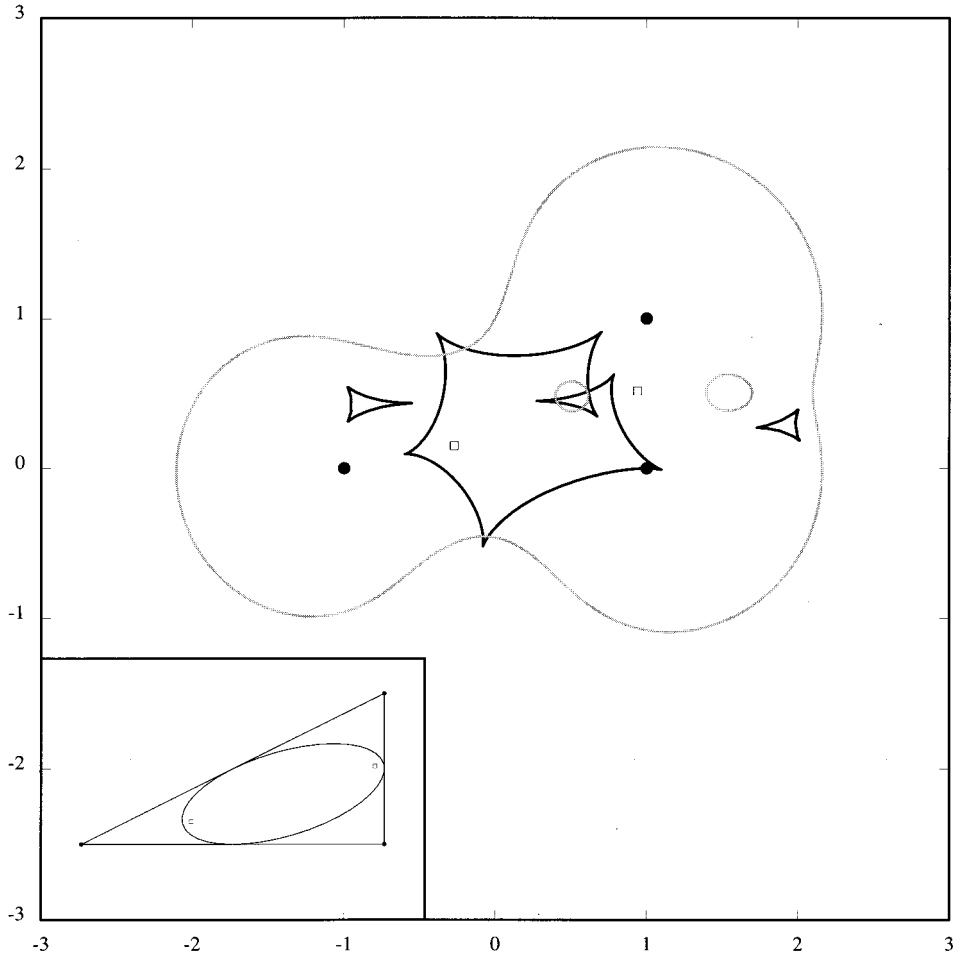

FIG. 3. A superposition of the critical curves (grey) and caustics (black) for three equal masses, $m_i=1$, and $\kappa_c=\gamma=0$. The positions of the masses are indicated by dots (●): $d_1=(-1,0)$, $d_2=(1,0)$, and $d_3=(1,1)$. The position of the fixed points are marked by boxes (□): (0.93912, 0.51675) and (−0.27245, 0.14992). The insert illustrates that the fixed points are foci of an ellipse inscribed in the triangle defined by the three masses and meeting the triangle at the pairwise centers of the sides.

C. Fixed points due to point masses with $0\leqslant\gamma\leqslant\kappa_c$ and $0<\kappa_c<\gamma$

1. Counting formulas: $0\leqslant\gamma\leqslant\kappa_c$ and $0<\kappa_c<\gamma$

For $0\leqslant\gamma<\kappa_c$, we define the ellipse parametrized by

$$\theta\rightarrow\left(\sqrt{\frac{2}{\kappa_c-\gamma}}\,R\,\cos(\theta),\sqrt{\frac{2}{\kappa_c+\gamma}}\,R\,\sin(\theta)\right),$$

whereas we choose

$$\theta\rightarrow\left(\sqrt{\frac{2}{\gamma-\kappa_c}}\,R\,\cos(\theta),\sqrt{\frac{2}{\kappa_c+\gamma}}\,R\,\sin(\theta)\right)$$

when $0\leqslant\kappa_c<\gamma$. Let $D(R)$ be the interior of the ellipse punctured by disjoint open discs $D_{\xi_1},\cdots,D_{\xi_g}$ centered at the point masses. By choosing the radii of the discs D_{ξ_l} sufficiently small, we can assume that at points of the boundary ∂D_{ξ_g}, grad ψ_- points into the exterior of $D(R)$. Moreover, for R sufficiently large we can also assume that all critical points of ψ_- lie in $D(R)$. *Henceforth, we shall assume the previous properties hold for D(R) and all critical points of* ψ_- *lie*

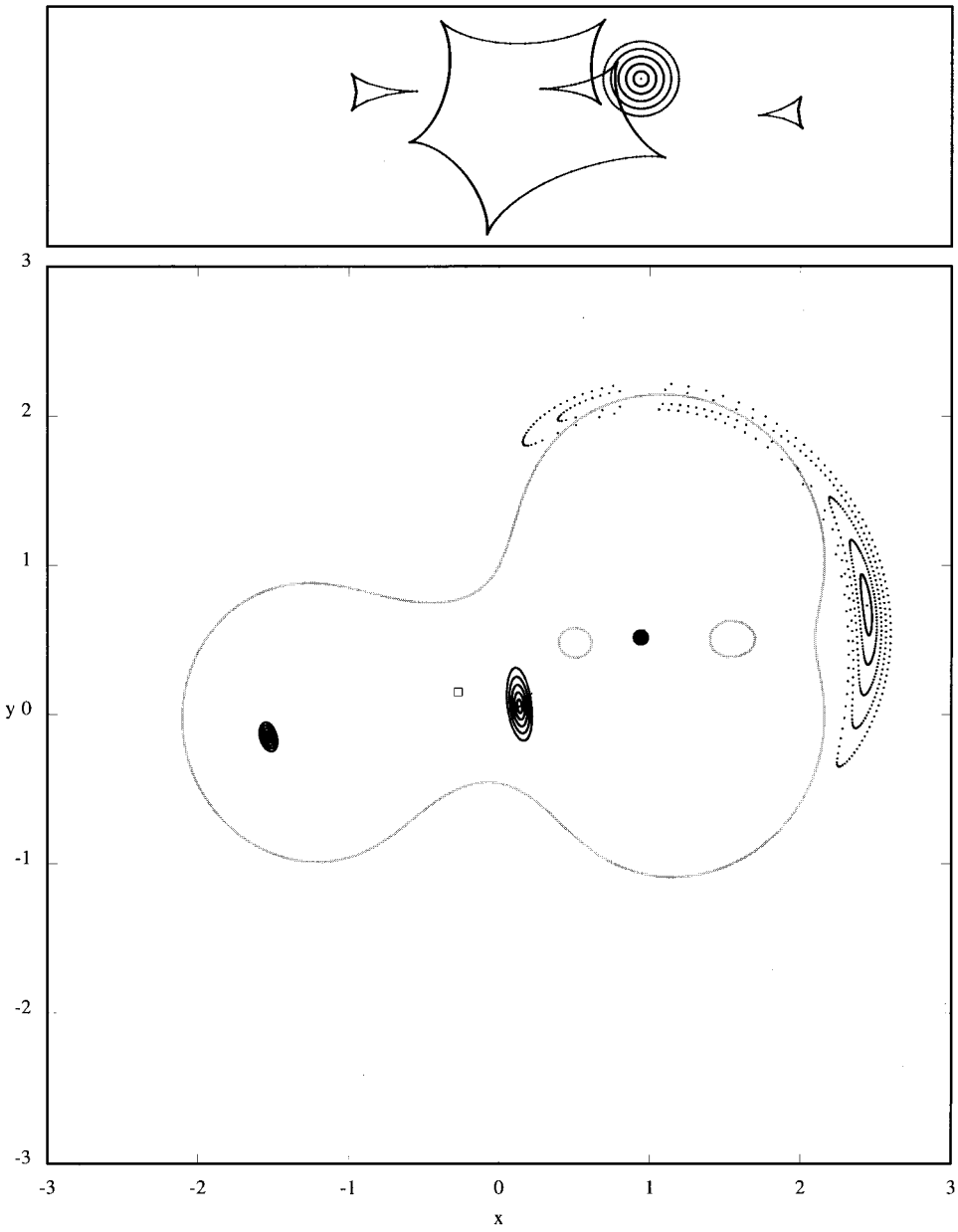

FIG. 4. (Upper) An extended light source, located at one of the fixed points, that crosses a caustic. (Lower) Four images of the source. The arclike image is due to the fact that the source crosses a caustic.

in the interior of D(R). Finally, let $D_-(R)$ be the union of all submanifolds of $\partial D(R)$ at whose points grad ψ_- points into the *exterior* of $D(R)$. In other words, $D_-(R)=D_-$.

Theorem 4: *Suppose that ψ obeys Morse boundary conditions B relative to D(R).*

(1) *If* $0\leqslant\gamma\leqslant\kappa_c$, *then* $|F(\boldsymbol{\eta})|=2M_{\min}(\psi)+g-1=2M_{\text{sad}}(\psi)-g+1$.
(2) *If* $0<\kappa_c<\gamma$, *then* $|F(\boldsymbol{\eta})|=2M_{\min}(\psi)+g+1=2M_{\text{sad}}(\psi)-g-1$.

The theorem shows that $|F(\boldsymbol{\eta})|$ and g have opposite parity (i.e., even or oddness).

Proof: We have $M_{\min}(\psi)=M_{\max}(\psi_-)$, $M_{\text{sad}}(\psi)=M_{\text{sad}}(\psi_-)$, $M'_{\min}(\psi|_{D_-(R)})=M'_{\max}(\psi_-|_{D_-(R)})$, and $M'_{\max}(\psi|_{D_-(R)})=M'_{\min}(\psi_-|_{D_-(R)})$. Theorem 1 implies that

$$|F(\boldsymbol{\eta})| = 2M_{\min}(\psi) + g - 1 + [M'_{\max}(\psi_-|_{D_-(R)}) - M'_{\min}(\psi_-|_{D_-(R)})]$$
$$= 2M_{\text{sad}}(\psi) - g + 1 - [M'_{\max}(\psi_-|_{D_-(R)}) - M'_{\min}(\psi_-|_{D_-(R)})]. \quad (10)$$

(1) For R sufficiently large, the hypothesis that $\gamma \leqslant \kappa_c$ implies that $\boldsymbol{\alpha}_-$ points into the interior of $D(R)$ at all points of the outer boundary of $\partial D(R)$. Recall that if $\gamma = \kappa_c$, then $\psi_-(x) \to -\infty$ as $|x| \to \infty$. In addition, $\boldsymbol{\alpha}_-$ points into the exterior of $D(R)$ at all points of the g inner-boundary curves of $\partial D(R)$. Consequently, $D_-(R)$ is a union of g disjoint circles S^1. Since $\psi_-|_{D_-(R)}$ is nondegenerate, the Morse inequalities for boundaryless manifolds (see Ref. 12) also apply and yield

$$M'_{\max}(\psi_-|_{D_-(R)}) - M'_{\min}(\psi_-|_{D_-(R)}) = B_1(D_-(R)) - B_0(D_-(R))$$
$$= g(B_1(S^1) - B_0(S^1)) = g(1-1) = 0.$$

Application of Eq. (10) yields the desired result.

(2) By restricting ψ_- to ∂D and employing our parametrization of the ellipse, we obtain

$$\psi_-|_{\partial D(R)}(\theta) = R^2 \cos(2\theta) - \sum_{l=1}^{g} m_l \ln\left(R\sqrt{\widetilde{G}_l(\theta)}\right),$$

where

$$\widetilde{G}_l(\theta) = \frac{2}{\gamma - \kappa_c}\cos^2\theta + \frac{2}{\kappa_c + \gamma}\sin^2\theta - \frac{2\xi_l^1}{R}\sqrt{\frac{2}{\gamma - \kappa_c}}\cos\theta - \frac{2\xi_l^2}{R}\sqrt{\frac{2}{\kappa_c + \gamma}}\sin\theta + \left(\frac{|\boldsymbol{\xi}_l|}{R}\right)^2.$$

For large values of R, the function $\psi_-|_{\partial D(R)}(\theta)$ is approximately $R^2\cos(2\theta)$. Hence, $D_-(R)$ is approximated by portions of ∂D for $-\pi/4 < \theta < \pi/4$ and $3\pi/4 < \theta < 5\pi/4$. Thus the map $\psi_-|_{D_-(R)}$ has two critical points, both maxima: $M'_{\max}(\psi_-|_{D_-(R)}) = 2$ and $M'_{\min}(\psi_-|_{D_-(R)}) = 0$. The result now follows from Eq. (10). q.e.d.

2. Bounds: $0 \leqslant \gamma \leqslant \kappa_c$ and $0 < \kappa_c < \gamma$

Theorem 5: *Suppose that ψ obeys Morse boundary conditions B.*

(1) *If $0 \leqslant \gamma \leqslant \kappa_c$, then*

$$g - 1 \leqslant |F(\boldsymbol{\eta})| \leqslant \begin{cases} (g+1)^2, & \text{if } \gamma > 0, \\ g^2 - 1, & \text{if } \gamma = 0. \end{cases}$$

(2) *If $0 < \kappa_c < \gamma$, then*

$$g + 1 \leqslant |F(\boldsymbol{\eta})| \leqslant (g+1)^2.$$

Proof: For the lower bounds in the theorem, note that Theorem 4(1) implies that if $0 \leqslant \gamma \leqslant \kappa_c$, then $|F(\boldsymbol{\eta})| = 2M_{\min}(\psi) + g - 1 \geqslant g - 1$. In addition, by Theorem 4(2), we have $|F(\boldsymbol{\eta})| = 2M_{\min}(\psi) + g + 1 \geqslant g + 1$ for $0 < \kappa_c < \gamma$.

In order to prove the upper bounds in the theorem, we must introduce some facts from the theory of resultants. The x_1-Sylvester resultant of two polynomials

$$P(x_1, x_2) = a_0(x_2) + a_1(x_2)x_1 + \cdots + a_m(x_2)x_1^m$$

and

$$Q(x_1, x_2) = b_0(x_2) + b_1(x_2)x_1 + \cdots + b_n(x_2)x_1^n,$$

is defined by

$$\text{Res}\ (x_2) = \det\ (S_{m+n}),$$

where

$$S_{m+n}=\begin{bmatrix} a_m & a_{m-1} & a_{m-2} & \cdot & \cdot & \cdot & a_0 & 0 & 0 & \cdots & 0 \\ 0 & a_m & a_{m-1} & \cdot & \cdot & \cdot & a_1 & a_0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & a_m & a_{m-1} & a_{m-2} & \cdot & \cdot & \cdot & a_1 & a_0 \\ b_n & b_{n-1} & \cdot & \cdot & \cdot & b_0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & b_n & b_{n-1} & \cdot & \cdot & \cdot & b_0 & \cdot & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdot & \cdot & \cdot & 0 & b_n & b_{n-1} & \cdot & \cdot & \cdot & b_0. \end{bmatrix} \tag{11}$$

Explicitly, the lth row of S_{m+n}, where $1\leqslant l\leqslant n$, consists of $l-1$ zeros, followed by a_m, $a_{m-1},\ldots,a_0$, and the remainder of the row filled with zeros; for $n+1\leqslant l\leqslant m+n$, the lth row has $l-n-1$ zeros followed by b_n, $b_{n-1},\ldots,b_0$, and the remaining entries are zeros.

The Resultant Theorem (see Refs. 13 and 14) states that β satisfies Res $(\beta)=0$ if and only if either $a_m(\beta)=b_n(\beta)=0$ or $P(\alpha,\beta)=Q(\alpha,\beta)=0$ for some α. Another fact we shall need is that the resultant is a polynomial of degree n in the coefficients a_l and of degree m in b_l. That is, if we write Res $(x_2)=$Res $(a_0,\ldots,a_m,b_0,\ldots,b_n)$, then

$$\text{Res }(a_0,\ldots,a_m,b_0,\ldots,b_n)=\sum S_{i_0\ldots i_m j_0\ldots j_n}a_0^{i_0}\ldots a_m^{i_m}b_0^{i_0}\cdots b_n^{j_n}$$

and $i_0+\cdots+i_m=n$ and $j_0+\cdots+j_n=m$. Therefore,

$$\deg[\text{Res }(x_2)]\leqslant(n)(\max\{\deg[a_l]_{l=0}^m\})+(m)(\max\{\deg[b_l]_{l=0}^n\}). \tag{12}$$

The previous results will now be applied to Eqs. (5) and (6) to obtain the upper bounds in the theorem. We begin with $\gamma=0$. We then consider the common roots of

$$p(z,\overline{z})=a_1(\overline{z})z+a_0(\overline{z})$$

and

$$\overline{p(z,\overline{z})}=b_g(\overline{z})z^g+\cdots+b_0(\overline{z}).$$

Denote the z-resulant of p and $\overline{p}$ by Res $(\overline{z})$. If a is a fixed point, then $(a,\overline{a})$ is a common zero of p and $\overline{p}$. By the Resultant Theorem, a is a solution of Res $(\overline{z})=0$. Since $\gamma=0$,

$$\max\{\deg[a_0],\deg[a_1]\}=g$$

and

$$\max\{\deg[b_l]_{l=0}^g\}=1,$$

Eq. (12) implies

$$|F(\boldsymbol{\eta})|\leqslant\deg[\text{Res }(\overline{z})]\leqslant(g)(g)+(1)(1)=g^2+1.$$

Expanding the coefficients $a_1(\overline{z})$ and $b_g(\overline{z})$ we find that

$$a_1(\overline{z})=-\kappa_c\prod_{l=1}^{g}(\overline{\boldsymbol{\xi}_l}-\overline{z})=(-1)^{g+1}\kappa_c(\overline{z^g}-\overline{s_1z^{g-1}}+\cdots+(-1)^{g-2}\overline{s_{g-1}}\overline{z}+(-1)^{g-1}\overline{s_g})$$

and

$$b_g(\overline{z})=(-1)^{g+1}\kappa_c\overline{z}.$$

Without loss of generality, assume that $\xi_1 = 0$. Then $\overline{z} = 0$ is a common zero of $a_1(\overline{z})$ and $b_g(\overline{z})$. By the Resultant Theorem, $\overline{z} = 0$ is a zero of Res $(\overline{z})$. Since the origin is the position of the first point mass, $\overline{z} = 0$ is not a fixed point; hence, $|F(\boldsymbol{\eta})| \leq g^2$. By Theorem 4(1), $|F(\boldsymbol{\eta})|$ is even if and only if g is odd. Thus, we can reduce the upper bound by one: $|F(\boldsymbol{\eta})| \leq g^2 - 1$.

The proof for the upper bound when $\gamma > 0$ and $\kappa_c > 0$ (either $\gamma \leq \kappa_c$ or $\gamma > \kappa_c$ may apply) follows the proof above, except that Eq. (5) now has coefficients (in $\overline{z}$) of maximal degree $g+1$ and Eq. (6) is now a degree $g+1$ polynomial in z. Hence, by Eq. (12),

$$|F(\boldsymbol{\eta})| \leq (g+1)(g+1) + (1)(1) = g^2 + 2g + 2.$$

Again we can reduce this bound by one by noting that $\overline{z} = 0$ is a common root of $a_1(\overline{z})$ and $b_g(\overline{z})$, but the parity argument no longer applies. Thus we may only claim that $|F(\boldsymbol{\eta})| \leq (g+1)^2$. q.e.d.

As a last example, we again consider the ternary point-mass system with continuous matter and shear. Theorem 4 predicts that there will be an even number of fixed points when $g = 3$. Theorem 5 predicts that $2 \leq |F(\boldsymbol{\eta})| \leq 8$ if $0 = \gamma < \kappa_c$. If we perturb the lensing system shown in Fig. 3 to, say, $\gamma = 0$ and $\kappa_c = 0.1$, then we find that the constellation is essentially the same, but that the two fixed points slightly perturb to the locations (0.92698, 0.52446) and (−0.28041, 0.14101). By contrast, when $0 < \kappa_c < \gamma$, Theorem 5 predicts that $4 \leq |F(\boldsymbol{\eta})| \leq 16$. If we perturb the lensing system shown in Fig. 3 to $\kappa_c = 0.1$ and $\gamma = 0.3$, then we find that two new fixed points appear ''from infinity,'' resulting in four fixed points: (0.96570, 0.54376), (−0.246848, 0.13689), (−3.81466, −100.95612), and (4.0949, 56.2398).

ACKNOWLEDGMENTS

The figures in this paper were produced using PISCES,[15] which was supported by The Geometry Center. A.P.'s research was supported in part by NSF Grant No. DMS-9404522.

[1] A. O. Petters, J. Math. Phys. **36**, 4263 (1995).
[2] A. O. Petters, J. Math. Phys. **38**, 1605 (1997).
[3] A. O. Petters and H. J. Witt, J. Math. Phys. **37**, 2920 (1996).
[4] V. Guillemin and A. Pollack, *Differential Topology* (Prentice–Hall, Englewood Cliffs, NJ, 1974). See p. 43.
[5] M. Morse and S. Cairns, *Critical Point Theory in Global Analysis and Differential Topology* (Academic, New York, 1969). See page 47.
[6] See Theorem 10.2 on p. 71 of Ref. 5.
[7] See, for example, the proof of Theorem 11 on page 1924 of A. O. Petters, J. Math. Phys. **33**, 1915 (1992).
[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vatterling, *Numerical Recipes in C* (Cambridge U.P. Cambridge, 1988).
[9] R. R. Bourassa, R. Kantowski, and T. D. Norton, Astrophys. J. **185**, 747 (1973).
[10] H. Witt, Astron. Astrophys. **236**, 311 (1990).
[11] M. Marden, *Geometry of Polynomials* (American Mathematical Society, New York, 1966).
[12] J. Milnor, *Morse Theory* (Princeton U.P., Princeton, 1963). See p. 30.
[13] S. Abhyankar, in *Proceedings of the International Symposium on Algebraic Geometry, Kyoto, 1977*, edited by M. Nagata (Kinokuniya, Tokyo, 1978), p. 391.
[14] S. Abhyankar, *Algebraic Geometry for Scientists and Engineers*, Mathematical Surveys and Monographs Vol. 35 (American Mathematical Society, Providence, 1990).
[15] F. J. Wicklin, The Geometry Center Technical Report #89, 1995.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Stationary generalized Kerr–Schild spacetimes

Carlos F. Sopuerta[a)]
Departament de Física Fonamental, Universitat de Barcelona,
Diagonal 647, 08028 Barcelona, Spain

(Received 11 June 1997; accepted for publication 4 November 1997)

In this paper we have applied the generalized Kerr–Schild transformation finding a new family of stationary perfect-fluid solutions of the Einstein field equations. The procedure used combines some well-known techniques of null and timelike vector fields, from which some properties of the solutions are studied in a coordinate-free way. These spacetimes are algebraically special being their Petrov types II and D. This family includes all the classical vacuum Kerr–Schild spacetimes, excepting the plane-fronted gravitational waves, and some other interesting solutions as, for instance, the Kerr metric in the background of the Einstein Universe. However, the family is much more general and depends on an arbitrary function of one variable.
© *1998 American Institute of Physics.* [S0022-2488(98)02402-5]

I. INTRODUCTION

Since the discovery of the Kerr metric[1] until now, the Kerr–Schild Ansatz[2–4] has been a powerful tool to find new explicit solutions of the Einstein equations. Roughly speaking, this technique consists in generating new solutions from the Minkowski spacetime and its null geodesic vector fields, and it can be applied to several types of energy–momentum content: vacuum, electromagnetic field, pure radiation, etc. With regard to the results obtained, we should note that some of the solutions found have been of crucial importance in general relativity and they have received too much attention, especially in relation to the description of black holes or radiating bodies. Well-known examples are the Kerr–Newman family of spacetimes, the Vaidya metric, etc. A complete review of the classical Kerr–Schild transformation can be found in Reference 5 (chapter 28).

Some generalizations of this generation technique have appeared, so that they overcome some of the limitations that the classical Kerr–Schild transformation had. In particular, the energy–momentum tensor of the classical Kerr-Schild spacetimes generated have the null vector field used in the transformation as a null eigenvector and hence, they cannot be perfect-fluid solutions. A generalization which allows us to find perfect-fluid spacetimes is the generalized Kerr–Schild (GKS hereafter) transformation, which has been studied widely in Refs. 6–15. The GKS transformation is a generation technique in which the metrics of the initial and the final spacetimes are related by the following expression:

$$\tilde{g}_{ab} = g_{ab} + 2\,\mathscr{H}\ell_a\ell_b\,, \tag{1}$$

where the starting metric g_{ab}, which we shall call the *seed metric*, is any metric, $\vec{\ell}$ is a null vector field for the metric g_{ab}, and $\mathscr{H}$ is a scalar field. In the case in which $\vec{\ell}$ is geodesic, an illuminating result was proved in Refs. 6 and 7 (see also Refs. 8–10). It states that the energy–momentum tensor of the GKS spacetimes has $\vec{\ell}$ as an eigenvector if and only if $\vec{\ell}$ is an eigenvector for the energy–momentum tensor of the seed spacetimes. Therefore, if we are interested in finding perfect–fluid spacetimes, we should start with a seed spacetime whose energy–momentum tensor does not have null eigenvectors. In this line, a particularization of the GKS transformation for the search of perfect–fluid solutions was developed in Refs. 6, 7, where the seed spacetime was taking to be conformally-flat. From the application of this technique to several cases, new perfect-fluid

[a)]Also at the Laboratori de Física Matemàtica, Societat Catalana de Física, IEC, Barcelona.
Electronic mail: carlos@hermes.ffn.ub.es

 © 1998 American Institute of Physics

solutions of Einstein equations have been obtained. In some of these applications the solutions represent inhomogeneous cosmological models while in others they are static and stationary spacetimes (see Refs. 6–9, 12, and 16).

The subject of this paper is to apply the GKS transformation, in combination with techniques of null and timelike vector fields, in order to find new perfect-fluid solutions of Einstein's equations. The result found is a wide new family of stationary perfect-fluid spacetimes.

In the development of this work we follow the spirit of the classical Kerr–Schild transformation, where the seed spacetime was very simple (the Minkowski spacetime) and the richness of the transformation lay on the choice of the null vector fields $\vec{\ell}$ used. In this sense, we apply the GKS transformation to the FLRW models with a constant scale factor and taking their most general shear-free geodesic null (SFGN hereafter) vector field as the vector field $\vec{\ell}$. These objects, which are the basic ingredients to carry out the GKS transformation, are studied in section II. In that situation, we study the perfect-fluid Einstein field equations for the GKS metrics, from which we obtain the form of the GKS energy–momentum tensor (energy density, pressure, and fluid velocity) and some partial differential equations for the function $\mathscr{H}$ (section III). Then, through the study of the kinematical quantities of the GKS fluid velocity we impose some rectrictions in order to find stationary perfect-fluid solutions. Moreover, these restrictions together with the choice of the null basis will allow us to integrate the equations for $\mathscr{H}$ without the explicit use of a coordinate system. In addition, we deduce the Petrov type of the GKS spacetimes as well as expressions for the non-zero kinematical quantities of the fluid velocity (section IV).

In section V, we find the explicit form of the solutions. As far as we know, this family was previously unknown. It is noted that it contains all the classical vacuum Kerr–Schild spacetimes, excepting the plane–fronted gravitational waves, and hence the Kerr metric[1] is included too. It is also pointed out that the generalization of the Kerr metric in the Einstein background, due to Vaidya,[16] is also included. However, the family of stationary spacetimes found is much more general and depends on an arbitrary function of one variable. On the other hand, the efficiency of the procedure is shown in section VI, where the Killing equations are completely solved without using coordinates. It is remarkable that, in general, these spacetimes have only one symmetry. Moreover, we discuss the conditions in which other symmetries can appear and their consequences on the GKS spacetimes.

The main equations in this paper are written in the Newman–Penrose formalism.[17] The conventions that we use here are the same as those in Ref. 5 with the only exception that the name for the main vector field of the null basis here is $\vec{\ell}$ whereas in Ref. 5 it is $\vec{k}$. Moreover, throughout this paper we have used units in which $8\pi G=c=1$. Latin indexes run from 0 to 3. The abbreviation ''c.c.'' will stand for complex conjugate.

II. BASIC INGREDIENTS FOR THE GKS TRANSFORMATION

In the classical Kerr–Schild transformation the seed spacetime was the simplest one, the Minkowski spacetime being the important object the null vector field of the transformation, which was the most general SFGN vector field for Minkowski spacetime. Following this line, we take as seeds the subclass of the FLRW spacetimes with a constant scale factor and their most general SFGN vector field as the vector field $\vec{\ell}$ for the transformation.

In what follows, we give the characterization and the explicit form of the seed spacetimes g_{ab} as well as of the SFGN vector fields $\vec{\ell}$.

A. The seed metrics

As we have said, the seed spacetimes are the FLRW spacetimes with a constant scale factor, or equivalently, without expansion. Taking into account the form of the FLRW spacetimes given in appendix A, the line element for the seed spacetimes can be written in the following explicitly conformally-flat form:

$$ds^2=\frac{a^2}{(1+\varepsilon U^2)(1+\varepsilon V^2)}\left\{-4\,dU dV+4\left(\frac{V-U}{1+\xi\bar{\xi}}\right)^2 d\xi d\bar{\xi}\right\}, \tag{2}$$

where a is the constant scale factor and a bar denotes complex conjugation. In the case $\varepsilon=0$ this line element corresponds to the Minkowski spacetime in double null coordinates; in the case $\varepsilon=1$ to the Einstein static Universe and, in the case $\varepsilon=-1$, it corresponds to a spacetime usually not considered because the energy density is negative, however we consider it as a seed spacetime because it may lead to GKS solutions with good physical properties.

These conformally-flat spacetimes have a perfect-fluid matter content and therefore, its energy–momentum tensor is given by

$$T_{ab}=(\varrho+p)u_a u_b+p g_{ab}, \quad g_{ab}u^a u^b=-1,$$

where the energy density ϱ and the pressure p are constant,

$$\varrho=\frac{3\,\varepsilon}{a^2}=-3\,p,$$

and in addition, the fluid velocity is

$$\vec{u}=u^a\frac{\partial}{\partial x^a}=\frac{\partial}{\partial t}=\frac{1}{2\,a}\left\{(1+\varepsilon U^2)\frac{\partial}{\partial U}+(1+\varepsilon V^2)\frac{\partial}{\partial V}\right\},$$

$$\mathbf{u}=u_a dx^a=-dt=-a\left\{\frac{dU}{1+\varepsilon U^2}+\frac{dV}{1+\varepsilon V^2}\right\}.$$

It can be checked that $\vec{u}$ is a constant vector field, that is $\nabla_a u_b=0$. On the other hand, the case $\varepsilon=0$ is Minkowski ($\varrho=p=0$), and although we can define $\vec{u}$, it is not a preferred vector field.

B. The shear-free geodesic null vector field $\vec{\ell}$

In analogy with the pioneering works[4] on the classical Kerr–Schild transformation, we restrict ourselves to the class of shear–free geodesic null vector fields. Although the Goldberg–Sachs theorem[18] tells us that such a kind of vector fields may not exist in general spacetimes, it is common knowledge that in the Minkowski spacetime there is a big family of such vector fields and all of them are known explicitly: this result constitutes the so-called *Kerr's theorem* (see Refs. 5, 19). Moreover, it is well known that a SFGN vector field for Minkowski is also a SFGN vector field for any conformally-flat spacetime, which means that we have the same large class of such vector fields for the seed spacetimes (2). Here, we are going to construct all of them and after this, we will construct an appropriate null basis associated with them.

To begin with, the most general null one-form field for the seed spacetimes (2) can be written in the following form:

$$\vec{\ell}=F\left\{dU+Y\overline{Y}dV+\frac{V-U}{1+\xi\overline{\xi}}(\overline{Y}d\xi+Yd\overline{\xi})\right\}, \tag{3}$$

where F and Y are real and complex arbitrary scalar fields, respectively. Then, in order that $\vec{\ell}$ be geodesic and shear-free, the complex function Y must be a solution of the following system of non-linear partial differential equations:

$$(1+\xi\overline{\xi})YY_{,\xi}=(V-U)Y_{,V}+Y(1+\overline{\xi}Y), \tag{4}$$

$$(1+\xi\overline{\xi})Y_{,\overline{\xi}}=(V-U)YY_{,U}-Y(Y-\xi), \tag{5}$$

where commas stand for partial derivatives with respect to the subscript that follows.

Now, in order to simplify further calculations, we take a null basis $\{\ell,\mathrm{k},\mathrm{m},\overline{\mathrm{m}}\}$ for the seed spacetimes (2) associated with $\vec{\ell}$. We choose this null basis in such a way that $\vec{\ell}$ be affinely parametrized and such that the unit timelike vector field $\vec{u}$ can be written in the following way:

$$\vec{u}=\frac{1}{\sqrt{2}}(\vec{\ell}+\vec{\mathrm{k}}),\quad \mathbf{u}=\frac{1}{\sqrt{2}}(\ell+\mathbf{k}).$$

Part of the remaining freedom is used for setting the imaginary part of the spin coefficient ϵ equals to zero. After some calculations, the explicit expressions for such a null basis $\{\ell,\mathrm{k},\mathrm{m},\overline{\mathrm{m}}\}$ are (3) and

$$\mathbf{k}=F\left\{\frac{1+\varepsilon V^2}{1+\varepsilon U^2}Y\overline{Y}dU+\frac{1+\varepsilon U^2}{1+\varepsilon V^2}dV-\frac{V-U}{1+\xi\overline{\xi}}(\overline{Y}d\xi+Yd\overline{\xi})\right\},$$

$$\mathbf{m}=-F\left\{\sqrt{Y\overline{Y}}\left(-\sqrt{\frac{1+\varepsilon V^2}{1+\varepsilon U^2}}dU+\sqrt{\frac{1+\varepsilon U^2}{1+\varepsilon V^2}}dV\right)\right.$$
$$\left.+\frac{V-U}{1+\xi\overline{\xi}}\sqrt{\frac{\overline{Y}}{Y}}\left(\sqrt{\frac{1+\varepsilon U^2}{1+\varepsilon V^2}}d\xi-Y^2\sqrt{\frac{1+\varepsilon V^2}{1+\varepsilon U^2}}d\overline{\xi}\right)\right\},$$

where the case $Y=0$ can be obtained by taking $Y/\overline{Y}=1$, and F is the following function:

$$F=\frac{-\sqrt{2}a}{(1+\varepsilon U^2)+(1+\varepsilon V^2)Y\overline{Y}}.$$

After long but straightforward calculations we find that the spin coefficients associated with this null basis satisfy the relations

$$\begin{aligned}&\kappa=\sigma=\epsilon=\pi=\lambda=0,\\ &\tau-\overline{\nu}=\rho-\overline{\mu}=\alpha+\overline{\beta}=\gamma+\overline{\gamma}=0.\end{aligned}\tag{6}$$

Furthermore, the Newman–Penrose symbols for the Riemann tensor are the following:

Ricci tensor,

$$\Phi_{01}=\Phi_{02}=\Phi_{12}=0,\quad \Phi_{00}=2\,\Phi_{11}=\Phi_{22}=\frac{1}{4}(\varrho+p)=\frac{\varepsilon}{2a^2},\quad \Lambda=\frac{1}{24}(\varrho-3p)=\frac{\varepsilon}{4a^2};$$

Weyl tensor,

$$\Psi_0=\Psi_1=\Psi_2=\Psi_3=\Psi_4=0.$$

III. THE CONSTRUCTION OF THE GKS SPACETIMES

In this section we study the GKS transformation (1) for the seed spacetimes (2) with (3) as the SFGN vector field $\vec{\ell}$. To that end, we use the fact that given a null basis $\{\ell,\mathrm{k},\mathrm{m},\overline{\mathrm{m}}\}$ associated with the seed metric, we can construct a null basis $\{\widetilde{\ell},\widetilde{\mathrm{k}},\widetilde{\mathrm{m}},\overline{\widetilde{\mathrm{m}}}\}$ associated with the GKS metric by taking (see Refs. 12 and 6)

$$\begin{aligned}&\widetilde{\ell}^a=\ell^a,\quad \widetilde{\ell}_a=\ell_a,\\ &\widetilde{\mathrm{k}}^a=\mathrm{k}^a+\mathscr{H}\ell^a,\quad \widetilde{\mathrm{k}}_a=\mathrm{k}_a-\mathscr{H}\ell_a,\\ &\widetilde{\mathrm{m}}^a=\mathrm{m}^a,\quad \widetilde{\mathrm{m}}_a=\mathrm{m}_a,\end{aligned}\tag{7}$$

where, from now on, objects with an overtilde are associated with the GKS spacetimes in contrast with objects without an overtilde which are associated with the seed spacetimes.

These relationships between null bases allow us to compute all objects associated with the GKS spacetimes as functions of the same objects associated with the seed spacetimes, and the function $\mathscr{H}$ and their derivatives. In this sense, we can compute all the Newman–Penrose symbols for the GKS spacetimes. Taking into account the expressions of the previous section, the result for the components of the Ricci tensor is the following:

$$\widetilde{\Phi}_{00}=\frac{\varepsilon}{2\,a^2},\quad \widetilde{\Phi}_{01}=\widetilde{\Phi}_{02}=0, \tag{8}$$

$$\widetilde{\Phi}_{11}=\frac{\varepsilon}{4\,a^2}+\frac{1}{2}\left[(\rho-\overline{\rho})^2+\frac{\varepsilon}{2\,a^2}\right]\mathscr{H}+\frac{1}{4}(\rho+\overline{\rho})\mathscr{V}-\frac{1}{4}D\mathscr{V}, \tag{9}$$

$$\widetilde{\Phi}_{12}=[\delta\overline{\rho}+(\rho-\overline{\rho})\tau]\mathscr{H}+\overline{\rho}\mathscr{H}+\frac{1}{2}(\tau\mathscr{V}-\delta\mathscr{V}), \tag{10}$$

$$\widetilde{\Phi}_{22}=\frac{\varepsilon}{2\,a^2}+\frac{1}{2}[2\,\triangle(\rho+\overline{\rho})-(\rho-\overline{\rho})^2]\mathscr{H}+\frac{1}{2}(\rho+\overline{\rho})\triangle\mathscr{H}+\frac{1}{2}(\rho+\overline{\rho})\mathscr{V}$$
$$+(\overline{\tau}+\alpha)\mathscr{U}+(\tau+\overline{\alpha})\overline{\mathscr{U}}-\frac{1}{2}(\delta\overline{\mathscr{U}}+\overline{\delta}\mathscr{U})+\frac{\varepsilon}{2\,a^2}\mathscr{H}^2, \tag{11}$$

$$\widetilde{\Lambda}=\frac{\varepsilon}{4\,a^2}+\frac{1}{6}\left[(\rho+\overline{\rho})^2+2\,\rho\overline{\rho}-\frac{3\,\varepsilon}{2\,a^2}\right]\mathscr{H}-\frac{1}{4}(\rho+\overline{\rho})\mathscr{V}+\frac{1}{12}D\mathscr{V}, \tag{12}$$

where $\mathscr{V}$ and $\mathscr{U}$ are defined by

$$\mathscr{V}\equiv D\mathscr{H}+(\rho+\overline{\rho})\mathscr{H},\quad \mathscr{U}\equiv\delta\mathscr{H}.$$

In the same way, we can obtain the components of the Weyl tensor. Their expressions are

$$\widetilde{\Psi}_0=0,\quad \widetilde{\Psi}_1=0, \tag{13}$$

$$3\,\widetilde{\Psi}_2=-(\rho-\overline{\rho})^2\mathscr{H}-\frac{3}{2}(\rho-\overline{\rho})\mathscr{V}-\frac{1}{2}D\mathscr{V}, \tag{14}$$

$$\widetilde{\Psi}_3=-(\rho-\overline{\rho})\overline{\mathscr{U}}+\frac{1}{2}(\overline{\tau}\mathscr{V}-\overline{\delta}\mathscr{V}),\quad \widetilde{\Psi}_4=2\,(\overline{\tau}-\alpha)\overline{\mathscr{U}}-\overline{\delta\mathscr{U}}. \tag{15}$$

As we can see directly from these equations, all the GKS spacetimes that we can obtain are algebraically special because the null vector field $\vec{\ell}=\vec{\ell}$ is a multiple null eigenvector of the Weyl tensor.

Now, we must study the Einstein field equations for the GKS metrics with a perfect–fluid source. They read as follows:

$$\widetilde{G}_{ab}+C\widetilde{g}_{ab}=\widetilde{T}_{ab}, \tag{16}$$

where $\widetilde{G}_{ab}$ is the Einstein tensor for the GKS metrics, C is the cosmological constant, and $\widetilde{T}_{ab}$ is the energy–momentum tensor, which has the form

$$\widetilde{T}_{ab}=(\widetilde{\varrho}+\widetilde{p})\widetilde{u}_a\widetilde{u}_b+\widetilde{p}\,\widetilde{g}_{ab},\quad \widetilde{g}_{ab}\widetilde{u}^a\widetilde{u}^b=-1, \tag{17}$$

where $\tilde{\varrho}$, $\tilde{p}$ and $\tilde{\vec{u}}$ are the energy density, the pressure, and the fluid velocity of the GKS perfect fluid, respectively. In this situation, we project the Einstein equations (16) onto the null basis $\{\tilde{\ell},\tilde{\mathrm{k}},\tilde{\mathrm{m}},\tilde{\overline{\mathrm{m}}}\}$ and then, using (17) we obtain other expressions for the Ricci tensor components,

$$\tilde{\Phi}_{00}=\frac{1}{2}(\tilde{\varrho}+\tilde{p})(\tilde{\ell}^a\tilde{u}_a)^2, \tag{18}$$

$$\tilde{\Phi}_{01}=\frac{1}{2}(\tilde{\varrho}+\tilde{p})(\tilde{\ell}^a\tilde{u}_a)(\tilde{\mathrm{m}}^b\tilde{u}_b), \quad \tilde{\Phi}_{02}=\frac{1}{2}(\tilde{\varrho}+\tilde{p})(\tilde{\mathrm{m}}^a\tilde{u}_a)^2, \tag{19}$$

$$\tilde{\Phi}_{11}=\frac{1}{4}(\tilde{\varrho}+\tilde{p})[(\tilde{\ell}^a\tilde{u}_a)(\tilde{\mathrm{k}}^b\tilde{u}_b)+(\tilde{\mathrm{m}}^a\tilde{u}_a)(\tilde{\overline{\mathrm{m}}}^b\tilde{u}_b)], \tag{20}$$

$$\tilde{\Phi}_{12}=\frac{1}{2}(\tilde{\varrho}+\tilde{p})(\tilde{\mathrm{k}}^a\tilde{u}_a)(\tilde{\mathrm{m}}^a\tilde{u}_a), \quad \tilde{\Phi}_{22}=\frac{1}{2}(\tilde{\varrho}+\tilde{p})(\tilde{\mathrm{k}}^a\tilde{u}_a)^2, \tag{21}$$

$$\tilde{\Lambda}=\frac{1}{24}(\tilde{\varrho}-3\,\tilde{p})+\frac{1}{6}C. \tag{22}$$

The next step in this process is to compare the expressions for the Ricci tensor (8)–(12) with the Einstein equations for the GKS metrics (18)–(22). We do this for the seed metrics with $\varepsilon\neq 0$, so that $\tilde{\varrho}+\tilde{p}\neq 0$, and later we will extend the results for the $\varepsilon=0$ case. The outcome of this comparison is the following.

• The components, in the null bases $\{\ell,\mathrm{k},\mathrm{m},\overline{\mathrm{m}}\}$ and $\{\tilde{\ell},\tilde{\mathrm{k}},\tilde{\mathrm{m}},\tilde{\overline{\mathrm{m}}}\}$, for the fluid velocity $\tilde{\vec{u}}$ of the GKS perfect fluid:

$$(\tilde{\ell}^a\tilde{u}_a)^2=(\ell^a\tilde{u}_a)^2=\frac{\varepsilon}{2a^2}\left[\frac{\varepsilon}{a^2}+2\left((\rho-\overline{\rho})^2+\frac{\varepsilon}{2a^2}\right)\mathscr{H}+(\rho+\overline{\rho})\mathscr{V}-D\mathscr{V}\right]^{-1}, \tag{23}$$

$$\tilde{\mathrm{k}}^a\tilde{u}_a=\mathrm{k}^a\tilde{u}_a+\mathscr{H}\ell^a\tilde{u}_a=[2\,(\tilde{\ell}^a\tilde{u}_a)]^{-1}, \tag{24}$$

$$\tilde{\mathrm{m}}^a\tilde{u}_a=\mathrm{m}^a\tilde{u}_a=0. \tag{25}$$

• The energy density and the pressure of the GKS perfect fluid:

$$\tilde{\varrho}=\frac{3\,\varepsilon}{a^2}-C+2\,(2\,\rho^2-\rho\,\overline{\rho}+2\,\overline{\rho^2})\mathscr{H}-D\mathscr{V}, \tag{26}$$

$$\tilde{p}=-\frac{\varepsilon}{a^2}+C+2\left(-3\,\rho\,\overline{\rho}+\frac{\varepsilon}{a^2}\right)\mathscr{H}+2\,(\rho+\overline{\rho})\mathscr{V}-D\mathscr{V}. \tag{27}$$

• Two second order partial differential equations for $\mathscr{H}$:

$$\delta\mathscr{V}=\tau\mathscr{V}+2\,\overline{\rho}\,\mathscr{U}+2\,[\delta\overline{\rho}+(\rho-\overline{\rho})\,\tau]\mathscr{H}, \tag{28}$$

$$\begin{aligned}(\rho+\overline{\rho})\triangle\mathscr{H}=&\left[(\rho-\overline{\rho})^2+\frac{\varepsilon}{a^2}-2\,\Delta(\rho+\overline{\rho})\right]\mathscr{H}-(\rho+\overline{\rho})\mathscr{V}+\delta\overline{\mathscr{U}}\\&+\overline{\delta}\mathscr{U}-2\,(\overline{\tau}+\alpha)\mathscr{U}-2\,(\tau+\overline{\alpha})\overline{\mathscr{U}}+\left(\frac{\varepsilon}{a^2}\right)^{-1}\mathscr{W},\end{aligned} \tag{29}$$

where $\mathscr{W}$ is given by

$$\mathscr{W}=[D\mathscr{V}-(\rho+\overline{\rho})\mathscr{V}-2(\rho-\overline{\rho})^2.\mathscr{H}]\left[D\mathscr{V}-(\rho+\overline{\rho})\mathscr{V}-2(\rho-\overline{\rho})^2.\mathscr{H}-\frac{2\,\varepsilon}{a^2}(1+\mathscr{H})\right]. \tag{30}$$

As we can see, equation (29) is a nonlinear partial differential equation for $\mathscr{H}$ due to the term with $\mathscr{W}$.

IV. CHARACTERIZATION OF THE GKS SPACETIMES THROUGH THE KINEMATICAL QUANTITIES

Until now, the only assumption we have made on the GKS spacetimes is that their energy–momentum tensor must be of the perfect-fluid type. Then, once we have a function $\mathscr{H}$ solution of the system (28)–(29), we have the explicit form of the GKS metric as well as of the matter content variables. However, there is not a systematic way of solving (28)–(29), especially because (29) is a nonlinear equation. For this reason, we are going to introduce some additional assumptions on the GKS spacetimes which can help us to integrate these equations. Of course, it would be advisable to impose conditions which have a physical meaning. In this sense, an interesting way to control these conditions is through the study of the kinematical quantities for the fluid velocity $\vec{\tilde{u}}$ of the GKS perfect fluid.

From equations (23)–(25) it follows that $\vec{\tilde{u}}$ lies in the two-planes generated by $\vec{\ell}$ and $\vec{\tilde{k}}$. Then, by making the following change of basis:

$$\vec{\ell}\to\vec{L}=\mathscr{A}\vec{\ell},\quad \vec{\tilde{k}}\to\vec{K}=\mathscr{A}^{-1}\,\vec{\tilde{k}},\quad \vec{\tilde{m}}\to\vec{\tilde{m}}, \tag{31}$$

where

$$\mathscr{A}^2\equiv[2\,(\ell^a\tilde{u}_a)^2]^{-1}, \tag{32}$$

we get the following form for $\vec{\tilde{u}}$:

$$\vec{\tilde{u}}=\frac{1}{\sqrt{2}}(\vec{L}+\vec{K}).$$

Therefore, we can now use the formulas given in Appendix B to obtain the kinematical quantities of $\vec{\tilde{u}}$. To that end, we compute the spin coefficients associated with the null basis $\{L,K,\tilde{m},\overline{\tilde{m}}\}$, which we denote with a hat. The most useful results are

$$\hat{\rho}+\hat{\tilde{\rho}}=(\rho+\overline{\rho}).\mathscr{A},\quad \hat{\mu}+\hat{\tilde{\mu}}=(\rho+\overline{\rho})(1-\mathscr{H}).\mathscr{A}^{-1},\quad \hat{\epsilon}+\hat{\tilde{\epsilon}}=D.\mathscr{A},$$

$$\hat{\gamma}+\hat{\tilde{\gamma}}=-\mathscr{A}^{-1}D.\mathscr{H}+(\triangle.\mathscr{A}+\mathscr{H}D.\mathscr{A}).\mathscr{A}^{-2},\quad \hat{\alpha}+\hat{\tilde{\beta}}=\mathscr{A}^{-1}\overline{\delta}.\mathscr{A},$$

$$\hat{\pi}+\hat{\tilde{\tau}}-\hat{\nu}-\hat{\tilde{\kappa}}=\overline{\tau}+[\,\overline{\delta}.\mathscr{H}-\overline{\tau}(1+\mathscr{H})].\mathscr{A}^{-2},\quad \hat{\sigma}=\hat{\lambda}=0,$$

$$\hat{\rho}-\hat{\overline{\tilde{\rho}}}=(\rho-\overline{\rho}).\mathscr{A},\quad \hat{\mu}-\hat{\tilde{\mu}}=-(\rho-\overline{\rho})(1+\mathscr{H}).\mathscr{A}^{-1}.$$

In this situation, we are going to impose conditions which can lead to stationary spacetimes. To that end, we impose the vanishing of the shear and expansion of the fluid velocity $\vec{\tilde{u}}$, that is,

$$\tilde{\theta}=\tilde{\sigma}_{ab}=0. \tag{33}$$

These conditions are necessary in order to have $\vec{\tilde{u}}$ proportional to a time-like Killing vector field. Then, from the vanishing of the expansion and the $m_{(a}\overline{m}_{b)}$-component of the shear tensor (see appendix B) we have

$$(\rho+\overline{\rho})(\mathscr{A}^2-1+\mathscr{H})=0,$$

so that, two different cases may appear: $\rho+\overline{\rho}=0$ or $\mathscr{A}^2=1-\mathscr{H}$. If $\rho+\overline{\rho}=0$ the SFGN vector field $\vec{\ell}$ is expansion–free. In addition, from the Newman–Penrose equations[17] we obtain that $\rho^2+\varepsilon/(2\,a^2)=0$, which means that only the case $\varepsilon=1$ is possible. However, further analysis of the equations for this case shows that there are not solutions in this case. Therefore, from now on we will consider that $\rho+\overline{\rho}\neq 0$, and then we must follow through the second possibility,

$$\mathscr{A}^2=1-\mathscr{H}. \tag{34}$$

From this equation and using the expressions (23), (32) we get another differential equation for $\mathscr{H}$,

$$D\mathscr{V}=(\rho+\overline{\rho})\mathscr{V}+2\left[(\rho-\overline{\rho})^2+\frac{\varepsilon}{a^2}\right]\mathscr{H}. \tag{35}$$

If we use this equation, we can see from (30) that $\mathscr{W}=-4\,(\varepsilon^2/a^2).\mathscr{H}$, and therefore (29) is a linear differential equation for $\mathscr{H}$. Moreover, from the remaining components of equations (33) we have the following two additional conditions on $\mathscr{H}$ and Y:

$$(D+\triangle).\mathscr{H}=0, \tag{36}$$

$$\tau=0, \tag{37}$$

respectively. The first is a further differential equation for $\mathscr{H}$, while the second one is a constraint on the possible complex functions Y that we can use to construct the SFGN vector fields $\vec{\ell}$. Then, replacing (34) into expressions (23)–(25) for the fluid velocity $\tilde{\vec{u}}$ we obtain

$$\tilde{\vec{u}}=\frac{1}{\sqrt{2\,(1-\mathscr{H})}}[(1-\mathscr{H})\tilde{\vec{\ell}}+\tilde{\vec{\mathrm{k}}}]=\frac{1}{\sqrt{2\,(1-\mathscr{H})}}(\vec{\ell}+\vec{\mathrm{k}})=\frac{1}{\sqrt{1-\mathscr{H}}}\vec{u}, \tag{38}$$

that is to say, $\tilde{\vec{u}}$ is proportional to $\vec{u}$. On the other hand, it can be shown that the vector field $\vec{\mathrm{u}}$, which is a Killing vector field for the seed spacetimes (2), is also a Killing vector field for the GKS spacetimes provided that (36) and (37) hold. Thus, $\tilde{\vec{u}}$ is parallel to a Killing vector field. We can see this fact from the following expression:

$$\tilde{\boldsymbol{\nabla}}_a\tilde{u}'_b+\tilde{\boldsymbol{\nabla}}_b\tilde{u}'_a=\sqrt{2}\{(D+\triangle).\mathscr{H}\ell_a\ell_b+2[\tau\ \mathrm{k}_{(a}\mathrm{m}_{b)}+\text{ c.c.}].\mathscr{H}=0, \tag{39}$$

where $\tilde{u}'_a=\tilde{g}_{ab}u^b$. Here, we have used (6) and the second equality follows from (36),(37). Notice that this is not a trivial result because, even though conditions (33) are necessary for $\tilde{u}$ to be proportional to a Killing, they are not sufficient (we also need the acceleration $\tilde{\mathbf{a}}$ to be an exact 1-form). Therefore, we have proved that (33) implies $\mathbf{d}\tilde{\mathbf{a}}=\mathbf{0}$ under our conditions.

On the other hand, in the last section we restricted ourselves to the cases $\varepsilon=\pm 1$. Now, we can include the case $\varepsilon=0$ by defining the vector field $\tilde{\vec{u}}$ through the expression (38), and by taking the same partial differential equations (28)–(29) and (35)–(36) for $\mathscr{H}$.

The next point is to find which are the integrability conditions for the system of *linear* partial differential equations for $\mathscr{H}$. In this process we find two different cases.

• *Case A:* $\overline{\delta\rho}=0\Rightarrow\rho-\overline{\rho}=0$. It can be shown that all the GKS metrics that we can find in this case are included in a particular case of Ref. 8, where the authors use the GKS transformation with the interior Schwarzschild metric as the seed metric. The resultant GKS spacetimes in that paper are stationary spherically symmetric and they contain a subfamily of regular solutions with equation of state $\tilde{\varrho}+3\,\tilde{p}=\text{const}$. Among these spacetimes we have the Whittaker metric[20] as a special case, which, at the same time, is the static limit of the Wahlquist metric.[21]

• *Case B:* $\overline{\delta\rho}\neq 0$. In this case the procedure to find the integrability conditions of the system of partial differential equations (28)–(29) and (35)–(36) for $\mathscr{H}$, leads us to the following equations for the first derivatives of $\mathscr{H}$:

$$D.\mathscr{H}=\frac{\rho^2+\overline{\rho^2}+(\varepsilon/a^2)}{\rho+\overline{\rho}}\mathscr{H}=-\Delta.\mathscr{H},\quad \delta.\mathscr{H}=\frac{\delta\overline{\rho}}{\rho+\overline{\rho}}\mathscr{H}. \tag{40}$$

These equations can be easily solved without the explicit use of coordinates and the solution is

$$\mathscr{H}=m(\rho+\overline{\rho}), \tag{41}$$

where m is an arbitrary constant.

The energy–momentum tensor of these GKS spacetimes can be obtained by substituting $\mathscr{H}$ in the explicit form of the fluid velocity, the energy density, and the pressure, given in (38) and (26)–(27), respectively. In particular, the energy density and the pressure are

$$\widetilde{\varrho}=\frac{3\,\varepsilon}{a^2}[1-m(\rho+\overline{\rho})]-C,\quad \widetilde{p}=\frac{\varepsilon}{a^2}[-1+m(\rho+\overline{\rho})]+C, \tag{42}$$

so that we have the following equation of state:

$$\widetilde{\varrho}+3\,\widetilde{p}=2C.$$

The most interesting case from the physical point of view is the one given by $C\geqslant 0$ and $\varepsilon=1$, for which there are spacetime regions where all the energy conditions are fulfilled (see Ref. 22). In this case, from the form (38) of $\tilde{\tilde{u}}$, it follows that the region of the spacetime in which we can define this unit timelike vector field is

$$1-\mathscr{H}>0.$$

Moreover, the case $\varepsilon=C=0$ corresponds to vacuum solutions. In the next section we will analyze to which solutions correspond each case.

On the other hand, from the expressions (13)–(15) for the Weyl tensor we already know that the GKS spacetimes are algebraically special. In this case, the Petrov type is II in general, except when the following relation holds:

$$3\,\widetilde{\Psi}_2\widetilde{\Psi}_4=2\,\widetilde{\Psi}_3^2\Rightarrow\left(\rho^2+\frac{\varepsilon}{2\,a^2}\right)(\overline{\delta\,\delta\rho}+2\,\alpha\,\overline{\delta\rho})=3\,\rho(\overline{\delta\rho})^2. \tag{43}$$

In that case the Petrov type is D and the fluid velocity does not lie in the preferred two-space spanned by the two multiple null eigenvectors of the Weyl tensor.

Finally, the non-vanishing kinematical quantities for $\tilde{\tilde{u}}$ are

$$\tilde{\tilde{a}}=-\frac{D.\mathscr{H}}{\sqrt{2\,(1-\mathscr{H})}}\tilde{\tilde{v}}-\left(\frac{\overline{\delta}.\mathscr{H}}{1-\mathscr{H}}\tilde{\tilde{m}}+\text{ c.c.}\right),$$

$$\widetilde{\omega}_{ab}=-\left(\frac{\overline{\delta}.\mathscr{H}}{1-\mathscr{H}}\widetilde{v}_{[a}\widetilde{m}_{b]}+\text{ c.c.}\right)-\frac{\sqrt{2}(\rho-\overline{\rho}).\mathscr{H}}{\sqrt{1-\mathscr{H}}}\widetilde{m}_{[a}\widetilde{\overline{m}}_{b]}, \tag{44}$$

where $\widetilde{v}^a$ is a unit spacelike vector field, orthogonal to $\widetilde{u}^a$, and is defined as follows:

$$\widetilde{v}^a\equiv\frac{(1-\mathscr{H})\widetilde{\gamma}^a-\widetilde{k}^a}{\sqrt{2\,(1-\mathscr{H})}}\Rightarrow\widetilde{v}^a\widetilde{v}_a=1,\quad \widetilde{v}^a\widetilde{u}_a=0.$$

It is interesting to note that from expression (44) and from equations (40)–(41) we can deduce that the rotation $\widetilde{\omega}_{ab}$ of $\widetilde{u}^a$ vanishes if and only if so does the rotation $\rho-\overline{\rho}$ of ℓ^a.

V. EXPLICIT FORM OF THE SOLUTIONS

After the integration of the equations for $\mathscr{H}$, we are going to find an explicit expression for the line element of the GKS spacetimes. For this we must find explicitly all the SFGN vector fields $\vec{\ell}$ satisfying the condition (37), that is, we must find the explicit solutions of the partial differential equations (4)–(5) together with (37). To do that, it is more convenient to consider a new function Ω instead of Y, which is defined by

$$Y(U,V,\xi,\overline{\xi})\equiv\sqrt{\frac{1+\varepsilon U^2}{1+\varepsilon V^2}}\Omega(t,\chi,\xi,\overline{\xi}),$$

and also, it is better to change from the coordinates $\{U,V\}$ to the $\{t,\chi\}$ ones (see appendix A). Once we have performed these changes, the equations (4)–(5) and (37) read as follows:

$$(1+\xi\overline{\xi})\Omega\Omega_{,\xi}-\overline{\xi}\Omega^2=\Sigma\Omega_{,\chi}+\Sigma'\Omega, \tag{45}$$

$$(1+\xi\overline{\xi})\Omega_{,\overline{\xi}}+\xi\Omega=-\Sigma\Omega\Omega_{,\chi}+\Sigma'\Omega^2, \tag{46}$$

$$\Omega_{,t}=0, \tag{47}$$

where $\Sigma(\varepsilon,\chi)$ is given in appendix A. The most general solution of these equations can be given implicitly by

$$G\left(\frac{e^{-\sqrt{-\varepsilon}\chi}\Omega-\xi}{1+\overline{\xi}\,e^{-\sqrt{-\varepsilon}\chi}\Omega},\frac{(1-\xi\overline{\xi})\Sigma'\Omega-\xi+\overline{\xi}\Omega^2}{(1+\xi\overline{\xi})\Sigma\Omega}\right)=0, \tag{48}$$

where $G(z_1,z_2)$ is any analytic complex function of two complex variables $\{z_1,z_2\}$. It is interesting to note that, in the same way, we could have found the general solution of only equations (45) and (46), then we would have obtained all the possible SFGN vector fields for the seed metrics (2), which constitutes the analogous result of the Kerr theorem. However, for the sake of brevity we do not give here the explicit form of the general solution. From the definition of the spin coefficient ρ, and using (45)–(47) we have

$$\rho=\begin{cases}\dfrac{1}{\sqrt{2}a}\dfrac{\Omega_{,\chi}}{\Omega}, & \text{for } \Omega\neq 0,\\[2ex] \dfrac{1}{\sqrt{2}a}\dfrac{\Sigma'}{\Sigma}, & \text{for } \Omega=0,\end{cases}$$

and then, the solution (41) for $\mathscr{H}$ is

$$\mathscr{H}=\begin{cases}\dfrac{m}{\sqrt{2}a}[\ln(\Omega\overline{\Omega})]_{,\chi}, & \text{for } \Omega\neq 0,\\[2ex] \dfrac{\sqrt{2}m}{a}\dfrac{\Sigma'}{\Sigma}, & \text{for } \Omega=0.\end{cases}$$

Now, making the substitution of the expressions (41) for the function $\mathscr{H}$ and (38) for the SFGN 1-form $\vec{\ell}$ into the definition of the GKS transformation (2), we find the following explicit form for the line element of the GKS metrics:

$$ds^2=-dt^2+a^2\left(d\chi^2+\frac{4\Sigma^2 d\xi d\overline{\xi}}{(1+\xi\overline{\xi})^2}\right)$$
$$+\frac{a^2\mathscr{H}}{(1+\Omega\overline{\Omega})^2}\left[(1+\Omega\overline{\Omega})\frac{dt}{a}-(1-\Omega\overline{\Omega})d\chi+\frac{2\Sigma}{1+\xi\overline{\xi}}(\overline{\Omega}d\xi+\Omega d\overline{\xi})\right]^2. \qquad (49)$$

It is interesting to point out that this family of stationary solutions of Einstein's equations has an arbitrary function, namely Ω. Moreover, as far as we know, these metrics were previously unknown.

With regard to the complex function Ω, if the function $G(z_1,z_2)$ does not depend on z_2 we have

$$\Omega=\mathrm{e}^{\sqrt{-\varepsilon}\chi}f(\xi,\overline{\xi}),$$

where f is a complex function that we can find by solving the system of differential equations (45)–(46). This form of Ω implies that

$$\mathscr{H}=\begin{cases}0, & \text{if } \varepsilon=0,1,\\ \dfrac{\sqrt{2}m}{a}, & \text{if } \varepsilon=-1,\end{cases}$$

and this means that in order to avoid $\mathscr{H}=0$ for the cases $\varepsilon=0,1$ (which would mean that the GKS transformation is trivial), we must impose that

$$\frac{\partial G}{\partial z_2}\neq 0,$$

and therefore, by the implicit function theorem, we can write (48) as follows:

$$\Upsilon\left(\frac{\mathrm{e}^{-\sqrt{-\varepsilon}\chi}\Omega-\xi}{1+\overline{\xi}\,\mathrm{e}^{-\sqrt{-\varepsilon}\chi}\Omega}\right)+\frac{(1-\xi\overline{\xi})\Sigma'\Omega-\xi+\overline{\xi}\Omega^2}{(1+\xi\overline{\xi})\Sigma\Omega}=0,$$

where $\Upsilon(z)$ is any analytic complex function of one complex variable z. Moreover, as we can see directly from the last equation, or from equations (45)–(47) and the expression for the line-element (49), the transformation

$$\Omega\to\Omega'=-\frac{1}{\overline{\Omega}},\quad \chi\to\chi'=-\chi,$$

is a *gauge* transformation since it induces the change $\vec{\ell}\to\vec{\ell}\,'=-\vec{\ell}$, and therefore both solutions of equations (45)–(47), Ω and Ω', determine the same GKS metric. Apart from this gauge, we have not been able to find any other. On the other hand, we can think, as Debney[23] did, that part of the complex functions Ω are gauges of the seed metric, or in other words, that the GKS metrics we obtain are the seed metrics. Looking at the expressions (13)–(15) for the Weyl tensor, this can only happen when $\rho+\overline{\rho}=0\Rightarrow\overline{\delta\rho}=0$, which corresponds to Case A. Therefore, the GKS metrics (49) are always different from the seed metrics.

As we said before, the family of solutions (49) is too wide in the sense that it depends on the arbitrary function Ω. Analyzing the form of the metric (see also Ref. 10) the solutions contained in each case are the following.

• Case $\varepsilon=0$: If we do not consider the cosmological constant ($C=0$), the metrics (49) are all the classical vacuum Kerr–Schild metrics[4] with the exception of the plane-fronted gravitational waves.[24,23,25] This exception is due to the fact that this class of solutions corresponds to the case $\rho=0$ (see Ref. 5). The Schwarzschild spacetime is obtained when $Y=\Omega=0$, and the Kerr spacetime when we choose Υ in the following way:

$$\Upsilon = -\frac{1}{\sqrt{2}\,\mathrm{i}c}, \tag{50}$$

where c is an arbitrary real constant. In this case, if we carry out the suitable changes (see Ref. 10), we can identify the mass and the angular momentum as measured from infinity with $m/(4\sqrt{2}a)$ and $(mc)/(4\sqrt{2}a)$, respectively.

• Case $\varepsilon=1$: All the spacetimes within this case were unknown with the exception of the case given by (50), corresponding to a metric due to Vaidya[16] (it is also a particular case of the Wahlquist metric[21,26]) which has been interpreted as the Kerr metric in the cosmological background of the Einstein Universe. In this sense, the metrics of this class can be interpreted as the vacuum Kerr–Schild configurations in the background of the Einstein Universe.

• Case $\varepsilon=-1$: These spacetimes were also unknown and they correspond to the Kerr–Schild configurations in the background of the seed spacetimes (2) with $\varepsilon=-1$. From the physical point of view these spacetimes have less interest since they do not satisfy the usual energy conditions.[22]

VI. SOLUTION OF THE KILLING EQUATIONS AND STUDY OF THE SYMMETRIES

The subject of this section is the study of the symmetries of the GKS spacetimes (49). A remarkable point in this study is the fact that it is possible to integrate the Killing equations without the explicit use of coordinates, as it happened with the equations for $\mathscr{H}$. To carry out this integration, we consider the Killing equations for these metrics,

$$\tilde{\nabla}_a\zeta_b+\tilde{\nabla}_b\zeta_a=0, \tag{51}$$

where $\vec{\zeta}$ is a hypothetical Killing vector field for the metrics (49), and we project them onto the null basis $\{\ell,\mathrm{k},\mathrm{m},\overline{\mathrm{m}}\}$ associated with the seed spacetimes, which allows us to use its spin coefficients and their properties (6).

After making the integration (see Ref. 10 for details), we find that in the general case, that is to say, without additional assumptions on the complex function Ω, there is only one independent Killing vector field. Obviously, this Killing vector field is the timelike one that we already know,

$$\vec{\zeta}=\frac{1}{\sqrt{2}}[(1-\mathscr{H})\tilde{\vec{\ell}}+\tilde{\vec{\mathbf{k}}}]=\frac{1}{\sqrt{2}}(\vec{\ell}+\vec{\mathbf{k}})=\vec{u}\propto\tilde{\vec{u}}. \tag{52}$$

Thus, the solution (49) is stationary in general with *no* further symmetries. The norm of this Killing vector field is $\mathscr{H}-1$, which for the cases $\varepsilon\pm1$ is always negative [otherwise we cannot define $\tilde{\vec{u}}$ (38)], but in the case $\varepsilon=0$ there are not restrictions and then, the sign of the norm divides the spacetime into three regions: (A) $1-\mathscr{H}>0$, where the Killing is timelike. (B) $1-\mathscr{H}<0$, where the Killing is spacelike. (C) $1-\mathscr{H}=0$, where the Killing is null. In fact, that is the stationary limit hypersurface (this can be checked easily for the Schwarzschild and Kerr spacetimes). This division according to the norm of the Killing vector field (52) is an interesting fact because these three regions are usually seen as different spacetimes (see Ref. 5).

Apart from the general case, there are specific cases with additional symmetries. These cases appear when the quantity $\Gamma+\overline{\Gamma}$ vanishes, where Γ is a complex function defined by

$$\Gamma\equiv\frac{1}{(\overline{\delta\rho})^2}\left\{\left(\rho^2+\frac{\varepsilon}{2\,a^2}\right)(\overline{\delta\,\delta\rho}+2\,\alpha\,\overline{\delta\rho})-3\,\rho(\overline{\delta\rho})^2\right\}. \tag{53}$$

Now, two different cases appear depending on whether the complex function Γ vanishes or not.

Case (i): $\Gamma\neq0$. After integrating the Killing equations (51) we find that there are two independent Killing vector fields, which we can take as ζ^a, given in (52), and ζ_1^a, given by

$$\vec{\zeta}_1=\frac{1}{\sqrt{|\Gamma|}}\left\{(1+\mathscr{H})\vec{\ell}-\vec{\mathrm{k}}-\frac{2(\overline{\rho^2}+\varepsilon/(2a^2))}{\delta\overline{\rho}}\vec{\tilde{\mathrm{m}}}-\frac{2(\rho^2+\varepsilon/(2a^2))}{\overline{\delta\rho}}\vec{\mathrm{m}}\right\}. \tag{54}$$

It is clear that we can write the most general Killing vector field as follows:

$$C_1\vec{\zeta}+C_2\vec{\zeta}_1, \tag{55}$$

where C_1 and C_2 are two arbitrary constants. It can be shown that this general Killing vector field commutes with the timelike one $\vec{\zeta}$ given in (52). Furthermore, we can study the possible existence of an axial Killing vector field. To that end we must check the regularity condition,

$$\frac{\tilde{\nabla}^a\zeta^2\tilde{\nabla}_a\zeta^2}{4\,\zeta^2}\overset{\text{axis}}{\to}1, \quad \zeta^2\equiv\zeta^a\zeta_a. \tag{56}$$

For the general Killing vector field (55) this condition is not satisfied for any value of C_1 and C_2 and therefore, there is no axial symmetry in this case.

Case (ii): $\Gamma=0$. In this case, equation (43) implies that the GKS spacetimes (49) are Petrov type D. When we integrate the Killing equations (51), we find that there are also two independent Killing vector fields, which can be taken as (52) and

$$\vec{\zeta}_2=\frac{1}{(\rho^2+\varepsilon/(2a^2))^{3/2}(\overline{\rho^2}+\varepsilon/(2a^2))^{3/2}}$$
$$\times\left\{\overline{\delta\rho}\,\delta\overline{\rho}[(1+\mathscr{H})\vec{\ell}-\vec{\mathrm{k}}]-2\left(\overline{\rho^2}+\frac{\varepsilon}{2\,a^2}\right)\overline{\delta\rho}\vec{\tilde{\mathrm{m}}}-2\left(\rho^2+\frac{\varepsilon}{2\,a^2}\right)\delta\rho\,\vec{\overline{\mathrm{m}}}\right\}. \tag{57}$$

In the same way as in case (i), we can check that the general expression for a Killing vector field $C_1\vec{\zeta}+C_2\vec{\zeta}_2$, where C_1 and C_2 are two arbitrary constants, commutes with the timelike Killing vector field $\vec{\zeta}$. In addition, and with regard to the possibility of axial symmetry, it can be shown that we can only have a spacelike Killing vector field vanishing in a certain region of the spacetime when $C_1=0$, and then, this region is defined by $|\overline{\delta\rho}|=0$. Moreover, it can be checked that if the following relation holds:

$$(\rho-\overline{\rho})^{-2}\left(\rho^2+\frac{\varepsilon}{2\,a^2}\right)\left(\overline{\rho^2}+\frac{\varepsilon}{2\,a^2}\right)\Bigg|_{|\overline{\delta\rho}|\to0}\to\text{const}, \tag{58}$$

and we choose the constant C_2 as follows:

$$C_2=\sqrt{\frac{(\rho^2+\varepsilon/(2\,a^2))(\overline{\rho^2}+\varepsilon/(2\,a^2))}{-8(\rho-\overline{\rho})^2}}\Bigg|_{|\overline{\delta\rho}|\to0},$$

the axis regularity condition (56) is satisfied for the Killing vector field $C_2\vec{\zeta}_2$, and therefore, it is an axial Killing vector field. In that case, the GKS spacetimes would be stationary axially symmetric spacetimes in *rigid rotation* because $\vec{\tilde{u}}$ is proportional to the timelike Killing vector field (52). For the Kerr and the Kerr–Vaidya metrics, which are given by (50) and $\varepsilon=0,1$, respectively, it can be checked that the condition (58) is fulfilled. In general, following Ref. 27, we can conclude that all the spacetimes in this case with axial symmetry are subcases of the Wahlquist family.[21]

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. J. M. M. Senovilla for helpful discussions, many valuable suggestions and for his collaboration in the preparation of this manuscript. Some calculations have been performed with the help of the computer program REDUCE.

APPENDIX A: CONFORMALLY FLAT FORM OF THE FLRW SPACETIMES

In this appendix we are going to give a coordinate change to pass from the following well-known form for the line element of the FLRW spacetimes,

$$ds^2=-dt^2+a^2(t)\{d\chi^2+\Sigma^2(\varepsilon,\chi)(d\theta^2+\sin^2\theta d\varphi^2)\}, \tag{A1}$$

where $\varepsilon^3=\varepsilon$ and $\Sigma(\varepsilon,\chi)$ satisfy the differential equation

$$\Sigma'^2+\varepsilon\Sigma^2=1\Leftrightarrow\Sigma(\varepsilon,\chi)=\begin{cases}\sin\chi, & \text{if } \varepsilon=1,\\ \chi, & \text{if } \varepsilon=0,\\ \sinh\chi & \text{if } \varepsilon=-1,\end{cases}$$

to the following explicitly conformally-flat form:

$$ds^2=\Phi^2(T,R)\{-dT^2+dR^2+R^2(d\theta^2+\sin^2\theta d\varphi^2)\}.$$

This transformation can be achieved by using the following coordinate change:

$$T=\frac{2\,\Sigma(\varepsilon,\psi)}{\Sigma'(\varepsilon,\psi)+\Sigma'(\varepsilon,\chi)},\quad R=\frac{2\,\Sigma(\varepsilon,\chi)}{\Sigma'(\varepsilon,\psi)+\Sigma'(\varepsilon,\chi)},$$

where ψ is the so-called *parametric time*, defined by

$$\dot{\psi}=\frac{d\psi}{dt}\equiv\frac{1}{a}.$$

After making this coordinate change, the conformal factor Φ becomes

$$\Phi^2=\frac{a^2}{[1+\varepsilon((T-R)/2)^2][1+\varepsilon((T+R)/2)^2]}.$$

In these coordinates the fact that the scale factor a depends only on t reads as follows:

$$a=a\left(\frac{T}{1-(\varepsilon/4)(T^2-R^2)}\right).$$

Another interesting form for the line element (A1) is obtained by passing to the following null spherical coordinates:

$$U=\frac{T-R}{2},\quad V=\frac{T+R}{2},\quad \xi=e^{i\varphi}\tan\left(\frac{\theta}{2}\right);$$

then we obtain

$$ds^2=\Phi^2(U,V)\left\{-4dUdV+4\left(\frac{V-U}{1+\xi\bar{\xi}}\right)^2 d\xi d\bar{\xi}\right\},\quad \Phi^2(U,V)=\frac{a^2}{(1+\varepsilon U^2)(1+\varepsilon V^2)},$$

where

$$a=a\left(\frac{V+U}{1-\varepsilon UV}\right).$$

APPENDIX B: KINEMATICAL QUANTITIES AND THE NEWMAN–PENROSE FORMALISM

In this appendix we give some useful formulae due to Wainwright,[28] for the kinematical quantities of a unit timelike vector field. Let $\vec{u}$ be any unit timelike vector field and let $\{\ell, \mathrm{k}, \mathrm{m}, \overline{\mathrm{m}}\}$ be a null basis such that we can write down $\vec{u}$ in the following way:

$$\vec{u} = \frac{1}{\sqrt{2}}(\vec{\ell} + \vec{\mathrm{k}}).$$

Then, we can write the kinematical quantities associated with $\vec{u}$ using only the null basis $\{\ell, \mathrm{k}, \mathrm{m}, \overline{\mathrm{m}}\}$ and the spin coefficients associated with them as follows:

Expansion,

$$\theta = \frac{1}{\sqrt{2}}(\epsilon + \overline{\epsilon} - \gamma - \overline{\gamma} + \mu + \overline{\mu} - \rho - \overline{\rho});$$

acceleration,

$$\vec{\mathrm{a}} = \frac{1}{\sqrt{2}}(\epsilon + \overline{\epsilon} + \gamma + \overline{\gamma})\vec{v} + \frac{1}{2}\{(\pi - \overline{\tau} + \nu - \overline{\kappa})\vec{\mathrm{m}} + \text{ c.c.}\};$$

shear,

$$\sigma_{ab} = \frac{1}{3\sqrt{2}}\{2(\epsilon + \overline{\epsilon}) + \rho + \overline{\rho} - 2(\gamma + \overline{\gamma}) - \mu - \overline{\mu}\}(v_a v_b - \mathrm{m}_{(a}\overline{\mathrm{m}}_{b)})$$

$$+ \frac{1}{2}\{(2(\alpha + \overline{\beta}) + \pi + \overline{\tau} - \nu - \overline{\kappa})v_{(a}\mathrm{m}_{b)} + \text{ c.c.}\} + \frac{1}{\sqrt{2}}\{(\lambda - \overline{\sigma})\mathrm{m}_a \mathrm{m}_b + \text{ c.c.}\};$$

rotation,

$$\omega_{ab} = \frac{1}{2}\{(2(\alpha + \overline{\beta}) + \nu + \overline{\kappa} - \pi - \overline{\tau})v_{[a}\mathrm{m}_{b]} + \text{ c.c.}\} + \frac{1}{\sqrt{2}}(\rho - \overline{\rho} + \mu - \overline{\mu})\mathrm{m}_{[a}\overline{\mathrm{m}}_{b]},$$

where

$$\vec{v} \equiv \frac{1}{\sqrt{2}}(\vec{\ell} - \vec{\mathrm{k}}) \Rightarrow v^a v_a = 1, \quad v^a u_a = 0.$$

[1] R. P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).
[2] R. P. Kerr and A. Schild, *Atti Del Convegno Sulla Relatività Generale: Problemi Dell'Energia E Onde Gravitazionali*, edited by G. Barbéra (Firenze, Italy, 1965), p. 173.
[3] R. P. Kerr and A. Schild, Proc. Symp. Appl. Math. **17**, 199 (1969).
[4] G. C. Debney, R. P. Kerr, and A. Schild, J. Math. Phys. **10**, 1842 (1969).
[5] D. Kramer, H. Stephani, E. Herlt, and M. A. H. MacCallum, *Exact solutions of Einstein's field equations* (Cambridge University Press, Cambridge, 1980).
[6] J. Martín and J. M. M. Senovilla, J. Math. Phys. **27**, 265 (1986).
[7] J. M. M. Senovilla, Ph.D. thesis, Universidad de Salamanca, 1986.
[8] F. Martín-Pascual and J. M. M. Senovilla, J. Math. Phys. **29**, 937 (1988).
[9] J. M. M. Senovilla and C. F. Sopuerta, Class. Quantum Grav. **11**, 2073 (1994).
[10] C. F. Sopuerta, Ph.D. thesis, Universitat de Barcelona, 1996.
[11] A. H. Thompson, Tensor **17**, 92 (1966).
[12] A. H. Taub, Ann. Phys. (N.Y.) **134**, 326 (1981).
[13] A. H. Bilge and M. Gürses, *XI International Colloquium on Group Theoretical Methods in Physics*, edited by M. Serdaroglu and E. Inönü (Springer, Istanbul, Turkey, 1982).
[14] B. C. Xanthopoulos, Ann. Phys. (N.Y.) **149**, 286 (1983).
[15] E. Nahmad-Achar, J. Math. Phys. **29**, 1879 (1988).
[16] P. C. Vaidya, Pramana **8**, 512 (1977).

[17] E. T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).
[18] J. N. Goldberg and R. K. Sachs, Acta Phys. Pol. **22**, 13 (1962).
[19] D. Cox and E. J. Flaherty, Commun. Math. Phys. **47**, 75 (1976).
[20] J. M. Whittaker, Proc. R. Soc. London, Ser. A **306**, 1 (1968).
[21] H. D. Wahlquist, Phys. Rev. **172**, 1291 (1968).
[22] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[23] G. C. Debney, J. Math. Phys. **15**, 992 (1974).
[24] A. Trautman, *Recent Developments in General Relativity* (Pergamon Press, New York, 1962), p. 459.
[25] H. Urbantke, Acta Phys. Austriaca **35**, 396 (1972).
[26] E. Herlt and H. Ch. Hermann, Exp. Tech. Phys. (Berlin) **28**, 97 (1980).
[27] J. M. M. Senovilla, Phys. Lett. A **123**, 214 (1987).
[28] J. Wainwright, J. Math. Phys. **18**, 672 (1977).

Vacuum expectation value asymptotics for second order differential operators on manifolds with boundary

Thomas P. Branson[a)]
Mathematics Department, University of Iowa, Iowa City, Iowa 52242

Peter B. Gilkey[b)]
Mathematics Department, University of Oregon, Eugene, Oregon 97403

Dmitri V. Vassilevich[c)]
Department of Theoretical Physics, St. Petersburg University, 198904 St. Petersburg, Russia

(Received 12 May 1997; accepted for publication 29 October 1997)

Let M be a compact Riemannian manifold with smooth boundary. We study the vacuum expectation value of an operator Q by studying $\mathrm{Tr}_{L^2}Qe^{-tD}$, where D is an operator of Laplace type on M, and where Q is a second order operator with scalar leading symbol; we impose Dirichlet or modified Neumann boundary conditions. © *1998 American Institute of Physics.* [S0022-2488(98)03102-8]

I. INTRODUCTION

Let M be a compact smooth Riemannian manifold of dimension m with smooth boundary ∂M. We say that a second order operator D on the space of smooth sections $C^\infty(V)$ of a smooth vector bundle over M has *scalar leading symbol* if the leading symbol is $h^{ij}I_V\xi_i\xi_j$ for some symmetric 2-tensor h. We say that D is of *Laplace type* if h^{ij} is the metric tensor on the cotangent bundle. Let D be an operator of Laplace type. If the boundary of M is non-empty, we impose Dirichlet or Neumann boundary conditions $\mathscr{B}$ to define the operator $D_{\mathscr{B}}$, see Sec. IV for further details. Let Q be an auxiliary second order partial differential operator on V with scalar leading symbol; if the order of Q is at most 1, then this hypothesis is satisfied trivially. As $t\downarrow 0$, there is an asymptotic expansion

$$\mathrm{Tr}_{L^2}(Qe^{-tD_{\mathscr{B}}})\sim\sum_{n=-2}^{\infty}a_n(Q,D,\mathscr{B})t^{(n-m)/2}$$

[see Ref. 1 (Lemma 1.9.1)] where a different numbering convention was used. The invariants $a_n(Q,D,\mathscr{B})$ are locally computable. We have $a_{-2}(Q,D,\mathscr{B})=0$ and $a_{-1}(Q,D,\mathscr{B})=0$ if Q has order at most 1. If the boundary of M is empty, the boundary condition $\mathscr{B}$ plays no role and we drop it from the notation; in this case, if n is odd, then $a_n(Q,D)=0$.

Our paper is motivated by several physical examples. First, consider a Euclidean quantum field theory with a propagator D^{-1} depending on external fields. Typically, D is a second order differential operator of Laplace type. In the one-loop approximation, the vacuum expectation value $\langle Q\rangle$ of a second order differential operator is given by $\langle Q\rangle=\mathrm{Tr}_{L^2}(QD^{-1})$. By formal manipulations, this can be represented in the form

$$\langle Q\rangle\sim\int_0^\infty dt\ \mathrm{Tr}_{L^2}(Qe^{-tD})\sim\int_0^\infty dt\sum_{n=-2}^{\infty}a_n(Q,D,\mathscr{B})t^{(n-m)/2}.$$

These integrals are divergent at the lower limit and need to be regularized. This can be done by replacing 0 by $1/\Lambda$ in the limits of integration; Λ is called the ultraviolet cutoff parameter. The coefficients $a_n(Q,D,\mathscr{B})$ define the asymptotics of $\langle Q\rangle$ as $\Lambda\to\infty$. The first m terms are divergent

a)Electronic mail: branson@math.uiowa.edu
b)Electronic mail: gilkey@math.uoregon.edu
c)Electronic mail: vasilev@snoopy.niif.spb.su

 © 1998 American Institute of Physics

and are essential for renormalization. The coefficients $a_n(Q,D,\mathscr{B})$ also define large mass asymptotics of $\langle Q\rangle$ in the theory of a massive quantum field; for details see, for example, Ref. 2.

A second example is provided by quantum anomalies. In the Fujikawa approach,[3] the anomaly $\mathscr{A}$ is defined as $\mathscr{A}=\lim_{\Lambda\to\infty}\operatorname{Tr}(Qe^{-D/\Lambda^2})$, where Q is the generator of an anomalous symmetry transformation, and D is a regulator. Usually divergent terms may be absorbed in renormalization and one has that $\mathscr{A}\sim a_{m-2}(Q,D,\mathscr{B})$. Other examples where these asymptotics arise naturally are the study of the anomaly for an arbitrary local symmetry transformation, and in the study of the vacuum expectation value of the stress-energy tensor.

In this paper, we will study the asymptotics $a_n(Q,D,\mathscr{B})$ in a general mathematical framework. In Sec. II, we review the geometry of operators of Laplace type and derive some variational formulas. The operator D determines the metric g, a connection ∇ on V, and an endomorphism E of V. Conversely, given these data, we can define an operator of Laplace type $D(g,\nabla,E)$; see Lemma II.1 for details. Let q_2 be a symmetric 2-tensor and let q_1 be a 1-form valued endomorphism of V. We shall use q_2 to define a variation of the metric $g(\varepsilon):=g+\varepsilon q_2$ and we shall use q_1 to define a variation of the connection $\nabla(\varepsilon):=\nabla+\varepsilon q_1$. Let $Q_2:=\partial_\varepsilon D(g(\varepsilon),\nabla,E)\big|_{\varepsilon=0}$ and let $Q_1:=\partial_\varepsilon D(g,\nabla(\varepsilon),E)\big|_{\varepsilon=0}$. Let Q be a second order operator with scalar leading symbol. We may decompose $Q=Q_2+Q_1+Q_0$ for $Q_0\in\operatorname{End}(V)$ and for Q_2 and Q_1 defined by suitably chosen q_2 and q_1. Since $a_n(Q,D)=\Sigma_i a_n(Q_i,D)$, it suffices to compute the $a_n(Q_i,D)$. In Lemma II.2, we will study the operators Q_2 and Q_1 and show that $\partial_\varrho a_{n+2}(1,D(\varrho),\mathscr{B})=-a_n(\partial_\varrho D(\varrho),D(\varrho),\mathscr{B})$ for any 1-parameter family of operators of Laplace type and fixed boundary condition $\mathscr{B}$. In Secs. III and IV we use this variational formula and apply results of Refs. 4 and 5 to study the invariants $a_n(Q,D,\mathscr{B})$; in Sec. III we consider manifolds without boundary and in Sec. IV we consider manifolds with boundary.

An operator A is said to be of Dirac type if A^2 is of Laplace type. Branson and Gilkey[4] studied the asymptotics of $\operatorname{Tr}_{L^2}(Ae^{-tA^2})$ for an operator A of Dirac type on a closed manifold. In Sec. V, we use the results of Sec. III to rederive these results and to compute some additional terms in the asymptotic expansion. The numbering convention we shall use in this paper differs from that used in Ref. 4; the invariants $a_n(A,A^2)$ of this paper were denoted by $a_{n-1}(A,A^2)$ in Ref. 4.

II. GEOMETRY OF OPERATORS OF LAPLACE TYPE

We adopt the following notational conventions. Greek indices μ, ν, etc. will range from 1 through $m=\dim(M)$ and index local coordinate frames ∂_ν and dx^ν for the tangent and cotangent bundles TM and T^*M. Roman indices i, j will also range from 1 through m and index local orthonormal frames e_i and e^i for TM and T^*M. We shall suppress the bundle indices for tensors arising from V. We adopt the Einstein convention and sum over repeated indices. Let D be an operator of Laplace type. This means that we can decompose D locally in the form

$$D=-(g^{\nu\mu}I_V\partial_\nu\partial_\mu+a^\sigma\partial_\sigma+b), \tag{2.1}$$

where a and b are local sections of $TM\otimes\operatorname{End}(V)$ and $\operatorname{End}(V)$, respectively. It is important to have a more invariant expression than that which is given in equation (2.1). Let Γ be the Christoffel symbols of the Levi-Civita connection of the metric g on M, let ∇ be an auxiliary connection on V, and let $E\in C^\infty(\operatorname{End}(V))$. Define

$$\begin{aligned}D(g,\nabla,E)&:=-(\operatorname{Tr}_g\nabla^2+E)\\&=-g^{\mu\sigma}\{I_V\partial_\mu\partial_\sigma+2\omega_\mu\partial_\sigma-\Gamma_{\mu\sigma}{}^\nu I_V\partial_\nu+\partial_\mu\omega_\sigma+\omega_\mu\omega_\sigma-\Gamma_{\mu\sigma}{}^\nu\omega_\nu\}-E.\end{aligned} \tag{2.2}$$

We compare equations (2.1) and (2.2) to prove the following Lemma.

Lemma II.1: If D is an operator of Laplace type, then there exists a unique connection ∇ on V and a unique endomorphism E of V so that $D=D(g,\nabla,E)$.

(1) If ω is the connection 1-form of ∇, then $\omega_\delta=g_{\nu\delta}(a^\nu+g^{\mu\sigma}\Gamma_{\mu\sigma}{}^\nu I_V)/2$.
(2) We have $E=b-g^{\nu\mu}(\partial_\mu\omega_\nu+\omega_\nu\omega_\mu-\omega_\sigma\Gamma_{\nu\mu}{}^\sigma)$.

Let $D=D(g,\nabla,E)$. We use the Levi-Civita connection of the metric g and the connection ∇ on V to covariantly differentiate tensors of all types. We shall let ';' denote multiple covariant differentiation. Thus, for example, $Df=-(f_{;kk}+Ef)$.

Lemma II.2: Let $D=D(g,\nabla,E)$ be an operator of Laplace type. Let $q_2=q_{2,ij}$ be a symmetric 2-tensor and let $q_1=q_{1,i}$ be an endomorphism valued 1-tensor. Then

(1) $Q_1f:=\partial_\varepsilon D(g,\nabla+\varepsilon q_1,E)f|_{\varepsilon=0}=-2q_{1,i}f_{;i}-q_{1,i;i}f$.
(2) $Q_2f:=\partial_\varepsilon D(g+\varepsilon q_2,\nabla,E)f|_{\varepsilon=0}=q_{2,ij}f_{;ij}+(2q_{2,ij;j}-q_{2,jj;i})f_{;i}/2$.
(3) *Let $D(\varrho)$ be a smooth 1-parameter family of operators of Laplace type and fix $\mathscr{B}$. Then* $a_n(\partial_\varrho D(\varrho),D(\varrho),\mathscr{B})=-\partial_\varrho a_{n+2}(1,D(\varrho),\mathscr{B})$.

Proof: Fix a point $x_0\in M$; we may assume that x_0 is in the interior of M. Choose coordinates centered at x_0 and a local frame for V so that $g_{\nu\mu}(x_0)=\delta_{\nu\mu}$, $\Gamma(x_0)=0$, and so that $\omega(x_0)=0$. We use equation (2.2) to compute

$$\partial_\varepsilon D(g,\nabla+\varepsilon q_1,E)(x_0)|_{\varepsilon=0}=-g^{\nu\sigma}(2q_{1,\nu}\partial_\sigma+\partial_\nu q_{1,\sigma})(x_0)$$

$$=(-2q_{1,i}\nabla_i-q_{1,i;i})(x_0),$$

$$\partial_\varepsilon\{g+\varepsilon q_2\}^{\nu\sigma}(x_0)|_{\varepsilon=0}=-q_{2,\nu\sigma}(x_0),$$

$$2\partial_\varepsilon\Gamma(g+\varepsilon q_2)_{\nu\sigma}{}^\mu(x_0)|_{\varepsilon=0}=(q_{2,\mu\nu;\sigma}+q_{2,\mu\sigma;\nu}-q_{2,\nu\sigma;\mu})(x_0),$$

$$\partial_\varepsilon D(g+\varepsilon q_2,\nabla,E)(x_0)|_{\varepsilon=0}=(q_{2,\nu\sigma}(\partial_\nu\partial_\sigma+\partial_\nu\omega_\sigma)+\partial_\varepsilon\Gamma(g+\varepsilon q_2)_{\mu\mu}{}^\nu\partial_\nu)(x_0)|_{\varepsilon=0}.$$

The first two assertions now follow. We use Ref. 1 (Lemma 1.9.3) to see that the asymptotic series of the variation is the variation of the asymptotic series. We equate coefficients in the following two asymptotic expansions to complete the proof:

$$\sum_n \partial_\varrho a_n(1,D(\varrho),\mathscr{B})t^{(n-m)/2}\sim\partial_\varrho\mathrm{Tr}_{L^2}(e^{-tD(\varrho)\mathscr{B}})$$

$$=\mathrm{Tr}_{L^2}(-t\partial_\varrho D(\varrho)e^{-tD(\varrho)\mathscr{B}})$$

$$\sim-\sum_k a_k(\partial_\varrho D(\varrho),D(\varrho),\mathscr{B})t^{(k+2-m)/2}.\qquad\square$$

To use Lemma II.2, we shall need some variational formulas. Let $R_{\mu\nu\sigma}{}^\delta$ be the curvature of the Levi-Civita connection with the sign convention that the Ricci tensor is given by $\rho_{\nu\sigma}:=R_{\mu\nu\sigma}{}^\mu$ and the scalar curvature is given by $\tau:=g^{\nu\sigma}\rho_{\nu\sigma}$. Let $\Delta_0=\delta d$ be the scalar Laplacian, let $d\mathrm{vol}$ be the Riemannian measure, and let $F_{\mu\nu}$ be the curvature tensor of ∇.

Lemma II.3: Let $\nabla(\varepsilon):=\nabla+\varepsilon q_1$ and let $g(\varrho):=g+\varrho q_2$. Let $\mathscr{T}_{ij}:=(\partial_\varepsilon|_{\varepsilon=0}F)_{ij}$, let $\mathscr{R}_{\mu\nu\sigma}{}^\delta:=(\partial_\varrho|_{\varrho=0}R)_{\mu\nu\sigma}{}^\delta$, and let $\mathscr{D}:=\partial_\varrho|_{\varrho=0}\Delta_0$. Then

(1) $\mathscr{T}_{ij}=q_{1,j;i}-q_{1,i;j}$ *and* $\partial_\varepsilon|_{\varepsilon=0}(\nabla F)_{ij;k}=\mathscr{T}_{ij;k}+[q_{1,k},\mathscr{T}_{ij}]$.
(2) $\mathscr{D}f=q_{2,ij}f_{;ij}+(2q_{2,ij;j}-q_{2,jj;i})f_{;i}/2$.
(3) $\partial_\varrho d\mathrm{vol}|_{\varrho=0}=q_{2,ii}d\mathrm{vol}/2$.
(4) $(\partial_\varrho\Gamma)_{\mu\nu}{}^\sigma|_{\varepsilon=0}=g^{\sigma\gamma}(q_{2,\mu\gamma;\nu}+q_{2,\nu\gamma;\mu}-q_{2,\mu\nu;\gamma})/2$.
(5) $\mathscr{R}_{\mu\nu\sigma}{}^\delta=g^{\delta\gamma}(q_{2,\mu\sigma;\gamma\nu}+q_{2,\nu\gamma;\sigma\mu}-q_{2,\mu\gamma;\sigma\nu}-q_{2,\nu\sigma;\gamma\mu}+q_{2,\sigma\beta}R_{\mu\nu\gamma}{}^\beta+q_{2,\gamma\beta}R_{\mu\nu\sigma}{}^\beta)/2$.
(6) $\partial_\varrho|_{\varrho=0}\tau=-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik}$.
(7) $\partial_\varrho|_{\varrho=0}|\rho|^2=2\mathscr{R}_{kijk}\rho_{ij}-2q_{2,ij}\rho_{ik}\rho_{jk}$.
(8) $\partial_\varrho|_{\varrho=0}|R|^2=2\mathscr{R}_{ijkl}R_{ijkl}-2q_{2,jn}R_{ijkl}R_{inkl}$.

Proof: The assertion (1) is immediate from the definition. Assertion (2) follows from Lemma II.2. Assertions (3) and (4) are straightforward calculations. Assertion (5) follows from assertion (4) and from the identity

$$q_{2,\sigma\gamma;\nu\mu}-q_{2,\sigma\gamma;\mu\nu}=q_{2,\gamma\rho}R_{\mu\nu\sigma}{}^\rho+q_{2,\sigma\rho}R_{\mu\nu\gamma}{}^\rho.$$

Raising and lowering indices does not commute with varying the metric so we emphasize that the tensor $\mathscr{R}$ is the variation of a tensor of type (3,1). The remaining assertions now follow. □

III. MANIFOLDS WITHOUT BOUNDARY

Lemma II.2 reduces the computation of $a_n(Q,D)$ to the special cases $a_n(Q_i,D)$ where $i=0,1,2$. Recall that $a_n(Q,D)=0$ for n odd; $a_{-2}(Q,D)=0$ if $\mathrm{ord}(Q)\leqslant 1$. If $\mathscr{P}$ is a scalar invariant, let $\mathscr{I}[M]:=\int_M \mathscr{P}(x)d\mathrm{vol}(x)$. Let tr_V be the fiber trace. We refer to Ref. 6 for the proof of the following result:

Theorem III.1: *Let M be a compact Riemannian manifold without boundary, let D be an operator of Laplace type, and let $Q_0\in \mathrm{End}(V)$. Then*

(1) $a_0(Q_0,D)=(4\pi)^{-m/2}\mathrm{tr}_V\{Q_0\}[M]$.
(2) $a_2(Q_0,D)=(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{Q_0(6E+\tau)\}[M]$.
(3) $a_4(Q_0,D)=(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V\{Q_0(60E_{;kk}+60\tau E+180E^2+12\tau_{;kk}+5\tau^2-2|\rho|^2+2|R|^2+30F_{ij}F_{ij})\}[M]$.
(4) $a_6(Q_0,D)=(4\pi)^{-m/2}\mathrm{tr}_V\{Q_0/7!(18\tau_{;iijj}+17\tau_{;k}\tau_{;k}-2\rho_{ij;k}\rho_{ij;k}-4\rho_{jk;n}\rho_{jn;k}+9R_{ijk\ell;n}R_{ijk\ell;n}+28\tau\tau_{;nn}-8\rho_{jk}\rho_{jk;nn}+24\rho_{jk}\rho_{jn;kn}+12R_{ijk\ell}R_{ijk\ell;nn}+35/9\tau^3-14/3\tau\rho^2+14/3\tau R^2-208/9\rho_{jk}\rho_{jn}\rho_{kn}-64/3\rho_{ij}\rho_{k\ell}R_{ikj\ell}-16/3\rho_{jk}R_{jn\ell i}R_{kn\ell i}-44/9R_{ijkn}R_{ij\ell p}R_{kn\ell p}-80/9R_{ijkn}R_{i\ell kp}R_{j\ell np})+360^{-1}Q_0(8F_{ij;k}F_{ij;k}+2F_{ij;j}F_{ik;k}+6F_{ij;kk}F_{ij}+6F_{ij}F_{ij;kk}-12F_{ij}F_{jk}F_{ki}-6R_{ijkn}F_{ij}F_{kn}-4\rho_{jk}F_{jn}F_{kn}+5\tau F_{kn}F_{kn}+6E_{;iijj}+30EE_{;ii}+30E_{;ii}E+30E_{;i}E_{;i}+60E^3+12EF_{ij}F_{ij}+6F_{ij}EF_{ij}+12F_{ij}F_{ij}E+10\tau E_{;kk}+4\rho_{jk}E_{;jk}+12\tau_{;k}E_{;k}-6E_{;j}F_{ij;i}+6F_{ij;i}E_{;j}+30EE\tau+12E\tau_{;kk}+5E\tau^2-2E\rho_{jk}\rho_{jk}+2ER^2)\}[M]$.

Next, we study the invariants $a_n(Q_1,D)$.

Theorem III.2: *Let M be a compact Riemannian manifold without boundary, let D be an operator of Laplace type, and let $Q_1=\partial_\varepsilon D(g,\nabla+\varepsilon q_1,E)|_{\varepsilon=0}$. Then*

(1) $a_0(Q_1,D)=0$.
(2) $a_2(Q_1,D)=-(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V\{60F_{ij}\mathscr{F}_{ij}\}[M]$.
(3) $a_4(Q_1,D)=-(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V\{-8F_{ij;k}\mathscr{F}_{ij;k}-8F_{ij;k}q_{1,k}F_{ij}+8F_{ij;k}F_{ij}q_{1,k}+4F_{ij;j}\mathscr{F}_{ik;k}+4F_{ij;j}q_{1,k}F_{ik}-4F_{ij;j}F_{ik}q_{1,k}-36F_{ij}F_{jk}\mathscr{F}_{ki}-12R_{ijkn}F_{ij}\mathscr{F}_{kn}-8\rho_{jk}F_{jn}\mathscr{F}_{kn}+10\tau F_{kn}\mathscr{F}_{kn}-60E_{;k}q_{1,k}E+60E_{;k}Eq_{1,k}+30EF_{ij}\mathscr{F}_{ij}+30E\mathscr{F}_{ij}F_{ij}\}[M]$.

Proof: We use Lemma II.2 and Theorem III.1. Note that $\Delta_0(f)=-f_{;kk}$ is independent of the connection ∇ for $f\in C^\infty(M)$. We compute the following.

(1) $\partial_\varepsilon \mathrm{tr}_V(60E_{;kk})|_{\varepsilon=0}=60\partial_\varepsilon \mathrm{tr}_V(E)_{;kk}|_{\varepsilon=0}=-60\partial_\varepsilon\Delta_0\mathrm{tr}_V(E)=0$.
(2) $\partial_\varepsilon \mathrm{tr}_V(30F^2)|_{\varepsilon=0}=\mathrm{tr}_V(60F_{ij}\mathscr{F}_{ij})$.
(3) $\partial_\varepsilon \mathrm{tr}_V(8F_{ij;k}F_{ij;k}+6F_{ij;kk}F_{ij}+6F_{ij}F_{ij;kk})|_{\varepsilon=0}$
$=\partial_\varepsilon\{\mathrm{tr}_V(-4F_{ij;k}F_{ij;k})+\mathrm{tr}_V(6F_{ij}F_{ij})_{;kk}\}|_{\varepsilon=0}$
$=\mathrm{tr}_V(-8F_{ij;k}\mathscr{F}_{ij;k}-8F_{ij;k}q_{1,k}F_{ij}+8F_{ij;k}F_{ij}q_{1,k})$.
(4) $\partial_\varepsilon \mathrm{tr}_V(30EE_{;ii}+30E_{;ii}E+30E_{;i}E_{;i})|_{\varepsilon=0}=\partial_\varepsilon\{\mathrm{tr}_V(-30E_{;i}E_{;i})+\mathrm{tr}_V(30EE)_{;ii}\}|_{\varepsilon=0}$
$=\mathrm{tr}_V(-60E_{;k}q_{1,k}E+60E_{;k}Eq_{1,k})$. □

We conclude this section by studying $a_n(Q_2,D)$. The following result is a consequence of Lemma II.2, Lemma II.3, and Theorem III.1. We omit the formula for $a_4(Q_2,D)$ in the interest of brevity.

Theorem III.3: *Let M be a compact Riemannian manifold without boundary, let D be an operator of Laplace type on $C^\infty(V)$, and let $Q_2=\partial_\varepsilon D(g+\varepsilon q_2,\nabla,E)|_{\varepsilon=0}$.*

(1) $a_{-2}(Q_2,D)=-\frac{1}{2}a_0(q_{2,ii},D)$.
(2) $a_0(Q_2,D)=-\frac{1}{2}a_2(q_{2,ii},D)-(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{(-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik})I_V\}[M]$.
(3) $a_2(Q_2,D)=-\frac{1}{2}a_4(q_{2,ii},D)-(4\pi)^{-m/2}360^{-1}\{-\mathscr{D}\mathrm{tr}_V(60E+12\tau I_V)+\mathrm{tr}_V\{10(-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik})\tau I_V-2(2\mathscr{R}_{kijk}\rho_{ij}-2q_{2,ij}\rho_{ik}\rho_{jk})I_V+2(2\mathscr{R}_{ijkl}R_{ijkl}-2q_{2,jn}R_{ijkl}R_{inkl})I_V+60(-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik})E-60q_{2,ik}F_{ij}F_{ik}\}[M]$.

IV. MANIFOLDS WITH BOUNDARY

We now suppose M has smooth non-empty boundary ∂M. Near ∂M, let e_i be a local orthonormal frame for TM where we normalize the choice so that e_m is the inward unit normal. We let indices $a,b,\dots$ range from 1 through $m-1$ and index the resulting orthonormal frame for the tangent bundle $T(\partial M)$ of the boundary. Let $f\in C^\infty(V)$. Let S be an endomorphism of V defined on ∂M. The Neumann boundary operator is defined by the operator $B_S^+ f:=(\nabla_m+S)f|_{\partial M}$ and the Dirichlet boundary operator is defined by the operator $B_S^- f=f|_{\partial M}$; we set $S=0$ with Dirichlet boundary conditions to have a uniform notation. Let $L_{ab}=(\nabla_{e_a}e_b,e_m)$ be the second fundamental form. Let ':' denote multiple covariant differentiation tangentially with respect to the Levi-Civita connection of the metric on the boundary and the connection ∇ on V. The difference between ';' and ':' is given by the second fundamental form. For example, $E_{;a}$ and $E_{:a}$ agree since there are no tangential indices in E to be differentiated while $E_{;ab}=E_{:ab}-L_{ab}E_{;m}$. There are some new features here which are not present in the case of manifolds without boundary in the following formulas. The invariants $a_n(Q,D,\mathscr{B})$ are non-zero for odd n and the normal derivatives of Q_0 enter. We still have $a_{-2}(Q,D,\mathscr{B})=0$ if $\mathrm{ord}(Q)\leqslant 1$.

Theorem IV.1: *Let M be a compact Riemannian manifold with smooth boundary, let D be an operator of Laplace type and let* $Q_0\in\mathrm{End}(V)$. *Then*

(1) $a_0(Q_0,D,\mathscr{B}_S^\pm)=(4\pi)^{-m/2}\mathrm{tr}_V\{Q_0\}[M]$.
(2) $a_1(Q_0,D,\mathscr{B}_S^\pm)=\pm(4\pi)^{-(m-1)/2}4^{-1}\mathrm{tr}_V\{Q_0\}[\partial M]$.
(3) $a_2(Q_0,D,\mathscr{B}_S^\pm)=(4\pi)^{-m/2}6^{-1}\{\mathrm{tr}_V\{Q_0(6E+\tau)\}[M]$
$+\mathrm{tr}_V\{Q_0(2L_{aa}+12S)+(3^+,-3^-)Q_{0;m}\}[\partial M]\}$.
(4) $a_3(Q_0,D,\mathscr{B}_S^\pm)=(4\pi)^{-(m-1)/2}384^{-1}\mathrm{tr}_V\{Q_0((96^+,-96^-)E+(16^+,-16^-)\tau$
$+(8^+,-8^-)R_{amam}+(13^+,-7^-)L_{aa}L_{bb}+(2^+,10^-)L_{ab}L_{ab}$
$+96SL_{aa}+192S^2)+Q_{0;m}((6^+,30^-)L_{aa}+96S)$
$+(24^+,-24^-)Q_{0;mm}\}[\partial M]$.
(5) $a_4(Q_0,D,\mathscr{B}_S^\pm)=(4\pi)^{-m/2}360^{-1}\{\mathrm{tr}_V\{Q_0(60E_{;kk}+60\tau E+180E^2+30F_{ij}F_{ij}+12\tau_{;kk}$
$+5\tau^2-2\rho^2+2R^2)\}[M]+\mathrm{tr}_V\{Q_0(240^+,-120^-)E_{;m}+(42^+,-18^-)\tau_{;m}$
$+24L_{aa:bb}+120EL_{aa}+20\tau L_{aa}+4R_{amam}L_{bb}-12R_{ambm}L_{ab}+4R_{abcb}L_{ac}$
$+360(SE+ES)+21^{-1}\{(280^+,40^-)L_{aa}L_{bb}L_{cc}+(168^+,$
$-264^-)L_{ab}L_{ab}L_{cc}+(224^+,320^-)L_{ab}L_{bc}L_{ac}\}+120S\tau+144SL_{aa}L_{bb}$
$+48SL_{ab}L_{ab}+480S^2L_{aa}+480S^3+120S_{:aa}+Q_{0;m}((180^+,-180^-)E$
$+(30^+,-30^-)\tau+(84^+,-180^-)/7\cdot L_{aa}L_{bb}+(84^+,60^-)/7\cdot L_{ab}L_{ab}$
$+72SL_{aa}+240S^2+Q_{0;mm}(24L_{aa}+120S)+(30^+,-30^-)Q_{0;iim}\}[\partial M]\}$.

Theorem IV.2: *Let M be a compact Riemannian manifold with smooth boundary, let D be an operator of Laplace type, and let* $Q_1=\partial_\varepsilon D(g,\nabla+\varepsilon q_1,E)|_{\varepsilon=0}$. *Then*

(1) $a_{-1}(Q_1,D,\mathscr{B}_S^\pm)=0$.
(2) $a_0(Q_1,D,\mathscr{B}_S^\pm)=-(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V((-12^+,0^-)q_{1,m})[\partial M]$.
(3) $a_1(Q_1,D,\mathscr{B}_S^\pm)=-(384)^{-1}(4\pi)^{-(m-1)/2}\mathrm{tr}_V((-96^+,0^-)q_{1,m}L_{aa}-384Sq_{1,m})[\partial M]$.
(4) $a_2(Q_1,D,\mathscr{B}_S^\pm)=-(4\pi)^{-m/2}360^{-1}\{\mathrm{tr}_V(60F_{ij}\cdot\mathscr{F}_{ij})[M]+\mathrm{tr}_V\{(-720^+,0^-)q_{1,m}E$
$+(-120^+,0^-)q_{1,m}\tau+(-144^+,0^-)q_{1,m}L_{aa}L_{bb}+(-48^+,0^-)q_{1,m}L_{ab}L_{ab}$
$-960Sq_{1,m}L_{aa}-1440S^2q_{1,m}\}[\partial M]\}$.
(5) $a_3(Q_1,D,\mathscr{B}^-)=5760^{-1}(4\pi)^{(m-1)/2}\mathrm{tr}_V\{240F_{ab}\cdot\mathscr{F}_{ab}-720F_{am}\cdot\mathscr{F}_{am}\}[\partial M]$.
(6) *If the boundary of M is totally geodesic, then*
$a_3(Q_1,D,\mathscr{B}_S^\pm)=-5760^{-1}(4\pi)^{(m-1)/2}\mathrm{tr}_V(-1440E_{;m}q_{1,m}+1440(q_{1,m}E-Eq_{1,m})S$
$+240F_{ab}\cdot\mathscr{F}_{ab}-960\tau Sq_{1,m}-240\rho_{mm}Sq_{1,m}+180F_{am}\cdot\mathscr{F}_{am}-270\tau_{;m}q_{1,m}$
$+720S_{:a}q_{1,m:a}+720S_{:a}(q_{1,a}S-Sq_{1,a})-2880E(Sq_{1,m}+q_{1,m}S)$
$-5760q_{1,m}S^3)[\partial M]$.

Proof: If $Q_0=q_0I_V$ for $q_0\in C^\infty(M)$ is a scalar operator, then Theorem IV.1 follows from Ref. 4. If Q_0 is not a scalar operator, we must worry about the lack of commutativity; the only point at which this enters is in the coefficient of $\mathrm{tr}_V(Q_0SE)$ and $\mathrm{tr}_V(Q_0ES)$. We express

$$a_4(Q_0,D,\mathscr{B}_S)=(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V(C_1Q_0SE+C_2Q_0ES)[\partial M]+\text{ other terms;}$$

the sum $C_1+C_2=720$ is determined by the scalar case. If D, Q_0, and S are real, then $\mathrm{Tr}_{L^2}(Q_0e^{-tD})$ is real; this shows that C_1 and C_2 are real. If Q_0, D, and S are self-adjoint, $\mathrm{Tr}_{L^2}(Q_0e^{-tD})$ is real so $\mathrm{tr}_V(Q_0(C_1SE+C_2ES))[\partial M]$ is real; this now shows $C_1=C_2$ and completes the proof of Theorem IV.1.

To keep boundary conditions constant, let $S(\varepsilon):=S-\varepsilon q_{1,m}$ so $\partial_\varepsilon S|_{\varepsilon=0}=-q_{1,m}$. Assertions (1)–(5) of Theorem IV.2 now follow directly from Lemma II.2, from Lemma II.3, and from Theorem IV.1. In Ref. 5, we showed that

$$\begin{aligned}a_5(1,D,\mathscr{B}_S^{\pm})=&\pm 5760^{-1}(4\pi)^{(m-1)/2}\mathrm{tr}_V\{(360E_{;mm}+1440E_{;m}S\\&+720E^2+240E_{:aa}+240\tau E+120F_{ab}F_{ab}+48\tau_{;ii}+20\tau^2-8\rho^2\\&+8R^2-120\rho_{mm}E-20\rho_{mm}\tau+480\tau S^2+(90^+,-360^-)F_{am}F_{am}\\&+12\tau_{;mm}+24\rho_{mm:aa}+15\rho_{mm;mm}+270\tau_{;m}S+120\rho_{mm}S^2\\&+960S_{:aa}S+600S_{:a}S_{:a}+16R_{ammb}\rho_{ab}-17\rho_{mm}\rho_{mm}\\&-10R_{ammb}R_{ammb}+2880ES^2+1440S^4)+\mathscr{E}\}[\partial M].\end{aligned}$$

The variation of the terms other than $\mathscr{E}$ gives rise to the expressions listed in Theorem IV.2 (5) and (6). The remainder term $\mathscr{E}$ is given below. It vanishes if the boundary is totally geodesic and involves 40 undetermined coefficients:

$$\begin{aligned}\mathscr{E}=&d_1^{\pm}L_{aa}E_{;m}+d_2^{\pm}L_{aa}\tau_{;m}+d_3^{\pm}L_{ab}R_{ammb;m}+d_4^{+}L_{aa}S_{:bb}+d_5^{+}L_{ab}S_{:ab}\\&+d_6^{+}L_{aa:b}S_{:b}+d_7^{+}L_{ab:a}S_{:b}+d_8^{+}L_{aa:bb}S+d_9^{+}L_{ab:ab}S+d_{10}^{\pm}L_{aa:b}L_{cc:b}\\&+d_{11}^{\pm}L_{ab:a}L_{cc:b}+d_{12}^{\pm}L_{ab:a}L_{bc:c}+d_{13}^{\pm}L_{ab:c}L_{ab:c}+d_{14}^{\pm}L_{ab:c}L_{ac:b}\\&+d_{15}^{\pm}L_{aa:bb}L_{cc}+d_{16}^{\pm}L_{ab:ab}L_{cc}+d_{17}^{\pm}L_{ab:ac}L_{bc}+d_{18}^{\pm}L_{aa:bc}L_{bc}+d_{19}^{\pm}L_{bc:aa}L_{bc}\\&+1440^{+}L_{aa}SE+d_{20}^{+}L_{aa}S\rho_{mm}+240^{+}L_{aa}S\tau+d_{21}^{+}L_{ab}S\rho_{ab}\\&+d_{22}^{+}L_{ab}SR_{mabm}+(195^{+},105^{-})L_{aa}L_{bb}E+(30^{+},150^{-})L_{ab}L_{ab}E\\&+(195^{+}/6,105^{-}/6)L_{aa}L_{bb}\tau+(5^{+},25^{-})L_{ab}L_{ab}\tau+d_{23}^{\pm}L_{aa}L_{bb}\rho_{mm}\\&+d_{24}^{\pm}L_{ab}L_{ab}\rho_{mm}+d_{25}^{\pm}L_{aa}L_{bc}\rho_{bc}+d_{26}^{\pm}L_{aa}L_{bc}R_{mbcm}+d_{27}^{\pm}L_{ab}L_{ac}\rho_{bc}\\&+d_{28}^{\pm}L_{ab}L_{ac}R_{mbcm}+d_{29}^{\pm}L_{ab}L_{cd}R_{acbd}+d_{30}^{+}L_{aa}S^3+d_{31}^{+}L_{aa}L_{bb}S^2\\&+d_{32}^{+}L_{ab}L_{ab}S^2+d_{33}^{+}L_{aa}L_{bb}L_{cc}S+d_{34}^{+}L_{aa}L_{bc}L_{bc}S+d_{35}^{+}L_{ab}L_{bc}L_{ac}S\\&+d_{36}^{\pm}L_{aa}L_{bb}L_{cc}L_{dd}+d_{37}^{\pm}L_{aa}L_{bb}L_{cd}L_{cd}+d_{38}^{\pm}L_{ab}L_{ab}L_{cd}L_{cd}\\&+d_{39}^{\pm}L_{aa}L_{bc}L_{cd}L_{db}+d_{40}^{\pm}L_{ab}L_{bc}L_{cd}L_{da}.\end{aligned}$$

The variation of $\mathscr{E}$ is zero for Dirichlet boundary conditions or if the boundary is totally geodesic. □

To study $a_n(Q_2,D,\mathscr{B}_S^{\pm})$ we need some additional formulas.

Lemma IV.3:

(1) *Let* $g(\varrho):=g+\varrho q_2$, $\mathscr{N}:=\partial_\varrho|_{\varrho=0}e_m(\varrho)$, and $\mathscr{L}_{\alpha\beta}:=\partial_\varrho|_{\varrho=0}L_{\alpha\beta}$. *Then*
   (a) $\mathscr{N}=-q_{2,am}e_a-q_{2,mm}e_m/2$.
   (b) $\mathscr{L}_{ab}=(q_{2,am;b}+q_{2,bm;a}-q_{2,ab;m}-q_{2,mm}L_{ab})/2$.
(2) $q_{2,am;a}=q_{2,am:a}-L_{aa}q_{2,mm}+L_{ab}q_{2,ab}$.
(3) $\mathscr{R}_{kiik}=q_{2,ki;ki}-q_{2,ii;kk}$.

Proof: Let $1\leqslant\alpha,\beta\leqslant m-1$. Let $y=(y^\alpha)$ be local coordinates on the boundary ∂M centered at y_0. We suppose $g_{\alpha\beta}(y_0)=\delta_{\alpha\beta}+O(|y|^2)$. Introduce coordinates $x=(y,x^m)$ so the curves $t\mapsto(y,t)$ are unit speed geodesics perpendicular to the boundary. Then $g_{mm}=1$ and $g_{\alpha m}=0$; ∂_m is the

inward geodesic normal vector field for g. Let $N(\varrho)$ be the inward geodesic normal vector field for the metric $g(\varrho)$. Expand $N(\varrho)(y_0)=\partial_m+\varrho(c^m\partial_m+c^\beta\partial_\beta)+O(\varrho^2)$. We prove the first assertion by solving the equations

$$0=g(\varrho)(N(\varrho),\partial_\alpha)(y_0)=\varrho(c^\alpha+q_{2,\alpha m})(y_0)+O(\varrho^2),$$

$$1=g(\varrho)(N(\varrho),N(\varrho))(y_0)=1+\varrho(2c^m+q_{2,mm})(y_0)+O(\varrho^2)$$

to see $c^\alpha(y_0)=-q_{2,m\alpha}(y_0)$ and $c^m(y_0)=-q_{2,mm}/2$. We use Lemma II.3 to compute the variation of the Christoffel symbols and complete the proof by computing:

$$L_{\alpha\beta}=\Gamma(\rho)_{\alpha\beta}{}^i g(N(\varrho),\partial_i),$$

$$\mathscr{L}_{\alpha\beta}(y_0)=(\dot\Gamma_{\alpha\beta}{}^m-q_{2,mm}L_{\alpha\beta}/2)(y_0).$$

The second assertion is immediate, the third follows from Lemma II.3. □

Dirichlet boundary conditions are unchanged by a variation of the metric g. The following result follows from Lemma II.2, Theorem IV.1, and Lemma IV.3. We omit the formula for a_2 in the interest of brevity.

Theorem IV.4: *Let M be a compact Riemannian manifold with smooth boundary, let D be an operator of Laplace type, and let $Q_2:=\partial_\varepsilon|_{\varepsilon=0}D(g+\varepsilon q_2,\nabla,E)$. Then*

(1) $a_{-2}(Q_2,D,\mathscr{B}^-)=-(4\pi)^{-m/2}\mathrm{tr}_V\{q_{2,ii}/2\}[M]$.
(2) $a_{-1}(Q_2,D,\mathscr{B}^-)=(4\pi)^{-(m-1)/2}4^{-1}\mathrm{tr}_V\{q_{2,aa}/2\}[\partial M]$.
(3) $a_0(Q_2,D,\mathscr{B}^-)=-(4\pi)^{-m/2}6^{-1}\{\mathrm{tr}_V\{q_{2,ii}(6E+\tau)/2-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik}\}[M]$
$+\mathrm{tr}_V\{q_{2,aa}L_{aa}+2\mathscr{L}_{aa}-2q_{2,ab}L_{ab}\}[\partial M]\}$.
(4) $a_1(Q_2,D,\mathscr{B}^-)=(4\pi)^{-(m-1)/2}384^{-1}\mathrm{tr}_V\{q_{2,aa}(96E+16\tau+8R_{amam}+7L_{aa}L_{bb}-10L_{ab}L_{ab})/2$
$+16(-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik})+8(\mathscr{R}_{amam}-q_{2,ab}R_{ambm}+2R_{ama\mathscr{N}})$
$+7(2\mathscr{L}_{aa}L_{bb}-2q_{2,ab}L_{ab}L_{cc})-10(2\mathscr{L}_{ab}L_{ab}-2q_{2,ac}L_{ab}L_{cb})\}[\partial M]$.
(5) *Suppose for simplicity that the metric g is flat, i.e., that $R_{ijkl}=0$. Then*
$a_2(Q_2,D,\mathscr{B}^-)=-(4\pi)^{-m/2}360^{-1}\{-\mathscr{D}\mathrm{tr}_V(60E)+\mathrm{tr}_V(60\mathscr{R}_{ijji}E+30q_{2,ii}E_{;kk}+90q_{2,ii}E^2$
$+15q_{2,ii}\Omega^2)\}[M]-(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V\{60E_{;m}q_{2,mm}+120E_{;a}q_{2,am}$
$-18\mathscr{R}_{ikki;m}+L_{aa}(20\mathscr{R}_{ikki}+4\mathscr{R}_{bmbm})+12\mathscr{R}_{ambm}L_{ab}+4\mathscr{R}_{abcb}L_{ac}$
$+q_{2,dd}(-60E_{;m}+60EL_{aa}+12L_{aa:bb}+20/21L_{aa}L_{bb}L_{cc}+44/7L_{ab}L_{ab}L_{cc}$
$+160/21L_{ab}L_{bc}L_{ac})+(120E+40/7L_{aa}L_{bb}+88/7L_{ab}L_{ab})(\mathscr{L}_{cc}-q_{2,cd}L_{cd})$
$+88/7(2\mathscr{L}_{ab}L_{ab}L_{cc}-2q_{2,bd}L_{ab}L_{ad}L_{cc})+320/7(\mathscr{L}_{ab}L_{bc}L_{ca}$
$-q_{2,ad}L_{ab}L_{bc}L_{cd})+12L_{bb:c}q_{2,aa:c}\}[\partial M]$.

We now study Neumann boundary conditions. The situation is quite different as Neumann boundary conditions are **not** invariant under general perturbations of the metric; if $q_{am}\neq 0$ on ∂M, $\mathscr{B}_S^+(\varrho)$ will involve tangential derivatives regardless of how S is varied. Thus Lemma II.2 is not directly applicable. Nevertheless, we can still compute the first three terms in the asymptotic expansion.

Theorem IV.5: *Let M be a compact Riemannian manifold with smooth boundary, let D be an operator of Laplace type, and let $Q_2:=\partial_\varepsilon D(g+\varepsilon q_2,\nabla,E)|_{\varepsilon=0}$. Then*

(1) $a_{-2}(Q_2,D,\mathscr{B}_S^+)=-(4\pi)^{-m/2}\mathrm{tr}_V\{q_{2,ii}/2\}[M]$.
(2) $a_{-1}(Q_2,D,\mathscr{B}_S^+)=-(4\pi)^{-(m-1)/2}4^{-1}\mathrm{tr}_V\{q_{2,aa}/2\}[\partial M]$.
(3) $a_0(Q_2,D,\mathscr{B}_S^+)=-(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{q_{2,ii}(6E+\tau)/2-q_{2,ij}\rho_{ij}\}[M]$
$-(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{q_{2,aa}L_{bb}-q_{2,ab}L_{ab}-6q_{2,mm}S+6q_{2,aa}\}[\partial M]$.

Proof: Define $\mathrm{ord}(q_{2,ij})=0$, $\mathrm{ord}(E)=2$, $\mathrm{ord}(R_{ijkl})=2$, $\mathrm{ord}(F)=2$, $\mathrm{ord}(L)=1$, and $\mathrm{ord}(S)=1$. Increase the order by 1 for each explicit covariant derivative which is present. Dimensional analysis then shows the interior integrands in the formula for $a_n(Q_2,D,\mathscr{B})$ are homogeneous of order $n+2$ while the boundary integrands are homogeneous of degree $n+1$. We use H. Weyl's theorem to write a spanning set for the set of invariants and express

$$a_{-2}(Q_2,D,\mathscr{B}_S^+)=-(4\pi)^{-m/2}\mathrm{tr}_V(b_1q_{2,ii}/2)[M],\tag{4.1}$$

$$a_{-1}(Q_2,D,\mathscr{B}_S^+)=-(4\pi)^{-(m-1)/2}4^{-1}\mathrm{tr}_V(c_1q_{2,aa}/2+c_2q_{2,mm})[\partial M], \tag{4.2}$$

$$\begin{aligned}a_0(Q_2,D,\mathscr{B}_S^+)=&-(4\pi)^{-m/2}6^{-1}\{\mathrm{tr}_V(q_{2,ii}(6b_2E+b_3\tau)/2-b_4q_{2,ij}\rho_{ij})[M]+\mathrm{tr}_V(c_3q_{2,aa}L_{bb}\\&+c_4q_{2,ab}L_{ab}+c_5q_{2,mm}L_{aa}+c_6q_{2,mm}S+c_7q_{2,aa}S+c_8q_{2,mm;m}\\&+c_9q_{2,aa;m})[\partial M]\}.\end{aligned} \tag{4.3}$$

Product formulas then show the constants are independent of the dimension m; these invariants form a basis for the integral invariants and are uniquely determined for m large. A word of explanation for the formula in equation (4.3) is in order. We can integrate by parts to replace the interior integrals $q_{2,ij;ij}$ and $q_{2,ii;jj}$ by boundary integrals of $q_{2,mm;m}$, $q_{2,am;a}$, and $q_{2,aa;m}$. Since $q_{2,am;a}[\partial M]=0$, we use Lemma IV.3 to omit the variable $q_{2,am;a}$. If we take ∂M empty, the boundary condition plays no role and $\mathscr{R}_{kiik}[M]=0$ (see equation (4.5) below). We use Theorem III.3 to see $b_1=b_2=b_3=b_4=1$; this completes the proof of assertion (1) and the first part of assertion (3).

The $q_{2,am;*}$ variables do not appear in equations (4.1), (4.2), and (4.3). Thus we may take a variation with $q_{2,am}=0$ on ∂M. This is an essential simplification since it means $N(\varrho)=g^{mm}(\varrho)^{1/2}\partial_m$. Thus the boundary conditions do not involve any tangential derivatives. We set $S(\varrho)=g^{mm}(\varrho)^{1/2}S$. Then the boundary condition is preserved; $\nabla_{N(\varrho)}+S(\varrho)=g^{mm}(\varrho)^{1/2}(\nabla_m+S)$. We have $\partial_\varrho S(\varrho)|_{\varrho=0}=-q_{2,mm}S/2$. By Lemma II.2, $a_{-1}(Q_2,D,B_S^+)=-(4\pi)^{-(m-1)/2}4^{-1}\mathrm{tr}_V(q_{2,aa}/2)[\partial M]$. This shows $c_1=1$ and $c_2=0$ and completes the proof of assertion (2).

We use Lemma II.2 and Theorem IV.1 to see

$$\begin{aligned}a_0(Q_2,D,B_S^+)=&-(4\pi)^{-(m-1)/2}6^{-1}\{\mathrm{tr}_V(q_{2,ii}(6E+\tau)/2-q_{2,ij}\rho_{ij}+\mathscr{R}_{kiik})[M]\\&+\mathrm{tr}_V(q_{2,aa}(L_{bb}+6S)-2q_{2,ab}L_{ab}+2\mathscr{L}_{aa}-6q_{2,mm}S)[\partial M]\}.\end{aligned} \tag{4.4}$$

We have that

$$\begin{aligned}\mathscr{R}_{kiik}[M]&=(q_{2,ki;ki}-q_{2,ii;kk})[M]\\&=(-q_{2,am;a}+q_{2,aa;m})[\partial M]\\&=(L_{aa}q_{2,mm}-L_{ab}q_{2,ab}+q_{2,aa;m})[\partial M],\\2\mathscr{L}_{aa}[\partial M]&=(2q_{2,am;a}-q_{2,aa;m}-q_{2,mm}L_{aa})[\partial M]\\&=(-3L_{aa}q_{2,mm}+2L_{ab}q_{2,ab}-q_{2,aa;m})[\partial M].\end{aligned} \tag{4.5}$$

We use equation (4.5) to compare equations (4.3) and (4.4). This shows

$$c_3=1,\quad c_4=-1,\quad c_5=-2,\quad c_6=-6,\quad c_7=6,\quad c_8=0,\quad c_9=0. \qquad \square$$

V. OPERATORS OF DIRAC TYPE

In this section, we study the invariants $a_n(HA,A^2)$, where M is a closed manifold, A is an operator of Dirac type, and H is a smooth endomorphism; we refer to Ref. 7 for a discussion of the case of manifolds with boundary. We begin with a technical result:

Lemma V.1: Let $A=\gamma^\nu\partial_\nu-\psi$ be an operator of Dirac type and let $D=A^2$ be the associated operator of Laplace type. Let $D=D(g,\nabla,E)$ and let $H\in C^\infty\mathrm{End}(V)$.

(1) $\omega_\mu=g_{\nu\mu}(-\gamma^\sigma\partial_\sigma\gamma^\nu+\psi\gamma^\nu+\gamma^\nu\psi+g^{\sigma\rho}\Gamma_{\sigma\rho}{}^\nu)/2$.
(2) *Let $\phi:=\psi+\gamma^\nu\omega_\nu$. Then $A=\gamma^\nu\nabla_\nu-\phi$, and ϕ is invariantly defined.*
(3) $\gamma_{i;j}+\gamma_{j;i}=0$.
(4) $E=-\psi^2+\gamma^\mu\partial_\mu\psi-g^{\nu\mu}(\partial_\mu\omega_\nu+\omega_\nu\omega_\mu-\omega_\sigma\Gamma_{\nu\mu}{}^\sigma)=-\gamma_i\gamma_jF_{ij}/2+\gamma_i\phi_{;i}-\phi^2$.
(5) *Let $q_{1,i}:=-H\gamma_i/2$, and $Q_0:=-H\phi-H_{;i}\gamma_i/2$, then $HA=Q_1+Q_0$.*

Proof: We compute

$$\begin{aligned} D=A^2&=(\gamma^\nu\partial_\nu-\psi)(\gamma^\mu\partial_\mu-\psi)\\ &=\gamma^\nu\gamma^\mu\partial_\nu\partial_\mu+(\gamma^\nu\partial_\nu\gamma^\mu-\gamma^\mu\psi-\psi\gamma^\mu)\partial_\mu+\psi^2-\gamma^\mu\partial_\mu\psi,\end{aligned}$$

$$a^\mu=-\gamma^\nu\partial_\nu\gamma^\mu+\gamma^\mu\psi+\psi\gamma^\mu, \text{ and } b=-\psi^2+\gamma^\mu\partial_\mu\psi.$$

Assertion (1) and the first assertion of (4) follows from Lemma II.1. We prove assertion (2) by computing: $\gamma^i\nabla_i-\phi=\gamma^i\partial_i+\gamma^i\omega_i-\phi=\gamma^i\partial_i-\psi$. We choose a system of coordinates and a local frame so that $\Gamma(x_0)=0$ and so that $\omega(x_0)=0$. Then at x_0, we have

$$\begin{aligned} D&=(\gamma^\nu\nabla_\nu-\phi)(\gamma^\mu\nabla_\mu-\phi)\\ &=(\gamma^\nu\gamma^\mu+\gamma^\mu\gamma^\nu)/2\nabla_\nu\nabla_\mu+(\gamma^\mu\gamma^\nu{}_{;\mu}-\phi\gamma^\nu-\gamma^\nu\phi)\nabla_\nu\\ &\quad-\gamma^\nu\phi_{;\nu}+\gamma^\nu\gamma^\mu\Omega_{\nu\mu}/2+\phi^2\\ &=-g^{\nu\mu}\nabla_\nu\nabla_\mu-E.\end{aligned}$$

We equate coefficients to derive the second part of assertion (4). Choose a coordinate system centered at x_0 so $g_{\mu\nu}=\delta_{\nu\mu}+O(|x|^2)$. We use Ref. 8 (Lemma 1.2) to see that we can choose a local frame for V so $\partial_\mu\gamma^\nu(x_0)=0$. Then we have that $\omega_\mu(x_0)=(\psi\gamma_\mu+\gamma_\mu\psi)(x_0)/2$. We prove assertion (3) by computing at x_0

$$\begin{aligned} 2(\gamma_{\nu;\mu}+\gamma_{\mu;\nu})&=[\psi\gamma_\mu+\gamma_\mu\psi,\gamma_\nu]+[\psi\gamma_\nu+\gamma_\nu\psi,\gamma_\mu]\\ &=\psi\gamma_\mu\gamma_\nu+\gamma_\mu\psi\gamma_\nu-\gamma_\nu\psi\gamma_\mu-\gamma_\nu\gamma_\mu\psi+\psi\gamma_\nu\gamma_\mu+\gamma_\nu\psi\gamma_\mu\\ &\quad-\gamma_\mu\psi\gamma_\nu-\gamma_\mu\gamma_\nu\psi\\ &=\psi\gamma_\mu\gamma_\nu+\psi\gamma_\nu\gamma_\mu-\gamma_\mu\gamma_\nu\psi-\gamma_\mu\gamma_\nu\psi\\ &=\psi\delta_{\mu\nu}-\delta_{\mu\nu}\psi=0.\end{aligned}$$

If we set $q_{1,i}=-H\gamma_i/2$, then $Q_1=H\gamma_i\nabla_i+H_{;i}\gamma_i/2$ by Lemma II.2 since $\gamma_{i;i}=0$. We must therefore take $Q_0=-H\phi-H_{;i}\gamma_i/2$. □

The following theorem now follows from Theorem III.1, Theorem III.2, and from Lemma III.3.

Theorem V.2: *Let $A=\gamma_\nu\partial_\nu-\psi$ be an operator of Dirac type on $C^\infty(V)$ over a closed manifold M. Adopt the notation of Lemma V.1.*

(1) $a_0(HA,A^2)=(4\pi)^{-m/2}\mathrm{tr}_V\{Q_0\}[M]$.
(2) $a_2(HA,A^2)=(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{Q_0(\tau+6E)-F_{ij}\mathscr{F}_{ij}\}[M]$.
(3) $a_4(HA,A^2)=(4\pi)^{-m/2}360^{-1}\mathrm{tr}_V\{Q_0(60E_{;kk}+60\tau E+180E^2+12\tau_{;kk}+5\tau^2-2|\rho|^2+2|R|^2$
$+30F_{ij}F_{ij})+8F_{ij;k}\mathscr{F}_{ij;k}+8F_{ij;k}q_{1,k}F_{ij}-8F_{ij;k}F_{ij}q_{1,k}-4F_{ij;j}\mathscr{F}_{ik;k}$
$-4F_{ij;j}q_{1,k}F_{ik}+4F_{ij;j}F_{ik}q_{1,k}+36F_{ij}F_{jk}\mathscr{F}_{ki}+12R_{ijkn}F_{ij}\mathscr{F}_{kn}+8\rho_{jk}F_{jn}\mathscr{F}_{kn}$
$-10\tau F_{kn}\mathscr{F}_{kn}+60E_{;k}q_{1,k}E-60E_{;k}Eq_{1,k}-30EF_{ij}\mathscr{F}_{ij}-30E\mathscr{F}_{ij}F_{ij}\}[M]$.

The first two authors Ref. 8 (Theorem 2.7) computed a_0 and a_2 for f scalar and showed that

$$a_0(fA,A^2)=-(4\pi)^{-m/2}\mathrm{tr}_V(f\phi)[M], \tag{5.1}$$

and

$$a_2(fA,A^2)=-(4\pi)^{-m/2}6^{-1}\mathrm{tr}_V\{f(\phi\tau+6\phi E-F_{ij;j}\gamma_i)\}[M]. \tag{5.2}$$

For scalar f, $\mathrm{Tr}_{L^2}(f_{;i}\gamma_ie^{-tA^2})=\mathrm{Tr}_{L^2}(Afe^{-tA^2}-fAe^{-tA^2})=0$ so $a_n(f_{;i}\gamma_i,A^2)=0$ for all n and we may replace Q_0 by $-f\phi$ in performing our computations. It then follows that Theorem V.2(1) agrees with equation (5.1). We see Theorem V.2 (2) agrees with equation (5.2) by using Lemma II.3 and integrating by parts:

$$-\mathrm{tr}_V(F_{ij}\mathscr{F}_{ij})[M]=\mathrm{tr}_V((f\gamma_j)_{;i}-(f\gamma_i)_{;j})F_{ij}[M]/2$$
$$=-\mathrm{tr}_V(f\gamma_jF_{ij;i}[M])=\mathrm{tr}_V(f\gamma_iF_{ij;j})[M].$$

ACKNOWLEDGMENTS

T.P.B. was partially supported by an international travel grant of the NSF (USA); P.B.G. was partially supported by the NSF (USA) and MPIM (Germany); and D.V.V. was partially supported by GRACENAS (Russia) and DAAD (Germany).

[1] P. Gilkey, *Invariance Theory, the Heat Equation, and the Atiyah-Singer Index Theorem*, 2nd ed. (CRC, Boca Raton, FL, 1994) (ISBN 0-8493-7874-4).
[2] N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge U. P., Cambridge, 1982).
[3] K. Fujikawa, ''Path integral measure for gauge-invariant fermion theories,'' Phys. Rev. Lett. **42**, 1195–1198 (1979).
[4] T. Branson and P. Gilkey, ''The asymptotics of the Laplacian on a manifold with boundary,'' Comm. PDE **15**, 245–272 (1990).
[5] T. Branson, P. Gilkey, and D. V. Vassilevich, ''The asymptotics of the Laplacian on a manifold with boundary II,'' Boll. Unione Mate. Ital. **7**, Suppl. fasc. 2 39–67 (1997).
[6] P. Gilkey, ''The spectral geometry of a Riemannian manifold,'' J. Diff. Geom. **10**, 601–618 (1975).
[7] T. Branson and P. Gilkey, ''Residues of the eta function for an operator of Dirac type with local boundary conditions,'' Diff. Geom. Appl. **2**, 249–269 (1992).
[8] T. Branson and P. Gilkey, ''Residues of the eta function for an operator of Dirac type,'' J. Funct. Anal. **108**, 47–87 (1992).

Breakdown of smooth solutions of the three-dimensional Euler–Poisson system

Uwe Brauer
Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85 740 Garching bei München, Federal Republic of Germany, and Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Avenida Complutense S/N, 28 040 Madrid, Spain

(Received 29 October 1996; accepted for publication 8 October 1997)

Results concerning the occurrence of (kinematical) singularities obtained by Majda *et al.* [Commun. Math. Phys. **94**, 61–66 (1984)] for the incompressible Euler equations and of Chemin [Commun. Math. Phys. **133**, 323–329 (1990)] for the compressible Euler equations are generalized for the compressible Euler–Poisson system. This generalization is applied to two situations of physical interest, namely, either the evolution of a compact body with a freely falling boundary or a cosmological solution with finite, spatially periodic, deviations of a Newtonian, Friedman-like cosmological model. Both situations are briefly reviewed. For the compact body the solutions belong to a special class, introduced by Makino [*Patterns and Waves* (North-Holland, Amsterdam, 1986), pp. 459–479]. In Sec. III, uniqueness is shown for these and therewith one of the severe disadvantages of these solution is eliminated. In both situations the qualitative behavior is similar to the gravitation free case in the sense that only some of the kinematical quantities of the fluid and the gradient of the matter variable diverge; in other words, no specific ''gravitation singularity'' appears. The differences between the two situations considered here is that, for technical reasons, a nonlinear function $w=M(\varrho)$ has to be introduced as a new matter variable for the compact body. Because ϱ has compact support the blow-up of grad w in the L^{∞}-norm implies two possibilities, one being that the singularity is in the interior of the body. In that case the blow-up of grad w implies the blow-up of grad ϱ. If, on the other hand, the singularity is near the boundary of the body, then no precise information is available. © *1998 American Institute of Physics.* [S0022-2488(98)03401-X]

I. INTRODUCTION

One of the most important problems in the relativistic theory of gravitation is the so-called cosmic censorship hypothesis which, in the version considered here, conjectures that all *physical relevant singularities* are not visible ''for a distant observer.'' Here the term ''for a distant observer'' will not be made more precise as it is of no relevance for the task of the following discussion. One serious drawback of the conjecture, as stated above, is the lack of a precise definition of a physical relevant singularity. It is common, though, to distinguish mainly two types of singularities, namely, *matter singularities* and *gravitational singularities*. Gravitational singularities should be those which are caused *only* due to the *presence* of the gravitational field, while matter singularities would also occur in the gravitational free case. One considers the case of a fluid as a typical example for *matter singularities* like shock waves or rotational singularities. There is, however, no mathematical result known which would allow a justification of this sort of distinction. Since the relativistic case is too difficult to attack directly with mathematical methods known at the moment, it may be useful and instructive to start with the analogous problem in the Newtonian case. The idea would then be to try to generalize results about singularities for the gravitational free case to the situation where self-gravitation is considered. The results of most relevance would be an article by Chemin,[1] in which he generalized results obtained by Beale *et al.*[2] earlier on.

 © 1998 American Institute of Physics

Here the following pragmatic point of view is adopted. If ''typical'' gravitational terms, such as, for example, the tidal tensor, appeared in the continuation criteria for the Euler–Poisson system, this then would be taken as evidence of a gravitational singularity. This, however, is not the case. As in the gravitational free case, only the kinematical quantities of the fluid control the breakdown of smooth solutions.

The Euler–Poisson system that will be considered in the following consists of the following equations:

$$\partial_t \varrho + v^a \partial_a \varrho + \varrho \partial_a v^a = 0, \tag{I.1a}$$

$$\varrho(\partial_t v^a + v^b \partial_b v^a) + \partial^a p = -\varrho \partial^a \phi, \tag{I.1b}$$

$$\Delta \phi = \varrho. \tag{I.1c}$$

Here $\varrho(\mathbf{x},t)$, $p(\mathbf{x},t)$, $\phi(\mathbf{x},t)$, and $\mathbf{v}(\mathbf{x},t)$ depend on the Cartesian coordinates ($\mathbf{x}\in\mathbb{R}^3$ and $\mathbf{x}\in\mathbb{T}^3$, respectively) and the absolute Newtonian time t. The mass density is denoted by $\varrho(\mathbf{x},t)$ and is connected with the pressure p via some equation of state, the only restriction on *the equation of state* being that for $\varrho>0$, p depends smoothly and monotonously on ϱ in such a way that $p=f(\varrho)$ is invertible. An example for this restriction is the so-called *barotropic* equation of state $p=K\varrho^\gamma$, K, $\gamma\in\mathbb{R}^+ 1<\gamma<\frac{5}{3}$, which describes the adiabatic changes of an ideal gas. The case of boundaries with $\varrho=0$ is somewhat subtle and requires further investigation—at the moment no more general statements about the equation of state are made.

A. Helmholtz decomposition and main idea

The gradient of the velocity can be decomposed into the antisymmetric rotation tensor Ω_{ab} and the symmetric deformation tensor θ_{ab}, which in turn can be split into the traceless shear tensor σ_{ab} and in the expansion θ, the trace of θ_{ab}:

$$\partial_a v_b = \tfrac{1}{2}(\partial_a v_b - \partial_b v_a) + \tfrac{1}{2}(\partial_a v_b + \partial_b v_a) = \theta_{ab} - \Omega_{ab} \tag{I.2a}$$

$$= \sigma_{ab} - \Omega_{ab} + \tfrac{1}{3}\delta_{ab}\theta. \tag{I.2b}$$

This is the so-called Helmholtz decomposition and combinations of these kinematical quantities will appear in the continuation criterion presented in Sec. V. Use will be made of the rotation vector Ω^a given by $\Omega^a=\epsilon^{abc}\Omega_{bc}$ which satisfies the propagation equation (V.2) expressing the conservation of rotation. This equation will be used to prove a continuation criterion in which the L^∞-norm of the deformation tensor and the gradient of the matter variable control the occurrence of singularities.

As indicated above, one of the main tasks of this article is to investigate whether gravitational singularities can be formed for self-gravitation fluids. The question is investigated in the following two situations:

(1) The time evolution of a compact body denoted as case I from now on, or
(2) the time evolution of finite spatially periodic deviations from a Newtonian Friedmann-like cosmological model, called case II from now on.

Local in time existence for the first situation has been shown by the author[3] and for the second it has been shown in Ref. 4.

This article proceeds as follows: In the next section some notation is fixed; appropriate function spaces and the definition of symmetric–hyperbolic systems are presented. The fact that the Euler equations, written as a symmetric–hyperbolic system, are not regular at $\varrho=0$ is discussed as well as the implications for describing a compact body. Furthermore, the possible

breakdowns of solutions of symmetric hyperbolic systems are discussed. In Sec. II we deal with the evolution of a compact body. As it turns out, the usual way of deriving boundary conditions via conservation laws is not sufficient to prove that, for the case of vanishing boundary density, the boundary is generated by the fluid lines. This fact never seems to be pointed out in the literature. However, the Makino solutions, treated here, have this ''boundary feature.''

Next uniqueness is proven for the Makino solutions and so an existing gap of these solutions is eliminated. Section IV contains the main results together with some fairly easy to prove corollaries. In Sec. V the proof of the main result is presented. In the last section the differences of the cosmological situation in the main result, weak points of the main result, and possible improvements and generalizations are discussed. In the Appendix finally some details of the proofs are given. Especially the nonlocal elliptic estimate for the gradient of the velocity, one of the key elements of the proof, had to be proved for the $\mathbb{T}^3$.

II. PRELIMINARIES

The symbol $H^s(M)$ denotes the Sobolev space of order s on M. The norm is

$$\|f\|_{H^s(M)}=\left\{\sum_{|\alpha|\leqslant s}\int_M(|\partial_\alpha f|)^2\,dV\right\}^{1/2},$$

where s in a non-negative integer, $s\geqslant 1$, α is a multi-index, M is either $\mathbb{R}^3$ or the torus $\mathbb{T}^3$ with the flat Euclidian metric δ_{ij}, and dV is the volume form associated with the metric δ_{ij}.

The norm of a tensor T in a given coordinate basis as $T=T^{ij\cdots}\partial_i\partial_j\partial\ldots$, is taken with respect to the metric δ_{ij}. For example, $\|U\|_{L^2}=\int_M\delta_{ij}U^iU^j\,dV$. We denote the space of continuously differentiable functions of order k by C^k. It is a Banach space with norm given by

$$\|u\|_{C^k}=\sum_{|\alpha|\leqslant k}\sup|D^\alpha u|. \tag{II.1}$$

Here L^∞ is a Banach space of almost everywhere defined measurable functions bounded almost everywhere (that means everywhere except on a set of measure zero). L^∞ has the norm

$$\|f\|_{L^\infty}=\operatorname{ess\,sup}|f(x)|.$$

$C^k([0,T];H^s)$ (abbreviated by $C^{k,s}_{T,M}$) denotes the space of all H^s-valued functions on $[0,T]$ that have continuous derivatives of order $\leqslant k$.

H^s is a *Banach algebra*, that is, the pointwise multiplication is continuous if $s>\frac{3}{2}$. Furthermore, use will be made of the Sobolev inequalities, that is, let k be a non-negative integer. Then for $f\in H^{s+k}$, the following holds:

$$\|f\|_{C^k}\leqslant\text{const}\,\|f\|_{H^s}\quad\text{if}\ \ s\geqslant 2+k.$$

A. Symmetric hyperbolic systems

Let $U^\alpha:V\subset M\times I\to G\subset\mathbb{R}^l$ be a vector-valued function, G denote an open set of the target space $\mathbb{R}^l$, and I be an open subset of $\mathbb{R}^+$. (In the following indices are used whenever it is convenient.) As will be pointed out later in case II, G consists of all those ϱ with $\varrho>0$. $U\in C(V,G)$ fulfills the system of quasilinear partial differential equations of the form

$$L[U]=A^0_{\alpha\beta}(U)\,\frac{\partial U^\beta}{\partial t}+\sum_{k=1}^{n}A^k_{\alpha\beta}(U)\,\frac{\partial U^\beta}{\partial x^k}+B_{\alpha\beta}(U), \tag{II.2}$$

where the matrix $A^\alpha:V\times G\to\mathscr{M}_{l\times l}$ (the symbol $\mathscr{M}_{l\times l}$ denoting the set of real matrices) is symmetric. The system (II.2) is *symmetric-hyperbolic* iff for an arbitrary $U^\alpha\in G$ there exists a

covector ξ_α such that $A^\alpha \xi_\alpha$ is positive definite. The covector ξ_α is spacelike with respect to the equation. Here ξ_α is chosen to be t_α and therefore A^0 has to be positive definite. Local existence on the strip $M\times[0,T]$ can be shown if A^α and B fulfill the conditions A^0, $A^k \in C^\infty(V \times G; \mathscr{M}_{l\times l})$, $B\in C^\infty(V\times G;\mathbb{R}^l)$, and especially

$$C\delta_{\alpha\beta}U^\alpha U^\beta \leqslant A^0_{\alpha\beta}U^\alpha U^\beta \leqslant C^{-1}\delta_{\alpha\beta}U^\alpha U^\beta, \quad C\in\mathbb{R}^+. \tag{II.3}$$

For details, see, for example, Ref. 5. From now on condition (II.3) will be called the ''A^0 *condition*.''

The Euler equations as a symmetric-hyperbolic system and their singular behavior at $\varrho=0$: To discuss this problem it is sufficient to use the Euler equations (with $\phi=0$) alone; the presence of gravitation does not influence the problem. The difficulty can be seen if system (I.1a) and (I.1b) is written as a symmetric–hyperbolic system in the ''*obvious way*,'' namely by multiplying (I.1a) and (I.1b) by suitable factors of $\varrho^{-n}(\partial p/\partial\varrho)^m$, for example, (I.1a) with $\varrho^{-2}\partial p/\partial\varrho$. However, this procedure leads to $A^0_{\alpha\beta}$, $A^k_{\alpha\beta}$, which contain terms of the form ϱ^{-2}.

$$\begin{pmatrix} \frac{1}{\varrho^2}p' & 0 \\ 0 & \delta_{ab} \end{pmatrix}\partial_l\begin{pmatrix}\varrho\\ v^b\end{pmatrix} + \begin{pmatrix} \frac{1}{\varrho^2}p'v^c & \frac{1}{\varrho}p'\delta^c_b \\ \frac{1}{\varrho}p'\delta^c_a & \delta_{ab}v^c \end{pmatrix}\partial_c\begin{pmatrix}\varrho\\ v^b\end{pmatrix}=0. \tag{II.4}$$

Therefore for $\varrho=0$ the ''A^0 *condition*'' is not fulfilled, the equations are singular at $\varrho=0$, and a local existence theorem cannot be proven on the strip $M\times[0,T]$. There seems to be no possibility to write the Euler system for the variables $(\varrho,\mathbf{v})$ in such a way that they are regular at $\varrho=0$ and the ''A^0 *condition*'' is satisfied, an exemption being that $\lim_{\varrho\to 0} p'/\varrho^2$ exists and is positive, for example, $p=K\varrho^3$. Another possibility is to introduce a new scalar matter variable $w=M(\varrho)$, as will be explained shortly.

III. TIME EVOLUTION OF A COMPACT BODY

In the following section a self-gravitating, compact, isolated body will be described and therefore the gravitational potential has to satisfy the following boundary condition:

$$\lim_{|x|\to\infty}\phi(x,t)=0. \tag{III.1}$$

The boundary conditions for a fluid ball are a special case of the Rankine–Huginot conditions in shock wave theory. However, there is a certain subtle worth discussing further. Let us start with the boundary conditions for the fluid body to be specializing the Rankine–Huginot in order to obtain

$$(\varrho)_+ N_t + (\varrho v^a)_+ N_a = 0, \tag{III.2}$$

$$\{(\varrho)_+ N_t + (\varrho v^a)_+ N_a\} v^b_+ = -p_+ N^b. \tag{III.3}$$

A couple of conclusions can be drawn from these equations:

(1) The first equation implies

$$p_+ = 0, \tag{III.4}$$

which is the usual boundary condition of a compact fluid ball.

(2) Now, two cases have be distinguished depending on the behavior of the density at the boundary.

(2a) The first possibility is that $\varrho_+ \neq 0$ (hard boundary). Then it follows from Eq. (III.2) that

$$N_t + v_+^a N_a = 0 \tag{III.5}$$

holds. This equation states that the normal component of the velocity U^α coincides with the velocity of the boundary or, in other words, the flow lines of the fluid generate the boundary of the body.

(2b) The second possibility is that the density at the boundary vanishes $\varrho_+ = 0$ (soft boundary), but then there are no conditions on the normal component of U^α. In other words, using the conservation laws it is not possible to conclude that the boundary is generated by the flow lines of the fluid supposing $\varrho_+ = 0$. This fact never seems to be stated in the literature.

As will be shown in the next section, the Makino solution has vanishing boundary density, but nevertheless it is possible to show that the boundary is generated by the flow lines of the fluid by means of other tools.

A. Introduction of the Makino variable

I will deal in the following with the situation of a body with vanishing boundary density and solve the problem mentioned above by introducing a new scalar matter variable w, which is related to the density ϱ via the equation

$$w = M(\varrho') = \int_0^{\varrho'} \frac{\sigma}{\varrho}\, d\varrho = \int_0^{\varrho'} \frac{1}{\varrho} \left(\frac{\partial p}{\partial \varrho}\right)^{1/2} d\varrho. \tag{III.6}$$

For the barotropic equation of state w is a kind of ''*weighted sound velocity*.'' With this variable our system takes the form

$$\begin{pmatrix} 1 & 0 \\ 0 & \delta_{ab} \end{pmatrix} \partial_t \begin{pmatrix} w \\ v^b \end{pmatrix} + \begin{pmatrix} v^c & \sigma \delta_b^c \\ \sigma \delta_a^c & \delta_{ab} v^c \end{pmatrix} \partial_c \begin{pmatrix} w \\ v^b \end{pmatrix} = \begin{pmatrix} -\partial_a \phi \\ 0 \end{pmatrix}, \tag{III.7a}$$

$$\Delta \phi = \varrho. \tag{III.7b}$$

Now the ''*regularized*'' evolution equations for $\mathbf{v}$ and the ''*unphysical*'' quantity w satisfy the ''A^0 *condition*'' even for $\varrho = 0$.

This new variable demands a couple of restrictions on the equation of state which are listed below. First these conditions will be stated in general, i.e., without specifying a concrete equation of state, and then applied to the barotropic equation of state.

(1) The first condition is related to the existence of the Makino variable

$$\int_0^{\varrho'} \frac{1}{\varrho} \left(\frac{\partial p}{\partial \varrho}\right)^{1/2} d\varrho < \infty. \tag{III.8}$$

For the barotropic equation of state $p = K\varrho^\gamma$ with K, $\gamma \in \mathbb{R}^+$, $p' = K\gamma\varrho^{\gamma-1}$, and $\sqrt{p'} = \sqrt{\gamma K}\varrho(\gamma-1)/2$, which implies $1 < \gamma$.

(2) The second condition deals with the differentiability conditions of σ and ϱ as functions of w, i.e., $\sigma = S(w)$ and $\varrho = M^{-1}(w)$. Two subproblems have to be considered.

(2a) One restriction on the equation of state is that $\sigma \in C^\infty(\mathbb{R})$ for $\varrho = 0$ which is needed to satisfy the regularity conditions for the local existence Theorem 1. For the barotropic equation of state $\sigma = [(\gamma-1)/2]w$ holds, which satisfies the needed requirements.

(2b) It is necessary for the proof of the existence theorem as well as for the proof of the continuation criterion to estimate ϱ by w even at $\varrho=0$ since the resulting system consists of the variables (w,v^i) for the evolution equations and ϱ for the Poisson equation. Therefore the general requirement for the equation of state is that, for $\varrho=0$,

$$\text{if } w(x,t)\in H^k \quad \text{it should follow that } \varrho(x,t)\in H^k. \tag{III.9}$$

Consider $p=K\varrho^\gamma$. Then the above restriction that $\varrho\in H^3$ even for $\varrho=0$ leads to a restriction of γ, namely, $1<\gamma\leqslant\frac{5}{3}$.

Taking all that into account one can prove the following local in time existence theorem for the Euler–Poisson system with the Makino variable (for details see Ref. 4).

Theorem 1 (Local existence of Makino solutions): *Let the equation of state be such that*

(1) *the Makino variable* $w=M(\rho)$ *exists*;
(2) *for* w *as a function of* (x,t), $w(x,0)\in H^s(\mathbb{R}^3)$ *with* $s\geqslant 3$ *holds*;
(3) *the sound velocity* σ *as a function of the Makino variable is* C^∞, *i.e.*, $\sigma=S(w)\rangle C^\infty(\mathbb{R})$ *holds*;
(4) $\varrho(0)$ *has compact support; and*
(5) $v_a(0,x)\rangle H_s(\mathbb{R}^3)$ *with* $s\geqslant 3$.

Then there exists a time $T>0$ *which depends on the* H^s *-norm of the initial data and there exists a unique solution* $U^\alpha=(\varrho,v^a)^t$ *of system (III.7a), and (III.7b) with*

$$U\in C^{0,s}_{T,\mathbb{R}^3}\cap C^{1,s-1}_{T,\mathbb{R}^3}. \tag{III.10}$$

The density ϱ *has compact support.*

Remarks: Here, contrary to Makino, the equation of state has not been specified. If one wants to do so, the restriction mentioned above has to be taken into account. Although the introduction of w allows us to prove a local (in time) existence theorem, this approach has some particularities which are listed below.

(1) Makino solutions have additional regularity conditions as one would expect to hold for usual regular (ϱ,v) solutions. This additional properties originated from the requirement that $w=\varrho^{(\gamma-1)/2}\rangle H^3$ even at $\varrho=0$ (which implies $w\in C^1$). These conditions are not obvious at all for the γ values considered (recall that $0<(\gamma-1)/2<\frac{5}{3}$). To emphasize these properties Makino denoted his solutions ''*sweet*,'' ''*tame*,'' or ''*gentle*'' solutions. Here these solutions will be denoted as *Makino* solutions. The central point is that one cannot draw, in general, conclusions from the behavior of the Makino solutions about the behavior of the regular (ϱ,v) ones. Insofar as the occurrence of singularities is concerned, the fact that $\|\nabla w\|_{L^\infty}$ blows up does not necessarily imply that $\|\nabla\varrho\|_{L^\infty}$ blows up.
(2) It has been shown by Makino that his solutions do not include the class of regular spherically symmetric static solutions of the Euler–Poisson system which are described for the equation of state $p=K\varrho^\gamma$ by

$$\frac{d^2u}{dz^2}+\frac{2}{z}\frac{du}{dz}+u^{1/(\gamma-1)}=0, \tag{III.11}$$

where $z=Ar$, $A=[(\gamma-1)/\gamma K]\varrho_c^{(2-\gamma)/\gamma}$, and $u=(\varrho/\varrho C)^{\gamma-1}$, ϱ_c being the central density; for details see Ref. 6 Chap. IV, p. 87 ff.
(3) Theorem 1 does not contain any statement about unique determination of a solution by the (physically meaningful) initial data. Of course, uniqueness does hold for the basic symmetric-hyperbolic system used to provide the solution, but the data for that include initial data for v^k outside of the compact support of ϱ. In other words $\varrho=0$ does not imply $v^k=0$. This property requires some further discussion. According to the local existence Theorem 1, w, ϱ, and v are C^1, ϱ has compact support and, according to Proposition 1, v is nonzero at the boundary.

According to its regularity properties v is nonzero also outside the compact support of the body. This situation of ''*velocity without matter*'' is a phenomenon which reminds one of the grin without a cat of the Cheshire Cat of *Alice in Wonderland*. Nevertheless, uniqueness holds for the physical relevant quantities. Using the fact that at the boundary $\sigma=0$ $\varrho=0$ holds, it will be shown below that the space–time region covered by the Makino solution has ''*a lens shaped form*'' which allows the use of energy estimates in order to show the desired uniqueness.

(4) In addition the Makino solutions do not only belong to the class of spherically symmetric static solutions, but they also have the property that their boundary is freely falling as stated in the following proposition:

Proposition 1 (free fall of the boundary): The Makino solutions with compact support describe a compact body with freely falling boundary and the following holds:

$$\left.\frac{\partial^a p}{\varrho}\right|_{\mathscr{R}}=\frac{p'}{\varrho}\,\partial^a\varrho|_{\mathscr{R}}=0. \tag{III.12}$$

Proof of Proposition 1: First consider the flow lines $x^k(x(t),t)$ of the vector field $v^k(x(t),t)$ with $dx^k/dt=v^k$ and

$$\frac{d^2x^k}{dt^2}=\partial_t v^k+v^i\partial_i v^k. \tag{III.13}$$

The Euler equation with the Makino variable has the following form:

$$\frac{d^2x^k}{dt^2}=\partial_t v^k+v^i\partial_i v^k=-Cw\partial^k w-\partial^k\phi. \tag{III.14}$$

Now $\lim_{\varrho\to 0} w\partial^k w=0$ holds, because $\lim_{\varrho\to 0} w=0$ and $|\partial^k w|<C$ is valid. The last inequality follows from the local existence Theorem 1. Hence Eq. (III.14) takes the form $d_t^2x^k=-\partial^k\phi$, which describes the free fall of the boundary.

Remark: In principle the class of solutions with freely falling boundary could be much larger than the class of Makino solutions.

B. Uniqueness for the Makino solutions

Now I want to prove, in spite of the remark (3) of the last section, uniqueness for the Makino system, i.e., that the solution is determined by the initial data on the support of ϱ alone. For this purpose, use will be made of the following properties of the Makino solution:

(1) As long as the solution exists, the flow lines are regular; that means they do not cross. Because the boundary is freely falling and consists of the flow lines, the boundary is regular.
(2) At the boundary, at $p=0$, $\sigma=0$ holds. Therefore the sound cone degenerates to a line there.

From (1) and (2) follows that the boundary is an outer characteristic hypersurface. This means the following. First of all, the boundary is a characteristic hypersurface because the covectors ξ_a belonging to the boundary which vanish on any vector of the boundary v^a lay in the conormal cone. Furthermore, these covectors are outer characteristic covectors in the sense that the expression $A^\alpha\xi_\alpha$ is positive semi-definite. The reason is that the covectors of the boundary are the limit of spacelike covectors, simply because the limit of positive matrices is positive semi-definite. Therefore the space–time region, called L from now on, covered by the collapsing body has a lenslike shape which will play a role in the energy estimates. The intersection between a spacelike

hypersurface Σ_t and L is denoted by $L_t=\Sigma_t\cap L$. The boundary ∂L is given by $\partial L=\partial L_{t_0}+\partial L_{t_1}+\partial L_{C/R}$, $\partial L_{C/R}$ consisting either of outer characteristics or spacelike hypersurfaces.

Proposition 2 (uniqueness of the Makino solutions): Under the conditions of Theorem 1 there exists a unique solution $U\in C^1([0,T];H^{s-1}(\mathbb{R}^3))\cap C^0([0,T];H^s(\mathbb{R}^3))$ *of system (III.7a) and (III.7b). Given two Makino solutions* f_1 *and* f_2 *with* $f_1(0)=f_2(0)$ *on the compact support of the initially density* L_0*, then the following holds*:

(1) *As long as the solution* f_1 *exists, there exists also* f_2*, and the compact support of* ρ *corresponding to* f_1L_1 *is equal to* L_2*, i.e., and* $L_{1t}=L_{2t}$ *holds.*
(2) *Furthermore,* $f_1(t)=f_2(t)$ *and* $\phi_1(t)=\phi_2(t)$ *hold on* L_t.

Remark: The functions ϕ_1 and ϕ_2 are identical not only on the support of the matter but also on the whole $\mathbb{R}^3$. The reason is that if the matter quantities are identical on the support, then, with the help of Eq. (III.23), it follows that the gradient of ϕ is identical for both solutions on the support. That is why ϕ_1 and ϕ_2 differ at most by a constant. Now because ϕ_1 and ϕ_2 solve the Poisson equation with the boundary condition (III.1) both functions vanish at infinity and hence the constant, in fact, is zero.

Proof of Proposition 2: Now shortly the method will be presented with which uniqueness will be proven. Let us recall the procedure for a linear symmetric–hyperbolic equation of the form (for reasons of convince the indices have been omitted)

$$u_t+A^k(x,t)\partial_k u+B(x,t)u=0, \tag{III.15}$$

where A^k and B have continuous derivatives. One gets uniqueness with the help of energy estimates for Eq. (III.15) by integration on a lens shape space–time region L (for details, see Ref. 7, p. 441 ff). One gets an estimate of the form

$$\|u\|_{D^2(L_{t_1})}\leqslant C\|u\|_{L^2(L_{t_0})}. \tag{III.16}$$

This inequality is already sufficient to prove uniqueness of Eq. (III.15) just by taking the difference $z={}^1u-;{}^2u$. In the case of the quasilinear symmetric–hyperbolic Euler equation (for the moment without gravitation) one has, instead of (III.15),

$$u_t+A^k(u)\partial_k u+B(u)=0. \tag{III.17}$$

In order to prove uniqueness one has, similar to the linear case, to consider ${}^1u$ and ${}^2u$ two solutions with the same initial data. Let their difference be called $z={}^1u-{}^2u$, insert it into (III.17), and form the difference of the two corresponding equations, which leads to

$$\partial_t z+A^k({}^2u)\partial_k z+[A^k({}^1u)-A^k({}^2u)]\partial_k^1 u+B^k({}^1u)-B^k({}^2u)=0. \tag{III.18}$$

Now we can apply the mean value theorem,

$$A^k({}^1u)-A^k({}^2u)=H^k({}^1u,{}^2u)z,\quad B({}^1u)-B({}^2u)=K({}^1u,{}^2u)z, \tag{III.19}$$

because A and B are C^1 functions of their variables. Now ${}^1u$, ${}^2u$, and ∂_k^u are considered as known expressions in x, t and these expressions are substituted in H, K, and $A({}^2u)$; thereby (III.18) becomes a linear differential equation for z of the form

$$z_t+az_x+bz=0 \tag{III.20}$$

and the result for the linear system can be applied. In our case, where (III.17) is replaced by

$$u_t+A(u,x,t)u_x=Q, \tag{III.21}$$

u consists of $(w(\varrho),v)$, and $G=\operatorname{grad}\phi$ is coupled with ϱ by $\Delta\phi=\varrho$, the situation is a little different. Applying the idea of considering the difference of solutions with the same initial data for Eq. (III.21) leads to

$$\partial_t z+A^k({}^2u)\partial_k z+[A^k({}^1u)-A^k({}^2u)]\partial_k^1 u=\delta Q. \tag{III.22}$$

So instead of (III.18) we obtain (III.22) where $\delta Q={}^1Q-{}^2Q$. The aim is now to estimate δQ by z or, to be more precise,

$$\|{}^1Q-{}^2Q\|_{L_2(L_t)}\leqslant C\|{}^1\varrho-{}^2\varrho\|_{L_2(L_t)}. \tag{III.23}$$

If such an estimate holds, the proof would be finished. The reason is that one would get an inequality of a special form for which one can derive via the Gronwall inequality an inequality of the form (III.16).

To do so, mainly two steps have to be done:

(1) estimate δQ by $\delta\varrho$ and then
(2) estimate $\delta\varrho$ by $\delta w(\varrho)$.

First of all weighted Sobolev spaces will be introduced, which are a generalization of the usual Sobolev spaces as introduced in Sec. II. They carry the norm. (In the following the quantity which usually is denoted by δ will be denoted by σ in order to avoid confusion with differences of quantities like δQ.)

$$\|f\|_{H_{s,\sigma}(\mathbb{R}^3)}=\left\{\sum_{|\alpha|\leqslant s}\int_{\mathbb{R}^3}((1+\|x\|^2)^{(\sigma+|\alpha|)/2}|\partial_\alpha f|)^2\,dV\right\}^{1/2}. \tag{III.24}$$

Details concerning the properties of these spaces can be found in Ref. 8. By using these spaces one obtains the following inequality:

$$\|\delta Q\|_{L_2(L_t)}\leqslant C\|\delta Q\|_{L_{2,\sigma}(L_t)}. \tag{III.25}$$

after choosing an appropriate constant C. Furthermore, for the Laplace operator it holds that for $s\geqslant 2$ and $-\frac{3}{2}<\sigma<-\frac{1}{2}$, the map

$$\Delta: H_{s,\sigma}\rightarrow: H_{s-2,\sigma+2} \tag{III.26}$$

is an isomorphism, i.e., from

$$\Delta\phi=\varrho \tag{III.27}$$

it follows

$$\|\phi\|_{H_{s,\sigma}(\mathbb{R}^3)}\leqslant C\|\varrho\|_{H_{s-2,\sigma+2}(\mathbb{R}^3)}, \tag{III.28a}$$

and

$$\|\nabla\phi\|_{H_{s,\sigma}(\mathbb{R}^3)}\leqslant C\|\varrho\|_{H_{s-1,\sigma+1}(\mathbb{R}^3)}. \tag{III.28b}$$

Obviously

$$\|\delta Q\|_{L_{2,\sigma}(L_t)}\leqslant\|\delta Q\|_{L_{2,\sigma}(\mathbb{R}^3)}\leqslant C\|\delta Q\|_{H_{1,\sigma}(\mathbb{R}^3)} \tag{III.29a}$$

holds. With the help of (III.28b) one obtains

$$\|\delta Q\|_{H_{1,\sigma}(\mathbb{R}^3)}=\|\nabla\phi\|_{H_{1,\sigma}(\mathbb{R}^3)}\leqslant\|\varrho\|_{H_{0,\sigma+1}(\mathbb{R}^3)}. \tag{III.29b}$$

Now because ϱ has compact support the estimate

$$\|\varrho\|_{H_{0,\sigma+1}(\mathbb{R}^3)}=\|\varrho\|_{H_{0,\sigma+1}(L_t)}\leqslant C\|\varrho\|_{H_0(L_t)} \tag{III.29c}$$

holds by choosing a suitable constant. Furthermore, on L the estimate

$$\|\delta Q\|_{L_2(L_t)}\leqslant C\|\delta Q\|_{L_{2,\sigma}(L_t)} \tag{III.29d}$$

is valid. It remains to prove the inequality

$$\|\delta\varrho\|_{L_2(L_t)}\leqslant C\|\delta w(\varrho)\|_{L_2(L_t)}. \tag{III.29e}$$

This follows from the remarks that were made shortly after the introduction of the Makino variable. Hence the mean value theorem can be applied to obtain

$$\|\delta\varrho\|_{L_2(L_t)}\leqslant\|f'\|_{L_\infty}\|\delta w(\varrho)\|_{L_2(L_t)}. \tag{III.29f}$$

Herewith Lemma 2 is proven. ■

IV. THE COSMOLOGICAL CASE

In the following, the cosmological case will be treated, but in a far less detailed manner than the Makino case. The conceptual problems, the method of describing finite deviations of a given Friedman solution, and some results concerning local existence and stability can be found in Ref. 3 so that here only the local existence theorem is quoted. The starting point is the Euler–Poisson system written for small but finite deviations from the given Friedmann solution given by

$$\begin{pmatrix} \dfrac{g'}{(\hat\varrho+\varrho)^2} & 0 \\ 0 & \delta_{ab} \end{pmatrix}\partial_t\begin{pmatrix}\varrho\\ v^a\end{pmatrix}+\begin{pmatrix} \dfrac{g'}{(\hat\varrho+\varrho)^2}\,v^k & \dfrac{1}{a}\dfrac{g'}{\hat\varrho+\varrho}\,\delta_a^k \\ \dfrac{1}{a}\dfrac{g'}{\hat\varrho+\varrho}\,\delta_b^k & \dfrac{1}{a}\,v^k\delta_{ab} \end{pmatrix}\partial_k\begin{pmatrix}\varrho\\ v^a\end{pmatrix}$$
$$+\begin{pmatrix} \dfrac{3Hg'}{(\hat\varrho+\varrho)^2} & 0 \\ 0 & H\delta_{ab} \end{pmatrix}\begin{pmatrix}\varrho\\ v^a\end{pmatrix}=\begin{pmatrix}0\\ -\partial_a\phi\end{pmatrix}, \tag{IV.1a}$$

$$\frac{1}{a^2}\,\Delta\phi=\varrho, \tag{IV.1b}$$

where $\hat\varrho$ denotes the density and $H(t)=\dot a/a$ is the Hubble constant of the background Friedman solution which satisfies the relations

$$(\varrho a^3)=\text{const},\quad \left(\frac{\dot a}{a}\right)^2=\frac{2}{3}\,\varrho-\frac{k}{a^2}. \tag{IV.2}$$

The difference for the pressure of the background and the ''*real*'' density is denoted by $f(\hat\varrho+\varrho)-f(\hat\varrho)=:g(\varrho,t)$ and g' denotes $dg/d\varrho$.

In order to apply the standard local existence theorem of quasilinear symmetric–hyperbolic systems the ''A^0 *condition*'' has to be satisfied. In the cosmological case, however, the background consists of an overall positive density $\hat\varrho$ so that the ''A^0 *condition*'' does not cause a

problem nor does it force the introduction of a Makino variable contrary to the compact body case. The open set G is not the complete target space as it is in the Makino case, but only an open subset, i.e., $G\subset\mathbb{R}^l$. Here G consists of all ϱ for which $\hat{\varrho}+\varrho>0$ holds. Especially for every compact subset $K\subset G$, $\hat{\varrho}+\varrho$ is bounded uniformly from below and above by a positive constant. This fact plays a role later on in the formulation of the continuation criterion. For this system the following local existence theorem can be proven (see Ref. 3).

Theorem 2 (local existence theorem for cosmological solutions): *Let* $\varrho(0)\in H^s(\mathbb{T}^3)$, $v^a(0)\in H^s(\mathbb{T}^3)$, $s\geqslant 3$, *and* $p=g(\varrho,t)$ *be an equation of state with* $g'>0$. *Let* $\varrho(0)$ *have vanishing integral over the torus, i.e.,* $\int_{\mathbb{T}^3}\varrho(0)\,dV=0$, *and let* $\varrho(0)$ *be chosen such that* $\hat{\varrho}(0)+\varrho(0)>0$ *holds. Then there exists a time* $T>0$ *which depends on the* H^s*-norm of the initial data and* $a(t)$, *and there exists for given initial data a unique solution* $U^\alpha=(\varrho,v^a)^t$ *of the system (IV.1a) and (IV.1b) with*

$$U\in C^{0,s}_{T,\mathbb{T}^3}\cap C^{1,s-1}_{T,\mathbb{T}^3}. \tag{IV.3}$$

The density ϱ *has vanishing integral over the torus, i.e.,* $\int_{\mathbb{T}^3}\varrho\,dV=0$.

V. CONTINUATION CRITERIA FOR THE EULER–POISSON SYSTEM

The following theorem is the main result of the present paper.

Theorem 3 (strong continuation criterion for Makino and cosmological solutions): *Let* U^α *be a maximal solution of the Euler–Poisson system (V.1a) and (V.1b) with the property* $U\in C^1([0,T^*]);H^{s-1}(M))\cap C^0([0,T^*]);H^s(M))$ *of*

$$A^0_{\alpha\beta}(U)\partial_t U^\beta+\sum_{k=1}^{3}A^k_{\alpha\beta}(U)\partial_k U^\beta+B_{\alpha\beta}U^\beta=G_\alpha, \tag{V.1a}$$

$$\Delta\phi=\varrho, \tag{V.1b}$$

where the quantities have the following form:

$$\textit{in case I}\quad U^\alpha=(w(\varrho),v^a)^T,\quad A^0=\delta_{\alpha\beta},\quad B_{\alpha\beta}=0,\quad G_\alpha=(0,\partial_a\phi)^T;$$

$$\textit{in case II}\quad U^\alpha=(\varrho,v^a)^T,\quad A^0\neq\delta_{\alpha\beta},\quad B_{\alpha\beta}\neq 0,\quad G_\alpha=(0,\partial_a\phi)^T,$$

and $A^0_{\alpha\beta}$ *is given in case II by (IV.1a). Then either*

$$(i)\quad T^*=\infty,$$

$$(ii)\ \textit{or case I}\quad T^*<\infty\ \textit{and}\ \int_0^{T^*}(\|\theta\|_{L^\infty}+\|\Omega_{ab}\|_{L^\infty}+\|\nabla w\|_{L^\infty})dt=\infty,$$

$$\textit{or case II}\quad T^*<\infty\ \textit{and}\ \int_0^{T^*}(\|\theta\|_{L^\infty}+\|\Omega_{ab}\|_{L^\infty}+\|\nabla\varrho\|_{L^\infty})dt=\infty.$$

Remarks: Theorem 3 states the not all components of $\varrho_a v_b$ are needed to continue the solution.

This theorem has the following two corollaries:

Corollary 1: Under the assumptions of Theorem 3 and the assumption that the fluid is pressure free and curl free, there are the following possibilities:

(*i*) *either* $T^*=\infty$,

(*ii*) *or, in case II,* $T^*<\infty$ *and* $\int_0^{T^*}(\|\theta\|_{L^\infty})dt=\infty$.

Proof of Corollary 1: If one considers the system (I.1a)–(I.1c), one sees that the equations decouple for $p=0$ *partly because the density does not appear in the Euler equation (I.1b) but* div v *does appear in the continuity equation (I.1a). If* $\int_0^T\|\theta\|_{L^\infty}$ *is bounded on the maximal interval of existence, then, by Lemma 2,* ϱ *is bounded from above and below. To show that* v *is bounded as well, one considers the Euler equation (I.1b) as an independent quasilinear symmetric–hyperbolic equation for* v. *One repeats the steps used for the proof of Theorem 3 and obtains Corollary 1 because* rot $v=0$. ■

Another corollary is the following:

Corollary 2: Under the assumptions of Theorem 3 there are the following possibilities:

(*i*) *either* $T^*=\infty$,

(*ii*) *or case I,* $T^*<\infty$ *and* $\int_0^{T^*}(\|\theta_{ab}\|_{L^\infty(\mathbb{R}^3)}+\|\nabla w(\varrho)\|_{L^\infty(\mathbb{R}^3)})\,dt=\infty$,

(*iii*) *or case II,* $T^*<\infty$ *and* $\int_0^{T^*}(\|\theta_{ab}\|_{L^\infty(\mathbb{T}^3)}+\|\nabla\varrho\|_{L^\infty(\mathbb{T}^{-3})})\,dt=\infty$

Proof of Corollary 2: In order to show the propagation equation for Ω_{ab},

$$\partial_t\Omega_{ab}+v^c\partial_c\Omega_{ab}+\theta_a{}^d\Omega_{db}+\Omega_a{}^d\theta_{db}=0, \tag{V.2}$$

is used. One integrates along the stream lines of $\mathbf{v}$ *and takes the* L^∞*-norm in order to obtain*

$$\|\Omega_{ab}(t)\|_{L^\infty}\leqslant\|\Omega_{ab}(0)\|_{L^\infty}+C\int_0^t\|\delta^{di}\Omega_{ai}\theta_b^i\|_{L^\infty}\,dt. \tag{V.3}$$

By using the Gronwall inequality one obtains

$$\|\Omega_{ab}(t)\|_{L^\infty}\leqslant\|\Omega_{ab}(0)\|_{L^\infty}e^{C\int_0^t\|\theta_{ab}\|_{L^\infty}\,dt}. \tag{V.4}$$

Suppose $\int_0^{T^*}\|\theta_{ab}\|_{L^\infty}\,dt$ *is bounded on the maximal interval of existence. Then* $\int_0^{T^*}\|\theta\|_{L^\infty}\,dt$ *is bounded as well, because* θ *is the trace of* θ_{ab}. *With the help of Eq. (V.4)* Ω_{ab} *is bounded and Theorem 3 can be applied and therefore Corollary 1 is proven.* ■

VI. PROOF OF THE MAIN THEOREM

The proof of Theorem 3 is very much in the line of Kato[2] and Chemin.[1] However, due to the more complicated character of the system, some modifications are necessary; first, the matrix $A^0\neq 0$, so time derivatives of the form $\|\partial_t U\|_{L^\infty}$ have to be controlled. Second, because the Poisson equation is coupled to the fluid equations system, additionally the term $\|U\|_{L^\infty}$ has to be controlled. Finally, one of the key elements of the proof of Kato[2] and Chemin[1] has been a nonlocal elliptic estimate of the gradient of the velocity in terms of its divergence and rotation. It turns out that the part for the $\mathbb{R}^3$ elaborated by Chemin can be completely adopted, while in the spatially periodic case, also sketched by Kato *et al.*, it had to be completely rewritten. This estimate is demonstrated in the Appendix. To sum up the demonstration, the proof consists in the following steps:

Step 1: A general *weak continuation criterion* (3) for the Euler–Poisson is derived. Such a criterion consists of the following two main possibilities.

Step 1.1: Nondivergence singularities: Either the the solutions ''leave their domain of definition'' (for the Euler equations that means that the density turns to zero) or the density/velocity itself diverges.

Step 1.2: Divergence singularities: Derivatives of the solution itself blow up.

Step 2: In order to prove Theorem 3, starting with Theorem 3 the following steps are done:

Step 2.1: Elimination of the possibility of a *non-divergence singularity*.

Step 2.2: The variety of *divergence singularities* is reduced by

Step 2.2.a: The time derivative $\|\partial_t U\|_{L^\infty}$ is eliminated.

Step 2.2.b: Finally some of the spatial derivatives of the velocity are eliminated by a nonlocal elliptical estimate.

Step 1: Weak continuation criterion for Makino and cosmological solutions: General continuation criterion for the compressible Euler equations has been first, formulated by Majda.[5] The following theorem is direct generalization of his result. The content, however, is a little different due to the fact that $A^0 \neq 0$ and due to the presence of the Poisson equation.

Theorem 4 (weak continuation criterion for Makino and cosmological solutions): *Under the conditions of Theorem 3, then either*

(i) $T^* = \infty$,

(ii) *or, in case I,* $T^* < \infty$ *and* $\int_0^{T^*} \|\nabla U^\alpha\|_{L^\infty}\, dt = \infty$,

(iii) *or, in case II,* $T^* < \infty$ *and* $\int_0^{T^*} (\|\partial_t U^\alpha\|_{L^\infty(M)} + \|\nabla U^\alpha\|_{L^\infty(M)})\, dt = \infty$,

(iv) *or* $T^* < \infty$ *and, for* $t \to T^*$,

U *leaves* K *for every compact subset* $K \subset G$.

Remarks: For case I $G = \mathbb{R}^l$, while for case II G consists of all ϱ for which $\hat{\varrho} + \varrho > 0$ holds. One example for possibility (iii) in Theorem 3 is that

(1) for cases I and II $\|U\|_{L^\infty}$ tends to infinity, and
(2) for case II the ''A^0 *condition*'' is violated, for example if ϱ tends to zero.

The proof of this theorem is based on one hand, on the local existence theorem in the cases which are considered here—these are 1 and 2—and, on the other hand, on, an *a priori* estimate on the interval $[0, T)$:

$$\text{case I:} \quad \|U\|_{H^s}^2 \leqslant \|U(0)\|_{H^s}^2 e^{C(\|U\|_{L^\infty}) \int_0^t \|\partial_t U\|_{L^\infty} + \|\nabla U\|_{L^\infty}\, dt}, \quad t \in [0,T), \tag{VI.1a}$$

$$\text{case II:} \quad \|U\|_{H^s}^2 \leqslant \|U(0)\|_{H^s}^2 e^{C(\|U\|_{L^\infty}) \int_0^t \|\nabla U\|_{L^\infty} dt}. \tag{VI.1b}$$

The derivation of these estimates will be outlined in the Appendix.

First a lemma useful for further considerations is presented.

Lemma 1 (L^∞ *estimate for the Poisson equation): Let*

$$\Delta\phi=\varrho \tag{VI.2}$$

and let

(1) case I: ϱ *have compact support,* $\lim_{|x|\to\infty}\phi=0$ *hold and*

(2) case II: ϱ *have vanishing integral over the torus, i.e.,* $\int_{\mathbb{T}^3}\varrho\ dV=0$.

Then the following inequality,

$$\|\text{grad }\phi\|_{L^\infty(M)}\leqslant C\|\varrho\|_{L^\infty(M)}, \tag{VI.3}$$

holds where for case I the constant C depends on the support of the function ρ.

The proof of this lemma will be outlined in the Appendix.

Step 2.1: Eliminating the nondivergence singularities: In the following $\|\varrho\|_{L^\infty(M)}$ and $\|v^a\|_{L^\infty(M)}$ will be estimated by $\int_0^{T^*}\|\theta\|_{L^\infty(M)}\ dt$, and by $\int_0^{T^*}\|\nabla\varrho\|_{L^\infty(M)}\ dt$, or $\int_0^{T^*}\|\nabla w\|_{L^\infty(M)}\ dt$, respectively, and the following lemma will be proven.

Lemma 2 (bounds for ϱ *and* v^a *): Let*

$$\textit{(1)}\quad \int_0^{T^*}\|\theta\|_{L^\infty(M)}\ dt<\infty. \tag{VI.4}$$

$$\textit{(2) In case I,}\quad \int_0^{T^*}\|\nabla w\|_{L^\infty(M)}\ dt<\infty\quad \textit{and, in case II,}\quad \int_0^{T^*}\|\nabla\varrho\|_{L^\infty(\mathbb{T}^3)}dt<\infty. \tag{VI.5}$$

Then the following holds:

$$\textit{(1a) for case I}\quad w<C; \tag{VI.6}$$

$$\textit{(1b) for case II}\quad C<\varrho<C^{-1}\ \textit{on}\ \mathbb{T}^3;\ \textit{and} \tag{VI.7}$$

$$\textit{(2) for both cases}\quad \int_0^{T^*}\|v\|_{L^\infty(M)}\ dt<\infty. \tag{VI.8}$$

Proof of Lemma 2: We first start with case II. To prove statement (1) one first makes use of the continuity equation (using the notation ϱ instead of $\hat{\varrho}+{}^p\varrho$)

$$\partial_t\varrho+v^a\partial_a\varrho+\varrho\,\theta=0, \tag{VI.9}$$

then integrates along the streamlines v^a *and divides by* ϱ *and one obtains*

$$D_t(\ln\varrho)=-\theta. \tag{VI.10}$$

Integrating with respect to time on the interval $[0,T^*)$ *and taking the* L^∞*-norm over* $\mathbb{T}^3$ *one obtains the inequality*

$$\|\ln\varrho\|_{L^\infty(\mathbb{T}^3)}\leqslant\|\ln\varrho(0)\|_{L^\infty(\mathbb{T}^3)}+C\int_0^{T^*}\|\theta\|_{L^\infty}\ dt<\infty. \tag{VI.11}$$

According to assumption (1) $\|\theta\|_{L^\infty}\ dt<\infty$ *holds. It then follows from Eq. (VI.11) that, supposing* $\|\ln\varrho(0)\|_{L^\infty(\mathbb{T}^3)}$ is bounded, *then* $\|\ln\varrho(0)\|_{L^\infty(\mathbb{T}^3)}$ *is bounded on the maximal interval of existence and therefore* ϱ *is bounded above and below by a positive constant. For case I one just uses the fact that it is a monotone function of* ϱ. *Then it follows that if* ϱ *is bounded, then* w *is bounded.*

In order to prove statement (2), one proceeds in a similar way, but uses the Euler equation which has, in case II, the form

$$\partial_t v^a + v^b \partial_b v^a + \frac{f'}{\varrho}\,\partial^a \varrho = -\partial^a \phi, \tag{VI.12}$$

and again one integrates along the streamlines v^a *where one obtains*

$$D_t v^a + \frac{f'}{\varrho}\,\partial^a \varrho = -\partial^a \phi. \tag{VI.13}$$

By integrating with respect to t *and taking the* L^∞*-norm using Lemma 1 one obtains*

$$\|v^a\|_{L^\infty} \leqslant \int_0^{T^*} \left\| \frac{1}{\varrho} \right\|_{L^\infty} \|f'\|_{L^\infty} \|\nabla \varrho\|_{L^\infty} + \|\varrho\|_{L^\infty}\, dt. \tag{VI.14}$$

According to (i), ϱ *is bounded above and below by a positive constant on the maximal interval of existence. Because, as stated in the Introduction, the equation of state* $p(\varrho)$ *will be* C^∞ *for positive* ϱ*,* f' *is bounded for bounded values of* ϱ*. According to the underlying assumptions, furthermore,* $\int_0^{T^*} \|\nabla \varrho\|_{L^\infty}\, dt$ *is bounded and therefore Lemma 2 is proven. For case I again one just uses the fact that* $w = M(\varrho)$ *is a monotone function of* ϱ*; it then follows that if* ϱ *is bounded, then* w *is bounded.*

Step 2.2a. Elimination of $\|\partial_t U\|_{L^\infty}$ *in Theorem 3:* The following result will be proven for case II:

Lemma 3 (Estimate by $\|\partial_t U\|_{L^\infty}$ *by* $\|\nabla U\|_{L^\infty}$*): Under the assumptions of Theorem 3 it holds that*

$$\int_0^T \|\partial_t U\|_{L^\infty}\, dt \leqslant C \int_0^T \|\nabla U\|_{L^\infty}\, dt. \tag{VI.15}$$

In other words, suppose $\int_0^T \|\nabla U\|_{L^\infty}\, dt$ *is bounded on the interval* $[0,T)$*. Then* $\int_0^T \|\partial_t U\|_{L^\infty}\, dt$ *is bounded.*

The key ingredient for the proof of the above lemma is the Lipschitz property of the matrices $\widetilde{A}^k_{\alpha\beta} = (A^0_{\alpha\nu})^{-1} A^k_{\nu\beta}$ of system (IV.1a), i.e., $|\widetilde{A}^k({}^1U) - \widetilde{A}^k({}^2U)| \leqslant C|{}^1U - {}^2U|$. This property can be checked easily for both cases. It follows from the Lipschitz property via the triangle inequality

$$|\widetilde{A}^k(U)| \leqslant C_1|U| + C_2. \tag{VI.16}$$

Proof of Lemma 3: First of all the evolution equation (IV.1a) will be multiplied by $(A^0)^{-1}$. Taking the L^∞-norm one obtains

$$\|\partial_t U\|_{L^\infty} = \|\widetilde{A}^k(U)\partial_k U + \widetilde{B}U + \widetilde{Q}\|_{L^\infty}. \tag{VI.17}$$

In order to estimate the single terms one uses the Lipschitz property of the matrix $\widetilde{A}^k(U)$ on each compact subset $K \subset G$. Furthermore, it will be used that the term $\widetilde{B}$ (which is zero in case II) only depends on t, as one can see easily in Eq. (IV.1b). Therefore one obtains an inequality of the form

$$\|\widetilde{B}U\|_{L^\infty} \leqslant C\|U\|_{L^\infty}. \tag{VI.18}$$

Furthermore, an estimate of the form

$$\|\widetilde{Q}\|_{L^\infty} \leqslant C\|U\|_{L^\infty} \tag{VI.19}$$

will be used which follows from Eq. (VI.3) in Lemma (1):

$$\|\partial_t U\|_{L^\infty} \leqslant \|\widetilde{A}^k(U)\partial_k U\|_{L^\infty} + \|\widetilde{B}U\|_{L^\infty} + \|\widetilde{Q}\|_{L^\infty} \tag{VI.20a}$$

$$\leqslant (C_1\|U\|_{L^\infty} + C_2)\|\nabla U\|_{L^\infty} + C\|U\|_{L^\infty} \tag{VI.20b}$$

$$= C_2\|\nabla U\|_{L^\infty} + C\|U\|_{L^\infty}(C_3\|\nabla U\|_{L^\infty} + 1) \tag{VI.20c}$$

or

$$\int_0^T \|\partial_t U\|_{L^\infty}\, dt \leqslant \int_0^T C_2\|\nabla U\|_{L^\infty} + C\|U\|_{L^\infty}(C_3\|\nabla U\|_{L^\infty} + 1)dt. \tag{VI.21}$$

According to Lemma 2, $\|U\|_{L^\infty}$ is bounded on the interval $[0,T)$ if $\int_0^T \|\theta\|_{L^\infty}\, dt$ and $\int_0^T \|\nabla \varrho\|_{L^\infty}\, dt$ are bounded. Hence, Lemma 3 is proved.

The *a priori* estimate now has the following form:

$$\|U\|_{H^s}^2 \leqslant C\|U(0)\|_{H^s}^2 e^{C\int_0^t \|\nabla U\|_{L^\infty}\, dt}, \quad t \in [0,T), \tag{VI.22}$$

where the constant C does not depend on $\|U\|_{L^\infty}$ any longer.

Step 2.2.b. Elimination of some spatial derivative of the velocity: The final step can be summarized as follows.

Lemma 4: Under the assumptions of Theorem 3 the following holds:

$$\|U\|_{H^s}^2 \leqslant \|U(0)\|_{H^s}^2 e^{C\int_0^t \|\nabla \varrho\|_{L^\infty} + \|A\|_{L^\infty}\, dt}, \quad t \in [0,T), \tag{VI.23}$$

where $\|A\|_{L^\infty} = \|\theta\|_{L^\infty} + \|\Omega_{ab}\|_{L^\infty}$.

Proof of Lemma 4: The central tool of the proof is the following nonlocal decomposition of the gradient of the velocity.

Lemma 5 (nonlocal decomposition of the gradient of) $\nabla \otimes \mathbf{v}$*: Let* $\mathbf{v} \in H^3(M)$*, where* $s \geqslant 3$*. Let* $\operatorname{div} \mathbf{v} = \theta$ *and* $\operatorname{rot} \mathbf{v} = \boldsymbol{\Omega}$*. Then the following inequalities hold*:

(1) For $\mathbb{T}^3$,

$$\|\boldsymbol{\nabla} \mathbf{v}\|_{L^\infty(\mathbb{T}^3)} \leqslant C\{1 + (1 + \ln^+ \|\mathbf{v}\|_{H^s(\mathbb{T}^3)})(\|\theta\|_{L^\infty(\mathbb{T}^3)} + \|\boldsymbol{\Omega}\|_{L^\infty(\mathbb{T}^3)})\}. \tag{VI.24}$$

(2) For $\mathbb{R}^3$,

$$\|\boldsymbol{\nabla} \mathbf{v}\|_{L^\infty(\mathbb{R}^3)} \leqslant C\{1 + (1 + \ln^+ \|\mathbf{v}\|_{H^s(\mathbb{R}^3)})(\|\theta\|_{L^\infty(\mathbb{R}^3)} + \|\boldsymbol{\Omega}\|_{L^\infty(\mathbb{R}^3)}) + \|\mathbf{v}\|_{L^\infty(\mathbb{R}^3)}\}, \tag{VI.25}$$

where

$$\ln^+ a = \begin{cases} \ln a, & \text{if } a \geqslant 1, \\ 0, & \textit{otherwise}. \end{cases}$$

Remark: The proof of (2) can be found in Ref. 1; the proof of (1) will be done in the Appendix.

The estimates (VI.24) and (VI.25) differ about the term $\|v\|_{L^\infty}$. As it has been shown in Lemma 2, $\|v\|_{L^\infty}$ is bounded by $\int_0^{T^*} \|\theta\|_{L^\infty}\, dt$ and $\int_0^{T^*} \|\nabla w\|_{L^\infty}\, dt$ in the maximal interval of existence. Therefore the difference in the proof of the main theorem concerning the use of Lemma 4 in both cases is marginal.

Use will be made of the following obvious inequalities,

$$\ln^+ v + 1 = \ln^+ v + \ln e \leqslant C \ln^+(v + e) \tag{VI.26}$$

and

$$0 \leqslant \int_0^t 1+(1+\ln^+ \|v\|_s)\|A\|_{L^\infty}+e \; d\tau, \tag{VI.27}$$

which lead to

$$\|\nabla v\|_{L^\infty} \leqslant C\{1+(1+\ln^+ \|v\|_s)\|A\|_{L^\infty}\} \leqslant C\{1+(\ln^+ \|v\|_s+e)\|A\|_{L^\infty}\}. \tag{VI.28}$$

Now the improved *a priori* estimate (VI.22) and Eq. (VI.28) will be used so that one gets the following estimates:

$$\|v\|_{H^s}^2 \leqslant \|v(0)\|_{H^s}^2 e^{\int_0^t \|\nabla v\|_{L^\infty}+\|\nabla \varrho\|_{L^\infty} d\tau} \leqslant \|v(0)\|_{H^s}^2 \underbrace{e^{\int_0^t \|\nabla \varrho\|_{L^\infty} d\tau}}_{=C} e^{\int_0^t C\{1+\|A\|_{L^\infty}(\ln^+(\|v\|_s+e))\}d\tau}, \tag{VI.29a}$$

$$\|v\|_{H^s}^2+e \leqslant (C\|v(0)\|_{H^s}^2+e) e^{\int_0^t C\{1+\|A\|_{L^\infty}(\ln^+(\|v\|_s+e))\}d\tau}. \tag{VI.29b}$$

In Eq. (VI.29a) the expression $e^{\int_0^t \|\nabla \varrho\|_{L^\infty} d\tau}$ has been replaced by a constant, because this expression appears on the continuation criterion. Defining $y(\tau)=\|v\|_{\mathrm{H^s}}^2+e$, Eq. (VI.29b) becomes

$$y(t) \leqslant y(0) e^{\int_0^t C\{1+\|A\|_{L^\infty}(\ln^+ y(\tau))\}d\tau}. \tag{VI.30a}$$

Let $z(\tau)=\ln y(\tau)$. Then

$$e^{z(t)} \leqslant e^{z(0)} e^{\int_0^t C\{1+\|A\|_{L^\infty} z(\tau)\}d\tau}, \tag{VI.30b}$$

and

$$z(t) \leqslant z(0)+\int_0^t C\{1+\|A\|_{L^\infty} z(\tau)\}d\tau, \tag{VI.30c}$$

hold, respectively. By using Gronwall's inequality on the interval $[0,t)$, one obtains

$$z(t) \leqslant z(0) e^{\int_0^t C\{1+\|A\|_{L^\infty}\}d\tau}. \tag{VI.30d}$$

Therefore $z(t)$ is bounded by a constant which depends on $\|v\|_{\mathrm{H^s}}^2$, t, and $\int_0^t \|\nabla \varrho\|_{L^\infty}+\|\Omega\|_{L^\infty}+\|\theta\|_{L^\infty}$. From the boundness of $\int_0^T \|\nabla \varrho\|_{L^\infty}+\|\Omega\|_{L^\infty}+\|\theta\|_{L^\infty} dt$ follows the boundness of $\int_0^t \|\nabla \varrho\|_{L^\infty}+\|\Omega\|_{L^\infty}+\|\theta\|_{L^\infty} dt$ and Lemma 4 is proven. ■

VII. DISCUSSIONS

To put the results obtained in context, the following remarks should be taken in account:

(1) *What conclusions can be drawn from the divergence of* $\int_0^{T^*} \|\nabla_x w\|_{L^\infty}$*?* First of all $\|\nabla_x w\|_{L^\infty}$ is given by $\|\nabla_x w\|_{L^\infty}=\|\varrho^{(\gamma-3)/2}\nabla_x \varrho\|_{L^\infty}$. Now two cases have to be distinguished:

(1a) The singularity occurs in the interior of the body; then $\int_0^{T^*} \|\nabla_x w\|_{L^\infty}=\infty$ implies that $\int_0^{T^*} \|\nabla_x \varrho\|_{L^\infty}=\infty$ holds.

(1b) The singularity appears near to the boundary; then $\int_0^{T^*} \|\nabla_x w\|_{L^\infty}=\infty$ implies one of the following possibilities:

(1b I) Either $\int_0^{T^*}\|\nabla_x w\|_{L^\infty}=\infty$ holds or

(1b II) ϱ has no suitable falloff condition such that $\int_0^{T^*}\|\nabla_x w\|_{L^\infty}=\infty$ diverges but not $\int_0^{T^*}\|\nabla_x \varrho\|_{L^\infty}=\infty$. Because one cannot decide which of the possibilities is realized, there is no information available about the the behavior of $\int_0^{T^*}\|\nabla_x \varrho\|_{L^\infty}=\infty$.

(2) *In the cosmological situation* the background solution contains a density that is positive everywhere. Therefore the introduction of the quantity w is not necessary.

(3) *Intuitive picture*: In a fluid free of vorticity the singularities which occur first are shock waves, i.e., $\|\theta\|_{L^\infty}\to\infty$ $\|\nabla\varrho\|_{L^\infty}\to\infty$. In that case Ref. 9 has obtained a lower bound of the form $Ce^{1/\epsilon}$ for the lifespan, where ϵ measures the size of the initial data. In a general flow it is expected that much sooner than that, in a time of order $1/\epsilon$, singularities due to concentration of vorticity will occur.

(4) *Localization of singularities*: To clarify this point let us consider the case of the compact body. The occurrence of singularities is determined by the initial data and by

$$\int_0^{T^*}(\|\theta\|_{L^\infty(\mathbb{R}^3)}+\|\Omega\|_{L^\infty(\mathbb{R}^3)}+\|\nabla\varrho\|_{D^\infty(\mathbb{R}^3)})dt,$$

as one can see from the energy estimate of the form

$$\|u\|_{H^s(L_t)}\leqslant\|u(0)\|_{H^s(L_t)}e^{\int_0^{T^*}(\|\theta\|_{L^\infty(\mathbb{R}^3)}+\|\Omega\|_{L^\infty(\mathbb{R}^3)}+\|\Omega\|_{L^\infty(\mathbb{R}^3)}+\|\nabla\varrho\|_{L^\infty(\mathbb{R}^3)})dt}.$$

This inequality is an equivalent form to express possibility (ii) of Theorem 3. Let the compact support of ϱ be denoted by L_t. Now the initial data for v are changed outside L_t. Then the solutions inside L_t are not influenced by this change. In that sense the occurrence of singularities is influenced by the initial data on L_t. On the other hand $\partial_b v^a$ was estimated by a *nonlocal* elliptic estimated of the form

$$\|\nabla v\|_{L^\infty(\mathbb{R}^3)}\leqslant C\{1+|v|_{L^\infty(\mathbb{R}^3)}+(1+\ln^+\|v\|_{H^s(\mathbb{R}^3)})(\|\Omega\|_{L^\infty(\mathbb{R}^3)}+\|\theta\|_{L^\infty(\mathbb{R}^3)}\}$$

on the whole $\mathbb{R}^3$. This estimate leads to the energy estimate mentioned above. This inequality implies that one has to control the L^∞-norm of θ, Ω, and $\nabla\varrho$ on the whole $\mathbb{R}^3$. This is not the desired estimate to localize the singularities. The estimated needed for this purpose must have the form

$$\|u\|_{H^s(L_t)}\leqslant\|u(0)\|_{H^s(L_t)}e^{\int_0^{T^*}(\|\theta\|_{L^\infty(L_t)}+\|\Omega\|_{L^\infty(L_t)}+\|\nabla\varrho\|_{L^\infty(L_t)})dt}.$$

In this estimate one has only to control the L^∞-norm of θ, Ω, and $\nabla\varrho$ on the support of the matter L_t. It was not possible to derive such an estimate up to now.

(5) *Comparison with known results:* As pointed out already, one weak point of the results obtained here is that they do not state positively that, in fact, singularities do occur. There are, however, two special subcases, both for a spherical symmetric fluid. For this situation Makino and Perthame[10] could show that the lifespan of every smooth spherically symmetric positive Makino solution is finite. In fact they proved the following.

Theorem 5 (finite lifespan of classical spherical symmetric Makino solutions): *Let $\varrho(t)$, $v^a(t)$ be a spherically symmetric Makino solution of the Euler–Poisson system on $0\leqslant t<T^*$. If the support of $\varrho(0)$, $v^a(0)$ is compact and if $\varrho(0)\neq 0$, then T^* is finite.*

Given even this result the continuation criteria are of limited value as far as the question of the structure of singularities is concerned. Let us discuss this somewhat subtle point in the following. For that, let us consider the two relevant criteria, that is, Theorem 3 and Corollary 2. The only difference between them is the different kinematical quantities controlling the possible breakdown. The following discussion hence will be focused only on these different objects. On one hand, if these criteria are used to continue a solution, in the case of spherical symmetry, the former is

superior to the latter because by Theorem 3 it is simply sufficient to control the expansion and not additionally the shear. On the other hand, as far as the appearance of singularities is concerned, each of these criteria is nearly as good or bad as the other. The point is that if we use the Makino–Perthame theorem together with criterion 3 or 2 it cannot be decided whether the expansion alone or the shear as well will blow up. However, at least we know that the shear definitively cannot diverge while the expansion stays bounded. So only singularities of the type *unbounded-expansion–unbounded-shear* or *unbounded-expansion–bounded-shear* but not *bounded-expansion–unbounded-shear* are possible.

More detailed information about the structure of the singularity, such as whether shock waves occur, clearly is not possible.

There is, however, a very special case, namely that of a spherically symmetric shear-free solution of the Euler–Poisson equation with the special equation of state $p=\varrho^{4/3}$. The time-independent version of this solution is a special subcase of the Lane–Emden solution. The time-dependent solution, on the contrary, has not been investigated over the years. Makino[11] recently proved the following result:

Theorem 6 (blow-up solutions of the Euler–Poisson for $p=\varrho^{4/3}$): *If $\gamma=\frac{4}{3}$, then there exists a family of classical solutions of the Euler–Poisson system which tends, in the distribution sense, to the delta distribution after finite time.*

In some sense this type of collapse is very dustlike. Because no shock waves are formed, the shear stays bounded by definition and only the density (and therefore, via the continuity equation, the expansion too) blows up. The reason for this behavior lies partly in the *soft* equation of state.

APPENDIX: ESSENTIAL ELLIPTICAL AND HYPERBOLIC ESTIMATES

In the following Lemma 5 for the three-dimensional torus $\mathbb{T}^3$ is proven and furthermore the proof of the *a priori* estimates VI.1a and VI.1b is sketched.

1. Proof of Lemma (5)

Use has been made of elliptic estimates on the flat three-dimensional torus $\mathbb{T}^3=\mathbb{S}^1\times\mathbb{S}^1\times\mathbb{S}^1$. To prove these estimates it is convenient to use methods that have been used for the $\mathbb{R}^3$ as far as possible. An essential tool is the Green's function for the Poisson equation. The existence of a Green's function and some of its properties have been shown in Ref. 12 for periodic functions on $\mathbb{R}^3$. In order to make use of these results it is necessary to clarify the connection between periodic functions on $\mathbb{R}^3$ and functions on the torus $\mathbb{T}^3$ which will be done in the following. Let us define a flat three-dimensional torus $\mathbb{T}^3$ as the quotient of $\mathbb{R}^3$ by the translations $x^i\mapsto x^i+1$ Here x^i denote the Cartesian coordinates. The metric and the volume form is the flat Riemannian metric and the volume form dV of $\mathbb{R}^3$. Let Q be the subset $(-\frac{1}{2},\frac{1}{2})^3$ of the $\mathbb{R}^3$. The set which is translated by z is denoted by $Q+z$ and is given by

$$z+Q=\{x\in\mathbb{R}^3: x-z\in Q\}. \tag{A1}$$

Let p be a map $p:W\to\mathbb{T}^3$ on the torus and let $U_z=p(z+Q)$. Since the torus is the quotient of $\mathbb{R}^3$ for all points $x\in\mathbb{R}^3$, every point $\overline{x}=p(x)\in\mathbb{T}^3$ is the equivalence class of x. Every set U_z is a open and dense subset of $\mathbb{T}^3$ and the inverse of the restriction of p on $z+Q$ defines a chart with the domain of definition U_z. These charts cover the torus. If now a function f is defined on the torus, then $\tilde{f}=f\circ p$ is a periodic function on $\mathbb{R}^3$. On the contrary, if a periodic function $\tilde{f}$ is given on $\mathbb{R}^3$, there exists a unique function f on $\mathbb{T}^3$ such that $\tilde{f}=f\circ p$ holds. Furthermore, the following relation exists between integrals for f and integrals for $\tilde{f}$:

$$\int_{\mathbb{T}^3} f(\overline{x})dV=\int_{U_z} f(\overline{x})dV=\int_{z+Q}\tilde{f}(x)dx. \tag{A2}$$

Here the point z is arbitrary.

In the following a theorem from Ref. 12 about the Green's function for periodic functions on $\mathbb{R}^3$ will be used.

Lemma 6 (Green's function for periodic functions on the $\mathbb{R}^3$): Let $\mathscr{P}(Q):=\{f:\mathbb{R}^3\to\mathbb{R};\ f(x+\alpha=f(x),x\in\mathbb{R}^3,\ \alpha\in\mathbb{Z}^3\}$ be the set of all periodic functions on $\mathbb{R}^3$ and $C^n_\pi(Q):=C^n(\mathbb{R}^3)\cap\mathscr{P}(Q)$ be the set of all differentiable and periodic functions on $\mathbb{R}^3$. There exists a function $\hat{G}\in C^\infty(\mathbb{R}^3\backslash\mathbb{Z}^3)\cap\mathscr{P}(Q)$ with the following properties:

(1) There exists a function $\hat{G}_0\in C^\infty(\mathbb{R}^3\backslash\mathbb{Z}^3)\cup\{(0,0,0)\}$, so that $\hat{G}:\mathbb{R}^3\to\mathbb{R}$ has the form

$$\hat{G}=\hat{G}_0+\hat{G}_1=\hat{G}_0-\frac{1}{4\pi|x|},\quad x\in\mathbb{R}^3\backslash Z^3. \tag{A3}$$

(2) For all $\varrho\in C^1_\pi(Q)$ with $\int_Q\varrho\ dV=0$, the function

$$\phi(x)=\int_Q\hat{G}(x-y)\varrho(y)dy=\int_{Q+z}\hat{G}(x-y)\varrho(y)dy \tag{A4}$$

with $\phi\in C^2_\pi(Q)$, and $\int_Q\phi dV=0$ is the unique solution of the equation

$$\Delta\phi=\varrho. \tag{A5}$$

To use these statements for functions on the torus one has to keep in mind that the expression

$$G(x-y):\mathbb{R}^3\to\mathbb{R} \tag{A6}$$

has no coordinate-independent meaning on the torus. Because of that the function $\widetilde{G}(x,y):\mathbb{R}^3\times\mathbb{R}^3\to\mathbb{R}$ will be defined by

$$\widetilde{G}(x,y)=\hat{G}(x-y). \tag{A7}$$

Now $\hat{G}(x-y)$ is periodic in x and y and therefore $\widetilde{G}(x,y)$ is periodic in x and y as well. As mentioned already in the discussion of Eq. (A2), there exists a unique local integrable function $G:\mathbb{T}^3\times\mathbb{T}^3\to\mathbb{R}$ with $\widetilde{G}(x,y)=G(p(x),p(y))$. Let ϱ be a (sufficiently differentiable) function on $\mathbb{T}^3$ with zero integral and let $\widetilde{\varrho}=\varrho\circ p$. Now $\widetilde{\phi}(x)$ is defined by

$$\widetilde{\phi}(x)=\int_{z+Q}\hat{G}(x-y)\widetilde{\varrho}(y)dy, \tag{A8}$$

and, according to Lemma (6), $\widetilde{\phi}$ fulfills the equation $\Delta\widetilde{\phi}=\widetilde{\varrho}$. Because $\hat{G}(x-y)$ is periodic in x, $\widetilde{\phi}$ is periodic as well. Therefore there exists a function ϕ on $\mathbb{T}^3$ with $\widetilde{\phi}=\phi\circ p$. This function satisfies $\Delta\phi=\varrho$. From Eqs. (A1) and (A2) it follows that

$$\phi(x)=\int_{\mathbb{T}^3}G(\overline{x,y})\varrho(\overline{y})dV \tag{A9}$$

holds.

Therefore

$$G(\overline{x,y}):\mathbb{T}^3\times\mathbb{T}^3\to\mathbb{R} \tag{A10}$$

has the desired properties of a Green's function on the torus.

The statement (ii) of the Lemma 6, i.e., $\hat{G}=\hat{G}_0+1/4\pi|x|$, has to be translated in the following way. Because $\hat{G}_0(x-y)$ is only defined if the coordinates x^i are not integers or if x, the origin of

$\hat{G}_0$, is not periodic and therefore cannot be defined on the torus. To get a similar decomposition on the torus one proceeds as follows. Let ψ be a smooth function on $\mathbb{R}^3$ whose compact support is in the subset $(-\frac{1}{2},\frac{1}{2})^3$ and who is identical to one in an open neighborhood $|x|<r_0$ of the origin. This neighborhood will be denoted by W. The function ψ can, without loss of generality, be chosen in such a way that the support of ψ lies in a region of the form $|x|<r_1$ where $r_0<r_1<\frac{1}{2}$ holds. Let

$$\overline{G_0}(x,y)=\psi(x-y)\hat{G}_0(x-y) \tag{A11}$$

for x,y in $(-1/2),1/2)^3$ and let $\widetilde{G}_0$ be periodic in x and y. The function $\widetilde{G}_0(x,y)$ is smooth and can be projected to a smooth function $G_0(\overline{x,y}):\mathbb{T}^3\times\mathbb{T}^3\to\mathbb{R}$. Let $d(\overline{x,y})$ be the distance between the points $\overline{x}$ and $\overline{y}$ on the torus. The construction is done in such a way that $G(\overline{x,y})=G_0(\overline{x,y})+1/d(\overline{x,y})$ holds in an open neighborhood Y of the diagonal which is given by

$$Y=\{(\overline{x,y})\in\mathbb{T}^3\times\mathbb{T}^3: d(\overline{x,y})<r_0\}. \tag{A12}$$

In other words, if one defines $G_1=G-G_0$, then $G_1(\overline{x,y})=1/d(\overline{x,y})$ holds in Y.

2. Nonlocal decomposition of ∇v

Starting from the relation $\Delta\mathbf{v}=\nabla\nabla\cdot\mathbf{v}-\nabla\times\nabla\times\mathbf{v}$—keeping in mind the relation $\int v\,dV=0$—and solving this relation with the help of the Green's function one obtains

$$\nabla\mathbf{v}(\overline{x})=\nabla\int_{\mathbb{T}^3}\theta\nabla G(\overline{x,y})d\overline{y}+\nabla\int_{\mathbb{T}^3}\Omega\times\nabla G(\overline{x,y})d\overline{y}. \tag{A13}$$

In the following $A(y)$ is used as an abbreviation for θ and Ω^a, and $K(\overline{x-y})$ as an abbreviation for $\nabla G(\overline{x-y})$ whenever it is convenient.

To estimate the first term the integral will be transformed back to $\mathbb{R}^3$ and basically the methods of Ref. 2 will be used. In order to calculate an L^∞-norm one has to calculate the absolute value of a function in an arbitrary point $\overline{x}\in\mathbb{T}^3$. Therefore, one chooses a z such that $p(z)=\overline{x}$. Then $\overline{x}$ is centralized in the chart on U_z:

$$\int_{\mathbb{T}^3}\nabla G_1(\overline{x,y})A(\overline{y})dV=\int_{z+V}\psi(z-y)A(p(z-y))\nabla\left(\frac{1}{|z-y|}\right)dy. \tag{A14}$$

Next a cutoff function is introduced which has the form

$$\xi_r=\begin{cases}1, & |z|<r,\\ 0, & |z|>2r,\end{cases} \tag{A15}$$

and the properties

$$\xi_r(z)\in C_0^\infty,\quad \|\xi_r(z)\|_{L^\infty}\leqslant\frac{C}{r}. \tag{A16}$$

It will be presumed that $2r\leqslant r_0$ holds so that $\psi(z)=1$ on the support on ξ_r. The integral (A14) will be written with the help of $\xi_r(z-y)+[1-\xi_r(z-y)]$:

$$\nabla_z v(z)=\nabla_z\int_{z+V}\psi(z-y)A(p(z-y))\nabla\left(\frac{1}{|z-y|}\right)dy=\nabla v^{(1)}+\nabla v^{(2)}+\nabla v^{(2)} \tag{A17}$$

where

$$\nabla v^{(1)}=\nabla_z\int_{z+V}\xi_r(z-y)A(p(z-y))\nabla\left(\frac{1}{|z-y|}\right)dy,$$

$$\nabla v^{(2)}=\nabla_z\int_{|z-y|<r_0}\left(1-\xi_r(z-y)\right)A(p(z-y))\nabla\left(\frac{1}{|z-y|}\right)dy,$$

$$\nabla v^{(3)}=\nabla_z\int_{|z-y|>r_0}\psi(z-y)(1-\xi_r(z-y))A(p(z-y))\nabla\left(\frac{1}{|z-y|}\right)dy.$$

The third integral can be estimated by a constant times $\|A\|_{L^\infty}$ because A is multiplied by an continuous function under the integral and it is integrated in a region with finite volume.

a. Estimation of $\nabla v^{(1)}$

Using the antisymmetry of $K_1(z-y)$ and $\xi(z-y)$ and integrating by parts leads to

$$\nabla_z v^{(1)}\int_{|z-y|<2r}[\xi_r(z-y)K_1(z-y)]\nabla_y A(y)dy. \tag{A18}$$

Using the Young inequalities for convolutions on Eq. (A18) leads to

$$\begin{aligned}\|\nabla v\|_{L^\infty}&\leqslant\|K_1(z-y)\|_{L^{4/3}}\|\nabla_y A(y)\|_{L^4}\leqslant\|(z-y)^{-2}\|_{L^{4/3}}\|\nabla_y A(y)\|_{L^4}\leqslant Cr^{1/4}\|\nabla_y A(y)\|_{L^4}\\&\leqslant Cr^{1/4}\|\nabla_y A(y)\|_{H^1}\leqslant Cr^{1/4}\|v\|_{H^3}.\end{aligned}\tag{A19}$$

In the second to last step the Sobolev inequality has been used. In the first step it was used that

$$K_1(z-y)\in L_{2p}(y:|z-y|<2r),\quad p<\tfrac{3}{2}, \tag{A20}$$

holds as one can easily check.

b. Estimation of $\nabla v^{(2)}$

This term again is estimated by the Young inequality where one obtains

$$\|\nabla v\|_{L^\infty}\leqslant\|A\|_{L^\infty}\|\nabla_z\{K_1(z-y)[1-\xi_r(z-y)]\}\|_{L^1}. \tag{A21}$$

Now the L^1-norm in that inequality can be transformed to

$$\begin{aligned}\|\nabla_z\{K_1(z-y)[1-\xi_r(z-y)]\}\|_{L^1}=&\int_{|z-y|>r}\{\nabla_z K_1(z-y)\}[1-\xi_r(z-y)]dy\\&+\int_{|z-y|>r}K_1(z-y)\nabla_z[1-\xi_r(z-y)]dy.\end{aligned}\tag{A22}$$

For the integration polar coordinates have been used and the integration was done between $0<r<2r<r_0$. Also the inequalities $|K_1(z-y)|\leqslant|z-y|^{-3}$ and $\|\xi_\tau(z)\|_{L^\infty}\leqslant C/r$ have been used. Furthermore, for $|z-y|>2r$, $[1-\xi_r(z-y)]=1$ holds and hence $\nabla[1-\xi_\tau(z-y)]=0$. One obtains

$$\begin{aligned}\|\nabla_z\{K_1(z-y)[1-\xi_r(z-y)]\}\|_{L^1} &= \int_{r<|z-y|<r_0}\{\nabla_z K_1(z-y)\}[1-\xi_r(z-y)]dy \\ &\quad + \int_{r<|z-y|<2r} K_1(z-y)\nabla_z[1-\xi_r(z-y)]dy \\ &= \left\{4\pi\int_r^{r_0} s^{-3}[1-\xi_r(z-y)]s^2ds + 4\pi\int_r^{2r} s^{-2}\nabla[1-\xi_r]s^2ds\right\} \\ &= \left\{4\pi\int_r^{1/2} s^{-1}ds + 4\pi\,\frac{1}{r}\int_r^{2r} ds\right\} = \{C(\ln(r_0)-\ln r+1)\}.\end{aligned} \tag{A23}$$

Combining Eqs. (A23) and (A19) one obtains

$$\|\nabla v\|_{L^\infty} \leqslant C\{r^{1/4}\|v\|_{H^3} + (\ln(r_0)+1-\ln r)\|A\|_{L^\infty}\}. \tag{A24}$$

Now two cases are treated separately. If $\|v\|_{H^3} \leqslant (r_0/2)^{-1/4}$, then $r = r_0/2$. Otherwise $r = \|v\|_{H^3}^{-4}$ and therefore $\log r = -4\log\|v\|_{H^3}$. In the first case one gets an estimate of the form

$$\|\nabla v\|_{L^\infty} \leqslant C(1+\|A\|_{L^\infty}), \tag{A25}$$

and in the second case

$$\|\nabla v\|_{L^\infty} \leqslant C(1+(1+\log\|v\|_{H^3})\|A\|_{L^\infty}). \tag{A26}$$

Now $C(1+(1+\log_+\|v\|_{H^3})\|A\|_{L^\infty})$ is greater than the expressions on the right-hand side of both inequalities. Therefore the estimate of Lemma 5 holds. Here $\log_+ = \max\{R_1, R_2\}$ holds where R_1 and R_2 are the right-hand sides of Eqs. (A26) and (A25). ■

Proof of the a priori estimate: For both situations the proof consists of the following two steps:

(1) energy estimates for the symmetric hyperbolic part, and
(2) elliptic estimates for the Poisson equation.

The first step is somehow standard and can be found, for example, in Ref. 5. Therefore it will not repeated here. One obtains, after differentiating the equation with respect to $x^{|\alpha|}$ (where $|\alpha|$ is a multiindex), and with respect to t,

$$\tfrac{1}{2}\partial_t\|U\|_{E,s} \leqslant \tfrac{1}{2}C(\|U\|_{L^\infty})\|(\partial_t U + \nabla U)\|_{L^\infty}\|U\|_{H^s}^2 + C\|U\|_{H^s}^2 + C\|Q\|_{H^s}\|U\|_{H^s}. \tag{A27}$$

Here the usal energy norm $\|U\|_E^2 = \langle A^0U, U\rangle_{L_2}$ and $\|D^{|\alpha|}U\|_E^2 = \langle A^0(D^{|\alpha|}U), (D^{|\alpha|}U)\rangle_{L_2}$ and, respectively, $\|U\|_{E,s}^2 = \Sigma_{\alpha\leqslant s}\|D^\alpha U\|_E^2$. The notation in the last expressions means that one takes the L^2-norm (and the energy norm) for every $D^\alpha f$ for all fixed $|\alpha|$. Both norms have the properties

$$C\langle D^\alpha U, D^\alpha U\rangle_{L_2} \leqslant \langle A^0 D^\alpha U, D^\alpha U\rangle_{L_2} \leqslant C^{-1}\langle D^\alpha U, D^\alpha U\rangle_{L_2}, \tag{A28}$$

$$C\|U\|_{H^s}^2 \leqslant \|u\|_{E,s}^2 \leqslant C^{-1}\|U\|_{H^s}^2. \tag{A29}$$

In case II the equations consist of a lower-order term $B(t)$ that can be estimated by a constant, therefore the term $C\|U\|_{H^s}^2$ appears. In case I there is no lower-order term and therefore inequality (A27) does not contain the term $C\|U\|_{H^s}^2$. It remains for us to derive an estimate for $\|Q\|_{H^s}$ of the form

$$\|Q\|_{H^s} \leqslant C\|U\|_{H^s}, \tag{A30}$$

where Q satisfies the equation

$$Q = \text{grad } \phi \quad \text{and} \quad \Delta\phi = \varrho. \tag{A31}$$

That estimate (A30) holds follows from the following lemma:

Lemma 7 (H^s *estimate for the Poisson equation): Let*

$$\Delta\phi = \varrho \tag{A32}$$

and

(1) *case I: let* ϱ *have compact support, let* ϕ *satisfy the boundary condition* $\lim_{|x|\to\infty} \phi = 0$, *and*
(2) *case II: let* ϱ *have vanishing integral, i.e.,* $\int_{\mathbb{T}^3} \varrho \, dV = 0$.

Then there exists a unique solution ϕ *with the property*

(1) *case I:* $\lim_{|x|\to\infty} \phi = 0$ *and*
(2) *case II:* $\int_{\mathbb{T}^3} \phi \, dV = 0$.

Furthermore,

$$\|\phi\|_{H^s(M)} \leqslant C\|\varrho\|_{H^{s-1}(M)}. \tag{A33}$$

holds, *where for case I the constant C depends on the compact support of* ϱ.

The proof in case I can be found in Ref. 4. The proof of case II is in Ref. 3

So, finally, one has

$$\tfrac{1}{2}\partial_t \|U\|^2_{E,N} \leqslant C(\|\partial_t U\|_{L^\infty} + \|\nabla U\|_{L^\infty} + 1)\|U\|^2_{H^s}. \tag{A34}$$

One uses the relation (A29) and Gronwall's inequality to obtain

$$\|U\|^2_{H^s} \leqslant \|U(0)\|^2_{H^s} e^{C(\|U\|_{L^\infty}) \int_0^T |\partial_t U|_{L^\infty} + |\nabla U|_{L^\infty} \, dt}. \tag{A35}$$

This is the desired estimate and therefore (VI.1b) is proven for case II.

Now let us mention the differences from case I. The main difference is that we have an evolution equation for the nonlinear function $w(\varrho)$ but the source $Q(=\text{grad } \phi)$ is coupled to ϱ. Therefore the same energy estimates are done for $U = (w(\varrho), v^a)$ as well as the elliptic estimate,

$$\|Q\|_{H^s} \leqslant C\|\varrho\|_{H^{s-1}} \leqslant C\|\varrho\|_{H^s}, \tag{A36}$$

and then ϱ has to be estimated by $w(\varrho)$ with the help of the Moser inequalities. ■

[1] J. T. Beale, T. Kato, and A. Majda, ''Remarks on the Breakdown of Smooth Solutions for the 3-D Euler Equations,'' Commun. Math. Phys. **94**, 61–66 (1984).
[2] J.-Y. Chemin, ''Remarques sur L'Apparition de Singularités dans les Ecoulements Euleriens Compressibles,'' Commun. Math. Phys. **133**, 323–329 (1990).
[3] T. Makino, ''On a Local Existence Theorem for the Evolution Equation of Gaseous Stars,'' in Patterns and Waves, edited by T. Nishida, M. Mimura, and H. Fujii (North-Holland, Amsterdam, 1986), pp. 459–479.
[4] U. Brauer, ''An existence theorem for perturbed Newtonian cosmological models,'' J. Math. Phys. **33**, 1224–1233 (1992).
[5] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables* (Springer, New York, 1984).
[6] S. Chandrasekhar, *An Introduction to the Study of Stellar Structure* (Dover, New York, 1939).
[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Wiley, New York, 1989), Vol. II.
[8] Y. Choquet-Bruhat and D. Christodoulou, ''Elliptic systems in $H_{s,\delta}$ spaces on manifolds which are euclidian at infinity,'' Acta Math. **146**, 129–150 (1981).

[9] T. C. Sideris, ''The Lifespan of Smooth Solutions to the Three-Dimensional Compressible Euler equations and the Incompressible Limit,'' Indiana Univ. Math. J. **40**(2) 535–550 (1991).
[10] T. Makino and B. Perthame, ''Sur les solutions à symétric sphérique de l'equation d'euler-poisson pour l'evolution d'etoiles gazeuses,'' Jpn. J. Appl. Math. **7**(1), 165–170 (1990).
[11] T. Makino, ''Blowing up solutions of the Euler-Poisson equations for the evolution of gaseous stars,'' in *Proceedings of the Fourth International Workshop on Mathematical Aspects of Fluid and Plasma Dynamics* (Kyoto, 1991). Transp. Theory Stat. Phys. **21**(4–6), 615–624 (1992).
[12] J. Batt and G. Rein, ''Global classical solutions of the periodic Vlasov–Poisson system in three dimensions,'' C. R. Acad. Sci. Paris **313**, 411–416 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Zeta functions on a product of Einstein manifolds, and the multiplicative anomaly

Andrei A. Bytsenko[a)]
Departamento de Fisica, Universidade Estadual de Londrina, Caixa, Postal 6001, Londrina-Parana, Brazil

Floyd L. Williams[b)]
Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01003

(Received 7 July 1997; accepted for publication 6 October 1997)

The global additive and multiplicative properties of Laplace-type operators acting on irreducible rank 1 symmetric spaces are considered. The explicit form of the zeta function on product spaces and of the multiplicative anomaly is derived. © *1998 American Institute of Physics.* [S0022-2488(98)00702-6]

I. INTRODUCTION

Among the most important geometric structures on manifolds[1–3] almost-product structures play an essential role. Structures of this kind appear in a natural way from a variational principle based on a general class of Lagrangians depending on the Ricci square invariant constructed out of a metric and a symmetric connection.[4,5] Geometric properties of (pseudo-) Riemannian almost-product structures have been studied in Refs. 1 and 6–9.

In this class of structures an almost-product structure on an Einstein manifold is associated with an Einstein metric $\mathbf{g}$ [i.e., $\mathrm{Ric}(\mathbf{g})=\gamma\mathbf{g}$ holds, where $\mathrm{Ric}(\mathbf{g})$ is the Ricci tensor and γ is a constant]. It is well known that gravitational Lagrangians which are nonlinear in the scalar curvature of a metric give rise to equations with higher derivatives or to the appearance of additional matter fields. An important example of a nonlinear Lagrangian leading to equations with higher derivatives is given by Calabi's variational principle.[10] The nonlinear gravitational Lagrangians which still generate Einstein equations are particularly important since they provide a general approach to governing topology in low-dimension models and can be adapted in string theory.

In higher derivative field theories (for example, in higher derivative quantum gravity) as a consequence one has to deal with the product of two (or more) elliptic differential operators.[11,12] In some cases an elliptic (pseudo-) differential operator has a well-defined zeta-regularized determinant. It is natural to investigate multiplicative properties of the determinants of differential operators, in particular the so-called multiplicative anomaly;[13–15] for the definition of the anomaly see Sec. III. The multiplicative anomaly can be expressed by means of the noncommutative residue associated with a classical pseudo-differential operator, the Wodzicki residue.[16]

Recently the important role of this residue has been recognized in physics. The Wodzicki residue, which is the unique extension of the Dixmier trace to the wider class of pseudo-differential operators (PDOs),[17,18] has been considered within the noncommutative geometrical approach to the standard model of the electroweak interactions.[19–24] This residue is also used to write down the Yang–Mills action functional. Other recent papers along these lines can be found in Refs. 25–27. The residue formulas have been used also for dealing with the singularity structure of zeta functions[28] and the commutator anomalies of current algebras.[29]

The purpose of the present paper is to investigate the global additive and multiplicative properties of invertible elliptic operators of Laplace type acting on Einstein manifolds, especially on a noncompact symmetric space, and related zeta functions. The contents of the paper are the following. In Sec. II we summarize some properties of the spectral zeta function $\zeta(s|\oplus_p \mathscr{L}_p)$ related to the Laplace-type operators $\mathscr{L}_p, p=1,2$, each of them acting on an irreducible rank 1 symmetric space of noncompact type. The explicit form of the zeta function is given for a wide

a)Electronic mail: abyts@fisica.uel.br; on leave from Sankt-Petersburg State Technical University, Russia.
b)Electronic mail: williams@math.umass.edu

 © 1998 American Institute of Physics

class of symmetric space products, in particular for the simplest case $H^2\otimes H^2$. In Secs. III and IV the spectral zeta function $\zeta(s|\otimes_p \mathscr{L}_p)$ and the corresponding multiplicative anomaly (for the spaces mentioned above) are studied. The explicit calculation is given for essentially all rank 1 symmetric spaces. We end with some conclusions in Sec. V.

II. SPECTRAL ZETA FUNCTIONS AND PRODUCT OF EINSTEIN MANIFOLDS

In this section we consider the problem of the global existence of zeta functions on pseudo-Riemannian product manifolds, a product of two Einstein manifolds[1,30]

$$(M,\mathbf{g},\mathscr{P})=(M_1,\mathbf{g}_1,\mathscr{P}_1)\otimes(M_2,\mathbf{g}_2,\mathscr{P}_2), \tag{2.1}$$

where $\mathbf{g}=\mathbf{g}_1\otimes\mathbf{g}_2$ and the metric $\mathbf{g}$ separates the variables, i.e.,

$$ds^2=\mathbf{g}_{\alpha\beta}(x)dx^\alpha\otimes dx^\beta+\mathbf{g}_{\mu\nu}(y)dy^\mu\otimes dy^\nu. \tag{2.2}$$

The tangent bundle splits as $TM=TM_1\oplus TM_2$ and $\mathscr{P}=\mathscr{P}_1+\mathscr{P}_2$, where $\mathscr{P}_p(p=1,2)$ are the corresponding projections on TM_p,

$$\mathrm{Ric}(\mathbf{g})=\gamma\mathbf{g}, \tag{2.3}$$

$$\mathscr{P}^2=\mathrm{Id},\quad \mathbf{g}(\mathscr{P}\mathscr{X},\mathscr{P}\mathscr{Y})=\mathbf{g}(\mathscr{X},\mathscr{Y}),\quad \mathscr{X},\mathscr{Y}\in\chi(M), \tag{2.4}$$

and where $\chi(M)$ is the Lie algebra of vector fields $\mathscr{X}$ and $\mathscr{Y}$ on M. The trivial examples of an almost-product structure are given by the choices $\mathscr{P}=\pm\mathrm{Id}$ ($\pm$identity).

Consider an almost-product (pseudo-) Riemannian structure $(M,\mathbf{g},\mathscr{P})$ which is integrable iff $\Delta_{\mathbf{g}}\mathscr{P}=0$ for the Levi–Civita connection $\Delta_{\mathbf{g}}$ of $\mathbf{g}$. The integrable complementary subbundles, i.e., both foliations, are totally geodesic.[1,30] Suppose M is a pseudo-Kähler manifold; the metric $\mathbf{g}$ splits as $\mathbf{g}=\mathbf{g}_1\otimes\mathbf{g}_2$. Such a manifold is an Einstein manifold iff in any adapted coördinates (x^α,y^α) both metrics $\mathbf{g}_1$ and $\mathbf{g}_2$ are Einstein metrics for the same constant γ,[1,30,31] i.e.,

$$R_{\alpha_p\beta_p}=\gamma g_{\alpha_p\beta_p},\quad p=1,2. \tag{2.5}$$

Our consideration will be restricted to only locally decomposable manifolds. A wide class of (pseudo-) Riemannian manifolds includes nonlocally decomposable manifolds as well, which are given by warped product space–times.[32–34] Note that many exact solutions of Einstein equations (associated with Schwarzschild, Robertson–Walker, Reissner–Nordström, de Sitter space–times) and p-brane solutions[35,5] are, in fact, warped product space–times.[31,34]

A. The spectral zeta function of a space form

We shall be working with irreducible rank 1 symmetric spaces $M_p\equiv X_p=G_p/K_p$ of noncompact type. Thus G_p will be connected noncompact simple split rank 1 Lie groups with finite center and $K_p\subset G_p$ will be maximal compact subgroups. Let $\Gamma_p\subset G_p$ be discrete, co-compact, torsion-free subgroups.

Let $L_p: C^\infty(V(M_p))\mapsto C^\infty(V(M_p))$ be partial differential operators acting on smooth sections of vector bundles $V(M_p)$. Let χ_p be a finite-dimensional unitary representation of Γ_p, let $\{\lambda_l(p)\}_{l=0}^\infty$ be the set of eigenvalues of the second-order operator of Laplace type $L_p=-\Delta_{\Gamma_p}$ acting on smooth sections of the vector bundle $V(M_p)$ over $\Gamma_p\backslash X_p$ induced by χ_p, and let $n_l(\chi_p)$ denote the multiplicity of $\lambda_l(p)$.

We shall need further a suitable regularization of the determinant of an elliptic differential operator, since the naive definition of the product of eigenvalues gives rise to a badly divergent quantity. We shall make the choice of zeta-function regularization [see (3.2)]. The zeta functions associated with the operators $\mathscr{L}_p\equiv L_p+b_p$ have the form

$$\zeta(s|\mathscr{L}_p)=\sum_l n_l(\chi_p)\{\lambda_l(p)+b_p\}^{-s}; \tag{2.6}$$

here the b_p are arbitrary constants [endomorphisms of the vector bundle $V(M_p)$], called in the physical literature the potential terms. $\zeta(s|\mathscr{L}_p)$ is a well-defined analytic function for $\mathrm{Re}\, s > \dim(M_p)/2$, and can be analytically continued to a meromorphic function on the complex plane $\mathbf{C}$, regular at $s=0$. One can define the heat kernel of the elliptic operator $\mathscr{L}_p$ by

$$\omega_{\Gamma_p}(t) \equiv \mathrm{Tr}(e^{-T\mathscr{L}_p}) = \frac{-1}{2\pi i}\,\mathrm{Tr}\int_{\mathscr{C}_0} e^{-zt}(z-\mathscr{L}_p)^{-1}\,dz, \tag{2.7}$$

where $\mathscr{C}_0$ is an arc in the complex plane $\mathbf{C}$. By standard results in operator theory there exist $\epsilon, \delta > 0$ such that for $0 < t < \delta$ the heat kernel expansion holds

$$\omega_{\Gamma_p}(t) = \sum_{l=0}^{\infty} n_l(\chi_p) e^{-(\lambda_l(p)+b_p)t} = \sum_{o \leqslant l \leqslant l_0} a_l(\mathscr{L}_p) t^{-l} + O(t^{\epsilon}). \tag{2.8}$$

$\zeta(s|\mathscr{L}_p)$ and $\omega_{\Gamma_p}(t)$ are related by

$$\Gamma(s)\zeta(s|\mathscr{L}_p) = \int_0^{\infty} \omega_{\Gamma_p}(t) t^{s-1}\,dt. \tag{2.9}$$

Eventually we would like also to take $b_p = 0$, but for now we consider only nonzero modes: $b_p + \lambda_l(p) > 0$, $\forall l: \lambda_0(p) = 0$, $b_p > 0$.

B. The explicit form of the zeta function

The following representations of X_p up to local isomorphism can be chosen:

$$X_p = \begin{bmatrix} \mathrm{SO}_1(n,1)/\mathrm{SO}(n) & (I) \\ \mathrm{SU}(n,1)/\mathrm{U}(n) & (II) \\ \mathrm{SP}(n,1)/(\mathrm{SP}(n)\otimes \mathrm{SP}(1)) & (III) \\ F_{4(-20)}/\mathrm{Spin}(9) & (IV) \end{bmatrix}, \tag{2.10}$$

where $n \geqslant 2$, and $F_{4(-20)}$ is the unique real form of F_4 (with Dynkin diagram $\circ - \circ = \circ - \circ$) for which the character $(\dim X - \dim K)$ assumes the value (-20).[36] We assume that if G_1 or $G_2 = \mathrm{SO}(m,1)$ or $\mathrm{SU}(q,1)$, then m is even and q is odd.

We study the zeta function

$$\zeta\left(s\left|\bigoplus_p \mathscr{L}_p\right.\right) = \zeta_{\Gamma_1\backslash X_1 \otimes \Gamma_2 \backslash X_2}(s; b_1, b_2) = \frac{1}{\Gamma(s)} \int_0^{\infty} \prod_p \omega_{\Gamma_p}(t) t^{s-1} dt, \mathrm{Re}\, s > \frac{d_1+d_2}{2}, \tag{2.11}$$

where $d_p = \dim X_p$; see Eq. (2.9). The suitable Harish–Chandra–Plancherel measure is given as follows:

$$|C_p(r)|^{-2} = C_{G_p} \pi r P_p(r) \tanh(a(G_p) r) = C_{G_p} \pi \sum_{l=0}^{d_p/2-1} a_{2l}(p) r^{2l+1} \tanh(a(G_p) r), \tag{2.12}$$

where

$$a(G_p) = \begin{bmatrix} \pi & \text{for } G_p = \mathrm{SO}_1(2n,1) \\ \dfrac{\pi}{2} & \text{for } G_p = \mathrm{SU}(q,1), \quad q \text{ odd} \\ & \text{or } G_p = \mathrm{SP}(m,1), \quad F_4(-20) \end{bmatrix}, \tag{2.13}$$

while C_{G_p} is some constant depending on G, and where the $P_p(r)$ are even polynomials [with suitable coefficients $a_{2l}(p)$] of degree $d-2$ for $G \neq \mathrm{SO}(2n+1,1)$, and of degree $d-1=2n$ for $G = \mathrm{SO}_1(2n+1,1)$.[37,38]

Before starting the main theorem of this section (Theorem 1 below) we shall need further notation, and a few more definitions, some of which arise in connection with Selberg's trace formula-a formula of crucial importance for our purposes which we review. For the data (G,K,Γ) as above, the trace formula[39,40] reads

$$\omega_\Gamma(t;b,\chi)=V\int_{\mathbf{R}} dr\ e^{-(r^2+b+\rho_0^2)t}|C(r)|^{-2}+\theta_\Gamma(t;b,\chi), \tag{2.14}$$

where

$$V\overset{\text{def}}{=}\frac{1}{4\pi}\chi(1)\ \text{vol}\ (\Gamma\backslash G), \tag{2.15}$$

χ is a finite-dimensional unitary representation (or a character) of Γ, and where the number ρ_0 is associated with the positive restricted (real) roots of G (with multiplicity) with respect to a nilpotent factor N of G in an Iwasawa decomposition $G=KAN$. One has $\rho_0=(n-1)/2,n$, $2n+1$, and 11 in the cases $(I)-(IV)$, respectively, in Eq. (2.10). Finally the function $\theta_\Gamma(t;b,\chi)$ is defined as follows:

$$\theta_\Gamma(t;b,\chi)\overset{\text{def}}{=}\frac{1}{\sqrt{4\pi t}}\sum_{\gamma\in C_\Gamma-\{1\}}\chi(\gamma)t_\gamma j(\gamma)^{-1}C(\gamma)e^{-(bt+\rho_0^2t+t_\gamma^2/(4t))}, \tag{2.16}$$

for a function $C(\gamma)$, $\gamma\in\Gamma$, defined on $\Gamma-\{1\}$ by

$$C(\gamma)\overset{\text{def}}{=}e^{-\rho_0 t_\gamma}|\det{}_{n_0}(\text{Ad}(m_\gamma e^{t_\gamma H_0})^{-1}-1)|^{-1}. \tag{2.17}$$

The notation used in Eqs. (2.16) and (2.17) is the following. Let a_0, n_0 denote the Lie algebras of A,N. Since the rank of G is 1, dim $a_0=1$ by definition, say $a_0=\mathbf{R}H_0$ for a suitable basis vector H_0. One can normalize the choice of H_0 by $\beta(H_0)=1$, where $\beta:a_0\mapsto\mathbf{R}$ is the positive root which defines n_0; for more detail see Ref. 38. Since Γ is torsion free, each $\gamma\in\Gamma-\{1\}$ can be represented uniquely as some power of a primitive element δ: $\gamma=\delta^{j(\gamma)}$ where $j(\gamma)\geqslant 1$ is an integer and δ cannot be written as γ_1^j for $\gamma_1\in\Gamma$, $j>1$ an integer. Taking $\gamma\in\Gamma$, $\gamma\neq 1$, one can find $t_\gamma>0$ and $m_\gamma\in K$ satisfying $m_\gamma a=am_\gamma$ for every $a\in A$ such that γ is G conjugate to $m_\gamma\exp(t_\gamma H_0)$, namely for some $g\in G$, $g\gamma g^{-1}=m_\gamma\exp(t_\gamma H_0)$. For Ad denoting the adjoint representation of G on its complexified Lie algebra, one can compute t_γ as follows:[39]

$$e^{t_\gamma}=\max\{|c|\,|\,c=\text{an eigenvalue of Ad}(\gamma)\}, \tag{2.18}$$

in case $G=\text{SO}_1(m,1)$, with $|c|$ replaced by $|c|^{1/2}$ in the other cases of Eq. (2.10).

We will write V_k for $G=G_k$ in (2.15), $k=1,2$, G_1,G_2 being one of the groups in (2.10). Similarly we write $\rho_0(k)$, $a_{2j}(k)$, C_{G_k}, etc., in place of ρ_0,a_{2j},C_G for $G=G_k$. Let

$$\begin{gathered}B\overset{\text{def}}{=}\rho_0^2(1)+\rho_0^2(2)+b_1+b_2,\quad B_p\overset{\text{def}}{=}\rho_0^2(p)+b_p,\quad p=1,2,\\ y_p(s;z)\overset{\text{def}}{=}\frac{s}{2}+(-1)^{p-1}z,\quad \text{for}\quad s,z\in\mathbf{C}.\end{gathered} \tag{2.19}$$

Also for $a,\delta>0$, $n\in\mathcal{N}$, and $s\in\mathbf{C}$, define entire functions $\mathcal{K}_n(s;\delta,a)$ of s by

$$\mathcal{K}_n(s;\delta,a)\overset{\text{def}}{=}\int_{\mathbf{R}}\frac{dr\ r^{2n}\ \text{sech}^2(ar)}{(\delta+r^2)^s}. \tag{2.20}$$

Finally we denote by $\psi_\Gamma(s;\chi)$ the logarithmic derivative of Selberg's zeta function $Z_\Gamma(s;\chi)$, suitably normalized, attached to (G,K,Γ,χ).[37,41–48] For convenience we have assumed (recall) that if G_1 or G_2 is $\mathrm{SO}_1(m,1)$ or $\mathrm{SU}(q,1)$, then m is even and q is odd. We can now state the following.

Theorem 1: *The function $\zeta(s|\oplus_p \mathscr{L}_p)$ admits an explicit meromorphic continuation to* **C** *with at most a simple pole at $s=1,2,\ldots,(d_1+d_2)/2$. In particular on the domain* Re $s<1$,

$$\begin{aligned}\zeta(s|\oplus_p \mathscr{L}_p)=&\frac{\pi^2}{2}\prod_k^2 V_k a(G_k) C_{G_k}\sum_{j=0}^{d_1/2-1}\sum_{l=0}^{j}\sum_{\mu=0}^{d_2/2-1}\sum_{\nu=0}^{\mu}\frac{a_{2j}(1)a_{2\mu}(2)j!\mu!}{(j-l)!(\mu-\nu)!}\\
&\times\frac{\int_0^\infty r^{2(j-l)}\ \mathrm{sech}^2(a(G_1)r)\mathscr{K}_{\mu-\nu}(s-l-\nu-2;r^2+B,a(G_1))dr}{(s-1)(s-2)\ldots(s-(l+1))(s-(l+2))\cdots(s-(l+1+\nu+1))}\\
&+C_{G_1}V_1\ \sin(\pi s)\sum_{j=0}^{d_1/2-1}a_{2j}(1)\int_{\mathbf{R}}dr\ r^{2j+1}\ \tanh(a(G_1)r)\\
&\times\left[\int_0^\infty dt\ \Psi_{\Gamma_2}(\rho_0(2)+t+\sqrt{r^2+B};\chi_2)(2t\sqrt{r^2+B}+t^2)^{-s}\right]\\
&+C_{G_2}V_2\ \sin(\pi s)\sum_{j=0}^{d_2/2-1}a_{2j}(2)\int_{\mathbf{R}}dr\ r^{2j+1}\ \tanh(a(G_2)r)\\
&\times\left[\int_0^\infty dt\ \psi_{\Gamma_1}(\rho_0(1)+t+\sqrt{r^2+B};\chi_1)(2t\sqrt{r^2+B}+t^2)^{-s}\right]\\
&+\frac{1}{2\pi^3 i\Gamma(s)}\int_{\mathrm{Re}\ z=\epsilon}dz\left[\sin\ \pi\left(z+\frac{s}{2}\right)\right]\\
&\times\left[\sin\ \pi\left(\frac{s}{2}-z\right)\right]\Gamma\left(z+\frac{s}{2}\right)\Gamma\left(\frac{s}{2}-z\right)\\
&\times\prod_k^2\left[\int_0^\infty dt\ \Psi_{\Gamma_k}(\rho_0(k)+t+B_k^{1/2};\chi_1)(2tB_k^{1/2}+t^2)^{-y_k(s,z)}\right],\end{aligned}\tag{2.21}$$

for any $-\frac{1}{2}\leqslant\epsilon\leqslant\frac{1}{2}$. All of the integrals are entire functions of s.

Here $\zeta(s|\oplus_p \mathscr{L}_p)$ is defined by Eq. (2.11), which involves the parameters b_1,b_2. In Theorem 1 we require that $b_1,b_2>0$. One can, however, consider zero modes. More precisely assume $b_1,b_2\geqslant 0$, but at least either $b_1>0$ or $b_2>0$. Then Theorem 1 holds on the domain Re $s<0$ when the condition $-\frac{1}{2}\leqslant\epsilon\leqslant\frac{1}{2}$ is replaced by the condition Re $s/2<\epsilon<-$Re $s/2$.

The proof of Theorem 1 involves quite a number of steps. One of the steps, among many, is an analysis of the Mellin transform

$$\hat{\theta}_\Gamma(s;b,\chi)\overset{\mathrm{def}}{=}\int_0^\infty dt\ \theta_\Gamma(t;b,\chi)t^{s-1}\tag{2.22}$$

of the theta function $\theta_\Gamma(t;b,\chi)$ in Eq. (2.16). Let

$$K_\nu(s)=\frac{1}{2}\int_0^\infty dt\ e^{-(s/2)(t+1/t)}t^{\nu-1}\tag{2.23}$$

be the modified Bessel function. Then by results of Refs. 37 and 38,

$$\hat{\theta}_\Gamma(s;b,\chi)=\frac{\sin(\pi s)}{2\pi}\Gamma(s)\int_0^\infty dt\ \psi_\Gamma(\rho_0+t+a(b);\chi)(2a(b)t+t^2)^{-s},\tag{2.24}$$

on $\mathrm{Re}\, s<1$, for $a(b)\overset{\text{def}}{=}\sqrt{b+\rho_0^2}$, $b>0$. Equation (2.24) holds for $b=0$ (the zero modes case) on the domain $\mathrm{Re}\, s<0$. Given the notation in Eq. (2.19) the following holds:

Proposition 1: One has the equality of entire functions of s

$$\int_0^\infty dt\ \prod_p\ \theta_{\Gamma_p}(t;b_p,\chi_p)t^{s-1}=\frac{1}{2\pi i}\ \int_{\mathrm{Re}\, z=c} dz\ \prod_p\ \hat{\theta}_{\Gamma_p}(y_p(s,z);b_p,\chi_p)$$

$$=\frac{1}{2i\pi^3}\ \int_{\mathrm{Re}\, z=c} dz\ \prod_l^2\ \Gamma(y_l(s;z))[\sin\ \pi(y_l(s,z))]$$

$$\times\prod_p\ \int_0^\infty dt\ \psi_{\Gamma_p}(\rho_0(p)+t+B_p^{1/2};\chi_p)(2B_p^{1/2}t+t^2)^{-y_p(s;z)}, \quad (2.25)$$

for $c\in\mathbf{R}$ *with* $\mathrm{Re}\, s/2<c<-\mathrm{Re}\, s/2$, *and* $b_p\geqslant 0$ *with at least* $b_1>0$ *or* $b_2>0$.

The proposition is a consequence of the first integral transformation of Eq. (2.25) as an integral in the complex plane with the help of the Mellin–Parseval identity

$$\int_0^\infty dt\ h(t)g(t)=\frac{1}{2\pi i}\ \int_{\mathrm{Re}\, z=c} dz\ \hat{h}(z)\hat{g}(1-z), \quad c\in\mathbf{R}.$$

To this end one can choose $h(t)=\theta_{\Gamma_1}(t;b_1,\chi_1)t^{s/2}$ and $g(t)=\theta_{\Gamma_1}(t;b_2,\chi_2)t^{s/2-1}$. One makes use of Eq. (2.24).

The meromorphic continuation of $\zeta(s|\oplus_p\mathscr{L}_p)$ to the full complex plane, whereby Theorem 1 follows in particular, can be expressed as follows. For $a,b>0$, $l=0,1,2,3,...$, set

$$f_l(t;a,G)\overset{\text{def}}{=}\int_0^\infty dr\ r^{2l+1}e^{-(r^2+a)t}\ \tanh(a(G)r),$$

$$H_\Gamma^{(1)}(s;a,b,\chi,G)\overset{\text{def}}{=}\int_0^\infty dt\ f_l(t;a,G)\,\theta_\Gamma(t;b,\chi)t^{s-1}, \quad (2.26)$$

for $t>0$ and $s\in\mathbf{C}$. $H_\Gamma^{(l)}(s;a,b,\chi,G)$ is an entire function of s. Then for $b_1,b_2\geqslant 0$ with either $b_1>0$ or $b_2>0$ we have the following.

Theorem 2: *Suppose* $G_p\neq \mathrm{SO}_1(m,1)$, $\mathrm{SU}(q,1)$ *with m odd and q even. Then for* $\mathrm{Re}\, s>d_1/2+d_2/2$,

$$\zeta(s|\underset{p}{\oplus}\mathscr{L}_p)\overset{\text{def}}{=}\frac{1}{\Gamma(s)}\ \int_0^\infty dt\ \prod_p\ \omega_{\Gamma_p}(t,b_p,\chi_p)t^{s-1}$$

$$=\frac{\pi^2}{2}\ \prod_k^2\ V_k a(G_k)C_{G_k}\ \sum_{j=0}^{d_1/2-1}\ \sum_{l=0}^{j}\ \sum_{\mu=0}^{d_2/2-1}\ \sum_{\nu=0}^{\mu}\ \frac{a_{2j}(1)a_{2\mu}(2)j!\mu!}{(j-l)!(\mu-\nu)!}$$

$$\times\frac{\int_0^\infty dr r^{2(j-l)}\ \mathrm{sech}^2(a(G_1)r).\mathscr{K}_{\mu-\nu}(s-l-\nu-2;r^2+B,a(G_1))dr}{(s-1)(s-2)\cdots(s-(l+1))(s-(l+2))\cdots(s-(l+1+\nu+1))}$$

$$+\frac{2\pi}{\Gamma(s)}\left\{V_1C_{G_1}\ \sum_{j=0}^{d_1/2-1}\ a_{2j}(1)H_{\Gamma_2}^{(j)}(s;b_1+B_1,b_2,\chi_2,G_1)\right.$$

$$\left.+V_2C_{G_2}\ \sum_{j=0}^{d_2/2-1}\ a_{2j}(2)H_{\Gamma_1}^{(j)}(s;b_2+B_2,b_1,\chi_1,G_2)\right\}$$

$$+\frac{1}{\Gamma(s)}\int_0^\infty dt\ \prod_k^2\ \theta_{\Gamma_k}(t,b_k,\chi_k)t^{s-1}. \tag{2.27}$$

The simplest case of Theorem 1 is, for example, $G_1=G_2=G=\mathrm{SO}_1(2,1)\simeq\mathrm{SL}(2,R)$; besides $X_1=X_2=H^2$ is a two-dimensional real hyperbolic space. Then we have $\rho_0^2=\rho_0^2(1)+\rho_0^2(2)=\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$, $a_{20}=1$, $C_G=1$, $a(G)=\pi$, $\Gamma_1=\Gamma_2=\Gamma$, and finally $|C(r)|^{-2}=\pi r\tanh(\pi r)$. Using Eq. (2.21) of Theorem 1 for $\mathrm{Re}\,s<1$ we have

$$\begin{aligned}\zeta(s|\underset{p}{\oplus}\mathscr{L}_p)=&\frac{\pi^3V_1^2}{2(s-1)(s-2)}\int_0^\infty dr\ \mathrm{sech}^2(\pi r)\mathscr{K}_0(s-2;r^2+B,\pi)\\&+2V_1\ \sin(\pi s)\int_R dr\ r\ \tanh(\pi r)\int_0^\infty dt\ \psi_\Gamma\left(\frac{1}{2}+t+\sqrt{r^2+B};\chi\right)(2t\sqrt{r^2+B}+t^2)^{-s}\\&+\frac{1}{2\pi^3i\Gamma(s)}\int_{\mathrm{Re}\,z=\epsilon}dz\left[\sin\ \pi\left(z+\frac{s}{2}\right)\right]\left[\sin\ \pi\left(\frac{s}{2}-z\right)\right]\Gamma\left(z+\frac{s}{2}\right)\\&\times\Gamma\left(\frac{s}{2}-z\right)\int_0^\infty dt\ \psi_\Gamma\left(\frac{1}{2}+t+B_1^{1/2};\chi\right)(2tB_1^{1/2}+t^2)^{-y_1(s;z)}\int_0^\infty dt\\&\times\psi_\Gamma\left(\frac{1}{2}+t+B_2^{1/2};\chi\right)(2tB_2^{1/2}+t^2)^{-y_2(s;z)},\end{aligned} \tag{2.28}$$

where $B=\frac{1}{2}+b_1+b_2$ and $B_p=\frac{1}{4}+b_p$ $(p=1,2)$.

C. The zeta function of the product of Laplace-type operators

The spectral zeta function associated with the product $\otimes\mathscr{L}_p$ has the form

$$\zeta(s|\underset{p}{\otimes}\mathscr{L}_p)=\sum_{j\geqslant 0}n_j\prod_p^2\ (\lambda_j+b_p)^{-s}. \tag{2.29}$$

We shall always assume that $b_1\neq b_2$, say $b_1>b_2$. If $b_1=b_2$, then $\zeta(s|\otimes\mathscr{L}_p)=\zeta(2s|\mathscr{L})$ is a well-known function. For $b_1,b_2\in\mathbf{R}$, set $b_+\underset{\mathrm{def}}{=}(b_1+b_2)/2$, $b_-\underset{\mathrm{def}}{=}(b_1-b_2)/2$; thus $b_1=b_++b_-$ and $b_2=b_+-b_-$. Also let

$$I_\nu(s)=\left(\frac{s}{2}\right)^\nu\sum_{m=0}^\infty\ \frac{(s/2)^{2m}}{m!\Gamma(m+\nu+1)}. \tag{2.30}$$

Theorem 3: *The spectral zeta function can be written as follows:*

$$\zeta(s|\underset{p}{\otimes}\mathscr{L}_p)=(2b_-)^{1/2-s}\frac{\sqrt{\pi}}{\Gamma(s)}\int_0^\infty dt\ \omega_\Gamma(t,b_+)I_{s-1/2}(b_-t), \tag{2.31}$$

where the integral converges absolutely for $\mathrm{Re}\,s>d/4$ $[d=\dim(G/K)]$.

This formula is a main starting point to study the zeta function. It expresses $\zeta(s|\otimes\mathscr{L}_p)$ in terms of the Bessel function $I_{s-1/2}(b_-t)$ and $\omega_\Gamma(t,b_+)$, where the trace formula applies to $\omega_\Gamma(t,b_+)$. Thus we can study the zeta function by the trace formula. One can use the trace formula for $\omega_\Gamma(t,b_+)$; see Eq. (2.14). As a result

$$\zeta(s|\underset{p}{\otimes}\mathscr{L}_p)=(2b_-)^{1/2-s}\frac{\sqrt{\pi}}{\Gamma(s)}\int_0^\infty dt\ t^{s-1/2}I_{s-1/2}(b_-t)$$

$$\times\left[\frac{\chi(1)\mathrm{vol}(\Gamma\backslash G)}{4\pi}\int_{\mathbf{R}}dr\ e^{-(r^2+b_++\rho_0^2)t}|C(r)|^{-2}+\theta_\Gamma(t)\right]. \tag{2.32}$$

Then for $\mathrm{Re}\,s>0$ Fubini's theorem gives

$$\int_0^\infty dt\ \frac{\chi(1)\mathrm{vol}(\Gamma\backslash G)}{4\pi}\int_{\mathbf{R}}dr\ e^{-(r^2+b_++\rho_0^2)t}|C(r)|^{-2}I_{s-1/2}(b_-t)t^{s-1/2}$$

$$=(2b_-)^{1/2-s}\ \frac{\chi(1)\mathrm{vol}(\Gamma\backslash G)\Gamma(s)}{4\pi^{3/2}}\int_{\mathbf{R}}dr|C(r)|^{-2}\prod_p\ (r^2+B_p)^{-s}. \tag{2.33}$$

In order to analyze the last integral in Eq. (2.33) (for the possibility of a meromorphic continuation) it is useful to rewrite the function $|C(r)|^{-2}$ [see Eq. (2.12)], using the identity $\tanh(ar)\equiv 1-2(1+e^{2ar})^{-1}$. Then one can calculate a suitable integral in terms of the hypergeometric function $F(\alpha,\beta;\gamma;z)$, namely,

$$\int_0^\infty dr\ r^{2j+1}\prod_p\ (r^2+B_p)^{-s}=\frac{\sqrt{\pi}\Gamma(2s-j-1)j!}{2^{2s}\Gamma(s)\Gamma(s+\frac{1}{2})}\ B_1^{-s}B_2^{j+1-s}\left(\frac{2B_1}{B_1+B_2}\right)^{j+1}$$

$$\times F\left(\frac{j+1}{2},\frac{j+2}{2};s+\frac{1}{2};\left(\frac{B_1-B_2}{B_1+B_2}\right)^2\right), \tag{2.34}$$

which is a holomorphic function on $\mathrm{Re}\,s>(j+1)/2$ and admits a meromorphic continuation to $\mathbf{C}$ with only simple poles at points $s=(j+1-n)/2$, $n\in\mathcal{N}$. Let us introduce the function

$$E_j(s)\overset{\mathrm{def}}{=}2\int_0^\infty\frac{dr\ r^{2j+1}}{(1+e^{2a(G)r})}\prod_p\ (r^2+B_p)^{-s},$$

which is an entire function of s. Then a consequence of Eqs. (2.31)–(2.34) is the following.

Theorem 4: *For* $\mathrm{Re}\,s>d/4$ *the explicit meromorphic continuation holds:*

$$\zeta(s|\underset{p}{\otimes}\mathscr{L}_p)=\mathscr{F}(s)-2A\sum_{j=0}^{d/2-1}\ a_{2j}E_j(s)+\mathscr{I}(s), \tag{2.35}$$

where

$$A\overset{\mathrm{def}}{=}\frac{1}{4}\ \chi(1)\mathrm{vol}(\Gamma\backslash G)C_G,$$

$$\mathscr{F}(s)\overset{\mathrm{def}}{=}A(B_1B_2)^{-s}\sum_{j=0}^{d/2-1}\ \frac{a_{2j}j!\left(\frac{2B_1B_2}{B_1+B_2}\right)^{j+1}F\left(\frac{j+1}{2},\frac{j+2}{2};s+\frac{1}{2};\left(\frac{B_1-B_2}{B_1+B_2}\right)^2\right)}{(2s-1)(2s-2)\cdots(2s-(j+1))}, \tag{2.36}$$

$$\mathscr{I}(s)\overset{\mathrm{def}}{=}(2b_-)^{1/2-s}\ \frac{\sqrt{\pi}}{\Gamma(s)}\int_0^\infty dt\ \theta_\Gamma(t,b_+)I_{s-1/2}(b_-t)t^{s-1/2}.$$

The goal now is to compute the derivative of the zeta function at $s=0$. Thus we have

$$\zeta'(0|\underset{p}{\otimes}\mathscr{L}_p)=A\sum_{j=0}^{d/2-1}\ a_{2j}\sum_l^4\ \mathscr{E}_l, \tag{2.37}$$

where

$$\mathscr{E}_1 = j!(B_1^{j+1}+B_2^{j+1})\sum_{k=0}^{j}\frac{(-1)^{k+1}}{k!(j-k)!(j+1-k)!},$$

$$\mathscr{E}_2 = B_2^{j+1}\left(\frac{B_1-B_2}{2B_1}\right)\frac{(-1)^j}{(j+1)!}\sum_{k=1}^{\infty}\frac{(j+k+1)!}{(k+1)!}\,\sigma_k\left(\frac{B_1-B_2}{B_1}\right)^k,$$

$$\mathscr{E}_3 = \log(B_1B_2)\,\frac{(-1)^j}{2(j+1)}\,(B_1^{j+1}+B_2^{j+1}) - 4\int_0^{\infty}\frac{dr\ r^{2j+1}\log\left(\frac{r^2+B_1}{r^2+B_2}\right)}{1+e^{2a(G)r}},$$

$$\mathscr{E}_4 \equiv \mathscr{T}'(s=0) = T_\Gamma(0,b_1,\chi_1)+T_\Gamma(0,b_2,\chi_2), \tag{2.38}$$

and

$$T_\Gamma(0,b_p,\chi_p) \overset{\text{def}}{=} \int_0^{\infty} dt\ \theta_\Gamma(t,b_p)t^{-1}, \quad \sigma_k \overset{\text{def}}{=} \sum_{k=1}^{n}\frac{1}{k}. \tag{2.39}$$

Similarly using results from Ref. 49 one can show

$$\zeta'(0|\mathscr{L}_p) = A\sum_{j=0}^{d/2} a_{2j}\left\{ j!B_p^{j+1}\left[\sum_{k=0}^{j}\frac{(-1)^{k+1}}{k!(j-k)!(j+1-k)^2}+\frac{(-1)^j}{(j+1)!}\log B_p\right]\right.$$
$$\left. +4\int_0^{\infty}\frac{dr\ r^{2j+1}\log(r^2+B_p)}{1+e^{2a(G)r}}\right\} + T_\Gamma(0,b_p,\chi_p). \tag{2.40}$$

Then we can put these results together to compute the anomaly

$$\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2) \overset{\text{def}}{=} \zeta'(0|\otimes\mathscr{L}_p)-\zeta'(0|\mathscr{L}_1)-\zeta'(0|\mathscr{L}_2). \tag{2.41}$$

Proposition 2: A preliminary form of the multiplicative anomaly is

$$\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2) = A\sum_{j=0}^{d/2-1} a_{2j}B_2^{j+1}\left(\frac{B_1-B_2}{2B_1}\right)\frac{(-1)^j}{(j+1)!}\sum_{k=1}^{\infty}\sigma_k\left(\frac{B_1-B_2}{B_1}\right)^k\frac{(j+k+1)!}{(k+1)!}$$
$$+A\sum_p \log(B_p)\sum_{j=0}^{d/2-1} a_{2j}\,\frac{(-1)^{j+p}}{2(j+1)}\,(B_1^{j+1}-B_2^{j+1}). \tag{2.42}$$

To obtain the final explicit form of the anomaly we must compute the infinite series appearing in the statement of Proposition 2. For $0<x<1$, $j=0,1,2,...$, let

$$S_j(x) = \sum_{k=1}^{\infty}\frac{(j+k+1)!}{(k+1)!}\,\sigma_k x^k. \tag{2.43}$$

Thus in Proposition 2 we choose $x=(B_1-B_2)/B_1$. After a tedious calculation one obtains the following result.

Theorem 5: *The function $S_j(x)$ admits the representation*

$$(1-x)^{j+1}S_j(x) = P_j(x)+Q_j(x)\log(1-x), \tag{2.44}$$

where $P_j(x)$ and $Q_j(x)$ are polynomials of degree j given by

$$\frac{P_j(x)}{(j+1)!} = \frac{jx}{2}\,(1-x)^{j-1}+\frac{j(j-1)}{4}\,x^2(1-x)^{j-2}+\sum_{p=3}^{j}\frac{j!}{(p+1)!(j-p)!}$$

$$\times\left[\frac{1}{p}+\frac{1}{p-1}+\sum_{q=1}^{p-2}\frac{1}{p-q-1}\right]x^p(1-x)^{j-p},$$
$$\frac{xQ_j(x)}{(j+1)!}=-\frac{1}{j+1}\,[1-(1-x)^{j+1}]. \tag{2.45}$$

III. THE MULTIPLICATIVE ANOMALY AND ASSOCIATED ONE-LOOP CONTRIBUTIONS

In this section the product of the second-order operators of Laplace type $\otimes\mathscr{L}_p$, $p=1,2$, will be considered. We are interested in multiplicative properties of determinants, the multiplicative anomaly,[50,13,14] related with one-loop approximation in quantum field theory. This approximation can be given in terms of the multiplicative anomaly $F(\mathscr{L}_1,\mathscr{L}_2)$, which has the form

$$F(\mathscr{L}_1,\mathscr{L}_2)=\det{}_\zeta\Big[\oplus_p\mathscr{L}_p\Big]\big[\det{}_\zeta(\mathscr{L}_1)\det{}_\zeta(\mathscr{L}_2)\big]^{-1}, \tag{3.1}$$

where we assume a ζ-regularization of determinants, i.e.,

$$\det{}_\zeta(\mathscr{L}_p)\overset{\text{def}}{=}\exp\left(-\frac{\partial}{\partial s}\,\zeta(s=0|\mathscr{L}_p)\right). \tag{3.2}$$

Generally speaking, if a multiplicative anomaly related to elliptic operators is nonvanishing, then the relation $\log\det(\otimes\mathscr{L}_p)=\mathrm{Tr}\,\log(\otimes\mathscr{L}_p)$ does not hold.

The operator product $\otimes\mathscr{L}_p$ can arise in higher-derivative field theories. Note also that the partition function Z, associated with the product of two elliptic differential operators, for the simplest $O(2)$ invariant model of self-interacting charged fields in $\mathbf{R}^4$ (Ref. 51) has been analyzed recently in Ref. 15:

$$\log Z\propto-\log\det(\otimes\mathscr{L}_p). \tag{3.3}$$

We come now to a key result of this paper. Recall Eqs. (2.37) and (2.40), definition (2.41), Proposition 2, Theorem 5, and the notation $B_p=b_p+\rho_0^2(p)$, $p=1,2$, of Eq. (2.19). We assume $b_1>b_2>0$, and if $G=\mathrm{SO}_1(m,1)$ or $\mathrm{SU}(q,1)$ in Eq. (2.9), then m is even and q is odd.

Theorem 6: *The explicit formula for the multiplicative anomaly in Eq. (2.41) is*

$$\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2)=A\sum_{j=0}^{d/2-1}\frac{a_{2j}(-1)^j}{2}\left\{\frac{j}{2}\,(B_1-B_2)^2B_2^{j-1}+\frac{j(j-1)}{4}\,(B_1-B_2)^3B_2^{j-2}\right.$$
$$\left.+\sum_{p=3}^{j}\frac{j!}{(p+1)p!(j-p)!}\left[\frac{1}{p}+\frac{1}{p-1}+\sum_{q=1}^{p-2}\frac{1}{p-q-1}\right](B_1-B_2)^{p+1}B_2^{j-p}\right\}, \tag{3.4}$$

where the coefficients a_{2j} are given by Eq. (2.11).

In the case $G=\mathrm{SO}_1(m,1)$ with m odd, $\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2)=0$. In the special case $G=\mathrm{SO}_1(2,1)$ ($d=\dim X=2$) Theorem 6 gives $\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2)=0$. In general for odd-dimensional manifolds [$X=\mathrm{SO}_1(m,1)/\mathrm{SO}(m)$ with m odd being a special case] one has $\mathscr{A}(\mathscr{L}_1,\mathscr{L}_2)=0$; see, for example, Ref. 15.

IV. THE RESIDUE FORMULA AND THE MULTIPLICATIVE ANOMALY

The value of $F(\mathscr{L}_1,\mathscr{L}_2)$ can be expressed by means of the noncommutative Wodzicki residue.[16] Let $\mathscr{O}_p$, $p=1,2$, be invertible elliptic PDOs of real nonzero orders α and β such that $\alpha+\beta\neq0$. Even if the zeta functions for operators $\mathscr{O}_1,\mathscr{O}_2$ and $\mathscr{O}_1\otimes\mathscr{O}_2$ are well defined and if their principal symbols obey the Agmon–Nirenberg condition (with appropriate spectra cuts) one has in general that $F(\mathscr{O}_1,\mathscr{O}_2)\neq1$. For such invertible elliptic operators the Wodzicki's formula for the anomaly of commuting operators holds:[15,16]

$$\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2)=\mathcal{A}(\mathcal{O}_2,\mathcal{O}_1)=\log\left(F(\mathcal{O}_1,\mathcal{O}_2)\right)=(2\alpha\beta(\alpha+\beta))^{-1}\,\mathrm{res}\left[(\log(\mathcal{O}_1^{\beta}\otimes\mathcal{O}_2^{-\alpha}))^2\right]. \tag{4.1}$$

More general formulas have been derived in Refs. 13 and 14. Furthermore, the anomaly can be iterated consistently. Indeed, using Eqs. (3.2) and (4.1) we have

$$\begin{aligned}
&\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2)=\zeta'(0|\mathcal{O}_1\mathcal{O}_2)-\zeta'(0|\mathcal{O}_1)-\zeta'(0|\mathcal{O}_2),\\
&\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\mathcal{O}_3)=\zeta'\left(0\left|\bigotimes_j^3\mathcal{O}_j\right.\right)-\sum_j^3\zeta'(0|\mathcal{O}_j)-\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2),\\
&\cdots\\
&\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_n)=\zeta'\left(0\left|\bigotimes_j^n\mathcal{O}_j\right.\right)-\sum_j^n\zeta'(0|\mathcal{O}_j)-\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2)-\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\mathcal{O}_3)\cdots\\
&\qquad-\mathcal{A}(\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_{n-1}).
\end{aligned} \tag{4.2}$$

In particular, for $n=2$ and $\mathcal{O}_p\equiv\mathcal{L}_p$ the explicit form of anomaly is given by Eq. (2.41) and Theorem 6. We note that for the four-dimensional space with $G=\mathrm{SO}_1(4,1)$, one derives from Theorem 6 the result

$$\mathcal{A}(\mathcal{L}_1,\mathcal{L}_2)=-A_G(b_1-b_2)^2,\quad d=4, \tag{4.3}$$

which also follows from Wodzicki's formula (4.1), where $A_G=\frac{1}{4}Aa_2$.

V. CONCLUSIONS

In this paper the global additive and multiplicative properties of Laplace-type operators and related zeta functions have been studied. We have considered product structures on Einstein manifolds, especially on an irreducible rank 1 symmetric space.

In fact the explicit form of the zeta function on product spaces (Theorem 1) is derived. As an example the zeta function associated with the Kronecker sum of Laplacians acting on two-dimensional real hyperbolic spaces is calculated.

We have obtained also the explicit formula for the multiplicative anomaly in the main theorem, Theorem 6. It has been shown that the anomaly is equal to zero for $d=2$ and for the odd-dimensional cases; this result is in agreement with the calculation given in Ref. 15. We have preferred to limit ourselves here to discussing in detail various particular cases and emphasizing the general picture. It seems to us that the explicit results for the anomaly (3.4) and (4.3) are not only interesting as mathematical results but are of physical interest, in view of future applications to concrete problems in field theory and gravity, both at classical and quantum level. Note also that spectral properties of products of differential operators related to higher spin fields might differ in principal from the properties considered in this paper; we hope to return to this problem elsewhere.

ACKNOWLEDGMENTS

We thank Professor E. Elizalde, Professor L. Vanzo, and Professor S. Zerbini for useful discussions. A. A. Bytsenko wishes to thank CNPq and the Department of Physics of Londrina University for financial support and kind hospitality. The research of A. A. Bytsenko was supported in part by Russian Foundation for Fundamental Research Grant No. 95-02-03568-a and by Russian Universities Grant No. 95-0-6.4-1.

[1] K. Yano, *Differential Geometry on Complex and Almost Complex Spaces* (Pergamon, Oxford, 1965).
[2] S. S. Chern, Bull. Am. Math. Soc. **72**, 167 (1966).
[3] S. Kobayashi, *Transformation Groups in Differential Geometry* (Springer-Verlag, Berlin, 1972).
[4] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, ''University of Einstein Equations for the Ricci Squared Lagrangians,'' gr-qc/9611067 (1996).
[5] A. Borowiec, M. Francaviglia, and I. Volovich, ''Almost Complex and Almost Einstein Manifolds from a Variational Principle,'' Preprint TO-JLL-P 7/96 (1996).

[6] B. L. Reinhart, *Differential Geometry of Foliations* (Springer-Verlag, Berlin, 1983).
[7] A. M. Naveira, Rend. Mat. **3**, 577 (1983).
[8] O. Gil-Medrano and A. M. Naveira, Can. Math. Bull. **26**, 358 (1983).
[9] A. H. Rocamora, Illinois J. Math. **32**, 654 (1988).
[10] E. Calabi, ''Extremal Kähler Metrics,'' in *Seminar on Differential Geometry*, edited by S. T. Yau, Ann. Math. Stud., **102**, 259 (1982).
[11] N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge U. P., Cambridge, 1982).
[12] I. L. Buchbinder, S. D. Odintsov, and I. L. Shapiro, *Effective Action in Quantum Gravity* (Hilger, Bristol, 1992).
[13] M. Kontsevich and S. Vishik, ''Determinants of Elliptic Pseudo-Differential Operators,'' Preprint MPI/94-30 (1994).
[14] M. Kontsevich and S. Vishik, ''Geometry of Determinants of Elliptic Operators,'' hep-th/9406140 (1994).
[15] E. Elizalde, L. Vanzo, and S. Zerbini, ''Zeta-Function Regularization, the Multiplicative Anomaly and the Wodzicki Residue,'' hep-th/9701060 (1997).
[16] M. Wodzicki, ''Non-Commutative Residue. Chapter I,'' in *Lecture Notes in Mathematics*, edited by Yu. I. Manin (Springer-Verlag, Berlin, 1987), **1289**, 320 (1987).
[17] A. Connes, Commun. Math. Phys. **117**, 673 (1988).
[18] D. Kastler, Commun. Math. Phys. **166**, 633 (1995).
[19] A. Connes and J. Lott, Nucl. Phys. B **18**, 29 (1990).
[20] A. Connes, *Non-Commutative Geometry* (Academic, New York, 1994).
[21] A. Connes, ''Gravity Coupled with Matter and the Foundation of non Commutative Geometry,'' hep-th/9603053 (1996).
[22] A. H. Chamseddine and A. Connes, ''The Spectral Action Principle,'' hep-th/9606001 (1996).
[23] A. H. Chamseddin and A. Connes, ''A Universal Action Formula,'' hep-th/9606056 (1996).
[24] C. P. Martin, J. M. Gracia-Bondia, and J. C. Varilly, ''The Standard Model as a Noncommutative Geometry: the Low Energy Regime,'' hep-th/9605001 (1996).
[25] W. Kalau and M. Walze, J. Geom. Phys. **16**, 327 (1995).
[26] T. Ackermann, ''A Note on the Wodzicki Residue,'' funct-an/9506006 (1995).
[27] T. Ackermann and J. Tolksdorf, ''A Generalized Lichnerowich Formula, the Wodzicki Residue and Gravity,'' hep-th/9503152 (1995).
[28] E. Elizalde, ''Complete Determination of the Singularity Structure of Zeta Functions,'' hep-th/9608056 (1996).
[29] J. Mickelsson, ''Wodzicki Residue and Anomalies of Current Algebras,'' hep-th/9404093 (1994).
[30] K. Yano and M. Kon, *Structures on Manifolds*, Series in Pure Mathematics (World Scientific, Singapore, 1984), Vol. 3.
[31] A. L. Besse, *Einstein Manifolds* (Springer-Verlag, New York, 1987).
[32] B. O'Neil, *Semi-Riemannian Geometry* (Academic, New York, 1983).
[33] R. Deszcz, L. Verstralen, and L. Vrancken, Gen. Relativ. Gravit. **23**, 671 (1991).
[34] J. Carot and J. da Costa, Class. Quantum Grav. **10**, 461 (1993).
[35] I. Ya. Aref'eva, K. S. Viswanathan, and I. V. Volovich, ''p-Brane Solutions in Diverse Dimensions,'' hep-th/9609225 (1996).
[36] S. Helgason, *Differential Geometry and Symmetric Spaces*, Pure and Applied Math. Series, (Academic, New York, 1962), Vol. 12.
[37] A. A. Bytsenko, G. Cognola, L. Vanzo, and S. Zerbini, Phys. Rep. **266**, 1 (1996).
[38] F. L. Williams, J. Math. Phys. **38**, 796 (1997).
[39] N. Wallach, Bull. Am. Math. Soc. **82**, 171 (1976).
[40] F. Williams, *Lectures on the Spectrum of* $L^2(\Gamma\backslash G)$, Pitman Rearch Notes in Mathematics (Wiley, New York, 1990), Vol. 242.
[41] A. Selberg, J. India Math. Soc. **20**, 47 (1956).
[42] D. Fried, Ann. Sci. E'cole Norm. Suppl. **10**, 133 (1977).
[43] R. Gangolli, Illinois J. Math. **21**, 1 (1977).
[44] R. Gangolli and G. Warner, Nagoya Math. J. **78**, 1 (1980).
[45] D. Scott, Math. Ann. **253**, 177 (1980).
[46] M. Wakayama, Hiroshima Math. J. **15**, 235 (1985).
[47] F. Williams, Manuscr. Math. **77**, 17 (1992).
[48] F. Williams, ''Some Zeta Functions Attached to $\Gamma\backslash G/K$,'' in *New Developments in Lie Theory and Their Applications*, edited by J. Tirao and N. Wallach, Birkhäuser Progress in Math. Ser. **105**, 163 (1992).
[49] F. L. Williams, ''Meromorphic Continuation of Minakshisundaram-Pleijel Series for Semisimple Lie Groups,'' to be published in Pac. J. Math. (1998).
[50] C. Kassel, ''Le Residue non Commutatif (d'apres M. Wodzicki),'' Semin. Bourbaki, 41 eme ann. 1988-1989, Exp., 708 (1989).
[51] K. Benson, J. Bernstein, and S. Dodelson, Phys. Rev. D **44**, 2480 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Clebsch–Gordan coefficients in the asymptotic limit

Hubert de Guise[a] and David J. Rowe
Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

(Received 24 March 1997; accepted for publication 12 September 1997)

We investigate the graded structure of the tensor product space $V^{\lambda_2}\otimes V^{\lambda_1}$, which arises in the coupling of two irreducible representations λ_1 and λ_2 of a semisimple Lie algebra $\mathcal{G}$, in the limit in which one of the highest weights becomes asymptotically large. The construction of asymptotic coupling coefficients is considered.
© *1998 American Institute of Physics.* [S0022-2488(98)01002-0]

I. INTRODUCTION

It is known that every unitary irreducible representation (unirrep) of a compact semisimple Lie group is characterized by a highest weight $\lambda=(\lambda^1,\lambda^2,\lambda^3,\ldots)$. Often, the unirreps of physical interest of noncompact semisimple Lie groups are likewise characterized by extremal (highest or lowest) weights.

In this paper, we consider the tensor product $V^{\lambda_2}\otimes V^{\lambda_1}$ of modules for two unirreps of highest weights λ_1 and λ_2 of a semisimple Lie algebra $\mathcal{G}$ in the limit in which λ_1 becomes asymptotically large. (A weight λ is said to be asymptotically large if one or more of its components is asymptotically large.) We show that $V^{\lambda_2}\otimes V^{\lambda_1}$ is expressible as a sum of graded subspaces which carry representations of a subalgebra $\mathcal{H}\subset\mathcal{G}$, where $\mathcal{H}$ is uniquely determined by specification of which components of λ_1 are asymptotically large. We also show that the graded subspaces of $V^{\lambda_2}\otimes V^{\lambda_1}$ are simple products of graded subspaces of V^{λ_1} and V^{λ_2}. This factorization implies simple expressions for the Clebsch–Gordan coupling coefficients of $\mathcal{G}$ in the asymptotic limit.

Our motivation for studying asymptotic limits arises because the limiting process is generally associated with group contractions. Thus if a noncompact group of interest is the asymptotic limit of a compact group, one can determine, for instance, Clebsch–Gordan coefficients for the noncompact group by taking the asymptotic limit of Clebsch–Gordan coefficients for the corresponding compact group.

An interesting example of this procedure can be found in the work of W. T. Sharp.[1] Sharp was able to obtain formulas for various products of Bessel functions by considering them as contractions of Legendre functions (see Ref. 2, Sec. 5.71). The coupling coefficients for Bessel functions then emerged naturally as asymptotic limits of ordinary SO(3) coupling coefficients.

Our analysis is appropriate for the less ambitious task of computing some coupling coefficients involving one representation of a group and one representation of its contraction. The need for such coupling coefficients arises in physics when quantum numbers become large. For example, it is well known, in nuclear physics, that the representations of su(3) approach those of the noncompact rigid rotor algebra for large values of the su(3) highest weight.[3] The asymptotic properties of tensor product spaces are also needed for other purposes. For example, they were recently considered by Rowe and Repka[4] in the construction of the coherent shift tensors of Flath and Towber.[5]

The nontrivial unitary representations of a noncompact group are usually infinite dimensional, even when their extremal weights are finite. To avoid confusion, we use the adjective ''asymptotic'' rather than ''infinite'' to describe properties which approach infinity as one or many components of an extremal weight become large. Thus, in the asymptotic limit, the number of basis states in a unirrep of a compact group becomes asymptotically large. Similarly, some chains of weights, which would be finite for finite highest weights, become asymptotic chains in the limit.

[a]Present address: Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128 Succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada.

 © 1998 American Institute of Physics

The paper is organized as follows. Sections II and III A, respectively, deal with the basics of grade space decompositions and provide definitions and examples of asymptotic representations. We give, in Sec. III B, results for highest weight states of semisimple compact Lie algebras, starting with su(2) and followed by the general case of any semisimple compact rank N Lie algebras (with N finite) in III C. In Sec. III D, the theory of the previous section is applied to the Lie algebra su(3). The main result on asymptotic Clebsch–Gordan coefficients is presented in Sec. III E.

Although close in spirit to the corresponding compact cases, the results for the noncompact cases only apply to some of the representations appearing in the decomposition of the tensor product of two unirreps of the positive (or negative) discrete series. The results for noncompact cases are specialized to sp(m,R) and presented in Sec. IV, starting with sp(1,R) and followed by sp(m,R).

Section V deals with the coupling of a finite dimensional, nonunitary representation to an asymptotic unirrep.

II. GRADE-SPACE DECOMPOSITIONS

In calculating Clebsch–Gordan coefficients for a semisimple Lie algebra $\mathscr{G}$, in a basis which reduces a subalgebra $\mathscr{H}\subset\mathscr{G}$, one encounters a need to decompose tensor products of $\mathscr{G}$-invariant spaces into irreducible $\mathscr{G}$- and $\mathscr{H}$-invariant subspaces. Thus, if V^λ is a module for a unirrep of $\mathscr{G}$ with highest weight λ, we need to express $W=V^{\lambda_2}\otimes V^{\lambda_1}$ as a sum

$$W=\sum_{\rho\lambda} W^{\rho\lambda}=\sum_{\rho\lambda\,\gamma\omega} W^{\rho\lambda}_{\gamma\omega}, \tag{1}$$

where ρ indexes the multiplicity of $\mathscr{G}$-invariant subspaces $\{W^{\rho\lambda}\}$ of highest weight λ in W and γ indexes the multiplicity of $\mathscr{H}$-invariant subspaces of highest weight ω in $W^{\rho\lambda}$.

Such a decomposition is simplest when $\mathscr{H}$ is a reductive subalgebra of $\mathscr{G}$ containing the Cartan subalgebra $\mathscr{T}$ and a subset of simple root vectors for $\mathscr{G}$. The Lie algebra $\mathscr{G}$ can then be expressed as a sum of graded subspaces

$$\mathscr{G}=\sum_k \mathscr{G}_k, \tag{2}$$

in which $\mathscr{H}=\mathscr{G}_0$ is the zero-grade subspace and $\mathscr{G}_1$ is the $\mathscr{H}$-invariant subspace of $\mathscr{G}$ containing all simple root vectors not in $\mathscr{H}$. It then follows that

$$[\mathscr{G}_k,\mathscr{G}_l]\subseteq\mathscr{G}_{k+l}. \tag{3}$$

This grading induces a corresponding grading $V=\Sigma_l V_l$ of any module for a representation of $\mathscr{G}$ with the property

$$\mathscr{G}_k: V_l\to V_{k+l}. \tag{4}$$

It will be convenient to introduce the notation $[\omega]$ to denote the grade of a weight ω. Of particular importance is the highest grade subspace $V^\lambda_{[\lambda]}$, since it is the subspace of V^λ containing the highest weight state. This is because the highest grade subspace $V^\lambda_{[\lambda]}$ of an irreducible $\mathscr{G}$-module is an irreducible $\mathscr{H}$-module having the same highest weight.

We denote by $W^k\subset W$ the sum of $\mathscr{G}$-invariant subspaces of $W=V^{\lambda_2}\otimes V^{\lambda_1}$ whose highest weight states are of grade k. Let $W^k=\Sigma_l W^k_l$ be the grade-space decomposition of W^k so that

$$W=\sum_k W^k=\sum_{kl} W^k_l. \tag{5}$$

W^k can in turn be expressed as

$$W^k=\sum_{\rho\lambda} W^{\rho\lambda}, \qquad [\lambda]=k, \tag{6}$$

where $W^{\rho\lambda}$ is as in Eq. (1) but the sum is restricted to highest weights which satisfy $[\lambda]=k$, and

$$W_l^k=\sum_{\rho\lambda} W_l^{\rho\lambda}=\sum_{\rho\lambda}\sum_{\gamma\omega} W_{\gamma\omega}^{\rho\lambda}, \qquad [\lambda]=k\ ,[\omega]=l\ , \tag{7}$$

where $W_{\gamma\omega}^{\rho\lambda}$ is also as in Eq. (1) but the sum is now restricted by the condition $[\omega]=l$.

The grade-space decomposition of W is the first step towards obtaining the full decomposition of Eq. (1). Moreover, if one already knows the Clebsch–Gordan coefficients for the subalgebra $\mathscr{H}$, it is the most important step in the evaluation of asymptotic Clebsch–Gordan coefficients for, as we show in the following, the subspaces W_l^k acquire the very simple expressions

$$W_l^k \to V^{\lambda_2}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]-k+l}\ , \tag{8}$$

in suitable asymptotic limits. A consequence of Eq. (8) is that asymptotic Clebsch–Gordan coefficients for $\mathscr{G}$ can be reduced to Clebsch–Gordan coefficients for $\mathscr{H}$.

III. COMPACT, SEMISIMPLE LIE ALGEBRAS

For a compact, semisimple Lie algebra, the highest weights are dominant integral (i.e., nonnegative and integer valued). [For su(2) it will be convenient to follow the physics convention of labeling irreps by the angular momentum j which can take integer or half-odd integer values. Thus the su(2) angular momentum is half the standard, integer-valued, weight.]

A. Asymptotic representations

We follow the convention of saying that a unirrep is asymptotic if its highest weight is asymptotically large, and simply write $\lambda\to\infty$ if one or more components of λ are asymptotically large. We will denote by $\Delta_+=\{\alpha_1,\alpha_2,\ldots,\alpha_N,\alpha_{N+1},\ldots,\alpha_q\}$ an ordered set of positive roots of $\mathscr{G}$ such that the first N are simple, i.e., $\{\alpha_1,\ldots,\alpha_N\}$ is a basis for Δ_+, and denote by $\{e_1,e_2,\ldots,e_N,e_{N+1},\ldots,e_q\}$ the corresponding root vectors.

If α_k is a root and ω is a weight, then we shall refer to the set of weights of the type $\omega+n\alpha_k$ (n is an integer) that occur within the weight space of a given irrep as an α_k-string. An α_k-string can be labeled by its highest weight. Then, if an α_k-string has highest weight $2j$, its weights are those of an su(2) irrep of angular momentum j. An α_k-string will be said to be asymptotic if and only if $j\to\infty$ as $\lambda\to\infty$.

Let Δ_+ be expressed as the sum of two subsets $\Delta_+=\mathscr{L}_+ +\mathscr{F}_+$, defined as follows. A positive root $\alpha_k\in\Delta_+$ is an element of $\mathscr{L}_+$ if the α_k-string of weights through the highest weight state $|\lambda\rangle$ of V^λ is an asymptotic string, whereas it is an element of $\mathscr{F}_+$ if the α_k-string is not asymptotic. Corresponding sets, $\Delta_-,\mathscr{L}_-$, and $\mathscr{F}_-$, contain the negatives of the roots in $\Delta_+,\mathscr{L}_+$, and $\mathscr{F}_+$, respectively.

If all the components of λ are finite, then $\mathscr{L}_+=\varnothing$ and $\mathscr{F}_+=\Delta_+$. Figure 1(b) is an example of a representation of a rank 2 compact algebra where λ^2, the second component of the highest weight λ, is asymptotically large. In Fig. 1(a), $\alpha_1\in\mathscr{F}_+$, while α_2 and all the other positive roots belong to $\mathscr{L}_+$. In a situation where all the components of the weight are asymptotically large, $\mathscr{F}_+=\varnothing$ and $\mathscr{L}_+=\Delta_+$.

The separation of Δ_+ into subsets $\Delta_+=\mathscr{L}_+ +\mathscr{F}_+$ leads to a grading of $\mathscr{G}$ of the type given in Eq. (2), in which $\mathscr{H}=\mathscr{G}_0$ contains a Cartan subalgebra $\mathscr{T}$ for $\mathscr{G}$ as well as all root vectors whose roots are in $\mathscr{F}_+$ and $\mathscr{F}_-$. The complementary nilpotent subalgebras

$$\mathscr{N}_+=\sum_{k\geqslant 1}\mathscr{G}_k \quad \text{and} \quad \mathscr{N}_-=\sum_{k\geqslant 1}\mathscr{G}_{-k} \tag{9}$$

are spanned by root vectors whose roots are in $\mathscr{L}_+$ and $\mathscr{L}_-$, respectively.

For example, for an asymptotic representation of $\mathscr{G}=$ su(2), $\mathscr{N}_\pm=\mathscr{G}_\pm$ are the one-dimensional algebras spanned by $J_\pm$, respectively, and $\mathscr{H}$ is the Cartan subalgebra of su(2) spanned by J_0. In the example of Fig. 1(a), $\mathscr{H}$ is spanned by $\mathscr{T}$, e_1, and f_1 while $\mathscr{N}_\pm$ contain the remaining positive and negative root vectors, respectively.

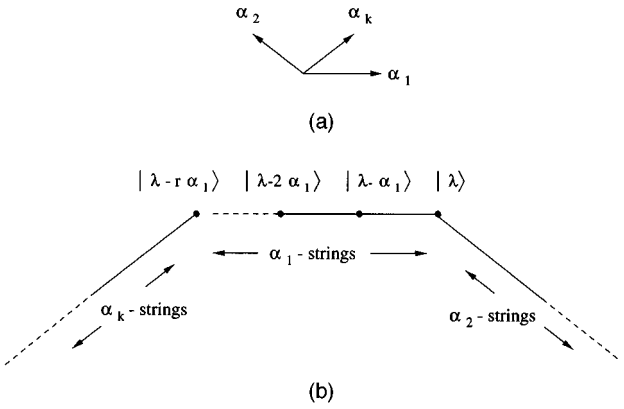

FIG. 1. Example of the weight diagram of a unirrep λ, with $\lambda^2 \to \infty$. (a) Some roots of a rank 2 algebra; (b) An asymptotically large irrep of this algebra.

Throughout this paper, we will assume that the Lie algebra $\mathscr{G}=\Sigma_k \mathscr{G}_k$ is graded such that $\mathscr{G}_1$ is spanned by simple root vectors $\{e_i\}$ for which the corresponding α_i-chains are asymptotically long in the limit where $\lambda_1 \to \infty$. Thus $\mathscr{G}_+$ will always contain positive roots of $\mathscr{G}$ that are asymptotic for the irrep λ_1.

B. Highest grade subspaces of su(2)

The generators of su(2) satisfy the well-known commutation relations

$$[J_0, J_\pm] = \pm J_\pm, \quad [J_+, J_-] = 2J_0, \tag{10}$$

with matrix elements given by

$$J_0|jm\rangle = m|jm\rangle, \quad m = j, j-1, \ldots, -j,$$
$$J_\pm|jm\rangle = \sqrt{(j \mp m)(j \pm m + 1)}\ |j, m+1\rangle. \tag{11}$$

Lemma 1: The highest weight state $|jj\rangle$ of an su(2) unirrep j appearing in the decomposition of the tensor product $j_2 \otimes j_1$ is given, when j_2 is finite and $j_1 \to \infty$, by

$$\lim_{j, j_1 \to \infty} |jj\rangle = |j_2\ \Delta\rangle\ |j_1 j_1\rangle, \qquad \Delta = j - j_1. \tag{12}$$

Proof: Consider the state

$$|jj\rangle = \sum_{m_1 m_2} (j_1\ m_1; j_2\ m_2 | j\ j) |j_2\ m_2\rangle |j_1\ m_1\rangle. \tag{13}$$

Since $m_1 + m_2 = j$, we must have $m_1 = j_1, m_2 = j_2$ when $j = j_1 + j_2$. There is then only one term in the sum and the appropriate coupling coefficient is 1. Thus Eq. (12) holds trivially when $\Delta = j_2$.

When $\Delta = j - j_1 \neq j_2$, we have, in general, more than one term in the sum (13). The equation

$$0 = [\langle j_2 m_2| \langle j_1 m_1|] L_+ |jj\rangle = (j_1\ m_1 - 1; j_2\ m_2 | j\ j) \sqrt{j_1(j_1+1) - m_1(m_1-1)}$$
$$+ (j_1\ m_1; j_2\ m_2 - 1 | j\ j) \sqrt{j_2(j_2+1) - m_2(m_2-1)} \tag{14}$$

implies that the Clebsch–Gordan coefficients satisfy the equation

$$\frac{(j_1\ m_1 - 1; j_2\ m_2 | j\ j)}{(j_1\ m_1; j_2\ m_2 - 1 | j\ j)} = -\sqrt{\frac{j_2(j_2+1) - m_2(m_2-1)}{j_1(j_1+1) - m_1(m_1-1)}}, \tag{15}$$

and hence that

$$\frac{(j_1\ j_1-n;j_2\ m_2|j\ j)}{(j_1\ j_1-n+1;j_2\ m_2-1|j\ j)}=-\sqrt{\frac{j_2(j_2+1)-m_2(m_2-1)}{2nj_1-n(n-1)}}\,, \tag{16}$$

$$\lim_{j_1\to\infty}\frac{(j_1\ j_1-n;j_2\ m_2|j\ j)}{(j_1\ j_1-n+1;j_2\ m_2-1|j\ j)}=-\sqrt{\frac{j_2(j_2+1)-m_2(m_2-1)}{2nj_1}}\,, \tag{17}$$

where n is in the range $0\leqslant n\leqslant 2j_2$. It follows that, for $\Delta<j_2$,

$$\lim_{j_1\to\infty}\frac{(j_1\ j_1-n;j_2\ m_2|j\ j)}{(j_1\ j_1;j_2\ \Delta|j\ j)}<\frac{(-1)^n}{\sqrt{n!}}(2j_1)^{-n/2}[j_2(j_2+1)-\Delta(\Delta+1)]^{n/2}\,. \tag{18}$$

The coupling coefficients are normalized such that

$$\sum_{m_1m_2}|\,(j_1\ m_1;j_2\ m_2|j\ j)|^2=1\,. \tag{19}$$

This sum contains finitely many terms because m_2 ranges over finitely many values. Thus, using Eq. (18), the only nonvanishing contribution to the above sum, in the limit $j_1\to\infty$, is from the term $(j_1\ j_1;j_2\ \Delta|j\ j)$. Furthermore, we must have $|(j_1\ j_1;j_2\ \Delta|j\ j)|=1$. Hence the highest weight state approaches the simple form of Eq. (12), i.e.,

$$|j\ j\rangle\overset{j_1,j\to\infty}{\to}|j_2\ \Delta\rangle|j_1\ j_1\rangle\,, \tag{20}$$

and the proof is complete.

The converse result, that every product state of the type $|j_2\Delta\rangle|j_1j_1\rangle$ becomes a highest weight state $|jj\rangle$ with $j=j_1+\Delta$ as $j_1\to\infty$, is easily shown starting from the general expression

$$|j_2\Delta\rangle|j_1j_1\rangle=\sum_j\,|j,j_1+\Delta\rangle\,(j_1\ j_1;j_2\ \Delta|j\ j_1+\Delta)\,. \tag{21}$$

All of these asymptotic properties can be verified directly from the exact expression[6]

$$(j_1\ m_1;j_2\ m_2|J\ J)=(-1)^{j_1-m_1}\times\sqrt{\frac{(2J+1)!(j_1+j_2-J)!(j_1+m_1)!(j_2+m_2)!}{(j_1+j_2+J+1)!(j_1-j_2+J)!(-j_1+j_2+J)!(j_1-m_1)!(j_2-m_2)!}}\,. \tag{22}$$

The above lemma can be seen as a special case of Eq. (8) by observing first that, for su(2), $\mathscr{H}$ is the u(1) $\subset$ su(2) subalgebra spanned by J_0, so that the graded subspaces of W contain states of a given m value. Since the su(2) coupling is multiplicity free, the subspaces W^k of Eq. (5) contain a single su(2) unirrep and the subspace W^j_j contains only the highest weight state of the unirrep j. Thus, in the notation of Sec. II, Eq. (12) is equivalent to the statement

$$\lim_{j,j_1\to\infty}W^j_j=V^{j_2}_{j-j_1}\otimes V^{j_1}_{j_1}\,. \tag{23}$$

C. Highest grade subspaces of rank *N* Lie algebras

Lemma 2: Let V^λ denote the carrier space for a unirrep of $\mathscr{G}$ of highest weight λ and let W be the tensor product $W=V^{\lambda_2}\otimes V^{\lambda_1}$. Let W^k be the $\mathscr{G}$-invariant subspace of W containing the highest weight states of grade k. Then, the highest grade subspace $W^k_k\subset W^k\subset W$ is given asymptotically by

$$W_k^k \to V^{\lambda_2}_{k-[\lambda_1]} \otimes V^{\lambda_1}_{[\lambda_1]}, \quad \text{as} \quad \lambda_1 \to \infty . \tag{24}$$

Proof: Since W_k^k is the highest grade subspace of the $\mathscr{G}$-invariant subspace $W^k \subset W$, it follows that, if $|\Psi\rangle \in W_k^k$, then $e_i|\Psi\rangle = 0$, $\forall e_i \in \mathscr{N}_+$. Let $\mathscr{A}_i$ denote an su(2) subalgebra of $\mathscr{G}$ with a positive root $\alpha_i \in \mathscr{L}_+$, where we recall that $\mathscr{L}_+$ contains the positive roots of $\mathscr{G}$ which are asymptotic for the irrep λ_1.

Since $e_i|\Psi\rangle = 0$ for any $|\Psi\rangle \in W_k^k$, it follows that W_k^k is spanned by a set of states which are all highest weight states for unirreps of $\mathscr{A}_i$. Moreover, since $\alpha_i \in \mathscr{L}_+$, it follows that such a highest weight state $|\Psi\rangle$ is an asymptotic su(2) highest weight state for the su(2) subalgebra generated by e_i and f_i. Therefore, by *Lemma 1*, $|\Psi\rangle = |\varphi\rangle|\psi\rangle$ is a product of a state $|\varphi\rangle \in V^{\lambda_2}$ and a state $|\psi\rangle \in \sigma_i$, where

$$\sigma_i = \{|\psi\rangle \in V^{\lambda_1}; \quad e_i|\psi\rangle = 0\} . \tag{25}$$

Since this result holds for all $\mathscr{A}_i \subset \mathscr{G}$ with $\alpha_i \in \mathscr{L}_+$, and since

$$V^{\lambda_1}_{[\lambda_1]} = \{|\psi\rangle \in V^{\lambda_1} \ ; e_i|\psi\rangle = 0, \quad \forall \alpha_i \in \mathscr{L}_+\} \tag{26}$$

it follows, in the asymptotic limit, that every state in W_k^k lies in $V^{\lambda_2}_{k-[\lambda_1]} \otimes V^{\lambda_1}_{[\lambda_1]}$. The converse likewise follows from the converse of *Lemma 1*. This completes the proof.

The decomposition of a highest grade subspace W_k^k is a special case of Eq. (7) since, in W_k^k, a highest weight state for a unirrep of $\mathscr{G}$ is also a highest weight state for a unirrep of $\mathscr{H}$. Therefore, we have, in the notation of Eq. (7),

$$W_k^k = \sum_{\rho\lambda} W_\lambda^{\rho\lambda}, \qquad [\lambda] = k . \tag{27}$$

By the same argument, $V^{\lambda_1}_{[\lambda_1]}$ is irreducible under $\mathscr{H}$, since V^{λ_1} is irreducible under $\mathscr{G}$. This implies that $V^{\lambda_1}_{[\lambda_1]} = V^{\lambda_1}_{\lambda_1}$. On the other hand, $V^{\lambda_2}_{k-[\lambda_1]}$ is a sum of $\mathscr{H}$-irreducible subspaces

$$V^{\lambda_2}_{k-[\lambda_1]} = \sum_{\gamma\omega_2} V^{\lambda_2}_{\gamma\omega_2}, \qquad [\omega_2] = k - [\lambda_1]. \tag{28}$$

Now, let $[V^{\lambda_2}_{\gamma\omega_2} \otimes V^{\lambda_1}_{\lambda_1}]_{\alpha\omega}$ denote the $\mathscr{H}$-coupled tensor product of irreducible $\mathscr{H}$-modules, where ω is a highest weight for an $\mathscr{H}$-unirrep and α indexes its multiplicity in the tensor product space $V^{\lambda_2}_{\gamma\omega_2} \otimes V^{\lambda_1}_{\lambda_1}$. It follows from the lemma that

$$W_k^k = \sum_{\rho\lambda} W_\lambda^{\rho\lambda} \to \sum_{\lambda} \sum_{\gamma\omega_2\alpha} [V^{\lambda_2}_{\gamma\omega_2} \otimes V^{\lambda_1}_{\lambda_1}]_{\alpha\lambda}, \qquad [\lambda] = k \tag{29}$$

and

$$W_\lambda^{\rho\lambda} \to \sum_{\gamma\omega_2\alpha} C^\lambda_{\rho,\gamma\omega_2\alpha} [V^{\lambda_2}_{\gamma\omega_2} \otimes V^{\lambda_1}_{\lambda_1}]_{\alpha\lambda}, \tag{30}$$

where C^λ is a $\rho \times \rho$ matrix which combines equivalent unirreps of $\mathscr{H}$.

Now, if τ indexes a basis $\{|\Psi_\tau^{\rho\lambda}\rangle\}$ for $W_\lambda^{\rho\lambda}$, we can write

$$|\Psi_\tau^{\rho\lambda}\rangle = \sum_{\substack{\gamma\omega_2\alpha \\ ij}} C^\lambda_{\rho,\gamma\omega_2\alpha} \ (\lambda_1 \ i; \omega_2 \ j | \alpha\lambda \ \tau) |\varphi^{\lambda_2}_{\gamma\omega_2 j}\rangle |\psi_i^{\lambda_1}\rangle , \tag{31}$$

where i indexes a basis $\{|\psi_i^{\lambda_1}\rangle\}$ for $V^{\lambda_1}_{\lambda_1}$, j indexes a basis $\{|\varphi^{\lambda_2}_{\gamma\omega_2 j}\rangle\}$ for $V^{\lambda_2}_{\gamma\omega_2}$, and $(\lambda_1 \ i; \omega_2 \ j | \alpha\lambda \ \tau)$ is a Clebsch–Gordan coupling coefficient for $\mathscr{H}$.

The matrix C^{λ} depends only on the way in which the multiplicities ρ and α of the $\mathscr{H}$-unirrep λ are separated. Since this separation is arbitrary, there exists a choice of ρ and α that will make C^{λ} diagonal. This implies a remarkable property of asymptotic coupling coefficients for basis states of $\mathscr{G}$ in W_k^k: they can always be chosen to be equal to coupling coefficients for the subalgebra $\mathscr{H}\subset\mathscr{G}$.

D. An example: su(3)

The coupling of two su(3) unirreps provides the simplest application of Eq. (31). For su(3), we have a Cartan subalgebra $\mathscr{T}$ with basis $\{h_1,h_2\}$ and positive roots $\Delta_+=\{\alpha_1,\alpha_2,\alpha_3\}$, where α_1 and α_2 are simple roots, and $\{e_1,e_2,e_3\}$ are the corresponding raising operators. Let $|\psi^{\lambda}_{\beta\sigma}\rangle$ denote a state of weight $\sigma=(\sigma^1,\sigma^2)$ of an su(3) unirrep of highest weight λ, where β indexes multiple occurrences of the weight σ; the label β will be suppressed when not needed. The highest weight state of the unirrep $\lambda=(\lambda^1,\lambda^2)$ will be denoted by $|\lambda\rangle=|\psi^{\lambda}_{\lambda}\rangle$.

We will consider the tensor product of two su(3) unirreps $\lambda_2\otimes\lambda_1$ in the limit in which, say, the second component λ_1^2 of λ_1 is asymptotically large. Then, $\mathscr{T}_+=\{\alpha_1\}$, $\mathscr{L}_+=\{\alpha_2,\alpha_3\}$, and $\mathscr{H}$ is the subalgebra of $\mathscr{G}$ spanned by $\{e_1,f_1,h_1,h_2\}$, where e_1 and f_1 are root vectors corresponding to the roots α_1 and α_{-1}, respectively.

We choose h_1 and h_2 such that

$$[h_1,e_1]=2e_1,\quad [h_1,f_1]=-2f_1,\quad [e_1,f_1]=h_1, \tag{32}$$

and

$$h_1|\psi^{\lambda}_{\beta\sigma}\rangle=\sigma^1|\psi^{\lambda}_{\beta\sigma}\rangle\ ,\quad h_2|\psi^{\lambda}_{\beta\sigma}\rangle=\sigma^2|\psi^{\lambda}_{\beta\sigma}\rangle\ . \tag{33}$$

$\mathscr{H}$ is then the direct sum $\mathscr{H}=\mathrm{su}(2)+\mathrm{u}(1)$, where su(2) is the algebra spanned by e_1,f_1, and h_1, and u(1) is spanned by h_2. Since states of a given grade have identical values of σ^2, we can identify the grade with σ^2.

The subspace $V^{\lambda_1}_{[\lambda_1]}$ contains states with the property

$$V^{\lambda_1}_{[\lambda_1]}=\{|\psi\rangle\in V^{\lambda_1};e_i|\psi\rangle=0,\ \ i=2,3\}\ . \tag{34}$$

States in $V^{\lambda_1}_{[\lambda_1]}$ are generated from the highest weight state $|\lambda_1\rangle$ by repeatedly acting on $|\lambda_1\rangle$ with the lowering operator f_1. Thus they carry a unirrep of angular momentum $j_1=\lambda_1^1/2$ of su(2) $\subset\mathscr{H}$ and a unirrep λ_1^2 of u(1) $\subset\mathscr{H}$. The state $|\lambda_1-r\alpha_1\rangle$ [see Fig. 1(b)], which is obtained by lowering r times from $|\lambda_1\rangle$, can be written, in an $\mathscr{H}=$ su(2)+u(1) basis,

$$|\lambda_1-r\alpha_1\rangle=|\psi^{\lambda_1}_{j_1m_1\lambda_1^2}\rangle\ , \tag{35}$$

where $m_1=j_1-r$.

The subspace $V_l^{\lambda_2}\subset V^{\lambda_2}$ has decomposition into $\mathscr{H}$-irreducible subspaces given, in the notation of Eq. (28), by

$$V_l^{\lambda_2}=\sum_{j_2}\ V^{\lambda_2}_{j_2\sigma^2}\ ,\quad \sigma^2=l\ , \tag{36}$$

where the multiplicity label γ of Eq. (28) is suppressed because the unirrep $j_2\sigma^2$ occurs at most once in $V_l^{\lambda_2}$. An su(2)+u(1) basis for $V^{\lambda_2}_{j_2\sigma^2}$ is then given by $\{|\varphi^{\lambda_2}_{j_2m_2\sigma^2}\rangle\ ;m_2=-j_2,\ldots,j_2.\}$.

From *Lemma 2*, we know that a basis for W_k^k is given by products of the type $|\varphi^{\lambda_2}_{j_2m_2\sigma^2}\rangle|\psi^{\lambda_1}_{j_1m_1\lambda_1^2}\rangle$, with $k=\lambda_1^2+\sigma^2$. These states can be combined to form a good su(2)+u(1)-coupled basis $\{|\Psi_{JMk}\rangle\}$ for W_k^k with

$$|\Psi_{JMk}\rangle=[|\varphi^{\lambda_2}_{j_2\sigma^2}\rangle\otimes|\psi^{\lambda_1}_{j_1\lambda_1^2}\rangle]^{Jk}_M=\sum_{m_1m_2}\ (j_1\ m_1;j_2\ m_2|J\ M)|\varphi^{\lambda_2}_{j_2m_2\sigma^2}\rangle|\psi^{\lambda_1}_{j_1m_1\lambda_1^2}\rangle\ , \tag{37}$$

where $(j_1\ m_1;j_2\ m_2|J\ M)$ is an ordinary su(2) coupling coefficient. The u(1) coupling is fulfilled by the requirement that $\lambda_1^2+\sigma^2=k$.

Since W_k^k is a highest grade subspace, an su(2)+u(1) highest weight state $|\Psi_{JJk}\rangle$ is also a highest weight state for an su(3) unirrep of highest weight $\lambda=(\lambda^1=2J,\lambda^2=k)$. We then have the identification $|\Psi_{JJk}\rangle=|\lambda\rangle$.

Although the su(2)+u(1) coupling is multiplicity-free, there may be more than one $|\Psi_{JJk}\rangle$ (and hence more than one occurrence of $|\lambda\rangle$) because, in general, $V_l^{\lambda_2}$ contains more than one value of j_2 such that the coupling $j_2\otimes j_1\rightarrow J$ exists. Since there is one copy of $|\Psi_{JJk}\rangle$ for every j_2 satisfying the above condition, an obvious way to distinguish these multiple copies of $|\Psi_{JJk}\rangle$ is to label them with j_2. Highest weight states for the su(2)+u(1) unirrep Jk will henceforth be denoted $|\Psi_{j_2JJk}\rangle$.

If ρ labels multiple occurrences of the su(3) highest weight state $|\lambda\rangle$, the set $\{|\rho\lambda\rangle\}$ may still differ from $\{|\Psi_{j_2JJk}\rangle\}$ by an arbitrary unitary transformation; this is the matrix C^λ of Eq. (31). Thus, the identification $\rho\leftrightarrow j_2$, which provides a convenient resolution of the multiplicities in the $\lambda_2\otimes\lambda_1\rightarrow\lambda$ coupling, makes this matrix diagonal. Furthermore, for this choice, we have

$$|j_2\lambda\rangle=\sum_{m_1m_2}(j_1\ m_1;j_2\ m_2|J\ J)|\varphi^{\lambda_2}_{j_2m_2\sigma^2}\rangle|\psi^{\lambda_1}_{j_1m_1\lambda_1^2}\rangle\ , \tag{38}$$

where $2J=\lambda^1$ and $\lambda_1^2+\sigma^2=\lambda^2$. Other states with $M\neq J$ are obtained by lowering with f_1, and we have, in general,

$$|j_2\lambda\,\tau\rangle=\sum_{m_1m_2}(j_1\ m_1;j_2\ m_2|J\ M)|\varphi^{\lambda_2}_{j_2m_2\sigma^2}\rangle|\psi^{\lambda_1}_{j_1m_1\lambda_1^2}\rangle\ , \tag{39}$$

where $2M=\tau$.

Thus, in the limit where one component of λ_1 becomes asymptotically large, we find that su(3) coupling reduces to su(2)+u(1) coupling, with

$$(\lambda_1j_1m_1\sigma_1^2,\lambda_2j_2m_2\sigma_2^2|\rho\lambda\,\tau)\rightarrow\delta_{\rho,j_2}\delta_{\sigma_1^2,\lambda_1^2}(j_1m_1,j_2m_2|JM)\ . \tag{40}$$

This conclusion is confirmed when we take the limit of the expressions obtained, for instance, by Hecht[7] for some su(3) coupling coefficients (cf. also Rowe and Repka[8]).

Suppose now that both components $(\lambda_1^1,\lambda_1^2)$ of λ_1 are asymptotically large. Since $2j_1=\lambda_1^1$ and $2J=\lambda^1$, the coupling coefficient appropriate to this case is obtained from Eq. (40) by taking the limit of the su(2) coefficient:

$$\lim_{\lambda_1,\lambda\rightarrow\infty}(\lambda_1j_1m_1\sigma_1^2,\lambda_2j_2m_2\sigma_2^2|\rho\lambda\,\tau)\rightarrow\delta_{\rho,j_2}\delta_{\sigma_1^2,\lambda_1^2}\ \times\ \lim_{j_1,J\rightarrow\infty}(j_1m_1,j_2m_2|JM)\ . \tag{41}$$

In the next section, we show how Clebsch–Gordan coefficients of the type $(j_1m_1,j_2m_2|JM)$ can be evaluated in the $j_1\rightarrow\infty$ limit.

E. Asymptotic Clebsch–Gordan coefficients

Let $\mathcal{G}$ be a compact semisimple Lie algebra of rank N, and let V^{λ_1}, V^{λ_2} be defined as before.

Lemma 3: Let $|\varphi\rangle\in V^{\lambda_2}$, $|\psi\rangle\in V^{\lambda_1}_{[\lambda_1]-p}$, and assume that p is finite. Then, in the limit in which $\lambda_1\rightarrow\infty$, the state $|\Psi\rangle=|\varphi\rangle|\psi\rangle\in V^{\lambda_2}\otimes V^{\lambda_1}$ is such that

$$f_k|\Psi\rangle\rightarrow|\varphi\rangle[f_k|\psi\rangle]\ ,\quad\forall f_k\in\mathcal{N}_-\ . \tag{42}$$

Proof: Recall that $\mathcal{N}_-$ contains root vectors whose corresponding roots are asymptotic for $\lambda_1\rightarrow\infty$. Then, we have

$$f_k|\Psi\rangle=[f_k|\varphi\rangle]|\psi\rangle+|\varphi\rangle[f_k|\psi\rangle]\ . \tag{43}$$

We may assume, without loss of generality, that $|\varphi\rangle$ and $|\psi\rangle$ are normalized states so that $\||\varphi\rangle\|^2=\||\psi\rangle\|^2=1$. Then

$$\|f_k|\Psi\rangle\|^2=\|f_k|\varphi\rangle\|^2+\|f_k|\psi\rangle\|^2 \ . \tag{44}$$

Now

$$\|f_k|\psi\rangle\|^2=\langle\psi|e_kf_k|\psi\rangle=\langle\psi|[e_k,f_k]|\psi\rangle+\|e_k|\psi\rangle\|^2=\langle\psi|h_k|\psi\rangle+\|e_k|\psi\rangle\|^2 \ , \tag{45}$$

which, since $\langle\psi|h_k|\psi\rangle\to\infty$ as $\lambda_1\to\infty$, implies that

$$\|f_k|\psi\rangle\|^2\to\infty \quad \text{as} \quad \lambda_1\to\infty \ . \tag{46}$$

On the other hand, $\|f_k|\varphi\rangle\|^2$ remains finite and the lemma is shown.

Two important properties for the norms of asymptotic states follow from this lemma. First, set $|\Psi_{ij}\rangle=|\varphi_i\rangle|\psi_j\rangle$, where i labels a basis $\{|\varphi_i\rangle\}$ for V^{λ_2}, and j labels a basis $\{|\psi_j\rangle\}$ for $V^{\lambda_1}_{[\lambda_1]-p}$, and consider the linear combination

$$|\Theta\rangle=\sum_{ij} c_{ij}|\Psi_{ij}\rangle \ , \tag{47}$$

where c_{ij} are complex coefficients satisfying $\Sigma_{ij}c_{ij}c^*_{ij}=1$, but otherwise arbitrary. Then, by *Lemma 3*, it follows that $\|f_k|\psi_j\rangle\|$ is independent of j. [*Proof:* act on $|\Theta\rangle$ to construct the normalized state $f_k|\Theta\rangle/\|f_k|\Theta\rangle\|$ in terms of $|\varphi_i\rangle$ and $|\psi_j\rangle$. Since $|\varphi_i\rangle$ is normalized, and since $\Sigma_{ij}c_{ij}c^*_{ij}=1$, we deduce that $\|f_k|\psi_j\rangle\|=\|f_k|\Theta\rangle\|\forall j$.]

Next, it follows immediately from Eq. (45) that $\|f_k|\Theta\rangle\|=\|f_k|\psi_j\rangle\|$ is also independent of any component of λ or λ_1 that is asymptotically large. [*Proof:* Eq. (45) is unchanged if we use the unirrep $\tilde{\lambda}_1$ whose highest weight is related to the highest weight of λ_1 by $\tilde{\lambda}_1=\lambda_1+\Delta$, where $\Delta=(\delta_1,\ldots,\delta_n)$ contains only finite integers.]

Theorem 1: Let p be a finite integer. Then,

$$W^k_{k-p}\to V^{\lambda_2}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]-p}, \quad \text{as} \quad \lambda_1\to\infty \ . \tag{48}$$

Proof: Let $|\varphi\rangle\in V^{\lambda_2}_{k-[\lambda_1]}$ and $|\psi\rangle\in V^{\lambda_1}_{[\lambda_1]-p}$. Assume the theorem holds for some p, and act on $|\Psi\rangle=|\varphi\rangle|\psi\rangle\in W^k_{k-p}$ with any $f_r\in\mathcal{G}_{-q}, q>0$. Since W^k is $\mathcal{G}$-invariant, we have, by Eq. (4), $f_r|\Psi\rangle\in W^k_{k-p-q}$, and, by *Lemma 3* and Eq. (4),

$$f_r|\Psi\rangle\to|\varphi\rangle[f_r|\psi\rangle]\in V^{\lambda_2}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]-p-q} \ . \tag{49}$$

Hence, if the theorem holds for p, it holds for $p+q$. The seed of the recursion is Lemma 2, which is Theorem 1 for $p=q=0$. Going over all f_r in all $\mathcal{G}_{-q}$ completes the proof.

Consider, for example, the su(2) coupling $j_2\otimes j_1$ in the limit in which $j_1\to\infty$ but j_2 remains finite. Application of the theorem implies that, for a finite value of p, the state

$$|j,j-p\rangle=\sum_{m_1,m_2}(j_1\ m_1;j_2\ m_2|j\ j-p)|j_2,m_2\rangle|j_1m_1\rangle \tag{50}$$

becomes, in the $j_1\to\infty$ limit,

$$|j,j-p\rangle\to|j_2,j-j_1\rangle|j_1,j_1-p\rangle \ . \tag{51}$$

It follows that, for finite p, the asymptotic SU(2) coupling coefficients are given by

$$(j_1\ m_1;j_2\ m_2|j\ j-p)\to\delta_{m_1,j_1-p}\quad\delta_{m_2,j-j_1} \ . \tag{52}$$

For an arbitrary compact semisimple Lie algebra, the highest grade states in the tensor product space $V^{\lambda_2}\otimes V^{\lambda_1}$ become, by Theorem 1, of the form

$$|\Psi^{\rho\lambda}_{\lambda\tau}\rangle \to \sum_{\omega_2\sigma} C^{\lambda}_{\rho,\gamma_2\omega_2\sigma}\ [|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle \otimes |\psi^{\lambda_1}_{\lambda_1}\rangle]_{\sigma\lambda\tau}, \tag{53}$$

in the asymptotic limit $\lambda_1 \to \infty$, where C^{λ} is a unitary transformation. If a particular resolution of the multiplicity is chosen, equivalent to setting $\rho \equiv (\gamma_2\omega_2\sigma)$, such that C^{λ} is an identity matrix, then we have

$$|\Psi^{\rho\lambda}_{\lambda\tau}\rangle \to [|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle \otimes |\psi^{\lambda_1}_{\lambda_1}\rangle]_{\sigma\lambda\tau}, \quad \rho \equiv (\gamma_2\omega_2\sigma) \tag{54}$$

and the $\mathscr{H}$-reduced asymptotic CG coefficients

$$(\lambda_1\gamma_1\omega_1, \lambda_2\gamma_2\omega_2 \| \rho\lambda\lambda)_{\sigma} \to \delta_{\omega_1,\lambda_1}\delta_{\rho,\gamma_2\omega_2\sigma}, \tag{55}$$

in accordance with Eq. (31).

To obtain the asymptotic limits of coefficients $(\lambda_1\gamma_1\omega_1, \lambda_2\gamma_2\omega_2 \| \rho\lambda\,\gamma\omega)_{\sigma}$ for finite values of $[\lambda]-[\omega]$, let the state $|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle$ be expressed in the form

$$|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle = [P_n(f) \otimes |\Psi^{\rho\lambda}_{\lambda}\rangle]_{\alpha\omega\tau}, \tag{56}$$

where $P_n(f)$ is an $\mathscr{H}$-tensor of highest weight n whose components are polynomials in the elements of $\mathscr{N}_-$ and we identify the multiplicity index γ with the pair of indices $(n\alpha)$. Then, in the asymptotic limit,

$$|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle \to [P_n(f) \otimes [|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle \otimes |\psi^{\lambda_1}_{\lambda_1}\rangle]_{\sigma\lambda}]_{\alpha\omega\tau}. \tag{57}$$

Let $\phi(\lambda_1, \omega_2; \sigma\lambda)$ be the phase factor for which

$$[|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle \otimes |\psi^{\lambda_1}_{\lambda_1}\rangle]_{\sigma\lambda} = \phi(\lambda_1, \lambda_2; \sigma\lambda)[|\psi^{\lambda_1}_{\lambda_1}\rangle \otimes |\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle]_{\sigma\lambda}. \tag{58}$$

Using Theorem 1 and setting $\rho \equiv (\gamma_2\omega_2\sigma)$, we have

$$\begin{aligned} |\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle &\to \phi(\lambda_1, \omega_2; \sigma\lambda) \sum_{\beta\kappa\omega_1} U(\omega_2\lambda_1\omega n; \lambda\sigma\alpha, \omega_1\beta\kappa)[[P_n(f) \otimes |\psi^{\lambda_1}_{\lambda_1}\rangle]_{\beta\omega_1} \otimes |\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle]_{\kappa\omega\tau}, \\ &\to \phi(\lambda_1, \omega_2; \sigma\lambda) \sum_{\beta\kappa\omega_1} \phi(\omega_2, \omega_1; \kappa\omega) U(\omega_2\lambda_1\omega n; \lambda\sigma\alpha, \omega_1\beta\kappa)[|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle \otimes |\psi^{\lambda_1}_{n\beta\omega_1}\rangle]_{\kappa\omega\tau}, \end{aligned} \tag{59}$$

where $U(\omega_2\lambda_1\omega n; \lambda\sigma\alpha, \omega_1\beta\kappa)$ is a Racah recoupling coefficient for $\mathscr{H}$, and κ labels multiple copies of ω in the coupling $\omega_1 \otimes \omega_2$. In deriving Eq. (59), we have used properties of the norms of asymptotic states discussed as corollaries of Lemma 3. Furthermore, we use $(n\beta)$ to label multiple occurrences of γ_1. Thus we find

$$\begin{aligned} (\lambda_1\gamma_1\omega_1, \lambda_2\gamma_2\omega_2 \| \rho\lambda\,\gamma\omega)_{\sigma} \to \sum_{\kappa}\ & \phi(\lambda_1, \omega_2; \sigma\lambda)\phi(\omega_2, \omega_1; \alpha\omega)\ \ U(\omega_2\lambda_1\omega n; \lambda\sigma\alpha, \omega_1\beta\kappa) \\ & \times\ M^{\omega}_{\gamma\kappa}\ \times\ \delta_{\rho,\gamma_2\omega_2\sigma}\ \ \delta_{\gamma,n\alpha}\ \ \delta_{\gamma_1,n\beta}, \quad \text{as}\ \ \lambda_1 \to \infty\ , \end{aligned} \tag{60}$$

where the matrix M^{ω} combines the equivalent $\mathscr{H}$-representations ω. Since this matrix depends only on the arbitrary way in which the multiplicities γ and κ of the $\mathscr{H}$-unirrep ω are separated, we set $\kappa = \gamma$ and choose the phases so that M^{ω} is the unit matrix. It then follows that the $\mathscr{H}$-reduced asymptotic coupling coefficients for basis states of $\mathscr{G}$ in $W^{k}_{[\lambda]-[\omega]}$, for finite values of $[\lambda] - [\omega]$, are given simply by a recoupling coefficient for the subalgebra $\mathscr{H}$ times some phases, i.e.,

$$(\lambda_1\gamma_1\omega_1,\lambda_2\gamma_2\omega_2\|\rho\lambda\,\gamma\omega)_\sigma \to \phi(\lambda_1,\omega_2;\sigma\lambda)\,\phi(\omega_2,\omega_1;\alpha\omega) \times U(\omega_2\lambda_1\omega n;\lambda\sigma\alpha,\omega_1\beta\gamma)\,\delta_{\rho,\gamma_2\omega_2\sigma}\ \ \delta_{\gamma,n\alpha}\ \ \delta_{\gamma_1,n\beta}, \quad \text{as } \lambda_1\to\infty. \tag{61}$$

Combining Eq. (59) with Eq. (61), we finally obtain the simple expression

$$|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle \to \phi(\lambda_1,\omega_2;\sigma\lambda)\sum_{\beta\omega_1}\phi(\omega_2,\omega_1;\kappa\omega)U(\omega_2\lambda_1\omega n;\lambda\sigma\alpha,\omega_1\beta\omega) \times[|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle\otimes|\psi^{\lambda_1}_{n\beta\omega_1}\rangle]_{\gamma\omega\tau}\delta_{\rho,\gamma_2\omega_2\sigma}\ \ \delta_{\gamma,n\alpha}, \quad \text{as } \lambda_1\to\infty\ . \tag{62}$$

IV. THE REAL SYMPLECTIC sp(m,R) ALGEBRAS

In this section, we discuss the generalization of the above results to noncompact semisimple Lie algebras. For simplicity we restrict considerations to sp(m,R), although the results are much more widely applicable. We also restrict considerations to products of harmonic series of representations.[9,10] These representations, which include the positive discrete series, have lowest weights but no highest weights. Thus we consider subspaces of $W=V^{\lambda_2}\otimes V^{\lambda_1}$ that are lowest (rather than highest) in grade. $W^k\subset W$ is now the sp(m,R)-invariant subspace containing the lowest weight states of grade k, and W^k_{k+p} is the subspace of W^k of grade $k+p$.

A. Factorization for sp(1,R)

The elements of sp(1,R) satisfy

$$[J_-,J_+]=2J_0, \quad [J_0,J_\pm]=\pm J_\pm, \tag{63}$$

with matrix elements given by

$$\begin{aligned} J_0|jm\rangle &= m|jm\rangle\ , \quad m=j,j+1,j+2,\ldots\ , \\ J_+|jm\rangle &= \sqrt{(m+j)(m-j+1)}\,|j,m+1\rangle\ , \\ J_-|jm\rangle &= \sqrt{(m-j)(m+j-1)}\,|j,m-1\rangle\ , \end{aligned} \tag{64}$$

where $j=\frac{1}{4},\frac{3}{4},\frac{5}{4},\ldots$. Now, consider the coupling of two harmonic series unirreps $j_2\otimes j_1$ with $j_1\to\infty$ and j_2 finite. It can be verified, e.g., by comparing the number of states of weight m $=m_1+m_2$ with the number of states of weight $m-1$, that

$$j_2\otimes j_1=(j_1+j_2)\oplus(j_1+j_2+1)\oplus(j_1+j_2+2)\oplus\cdots\ . \tag{65}$$

Next, consider the lowest weight state $|jj\rangle$ of a subrepresentation in Eq. (65), with $j=j_1+j_2$ $+s$, where j_2 and $s\geqslant 0$ are finite, in the limit $j_1\to\infty$. Write

$$|jj\rangle=\sum_{m_1m_2}\ (j_1\ m_1;j_2\ m_2|j\ j)\ \ |j_2m_2\rangle|j_1m_1\rangle\ , \tag{66}$$

where $(j_1\ m_1;j_2\ m_2|j\ j)$ is now an sp(1,R) coupling coefficient. If we use $J_-|jj\rangle=0$, we get

$$\begin{aligned} 0=&\ (j_1\ m_1+1;j_2\ m_2-1|j\ j)\ \ \sqrt{(m_1+1-j_1)(m_1+j_1)} \\ &+(j_1\ m_1;j_2\ m_2|j\ j)\ \ \sqrt{(m_2-j_2)(m_2+j_2-1)}\ . \end{aligned}$$

This equation can then be rewritten as

$$\frac{(j_1\ m_1+1;j_2\ m_2-1|j\ j)}{(j_1\ m_1;j_2\ m_2|j\ j)}=-\sqrt{\frac{(m_2-j_2)(m_2+j_2-1)}{(m_1+1-j_1)(m_1+j_1)}}\ . \tag{67}$$

Since $m_1+j_1\geqslant 2j_1$ and $j_1\to\infty$ while j_2 and m_2 remain finite, it follows that

$$\frac{(j_1\ m_1+1;j_2\ m_2-1|j\ j)}{(j_1\ m_1;j_2\ m_2|j\ j)}\to 0 \quad \text{as } j_1\to\infty\ . \tag{68}$$

If we now suppose that $|j,j+\delta\rangle\to|j_2,j-j_1\rangle|j_1,j_1+\delta\rangle$ as $j_1\to\infty$, when δ is some finite integer, then, by acting on $|j,j+\delta\rangle$ with J_+, we find, on the one hand,

$$J_+|j,j+\delta\rangle=\sqrt{(2j+\delta)(\delta+1)}|j,j+\delta+1\rangle\ , \tag{69}$$

while, on the other hand,

$$\begin{aligned}J_+[|j_2,j-j_1\rangle|j_1,j_1+\delta\rangle]&=\sqrt{(2j_1+\delta)(\delta+1)}|j_2,j-j_1\rangle|j_1,j_1+\delta+1\rangle\\&\quad+\sqrt{(2j_2+s)(s+1)}|j_2,j-j_1+1\rangle|j_1,j_1+\delta\rangle\ .\end{aligned} \tag{70}$$

Comparison of these equations gives

$$\begin{aligned}|j,j+\delta+1\rangle&=\sqrt{\frac{(2j_1+\delta)}{(2j+\delta)}}|j_2,j-j_1\rangle|j_1,j_1+\delta+1\rangle\\&\quad+\sqrt{\frac{(2j_2+s)(s+1)}{(2j+\delta)(\delta+1)}}|j_2,j-j_1+1\rangle|j_1,j_1+\delta\rangle\ ,\end{aligned} \tag{71}$$

and implies that

$$|j,j+\delta+1\rangle\to|j_2,j-j_1\rangle|j_1,j_1+\delta+1\rangle\ ,\quad \text{as}\ \ j_1\to\infty\ . \tag{72}$$

Hence we have shown by recursion that the following holds:

Lemma 4: Let $j=j_1+j_2+s$ with $s\geqslant 0$. Then, if j_2 and s are finite,

$$|j,j+\delta\rangle\to|j_2,j-j_1\rangle|j_1,j_1+\delta\rangle\ ,\quad \text{as}\ \ j_1\to\infty\ . \tag{73}$$

B. Asymptotic unirreps of sp(m,R)

Every harmonic series representation of sp(m,R) is contained within the space of some A-particle harmonic oscillator in m dimensions.[9,10] For these representations, the sp(m,R) algebra is realized as the set of bilinear products of creation and destruction operators:

$$\begin{aligned}A_{ij}&=\sum_{a=1}^{\mathrm{A}} b_{ai}^\dagger b_{aj}^\dagger,\\C_{ij}&=\sum_{a=1}^{\mathrm{A}}\frac{1}{2}\,(b_{ai}^\dagger b_{aj}+b_{aj}b_{ai}^\dagger)\ ,\\B_{ij}&=\sum_{a=1}^{\mathrm{A}} b_{ai}b_{aj},i,\quad j=1,\ldots,m\ .\end{aligned} \tag{74}$$

Sp(m,R) and u(m) have a common Cartan algebra, spanned by $\{C_{ii},i=1,\ldots,m\}$. Thus a u(m) unirrep is labeled by a highest weight $\lambda=(\lambda^1,\ldots,\lambda^m)$ where λ^i is an eigenvalue of C_{ii} and $\lambda^i\geqslant\lambda^{i+1}$. However, the sp($m$,R) unirreps we consider have no highest weight. The carrier spaces of these sp(m,R) unirreps comprise infinitely many u(m) invariant subspaces each labeled by a u(m) highest weight. The lowest of these u(m) highest weights uniquely characterizes the sp(m,R) unirrep and will be referred to as the lowest weight of the sp(m,R) irrep.

We will consider states of sp(m,R) unirreps in the limit where $\lambda^m\to\infty$. This implies that all the other components of λ are also asymptotically large.

There are two types of sp(m,R) raising operators: the compact su(m) raising operators $\{C_{ij};1\leqslant j<i\leqslant m\}$ and the noncompact raising operators $\{A_{ij};1\leqslant i,j\leqslant m\}$. The action of an A_{ij} operator connects weights belonging to an sp(1,R) chain. States of an sp(1,R) chain carry a unirrep

of an sp(1,R) subalgebra of sp(m,R). If such a unirrep is labeled by j, we define the corresponding sp(1,R) chain of weights to be an asymptotic chain if and only if $j\to\infty$ as $\lambda\to\infty$. Moreover, we note that, when $\lambda^m\to\infty$, all the $\{A_{ij}\}$ operators generate asymptotic chains. Thus, when $\lambda^m\to\infty$, all the noncompact roots are in $\mathscr{S}_+$.

Chains of weights associated with the compact su(m) roots are conveniently analyzed by reexpressing a u(m) weight λ as an su(m)+u(1) weight with components $(\lambda^1-\lambda^2,\lambda^2-\lambda^3,\ldots,\lambda^{m-1}-\lambda^m)\lambda^m$. The chains of weights associated with the compact su(m) roots then become asymptotic when the differences $\lambda^i-\lambda^{i+1}$ of consecutive components of a u(m) weight become large.

We assume, for the moment, that $\lambda^i-\lambda^{i+1}$ remains finite for all i, so that $\mathscr{T}_+$ contains all the positive su(m) roots; i.e., $\mathscr{T}_+$ contains the compact positive roots and $\mathscr{S}_+$ contains the noncompact positive roots. We then have $\mathscr{K}=$ u(m), and sp(m,R) (or, more precisely, its complex extension) has the graded decomposition

$$\mathrm{sp}(m,\mathrm{R})=\mathscr{G}_0+\mathscr{G}_+ +\mathscr{G}_-\,, \tag{75}$$

where $\mathscr{G}_0$ is the u(m) subalgebra spanned by the $\{C_{ij}\}$, $\mathscr{G}_+$ is spanned by the raising operators $\{A_{ij}\}$, and $\mathscr{G}_-$ is spanned by the lowering operators $\{B_{ij}\}$. The grade is conveniently identified with the eigenvalue of the operator

$$\hat{N}=\sum_{i=1}^{m} C_{ii}\,. \tag{76}$$

The lowest grade subspace $V^\lambda_{[\lambda]}$ of an sp(m,R)-irreducible vector space V^λ is defined as the subset of states that are annihilated by the elements of $\mathscr{G}_-$:

$$V^\lambda_{[\lambda]}=\{|\psi\rangle\in V^\lambda;\ \ B_{ij}|\psi\rangle=0,\ \forall B_{ij}\in\mathscr{G}_-\}\ . \tag{77}$$

If one or more of the su(m) labels $\lambda^i-\lambda^{i+1}$ becomes asymptotically large, then $\mathscr{T}_+$ and $\mathscr{S}_+$ must be shrunk and expanded accordingly. However, as the process of taking $\lambda^m\to\infty$ ''commutes'' with the process of taking any su(m) label to ∞, we will assume henceforth that all the $\lambda^i-\lambda^{i+1}$ are finite, knowing that the extra simplifications that arise should one or more su(m) labels become asymptotically large can be made once the analysis of the large-λ^m limit has been completed.

C. Lowest grade subspaces of sp(m,R)

Lemma 5: Let V^λ denote the carrier space for a unirrep of sp(m,R) of lowest weight λ. Let $W^k\subset W$ be the sp(m,R) invariant subspace of $W=V^{\lambda_2}\otimes V^{\lambda_1}$ containing the lowest weight states of grade k, where $k=[\lambda_1]+[\lambda_2]+s$, with s a finite integer. Then, the lowest grade subspace $W^k_k\subset W^k$ is given asymptotically by

$$W^k_k\to V^{\lambda_2}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]}\,,\qquad \text{as } \lambda_1^m\to\infty\ . \tag{78}$$

Proof: The proof parallels that given for Lemma 2 and is omitted.

It is possible to use this lemma to derive asymptotic sp(m,R) coupling coefficients. Since $\mathscr{K}=$ u(m) in the $\lambda_1^m\to\infty$ limit, it follows, in the notation of Eq. (31), that a basis $\{|\Psi^{\rho\lambda}_\tau\rangle\}$ for $W^{\rho\lambda}_\lambda\subset W^k_k$ is given by the states

$$|\Psi^{\rho\lambda}_\tau\rangle=\sum_{\substack{\gamma\omega_2\alpha\\ ij}} C^\lambda_{\rho,\gamma\omega_2\alpha}\ (\lambda_1\ i;\omega_2\ j|\alpha\omega\ \tau)|\varphi^{\lambda_2}_{\gamma\omega_2 j}\rangle|\psi^{\lambda_1}_i\rangle, \tag{79}$$

where α labels the multiplicity of the u(m) coupling $\omega_2\otimes\lambda_1\to\lambda$, i indexes a basis $\{|\psi^{\lambda_1}_i\rangle\}$ for the u(m) unirrep λ_1 in $V^{\lambda_1}_{[\lambda_1]}$, j indexes a basis $\{|\varphi^{\lambda_2}_{\gamma\omega_2 j}\rangle\}$ for the γth copy of the u(m) unirrep ω_2 in $V^{\lambda_2}_{k-[\lambda_1]}$, and $(\lambda_1\ i;\omega_2\ j|\alpha\omega\ \tau)$ is a u(m) Clebsch–Gordan coupling coefficient.

The derivation of Eq. (79) is identical to the derivation of Eq. (31) and therefore omitted.

D. Asymptotic states of sp(m,R)

Lemma 6: Let $|\psi\rangle \in V^{\lambda_1}$ and $|\varphi\rangle \in V^{\lambda_2}_{[\lambda_2]+q}$ be normalized states, where q is a finite positive integer. Then, in the $\lambda_1^m \to \infty$ limit, the state $|\Psi\rangle = |\varphi\rangle|\psi\rangle \in V^{\lambda_2} \otimes V^{\lambda_1}$ satisfies

$$A_{ij}|\Psi\rangle \to |\varphi\rangle[A_{ij}|\psi\rangle] \ , \quad \forall ij \ . \tag{80}$$

Proof: We have

$$A_{ij}|\Psi\rangle = [A_{ij}|\varphi\rangle] \ |\psi\rangle + |\varphi\rangle[A_{ij}|\psi\rangle] \tag{81}$$

and, since $|\varphi\rangle$ and $|\psi\rangle$ are normalized states,

$$\|A_{ij}|\Psi\rangle\| = \|A_{ij}|\varphi\rangle\| + \|A_{ij}|\psi\rangle\| \ . \tag{82}$$

To evaluate $\|A_{ij}|\psi\rangle\|$, observe that

$$\begin{aligned} \|A_{ij}|\psi\rangle\|^2 &= \langle\psi|[B_{ij},A_{ij}]|\psi\rangle + \|B_{ij}|\psi\rangle\|^2 \\ &= \langle\psi|C_{ii}+C_{jj}+2\delta_{ij}C_{ii}|\psi\rangle + \|B_{ij}|\psi\rangle\|^2 \ . \end{aligned} \tag{83}$$

Since

$$\langle\psi|C_{ii}|\psi\rangle \geqslant \langle\psi|C_{mm}|\psi\rangle \geqslant \lambda_1^m , \tag{84}$$

it follows that $\|A_{ij}|\psi\rangle\|^2 \to \infty$ as $\lambda_1^m \to \infty$. On the other hand, for $|\varphi\rangle \in V^{\lambda_2}_{[\lambda_2]+q}$, with q finite, the norm $\|A_{ij}|\varphi\rangle\|$ remains finite. Hence we have $\|A_{ij}|\Psi\rangle\| \to \|A_{ij}|\psi\rangle\|$ and the proof is complete.

Theorem 2: In the $\lambda_1^m \to \infty$ limit,

$$W^k_{k+p} \to V^{\lambda_2}_{k-[\lambda_1]} \otimes V^{\lambda_1}_{[\lambda_1]+p} \tag{85}$$

for finite values of p.

Proof: The proof is by induction, starting with Lemma 5. It parallels that for Theorem 1, but uses Lemma 6 rather than Lemma 3 to iterate between the graded subspaces.

Asymptotic coupling coefficients for the large-λ_1^m limit of sp(m,R) can be derived in the manner of Sec. III E, provided that the polynomial $P_n(f)$ appearing in Eqs. (56), (57), and (59) is replaced by the corresponding polynomial $P_n(A)$ in terms of raising operators. Thus in the $\lambda_1^m \to \infty$ limit and for finite values of $[\omega]-[\lambda]$, where ω and λ label unirreps of u(m), the state $|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle$ is given by Eq. (62), where all coupling and recoupling coefficients are u(m) coefficients.

V. sp(m,R) TENSOR OPERATORS

In this section, we extend the results on sp(m,R) to the coupling of a nonunitary finite-dimensional representation of sp(m,R) to a unitary infinite-dimensional representation. Interest in such couplings arises because tensor operators often belong to finite-dimensional irreps.

A. sp(1,R) tensor operators

The components $\{T^{\tilde{j}}_{\tilde{m}}; \tilde{m} = -\tilde{j}, \ldots, +\tilde{j}\}$ of a finite-dimensional (nonunitary) sp(1,R) tensor operator $T^{\tilde{j}}$ are related[10] by the equations

$$\begin{aligned} [J_\pm, T^{\tilde{j}}_{\tilde{m}}] &= \mp\sqrt{(\tilde{j}\mp\tilde{m})(\tilde{j}\pm\tilde{m}+1)}\,T^{\tilde{j}}_{\tilde{m}\pm 1}, \\ [J_0, T^{\tilde{j}}_{\tilde{m}}] &= \tilde{m}T^{\tilde{j}}_{\tilde{m}}. \end{aligned} \tag{86}$$

The decomposition of the coupling $\tilde{j} \otimes j$, where j is the lowest weight for a unirrep, can be inferred by simply counting the number of states with a given weight $M = \tilde{m} + m$. This shows that $\tilde{j} \otimes j$ is the finite sum of unitary irreducible representations

$$\tilde{j}\otimes j=|j-\tilde{j}|\oplus|j-\tilde{j}+1|\oplus\cdots\oplus|j+\tilde{j}| \ . \tag{87}$$

We are interested in the lowest weight states of these couplings, which we express in the form

$$[T^{\tilde{j}}\otimes|j\rangle]^J_J=\sum_{m,\tilde{m}} (j\ m;\tilde{j}\ \tilde{m}|J\ J)T^{\tilde{j}}_{\tilde{m}}\ |jm\rangle \ . \tag{88}$$

Applying the lowering operator J_- to Eq. (88), we get

$$\begin{aligned}0=&\sum_{m,\tilde{m}} (j\ m-1;\tilde{j}\ \tilde{m}+1|J\ J)[J_-,T^{\tilde{j}}_{\tilde{m}+1}]|jm-1\rangle\\ &+\sum_{m,\tilde{m}} (j\ m;\tilde{j}\ \tilde{m}|J\ J)T^{\tilde{j}}_{\tilde{m}}\ [J_-|jm\rangle] \ ,\end{aligned} \tag{89}$$

from which it follows, using Eqs. (64) and (86), that

$$\begin{aligned}0=&\sqrt{(\tilde{j}+\tilde{m}+1)(\tilde{j}-\tilde{m})}\ (j\ m-1;\tilde{j}\ \tilde{m}+1|J\ J)\\ &+\sqrt{(m-j)(m+j-1)}\ (j\ m;\tilde{j}\ \tilde{m}|J\ J) \ .\end{aligned} \tag{90}$$

Combined with the fact that there are finitely many coupling coefficients when $\tilde{j}$ is finite, Eq. (90) can be used to establish that

$$(j\ m;\tilde{j}\ \tilde{m}|J\ J)\to k\left(\frac{1}{\sqrt{2j}}\right)^{m-j}(j\ j;\tilde{j}\ J-j|J\ J) \ , \quad \text{as } j\to\infty \ , \tag{91}$$

where k is some finite constant. It then follows that

$$[T^{\tilde{j}}\otimes|j\rangle]^J_J\to T^{\tilde{j}}_{J-j}|jj\rangle \ , \quad \text{as } j\to\infty \ . \tag{92}$$

Furthermore, since

$$\begin{aligned}J_+[T^{\tilde{j}}\otimes|j\rangle]^J_J=&\sqrt{2J}[T^{\tilde{j}}\otimes|j\rangle]^J_{J+1}\to[J_+,T^{\tilde{j}}_{J-j}]|jj\rangle+T^{\tilde{j}}_{J-j}[J_+|jj\rangle]\\ =&-\sqrt{(\tilde{j}-J+j)(\tilde{j}+J-j+1)}T^{\tilde{j}}_{J-j+1}|jj\rangle+\sqrt{2j}T^{\tilde{j}}_{J-j}|j,j+1\rangle \ ,\end{aligned} \tag{93}$$

we have

$$[T^{\tilde{j}}\otimes|j\rangle]^J_{J+1}\to T^{\tilde{j}}_{J-j}|j,j+1\rangle \ , \quad \text{as} \quad j\to\infty \ . \tag{94}$$

By acting now on $[T^{\tilde{j}}\otimes|j\rangle]^J_{J+1}$ with J_+, we find, in the same manner, that

$$[T^{\tilde{j}}\otimes|j\rangle]^J_{J+2}\to T^{\tilde{j}}_{J-j}|j_1,j_1+2\rangle \ , \quad \text{as } j\to\infty \ . \tag{95}$$

These are the first two steps of an inductive proof of the following lemma.

Lemma 7: For finite values of $\tilde{j}$,

$$[T^{\tilde{j}}\otimes|j\rangle]^J_{J+\delta}\to T^{\tilde{j}}_{J-j}|j,j+\delta\rangle \ , \quad \text{as} \quad j\to\infty \ . \tag{96}$$

B. sp(3,R) tensor operators

The coupling of an sp(3,R) tensor operator $T^{\tilde{\lambda}}$, that transforms according to a nonunitary, finite dimensional representation $\tilde{\lambda}$, to a unitary asymptotic irrep λ of sp(3,R) can be reduced as follows.

Let us denote by $\langle\lambda\rangle$ the character of the Sp(3,R) irrep with lowest weight λ and by $\{\lambda\}=\{\lambda^1,\lambda^2,\lambda^3\}$ the character of the U(3) irrep with highest weight $\lambda=(\lambda^1,\lambda^2,\lambda^3)$. Then, for $\lambda^3\geqslant 3$, the Sp(3,R) $\rightarrow$ U(3) branching rule is given in terms of characters,[10] by

$$\mathrm{Sp}(3,\mathrm{R})\rightarrow \mathrm{U}(3)\ \ ;\ \ \langle\lambda\rangle\mapsto\{\lambda\}\cdot\{D\}\ , \tag{97}$$

where D denotes the sum of U(3) irreps with character

$$\{D\}=\{0\}+\{2\}+\{4\}+\{2,2\}+\{6\}+\{4,2\}+\{2,2,2\}+\cdots\ . \tag{98}$$

Let

$$\langle\lambda\rangle\mapsto\sum_p\ m_p\{\omega_p\}=\sum_p\ m_p\{\omega_p^1,\omega_p^2,\omega_p^3\}\ , \tag{99}$$

denote the decomposition of the character of $\tilde{\lambda}$ into its U(3) characters $\{\omega_p\}=\{\omega_p^1,\omega_p^2,\omega_p^3\}$, where m_p is the multiplicity of ω_p in $\tilde{\lambda}$. Since $\tilde{\lambda}$ is finite dimensional, the sum in Eq. (99) will contain finitely many U(3) unirreps. Then, if $\langle\tilde{\lambda}\rangle\cdot\langle\lambda_1\rangle$ denotes the character for the tensor product $\tilde{\lambda}\otimes\lambda_1$ of a finite Sp(3,R) irrep $(\tilde{\lambda})$ and an infinite unitary irrep (λ_1), we have

$$\mathrm{Sp}(3,\mathrm{R})\rightarrow \mathrm{U}(3)\ \ ;\ \ \langle\tilde{\lambda}\rangle\cdot\langle\lambda_1\rangle\mapsto\sum_p\ m_p\{\omega_p\}\cdot\{\lambda_1\}\cdot\{D\}\ . \tag{100}$$

By comparing this expression with Eq. (97), we deduce that an Sp(3,R) unirrep with lowest weight λ will occur $\Sigma_p m_p\times\rho_\lambda$ times, where ρ_λ is the multiplicity of the U(3) highest weight λ in the reduction of the U(3) product $\omega_p\otimes\lambda_1$; i.e., if

$$\langle\lambda\rangle\mapsto\sum_p\ m_p\{\omega_p\}\quad\text{and}\quad\{\omega_p\}\cdot\{\lambda_1\}=\sum_\lambda\ \rho_\lambda\{\lambda\}\ , \tag{101}$$

then

$$\langle\tilde{\lambda}\rangle\cdot\langle\lambda_1\rangle=\sum_{p\lambda}\ m_p\rho_\lambda\langle\lambda\rangle\ . \tag{102}$$

Lemma 8: Let $T^{\tilde{\lambda}}$ be an sp(3,R) tensor operator, and let V^{λ_1} be defined as usual. Let $W^k\subset W$ be the sp(3,R) invariant subspace of $W=T^{\tilde{\lambda}}\otimes V^{\lambda_1}$ containing the lowest weight of grade k. Then, the lowest grade subspace $W_k^k\subset W^k$ is given, in the limit where $\lambda_1^3\rightarrow\infty$, by

$$W_k^k\rightarrow T^{\tilde{\lambda}}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]}\ , \tag{103}$$

where $T^{\tilde{\lambda}}_{k-[\lambda_1]}$ is the subspace of $T^{\tilde{\lambda}}$ spanned by the components of $T^{\tilde{\lambda}}$ with grade $k-[\lambda_1]$.

Proof: Let $\{A_{ij},B_{ij},[A_{ij},B_{ij}]\}$ denote a basis for $\mathscr{A}_{ij}$, the sp(1,R) subalgebra of sp(3,R) generated by A_{ij} and B_{ij}. Since W_k^k is a lowest grade subspace of the sp(3,R) invariant subspace $W^k\subset W$, it follows that $B_{ij}|\Psi\rangle=0$ for any $|\Psi\rangle\in W_k^k$. Thus W_k^k is spanned by a set of states which are all lowest weight states for unirreps of $\mathscr{A}_{ij}$. If $|\Psi\rangle$ is such a lowest weight state, it is, by Lemma 7, a product $|\Psi\rangle=T^{\tilde{\lambda}}_q\ \ |\psi\rangle$ of a state $|\psi\rangle\in\sigma_{ij}$ and a component of $T^{\tilde{\lambda}}$, where

$$\sigma_{ij}=\{|\psi\rangle\in V^{\lambda_1};\ \ B_{ij}|\psi\rangle=0\}\ . \tag{104}$$

Since this result holds for all $\mathscr{A}_{ij}$, and since

$$V^{\lambda_1}_{[\lambda_1]}=\{|\psi\rangle\ ;B_{ij}|\psi\rangle=0\ ;\forall\ ij\}\ , \tag{105}$$

it follows that, in the asymptotic limit, every state in W_k^k lies in $T^{\tilde{\lambda}}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]}$, and the proof is complete.

Now, if α labels the components $\{T^{\tilde{\lambda}}_{\gamma_p\omega_p\alpha}\}$ of the U(3) tensor $T^{\tilde{\lambda}}_{\gamma_p\omega_p}$ whose components are a subset of components of the sp(3,R) tensor $T^{\tilde{\lambda}}$, we have

$$T^{\tilde{\lambda}}_{\gamma_p\omega_p\alpha}|\psi^{\lambda_1}_{\beta}\rangle \in W^k, \quad k=[\omega_p]+[\lambda_1]\ , \tag{106}$$

where β labels basis states $\{|\psi^{\lambda_1}_{\beta}\rangle\}$ in $V^{\lambda_1}_{[\lambda_1]}$. Furthermore, since

$$[A_{ij},T^{\tilde{\lambda}}_{\gamma_p\omega_p\alpha}]=\sum_{\gamma'\omega'\beta'} c_{\gamma_p\omega_p\alpha,\gamma'\omega'\beta'}T^{\tilde{\lambda}}_{\gamma'\omega'\beta'}\ , \tag{107}$$

and, for finite values of $\tilde{\lambda}$, the coefficients $c_{\gamma_p\omega_p\alpha,\gamma'\omega'\beta'}$ are finite for all values of the indices, we immediately find that

$$A_{ij}T^{\tilde{\lambda}}_{\gamma_p\omega_p\alpha}|\psi^{\lambda_1}_{\beta}\rangle \to T^{\tilde{\lambda}}_{\gamma_p\omega_p\alpha}A_{ij}|\psi^{\lambda_1}_{\beta}\rangle\ , \quad \forall|\psi^{\lambda_1}_{\beta}\rangle\in V^{\lambda_1}\ , \tag{108}$$

since, as seen before, $\|A_{ij}|\psi\rangle\|\to\infty$ in the $\lambda_1^3\to\infty$ limit. From this, we now have

Theorem 3: In the $\lambda_1^3\to\infty$ limit of λ_1,

$$W^k_{k+q}\to T^{\tilde{\lambda}}_{k-[\lambda_1]}\otimes V^{\lambda_1}_{[\lambda_1]+q}\ . \tag{109}$$

Proof: The proof is once again inductive starting this time with Lemma 8 and using Eq. (108) to step between the graded subspaces. The details are omitted.

Using the theorem, we find, in the notation of Eq. (62), that

$$\begin{aligned}|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle \to \phi(\lambda_1,\omega_p;\sigma\lambda)\sum_{\beta\omega_1}\ \phi(\omega_p,\omega_1;\gamma\omega)\ \ U(\omega_2\lambda_1\omega n;\lambda\sigma\alpha,\omega_1\beta\gamma)\\ \times[T^{\tilde{\lambda}}_{\gamma_p\omega_p}\otimes|\psi^{\lambda_1}_{n\beta\omega_1}\rangle]_{\gamma\omega\tau}\ \ \delta_{\rho,\gamma_p\omega_p\sigma}\ \ \delta_{\gamma,n\alpha},\ \text{ as }\ \lambda_1^3\to\infty\ ,\end{aligned} \tag{110}$$

where the coupling is a u(3) coupling, and U is a Racah recoupling coefficient for u(3).

VI. DISCUSSION AND CONCLUSION

In this paper we have investigated the properties of asymptotic representations by looking at matrix elements of some ladder operators of semisimple Lie algebras. Ladder operators are naturally associated with gradings of representations. Thus we have made use of the fact that a contraction, corresponding to an asymptotic limit $\lambda\to\infty$, preserves a suitably defined graded structure of a Lie algebra and its ladder representations.

The formalism singles out a subalgebra $\mathscr{H}\subset\mathscr{G}$ which, by construction, contains all the ladder operators of $\mathscr{G}$ with finite matrix elements. We have shown that, in the asymptotic limit as $\lambda_1\to\infty$, the basis states $\{|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle\}$, for which $[\lambda]-[\omega]$ is finite, for the tensor product space $W=V^{\lambda_2}\otimes V^{\lambda_1}$, depend only on the coupling and recoupling coefficients of $\mathscr{H}$. Thus a major result of this paper is given by Eq. (62)

$$\begin{aligned}|\Psi^{\rho\lambda}_{\gamma\omega\tau}\rangle \to \phi(\lambda_1,\omega_2;\sigma\lambda)\sum_{\beta\omega_1}\ \phi(\omega_2,\omega_1;\kappa\omega)U(\omega_2\lambda_1\omega n;\lambda\sigma\alpha,\omega_1\beta\omega)\\ \times[|\varphi^{\lambda_2}_{\gamma_2\omega_2}\rangle\otimes|\psi^{\lambda_1}_{n\beta\omega_1}\rangle]_{\gamma\omega\tau}\ \ \delta_{\rho,\gamma_2\omega_2\sigma}\ \delta_{\gamma,n\alpha},\ \text{as}\ \lambda_1\to\infty\ ,\end{aligned}$$

where $\phi(\lambda_1,\omega_2;\sigma\lambda)$ and $\phi(\omega_2,\omega_1;\gamma\omega)$ are phase factors and $U(\omega_2\lambda_1\omega n;\lambda\sigma\alpha,\omega_1\beta\gamma)$ is a Racah coefficient for $\mathscr{H}$. In deriving this explicit expression, we have made use of a result, given in Sec. III C, that the arbitrariness of separating multiple copies of equivalent irreducible subrepresentations in the tensor product $\lambda_2\otimes\lambda_1$ has a natural resolution in the $\lambda_1\to\infty$ limit.

In the $\lambda_1 \to \infty$ limit, a unirrep of a semisimple Lie algebra $\mathscr{G}$ approaches a (possibly reducible) representation of a contraction $\mathscr{G}^c$ of $\mathscr{G}$. Consider, for example, a representation of the su(2) Lie algebra

$$[J_+, J_-] = 2J_0, \quad [J_0, J_\pm] = \pm J_\pm \tag{111}$$

with angular momentum j and basis states $\{|jm\rangle\}$. Let $\mathscr{J}_\pm$ and $\mathscr{I}$ denote the renormalized operators

$$\mathscr{J}_\pm = \frac{J_\pm}{\sqrt{2j}}, \quad \mathscr{I} = \frac{J_0}{j}. \tag{112}$$

In terms of these operators, the su(2) commutation relations become

$$[\mathscr{J}_+, \mathscr{J}_-] = \mathscr{I}, \quad [\mathscr{I}, \mathscr{J}_\pm] = \pm \frac{\mathscr{J}_\pm}{j}. \tag{113}$$

Thus, in the asymptotic limit as $j \to \infty$, we obtain the İnönü–Wigner[12] contraction of su(2) to a Heisenberg–Weyl algebra with

$$[\mathscr{J}_+, \mathscr{J}_-] = \mathscr{I}, \quad [\mathscr{I}, \mathscr{J}_\pm] = 0. \tag{114}$$

The latter algebra is more usually expressed in terms of harmonic oscillator raising and lowering operators

$$[c, c^\dagger] = I, \quad [\mathscr{I}, c] = [I, c^\dagger] = 0. \tag{115}$$

Note also that we still have the commutation relation

$$[J_0, \mathscr{J}_\pm] = \pm \mathscr{J}_\pm \tag{116}$$

which can be compared with the harmonic oscillator equations

$$[H, c^\dagger] = c^\dagger, \quad [H, c] = -c, \tag{117}$$

with $H = c^\dagger c$.

The way in which the states $\{|jm\rangle\}$ of the su(2) representation approach those of a harmonic oscillator is given by the identification

$$|jm\rangle \equiv |n\rangle, \quad \text{with} \quad n = j - m. \tag{118}$$

As $j \to \infty$, we have, for small values of n,

$$\mathscr{J}_+ |n\rangle = \sqrt{\frac{1}{2j} n(2j+1)}\, |n-1\rangle \to \sqrt{n}\, |n-1\rangle,$$

$$\mathscr{J}_- |n\rangle = \sqrt{\frac{1}{2j}(2j-n)(n+1)}\, |n+1\rangle \to \sqrt{n+1}\, |n+1\rangle, \qquad \mathscr{J}_0 |n\rangle = n|n\rangle. \tag{119}$$

Thus we obtain the correspondence

$$\mathscr{J}_+ \to c, \quad \mathscr{J}_- \to c^\dagger, \quad \mathscr{I} \to I, \quad \mathscr{J}_0 \to c^\dagger c \tag{120}$$

valid whenever $n \ll j$. In terms of special functions, the contraction of spherical harmonics to Hermite polynomials can be found in Ref. 13.

Note that such a contraction of a semisimple Lie algebra is not semisimple. Since representations of the non-semisimple algebra can be obtained as asymptotic limits of those of the semisimple algebra $\mathscr{G}$, we can *define* the coupling of an irrep of $\mathscr{G}^c$ to an irrep of $\mathscr{G}$ as the asymptotic limit of the coupling of irreps of $\mathscr{G}$. Asymptotic coupling coefficients of $\mathscr{G}$ are precisely the

coupling coefficients for the coupling $\mathscr{G}^c \times \mathscr{G} \to \mathscr{G}^c$. Furthermore, the renormalization of the Lie algebra implies a renormalization of the representation labels of $\mathscr{G}$, so that the labels specifying representations of $\mathscr{G}^c$ can take finite values. Thus we have the remarkable fact that the asymptotic coupling coefficients become exact for $\mathscr{G}^c \times \mathscr{G} \to \mathscr{G}^c$, even for finite values of the representation labels of $\mathscr{G}^c$.

The Clebsch–Gordan coefficients for the coupling of two irreps, $\lambda_1 \otimes \lambda_2$, of $\mathscr{G}^c$ are given precisely by the double limit $\lambda_1, \lambda_2 \to \infty$ of Clebsch–Gordan coefficients for $\mathscr{G}$. Unfortunately, it appears difficult to compute such coefficients within the current formalism.

In a previous work,[13] we obtained asymptotic Clebsch–Gordan coefficients for the su(3) $\supset$ so(3) subalgebra chain by a projection method. This projection method is different from the technique presented in Sec. III E. Nevertheless, it shares a ''factorization'' property with the current work. To be precise, it is implicit in Ref. 14 that

$$\frac{\|Q_0|\varphi\rangle\|}{\|Q_0|\psi\rangle\|} \to 0 \quad \text{as } \lambda_1^1 \to \infty \ , \tag{121}$$

where $|\psi\rangle \in V^{\lambda_1}$ and $|\varphi\rangle \in V^{\lambda_2}$ as usual, and where Q_0 is the $\nu = 0$ component of an so(3) $\subset$ su(3) tensor operator. This leads immediately to the factorization

$$Q_0|\Psi\rangle \to |\varphi\rangle[Q_0|\psi\rangle] \quad \text{as } \lambda_1^1 \to \infty \ , \tag{122}$$

which can be compared with Eq. (42) for components of $\mathscr{H} \subset \mathscr{G}$ tensors. However, unlike the $\mathscr{H} \subset \mathscr{G}$ tensors, not all components of the su(3) quadrupole moment Q_ν satisfy Eq. (121) in general. Thus it would be interesting to know if asymptotic coupling always implies such a factorization of some matrix elements and if, conversely, such a factorization implies asymptotic coupling.

Asymptotic Clebsch–Gordan coefficients have so far been used to analyze coupled systems having two different scales. For instance, in some core-plus-particle models of the nucleus, the so-called ''collective'' part of multipole operators can dwarf the single particle contribution, so that Eqs. (42) or (122) are true to a first approximation. The eigenfunctions of such two-scale systems are often found, to leading order, by using the Born–Oppenheimer (BO) approximation.[15] This suggests a useful parallel between the physically insightful BO approach and the mathematical technique of asymptotic coupling.[16] It remains to see if this parallel extends to higher orders, i.e., if corrections to the BO wave functions have a corresponding group-theoretical interpretation in terms of corrections to the asymptotic limit.

Finally, an obvious question which remains unanswered in the present work is the evaluation and properties of asymptotic coupling coefficients for finite values of the grade $[\omega]$. In Ref. 14, we computed su(2) $\supset$ u(1) Clebsch–Gordan coefficients for finite value of the projection m in the limit where $j \to \infty$. This was done by embedding the u(1) subalgebra so that the set of states obtained from the highest weight state by the action of all elements of the corresponding U(1) subgroup spanned the whole su(2) representation. In contrast, the theory and examples presented in the present paper are closely related to the work found in Ref. 8, where the subalgebra $\mathscr{H}$ does not connect states in a unitary representation of $\mathscr{G}$ having different grades. This suggests that different embeddings of $\mathscr{H}$, when equivalent, will yield results applicable to different ranges of values of $[\omega]$.

ACKNOWLEDGMENTS

This paper was supported in part by N.S.E.R.C. of Canada.

[1] W. T. Sharp, *Racah Algebra and the Contraction of Groups* (University of Toronto Press, Toronto, 1984).
[2] G. N. Watson, *Theory of Bessel Functions* (Cambridge University Press, Cambridge, England, 1966), 2nd ed.
[3] D. J. Rowe, R. LeBlanc, and J. Repka, J. Phys. A **22**, L209 (1989); D. J. Rowe, M. G. Vassanji, and J. Carvalho, Nucl. Phys. A **504**, 76 (1989).
[4] D. J. Rowe and J. Repka, J. Math. Phys. **37**, 1 (1996).
[5] D. E. Flath and J. Towber, Commun. Algebra **18**, 4047 (1990); **20**, 2903 (1992); D. Flath and J. Towber, J. Math. Phys. **34**, 1523 (1993); Contemp. Math. **160**, 75 (1994).
[6] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).

[7] K. T. Hecht, Nucl. Phys. **62**, 1 (1965).
[8] D. J. Rowe and J. Repka, J. Math. Phys. **38**, 4363 (1997).
[9] M. Kashiwara and M. Vergne, Invent. Math. **44**, 1 (1978).
[10] D. J. Rowe, B.G. Wybourne, and P. H. Butler, J. Phys. A **18**, 939 (1985).
[11] H. Ui, Prog. Theor. Phys. **44**, 689 (1970).
[12] E. Inönü and E. P. Wigner, Proc. Natl. Acad. Sci. (N.Y.) **39**, 510 (1953).
[13] F. T. Arecchi *et al.*, Phys. Rev. A **6**, 2211 (1972).
[14] H. de Guise and D. J. Rowe, J. Math. Phys. **36**, 6991 (1995).
[15] M. Born and J. R. Oppenheimer, Ann. der Physik **84**, 457 (1927); F. M. H. Villars, Nucl. Phys. A **420**, 61 (1984); **473**, 539 (1987).
[16] H. de Guise and D. J. Rowe, Nucl. Phys. A **589**, 58 (1995).

Invariant polynomials and Molien functions

Michael Forger
Departamento de Matemática Aplicada, Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281, BR-05315-970 São Paulo, Brazil

(Received 12 February 1997; accepted for publication 9 June 1997)

The Molien function associated with a finite-dimensional representation of a compact Lie group is a useful tool in representation theory because it acts as a generating function for counting the number of invariant polynomials on the representation space. The main purpose of this paper is to introduce a more general (and apparently new) generating function which, in the same sense, counts not only the number of invariant real polynomials in a real representation or the number of invariant complex polynomials in a complex representation, as a function of their degree, but encodes the number of invariant real polynomials in a complex representation, as a function of their bidegree (the first and second component of this bidegree being the number of variables in which the polynomial is holomorphic and antiholomorphic, respectively). This is obviously an additional and non-trivial piece of information for representations which are truly complex (i.e., not self-conjugate) or are pseudo-real, but it provides additional insight even for real representations. In addition, we collect a number of general formulas for these functions and for their coefficients and calculate them in various irreducible representations of various classical groups, using the software package MAPLE.
© *1998 American Institute of Physics.* [S0022-2488(98)02901-6]

I. INTRODUCTION

Determining the ring of invariant polynomials in arbitrary representations is perhaps one of the most important open problems in group theory. Trying to solve this problem in full generality—for example, in the framework of arbitrary finite-dimensional representations of compact Lie groups (including finite groups)—is presently considered to be a hopeless enterprise. What is, however, a tractable problem is to determine at least the number of (linearly independent) invariant polynomials, or more precisely, the dimension of the space of invariant homogeneous polynomials of any given degree, and this is often an extremely useful piece of information when it comes to calculating invariant polynomials for a concrete representation of a concrete group.

The present investigation originated from the recent work of J. E. M. Hornos and Y. M. M. Hornos[1] on the origin of the genetic code that has found great repercussion in the international scientific literature.[2,3] According to their proposal, the degeneracy of the universal genetic code for protein synthesis is not (as many molecular biologists used to and some continue to believe) purely accidental, but can be understood as resulting from an evolutionary process which involves symmetry breaking: evolution from a highly symmetric initial state to a final state in which this symmetry is strongly broken. This evolution must have occurred, in several consecutive steps, far back in earth's early history, and so is not accessible to direct observation. For the time being, the scheme proposed by Hornos and Hornos is purely group theoretical, its main virtue being that—within the limits of the originally proposed scheme—the initial symmetry and all intermediate steps in the sequence can be uniquely reconstructed from presently available data. (More recently, the scope of the scheme has been extended and, as a result, a second possibility has emerged.)[4]

The great challenge for the future is to identify a dynamical system modelling the underlying evolutionary process, so that the sequence of symmetry breakings found can be associated with a sequence of (generic) bifurcations. In fact, it is well known that in dynamical systems with symmetry and with external parameters, bifurcations that occur under appropriate variations of the parameters almost unavoidably lead to symmetry breaking.[5] In this general framework, of course, the variety of possibilities is enormous, so that for the time being, we have decided to perform our search in the more restricted class of Hamiltonian dynamical systems. But the most natural can-

 © 1998 American Institute of Physics

didates for a Hamiltonian function capable of reproducing the desired symmetry breaking pattern are just the polynomial functions on the representation space which are invariant under the full symmetry group and which, if possible, should be of degree $\leqslant 4$. (This would correspond to some kind of anharmonic oscillator type model.) In the case at hand, the full symmetry group is the rank 3 symplectic group Sp(6) and the representation is the 64-dimensional irreducible representation of highest weight (1,1,0)—a complex representation which is self-conjugate but is pseudo-real, rather than real, and which has come to be called the codon representation of Sp(6).

Before attempting to explicitly construct all invariant polynomials up to a given degree, in any given representation of any group, it is obviously of great help to know precisely how many such polynomials there are. To see how useful this information can be and how it may help to avoid unpleasant surprises, consider as an example the problem of finding the invariant polynomials of degree 4 in, say, the spin 3 representation of the ordinary rotation group. (This example has been chosen because of certain similarities with the codon representation.) Given the fact that vectors in this representation space can be realized as totally symmetric tensors t of rank 3 over three-dimensional Euclidean space which are traceless in any pair of indices, it is easy to construct invariant polynomials of degree 4 by considering all possibilities of contracting indices in a product of four such tensors, using the invariant scalar product. Symmetry implies that the only relevant information is how many indices of any given tensor are contracted with how many indices of any other given tensor (not which with which), while tracelessness forbids contraction of two indices that appear within the same tensor. Therefore, there are three different possibilities to contract the 12 indices (each one ranging from 1 to 3) in the product

$$t_{i_1 j_1 k_1} t_{i_2 j_2 k_2} t_{i_3 j_3 k_3} t_{i_4 j_4 k_4}$$

which come to mind.

(1) Every tensor has only one partner for the contraction. Contract, for example, i_1 with i_2, j_1 with j_2, k_1 with k_2 and i_3 with i_4, j_3 with j_4, k_3 with k_4:

$$P_{4,1}(t) = P_2(t)^2, \quad P_2(t) = \sum_{i,j,k=1}^{3} t_{ijk}^2.$$

(2) Every tensor has two partners for the contraction. Contract, for example, i_1 with i_2, j_1 with j_2, k_1 with i_3, k_2 with i_4, j_3 with j_4, k_3 with k_4:

$$P_{4,2}(t) = \sum_{i,j,k,l,m,n=1}^{3} t_{ijk} t_{ijl} t_{kmn} t_{lmn}\,.$$

(3) Every tensor has three partners for the contraction. Contract, for example, i_1 with i_2, j_1 with i_3, k_1 with i_4, j_2 with j_3, k_2 with j_4, k_3 with k_4:

$$P_{4,3}(t) = \sum_{i,j,k,l,m,n=1}^{3} t_{ijk} t_{ilm} t_{jln} t_{kmn}\,.$$

The first possibility corresponds to the square of the quadratic polynomial stemming from the invariant scalar product in this representation, while the other two are genuinely quartic and apparently independent. Therefore, it comes as a surprise that the number of invariant polynomials of degree 4 in this representation, as computed by the techniques to be discussed (and further developed) in the present paper, turns out to be 2, and not 3. This means that the three polynomials obtained above must be linearly dependent! And indeed, writing out these polynomials as explicit functions of the seven variables $t_{112}, t_{122}, t_{113}, t_{133}, t_{223}, t_{233}, t_{123}$, we find, using MAPLE, the following linear relation:

$$P_{4,1} = 2(P_{4,2} + P_{4,3}).$$

This simple example shows that independent information on the correct number of invariant polynomials is crucial if one wants to avoid naive overcounting. It is equally crucial if one wants to avoid undercounting, which may occur as a result of overlooking non-obvious, ''hidden''

invariants. (One fairly well-known example that comes to mind is the Pfaffian—an invariant in the adjoint representation of the special orthogonal groups in even dimensions which is hard to detect by tensorial methods.)

The basic strategy for determining the number of invariant polynomials of any given degree is to encode all of them into a generating function, which is commonly called the Molien function and which can often be calculated in closed form, at least in sufficiently simple situations. But even when this is not possible in practice, the existing formulas for the generating function can be exploited to compute at least the first few coefficients.

Unfortunately, the standard Molien function M_π associated with a given finite-dimensional representation π of a given compact Lie group G is inadequate for handling the problem at hand in its full generality, due to a discrepancy between ground fields. On the one hand, the representation spaces encountered in group theory are always assumed to be complex (this guarantees that one can simultaneously diagonalize maximal commuting sets of linear transformations); the usual convention for handling real representations is then to view them as complex representations possessing an invariant antilinear involution. On the other hand, we are typically interested in finding all invariant real polynomials and not just the complex ones. The standard Molien function, however, does not allow one to identify the extent to which a real polynomial on the space of a complex representation is holomorphic or antiholomorphic in its variables. What is worse, it does not detect invariant polynomials of mixed type. The obvious prototype of such a polynomial is the invariant scalar product—a quadratic polynomial on the representation space, holomorphic in one variable and antiholomorphic in the other—which exists in any finite-dimensional representation of any compact Lie group and, in addition, is the only polynomial of its kind (up to a constant multiple) in case the representation is irreducible. The fact that the standard Molien function captures only purely holomorphic (or purely antiholomorphic) invariants, but fails to detect mixed invariants, including the invariant scalar product, can already be illustrated by looking at the simplest of all representations: the fundamental spin 1/2 representation of the ordinary rotation group (or rather its universal covering group SU(2)).

The natural way out of this dilemma, proposed and elaborated in the present paper, is to invent a new generating function F_π which generalizes the usual Molien function and is specifically designed to capture all real polynomials in complex representations, discriminating between holomorphic ones, purely antiholomorphic ones and mixed ones, according to their bidegree.

As far as the specific case of the codon representation of Sp(6) is concerned, the techniques developed in the present work allow one to conclude that: (a) there are no invariant quadratic polynomials of bidegree (2,0) (purely holomorphic) or of bidegree (0,2) (purely antiholomorphic), while there is one invariant quadratic polynomial of bidegree (1,1) (the scalar product), (b) there are no invariant cubic polynomials of any kind and (c) the numbers of invariant quartic polynomials are as follows: 3 of bidegree (4,0) (purely holomorphic), 3 of bidegree (0,3) (purely antiholomorphic), 6 of bidegree (3,1), 6 of bidegree (1,3) and finally 15 of bidegree (2,2). Since a Hamiltonian function must be real, we may therefore conclude that the general candidate for a Hamiltonian capable of describing the evolution of the genetic code through an anharmonic oscillator type model must be a linear combination of the invariant scalar product, its square and another 14 genuinely quartic invariant polynomials of bidegree (2,2). What remains to be determined are the explicit form of these polynomials and the conditions to be imposed on their coefficients in order to guarantee positivity of the energy. (The remaining final freedom of modifying the Hamiltonian by an additive constant may then be used to normalize its minimum value to 0.)

The paper is organized as follows. In Sec. II, we briefly review the definition of the Molien function, whereas in Sec. III we define our new generating function for counting invariant real polynomials in complex representations. Both sections contain comments on the relations between the analytic form of the generating functions and the structure of the (graded or bigraded) algebra of invariant polynomials, in terms of generators and relations. In Sec. IV, we derive integral formulas for both generating functions, with emphasis on their explicit form for unitary representations of compact connected Lie groups, in terms of roots and weights. In Sec. V, we present purely combinatorial formulas for the coefficients, involving the multiplicities of the weights and a set of integer coefficients called ''decomposition indices'' associated with the vectors in the root lattice. In Sec. VI, we discuss as an example the results we have obtained for the simplest among

all compact connected simple Lie groups: SU(2). Finally, in Sec. VII, we present calculations for various irreducible representations of the rank 2 symplectic group Sp(4) and the rank 3 symplectic group Sp(6), including the fundamental representations and the other irreducible representations which appear in the symmetry breaking scheme of Hornos and Hornos.[1]

II. THE MOLIEN FUNCTION: DEFINITION AND ELEMENTARY PROPERTIES

Given an arbitrary finite-dimensional representation π of a compact Lie group G on some n-dimensional vector space V, and denoting by $c_k(\pi)$ the number of (linearly independent) G-invariant polynomials of degree k on V, one defines the corresponding Molien function M_π by the power series

$$M_\pi(z)=\sum_{k=0}^{\infty} c_k(\pi)z^k. \tag{1}$$

Note that identifying homogeneous polynomials of degree k on V with totally symmetric tensors of degree k over V^*, we easily obtain the estimate

$$0\leqslant c_k(\pi)\leqslant\binom{n+k-1}{k}=\binom{n+k-1}{n-1},$$

so $c_k(\pi)$ grows at most polynomially as $k\to\infty$ (the highest power being k^{n-1}); therefore, the above power series is absolutely convergent on the open unit disk in the complex z plane and hence M_π is a complex analytic function there—a function from which we may obviously recover all the numbers $c_k(\pi)$ as Taylor coefficients:

$$c_k(\pi)=\frac{1}{k!}\frac{d^k}{dz^k}M_\pi(z)\bigg|_{z=0}. \tag{2}$$

Note also that this definition can be used both in the real and in the complex setting, and more generally, for representations π of G by $\mathbb{F}$-linear transformations on finite-dimensional vector spaces V over $\mathbb{F}$, where $\mathbb{F}$ is an arbitrary field of characteristic 0.

Some initial information on the structure of the Molien function can be gained by describing the graded ring of G-invariant polynomials on V in terms of generators and relations. In fact, the Hilbert–Weyl theorem guarantees that this graded ring is finitely generated, i.e., that there exists a finite set $\{P_1,\ldots,P_N\}$ of homogeneous G-invariant polynomials on V such that every (homogeneous) G-invariant polynomial P on V can be written in the form

$$P(v)=p(P_1(v),\ldots,P_N(v)). \tag{3}$$

where p is some (homogeneous) polynomial on $\mathbb{F}^N$, provided we define homogeneity of polynomials on $\mathbb{F}^N$ as referring to a modified notion of degree, namely,

$$\deg\ p_{m_1,\ldots,m_N}=m_1\ \deg\ P_1+\ldots+m_N\ \deg\ P_N, \tag{4}$$

for the monomial $p_{m_1},\ldots,m_N$ given by

$$p_{m_1,\ldots,m_N}(u_1,\ldots,u_N)=u_1^{m_1}\ldots u_N^{m_N}. \tag{5}$$

(See, for example, Ref. 5, p. 46 for a statement and pp. 54–58 for a proof. However, I prefer to avoid the term ''Hilbert basis'' used in Ref. 5, which I consider to be potentially misleading and therefore unfortunate.) Note that the polynomial p is in general not uniquely fixed by the polynomial P because there may be relations, i.e., polynomials R on $\mathbb{F}^N$ satisfying

$$R(P_1(v),\ldots,P_N(v))\equiv 0. \tag{6}$$

Note also that the relations R form a graded ring which is nothing but the kernel of the degree-preserving homomorphism, defined by (3), from the graded ring of polynomials p on $\mathbb{F}^N$ onto the graded ring of G-invariant polynomials P on V. Being a graded ideal in a polynomial ring, this kernel is finitely generated, i.e., there exists a finite set $\{R_1,\ldots,R_M\}$ of homogeneous polynomials on $\mathbb{F}^N$ such that every (homogeneous) polynomial R on $\mathbb{F}^N$ satisfying (6) can be written in the form

$$R(u_1,\ldots,u_N)=r(R_1(u_1,\ldots,u_N),\ldots,R_M(u_1,\ldots,u_n)), \tag{7}$$

where r is some (homogeneous) polynomial on $\mathbb{F}^M$, provided we define homogeneity of polynomials on $\mathbb{F}^M$ as referring to a modified notion of degree, namely,

$$\deg\ r_{n_1,\ldots,n_M}=n_1\ \deg\ R_1+\ldots+n_M\ \deg\ R_M, \tag{8}$$

for the monomial $r_{n_1,\ldots,n_M}$ given by

$$r_{n_1,\ldots,n_M}(v_1,\ldots,v_M)=v_1^{n_1}\ldots v_M^{n_M}. \tag{9}$$

(See, for example, Ref. 5, Corollary 6.2, p. 54.) In general, neither the set $\{P_1,\ldots,P_N\}$ of generators nor the set $\{R_1,\ldots,R_M\}$ of relations is unique, but we can at least fix the number N of generators and the number M of relations by requiring both N and M to be minimal.

The simplest case is, of course, when $M=0$, i.e., there are no relations. Counting the number $c_k(\pi)$ of G-invariant polynomials of degree k on V gives

$$c_k(\pi)=\mathrm{card}\left\{(m_1,\ldots,m_N)\in\mathbb{N}_0^N \Big/ \sum_{j=1}^N m_j\ \deg\ P_j=k\right\}.$$

But this is just the coefficient of z^k in the power series expansion of the function

$$M_\pi(z)=\frac{1}{\prod_{j=1}^N(1-z^{\deg\ P_j})}. \tag{10}$$

Conversely, it is clear that if the Molien function for the representation π of G on V has this form, then there can be no relations, because otherwise the number of G-invariant polynomial functions on V of degree k would have to be strictly less than the coefficient of z^k in the power series expansion of Eq. (10), at least for some k. Thus the Molien function detects the presence or absence of relations among the generators of the ring of invariant polynomials.

For later use, it is also of some interest to write down the corresponding result for the next-simplest case $M=1$, i.e., when there is a single relation $R=r(P_1,\ldots,P_N)$. Due to the fact that polynomials of the form $p(P_1,\ldots,P_N)\ r(P_1,\ldots,P_N)$ will vanish identically, counting the number $c_k(\pi)$ of G-invariant polynomials of degree k on V now gives

$$\begin{aligned}c_k(\pi)=&\,\mathrm{card}\left\{(m_1,\ldots,m_N)\in\mathbb{N}_0^N \Big/ \sum_{j=1}^N m_j\ \deg\ P_j=k\right\}\\ &-\mathrm{card}\left\{(m_0;m_1,\ldots,m_N)\in\mathbb{N}_0^{N+1} \Big/ \sum_{j=1}^N m_0 m_j\ \deg\ R\ \deg\ P_j=k\right\}.\end{aligned}$$

But this is just the coefficient of z^k in the power series expansion of the function

$$M_\pi(z)=\frac{1-z^{\deg\ R}}{\prod_{j=1}^{N}(1-z^{\deg\ P_j})}. \tag{11}$$

Again, it is clear that if the Molien function for the representation π of G on V has this form, then there can be no other relations, because otherwise the number of G-invariant polynomial functions on V of degree k would have to be strictly less than the coefficient of z^k in the power series expansion of Eq. (11), at least for some k.

III. A NEW GENERATING FUNCTION

As remarked previously, the above definition of the Molien function applies equally well in the real and in the complex setting. Often, however, it is of interest to also determine the number of (linearly independent) G-invariant real polynomials in a complex representation, the typical example for a quadratic polynomial of this kind being the square of the norm in a unitary representation. Therefore, it is useful to introduce a generating function for counting the number of such invariants as well. The main new feature that must be taken into account is the fact that real polynomials over a complex vector space carry, over and above their usual degree, a bidegree that counts the number of variables in which they are holomorphic and antiholomorphic, respectively. This will lead to a generating function which depends on two variables, rather than one.

Indeed, given an arbitrary representation π of a compact Lie group G on some n-dimensional complex vector space V, let us first of all define $\bar{V}$ to be the n-dimensional complex vector space which is ''V with the opposite complex structure'' and $V_r=\bar{V}_r$ to be the $2n$-dimensional real vector space obtained from V or $\bar{V}$ by ''forgetting the complex structure''. In other words, V, $\bar{V}$ and V_r are identical as sets and as real vector spaces, while the complex structures on V and on $\bar{V}$ are in this picture encoded into real linear transformations $J:V_r\to V_r$ and $J:V_r\to\bar{V}_r$, which are nothing but multiplication by i in V and in $\bar{V}$, respectively, so $J^2=-1$ and $J^2=-1$; then ''opposite'' means that $J=-J$. (The idea behind this construction is that it enables us to identify, for any complex vector space W, complex antilinear maps from V to W with complex linear maps from $\bar{V}$ to W.) Next, recall that homogeneous real polynomials of degree k on V can be identified with totally symmetric $\mathbb{R}$-multilinear mappings from $V_r\times\ldots\times V_r$ (k copies) to $\mathbb{R}$—or to $\mathbb{C}$ if we allow such polynomials to be complex-valued, as will be assumed throughout the following. We shall say that such a polynomial is homogeneous of bidegree (p,q), with $p+q=k$, if under this identification it corresponds to a totally symmetric $\mathbb{C}$-multilinear mapping from $V\times\ldots\times V\times\bar{V}\times\ldots\times\bar{V}$ to $\mathbb{C}$, with p copies of V and q copies of $\bar{V}$. For such a polynomial P, we call the number p its holomorphic degree, denoted by $\deg_h P$, and the number q its antiholomorphic degree, denoted by $\deg_a P$. It is easy to show that any homogeneous real polynomial of degree k may be uniquely decomposed into a sum of homogeneous polynomials of bidegree (p,q), as follows:

$$P_k=\sum_{\substack{p,q=0\\ p+q=k}}^{k} P_{p,q}.$$

Namely, given P_k, we may set

$$\widetilde{P}_{p,q}(v_1,\ldots,v_k)=\frac{1}{2^k}\sum_{l_1,\ldots l_k=0}^{1}(-i)^{l_1+\ldots+l_p}\,i^{l_{p+1}+\ldots+l_{p+q}}\,P_k(i^{l_1}v_1,\ldots,i^{l_k}v_k),$$

which defines a real multilinear function, complex linear and symmetric in the first p variables and complex antilinear and symmetric in the last q variables; then the above decomposition holds with $P_{p,q}$ obtained from $\widetilde{P}_{p,q}$ by symmetrization in all k arguments,

$$P_{p,q}(v_1,\ldots,v_k)=\frac{1}{k!}\sum_{\sigma\in S_k}\widetilde{P}_{p,q}(v_{\sigma(1)},\ldots,v_{\sigma(k)}).$$

This decomposition can be stated in more concrete terms by introducing an arbitrary basis $\{v_1,\ldots,v_n\}$ of V (over $\mathbb{C}$), together with the induced basis $\{v_1,iv_1,\ldots,v_n,iv_n\}$ of V_r (over $\mathbb{R}$), and expanding vectors in the representation space into components:

$$v=\sum_{j=1}^{n}\zeta_j v_j=\sum_{j=1}^{n}(\xi_j+i\eta_j)v_j.$$

Then any polynomial P on V_r can be written either as a linear combination of monomials which are products of powers of the real coordinates ξ_j and η_j or as a linear combination of monomials which are products of powers of the complex coordinates ζ_j and their complex conjugates $\overline{\zeta_j}$. Using the latter representation and employing multi-index notation, we have

$$P(v)=\sum_{\alpha,\beta}a_{\alpha,\beta}\zeta^{\alpha}\overline{\zeta^{\beta}},$$

so

$$P(v)=\sum_{k}P_k(v)=\sum_{p,q}P_{p,q}(v),$$

with

$$P_k(v)=\sum_{|\alpha|+|\beta|=k}a_{\alpha,\beta}\zeta^{\alpha}\overline{\zeta^{\beta}},\quad P_{p,q}(v)=\sum_{\substack{|\alpha|=p\\|\beta|=q}}a_{\alpha,\beta}\zeta^{\alpha}\overline{\zeta^{\beta}}.$$

(All sums are supposed to be finite.)

The crucial point is now that since G acts on V by complex linear transformations $\pi(g)$, these decompositions preserve G-invariance, that is, if P is G-invariant, so are not only the P_k but also the $P_{p,q}$. Therefore, denoting by $c_{p,q}(\pi)$ the number of (linearly independent) G-invariant polynomials of bidegree (p,q), we define the following generating function F_π of two variables, which for later convenience we shall assume to be mutually complex conjugate:

$$F_\pi(z,\overline{z})=\sum_{p,q=0}^{\infty}c_{p,q}(\pi)z^p\overline{z^q}. \tag{12}$$

As before, this power series is absolutely convergent on the open unit disk in the complex z plane and hence F_π is a real analytic function there—a function from which we may obviously recover all the numbers $c_{p,q}(\pi)$ as Taylor coefficients:

$$c_{p,q}(\pi)=\frac{1}{p!}\frac{1}{q!}\frac{\partial^p}{\partial z^p}\frac{\partial^q}{\partial\overline{z^q}}F_\pi(z,\overline{z})\big|_{z=0,\overline{z}=0}. \tag{13}$$

As a first elementary property of this new generating function, note that it behaves naturally under complex conjugation. Namely, introducing an arbitrary conjugation σ on V, that is, an involutive antilinear transformation $\sigma:V\to V$ to define the complex conjugate representation $\overline{\pi}$ of π according to

$$\overline{\pi}(g)=\sigma\pi(g)\sigma\quad\text{for } g\in G, \tag{14}$$

we note that $c_{p,q}(\overline{\pi})=c_{q,p}(\pi)$ and hence

$$F_{\overline{\pi}}(z,\overline{z})=F_\pi(\overline{z},z). \tag{15}$$

Obviously, this relation does not depend on the choice of the conjugation σ because the representations $\overline{\pi_1}$ and $\overline{\pi_2}$ defined by means of two different conjugations σ_1 and σ_2 are equivalent (with $\sigma_2\sigma_1^{-1}$ as the intertwining operator). Note also that F_π contains the Molien functions M_π and $M_{\overline{\pi}}$ as special cases:

$$M_\pi(z)=F_\pi(z,0), \quad M_{\overline{\pi}}(z)=F_\pi(0,z). \tag{16}$$

As in the case of the ordinary Molien function M_π, the generating function F_π allows one to read off important information about the generators of the bigraded ring of G-invariant real polynomials on V and about the relations that exist between them. Indeed, observe first of all that the set $\{P_1,\ldots,P_N\}$ of generators and the set $\{R_1,\ldots,R_M\}$ of relations may without loss of generality (and at most at the expense of increasing the ''minimum'' number N of generators and the '' minimum'' number M of relations required) be assumed to consist of polynomials which are homogeneous in bidegree. Then it is not difficult to see that in the simplest case $M=0$ (no relations),

$$c_{p,q}(\pi)=\mathrm{card}\left\{(m_1,\ldots,m_N)\in\mathbb{N}_0^N \;\middle/\; \sum_{j=1}^N m_j\ \deg_h P_j=p,\ \sum_{j=1}^N m_j\ \deg_a P_j=q\right\}.$$

and hence

$$F_\pi(z,\overline{z})=\frac{1}{\prod_{j=1}^N(1-z^{\deg_h P_j}\overline{z}^{\deg_a P_j})}, \tag{17}$$

while in the next-simplest case $M=1$ (a single relation),

$$\begin{aligned} c_{p,q}(\pi)=&\mathrm{card}\left\{(m_1,\ldots,m_N)\in\mathbb{N}_0^N \;\middle/\; \sum_{j=1}^N m_j\ \deg_h P_j=p,\ \sum_{j=1}^N m_j\ \deg_a P_j=q\right\}\\ &-\mathrm{card}\left\{(m_0;m_1,\ldots,m_N)\in\mathbb{N}_0^{N+1} \;\middle/\; \begin{array}{l}\displaystyle\sum_{j=1}^N m_0 m_j\ \deg_h R\ \deg_h P_j=p\\ \displaystyle\sum_{j=1}^N m_0 m_j\ \deg_a R\ \deg_a P_j=q\end{array}\right\}. \end{aligned}$$

and hence

$$F_\pi(z,\overline{z})=\frac{1-z^{\deg_h R}\,\overline{z}^{\deg_a R}}{\prod_{j=1}^N(1-z^{\deg_h P_j}\,\overline{z}^{\deg_a P_j})}. \tag{18}$$

IV. INTEGRAL FORMULAS

To begin with, we quote a well-known integral formula which allows one to compute the Molien function M_π in terms of an integral over the group. Namely, let μ_G be the biinvariant Haar measure on G, normalized so that the total volume of G with respect to μ_G is 1. Then

$$M_\pi(z)=\int_G d\mu_G(g)\ \frac{1}{\det(1-z\,\pi(g))}. \tag{19}$$

This formula is easily generalized to an integral formula for the generating function F_π; it reads

$$F_\pi(z,\overline{z})=\int_G d\mu_G(g)\ \frac{1}{\det(1-z\,\pi(g))}\ \frac{1}{\det(1-\overline{z}\,\overline{\pi(g)})}. \tag{20}$$

The proof is similar to that for the usual Molien function (see, e.g., Ref. 6 p. 204) and is based on calculating the character χ_P of the representation π_P of G on the algebra of polynomial functions on V induced by the given representation π of G on V according to

$$(\pi_P(g)P)(v)=P(\pi(g)^{-1}v),$$

as follows: Since G is a compact Lie group, the given representation π of G on V may without loss of generality be assumed to be unitary (starting from some arbitrary scalar product on V, a G-invariant scalar product on V is obtained by integration over the group), so for fixed $g\in G$, $\pi(g)^{-1}$ can be diagonalized, i.e., there exists a basis $v_1,\ldots,v_n$ of V consisting of eigenvectors of $\pi(g)^{-1}$ with eigenvaleus $\lambda_1,\ldots,\lambda_n$. As a result, the monomials $\zeta^\alpha\overline{\zeta^\beta}$ (see above) form a basis of the space of polynomials on V of bidegree (p,q) consisting of eigenvectors of $\pi_P(g)$ with eigenvalues $\lambda^\alpha\overline{\lambda^\beta}$, so that the character $\chi_{p,q}$ of the representation $\pi_{p,q}$ on the space of homogeneous polynomials on V of bidegree (p,q) induced by the given representation π of G on V is given by

$$\chi_{p,q}(g)=\text{trace }\pi_{p,q}(g)=\sum_{\substack{|\alpha|=p\\|\beta|=q}}\lambda^\alpha\overline{\lambda^\beta}.$$

Multiplying by $z^p\overline{z^q}$ and summing over p and q gives, in the sense of formal power series,

$$\sum_{p,q=0}^{\infty}\chi_{p,q}(g)z^p\overline{z^q}=\sum_{p,q=0}^{\infty}\sum_{\substack{|\alpha|=p\\|\beta|=q}}(z\lambda)^\alpha(\overline{z\lambda})^\beta$$
$$=\prod_{j=1}^{n}(1-z\lambda_j)^{-1}\prod_{j=1}^{n}(1-\overline{z\lambda}_j)^{-1}=\det(1-z\pi(g))^{-1}\det(1-\overline{z\pi(g)})^{-1}.$$

The result now follows due to a standard fact from the representation theory of compact groups, namely, that the dimension of the fixed subspace of a given representation—or to put it differently, the multiplicity with which the trivial representation occurs in a given representation—is equal to the integral of the character of that representation over the group.

An important aspect of Eqs. (19) and (20) which greatly facilitates the evaluation of the integrals is the fact that the determinants appearing under the integral signs are central functions on the group (i.e., are invariant under conjugation), so that the integral over the whole group can be reduced to an integral over the space of conjugacy classes.

Before performing this reduction, we note that the integral representations (19) and (20) are valid in two special cases which are at opposite extremes. One of these occurs when G is discrete, that is, a finite group, so that the integrals reduce to finite sums,

$$M_\pi(z)=\frac{1}{|G|}\sum_{g\in G}\frac{1}{\det(1-z\pi(g))},\tag{21}$$

$$F_\pi(z,\overline{z})=\frac{1}{|G|}\sum_{g\in G}\frac{1}{\det(1-z\pi(g))}\frac{1}{\det(1-\overline{z\pi(g)})},\tag{22}$$

which can be reduced to sums over conjugacy classes; their explicit evaluation, by means of various techniques, has been studied in the literature (see, e.g., Refs. 6, pp. 204–207, 7, and 8). The other and apparently much less studied case occurs when G is a compact connected Lie group, so that the integrals over the whole group G can be reduced to integrals over a maximal torus T: it is this situation that we shall now investigate in some detail.

Thus let G be a compact connected Lie group, let T be a maximal torus in G and let μ_T be the bi-invariant Haar measure on T, normalized so that the total volume of T with respect to μ_T is 1.

Moreover, let $\mathfrak{g}$ be the Lie algebra of G and $\mathfrak{t}\subset\mathfrak{g}$ be the Lie algebra of $T\subset G$. Introducing a G-invariant inner product $(.,.)$ on $\mathfrak{g}$, we may decompose $\mathfrak{g}$ into the orthogonal direct sum

$$\mathfrak{g}=\mathfrak{t}\oplus\mathfrak{t}^{\perp}$$

of $\mathfrak{t}$ and its orthogonal complement $\mathfrak{t}^{\perp}$; this decomposition is $\mathrm{Ad}(T)$-invariant and does not depend on the choice of the inner product $(.,.)$. Finally, let W_G be the Weyl group of G ($W_G=N_G(T)/T$ where $N_G(T)$ is the normalizer of T in G, defined by $N_G(T)=\{g\in G/gtg^{-1}\in T$ for all $t\in T\}$) and $|W_G|$ be its order. Then

$$\mathrm{Ad}(t)=1+\mathrm{Ad}^{\perp}(t)\quad\text{for } t\in T$$

and (see, e.g., Ref. 9, pp. 101–103)

$$M_{\pi}(z)=\frac{1}{|W_G|}\int_T d\mu_T(t)\,\frac{\det(1-\mathrm{Ad}^{\perp}(t))}{\det(1-z\,\pi(t))}, \tag{23}$$

$$F_{\pi}(z,\overline{z})=\frac{1}{|W_G|}\int_T d\mu_T(t)\,\frac{\det(1-\mathrm{Ad}^{\perp}(t))}{\det(1-z\,\pi(t))\det(1-\overline{z\,\pi(t)})}. \tag{24}$$

These integrals can be further evaluated in terms of the root system Δ of $\mathfrak{g}$ and the weight system Φ for the representation π. The procedure is standard when $\mathfrak{g}$ is semisimple, but to a certain degree it works just as well in the more general case when $\mathfrak{g}$ has a non-trivial center. The starting point is the fact that T being Abelian, the restriction from G to T of any unitary representation of G, such as the complexified adjoint representation Ad on the complexification $\mathfrak{g}^c$ of $\mathfrak{g}$ or the representation π on V, splits into the direct sum of irreducible one-dimensional representations. Grouping together all subspaces characterized by the same eigenvalues under all elements of T leads to the well-known root space decomposition

$$\mathfrak{g}^c=\mathfrak{t}^c\oplus\bigoplus_{\alpha\in\Delta}\mathfrak{g}_{\alpha}$$

of $\mathfrak{g}^c$ and to the weight space decomposition

$$V=\bigoplus_{\lambda\in\Phi}V_{\lambda}$$

of V; the dimension of the subspace V_{λ} is commonly known as the multiplicity $m(\lambda)$ of λ. The action of T on each of these subspaces is given by a character of T, i.e., a Lie group homomorphism from T to the unit circle, written as $t\to t^{\alpha}$ and $t\to t^{\lambda}$, respectively, according to

$$\mathrm{Ad}(t)(X)=t^{\alpha}X\quad\text{for } t\in T \text{ and } X\in\mathfrak{g}_{\alpha}, \tag{25}$$

and

$$\pi(t)(v)=t^{\lambda}v\quad\text{for } t\in T \text{ and } v\in V_{\lambda}, \tag{26}$$

respectively. At the Lie algebra level, Eqs. (25) and (26) imply the usual relations

$$\mathrm{ad}(H)(X)=\alpha(H)X\quad\text{for } H\in\mathfrak{t} \text{ and } X\in\mathfrak{g}_{\alpha}, \tag{27}$$

and

$$\dot{\pi}(H)(v)=\lambda(H)v\quad\text{for } H\in\mathfrak{t} \text{ and } v\in V_{\lambda}, \tag{28}$$

respectively. By complex conjugation of Eqs. (26) and (28), we obtain

$$\overline{\pi}(t)(v)=t^{-\lambda}v \quad \text{for } t\in T \text{ and } v\in\sigma V_\lambda, \tag{29}$$

$$\dot{\overline{\pi}}(H)(v)=-\lambda(H)v \quad \text{for } H\in\mathfrak{t} \text{ and } v\in\sigma V_\lambda. \tag{30}$$

We follow here the standard mathematical convention of considering roots α and weights λ as real linear forms on $\mathfrak{t}$, or as complex linear forms on the complexification $\mathfrak{t}^c$ of $\mathfrak{t}$, which—in accordance with the fact that $\mathfrak{g}$ is the compact real form of $\mathfrak{g}^c$—take purely imaginary values on $\mathfrak{t}$ (eigenvalues of antihermitean matrices are purely imaginary). Moreover, roots α and weights λ are transferred to generators H_α and H_λ using the isomorphism induced by the G-invariant nondegenerate complex bilinear form $(.,.)$ on $\mathfrak{g}^c$ obtained from the G-invariant inner product $(.,.)$ on $\mathfrak{g}$ by complex bilinear extension:

$$(H_\alpha,H)=\alpha(H), \quad (H_\lambda H)=\lambda(H) \quad \text{for all } H\in\mathfrak{t}^c.$$

This isomorphism is, by definition, an isometry:

$$(\alpha,\beta)=(H_\alpha,H_\beta), \quad (\lambda,\mu)=(H_\lambda,H_\mu).$$

The reality properties may then be summarized in the statement that roots α and weights λ belong to the real vector space $i\mathfrak{t}^*$, while the vectors H_α and H_λ belong to the real vector space $i\mathfrak{t}$. With this notation, we can rewrite the integrals in Eqs. (23) and (24) as follows:

$$M_\pi(z)=\frac{1}{|W_G|}\int_T d\mu_T(t)\,\frac{\Pi_{\alpha\in\Delta}(1-t^\alpha)}{\Pi_{\lambda\in\Phi}(1-zt^\lambda)^{m(\lambda)}}, \tag{31}$$

$$F_\pi(z,\overline{z})=\frac{1}{|W_G|}\int_T d\mu_T(t)\,\frac{\Pi_{\alpha\in\Delta}(1-t^\alpha)}{\Pi_{\lambda\in\Phi}(1-zt^\lambda)^{m(\lambda)}(1-\overline{zt^{-\lambda}})^{m(\lambda)}}. \tag{32}$$

A further condition to be employed is that roots α and weights λ must be integral linear forms in the sense of taking values in $2\pi i\mathbb{Z}$ on the so-called unit lattice

$$L_1=\{H\in\mathfrak{t}/\exp(H)=1\} \tag{33}$$

of G: such integral linear forms are precisely the ones that arise as differentials of characters of T (Ref. 9, pp. 94–95). This lattice is the essential ingredient for understanding how to convert the integrals (31) and (32) into integrals over the product of r unit circles ($r=\text{rank } G=\dim T$), which can then be evaluated by an r-fold successive application of the residue theorem. Indeed, let us assume that $\{2\pi H_1,\ldots,2\pi H_r\}$ is a basis of L_1 and define

$$\begin{aligned} h(\alpha)&=(h_1(\alpha),\ldots,h_r(\alpha))=(-i\alpha(H_1),\ldots,-i\alpha(H_r)) \quad \text{for } \alpha\in\Delta,\\ h(\lambda)&=(h_1(\lambda),\ldots,h_r(\lambda))=(-i\lambda(H_1),\ldots,-i\lambda(H_r)) \quad \text{for } \lambda\in\Phi. \end{aligned} \tag{34}$$

Then the map

$$\begin{aligned} &S^1\times\ldots\times S^1\to T,\\ &w=(w_1,\ldots,w_r)=(e^{i\theta_1},\ldots,e^{i\theta_r})\mapsto t=\exp\left(\sum_{j=1}^r\theta_jH_j\right) \end{aligned} \tag{35}$$

is an isomorphism of compact Abelian Lie groups such that, in multi-index notation (generalized to include negative integer powers),

$$t^{\alpha}=w^{h(\alpha)}=w_1^{h_1(\alpha)}\ldots w_r^{h_r(\alpha)} \quad \text{for } \alpha\in\Delta,$$
$$t^{\pm\lambda}=w^{\pm h(\lambda)}=w_1^{\pm h_1(\lambda)}\ldots w_r^{\pm h_r(\lambda)} \quad \text{for } \lambda\in\Phi. \tag{36}$$

Thus we obtain

Theorem 1: *Assume π is a finite-dimensional unitary representation of a compact connected Lie group G. Then in terms of the multi-index notation for roots and weights with respect to a basis of the unit lattice of G, as introduced in Eqs. (33–36), the generating functions M_π and F_π are given by*

$$M_\pi(z)=\frac{1}{|W_G|}\prod_{j=1}^{r}\oint\frac{dw_j}{2\pi i w_j}\frac{\Pi_{\alpha\in\Delta}(1-w^{h(\alpha)})}{\Pi_{\lambda\in\Phi}(1-zw^{h(\lambda)})^{m(\lambda)}}, \tag{37}$$

$$F_\pi(z,\overline{z})=\frac{1}{|W_G|}\prod_{j=1}^{r}\oint\frac{dw_j}{2\pi i w_j}\frac{\Pi_{\alpha\in\Delta}(1-w^{h(\alpha)})}{\Pi_{\lambda\in\Phi}(1-zw^{h(\lambda)})^{m(\lambda)}(1-\overline{zw}^{-h(\lambda)})^{m(\lambda)}}. \tag{38}$$

The expressions $w^{h(\alpha)}$ and $w^{\pm h(\lambda)}$ will often be abbreviated to w^{α} and $w^{\pm\lambda}$, respectively.

When G is semisimple, which is by far the most important case for applications, the exponents $h(\alpha)$ and $h(\lambda)$ are easily calculated from the root system Δ of $\mathfrak{g}$ and the weight system Φ for the representation π. To this end, it is convenient to introduce the following two lattices in $\mathfrak{t}$:

(a) the coroot lattice L_{cr}, which is dual to the standard weight lattice (Ref. 10, p. 67), in the sense that

$$L_{cr}=\{\check{\alpha}\in\mathfrak{t}/\lambda(\check{\alpha})\in 2\pi i\mathbb{Z} \quad \text{for all weights } \lambda\}, \tag{39}$$

and identical with the Δ-lattice which forms the translation part of the affine Weyl group (Ref. 11, p. 314) and generated by the vectors $4\pi i H_\alpha/(H_\alpha,H_\alpha)$ with $\alpha\in\Delta$ (Ref. 11, pp. 317–318),

(b) the coweight lattice L_{cw}, which is dual to the standard root lattice (Ref. 10, p. 67), in the sense that

$$L_{cw}=\{\check{\lambda}\in\mathfrak{t}/\alpha(\check{\lambda})\in 2\pi i\mathbb{Z} \quad \text{for all roots } \alpha\}, \tag{40}$$

and identical with the central lattice, defined as

$$L_c=\{H\in\mathfrak{t}/\exp(H)\in Z\}, \tag{41}$$

where Z is the center of G (Ref. 9, p. 95, Ref. 11, p. 311).

Obviously, the coroot lattice is contained in the coweight lattice, and the unit lattice lies in between:

$$L_{cr}\subset L_1\subset L_{cw}. \tag{42}$$

Note also that the unit lattice is sensitive to coverings, while the coroot lattice and the coweight lattice are not: they depend on the Lie group G only through its Lie algebra $\mathfrak{g}$. In fact, we have two extreme cases, between which the general case is intermediate:

(i) When G is simply connected, the unit lattice is minimal and coincides with the coroot lattice.
(ii) When G has trivial center, the unit lattice is maximal and coincides with the coweight lattice.

Therefore, when G is simply connected (which we may always assume to be the case, without loss of generality), one possible choice of the basis $\{2\pi H_1,\ldots,2\pi H_r\}$ is to set

$$H_j=\frac{2iH_{\alpha_j}}{(H_{\alpha_j},H_{\alpha_j})}, \tag{43}$$

where $\{\alpha_1,\ldots,\alpha_r\}$ is a basis of simple roots. In this case, the exponents $h(\alpha)$ and $h(\lambda)$ are precisely the coefficients in the expansion

$$\alpha=\sum_{j=1}^{r} h_j(\alpha)\lambda_j, \quad \lambda=\sum_{j=1}^{r} h_j(\lambda)\lambda_j \tag{44}$$

of a root $\alpha\in\Delta$ and of a weight $\lambda\in\Phi$ in terms of the basis $\{\lambda_1,\ldots,\lambda_r\}$ of fundamental weights, which is dual to the basis of simple roots in the usual sense:

$$\frac{2(\lambda_j,\alpha_k)}{(\alpha_k,\alpha_k)}=\delta_{jk}. \tag{45}$$

Observe that the relevant parts of this construction do not depend on the choice of the scalar product (.,.): a different choice would simply amount to a change of an overall normalization factor on each simple ideal which drops out of the definition of the generators appearing on the rhs of Eq. (43) or the definition (45) of the fundamental weights. When we want to be specific about normalization, we shall not use the Killing form, but rather the so-called standard form, which is normalized so that the long roots have length $\sqrt{2}$.

The additional assumption that G should be semisimple is less restrictive than it may seem. Indeed, when G is not semisimple, that is, has a non-discrete center Z, consider the orthogonal direct decomposition $\mathfrak{g}=\mathfrak{z}\oplus\mathfrak{g}_s$ of $\mathfrak{g}$ into its center $\mathfrak{z}$ and its maximal semisimple ideal $\mathfrak{g}_s=[\mathfrak{g},\mathfrak{g}]$, together with the corresponding orthogonal direct decomposition $\mathfrak{t}=\mathfrak{z}\oplus\mathfrak{t}_s$ of the maximal Abelian subalgebra $\mathfrak{t}$ of $\mathfrak{g}$ into the center $\mathfrak{z}$ of $\mathfrak{g}$ and a maximal Abelian subalgebra $\mathfrak{t}_s$ of $\mathfrak{g}_s$. Then the roots α ($\alpha\in\Delta$) only generate the subspace $i\mathfrak{t}_s^*$ of $i\mathfrak{t}^*$ and the vectors H_α ($\alpha\in\Delta$) only generate the subspace $i\mathfrak{t}_s$ of $i\mathfrak{t}$, as real vector spaces. We can still define the unit lattice L_1 (cf. Eq. (33)) and introduce a basis $\{2\pi H_1,\ldots,2\pi H_r\}$ as before, as well as the exponents $h(\alpha)$ and $h(\lambda)$ (cf. Eq. (34)), but the unit lattice is now very flexible: any lattice in $\mathfrak{g}$ which contains the coroot lattice of $\mathfrak{g}_s$ and whose orthogonal projection to $\mathfrak{g}_s$ is contained in the coweight lattice of $\mathfrak{g}_s$ is admissible (Ref. 9, p. 97). Therefore, there is now no general way to proceed beyond Eqs. (37) and (38); their explicit evaluation must for each representation π be carried out separately.

A final important observation concerns the form of the generating functions M_π and F_π when π is the adjoint representation Ad. In this case, we may appeal to Chevalley's theorem, which provides a complete description of the ring of invariant polynomials on a semisimple Lie algebra $\mathfrak{g}$: it is freely generated by $r=\mathrm{rank}(\mathfrak{g})$ elementary polynomials $P_1,\ldots,P_r$, whose degrees $p_1,\ldots,p_r$ are commonly known as the exponents of $\mathfrak{g}$. Therefore, the usual Molien function for the adjoint representation reads

$$M_{\mathrm{Ad}}(z)=\prod_{j=1}^{r}\frac{1}{1-z^{p_j}}. \tag{46}$$

Explicitly, for the classical groups, the polynomials $P_1,\ldots,P_r$ can (with one exception) be written as trace polynomials in the defining representation:

(i) $A_r: \mathfrak{g}=\mathfrak{sl}(r+1)$ or $\mathfrak{su}(r+1)$ (compact real form):

$$P_j(X)=\mathrm{tr}(X^{j+1}) \quad \text{for } X\in\mathfrak{g},\ j=1,...,r.$$

(ii) $B_r: \mathfrak{g}=\mathfrak{so}(2r+1,\mathbb{C})$ or $\mathfrak{so}(2r+1)$ (compact real form):

$$P_j(X)=\mathrm{tr}(X^{2j}) \quad \text{for } X\in\mathfrak{g},\ j=1,...,r.$$

(iii) $C_r: \mathfrak{g}=\mathfrak{sp}(2r,\mathbb{C})$ or $\mathfrak{sp}(2r)$ (compact real form):

$$P_j(X)=\mathrm{tr}(X^{2j}) \quad \text{for } X\in\mathfrak{g},\ j=1,...,r.$$

(iv) $D_r: \mathfrak{g}=\mathfrak{so}(2r,\mathbb{C})$ or $\mathfrak{so}(2r)$ (compact real form):

$$P_j(X)=\mathrm{tr}(X^{2j}) \quad \text{for } X\in\mathfrak{g},\ j=1,...,r-1,$$

$$P_r(X)=Pf(X) \quad \text{for } X\in\mathfrak{g},$$

where $Pf(X)$ denotes the Pfaffian of X. See, for example, Ref. 12, pp. 253–263.

It should be pointed out that Chevalley's theorem refers to invariant complex polynomials on complex semisimple Lie algebras, or equivalently, to invariant real polynomials on real semisimple Lie algebras (including compact real forms), but not to invariant real polynomials on complex semisimple Lie algebras. There is thus no reason to believe that this polynomial ring has an equally simple structure. In fact, it does not. To show this, we have calculated the generating function F_{Ad} for the simple Lie algebra $B_2=C_2$: the result (cf. Eq. (125) in Sec. VII below) exhibits a complicated structure, with lots of generators and relations.

V. COMBINATORIAL FORMULAS

In the following, we shall derive combinatorial formulas which allow one to determine the coefficients c_k and $c_{p,q}$ of the generating functions M_π and F_π solely in term of the root system Δ and the weight system Φ of the representation π.

Our starting point will be the integral representations (37) of M_π and (38) of F_π. First of all we need the following

Proposition: Let $f_1,\ldots,f_n$ be polynomials of degree 1 in the variable x (with coefficients that are rational functions of other variables $w_1,\ldots,w_r$). Then for non-negative integers $m_1,\ldots,m_n$ and k,

$$\frac{1}{k!}\frac{\partial^k}{\partial x^k}\left(\prod_{i=1}^n f_i^{-m_i}\right)=(-1)^k\sum_{|l|=k}\prod_{i=1}^n\binom{m_i-1+l_i}{l_i}\left(f_i^{-m_i-l_i}\right)\left(\frac{\partial f_i}{\partial x}\right)^{l_i}, \tag{47}$$

where we use multi-index notation, i.e., $l=(l_1,\ldots,l_n)$, where the l_i are non-negative integers, and $|l|=l_1+\ldots+l_n$.

Proof: For $k=1$, the above formula reduces to the statement that

$$\frac{\partial}{\partial x}\left(\prod_{i=1}^n f_i^{-m_i}\right)=-\sum_{j=1}^n\prod_{i=1}^n\left(f_i^{-m_i}\right)\left(\frac{m_j}{f_j}\frac{\partial f_j}{\partial x}\right),$$

which is obvious. The general case is proved by induction on k, using the hypothesis that

$$\frac{\partial^2 f_i}{\partial x}=0,$$

as follows:

$$\frac{1}{(k+1)!}\frac{\partial^{k+1}}{\partial x^{k+1}}\left(\prod_{i=1}^{n} f_i^{-m_i}\right)=\frac{(-1)^{k+1}}{k+1}\sum_{|p|=k}\sum_{j=1}^{n}\left[\prod_{\substack{i=1\\ i\neq j}}^{n}\binom{m_i-1+p_i}{p_i}(f_i^{-m_i-p_i})\left(\frac{\partial f_i}{\partial x}\right)^{p_i}\right]$$
$$\times\binom{m_j-1+p_j}{p_j}(m_j+p_j)(f_j^{-m_j-p_j-1})\left(\frac{\partial f_j}{\partial x}\right)^{p_j+1}$$
$$=(-1)^{k+1}\sum_{|p|=k}\sum_{j=1}^{n}\left[\prod_{\substack{i=1\\ i\neq j}}^{n}\binom{m_i-1+l_i}{l_i}(f_i^{-m_i-l_i})\left(\frac{\partial f_i}{\partial x}\right)^{l_i}\right]$$
$$\times\binom{m_j-1+l_j}{l_j}\frac{l_j}{k+1}(f_j^{-m_j-l_j})\left(\frac{\partial f_j}{\partial x}\right)^{l_j},$$

where l is defined in terms of p and j by putting $l_i=p_i$ for $i\neq j$ and $l_j=p_j+1$. Converting the double sum over p and j to a single sum over l yields the desired result.

As a result, we can explicitly differentiate the integrands of Eqs. (37) and (38):

$$\frac{1}{k!}\frac{\partial^k}{\partial z^k}\frac{1}{\prod_{\lambda\in\Phi}(1-zw^{\lambda})^{m(\lambda)}}\bigg|_{z=0}=\sum_{|l|=k}N(l,m)w^{\Lambda(l)},$$

$$\frac{1}{p!}\frac{1}{q!}\frac{\partial^p}{\partial z^p}\frac{\partial^q}{\partial\overline{z}^q}\frac{1}{\prod_{\lambda\in\Phi}(1-zw^{\lambda})^{m(\lambda)}(1-\overline{z}\overline{w}^{-\lambda})^{m(\lambda)}}\bigg|_{z=0,\overline{z}=0}=\sum_{|r|=p,|s|=q}N(r,m)N(s,m)w^{\Lambda(r-s)},$$

where

$$N(l,m)=\prod_{\lambda\in\Phi}\binom{m(\lambda)+l(\lambda)-1}{l(\lambda)},$$

and

$$\Lambda(l)=\sum_{\lambda\in\Phi}l(\lambda)\lambda.$$

In order to carry out the residue integrals in Eqs. (37) and (38), we must also expand the numerator in powers of w. The net result is most conveniently formulated in terms of the following concepts:

Definition 1: The extended root system $\tilde{\Delta}$ associated with a given root system Δ is the set of all linear combinations

$$\tilde{\alpha}=\sum_{\alpha\in\Delta}a(\alpha)\alpha \tag{48}$$

of roots with coefficients $a(\alpha)$ which are either 0 or 1. (Thus $\tilde{\Delta}$ is a (finite) subset of the root lattice generated by Δ.) To any such extended root $\tilde{\alpha}$, we associate its decomposition index $i(\tilde{\alpha})$, defined as the difference

$$i(\tilde{\alpha})=n_+(\tilde{\alpha})-n_-(\tilde{\alpha}), \tag{49}$$

between the number $n_+(\tilde{\alpha})$ of such decompositions of $\tilde{\alpha}$ into a sum of roots with an even number of nonzero coefficients and the number $n_-(\tilde{\alpha})$ of such decompositions of $\tilde{\alpha}$ into a sum of roots with an odd number of nonzero coefficients. In other words,

$$i(\tilde{\alpha}) = \sum_a (-1)^{|a|}, \tag{50}$$

where the sum is over all sequences $(a(\alpha))_{\alpha \in \Delta}$ of coefficients $a(\alpha) \in \{0,1\}$ satisfying Eq. (48), and

$$|a| \equiv \sum_{\alpha \in \Delta} a(\alpha). \tag{51}$$

We extend the definition of the decomposition index to the whole root lattice by setting

$$i(\tilde{\alpha}) = 0 \quad \text{if} \quad \tilde{\alpha} \notin \tilde{\Delta}. \tag{52}$$

In these terms, we have

$$\prod_{\alpha \in \Delta} (1 - w^{\alpha}) = \sum_{\tilde{\alpha} \in \tilde{\Delta}} i(\tilde{\alpha}) w^{\tilde{\alpha}}. \tag{53}$$

Note that just like the usual root system, the extended root system is invariant under the action of the corresponding Weyl group W, and so is the decomposition index (it is constant along Weyl group orbits). Two particular values that can be computed immediately are

$$i(2\rho) = (-1)^{|\Delta|/2}, \quad i(0) = |W|, \tag{54}$$

where 2ρ is the vector obtained as the sum of all positive roots. (The first formula follows by observing that in this case there is only one possible sequence, namely (1,...,1), which has the parity stated above, while the second formula follows by combining the previous formula at $w = 0$ with the fact that $M_{\pi}(0) = 1$.) For the simplest rank 1 algebra A_1, for example, we have $\Delta = \{\alpha, -\alpha\}$ and $\tilde{\Delta} = \{\alpha, 0, -\alpha\}$ with $i(\alpha) = i(-\alpha) = -1$ and $i(0) = 2$. The result for the simple rank 2 algebras A_2, $B_2 = C_2$ and G_2 is shown in Figures 1–3.

Definition 2: For any positive integer k, the k-extended weight system $\tilde{\Phi}(k)$ associated with a given weight system Φ is the set of all linear combinations

$$\tilde{\lambda} = \sum_{\lambda \in \Phi} l(\lambda) \lambda \tag{55}$$

of weights with coefficients $l(\lambda)$ which are non-negative integers, such that

$$|l| \equiv \sum_{\lambda \in \Phi} l(\lambda) = k. \tag{56}$$

(Thus $\tilde{\Phi}(k)$ is a (finite) subset of the weight lattice.) To any such k-extended weight $\tilde{\lambda}$, we associate its k-extended multiplicity $m_k(\tilde{\lambda})$, defined as the sum of the combinatorial coefficients

$$N(l,m) = \prod_{\lambda \in \Phi} \binom{m(\lambda) - 1 + l(\lambda)}{l(\lambda)} = \prod_{\lambda \in \Phi} \binom{m(\lambda) - 1 + l(\lambda)}{m(\lambda) - 1} \tag{57}$$

over all such representations of $\tilde{\lambda}$. In other words,

$$m_k(\tilde{\lambda}) = \sum_l N(l,m), \tag{58}$$

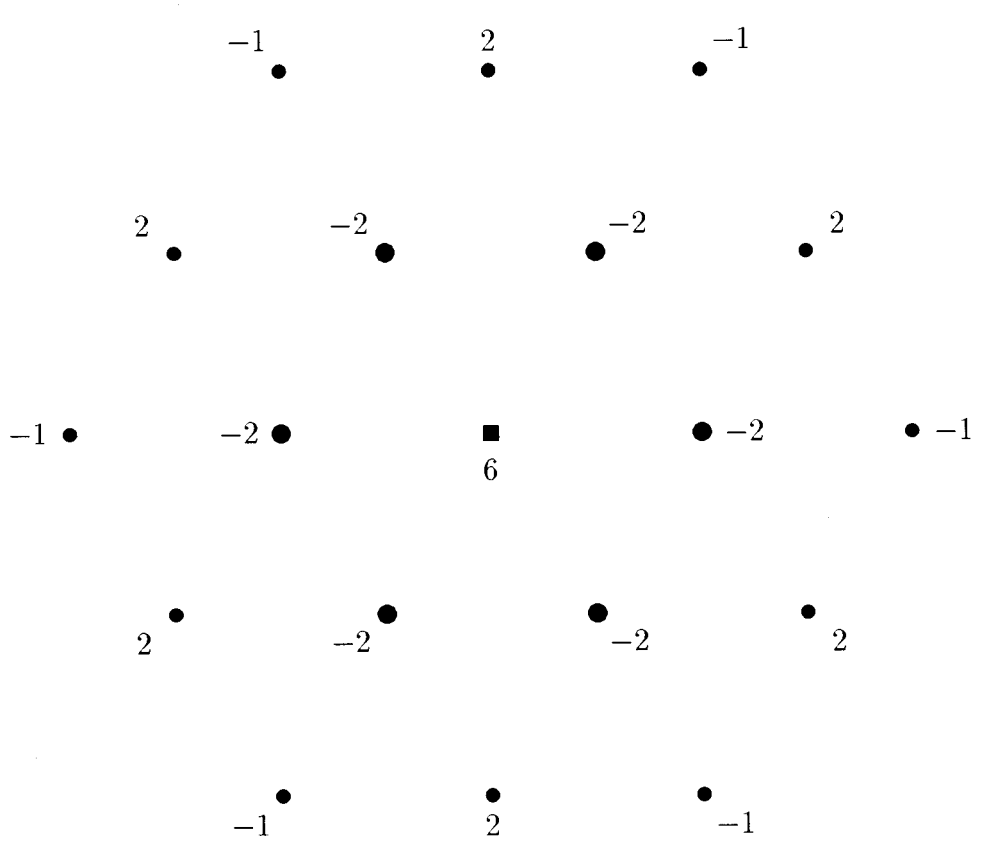

FIG. 1. Root system and extended root system, with decompostion indices: A_2.

where the sum is over all sequences $(l(\lambda))_{\lambda\in\Phi}$ of coefficients $l(\lambda)\in\mathbb{N}$ satisfying Eqs. (55) and (56).

Note that just like the usual weight system and the usual multiplicity, the k-extended weight system and the k-extended multiplicity are invariant under the action of the corresponding Weyl group W, and that $\tilde{\Phi}(1)=\Phi$, $m_1(\lambda)=m(\lambda)$.

Now we are ready to formulate the main result of this section:

Theorem 2: *Assume π is a finite-dimensional unitary representation of a compact connected Lie group G. Then in terms of the multi-index notation for roots and weights with respect to a basis of the unit lattice of G, as introduced in Eqs. (33–36), and with the notation introduced above, the number $c_k(\pi)$ of (linearly independent) G-invariant complex polynomials of degree k and the number $c_{p,q}(\pi)$ of (linearly independent) G-invariant real polynomials of bidegree (p,q) on the carrier space of π are given by the combinatorial formulas*

$$c_k(\pi)=\frac{1}{|W_G|}\sum_{\tilde{\lambda}\in\tilde{\Phi}(k)} i(\tilde{\lambda})m_k(\tilde{\lambda}), \tag{59}$$

and

$$c_{p,q}(\pi)=\frac{1}{|W_G|}\sum_{\tilde{\lambda}\in\tilde{\Phi}(p),\tilde{\mu}\in\tilde{\Phi}(q)} i(\tilde{\lambda}-\tilde{\mu})m_p(\tilde{\lambda})m_q(\tilde{\mu}), \tag{60}$$

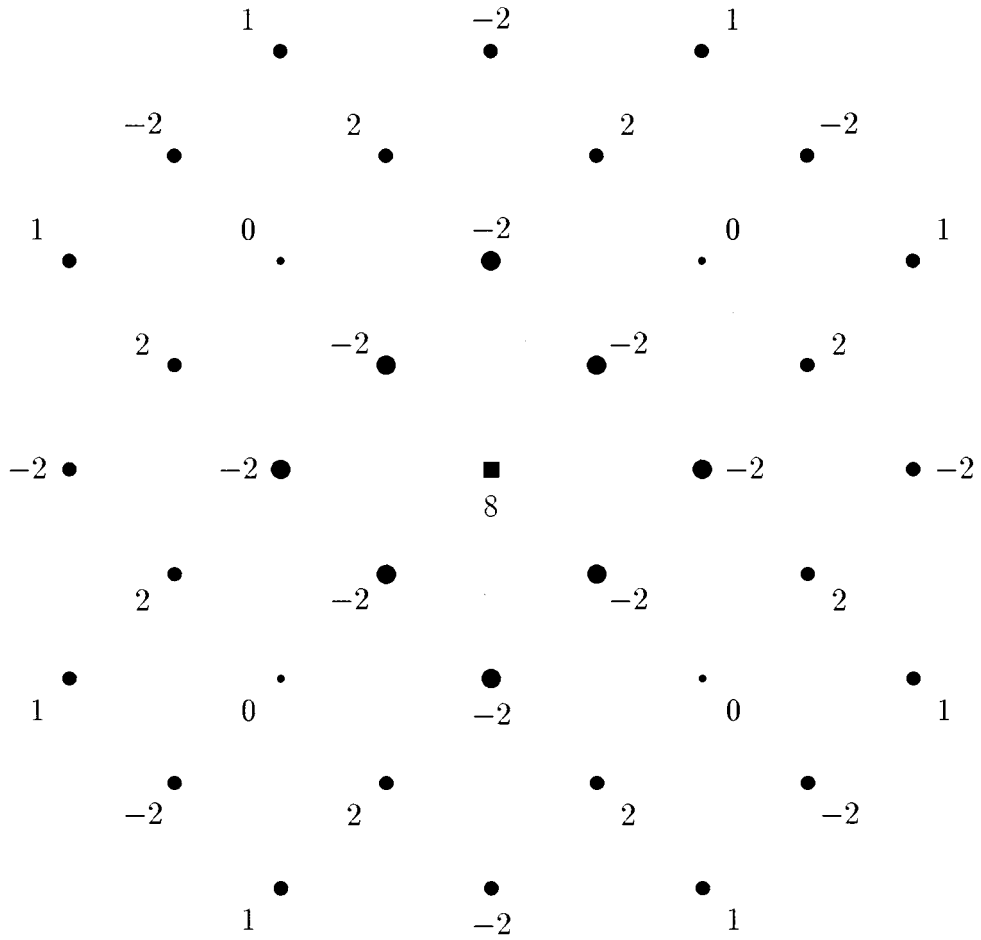

FIG. 2. Root system and extended root system, with decomposition indices: C_2.

respectively. (Note that the terms on the rhs of these equations yield non-vanishing contributions only when $\tilde{\lambda}$ *and* $\tilde{\lambda}-\tilde{\mu}$*, respectively, belongs to the extended root system* $\tilde{\Delta}$*.)*

We believe that on the basis of this theorem, it should be possible to develop a computer program for calculating the numbers c_k and $c_{p,q}$, up to reasonably high orders, for arbitrary groups and representations. The amount of computing time can be reduced by a factor of the order of $|W_G|$ by an appropriate implementation of the Weyl group symmetry.

VI. EXAMPLE: SU(2)

As a first example, let us apply Eqs. (37) and (38) to the case where $G=\mathrm{SU}(2)$, with maximal torus $T=\mathrm{U}(1)$ and $\pi=\pi_s$ the irreducible spin s representation. Then $\mathfrak{g}=\mathfrak{su}(2)$ is the Lie algebra of complex traceless antihermitean matrices and $\mathfrak{t}=\mathfrak{u}(1)$ the maximal Abelian subalgebra of imaginary traceless diagonal matrices, with invariant scalar product $(.,.)$ given by

$$(X,Y)=\mathrm{tr}(XY) \quad \text{for } X,Y\in\mathfrak{su}(2). \tag{61}$$

Using the Pauli matrices

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \quad \sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{62}$$

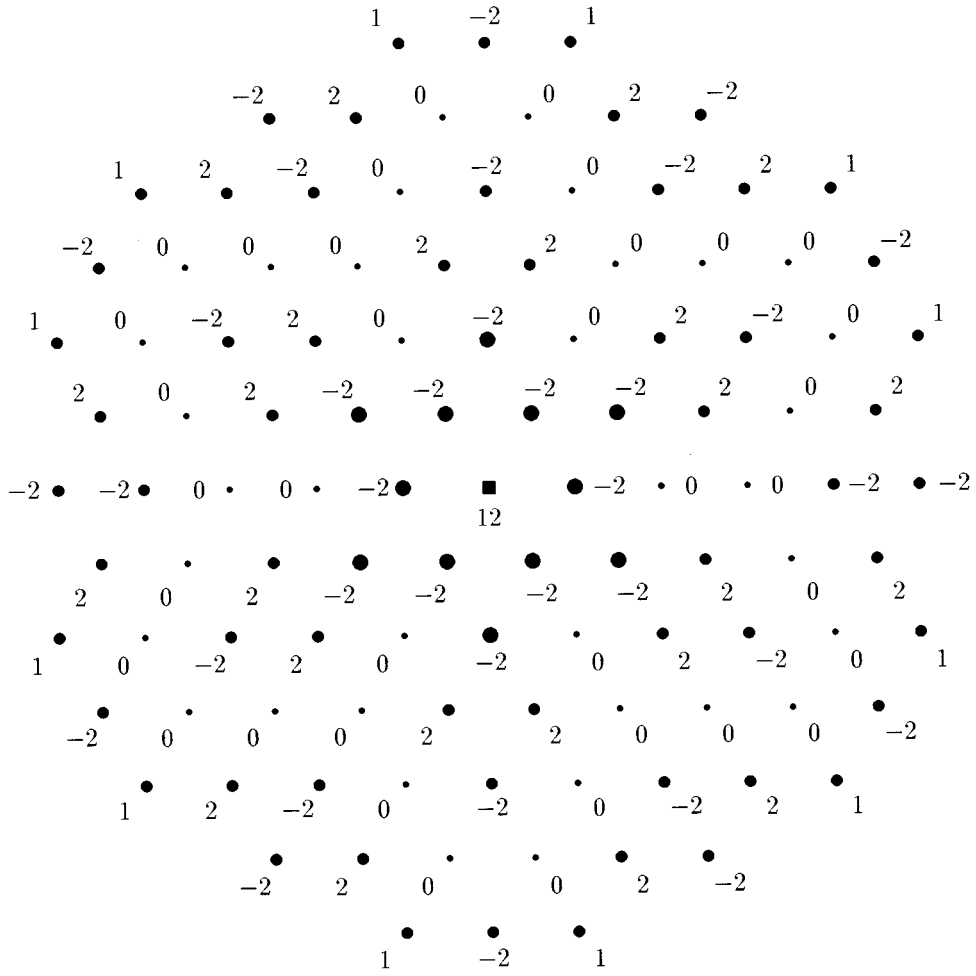

FIG. 3. Root system and extended root system, with decomposition indices: G_2.

which satisfy

$$\sigma_j \sigma_k = \delta_{jk} + i \epsilon_{jkl} \sigma_l , \tag{63}$$

we see that the nonzero roots are $\pm \alpha$, where

$$\alpha(H) = H_{11} - H_{22} = \mathrm{tr}(H \sigma_3) \quad \text{for } H \in \mathfrak{u}(1), \tag{64}$$

with root vectors

$$E_{\pm \alpha} = \sigma_1 \pm i \sigma_2 , \tag{65}$$

implying

$$H_\alpha = \sigma_3 , \tag{66}$$

which does have length $\sqrt{2}$, in accordance with our previous convention. The fundamental weight is given by

$$\lambda(H)=\tfrac{1}{2}(H_{11}-H_{22})=\tfrac{1}{2}\mathrm{tr}(H\sigma_3) \quad \text{for } H\in\mathfrak{u}(1), \tag{67}$$

so

$$H_\lambda=\tfrac{1}{2}\,\sigma_3. \tag{68}$$

Note that the root lattice is generated by α and the weight lattice is generated by λ, whereas the coroot lattice (which coincides with the unit lattice for SU(2) since SU(2) is simply connected) is generated by $2\pi H_\alpha$ and the coweight lattice (which coincides with the unit lattice for $\mathrm{SO}(3)=\mathrm{SU}(2)/\mathbb{Z}_2$ since $\mathbb{Z}_2$ is the center of SU(2)) is generated by $2\pi H_\lambda$. The highest weight of the irreducible spin s representation is precisely $2s\lambda$, and its complete weight system consists of the multiples $2m\lambda$ with m taking all integer (half-integer) values between $-s$ and s (inclusive) when s is integer (half-integer). Therefore, Eqs. (37) and (38) yield for this case

$$M_s(z)=\frac{1}{4\pi i}\oint\frac{dw}{w}\frac{(1-w^2)(1-w^{-2})}{\prod_{m=-s}^{s}(1-zw^{2m})}, \tag{69}$$

$$F_s(z,\overline{z})=\frac{1}{4\pi i}\oint\frac{dw}{w}\frac{(1-w^2)(1-w^{-2})}{\prod_{m=-s}^{s}(1-zw^{2m})(1-\overline{z}\,\overline{w}^{-2m})}. \tag{70}$$

(We have abbreviated M_{π_s} to M_s and F_{π_s} to F_s.) See Ref. 6, p. 94, noting that the formula given there is correct only for integer spin, in which case it can be deduced from Eq. (69) by the variable transformation $w=\exp(i\theta/2)$, together with the observation that the resulting integral from 0 to 4π may actually be reduced to an integral from 0 to π because the argument is a periodic function of θ with period 2π and, in addition, invariant under the reflection $\theta\to-\theta$.

For integer spin s, these equations are equivalent to

$$M_s(z)=-\frac{1}{2(1-z)}\frac{1}{2\pi i}\oint dw\,\frac{w^{s(s+1)-3}(1-w^2)^2}{\prod_{k=1}^{s}(1-zw^{2k})(w^{2k}-z)}, \tag{71}$$

$$F_s(z,\overline{z})=-\frac{1}{2(1-z)(1-\overline{z})}\frac{1}{2\pi i}\oint dw\,\frac{w^{2s(s+1)-3}(1-w^2)^2}{\prod_{k=1}^{s}(1-zw^{2k})(1-\overline{z}\,\overline{w}^{2k})(w^{2k}-z)(w^{2k}-\overline{z})}, \tag{72}$$

while for half-integer spin s, they are equivalent to

$$M_s(z)=-\frac{1}{2}\frac{1}{2\pi i}\oint dw\,\frac{w^{(s+1/2)^2-3}(1-w^2)^2}{\prod_{k=1}^{s+1/2}(1-zw^{2k-1})(w^{2k-1}-z)}, \tag{73}$$

$$F_s(z,\overline{z})=-\frac{1}{2}\frac{1}{2\pi i}\oint dw\,\frac{w^{2(s+1/2)^2-3}(1-w^2)^2}{\prod_{k=1}^{s+1/2}(1-zw^{2k-1})(1-\overline{z}\,\overline{w}^{2k-1})(w^{2k-1}-z)(w^{2k-1}-\overline{z})}. \tag{74}$$

For integer spin, however, it is more convenient to work with the complex variable $u=w^2$; then Eqs. (69), (70) and (71), (72) become

$$M_s(z)=\frac{1}{4\pi i}\oint\frac{du}{u}\frac{(1-u)(1-u^{-1})}{\prod_{m=-s}^{s}(1-zu^{m})}, \tag{75}$$

$$F_s(z,\overline{z})=\frac{1}{4\pi i}\oint\frac{du}{u}\frac{(1-u)(1-u^{-1})}{\prod_{m=-s}^{s}(1-zu^{m})(1-\overline{z}\,\overline{u}^{-m})}, \tag{76}$$

and

$$M_s(z) = -\frac{1}{2(1-z)}\frac{1}{2\pi i}\oint du\,\frac{u^{s(s+1)/2-2}(1-u)^2}{\Pi_{k=1}^{s}(1-zu^k)(u^k-z)}, \tag{77}$$

$$F_s(z,\overline{z}) = -\frac{1}{2(1-z)(1-\overline{z})}\frac{1}{2\pi i}\oint du\,\frac{u^{s(s+1)-2}(1-u)^2}{\Pi_{k=1}^{s}(1-zu^k)(1-\overline{zu}^k)(u^k-z)(u^k-\overline{z})}, \tag{78}$$

respectively. (Note that in the last four equations, a factor of 2 has disappeared because we must take into account that $u=w^2$ winds twice around the unit circle when w winds around once.)

We now proceed to calculate the generating functions M_s and F_s for a few irreducible representations of low spin, by applying the residue theorem (and remembering that $|z|<1$).

A. Spin 0

For the trivial representation, the integrands in Eqs. (77), (78) both have a double pole at $u=0$, and

$$M_0(z) = -\frac{1}{2(1-z)}\frac{1}{2\pi i}\oint du\,\frac{1-2u+u^2}{u^2},$$

$$F_0(z,\overline{z}) = -\frac{1}{2(1-z)(1-\overline{z})}\frac{1}{2\pi i}\oint du\,\frac{1-2u+u^2}{u^2},$$

that is,

$$M_0(z) = \frac{1}{1-z}, \tag{79}$$

$$F_0(z,\overline{z}) = \frac{1}{(1-z)(1-\overline{z})}. \tag{80}$$

This corresponds to the fact that for the one-dimensional trivial representation, every polynomial is invariant, and the ring of all (complex and real, respectively) polynomials in one variable is generated by the linear monomial(s) ζ and, ζ, $\overline{\zeta}$, respectively.

B. Spin 1/2

For spin 1/2, the integrand in Eq. (73) has a double pole at $w=0$, the coefficient of w in the Taylor expansion of the remaining factor around this pole being

$$\begin{aligned}
&\frac{d}{dw}\frac{(1-w^2)^2}{(1-zw)(w-z)}\bigg|_{w=0} \\
&\quad = \frac{(1-zw)(w-z)2(1-w^2)(-2w)-(1-w^2)^2(-z(w-z)+(1-zw))}{(1-zw)^2(w-z)^2}\bigg|_{w=0} \\
&\quad = -\frac{1+z^2}{z^2},
\end{aligned}$$

and a simple pole at $w=z$, whereas the integrand in Eq. (74) has simple poles at $w=0$, $w=z$ and $w=\overline{z}$, so

$$M_{1/2}(z) = -\frac{1}{2}\frac{1}{2\pi i}\oint dw\,\frac{(1-w^2)^2}{w^2(1-zw)(w-z)}$$
$$= -\frac{1}{2}\left\{-\frac{1+z^2}{z^2}+\frac{(1-z^2)^2}{z^2(1-z^2)}\right\} = \frac{1}{2}\left\{\frac{1+z^2}{z^2}-\frac{1-z^2}{z^2}\right\},$$

whereas

$$F_{1/2}(z,\overline{z}) = -\frac{1}{2}\frac{1}{2\pi i}\oint dw\,\frac{(1-w^2)^2}{w(1-zw)(1-\overline{z}w)(w-z)(w-\overline{z})}$$
$$= -\frac{1}{2}\left\{\frac{1}{z\overline{z}}+\frac{(1-z^2)^2}{z(1-z^2)(1-z\overline{z})(z-\overline{z})}+\frac{(1-\overline{z}^2)^2}{\overline{z}(1-z\overline{z})(1-\overline{z}^2)(\overline{z}-z)}\right\}$$
$$= -\frac{1}{2z\overline{z}(1-z\overline{z})(z-\overline{z})}\{(1-z\overline{z})(z-\overline{z})+\overline{z}(1-z^2)-z(1-\overline{z}^2)\},$$

that is

$$M_{1/2}(z) = 1, \tag{81}$$

$$F_{1/2}(z,\overline{z}) = \frac{1}{1-z\overline{z}}. \tag{82}$$

This confirms the idea that for the two-dimensional spinor representation, there are no invariant complex polynomials except 1—in accordance with the fact that the jth symmetric tensor power of this fundamental representation is just the irreducible representation of spin $j/2$ and therefore cannot contain the trivial representation as a subrepresentation, except when $j=0$—whereas the ring of invariant real polynomials is freely generated by the quadratic form $\zeta\cdot\zeta$, which is nothing but the invariant scalar product used in the definition of the group SU(2).

C. Spin 1

For the vector representation, the integrand in Eq. (77) has simple poles at $u=0$ and at $u=z$, whereas the integrand in Eq. (78) has simple poles at $u=z$ and $u=\overline{z}$, so

$$M_1(z) = -\frac{1}{2(1-z)}\frac{1}{2\pi i}\oint du\,\frac{(1-u)^2}{u(1-zu)(u-z)} = \frac{1}{2(1-z)}\left\{\frac{1}{z}-\frac{(1-z)^2}{z(1-z^2)}\right\},$$

whereas

$$F_1(z,\overline{z}) = -\frac{1}{2(1-z)(1-\overline{z})}\frac{1}{2\pi i}\oint du\,\frac{(1-u)^2}{(1-zu)(1-\overline{z}u)(u-z)(u-\overline{z})}$$
$$= -\frac{1}{2(1-z)(1-\overline{z})}\left\{\frac{(1-z)^2}{(1-z^2)(1-z\overline{z})(z-\overline{z})}+\frac{(1-\overline{z})^2}{(1-z\overline{z})(1-\overline{z}^2)(\overline{z}-z)}\right\}$$
$$= -\frac{1}{2(1-z)(1-\overline{z})(1-z\overline{z})(z-\overline{z})}\left\{\frac{1-z}{1+z}-\frac{1-\overline{z}}{1+\overline{z}}\right\},$$

that is,

$$M_1(z) = \frac{1}{1-z^2}, \tag{83}$$

$$F_1(z,\overline{z}) = \frac{1}{(1-z^2)(1-z\overline{z})(1-\overline{z^2})}. \tag{84}$$

This confirms the idea that for the three-dimensional vector representation, the ring of invariant (complex and real, respectively) polynomials is freely generated by the quadratic forms $\boldsymbol{\zeta}^2$ and $\boldsymbol{\zeta}^2$, $\boldsymbol{\zeta}\cdot\overline{\boldsymbol{\zeta}}$, $\overline{\boldsymbol{\zeta}^2}$, respectively. Alternatively, representing three-dimensional vectors as anti-symmetric (3×3)-matrices A, these generators may be written as $\mathrm{tr}(A^2)$ and $\mathrm{tr}(A^2)$, $\mathrm{tr}(AA)$, $\mathrm{tr}(\overline{A^2})$, respectively.

For higher spin, these calculations become increasingly cumbersome because, according to Eqs. (77), (78) and (73), (74), the integrands have poles at $|z|^{1/p}$ times the pth roots of unity, for all integers p from 1 to s if s is integer and for all odd integers p from 1 to $2s$ if s is half-integer. They can, however, be simplified by combining a decomposition of the integrand into partial fractions with the fact that the residue integral with a single factor in the denominator can be easily evaluated, even when the numerator is a complicated polynomial in the integration variable, without having to sum over roots of unity. In fact, we may use the following elementary

Proposition: Let P be a polynomial in w with coefficients that are rational functions of z (and possibly of other variables $z_1,\ldots,z_r$):

$$P(z_1,\ldots,z_r,z,w) = \sum_{n=1}^{N} a_n(z_1,\ldots,z_r,z)w^{n-1}. \tag{85}$$

For any integer $k\geqslant 1$, let M be the largest integer such that $kM\leqslant N$, and define

$$Q_k(z_1,\ldots,z_r,z) = \sum_{m=1}^{M} a_{km}(z_1,\ldots,z_r,z)z^{m-1}. \tag{86}$$

Then

$$\oint \frac{dw}{2\pi i}\,\frac{P(z_1,\ldots,z_r,z,w)}{w^k-z} = Q_k(z_1,\ldots,z_r,z). \tag{87}$$

Proof: Let a be any kth root of z and ϵ be any primitive kth root of unity, e.g., $\epsilon=\exp(2\pi i/k)$. Then

$$\sum_{l=0}^{k-1} x^l = \frac{x^k-1}{x-1} = \prod_{j=1}^{k-1}(x-\epsilon^j),$$

so taking the limit $x\to 1$, we obtain

$$\prod_{j=1}^{k-1}(1-\epsilon^j) = k,$$

while putting $x=\epsilon$ gives

$$\sum_{l=0}^{k-1}\epsilon^l = 0.$$

More generally, we have

$$\sum_{l=0}^{k-1}\epsilon^{ln} = 0$$

if k and n are relatively prime, because in this case, ln mod k will assume every value between 0 and $k-1$ exactly once when l ranges from 0 to $k-1$. Still more generally, we have

$$\sum_{l=0}^{k-1} \epsilon^{ln} = \begin{cases} 0 & \text{if } n \text{ is not a multiple of } k \\ k & \text{if } n \text{ is a multiple of } k \end{cases},$$

because denoting by m the greatest common divisor of k and n, $p=k/m$ and $r=n/m$ are relatively prime while ϵ^m will be a primitive pth root of unity, so decomposing the summation variable l according to $l=pj+i$ and using the previous equation,

$$\sum_{l=0}^{k-1} \epsilon^{ln} = \sum_{l=0}^{k-1} (\epsilon^m)^{lr} = \sum_{j=0}^{m-1} \sum_{i=0}^{p-1} (\epsilon^m)^{ir} = m \sum_{i=0}^{p-1} (\epsilon^m)^{ir} = \begin{cases} 0 & \text{if } m \neq k \\ k & \text{if } m = k \end{cases}.$$

Now we are ready to prove the proposition (for simplicity, we suppress the variables $z_1,\ldots,z_r$):

$$\oint \frac{dw}{2\pi i} \frac{P(z,w)}{w^k - z} = \oint \frac{dw}{2\pi i} \frac{P(z,w)}{\prod_{m=0}^{k-1}(w-\epsilon^m a)} = \sum_{l=0}^{k-1} \left. \frac{P(z,w)}{\prod_{\substack{m=0 \\ m\neq l}}^{k-1}(w-\epsilon^m a)} \right|_{w=\epsilon^l a}$$

$$= \sum_{l=0}^{k-1} \frac{P(z,\epsilon^l a)}{\prod_{\substack{m=0 \\ m\neq l}}^{k-1}(\epsilon^l a - \epsilon^m a)} = \sum_{l=0}^{k-1} \frac{P(z,\epsilon^l a)}{(\epsilon^l a)^{k-1} \prod_{j=1}^{k-1}(1-\epsilon^j)} = \frac{1}{kz} \sum_{l=0}^{k-1} \epsilon^l a P(z,\epsilon^l a)$$

$$= \frac{1}{kz} \sum_{n=1}^{N} \sum_{l=0}^{k-1} a_n(z) \epsilon^{ln} a^n = \frac{1}{z} \sum_{\substack{n=1 \\ n \text{ multiple of } k}}^{N} a_n(z) z^{n/k} = Q_k(z).$$

This proposition is combined with partial fraction decompositions of the form

$$\prod_{k=1}^{s} \frac{1}{(w^k - z)(1 - z w^k)} = \sum_{k=1}^{s} \left(\frac{a_k(z,w)}{w^k - z} + \frac{b_k(z,w)}{1 - z w^k} \right),$$

$$\prod_{k=1}^{s} \frac{1}{(w^k - z)(w^k - \overline{z})(1 - z w^k)(1 - \overline{z}\,\overline{w}^k)}$$

$$= \sum_{k=1}^{s} \left(\frac{a_k(\overline{z},z,w)}{w^k - z} + \frac{a_k(z,\overline{z},w)}{w^k - \overline{z}} + \frac{b_k(\overline{z},z,w)}{1 - z w^k} + \frac{b_k(z,\overline{z},w)}{1 - \overline{z}\,\overline{w}^k} \right),$$

for integer spin s, where the a_k and b_k are polynomials in w of degree strictly less than k whose coefficients are rational functions of z and of z, $\overline{z}$, respectively, and of the form

$$\prod_{k=1}^{s+1/2} \frac{1}{(w^{2k-1} - z)(1 - z w^{2k-1})} = \sum_{k=1}^{s+1/2} \left(\frac{a_k(z,w)}{w^{2k-1} - z} + \frac{b_k(z,w)}{1 - z w^{2k-1}} \right),$$

$$\prod_{k=1}^{s+1/2} \frac{1}{(w^{2k-1} - z)(w^{2k-1} - \overline{z})(1 - z w^{2k-1})(1 - \overline{z}\,\overline{w}^{2k-1})}$$

$$= \sum_{k=1}^{s+1/2} \left(\frac{a_k(\overline{z},z,w)}{w^{2k-1} - z} + \frac{a_k(z,\overline{z},w)}{w^{2k-1} - \overline{z}} + \frac{b_k(\overline{z},z,w)}{1 - z w^{2k-1}} + \frac{b_k(z,\overline{z},w)}{1 - \overline{z}\,\overline{w}^{2k-1}} \right),$$

for half-integer spin s, where the a_k and b_k are polynomials in w of degree strictly less than $2k-1$ whose coefficients are rational functions of z and of z, $\overline{z}$, respectively; these functions and

coefficients can be calculated recursively, by induction on the number of factors (that is, induction on s) and repeated application of the Euclid algorithm. Using these two techniques, we have developed, within MAPLE, a program which allows to calculate the function F up to spin 2 and the function M up to spin 4 within reasonable limits of a few minutes of computing time on a standard 386 PC, with the following results:

Spin 3/2:

$$M_{3/2}(z)=\frac{1}{1-z^4}, \tag{88}$$

$$F_{3/2}(z,\overline{z})=\frac{(1-z^4\overline{z^4})(1-z^6\overline{z^6})}{(1-z\overline{z})(1-z^4)(1-z^3\overline{z})(1-z\overline{z^3})(1-\overline{z^4})(1-z^3\overline{z^3})}. \tag{89}$$

Spin 2:

$$M_2(z)=\frac{1}{(1-z^2)(1-z^3)}, \tag{90}$$

$$F_2(z,\overline{z})=\frac{1-z^6\overline{z^6}}{(1-z^2)(1-z\overline{z})(1-\overline{z^2})(1-z^3)(1-z^2\overline{z})(1-z\overline{z^2})(1-\overline{z^3})(1-z^2\overline{z^2})}. \tag{91}$$

Spin 5/2:

$$M_{5/2}(z)=\frac{1-z^{36}}{(1-z^4)(1-z^8)(1-z^{12})(1-z^{18})}. \tag{92}$$

Spin 3:

$$M_3(z)=\frac{1-z^{30}}{(1-z^2)(1-z^4)(1-z^6)(1-z^{10})(1-z^{15})}. \tag{93}$$

Spin 7/2:

$$\begin{aligned}\mathrm{numer}(M_{7/2}(z))=&1+2z^8+4z^{12}+4z^{14}+5z^{16}+9z^{18}+6z^{20}+9z^{22}+8z^{24}+9z^{26}+6z^{28}+9z^{30}\\&+5z^{32}+4z^{34}+4z^{36}+2z^{40}+z^{48},\end{aligned} \tag{94}$$

$$\mathrm{denom}(M_{7/2}(z))=(1-z^4)(1-z^8)(1-z^{12})^2(1-z^{20}).$$

Spin 4:

$$M_4(z)=\frac{1+z^8+z^9+z^{10}+z^{18}}{(1-z^2)(1-z^3)(1-z^4)(1-z^5)(1-z^6)(1-z^7)}. \tag{95}$$

These formulas lead to several interesting observations.

To begin with, let us comment on the result for the five-dimensional spin 2 representation, which can be interpreted most conveniently by realizing five-dimensional vectors as traceless symmetric (3×3)-matrices A. Equation (90) confirms the idea that the ring of invariant complex polynomials in this representation is freely generated by the quadratic form $\mathrm{tr}(A^2)$ and the cubic form $\mathrm{tr}(A^3)$, whereas Eq. (91) states that the ring of invariant real polynomials in this representation is generated by three quadratic forms of bidegree (2,0), (1,1) and (0,2), respectively, together with four cubic forms of bidegree (3,0), (2,1), (1,2) and (0,3), respectively, plus an extra quartic form of bidegree (2,2), and that these generators should satisfy a relation of bidegree (6,6).

Obviously, the quadratic forms are given by $\mathrm{tr}(A^2)$, $\mathrm{tr}(A\overline{A})$ and $\mathrm{tr}(\overline{A^2})$, respectively, and the cubic forms by $\mathrm{tr}(A^3)$, $\mathrm{tr}(A^2\overline{A})$, $\mathrm{tr}(A\overline{A^2})$ and $\mathrm{tr}(\overline{A^3})$, respectively, but the extra generator of bidegree (2,2) comes as a surprise, especially in view of the fact that the usual Molien function provides no hint toward its existence. To see why there should be such an extra generator, note first that the natural invariant quartic forms $\mathrm{tr}(A^4)$ of bidegree (4,0), $\mathrm{tr}(A^3\overline{A})$ of bidegree (3,1), $\mathrm{tr}(A\overline{A^3})$ of bidegree (1,3) and $\mathrm{tr}(\overline{A^4})$ of bidegree (0,4) are not independent, but can be expressed in terms of the natural invariant quadratic forms $\mathrm{tr}(A^2)$, $\mathrm{tr}(A\overline{A})$ and $\mathrm{tr}(\overline{A^2})$:

$$\mathrm{tr}(A^4) = \tfrac{1}{2}\,(\mathrm{tr}(A^2))^2,$$

$$\mathrm{tr}(A^3\overline{A}) = \tfrac{1}{2}\,\mathrm{tr}(A^2)\mathrm{tr}(A\overline{A}),$$

$$\mathrm{tr}(A\overline{A^3}) = \tfrac{1}{2}\,\mathrm{tr}(A\overline{A})\mathrm{tr}(\overline{A^2}),$$

$$\mathrm{tr}(\overline{A^4}) = \tfrac{1}{2}\,(\mathrm{tr}(\overline{A^2}))^2.$$

On the other hand, there are four natural invariant quartic forms of bidegree (2,2), namely

$$\mathrm{tr}(A^2\overline{A^2}),\ \ \mathrm{tr}((A\overline{A})^2),\ \ \mathrm{tr}(A^2)\mathrm{tr}(\overline{A^2}),\ \ \mathrm{tr}(A\overline{A})^2,$$

between which there exists precisely one linear relation:

$$4\,\mathrm{tr}(A^2\overline{A^2}) + 2\,\mathrm{tr}((A\overline{A})^2) = \mathrm{tr}(A^2)\mathrm{tr}(\overline{A^2}) - 2\ \ \mathrm{tr}(A\overline{A})^2.$$

The extra generator of bidegree (2,2) can therefore be chosen to be any linear combination of $\mathrm{tr}(A^2\overline{A^2})$ and $\mathrm{tr}((A\overline{A})^2)$ which is not proportional to the lhs of this equation. As far as the relation of bidegree (6,6) is concerned, we have not been patient enough to determine its explicit form: this seems a formidable task in view of the fact that a power series expansion of Eq. (91) shows, using MAPLE, that the coefficient of $z^6\overline{z^6}$ in $F_2(z,\overline{z})$ is 36, so one has to find exactly one linear relation between 37 polynomials of bidegree (6,6) in 10 variables (5 holomorphic and 5 antiholomorphic)!

A similar situation, though somewhat more complicated, occurs for the four-dimensional spin 3/2 representation. Here, we encounter one invariant quadratic form of bidegree (1,1) (the invariant scalar product, as usual) and four invariant quartic forms of bidegree (4,0), (3,1), (1,3) and (0,4), respectively, plus an extra invariant form of bidegree (3,3) (besides the cube of the invariant scalar product), as generators. There are two relations: one relation of bidegree (4,4), expressing a linear dependence between the four invariant polynomials $P(4,0)\cdot P(0,4)$, $P(3,1)\cdot P(1,3)$, $P(1,1)^4$ and $P(1,1)\cdot P(3,3)$, and one relation of bidegree (6,6), expressing a linear dependence between the seven invariant polynomials $P(4,0)\cdot P(0,4)\cdot P(1,1)^2$, $P(3,1)\cdot P(1,3)\cdot P(1,1)^2$, $P(1,1)^6$, $P(1,1)^3\cdot P(3,3)$, $P(4,0)\cdot P(1,3)^2$, $P(0,4)\cdot P(3,1)^2$ and $P(3,3)^2$, over and above the relation obtained by multiplying the previous one by $P(1,1)^2$. These two relations are, however, not independent, because the presence of an additional generator of bidegree (10,10) suggests that their product reduces to a trivial identity.

As far as the Molien functions for representations of spin >2 are concerned, the results indicate that for spin 5/2 and spin 3, the generators of the ring of invariant complex polynomials are subject to a single relation, while for spin 7/2 and spin 4, the structure of the relations themselves becomes complicated and no longer fits into the relatively simple scheme given by Eq. (11): there are lots of additional generators satisfying complicated relations, relations between the relations, etc. It is not even clear what is in general the most adequate way to present these functions, since numerator and denominator may have common factors. For example, the numerator and the denominator of the Molien function for spin 7/2 as given in Eq. (94) have a common factor $(1+z^6)(1+z^{10})$ which has been introduced to eliminate factors $1-z^6$ and $1-z^{10}$ from the denominator, so as to comply with the fact that there are no invariant polynomials of degree 6 and

of degree 10 in this representation (as can be seen upon Taylor expansion). At any rate, the polynomials in the numerator of Eq. (94) and of Eq. (95) have roots of modulus $\neq 1$ and hence cannot possibly be reduced to an expression of the form $1-z^k$ or to a product of such expressions. (More precisely, a numerical calculation, using MAPLE, shows that the quotient of the numerator of Eq. (94) by $(1+z^6)(1+z^{10})$, considered as a polynomial of degree 16 in z^2, has 8 roots of modulus 1, 2 roots of modulus 1.46292, 2 roots of modulus 1.36453, 2 roots of modulus 0.73285 and 2 roots of modulus 0.68356, while the numerator of Eq. (95) has 14 roots of modulus 1, 2 roots of modulus 1.10697 and 2 roots of modulus 0.90377.)

In summary, everything indicates that with increasing spin, the situation becomes extremely complex. We shall therefore not pursue this matter any further and instead pass to other groups and representations.

VII. THE CODON REPRESENTATION AND ITS REDUCTIONS

Apart from the circle group, U(1), and the ordinary rotation group (or rather its universal covering group), SU(2), the compact simple Lie groups appearing in the symmetry breaking scheme of Hornos and Hornos[1] for describing the degeneracy of the genetic code involve the symplectic groups Sp(4) and Sp(6). With these applications in mind, we begin by collecting a few pertinent facts about the symplectic groups $\mathrm{Sp}(2r)$ and some of their irreducible representations, especially for the cases $r=2$ and $r=3$. (Note that we shall be dealing exclusively with the compact real form $\mathrm{Sp}(2r)$ of the complex symplectic group $\mathrm{Sp}(2r,\mathbb{C})$, which can be defined as a group of $(r\times r)$-matrices with quaternionic entries, not with the normal real form $\mathrm{Sp}(2r,\mathbb{R})$ that appears, e.g., in Hamiltonian mechanics.

The symplectic group $\mathrm{Sp}(2r)$ is a compact, connected, simply connected Lie group with center $\mathbb{Z}_2$, and its Lie algebra $\mathfrak{sp}(2r)$ is the compact real form of the complex simple Lie algebra $\mathfrak{sp}(2r,\mathbb{C})$; in the Cartan classification this is C_r, of rank r and dimension $r(2r+1)$. To construct its root system and the weight systems of various other irreducible representations besides the adjoint, we identify the spaces $\mathfrak{t}$ and $i\mathfrak{t}^*$ used before (cf. Sec. 4) with $\mathbb{R}^r$ by introducing bases $\{H_1,\ldots,H_r\}$ of $\mathfrak{t}$ and $\{e_1,\ldots,e_r\}$ of $i\mathfrak{t}^*$, dual to each other in the sense that

$$e_j(H_k)=i\,\delta_{jk}, \tag{96}$$

and orthonormal except for an overall normalization factor of $\sqrt{2}$; more precisely, we assume that

$$(e_j,e_k)=\tfrac{1}{2}\,\delta_{jk}, \quad (H_j,H_k)=2\,\delta_{jk}. \tag{97}$$

Then the root system Δ of $\mathfrak{sp}(2r)$, when written as the disjoint union

$$\Delta=\Delta_l\cup\Delta_s \tag{98}$$

of the set Δ_l of long roots (of length $\sqrt{2}$) and the set Δ_s of short roots (of length 1), is given by

$$\Delta_l=\{\pm 2e_j/1\leqslant j\leqslant r\}, \tag{99}$$

$$\Delta_s=\{\pm e_j\pm e_k/1\leqslant j<k\leqslant r\}. \tag{100}$$

(All signs are to be read independently.) We choose an ordering in this root system such that the set of positive roots becomes

$$\Delta^+=\Delta_l^+\cup\Delta_s^+, \tag{101}$$

where

$$\Delta_l^+=\{2e_j/1\leqslant j\leqslant r\}, \tag{102}$$

$$\Delta_s^+ = \{e_j \pm e_k / 1 \leqslant j < k \leqslant r\}, \tag{103}$$

leading to the following basis $\{\alpha_1, \ldots, \alpha_r\}$ of simple roots:

$$\alpha_1 = e_1 - e_2, \quad \ldots, \quad \alpha_{r-1} = e_{r-1} - e_r, \quad \alpha_r = 2e_r. \tag{104}$$

The highest root is

$$\delta = 2e_1. \tag{105}$$

Moreover, the vector ρ defined as half the sum of the positive roots, or equivalently, as the sum of the fundamental weights, which plays an important role in representation theory, is given by

$$\rho = re_1 + (r-1)e_2 + \ldots + 2e_{r-1} + e_r. \tag{106}$$

Passing to irreducible representations, we first compute the fundamental weights, defined by the condition (45), which in the present case leads to

$$\lambda_1 = e_1, \quad \lambda_2 = e_1 + e_2, \quad \ldots, \quad \lambda_{r-1} = e_1 + \ldots + e_{r-1}, \quad \lambda_r = e_1 + \ldots + e_r. \tag{107}$$

This implies that $\{e_1, \ldots, e_r\}$ is a basis of the weight lattice and $\{2\pi H_1, \ldots, 2\pi H_r\}$ is a basis of the coroot lattice (which coincides with the unit lattice for $\mathrm{Sp}(2r)$ since $\mathrm{Sp}(2r)$ is simply connected); these are much more convenient than the basis $\{\lambda_1, \ldots, \lambda_r\}$ of fundamental weights and the basis $\{2\pi\check{\alpha}_1, \ldots, 2\pi\check{\alpha}_r\}$ formed by the simple coroots, respectively, because they are orthonormal (except for the aforementioned overall normalization factor of $\sqrt{2}$).

With these generalities out of the way, we can proceed to write down the weight systems and, as a consequence, the generating functions M_π and F_π for the irreducible representations of Sp(4) and Sp(6) that appear in Ref. 1. For completeness, we also list their dimension and height, recalling that all irreducible representations of $\mathrm{Sp}(2r)$ are self-conjugate and that the height $\mathrm{ht}(\Lambda)$ of a self-conjugate representation of highest weight Λ allows one to decide whether the representation is real or pseudo-real: it is real iff $\mathrm{ht}(\Lambda)$ is even and pseudo-real iff $\mathrm{ht}(\Lambda)$ is odd (see Ref. 13, pp. 31–33). Finally, we list the coefficients $c_{p,q}$ in the Taylor expansion of F_π, up to fourth order, which have been calculated by differentiating under the integral sign and then computing the residues, using MAPLE. The results obtained for $c_{0,0}$, $c_{1,0}$, $c_{0,1}$ and $c_{1,1}$ are not listed because they come out to be what they must be for any irreducible representation:

$$c_{0,0} = 1, \quad c_{1,0} = 0 = c_{0,1}, \quad c_{1,1} = 1. \tag{108}$$

The result for $c_{0,0}$ reflects the correct normalization: there is (up to a constant multiple) always precisely one invariant polynomial of bidegree (0,0), namely the constant 1. The results for $c_{1,0}$ and for $c_{0,1}$ reflect the fact that an irreducible representation does not admit any invariant vectors, while the result for $c_{1,1}$ corresponds to the theorem that in an irreducible representation, the invariant scalar product is unique (up to a constant multiple), due to Schur's lemma.

A. Sp(4)

For an irreducible representation π_Λ of highest weight $\Lambda = a_1\lambda_1 + a_2\lambda_2$, we have

$$\dim(\pi_\Lambda) = (1+a_1)(1+a_2)\left(1 + \frac{a_1+a_2}{2}\right)\left(1 + \frac{a_2+2a_3}{3}\right), \tag{109}$$

$$\mathrm{ht}(\pi_\Lambda) = 3a_1 + 4a_2. \tag{110}$$

The Weyl group of Sp(4) is $\mathbb{Z}_2 \times \mathbb{Z}_2 \times S_2$, generated by the reflections $e_1 \leftrightarrow -e_1$, $e_2 \leftrightarrow -e_2$ and the permutation of e_1 with e_2; it has order 8. Therefore,

$$M_\Lambda(z)=\frac{1}{8}\oint\frac{dw_1}{2\pi i}\oint\frac{dw_2}{2\pi i}\frac{w_1^{-5}w_2^{-5}f_N(w_1,w_2)^2}{f_{\Lambda,D}(z,w_1,w_2)},\tag{111}$$

$$F_\Lambda(z,\overline{z})=\frac{1}{8}\oint\frac{dw_1}{2\pi i}\oint\frac{dw_2}{2\pi i}\frac{w_1^{-5}w_2^{-5}f_N(w_1,w_2)^2}{f_{\Lambda,D}(z,w_1,w_2)f_{\Lambda,D}(\overline{z},w_1,w_2)},\tag{112}$$

with

$$f_N(w_1,w_2)=(1-w_1^2)(1-w_2^2)(1-w_1w_2)(w_1-w_2).\tag{113}$$

(The definition of $f_{\Lambda,D}$ follows below.)

1. First fundamental representation (1,0)

The highest weight is $\Lambda=\lambda_1=e_1$, the dimension is 4, the height is 3 (so the representation is pseudo-real), the complete weight system (cf. Figure 4) is

$$\Phi_{(1,0)}=\{\pm e_1,\pm e_2\}\tag{114}$$

(all weights have multiplicity 1; the signs are to be read independently), so

$$f_{(1,0),D}(x,w_1,w_2)=(1-xw_1)(1-xw_1^{-1})(1-xw_2)(1-xw_2^{-1}).\tag{115}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=0=c_{0,2},$$

$$c_{3,0}=0=c_{0,3},\quad c_{2,1}=0=c_{1,2},\tag{116}$$

$$c_{4,0}=0=c_{0,4},\quad c_{3,1}=0=c_{1,3},\quad c_{2,2}=1.$$

Without much difficulty, the generating function F can be computed in closed form; the result is

$$F_{(1,0)}(z,\overline{z})=\frac{1}{1-z\overline{z}}\tag{117}$$

2. Second fundamental representation (0,1)

The highest weight is $\Lambda=\lambda_2=e_1+e_2$, the dimension is 5, the height is 4 (so the representation is real), the complete weight system (cf. Figure 4) is

$$\Phi_{(0,1)}=\{0,\pm e_1\pm e_2\}\tag{118}$$

(all weights have multiplicity 1; the signs are to be read independently), so

$$f_{(0,1),D}(x,w_1,w_2)=(1-x)(1-xw_1w_2)(1-xw_1^{-1}w_2)(1-xw_1w_2^{-1})(1-xw_1^{-1}w_2^{-1}).\tag{119}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=1=c_{0,2},$$

$$c_{3,0}=0=c_{0,3},\quad c_{2,1}=0=c_{1,2},\tag{120}$$

$$c_{4,0}=1=c_{0,4}, \quad c_{3,1}=1=c_{1,3}, \quad c_{2,2}=2.$$

Again, the generating function F can be computed in closed form; the result is

$$F_{(0,1)}(z,\overline{z})=\frac{1}{(1-z^2)(1-z\overline{z})(1-\overline{z^2})}. \tag{121}$$

3. Adjoint representation (2,0)

The highest weight is $\Lambda=2\lambda_1=2e_1$, the dimension is 10, the height is 6 (so the representation is real), the complete weight system (cf. Figure 4) is the union of $\{0\}$ with the root system Δ,

$$\Phi_{(2,0)}=\{0\}\cup\{\pm 2e_1,\pm 2e_2,\pm e_1\pm e_2\} \tag{122}$$

(0 has multiplicity 2 and the roots have multiplicity 1; the signs are to be read independently), so

$$\begin{aligned} f_{(2,0),D}(x,w_1,w_2)=&(1-x)^2(1-xw_1^2)(1-xw_1^{-2})(1-xw_2^2)(1-xw_2^{-2})\\ &\times(1-xw_1w_2)(1-xw_1^{-1}w_2)(1-xw_1w_2^{-1})(1-xw_1^{-1}w_2^{-1}). \end{aligned} \tag{123}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=1=c_{0,2},$$

$$c_{3,0}=0=c_{0,3}, \quad c_{2,1}=0=c_{1,2}, \tag{124}$$

$$c_{4,0}=2=c_{0,4}, \quad c_{3,1}=2=c_{1,3}, \quad c_{2,2}=4.$$

With considerable effort, the generating function F can be computed in closed form; the result is

$$\begin{aligned} \mathrm{numer}(F_{(2,0)}(z,\overline{z}))=&1+z^2\overline{z^2}+z^3\overline{z^3}+z^4\overline{z^3}+z^3\overline{z^4}+z^4\overline{z^4}+z^6\overline{z^3}+z^5\overline{z^4}\\ &+z^4\overline{z^5}+z^3\overline{z^6}+z^5\overline{z^5}+z^6\overline{z^5}+z^5\overline{z^6}+z^6\overline{z^6}+z^7\overline{z^7}+z^9\overline{z^9}, \end{aligned} \tag{125}$$

$$\begin{aligned} \mathrm{denom}(F_{(2,0)}(z,\overline{z}))=&(1-z^2)(1-z\overline{z})(1-\overline{z^2})(1-z^4)(1-z^3\overline{z})(1-z^2\overline{z^2})\\ &\times(1-z\overline{z^3})(1-\overline{z^4})(1-z^4\overline{z^2})(1-z^2\overline{z^4}). \end{aligned}$$

4. Reduced codon representation (1,1)

The highest weight is $\Lambda=\lambda_1+\lambda_2=2e_1+e_2$, the dimension is 16, the height is 7 (so the representation is pseudo-real), the complete weight system (cf. Figure 4) is the union

$$\Phi(1,1)=\Phi^{(1)}_{(1,1)}\cup\Phi^{(2)}_{(1,1)}, \tag{126}$$

where

$$\Phi^{(1)}_{(1,1)}=\{\pm 2e_1\pm e_2,\pm 2e_2\pm e_1\} \tag{127}$$

(these weights have multiplicity 1; the signs are to be read independently) and

$$\Phi^{(2)}_{(1,1)}=\{\pm e_1,\pm e_2\} \tag{128}$$

(these weights have multiplicity 2; the signs are to be read independently), so

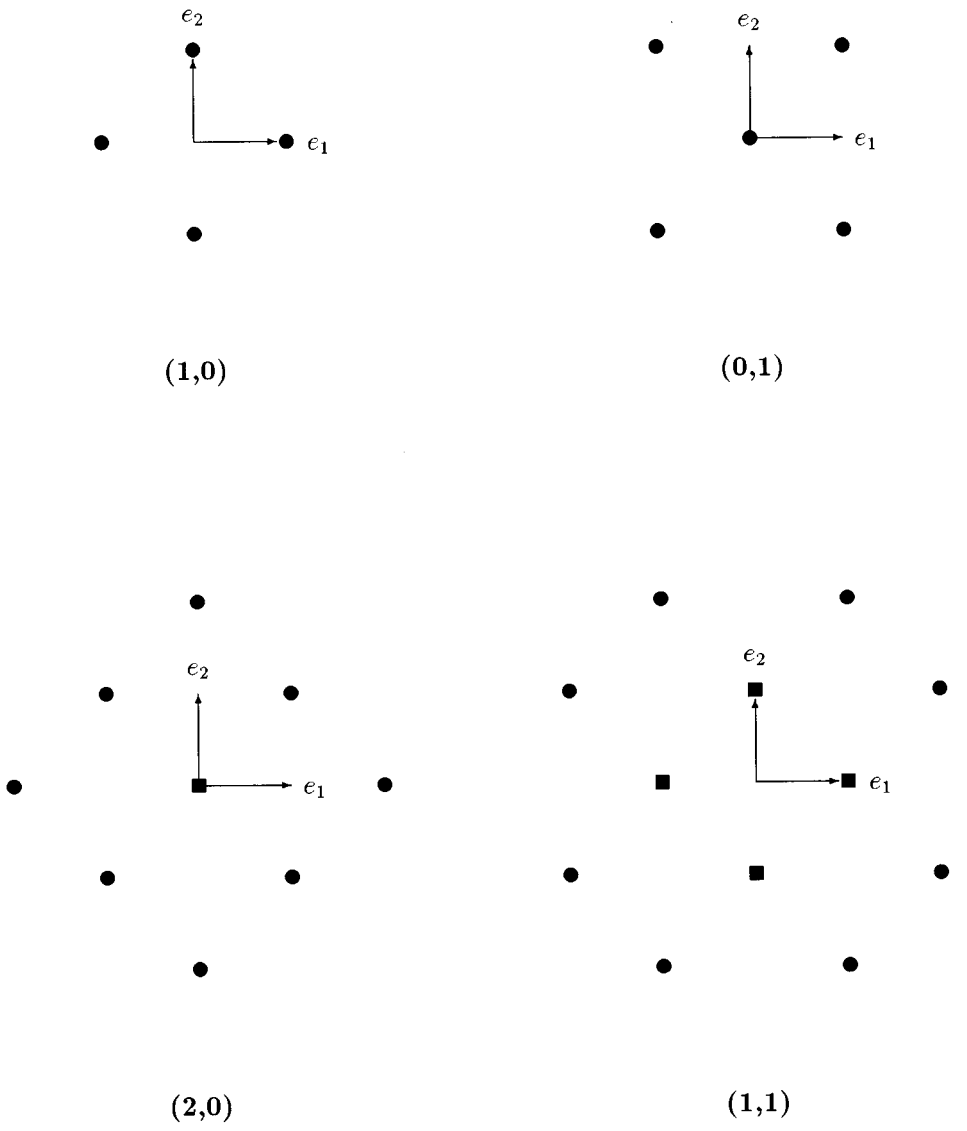

FIG. 4. Weight diagrams for some irreducible representations of Sp(4).

$$f_{(1,1),D}(x,w_1,w_2)=(1-xw_1)^2(1-xw_1^{-1})^2(1-xw_2)^2(1-xw_2^{-1})^2$$
$$\times(1-xw_1^2w_2)(1-xw_1^{-2}w_2)(1-xw_1^2w_2^{-1})(1-xw_1^{-2}w_2^{-1})$$
$$\times(1-xw_1w_2^2)(1-xw_1^{-1}w_2^2)(1-xw_1w_2^{-2})(1-xw_1^{-1}w_2^{-2}). \quad (129)$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=0=c_{0,2},$$

$$c_{3,0}=0=c_{0,3}, \quad c_{2,1}=0=c_{1,2}, \quad (130)$$

$$c_{4,0}=1=c_{0,4}, \quad c_{3,1}=2=c_{1,3}, \quad c_{2,2}=6.$$

B. Sp(6)

For an irreducible representation π_Λ of highest weight $\Lambda = a_1\lambda_1 + a_2\lambda_2 + a_3\lambda_3$, we have

$$\dim(\pi_\Lambda) = (1+a_1)(1+a_2)(1+a_3)\left(1+\frac{a_1+a_2}{2}\right)\left(1+\frac{a_2+a_3}{2}\right)\left(1+\frac{a_2+2a_3}{3}\right)$$
$$\times\left(1+\frac{a_1+a_2+a_3}{3}\right)\left(1+\frac{a_1+a_2+2a_3}{4}\right)\left(1+\frac{a_1+2a_2+2a_3}{5}\right), \tag{131}$$

$$\mathrm{ht}(\pi_\Lambda) = 5a_1 + 8a_2 + 9a_3. \tag{132}$$

The Weyl group of Sp(6) is $\mathbb{Z}_2\times\mathbb{Z}_2\times\mathbb{Z}_2\times S_3$, generated by the reflections $e_i \leftrightarrow -e_i$ and the permutations of the e_i $(i=1,2,3)$; it has order 48. Therefore,

$$M_\Lambda(z) = \frac{1}{48}\oint\frac{dw_1}{2\pi i}\oint\frac{dw_2}{2\pi i}\oint\frac{dw_3}{2\pi i}\frac{w_1^{-7}w_2^{-7}w_3^{-7}f_N(w_1,w_2,w_3)^2}{f_{\Lambda,D}(z,w_1,w_2,w_3)}, \tag{133}$$

$$F_\Lambda(z,\overline{z}) = \frac{1}{48}\oint\frac{dw_1}{2\pi i}\oint\frac{dw_2}{2\pi i}\oint\frac{dw_3}{2\pi i}\frac{w_1^{-7}w_2^{-7}w_3^{-7}f_N(w_1,w_2,w_3)^2}{f_{\Lambda,D}(z,w_1,w_2,w_3)f_{\Lambda,D}(\overline{z},w_1,w_2,w_3)}, \tag{134}$$

with

$$f_N(w_1,w_2,w_3) = -(1-w_1^2)(1-w_2^2)(1-w_3^2)(1-w_1w_2)(1-w_1w_3)$$
$$\times(1-w_2w_3)(w_1-w_2)(w_1-w_3)(w_2-w_3). \tag{135}$$

(The definition of $f_{\Lambda,D}$ follows below.)

1. First fundamental representation (1,0,0)

The highest weight is $\Lambda = \lambda_1 = e_1$, the dimension is 6, the height is 5 (so the representation is pseudo-real), the complete weight system is

$$\Phi_{(1,0,0)} = \{\pm e_1, \pm e_2, \pm e_3\} \tag{136}$$

(all weights have multiplicity 1; the signs are to be read independently), so

$$f_{(1,0,0),D}(x,w_1,w_2,w_3) = (1-xw_1)(1-xw_1^{-1})(1-xw_2)(1-xw_2^{-1})(1-xw_3)(1-xw_3^{-1}). \tag{137}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0} = 0 = c_{0,2},$$

$$c_{3,0} = 0 = c_{0,3}, \quad c_{2,1} = 0 = c_{1,2},$$

$$c_{4,0} = 0 = c_{0,4}, \quad c_{3,1} = 0 = c_{1,3}, \quad c_{2,2} = 1. \tag{138}$$

2. Second fundamental representation (0,1,0)

The highest weight is $\Lambda = \lambda_2 = e_1 + e_2$, the dimension is 14, the height is 8 (so the representation is real), the complete weight system is

$$\Phi_{(0,1,0)} = \{0\} \cup \{\pm e_1 \pm e_2, \pm e_1 \pm e_3, \pm e_2 \pm e_3\} \tag{139}$$

(0 has multiplicity 2 and all other weights have multiplicity 1; the signs are to be read independently), so

$$\begin{aligned} f_{(0,1,0),D}(x,w_1,w_2,w_3) &= (1-x)^2(1-xw_1w_2)(1-xw_1^{-1}w_2)(1-xw_1w_2^{-1})(1-xw_1^{-1}w_2^{-1}) \\ &\times(1-xw_1w_3)(1-xw_1^{-1}w_3)(1-xw_1w_3^{-1})(1-xw_1^{-1}w_3^{-1}) \\ &\times(1-xw_2w_3)(1-xw_2^{-1}w_3)(1-xw_2w_3^{-1})(1-xw_2^{-1}w_3^{-1}). \end{aligned} \tag{140}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=1=c_{0,2},$$

$$c_{3,0}=1=c_{0,3}, \quad c_{2,1}=1=c_{1,2}, \tag{141}$$

$$c_{4,0}=1=c_{0,4}, \quad c_{3,1}=1=c_{1,3}, \quad c_{2,2}=3.$$

3. Third fundamental representation (0,0,1)

The highest weight is $\Lambda=\lambda_3=e_1+e_2+e_3$, the dimension is 14, the height is 9 (so the representation is pseudo-real), the complete weight system is

$$\Phi_{(0,0,1)}=\{\pm e_1,\pm e_2,\pm e_3,\pm e_1\pm e_2\pm e_3\} \tag{142}$$

(all weights have multiplicity 1; the signs are to be read independently), so

$$\begin{aligned} f_{(0,0,1),D}(x,w_1,w_2,w_3) &= (1-xw_1)(1-xw_1^{-1})(1-xw_2)(1-xw_2^{-1})(1-xw_3)(1-xw_3^{-1}) \\ &\times(1-xw_1w_2w_3)(1-xw_1^{-1}w_2w_3)(1-xw_1w_2^{-1}w_3) \\ &\times(1-xw_1^{-1}w_2^{-1}w_3)(1-xw_1w_2w_3^{-1})(1-xw_1^{-1}w_2^{-1}w_3^{-1}). \end{aligned} \tag{143}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=0=c_{0,2},$$

$$c_{3,0}=0=c_{0,3}, \quad c_{2,1}=0=c_{1,2}, \tag{144}$$

$$c_{4,0}=1=c_{0,4}, \quad c_{3,1}=1=c_{1,3}, \quad c_{2,2}=1.$$

4. Codon representation (1,1,0)

The highest weight is $\Lambda=\lambda_1+\lambda_2=2e_1+e_2$, the dimension is 64, the height is 13 (so the representation is pseudo-real), the complete weight system is the union

$$\Phi_{(1,1,0)}=\Phi^{(1)}_{(1,1,0)}\cup\Phi^{(2)}_{(1,1,0)}\cup\Phi^{(4)}_{(1,1,0)}, \tag{145}$$

where

$$\Phi^{(1)}_{(1,1,0)}=\{\pm 2e_1\pm e_2,\pm 2e_2\pm e_1,\pm 2e_1\pm e_3,\pm 2e_3\pm e_1,\pm 2e_2\pm e_3,\pm 2e_3\pm e_2\} \tag{146}$$

(these weights have multiplicity 1; the signs are to be read independently),

$$\Phi^{(2)}_{(1,1,0)}=\{\pm e_1\pm e_2\pm e_3\} \tag{147}$$

(these weights have multiplicity 2; the signs are to be read independently), and

$$\Phi^{(4)}_{(1,1,0)}=\{\pm e_1,\pm e_2,\pm e_3\} \tag{148}$$

(these weights have multiplicity 4; the signs are to be read independently), so

$$\begin{aligned} f_{(1,1,0),D}(x,w_1,w_2,w_3)=&(1-xw_1)^4(1-xw_1^{-1})^4(1-xw_2)^4(1-xw_2^{-1})^4(1-xw_3)^4(1-xw_3^{-1})^4\\ &\times(1-xw_1w_2w_3)^2(1-xw_1^{-1}w_2w_3)^2(1-xw_1w_2^{-1}w_3)^2\\ &\times(1-xw_1^{-1}w_2^{-1}w_3)^2(1-xw_1w_2w_3^{-1})^2(1-xw_1^{-1}w_2^{-1}w_3^{-1})^2\\ &\times(1-xw_1^2w_2)(1-xw_1^{-2}w_2)(1-xw_1^2w_2^{-1})(1-xw_1^{-2}w_2^{-1})\\ &\times(1-xw_1w_2^2)(1-xw_1^{-1}w_2^2)(1-xw_1w_2^{-2})(1-xw_1^{-1}w_2^{-2})\\ &\times(1-xw_1^2w_3)(1-xw_1^{-2}w_3)(1-xw_1^2w_3^{-1})(1-xw_1^{-2}w_3^{-1})\\ &\times(1-xw_1w_3^2)(1-xw_1^{-1}w_3^2)(1-xw_1w_3^{-2})(1-xw_1^{-1}w_3^{-2})\\ &\times(1-xw_2^2w_3)(1-xw_2^{-2}w_3)(1-xw_2^2w_3^{-1})(1-xw_2^{-2}w_3^{-1})\\ &\times(1-xw_2w_3^2)(1-xw_2^{-1}w_3^2)(1-xw_2w_3^{-2})(1-xw_2^{-1}w_3^{-2}). \end{aligned} \tag{149}$$

The non-trivial coefficients $c_{p,q}$ up to fourth order are

$$c_{2,0}=0=c_{0,2},$$

$$c_{3,0}=0=c_{0,3},\quad c_{2,1}=0=c_{1,2},$$

$$c_{4,0}=3=c_{0,4},\quad c_{3,1}=6=c_{1,3},\quad c_{2,2}=15. \tag{150}$$

ACKNOWLEDGMENTS

The author would like to thank J. E. M. Hornos and Y. M. M. Hornos for introducing him to their work on the origin of the genetic code, which provided the motivation for the present investigation, and for generously sharing their insights into this fascinating subject. Thanks are also due to Renato Zornoff Táboas and Viktor Bojarczuk for discussions, and last but not least to all the people who developed the software package MAPLE, without which many of the calculations reported here would never have been performed. Finally, the financial support by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), Brazil, is gratefully acknowledged.

[1] J. E. M. Hornos and Y. M. M. Hornos, ''Algebraic Model for the Evolution of the Genetic Code,'' Phys. Rev. Lett. **71**, 4401–4404 (1993).
[2] J. Maddox, ''The Genetic Code by Numbers,'' Nature (London) **367**, 111 (1994).
[3] I. Stewart, ''Broken Symmetry in the Genetic Code?'' New Sci. **141**, 16 (1994).
[4] M. Forger, Y. M. M. Hornos, and J. E. M. Hornos, ''Global Aspects in the Algebraic Approach to the Genetic Code,'' Phys. Rev. E (to be published).
[5] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory* (Springer-Verlag, Berlin, 1988), Vol. II.
[6] D. H. Sattinger and O. L. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry and Mechanics* (Springer-Verlag, Berlin, 1986).

[7] M. V. Jaric and J. L. Birman, ''New Algorithms for the Molien Functions,'' J. Math. Phys. **18**, 1456–1458 (1977); ''Calculation of the Molien Generating Function for Invariants of Space Groups,'' **18**, 1459–1465 (1977).
[8] L. Michel, ''Invariants Polynomiaux des Groupes de Symétrie Moleculaire et Cristallographique,'' in ''Group Theoretical Methods in Physics,'' *Proceedings of the Fifth International Colloquium on Group Theoretical Methods in Physics*, edited by R. Sharp and B. Kolman, Montréal, Canada 1976 (Academic, New York, 1977).
[9] N. Wallach, *Harmonic Analysis on Homogeneous Spaces* (Marcel Dekker, New York, 1973).
[10] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer-Verlag, Berlin, 1972).
[11] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces* (Academic, New York, 1978).
[12] W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature and Cohomology* (Academic, New York, 1976), Vol. 3.
[13] R. Slansky: ''Group Theory for Unified Model Building,'' Phys. Rep. **79**, 1–128 (1981).

Generalization of the algebra of the matrix pseudodifferential operators

Jarosław Gawrylczyk[a)]
Institute for Theoretical Physics, University of Wroclaw, Pl.Maxa Borna 9, 50-205 Wroclaw, Poland

(Received 6 March 1997; accepted for publication 6 August 1997)

We introduce the new nonhomogeneous graded matrix algebra $\mathrm{Mat}_n(\mathfrak{R})$ over the ring $\mathfrak{R}$ of the pseudodifferential operators of the arbitrary orders. This algebra forms the generalization of the algebra of the standard matrix pseudodifferential operators. We show that the $\mathrm{Mat}_n(\mathfrak{R})$ algebra admits the nontrivial central extension. The case $n=2$ is considered in some detail. As the example we consider the $\mathrm{Mat}_2(\mathrm{W}_{1+\infty})$, $\mathrm{Mat}_2(\mathrm{Virasoro})$ and $\mathrm{Mat}_2(\mathrm{sl}(2,\mathbb{C}))$ algebras. © *1998 American Institute of Physics.* [S0022-2488(97)03111-3]

I. INTRODUCTION

It is well known that the algebras of the differential and pseudodifferential (psd) operators play the crucial role in the wide classes of the physical models, e.g., theory of the integrable systems,[1,2] conformal field theory,[3] matrix models,[4] two dimensional gravity,[5] and string theory.[6] Among other things, it was recognized that there exists the close relationship between the algebras of the psd operators and so-called W algebras[7–9] which are the higher-spin extensions of the Virasoro algebra. It was shown[10,11] that the Lie algebras of the differential operators admit the nontrivial central extensions which are isomorphic with the central extensions of the W algebras. The systematic way to obtain the suitable 2-cocycles (so-called ''logarithmic cocycles'') on these algebras was first considered in Ref. 12 and was further studied by several authors.[13–15]

One of the classes of the psd operators is the class of the matrix psd operators of the integer orders which has been investigated in the context of the examination of the multicomponent KP[16,17] and KdV[18] hierarchies as well as of the theory of the representations of the $\mathrm{W}_{1+\infty}$ algebra.[19–23] The generalizations of the matrix psd operators to the case of the complex orders were considered in Ref. 14.

The aim of this paper is the presentation of the another generalization of the algebra of the matrix psd operators to the case of the nonhomogeneous gradation. In our approach we consider the algebras which are generated by the matrices whose matrix elements are determined by the psd operators of the arbitrary orders. The essential point is that the orders of the each psd operator (which form the matrix elements of the single generator) can in general be different. In the case of the standard, homogeneous gradation the orders of such operators are the same. We show that the nonhomogeneous graded algebra of the $n\times n$-matrix differential operators admit the central extension and that this algebra is isomorphic with the central extended $n\times n$-matrix $\mathrm{W}_{1+\infty}$ algebra (which is different from the ones considered in Refs. 19–24). Next, we reduce the obtained in the explicit form 2×2-matrix $\mathrm{W}_{1+\infty}$ algebra to the $\mathrm{W}_{1+\infty}$ Virasoro subalgebra and the latter to the finite dimensional 2-index $\mathrm{sl}(2,\mathbb{R})$ algebra.

The paper is organized as follows. In Sec. II, we review some of the standard facts on the algebra of the pseudodifferential operators of the integer and the complex orders. In Sec. III, we introduce the nonhomogeneous graded algebra of the 2×2-matrix psd operators. Section IV is devoted to the study of the central extension of this algebra and their generalization to the $n\times n$ case. In Sec. V, we define the nonhomogeneous graded 2×2-matrix $\mathrm{W}_{1+\infty}$ algebra and their two subalgebras. Finally, we present some concluding remarks in Sec. VI.

[a)]Permanent address: Air Force Institute of Technology, Księcia Janusza, PO Box 60, 00-961 Warszawa, Poland.

 © 1998 American Institute of Physics

II. THE ALGEBRA OF THE PSEUDODIFFERENTIAL OPERATORS ON A CIRCLE

Let $\mathfrak{R}(\mathbb{C},\mathbf{k})$ denote the ring of the pseudodifferential operators of order $\mathbb{C}$ with the coefficients belonging to the algebra $C^\infty(\mathbb{S}^1,\mathbf{k})$ of smooth $\mathbf{k}$-valued functions on $\mathbb{S}^1$ ($\mathbf{k}=\mathbb{R}$ or $\mathbb{C}$):

$$\mathfrak{R}(\mathbb{C},\mathbf{k})=\left\{A(x,\partial)=\sum_{i=0}^{\infty} a_i(x)\partial_x^{\alpha-i}\,\middle|\, a_i(x)\in C^\infty(\mathbb{S}^1,\mathbf{k}),\alpha\in\mathbb{C}\right\}. \tag{2.1}$$

The multiplication law in $\mathfrak{R}(\mathbb{C},\mathbf{k})$ is determined by the Leibniz rule

$$A(x,\partial)\circ B(x,\partial)=\sum_{i,j=0}^{\infty}\sum_{k=0}^{\infty}\binom{\alpha-i}{k}a_i(x)b_j^{(k)}(x)\partial_x^{\alpha+\beta-i-j-k}, \tag{2.2}$$

where

$$\binom{\alpha}{k}\equiv\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}\quad\text{and}\quad b_j^{(k)}(x)\equiv\partial_x^k b_j(x).$$

The law (2.2) determines the Lie algebra structure on $\mathfrak{R}(\mathbb{C},\mathbf{k})$

$$[A,B]=A\circ B-B\circ A=\sum_{i,j=0}^{\infty}\sum_{k=0}^{\infty}\left[\binom{\alpha-i}{k}a_i(x)b_j^{(k)}(x)-\binom{\beta-j}{k}b_j(x)a_i^{(k)}(x)\right]\partial_x^{\alpha+\beta-i-j-k}. \tag{2.3}$$

Restricting the orders $\mathrm{ord}A(x,\partial)$ of the pseudodifferential operators on those satisfying the condition $\mathrm{ord}A(x,\partial)\in\mathbb{Z}$, we obtain the algebra $\mathfrak{R}(\mathbb{Z},\mathbf{k})=\mathfrak{R}(\mathbb{Z}_+,\mathbf{k})\oplus\mathfrak{R}(\mathbb{Z}_-,\mathbf{k})$ consisting the standard algebra of the differential and integral (Voltera's) operators.

It is well known that the algebra $\mathfrak{R}(\mathbb{Z}_+,\mathbf{k})$ allows the central extension given by the following 2-cocycle:[19]

$$c(a_i(x)\partial_x^i,b_i(x)\partial_x^j)=\frac{i!j!}{(i+j+1)!}\int_{S^1}a_i^{(j)}(x)b_j^{(i+1)}(x)dx,\quad (i,j=0,1,2,...), \tag{2.4}$$

where $a_i\partial_x^i$ and $b_j\partial_x^j$ are the purely differential operators. The relation (2.4) can be obtained by using, for example, the Khesin–Kravchenko construction.[12]

It is well known that the $\mathfrak{R}(\mathbb{Z}_+,\mathbf{k})$ algebra is isomorphic with the $W_{1+\infty}$ algebra.[19] Restricting the differential operators on the operators of the first order one can obtain the Virasoro algebra.

The algebra $\mathfrak{R}(\mathbb{C},\mathbf{k})$ can be generalized to the matrix case.[13,14] Let $\mathcal{G}$ be a matrix Lie algebra. Then the algebra $\mathfrak{R}(\mathbb{C},\mathcal{G})$ consists the matrix pseudodifferential operators of the form

$$\mathfrak{R}(\mathbb{C},\mathcal{G})=\left\{A(x,\partial)=\sum_{i=0}^{\infty} a_i(x)\partial_x^{\alpha-i}\,\middle|\, a_i(x)\in C^\infty(\mathbb{S}^1,\mathcal{G}),\alpha\in\mathbb{C}\right\}. \tag{2.5}$$

The multiplication law in $\mathfrak{R}(\mathbb{C},\mathcal{G})$ is the same as (2.2) but, in this case, the matrix product of the coefficients must be taking into account. If the condition $\mathrm{ord}A(x,\partial)\in\mathbb{Z}$ is fulfilled, we obtain the algebra $\mathfrak{R}(\mathbb{Z},\mathcal{G})=\mathfrak{R}(\mathbb{Z}_+,\mathcal{G})\oplus\mathfrak{R}(\mathbb{Z}_-,\mathcal{G})$ of the matrix differential and integral operators.

It has been shown[13] that the algebra $\mathfrak{R}(\mathbb{Z}_+,\mathcal{G})$ admits the nontrivial central extension given by the 2-cocycle

$$c(a_i(x)\partial_x^i,b_i(x)\partial_x^j)=\mathrm{tr}\left[\frac{i!j!}{(i+j+1)!}\int_{S^1}a_i^{(j)}(x)b_j^{(i+1)}(x)dx\right], \tag{2.6}$$

where the symbol tr denotes the standard trace of a matrix.

The particular cases of the $\mathfrak{R}(\mathbb{Z}_+,\mathcal{G})$ algebras are, for example, the colored W_∞^p algebra[20] and non-abelian current version of $W_{1+\infty}$.[24]

III. THE NONHOMOGENEOUS GRADED $\mathfrak{R}(\mathbb{C},\mathscr{G})$ ALGEBRA

Let us consider the algebra $\mathfrak{R}(\mathbb{C},\mathscr{G})$ (2.5). Let us assume that, for simplicity, $\mathscr{G}$ is some space of the 2×2 matrices. In this case, the generators of the $\mathfrak{R}(\mathbb{C},\mathscr{G})$ algebra are the matrix pseudodifferential operators of the form

$$A(x,\partial_x)=\sum_{i=0}^{\infty}\begin{pmatrix} a_i^{11}(x) & a_i^{12}(x)\\ a_i^{21}(x) & a_i^{22}(x)\end{pmatrix}\begin{pmatrix}\partial^{\alpha-1} & 0\\ 0 & \partial^{\alpha-1}\end{pmatrix}=\sum_{i=0}^{\infty}\begin{pmatrix} a_i^{11}(x)\partial_x^{\partial-1} & a_i^{12}(x)\partial_x^{\alpha-1}\\ a_i^{21}(x)\partial_x^{\alpha-i} & a_i^{22}(x)\partial_x^{\alpha-i}\end{pmatrix}. \quad (3.1)$$

It easy to see that the orders of the matrix elements of the operator (3.1) are the same. The operators which have this property, we shall call the homogeneous graded operators. Now, we would like to extend the definition of the admissible matrix pseudodifferential operators by assumption that the matrix elements of such operators can be chosen to be the psd operators of the arbitrary orders. In this way, we obtain what will be referred to as the nonhomogeneous graded algebra of the matrix pseudodifferential operators $\mathrm{Mat}_2(\mathfrak{R})$. Indeed, let us consider the operators of the form

$$\mathbb{A}(x,\partial_x)=\begin{pmatrix}\sum_{i=0}^{\infty} a_i^{11}(x)\partial_x^{\alpha_{11}-i} & \sum_{i=0}^{\infty} a_i^{12}(x)\partial_x^{\alpha_{12}-i}\\ \sum_{i=0}^{\infty} a_i^{21}(x)\partial_x^{\alpha_{21}-i} & \sum_{i=0}^{\infty} a_i^{22}(x)\partial_x^{\alpha_{22}-i}\end{pmatrix}\equiv\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}, \quad (3.2)$$

where, in general, $\alpha_{11}\neq\alpha_{12}\neq\alpha_{21}\neq\alpha_{22}$, and the coefficients a's belong to the space

$$\text{case 1: } a_i^{mn}(x)\in C^{\infty}(\mathbb{S}^1,\mathbf{k}), \text{ or} \quad (3.3)$$

$$\text{case 2: } a_i^{mn}(x)\in C^{\infty}(\mathbb{S}^1,\mathscr{G}), \ (m,n=1,2). \quad (3.4)$$

Definition: The operators of the form (3.2) generates the matrix algebra:
case 1: $\mathrm{Mat}_2(\mathfrak{R}(\mathbb{C},\mathbf{k}))$, if the coefficients satisfy the condition (3.3), or
case 2: $\mathrm{Mat}_2(\mathfrak{R}(\mathbb{C},\mathscr{G}))$, if the coefficients satisfy the condition (3.4).

Using the multiplication law (2.2) one can check that the algebra $G_{\cdot,*}\equiv\mathrm{Mat}_2(\mathfrak{R}(\cdot,*))$, ($*\equiv\mathbf{k}$ or $\mathscr{G}$ and $\cdot\equiv\mathbb{C}$ or $\mathbb{Z}$) has the form

$$[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]=\mathbb{C}(x,\partial_x), \quad (3.5)$$

where $\mathbb{C}(x,\partial_x)\equiv\begin{pmatrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{pmatrix}$, and

$$\begin{aligned} c_{11}&\equiv K(\alpha_{11},\beta_{11})-L(\beta_{11},\alpha_{11})+K(\alpha_{12},\beta_{21})-L(\beta_{12},\alpha_{21}),\\ c_{12}&\equiv K(\alpha_{11},\beta_{12})-L(\beta_{12},\alpha_{22})+K(\alpha_{12},\beta_{22})-L(\beta_{11},\alpha_{12}),\\ c_{21}&\equiv K(\alpha_{21},\beta_{11})-L(\beta_{22},\alpha_{21})+K(\alpha_{22},\beta_{21})-L(\beta_{21},\alpha_{11}),\\ c_{22}&\equiv K(\alpha_{22},\beta_{22})-L(\beta_{22},\alpha_{22})+K(\alpha_{21},\beta_{12})-L(\beta_{21},\alpha_{12}),\end{aligned} \quad (3.6)$$

where

$$K(\alpha,\beta)=\sum_{i,j=0}^{\infty}\sum_{k=0}^{\alpha-j}\binom{\alpha-j}{k}a_i(x)b_j^{(k)}(x)\partial_x^{\alpha+\beta-i-j-k},$$

$$L(\beta,\alpha)=\sum_{i,j=0}^{\infty}\sum_{k=0}^{\beta-j}\binom{\beta-j}{k}b_j(x)a_i^{(k)}(x)\partial_x^{\alpha+\beta-i-j-k}.$$

Let us consider the special cases of the algebra $G_{\cdot,*}$.

(i) If $A_{12}=A_{21}=A_{22}=0$ and $A_{11}\neq 0$ as well as $\alpha_{11}\in\mathbb{C}$ and $\beta\in\mathbb{C}$ then $G_{\mathbb{C},\mathcal{G}}\equiv\mathfrak{R}(\mathbb{C},\mathcal{G})$ and $G_{\mathbb{C},\mathbf{k}}\equiv\mathfrak{R}(\mathbb{C},\mathbf{k})$.

(ii) If $A_{11}\neq 0$, $A_{12}\neq 0$, $A_{21}\neq 0$ and $A_{22}\neq 0$ as well as $\alpha_{11}=\alpha_{12}=\alpha_{21}=\alpha_{22}\in\mathbb{C}$ and $\beta_{11}=\beta_{12}=\beta_{21}=\beta_{22}\in\mathbb{C}$ then $G_{\mathbb{C},\mathbf{k}}\equiv\mathfrak{R}(\mathbb{C},\mathcal{G})$. (The homogeneous graded algebra.)

(iii) If $A_{11}=A_{21}=A_{22}=0$ and $A_{11}\neq 0$ as well as $\alpha_{11}\in\mathbb{Z}$ and $\beta_{11}\in\mathbb{Z}$ then $G_{\mathbb{Z},\mathcal{G}}\equiv\mathfrak{R}(\mathbb{Z},\mathcal{G})$ and $G_{\mathbb{Z},\mathbf{k}}\equiv\mathfrak{R}(\mathbb{Z},\mathbf{k})$.

(iv) If $A_{11}\neq 0$, $A_{12}\neq 0$, $A_{21}\neq 0$ and $A_{22}\neq 0$ as well as $\alpha_{11}=\alpha_{12}=\alpha_{21}=\alpha_{22}\in\mathbb{Z}$ and $\beta_{11}=\beta_{12}=\beta_{21}=\beta_{22}\in\mathbb{Z}$ then $G_{\mathbb{Z},\mathbf{k}}\equiv\mathfrak{R}(\mathbb{Z},\mathcal{G})$. (The homogeneous graded algebra.)

For the other values of the orders of the operators which form the matrix elements of (3.2), we obtain the new non-homogeneous graded matrix algebras.

IV. THE CENTRAL EXTENSION OF THE $\mathrm{Mat}_n(\mathfrak{R}(\mathbb{Z},\mathbf{k}))$ ALGEBRA

Using the Kravchenko–Khesin construction[12,13] we show that the $n\times n$ matrix algebra $G_{\mathbb{Z},\mathbf{k}}\equiv\mathrm{Mat}_n(\mathfrak{R}(\mathbb{Z},\mathbf{k}))$ over the ring of the pseudodifferential operators of the integer orders admits the nontrivial central extension.

First, let us consider the 2×2 matrix $G_{\mathbb{Z},\mathbf{k}}$ algebra generated by the operators of the integer form

$$\mathbb{A}(x,\partial_x)=\begin{pmatrix}\sum_{i=-\infty}^{n_{11}} a_i^{11}(x)\partial_x^i & \sum_{i=-\infty}^{n_{12}} a_i^{12}(x)\partial_x^i\\ \sum_{i=-\infty}^{n_{21}} a_i^{21}(x)\partial_x^i & \sum_{i=-\infty}^{n_{22}} a_i^{22}(x)\partial_x^i\end{pmatrix}\equiv\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix},\tag{4.1}$$

where $a_i^{mn}(x)\in\mathbb{C}^\infty(\mathbb{S}^1,\mathbf{k})$ $(m,n=1,2)$.

The algebra of the symbols of the operators (4.1) is defined by the elements

$$\mathbb{A}(x,\xi)=\begin{pmatrix}\sum_{i=-\infty}^{n_{11}} a_i^{11}(x)\xi^i & \sum_{i=-\infty}^{n_{12}} a_i^{12}(x)\xi^i\\ \sum_{i=-\infty}^{n_{21}} a_i^{21}(x)\xi^i & \sum_{i=-\infty}^{n_{22}} a_i^{22}(x)\xi^i\end{pmatrix},\tag{4.2}$$

and the multiplication law

$$\left(\sum_{i=-\infty}^{n} a_i(x)\xi^i\right)\circ\left(\sum_{j=-\infty}^{m} b_j(x)\xi^j\right)=\sum_{i=-\infty}^{n}\sum_{j=-\infty}^{m}\sum_{k\geqslant 0}\frac{1}{k!}[\partial_\xi^k(a_i(x)\xi^i)][\partial_x^k(b_j(x)\xi^j)].\tag{4.3}$$

Let us introduce the operator

$$\mathbb{R}\mathrm{es}:\begin{pmatrix}\mathfrak{R}(\mathbb{Z},\mathbf{k}) & \mathfrak{R}(\mathbb{Z},\mathbf{k})\\ \mathfrak{R}(\mathbb{Z},\mathbf{k}) & \mathfrak{R}(\mathbb{Z},\mathbf{k})\end{pmatrix}\to\begin{pmatrix}C^\infty(\mathbb{S},\mathbf{k}) & C^\infty(\mathbb{S},\mathbf{k})\\ C^\infty(\mathbb{S},\mathbf{k}) & C^\infty(\mathbb{S},\mathbf{k})\end{pmatrix},\tag{4.4}$$

defined by the relation

$$\mathbb{R}\mathrm{es}\ \mathbb{A}(x,\partial_x)=\begin{pmatrix}\mathrm{res}\ A_{11} & \mathrm{res}\ A_{12}\\ \mathrm{res}\ A_{21} & \mathrm{res}\ A_{22}\end{pmatrix}=\begin{pmatrix}a_{-1}^{11}(x) & a_{-1}^{12}(x)\\ a_{-1}^{21}(x) & a_{-1}^{22}(x)\end{pmatrix}.\tag{4.5}$$

Lemma 1: Let $\mathbb{A}(x,\partial_x)$ and $\mathbb{B}(x,\partial_x)\in G_{\mathbb{Z},\mathbf{k}}$ and $A_{11}=A_{22}$, $A_{12}=A_{21}$, $B_{11}=B_{22}$, and $B_{12}=B_{21}$ then the following relation:

$$\int_{\mathbb{S}^1}\mathbb{R}\mathrm{es}[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]=0\tag{4.6}$$

holds.

Proof: From (4.1) and (4.5) it follows that:

$$\int_{S^1}\mathbb{R}\mathrm{es}[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]$$

$$=\begin{pmatrix} \int \mathrm{res}[A_{11},B_{11}]+\int \mathrm{res}[A_{12},B_{12}] & \int \mathrm{res}[A_{11},B_{12}]+\int \mathrm{res}[A_{12},B_{11}] \\ \int \mathrm{res}[A_{12},B_{11}]+\int \mathrm{res}[A_{11},B_{12}] & \int \mathrm{res}[A_{12},B_{12}]+\int \mathrm{res}[A_{11},B_{11}] \end{pmatrix}=0,$$

since $\int_{s^1}\mathrm{res}[A_{ij},B_{kl}]=0$ $(i,j,k,l,=1,2)$. □

Let us consider the formal expression

$$\mathbb{L}\mathrm{og}\ \partial_x:=\begin{pmatrix} \log\partial_x & 0 \\ 0 & \log\partial_x \end{pmatrix}. \tag{4.7}$$

One can show that the formal commutator $[\mathbb{L}\mathrm{og}\ \partial_x,\mathbb{A}(x,\partial_x)]=\mathbb{L}\mathrm{og}\ \partial_x\circ\mathbb{A}(x,\partial_x)-\mathbb{A}(x,\partial_x)\circ\mathbb{L}\mathrm{og}\ \partial_x$ belongs to $G_{Z,\mathbf{k}}$. Indeed, using the algebra of the symbols, it is easy to see that:

$$[\mathbb{L}\mathrm{og}\ \partial_x,\mathbb{A}(x,\partial_x)]=\begin{pmatrix} [\log\partial_x,A_{11}] & [\log\partial_x,A_{12}] \\ [\log\partial_x,A_{21}] & [\log\partial_x,A_{22}] \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{k\geqslant 1}\frac{(-1)^{k-1}}{k}(\partial_x^k A_{11})\partial_x^{-k} & \sum_{k\geqslant 1}\frac{(-1)^{k-1}}{k}(\partial_x^k A_{12})\partial_x^{-k} \\ \sum_{k\geqslant 1}\frac{(-1)^{k-1}}{k}(\partial_x^k A_{21})\partial_x^{-k} & \sum_{k\geqslant 1}\frac{(-1)^{k-1}}{k}(\partial_x^k A_{22})\partial_x^{-k} \end{pmatrix}. \tag{4.8}$$

Lemma 2: The following relation

$$[\mathbb{L}\mathrm{og}\ \partial_x,\mathbb{A}(x,\partial_x)\circ\mathbb{B}(x,\partial_x)]$$

$$=[\mathbb{L}\mathrm{og}\ \partial_x,\mathbb{A}(x,\partial_x)]\circ\mathbb{B}(x,\partial_x)+\mathbb{A}(x,\partial_x)\circ[\mathbb{L}\mathrm{og}\ \partial_x,\mathbb{B}(x,\partial_x)] \tag{4.9}$$

holds.

Proof: It follows immediately from the formulas (4.1)–(4.3). □

Now, we can formulate the main result of this section.

Theorem: The central extended algebra

$$\hat{G}_{Z,\mathbf{k}}=G_{Z,\mathbf{k}}\oplus\begin{pmatrix} \mathbb{C} & \mathbb{C} \\ \mathbb{C} & \mathbb{C} \end{pmatrix}$$

has the form

$$[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]_c=[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]+c(\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)), \tag{4.10}$$

where $\mathbb{A}(x,\partial_x)$ and $\mathbb{B}(x,\partial_x)\in G_{Z,\mathbf{k}}$ and $c(\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x))$ is defined by the relation

$$c(\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x))=\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}; \tag{4.11}$$

where

$$\begin{aligned} c_{11}=c_{22}&\equiv[\mathscr{P}(A_{11},B_{11})+\mathscr{P}(A_{12},B_{12})]\Delta_1, \\ c_{12}=c_{21}&\equiv[\mathscr{P}(A_{11},B_{12})+\mathscr{P}(A_{12},B_{11})]\Delta_2, \end{aligned} \tag{4.12}$$

where

$$\mathscr{P}(A_{ij},B_{mn})\equiv\int_{S^1}\mathrm{res}([A_{ij},\log\,\partial_x]\circ B_{mn})=\int_{S^1}\mathrm{res}\left(\sum_{k\geqslant 1}\frac{(-1)^k}{k}(\partial_x^k A_{ij})\partial_x^{-k}B_{mn}\right),\quad(4.13)$$

$(i,j,m,n\in\{1,2\})$, and

$$\Delta_1\equiv\delta_{A_{11},A_{22}}\delta_{A_{12},A_{21}}\delta_{B_{11},B_{22}}\delta_{B_{12},B_{21}},$$

$$\Delta_2\equiv\delta_{A_{11},A_{12}}\delta_{A_{11},A_{21}}\delta_{A_{11},A_{22}}\delta_{B_{11},B_{12}}\delta_{B_{11},B_{21}}\delta_{B_{11},B_{22}},\quad(4.14)$$

$$\delta_{A_{ij},A_{mn}}=\begin{cases}1 & \text{if ord } A_{ij}=\text{ord } A_{mn}\\ 0 & \text{in the other}\end{cases}$$

and B_{mn} $(m,n=1,2)$ are the matrix elements of the operator $\mathbb{B}(x,\partial_x)$.

Proof: By direct calculations [using the Jacobi identity and the relations (4.6), (4.8), and (4.9)] one can check that the function (4.11) satisfy the 2-cocycle condition

$$\sum_{\text{cyclic}}c(\mathbb{A},[\mathbb{B},\mathbb{C}])=\int_{S^1}\mathbb{R}\mathrm{es}([\mathbb{A},\mathbb{L}\mathrm{og}\,\partial_x]\circ[\mathbb{B},\mathbb{C}]+[\mathbb{B},\mathbb{L}\mathrm{og}\,\partial_x]\circ[\mathbb{C},\mathbb{A}]+[\mathbb{C},\mathbb{L}\mathrm{og}\,\partial_x]\circ[\mathbb{A},\mathbb{B}])=0.$$

□

The restriction of the 2-cocycle (4.11) on the purely differential operators $\mathbb{A}_+(x,\partial_x)$ and $\mathbb{B}_+(x,\partial_x)$ (which satisfy the conditions ord $\mathbb{A}_+\in\mathbb{Z}_+$ and ord $\mathbb{B}_+\in\mathbb{Z}_+$) determines the central extension $\hat{G}_{Z_+,\mathbf{k}}$ of the $G_{Z_+,\mathbf{k}}$ algebra

$$[\mathbb{A}_+(x,\partial_x),\mathbb{B}_+(x,\partial_x)]_c=[\mathbb{A}_+(x,\partial_x),\mathbb{B}_+(x,\partial_x)]+c(\mathbb{A}_+(x,\partial_x),\mathbb{B}_+(x,\partial_x)).\quad(4.15)$$

Using the relations (4.1) and (4.10)–(4.14) one can compute the explicit form of the 2-cocycle $c(\mathbb{A}_+,\mathbb{B}_+)$

$$c(\mathbb{A}_+(x,\partial_x),\mathbb{B}_+(x,\partial_x))=\begin{pmatrix}\tilde{c}_{11} & \tilde{c}_{12}\\ \tilde{c}_{21} & \tilde{c}_{22}\end{pmatrix},\quad(4.16)$$

where

$$\tilde{c}_{11}=\tilde{c}_{22}\equiv[\tilde{\mathscr{P}}(n_{11},m_{11})+\tilde{\mathscr{P}}(n_{12},m_{12})]\tilde{\Delta}_1,$$
$$\tilde{c}_{12}=\tilde{c}_{21}\equiv[\tilde{\mathscr{P}}(n_{11},m_{12})+\tilde{\mathscr{P}}(n_{12},m_{11})]\tilde{\Delta}_2,\quad(4.17)$$

where

$$\tilde{\mathscr{P}}(n,m)\equiv\sum_{i=0}^{n}\sum_{j=0}^{m}\frac{i!j!}{(i+j+1)!}\int_{S^1}[\partial_x^j a_i(x)][\partial_x^{i+1}b_j(x)],\quad(4.18)$$

and

$$\tilde{\Delta}_1\equiv\delta_{n_{11},n_{22}}\delta_{n_{12},n_{21}}\delta_{m_{11},m_{22}}\delta_{m_{12},m_{21}},$$
$$\tilde{\Delta}_2\equiv\delta_{n_{11},n_{12}}\delta_{n_{11},n_{21}}\delta_{n_{11},n_{22}}\delta_{m_{11},m_{12}}\delta_{m_{11},m_{21}}\delta_{m_{11},m_{22}};\quad(4.19)$$

n_{ij} and m_{kl} $(i,j,k,l\in\{1,2\})$ are the orders of the elements of the operators $\mathbb{A}_+(x,\partial_x)$ and $\mathbb{B}_+(x,\partial_x)$.

The relations (4.16)–(4.19) describe the generalization of the Kac–Peterson 2-cocycles.[10] In particular, putting $n_{12}=n_{21}=n_{22}=m_{12}=m_{21}=m_{22}=0$ into these relations, we obtain the standard 2-cocycle on $\mathfrak{R}(\mathbb{Z}_+,\mathbf{k})$ (2.4). In the case $n_{11}=n_{12}=n_{21}=n_{22}\neq 0$ and $m_{11}=m_{12}=m_{21}=m_{22}\neq 0$, we obtain the 2-cocycle on $\mathfrak{R}(\mathbb{Z}_+,\mathscr{G})$ (2.6).

The above results one can generalize to the case of the $n\times n$ matrices. It follows from the Jacobi identity that the $n\times n$ matrix generalization of the $\hat{G}_{Z,\mathbf{k}}$ algebra (4.10) has the form:

$$[\mathbb{A}(x,\partial_x),\mathbb{B}(x,\partial_x)]_c=\begin{pmatrix} a_1 & a_2 & \cdots & a_2 \\ a_2 & a_1 & & \vdots \\ \vdots & & \ddots & \vdots \\ a_2 & \cdots & \cdots & a_1 \end{pmatrix}+\begin{pmatrix} c_1 & c_2 & \cdots & c_2 \\ c_2 & c_1 & & \vdots \\ \vdots & & \ddots & \vdots \\ c_2 & \cdots & \cdots & c_1 \end{pmatrix}, \tag{4.20}$$

where

$$\mathbb{A}(x,\partial_x)\equiv\begin{pmatrix} A_1 & A_2 & \cdots & A_2 \\ A_2 & A_1 & & \vdots \\ \vdots & & \ddots & \vdots \\ A_2 & \cdots & \cdots & A_1 \end{pmatrix}, \tag{4.21}$$

$$A_j=\sum_{i=0}^{\infty} a_i^j(x)\partial_x^{\alpha_j-i}, \quad (j=1,2),$$

and

$$\begin{aligned} a_1&=[A_1,B_1]+(n-1)[A_2,B_2], \\ a_2&=[A_1,B_2]+[A_2,B_1]+(n-2)[A_2,B_2], \end{aligned} \tag{4.22}$$

and

$$\begin{aligned} c_1&\equiv\mathscr{P}(A_1,B_1)+(n-1)\mathscr{P}(A_2,B_2), \\ c_2&\equiv\mathscr{P}(A_1,B_2)+\mathscr{P}(A_2,B_1)+(n-2)\mathscr{P}(A_2,B_2). \end{aligned} \tag{4.23}$$

The symbol $\mathscr{P}(A_i,B_j)$ is defined by the relation (4.13) and n denotes the dimension of the matrices $\mathbb{A}$ and $\mathbb{B}$. We note that the commutator of the $n\times n$ matrices other than (4.21) cannot be extended by some 2-cocycles because their elements are not the commutators of the operators A_j and does not satisfy the assumptions of the lemmas 1 and 2. Therefore the $n\times n$ matrix central extended $\hat{G}_{Z,\mathbf{k}}$ algebra is generated by only two classes of the pseudodifferential operators of the different orders. If orders of these operators are the same then the relation (4.20) describe the standard algebra of the $n\times n$ matrix pseudodifferential operators.

V. THE EXAMPLE OF THE $\mathrm{Mat}_2(\mathfrak{R})$-TYPE ALGEBRAS

Example 1: The $\mathrm{Mat}_2(W_{1+\infty})$ algebra.

Let us consider the following differential operators on $\mathbb{S}^1$, which form the basis of the standard $W_{1+\infty}$ algebra[13]

$$V_m^s=-B(s)\sum_{k=1}^{s}\alpha_k^s\binom{m+s-1}{k-1}z^{m+s-k}D^{s-k}, \tag{5.1}$$

where $z=e^{ix}$ $(x\in S^1)$, $D\equiv\partial_z$ and

$$B(s)=\frac{2^{s-3}(s-1)!}{(2s-3)!!} \quad \text{and} \quad \alpha_k^s=\frac{(2s-k-1)!}{[(s-k)!]}. \tag{5.2}$$

From the relations (4.15)–(4.19) and (5.1) it follows that the nonhomogeneous graded matrix $W_{1+\infty}$ algebra has the form

$$\left[\begin{pmatrix} V^{s_1}_{m_1} & V^{s_2}_{m_2} \\ V^{s_3}_{m_3} & V^{s_4}_{m_4} \end{pmatrix}, \begin{pmatrix} V^{r_1}_{n_1} & V^{r_2}_{n_2} \\ V^{r_3}_{n_3} & V^{r_4}_{n_4} \end{pmatrix}\right] = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} + \begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix}, \tag{5.3}$$

where

$$\begin{aligned} A_1 &= V^{s_1}_{m_1} * V^{r_1}_{n_1} - V^{r_1}_{n_1} * V^{s_1}_{m_1} + V^{s_2}_{m_2} * V^{r_3}_{n_3} - V^{r_2}_{n_2} * V^{s_3}_{m_3}, \\ A_2 &= V^{s_1}_{m_1} * V^{r_2}_{n_2} - V^{r_2}_{n_2} * V^{s_4}_{m_4} + V^{s_2}_{m_2} * V^{r_4}_{n_4} - V^{r_1}_{n_1} * V^{s_2}_{m_2}, \\ A_3 &= V^{s_3}_{m_3} * V^{r_1}_{n_1} - V^{r_4}_{n_4} * V^{s_3}_{m_3} + V^{s_4}_{m_4} * V^{r_3}_{n_3} - V^{r_3}_{n_3} * V^{s_1}_{m_1}, \\ A_4 &= V^{s_4}_{m_4} * V^{r_4}_{n_4} - V^{r_4}_{n_4} * V^{s_4}_{m_4} + V^{s_3}_{m_3} * V^{r_2}_{n_2} - V^{r_3}_{n_3} * V^{s_2}_{m_2}, \end{aligned} \tag{5.4}$$

and the symbol $*$ denotes the lone star product defined by the relation[25]

$$V^s_m * V^r_n = \frac{1}{2} \sum_{i \geqslant 0} g^{sr}_i(m,n) V^{s+r-i}_{m+n}, \tag{5.5}$$

where

$$g^{sr}_i(m,n) = \frac{\varphi^{sr}_i}{2(i+1)!} N^{sr}_i(m,n), \tag{5.6}$$

$$\varphi^{sr}_i = \sum_{k=0} \frac{(\frac{1}{2})_k (\frac{1}{2})_k (-(i+1)/2)_k (-\frac{1}{2})_k}{k!(-s+\frac{3}{2})_k (-r+\frac{3}{2})_k (s+r-i-\frac{3}{2})_k}, \tag{5.7}$$

and

$$N^{sr}_i(m,n) = \sum_{k=0}^{i+1} (-1)^k \binom{i+1}{k} (2s-i-2)_k [2r-k-2]_{i+1-k} [s-1+m]_{i+1-k} [r-1+n]_k, \tag{5.8}$$

where

$$(a)_k = a(a+1)\cdots(a+k-1); \quad [a]_k = a(a-1)\cdots(a-k+1). \tag{5.9}$$

By using the lone star product (5.5) the commutation relations of $W_{1+\infty}$ algebra can be defined as $[V^s_m, V^r_n] = V^s_m * V^r_n - V^r_n * V^s_m$. In the relation (5.3) the matrix $\left(\begin{smallmatrix} C_1 & C_2 \\ C_3 & C_4 \end{smallmatrix}\right)$ denotes the central terms in which the elements C's have the form

$$\begin{aligned} C_1 &= [\beta(s_1,m_1)\delta_{s_1,r_1}\delta_{m_1+n_1,0} + \beta(s_2,m_2)\delta_{s_2,r_3}\delta_{m_2+n_3,0}]\Delta_1, \\ C_2 &= [\beta(s_1,m_1)\delta_{s_1,r_2}\delta_{m_1+n_2,0} + \beta(s_2,m_2)\delta_{s_2,r_4}\delta_{m_2+n_4,0}]\Delta_2, \\ C_3 &= [\beta(s_3,m_3)\delta_{s_3,r_1}\delta_{m_3+n_1,0} + \beta(s_4,m_4)\delta_{s_4,r_3}\delta_{m_4+n_3,0}]\Delta_2, \\ C_4 &= [\beta(s_3,m_3)\delta_{s_3,r_2}\delta_{m_3+n_2,0} + \beta(s_4,m_4)\delta_{s_4,r_4}\delta_{m_4+n_4,0}]\Delta_1, \end{aligned} \tag{5.10}$$

where

$$\beta(s,m) = -\frac{B(s)^2}{2s-1} \frac{(m+s-1)!}{(m-s)!}, \tag{5.11}$$

and

$$\Delta_1=\delta_{s_1,s_4}\delta_{s_2,s_3}\delta_{r_1,r_4}\delta_{r_2,r_3}\delta_{m_1,m_4}\delta_{m_2,m_3}\delta_{n_1,n_4}\delta_{n_2,n_3},$$
$$\Delta_2=\delta_{s_1,s_2}\delta_{s_1,s_3}\delta_{s_1,s_4}\delta_{r_1,r_2}\delta_{r_1,r_3}\delta_{r_1,r_4}\delta_{m_1,m_2}\delta_{m_1,m_3}\delta_{m_1,m_4}\delta_{n_1,n_2}\delta_{n_1,n_2}\delta_{n_1,n_4}. \tag{5.12}$$

It is known that the generator V_m^s of the $W_{1+\infty}$ algebra corresponds to the mth Fourier mode of a conformal spin s field ($s\geqslant 1$). From the above fact and from the relations (5.3)–(5.12) it follows that the central extended $\mathrm{Mat}_2(W_{1+\infty})$ algebra (5.3) describe the algebra of the Fourier modes of the systems of conformal fields of the following conformal dimensions (spins) Δ:

$$\begin{pmatrix}\Delta_{s_1} & \Delta_{s_2}\\ \Delta_{s_2} & \Delta_{s_1}\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}\Delta_{r_1} & \Delta_{r_2}\\ \Delta_{r_2} & \Delta_{r_1}\end{pmatrix},$$

where $\Delta_{s_i}=s_i$ and $\Delta_{r_i}=r_i$ (s_i and $r_i=1,2,3,\ldots$; $i=1,2$). We note that the constraints $s_1=s_4$, $s_2=s_3$, $r_1=r_4$ and $r_2=r_3$ arise from the 2-cocycle conditions (5.10).

Example 2: The Mat_2(Virasoro) algebra.

It is known that the Virasoro algebra[7] is the subalgebra of the standard $W_{1+\infty}$ algebra. It follows from the relation (5.3) that the closed subalgebra of $\mathrm{Mat}_2(W_{1+\infty})$ corresponding with the matrix Virasoro algebra Mat_2(Virasoro) has the form

$$\left[\begin{pmatrix}V_{m_1} & V_{m_2}\\ V_{m_2} & V_{m_1}\end{pmatrix},\begin{pmatrix}V_{n_1} & V_{n_2}\\ V_{n_2} & V_{n_1}\end{pmatrix}\right]=\begin{pmatrix}A_1 & A_2\\ A_2 & A_1\end{pmatrix}+\begin{pmatrix}C_1 & C_2\\ C_2 & C_1\end{pmatrix}, \tag{5.13}$$

where

$$A_1=(n_1-m_1)V_{m_1+n_1}+(n_2-m_2)V_{m_2+n_2},$$
$$A_2=(n_2-m_1)V_{m_1+n_2}+(n_1-m_2)V_{m_2+n_1}, \tag{5.14}$$

and

$$C_1=\alpha_1\beta(m_1)\,\delta_{m_1+n_1,0}+\alpha_2\beta(m_2)\,\delta_{m_2+n_2,0},$$
$$C_2=[\,\alpha_1\beta(m_1)\,\delta_{m_1+n_2,0}+\alpha_2\beta(m_2)\,\delta_{m_2+n_1,0}]\,\delta_{m_1,m_2}\delta_{n_1,n_2}, \tag{5.15}$$

where $\beta(m)=m^3-m$ as well as α_1 and α_2 are the arbitrary constants.

It is easy to see, that the algebra (5.13) admits the representation described in the terms of the following differential operator:

$$\begin{pmatrix}V_{m_1} & V_{m_2}\\ V_{m_2} & V_{m_1}\end{pmatrix}\equiv\begin{pmatrix}z^{m_1+1}D & z^{m_2+1}D\\ z^{m_2+1}D & z^{m_1+1}D\end{pmatrix}. \tag{5.16}$$

Since the orders of the differential operators in (5.16) are the same, we conclude from this fact that the Mat_2(Virasoro) algebra (5.13) has the homogeneous gradation.

Example 3: The $\mathrm{Mat}_2(\mathrm{s1}(2,\mathbb{R}))$ algebra.

It is well known that the standard Virasoro algebra (generated by $V_m=z^{m+1}D$) contains the finite dimensional subalgebra isomorphic with the $\mathrm{s1}(2,\mathbb{R})$ algebra, which is generated by the set $\{V_{-1},V_0,V_1\}$. Analogously, the algebra Mat_2(Virasoro) (5.13) also contains the finite subalgebra which we shall call the $\mathrm{Mat}_2(\mathrm{sl}(2,\mathbb{R}))$ algebra. Indeed, using the relations (5.13) and (5.14) one can check that the generators

$$V_{i,j}=\begin{pmatrix}V_i & V_j\\ V_j & V_i\end{pmatrix}, \quad (i,j=-1,0,1), \tag{5.17}$$

close the 2-index algebra of the form

$$[V_{0,0},V_{0,0}]=[V_{0,1},V_{0,1}]=[V_{0,1},V_{1,0}]=[V_{0,-1},V_{0,-1}]=[V_{0,-1},V_{-1,0}]=[V_{1,0},V_{1,0}]$$

$$[V_{1,1},V_{1,1}]=[V_{1,-1},V_{1,-1}]=[V_{1,-1},V_{-1,1}]=[V_{-1,0},V_{-1,0}]=[V_{-1,1},V_{-1,1}]=[V_{-1,-1},V_{-1,-1}]$$
$$=0,$$

$$[V_{0,0},V_{0,1}]=[V_{0,0},V_{1,0}]=[V_{0,1},V_{1,1}]=[V_{1,0},V_{1,1}]=V_{1,1},$$

$$[V_{0,0},V_{0,-1}]=[V_{0,0},V_{-1,0}]=[V_{0,-1},V_{-1,-1}]=[V_{-1,0},V_{-1,-1}]=-V_{-1,-1},$$

$$[V_{1,1},V_{1,-1}]=[V_{1,1},V_{-1,1}]=[V_{1,-1},V_{-1,-1}]=[V_{-1,1},V_{-1,-1}]=-2V_{0,0},$$

$$[V_{0,0},V_{1,-1}]=[V_{0,0},V_{-1,1}]=V_{1,1}-V_{-1,-1},$$

$$\begin{aligned}
&[V_{0,0},V_{1,1}]=2V_{1,1}, \quad [V_{0,0},V_{-1,-1}]=-2V_{-1,-1},\\
&[V_{0,1},V_{0,-1}]=-V_{0,-1}-V_{0,1}, \quad [V_{0,1},V_{-1,0}]=-V_{-1,0}-V_{1,0},\\
&[V_{0,1},V_{-1,0}]=-V_{-1,0}-V_{1,0}, \quad [V_{0,1},V_{-1,0}]=-V_{-1,0}-V_{1,0},\\
&[V_{1,0},V_{-1,1}]=V_{1,-1}-2V_{0,-1}, \quad [V_{0,1},V_{-1,1}]=V_{-1,1}-2V_{-1,0},\\
&[V_{1,0},V_{-1,-1}]=-V_{-1,-1}-2V_{0,0}, \quad [V_{0,1},V_{-1,-1}]=-V_{-1,-1}-2V_{0,0},\\
&[V_{1,1},V_{-1,0}]=-V_{1,1}-2V_{0,0}, \quad [V_{0,-1},V_{1,0}]=V_{1,0}+V_{-1,0},\\
&[V_{1,-1},V_{-1,0}]=V_{-1,1}-2V_{0,1}, \quad [V_{0,-1},V_{1,1}]=V_{1,1}+2V_{0,0}\\
&[V_{-1,0},V_{-1,-1}]=-V_{1,-1}+2V_{1,-1}, \quad [V_{0,-1},V_{1,-1}]=-V_{1,-1}+2V_{1,0},\\
&[V_{1,0},V_{-1,0}]=-2V_{0,-1}, \quad [V_{0,-1},V_{-1,1}]=-V_{-1,1}+2V_{0,1},\\
&[V_{1,1},V_{-1,-1}]=-4V_{0,0}.
\end{aligned} \tag{5.18}$$

We also point out that the algebra (5.18) has the homogeneous gradation.

VI. FINAL REMARKS

In this paper we have presented the generalization of the central extended algebra $\hat{G}_{\mathbb{Z},\mathbf{k}}$ of the $n\times n$-matrix psd operators on the circle to the case of the nonhomogeneous gradation ($\mathrm{Mat}_n(\hat{G})$). We have shown that the $\mathrm{Mat}_n(\hat{G})$ algebra is isomorphic with the central extended algebra $\mathrm{Mat}_n(W_{1+\infty})$ whose matrix elements are determined by the compositions of two arbitrary generators of the $W_{1+\infty}$. Therefore, the $\mathrm{Mat}_n(W_{1+\infty})$ algebra can be interpreted as the algebra of the Fourier modes of the systems of two conformal fields with the arbitrary (integer) dimensions. However, in the case of the homogeneous gradation the conformal dimensions of the each pair of the fields are the same.

As the example, we have also shown that the $\mathrm{Mat}_2(W_{1+\infty})$ algebra contains the two, homogeneous graded subalgebras, namely, the Mat_2(Virasoro) and the finite dimensional $\mathrm{Mat}_2(\mathrm{sl}(2,\mathbb{R}))$ algebras. It is known[25] that the $W_{1+\infty}$ algebra can be viewed as extension of the $\mathrm{sl}(2,\mathbb{R})$ tensor algebras $\tau(\mu)$, where the parameter μ is the value of the quadratic Casimir Q of $\mathrm{sl}(2,\mathbb{R})$. Thus one may ask whether the nonhomogeneous graded $\mathrm{Mat}_n(W_{1+\infty})$ algebra can also be obtained as the universal enveloping algebra of the homogeneous graded Mat_n(Virasoro) or $\mathrm{Mat}_n(\mathrm{sl}(2,\mathbb{R}))$ algebra?

It seems to be interesting to find the representations of the central extended $\mathrm{Mat}_n(\hat{G})$ algebra ($n\geqslant 2$) in the terms of the bosonic and fermionic fields. Further, one can study their connections with the representations of the homogeneous graded matrix $W_{1+\infty}$ algebras.[19–24]

Another problem worth studying is determining the completely integrable hierarchies of the nonlinear equations (KdV and KP-type) associated with the nonhomogeneous graded algebras of the matrix psd operators.

ACKNOWLEDGMENT

The author wishes to thank the referee for many helpful suggestions on an earlier version of this paper and W. Ziubiński for fruitful discussions.

[1] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, in *Proceedings of the RIMS Symposium Nonlinear Integrable Systems*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983).
[2] M. Adler, Invent. Math. **50**, 219 (1979).
[3] I. Bakas, Commun. Math. Phys. **123**, 627 (1989).
[4] L. Bonora and C. S. Xiong, Phys. Lett. B **285**, 191 (1992).
[5] M. Adler and P. van Moerbeke, Commun. Math. Phys. **147**, 25 (1993).
[6] M. Douglas, Phys. Lett. B **238**, 176 (1990).
[7] P. Bouwknegt and K. Schoutens, Phys. Rev. B **223**(4), 183 (1993).
[8] A. O. Radul, Funkc. Anal. Priloz. **25**(1), 33 (1991).
[9] I. Vaysburd and A. O. Radul, Phys. Lett. B **274**, 317 (1992).
[10] V. Kac and D. Peterson, Proc. Natl. Acad. Sci. USA **78**(6), 3308 (1981).
[11] A. O. Radul, Phys. Lett. B **265**, 86 (1991).
[12] O. S. Kravchenko and B. A. Khesin, Funct. Anal. Priloz. **23**, 78 (1989).
[13] I. Bakas, B. Khesin, and E. Kiritsis, Commun. Math. Phys. **151**, 233 (1993).
[14] B. Khesin and I. Zakharevich, C. R. Acad. Sci. Paris, Ser. I **316**, 621 (1993); ''Poisson–Lie group of pseudodifferential symbols'' (preprint, 1993).
[15] J. M. Figueroa-O'Farrill, J. Mas, and E. Ramos, Commun. Math. Phys. **158**, 17 (1993); Phys. Lett. B **299**, 41 (1993).
[16] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, J. Phys. Soc. Jpn. **50**, 3806 (1981).
[17] L. A. Dickey, Commun. Math. Phys. **88**, 27 (1983); **163**, 509 (1994); J. Math. Phys. **32**, 2996 (1991); ''On τ-functions of Zakharov–Shabat and other matrix hierarchies of integrable equations'' (preprint OK 73019).
[18] A. Bilal, Lett. Math. Phys. **32**, 103 (1994); ''Non-local matrix generalizations of W-algebras'' (preprint PUPT-1452, 1994); ''Non-local extensions of the conformal algebra: matrix W algebra, matrix KdV hierarchies and non-Abelian Toda theories'' (preprint SLAC-PUB-LPTENS-95/1, 1995).
[19] V. Kac and A. Radul, Commun. Math. Phys. **157**, 429 (1993).
[20] I. Bakas and E. Kiritsis, Mod. Phys. Lett. A **5**(25), 2039 (1990); Prog. Theor. Phys. Suppl. **102**, 15 (1990); Nucl. Phys. B **343**, 185 (1990).
[21] M. Adler, T. Shiota, and P. van Moerbeke, Commun. Math. Phys. **171**, 547 (1995); Phys. Lett. A **194**, 33 (1994).
[22] J. van de Leur, ''The $W_{1+\infty}(gl_s)$-symmetries of the s-component KP hierarchy'' (preprint, 1994).
[23] S. Odake, Int. J. Mod. Phys. A **7**, 6339 (1992).
[24] S. Odake and T. Sano, Phys. Lett. B **258**, 369 (1991).
[25] C. N. Pope, L. J. Romans, and X. Shen, Phys. Lett. B **242**, 401 (1990).

Enlarged test-function spaces for the global free folia dynamics on the CCR-algebra

Reinhard Honegger
Institut für Theoretische Physik, Universität Tübingen, D-72076 Tübingen, Germany

(Received 19 May 1997; accepted for publication 9 July 1997)

The global free dynamics for the boson field is investigated for the test-function space $\mathscr{D}(\mathbb{R}^3)$ of infinitely differentiable functions with compact support in the position space $\mathbb{R}^3$. $\mathscr{D}(\mathbb{R}^3)$ is not invariant with respect to the one-boson dynamics, and thus the free dynamics cannot be formulated in terms of Bogoliubov automorphisms on the CCR-algebra over $\mathscr{D}(\mathbb{R}^3)$. In the Schrödinger picture the free dynamics here is constructed in terms of a group of affine bijections on suitable folia of the CCR-algebra so that the expectations are continuous in time. These folia contain a continuum of disjoint subfolia, namely the squeezed vacuum sectors. Those squeezed vacuum sectors are determined which give rise to covariant dynamical subdescriptions, that is, for which the time evolution of the states satisfies the stronger time-continuity with respect to the norm, and thus in the Heisenberg picture lead to W^*-dynamical systems. © *1998 American Institute of Physics.* [S0022-2488(98)02101-X]

I. INTRODUCTION

Infinite boson resp. photon systems occur in various algebraic investigations in statistical physics, QED, quantum optics, solid state physics, and quantum chemistry (cf., e.g., Refs. 1–10, and references therein). In these works, however, the reader may find a wide range of different choices for the one-boson test-function space E for the same infinite boson resp. photon field.[11] But the bosonic resp. photonic C^*-Weyl algebra $\mathrm{CCR}(E)$ of the *c*anonical *c*ommutation *r*elations depends sensitively from the selection of the test-function space. The inclusion $E \subseteq \mathscr{E}$ implies $\mathrm{CCR}(E) \subseteq \mathrm{CCR}(\mathscr{E})$, where $\mathrm{CCR}(E) = \mathrm{CCR}(\mathscr{E})$, if and only if $E = \mathscr{E}$.[2]

For example, in the whole Euclidean space $\mathbb{R}^3$ (position space) for the massive (spinless) boson field, resp. for the photon field with one direction of polarization, the one-particle test-function space E is a dense complex subspace of the Hilbert space of complex-valued square integrable functions $L^2(\mathbb{R}^3)$. In statistical physics for thermodynamical reasons E should consist of functions with compact support. So E usually is chosen as the complex vector space $\mathscr{D}(\mathbb{R}^3)$ of infinitely differentiable functions with compact support in $\mathbb{R}^3$, but it is also taken the subspace of those functions $f \in L^2(\mathbb{R}^3)$ which have compact support.[1–3] Unfortunately here E is not invariant with respect to the one-particle dynamics $\exp\{it(-\Delta)^{q/2}\}$, where $-\Delta \geqslant 0$ is the usual self-adjoint Laplacian on $L^2(\mathbb{R}^3)$, and $q = 1$ stands for the radiation field, whereas one has $q = 2$ for the boson field.

In other works the test-function space E is chosen to be invariant with respect to the one-particle dynamics.[4,5]

The time evolution of a quantum mechanical system can be described either in the Heisenberg or in the Schrödinger picture. For infinite C^*-algebraic systems, however, these pictures no longer are equivalent (cf. Refs. 12, 2, and references therein). The Schrödinger picture is the more general one, but the Heisenberg picture is easier to handle. Usually the Heisenberg time evolution is described by a time-continuous one-parameter group of automorphisms.

For the infinite boson resp. photon system there exists a group of Bogoliubov automorphisms on $\mathrm{CCR}(E)$ describing the free time evolution in the Heisenberg picture, if and only if the test-function space E is invariant with respect to the one-particle dynamics. However, in case of existence this automorphism group lacks any continuity in time.

The present investigation is devoted to a formulation of the global free dynamics for the massive boson field with the test-function space $E = \mathscr{D}(\mathbb{R}^3)$, which is not invariant with respect to

 © 1998 American Institute of Physics

the one-boson dynamics $\exp\{it(-\Delta)\}$. The dynamics is constructed in the Schrödinger picture not on the whole state space of CCR(E), but on suitable subsets thereof, more exactly on folia, which are large enough to contain the physically relevant states [e.g., thermodynamic equilibrium (limiting Gibbs) states for the Bose–Einstein condensation, classical states, coherent states, and also squeezed states]. Hereby we follow Roos' general global dynamical concept,[12] which in detail is outlined in Sec. II.

In the previous work [13] we studied the unique extension of certain classes of states (and representations) from CCR(E) to the enlarged CCR-algebra CCR($\mathscr{E}$) in the case where the original E is a proper subspace of the enlarged test-function space $\mathscr{E}$. Such extensions require a topology τ (stronger than or equal to the norm) so that E is τ-dense in $\mathscr{E}$. Here $\mathscr{E}:=\mathscr{S}(\mathbb{R}^3)$ is chosen as the Schwartz space, whereon the one-boson dynamics $\exp\{it(-\Delta)\}$ is τ-continuous for suitable topologies τ. Then the techniques from Ref. 13 apply to the construction of the free dynamics in terms of a group of affine bijections ν_t, $t\in\mathbb{R}$, on the folium $\mathscr{F}_\tau$ of the τ-continuous states on CCR(E) (Schrödinger picture).

The global free dynamics ν_t turns out to be time-continuous with respect to the expectations

$$\lim_{t\to 0}\langle \nu_t(\omega);A\rangle=\langle \omega;A\rangle \quad \forall A\in \mathrm{CCR}(E) \quad \forall \omega\in\mathscr{F}_\tau. \tag{1}$$

In the Heisenberg picture the free dynamics is given by a group of W^*-automorphisms α_t, $t\in\mathbb{R}$, on the weak closure of the represented CCR-algebra CCR(E), corresponding to the representation Π_τ of the folium $\mathscr{F}_\tau$. The group α_t does not leave the represented CCR-algebra invariant and does not define a W^*-dynamical system because of the absence of the W^*-continuity in time. A W^*-dynamical system, however, is only obtained when restricting the Schrödinger dynamics ν_t to an invariant subfolium $\mathscr{F}$ of $\mathscr{F}_\tau$ such that the time evolution $\mathbb{R}\ni t\mapsto\nu_t(\omega)$ is norm-continuous for every $\omega\in\mathscr{F}$—the covariance of the dynamical subdescription on $\mathscr{F}$—and then turning over to the Heisenberg subpicture.

Finally, we show that the folium $\mathscr{F}_\tau$ contains a continuum of disjoint subfolia, namely the squeezed vacuum sectors for bounded squeezing transformations. We completely determine those squeezed vacuum sectors which give rise to a covariant dynamical subdescription. The free dynamics acts noncovariant on the other ones, that is, for those states the evolution $\mathbb{R}\ni t\mapsto\nu_t(\omega)$ is not norm-continuous, but nevertheless satisfies the time-continuity of the evolving expectations of Eq. (1).

The presented results show that the global free folia dynamics in a certain sense is independent from the specific chosen test-function space E. They also outline how the global free dynamics may be formulated for other bosonic systems, especially for the photon field.

II. FOLIA, REPRESENTATIONS, AND GLOBAL DYNAMICS

Let $\mathscr{A}$ be the C^*-algebra describing a physical system. $\mathscr{S}(\mathscr{A})$ denotes the state space of $\mathscr{A}$, and for the state $\omega\in\mathscr{S}(\mathscr{A})$ the expectation at the observable $A\in\mathscr{A}$ is written as $\langle\omega;A\rangle$.

A. Folia and representations

A folium $\mathscr{F}$ of $\mathscr{A}$ is a norm-closed, convex subset of the state space $\mathscr{S}(\mathscr{A})$, which is invariant under ''perturbations,'' that is, $\omega\in\mathscr{F}$ implies $\omega_B\in\mathscr{F}$ for all $B\in\mathscr{A}$ with $\langle\omega;B^*B\rangle\neq 0$, where ω_B is the state on $\mathscr{A}$ given by $\langle\omega_B;.\rangle=\langle\omega;B^*.B\rangle\langle\omega;B^*B\rangle^{-1}$. $V\subseteq\mathscr{S}(\mathscr{A})$ is called a generating set of states for the folium $\mathscr{F}$, if $\mathscr{F}$ is the smallest folium of $\mathscr{A}$ containing all the states $\omega\in V$. Obviously, if $V\subseteq\mathscr{F}_1$ for the folium $\mathscr{F}_1$, then $\mathscr{F}\subseteq\mathscr{F}_1$.

There is a one-to-one correspondence between the folia of $\mathscr{A}$ and the (quasiequivalence classes of) representations of $\mathscr{A}$, which preserves the partial orderings, respectively; that is, if $\Pi_\mathscr{F}$ is the representation of $\mathscr{A}$ uniquely corresponding to the folium $\mathscr{F}$, then

$$\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\mathscr{S}(\mathscr{A})\Leftrightarrow\Pi_{\mathscr{F}_1}\leqslant\Pi_{\mathscr{F}_2}\leqslant\Pi_u,$$

where Π_u is the universal representation of $\mathscr{A}$. The folium $\mathscr{F}$ consists just of the $\Pi_\mathscr{F}$-normal states, that is, $\mathscr{M}_\mathscr{F}:=\overline{\Pi_\mathscr{F}(\mathscr{A})}^w=LH(\mathscr{F})^*$, where $\mathscr{N}^w$ is the weak closure of $\mathscr{N}$. The representation $\Pi_\mathscr{F}$ is quasiequivalent to the direct sum of the GNS representations Π_ω of all the states ω

in $\mathscr{F}$, resp. in a generating set V, $\Pi_{\mathscr{F}}=\oplus_{\omega\in\mathscr{F}}\Pi_\omega=\oplus_{\omega\in V}\Pi_\omega$. Especially, the GNS folium $\mathscr{F}_\omega$, that is the smallest folium containing $\omega\in\mathscr{S}(\mathscr{A})$, corresponds to the GNS representation Π_ω, i.e., Π_ω is quasiequivalent to $\Pi_{\mathscr{F}_\omega}$.

A folium constitutes a face in the state space of $\mathscr{A}$. Every folium $\mathscr{F}$ expresses a certain range of classical, macroscopic aspects of the physical system described by $\mathscr{A}$. These one also may refind in the center of the W^*-algebra $\mathscr{M}_{\mathscr{F}}$. Thus, dealing with a subfolium instead of $\mathscr{F}$ itself, means a restriction of the considered collective, macroscopic structure of the physical system. Disjoint folia belong to macroscopically totally different aspects and thus define two different superselection sectors, then

$$\|\omega_1-\omega_2\|=2\ \forall\,\omega_1\in\mathscr{F}_1\forall\,\omega_2\in\mathscr{F}_2,\quad\text{where}\quad\mathscr{F}_1\cap\mathscr{F}_2=\emptyset. \tag{2}$$

For more details to the present subsection we refer to Ref. 14, Sec. 1 and Ref. 15. A physical interpretation of the basic concepts of properties resp. filters, classical properties (that are split-faces resp. folia), disjointness (i.e., nonoverlapping classical properties) in the convex state space approach may be found in Ref. 16 for the Josephson junction and the SQUID.

B. Global Schrödinger dynamics on folia

1. Roos' dynamical concept

In the Schrödinger picture Roos[12] introduced the dynamical description of our physical system not on the whole state space, but on a folium $\mathscr{F}$ of $\mathscr{A}$.

Definition 1 (dynamical description): *Let the time evolution of the states in the folium $\mathscr{F}$ of $\mathscr{A}$ be given in terms of a one-parameter group of affine bijections $\nu_t:\mathscr{F}\to\mathscr{F}$, $t\in\mathbb{R}$, which obeys the time-continuity of the expectations*

$$\lim_{t\to 0}\langle\nu_t(\omega);A\rangle=\langle\omega;A\rangle\quad\forall A\in\mathscr{A} \tag{3}$$

for every state $\omega\in\mathscr{F}$ (group property: ν_0 is the identity, and $\nu_s\circ\nu_t=\nu_{s+t}\forall s,t\in\mathbb{R}$). Then the triple $(\mathscr{A},\mathscr{F},\nu_t)$ is denoted a dynamical description of our physical system.

In our notation we do not distinguish between $\omega\in\mathscr{F}$ as a state on $\mathscr{A}$ and its unique normal extension to $\mathscr{M}_{\mathscr{F}}$. By the duality

$$\langle\nu_t(\omega);A\rangle=\langle\omega;\alpha_t(A)\rangle\quad\forall\,\omega\in\mathscr{F}\ \forall A\in\mathscr{M}_{\mathscr{F}} \tag{4}$$

[observe that the linear hull LH($\mathscr{F}$) is the predual of $\mathscr{M}_{\mathscr{F}}$ by Sec. II A], there exists for each $t\in\mathbb{R}$ a unique Jordan automorphism α_t on $\mathscr{M}_{\mathscr{F}}$ (Ref. 2, Theorem 3.2.8). In applications, however, the group α_t actually consists of automorphisms (which is assumed in the sequel) describing the dynamics in the Heisenberg picture on the W^*-algebra $\mathscr{M}_{\mathscr{F}}$.[18] In general, however, α_t may not leave the represented C^*-algebra $\Pi_{\mathscr{F}}(\mathscr{A})$ invariant.

Roos' dynamical concept is a generalization of Kadison's original definition of a dynamical system,[17] and it is also more general than the concepts of C^*-, resp. W^*-dynamical systems given, e.g., in Ref. 2, Definition 2.7.1. $(\mathscr{M}_{\mathscr{F}},\alpha_t)$ defines a W^*-dynamical system, if and only if (3) is valid for each $A\in\mathscr{M}_{\mathscr{F}}$ (and not only for $A\in\mathscr{A}$), which is equivalent to the norm-continuity $\lim_{t\to 0}\|\nu_t(\omega)-\omega\|=0$ for all $\omega\in\mathscr{F}$. The C^*-dynamical system $(\mathscr{A},\alpha_t)$ is obtained, if and only if the folium $\mathscr{F}$ in Definition 1 is the whole state space $\mathscr{S}(\mathscr{A})$. (See Refs. 12 and 14 for more details.)

2. Covariant dynamical (sub-) descriptions and W^*-dynamical systems

The subfolium $\mathscr{F}_0\subseteq\mathscr{F}$ implies the existence of a unique orthogonal projection $c_{\mathscr{F}_0}$ in the center of $\mathscr{M}_{\mathscr{F}}$, so that $\Pi_{\mathscr{F}_0}(A)=c_{\mathscr{F}_0}\Pi_{\mathscr{F}}(A)\forall A\in\mathscr{A}$, and thus $\mathscr{M}_{\mathscr{F}_0}=c_{\mathscr{F}_0}\mathscr{M}_{\mathscr{F}}$ for the associated W^*-algebras. It holds: $\mathscr{F}_0$ is ν_t-invariant [that is $\nu_t(\mathscr{F}_0)\subseteq\mathscr{F}_0\forall t\in\mathbb{R}$], if and only if $c_{\mathscr{F}_0}$ resp. $\mathscr{M}_{\mathscr{F}_0}$ are α_t-invariant [that is $\alpha_t(c_{\mathscr{F}_0})=c_{\mathscr{F}_0}\forall t\in\mathbb{R}$]. In this case, ν_t restricted to $\mathscr{F}_0$, resp. α_t restricted to $\mathscr{M}_{\mathscr{F}_0}$, determines a dynamical subdescription of our physical system. Furthermore, for the dynamical subdescription $(\mathscr{A},\mathscr{F}_0,\nu_t)$ we have the following equivalences:

(i) The restricted Heisenberg dynamics α_t defines a W^*-dynamical system on $\mathscr{M}_{\mathscr{F}_0}$;

(ii) The more stringent time-continuity $\lim_{t\to 0}\|\nu_t(\omega)-\omega\|=0 \forall \omega\in\mathscr{F}_0$ is fulfilled;

(iii) There exists a normal representation of the W^*-algebra $\mathscr{M}_{\mathscr{F}_0}$ on some Hilbert space $\mathscr{H}_{\mathscr{F}_0}$ such that the restricted Heisenberg dynamics α_t is implementable by a strongly continuous unitary group e^{itH} acting on $\mathscr{H}_{\mathscr{F}_0}$, i.e., $\alpha_t(A)=e^{itH}Ae^{-itH}\forall A\in\mathscr{M}_{\mathscr{F}_0}$.

(The proof uses duality arguments and standard representations of W^*-algebras.[19]) Let us remark, that in general the selfadjoint Hamiltonian H in (iii) is not affiliated with $\mathscr{M}_{\mathscr{F}_0}$.

Definition 2 (covariant dynamical description): If the equivalent conditions (i)–(iii) are fulfilled, then the dynamical (sub-) description $(\mathscr{A},\mathscr{F}_0,\nu_t)$ *is called* covariant.

Suppose $\omega\in\mathscr{F}$ is a ν_t-invariant state [i.e., $\nu_t(\omega)=\omega\forall t\in\mathbb{R}$], then the GNS folium $\mathscr{F}_\omega\subseteq\mathscr{F}$ is ν_t-invariant, and thus constitutes a dynamical subdescription. Because of the uniqueness of the GNS representation $(\Pi_\omega,\mathscr{H}_\omega,\Omega_\omega)$ up to unitary equivalence, there exists a group of unitary operators U_t on $\mathscr{H}_\omega$ with $\alpha_t(A)=U_tAU_t^*\forall A\in\mathscr{M}_{\mathscr{F}_\omega}$ and $U_t\Omega_\omega=\Omega_\omega$ for each $t\in\mathbb{R}$. Equation (3) implies the strong continuity of $t\mapsto U_t$, which yields the dynamical subdescription $(\mathscr{A},\mathscr{F}_\omega,\nu_t)$ to be covariant.

III. GLOBAL QUASIFREE DYNAMICS ON THE CCR-ALGEBRA

A. Enlarged test-function space for the CCR

The one-boson test-function space is chosen as the symplectic space (E,σ), that is, E is a real vector space and $\sigma: E\times E\to\mathbb{R}$ is a (real-) bilinear mapping on E satisfying $\sigma(f,g)=-\sigma(g,f)\forall f,g\in E$. The symplectic form σ is denoted nondegenerate, if $\sigma(f,g)=0\forall g\in E$ implies $f=0$. But here σ also may be degenerate.

The associated bosonic C^*-algebra is the Weyl resp. CCR-algebra $\mathscr{W}(E,\sigma)$ over (E,σ), which is generated (with the minimal regular norm) by the unitary Weyl operators $W(f)$, $f\in E$, satisfying the Weyl relations

$$W(f)W(g)=\exp\left\{-\frac{i}{2}\,\sigma(f,g)\right\}W(f+g),\quad W(f)^*=W(-f),\quad \forall f,g\in E. \tag{5}$$

$\mathscr{W}(E,\sigma)$ is simple, if and only if σ is nondegenerate.[20,2] The state space of $\mathscr{W}(E,\sigma)$ is denoted by $\mathscr{S}(E,\sigma)$. For more details, proofs and references to the present Subsection we refer to Refs. 13,14.

1. The folium of τ-continuous states

Let τ be a vector space topology on E, such that σ is *separately* τ-continuous [i.e., $E\ni g\mapsto\sigma(f,g)=-\sigma(g,f)$ is τ-continuous for every $f\in E$]. Then the set of all τ-continuous states,

$$\mathscr{F}_\tau(E,\sigma):=\{\omega\in\mathscr{S}(E,\sigma)\,|\,E\ni f\mapsto\langle\omega;W(f)\rangle \text{ is } \tau\text{-continuous}\}, \tag{6}$$

defines a folium of the CCR-algebra $\mathscr{W}(E,\sigma)$.[13,14]

Moreover, the representation Π_τ of $\mathscr{W}(E,\sigma)$, which uniquely associates to the folium $\mathscr{F}_\tau(E,\sigma)$ by Sec. II A, is τ-continuous (and thus each subrepresentation $\Pi\leqslant\Pi_\tau$), i.e., $E\ni f\mapsto\Pi_\tau(W(f))$ is continuous with respect to τ on E and the strong operator topology on the W^*-algebra $\mathscr{M}_\tau:=\Pi_\tau(\mathscr{W}(E,\sigma))^w$.

2. Enlarged one-boson test-function spaces

An *enlarged* one-boson test-function space is a symplectic space $(\mathscr{E},\sigma)$, so that E is a subspace of $\mathscr{E}$ and the restriction of σ to E gives our genuine one-boson test-function space (E,σ). Then $\{W(f)|f\in E\}$ generates a C^*-subalgebra of $\mathscr{W}(\mathscr{E},\sigma)$, which agrees with the Weyl algebra $\mathscr{W}(E,\sigma)$, that is, we have the implication

$$(E,\sigma)\subseteq(\mathscr{E},\sigma)\Rightarrow\mathscr{W}(E,\sigma)\subseteq\mathscr{W}(\mathscr{E},\sigma).$$

It holds $\mathcal{W}(E,\sigma)=\mathcal{W}(\mathcal{E},\sigma)$, if and only if $E=\mathcal{E}$.[20,2] Thus, if E is a proper subspace of $\mathcal{E}$, then $\mathcal{W}(E,\sigma)$ is a proper C^*-subalgebra of $\mathcal{W}(\mathcal{E},\sigma)$.

3. Unique extension of τ-continuous states

Let us here suppose the enlarged test-function space $(\mathcal{E},\sigma)$ to be equipped with a vector space topology τ such that

(a) E is a *proper but τ-dense* subspace of $\mathcal{E}$, and
(b) σ is *jointly* τ-continuous [i.e., $\sigma:\mathcal{E}\times\mathcal{E}\to\mathbb{R}$ is continuous with respect to the product τ-topology on $\mathcal{E}\times\mathcal{E}$, which is equivalent to $\sigma(f_\alpha,g_\alpha)\to\sigma(f,g)$ for two arbitrary converging nets $f_\alpha\to f\in\mathcal{E}$ and $g_\alpha\to g\in\mathcal{E}$].

The original test-function space E is supposed to be equipped with the relative τ-topology. From Ref. 13 the following result is known.

Theorem 3 (τ-continuous extension of states): *Let all be as above. Then each state $\omega\in\mathcal{F}_\tau(E,\sigma)$ extends uniquely to a state $\tilde{\omega}\in\mathcal{F}_\tau(\mathcal{E},\sigma)$, that is, ω is the restriction of $\tilde{\omega}$ from $\mathcal{W}(\mathcal{E},\sigma)$ to the C^*-subalgebra $\mathcal{W}(E,\sigma)$, $\omega=\tilde{\omega}|_{\mathcal{W}(E,\sigma)}$.*

In this sense the folia $\mathcal{F}_\tau(E,\sigma)$ resp. $\mathcal{F}_\tau(\mathcal{E},\sigma)$ of the two CCR-algebras $\mathcal{W}(E,\sigma)\subset\mathcal{W}(\mathcal{E},\sigma)$ coincide,

$$\mathcal{F}_\tau\equiv\mathcal{F}_\tau(E,\sigma)\cong\mathcal{F}_\tau(\mathcal{E},\sigma). \tag{7}$$

As a consequence, the representation Π_τ (and thus each subrepresentation $\Pi\leqslant\Pi_\tau$) extends τ-continuously from $\mathcal{W}(E,\sigma)$ to $\mathcal{W}(\mathcal{E},\sigma)$, where

$$\mathcal{M}_\tau=\overline{\Pi_\tau(\mathcal{W}(E,\sigma))}^{w}=\overline{\Pi_\tau(\mathcal{W}(\mathcal{E},\sigma))}^{w}. \tag{8}$$

Example 4 (jointly continuous symplectic form): In physical applications (cf. e.g. Refs. 1–10, and Sec. IV below) the test-function space E in general is a complex pre-Hilbert space, and $\sigma(.,.)=\mathrm{Im}\langle .|.\rangle$ is defined to be the imaginary part of the complex scalar product $\langle\cdot|\cdot\rangle$, but on E, only. Let $\mathcal{E}\supset E$ be an enlarged real vector space equipped with a vector space topology τ, which is stronger than (or equal to) the norm $\|\cdot\|$ on E and for which E is τ-dense in $\mathcal{E}$. Then σ is jointly continuous with respect to the relative τ-topology on E (since σ is jointly $\|\cdot\|$-continuous), and has a unique jointly τ-continuous extension to a symplectic form on $\mathcal{E}$. Whereas σ on E is nondegenerate, its extension to $\mathcal{E}$ may lose this property (for examples, see Proposition 9 below).

B. The formulation of the global quasifree dynamics

1. Bogoliubov transformations for noninvariant test-function space

In statistical physics, and in quantum optics, resp. in nonrelativistic QED, one often is confrontated with the following situation:[13]

(a) The boson (resp. photon) system is determined in terms of a test-function space (E,σ).
(b) There exists an enlarged test-function space $(\mathcal{E},\sigma)\supset(E,\sigma)$, such that E is a proper subspace of $\mathcal{E}$.
(c) On the enlarged test-function space $(\mathcal{E},\sigma)$ there is given a symplectic transformation T, that is, T is a (real-) linear bijection $T:\mathcal{E}\to\mathcal{E}$ fulfilling the symplectic property

$$\sigma(Tf,Tg)=\sigma(f,g)\quad\forall f,g\in\mathcal{E}. \tag{9}$$

The Bogoliubov automorphism β_T on the extended Weyl algebra $\mathcal{W}(\mathcal{E},\sigma)$, which is associated with the symplectic T, is given in the Heisenberg picture by $\beta_T(W(f))=W(Tf)\forall f\in\mathcal{E}$ (cf. the Ref. 20 and 2). β_T lifts in the Schrödinger picture to the affine bijection ν_T on the state space $\mathcal{S}(\mathcal{E},\sigma)$ satisfying

$$\langle\nu_T(\omega);W(f)\rangle=\langle\omega;W(Tf)\rangle=\langle\omega;\beta_T(W(f))\rangle,\quad\forall\omega\in\mathcal{S}(\mathcal{E},\sigma)\forall f\in\mathcal{E}. \tag{10}$$

The inverse T^{-1} is a symplectic transformation on $(\mathcal{E},\sigma)$, too. Thus $(\nu_T)^{-1}=\nu_{T^{-1}}$.

However, the physically relevant one-particle test-function space is (E,σ) and not $(\mathscr{E},\sigma)$, which implies $\mathscr{W}(E,\sigma)$ to be the correct C^*-algebra describing our boson system and not $\mathscr{W}(\mathscr{E},\sigma)$. Hence it is necessary to transfer the Bogoliubov transformation ν_T (resp. β_T) from the enlarged $(\mathscr{E},\sigma)$ to the original test-function space (E,σ). Two cases have to be distinguished.

(1) Let us first assume the rare case $T(E)=E$. Then the automorphism β_T leaves the C^*-subalgebra $\mathscr{W}(E,\sigma)\subset\mathscr{W}(\mathscr{E},\sigma)$ invariant. Thus restricting β_T and the states $\nu_T(\omega)\in\mathscr{S}(\mathscr{E},\sigma)$ and $\omega\in\mathscr{S}(\mathscr{E},\sigma)$ in (10) to $\mathscr{W}(E,\sigma)$ gives the Bogoliubov transformation ν_T and β_T as a dual pair on $\mathscr{S}(E,\sigma)$ resp. $\mathscr{W}(E,\sigma)$.
(2) Unfortunately, in general the relevant one-particle test-function space E is not T-invariant [i.e., $T(f)\notin E$ for some $f\in E$], which implies that β_T does not leave the genuine CCR-algebra $\mathscr{W}(E,\sigma)$ invariant, that is, $\beta_T(A)\notin\mathscr{W}(E,\sigma)$ for some $A\in\mathscr{W}(E,\sigma)$. Thus by (10) it is not possible to transfer the Bogoliubov bijection ν_T resp. β_T uniquely from $(\mathscr{E},\sigma)$ to the original (E,σ). However, if E, $\mathscr{E}$, and T are connected with a suitable topology τ, then such a transfer can be carried through. The following result is valid for both cases.

Theorem 5: *On the enlarged test-function space $(\mathscr{E},\sigma)\supseteq(E,\sigma)$ let be given the symplectic transformation T. Suppose a vector space topology τ on the enlarged $\mathscr{E}$ such that*

(a) *E is τ-dense in $\mathscr{E}$ and σ is jointly τ-continuous.*
(b) *T and T^{-1} are τ-continuous, i.e., T is a τ-homeomorphism.*

Then it follows $\nu_T(\mathscr{F}_\tau)=\mathscr{F}_\tau$ for the folium $\mathscr{F}_\tau$ of τ-continuous states on the extended CCR-algebra $\mathscr{W}(\mathscr{E},\sigma)$.

But by Sec. III A 3, especially Eq. (7), $\mathscr{F}_\tau$ also is a folium on the genuine physically relevant CCR-algebra $\mathscr{W}(E,\sigma)$. Hence, in the restriction of ν_T to $\mathscr{F}_\tau$ we have found the Bogoliubov transformation associated with the symplectic T for our boson system with test-function space (E,σ) (Schrödinger picture).

Proof: Applying (10) to T and T^{-1} we obtain the relations $\langle\nu_T(\omega);\,W(f)\rangle=\langle\omega;W(Tf)\rangle$ and $\langle\nu_{T^{-1}}(\omega);W(f)\rangle=\langle\omega;W(T^{-1}f)\rangle$ for all $f\in\mathscr{E}$. Since by assumption T and T^{-1} are τ-continuous, these relations and the definition of $\mathscr{F}_\tau$ imply $\nu_T(\omega)\in\mathscr{F}_\tau$ and $\nu_{T^{-1}}(\omega)\in\mathscr{F}_\tau$ for every $\omega\in\mathscr{F}_\tau$. Thus $\nu_T(\mathscr{F}_\tau)\subseteq\mathscr{F}_\tau$ and $\nu_{T^{-1}}(\mathscr{F}_\tau)\subseteq\mathscr{F}_\tau$. Now $(\nu_T)^{-1}=\nu_{T^{-1}}$ gives the result. ■

In both cases the Heisenberg picture may be formulated on the W^*-algebra $\mathscr{M}_\tau$ from Eq. (8), but only in the case (a) it exists on the genuine C^*-algebra $\mathscr{W}(E,\sigma)$. By the duality $\langle\nu_T(\omega);B\rangle=\langle\omega;\alpha_T(B)\rangle,\omega\in\mathscr{F}_\tau,B\in\mathscr{M}_\tau$ [cf. Eq. (4)], we obtain the automorphism α_T on $\mathscr{M}_\tau$, which satisfies $\alpha_T(\Pi_\tau(A))=\Pi_\tau(\beta_T(A))\forall A\in\mathscr{W}(\mathscr{E},\sigma)$. Thus α_T extends the Bogoliubov transformation β_T normally in the representation Π_τ from $\mathscr{W}(\mathscr{E},\sigma)$ to $\mathscr{M}_\tau$.

Lemma 6 (joint continuity and symplectic τ-homeomorphism): Let $\mathscr{E}$ with τ constitute a Fréchet space (τ is a metrizable locally convex topology, and $\mathscr{E}$ is τ-complete). Furthermore, suppose σ to be a separately τ-continuous symplectic form on $\mathscr{E}$. Then it follows:

(a) *σ is jointly τ-continuous.*
(b) *In addition let σ be nondegenerate. Then each symplectic transformation T on $(\mathscr{E},\sigma)$ (and thus its inverse T^{-1}) is τ-continuous.*

Proof: Part (a) is a consequence of the principle of uniform boundedness (see Ref. 21, Sec. V.2). For proving part (b) we show that the graph $G(T)=\{(f,Tf)\,|\,f\in\mathscr{E}\}$ of T is a closed linear subspace of the Cartesian product Fréchet space $\mathscr{E}\times\mathscr{E}$ (product topology). Then the result follows from the closed graph theorem. Let $(f,g)\in G(T)$, that is, there is a sequence $f_n\in\mathscr{E}$, $n\in\mathbb{N}$, with $f_n\to f$ such that $Tf_n\to g$ converges, too. From $\sigma(Th_1,h_2)=\sigma(h_1,T^{-1}h_2)\forall h_j\in\mathscr{E}$, which follows from (9), and the τ-continuity of our symplectic form σ it follows

$$\sigma(g,h)=\lim_{n\to\infty}\sigma(Tf_n,h)=\lim_{n\to\infty}\sigma(f_n,T^{-1}h)=\sigma(f,T^{-1}h)=\sigma(Tf,h),\quad\forall h\in\mathscr{E}.$$

The nondegeneracy of σ gives $g=Tf$, that is $(f,g)=(f,Tf)\in G(T)$. Thus $G(T)$ is closed. ■

2. The global quasifree dynamics for noninvariant test-function space

The above concept of Bogoliubov transformations for noninvariant test-function space E especially applies to one-particle dynamics defined on enlarged test-function spaces. We consider the following situation:

(a) The boson (resp. photon) system is determined in terms of a test-function space (E,σ).
(b) On an enlarged test-function space $(\mathscr{E},\sigma)\supseteq(E,\sigma)$ there is given the one-boson dynamics, that is a group of symplectic transformations $v_t, t\in\mathbb{R}$, on $(\mathscr{E},\sigma)$ (v_0 is the identity on $\mathscr{E}$, and $v_s\circ v_t=v_{s+t}$ for all times $s,t\in\mathbb{R}$).

Let us remark here, that physicists are interested in a Schrödinger dynamics for physically relevant states, only, and not necessarily for the whole state space. A dynamics in addition should obey some continuity in time, at least for the expectations as in Eq. (3). For nontrivial one-particle dynamics such a time-continuity is not given by the (above) group of Bogoliubov transformations $\nu_t:=v_{v_t}$ on $\mathscr{S}(\mathscr{E},\sigma)$ (Schrödinger picture), resp. $\beta_t:=\beta_{v_t}$ on $\mathscr{W}(\mathscr{E},\sigma)$ (Heisenberg picture), since $\|W(f)-W(g)\|=2\forall f\neq g$. As in Theorem 5 let us connect E, $\mathscr{E}$, and ν_t with a suitable topology τ.

Theorem 7 (global quasifree dynamics): *Let all be as above, and suppose a vector space topology τ on the enlarged $\mathscr{E}$ such that*

(a) *E is τ-dense in $\mathscr{E}$ and σ is jointly τ-continuous.*
(b) *$\mathscr{E}\ni f\mapsto v_t f\in\mathscr{E}$ is τ-continuous for each $t\in\mathbb{R}$. [Since $\nu_{-t}=(\nu_t)^{-1}\forall t\in\mathbb{R}$ by the group property, it follows that each ν_t is a τ-homeomorphism.]*
(c) *$\mathbb{R}\ni t\mapsto v_t f\in\mathscr{E}$ is τ-continuous for each $f\in\mathscr{E}$.*

Then it holds $\nu_t(\mathscr{T}_\tau)=\mathscr{T}_\tau\forall t\in\mathbb{R}$, and the evolving expectations $\mathbb{R}\ni t\mapsto\langle\nu_t(\omega);A\rangle$ are continuous for all $\omega\in\mathscr{T}_\tau$ and all $A\in\mathscr{W}(\mathscr{E},\sigma)\supset\mathscr{W}(E,\sigma)$.

Since $\mathscr{T}_\tau$ is also a folium on the genuine CCR-algebra $\mathscr{W}(E,\sigma)$, we arrive at the dynamical description $(\mathscr{W}(E,\sigma),\mathscr{T}_\tau,\nu_t)$, which defines the time-continous quasifree Schrödinger dynamics for our boson system with test-function space (E,σ) and one-boson dynamics ν_t on the enlarged $(\mathscr{E},\sigma)$.

Proof: $\nu_t(\mathscr{T}_\tau)=\mathscr{T}_\tau$ follows from Theorem 5. (10) implies $\langle\nu_t(\omega);W(f)\rangle=\langle\omega;W(v_t f)\rangle$. The τ-continuity of $t\mapsto v_t f$ for $f\in\mathscr{E}$ yields the continuity of the expectations $t\mapsto\langle\nu_t(\omega);W(f)\rangle=\langle\omega;W(v_t f)\rangle$ for all $\omega\in\mathscr{T}_\tau$ and $f\in\mathscr{E}$. With an $\epsilon/3$-argument one obtains the continuity of an arbitrary evolving expectation $t\mapsto\langle\nu_t(\omega);A\rangle$ with $\omega\in\mathscr{T}_\tau$ and $A\in\mathscr{W}(\mathscr{E},\sigma)\supset\mathscr{W}(E,\sigma)$. ■

The quasifree dynamics ν_t is an example of Roos' dynamical concept outlined in Sec. II B 1, especially Definition 1, and it generalizes the idea in Ref. 12, Appendix B. However, in general $(\mathscr{W}(E,\sigma),\mathscr{T}_\tau,v_t)$ is not covariant, which is shown in Sec. IV H 4 below.

The present formulation of a quasifree dynamics already has been used in our recent works[14,22] on the global dynamics for a class of interactions between the boson gas and an infinite mean field quantum lattice system. Applications are found in the Dicke model[6] and in the Josephson oscillator model,[7] where the radiation field couples weakly to infinitely many two-level atoms resp. to two superconductors. There the one-boson resp. one-photon dynamics on complete enlarged test-function spaces plays an important role for the interacting dynamics and for the emitted radiation, which is quantum optically coherent and which reflects the classical, collective ordering of the atoms resp. superconductors.

IV. THE GLOBAL FREE DYNAMICS FOR THE BOSON FIELD

Let us make some notational remarks. Throughout the present section the one-boson test-function space is a complex vector space. On the one-particle level, with the expressions linear and antilinear we always mean complex-linear resp. complex-antilinear mappings, and in the notations operator and unitary we include complex-linearity. (Complex) scalar products always are chosen to be (complex-) linear in the right factor. Real-linear one-particle objects (operators resp. scalar products) always are denoted so.

A. The one-boson test-function space

Let us from now on consider massive spinless bosons in the euclidean space $\mathbb{R}^p$, where $p \in \mathbb{N}$. In statistical physics for thermodynamical reasons one usually takes the complex vector space $\mathscr{D}(\mathbb{R}^p)$ of infinitely differentiable functions with compact support in $\mathbb{R}^p$ (position space) for the test-function space E. The symplectic form σ on E is chosen as the imaginary part of the usual scalar product $\langle f|g\rangle = \int_{\mathbb{R}^p} f g\, d^p x$. This ensures the quasilocal structure, which decomposes the bosonic Weyl algebra $\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|.\rangle)$ into the local C^*-algebras $\mathscr{W}(\mathscr{D}(\Lambda), \mathrm{Im}\langle .|.\rangle)$ for certain open bounded regions $\Lambda \subset \mathbb{R}^p$. Thus, defining the local Gibbs states on the local algebras one may perform the thermodynamic limit $\Lambda \to \mathbb{R}^p$, e.g., for obtaining Bose–Einstein condensation.[1,2]

B. The one-boson dynamics on the enlarged symplectic space

The one-boson dynamics $v_t = \exp\{-it\Delta\}$ is obtained as the thermodynamic limit of the local one-boson time translations. Here the one-boson Hamiltonian $-\Delta \geqslant 0$ is the usual selfadjoint Laplacian on the complex Hilbert space $L^2(\mathbb{R}^p)$ of square integrable functions with respect to the Lebesgue measure $d^p x$, Ref. 21, Sec. IX 7.

It is well known that for $0 \neq f \in \mathscr{D}(\mathbb{R}^p)$ the evolved $v_t(f)$ no longer has compact support, i.e., $v_t(f) \notin \mathscr{D}(\mathbb{R}^p)$ for $t \neq 0$ [this may be shown by use of Fourier transformation, and analyticity arguments, and the Paley–Wiener theorem (Ref. 21, Theorem IX.11)]. Since the unitaries v_t are symplectic on $(L^2(\mathbb{R}^p), \mathrm{Im}\langle .|.\rangle)$, we are in the situation of Sec. III B 2. We need an enlarged test-function space $(\mathscr{E}, \mathrm{Im}\langle .|.\rangle)$ with a complex, v_t-invariant space $\mathscr{D}(\mathbb{R}^p) \subset \mathscr{E} \subseteq L^2(\mathbb{R}^p)$, which is connected with $\mathscr{D}(\mathbb{R}^p)$ by a suitable topology τ on $\mathscr{E}$. A suitable choice is the Schwartz space $\mathscr{S}(\mathbb{R}^p)$, the complex-valued functions of rapid decrease on $\mathbb{R}^p$, which is v_t-invariant. Thus, for our enlarged test-function space $(\mathscr{E}, \sigma)$ let us take here $(\mathscr{S}(\mathbb{R}^p) \mathrm{Im}\langle .|.\rangle)$. Summarizing we have the following situation:

(a) Our boson system is determined by the test-function space $(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|.\rangle)$.
(b) On the enlarged test-function space $(\mathscr{S}\,\mathbb{R}^p), \mathrm{Im}\langle .|.\rangle)$ there is given the one-boson dynamics in terms of the group of symplectic transformations $v_t = \exp\{-it\Delta\}$, $t \in \mathbb{R}$.

C. Topologies on the enlarged symplectic space

The natural Fréchet topology on $\mathscr{S}(\mathbb{R}^p)$ is denoted by τ_{Fr}. The original test-function space $\mathscr{D}(\mathbb{R}^p)$ is τ_{Fr}-dense in $\mathscr{S}(\mathbb{R}^p)$.

As usual the τ_{Fr}-dual is denoted by $\mathscr{S}'(\mathbb{R}^p)$, the space of the tempered distributions (Ref. 21, Sec. V.3). Let $\mathscr{S}'_m(\mathbb{R}^p)$ be the set of those positive measures μ on $\mathbb{R}^p$, for which $\mathscr{S}(\mathbb{R}^p) \ni f \to \int_{\mathbb{R}^p} f\, d\mu$ is τ_{Fr}-continuous, a tempered distribution, in which sense $\mathscr{S}'_m(\mathbb{R}^p)$ is a subspace of $\mathscr{S}'(\mathbb{R}^p)$. Using the Fourier transformation $f \mapsto \hat{f}$ we construct for each $\mu \in \mathscr{S}'_m(\mathbb{R}^p)$ the new scalar product

$$\langle f|g\rangle_\mu := \langle f|g\rangle + \int_{\mathbb{R}^p} \overline{\hat{f}(k)}\,\hat{g}(k)\, d\mu(k), \quad f,g \in \mathscr{S}(\mathbb{R}^p).$$

Obviously, $\mu = 0$ gives $\langle .|.\rangle_0 = \langle .|.\rangle$, the usual scalar product.

Proposition 8: For each $\mu \in \mathscr{S}'_m(\mathbb{R}^p)$ we have the following orderings:

$$\|.\| \leqslant \|.\|_\mu < \tau_{\mathrm{Fr}}, \quad \mathscr{T}_{\|.\|} \subseteq \mathscr{T}_{\|.\|_\mu} \subset \mathscr{T}_{\mathrm{Fr}}$$

for the topologies $\|.\|$, and $\|.\|_\mu$, and τ_{Fr}, resp. for the associated folia of continuous states. The last inclusion is proper. It is $\mathscr{T}_{\|.\|} \neq \mathscr{T}_{\|.\|_\mu}$, if and only if the norm-topology $\|.\|_\mu$ is strictly stronger than the usual norm $\|.\|$.

Proof: First recall that the multiplication $(f,g) \to fg$ of functions on $\mathscr{S}(\mathbb{R}^p)$ is jointly τ_{Fr}-continuous, and that the Lebesgue measure $d^p k$ is an element of $\mathscr{S}'_m(\mathbb{R}^p)$. Thus $\|.\|_\mu$ is strictly weaker than τ_{Fr}. The implication for the folia of continuous states immediately follows from the ordering of the topologies and the definition (6). Let us prove that $\mathscr{T}_{\|.\|_\mu} \neq \mathscr{T}_{\mathrm{Fr}}$. For each complex-linear form $L: \mathscr{S}(\mathbb{R}^p) \to \mathbb{C}$ consider the fully coherent state ω_L defined in terms of the characteristic

function $\langle \omega_L ; W(f)\rangle = \exp\{-\frac{1}{4}\| f \|^2 + i\sqrt{2}\,\mathrm{Re}(L(f))\}$.[8] Since there exists a linear form L, which is τ_{Fr}-continuous, but not $\|.\|_\mu$-continuous, it follows that $\omega_L \in \mathscr{F}_{\mathrm{Fr}}$ but $\omega_L \notin \mathscr{F}_{\|.\|_\mu}$. ■

D. The global free boson dynamics in the Schrödinger picture

Since in momentum space (Fourier transformation) the operator $v_t = \exp\{it(-\Delta)\}$ is the multiplication with the function $\mathbb{R}^p \ni k \to \exp\{it\|k\|^2\}$, it is easily seen that the one-boson dynamics v_t on $\mathscr{S}(\mathbb{R}^p)$ is continuous with respect to every topology $\|.\|_\mu$, $\mu \in \mathscr{S}'_m(\mathbb{R}^p)$, and the Fréchet topology τ_{Fr}, that is, $f \to v_t f$ is continuous for each $t \in \mathbb{R}$ and $t \to v_t f$ is continuous for each $f \in \mathscr{S}(\mathbb{R}^p)$. The joint continuity of the symplectic form $\mathrm{Im}\langle .|. \rangle$ with respect to $\|.\|_\mu$ resp. τ_{Fr} follows from the joint continuity of the scalar product $\langle .|. \rangle$ with respect to the usual norm $\|.\|$ and from the relation $\|.\| \leqslant \|.\|_\mu < \tau_{\mathrm{Fr}}$ (cf. also Example 4).

Thus all assumptions of Theorem 7 are fulfilled. According to Sec. III B 2 and by use of the topology τ_{Fr} we now construct the global free boson dynamics in the Schrödinger picture in terms of the dynamical description $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle), \mathscr{F}_{\mathrm{Fr}}, \nu_t)$ for the original one-boson test-function space $(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$. By construction it holds

$$\langle \nu_t(\omega); W(f)\rangle = \langle \omega; W(\exp\{-it\Delta\} f)\rangle \quad \forall \omega \in \mathscr{F}_{\mathrm{Fr}} \quad \forall f \in \mathscr{S}(\mathbb{R}^p). \tag{11}$$

Since the one-boson dynamics v_t also is $\|.\|_\mu$-continuous, it is immediate to check that the subfolium $\mathscr{F}_{\|.\|_\mu}$ is ν_t-invariant for each $\mu \in \mathscr{S}'_m(\mathbb{R}^p)$, and thus defines the dynamical subdescription $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle), \mathscr{F}_{\|.\|_\mu}, \nu_t)$. This dynamical subdescription obviously agrees with the dynamical description constructed according to Theorem 7 directly on $\mathscr{F}_{\|.\|_\mu}$ by use of the topology $\|.\|_\mu$.

E. Further enlarged test-function spaces

The Schwartz space $\mathscr{S}(\mathbb{R}^p)$ is complete with respect to τ_{Fr}, but not with respect to $\|.\|_\mu$, $\mu \in \mathscr{S}'_m(\mathbb{R}^p)$. The one-boson dynamics ν_t continuously extends from $\mathscr{S}(\mathbb{R}^p)$ to a strongly continuous unitary group on the $\|.\|_\mu$-completion $\mathscr{K}(\mu)$. Thus the original test-function space $\mathscr{D}(\mathbb{R}^p)$ may be enlarged to the complex Hilbert space $\mathscr{K}(\mu)$, where the symplectic form $\sigma \equiv \mathrm{Im}\langle .|. \rangle$ also extends from $\mathscr{D}(\mathbb{R}^p)$, resp. $\mathscr{S}(\mathbb{R}^p)$ to $\mathscr{K}(\mu)$ by joint $\|.\|_\mu$-continuity, its extension is denoted by σ_μ. Thus, according to Eq. (7) it holds

$$\mathscr{F}_{\|.\|_\mu} \equiv \mathscr{F}_{\|.\|_\mu}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle) \cong \mathscr{F}_{\|.\|_\mu}(\mathscr{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle) \cong \mathscr{F}_{\|.\|_\mu}(\mathscr{K}(\mu), \sigma_\mu). \tag{12}$$

Obviously, since $\|.\| = \|.\|_0$ it is $\mathscr{K}(0) = L^2(\mathbb{R}^p)$ and $\sigma_0 = \mathrm{Im}\langle .|. \rangle$ for $\mu = 0$. Using the topology $\|.\|_\mu$ we here also may take $(\mathscr{K}(\mu), \sigma_\mu)$ as the enlarged test-function space. However, the free dynamics constructed on $\mathscr{F}_{\|.\|_\mu}$ this way agrees with the one constructed by use of $(\mathscr{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$ as enlarged test-function space.

F. Independence from the test-function spaces

If E is an arbitrary τ_{Fr}-dense complex subspace of $\mathscr{S}(\mathbb{R}^p)$, then by the above construction one also may take $(E, \mathrm{Im}\langle .|. \rangle)$ for the test-function space,

$$\mathscr{F}_{\mathrm{Fr}} \equiv \mathscr{F}_{\mathrm{Fr}}(E, \mathrm{Im}\langle .|. \rangle) \cong \mathscr{F}_{\mathrm{Fr}}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle) \cong \mathscr{F}_{\mathrm{Fr}}(\mathscr{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle).$$

Similarly, if $(E, \sigma_\mu) \subseteq (\mathscr{K}(\mu), \sigma_\mu)$ is a $\|.\|_\mu$-dense symplectic subspace, then in addition to (12) it holds $\mathscr{F}_{\|.\|_\mu} \cong \mathscr{F}_{\|.\|_\mu}(E, \sigma_\mu)$. In both cases the free dynamics ν_t is well-defined.

This demonstrates some kind of independence of the chosen test-function space. But subsequently, as above, we take $(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$ for our test-function space.

G. Physically relevant Fréchet-continuous states

Here let us illustrate, that the folium $\mathscr{F}_{\mathrm{Fr}}$ contains a whole range of physically relevant states.

1. The (Fock) vacuum sector

Let $\mathscr{F}_{\rm vac}$ be the GNS folium of the (Fock) vacuum state $\omega_{\rm vac}$ (the smallest folium containing $\omega_{\rm vac}$). By Sec. II A the so-called vacuum sector $\mathscr{F}_{\rm vac}$ corresponds to the Fock representation $\Pi_{\rm Fock}$ of the CCR-algebra (the GNS representation of $\omega_{\rm vac}$). Thus $\mathscr{F}_{\rm vac}$ consists of all states, which are Fock-normal, i.e., which are given by density operators on Fock space.

The vacuum state $\omega_{\rm vac}$ has the characteristic function[2]

$$\langle \omega_{\rm vac}; W(f)\rangle = \exp\{-\tfrac{1}{4}\|f\|^2\}, \quad f\in\mathscr{D}(\mathbb{R}^p).$$

Hence $\omega_{\rm vac}\in\mathscr{F}_{\|.\|}$, and thus $\mathscr{F}_{\rm vac}\subseteq\mathscr{F}_{\|.\|}$.

The characteristic function of the vacuum implies $\nu_t(\omega_{\rm vac})=\omega_{\rm vac}\forall t\in\mathbb{R}$ by (11). Thus from Sec. II B 2 it follows that the vacuum sector $\mathscr{F}_{\rm vac}$ defines the covariant dynamical subdescription $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p),{\rm Im}\langle .|.\rangle),\mathscr{F}_{\rm vac},\nu_t)$, see Definition 2. It is well known that $\exp\{itd\Gamma(-\Delta)\}$ unitarily implements the free Heisenberg dynamics α_t restricted to the Fock vacuum sector, where $d\Gamma(-\Delta)$ means the self-adjoint second quantization of the one-boson Hamiltonian $-\Delta$ [cf. condition (iii) of Sec. II B 2, and Ref. 2, Sec. 5.2].

2. The limiting Gibbs states for the Bose–Einstein condensation

The local Gibbs states by construction are elements of $\mathscr{F}_{\rm vac}$ [they are given on Fock space by density operators which are constructed in terms of the second quantization of the Laplacians $-\Delta_\Lambda$ for the local regions $\Lambda\subset\mathbb{R}^p$ (Ref. 2, Sec. 5.2.5)]. Folia are norm-closed subsets of the state space $\mathscr{S}(\mathscr{D}(\mathbb{R}^p),{\rm Im}\langle .|.\rangle)$. But the thermodynamic limiting procedures are done in the weak*-topology, which is strictly weaker than the norm. Thus one may not expect that the limiting Gibbs states are Fock normal, too. Indeed, by the subsequent result the limiting Gibbs states are not contained in $\mathscr{F}_{\|.\|}$, and hence the thermodynamic limits lead beyond the folia $\mathscr{F}_{\rm vac}$, resp. $\mathscr{F}_{\|.\|}$. We set $\mathbb{R}_+:=[0,\infty[$.

Proposition 9: The following assertions hold:

(a) *Let $F:\mathbb{R}_+\to\mathbb{R}$ be a measurable function such that there exist constants $c>0$ and $n\in\mathbb{N}$ with $|F(\lambda)|^2\leqslant c\lambda^{(1-p)/2}$ for $0<\lambda<1$ and $|F(\lambda)|^2\leqslant c\lambda^n$ for $\lambda\geqslant 1$. Then $\mu_F\in\mathscr{S}'_m(\mathbb{R}^p)$ for the measure $d\mu_F(k):=|F(\|k\|^2)|^2d^pk$. Moreover, $\|.\|_{\mu_F}$ is the graph norm of the self-adjoint operator $F(-\Delta)$ on $L^2(\mathbb{R}^p)$. $\mathscr{S}(\mathbb{R}^p)$ is a core for $F(-\Delta)$ and the $\|.\|_{\mu_F}$-completion $\mathscr{K}(\mu_F)\subseteq L^2(\mathbb{R}^p)$ is the domain of $F(-\Delta)$. The extended symplectic form $\sigma_{\mu_F}(f,g)={\rm Im}\langle f|g\rangle$, $f,g\in\mathscr{K}(\mu_F)$, is nondegenerate.*

(b) *Let F be as in (a). Denote by δ_{k_0} the point measure $\delta_{k_0}(\hat f)=\hat f(k_0)\forall f\in\mathscr{S}(\mathbb{R}^p)$ at some $k_0\in\mathbb{R}^p$ in momentum space (Fourier transformation $f\mapsto\hat f$). Then $\mu_F+\delta_{k_0}\in\mathscr{S}'_m(\mathbb{R}^p)$, and $\mathscr{K}(\mu_F+\delta_{k_0})=\mathscr{K}(\mu_F)\oplus\mathbb{C}$ with the embedding $\mathscr{S}(\mathbb{R}^p)\ni f\to f\oplus\hat f(k_0)$. Moreover, the extended symplectic form $\sigma_{\mu_F+\delta_{k_0}}(f\oplus\alpha,g\oplus\beta)={\rm Im}\langle f|g\rangle$, $f\oplus\alpha,g\oplus\beta\in\mathscr{K}(\mu_F)\oplus\mathbb{C}$, is degenerate.*

(c) *Each limiting Gibbs state $\omega_{\beta,\gamma}\in\mathscr{S}(\mathscr{D}(\mathbb{R}^p),{\rm Im}\langle .|.\rangle)$ for the Bose–Einstein condensation (Refs. 1 and 2, Sec. 5.2.5, also Refs. 23, 24) has the characteristic function*

$$\langle \omega_{\beta,\gamma}; W(f)\rangle = \exp\{-\tfrac{1}{4}\|f\|^2-\tfrac{1}{2}\|T_\beta^{1/2}f\|^2-\gamma|\hat f(0)|^2\}, \quad f\in\mathscr{D}(\mathbb{R}^p),$$

where $T_\beta:=e^{\beta\Delta}(\mathbb{1}-e^{\beta\Delta})^{-1}$ at the inverse temperature $\beta>0$. No condensation implies $\gamma=0$. With condensation it is $\gamma>0$, which depends on the particle density.

Let $G(\lambda):=\lambda^{-1/2}$ for $0<\lambda<\infty$ and dimension $p\geqslant 2$. Then for each $\beta>0$ and $\gamma\geqslant 0$ it follows that $\omega_{\beta,\gamma}\in\mathscr{F}_{\|.\|_\mu}$, where $\mu:=\mu_G+\delta_0$. But it holds $\omega_{\beta,\gamma}\notin\mathscr{F}_{\|.\|}$.

Proof: (a) For each $f\in\mathscr{S}(\mathbb{R}^p)$ it is

$$\left|\int_{\mathbb{R}^p} f(k)d\mu_F(k)\right| \leqslant c\int_{\|k\|<1}\frac{|f(k)|}{\|k\|^{p-1}}\,d^pk \;+\; c\int_{\|k\|\geqslant 1}(|f(k)|\|k\|^{2n+p+1})\,\frac{1}{\|k\|^{p+1}}\,d^pk$$
$$\leqslant c'\|f\|_\infty + c'\|\,|.|^{2n+p+1}f\|_\infty,$$

for some constant $c'>0$, since $\int_{\|k\|<1}\|k\|^{-p+1}d^pk<\infty$ and $\int_{\|k\|\geq 1}\|k\|^{-p-1}d^pk<\infty$. But $\|f\|_\infty$ and $\|\|.\|^{2n+p+1}f\|_\infty$ are τ_{Fr}-continuous seminorms on $\mathscr{S}(\mathbb{R}^p)$, which ensures $\mu_F\in\mathscr{S}'_m(\mathbb{R}^p)$. In momentum space (Fourier transformation) the operator $F(-\Delta)$ is the multiplication with $k\to F(\|k\|^2)$. It is well known that $\mathscr{S}(\mathbb{R}^p)$ is a core for locally bounded multiplication operators on $L^2(\mathbb{R}^p)$ (Ref. 25, p. 92).

(b) Let us show that the embedding of $\mathscr{S}(\mathbb{R}^p)$, $f\mapsto f\oplus\hat{f}(k_0)=:E(f)$, is dense in the Hilbert space $\mathscr{K}(\mu_F)\oplus\mathbb{C}$. Let $h\in\mathscr{S}(\mathbb{R}^p)$ so that $\hat{h}$ has support in $\{k\in\mathbb{R}^p|\|k\|\leq 1\}$ and $\hat{h}(0)=1$. Define the sequence $h_n\in\mathscr{S}(\mathbb{R}^p)$ by $\widehat{h_n}(k):=\hat{h}(n(k-k_0))$, $n\in\mathbb{N}$. Let first $k_0=0$. Then with $q:=nk$ we get

$$\|h_n\|^2_{\mu_F}\leq\int_{\mathbb{R}^p}|\hat{h}(nk)|^2\left(1+\frac{c}{\|k\|^{p-1}}\right)d^pk\overset{q=nk}{=}\frac{\|h\|^2}{n^p}+\frac{c}{n}\int_{\mathbb{R}^p}\frac{|\hat{h}(q)|^2}{\|q\|^{p-1}}d^pq.$$

If $k_0\neq 0$, then there is an $n_0\in\mathbb{N}$ with $0\notin\{k|\|k-k_0\|\leq 1/n_0\}$, and there is a $\tilde{c}>0$ with $|F(\|k\|^2)|^2\leq\tilde{c}$ for $\|k-k_0\|\leq 1/n_0$. For all $n\geq n_0$ we obtain

$$\|h_n\|^2_{\mu_F}\leq\int_{\mathbb{R}^p}|\hat{h}(n(k-k_0))|^2(1+\tilde{c})d^pk=\frac{(1+\tilde{c})\|h\|^2}{n^p}.$$

Thus in both cases $\|h_n\|_{\mu_F}\to 0$, but $\widehat{h_n}(k_0)=1\,\forall n\in\mathbb{N}$. [That is, the linear form $f\mapsto\hat{f}(k_0)$ is unbounded with respect to $\|.\|_{\mu_F}$.] It follows $h_n\oplus\widehat{h_n}(k_0)\to 0\oplus 1$, which implies $0\oplus\mathbb{C}$ to be a subspace of the closure of the embedded E($\mathscr{S}(\mathbb{R}^p)$), and thus $\mathscr{S}(\mathbb{R}^p)\oplus 0$ is so, too. Now observe that by part (a) it holds that $\mathscr{S}(\mathbb{R}^p)$ is $\|.\|_{\mu_F}$-dense in $\mathscr{K}(\mu_F)$.

(c) We show that for each $\beta>0$ the graph norm of $T_\beta^{1/2}$ is equivalent to the norm $\|.\|_{\mu_G}$, the graph norm of $G(-\Delta)=(-\Delta)^{-1/2}$. The unboundedness of the operators $T_\beta^{1/2}$ and $(-\Delta)^{-1/2}$ on $L^2(\mathbb{R}^p)$ arises at the spectral value zero. Thus it sufficess to prove the equivalence around zero. It is $\beta\lambda\leq e^{\beta\lambda}-1\leq b\lambda$, and thus $(\beta\lambda)^{-1}\geq e^{-\beta\lambda}(1-e^{-\beta\lambda})^{-1}=(e^{\beta\lambda}-1)^{-1}\geq(b\lambda)^{-1}$ for $0<\lambda\leq\epsilon$ for some $b>\beta$ and $\epsilon>0$. Then the equivalence of the graph norms follows from the spectral calculus, especially the domains of $T_\beta^{1/2}$ and $(-\Delta)^{-1/2}$ agree to be $\mathscr{K}(\mu_G)$. Finally, use part (b). ■

The limiting Gibbs states $\omega_{\beta,\gamma}$, $\beta>0,\gamma\geq 0$, are ν_t-invariant, which, e.g., is seen from applying (11) to their characteristic functions. Thus by Sec. II B 2 their GNS folia define covariant dynamical subdescriptions of $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p),\mathrm{Im}\langle.|.\rangle),\mathscr{T}_{\mathrm{Fr}},\nu_t)$.

3. Quasifree states

The above local and limiting Gibbs states are examples of quasifree states, also called Gaussian states. Each *quasifree* state ω on $\mathscr{W}(\mathscr{D}(\mathbb{R}^p),\mathrm{Im}\langle.|.\rangle)$ has the characteristic function

$$\langle\omega;\mathscr{W}(f)\rangle=\exp\{i\sqrt{2}\ \mathrm{Re}(L(f))-\tfrac{1}{4}s(f,f)\},\quad f\in\mathscr{D}(\mathbb{R}^p),$$

where $L:\mathscr{D}(\mathbb{R}^p)\to\mathbb{R}$ is a complex-linear form and $s:\mathscr{D}(\mathbb{R}^p)\times\mathscr{D}(\mathbb{R}^p)\to\mathbb{R}$ is a positive symmetric real-bilinear form satisfying $|\mathrm{Im}\langle f|g\rangle|^2\leq s(f,f)s(g,g)\,\forall f,g\in\mathscr{D}(\mathbb{R}^p)$. Conversely, for each such L and s, there exists a unique quasifree state ω of the above form.[2,24,10]

If L is τ_{Fr}-continuous and s is jointly τ_{Fr}-continuous, then $\omega\in\mathscr{T}_{\mathrm{Fr}}$. Observe, if s is a symmetric real-bilinear form on the enlarged, τ_{Fr}-complete $\mathscr{S}(\mathbb{R}^p)$, then the joint τ_{Fr}-continuity of s arises from the separate τ_{Fr}-continuity, i.e., $\mathscr{S}(\mathbb{R}^p)\ni g\mapsto s(f,g)=s(g,f)$ is τ_{Fr}-continuous for every $f\in\mathscr{S}(\mathbb{R}^p)$ [this is a consequence of the principle of uniform boundedness (see. Ref. 21, Sec. V.2)].

4. Classical states

The above local and limiting Gibbs states also are examples for classical states. In quantum optics a classical state of the photon field is defined by the positivity of its P-representation (for more informations and references we refer to Refs. 10 and 26). The generalization to our infinite

dimensional test-function space leads to the following definition: A state ω on the CCR-algebra $\mathscr{W}(\mathscr{D}(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$ is called *classical*, if its characteristic function decomposes multiplicatively according to

$$\langle \omega; W(f)\rangle = \langle \omega_{\mathrm{vac}}; W(f)\rangle P_\omega(f), \quad f\in\mathscr{D}(\mathbb{R}^p), \tag{13}$$

with a positive-definite function $P_\omega:\mathscr{D}(\mathbb{R}^p)\to\mathbb{C}$ on the test-function space $\mathscr{D}(\mathbb{R}^p)$ regarded as an additive group. Conversely, if $P:\mathscr{D}(\mathbb{R}^p)\to\mathbb{C}$ is a positive-definite function with the normalization $P(0)=1$, then there exists a unique classical state ω with $P=P_\omega$.

If P_ω is τ_{Fr}-continuous, then $\omega\in\mathscr{F}_{\mathrm{Fr}}$. In this case we may decompose the positive-definite P_ω with help of the Bochner–Minlos theorem,

$$P_\omega(f)=\int_{\mathscr{S}'(\mathbb{R}^p)} \exp\{i\sqrt{2}\ \mathrm{Re}(L(f))\}\ \ d\rho_\omega(L), \quad f\in\mathscr{S}(\mathbb{R}^p), \tag{14}$$

[ω extends to $\mathscr{S}(\mathbb{R}^p)$, thus P_ω does so, too] with respect to a unique probability measure ρ_ω on the tempered distributions $\mathscr{S}'(\mathbb{R}^p)$.[27] Equations (13) and (14) imply the decomposition of the classical state $\omega\in\mathscr{F}_{\mathrm{Fr}}$,

$$\omega=\int_{\mathscr{S}'(\mathbb{R}^p)} \omega_L\ \ d\rho_\omega(L), \tag{15}$$

into the pure, fully coherent states $\omega_L\in\mathscr{F}_{\mathrm{Fr}}$, $L\in\mathscr{S}'(\mathbb{R}^p)$, with respect to the weak*-topology. For arbitrary linear form $L:\mathscr{D}(\mathbb{R}^p)\to\mathbb{C}$ the pure, fully coherent ω_L is given with the characteristic function (see Ref. 8 and Sec. IV G 5 below),

$$\langle \omega_L; W(f)\rangle = \langle \omega_{\mathrm{vac}}; W(f)\rangle \exp\{i\sqrt{2}\ \mathrm{Re}(L(f))\}, \quad f\in\mathscr{D}(\mathbb{R}^p).$$

These states have been used already in the proof of Proposition 8.

The measure ρ_ω is the P-representation of our classical state $\omega\in\mathscr{F}_{\mathrm{Fr}}$. And for $\|.\|$-continuous linear forms L, the states ω_L are just the Glauber vector states, which are well known from quantum optics.

5. Coherent states

A smearing procedure of Glauber's original factorization condition[28] leads to the algebraic formulation of quantum optical coherence, where Glauber's factorizing coherence function is replaced by a (complex-) linear form $G:\mathscr{D}(\mathbb{R}^p)\to\mathbb{C}$. For more details on coherent states we refer to Ref. 8. An analytic state ω on $\mathscr{W}(\mathscr{D}(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$ is called *G-coherent in nth order*, $n\in\mathbb{N}\cup\{\infty\}$, if the normally ordered expectations factorize up to degree n,

$$\langle \omega; a^*(f_1)\cdots a^*(f_j)a(g_1)\cdots a(g_j)\rangle = G(f_1)\cdots G(f_j)\overline{G(g_1)}\cdots\overline{G(g_j)}$$

for all f_k, $g_t\in\mathscr{D}(\mathbb{R}^p)$ and each $1\leqslant j\leqslant n$ [here, $a(f)$ and $a^*(f)$ are the annihilation resp. creation operators for the test-function f].

For $\|.\|$-continuous linear form G, the G-coherent states in every order are Fock normal, i.e., they are elements of the vacuum sector $\mathscr{F}_{\mathrm{vac}}$. If G is unbounded with respect to the norm $\|.\|$, then each G-coherent state ω automatically is classical and decomposes with a suitable probability measure ρ_ω^G on $\mathbb{C}$ into the pure coherent states ω_{zG},

$$\omega=\int_{\mathbb{C}} \omega_{zG}\ \ d\rho_\omega^G(z). \tag{16}$$

Obviously, if $G\in\mathscr{S}'(\mathbb{R}^p)$, then this decomposition agrees with (15) with the measure $\rho_\omega^G\cong\rho_\omega$ supported by $\mathbb{C}\cong\mathbb{C}G\subset\mathscr{S}'(\mathbb{R}^p)$.

Summarizing, if the linear form G is a tempered distribution, $G\in\mathscr{S}'(\mathbb{R}^p)$, then the G-coherent states in all orders are elements of the folium $\mathscr{F}_{\mathrm{Fr}}$.

6. Squeezed states

Squeezing operations in quantum optics usually are described in terms of Bogoliubov transformations.[29,9,10,26,30] We adopt this notion also for our massive bosons.

Let T be a symplectic transformation on $(\mathcal{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$. Then T and T^{-1} are τ_{Fr}-continuous by Lemma 6(b). In Proposition 13(c) below we give an example for such a τ_{Fr}-continuous symplectic transformation. By Theorem 5 the associated (squeezing) Bogoliubov transformation ν_T is an affine bijection on the folium $\mathcal{F}_{\mathrm{Fr}}$ that is $\nu_T(\mathcal{F}_{\mathrm{Fr}}) = \mathcal{F}_{\mathrm{Fr}}$.

Thus applying the squeezing operation ν_T to the states from the previous subsections, we obtain squeezed quasifree, classical, resp. coherent states, which are τ_{Fr}-continuous, too.

H. Squeezed vacua and their time evolutions

In contrast to Sec. IV G 6 we here exclusively consider symplectic transformations T on the enlarged test-function spaces $(L^2(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$ as candidates for (squeezing) Bogoliubov transformations ν_T.

Let us denote by $\mathcal{B}$ the group of all symplectic transformations T on $(L^2(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$. Remark, in general $T \in \mathcal{B}$ is only real-linear; T commutes with the multiplication with the complex "i" (that is the complex-linearity of T), if and only if T is a (complex-linear) unitary on $L^2(\mathbb{R}^p)$. By Lemma 6(b) each $T \in \mathcal{B}$ is $\|.\|$-bounded, and thus $T^{-1} \in \mathcal{B}$ is so, too. Consequently, according to Theorem 5 the Bogoliubov transformations $\nu_T, T \in \mathcal{B}$, act on the folium $\mathcal{F}_{\|.\|}$ of $\|.\|$-continuous states on the CCR-algebra $\mathcal{W}(\mathcal{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$.

1. The polar decomposition of symplectic transformations

With the following procedure one may construct all $T \in \mathcal{B}$. Let $S \geqslant 0$ be a bounded self-adjoint operator on $L^2(\mathbb{R}^p)$ commuting with the antilinear involution J, and let U be an arbitrary unitary. Then the real-linear operator

$$T = U \exp\{JS\} = U(\cosh(S) + J \sinh(S)) \tag{17}$$

on $L^2(\mathbb{R}^p)$ is an element of $\mathcal{B}$, i.e., T is a symplectic transformation with respect to Im $\langle .|. \rangle$. Conversely, each $T \in \mathcal{B}$ is of the form (17) with unique bounded self-adjoint $S \geqslant 0$ commuting with the antilinear involution J, and unitary U. (17) is the polar decomposition of $T \in \mathcal{B}$.[9]

2. The squeezed vacuum sectors

It is $\omega_{\mathrm{vac}} \in \mathcal{F}_{\mathrm{vac}} \subseteq \mathcal{F}_{\|.\|}$. We denote the transformed state $\nu_T(\omega_{\mathrm{vac}})$ a *squeezed vacuum*, and the transformed folium $\nu_T(\mathcal{F}_{\mathrm{vac}})$ a *squeezed vacuum sector*, where $T \in \mathcal{B}$. Obviously, the squeezed vacuum sector $\nu_T(\mathcal{F}_{\mathrm{vac}})$ is the GNS folium of the squeezed vacuum state $\nu_T(\omega_{\mathrm{vac}})$, the smallest folium containing $\nu_T(\omega_{\mathrm{vac}})$, which is also a subfolium of $\mathcal{F}_{\|.\|}$.

Let us denote by $\mathcal{B}_{\mathrm{vac}}$ the subgroup of $\mathcal{B}$ consisting of those $T \in \mathcal{B}$ for which $\nu_T(\mathcal{F}_{\mathrm{vac}}) = \mathcal{F}_{\mathrm{vac}}$. By the Appendix it holds $T \in \mathcal{B}_{\mathrm{vac}}$, if and only if the self-adjoint $S \geqslant 0$ occurring in the polar decomposition $T = U \exp\{JS\}$ of (17) is a Hilbert–Schmidt operator on $L^2(\mathbb{R}^p)$. Especially, for the one-boson dynamics it is $v_t \in \mathcal{B}_{\mathrm{vac}}$ for all times $t \in \mathbb{R}$.

Proposition 10: Let $T_1, T_2 \in \mathcal{B}$. *Then the following assertions hold:*

(a) $T_1 T_2^{-1} \in \mathcal{B}_{\mathrm{vac}} \Leftrightarrow \nu_{T_1}(\omega_{\mathrm{vac}}) \in \nu_{T_2}(\mathcal{F}_{\mathrm{vac}}) \Leftrightarrow \nu_{T_1}(\mathcal{F}_{\mathrm{vac}}) = \nu_{T_2}(\mathcal{F}_{\mathrm{vac}})$.

(b) $T_1 T_2^{-1} \notin \mathcal{B}_{\mathrm{vac}} \Leftrightarrow \nu_{T_1}(\mathcal{F}_{\mathrm{vac}}) \cap \nu_{T_2}(\mathcal{F}_{\mathrm{vac}}) = \emptyset$.

Especially, it holds $\nu_T(\mathcal{F}_{\mathrm{vac}}) \cap (\mathcal{F}_{\mathrm{vac}}) = \emptyset \Leftrightarrow T \in \mathcal{B} \setminus \mathcal{B}_{\mathrm{vac}}$.

Proof: (a) $\nu_{T_1}(\mathcal{F}_{\mathrm{vac}}) = \nu_{T_2}(\mathcal{F}_{\mathrm{vac}})$ is equivalent to $\nu_V(\mathcal{F}_{\mathrm{vac}}) = \mathcal{F}_{\mathrm{vac}}$ for $V = T_1 T_2^{-1}$ (observe $\nu_V = (\nu_{T_2}{}^{-1} \circ \nu_{T_1})$. Now use the Appendix. (b) Suppose $\omega \in \nu_V(\mathcal{F}_{\mathrm{vac}}) \cap \mathcal{F}_{\mathrm{vac}} \neq \emptyset$. Then $\mathcal{F}_\omega \in \nu_V(\mathcal{F}_{\mathrm{vac}}) \cap \mathcal{F}_{\mathrm{vac}}$ for the GNS folium $\mathcal{F}_\omega$ of ω. But the irreducibility of the GNS representations of $\nu_V(\omega_{\mathrm{vac}})$ and ω_{vac} (the Fock representation is irreducible) implies $\nu_V(\mathcal{F}_{\mathrm{vac}}) = \mathcal{F}_\omega = \mathcal{F}_{\mathrm{vac}}$. ■

3. Squeezed vacuum sectors as covariant dynamical (sub-) descriptions

Let us denote by $\mathscr{B}_{\mathrm{cov}}$ the set of those symplectic transformations $T\in\mathscr{B}$, for which the squeezed vacuum sector $\nu_T(\mathscr{F}_{\mathrm{vac}})$ is invariant with respect to our free boson dynamics ν_t, i.e., $\nu_t(\nu_T(\mathscr{F}_{\mathrm{vac}}))=\nu_T(\mathscr{F}_{\mathrm{vac}})\forall\, t\in\mathbb{R}$. Especially, by Sec. IV G 1 the vacuum sector $\mathscr{F}_{\mathrm{vac}}$ is ν_t-invariant, but it is $\nu_T(\mathscr{F}_{\mathrm{vac}})=\mathscr{F}_{\mathrm{vac}}$ for all $T\in\mathscr{B}_{\mathrm{vac}}$. Consequently, $\mathscr{B}_{\mathrm{vac}}\subseteq\mathscr{B}_{\mathrm{cov}}$.

All $T\in\mathscr{B}_{\mathrm{cov}}$ are determined by Theorem 11.

Theorem 11: *For $T\in\mathscr{B}$ we have the following equivalences*:

(i) $T\in\mathscr{B}_{\mathrm{cov}}$, *that is,* $\nu_t(\nu_T(\mathscr{F}_{\mathrm{vac}}))=\nu_T(\mathscr{F}_{\mathrm{vac}})$ *for all* $t\in\mathbb{R}$;

(ii) $Tv_tT^{-1}\in\mathscr{B}_{\mathrm{vac}}$ *for all* $t\in\mathbb{R}$;

(iii) $\nu_t(\nu_T(\omega_{\mathrm{vac}}))\in\nu_T(\mathscr{F}_{\mathrm{vac}})$ *for all* $t\in\mathbb{R}$;

(iv) *for some* $a>0$ *both* $P_a(T^+T-\mathbb{I})$ *and* $[T^+T,i(\mathbb{I}-P_a)\Delta]$ *are real-linear Hilbert–Schmidt operators on the Hilbert space* $L^2(\mathbb{R}^p)$ *regarded as a real Hilbert space with the real scalar product* $\mathrm{Re}\,\langle .|.\rangle$ *(here the real adjoint* T^+ *is with respect to* $\mathrm{Re}\langle .|.\rangle$, *and* P_a *is the complex-linear spectral projection of* $-\Delta$ *for the spectral values larger than a; with the decomposition (17) it is* $T^+T=\exp\{2JS\}=\cosh(2S)+J\sinh(2S)$).

Furthermore, the dynamical subdescription $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle),\nu_T(\mathscr{F}_{\mathrm{vac}}),\nu_t)$ *is covariant (cf. Definition 2) for every* $T\in\mathscr{B}_{\mathrm{cov}}$.

Proof: Let us extend v_t to $L^2(\mathbb{R}^p)$, so that the Bogoliubov automorphisms β_T and $\beta_t=\beta_{v_t}$ act on $\mathscr{W}(L^2(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$. Then $\Pi_T:=\Pi_{\mathrm{Fock}}\circ\beta_T$ is the representation of $\mathscr{W}(L^2(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$ which corresponds to the folium $\nu_T(\mathscr{F}_{\mathrm{vac}})$ by Sec. II A, and which commonly is called symplectically related to the Fock representation. Observe, $\nu_V=(\nu_T)^{-1}\circ\nu_t\circ\nu_T$ with the symplectic $V:=Tv_tT^{-1}\in\mathscr{B}$ is the dual Bogoliubov bijection to the automorphism $\beta_V=\beta_T\beta_t(\beta_T)^{-1}$ on $\mathscr{W}(L^2(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$. Thus, applying Theorem 14 of the Appendix yields (ii) to be equivalent to the unitary implementability of β_V in the Fock representation Π_{Fock}, which agrees with the unitary implementability of β_t in Π_T. Now the theorem is a reformulation of Warchall's results.[31] Warchall analyzed the implementability of Bogoliubov transformations associated to a unitary group $\exp\{itC\}$ with self-adjoint generator $C\geqslant 0$ on the complex separable Hilbert space $\mathscr{V}$ [thus symplectic on $(\mathscr{V},\mathrm{Im}\,\langle .|.\rangle)$] in representations of $\mathscr{W}(\mathscr{V},\mathrm{Im}\langle .|.\rangle)$ symplectically related to Fock. An important feature of Warchall's results is the fact that the ν_t-invariance of a squeezed vacuum sector $\nu_T(\mathscr{F}_{\mathrm{vac}})$ for every $t\in\mathbb{R}$, that is $T\in\mathscr{B}_{\mathrm{cov}}$, automatically gives its covariance. ■

In Ref. 31 there are some examples for $T\in\mathscr{B}_{\mathrm{cov}}$ but $T\notin\mathscr{B}_{\mathrm{vac}}$, and in Proposition 13(b) below there are constructed examples for $T\in\mathscr{B}\setminus\mathscr{B}_{\mathrm{cov}}$. Hence we have the proper inclusions

$$\mathscr{B}_{\mathrm{vac}}\subset\mathscr{B}_{\mathrm{cov}}\subset\mathscr{B}.$$

4. Noncovariant free dynamics on squeezed vacuum sectors

As a consequence of Theorem 11 we obtain Corollary 12.

Corollary 12: The following assertions are valid:

(a) *For* $T\in B_{\mathrm{cov}}$ *it is* $\lim_{t\to 0}\|\nu_t(\omega)-\omega\|=0$ *for all* $\omega\in\nu_T(\mathscr{F}_{\mathrm{vac}})$.

(b) *Let* $T\in B$ *but* $T\notin B_{\mathrm{cov}}$. *Then there exists a sequence* $t_n\neq 0$, $n\in\mathbb{N}$, *with* $\lim_{n\to\infty}t_n=0$, *but it holds* $\|\nu_{t_n}(\omega)-\omega\|=2$ *for all* $n\in\mathbb{N}$ *and all* $\omega\in\nu_T(\mathscr{F}_{\mathrm{vac}})$.

Proof: (a) expresses the covariance, see Sec. II B 2. (b) If there would exist a neigborhood $\mathscr{N}$ of $t=0$ with $\nu_t(\nu_T(\omega_{\mathrm{vac}}))\in\nu_T(\mathscr{F}_{\mathrm{vac}})\forall\, t\in\mathscr{N}$, then the group property of ν_t implies that this is valid for all $t\in\mathbb{R}$. Thus by Theorem 11 in each neighborhood of $t=0$ there exist some $t\neq 0$ with $\nu_t(\nu_T(\omega_{\mathrm{vac}}))\notin\nu_T(\mathscr{F}_{\mathrm{vac}})$, and thus $\nu_t(\nu_T(\mathscr{F}_{\mathrm{vac}}))\cap\nu_T(\mathscr{F}_{\mathrm{vac}})=\emptyset$. With (2) the result follows. ■

Let us construct an example to part (b) of the above Corollary. For the bounded continuous function $\zeta:\mathbb{R}_+\to\mathbb{R}_+$ define the bounded self-adjoint $S:=\zeta(-\Delta)\geqslant 0$ with the one-boson Hamiltonian $-\Delta$, and let the antilinear involution J be the ordinary complex conjugation on $L^2(\mathbb{R}^p)$, $Jf=\bar{f}$. Then by Sec. IV H 1 the real-linear operator $T_\zeta:=\exp\{JS\}=\cosh(\zeta(-\Delta))+J\sinh(\zeta(-\Delta))$ is symplectic on $(L^2(\mathbb{R}^p),\mathrm{Im}\langle .|.\rangle)$, that is $T_\zeta\in\mathscr{B}$. Here, J, S, and v_t, resp $-\Delta$

mutually commute (observe, $v_t J = J v_{-t}$), a situation which is similarly to our investigation[30] concerning unitary implementations of Bogoliubov automorphism groups in representations symplectically related to Fock.

Proposition 13: Let $\zeta, \eta: \mathbb{R}_+ \to \mathbb{R}_+$ *be bounded and continuous. Then it holds*:

(a) *It is* $\nu_{T_\zeta}(\mathscr{F}_{\rm vac}) \cap \nu_{T_\eta}(\mathscr{F}_{\rm vac}) = \emptyset$, *if and only if* $\zeta \neq \eta$. *Especially, if* $\zeta \neq 0$, *then* $\nu_{T_\zeta}(\mathscr{F}_{\rm vac}) \cap \mathscr{F}_{\rm vac} = 0$, *that is,* $T_\zeta \in \mathscr{B} \setminus \mathscr{B}_{\rm vac}$.

(b) *Let* $\zeta \neq 0$. *Then* $\nu_s(\nu_{T_\zeta}(\mathscr{F}_{\rm vac})) \cap \nu_t(\nu_{T_\zeta}(\mathscr{F}_{\rm vac})) = \emptyset$ *for all times* $s \neq t$, *especially it holds* $T_\zeta \in \mathscr{B} \setminus \mathscr{B}_{\rm cov}$. *Thus* $\|\nu_s(\omega_1) - \nu_t(\omega_2)\| = 2 \quad \forall\, \omega_1, \omega_2 \in \nu_{T_\zeta}(\mathscr{F}_{\rm vac})$ *for each* $s \neq t$ *by Eq. (2).*

(c) *If* ζ *is infinitely differentiable with polynomially bounded derivatives, then* T_ζ *leaves the Schwartz space* $\mathscr{S}(\mathbb{R}^p)$ *invariant,* $T_\zeta(\mathscr{S}(\mathbb{R}^p)) = \mathscr{S}(\mathbb{R}^p)$, *and the restrictions* $T_\zeta|_{\mathscr{S}(\mathbb{R}^p)}$ *and* $T_\zeta^{-1}|_{\mathscr{S}(\mathbb{R}^p)}$ *are* $\tau_{\rm Fr}$*-continuous symplectic transformations on* $(\mathscr{S}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$. *Thus the Bogoliubov bijection* ν_{T_ζ} *may be defined to act on the folium* $\mathscr{F}_{\rm Fr}$ *of* $\tau_{\rm Fr}$*-continuous states on the original CCR-algebra* $\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$.

Proof: (a) $\Leftrightarrow \nu_{T_\zeta T_\eta^{-1}}(\mathscr{F}_{\rm vac}) \cap \mathscr{F}_{\rm vac} = \emptyset \Leftrightarrow$ the antilinear part $(T_\zeta T_\eta^{-1})_a$ is not Hilbert–Schmidt on $L^2(\mathbb{R}^p)$, by the Appendix. But $(T_\zeta T_\eta^{-1})_a = J \sinh(\zeta(-\Delta) - \eta(-\Delta))$.

(b) $\Leftrightarrow \nu_{T_\zeta v_{t-s} T_\zeta^{-1}}(\mathscr{F}_{\rm vac}) \cap \mathscr{F}_{\rm vac} = \varnothing \Leftrightarrow$ the antilinear part $(T_\zeta v_{t-s} T_\zeta^{-1})_a$ is not Hilbert–Schmidt on $L^2(\mathbb{R}^p)$, by the Appendix. But $(T_\zeta v_{t-s} T_\zeta^{-1})_a = iJ \sinh(2\zeta(-\Delta)) \sin((t-s)\Delta)$.

(c) By (Ref. 21, Sec. V.3) the operator of multiplication with $\mathbb{R}^p \ni k \mapsto \exp\{\pm\zeta(\|k\|^2)\}$ acts $\tau_{\rm Fr}$-continuous on the Schwartz space. Thus with Fourier transformation it follows that $\cosh(\zeta(-\Delta))$ and $\sinh(\zeta(-\Delta))$ are $\tau_{\rm Fr}$-continuous on $\mathscr{S}(\mathbb{R}^p)$. But then $T_\zeta = \exp\{JS\} = \cosh(S) + J \sinh(S)$ and $T_\zeta^{-1} = \exp\{-JS\} = \cosh(S) - J \sinh(S)$ are $\tau_{\rm Fr}$-continuous on $\mathscr{S}(\mathbb{R}^p)$, too. This especially implies $T_\zeta(\mathscr{S}(\mathbb{R}^p)) \subseteq \mathscr{S}(\mathbb{R}^p)$ and $T_\zeta^{-1}(\mathscr{S}(\mathbb{R}^p)) \subseteq \mathscr{S}(\mathbb{R}^p)$, which yields the result.

■

5. Summary

We have seen that the folium $\mathscr{F}_{\|.\|} \subset \mathscr{F}_{\rm Fr}$ of $\|.\|$-continuous states on $\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle)$ contains a continuum of disjoint subfolia, namely the squeezed vacuum sectors $\nu_T(\mathscr{F}_{\rm vac}), T \in \mathscr{B}$, see Proposition 10 and Proposition 13(a).

By the construction in Sec. IV D the global free dynamics ν_t on $\mathscr{F}_{\|.\|}$ is time-continuous with respect to the expectations (3),

$$\lim_{t \to 0} \langle \nu_t(\omega); A \rangle = \langle \omega; A \rangle \quad \forall A \in \mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle) \quad \forall\, \omega \in \mathscr{F}_{\|.\|}. \tag{18}$$

The set of those symplectic transformations $T \in \mathscr{B}$, for which the squeezed vacuum sector $\nu_T(\mathscr{F}_{\rm vac})$ is invariant with respect to the free dynamics ν_t, has been denoted by $\mathscr{B}_{\rm cov}$. Let $\mathscr{F}_{\rm sq,cov}$ be the folium generated by $\{\nu_T(\mathscr{F}_{\rm vac}) | T \in \mathscr{B}_{\rm cov}\}$. Then, restricting in (18) the states ω to the ν_t-invariant subfolium $\mathscr{F}_{\rm sq,cov} \subset \mathscr{F}_{\|.\|}$, we obtain from Corollary 12(a) the stronger time-continuity with respect to the norm,

$$\lim_{t \to 0} \|\nu_t(\omega) - \omega\| = 0 \quad \forall\, \omega \in \mathscr{F}_{\rm sq,cov},$$

which ensures the dynamical subdescription $(\mathscr{W}(\mathscr{D}(\mathbb{R}^p), \mathrm{Im}\langle .|. \rangle), \mathscr{F}_{\rm sq,cov}, \nu_t)$ to be covariant according to Definition 2.

For every state ω from the squeezed vacuum sectors $\nu_T(\mathscr{F}_{\rm vac})$ with $T \in \mathscr{B} \setminus \mathscr{B}_{\rm cov}$, however, the norm-continuity of $\mathbb{R} \ni t \mapsto \nu_t(\omega)$ fails by Corollary 12(b).

APPENDIX: UNITARY IMPLEMENTATIONS ON FOCK SPACE

Let us present a result, which is used in Sec. IV H. It is due to Shale,[32] and generalized in Ref. 9, Theorem 3.1. Let $\mathscr{V}$ be a complex Hilbert space with scalar product $\langle .|. \rangle$, and V a symplectic

transformation on $(\mathscr{V},\mathrm{Im}\langle .|.\rangle)$. The associated Bogoliubov automorphism on the CCR-algebra $\mathscr{W}(\mathscr{V},\mathrm{Im}\langle .|.\rangle)$ is denoted by β_V, that is, $\beta_V(W(f))=W(Vf)\forall f\in\mathscr{V}$, its dual transformation is ν_V according to (10).

Theorem 14: *The following assertions are equivalent*:

(i) $\nu_V(\omega_{\mathrm{vac}})\in\mathscr{F}_{\mathrm{vac}}$,
(ii) $\nu_V(\mathscr{F}_{\mathrm{vac}})=\mathscr{F}_{\mathrm{vac}}$,
(iii) *β_V is unitarily implementable in the Fock representation Π_{Fock}, that is, there exists a unitary U on Fock space with $\Pi_{\mathrm{Fock}}(\beta_V(A))=U\Pi_{\mathrm{Fock}}(A)U^*\forall A\in\mathscr{W}(\mathscr{V},\mathrm{Im}\langle .|.\rangle)$*;
(iv) *the antilinear part $V_a=\frac{1}{2}(V+iVi)$ of the symplectic V is Hilbert–Schmidt on $\mathscr{V}$, or equivalently, $V^+V-\mathbb{1}$ is Hilbert–Schmidt on $\mathscr{V}$ regarded as real Hilbert space with the real scalar product* $\mathrm{Re}\,\langle .|.\rangle$ *(here the real adjoint V^+ is taken with respect to* $\mathrm{Re}\,\langle .|.\rangle$*)*;
(v) *the self-adjoint operator $S\geqslant 0$ is Hilbert–Schmidt, where S is from the polar decomposition $V=U\exp\{JS\}$ of (17).*

The irreducibility of the Fock representation[2] *implies U in (iii) to be unique up to a phase.*

ACKNOWLEDGMENT

The author is grateful to Professor A. Rieckers for very useful discussions, and to the Deutsche Forschungsgemeinschaft for financial support.

[1] J. T. Cannon, ''Infinite volume limits of the canonical free Bose gas states on the Weyl algebra,'' Commun. Math. Phys. **29**, 89–104 (1973); J. T. Lewis and J. V. Pulè, ''The equilibrium states of the free boson gas,'' *ibid.* **36**, 1–18 (1974); D. A. Dubin, *Solvable Models in Algebraic Statistical Mechanics* (Clarendon, Oxford, 1974); R. Brendle, ''Einstein condensation in a macroscopic field,'' Z. Naturforsch. Teil A **40**, 1189–1198 (1985); J. Hertle, ''Macroscopically inhomogeneous bose Einstein condensation,'' in *Proceedings of NATO ASI on Large Scale Molecular Systems: Quantum and Stochastical Aspects*, edited by W. Gans, A. Blumen, and A. Amann (Plenum, London, 1991).
[2] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I, II* (Springer, Berlin, Heidelberg, New York, 1979, 1981).
[3] J. Hertle and R. Honegger, ''Limiting Gibbs states and dynamics for thermal photons,'' J. Math. Phys. **33**, 143–151 (1992).
[4] P. Pfeifer, *Chiral Molecules—a Superselection Rule Induced by the Radiation Field*, Ph.D. thesis, ETH-Zürich, No. 6551, ok Gotthard S+D AG, 1980; M. Fannes, B. Nachtergaele, and A. Verbeure, ''The equilibrium states of the spin-boson model,'' Commun. Math. Phys. **114**, 537–548 (1988); H. Spohn, ''Ground state(s) of the spin-boson Hamiltonian,'' *ibid.* **123**, 277–304 (1989); A. Amann, ''Molecules coupled to their environment,'' in *Proceedings of NATO ASI on Large Scale Molecular Systems: Quantum and Stochastical Aspects*, edited by W. Gans, A. Blumen, and A. Amann (Plenum, London, 1991); A. Amann, ''Chirality: A superselection rule generated by the molecular environment?,'' J. Math. Chem. **6**, 1–15 (1991); R. Honegger, ''On the dynamics and the temperature states of the spin-boson model,'' Lett. Math. Phys. **21**, 351–359 (1991).
[5] R. Honegger, A. Rieckers, and Th. Unnerstall, ''Perturbational dynamics of the infinite Dicke model,'' Z. Naturforsch. Teil A **48**, 1043–1053 (1993); R. Honegger and A. Rieckers, ''Quantized radiation states from the infinite Dicke model,'' Publ. RIMS Kyoto Univ. **30**, 111–138 (1994).
[6] R. Honegger, ''The weakly coupled infinite Dicke model,'' Physica A **225**, 391–411 (1996).
[7] Th. Gerisch, R. Honegger, and A. Rieckers, ''The Josephson microwave radiation states from tunneling cooper pairs,'' Tübingen (preprint, 1997).
[8] R. Honegger and A. Rapp ''General Glauber coherent states on the Weyl algebra and their phase integrals,'' Physica A **167**, 945–961 (1990); R. Honegger and A. Rieckers, ''The general form of non-Fock coherent boson states,'' Publ. RIMS Kyoto Univ. **26**, 397–417 (1990); ''First order coherent boson states,'' Helv. Phys. Acta **65**, 966–984 (1992); ''On higher order coherent states on the Weyl algebra,'' Lett. Math. Phys. **24**, 221–225 (1992); R. Honegger, ''The general form of the microscopic coherent boson states,'' Physica A **198,** 179–209 (1993); ''The extremal microscopic coherent boson states,'' Lett. Math. Phys. **28**, 155–164 (1993).
[9] R. Honegger and A. Rieckers, ''Squeezing Bogoliubov transformations on the infinite mode CCR-algebra,'' J. Math. Phys. **37**, 4292–4309 (1996).
[10] R. Honegger and A. Rieckers, ''Squeezed variances of smeared boson fields,'' Helv. Phys. Acta **70**, 507–541 (1997).
[11] Usually in physics (cf., e.g., Refs. 1–10 and Sec. IV) the one-boson test-function space E (resp. $\mathscr{E}$) is a (in general noncomplete) complex pre-Hilbert space with complex scalar product $\langle .|.\rangle$. The symplectic form σ on E, which is necessary for the Weyl form of the CCR in Eq. (5), is chosen as the imaginary part of the scalar product, $\sigma(f,g)=\mathrm{Im}\langle f|g\rangle\forall f,g\in E$. With the notation used in Sec. III A the CCR-algebra is given by $\mathrm{CCR}(E)=\mathscr{W}(E,\sigma)$.
[12] H. Roos, ''KMS condition in a Schrödinger picture of the dynamics,'' Physica A **100**, 183–195 (1980).
[13] R. Honegger, ''On the continuous extension of states on the CCR-algebra,'' Lett. Math. Phys. **42**, 11–25 (1997).
[14] R. Honegger, ''Global cocycle dynamics for infinite mean field quantum systems interacting with the boson gas,'' J. Math. Phys. **37**, 263–282 (1996).
[15] R. Haag, R. V. Kadison, and D. Kastler, ''Nets of C^*-algebras and classification of states,'' Commun. Math. Phys. **16**, 81–104 (1970); G. L. Sewell, ''States and dynamics of infinitely extended physical systems,'' *ibid.* **33**, 43–51 (1973); R.

V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras I, II* (Academic, New York, London, 1983, 1986).
[16] A. Rieckers, ''Concepts of macroscopic quantum phenomena at the SQUID,'' Tübingen (preprint, 1997).
[17] R. V. Kadison, *Topology 3* (Pergamon, Oxford, 1965), Suppl. 2, pp. 177–198.
[18] G. L. Sewell, ''W^*-dynamics of infinite quantum systems,'' Lett. Math. Phys. **6**, 209–213 (1982).
[19] U. Haagerup, ''The standard form of von Neumann algebras,'' Math. Scand. **37**, 271–283 (1975); S. Strătilă and L. Zsidó, *Lectures on von Neumann Algebras* (Abacus, Turnbridge Wells, Kennt, 1979); A. Rieckers, ''On the covariance representation of global quantum dynamics and its symmetries,'' in *Proceedings of the XV International Colloquium on Group Theoretical Methods in Physics*, 1986, edited by R. Gilmore (World Scientific, Singapore, 1987); A. Rieckers and H. Roos, ''Implementations of Jordan-isomorphisms for general von Neumann algebras,'' Ann. Inst. Henri Poincaré **50**, 95–113 (1989).
[20] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, ''The smallest C^*-algebra for canonical commutation relations,'' Commun. Math. Phys. **32**, 231–243 (1973).
[21] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I, II* (Academic, New York, London, 1980, 1975).
[22] R. Honegger, ''Time-asymptotic boson states from infinite mean field quantum systems coupled to the boson gas,'' Lett. Math. Phys. **27**, 191–203 (1993).
[23] H. Araki and E. J. Woods, ''Representations of the canonical commutation relations describing a nonrelativistic infinite free bose gas,'' J. Math. Phys. **4**, 637–662 (1963).
[24] R. Honegger, ''Decomposition of positive sesquilinear forms and the central decomposition of gauge-invariant quasi-free states on the Weyl-algebra,'' Z. Naturforsch. Teil A **45**, 17–28 (1990).
[25] J. Weidmann, *Linear Operators in Hilbert Spaces* (Springer, Berlin, New York, 1980).
[26] R. Honegger and A. Rieckers, ''Squeezing of optical states on the CCR-algebra,'' (to be published); R. Honegger and A. Rieckers, ''Squeezing operations in Fock space and beyond,'' Physica A **242**, 423–438 (1997).
[27] Y. Umemura, ''Measures on infinite dimensional vector spaces,'' Publ. RIMS Kyoto Univ. **A1**, 1–47 (1965); H. H. Schäfer, *Topological Vector Spaces* (MacMillan, New York, 1966); T. Hida, *Brownian Motion* (Springer, Berlin, New York, 1980).
[28] R. J. Glauber, ''The quantum theory of optical coherence,'' Phys. Rev. **130**, 2529–2539 (1963); ''Coherent and incoherent states of the radiation field,'' *ibid.* **131**, 2766–2788 (1963); in *Quantum Optics and Electronics*, Les Houches 1964, edited by C. de Witt, A. Blandin, and C. Cohen-Tannoudji (Gordon and Breach, New York, 1965); U. M. Titulaer and R. J. Glauber, ''Density operators for coherent fields,'' Phys. Rev. **145**, 1041–1050 (1966).
[29] H. P. Yuen, ''Two-photon coherent states of the radiation field,'' Phys. Rev. A **13**, 2226–2243 (1976); D. F. Walls, ''Squeezed states of light,'' Nature (London) **306**, 141–146 (1983); R. Loudon and P. L. Knight, ''Squeezed light,'' J. Mod. Opt. **34**, 709–759 (1987); X. Ma and W. Rhodes, ''Multimode squeeze operators and squeezed states,'' Phys. Rev. A **41**, 4625–4631 (1990).
[30] R. Honegger and A. Rieckers, *Unitary Implementations of Squeezing Groups* (to be published).
[31] H. A. Warchall, ''Implementation of automorphism groups in certain representations of the canonical commutation relations,'' J. Math. Phys. **23**, 2221–2228 (1982).
[32] D. Shale, ''Linear symmetries of free boson fields,'' Trans. Am. Math. Soc. **103**, 149–167 (1962).

Group analysis and renormgroup symmetries

Vladimir F. Kovalev
Institute for Mathematical Modelling of Russian Academy of Sciences, Miusskaya 4-A, Moscow, 125047, Russia

Veniamin V. Pustovalov
P. N. Lebedev Physical Institute, Leninskii pr., 53, Moscow, 117924, Russia

Dmitrii V. Shirkov
Bogoliubov Laboratory, Joint Institute for Nuclear Research, Dubna, 141980, Russia

(Received 1 July 1996; accepted for publication 6 August 1997)

An original regular approach to constructing special type symmetries for boundary value problems, namely renormgroup symmetries, is presented. Different methods of calculating these symmetries based on modern group analysis are described. An application of the approach to boundary value problems is demonstrated with the help of a simple mathematical model. © *1998 American Institute of Physics.* [S0022-2488(98)03901-2]

I. INTRODUCTION

The paper is devoted to the problem of constructing a special class of symmetries for boundary value problems (BVPs) in mathematical physics, namely *renormalization group symmetries* (hereafter referred to as RG-symmetries).

Symmetries of this type appeared about forty years ago in the context of the renormalization group (RG) concept. This concept originally arose[1–4] in the ''depth'' of quantum field theory (QFT) and was connected with a complicated procedure of renormalization, that is ''removing of ultra-violet infinities.'' In QFT, the renormalization group was based upon finite Dyson transformations and appeared as a continuous group in a usual mathematical sense. It was successfully used for improving an approximate perturbation solution to restore a correct structure of a solution singularity.

In the seventies, it was found that the RG concept was fruitful in some other fields of microscopic physics: phase transitions in large statistical systems, polymers, turbulence, and so on. However, in some cases, following Wilson's approach[5] to spin lattice, the original *exact symmetry* underlying the renormalization group notion in QFT was changed to an approximate one with the corresponding transformations forming a semi-group (not a group as in the QFT case). Here, in this paper, by RG-symmetry we mean the original exact property of a solution—as it was formulated in Refs. 3, 4 (see also Ref. 6) by Bogoliubov and one of the present authors. Thus, by RG-symmetry we mean a symmetry that characterizes a solution of a BVP and corresponds to *transformations* involving both ''dynamical'' (i.e., equation) *variables* and *parameters* entering into a solution via equations and boundary conditions.

For a simple illustration we consider some BVP that produces a family of solutions. The simplest variant of RG transformation is given by a simultaneous one-parameter point transformation,

$$T_a:\{x\to x'=x/a,\ \ g\to g'=G(a,g)\},\quad G(1,g)=g, \tag{1}$$

of a dimensionless ''coordinate'' x and a one-argument ''characteristic'' (e.g., initial value) g of each solution, the quantity of a direct physical interest. The transformation function $G(x,g)$, which depends upon two arguments,[7] should satisfy the functional equation,

$$G(x,g)=G(x/a,G(a,g)), \tag{2}$$

that guarantees the group property $T_a\cdot T_b=T_{ab}$ fulfillment.

 © 1998 American Institute of Physics

The functional equation (2) and transformation (1) arise, for example, in the massless QFT with one coupling. In that case $x=Q^2/\mu^2$ is the ratio of a 4-momentum Q squared to a ''reference momentum'' μ squared, g is the coupling constant, and G is the so-called *effective coupling*.

Later on the (1)–(2)-type symmetry underlying the renormgroup invariance was also found in a number of problems of macroscopic physics like, e.g., mechanics, transfer theory, hydrodynamics, and a close relation of RG-symmetry to the notion of self-similarity was established.[9,10]

The infinitesimal form of transformation (1) can be written down as a differential equation,

$$R\ G=0\,,\quad \text{with}\quad R=x\partial_x-\beta(g)\partial_g\,,\quad \beta(g)=\left.\frac{\partial G(a;g)}{\partial a}\right|_{a=1}\,, \tag{3}$$

where R is the infinitesimal operator of RG-symmetry (or, simply, *RG-operator*) with the coordinate $\beta(g)$ defined by the derivative of the function G.

Therefore, instead of relations (1) and (2), the RG transformation can also be introduced by means of an RG-operator. And vice versa, being given an RG-operator one can reconstruct the functional equation for the solution G with the help of the characteristic equation for (3). Moreover, for a given β-function or, in other words, for the given RG-symmetry, one can get an explicit expression for the invariant of the group transformation $G(x,g)$ by solving the corresponding Lie equations,[11]

$$-\frac{dx'}{x'}=\frac{dg'}{\beta(g')}=\frac{da}{a}, \tag{4}$$

with the initial conditions $x'|_{a=1}=x,\ \ g'|_{a=1}=g$.

Along with (3) a different form of the *invariance condition* for the function $G(x,g)$ is often employed,

$$\frac{d}{da}\,G(x/a,G(a,g))|_{a=1}=0\,. \tag{5}$$

Equation (3), reflecting the invariance of G under the RG transformation can be treated as a vanishing condition for the coordinate æ of the RG-operator (3) in the canonical form,[12]

$$\overline{R=}\ \text{æ}\,\partial_G\,,\qquad \text{æ}\equiv xG_x-\beta(g)G_g=0\,, \tag{6}$$

identically valid on a particular BVP solution $G=G(x,g)$.

At the same time, the relation

$$RS(x,g)\equiv(x\partial_x-\beta(g)\partial_g)S(x,g)=\gamma(g)S(x,g) \tag{7}$$

corresponds to the function $S(x,g)$ that is a *covariant*[13] of the RG transformation. In the QFT case, this relates, e.g., to a propagator amplitude (see Refs. 4 and 6). Here, $\gamma(g)$ is known as the *anomalous dimension* of $S(x,g)$.

Generally, the differential equation akin to (3),

$$xf_x-\beta(g)f_g=0, \tag{8}$$

states an invariance of a function f under the RG transformation (1). Its solution $f(x,g)=F(G(x,g))$ precisely corresponds to the same property emphasized by (2).

In a particular case, when the function G is linear in the last argument, $G\sim kg$, equation (2) defines a solution that has a power x dependence, i.e., $G(x,g)=gx^k$ with k being an arbitrary number. Then, equation (2) takes a form of power scaling (or *power self-similarity*) transformation,

$$x'=x/a,\qquad g'=ga^k,$$

that is well-known in mathematical physics and widely used in the problems of hydrodynamics of liquids and gases.

Therefore, transformation (2) can be considered[10] as a functional generalization $gx^k \to G(x,g)$ of a usual (i.e., power) self-similarity transformation. One can refer to it as to *functional self-similarity transformation*: this term was first introduced in Ref. 9 as a synonym of the RG transformation as defined above.

It is widely known that in QFT, as well as in other mentioned fields of theoretical physics, the RG method allows one to improve the perturbation theory results and to simplify the analysis of a singular behavior of a solution which becomes scale-invariant in the vicinity of a singularity. The latter reminds a situation, which is typical of asymptotic analysis of solutions of differential equations (DEs): long-time asymptotics demonstrate self-similar regimes.[14] Hence, it looks natural to use the RG methods to study strong nonlinear regimes and to investigate asymptotic behavior of physical systems described by DEs. We have no possibility to discuss this in detail here and would limit ourselves to mentioning some successful attempts of using the RG ideas in mathematical physics.

To our knowledge, the very first results in this field were obtained about a decade ago by two of the co-authors[15] of this paper by applying RG ideas to a problem of generating of higher harmonics in plasma. This problem, after some simplification, was reduced to a couple of partial DEs with the boundary parameter—solution ''characteristic''—explicitly included. It was proved that these DEs admit an exact symmetry group similar to that, defined by Eq. (3). The group obtained was then utilized to construct the desired nonlinear solution of the BVP. This approach has further been developed (see Ref. 16 and references therein) and we shall describe it in some detail in the next Section.

The methods of QFT RG were exploited by Goldenfeld, Martin, and Oono with co-authors (Urbana group) (see, e.g., Refs. 17 and 18) to find asymptotics of the solutions of parabolic-type nonlinear differential equations, that describe a variety of physical phenomena, such as ground-water flow under gravity, shock waves dynamics, radiative heat transfer, and so on. As an auxiliary tool, they used the concept of *intermediate asymptotics* first introduced by Barenblatt and Zeldovich[19]—see also the review monographs, Refs. 14 and 20. In this way, the Urbana team was able to determine values of exponents in the ratios of invariants forming arguments of self-similar solutions.

Later on, with the goal to a global asymptotic analysis they developed and illustrated, by numerous examples, the ''perturbative renormalization group theory'' (see Ref. 21 and references therein) that exploited the form of the invariance condition akin to that (5) used in QFT. The geometrical formulation of the perturbative RG theory for global analysis was presented by Kunihiro[22] on the basis of a classical theory of envelopes.

On the other hand, Bricmont and Kupiainen[23–25] attracted RG ideas for nonlinear DEs analyzing in a bit different manner. They used an iterative set of rescalings borrowed from the Wilson version of renormalization group, that is semi-group. On the basis of that RG-mapping procedure they succeeded in proving the global existence and detailed long time asymptotics for classes of nonlinear parabolic equations.

Generally, the procedure of revealing RG transformations, or some group features, similar to RG regularities, in any partial case (QFT, spin lattice, polymers, turbulence, and so on) up to now is not a regular one. In practice, it needs some imagination and atypical manipulation (see the discussion in Refs. 8, 26, 17) ''invented'' for every particular case. For example, the above described RG methods applied to an asymptotic analysis of differential equations were based on the *a priori* assumption of the existence of some scaling transformations or on the invariance condition of an approximate solution. By this reason, the possibility to find a regular approach to constructing RG-symmetries is of principal interest. In this paper we give an account of our efforts for creating a possible scheme of this kind in application to physical systems that are described by DEs. The leading idea in this case is based on the fact that symmetries of such systems can be found in a regular manner by using the well-developed methods of modern group analysis.

The paper is organized as follows: in Section II we describe the general scheme of constructing RG-symmetry for a BVP. It appears that the implementation of this scheme strongly depends on the mathematical model used and on the form of boundary conditions. As a result, different approaches to finding RG-symmetries are possible, and these are illustrated by examples in the

following five sections. In Section III, RG-symmetries are calculated using the classical Lie symmetries. In Section IV RG-symmetries are obtained on the basis of Lie–Bäcklund symmetries. In the next two sections RG-symmetries are found when boundary conditions are presented either in the form of a differential constraint (Section V), or in the form of an embedding equation (Section VI). In Section VII one more approach to RG-symmetries constructing is presented which is based on an approximate group symmetry. In conclusion, we make a summary of the approach and discuss some further applications.

II. APPROACH TO CONSTRUCTING RG-SYMMETRIES

First of all, we emphasize that the desired regular approach to constructing RG-symmetries turns out to be possible for those mathematical models of physical systems that are based on differential or, in some particular cases, integro-differential equations. The key idea uses the fact[27,28] that such models can be analyzed by algorithms of modern group analysis.

The proposed scheme comprises a sequence of the four steps.

(I) A specific manifold (differential, integro-differential, etc.) should be primarily constructed. This manifold that will be referred to as *renormgroup manifold* (RG-manifold) generally differs (see below) from the manifold given by the original system of DEs.

(II) The second step consists in calculating the most general symmetry group $\mathscr{G}$ admitted by the RG-manifold.

(III) The restriction of the group $\mathscr{G}$ on the desired BVP solution (exact or approximate) constitutes the next step. The group of transformations thus obtained (renormgroup) is characterized by a set of infinitesimal operators (RG-operators), each containing the solution of a BVP in its invariant manifold.

(IV) The last, fourth step implies utilization of RG-operators to find analytical expressions for solutions of the BVP.

Being formulated in a concise form these steps deserve further comments.

Comment to I. In the scheme described above the first step, constructing the RG-manifold is of fundamental importance. The form of its realization depends both on a mathematical model and on a form of a boundary condition.

Here we show the following different approaches to RG-manifold constructing.

(a) In the first more simple case, the RG-manifold, as is usual in classical group analysis, is presented by a system of basic DEs with only one substantial difference: parameters, entering into a solution via the equation and boundary conditions are included in the list of independent variables.

(b) Another approach to constructing the RG-manifold implies an extension of a space of variables involved in group transformations, for example, by including differential variables of higher order and nonlocal variables. It means that in this case Lie–Bäcklund transformation groups and nonlocal transformation groups should be invoked.[12,29]

(c) In the third approach, the procedure of construction of the RG-manifold is based on the invariant embedding method.[30] Here the RG-manifold is given by a system of equations that consists of original DEs and/or embedding equations which correspond to the BVP under consideration.

(d) The fourth approach, to some extent, is similar to the previous one. In this event boundary conditions are reformulated in terms of a differential constraint which is then combined with original equations to form the desired RG-manifold.

(e) The last approach utilizes approximate transformation groups. Here, the RG-manifold is given by a system of DEs with small parameters and can be analyzed by perturbation methods.[31]

Comment to II. Searching the symmetry of the RG-manifold is the main problem of the second step. The term ''symmetry'' as used in the classical group analysis means the property of a system of DEs to admit a Lie group of point transformations in the basic space of all independent and dependent (differential) variables entering these DEs. The Lie calculational algorithm of finding such symmetries is reduced to constructing tangent vector fields with coordinates that are functions of these basic group variables and can be defined from the solution of an overdetermined system of DEs, called *determining equations*. In modern group analysis different modification of a classical Lie scheme are in use (see, e.g., Refs. 29, 32, and references therein). If the problem of finding symmetries for a given system of DEs (RG-manifold) is solved, then the result is presented

in the form of Lie algebra of infinitesimal operators (also known as group generators), which correspond to the admitted vector field. In what follows these operators will be denoted by X.

Comment to III. The goal of a group restriction is the construction of a transformation group with a tangent vector field (point, Lie–Bäcklund, etc.) infinitesimal operators of which (hereinafter referred as R) contain the desired BVP solution in an invariant manifold. This means that the coordinate of the canonical operator of RG-symmetry vanishes on the BVP solution and on its differential consequences.

Mathematically, the procedure of a group restriction appears as a ''combining'' of different coordinates of group generators X admitted by the RG-manifold. The vanishing condition for this combination on a solution of the BVP leads to algebraic equalities that couple different coordinates and give rise to desired RG-symmetries. In a particular case, when RG is constructed from a Lie group admitted by the original system of DEs, it turns out to be a subgroup of this group and a solution of the BVP appears as an invariant solution with respect to the point RG obtained (compare with Ref. 11). In the general case, not only Lie point group, but Lie–Bäcklund groups, approximate groups, nonlocal transformation groups, etc. (see, e.g., Ref. 29), are also employed as basic groups which are then to be restricted on the solution of a BVP.

Comment to IV. A technique for constructing group invariant solutions corresponding to a symmetry group when its infinitesimal operators are known has been detailed in various monographs (see, e.g., Refs. 11, 32, 29). Therefore, the final step is performed in a usual way and needs no specific comments.

Before proceeding any further, we make a short review of results, that were obtained on the basis of the formulated scheme. The first application of the RG-approach to a particular problem of laser plasma was announced in Ref. 15. This problem, namely the problem of a nonlinear interaction of a powerful laser radiation with inhomogeneous plasma, has been detailed in subsequent publications.[16,33,34] A mathematical model was given by a system of nonlinear DEs for components of electron velocity, electron density, and the electric and magnetic fields. The presence of small parameters (such as weak inhomogeneity of the ion density, low electron thermal pressure, and small angles of incidence of a laser beam on plasma surface) in the initial system of equations provided a way for constructing the RG-manifold using the (e) approach, based on approximate group methods. The desired RG-symmetry appears as Lie point symmetry that takes account transformations of a boundary parameter [common to the (a) approach], which is related to the amplitude of the magnetic field at a critical density point. RG-symmetry obtained made it possible to get the exact solution of original equations, that was then used to evaluate the efficiency of harmonics generation in cold and hot plasma (see Ref. 16).

The advantageous use of the RG-approach in solving the above particular problem gave the promise that it may work in other cases. This was illustrated in Ref. 35 by a series of examples for different BVPs. Various methods of constructing RG-symmetries were described, based on the use of point symmetries, approximate symmetries, embedding equations, and transformations of Fourier components. As is shown in Ref. 35, different formulations of BVPs give rise to various methods of finding RG-symmetries. Thus, further development of the scheme was concentrated on analyzing these methods.

The first one was concerned with the initial value problem for the modified Burgers equation with parameters of nonlinearity and dissipation included explicitly. This example[36] yielded a detailed illustration of the method of constructing RG-symmetries when a basic RG-manifold is given by an original DE with parameters included in the list of independent variables [the (a) approach]. It was argued that the exact solution can be reconstructed from the perturbative solution with the help of any of the admitted RG-symmetry operators which form an eight-dimensional algebra. Two illustrative examples were given, dealing with perturbation theory in time and in nonlinearity parameter.

To demonstrate the method of constructing Lie–Bäcklund RG-symmetries that uses the (Ib) approach, the initial value problem for a linear parabolic equation was considered in Ref. 27. It was shown that appending Lie–Bäcklund RG-symmetries to point RG-symmetries extends the algebra of RG-symmetries up to an arbitrary order.

The same mathematical model was also employed within the (c) approach when the boundary condition is described by a differential constraint.[37] It was found that RG-symmetries obtained can not be reduced to point RG-symmetries which arise from the (a) case. However, some of them can

be reformulated in terms of RG-symmetries previously found in the (a) approach while the others can be constructed from the Lie–Bäcklund symmetries of basic equations in view of the given differential constraint.

An idea of the (d) approach to constructing the RG-manifold based on the invariant embedding method was realized in Ref. 35 for ordinary DEs. Here, embedding equations can be treated as a specific form of a differential constraint, that takes boundary data into account. The method of finding RG-symmetry using the (d) approach proves to be of particular interest for the first order ordinary DEs when using the (a) approach faces standard problems in calculating point symmetries admitted by the RG-manifold. Provided the embedding equation has the form of a first order DE of an evolutionary type there appear no difficulties in the group analysis of a joint system of the basic and the embedding equations.

Worthy of mention is an example that demonstrates the utilization of RG-symmetries to constructing solutions of the BVP for a system of two first order partial DEs that describes the propagation of a laser beam in a nonlinear focusing medium.[38–43] It was revealed that RG-symmetries are related to formal symmetries that are constructed in the form of an infinite series in the medium nonlinearity parameter. For a specific form of boundary data infinite series are truncated with RG-symmetries presented by finite sums. Generally, for arbitrary boundary data this is not the case and in that event a finite sum describes approximate RG-symmetry for a small nonlinearity parameter. Based on the (a), (b), and (e) approaches both point and Lie–Bäcklund (exact and approximate) RG-symmetries were obtained and then used to find an analytical solution of the problem.

To clarify the idea of constructing RG-symmetries several examples are given below which demonstrate different approaches to the problem. To gain better understanding of these approaches, a simple mathematical model is used.[44] This model corresponds to BVP for a system of two first-order partial DEs that were studied by Chaplygin[45] in gas dynamics,

$$v_t+vv_x-a\varphi(n)n_x=0\,,\quad n_t+vn_x+nv_x=0\,;$$
$$v(0,x)=V(x)\,,\quad n(0,x)=N(x)\,, \tag{9}$$

where $\varphi(n)$ is an arbitrary function of n and a is a nonlinearity parameter. Despite its simplicity, this mathematical model has a wide field of application and was used to describe various physical phenomena (in the so-called quasi-gaseous media[46]). In such a case the physical meaning of variables t, x, v, and n may differ from that in gas dynamics. For example, in nonlinear geometrical optics, t and x are coordinates, respectively, along and transverse to the direction of laser beam propagation, v is the derivative of eikonal with respect to x, and n is a laser beam intensity. In this case, functions V and N characterize the curvature of the wave front and the beam intensity distribution upon the coordinate x and the entrance of a medium $t=0$.

Along with (9) another form of basic equations will be used,

$$a\tau_v-(n/\varphi(n))\chi_n=0\,,\quad \chi_v+\tau_n=0\,. \tag{10}$$

These linear equations for new variables $\tau=nt$ and $\chi=x-vt$ results from (9) under hodograph transformations.

III. RG AS LIE POINT SUBGROUP

In this section we present an illustration of the method of constructing RG-symmetries when a basic RG-manifold is given by the original DEs with parameters included in the list of independent variables. Boundary conditions are taken into account while restricting the group admitted by RG-manifolds up to the desired RG on the exact or approximate solution of a BVP which thus appears as an invariant solution with respect to any of RG operators obtained.

3.1. First we shall consider a particular case of equations (9) when the nonlinearity parameter a is equal to zero: in the application to optical equations discussed above this means that nonlinear effects are neglected. Then by introducing a new variable $v=\varepsilon u$, the system of equations (9) is rewritten in the following form:

$$u_t+\varepsilon u u_x=0\,,\quad n_t+\varepsilon u n_x+\varepsilon n u_x=0\,; \tag{11}$$

$$u(0,x)=U(x)\,,\quad n(0,x)=N(x)\,. \tag{12}$$

The continuous point Lie group admitted by the differential manifold (11) (RG-manifold) is given by the infinitesimal operator (a general element of Lie algebra) with six independent terms:

$$X=\xi^1\partial_t+\xi^2\partial_x+\xi^3\partial_\varepsilon+\eta^1\partial_u+\eta^2\partial_n\equiv\sum_{i=1}^{6}X_i\,, \tag{13}$$

$$X_1=(1/\varepsilon)\ \Delta J^1\ \partial_t+(J^1+u\ \Delta J^1)\partial_x-nJ^1_\chi\partial_n\,,\quad \Delta J^k\equiv(\varepsilon t J^k_\chi-J^k_u)\,,$$

$$X_2=(1/n)\ J^2(\partial_t+\varepsilon u\partial_x)\,,\quad X_3=nJ^3\partial_n\,,\quad X_4=J^4(-t\partial_t+n\partial_n+\varepsilon\partial_\varepsilon)\,,$$

$$X_5=\Delta J^5 D+J^5(\varepsilon t\partial_x+\partial_u)\,,\quad X_6=-(1/n)\ J^6 D\,,\quad D\equiv(t\partial_t+\varepsilon u t\partial_x-n\partial_n)\,.$$

Coordinates ξ and η of this infinite-dimensional group operator depend upon five functions $J^i(\chi,u,\varepsilon)$, $i=1,2,3,5,6$ which appear as arbitrary functions of their arguments $\chi=x-vt$, u and ε. The sixth one, J^4, that enters the operator which describes the group transformation of parameter ε, is an arbitrary function of this parameter only. The restriction of the group admitted by RG-manifold (11) on the solution of the BVP $u=\overline{u}(t,x,\varepsilon)$, $n=\overline{n}(t,x,\varepsilon)$ leads to zero equalities for two coordinates of the operator (13) in the canonical form—the conditions of functional self-similarity:

$$\eta^1+\xi^1\varepsilon\,\overline{u\,u_x}-\xi^2\overline{u_x}-\xi^3\overline{u_\varepsilon}=0\,,\quad \eta^2+\xi^1\varepsilon(\overline{n\,u})_x-\xi^2\overline{n_x}-\xi^3\overline{n_\varepsilon}=0\,. \tag{14}$$

These equalities should be valid for any values of t, and certainly for $t=0$, when dependencies $\overline{u}$ and $\overline{n}$ upon x are given by boundary conditions (12). This yields two linear relations between J^i and J^1_χ:

$$J^5=U_x J^1\,,\quad J^6=N_x J^1+NJ^1_\chi-N(U_x)_u J^1-\varepsilon U_x J^2-NJ^3-NJ^4\,. \tag{15}$$

Here, and in what follows functions U and N and their derivatives with respect to x should be expressed either in terms of u or in terms of χ. Substituting (15) in (13) gives the desired RG-symmetries with the RG-operator,

$$R=\sum_{i=1}^{4}R_i\,, \tag{16}$$

$$R_1=X_1+\left[\left(\varepsilon t(U_x)_\chi-\left(1-\frac{N}{n}\right)(U_x)_u-\frac{N_x}{n}\right)J^1\right.$$

$$\left.+\left(\varepsilon tU_x-\frac{N}{n}\right)J^1_\chi-U_xJ^1_u\right]D+U_xJ^1(\varepsilon t\partial_x+\partial_u)\,,$$

$$R_2=X_2+\frac{\varepsilon U_x}{n}J^2D,\quad R_k=X_k+\frac{N}{n}J^kD\,,\quad k=3,4\,.$$

We see that RG-symmetries for (11), (12) are presented as a combination of symmetries of infinite-dimensional algebra with the infinitesimal operator (13). Any of the four operators R_k (and their linear combinations with coefficients that are arbitrary functions of ε) contains the BVP solution $u=\overline{u}(t,x,\varepsilon)$ and $n=\overline{n}(t,x,\varepsilon)$ in the invariant manifold and enables us to obtain group transformation of both group variables and different functionals of the solution (for a method of calculating transformation of a functional see Ref. 47).

Generally, the renormalization group using is capable of improving a perturbation theory solution. As an example, consider the perturbative solution of (11), (12) for small value of $\varepsilon t \ll 1$,

$$u = U(x) - (\varepsilon t) U U_x + O(\varepsilon^2 t^2) \ , \quad n = N(x) - (\varepsilon t)(U N_x + N U_x) + O(\varepsilon^2 t^2) \ . \tag{17}$$

This approximate solution in the limit $(\varepsilon t) \to 0$ is invariant with respect to RG transformation defined by the operator R_2 with arbitrary $\varepsilon J^2 \neq 0$. Assuming $J^2 = 1/\varepsilon$, we obtain the explicit expression for the RG-operator,

$$R = \frac{1}{n}[(1 + \varepsilon t U_x)(\partial_t + \varepsilon u \partial_x) - \varepsilon U_x n \partial_n] \ , \tag{18}$$

and invariance conditions written in the form of two first order DEs:

$$u_t + \varepsilon u u_x = 0 \ , \ (1 + \varepsilon t U_x)(n_t + \varepsilon u n_x) + \varepsilon n U_x = 0 \ . \tag{19}$$

Solving Lie equations which correspond to the RG-operator (18) [and coincide with characteristics equations for (19)] enables us to reconstruct the desired exact solution of (11), (12) from the perturbative solution (17),

$$u = U(x - \varepsilon u t) \ , \quad n = \frac{1}{1 + \varepsilon t U_x} N(x - \varepsilon t u) \ , \tag{20}$$

where U_x should be expressed in terms of u. For example, in the particular case of $N(x) = N_0 \exp(-x^2)$, $U(x) = -x$ and $\varepsilon = 1/T$ the latter formulas describe the focusing of a Gaussian laser beam in geometrical optics,

$$n = \frac{T}{T - t} N_0 \exp\left(-x^2 \left(\frac{T}{t - T} \right)^2 \right) \ , \quad u = x \frac{T}{t - T} \ , \quad t \leqslant T. \tag{21}$$

3.2. Now let us turn to a more general case of $a \neq 0$. The Lie point symmetry group, admitted by the RG-manifold (10), is characterized by a canonical infinitesimal operator[12] with six independent terms X_i, $i = 1, \ldots, 5$ and X_∞,

$$X = X_\infty + \sum_{i=1}^{5} c_i X_i \equiv \left(\overline{f} + \sum_{i=1}^{5} c_i f_i \right) \partial_\tau + \left(\overline{g} + \sum_{i=1}^{5} c_i g_i \right) \partial_\chi , \tag{22}$$

where coordinates f_i and g_i are linear combinations of τ and χ and their first derivatives $\tau_1 = (\partial \tau / \partial n)$ and $\chi_1 = (\partial \chi / \partial n)$ with coefficients depending only on v and n.[40,41] For a particular case $\varphi = 1$ they are

$$\begin{aligned}
&f_1 = \tau, \quad g_1 = \chi; \quad f_2 = -(1/a) n \chi_1, \quad g_2 = \tau_1; \\
&f_3 = -\tau/2 + n \tau_1 + (1/2a) n v \chi_1, \quad g_3 = -(v/2) \tau_1 + n \chi_1; \\
&f_4 = -(1/2) n \chi + v n \tau_1 + [(1/4a) v^2 - n] n \chi_1, \\
&g_4 = (a/2) \tau + (1/2) v \chi + v n \chi_1 + [a n - (1/4) v^2] \tau_1; \\
&f_5 = (n \tau_1 - \tau) - a \tau_a, \quad g_5 = n \chi_1 - a \chi_a .
\end{aligned} \tag{23}$$

''Evident'' symmetries f_1, g_1 and f_2, g_2 describe dilations of τ and χ and translations along the v-axis, respectively, for an arbitrary nonlinearity $\varphi(n)$. Two more symmetries f_3, g_3 and f_4, g_4 appear due to a special form of the function $\varphi = 1$ under consideration. The symmetry f_5, g_5 involves the parameter a transformation along with transformations of dynamic variables.

The operator X_∞ with coordinates $\overline{f}=\xi^1(v,n), \overline{g}=\xi^2(v,n)$ that are arbitrary solutions of partial DEs,

$$\xi^1_v-(n/a)\xi^2_n=0, \quad \xi^2_v+\xi^1_n=0\,, \tag{24}$$

results from the linearity of basic Equations (10); it is an ideal of an infinite-dimensional Lie algebra L_∞ formed by operators $X_1,\ldots,X_5$ and X_∞.

The restriction of the group (23) on the BVP solution means that coordinates f and g of the canonical operator (22) turns to zero on this solution, that is,

$$\overline{f}=-\sum_{i=1}^{5} c_i f_i\,, \quad \overline{g}=-\sum_{i=1}^{5} c_i g_i\,. \tag{25}$$

These relations express functions $\overline{f}$ and $\overline{g}$ in terms of $f_i,\ g_i,\ i=1,\ldots,5$ taken on a solution $\tau=\overline{\tau}(v,n),\ \chi=\overline{\chi}(v,n)$ of a BVP (exact or approximate). Substitution of (25) in (22) gives five RG-operators,

$$R=\sum_{i=1}^{5} c_i(a)R_i\,, \tag{26}$$

each being determined by corresponding coordinates $f_i,\ g_i$ and by a pair of functions $A^i,\ B^i$,

$$\begin{aligned}
&R_1=(\tau-A^1)\partial_\tau+(\chi-B^1)\partial_\chi, \quad R_2=-A^2\partial_\tau-B^2\partial_\chi+\partial_v\,,\\
&R_3=(-(\tau/2)-A^3)\partial_\tau-B^3\partial_\chi-(v/2)\partial_v-n\partial_n\,,\\
&R_4=(-(n/2)\chi-A^4)\partial_\tau+((a/2)\tau+(v/2)\chi-B^4)\partial_\chi\\
&\qquad +(-(1/4)v^2+an)\partial_v+vn\partial_n\,,\\
&R_5=(-\tau-A^5)\partial_\tau-B^5\partial_\chi-n\partial_n+a\partial_a.
\end{aligned} \tag{27}$$

Here, ten functions A^i,B^i are defined by expressions (23) for f^i and g^i where one should replace $\tau,\ \chi$ by $\overline{\tau}(n,v),\ \overline{\chi}(n,v)$. Explicit formulas for RG-operators depend upon the specific solution of the BVP. For example, for the particular solution of the BVP (9) with $V=0$ and $N(x)=\cosh^{-2}(x)$, described by[48]

$$\begin{aligned}
&\tau=\frac{(v/2)^{1/2}}{a^{3/4}}(\sqrt{\kappa^2+1}-\kappa)^{1/2}\,, \quad \kappa=\frac{\sqrt{a}}{v}\left(1-n-\frac{v^2}{4a}\right),\\
&\chi=-\frac{1}{2}\ln\frac{(v/2\sqrt{a})^{1/2}+(\sqrt{\kappa^2+1}-\kappa)^{1/2}}{-(v/2\sqrt{a})^{1/2}+(\sqrt{\kappa^2+1}-\kappa)^{1/2}}\,,
\end{aligned} \tag{28}$$

functions $A^5,\ B^5$ in (27) are expressed as follows:[40]

$$A^5=-\frac{(v/2)^{1/2}}{4a^{3/4}\sqrt{1+\kappa^2}}(\sqrt{\kappa^2+1}-\kappa)^{1/2}\left(\kappa+\sqrt{1+\kappa^2}-2\frac{\sqrt{a}}{v}\right)\,,$$

$$B^5=\frac{(v/2)^{1/2}}{4a^{1/4}\sqrt{1+\kappa^2}}\,\frac{(\sqrt{\kappa^2+1}-\kappa)^{1/2}}{(\sqrt{\kappa^2+1}-\kappa-(v/2\sqrt{a}))}\left(\sqrt{1+\kappa^2}-3\kappa+2\frac{\sqrt{a}}{v}-\frac{v}{\sqrt{a}}\right)\,.$$

It should be noticed that the solution of the presented above BVP is unique, but a number of RG-operators that give rise to this solution is different from one [in the first example case we have four RG-operators with arbitrary functions of $(n,\ \chi)$, and in the second example five RG-operators with arbitrary functions of a]. In the next section we will show that the number of RG-operators may be enlarged to an arbitrary value, provided not only point but Lie–Bäcklund groups are taken into account.

IV. RG AS LIE–BÄCKLUND SUBGROUP

The method of constructing RG-symmetries from Lie point symmetries admitted by the original DE is naturally generalized to include Lie–Bäcklund (L–B) symmetries. The extension of the space of differential variables increases the amount of BVPs that allow restriction of a group on their solution. A complete set of RG-symmetries is obtained by appending L–B RG-symmetries to point RG-symmetries. In this section we present an example of constructing L–B RG-symmetries of the second order for the BVP (9). As in the previous section we use a transformed form of the basic equations (10).

L–B symmetries admitted by the RG-manifold (10) are characterized by the same canonical infinitesimal operator (22) where additional terms proportional to higher-order derivatives of τ and χ should be added in coordinates f and g. Similar to first-order symmetries, these terms are linear combinations of τ and χ and their derivatives $\tau_i=(\partial^i\tau/\partial n^i)$ and $\chi_i=(\partial^i\chi/\partial n^i)$ with coefficients that depend only on v and n.[39–41] For the second order Lie–Bäcklund symmetries in a particular case $\varphi(n)=1$, we have five additional operators X_i with $i=7,\ldots,11$ (the term with $i=6$ corresponds to X_∞ and is omitted in the sum, i.e., $c_6=0$),

$$X=X_\infty+\sum_{i=1}^{11} c_iX_i\equiv\left(\overline{f}+\sum_{i=1}^{11} c_if_i\right)\partial_\tau+\left(\overline{g}+\sum_{i=1}^{11} c_ig_i\right)\partial_\chi. \tag{29}$$

It should be noted that expressions for all coordinates in (29) can be obtained by the action of the following three recursive operators[41] L_i, $i=1,2,3$,

$$L_1=\begin{pmatrix} 0 & -(n/a)D_n \\ D_n & 0 \end{pmatrix},\quad L_2=\begin{pmatrix} 2nD_n-1 & (n/a)vD_n \\ -vD_n & 2nD_n \end{pmatrix},$$
$$L_3=\begin{pmatrix} 2nvD_n & n(v^2/2a-2n)D_n-n \\ (-v^2/2+2an)D_n+a & 2nvD_n+v \end{pmatrix}, \tag{30}$$

on the ''trivial'' operator with $f=\tau$ and $g=\chi$ (here, D_n is the operator of total differentiation with respect to n). Below, we present only three of these five second-order L–B operators,

$$f_7=n\tau_2,\ g_7=\chi_1+n\chi_2;$$
$$f_8=(1/2a)n[-\chi_1+v\tau_2-2n\chi_2],\ g_8=(1/2a)v\chi_1+n\tau_2+\frac{1}{2a}nv\chi_2; \tag{31}$$
$$f_9=(1/4)\tau-n\tau_1-(5/4a)vn\chi_1+(-n+(1/4a)v^2)n\tau_2-(1/a)vn^2\chi_2,$$
$$g_9=(3/4)v\tau_1-(2n-(1/4a)v^2)\chi_1+vn\tau_2+(-n+(1/4a)v^2)n\chi_2.$$

The procedure of restriction of the L–B group obtained on the solution of the BVP leads to expressions for $\overline{f}$ and $\overline{g}$ akin to (25),

$$\overline{f} = -\sum_{i=1}^{11} c_i f_i, \quad \overline{g} = -\sum_{i=1}^{11} c_i g_i. \tag{32}$$

Substitution of (32) in (29) yields additional terms in the expression (26) for the RG-operator R that depends on higher-order derivatives of τ and χ,

$$R = \sum_{i=1}^{11} c_i(a) R_i \equiv \sum_{i=1}^{11} c_i(a)((f_i - A^i)\partial_\tau + (g_i - B^i)\partial_\chi) \ . \tag{33}$$

Here functions A^i and B^i are given by the corresponding formulas for coordinates f_i and g_i to be evaluated on the solution $\overline{\tau}(n,v)$ and $\overline{\chi}(n,v)$. It appears that coordinates of L–B RG-operators are obtained from point RG-operators with the help of the above-mentioned recursive operators, hence, one can obtain L–B RG-operators of an arbitrary high order. Despite an unusual form, we still call them RG-operators since they possess the main property of RG-operators, namely, they contain a solution of the BVP in their invariant manifold.

The procedure of using L–B RG-operators is not as simple as for point RG-operators. Yet we can describe two possible ways.

Firstly, coordinates of canonical L–B RG-operators can be used to construct a set of relations, differential constraints, that are compatible with the original DEs and satisfy specific boundary conditions. The use of such constraints is described in Section VI. In the general case, for an arbitrary L–B group of a given order, coordinates of the corresponding canonical operator can be treated as a set of differential expressions, zero equalities for which impose appropriate restrictions on the basic DEs, consistent either with physical or with symmetry conditions. These equalities can also be treated as embedding equations (see Ref. 27).

Secondly, L–B RG-operators can be used to construct invariant solutions that automatically fit boundary conditions. It should be noticed that in some particular cases, L–B RG-symmetries can be constructed from a L–B group with a finite number of operators. For example, RG-symmetry for (9) with boundary conditions $V=0$ and $N=\cosh^{-2}(x)$ appear as a linear combination of three L–B symmetries,

$$R = (f_3 + 2(f_7 + f_9))\partial_\tau + (g_3 + 2(g_7 + g_9))\partial_\chi. \tag{34}$$

The desired solution of the BVP can be found as the invariant solution with respect to the RG-operator (34) and is presented by formulas (28).

The recipe of constructing the L–B renormgroup formulated in this section goes far beyond a simple illustrative example for the BVP (9). In a similar way, L–B RG-operators are constructed for different BVPs of mathematical physics that admit L–B symmetries; other examples are presented in Ref. 27 for the linear parabolic and modified Burgers equation. It is essential that when parameters entering into the equation and boundary conditions are involved in group transformations, coordinates of canonical L–B RG-operators contain not only first but higher-order derivatives with respect to these parameters. This means that in addition to recursive operators containing operators of total differentiation with respect to n [for BVP (9)], new recursive operators comprise operators of differentiation with respect to parameters, as well [$\propto D_\epsilon$ and D_a in the case of BVPs (11) and (10)].

V. RG DEVISING BASED ON EMBEDDING EQUATIONS

In this section we present a specific method of constructing RG-symmetries,[35,27] which is based on embedding equations.[30] It is of prime interest for physical systems described by ordinary differential equations (ODEs). In the context of the discussed model of quasi-Chaplygin media such equations arise, e.g., when constructing invariant solutions with respect to symmetries obtained. We demonstrate the idea of this method for the very simple BVP,

$$u_t = f(t,u,a) \ ; \quad t=\tau \ , \quad u=x \ . \tag{35}$$

Extension of the original differential manifold by adding, to the original equation, the embedding equation that appears as a linear first-order partial DE,

$$u_\tau + f(\tau,x,a)u_x = 0\,, \tag{36}$$

gives the desired RG-manifold, where u is now treated as the function of four variables $\{t,\ \tau,\ x,\ a\}$. Performing the group analysis for this RG-manifold involves boundary data and parameter a in group transformations, while the subsequent restriction of the group obtained on any solution of the BVP yields the desired RG-symmetries. We give two examples of such calculations for $f=au^2$ and $f=u^2+au^3$.

5.1. In the event of $f=au^2$ the RG-manifold (35)–(36) is given by two equations:

$$u_t = au^2\,, \quad u_\tau + ax^2u_x = 0, \tag{37}$$

that admit an infinite-dimensional Lie point algebra with five independent elements,

$$X=\sum_{i=1}^{5}\alpha_i X_i\,,$$

$$X_1=\partial_t+au^2\partial_u\,, \quad X_2=\partial_\tau+ax^2\partial_x\,, \tag{38}$$

$$X_3=u^2\partial_u\,, \quad X_4=x^2\partial_x\,, \quad X_5=x^2\tau\partial_x+u^2t\partial_u+\partial_a\,.$$

Here, functions α_1 and α_2 depend upon five variables $\{t,\ \tau,\ x,\ a,\ u\}$, whereas α_i, $i=3,4,5$ are arbitrary functions of three combinations $at+(1/u)$, $a\tau+(1/x)$, a.

The procedure of restriction of the group obtained leads to the invariance condition

$$U^2(\alpha_3+a\alpha_1+\alpha_5 t)-\alpha_1U_t-\alpha_2U_\tau-x^2(\alpha_4+a\alpha_2+\alpha_5\tau)U_x-\alpha_5U_a=0, \tag{39}$$

to be fulfilled on an exact or approximate solution $u=U(t,x,\tau,a)$ of the BVP (35)–(36); for example, one can take the perturbative solution as an expansion in powers of a,

$$u=U(t,x,\tau,a)\equiv x+ax^2(t-\tau)+O(a^2), \quad a\ll 1\,. \tag{40}$$

Substituting (40) into (39) shows that the invariance condition (39) is fulfilled for $\alpha_3=\alpha_4\equiv\alpha$ and arbitrary α_1,α_2, and α_5. Assuming $\alpha_1=\alpha_2=\alpha=0$ and $\alpha_5=1$ in (38) yields one of the RG-operators,

$$R=x^2\tau\partial_x+\partial_a+u^2t\partial_u\,, \tag{41}$$

which enables us to transform the perturbative solution of (35) for small $a\ll 1$ to the following exact solution:

$$u=\frac{x}{1-ax(t-\tau)}\,.$$

This result is found by solving the Lie equations, that correspond to the RG-operator (41).

5.2. For another value of the function $f=u^2+au^3$, the RG-operator that is similar to (41) is given as follows:

$$R=(x^2(1+ax)\tau+x)\partial_x+(u^2(1+au)t+u)\partial_u-a\partial_a\,. \tag{42}$$

The invariance condition for the solution of the BVP with respect to the RG-operator (42) has the form of the first order partial DE,

$$-(x^2(1+ax)\tau+x)u_x+au_a+u^2(1+au)t+u=0\,. \tag{43}$$

Solving the characteristic equations for (43) (Lie equations) yields the following exact solution of the BVP (35) with $f=u^2+au^3$:

$$t-\tau=\frac{1}{x}-\frac{1}{u}+a\ln\left|\frac{x}{u}\frac{(1+au)}{(1+ax)}\right| .$$

What all renormgroups obtained for the BVPs for the first-order ODE in the above examples have in common is that their operators depend upon arbitrary functions α_i, which means that RG can be expressed in terms of different RG-operators with various particular expressions for their coordinates. This situation is the same as that obtained for the BVP in the case of partial DE: different RG-operators yield the same unique specific solution of the given BVP contained in the invariant manifold of RG-operators. The previous procedure of RG constructing for the BVP for the ODE was based on the use of point groups. However, L–B groups can also be employed for constructing RG-symmetries for the first-order ODE, especially, in view of embedding equations (see the references in Ref. 27).

The structure of embedding equations depends not only on the form of the original equation, but also on the boundary conditions. This means that for given basic equations we may obtain different embedding equations. For example, if the function f in the r.h.s. of (35) depends upon x,

$$u_t=f(t,x,a,u)\ ; \quad t=\tau\ , \ u=x, \tag{44}$$

we arrive at the embedding equation

$$u_\tau+f(\tau,x,a,x)u_x=f(\tau,x,a,x) + \int_\tau^t d\ t' f_x(t',x,a,u(t'))\exp\left[-\int_t^{t'} d\ t'' f_u(t'',x,a,u(t''))\right] . \tag{45}$$

Hence, the RG-manifold in this case is defined by a system of integro-differential equations (44) and (45) and one should employ the modern group analysis techniques which give a possibility of analyzing such equations, as well.[49,50]

VI. RG AND DIFFERENTIAL CONSTRAINT

In the previous section RG-manifold was obtained by combining an original DE and an embedding equation. More generally instead of an embedding equation, an additional differential constraint can be used that satisfy two conditions: firstly, it must be compatible with the original DE and, secondly, it should explicitly take boundary conditions into account. This constraint naturally emerges when a coordinate of a canonical operator of the L–B RG admitted by BVP is assumed to be equal to zero. Adding this constraint to original equations we obtain the RG-manifold.

6.1. To illustrate, consider first a BVP (9) with $a=0$ which we rewrite using hodograph transformations in a simple form [compare with (10)]

$$\chi_n=0\ , \quad \chi_v+\tau_n=0\ . \tag{46}$$

L–B symmetries of this system of DEs are given by a canonical operator,

$$X=f\partial_\tau+g\partial_\chi, \tag{47}$$

with coordinates f and g depending upon v and derivatives $\tau_s+n\chi_{s+1}$, χ_s of an arbitrary order $s\geqslant 0$,

$$f=F(v,\chi_s,\tilde{\tau}_s)-n\left[\partial_v+\sum_{k=0}^{\infty}(\tilde{\tau}_{k+1}\partial_{\tilde{\tau}_s}+\chi_{k+1}\partial_{\chi_k})\right]G, \quad g=G(v,\chi_s,\tilde{\tau}_s)\ , \tag{48}$$

$$\tilde{\tau}_s=\tau_s+n\chi_{s+1}, \quad \tau_s=(\partial^s\tau/\partial v^s)\ , \quad \chi_s=(\partial^s\chi/\partial v^s)\ .$$

Consider a particular case of a BVP (9) with boundary conditions defined by $V(x)=-\varepsilon x$ and arbitrary $N(x)$. In terms of the variables τ and χ, these conditions are described, for example, by a pair of differential constraints,

$$\chi_{vv}=0\,,\quad \tau_{vv}-N_{vv}\chi_v-N_v\chi_{vv}=0\,. \tag{49}$$

Here the dependence of N upon x is given in terms of v with the use of the above boundary condition.

It is easily checked by direct substituting into (48) that left-hand sides of these equalities are the corresponding coordinates g and f of the second-order L–B symmetry operator (47). Adding differential constraints (49) to the original equation (46), we obtain the desired RG-manifold,

$$\chi_n=0\,,\quad \chi_v+\tau_n=0\,,\quad \chi_{vv}=0\,,\quad \tau_{vv}-N_{vv}\chi_v=0\,. \tag{50}$$

The latter admits a 17-parameter group of point transformations given by the following operators:

$$X=\sum_{i=1}^{m} c_iX_i\,,\quad m=17\,, \tag{51}$$

$$X_1=v^2\partial_v+v(2(n-N)+vN_v)\partial_n+(\chi(N-n)+\tau v)\partial_\tau+v\chi\partial_\chi,$$

$$X_2=v\chi\partial_v+(\chi(n-N)+v(\chi N_v-\tau))\partial_n+2\tau\chi\partial_\tau+\chi^2\partial_\chi,$$

$$X_3=-v\partial_v+(N-n-vN_v)\partial_n,\quad X_4=v\chi\partial_n-\chi^2\partial_\tau,$$

$$X_5=v\partial_n,\quad X_6=(N-n)\partial_n+\chi\partial_\chi,\quad X_7=(n-N)\partial_n+\tau\partial_\tau,$$

$$X_8=\partial_\tau,\quad X_9=v\partial_\tau,\quad X_{10}=(N-n)\partial_\tau+v\partial_\chi,$$

$$X_{11}=\partial_\chi,\quad X_{12}=\chi\partial_\tau,\quad X_{13}=-v^2\partial_n+v\chi\partial_\tau.$$

$$X_{14}=-\partial_v-N_v\partial_n,\quad X_{15}=\partial_n,\quad X_{16}=\chi\partial_n,\quad X_{17}=-\chi\partial_v+(\tau-\chi N_v)\partial_n.$$

The usual procedure of restriction of the group obtained on a solution of the BVP (46) relates different coefficients in the sum (51) and gives the desired RG-operators,

$$R=\sum_{i=1}^{13} c_iR_i\,, \tag{52}$$

$$R_1=X_1,\quad R_2=X_2,\quad R_3=X_3+\varepsilon X_{17},\quad R_4=X_4,$$

$$R_5=X_5+\epsilon X_{16},\quad R_6=X_6+\varepsilon X_{17},\quad R_7=X_7+\varepsilon X_{16},$$

$$R_8=X_8+\varepsilon X_{15},\quad R_9=X_9-\varepsilon^2X_{16},\quad R_{10}=X_{10}-\varepsilon^2X_{17},$$

$$R_{11}=X_{11}+\varepsilon X_{14},\quad R_{12}=X_{12}+\varepsilon X_{16},\quad R_{13}=X_{13}.$$

The exact solution of the BVP $\chi=-v/\varepsilon$, $\tau=(1/\varepsilon)(n-N)$ is found either by solving Lie equations corresponding to any of these RG-operators, or as the intersection of all invariant manifolds.

6.2. Now let us turn to a general case of a BVP (9) with $a\neq 0$. We shall consider the problem of constructing RG-symmetries using the RG-manifold given by basic equations in the form (10)

and the most simple differential constraint yielded by the linear combination of the second-order L–B symmetry (31) f_7, g_7 and trivial infinite-dimensional symmetry $f_\infty=0$, $g_\infty=-1$,

$$a\tau_v-n\chi_n=0\,,\quad \chi_v+\tau_n=0\,,\quad n\tau_{nn}=0\,,\quad \chi_n+n\chi_{nn}-1=0\,. \tag{53}$$

This differential constraint describes, in particular, a linear dependence of N upon x and $V(x)=0$. The Lie point group admitted by the RG-manifold (53) is characterized by seven infinitesimal operators [use the formula (51) for $m=7$],

$$X_1=-2v\partial_v-4n\partial_n-6n(v/a)\partial_\tau+(2\chi-6n+3v^2/a)\partial_\chi,$$

$$X_2=v\partial_v+2n\partial_n+(\tau+2nv/a)\partial_\tau+(2n-v^2/a)\partial_\chi,\quad X_3=\partial_v\,,$$

$$X_4=n\partial_\tau-v\partial_\chi,\quad X_5=(v/a)\partial_\tau+\ln n\partial_\chi,\quad X_6=\partial_\tau,\quad X_7=\partial_\chi\,.$$

The restriction of this group on the solution of the BVP with the above-mentioned boundary conditions leads to the three-parameter RG,

$$R_1=X_1,\quad R_2=X_2,\quad R_3=aX_3+X_4.$$

As in the previous case, the exact solution of the BVP $\tau=nw$, $\chi=n-aw^2/2$ appears as an intersection of all invariant manifolds that correspond to these RG-operators.

The characteristic feature of the described approach is the formulation of boundary data in the form of a differential constraint and the subsequent search of the group admitted by this constraint and basic equations. It is evident that there exists an infinite number of other differential constraints that adequately describe the same boundary data and the use of which leads to different RG algebras. As an example, we can point to differential constraints that arise from the zero equality of appropriate coordinates of the infinite L–B algebra.

The example of RG-symmetry constructing on the basis of L–B symmetry reveals the practical importance of the latter and, on the other hand, demonstrates point symmetries that are not admitted by the original equation. The procedure of construction of RG-symmetries with the help of a differential constraint was also carried out in Ref. 37 for the linear parabolic equation.

VII. RG AS A SUBGROUP OF AN APPROXIMATE SYMMETRY GROUP

An attractive method of RG constructing is that based on approximate symmetries.[31] This method can be applied to systems described in terms of models based on DEs with small parameters. These small parameters allows us to consider a simple subsystem of the original DEs that usually admits an extended symmetry group inherited by the original DEs. Restricting this approximate group on the solution of the BVP yields the desired RG-symmetries. The merits of the described method is illustrated below for the BVP (9) with a small nonlinearity parameter $a\ll 1$. In terms of the variable $w=v/a$ the following basic system of linear DEs is obtained instead of (10):

$$\tau_w-(n/\varphi(n))\chi_n=0\,,\quad \chi_w+a\tau_n=0\,. \tag{54}$$

For $a=0$, it admits an infinite L–B symmetry group,

$$X=f\partial_\tau+g\partial_\chi, \tag{55}$$

characterized by an arbitrary dependence of the zeroth-order coordinates $f=f^0$ and $g=g^0$ upon n, τ, χ, and the derivatives $\widetilde{\tau}_s$, χ_s of an arbitrary order,

$$f^0=F^0+\int dw\{(n/\varphi)Yg^0\},\quad g^0=G^0\,. \tag{56}$$

Here and below in (57),

$$Y=\partial_n+\sum_{s=0}^{\infty}\left(\tau_{s+1}\partial_{\tau_s}+\chi_{s+1}\partial_{\chi_s}\right),$$

$$\tau_s=\frac{\partial^s\tau}{\partial n^s},\quad \chi_s=\frac{\partial^s\chi}{\partial n^s},\quad \tilde{\tau}_s=\tau_s-w\sum_{p=0}^{s}\binom{s}{p}\frac{\partial^p(n/\varphi)}{\partial n^p}\chi_{s-p+1},$$

$F^i(n,\chi_s,\tilde{\tau}_s)$ and $G^i(n,\chi_s,\tilde{\tau}_s)$ are arbitrary functions of their arguments, and expressions in curly brackets before integrating over w should be given in terms of $\tilde{\tau}_s$, χ_s, n, w.

For $0<a\ll 1$, this symmetry is inherited as an approximate one by equations (54) which thus represent an approximate RG-manifold. For example, for $\varphi(n)=1$ the following result is obtained:

$$f^i=F^i+\int dw\left\{Zf^{i-1}+nYg^i\right\},\quad g^i=G^i+\int dw\{Zg^{i-1}-Yf^{i-1}\},$$
$$Z=\sum_{s=0}^{\infty}\tau_{s+1}\partial_{\chi_s},\quad \tilde{\tau}_s=\tau_s-w(n\chi_{s+1}+s\chi_s),\quad i\geqslant 1. \tag{57}$$

One can see that the symmetry of equations (54) for $a=0$ is inherited by the symmetry of these equations for $a\neq 0$ up to an arbitrary order of this parameter. It should be noticed that both zeroth-order and higher-order approximate symmetries may appear as Lie point symmetries or L–B symmetries, and this parameter may be involved in group transformations as well.

The restriction of the approximate group obtained on a particular solution of the BVP defines the specific form of the zeroth-order symmetries. It means that while constructing RG-symmetries for the BVP (54) in view of the boundary data from (9), coordinates f^0, g^0 and ''integration constants'' F^i, G^i, $i\geqslant 1$ are not arbitrary functions, but should be chosen so that relations $f=0$, $g=0$ satisfy desired boundary conditions $\tau_s=0$, $\chi=H(n)$ at $w=0$. Provided that the functions F^i and G^i, $i\geqslant 1$ are also equal to zero in this case, boundary conditions are correlated with the form of functions f^0 and g^0. In general, invariance conditions $f=0$ and $g=0$ appear as differential constraints (or algebraic relations) to be fitted by boundary data.

Of special interest are such zeroth-order functions f^0 and g^0 for which infinite series (57) are truncated for some finite value of $i=i_{max}$, and we arrive at finite sums. In this case, instead of an approximate group with respect to a small parameter a we obtain the exact symmetry group (compare with Ref. 31, Sec. 11). A simple example of this is given by the RG-operator (34). It is easily checked that in terms of n and w, the combinations of coordinates $f_3+2(f_7+f_9)$ and $g_3+2(g_7+g_9)$ are expressed as a binomial in a, i.e., expressions for f and g are represented as zeroth-order and first-order terms $f=f^0+af^1$ and $g=g^0+ag^1$, where f^0, g^0 and f^1, g^1 according to (56) and (57) are defined by the formulas

$$f^0=2n(1-n)\tau_2-n\tau_1-2nw(\chi_1+n\chi_2),$$
$$g^0=2n(1-n)\chi_2+(2-3n)\chi_1, \tag{58}$$
$$f^1=\tfrac{1}{2}nw^2\tau_2,\quad g^1=2nw\tau_2+w\tau_1+\tfrac{1}{2}(nw^2\chi_2+w^2\chi_1).$$

From here, in view of (57), it follows that higher-order corrections vanish, and we obtain an exact second-order L–B symmetry of DEs (54) at $\varphi=1$ for arbitrary $a\neq 0$; this symmetry gives rise to the exact solution[38,39] satisfying the boundary condition $N=\cosh^{-2}(x)$ defined by the zeroth-order term g^0.

The arbitrariness in functions f^0, g^0 enables us to construct RG-symmetries for any boundary conditions. As an illustration, we present RG-symmetries for the BVP with

$$H(n)=(\ln(1/n))^{1/2}\,, \tag{59}$$

describing space evolution (self-focusing) of the Gaussian beam with the originally plane phase front at $\tau=0$. To satisfy the initial distribution (59), one can choose the following functions $f^0=1+2n\chi\chi_1$ and $g^0=0$. For this value of f^0 the inherited point group of the BVP is constructed with the help of formulas (57) and is given by the operator

$$R=-2\chi\partial_w+2a\tau\partial_n+\left(1+\frac{a\tau^2}{n}\right)\partial_\tau. \tag{60}$$

The invariance condition for the solution of the BVP with respect to RG with this operator is presented in the form of two partial DEs,

$$\chi\chi_w-a\tau\chi_n=0\,,\quad 2\chi\tau_w-2a\tau\tau_n+1+(a\tau^2/n)=0\,,$$

the solution of which yields the desired approximate analytical solution of the problem

$$x^2=(ant^2-\ln n)[1-2Q(\sqrt{ant^2})]^2\,,\quad v=-2\frac{x}{t}\,\frac{Q(\sqrt{ant^2})}{[1-2Q(\sqrt{ant^2})]}\,. \tag{61}$$

Here the function $Q(z)$ is expressed as follows:

$$Q(z)=ze^{-z^2/2}\int_0^z dt\ e^{t^2/2}\,.$$

The first-order approximate symmetry obtained can be used to calculate a higher-order approximation of the RG-operator (60) and, thus, to improve the analytical solution (61). One can also obtain new-type RG-operators just by substituting the approximate solution (61) into formulas (27).

VIII. CONCLUSION

In this paper we present a new approach to constructing RG-symmetries based on the mathematical apparatus of classical and modern group analysis. It differs from the traditionally used methods of constructing RGs in theoretical physics and is formulated as a sequence of the following steps

(I) *Constructing the RG-manifold*, that takes into account both basic equations and the corresponding boundary conditions.

(II) *Calculating the symmetry group*, admitted by the RG-manifold.

(III) *Restricting the group* obtained *on the solution of the BVP*.

(IV) Utilizing the RG-operators to *find analytical expressions for solutions*.

As it was shown there exists a set of different algorithms for finding RG-symmetries. The choice of a particular one for a given physical problem depends on a mathematical model used for the problem description.

It should be noted that different methods of constructing RG-symmetries described above do not exhaust the suggested approach (see, e.g., Refs. 42, 43). The procedure of constructing RG-symmetries may combine different algorithms; for example, of interest is a simultaneous use of the method based on approximate symmetries and the invariant embedding method, and so on.

Our approach reveals a close relation of the *functional self-similarity* property (i.e., ''classical'' RG–symmetry as an exact property of a solution) to an invariance condition of a BVP solution with respect to the RG-operator. Mathematically, the latter is formulated as the vanishing condition for the coordinate of a canonical RG-operator on a solution of BVP.

One can readily see that RG-operators may appear in the form that is different from the QFT case,[6] e.g., operators of Lie–Bäcklund RG-symmetries. However, in some cases ''our'' RG-operators can look like the ones in the QFT renormalization group. For example, a linear combination of operators α_1X_1 and $\alpha_3(X_3+X_4)$ for the BVP (35) with $\alpha_1=1$, $\alpha_3=-a$ gives

$$R = \partial_t - ax^2\partial_x, \tag{62}$$

which is formally equivalent [with an appropriate change of variables $t = \ln x$, $x = g$ and $\beta(g) = ag^2$] to the differential operator for the one-coupling massless QFT model in the one-loop approximation.

Up to now this approach is feasible for systems that can be described by DEs and is based on the formalism of modern group analysis.

It seems also possible to extend our approach on physical systems that are not described just by differential equations. A chance of such an extension is based on recent advances in group analysis of systems of integro-differential equations[49,50] that allow transformations of both dynamical variables and functionals of a solution to be formulated.[47] More intriguing is the issue of a possibility of constructing a regular approach for more complicated systems, in particular to that ones having an infinite number of degrees of freedom. The former can be represented in a compact form by functional integrals (or *path integrals*).

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Fundamental Research (Project No. 96-01-00195 and partially Projects No. 96-01-01297 and No. 96-15-96030).

[1] E. E. C. Stueckelberg and A. Petermann, ''La normalisation des constantes dans la theorie des quanta,'' Helv. Phys. Acta **22**, 499–520 (1953) (in French).
[2] M. Gell-Mann and F. Low, ''Quantum electrodynamics at small distances,'' Phys. Rev. **95**, 1300–1312 (1954).
[3] N. N. Bogoliubov and D. V. Shirkov, ''On renormalization group in quantum electrodynamics,'' Dokl. Akad. Nauk. **103**, 203–206 (1955); **103**, 391–394 (1955) (in Russian).
[4] N. N. Bogoliubov and D. V. Shirkov, ''Charge renormalization group in quantum field theory,'' Nuovo Cimento **3**, 845–863 (1956); see also Sov. Phys. JETP **3**, 57 (1956).
[5] K. Wilson, ''Renormalization group and critical phenomena,'' Phys. Rev. B **4**, 3184–3205 (1971).
[6] N. Bogoliubov and D. Shirkov, *Introduction to the Theory of Quantized Fields*, translation of 3rd Russian ed. (Wiley-Interscience, New York, 1980).
[7] The generalization to the case of several parameters is straightforward—see, e.g., Refs. 4, 8.
[8] D. V. Shirkov, ''Renormalization group in modern physics,'' Int. J. Mod. Phys. A **3**, 1321–1342 (1988).
[9] D. V. Shirkov, ''Renormalization group, invariance principle and functional self-similarity,'' Sov. Phys. Dokl. **27**, 197–200 (1982).
[10] D. V. Shirkov, ''Renormgroup and functional self-similarity in different branches of physics,'' Theor. Math. Phys. **60**, 778–782 (1984).
[11] L. V. Ovsyannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[12] N. H. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Riedel, Dordrecht, 1985).
[13] In general this means that S transforms according to some representation of a group.
[14] G. I. Barenblatt, *Similarity, Self-Similarity and Intermediate Asymptotics* (Plenum, New York, 1979).
[15] V. F. Kovalev, V. V. Pustovalov, ''Strong nonlinearity and generation of high harmonics in laser plasma,'' in Proceedings of contributed papers of the International Conference on Plasma Physics, Kiev, USSR, 6–12 April 1987, edited by A. G. Sitenko (Naukova Dumka, Kiev, 1987), Vol. 1, pp. 271–273; ''Influence of laser plasma temperature on the high harmonics generation process,'' *ibid.* pp. 274–277; for details see V. F. Kovalev and V. V. Pustovalov, ''Strong nonlinearity and generating of higher harmonics in inhomogeneous plasma,'' P. N. Lebedev Physical Institute, preprint No. 78, February, 1987 (in Russian).
[16] V. F. Kovalev and V. V. Pustovalov, ''Functional self-similarity in a problem of plasma theory with electron nonlinearity,'' Theor. Math. Phys. **81**, 1060–1071 (1990). A detailed discussion of related interesting physical results was presented by these authors in Sov. J. Quantum Electron. **18**, 463 (1988); **19**, 1454 (1989); Sov. J. Plasma Phys. **15**, 27 (1989); **15**, 327 (1989).
[17] N. Goldenfeld, O. Martin, and Y. Oono, ''Intermediate asymptotics and renormalization group theory,'' J. Sci. Comput. **4**, 355–372 (1989).
[18] N. Goldenfield, O. Martin, Y. Oono, and F. Lui, ''Anomalous dimensions and the renormalization group in a nonlinear diffusion process,'' Phys. Rev. Lett. **64**, 1361–1364 (1990).
[19] G.I. Barenblatt and Ya.B. Zeldovich, ''Intermediate Asymptotics in math. physics,'' Russ. Math. Surveys **26**, 45–61 (1971); ''Self-similar solutions as intermediate asymptotics,'' Annu. Rev. Fluid Mech. **4**, 285–312 (1972).
[20] G.I. Barenblatt, *Scaling, Self-Similarity and Intermediate Asymptotics* (Cambridge University Press, 1996).
[21] L.-Y. Chen, N. Goldenfeld, and Y. Oono, ''The renormalization group and singular perturbations: Multiple-scales, boundary layers and reductive perturbation theory,'' Phys. Rev. E **54**, 376–394 (1996).
[22] T. Kunihiro, ''A geometrical formulation of the renormalization group method for global analysis,'' Prog. Theor. Phys. **94**, 503–514 (1995).
[23] J. Bricmont and A. Kupiainen, ''RG and the Ginzburg–Landau equation,'' Commun. Math. Phys. **150**, 193–203 (1992).
[24] J. Bricmont, A. Kupiainen, and G. Lin, ''RG and asymptotics of solutions of nonlinear parabolic equations,'' Commun. Pure Appl. Math. **47**, 893–922 (1994).
[25] J. Bricmont, A. Kupiainen and J. Xin, ''Global large time self-similarity of a thermal-diffusive combustion system with critical nonlinearity,'' J. Diff. Eqns. **130**, 9–35 (1996).

[26] D.V. Shirkov, ''Bogolyubov renormgroup,'' Russ. Math. Surveys **49:5**, 155–176 (1994)—with numerous misprints; for the corrected version see ''The Bogoliubov renormalization group (second English printing),'' JINR Commun. E2-96-15; also in hep-th/9602024.
[27] V.F. Kovalev, V.V. Pustovalov, and D.V. Shirkov, ''Group analysis and renormgroup,'' Commun. JINR, Dubna (1995), P5-95-447 (in Russian).
[28] D. V. Shirkov, ''Renormalization group symmetry and sophus Lie group analysis,'' Int. J. Mod. Phys. C **6**, 503–512 (1995).
[29] *CRC Handbook of Lie Group Analysis of Differential Equations*, edited by N. H. Ibragimov (CRC Press, Boca Raton, FL). Vol. 1: ''Symmetries, exact solutions and conservation laws,'' Vol. 2: ''Applications in engineering and physical sciences,'' Vol. 3: ''New trends in theoretical developments and computational methods.''
[30] To our knowledge the embedding method was originally used by V. A. Ambartzumyan, ''On the light scattering by atmospheres of planets,'' Astronom. J. **19**, 30 (1942) (in Russian); later, this method enjoyed wide application to different problems — see, e.g., J. Casti and R. Kalaba, *Imbedding Methods in Applied Mathematics* (Addison-Wesley, Reading, MA, 1973), and references therein.
[31] V. A. Baikov, R. K. Gazizov, and N. H. Ibragimov, ''Perturbation methods in group analysis,'' J. Sov. Math. **55**, 1450 (1991).
[32] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
[33] V. F. Kovalev and V. V. Pustovalov, ''Functional self-modeling in a nonlinear plasma theory,'' Soviet Physics—Lebedev Institute Reports, No. 3, 1989, pp. 54–57.
[34] V. F. Kovalev, V. V. Pustovalov, ''Symmetry group of nonlinear one-dimensional equations of electron plasma,'' Soviet Physics—Lebedev Institute Reports, No. 2, 1991, pp. 28–31.
[35] V. F. Kovalev, S. V. Krivenko, and V. V. Pustovalov, ''The renormalization group method based on group analysis,'' in *Renormalization Group-91*, edited by D. V. Shirkov and V. B. Priezzev (World Scientific, Singapore, 1992), pp. 300–314.
[36] V. F. Kovalev and V. V. Pustovalov, ''Lie algebra of renormalization group admitted by initial value problem for Burgers equation,'' Lie Groups Appl. **1**, 104–120 (1994).
[37] V. F. Kovalev, S. V. Krivenko, and V. V. Pustovalov, ''Lie symmetry and a group on a solution of a boundary-value problem,'' P. N. Lebedev Physical Institute, preprint No. 13, April, 1995 (in English).
[38] V. F. Kovalev, V. V. Pustovalov, and S. I. Senashov, ''Lie–Bäcklund symmetry of nonlinear geometrical optics equations,'' Differential Equations **29**, 1521–1531 (1993).
[39] V. F. Kovalev, ''Computer algebra tools in a group analysis of quasi-Chaplygin system of equations,'' in *New Computing Techniques in Physics Research IV*, edited by B. Denby and D. Perret-Gallix (World Scientific, Singapore, 1995), p. 229.
[40] V. F. Kovalev, ''Group and renormgroup symmetry of quasi-Chaplygin media,'' J. Nonlin. Math. Phys. **3**, 351–356 (1996).
[41] V. F. Kovalev and V. V. Pustovalov, ''Group and renormgroup symmetry of a simple model for nonlinear phenomena in optics, gas dynamics and plasma theory,'' Math. Comput. Mod. **25**, 165–179 (1997).
[42] V. F. Kovalev, ''Renormgroup symmetries in problems of nonlinear geometrical optics,'' Theor. Math. Physics **111** (in press).
[43] V. F. Kovalev and D. V. Shirkov, ''Renormalization group in mathematical physics and some problems of laser optics,'' JINR Preprint, Dubna, 1997, E5-97-41 (in English); J. Nonlin. Opt. Phys. Mat. **6** (in press).
[44] A disadvantage of this is that some examples have a methodological, rather than a physical, significance.
[45] S. A. Chaplygin, *On Gas Jets* (GITTL, Moscow–Leningrad, 1949) (in Russian).
[46] B. A. Trubnikov and S. K. Zhdanov, ''Unstable quasi-gaseous media,'' Phys. Rep. **155**, 137–230 (1987).
[47] V. F. Kovalev, S. V. Krivenko, and V. V. Pustovalov, ''Symmetry group of Vlasov–Maxwell equations in plasma theory,'' J. Nonlin. Math. Phys. **3**, 175–180 (1996).
[48] S. A. Akhmanov, R. V. Khokhlov, and A. P. Sukhorukov, ''On the self-focusing and self-channeling of intense laser beams in nonlinear medium,'' Sov. Phys. JETP **23**, 1025–1033 (1966).
[49] Yu. N. Grigoryev and S. V. Meleshko, ''Group analysis of integro-differential Boltzmann equation,'' Sov. Phys. Dokl. **32**, 874–876 (1987).
[50] V. F. Kovalev, S. V. Krivenko, and V. V. Pustovalov, ''Group analysis of the Vlasov kinetic equations,'' Differential Equations **29**, 1568–1578 (1993); **29**, 1712–1721 (1993).

Three-dimensional metrics with a spherical homogeneous model

Victor Patrangenaru[a)]
Indiana University, Department of Mathematics, Bloomington, Indiana 47405-5701

(Received 9 July 1997; accepted for publication 18 August 1997)

One finds the moduli spaces of isocurved homogeneous metrics on S^3. Then, one determines the germs of three-dimensional analytic metrics with a spherical homogeneous model. © *1998 American Institute of Physics.* [S0022-2488(98)00202-3]

I. INTRODUCTION

Assume $(\mathbf{M}_0,\mathbf{g}_0)$ is a homogeneous pseudo-Riemannian manifold of curvature tensor field ${}_0\mathbf{R}$. Following Ref. 1, we will say that $(\mathbf{M}_0,\mathbf{g}_0)$ is a *homogeneous model* of $(\mathbf{M},\mathbf{g})$ if for each point $\mathbf{p}$ in $\mathbf{M}$, there is a point $\mathbf{p}_0$ in $\mathbf{M}_0$ and an isometry $\mathbf{F}:\mathbf{T}_p\mathbf{M}\to\mathbf{T}_{p_0}\mathbf{M}_0$ such that $\mathbf{F}^*{}_0\mathbf{R}_{p_0}=\mathbf{R}_p$, where $\mathbf{R}$ is the curvature tensor field of $\mathbf{M}$.

In this paper one is concerned with those germs of analytic manifolds whose homogeneous model $\mathbf{M}_0$ is $\mathbb{S}^3$.

In Sec. II, based on the $\mathfrak{h}$-triples algorithm in Ref. 2 and of the list of invariant Lorentz structures on three-dimensional unimodular Lie groups in Ref. 3, one takes care of the homogeneous spherical model, that is the moduli space of homogeneous Lorentz structures on $\mathbb{S}^3$; this is a three-dimensional stratified set, slightly more complicated than its Riemannian analog:[4] Let α be the action of $\mathbb{Z}_2$ on the first quadrant of $\mathbb{R}^2$, generated by the symmetry about the diagonal. Let Q be the quotient space $\mathbb{R}_+{}^2/\alpha$ and $L=Q\times(0,\infty)$. A point λ in L has a unique representative, $(\lambda_1,\lambda_2,\lambda_3)\in\mathbb{R}^3$, $\lambda_1\geqslant\lambda_2>0$, $\lambda_3>0$.

Let $\sigma=\sigma^1e_1+\sigma^2e_2+\sigma^3e_3$ be the dual form of the unit quaternions, which is given in real coordinates of a point $\mathbf{x}$ in $\mathbb{S}^3\subseteq\mathbb{R}^4$ is by

$$\begin{aligned}\sigma^1&=-x^2dx^1+x^1dx^2-x^4dx^3+x^4dx^3,\\ \sigma^2&=-x^3dx^1+x^4dx^2-x^1dx^3+x^2dx^4,\\ \sigma^3&=-x^4dx^1+x^3dx^2-x^2dx^3+x^1dx^4.\end{aligned}\tag{1.1}$$

Theorem 1.1: *There is a one to one map* $\mathbf{G}$ *from* L *onto isometry classes of homogeneous Lorentz structures on* $\mathbf{S}^3$, *given by* $\mathbf{G}(\lambda)=$ *isometry class of* $\mathbf{g}_\lambda$,

$$\mathbf{g}_\lambda=4((\lambda_2\lambda_3)^{-1}\sigma_1^2+(\lambda_3\lambda_1)^{-1}\sigma_2^2-(\lambda_1\lambda_2)^{-1}\sigma_3^2).\tag{1.2}$$

The dimension of the total isometry group I_λ *of* $(\mathbf{S}^3,\mathbf{g}_\lambda)$, is 4, *when* $\lambda_1=\lambda_2$, *respectively*, 3 *when* $\lambda_1\neq\lambda_2$.

We shall also prove:

Theorem 1.2: (a) *The local geometry of a homogeneous Lorentz structure* $\mathbf{g}$ *on* $\mathbf{S}^3$ *is uniquely determined by the value at one point of the curvature tensor, whenever zero is a regular value of the Ricci operator.*

(b) *Assume zero is one of the principal Ricci curvatures of a homogeneous Lorentz structure on* $\mathbb{S}^3$. *Then zero is a double root of the Ricci polynomial, and the third principal Ricci curvature is a positive constant* k^2. *For each positive* k, *there is a one-parameter family of nonisometric homogeneous Lorentz structures on* $\mathbb{S}^3$, *with same curvature tensor of Ricci spectrum* $(\mathbf{0},k^2,\mathbf{0})$.

a)Electronic mail: vpatrang@indiana.edu

 © 1998 American Institute of Physics

Assume (**M**,**g**) has a spherical model. In this case **g** is a *curvature homogeneous* (**ch**) metric on **M**, that is, its curvature tensor field is punctually self-equivalent. In dimension 3, a curvature tensor is uniquely determined by its Ricci tensor. The Ricci operator is self-adjoint and in the Lorentz case there are four types of Ricci tensors,[5–7] among which those that are diagonalizable w.r.t. an orthobasis are said to be of *type I*. Three-dimensional chR metrics with a generic Ricci tensor, have been studied in Refs. 8–12. The nongeneric three-dimensional Lorentz case was investigated in Refs. 13–15.

Using the technique of homotopy operators,[12] one may determine the structure of chL generic metrics in general, but homogeneous Lorentz structure on $\mathbb{S}^3$ have a type I Ricci tensor, as shown in Sec. II. As such in Sec. III, we prove the following.

Theorem 1.3: *Assume* **R** *is a generic Ricci tensor of type I. Then the space of germs of analytic ch metrics whose Ricci spectrum is* **R**, *depends on three functions of three variables and three functions of two variables*. This result also holds in the generic Riemannian analytic case; for other less general, still more definitive results, see Ref. 16, where the ''local moduli space'' of such metrics depends on functions in one variable only. Notice that the topologies of **M** and its model may differ drastically. It suffices to recall examples of complete noncompact manifolds with a compact homogeneous model.[15,17] All these examples have a nongeneric Ricci tensor. Bryant conjectured that there are no nonhomogeneous chR metric on $\mathbb{S}^3$, of a nongeneric Ricci tensor.[9] ■

II. HOMOGENEOUS METRICS ON $\mathbb{S}^3$

A partial classification of isometry groups of compact Lorentz manifolds has been given in Ref. 18 based on some previous results of D'Ambra.[19] Yet, even for three-dimensional compact manifolds **M**, despite new techniques, such as those of flows introduced in Ref. 20, the homogeneous Lorentz metrics on $\mathbb{S}^3$, with a multiply transitive group of isometries were omitted.

In this section we classify homogeneous Lorentz structures on $\mathbb{S}^3$. The metric classification of homogeneous Riemannian structures on $\mathbb{S}^3$, is given in Ref. 4. Our approach differs also from the classical characterization of Killing vector fields.[21]

Assume **g** is a homogeneous Lorentz metric on $\mathbb{S}^3$. From Ref. 18, one may infer that the isometry group **K** of $(\mathbb{S}^3,\mathbf{g})$ is compact. As such, the isotropy group **H** is a compact subgroup of $SO(2,1)$, thus being trivial or isomorphic to $SO(2)$. By considering the Levi decomposition, one sees that up to finite covers, **K** is isomorphic to $\mathbb{S}^3$ or to $\mathbb{S}^3\times\mathbb{S}^1$, since tori do not fiber over $\mathbb{S}^3$.

In order to determine the moduli space of **K**-invariant Lorentz metrics on $\mathbb{S}^3=\mathbf{K}/\mathbf{H}$, we will use the technique of $\mathfrak{h}$-triples introduced in Ref. 2. There $\mathfrak{h}$ stands for the Lie algebra of **H**.

Assume the restriction of the Killing form of $\mathfrak{o}(2,1)$ to $\mathfrak{h}$ is nondegenerated; then $\mathfrak{h}^\perp$ is the orhtocomplement of $\mathfrak{h}$ in $\mathfrak{o}(2,1)$ is w.r.t. the Killing form.

Let $\boldsymbol{\nabla},\boldsymbol{\nabla}^0$ be Levi-Civita and Ambrose–Singer (canonical) connection on $(\mathbb{S}^3,\mathbf{g})$, and let **G** be their difference tensor $\boldsymbol{\nabla}-\boldsymbol{\nabla}^0$. If **p** is a point of $\mathbb{S}^3$, **G** yields a map $\mathbf{G}_p:\mathbf{T}_p\mathbb{S}^3\to\mathrm{End}\,\mathbf{T}_p\mathbb{S}^3$, and if one selects an orthoframe $\mathbf{u}=(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ in $\mathbf{T}_p\mathbb{S}^3$, one may identify $\mathbf{G}_p$ with the so-called *Cartan–Singer map* $\Gamma:\mathbb{R}^3\to\mathfrak{h}^\perp$.[2]

With respect to the splitting $\mathfrak{o}(2,1)=\mathfrak{h}\oplus\mathfrak{h}^\perp$, the curvature at **p** has the components

$$\overline{\Omega}:\mathbb{R}^3\times\mathbb{R}^3\to\mathfrak{h}$$

and

$$\Omega_\perp=[\Gamma(X),\Gamma(Y)]_{\mathfrak{h}^\perp}+\Gamma(T(X,Y)), \tag{2.1}$$

where $[\ ,\ ]_{\mathfrak{h}^\perp}$ is the $\mathfrak{h}^\perp$ part of the commutator and

$$T(X,Y)=\Gamma(Y)X-\Gamma(X)Y.$$

The curvature operator at **p** is

$$\hat{\Omega}=\overline{\Omega}\oplus\Omega_\perp. \tag{2.2}$$

The Ricci form associated with $(\mathfrak{k},\mathbf{E})$ is the bilinear symmetric form $\rho:\mathbb{R}^3\times\mathbb{R}^3\to\mathbb{R}$, given by

$$\rho(\mathbf{e}_i,\mathbf{e}_j)=\mathrm{Tr}(\mathbf{x}\to\hat{\Omega}(\mathbf{x},\mathbf{e}_i)\mathbf{e}_j), \tag{2.3}$$

and the *Ricci polynomial* Ric is defined by

$$\mathrm{Ric}(t)=\det(\boldsymbol{\rho}(\mathbf{e}_i,\mathbf{e}_j)-t\ {}_1\delta_{ij}), \tag{2.4}$$

where ${}_1\delta_{1j}=\delta_{1j}$ for $i=1,2$ and ${}_1\delta_{3j}=-\delta_{3j}$; Ric(**t**) is a local metric invariant of **g**.

Another such invariant is $\|\Gamma\|^2$, the *square pseudonorm* of the Cartan–Singer map (see Ref. 4, for the Riemannian case),

$$\|\Gamma\|^2={}_1\delta^{ij}k(\Gamma(\mathbf{e}_i),\Gamma(\mathbf{e}_j)), \tag{2.5}$$

where Einstein's summation convention applies.

We recall that $\mathbf{f}^i_j=\mathbf{E}^i_j-{}_1\delta_{1j}\mathbf{E}^j_i$, where $\mathbf{E}^i_j(\mathbf{e_k})=\delta^i_{\mathbf{k}}\mathbf{e}_j$, $1\leqslant j<i\leqslant 3$, is a basis of $\mathfrak{o}_1(3)$.

(a) *Assume* $\mathbf{H}=\mathbb{S}^1$, and select an orthoframe **u** at **p**, such that $\mathbb{S}^1$ fixes $\mathbf{u}_3$, that is, $\mathfrak{h}=\mathfrak{o}(2)$.

Straightforward computations based on this $\mathfrak{h}$-triple technique, show that

$$\begin{gathered}\Gamma(e_1)=b\mathbf{f}^3_2,\quad \Gamma(e_2)=-b\mathbf{f}^3_1,\quad \Gamma(e_3)=0,\\ \overline{\Omega}(e_1,e_2)=\Omega_{12}\mathbf{f}^2_1,\quad \overline{\Omega}(e_1,e_3)=\overline{\Omega}(e_2,e_3)=\mathbf{0},\quad \Omega_{12}>3b^2,\quad b\neq 0,\end{gathered} \tag{2.6}$$

and one finds the following Ricci polynomial:

$$\mathrm{Ric}(t)=(t+2b^2)(t+b^2-\Omega_{12})^2. \tag{2.7}$$

Note that the restriction of the action of $\mathbf{K}=\mathbb{S}^1\times\mathbb{S}^3$ on $\mathbb{S}^3$ to $\mathbb{S}^1\times 1$ is not trivial. At Lie algebra level, $\mathfrak{k}$, the Lie algebra of **K** is isomorphic to $(\mathfrak{o}(2)\oplus\mathbb{R}^3,[\ ,\ ])$, where the bracket $[\ ,\ ]$ is defined by

$$\begin{gathered}[\mathbf{e}_1,\mathbf{e}_2]=2b\mathbf{e}_3+(b^2-\Omega_{12})\mathbf{f}^2_1,\quad [\mathbf{e}_1,\mathbf{e}_3]=b\mathbf{e}_2,\\ [\mathbf{e}_2,\mathbf{e}_3]=-b\mathbf{e}_1,\quad [\mathbf{f}^2_1,\mathbf{e}_1]=-\mathbf{e}_2,\\ [\mathbf{f}^2_1,\mathbf{e}_2]=\mathbf{e}_1,\quad [\mathbf{f}^2_1,\mathbf{e}_3]=0.\end{gathered} \tag{2.8}$$

The last three equations in (2.8) stand for the linear isotropic representation of $\mathbb{S}^1\times\mathbf{1}$ on $\mathbf{T}_1\mathbb{S}^3$.

On the other hand, $\mathrm{Span}(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3+(2b)^{-1}(b^2-\Omega_{12})\mathbf{f}^2_1)$ is a transitive Killing subalgebra of $\mathfrak{k}$ isomorphic to $\mathfrak{o}(3)$, showing that the homogeneous metric of $\mathbb{S}^3$ is actually a left-invariant metric on $\mathbb{S}^3$ regarded as a Lie group.

(b) *Assume* now $\mathfrak{h}=0$, that is, we look for left-invariant Lorentz metrics **g** on $\mathbb{S}^3$. Set $\mathbf{p}=\mathbf{1}_{\mathbb{S}^3}$. Using a unimodular-type argument (Ref. 3 or 5), it follows that there a **g** orthoframe $\mathbf{u}=(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ of the Lie algebra of $\mathbb{S}^3$, w.r.t. which the structure equations take the form

$$[\mathbf{e}_1,\mathbf{e}_2]=\lambda_3\mathbf{e}_3,\quad [\mathbf{e}_2,\mathbf{e}_3]=\lambda_1\mathbf{e}_1,\quad [\mathbf{e}_3,\mathbf{e}_1]=\lambda_2\mathbf{e}_2, \tag{2.9}$$

with positive $\lambda-s$. W.l.o.g. one may assume that $\lambda_1\geqslant\lambda_2>0$, $\lambda_3>0$. Indeed, if $\lambda_1<\lambda_2$, we set $\mathbf{e}'_1=\mathbf{e}_2$, $\mathbf{e}'_2=\mathbf{e}_1$, $\mathbf{e}'_3=-\mathbf{e}_3$, and then get an orthobasis $(\mathbf{e}'_1,\mathbf{e}'_2,\mathbf{e}'_3)$ that satisfies similar equations with structure constants λ' such that $\lambda'_1>\lambda'_2$.

u extends to a global moving frame on $\mathbb{S}^3$, and by the moving frame method, one may determine the components of the Ricci curvature w.r.t. this frame field. We set

$$\lambda_3=\mu_1-\mu_2,\quad \lambda_1=\mu_2-\mu_3,\quad \lambda_2=-\mu_1-\mu_3. \tag{2.10}$$

The Cartan–Singer map is

$$\Gamma(\mathbf{e}_1)=\mu_1\mathbf{f}^3_2,\quad \Gamma(\mathbf{e}_2)=\mu_2\mathbf{f}^3_1,\quad \Gamma(\mathbf{e}_3)=\mu_3\mathbf{f}^2_1. \tag{2.11}$$

Then, up to a permutation of the indices, the possible nonzero components of the Riemannian curvature tensor are

$$R^1_{212}=-\mu_1\mu_2-\mu_3\lambda_3,\quad R^1_{313}=-\mu_1\mu_3+\mu_2\lambda_2,\quad R^2_{323}=\mu_2\mu_3-\mu_1\lambda_1. \tag{2.12}$$

Our basis diagonalizes the Ricci curvature:

$$\rho_{11}=2\mu_2\mu_3,\quad \rho_{22}=-2\mu_1\mu_3,\quad \rho_{33}=-2\mu_1\mu_2. \tag{2.13}$$

We are ready for the following proof.

Proof of Theorem 1.1: We have noticed that the set of roots of the Ricci polynomial is an invariant of the isometry class of a given homogeneous Lorentz manifold. If M is the geometric realization of a **0**-triple, defined by $\lambda\in L$, this set is $\Sigma=\{\rho_{11},\rho_{22},-\rho_{33}\}$. The relative position of the elements in Σ on the real line is shown below:

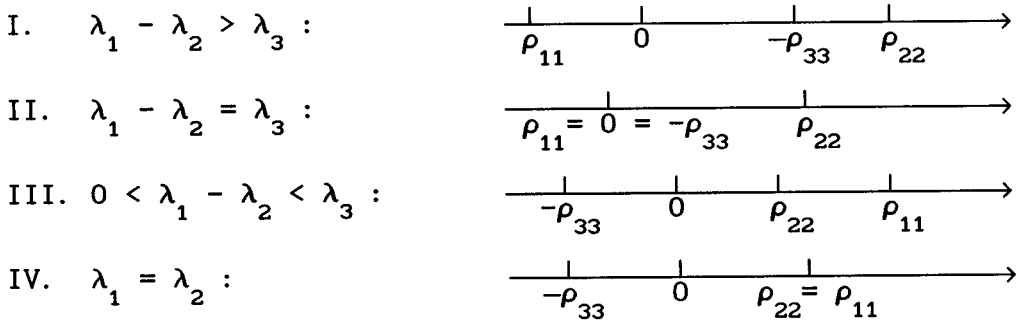

We proved the following.

Proposition 2.1: The Ricci tensor of a homogeneous Lorentz metric on $\mathbb{S}^3$ is of type I. In the generic situation, the Ricci operator has two negative eigenvalues, whose eigenvectors are one spatial and the other temporal, and the third eigenvalue is positive.

We would like to identify those Lorentz–Lie group structures on $\mathbb{S}^3$, with a nontrivial isotropy group of isometries. The Killing algebra $\mathfrak{k}(\mathbb{S}^3,\mathbf{g})$ is four dimensional. From (2.7), it follows that in this case there are some constants b, Ω_{12}, with $\Omega_{12}-3b^2>0$, and the Ricci polynomial has one positive double root $\Omega_{12}-b^2$, and one nonpositive simple root $-2b^2\leqslant 0$. We compare with the above chart; this may happen only when we are in case IV, and $\rho_{11}=\rho_{22}=\Omega_{12}-b_2$, $\rho_{33}=2b^2$. Then $\lambda_1=\lambda_2=(\Omega_{12}-3b^2)/2b$, $\lambda_3=2b$, which shows that our Killing algebra is indeed isomorphic to the subalgebra $\mathrm{Span}(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3+(2b)^{-1}(b^2-\Omega_{12})\mathbf{f}_1^2)$ of $\mathfrak{k}$ in (2.8).

We consider now the other cases I, II, III. If $\lambda,\lambda'\in\Lambda$, define two 0-triples, whose geometric realizations are locally isometric geometric models, an inspection of their equal Ricci spectra Σ,Σ', shows that either (a) λ,λ' are in the same case; or (b) λ is in case I and λ' is of case III.

Let us examine (b) first. Then the chart shows that

$$\rho_{11}=-\rho'_{33},\quad \rho_{22}=\rho'_{11},\quad \rho_{33}=-\rho'_{22}.$$

This is

$$\mu_2\mu_3=\mu'_1\mu'_2,\quad \mu_1\mu_3=\mu'_2\mu'_3,\quad \mu_1\mu_2=-\mu'_1\mu'_3.$$

We manipulate these equalities and find $\mu_1^2=\mu_3'^2$, $\mu_2^2=\mu_1'^2$, $\mu_3^2=\mu_2'^2$, or $\mu_1=\mu'_3$, $\mu_2=\mu'_1$, $\mu_3=-\mu'_2$, and from here $-\lambda_3=\lambda'_1+\lambda'_3$, which yields a contradiction, since both sides are positive.

It is obvious that if both λ,λ' are in same case, I, or III, then $\Sigma=\Sigma'$ implies $\mu=\mu'$. That is, $\lambda=\lambda'$.

We are left with the case when λ,λ' are both in case II. Then by (2.11), the corresponding Cartan–Singer maps are defined by

$$\Gamma(\mathbf{e}_1)=\mu_1\mathbf{f}_2^3,\quad \Gamma(\mathbf{e}_2)=0,\quad \Gamma(\mathbf{e}_3)=\mu_3\mathbf{f}_1^2,$$

$$\Gamma'(\mathbf{e}_1)=\mu'_1\mathbf{f}_2^3,\quad \Gamma'(\mathbf{e}_2)=0,\quad \Gamma'(\mathbf{e}_3)=\mu'_3\mathbf{f}_1^2,$$

and are $\mathbf{0}_1(3)$ conjugated. The square pseudonorms of Γ and Γ' being equal, it turns out that $\mu_1{}^2+\mu_3{}^2=\mu_1'{}^2+\mu_3'{}^2$. But in case II, $\mu_1=\lambda_3$, $\mu_3=-\lambda_1$, the nonzero Ricci curvature is $2\lambda_1\lambda_3$, and $\lambda_2=\lambda_1-\lambda_3$ is positive. As such, we have

$$\lambda_1{}^2+\lambda_3{}^2=\lambda_1'{}^2+\lambda_3'{}^2,\lambda_1\lambda_3=\lambda_1'\lambda_3', \quad \lambda_1>\lambda_3, \quad \lambda_1'>\lambda_3'. \tag{2.14}$$

The identities (2.14) clearly imply that $\lambda=\lambda'$.
The structure equations expressed in the dual coframe are

$$d\theta^1+\lambda_1\theta^2\wedge\theta^3=0, \quad d\theta^2+\lambda_2\theta^3\wedge\theta^1=0, \quad d\theta^3+\lambda_3\theta^1\wedge\theta^2=0, \tag{2.15}$$

we may find constants α_i, $i=1,2,3$ such that after the transformation $\theta^i=\alpha_i\sigma^i$, we obtain in the σ's some equations of the form (2.15) where all the coefficients of $\sigma^i\wedge\sigma^j$ are 2. Up to a diffeomorphism these are the dual forms of the moving frame that trivializes $\mathbf{T}S^3$ given in (1.1). ■

Proof of Theorem 1.2. Part a: Since $\mathbf{0}$ is not a Ricci principal curvature, we are in one of the cases I, III, or IV outlined in the previous proof. Due to (2.11), it follows that if I or III holds true, then $\hbar=\mathrm{Span}\,(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ is the Lie algebra of $\mathbf{I}(\mathbf{M})$. In each of these cases, we have seen in the previous proof that the curvature tensor uniquely determines the parameters λ, and therefore the geometries are determined by the curvature at one point. In case IV, the two distinct Ricci principal curvatures, ρ_{11} and ρ_{33}, uniquely determine the parameters Ω_{12} and b, as shown in the analysis of this case does in the proof of Theorem 1. These two parameters ''separate'' two such nonisometric Lorentz manifolds.

Part b. It is obvious from the chart in the proof of Theorem 1.1 that simple Ricci principal curvatures cannot occur. If the Ricci curvature operator has a zero eigenvalue, then we are in case II of the aforementioned chart, and the third eigenvalue is positive, say k^2. Assume the Lorentz–Lie groups have the Ricci spectrum $(\mathbf{0},k^2,\mathbf{0})$. The group is unimodular and all the nonzero components curvatures [see (3.4)] are determined by $k^2=\rho_{22}=\lambda_3(\lambda_2+\lambda_3)$. Therefore no other restriction on λ_2,λ_3 is imposed if this Lorentz Lie group has a homogeneous model in curvature with the Ricci spectrum $(\mathbf{0},k^2,\mathbf{0})$. From Theorem 1.1, it follows that for each t in $(\mathbf{0},k)$, there is a unique unimodular Lorentz Lie group, which is itself a total isometry group, corresponding with λ in L with

$$(\lambda_1,\lambda_2,\lambda_3)=\left(\frac{k^2}{t},\ \frac{k^2}{t}-t,t\right).$$

These are the only nonisometric three-dimensional homogeneous Lorentz structures on S^3, with the same homogeneous model in curvature. ■

Note. Interestingly enough, there are noncompact Lorentz–Lie groups with a spherical homogeneous model of the Ricci spectrum $(\mathbf{0},k^2,\mathbf{0})$. Such an example is provided by the Lorentz–Lie structures. $\mathbf{g}_t$ on $S1(2,\mathbb{R})$, whose Lie algebra of a $\mathbf{g}_t$ orthobasis $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ are the structure equations (2.9), where $\lambda_1=t<0$, $\lambda_2=t-k^2/2t$, $\lambda_3=k^2/2t$. ■

The first nonhomogeneous Lorentz examples with a homogeneous model in curvature also have double Ricci principal curvatures.[14]

III. CURVATURE HOMOGENEOUS MANIFOLDS WITH A GENERIC RICCI TENSOR OF TYPE I

Although we consider only the Lorentz case, the results (see also Ref. 9 for an EDS approach) or techniques (compare with Ref. 12) involved in the Riemannian case should be the same.

As far as the curvature is concerned, we have the following hierarchy:

constant curvature	R isotropic
locally symmetric	R LC-parallel ⇒ (LC = AS)
locally homogeneous	R,T AS-parallel, T = LC - AS
curvature homogeneous	R punctually self equivalent
Ricci curvature homogeneous	Ric punctually self equivalent
constant scalar curvature	

Unlike the top three structures listed in this chart that admit an algebraic approach (Refs. 21 and 22), constant scalar curvature manifolds do not, even in the analytic situation. Their study amounts to solving an EDS.

Four-dimensional chL's are reasonable-models for intergalactic travelers.[23]

In three dimensions, there is no distinction between chm's and the Ricci curvature homogeneous manifolds.

Three-dimensional chm's with a nongeneric Ricci curvature tensor, were studied in Refs. 8 and 24 in the Riemannian case, and in Refs. 13 and 15 in the Lorentz case. The Riemannian three-dimensional generic case was investigated in Refs. 9, 10, and 12.

In this section, we will determine the three-dimensional germs of generic chL metrics of Ricci tensor of type I, that is, those of distinct principal Ricci curvatures $\rho_1,\rho_2,-\rho_3$.

The local structure equations of an analytic three-dimensional chLm (**M**,**g**), in the moving frame approach:

$$d\theta+\omega\wedge\theta=0,\quad d\omega+\omega\wedge\omega=\Omega, \tag{3.1}$$

under the additional hypothesis that the frame field **X** consists of eigenvectors of a generic Ricci operator of type I, yield the following integrability conditions of (3.1):

$$\begin{aligned}&(\rho_2+\rho_3)\,\theta^1\wedge\theta^3\wedge\omega_3^2-(\rho_1+\rho_3)\,\theta^2\wedge\theta^3\wedge\omega_3^1=0,\\&(\rho_2+\rho_3)\,\theta^1\wedge\theta^2\wedge\omega_3^2-(\rho_1-\rho_2)\,\theta^2\wedge\theta^3\wedge\omega_2^1=0,\\&(\rho_1-\rho_2)\,\theta^1\wedge\theta^3\wedge\omega_2^1+(\rho_1+\rho_3)\,\theta^1\wedge\theta^2\wedge\omega_3^1=0.\end{aligned} \tag{3.2}$$

Let α be a vector-valued differential form. We denote by α_p the homogeneous polynomial component of α of degree p around a fixed point **x**. With this notation, $\theta=\Sigma_{p\geqslant 0}\theta_p$, $\omega=\Sigma_{p\geqslant 0}\omega_p$ and $\Omega=\Sigma_{p\geqslant 0}\Omega_p$ are the Taylor series around a fixed point **x** of **M** (the equations are restricted to a neighborhood **U** of **x**). The EDS (3.1)+(3.2) breaks into the following pieces:

$$d\theta_{p+1}=-\sum_{r+s=p}\omega_r\wedge\theta_s,\quad d\omega_{p+1}=-\sum_{r+s=p}\omega_r\wedge\omega_s+\Omega_p, \tag{3.3}$$

and

$$\begin{aligned}&(\rho_2+\rho_3)\sum_{r+s+t=p+1}\theta^1{}_r\wedge\theta^3{}_s\wedge\omega^2_{3,t}-(\rho_1+\rho_3)\sum_{r+s+t=p+1}\theta^2{}_r\wedge\theta^3{}_s\wedge\omega^1_{3,t}=0,\\&(\rho_2+\rho_3)\sum_{r+s+t=p+1}\theta^1{}_r\wedge\theta^2{}_s\wedge\omega^2_{3,t}-(\rho_1-\rho_2)\sum_{r+s+t=p+1}\theta^2{}_r\wedge\theta^3{}_s\wedge\omega^1_{2,t}=0,\\&(\rho_1-\rho_2)\sum_{r+s+t=p+1}\theta^1{}_r\wedge\theta^3{}_s\wedge\omega^1_{2,t}+(\rho_1+\rho_3)\sum_{r+s+t=p+1}\theta^1{}_r\wedge\theta^2{}_s\wedge\omega^1_{3,t}=0,\end{aligned} \tag{3.4}$$

W.l.o.g., we may assume that $\mathbf{x}=\mathbf{0}$ and $\theta|_{\mathbf{x}=0}=d_0id=id$, $\omega_0|_{\mathbf{x}=0}={}_0\Gamma^i_{jk}\,dx^k\,f^j_i$. Let **U** be an open set (an open neighborhood of the origin in $\mathbb{R}^3$), where the Taylor series are assumed to be convergent. Also, let **V** be a real vector space. $\Omega(\mathbf{U};\mathbf{V})$, the set of all **V**-valued analytic forms is first quadrant **V**-valued double complex: $\Omega(\mathbf{U};\mathbf{V})=\Sigma_{r,s}\Omega^{r,s}$, where $\Omega^{r,s}$ is the set of all differential forms of degree s with coefficients of homogeneous polynomials of degree r. In particular, $\Omega(\mathbf{U}):=\Omega(\mathbf{U};\mathbb{R})$ is the tensor product of a symmetric algebra with an exterior algebra.

The differential $d=\Sigma d^{r,s}$, $d^{r,s}:\Omega^{r,s}\to\Omega^{r-1,s+1}$ has total degree 0.
We will consider a variant of homotopy operators as defined in classical textbooks, such as Ref. 25: more precisely, assume the radial homotopy defined in a neighborhood of the origin in $\mathbb{R}\times\mathbf{U}$, is $\mathbf{H}:\mathbb{R}\times\mathbf{U}\to\mathbf{U}$, $\mathbf{h}(\mathbf{s},\mathbf{p})=\mathbf{sp}$, and for any differential form α on $\mathbf{U}$, let $\mathbf{H}^*\alpha=\mathbf{ds}\wedge\beta+\gamma$, where β and γ do not contain the differential $\mathbf{ds}$. Then $\mathbf{h}\alpha=\int_0^1\beta$.

The homotopy operators are the components of $\mathbf{h}$, $\mathbf{h}^{r-1,s+1}:\Omega^{r-1,s+1}\to\Omega^{r,s}$, and they act on an ordinary form $(i_1+\cdots+i_n=r-1, j_1<\cdots<j_{s+1})$ by

$$h^{r-1,s+1}((x^1)^{i_1}\cdots(x^n)^{i_n}\,dx^{j_1}\wedge\cdots\wedge dx^{j_{s+1}})$$

$$=\frac{1}{(r+s)}\,(x^1)^{i_1}\cdots(x^n)^{i_n}\left(\sum_{\alpha=1}^{s+1}(-1)^{\alpha-1}x^{j_\alpha}dx^{j_1}\wedge\cdots\wedge dx^{j_\alpha}\wedge\cdots\wedge dx^{j_{s+1}}\right). \tag{3.5}$$

Since

$$\mathbf{dh}+\mathbf{hd}=\mathbf{id}, \tag{3.6}$$

it follows that locally $\alpha\in\Omega^{r-1,s+1}$ $d\alpha=\beta$, iff

$$\alpha=h\beta+dh\ \ \alpha=h\beta+d\phi,\quad \phi\in\Omega^{r,s}. \tag{3.7}$$

The key point is that one may solve (3.3)–(3.4) inductively: from (3.6), the formal solution is

$$\theta_{p+1}=-\sum_{r+s=p}h(\omega_r\wedge\theta_s)+d\phi_{p+2}, \tag{3.8}$$

$$\omega_{p+1}=h\left(\Omega_p-\sum_{r+s=p}\omega_r\wedge\omega_s\right)+d\psi_{p+2}, \tag{3.9}$$

where $\phi_{p+2}\in\Omega(\mathbf{U};\mathbb{R}^3)$, $\psi_{p+2}\in\Omega(\mathbf{U};\mathfrak{o}_1(3))$ are homogeneous polynomial forms, such that (3.4) holds true.

In order to prove that $\theta=\Sigma_{p\geqslant 0}\ \theta_p$, $\omega=\Sigma_{p\geqslant 0}\ \omega_p$ given by (3.8), (3.9), and (3.4) is a solution of (3.3), (3.4), one has to find inductively ϕ_{p+1}, ψ_{p+1} satisfying these equations, such that the corresponding series $\Sigma_{p\geqslant 0}\ \theta_p$, $\Sigma_{p\geqslant 0}\ \omega_p$ are convergent. To be more precise, at each point $\mathbf{p}$ in $\mathbf{U}$, one has the pointwise defined scalar products $\langle\alpha,\beta\rangle_{\mathbf{p}}$ on differential forms, at the point $\mathbf{p}$, given a structure of a Riemannian vector bundle of $\Omega^s(\mathbf{U};\mathbf{V})$ of forms of degree $\mathbf{s}$. For each radius $\mathbf{r}$, define

$${}_{\mathbf{r}}\langle\alpha,\beta\rangle=\int_{B_{\mathbf{r}}(0)}\langle\alpha,\beta\rangle_{\mathbf{p}}\,\mathbf{dp}.$$

Assume $j_1<\cdots<j_{s+1}$. Then, due to (3.5), it follows that for any homogeneous polynomial f_{r-1} of degree $r-1$,

$${}_{\mathbf{r}}\|\mathbf{h}^{r-1,s+1}(f_{r-1}(x)dx^{j_1}\wedge\cdots\wedge dx^{j_{s+1}})\|^2<\frac{s+1}{(r+s)^2}\,\mathbf{r}^2{}_{\mathbf{r}}\|f_{r-1}(x)dx^{j_1}\wedge\cdots\wedge dx^{j_{s+1}}\|^2,$$

which for any form α in $\Omega^{r-1,s+1}$ imply

$${}_{\mathbf{r}}\|\mathbf{h}^{r-1,s+1}\alpha\|^2<\frac{s+1}{(r+s)^2}\,\mathbf{r}^2{}_{\mathbf{r}}\|\alpha\|^2. \tag{3.10}$$

For $r=p$, $s=2$ this inequality can be bettered,

$${}_{\mathbf{r}}\|\mathbf{h}^{1,2}\alpha\|<\frac{1}{p+2}\,\mathbf{r}_{\mathbf{r}}\|\alpha\|. \tag{3.11}$$

From (3.8), it follows that

$$_{\mathbf{r}}\|\theta_{p+1}\|<\frac{2p+2}{p+2}\,\mathbf{r}\,\max(_{\mathbf{r}}\|\omega_r\|,\quad _{\mathbf{r}}\|\theta_s\|)^2+{}_{\mathbf{r}}\|d\phi_{p+2}\|.$$

Since

$$\Omega_p=k_3\left(\sum_{r+s=p}\theta^1{}_r\wedge\theta^2{}_s\right)f_1^2-k_2\left(\sum_{r+s=p}\theta^1{}_r\wedge\theta^3 s\right)f_1^3+k_1\left(\sum_{r+s=p}\theta^2{}_r\wedge\theta^3{}_s\right)f_2^3.$$

from (3.9) and a number of inequalities, one gets

$$_{\mathbf{r}}\|\omega_{p+1}\|<\frac{2p+2}{p+2}\,\mathbf{r}\,\max(_{\mathbf{r}}\|\omega_r\|,{}_{\mathbf{r}}\|\theta_s\|)^2(|k_1|+|k_2|+|k_3|+1)+{}_{\mathbf{r}}\|d\psi_{p+2}\|.$$

Let $a_p=\max(_{\mathbf{r}}\|\omega_r\|,{}_{\mathbf{r}}\|\theta_s\|,r,s\leqslant p)$. The above inequalities can be rewritten as

$$\begin{gathered}
_{\mathbf{r}}\|\theta_{p+1}\|<\frac{2p+2}{p+2}\,\mathbf{r}a_p{}^2+{}_{\mathbf{r}}\|d\phi_{p+2}\|,\\
_{\mathbf{r}}\|\omega_{p+1}\|<\frac{2p+2}{p+2}\,\mathbf{r}a_p{}^2(|k_1|+|k_2|+|k_3|+1)+{}_{\mathbf{r}}\|d\psi_{p+2}\|.\\
\text{If}\quad b_p=\max(_{\mathbf{r}}\|d\phi_{p+1}\|,\quad _{\mathbf{r}}\|d\psi_{p+1}\|),
\end{gathered}\tag{3.12}$$

then

$$a_{p+1}<\frac{2p+2}{p+2}\,\mathbf{r}a_p{}^2(|k_1|+|k_2|+|k_3|+1)+b_{p+1}.\tag{3.13}$$

Let $\mathbf{K}$ be such that $[(2p+2)/(p+2)](|k_1|+|k_2+|k_3|+1)<\mathbf{K}$.

Theorem 3.1: *If* $\mathbf{c}=\mathbf{Kr}$ *is a sufficiently small and the series* $\phi=\Sigma\phi_p$, $\psi=\Sigma\psi_p$ *are uniformly convergent on* $\mathbf{B}(0,\mathbf{r})$, *then so are the series* $\theta=\Sigma_{p\geqslant 0}\theta_p$ and $\omega=\Sigma_{p\geqslant 0}\omega_p$, *whose general terms are solutions of (3.8) and (3.9).*

Proof: $a_{p+1}<\mathbf{K}\mathbf{r}a_p{}^2+b_{p+1}$.

Assumption: Let α_p, be defined by $\alpha_{p+1}=c\alpha_p{}^2+b_{p+1}$, $\alpha_1=a_1$. Since all the sequences are positive, one may prove by induction that $a_{p+1}<\alpha_{p+1}$.

Claim: The series $\Sigma\alpha_p$ is convergent.

Let $\beta_p=\max(b_s,s\geqslant p)$. Then β_p is decreasing and $\beta_p\geqslant b_p$. The series $\Sigma\beta_p$ is as well convergent, and if one defines the sequence α'_p, by $\alpha'_{p+1}=c\alpha'_p{}^2+\beta_{p+1}$, $\alpha'_1=a_1$, it is obvious that $\alpha_p\leqslant\alpha'_p$. Therefore w.l.o.g. we may assume that b_p is decreasing. Also, since b_p approaches 0, one may select $\mathbf{c}<1$, such that $\alpha_{p+1}<\max(\alpha_p,1)$ for some p, and therefore α_p is decreasing, and w.l.o.g. one may assume that $\alpha_p<1$. Then $\alpha_{p+1}<\mathbf{c}\alpha_p+b_{p+1}$, and therefore $(1-\mathbf{c})\Sigma_p^n\alpha_s<\Sigma_p^n b_s-\alpha_{n+1}$, which proves the claim. ■

The inductive step is the following.

Lemma 3.1: Assume $\theta_{p+1},\omega_{p+1}$ *defined in (3.8) and (3.9) is a solution of (3.3) and (3.4), for* $p=\overline{0,q}$ *and let* θ_{q+2} *given by (3.8) for* $p=q$, *with* ϕ_{q+2} *arbitrary. Then there is a solution* ω_{q+2} *given in* (3.9) *for* $p=q+1$, *which satisfies to the integrability conditions (3.4) for* $p=q+1$, *if the subspace of homogeneous polynomial solutions* $\psi^0_{q+2}=\psi^{0\ j}_{q+2,\mathbf{i}}\mathbf{f}^i_j\in\Omega^{q+2,0}$ *of the system of linear PDE's,*

$$\begin{gathered}
(\rho_1+\rho_3)\frac{\partial}{\partial x^1}\Psi^0_{q+2,3}{}^1-(\rho_2+\rho_3)\frac{\partial}{\partial x^2}\Psi^0_{q+2,3}{}^2=0,\\
(\rho_2+\rho_3)\frac{\partial}{\partial x^3}\Psi^0_{q+2,3}{}^2-(\rho_1-\rho_2)\frac{\partial}{\partial x^1}\Psi^0_{q+2,2}{}^1=0,
\end{gathered}\tag{3.14}$$

$$(\rho_1-\rho_2)\frac{\partial}{\partial x^2}\Psi^0_{q+2,2}{}^1+(\rho_1+\rho_3)\frac{\partial}{\partial x^3}\Psi^0_{q+2,3}{}^1=0,$$

is $3(q+3)$ *dimensional.*

Proof: If we make the assumption following (3.4), and identify the coefficients of $dx^1\wedge dx^2\wedge dx^3$ in the integrability conditions for $p=q$, we obtain the following linear partial differential system of equations in the components of the new added 2-form $d\Psi_{q+2}\in\Omega(\mathbf{U};\mathfrak{o}_1(3))$ with homogeneous polynomial coefficients:

$$(\rho_1+\rho_3)\frac{\partial}{\partial x^1}\Psi_{q+2,3}{}^1-(\rho_2+\rho_3)\frac{\partial}{\partial x^2}\Psi_{q+2,3}{}^2+F_{q,2}{}^1=0,$$

$$(\rho_2+\rho_3)\frac{\partial}{\partial x^3}\Psi_{q+2,3}{}^2-(\rho_1-\rho_2)\frac{\partial}{\partial x^1}\Psi_{q+2,2}{}^1+F_{q,3}{}^1=0,$$

$$(\rho_1-\rho_2)\frac{\partial}{\partial x^2}\Psi_{q+2,2}{}^1+(\rho_1+\rho_3)\frac{\partial}{\partial x^3}\Psi_{q+2,3}{}^1+F_{q,3}{}^2=0, \tag{3.15}$$

where F_q is a homogeneous $\mathfrak{o}_1(3)$-valued function, depending on $d\phi_s$, $s\leqslant q+2$, and $d\psi_s$, $s\leqslant q+1$. (3.14) is equivalent to the equation $F(d\Psi_{q+2})=F_q$, where F is the linear p.d.o. from **S** to **T** whose action on $d\Psi_{q+2}$ is given by the left-hand side of (3.15), with F_q identically 0. The result follows, given that both the source and target of F are finite dimensional, and then the Fredholm alternative applies. **T** is isomorphic to the product of three copies of the space of homogeneous polynomials of degree $q+1$ in three variables, and **S** to the product if three copies of the space of homogeneous polynomials of degree $q+2$ in three variables, and obviously $\dim S-\dim \mathbf{T}=3(q+3)$. ■

Proof of Theorem 1.3. Indeed, assume ψ^0_{q+2} is in $\ker F$. Since the Ricci operator is generic, $(\rho_1+\rho_3)(\rho_2+\rho_3)(\rho_1-\rho_2)\neq 0$. Let us set $X=-(\rho_2+\rho_3)\Psi^0_{q+2,3}{}^2$, $Y=(\rho_1+\rho_3)\Psi^0_{q+2,3}{}^1$, $Z=(\rho_1-\rho_2)\Psi^0_{q+2,2}{}^1$. Then $\partial Y/\partial x^1+\partial X/\partial x^2=0$, $\partial X/\partial x^3+\partial Z/\partial x^1=0$, $\partial Z/\partial x^2+\partial Y/\partial x^3=0$. If one takes on more partial derivative, one finds that $\partial X/\partial x^2=A(x^1,x^2)$, $\partial Y/\partial x^3=B(x^2,x^3)$, $\partial Z/\partial x^1=C(x^1,x^3)$, where A,B,C are arbitrary homogeneous functions in two variables of degree $q+2$, proving that $\dim\ker F=3(q+3)$. Theorem 1.3, follows from Theorem 3.1 and Lemma 3.1. ■

ACKNOWLEDGMENTS

The author would like to thank Ji-Ping Sha and Izu Vaisman for useful conversations. I would also like to thank Scot Adams, Peter Bueken, Robert Bryant, Luis Cordero, Phil Parker, Oldrich Kowalski, Lieven Vanhecke, and Abdelghani Zeghib for preprints, reprints, etc.

[1] O. Kowalski, F. Tricerri, and L. Vanhecke, ''Curvature homogeneous spaces with a solvable Lie group as homogeneous model,'' J. Math. Soc. Jpn. **4**, 461–484 (1992).
[2] V. Patrangenaru, ''Locally homogeneous pseudo-Riemannian manifolds,'' J. Geom. Phys. **17**, 59–72 (1995).
[3] S. Rahmani, ''Métriques de Lorentz sur les groupes de Lie unimodulaires de dimension trois,'' J. Geom. Phys. **9**, 295–302 (1992).
[4] V. Patrangenaru, ''Classifying 3 and 4 dimensional homogeneous spaces by Cartan triples,'' Pacific J. Math. **173**, 511–532 (1996).
[5] L. A. Cordero and P. E. Parker, ''Left invariant Lorentzian metrics on 3 dimensional Lie groups,'' Rend. Math. Appl. **7**, 129–155 (1997).
[6] G. Hall, T. Morgan, and Z. Perjes, ''Three dimensional spacetimes,'' Gen. Relativ. Gravit. **19**, 1137–1147 (1987).
[7] B. O'Neill, *Semi-Riemannian Geometry (With Applications to Relativity)* (Academic, New York, 1983).
[8] P. Bueken, ''Three-dimensional Riemannian manifolds with constant principal Ricci curvatures $\rho_1=\rho_2\neq\rho_3$,'' J. Math. Phys. **37**, 4062–4075 (1996).
[9] R. Bryant, private communication (1995).
[10] O. Kowalski and F. Prüfer, ''A classification of special Riemannian manifolds with distinct constant eigenvalues,'' J. Anal. Appl. **14**, 43–58 (1995).
[11] D. J. McManus, ''Riemannian three-metrics with degenerate Ricci tensors,'' J. Math. Phys. **36**, 362–369 (1995).

[12] A. Spiro and F. Tricerri, ''3-dimensional Riemannian metrics with prescribed Ricci principal curvatures,'' J. Math. Pure Appl. **74**, 253–271 (1995).
[13] P. Bueken, ''Three-dimensional Lorenzian manifolds with constant principal Ricci curvatures $\rho_1=\rho_2\neq\rho_3$,'' J. Math. Phys. **38**, 1000–1013 (1997).
[14] M. Cahen, J. Leroy, M. Parker, F. Tricerri, and L. Vanhecke, ''Lorentz manifolds modeled on a Lorentz symmetric space,'' J. Geom. Phys. **7**, 571–591 (1990).
[15] D. J. McManus, ''Lorentzian three-metrics with degenerate Ricci tensors,'' J. Math. Phys. **36**, 1353–1365 (1995).
[16] O. Kowalski and F. Prüfer, ''On Riemmanian 3-manifolds with distinct constant Ricci curvatures,'' Math. Ann. **300**, Heft 1, 17–29 (1994).
[17] F. Lastaria, ''Homogeneous metrics with the same curvature,'' Simon Stevin, **65**, 267–281 (1991).
[18] S. Adams and G. Stuck, ''The isometry group of a compact Lorentz manifold,'' Invent. Math. **129**, 239–287 (1997).
[19] G. D'Ambra, ''Isometry groups of Lorentz manifolds,'' Invent. Math. **92**, 555–565 (1988).
[20] A. Zeghib, ''Killing fields in compact Lorentz 3-manifolds,'' J. Diff. Geom. **43**, 859–894 (1996).
[21] C. Bona and B. Coll, ''Isometry groups of three-dimensional Lorentzian metrics,'' J. Math. Phys. **35**, 873–884 (1994).
[22] M. Cahen and N. Wallach, ''Lorentzian symmetric spaces,'' Bull. Am. Math. Soc. **76**, 585–591 (1970).
[23] V. Patrangenaru, ''Constant gravitational fields and redshift of light,'' to appear in J. Geom. Phys.
[24] O. Kowalski, ''A classification of Riemannian 3-manifolds with constant principal Ricci curvatures $\rho_1=\rho_2\neq\rho_3$,'' Nagoya Math. J. **132**, 1–36 (1993).
[25] J. Dupont, *Curvature and characteristic classes*, Lect. Notes Math. **640** (1978).

Duals of colored quantum universal enveloping algebras and colored universal $\mathscr{T}$-matrices

C. Quesne[a)]
Physique Nucléaire Théorique et Physique Mathématique, Université Libre de Bruxelles, Campus de la Plaine CP229, Boulevard du Triomphe, B-1050 Brussels, Belgium

(Received 4 June 1997; accepted for publication 10 September 1997)

We extend the notion of dually conjugate Hopf (super)algebras to the colored Hopf (super)algebras $\mathscr{H}^c$ that we recently introduced. We show that if the standard Hopf (super)algebras $\mathscr{H}_q$ that are the building blocks of $\mathscr{H}^c$ have Hopf duals $\mathscr{H}_q^*$, then the latter may be used to construct coloured Hopf duals $\mathscr{H}^{c*}$, endowed with colored algebra and antipode maps, but with a standard coalgebraic structure. Next, we review the case where the $\mathscr{H}_q$'s are quantum universal enveloping algebras of Lie (super)algebras $U_q(g)$, so that the corresponding $\mathscr{H}_q^*$'s are quantum (super) groups G_q. We extend the Fronsdal and Galindo universal $\mathscr{T}$-matrix formalism to the colored pairs $(\mathrm{U}^c(g),G^c)$ by defining colored universal $\mathscr{T}$-matrices. We then show that together with the colored universal $\mathscr{R}$-matrices previously introduced, the latter provide an algebraic formulation of the colored RTT-relations, proposed by Basu-Mallick. This establishes a link between the colored extensions of Drinfeld–Jimbo and Faddeev–Reshetikhin–Takhtajan pictures of quantum groups and quantum algebras. Finally, we illustrate the construction of colored pairs by giving some explicit results for the two-parameter deformations of (U(gl(2)),Gl(2)), and (U(gl(1/1)),Gl(1/1)). © *1998 American Institute of Physics.*
[S0022-2488(98)03601-9]

I. INTRODUCTION

In a recent paper[1] (henceforth referred to as I and whose equations will be subsequently quoted by their number preceded by I), we did introduce some new algebraic structures $\mathscr{H}^c=(\mathscr{H},\mathscr{C},\mathscr{G})$, termed colored Hopf algebras (see also Ref. 2). They are constructed by starting from a standard Hopf algebra set $\mathscr{H}=\{\mathscr{H}_q|q\in\mathscr{Q}\}$, corresponding to some parameter set $\mathscr{Q}$, and by generalizing the definitions of coalgebra maps, and antipodes, by combining the standard ones with the transformations of an algebra isomorphism group $\mathscr{G}$, called color group. Such transformations are labeled by some color parameters, taking values in a color set $\mathscr{C}$. Whenever the starting Hopf algebras $\mathscr{H}_q$ are quasitriangular, the resulting colored one $\mathscr{H}^c$ is characterized by the existence of a colored universal $\mathscr{R}$-matrix, denoted by $\mathscr{R}^c$, and satisfying the colored Yang–Baxter equation (YBE), i.e., the YBE with nonadditive spectral parameters.[3,4]

In I, we applied these new concepts to the Drinfeld–Jimbo formulation of quantum algebras,[5] and constructed various examples of colored quantum universal enveloping algebras (QUEA's) of both semisimple and nonsemisimple Lie algebras.

In a more recent paper[6] (henceforth referred to as II), we showed that our definitions can be easily extended to deal with graded algebraic structures,[7] thereby leading to colored Hopf superalgebras in general, and to colored QUEA's of Lie superalgebras in particular. In such cases, quasitriangularity implies that $\mathscr{R}^c$ is a solution of the colored graded YBE.

In the literature, there exist other extensions of Hopf algebras related to the colored YBE, and using either the Drinfeld–Jimbo,[5] or the Faddeev–Reshetikhin–Takhtajan[8] approach to quantum groups and quantum algebras. In the former class, one should mention two previous introductions of some algebraic structures called colored Hopf algebras, the first one by Ohtsuki[9] in the context of knot theory, and the second one by Bonatsos *et al.*[10] in the study of some nonlinear deformation of su(2). In the latter class, some extensions of the Faddeev–Reshetikhin–Takhtajan RLL- and

[a)]Directeur de recherches FNRS; Electronic mail: cquesne@ulb.ac.be

 © 1998 American Institute of Physics

RTT-relations, where R is replaced by a colored R-matrix, have been considered by Kundu and Basu-Mallick,[11,12] thereby leading to colored $U(R)$ and $A(R)$ Hopf algebras, respectively.

As shown in I, some direct connections exist between our colored Hopf algebras, and those previously introduced by Ohtsuki,[9] and by Bonatsos *et al.*,[10] as the former may be considered as generalizations of the latter. On the contrary, possible relationships with colored $U(R)$ and $A(R)$ Hopf algebras have not been investigated so far.

It is one of the aims of the present paper to fill in this gap for the colored $A(R)$ algebras, introduced by Basu-Mallick.[12] For such a purpose, we shall have to extend the notion of dually conjugate Hopf algebras[13] to the colored context.

In the case of standard Hopf algebras, the interest of such a duality concept was recently highlighted by Fronsdal and Galindo,[14] who constructed dual bases for the two-parameter deformations of Gl(2) and U(gl(2)), and used them to build a universal $\mathscr{T}$-matrix, providing an algebraic formulation of the RTT-relations, and yielding the quantum group generalization of the familiar exponential map from a Lie algebra to a Lie group. A similar construction was carried out for some other quantum group and algebra pairs,[15–18] and used to get representations of one member of the pair from those of the other.[19,20] Moreover, the Fronsdal and Galindo approach was also recently generalized to quantum supergroups and superalgebras.[21]

We plan to show here that this universal $\mathscr{T}$-matrix formalism can be extended to colored QUEA's and their duals, and that the resulting colored universal $\mathscr{T}$-matrix, to be denoted by $\mathscr{T}^c$, provides the key notion for establishing a link between such duals and Basu-Mallick's colored $A(R)$ algebras.

This paper is organized as follows. In Sec. II, duals $\mathscr{H}^{c*}$ of colored Hopf (super)algebras $\mathscr{H}^c$ are defined. In Sec. III, the case of colored QUEA's of Lie (super)algebras is considered, and colored universal $\mathscr{T}$-matrices are introduced. In Secs. IV and V, the theory developed in the previous section is applied to construct pairs $(\mathscr{H}^c,\mathscr{H}^{c*})$ for the two-parameter deformations of (U(gl(2)),Gl(2)), and (U(gl(1/1)),Gl(1/1)), respectively. Section VI contains some concluding remarks.

II. DUALS OF COLORED HOPF (SUPER)ALGEBRAS

Let $\mathscr{H}^c=(\mathscr{H},\mathscr{C},\mathscr{G})=(\mathscr{H}_q,+,m_q,\iota_q,\Delta_{q,\nu}^{\lambda,\mu},\epsilon_{q,\nu},S_{q,\nu}^{\mu};k,\mathscr{Q},\mathscr{C},\mathscr{G})$ be a colored Hopf algebra over some field k ($=\mathbb{C}$ or $\mathbb{R}$).[1,2] Here $\mathscr{Q}$, $\mathscr{C}$, and $\mathscr{G}=\{\sigma^\nu:\mathscr{H}_q\to\mathscr{H}_{q^\nu}|q,q^\nu\in\mathscr{Q},\nu\in\mathscr{C}\}$ denote the parameter set, the color set, and the color group, respectively. Hence, by definition, the σ^ν's satisfy Eqs. (I2.1)–(I2.4). The maps $m_q:\mathscr{H}_q\otimes\mathscr{H}_q\to\mathscr{H}_q$, and $\iota_q:k\to\mathscr{H}_q$ are standard multiplication and unit maps, whereas $\Delta_{q,\nu}^{\lambda,\mu}\equiv(\sigma^\lambda\otimes\sigma^\mu)\circ\Delta_q\circ\sigma_\nu:\mathscr{H}_{q^\nu}\to\mathscr{H}_{q^\lambda}\otimes\mathscr{H}_{q^\mu}$, $\epsilon_{q,\nu}\equiv\epsilon_q\circ\sigma_\nu:\mathscr{H}_{q^\nu}\to k$, and $S_{q,\nu}^{\mu}\equiv\sigma^\mu\circ S_q\circ\sigma_\nu:\mathscr{H}_{q^\nu}\to\mathscr{H}_{q^\mu}$, defined in terms of standard comultiplication Δ_q, counit ϵ_q, and antipode S_q, with $\sigma_\nu\equiv(\sigma^\nu)^{-1}$, are called colored comultiplication, counit, and antipode maps, respectively, and satisfy some generalized axioms, stated in Proposition II.2 of I.

Let us now assume that the standard Hopf algebras $\mathscr{H}_q$, belonging to $\mathscr{H}$, have Hopf duals $(\mathscr{H}_q^*,+,\widetilde{m}_q,\widetilde{\iota}_q,\widetilde{\Delta}_q,\widetilde{\epsilon}_q,\widetilde{S}_q;k)$ (or in short $\mathscr{H}_q^*$), and let us set $\mathscr{H}^*=\{\mathscr{H}_q^*|q\in\mathscr{Q}\}$. This means[13] that for any $q\in\mathscr{Q}$, there exists a doubly nondegenerate bilinear form $\langle,\rangle_q$, such that the bialgebra and antipode maps of the dual pair $(\mathscr{H}_q,\mathscr{H}_q^*)$ are related by

$$\langle\widetilde{m}_q(x\otimes y),X\rangle_q=\langle x\otimes y,\Delta_q(X)\rangle_q,\quad \langle\widetilde{\iota}_q(1),X\rangle_q=\langle\widetilde{1}_q,X\rangle_q=\epsilon_q(X), \tag{2.1}$$

$$\langle\widetilde{\Delta}_q(x),X\otimes Y\rangle_q=\langle x,m_q(X\otimes Y)\rangle_q,\quad \widetilde{\epsilon}_q(x)=\langle x,\iota_q(1)\rangle_q=\langle x,1_q\rangle_q, \tag{2.2}$$

$$\langle\widetilde{S}_q(x),X\rangle_q=\langle x,S_q(X)\rangle_q, \tag{2.3}$$

for any $x,y\in\mathscr{H}_q^*$, and any $X,Y\in\mathscr{H}_q$.

Let us next introduce

Definition II.1: Let $\rho^\nu:\mathscr{H}_q^*\to\mathscr{H}_{q^\nu}^*$ *be defined by*

$$\langle \rho^{\nu}(x), \sigma^{\nu}(X) \rangle_{q^{\nu}} = \langle x, X \rangle_{q}, \tag{2.4}$$

for any $x \in \mathscr{H}_q^*$, $X \in \mathscr{H}_q$, $q \in \mathscr{Q}$, *and* $\nu \in \mathscr{C}$.

From the properties of the σ^{ν}'s and the nondegeneracy of $\langle , \rangle_q$, it is then straightforward to show

Proposition II.2: The maps ρ^{ν}, *defined in Eq. (2.4), are coalgebra isomorphisms, and the set* $\{\rho^{\nu} | \nu \in \mathscr{C}\}$ *is a group isomorphic to* $\mathscr{G}$ *(and denoted by the same symbol) with respect to the composition of maps. In other words, the* ρ^{ν}'s *are one-to-one, and satisfy the properties*

$$\widetilde{\Delta}_{q^{\nu}} \circ \rho^{\nu} = (\rho^{\nu} \otimes \rho^{\nu}) \circ \widetilde{\Delta}_q, \quad \widetilde{\epsilon}_{q^{\nu}} \circ \rho^{\nu} = \widetilde{\epsilon}_q, \tag{2.5}$$

and

$$\forall \nu, \nu' \in \mathscr{C}: \rho^{\nu' \circ \nu} = \rho^{\nu'} \circ \rho^{\nu}: \mathscr{H}_q^* \to \mathscr{H}_{q^{\nu' \circ \nu}}^*,$$

$$\rho^{\nu^0} = \mathrm{id}: \mathscr{H}_q^* \to \mathscr{H}_{q^{\nu^0}}^* = \mathscr{H}_q^*, \tag{2.6}$$

$$\forall \nu \in \mathscr{C}: \rho^{\nu^i} = \rho_{\nu} \equiv (\rho^{\nu})^{-1}: \mathscr{H}_{q^{\nu}}^* \to \mathscr{H}_q^*,$$

where $\nu' \circ \nu$, ν^0, *and* ν^i *have the same meaning as in Eqs. (I2.2)–(I2.4).*

Proceeding as for $\mathscr{H}$ in I, we may now combine the definitions of $\mathscr{H}^*$, $\mathscr{C}$, and $\mathscr{G}$ into

Definition II.3: The maps $\widetilde{m}^{\nu}_{q,\lambda,\mu}: \mathscr{H}_{q^{\lambda}}^* \otimes \mathscr{H}_{q^{\mu}}^* \to \mathscr{H}_{q^{\nu}}^*$, $\widetilde{\iota}_q^{\nu}: k \to \mathscr{H}_{q^{\nu}}^*$, *and* $\widetilde{S}^{\nu}_{q,\mu}: \mathscr{H}_{q^{\mu}}^* \to \mathscr{H}_{q^{\nu}}^*$, *defined by*

$$\widetilde{m}^{\nu}_{q,\lambda,\mu} \equiv \rho^{\nu} \circ \widetilde{m}_q \circ (\rho_{\lambda} \otimes \rho_{\mu}), \quad \widetilde{\iota}_q^{\nu} \equiv \rho^{\nu} \circ \widetilde{\iota}_q, \quad \widetilde{S}^{\nu}_{q,\mu} \equiv \rho^{\nu} \circ \widetilde{S}_q \circ \rho_{\mu}, \tag{2.7}$$

for any $q \in \mathscr{Q}$, *and any* λ, μ, $\nu \in \mathscr{C}$, *are called colored multiplication, unit, and antipode, respectively.*

By using the dual pairing (2.1)–(2.3), Definitions II.1 and II.3, as well as Definition II.1 and Proposition II.2 of I, it is easy to establish

Proposition II.4: The colored multiplication, unit, and antipode maps, defined in Eq. (2.7), are dual to the colored comultiplication, counit, and antipode of $\mathscr{H}^c$, *i.e.,*

$$\forall x \in \mathscr{H}_{q^{\lambda}}^*, \forall y \in \mathscr{H}_{q^{\mu}}^*, \forall X \in \mathscr{H}_{q^{\nu}}: \langle \widetilde{m}^{\nu}_{q,\lambda,\mu}(x \otimes y), X \rangle_{q^{\nu}} = \langle x \otimes y, \Delta^{\lambda,\mu}_{q,\nu}(X) \rangle_{q^{\lambda},q^{\mu}},$$

$$\forall X \in \mathscr{H}_{q^{\nu}}: \langle \widetilde{\iota}_q^{\nu}(1), X \rangle_{q^{\nu}} = \epsilon_{q,\nu}(X), \tag{2.8}$$

$$\forall x \in \mathscr{H}_{q^{\mu}}^*, \forall X \in \mathscr{H}_{q^{\nu}}: \langle \widetilde{S}^{\nu}_{q,\mu}(x), X \rangle_{q^{\nu}} = \langle x, S^{\mu}_{q,\nu}(X) \rangle_{q^{\mu}}.$$

In addition, they transform under $\mathscr{G}$ *as*

$$\widetilde{m}^{\nu}_{q,\alpha,\beta} \circ (\rho^{\alpha}_{\lambda} \otimes \rho^{\beta}_{\mu}) = \widetilde{m}^{\nu}_{q,\lambda,\mu} = \rho^{\nu}_{\gamma} \circ \widetilde{m}^{\gamma}_{q,\lambda,\mu},$$

$$\rho^{\nu}_{\alpha} \circ \widetilde{\iota}_q^{\alpha} = \widetilde{\iota}_q^{\nu}, \tag{2.9}$$

$$\widetilde{S}^{\nu}_{q,\alpha} \circ \rho^{\alpha}_{\mu} = \widetilde{S}^{\nu}_{q,\mu} = \rho^{\nu}_{\beta} \circ \widetilde{S}^{\beta}_{q,\mu},$$

and satisfy generalized associativity, unit, and antipode axioms

$$\widetilde{m}^{\nu}_{q,\lambda,\mu} \circ (\widetilde{m}^{\lambda}_{q,\alpha,\beta} \otimes \rho^{\mu}_{\gamma}) = \widetilde{m}^{\nu}_{q,\lambda',\mu'} \circ (\rho^{\lambda'}_{\alpha} \otimes \widetilde{m}^{\mu'}_{q,\beta,\gamma}),$$

$$\widetilde{m}^{\nu}_{q,\lambda,\mu} \circ (\widetilde{\iota}_q^{\lambda} \otimes \rho^{\mu}_{\alpha}) = \widetilde{m}^{\nu}_{q,\lambda',\mu'} \circ (\rho^{\lambda'}_{\alpha} \otimes \widetilde{\iota}_q^{\mu'}) = \rho^{\nu}_{\alpha}, \tag{2.10}$$

$$\widetilde{m}^{\nu}_{q,\lambda,\mu} \circ (\widetilde{S}^{\lambda}_{q,\alpha} \otimes \rho^{\mu}_{\alpha}) \circ \widetilde{\Delta}_{q^{\alpha}} = \widetilde{m}^{\nu}_{q,\lambda',\mu'} \circ (\rho^{\lambda'}_{\alpha} \otimes \widetilde{S}^{\mu'}_{q,\alpha}) \circ \widetilde{\Delta}_{q^{\alpha}} = \widetilde{\iota}_q^{\nu} \circ \widetilde{\epsilon}_{q^{\alpha}},$$

as well as generalized bialgebra axioms

$$\begin{gathered}\widetilde{\Delta}_{q^\nu}\circ\widetilde{m}^{\ \nu}_{q,\lambda,\mu}=(\widetilde{m}^{\ \nu}_{q,\lambda,\mu}\otimes\widetilde{m}^{\ \nu}_{q,\lambda,\mu})\circ(\mathrm{id}\otimes\tau\otimes\mathrm{id})\circ(\widetilde{\Delta}_{q^\lambda}\otimes\widetilde{\Delta}_{q^\mu}),\\ \widetilde{\Delta}_{q^\nu}\circ\widetilde{\iota}^{\nu}_{q}=\widetilde{\iota}^{\nu}_{q}\otimes\widetilde{\iota}^{\nu}_{q},\\ \widetilde{\epsilon}_{q^\nu}\circ\widetilde{m}^{\ \nu}_{q,\lambda,\mu}=\widetilde{\epsilon}_{q^\lambda}\otimes\widetilde{\epsilon}_{q^\mu},\\ \widetilde{\epsilon}_{q^\nu}\circ\widetilde{\iota}^{\nu}_{q}=\mathrm{id},\end{gathered}\tag{2.11}$$

where $\rho^{\lambda}_{\mu}\equiv\rho^{\lambda}\circ\rho_{\mu}$, τ *is the twist map, and no summation is implied over repeated indices.*

Remarks: (1) As usual for Hopf algebras, the twist map is defined by $\tau(x\otimes y)=y\otimes x$ for any x, y belonging to appropriate spaces. (2) As shown in Eqs. (2.10) and (I2.7), the colored antipodes $\widetilde{S}^{\ \nu}_{q,\mu}$, and $S^{\ \mu}_{q,\nu}$ satisfy distinct generalized axioms. A similar remark is valid for the generalized bialgebra structures given in Eqs. (2.11) and (I2.8), respectively.

From Proposition II.4, one obtains

Corollary II.5: For any $q\in\mathcal{Q}$, *any* $\nu\in\mathcal{C}$, *and* $q_\nu\equiv q^{\nu^i}$, $(\mathcal{H}^{\ *}_{q},+,\widetilde{m}^{\ \nu}_{q_\nu,\nu,\nu},\widetilde{\iota}^{\nu}_{q_\nu},\widetilde{\Delta}_{q},\widetilde{\epsilon}_{q},\widetilde{S}^{\ \nu}_{q_\nu,\nu};k)$ *is a Hopf algebra over k with multiplication* $\widetilde{m}^{\ \nu}_{q_\nu,\nu,\nu}$, *unit* $\widetilde{\iota}^{\nu}_{q_\nu}$, *and antipode* $\widetilde{S}^{\ \nu}_{q_\nu,\nu}$, *defined by particularizing Eq. (2.7). Moreover, it is the Hopf dual of the Hopf algebra* $(\mathcal{H}_q,+,m_q,\iota_q,\Delta^{\nu,\nu}_{q_\nu,\nu},\epsilon_{q_\nu,\nu},S^{\ \nu}_{q_\nu,\nu};k)$, *considered in Corollary II.3 of I.*

Remark: In particular, for $\nu=\nu^0$, we get back the original Hopf structures of $\mathcal{H}^{\ *}_{q}$ and $\mathcal{H}_q$.

This result can be generalized as follows:

Definition II.6: If the standard Hopf algebras $\mathcal{H}_q$ *that are the building blocks of a colored Hopf algebra* $\mathcal{H}^c$ *have Hopf duals* $\mathcal{H}^{\ *}_{q}$, *then the set* $\mathcal{H}^*=\{\mathcal{H}^{\ *}_{q}\,|\,q\in\mathcal{Q}\}$, *endowed with colored multiplication, unit, and antipode maps* $\widetilde{m}^{\ \nu}_{q,\lambda,\mu}$, $\widetilde{\iota}^{\nu}_{q}$, $\widetilde{S}^{\ \nu}_{q,\mu}$, *as defined in (2.7), is called colored Hopf dual of* $\mathcal{H}^c$, *and denoted by any one of the symbols* $(\mathcal{H}^{\ *}_{q},+,\widetilde{m}^{\ \nu}_{q,\lambda,\mu},\widetilde{\iota}^{\nu}_{q},\widetilde{\Delta}_{q},\widetilde{\epsilon}_{q},\widetilde{S}^{\ \nu}_{q,\mu};k,\mathcal{Q},\mathcal{C},\mathcal{G})$, $(\mathcal{H}^*,\mathcal{C},\mathcal{G})$, *or* $\mathcal{H}^{c*}$.

The colored antipode $\widetilde{S}^{\ \nu}_{q,\mu}$ satisfies some additional properties, which are again distinct from the corresponding ones of $S^{\ \mu}_{q,\nu}$, given in Proposition II.5 of I.

Proposition II.7: The colored antipode $\widetilde{S}^{\ \nu}_{q,\mu}$ *of a colored Hopf dual* $\mathcal{H}^{c*}$ *fulfills the relations*

$$\begin{gathered}\widetilde{S}^{\ \nu}_{q,\gamma}\circ\widetilde{m}^{\ \gamma}_{q,\alpha,\beta}=\widetilde{m}^{\ \nu}_{q,\mu,\lambda}\circ\tau\circ(\widetilde{S}^{\lambda}_{q,\alpha}\otimes\widetilde{S}^{\mu}_{q,\beta}),\quad \widetilde{S}^{\ \nu}_{q,\mu}\circ\widetilde{\iota}^{\mu}_{q}=\widetilde{\iota}^{\nu}_{q},\\ (\widetilde{S}^{\ \nu}_{q,\mu}\otimes\widetilde{S}^{\ \nu}_{q,\mu})\circ\widetilde{\Delta}_{q^\mu}=\tau\circ\widetilde{\Delta}_{q^\nu}\circ\widetilde{S}^{\ \nu}_{q,\mu},\quad \widetilde{\epsilon}_{q^\nu}\circ\widetilde{S}^{\ \nu}_{q,\mu}=\widetilde{\epsilon}_{q^\mu}.\end{gathered}\tag{2.12}$$

The case where the $\mathcal{H}_q$'s are Hopf superalgebras $(\mathcal{H}_q,+,\gamma_q,m_q,\iota_q,\Delta_q,\epsilon_q,S_q;k)$ (Ref. 7) can be dealt with only some minor changes. Here $\gamma_q:\mathcal{H}_q\to\mathcal{H}_q$ denotes their grading automorphism, i.e.,

$$\gamma_q(X)=(-1)^{\deg X}X,\tag{2.13}$$

for any homogeneous $X\in\mathcal{H}_q$, where $\deg X=0$ or 1 according to whether X is even or odd. As all vector spaces are now graded, the tensor product and the twist map are such that

$$(X\otimes Y)(Z\otimes T)=(-1)^{(\deg Y)(\deg Z)}XZ\otimes YT,\quad \tau(X\otimes Y)=(-1)^{(\deg X)(\deg Y)}Y\otimes X,\tag{2.14}$$

for any homogeneous X, Y, Z, $T\in\mathcal{H}_q$.

Then, according to Definition 3.1 of II, a colored Hopf superalgebra $(\mathcal{H}_q,+,\gamma_q,m_q,\iota_q,\Delta^{\lambda,\mu}_{q,\nu},\epsilon_{q,\nu},S^{\ \mu}_{q,\nu};k,\mathcal{Q},\mathcal{C},\mathcal{G})$ is a colored Hopf algebra, whose building blocks are Hopf superalgebras, and whose color group elements σ^ν are superalgebra isomorphisms. In other words, the σ^ν's satisfy Eqs. (I2.1)–(I2.4), and in addition are grade-preserving maps, i.e.,

$$\sigma^\nu\circ\gamma_q=\gamma_{q^\nu}\circ\sigma^\nu,\tag{2.15}$$

for any $q\in\mathcal{Q}$, and $\nu\in\mathcal{C}$. Such colored Hopf superalgebras, which will still be denoted by $\mathcal{H}^c$, fulfill similar properties as ordinary colored Hopf algebras, provided Eq. (2.14) is taken into

account. In particular, if the $\mathscr{H}_q$'s are quasitriangular, then the colored universal $\mathscr{R}$-matrix $\mathscr{R}^c = \{\mathscr{R}_q^{\lambda,\mu} | q \in \mathscr{Q}, \lambda, \mu \in \mathscr{C}\}$ of $(\mathscr{H}^c, \mathscr{R}^c)$ satisfies the colored graded YBE,

$$R_{q,12}^{\lambda,\mu} R_{q,13}^{\lambda,\nu} R_{q,23}^{\mu,\nu} = R_{q,23}^{\mu,\nu} R_{q,13}^{\lambda,\nu} R_{q,12}^{\lambda,\mu}, \tag{2.16}$$

which has the same form as the nongraded one, but wherein the first relation in Eq. (2.14) is used to evaluate the products.

Turning now to dual structures, let us assume that the Hopf superalgebras $\mathscr{H}_q$ have Hopf duals $(\mathscr{H}_q^*, +, \tilde{\gamma}_q, \tilde{m}_q, \tilde{\iota}_q, \tilde{\Delta}_q, \tilde{\epsilon}_q, \tilde{S}_q; k)$ with corresponding grading maps denoted by $\tilde{\gamma}_q$, i.e.,

$$\tilde{\gamma}_q(x) = (-1)^{\deg x} x, \tag{2.17}$$

for any homogeneous $x \in \mathscr{H}_q^*$, and $\deg x = 0$ or 1 for x even or odd, respectively. The doubly nondegenerate bilinear forms $\langle , \rangle_q$ not only satisfy Eqs. (2.1)–(2.3), but are also consistent or, in other words, fulfill the relation

$$\langle \tilde{\gamma}_q(x), X \rangle_q = \langle x, \gamma_q(X) \rangle_q, \tag{2.18}$$

for any homogeneous $x \in \mathscr{H}_q^*$, and $X \in \mathscr{H}_q$.

It is then straightforward to show

Proposition II.8: In the case of Hopf superalgebras, the maps $\rho^\nu : \mathscr{H}_q^* \to \mathscr{H}_{q^\nu}^*$, *defined in Eq. (2.4), preserve the grading of the* $\mathscr{H}_q^*$'s, *i.e.*,

$$\rho^\nu \circ \tilde{\gamma}_q = \tilde{\gamma}_{q^\nu} \circ \rho^\nu, \tag{2.19}$$

for any $q \in \mathscr{Q}$, $\nu \in \mathscr{C}$, *in addition to satisfying Proposition II.2. Moreover the colored multiplication, unit, and antipode maps* $\tilde{m}_{q,\lambda,\mu}^{\nu}$, $\tilde{\iota}_q^{\nu}$, $\tilde{S}_{q,\mu}^{\nu}$, *defined in Eq. (2.7), satisfy Propositions II.4 and II.7, provided the counterpart of Eq. (2.14) for the duals* $\mathscr{H}_q^*$ *is taken into account.*

Finally, Definition II.6 can be extended to colored Hopf duals of colored Hopf superalgebras with $\mathscr{H}^{c*} = (\mathscr{H}_q^*, +, \tilde{\gamma}_q, \tilde{m}_{q,\lambda,\mu}^{\nu}, \tilde{\iota}_q^{\nu}, \tilde{\Delta}_q, \tilde{\epsilon}_q, \tilde{S}_{q,\mu}^{\nu}; k, \mathscr{Q}, \mathscr{C}, \mathscr{G})$.

III. DUALS OF COLORED QUEA'S OF LIE (SUPER)ALGEBRAS

Whenever the Hopf (super)algebras $\mathscr{H}_q$ are QUEA's of Lie (super)algebras $U_q(g)$, their Hopf duals $\mathscr{H}_q^*$ (if they exist) are Hopf (super)algebras of quantized functions on the corresponding Lie (super)groups, $\mathrm{Fun}_q(G) = G_q$.

Let $\{X_A\}$ and $\{x^A\}$ be dual bases of $\mathscr{H}_q$ and $\mathscr{H}_q^*$, respectively. Hence

$$\langle x^A, X_B \rangle_q = \delta_B^A. \tag{3.1}$$

The universal $\mathscr{T}$-matrix of G_q (Refs. 14,16) is defined as the element of $G_q \otimes U_q(g)$ given by

$$\mathscr{T}_q = \sum_A x^A \otimes X_A. \tag{3.2}$$

In terms of the bases $\{X_A\}$ and $\{x^A\}$, the duality relations (2.2), and (2.8), between $\mathscr{H}^c = U^c(g)$ and $\mathscr{H}^{c*} = G^c$ can be written as

$$m_q(X_A \otimes X_B) = \sum_C f_{AB}^C(q) X_C, \quad 1_q = \iota_q(1) = \sum_A g^A(q) X_A,$$

$$\Delta_{q,\nu}^{\lambda,\mu}(X_A) = \sum_{BC} h_A^{BC}(q^\lambda, q^\mu, q^\nu) X_B \otimes X_C, \quad \epsilon_{q,\nu}(X_A) = \tilde{g}_A(q^\nu), \tag{3.3}$$

$$S_{q,\nu}^{\mu}(X_A) = \sum_B s_A^B(q^\mu, q^\nu) X_B,$$

and

$$\widetilde{m}^{\nu}_{q,\lambda,\mu}(x^B\otimes x^C)=\sum_A h_A^{BC}(q^\lambda,q^\mu,q^\nu)x^A,\quad \widetilde{\iota}^{\nu}_q(1)=\sum_A \widetilde{g}_A(q^\nu)x^A,$$

$$\widetilde{\Delta}_q(x^C)=\sum_{AB} f^C_{AB}(q)x^A\otimes x^B,\quad \widetilde{\epsilon}_q(x^A)=g^A(q), \tag{3.4}$$

$$\widetilde{S}^{\nu}_{q,\mu}(x^B)=\sum_A s^B_A(q^\mu,q^\nu)x^A,$$

where the structure constants for the multiplication (resp. colored comultiplication) in $\mathrm{U}^c(g)$ become the structure constants for the comultiplication (resp. colored multiplication) in G^c, and similarly for the unit and counit, or the antipodes. Note that here we denote by the same symbol basis elements X_A (resp. x^A) belonging to different members $\mathscr{H}_{q^\nu}=\mathrm{U}_{q^\nu}(g)$ (resp. $\mathscr{H}^*_{q^\nu}=G_{q^\nu}$) of the Hopf algebra set $\mathscr{H}$ (resp. $\mathscr{H}^*$) in order not to overload the notation by adding an extra index referring to the corresponding algebra. This should cause no confusion since from the context, it is always clear to which Hopf algebra X_A (resp. x^A) belongs.

Let us now introduce

Definition III.1: The set $\mathscr{T}^c=\{\mathscr{T}^\lambda_q | q\in\mathscr{Q},\lambda\in\mathscr{C}\}$, *where* $\mathscr{T}^\lambda_q\equiv\mathscr{T}_{q^\lambda}$ *is defined in Eq. (3.2), is called the universal* $\mathscr{T}$*-matrix of* G^c.

Definition III.2: For any $q\in\mathscr{Q}$, *and any* $\lambda,\mu,\nu,\alpha,\beta\in\mathscr{C}$, *let* $\mathscr{U}^{\lambda,\mu}_q\in G_{q^\lambda}\otimes U_{q^\mu}(g)$, $\mathscr{A}^{\alpha,\mu}_q\dot{\otimes}\mathscr{B}^{\beta,\mu}_q\in G_{q^\alpha}\otimes G_{q^\beta}\otimes U_{q^\mu}(g)$, *and* $\mathscr{A}^{\lambda,\alpha}_q\overset{\nu}{\ddot{\otimes}}\mathscr{B}^{\mu,\beta}_q\in G_{q^\nu}\otimes U_{q^\alpha}(g)\otimes U_{q^\beta}(g)$ *be defined by*

$$\mathscr{U}^{\lambda,\mu}_q\equiv(\mathrm{id}\otimes S^\mu_{q,\lambda})(\mathscr{T}^\lambda_q),$$

$$\mathscr{A}^{\alpha,\mu}_q\dot{\otimes}\mathscr{B}^{\beta,\mu}_q\equiv\sum_{ABCD} a^C_A b^D_B x^A\otimes x^B\otimes m_{q^\mu}(X_C\otimes X_D), \tag{3.5}$$

$$\mathscr{A}^{\lambda,\alpha}_q\overset{\nu}{\ddot{\otimes}}\mathscr{B}^{\mu,\beta}_q\equiv\sum_{ABCD} a^C_A b^D_B \widetilde{m}^{\nu}_{q,\lambda,\mu}(x^A\otimes x^B)\otimes X_C\otimes X_D,$$

where

$$\mathscr{A}^{\lambda,\mu}_q\equiv\sum_{AC} a^C_A x^A\otimes X_C,\quad \mathscr{B}^{\lambda,\mu}_q\equiv\sum_{BD} b^D_B x^B\otimes X_D, \tag{3.6}$$

with a^C_A, $b^D_B\in k$, *denote any two elements of* $G_{q^\lambda}\otimes U_{q^\mu}(g)$.

From Definitions III.1, III.2, Eqs. (3.3), (3.4), and the generalized antipode axiom for $\widetilde{S}^{\nu}_{q,\mu}$, given in Eq. (2.10), it is straightforward to show

Proposition III.3: The elements of the colored universal $\mathscr{T}$*-matrix of* G^c *satisfy the relations*

$$\mathscr{T}^\lambda_q\dot{\otimes}\mathscr{T}^\lambda_q=(\widetilde{\Delta}_{q^\lambda}\otimes\mathrm{id})(\mathscr{T}^\lambda_q),$$

$$\mathscr{T}^\lambda_q\overset{\nu}{\ddot{\otimes}}\mathscr{T}^\mu_q=(\mathrm{id}\otimes\Delta^{\lambda,\mu}_{q,\nu})(\mathscr{T}^\nu_q),$$

$$\mathscr{U}^{\lambda,\mu}_q=(\widetilde{S}^\lambda_{q,\mu}\otimes\mathrm{id})(\mathscr{T}^\mu_q), \tag{3.7}$$

$$(\widetilde{m}^{\nu}_{q,\lambda,\mu}\otimes\mathrm{id})(\mathscr{T}^\lambda_q\dot{\otimes}\mathscr{U}^{\mu,\lambda}_q)=(\widetilde{m}^{\nu}_{q,\mu',\lambda}\otimes\mathrm{id})(\mathscr{U}^{\mu',\lambda}_q\dot{\otimes}\mathscr{T}^\lambda_q)=\widetilde{\iota}^\nu_q(1)\otimes 1_{q^\lambda},$$

$$(\mathrm{id}\otimes m_{q^\lambda})(\mathscr{T}^\lambda_q\overset{\nu}{\ddot{\otimes}}\mathscr{U}^{\mu,\lambda}_q)=(\mathrm{id}\otimes m_{q^\lambda})(\mathscr{U}^{\mu',\lambda}_q\overset{\nu}{\ddot{\otimes}}\mathscr{T}^\lambda_q)=\widetilde{\iota}^\nu_q(1)\otimes 1_{q^\lambda}.$$

Remark: The first relation in Eq. (3.7) is a well-known property of the universal $\mathscr{T}$-matrix of G_{q^λ},[16] but the remaining relations are new. Whenever the color parameters reduce to that characterizing the unit element of $\mathscr{G}$, they go over into corresponding relations for the universal $\mathscr{T}$-matrix, with $\mathscr{U}_q^{\lambda,\mu}$ becoming the element of $G_q \otimes \mathrm{U}_q(g)$ denoted as $\mathscr{T}_q^{-1}$ in Ref. 16.

With the purpose of generalizing the algebraic formulation of the RTT-relations,[16] let us introduce the colored counterparts of $\mathscr{T}_{q1} \equiv \Sigma_A x^A \otimes (X_A \otimes 1_q)$, and $\mathscr{T}_{q2} \equiv \Sigma_A x^A \otimes (1_q \otimes X_A)$.

Definition III.4: Let $\mathscr{T}_{q,1\mu}^{\lambda} \in G_{q^\lambda} \otimes U_{q^\lambda}(g) \otimes U_{q^\mu}(g)$, *and* $\mathscr{T}_{q,2\lambda}^{\mu} \in G_{q^\mu} \otimes U_{q^\lambda}(g) \otimes U_{q^\mu}(g)$ *be defined by*

$$\mathscr{T}_{q,1\mu}^{\lambda} \equiv \sum_A x^A \otimes (X_A \otimes 1_{q^\mu}), \quad \mathscr{T}_{q,2\lambda}^{\mu} \equiv \sum_B x^B \otimes (1_{q^\lambda} \otimes X_B), \tag{3.8}$$

where $\mathscr{T}_q^{\lambda} = \Sigma_A x^A \otimes X_A$, *and* $\mathscr{T}_q^{\mu} = \Sigma_B x^B \otimes X_B$ *are the universal* $\mathscr{T}$*-matrices of* G_{q^λ} *and* G_{q^μ}, *respectively.*

The searched for result is expressed in

Proposition III.5: If the colored Hopf (super)algebra $\mathrm{U}^c(g)$ *has a colored universal* $\mathscr{R}$*-matrix, then the colored universal* $\mathscr{T}$*-matrix of the colored Hopf dual* G^c *satisfies the relation*

$$(\tilde{1}_{q^\nu} \otimes \mathscr{R}_q^{\lambda,\mu})(\mathscr{T}_{q,1\mu}^{\lambda} \overset{\nu}{\cdot} \mathscr{T}_{q,2\lambda}^{\mu}) = (\mathscr{T}_{q,2\lambda}^{\mu} \overset{\nu}{\cdot} \mathscr{T}_{q,1\mu}^{\lambda})(\tilde{1}_{q^\nu} \otimes \mathscr{R}_q^{\lambda,\mu}), \tag{3.9}$$

where $\overset{\nu}{\cdot}$ *denotes a generalized multiplication such that*

$$\mathscr{T}_{q,1\mu}^{\lambda} \overset{\nu}{\cdot} \mathscr{T}_{q,2\lambda}^{\mu} \equiv \sum_{AB} \tilde{m}_{q,\lambda,\mu}^{\nu}(x^A \otimes x^B) \otimes (X_A \otimes X_B), \tag{3.10}$$

$$\mathscr{T}_{q,2\lambda}^{\mu} \overset{\nu}{\cdot} \mathscr{T}_{q,1\mu}^{\lambda} \equiv \sum_{AB} \tilde{m}_{q,\mu,\lambda}^{\nu}(x^B \otimes x^A) \otimes (X_A \otimes X_B). \tag{3.11}$$

Proof: Comparing Eq. (3.10) with the last relation in Eq. (3.5), and taking Proposition III.3 into account, we successively get

$$\mathscr{T}_{q,1\mu}^{\lambda} \overset{\nu}{\cdot} \mathscr{T}_{q,2\lambda}^{\mu} = \mathscr{T}_q^{\lambda} \overset{\nu}{\ddot{\otimes}} \mathscr{T}_q^{\mu} = (\mathrm{id} \otimes \Delta_{q,\nu}^{\lambda,\mu})(\mathscr{T}_q^{\nu}). \tag{3.12}$$

Moreover, from Eqs. (3.10)–(3.12), we also obtain

$$\mathscr{T}_{q,2\lambda}^{\mu} \overset{\nu}{\cdot} \mathscr{T}_{q,1\mu}^{\lambda} = (\mathrm{id} \otimes \tau)(\mathscr{T}_{q,1\lambda}^{\mu} \overset{\nu}{\cdot} \mathscr{T}_{q,2\mu}^{\lambda}) = (\mathrm{id} \otimes \tau \circ \Delta_{q,\nu}^{\mu,\lambda})(\mathscr{T}_q^{\nu}). \tag{3.13}$$

Equation (3.9) then directly follows from Eq. (I2.15), expressing that $\mathrm{U}^c(g)$ is almost cocommutative. ∎

To explicitly show that Proposition III.5 provides an algebraic formulation of the colored RTT-relations, introduced by Basu-Mallick,[12] one has to consider matrix representations of Eq. (3.9). For such a purpose, let us denote by $D_q^{(i)}$ matrix representations of $U_q(g)$ in some n_i-dimensional k-modules $V_q^{(i)}$, where index i distinguishes between inequivalent representations. We may now represent the elements of the colored universal $\mathscr{R}$- and $\mathscr{T}$-matrices by

$$R_q^{\lambda(i),\mu(j)} \equiv (D_{q^\lambda}^{(i)} \otimes D_{q^\mu}^{(j)})(\mathscr{R}_q^{\lambda,\mu}), \quad T_q^{\lambda(i)} \equiv \sum_A x^A D_{q^\lambda}^{(i)}(X_A), \tag{3.14}$$

which are $n_i n_j \times n_i n_j$ and $n_i \times n_i$ matrices, valued in k and G_{q^λ}, respectively. Such definitions can be extended to provide representations of the operators in Eqs. (3.8), (3.10), and (3.11), as $n_i n_j \times n_i n_j$ matrices valued in G_{q^λ}, G_{q^μ}, G_{q^ν}, and G_{q^ν}, respectively,

$$T_{q,1\mu(j)}^{\lambda(i)} \equiv \sum_A x^A (D_{q^\lambda}^{(i)}(X_A) \otimes I^{(j)}),$$

$$T^{\mu(j)}_{q,2\lambda(i)} \equiv \sum_B x^B (I^{(i)} \otimes D^{(j)}_{q^\mu}(X_B)),$$

$$T^{\lambda(i)}_{q,1\mu(j)} \overset{\nu}{\cdot} T^{\mu(j)}_{q,2\lambda(i)} \equiv \sum_{AB} \widetilde{m}^{\nu}_{q,\lambda,\mu} (x^A \otimes x^B)(D^{(i)}_{q^\lambda}(X_A) \otimes D^{(j)}_{q^\mu}(X_B)), \tag{3.15}$$

$$T^{\mu(j)}_{q,2\lambda(i)} \overset{\nu}{\cdot} T^{\lambda(i)}_{q,1\mu(j)} \equiv \sum_{AB} \widetilde{m}^{\nu}_{q,\mu,\lambda} (x^B \otimes x^A)(D^{(i)}_{q^\lambda}(X_A) \otimes D^{(j)}_{q^\mu}(X_B)).$$

Here $I^{(i)}$, and $I^{(j)}$ denote $n_i \times n_i$, and $n_j \times n_j$ unit matrices.

In the case of QUEA's of Lie superalgebras, since the $D^{(i)}_q$'s are graded matrices, their tensor product has to be evaluated in conformity with Eq. (2.14). According to the usual convention,[7] matrix multiplication of graded matrices is assumed to be the same as that of nongraded ones, whereas the tensor product of two graded matrices is defined by

$$(A \otimes B)_{ij,kl} \equiv (-1)^{\pi_k(\pi_j + \pi_l)} A_{i,k} B_{j,l}. \tag{3.16}$$

Here π_k denotes the $\mathbb{Z}_2$-grade of the kth row or column of matrix A or B. In particular, if A is homogeneous, its degree is given by $\deg A = \pi_k + \pi_l$ for any row index k, and any column index l. Special cases of Eq. (3.16) are

$$(A \otimes I)_{ij,kl} = A_{i,k} \delta_{j,l}, \quad (I \otimes A)_{ij,kl} = (-1)^{\pi_i(\pi_j + \pi_l)} \delta_{i,k} A_{j,l}. \tag{3.17}$$

From Proposition III.5, it is now straightforward to obtain

Corollary III.6: In the representation $D^{(i)}_{q^\lambda} \otimes D^{(j)}_{q^\mu}$ of $U_{q^\lambda}(g) \otimes U_{q^\mu}(g)$, Eq. (3.9) becomes

$$R^{\lambda(i),\mu(j)}_q (T^{\lambda(i)}_{q,1\mu(j)} \overset{\nu}{\cdot} T^{\mu(j)}_{q,2\lambda(i)}) = (T^{\mu(j)}_{q,2\lambda(i)} \overset{\nu}{\cdot} T^{\lambda(i)}_{q,1\mu(j)}) R^{\lambda(i),\mu(j)}_q, \tag{3.18}$$

where the various matrices are defined in Eqs. (3.14), and (3.15).

Remark: If we consider, in particular, the defining representations of $U_{q^\lambda}(g)$ and $U_{q^\mu}(g)$, Eq. (3.18) can be rewritten in a simplified notation as

$$R^{\lambda,\mu}_q (T^{\lambda}_{q1} \overset{\nu}{\cdot} T^{\mu}_{q2}) = (T^{\mu}_{q2} \overset{\nu}{\cdot} T^{\lambda}_{q1}) R^{\lambda,\mu}_q. \tag{3.19}$$

This relation is formally identical with the colored RTT-relations,

$$R^{\lambda,\mu}_q T^{\lambda}_{q1} T^{\mu}_{q2} = T^{\mu}_{q2} T^{\lambda}_{q1} R^{\lambda,\mu}_q, \tag{3.20}$$

introduced by Basu-Mallick.[12] The dependence upon ν in Eq. (3.19), which is absent in Eq. (3.20), only means that we may evaluate the latter in various dual Hopf algebras G_{q^ν}, with ν running over $\mathscr{C}$.

In the next two sections, we shall proceed to illustrate the new concepts and results presented here on two simple, but nevertheless significant examples.

IV. THE COLORED TWO-PARAMETER QUEA $U^c(gl(2))$ AND ITS DUAL $Gl^c(2)$

In I, we constructed a colored QUEA by starting from the two-parameter deformation of U(gl(2)) (Refs. 22,23) in the Burdík and Hellinger formulation.[24] On the other hand, Fronsdal and Galindo[14] analyzed the duality relationship between the pair of Hopf algebras $U_{p,q}(gl(2))$ and $Gl_{p,q}(2)$, and constructed the universal $\mathscr{T}$-matrix of the latter. It is therefore interesting to study how this duality picture can be extended to the colored context. For such a purpose, it will prove convenient to use the approach to $U_{p,q}(gl(2))$ considered in Refs. 14,20, instead of that of Ref. 24. The resulting colored QUEA will therefore differ from that constructed in I. One should remember in this respect that there exist many ways of transforming a given Hopf algebra into a colored one, depending upon the set of generators and the color group that are used.

A. The dual Hopf algebras $U_{p,q}(gl(2))$ and $Gl_{p,q}(2)$

In the standard formulation,[20,22] the quantum algebra $U_{p,q}(gl(2))$ is defined as the algebra generated by the elements $\{J_0, J_\pm, Z\}$, subject to the relations

$$[J_0, J_\pm] = \pm J_\pm, \quad [J_+, J_-] = [2J_0]_Q, \quad [Z, J_0] = [Z, J_\pm] = 0, \tag{4.1}$$

where as usual

$$[X]_t \equiv \frac{t^X - t^{-X}}{t - t^{-1}}, \tag{4.2}$$

and the following combinations of p, q are considered:

$$Q \equiv \sqrt{pq}, \quad P \equiv \sqrt{p/q}. \tag{4.3}$$

Whereas the algebraic relations (4.1) only depend upon the first of these combined parameters, the coalgebraic structure depends upon both of them.

To describe the duality relationship between $U_{p,q}(gl(2))$ and $Gl_{p,q}(2)$, one has to consider another set of $U_{p,q}(gl(2))$ generators $\{\hat{J}_0, \hat{J}_\pm, \hat{Z}\}$, defined in terms of the first one by [20]

$$\hat{J}_0 = J_0, \quad \hat{J}_\pm = J_\pm Q^{\mp J_0 - 1/2} P^{Z - 1/2}, \quad \hat{Z} = Z, \tag{4.4}$$

and satisfying the commutation relations

$$[\hat{J}_0, \hat{J}_\pm] = \pm \hat{J}_\pm, \quad [\hat{J}_+, \hat{J}_-] = P^{2\hat{Z}-1}[2\hat{J}_0]_Q, \quad [\hat{Z}, \hat{J}_0] = [\hat{Z}, \hat{J}_\pm] = 0. \tag{4.5}$$

The remaining Hopf algebra maps are

$$\begin{gathered}
\Delta_{p,q}(\hat{J}_0) = \hat{J}_0 \otimes 1 + 1 \otimes \hat{J}_0, \quad \Delta_{p,q}(\hat{Z}) = \hat{Z} \otimes 1 + 1 \otimes \hat{Z}, \\
\Delta_{p,q}(\hat{J}_+) = \hat{J}_+ \otimes Q^{-2\hat{J}_0} P^{2\hat{Z}} + 1 \otimes \hat{J}_+, \quad \Delta_{p,q}(\hat{J}_-) = \hat{J}_- \otimes 1 + Q^{2\hat{J}_0} P^{2\hat{Z}} \otimes \hat{J}_-, \\
\epsilon_{p,q}(X) = 0, \quad X \in \{\hat{J}_0, \hat{J}_\pm, \hat{Z}\}, \\
S_{p,q}(\hat{J}_0) = -\hat{J}_0, \quad S_{p,q}(\hat{Z}) = -\hat{Z}, \\
S_{p,q}(\hat{J}_+) = -\hat{J}_+ Q^{2\hat{J}_0} P^{-2\hat{Z}}, \quad S_{p,q}(\hat{J}_-) = -Q^{-2\hat{J}_0} P^{-2\hat{Z}} \hat{J}_-,
\end{gathered} \tag{4.6}$$

and the universal $\mathscr{R}$-matrix can be written as[20,25]

$$\mathscr{R}_{p,q} = Q^{-2\hat{J}_0 \otimes \hat{J}_0} P^{2(\hat{Z} \otimes \hat{J}_0 - \hat{J}_0 \otimes \hat{Z})} \sum_{n=0}^{\infty} \frac{(1 - Q^2)^n P^n}{[n]_Q!} Q^{-n(n+1)/2} \hat{J}_+^{\,n} \otimes \hat{J}_-^{\,n}, \tag{4.7}$$

where $[n]_Q! \equiv [n]_Q [n-1]_Q \cdots [1]_Q$ for $n \in \mathbb{N}^+$, and $[0]_Q! \equiv 1$.

In the 2×2 defining representation $D_{p,q}$ of $U_{p,q}(gl(2))$, given by

$$\begin{gathered}
D_{p,q}(\hat{J}_0) = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad D_{p,q}(\hat{Z}) = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\
D_{p,q}(\hat{J}_+) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad D_{p,q}(\hat{J}_-) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},
\end{gathered} \tag{4.8}$$

the universal $\mathscr{R}$-matrix (4.7) is represented by the 4×4 matrix

$$R_{p,q}\equiv(D_{p,q}\otimes D_{p,q})(\mathscr{R}_{p,q})=Q^{1/2}\begin{pmatrix} Q^{-1} & 0 & 0 & 0\\ 0 & P^{-1} & Q^{-1}-Q & 0\\ 0 & 0 & P & 0\\ 0 & 0 & 0 & Q^{-1}\end{pmatrix}. \tag{4.9}$$

For the corresponding quantum group $\mathrm{Gl}_{p,q}(2)$, the defining T-matrix is specified by[14,20]

$$T_{p,q}=\begin{pmatrix} a & b\\ c & d\end{pmatrix}, \tag{4.10}$$

with the commutation relations

$$\begin{gathered} ab=qba,\quad ac=pca,\quad bd=pdb,\quad cd=qdc,\\ bc=(p/q)cb,\quad ad-da=(q-p^{-1})bc,\end{gathered} \tag{4.11}$$

following from the RTT-relations corresponding to the R-matrix (4.9). The coalgebra maps are given by

$$\widetilde{\Delta}_{p,q}(T_{p,q})=T_{p,q}\dot{\otimes}T_{p,q}=\begin{pmatrix} a\otimes a+b\otimes c & a\otimes b+b\otimes d\\ c\otimes a+d\otimes c & c\otimes b+d\otimes d\end{pmatrix},\quad \widetilde{\epsilon}_{p,q}(T_{p,q})=I. \tag{4.12}$$

The quantum determinant, defined by

$$\mathscr{D}\equiv ad-qbc=ad-pcb=da-p^{-1}bc=da-q^{-1}cb, \tag{4.13}$$

is a group-like element, i.e.,

$$\widetilde{\Delta}_{p,q}(\mathscr{D})=\mathscr{D}\otimes\mathscr{D},\quad \widetilde{\epsilon}_{p,q}(\mathscr{D})=1, \tag{4.14}$$

but is not central (except for $p=q$), since

$$\mathscr{D}a=a\mathscr{D},\quad \mathscr{D}b=P^{-2}b\mathscr{D},\quad \mathscr{D}c=P^{2}c\mathscr{D},\quad \mathscr{D}d=d\mathscr{D}. \tag{4.15}$$

With the assumption that $\mathscr{D}$ is invertible, $\mathrm{Gl}_{p,q}(2)$ is endowed with an antipode map

$$\widetilde{S}_{p,q}(T_{p,q})=(T_{p,q})^{-1}=\mathscr{D}^{-1}\begin{pmatrix} d & -p^{-1}b\\ -pc & a\end{pmatrix}=\begin{pmatrix} d & -q^{-1}b\\ -qc & a\end{pmatrix}\mathscr{D}^{-1}, \tag{4.16}$$

such that $\widetilde{S}_{p,q}(\mathscr{D})=\mathscr{D}^{-1}$. This completes the Hopf algebraic structure of $\mathrm{Gl}_{p,q}(2)$.

Following Fronsdal and Galindo,[14] one may consider a Gauss decomposition of the T-matrix (4.10),

$$T_{p,q}=\begin{pmatrix} 1 & 0\\ \gamma & 1\end{pmatrix}\begin{pmatrix} a & 0\\ 0 & \hat{d}\end{pmatrix}\begin{pmatrix} 1 & \beta\\ 0 & 1\end{pmatrix}, \tag{4.17}$$

where a and $\hat{d}$ are assumed invertible, thence $\beta=a^{-1}b$, $\gamma=ca^{-1}$, $\hat{d}=d-ca^{-1}b$. By assuming that the algebra can be augmented with the logarithms of a and $\hat{d}$, and by using the maps

$$a=e^{\alpha},\quad \hat{d}=e^{-\delta}, \tag{4.18}$$

and the reparametrization

$$P=e^{\theta},\quad Q=e^{\varphi}, \tag{4.19}$$

the quantum determinant (4.13) is transformed into $\mathscr{D}=\exp(\alpha-\delta)$, and the new variables $\{\alpha,\beta,\gamma,\delta\}$ are seen to satisfy the commutation relations of a solvable Lie algebra

$$[\alpha,\beta]=(\varphi-\theta)\beta, \quad [\alpha,\gamma]=(\varphi+\theta)\gamma, \quad [\alpha,\delta]=0,$$
$$[\delta,\beta]=(\varphi+\theta)\beta, \quad [\delta,\gamma]=(\varphi-\theta)\gamma, \quad [\beta,\gamma]=0, \tag{4.20}$$

with a noncocommutatitve coproduct structure. As a consequence, $\mathrm{Gl}_{p,q}(2)$ can be embedded into the enveloping algebra $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ of such a Lie algebra. As shown by Fronsdal and Galindo,[14] $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ is dual to the QUEA $\mathrm{U}_{p,q}(\mathrm{gl}(2))$. Following common practice, we shall mean the larger enveloping algebra $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ whenever speaking of the duality between $\mathrm{U}_{p,q}(\mathrm{gl}(2))$ and $\mathrm{Gl}_{p,q}(2)$.

Dual bases of $\mathrm{U}_{p,q}(\mathrm{gl}(2))$ and $\mathrm{Gl}_{p,q}(2)$ are given by[14]

$$X_A=\frac{Q^{a_1(a_1-1)/2}\hat{J}_-^{a_1}}{[a_1]_Q!}\frac{H^{a_2}}{a_2!}\frac{\widetilde{H}^{a_3}}{a_3!}\frac{Q^{-a_4(a_4-1)/2}\hat{J}_+^{a_4}}{[a_4]_Q!}, \tag{4.21}$$

and

$$x^A=\gamma^{a_1}\alpha^{a_2}\delta^{a_3}\beta^{a_4}, \tag{4.22}$$

respectively, where

$$H\equiv\hat{J}_0+\hat{Z}, \quad \widetilde{H}\equiv\hat{J}_0-\hat{Z}, \tag{4.23}$$

$A=(a_1,a_2,a_3,a_4)$, and a_1, a_2, a_3, $a_4\in\mathbb{N}$. From Eqs. (3.2), (4.21), and (4.22), the universal $\mathscr{T}$-matrix of $\mathrm{Gl}_{p,q}(2)$ can be written as

$$\mathscr{T}_{p,q}=\mathscr{E}xp_{Q^{-2}}(\gamma\otimes\hat{J}_-)\exp(\alpha\otimes H+\delta\otimes\widetilde{H})\mathscr{E}xp_{Q^2}(\beta\otimes\hat{J}_+), \tag{4.24}$$

in terms of the basic exponential function

$$\mathscr{E}xp_{t^2}(z)=\sum_{n=0}^{\infty}\frac{t^{-n(n-1)/2}}{[n]_t!}z^t. \tag{4.25}$$

In the next subsection, it will prove convenient to consider another set of dual bases

$$\overline{X_A}=\sum_B c_A^B X_B=\frac{Q^{a_1(a_1-1)/2}\hat{J}_-^{a_1}}{[a_1]_Q!}\frac{\hat{J}_0^{a_2}}{a_2!}\frac{\hat{Z}^{a_3}}{a_3!}\frac{Q^{-a_4(a_4-1)/2}\hat{J}_+^{a_4}}{[a_4]_Q!}, \tag{4.26}$$

$$\overline{x^A}=\sum_B (c^{-1})_B^A x^B=\gamma^{a_1}h^{a_2}\widetilde{h}^{a_3}\beta^{a_4}. \tag{4.27}$$

Here, A has the same meaning as in Eqs. (4.21), (4.22),

$$h\equiv\alpha+\delta, \quad \widetilde{h}\equiv\alpha-\delta, \tag{4.28}$$

and

$$c_A^B=2^{-a_2-a_3}(c^{-1})_A^B=\delta_{a_1}^{b_1}\delta_{a_2+a_3}^{b_2+b_3}\delta_{a_4}^{b_4}2^{-a_2-a_3}\sum_{t=\max(0,b_2-a_2)}^{\min(b_2,a_3)}(-1)^{a_3-t}\binom{b_2}{t}\binom{b_3}{a_3-t}, \tag{4.29}$$

where $\binom{s}{t}$ denotes a standard binomial coefficient. In terms of such dual bases, the universal $\mathscr{T}$-matrix (4.24) can be written as

$$\mathscr{T}_{p,q}=\mathscr{E}xp_{Q^{-2}}(\gamma\otimes\hat{J}_-)\exp(h\otimes\hat{J}_0+\widetilde{h}\otimes\hat{Z})\mathscr{E}xp_{Q^2}(\beta\otimes\hat{J}_+). \tag{4.30}$$

B. The colored Hopf algebra $U^c(gl(2))$ and its dual $Gl^c(2)$

The defining relations (4.5) of the $U_{p,q}(gl(2))$ algebra are left invariant under the transformations

$$\sigma^\nu(\hat{J}_0)=\hat{J}_0, \quad \sigma^\nu(\hat{J}_\pm)=P^{(\nu-1)/2}\hat{J}_\pm, \quad \sigma^\nu(\hat{Z})=\nu\hat{Z}, \tag{4.31}$$

where $\nu\in\mathscr{C}=\mathbb{C}\backslash\{0\}$, provided P is replaced by its νth power, while Q is left unchanged, or equivalently[26]

$$(p,q)\rightarrow(p^{(\nu)},q^{(\nu)})=(p^{(1+\nu)/2}q^{(1-\nu)/2},p^{(1-\nu)/2}q^{(1+\nu)/2}). \tag{4.32}$$

Hence, the σ^ν's are isomorphic mappings between two $U_{p,q}(gl(2))$ algebras with different parameters, as given in Eq. (4.32), and they define a color group $\mathscr{G}=Gl(1,\mathbb{C})$, since $\nu'\circ\nu=\nu'\nu$, $\nu^0=1$, and $\nu^i=\nu^{-1}$.

By choosing the parameter set $Q=(\mathbb{C}\backslash\{0\})\times(\mathbb{C}\backslash\{0\})$, we therefore obtain a colored quasitriangular Hopf algebra $U^c(gl(2))$, whose colored comultiplication, counit, antipode, and $\mathscr{R}$-matrix are given by

$$\Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_0)=\hat{J}_0\otimes 1+1\otimes\hat{J}_0, \quad \Delta^{\lambda,\mu}_{p,q,\nu}(\hat{Z})=\frac{\lambda}{\nu}\hat{Z}\otimes 1+\frac{\mu}{\nu}1\otimes\hat{Z},$$

$$\Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_+)=P^{(\lambda-\nu)/2}\hat{J}_+\otimes Q^{-2\hat{J}_0}P^{2\mu\hat{Z}}+P^{(\mu-\nu)/2}1\otimes\hat{J}_+,$$

$$\Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_-)=P^{(\lambda-\nu)/2}\hat{J}_-\otimes 1+P^{(\mu-\nu)/2}Q^{2\hat{J}_0}P^{2\lambda\hat{Z}}\otimes\hat{J}_-,$$

$$\epsilon_{p,q,\nu}(X)=0, \quad X\in\{\hat{J}_0,\hat{J}_\pm,\hat{Z}\},$$

$$S^{\mu}_{p,q,\nu}(\hat{J}_0)=-\hat{J}_0, \quad S^{\mu}_{p,q,\nu}(\hat{J}_+)=-P^{(\mu-\nu)/2}\hat{J}_+Q^{2\hat{J}_0}P^{-2\mu\hat{Z}}, \tag{4.33}$$

$$S^{\mu}_{p,q,\nu}(\hat{J}_-)=-P^{(\mu-\nu)/2}Q^{-2\hat{J}_0}P^{-2\mu\hat{Z}}\hat{J}_-, \quad S^{\mu}_{p,q,\nu}(\hat{Z})=-\frac{\mu}{\nu}\hat{Z},$$

$$\mathscr{R}^{\lambda,\mu}_{p,q}=Q^{-2\hat{J}_0\otimes\hat{J}_0}P^{2(\lambda\hat{Z}\otimes\hat{J}_0-\mu\hat{J}_0\otimes\hat{Z})}\sum_{n=0}^{\infty}\frac{(1-Q^2)^n}{[n]_Q!}Q^{-n(n+1)/2}P^{(\lambda+\mu)n/2}\hat{J}_+^n\otimes\hat{J}_-^n,$$

respectively.

In the defining representation (4.8), the colored universal $\mathscr{R}$-matrix is represented by the 4×4 matrix,

$$R^{\lambda,\mu}_{p,q}=Q^{1/2}\begin{pmatrix} Q^{-1}P^{(\lambda-\mu)/2} & 0 & 0 & 0 \\ 0 & P^{-(\lambda+\mu)/2} & Q^{-1}-Q & 0 \\ 0 & 0 & P^{(\lambda+\mu)/2} & 0 \\ 0 & 0 & 0 & Q^{-1}P^{-(\lambda-\mu)/2} \end{pmatrix}, \tag{4.34}$$

which is a matrix solution of the colored YBE (I1.5), and gives back matrix (4.9) whenever λ, $\mu\rightarrow1$.

Under transformation (4.31), the basis elements $\overline{X_A}$ of $U_{p,q}(gl(2))$, defined in Eq. (4.26), become

$$\sigma^\nu(\overline{X_A})=P^{(\nu-1)(a_1+a_4)/2}\nu^{a_3}\overline{X_A}. \tag{4.35}$$

Hence, from Eqs. (2.4) and (3.1), it follows that

$$\rho^\nu(\overline{x^A})=P^{(1-\nu)(a_1+a_4)/2}\nu^{-a_3}\overline{x^A} \tag{4.36}$$

defines an isomorphic mapping between the $\mathrm{Gl}_{p,q}(2)$ coalgebras whose parameters are given in Eq. (4.32). By taking Eqs. (4.27) and (4.28) into account, and by summing both sides of Eq. (4.36) over a_2 and a_3, we also obtain the relation

$$\rho^{\nu}(\gamma^{a_1}e^{r(\alpha+\delta)}e^{s(\alpha-\delta)}\beta^{a_4})=P^{(1-\nu)(a_1+a_4)/2}\gamma^{a_1}e^{r(\alpha+\delta)}e^{s(\alpha-\delta)/\nu}\beta^{a_4}, \tag{4.37}$$

valid for any a_1, $a_4\in\mathbb{N}$, and any r, $s\in\mathbb{C}$.

The action of ρ^{ν}, and of its inverse ρ_{ν}, on α, β, γ, δ,

$$\begin{aligned}
&\rho^{\nu}(\alpha)=\frac{\nu+1}{2\nu}\,\alpha+\frac{\nu-1}{2\nu}\,\delta, \quad \rho^{\nu}(\beta)=P^{(1-\nu)/2}\beta,\\
&\rho^{\nu}(\gamma)=P^{(1-\nu)/2}\gamma, \quad \rho^{\nu}(\delta)=\frac{\nu-1}{2\nu}\,\alpha+\frac{\nu+1}{2\nu}\,\delta,
\end{aligned} \tag{4.38}$$

$$\begin{aligned}
&\rho_{\nu}(\alpha)=\frac{1+\nu}{2}\,\alpha+\frac{1-\nu}{2}\,\delta, \quad \rho_{\nu}(\beta)=P^{(\nu-1)/2}\beta,\\
&\rho_{\nu}(\gamma)=P^{(\nu-1)/2}\gamma, \quad \rho_{\nu}(\delta)=\frac{1-\nu}{2}\,\alpha+\frac{1+\nu}{2}\,\delta,
\end{aligned} \tag{4.39}$$

and on a, b, c, d, $\mathscr{D}$,

$$\begin{aligned}
&\rho^{\nu}(a)=\mathscr{D}^{(1-\nu)/(2\nu)}a, \quad \rho^{\nu}(b)=(P^{\nu}\mathscr{D})^{(1-\nu)/(2\nu)}b,\\
&\rho^{\nu}(c)=(P^{-\nu}\mathscr{D})^{(1-\nu)/(2\nu)}c, \quad \rho^{\nu}(d)=\mathscr{D}^{(1-\nu)/(2\nu)}d, \quad \rho^{\nu}(\mathscr{D})=\mathscr{D}^{1/\nu},
\end{aligned} \tag{4.40}$$

$$\begin{aligned}
&\rho_{\nu}(a)=\mathscr{D}^{(\nu-1)/2}a, \quad \rho_{\nu}(b)=(P\mathscr{D})^{(\nu-1)/2}b,\\
&\rho_{\nu}(c)=(P^{-1}\mathscr{D})^{(\nu-1)/2}c, \quad \rho_{\nu}(d)=\mathscr{D}^{(\nu-1)/2}d, \quad \rho_{\nu}(\mathscr{D})=\mathscr{D}^{\nu},
\end{aligned} \tag{4.41}$$

can be obtained as special cases of Eqs. (4.36), and (4.37), respectively.[27]

By using Eqs. (2.7), (4.11), (4.15), (4.16), (4.17), (4.18), (4.37), and (4.41), it is straightforward to determine the colored maps of $\mathrm{Gl}^{c}(2)$. The results are listed in Appendix A.

We shall now proceed to show that the relations obtained for $\widetilde{m}^{\nu}_{p,q,\lambda,\mu}$ can be rewritten in an alternative way. For such a purpose, in analogy with Eq. (3.15), let us define

$$x(\lambda)\overset{\nu}{\cdot}y(\mu)\equiv\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes y), \tag{4.42}$$

where to avoid ambiguity, we specify the algebra to which each element belongs by the corresponding color parameter, e.g., $x(\lambda)\in\mathrm{Gl}_{p^{(\lambda)},q^{(\lambda)}}(2)$.

The 16 relations expressing $\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes y)$, x, $y\in\{a,b,c,d\}$, as elements of $\mathrm{Gl}_{p^{(\nu)},q^{(\nu)}}(2)$, and given in Eqs. (A1), (A5), can be combined into two different sets. The first one contains the relations

$$\begin{aligned}
&a(\lambda)\overset{\nu}{\cdot}b(\mu)=q^{(\lambda)}b(\mu)\overset{\nu}{\cdot}a(\lambda), \quad a(\lambda)\overset{\nu}{\cdot}c(\mu)=p^{(\lambda)}c(\mu)\overset{\nu}{\cdot}a(\lambda),\\
&b(\lambda)\overset{\nu}{\cdot}d(\mu)=p^{(\mu)}d(\mu)\overset{\nu}{\cdot}a(\lambda), \quad c(\lambda)\overset{\nu}{\cdot}d(\mu)=q^{(\mu)}d(\mu)\overset{\nu}{\cdot}c(\lambda),\\
&b(\lambda)\overset{\nu}{\cdot}c(\mu)=\left(\frac{p^{(\lambda)}p^{(\mu)}}{q^{(\lambda)}q^{(\mu)}}\right)^{1/2}c(\mu)\overset{\nu}{\cdot}b(\lambda),
\end{aligned} \tag{4.43}$$

$$a(\lambda)\overset{\nu}{\cdot}d(\mu)-d(\mu)\overset{\nu}{\cdot}a(\lambda)=\{(q^{(\lambda)}q^{(\mu)})^{1/2}-(p^{(\lambda)}p^{(\mu)})^{-1/2}\}b(\lambda)\overset{\nu}{\cdot}c(\mu),$$

giving back the $\mathrm{Gl}_{p,q}(2)$ defining relations (4.11) in the λ, μ, $\nu\to1$ limit, while the second set is given by the equations

$$
\begin{gathered}
a(\lambda)\overset{\nu}{\cdot}a(\mu)=a(\mu)\overset{\nu}{\cdot}a(\lambda), \quad a(\lambda)\overset{\nu}{\cdot}b(\mu)=P^{(\mu-\lambda)/2}a(\mu)\overset{\nu}{\cdot}b(\lambda), \\
a(\lambda)\overset{\nu}{\cdot}c(\mu)=P^{(\lambda-\mu)/2}a(\mu)\overset{\nu}{\cdot}c(\lambda), \quad a(\lambda)\overset{\nu}{\cdot}d(\mu)=a(\mu)\overset{\nu}{\cdot}d(\lambda), \\
b(\lambda)\overset{\nu}{\cdot}b(\mu)=P^{\mu-\lambda}b(\mu)\overset{\nu}{\cdot}b(\lambda), \quad b(\lambda)\overset{\nu}{\cdot}c(\mu)=b(\mu)\overset{\nu}{\cdot}c(\lambda), \\
b(\lambda)\overset{\nu}{\cdot}d(\mu)=P^{(\mu-\lambda)/2}b(\mu)\overset{\nu}{\cdot}d(\lambda), \quad c(\lambda)\overset{\nu}{\cdot}c(\mu)=P^{\lambda-\mu}c(\mu)\overset{\nu}{\cdot}c(\lambda), \\
c(\lambda)\overset{\nu}{\cdot}d(\mu)=P^{(\lambda-\mu)/2}c(\mu)\overset{\nu}{\cdot}d(\lambda), \quad d(\lambda)\overset{\nu}{\cdot}d(\mu)=d(\mu)\overset{\nu}{\cdot}d(\lambda),
\end{gathered}
\tag{4.44}
$$

which have no standard counterpart.

Similarly, the relations for the colored multiplication involving the quantum determinant, given in Eqs. (A2), (A3), (A4), (A6), and (A7), can be rewritten as

$$
\begin{gathered}
\mathscr{D}(\lambda)\overset{\nu}{\cdot}a(\mu)=a(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda), \quad \mathscr{D}(\lambda)\overset{\nu}{\cdot}b(\mu)=P^{-2\lambda}b(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda), \\
\mathscr{D}(\lambda)\overset{\nu}{\cdot}c(\mu)=P^{2\lambda}c(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda), \quad \mathscr{D}(\lambda)\overset{\nu}{\cdot}d(\mu)=d(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda),
\end{gathered}
\tag{4.45}
$$

and

$$
\begin{gathered}
\mathscr{D}(\lambda)\overset{\nu}{\cdot}a(\mu)=(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(\mathscr{D}(\mu)\overset{\nu}{\cdot}a(\lambda)), \\
\mathscr{D}(\lambda)\overset{\nu}{\cdot}b(\mu)=P^{(\mu-\lambda)/2}(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(\mathscr{D}(\mu)\overset{\nu}{\cdot}b(\lambda)), \\
\mathscr{D}(\lambda)\overset{\nu}{\cdot}c(\mu)=P^{(\lambda-\mu)/2}(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(\mathscr{D}(\mu)\overset{\nu}{\cdot}c(\lambda)), \\
\mathscr{D}(\lambda)\overset{\nu}{\cdot}d(\mu)=(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(\mathscr{D}(\mu)\overset{\nu}{\cdot}d(\lambda)), \\
\mathscr{D}(\lambda)\overset{\nu}{\cdot}\mathscr{D}(\mu)=\mathscr{D}(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda),
\end{gathered}
\tag{4.46}
$$

where the first set leads to Eq. (4.15) in the λ, μ, $\nu\rightarrow 1$ limit, while the second has again no standard counterpart.

The simplicity of Eqs. (4.43), and (4.44) should be stressed. It is also remarkable that these equations are entirely independent of the final Hopf algebra color parameter ν. Such a property reflects the fact that the two sets of equations (4.43), (4.44) are equivalent to the single matrix equation (3.19), where $R^{\lambda,\mu}_{p,q}$ is given by Eq. (4.34), and

$$
T^{\lambda}_{p,q}=T_{p^{(\lambda)},q^{(\lambda)}}=\begin{pmatrix} a(\lambda) & b(\lambda) \\ c(\lambda) & d(\lambda) \end{pmatrix}.
$$

It can indeed be checked by expressing Eq. (3.19) in elementwise form that the resulting independent relations can be grouped together in two sets, given by Eqs. (4.43), and (4.44). Since, moreover, the coalgebraic structure of $\mathrm{Gl}^c(2)$ coincides with that of each individual standard quantum group $\mathrm{Gl}_{p,q}(2)$, for (p,q) running over $\mathcal{Q}$, we did show in a very explicit way that our colored $\mathrm{Gl}^c(2)$ is equivalent to the colored extension of $\mathrm{Gl}_{p,q}(2)$, defined through colored RTT-relations, according to Basu-Mallick's prescription.[12]

By the way, it is worth noting that since parameter Q is left unchanged by the color group transformations, the elements $\mathscr{T}^{\lambda}_{p,q}$ of the colored universal $\mathscr{T}$-matrix, considered in Definition III.1, have no explicit λ-dependence, so that they are all given by Eq. (4.30) with $\hat{J}_0$, $\hat{J}_{\pm}$, $\hat{Z}$ $\in \mathrm{U}_{p^{(\lambda)},q^{(\lambda)}}(\mathrm{gl}(2))$, and h, $\tilde{h}$, β, $\gamma \in \mathrm{Gl}_{p^{(\lambda)},q^{(\lambda)}}(2)$.

As a final point of this section, let us mention that the colored maps $\tilde{m}^{\nu}_{p,q,\lambda,\mu}$, $\tilde{\iota}^{\nu}_{p,q}$, and $\tilde{S}^{\nu}_{p,q,\mu}$ may be extended from $\mathrm{Gl}^c(2)$ to the whole dual of $\mathrm{U}^c(\mathrm{gl}(2))$ by considering the enveloping algebras $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ for (p,q) running over Q. For the colored multiplication, for instance, one finds that the 16 relations giving $\tilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes y)$, $x,\ y\in\{\alpha,\beta,\gamma,\delta\}$, can be combined into six generalized commutation relations

$$
\begin{aligned}
&[\alpha(\lambda),\beta(\mu)]^{\nu}=P^{(\mu-\nu)/2}(\varphi-\lambda\theta)\beta(\nu), \quad [\alpha(\lambda),\gamma(\mu)]^{\nu}=P^{(\mu-\nu)/2}(\varphi+\lambda\theta)\gamma(\nu),\\
&[\delta(\lambda),\beta(\mu)]^{\nu}=P^{(\mu-\nu)/2}(\varphi+\lambda\theta)\beta(\nu), \quad [\delta(\lambda),\gamma(\mu)]^{\nu}=P^{(\mu-\nu)/2}(\varphi-\lambda\theta)\gamma(\nu),\\
&[\alpha(\lambda),\delta(\mu)]^{\nu}=0, \quad [\beta(\lambda),\gamma(\mu)]^{\nu}=0,
\end{aligned}
\tag{4.47}
$$

where

$$
[x(\lambda),y(\mu)]^{\nu}\equiv x(\lambda)\overset{\nu}{\cdot}y(\mu)-y(\mu)\overset{\nu}{\cdot}x(\lambda), \tag{4.48}
$$

and ten additional relations

$$
\begin{aligned}
&[\alpha(\lambda),\alpha(\mu)]^{\nu}=[\beta(\lambda),\beta(\mu)]^{\nu}=[\gamma(\lambda),\gamma(\mu)]^{\nu}=[\delta(\lambda),\delta(\mu)]^{\nu}=0,\\
&\alpha(\lambda)\overset{\nu}{\cdot}\beta(\mu)=P^{(\mu-\lambda)/2}(2\mu)^{-1}((\lambda+\mu)\alpha(\mu)\overset{\nu}{\cdot}\beta(\lambda)-(\lambda-\mu)\delta(\mu)\overset{\nu}{\cdot}\beta(\lambda)),\\
&\gamma(\lambda)\overset{\nu}{\cdot}\alpha(\mu)=P^{(\lambda-\mu)/2}(2\lambda)^{-1}((\lambda+\mu)\gamma(\mu)\overset{\nu}{\cdot}\alpha(\lambda)+(\lambda-\mu)\gamma(\mu)\overset{\nu}{\cdot}\delta(\lambda)),\\
&\alpha(\lambda)\overset{\nu}{\cdot}\delta(\mu)=\alpha(\mu)\overset{\nu}{\cdot}\delta(\lambda)+(\lambda-\mu)(\lambda+\mu)^{-1}(\alpha(\mu)\overset{\nu}{\cdot}\alpha(\lambda)-\delta(\mu)\overset{\nu}{\cdot}\delta(\lambda)),\\
&\gamma(\lambda)\overset{\nu}{\cdot}\beta(\mu)=\gamma(\mu)\overset{\nu}{\cdot}\beta(\lambda),\\
&\delta(\lambda)\overset{\nu}{\cdot}\beta(\mu)=P^{(\mu-\lambda)/2}(2\mu)^{-1}(-(\lambda-\mu)\alpha(\mu)\overset{\nu}{\cdot}\beta(\lambda)+(\lambda+\mu)\delta(\mu)\overset{\nu}{\cdot}\beta(\lambda)),\\
&\gamma(\lambda)\overset{\nu}{\cdot}\delta(\mu)=P^{(\lambda-\mu)/2}(2\lambda)^{-1}((\lambda-\mu)\gamma(\mu)\overset{\nu}{\cdot}\alpha(\lambda)+(\lambda+\mu)\gamma(\mu)\overset{\nu}{\cdot}\delta(\lambda)).
\end{aligned}
\tag{4.49}
$$

Whenever $\lambda,\ \mu,\ \nu\to 1$, the former set reproduces the defining relations (4.20) of $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$, whereas the latter has no standard counterpart. Let us stress that contrary to what happens in Eqs. (4.43) and (4.44), there appears an explicit ν-dependence in Eqs. (4.47), and (4.49).

V. THE COLORED TWO-PARAMETER QUEA $\mathrm{U}^c(\mathrm{gl}(1/1))$ AND ITS DUAL $\mathrm{Gl}^c(1/1)$

One of the simplest examples of dual pairs made of a QUEA of a Lie superalgebra and of a quantum supergroup consists in the two-parameter deformations of U(gl(1/1)) and Gl(1/1),[28–32] which find some interesting applications to the multivariable Alexander–Conway polynomial and the free fermion model,[33] and, in another context, to the XY quantum chain in a magnetic field.[34]

In II, a colored QUEA was constructed by starting from the two-parameter deformation of U(gl(1/1)) in the Bednar *et al.* formulation.[30,31] Moreover, Chakrabarti and Jagannathan[21] recently extended the Fronsdal and Galindo universal T-matrix formalism[14] to the pair of Hopf superalgebras $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$ and $\mathrm{Gl}_{p,q}(1/1)$. So all the ingredients needed to build an example of colored Hopf dual in the graded case are available. To solve such a problem, we shall use the approach to $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$ employed in Ref. 21, instead of that of Refs. 30,31; hence, the results for $\mathrm{U}^c(\mathrm{gl}(1/1))$ derived here will slightly differ from those obtained in II.

A. The dual Hopf superalgebras $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$ and $\mathrm{Gl}_{p,q}(1/1)$

The quantum superalgebra $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$ is defined[21] as the algebra generated by two even elements $\{J_0,Z\}$, and two odd ones $\{J_+,J_-\}$, subject to the relations

$$
[J_0,J_\pm]=\pm J_\pm, \quad \{J_+,J_-\}=[2Z]_Q, \quad J_\pm^2=0, \quad [Z,J_0]=[Z,J_\pm]=0, \tag{5.1}
$$

where $[X]_t$, and Q, P are still defined by Eqs. (4.2) and (4.3), respectively. As in the $U_{p,q}(gl(2))$ case, the coalgebraic structure depends upon both P and Q.

The duality relationship between $U_{p,q}(gl(1/1))$ and $Gl_{p,q}(1/1)$ is most easily described in terms of another set of $U_{p,q}(gl(1/1))$ generators $\{\hat{J}_0,\hat{J}_\pm,\hat{Z}\}$, with $\hat{J}_0$, $\hat{Z}$ even, and $\hat{J}_\pm$ odd. The new generators are defined in terms of the old ones by

$$\hat{J}_0=J_0, \quad \hat{J}_\pm=J_\pm Q^{\pm Z}P^{Z-1/2}, \quad \hat{Z}=Z, \tag{5.2}$$

and satisfy the superalgebra (anti)commutation relations

$$\begin{gathered}[\hat{J}_0,\hat{J}_\pm]=\pm\hat{J}_\pm, \quad \{\hat{J}_+,\hat{J}_-\}=\frac{p^{2\hat{Z}}-q^{-2\hat{Z}}}{p-q^{-1}}, \quad \hat{J}_\pm^2=0,\\ [\hat{Z},\hat{J}_0]=[\hat{Z},\hat{J}_\pm]=0.\end{gathered} \tag{5.3}$$

The remaining Hopf superalgebra maps are given by

$$\begin{gathered}\Delta_{p,q}(\hat{J}_0)=\hat{J}_0\otimes 1+1\otimes\hat{J}_0, \quad \Delta_{p,q}(\hat{Z})=\hat{Z}\otimes 1+1\otimes\hat{Z},\\ \Delta_{p,q}(\hat{J}_+)=\hat{J}_+\otimes p^{2\hat{Z}}+1\otimes\hat{J}_+, \quad \Delta_{p,q}(\hat{J}_-)=\hat{J}_-\otimes 1+q^{-2\hat{Z}}\otimes\hat{J}_-,\\ \epsilon_{p,q}(X)=0, \quad X\in\{\hat{J}_0,\hat{J}_\pm,\hat{Z}\},\\ S_{p,q}(\hat{J}_0)=-\hat{J}_0, \quad S_{p,q}(\hat{Z})=-\hat{Z},\\ S_{p,q}(\hat{J}_+)=-p^{-2\hat{Z}}\hat{J}_+, \quad S_{p,q}(\hat{J}_-)=-q^{2\hat{Z}}\hat{J}_-,\end{gathered} \tag{5.4}$$

and the universal $\mathscr{R}$-matrix can be written as[31,32]

$$\mathscr{R}_{p,q}=p^{2\hat{Z}\otimes\hat{J}_0}q^{2\hat{J}_0\otimes\hat{Z}}\{1\otimes 1-(p-q^{-1})\hat{J}_+\otimes\hat{J}_-\}. \tag{5.5}$$

In the 2×2 defining representation $D_{p,q}$ of $U_{p,q}(gl(1/1))$, given by Eq. (4.8), the universal $\mathscr{R}$-matrix (5.5) is represented by the 4×4 matrix,

$$R_{p,q}\equiv(D_{p,q}\otimes D_{p,q})(\mathscr{R}_{p,q})=\begin{pmatrix} Q & 0 & 0 & 0\\ 0 & P^{-1} & Q-Q^{-1} & 0\\ 0 & 0 & P & 0\\ 0 & 0 & 0 & Q^{-1}\end{pmatrix}. \tag{5.6}$$

In deriving Eq. (5.6), use is made of Eq. (3.16) with the $\mathbb{Z}_2$-grade of the first (resp. second) row or column of $D_{p,q}(X)$ defined as zero (resp. one).

The defining T-matrix of the corresponding quantum supergroup $Gl_{p,q}(1/1)$ is still given by Eq. (4.10), where a, d, and b, c are now even and odd, respectively, and satisfy the relations

$$\begin{gathered}ab=p^{-1}ba, \quad ac=q^{-1}ca, \quad bd=pdb, \quad cd=qdc,\\ bc=-(p/q)cb, \quad ad-da=(q-p^{-1})bc, \quad b^2=c^2=0.\end{gathered} \tag{5.7}$$

The latter follow from the RTT-relations corresponding to the R-matrix (5.6), with convention (3.17) taken into account. The coalgebra maps are given by Eq. (4.12), while the antipode one is

$$\widetilde{S}_{p,q}(T_{p,q})=(T_{p,q})^{-1}=\begin{pmatrix} a^{-1}+a^{-1}bd^{-1}ca^{-1} & -a^{-1}bd^{-1}\\ -d^{-1}ca^{-1} & d^{-1}+d^{-1}ca^{-1}bd^{-1}\end{pmatrix}, \tag{5.8}$$

where a and d are assumed invertible. The quantum superdeterminant, defined by

$$\mathscr{D}\equiv ad^{-1}-bd^{-1}cd^{-1}, \tag{5.9}$$

is both central and group-like, with $\widetilde{S}_{p,q}(\mathscr{D})=\mathscr{D}^{-1}=da^{-1}+ba^{-1}ca^{-1}$.

Gauss decomposition (4.17) still holds, but now a and $\hat{d}$ are invertible even elements, while β and γ are odd. The antipode map and the quantum superdeterminant may be rewritten as

$$\widetilde{S}_{p,q}(T_{p,q})=\begin{pmatrix} a^{-1}-p^{-1}qca^{-2}\hat{d}^{-1}b & -p^{-1}a^{-1}\hat{d}^{-1}b \\ -qca^{-1}\hat{d}^{-1} & \hat{d}^{-1} \end{pmatrix}, \tag{5.10}$$

and

$$\mathscr{D}=a\hat{d}^{-1}, \tag{5.11}$$

respectively.

By assuming that the superalgebra can be augmented with the logarithms of a and $\hat{d}$, and by using the maps

$$a=e^{\alpha}, \quad \hat{d}=e^{\delta}, \tag{5.12}$$

and the reparametrization

$$p=e^{-\omega}, \quad q=e^{-\eta}, \tag{5.13}$$

the quantum superdeterminant is transformed into $\mathscr{D}=\exp(\alpha-\delta)$, and the new variables $\{\alpha,\beta,\gamma,\delta\}$ satisfy the (anti)commutation relations of a solvable Lie superalgebra

$$\begin{gathered} [\alpha,\beta]=\omega\beta, \quad [\alpha,\gamma]=\eta\gamma, \quad [\delta,\beta]=\omega\beta, \quad [\delta,\gamma]=\eta\gamma, \\ [\alpha,\delta]=0, \quad \{\beta,\gamma\}=0, \quad \beta^2=0, \quad \gamma^2=0, \end{gathered} \tag{5.14}$$

with a noncocommutative coproduct structure.[21] This shows that $\mathrm{Gl}_{p,q}(1/1)$ can be embedded into the enveloping algebra $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ of such a Lie superalgebra. According to Chakrabarti and Jagannathan,[21] $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ is dual to the QUEA $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$.[35]

Dual bases of $\mathrm{U}_{p,q}(\mathrm{gl}(1/1))$ and $\mathrm{Gl}_{p,q}(1/1)$ are given by[21]

$$X_A=\hat{J}_{-}^{a_1}\frac{H^{a_2}}{a_2!}\frac{\widetilde{H}^{a_3}}{a_3!}\hat{J}_{+}^{a_4}, \quad x^A=\gamma^{a_1}\alpha^{a_2}\delta^{a_3}\beta^{a_4}, \tag{5.15}$$

respectively, where $H\equiv\hat{Z}+\hat{J}_0$, $\widetilde{H}\equiv\hat{Z}-\hat{J}_0$, $A=(a_1,a_2,a_3,a_4)$, a_1, $a_4\in\{0,1\}$, and a_2, $a_3\in\mathbb{N}$. From Eqs. (3.2), and (5.15), the universal $\mathscr{T}$-matrix of $\mathrm{Gl}_{p,q}(1/1)$ can be written as

$$\mathscr{T}_{p,q}=\exp(\gamma\otimes\hat{J}_{-})\exp(\alpha\otimes H+\delta\otimes\widetilde{H})\exp(\beta\otimes\hat{J}_{+}), \tag{5.16}$$

in terms of standard exponentials.

In the next subsection, it will prove convenient to consider another set of dual bases

$$\overline{X_A}=\sum_B c_A^B X_B=\hat{J}_{-}^{a_1}\frac{\hat{J}_0^{a_2}}{a_2!}\frac{\hat{Z}^{a_3}}{a_3!}\hat{J}_{+}^{a_4}, \tag{5.17}$$

$$\overline{x^A}=\sum_B (c^{-1})_B^A x^B=\gamma^{a_1}h^{a_2}\widetilde{h}^{a_3}\beta^{a_4}, \tag{5.18}$$

where A has the same meaning as in Eq. (5.15),

$$h\equiv\alpha-\delta, \quad \widetilde{h}\equiv\alpha+\delta, \tag{5.19}$$

and

$$c_A^B=2^{-a_2-a_3}(-1)^{a_2-b_2}(c^{-1})_A^B$$

$$= \delta^{b_1}_{a_1}\delta^{b_2+b_3}_{a_2+a_3}\delta^{b_4}_{a_4} 2^{-a_2-a_3} \sum_{t=\max(0,b_2-a_2)}^{\min(b_2,a_3)} (-1)^{a_2-b_2+t} \begin{pmatrix} b_2 \\ t \end{pmatrix} \quad \begin{pmatrix} b_3 \\ a_3-t \end{pmatrix}. \tag{5.20}$$

In terms of them, the universal $\mathscr{T}$-matrix (5.16) can be rewritten as

$$\mathscr{T}_{p,q} = \exp(\gamma \otimes \hat{J}_-) \exp(h \otimes \hat{J}_0 + \tilde{h} \otimes \hat{Z}) \exp(\beta \otimes \hat{J}_+). \tag{5.21}$$

B. The colored Hopf superalgebra $U^c(gl(1/1))$ and its dual $Gl^c(1/1)$

The defining relations (5.3) of the $U_{p,q}(gl(1/1))$ superalgebra are left invariant under the grade-preserving transformations,

$$\sigma^\nu(\hat{J}_0) = \hat{J}_0, \quad \sigma^\nu(\hat{J}_\pm) = \left(\frac{p^\nu - q^{-\nu}}{p - q^{-1}} \right)^{1/2} \hat{J}_\pm, \quad \sigma^\nu(\hat{Z}) = \nu \hat{Z}, \tag{5.22}$$

where $\nu \in \mathscr{C} = \mathbb{C} \setminus \{0\}$, provided p and q are replaced by their νth powers, p^ν and q^ν (and the same for P and Q). The σ^ν's are therefore isomorphic mappings between the $U_{p,q}(gl(1/1))$ and $U_{p^\nu,q^\nu}(gl(1/1))$ superalgebras, and define a color group $\mathscr{G} = Gl(1,\mathbb{C})$.

By choosing the parameter set $\mathscr{Q} = (\mathbb{C} \setminus \{0\}) \times (\mathbb{C} \setminus \{0\})$, we obtain a colored quasitriangular Hopf superalgebra $U^c(gl(1/1))$, whose colored comultiplication, counit, antipode, and $\mathscr{R}$-matrix are given by

$$\begin{gathered}
\Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_0) = \hat{J}_0 \otimes 1 + 1 \otimes \hat{J}_0, \quad \Delta^{\lambda,\mu}_{p,q,\nu}(\hat{Z}) = \frac{\lambda}{\nu} \hat{Z} \otimes 1 + \frac{\mu}{\nu} 1 \otimes \hat{Z}, \\
\Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_+) = A^\lambda_\nu \hat{J}_+ \otimes p^{2\mu\hat{Z}} + A^\mu_\nu 1 \otimes \hat{J}_+, \quad \Delta^{\lambda,\mu}_{p,q,\nu}(\hat{J}_-) = A^\lambda_\nu \hat{J}_- \otimes 1 + A^\mu_\nu q^{-2\lambda\hat{Z}} \otimes \hat{J}_-, \\
\epsilon_{p,q,\nu}(X) = 0, \quad X \in \{\hat{J}_0, \hat{J}_\pm, \hat{Z}\}, \\
S^\mu_{p,q,\nu}(\hat{J}_0) = -\hat{J}_0, \quad S^\mu_{p,q,\nu}(\hat{J}_+) = -A^\mu_\nu p^{-2\mu\hat{Z}} \hat{J}_+, \\
S^\mu_{p,q,\nu}(\hat{J}_-) = -A^\mu_\nu q^{2\mu\hat{Z}} \hat{J}_-, \quad S^\mu_{p,q,\nu}(\hat{Z}) = -\frac{\mu}{\nu} \hat{Z}, \\
\mathscr{R}^{\lambda,\mu}_{p,q} = p^{2\lambda\hat{Z}\otimes\hat{J}_0} q^{2\mu\hat{J}_0\otimes\hat{Z}} \{1 \otimes 1 - \sqrt{(p^\lambda - q^{-\lambda})(p^\mu - q^{-\mu})} \hat{J}_+ \otimes \hat{J}_-\},
\end{gathered} \tag{5.23}$$

where $A^\lambda_\nu \equiv ((p^\lambda - q^{-\lambda})/(p^\nu - q^{-\nu}))^{1/2}$.

In the defining representation (4.8), the colored universal $\mathscr{R}$-matrix is represented by the 4×4 matrix,

$$R^{\lambda,\mu}_{p,q} = \begin{pmatrix} Q^{(\lambda+\mu)/2} P^{(\lambda-\mu)/2} & 0 & 0 & 0 \\ 0 & Q^{-(\lambda-\mu)/2} P^{-(\lambda+\mu)/2} & Q^{-(\lambda-\mu)/2} B^{\lambda,\mu} & 0 \\ 0 & 0 & Q^{(\lambda-\mu)/2} P^{(\lambda+\mu)/2} & 0 \\ 0 & 0 & 0 & Q^{-(\lambda+\mu)/2} P^{-(\lambda-\mu)/2} \end{pmatrix}, \tag{5.24}$$

with $B^{\lambda,\mu} \equiv \sqrt{(Q^\lambda - Q^{-\lambda})(Q^\mu - Q^{-\mu})}$. Such a matrix is a solution of the colored graded YBE, and gives back matrix (5.6) whenever $\lambda, \mu \to 1$.

Under transformation (5.22), the basis elements $\overline{X_A}$ of $U_{p,q}(gl(1/1))$, defined in Eq. (5.17), become

$$\sigma^\nu(\overline{X_A}) = (A^\nu_1)^{(a_1+a_4)/2} \nu^{a_3} \overline{X_A}. \tag{5.25}$$

Hence, from Eqs. (2.4) and (3.1), it follows that

$$\rho^{\nu}(\overline{x^A}) = (A^{\ 1}_{\nu})^{(a_1+a_4)/2}\nu^{-a_3}\overline{x^A}, \tag{5.26}$$

and

$$\rho^{\nu}(\gamma^{a_1}e^{r(\alpha-\delta)}e^{s(\alpha+\delta)}\beta^{a_4}) = (A^{\ 1}_{\nu})^{(a_1+a_4)/2}\gamma^{a_1}e^{r(\alpha-\delta)}e^{s(\alpha+\delta)/\nu}\beta^{a_4}, \tag{5.27}$$

where $r,\ s \in \mathbb{C}$.

As special cases of Eqs. (5.26) and (5.27), we obtain[27]

$$\begin{gathered}\rho^{\nu}(\alpha) = \frac{1+\nu}{2\nu}\,\alpha + \frac{1-\nu}{2\nu}\,\delta, \quad \rho^{\nu}(\beta) = A^{\ 1}_{\nu}\beta, \\ \rho^{\nu}(\gamma) = A^{\ 1}_{\nu}\gamma, \quad \rho^{\nu}(\delta) = \frac{1-\nu}{2\nu}\,\alpha + \frac{1+\nu}{2\nu}\,\delta,\end{gathered} \tag{5.28}$$

$$\begin{gathered}\rho_{\nu}(\alpha) = \frac{\nu+1}{2}\,\alpha + \frac{\nu-1}{2}\,\delta, \quad \rho_{\nu}(\beta) = A^{\ \nu}_{1}\beta, \\ \rho_{\nu}(\gamma) = A^{\ \nu}_{1}\gamma, \quad \rho_{\nu}(\delta) = \frac{\nu-1}{2}\,\alpha + \frac{\nu+1}{2}\,\delta,\end{gathered} \tag{5.29}$$

and

$$\begin{gathered}\rho^{\nu}(a) = \mathscr{D}^{(\nu-1)/(2\nu)}a^{1/\nu}, \quad \rho^{\nu}(b) = A^{\ 1}_{\nu}\mathscr{D}^{(\nu-1)/(2\nu)}a^{(1-\nu)/\nu}b, \\ \rho^{\nu}(c) = q^{1-\nu}A^{\ 1}_{\nu}\mathscr{D}^{(\nu-1)/(2\nu)}a^{(1-\nu)/\nu}c, \\ \rho^{\nu}(d) = \mathscr{D}^{(\nu-1)/(2\nu)}a^{(1-\nu)/\nu}\{d - p(A^{\ \nu-1}_{\nu})^2 ca^{-1}b\}, \quad \rho^{\nu}(\mathscr{D}) = \mathscr{D},\end{gathered} \tag{5.30}$$

$$\begin{gathered}\rho_{\nu}(a) = \mathscr{D}^{(1-\nu)/2}a^{\nu}, \quad \rho_{\nu}(b) = A^{\ \nu}_{1}\mathscr{D}^{(1-\nu)/2}a^{\nu-1}b, \\ \rho_{\nu}(c) = q^{\nu-1}A^{\ \nu}_{1}\mathscr{D}^{(1-\nu)/2}a^{\nu-1}c, \\ \rho_{\nu}(d) = \mathscr{D}^{(1-\nu)/2}a^{\nu-1}\{d - p^{\nu}(A^{\ 1-\nu}_{1})^2 ca^{-1}b\}, \quad \rho_{\nu}(\mathscr{D}) = \mathscr{D}.\end{gathered} \tag{5.31}$$

By proceeding as in Sec. IV, it is straightforward to determine the colored maps of $\mathrm{Gl}^c(1/1)$, listed in Appendix B, and to combine the results for the colored multiplication $\widetilde{m}^{\ \nu}_{p,q,\lambda,\mu}(x\otimes y)$, $x,\ y \in \{a,b,c,d\}$, given in Eqs. (B1)–(B4), and (B10), into two sets of eight relations each,

$$\begin{gathered}a(\lambda)\overset{\nu}{\cdot}b(\mu) = p^{-\lambda}b(\mu)\overset{\nu}{\cdot}a(\lambda), \quad a(\lambda)\overset{\nu}{\cdot}c(\mu) = q^{-\lambda}c(\mu)\overset{\nu}{\cdot}a(\lambda), \\ b(\lambda)\overset{\nu}{\cdot}d(\mu) = p^{\mu}d(\mu)\overset{\nu}{\cdot}a(\lambda), \quad c(\lambda)\overset{\nu}{\cdot}d(\mu) = q^{\mu}d(\mu)\overset{\nu}{\cdot}c(\lambda), \\ b(\lambda)\overset{\nu}{\cdot}c(\mu) = -p^{\mu}q^{-\lambda}c(\mu)\overset{\nu}{\cdot}b(\lambda), \quad b(\lambda)\overset{\nu}{\cdot}b(\mu) = c(\lambda)\overset{\nu}{\cdot}c(\mu) = 0, \\ a(\lambda)\overset{\nu}{\cdot}d(\mu) - d(\mu)\overset{\nu}{\cdot}a(\lambda) = p^{-\mu}q^{\lambda}\sqrt{(p^{\lambda}-q^{-\lambda})(p^{\mu}-q^{-\mu})}\,b(\lambda)\overset{\nu}{\cdot}c(\mu),\end{gathered} \tag{5.32}$$

and

$$\begin{gathered}a(\lambda)\overset{\nu}{\cdot}a(\mu) = a(\mu)\overset{\nu}{\cdot}a(\lambda), \quad a(\lambda)\overset{\nu}{\cdot}b(\mu) = A^{\ \mu}_{\lambda}a(\mu)\overset{\nu}{\cdot}b(\lambda), \\ a(\lambda)\overset{\nu}{\cdot}c(\mu) = q^{\mu-\lambda}A^{\ \mu}_{\lambda}a(\mu)\overset{\nu}{\cdot}c(\lambda), \\ a(\lambda)\overset{\nu}{\cdot}d(\mu) - a(\mu)\overset{\nu}{\cdot}d(\lambda) = q^{\lambda-\mu}A^{\ \lambda-\mu}_{\lambda}A^{\ \lambda-\mu}_{\mu}b(\lambda)\overset{\nu}{\cdot}c(\mu),\end{gathered} \tag{5.33}$$

$$b(\lambda)\overset{\nu}{\cdot}c(\mu)=(pq)^{\mu-\lambda}b(\mu)\overset{\nu}{\cdot}c(\lambda),\quad b(\lambda)\overset{\nu}{\cdot}d(\mu)=p^{\mu-\lambda}A^{\lambda}_{\mu}b(\mu)\overset{\nu}{\cdot}d(\lambda),$$

$$c(\lambda)\overset{\nu}{\cdot}d(\mu)=A^{\lambda}_{\mu}c(\mu)\overset{\nu}{\cdot}d(\lambda),\quad d(\lambda)\overset{\nu}{\cdot}d(\mu)=d(\mu)\overset{\nu}{\cdot}d(\lambda).$$

Similarly, the relations for the colored multiplication involving the quantum superdeterminant, given in Eqs. (B5)–(B9), and (B11), can be rewritten as

$$\mathscr{D}(\lambda)\overset{\nu}{\cdot}x(\mu)=x(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda),\quad x\in\{a,b,c,d\},\tag{5.34}$$

and

$$\mathscr{D}(\lambda)\overset{\nu}{\cdot}a(\mu)=(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(a(\nu))^{(\mu-\lambda)/\nu}(\mathscr{D}(\mu)\overset{\nu}{\cdot}a(\lambda)),$$

$$\mathscr{D}(\lambda)\overset{\nu}{\cdot}b(\mu)=A^{\mu}_{\lambda}(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(a(\nu))^{(\mu-\lambda)/\nu}(\mathscr{D}(\mu)\overset{\nu}{\cdot}b(\lambda)),$$

$$\mathscr{D}(\lambda)\overset{\nu}{\cdot}c(\mu)=q^{\mu-\lambda}A^{\mu}_{\lambda}(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(a(\nu))^{(\mu-\lambda)/\nu}(\mathscr{D}(\mu)\overset{\nu}{\cdot}c(\lambda)),\tag{5.35}$$

$$\begin{aligned}\mathscr{D}(\lambda)\overset{\nu}{\cdot}d(\mu)=&q^{\mu-\lambda}(A^{\mu-\nu}_{\lambda-\nu})^2(\mathscr{D}(\nu))^{(\lambda-\mu)/(2\nu)}(a(\nu))^{(\mu-\lambda)/\nu}(\mathscr{D}(\mu)\overset{\nu}{\cdot}d(\lambda))\\&+p^{\mu-\nu}(A^{\lambda-\mu}_{\lambda-\nu})^2(\mathscr{D}(\nu))^{(3\nu-\mu)/(2\nu)}(a(\nu))^{(\mu-\lambda)/\nu}d(\nu),\end{aligned}$$

$$\mathscr{D}(\lambda)\overset{\nu}{\cdot}\mathscr{D}(\mu)=\mathscr{D}(\mu)\overset{\nu}{\cdot}\mathscr{D}(\lambda).$$

In the λ, μ, $\nu\to 1$ limit, Eqs. (5.32) and (5.34) give back the $\mathrm{Gl}_{p,q}(1/1)$ defining relations (5.7) and the central property of $\mathscr{D}$, while Eqs. (5.33), and (5.35) have no standard counterpart.

The two ν-independent sets of equations, contained in (5.32) and (5.33), can be explicitly shown to be equivalent to the single matrix equation (3.19), where $R^{\lambda,\mu}_{p,q}$ is given by Eq. (5.24). Moreover, as in Sec. IV, the elements $\mathscr{T}^{\lambda}_{p,q}$ of the colored universal $\mathscr{T}$-matrix have no explicit λ-dependence.

The counterparts of Eqs. (4.47) and (4.49), for the colored multiplication extended to the whole dual $\mathrm{U}_{p,q}(\{\alpha,\beta,\gamma,\delta\})$ of $\mathrm{U}^c\mathrm{gl}(1/1))$, are

$$[\alpha(\lambda),\beta(\mu)]^{\nu}=\lambda\omega A^{\mu}_{\nu}\beta(\nu),\quad [\alpha(\lambda),\gamma(\mu)]^{\nu}=\lambda\eta A^{\mu}_{\nu}\gamma(\nu),$$

$$[\delta(\lambda),\beta(\mu)]^{\nu}=\lambda\omega A^{\mu}_{\nu}\beta(\nu),\quad [\delta(\lambda),\gamma(\mu)]^{\nu}=\lambda\eta A^{\mu}_{\nu}\gamma(\nu),\tag{5.36}$$

$$[\alpha(\lambda),\delta(\mu)]^{\nu}=0,\quad \{\beta(\lambda),\gamma(\mu)\}^{\nu}=\beta(\lambda)\overset{\nu}{\cdot}\beta(\mu)=\gamma(\lambda)\overset{\nu}{\cdot}\gamma(\mu)=0,$$

and

$$[\alpha(\lambda),\alpha(\mu)]^{\nu}=[\delta(\lambda),\delta(\mu)]^{\nu}=0,$$

$$\alpha(\lambda)\overset{\nu}{\cdot}\beta(\mu)=A^{\mu}_{\lambda}(2\mu)^{-1}((\lambda+\mu)\alpha(\mu)\overset{\nu}{\cdot}\beta(\lambda)+(\lambda-\mu)\delta(\mu)\overset{\nu}{\cdot}\beta(\lambda)),$$

$$\gamma(\lambda)\overset{\nu}{\cdot}\alpha(\mu)=A^{\lambda}_{\mu}(2\lambda)^{-1}((\lambda+\mu)\gamma(\mu)\overset{\nu}{\cdot}\alpha(\lambda)-(\lambda-\mu)\gamma(\mu)\overset{\nu}{\cdot}\delta(\lambda)),$$

$$\alpha(\lambda)\overset{\nu}{\cdot}\delta(\mu)=\alpha(\mu)\overset{\nu}{\cdot}\delta(\lambda)+(\lambda-\mu)(\lambda+\mu)^{-1}(-\alpha(\mu)\overset{\nu}{\cdot}\alpha(\lambda)+\delta(\mu)\overset{\nu}{\cdot}\delta(\lambda)),$$

$$\gamma(\lambda)\overset{\nu}{\cdot}\beta(\mu)=\gamma(\mu)\overset{\nu}{\cdot}\beta(\lambda),\tag{5.37}$$

$$\delta(\lambda)\overset{\nu}{\cdot}\beta(\mu)=A^{\mu}_{\lambda}(2\mu)^{-1}((\lambda-\mu)\alpha(\mu)\overset{\nu}{\cdot}\beta(\lambda)+(\lambda+\mu)\delta(\mu)\overset{\nu}{\cdot}\beta(\lambda)),$$

$$\gamma(\lambda)\overset{\nu}{\cdot}\delta(\mu)=A^{\lambda}_{\mu}(2\lambda)^{-1}(-(\lambda-\mu)\gamma(\mu)\overset{\nu}{\cdot}\alpha(\lambda)+(\lambda+\mu)\gamma(\mu)\overset{\nu}{\cdot}\delta(\lambda)),$$

where

$$\{x(\lambda),y(\mu)\}^{\nu}\equiv x(\lambda)\overset{\nu}{\cdot}y(\mu)+y(\mu)\overset{\nu}{\cdot}x(\lambda). \tag{5.38}$$

Equation (5.36) gives back Eq. (5.14) whenever λ, μ, $\nu\rightarrow 1$.

All the results obtained for the dual pair ($U^c(gl(1/1))$, $Gl^c(1/1)$) are therefore entirely similar to those derived for ($U^c(gl(2))$, $Gl^c(2)$) in Sec. IV.

VI. CONCLUSION

In the present paper, we extended the notion of dually conjugate Hopf (super)algebras to the colored Hopf (super)algebras $\mathscr{H}^c$, introduced in I and II. We showed that if the standard Hopf (super)algebras $\mathscr{H}_q$ that are the building blocks of $\mathscr{H}^c$ have Hopf duals $\mathscr{H}_q^*$, then the latter may be used to construct colored Hopf duals $\mathscr{H}^{c*}$, endowed with colored algebra and antipode maps, but with a standard coalgebraic structure.

Next, we reviewed the case where the $\mathscr{H}_q$'s are QUEA's of Lie (super)algebras $U_q(g)$, so that the $\mathscr{H}_q^*$'s are quantum (super)groups G_q. We extended the Fronsdal and Galindo universal $\mathscr{T}$-matrix formalism[14] to the colored pairs $(U^c(g),G^c)$ by introducing colored universal $\mathscr{T}$-matrices. We then proved that the colored RTT-relations, defining colored $A(R)$ Hopf (super) algebras in Basu-Mallick's approach,[12] may be considered as the realization in the $U_q(g)$ defining representation of a representation-free relation satisfied by the colored universal $\mathscr{R}$- and $\mathscr{T}$-matrices.

Such results were finally illustrated by constructing two physically-relevant examples of colored pairs, corresponding to the two-parameter deformations of (U(gl(2)), Gl(2)), and (U(gl(1/1)), Gl(1/1)), respectively.

In conclusion, we did prove that the formalism developed in I, II, and the present paper, provides an algebraic formulation of the colored RTT-relations, and establishes a link between the colored extensions of Drinfeld–Jimbo[5] and Faddeev–Reshetikhin–Takhtajan[8] pictures of quantum groups and quantum algebras. Since transfer matrices of quantum integrable models may be obtained from $\mathscr{T}$ by specialization to given representations, we do think that the colored extension of $\mathscr{T}$, and the related new algebraic structures introduced here, may have some interesting applications to such models.

There remain some open questions, which might be interesting topics for future study. We would like to mention here two of them. The first one consists in investigating the complementary approach to that considered in the present paper, namely trying to transform a set of quantum (super)groups G_q into a colored Hopf (super)algebra $\mathscr{H}^c$ by defining an appropriate color group, then transferring the colored structure to the duals $U_q(g)$ to build a colored Hopf dual $\mathscr{H}^{c*}$. The second problem is to understand the relation, if any, between the latter and the colored $U(R)$ Hopf (super)algebras introduced by Kundu and Basu-Mallick, and defined in terms of colored RLL-relations.[11]

APPENDIX A: COLORED MAPS OF $Gl^c(2)$

In this Appendix, we list the results obtained for the colored multiplication, unit, and antipode of $Gl^c(2)$. For their derivation, we used Eqs. (2.7), (4.11), (4.15), (4.16), (4.17), (4.18), (4.37), (4.41), and the fact that $\overline{x^{0000}}$ is the unit of $Gl_{p,q}(2)$.

The colored multiplication is given by

$$\tilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes y)=P^{t\,\nu}_{\lambda,\mu}{}^{(x,y)/2}\mathscr{D}^{(\lambda+\mu-2\nu)/(2\nu)}xy, \tag{A1}$$

$$\tilde{m}^{\nu}_{p,q,\lambda,\mu}(\mathscr{D}\otimes x)=P^{t\,\nu}_{\lambda,\mu}{}^{(\mathscr{D},x)/2}\mathscr{D}^{(2\lambda+\mu-\nu)/(2\nu)}x, \tag{A2}$$

$$\tilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes\mathscr{D})=P^{t\,\nu}_{\lambda,\mu}{}^{(x,\mathscr{D})/2}\mathscr{D}^{(\lambda+2\mu-\nu)/(2\nu)}x, \tag{A3}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(\mathscr{D}\otimes\mathscr{D})=\mathscr{D}^{(\lambda+\mu)/\nu}, \tag{A4}$$

where

$$t^{\nu}_{\lambda,\mu}(x,y)=0,\mu-\nu,-\mu+\nu,0,\lambda+2\mu-3\nu,\lambda+3\mu-4\nu,\lambda+\mu-2\nu,\lambda+2\mu-3\nu,-\lambda-2\mu+3\nu,$$
$$-\lambda-\mu+2\nu,-\lambda-3\mu+4\nu,-\lambda-2\mu+3\nu,0,\mu-\nu,-\mu+\nu,0, \tag{A5}$$

$$t^{\nu}_{\lambda,\mu}(\mathscr{D},x)=0,\mu-\nu,-\mu+\nu,0, \tag{A6}$$

$$t^{\nu}_{\lambda,\mu}(x,\mathscr{D})=0,\lambda+4\mu-\nu,-\lambda-4\mu+\nu,0, \tag{A7}$$

whenever x,y run over $\{a,b,c,d\}$, and are listed in lexicographical order.

For the colored unit and antipode, the results read

$$\widetilde{\iota}^{\nu}_{p,q}(1_k)=\widetilde{\iota}_{p^{(\nu)},q^{(\nu)}}(1_k), \tag{A8}$$

and

$$\widetilde{S}^{\nu}_{p,q,\mu}(a)=\mathscr{D}^{-(\mu+\nu)/(2\nu)}d,\quad \widetilde{S}^{\nu}_{p,q,\mu}(b)=-Q^{-1}(P^{\nu}\mathscr{D})^{-(\mu+\nu)/(2\nu)}b, \tag{A9}$$

$$\widetilde{S}^{\nu}_{p,q,\mu}(c)=-Q(P^{-\nu}\mathscr{D})^{-(\mu+\nu)/(2\nu)}c,\quad \widetilde{S}^{\nu}_{p,q,\mu}(d)=\mathscr{D}^{-(\mu+\nu)/(2\nu)}a, \tag{A10}$$

$$\widetilde{S}^{\nu}_{p,q,\mu}(\mathscr{D})=\mathscr{D}^{-\mu/\nu}, \tag{A11}$$

respectively.

APPENDIX B: COLORED MAPS OF $\mathrm{Gl}^c(1/1)$

In this Appendix, we list the results obtained for the coloured multiplication, unit, and antipode of $\mathrm{Gl}^c(1/1)$. For their derivation, we used Eqs. (2.7), (4.17), (5.7), (5.10), (5.12), (5.27), (5.31), and the fact that $\overline{x^{0000}}$ is the unit of $\mathrm{Gl}_{p,q}(1/1)$.

The colored multiplication is given by

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes y)=C^{\nu}_{\lambda,\mu}(x,y)\mathscr{D}^{(2\nu-\lambda-\mu)/(2\nu)}a^{(\lambda+\mu-2\nu)/\nu}xy, \tag{B1}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(a\otimes d)=\mathscr{D}^{(2\nu-\lambda-\mu)/(2\nu)}a^{(\lambda+\mu-2\nu)/\nu}\{ad-p^{\mu}q^{-\nu}(A^{\nu-\mu}_{\nu})^2cb\}, \tag{B2}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(d\otimes a)=\mathscr{D}^{(2\nu-\lambda-\mu)/(2\nu)}a^{(\lambda+\mu-2\nu)/\nu}\{da+p^{\mu}(q^{-\nu}(A^{\nu-\mu}_{\nu})^2-p^{\lambda}(A^{2\nu-\lambda-\mu}_{\nu})^2)cb\}, \tag{B3}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(d\otimes d)=\mathscr{D}^{(2\nu-\lambda-\mu)/(2\nu)}a^{(\lambda+\mu-2\nu)/\nu}\{d^2-p^{\lambda+\mu}(A^{2\nu-\lambda-\mu}_{\nu})^2ca^{-1}db\}, \tag{B4}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(\mathscr{D}\otimes x)=C^{\nu}_{\lambda,\mu}(\mathscr{D},x)\mathscr{D}^{(3\nu-\mu)/(2\nu)}a^{(\mu-\nu)/\nu}x, \tag{B5}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(\mathscr{D}\otimes d)=\mathscr{D}^{(3\nu-\mu)/(2\nu)}a^{(\mu-\nu)/\nu}\{d-p^{\mu}(A^{\nu-\mu}_{\nu})^2ca^{-1}b\}, \tag{B6}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(x\otimes\mathscr{D})=C^{\nu}_{\lambda,\mu}(x,\mathscr{D})\mathscr{D}^{(3\nu-\lambda)/(2\nu)}a^{(\lambda-\nu)/\nu}x, \tag{B7}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(d\otimes\mathscr{D})=\mathscr{D}^{(3\nu-\lambda)/(2\nu)}a^{(\lambda-\nu)/\nu}\{d-p^{\lambda}(A^{\nu-\lambda}_{\nu})^2ca^{-1}b\}, \tag{B8}$$

$$\widetilde{m}^{\nu}_{p,q,\lambda,\mu}(\mathscr{D}\otimes\mathscr{D})=\mathscr{D}^2. \tag{B9}$$

In Eq. (B1), x,y run over $\{a,b,c,d\}$ with the exceptions of $(x,y)=(a,d)$, (d,a), (d,d), and

$$C^{\nu}_{\lambda,\mu}(x,y)=1,A^{\mu}_{\nu},q^{\mu-\nu}A^{\mu}_{\nu},p^{\mu-\nu}A^{\lambda}_{\nu},1,(pq)^{\mu-\nu}A^{\lambda}_{\nu}A^{\mu}_{\nu},p^{\mu-\nu}A^{\lambda}_{\nu},q^{\lambda+\mu-2\nu}A^{\lambda}_{\nu},$$
$$q^{\lambda+\mu-2\nu}A^{\lambda}_{\nu}A^{\mu}_{\nu},1,q^{\lambda+\mu-2\nu}A^{\lambda}_{\nu},A^{\mu}_{\nu},q^{\mu-\nu}A^{\mu}_{\nu}, \tag{B10}$$

for (x,y) listed in lexicographical order. Similarly, in Eqs. (B5) and (B7), x runs over $\{a,b,c\}$, and

$$C^{\nu}_{\lambda,\mu}(\mathscr{D},x)=1,A^{\mu}_{\nu},q^{\mu-\nu}A^{\mu}_{\nu}, \quad C^{\nu}_{\lambda,\mu}(x,\mathscr{D})=1,A^{\lambda}_{\nu},q^{\lambda-\nu}A^{\lambda}_{\nu}. \tag{B11}$$

For the colored unit and antipode, the results read

$$\tilde{\iota}^{\nu}_{p,q}(1_k)=\tilde{\iota}_{p^{\nu},q^{\nu}}(1_k), \tag{B12}$$

and

$$\tilde{S}^{\nu}_{p,q,\mu}(a)=\mathscr{D}^{(\mu-\nu)/(2\nu)}\{a^{-\mu/\nu}-p^{-\mu}q^{\mu}(A^{\mu}_{\nu})^2ca^{-(\mu+\nu)/\nu}\hat{d}^{-1}b\}, \tag{B13}$$

$$\tilde{S}^{\nu}_{p,q,\mu}(b)=-p^{-\mu}A^{\mu}_{\nu}a^{-(\mu+\nu)/(2\nu)}\hat{d}^{-(\mu+\nu)/(2\nu)}b, \tag{B14}$$

$$\tilde{S}^{\nu}_{p,q,\mu}(c)=-q^{\mu}A^{\mu}_{\nu}ca^{-(\mu+\nu)/(2\nu)}\hat{d}^{-(\mu+\nu)/(2\nu)}, \tag{B15}$$

$$\tilde{S}^{\nu}_{p,q,\mu}(d)=\mathscr{D}^{-(\mu-\nu)/(2\nu)}\hat{d}^{-\mu/\nu}, \tag{B16}$$

$$\tilde{S}^{\nu}_{p,q,\mu}(\mathscr{D})=\mathscr{D}^{-1}, \tag{B17}$$

respectively.

[1] C. Quesne, J. Math. Phys. **38**, 6018 (1997).
[2] C. Quesne, in *Quantum Group Symposium at Group 21 (Proceedings of the Quantum Group Symposium at the XXI International Colloquium on Group Theoretical Methods in Physics, Goslar, 1996*), edited by H. D. Doebner and V. K. Dobrev (Heron, Sofia, 1997), p. 219.
[3] V. V. Bazhanov and Yu. G. Stroganov, Theor. Math. Phys. **62**, 253 (1985); L. Hlavatý, J. Phys. A **20**, 1661 (1987); R. J. Baxter, J. H. Perk, and H. Au-Yang, Phys. Lett. A **128**, 138 (1988); B. S. Shastry, J. Stat. Phys. **50**, 57 (1988).
[4] The colored YBE, considered here and in I, should not be confused with the color YBE (Ref. 36) that arises when extending the graded YBE (Ref. 7) to more general gradings than that determined by $\mathbb{Z}_2$ (Ref. 37). As a consequence, our colored Hopf algebras are distinct from the Hopf color algebras (Ref. 36), generalizing Hopf superalgebras (Ref. 7) to such more general gradings.
[5] V. G. Drinfeld, in *Proceedings of the International Congress of Mathematicians*, Berkeley, CA, 1986, edited by A. M. Gleason (AMS, Providence, RI, 1987), p. 798; M. Jimbo, Lett. Math. Phys. **10**, 63 (1985); **11**, 247 (1986).
[6] C. Quesne, ''Coloured Hopf algebras,'' in *Proceedings of the Workshop on Special Functions and Differential Equations*, Chennai, India, January 13–24, 1997 (in press), q-alg/9705022.
[7] M. Chaichian and P. Kulish, Phys. Lett. B **234**, 72 (1990); W. B. Schmidke, S. P. Volos, and B. Zumino, Z. Phys. C **48**, 249 (1990).
[8] L. Faddeev, N. Reshetikhin, and L. Takhtajan, in *Algebraic Analysis*, edited by M. Kashiwara and T. Kawai (Academic, New York, 1988), Vol. 1, p. 129; in *Braid Group, Knot Theory and Statistical Mechanics*, edited by C. N. Yang and M. L. Ge (World Scientific, Singapore, 1989), p. 97.
[9] T. Ohtsuki, J. Knot Theor. Rami. **2**, 211 (1993).
[10] D. Bonatsos, P. Kolokotronis, C. Daskaloyannis, A. Ludu, and C. Quesne, Czech. J. Phys. **46**, 1189 (1996); D. Bonatsos, C. Daskaloyannis, P. Kolokotronis, A. Ludu, and C. Quesne, J. Math. Phys. **38**, 369 (1997).
[11] A. Kundu and B. Basu-Mallick, J. Phys. A **27**, 3091 (1994); B. Basu-Mallick, Mod. Phys. Lett. A **9**, 2733 (1994).
[12] B. Basu-Mallick, Int. J. Mod. Phys. A **10**, 2851 (1995).
[13] S. Majid, Int. J. Mod. Phys. A **5**, 1 (1990); V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994).
[14] C. Fronsdal and A. Galindo, Lett. Math. Phys. **27**, 59 (1993).
[15] C. Fronsdal and A. Galindo, Contemp. Math. **175**, 73 (1994); C. Fronsdal, in *Noncompact Lie Groups and Some of Their Applications (San Antonio, TX, 1993)*, NATO Advanced Science Institute Series C Math. Phys. Sci. 429 (Kluwer, Dordrecht, 1994), p. 423.
[16] F. Bonechi, E. Celeghini, R. Giachetti, C. M. Pereña, E. Sorace, and M. Tarlini, J. Phys. A **27**, 1307 (1994).
[17] A. Morozov and L. Vinet, ''Free-field representation of group element for simple quantum groups,'' Université de Montréal preprint CRM-2202, hep-th/9409093, 1994.
[18] R. Chakrabarti and R. Jagannathan, Z. Phys. C **72**, 519 (1996).
[19] R. J. Finkelstein, Lett. Math. Phys. **29**, 75 (1993).
[20] R. Jagannathan and J. Van der Jeugt, J. Phys. A **28**, 2819 (1995); J. Van der Jeugt and R. Jagannathan, Czech. J. Phys. **46**, 269 (1996).
[21] R. Chakrabarti and R. Jagannathan, Lett. Math. Phys. **37**, 191 (1996).
[22] A. Schirrmacher, J. Wess, and B. Zumino, Z. Phys. C **49**, 317 (1991).
[23] V. K. Dobrev, J. Math. Phys. **33**, 3419 (1992).
[24] Č. Burdík and P. Hellinger, J. Phys. A **25**, L629 (1992).
[25] R. Chakrabarti and R. Jagannathan, J. Phys. A **27**, 2023 (1994).

[26] In I, as well as in Secs. II and III of the present paper, we used the generic symbol q^{ν}, where ν behaved as a contravariant index, to denote the parameters of the final algebra under color group transformations. Since in this section, and in Sec. V, powers of the parameters make their appearance, to avoid confusion we slightly modify our notation q^{ν} into $q^{(\nu)}$.
[27] It should be noted that the images of a, b, c, d under ρ^{ν} or ρ_{ν} belong to the larger enveloping algebra $U_{p,q}(\{\alpha,\beta,\gamma,\delta\})$.
[28] L. Dabrowski and L. Wang, Phys. Lett. B **266**, 51 (1991).
[29] R. Chakrabarti and R. Jagannathan, J. Phys. A **24**, 5683 (1991).
[30] M. Bednář, Č. Burdík, M. Couture, and L. Hlavatý, J. Phys. A **25**, L341 (1992); Č. Burdík and R. Tomášek, Lett. Math. Phys. **26**, 97 (1992).
[31] Č. Burdík and P. Hellinger, J. Phys. A **25**, L1023 (1992).
[32] R. Chakrabarti and R. Jagannathan, Z. Phys. C **66**, 523 (1995).
[33] L. H. Kauffman and H. Saleur, Commun. Math. Phys. **141**, 293 (1991); L. Rozanskey and H. Saleur, Nucl. Phys. B **376**, 461 (1992).
[34] H. Hinrichsen and V. Rittenberg, Phys. Lett. B **275**, 350 (1992).
[35] Although we use the same symbol to denote the duals of $U_{p,q}(\mathrm{gl}(2))$ and $U_{p,q}(\mathrm{gl}(1/1))$, there is of course no relation between them.
[36] D. S. McAnally, in *Proceedings of the Yamada Conference XL, XX International Colloquium on Group Theoretical Methods in Physics*, Toyonaka, Japan, July 4–9, 1994, edited by A. Arima, T. Eguchi, and N. Nakanishi (World Scientific, Singapore, 1995), p. 339.
[37] V. Rittenberg and D. Wyler, Nucl. Phys. B **139**, 189 (1978); J. Math. Phys. **19**, 2193 (1978); J. Lukierski and V. Rittenberg, Phys. Rev. D **18**, 385 (1978); M. Scheunert, J. Math. Phys. **20**, 712 (1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 2 FEBRUARY 1998

Yang–Mills field equations for the Coriolis tensor in the *n*-body problem

Matthias Reinsch
Department of Physics, University of California, Berkeley, California 94720

(Received 3 September 1996; accepted for publication 3 November 1997)

The configuration space of a system of n particles has the structure of a principal fiber bundle under the action of the rotation group. In this paper, identities are derived for the curvature of the natural connection on this bundle. These identities are translated into statements about the Coriolis tensor, the local expression for this curvature on the base space of the bundle. The Coriolis tensor is shown to obey field equations of the Yang–Mills type. © *1998 American Institute of Physics.* [S0022-2488(98)02502-X]

I. INTRODUCTION

The subject of gauge fields in the separation of rotations and internal motions in the n-body problem has attracted much attention in recent years.[1,2] Early works in this area include those of Guichardet[3] and Iwai,[4] and numerous authors have contributed since then. The study is especially relevant to molecular physics, but could also be applied to atomic or nuclear physics, or even celestial mechanics. The reason for this large range of applications is that the system under consideration is quite general. It is that of n distinguishable nonrelativistic particles interacting via some potential that is invariant under overall translations and rotations of the system. Both the classical and quantum mechanical versions of the problem can be treated using the same gauge fields. This subject has been reviewed in Ref. 1, in which it is shown that the problem involves certain static fields that arise in a natural way. There are fields defined on the configuration space of the particles and fields defined on the internal space (shape space,[1,5] the space of all possible shapes of the n-particle system, explained below). The moment of inertia tensor with respect to the laboratory frame is an example of a field on configuration space. Other fields include the mechanical connection[6] and its curvature, to be discussed in the next section. This paper contains the derivation of several identities for this curvature. One of these identities has the form of Yang–Mills field equations.

The motivation for this study originated from an analysis of the n-body problem[1] in a way that involves fields on shape space. A field called the Coriolis tensor $\mathbf{B}_{\mu\nu}$ arises naturally in the dynamics. This is a static field defined on shape space. It is the curvature form in a non-Abelian gauge theory, in analogy with Yang–Mills theories. For the three-body problem, the Coriolis tensor has the form of a monopole on shape space.[4] It satisfies the identity $\mathbf{B}_{\mu\nu}{}^{;\nu}=0$ (this notation is explained later in this paper) in addition to the Bianchi identities that are known to hold for all values of n. We calculated $\mathbf{B}_{\mu\nu}{}^{;\nu}$ for the four-body problem and for several larger values of the number of particles and found that it vanished in these cases as well. Eventually we were able to prove the $\mathbf{B}_{\mu\nu}{}^{;\nu}=0$ identity for arbitrary values of the number of particles n, but the original proof was rather complex and required lengthy calculations. In this paper we present a simpler proof of the identity and also include proofs of new identities for the curvature of the mechanical connection.

The proof of the identities consists of two main parts. The first part deals with certain differential forms defined on the configuration space of the particles. These forms are the connection 1-form[7,8] and the curvature 2-form, defined in the next section. Such Lie-algebra-valued forms are common today in mathematics and theoretical physics, and it is interesting to see them appear in an area with as much history as the n-body problem. The second part of the proof involves a transition from quantities defined on configuration space to quantities defined on shape space. The identities derived in the first part of the paper are translated to identities for the Coriolis tensor.

In an Appendix, the relationship between covariant differentiation of tensors on shape space

 © 1998 American Institute of Physics

and corresponding quantities on configuration space is explained. This result is then used to state the Yang–Mills type identity for the Coriolis tensor in a way which involves simpler types of differentiation.

II. THE *n*-BODY PROBLEM

The recent realizations about gauge fields in the n-body problem are closely tied to a certain mathematical structure called a principal fiber bundle with a connection. We assume some familiarity with these concepts[7–9] on the part of the reader, but we do recall the basic definitions here. We begin by explaining how they relate to the problem at hand.

As stated above, the system under consideration is that of n nonrelativistic distinguishable particles interacting via a potential that is invariant under overall translations and rigid rotations of the system. The particles are in a three-dimensional Cartesian physical space with the usual Euclidean metric. The configuration space of the n-particle system is thus $\mathbb{R}^{3n}$. References 1 and 5 explain how the kinetic energy of the center-of-mass motion can be separated out by performing a linear change of coordinates on configuration space. The new coordinates are the center of mass of the n-particle system (a three-component vector) and $n-1$ mass-weighted Jacobi vectors, each of which is a three-component vector. The details of this transformation are not relevant to the present paper; we may take as our starting point the problem of $n-1$ unit-mass particles with only a rotational symmetry about a laboratory origin. References 1 and 5 explain how this problem is related to the original n-body problem. Briefly, the Jacobi vectors are linear combinations of the interparticle displacement vectors and are defined in such a way that the kinetic energy of the system is equal to the kinetic energy of the center of mass plus one-half the sum of the squares of the time derivatives of the $n-1$ Jacobi vectors. We call the configuration space of these $n-1$ unit-mass particles translation-reduced configuration space, or sometimes just ''configuration space'' for brevity. We use the symbol $\mathscr{C}$ for this space, and it is isomorphic to $\mathbb{R}^{3n-3}$. The Jacobi vectors are denoted by $\boldsymbol{\rho}_{s\alpha}, \alpha=1,\ldots,n-1$, with the subscript ''$s$'' indicating that the components of these vectors are referred to the space or laboratory frame, an inertial frame. Boldface is used to indicate three-component vectors. We use lower-case letters from the beginning of the Greek alphabet $\alpha,\beta,\ldots=1,\ldots,n-1$ to index the $n-1$ Jacobi vectors. The indices $i,j,k,\ldots=1,2,3$ are used to index components of vectors and matrices with respect to axes in three-dimensional physical space. In this notation, translation-reduced configuration space is parametrized by $3n-3$ coordinates, $\{\rho_{s\alpha i};\alpha=1,\ldots,n-1;i=1,2,3\}$, each of which runs from $-\infty$ to ∞.

The group of rotations in three-dimensional physical space [which is isomorphic to the matrix group $SO(3)$] acts on configuration space by left multiplication. An element R of $SO(3)$ defines an automorphism on configuration space by mapping a point $\{\rho_{s\alpha i};\alpha=1,\ldots,n-1;i=1,2,3\}$ to the point $\{R_{ij}\rho_{s\alpha j};\alpha=1,\ldots,n-1;i=1,2,3\}$ (here a sum on j is implied). For $n\geqslant 3$ the resulting structure is that of a principal fiber bundle, as discussed in Ref. 1. Shape space is the space of equivalence classes of configurations under the action of the rotation group, i.e., two configurations are defined to be equivalent if they differ by a proper rotation. Because $SO(3)$ is a three-dimensional Lie group, the dimensionality of shape space is $3n-6$. A technicality is that collinear configurations must be discarded from configuration space in order to obtain a principal fiber bundle. Aside from this technicality, the base space of the bundle is shape space. A pedagogical introduction to the fields defined on shape space that are relevant to the n-body problem is given in Ref. 1. In this paper, we will briefly recall the definitions of these fields as we encounter them in the proofs.

We will begin this discussion with fields defined on configuration space. An example is the moment of inertia tensor with respect to the laboratory axes. This is called the space moment of inertia tensor, M_s. The tensorial nature of this object is explained below. The definition of the usual space moment of inertia tensor can be expressed in terms of the mass-weighted Jacobi vectors,

$$(\mathsf{M}_s)_{ij}=\delta_{ij}\boldsymbol{\rho}_{s\alpha}\cdot\boldsymbol{\rho}_{s\alpha}-\rho_{s\alpha i}\rho_{s\alpha j}. \tag{2.1}$$

We use a summation convention in which repeated indices are summed over, so, for example, in the preceding equation there is a sum on α. As explained above, the subscript ''s'' is not an index.

The components $(\mathsf{M}_s)_{ij}$ of M_s would transform in the usual way under a rotation of the laboratory axes, and M_s would be considered a second rank tensor under rotations. However, we will not consider such redefinitions of the laboratory axes; they are assumed to be a fixed choice of axes in three-dimensional physical space. Thus, M_s is a 3×3 array of zero-forms (a matrix-valued function) on configuration space. It is a matrix-valued rank zero tensor on configuration space. We use the term ''moment of inertia tensor,'' because this is the standard term used in the literature. The moment of inertia tensor with respect to body-fixed axes, to be discussed later in this paper, is a different object and has more complicated transformation properties.

Another quantity defined on configuration space is the connection. The connection relevant to the n-body problem is called the mechanical connection in Ref. 6. A connection is a definition of a ''horizontal'' subspace of the tangent space of configuration space at each point in configuration space. The horizontal subspace at a point in configuration space is defined to be the set of vectors which correspond to instantaneous motion of the system with zero total angular momentum. In other words, the horizontal subspace is the set of vectors which give zero when acted on by the three dual vectors obtained by evaluating the three 1-forms,

$$\boldsymbol{\rho}_{s\alpha}\times d\boldsymbol{\rho}_{s\alpha}, \tag{2.2}$$

at the point in configuration space. The cross in this expression denotes the usual cross product of three-component vectors, and the result is a three-component vector (an ordered triple) of one-forms on configuration space. Written out explicitly, the three 1-forms are

$$\epsilon_{ijk}\rho_{s\alpha j}d\rho_{s\alpha k}, i=1,2,3, \tag{2.3}$$

the ϵ_{ijk} tensor being the usual totally antisymmetric tensor with $\epsilon_{123}=1$. For each value of the index i, there is a sum on α, j and k, and the result is a 1-form on configuration space. Since we have three conditions, the dimensionality of each horizontal subspace is three less than that of configuration space; i.e., it is $3n-6$.

A. An expression for the connection

In the previous section, the connection was introduced as a definition of horizontal subspaces. In order to calculate its curvature, it will be useful to describe the connection using a ''connection 1-form,'' as described in references such as Refs. 7 and 8. We use the symbol ϕ to denote the connection 1-form (and later Φ for the curvature 2-form), as in Ref. 8. The differential form ϕ is a Lie-algebra-valued 1-form on configuration space with the property that when acting on a tangent vector it returns the Lie-algebra element that corresponds to the vertical part of the tangent vector. The definition of the vertical part of a tangent vector and the correspondence between Lie-algebra elements and vertical vectors is as follows. The Lie algebra of the group $SO(3)$ is isomorphic to the set of antisymmetric 3×3 matrices. Given an antisymmetric 3×3 matrix X, the function $\exp(t\mathsf{X})$ is a parametrized curve in the group $SO(3)$ with parameter t. If one takes a point in configuration space and lets $\exp(t\mathsf{X})$ act on it using the left action defined earlier, one obtains a curve in configuration space. At the parameter value $t=0$, this curve identifies a vector in the tangent space of configuration space at the specified point. This vector is defined to be the vertical vector corresponding to X. An arbitrary vector in the tangent space of configuration space can be written as the sum of a horizontal and vertical part in a unique way, so the map that takes a vector to the Lie algebra element corresponding to the vector's vertical part is well-defined. In Sec. III we will show that there is a simple relationship between the connection 1-form and the vector potential $\mathbf{A}_\mu$ used in Ref. 1.

Next, we work toward an explicit expression for the connection 1-form. We start by introducing some convenient notation. As a basis for Lie algebra of $SO(3)$ we use the three elements corresponding to the antisymmetric 3×3 matrices $\mathsf{J}_\ell, \ell=1,2,3$. The J_ℓ matrices are the usual generators of rotations, whose entries are

$$(\mathsf{J}_\ell)_{ij}=-\epsilon_{\ell ij}. \tag{2.4}$$

The choice of a basis for the Lie algebra of $SO(3)$ results in a one-to-one correspondence between ordered triples of real numbers and the elements of the Lie algebra. Since the connection 1-form

is a Lie-algebra-valued form, this correspondence can be used to define an $\mathbb{R}^3$-valued 1-form, which we denote by boldface $\boldsymbol{\phi}$. The explicit expression for $\boldsymbol{\phi}$ works out to be

$$\boldsymbol{\phi} = \mathsf{M}_s^{-1} \cdot (\boldsymbol{\rho}_{s\alpha} \times d\boldsymbol{\rho}_{s\alpha}), \tag{2.5}$$

as in Ref. 6. In order to verify this expression for $\boldsymbol{\phi}$, one must show that it returns the correct values when acting on a basis of the tangent space of configuration space at a point in configuration space. That the expression correctly returns a zero for horizontal vectors follows from the definition of the horizontal subspace in Eq. (2.2). As for the vertical vectors, we calculate the result of $\mathsf{M}_s^{-1} \cdot (\boldsymbol{\rho}_{s\alpha} \times d\boldsymbol{\rho}_{s\alpha})$ acting on the vertical vector corresponding to J_{ℓ}. When $d\boldsymbol{\rho}_{s\alpha}$ acts on this vector the result is $\mathsf{J}_{\ell}\boldsymbol{\rho}_{s\alpha}$, as can be seen from the discussion of vertical vectors in the previous paragraph and from the definition of the left action of $SO(3)$ on configuration space, defined earlier. Thus, the result of letting $\mathsf{M}_s^{-1} \cdot (\boldsymbol{\rho}_{s\alpha} \times d\boldsymbol{\rho}_{s\alpha})$ act on this vector is $\mathsf{M}_s^{-1} \cdot [\boldsymbol{\rho}_{s\alpha} \times (\mathsf{J}_{\ell}\boldsymbol{\rho}_{s\alpha})]$. Using the ''BAC-CAB'' rule, Eq. (A3), this simplifies to a unit vector with a one at position ℓ, and completes the verification of Eq. (2.5). In this proof of Eq. (2.5), we have used the compact three-vector notation. Alternatively, one can work with $\mathbb{R}$-valued differential forms. The i-th component of the right-hand side of Eq. (2.5) is $(\mathsf{M}_s^{-1})_{ij}\epsilon_{jkm}\rho_{s\alpha k}d\rho_{s\alpha m}$, an $\mathbb{R}$-valued 1-form. The result of this 1-form acting on the vertical vector is $(\mathsf{M}_s^{-1})_{ij}\epsilon_{jkm}\rho_{s\alpha k}(\mathsf{J}_{\ell}\boldsymbol{\rho}_{s\alpha})_m$ which works out to be $\delta_{i\ell}$, using Eq. (2.1) and Eq. (A3). Physically speaking, the right-hand side of Eq. (2.5) may be thought of as a mapping which maps a specification of particle positions and velocities to an angular-velocity three-vector. This angular velocity has the property that the total angular momentum of the system is obtained by multiplying it by M_s.

B. Curvature of the connection

Having defined the connection, we can move on to its curvature. The curvature is defined below and expresses the fact that if a closed path in the base space is lifted horizontally[1,7–10] into configuration space, the resulting path in general will not be closed. A standard way to describe this curvature is with a curvature 2-form. This is a Lie-algebra-valued 2-form on configuration space. Again we will use the correspondence between elements of the Lie algebra and elements of $\mathbb{R}^3$ (induced by our choice of basis) to define an $\mathbb{R}^3$-valued curvature 2-form, which we will denote by boldface $\boldsymbol{\Phi}$. The curvature 2-form is defined to be the exterior derivative of the connection 1-form, composed with projection onto the horizontal subspace. This definition results in a standard formula for the curvature 2-form $\boldsymbol{\Phi}$ in terms of the connection 1-form $\boldsymbol{\phi}$,

$$\Phi^i = d\phi^i - \frac{1}{2}\epsilon_{ijk}\phi^j \wedge \phi^k. \tag{2.6}$$

Here Φ^i denotes the i-th component of the vector $\boldsymbol{\Phi}$, and ϕ^i denotes the i-th component of the vector $\boldsymbol{\phi}$.

Next, we consider writing $\boldsymbol{\phi}$ and $\boldsymbol{\Phi}$ with respect to a coordinate basis of differential forms on configuration space. We start with arbitrary coordinates x^a on configuration space. The index a runs from 1 to $3n-3$, and we write

$$\boldsymbol{\phi} = \boldsymbol{\phi}_a dx^a,$$

$$\boldsymbol{\Phi} = \frac{1}{2}\boldsymbol{\Phi}_{ab} dx^a \wedge dx^b. \tag{2.7}$$

For fixed values of the indices a and b, the quantities $\boldsymbol{\phi}_a$ and $\boldsymbol{\Phi}_{ab}$ are three-component vector-valued functions on configuration space. Now we use

$$d\boldsymbol{\phi} = \boldsymbol{\phi}_{a,b} dx^b \wedge dx^a$$

$$= \frac{1}{2}(\boldsymbol{\phi}_{b,a} - \boldsymbol{\phi}_{a,b}) dx^a \wedge dx^b, \tag{2.8}$$

where the comma denotes differentiation, to obtain

$$\mathbf{\Phi}_{ab} = \boldsymbol{\phi}_{b,a} - \boldsymbol{\phi}_{a,b} - \boldsymbol{\phi}_a \times \boldsymbol{\phi}_b \,. \tag{2.9}$$

This expression will be used to obtain a coordinate expression for the curvature 2-form.

C. A coordinate expression for the curvature of the connection

The coordinates $\{\rho_{s\alpha i}\}$ on configuration space are useful for doing calculations. Thus, we now take the x^a coordinates in Eq. (2.7) to be the $\rho_{s\alpha i}$ coordinates. The index a becomes a pair, (αi), written without a comma in order to avoid confusion with a notation for differentiation. From Eq. (2.5) we see that $\phi^i_{(\alpha \not{)}} = (\mathsf{M}_s^{-1})_{ij} \epsilon_{jk \not{)}} \rho_{s\alpha k}$, so that

$$\boldsymbol{\phi}_{(\alpha \not{)}} = -\mathsf{M}_s^{-1} \mathsf{J}_{\not{)}} \boldsymbol{\rho}_{s\alpha} \,. \tag{2.10}$$

With this expression at hand, Eq. (2.9) can be used to calculate the coefficients $\mathbf{\Phi}_{(\alpha j)(\beta k)}$ of the curvature 2-form with respect to the coordinate basis. This is done in Appendix A. The result may be written in the form

$$\Phi^i_{(\alpha j)(\beta k)} = 2(\mathsf{M}_s^{-1})_{im} \epsilon_{mjk} \delta_{\alpha\beta} + \boldsymbol{\rho}_{s\alpha} \cdot \mathsf{C}^i_{jk} \cdot \boldsymbol{\rho}_{s\beta} \,, \tag{2.11}$$

where $\delta_{\alpha\beta}$ is the Kronecker delta symbol, and the quantity C^i_{jk} is a 3×3 matrix for each set of values of the indices i, j and k, given by

$$\begin{aligned} \mathsf{C}^i_{jk} = {} & [2\mathbf{e}_j \otimes \mathbf{e}_i \mathsf{M}_s^{-1} - \mathsf{M}_s^{-1} \mathbf{e}_i \otimes \mathbf{e}_j - (\mathsf{M}_s^{-1})_{ij} \mathsf{I}] \mathsf{M}_s^{-1} \mathsf{J}_k \\ & + \mathsf{J}_j \mathsf{M}_s^{-1} [2\mathsf{M}_s^{-1} \mathbf{e}_i \otimes \mathbf{e}_k - \mathbf{e}_k \otimes \mathbf{e}_i \mathsf{M}_s^{-1} - (\mathsf{M}_s^{-1})_{ik} \mathsf{I}] - \mathsf{J}_j \mathsf{M}_s^{-1} \mathsf{J}_i \mathsf{M}_s^{-1} \mathsf{J}_k \,. \end{aligned} \tag{2.12}$$

Here the symbol $\mathbf{e}_i$ is a three-component vector that has a one at position i and zeroes for the other two entries; in other words, it is a unit vector in the direction of the i-th laboratory or ''space'' axis (these axes were fixed at the beginning of the problem, as described at the beginning of Sec. II). The symbol $\otimes$ denotes a tensor product, so that $\mathbf{e}_i \otimes \mathbf{e}_j$ is a 3×3 matrix whose entries are all zeroes except the entry at position (i,j), which is a one. The symbol I is the 3×3 identity matrix.

D. Identities for the curvature

The expression for the coefficients $\Phi^i_{(\alpha j)(\beta k)}$ given in Eqs. (2.11) and (2.12) can be used to derive some simple identities for the curvature 2-form. We start by setting j equal to i in Eq. (2.12) and summing on i. This results in

$$\begin{aligned} \mathsf{C}^i_{ik} = {} & [\mathsf{M}_s^{-1} - \mathrm{Tr}(\mathsf{M}_s^{-1}) \mathsf{I}] \mathsf{M}_s^{-1} \mathsf{J}_k \\ & + \mathsf{J}_i \mathsf{M}_s^{-1} [-\mathbf{e}_k \otimes \mathbf{e}_i \mathsf{M}_s^{-1} - (\mathsf{M}_s^{-1})_{ik} \mathsf{I}] - \mathsf{J}_i \mathsf{M}_s^{-1} \mathsf{J}_i \mathsf{M}_s^{-1} \mathsf{J}_k \,, \end{aligned} \tag{2.13}$$

where the symbol Tr denotes the trace of a matrix. Now we use the ''BAC-CAB'' rule in the form $\mathsf{J}_i \mathsf{M}_s^{-1} \mathsf{J}_i = \mathsf{M}_s^{-1} - \mathrm{Tr}(\mathsf{M}_s^{-1}) \mathsf{I}$ to obtain

$$\mathsf{C}^i_{ik} = 0. \tag{2.14}$$

This equation will be used in proving identities about the curvature 2-form which will then be used to prove the Yang–Mills type fields equations.

The three identities for the curvature 2-form that will be of use to us are

$$\Phi^j_{(\alpha j)(\beta k)} = 0, \tag{2.15}$$

$$\Phi^i_{(\alpha j)(\beta k)} \boldsymbol{\rho}_{s\alpha j} = 0, \tag{2.16}$$

$$\Phi^i_{(\alpha j)(\beta k)} \boldsymbol{\rho}_{s\alpha i} = 0. \tag{2.17}$$

As always, the repeated indices in these equations are summed over. Equation (2.15) follows from the $\mathsf{C}^i_{ik}=0$ identity and the coordinate expression for the curvature 2-form, Eq. (2.11). The definition of the space moment of inertia tensor M_s, given in Eq. (2.1), is used in the proof of Eq. (2.16). A proof of Eq. (2.17) is given in Appendix B.

The identities in Eqs. (2.15), (2.16) and (2.17) make reference to the coefficients $\Phi^i_{(\alpha j)(\beta k)}$ of the curvature 2-form with respect to a coordinate basis. These identities may also be written in ways that do not make reference to these expansion coefficients. Equation (2.15) is equivalent to

$$\frac{\partial}{\partial \boldsymbol{\rho}_{s\alpha j}} \lrcorner \Phi^j = 0, \tag{2.18}$$

where the symbol $\lrcorner$ is used for the inner product of a vector and a differential form, and a common notation for vectors has been used: $\partial/\partial\rho_{s\alpha j}$ is the vector obtained by holding all of the coordinates except $\rho_{s\alpha j}$ fixed and using the coordinate $\rho_{s\alpha j}$ as the parameter for a parametrized curve.

The identity in Eq. (2.16) may also be written in a form similar in appearance to Eq. (2.18). We start by considering the quantity $\mathrm{Tr}\mathsf{M}_s$, a scalar field on configuration space. The exterior derivative of this quantity is

$$d\mathrm{Tr}\mathsf{M}_s = d(2\boldsymbol{\rho}_{s\alpha}\cdot\boldsymbol{\rho}_{s\alpha}) = 4\boldsymbol{\rho}_{s\alpha}\cdot d\boldsymbol{\rho}_{s\alpha}. \tag{2.19}$$

The Euclidean metric on configuration space[1] makes possible a correspondence between 1-forms and vectors. The symbol $\sharp$ is used to indicate the vector corresponding to a 1-form. The vector corresponding to $d\mathrm{Tr}\mathsf{M}_s$ is the gradient of $\mathrm{Tr}\mathsf{M}_s$ and is given by

$$(d\mathrm{Tr}\mathsf{M}_s)^\sharp = 4\boldsymbol{\rho}_{s\alpha}\cdot\frac{\partial}{\partial\boldsymbol{\rho}_{s\alpha}}. \tag{2.20}$$

Now the identity in Eq. (2.16) may be written in the form

$$(d\mathrm{Tr}\mathsf{M}_s)^\sharp \lrcorner \boldsymbol{\Phi} = 0. \tag{2.21}$$

The identity in Eq. (2.17) can be written as a sum of inner products of vectors with 2-forms,

$$\rho_{s\alpha i}\frac{\partial}{\partial\boldsymbol{\rho}_{s\alpha j}} \lrcorner \Phi^i = 0. \tag{2.22}$$

In other words, for any fixed value of the index j, the sum on i of the inner product of the vector $\rho_{s\alpha i}(\partial/\partial\rho_{s\alpha j})$ with the 2-form Φ^i vanishes.

III. TRANSITION TO QUANTITIES ON SHAPE SPACE

The identities for the curvature 2-form on configuration space derived in the previous section can be translated into statements about fields defined on shape space, the base space of the rotation-group fiber bundle described at the beginning of Sec. II. The Coriolis tensor,[1] to be described below, is the field on shape space that corresponds to the curvature 2-form. In Ref. 1 the Coriolis tensor is derived from a vector potential. We start by showing how this vector potential is related to the connection 1-form and how the Coriolis tensor is related to the curvature 2-form. These relationships are standard in the study of connections on fiber bundles. The concept of a local section of the fiber bundle[1,7,8] is central to the understanding of these relationships. A local section is a differentiable mapping from an open subset U of shape space to configuration space $\mathscr{C}$, with the property that a point in shape space is mapped to a point in configuration space that is in the equivalence class identified by the point in shape space. We use the symbol σ for such a mapping, and use the notation $\{\boldsymbol{\rho}_\alpha(q)\}$ (not to be confused with the $\rho_{s\alpha i}$ coordinates on configuration space) for the image of a shape q under this mapping,

$$\sigma: U \to \mathscr{C}, \tag{3.1}$$

$$q \mapsto \{\boldsymbol{\rho}_\alpha(q)\}.$$

As in Ref. 1, we use the notation $q^\mu, \mu = 1, \ldots, 3n-6$ for arbitrary coordinates on shape space.

The local section may be thought of as an assignment of a body frame[1] to each shape in the domain of definition of σ. This makes it possible to define a body-frame moment of inertia tensor M,

$$\mathsf{M} = \mathsf{M}_s \circ \sigma. \tag{3.2}$$

The $\circ$ denotes functional composition. M is a 3×3 matrix-valued function on shape space that depends on the choice of local section. We note that since M_s is an array of 0-forms, Eq. (3.2) may also be written as $\mathsf{M} = \sigma^* \mathsf{M}_s$, where σ^* denotes the pullback operation, which is discussed in the next section.

The components of the body-frame moment of inertia tensor M are found from the definition of M_s, Eq. (2.1), to be

$$M_{ij} = \delta_{ij} \boldsymbol{\rho}_\alpha \cdot \boldsymbol{\rho}_\alpha - \rho_{\alpha i} \rho_{\alpha j}. \tag{3.3}$$

A. Local expressions for the connection and its curvature

Other fields on configuration space can be pulled back[7–9] by a local section to give fields on shape space. First we will consider the pullback of the connection 1-form $\boldsymbol{\phi}$ by the local section σ. This is denoted $\sigma^* \boldsymbol{\phi}$ and is a Lie algebra valued 1-form on shape space. This form evaluated on a vector in the tangent space of shape space at a point in shape space is found in the following way. If the vector is called $v = v^\mu \partial/\partial q^\mu$, one first pushes v forward by the local section σ to a vector on configuration space,

$$v^\mu \frac{\partial \boldsymbol{\rho}_\alpha}{\partial q^\mu} \cdot \frac{\partial}{\partial \boldsymbol{\rho}_{s\alpha}}. \tag{3.4}$$

Then one lets the connection 1-form $\boldsymbol{\phi}$ act on this vector,

$$(\sigma^* \boldsymbol{\phi}) v = \boldsymbol{\phi}\left(v^\mu \frac{\partial \boldsymbol{\rho}_\alpha}{\partial q^\mu} \cdot \frac{\partial}{\partial \boldsymbol{\rho}_{s\alpha}} \right). \tag{3.5}$$

The right-hand side of this equation can be evaluated using the expression for the connection 1-form given in Eq. (2.5). The result is

$$(\sigma^* \boldsymbol{\phi}) v = \mathsf{M}^{-1} \cdot \left(\boldsymbol{\rho}_\alpha \times \frac{\partial \boldsymbol{\rho}_\alpha}{\partial q^\mu} \right) v^\mu. \tag{3.6}$$

We recognize the right-hand side of Eq. (3.6) as the vector potential,[1]

$$\mathbf{A} = \mathbf{A}_\mu dq^\mu = \mathsf{M}^{-1} \cdot \left(\boldsymbol{\rho}_\alpha \times \frac{\partial \boldsymbol{\rho}_\alpha}{\partial q^\mu} \right) dq^\mu, \tag{3.7}$$

acting on the vector v. Thus, we have shown that the relationship between the connection 1-form $\boldsymbol{\phi}$ and the vector potential $\mathbf{A}$ is

$$\mathbf{A} = \sigma^* \boldsymbol{\phi}. \tag{3.8}$$

Next, we move on to the relationship between the Coriolis tensor and the curvature 2-form. We start by letting the pullback of $\boldsymbol{\Phi}$ by the local section act on the vectors $\partial/\partial q^\mu$ and $\partial/\partial q^\nu$. Using a procedure similar to that used above for the connection 1-form, one obtains

$$(\sigma^*\mathbf{\Phi})\left(\frac{\partial}{\partial q^\mu},\frac{\partial}{\partial q^\nu}\right)=\mathbf{A}_{\nu,\mu}-\mathbf{A}_{\mu,\nu}-\mathbf{A}_\mu\times\mathbf{A}_\nu,$$

$$=\mathbf{B}_{\mu\nu}. \tag{3.9}$$

Equation (2.9) has been used in this calculation. The final equality in Eq. (3.9) expresses the definition of the Coriolis tensor given in Ref. 1. Equation (3.9) shows that

$$\mathbf{B}=\sigma^*\mathbf{\Phi}, \tag{3.10}$$

where the Lie-algebra-valued 2-form $\mathbf{B}$ on shape space is defined to be

$$\mathbf{B}=\frac{1}{2}\mathbf{B}_{\mu\nu}dq^\mu\wedge dq^\nu. \tag{3.11}$$

B. Identities for quantities on shape space

The relationships between fields on shape space and fields on configuration space derived in the previous section can be used to translate the identities for the curvature derived in Sec. II into identities for the fields on shape space. There is one more equation that is necessary in order to carry out this process. This is a certain statement about the vector fields $\partial/\partial\rho_{s\alpha i}$ on configuration space and the vector fields $\partial/\partial q^\mu$ on shape space. The statement also involves the projection π from configuration space to shape space. This projection is part of the definition of configuration space as a rotation group fiber bundle. It maps a point in configuration space to the equivalence class to which it belongs. A vector on configuration space such as $\partial/\partial\rho_{s\alpha i}$ can be projected onto shape space. The notation for this projected vector (the ''push-forward'' by the projection) is

$$\pi_*\frac{\partial}{\partial\rho_{s\alpha i}}.$$

The point at which the vector $\partial/\partial\rho_{s\alpha i}$ is based is suppressed in this notation. We will consider the case in which this point is ''on the gauge surface,'' that is, the point is the image under the local section σ of a point q in shape space. The projected vector may be expressed in terms of the basis $\{\partial/\partial q^\mu,\mu=1,\ldots,3n-6\}$ of the tangent space of shape space at the point q. This can be done by calculating the Jacobian of the coordinate transformation on configuration space from the $\rho_{s\alpha i}$ coordinates to the coordinates that are the $3n-6$ shape coordinates q^μ and three coordinates on the rotation group. Such a local coordinate transformation is possible once a local section of the bundle has been chosen. The Jacobian of this coordinate transformation is presented in Ref. 1, and the result for the projected vector is

$$\pi_*\frac{\partial}{\partial\rho_{s\alpha i}}=g^{\mu\nu}\rho_{\alpha i;\nu}\frac{\partial}{\partial q^\mu}. \tag{3.12}$$

In this equation, $g^{\mu\nu}$ denotes the inverse of the metric tensor on shape space, and $\boldsymbol{\rho}_{\alpha;\nu}=\boldsymbol{\rho}_{\alpha,\nu}-\mathbf{A}_\nu\times\boldsymbol{\rho}_\alpha$ is the covariant derivative of $\boldsymbol{\rho}_\alpha$ with respect to the coordinate q^ν, as explained in Ref. 1. The metric $g_{\mu\nu}dq^\mu\otimes dq^\nu$ is the metric induced on shape space by the (rotationally invariant) Euclidean metric on configuration space. Covariant differentiation is also discussed in Appendix C.

Next, we consider projecting other vectors from configuration space to shape space. The vector $(d\mathrm{TrM}_s)^\sharp$ that appears in Eq. (2.20) is a linear combination of the vectors in the $\{\partial/\partial\rho_{s\alpha i}\}$ basis,

$$(d\mathrm{TrM}_s)^\sharp=4\rho_{s\alpha i}\frac{\partial}{\partial\rho_{s\alpha i}}. \tag{3.13}$$

Thus, it projects according to

$$\pi_*[(d\mathrm{Tr}\mathsf{M}_s)^\sharp]=4\rho_{\alpha i}\pi_*\frac{\partial}{\partial\rho_{s\alpha i}}=g^{\mu\nu}(2\rho_{\alpha i}\rho_{\alpha i})_{;\nu}\frac{\partial}{\partial q^\mu}=(\mathrm{Tr}\mathsf{M})^{;\mu}\frac{\partial}{\partial q^\mu}, \tag{3.14}$$

where the index in the final expression has been raised with $g^{\mu\nu}$, as is standard in tensor analysis, and we note that the covariant derivative of a scalar such as $\rho_{\alpha i}\rho_{\alpha i}$ is equal to the ordinary derivative of the scalar.

We are now in a position to calculate $\mathbf{B}_{\mu\nu}(\mathrm{Tr}\mathsf{M})^{;\nu}$, a quantity that will be needed later in the proof. The question is what happens when the 2-form $\mathbf{B}$ acts on the vector $(\mathrm{Tr}\mathsf{M})^{;\nu}\partial/\partial q^\nu$ and any other vector. According to Eq. (3.10), this equals $\mathbf{\Phi}$ acting on the vectors pushed forward by the local section σ. Thus, we are interested in the inner product of $\mathbf{\Phi}$ with the push-forward of $(\mathrm{Tr}\mathsf{M})^{;\nu}\partial/\partial q^\nu$ by σ. However, since the original definition of $\mathbf{\Phi}$ involved projection onto the horizontal subspace (the form $\mathbf{\Phi}$ is ''horizontal'' in the terminology of Ref. 8), we can consider the inner product of $\mathbf{\Phi}$ with any vector that projects to the vector $(\mathrm{Tr}\mathsf{M})^{;\nu}\partial/\partial q^\nu$ on shape space. Equation (3.14) shows that $(d\mathrm{Tr}\mathsf{M}_s)^\sharp$ is such a vector. The vanishing of the inner product of $\mathbf{\Phi}$ with this vector, shown in Eq. (2.21), thus results in

$$\mathbf{B}_{\mu\nu}(\mathrm{Tr}\mathsf{M})^{;\nu}=0. \tag{3.15}$$

In a similar manner, we now calculate $\mathbf{B}_{\mu\nu}\cdot\boldsymbol{\rho}_\alpha{}^{;\nu}$. The center dot denotes a contraction on the three-vector indices; written out this is $B^i_{\mu\nu}\rho_{\alpha i}{}^{;\nu}$. Thus, we are considering a sum on i of the inner product of the 2-form B^i on shape space with the vector $\rho_{\alpha i}{}^{;\nu}\partial/\partial q^\nu$. As in the previous calculation, Eq. (3.12) leads us to the inner product of the 2-form Φ^i on configuration space with the vector $\partial/\partial\rho_{s\alpha i}$. Equation (2.18) shows that when this is summed on i, the result is zero. Therefore, we have shown

$$\mathbf{B}_{\mu\nu}\cdot\boldsymbol{\rho}_\alpha{}^{;\nu}=0. \tag{3.16}$$

This result can be obtained by other means, but the calculations are lengthy and do not use geometrical ideas.

The third of the identities for the curvature 2-form on configuration space, Eq. (2.17), enables us to show

$$B^i_{\mu\nu}\rho_{\alpha i}\rho_{\alpha j}{}^{;\nu}=0, \tag{3.17}$$

which may be seen as follows. For any value of the free index j, the left-hand side of this equation, contracted with dq^μ, is a 1-form. This 1-form is a sum on i of the inner product of the 2-form B^i on shape space with the vector $\rho_{\alpha i}\rho_{\alpha j}{}^{;\nu}\partial/\partial q^\nu$. As above, for each value of the index i, Eq. (3.12) leads us to the inner product of the 2-form Φ^i on configuration space with the vector $\rho_{s\alpha i}\partial/\partial\rho_{s\alpha j}$. Equation (2.22) shows that when this is summed on i, the result is zero, providing the proof of Eq. (3.17).

Equations (3.15), (3.16) and (3.17) make it possible to prove the Yang–Mills type identity $\mathbf{B}_{\mu\nu}{}^{;\nu}=0$ for the Coriolis tensor on shape space. One begins by calculating $\mathsf{M}^{;\nu}\cdot\mathbf{B}_{\mu\nu}$. The covariant derivative of the moment of inertia tensor on shape space $\mathsf{M}_{;\mu}$ is given by

$$M_{ij;\mu}=\frac{1}{2}\delta_{ij}(\mathrm{Tr}\mathsf{M})_{;\mu}-\rho_{\alpha i;\mu}\rho_{\alpha j}-\rho_{\alpha i}\rho_{\alpha j;\mu}, \tag{3.18}$$

[see Eq. (3.3)]. Thus

$$\begin{aligned}M_{ij}{}^{;\nu}B^j_{\mu\nu}&=\frac{1}{2}B^i_{\mu\nu}(\mathrm{Tr}\mathsf{M})^{;\nu}-B^j_{\mu\nu}\rho_{\alpha i}{}^{;\nu}\rho_{\alpha j}-B^j_{\mu\nu}\rho_{\alpha i}\rho_{\alpha j}{}^{;\nu},\\&=0,\end{aligned} \tag{3.19}$$

where Eqs. (3.15), (3.16) and (3.17) have been used.

Having proved the identity $\mathsf{M}^{;\nu}\cdot\mathbf{B}_{\mu\nu}=0$, we now use the fact that configuration space is Riemann flat, in order to arrive at the identity $\mathbf{B}_{\mu\nu}{}^{;\nu}=0$. This involves a translation of the Rie-

mann flatness of configuration space into a number of identities for the fields on shape space. Such calculations are done in Ref. 10 in a general setting, and in Ref. 1 for the case of the rotational symmetry in the n-body problem. Among the resulting identities is the relation

$$2(\mathsf{M}\cdot\mathbf{B}_{\mu\tau})_{;\nu}-\mathsf{M}_{;\mu}\cdot\mathbf{B}_{\tau\nu}-\mathsf{M}_{;\tau}\cdot\mathbf{B}_{\nu\mu}=0. \tag{3.20}$$

Contracting this with $g^{\tau\nu}$ results in

$$(\mathsf{M}\cdot\mathbf{B}_{\mu\nu})^{;\nu}=0, \tag{3.21}$$

which may be expanded into two terms using the Leibnitz rule. The relation $\mathsf{M}^{;\nu}\cdot\mathbf{B}_{\mu\nu}=0$ implies that in fact both of these terms vanish. This completes the proof of the Yang–Mills type identity for the Coriolis tensor on shape space,

$$\mathbf{B}_{\mu\nu}{}^{;\nu}=0. \tag{3.22}$$

In summary, we have derived several simple identities, Eqs. (2.15), (2.16) and (2.17), for the curvature 2-form on configuration space, and shown that together with the Riemann flatness of configuration space these identities imply that the Coriolis tensor $\mathbf{B}_{\mu\nu}$ obeys field equations of the Yang–Mills type.

ACKNOWLEDGMENTS

The author would like to thank Jerrold Marsden and Claudio Emmrich for useful conversations. This work was supported by the U. S. Department of Energy under contract No. DE-AC03-76SF00098 and by the Alice J. Rosenberg Fellowship for graduate study in Physics at the University of California at Berkeley.

APPENDIX A: A COORDINATE EXPRESSION FOR THE CURVATURE 2-FORM

The subject of this appendix is the proof of the coordinate expression for the curvature 2-form given in Eq. (2.11). Using the coordinates $\{\rho_{s\alpha i}\}$ on configuration space, Eq. (2.9) reads as

$$\mathbf{\Phi}_{(\alpha j)(\beta k)}=\boldsymbol{\phi}_{(\beta k),(\alpha j)}-\boldsymbol{\phi}_{(\alpha j),(\beta k)}-\boldsymbol{\phi}_{(\alpha j)}\times\boldsymbol{\phi}_{(\beta k)}. \tag{A1}$$

When the coordinate expression for $\boldsymbol{\phi}$ given in Eq. (2.10) is substituted into this equation, it is necessary to differentiate the moment of inertia tensor M_s. In doing this, it is helpful to write the definition of M_s in the form

$$\mathsf{M}_s=-\rho_{s\alpha i}\rho_{s\alpha j}\mathsf{J}_i\mathsf{J}_j. \tag{A2}$$

The equivalence of this expression and the definition of the moment of inertia tensor given in Eq. (2.1) can be seen from the definition of the J matrices [Eq. (2.4)] and the ''BAC-CAB'' rule, Eq. (A3). The expression that results for $\Phi^i_{(\alpha j)(\beta k)}$ may be separated into two parts. The first part is the first term on the right-hand side of Eq. (2.11). The second part is a linear combination of the nine quantities $\rho_{s\alpha i}\rho_{s\beta j}; i,j=1,2,3$. The coefficients in this linear combination are functions of the location in configuration space and can be expressed in terms of the moment of inertia tensor. One way to write the final expression is as a matrix sandwiched between the vectors $\boldsymbol{\rho}_{s\alpha}$ and $\boldsymbol{\rho}_{s\beta}$. The ''BAC-CAB'' rule,

$$\epsilon_{ijk}\epsilon_{i\ell m}=\delta_{j\ell}\delta_{km}-\delta_{jm}\delta_{k\ell}, \tag{A3}$$

is useful in doing this. The final result for the coordinate expression for Φ^i is given in Eq. (2.11).

APPENDIX B: PROOF OF AN IDENTITY FOR THE CURVATURE

This appendix contains a summary of the calculation involved in proving the identity Eq. (2.17) for the curvature 2-form. One starts with the coordinate expression for the curvature 2-form given in Eq. (2.11). This is then contracted with $\rho_{s\alpha i}$ as shown in Eq. (2.17). The definition of the

space moment of inertia tensor, Eq. (2.1), is used to eliminate the explicit occurrences of sums over Jacobi vectors. One is left with an expression that may be written as a matrix contracted with the vectors $\mathbf{e}_j$ and $\boldsymbol{\rho}_{s\beta}$. This matrix,

$$[\mathrm{Tr}\mathsf{M}_s(\mathsf{M}_s^{-1}-\mathsf{I}\mathrm{Tr}\mathsf{M}_s^{-1})+2\mathsf{I}-M_{s\not m}\mathsf{J}_{\not m}\mathsf{M}_s^{-1}\mathsf{J}_m]\mathsf{M}_s^{-1}\mathsf{J}_k+(\mathsf{M}_s^{-1})_{\not k}\mathsf{J}_{\not}+\mathsf{J}_k\mathsf{M}_s^{-1},$$

can be shown to vanish using only the fact that M_s is a 3×3 symmetric matrix. In doing this, one uses the Cayley–Hamilton Theorem, which may be written as

$$\mathsf{M}_s^3-(\mathrm{Tr}\mathsf{M}_s)\mathsf{M}_s^2=\det(\mathsf{M}_s)\{\mathsf{I}-[\mathrm{Tr}(\mathsf{M}_s^{-1})]\mathsf{M}_s\}, \tag{B1}$$

for a 3×3 matrix. One also uses the fact that $\mathsf{M}_s\mathsf{J}_k+\mathsf{J}_k\mathsf{M}_s$ is antisymmetric and can therefore be expressed in terms of the J matrices,

$$\mathsf{M}_s\mathsf{J}_k+\mathsf{J}_k\mathsf{M}_s=\mathsf{J}_{\not}(\delta_{\not k}\mathrm{Tr}\mathsf{M}_s-\mathsf{M}_{s\not k}). \tag{B2}$$

APPENDIX C: COVARIANT DIFFERENTIATION

The subject of this appendix is the relationship between covariant differentiation of fields defined on shape space[1] and covariant differentiation of corresponding quantities on configuration space. Covariant differentiation on configuration space has a simple definition, involving only the metric connection,[7,8] and it is easily carried out in the $\{\rho_{s\alpha i}\}$ coordinates because this space is Riemann flat.

We are especially interested in the curvature 2-form; the projection onto shape space of the covariantly differentiated $\boldsymbol{\Phi}$ is equal to the covariant derivative of the Coriolis tensor on shape space. This relationship is used to translate the Yang–Mills identity for the Coriolis tensor, Eq. (3.22), into a statement about the curvature 2-form on configuration space.

1. Covariant differentiation of the curvature 2-form on configuration space

We begin with the covariant derivative of the curvature 2-form on configuration space. The first step is to express this Lie-algebra-valued 2-form with respect to the basis of the Lie algebra of $SO(3)$ defined in Eq. (2.4), as discussed in Sec. II. This results in three real-valued 2-forms $\Phi^i, i=1,2,3$. The covariant derivative of such a 2-form is defined using the metric connection on configuration space. In the language of tensor analysis, this means that for a choice of coordinates on configuration space, one calculates Christoffel symbols and uses these to obtain the covariant derivative of Φ^i_{ab} as a second-rank tensor. This covariant derivative involves a term that contains ordinary directional differentiation and two terms that involve Christoffel symbols. If $\{\rho_{s\alpha i}\}$ coordinates are used, the Christoffel symbols vanish because the metric tensor is constant in these coordinates. Thus, we may describe covariant differentiation of the curvature 2-form on configuration space as follows. Given a vector U in the tangent space of configuration space at a point, one obtains $\nabla_U\boldsymbol{\Phi}$, the covariant derivative of $\boldsymbol{\Phi}$ in the direction U, by first expressing $\boldsymbol{\Phi}$ with respect to the basis $d\rho_{s\alpha i}\wedge d\rho_{s\beta j}$ as in Eq. (2.7),

$$\boldsymbol{\Phi}=\frac{1}{2}\boldsymbol{\Phi}_{(\alpha i)(\beta j)}d\rho_{s\alpha i}\wedge d\rho_{s\beta j}, \tag{C1}$$

and then applying directional differentiation to these coefficients,

$$\nabla_U\boldsymbol{\Phi}=\frac{1}{2}U[\boldsymbol{\Phi}_{(\alpha i)(\beta j)}]d\rho_{s\alpha i}\wedge d\rho_{s\beta j}, \tag{C2}$$

where the square brackets denote directional differentiation.

2. Relation between covariant differentiation on configuration space and on shape space

Next we turn to the subject of the relationship between covariant differentiation on shape space and the simple covariant differentiation on configuration space. The result is

$$\mathbf{B}_{;\mu}=\sigma^{*}[(\nabla_{e_\mu}\mathbf{\Phi})\circ\Pi_h], \tag{C3}$$

where e_μ is the horizontal vector defined in Ref. 1 (the horizontal lift of the vector $\partial/\partial q^\mu$), $\mathbf{B}_{;\mu}$ is defined by

$$\mathbf{B}_{;\mu}=\frac{1}{2}\mathbf{B}_{\lambda\nu;\mu}dq^\lambda\wedge dq^\nu, \tag{C4}$$

and Π_h denotes projection onto the horizontal subspaces (the orthogonal complements of the vertical subspaces). The methods in Ref. 1 can be used to verify Eq. (C3). This equation expresses the fact that the projection onto shape space of the covariant derivative of $\mathbf{\Phi}$ on configuration space is the covariant derivative of the Coriolis tensor on shape space.

We recall the definition of the covariant derivative of the Coriolis tensor on shape space from Ref. 1,

$$\mathbf{B}_{\mu\nu;\sigma}=\mathbf{B}_{\mu\nu,\sigma}-\mathbf{A}_\sigma\times\mathbf{B}_{\mu\nu}-\Gamma^\tau_{\sigma\mu}\mathbf{B}_{\tau\nu}-\Gamma^\tau_{\sigma\nu}\mathbf{B}_{\mu\tau}. \tag{C5}$$

In this equation, the quantities $\Gamma^\tau_{\sigma\mu}$ are the Christoffel symbols calculated from the metric tensor $g_{\mu\nu}$ on shape space. Each of the terms in this equation can be obtained by carrying out the operations on the right-hand side of Eq. (C3).

3. The Yang–Mills identity

It is possible to express the Yang–Mills field equations for the Coriolis tensor, Eq. (3.22), as a statement about quantities defined on configuration space, i.e., without the use of local sections of the bundle. One way to do this is to start with an arbitrary orthonormal basis $\{u_{(i)}, i=1,\ldots,3n-6\}$ of the tangent space of shape space, at some point q. This means that $u^\mu_{(i)}g_{\mu\nu}u^\nu_{(j)}=\delta_{ij}$, which implies $u^\mu_{(i)}u^\nu_{(i)}=g^{\mu\nu}$. Thus, the $\mathbf{B}_{\mu\nu}{}^{;\nu}=0$ identity may be written as

$$\mathbf{B}_{\mu\nu;\lambda}u^\nu_{(i)}u^\lambda_{(i)}=0. \tag{C6}$$

This equation translates into a statement about the curvature 2-form with the help of Eq. (C3),

$$u_{(j)}\lrcorner\sigma^{*}[(\nabla_{u^\mu_{(j)}e_\mu}\Phi)\circ\Pi_h]=0. \tag{C7}$$

The horizontal lifts of the vectors $\{u_{(i)}, i=1,\ldots,3n-6\}$ form an orthonormal basis of the horizontal subspace of the tangent space. This follows from the definition of the induced metric on shape space. The Yang–Mills identity may now be stated as follows. Given any orthonormal basis $\{U_{(i)}, i=1,\ldots,3n-6\}$ of the horizontal subspace of the tangent space of configuration space, we have

$$(U_{(j)}\lrcorner\nabla_{U_{(j)}}\Phi)\circ\Pi_h=0. \tag{C8}$$

We note that this may also be written as $U_{(j)}\lrcorner[(\nabla_{U_{(j)}}\Phi)\circ\Pi_h]=0$. An equivalent way to state the contents of Eq. (C8) is to say that the form $U_{(j)}\lrcorner\nabla_{U_{(j)}}\Phi$ is vertical.[8] Also, we recall again that the covariant differentiation in Eq. (C8) is simple, involving only directional differentiation if the $\{\rho_{sai}\}$ coordinates are used.

[1] R. G. Littlejohn and M. Reinsch, ''Gauge fields in the separation of rotations and internal motions in the *n*-body problem,'' Rev. Mod. Phys. **69**, 213–275 (1997).
[2] A. Shapere and F. Wilczek, in *Geometric Phases in Physics* (World Scientific, Singapore, 1989).
[3] A. Guichardet, Ann. Inst. Henri Poincaré **40**, 329 (1984).
[4] T. Iwai, J. Math. Phys. **28**, 1315 (1987).
[5] R. G. Littlejohn and M. Reinsch, Phys. Rev. A **52**, 2035 (1995).
[6] J. E. Marsden, *Lectures on Mechanics* (Cambridge University Press, Cambridge, 1992).
[7] M. Nakahara, *Geometry, Topology, and Physics* (Hilger, New York, 1990).

[8] R. L. Bishop and R. J. Crittenden, *Geometry of Manifolds* (Academic, NewYork, 1964).
[9] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[10] R. Coquereaux, *Riemannian Geometry, Fiber Bundles, Kaluza-Klein Theories and All That*, World Scientific Lecture Notes in Physics (World Scientific, Singapore, 1988), Vol. 16.

The inverse loop transform

T. Thiemann[a)]
Albert Einstein Institut, Max Planck Institut für Gravitationsphysik, Schlaatzweg 1, 14473 Potsdam, Germany

(Received 3 March 1997; accepted for publication 7 October 1997)

The loop transform in quantum gauge field theory can be recognized as the Fourier transform (or characteristic functional) of a measure on the space of generalized connections modulo gauge transformations. Since this space is a compact Hausdorff space, conversely, we know from the Riesz–Markov theorem that every positive linear functional on the space of continuous functions thereon qualifies as the loop transform of a regular Borel measure on the moduli space. In the present article we show how one can compute the finite joint distributions of a given characteristic functional, that is, we derive the inverse loop transform. © *1998 American Institute of Physics.* [S0022-2488(98)00302-8]

I. INTRODUCTION

Quite some time ago[1] it became clear that for the canonical formulation of quantum gauge field theories it seems to be of advantage to work in a so-called ''loop representation.'' By this is meant the following: Let A be a smooth connection of a compact gauge group G on some manifold Σ and let α be a loop in Σ. From these data one can construct the holonomy (or parallel transport) $h_\alpha(A)$ of A along α. One knows[2] that the gauge invariant information of a smooth connection is encoded in the holonomies of A as one considers all possible loops. Thus, a gauge invariant function f of A may be written as a function of holonomies. The idea behind the loop representation is then to define a measure μ on a space of connections modulo gauge transformations ''$\mathcal{A}/\mathcal{G}$'' (we set inverted commas because, as one can imagine, the support of the measure will not be contained in the space of smooth connections modulo gauge transformations $\mathcal{A}/\mathcal{G}$) and to define the loop transform of f as a kind of Fourier transform via $\tilde{f}(\alpha) := \int d\mu(A)\overline{\mathrm{tr}[h_\alpha(A)]}f(A)$.

This idea has been rediscovered by Rovelli and Smolin[3] in the context of canonical quantum gravity based on the observation by Ashtekar[4] that general relativity can be formulated as a theory of SL(2, $\mathbb{C}$) connections. The virtue of these complex-valued variables is that they allow for a polynomial formulation of the theory that for a highly nonpolynomial theory as general relativity is a major breakthrough.

The status of the loop transform had been rather formal then for several reasons: First of all, for gauge theories for compact gauge groups, the measure μ was never specified, partly because it was unknown over which space of connections to integrate. Second, for quantum gravity the situation was even worse since the gauge group was noncompact, and this made several of the techniques developed for gauge theories inaccessible. Finally, in Ref. 5 a quantum configuration space $\overline{\mathcal{A}/\mathcal{G}}$ of generalized (distributional) connections was defined and in Refs. 6 and 7 measures on this space were constructed. This put the loop transform on a mathematically rigorous foundation, at least for compact gauge groups. Expectedly, the space of smooth connections modulo gauge transformations $\mathcal{A}/\mathcal{G}$ was seen to be contained in a measure zero measurable set with respect to the measures μ constructed.[8]

On the mathematical side, several additional quite powerful techniques associated with $\overline{\mathcal{A}/\mathcal{G}}$ have been constructed, such as differential geometry[9] on $\overline{\mathcal{A}/\mathcal{G}}$ based on its description as a projective limit,[8,10] the complete set of diffeomorphism invariant states[11] for any diffeomorphism invariant quantum gauge field theory for a compact gauge group and a coherent state transform from a theory for compact gauge groups to a theory with noncompact gauge groups.[12]

[a)]Electronic mail: thiemann@aei-potsdam.mpg.de

 © 1998 American Institute of Physics

On the physical side, in applications to quantum gravity a Wick rotation transform was developed[13,14] that uses the techniques of Ref. 12, but, in addition, solves the problem of the nonpolynomial reality conditions on the complex-valued Ashtekar variables. Suffice it to say that due to a recently developed technique (see Ref. 15 and references therein), following Ref. 16, a Wick rotation transform is not necessary anymore so that the framework developed in Refs. 5–8 can be immediately applied to equip quantum gravity with a Hilbert space structure. For an account of these various transforms, including the one discussed in the present paper, the reader is referred to Ref. 17. In applications to gauge field theory a new constructive quantum field theory framework in the sense of Ref. 18 was developed,[19,20] and, in particular, applied to two-dimensional Yang–Mills theory.[21]

In this paper we make yet another contribution to this subject, namely we construct the analog of the inverse Fourier transform on $\mathbb{R}^n$,[22] which we call the ''inverse loop transform.'' This will allow us to reconstruct a measure as defined by its finite joint distributions from its characteristic functional. In particular, given a (singular) knot invariant satisfying certain additional conditions, we can find out to what measure it corresponds.

The plan of the paper is as follows.

In Sec. II we recall the basic notions from calculus on $\overline{\mathcal{A}/\mathcal{G}}$ as far as necessary for the present context. The interested reader is referred to the article in Ref. 11 for further details.

In Sec. III we first formulate and prove the analog of the inverse Fourier transform theorem for compact gauge groups and then give a new definition of the loop transform based on the notion of ''loop networks.'' [23,24] This notion will prove useful in proving the inverse loop transform.

In Sec. IV we display the inverse loop transform of physically interesting characteristic functionals.

In an Appendix we indicate how all the results of this paper can be rewritten in terms of edges rather than loops, which is sometimes more convenient in applications beyond integration on $\overline{\mathcal{A}/\mathcal{G}}$. In particular, we introduce the notion of an ''edge network'' that generalizes the notion of a ''spin network'' [23,24] to an arbitrary compact gauge group and also allows us to give (a) a closed and compact expression for a spin-network state without referring to a graphical notation,[23] and (b) an explicit algebraic expression for the in[24] only abstractly defined ''intertwiners'' for an arbitary compact gauge group.

II. PRELIMINARIES

We give here only the absolutely necessary information in order to fix the notation. For further details see Ref. 11 and references therein.

The space $\overline{\mathcal{A}/\mathcal{G}}$ of generalized connections modulo gauge transformations is the Gel'fand spectrum of the Abelian C^* algebra generated by the Wilson-loop functionals for smooth connections, that is, traces of the holonomy for piecewise analytic loops in the base manifold Σ. As such, it is a compact Hausdorff space and therefore measures on that space are in one to one correspondence with positive linear functionals on $C(\overline{\mathcal{A}/\mathcal{G}})$.

A certain natural measure μ_0 will play a very crucial role in this article so that we go now into more detail.

In what follows, $\gamma \subset \Sigma$ will always denote a finite, unoriented, piecewise analytic graph, meaning that it is the union of a finite number of analytic edges and vertices. Its fundamental group $\pi_1(\gamma)$ is then finitely generated by some loops $\beta_1(\gamma),\ldots,\beta_{n(\gamma)}(\gamma)$ that we fix once and for all together with some orientation and that are based at some arbitrary but fixed base point $p \in \Sigma$, $n(\gamma) := \dim(\pi_1(\gamma))$ being the number of independent generators of the fundamental group of γ. A function f on $\overline{\mathcal{A}/\mathcal{G}}$ is said to be cylindrical with respect to a graph $\gamma, f \in \mathrm{Cyl}_\gamma(\overline{\mathcal{A}/\mathcal{G}})$, if it is a function only of the finite set of arguments $p_\gamma(A) := (h_{\beta_1}(A),\ldots,h_{\beta_n}(A))$, where $h_\alpha(A)$ is the holonomy of A along the loop α. A measure μ is now specified by its finite joint distributions μ_γ, which are defined by

$$\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu(A) f(A) = \int_{G^n} d\mu_\gamma(g_1,\ldots,g_n) f_\gamma(g_1,\ldots,g_n), \tag{II.1}$$

where $f=f_\gamma\circ p_\gamma$ and $f_\gamma: G^m\to\mathbb{C}$ is a gauge invariant function. In order that this definition makes sense, we have to make sure that if we write $f=f_\gamma\circ p_\gamma=f_{\gamma'}\circ p_{\gamma'}$ in two different ways as a cylindrical function where $\gamma\subset\gamma'$ is a subgraph of γ', then we should have that the so-called consistency conditions

$$\int_{G^n} d\mu_\gamma\, f_\gamma=\sum_{G^{n'}} d\mu_{\gamma'}\, f_{\gamma'}, \tag{II.2}$$

are satisfied.

The natural measure μ_0 is the induced Haar measure, meaning that $d\mu_{0,\gamma}(g_1,\ldots,g_n)=d\mu_H(g_1)\ldots d\mu_H(g_n)$. One can check that the consistency conditions are satisfied[6] and that the so-defined cylindrical measure has a σ-additive extension μ_0 on the projective limit measurable space of the family of measurable spaces $\overline{\mathcal{A}/\mathcal{G}}_\gamma$.[8] The space $\overline{\mathcal{A}/\mathcal{G}}_\gamma$ is defined to be the set of all homomorphisms from the group of based loops restricted to γ into the gauge group modulo conjugation while $\overline{\mathcal{A}/\mathcal{G}}$ is the set of all homomorphisms from the whole loop group into G modulo conjugation. Note that the semigroup of loops with respect to compositions of loops can be given a group structure by identifying paths that are traversed in the opposite direction with the inverse of the original path.

III. THE INVERSE LOOP TRANSFORM

A. The inverse Fourier transform for compact groups

Let us recall some basic facts from harmonic analysis on compact gauge groups.[25]

Definition III.1: Let $\{\pi\}$ denote the set of all finite-dimensional, nonequivalent (we fix one representant from each equivalence class once and for all), unitary, irreducible representations of the compact gauge group G, let d_π be the dimension of π, and let μ_H be the normalized Haar measure on G.

For any $f\in L_1(G,d\mu_H)$ define the Fourier transform of f by

$$\hat{f}^{ij}_\pi:=\int_G d\mu_{\rm H}(g)\sqrt{d_\pi}\,\overline{\pi_{ij}}(g)f(g),\quad i,j=1,\ldots,d_\pi, \tag{III.1}$$

where $\pi_{ij}(g)$ denotes the matrix elements of $\pi(g)$.

Note that this definition makes sense because the matrix elements of $\pi(g)$ are bounded by 1.

Definition III.2: The Fourier transform of a function is said to be ℓ_1 or ℓ_2, respectively, iff,

$$\|\hat{f}\|_1:=\sum_\pi\sum_{i,j=1}^{d_\pi}\sqrt{d_\pi}|f^{ij}_\pi|<\infty\quad\text{or}\quad\|\hat{f}\|_2:=\sum_\pi\sum_{i,j=1}^{d_\pi}|f^{ij}_\pi|^2<\infty. \tag{III.2}$$

The Fourier series associated with a function f on G such that $\hat{f}\in\ell_1$ is given by

$$\widetilde{f}(g):=\sum_\pi\sum_{i,j=1}^{d_\pi}\hat{f}^{ij}_\pi\pi_{ij}(g)\sqrt{d_\pi}. \tag{III.3}$$

The analog of the Plancherel theorem for $\mathbb{R}^n$ is the Peter and Weyl theorem.

Theorem III.1 (Peter and Weyl): (1) *The functions $g\to\sqrt{d_\pi}\pi_{ij}(g)$, $i,j=1,\ldots,d_\pi$ form a complete and orthonormal system on $L_2(G,d\mu_H)$.*

(2) *For any $f\in L_2(G,d\mu_H)$ it holds that $f=\widetilde{f}$ in the sense of L_2 functions and the Fourier transform is a unitary map $\wedge: L_2(G,d\mu_H)\to\ell_2$.*

The author was unable though to find the analog of the inverse Fourier transform theorem for groups in the literature that we therefore prove here. This theorem answers the question of whether a function that is only L_1 can be represented, in the L_1 sense, by its Fourier transform. Although the proof that follows applies directly to a compact group only for which a different, less complicated proof is available, the method of proof below hints at how to extend the result to locally compact groups and is therefore of interest.

Theorem III.2: *Let $f\in L_1(g,d\mu_H)$ such that also $\hat{f}\in\ell_1$. Then $f(g)=\widetilde{f}(g)$ on L_1.*

Proof: Let be given any $\hat{k} \in \ell_1$. Then the Fourier series

$$\tilde{f}_k := \sum_{\pi,i,j} \left[\sum_k f_\pi^{ij} \overline{k_\pi^{jk}} \right] \pi_{ij}(g) \sqrt{d_\pi} \tag{III.4}$$

still converges absolutely, as can be seen from the following considerations: Upon using the Schwarz inequality we obtain the estimate

$$\left| \sum_k f_\pi^{ik} \overline{k_\pi^{jk}} \right| \leq \sum_k |f_\pi^{ik}||k_\pi^{jk}| \leq \sqrt{\sum_k |f_\pi^{ik}|^2} \sqrt{\sum_l |k_\pi^{jk}|^2} \leq \left[\sum_k |f_\pi^{ik}| \right] \left[\sum_l |k_\pi^{jl}| \right],$$

so that

$$\|\tilde{f}_k\|_1 \leq \sum_\pi \left[\sum_{i,k} |f_\pi^{ik}| \sqrt{d_\pi} \right] \left[\sum_{j,l} |k_\pi^{jl}| \sqrt{d_\pi} \right] \leq \left[\sum_{\pi,i,k} |f_\pi^{ik}| \sqrt{d_\pi} \right] \left[\sum_{\sigma,j,l} |k_\sigma^{jl}| \sqrt{d_\sigma} \right] = \|\hat{f}\|_1 \|\hat{k}\|_1 . \tag{III.5}$$

Hence (III.4) is well defined, and we may write

$$\begin{aligned} \tilde{f}_k(g) &= \sum_{\pi,j,k} \sqrt{d_\pi} \overline{k_\pi^{jk}} \left[\int d\mu_{\mathrm{H}}(h) \pi(h^{-1}g)_{kj} f(h) \right] \\ &= \sum_{\pi,j,k} \overline{k_\pi^{jk}} \sqrt{d_\pi} \left[\int d\mu_{\mathrm{H}}(h) \pi(h^{-1})_{kj} (L_g f)(h) \right] \\ &= \int d\mu_{\mathrm{H}}(h) (L_g f)(h) \overline{\sum_{\pi,j,k} \sqrt{d_\pi} k_\pi^{jk} \pi_{jk}(h)} \\ &= \int d\mu_{\mathrm{H}}(h) (L_g f)(h) \bar{\tilde{k}}(h). \end{aligned} \tag{III.6}$$

In the second line we have made use of the translation invariance of the Haar measure and the unitarity of the representation, $(L_g f)(h) = f(gh)$ is the definition of the left regular representation of G, in the third line we could switch integration and summation because $\hat{k} \in \ell_1$ and $\overline{\pi_{jk}} L_g f \in L_1(G, d\mu_{\mathrm{H}})$, and finally in the last step we have used the definition of the Fourier series (III.3).

We now choose $k_\pi^{ij} := \sqrt{d_\pi} \pi_{ij}(1) e^{-t\lambda_\pi}$, where λ_π are the eigenvalues of the Casimir operator Δ in the representation π. Then $\tilde{k}(g) = \rho_t(g)$ becomes the heat kernel on G,[26] that is, the fundamental solution of the equation

$$\left[\frac{\partial}{\partial t} - \Delta \right] \rho_t(g) = 0, \quad \rho_0(g) = \delta_{\mu_{\mathrm{H}}}(g, 1). \tag{III.7}$$

The motivation for this choice is of course that for $t \to 0$ the lhs of (III.6) tends to $\tilde{f}(g)$ while the rhs should tend to $f(g)$. Indeed, since ρ_t tends to the δ distribution on G wrt the Haar measure, this would be straightforward to see if $f \in C^\infty(G)$. To show that this is true even for $f \in L_1(G, d\mu_{\mathrm{H}})$, we argue as follows: denote $(f \star \rho_t)(g) := \int d\mu_{\mathrm{H}}(h) (L_g f)(h) \rho_t(h)$; then we have

$$\begin{aligned} \|f \star \rho_t - f\|_1 &= \int d\mu_{\mathrm{H}}(g) \left| \int d\mu_{\mathrm{H}}(h) [R_h f - f](g) \rho_t(h) \right| \\ &\leq \int d\mu_{\mathrm{H}}(g) \int d\mu_{\mathrm{H}}(h) |R_h f - f|(g) \rho_t(h), \end{aligned} \tag{III.8}$$

where we have used the normalization and positivity of the heat kernel and $(R_h f)(g) = f(gh)$.

The idea is now to split the integration domain of the inner integral into a compact neighborhood U of the identity of μ_{H} volume δ and its complement $G - U$ in G:

$$\|f\star\rho_t - f\|_1 \leq \int_G d\mu_{\rm H}(g) \int_U d\mu_{\rm H}(h) |R_h f - f|(g)\rho_t(h)$$
$$+ \int_G d\mu_{\rm H}(g) \int_{G-U} d\mu_{\rm H}(h) |R_h f - f|(g)\rho_t(h)$$
$$=: {\rm I} + {\rm II}. \tag{III.9}$$

Since the heat kernel gets concentrated at the identity for $t\to 0$, we can estimate the second integral while we will estimate the first integral by a compactness argument.

The details are as follows. We first assume that $f\in C(G)$. This assumption will be dropped later again. Then f is uniformly continuous on the compact set $gU=\{R_h g; h\in U\}$ and therefore there is a non-negative function $\omega(\delta)$, $\lim_{\delta\to 0}\omega(\delta)=0$, such that $|R_h f - f|(g)\leq\omega(\delta)$, $\forall h\in U$. Accordingly,

$${\rm I}\leq\omega(\delta)\int_U d\mu_{\rm H}(h)\rho_t(h)\leq\omega(\delta), \tag{III.10}$$

due to the normalization of the heat kernel.

The continuous function $(g,h)\to(R_h f - f)(g)\rho_t(h)$ is measurable on $G\times G$ wrt the Borel measure $\mu_{\rm H}\times\mu_{\rm H}$ (recall Ref. 26 that ρ_t is even real analytic). Since $|\int_G d\mu_{\rm H}(h)\rho_t(h)\int d\mu_{\rm H}(g) \times(R_h f - f)(g)|\leq 2\|f\|_1<\infty$, the theorem of Fubini allows us to switch the integrations in the second integral, and we arrive at

$${\rm II} = \int_{G-U} d\mu_{\rm H}(h)\rho_t(h)\int_G d\mu_{\rm H}(g)|R_h f - f|(g)$$
$$= \int_{G-U} d\mu_{\rm H}(h)\rho_t(h)\|R_h f - f\|_1$$
$$\leq 2\|f\|_1\int_{G-U} d\mu_{\rm H}(h)\rho_t(h). \tag{III.11}$$

Now for any $\epsilon>0$ we find a $\delta(\epsilon)$ such that $\omega(\delta(\epsilon))<\epsilon/2$ and for this so chosen $\delta(\epsilon)$ we find a $t(\epsilon,f)$ so that $2\|f\|_1\int_{G\text{-}U} d\mu_{\rm H}\rho_t<\epsilon/2$, since the support of ρ_t gets more and more concentrated at the identity for $t\to 0$. Therefore we conclude that for any $f\in C(G)$, $\epsilon>0$ there exists a $t(\epsilon)>0$ such that $\|\rho_t\star f - f\|_1<\epsilon$.

Now let us focus on a general $f\in L_1(G,d\mu_{\rm H})$ and consider an arbitrary $k\in C(G)$. Then we expand

$$\|\rho_t\star f - f\|_1 \leq \|\rho_t\star(f-k)-(f-k)\|_1 + \|\rho_t\star k - k\|_1. \tag{III.12}$$

Now G is a compact Hausdorff space and $\mu_{\rm H}$ is a Borel measure on G so that $C(G)$ is dense in $L_1(G,d\mu_{\rm H})$.[27] We can therefore find a k such that $\|f-k\|_1<\epsilon/4$ and therefore $\|\rho_t\star(f-k)-(f-k)\|\leq 2\|f-k\|_1\|\rho_t\|_1\leq\epsilon/2$ for any $t>0$ while we have shown above that for any given $k\in C(G)$ we can always choose t sufficiently small such that $\|\rho_t\star k - k\|_1<\epsilon/2$.

This furnishes the proof. □

The theorem can obviously be extended to functions of more than one variable. This will enable us to define the finite-dimensional joint distributions of a measure.

B. Computation of the finite joint distributions of a measure

Recall[28] that every representation of a compact group is equivalent to a unitary one, so that we may restrict ourselves to unitary representations in the sequel. Also, every such representation is completely reducible. In what follows we will assume that we have fixed, in each equivalence class of irreducible representations that arise in the decomposition into irreducibles of a tensor product of irreducible representations (the ones that were fixed in Definition 3.1), a standard base

of independent representations that project onto orthogonal representation spaces. For the case of SU(2) this is the familiar Clebsh–Gordan decomposition and for SU(n) this can be established, for instance, by choosing the representations associated with the standard tableaux of the corresponding Young diagrams.[28] For the general case of a compact group we assume to have made a similar choice.

First we introduce a new notion.

Definition III.3: (i) A loop network is a triple $(\gamma, \boldsymbol{\pi}, \pi)$ *consisting of a graph* γ, *a vector* $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_{n(\gamma)})$ *of irreducible representations of* G *and an irreducible representation* π *of* G *that takes values in the set of irreducible representations of* G *contained in the decomposition into irreducibles of the tensor product* $\otimes_{k=1}^{n}\pi_k$.

(ii) A loop network state is a map from $\overline{\mathcal{A}/\mathcal{G}}$ *into* $\mathbb{C}$ *defined by*

$$T_{\gamma,\boldsymbol{\pi},\pi}(A):=\mathrm{tr}[\otimes_{k=1}^{n(\gamma)}\pi_k(h_{\beta_k(\gamma)}(A))\cdot c(\boldsymbol{\pi},\pi)], \tag{III.13}$$

where the matrix c *is defined by*

$$c(\boldsymbol{\pi},\pi):=\sqrt{\frac{\Pi_{k=1}^{n(\gamma)}d_{\pi_k}}{d_\pi}}\,\pi(1). \tag{III.14}$$

Loop-network states satisfy the following important properties.

Lemma III.1: (i) Given a graph γ, *the set of all loop network states provides an orthonormal basis of* $L_2(\overline{\mathcal{A}/\mathcal{G}}_\gamma, d\mu_{0,\gamma})=L_2(\overline{\mathcal{A}/\mathcal{G}}, d\mu_0)\cap Cyl_\gamma(\overline{\mathcal{A}/\mathcal{G}})$.

(ii) Given a graph γ', *consider all its subgraphs* $\gamma<\gamma'$. *Remove all the loop-network states on* γ' *that are pull-backs of loop-network states on* γ. *The collection of all loop-network states so obtained provides an orthonormal basis of* $L_2(\overline{\mathcal{A}/\mathcal{G}}, d\mu_0)$.

Proof: (i) The orthogonality relations for loop-network states on a given graph γ follow easily from basic group integration theory. By the Peter and Weyl theorem we have

$$\langle T_{\gamma,\boldsymbol{\pi},\pi}, T_{\gamma,\boldsymbol{\pi}',\pi'}\rangle=\overline{c(\boldsymbol{\pi},\pi)}^{(i_1,j_1),\ldots,(i_n,j_n)}c(\boldsymbol{\pi}',\pi')^{(k_1,l_1),\ldots,(k_n,l_n)}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\,\frac{1}{\Pi_{k=1}^{n}d_{\pi_k}}\,\delta_{i_1,k_1}\delta_{j_1,l_1}$$

$$\cdot\,\delta_{i_n,k_n}\delta_{j_n,l_n}=\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\,\frac{1}{\sqrt{d_\pi d_{\pi'}}}\,\mathrm{tr}[\pi^\dagger(1)\pi'(1)]=\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\delta_{\pi,\pi'}, \tag{III.15}$$

where we have used that the nonequivalent irreducible—as well as our choice of equivalent—representations of a compact gauge groups are orthogonal, that is, $\pi(1)$ is a projector. The completeness of these states on $L_2(\overline{\mathcal{A}/\mathcal{G}}_\gamma, d\mu_{0,\gamma})$ follows also from the Peter and Weyl theorem together with a gauge-invariance argument. We know that the states

$$T_{\gamma,\boldsymbol{\pi}}^{(i_1,j_1)..(i_n,j_n)}:=[\otimes_{k=1}^{n}\pi_k(h_{\beta_k(\gamma)}(A))]^{(i_1,j_1),\ldots(i_n,j_n)}, \tag{III.16}$$

contain an overcomplete set of states for $L_2(\overline{\mathcal{A}/\mathcal{G}}_\gamma, d\mu_{0,\gamma})$, and thus we only need to select all the independent gauge-invariant combinations of those, that is, we need to find all the matrices c, called contractors, which turn (III.16) into gauge invariant states when being contracted with them. Notice that all the generators β are based loops. Therefore under a gauge transformation,

$$\mathrm{tr}(T_{\gamma,\boldsymbol{\pi}}\cdot c)\to\mathrm{tr}[T_{\gamma,\boldsymbol{\pi}}\cdot(\otimes_{k=1}^{n}\pi_k)(g^{-1})\cdot c\cdot(\otimes_{k=1}^{n}\pi_k)(g)], \tag{III.17}$$

and gauge invariance requires choosing c such that

$$(\otimes_{k=1}^{n}\pi_k)(g^{-1})\cdot c\cdot(\otimes_{k=1}^{n}\pi_k)(g)=c \quad \text{for all } g\in G. \tag{III.18}$$

Now notice that the matrix c in (III.17) is already projected on the reducible representation space defined by the tensor product representation $\otimes_{k=1}^{n}\pi_k$ because the matrix $(\otimes_{k=1}^{n}\pi_k)(1)$ is a projector on that space and leaves the matrix $T_{\gamma\boldsymbol{\pi}}$ in (III.16) invariant under multiplication from both sides. It follows that we can expand

$$c=\sum_{\pi\in\otimes_{k=1}^{n}\pi_k} c_\pi, \tag{III.19}$$

where the sum is over the irreducibles contained in the decomposition of $\otimes_{k=1}^{n}\pi_k$ into irreducibles and the matrix c_π is projected onto the representation space labeled by π, namely, $\pi(1)c_\pi = c_\pi\pi(1)=c_\pi$. Let us also decompose

$$(\otimes_{k=1}^{n}\pi_k(g))=\sum_{\pi\in\otimes_{k=1}^{n}\pi_k}\pi(g). \tag{III.20}$$

We now plug (III.19) and (III.20) into (III.18) and obtain

$$\sum_{\pi\in\otimes_{k=1}^{n}\pi_k}\pi(g)c_\pi\pi(g)^{-1}=\sum_{\pi\in\otimes_{k=1}^{n}\pi_k}c_\pi, \tag{III.21}$$

which we multiply by $\pi(1)$ to obtain

$$\pi(g)c_\pi\pi(g)^{-1}=c_\pi, \tag{III.22}$$

that is, c_π commutes with the representation and therefore must be proportional to $\pi(1)$ by the lemma of Schur.

(ii) It follows immediately from (i) that the union of all the loop-network states for all the graphs γ is an overcomplete (uncountable) set of states on $L_2(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$ (the graphs label cylindrical functions that are dense in $L_2(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$; compare also Ref. 24). The redundant states are eliminated by the recipe stated in the lemma. In particular, then all the representations involved in $\boldsymbol{\pi}$ are required to be nontrivial (except for the empty graph) since any loop network with trivial representations can be realized already on a smaller graph.

It remains to show that then two loop-network states that are defined on different graphs are orthogonal. But this is trivial because for two graphs $\gamma\neq\gamma'$ there is at least one generator β in which they differ and the representation π associated with that generator is nontrivial. Therefore the inner product between these loop-network states will contain the integral $\int d\mu_{\mathrm{H}}(g)\pi(g)=0$. Therefore we get altogether,

$$\langle T_{\gamma,\boldsymbol{\pi},\pi},T_{\gamma',\boldsymbol{\pi}',\pi'}\rangle=\delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\delta_{\pi,\pi'}. \tag{III.23}$$

Remark: More concretely, the redundant states can be removed by imposing the following constraints on the vector $\boldsymbol{\pi}$: given an edge e of γ (that is, a maximally analytic piece of γ) determine the generators, say $\beta_1,\ldots,\beta_k$, that contain e. If β_i is cloured with the representation π_i then require that the tensor product $\pi_1\otimes\ldots\otimes\pi_k$ does not contain any trivial representation, otherwise the loop-network state would contain a piece defined on a smaller graph.

This restriction leads to the definition of edge-network states (compare the Appendix).

The next thing to do is to define the Fourier transform of a measure on $\overline{\mathcal{A}/\mathcal{G}}$.

Definition III.4: The loop transform (Fourier transform, characteristic functional) of a measure μ on $\overline{\mathcal{A}/\mathcal{G}}$ is defined by

$$\chi_\mu(\gamma,\boldsymbol{\pi},\pi):=\langle\overline{T}_{\gamma,\boldsymbol{\pi},\pi}\rangle:=\int_{\overline{\mathcal{A}/\mathcal{G}}}d\mu(A)\overline{T}_{\mu,\boldsymbol{\pi},\pi}(A). \tag{III.24}$$

This definition differs from the one given in Refs. 5 and 19, however, both definitions are equivalent in the sense that they allow for a reconstruction of μ according to the Riesz–Markov theorem.[27] Namely, the former definition is based on the vacuum expectation value of products of Wilson loop functionals, and according to Refs. 5 and 2, these functions are an overcomplete set of functions on $\overline{\mathcal{A}/\mathcal{G}}$ (that is, they are subject to Mandelstam identities) so that we can reexpress them in terms of loop networks and vice versa.

Now let be given a functional χ on loop networks. Provided it is positive (note that there are no Mandelstam relations between loop-network states anymore and that the product of loop-

network states is a linear combination of loop-network states), we know by the Riesz–Markov theorem that there is a measure μ whose Fourier transform is given by χ. This measure will be known if we know its finite joint distributions that automatically form a self-consistent system of measures whose projective limit (known to exist) gives us back μ. We now compute these joint distributions.

Lemma III.2: If the Fourier transform of a (complex) regular Borel measure μ on a compact gauge group G is in ℓ_1, then it is absolutely continuous with respect to the Haar measure on G.

Proof: Given the Fourier coefficients χ_π^{ij} of the measure μ the Fourier series $\hat{\chi}$ associated with these coefficients is an $L_1(G,d\mu_{\rm H})$ function due to the anticipated ℓ_1 property of the Fourier coefficients. Moreover, the measure $d\hat{\mu}(g):=\hat{\chi}(g)d\mu_{\rm H}(g)$ has the same Fourier transform as μ. Since the functions defined by finite linear combinations of the functions $\sqrt{d_\pi}\pi_{ij}$ form a dense set in $C(G)$, G being a compact Hausdorff group, it follows that both measures define the same bounded linear functional. Now we infer from the uniqueness part of the Riesz–Markov theorem that indeed $\mu=\hat{\mu}$ from which absolute continuity follows. If μ is even a positive measure then $\hat{\chi}$ is positive.

The theorem can obviously extended to any finite number of variables. □

Theorem III.3: *Let χ be a positive linear functional on $C(\overline{\mathcal{A}/\mathcal{G}})$. Then χ is the loop transform of a positive regular Borel measure μ on $\overline{\mathcal{A}/\mathcal{G}}$. If for a given graph γ with n generators the sequence $\{\chi(\gamma,\boldsymbol{\pi},\pi)\sqrt{d_\pi\Pi_{k=1}^n d_{\pi_k}}\}$ is in ℓ_1, then the finite joint distributions of μ_γ are [in the sense of $L_1(\overline{\mathcal{A}/\mathcal{G}}_\gamma,\mu_{0,\gamma})$] given by*

$$\frac{d\mu_\gamma(A)}{d\mu_{0,\gamma}(A)}=\sum_{\boldsymbol{\pi}}\sum_{\pi\in\otimes_{k=1}^n\pi_k}\chi(\gamma,\boldsymbol{\pi},\pi)T_{\gamma,\boldsymbol{\pi},\pi}(A). \tag{III.25}$$

Proof: The proof is a straightforward application of the inverse Fourier transform, Theorem 3.2. The convergence condition on the characteristic functional mentioned in the theorem together with Lemma 3.2 allows us to conclude that on cylindrical subspaces the measure μ_γ is absolutely continuous with respect to the induced Haar measure $\mu_{0,\gamma}$. Therefore there exists a positive $L_1(G^n,d^n\mu_H)$ function $\rho_\gamma(g_1,\ldots,g_n)$ such that it is the Radon–Nikodym derivative of $d\mu_\gamma$ with respect to $d\mu_{0,\gamma}$.[27] Since μ is a gauge-invariant measure, the Fourier coefficients of ρ_γ satisfy

$$(\otimes_{k=1}^n\pi_k(g^{-1}))_{(i_1,k_1)\cdots(i_n,k_n)}\rho_{\gamma,\boldsymbol{\pi}}^{(k_1,l_1)\cdots(k_n,l_n}(\otimes_{k=1}^n\pi_k(g))_{(l_1,j_1)\cdots(l_n,j_n)}=\rho_{\gamma,\boldsymbol{\pi}}^{(i_1,j_1)\cdots(i_n,j_n)}, \tag{III.26}$$

so they lie in the invariant subspace spanned by the matrices $\pi(1)$, where $\pi\in\otimes_{k=1}^n\pi_k$. Thus

$$\begin{aligned}\rho_{\gamma,\boldsymbol{\pi}}^{(i_1,j_1)..(i_n,j_n)}&=\sum_{\pi\in\otimes_{k=1}^n\pi_k}\frac{1}{d_\pi}\,\mathrm{tr}[\rho_{\gamma,\boldsymbol{\pi}}\pi(1)]\pi(1)_{(i_1,j_1),\cdots,(i_n,j_n)}(1)\\&=\sum_{\pi\in\otimes_{k=1}^n\pi_k}\frac{1}{\sqrt{d_\pi}}\,\chi(\gamma,\boldsymbol{\pi},\pi)\pi_{(i_1,j_1),\cdots,(i_n,j_n)}(1).\end{aligned} \tag{III.27}$$

Now

$$\begin{aligned}&\sum_{i_1,j_1=1}^{d_{\pi_1}}\sqrt{d_{\pi_1}}\cdots\sum_{i_n,j_n=1}^{d_{\pi_n}}\sqrt{d_{\pi_n}}|\rho_{\gamma,\boldsymbol{\pi}}^{(i_1,j_1)\cdots(i_n,j_n)}|\\&\quad\leqslant\sum_{\pi\in\otimes_{k=1}^n\pi_k}|\chi(\gamma,\boldsymbol{\pi},\pi)|\sum_{i_1,j_1=1}^{d_{\pi_n}}\cdots\sum_{i_n,j_n=1}^{d_{\pi_n}}|c(\boldsymbol{\pi},\pi)_{(i_1,j_1),\cdots,(i_n,j_n)}(1)|\\&\quad=\sum_{\pi\in\otimes_{k=1}^n\pi_k}|\chi(\gamma,\boldsymbol{\pi},\pi)|\sqrt{d_\pi\prod_{k=1}^n d_{\pi_k}}.\end{aligned} \tag{III.28}$$

Therefore the convergence condition on χ mentioned in the theorem implies that $\hat{\rho}_\gamma \in \ell_1$ and the theorem on the inverse Fourier transform tells us that in the sense of L_1,

$$\rho_\gamma(g_1,\ldots,g_n)=\sum_{\vec{\pi}} \sum_{i_1,j_1=1}^{d_{\pi_1}} \sqrt{d_{\pi_1}}\cdots \sum_{i_n,j_n=1}^{d_{\pi_n}} \sqrt{d_{\pi_n}}\rho_{\gamma,\vec{\pi}}^{(i_1,j_1)\cdots(i_n,j_n)}(\otimes_{k=1}^n \pi_k)_{(i_1,j_1)\cdots(i_n,j_n)}(g_1,\ldots,g_n)$$

$$=\sum_{\vec{\pi}} \sum_{\pi\in\otimes_{k=1}^n \pi_k} \chi(\gamma,\vec{\pi},\pi)T_{\gamma,\vec{\pi},\pi}, \tag{III.29}$$

where we used (III.27) and the definition of a loop network. This furnishes the proof. □

IV. EXAMPLES OF INVERSE LOOP TRANSFORMS

In order to determine whether a given function χ from (singular) knots into the complex numbers arises as the loop transform of a measure one has to check two things.

(1) All the identities that are satisfied by products of traces of holonomies of loops have to be satisfied (Mandelstam identities; see Ref. 2). Alternatively, it has to be true that χ can be written purely in terms of loop-network states.

(2) It is a positive linear functional on any cylindrical subspace of $C(\overline{\mathcal{A}/\mathcal{G}})$.

(a) An example of a (singular) knot function that satisfies these criteria is of course the Fourier transform of any σ-additive measure on $\overline{\mathcal{A}/\mathcal{G}}$. Let us look at the Fourier transform of the measure μ_0 that is even diffeomorphism invariant so that χ is a singular knot invariant:

$$\chi_{\mu_0}(\gamma,\vec{\pi},\pi)=\delta_{\vec{\pi},\vec{0}}\delta_{\pi,0}, \tag{IV.1}$$

where 0 denotes the trivial representation. In other words, χ is nonvanishing only on the trivial loop network 1. Therefore we find for the finite joint distribution precisely $\rho_\gamma(g_1,\ldots,g_n)=1$ according to (III.25).

(b) A second example of a singular knot invariant is given as follows.

Let K be a regular knot invariant and let $\mathrm{orb}(\alpha_0)$ be the orbit of the regular knot α_0 under the diffeomorphism group under question. Then the formal expression

$$d\mu(A):=\sum_{\alpha\in \mathrm{orb}(\alpha_0)} \overline{T_\alpha}(A)\,d\mu_0(A) \tag{IV.2}$$

is a diffeomorphism invariant measure on $\overline{\mathcal{A}/\mathcal{G}}$, where $T_\alpha=\mathrm{tr}h_\alpha$, namely its Fourier transform,

$$\chi_\mu(\gamma,\vec{\pi},\pi)=\chi_{\mathrm{orb}(\alpha_0)}(\gamma)\,\delta_{\vec{\pi},\mathrm{def}}\delta_{\pi,\mathrm{def}} \tag{IV.3}$$

(χ_S means the characteristic function of a set S) is a bounded linear functional on $C(\overline{\mathcal{A}/\mathcal{G}})$ and thus by the extension of the Riesz–Markov theorem[27] we know that it corresponds to a unique complex regular Borel measure on $\overline{\mathcal{A}/\mathcal{G}}$ that is rigorously defined. So we get a new singular knot invariant from a regular one!

(c) Now we provide an example of a Fourier transform for a nondiffeomorphism invariant measure.

In two Euclidean space–time dimensions one can choose the generators of a graph to be simple, meaning that they do not have self-intersections, and nonoverlapping, meaning that the intersection of the surfaces that any two of them enclose have zero Euclidean area.[21] Then the characteristic functional (in the continuum) for pure Yang–Mills theory on the Euclidean plane is given by

$$\chi_{\mu_{\mathrm{YM}}^{(2)}}(\gamma,\vec{\pi},\pi)=\exp\left(-\frac{g_0^2}{2}\sum_{k=1}^n \lambda_{\pi_k}\mathrm{Ar}(\beta_k)\right)\sqrt{d_\pi\prod_{k=1}^n d_{\pi_k}}, \tag{IV.4}$$

where g_0 is the bare coupling constant, $-\lambda_\sigma$ is the eigenvalue of the Casimir operator on the matrix element functions $\sigma_{ij}(g)$ for an irreducible representation σ, and $\mathrm{Ar}(\alpha)$ is the area of the surface enclosed by a simple loop α.

Let us compute, for instance, the one-dimensional joint distributions. We find from (III.25) (if $\boldsymbol{\pi}$ is one dimensional then of course also $\pi=\boldsymbol{\pi}$ is the only possible choice),

$$\rho_{\gamma=\beta}(g)=\sum_{\pi} d_\pi \exp\left(-\frac{g_0^2}{2}\mathrm{Ar}(\beta)\lambda_\pi\right)\chi_\pi(g), \tag{IV.5}$$

therefore $\rho_{\gamma=\beta}(g)=\rho_{g_0^2\mathrm{Ar}(\beta)}(g)$, where, as before, ρ_t is the heat kernel on G.

(d) In the literature it has been argued[29] that the exponential of the Chern–Simons action S is a (formal) solution to all constraints of canonical quantum gravity in four space–time dimensions with nonzero cosmological constant λ in the connection representation. Rigorously speaking, this could only be true for Euclidean gravity: for Lorentzian gravity it could at most be a rigorous solution in the complex Ashtekar representation, however, no Hilbert space has so far been defined for this theory. We will therefore understand in the sequel that this claim was made for the Euclidean signature only for which the gauge group is SU(2).

However, even for the Euclidean theory there is an immediate problem: the Chern–Simons action is not invariant under all, possibly nonsmooth (as required on $\overline{\mathcal{A}/\mathcal{G}}$) gauge transformations, and so it cannot even be written as a measurable function on $\overline{\mathcal{A}/\mathcal{G}}$. Thus, it is not an element of the Hilbert space and, even worse, not even an element of $L_1(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$ and so one cannot compute its Fourier transform with respect to $d\mu_0$. On the other hand, what one has tried to do in the literature is the following (see Ref. 30 and references therein): If one pretends that there is an infinite-dimensional Lebesgue measure ν on (some extension of) $\mathcal{A}/\mathcal{G}$ then one can just compute the integral of $e^{iS/\lambda}$ with respect to the ''measure'' $d\nu\,\mathrm{tr}(h_\alpha(A))$ and get an answer very much in the same way as perturbative calculations in quantum field theory are performed. The result is a formal series expansion of the form $e^{w(\alpha)}\sum_{n=0}^{\infty}\lambda^{-n}a_n(\alpha)$. Here, w is depends on the writhe of the loop and captures the framing dependence of the integral computed (the Chern–Simons propagator in three dimensions is too singular on intersecting loops and one introduces a regulator in the form of a framing of the loop), but the coefficients $a_n(\alpha)$ are truly diffeomorphism invariant, they are singular knot invariants, and the lowest orders were verified to be solutions to a Hamiltonian constraint in the loop representation in various contexts (see, e.g., Ref. 30 and references therein). For nonintersecting and singly retraced loops they are related to the Jones polynomial. It is therefore meaningful to ask whether the $a_n(\alpha)$ qualify as the loop transform of a measure on $\overline{\mathcal{A}/\mathcal{G}}$. The problem is that one would have to check an infinite number of Mandelstam identities that has not been done to the present and, moreover, one needs to extend the definition of a regular knot invariant to arbitrarily singular knots before a decision can be made.

(e) In Ref. 31, which was published later than Ref. 17 or the present paper on the net, the authors show that if one defines an inner product in the loop representation by requiring that spin-network states (compare the Appendix) $\tilde{T}_{\gamma,\boldsymbol{\pi},\pi}$ form an orthonormal basis then one gets the same inner product as in the connection representation. This result is, from our perspective of course, a triviality: in the loop representation a state $\tilde{f}$ is characterized completely by evaluating it on a spin-network γ, $\boldsymbol{\pi}$, π. A spin-network state in the loop representation $\tilde{T}_{\gamma,\boldsymbol{\pi},\pi}$ is then defined by $\tilde{T}_{\gamma,\boldsymbol{\pi},\pi}(\gamma',\boldsymbol{\pi}',\pi')=\delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\delta_{\pi,\pi'}$. However, this is precisely the inner product between the loop-network states $T_{\gamma,\boldsymbol{\pi},\pi}, T_{\gamma',\boldsymbol{\pi}',\pi'}$ in the connection representation, which, in turn, is the loop transform of the state $T_{\gamma,\boldsymbol{\pi},\pi}$. That therefore the $\tilde{T}_{\gamma,\boldsymbol{\pi},\pi}$ form an orthonormal basis in the loop representation is the statement that the loop transform (Fourier transform) is a unitary map between L_2 and l_2, which, since μ_0 is effectively a product of Haar measures on cylindrical subspaces, is merely the statement of the Peter and Weyl theorem, and so is nothing new.

ACKNOWLEDGMENTS

This research project was partially supported by National Science Foundation Grant No. PHY93-96246, the Eberly research fund of the Pennsylvania State University, and Department of Energy Grant No. DE-FG02-94ER25228 to Harvard University.

APPENDIX: EDGE NETWORKS

In this appendix we show how the developments of this paper can be written in terms of edges that are more convenient to deal with if
(a) one wants to write down a complete orthonormal basis[23,24] of $L_2(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$ without having to make use of the recipe mentioned in (ii) of Lemma 3.1, and
(b) if one is interested in applications to quantum gravity, in particular, if one is to obtain the spectrum of certain area and volume operators.[32]

Again we consider an unoriented graph γ and fix an orientation for each of its edges once and for all. Note that we do not have to make a choice of edges in this case because they are defined to be the maximally analytic pieces of the given graph. Denote by E_γ and V_γ the set of edges and vertices of γ, respectively, $n_E(\gamma)$ and $n_V(\gamma)$ are the number of these edges and vertices, respectively, and, in general, we will denote edges by the symbol e and vertices by the symbol v.

Definition A1: (i) An edge network is a triple $(\gamma,\boldsymbol{\pi},\boldsymbol{\sigma})$ *consisting of a graph* γ *an edge vector of irreducible nontrivial representations* $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_{n_E(\gamma)})$ *and a vertex vector of irreducible trivial representations* $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_{n_V(\gamma)})$. *The irreducible representation* σ_v *takes values in the set of trivial irreducible representations that are contained in the decomposition into irreducibles of* $\otimes_{e^-=v}\overline{\pi_e}\otimes_{e^+=v}\pi_e$, $e^\pm$ *being the starting or ending point of* e *(these representations are automatically orthogonal to each other). We assume that* $\boldsymbol{\pi}$ *is such that the space of possible* $\boldsymbol{\sigma}$ *is nonempty.*

(ii) An edge network state is a function from $\overline{\mathcal{A}/\mathcal{G}}$ *into the complex numbers defined by*

$$T_{\gamma,\boldsymbol{\pi},\boldsymbol{\sigma}}(A):=\mathrm{tr}[\otimes_{e\in E_\gamma}\pi_e(h_e(A))\cdot c(\gamma,\boldsymbol{\pi},\boldsymbol{\sigma})], \tag{A1}$$

where the matrix c *is defined as follows.*

There exist permutation matrices $P_\gamma^\pm$ such that

$$(P_\gamma^\pm)^{-1}\cdot\otimes_{e\in E_\gamma}\pi_e(h_e)\cdot P_\gamma^\pm=\otimes_{v\in V_\gamma}\otimes_{e^\pm=v}\pi_e(h_e); \tag{A2}$$

then

$$c(\gamma,\boldsymbol{\pi},\boldsymbol{\sigma}):=\sqrt{\prod_{e\in E_\gamma}d_{\pi_e}}P_\gamma^-\cdot\otimes_{v\in V_\gamma}c_v(\boldsymbol{\sigma})\cdot(P_\gamma^+)^{-1}, \tag{A3}$$

where the vertex contractor is given by

$$c_v(\boldsymbol{\sigma})_{ij}:=\frac{\sigma_v(1)(i_0j_0),(ij)}{\sqrt{\sigma_v(1)_{(i_0j_0),(i_0j_0)}}}, \tag{A4}$$

and where (i_0j_0) *is an arbitrary but fixed choice of index pairs such that the denominator of (A4) is nonvanishing (since* σ_v *is a trivial representation, any choice of* i_0,j_0 *leads to the same vector* c_v *up to a multiple constant). Here the index structure comes from* $[\otimes_{e^-=v}\overline{\pi_e}]\otimes[\otimes_{e^+=v}\pi_e]_{(ij),(i_0j_0)}$.

The composition of loops can result in a loop that is defined already on a smaller graph. This is the source of the redundancy mentioned in Lemma 3.1. Something similar cannot happen with edges whence there is no redundancy in the definition of edge networks.

In the special case of $G=\mathrm{SU}(2)$ the notion of edge networks coincides with the notion of spin networks.[23,24] In this case the vertex contractors can easily be seen to be the usual Clebsh–Gordan coefficients $c_v=\langle 0,\{(j_v,l_v)\}\rangle$ for the addition of all the angular momenta corresponding to the irreducible representations with which the edges starting and ending at v are colored. Note that our analytical expression (A1) for an edge-network state does not need any graphical visualization and no preferred role is played by trivalent graphs.[23]

Theorem A1: *The set of all edge networks provides for a complete orthonormal basis of* $L_2(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$.

Proof: (1) Orthonormality: Note that $P_\gamma^T=P_\gamma^\dagger=P_\gamma^{-1}$ for any permutation matrix P_γ, which merely reshuffles the order of the factors in the tensor product $\otimes_{e\in E_\gamma}\pi_e$. Therefore

$$\begin{aligned}\langle T_{\gamma,\boldsymbol{\pi},\boldsymbol{\sigma}}, T_{\gamma',\boldsymbol{\pi}',\boldsymbol{\sigma}'}\rangle &= \delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\ \mathrm{tr}[c(\gamma,\boldsymbol{\pi},\boldsymbol{\sigma})^\dagger c(\gamma,\boldsymbol{\pi},\boldsymbol{\sigma})]\\ &= \delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\ \mathrm{tr}[((P^-_\gamma\cdot\otimes_{v\in V_\gamma}c_v(\boldsymbol{\sigma})\cdot(P^+_\gamma)^{-1})^\dagger\times P^-_\gamma\cdot\otimes_{v\in V_\gamma}c_v(\boldsymbol{\sigma}')\cdot(P^+_\gamma)^{-1})]\\ &= \delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\ \mathrm{tr}[\otimes_{v\in V_\gamma}(c_v(\boldsymbol{\sigma}))^\dagger c_v(\boldsymbol{\sigma}')]\\ &= \delta_{\gamma,\gamma'}\delta_{\boldsymbol{\pi},\boldsymbol{\pi}'}\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'},\end{aligned} \tag{A5}$$

we have orthonormality. We have used that two different graphs differ in at least one edge which carries a nontrivial irreducible representation and therefore the integral (A5) with respect to the Haar measure vanishes as well as the reality and symmetry of the projectors $\pi(1)$ for any irreducible representation π.

(2) Completeness: As in (III.16) we start from the observation that the states

$$(T_{\gamma,\boldsymbol{\pi}})_{(i_1j_1)\cdots(i_{n(E_\gamma)}j_{n(E_\gamma)})} = (\otimes_{e\in E_\gamma}\pi_e(h_e))_{(i_1j_1)\cdots(i_{n(E_\gamma)}j_{n(E_\gamma)})} \tag{A6}$$

contain a complete set of states for $L_2(\overline{\mathscr{A}/\mathscr{G}}_\gamma, d\mu_{0,\gamma})$ so that we only need to contract (A6) in all the possible gauge-invariant ways. Making use of (A2), we see that under a gauge transformation g_v at every vertex v, we have

$$\mathrm{tr}[T_{\gamma,\boldsymbol{\pi}}\cdot c]\to\mathrm{tr}[T_{\gamma,\boldsymbol{\pi}}\cdot P^-_\gamma\cdot[\otimes_v\otimes_{e^-=v}\pi_e(g_v^{-1})][(P^-_\gamma)^{-1}\cdot c\cdot P^+_\gamma]\cdot[\otimes_v\otimes_{e^+=v}\pi_e(g_v)]\cdot(P^+_\gamma)^{-1}] \tag{A7}$$

so that gauge invariance requires that

$$[\otimes_v\otimes_{e^-=v}\pi_e(g_v^{-1})]\cdot[(P^{-1}_\gamma)^{-1}\cdot c\cdot P^+_\gamma][\otimes_v\otimes_{e^+=v}\pi_e(g_v)] = [(P^{-1}_\gamma)^{-1}\cdot c\cdot P^+_\gamma]. \tag{A8}$$

Without loss of generality we write $c = P^-_\gamma\cdot\Sigma_{i=1}^n\otimes_v s_v^{(i)}\cdot(P^+_\gamma)^{-1}$ for some n and arrive at

$$\sum_i\otimes_v s_v^{(i)} = \sum_i\otimes_v[\otimes_{e^-=v}\pi_e(g_v^{-1})]\cdot s_v^{(i)}\cdot[\otimes_{e^+=v}\pi_e(g_v)], \tag{A9}$$

which we can write more conveniently as the eigenvalue equation,

$$\sum_i\otimes_v s_v^{(i)} = \left[\sum_i\otimes_v s_v^{(i)}\right]\cdot[\otimes_v\nu_v(g_v)], \tag{A10}$$

where we have abbreviated

$$\nu_v(g_v) = [\otimes_{e^-=v}\overline{\pi_e}\otimes_{e^+=v}\pi_e](g_v). \tag{A11}$$

Since $\nu_v(1)$ is a projector, it follows that the eigenvector on the lhs of (A10) lies in the (reducible) subspace corresponding to $\otimes_v\nu_v$, that is, each $s_v^{(i)}$ lies in the (reducible) subspace corresponding to ν_v.

We now ask for the possible solutions x of the equation $x[\otimes_v\nu_v(g_v)] = x$, $\forall(g_v)_{v\in V_\gamma}\in G^{n(V_\gamma)}$. Setting $g_{v'}=1$ for all $v'\in V_\gamma$ except for one, v, say, shows that x has to project onto one of the trivial representations σ_v contained in ν_v. Therefore x has the general form $\Sigma_{\boldsymbol{\sigma}\in\boldsymbol{\nu}=\text{trivial}}\,k_{\boldsymbol{\sigma}}\otimes_v c_v(\boldsymbol{\sigma})$ for some complex numbers $k_{\boldsymbol{\sigma}}$, which is precisely of the form (A4). Since the space of vectors spanned by the matrix elements of the trivial representation is one-dimensional the choice made in (A4) means no loss of generality. This proves completeness. □

It is straightforward to see how the inverse loop transform theorem can be translated into an ''inverse edge transform'' theorem.

[1] R. Gambini and A. Trias, Phys. Rev. D **22**, 1380 (1980); C. Di Bartolo, F. Nori, R. Gambini, and A. Trias, Lett. Nuovo Cimento **38**, 497 (1983).
[2] R. Giles, Phys. Rev. D **8**, 2160 (1981).
[3] C. Rovelli and L. Smolin, Nucl. Phys. B **331**, 80 (1990).
[4] A. Ashtekar, Phys. Rev. Lett. **57**, 2244 (1986); Phys. Rev. D **36**, 1587 (1987).
[5] A. Ashtekar and C. J. Isham, Class. Quantum Grav. **9**, 1433 (1992).

[6] A. Ashtekar and J. Lewandowski, ''Representation theory of analytic holonomy C^* algebras,'' in *Knots and Quantum Gravity*, edited by J. Baez (Oxford University Press, Oxford, 1994); J. Lewandowski, Int. J. Mod. Phys. D **3**, 207 (1994).
[7] J. Baez, Lett. Math. Phys. **31**, 213 (1994); ''Diffeomorphism invariant generalized measures on the space of connections modulo gauge transformations,'' hep-th/9305045, in *Proceedings of the Conference on Quantum Topology*, edited by D. Yetter (World Scientific, Singapore, 1994).
[8] D. Marolf and J. M. Mourão, Commun. Math. Phys. **170**, 583 (1995); A. Ashtekar, D. Marolf, and J. Mourão, ''Integration on the space of connections modulo gauge transformations'' in *The Proceedings of the Lanczos International Centenary Conference*, edited by J. D. Brown *et al.* (SIAM, Philadelphia, 1994).
[9] J. Lewandowski, ''Differential geometry for the space of connections modulo gauge transformations'', in *The Proceedings of the Lanczos International Centenary Conference*, edited by J. D. Brown *et al.* (SIAM, Philadelphia, 1994); A. Ashtekar and J. Lewandowski, J. Geom. Phys. **17**, 191 (1995).
[10] A. Ashtekar and J. Lewandowski, J. Math. Phys. **36**, 2170 (1995).
[11] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, J. Math. Phys. **36**, 6456 (1995).
[12] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, J. Funct. Anal. **135**, 519 (1996).
[13] T. Thiemann, Class. Quantum Grav. **13**, 1383 (1996).
[14] A. Ashtekar, Phys. Rev. D **53**, R2865 (1996).
[15] T. Thiemann, Phys. Lett. B **380**, 1383 (1996); ''Quantum Spin Dynamics (QSD),'' Harvard University preprint HUTMP-96/B-359, gr-qc/9606089 (to appear in Class. Quantum Grav.).
[16] F. Barbero, Phys. Rev. D **51**, 5507 (1995).
[17] T. Thiemann, Acta Cosmol. **21**, 145 (1995).
[18] J. Glimm and A. Jaffe, *Quantum Physics* (Springer-Verlag, New York, 1987); Y. Yamasaki, *Measures on Infinite Dimensional Spaces* (World Scientific, Singapore, 1985).
[19] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, ''A manifestly gauge invariant approach to quantum theories of gauge fields,'' in *Geometry of Constrained Dynamical Systems*, edited by J. Charap (Cambridge University Press, Cambridge, 1994).
[20] T. Thiemann, Banach Center Publications, 1996, Vol. 39, p. 389.
[21] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, ''SU(N) Yang–Mills Theory in two dimensions: A complete solution,'' J. Math. Phys. (in press).
[22] O. Foster, *Analysis 3* (Vieweg, Braunschweig/Wiesbaden, 1987).
[23] C. Rovelli and L. Smolin, Phys. Rev. D **53**, 5743 (1995).
[24] J. Baez, Adv. Math. **117**, 253 (1996); ''Spin networks in non-perturbative quantum gravity,'' in *The Interface of Knots and Physics*, edited by L. Kauffman (American Mathematical Society, Providence, RI, 1996).
[25] N. J. Vilenkin, *Special Functions and the Theory of Group Representations* (American Mathematical Society, Providence, RI, 1968).
[26] B. C. Hall, J. Funct. Anal. **122**, 103 (1994).
[27] W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1987).
[28] H. Boerner, *Representations of Groups* (North-Holland, Amsterdam, 1970).
[29] H. Kodama, Phys. Rev. D **42**, 2548 (1990).
[30] R. Gambini and J. Pullin, Phys. Rev. D **54**, 5935 (1996); H. Fort, R. Gambini, and J. Pullin, ''Lattice knot theory and quantum gravity in the loop representation,'' gr-qc/9608033; R. Gambini, J. Griego, and J. Pullin, ''Chern–Simons states in spin-network quantum gravity,'' gr-qc/9703042.
[31] R. De Pietri and C. Rovelli, ''Geometry eigenvalues and scalar product from recoupling theory in loop quantum gravity,'' Phys. Rev. D **54**, 2664 (1996).
[32] C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995); **456**, 5743 (1995); A. Ashtekar and J. Lewandowski, Class. Quantum Grav. **14**, A55 (1997); T. Thiemann, ''A length operator for canonical quantum gravity,'' Harvard University preprint HUTMP-96/B-354, gr-qc/9606092.

Erratum: Spectral transform and solitons for the three-wave coupling model with nontrivial boundary conditions [J. Math. Phys. 38, 4138–4150 (1997)]

E. V. Doktorov
B. I. Stepanov Institute of Physics, 70 F. Skaryna Ave., 220072 Minsk, Republic of Belarus

(Received 22 September 1997; accepted for publication 22 September 1997)

[S0022-2488(98)01502-3]

Due to the author's inaccuracy, there are misprints in Eqs. (8), (19), and (25), where $|C|^2$ should be replaced by $\rho|C|^2$. The same should be done in the expression for ν after Eq. (4). The second term in the expression for w after Eq. (7) should read $2\rho\gamma|C|^2$. Equation (26) takes the form $\sigma=0$ which should also be inserted into Eqs. (27) and (28).

 © 1998 American Institute of Physics

Quantization of diffeomorphism-invariant theories with fermions

John C. Baez[a)]
Department of Mathematics, University of California, Riverside, California 92521

Kirill V. Krasnov[b)]
Center for Gravitational Physics and Geometry, Pennsylvania State University, University Park, Pennsylvania 16802
and Bogolyubov Institute for Theoretical Physics, Kiev 143, Ukraine

(Received 16 April 1997; accepted for publication 21 November 1997)

We extend ideas developed for the loop representation of quantum gravity to diffeomorphism-invariant gauge theories coupled to fermions. Let $P \to \Sigma$ be a principal G-bundle over space and let F be a vector bundle associated to P whose fiber is a sum of continuous unitary irreducible representations of the compact connected gauge group G, each representation appearing together with its dual. We consider theories whose classical configuration space is $\mathcal{A}\times\mathcal{F}$, where $\mathcal{A}$ is the space of connections on P and $\mathcal{F}$ is the space of sections of F, regarded as a collection of Grassmann-valued fermionic fields. We construct the ''quantum configuration space'' $\overline{\mathcal{A}}\times\overline{\mathcal{F}}$ as a completion of $\mathcal{A}\times\mathcal{F}$. Using this, we construct a Hilbert space $L^2(\overline{\mathcal{A}}\times\overline{\mathcal{F}})$ for the quantum theory on which all automorphisms of P act as unitary operators, and determine an explicit ''spin network basis'' of the subspace $L^2((\overline{\mathcal{A}}\times\overline{\mathcal{F}})/\overline{\mathcal{G}})$ consisting of gauge-invariant states. We represent observables constructed from holonomies of the connection along paths together with fermionic fields and their conjugate momenta as operators on $L^2((\overline{\mathcal{A}}\times\overline{\mathcal{F}})/\overline{\mathcal{G}})$. We also construct a Hilbert space $\mathcal{H}_{\mathrm{diff}}$ of diffeomorphism-invariant states using the group averaging procedure of Ashtekar, Lewandowski, Marolf, Mourão and Thiemann. © *1998 American Institute of Physics.* [S0022-2488(98)01403-0]

I. INTRODUCTION

In this paper we address the problem of the quantization of a wide class of diffeomorphism-invariant theories. We consider theories whose degrees of freedom are described in the Hamiltonian framework by a connection on some principal G-bundle on space together with a set of fermionic fields on space taking values in representations of the compact connected gauge group G. Theories of this class can be thought of as giving a diffeomorphism-invariant description of fermionic matter interacting by means of gauge fields. One of the main motivations for considering such theories is that general relativity can be reformulated as a diffeomorphism-invariant theory having as canonically conjugate variables a connection on space and a densitized triad field.[1] This means that the gravitational force coupled to Yang-Mills gauge fields and fermionic matter fits into our framework.

Our approach builds upon existing work on the loop representation of quantum gravity and other diffeomorphism-invariant theories of connections.[2] The loop quantization of theories of connections is based on an assumption that Wilson loop functionals of connection become well-defined quantum operators. Similarly, our approach is based on an assumption that certain fermionic analogs of Wilson observables become well-defined operators. An important feature of our scheme is that we do not introduce closed Wilson loops as fundamental observables, as typical in the loop representation of a pure gauge theory. Instead, we use open Wilson paths with fermions or their conjugate momenta at the endpoints, which we call ''fermionic path observables.'' Closed loops arise as secondary quantities.

a)Electronic mail: baez@math.ucr.edu
b)Electronic mail: krasnov@phys.psu.edu

 © 1998 American Institute of Physics

The basic strategy of the paper can be summarized as follows. We introduce a set of path observables on the classical phase space of our theory that forms a closed algebra with respect to Poisson brackets. The main idea then is to construct a ''kinematical'' Hilbert space for the theory and represent our algebra by operators on this Hilbert space. This is not the space of physical states, because the constraints—gauge-invariance, spatial diffeomorphism-invariance, and the Hamiltonian constraint—have not yet been implemented. The quantization procedure we adopt satisfies some natural requirements; for example, gauge-invariant classical quantities become operators that preserve the gauge-invariant subspace of the kinematical Hilbert space, ''real'' quantities become self-adjoint operators, and so on. It turns out, however, that representation of our algebra is, in general, reducible even on the subspace of gauge-invariant states.

An important issue in quantizing diffeomorphism-invariant gauge theories coupled to fermions is that of respecting the classical reality conditions on the quantum level. In field theories on Minkowski spacetime the operator † of Hermitian conjugation of fermionic fields sends configuration fermionic fields into their canonically conjugate momenta. However, in diffeomorphism-invariant theories the † operation is more complicated. Consider, for example, a theory in which the only gauge field is the gravitational field, described by the connection field A_a canonically conjugate to the densitized triad field.[1] The configuration and momentum fermionic fields (that we shall denote here by ξ and $\tilde{\pi}$, respectively) have different density weights, and the action of † operator involves the square root of the determinant of the metric:

$$\tilde{\pi}^{\dagger} = -i\sigma\xi. \tag{1}$$

Here σ is the square root of the determinant of the metric, defined in terms of the densitized triad field.

In the quantum theory, one wants to impose this ''reality condition'' in an appropriate form. However, the following difficulties may arise. In general relativity the triad field is a dynamical variable. In the quantum theory, therefore, the quantity σ constructed from the triad field becomes an operator. One might require that the adjoint of the operator $\hat{\tilde{\pi}}$ corresponding to the fermionic momentum be given by

$$(\hat{\tilde{\pi}})^{\dagger} = -i\hat{\sigma}\hat{\xi}.$$

Let us now consider the commutator

$$\left[\int d^n x \hat{\tilde{\pi}}(x) f(x), \int d^n x \ \mathrm{Tr}(\hat{A}_a(x) g^a(x))\right]$$

of operators $\hat{\tilde{\pi}}$ and $\hat{A}_a$ smeared with test functions. This commutator should vanish, since the corresponding Poisson bracket is zero in the classical theory. However, if one assumes that the connection A_a is real, and applies the † operation to this commutator, one obtains

$$\left[\int d^n x(-i\hat{\sigma})\hat{\xi}(x), \int d^n x \ \mathrm{Tr}(\hat{A}_a(x) g^a(x))\right]$$

which appears to be nonzero, since the operator $\hat{\sigma}$ does not commute with $\hat{A}_a$. This inconsistency, however, may not arise if one does not assumes the connection to be real. In fact, it is known[3] that in the case when fermionic fields are coupled to the self-dual gravity the connection must be complex, and its real and imaginary parts depend on the fermionic fields. Unfortunately, the functional analysis required to quantize theories of complex connections is not fully developed.

The strategy we adopt in this paper for dealing with the issue of reality conditions is as follows. We assume that the theories of interest can be cast into a form in which the connection becomes real by an appropriate canonical transformation, as it happens, for instance, in the case of pure gravity.[4] After this is achieved, no non-trivial reality conditions on the connection should be imposed in the quantum theory. We then impose the fermionic reality conditions by requiring that classical real quantities are represented by self-adjoint operators at the quantum level. One has a satisfactory quantum theory if a sufficient number of physically interesting real observables are

represented by self-adjoint operators. The quantization scheme adopted in this paper respects the classical reality conditions in the sense that a certain set of real observables is represented by self-adjoint operators. Note, however, that in this paper only a rather small set of real observables is treated. For instance, the quantization of the Hamiltonian constraint, which is itself a real functional on the phase space of the theory (in its ''real'' formulation used in Ref. 5), is not considered in this paper.

Note that our way of dealing with the reality conditions (1) does not allow us to prove that the inner product we construct is selected by these reality conditions. Indeed, we treat only a rather small set of real observables, while to prove that the inner product is selected by the classical reality conditions one has to show that a sufficiently large set of real observables become self-adjoint operators.

The organization of this paper is as follows. In Sec. II we start by reviewing the main results of the loop representation approach to quantizing diffeomorphism-invariant theories of connections. In Sec. III we specify the class of theories which are of interest to us here and give some examples of theories belonging to this class. In Sec. IV we describe the path observables which are the basic building blocks of our quantum theory. In Sec. V we describe the classical configuration space and a certain completion of it which we call the ''quantum configuration space.'' We then construct the kinematical Hilbert space as a space of functions on the quantum configuration space. In Sec. VI we construct the Hilbert space of gauge-invariant states, and describe a basis of this space given by ''open spin networks'' with ends labelled by fermionic fields. In Sec. VII we define operators representing the path observables. The issue of diffeomorphism-invariant states is discussed in Sec. VIII. We conclude with a discussion of the results obtained, focusing on the issue of superselection sectors.

Despite its great importance, we do not discuss dynamics in this paper, as the correct treatment of the Hamiltonian constraint is still a matter of controversy in the loop representation of pure gravity, despite recent progress on this front.[5] After this work was completed, papers by Thiemann[6] appeared in which both kinematical and dynamical aspects of diffeomorphism-invariant theories with fermions are treated. Thiemann proposes a different strategy to deal with the fermionic reality conditions (1) on the quantum level. His work also contains a proposal for the quantum Hamiltonian constraint for theories with fermions.

II. QUANTIZATION OF DIFFEOMORPHISM-INVARIANT THEORIES OF CONNECTIONS

This section reviews the main steps of quantization of diffeomorphism-invariant theories of connections. See references 2,7,8 for more details.

In the Hamiltonian formalism, the kinematical phase space of such a theory consists of pairs $(A,\tilde{E})$ satisfying suitable regularity conditions. Here A is a connection on a principal G-bundle P over a manifold Σ which represents ''space,'' and the field $\tilde{E}$ is the canonically conjugate momentum (the tilde means that it is a densitized vector field). For technical reasons we assume that Σ is real-analytic and G is a compact connected Lie group. Actually the assumption that Σ is analytic is unnecessary;[9] much of what we do in this paper generalizes to the smooth case, but analyticity makes things a bit simpler.

The classical configuration space of our theory is the space of smooth connections on P, which we denote as $\mathscr{A}$. In the case of canonical quantization of a theory with a finite number of degrees of freedom, one normally represents quantum states by square-integrable functions on the classical configuration space. The Hilbert space of states is then constructed as the space of such functions with the inner product being defined by the usual integral over the configuration space, $\langle\psi|\phi\rangle=\int\bar{\psi}\phi d\mu$, where $d\mu$ is some measure on the configuration space. However, in the case of theories with an infinite number of degrees of freedom, which is of interest to us here, there often does not exist an appropriate measure on the classical configuration space. For example, there is no ''Lebesgue measure'' on an infinite-dimensional vector space. However, one can often complete the classical configuration space in some topology and construct a suitable measure on the resulting ''quantum configuration space.'' In field theories, this larger space usually includes distributional fields in addition to smooth fields. Also, it is common in field theories for the classical configuration space to have zero measure with respect to the measure on the quantum configuration space.

The quantum configuration space depends on a choice of functions on the classical configuration space that play the role of distinguished observables. In our case we work with (smooth) ''cylinder functions'' on $\mathscr{A}$, that is, functions depending smoothly on the holonomy of the connection A along finitely many analytic paths. In other words, Ψ is a cylinder function if it is of the form

$$\Psi(A)=\psi\left(\mathscr{P}\exp\int_{e_1}A,\ldots,\mathscr{P}\exp\int_{e_n}A\right)$$

for some analytic paths $e_1,\ldots,e_n$ in Σ and some smooth function ψ on G^n. The most well-known functions of this form are the Wilson loops, introduced as observables for quantum gravity by Rovelli and Smolin[10] in their original paper on the loop representation. Unlike the Wilson loops, the above cylinder functions are not necessarily gauge-invariant. This allows them to serve as a complete set of functions on $\mathscr{A}$.

One can complete the algebra of cylinder functions with respect to the sup norm to obtain a commutative C*-algebra Fun($\mathscr{A}$) for which the $*$-operation is just pointwise complex conjugation. The Gelfand-Naimark theorem then tells us that this C*-algebra is isomorphic to the algebra of all continuous functions on its spectrum. We take the spectrum of Fun($\mathscr{A}$) to be the quantum configuration space of our theory. We denote this space by $\overline{\mathscr{A}}$, because it contains the classical configuration space $\mathscr{A}$ as a dense subset.

Elements of the quantum configuration space are called ''generalized connections,'' and can be described as follows. First, define a ''transporter'' from the point p to the point q to be a map from P_p to P_q that commutes with the right action of G on the bundle P. If we trivialize the bundle over p and q, we can think of such a transporter simply as an element of G. A ''generalized connection'' A is a map assigning to each oriented analytic path e in Σ a parallel transporter $A_e:P_p\to P_q$, where p is the initial point of the path e and q is the final point. We require that A satisfy certain obvious consistency conditions: A should assign the same transporter to two paths that differ only by an orientation-preserving reparametrization, it should assign to the inverse of any path the inverse transporter, and it should assign to the composite of two paths the composite transporter.

An ordinary smooth connection A gives a generalized connection where the parallel transporter A_e along any path e is simply the holonomy of A along this path, so $\mathscr{A}\subseteq\overline{\mathscr{A}}$. Also, any cylinder function extends to a continuous function on $\overline{\mathscr{A}}$ by setting

$$\Psi(A)=\psi(A_{e_1},\ldots,A_{e_n})$$

for any generalized connection A. This lets us think of cylinder functions as functions on the quantum configuration space.

There is a natural measure μ_0 on $\overline{\mathscr{A}}$, which comes from Haar measure on the group G. One can show that any reasonable measure on $\overline{\mathscr{A}}$ is determined by the values of the integrals of all cylinder functions. Thus, to define the measure μ_0 we specify the values of the integrals $\int\Psi\,d\mu_0$ for all cylinder functions Ψ given as above. Such an integral is defined as the integral over the corresponding copies of the group G:

$$\int\Psi\,d\mu_0=\int\psi(g_1,\ldots,g_n)\,dg_1\cdots dg_n,$$

where dg denotes the normalized Haar measure on G. The measure μ_0 has three important properties. First, it is gauge-invariant. More precisely, note that any gauge transformation g acts on $A\in\overline{\mathscr{A}}$ to give a generalized connection A' with $A'_e=g(q)^{-1}\circ A_e\circ g(p)$, where p,q are the final and the initial point of the path e, respectively. This gives an action of the group $\mathscr{G}$ of gauge transformations of P on the space $\overline{\mathscr{A}}$, and this action preserves the measure μ_0. Second, μ_0 is diffeomorphism-invariant. More precisely, μ_0 is invariant under all automorphisms of the bundle P, not necessarily acting as the identity on the base space Σ. Third, μ_0 is strictly positive, meaning that $\int\Psi\,d\mu_0>0$ for any non-negative integrable function on $\overline{\mathscr{A}}$ except the function 0.

We define the kinematical Hilbert space $L^2(\overline{\mathscr{A}})$ to be the space of functions on $\overline{\mathscr{A}}$ that are square-integrable with respect to the measure $d\tilde{\mu}_0$. This space is not the physical state space, since it contains states that are not invariant under the ''gauge'' symmetries of our theory, that is, gauge transformations and diffeomorphisms of spacetime. One way to try to find physical states is to look for solutions of quantum constraints in $L^2(\overline{\mathscr{A}})$. In general, the solutions may not live in $L^2(\overline{\mathscr{A}})$, but in some completion thereof, but for the Gauss law constraint this problem does not occur: there is a large subspace of $L^2(\overline{\mathscr{A}})$, the ''gauge-invariant Hilbert space,'' consisting of gauge-invariant square-integrable functions on $\overline{\mathscr{A}}$. Alternatively, since the measure μ_0 is gauge-invariant, it gives rise to a well-defined measure on the space $\overline{\mathscr{A}}/\mathscr{G}$ of generalized connections modulo gauge transformations. Then we may equivalently define the gauge-invariant Hilbert space to be the space $L^2(\overline{\mathscr{A}}/\mathscr{G})$ of square-integrable functions on $\overline{\mathscr{A}}/\mathscr{G}$.

We can construct an explicit basis of $L^2(\overline{\mathscr{A}}/\mathscr{G})$ using ''spin networks.''[8,11] A spin network is a triple (γ,ρ,ι) consisting of:

(1) a graph γ analytically embedded in Σ,
(2) a labelling ρ of each edge e of γ with an irreducible representation ρ_e of G,
(3) a labelling ι of each vertex v of γ with an intertwining operator ι_v.
Here the edges of the graph γ are assumed to be unparametrized but oriented, and ι_v is an interwining operator from the tensor product of the representations corresponding to the incoming edges at the vertex v to the tensor product of the representations labelling the outgoing edges. (For more details see references 7,8. Here we allow graphs with isolated vertices, i.e., vertices that are not vertices of any edge. This is unimportant now but will be important when we come to fermions.) Without loss of generality, spin networks Γ are always assumed to satisfy a fourth non-degeneracy condition:
(4) All representations ρ_e are non-trivial and γ is a ''minimal'' graph, in the sense that it cannot be obtained from another graph γ' by subdividing edges of γ'.

The ''spin network state'' Ψ_Γ is a gauge-invariant cylinder function on $\overline{\mathscr{A}}$ constructed from the spin network Γ as follows:

$$\Psi_\Gamma(A)=[\otimes_e \rho_e(A_e)]\cdot[\otimes_v \iota_v],$$

where ''$\cdot$'' stands for contracting, at each vertex v of γ, the upper indices of the matrices corresponding to the incoming edges, the lower indices of the matrices assigned to the outgoing edges, and the corresponding indices of the intertwining operator ι_v. Being gauge invariant, such states lie in $L^2(\overline{\mathscr{A}}/\mathscr{G})$. We obtain an orthonormal basis of $L^2(\overline{\mathscr{A}}/\mathscr{G})$ if we use spin network states corresponding to all possible choices of γ and ρ and a choice of an orthonormal basis of intertwining operators ι_v for each vertex v and each choice of representations labeling incident edges.

The states in $L^2(\overline{\mathscr{A}}/\mathscr{G})$ are still not invariant under the action of diffeomorphisms of Σ. There is a natural unitary representation of Diff(Σ) on $L^2(\overline{\mathscr{A}}/\mathscr{G})$. For instance, the action of the operator $U(\phi)$ corresponding to a diffeomorphism ϕ of Σ on any spin network state is given by

$$U(\phi)\Psi_{(\gamma,\rho,\iota)}=\Psi_{(\phi\gamma,\phi\rho,\phi\iota)}.$$

Here $\phi\gamma$ is the image of the graph γ under the diffeomorphism ϕ, and $\phi\rho,\phi\iota$ are the corresponding representations and intertwining operators associated with the new graph $\phi\gamma$.

To find diffeomorphism-invariant states one has to impose the quantum diffeomorphism constraint. There are very few diffeomorphism-invariant states in $L^2(\overline{\mathscr{A}}/\mathscr{G})$, so here we have to follow the strategy of looking for solutions in some larger space. One often needs to do this when solving quantized constraint equations, and the larger space is usually chosen to be a space of linear functionals on some dense subspace of the initial Hilbert space. That is, we should choose some dense subspace $\mathscr{C}$ of the gauge-invariant Hilbert space $L^2(\overline{\mathscr{A}}/\mathscr{G})$, equip it with some topology in which it is complete, and look for solutions in the topological dual $\mathscr{C}^*$. Note that $\mathscr{C}\subset L^2(\overline{\mathscr{A}}/\mathscr{G})\subset\mathscr{C}^*$.

Since there is a simple geometrical action of diffeomorphisms on cylinder functions, it is natural to choose the subspace of gauge-invariant cylinder functions as $\mathscr{C}$. This has a natural topology in which it is complete, the inductive limit topology. To solve the diffeomorphism constraint, we then seek diffeomorphism-invariant vectors in $\mathscr{C}^*$.

We obtain these by averaging gauge-invariant cylinder functions over the action of the diffeomorphism group, following the procedure of Ashtekar, Lewan-dowski, Marolf, Mourão and Thiemann.[2] It is easiest to average a special sort of gauge-invariant cylinder function, namely, a spin network state Ψ_Γ, where $\Gamma=(\gamma,\rho,\iota)$ is some spin network. For technical reasons we assume γ is ''type I,'' i.e., that for every edge of γ there is an analytic function on Σ that vanishes only on the maximal analytic curve extending that edge. Let $S(\gamma)$ be any set of diffeomorphisms of Σ with the following property: For any graph γ' which equals $\phi\gamma$ for some $\phi\in\mathrm{Diff}(\Sigma)$, there is a unique diffeomorphism $\phi\in S(\gamma)$ with $\gamma'=\phi\gamma$. Also let $GS(\gamma)$ be the group of ''graph symmetries'' of γ, that is, the group $\mathrm{Iso}(\gamma)/\mathrm{TA}(\gamma)$, where $\mathrm{Iso}(\gamma)$ is the group of diffeomorphisms mapping γ to itself, and $\mathrm{TA}(\gamma)$ is the subgroup fixing each edge of γ. We may define an element $\overline{\Psi}_\Gamma\in\mathscr{C}^*$ by

$$\overline{\Psi}_\Gamma(\Phi)=\sum_{\phi_1\in S(\gamma)}\ \sum_{[\phi_2]\in GS(\gamma)}\langle\Psi_{(\phi_1\circ\phi_2)\Gamma}|\Phi\rangle,$$

where $\Phi\in\mathscr{C}$, and where we choose one representative ϕ_2 for each equivalence class $[\phi_2]\in GS(\gamma)$. It is easy to check that $\overline{\Psi}_\Gamma$ is diffeomorphism-invariant. More generally, suppose the cylinder function Ψ is a (possibly infinite) linear combination of such spin network states $\Sigma_\Gamma a_\Gamma\Psi_\Gamma$. Then we define $\overline{\Psi}\in\mathscr{C}^*$ by

$$\overline{\Psi}(\Phi)=\sum_\Gamma\ a_\Gamma\overline{\Psi}_\Gamma(\Phi).$$

It is easy to see that the sum converges and defines a diffeomorphism-invariant element $\overline{\Psi}\in\mathscr{C}^*$.

We may define the inner product of diffeomorphism-invariant vectors of this form by

$$\langle\overline{\Psi}|\overline{\Phi}\rangle=\overline{\Psi}(\Phi).$$

Completing the space of such vectors in this inner product we thus obtain a Hilbert space, the ''diffeomorphism-invariant Hilbert space'' $\mathscr{H}_{\mathrm{diff}}$. The diffeo-morphism-invariant spin network states $\overline{\Psi}_\Gamma$ form an orthogonal (but not orthonormal) basis of $\mathscr{H}_{\mathrm{diff}}$ as we let Γ range over a set of spin networks containing one from each diffeomorphism equivalence class (but restricting ourselves to those whose graph is type I). For any spin network Γ, the diffeomorphism-invariant state $\overline{\Psi}_\Gamma$ takes a zero value on spin networks that do not belong to the same equivalence class, and a nonzero value on spin networks from the same equivalence class.

In short, the main steps of the quantization procedure include: (i) construction of the kinematical Hilbert space of states $L^2(\overline{\mathscr{A}})$, which requires the specification of a quantum configuration space together with an appropriate measure on this space; (ii) construction of the gauge-invariant Hilbert space $L^2(\overline{\mathscr{A}}/\mathscr{G})$ (and, conveniently, an explicit basis for this space), and (iii) construction of the diffeomorphism-invariant Hilbert space $\mathscr{H}_{\mathrm{diff}}$. In what follows we generalize all these steps to the case of theories involving fermions. Finally, there is the problem of imposing the Hamiltonian constraint and the problem of finding solutions to this constraint. In this paper we do not enter into this all-important problem.

In the next section we describe the class of theories which are considered in this paper and give several examples of physically interesting theories belonging to this class. Their quantization will be discussed in the following sections.

III. THE CLASS OF THEORIES

In this paper we consider a special class of theories with fermionic degrees of freedom. Each theory belonging to this class is specified, first, by (i) a principal G-bundle $P\to\Sigma$, where the

gauge group G is a compact connected Lie group and Σ is a real-analytic manifold. We also need to specify (ii) a finite list I of irreducible continuous unitary representations of G. Each representation in I corresponds to an elementary fermion appearing in the theory. Using any representation $\rho \in I$ we may associate to P a vector bundle $P \times_G \rho$ over Σ. It will be convenient to lump these together by forming a single vector bundle

$$F = P \times_G (\oplus_{\rho \in I} \rho),$$

sections of which simultaneously describe all the fermionic fields in the theory.

From this data we can construct the classical configuration space and phase space of the theory. We have already described the classical configuration space for the gauge fields; this is the space $\overline{\mathscr{A}}$ of smooth connections on P. The corresponding classical phase space is the space of pairs $(A, \widetilde{E})$, where A is a smooth connection on P and $\widetilde{E}$ is a smooth Ad P-valued vector field of density weight 1. We denote this classical phase space as $T^*\mathscr{A}$, since we can think of $\widetilde{E}$ as a cotangent vector to $\mathscr{A}$ satisfying certain smoothness conditions. Both $\mathscr{A}$ and $T^*\mathscr{A}$ become infinite-dimensional smooth manifolds in a natural way.

Similarly, the classical configuration space for the fermionic fields is the space $\mathscr{F}$ of smooth sections of F, and the classical phase space for the fermionic fields is space of pairs consisting of a smooth section of F together with a densitized smooth section of F^* with density weight 1. We denote this classical phase space as $T^*\mathscr{F}$. Both $\mathscr{F}$ and $T^*\mathscr{F}$ become infinite-dimensional topological vector spaces in a natural way.

The classical configuration space for the whole theory is thus the product $\mathscr{A} \times \mathscr{F}$, while the classical phase space is $T^*\mathscr{A} \times T^*\mathscr{F}$. In down-to-earth terms, a point in the classical phase space is simply a list of pairs $(A, \widetilde{E}), (\xi, \widetilde{\pi}), \ldots, (\eta, \widetilde{\omega})$. Here A is the connection on P and $\widetilde{E}$ is its canonically conjugate momentum. Similarly, the fields $\xi, \ldots, \eta$ are fermionic configuration fields corresponding to the representations in I, while the fields $\widetilde{\pi}, \ldots, \omega$ are their canonically conjugate momenta.

In what follows we shall treat the fermionic fields as Grassmann-valued, in order to guarantee that the Pauli principle holds for our fermions. This amounts to treating $\mathscr{F}$ as a supermanifold with all odd coordinates.[12] Thus we define the algebra $C^\infty(\mathscr{F})$ of ''smooth functions'' on $\mathscr{F}$ to be the exterior algebra $\Lambda \mathscr{F}^*$ generated by the continuous linear functionals on $\mathscr{F}$. Note that these are not functions on $\mathscr{F}$ in the standard sense, but only in the sense of supermanifold theory. Similarly, we define the algebra $C^\infty(T^*\mathscr{F})$ of ''smooth functions'' on the fermionic classical phase space $T^*\mathscr{F}$ to be the exterior algebra generated by the continuous linear functions on $T^*\mathscr{F}$. We also define ''smooth functions'' on the configuration space and phase space of the whole theory as follows:

$$C^\infty(\mathscr{A} \times \mathscr{F}) = C^\infty(\mathscr{A}) \otimes C^\infty(\mathscr{F}),$$

$$C^\infty(T^*\mathscr{A} \times T^*\mathscr{F}) = C^\infty(T^*\mathscr{A}) \otimes C^\infty(T^*\mathscr{F}).$$

The latter algebra is the algebra of classical observables of the theory.

Theories may have the same phase space and differ only in the form of the Hamiltonian, so a theory is finally determined by (iii) the Hamiltonian. Since we shall not actually treat dynamics in this paper, we shall not be very precise about the class of allowed Hamiltonians, but we are interested in those for which the action can be written as follows:

$$S = \int dt \int d^3x (\mathrm{Tr}\ \widetilde{E}^a \mathscr{L}_t A_a + \widetilde{\pi} \mathscr{L}_t \xi + \widetilde{\omega} \mathscr{L}_t \eta + \cdots + \underset{\sim}{N} \tilde{\tilde{H}} + N^a \widetilde{H}_a + \mathrm{Tr}\ \mathbf{N} \widetilde{G}. \tag{2}$$

Here $\mathscr{L}_t$ stands for the time derivative and the quantities $\underset{\sim}{N}$, N^a, and $\mathbf{N}$ are Lagrange multipliers. The quantities $\tilde{\tilde{H}}$, $\widetilde{H}_a$, and $\widetilde{G}$ are functionals on the phase space corresponding to the Hamiltonian, diffeomorphism and Gauss law constraints. Here tildes over the momentum fields keep track of density weights: a single tilde over a symbol stands for a density of weight 1, a double tilde over a symbol stands for a density of weight 2, and a single tilde under a symbol denotes a density of weight -1.

Once the canonical variables are chosen (or in other words, once the phase space is chosen), a theory is determined solely by the Hamiltonian. The other two constraint functionals are determined by the requirement that they generate diffeomorphisms and gauge transformations on the phase space of the theory.

Let us give several examples.

a) An example of such a theory is Husain-Kuchař model with fermions.[13] In its simplest version it describes a massless SU(2) fermionic field interacting with the gravitational field, with the 4-metric being degenerate in a very special way.[14] In the canonical formulation the Lagrangian of this model is of the form (2), *without* the Hamiltonian constraint term. Thus, the only constraints in the theory are Gauss and diffeomorphism constraints. All fields in the model are real, so there are no reality conditions to deal with. Thus the quantization scheme developed in the present paper gives a complete quantization of this model.

b) Another example was considered in Ref. 15. It describes a massless fermionic field interacting with a gravitational field in four-dimensional (Riemannian) spacetime, with the gravitational field being described using a chiral spin connection. The gauge group is SU(2), and the gravitational degees of freedom are described by the SU(2) connection A_a and its canonically conjugate momentum field $\tilde{E}^a$. The only fermionic field of the theory, ξ, takes values in the spin-1/2 representation. The Hamiltonian constraint of the theory consists of two parts:

$$\tilde{\tilde{H}} = \tfrac{1}{2}\,\mathrm{Tr}(\tilde{E}^a \tilde{E}^b F_{ab}) + \tilde{E}^{aB}_A \mathscr{D}_a \xi_B \tilde{\pi}^A,$$

where the first part is the Hamiltonian of the gauge field and the second part is responsible for the dynamics of the fermionic field. Here F_{ab} is the curvature of the connection field A_a, and $\mathscr{D}_a = \partial_a + A_a$ is the covariant derivative operator. This theory is the simplest theory of matter coupled to Einstein gravity.

c) One more example is the theory describing massive fermions interacting both gravitationally and electromagnetically, which was considered in Ref. 16. The gauge group here is SU(2) $\times$U(1). The gravitational and electromagnetic degrees of freedom are described together by a SU(2)$\times$U(1) connection field A_a and canonically conjugate momentum field $\tilde{E}^a$. The fermionic degrees of freedom are described by two Grassmann-valued fields ξ, η taking values in the representations of spin 1/2 and charge ± 1. The Hamiltonian of the theory consists of the following three parts. The part generating the dynamics of the gravitational field is

$$\tilde{\tilde{H}}_{\mathrm{grav}} = \tfrac{1}{2}\,\underset{\sim}{\eta}_{abc}\,\mathrm{Tr}(\tilde{E}^a \tilde{E}^b \tilde{B}^c),$$

where the magnetic field $\tilde{B}^a$ is the dual of the curvature $F_{ab} = \underset{\sim}{\eta}_{abc}\tilde{B}^c$, $\underset{\sim}{\eta}_{abc}$ being the totally antisymmetric tensor of weight -1. The part generating the dynamics of the electromagnetic field is

$$\begin{aligned}\tilde{\tilde{H}}_{\mathrm{em}} = \tfrac{1}{32}\sigma^{-2}\,\underset{\sim}{\eta}_{abe}\underset{\sim}{\eta}_{cdf}\,\mathrm{Tr}(\tilde{E}^a\tilde{E}^c)[&\mathrm{Tr}(\tilde{E}^b\tilde{E}^d)\mathrm{Tr}(\tilde{B}^e)\mathrm{Tr}(\tilde{B}^f) - \mathrm{Tr}(\tilde{E}^b\tilde{E}^d)\mathrm{Tr}(\tilde{E}^e)\mathrm{Tr}(\tilde{E}^f)\\ &- \mathrm{Tr}(\tilde{E}^b)\mathrm{Tr}(\tilde{E}^d)\mathrm{Tr}(\tilde{B}^e)\mathrm{Tr}(\tilde{B}^f)],\end{aligned}$$

where σ is the square root of the determinant of the metric, as defined using the $\tilde{E}$ field. Finally, the part of the Hamiltonian responsible for the fermionic fields is

$$\tilde{\tilde{H}}_{\mathrm{ferm}} = [\tilde{E}^{aB}_A - \tfrac{1}{2}\,\mathrm{Tr}(\tilde{E}^a)\,\delta^B_A](\mathscr{D}_a \xi_B \tilde{\pi}^A + \mathscr{D}_a \eta^A \tilde{\omega}_B) + im[\tilde{\pi}^A \tilde{\omega}_A - (\sigma)^2 \eta^A \xi_A],$$

where m is the fermion mass.

In all the theories above particles appear in the theory with their antiparticles. Mathematically this is manifested by the fact that each irreducible representation that appears in the theory appears with its dual. However, this comes about in very different ways for the different theories above. The SU(2) theory in examples a) and b) contains only one fermionic field, but this field transforms under a representation that is isomorphic to its dual. It describes a particle that is its

own antiparticle. In example c) we have two different fermionic fields transforming under dual representations of the group $\mathrm{SU}(2)\times\mathrm{U}(1)$. This theory describes a particle that is distinct from its antiparticle.

One can see that representations must appear with their duals in order to construct a gauge-invariant mass term in the Hamiltonian, or any other gauge-invariant expression bilinear in the configuration fermionic fields. It is sensible therefore to limit our attention to theories for which each irreducible representation of G appears in the list I together with its dual. Actually, to streamline the exposition, we shall consider only the case where particles are distinct from their antiparticles. That is, we assume: (iia) the list I of representations of G is of the form $(\rho_1,\rho_1^*,\ldots,\rho_n,\rho_n^*)$. It is not difficult to extend our analysis to the case of particles that are their own antiparticles, and we mention a few of the changes that need to be made as we come to them.

Having described the class of theories which are the subject of this paper we are now ready to proceed with the quantization program. For this we need to construct fermionic path observables, which are the basic building blocks of the quantum theory.

IV. FERMIONIC PATH OBSERVABLES

We now introduce ''path observables,'' which play in our theories a role similar to that of the standard Wilson loop observables in theories of connections. We start by describing some path observables that are functions on the classical configuration space $\mathscr{A}\times\mathscr{F}$. These observables are built from the holonomy of the connection along a path in space together with fermion fields at the endpoints of this path.

Note first that given two fermionic fields ξ,η transforming under dual representations ρ and ρ^* of the gauge group, one can construct a gauge-invariant quantity from these fields using the G-invariant bilinear pairing between ρ^* and ρ. This quantity depends on the values ξ_p,η_p of these fields at a point $p\in\Sigma$ and is given simply by (ξ_p,η_p). More generally, given points p and p' and a path e from p to p', one can parallel transport ξ_p to p' along the path e using the connection A and then pair it with $\eta_{p'}$, obtaining the gauge-invariant quantity

$$(\xi_p|e|\eta_{p'})=\left(\rho\left(\mathscr{P}\exp\int_e A\right)\xi_p,\eta_{p'}\right)$$

which we call a ''configurational path observable.'' It is also helpful to have a graphical notation for a configurational path observable, in which we represent the oriented path e by a line with an arrow on it, and draw the fermionic fields ξ and γ as dots at the endpoints:

$$(\xi_p|e|\eta_{p'}) \qquad \xi \bullet\!\!\xrightarrow{\;e\;}\!\!\bullet\ \eta \tag{3}$$

We may think of the dot labelled by ξ as a particle and the the dot labelled by η as its antiparticle. Configurational path observables can be introduced for any representation in the list I together with its dual, or in other words, for any particle in the theory together with its antiparticle.

Note that the configurational path observables are even since they involve a product of two fermion fields. As a result they commute. We may also think of them as functions on the classical phase space, since they involve no momenta they all Poisson-commute. Let us now introduce the basic path observables involving fermionic momentum fields and describe the Poisson algebra these observables generate.

Recall that the momentum field $\tilde{\pi}$ canonically conjugate to ξ transforms under the dual representation. This means that one can apply the above construction to the $\xi,\tilde{\pi}$ fields. Namely, given two points p,p' and a path e connecting them, one can construct the quantity

$$(\xi_p|e|\tilde{\pi}_{p'})=\left(\rho\left[\mathscr{P}\exp\int_e A\right]\xi_p,\tilde{\pi}_{p'}\right),$$

which is gauge invariant and depends only on the values of fields $\xi, \widetilde{\pi}$ at the endpoints p, p' of the path e and on the holonomy of the connection field along e.

The quantity $(\xi|\gamma|\widetilde{\pi})$ is almost the one which we need. However, the Poisson bracket of this quantity with a configurational path observable $(\xi|\gamma|\eta)$ is a distribution. Following Ref. 16, let us introduce some ''averaged'' momentum observables in such a way that the Poisson bracket of a configurational path observable with a momentum observable is again a configurational path observable. In other words, let us introduce momentum observables in such a way that the resulting Poisson algebra contains no distributions. To do this we choose an arbitrary rule which specifies a path $e_{pp'}$ from p to p' given points p, p' in some region $\mathscr{R} \subset \Sigma$, an open set whose closure is compact. Having this rule at our disposal, we can integrate $(\xi|e_{pp'}|\widetilde{\pi})$ over $\mathscr{R}$ as a function of p'. We obtain the quantity

$$\int_{\mathcal{R}} d^n p' \, (\xi_p | e_{pp'} | \tilde{\pi}_{p'}) \qquad (4)$$

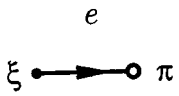

Here in our graphical notation we use a dot as before to stand for the fermionic configuration field, but use a little circle to represent the averaged fermionic momentum field. The result is a gauge-invariant quantity, which depends on values of the momentum field $\widetilde{\pi}$ and the connection field A over a region of the spatial manifold, on the value of ξ at the point p, and on a rule $(p, p') \mapsto e_{pp'}$.

Similarly, one can introduce a quantity

$$\int_{\mathcal{R}} d^n p' \, (\tilde{\omega}_{p'} | e_{p'p} | \eta_p) \qquad (5)$$

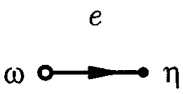

depending on the momentum field $\widetilde{\omega}$ canonically conjugate to η (or, physically, the momentum field of the antiparticle). One can also construct quantities involving two momentum fields

$$\int_{\mathcal{R}} d^n p \int_{\mathcal{R}'} d^n p' (\tilde{\omega}_p | e_{pp'} | \tilde{\pi}_{p'}), \qquad (6)$$

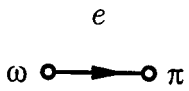

Again we represent averaged momentum fields by little circles.

Note that were we to consider a theory in which ξ coincided with its antiparticle η, there would be only one path observable linear in the momentum field, instead of two different quantities (4) and (5). Some of the Poisson brackets below would be different in this case, but the necessary modifications are not difficult.

The quantities we have described so far are not invariant under the Hermitian conjugation operation ''†.'' This operation is defined when the gauge fields of our theory include gravity, and the crucial property it satisfies is given by equation (1). Let us now introduce a set of ''real'' observables, by which we mean functions on the classical phase space that are preserved by this † operation. Since Hermitian conjugation sends fermionic momentum fields into configuration fields, real observables should involve both fields. Let us introduce the following observables which are linear in the fermionic momenta

$$\int_{\mathcal{R}} d^n p \, (\xi_p, \tilde{\pi}_p), \qquad (7)$$

$$\int_{\mathcal{R}} d^n p \, (\tilde{\omega}_p, \eta_p), \qquad (8)$$

where $(\cdot,\cdot)$ is the bilinear pairing between the dual representations ρ and ρ^*. As we shall see, when quantized these observables play the role of number operators measuring the number of particles of the appropriate sort in the region $\mathscr{R}$. The graphical representation that we use for these observables is as shown above.

The quantities we have introduced form a set of functions on the phase space of the theory. Let us now describe the Poisson algebra which these quantities generate (we also give a graphical description). First, the Poisson brackets of the observables linear in the momentum (4–5) with the configurational path observables (3) are again configurational path observables. The following identities can be verified:[16]

$$\left\{ \int_{\mathscr{R}} d^n p'(\xi_p|e_{pp'}|\tilde{\pi}_{p'}),(\xi_q|f|\eta_{q'}) \right\} = (\xi_p|e_{pq}\circ f|\eta_{q'}),$$

$$\left\{ \begin{array}{ccc} & e & \\ \xi \bullet & \longrightarrow & \circ\ \pi \end{array} , \begin{array}{ccc} & f & \\ \xi \bullet & \longrightarrow & \bullet\ \eta \end{array} \right\} = \begin{array}{ccc} & e\circ f & \\ \xi \bullet & \longrightarrow & \bullet\ \eta \end{array} \tag{9}$$

$$\left\{ \int_{\mathcal{R}} d^n p'(\tilde{\omega}_{p'}|e_{p'p}|\eta_p)\ ,\ (\xi_q|f|\eta_{q'}) \right\} = (\xi_q|f\circ e_{q'p}|\eta_p).$$

$$\left\{ \begin{array}{ccc} & e & \\ \omega \circ & \longrightarrow & \bullet\ \eta \end{array} , \begin{array}{ccc} & f & \\ \xi \bullet & \longrightarrow & \bullet\ \eta \end{array} \right\} = \begin{array}{ccc} & f\circ e & \\ \xi \bullet & \longrightarrow & \bullet\ \eta \end{array} \tag{10}$$

Here we use ''$\circ$'' to denote composition of paths. Also, it is assumed in (9) that $q \in \mathscr{R}$. If the point q lies outside of the region $\mathscr{R}$ the Poisson bracket is zero. Similarly, in (10) we assume $q' \in \mathscr{R}$; otherwise the Poisson bracket is zero. Note that if the result is not zero, in both cases it is given by a configurational path observable with the path being the composite of the paths e,f compatible with their orientation.

The Poisson bracket of a momentum observable (4) with another observable of this sort gives

$$\left\{ \int_{\mathscr{R}} d^n p'(\xi_p|e_{pp'}|\tilde{\pi}_{p'}), \int_{\mathscr{R}'} d^n q'(\xi_q|f_{qq'}|\tilde{\pi}_{q'}) \right\}$$

$$= \int_{\mathscr{R}'} d^n q'(\xi_p|e_{pp'}\circ f_{p'q'}|\tilde{\pi}_{q'}) + \int_{\mathscr{R}} d^n p'(\xi_q|f_{qq'}\circ e_{q'p'}|\tilde{\pi}_{p'}). \tag{11}$$

$$\left\{ \begin{array}{ccc} & e & \\ \xi \bullet & \longrightarrow & \circ\ \pi \end{array} , \begin{array}{ccc} & f & \\ \xi \bullet & \longrightarrow & \circ\ \pi \end{array} \right\} = \begin{array}{ccc} & e\circ f & \\ \xi \bullet & \longrightarrow & \circ\ \pi \end{array} + \begin{array}{ccc} & f\circ e & \\ \xi \bullet & \longrightarrow & \circ\ \pi \end{array}$$

Similarly, for (5) we get

$$\left\{ \int_{\mathscr{R}} d^n p'(\tilde{\omega}_{p'}|e_{p'p}|\eta_p), \int_{\mathscr{R}'} d^n q'(\tilde{\omega}_{q'}|f_{q'q}|\eta_q) \right\}$$

$$= \int_{\mathscr{R}} d^n p'(\tilde{\omega}_{p'}|e_{p'p}\circ f_{pq}|\eta_q) + \int_{\mathscr{R}'} d^n q'(\tilde{\omega}_{q'}|f_{q'q}\circ e_{qp}|\eta_p). \tag{12}$$

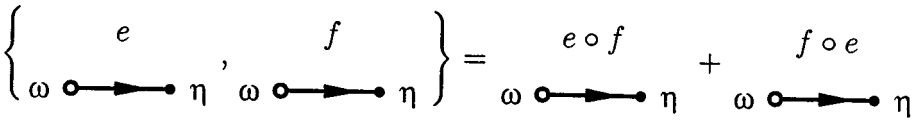

where we assume that the point where the configuration field in each of the observables is evaluated lies inside the region over which the momentum field in the other is smeared; otherwise one or both terms of the Poisson bracket vanishes.

The Poisson bracket of (6) and a configuration observable also consists of two terms:

$$\left\{ \int_{\mathscr{R}} d^n p \int_{\mathscr{R}'} d^n p' (\tilde{\omega}_p | e_{pp'} | \tilde{\pi}_{p'}), (\xi_q | f | \eta_{q'}) \right\}$$
$$= \int_{\mathscr{R}} d^n p (\tilde{\omega}_p | e_{pq} \circ f | \eta_{q'}) + \int_{\mathscr{R}'} d^n p' (\xi_q | f \circ e_{q'p'} | \tilde{\pi}_{p'}). \tag{13}$$

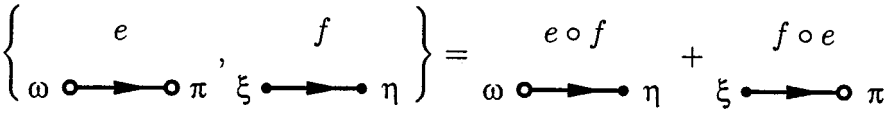

where again we assume $q \in \mathscr{R}'$ and $q' \in \mathscr{R}$. For the Poisson brackets between (6) and observables linear in momentum we have

$$\left\{ \int_{\mathscr{R}} d^n p \int_{\mathscr{R}'} d^n p' (\tilde{\omega}_p | e_{pp'} | \tilde{\pi}_{p'}), \int_{\mathscr{R}''} d^n q' (\xi_q | f_{qq'} | \tilde{\pi}_{q'}) \right\}$$
$$= \int_{\mathscr{R}} d^n p \int_{\mathscr{R}''} d^n q' (\tilde{\omega}_p | e_{pq} \circ f_{qq'} | \tilde{\pi}_{q'}), \tag{14}$$

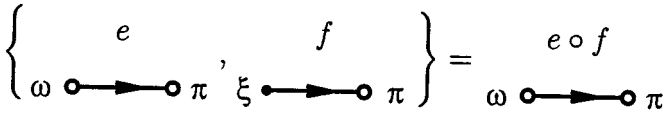

$$\left\{ \int_{\mathscr{R}} d^n p \int_{\mathscr{R}'} d^n p' (\tilde{\omega}_p | e_{pp'} | \tilde{\pi}_{p'}), \int_{\mathscr{R}''} d^n q' (\tilde{\omega}_{q'} | f_{q'q} | \eta_q) \right\}$$
$$= \int_{\mathscr{R}''} d^n q' \int_{\mathscr{R}'} d^n p' (\tilde{\omega}_{q'} | f_{q'q} \circ e_{qp'} | \tilde{\pi}_{p'}). \tag{15}$$

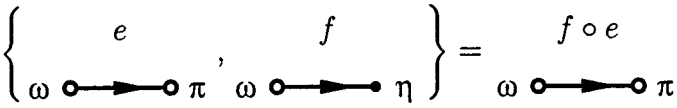

where in (14) we assume $q' \in \mathscr{R}'$, and in (15) we assume $q \in \mathscr{R}$; otherwise the brackets vanish.

One can also write down Poisson brackets between (3–6) and the real quantities (7–8). For example, the Poisson brackets between configurational path observables and these real quantities

are given by

$$\left\{ \int_{\mathcal{R}} d^n p(\xi_p, \tilde{\pi}_p) \ , \ (\xi_q|e|\eta_{q'}) \right\} = (\xi_q|e|\eta_{q'}),$$

$$\left\{ \xi \circledcirc \pi \ , \ \xi \xrightarrow{e} \eta \right\} = \xi \xrightarrow{e} \eta \tag{16}$$

$$\left\{ \int_{\mathcal{R}} d^n p(\tilde{\omega}_p, \eta_p) \ , \ (\xi_q|e|\eta_{q'}) \right\} = (\xi_q|e|\eta_{q'}).$$

$$\left\{ \omega \circledcirc \eta \ , \ \xi \xrightarrow{e} \eta \right\} = \xi \xrightarrow{e} \eta \tag{17}$$

Here in (16) we assume q is contained in $\mathcal{R}$, and in (17) we assume q' is contained in $\mathcal{R}$; otherwise the Poisson brackets vanish.

A nice feature of the graphical notation is that it gives a simple mnemonic for the Poisson bracket formulas (9–17) above. Note that in every case, the terms appearing in the Poisson bracket of two of our observables correspond to the ways of attaching them together by filling little empty circles (i.e., momentum fields) of one observable with dots (i.e., configurational fields) of the other, in a manner compatible with the orientations of the paths. (In theories for which particles coincided with their antiparticles, we would not include orientation arrows on the paths in our graphical notation, and not require compatibility of orientations when attaching paths together. This would give precisely the extra terms in the Poisson bracket relations that actually appear in such theories.)

We see that the observables we have introduced are closed with respect to Poisson brackets: A Poisson bracket of any two our observables is a linear combination of these observables. Note, however, that the path observables involving fermionic momenta depend on a rule assigning to any two points $p,p' \in \Sigma$ a path $e_{pp'}$. The rules that appear on the right hand side of (9)-(15) are of the form $e_{pq} \circ f_{qq'}$, i.e., they are formed by composition from the rules we started with, so if we fix a rule we do not obtain a Poisson algebra. Instead, we work with the ''big'' algebra generated by all observables of the form (3–8), allowing all possible choices of a rule $e_{pp'}$ and all possible regions $\mathcal{R}$ (open sets whose closure is compact). We denote this algebra as $\mathcal{B}$.

We may summarize by saying that we have introduced a Poisson algebra $\mathcal{B}$ of smooth functions on the classical phase space $T^*\mathcal{A} \times T^*\mathcal{F}$. This algebra is not closed under the Hermitian conjugation operator †. However, it has a distinguished subalgebra of observables preserved by the † operation, namely the real-linear combinations of products of the observables (7–8).

Our aim now is to quantize the observables (3–8), finding operators on some Hilbert space whose commutators mimic the classical Poisson bracket relations. We proceed in several steps. We first construct a kinematical Hilbert space. Then we describe a subspace of gauge invariant states in this space, which can be thought of as a Hilbert space of solution of Gauss law constraint, and represent the observables in $\mathcal{B}$ as operators on this space.

V. QUANTUM CONFIGURATION SPACE AND KINEMATICAL HILBERT SPACE

Recall that for diffeomorphism-invariant theories of connections, the quantum configuration space was most efficiently obtained by completing the algebra of cylinder functions to obtain a commutative C*-algebra, and taking the spectrum of this C*-algebra. We would like to follow a similar strategy for our theories with fermions. However, in our case the Gelfand-Naimark theorem is not applicable because of the presence of Grassmann-valued fields. Another strategy is needed.

Recall that in the case of theories of connections, we could also describe the quantum configuration space as a certain completion of the classical configuration space. We can implement this idea in our case and construct the fermionic quantum configuration space $\overline{\mathcal{F}}$ as a completion

of the classical configuration space $\mathscr{F}$. Namely, we define $\overline{\mathscr{F}}$ to be the space of all (not necessarily smooth) sections of the vector bundle F over Σ. The space $\mathscr{F}$ is dense in $\overline{\mathscr{F}}$ in the topology of pointwise convergence.

Alternatively, one can describe $\overline{\mathscr{F}}$ as the projective limit of configuration spaces of certain fermionic systems with finitely many degrees of freedom. Namely, given any finite subset of points $V \subset \Sigma$ we may consider the configuration space of a system of fermions living at the points of V; this is the product

$$\mathscr{F}_V = \prod_{p \in V} F_p$$

of copies of fibers of F, one for each point in V. Given finite subsets $V \subseteq V'$ of points in Σ there is a natural projection from $\mathscr{F}_{V'}$ to $\mathscr{F}_V$, and the projective limit of all these spaces $\mathscr{F}_V$ is $\overline{\mathscr{F}}$.

We also wish to have a kinematical Hilbert space for the fermion fields. Although we shall denote this Hilbert space by $L^2(\overline{\mathscr{F}})$, it is not a space of square-integrable functions on $\overline{\mathscr{F}}$ in the standard sense. We use the notation $L^2(\overline{\mathscr{F}})$ just to emphasize the similarity in notations between the fermionic and the gauge fields degrees of freedom.

The space $L^2(\overline{\mathscr{F}})$ can be constructed in a variety of equivalent ways. Perhaps the easiest is to use the fact that the fibers F_p are finite-dimensional Hilbert spaces and let $L^2(\overline{\mathscr{F}})$ be the fermionic Fock space over the Hilbert space direct sum $\oplus_{p \in \Sigma} F_p$. This is the natural Hilbert space completion of the exterior algebra

$$\Lambda(\oplus_{p \in \Sigma} F_p).$$

Alternatively, we can think of $L^2(\overline{\mathscr{F}})$ as the infinite ''grounded'' tensor product

$$\otimes_{p \in \Sigma} \Lambda F_p$$

of the fermionic Fock spaces ΛF_p, one for each point p in space. (For an introduction to the mathematics of fermionic Fock spaces and infinite grounded tensor products, see Ref. 17.)

In the case of connection fields we can define the kinematical Hilbert space as a completion of an algebra of cylinder functions. An analogous fact holds for fermion fields if we define a ''cylinder function'' on $\mathscr{F}$ to be a smooth function on $\mathscr{F}$ depending only on the value of the fermion fields at finitely many points of Σ. The space of cylinder functions is thus

$$\Lambda(\oplus_{p \in \Sigma} F_p^*) \subseteq \Lambda \mathscr{F}^*,$$

the exterior algebra over the algebraic direct sum $\oplus_{p \in \Sigma} F_p^*$. Since each F_p is a Hilbert space, this space is naturally isomorphic to

$$\Lambda(\oplus_{p \in \Sigma} F_p),$$

and by the above remarks we see that it can be completed to obtain $L^2(\overline{\mathscr{F}})$.

Finally, we may also describe $L^2(\overline{\mathscr{F}})$ in the following physically appealing way. For each finite set V of points of Σ we may start with the theory of fermions living at these points. The configuration space for this theory is the finite-dimensional space $\mathscr{F}_V$ described above. Quantizing this system we obtain the fermionic Fock space over $\mathscr{F}_V$, which is just the exterior algebra $\Lambda \mathscr{F}_V$ equipped with its standard inner product. Given finite subsets $V \subset V'$ of points in Σ there is a natural projection from $\Lambda \mathscr{F}_{V'}$ to $\Lambda \mathscr{F}_V$, and the projective limit of these Hilbert spaces is the kinematical Hilbert space $L^2(\overline{\mathscr{F}})$.

Putting together the quantum configuration spaces for the gauge field and fermionic degrees of freedom, we obtain the quantum configuration space $\overline{\mathscr{A}} \times \overline{\mathscr{F}}$ for the whole theory. Similarly, the kinematical Hilbert space for the full theory is $L^2(\overline{\mathscr{A}}) \otimes L^2(\overline{\mathscr{F}})$, which we denote by $L^2(\overline{\mathscr{A}} \times \overline{\mathscr{F}})$. Note that cylinder functions on $\mathscr{A}$ are dense in $L^2(\overline{\mathscr{A}})$, and cylinder functions on $\mathscr{F}$ are

dense in $L^2(\overline{\mathscr{F}})$. Thus if we define the algebra of ''cylinder functions'' on $\mathscr{A}\times\mathscr{F}$ to be the tensor product of the algebras of cylinder functions on $\mathscr{A}$ and cylinder functions on $\mathscr{F}$, it follows immediately that cylinder functions on $\mathscr{A}\times\mathscr{F}$ are dense in $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$.

In simple terms, a cylinder function on $\mathscr{A}\times\mathscr{F}$ is a wavefunction depending on the connection only via its holonomies along finitely many paths in Σ and on the fermionic fields only via their values at finitely many points of Σ. As we shall see in the next section, cylinder functions may be thought of as states of gauge theories coupled to fermions living on graphs in Σ with finitely many edges and vertices. Cylinder functions play an important role in the quantum theory, for it is natural to define quantum operators first on the dense subspace of cylinder functions, and then to extend them to the entire Hilbert space $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$.

The Hilbert space $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$ serves only as an auxiliary Hilbert space of the theory, since its states are typically not invariant under gauge transformations, diffeomorphisms of space, or time evolution. To find physical states we must impose constraints corresponding to invariance under these symmetries. In the next section we describe a basis of solutions to the Gauss law constrant, that is, gauge-invariant elements of $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$.

VI. GAUGE-INVARIANT STATES

There is a natural unitary representation of the group $\mathscr{G}$ of smooth gauge transformations on the Hilbert space of quantum states $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$. In this section we describe the subspace of $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$ consisting of the states which are invariant under these gauge transformations. We denote this space by $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$. We also describe an orthonormal basis for $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$ that is a generalization of the spin network basis of $L^2(\overline{\mathscr{A}}/\mathscr{G})$.

Given any graph γ analytically embedded in Σ, let E be its set of edges and let V be its set of vertices. Any edge $e\in E$ begins at some vertex $s(e)$ called its ''source'' and ends at some vertex $t(e)$ called its ''target''. We define the space of connections on γ to be

$$\mathscr{A}_\gamma=\prod_{e\in E}\mathscr{A}_e,$$

where $\mathscr{A}_e$ is the space of transporters from $s(e)$ to $t(e)$. If we trivialize P at all the vertices of γ we can think of $\mathscr{A}_\gamma$ as a product of copies of the gauge group G, one for each edge. Similarly, we define the classical configuration space for fermion fields on γ to be

$$\mathscr{F}_\gamma=\prod_{p\in V}F_p.$$

We define $L^2(\mathscr{A}_\gamma)$ using normalized Haar measure on G, define $L^2(\mathscr{F}_\gamma)$ to be the fermionic Fock space over $\mathscr{F}_\gamma$, and set

$$L^2(\mathscr{A}_\gamma\times\mathscr{F}_\gamma)=L^2(\mathscr{A}_\gamma)\otimes L^2(\mathscr{F}_\gamma).$$

We may think of this as a subspace of $L^2(\overline{\mathscr{A}}\times\overline{\mathscr{F}})$.

For any graph γ there is a unitary representation of $\mathscr{G}$ on $L^2(\mathscr{A}_\gamma\times\mathscr{F}_\gamma)$, and we denote the subspace consisting of gauge-invariant elements by $L^2((\mathscr{A}_\gamma\times\mathscr{F}_\gamma)/\mathscr{G})$. As in the case of pure gauge fields,[7,8] the union of these spaces as γ ranges over all graphs is dense in $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\overline{\mathscr{G}})$. Thus to describe $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$, it suffices to describe these spaces for arbitrary graphs γ.

Trivializing P at all the vertices of γ we have

$$L^2(\mathscr{A}_\gamma\times\mathscr{F}_\gamma)=\otimes_{e\in E}L^2(G)\otimes\otimes_{v\in V}\Lambda F_v. \tag{18}$$

Let us decompose the Grassmann algebra ΛF_v at each vertex as an orthogonal direct sum of irreducible representations of G: $\Lambda F_v=\oplus_{\rho\in S}\rho$, where we have denoted by S the list of irreduc-

ible representations appearing in ΛF_v. Similarly, the Peter-Weyl theorem says that $L^2(G) \cong \oplus_{\rho \in \mathrm{Rep}(G)} \rho \otimes \rho^*$, where the set $\mathrm{Rep}(G)$ contains one irreducible continuous unitary representation of G from each equivalence class, so that (18) implies

$$L^2(\mathscr{A}_\gamma \times \mathscr{F}_\gamma) = \otimes_{e \in E}(\oplus_{\rho \in \mathrm{Rep}(G)} \rho \otimes \rho^*) \otimes \otimes_{v \in V}(\oplus_{\rho \in S} \rho). \tag{19}$$

The right hand side of (19) can be rewritten as follows:

$$L^2(\mathscr{A}_\gamma \times \mathscr{F}_\gamma) = \oplus_{\rho_e \in \mathrm{Rep}(G), \rho_v \in S} \otimes_{v \in V}((\otimes_{t(e)=v} \rho_e^* \otimes \otimes_{s(e)=v} \rho_e) \otimes \rho_v). \tag{20}$$

The sum here runs over all labellings of edges e by irreducible representations ρ_e of the group G, and by all labellings of vertices by irreducible representations from the list S. It follows that

$$L^2((\mathscr{A}_\gamma \times \mathscr{F}_\gamma)/\mathscr{G}) = \oplus_{\rho_e \in \mathrm{Rep}(G), \rho_v \in S} \otimes_v \mathrm{Inv}((\otimes_{t(e)=v} \rho_e^* \otimes \otimes_{s(e)=v} \rho_e) \otimes \rho_v), \tag{21}$$

where Inv denotes the G-invariant subspace of the given representation. Note that $\mathrm{Inv}(\otimes_{t(e)=v} \rho_e^* \otimes \otimes_{s(e)=v} \rho_e \otimes \rho_v)$ has a natural inner product, and is isomorphic to the space of intertwining operators from $\otimes_{t(e)=v} \rho_e$ to $(\otimes_{s(e)=v} \rho_e) \otimes \rho_v$. We denote this space of intertwining operators by

$$\mathrm{Hom}(\otimes_{t(e)=v} \rho_e, (\otimes_{s(e)=v} \rho_e) \otimes \rho_v).$$

As in the case of pure gauge fields, the above discussion also gives an orthonormal basis of spin network states for $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$. Unlike spin networks for theories describing only gauge fields, the spin networks for theories with fermions can contain edges with ''open ends.'' Such open ends represent fermionic degrees of freedom. We describe the spin network basis for $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$ in the following theorem summarizing the results of this section:

Theorem 1: *The space $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$ has an orthonormal basis of ''fermionic spin network states,'' each such state being specified by a choice of*:

1. *a graph γ analytically embedded in Σ,*
2. *a labelling of each edge e of γ by an irreducible representation $\rho_e \in \mathrm{Rep}(G)$,*
3. *a labelling of each vertex v of γ by an irreducible representation $\rho_v \in S$,*
4. *a labelling of each vertex v of γ by an intertwining operator*

$$\iota_v \in \mathrm{Hom}(\otimes_{t(e)=v} \rho_e, (\otimes_{s(e)=v} \rho_e) \otimes \rho_v)$$

chosen from an orthonormal basis of such intertwining operators.

We call the data $\Gamma = (\gamma, \rho, \iota)$ a ''fermionic spin network,'' and denote the fermionic spin network state corresponding to Γ as above by Ψ_Γ.

It is interesting to note the similarity between these fermionic spin networks and the extended spin networks introduced by Ashtekar and Lewandowski[18] as a basis of $L^2(\overline{\mathscr{A}})$. Note, however, that unlike extended spin networks of theories of connections, fermionic spin networks describe fermionic degrees of freedom and are gauge invariant.

VII. QUANTIZATION OF FERMIONIC PATH OBSERVABLES

We have now constructed the kinematical Hilbert space $L^2(\overline{\mathscr{A}} \times \overline{\mathscr{F}})$ as a completion of the space of cylinder functions on $\overline{\mathscr{A}} \times \overline{\mathscr{F}}$, and found a spin network basis for the gauge-invariant subspace $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$. Our construction so far has nothing directly to do with the fermionic path observables introduced in Sec. IV. Thus, we have to make sure that there is a way to quantize these observables and obtain operators on $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$. In fact, we shall obtain operators on $L^2(\overline{\mathscr{A}} \times \overline{\mathscr{F}})$ that commute with gauge transformations, and thus map $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$ to itself.

There is an obvious way to promote the configurational path observables into operators. We represent the quantities $(\xi | e | \eta)$ by multiplication operators $(\xi | \hat{e} | \eta)$ on the space of cylinder functions on $\mathscr{A} \times \mathscr{F}$ as follows:

$$(\xi|\hat{e}|\eta)\Psi = (\xi|e|\eta)\Psi, \tag{22}$$

where $(\xi|e|\eta)$ on the right hand side is considered as a cylinder function on $\mathcal{A}\times\mathcal{F}$. Note that if Ψ is a cylindrical function so is $(\xi|e|\eta)\Psi$. One has, therefore, a well-defined action of $(\xi|\hat{e}|\eta)$ on the space of cylinder functions.

The observables (4–8) involving fermionic momenta can also be first promoted to operators on the space of cylinder functions. The natural way to do this is to replace the fermionic momentum field $\tilde{\pi}$ in such observables by the corresponding Berezin derivative $\delta/\delta\xi$. Berezin derivatives[19] are usually used in the context of field theories in Minkowski spacetime. However, we can also make sense of this notion in the context of diffeomorphism-invariant theories.

We introduce Berezin derivatives as operators on the space of cylinder functions as follows. Recall that the algebra of cylinder functions on $\mathcal{F}$ is the exterior algebra

$$\Lambda(\oplus_{p\in\Sigma}F_p^*).$$

For any point p choose an orthonormal basis $\xi^\alpha(p)$ of the subspace of F_p^* corresponding to the representation $\rho\in I$ under which the field ξ transforms. Note that $\xi^\alpha(p)$ is a cylinder function on $\mathcal{F}$, and the algebra of cylinder functions on $\mathcal{F}$ is generated by functions of this form. Thus, formally speaking, we may define the Berezin derivative operator to be the superderivation (i.e., graded derivation) of the algebra of cylinder functions such that

$$\frac{\delta}{\delta\xi^\alpha(p)}\xi^\beta(p') = \delta_\alpha^\beta\tilde{\delta}^n(p,p'), \tag{23}$$

where δ_α^β is the Kroneker delta. However, the right hand side is distributional, so strictly speaking we should smear this equation with a ''test function.'' In other words, for any section f of $P\times_G\rho$, we define

$$\int_\Sigma d^n p f^\beta(p)\frac{\delta}{\delta\xi^\beta(p)}$$

to be the unique superderivation on the algebra of cylinder functions such that

$$\left(\int_\Sigma d^n p f^\beta(p)\frac{\delta}{\delta\xi^\beta(p)}\right)\xi^\alpha(p') = f^\alpha(p').$$

To describe the action of this operator on an arbitrary cylinder function, let us introduce the interior product operator $i[\xi]_\alpha(p)$, namely, the adjoint of exterior multiplication by $\xi^\alpha(p)$. We then have

$$\left(\int_\Sigma d^n p f^\beta(p)\frac{\delta}{\delta\xi^\beta(p)}\right)\Psi = \sum_{p\in\Sigma} f^\alpha(p) i[\xi]_\alpha(p)\Psi,$$

where the right hand side, for any cylinder function Ψ, contains only a finite number of non-zero terms.

Using this definition of Berezin derivative one can proceed with the quantization of observables (4–6). However, there is an operator ordering ambiguity that needs to be resolved in this procedure. Since the observables (4–6) involve both fermion fields and their canonically conjugate momenta, which do not anticommute, we must specify which acts first. Note that one of our classical quantities, namely, $\int_{\mathcal{R}} d^n p(\xi_p, \tilde{\pi}_p)$, has a natural interpretation as the number of particles of type ξ within the region $\mathcal{R}$. To obtain this number operator when we quantize, we need to choose the operator ordering in which the Berezin derivative corresponding to the momentum field $\tilde{\pi}_p$ acts before the multiplication operator corresponding to the configuration field ξ_p. Let us therefore demand that derivatives act before multiplication operators for all our observables, which amounts to a ''normal ordering'' prescription.

Following this choice of operator ordering we can now quantize all our fermionic observables, obtaining operators on the space of cylinder functions. More precisely, we define the quantum analogs of the observables (4–6) as follows:

$$\int_{\mathscr{R}} d^n p'(\xi_p|\hat{e}_{pp'}|\tilde{\pi}_{p'}) = \sum_{p' \in \mathscr{R}} \xi^\alpha(p) \left(\mathscr{P} \int_e A \right)_\alpha^\beta i[\xi]_\beta(p'), \tag{24}$$

$$\int_{\mathscr{R}} d^n p'(\tilde{\omega}_{p'}|\hat{e}_{p'p}|\eta_p) = \sum_{p' \in \mathscr{R}} \eta_\alpha(p) \left(\mathscr{P} \int_e A \right)_\beta^\alpha i[\eta]^\beta(p'), \tag{25}$$

$$\int_{\mathscr{R}} d^n p \int_{\mathscr{R}'} d^n p'(\tilde{\omega}_p|\hat{e}_{pp'}|\tilde{\pi}_{p'}) = \sum_{p \in \mathscr{R}} \sum_{p' \in \mathscr{R}'} i[\eta]^\alpha(p) \left(\mathscr{P} \int_e A \right)_\alpha^\beta i[\xi]_\beta(p'), \tag{26}$$

$$\int_{\mathscr{R}} d^n p(\xi_p, \tilde{\pi}_p) = \sum_{p \in \mathscr{R}} \xi^\alpha(p) i[\xi]_\alpha(p), \tag{27}$$

$$\int_{\mathscr{R}} d^n p(\tilde{\omega}_p, \eta_p) = \sum_{p \in \mathscr{R}} \eta_\alpha(p) i[\eta]^\alpha(p). \tag{28}$$

The operators (22–28) map cylinder functions to cylinder functions, and extend uniquely to bounded linear operators on $L^2(\bar{\mathscr{A}} \times \bar{\mathscr{F}})$. Since they commute with gauge transformations they may also be thought of as operators on $L^2((\bar{\mathscr{A}} \times \bar{\mathscr{F}})/\mathscr{G}$. Finite linear combinations of fermionic spin network states are dense in $L^2((\bar{\mathscr{A}} \times \bar{\mathscr{F}})/\mathscr{G})$, so the actions of these operators on $L^2((\bar{\mathscr{A}} \times \bar{\mathscr{F}})/\mathscr{G})$ are determined by their actions on fermionic spin network states. On such states the action is simply described in our graphical notation, which is very nice for explicit computations. Let us give several simple examples; generalization to arbitrary spin network states is straightforward. We consider the action of our operators on a fermionic spin network state containing a single path, that is, a cylinder function of the form $(\xi|e|\eta)$. First, the fermionic path observables linear in momentum, (25–26), act to compose the path e with a path f. That is,

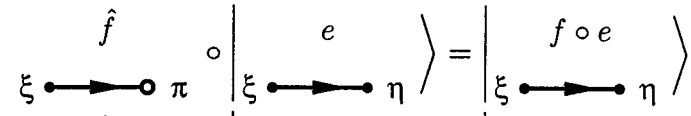

(29)

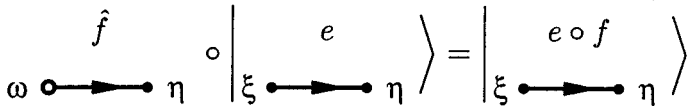

(30)

if the relevant endpoint of e lies within the region $\mathscr{R}$; otherwise we get zero. The operators (27–28) on the same state are given as follows:

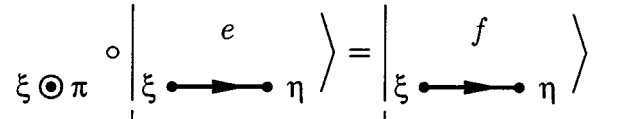

(31)

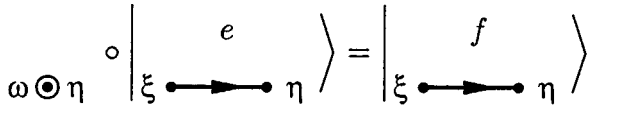

(32)

if the relevant endpoint of e lies within the region $\mathscr{R}$; otherwise we get zero. In general, these operators count the number of the fermions (resp. antifermions) in the region $\mathscr{R}$. One can easily see that all fermionic spin network states are eigenstates of these operators.

The operator corresponding to a configurational path observable acts to create a path:

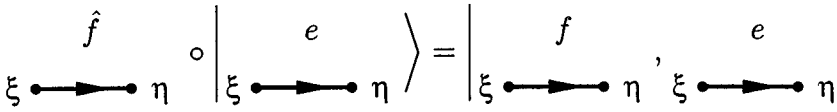

(33)

The result is a spin network state containing two paths. The operator involving two momentum fields acts to close off a path:

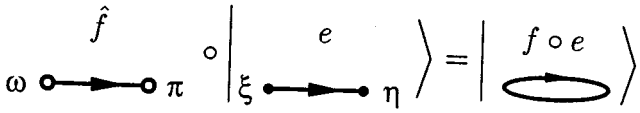

(34)

if the endpoints of e lie within the corresponding regions of integration in the operator $\int_{\mathscr{R}} d^n p \int_{\mathscr{R}'} d^n p' (\tilde{\omega}_p | \hat{e}_{pp'} | \tilde{\pi}_{p'})$; otherwise we get zero. Note that the result here is the simple loop state so familiar from the loop representation of pure gravity.

In short, we can quantize our fermionic path observables, implementing them as operators on $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$. As usual in quantization, we do not obtain a true representation of the Poisson algebra $\mathscr{B}$, in which Poisson brackets of arbitrary observables in $\mathscr{B}$ are mapped to commutators of the corresponding operators. However, if we restore the factors of $\hbar$ which we have suppressed above, and work with $\hbar\, \delta/\delta\xi^\alpha(p)$ where above we use $\delta/\delta\xi^\alpha(p)$, we find that the commutators of operators (22–28) match the Poisson brackets of the observables (3–8) up to terms of order $\hbar^2$.

VIII. DIFFEOMORPHISM-INVARIANT STATES

The Hilbert space $L^2((\overline{\mathscr{A}} \times \overline{\mathscr{F}})/\mathscr{G})$ of gauge-invariant states is still too large to be the space of physical states of the theory. The group of diffeomorphisms $\mathrm{Diff}(\Sigma)$ acts non-trivially on the elements of $L^2((\overline{\mathscr{A}} \times \mathscr{F})/\mathscr{G})$. To find diffeomorphism-invariant states one should adopt the strategy described in Sec. II, namely the ''group averaging'' technique of Ashtekar, Lewandowski, Marolf, Mourão and Thiemann.[2] Since the details of the construction are very similar to those in the case of theories with pure connection degrees of freedom, we only sketch the main steps here.

We look for diffeomorphism-invariant elements of the space $\mathscr{C}^*$, the topological dual of the space $\mathscr{C}$ of gauge-invariant cylindrical functionals. We construct such solutions by averaging fermionic spin network states, just as described in Sec. II. Any state $\Psi \in \mathscr{C}$ that is a linear combination of such states defines in this way a diffeomorphism-invariant element $\overline{\Psi}$ of $\mathscr{C}^*$. There is an inner product on the space of such diffeomorphism-invariant states given by

$$\langle \overline{\Psi} | \overline{\Phi} \rangle = \overline{\Psi}(\Phi).$$

Completing this space in this inner product, we obtain the Hilbert space $\mathscr{H}_{\mathrm{diff}}$ of diffeomorphism-invariant states. This space has an orthogonal basis given by states of the form Ψ_Γ, where Γ as we let Γ range over a set of fermionic spin networks containing one from each diffeomorphism equivalence class (again restricting ourselves to graphs of type I).

IX. CONCLUSIONS

Having set up a kinematical framework for diffeomorphism-invariant gauge theories coupled to fermions, and dealt with the Gauss law and diffeomorphism constraints, there are still some interesting questions concerning superselection sectors to deal with before moving on to the much deeper problem of the Hamiltonian constraint. In general, one can obtain superselection sectors (invariant subspaces) for the action of an algebra on a Hilbert space by studying the orbits of

various states under the action of this algebra. Let $\hat{\mathscr{B}}$ denote the algebra of operators generated by the quantized fermionic path observables (22–28) of Sec. (VII). It is especially interesting to consider the orbit of the ''empty state'' under this algebra. By the ''empty state'' we mean the vector in $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$ given by

$$1\otimes 1 \in L^2(\overline{\mathscr{A}})\otimes\Lambda(\oplus_{p\in\Sigma}F_p).$$

This is the fermionic spin network state corresponding to a graph with no vertices and no edges. In general the empty state is not a cyclic vector for $\hat{\mathscr{B}}$; i.e., the orbit $\hat{\mathscr{B}}(1\otimes 1)$ is not dense in $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$. Thus $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$ has non-trivial superselection sectors.

There are various reasons why the empty state may not be a cyclic vector. First, there may be a non-trivial subgroup $H\subseteq G$ that acts trivially on all the representations in our list I corresponding to the fermions in our theory. If this occurs, the empty state will never be a cyclic vector for $\hat{\mathscr{B}}$. To see this, note that H is a closed normal subgroup of G so $G'=G/H$ is a compact connected Lie group. We may naturally construct from P a principal G'-bundle P' and think of F as associated to P'. Letting $\mathscr{G}'$ denote the group of gauge transformations of P', and letting $\mathscr{A}'$ denote the space of smooth connections on P', one may check that $L^2(\overline{\mathscr{A}}'\times\overline{\mathscr{F}}')/\mathscr{G}')$ is in a natural way a proper closed subspace of $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$. One may also check that the orbit of the empty state lies in $L^2(\overline{\mathscr{A}}'\times\overline{\mathscr{F}}')/\mathscr{G}')$, so the empty state is not a cyclic vector.

It is easy to get around this problem, though, because in this situation there is no harm in replacing the gauge group G of our theory by G' and taking $L^2(\overline{\mathscr{A}}'\times\overline{\mathscr{F}}')/\mathscr{G}')$ as our space of gauge-invariant states. For example, consider the gravitational and electromagnetic fields coupled to a charged fermionic field as in example c) of Sec. III. The particle in this theory transform according to the spin 1/2, charge 1 representation of $G=\mathrm{SU}(2)\times\mathrm{U}(1)$, while its antiparticle transforms according to the spin 1/2, charge -1 representation. The group H thus consists of the elements $\pm(1,1)$, so there is no harm in replacing G by the group $G/H=\mathrm{U}(2)$.

Suppose therefore that this problem does not occur: only the identity element of G acts trivially on all the representations in the list I. The empty state may still fail to be a cyclic vector. In fact, this always occurs if $G=\mathrm{SU}(n)$ for odd values of $n>2$ when the list I consists only of the fundamental representation and its dual. The reason is that the nth exterior power of the fundamental representation is the trival representation, so one may construct a gauge-invariant quantity of order n in the fermion field. Applying this to the empty state gives a fermionic spin network state living on the graph with one vertex and no edges. This element $L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$ cannot be in the orbit of the empty state because the operators (22–28) always change the total fermion number by an even integer.

However, the question of whether the empty state is a cyclic vector remains open for the interesting cases of $G=\mathrm{SU}(2)$ and $G=\mathrm{SU}(2)\times\mathrm{U}(1)$. Another interesting open question is as follows. At least for $G=\mathrm{U}(n)$ and $G=\mathrm{SU}(n)$ one knows that products of Wilson loop operators labelled by the fundamental representation span $L^2(\overline{\mathscr{A}}/\mathscr{G})$. As noted in Sec. VII, such Wilson loops can be expressed as products of fermionic path observables. Therefore in theories with these gauge groups and fermions tranforming under the fundamental representation one knows that the closure of the orbit of the empty state at least contains $L^2(\overline{\mathscr{A}}/\mathscr{G})\subseteq L^2((\overline{\mathscr{A}}\times\overline{\mathscr{F}})/\mathscr{G})$. For what other groups and choices of representations does this result hold?

ACKNOWLEDGMENTS

We would like to thank H. Morales-Técotl and Y. Shtanov for many stimulating discussions when the basic ideas behind this work were being developed. K. K. also thanks A. Ashtekar for a useful discussion. We are grateful to the Erwin Schrödinger Institute for Mathematical Physics in Vienna for their hospitality while much of this paper was written, and especially to Peter Aichelburg and Abhay Ashtekar for running the Workshop on Mathematical Problems of Quantum Gravity held there. J. B. also thanks the Center for Gravitational Physics and Geometry for their

hospitality while this paper was being completed. K. K. was supported, in part, by the International Soros Science Education Program (ISSEP) through grant No. PSU062052, and this research was also supported in part by NSF Grant No. PHY95-14240.

[1] A. Ashtekar, ''New variables for classical and quantum gravity,'' Phys. Rev. Lett. **57**, 2244–2247 (1986); ''New Hamiltonian formulation of general relativity,'' Phys. Rev. D **36**, 1587–1602 (1987).
[2] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, ''Quantization of diffeomorphism invariant theories of connections with local degrees of freedom,'' J. Math. Phys. **36**, 6456–6493 (1995).
[3] T. Jacobson, ''Fermions in canonical gravity,'' Class. Quantum Grav. **5**, L143–L148 (1988).
[4] T. Thiemann, ''Reality conditions including transforms for quantum gauge field theory and quantum gravity,'' Class. Quantum Grav. **13**, 1383–1403 (1996).
[5] T. Thiemann, ''Anomaly-free formulation of non-perturbative, four-dimensional Lorentzian quantum gravity,'' Phys. Lett. B **380**, 257–264 (1996).
[6] T. Thiemann, ''Kinematical Hilbert spaces for fermionic and Higgs quantum field theories,'' preprint HUTMP-97/B364, also available as gr-qc/9705021; T. Thiemann, ''QSD V: Quantum gravity as the natural regulator of matter quantum field theories,'' preprint HUTMP-97/B-357, also available as gr-qc/9705019.
[7] A. Ashtekar and J. Lewandowski, ''Differential geometry on the space of connections via graphs and projective limits,'' J. Geom. Phys. **17**, 191–230 (1995); ''Representation theory of analytic holonomy C*-algebras,'' in *Knots and Quantum Gravity*, edited by J. Baez (Oxford University Press, Oxford, 1994); ''Projective techniques and functional integration,'' J. Math. Phys. **36**, 2170–2191 (1995).
[8] J. Baez, ''Diffeomorphism-invariant generalized measures on the space of connections modulo gauge transformations,'' in *Proceedings of the Conference on Quantum Topology*, edited by David N. Yetter (World Scientific, Singapore, 1994); ''Generalized measures in gauge theory,'' Lett. Math. Phys. **31**, 213–223 (1994); ''Spin networks in gauge theory,'' Adv. Math. **117**, 253–272 (1996); ''Spin networks and nonperturbative quantum gravity,'' in *The Interface of Knots and Physics*, edited by Louis Kauffman (American Mathematical Society, Providence, Rhode Island, 1996).
[9] J. Baez and S. Sawin, ''Functional integration on spaces of connections,'' to appear in J. Funct. Anal. preprint available as q-alg/9507023.
[10] C. Rovelli and L. Smolin, ''Loop representation for quantum general relativity,'' Nucl. Phys. B**331**, 80–152 (1990).
[11] C. Rovelli and L. Smolin, ''Spin networks in quantum gravity,'' Phys. Rev. D **52**, 5743–5759 (1995).
[12] B. DeWitt, *Supermanifolds* (Cambridge University Press, Cambridge, 1992).
[13] V. Husain (private communication to K. Krasnov).
[14] V. Husain and K. Kuchař, ''General covariance, the new variables, and dynamics without dynamics,'' Phys. Rev. D **42**, 4070–4077 (1990).
[15] C. Rovelli and H. Morales-Tecótl, ''Fermions in quantum gravity,'' Phys. Rev. Lett. **72**, 3642–3645 (1995); ''Loop space representation of quantum fermions and gravity,'' Nucl. Phys. B**451**, 325–361 (1995).
[16] K. Krasnov, ''Quantum loop representation for fermions coupled to Einstein-Maxwell field,'' Phys. Rev. D **53**, 1874–1888 (1996).
[17] J. Baez, I. Segal and Z. Zhou, *An Introduction to Algebraic and Constructive Quantum Field Theory* (Princeton University Press, Princeton, 1992).
[18] A. Ashtekar and J. Lewandowski, ''Quantum Theory of Geometry I: Area Operators,'' Class. Quantum Grav. **14**, A55–A81 (1997).
[19] F. A. Berezin, *Introduction to Superanalysis* (Kluwer, Boston, 1987).

Onset of superconductivity in decreasing fields for general domains

Andrew Bernoff[a)]
Department of Mathematics, Duke University, Durham, North Carolina 27708

Peter Sternberg[b)]
Department of Mathematics, Indiana University, Bloomington, Indiana 47405

(Received 10 September 1997; accepted for publication 12 November 1997)

Ginzburg–Landau theory has provided an effective method for understanding the onset of superconductivity in the presence of an external magnetic field. In this paper we examine the instability of the normal state to superconductivity with decreasing magnetic field for a closed smooth cylindrical region of arbitrary cross-section subject to a vertical magnetic field. We examine the problem asymptotically in the boundary layer limit (i.e., when the Ginzburg–Landau parameter, k, is large). We demonstrate that instability first occurs in a region exponentially localized near the point of maximum curvature on the boundary. The transition occurs at a value of the magnetic field associated with the half-plane at leading order, with a small positive correction due to the curvature (which agrees with the transition problem for the disc), and a smaller correction due to the second derivative of the curvature at the maximum. © *1998 American Institute of Physics.*
[S0022-2488(98)00203-5]

I. INTRODUCTION

The application of a sufficiently large magnetic field is known to destroy the superconducting property of a superconductor. If one keeps a sample below some critical temperature, however, then experiments reveal the emergence of a superconducting state once the external field is lowered below some critical value. In addition to a loss of normal resistivity, the material in the superconducting state gains the ability to expel the applied magnetic field. Here we investigate this bifurcation using the Ginzburg–Landau theory of superconductivity as a model.[1]

This problem has been treated analytically in various settings in which the sample is cylindrical (i.e., of constant cross section), and the applied magnetic field points in the axial direction. This assumption allows for a two-dimensional approach to the analysis. Saint-James and de Gennes[2] consider the case of a half-plane and infinite slab, and using the Ginzburg–Landau theory they linearize the equations about the so-called ''normal state'' in which the applied field totally permeates the medium. Through an explicit solution, they find the critical value of the applied field where the normal state loses stability—frequently referred to as H_{c_3} in the literature—to be approximately $1.7k$. Here k is the Ginzburg–Landau parameter, a dimensionless material constant defined as the ratio of two important length scales in the system. In their analysis and in ours as well, k is taken to be greater than $1/\sqrt{2}$, so that the material is a so-called type-II superconductor. Saint-James and de Gennes also find that the superconducting state concentrates along a thin neighborhood of the boundary of the sample and then tails off exponentially in the interior, a property referred to as ''surface superconductivity.'' More recently, Chapman[3,4] carried out a more detailed formal mathematical treatment of the half-plane problem as part of a general analysis of onset for decreasing fields, starting from a perturbation theory developed by Millman and Keller.[5] (See also Refs. 6–8 for related rigorous work and more recently,[9] for rigorous estimates on the upper critical field, above which only normal states solve the Ginzburg–Landau equations.)

Following these studies, Bauman, Phillips, and Tang[10] studied this bifurcation when the cross section of the sample is a disc. By separating variables and using a highly nontrivial O.D.E.

a)Electronic mail: ajb@nwu.edu
b)Electronic mail: sternber@indiana.edu

 © 1998 American Institute of Physics

analysis, they rigorously show that the value of H_{c_3} is raised above the half-plane value of $1.7k$ by a quantity proportional to the curvature of the disc. As in the case of a half-plane, they find the superconducting state concentrates near the boundary of the disc with an exponentially small tail.

It was this last result which caught our attention and lead us to ask the question: what is the nature of the bifurcating superconducting state when the cross section of the sample has an arbitrary shape? Here we investigate this problem using formal asymptotic expansions in the regime where k is large. We discover that the transition occurs at a value of the external field associated with the half-plane at leading order, with a small positive correction proportional to the maximum of curvature (which agrees with the transition problem for the disc), and a smaller correction due to the second derivative of the curvature at the maximum. Furthermore, the amplitude of the order parameter measuring the presence of superconductivity is exponentially localized about the point on the boundary where curvature is a maximum. We note that recently Chapman[11] studied a complementary bifurcation problem in which a small applied magnetic field is gradually raised until it reaches a critical field (H_{c_1}) at which the purely superconducting state loses stability to a vortex state. In that setting it turns out that the bifurcating vortex state emerges near the point of maximum curvature of the boundary as well.

In the section which follows we introduce the Ginzburg–Landau model and then present the asymptotic analysis leading to the conclusions stated above.

II. ASYMPTOTIC ANALYSIS

We consider a sample occupying an infinite cylinder with two-dimensional cross section consisting of an arbitrary simply connected smooth bounded region $\Omega \subset \mathbf{R}^2$. The sample is adjacent to a vacuum and is subjected to an applied magnetic field $\mathbf{H}$ pointing in the axial direction, so that $\mathbf{H}=H\hat{\mathbf{z}}$ for some positive scalar H. We use the Ginzburg–Landau theory to model the superconductor, thus the system is described by two quantities, an order parameter $\Psi:\Omega\to\mathbf{C}$ and the magnetic potential $\mathbf{A}:\Omega\to\mathbf{R}^2$. The pair $(\Psi,\mathbf{A})$ is an equilibrium state for the energy

$$G(\Psi,\mathbf{A})=\int_\Omega\left(\frac{1}{2}\left|\left(\frac{i}{k}\nabla+\mathbf{A}\right)\Psi\right|^2+\frac{1}{4}(1-|\Psi|^2)^2+\frac{1}{2}\left|\text{curl }\mathbf{A}-H\hat{\mathbf{z}}\right|^2\right)dx, \tag{2.1}$$

where curl $\mathbf{A}=(\partial_x A_2-\partial_y A_1)\hat{\mathbf{z}}$, k is the Ginzburg–Landau parameter and we have used the penetration depth as a length scale. This energy is well-defined for Ψ belonging to the Sobolev space $H^1(\Omega;C)$ and $\mathbf{A}$ belonging to the Sobolev space $H^1(\Omega;\mathbf{R}^2)$. Setting the first variation of G to zero, one finds that the pair $(\Psi,\mathbf{A})$ satisfies the Ginzburg–Landau equations

$$\left(\frac{i}{k}\nabla+\mathbf{A}\right)^2\Psi-\Psi+|\Psi|^2\Psi=0 \quad \text{in } \Omega, \tag{2.2}$$

$$(\text{curl})^2\mathbf{A}+\frac{i}{2k}(\Psi^*\nabla\Psi-\Psi\nabla\Psi^*)+|\Psi^2|\mathbf{A}=0 \quad \text{in } \Omega, \tag{2.3}$$

along with the boundary conditions

$$\hat{\mathbf{n}}\cdot\left(\frac{i}{k}\nabla+\mathbf{A}\right)\Psi=0 \quad \text{on } \partial\Omega, \tag{2.4}$$

$$\text{curl}\mathbf{A}=H\hat{\mathbf{z}} \quad \text{or } \partial\Omega. \tag{2.5}$$

Here $\hat{\mathbf{n}}$ is the inward unit normal to $\partial\Omega$ and $(\cdot)^*$ denotes complex conjugation. We refer to Ref. 12 for more background on the model. We also note that in light of our assumption of smoothness on $\partial\Omega$, standard elliptic regularity theory shows that solutions to (2.2)–(2.5) are classical (C^2) solutions.

The problem is invariant under the gauge transformation $(\Psi,\mathbf{A})\to(\Psi',\mathbf{A}')$ where

$$\Psi'\equiv\Psi e^{ik\phi}, \quad \mathbf{A}'\equiv\mathbf{A}+\nabla\phi \tag{2.6}$$

for an arbitrary real-valued $H^2(\Omega)$ function ϕ. Throughout this analysis it will be convenient to impose the additional conditions

$$\text{div } \mathbf{A}=0 \text{ in } \Omega, \quad \mathbf{A}\cdot\hat{\mathbf{n}}=0 \text{ on } \partial\Omega. \tag{2.7}$$

This amounts to choosing a gauge, thus eliminating the degeneracy associated with the gauge invariance of our problem.

Within the Ginzburg–Landau model the normal state corresponds to the conditions

$$\Psi\equiv 0, \quad \mathbf{A}=\overline{\mathbf{A}},$$

where $\overline{\mathbf{A}}$ satisfies

$$(\text{curl})^2\,\overline{\mathbf{A}}=0 \text{ in } \Omega,$$

$$\text{curl } \overline{\mathbf{A}}=H\hat{\mathbf{z}} \text{ on } \partial\Omega,$$

$$\text{div } \overline{\mathbf{A}}=0 \text{ in } \Omega,$$

$$\overline{\mathbf{A}}\cdot\hat{\mathbf{n}}=0 \text{ on } \partial\Omega.$$

For a vector with only two components the condition $(\text{curl})^2\overline{\mathbf{A}}=0$ implies that curl $\overline{\mathbf{A}}$ is constant so the conditions above reduce to simply

$$\text{curl } \overline{\mathbf{A}}=H\hat{\mathbf{z}} \text{ in } \Omega,$$

$$\text{div } \overline{\mathbf{A}}=0 \text{ in } \Omega,$$

$$\overline{\mathbf{A}}\cdot\hat{\mathbf{n}}=0 \text{ on } \partial\Omega.$$

It is convenient to introduce the vector field $\overline{\mathbf{a}}$ through the relation $\overline{\mathbf{A}}=H\overline{\mathbf{a}}$, so that

$$\text{curl } \overline{\mathbf{a}}=\hat{\mathbf{z}} \text{ in } \Omega,$$

$$\text{div } \overline{\mathbf{a}}=0 \text{ in } \Omega,$$

$$\overline{\mathbf{a}}\cdot\hat{\mathbf{n}}=0 \text{ on } \partial\Omega. \tag{2.8}$$

We note that obtaining the existence of a classical solution to (2.8) in a smooth bounded domain is simple. For example, one can obtain $\overline{\mathbf{a}}$ in the form $\overline{\mathbf{a}}=\frac{1}{2}(-y,x)+\nabla u$, where u is a harmonic function satisfying the Neumann condition $\nabla u\cdot\hat{\mathbf{n}}=-\frac{1}{2}(-y,x)\cdot\hat{\mathbf{n}}$ on $\partial\Omega$. The uniqueness of $\overline{\mathbf{a}}$ follows from the assumption of simple connectivity on Ω. That is, letting say $\mathbf{b}$ denote the difference between any two solutions to (2.8), one finds that $\mathbf{b}$ must be the gradient of a harmonic scalar function whose normal derivative vanishes on $\partial\Omega$. Thus the scalar function is constant and $\mathbf{b}\equiv 0$.

Since we seek a description of the initial bifurcation of the solution from the normal state $(0,H\overline{\mathbf{a}})$ as the applied magnetic field is decreased, we begin our formal analysis by introducing a small parameter ϵ (which measures the magnitude of $|\Psi|$) and assuming asymptotic expansions for the quantities Ψ, $\mathbf{A}$, and H of the form

$$\Psi=0+\epsilon\Psi^{(1)}+\cdots,$$

$$\mathbf{A}=h\overline{\mathbf{a}}+\epsilon^2\mathbf{A}^{(2)}+\cdots,$$

$$H=h+\epsilon^2H^{(1)}+\cdots$$

(cf. Refs. 3 and 5). The relative scalings of Ψ, $\mathbf{A}$, and H are motivated by the need to balance the term $|\Psi|^2\Psi$ with the terms involving $\mathbf{A}-H\overline{\mathbf{a}}$ in (2.2). Inserting these expansions into (2.2)–(2.5) yields a linear problem for $\Psi^{(1)}$ at leading order,

$$\left(\frac{i}{k}\nabla+h\overline{\mathbf{a}}\right)^2\Psi^{(1)}-\Psi^{(1)}=0 \quad \text{in} \quad \Omega, \tag{2.9}$$

$$\hat{\mathbf{n}}\cdot\left(\frac{i}{k}\nabla+h\overline{\mathbf{a}}\right)\Psi^{(1)}=0 \quad \text{on} \quad \partial\Omega. \tag{2.10}$$

As h is decreased, one expects the instability of the normal state to occur when (2.9)–(2.10) has a nontrivial solution. Here, h plays the role of an eigenvalue and we define H_{c_3} to be the largest positive eigenvalue for the linearized problem (2.9)–(2.10). When k is sufficiently large, nonlinear analysis suggests that this bifurcation is a supercritical pitchfork (cf. Refs. 3 and 4).

At this point, we wish to analyze (2.9)–(2.10) via formal matched asymptotic expansions in the regime where the Ginzburg–Landau parameter k is large. Experiments (both physical and numerical) indicate that the superconducting state nucleates on the boundary of Ω and the factor of $(1/k)$ in (2.9) suggests that the order parameter $\Psi^{(1)}$ will vary there with a characteristic length scale $1/k$. Thus, we will seek a boundary layer solution which matches to the normal state in the interior of the domain. (Indeed, we will find that the modulus of this boundary layer solution decays exponentially to zero as we depart from $\partial\Omega$.)

To this end, we introduce coordinates (s,η) valid in a neighborhood of $\partial\Omega$. Here s is arclength along $\partial\Omega$ and η is the distance from a point within Ω to the boundary. We can arbitrarily select a point on $\partial\Omega$ corresponding to $s=0$; without loss of generality we will choose this point to coincide with the maximum of curvature on the boundary. We will use $\kappa=\kappa(s)$ to denote the curvature of $\partial\Omega$ (not to be confused with the Ginzburg–Landau parameter k!).

This local coordinate system will be well-defined in the rectangle

$$S\equiv\{(s,\eta):-L/2<s<L/2,\ 0<\eta<C\},$$

where L denotes the arclength of the boundary, and C is any positive constant smaller than $[1/\kappa(0)]$. We adopt the convention that for Ω a disc κ is positive. We denote by $\hat{\mathbf{t}}=\hat{\mathbf{t}}(s)$ a unit tangent vector to $\partial\Omega$ and we recall that $\hat{\mathbf{n}}=\hat{\mathbf{n}}(s)$ denotes the inner unit normal vector. Thus, in particular, any vector field $\mathbf{F}$ defined in a neighborhood of $\partial\Omega$ can be expressed as $\mathbf{F}(s,\eta)=F_1(s,\eta)\hat{\mathbf{t}}(s)+F_2(s,\eta)\hat{\mathbf{n}}(s)$. We now proceed to rewrite problem (2.9)–(2.10) in $s-\eta$ coordinates so that we may pursue a boundary layer construction in the region S.

For this purpose, we record here that for any scalar-valued function $f=f(s,\eta)$ one has the identities,

$$\nabla f=\partial_\eta f\hat{\mathbf{n}}+\frac{1}{1-\kappa\eta}\,\partial_s f\hat{\mathbf{t}}, \tag{2.11}$$

$$\nabla^2 f=\partial_{\eta\eta}f-\frac{\kappa}{1-\kappa\eta}\,\partial_\eta f+\frac{1}{1-\kappa\eta}\,\partial_s\left(\frac{1}{1-\kappa\eta}\right)\partial_s f, \tag{2.12}$$

and for any vector field $\mathbf{F}=F_1(s,\eta)\hat{\mathbf{t}}+F_2(s,\eta)\hat{\mathbf{n}}$ we have

$$\text{div}\ \mathbf{F}=\frac{1}{1-\kappa\eta}\,\partial_s F_1+\frac{1}{1-\kappa\eta}\,\partial_\eta[(1-\kappa\eta)F_2], \tag{2.13}$$

$$\text{curl}\ \mathbf{F}=\text{div}(F_2\hat{\mathbf{t}}-F_1\hat{\mathbf{n}})\hat{\mathbf{z}}. \tag{2.14}$$

These identities can be computed using standard planar differential geometry (see e.g., Ref. 13) and are explicitly derived for example in Ref. 14.

For a general domain Ω, although $\overline{\mathbf{a}}\cdot\hat{\mathbf{n}}=0$ on $\partial\Omega$, $\overline{\mathbf{a}}$ is not purely tangential. That is, writing $\overline{\mathbf{a}}=\overline{\mathbf{a}}(s,\eta)$ we have $\overline{\mathbf{a}}(s,\eta)\cdot\hat{\mathbf{n}}(s)=0$ only for $\eta=0$ but not for $0<\eta<C$. However in the coordinate rectangle S, we can decompose $\overline{\mathbf{a}}$ into components $\mathbf{p}$ and $\mathbf{q}$,

$$\overline{\mathbf{a}}=\mathbf{p}+\mathbf{q}$$

as follows. We first define $\mathbf{q}$ as the solution to the problem

$$\text{curl } \mathbf{q}=\hat{\mathbf{z}} \text{ in } S, \quad \mathbf{q}\cdot\hat{\mathbf{n}}=0 \text{ in } S, \quad \text{and } \mathbf{q}(s,0)=0. \tag{2.15}$$

Note in particular that $\mathbf{q}$ will be purely tangential is the sense mentioned above.

Seeking a solution to (2.15) in the form $\mathbf{q}=q(s,\eta)\hat{\mathbf{t}}$, we can solve for the scalar q through the use of (2.13) and (2.14). We find that q must satisfy the first order differential equation

$$\frac{-1}{1-\kappa\eta}\,\partial_\eta((1-\kappa\eta)q)=1, \quad q(s,0)=0,$$

so that

$$\mathbf{q}(s,\eta)=q(s,\eta)\hat{\mathbf{t}}(s)=-\eta\left(\frac{1-\kappa(s)\,\eta/2}{1-\kappa(s)\,\eta}\right)\hat{\mathbf{t}}(s). \tag{2.16}$$

We then define $\mathbf{p}$ by the relation $\overline{\mathbf{a}}=\mathbf{p}+\mathbf{q}$. Note in particular that $\mathbf{p}$ will then be conservative,

$$\text{curl } \mathbf{p}=0 \text{ in } S. \tag{2.17}$$

Motivated by the gauge invariance, and utilizing (2.17), we now introduce a phase Φ on the rectangle S through the relation $\nabla\Phi=h\mathbf{p}$. Hence, for any $(s,\eta)\in S$, we let

$$\Phi(s,\eta)=h\int_\gamma \mathbf{p}\cdot d\mathbf{r}, \tag{2.18}$$

where γ is any path in S joining (0,0) to (s,η). We then introduce the function ψ by writing the solution $\Psi^{(1)}$ to (2.9)–(2.10) near $\partial\Omega$ in the form

$$\Psi^{(1)}=\psi e^{ik\Phi}e^{ik\beta s}, \tag{2.19}$$

where β is a constant to be determined. Through the gauge transformation it becomes apparent that the complex exponential of Φ eliminates $\mathbf{p}$, the conservative component of $\overline{\mathbf{a}}$. The factor $e^{ik\beta s}$ represents a contribution to the phase evident in the earlier investigations of Saint-James and de Gennes[2] and Chapman[3] that arises when the problem is posed on the half-plane. Our motivation for this change of variables is to make the leading-order balance conform with the known results for the half-plane problem. In particular this is not a polar decomposition; although ψ will have real solutions at leading order, the higher-order corrections may be complex in nature.

Using (2.6), we find that in a neighborhood of $\partial\Omega$, the problem (2.9)–(2.10) for $\Psi^{(1)}$ as a function of x transforms to the following problem for ψ as a function of s and η,

$$\left(\frac{i}{k}\,\boldsymbol{\nabla}+h\mathbf{q}-\beta\boldsymbol{\nabla}s\right)^2\psi-\psi=0 \text{ in } S, \tag{2.20}$$

$$\hat{\mathbf{n}}\cdot\left(\frac{i}{k}\,\nabla+h\mathbf{q}-\beta\nabla s\right)\psi=0 \text{ on } \eta=0. \tag{2.21}$$

Then using (2.11) to compute ∇s and exploiting the fact that $h\mathbf{q}-\beta\nabla s$ has no $\hat{\mathbf{n}}$ component, we see that (2.20)–(2.21) can be written as

$$\left(\frac{i}{k}\,\nabla+\left(hq-\frac{\beta}{1-\kappa\eta}\right)\hat{\mathbf{t}}\right)^2\psi-\psi=0 \text{ in } S,$$

$$\nabla\psi\cdot\hat{\mathbf{n}}=0 \text{ on } \eta=0.$$

A further expansion of this equation yields

$$\left(\frac{1}{k^2}\,\Delta+\frac{i}{k}\,\frac{2V}{1-\kappa\eta}\,\partial_s+\frac{i}{k}\,\frac{V_s}{1-\kappa\eta}+1-V^2\right)\psi=0 \quad \text{in} \quad S, \tag{2.22}$$

$$\psi_\eta=0 \quad \text{on} \quad \eta=0, \tag{2.23}$$

where

$$V(\eta,s)=\frac{h\,\eta(1-\kappa\eta/2)+\beta}{1-\kappa\eta}. \tag{2.24}$$

So far we have simply rewritten the linearized problem (2.9)–(2.10) in terms of the function ψ (introduced via the changes of variables (2.19)) and in terms of the independent variables s and η to obtain a convenient formulation of the problem valid near $\partial\Omega$. At this point, we wish to begin a process of formal matched asymptotic expansion (based on the assumption $k\gg 1$) by introducing stretched coordinates to solve a boundary layer problem which will then be matched to the normal state in the interior.

As we are seeking a boundary layer solution and we have a factor of $(1/k^2)$ multiplying the Laplacian in (2.22) it is natural to seek a solution which depends not on the normal variable η but on the stretched variable

$$\xi=k\,\eta. \tag{2.25}$$

What is not so obvious, but nonetheless necessary, is that we must stretch the tangential variable as well, by introducing

$$\tau=k^{1/4}(s-s_0). \tag{2.26}$$

Here s_0 is a value of arc length along the boundary which remains to be determined. The transverse length scale $(1/k^{1/4})$ represents a balance between the quadratic nature of the curvature at the maximum and the transverse modulation of the amplitude, a fact best explained *a posteriori*. Putting it another way, any expansion which fails to incorporate this stretched tangential dependence—that is, which fails to sufficiently promote the appearance of tangential derivatives of the solution as we expand—will lead to inconsistencies later in the asymptotic analysis.

Once one resolves to seek a solution depending on ξ and τ, however, one is obligated to pursue a solution which is not only localized near $\partial\Omega$, but localized near the (so far arbitrary) point corresponding to $s=s_0$ as well. Accordingly, this solution must be matched to the normal state both in the interior of Ω (ξ large) *and along the boundary away from* $s=s_0$ ($|\tau|$ large).

Thus, we assume an expansion for ψ of the form

$$\psi=\psi_0(\xi,\tau)+\frac{1}{k^{1/4}}\,\psi_1(\xi,\tau)+\frac{1}{k^{1/2}}\,\psi_2(\xi,\tau)+\frac{1}{k^{3/4}}\,\psi_3(\xi,\tau)+\cdots \tag{2.27}$$

and an expansion for $h=h(k)$ of the form

$$h=h_0k+h_1k^{3/4}+h_2k^{1/2}+h_3k^{1/4}+h_4+\cdots\ . \tag{2.28}$$

To carry out the analysis, we need an expansion for the expression V in powers of k. To this end, we first express the curvature κ as a function of τ, say $K(\tau)$, and Taylor expand about $s=s_0$ using (2.26) to obtain

$$K(\tau)=\kappa(s_0)+\frac{1}{k^{1/4}}\,\kappa_s(s_0)\,\tau+\frac{1}{k^{1/2}}\,\kappa_{ss}(s_0)\,\frac{\tau^2}{2}+\mathcal{O}\left(\frac{1}{k^{3/4}}\right)=\overline{\kappa}+\frac{1}{k^{1/4}}\,\overline{\kappa_s}\,\tau+\frac{1}{k^{1/2}}\,\overline{\kappa_{ss}}\,\frac{\tau^2}{2}+\mathcal{O}\left(\frac{1}{k^{3/4}}\right), \tag{2.29}$$

where $\overline{\kappa}$, $\overline{\kappa_s}$, and $\overline{\kappa_{ss}}$ denote $\kappa(s_0)$, $\kappa_s(s_0)$, and $\kappa_{ss}(s_0)$, respectively.

To expand V, we substitute (2.25) and (2.29) in (2.24) to find

$$V=(h_0\xi+\beta)+\frac{1}{k^{1/4}}(h_1\xi)+\frac{1}{k^{1/2}}(h_2\xi)+\frac{1}{k^{3/4}}(h_3\xi)$$
$$+\frac{1}{k}\left\{\left(h_4\xi+\left(\frac{h_0}{2}\xi^2+\beta\xi\right)\overline{\kappa}\right\}+\frac{1}{k^{5/4}}\left\{\left(h_5\xi+\frac{h_1}{2}\xi^2\overline{\kappa}+\left(\frac{h_0}{2}\xi^2+\beta\xi\right)\overline{\kappa_s}\tau\right\}\right.$$
$$+\frac{1}{k^{3/2}}\left\{\left(h_6\xi+\frac{h_2}{2}\xi^2\overline{\kappa}+\frac{h_1}{2}\xi^2\overline{\kappa_s}\tau+\left(\frac{h_0}{2}\xi^2+\beta\xi\right)\overline{\kappa_{ss}}\frac{\tau^2}{2}\right\}+\mathcal{O}\left(\frac{1}{k^{7/4}}\right).\right. \tag{2.30}$$

Using (2.12) and (2.27), we expand the Laplacian

$$\frac{1}{k^2}\Delta=\partial^2_{\xi\xi}-\frac{1}{k}\overline{\kappa}\partial_\xi-\frac{1}{k^{5/4}}\overline{\kappa_s}\tau\partial_\xi+\frac{1}{k^{3/2}}\left\{\partial_{\tau\tau}-\overline{\kappa_{ss}}\frac{\tau^2}{2}\partial_\xi\right\}+\mathcal{O}\left(\frac{1}{k^{7/4}}\right). \tag{2.31}$$

The imaginary terms in (2.22) can be expanded using (2.13), (2.25) and (2.26)

$$\frac{i}{k}\frac{2V}{1-\kappa\eta}\partial_s+\frac{i}{k}\frac{V_s}{1-\kappa\eta}=\frac{i}{k^{3/4}}2V\partial_\tau+\frac{i}{k^{3/4}}V_\tau+\mathcal{O}\left(\frac{1}{k^{7/4}}\right). \tag{2.32}$$

We can now proceed to expand and solve for successive orders in the expansion for ψ.

A. Leading order balance: The half-plane problem

The basis for our expansion is that the leading-order balance in the problem is exactly equivalent to the instability problem for the half-plane. This problem was first considered by Saint-James and de Gennes,[2] and yields an exponentially localized superconducting region near the boundary. In the general domain we will construct a solution whose leading order structure is equivalent to the solution of the half-plane problem with an amplitude that varies along the boundary.

We can now expand the problem at leading order ($\mathcal{O}(1)$); Eqs. (2.22) and (2.23) yield

$$\psi_{0_{\xi\xi}}+(1-(h_0\xi+\beta)^2)\psi_0=0 \quad \text{for} \quad 0<\xi<\infty,$$
$$\psi_{0_\xi}(0)=0, \quad \psi_{0_\xi}(\infty)=0, \tag{2.33}$$

where the second boundary condition allows us to match the boundary layer solution to the normal state in the interior of Ω. This is exactly the half-plane problem considered formally in Refs. 2 and 3, and treated rigorously in Refs. 6 and 7.

Recall that we seek the first appearance of a nonzero state as the magnetic field is decreased. Hence we view (2.33) as a double eigenvalue problem for h_0 and β in which β is to be determined so as to yield $\underline{h}$, the largest possible value of h_0. Note that β corresponds to a transverse (i.e., parallel to the boundary) wave number, which in the half-plane problem arises naturally through separation of variables. We now tabulate a set of results known for this problem, which will form the basis for our perturbation expansion. With the exception of identities (2.34)–(2.36) below, the proof of the Lemma can be found, for example, in Ref. 15. See also Refs. 6 and 7. The derivation of (2.34)–(2.36) can be found in the appendix.

Lemma: There exists a largest value of $h_0(\beta)$, denoted by $\underline{h}$, for which (2.33) possesses a nontrivial solution. This occurs when β takes the value -1. The corresponding eigenfunction, denoted henceforth by $\underline{\psi}(\xi)$, satisfies

$$\mathcal{L}\underline{\psi}=0 \quad \text{for} \quad 0<\xi<\infty,$$
$$\underline{\psi}_\xi(0)=0, \quad \underline{\psi}_\xi(\infty)=0,$$

with the normalization $\underline{\psi}(0)=1$.

The operator $\mathcal{L}$ is defined by

$$\mathcal{L}(\cdot)\equiv\partial_{\xi\xi}(\cdot)+(1-(\underline{h}\xi-1)^2)(\cdot),$$

and is self-adjoint in the L^2 inner-product.

The solution ψ decays exponentially as $\xi\to\infty$ and, letting $J_0\equiv\int_0^\infty \overline{\psi^2}\,d\xi$, we have that $\overline{\psi}$ satisfies the following identities:

$$J_1\equiv\int_0^\infty \xi\overline{\psi^2}\,d\xi=\frac{1}{\overline{h}}J_0, \tag{2.34}$$

$$J_2\equiv\int_0^\infty \xi^2\overline{\psi^2}\,d\xi=\frac{3}{2\overline{h^2}}\,J_0, \tag{2.35}$$

$$J_3\equiv\int_0^\infty \xi^3\overline{\psi^2}\,d\xi=\frac{1}{6\overline{h^2}}+\frac{5}{2\overline{h^3}}\,J_0. \tag{2.36}$$

Remark: Through numerical approximation one finds $\overline{h}=1.69461+$ and $J_0=1.00784+$.

In light of the absence of any τ dependence in (2.33), we may write

$$\psi_0(\xi,\tau)=W_0(\tau)\overline{\psi}(\xi),$$

where $W_0(\tau)$ is a (perhaps complex) amplitude corresponding to a transverse modulation of $\overline{\psi}(\xi)$ on the τ length scale whose behavior will be determined at higher order. This type of modulation or multiple scale expansion is familiar from many physical systems.[16]

B. Higher order corrections

At orders $\mathcal{O}(1/k^{1/4})$ and $\mathcal{O}(1/k^{1/2})$ we find

$$\mathcal{L}(\psi_i)=2h_i\xi(\overline{h}\xi-1)\overline{\psi}(\xi)W_0(\tau)\quad\text{for}\quad i=1,2,$$

$$(\psi_i)_\xi(0)=0,\quad (\psi_i)_\xi(\infty)=0. \tag{2.37}$$

If we define the L^2 inner-product,

$$\langle\psi,\phi\rangle=\int_0^\infty \psi^*\phi\,d\xi,$$

then the Fredholm Alternative Theorem yields necessary and sufficient solvability conditions for (2.37). That is, orthogonality to $\overline{\psi}$,

$$\langle\overline{\psi},\mathcal{L}(\psi_i)\rangle=0\quad\text{for}\quad i=1,2,$$

implies

$$h_iW_0\,\frac{J_0}{\overline{h}}=0.$$

Thus we conclude $h_1=h_2=0$ so that $\mathcal{L}(\psi_i)=0$ and we have

$$\psi_1=W_1(\tau)\overline{\psi}(\xi)\quad\text{and}\quad\psi_2=W_2(\tau)\overline{\psi}(\xi)$$

with the functions W_1 and W_2 yet to be determined.

Proceeding to $\mathcal{O}(1/k^{3/4})$,

$$\mathcal{L}(\psi_3)=2h_3\xi(\overline{h}\xi-1)\overline{\psi}W_0-2i(\overline{h}\xi-1)\overline{\psi}W_{0\tau}\quad\text{for}\quad 0<\xi<\infty,$$

$$\psi_{3_\xi}(0)=0,\quad \psi_{3_\xi}(\infty)=0. \tag{2.38}$$

Application of the solvability condition at this order yields

$$\langle \overline{\psi}, \mathscr{L}(\psi_3)\rangle = 0 = h_3 W_0 \frac{J_0}{\overline{h}},$$

hence $h_3 = 0$. Now note that since $\partial_\xi(\mathscr{L}(\overline{\psi})) = 0$, one has

$$\mathscr{L}(\overline{\psi}_\xi) = 2\overline{h}(\overline{h}\xi - 1)\overline{\psi}. \tag{2.39}$$

One also readily verifies that $\overline{\psi}_{\xi\xi}(0) = \overline{\psi}_{\xi\xi}(\infty) = 0$ (recall that in fact $\overline{\psi}$ decays exponentially at infinity) so that we find

$$\psi_1 = W_3(\tau)\overline{\psi}(\xi) - \frac{i}{\overline{h}}\overline{\psi}_\xi W_{0_\tau}, \tag{2.40}$$

where W_3 remains to be determined.

We now proceed to order $\mathscr{O}(1/k)$,

$$\mathscr{L}(\psi_4) = h_4[2\xi(\overline{h}\xi - 1)\overline{\psi}]W_0 + \overline{\kappa}[\xi(\overline{h}\xi - 1)(\overline{h}\xi - 2)\overline{\psi} + \overline{\psi}_\xi]W_0 - [2i(\overline{h}\xi - 1)\overline{\psi}]W_{1_\tau},$$

$$\psi_{4_\xi}(0) = 0, \quad \psi_{4_\xi}(\infty) = 0. \tag{2.41}$$

Solvability at this order takes the form,

$$\left(h_4 \frac{J_0}{\overline{h}} - \frac{\overline{\kappa}}{3}\right) W_0 = 0, \tag{2.42}$$

where we have applied the identities (2.34)–(2.35), and integration by parts. Thus

$$h_4 = \frac{\overline{h}}{3J_0}\overline{\kappa} = \frac{\overline{h}}{3J_0}\kappa(s_0). \tag{2.43}$$

Note that we seek the largest value of h (hence of h_4) at which a superconducting solution emerges; clearly we can maximize h_4 by choosing to expand around the point where $\kappa(s_0)$ is a maximum, i.e., by choosing $s_0 = 0$. Consequently, we conclude that *instability first occurs at the point of maximum curvature*, so that

$$\overline{\kappa} = \max_{s_0} \kappa(s_0) = \kappa(0),$$

$$\overline{\kappa}_s = 0,$$

$$\overline{\kappa}_{ss} \leqslant 0.$$

We shall make the (generic) assumption that

$$\overline{\kappa}_{ss} < 0.$$

We have also tacitly assumed that the curvature reaches a unique maximum; in the case of multiple isolated maximums (such as an ellipse), we could construct an exponentially localized solution at each which would interact only exponentially weakly. In the case when the second derivative vanishes at the maximum (such as the disc), our scaling ansatz will prove to be invalid at higher order.

In addition, formula (2.43)—specifically the fact that h_4 is positive—bears the information that onset of superconductivity in a bounded sample occurs at a higher value of applied field than for the case where Ω is a half-space. This correction agrees with that computed by Bauman *et al.*[10] for the disc.

We could proceed to solve for ψ_4 at this point (an explicit solution can be obtained via variation of parameters), however we omit this calculation as it will play no further role in our analysis.

We proceed to the next order and collect terms of $\mathcal{O}(1/k^{5/4})$,

$$\mathcal{L}(\psi_5) = h_5[2\xi(\overline{h}\xi - 1)\overline{\psi}]W_0 + h_4[2\xi(\overline{h}\xi - 1)\overline{\psi}]W_1 + \overline{\kappa}[\xi(\overline{h}\xi - 1)(\overline{h}\xi - 2)\overline{\psi} + \overline{\psi}_\xi]W_1 - [2i(\overline{h}\xi - 1)\overline{\psi}]W_{2_\tau},$$

$$\psi_{5_\xi}(0) = 0, \quad \psi_{5_\xi}(\infty) = 0. \tag{2.44}$$

Once again, integration against $\overline{\psi}$ yields the solvability condition. The terms in W_1 and W_{2_τ} in the solvability condition vanish in an analogous fashion to those in W_0 and W_{1_τ} at the previous order (taking into account the value of h_4 determined by (2.43)). Hence, $h_5 = 0$. Once again, we neglect the calculation of ψ_5 as it plays no further role in our analysis.

Finally, we proceed to the next order and collect terms of $\mathcal{O}(1/k^{3/2})$,

$$\mathcal{L}(\psi_6) = h_6[2\xi(\overline{h}\xi - 1)\overline{\psi}]W_0 + \overline{\kappa}_{ss}\frac{\tau^2}{2}[\xi(\overline{h}\xi - 1)(\overline{h}\xi - 2)\overline{\psi} + \overline{\psi}_\xi]W_0 - \left[\frac{2}{\overline{h}}(\overline{h}\xi - 1)\overline{\psi}_\xi + \overline{\psi}\right]W_{0_{\tau\tau}} + h_4[2\xi(\overline{h}\xi - 1)\overline{\psi}]W_2 + \overline{\kappa}[\xi(\overline{h}\xi - 1)(\overline{h}\xi - 2)\overline{\psi} + \overline{\psi}_\xi]W_2 - [2i(\overline{h}\xi - 1)\overline{\psi}]W_{3_\tau},$$

$$\psi_{6_\xi}(0) = 0, \quad \psi_{6_\xi}(\infty) = 0. \tag{2.45}$$

Once again, we integrate against $\overline{\psi}$. The terms proportional to W_2 and W_{3_τ} in the solvability condition vanish in an analogous fashion to those proportional to W_1 and W_{2_τ} respectively at the previous order. However, we obtain a nontrivial solvability condition for $W_0(\tau)$. Applying the identities (2.34)–(2.36) and integrating by parts yields

$$-\frac{1}{\overline{h}}W_{0_{\tau\tau}} + \left(h_6\frac{J_0}{\overline{h}} - \frac{\overline{\kappa}_{ss}}{6}\tau^2\right)W_0 = 0. \tag{2.46}$$

At this point we see the motivation for the introduction of the τ lengthscale; on this length scale the tranverse variation of the amplitude ($W_{0_{\tau\tau}}$) balances the quadratic variation of the curvature ($\overline{\kappa}_{ss}\tau^2/2$).

Boundary conditions on $W_0(\tau)$ are determined by a need to match to the normal state at large $|\tau|$, for which $\Psi^{(1)}$ and consequently W_0 vanish. Thus, we apply the condition that $W_0(\tau)$ vanishes in the far-field,

$$W_0(\tau) \to 0 \quad \text{as} \quad |\tau| \to \infty \tag{2.47}$$

with the understanding that τ is in the matching region, $1 \ll |\tau| \ll k^{1/4}$. Fortunately, $W_0(\tau)$ decays exponentially and consequently this matching will only introduce exponentially small corrections in the expansion.

Fortuitously, Eq. (2.46) is real, and consequently admits real solutions for W_0 (we expect W_n to be complex in general). We can rescale (2.46) into the standard form for the harmonic oscillator equation; define

$$\zeta = \alpha\tau, \quad w(\zeta) = W_0(\zeta/\alpha), \quad \lambda = -\frac{J_0 h_6}{\alpha^2},$$

where

$$\alpha = \left(\frac{-2\overline{\kappa}_{ss}\overline{h}}{3}\right)^{1/4},$$

remembering that at the maximum of curvature, we have $\overline{\kappa}_{ss}<0$. Equations (2.46) and (2.47) become

$$w_{\zeta\zeta}+\left(\lambda-\frac{1}{4}\,\zeta^2\right)w=0, \quad w(\zeta)\to 0 \quad \text{as} \quad |\zeta|\to\infty.$$

The eigenvalues and eigenfunctions of this problem are given by

$$\lambda_n=n+\tfrac{1}{2}, \quad w_n(\zeta)=\mathrm{He}_n(\zeta)e^{-\zeta^2/4}, \quad n=0,1,2,3,...,$$

where $\mathrm{He}_n(\zeta)$ are the Hermite polynomials (cf. Ref. 17). Note that the smallest eigenvalue is $\lambda_0=1/2$, with corresponding eigenfunction $w_0(\zeta)=e^{-\zeta^2/4}$. Hence, the largest (least negative) eigenvalue for (2.46) is given by

$$h_6=-\frac{1}{2J_0}\left(\frac{-2\overline{\kappa}_{ss}\overline{h}}{3}\right)^{1/2}, \tag{2.48}$$

and the corresponding eigenfunction is

$$W_0(\tau)=e^{-\alpha^2\tau^2/4}. \tag{2.49}$$

The exponential decay of W_0 away from $\tau=0$ reveals that the phenomenon of onset is highly localized near the point of maximum curvature for $\partial\Omega$. We have now completely determined the leading order solution for $\Psi^{(1)}$ to the linearized problem (2.9), (2.10) in the asymptotic regime $\kappa\gg 1$.

III. DISCUSSION

In this paper we have examined the instability of the normal state to superconductivity with decreasing magnetic field for a closed smooth cylindrical region subject to a vertical magnetic field. We have examined the problem asymptotically in the boundary layer limit ($k\gg 1$). Instability first occurs in a region exponentially localized near the point of maximum curvature on the boundary. The transition occurs at a value of the magnetic field,

$$H_{c_3}=\overline{h}k+\frac{\overline{h}}{3J_0}\,\overline{\kappa}-\frac{1}{2J_0}\left(\frac{-2\overline{\kappa}_{ss}\overline{h}}{3}\right)^{1/2}\frac{1}{k^{1/2}}+\mathcal{O}\left(\frac{1}{k}\right), \tag{3.1}$$

where $\overline{h}k$ is the transition value associated with the half-plane (cf. Refs. 2 and 3), the second term is a positive constant times the maximum curvature ($\overline{\kappa}$) that agrees with the correction computed for the disc (cf. Ref. 10), the third term is a negative correction of order $k^{-1/2}$, associated with the quadratic variation of the curvature at the maximum, and the error term follows from continuing the expansion documented in the previous section.

At leading order, the superconducting order-parameter satisfies

$$|\Psi^{(1)}|\sim\overline{\psi}(\xi)e^{-\alpha^2\tau^2/4},$$

where

$$\xi=k\eta, \quad \tau=k^{1/4}s, \quad \text{and} \quad \alpha=\left(\frac{-2\overline{\kappa}_{ss}\overline{h}}{3}\right)^{1/4}.$$

Here $\overline{\psi}$ is the profile associated with transition in the half-plane problem—it is an exponentially decaying function of the normal distance to the boundary (η) scaled by the boundary layer thickness ($1/k$). The amplitude along the boundary is modulated by a Gaussian, $e^{-\alpha^2\tau^2/4}$, where τ

is the arclength measured from the point of maximum curvature (at $s=0$) scaled by a length scale intermediate between the boundary layer thickness and the characteristic length scale of the domain (unity).

Physically, this result is a refinement of the corresponding results for the half-plane and the disc. The half-plane result yields the leading order transition value for the magnetic field and exponentially decaying boundary layer profile, kh and ψ, respectively. The disc calculation shows that transition occurs at a higher value of the magnetic field where the correction is of order unity and proportional to the curvature (i.e., the inverse of the radius). In contrast to the half-plane and disc calculation where onset occurs along the entire boundary simultaneously, in a general domain the profile is exponentially localized near the point of maximum curvature on the boundary on the τ length scale; the transition value is slightly below that of the oscillating disc at the point of maximum curvature.

An additional effect of this exponential localization of the transition profile is that the periodicity of the boundary apparently has an exponentially small effect on the critical magnetic field and the various eigenmodes are never degenerate. This is in sharp contrast to the theoretical calculation for the disc where periodicity ''quantizes'' the transverse wave numbers and at various critical parameter values there is a transition between different modes as the most unstable. We note that this quantization in a disc is captured via formal matched asymptotics in Ref. 18. These characteristics are also observed experimentally.[19,20] We suggest here that this degeneracy occurs only in the presence of symmetry in the sample geometry.

Finally, we note that the approach here is quite general, and should allow for the incorporation of nonlinearity, temporal evolution through relaxational effects, etc. The result we believe will be a Newell–Whitehead–Segal-type equation (cf. Ref. 16),

$$W_T = W_{SS} + [h - \kappa(S)]W - W|W|^2,$$

where W is a transverse modulation amplitude of the boundary layer profile, S is a scaled arc length, T is a slow time scale associated with relaxational effects (cf. Ref. 4), h is a scaled magnetic field, and $\kappa(S)$ measures curvature variation along the boundary. A model of this sort could be used to study hysteresis, symmetry breaking and a host of other phenomena which may lead to better understanding and prediction of the behavior of superconductors.

Note added in proof: After completing our work we learned about Reference 21.

ACKNOWLEDGMENTS

The authors gratefully acknowledge fruitful discussions with Joe Keller, John Neu, and Jacob Rubinstein on the topic of this paper. A.J.B. wishes to thank the MSRI and P.S. wishes to thank the Department of Mathematics at Stanford University for the hospitality extended during visits covering the period in which this research was conducted.

The work of P.S. was supported by the National Science Foundation under Grants Nos. DMS-9322617 and DMS-9705774 and by a U.S.–Israel Bi-National Grant.

APPENDIX: DERIVATION OF MOMENT IDENTITIES (2.34)–(2.36)

To derive identities (2.34)–(2.36), we observe that the operator $\mathscr{L}$ maps the vector space of functions of the form $a(\xi)\psi + b(\xi)\psi_\xi$, where $a(\xi)$ and $b(\xi)$ are polynomials, into itself. A little algebra shows that for a polynomial $c(\xi)$,

$$\mathscr{L}(2c\bar{\psi}_\xi - c_\xi\bar{\psi}) = -\{c_{\xi\xi\xi} + 4\bar{h}\xi(2-\bar{h}\xi)c_\xi + 4\bar{h}(1-\bar{h}\xi)c\}\bar{\psi}. \tag{A1}$$

Multiplying by $\bar{\psi}$, integrating over the range and applying integration by parts, we reach the identity

$$\int_0^\infty \bar{\psi}\mathscr{L}(2c\bar{\psi}_\xi - c_\xi\bar{\psi})\,d\xi = c_{\xi\xi}(0). \tag{A2}$$

If we now define the moments,

$$J_n \equiv \int_0^\infty \xi^n \overline{\psi^2} d\xi,$$

we can derive a recursion relationship for J_n by choosing $c=\xi^n$, substituting into (A1) and applying the identity (A2)

$$c=1: \quad 0=-4\overline{h}J_0+4\overline{h^2}J_1,$$

$$c=\xi: \quad 0=-12\overline{h}J_1+8\overline{h^2}J_2,$$

$$c=\xi^2: \quad 2=-20\overline{h}J_2+12\overline{h^2}J_3,$$

$$c=\xi^n: \quad 0=-n(n-1)(n-2)J_{n-3}-4\overline{h}(2n+1)J_n+4(n+1)\overline{h^2}J_{n+1} \quad n\geqslant 3.$$

The identities (2.34)–(2.36) and expressions for J_n in terms of J_0 can now be found by solving these relations. ■

[1] V. L. Ginzburg and L. D. Landau, ''On the theory of superconductivity,'' JETP **20**, 1064–1082 (1950).
[2] D. Saint-James and P. G. de Gennes, ''Onset of superconductivity in decreasing fields,'' Phys. Lett. **7**, 306–308 (1963).
[3] S. J. Chapman, ''Nucleation of superconductivity in decreasing fields, I,'' Eur. J. Appl. Math. **5**, 449–468 (1994).
[4] S. J. Chapman, ''Nucleation of superconductivity in decreasing fields, II,'' Eur. J. Appl. Math. **5**, 469–494 (1994).
[5] M. H. Millman and J. B. Keller, ''Perturbation theory of nonlinear boundary value problems,'' J. Math. Phys. **10**, 342–361 (1969).
[6] C. Bolley, ''Modélisation de champ de retard à la condensation d'un supraconducteur par un problème de bifurcation,'' M^2 AN **26**, 235–287 (1992).
[7] C. Bolley and B. Helffer ''An application of semi-classical analysis to the asymptotic study of the supercooling field of a superconducting material,'' Ann. Inst. Henri Poincaré: Phys. Theor. **58**, 189–233 (1993).
[8] F. Odeh, ''Existence and bifurcation theorems for the Ginzburg-Landau equations,'' J. Math. Phys. **8**, 2351–2357 (1967).
[9] T. Giorgi, and D. Phillips, ''The breakdown of superconductivity due to strong fields for the Ginzburg Landau model'' (preprint).
[10] P. Bauman, D. Phillips, and Q. Tang, ''Stable nucleation for the Ginzburg Landau system with an applied magnetic field,'' Arch. Ration. Mech. Anal. (to be published).
[11] S. J. Chapman, ''Nucleation of vortices in type-II superconductors in increasing magnetic fields,'' Appl. Math. Lett. **10** 29–31 (1997).
[12] Q. Du, M. D. Gunzburger, J. S. Peterson, ''Analysis and approximation of the Ginzburg Landau model of superconductivity,'' SIAM (Soc. Ind. Appl. Math.) Rev. **34**, 45–81 (1992).
[13] R. Aris, *Vectors, Tensors and the Basic Equations of Fluid Mechanics* (Prentice–Hall Englewood Cliffs, 1962).
[14] D. Sarocka and A. Bernoff, ''An intrinsic equation of interfacial motion for the solidification of a pure hypercooled melt,'' Physica D **85**, 348–374 (1995).
[15] M. Dauge and B. Helffer ''Eigenvalues variation I, Neumann problem for Sturm–Liouville operators,'' J.D.E. **104** 243–262 (1993).
[16] M. C. Cross and P. C. Hohenberg, ''Pattern formation outside of equilibrium,'' Rev. Mod. Phys. **65**, 854–1112 (1993).
[17] M. Abramowitz and I. A. Stegun *Handbook of Mathematical Functions* (Dover, New York, (1964).
[18] J. Rubinstein and P. Sternberg ''Onset of superconductivity in a disc'' (in preparation).
[19] O. Buisson, P. Gandit, R. Rammal, Y. Y. Wang, and B. Pannetier, ''Magnetization oscillations of a superconducting disk,'' Phys. Lett. A **A 150**, 36–42 (1990).
[20] V. V. Moschchalkov, L. Gielen, C. Strunk, R. Jonckheere, X. Qiu, C. Van Haesendonck, and Y. Bruynseraede, ''Effect of sample topology on the critical fields of mesoscopic superconductors,'' Nature (London) **373**, 319–322 (1995).
[21] K. Lu and X. B. Pan, ''Estimates of upper critical external magnetic field for the Ginzburg–Landau equation,'' preprint.

Hidden algebras of the (super) Calogero and Sutherland models

Lars Brink[a)]
Institute of Theoretical Physics, S-412 96 Göteborg, Sweden

Alexander Turbiner[b)]
Instituto de Ciencias Nucleares, UNAM, Apartado Postal 70-543, 04510 Mexico D.F., Mexico

Niclas Wyllard[c)]
Institute of Theoretical Physics, S-412 96 Göteborg, Sweden

(Received 12 June 1997; accepted for publication 29 October 1997)

We propose to parametrize the configuration space of one-dimensional quantum systems of N identical particles by the elementary symmetric polynomials of bosonic and fermionic coordinates. It is shown that in this parametrization the Hamiltonians of the A_N, BC_N, B_N, C_N and D_N Calogero and Sutherland models, as well as their supersymmetric generalizations, can be expressed—for *arbitrary* values of the coupling constants—as quadratic polynomials in the generators of a Borel subalgebra of the Lie algebra $gl(N+1)$ or the Lie superalgebra $gl(N+1|N)$ for the supersymmetric case. These algebras are realized by first order differential operators. This fact establishes the exact solvability of the models according to the general definition given by Turbiner, and implies that the Calogero and Jack–Sutherland polynomials, as well as their supersymmetric generalizations, are related to finite-dimensional irreducible representations of the Lie algebra $gl(N+1)$ and the Lie superalgebra $gl(N+1|N)$. © *1998 American Institute of Physics.* [S0022-2488(98)01003-2]

I. INTRODUCTION

Exact solutions of non-trivial problems are always of great importance. They give a hint about the structure of real problems and also provide a laboratory for testing approximate methods. The non-relativistic many-body Calogero–Sutherland models[1,2] together with their supersymmetric extensions[3–5] provide one of the most valuable sources of exact solutions to one-dimensional many-body quantum mechanical systems. The goal of this article is to try to uncover a hidden reason for the solvability of these models in order to answer the question what is special about them and whether there are other exactly solvable or quasi-exactly solvable many-body problems.

The remarkable discovery of the solvability of the bosonic N-body Calogero[1] and Sutherland[2] models was at the time a state-of-the-art achievement. Of course, this raised the question of whether there existed a regular procedure to generate these models. After a few years, it was found that the models are connected with the root systems of the A_{N-1} Lie algebras,[6] and can be obtained from the Laplace–Beltrami operators defined on the symmetric spaces; this procedure was called ''the method of Hamiltonian reduction.''[6–9] For a contemporary viewpoint on the Hamiltonian reduction method see, for example, Ref. 10 and Ref. 27. The Hamiltonian reduction method provides a regular basis for an explanation of the solvability of the Calogero and Sutherland models, at least for a selected set of values of the coupling constant(s). Considerations of the root systems of the other simple Lie algebras made it possible to find several additional families of exactly solvable multi-dimensional Schrödinger equations and to prove the complete integrability of all these models.

[a)]Electronic mail: tfelb@fy.chalmers.se
[b)]On leave of absence from the Institute for Theoretical and Experimental Physics, Moscow 117259, Russia; electronic mail: turbiner@xochitl.nuclecu.unam.mx
[c)]Electronic mail: wyllard@fy.chalmers.se

 © 1998 American Institute of Physics

Recently, another explanation of the exact-solvability of bosonic many-body problems was presented, which was based on the finding that the eigenfunctions of the N-body A_{N-1} Calogero–Sutherland models form a flag coinciding with the flag of the finite-dimensional representation spaces of the Lie algebra $gl(N)$.[11,12] It was shown that the Hamiltonians of the Calogero and Sutherland models are nothing but different non-linear elements of the universal enveloping algebra of a Borel subalgebra of the $gl(N)$ algebra. Unlike the ''method of reduction,'' in the second method the coupling constants appear as certain parameters fixing the element of the universal enveloping algebra. The second method can be used to explain the solvability of the above models for *all* allowed values of the coupling constant(s). In the present paper we show that the BC_N, B_N, C_N and D_N Calogero (rational) and Sutherland (trigonometric) Hamiltonians, for *any* values of the coupling constants, are second order polynomials in the generators of the $gl(N+1)$ algebra. These results were made possible by exploiting the fact that the configuration space of the above quantum-mechanical systems can be parametrized by variables in which the permutation symmetry is already encoded. The most suitable variables are given by the elementary symmetric polynomials of the coordinates of the particles. In these variables the eigenfunctions have an especially simple form.

Other approaches have also been used to study the Calogero and Sutherland models. Some of these have been directed toward obtaining closed expressions for the wave functions. The bosonic A_N Calogero model was solved using an operator method in Ref. 13, a similar method was applied to the bosonic A_N Sutherland model in Ref. 14. The eigenfunctions of the BC_N Sutherland model has been studied in Refs. 9,15 (ground-state) and 16. It is worth mentioning the remarkable observation that the wave functions of the bosonic A_N Sutherland models are in correspondence with singular vectors of W-algebras (see, for example, Ref. 17).

Quite recently, supersymmetric extensions of the many-body A_N Calogero[3,4] and Sutherland[5] models were constructed using the standard prescription of supersymmetrization. In the present article we derive the supersymmetric extensions of the BC_N, B_N, C_N and D_N Calogero and Sutherland models and show that their solvability can be explained by the existence of the hidden Lie superalgebra $gl(N+1|N)$. It is worth mentioning that a supersymmetric analogue of the Hamiltonian reduction method has not been constructed for these models so far.

The paper is organized as follows: In the next section we briefly review the bosonic many-body A_N Calogero and Sutherland models with emphasis on the property of exact solvability, and also set up our notation. In Section III we study the supersymmetric extension of the Calogero model and show that it is exactly solvable. In Section IV the same analysis is carried out for the supersymmetric Sutherland model. Section V is devoted to the Calogero models connected with the Lie algebras BC_N, B_N, C_N and D_N as well as their supersymmetric extensions; it is shown that those models are exactly solvable. In Section VI the BC_N, B_N, C_N and D_N Sutherland models and their supersymmetric extensions are explored and their exact solvability is established. The Conclusion contains a summary of the results obtained and a discussion of some possible directions for future investigations. In Appendix A we present a realization of the Lie algebra $gl(N)$ and the Lie superalgebra $gl(N|M)$ in terms of first order differential operators. Appendix B contains the Lie algebraic forms of the Hamiltonians discussed in this paper. Finally, in Appendix C we give some details on the derivation of the Lie algebraic forms of the Hamiltonians for the Calogero and Sutherland models.

II. BOSONIC (MANY-BODY) A_{N-1} CALOGERO AND SUTHERLAND MODELS (REVIEW)

In this section we briefly review, following Ref. 12, the *exact-solvability* of the bosonic many-body Calogero and Sutherland models or, more precisely, the A_{N-1} Calogero and Sutherland models. For the sake of simplicity, in Sections II–IV we refer to these models as the Calogero and Sutherland models.

The Calogero and Sutherland models describe systems of N identical particles situated on the line and the circle, respectively. The degrees of freedom of these models are parametrized by N real coordinates x_i. The Hamiltonian of the Calogero model is defined by

$$\mathscr{H}_{\mathrm{Cal}}=\frac{1}{2}\sum_{i=1}^{N}\left[-\frac{\partial^2}{\partial x_i^2}+\omega^2 x_i^2\right]+g\sum_{i<j}\frac{1}{(x_i-x_j)^2}, \tag{2.1}$$

where $g=\nu(\nu-1)>-\frac{1}{4}$ is the coupling constant and ω is the harmonic oscillator frequency. For convenience only, we place the center-of-mass under the influence of the harmonic oscillator force. The ground state eigenfunction is given by

$$\Psi_0^{(c)}(x)=\Delta^{\nu}(x)e^{-\omega(X^2/2)}, \tag{2.2}$$

where $\Delta(x)=\Pi_{i<j}|x_i-x_j|$ is the Vandermonde determinant and $X^2=\Sigma_i x_i^2$. One should stress the point that to a fixed value of the coupling constant g there correspond two different values of the parameter ν, namely, $\nu=\alpha$ and $\nu=(1-\alpha)$, giving rise to two families of eigenfunctions. Of course, α should be chosen in such a way as to minimize the eigenvalue; then (2.2) corresponds to the ground state. The value $\nu=(1-\alpha)$ inserted in (2.2) describes the ground state (if normalizable) but of *another* family. If $g=0$, this family comprises the states of negative parity with respect to permutations. As has been shown by F. Calogero, any eigenstate of the system can be written in the form

$$\Psi(x)=\Psi_0^{(c)}(x)P_c(x), \tag{2.3}$$

where $P_c(x)$ is a certain polynomial in the x_i's, symmetric under the permutation of any two particles. We refer to these polynomials as *Calogero polynomials*. The operator having the Calogero polynomials as eigenfunctions is obtained from the Hamiltonian (2.1) by a ''gauge'' rotation, by which we mean the similarity transformation

$$h_{\mathrm{Cal}}=(\Psi_0^{(c)})^{-1}\mathscr{H}_{\mathrm{Cal}}\Psi_0^{(c)}. \tag{2.4}$$

The Hamiltonian of the Sutherland model is defined by

$$\mathscr{H}_{\mathrm{Suth}}=-\frac{1}{2}\sum_{k=1}^{N}\frac{\partial^2}{\partial x_k^2}+\frac{g}{4}\sum_{k<l}\frac{1}{\sin^2(\frac{1}{2}(x_k-x_l))}, \tag{2.5}$$

where $g=\nu(\nu-1)>-\frac{1}{4}$ is the coupling constant. There are two families of eigenfunctions, cf. the discussion above. The ground state eigenfunction is given by

$$\Psi_0^{(s)}(x)=(\Delta^{(trig)}(x))^{\nu}, \tag{2.6}$$

where $\Delta^{(trig)}(x)=\Pi_{i<j}|\sin(\frac{1}{2}(x_i-x_j)|$ is a trigonometric analogue of the Vandermonde determinant. It was shown by B. Sutherland that similarly to the Calogero model any eigenfunction of the Hamiltonian (2.5) can be written in the form

$$\Psi(x)=\Psi_0^{(s)}(x)P_s(e^{ix}), \tag{2.7}$$

where $P_s(e^{ix})$ is a certain polynomial in e^{ix_j}, symmetric under the permutation of any two particles. These polynomials are the so-called Jack–Sutherland polynomials[18,2] (for a general description, see, for example, Refs. 19 and 20). Performing a gauge rotation of $\mathscr{H}_{\mathrm{Suth}}$ with the gauge factor $\Psi_0^{(s)}$ we arrive at the operator

$$h_{\mathrm{Suth}}=(\Psi_0^{(s)})^{-1}\mathscr{H}_{\mathrm{Suth}}\Psi_0^{(s)}, \tag{2.8}$$

which has the Jack–Sutherland polynomials as eigenfunctions.

In order to study the internal dynamics we introduce the center-of-mass coordinate $Y=\Sigma_{j=1}^{N}x_j$ and the translation-invariant relative Perelomov coordinates:[21]

$$y_i=x_i-\frac{1}{N}Y, i=1,2,\ldots,N, \tag{2.9}$$

which obey the constraint $\Sigma_{i=1}^{N}y_i=0$. In order to incorporate the permutation symmetry and the translation invariance we consider two sets of coordinates:[12]

$$(x_1,x_2,\ldots,x_N)\to(Y,\tau_n(x)=\sigma_n(y(x))|n=2,3,\ldots,N), \tag{2.10}$$

and

$$(x_1,x_2,\ldots,x_N)\to(e^{iY},\eta_n(x)=\sigma_n(e^{iy(x)})|n=1,2,\ldots,(N-1)), \tag{2.11}$$

where

$$\sigma_k(x)=\sum_{i_1<i_2<\cdots<i_k} x_{i_1}x_{i_2}\cdots x_{i_k} \tag{2.12}$$

are the elementary symmetric polynomials (see, for example, Ref. 20). $\sigma_k(y(x))$ are thus the elementary symmetric polynomials with translation-invariant arguments (due to the constraint $\Sigma_{i=1}^N y_i=0$, the symmetric polynomials can be considered as defined in $\mathbf{R}^N$ and then restricted to the hyperplane $\Sigma_{i=1}^N y_i=0$). Then the following statement holds:[12]

After separation of the center-of-mass, the operators h_{Cal} and h_{Suth}, when written in the coordinates τ (2.10) and η (2.11), respectively, are quadratic polynomials in the generators of the Borel subalgebra of $gl(N)$ spanned by the operators J_{ij}^0 and J_i^- realized as first order differential operators (see Appendix A, Eq. (A1)):

$$h=\sum_{i,j,k,l=2}^{N} A_{ijkl}J_{ij}^0J_{kl}^0+\sum_{i,j,k=2}^{N} B_{ijk}J_{ij}^0J_k^-+\sum_{i,j=2}^{N} C_{ij}J_{ij}^0+\sum_{i=2}^{N} D_iJ_i^-$$

(see Appendix B). The operators J_{ij}^0 and J_i^- can be represented by triangular matrices and preserve the flag of spaces of inhomogeneous polynomials

$$\mathscr{P}_n=\ \mathrm{span}\left\{v_2^{n_2}v_3^{n_3}v_4^{n_4}\ldots v_N^{n_N}\middle|0\leqslant\sum\ n_i\leqslant n\right\},$$

where $v_k=\tau_k$ and $v_k=\eta_{k+1}$, respectively, and $k=2,\ldots,N$. The coupling constants g (see (2.1) and (2.5)) appear only in the coefficients C_{ij} and D_i and are not related to the dimension of the representation of $gl(N)$. Consequently, each Hamiltonian (2.1), (2.5) has one or several infinite families of polynomial eigenfunctions.

This statement leads to the important conclusion that the Calogero and Jack–Sutherland polynomials are intimately connected with the finite-dimensional irreducible representations of the Lie algebra $gl(N)$. It also provides a simple computational tool for deriving the explicit form of the Calogero and Jack–Sutherland polynomials.

III. THE SUPERSYMMETRIC MANY-BODY CALOGERO MODEL

In order to proceed to the problem of the supersymmetric generalizations of (2.1) and (2.5), let us first recall[3] that in a supersymmetric system of particles each bosonic degree of freedom x_i is accompanied by the fermionic variables ϑ_i and $\vartheta_i^\dagger$, which obey the standard anti-commutation rules $\{\vartheta_i,\vartheta_j\}=\{\vartheta_i^\dagger,\vartheta_j^\dagger\}=0$ and $\{\vartheta_i,\vartheta_j^\dagger\}=\delta_{ij}$. Throughout the paper we use a concrete realization of this algebra:

$$\vartheta_i=\theta_i,\ \vartheta_i^\dagger=\frac{\partial}{\partial\theta^i}.$$

It is convenient to introduce the ''cmino'' Ψ—the fermionic analogue of the center-of-mass coordinate as

$$\Psi=\sum_{i=1}^{N}\ \theta_i, \tag{3.1}$$

which is the super-partner of Y. We can also define the fermionic analogue of the Perelomov coordinates (2.9)

$$\lambda_i = \theta_i - \frac{1}{N}\Psi. \tag{3.2}$$

In order to construct the supersymmetric many-body Calogero model let us introduce the supercharges Q and $Q^\dagger$ as defined in Ref. 3:

$$Q = \sum_k \frac{\partial}{\partial\theta_k}\left(p_k - i\frac{\partial W}{\partial x^k}\right), \quad Q^\dagger = \sum_k \theta_k\left(p_k + i\frac{\partial W}{\partial x^k}\right), \tag{3.3}$$

where $Q^2 = Q^{\dagger 2} = 0$, $p_k = -i(\partial/\partial x_k)$, and the superpotential is

$$W(x_1, x_2, \ldots, x_N) = -\frac{\omega}{2}\sum_{i=1}^N x_i^2 + \nu\sum_{i<j}\log|x_i - x_j|. \tag{3.4}$$

Then the supersymmetric Hamiltonian $\mathscr{H}_{\mathrm{sCal}} = \frac{1}{2}\{Q, Q^\dagger\}$ has the form[3]

$$\mathscr{H}_{\mathrm{sCal}} = \frac{1}{2}\sum_{i=1}^N\left[-\frac{\partial^2}{\partial x_i^2} + \omega^2 x_i^2\right] + \omega\sum_{i=1}^N \theta_i\frac{\partial}{\partial\theta^i} + \sum_{i<j}\frac{\nu}{(x_i - x_j)^2}\left[(\nu - 1) + \theta_i - \theta_j\left(\frac{\partial}{\partial\theta_i} - \frac{\partial}{\partial\theta_j}\right)\right] + C, \tag{3.5}$$

where $C = -\frac{1}{2}\nu\omega N(N-1) - \frac{1}{2}N\omega$. The ground state eigenfunction remains the same (cf. (2.2)) as for the bosonic many-body Calogero model (2.1). It is easy to see that a gauge rotation of $\mathscr{H}_{\mathrm{sCal}}$ with the ground state wave function $\Psi_0^{(c)}$ as a gauge factor affects only the bosonic part of the Hamiltonian. Defining

$$h_{\mathrm{sCal}} = -2(\Psi_0^{(c)})^{-1}\mathscr{H}_{\mathrm{sCal}}\Psi_0^{(c)} \tag{3.6}$$

[cf. (2.4)], we obtain

$$h_{\mathrm{sCal}} = \sum_{i=1}^N \frac{\partial^2}{\partial x_i^2} - 2\omega\sum_{i=1}^N\left[x_i\frac{\partial}{\partial x_i} + \theta_i\frac{\partial}{\partial\theta^i}\right] + 2\nu\sum_{i<j}\left[\frac{1}{x_i - x_j}\left(\frac{\partial}{\partial x_i} - \frac{\partial}{\partial x_j}\right) - \frac{\theta_i - \theta_j}{(x_i - x_j)^2}\left(\frac{\partial}{\partial\theta_i} - \frac{\partial}{\partial\theta_j}\right)\right]. \tag{3.7}$$

This expression can be called the *rational* form of the supersymmetric many-body Calogero model. Similar to the bosonic case the eigenfunctions of the operator (3.7) after separation of the center-of-mass remain polynomials but now in (y^i, λ^i). These polynomials are symmetric under the permutation of any pair of particles $(x^i, \theta^i) \leftrightarrow (x^j, \theta^j)$, and can be considered as the supersymmetric analogue of the Calogero polynomials.

Now let us introduce permutation-symmetric variables. One can construct two sets of such variables: (i) a set analogous to the Newton polynomials

$$s_n = \sum_{i=1}^N x_i^n, \rho_n = \sum_{i=1}^N \theta_i x_i^{n-1} \tag{3.8}$$

and (ii) a set analogous to the elementary symmetric polynomials

$$\sigma_k = \sum_{i_1 < i_2 < \cdots < i_k} x_{i_1}x_{i_2}\cdots x_{i_k}, \zeta_k = \frac{1}{(k-1)!}\sum_{i_1 \neq i_2 \neq \cdots \neq i_k} \theta_{i_1}x_{i_2}\cdots x_{i_k}. \tag{3.9}$$

The variables s_n and ρ_n as well as σ_k and ζ_k are symmetric under the permutation, $(x^i, \theta^i) \leftrightarrow (x^j, \theta^j)$. However, the property $\sigma_{N+k} = 0, \zeta_{N+k} = 0, k = 1, 2, \ldots$ implies that the variables σ_k and ζ_k, has the advantage that they avoid the difficulties associated with the overcompleteness of the basis, which plague the variables s_n and ρ_n. Therefore, in the sequel we make use of the variables σ_k and ζ_k.

It is worth mentioning the following relations between the two sets of variables:

$$\sum_{k=0}^{N}\sigma_k(x)t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_n(x)t^n\right],$$
$$\sum_{k=0}^{N}\zeta_k(x)t^k=\left(\sum_{m=1}^{\infty}(-1)^{m+1}\rho_m(x,\theta)t^m\right)\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_n(x)t^n\right]. \tag{3.10}$$

A convenient way of succinctly writing these relations is obtained if one introduces the ''superspace'' coordinates

$$\chi_i=\sigma_i+\alpha\zeta_i,\ \phi_i=x_i+\alpha\theta_i, \tag{3.11}$$

where α is a Grassmann number, $\alpha^2=0$; then

$$\sum_{k=0}^{N}\chi_k(\phi_i)t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_n(\phi_i)t^n\right], \tag{3.12}$$

encodes both the relations (3.10).

The ''superspace'' formulation makes it possible to write formulas in a more compact way. The ''supercoordinate'' ϕ_i is clearly an $N=1$ superfield. However, one should stress that the models we study possess $N=2$ supersymmetry, while the ''superspace'' formulation is not manifestly $N=2$ supersymmetric. Nevertheless, later on we will see explicitly that the superspace (we drop the quotation marks hereafter) formulation turns out to be a useful computational aid. For example, one can rewrite the Hamiltonian h_{sCal} in the superspace coordinates as follows:

$$h_{\mathrm{sCal}}=\int d\alpha d\bar{\alpha}\sum_{i=1}^{N}\frac{\delta}{\delta\bar{\phi}_i}\frac{\delta}{\delta\phi_i}-\int d\alpha\mathscr{W}-\int d\bar{\alpha}\bar{\mathscr{W}}. \tag{3.13}$$

Here $\bar{\alpha}$ is a Grassmann variable independent of α and

$$\frac{\delta}{\delta\phi_i}=\frac{\partial}{\partial\theta_i}+\alpha\frac{\partial}{\partial x_i},\ \frac{\delta}{\delta\bar{\phi}_i}=\frac{\partial}{\partial\theta_i}+\bar{\alpha}\frac{\partial}{\partial x_i}, \tag{3.14}$$

while

$$\mathscr{W}=\omega\sum_{i=1}^{N}\phi_i\frac{\delta}{\delta\phi_i}-\nu\sum_{i<j}\frac{1}{\phi_i-\phi_j}\left(\frac{\delta}{\delta\phi_i}-\frac{\delta}{\delta\phi_j}\right),\ \bar{\mathscr{W}}=\mathscr{W}(\bar{\phi}), \tag{3.15}$$

where $\bar{\phi}_i=x_i+\bar{\alpha}\theta_i$. Notice that we have not introduced any superderivatives; $\delta/\delta\phi_i$ and $\delta/\delta\bar{\phi}_i$ should be considered as ''superfields'' (since they do not contain derivatives with respect to t, α and/or $\bar{\alpha}$).

Like we did in Section II we would like to separate the dynamics of the center-of-mass and the relative motion, and encode in relative coordinates the translation invariance and permutation symmetry of the system. In order to implement these requirements we introduce the translation-invariant and permutation-symmetric coordinates of the relative motion

$$\tau_k(y)=\sigma_k(y(x)),\kappa_k(y,\lambda)=\zeta_k(y(x),\lambda(\theta)), \tag{3.16}$$

for example, in the symmetric polynomials (3.9) the arguments (x's) are replaced by the translation-invariant y's (2.9) and the θ's are replaced by the translation-invariant λ's (3.2). Here $\tau_0=1,\tau_1=0$ and $\kappa_0=\kappa_1=0$. The next step is to make the change of variables

$$(x_i,\theta_i|i=1,2,\ldots,N)\to(Y,\tau_k;\Psi,\kappa_k|k=2,3,\ldots,N) \tag{3.17}$$

in the Hamiltonian h_{sCal}. In general, this is a tedious and cumbersome calculation. Fortunately, the superspace formulation allows us to avoid most of the tiresome algebraic calculations. Let us define

$$\tilde{\phi}_i = \phi_i - \frac{1}{N}\sum_{j=1}^{N} \phi_j . \tag{3.18}$$

Then the expression for $\psi_i = \tau_i + \alpha\kappa_i$ can be obtained from the relation

$$\sum_{k=0}^{N} \psi_k(\tilde{\phi}_i) t^k = \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^n}{n} s_n(\tilde{\phi}_i) t^n\right]. \tag{3.19}$$

The derivative

$$\frac{\delta}{\delta\phi_i(\alpha)} = \frac{\partial}{\partial\theta_i} + \alpha\frac{\partial}{\partial x_i}$$

satisfies

$$\frac{\delta\phi_j(\beta)}{\delta\phi_i(\alpha)} = \delta_{ij}(\alpha-\beta) = \delta_{ij}\delta(\alpha-\beta), \tag{3.20}$$

where we have used that $\int d\alpha\,\delta(\alpha-\beta) f(\alpha) = f(\beta)$. We now make the change of variables

$$(\phi_i | i = 1, \ldots, N) \to \left(\chi_1 = \sum_{k=1}^{N} \phi_k ; \psi_j \,\middle|\, j = 2, \ldots, N\right).$$

Under such a change of variables the derivatives transform as (from now on we suppress the center-of-mass coordinate χ_1)

$$\frac{\delta}{\delta\phi_i(\alpha)} = -\int d\beta \frac{\delta\psi_j(\beta)}{\delta\phi_i(\alpha)} \frac{\delta}{\delta\psi_j(\beta)}, \tag{3.21}$$

where

$$\frac{\delta}{\delta\psi_i(\alpha)} = \frac{\partial}{\partial\kappa_i} + \alpha\frac{\partial}{\partial\tau_i},$$

and summation over j is assumed. To prove the relation (3.21) one simply writes both sides of the equality in components. It is convenient to perform the change of variables in two steps: to write,

$$\frac{\delta}{\delta\phi_i(\alpha)} = \int d\beta d\gamma \frac{\delta\tilde{\phi}_j(\gamma)}{\delta\phi_i(\alpha)} \frac{\delta\psi_k(\beta)}{\delta\tilde{\phi}_j(\gamma)} \frac{\delta}{\delta\psi_k(\beta)}, \tag{3.22}$$

and then use the definition of $\tilde{\phi}_i$ together with (3.21) which leads to $\delta\tilde{\phi}_j(\gamma)/\delta\phi_i(\alpha) = (\delta_{ij} - 1/N)\delta(\alpha-\gamma)$. Let us give as an example the expression for the Laplace operator in the new coordinates

$$\sum_{i=1}^{N}\int d\alpha d\beta \frac{\delta}{\delta\phi_i(\beta)}\frac{\delta}{\delta\phi_i(\alpha)} \to \sum_{i=1}^{N}\sum_{j,k=2}^{N}\sum_{l,m=1}^{N}\int d\alpha d\beta \int d\gamma\left(\delta_{il}-\frac{1}{N}\right) \times \frac{\delta\psi_j(\gamma)}{\delta\tilde{\phi}_l(\beta)}\frac{\delta}{\delta\psi_j(\gamma)}\int d\epsilon\left(\delta_{im}-\frac{1}{N}\right)\frac{\delta\psi_k(\epsilon)}{\delta\tilde{\phi}_m(\alpha)}\frac{\delta}{\delta\psi_k(\epsilon)}. \tag{3.23}$$

The above considerations imply that within the framework of the superspace formalism the calculations for the supersymmetric models follow closely the calculations carried out in the bosonic case.[12] In fact, for most of the terms the supersymmetric results can be obtained from the bosonic ones by judiciously replacing the bosonic coordinates by superspace coordinates. More details can be found in Appendix C. The derivation makes extensive use of the generating function (3.12).

In the superspace coordinates the Laplace operator becomes (after reinserting the center-of-mass coordinate)

$$\sum_{k=1}^{N}\int d\alpha d\bar{\alpha}\frac{\delta}{\delta\bar{\phi}_k}\frac{\delta}{\delta\phi_k} = \int d\alpha d\bar{\alpha}\left[N\frac{\delta}{\delta\bar{\chi}_1}\frac{\delta}{\delta\chi_1}+\sum_{i,j=2}^{N}\mathscr{A}_{ij}\frac{\delta}{\delta\bar{\psi}_j}\frac{\delta}{\delta\psi_i}\right] + \int d\alpha\sum_{i=2}^{N}\mathscr{B}_i\frac{\delta}{\delta\psi_i}+\int d\bar{\alpha}\sum_{i=2}^{N}\bar{\mathscr{B}}_i\frac{\delta}{\delta\bar{\psi}_i}, \tag{3.24}$$

where

$$\psi_i=\tau_i+\alpha\kappa_i,\ \frac{\delta}{\delta\psi_i}=\frac{\partial}{\partial\kappa_i}+\alpha\frac{\partial}{\partial\tau_i}, \tag{3.25}$$

and

$$\mathscr{A}_{ij}=\frac{(N-i+1)(j-1)}{N}\psi_{i-1}\bar{\psi}_{j-1}+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\psi_{i+l-1}\bar{\psi}_{j-l-1} -\sum_{l\geqslant1} l[\bar{\psi}_{i+l}\psi_{j-2-l}-\psi_{i+l}\bar{\psi}_{j-2-l}], \tag{3.26}$$

$$\mathscr{B}_i(\psi)=-\frac{1}{2N}\sum_{i=2}^{N}(N-i+2)(N-i+1)\psi_{i-2},\ \bar{\mathscr{B}}_i=\mathscr{B}_i(\bar{\psi}). \tag{3.27}$$

The final expression for the Laplace operator in components is given by

$$\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2}=N\frac{\partial^2}{\partial Y^2}+\sum_{i,j=2}^{N}\left[A_{ij}^{\tau\tau}\frac{\partial^2}{\partial\tau_i\partial\tau_j}+A_{ij}^{\tau\kappa}\frac{\partial^2}{\partial\tau_i\partial\kappa_j}+A_{ij}^{\kappa\tau}\frac{\partial^2}{\partial\tau_j\partial\kappa_i}+A_{ij}^{\kappa\kappa}\frac{\partial^2}{\partial\kappa_j\partial\kappa_i}\right] +\sum_{i=2}^{N}\left[B_i^{\tau}\frac{\partial}{\partial\tau_i}+B_i^{\kappa}\frac{\partial}{\partial\kappa_i}\right], \tag{3.28}$$

where

$$A_{ij}^{\tau\tau}=\frac{(N-i+1)(j-1)}{N}\tau_{i-1}\tau_{j-1}+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\tau_{i+l-1}\tau_{j-l-1},$$

$$A_{ij}^{\tau\kappa}=\frac{(N-i+1)(j-1)}{N}\tau_{i-1}\kappa_{j-1}+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\tau_{i+l-1}\kappa_{j-l-1} -\sum_{l\geqslant1} l[\kappa_{i+l}\tau_{j-2-l}-\tau_{i+l}\kappa_{j-2-l}],$$

$$A_{ij}^{\kappa\tau}=\frac{(N-i+1)(j-1)}{N}\kappa_{i-1}\tau_{j-1}+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\kappa_{i+l-1}\tau_{j-l-1}$$
$$+\sum_{l\geqslant 1} l[\kappa_{i+l}\tau_{j-2-l}-\tau_{i+l}\kappa_{j-2-l}], \tag{3.29}$$

$$A_{ij}^{\kappa\kappa}=\frac{(N-i+1)(j-1)}{N}\kappa_{i-1}\kappa_{j-1}+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\kappa_{i+l-1}\kappa_{j-l-1}+2\sum_{l\geqslant 1} l\kappa_{i+l}\kappa_{j-2-l},$$

$$B_i^{\tau}=-\frac{(N-i+2)(N-i+1)}{N}\tau_{i-2},$$

$$B_i^{\kappa}=-\frac{(N-i+2)(N-i+1)}{N}\kappa_{i-2}.$$

As expected, $A_{ij}^{\tau\tau}$ and B_i^{τ} coincide with the expressions found in Ref. 12 for the bosonic Calogero model. Let us note that the coefficient functions A_{ij} and B_i are second and first order polynomials in τ_i and κ_i, respectively.

Dropping the decoupled center-of-mass terms one can show that in superspace coordinates the remaining part of the Hamiltonian $h_{\mathrm{sCal}}^{(\mathrm{rel})}$, which describes the relative motion, becomes

$$h_{\mathrm{sCal}}^{(\mathrm{rel})}=\int d\alpha d\bar{\alpha}\sum_{i,j=2}^{N}\mathscr{A}_{ij}\frac{\delta}{\delta\bar{\psi}_j}\frac{\delta}{\delta\psi_i}-\int d\alpha\mathscr{W}-\int d\bar{\alpha}\bar{\mathscr{W}}, \tag{3.30}$$

where $\mathscr{A}_{ij}$ is given in (3.26) and

$$\mathscr{W}=\omega\sum_{i=2}^{N} i\psi_i\frac{\delta}{\delta\psi_i}+\frac{1}{2}\left(\frac{1}{N}+\nu\right)\sum_{i=2}^{N}(N-i+2)(N-i+1)\psi_{i-2}\frac{\delta}{\delta\psi_i},\ \bar{\mathscr{W}}=\mathscr{W}(\bar{\psi}). \tag{3.31}$$

Finally, in component form $h_{\mathrm{sCal}}^{(\mathrm{rel})}$ looks like

$$h_{\mathrm{sCal}}^{(\mathrm{rel})}=\sum_{i,j=2}^{N}\left[A_{ij}^{\tau\tau}\frac{\partial^2}{\partial\tau_i\partial\tau_j}+A_{ij}^{\tau\kappa}\frac{\partial^2}{\partial\tau_i\partial\kappa_j}+A_{ij}^{\kappa\tau}\frac{\partial^2}{\partial\kappa_i\partial\tau_j}+A_{ij}^{\kappa\kappa}\frac{\partial^2}{\partial\kappa_j\partial\kappa_i}\right]$$
$$-\left(\frac{1}{N}+\nu\right)\sum_{i=2}^{N}(N-i+2)(N-i+1)\left[\tau_{i-2}\frac{\partial}{\partial\tau_i}+\kappa_{i-2}\frac{\partial}{\partial\kappa_i}\right]-2\omega\sum_{i=2}^{N} i\left[\tau_i\frac{\partial}{\partial\tau_i}+\kappa_i\frac{\partial}{\partial\kappa_i}\right], \tag{3.32}$$

where $A_{ij}^{\tau\tau}, A_{ij}^{\tau\kappa}, A_{ij}^{\kappa\tau}, A_{ij}^{\kappa\kappa}, B_i^{\tau}, B_i^{\kappa}$ are given in (3.29). This form can be called the *algebraic* form of the supersymmetric many-body Calogero Hamiltonian.

It turns out that we are able to rewrite the supersymmetric Calogero Hamiltonian $h_{\mathrm{sCal}}^{(\mathrm{rel})}$ in terms of generators of the Borel subalgebra of the algebra $gl(N|N-1)$ spanned by first order differential operators in the variables $(\tau_k,\kappa_k), k=2,\ldots,N$ (see Appendix A2). The result is given in Appendix B, Eq. (B2). Since the expression (B2) contains no positive-root generators J_i^+, Q_γ^+, then in accordance with the general definition given in Ref. 11 we conclude that the supersymmetric many-body Calogero model (3.5) is exactly solvable. The existence of the representation (B2) proves that there are infinitely many eigenfunctions of the operator (3.7) having the form of polynomials in the variables (τ_k,κ_k). It also implies that eigenfunctions of the supersymmetric many-body Calogero model (3.5) have a factorizable form being the product of the ground-state eigenfunction multiplied by a polynomial in the variables (τ_k,κ_k). These polynomials are related to finite-dimensional irreducible representations of the algebra $gl(N|N-1)$ in the realization (A3).

IV. THE SUPERSYMMETRIC MANY-BODY SUTHERLAND MODEL

In this section we analyze the exact-solvability of the supersymmetric Sutherland model.

In order to define the supersymmetric extension of the bosonic many-body Sutherland model (2.5) we use the standard prescription of supersymmetrization already used in the previous section. Let us take the supercharges $Q, Q^\dagger$ (see (3.3)) with a superpotential

$$W(x_1, x_2, \ldots, x_N) = \nu \sum_{i<j} \log \left| \sin\left(\frac{1}{2}(x_i - x_j) \right) \right| = \nu \log \Delta^{(trig)}, \tag{4.1}$$

[cf. (3.4)] and then construct the supersymmetric Hamiltonian $\mathscr{H}_{\mathrm{sCal}} = \frac{1}{2}\{Q, Q^\dagger\}$. After carrying out the calculation the Hamiltonian of the supersymmetric many-body Sutherland model emerges:[5]

$$\mathscr{H}_{\mathrm{sSuth}} = -\frac{1}{2}\sum_{k=1}^{N} \frac{\partial^2}{\partial x_k^2} + \frac{1}{4}\sum_{k<l} \frac{\nu}{\sin^2(\frac{1}{2}(x_k - x_l))} \left[\nu - 1 + (\theta_k - \theta_l)\left(\frac{\partial}{\partial \theta_k} - \frac{\partial}{\partial \theta_l} \right) \right] + C, \tag{4.2}$$

where $C = -N(N^2-1)\nu^2/3$. The ground state wave function remains the same as in the bosonic case, $\Psi_0^{(s)} = (\Delta^{(trig)}(x))^\nu$, where $\Delta^{(trig)}(x) = \Pi_{k<l} |\sin(\frac{1}{2}(x_k - x_l))|$. Introducing the gauge-rotated Hamiltonian $h_{\mathrm{sSuth}} = -2(\Psi_0^{(s)})^{-1} \mathscr{H} \Psi_0^{(s)}$, we get

$$h_{\mathrm{sSuth}} = \sum_{k=1}^{N} \frac{\partial^2}{\partial x_k^2} + \nu \sum_{k<l} \left[\cot\left(\frac{1}{2}(x_k - x_l) \right) \left(\frac{\partial}{\partial x_k} - \frac{\partial}{\partial x_l} \right) - \frac{1}{2} \frac{(\theta_k - \theta_l)}{\sin^2(\frac{1}{2}(x_k - x_l))} \left(\frac{\partial}{\partial \theta_k} - \frac{\partial}{\partial \theta_l} \right) \right]. \tag{4.3}$$

It is worth mentioning that if the operator (4.3) is written in the coordinates e^{ix_j}, it appears in its *rational* form, which is a supersymmetric generalization of the rational form of the Sutherland Hamiltonian (see Ref. 2).

One can show that like in the bosonic case [cf. (2.8)] after separation of the center-of-mass motion the eigenfunctions of the operator (4.3) remain polynomials but now in the coordinates (e^{iy_k}, λ_k). These polynomials are symmetric under the permutation of any pair of particles $(x_i, \theta_i) \leftrightarrow (x_j, \theta_j)$ and can be considered as a supersymmetric analogue of the Jack–Sutherland polynomials.

In the superspace coordinates (3.11), (3.14) the Hamiltonian h_{sSuth} can be written

$$h_{\mathrm{sSuth}} = \int d\alpha d\bar{\alpha} \sum_{k=1}^{N} \frac{\delta}{\delta \bar{\phi}_i} \frac{\delta}{\delta \phi_i} + \nu \int d\alpha \mathscr{W} + \nu \int d\bar{\alpha} \bar{\mathscr{W}}, \tag{4.4}$$

where

$$\mathscr{W}(\phi) = \frac{1}{2}\sum_{k<l} \cot\left(\frac{1}{2}(\phi_k - \phi_l) \right) \left(\frac{\delta}{\delta \phi_k} - \frac{\delta}{\delta \phi_l} \right), \bar{\mathscr{W}} = \mathscr{W}(\bar{\phi}). \tag{4.5}$$

Next, we introduce the new variables [cf. (3.16)]

$$\eta_n + \alpha \mu_n = \sigma_n(\exp[i(y_k + \alpha \lambda_k)]), n = 1, \ldots, N-1, \tag{4.6}$$

and

$$\sigma_N + \alpha \zeta_N = \sigma_N(\exp[i(x_k + \alpha \theta_k)]), n = N, \tag{4.7}$$

where $\alpha^2 = 0$, σ_n is defined in (3.9) and y_i, λ_i are given in (2.9), (3.2), respectively. Furthermore,

$$\begin{aligned} \sum_{k=0}^{N} \eta_k t^k &= \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} s_n(e^{iy_k}) t^n \right], \\ \sum_{k=0}^{N} \mu_k t^k &= \left(\sum_{m=1}^{\infty} (-1)^{m+1} \rho_m(e^{iy_k}, \lambda_i e^{iy_k}) t^m \right) \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} s_n(e^{iy_k}) t^n \right], \end{aligned} \tag{4.8}$$

[cf. (3.10)], where we have set $\eta_N=1$ and $\mu_N=0$.

Following the same procedure as in Section III one can rewrite the Laplace operator in the new coordinates (4.6), (4.7):

$$-\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2}=N\left(\sigma_N\frac{\partial}{\partial\sigma_N}\right)^2+\sum_{i,j=1}^{N-1}\left[A_{ij}^{\eta\eta}\frac{\partial^2}{\partial\eta_i\partial\eta_j}+A_{ij}^{\eta\mu}\frac{\partial^2}{\partial\eta_i\partial\mu_j}+A_{ij}^{\mu\eta}\frac{\partial^2}{\partial\eta_j\partial\mu_i}+A_{ij}^{\mu\mu}\frac{\partial^2}{\partial\mu_j\partial\mu_i}\right]$$
$$+\sum_{i=1}^{N-1}\left[B_i^{\eta}\frac{\partial}{\partial\eta_i}+B_i^{\mu}\frac{\partial}{\partial\mu_i}\right], \tag{4.9}$$

where

$$A_{ij}^{\eta\eta}=\frac{j(N-i)}{N}\eta_i\eta_j+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\,\eta_{i+l}\eta_{j-l},$$

$$A_{ij}^{\eta\mu}=\frac{j(N-i)}{N}\eta_i\mu_j+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\,\eta_{i+l}\mu_{j-l}-\sum_{l\geqslant1}l[\mu_{i+l-1}\eta_{j-l-1}-\eta_{i+l-1}\mu_{j-l-1}],$$

$$A_{ij}^{\mu\eta}=\frac{j(N-i)}{N}\mu_i\eta_j+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\mu_{i+l}\eta_{j-l}+\sum_{l\geqslant1}l[\mu_{i+l-1}\eta_{j-l-1}-\eta_{i+l-1}\mu_{j-l-1}], \tag{4.10}$$

$$A_{ij}^{\mu\mu}=\frac{j(N-i)}{N}\mu_i\mu_j+\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\mu_{i+l}\mu_{j-l}+2\sum_{l\geqslant1}l\mu_{i+l-1}\mu_{j-l+1},$$

$$B_i^{\eta}=\frac{i(N-i)}{N}\eta_i,$$

$$B_i^{\mu}=\frac{i(N-i)}{N}\mu_i.$$

As expected, the coefficients $A_{ij}^{\eta\eta}$ and B_i^{η} coincide with the expressions found in Ref. 12 for the bosonic Sutherland model. Furthermore, the coefficient functions A_{ij} and B_i are second and first order polynomials in η_i and μ_i, respectively. Similar to what happens in the bosonic case the Laplace operator (4.9) possesses infinitely many polynomial eigenfunctions.[12] These have the form of supersymmetric analogies of the Bethe-ansatz wave functions.

After omitting the center-of-mass motion associated with σ_N and ζ_N, the final expression for the supersymmetric many-body Sutherland Hamiltonian h_{sSuth} takes the form [cf. Section III, Eq. (3.32)]

$$h_{\mathrm{sSuth}}^{(\mathrm{rel})}=\sum_{i,j=1}^{N-1}\left[A_{ij}^{\eta\eta}\frac{\partial^2}{\partial\eta_i\partial\eta_j}+A_{ij}^{\eta\mu}\frac{\partial^2}{\partial\eta_i\partial\mu_j}+A_{ij}^{\mu\eta}\frac{\partial^2}{\partial\eta_j\partial\mu_i}+A_{ij}^{\mu\mu}\frac{\partial^2}{\partial\mu_j\partial\mu_i}\right]$$
$$+\left(\frac{1}{N}+\nu\right)\sum_{i=1}^{N-1}i(N-i)\left[\eta_i\frac{\partial}{\partial\eta_i}+\mu_i\frac{\partial}{\partial\mu_i}\right], \tag{4.11}$$

where $A_{ij}^{\eta\eta},A_{ij}^{\eta\mu},A_{ij}^{\mu\eta},A_{ij}^{\mu\mu},B_i^{\eta},B_i^{\mu}$ are given in (4.10). This expression can be called the *algebraic* form of the supersymmetric many-body Sutherland Hamiltonian.

In superspace coordinates the Hamiltonian $h_{\mathrm{sSuth}}^{(\mathrm{rel})}$ describing the relative motion can be written

$$-h_{\mathrm{sSuth}}^{(\mathrm{rel})}=\int d\alpha d\bar{\alpha}\sum_{i,j=1}^{N-1}\mathscr{A}_{ij}\frac{\delta}{\delta\bar{\psi}_j}\frac{\delta}{\delta\psi_i}+\int d\alpha\mathscr{W}+\int d\bar{\alpha}\bar{\mathscr{W}}, \tag{4.12}$$

where

$$\mathscr{W}=\frac{1}{2}\left(\frac{1}{N}+\nu\right)\sum_{i=1}^{N-1} i(N-i)\psi_i\frac{\delta}{\delta\psi_i},\ \overline{\mathscr{W}}=\mathscr{W}(\overline{\psi}), \tag{4.13}$$

$$\psi_i=\eta_i+\alpha\mu_i,\ \frac{\delta}{\delta\psi_i}=\frac{\partial}{\partial\mu_i}+\alpha\frac{\partial}{\partial\eta_i}, \tag{4.14}$$

[cf. (3.25)], while

$$\mathscr{A}_{ij}=\frac{j(N-i)}{N}\psi_i\overline{\psi}_j-\sum_{l\geqslant\max(1,j-i)}(j-i-2l)\psi_{i+l}\overline{\psi}_{j-l}+\sum_{l\geqslant 1} l[\overline{\psi}_{i+l-1}\psi_{j-l+1}-\psi_{i+l-1}\overline{\psi}_{j-l+1}], \tag{4.15}$$

[cf. (3.26)]. The method of the calculation of the coefficients $\mathscr{A}_{ij}$ is similar to that presented in Appendix C.

It turns out that the Hamiltonian governing the relative motion can be rewritten in terms of the generators of a Borel subalgebra of the Lie superalgebra $gl(N|N-1)$ (see Appendix A 2) after substituting $\kappa_i\to\mu_{i-1}$ and $\tau_i\to\eta_{i-1}$. The final result is given in Appendix B, Eq. (B4). Since the expression (B4) contains no positive-root generators T_i^+, Q_γ^+, then in accordance with the general definition[11] we conclude that the supersymmetric many-body Sutherland model (4.2) is exactly solvable. The existence of the representation (B4) proves that there are infinitely many eigenfunctions of the operator (4.3) having the form of polynomials in the variables (η_k,μ_k). It also leads to the conclusion that eigenfunctions of the supersymmetric many-body Sutherland model (4.2) have a factorizable form being the product of the ground-state eigenfunction multiplied by a polynomial in the variables (η_k,μ_k). These polynomials are related to finite-dimensional irreducible representations of the algebra $gl(N|N-1)$ in the realization (A3).

To conclude, the supersymmetric many-body Sutherland models possess the algebraic form (4.11) and also the Lie-algebraic form (B4) represented by second-order polynomials in the generators of the of the algebra $gl(N|N-1)$ with certain coefficients.

V. THE BC_N, B_N, C_N AND D_N CALOGERO MODELS AND THEIR SUPERSYMMETRIC EXTENSIONS

It is well-known that there is a deep connection between the integrable Calogero–Sutherland systems and Lie algebras (see the discussion in the Introduction). This connection is explicitly realized in the Hamiltonian reduction method[6,7] (see also Refs. 8–10). In particular, the celebrated many-body Calogero and Sutherland models discussed in Section II are related to the Lie algebras A_{N-1} and are therefore known as the Calogero and Sutherland systems of A_{N-1}-type. A natural question emerges: What are the integrable systems corresponding to the other simple Lie algebras like B_N, C_N, D_N, etc., appearing in the Hamiltonian reduction method? The answer and a complete classification of these systems is given in Ref. 6 (for a review, see Refs. 8 and 9). In the present paper we focus on the quantum Calogero (rational) and Sutherland (trigonometric) systems of the BC_N, B_N, C_N and D_N types leaving for a future publication the case of the exceptional algebras.

A. The bosonic case

Unlike the A_{N-1} many-body Calogero model, the BC_N, B_N, C_N and D_N Calogero models are not translation invariant, and describe systems with boundaries. However, they are still permutation invariant with respect to the interchange of any pair of coordinates. The configuration space of these models is $\{x_i | x_i>0; j<k: x_j<x_k\}$. In this section we will show that like the A_{N-1} many-body Calogero system, all the BC_N, B_N, C_N, D_N quantum Calogero systems are not only completely integrable, but also exactly-solvable possessing a hidden algebra $gl(N+1)$.

It is well known that the Hamiltonians of the BC_N, B_N, C_N Calogero models are all given by (see Ref. 6)

$$\mathscr{H}^{(c)}_{BCD}=\frac{1}{2}\sum_{i=1}^{N}[-\partial_i^2+\omega^2 x_i^2]+g\sum_{i<j}\left[\frac{1}{(x_i-x_j)^2}+\frac{1}{(x_i+x_j)^2}\right]+\frac{g_2}{2}\sum_{i=1}^{N}\frac{1}{x_i^2}, \tag{5.1}$$

where $g=\nu(\nu-1)$ and $g_2=\nu_2(\nu_2-1)$. When the coupling constant g_2 tends to zero, the Hamiltonian $\mathscr{H}^{(c)}_{BCD}$ degenerates to the Hamiltonian of the D_N Calogero model. The ground state eigenfunction of the Hamiltonian (5.1) is given by

$$\Psi_0=\left[\prod_{i<j}|x_i-x_j|^{\nu}|x_i+x_j|^{\nu}\prod_{i=1}^{N}|x_i|^{\nu_2}\right]e^{-\omega/2\sum_{i=1}^{N}x_i^2}, \tag{5.2}$$

[cf. (2.2)].

One should stress the point (see also the discussion in Section II) that to fixed values of the coupling constants g,g_2 there corresponds two different values of the parameter $\nu(\nu_2)$: $\nu(\nu_2)=\alpha(\alpha_2)$ and $(1-\alpha(\alpha_2))$ giving rise to four families of eigenfunctions. Of course, $\alpha(\alpha_2)$ should be chosen in such a way as to minimize the eigenvalue and then (5.2) corresponds to the ground state. The ground state can be denoted as (α,α_2). The other values $\nu(\nu_2)$ inserted in (5.2) describe the ground states (if normalizable) of three *other* families: $(1-\alpha,\alpha_2),(\alpha,1-\alpha_2)$, $(1-\alpha,1-\alpha_2)$. If $g_2=0$, the sectors $(\alpha,1-\alpha_2),(1-\alpha,1-\alpha_2)$ correspond to totally antisymmetric states with respect to the reflections $x_i\leftarrow -x_i$. In what follows, for the sake of simplicity, we call (5.2) the ground state, of course keeping in mind the above discussion.

Following the same approach as in Sections II–IV, we make a gauge rotation of (5.1) with the ground-state eigenfunction as the gauge factor; $h^{(c)}_{BCD}=-2\Psi_0^{-1}\mathscr{H}\Psi_0$. A straightforward calculation gives (we omit an additive constant)

$$h^{(c)}_{BCD}=\sum_{i=1}^{N}\partial_i^2+2\nu\sum_{i<j}\left[\frac{1}{x_i-x_j}(\partial_i-\partial_j)+\frac{1}{x_i+x_j}(\partial_i+\partial_j)\right]+\nu_2\sum_{i=1}^{N}\frac{1}{x_i}\partial_i-2\omega\sum_{i=1}^{N}x_i\partial_i. \tag{5.3}$$

We call this operator the *rational* form of the BC_N,B_N,C_N and D_N Calogero systems.

The Hamiltonian (5.1) as well as the operator (5.3) is invariant under the permutation of any pair of the coordinates $x_i\leftrightarrow x_j$ and also under the action of the reflection group $\mathbf{Z}_2^{\oplus N}$: $x_i\rightarrow -x_i, i=1,2,\ldots,N$. An infinite set of eigenfunctions of (5.3) are given as polynomials in the x's and are classified by the representations of the reflection group $\mathbf{Z}_2^{\oplus N}$. For the sake of simplicity, we consider in what follows only the eigenfunctions of the Hamiltonian (5.3), which are *totally symmetric* under the reflection and permutation group actions. The *totally (anti)symmetric-(anti)symmetric* eigenfunctions under the reflection and/or permutation group actions are reproduced by a change of parameters ν,ν_2 in (5.1). This fact implies that it is sufficient to look for eigenfunctions to the operator (5.3) of the form

$$\Psi^{\nu,\nu_2}=F(x_1^2,x_2^2,\ldots,x_N^2), \tag{5.4}$$

Now let us construct variables which encode the permutation invariance of the system. Analogous to what was done for the A_{N-1} Calogero model[12] we use as new variables the elementary symmetric polynomials σ_k [see (2.12)] but with x_i^2 as arguments:

$$\sum_{k=0}^{N}\widetilde{\sigma}_k(x_i^2)t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_{2n}(x_i)t^n\right], \tag{5.5}$$

[cf. (2.10), (3.10)]. Finally, in the new variables the Hamiltonian $h^{(c)}_{BCD}$ becomes

$$h^{(c)}_{BCD}=\sum_{i,j=1}^{N}A_{ij}\frac{\partial}{\partial\widetilde{\sigma}_i}\frac{\partial}{\partial\widetilde{\sigma}_j}+\sum_{j=1}^{N}B_j\frac{\partial}{\partial\widetilde{\sigma}_j} \tag{5.6}$$

where [cf. (3.28)]

$$\begin{aligned}A_{ij}&=4\sum_{l\geqslant 0}(2l+1+j-i)\widetilde{\sigma}_{i-l-1}\widetilde{\sigma}_{j+l},\\ B_j&=2[1+\nu_2+2\nu(N-j)](N-j+1)\widetilde{\sigma}_{j-1}-4\omega j\widetilde{\sigma}_j,\end{aligned} \tag{5.7}$$

and $\widetilde{\sigma}_k=0$, when $k>N$ or $k<0$. The method of the calculation of the coefficients A_{ij}, B_j is similar to that presented in Appendix C. This expression can be called the *algebraic* form of the BC_N, B_N, C_N and D_N Calogero Hamiltonians.

Similar to what happened in all previously discussed cases the coefficients A_{ij}, B_j are polynomials of second and first degree in $\widetilde{\sigma}_k$, respectively. Hence, there exists a Lie algebraic form of the Hamiltonian and $h^{(c)}_{BCD}$ can be written in terms of generators of a Borel subalgebra of $gl(N+1)$ (see Appendix A 1) as in Section II. The result is presented in Appendix B, Eq. (B5). Then in accordance with the general definition given in Ref. 11 we can conclude that all BC_N, B_N, C_N, D_N Calogero models (5.1) are exactly solvable. The existence of the representation (5.6) proves that there are infinitely many eigenfunctions of the operator (5.3) having the form of polynomials in the variables $\widetilde{\sigma}_k$. It also implies that totally (anti)symmetric-(anti)symmetric eigenfunctions with respect to permutations and reflections of the BC_N, B_N, C_N and D_N Calogero models (5.1) have a factorizable form being the product of the ground-state eigenfunction (5.2) multiplied by a polynomial in the variables $\widetilde{\sigma}_k$. These polynomials are related to finite-dimensional irreducible representations of the Lie algebra $gl(N+1)$ in the realization (A1) and can be called the BC_N *Calogero* polynomials.

B. The supersymmetric extensions

The BC_N, B_N, C_N and D_N Calogero models discussed above have also natural supersymmetric extensions. These can be constructed in a straightforward way. Let us consider the supercharges (3.3) with the superpotential W given by

$$W=\nu\sum_{i<j}\left(\log|x_i-x_j|+\log|x_i+x_j|\right)+\nu_2\sum_{i=1}^{N}\log|x_i|-\frac{\omega}{2}\sum_{i=1}^{N}x_i^2 \tag{5.8}$$

[cf. (3.4)]. After some calculations the supersymmetric Hamiltonian $\mathscr{H}^{(c)}_{\mathrm{sBCD}}=\frac{1}{2}\{Q,Q^\dagger\}$ of the supersymmetric BC_N, B_N, C_N, D_N Calogero models emerges:

$$\begin{aligned}\mathscr{H}^{(c)}_{\mathrm{sBCD}}=&\frac{1}{2}\sum_{i=1}^{N}\left[-\frac{\partial^2}{\partial x_i^2}+\omega^2x_i^2\right]+\frac{\nu_2}{2}\sum_{i=1}^{N}\frac{1}{x_i^2}\left[\nu_2-1+\theta_i\frac{\partial}{\partial\theta_i}\right]\\&+\omega\sum_{i=1}^{N}\theta_i\frac{\partial}{\partial\theta_i}+\nu\sum_{i<j}\frac{1}{(x_i-x_j)^2}\left[\nu-1+(\theta_i-\theta_j)\left(\frac{\partial}{\partial\theta_i}-\frac{\partial}{\partial\theta_i}\right)\right]\\&+\nu\sum_{i<j}\frac{1}{(x_i+x_j)^2}\left[(\nu-1)+(\theta_i+\theta_j)\left(\frac{\partial}{\partial\theta_i}+\frac{\partial}{\partial\theta_i}\right)\right]+C,\end{aligned} \tag{5.9}$$

where the constant $C=-2N[\nu(N-1)-\nu_2-1]$. Again, like for the other supersymmetric extensions the ground-state eigenfunction remains the same as in the bosonic case and is given by (5.4). Making the gauge rotation of the Hamiltonian (5.9):
$h^{(c)}_{sBCD}=-2\Psi_0^{-1}\mathscr{H}^{(c)}_{\mathrm{sBCD}}\Psi_0$, we get

$$\begin{aligned}h^{(c)}_{\mathrm{sBCD}}=&\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2}+2\nu\sum_{i<j}\left[\frac{1}{x_i-x_j}\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}\right)-\frac{(\theta_i-\theta_j)}{(x_i-x_j)^2}\left(\frac{\partial}{\partial\theta_i}-\frac{\partial}{\partial\theta_j}\right)\right]\\&+2\nu\sum_{i<j}\left[\frac{1}{x_i+x_j}\left(\frac{\partial}{\partial x_i}+\frac{\partial}{\partial x_j}\right)-\frac{(\theta_i+\theta_j)}{(x_i+x_j)^2}\left(\frac{\partial}{\partial\theta_i}+\frac{\partial}{\partial\theta_j}\right)\right]\\&-2\omega\sum_{i=1}^{N}\left[x_i\frac{\partial}{\partial x_i}+\theta_i\frac{\partial}{\partial\theta_i}\right]+\nu_2\sum_{i=1}^{N}\left[\frac{1}{x_i}\frac{\partial}{\partial x_i}-\frac{\theta_i}{x_i^2}\frac{\partial}{\partial\theta_i}\right].\end{aligned} \tag{5.10}$$

We call this operator the *rational* form of the supersymmetric BC_N, B_N, C_N Calogero systems.

The superspace coordinates χ_i which are invariant under the permutation and parity transformations are defined through the relation

$$\sum_{k=0}^{N} \chi_k(\phi_i^2)t^k = \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} s_{2n}(\phi_i)t^n\right], \tag{5.11}$$

[cf. (5.5) and (3.19)], where $\phi_i = x_i + \alpha\theta_i$ as in Sections III and IV, and $\chi_k(\phi_i^2) = \tilde{\sigma}_k(x_i^2) + \alpha\tilde{\zeta}_k(x_i^2, 2x_i\theta_i)$. Here $\tilde{\sigma}$ and $\tilde{\zeta}$ are the elementary symmetric polynomials defined in (3.9) but of the new arguments: $x \to x^2, \theta \to 2x\theta$. Using the same technique as in Sections III and IV, one can find the representation of the supersymmetric models directly from the analogous representation for the bosonic cases (for details see Appendix C). Finally, in the superspace coordinates (5.11) the Hamiltonian $h^{(c)}_{\mathrm{sBCD}}$ has the form

$$h^{(c)}_{\mathrm{sBCD}} = \int d\alpha d\bar{\alpha} \sum_{i,j=1}^{N} \mathscr{A}_{ij} \frac{\partial}{\partial\bar{\chi}_i} \frac{\partial}{\partial\chi_j} + \int d\alpha \sum_{j=1}^{N} \mathscr{B}_j \frac{\partial}{\partial\chi_j}, \tag{5.12}$$

where

$$\begin{aligned} \mathscr{A}_{ij} &= 2\sum_{l\geqslant 0} \{2(2l+1+j-i)\chi_{i-l-1}\bar{\chi}_{j+l} - l \\ &\quad \times [\chi_{j+l}\bar{\chi}_{i-l-1} - \bar{\chi}_{j+l}\chi_{i-l-1} + \chi_{i+l-1}\bar{\chi}_{j-l} - \bar{\chi}_{i+l-1}\chi_{j-l}]\}, \\ \mathscr{B}_j &= 2[1+\nu_2+2\nu(N-j)](N-j+1)\chi_{j-1} - 4\omega j\chi_j . \end{aligned} \tag{5.13}$$

For completeness we give the form of $h^{(c)}_{\mathrm{sBCD}}$ in components

$$\begin{aligned} h^{(c)}_{\mathrm{sBCD}} &= \sum_{i,j=1}^{N} \left[A_{ij}^{\tilde{\sigma}\tilde{\sigma}} \frac{\partial^2}{\partial\tilde{\sigma}_i\partial\tilde{\sigma}_j} + A_{ij}^{\tilde{\sigma}\tilde{\zeta}} \frac{\partial^2}{\partial\tilde{\sigma}_i\partial\tilde{\zeta}_j} + A_{ij}^{\tilde{\zeta}\tilde{\sigma}} \frac{\partial^2}{\partial\tilde{\zeta}_i\partial\tilde{\sigma}_j} + A_{ij}^{\tilde{\zeta}\tilde{\zeta}} \frac{\partial^2}{\partial\tilde{\zeta}_j\partial\tilde{\zeta}_i} \right] \\ &\quad + \sum_{j=1}^{N} \left\{ 2[1+\nu_2+2\nu(N-j)](j-N+1) \left[\tilde{\sigma}_{j-1}\frac{\partial}{\partial\tilde{\sigma}_j} + \tilde{\zeta}_{j-1}\frac{\partial}{\partial\tilde{\zeta}_j} \right] \right. \\ &\quad \left. - 4\omega j \left[\tilde{\sigma}_j\frac{\partial}{\partial\tilde{\sigma}_j} + \tilde{\zeta}_j\frac{\partial}{\partial\tilde{\zeta}_j} \right] \right\}, \end{aligned} \tag{5.14}$$

where

$$A_{ij}^{\tilde{\sigma}\tilde{\sigma}} = 4\sum_{l\geqslant 0} (2l+1+j-i)\tilde{\sigma}_{i-l-1}\tilde{\sigma}_{j+l},$$

$$\begin{aligned} A_{ij}^{\tilde{\sigma}\tilde{\zeta}} &= 2\sum_{l\geqslant 0} [2(2l+1+j-i)\tilde{\sigma}_{i-l-1}\tilde{\zeta}_{j+l} \\ &\quad - l[\tilde{\sigma}_{j-l-1}\tilde{\zeta}_{i+l} - \tilde{\sigma}_{i+l}\tilde{\zeta}_{j-l-1} + \tilde{\sigma}_{j-l}\tilde{\zeta}_{i+l-1} - \tilde{\sigma}_{i+l-1}\tilde{\zeta}_{j-l}]] \\ A_{ij}^{\tilde{\zeta}\tilde{\sigma}} &= 2\sum_{l\geqslant 0} [2(2l+1+j-i)\tilde{\zeta}_{i-l-1}\tilde{\sigma}_{j+l} \\ &\quad + l[\tilde{\sigma}_{j-l-1}\tilde{\zeta}_{i+l} - \tilde{\sigma}_{i+l}\tilde{\zeta}_{j-l-1} + \tilde{\sigma}_{j-l}\tilde{\zeta}_{i+l-1} - \tilde{\sigma}_{i+l-1}\tilde{\zeta}_{j-l}]] \\ A_{ij}^{\tilde{\zeta}\tilde{\zeta}} &= 4\sum_{l\geqslant 0} [(2l+1+j-i)\tilde{\zeta}_{i-l-1}\tilde{\zeta}_{j+l} - l\tilde{\zeta}_{i+l}\tilde{\zeta}_{j-l-1} - l\tilde{\zeta}_{i+l-1}\tilde{\zeta}_{j-l}]. \end{aligned} \tag{5.15}$$

From the expression (5.14) with the coefficients (5.15) it follows that $h^{(c)}_{sBCD}$ can be written as a quadratic polynomial in the generators of a Borel subalgebra of $gl(N+1|N)$ and, hence, the BC_N Calogero model is exactly solvable.

The existence of the representation of (5.14) in terms of the generators of a Borel subalgebra of $gl(N+1|N)$ (see (B6)), proves that the operator $h^{(c)}_{sBCD}$ has infinitely many eigenfunctions having the form of polynomials. These polynomials are related to finite-dimensional irreducible representations of the algebra $gl(N+1|N)$ in the realization (A3) and can be called the supersymmetric BC_N Calogero polynomials.

So the supersymmetric BC_N, B_N, C_N and D_N Calogero models possess the algebraic form and also the Lie-algebraic form (B6) being represented by second-order polynomials in the generators of the of the algebra $gl(N+1|N)$ with certain coefficients.

VI. THE BC_N, B_N, C_N AND D_N SUTHERLAND MODELS AND THEIR SUPERSYMMETRIC EXTENSIONS

Similar to what was done in Section V for the BC_N, B_N, C_N and D_N Calogero models the present section is devoted to a consideration of the BC_N, B_N, C_N and D_N Sutherland models.

A. The bosonic case

The Hamiltonians for BC_N, B_N, C_N, D_N Sutherland models are special cases of the general BC_N Hamiltonian[6]

$$\mathscr{H}^{(s)}_{BCD} = -\frac{1}{2}\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2} + \frac{g}{4}\sum_{i<j}\left[\frac{1}{\sin^2(\frac{1}{2}(x_i-x_j))} + \frac{1}{\sin^2(\frac{1}{2}(x_i+x_j))}\right] + \frac{g_2}{4}\sum_{i=1}^{N}\frac{1}{\sin^2(x_i)} + \frac{g_3}{8}\sum_{i=1}^{N}\frac{1}{\sin^2\left(\frac{x_i}{2}\right)}, \tag{6.1}$$

where $g=\nu(\nu-1)$, $g_2=\nu_2(\nu_2-1)$ and $g_3=\nu_3(\nu_3+2\nu_2-1)$. From the general Hamiltonian the B_N, C_N and D_N cases are obtained as follows:

(i) B_N case: $\nu_2=0$,
(ii) C_N case: $\nu_3=0$, and
(iii) D_N case: $\nu_2=\nu_3=0$.

The ground state wave function is[9,15]

$$\Psi_0 = \left[\prod_{i<j}\left|\sin\left(\frac{1}{2}(x_i-x_j)\right)\right|^{\nu}\left|\sin\left(\frac{1}{2}(x_i+x_j)\right)\right|^{\nu}\prod_{i=1}^{N}|\sin(x_i)|^{\nu_2}\left|\sin\left(\frac{x_i}{2}\right)\right|^{\nu_3}\right]. \tag{6.2}$$

Again, one should emphasize (see also the discussion in Sections II and V A) that to a fixed value of the coupling constant $g(g_2)[g_3]$ there corresponds two different values of the parameter $\nu(\nu_2)[\nu_3]$: $\nu(\nu_2)[\nu_3]=\alpha(\alpha_2)[\alpha_3]$ and $1-\alpha(1-\alpha_2)[1-\alpha_3]$ giving rise to eight families of eigenfunctions. In order to get the ground state the parameters α,α_2,α_3 should be chosen in such a way as to minimize the eigenvalue and then (6.2) corresponds to the ground state. It can be denoted $(\alpha,\alpha_2,\alpha_3)$. The other values of ν,ν_2,ν_3, if inserted in (6.2), describe (provided that the corresponding wave functions are normalizable) the ground states of the remaining seven *other* families of eigenstates:

$$(\alpha,\alpha_2,1-\alpha_3),(\alpha,1-\alpha_2,\alpha_3),(\alpha,1-\alpha_2,1-\alpha_3),$$

$$(1-\alpha,\alpha_2,\alpha_3),(1-\alpha,1-\alpha_2,\alpha_3),(1-\alpha,\alpha_2,1-\alpha_3),$$

$$(1-\alpha,1-\alpha_2,1-\alpha_3).$$

So taking different values of ν,ν_2,ν_3 in (6.2) one can exhaust all the types of ground states corresponding to the different families mentioned above. In what follows, for the sake of simplicity, we continue to call (6.2) the ground state, of course, keeping in mind the above discussion.

Using the same approach as in Sections II–IV, we make a gauge rotation of (6.1) with the ground-state eigenfunction as gauge factor, $h^{(s)}_{BCD}=-2\Psi_0^{-1}\mathscr{H}^{(s)}_{BCD}\Psi_0$. A straightforward calculation leads to the operator (we omit an additive constant)

$$h^{(s)}_{BCD}=\sum_{i=1}^{N}\partial_i^2+\nu\sum_{i<j}\left[\cot\left(\frac{1}{2}(x_i-x_j)\right)(\partial_i-\partial_j)+\cot\left(\frac{1}{2}(x_i+x_j)\right)(\partial_i+\partial_j)\right]$$
$$+\nu_2\sum_{i=1}^{N}\cot(x_i)\partial_i+\frac{\nu_3}{2}\sum_{i=1}^{N}\cot\left(\frac{x_i}{2}\right)\partial_i. \tag{6.3}$$

It is worth mentioning that if the operator (6.3) is written in the coordinates e^{ix}, it appears in its *rational* form[15] (cf. the rational form of the A_{N-1} Sutherland Hamiltonian in Ref. 2).

According to the above discussion about eight different families of eigenfunctions, it is sufficient to study the operator (6.3) for generic ν,ν_2,ν_3 and consider only eigenfunctions which are symmetric under reflections. In particular, this implies that eigenfunctions of (6.3) have the form

$$P^{\nu,\nu_2,\nu_3}=F(\cos x_1,\cos x_2,\ldots,\cos x_N), \tag{6.4}$$

where F is permutationally symmetric.

Next we encode the symmetry properties of the problem studied by introducing the permutation and reflection-invariant periodic variables $\hat\sigma_k(\cos x)$, which satisfy

$$\sum_{k=0}^{N}\hat\sigma_k(\cos(x_i))t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_n(\cos(x_i))t^n\right], \tag{6.5}$$

[cf. (4.8), (5.11)]. Let us emphasize that these variables (6.5) are characterized by the same period as the original Hamiltonian (6.1). In these variables the Hamiltonian $h^{(s)}_{BCD}$ becomes

$$h^{(s)}_{BCD}=\sum_{i,j=1}^{N}A_{ij}\frac{\partial}{\partial\hat\sigma_i}\frac{\partial}{\partial\hat\sigma_j}+\sum_{j=1}^{N}B_j\frac{\partial}{\partial\hat\sigma_j}, \tag{6.6}$$

where

$$A_{ij}=N\hat\sigma_{i-1}\hat\sigma_{j-1}-\sum_{l\geqslant 0}[(i-l)\hat\sigma_{i-l}\hat\sigma_{j+l}+(l+j-1)\hat\sigma_{i-l-1}\hat\sigma_{j+l-1}$$
$$-(i-2-l)\hat\sigma_{i-2-l}\hat\sigma_{j+l}-(l+j+1)\hat\sigma_{i-l-1}\hat\sigma_{j+l+1}]$$
$$B_j=\frac{\nu_3}{2}(j-N-1)\hat\sigma_{j-1}-\left[\nu_2+\frac{\nu_3}{2}+1+\nu(2N-j-1)\right]j\hat\sigma_j-\nu(N-j+1)(N-j+2)\hat\sigma_{j-2}, \tag{6.7}$$

and $\hat\sigma_k=0$, when $k<0$ or $k>N$. The method of the calculation of the coefficients A_{ij},B_j is similar to that presented in Appendix C. This expression can be called the *algebraic* form of the BC_N,B_N,C_N and D_N Sutherland Hamiltonians. Similar to what happened in all previously discussed bosonic cases the coefficients A_{ij},B_j are polynomials of second and first degree in $\hat\sigma_k$, respectively. Hence, the Hamiltonian $h^{(s)}_{BCD}$ can be written in terms of generators of the Borel subalgebra of $gl(N+1)$ realized as first order differential operators (see Appendix A1) as in Sections II and V A. The result is given in Eq. (B7) (see Appendix B).

Then in accordance with the general definition given in Ref. 11 we conclude that all BC_N,B_N,C_N,D_N Sutherland models (6.1) are exactly solvable. The existence of the representation (6.6) proves that there are infinitely many eigenfunctions of the operator (5.3) having the form of polynomials in the variables $\hat\sigma_k$. It also implies that totally (anti)symmetric-(anti)symmetric

eigenfunctions with respect to permutations and reflections of the BC_N, B_N, C_N and D_N Sutherland models (6.1) have a factorizable form being the product of the ground-state eigenfunction (6.2) multiplied by a polynomial in the variables $\hat{\sigma}_k$. These polynomials are related to finite-dimensional irreducible representations of the Lie algebra $gl(N+1)$ in the realization (A1). They can be called the BC_N Jack–Sutherland polynomials.

B. The supersymmetric extensions

The bosonic BC_N, B_N, C_N and D_N Sutherland models (6.1) have natural supersymmetric extensions. In this subsection we will construct these models and show that, as was the case for the other supersymmetric extensions discussed in Sections III, IV and V B, these models are also exactly solvable. Let us introduce the supercharges (3.3) with the superpotential W given by

$$W=\nu\sum_{i<j}\left[\log\left|\sin\left(\frac{1}{2}(x_i-x_j)\right)\right|+\log\left|\sin\left(\frac{1}{2}(x_i+x_j)\right)\right|\right]+\nu_2\sum_{i=1}^{N}\log|\sin(x_i)|+\nu_3\sum_{i=1}^{N}\log\left|\sin\left(\frac{x_i}{2}\right)\right| \tag{6.8}$$

[cf. (3.4), (5.8)]. From the general formula $\mathscr{H}^{(s)}_{sBCD}=\frac{1}{2}\{Q,Q^\dagger\}$, we derive the supersymmetric Hamiltonians of the BC_N, B_N, C_N and D_N Sutherland models:

$$\begin{aligned}\mathscr{H}^{(s)}_{sBCD}=&-\frac{1}{2}\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2}+\frac{\nu}{4}\sum_{i<j}\left[\frac{1}{\sin^2(\frac{1}{2}(x_i-x_j))}\left[\nu-1+(\theta_k-\theta_l)\left(\frac{\partial}{\partial\theta_k}-\frac{\partial}{\partial\theta_l}\right)\right]\right.\\&\left.+\frac{1}{\sin^2(\frac{1}{2}(x_i+x_j))}\left[\nu-1+(\theta_k+\theta_l)\left(\frac{\partial}{\partial\theta_k}+\frac{\partial}{\partial\theta_l}\right)\right]\right]+\sum_{i=1}^{N}\left\{\frac{\nu_3}{8}\left[\frac{1}{\sin^2\left(\frac{x_i}{2}\right)}\right.\right.\\&\left.\left.\times\left[\nu_3-1\,\theta_i\frac{\partial}{\partial\theta_i}\right]\right]+\frac{\nu_2}{2}\left[\frac{1}{\sin^2(x_i)}\left[\nu_2+2\nu_3-1+\theta_i\frac{\partial}{\partial\theta_i}\right]\right]\right\}+C,\end{aligned} \tag{6.9}$$

where $C=-\frac{1}{2}\sum_{i=1}^{N}[\nu(N-i)+\nu_2+\nu_3/2]^2$. As should be familiar by now, the next step is to introduce the gauge-rotated Hamiltonian $h^{(s)}_{sBCD}=-2\Psi_0^{-1}\mathscr{H}^{(s)}_{sBCD}\Psi_0$. After some calculations we get

$$\begin{aligned}h^{(s)}_{BCD}=&\sum_{i=1}^{N}\partial_i^2+\nu\sum_{i<j}\left[\cot\left(\frac{1}{2}(x_i-x_j)\right)\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}\right)-\frac{1}{2}\frac{(\theta_i-\theta_j)}{\sin^2(\frac{1}{2}(x_i-x_j))}\left(\frac{\partial}{\partial\theta_i}-\frac{\partial}{\partial\theta_j}\right)\right]\\&+\nu\sum_{i<j}\left[\cot\left(\frac{1}{2}(x_i+x_j)\right)\left(\frac{\partial}{\partial x_i}+\frac{\partial}{\partial x_j}\right)-\frac{1}{2}\frac{(\theta_i+\theta_j)}{\sin^2(\frac{1}{2}(x_i+x_j))}\left(\frac{\partial}{\partial\theta_i}+\frac{\partial}{\partial\theta_j}\right)\right]\\&+\nu_2\sum_{i=1}^{N}\left[\cot(x_i)\frac{\partial}{\partial x_i}-\frac{\theta_i}{\sin^2(x_i)}\frac{\partial}{\partial\theta_i}\right]+\frac{\nu_3}{2}\sum_{i=1}^{N}\left[\cot\left(\frac{x_i}{2}\right)\frac{\partial}{\partial x_i}-\frac{1}{2}\frac{\theta_i}{\sin^2\left(\frac{x_i}{2}\right)}\frac{\partial}{\partial\theta_i}\right].\end{aligned} \tag{6.10}$$

It is worth mentioning that if the operator (6.10) is written in the coordinates e^{ix} it appears in its *rational* form. This expression should be compared to the rational form of the supersymmetric A_N Sutherland model (4.3) and the BC_N Sutherland model (6.3).

As in Sections III, IV and V B it is convenient to use the superspace formalism and to introduce the superspace coordinates χ_i, which satisfies:

$$\sum_{k=0}^{N}\chi_k(\cos(\phi_i))t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_n(\cos(\phi_i))t^n\right]. \tag{6.11}$$

Here $\chi_k(\cos(\phi_i))=\hat{\sigma}_k(\cos x_i)+\alpha\hat{\zeta}_k(\cos x_i,-\theta_i \sin x_i)$, where $\hat{\sigma}_k$ and $\hat{\zeta}_k$ are the elementary symmetric polynomials defined in (3.9) but of new arguments: $x\to\cos x$, $\theta\to-\theta \sin x$. In terms of these variables the Hamiltonian (6.10) becomes

$$h^{(s)}_{sBCD}=\int d\alpha d\overline{\alpha}\sum_{i,j=1}^{N}\mathscr{A}_{ij}\frac{\partial}{\partial\overline{\chi}_i}\frac{\partial}{\partial\chi_j}+\int d\alpha\sum_{j=1}^{N}\mathscr{B}_j\frac{\partial}{\partial\chi_j}, \tag{6.12}$$

where

$$\begin{aligned}\mathscr{A}_{ij}=&N\chi_{i-1}\overline{\chi}_{j-1}-\sum_{l\geqslant 0}[(i-l)\chi_{i-l}\overline{\chi}_{j+l}+(l+j-1)\chi_{i-l-1}\overline{\chi}_{j+l-1}\\&-(i-2-l)\chi_{i-2-l}\overline{\chi}_{j+l}-(l+j+1)\chi_{i-l-1}\overline{\chi}_{j+l+1}+l[\chi_{i+l-1}\overline{\chi}_{j-l-1}-\overline{\chi}_{i+l-1}\overline{\chi}_{j-l-1}\\&+\chi_{i+l+1}\chi_{j-l-3}-\chi_{i+l+1}\overline{\chi}_{j-l-3}+\chi_{i+l+1}\overline{\chi}_{j-l-1}-\overline{\chi}_{i+l+1}\chi_{j-l-1}]],\\ \mathscr{B}_j=&\frac{\nu_3}{2}(j-N-1)\chi_{j-1}-\left(\nu_2+\frac{\nu_3}{2}+1+\nu(2N-j-1)\right)j\chi_j-\nu(N-j+1)(N-j+2)\chi_{j-2}.\end{aligned} \tag{6.13}$$

The method of the calculation of the coefficients $\mathscr{A}_{ij}$, $\mathscr{B}_j$ is similar to that presented in Appendix C. In components $h^{(s)}_{sBCD}$ becomes

$$\begin{aligned}h^{(s)}_{sBCD}=&\sum_{i,j=1}^{N}\left[A^{\hat{\sigma}\hat{\sigma}}_{ij}\frac{\partial^2}{\partial\hat{\sigma}_i\partial\hat{\sigma}_j}+A^{\hat{\sigma}\hat{\zeta}}_{ij}\frac{\partial^2}{\partial\hat{\sigma}_i\partial\hat{\zeta}_j}+A^{\hat{\zeta}\hat{\sigma}}_{ij}\frac{\partial^2}{\partial\hat{\zeta}_i\partial\hat{\sigma}_j}+A^{\hat{\zeta}\hat{\zeta}}_{ij}\frac{\partial^2}{\partial\hat{\zeta}_j\partial\hat{\zeta}_i}\right]\\&+\sum_{j=1}^{N}\left\{\frac{\nu_3}{2}(j-N-1)\left[\hat{\sigma}_{j-1}\frac{\partial}{\partial\hat{\sigma}_j}+\hat{\zeta}_{j-1}\frac{\partial}{\partial\hat{\zeta}_j}\right]-\left(\nu_2+\frac{\nu_3}{2}+1+\nu(2N-j-1)\right)j\right.\\&\left.\times\left[\hat{\sigma}_j\frac{\partial}{\partial\hat{\sigma}_j}+\hat{\zeta}_j\frac{\partial}{\partial\hat{\zeta}_j}\right]-\nu(N-j+1)(N-j+2)\left[\hat{\sigma}_{j-2}\frac{\partial}{\partial\hat{\sigma}_j}+\hat{\zeta}_{j-2}\frac{\partial}{\partial\hat{\zeta}_j}\right]\right\},\end{aligned} \tag{6.14}$$

where

$$\begin{aligned}A^{\hat{\sigma}\hat{\sigma}}_{ij}=&N\hat{\sigma}_{i-1}\hat{\sigma}_{j-1}-\sum_{l\geqslant 0}[(i-l)\hat{\sigma}_{i-l}\hat{\sigma}_{j+l}+(l+j-1)\hat{\sigma}_{i-l-1}\hat{\sigma}_{j+l-1}\\&-(i-2-l)\hat{\sigma}_{i-2-l}\hat{\sigma}_{j+l}-(l+j+1)\hat{\sigma}_{i-l-1}\hat{\sigma}_{j+l+1}],\\ A^{\hat{\sigma}\hat{\zeta}}_{ij}=&N\hat{\sigma}_{i-1}\hat{\zeta}_{j-1}-\sum_{l\geqslant 0}[(i-l)\hat{\sigma}_{i-l}\hat{\zeta}_{j+l}+(l+j-1)\hat{\sigma}_{i-l-1}\hat{\zeta}_{j+l-1}\\&-(i-2-l)\hat{\sigma}_{i-2-l}\hat{\zeta}_{j+l}-(l+j+1)\hat{\sigma}_{i-l-1}\hat{\zeta}_{j+l+1}\\&-l[\hat{\sigma}_{j-l-1}(\hat{\zeta}_{i+l-1}+\hat{\zeta}_{i+l+1})+\hat{\sigma}_{i+l+1}(\hat{\zeta}_{j-l-3}-\hat{\zeta}_{j-l-1})]]\\ A^{\hat{\zeta}\hat{\sigma}}_{ij}=&N\hat{\zeta}_{i-1}\hat{\sigma}_{j-1}-\sum_{l\geqslant 0}[(i-l)\hat{\zeta}_{i-l}\hat{\sigma}_{j+l}+(l+j-1)\hat{\zeta}_{i-l-1}\hat{\sigma}_{j+l-1}\\&-(i-2-l)\hat{\zeta}_{i-2-l}\hat{\sigma}_{j+l}-(l+j+1)\hat{\zeta}_{i-l-1}\hat{\sigma}_{j+l+1}\\&+l[\hat{\sigma}_{j-l-1}(\hat{\zeta}_{i+l-1}+\hat{\zeta}_{i+l+1})+\hat{\sigma}_{i+l+1}(\hat{\zeta}_{j-l-3}-\hat{\zeta}_{j-l-1})]],\end{aligned} \tag{6.15}$$

$$A_{ij}^{\hat{\xi}\hat{\xi}}=N\hat{\xi}_{i-1}\hat{\xi}_{j-1}-\sum_{l\geqslant 0}\,[(i-l)\hat{\xi}_{i-l}\hat{\xi}_{j+l}+(l+j-1)\hat{\xi}_{i-l-1}\hat{\xi}_{j+l-1}$$

$$-(i-2-l)\hat{\xi}_{i-2-l}\hat{\xi}_{j+l}-(l+j+1)\hat{\xi}_{i-l-1}\hat{\xi}_{j+l+1}$$

$$+l[\hat{\xi}_{i+l-1}\hat{\xi}_{j-l-1}-\hat{\xi}_{i+l+1}\hat{\xi}_{j-l-3}+2\hat{\xi}_{i+l+1}\hat{\xi}_{j-l-1}]].$$

From the expression (6.14) with the coefficients (6.15) it follows, that $h^{(s)}_{sBCD}$ can be written as a quadratic polynomial in the generators of a Borel subalgebra of $gl(N+1|N)$ and, hence, the BC_N Sutherland model is exactly solvable.

The existence of the representation of (6.14) in terms of the generators of a Borel subalgebra of $gl(N+1|N)$ (see (B8)) proves that there are infinitely many eigenfunctions of (6.14) having the form of polynomials. These polynomials are related to finite−dimensional irreducible representations of the algebra $gl(N+1|N)$ in the realization (A3). These polynomials can be called the supersymmetric BC_N Sutherland polynomials.

So the supersymmetric BC_N, B_N, C_N and D_N Sutherland models possess an algebraic form and also the Lie-algebraic form (B8) being represented by second-order polynomials in the generators of the of the algebra $gl(N+1|N)$ with certain coefficients.

VII. CONCLUSION

In this paper we have described the rational, algebraic and Lie−algebraic forms of the integrable $A_N-B_N-C_N-D_N$ rational (Calogero) and trigonometric (Sutherland) Hamiltonians in addition to the superspace expressions for the supersymmetric generalizations of the these models. We have shown that

All Hamiltonians of the integrable $A_N-B_N-C_N-D_N$ rational (Calogero) and trigonometric (Sutherland) models possess the **same** *hidden algebra $gl(N+1)$ and can be represented by second-degree polynomials in the generators of a Borel subalgebra of the $gl(N+1)$−algebra. If the configuration space is parametrized by permutationally symmetric coordinates v, the Hamiltonians have a triangular form and preserve the flag of spaces of inhomogeneous polynomials*

$$\mathscr{P}_n=\ \mathrm{span}\left\{ v_1^{n_1}v_2^{n_2}v_3^{n_3}\ldots v_N^{n_N}\,\middle|\, 0\leqslant\sum\ n_i\leqslant n\right\}.$$

Consequently, each Hamiltonian possesses one or several infinite families of polynomial eigenfunctions.

All Hamiltonians of the supersymmetric generalizations of the $A_N-B_N-C_N-D_N$ rational (Calogero) and trigonometric (Sutherland) models possess the **same** *hidden algebra $gl(N+1|N)$ and can be represented by second-degree polynomials in the generators of a Borel subalgebra of the $gl(N+1|N)$-algebra. If the configuration space is parametrized by permutationally symmetric coordinates v,κ (where $\kappa^2=0$), the Hamiltonians have a triangular form and preserve the flag of spaces of inhomogeneous polynomials*

$$\mathscr{P}_n=\ \mathrm{span}\left\{ v_1^{n_1}v_2^{n_2}\ldots v_N^{n_N}\kappa_1^{m_1}\kappa_2^{m_2}\ldots\kappa_N^{m_N}\,\middle|\, 0\leqslant\sum\ n_i+\sum\ m_i\leqslant n, m_i=0,1\right\}.$$

Consequently, each Hamiltonian has one or several infinite families of polynomial eigenfunctions. The integrability of the A_N supersymmetric system was proven in Ref. 5; *as for the $B_N-C_N-D_N$ systems, it is an open question.*

As an interesting issue for future study we would like to mention the question about whether there exists a Lie algebraic description of the higher $A_N-B_N-C_N-D_N$ rational (Calogero) and trigonometric (Sutherland) Hamiltonians. It is very probable that this will be the case and, perhaps, even the Lax operator can be represented in terms of $gl(N)$-generators. There are almost no doubts that it will be possible to extend the analysis of Sections II–VI to the case of the exceptional simple Lie algebras. In particular, a special case of the one-parametric $G(2)$ rational and trigonometric Hamiltonians corresponding to three-body interactions only[22,23] possesses a hidden

$gl(3)$-algebra and, hence, is expressible in terms of the $gl(3)$ generators.[23] Recently, a Lie algebraic form for the general $G(2)$ rational and trigonometric models was found.[24]

The question about the integrability of the above supersymmetric models leads naturally to the question about the possibility to develop the Hamiltonian reduction method for the supersymmetric cases and attempt to connect these models to $2d$ supersymmetric gauge theories along the lines of Ref. 10.

An alternative description of the supersymmetric models discussed in this paper can be obtained by using the matrix representation of the θ^α's in terms of the Pauli matrices, $\sigma^{\pm,3}$:

$$\theta^1=\sigma^-\otimes 1\otimes 1\otimes\cdots,\quad \frac{\partial}{\partial\theta^1}=\sigma^+\otimes 1\otimes 1\otimes\cdots,$$

$$\theta^2=\sigma^3\otimes\sigma^-\otimes 1\otimes\cdots,\quad \frac{\partial}{\partial\theta^2}=\sigma^3\otimes\sigma^+\otimes 1\otimes\cdots \tag{7.1}$$

$$\vdots$$

In this formalism our supersymmetric models become matrix generalizations of the ''scalar'' $A_N-B_N-C_N-D_N$ Calogero and Sutherland models. For the A_N case our matrix models are particular cases of the general construction with internal degrees of freedom proposed in Ref. 25.

The existence of the Lie algebraic form for the exactly solvable $A_N-B_N-C_N-D_N$ Calogero and Sutherland models can be considered as a good starting point to investigate whether there exist other exactly solvable problems and/or quasi-exactly solvable generalizations (for definitions see, for instance, Refs. 26 and 27). The first step in applying the Lie algebraic formalism to the search for quasi-exactly-solvable problems was taken in Ref. 28.

As a final comment we would like to mention that a great challenge in the subject is the search for solutions of the N-body elliptic model (for a description see, for example, Ref. 9). For the two-body case, the rational and algebraic forms have been known since Hermite's days. Recently, it was found that the 2-body case (the Lame operator) admits a Lie algebraic form with the hidden algebra $gl(2)$ and is a *quasi-exactly-solvable* problem[29] (see also Ref. 11). It gives a hint that the three-body (and more generally the N-body) elliptic problem has to be quasi-exactly solvable if any analytic solution exists whatsoever. However, all attempts to find a rational, algebraic or Lie algebraic form even for the three-body elliptic model have failed so far.

ACKNOWLEDGMENTS

A. T. thanks the Institute for Theoretical Physics, Chalmers University of Technology, where this work was initiated, for its kind hospitality. A.T. is grateful to the NORDITA for financial support during his visit to Göteborg.

APPENDIX A: REPRESENTATIONS OF THE LIE ALGEBRA $gl(N)$ AND THE LIE SUPERALGEBRA $gl(N|M)$

1. $gl(N)$

The generators of the Lie algebra $gl(N)$ can be realized by first order differential operators. For our purposes we need one of the simplest realizations of $gl(N)$ acting on the real (complex) space of dimension $(N-1)$. This is the space of minimal dimension where this algebra can act. The generators can be represented in the following form:

$$
\begin{aligned}
E_{0i} &= J_i^- = \frac{\partial}{\partial \tau_i}, i = 2,3,\ldots,N, \\
E_{ij} &= J_{i,j}^0 = \tau_i J_j^- = \tau_i \frac{\partial}{\partial \tau_j}, i,j = 2,3,\ldots,N, \\
E_{00} &= J^0 = n - \sum_{k=2}^{N} \tau_k \frac{\partial}{\partial \tau_k}, \\
E_{i0} &= J_i^+ = \tau_i J^0, i = 2,3,\ldots,N,
\end{aligned}
\tag{A1}
$$

where the parameter $n \in \mathbf{R}$ ($\mathbf{C}$). One of the generators, namely $J^0 + \Sigma_{p=2}^{N} J_{p,p}^0$, is proportional to a constant and, if it is removed, we end up with the algebra $sl(N)$. The generators $J_{i,j}^0$ form the algebra of the vector fields of $sl(N-1)$, which is a subalgebra of $gl(N)$. If n is a non-negative integer, the representation (A1) becomes the finite-dimensional representation acting on the space of polynomials in $(N-1)$ variables of the following type:

$$
V_n(t) = \text{span}\left\{ \tau_2^{n_2} \tau_3^{n_3} \tau_4^{n_4} \ldots \tau_N^{n_N} \middle| 0 \leqslant \sum n_i \leqslant n \right\}. \tag{A2}
$$

This representation corresponds to a Young tableau with one row and n blocks and is irreducible. The positive-root generators J_i^+ define the highest-weight vector. If the J_i^+'s are removed, the remaining generators form a Borel subalgebra.

2. $gl(N|M)$

Similar to what was done for the algebra $gl(N)$ one can construct a representation of the Lie superalgebra $gl(N|M)$ in terms of first order differential operators over the direct sum of a even space and an odd space. Again the simplest realization of $gl(N|M)$ act on the space $C^{(N-1|M)}$ spanned by the even variables $(\tau^i | i = 1,2,\ldots,N-1))$ and the odd variables $(\kappa^\gamma | \gamma = 1,2,\ldots,M)$ and is given by

$$
\begin{aligned}
E_{ij} &= J_{ij}^0 = \tau^i \frac{\partial}{\partial \tau^j}, \quad E_{00} = J^0 = -n + \sum_i \tau^j \frac{\partial}{\partial \tau^j} + \sum_\gamma \kappa^\gamma \frac{\partial}{\partial \kappa^\gamma}, \\
E_{i0} &= J_i^+ = \tau^i J^0, \quad E_{0i} = J_i^- = \frac{\partial}{\partial \tau^i}, \\
E_{\gamma i} &= Q_{\gamma i} = \kappa^\gamma \frac{\partial}{\partial \tau^i}, \quad E_{i\gamma} = \overline{Q}_{i\gamma} = \tau^i \frac{\partial}{\partial \kappa^\gamma}, \\
E_{0\gamma} &= Q_\gamma^- = \frac{\partial}{\partial \kappa^\gamma}, \quad E_{\gamma 0} = Q_\gamma^+ = \kappa^\gamma T^0, \\
E_{\gamma\beta} &= T_{\gamma\beta}^0 = \kappa^\gamma \frac{\partial}{\partial \kappa^\beta},
\end{aligned}
\tag{A3}
$$

where the parameter $n \in \mathbf{R}$ ($\mathbf{C}$). Here the indices run over the following values, $i = 1 \ldots N-1$ and $\gamma = 1 \ldots M$. The subset $\{E_{ij}, E_{\gamma i}, E_{i\gamma}, E_{\gamma\beta}\}$ form the ''vector field'' representation of $gl(N-1|M)$. The algebra has a $\mathbf{Z}_2$ gradation and it is clear that the elements with an odd number of Greek indices are odd and the elements with an even number are even. Using the composite index notation $I = (0,i,\gamma)$, the commutation relations of the superalgebra can be written [cf. the commutation relations in the defining $(N+M)\times(N+M)$ supermatrix representation, with $(E_{IJ})_{MN} = \delta_{IN}\delta_{JM}$]

$$[E_{IJ},E_{KL}]=\delta_{JK}E_{IL}-\delta_{IL}E_{KJ} \quad \text{(both even)},$$

$$[E_{IJ},E_{KL}]=\delta_{JK}E_{IL}-\delta_{IL}E_{KJ} \quad (E_{IJ}\text{even},E_{KL}\text{odd}),$$

$$\{E_{IJ},E_{KL}\}=\delta_{JK}E_{IL}+\delta_{IL}E_{KJ} \quad \text{(both odd)}. \tag{A4}$$

The Lie superalgebra $gl(N|M)$ is not simple. There are two different cases: (i) $N\neq M$ and (ii) $N=M$. Let us consider the first case, when $N\neq M$. Removing the unit element $\Sigma_{I=0}^{N+M}E_{II}$ we are left with the superalgebra $sl(N|M)$, which is simple. The ''diagonal'' generators for this case are given by

$$H_{II}=\begin{cases} E_{00}-E_{11}, & I=0 \\ E_{II}-E_{I+1,I+1}, & 1\leqslant I\leqslant N-1 \\ E_{II}+E_{I+1,I+1}, & I=N \\ E_{II}-E_{I+1,I+1}, & N+1\leqslant I\leqslant N+M-1. \end{cases} \tag{A5}$$

When $N=M$ the situation is slightly more complicated. There are two one-dimensional Abelian ideals, which must be removed in order to make the algebra simple.

The generators of $gl(N|M)$ defined in (A3) act in the space spanned by the monomials $\{\tau_1^{i_1}\cdots\tau_{N-1}^{i_{N-1}}\kappa_1^{\delta_1}\cdots\kappa_M^{\delta_M}\}$. Here δ_k equals either 0 or 1. When n (in the expression for T^0 above) is an integer we have finite-dimensional highest weight representations whose highest weight vectors satisfy $\Sigma_{k=1}^{N-1}i_k+\Sigma_{k=1}^{M}\delta_k=n$. These representations form a flag of spaces.

APPENDIX B: THE LIE ALGEBRAIC FORMS OF THE A_{N-1}, BC_N, B_N, C_N AND D_N HAMILTONIANS

In this Appendix we collect all Lie algebraic forms of the Hamiltonians for the (supersymmetric) A_{N-1}, BC_N, B_N, C_N and D_N Calogero and Sutherland models found in Ref. 12 and in the present paper.

(i) A_{N-1} *Calogero model*

The Lie algebraic form of the many-body Calogero model is the following:[12]

$$\begin{aligned} h_{\text{Cal}}=&\sum_{j=2}^{N}\left\{\frac{(N-j+1)(j-1)}{N}(J_{j-1,j})^2-2\sum_{\ell=1}^{j-1}\ell J_{j+\ell-1,j}J_{j-\ell-1,j}\right\} \\ &+2\sum_{2\leqslant k<j\leqslant N}\left\{\frac{(N-j+1)(k-1)}{N}J_{j-1,j}J_{k-1,k}-\sum_{\ell=1}^{k-1}(j-k+2\ell)J_{j+\ell-1,j}J_{k-\ell-1,k}\right\} \\ &-2\omega\sum_{k=2}^{N}kJ_{k,k}-\left(\frac{1}{N}+\nu\right)\sum_{k=2}^{N}(N-k+2)(N-k+1)J_{k-2,k}, \end{aligned} \tag{B1}$$

where the following notations are introduced for the $gl(N)$ generators (A1): $J_{ij}\equiv J_{ij}^0$, $J_{0,k}^0\equiv J_k^-$, $J_{1,k}^0\equiv 0$.

The supersymmetric Calogero model has the Lie algebraic form:

$$h^{(\mathrm{rel})}_{\mathrm{sCal}}=\sum_{i,j=2}^{N}\left\{\frac{(N-i+1)(j-1)}{N}[J_{i-1,j}J_{j-1,i}+J_{i-1,i}T_{j-1,j}+J_{j-1,j}T_{i-1,i}-T_{i-1,j}T_{j-1,i}]\right.$$
$$+\sum_{l>\max(1,j-i)}(j-i-2l)[J_{i+l-1,j}J_{j-l-1,i}+J_{i+l-1,i}T_{j-l-1,j}+J_{j-l-1,j}T_{i+l-1,i}$$
$$\left.-T_{i+l-1,j}T_{j-l-1,i}]\right\}+\sum_{l\geqslant 1}l[T_{j-2-l,j}J_{i+l,i}-T_{i+l,j}J_{j-2-l,i}+T_{i+l,i}J_{j-2-l,j}-T_{j-2-l,i}J_{i+l,j}$$
$$-2T_{i+l,j}T_{j-2-l,i}]-2\omega\sum_{j=2}^{N}j[J_{jj}+T_{jj}]-\left(\frac{1}{N}+\nu\right)\sum_{j=2}^{N}(N-j+2)(N-j+1)$$
$$\times[J_{j-2,j}+T_{j-2,j}], \tag{B2}$$

where the following conventions are used for the $gl(N|N-1)$ generators (A3): $J_{ij}\equiv J^0_{ij}=0$ and $T_{ij}\equiv T^0_{ij}0$, if i and/or j is less than 2 or greater than N.

(ii) A_{N-1} *Sutherland model*

The Lie algebraic form of the many-body Sutherland model in terms of the $gl(N)$ generators is the following:[12]

$$-h_{\mathrm{Suth}}=\sum_{j=1}^{N-1}\left\{\frac{j(N-j)}{N}(J_{j,j})^2-2\sum_{l=1}^{j}lJ_{j+l,j}J_{j-l,j}\right\}+\nu\sum_{l=1}^{N-1}l(N-l)J_{l,l}$$
$$+2\sum_{1\leqslant k<j\leqslant N-1}\left\{\frac{k(N-j)}{N}J_{j,j}J_{k,k}-\sum_{l=1}^{k}(j-k+2l)J_{j+l,j}J_{k-l,k}\right\}. \tag{B3}$$

Here we have used the identifications $J_{ij}\equiv J^0_{ij}$, $J_{0,i}\equiv J^-_i\equiv J_{N,i}$, $J_{1,i}=0$. The supersymmetric many-body Sutherland model written in terms of the $gl(N|N-1)$ generators has the form

$$-h^{(\mathrm{rel})}_{\mathrm{sSuth}}=\sum_{i,j=1}^{N-1}\left\{\frac{j(N-i)}{N}[J_{i,j}J_{j,i}+J_{i,i}T_{j,j}+J_{j,j}T_{i,i}-T_{i,j}T_{j,i}]\right.$$
$$\left.+\sum_{l>\max(1,j-i)}(j-i-2l)[J_{i+l,j}J_{j-l,i}+J_{i+l,i}T_{j-l,j}+J_{j-l,j}T_{i+l,i}-T_{i+l,j}T_{j-l,i}]\right\}$$
$$+\sum_{l\geqslant 1}l[T_{j-l-1,j}J_{i+l-1,i}-T_{i+l-1,j}J_{j-l-1,i}+T_{i+l-1,i}J_{j-l-1,j}-T_{j-l-1,i}J_{i+l-1,j}$$
$$-2T_{i+l-1,j}T_{j-l+1,i}]+\sum_{i=1}^{N-1}i(N-i)\left[\nu J_{i,i}+\left(\frac{2}{N}+\nu\right)T_{i,i}\right]. \tag{B4}$$

In (B4) $J_{ij}\equiv j^0_{ij}=0$ and $T_{i,j}\equiv T^0_{ij}=0$ are by definition zero when i and/or j is greater than $N-1$ or less than 1.

(iii) BC_N, B_N, C_N *and* D_N *Calogero models*

The bosonic BC_N, B_N, C_N and D_N Calogero models have the following Lie algebraic form in terms of the $gl(N+1)$ generators:

$$h^{(c)}_{BCD}=4\sum_{i,j=1}^{N}\sum_{l\geqslant 0}(2l+1+j-i)J_{i-l-1,j}J_{j+l,i}-4\omega\sum_{i=1}^{N}iJ_{i,i}$$
$$+2\sum_{i=1}^{N}\left\{[1+\nu_2+2\nu(N-j)](N-j+1)-\frac{N}{2}(N+3-2i)\right\}J_{i-1,i}, \tag{B5}$$

and the supersymmetric extensions have the form

$$h^{(c)}_{sBCD}=2\sum_{i,j=1}^{N}\sum_{l\geqslant 0}[2(2l+1+j-i)[J_{i-l-1,j}J_{j+l,i}+J_{i-l-1,j}T_{j+l,i}+T_{i-l-1,j}J_{j+l,i}$$
$$-T_{i-l-1,j}T_{j+l,i}]+l[T_{j-l-1,j}J_{i+l,i}-T_{i+l,j}J_{j-l-1,i}-T_{i+l-1,j}J_{j-l,i}+T_{j-l,j}J_{i+l-1,i}$$
$$-T_{j-l-1,i}J_{i+l,j}+T_{i+l,i}J_{j-l-1,j}+T_{i+l-1,i}J_{j-l,j}-T_{j-l,i}J_{i+l-1,j}+T_{i+l,j}T_{j-l-1,i}$$
$$+T_{i+l-1,j}T_{j-l,i}]]+\sum_{i=1}^{N}2[1+\nu_2+2\nu(N-j)](N-j+1)\cdot\left\{\frac{N}{2}(N+3-2i)T_{i-1,i}\right.$$
$$\left.-\frac{N}{2}(N+3-2i)J_{i-1,i}\right\}-4\omega\sum_{i=1}^{N}i[J_{i,i}+T_{i,i}] \tag{B6}$$

in terms of the $gl(N+1|N)$ generators. In the above two formulas, as well as in the two formulas below, we have used that $J_{ij}\equiv J^0_{ij}=0$ and $T_{ij}\equiv T^0_{ij}=0$ when i and/or j is greater than N or less than 1.

(iv) *BC_N, B_N, C_N and D_N Sutherland models*

Finally, the Lie algebraic form of the bosonic BC_N, B_N, C_N and D_N Sutherland models is

$$h^{(s)}_{BCD}=\sum_{i,j=1}^{N}\left[NJ_{i-1,j}J_{j-1,i}-\sum_{l\geqslant 0}[(i-l)J_{i-l,j}J_{j+l,i}+(l+j-1)J_{i-l-1,j}J_{j+l-1,i}-(i-2-l)\right.$$
$$\times J_{i-2-l,j}J_{j+l,i}-(l+j+1)J_{i-l-1,j}J_{j+l+1,i}]$$
$$+\sum_{j=1}^{N}\left[\frac{\nu_3}{2}(j-N-1)J_{j-1,j}-\left(\nu_2+\frac{\nu_3}{2}+1+N+\nu(2N-j-1)\right)jJ_{j,j}\right.$$
$$\left.\left.-[\nu(N-j+1)(N-j+2)+2N]J_{j-2,j}\right]\right] \tag{B7}$$

in terms of the $gl(N+1)$ generators, and the Lie algebraic form of their supersymmetric extensions is

$$h^{(s)}_{sBCD}=\sum_{i,j=1}^{N}\left\{N(J_{i-1,j}J_{j-1,i}-T_{i-1,j}T_{j-1,i})-\sum_{l\geqslant 0}[(i-l)[J_{i-l,j}J_{j+l,i}+T_{i-l,j}J_{j+l,i}+J_{i-l,j}T_{j+l,i}\right.$$
$$-T_{i-l,j}T_{j+l,i}]+(l+j-1)[J_{i-l-1,j}J_{j+l-1,i}+J_{i-l-1,j}T_{j+l-1,i}+T_{i-l-1,j}J_{j+l-1,i}$$
$$-T_{i-l-1,j}T_{j+l-1,i}]-(i-2-l)[J_{i-2-l,j}J_{j+l,i}+T_{i-2-l,j}J_{j+l,i}+J_{i-2-l,j}T_{j+l,i}$$
$$-T_{i-2-l,j}T_{j+l,i}]-(l+j+1)[J_{i-l-1,j}J_{j+l+1,i}+T_{i-l-1,j}J_{j+l+1,i}+J_{i-l-1,j}T_{j+l+1,i}$$
$$-T_{i-l-1,j}T_{j+l+1,i}]+l[J_{j-l-1,j}(T_{i+l-1,i}+T_{i+l+1,i})+J_{i+l+1,j}(T_{j-l-3,i}-T_{j-l-1,i})$$
$$-J_{j-l-1,i}(T_{i+l-1,j}+T_{i+l+1,j})-J_{i+l+1,i}(T_{j-l-3,j}-T_{j-l-1,j})-T_{I+l+1,j}(T_{j-l-1,i}$$
$$\left.-T_{j-l-3,i})-(T_{i+l-1,j}+T_{i+l-1,j})T_{j-l-1,i}]]\right\}+\sum_{j=1}^{N}\left[\frac{\nu_3}{2}(j-N-1)[J_{j-1,j}+T_{j-1,j}]\right.$$
$$-\left(\nu_2+\frac{\nu_3}{2}+1+\nu(2N-j-1)\right)j[J_{j,j}+T_{j,j}]-Nj[J_{j,j}-T_{j,j}]$$
$$\left.-[\nu(N-j+1)(N-j+2)](J_{j-2,j}+T_{j-2,j})-2N(J_{j-2,j}-T_{j-2,j})\right] \tag{B8}$$

in terms of the $gl(N+1|N)$ generators.

In conclusion, let us emphasize the fact that all Lie algebraic forms of the supersymmetric models (B2), (B4), (B6), (B8) contain no fermionic (odd) generators.

APPENDIX C: DERIVING THE LIE ALGEBRAIC FORMS: TECHNICAL DETAILS

In this Appendix we will give some details of the derivation of the Lie algebraic forms of the Hamiltonians discussed in this paper. We will exemplify the method for the BC_N Calogero models. The other cases are tackled in a similar way.

1. Bosonic BC_N Calogero model

Our final goal is to derive (5.6)–(5.7). Let us recall that the gauge-rotated BC_N Hamiltonian is given by

$$h^{(c)}_{BCD}=\sum_{i=1}^{N}\partial_i^2+2\nu\sum_{i<j}\left[\frac{1}{x_i-x_j}(\partial_i-\partial_j)+\frac{1}{x_i+x_j}(\partial_i+\partial_j)\right]+\nu_2\sum_{i=1}^{N}\frac{1}{x_i}\partial_i-2\omega\sum_{i=1}^{N}x_i\partial_i. \quad \text{(C1)}$$

After the change of variables $x_i\to\widetilde{\sigma}_i$ the Hamiltonian becomes

$$h^{(c)}_{BCD}=\sum_{\ell,k=1}^{N}\sum_{i=1}^{N}\frac{\partial\widetilde{\sigma}_\ell}{\partial x_i}\frac{\partial\widetilde{\sigma}_k}{\partial x_i}\frac{\partial}{\partial\widetilde{\sigma}_k}\frac{\partial}{\partial\widetilde{\sigma}_l}+\sum_{k=1}^{N}\sum_{i=1}^{N}\frac{\partial^2\widetilde{\sigma}_k}{\partial x_i^2}\frac{\partial}{\partial\widetilde{\sigma}_k}-2\omega\sum_{k=1}^{N}\sum_{i=1}^{N}x_i\frac{\partial\widetilde{\sigma}_k}{\partial x_i}\frac{\partial}{\partial\widetilde{\sigma}_k}$$
$$+\nu\sum_{k=1}^{N}\sum_{i\neq j}\left[\frac{1}{x_i+x_j}\left(\frac{\partial\widetilde{\sigma}_k}{\partial x_i}+\frac{\partial\widetilde{\sigma}_k}{\partial x_j}\right)+\frac{1}{x_i-x_j}\left(\frac{\partial\widetilde{\sigma}_k}{\partial x_i}-\frac{\partial\widetilde{\sigma}_k}{\partial x_j}\right)\right]\frac{\partial}{\partial\widetilde{\sigma}_k}+\nu_2\sum_{k=1}^{N}\sum_{i=1}^{N}\frac{1}{x_i}\frac{\partial\widetilde{\sigma}_k}{\partial x_i}\frac{\partial}{\partial\widetilde{\sigma}_k}. \quad \text{(C2)}$$

As an example, let us express

$$\widetilde{B}_k=\sum_{i\neq j}^{N}\left[\frac{1}{x_i+x_j}\left(\frac{\partial\widetilde{\sigma}_k}{\partial x_i}+\frac{\partial\widetilde{\sigma}_k}{\partial x_j}\right)+\frac{1}{x_i-x_j}\left(\frac{\partial\widetilde{\sigma}_k}{\partial x_i}-\frac{\partial\widetilde{\sigma}_k}{\partial x_j}\right)\right] \quad \text{(C3)}$$

and

$$\widetilde{A}_{\ell,k}=\sum_{i=1}^{N}\frac{\partial\widetilde{\sigma}_\ell}{\partial x_i}\frac{\partial\widetilde{\sigma}_k}{\partial x_i} \quad \text{(C4)}$$

in terms of the $\widetilde{\sigma}$ variables. The approach we will follow is to represent the generating functions of (C3) and (C4)

$$\widetilde{A}(t,p)=\sum_{k=0}^{N}\widetilde{A}_{\ell,k}t^\ell p^k,\quad \widetilde{B}(t)=\sum_{k=0}^{N}\widetilde{B}_k t^k,$$

in terms of the generating function of the $\widetilde{\sigma}$ variables

$$G(t)=\sum_{k=0}^{N}\widetilde{\sigma}_k(x)t^k=\exp\left[\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}s_{2n}(x)t^n\right]. \quad \text{(C5)}$$

Let us begin by calculating the generating function $\widetilde{B}(t)$ for (C3)

$$\widetilde{B}(t)=\sum_{i\neq j}\left[\frac{1}{x_i-x_j}\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}\right)+\frac{1}{x_i+x_j}\left(\frac{\partial}{\partial x_i}+\frac{\partial}{\partial x_j}\right)\right]G(t)$$
$$=2\sum_{i\neq j}\left\{\sum_{m=1}^{\infty}(-1)^{m+1}\left[\frac{x_i^{2m-1}+x_j^{2m-1}}{x_i+x_j}+\frac{x_i^{2m-1}-x_j^{2m-1}}{x_i-x_j}\right]t^m\right\}G(t). \quad \text{(C6)}$$

Using the relation

$$\frac{x_i^{2n-1}+x_j^{2n-1}}{x_i+x_j}+\frac{x_i^{2n-1}-x_j^{2n-1}}{x_i-x_j}=2\sum_{l=0}^{n-1}x_i^{2l}x_j^{2(n-1)-2l}, \quad \text{(C7)}$$

and making rather obvious mathematical transformations, one can show that

$$\widetilde{B}(t)=\left[4t^3\frac{\partial^2}{\partial t^2}-4(2N-2)t^2\frac{\partial}{\partial t}+4tN(N-1)\right]G(t). \tag{C8}$$

Substituting (C5) in (C8) we find that $\widetilde{B}_k$ in (C3) is

$$\widetilde{B}_k=4(N-k+1)(N-k)\widetilde{\sigma}_{k-1}. \tag{C9}$$

Next we proceed to the calculation of (C4). The corresponding generating function can be written as

$$\widetilde{A}(t,p)=\sum_{i=1}^{N}\frac{\partial G(p)}{\partial x_i}\frac{\partial G(t)}{\partial x_i}=4\sum_{i=1}^{N}\sum_{k,\ell=1}^{\infty}(-1)^{k+\ell}x_i^{2(k+\ell)-2}t^k p^\ell G(t)G(p). \tag{C10}$$

Changing the summation variables to $r=k+\ell$ and $u=k-\ell$ one obtains that

$$\widetilde{A}(t,p)=4\sum_{i=1}^{N}\sum_{r=2}^{\infty}(-1)^r s_{2(r-1)}(tp)^{\frac{r}{2}}\sum_{u}\left(\frac{t}{p}\right)^{u/2}G(t)G(p). \tag{C11}$$

The sum over u is easily determined

$$\sum_{u}\left(\frac{t}{p}\right)^{u/2}=\left(\frac{t}{p}\right)^{-r-2/2}\left(\frac{1-\left(\frac{t}{p}\right)^{r-1}}{1-\frac{t}{p}}\right). \tag{C12}$$

Using this formula it is easy to transform (C11) to the following expression for the generating function:

$$\widetilde{A}(t,p)=\frac{4t}{1-\frac{t}{p}}\left[t\frac{\partial}{\partial t}-p\frac{\partial}{\partial p}\right]G(t)G(p). \tag{C13}$$

Expanding the rhs of (C13) we arrive at the final expression for (C4),

$$\widetilde{A}_{\ell,k}=4\sum_{q\geqslant 0}(2q+1+l-k)\widetilde{\sigma}_{k-q-1}\widetilde{\sigma}_{\ell+q}. \tag{C14}$$

The other terms in (C2) can be calculated using the same method which finally leads to the results presented in (5.6) and (5.7).

2. Supersymmetric BC_N Calogero model

For the supersymmetric models the calculations are similar to those carried out for the bosonic models. In terms of the superspace ϕ-coordinates (3.11) the gauge rotated Hamiltonian (5.10) is written as

$$\begin{aligned}h^{(c)}_{sBCD}=&\int d\alpha d\bar{\alpha}\sum_{i=1}^{N}\frac{\delta}{\delta\phi_i(\bar{\alpha})}\frac{\delta}{\delta\phi_i(\alpha)}+2\nu\int d\alpha\sum_{i<j}\left[\frac{1}{\phi_i(\alpha)-\phi_j(\alpha)}\left(\frac{\delta}{\delta\phi_i(\alpha)}-\frac{\delta}{\delta\phi_j(\alpha)}\right)\right.\\&\left.+\frac{1}{\phi_i(\alpha)+\phi_j(\alpha)}\left(\frac{\delta}{\delta\phi_i(\alpha)}+\frac{\delta}{\delta\phi_j(\alpha)}\right)\right]+\nu_2\int d\alpha\sum_{i=1}^{N}\frac{1}{\phi_i(\alpha)}\frac{\delta}{\delta\phi_i(\alpha)}\\&-2\omega\int d\alpha\sum_{i=1}^{N}\phi_i(\alpha)\frac{\delta}{\delta\phi_i(\alpha)}.\end{aligned} \tag{C15}$$

Now we make the change of variables $\phi_i \to \chi_i$ (see (5.11)) using the relation (3.21). Finally, an expression similar to (C2) emerges except for the fact that the bosonic coordinates are replaced by supercoordinates with integration over the Grassmann variables:

$$h^{(c)}_{sBCD} = \sum_{i=1}^{N} \int d\alpha d\overline{\alpha} d\beta d\overline{\beta} \frac{\delta\chi_k(\overline{\beta})}{\delta\phi_i(\overline{\alpha})} \frac{\delta\chi_l(\beta)}{\delta\phi_i(\alpha)} \frac{\delta}{\delta\chi_l(\overline{\beta})} \frac{\delta}{\delta\chi_l(\beta)}$$
$$- 2\nu \int d\alpha d\beta \sum_{i<j} \sum_{k=1}^{N} \left[\frac{1}{\phi_i(\alpha) - \phi_j(\alpha)} \left(\frac{\delta\chi_k(\beta)}{\delta\phi_i(\alpha)} + \frac{\delta\chi_k(\beta)}{\delta\phi_j(\alpha)} \right) + \frac{1}{\phi_i(\alpha) + \phi_j(\alpha)} \right.$$
$$\left. \times \left(\frac{\delta\chi_k(\beta)}{\delta\phi_i(\alpha)} + \frac{\delta\chi_k(\beta)}{\delta\phi_j(\alpha)} \right) \right] \frac{\delta}{\delta\chi_k(\beta)} - \nu_2 \int d\alpha d\beta \sum_{i=1}^{N} \sum_{k=1}^{N} \frac{1}{\phi_i(\alpha)} \frac{\delta\chi_k(\beta)}{\delta\phi_i(\alpha)} \frac{\delta}{\delta\chi_k(\beta)}$$
$$+ 2\omega \int d\alpha \sum_{i=1}^{N} \sum_{k=1}^{N} \phi_i(\alpha) \frac{\delta\chi_k(\beta)}{\delta\phi_i(\alpha)} \frac{\delta}{\delta\chi_k(\beta)}. \tag{C16}$$

For terms linear in derivatives the process of calculation is exactly the same as for the bosonic case. The final result can be obtained from the bosonic calculation simply by replacing the σ-coordinates by χ-supercoordinates. However, for the term quadratic in derivatives:

$$\sum_{i=1}^{N} \int d\alpha d\overline{\alpha} d\beta d\overline{\beta} \frac{\delta\chi_k(\overline{\beta})}{\delta\phi_i(\overline{\alpha})} \frac{\delta\chi_l(\beta)}{\delta\phi_i(\alpha)} \frac{\delta}{\delta\chi_l(\overline{\beta})} \frac{\delta}{\delta\chi_l(\beta)} \equiv \int d\beta d\overline{\beta} \widetilde{\mathscr{H}}^{(s)}_{k,\ell} \frac{\delta}{\delta\chi_l(\overline{\beta})} \frac{\delta}{\delta\chi_l(\beta)}, \tag{C17}$$

a slight complication arises. Again our strategy is to make use of a generating function of the χ coordinates

$$\mathscr{G}(\beta,p) = \sum_{k=0}^{N} \chi_k(\phi_i(\beta)) p^k = \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} s_{2n}(\phi_i(\beta)) p^n \right] \tag{C18}$$

and try to represent the generating function for

$$\widetilde{\mathscr{H}}^{(s)}_{k,\ell} = \sum_{i=1}^{N} \int d\alpha d\overline{\alpha} \frac{\delta\chi_k(\overline{\beta})}{\delta\phi_i(\overline{\alpha})} \frac{\delta\chi_\ell(\beta)}{\delta\phi_i(\alpha)},$$

i.e.,

$$\widetilde{\mathscr{H}}^{(s)}(t,p) = \sum_{k,\ell=1}^{N} \widetilde{\mathscr{H}}^{(s)}_{k,\ell} t^k p^\ell, \tag{C19}$$

in terms of $\mathscr{G}(\beta,p)$. In what follows we use the notation $\phi_i := \phi_i(\beta)$ and $\overline{\phi_i} := \phi_i(\overline{\beta})$.

Let us first note that

$$\widetilde{\mathscr{H}}^{(s)}(t,p) = 4 \sum_{i=1}^{N} \sum_{m,n=1}^{\infty} (-1)^{n+m} \overline{\phi}_i^{2n-1} \phi_i^{2m-1} t^n p^m \mathscr{G}(\beta,p) \mathscr{G}(\overline{\beta},t), \tag{C20}$$

[cf. (C10)]. After some simple mathematical transformations similar to those leading to (C11) $\widetilde{\mathscr{H}}^{(s)}(t,p)$ can be rewritten as

$$2 \sum_{r=2}^{\infty} (-1)^r (tp)^{r/2} \sum_{u} [s_{2r-2}(\overline{\phi}) + s_{2r-2}(\phi)] \left(\frac{t}{p} \right)^{u/2} \mathscr{G}(\beta,p) \mathscr{G}(\overline{\beta},t)$$
$$+ 4 \sum_{r=2}^{\infty} (-1)^r (tp)^{r/2} \sum_{u} u \left[\frac{s_{2r-2}(\overline{\phi}) - s_{2r-2}(\phi)}{2(r-1)} \right] \left(\frac{t}{p} \right)^{u/2} \mathscr{G}(\beta,p) \mathscr{G}(\overline{\beta},t). \tag{C21}$$

Using (C12) together with the relation

$$\sum_u u\left(\frac{t}{p}\right)^{u/2}=\left(\frac{t}{p}\right)^{-(r-2/2)}\frac{1}{\left(1-\frac{t}{p}\right)^2}\left[-(r-2)+r\frac{t}{p}-r\left(\frac{t}{p}\right)^{r-1}+(r-2)\left(\frac{t}{p}\right)^r\right], \tag{C22}$$

(C21) takes the form

$$4\frac{t}{1-\frac{t}{p}}\sum_{r=1}^{\infty}(-1)^{r+1}[s_{2r}(\phi)p^r-s_{2r}(\bar{\phi})t^r]\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},t)$$

$$-2\frac{t}{p}\frac{t+p}{\left(1-\frac{t}{p}\right)^2}\sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r}[s_{2r}(\bar{\phi})-s_{2r}(\phi)](t^r-p^r)\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},t). \tag{C23}$$

Evidently, the first term in (C23) is equal to

$$\frac{4t}{1-\frac{t}{p}}\left[t\frac{\partial}{\partial t}-p\frac{\partial}{\partial p}\right]\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},t) \tag{C24}$$

[cf. (C13)]. In order to calculate the second term in (C23) we need to use two relations:

$$-\sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r}[s_{2r}(\bar{\phi})-s_{2r}(\phi)](t^r-p^r)=\left[\exp\left[-\sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r}[s_{2r}(\bar{\phi})-s_{2r}(\phi)](t^r-p^r)\right]-1\right], \tag{C25}$$

and

$$\mathcal{G}(\beta,t)\mathcal{G}(\bar{\beta},p)=\exp\left[\frac{1}{2}\sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r}[s_{2r}(\bar{\phi})+s_{2r}(\phi)](t^r+p^r)\right]$$

$$\times\exp\left[\frac{1}{2}\sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r}[s_{2r}(\bar{\phi})-s_{2r}(\phi)](t^r-p^r)\right]. \tag{C26}$$

Finally, the second term in (C23) becomes

$$2\frac{t}{p}\frac{t+p}{\left(1-\frac{t}{p}\right)^2}[\mathcal{G}(\beta,t)\mathcal{G}(\bar{\beta},p)-\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},p)]. \tag{C27}$$

We should emphasize that the bosonic part of this expression vanishes. Combining (C23) and (C26) we obtain the final expression for the generating function $\tilde{\mathcal{H}}^{(s)}(t,p)$ in terms of the generating function $\mathcal{G}(\beta,p)$:

$$\tilde{\mathcal{H}}^{(s)}(t,p)=\frac{4t}{1-\frac{t}{p}}\left[t\frac{\partial}{\partial t}-p\frac{\partial}{\partial p}\right]\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},t)-2\frac{t}{p}\frac{t+p}{\left(1-\frac{t}{p}\right)^2}[\mathcal{G}(\beta,t)\mathcal{G}(\bar{\beta},p)$$

$$-\mathcal{G}(\beta,p)\mathcal{G}(\bar{\beta},p)] \tag{C28}$$

[cf. (C13)]. Expanding $\tilde{\mathcal{H}}^{(s)}(t,p)$ in powers of t,p we arrive at the explicit expression for the coefficients in front of the second derivative terms

$$\tilde{\mathscr{H}}^{(s)}_{\ell,k}=2\sum_{q\geqslant 0}\{2(2q+1+k-\ell)\chi_{\ell-q-1}\overline{\chi}_{k+q}-l[\chi_{k+q}\overline{\chi}_{\ell-q-1}-\overline{\chi}_{k+q}\chi_{\ell-q-1}$$
$$+\chi_{\ell+q-1}\overline{\chi}_{k-q}-\overline{\chi}_{\ell+q-1}\chi_{k-q}]\} \tag{C29}$$

[cf. (C14)] given in (5.12). Making similar calculations for the coefficients in front of the terms linear in derivatives we come to the conclusion that these coefficients are equal to those given by (5.7) with replacement of σ's by χ's [see (5.7) and (5.13)].

[1] F. Calogero, ''Solution of a three-body problem in one dimension,'' J. Math. Phys. **10**, 2191–2196 (1996); ''Ground state of a one-dimensional *N*-body problem,'' *ibid.* **10**, 2197–2200 (1969); ''Solution of the one-dimensional *N*-body problem with quadratic and/or inversely quadratic pair potentials,'' *ibid.* **12**, 419–436 (1971).

[2] B. Sutherland, ''Exact results for a quantum many-body problem in one dimension,'' Phys. Rev. A **4**, 2019–2021 (1971); ''Exact results for a quantum many-body problem in one dimension, II,'' *ibid.* **5**, 1372–1376 (1972).

[3] D. Z. Freedman and P. F. Mende, ''An exactly solvable *N* particle system in supersymmetric quantum mechanics,'' Nucl. Phys. B **344**, 317–343 (1990).

[4] L. Brink, T. H. Hansson, S. Konstein, and M. A. Vasiliev, ''The Calogero model: Anyonic representation, fermionic extension and supersymmetry,'' Nucl. Phys. B 401, 591–612 (1993), .

[5] B. S. Shastry and B. Sutherland, ''Superlax pairs and infinite symmetries in the $1/r^2$ system,'' Phys. Rev. Lett. **70**, 4029–4033 (1993), cond-mat/9212029.

[6] M. A. Olshanetsky and A. M. Perelomov, ''Quantum completely integrable systems connected with semi-simple Lie algebras,'' Lett. Math. Phys. **2**, 7–13 (1977).

[7] D. Kazhdan, B. Kostant, and S. Sternberg, ''Hamiltonian group actions and dynamical systems of Calogero type,'' Commun. Pure Appl. Math. **31**, 481–507 (1978).

[8] M. A. Olshanetsky and A. M. Perelomov, ''Classical integrable finite dimensional systems related to Lie algebras,'' Phys. Rep. **71**, 313 (1981).

[9] M. A. Olshanetsky and A. M. Perelomov, ''Quantum integrable systems related to Lie algebras,'' Phys. Rep. **94**, 313 (1983).

[10] A. Gorsky and N. Nekrasov, ''Hamiltonian systems of Calogero type and two-dimensional Yang-Mills theory,'' Nucl. Phys. B **414**, 213–238 (1994), hep-th/9304047; ''Relativistic Calogero-Moser model as gauged WZW theory,'' *ibid.* **436**, 582–608 (1995), hep-th/9401017.

[11] A. V. Turbiner, ''Lie algebras and linear operators with invariant subspace,'' in *Lie Algebras, Cohomologies and New Findings in Quantum Mechanics*, edited by N. Kamran and P. J. Olver (AMS, 1994), Vol. 160, pp. 263–310; ''Lie-algebras and Quasi-exactly-solvable Differential Equations,'' in *CRC Handbook of Lie Group Analysis of Differential Equations*, edited by N. Ibragimov (CRC, Boca Raton, FL, 1995), Vol. 3, Chap. 12, pp. 331–366, 1995; hep-th/9409068.

[12] W. Rühl and A. Turbiner, ''Exact solvability of the Calogero and Sutherland models,'' Mod. Phys. Lett. A **10**, 2213–2222 (1995), hep-th/9506105.

[13] L. Brink, T. H. Hansson, and M. A. Vasiliev, ''Explicit solution to the *N* body Calogero problem,'' Phys. Lett. B **286**, 109–111 (1992), hep-th/9206049.

[14] L. Lapointe and L. Vinet, ''Exact operator solutions of the Calogero–Sutherland model,'' Commun. Math. Phys. **178**, 425–452 (1996), .

[15] D. Bernard, V. Pasquier, and D. Serban, ''Exact solution of long-range interacting spin chains with boundaries,'' Europhys. Lett. **30**, 301–305 (1995), hep-th/9501044.

[16] D. Serban, ''Some properties of Calogero–Sutherland model with reflections,'' J. Phys. A **30**, 4215–4226 (1997), hep-th/9701118.

[17] H. Awata, Y. Matsuo, S. Odake, and J. Shiraishi, ''Collective field theory, Calogero–Sutherland model and generalized matrix models,'' Phys. Lett. B **347**, 49–55 (1995), hep-th/9411053; ''Excited states of Calogero–Sutherland model and singular vectors of the $W(N)$ algebra,'' Nucl. Phys. B **449**, 347–374 (1995), hep-th/9503043; K. Mimachi and Y. Yamada, ''Singular vectors of the Virasoro algebra in terms of Jack symmetric polynomials,'' Commun. Math. Phys. **174**, 447–455 (1995).

[18] H. Jack, ''A class of symmetric polynomials with a parameter,'' Proc. R. Soc. Edinburgh Sect. A **69**, 1–17 (1969–1970).

[19] R. Stanley, ''Some combinatorial properties of Jack symmetric functions,'' Adv. Math. **77**, 76–115 (1989).

[20] I. G. MacDonald, *Symmetric Functions and Hall Polynomials*, 2nd ed. (Clarendon, Oxford, 1995).

[21] A. M. Perelomov, ''Algebraic approach to the solution of a one-dimensional model of *N* interacting particles,'' Teor. Mat. Fiz. **6**, 364 (1971) (in Russian); English translation: Sov. Phys.—Theor Math. Phys. **6**, 263 (1971).

[22] J. Wolfes, ''On the three-body linear problem with three-body interaction,'' J. Math. Phys. **15**, 1420–1424 (1974).

[23] C. Quesne, ''An exactly solvable three particle problem with three body interaction,'' Phys. Rev. A **55**, 3391–3394 (1997).

[24] A. Capella, M. Rosenbaum, and A. Turbiner, ''Solvability of $G(2)$ integrable system'' preprint ICN-UNAM 97/05 1997, 16, (to appear in Int. J. Mod. Phys. A).

[25] J. A. Minahan and A. P. Polychronakos, ''Integrable systems for particles with internal degrees of freedom,'' Phys. Lett. B **302**, 265–270 (1993).

[26] A. Turbiner, ''Quasi-Exactly-Solvable Problems and the $SL(2,R)$ Group, preprint ITEP-197 (1987); Commun. Math. Phys. **118**, 467–474 (1988).

[27] A. Morozov, A. Perelomov, A. Rosly, M. Shifman, and A. Turbiner, ''Quasi-Exactly-Solvable Problems: One-Dimensional Analogue of Rational Conformal Field Theories,'' Int. J. Mod. Phys. A **5**, 803–843 (1990).

[28] A. Minzoni, M. Rosenbaum, and A. Turbiner, ''Quasi-Exactly-Solvable Many-Body Problems,'' Mod. Phys. Lett. A **11**, 1977–1984 (1996), hep-th/9606092.
[29] A. Turbiner, ''Lame Equation and $sl(2)$,'' preprint ITEP-109 (1988); ''Lame Equation, $sl(2)$ and Isospectral Deformation,'' J. Phys. A **22**, L1–L3 (1989).

BRST symmetries for the tangent gauge group

Alberto S. Cattaneo[a)]
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138

Paolo Cotta-Ramusino[b)]
Department of Mathematics, University of Milano, Via Saldini 50, 20133 Milano, Italy and I.N.F.N., Sezione di Milano, Milano, Italy

Maurizio Rinaldi[b)]
Department of Mathematics, University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy

(Received 26 June 1997; accepted for publication 27 August 1997)

For any principal bundle P, one can consider the subspace of the space of connections on its tangent bundle TP given by the tangent bundle T.$\mathscr{A}$ of the space of connections $\mathscr{A}$ on P. The tangent gauge group acts freely on T.$\mathscr{A}$. Appropriate BRST operators are introduced for quantum field theories that include as fields elements of T.$\mathscr{A}$, as well as tangent vectors to the space of curvatures. As the simplest application, the BRST symmetry of the so-called BF-Yang–Mills theory is described and the relevant gauge fixing conditions are analyzed. A brief account on the topological BF theories is also included and the relevant Batalin–Vilkovisky operator is described. © *1998 American Institute of Physics.*
[S0022-2488(98)00802-0]

I. INTRODUCTION

In this paper we study the BRST complexes arising from the action of the (iterated) tangent bundle of the group of gauge transformations $\mathscr{G}$ on the (iterated) tangent bundle of the space of connections $\mathscr{A}$.

Our main original motivation for such a study was the understanding of the BRST and the Batalin–Vilkovisky (BV) structure of four-dimensional BF theories. These topological theories have been shown[1] to be strictly related to the four-dimensional Yang–Mills theory.

Let us first recall what BF action functionals are in both three and four dimensions. We consider ''space–time'' to be given by a closed, oriented Riemannian manifold M with a gauge structure given by a G-principal bundle P with base M. The group G is a simple compact connected Lie group [typically SU(N)] with Lie algebra denoted by $\mathfrak{g}$. Unless otherwise stated, the forms we are considering on M will take values in the adjoint bundle and hence their local expression will take values in $\mathfrak{g}$.

When $M \equiv M^3$ is a *three-dimensional* space, then the most known topological theory is that defined in terms of the Chern–Simons action functional,

$$\mathrm{CS}(A) \equiv \int_{M^3} \mathrm{Tr}(A \wedge dA + \tfrac{1}{3} A \wedge [A,A]),$$

where A is (the local expression of) a connection on $P(M,G)$.

We can also introduce a field $\hat{B}$ given by a 1-form. The 1-form $A + t\hat{B}$ is again a connection.

The three-dimensional topological BF action functional is then defined in terms of the Chern–Simons action functional by

$$S_{BF,t}(A,\hat{B}) \equiv \frac{1}{2t}\,(\mathrm{CS}(A+t\hat{B}) - \mathrm{CS}(A-t\hat{B})) = \int_{M^3} \mathrm{Tr}\left(2(\hat{B} \wedge F_A) + \frac{t^2}{3}\,\hat{B} \wedge [\hat{B},\hat{B}]\right), \quad \text{(I.1)}$$

[a)]Supported by I.N.F.N. Grant No. 5565/95 and DOE Grant No. DE-FG02-94ER25228, Amendment No. A003.
[b)]Supported in part by research grants of the Ministry of University and Scientific Research (MURST).

 © 1998 American Institute of Physics

where F_A denotes the curvature of A and t is a real parameter. The action functional (I.1) is called the *BF action functional with cosmological constant*.[2]

In the limit $t\to 0$ we obtain, up to a factor 2, the so-called *pure BF action functional*,

$$S_{BF}=\int_{M^3}\mathrm{Tr}(\hat{B}\wedge F_A).$$

One can define observables related to links in M^3 for both $S_{BF,t}$ and S_{BF}; it has been shown in Ref. 2 that the corresponding expectation values give the coefficients of the Homfly–Jones polynomials in the first case and are related to the Alexander–Conway polynomials in the second case.

In *four dimensions* we first consider the topological Chern action functional

$$\int_M \mathrm{Tr}(F_A\wedge F_A).$$

Then we consider a field B given by a 2-form. The four-dimensional topological *BF action functional (with cosmological constant)* is then defined as

$$S_{BF,t}\equiv\frac{1}{2t}\int_M \mathrm{Tr}((F_A+tB)\wedge(F_A+tB)-F_A\wedge F_A)=\int_M \mathrm{Tr}\left(B\wedge F_A+\frac{t}{2}B\wedge B\right), \tag{I.2}$$

and again the *pure BF action functional* S_{BF} is defined as the limit of the above one when t goes to 0.

In four dimensions one may also start with the Yang–Mills action functional

$$S_{\mathrm{YM}}=\int_M \mathrm{Tr}(F_A\wedge *F_A),$$

where $*$ is the Hodge operator. (We omit the coupling constant here.)

The corresponding BF-Yang–Mills (BFYM) action functional reads

$$S_{BF\mathrm{YM},t}\equiv\int_M \mathrm{Tr}\left(B\wedge F_A+\frac{t}{2}B\wedge *B\right), \tag{I.3}$$

whose limit $t\to 0$ gives again the pure BF action functional.

When $t=1$, (I.3) *is equivalent, by Gaussian integration, to the Yang–Mills action functional*.

When $t=0$, (I.3) has an extra-symmetry, namely, the translation $B\to B+d_A\tau$ where τ is a 1-form. In order to extend this symmetry for a generic t and carry on the perturbative expansion around the critical solutions for the BF-Yang–Mills theory, it is necessary to deform the action functional (I.3) into the following action functional:

$$S_{BF\mathrm{YM},\eta,t}\equiv\int_M \mathrm{Tr}\left(B\wedge F_A+\frac{t}{2}(B-d_A\eta)\wedge *(B-d_A\eta)\right). \tag{I.4}$$

Here η is a 1-form and the new symmetry reads, for a generic 1-form τ,

$$B\to B+d_A\tau,\quad \eta\to\eta+\tau. \tag{I.5}$$

It is now possible to show that in perturbation theory, the action functional (I.4) is equivalent to the Yang–Mills action functional and we refer to Ref. 1 for such a discussion.

By now it is clear, just by inspecting all the action functionals written above, that in both the three- and the four-dimensional BF theories, the field $\hat{B}$ and, respectively, B belong to some tangent space to the space of fields; the connections in three dimensions and the curvatures in four.

Some care is to be applied in four dimensions, for the space of ''curvatures'' is not the space of fields where one performs the functional integration.

Specifically, in the BFYM theories with action functional (I.4), one works with triples (A,η,B) where the pair (A,η) belongs to the tangent bundle of the space of connections and B is a tangent vector at F_A to the vector space of 2-forms.

The relevant BRST cohomology should accommodate both the gauge invariance and the invariance under translations (I.5); the latter condition implies that eventually no physical quantity should depend on the field η; that is, the theory should behave like a topological field theory as far as the field η is concerned.

The existence of such a BRST cohomology is *a priori* not obvious; the purpose of this paper is to describe in details such a BRST cohomology, which is nothing but a straightforward extension of the BRST cohomology arising from the action of the tangent gauge group on the tangent bundle of connections.

The plan of the paper is as follows:

In Sec. II we recall some aspects of the topological field theories of cohomological type as discussed in Ref. 3.

In Secs. III and IV we study the space of connections of the tangent bundle of a principal bundle (tangent connections) and the relevant gauge group.

In Secs. V and VI we discuss the BRST cohomology arising from the action of the tangent gauge group on the tangent bundle of connections.

In Sec. VII we show that it is possible to extend the above BRST operator to the 2-form B.

In Sec. VIII we show that the construction can be indefinitely iterated by including 2^n 2-forms B_i, 2^n 1-forms η_i and the relevant ghosts.

In Sec. IX we consider the BV construction corresponding to the action functional (I.2).

In Sec. X we discuss the orbit space and the moduli space relevant to the action functional (I.4).

Some of the results of Secs. VI, VII, X have been anticipated in Ref. 1.

As a final remark, we notice that the 2-form B in four-dimensional BF theories can be interpreted geometrically as a tangent vector to the space of connections over the following bundle.

Take the space $\mathscr{P}_A(P)$ of (free) horizontal paths on P with respect to a connection A. This is a principal G-bundle, whose base space is the space of the free paths on M. The initial-point map $ev_0:\mathscr{P}_A(P)\to P$ is a bundle morphism, so ev_0^*A is a connection on $\mathscr{P}_A(P)$. Any form $B \in \Omega^2(M,\mathrm{ad}\, P)$, once integrated over the paths, represents a tangent vector to the space of connections on $\mathscr{P}_A(P)$. The computation of the relevant holonomies allows us to define observables corresponding to imbedded tori in a (simply connected) 4-manifold M, for which expectation values with respect to the different BF action functionals can be computed. We refer to a separate paper[4] for such a discussion.

II. SYMMETRIES IN GAUGE AND TOPOLOGICAL FIELD THEORIES

In this section we review some well known facts concerning gauge invariance in topological (cohomological) theories and in Yang–Mills theories and the BRST structure of such theories.

The main symmetry group we are considering is the gauge group $\mathscr{G}$ that acts on the space $\mathscr{A}$ of (irreducible) smooth connections on P. We assume that we have divided $\mathscr{G}$ by its center, so that $\mathscr{G}$ acts freely on $\mathscr{A}$ yielding the principal bundle of gauge-orbits.

As usual, we denote by ad P the associated bundle $P\times_{Ad}\mathfrak{g}$ and by $\Omega^*(M,\mathrm{ad}\, P)$ the graded Lie algebra of tensorial form on P of the adjoint type, or equivalently, of forms on M with values in ad P. For any $A\in\mathscr{A}$ the covariant exterior derivative will be a linear map,

$$d_A:\Omega^*(M,\mathrm{ad}\, P)\to\Omega^{*+1}(M,\mathrm{ad}\, P).$$

The space $\mathscr{A}$ is an affine space and its tangent space $\mathrm{T}_A\mathscr{A}$ is the vector space $\Omega^1(M,\mathrm{ad}\, P)$.

A quantity (observable, action functional) is gauge invariant if it is invariant under the action of the group $\mathscr{G}$. It is moreover independent of the connection if it is invariant under the translation group $\Omega^1(M,\mathrm{ad}\, P)$. In quantum field theories one may wish to consider simultaneously both kinds of invariance.

In this case one is naturally led to consider the semidirect product

$$\mathcal{G}_T \triangleq \mathcal{G} \ltimes \Omega^1(M, \text{ad } P). \tag{II.1}$$

In this group the product of two elements (g_1, ρ_1) and (g_2, ρ_2) is defined by

$$(g_1, \rho_1)(g_2, \rho_2) = (g_1 g_2, \text{Ad}_{g_2^{-1}} \rho_1 + \rho_2).$$

Here $g_i \in \mathcal{G}$ and $\rho_i \in \Omega^1(M, \text{ad } P)$.

The group $\mathcal{G}_T$ acts on the space of connections as follows:

$$A \cdot (g, \rho) = A^g + \rho, \tag{II.2}$$

where A^g denotes the gauge-transformed connection.

This action is not free since $(g, A - A^g) \in \mathcal{G}_T$ leaves the connection A fixed. At the infinitesimal level, we see that the freedom of the action is missing since we have nontrivial solutions of the equation

$$\rho + d_A \chi = 0, \quad \rho \in \Omega M^1(M, \text{ad } P), \quad \chi \in \Omega^0(M, \text{ad } P),$$

where d_A is the covariant exterior derivative.

From a field-theoretical point of view, dealing with a symmetry that does not correspond to a free action is troublesome since a gauge fixing mechanism and a consistent definition of ghost fields may not be available.

The cure for this problem is to require ρ to belong to a complementary space of the image of $d_A : \Omega^0(M, \text{ad } P) \to \Omega^1(M, \text{ad } P)$. In other words we have to choose a *connection on the bundle of gauge orbits*,

$$\mathcal{A} \mapsto \frac{\mathcal{A}}{\mathcal{G}}. \tag{II.3}$$

Choosing such a connection is also called *fixing the gauge*. We denote the above connection on the bundle of gauge orbits by the symbol c.

In turn any connection c on $\mathcal{A}$, determines a connection on the bundle

$$\frac{P \times \mathcal{A}}{\mathcal{G}} \mapsto M \times \frac{\mathcal{A}}{\mathcal{G}}. \tag{II.4}$$

In fact the 1-form on $P \times \mathcal{A}$ given by

$$(A + c)_{(p,A)}(X, \eta) \triangleq A(X)_p + c(\eta)_{(A,p)}, \quad X \in \text{T}_p P, \eta \in \text{T}_A \mathcal{A} \tag{II.5}$$

defines a $\mathcal{G}$-invariant connection on $P \times \mathcal{A}$,[5] i.e., a connection on the principal G-bundle (II.4).

The exterior derivative on forms on $P \times \mathcal{A}$ which are of order (k, s) (i.e., of order k as forms on P and order s as forms on $\mathcal{A}$) can be written as $d_{\text{tot}} = d + (-1)^k \delta$, where d and δ are the exterior derivatives on P and $\mathcal{A}$, respectively. Analogously we set the following convention for the commutator:

$$[\omega_1, \omega_2] = (-1)^{k_1 k_2 + s_1 s_2 + 1} [\omega_2, \omega_1],$$

where (k_i, s_i) denotes the order of the form ω_i. As a consequence we have the following expression for the total exterior covariant derivative,

$$d_{A+c} = d_A + (-1)^k \delta_c, \tag{II.6}$$

where k denotes again the order of the form on P. In particular we have to remember Eq. (II.6) when computing curvatures and their Bianchi identity.

The order of a form with respect to the space $\mathcal{A}$ is called the *ghost number*.

The operator δ (or some of its restrictions or extensions) is as an example of what in quantum field theory is called a BRST (or a BRS) operator; it describes the symmetry of the fields that one considers in the theory.

We have in mind essentially two cases.

The first case is that of the *topological field theory* à la Baulieau–Singer[3] or, according to Witten's terminology,[6] of a *cohomological field theory*. In this case the theory has no degrees of freedom and the action functional is invariant under (infinitesimal) translations of the space of fields. The corresponding BRST operator is then the exterior derivative on the space of fields ($\mathscr{A}$ in the previous case).

The second case is that of a field theory (generally not topological) in which there is a symmetry group acting on the space of fields and the theory is only required to be invariant under such a symmetry group. In this last case the BRST operator is just the restriction of the exterior derivative on the space of fields to the orbits of that symmetry group.

In the Yang–Mills theory, the BRST operator is the exterior derivative along the gauge orbits.

Mixed cases will also (and especially) be considered in this paper, namely, theories for which the BRST operator is the exterior derivative for some of the fields and the derivative along the fibers (orbits) for other fields.

We may say that these theories are ''semitopological'' or topological in some fields and nontopological in others. We use this terminology, even though topological field theories are not only those of cohomological type.

In a topological theory (of the cohomological type) defined for an even-dimensional space, the typical action functional is defined in terms of characteristic classes, so is independent of the choice of the connection.

Let us come back to the Baulieu–Singer theory. Such a theory is invariant under translations in the space of connections, the action functional is represented by characteristic classes, and the field equations are obtained by considering the structure equation for the connection (II.5) and the relevant Bianchi identities.

The structure equations read

$$F_A+\psi+\phi=d_A A+d_A c+\tfrac{1}{2}[c,\ c]-\delta A+\delta c, \tag{II.7}$$

where the left-hand side gives the various components of the curvature of the connection (II.5); F_A,ψ,ϕ are respectively forms of degree (2,0), (1,1), and (0,2) on the product space,

$$P\times\mathscr{A}.$$

The previous equation can be rewritten as

$$\delta A=d_A c-\psi,\quad \delta c=-\tfrac{1}{2}[c,\ c]+\phi. \tag{II.8}$$

Notice that the form ψ is minus the projection of $\Omega^1(M,\mathrm{ad}\ P)$ onto the horizontal subspace at A.

The Bianchi identity defines the transformation properties of ψ and ϕ (and F_A),

$$\delta\psi=d_A\phi-[\psi,\ c],\quad \delta\phi=[\phi,\ c], \tag{II.9}$$

$$d_A F_A=0,\quad \delta F_A=[F_A,\ c]-d_A\psi. \tag{II.10}$$

By comparing the first equation in (II.8) with the action (II.2), we see that we have successfully turned around the problem of the nonfreedom of the action (II.2).

As was recalled before, the *Yang–Mills* theory, as opposed to the previous topological theory, involves the restriction of the transformation (II.8) to the orbit passing through $A\in\mathscr{A}$. The relevant equations are then

$$\delta A=d_A c,\quad \delta c=-\tfrac{1}{2}[c,\ c], \tag{II.11}$$

where c in (II.11) is now the *Maurer–Cartan form* on $\mathscr{G}$, and δ is the exterior derivative on $\mathscr{G}$ or, more precisely, on the $\mathscr{G}$-orbit passing through A.

The equations (II.11) are the classical BRST equations.[7]

A final remark is in order. Whenever we have a nonfree action of a group over a manifold, one is naturally led to consider the corresponding equivariant cohomology. In our case the group is $\mathscr{G}_T$

with its nonfree action on $\mathscr{A}$. But $\mathscr{G}_T$ is the semidirect product of an Abelian (contractible) subgroup $\Omega^1(M,\mathrm{ad}\,P)$ times $\mathscr{G}$, which behaves in many ways like a ''compact group'' and acts freely on (the contractible space) $\mathscr{A}$.

So one has just to consider this last action, or the cohomology of the principal bundle (II.4) (which in turn is just the equivariant cohomology corresponding to the nonfree action of $\mathscr{G}$ on P).

In the following sections we will extend this procedure by considering the action of the tangent bundle of the gauge group (which also is a Lie group) on the tangent space to the space of connections. This may look irrelevant from a purely topological point of view since, as will be recalled below, the tangent gauge group is again the semidirect product of $\mathscr{G}$ times an Abelian contractible group. This implies that the classifying spaces for the gauge group and for its tangent bundle are homotopically equivalent.

But the BRST structure of the tangent gauge group will show an interesting degree of flexibility, due exactly to the fact that there is a large Abelian normal subgroup of the tangent gauge group. This will allow us to include into the BRST algebra the 2-form fields B of the four-dimensional BF theories.

However, before doing so, we need to discuss some basic facts of the differential geometry of the tangent bundle of a principal bundle.

III. TANGENT PRINCIPAL BUNDLES AND THEIR GAUGE GROUPS

Let G be a Lie group and let $\mathfrak{g}$ its Lie algebra. The tangent bundle $\mathrm{T}G$ is a Lie group with respect to the multiplication given by the tangent map of the multiplication $m:G\times G\mapsto G$

$$\mathrm{T}m:\mathrm{T}G\times\mathrm{T}G\mapsto\mathrm{T}G.$$

Explicitly this multiplication is given by

$$(k,u_k)(h,v_h)\equiv(kh,kv_h+u_kh)\quad k,h\in G;u_k\in\mathrm{T}_kG;v_h\in\mathrm{T}_hG.$$

Here the right and left multiplications of a vector by an element of G denote the push-forward of the corresponding multiplications in G.

Since we can represent any vector u_k as $u_k=kx$ for a unique $x\in\mathfrak{g}=\mathrm{T}_eG$, there is an isomorphism

$$\mathrm{T}G\tilde{\to}G\ltimes\mathfrak{g},\quad(k,uk)\rightsquigarrow(k,k^{-1}u_k)\tag{III.1}$$

and on the semidirect product $G\ltimes\mathfrak{g}$ the product of two elements is defined as

$$(k,x)(h,y)\equiv(kh,\mathrm{Ad}_{h^{-1}}(x)+y)\quad k,h\in G;x,y\in\mathfrak{g}.$$

The inverse of the element $(k,x)\in(G\ltimes\mathfrak{g})$ is given by

$$(k,x)^{-1}=(k^{-1},-\mathrm{Ad}_k(x)),\tag{III.2}$$

and the conjugation on $G\ltimes\mathfrak{g}$ is given by

$$(k,x)(h,y)(k,x)^{-1}=(khk^{-1},\mathrm{Ad}_k\mathrm{Ad}_{h^{-1}}(x)+\mathrm{Ad}_k(y)-\mathrm{Ad}_k(x)).\tag{III.3}$$

The adjoint action of $G\ltimes\mathfrak{g}$ on $T_{(e,0)}(G\ltimes\mathfrak{g})=\mathfrak{g}\times\mathfrak{g}$ is given by

$$\mathrm{Ad}_{(k,z)}(x,y)=(\mathrm{Ad}_k(x),\mathrm{Ad}_k([z,x]+y)),\tag{III.4}$$

whereas the commutator is given by

$$[(x_1,y_1),\ (x_2,y_2)]=([x_1,\ y_2],[x_1,\ y_2]+[y_1,\ x_2]).\tag{III.5}$$

So $\mathrm{Lie}(\mathrm{T}G)$ is isomorphic as a Lie algebra to the semidirect sum $\mathfrak{g}\oplus_s\mathfrak{g}$. We now consider a G-principal bundle $P\mapsto M$, with base space M. Its tangent bundle is a $\mathrm{T}G$-principal bundle with right action of $\mathrm{T}G$ on $\mathrm{T}P$ given by the tangent map to the right action $R:P\times G\mapsto P$,

$$\begin{aligned} &\mathrm{T}R:\mathrm{T}P\times\mathrm{T}G\mapsto\mathrm{T}P,\\ &((p,X),(k,u_k))\rightsquigarrow(pk,Xk+pu_k). \end{aligned} \tag{III.6}$$

Here and in what follows, Xk and pu_k (with $p\in P$, $X\in \mathrm{T}_pP$, $k\in G$, $u_k\in \mathrm{T}_kG$) are simplified notations for the partial tangent maps $\mathrm{T}_1R:\mathrm{T}P\times G\to\mathrm{T}P$ and $\mathrm{T}_2R:P\times\mathrm{T}G\to\mathrm{T}P$. We now rewrite the previous action, by considering the isomorphism (III.1). If $(k,x)\in G\ltimes\mathfrak{g}$, and $i(x)$ denotes the fundamental vector field in P generated by $x\in\mathfrak{g}$, then the action (III.6) becomes

$$(p,X)(k,x)=(pk,Xk+i(x)|_{pk}). \tag{III.7}$$

It is easy to check directly that (III.7) defines a good action

$$\begin{aligned} [(p,X)(k,x)](h,y)&=(pkh,(Xk)h+(i(x)|_{pk})h+i(y)|_{pkh})\\ &=(pkh,X(kh)+i(\mathrm{Ad}_{h^{-1}}x)|_{pkh}+i(y)|_{pkh})\\ &=(p,X)[(k,x)(h,y)]. \end{aligned}$$

The action (III.7) is free and moreover $\pi:\mathrm{T}P\mapsto\mathrm{T}M$ is a $(G\ltimes\mathfrak{g})$-principal fiber bundle.

We now shift to the infinite-dimensional case. We consider the tangent bundle $\mathrm{T}\mathscr{G}$ of the group of gauge transformations.

This can be identified with the semidirect product of $\mathscr{G}\ltimes\Omega^0(M,\mathrm{ad}\,P)$, where the product of two elements is defined by

$$(g_1,\zeta_1)(g_2,\zeta_2)=(g_1g_2,\mathrm{Ad}_{g_2^{-1}}\zeta_1+\zeta_2). \tag{III.8}$$

Its Lie algebra is given by the semidirect sum $\Omega^0(M,\mathrm{ad}\,P)\oplus_s\Omega^0(M,\mathrm{ad}\,P)$ whose commutator is defined as follows:

$$[(\chi_1,\zeta_1),\,(\chi_2,\zeta_2)]=([\chi_1,\,\chi_2],[\chi_1,\,\zeta_2]+[\zeta_1,\,\zeta_2]).$$

The tangent map to the action of $\mathscr{G}$ on $\mathscr{A}$ defines the action of $\mathrm{T}\mathscr{G}$ on $\mathrm{T}\mathscr{A}\sim\mathscr{A}\times\Omega^1(M,\mathrm{ad}\,P)$ as follows:

$$(A,\eta)\cdot(g,\zeta)=(A^g,\mathrm{Ad}_{g^{-1}}(\eta)+d_{A^g}\zeta). \tag{III.9}$$

The fundamental vector field at $(A,\eta)\in\mathrm{T}\mathscr{A}$ corresponding to $(\chi,\zeta)\in\mathrm{Lie}(\mathrm{T}\mathscr{G})$ is given by

$$(d_{A\chi},[\eta,\,\chi]+d_A\zeta). \tag{III.10}$$

The action (III.9) is free when we understand that only irreducible connections on P are taken into account and that the group $\mathscr{G}$ has been divided by its center.

A *gauge transformation* for $\mathrm{T}P$ is, by definition, a map

$$f:\mathrm{TP}\to G\ltimes\mathfrak{g}\ \text{such that}\ f[(p,X)(k,x)]=(k,x)^{-1}f(p,X)(k,x).$$

We have the following:

Theorem III.1: *The tangent bundle* $\mathrm{T}\mathscr{G}$ *of the gauge Lie group* $\mathscr{G}$ *is a proper subgroup of the group* $\mathscr{G}_{\mathrm{TP}}$ *of gauge transformations for* $\mathrm{T}P$.

Proof of Theorem III.1: We consider the tangent map to the evaluation map $ev:P\times\mathscr{G}\to G$ at $(g,\chi)\in\mathscr{G}\ltimes\Omega^0(M,\mathrm{ad}\,P)\sim\mathrm{T}\mathscr{G}$.

It is a map,

$$\mathrm{T}P\to G\ltimes\mathfrak{g},$$

given by

$$(g(p),g^{-1}\,dg(p,X)+\chi(p))\in G\ltimes\mathfrak{g},$$

where we have to remember that the logarithmic derivative $g^{-1}\,dg$ is a map defined on TP with values in $\mathfrak{g}$.

In order to show that it is a gauge transformation for TP, it is enough to see that for any $(p,X)\in \mathrm{T}P$ and $(k,x)\in G\ltimes\mathfrak{g}$ we have the equation,

$$\begin{aligned} g^{-1}\,dg\{(p,X)(k,x)\} &= g^{-1}\,dg\{(pk, Xk + i(x)|_{pk})\} \\ &= \mathrm{Ad}_{k^{-1}}[g^{-1}\,dg(p,X)] + x - \mathrm{Ad}_{k^{-1}}\mathrm{Ad}_{g^{-1}(p)}\mathrm{Ad}_k x \\ &= p_2(\mathrm{Ad}_{(k,x)^{-1}}\{g^{-1}\,dg(p,X)\}), \end{aligned}$$

where $p_2:\mathfrak{g}\oplus_s\mathfrak{g}\to\mathfrak{g}$ is the projection onto the second component. For any $\eta\in\Omega^1(M,\mathrm{ad}\,P)$ and for any $(g,\chi)\in\mathscr{G}\ltimes\mathrm{Lie}(\mathscr{G})$ the map

$$(p,X)\rightsquigarrow(g(p), g^{-1}\,dg(p,X)+\chi(p)+\eta(p,X))$$

is also a gauge transformation on TP and this shows that the inclusion $\mathrm{T}\mathscr{G}\hookrightarrow\mathscr{G}_{\mathrm{T}P}$ is proper. □

We have also shown that the group $\mathscr{G}_{\mathrm{T}P}$ includes the group $\mathscr{G}_{\mathrm{aff}}$ of *affine gauge transformations*, defined as the semidirect product $\mathrm{T}\mathscr{G}\ltimes\Omega^1(M,\mathrm{ad}\,P)$, whose elements are triples $(g,\chi,\eta)\in\mathscr{G}\ltimes\Omega^0(M,\mathrm{ad}\,P)\ltimes\Omega^1(M,\mathrm{ad}\,P)$. The product of two such triples is given by

$$(g_1,\chi_1,\tau_1)(g_2,\chi_2,\tau_2)=(g_1g_2,\mathrm{Ad}_{g_2^{-1}}\chi_1+\chi_2,\mathrm{Ad}_{g_2^{-1}}\tau_1+\tau_2,). \tag{III.11}$$

The Lie algebra of $\mathscr{G}_{\mathrm{aff}}$ is

$$(\Omega^0(M,\mathrm{ad}\,P)\oplus_s\Omega_0(M,\mathrm{ad}\,P))\oplus_s\Omega^1(M,\mathrm{ad}\,P)$$

and the relevant commutator is given by

$$[(\zeta_1,\chi_1,\tau_1),(\zeta_2,\chi_2,\tau_2)]=([\zeta_1,\,\zeta_2],[\zeta_1,\,\chi_2]+[\chi_1,\,\zeta_2],[\zeta_1,\,\tau_2]+[\tau_1,\,\zeta_2]).$$

IV. TANGENT CONNECTIONS

Now we discuss the structure of the space of connections of the principal TG-bundle TP.

For any manifold M we can consider its double tangent bundle TTM,

$$\begin{array}{ccc} \mathrm{TT}M & \xrightarrow{\mathrm{T}\pi_M} & \mathrm{T}M \\ \downarrow{\scriptstyle \pi_{\mathrm{T}M}} & & \downarrow{\scriptstyle \pi_M} \\ \mathrm{T}M & \xrightarrow{\pi_M} & M. \end{array}$$

Let us consider a rectangle in M centered at x, i.e., a C^2-map $\underline{x}:[a,b]\times[a',b']\mapsto M$, with $(0,0)\in(a,b)\times(a',b')$, $\underline{x}(0,0)=x$. We will denote by s and t the two variables and by a prime the derivative with respect to s, by a dot the derivative with respect to t. A tangent vector in TTM at $(\underline{x}(0,0),\underline{x}'(0,0))$ can be represented as $(\dot{\underline{x}}(0,0),(d\underline{x}'/dt)(0,0))$. When it is not otherwise specified, the derivatives are understood to be computed at (0,0). Notice that two rectangles are meant to be equivalent if they have the same first order derivatives as well as the mixed derivatives of order 2. Everything that follows is defined on the equivalence classes. Identifying double tangents with equivalence classes of rectangles provides us also with a (noncanonical) extension of the double tangent vector to the image of the rectangle.

In our notation we have, for a given $[\underline{x}]\in\mathrm{TT}M$, $\mathrm{T}\pi_M[\underline{x}]=\dot{\underline{x}}$ and $\pi_{\mathrm{T}M}[\underline{x}]=\underline{x}'$.

There is a canonical involution $\alpha_M:\mathrm{TT}M\mapsto\mathrm{TT}M$ given by

$$\alpha_M\left(x,\underline{x}',\dot{\underline{x}},\frac{d\underline{x}'}{dt}\right)=\left(x,\dot{\underline{x}},\underline{x}',\frac{d\dot{\underline{x}}}{ds}\right). \tag{IV.1}$$

It is well known[8] that for any connection A on P, one can induce a connection $\mathring{A}$ on TP. If $\underline{p}$ is any rectangle in P centered at p, then $\mathring{A}$ is, by definition, the 1-form on TP with values in $\mathfrak{g} \oplus_s \mathfrak{g}$, given by

$$\mathring{A}_{p,p'}\left(\dot{\underline{p}}, \frac{d\underline{p}'}{dt}\right) = \left(A_p(\dot{\underline{p}}), \left.\frac{d}{ds}\right|_{s=0} [A_{\underline{p}(s,0)}(\dot{\underline{p}}(s,0))]\right). \tag{IV.2}$$

Notice that the canonical involution has been used in the definition of the above connection.

A more general connection on TP can be obtained by considering the evaluation map,

$$ev: \mathscr{A} \times TP \to \mathfrak{g},$$

$$(A; p, X) \rightsquigarrow A_p(X). \tag{IV.3}$$

The tangent map of (IV.3)

$$Tev: T\mathscr{A} \times TTP \to \mathfrak{g} \oplus_s \mathfrak{g}, \tag{IV.4}$$

evaluated at $(A, \eta) \in T\mathscr{A}$ and composed with the canonical involution α_P gives a connection on TP. In fact we have the following:

Theorem IV.1: *For any* $(A, \eta) \in T\mathscr{A}$*, the 1-form* $\omega(A, \eta)$ *on* TP *given by*

$$\omega(A, \eta)([\underline{p}]) \equiv Tev(A, \eta; \alpha_P[\underline{p}]) = \left(A_p(\dot{\underline{p}}), \left.\frac{d}{ds}\right|_{s=0} A_{\underline{p}(s,0)}(\dot{\underline{p}}(s,0)) + \eta(\dot{\underline{p}})\right) \tag{IV.5}$$

defines a connection on TP.

Proof of Theorem IV.1: The adjoint action of the group on its Lie algebra, extends, in an obvious way, to Lie-algebra valued forms. So we have a natural map $\mathrm{ad}: G \times \Omega^*(P, \mathfrak{g}) \mapsto \Omega^*(P, \mathfrak{g})$ and its derivative $T\,\mathrm{ad}_G = \mathrm{ad}_{TG}$ [see (III.4)]. If we consider again the derivative in the first variable of the right multiplication on P, i.e., the map $T_1R: TP \times G \to TP$ and we denote by $\Delta_G: G \to G \times G$ the diagonal map in G and by $p_V: V \times U \to V$ the projection on V for any pair of spaces (V, U), then the ad-equivariance of the connections is given by the following equation:

$$ev \circ p_{\mathscr{A} \times TP} = ev \circ (\mathrm{ad} \times T_1R) \circ (\mathbb{I} \times \Delta_G): \mathscr{A} \times TP \times G \to \mathfrak{g}, \tag{IV.6}$$

where the evaluation map is given by (IV.3). The derivative of both sides of (IV.6) gives

$$Tev \circ p_{T\mathscr{A} \times TTP} = Tev \circ (\mathrm{ad}_{TG} \times T_1R) \circ (\mathbb{I} \times \Delta_{TG}). \tag{IV.7}$$

The equivariance of (IV.5) follows when we notice that

$$T\,T_1R = \alpha_P \circ T_1TR \circ \alpha_P: TTP \times TG \to TTP,$$

where α_P is the canonical involution. Finally we notice that the difference $\omega(A, \eta) - \mathring{A}$ evaluated at $[\underline{p}]$ gives $\eta(\dot{\underline{p}})$. But the projection $T\pi_P$ applied to the fundamental vector field in $(p, X) \in TP$ determined by $(a, b) \in \mathfrak{g} \oplus_s \mathfrak{g}$ yields $i(a)_p$, i.e., the fundamental vector field in P corresponding to a evaluated at p. The theorem follows from the fact that $\mathring{A}$ is a connection and η is a tensorial 1-form. □

When we consider the inclusion

$$\iota: P \hookrightarrow TP, \quad \iota(p) = (p, 0), \tag{IV.8}$$

then the pullback of the connection $\omega(A, \eta)$ via (IV.8) is simply given by

$$[\iota^* \omega(A, \eta)]_p(X) = (A(X)_p, \eta(X)_p) \in \mathfrak{g} \oplus_s \mathfrak{g}, \quad X \in T_pP. \tag{IV.9}$$

Also we have the following:

Theorem IV.2: *The inclusion of* $T\mathscr{A}$ *into the space of connections of* TP *is proper.*

Proof of Theorem IV.2: The bundle $\mathrm{T}.\mathscr{A}$ can be identified with an affine space modeled on

$$\Omega^1(M,\mathrm{ad}\ P)\oplus\Omega^1(M,\mathrm{ad}\ P)\sim\Omega^1(M,\mathrm{ad}\ P\oplus\mathrm{ad}\ P),$$

while the space of connections on $\mathrm{T}P$ can be identified with an affine space modeled on $\Omega^1(\mathrm{T}M,\mathrm{ad}\,\mathrm{T}P)$. The theorem is proved when we notice that the projection $M\to\mathrm{T}M$ determines a proper inclusion of $\Omega^*(M)$ into $\Omega^*(\mathrm{T}M)$ and that the pair of projections $\mathrm{T}P\to P$, $\mathrm{T}G\to G$ determines

(1) a homomorphism of the $\mathrm{T}G$-principal bundle $\mathrm{T}P\times\mathfrak{g}\times\mathfrak{g}\to\mathrm{ad}\,\mathrm{T}P$ (where the left $\mathrm{T}G$-adjoint action on $\mathfrak{g}\times\mathfrak{g}$ is considered) onto the G-principal bundle $P\times\mathfrak{g}\times\mathfrak{g}\to\mathrm{ad}\ P\oplus\mathrm{ad}\ P$ (where the left G-adjoint action on $\mathfrak{g}\times\mathfrak{g}$ is considered), and hence
(2) a proper inclusion of the space of sections $\Gamma(\mathrm{ad}\ P\oplus\mathrm{ad}\ P)$ into the space of sections $\Gamma(\mathrm{ad}\,\mathrm{T}P)$. □

Two final remarks are in order.

(A) By iterating the procedure, we obtain an inclusion

$$\mathrm{T}^n.\mathscr{A}(P)\hookrightarrow.\mathscr{A}(\mathrm{T}^nP),$$

where T^n denotes the n-iterated tangent bundle and $.\mathscr{A}(\bullet)$ denotes the space of connections of a given principal bundle.

(B) When we consider only irreducible connections, then $\mathscr{A}\mapsto\mathscr{A}/\mathscr{G}$ is a principal bundle and so is $\mathrm{T}^n.\mathscr{A}\mapsto\mathrm{T}^n.\mathscr{A}/\mathrm{T}^n\mathscr{G}$.

V. TOPOLOGICAL FIELD THEORIES FOR TANGENT PRINCIPAL BUNDLES

Here we want to extend the topological field theory discussed in Sec. II to the case when the principal bundle is the *tangent bundle* $\mathrm{T}P$. The purpose of this section is to write the BRST equations for this case.

Here we will consider only those special connections for $\mathrm{T}P$ that are determined by elements of $\mathrm{T}.\mathscr{A}$, i.e., by pairs (A,η) where A is a connection on P and η is any element of $\Omega^1(M,\mathrm{ad}\ P)$.

In analogy with the discussion of Sec. II we will begin by including, among the admissible symmetries, besides the tangent gauge group $\mathrm{T}.\mathscr{G}$, the full translation group on the affine space $\mathrm{T}.\mathscr{A}$.

Namely, we want to consider the action of the group $\mathrm{T}\mathscr{G}_{\mathrm{T}}$ [tangent bundle of the group $\mathscr{G}_{\mathrm{T}}$ defined in (II.1)] on $\mathrm{T}.\mathscr{A}$. This action is given by the derivative of (II.2)

$$\begin{aligned}&\mathrm{T}.\mathscr{A}\times\mathrm{T}.\mathscr{G}_T\to\mathrm{T}.\mathscr{A},\\&(A,\eta)(g,\rho,\xi,\tau)\rightsquigarrow(A^g+\rho,\mathrm{Ad}_{g^{-1}}\eta+d_{A^g}\xi+\tau).\end{aligned}\tag{V.1}$$

Since (II.2) is not free, neither is (V.1).

We have an obvious infinite-dimensional analog of Theorem IV.1,

Theorem V.1: *Any pair $(c,\hat{c})$ where c is a connection on (II.3) and $\hat{c}$ is a tangent vector to the space of connections on (II.3), defines a connection $\omega(c,\hat{c})$ on*

$$\mathrm{T}.\mathscr{A}\mapsto\frac{\mathrm{T}.\mathscr{A}}{\mathrm{T}.\mathscr{G}}.\tag{V.2}$$

Explicitly $\hat{c}$ is an assignment to each connection $A\in\mathscr{A}$ of a map $\hat{c}_A:\Omega^1(M,\mathrm{ad}\ P)\mapsto\Omega^0(M,\mathrm{ad}\ P)$ with the property of $\mathscr{G}$-equivariance,

$$\hat{c}_{A^g}(\mathrm{Ad}_{g^{-1}}\tau)=\mathrm{Ad}_{g^{-1}}(\hat{c}_A(\tau)),$$

and of tensoriality

$$\mathrm{Im}(d_A|_{\Omega^0(M,\mathrm{ad\,P})})\subset\ker(\hat{c}_A).$$

In physics $\hat{c}$ is an *infinitesimal variation of the gauge fixing*.

By using the same notation of the previous theorem, we have also the following:

Theorem V.2: *The pair* $(c,\hat{c})$ *determines a connection on the* TG *principal bundle* T$P\times$T.$\mathscr{A}\to$T$M\times$T.$\mathscr{A}$ *that is* T.$\mathscr{G}$*-invariant, i.e., determines a connection on the principal* TG*-bundle,*

$$\frac{\mathrm{T}P\times\mathrm{T}.\mathscr{A}}{\mathrm{T}.\mathscr{G}}\mapsto\mathrm{T}M\times\frac{\mathrm{T}.\mathscr{A}}{\mathrm{T}.\mathscr{G}}. \tag{V.3}$$

If we consider a rectangle $([\underline{p}],[\underline{A}])$ on T$P\times$T.$\mathscr{A}$ centered at $(p,A)\in P\times\mathscr{A}$, with $X=\dot{\underline{p}}$, and $\eta=\dot{\underline{A}}$, then the explicit expression of such a connection at $(p,X,A,\eta)\in\mathrm{T}P\times\mathrm{T}.\mathscr{A}$ is given by

$$\omega(A,\eta)([\underline{p}])+\omega(c,\hat{c})([\underline{A}]), \tag{V.4}$$

where the notation here is completely analogous to the one of Theorem IV.1.

A simple calculation shows

Theorem V.3: *When we identify the double tangent bundle* TT.$\mathscr{A}$ *with* $\mathscr{A}\times\Omega^1(M,\mathrm{ad}\,P)^{\times 3}$, *then the connection determined by* $(c,\hat{c})$ *is a map*

$$(A,\eta,\gamma,\sigma)\rightsquigarrow\left(c_A(\gamma),\left.\frac{d}{ds}\right|_{s=0}c_{A+s\eta}(\gamma)+\hat{c}_A(\gamma)+c_A(\sigma)\right).$$

From now on we set

$$\xi_{A,\eta}(\gamma)\equiv\left.\frac{d}{ds}\right|_{s=0}c_{A+s\eta}(\gamma)+\hat{c}_A(\gamma). \tag{V.5}$$

Since space–time for us is M and not TM, we want to avoid dealing with forms on TP, as opposed to forms on P.

We consider then again the inclusion map (IV.8) $P\to\mathrm{T}P$ and establish the following:

Convention i: All the forms on T$P\times$T.$\mathscr{A}$ *are pulled back to forms on* $P\times$T.$\mathscr{A}$.

In the use of Convention i we should be aware that, even though (IV.8) is a morphism of principal bundles, the pullback of a connection on T$P\times$T.$\mathscr{A}$ is *not* (strictly speaking) a connection on $P\times$T.$\mathscr{A}$. It is a $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued 1-form that would become a connection only if we decided to disregard the second copy of $\mathfrak{g}$.

Moreover, when considering the decomposition T.$\mathscr{A}\sim\mathscr{A}\times\Omega^1(M,\mathrm{ad}\,P)$ we also assume the following:

Convention ii: We will omit the components of the forms on $P\times$T.$\mathscr{A}$ *which have degree higher than 0 as forms on* $\Omega^1(M,\mathrm{ad}\,P)$.

When we assume both the previous conventions, then the connection (V.4) becomes the following $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued 1-form on $P\times\mathscr{A}$, depending on $\eta\in\Omega^1(M,\mathrm{ad}\,P)$,

$$(A+c;\eta+\xi)_{p,A,\eta}(X,\gamma)\equiv(A_p(X)+c_A(\gamma);\eta_p(X)+\xi_{A,\eta}(\gamma)),\quad X\in\mathrm{T}_pP,\gamma\in\mathrm{T}_A\mathscr{A}, \tag{V.6}$$

where (V.5) has been used.

We will, in the future, refer to (V.6) as a ''connection'' with an associate ''covariant exterior derivative,'' ''curvature'' and ''Bianchi identities;'' but, as has been explained above, this will clearly be an abuse of language.

The covariant exterior derivative for $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued forms on $P\times\mathscr{A}$ is given by

$$d_{(A,\eta)}+\pm\delta_{(c,\xi)}. \tag{V.7}$$

More precisely the covariant exterior derivative applied to the pair of $\mathfrak{g}$-valued forms (ω,ω'), with degree (k,k') as forms on P, is given by

$$(d_A\omega+(-1)^k\delta\omega+(-1)^k[c,\ \omega],d_A\omega'+(-1)^{k'}\delta\omega'+[\,\eta+(-1)^k\xi\ ,\omega]+[(-1)^{k'}c,\ \omega']), \tag{V.8}$$

where we used (III.5). The curvature of (V.6) is a $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued 2-form on $P\times\mathscr{A}$ which we can compute in a similar way as

$$(F_A+\psi+\phi;\ \ d_A\eta+\tilde{\psi}+\tilde{\phi}), \tag{V.9}$$

where $(F_A, d_A\eta)$, $(\psi,\tilde{\psi})$ and $(\phi,\tilde{\phi})$ are, respectively, the (2,0), the (1,1) and the (0,2) components.

We now compute explicitly the structure equation and the Bianchi identities.

First we obtain once again the Eqs. (II.8), (II.9), and (II.10).

The structure equations give moreover

$$\tilde{\psi}=-\delta\eta+d_A\xi+[\eta,\ c], \tag{V.10}$$

$$\tilde{\phi}=\delta\xi+[c,\ \xi]. \tag{V.11}$$

Finally the Bianchi identities for the second component of the curvature give

$$\delta\tilde{\psi}=-[\tilde{\psi},\ c]+d_A\tilde{\phi}+[\eta,\ \phi]-[\xi,\ \phi],\quad \delta\tilde{\phi}=[\tilde{\phi},\ c]+[\phi,\ \xi]. \tag{V.12}$$

So we have:

Theorem V.4: *The transformations for the set of fields* $A,\eta,c,\xi,\psi,\phi,\tilde{\psi},\tilde{\phi}$ *given by the components of the connection (V.6) and the curvature (V.9) are as follows*:

$$\delta A=d_Ac-\psi, \tag{V.13}$$

$$\delta\eta=-\tilde{\psi}+d_A\xi+[\eta,\ c], \tag{V.14}$$

$$\delta F_A=-d_A\psi+[F_A,\ c], \tag{V.15}$$

$$\delta(d_A\eta)=-d_A\tilde{\psi}+[F_A,\ \xi]+[d_A\eta,\ c]-[\psi,\ \eta], \tag{V.16}$$

$$\delta c=-\tfrac{1}{2}[c,\ c]+\phi, \tag{V.17}$$

$$\delta\xi=\tilde{\phi}-[c,\ \xi], \tag{V.18}$$

$$\delta\psi=d_A\phi-[\psi,\ c], \tag{V.19}$$

$$\delta\tilde{\psi}=-[\tilde{\psi},\ c]+d_A\tilde{\phi}+[\eta,\ \phi]-[\xi,\ \psi], \tag{V.20}$$

$$\delta\phi=[\phi,\ c], \tag{V.21}$$

$$\delta\tilde{\phi}=[\tilde{\phi},\ c]+[\phi,\ \xi]. \tag{V.22}$$

The transformation laws in Theorem V.4 show how to overcome the problem of the lack of freedom of (V.1). As before the key element has been the introduction of a (equivariant) coboundary operator relevant to a principal bundle (in this case (V.3)).

VI. RESTRICTIONS TO THE ORBITS

Two possible fiber imbeddings can be considered,

$$j_A:\mathscr{G}\hookrightarrow\mathscr{A},\ \ j_A(g)=A^g, \tag{VI.1}$$

$$j_{(A,\eta)}:\mathrm{T}\mathscr{G}\hookrightarrow\mathrm{T}\mathscr{A},\ \ j_{(A,\eta)}(g,\chi)=(A^g,\mathrm{Ad}_{g^{-1}}(\eta)+d_{A^g}\chi). \tag{VI.2}$$

First we can pull back the bundle

$$P\times\mathrm{T}\mathscr{A}\mapsto M\times\mathrm{T}\mathscr{A} \tag{VI.3}$$

to

$$P\times \mathrm{T}\mathscr{G}\mapsto M\times \mathrm{T}\mathscr{G}$$

via (VI.2).

Then we pull back the 1-form (c,ξ) (without changing notation). This form then becomes the Maurer–Cartan form on $\mathrm{T}\mathscr{G}$, and the structure equations and Bianchi identities become simply

$$\delta A = d_A c,\quad \delta c = -\tfrac{1}{2}[c,\,c],\quad \delta\xi = -[c,\xi],\quad \delta\eta = d_A\xi + [\,\eta,\,c\,]. \tag{VI.4}$$

This is the generalization to the tangent bundle of the ordinary BRST equations.[7]

Starting from (VI.1) we obtain another restriction that will allow us to have a BRST symmetry that includes the 2-form field B considered in four-dimensional BF theories.

We begin by noticing that the group $\mathscr{G}_{\mathrm{aff}}$ considered in Sec. III is the right symmetry group for the action functional (I.4). Here the action of $\mathscr{G}_{\mathrm{aff}}$ on triples (A,η,B) with $(A,\eta)\in \mathrm{T}\mathscr{A}$ and $B\in\Omega^2(M,adP)$ is defined as follows [Notice that this action of $\mathscr{G}_{\mathrm{aff}}$ on $\mathrm{T}\mathscr{A}$ does *not* coincide with the restriction of the action of $\mathscr{G}_{TP}$ on $\mathscr{A}(\mathrm{T}P)$.]

$$(A,\eta)\cdot(g,\zeta,\tau) = (A^g,\tau + Ad_{g^{-1}}\eta + d_{A^g}\zeta), \tag{VI.5}$$

$$B\cdot(g,\zeta,\tau) = Ad_{g^{-1}}B + d_{A^g}\tau + d^2_{A^g}\zeta. \tag{VI.6}$$

The group $\mathscr{G}_{\mathrm{aff}}$ is just a subgroup of $\mathrm{T}\mathscr{G}_{\mathrm{T}}$ since we have

$$\mathscr{G}_{\mathrm{aff}} \equiv j^*(\mathrm{T}\mathscr{G}_T),$$

where j is the inclusion $j:\mathscr{G}\hookrightarrow\mathscr{G}_{\mathrm{T}}$. The action (VI.5) is the corresponding restriction of (V.1).

Requiring the invariance under the group of affine gauge transformations is tantamount to considering a ''*semitopological*'' *field theory where both the invariance under* $\mathrm{T}\mathscr{G}$ *and the independency of the choice of the tangent vector* $\eta\in \mathrm{T}_A\mathscr{A}$ *is considered*.

To turn around the problem of the nonfreedom of (VI.5) we proceed along the same lines of the previous section.

First, we consider the pulled-back bundle $j_A^*\mathrm{T}\mathscr{A}$ and pull back to it the 1-form (c,ξ) (again without changing notation).

Now, as in (VI.4), c becomes the Maurer–Cartan form on $\mathscr{G}$, but ξ, differently from (VI.4), does not become any more the second component of the Maurer–Cartan form on $\mathrm{T}\mathscr{G}$. The form ξ is defined as in (V.5), provided that the connection on P belongs to the $\mathscr{G}$-orbit passing through A.

The corresponding pulled-back ''connection'' on the bundle

$$P\times j_A^*\mathrm{T}\mathscr{A}\mapsto M\times j_A^*\mathrm{T}\mathscr{A}, \tag{VI.7}$$

is written again as

$$(A+c,\eta+\xi) \tag{VI.8}$$

and the corresponding ''curvature'' is

$$(F_A, d_A\eta + \tilde{\psi} + \tilde{\phi}). \tag{VI.9}$$

The Bianchi identities become

$$[d_{(A,\eta)} \pm \delta_{(c,\xi)}](F_A, d_A\eta + \tilde{\psi} + \tilde{\phi}) = 0$$

and altogether we obtain (see Ref. 1)

Theorem VI.1: *The transformation laws for the set of fields* $A,\eta,c,\xi,\tilde{\psi},\tilde{\phi}$ *given by the components of the connection (VI.8) and of the curvature (VI.9) are as follows*:

$$\delta A = d_A c, \tag{VI.10}$$

$$\delta\eta = -\tilde{\psi} + d_A\xi + [\eta,\ c], \tag{VI.11}$$

$$\delta F_A = [F_A,\ c], \tag{VI.12}$$

$$\delta(d_A\eta) = -d_A\tilde{\psi} + [F_A,\ \xi] + [d_A\eta,\ c], \tag{VI.13}$$

$$\delta c = -\tfrac{1}{2}[c,\ c], \tag{VI.14}$$

$$\delta\xi = \tilde{\phi} - [c,\ \xi], \tag{VI.15}$$

$$\delta\tilde{\psi} = -[\tilde{\psi},\ c] + d_A\tilde{\phi}, \tag{VI.16}$$

$$\delta\tilde{\phi} = [\tilde{\phi},\ c]. \tag{VI.17}$$

By comparing (VI.10), (VI.11) with (VI.5) we see that also this last BRST operator δ allowed us to overcome the problem of the lack of freedom of (VI.5). We have still to find how to read in our BRST complex the transformations of the field B corresponding to the action (VI.6). This will accomplished in the next section.

VII. A BRST COMPLEX THAT INCLUDES 2-FORMS

Now we will show that the BRST operator δ can consistently be extended to an operator s satisfying $s^2=0$ and that a double complex with operators (d,s) can be constructed with the following properties:

(1) s acts on the space $\Omega^2(M,\mathrm{ad}\ P)$;
(2) the gauge-equivariance is preserved, and
(3) the equations considered in Theorem VI.1 are preserved.

We use the synthetic notation $\mathscr{B} \triangleq \Omega^2(M,\mathrm{ad}\ P)$ and consider the tangent bundle $\mathrm{T}\mathscr{B} \sim \Omega^2(M,\mathrm{ad}\ P)\times\Omega^2(M,\mathrm{ad}\ P)$.

The group $\mathscr{G}$ acts on $\mathscr{A}\times\mathscr{B}$ yielding a $\mathscr{G}$-principal bundle. Moreover, the tangent map to the above action gives $\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}$ the structure of a $\mathrm{T}\mathscr{G}$-principal bundle.

Explicitly this last action is given by

$$(A,\eta;C,E)\cdot(g,\zeta) = (A^g, \mathrm{Ad}_{g^{-1}}\eta + d_{A^g}\zeta; \mathrm{Ad}_{g^{-1}}C, \mathrm{Ad}_{g^{-1}}E + [\mathrm{Ad}_{g^{-1}}C,\ \zeta]). \tag{VII.1}$$

It is evident that the projection $\mathscr{A}\times\mathscr{B}\mapsto\mathscr{A}$ is a morphism of $\mathscr{G}$-bundles. Therefore, the connection c on (II.3) can also be considered as a connection on $\mathscr{A}\times\mathscr{B}$, and the connection (c,ξ) on (V.2) is also a connection on $\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}$.

Moreover, $(A+c,\eta+\xi)$ is a (pulled-back) $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued connection on the bundle

$$P\times\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}\mapsto M\times\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}. \tag{VII.2}$$

The corresponding covariant exterior derivative is given by (V.7). Forms on $P\times\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}$ will be characterized by three indices (m,e,p) which represent the degree with respect to the three spaces P, $\mathrm{T}\mathscr{A}$, $\mathrm{T}\mathscr{B}$. The integer e is, as before, called the *ghost number*.

The pair $(C,E)\in\mathrm{T}\mathscr{B}$ is a $(\mathfrak{g}\oplus_s\mathfrak{g})$-valued (2,0,0)-form that is independent of $\mathrm{T}\mathscr{A}$.

We assume Convention i and Convention ii, and also the following:

Convention iii: We will omit the components of the forms on $P\times\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}$ which have a degree higher than 0 as forms on $\mathrm{T}\mathscr{B}$.

Under the action of the covariant exterior derivative (V.7), the pair (C,E) is transformed into

$$(d_AC + [c,\ C], d_AE + [\eta,\ C] + [c,\ E] + [\xi,\ C]), \tag{VII.3}$$

owing to the fact that (C,E) does not depend on the space $\mathscr{A}$.

Definition VII.1: We define s to be equal to δ when computed on forms over $P\times\mathrm{T}\mathscr{A}$; moreover, we set for $(C,E)\in\mathrm{T}\mathscr{B}$,

$$s(C,E)\triangleq([C,c],[C,\xi]+[E,c]). \tag{VII.4}$$

We have hence set $s(C,E)$ to be equal, up to a sign, to $\delta_{(c,\xi)}(C,E)$. This takes care of requirement 2 (preservation of gauge-equivariance).

As a consequence we have

$$\delta^2_{(c,\xi)}(C,E)=[(\phi,\tilde{\phi}),(C,E)]=([\phi,C],[\tilde{\phi},C]+[\phi,E]),$$

where $(\phi,\tilde{\phi})$ is the curvature of the connection (c,ξ) on the $\mathrm{T}\mathscr{G}$ bundle $\mathrm{T}\mathscr{A}$. Here Convention i and Convention ii are always assumed.

In conclusion we have:

Theorem VII.2: *Let us assume Definition VII.1. The operator s is nilpotent in the following cases*:

(1) *always in the case of Eqs. (VI.4)*;
(2) *never in the case of the equations of Theorem V.4*;
(3) *only when pairs* $(0,E)\in \mathrm{T}_0\mathscr{B}$ *are considered in the case of the equations of Theorem VI.1*.

Remark VII.3: The reason why, when we consider the bundle (VII.2), we have nontrivial solutions for the compatibility problem of Theorem VII.2, is the fact that $\{1\}\times\mathfrak{g}\triangleleft G\ltimes\mathfrak{g}$ *in an Abelian normal subgroup.*

We now translate the field E and define

$$B\equiv E+d_A\eta. \tag{VII.5}$$

By combining (VI.13) with (VII.4) we find the following transformation property of B (see Refs. 1 and 9):

$$sB=-d_A\tilde{\psi}+[F_A,\xi]+[B,c]. \tag{VII.6}$$

Equation (VII.6) may be added to Eqs. (VI.10)→(VI.17), namely,

$$sA=d_Ac,\quad s\eta=-\tilde{\psi}+d_A\xi+[\eta,c],$$

$$sc=-\tfrac{1}{2}[c,c],\quad s\xi=\tilde{\phi}-[c,\xi],$$

$$s\tilde{\psi}=-[\tilde{\psi},c]+d_A\tilde{\phi},\quad s\tilde{\phi}=[\tilde{\phi},c].$$

While E is an element of $\mathrm{T}_0\mathscr{B}$, B is naturally an element of $\mathrm{T}_{F_A}\mathscr{B}$. In particular, triples (A,η,B) are elements of the pulled-back bundle,

$$K^*(\mathrm{T}\mathscr{A}\times\mathrm{T}\mathscr{B}),$$

where K is the curvature map

$$K:\mathscr{A}\mapsto\mathscr{A}\times\mathscr{B},\quad K(A)\triangleq(A,F_A). \tag{VII.7}$$

As a final remark, notice that once again we have overcome the problem of the lack of freedom of (VI.6).

VIII. INFINITE BRST SYMMETRIES

The procedure of the previous two sections can be iterated, so as to incorporate an arbitrary number of fields. More precisely, given any integer n we can construct field equations, similar to those considered in the previous section, depending on a connection A, on 2^n-1 fields $\gamma_i\in\Omega^1(M,\mathrm{ad}\,P)$, on 2^n fields $\eta_i\in\Omega^1(M,\mathrm{ad}\,P)$, on 2^n fields $B_i\in\Omega^2(M,\mathrm{ad}\,P)$, and on the relevant ghost-fields.

A binary notation turns out to be helpful to describe iterated tangent spaces. In fact given a manifold M we will describe an element $\mathbf{X}\in\mathrm{T}^nM$ by a 2^n-tuple as follows:

$$\mathbf{X} = \left(X_{\underbrace{0\ldots00}_{n-\text{times}}}, X_{\underbrace{0\ldots01}_{n-\text{times}}}, X_{\underbrace{0\ldots10}_{n-\text{times}}}, \ldots, X_{\underbrace{1\ldots11}_{n-\text{times}}} \right).$$

For instance in the case $n=3$, X_{000} denotes a point of M, (X_{000},X_{001}) denotes a point in $\mathrm{T}M$, $(X_{000},X_{001},X_{010},X_{011})$ denotes a point in $\mathrm{TT}M$ whose tangent vector is represented by $(X_{100},X_{101},X_{110},X_{111})$.

Elements in $\mathrm{Lie}(\mathrm{T}^nG)$ can analogously be represented as 2^n-tuples $\xi=(\xi_{i_{1,\ldots,i_n}})$, where each index i_p can be either 0 or 1, $\forall p\in\{1,\ldots,n\}$.

We now want to describe the bracket in $\mathrm{Lie}(\mathrm{T}^nG)$ induced by the bracket in $\mathfrak{g}$. We have

Theorem VIII.1: *The Lie bracket of* $\boldsymbol{\xi}$ *and* $\boldsymbol{\eta}$ *in* $\mathrm{Lie}(\mathrm{T}^nG)$ *is given by*

$$[\boldsymbol{\xi},\boldsymbol{\eta}]_{i_1\cdots i_n}=\sum_{j_i+k_i=i_i}[\xi_{j_1\cdots j_n},\eta_{k_1\cdots k_n}], \tag{VIII.1}$$

where the previous notation is understood.

Proof of Theorem VIII.1: We will prove (VIII.1) by induction on n. Formula (VIII.1) holds for $n=0$, i.e., for $\mathfrak{g}$. We assume now that it is true for n. We can represent any element $\boldsymbol{\xi}$ of $\mathrm{Lie}(\mathrm{T}^{n+1}G)$ as $(\xi_{0j_1\cdots j_n},\xi_{1j_1\cdots j_n})$ where the first component represents the base point in $\mathrm{Lie}(\mathrm{T}^nG)$, while the second component represents the relevant tangent vector. We can therefore use the expression of the bracket for $\mathrm{Lie}(\mathrm{T}G)\sim\mathfrak{g}\oplus_s\mathfrak{g}$ thus getting

$$[(\xi_{0j_1\cdots j_n},\xi_{1j_1\cdots j_n}),(\eta_{0k_1\cdots k_n},\eta_{1k_1\cdots k_n})]=([\xi_{0j_1\cdots j_n},\eta_{0k_1\cdots k_n}],[\xi_{0j_1\cdots j_n},\eta_{1k_1\cdots k_n}]+[\xi_{1j_1\cdots j_n},\eta_{0k_1\cdots k_n}]).$$

Therefore,

$$[\boldsymbol{\xi},\boldsymbol{\eta}]_{i_0i_1\cdots i_n}=\sum_{j_i+k_i=i_i}[\xi_{j_0j_1\ldots j_n},\ \eta_{k_0k_1\ldots k_n}].$$

□

We now take into account the iterated gauge group $\mathrm{T}^n\mathscr{G}$.

We can consider the Maurer–Cartan form for $\mathrm{T}^n\mathscr{G}$. This is a form with values in $\mathrm{Lie}(\mathrm{T}^n\mathscr{G})$, so it can be represented by a vector $\mathbf{c}$ with 2^n components each having ghost number one.

We set $c_0\equiv c_{0\cdots0}$, and $c_I\equiv c_{i_1\cdots i_n}$ for $I=\{i_1,\ldots,i_n\}\neq\{0\ldots0\}$, and define $I+J\triangleq\{i_1+j_1,\ldots,i_n+j_n\}$ (*not* modulo 2).

The *Maurer–Cartan* equations are as follows:

$$\delta c_0=-\tfrac{1}{2}[c_0,c_0], \tag{VIII.2}$$

$$\delta c_I=-\frac{1}{2}\sum_{J+K=I}[c_J,\ c_K]. \tag{VIII.3}$$

The group $\mathrm{T}^n\mathscr{G}$ acts (freely) on the iterated tangent space $\mathrm{T}^n\mathscr{A}$. We denote the corresponding fiber imbedding by

$$j^{(n)}_{A,\boldsymbol{\gamma}}:\mathrm{T}^n\mathscr{G}\to\mathrm{T}^n\mathscr{A},$$

where (A,γ) represents an element of $\mathrm{T}^n\mathscr{A}$ with $\boldsymbol{\gamma}\in\Omega^1(M,\mathrm{ad}\,P)^{\times 2^n-1}$.

In this section we consistently apply Convention i, Convention ii, and Convention iii. The equations describing the infinitesimal action of $\mathrm{T}^n\mathscr{G}$ and corresponding to (VI.4) are as follows:

$$\delta A=d_Ac_0, \tag{VIII.4}$$

$$\delta\gamma_I = d_A c_I + \sum_{J+K=I} [\gamma_J,\ c_K]. \tag{VIII.5}$$

In analogy with (VI.7) we apply the tangent functor once more and consider the bundle

$$P\times (j^{(n)}_{A,\boldsymbol{\gamma}})^*\mathrm{T}(\mathrm{T}^n\mathscr{A}) \to M\times (j^{(n)}_{A,\boldsymbol{\gamma}})^*\mathrm{T}(\mathrm{T}^n\mathscr{A}). \tag{VIII.6}$$

We have a vector $\boldsymbol{\eta}$ with 2^n components in $\Omega^1(M,\mathrm{ad}\,P)$ with ghost number zero and a vector $\boldsymbol{\xi}$ with 2^n components in $\Omega^0(M,\mathrm{ad}\,P)$ with ghost number one. In analogy with (VI.8) on (VIII.6) we have a (pulled-back) connection [We have to keep in mind that each of the four vectors $(A,\boldsymbol{\gamma});\boldsymbol{\eta};\mathbf{c};\boldsymbol{\xi}$ is a form on $P\times T^n\mathscr{A}$ with values in $\mathrm{Lie}(\mathrm{T}^nG)$.]

$$((A,\boldsymbol{\gamma})+\mathbf{c},\boldsymbol{\eta}+\boldsymbol{\xi}) \tag{VIII.7}$$

with values in $\mathrm{Lie}(\mathrm{T}^{n+1}G)$.

The curvature of (VIII.7) is given by

$$(F_A + d_A\boldsymbol{\gamma} + \tfrac{1}{2}[\boldsymbol{\gamma},\ \boldsymbol{\gamma}], d_A\boldsymbol{\eta} + [\boldsymbol{\gamma},\ \boldsymbol{\eta}] + \tilde{\boldsymbol{\psi}} + \tilde{\boldsymbol{\phi}}), \tag{VIII.8}$$

where the Ith component of the vector $[\boldsymbol{\gamma},\boldsymbol{\eta}]$ is defined to be $\Sigma_{J+K=I}[\gamma_J,\eta_K]$, the vector $\tilde{\boldsymbol{\psi}}$ has 2^n components in $\Omega^1(M,\mathrm{ad}\,P)$ and ghost number one, while the vector $\tilde{\boldsymbol{\phi}}$ has 2^n components, respectively, in $\Omega^0(M,\mathrm{ad}\,P)$ and ghost number two.

The results of the previous section, suggest us to introduce a vector $\mathbf{B}$ with 2^n components in $\Omega^2(M,\mathrm{ad}\,P)$ and ghost number zero.

By performing calculations similar to those carried on in the proof of Theorem VI.1 and (VII.6), we obtain the following:

Theorem VIII.2: The transformations for the set of fields A, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, $\mathbf{c}$, $\boldsymbol{\xi}$, $\tilde{\boldsymbol{\psi}}$, $\tilde{\boldsymbol{\phi}}$, $\mathbf{B}$ *are given by (VIII.2), (VIII.3), (VIII.4), (VIII.5), and*

$$\delta\eta_I = -\tilde{\psi}_I + d_A\xi_I + \sum_{J+K=I}[\gamma_J,\ \xi_K] + \sum_{J+K=I}[\eta_J,\ c_K], \tag{VIII.9}$$

$$\delta B_I = -d_A\tilde{\psi}_I - \sum_{J+K=I}[\gamma_J,\ \tilde{\psi}_K] + [F_A,\ \xi_I] + \sum_{J+K=I}[d_A\gamma_J,\ \xi_K] + \sum_{J+K=I}[B_J,\ c_K], \tag{VIII.10}$$

$$\delta\xi_I = \tilde{\phi}_I - \sum_{J+K=I}[c_J,\ \xi_K], \tag{VIII.11}$$

$$\delta\tilde{\psi}_I = -\sum_{J+K=I}[\tilde{\psi}_J,\ c_K] + d_A\tilde{\phi}_I + \sum_{J+K=I}[\gamma_J,\ \tilde{\phi}_K], \tag{VIII.12}$$

$$\delta\tilde{\phi}_I = \sum_{J+K=I}[\tilde{\phi}_J,\ c_K]. \tag{VIII.13}$$

IX. THE BATALIN–VILKOVISKY ALTERNATIVE: TOPOLOGICAL *BF* THEORIES

In Secs. VI and VII we have considered the BRST structure of a field theory with fields $(A,\eta)\in\mathrm{T}\mathscr{A}$ and $B\in\mathscr{B}$ that is gauge invariant and can be considered *topological* only as far as the translations of the field η are concerned. Equivalently such a theory admits, as its basic symmetry group, the group $\mathscr{G}_{\mathrm{aff}}$.

For a full topological BF field theory, where translations in the space $\mathscr{A}$ have to be included as symmetries, one should consider instead the action of the group $\mathscr{G}_T$ on $\mathscr{A}$ given by (II.2) and extended to $\mathscr{A}\times\mathscr{B}$ as follows:

$$(A,B)\cdot(g,\tau) = (A^g+\tau, \mathrm{Ad}_{g^{-1}}B - d_{A^g}\tau - \tfrac{1}{2}[\tau,\ \tau]). \tag{IX.1}$$

Notice that (IX.1) implies that the action of $\mathscr{G}_{\mathrm{T}}$ on $B+F_A$ is simply the adjoint action of $\mathscr{G}$.

An example of such a theory is given by the action functional (I.2) when $t=1$, i.e., by

$$S_{BF,1}=\int_M \mathrm{Tr}(F_A\wedge B+\tfrac{1}{2}B\wedge B). \tag{IX.2}$$

Again the action (IX.1) is not free, but now this problem cannot be solved by considering the coboundary (BRST) operator of a principal bundle.

In fact one would have to extend the coboundary operator δ defined on (II.4) to a nilpotent operator s acting on B, while preserving the gauge equivariance.

This would mean to consider a covariant derivative acting on $\mathscr{B}$ corresponding to a connection c on $\mathscr{A}$. The requirement of the nilpotency of such an operator would then be equivalent to the requirement that, for any $B\in\mathscr{B}$, the following condition should hold:

$$[\phi,\ B]=0, \tag{IX.3}$$

where ϕ is the curvature of c.

But we know that condition (IX.3) is not satisfied.

The so-called Batalin–Vilkovisky mechanism[10] now comes into the picture by providing us with a nilpotent coboundary operator that has the same properties of a coboundary operator for a principal bundle.

The price we have to pay is that we have to introduce fields with negative ghost-number or, equivalently, that we have to consider fields that depend explicitly on the (functional integral with respect to the) action that we are considering.

The functional integral *formally* provides us with a volume form on the space of fields

$$\omega\equiv\exp(-S)*1\equiv\exp(-S)\Pi_x\mathscr{D}A(x)\mathscr{D}B(x), \tag{IX.4}$$

where S is the action functional (IX.2). The understanding that such a volume form exists is at the heart of the Batalin–Vilkovisky mechanism, which is the same to say that the Batalin–Vilkovisky mechanism is a device that allows *integration by parts in field theories*.

Let us denote by $\mathfrak{F}$ the ''manifold'' given by $\Pi_j\Pi_{x\in M}\mathscr{F}_j(x)$, where $\{\mathscr{F}_j\}$ is the set of all fields of the theory (e.g., $A^b,B^c,c^a,\psi^d,\phi^e$, with c,ψ,ϕ being defined as in Sec. II) and the indices a,b,c,d,e labeling a basis for the Lie algebra $\mathfrak{g}$. We work here in local coordinates as far as the bundle $P(M,G)$ is concerned.

We denote by the symbol B^* the ($\mathfrak{g}$-valued) contraction with respect to the vector field corresponding to the coordinate $B^a(x)$, $x\in M$ of the ''manifold'' $\mathfrak{F}$. Namely, we set

$$(B^*)^a\triangleq\frac{\partial}{\partial B^a(x)}. \tag{IX.5}$$

Being the ghost number defined as the degree of a form on $\mathfrak{F}$ (and particularly on the space of connections $\mathscr{A}$), a vector field on $\mathfrak{F}$ will have ghost number -1. More generally, given a field σ with form degree k (on M) and ghost number s, its antifield σ^* will have form degree $4-k$ and ghost number $-s-1$. The duality between the σ and σ^* will be given by the following condition:

$$\{(\sigma^*)^a(x),\sigma^b(y)\}=\delta^a_b\delta(x-y),$$

where the Dirac volume form δ on M has been used, and $\{\cdot,\cdot\}$ denotes the Poisson bracket for the cotangent bundle $T^*\mathfrak{F}$, which is mapped onto the tangent bundle via the metric (which gives the $*$-operator).

We now denote by δ the exterior derivative on $\mathfrak{F}$. By definition of B^* we have

$$\delta(B^*(*1))=0$$

and so

$$\delta(B^*\omega)=(-B(x)-F_A(x))\omega;$$

in other words we have the formal equation

$$\delta(B^*) = -B - F_A .$$

If we require equivarance, i.e., if we want to take into account the action of the group of gauge transformations $\mathscr{G}$, then we have to replace the exterior derivative with the covariant exterior derivative which we denote again by the symbol s. So we have

$$s(B^*) = -B - F_A - [B^*, c],$$

and the ensuing transformation

$$sB = [B,\ c] + d_A\psi - [B^*,\ \phi]. \tag{IX.6}$$

It is interesting to compare (IX.6) with (VII.6). The main difference between the two equations is that, in the second case, the operator s applied to B produces *only fields*, while it produces *both fields and antifields* in the first case. In the BV mechanism, antifields are added so to render the operator s nilpotent, i.e., so as to formally reconstruct a BRST operator. This is the case of the topological BF theory. On the contrary, in the BFYM theory, antifields are unnecessary since the BRST operator is nilpotent per se.

From the previous equations we have

$$\int [(B + F_A)\omega]\mathscr{D}\mathrm{A}\mathscr{D}B = 0,$$

and in particular,

$$\frac{\partial \exp S}{\partial B} \approx 0,$$

$$\begin{aligned} F_A \exp(-S) &= \exp\left(-\frac{1}{2}\int_M \mathrm{Tr}(B\wedge B)\right) B^*\left[\exp\left(-\int_M \mathrm{Tr}(B\wedge F_A)\right)\right] \\ &\approx -\exp\left(-\int_M \mathrm{Tr}(B\wedge F_A)\right) B^*\left[\exp\left(-\frac{1}{2}\int_M \mathrm{Tr}(B\wedge B)\right)\right] \\ &= -B\ \exp(-S), \end{aligned}$$

where $\approx$ means up to a total derivative. When the action functional is the topological BF action functional (IX.2), then $B^*(x)$ is equivalent to the transformation (concentrated at x)

$$F_A(x) \rightsquigarrow -B(x). \tag{IX.7}$$

We now include the fields (with negative ghost number) A^*, c^*, ψ^*, ϕ^* conjugate respectively to A, c, ψ, ϕ. The BV equations are written in the following double-complex:

4				$sc^* = [c^*,c] - d_A A^* + [B^*,B] - [\psi^*,\psi] - [\phi^*,\phi]$	$c^* = s\phi^* + [\phi^*,c] + d_A\psi^* + \frac{1}{2}[B^*,B^*]$	ϕ^*
3			$sA^* + [A^*,c] = -d_A B - [B^*,\psi] + [\psi^*,\phi]$	$A^* = -s\psi^* + [\psi^*,c] + d_A B^*$	ψ^*	
2		$sB + [c,B] = d_A\psi + [B^*,\phi]$	$-B = sB^* + [B^*,c] + F_A$	B^*		
1	$-s\psi - [c,\psi] = -d_A\phi$	$\psi = d_A c - sA$	A			
0	$\phi = sc + \frac{1}{2}[c,c]$	c				
	2	1	0	-1	-2	-3
			s			

The degrees corresponding to rows and columns in the above diagrams are, respectively, the form degree and the ghost number. Notice that the form

$$c+A+B^*+\psi^*+\phi^*$$

is formally a ''connection'' with ''curvature''

$$\phi+\psi-B+A^*+c^*.$$

So the field B^* allows the replacement, in the BV complex, of F_A with $-B$ in the structure equations. This is consistent with transformation (IX.7) and with integration by parts in the functional integral.

At the critical point $B=-F_A$, or equivalently, if we set $B^*=0$, we obtain, from the previous complex, Eqs. (II.9) and (II.10).

In conclusion topological field theories of the cohomological type are described by the structure equations and the Bianchi identities of a suitable bundle whose cohomology is some kind of equivariant cohomology. In topological field theories which are not of cohomological type, the above description is not possible, but the BV mechanism, with the introduction of fields with negative ghost number, allows us to recover, formally *off-shell*, the algebraic structure of equivariant cohomology.

X. GAUGE FIXING AND ORBITS

In this section we would like to report briefly on the gauge-fixing problem for the action functional (I.4).

The results of Secs. VI and VII imply that we have an exact BRST symmetry when we consider the space of fields $K^*(\mathrm{T}.\mathscr{A}\times\mathrm{T}.\mathscr{B})$ and the symmetry group $\mathrm{T}.\mathscr{G}$.

The action (I.4) is invariant under $\mathscr{G}_{\mathrm{aff}}$ and hence under its subgroup $\mathrm{T}.\mathscr{G}$.

So the action (I.4) is naturally defined on the space

$$\frac{K^*(\mathrm{T}.\mathscr{A}\times\mathrm{T}.\mathscr{B})}{\mathrm{T}.\mathscr{G}}=K^*\left(\frac{\mathrm{T}.\mathscr{A}\times\mathrm{T}.\mathscr{B}}{\mathrm{T}.\mathscr{G}}\right);$$

the last identity is a consequence of the fact that the curvature map K (VII.7) commutes with the action of $\mathscr{G}$ and hence descends to a map

$$K\colon \frac{\mathscr{A}}{\mathscr{G}}\to\frac{\mathscr{A}\times.\mathscr{B}}{\mathscr{G}}.$$

We notice that the following bundles are isomorphic:

$$\frac{K^*(\mathrm{T}.\mathscr{A}\times\mathrm{T}.\mathscr{B})}{\mathrm{T}.\mathscr{G}}\simeq\frac{K^*(\mathrm{H}.\mathscr{A}\times\mathrm{T}.\mathscr{B})}{\mathscr{G}},\tag{X.1}$$

the isomorphism being induced by the linear map

$$[A,\eta,B]_{\mathrm{T}.\mathscr{G}}\rightsquigarrow[A,\eta^{\mathrm{H}},B-d_A\eta^{\mathrm{V}}]_{\mathscr{G}}.\tag{X.2}$$

Here $\mathrm{H}.\mathscr{A}$ denotes the horizontal bundle defined by a connection c on $\mathscr{A}$ and the superscript V and H denote respectively the vertical and the horizontal projections.

In order to show that this map is well-defined on the equivalence classes we compute

$$\begin{aligned}&(A^g,\mathrm{Ad}_{g^{-1}}\eta+d_A\xi,\mathrm{Ad}_{g^{-1}}B+[F_{A^g},\xi])\\&\quad\rightsquigarrow(A^g,(\mathrm{Ad}_{g^{-1}}\eta)^{\mathrm{H}},\mathrm{Ad}_{g^{-1}}B+[F_{A^g},\xi]-d_{A^g}((\mathrm{Ad}_{g^{-1}}\eta)^{\mathrm{V}}+d_{A^g}\xi))\\&\quad=(A^g,(\mathrm{Ad}_{g^{-1}}\eta)^{\mathrm{H}},\mathrm{Ad}_{g^{-1}}B-\mathrm{Ad}_{g^{-1}}(d_A\eta^{\mathrm{V}})).\end{aligned}$$

It is immediate to check that (X.2) is injective and surjective.

The action functional (I.4) is also invariant under the subgroup of $\mathscr{G}_{\text{aff}}$ given by the translations by $\Omega^1(M,\text{ad } P)$,

$$(A,\eta,B)\rightsquigarrow(A,\eta+\tau,B+d_A\tau) \quad \tau\in\Omega^1(M,\text{ad } P). \tag{X.3}$$

When our space of fields is represented by $[K^*(\mathrm{H}\mathscr{A}\times\mathrm{T}\mathscr{B})]/\mathscr{G}$, then we have to consider triples (A,η,B) with $\eta\in\mathrm{H}_A\mathscr{A}$. The action (X.3) now becomes

$$(A,\eta,B)\rightsquigarrow(A,\eta+\tau^{\mathrm{H}},B+d_A\tau^{\mathrm{H}}) \quad \tau\in\Omega^1(M,\text{ad } P). \tag{X.4}$$

The above action is of course not free, but the space of orbits is still a manifold, namely it is

$$\frac{\mathscr{A}\times\mathscr{B}}{\mathscr{G}}\sim\frac{K_2^*\mathrm{T}\mathscr{B}}{\mathscr{G}}, \tag{X.5}$$

where K_2 denotes the second component of the curvature map (VII.7).

Different gauges are possible to describe the above orbit spaces. Let us clarify first what we mean by ''choice of the gauge'' in the present framework.

We have the vector bundle

$$\mathscr{V}\triangleq K^*(\mathrm{H}\mathscr{A}\times\mathrm{T}\mathscr{B})\to\mathscr{A} \tag{X.6}$$

which is isomorphic to the vector bundle

$$(\mathrm{H}\mathscr{A}\times\mathscr{B})\to\mathscr{A}.$$

The $\mathscr{G}$-invariant fiber metric in $\mathrm{H}\mathscr{A}$ and in $\mathscr{B}$ induces a $\mathscr{G}$-invariant fiber metric in $\mathscr{V}$.

Fixing a gauge means finding a vector bundle $\mathscr{W}$ so that we have a $\mathscr{G}$-equivariant splitting

$$\mathscr{V}=\mathrm{H}\mathscr{A}\oplus\mathscr{W}.$$

The first obvious choice is to choose

$$\mathscr{W}=\mathscr{A}\times\mathscr{B}. \tag{X.7}$$

In other words the gauge condition reads: $\eta=0$, meaning that the elements of $\mathscr{W}$ have zero projection on $\mathrm{H}\mathscr{A}$. We refer to the above gauge as to the *trivial gauge*.

In order to discuss the second choice of the gauge, we restrict A to belong to the open subset of $\mathscr{A}$, where the dimension of the space $\mathrm{Harm}^1_A(M,\text{ad } P)$ of harmonic 1-forms has the lowest dimension (possibly zero). Let us denote this open set by the symbol $\hat{\mathscr{A}}$.

The symmetry (X.4) indicates that we have at our disposal a second gauge fixing, namely, we can impose on B the following condition $\langle B,d_A\eta\rangle=0$, i.e., $\langle d_A^*B,\eta\rangle=0$, $\forall\,\eta\in\mathrm{H}\mathscr{A}$. Here $\langle\cdot,\cdot\rangle$ denotes the inner product in $\Omega^*(M,\text{ad } P)$.

Our second gauge fixing is then a condition on both η and $B\in\mathscr{B}$; namely, we require $\eta\in\mathrm{H}_A\hat{\mathscr{A}}$ and $B\in\hat{\mathscr{B}}$, where $\hat{\mathscr{B}}$ is a vector bundle of 2-forms over $\hat{\mathscr{A}}$ defined by the condition

$$\hat{\mathscr{B}}_A\triangleq\{B\in\mathscr{B}\,|\,d_A^*B\in\ \mathrm{Im}(d_A|_{\Omega^0(M,\text{ad } P)})\}. \tag{X.8}$$

This choice corresponds to defining

$$\mathscr{W}_A=(\mathrm{H}_A\hat{\mathscr{A}}\ominus\mathrm{Harm}^1_A(M,\text{ad } P))\oplus\hat{\mathscr{B}}_A\,.$$

Notice that the orthogonal projection of $\mathscr{W}_A$ on $\mathrm{H}_A\hat{\mathscr{A}}$ is not any more zero, but it is given by the orthogonal complement of the harmonic 1-forms.

We refer to the above gauge as the *covariant gauge*.

Finally we consider the *self-dual gauge*.

We start by considering the projection on self-dual 2-forms;[11] $P_+:\Omega^2(M,\text{ad } P)\to\Omega^2(M,\text{ad } P)^+$. We now restrict ourselves to the open set of connections where the operator $D_A\triangleq\sqrt{2}P_+d_A:\Omega^1(M,\text{ad } P)\to\Omega^2(M,\text{ad } P)^+$ is surjective. We denote by $\tilde{\mathscr{A}}$ such an open set.

If we define also

$$D_A \triangleq \sqrt{2} d_A : \Omega^2(M, \mathrm{ad}\ P)^+ \to \Omega^3(M, \mathrm{ad}\ P)$$

and

$$D_A \triangleq d_A : \Omega^0(M, \mathrm{ad}\ P) \to \Omega^1(M, \mathrm{ad}\ P),$$

then we can consider the forms that are harmonic with respect to the operator D_A, for which we use the notation $\widetilde{\mathrm{Harm}}^*_A(M, \mathrm{ad}\ P)$.

The self-dual gauge condition is now defined as follows: $\eta \in \mathrm{H}_A \cdot \widetilde{\mathscr{A}}$ and $P_+ B = 0$.

In this way we have defined

$$\mathscr{W}_A = (\mathrm{H}_A \cdot \widetilde{\mathscr{A}} \ominus \widetilde{\mathrm{Harm}}^1_A(M, \mathrm{ad}\ P)) \oplus \cdot \mathscr{B}^-,$$

where $\mathscr{B}^-$ is the space of anti-self-dual elements of $\mathscr{B}$.

Again the orthogonal projection of $\mathscr{W}_A$ on $\mathrm{H}_A \cdot \hat{\mathscr{A}}$ is not zero, but it is given by the orthogonal complement of the 1-forms that are harmonic with respect to the operator D_A.

Finally, we consider the equations of motion relevant to the action functional (I.4) and the corresponding extrema.

Taking the directional derivatives at (A, η, B) in the directions (τ, μ, C) we get the equations

$$-t\langle [\tau, \eta], B - d_A \eta \rangle + \langle B, *d_A \tau \rangle = 0,$$

$$\langle d_A \mu, B - d_A \eta \rangle = 0,$$

$$t\langle C, B - d_A \eta \rangle + \langle C, *F \rangle = 0.$$

Noticing that

$$\langle [\tau, \eta], U \rangle = -\langle \tau, *[*U, \eta] \rangle$$

for a generic 2-form U, we get

$$t[*B - *d_A \eta, \eta] + d_A B = 0,$$

$$d_A^*(B - d_A \eta) = 0,$$

$$t(B - d_A \eta) + *F_A = 0.$$

Owing to the Bianchi identity, the second equation is redundant and the first equation is equivalent to the Yang–Mills equation $d_A^* F_A = 0$. In summary, the equations of motions of (I.4) read

$$\begin{aligned} d_A^* F_A &= 0, \\ t(B - d_A \eta) + *F_A &= 0. \end{aligned} \tag{X.9}$$

At $t = 1$ the action functional (I.4) reads

$$S_{BF\mathrm{YM}, \eta} = \tfrac{1}{2}\langle B - d_A \eta + *F_A, B - d_A \eta + *F_A \rangle - \mathrm{YM}(A). \tag{X.10}$$

The points (A, η, B) satisfying the equation

$$B = d_A \eta - *F_A \tag{X.11}$$

are the minima of (X.10) once we have fixed the connection A.

When the condition (X.11) is satisfied, then (X.10) reduces to the Yang–Mills action functional (up to a sign).

XI. CONCLUSIONS

In this paper we have first discussed the differential geometry of the tangent bundle TP of a G-principal bundle P. It is known that TP is a TG-principal bundle.

We have shown that the group of gauge transformations of TP includes, as a proper subgroup, the tangent gauge group T$\mathcal{G}$, namely, the tangent bundle of the group of gauge transformations on P.

The tangent gauge group acts on the tangent bundle of the space of connections $\mathcal{A}$ of P, and we have identified this tangent bundle with a subspace of the space of connections of TP.

Then we have considered topological and nontopological quantum field theories that admit the tangent gauge group as their symmetry group.

It is known that:

(A) The structure equations and the Bianchi identities for the bundle

$$\frac{P\times\mathcal{A}}{\mathcal{G}}$$

can be interpreted as the BRST equations for a topological quantum field theory (see Ref. 3).

(B) By taking the restriction along a $\mathcal{G}$-fiber of the bundle $P\times\mathcal{A}$, one obtains from the above structure equations the BRST equations for the Yang–Mills theory.

In this paper we have first applied the ''tangent functor'' to the above cases (A) and (B). Moreover, a special mixing of these two cases has been considered by taking the restriction to the $\mathcal{G}$-orbit of the bundle T$P\times$T$\mathcal{A}$.

In this way we have defined a theory that is ''topological'' when we consider the fields that are represented by tangent vectors to $\mathcal{A}$ and coincides with the Yang–Mills theory when we consider the connection itself.

The remarkable fact is that, in the above ''semitopological'' theory, one is able to extend the BRST operator to fields represented by 2-forms. In other words one is allowed to consider field theories that include not only tangent vectors to the space of connections, but also tangent vectors to the space of the curvatures of connections.

The most important nontrivial example of these field theories is the four-dimensional BFYM theory, which, in a separate paper,[1] has been shown to be equivalent to the ordinary Yang–Mills theory. The fact that 2-forms are among the fields of this theory can be exploited by considering vacuum expectation values of observables depending on iterated loops (loops of loops) or on imbedded surfaces in a four-dimensional manifold.[4]

The gauge fixing problem for the BFYM theory is discussed in the last section. In particular the self-dual gauge and the covariant gauge are compared.

It is rather straightforward to extend all the above calculations to iterated tangent spaces. Hence we have constructed a BRST complex that includes—besides the connection A—$2^{n+1}-1$ fields which are 1-forms, 2^n fields which are 2-forms and the relevant ghost fields. We are investigating possible relations between this infinite BRST complex and SUSY models.

Finally, we have shown that (pure) *topological* field theories of the *BF* type do not have a BRST operator related to some Lie algebra cohomology. In other words the BRST operator considered in Ref. 3 does *not* extend to 2-forms. The way of dealing with topological *BF* theories, from a field-theoretical point of view, is to resort to the Batalin–Vilkovisky formalism.

ACKNOWLEDGMENTS

We would like to thank especially our collaborators of Ref. 1, namely, F. Fucito, M. Martellini, A. Tanzini, and M. Zeni for very useful discussions. The original motivation for this paper and some of the results reported here are a result of our common work. We also thank Jim Stasheff and S. Guerra for reading the manuscript and suggesting some corrections.

[1] A. S. Cattaneo, P. Cotta-Ramusino, F. Fucito, M. Martellini, M. Rinaldi, A. Tanzini, and M. Zeni, ''Four-Dimensional Yang–Mills Theory as a Deformation of Topological BF Theory,'' preprint HUTMP-97/B362-IFUM 547/FT-ROM2F/96/61-hep-th/9705123.

[2] A. S. Cattaneo, P. Cotta-Ramusino, and M. Martellini, ''Three-Dimensional *BF* Theories and the Alexander–Conway Invariant of Knots,'' Nucl. Phys. B **346**, 355–382 (1995); A. S. Cattaneo, P. Cotta-Ramusino, J. Fröhlich, and M.

Martellini, ''Topological *BF* Theories in 3 and 4 Dimensions,'' J. Math. Phys. **36**, 6137–6160 (1995); A. S. Cattaneo, ''Cabled Wilson Loops in *BF* Theories,'' *ibid.* **37**, 3684–3703 (1996); A. S. Cattaneo, ''Abelian *BF* Theories and Knot Invariants,'' Commun. Math. Phys. **189**, 795–828 (1997).

[3] L. Baulieau and I. M. Singer, ''Topological Yang–Mills Symmetry,'' Nucl. Phys. B (Proc. Suppl.) **5B**, 12 (1988).

[4] A. S. Cattaneo, P. Cotta-Ramusino, and M. Rinaldi, ''Connections on Loop and Path Spaces and Four-Dimensional *BF* Theories'' (in preparation).

[5] M. F. Atiyah and I. M. Singer, ''Dirac Operators coupled to Vector Potentials,'' Proc. Natl. Acad. Sci. USA **81**, 2597–2599 (1984).

[6] E. Witten, ''Introduction to Cohomological Field Theories,'' Int. J. Mod. Phys. A **6**, 2775–2792 (1991).

[7] L. Bonora and P. Cotta-Ramusino, ''Some Remarks on BRS Transformations, Anomalies and the Cohomology of the Lie Algebra of the Group of Gauge Transformations,'' Commun. Math. Phys. **87**, 589–603 (1983); C. Becchi, A. Rouet, and R. Stora, ''Renormalization of the Abelian Higgs-Kibble Model,'' *ibid.* **42**, 127 (1975); I. V. Tyutin, Lebedev Institute, preprint No. 39, 1975.

[8] S. Kobayashi, ''Theory of Connections,'' Ann. Mater. Pure Appl. **43**, 119–194 (1957).

[9] F. Fucito, M. Martellini, A. Tanzini, and M. Zeni (unpublished).

[10] A. Batalin and G. A. Vilkovisky, ''Relativistic S-Matrix of Dynamical Systems with Boson and Fermion Constraints,'' Phys. Lett. B **69**, 309–312 (1977); E. S. Fradkin and T. E. Fradkina, ''Quantization of Relativistic Systems with Boson and Fermion First- and Second-Class Constraints,'' *ibid.* **72**, 343–348 (1978).

[11] S. K. Donaldson and P. B. Kronheimer, *The Geometry of Four-Manifolds* (Oxford University Press, Oxford, 1990).

Discrete time adiabatic theorems for quantum mechanical systems

A. Dranov, J. Kellendonk, and R. Seiler
Fachbereich Mathematik, Technische Universität Berlin, 10623 Berlin, Germany

(Received 17 June 1997; accepted for publication 22 October 1997)

The theory of adiabatic asymptotics is adapted to systems with discrete time evolution. The corresponding theorems about the approximation of physical time evolution by the adiabatic time evolution are shown to hold true in a discrete setting.
© *1998 American Institute of Physics.* [S0022-2488(98)01602-8]

I. INTRODUCTION

In recent years it has become more and more evident that many concepts and results of continuous analysis have interesting counterparts in a discrete setting typically on graphs or lattices. (A paradigm in this category is the discrete version of the theory of surfaces.[1]) An example of this kind is the one described here, the discrete version of the adiabatic theorem of quantum mechanics.

Originally formulated in the 1920s by Born and Fock,[2] the adiabatic theorem has evolved through a sequence of steps. Kato[3,4] has used the theorem in an important way for his analysis of perturbation theory of linear operators. Berry[5] and later Simon[6] noticed an interesting aspect of adiabatic dynamics: Adiabatic dynamics is typically geometric in the sense that it can be interpreted as parallel transport with respect to a natural connection. This point of view was important in an extension of this theorem which was motivated by the analysis of the Quantum Hall effect.[7] That article and a follow-up[8] are used in the present article as a guideline for the analysis of the discrete version of the adiabatic theorem.

In the continuous time-setting, adiabatic asymptotics applies to Hamiltonians which are time dependent in a particular way. It applies to those Hamiltonians for which there exists a large T such that the variation of the Hamiltonian in time depends essentially only on the rescaled time $s=t/T$. The time scale set up by T is supposed to be very large compared with the inverse frequencies corresponding to energy levels of the Hamiltonian at a fixed time t. It then makes sense to consider the limit $T\to\infty$, which is called the adiabatic limit.[2] For that, we have to consider a whole family $\{H(t,T)\}_T$ of time dependent Hamiltonians which is parametrized by T. The time dependent Schrödinger equation

$$i\partial_t U(t,T)=H(t,T)U(t,T), \quad U(0,T)=1,$$

is then more conveniently written in rescaled time $s=t/T$,

$$i\partial_s U_T(s)=TH_T(s)U_T(s), \quad U_T(0)=1, \tag{1}$$

where $U_T(s)=U(t,T)$ and $H_T(s)=H(t,T)$. In particular, $(d/ds)H_T(s)=T(d/dt)H(t,T)$. To repeat, the natural set up for adiabatic asymptotics is one in which $H_T(s)$ is independent of T.

The adiabatic theorems, which were considered in Refs. 7 and 8, apply to a situation in which the spectrum of $H(t,T)$ consists of two parts which are well separated by a gap. In that case one considers the adiabatic time evolution, which is the best approximation of the time evolution under the constraint that the spectral subspaces corresponding to the two parts are conserved. Its corresponding group of unitaries $U_T^A(s)$ is defined by the initial value problem

$$i\partial_s U_T^A(s)=TH_T^A(s)U_T^A(s), \quad U_T^A(0)=1, \tag{2}$$

where

 © 1998 American Institute of Physics

$$H_T^A(s)=H_T(s)+\frac{i}{T}[\partial_s P_T(s),P_T(s)], \tag{3}$$

and $P_T(s)$ is the spectral projection of $H_T(s)$ onto one part of the spectrum. Next to the result that $U_T^A{}^*U_T$ is of order T^{-1} the authors in Refs. 7 and 8 show that, under the additional requirement that $H_T(s)$ is infinitely often differentiable, the tunneling probability between the two parts of the spectrum is of order $T^{-\infty}$.

Our aim here is to develop the corresponding adiabatic theory for systems with discrete time evolution. For that we consider the discrete form of the Schrödinger equation, $n\in\mathbf{Z}$,

$$U(n+1,T)=W(n,T)U(n,T),\quad U(0,T)=1, \tag{4}$$

where $W(n,T)$ is a sequence of unitaries parametrized by T, and so is $U(n,T)$. The rescaled equation then reads, with $N=n/T$ and $W_T(N)=W(n,T)$,

$$U_T\left(N+\frac{1}{T}\right)=W_T(N)U_T(N),\quad U_T(0)=1. \tag{5}$$

One way to obtain a discrete system is by discretizing the solution of a continuous system. In fact, upon using the time ordered product ::,

$$W(n,T)=:e^{-i\int_n^{n+1}H(t,T)dt}: \tag{6}$$

gives by (4) the continuous solution at discrete times. This correspondence is useful to motivate our setup. [Alternatively, we could have started with the discrete form of equation (2) which is (5). The identity $W_T(N)=W(n,T)$ mirrors the identity $\int_N^{N+1/T}TH_T(s)ds=\int_{NT}^{NT+1}H(t,T)dt$ in the continuous version.]

Like in the continuous problem, adiabatic asymptotics makes sense if one considers the (adiabatic) limit $T\to\infty$ for a family of unitaries $W_T(N)$ which depend weakly enough on T. This will mean, as we will elaborate and specify, that $W_T(N+1/T)-W_T(N)$ is of order T^{-1}. It should be kept in mind that, in the adiabatic limit, the rescaled time $N=n/T$ remains finite.

To consider an adiabatic evolution in the discrete setting we assume that part of the spectrum of $W_T(N)$ is well separated from the rest by two gaps (in a sense specified below). Let $P_T(N)$ then denote the spectral projection onto that part. The group of unitaries $U_T^A(N)$ describing the adiabatic evolution associated with this spectral projection ought to be defined by

$$U_T^A\left(N+\frac{1}{T}\right)=W_T^A(N)U_T^A(N),\quad U_T^A(0)=1, \tag{7}$$

for a family of unitaries $W_T^A(N)$, which shall now be determined. As in the continuous setting U_T^A should preserve the eigenspace of $P_T(N)$, i.e.,

$$U_T^A(N)P_T(0)=P_T(N)U_T^A(N). \tag{8}$$

Furthermore, it should be close to U_T in the sense that for arbitrary N, $P_T(N)(U_T(N)-U_T^A(N))P_T(0)$ and $Q_T(N)(U_T(N)-U_T^A(N))Q_T(0)$ are of order T^{-1}. This can be achieved if we find an operator $v(N,M)$ such that, for arbitrary M and N, $v(N,M)-1$ is of order T^{-1} and

$$P_T(N)U_T(N)U_T^*(M)P_T(M)=P_T(N)v_T(N,M)U_T^A(N)U_T^{A*}(M)P_T(M),$$

$$Q_T(N)U_T(N)U_T^*(M)Q_T(M)=Q_T(N)v_T(N,M)U_T^A(N)U_T^{A*}(M)Q_T(M).$$

From equations (5) and (7) and the above ansatz one obtains

$$P_T\left(N+\frac{1}{T}\right)W_T(N)P_T(N)=P_T\left(N+\frac{1}{T}\right)v_T\left(N+\frac{1}{T},N\right)W_T^A(N)P_T(N),$$
$$Q_T\left(N+\frac{1}{T}\right)W_T(N)Q_T(N)=Q_T\left(N+\frac{1}{T}\right)v_T\left(N+\frac{1}{T},N\right)W_T^A(N)Q_T(N). \tag{9}$$

Provided that $v_T(N+1/T,N)$ is invertible, these equations are solved by

$$W_T^A(N)=v_T\left(N+\frac{1}{T},N\right)^{-1}S_T\left(N+\frac{1}{T},N\right)W_T(N), \tag{10}$$

where

$$S_T(N,M)=P_T(N)P_T(M)+Q_T(N)Q_T(M).$$

This follows since $W_T(N)$ commutes with its spectral projection $P_T(N)$ and $S_T(N,M)P_T(M)=P_T(N)P_T(M)$. In addition to solving (9) we want that $W_T^A(N)$ is unitary and satisfies $P_T(N+1/T)W_T^A(N)=W_T^A(N)P_T(N)$ (which insures (8)), and that $v_T(N+1/T,N)-1$ is of order T^{-2}. The latter then implies that $v_T(M,N)-1$ is of order T^{-1} for arbitrary M and N. This can all be achieved by setting $v_T(N,M)=\sqrt{S^*(N,M)S(N,M)}$. In fact,

$$V_T(N,M):=v_T(N,M)^{-1}S_T(N,M) \tag{11}$$

is then the unitary of the polar decomposition of $S_T(N,M)$. We will show below that the requirement on the dynamics, namely that $W_T(N+1/T)-W_T(N)$ is of order T^{-1}, implies for $P_T(N+1/T)-P_T(N)$ to be of order T^{-1} as well. The estimate $(1+\lambda^2)^{-1/2}=1+O(\lambda^2)$ then implies that $v_T(N+1/T,N)-1=\sqrt{1-(P_T(N)-P_T(N+1/T))^2}-1$ is of order T^{-2}. In particular, for large enough T, $v_T(M,N)$ is then invertible. The unitary operator $V_T(N,M)$ has been introduced by Kato.[4] It has been used also in Ref. 9 for the purpose of providing an explicit unitary equivalence between two projections P_1,P_2 which are close in the sense that $\|P_1-P_2\|<1$ (in the present case $P_1=P_T(N)$ and $P_2=P_T(M)$).

The main result of this article is the extension of the adiabatic theorems to the discrete setting. In particular, we shall first prove that the wave operator,

$$\Omega_T(N):=U_T^{A*}(N)U_T(N), \tag{12}$$

behaves in the adiabatic limit as $\Omega_T(N)=1+O(T^{-1})$. The notation $X=O(T^{-k})$ means, as usual, that there is a constant c such that $\|X\|\leqslant cT^{-k}$, one says that X is of order T^{-k}.

Second, defining for a sequence of operators $X(N)$, $N\in\mathbf{Z}/T$, the difference operator D as usual by

$$DX(N):=X\left(N+\frac{1}{T}\right)-X(N),$$

and $D^{(2)}=D\circ D$, etc., we shall prove that if, for any k, $D^{(k)}W_T(N)$ is of order T^{-k}, then the tunneling probability out of the spectral subspace is smaller than any inverse power of the time scale,

$$Q_T(0)\Omega_T(N)P_T(0)=O(T^{-\infty}). \tag{13}$$

II. ADIABATIC THEOREMS

A. The setup

We consider a sequence of unitary operators $W_T(N)$, $N\in\mathbf{Z}/T$, on some Hilbert-space $\mathscr{H}$ which is parametrized by a large time T. $N=n/T$ plays the role of a rescaled discrete time. This sequence is supposed to satisfy the following conditions.

(1) For a positive integer k, the value of which will be specified below, the k difference of W_T, $D^{(k)}W_T(N)$, is of order T^{-k}.

(2) The spectrum of $W_T(N)$ consists of two parts which are separated by two gaps in the following sense: There are two continuous functions $g_i:\mathbf{R}\to S^1$ and an $\epsilon>0$ such that, first, for all N and T,

$$\mathrm{dist}(\{g_1(N),g_2(N)\},\sigma(W_T(N)))>\epsilon,$$

and, second, g_1 and g_2 do not level-cross, i.e., $\forall r\in\mathbf{R}: g_1(r)\neq g_2(r)$. The two parts of the spectrum of $W_T(N)$ are then the parts which belong to the two different components of $S^1\setminus\{g_1(N),g_2(N)\}$.

Let us compare the above setup with the setup used in Ref. 7. In that work, a family of self-adjoint operators $H(s), s\in I\subset\mathbf{R}$, is investigated, the members of which are bounded from below, k times continuously differentiable and have gaps in the spectrum in the following sense: there are two continuous functions $g_i:\mathbf{R}\to\mathbf{R}$ and an $\epsilon>0$ such that, for all s, $\mathrm{dist}(\{g_1(s),g_2(s)\},\sigma(H(s)))>\epsilon$. These operators $H(s)$ are in fact nothing else than the $H_T(s)$ from (1), but they are, according to the natural adiabatic setup, independent of T.

The regularity conditions which ensure the existence of a unique dynamical evolution are already built in by requiring that all $W_T(N)$ are unitary operators. The natural adiabatic setup, in which H_T is independent of T, however does not lead to a sequence W_T of T-independent unitaries. In fact, assuming for simplicity that $H(t,T)$ commutes at different times, one obtains, formally, from (6)

$$W_T(N)=\exp(-iH_T(N)+O(T^{-1})), \tag{14}$$

the $O(T^{-1})$-term vanishing only under special circumstances. However, if $H(t,T)$ is independent of T (and differentiable), then (14) shows that $DW_T(N)=O(T^{-1})$. This is covered by our first condition with $k=1$. The first condition of our setup is also supported by the observation that boundedness of the kth derivative of $H_T(s)$ (w.r.t. s) requires the kth derivative of $H(t,T)$ (w.r.t. t) to be of order T^{-k}, and hence, to require that the k-difference of $W_T(N)$ be of order T^{-k} is the natural analog for the discrete case.

It is clear that, due to the topology of S^1, one needs two functions to separate a part of the spectrum. This is not really more restrictive than the gap condition in the continuous case, however, it should be mentioned that our second condition is *a priori* stronger than that gap condition in case the continuous Hamiltonian $H_T(s)$ were unbounded, because then the spectrum of $W_T(N)$ could be all of S^1.

B. Direct consequences of the setup

If $\Gamma_T(N)$ is a contour (traversed clockwise) around the part of spectrum of $W_T(N)$ which lies between $g_1(N)$ and $g_2(N)$, then the spectral projection $P_T(N)$ onto that part is given by the integral (which is to be understood in the strong sense)

$$P_T(N)=\frac{1}{2\pi i}\int_{\Gamma_T(N)} R_T(N,z)dz, \tag{15}$$

where $R_T(N,z)=(W_T(N)-z)^{-1}$ is the resolvent of $W_T(N)$. The large T behavior of $D^{(k)}P_T$ is of the same order in T as that of $D^{(k)}W_T$, since for large T

$$DP_T(N)=-\frac{1}{2\pi i}\int_{\Gamma_T} R_T\left(N+\frac{1}{T},z\right)DW_T(N)R_T(N,z)dz. \tag{16}$$

Here, due to condition 2 and in particular the continuity of g_i we can choose for large enough T either $\Gamma_T=\Gamma_T(N)$ or $\Gamma_T=\Gamma_T(N+1/T)$. This proves the case $k=1$, the one for larger k follows by a similar but more complicated argument.

Likewise the large T behavior of the k-differences of $V_T(N)=V_T(N+1/T,N)$ and $W_T^A(N)$ are of the same order in T as that of $D^{(k)}W_T$, as they are defined by infinitely differentiable functions

in P and one may use the discrete analog of the Leibniz rule, $D(XY)(N)=DX(N)Y(N+1/T)+X(N)DY(N)$. Furthermore, since $(1+\lambda^2)^{-1/2}=1+O(\lambda^2)$, the large T-behavior of $V_T(N)$ is given by

$$V_T(N)=1+\left(P_T\left(N+\frac{1}{T}\right)-P_T(N)\right)(2P_T(N)-1)+O(T^{-2}). \tag{17}$$

The unitary operator which describes the discrete time evolution of the wave operator,

$$\Theta_T(N):=\Omega_T\left(N+\frac{1}{T}\right)\Omega_T^*(N), \tag{18}$$

is easily shown to be

$$\Theta_T(N)=U_T^{A*}\left(N+\frac{1}{T}\right)V_T^*\left(N+\frac{1}{T},N\right)U_T^A\left(N+\frac{1}{T}\right). \tag{19}$$

Instead of using Θ_T we shall make use of the operator $K_T(N)=T(1-\Theta_T(N))$ and formulate equation (18) in the way $D\Omega_T=-T^{-1}K_T\Omega_T$, or, summed up,

$$\Omega_T\left(\frac{l}{T}\right)=1-\frac{1}{T}\sum_{n=0}^{l-1}K_T\left(\frac{n}{T}\right)\Omega_T\left(\frac{n}{T}\right). \tag{20}$$

Note that $K_T(N)=O(1)$.

C. The theorem about partial summing up

The theorem and corollary of this section furnish the technical basis of this work.

Theorem 1: *Let $W_T(N)$, $N\in\mathbf{Z}/T$, be a sequence of unitaries satisfying the conditions of our setup, condition 1 with $k=1$, and suppose that $X(N)$ and $Y(N)$ are sequences of operators such that $DX(N)$ and $DY(N)$ are of order T^{-1}. Let*

$$\widetilde{X}(N)=-\frac{1}{2\pi i}\int_{\Gamma_T(N)}R_T(N,z)X(N)R_T(N,z)dz, \tag{21}$$

where $R_T(N,z)$ is the resolvent of $W_T(N)$. Then

$$\sum_{n=1}^{l}Q_0U_T^{A*}\left(\frac{n}{T}\right)X\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right)P_0Y\left(\frac{n}{T}\right)=\mathscr{B}+\frac{1}{T}\mathscr{S}, \tag{22}$$

where $P_0=P_T(0)$ and $Q_0=Q_T(0)$,

$$\mathscr{B}=Q_0U_T^{A*}(0)\widetilde{X}\left(\frac{1}{T}\right)U_T^A\left(\frac{1}{T}\right)P_0Y\left(\frac{1}{T}\right)-Q_0U_T^{A*}\left(\frac{l}{T}\right)\widetilde{X}\left(\frac{l+1}{T}\right)U_T^A\left(\frac{l+1}{T}\right)P_0Y\left(\frac{l+1}{T}\right) \tag{23}$$

is a boundary term and

$$\mathscr{S}=\sum_{n=1}^{l}Q_0U_T^{A*}\left(\frac{n}{T}\right)\left(Z\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right)P_0Y\left(\frac{n}{T}\right)+\widetilde{X}\left(\frac{n+1}{T}\right)W_T^A\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right)P_0TDY\left(\frac{n}{T}\right)\right) \tag{24}$$

is a sum, and

$$Z\left(\frac{n}{T}\right)=T\left(\left[\left(V_T^*\left(\frac{n}{T}\right)-1\right)W_T^A\left(\frac{n}{T}\right),\widetilde{X}\left(\frac{n}{T}\right)\right]+DW_T^A\left(\frac{n-1}{T}\right)\widetilde{X}\left(\frac{n}{T}\right)+D\widetilde{X}\left(\frac{n}{T}\right)W_T^A\left(\frac{n}{T}\right)\right). \tag{25}$$

Moreover, $Z(n/T)$ and $TDY(n/T)$ remain bounded in the adiabatic limit.

Proof: Note first the identity

$$Q_T(N)X(N)P_T(N) = -Q_T(N)[W_T(N),\widetilde{X}(N)]P_T(N) \tag{26}$$

which follows from

$$\begin{aligned}[W_T(N),\widetilde{X}(N)] &= -\frac{1}{2\pi i}\int_{\Gamma_T(N)}[W_T(N)-z,R_T(N,z)X(N)R_T(N,z)]dz \\ &= [P_T(N),X(N)].\end{aligned}$$

Substituting W_T by the expression of the adiabatic time evolution one gets

$$Q_T(N)X(N)P_T(N) = -Q_T(N)[W_T^A(N),\widetilde{X}(N)]P_T(N) + Q_T(N)[A(N),\widetilde{X}(N)]P_T(N), \tag{27}$$

where $A=(V_T^*-1)W_T^A=O(T^{-1})$. Using the intertwining properties (8) to obtain

$$Q_0U_T^{A^*}XU_T^AP_0 = Q_0U_T^{A^*}[W_T^A,\widetilde{X}]U_T^AP_0 + Q_0U_T^{A^*}[A,\widetilde{X}]U_T^AP_0$$

and the adiabatic equation of motion one arrives at

$$\begin{aligned}U_T^{A^*}\left(\frac{n}{T}\right)\left[W_T^A\left(\frac{n}{T}\right),\widetilde{X}\left(\frac{n}{T}\right)\right]U_T^A\left(\frac{n}{T}\right) &= U_T^{A^*}\left(\frac{n}{T}\right)W_T^A\left(\frac{n-1}{T}\right)\widetilde{X}\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right) - U_T^{A^*}\left(\frac{n}{T}\right)\widetilde{X}\left(\frac{n}{T}\right)U_T^A\left(\frac{n+1}{T}\right) \\ &\quad + U_T^{A^*}\left(\frac{n}{T}\right)\left(W_T^A\left(\frac{n}{T}\right) - W_T^A\left(\frac{n-1}{T}\right)\right)\widetilde{X}\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right) \\ &= U_T^{A^*}\left(\frac{n-1}{T}\right)\widetilde{X}\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right) - U_T^{A^*}\left(\frac{n}{T}\right)\widetilde{X}\left(\frac{n+1}{T}\right)U_T^A\left(\frac{n+1}{T}\right) \\ &\quad + U_T^{A^*}\left(\frac{n}{T}\right)B\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right),\end{aligned}$$

where

$$B\left(\frac{n}{T}\right) = (D\widetilde{X})\left(\frac{n}{T}\right)W_T^A\left(\frac{n}{T}\right) + (DW_T^A)\left(\frac{n-1}{T}\right)\widetilde{X}\left(\frac{n}{T}\right).$$

By hypothesis DX, and hence also $D\widetilde{X}$, and DW_T^A are of order T^{-1}. Thus $Z(n/T) = T([A(n/T),\widetilde{X}(n/T)]+B(n/T))$ is of order 1. Summing up yields the statement of the theorem. q.e.d.

Note that the sum $\mathscr{S}$ (24) is of the same form as the l.h.s. of (22). One might therefore apply the theorem iteratively if the boundary terms vanish. This is the content of the next corollary.

Corollary 1: Let in addition to the conditions of the last theorem $D^{(k)}W_T(N)$, $D^{(k)}X(N)$, and $D^{(k)}Y(N)$ be of order T^{-k} for all k and X or Y be supported inside $(0,s)$. Then for $l \geqslant sT$,

$$\frac{1}{T}\sum_{n=1}^{l} Q_0U_T^{A^*}\left(\frac{n}{T}\right)X\left(\frac{n}{T}\right)U_T^A\left(\frac{n}{T}\right)P_0Y\left(\frac{n}{T}\right) = O(T^{-\infty}). \tag{28}$$

Proof: Note first that, if $D^{(k)}W_T$, $D^{(k)}X$, and $D^{(k)}Y$ are of order T^{-k}, then $D^{(k-1)}Z$ is of order T^{-k+1}. Therefore one can in this case apply the above theorem k times iteratively, each time gaining one order in T^{-1}, provided the boundary terms $\mathscr{B}$ on the r.h.s. of (22) vanish. We assert that this is the case under the assumptions on the support of X or Y. For that, note that, for arbitrary X, $\mathrm{supp}X\subset[a,b]$, implies that $\mathrm{supp}DX\subset[a,b+1/T]$. By assumption, the support of X or that of Y is contained in $[\epsilon,s-\epsilon]$ for some $\epsilon>0$. This implies that $\mathrm{supp}Z\subset[\epsilon,s-\epsilon+1/T]$ or $\mathrm{supp}DY\subset[\epsilon,s-\epsilon+1/T]$. Therefore, if $\epsilon>1/T$, then the boundary terms on the r.h.s. of (22), which could arise upon one-fold application of the theorem, actually vanish. Inductively one concludes that if $T>k/\epsilon$, no boundary terms arise upon k-fold application of the theorem. Thus, after k-fold application of the theorem, the sum of interest has the form

$$\frac{1}{T}\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) X\left(\frac{n}{T}\right) U_T^{A}\left(\frac{n}{T}\right) P_0 Y\left(\frac{n}{T}\right) = \frac{1}{T^{k+1}}\sum_{i=1}^{2^k}\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) X_i\left(\frac{n}{T}\right) U_T^{A}\left(\frac{n}{T}\right) P_0 Y_i\left(\frac{n}{T}\right), \tag{29}$$

where X_i and Y_i are sequences of bounded operators, provided $T>k/\epsilon$. In particular, the large T behavior of the sum is bounded by any power of T^{-1}. q.e.d.

D. The first adiabatic theorem

The first adiabatic theorem states that the physical time evolution is approximated by the adiabatic time evolution. It is obtained upon estimating the large T behavior of Ω_T. For that we shall first investigate a sum of the form $T^{-1}\Sigma_{n=0}^{l-1}K_T(n/T)Y(n/T)$ where l is of the order of T.

Let us begin with the ''diagonal'' term $P_0K_T(n/T)P_0Y(n/T)$. Since, for any two projections p,q, holds $p(p-q)p=p(p-q)^2p$ equation (17) implies that $P_T(N)(1-V_T^*(N))P_T(N)=O(T^{-2})$, and therefore

$$\left\|\frac{1}{T}\sum_{n=0}^{[sT]} P_0K_T\left(\frac{n}{T}\right)P_0Y\left(\frac{n}{T}\right)\right\| = \left\|\sum_{n=1}^{[sT]+1} U_T^{A*}\left(\frac{n}{T}\right)P_T\left(\frac{n}{T}\right)\left(1-V^*\left(\frac{n}{T}\right)\right)P_T\left(\frac{n}{T}\right)U_T^{A}\left(\frac{n}{T}\right)Y\left(\frac{n-1}{T}\right)\right\|$$

$$\leqslant \frac{\text{const}}{T}\|Y\|_s, \tag{30}$$

where

$$\|Y\|_s = \sup_{0\leqslant n\leqslant sT}\left\|Y\left(\frac{n}{T}\right)\right\|. \tag{31}$$

The same estimate holds for the norm of $T^{-1}\Sigma_{n=0}^{[sT]}Q_0K_T(n/T)Q_0Y(n/T)$. To treat the ''off diagonal'' terms we apply Theorem 1, taking $X(N)=T(1-V_T^*(N))=TDP_T(N)(2P_T(N)-1)+O(T^{-1})$. In particular, $DX(N)$ is of order T^{-1} in the case W_T satisfies condition 1 with $k\geqslant 2$. The terms of

$$\frac{1}{T}\sum_{n=0}^{[sT]} Q_0K_T\left(\frac{n}{T}\right)P_0Y\left(\frac{n}{T}\right) = \frac{1}{T}\left(\mathscr{B}+\frac{1}{T}\mathscr{S}\right),$$

where $\mathscr{B}$ and $\mathscr{S}$ are as in Theorem 1, are bounded from above by

$$\|\mathscr{B}\|\leqslant \text{const}\|Y\|_s$$

and

$$\|\mathscr{S}\|\leqslant \text{const}T\|\|Y\|\|_s$$

with

$$\|\|Y\|\|_s = \max\{\|Y\|_s, T\|DY\|_s\}$$

and T-independent constants. Again, the same estimate holds for the norm of the other ''off diagonal'' $T^{-1}\Sigma_{n=0}^{[sT]}P_0K_T(n/T)Q_0Y(n/T)$. Summarizing we obtain

$$\left\|\frac{1}{T}\sum_{n=0}^{[sT]} K_T\left(\frac{n}{T}\right)Y\left(\frac{n}{T}\right)\right\| \leqslant \frac{\text{const}}{T}\|Y\|\|_s. \tag{32}$$

Theorem 2: *If $W_T(N)$, $N\in\mathbf{Z}/T$, is a sequence of unitaries satisfying the conditions of our setup, condition 1 with $k\geqslant 2$, then*

$$\Omega_T(N)=1+O(T^{-1}). \tag{33}$$

Proof: Define $\Omega_0 := 1$ and, for $j \geqslant 1$,

$$\Omega_j\left(\frac{l}{T}\right) := -\frac{1}{T}\sum_{n=0}^{l-1} K_T\left(\frac{n}{T}\right)\Omega_{j-1}\left(\frac{n}{T}\right). \tag{34}$$

Then, by (32) we have, for $j \geqslant 1$,

$$\left\|\Omega_j\left(\frac{l}{T}\right)\right\| \leqslant \frac{\text{const}}{T}\||\Omega_{j-1}\||_{l/T}\|. \tag{35}$$

But $D\Omega_j = -(1/T)K_T\Omega_{j-1}$ so that, for $j \geqslant 1$,

$$\||\Omega_j\||_s = \max\{\|\Omega_j\|_s, \|\Omega_{j-1}\|_s\}.$$

Since $\||\Omega_0\||_s = 1$ and hence, by (35), $\|\Omega_1\|_s = O(T^{-1})$, one obtains by induction on j, for all $j > 1$,

$$\|\Omega_{2j-1}\|_s + \|\Omega_{2j}\|_s = O(T^{-j}). \tag{36}$$

Furthermore, let

$$R_M := \Omega_T - \sum_{j=0}^{M} \Omega_j\,. \tag{37}$$

Then

$$\begin{aligned} R_M\left(\frac{l}{T}\right) &= -\frac{1}{T}\sum_{n=0}^{l-1} K_T\left(\frac{n}{T}\right)\Omega_T\left(\frac{n}{T}\right) + \frac{1}{T}\sum_{j=1}^{M}\sum_{n=0}^{l-1} K_T\left(\frac{n}{T}\right)\Omega_{j-1}\left(\frac{n}{T}\right) \\ &= -\frac{1}{T}\sum_{n=0}^{l-1} K_T\left(\frac{n}{T}\right)R_{M-1}\left(\frac{n}{T}\right), \end{aligned}$$

which shows that R_M satisfies the same equation as Ω_k. Hence, for $M \geqslant 1$,

$$\left\|R_M\left(\frac{l}{T}\right)\right\| \leqslant \frac{\text{const}}{T}\||R_{M-1}\||_{l/T}\,.$$

Since $R_0 = \Omega_T - 1$ is of order 1 in T and $DR_0 = -(1/T)K_T\Omega_T$ is of order T^{-1}, the above formula shows that R_1 is of order T^{-1} and, analogously, for all $M > 1$,

$$\|R_{2M-1}\|_s + \|R_{2M}\|_s = O(T^{-M}). \tag{38}$$

As $\Omega_T = 1 + R_1 + \Omega_1$, and $\|\Omega_T(N) - 1\| \leqslant \|\Omega_T - 1\|_N$, the statement of the theorem follows. q.e.d.

E. The second adiabatic theorem

We have just seen that the adiabatic time evolution approximates the physical time evolution. By construction, it preserves the spectral subspace corresponding to the projection $P_T(N)$. The quantities $\|Q_0\Omega_T P_0\|$ and $\|P_0\Omega_T Q_0\|$ are therefore upper bounds to the probabilities for the physical system to tunnel out and into, respectively, that spectral subspace. In case $D^{(k)}W_T(N)$ is of order T^{-k} for all k we shall show now that these probabilities are smaller than any inverse power of the time scale. For the remainder of this section we assume that $W_T(N)$, $N \in \mathbf{Z}/T$, is a sequence of unitaries satisfying the conditions of our setup, but condition 1 for all $k \geqslant 1$.

Recall that $\Omega_0 = 1$ and, for $j \geqslant 1$,

$$Q_0\Omega_j\left(\frac{l}{T}\right)P_0 = -\frac{1}{T}\sum_{n=1}^{l} Q_0 K_T\left(\frac{n-1}{T}\right)\Omega_{j-1}\left(\frac{n-1}{T}\right)P_0\,. \tag{39}$$

Lemma 1: Fix an integer $j \geqslant 1$. *Let* X *be supported inside* $(0,s)$ *and* $D^k X(N) = O(T^{-k})$ *for all* k, *and assume that, for* $l \geqslant sT$,

$$Q_0 \Omega_m\left(\frac{l}{T}\right) P_0 = O(T^{-\infty}) \quad \text{for all } 0 \leqslant m \leqslant j-1. \tag{40}$$

Then, for all $1 \leqslant j' \leqslant j$ *holds*

$$\begin{aligned}
&\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) X\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 \\
&\quad = \gamma(j')\frac{1}{T}\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) Z\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 + \gamma(j'-1)\frac{1}{T} \\
&\qquad \times \sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) Z_1\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-2}\left(\frac{n}{T}\right) P_0 + O(T^{-\infty}),
\end{aligned} \tag{41}$$

where $\gamma(j') = 1$ *for* $j' > 1$ *and* $\gamma(j') = 0$ *otherwise,*

$$Z_1\left(\frac{n}{T}\right) = \widetilde{X}\left(\frac{n+1}{T}\right) W_T^A\left(\frac{n}{T}\right) P_T\left(\frac{n}{T}\right) T\left(V_T^*\left(\frac{n-1}{T}\right) - 1\right),$$

and Z *and* $\widetilde{X}$ *are the same as in Theorem 1, cf. (21) and (25). Furthermore,* $D^k Z(n/T) = O(T^{-k})$ *and* $D^k Z_1(n/T) = O(T^{-k})$ *for all* k.

Proof: For $j' = 1$ the announced result is a special case of Corollary 1, namely that where $Y(N) = 1$ for all N. So let us consider the case $j' > 1$. Since $Q_0 \Omega_{j'-1}(n/T) P_0 = O(T^{-\infty})$,

$$\begin{aligned}
&\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) X\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 \\
&\quad = \sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) X\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) P_0 \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 + O(T^{-\infty}) \\
&\quad = \frac{1}{T}\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) Z\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 \\
&\qquad + \frac{1}{T}\sum_{n=1}^{l} Q_0 U_T^{A*}\left(\frac{n}{T}\right) \widetilde{X}\left(\frac{n+1}{T}\right) W_T^A\left(\frac{n}{T}\right) U_T^A\left(\frac{n}{T}\right) P_0 T D \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 + O(T^{-\infty}),
\end{aligned}$$

the last equality following by Theorem 1 with $Y(n/T) = \Omega_{j'-1}(n/T) P_0$ provided T is large enough. In fact, as for Corollary 1 one shows that the boundary terms vanish if $T > 1/\epsilon$ where $\epsilon > 0$ is chosen such that $\mathrm{supp} X(N) \subset [\epsilon, s-\epsilon]$. Using that, for $j' > 1$, $D\Omega_{j'-1}(n/T) P_0 = -U_T^{A*}(n/T)(1 - V_T^*((n\text{-}1)/T)) U_T^A(n/T) \Omega_{j'-2}(n/T) P_0$ and the intertwining properties (8) one gets

$$\begin{aligned}
U_T^A\left(\frac{n}{T}\right) P_0 T D \Omega_{j'-1}\left(\frac{n}{T}\right) P_0 &= U_T^A\left(\frac{n}{T}\right) P_0 U_T^{A*}\left(\frac{n}{T}\right) T\left(V_T^*\left(\frac{n-1}{T}\right) - 1\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-2}\left(\frac{n}{T}\right) P_0 \\
&= P_T\left(\frac{n}{T}\right) T\left(V_T^*\left(\frac{n-1}{T}\right) - 1\right) U_T^A\left(\frac{n}{T}\right) \Omega_{j'-2}\left(\frac{n}{T}\right) P_0 .
\end{aligned} \tag{42}$$

This shows (41) for $j' > 2$. If $j' = 2$ then we may apply Corollary 1 to the term containing $\Omega_{j'-2}(n/T)$, obtaining that this term is $O(T^{-\infty})$. It was already shown in the proof of Corollary 1 that $D^k Z(n/T) = O(T^{-k})$. That $D^k Z_1(n/T) = O(T^{-k})$ follows from the discrete Leibniz rule, since all factors satisfy this condition. q.e.d.

Theorem 3: *Suppose that* $W_T(N)$ *satisfies the conditions 1 and 2 for all* k *and that* $V_T^*(N)-1$ *be supported within* $(0,s)$. *Then for* $N\notin(0,s)$

$$Q_0\Omega_T(N)P_0=O(T^{-\infty}). \tag{43}$$

Proof: Recall that, for all M, $Q_0\Omega_T P_0=Q_0R_MP_0+\Sigma_{j=1}^{M}Q_0\Omega_jP_0$. We show by induction on j that, for $l\geqslant sT$ and $j\geqslant 1$,

$$Q_0\Omega_j\left(\frac{l}{T}\right)P_0=O(T^{-\infty}). \tag{44}$$

The theorem then follows from (38). Equation (44) is true for $j=1$, since, by Corollary 1 applied to $X=1-V_T^*$ and $Y=1$,

$$\begin{aligned}Q_0\Omega_1\left(\frac{l}{T}\right)P_0&=-\frac{1}{T}\sum_{n=1}^{l}Q_0K_T\left(\frac{n-1}{T}\right)P_0\\&=-\sum_{n=1}^{l}Q_0U_T^{A*}\left(\frac{n}{T}\right)\left[1-V_T^*\left(\frac{n}{T}\right)\right]U_T^{A}\left(\frac{n}{T}\right)P_0=O(T^{-\infty}).\end{aligned}$$

To prove (44) for $j>1$ we assume (40) and use Lemma 1 repeatedly. For that we take first, in view of (39), $X=V_T^*-1$, and in each further step apply that lemma separately with X being equal to the arising Z's and Z_1's from the last step. This is possible as long as the resulting Z and Z_1 satisfy also the remaining condition for the lemma, that is, as long as they have their support inside $(0,s)$. But as for Corollary 1 one shows that this is indeed the case if $T>k/\epsilon$, where $\mathrm{supp}X(N)\subset[\epsilon,s-\epsilon]$ for all N. Thus if $T>k/\epsilon$, k-fold application of Lemma 1 results in

$$Q_0\Omega_j\left(\frac{l}{T}\right)P_0=\frac{1}{T^k}\sum_{i=1}^{\min\{k,j-1\}}\sum_{n=1}^{l}Q_0U_T^{A*}\left(\frac{n}{T}\right)X_i\left(\frac{n}{T}\right)U_T^{A}\left(\frac{n}{T}\right)\Omega_{j-i}\left(\frac{n}{T}\right)P_0+O(T^{-\infty}), \tag{45}$$

where each X_i contains at most 2^k terms which all remain bounded in the adiabatic limit. Since also all Ω_j remain bounded in the adiabatic limit it follows that $Q_0\Omega_j(l/T)P_0=O(T^{-k})$ for all k. q.e.d.

That $P_0\Omega_T(N)Q_0=O(T^{-\infty})$ is proven likewise.

ACKNOWLEDGMENTS

The authors wish to thank Markus Klein for helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft.

[1] A. Bobenko and U. Pinkall, ''Discrete surfaces with constant negative Gaussian curvature and the Hirota equation,'' J. Diff. Geom. **43**, 527–611 (1996).
[2] M. Born and V. Fock, ''Beweis des Adiabatensatzes,'' Z. Phys. **51**, 165–169 (1928).
[3] T. Kato, ''On the adiabatic theorem of quantum mechanics,'' J. Phys. Soc. Jpn. **5**, 435–439 (1950).
[4] T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, Berlin, 1966).
[5] M. V. Berry, ''Quantal phase factors accompanying adiabatic changes,'' Proc. R. Soc. London, Ser. A **392**, 45–57 (1984).
[6] B. Simon, ''Holonomy, the quantum adiabatic theorem and Berry's phase,'' Phys. Rev. Lett. **51**, 2167–2170 (1983).
[7] J. Avron, R. Seiler, and L. Yaffe, ''Adiabatic theorems and applications to the quantum Hall effect,'' Commun. Math. Phys. **110**, 33–49 (1987).
[8] M. Klein and R. Seiler, ''Power-law corrections to the Kubo formula vanish in quantum Hall systems,'' Commun. Math. Phys. **128**, 141–160 (1990).
[9] J. Avron, R. Seiler, and B. Simon, ''The index of a pair of projections,'' J. Funct. Anal. **120**, 220–237 (1994).

Microscopic $SU_\pi(3)\otimes SU_\nu(3)$ interpretation of a two triaxial rotor model

G. F. Filippov
Bogolyubov Institute for Theoretical Physics, 252143, Kiev-143, Ukraine

A. F. Lisetskyi
Taras Shevchenko Kiev University, 252022, Kiev–22, Ukraine

J. P. Draayer
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803

(Received 30 July 1997; accepted for publication 5 November 1997)

A description of the relative motion of neutrons and protons in strongly deformed even-even nuclei is reduced to that of two coupled triaxial rotors, one representing the neutrons and the other the protons, by using generating function methods. The two rotor theory is interpreted as a generalization of the Elliott SU(3) model with a neutron–proton direct product $SU_\nu(3)\times SU_\pi(3)$ substructure. The second order Casimir operator for this system is constructed and some of its eigenfunctions are found in explicit form in the Fock–Bargmann space representation. Analytical results for isovector $M1$ transitions probabilities are obtained and the results compared with experimental data and pseudo-SU(3) calculations. Graphs picturing the relative orientation of neutron-proton rotors in ^{22}Ne are presented. © *1998 American Institute of Physics.* [S0022-2488(98)03202-2]

I. INTRODUCTION

A ''giant resonance'' description of a nucleus[1,2] suggests the possible existence of new collective modes of excitations in deformed nuclei. In analogy with the relative vibrational motion between the proton and neutron distributions within a nucleus, which gives rise to the well-known giant dipole resonance ($E1$ mode), there is the possibility of relative angular oscillations of these two distributions. The simplest collective orbital excitation of this type has been denoted the ''scissors mode'' since the corresponding motion mimics the opening and closing of a scissors. The actual existence of the scissors mode in deformed nuclei was first predicted by Lo Iudice and Palumbo in terms of the Two Rotor Model (TRM).[3] In the TRM the protons and neutrons are assumed to rotate independently except for an interaction through a harmonic potential.

Approximately ten years ago the scissors mode was observed by high resolution inelastic electron scattering experiments as a low-lying collective 1^+ state.[4–6] The experiment results indicate that these states have mainly an orbital and not a spin-flip character, because corresponding peaks were not found in inelastic proton scattering data.[5,6] Following the announcement of the experimental findings, many studies were published attempting to explain the microscopic structure of this mode.[7–17] Models used in various efforts to describe these collective $M1$ excitations include the IBM,[7–9] pseudo-SU(3) model,[10] RPA,[11–13] quasiparticle RPA,[14] VPM[15] and SRA.[16] Some of these calculations even predict a fragmentation of the scissors mode.[10,11,13] Most of the theoretical descriptions assign an isovector character to the mode and emphasize the algebraic structure of the excitation.

The purpose of this paper is to begin to build a bridge between the algebraic and geometric pictures of the scissors mode in deformed nuclei. In so doing, the underlying geometry of the SU(3) model will emerge as a key ingredient in understanding the nature of these collective excitations, both the summed strength and its fragmentation across several final states. Subsequent papers will deal with more in depth analyses and applications of the theory.

The geometric TRM developed by Lo Iudice *et al.*[3,17] can be considered to be a semiclassical explanation of the nature of a new collective mode. The original version of the model describes a nucleus as simply a system of two axial rotors. The earliest version of the theory gives wave

 © 1998 American Institute of Physics

functions for small values of the angle between the symmetry axes of the neutron and proton distributions. This picture also makes predictions for $B(M1)$ and $B(E2)$ strengths, excitation energies and TRM form-factors. Although the qualitative features of this version of the TRM are correct, the quantitative estimations for $B(M1)$ values often differ strongly the experimental data.[4]

A fully microscopic version of the TRM based on a systematic implementation of Fock–Bargmann space concepts has been recently developed in Refs. 18–20. In these studies the TRM is interpreted as a generalization of the Elliott SU(3) model.[21] Inasmuch as the simplest Elliott model describes the dynamics of only three collective degrees of freedom for the valence nucleons (whereas for the case of two rotors the number of degrees of freedom can range between four and six depending upon specifics of the two rotor geometries), it is necessary to extend the model to introduce a more sophisticated classification of the basis states. Basis states constructed as tensor products of the separate neutron and proton subdistributions, with the corresponding symmetry $\mathrm{SU}_\pi(3)\otimes\mathrm{SU}_\nu(3)$,[28] are required. A subsequent reduction of the tensor product with respect to SU(3) makes it possible to label the wave functions by the quantum numbers (λ,μ) characterizing the SU(3) symmetry of the neutron–proton system as a whole. The further reduction $\mathrm{SU}(3)\rightarrow\mathrm{SO}(3)$ results in an L,M,K classification of the states, where L is an orbital angular momentum, M is its projection on an external (laboratory) z-axis, and K is the angular momentum projection onto an intrinsic ζ-axis. (In general, when the λ and μ of both the proton and neutron subdistributions are nonzero, an additional label is required to uniquely identify the coupled SU(3) representation. This additional quantum number is called the outer multiplicity label. In this paper only multiplicity-free couplings will be considered; the physical significance of the multiplicity label will be considered separately.)

Since the Hamiltonian is always taken to be rotationally invariant and space is isotropic, the only exact symmetries of the system are the angular momentum L and its projection M in the laboratory system. In general, the projection of the angular momentum on the intrinsic axis, K, is not an integral of the motion so eigenstates are expected to be a superposition of states with different K values. It should be noted, however, that this label is only required when both λ and μ are greater than or equal to two, and in this case it can be chosen, for example, as the eigenvalue of the Bargmann-Moshinsky ω operator.[23,24] Indeed, it was shown by Elliott[21] that the (λ,μ) irreducible representation (irrep) can be reduced into irreps of SO(3) according to the following rule:

$$L=N,N+1,N+2,\ldots,N+B,$$

where $N=\min\{\lambda,\mu\}$, $\min\{\lambda,\mu\}-2$, $\min\{\lambda,\mu\}-4,\ldots,1$ or 0; and $B=\max\{\lambda,\mu\}$, except for the case $N=0$, when $L=B,B-2,\ldots,1$ or 0. In the present study we consider only the problem of configurations that do not require the use of the additional K quantum label.

In the present paper we extend the microscopic version of the TRM[18–20] to the case of two triaxial rotors. This extension, which is complementary to the pseudo-SU(3) approach, gives rise to a fragmentation of the scissors mode which is in keeping with experimental observations. The outline of the paper is as follows: In Sec. II we describe the basic ideas behind the generating function method. The construction procedure for the generators and second order Casimir operator of U(3) in the generating parameter space is considered in Sec. III. In Sec. IV some examples of basis states in Fock–Bargmann space are presented. In Sec. V we get reduced transition probabilities for the scissors mode in the multiplicity free case. In Sec. VI some pictures showing the relative orientation of one axial and one nonaxial rotor are discussed.

II. BASIC IDEAS BEHIND THE GENERATING FUNCTION METHOD

A microscopic analysis of the TRM can be established using the generating function method.[25–27] The analysis is limited to a consideration of only the collective modes of the valence nucleons within the same opened shell. A starting point for the analysis is an explanation of the way in which the dynamics of the relative motion of protons and neutrons can be reduced to a two rotor problem.

For this purpose consider the Slater determinant Φ constructed of proton

$$|i,\pi\rangle=H_{n_x}(\mathbf{u}_1\cdot\mathbf{r}_i)H_{n_y}(\mathbf{u}_2\cdot\mathbf{r}_i)H_{n_z}(\mathbf{u}_3\cdot\mathbf{r}_i)\exp(-r_i^2/2)\xi(\sigma_i) \tag{1}$$

and neutron

$$|i,\nu\rangle = H_{n_x}(\mathbf{v}_1\cdot\mathbf{r}_i)H_{n_y}(\mathbf{v}_2\cdot\mathbf{r}_i)H_{n_z}(\mathbf{v}_3\cdot\mathbf{r}_i)\exp(-r_i^2/2)\xi(\sigma_i) \tag{2}$$

orbitals in a harmonic oscillator potential. In these expressions the $H_n(x_i)$ are Hermite polynomials and the $\xi(\sigma_i)$ spin functions. We assume known rules for the filling of the shells[25] leading to a given SU(3) symmetry (λ_π,μ_π) for the proton subsystem and (λ_ν,μ_ν) for the neutron subsystem apply. The proton orbitals (1) contain a set of three orthonormalized vectors $\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3$ (denoted $\mathbf{u}_s$, where $s=1,2,3$, below); likewise the orbitals for the neutrons (2) contain another set of three vectors $\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3$ (denoted $\mathbf{v}_s$, where $s=1,2,3$, below) with the same properties. The Slater determinant is a generating function and can be written in the following form:

$$\begin{aligned}&\Phi(\{\mathbf{u}_s,\mathbf{v}_s\};\{\mathbf{r}_i,\sigma_i,\tau_i\})\\&=\sum_{L_1,\omega_1,M_1}\sum_{L_2,\omega_2,M_2}\sum_{K_1,K_2} C^{\omega_1\omega_2}(L_1K_1;L_2K_2)D^{L_1}_{K_1M_1}(\alpha_1\beta_1\gamma_1)D^{L_2}_{K_2M_2}(\alpha_2\beta_2\gamma_2)\\&=\sum_{\widetilde{K}_1\widetilde{K}_2}\widetilde{C}^{\omega_1\omega_2}(L_1\widetilde{K}_1;L_2\widetilde{K}_2)\Psi_{L_1\widetilde{K}_1M_1;L_2\widetilde{K}_2M_2}(\{\mathbf{r}_i,\sigma_i,\tau_i\}).\end{aligned} \tag{3}$$

As usual the τ_i is isospin label of i-th particle and the $D^{L_i}_{K_iM_i}(\alpha_i\beta_i\gamma_i)$ are spherical Wigner functions with $\alpha_1,\beta_1,\gamma_1$ and $\alpha_2,\beta_2,\gamma_2$ two sets of Euler angles defining, respectively, the orientation of the two sets of vectors $\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3$ and $\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3$, the letters ω_1,ω_2 are the additional quantum numbers (for instance, the Bargmann–Moshinsky integrals). This expansion (3) is of a great importance because it establishes the relation between many-particle basis functions

$$\sum_{\widetilde{K}_1\widetilde{K}_2}\widetilde{C}^{\omega_1\omega_2}(L_1\widetilde{K}_1;L_2\widetilde{K}_2)\Psi_{L_1\widetilde{K}_1M_1;L_2\widetilde{K}_2M_2}(\{\mathbf{r}_i,\sigma_i,\tau_i\})$$

and the images of this functions

$$\sum_{K_1,K_2} C^{\omega_1\omega_2}(L_1K_1;L_2K_2)D^{L_1}_{K_1M_1}(\alpha_1\beta_1\gamma_1)D^{L_2}_{K_2M_2}(\alpha_2\beta_2\gamma_2)$$

in the space of generating parameters which are the Euler angles α_1, β_1, γ_1, α_2, β_2, γ_2. The symbols $\widetilde{K}_1,\widetilde{K}_2$ are the angular momentum projections onto the intrinsic axis for the functions Ψ, defined in the usual coordinate space, and K_1,K_2 are the same, however for the function, defined in the generating parameter space. Finally, $\widetilde{C}^{\omega_1\omega_2}(L_1\widetilde{K}_1;L_2\widetilde{K}_2)$ and $C^{\omega_1\omega_2}(L_1K_1;L_2K_2)$ are the coefficients of the linear superposition for the wave function and its map.

In essence, the dynamical variables of the system are separated in the process of mapping these basis functions onto their images which are defined in Fock–Bargmann space.[28,27] The generating vector space can be identified with the Fock–Bargmann space if these vectors are considered as complex objects. Of course, the Euler angles will then be complex as well. However, there is no need to make such an extension to the complex space for the solving many problems. The latter applies here so we take all generating parameters to be real, even if this is not expressly stated. Note that the basis functions, which have a complicated form in a coordinate space, are much simpler in the Fock–Bargmann space, being homogeneous polynomials in the generating vectors $\{\mathbf{u}_s,\ \mathbf{v}_s\}$. The generating function Φ is a kernel of the integral transformation

$$\Psi(\{\mathbf{u}_s,\mathbf{v}_s\}) = \int\cdots\int \Phi(\{\mathbf{u}_s,\mathbf{v}_s\};\{\mathbf{r}_i,\sigma_i,\tau_i\})\Psi(\{\mathbf{r}_i,\sigma_i,\tau_i\})d\mathbf{r}_1\cdots d\mathbf{r}_A \tag{4}$$

that maps the model wave functions $\Psi(\{\mathbf{r}_i,\sigma_i,\tau_i\})$ onto their images $\Psi(\{\mathbf{u}_s,\mathbf{v}_s\})$ in the space of states reproducing the dynamics of the two rotors. The images of the basis functions can be interpreted as the wave functions of two (in general, triaxial) rotors. The angular momentum

values L_1, L_2 of these two rotors and their projections K_1, K_2 onto their respective internal symmetry axis are limited in accordance with the rules established by Elliott for basis states of the SU(3) irreps (λ_π, μ_π) and (λ_ν, μ_ν).

A study of the dynamics of two quantum rotors is a relatively simple, albeit physically significant problem that can be solved with the help of heuristic arguments focused on an analysis of the mechanics of such a system. It is very important to note, however, that the practical importance of the image concept is not restricted by going to this simplification. By further exploiting the image concept, a Hamiltonian for realistic nucleon-nucleon interactions can be written down, one that includes terms that may break the underlying unitary symmetries of the problem. This is important because it means the underlying microscopic substructure can be incorporated into the analysis.

The images and their normalized coefficients can actually be found from the overlap integral of the Slater determinant Φ with another determinant,

$$\widetilde{\Phi} = \Phi(\{\widetilde{\mathbf{u}}_s, \widetilde{\mathbf{v}}_s\}; \{\mathbf{r}_i, \sigma_i, \tau_i\}). \tag{5}$$

The six orthonormalized vectors in this last expression, $\{\widetilde{\mathbf{u}}_s, \widetilde{\mathbf{v}}_s\}$ differ in direction from the vectors of the generating function Φ. Then the overlap integral

$$\langle \widetilde{\Phi} | \Phi \rangle = \int \cdots \int \Phi(\{\widetilde{\mathbf{u}}_s, \widetilde{\mathbf{v}}_s\}; \{\mathbf{r}_i, \sigma_i, \tau_i\}). \Phi(\{\mathbf{u}_s, \mathbf{v}_s\}; \{\mathbf{r}_i, \sigma_i, \tau_i\}). d\mathbf{r}_1 \cdots d\mathbf{r}_A$$

can be calculated as a determinant of overlap integrals of single-particle orbitals. Most of its elements vanishes because of the orthogonality of the spin functions. The total generating overlap integral has a multiplicative form

$$\langle \widetilde{\Phi} | \Phi \rangle = \prod_k^{n^{open}} \langle \widetilde{\Phi}_k | \Phi_k \rangle,$$

where k, the oscillator shell number, runs over the n^{open} open shells and $\langle \widetilde{\Phi}_k | \Phi_k \rangle$ is an overlap integral of the k-th shell. It is important to note that the terms representing closed oscillator shells are constants. Thus the structure of the generating invariant $\langle \widetilde{\Phi} | \Phi \rangle$ is determined only by valence configurations of the system. It also means that nucleons in closed shells form an inert core in a very natural way with the dynamics determined only by the valence nucleons. While some approaches employ an artificial truncation of shell-model space, the present approach naturally leads to this type of factorization.

Taking into account the fact that protons and neutrons are filling different harmonic oscillator shells we obtain the very compact result

$$\langle \widetilde{\Phi} | \Phi \rangle = (\mathbf{u}_1 \cdot \widetilde{\mathbf{u}}_1)^{\lambda_\pi} (\mathbf{u}_3 \cdot \widetilde{\mathbf{u}}_3)^{\mu_\pi} (\mathbf{v}_1 \cdot \widetilde{\mathbf{v}}_1)^{\lambda_\nu} (\mathbf{v}_3 \cdot \widetilde{\mathbf{v}}_3)^{\mu_\nu}. \tag{6}$$

Notice that the overlap integral of generating functions $\widetilde{\Phi}$ and Φ with the Hamiltonian of a system supply us with constructions that are necessary for a complete study of the dynamics of the valence nucleons. They are matrix elements of the Hamiltonian in the basis states of the direct product $(\lambda_\pi, \mu_\pi) \times (\lambda_\nu, \mu_\nu)$. Working within the basis of SU(3) irreps and even-even nucleus we can assume that spin S_π of the proton and S_ν of the neutron subsystems are each zero. Therefore in this scheme orbital angular momentum L is an integral of motion and the overlap integral of generating functions can be written as

$$\langle \widetilde{\Phi} | \Phi \rangle = \sum_{\lambda \mu \nu} \sum_{L \omega M} \widetilde{\Psi}^{N \lambda \mu \nu}_{L \omega M} \Psi^{N \lambda \mu \nu}_{L \omega M}. \tag{7}$$

This expression is the realization of the well-known projection method of wave packet onto states with fixed values of angular momentum.

The overlap integral of these functions with the Hamiltonian is as follows:

$$\langle\tilde{\Phi}|\hat{H}|\Phi\rangle=\sum_{\tilde{\lambda}\tilde{\mu}\tilde{\nu}}\sum_{\lambda\mu\nu}\sum_{L\omega\tilde{\omega}M}\tilde{\Psi}^{N\tilde{\lambda}\tilde{\mu}\tilde{\nu}}_{L\tilde{\omega}M}\langle\tilde{\lambda}\tilde{\mu}\tilde{\nu}L\tilde{\omega}|\hat{H}|\lambda\mu\nu L\omega\rangle\Psi^{N\lambda\mu\nu}_{L\omega M}.$$

It is relatively easy to find the overlap integral $\langle\tilde{\Phi}|\hat{H}|\Phi\rangle$ even though the potential energy operator may include a sum of nucleon–nucleon interactions with the Coulomb interaction properly taken into account. However, to select a specific matrix element,

$$\langle\tilde{\lambda}\tilde{\mu}\tilde{\nu}L\tilde{\omega}|\hat{H}|\lambda\mu\nu L\omega\rangle,$$

from this overlap integral it is necessary to apply, for example, $\Psi^{N\lambda\mu\nu}_{L\omega M}$ to the result. The next goal of this study is to find a representation in which we can extract directly from the result required information for calculating isovector $M1$ transitions.

III. GENERATORS AND SECOND-ORDER CASIMIR OPERATOR OF U(3)

It is well-known that the quadrupole-quadrupole interaction is the simplest effective Hamiltonian for describing interactions among valence nucleons and that the corresponding eigenstates have good SU(3) symmetry.[21] In order to construct these states we should perform the reduction of the basis of direct products $(\lambda_\pi,\mu_\pi)\times(\lambda_\nu,\mu_\nu)$ onto the coupled (λ,μ) basis. The standard procedure for decomposing a product of two SU(3) representations into its irreducible parts is derived from the Littlewood rules for the coupling of two-rowed Young diagrams. This procedure is described in detail in Ref. 29. Note that our additional parameters differ from labels introduced in Ref. 29. In our scheme it is convenient to use the Young diagram assuming the following structure:

first row:	$f_1=\lambda_\nu+\lambda_\pi+\mu_{O\nu}+\mu_\pi-m-\nu_1-\nu_2$	boxes,
second row:	$f_2=\mu_\nu+\mu_\pi+m-k$	boxes,
third row:	$f_3=k+\nu_1+\nu_2$	boxes.

This is equivalent to

$$\lambda=\lambda_\nu+\lambda_\pi-2m-\nu_1-\nu_2+k,$$

$$\mu=\mu_\nu+\mu_\pi-2k-\nu_1-\nu_2+m,$$

$$\nu=k+\nu_1+\nu_2,$$

where (λ,μ) are the usual SU(3) irrep labels, and ν is the so-called background number of oscillator quanta. When we use only three parameters, we get

$$(\lambda,\mu)=(\lambda_\nu+\lambda_\pi-2(m-k)-\nu,\mu_\nu+\mu_\pi+(m-k)-\nu).$$

It is easy to see that total number of oscillator quanta N is

$$N=\lambda+2\mu+3\nu=\lambda_\nu+\lambda_\pi+2\mu_\nu+2\mu_\pi.$$

Note that m and k appear as a difference in these expressions and the final SU(3) representations are therefore the same for different combinations of m and $k$as long as their difference remains constant. But there is an additional source for this multiplicity; namely, the labels ν_1 and ν_2. The simplest case when we should use the label ν_1 is the reduction of the direct product $(\lambda_\pi,0)\otimes(0,\mu_\nu)$, when the irreps $(\lambda_\pi-\nu_1,\mu_\nu-\nu_1)$ appears. The direct product $(0,\mu_\pi)\otimes(\lambda_\nu,0)$ is the source for the irreps $(\lambda_\nu-\nu_2,\mu_\pi-\nu_2)$. We should employ both labels in the general case. Even though ν and k may be constant, the parameters ν_1 and ν_2 can vary. As an example, consider the ^{160}Gd nucleus with leading SU(3) irrep (28,8). In this case the following irreps contain the lowest scissors-like $J=1^+$ states:

$$(26,9)\rightarrow m=1,\quad \nu_1=0,\quad \nu_2=0,\quad k=0,$$

$$(27,7)\rightarrow m=0,\quad \nu_1=1,\quad \nu_2=0,\quad k=0,$$

$$(27,7)\rightarrow m=0,\quad \nu_1=0,\quad \nu_2=1,\quad k=0,$$

$$(29,6)\rightarrow m=0,\quad \nu_1=0,\quad \nu_2=0,\quad k=1.$$

The first case corresponds to an excitation that appears for two axial rotors with principal axis $\mathbf{u}_1$ and $\mathbf{v}_1$. States with (27,7) can be excited in two ways: $\nu_1=1,\nu_2=0$ or $\nu_1=0,\nu_2=1$. The fourth irrep is also a new one. It corresponds to the case of two axial rotors with symmetry axis directed along $\mathbf{u}_3$ and $\mathbf{v}_3$, respectively. (This reduction can also be achieved using Clebsch–Gordan series—here we choose to construct the basis functions with the quantum numbers $(\lambda,\mu);L,M,\nu$ in explicit form as functions of six variables describing the motion of two rotors.)

First we construct the generators of U(3) in the space of the three vectors $\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3$. This is easy to do because these vectors are mutually orthogonal. It is enough to take vector $\mathbf{u}_1$ and add another to it, $\mathbf{w}^{\pi}$, which is related to $\mathbf{u}_3$ as follows:

$$\mathbf{u}_3=\mathbf{u}_1\times\mathbf{w}^{\pi}. \tag{8}$$

Then if we introduce Cartesian components of the vectors $\mathbf{u}_1$ and $\mathbf{w}$,

$$\mathbf{u}_1=\{x_1^{\pi},x_2^{\pi},x_3^{\pi}\},\quad \mathbf{w}^{\boldsymbol{\pi}}=\{y_1^{\pi},y_2^{\pi},y_3^{\pi}\}, \tag{9}$$

and define the generators $a_{i,j}$ of U(3) for protons in the following way:

$$a_{i,j}(\pi)=x_i^{\pi}\frac{\partial}{\partial x_j^{\pi}}+y_i^{\pi}\frac{\partial}{\partial y_j^{\pi}}. \tag{10}$$

In an analogous way we define the generators $a_{i,j}(\nu)$ of U(3) for the neutron subsystem. And finally we have the generators $A_{i,j}$ of U(3) for the neutron-proton system as a whole:

$$A_{i,j}=a_{i,j}(\pi)+a_{i,j}(\nu). \tag{11}$$

It is convenient to write generators $a_{i,j}(\pi)$, $a_{i,j}(\nu)$ and $A_{i,j}$ in the intrinsic coordinate frame defined by the directions of the vectors $\mathbf{u}_1,\mathbf{u}_3,\mathbf{v}_1,\mathbf{v}_3$. In Ref. 26 it is shown how this can be done using only the two vectors $\mathbf{u}_1$ and $\mathbf{v}_1$. The eigensystem $\{\xi,\eta,\zeta\}$ is defined as one where the two vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ lie in the (ξ,η) plane and have the following components:

$$\mathbf{u}_1=\left\{u_1\cos\frac{\theta}{2},-u_1\sin\frac{\theta}{2},0\right\},\quad \mathbf{v}_1=\left\{v_1\cos\frac{\theta}{2},v_1\sin\frac{\theta}{2},0\right\}. \tag{12}$$

Obviously the ξ axis coincides with the bisectrix of the angle θ between $\mathbf{u}_1$ and $\mathbf{v}_1$. As for the vectors $\mathbf{u}_3$ and $\mathbf{v}_3$, their components in this system are parametrized as follows:

$$\begin{aligned}\mathbf{u_3}&=\left\{u_3\sin\frac{\theta}{2}\cos\phi_{\nu},u_3\cos\frac{\theta}{2}\cos\phi_{\nu},u_3\sin\phi_{\nu}\right\},\\ \mathbf{v_3}&=\left\{-v_3\sin\frac{\theta}{2}\cos\phi_{\pi},v_3\cos\frac{\theta}{2}\cos\phi_{\pi},v_3\sin\phi_{\pi}\right\}.\end{aligned} \tag{13}$$

As a result there are two additional angles ϕ_{ν}, ϕ_{π}. The orientation of the intrinsic frame is specified in the usual way by three Euler angles α, β, γ. Thus the rotation of the system as a whole is separated into external and internal degrees of freedom, the former being the Euler angles α, β, γ and the latter being θ, ϕ_{ν}, ϕ_{π}. We use these six dynamical variables, α, β, γ and θ, ϕ_{ν}, ϕ_{π}, as generating coordinates instead of the Euler angles α_1, β_1, γ_1, α_2, β_2, γ_2. In order to work

with generators of U(3) we ignore the fact that $\mathbf{u}_1,\mathbf{u}_3,\mathbf{v}_1,\mathbf{v}_3$ are unit vectors, that is, that u_1,u_3,v_1,v_3 are actually fixed and equal unity. Thus we have the following expressions for the U(3) generators for the proton subsystem:

$$a_{11}(1)=\cos^2\frac{\theta}{2}\,u_1\frac{\partial}{\partial u_1}-\cos\frac{\theta}{2}\sin\frac{\theta}{2}\left[\frac{\partial}{\partial\theta}+\frac{1}{2}M_\zeta\right]+\left(1-\cos^2\phi_\pi\sin^2\frac{\theta}{2}\right)u_3\frac{\partial}{\partial u_3}$$
$$+\sin^2\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{12}(1)=-\cos\frac{\theta}{2}\sin\frac{\theta}{2}\,u_1\frac{\partial}{\partial u_1}-\cos^2\frac{\theta}{2}\left[\frac{\partial}{\partial\theta}+\frac{1}{2}M_\zeta\right]-\cos\frac{\theta}{2}\sin\frac{\theta}{2}\cos^2\phi_\pi u_3\frac{\partial}{\partial u_3}$$
$$+\cos\frac{\theta}{2}\sin\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{13}(1)=\frac{1}{2}\cot\frac{\theta}{2}M_\xi+\frac{1}{2}M_\eta-\sin\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi u_3\frac{\partial}{\partial u_3}+\left[\sin\frac{\theta}{2}\sin^2\phi_\pi+\cos\frac{\theta}{2}\cot\theta\right]\frac{\partial}{\partial\phi_\pi}$$
$$+\frac{1}{2}\frac{1}{\sin\theta/2}\frac{\partial}{\partial\phi_\nu};$$

$$a_{21}(1)=-\cos\frac{\theta}{2}\sin\frac{\theta}{2}u_1\frac{\partial}{\partial u_1}+\sin^2\frac{\theta}{2}\left[\frac{\partial}{\partial\theta}+\frac{1}{2}M_\zeta\right]-\cos\frac{\theta}{2}\sin\frac{\theta}{2}\cos^2\phi_\pi u_3\frac{\partial}{\partial u_3}$$
$$+\cos\frac{\theta}{2}\sin\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{22}(1)=\sin^2\frac{\theta}{2}\,u_1\frac{\partial}{\partial u_1}+\cos\frac{\theta}{2}\sin\frac{\theta}{2}\left[\frac{\partial}{\partial\theta}+\frac{1}{2}M_\zeta\right]+\left(1-\cos^2\phi_\pi\cos^2\frac{\theta}{2}\right)u_3\frac{\partial}{\partial u_3}$$
$$+\cos^2\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{23}(1)=-\frac{1}{2}\tan\frac{\theta}{2}M_\eta-\frac{1}{2}M_\xi-\cos\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi u_3\frac{\partial}{\partial u_3}+\left[\cos\frac{\theta}{2}\sin^2\phi_\pi-\sin\frac{\theta}{2}\cot\theta\right]\frac{\partial}{\partial\phi_\pi}$$
$$-\frac{1}{2}\frac{1}{\cos\theta/2}\frac{\partial}{\partial\phi_\nu};$$

$$a_{31}(1)=-\sin\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi u_3\frac{\partial}{\partial u_3}-\sin\frac{\theta}{2}\cos^2\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{32}(1)=-\cos\frac{\theta}{2}\sin\phi_\pi\cos\phi_\pi u_3\frac{\partial}{\partial u_3}-\cos\frac{\theta}{2}\cos^2\phi_\pi\frac{\partial}{\partial\phi_\pi};$$

$$a_{33}(1)=\cos^2\phi_\pi u_3\frac{\partial}{\partial u_3}-\sin\phi_\pi\cos\phi_\pi\frac{\partial}{\partial\phi_\pi}.$$

(14)

The set of the generators for the neutron subsystem can be obtained from the above written ones (14) in a result of the simple substitution: $u_1\to v_1$, $u_3\to v_3$, $\theta\to-\theta$, $\phi_\pi\to\phi_\nu$, $\phi_\nu\to\phi_\pi$. The generators A_{ij} for the neutron–proton system as a whole can be put down in compact form:

$$A_{11}=\frac{1}{2}(1+t)(\lambda_\pi+\lambda_\nu)+(1-t^2)\frac{\partial}{\partial t}+\mu_\pi+\mu_\nu-\frac{1-t}{2}[\chi^+_{\tau_\pi}(\mu_\pi)+\chi^+_{\tau_\nu}(\mu_\nu)],$$

$$A_{12}=-\frac{1}{2}\{\sqrt{1-t^2}(\lambda_\pi-\lambda_\nu)+(1+t)M_\zeta+\sqrt{1-t^2}[\chi^+_{\tau_\pi}(\mu_\pi)-\chi^+_{\tau_\nu}(\mu_\nu)]\},$$

$$A_{13}=M_\eta-\sqrt{\frac{1-t}{2}}\,[\chi^-_{\tau_\pi}(\mu_\pi)-\chi^-_{\tau_\nu}(\mu_\nu)]\},$$

$$A_{21}=-\frac{1}{2}\{\sqrt{1-t^2}(\lambda_\pi-\lambda_\nu)-(1-t)M_\zeta+\sqrt{1-t^2}[\chi^+_{\tau_\pi}(\mu_\pi)-\chi^+_{\tau_\nu}(\mu_\nu)]\},$$

$$A_{22}=\frac{1}{2}(1-t)(\lambda_\pi+\lambda_\nu)-(1-t^2)\frac{\partial}{\partial t}+\mu_\pi+\mu_\nu-\frac{1+t}{2}[\chi^+_{\tau_\pi}(\mu_\pi)+\chi^+_{\tau_\nu}(\mu_\nu)], \tag{15}$$

$$A_{23}=-M_\xi-\sqrt{\frac{1+t}{2}}\,[\chi^-_{\tau_\pi}(\mu_\pi)+\chi^-_{\tau_\nu}(\mu_\nu)],$$

$$A_{31}=-\sqrt{\frac{1-t}{2}}\,[\chi^-_{\tau_\pi}(\mu_\pi)-\chi^-_{\tau_\nu}(\mu_\nu)],$$

$$A_{32}=-\sqrt{\frac{1+t}{2}}\,[\chi^-_{\tau_\pi}(\mu_\pi)+\chi^-_{\tau_\nu}(\mu_\nu)],$$

$$A_{33}=\chi^+_{\tau_\pi}(\mu_\pi)+\chi^+_{\tau_\nu}(\mu_\nu),$$

where $t=\cos\theta$, $\tau_\nu=\sin\phi_\nu$, $\tau_\pi=\sin\phi_\pi$, and

$$\chi^+_\tau(\mu)=(1-\tau^2)\left[\mu-\tau\frac{\partial}{\partial\tau}\right],\quad \chi^-_\tau(\mu)=\sqrt{1-\tau^2}\left[\mu\tau+(1-\tau^2)\frac{\partial}{\partial\tau}\right]. \tag{16}$$

The operators M_ξ, M_η, M_ζ are angular momentum projections.

The next step is the construction of the second order Casimir invariant C_2 of U(3) in terms of the U(3) group generators A_{ij}. This operator is a contraction of SU(3) generators:

$$C_2=d_{il}d_{km}A_{lm}d_{kj}d_{ir}A_{jr}, \tag{17}$$

where the d_{ij} are the elements of the transition matrix from the laboratory frame of reference into the intrinsic one. This expression is a second order differential form in terms of the internal variables and angular momentum projections M_ξ, M_η, M_ζ:

$$\begin{aligned}\hat{C}_2=2\Bigg\{&\chi^0_t(\lambda_\nu,\lambda_\pi)+\chi^0_{\tau_\nu}(\mu_\nu,\mu_\pi)+\chi^0_{\tau_\pi}(\mu_\nu,\mu_\pi)+\chi^+_t(\lambda_\nu)\chi^+_{\tau_\pi}(\mu_\pi)+\chi^+_t(\lambda_\pi)\chi^+_{\tau_\nu}(\mu_\nu)\\&+(1+t^2)\chi^+_{\tau_\nu}(\mu_\nu)\chi^+_{\tau_\pi}(\mu_\pi)+2t\chi^-_{\tau_\nu}(\mu_\nu)\chi^-_{\tau_\pi}(\mu_\pi)+\sqrt{\frac{1+t}{2}}\,[\chi^-_{\tau_\nu}(\mu_\nu)+\chi^-_{\tau_\pi}(\mu_\pi)]M_\xi\\&+\sqrt{\frac{1-t}{2}}\,[\chi^-_{\tau_\nu}(\mu_\nu)-\chi^-_{\tau_\pi}(\mu_\pi)]M_\eta+\frac{1}{2}\{\chi^-_{\tau_\nu}(\mu_\nu)-\chi^-_{\tau_\pi}(\mu_\pi)+\lambda_\nu-\lambda_\pi\}t\sqrt{1-t^2}M_\zeta\\&-\frac{1}{4}(1-t^2)M^2_\zeta\Bigg\}+(\lambda_\nu+\lambda_\pi+2)(\lambda_\nu+\lambda_\pi)+2(\mu_\nu+\mu_\pi)(\lambda_\nu+\lambda_\pi+\mu_\nu+\mu_\pi+1),\end{aligned} \tag{18}$$

where the operators $\chi^+_t(\lambda),\chi^-_t(\lambda)$ are defined by the relations (16) and

$$\chi^0_t(\lambda_\nu,\lambda_\pi)=(1-t^2)\left[(1-t^2)\frac{\partial^2}{\partial t^2}+(\lambda_\nu+\lambda_\pi-1)t\frac{\partial}{\partial t}-\lambda_\nu\lambda_\pi\right]. \tag{19}$$

The eigenfunctions of this differential form,

$$\Psi^{N\lambda\mu\nu}_{L\omega M}(\theta,\phi_\pi,\phi_\nu;\alpha,\beta,\gamma),$$

satisfy the following relation:

$$C_2\Psi^{N\lambda\mu\nu}_{L\omega M}=C^{\lambda\mu}_{\nu}\Psi^{N\lambda\mu\nu}_{L\omega M}, \tag{20}$$

where N is number of quanta, ν and ω are additional quantum numbers, and $C^{\lambda\mu}_{\nu}$ are eigenvalues of Casimir operator C_2:

$$C^{\lambda\mu}_{\nu}=\lambda^2+2\mu^2+2\lambda\mu+2(\lambda+\mu)+2\nu(\lambda+2\mu)+3\nu^2. \tag{21}$$

(Details involved in solving Eq. (20) are left for a more mathematical discussion of this matter.) In the next section we present some examples of basis functions for such a system.

IV. BASIS STATES IN FOCK–BARGMANN SPACE

In the present study we focus only on configurations that do not require the additional Bargmann–Moshinsky quantum number ω. For simplicity we denote

$$\Psi^{N\lambda\mu\nu}_{L\omega M}(\mathbf{u}_1,\mathbf{u}_3,\mathbf{v}_1,\mathbf{v}_3)\equiv\Psi^{N\lambda\mu\nu}_{L}.$$

As mentioned above, the eigenfunctions $\Psi^{N\lambda\mu\nu}_{L}$ are homogeneous polynomials in $\mathbf{u}_1,\mathbf{u}_3,\mathbf{v}_1,\mathbf{v}_3$ in Fock–Bargmann space. For example, the simplest solutions of Eq. (20) can be written in terms of the generating vectors:

$$\Psi^{2,2,0,0}_{0}=(\mathbf{v}_1\cdot\mathbf{u}_1),\quad \Psi^{3,0,0,1}_{0}=(\mathbf{v}_1\cdot\mathbf{u}_3),$$

$$\Psi^{3,0,0,1}_{0}=(\mathbf{v}_3\cdot\mathbf{u}_1),\quad \Psi^{2,0,2,0}_{0}=(\mathbf{v}_3\cdot\mathbf{u}_3),$$

$$\Psi^{4,2,2,0}_{0}=\sqrt{\frac{2}{3}}\left\{(\mathbf{v}_3\cdot\mathbf{u}_3)(\mathbf{v}_1\cdot\mathbf{u}_1)+\frac{1}{4}(\mathbf{v}_1\cdot\mathbf{u}_3)(\mathbf{v}_3\cdot\mathbf{u}_1)\right\},$$

$$\Psi^{6,1,1,1}_{1}=\sqrt{\frac{2}{5}}\,(\mathbf{v}_1\cdot\mathbf{u}_3)[\mathbf{v}_1\times\mathbf{u}_3].$$

To explain the normalization of the wave functions $\Psi^{N\lambda\mu\nu}_{L}$, we can go back to the overlap integral (3). Having defined the operator $\hat{\mathscr{A}}$

$$\hat{\mathscr{A}}=C_2-C^{\lambda_0\mu_0}_{0},$$

where $C^{\lambda_0\mu_0}_{0}$ is eigenvalue of Casimir operator for the leading SU(3) irrep

$$(\lambda_0,\mu_0)=(\lambda_\nu+\lambda_\pi,\mu_\nu+\mu_\pi),$$

and denoting

$$J(\lambda_\pi,\lambda_\nu,\mu_\pi,\mu_\nu)=(\mathbf{u}_1\cdot\tilde{\mathbf{u}}_1)^{\lambda_\pi}(\mathbf{u}_3\cdot\tilde{\mathbf{u}}_3)^{\mu_\pi}(\mathbf{v}_1\cdot\tilde{\mathbf{v}}_1)^{\lambda_\nu}(\mathbf{v}_3\cdot\tilde{\mathbf{v}}_3)^{\mu_\nu},$$

we observe that

$$\begin{aligned}\hat{\mathscr{A}}J(\lambda_\pi,\lambda_\nu,\mu_\pi,\mu_\nu)=&-\lambda_\pi\lambda_\nu[\mathbf{u}_1\times\mathbf{v}_1]\cdot[\tilde{\mathbf{u}}_1\times\tilde{\mathbf{v}}_1]J(\lambda_\pi-1,\lambda_\nu-1,\mu_\pi,\mu_\nu)\\&-\mu_\pi\mu_\nu[\mathbf{u}_3\times\mathbf{v}_3]\cdot[\tilde{\mathbf{u}}_3\times\tilde{\mathbf{v}}_3]J(\lambda_\pi,\lambda_\nu,\mu_\pi-1,\mu_\nu-1)\\&-\lambda_\nu\mu_\pi(\mathbf{u}_1\cdot\mathbf{v}_3)(\tilde{\mathbf{u}}_1\cdot\tilde{\mathbf{v}}_3)J(\lambda_\pi,\lambda_\nu-1,\mu_\pi-1,\mu_\nu)\\&-\lambda_\pi\mu_\nu(\mathbf{v}_1\cdot\mathbf{u}_3)(\tilde{\mathbf{v}}_1\cdot\tilde{\mathbf{u}}_3)J(\lambda_\pi-1,\lambda_\nu,\mu_\pi,\mu_\nu-1).\end{aligned} \tag{22}$$

On the other hand

$$J(\lambda_\pi,\lambda_\nu,\mu_\pi,\mu_\nu)=\sum_{\lambda,\mu} B^2_{mk\nu_1\nu_2}(\lambda,\mu)\langle(\lambda,\mu)|\widetilde{(\lambda,\mu)}\rangle, \tag{23}$$

where $B_{mk\nu_1\nu_2}(\lambda,\mu)$ are the weights of corresponding irreps. Therefore acting with the operator $\hat{\mathscr{A}}$ on expansion (23) and taking into account the relation (20) we have

$$\hat{\mathscr{A}}J(\lambda_\pi,\lambda_\nu,\mu_\pi,\mu_\nu)=\sum_{\lambda,\mu} B^2_{mk\nu_1\nu_2}(\lambda,\mu)(C^{\lambda,\mu}_\nu-C^{\lambda_0,\mu_0}_0)\langle(\lambda,\mu)|\widetilde{(\lambda,\mu)}\rangle, \tag{24}$$

where the overlap integral $\langle(\lambda,\mu)|\lambda,\mu)\rangle$ can be written as a product

$$|\mathbf{v}_1\times\mathbf{u}_1|^m|\mathbf{v}_3\times\mathbf{u}_3|^k(\mathbf{v}_1\cdot\mathbf{u}_3)^{\nu_1}(\mathbf{v}_3\cdot\mathbf{u}_1)^{\nu_2}\times|\tilde{\mathbf{v}}_1\times\tilde{\mathbf{u}}_1|^m|\tilde{\mathbf{v}}_3\times\tilde{\mathbf{u}}_3|^k(\tilde{\mathbf{v}}_1\cdot\tilde{\mathbf{u}}_3)^{\nu_1}(\tilde{\mathbf{v}}_3\cdot\tilde{\mathbf{u}}_1)^{\nu_2}I(\lambda,\mu). \tag{25}$$

In this expression $I(\lambda,\mu)$ is normalized so that when vectors $\mathbf{u}_1$ and $\mathbf{v}_1$, $\mathbf{u}_3$ and $\mathbf{v}_3$ are collinear, it takes on the well-known form first introduced by Elliott,[21]

$$I(\lambda,\mu)=d^\lambda_{11}d^\mu_{33}, \tag{26}$$

where $d_{ij}\equiv d_{ij}(\alpha',\beta',\gamma')$ are matrix elements of the transition matrix from the $\{\xi,\eta,\zeta\}$ coordinate frame into the $\{\tilde{\xi},\tilde{\eta},\tilde{\zeta}\}$ one.

Taking this property of the overlap integral into account and comparing relations (22) and (24), one can extract the coefficients $B^2_{mk\nu_1\nu_2}(\lambda,\mu)$. As an example, we now apply this procedure to find $B^2_{m=1}(\lambda,\mu)$. Using Eqs. (22), (18) and (25) we get

$$\begin{aligned}\lambda_\pi\lambda_\nu[\mathbf{u}_1\times\mathbf{v}_1]\cdot[\tilde{\mathbf{u}}_1\times\tilde{\mathbf{v}}_1]J(\lambda_\pi-1,\lambda_\nu-1,\mu_\pi,\mu_\nu)+\ldots\\ =\ldots+B^2_{m=1}(\lambda_0-2,\mu_0+1)(C_0^{\lambda_0-2,\mu_0+1}-C_0^{\lambda_0,\mu_0})\\ \times|\mathbf{u}_1\times\mathbf{v}_1||\tilde{\mathbf{u}}_1\times\tilde{\mathbf{u}}_1|I(\lambda_0-2,\mu_0+1)+\ldots,\end{aligned} \tag{27}$$

where dots denote the terms which do not contain the factor

$$|\mathbf{u}_1\times\mathbf{v}_1||\tilde{\mathbf{u}}_1\times\tilde{\mathbf{u}}_1|.$$

Putting in the functions $J(\lambda_\pi-1,\lambda_\nu-1,\mu_\pi,\mu_\nu)$ and $I(\lambda_0-2,\mu_0+1)$ with $(\mathbf{u}_1\cdot\mathbf{v}_1)=(\mathbf{u}_3\cdot\mathbf{v}_3)=1$ we find

$$B^2_{m=1}(\lambda_0-2,\mu_0+1)=\frac{\lambda_\nu\lambda_\pi}{\lambda_\nu+\lambda_\pi}.$$

Other weights can be found in the same way:

$$B^2_{k=1}(\lambda_0+1,\mu_0-2)=\frac{\mu_\nu\mu_\pi}{\mu_\nu+\mu_\pi},$$

$$B^2_{\nu_1=1}(\lambda_0-1,\mu_0-1)=\frac{\mu_\nu\lambda_\pi}{\mu_0+\lambda_0+1},\quad B^2_{\nu_2=1}(\lambda_0-1,\mu_0-1)=\frac{\mu_\pi\lambda_\nu}{\mu_0+\lambda_0+1}.$$

In the second step we project the overlap integral $\langle(\lambda,\mu)|\widetilde{(\lambda,\mu)}\rangle$ onto the states with certain values of orbital angular momentum L:

$$\langle(\lambda,\mu)|\widetilde{(\lambda,\mu)}\rangle=\sum_{L,M} A^2_{mk\nu_1\nu_1}(\lambda,\mu;L)\langle(\lambda,\mu);L,M|\widetilde{(\lambda,\mu)};L,M\rangle.$$

Passing from this expression to the Elliott case $((\mathbf{u}_1\cdot\mathbf{v}_1)=(\mathbf{u}_3\cdot\mathbf{v}_3)=1)$ we find the expansion coefficients $A_{mk\nu_1\nu_1}(\lambda,\mu;L)$:

$$A^2(\lambda_0,\mu_0;L=0)=\frac{1}{2\sqrt{\pi}}\frac{\Gamma((\lambda_0+1)/2)\Gamma((\lambda_0+\mu_0)/2+1)\Gamma((\mu_0+1)/2)}{\Gamma((\lambda_0)/2+1)\Gamma((\lambda_0+\mu_0+3)/2)\Gamma(\mu_0/2+1)};$$

$$A^2_{m=1}(\lambda_0-2,\mu_0+1;L=1)=\frac{3\lambda_0(\mu_0+1)}{(\lambda_0-1)(\mu_0+2)}A^2(\lambda_0,\mu_0;L=0);$$

$$A^2_{k=1}(\lambda_0+1,\mu_0-2;L=1)=\frac{3(\lambda_0+1)\mu_0}{(\lambda_0+2)(\mu_0-1)}A^2(\lambda_0,\mu_0;L=0); \tag{28}$$

$$A^2_{\nu_1=1,(\nu_2=1)}(\lambda_0-1,\mu_0-1;L=1)=\frac{3(\lambda_0+\mu_0+1)}{(\lambda_0+\mu_0)}A^2(\lambda_0,\mu_0;L=0).$$

Finally, we can write down expressions for the normalizing coefficients

$$N_{mk\nu_1\nu_1}(\lambda,\mu;L)=A_{mk\nu_1\nu_1}(\lambda,\mu;L)B_{mk\nu_1\nu_2}(\lambda,\mu).$$

Below we present one example of normalized wave function written in terms of dynamical variables θ, ϕ_ν, ϕ_π for $L=0$, $\mu_\nu=\mu_\pi=1$ and arbitrary even λ_ν,λ_π:

$$\begin{aligned}\Psi^{\lambda_0+4,(\lambda_0,2),0}_{L=0}=&\sqrt{\frac{\lambda_0+2}{2(\lambda_0+1)(\lambda_0+3)}}\Bigg\{\tau F\left(-\frac{\lambda_\nu}{2},-\frac{\lambda_\pi}{2},\frac{1}{2}-\frac{\lambda_0}{2};1-t^2\right)\\&+\frac{\lambda_\nu(\lambda_\nu-1)}{2(\lambda_0+2)(\lambda_0-1)}t(1-t^2)F\left(-\frac{\lambda_\nu}{2}+1,-\frac{\lambda_\pi}{2}+1,\frac{3}{2}-\frac{\lambda_0}{2};1-t^2\right)\\&\times\sqrt{1-\tau_\nu^2}\sqrt{1-\tau_\pi^2}\Bigg\}.\end{aligned} \tag{29}$$

In this expression the $F(a,b,c;z)$ are hypergeometric functions, $t=\cos\theta$, $\tau_\nu=\sin\phi_\nu$, $\tau_\pi=\sin\phi_\pi$ and $\tau=(\mathbf{u}_3\cdot\mathbf{v}_3)=\tau_\nu\tau_\pi+t\sqrt{1-\tau_\pi^2}\sqrt{1-\tau_\nu^2}$. The expression for the wave function for the most general case of arbitrary even μ_ν,μ_π, λ_ν,λ_π and $L=0$ is too spacious to be written wholly. Therefore we present here only first four terms of this function in order to show its structure:

$$\begin{aligned}\Psi^{N,(\lambda_0,\mu_0),0}_{L=0}=&F\left(-\frac{\mu_\pi}{2},-\frac{\mu_\nu}{2},\frac{1-\mu_0}{2};1-\tau^2\right)F\left(-\frac{\lambda_\pi}{2},-\frac{\lambda_\nu}{2},\frac{1-\lambda_0}{2};1-t^2\right)\\&+\frac{2\lambda_\pi\lambda_\nu\mu_\pi\mu_\nu}{(\lambda_0+\mu_0)(\lambda_0-1)(\mu_0-1)}(1-t^2)\sqrt{1-\tau_\pi^2}\sqrt{1-\tau_\nu^2}t\tau\\&\times F\left(-\frac{\lambda_\pi}{2}+1,-\frac{\lambda_\nu}{2}+1,\frac{3-\lambda_0}{2};1-t^2\right)F\left(-\frac{\mu_\pi}{2}+1,-\frac{\mu_\nu}{2}+1,\frac{3-\mu_0}{2};1-\tau^2\right)\\&-\frac{\lambda_\nu\mu_\pi(\lambda_\nu-1)(\mu_\pi-1)}{2(\lambda_0+\mu_0)(\lambda_0-1)(\mu_0-1)}(1-t^2)(1-\tau_\pi^2)\\&\times F\left(-\frac{\lambda_\pi}{2},-\frac{\lambda_\nu}{2}+1,\frac{3-\lambda_0}{2};1-t^2\right)F\left(-\frac{\mu_\pi}{2}+1,-\frac{\mu_\nu}{2},\frac{3-\mu_0}{2};1-\tau^2\right)\\&\times-\frac{\lambda_\pi\mu_\nu(\lambda_\pi-1)(\mu_\nu-1)}{2(\lambda_0+\mu_0)(\lambda_0-1)(\mu_0-1)}(1-t^2)(1-\tau_\nu^2)\end{aligned}$$

$$F\left(-\frac{\lambda_\pi}{2}+1,-\frac{\lambda_\nu}{2},\frac{3-\lambda_0}{2};1-t^2\right)F\left(-\frac{\mu_\pi}{2},-\frac{\mu_\nu}{2}+1,\frac{3-\mu_0}{2};1-\tau^2\right)+\ldots. \tag{30}$$

V. THE TRANSITION PROBABILITIES FOR THE SCISSORS MODE

To find transition probabilities, one must first construct the isovector operator $\mathbf{S}_M$ of the magnetic dipole transition. In the rotating frame this operator is connected with the angular momentum operators of the proton and neutron subsystems through the simple relation

$$\mathbf{S}_k=\mathbf{l}_k(\pi)-\mathbf{l}_k(\nu).$$

It is easy to obtain expressions for the spherical components of the operator $\mathbf{S}_k$ because the relation between operators $\mathbf{l}_k(\pi)$ and $\mathbf{l}_k(\nu)$ and the generators $a_{ij}(\pi)$ and $a_{ij}(\nu)$ are known:

$$\mathbf{l}_k(\pi)=e_{kij}a_{ij}(\pi)$$

and

$$\mathbf{l}_k(\nu)=e_{kij}a_{ij}(\nu).$$

Without giving details, we present the expression for the operator in the laboratory system that is necessary for the subsequent analysis. In going from the intrinsic to the laboratory frame of reference, the operator $\mathbf{S}_k$ transforms as components of a rank 1 tensor, that is,

$$\mathbf{S}'_M=\sum_k D^1_{M,k}\mathbf{S}_k.$$

Again, with the given details we now present an expression for the matrix elements of this operator. We confine ourselves to the analysis of the transitions (involving a change in the isospin) from the state with $L=0$ of the ground band $(\lambda_0,\mu_0)=(\lambda_\nu+\lambda_\pi,\mu_\nu+\mu_\pi)$ to the states with $L=1$ of the band with quantum numbers (λ_0-2,μ_0+1) and the band with (λ_0+1,μ_0-2), that is, we consider only multiplicity free cases for two triaxial rotors.

To find the matrix elements of a transition operator, it is necessary to extract functions $\Psi^{N,(\lambda_0-2,\mu_0+1),0}_{L=1}$ and $\Psi^{N,(\lambda_0+1,\mu_0-2),0}_{L=1}$ from a superposition of functions obtained as the result of the action of this operator on the normalized wave function $\Psi^{N\lambda_0\mu_0 0}_{L=0}$. We obtain

$$\mathbf{S}'_M\Psi^{N,(\lambda_0,\mu_0),0}_{L=0}=-i\frac{2\lambda_\nu\lambda_\pi}{(\lambda_\nu+\lambda_\pi-1)}\frac{N_0(\lambda_0,\mu_0;L=0)}{N_{m=1}(\lambda_0-2,\mu_0+1;L=1)}\Psi^{N,(\lambda_0-2,\mu_0+1),0}_{L=1}$$
$$-i\frac{2\mu_\nu\mu_\pi}{(\mu_\nu+\mu_\pi-1)}\frac{N_0(\lambda_0,\mu_0;L=0)}{N_{k=1}(\lambda_0+1,\mu_0-2;L=1)}\Psi^{N,(\lambda_0+1,\mu_0-2),0}_{L=1}+\ldots, \tag{31}$$

where $N_{mk\nu_1\nu_2}(\lambda,\mu;L)$ are normalizing coefficients and dots denote other functions. Taking into account the expressions given above we find

$$\langle(\lambda_0-2,\mu_0+1),L=1,M|\mathbf{S}'_M|(\lambda_0,\mu_0),L=0\rangle=-2i\sqrt{\frac{(\mu_\pi+\mu_\nu+2)\lambda_\nu\lambda_\pi}{3(\mu_\pi+\mu_\nu+1)(\lambda_\nu+\lambda_\pi-1)}}; \tag{32}$$

$$\langle(\lambda_0+1,\mu_0-2),L=1,M|\mathbf{S}'_M|(\lambda_0,\mu_0),L=0\rangle=-2i\sqrt{\frac{(\lambda_\nu+\lambda_\pi+2)\mu_\nu\mu_\pi}{3(\lambda_\nu+\lambda_\pi+1)(\mu_\nu+\mu-1)}}. \tag{33}$$

Using the well-known relation

$$\langle\alpha_2J_2M_2|Q_{\lambda,\mu}|\alpha_1J_1M_1\rangle=\frac{(J_1\lambda M_1\mu|J_2M_2)}{\sqrt{2J_2+1}}\langle\alpha_2J_2||Q_{\lambda,\mu}||\alpha_1J_1\rangle, \tag{34}$$

TABLE I. Theoretical and experimental M1 transition strengths from the ground state to the first excited 1^+ state. The results given in column Th^1 correspond to the approach used in this work and coincide with those of a pseudo-SU(3) theory. The SU(3) irrep (λ_0,μ_0) for the ground band and the irreps associated to the 1^+ states, (λ',μ'), for each is also given.

			$B(M1;0^+\to1^+)/(\mu_n^2)$	
Nucleus	(λ_0,μ_0)	(λ',μ')	Th^1	Exp
^{154}Sm	(30,0)	(28,1)	3.56	0.8(2)
^{156}Gd	(28,4)	(26,5)	1.91	1.5(3)
^{158}Gd	(28,8)	(26,9)	1.77	1.4(3)
		(29,6)	0.56	0.9(2)
^{160}Gd	(28,8)	(26,9)	1.77	0.8(2)
		(29,6)	0.56	0.60(17)
^{168}Er	(30,8)	(28,9)	1.83	
		(31,6)	0.56	0.9(2)
^{174}Yb	(32,6)	(30,7)	2.11	0.8(2)
^{232}Th	(42,6)	(40,7)	2.40	
		(43,4)	0.39	
$^{234-236}$U	(48,4)	(46,5)	3.29	
^{238}U	(54,0)	(52,1)	5.84	
^{242}Pu	(52,6)	(50,7)	3.27	

we also find the values of the reduced matrix elements of the transition operator. To calculate the reduced probabilities we use the well-known formula:

$$B(M1;1\to2)=\frac{3}{16\pi}\frac{1}{2J_1+1}|\langle\alpha_2J_2||\mathbf{S}'_M||\alpha_1J_1\rangle|^2\mu_n^2.$$

The reduced probabilities calculated by this formula have the values

$$B(M1;0^+\to1^+)=\frac{3}{4\pi}\left[\frac{\lambda_\nu\lambda_\pi}{(\lambda_\nu+\lambda_\pi-1)}\frac{(\mu_\pi+\mu_\nu+2)}{(\mu_\pi+\mu_\nu+1)}\right]\mu_n^2 \tag{35}$$

for the state $|(\lambda_\nu+\lambda_\pi-2,\mu_\nu+\mu_\pi+1),L=1,M\rangle$ and

$$B(M1;0^+\to1^+)=\frac{3}{4\pi}\left[\frac{\mu_\nu\mu_\pi}{(\mu_\nu+\mu_\pi-1)}\frac{(\lambda_\nu+\lambda_\pi+2)}{(\lambda_\nu+\lambda_\pi+1)}\right]\mu_n^2 \tag{36}$$

for the state $|(\lambda_\nu+\lambda_\pi+1,\mu_\nu+\mu_\pi-2),L=1,M\rangle$. The numerical values of $B(M1)$ strengths for the transitions $0^+\to1^+$ are the same that arise in a pseudo-SU(3) theory of collective M1 bands.[10] Note that proton and neutron g-factors are $g_p=1$ and $g_n=0$, respectively. The calculated strengths are given in Table I for some nuclei of the rare earth region and actinide species.

VI. RELATIVE ORIENTATION OF TWO ROTORS

We can analyze the dynamics of two rotors using the wave functions in explicit form. Here we study the dependence of basis states on the dynamical variables. An analysis of this dependence on the Euler angles α, β, γ is not of interest because these describe the well-studied rotation of the nucleus as a whole. Information about the relative motions of two rotors is, however, interesting. It can be obtained through an anlysis of the dependence of the basis states on the θ, ϕ_ν, and ϕ_π angles. As an example we consider in this section the nucleus ^{22}Ne $(\lambda_\nu=\lambda_\pi=4,\mu_\nu=2,\mu_\pi=0)$. We analyze the dependence of the density of the probability function on the θ and ϕ_ν angles. For this purpose we integrate the probability function

$$\rho^{L,M}_{(\lambda,\mu)}(\theta,\phi_\nu,\alpha,\beta,\gamma)\equiv|\Psi^{L,M}_{(\lambda,\mu)}|^2$$

over the Euler angles α, β, γ to get the function $\rho^{L}_{(\lambda,\mu)}(\theta,\phi_\nu)$. This function for ^{22}Ne for specific values of L,m,ν is

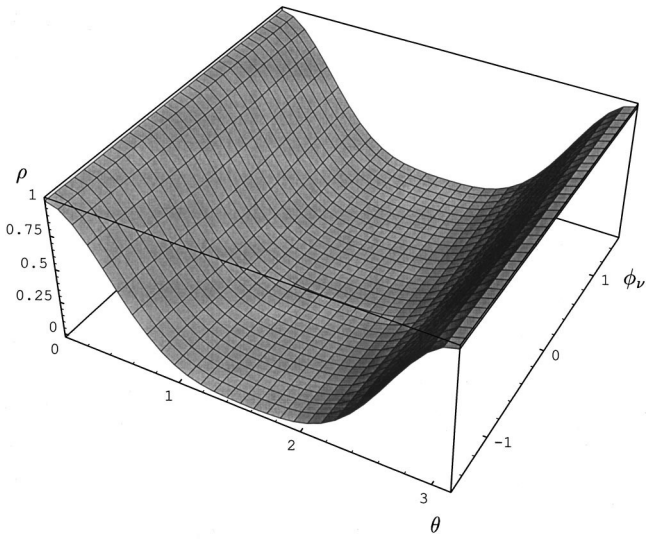

FIG. 1. The function $\rho^0_{(8,2)}(\theta,\phi_\nu)$ for ^{22}Ne.

$$\rho^0_{(8,2)}(\theta,\phi_\nu)=\left[1-\frac{8}{7}\sin^2\theta+\frac{8}{35}\sin^4\theta-\frac{6}{35}\sin^2\theta\cos^2\phi_\nu\left[1-\frac{4}{5}\sin^2\theta\right]\right]^2;$$

$$\rho^1_{(6,3)}(\theta,\phi_\nu)=\sin^2 2\theta\left[1-\frac{2}{5}\sin^2\theta-\frac{1}{5}\cos^2\phi_\nu-\frac{16}{175}\sin^2\theta\cos^2\phi_\nu\right]^2$$
$$+\frac{1}{400}\sin^2 2\phi_\nu\sin^2 2\theta\left[1-\frac{2}{5}\sin^2\theta\right]^2; \tag{37}$$

$$\rho^1_{(7,1)}(\theta,\phi_\nu)=\sin^2\theta\cos^2\phi_\nu\left\{\sin^2\phi_\nu\left(1-\frac{8}{7}\sin^2\theta+\frac{8}{35}\sin^4\theta\right)^2\right.$$
$$\left.+\cos^2\theta\cos^2\phi_\nu\left[1-\frac{4}{7}\sin^2\theta\right]^2\right\}$$

The behavior of these functions is shown in Figs. 1–3. {The values of $\rho^L_{(\lambda,\mu)}(\theta,\phi_\nu)$ are plotted on the vertical axis; the abscissa is the angle θ over the range $[0,\pi]$ where the ordinate is the angle

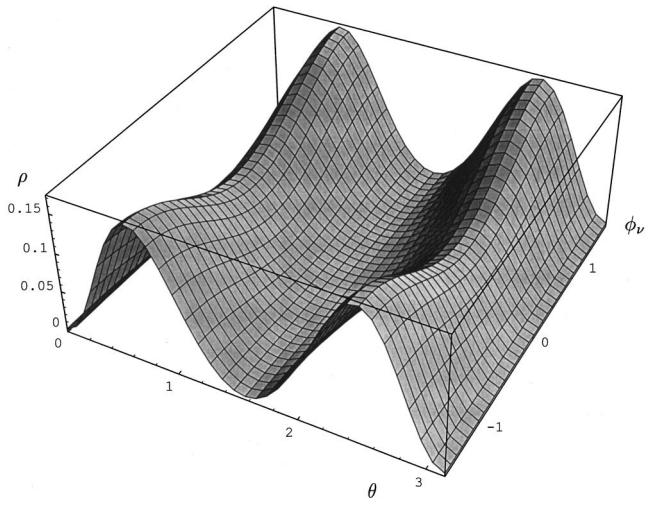

FIG. 2. The function $\rho^1_{(6,3)}(\theta,\phi_\nu)$ for ^{22}Ne.

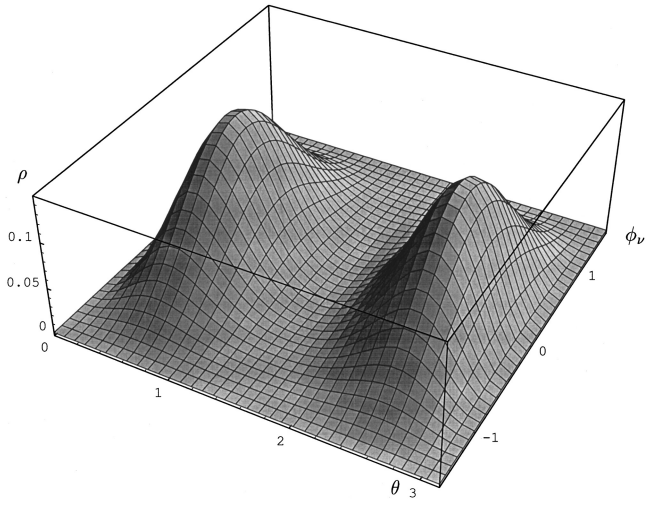

FIG. 3. The function $\rho^1_{(7,1)}(\theta,\phi_\nu)$ for ^{22}Ne.

ϕ_ν over the range $[-\pi/2,\pi/2]$.} Figure 1 indicates that the ground state corresponds to the maximum overlap of the proton and neutron distributions in the nucleus ($\theta=0$). In this case the function $\rho^0_{(\lambda,\mu)}(\theta,\phi_\nu)$ changes very slowly with angle ϕ_ν; nevertheless, it has maximum at $\phi_\nu=\pi/2$, which is difficult to observe in figures. The point $(\theta_{max},\phi_{\nu_{max}})=(0,\pi/2)$, for which $\rho^0_{(\lambda,\mu)}(\theta,\phi_\nu)$ has a maximum, corresponds to the limiting case when the system of one axial and one nonaxial rotor is transformed to one rotor.

The excitation of the nucleus to states with $L=1^+$ decreases the overlap of the two subsystems. The θ maximum of the $\rho^1_{(\lambda,\mu)}(\theta,\phi_\nu)$ function is shifted from $\theta=0$ to another θ_{max} in both the $m=1$ and $\nu_2=1$ cases. But one can see from Figs. 2 and 3 that the excited states with $m=1,\nu_2=0$ and $m=0,\nu_2=1$ differ one from another by the value of angle ϕ_ν ($\phi_{\nu_{max}}$) in which case $\rho^1_{(\lambda,\mu)}(\theta,\phi_\nu)$ has a maximum. When $m=1,\nu_2=0$ (state $|(\lambda_0-2,\mu_0+1),L=1,M\rangle$) the maximum is observed at $\phi_{\nu_{max}}=\pi/2$. This value of ϕ_ν corresponds to the transition from a system of one axial and one non-axial rotor into a system of two axial rotors. Actually, the state $|(\lambda-2,1),L=1,M\rangle$ appeared in the calculations of two axial rotors as the single lowest isovector state. The new state $|(\lambda_0-1,\mu_0-1),L=1,M\rangle$, that was absent in the case of two axial rotors is characterized by the maximum of the $\rho(\theta,\phi_\nu)$ function at $\phi_{\nu_{max}}=0$. This means that the most probable orientation of the neutron and proton subsystems is the one when $(\theta,\phi_\nu)=(\theta_{max},0)$, that is, when the three vectors $\mathbf{u}_1,\mathbf{v}_1,\mathbf{v}_3$ lie in a plane.

VII. CONCLUSION

In this paper we proposed a microscopic shell-model interpretation of the TRM. The theory is a neutron–proton version of the Elliott SU(3) model, where the final coupled SU(3) symmetry is constructed as a tensor product, $\mathrm{SU}_\pi(3)\otimes\mathrm{SU}_\nu(3)$. A basis for this new representation was constructed for $L=0$ and 1 states in Fock–Bargmann space. This representation is equivalent to a basis in coordinate space but has a simpler form that depends on only six variables. This simplification was achieved by constructing U(3) generators for the proton and neutron subsystems as well as for the total system. Through the construction of the second order Casimir operator of U(3) we were able to find by using generating function methods normalized wave functions for the ground state and two scissors-like states with $L=1^+$.

The efficiency of the method was illustrated with specific calculations for magnetic dipole transitions between nuclear states having quantum numbers $(\lambda_\nu+\lambda_\pi,\mu_\nu+\mu_\pi),L=0^+$, and $(\lambda_\nu+\lambda_\pi-2,\mu_\nu+\mu_\pi+1),L=1^+,M$, and $(\lambda_\nu+\lambda_\pi+1,\mu_\nu+\mu_\pi-2),L=1^+,M$. The results show that the approach predicts a fragmentation of the scissors mode into four transitions and we have obtained analytic expressions for the $B(M1)$–strengths corresponding to two of them. The results of calculations are compared with the corresponding pseudo-SU(3) predictions as well as experimental data for some rare earth and actinide nuclei.

We have also given a two rotor interpretation of the ground state and the excited isovector states and have connected changes in the underlying geometry with the transition operator. This has all been done without addressing the question of the excitation energies of the states involved. The reason is that it is known from Ref. 10 that there exists a simple Hamiltonian form built out of polynomial functions of the SU(3) generators that are SO(3) scalar that can be used to fit the experimental data. The Hamiltonian contains both kinetic (rotational spectrum within irreps) and potential (relative displacement of irreps) energy parts that can be used to reproduce rotor spectra, including a correct placement of the $M1$ band.

Only even–even nuclei have been considered in this paper. The approach can be extended to the general case when the total spin of a nucleus is nonzero. This case, as well as more complicated shapes that involve multiplicities, will be considered in a later publication.

ACKNOWLEDGMENT

This work was supported in part by ISSEP of the International Renaissance Foundation through Grant No. PSU 072033.

[1] A. B. Migdal, JETP **15**, 81 (1945).
[2] M. Goldhaber and E. Teller, Phys. Rev. **74**, 1046 (1948).
[3] N. Lo Iudice and F. Palumbo, Nucl. Phys. A **326**, 193 (1979).
[4] D. Bohle, A. Richter, W. Steffen, A. E. Dieperink, N. Lo Iudice, F. Palumbo, and O. Scholten, Phys. Lett. B **137**, 141 (1984); D. Bohle, G. Kuchler, A. Richter and W. Steffen, Phys. Lett. B **148**, 260 (1984).
[5] U. E. P. Berg *et al.*, Phys. Lett. B **149**, 59 (1984); C. Wesselborg *et al.*, Z. Phys. A **323**, 485 (1986).
[6] C. Djalali *et al.*, Phys. Lett. B **164**, 269 (1985).
[7] F. Iachello, Nucl. Phys. A **358**, 89 (1981).
[8] A. E. L. Dieperink, Prog. Part. Nucl. Phys. **9**, 121 (1983).
[9] M. Sambataro and A. E. L. Dieperink, Phys. Lett. B **107**, 249 (1981); M. Sambataro, O. Sholten, A. E. L. Dieperink, and G. Piccitto, Nucl. Phys. A **423**, 233 (1984).
[10] J. P. Draayer and K. J. Weeks, Ann. Phys. **156**, 41 (1984); O. Castanos, J. P. Draayer, and Y. Leschber, Ann. Phys. **180**, 290 (1987).
[11] D. R. Bes and R. A. Broglia, Phys. Lett. B **137**, 141 (1983).
[12] S. Iwasaki and K. Hara, Phys. Lett. B **144**, 9 (1984).
[13] K. Sugaware-Tanabe and A. Arima, Phys. Lett. B **206**, 573 (1988).
[14] J. Speth and B. Zawischa, Phys. Lett. B **211**, 247 (1988).
[15] T. Suzuki and D. J. Rowe, Nucl. Phys. A **289**, 461 (1977).
[16] E. Lipparini and S. Stringari, Phys. Lett. B **130**, 139 (1983).
[17] G. De Franceschi, F. Palumbo, and N. Lo Iudice, Phys. Rev. C **29**, 1496 (1984).
[18] G. F. Filippov and I. S. Dotsenko, Yad. Fiz. **57**, 2181 (1994) [Phys. At. Nucl. **57**, 2099 (1994)].
[19] G. F. Filippov, I. S. Dotsenko, and A. F. Lisetskyi, preprint ITP-94-3R (1994) (in Russian).
[20] G. F. Filippov, I. S. Dotsenko, and A. F. Lisetskyi, preprint ITP-95-13E (1995).
[21] J. P. Elliott, Proc. R. Soc. London **245**, 128, 562 (1958).
[22] G. F. Filippov and A. L. Blokhin, J. Math. Phys. **32**, 1683 (1991); V. I. Avramenko and A. L. Blokhin, J. Math. Phys. **32**, 1696 (1991).
[23] V. Bargmann and M. Moshinsky, Nucl. Phys. **18**, 697 (1960).
[24] I. S. Dotsenko and G. F. Filippov. Sov. J. Part. Nucl. **27**, 335 (1996); G. F. Filippov, I, S. Dotsenko, and A. F. Lisetskyi, preprint ITP-95-7R (1995) (in Russian).
[25] G. F. Filippov, V. S. Vasilevsky, and L. L. Chopovsky, Sov. J. Part. Nucl. **15**, 1338 (1984) (in Russian).
[26] G. F. Filippov, A. V. Nesterov, I. Yu. Rybkin, and S. V. Korennov, Sov. J. Part. Nucl. **25**, 1347 (1994).
[27] G. F. Filippov, I. Yu. Rybkin, S. V. Korennov, and K. Kato, J. Math. Phys. **36**, 4571 (1995).
[28] A. M. Perelomov, *Generalized Coherent States and their Applications* (Springer, Berlin, 1987).
[29] D. Rompf, J. P. Draayer, D. Troltenier, and W. Scheid, Z. Phys. A **354**, 359 (1996).

Generalized annihilation operator coherent states

Gautam Ghosh[a)]
Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Calcutta 700 064, India

(Received 21 July 1997; accepted for publication 4 September 1997)

A set of generalized coherent states as eigenstates of the annihilation operator is proposed. These states are analytic functions of a complex variable and admit a resolution of identity with positive measure. Guided by the classical action angle transformation and the correspondence principle a formalism is developed for the construction of the annihilation operator for a given Hamiltonian. © *1998 American Institute of Physics.* [S0022-2488(98)00603-3]

I. INTRODUCTION

The coherent states of the harmonic oscillator[1,2] are a set of states satisfying a large number of interesting and useful properties. They are (1) eigenfunctions of the annihilation operator, (2) entire functions of a complex variable, (3) a complete set of basis vectors and (4) minimum uncertainty states. Moreover, (5) these states can be constructed from the ground state by the action of a unitary displacement operator and (6) they are temporally stable, i.e., under time evolution they remain coherent. Historically, these states were introduced in an attempt to derive classical behavior viz., a nonspreading wave packet traveling coincidentally with the classical particle from the underlying quantum dynamics.[3] More general coherent states which include the oscillator coherent states as a special case were constructed. These are the generalized coherent states (GCS) of Perelomov and Gilmore,[4,5] the minimum uncertainty coherent states (MUCS) of Nieto and Simmons[6,7] and the very recent temporally stable coherent states (TSCS) of Klauder.[8] The GCS, defined through property (5), are richly endowed with a variety of interesting characteristics[9,10] including completeness arising from the group property of the unitary displacement operator. In general, they are neither minimum uncertainty states nor temporally stable. The TSCS of Klauder satisfy property (6) by construction and they are also complete. However like the GCS they are in general neither minimum uncertainty states nor eigenstates of the annihilation operator. The MUCS were proposed to have minimum uncertainty products with respect to operator pairs called natural quantum operators that are counterparts of classical variables which evolve like oscillator solutions. The MUCS have not been shown to be complete in general. Both the GCS and MUCS can be constructed for both finite and infinite spectra (including possible degeneracy) while the TSCS are defined only for infinite spectra.

In this work we propose a set of coherent states which are eigenstates of suitably constructed annihilation operators called generalized annihilation operators (GAO). The GAO can be obtained for both finite and infinite spectra through mappings to the linear spectrum of the oscillator so that they are direct generalizations of the oscillator annihilation operator. The coherent states defined only for infinite spectra will be analytic functions of complex variables and will be complete with respect to suitably chosen measures. Finally, these states will minimize the uncertainty products of operator pairs in a sense to be explained later.

II. FORMALISM

The construction of the GAO will be achieved within an action-angle framework. From the commutation relation (we take $\hbar=1$),

$$[J,\ e^{\pm i\phi}]=\pm e^{\pm i\phi}, \tag{1}$$

[a)]Electronic mail: gghosh@tnp.saha.ernet.in

 © 1998 American Institute of Physics

the eigenvalues (n) of J may be taken to be integers (modulo an inessential shift), unbounded both above and below while the unitary operators $e^{\pm i\phi}$ act as shift operators within the full eigenspace of J. However, identifying the ground state of the Hamiltonian with the state $n=0$, we note that the eigenspace ($\mathscr{H}:n\geqslant 0$) of the Hamiltonian is not invariant under the action of the shift operators. The problem is resolved for the oscillator where the creation and annihilation operators ($a^\dagger$ and a) turn out to be $\sqrt{J}e^{i\phi}$ and $e^{-i\phi}\sqrt{J}$ so that the ground state is indeed destroyed by a. The GAO are natural generalizations in the sense that for a given Hamiltonian we will look for operators of the form $e^{-i\phi}f(J)$ with the demand that $f(0)=0$, so as to ensure the aforementioned invariance and bypass the problem of the nonunitarity[11–16] of $e^{\pm i\phi}$ within $\mathscr{H}$. It turns out that the function $f(J)$ is essentially uniquely determined by the following requirements:

(i) The Hamiltonian operator must reduce to a function of J alone. This will directly lead to the spectra once J is replaced by n.

(ii) The relations between the quantum operators $(J,e^{\pm i\phi})$ and (x,p) must reduce to the classical relations in the leading order of J.

(iii)
$$e^{-in\phi}g(J)e^{in\phi}=g(J+n), \tag{2}$$

where $g(J)$ is any given function of J and $n\geqslant 0$. This is the same thing as (1) written in a form such that both the left- and right-hand sides of (2) are well defined in $\mathscr{H}$.

Once the generalized annihilation operator is constructed, the ground state in the coordinate representation may be obtained by solving a first order differential equation. Successive applications of the creation operator which is a first order differential operator then yields the excited state wave functions.[17] Since the formalism employs the classical action-angle transformation as the guiding principle, the construction of the GAO is limited to separable systems only. In the following we illustrate the method by working out the GAO and the corresponding coherent states in a number of standard problems involving an infinite number of bound states.

III. EXAMPLES OF COHERENT STATES

A. The harmonic oscillator with centripetal barrier (HOCB) potential

This potential supports a simple linear spectrum. The Hamiltonian (in appropriate units),

$$H=p^2/2+U\left(x-\frac{1}{x}\right)^2, \tag{3}$$

leads to the following classical relations between x, p and $J\equiv(1/2\pi)\oint p(E,x)dx$ and $\phi=(\partial/\partial J)\int^x p(J,x)dx$,

$$H=\sqrt{8U}J, \tag{4}$$

$$x^2=\left(1+\sqrt{\frac{2}{U}}\,J\right)+\sqrt{\frac{2}{U}}\,(J^2+\sqrt{2U}J)^{1/2}\sin\,\phi. \tag{5}$$

In conformity with the requirements mentioned above, we write the corresponding quantum relations as

$$H=\sqrt{8U}(J+\alpha), \tag{6}$$

$$x^2=a(J)+\frac{1}{2i}\,[b(J)e^{i\phi}-e^{-i\phi}b(J)], \tag{7}$$

where the constant α and the functions $a(J)$ and $b(J)$ are to be determined.

Equation (6) has been written so as to agree with (4) in the leading order in J, the term involving α having been introduced as a possible quantum correction. Similarly, the correspondence principle places restrictions on $a(J)$, $b(J)$ for $J\to\infty$. Furthermore for $b(J)\exp(i\phi)$ and $\exp(-i\phi)b(J)$ to be well defined in $\mathscr{H}$, $b(J)$ must satisfy $b(0)=0$.

The operators occurring on the left-hand side of (6) and (7) satisfy

$$[[H,\ x^2],\ x^2]=-4x^2 \tag{8a}$$

and

$$[H,\ x^2]^2=-4\{H,x^2\}+8U(x^2-1)^2-3, \tag{8b}$$

where {,} implies an anticommutator. Replacing these by corresponding operators on the right-hand side, and imposing (2) we get a set of consistency conditions determining the functions $a(J),b(J)$ through a set of difference equations. For instance, equating coefficients of $\exp(i\phi)$ in (8a) and (8b) one gets,

$$a(J)=1+\sqrt{\frac{2}{U}}\,(J+\alpha), \tag{9}$$

whereas equating the constants on both sides of (8a) and (8b) yields,

$$b(J)=\sqrt{\frac{2}{U}}\left[\left(J+\alpha+\sqrt{\frac{U}{2}}\right)\left(J+\alpha+\sqrt{\frac{U}{2}}-1\right)-\frac{U}{2}+\frac{3}{16}\right]^{1/2}, \tag{10}$$

so that the condition $b(0)=0$ yields,

$$\alpha=\frac{1}{2}\,(1-\sqrt{2U})+\left(\frac{U}{2}+\frac{1}{16}\right)^{1/2}. \tag{11}$$

It follows that (7) indeed reduces to the classical relation (5) in the leading order of J. We now define our GAO to be $e^{-i\phi}b(J)$ so that upon substitution for α from (11) in Eq. (10) we may write,

$$\mathscr{A}(J)\equiv\sqrt{\frac{2}{U}}\,e^{-i\phi}\sqrt{J(J+2\beta)}, \tag{12}$$

where $\beta=[(U/2)+(1/16)]^{1/2}$ and satisfying,

$$\mathscr{A}(J)|n\rangle=\sqrt{\frac{2}{U}}\,\sqrt{n(n+2\beta)}|n-1\rangle. \tag{13}$$

We can immediately write our desired coherent state as an eigenstate of $\mathscr{A}(J)$ belonging to eigenvalue z viz.

$$|z\rangle=[f(|z|^2)]^{-1/2}\sum\ \frac{z^n}{h_n}\,|n\rangle, \tag{14}$$

where $h_0\equiv 1$, $h_n=\Pi_{r=1}^{n}\sqrt{r(r+2\beta)}$ and the function f is defined as

$$f(s)=\sum_{n=0}^{\infty}\ \frac{s^n}{h_n^2}. \tag{15}$$

We see that (14) can be rewritten as,

$$|z\rangle=[f(|z|^2)]^{-1/2}f(z\mathscr{A}^{\dagger})|0\rangle. \tag{16}$$

The function f which embodies all the characteristics of the coherent state can be explicitly worked out to be,

$$f(s)=\Gamma(2\beta+1)\left(\frac{2}{Us}\right)^{\beta}I_{2\beta}(\sqrt{2Us}), \tag{17}$$

where $I_{2\beta}$ is the modified Bessel function of order 2β. In the coordinate representation Eq. (14) reads,

$$\psi_z(x)=\left[\frac{2\sqrt{2U}}{I_{2\beta}(|z|\sqrt{2U})}\right]^{1/2}\left(\frac{|z|}{z}\right)^{\beta}e^{\sqrt{U/2}(z-x^2)}\sqrt{x}J_{2\beta}(2x\sqrt{Uz}). \tag{18}$$

We now look for an isotropic weight function $\mu(|z|)$ such that the resolution of the identity operator reads,

$$\int |z\rangle\langle z|\mu(|z|)d^2z=1. \tag{19}$$

Taking diagonal matrix elements and using the orthonormality of the $|n\rangle$ states we find,

$$\int_0^{\infty}[f(s)]^{-1}s^n\mu(s)ds=\frac{h_n^2}{\pi}, \tag{20}$$

which is a moment problem for $\mu(s)/f(s)$ with $s=|z|^2$ and can be solved by the method of characteristic function to yield,

$$\mu(s)=\frac{U}{\pi}I_{2\beta}(\sqrt{2Us})K_{2\beta}(\sqrt{2Us}), \tag{21}$$

where $K_{2\beta}$ is another modified Bessel function.

The state so obtained is the same as the one obtained by Nieto and Simmons in their research on MUCS. To investigate the minimum uncertainty property of $\psi_z(x)$, we observe that an eigenstate of the operator $\mathscr{A}(J)$ minimizes the uncertainty product of the real and imaginary parts of $\mathscr{A}(J)$, i.e., if we write $\mathscr{A}(J)=A+iB$, where A and B are Hermitian operators, then $\Delta A\Delta B=\frac{1}{2}|\langle[A,B]\rangle|$, where the dispersions are calculated in the state $\psi_z(x)$. The operators A and B turn out to be $A=xp+px$ and $B=H-2Ux^2$ (modulo constant factors) which are essentially the same as the natural quantum operators of Nieto and Simmons.

As is expected the state $|z\rangle$ evolves as $|z(t)\rangle=|z(0)e^{-i\omega t}\rangle$, where $\omega(=\sqrt{8U})$ is the same as the classical angular frequency. We observe that mappings at the classical level (Nieto and Simmons) and quantum level (present work) yield the same coherent state for this linear spectrum case.

B. The Pöschl–Teller potential

This potential supports a quadratic spectrum and we will summarize our results below.

Classical:

$$H=\frac{p^2}{2}+U\ \mathrm{cosec}^2\ x=\frac{1}{2}\ (J+\sqrt{2U})^2,$$

$$\cos\ x=-\frac{[J^2+2\sqrt{2U}J]^{1/2}}{J+\sqrt{2U}}\ \sin\ \phi.$$

Quantum:

$$H=\tfrac{1}{2}\ (J+\alpha)^2,$$

$$\alpha=\tfrac{1}{2}\ (1+\sqrt{1+8U}),$$

$$\cos x=\frac{1}{2i}\,[b(J)e^{i\phi}-e^{-i\phi}b(J)],$$

$$b(J)=-\frac{\sqrt{J(J+2\alpha-1)}}{\sqrt{(J+\alpha)(J+\alpha-1)}}.$$

Dropping a minus sign in the definition, the GAO therefore satisfies,

$$\mathscr{A}(J)|n\rangle=\frac{\sqrt{n(n+2\alpha-1)}}{\sqrt{(n+\alpha)(n+\alpha-1)}}\,|n-1\rangle, \tag{22}$$

and the corresponding coherent state is written as,

$$|z\rangle=[f(|z|^2)]^{-1/2}\left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha+1)}\right]^{1/2}\sum_{n=0}^{\infty}\left[\frac{\Gamma(n+\alpha)\Gamma(n+\alpha+1)}{n!\Gamma(n+2\alpha)}\right]^{1/2}z^n|n\rangle, \tag{23}$$

where the normalization is given in terms of the hypergeometric function as,

$$f(s)={}_2F_1(\alpha,\alpha+1,2\alpha;s), \tag{24}$$

which diverges at $s=1$. The coherent states are therefore defined strictly within the unit circle ($|z|<1$).

The wave functions $\langle x|n\rangle$ involve Gegenbauer polynomials and the series in Eq. (23) can be summed by using a generating function of these polynomials[18] to yield the coordinate representation in the variable $y(=\cos x)$ viz.

$$\psi_z(y)\approx\frac{\Gamma(\alpha+1)}{[\pi\alpha 2^{1-2\alpha}\Gamma(2\alpha)]^{1/2}}\,(1-y^2)^{\alpha/2}(1-yz)^{-(\alpha+1)}{}_2F_1\left[\frac{\alpha+1}{2},\,\frac{\alpha+2}{2},\alpha+\frac{1}{2};\,\frac{(y^2-1)z^2}{(1-yz)^2}\right], \tag{25}$$

with the same normalization as in (23). To obtain a resolution of the identity we observe that the weight function satisfies the following well defined moment problem viz.,

$$\int_0^{\infty}g(s)s^n ds=\frac{1}{\pi}\,\frac{\Gamma(\alpha)\Gamma(\alpha+1)}{\Gamma(2\alpha)}\,\frac{n!\Gamma(n+2\alpha)}{\Gamma(n+\alpha)\Gamma(n+\alpha+1)},$$

where $g(s)[=\mu(s)/f(s)]$ is nonzero only in the interval $0\leqslant s<1$ and $s=|z|^2$. These coherent states differ from the MUCS of Nieto and Simmons. In fact these states minimize the uncertainty product of the operators $\cos x$ and $i[\sqrt{H},\cos x]$. However, like the MUCS the states in question do not remain coherent under time evolution, the equation of motion for the annihilation operator being given by,

$$\dot{\mathscr{A}}(J)=-i\mathscr{A}(J)(J-\tfrac{1}{2}+\alpha).$$

C. The Coulomb potential

We use squared parabolic coordinates[19] (μ,ν,ϕ) which are related to the rectangular coordinates (x,y,z) as,

$$x=\mu\nu\,\cos\,\phi,\;\; y=\mu\nu\,\sin\,\phi,\;\; z=\tfrac{1}{2}(\mu^2-\nu^2).$$

The Hamiltonian may therefore be written as,

$$H=\frac{1}{2}\,\frac{p_\mu^2+p_\nu^2}{\mu^2+\nu^2}+\frac{p_\phi^2}{2\mu^2\nu^2}-\frac{2e^2}{\mu^2+\nu^2}, \tag{26}$$

e being the electronic charge. Replacing p_ϕ by its quantized value m and remembering that $H=-E$ for a bound state we will consider the pair of Hamiltonians defined as,

$$H_\mu=\frac{p_\mu^2}{2}+\frac{m^2}{2\mu^2}+E\mu^2, \tag{27a}$$

$$H_\nu=\frac{p_\nu^2}{2}+\frac{m^2}{2\nu^2}+E\nu^2, \tag{27b}$$

with the condition that,

$$H_\mu+H_\nu=2e^2. \tag{28}$$

We recognize H_μ and H_ν to be a 2D harmonic oscillator Hamiltonians in polar coordinates. Proceeding along the lines by now familiar we obtain,

$$H_\mu=2\sqrt{2E}\left(J_\mu+\frac{1}{2}+\frac{|m|}{2}\right),$$

$$\mu^2=a(J_\mu)+\frac{1}{2i}\left[b(J_\mu)e^{i\phi_\mu}-e^{-i\phi_\mu}b(J_\mu)\right],$$

$$a(J_\mu)=\sqrt{\frac{2}{E}}\left(J_\mu+\frac{1}{2}+\frac{|m|}{2}\right),$$

$$b(J_\mu)=\sqrt{\frac{2}{E}}\left[J_\mu(J_\mu+|m|)\right]^{1/2},$$

and similar ones for the index ν. Imposition of (28) yields the energy eigenvalues. The coherent states are now constructed as eigenstates of the commuting GAO viz.,

$$\mathcal{A}(J_{\mu,\nu})=e^{-i\phi_{\mu,\nu}}[J_{\mu,\nu}(J_{\mu,\nu}+|m|)]^{1/2},$$

with eigenvalues z_1 and z_2, respectively.

The normalized state is obtained as

$$|z_1,z_2,m\rangle=\left[\frac{I_{|m|}(2|z_1|)I_{|m|}(2|z_2|)}{|z_1|^{|m|}|z_2|^{|m|}}\right]^{-1/2}\sum_{n_1n_2}\frac{z_1^{n_1}z_2^{n_2}}{\sqrt{n_1!n_2!(n_1+|m|)!(n_2+|m|)!}}|n_1,n_2,m\rangle, \tag{29}$$

where n_1 and n_2 are the eigenvalues of J_μ and J_ν, respectively. The resolution of the identity reads,

$$\sum_m\int d^2z_1d^2z_2\mu(z_1)\mu(z_2)|z_1,z_2,m\rangle\langle z_1,z_2,m|=1, \tag{30}$$

where the weight function μ is given by,

$$\mu(s)=\frac{2}{\pi}I_{|m|}(2|s|)K_{|m|}(2|s|). \tag{31}$$

IV. CONCLUDING REMARKS

In summary, the classical action angle transformation together with the correspondence principle has been used to transform to quantum action angle operators and thereby to construct the creation and annihilation operators and the corresponding coherent states for some exactly soluble and separable problems of quantum mechanics admitting infinite spectra. The states in question

are analytic functions of complex variables and form a complete set of basis vectors. In a non-confining potential with a finite number of bound states we could also construct coherent states as eigenstates of the GAO in the coordinate representation. However, these states will necessarily contain contributions from the continuum. The usual problems with the angle operator do not arise since the formulation involves only those operators that explicitly leave the Hilbert space of the problem invariant. The Coulomb coherent states have been defined for fixed m and should prove useful in the study of the highly excited states of atoms in magnetic fields.[20] Further investigations on these states as also relevant applications will be published elsewhere.

ACKNOWLEDGMENT

I thank Dr. Abhijit Lahiri for many helpful discussions.

[1] R. J. Glauber, Phys. Rev. **131**, 2766 (1963).
[2] E. C. G. Sudarshan, Phys. Rev. Lett. **10**, 84 (1963).
[3] E. Schrödinger, Naturwissenschaften **14**, 664 (1926).
[4] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[5] R. Gilmore, Ann. Phys. (N.Y.) **74**, 391 (1972).
[6] M. M. Nieto and L. M. Simmons, Jr., Phys. Rev. Lett. **41**, 207 (1978).
[7] M. M. Nieto, Phys. Rev. D **30**, 770 (1984), and references therein.
[8] J. R. Klauder, J. Phys. A **29**, L293 (1996).
[9] A. M. Perelomov, *Generalized Coherent States and Their Applications* (Springer, Berlin, 1986).
[10] W. Zhang, D. H. Feng, and R. Gilmore, Rev. Mod. Phys. **62**, 867 (1990).
[11] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **114**, 243 (1927).
[12] L. Susskind and J. Glogower, Physica (Amsterdam) **1**, 49 (1964).
[13] P. Carruthers and M. M. Nieto, Phys. Rev. Lett. **14**, 387 (1965).
[14] P. Carruthers and M. M. Nieto, Rev. Mod. Phys. **40**, 411 (1968).
[15] M. M. Nieto, Phys. Scr. **48**, 5 (1993). This issue is wholly devoted to various aspects of quantum phase and phase dependent measurements.
[16] D. T. Pegg and S. M. Barnett, Europhys. Lett. **6**, 483 (1988); Phys. Rev. A **39**, 1665 (1989). These authors provide an ingenius way of solving the unitarity problem.
[17] A. Lahiri, G. Ghosh, and T. K. Kar (unpublished).
[18] H. M. Srivastava and H. L. Manocha, *A Treatise on Generating Functions* (Ellis Horwood, Chichester, 1984).
[19] F. Ravndahl and T. Toyoda, Nucl. Phys. B **3**, 312 (1967).
[20] K. Bhattacharya, A. Lahiri, D. Bhaumik, and B. Dutta-Roy (unpublished).

Radon–Nikodym derivatives of quantum instruments

A. S. Holevo[a)]
Steklov Mathematical Institute, Gubkina 8, 117966 Moscow, Russia

(Received 15 April 1997; accepted for publication 8 October 1997)

A convenient representation for Radon–Nikodym derivatives of completely positive (c.p.) instruments on $B(\mathscr{H})$ with respect to a scalar measure is suggested, similar to the Stinespring–Kraus representation for c.p. maps, but involving possibly nonclosable unbounded operators. The structure of covariant c.p. instruments is studied in detail. In particular, an exhaustive description is given to instruments covariant with respect to shifts or rotations, corresponding to ''position'' or ''angle'' measurements. © *1998 American Institute of Physics.*
[S0022-2488(98)01802-7]

I. INTRODUCTION

The notion of instrument was introduced by Davies and Lewis[1] in order to treat the repeatability issue for observables with continuous spectrum in quantum measurement theory. Ozawa[2] introduced complete positivity into the definition of Ref. 1 and showed that *any* completely positive (c.p.) instrument can be obtained from a measurement procedure consisting of interaction with apparatus and subsequent observation of the apparatus. Since then the notion of c.p. instrument was successfully applied in a number of concrete problems such as counting processes in quantum optics and quantum communication,[3] the standard quantum limit in high-precision experiments,[4] quantum Zeno paradox and time continuous measurement processes,[5,6] etc. The importance of the concept of instrument lies in the fact that it gives a condensed linear description, in terms of the system alone, of the measurement statistics and of posterior states of the system, i.e., the states conditioned upon certain outcome of the measurement. It should be stressed that the abstractness of the notion of instrument, which may count for disadvantage from point of view of a physicists, is due to the highly reduced nature of the statistical description of a measuring procedure[7] and is just the reverse side of its mathematical advantages. The property of complete positivity is crucial for relating the abstract notion of the instrument to a measurement process.

From a mathematical point of view, an instrument is a measure with values in the cone of operations—positive linear contractions of the algebra of observables.[8] In the case of finite-dimensional Hilbert space, a Radon–Nikodym theorem for such measures was proved by Davies,[9] giving operation-valued density (Radon–Nikodym derivative) with respect to a scalar measure. The main objective of this paper is to prove results of this type for c.p. instruments in arbitrary Hilbert space. The resulting density has structure formally similar to Stinespring–Kraus representation of c.p. maps, but involves in general nonclosable unbounded operators (Sec. II). Particularly interesting is the case where the instrument is covariant under unitary action of a symmetry group (Sec. III). As an important application, we give in Secs. IV and V an exhaustive description of instruments, covariant with respect to rotations and shifts, corresponding to quantum measurements of ''angle'' or ''position'' variables.

II. A GENERAL RADON–NIKODYM THEOREM FOR COMPLETELY POSITIVE INSTRUMENTS

Let $\mathscr{H}$ be a separable Hilbert space, $\mathscr{A}\subset B(\mathscr{H})$ be a von Neumann algebra with the predual $\mathscr{A}_*$, and $(\mathscr{X},\mathscr{B})$ be a Borel space (where $\mathscr{B}$ is σ-algebra of measurable subsets of the set $\mathscr{X}$). A *completely positive (c.p.) instrument in* $\mathscr{A}$ *with the value space* $\mathscr{X}$ is a measure $B\to\mathscr{M}(B)[\,\cdot\,]on\mathscr{B}$, whose values are bounded linear normal c.p. maps $X\to\mathscr{M}(B)[X]$ of $\mathscr{A}$, such that for

[a)]Electronic mail: holevo@genesis.mi.ras.ru

 © 1998 American Institute of Physics

any $S\in\mathcal{A}_*$, $X\in\mathcal{A}$, the scalar measure $B\to\langle S,\mathcal{M}(B)[X]\rangle$ is σ-additive; moreover, the measure $\mathcal{M}(\cdot)$ must satisfy the normalization condition $\mathcal{M}(\mathcal{X})[I]=I$, where I is the unit operator in $\mathcal{H}$.

In quantum measurement theory the notion of instrument is used as an equivalent description of the measurement probability distribution and posterior states for an arbitrary initial state of the system. If S is the initial state of the system, then the measurement probability distribution is

$$\mu_S(B)=\langle S,\mathcal{M}(B)[I]\rangle,\quad B\in\mathcal{B},$$

and is equivalently given by the associated *probability operator-valued measure* (p.o.m.)

$$M(B)=\mathcal{M}(B)[I],\quad B\in\mathcal{B}$$

(see Refs. 9 and 10). Posterior states S_x corresponding to measurement outcomes $x\in\mathcal{X}$ are determined from the relation

$$\langle S,\mathcal{M}(B)[X]\rangle=\int_B\langle S_x,X\rangle\mu_S(dx)$$

(see Ref. 11 for more detail).

The following result was proved in Ref. 2 (Proposition 4.2).

Proposition 1: For arbitrary c.p. instrument $\mathcal{M}$ there exists a Hilbert space $\mathcal{K}$, a spectral measure $E:B\to E(B)$, $B\in\mathcal{B}$, in $\mathcal{K}$; a nondegenerate normal $*$-representation $\pi:X\to\pi[X]$, $X\in\mathcal{A}$, in $\mathcal{K}$; and a linear isometry V from $\mathcal{H}$ to $\mathcal{K}$ such that

$$\mathcal{M}(B)[X]=V^*E(B)\pi[X]V, \tag{2.1}$$

$$E(B)\pi[X]=\pi[X]E(B), \tag{2.2}$$

for any $B\in\mathcal{B}$. $X\in\mathcal{A}$. □

In the case $\mathcal{A}=B(\mathcal{H})$, which will be our main concern, one has

$$\mathcal{K}=\mathcal{H}\otimes\mathcal{H}_0,\quad \pi[X]=X\otimes I_0,$$

where I_0 is the unit operator in $\mathcal{H}_0$ (see e.g., Ref. 9, Lemma 9.2.2). Then (2.2) implies

$$E(B)=I\otimes E_0(B),$$

where E_0 is a spectral measure in $\mathcal{H}_0$, so that the representation (2.1) reads

$$\mathcal{M}(B)[X]=V^*(X\otimes E_0(B))V. \tag{2.3}$$

By using this we can deduce the following analog of the Stinespring–Kraus representation of a c.p. map on $B(\mathcal{H})$ (for related results see also Ref. 12).

Theorem 1: There exist a positive σ-finite measure μ on $(\mathcal{X},\mathcal{B})$, a dense domain $\mathcal{D}\subset\mathcal{H}$, and a countable family of functions $x\to A_k(x)$ defined for almost all x, such that $A_k(x)$ are linear operators from $\mathcal{D}$ to $\mathcal{H}$, satisfying

$$\int_{\mathcal{X}}\sum_k\|A_k(x)\psi\|^2\mu(dx)=\|\psi\|^2,\quad \psi\in\mathcal{D}, \tag{2.4}$$

and

$$(\psi|\mathcal{M}(B)[X]\psi)=\int_B\sum_k(A_k(x)\psi|XA_k(x)\psi)\mu(dx),\quad \psi\in\mathcal{D}. \tag{2.5}$$

Proof: According to von Neumann's spectral theorem, $\mathcal{H}_0$ can be decomposed into the direct integral

$$\mathscr{H}_0 = \int_{\mathscr{X}}^{\oplus} \mathscr{H}(x)\mu(dx) \tag{2.6}$$

with respect to some positive σ-finite measure μ, diagonalizing the spectral measure E_0:

$$E_0(B)\phi = \int_B^{\oplus} \phi(x))\mu(dx), \tag{2.7}$$

where $\phi(x) \in \mathscr{H}(x)$ are the components of the vector $\phi \in \mathscr{H}_0$ in the decomposition (2.6). In what follows we shall fix a measurable field of orthonormal bases $\{e_{k,x}\}$ in the direct integral (2.6) and denote $\phi_k(x) = (e_{k,x}|\phi(x))$, where the inner product is in $\mathscr{H}(x)$ [note that $\phi_k(x)$ are defined μ-almost everywhere].

The tensor product $\mathscr{K} = \mathscr{H} \otimes \mathscr{H}_0$ has the direct integral decomposition

$$\mathscr{H} \otimes \mathscr{H}_0 = \int_{\mathscr{X}}^{\oplus} (\mathscr{H} \otimes \mathscr{H}(x))\mu(dx).$$

Let $\psi \in \mathscr{H}$. The decomposition of the vector $V\psi \in \mathscr{H} \otimes \mathscr{H}_0$ then reads

$$V\psi = \int_{\mathscr{X}}^{\oplus} \sum_k (V\psi)_k(x) e_{k,x}\mu(dx),$$

where $(V\psi)_k(x)$ are (equivalence classes of) μ-measurable functions with values in $\mathscr{H}$ satisfying

$$\int_{\mathscr{X}} \sum_k \|(V\psi)_k(x)\psi\|^2 \mu(dx) = \|\psi\|^2, \quad \psi \in \mathscr{H}, \tag{2.8}$$

since V is isometric. Since V is linear, we have for μ-almost all x

$$\left(V\left(\sum_j \lambda_j \psi_j \right) \right)_k (x) = \sum_j \lambda_j (V\psi_j)_k(x), \tag{2.9}$$

where $\{\psi_j\} \subset \mathscr{H}$ is a fixed countable system of vectors, and λ_j are complex numbers, only a finite number of which are nonzero.

Let us now fix an orthonormal basis $\{\psi_j\} \subset \mathscr{H}$ and let $\mathscr{D} = \mathrm{lin}\{\psi_j\}$ be its linear span. We define linear operators $A_k(x)$ on $\mathscr{D}$ by the relation

$$A_k(x)\left(\sum_j \lambda_j \psi_j \right) = \sum_j \lambda_j (V\psi_j)_k^0(x),$$

where $(V\psi_j)_k^0$ is a fixed representative of the equivalence class $(V\psi_j)_k$. Then by (2.9), for any fixed $\psi \in \mathscr{D}$ there is a subset $\mathscr{X}_\psi \subset \mathscr{X}$ such that $\mu(\mathscr{X} \setminus \mathscr{X}_\psi) = 0$ and

$$A_k(x)\psi = (V\psi)_k(x) \quad \text{for all } k, \quad \text{all } x \in \mathscr{X}_\psi. \tag{2.10}$$

Combining (2.3), (2.7), and (2.10), we obtain (2.5). The normalization condition (2.4) follows from (2.7) and (2.10). □

The representation (2.5) is especially convenient from the point of view of calculation of the measurement probability distribution and posterior states. Take pure initial state $S = |\psi)(\psi|$, where $\psi \in \mathscr{D}$. Then the probability distribution of outcomes is

$$\mu_\psi(B) = (\psi|\mathscr{M}(B)[I]\psi) = \int_B \sum_k \|A_k(x)\psi\|^2 \mu(dx). \tag{2.11}$$

The posterior state corresponding to the outcome x is

$$S_\psi(x)=\frac{\Sigma_k|A_k(x)\psi)(A_k(x)\psi|}{\Sigma_k\|A_k(x)\psi\|^2}. \tag{2.12}$$

In the case where the multiplicity of E_0 is one, the range of k is singleton, and formulas (2.5), (2.11), and (2.12) become

$$(\psi|\mathscr{M}(B)[X]\psi)=\int_B(A(x)\psi|XA(x)\psi)\mu(dx), \tag{2.13}$$

$$\mu_\psi(dx)=\|A(x)\psi\|^2\mu(dx), \tag{2.14}$$

$$S_\psi(x)=\frac{|A(x)\psi)(A(x)\psi|}{\|A(x)\psi\|^2}, \tag{2.15}$$

where $A(x)$ is a family of operators from $\mathscr{D}$ to $\mathscr{H}$, satisfying

$$\int\|A(x)\psi\|^2\mu(dx)=\|\psi\|^2,\quad \psi\in\mathscr{D}.$$

In this case the pure initial state gives rise to pure posterior states, and the corresponding instrument may be called *pure* (see Ref. 11 for an information-theoretic characterization of pure instruments).

Note that if the measure μ is nonatomic, operators $A_k(x)$ in the representation (2.5) need not be bounded, nor even closable, as the following example shows.

Example 1: The idea behind this example goes back to Gordon and Louisell[13] (see also Ref. 5). Let $\mathscr{H}=L^2(\mathrm{R})$ and Q be the position observable in $\mathscr{H}$ given by the operator of multiplication by the independent variable $x\in\mathrm{R}$. Let ψ_x, $x\in\mathrm{R}$, be a measurable family of unit vectors in $\mathscr{H}$. We take $\mathscr{D}=C_0(\mathrm{R})$ (the subspace of continuous functions with compact support), and define

$$A(x)\psi=\Psi_x\cdot\psi(x),\quad \psi\in\mathscr{D}. \tag{2.16}$$

In Dirac's notations $A(x)=|\Psi_x)(x|$, where $(x|$ is the linear unbounded nonclosable form on $\mathscr{D}$ satisfying $(x|\psi)=\psi(x)$ for $\psi\in\mathscr{D}$. The relation (2.13) with $\mu(dx)=dx$ defines a pure c.p. instrument, whose probability distribution is

$$\mu_\psi(dx)=\|\psi(x)\|^2\,dx$$

and the posterior states are

$$S_\psi(x)=|\psi_x)(\psi_x|.$$

This instrument thus corresponds to precise measurement of the position, accompanied by the preparation of the vector state ψ_x, depending on the outcome x of the measurement.

III. COVARIANT INSTRUMENTS

Let G be a locally compact second countable group, and $g\to V_g$ a continuous unitary representation of G in $\mathscr{H}$, leaving invariant the algebra $\mathscr{A}$. An instrument $\mathscr{M}$ with the value space G is called *covariant* if

$$V_g\mathscr{M}(B)[V_g^*XV_g]V_g^*=\mathscr{M}(gB)[X] \tag{3.1}$$

for all $g\in G$, $X\in\mathscr{A}$, $B\in\mathscr{B}$, where $\mathscr{B}$ is the σ-algebra of Borel subsets of G. The definition (in a slightly more general form, allowing a transitive G-space as the value space of the instrument) was given by Davies (see Ref. 9), who studied covariant (not necessarily c.p.) instruments in the case of compact G and $\dim\mathscr{H}<\infty$.

The importance of the condition of covariance lies in the fact that it provides a correspondence principle between classical kinematical variables and the quantum instruments. Let $g\in G$ be

such a variable. Then the transformation $x \to gx$, where $x \in G$, describes a motion in a classical space G. The representation V_g gives the corresponding transformation of quantum states $\psi \to \psi_g = V_g \psi$. Equation (3.1) is equivalent to the following properties of the measurement probability distribution and the posterior states:

$$\mu_{\psi_g}(B) = \mu_\psi(g^{-1}B),$$

$$S_{\psi_g}(x) = V_g S_\psi(g^{-1}x) V_g^*,$$

expressing the principle of correspondence.

The following result, proved in Ref. 14, specifies the representation (2.1) in the case of covariant c.p. instruments.

Proposition 2: If $\mathscr{M}$ is a covariant c.p. instrument, then the representation (2.1) can be chosen in such a way that there exists a continuous unitary representation $g \to \widetilde{V}_g$ of G in $\mathscr{K}$, such that

$$\widetilde{V}_g V = V V_g, \tag{3.2}$$

$$\pi[V_g^* X V_g] = \widetilde{V}_g^* \pi[X] \widetilde{V}_g, \tag{3.3}$$

$$\widetilde{V}_g E(B) \widetilde{V}_g^* = E(gB), \tag{3.4}$$

for all $g \in G$, $X \in \mathscr{A}$, $B \in \mathscr{B}$. □

In the case $\mathscr{A} = B(\mathscr{H})$ this leads to the representation (2.3) with V satisfying (3.2) where

$$\widetilde{V}_g = V_g \otimes D_g,$$

and $g \to D_g$ is a continuous representation of G in $\mathscr{H}_0$. This fulfills the condition (3.3), while (3.2) and (3.4) take the forms

$$(V_g \otimes D_g) V = V V_g, \tag{3.5}$$

$$D_g E_0(B) D_g^* = E_0(gB). \tag{3.6}$$

The last equation means that (D, E_0) is an *imprimitivity system* in $\mathscr{H}_0$ (see, e.g., Ref. 15). By the Mackey imprimitivity theorem (D, E_0) is unitarily equivalent to the special system in the space $L_h^2(G)$ of square integrable functions $f(x)$; $x \in G$, with values in a separable Hilbert space h, defined as

$$D_g f(x) = f(g^{-1}x), \quad E_0(B) f(x) = I_B(x) f(x),$$

where I_B is the indicator of a set $B \in \mathscr{B}$. Let us fix an orthonormal basis $\{e_k\}$ in h. Then $\mathscr{K} = \mathscr{H} \otimes L_h^2(G)$ is represented as the space of functions $\psi_k(x)$; $k = 1, \ldots, \dim h$; $x \in G$, with values in $\mathscr{H}$, satisfying

$$\int_G \sum_k \|\psi_k(x)\|^2 d_l x < \infty,$$

where $d_l x$ is the left Haar measure on G. The *intertwining equation* (3.5) is then equivalent to

$$V_g (V V_g^* \psi)_k(x) = (V\psi)_k(gx) \tag{3.7}$$

for all g and almost all $x \in G$.

We conjecture that at least in the case of a type I group G there exists a dense domain $\mathscr{D} \subset \mathscr{H}$, invariant under the representation and such that the functions $x \to (V\psi)_k(x)$ are continuous for $\psi \in \mathscr{D}$, and therefore have unambiguously defined values for every $x \in G$. Then there is a sequence of linear operators A_k from $\mathscr{D}$ to $\mathscr{H}$ such that

$$A_k \psi = (V\psi)_k(e), \quad \psi \in \mathscr{D},$$

TABLE I. Canonical pairs in quantum mechanics.

G	R		T		Z_n
$\hat{G}$	R		Z		Z_n
Λ	R	$-R_+$	Z	$-Z_+$	Z_n
x	position	time	angle	phase	discretized phase
A	momentum P	$-$energy $-H$	angular momentum	$-$number $-N$	$-$truncated number

where e is the neutral element of G, and the equation (3.7) has the unique solution

$$(V\psi)_k(g)=V_gA_kV_g^*\psi.$$

If this is so, we have the representation of the type (2.5) with explicitly covariant $A_k(x)=V_xA_kV_x^*$, where $\mu(dx)$ is the Haar measure on G.

In what follows we give the proofs of this conjecture for two particularly important cases, where G is a one-dimensional Lie group (i.e., T or R). The proof for the case $G=\mathrm{R}$ can be generalized to an arbitrary locally compact Abelian group by using techniques developed in Ref. 15. However, the case of the elementary locally compact Abelian groups is particularly important for quantum mechanical applications as it gives a key for quantization of principal canonical pairs (see Table I). In this table $x\in G$ is a classical kinematical parameter and $x\to V_x=\mathrm{e}^{-ixA}$ is a unitary representation with the generator A. The spectrum Λ of A is a subset of the dual group $\hat{G}$. For the description of quantum observables canonically conjugated to A in terms of covariant probability operator-valued measures, including generalizations of the canonical commutation relation and of the uncertainty relation, see Ref. 16. The last column where Z_n is the cyclic group of order n corresponds to the approximate approach to quantum oscillator phase proposed by Pegg and Barnett.[17]

IV. THE CASE $G=\mathrm{T}$

In this Section we find the general form of a c.p. instrument, covariant with respect to one-dimensional rotations, describing measurements of cyclic quantities like angle, phase, or any other periodic coordinate. Let $G=[0,2\pi)$ with the operation of addition modulo 2π, denoted as $\dot{+}$, and let $\theta\to V_0$; $\theta\in[0,2\pi)$, be a continuous unitary representation of G in $\mathscr{H}$. The intertwining equation (3.7) reads

$$V_0(VV_0^*\psi)_k(\phi)=(V\psi)_k(\phi\dot{+}\theta),\quad \theta\in[0,2\pi). \tag{4.1}$$

By Stone's theorem, $V_0=\exp(-i\theta L)$, where L is a self-adjoint operator with purely point spectrum $\Lambda\subset G=Z$. Then there is an orthonormal basis $\{|l,j);l\in\Lambda\}$ in $\mathscr{H}$, consisting of eigenvectors of L:

$$L|l,j)=l|l,j),\quad l\in\Lambda.$$

Theorem 2: Let $\mathscr{D}=\mathrm{lin}\{|l,j);l\in\Lambda\}$. An arbitrary c.p. instrument, covariant with respect to the representation $\theta\to V_\theta$, has the form

$$(\psi|\mathscr{M}(d\theta)[X]\psi)=\sum_k\,(A_k(\theta)\psi|XA_k(\theta)\psi)\,\frac{d\theta}{2\pi},\quad \psi\in\mathscr{D}, \tag{4.2}$$

where $A_k(\theta)=V_\theta\,A_kV_\theta^*$ and A_k are linear operators from $\mathscr{D}$ to $\mathscr{H}$, satisfying

$$\sum_k\,(\psi_{l,j,k}|\psi_{l,j',k})=\delta_{j,j'},\quad l\in\Lambda, \tag{4.3a}$$

with $\psi_{l,j,k}=A_k|l,j)$.

Proof: To simplify notations we assume throughout that *the multiplicity of the spectrum is one* (this assumption can be easily relaxed). Then we can drop the index j and relation (4.3a) takes the form

$$\sum_k \|A_k|l)\|^2 = I; \quad l \in \Lambda. \tag{4.3b}$$

Apparently $\mathscr{D}$ is invariant under V_θ. Consider the function

$$c_{m,l,k}(\theta) = (m|(V|l))_k(\theta). \tag{4.4}$$

The equation (4.1) implies

$$c_{m,l,k}(\phi \dot{+} \theta) = \mathrm{e}^{i\theta(l-m)} c_{m,l,k}(\phi) \tag{4.5}$$

for all $\theta \in [0,2\pi)$ and almost all ϕ. The normalization condition (2.8) with $\mu(d\theta) = d\theta/2\pi$ implies

$$\int_0^{2\pi} \sum_k \sum_m |c_{m,l,k}(\theta)|^2 \frac{d\theta}{2\pi} = 1, \quad l \in \Lambda. \tag{4.6}$$

Denoting

$$c^\epsilon_{m,l,k}(\theta) = \epsilon^{-1} \int_0^\epsilon c_{m,l,k}(\phi \dot{+} \theta) d\phi,$$

we have that $c^\epsilon_{m,l,k}(\theta)$ is continuous for every $\epsilon > 0$. By integrating Eq. (4.5) we see that

$$c^\epsilon_{m,l,k}(\theta) = \mathrm{e}^{i\theta(l-m)} c^\epsilon_{m,l,k}(0).$$

Since $c^\epsilon_{m,l,k}(\theta)$ converges to $c_{m,l,k}(\theta)$ in the mean as $\epsilon \to 0$, we have

$$|c^\epsilon_{m,l,k}(0) - c^{\epsilon'}_{m,l,k}(0)| \leqslant \int_0^{2\pi} |c^\epsilon_{m,l,k}(\theta) - c^{\epsilon'}_{m,l,k}(\theta)| \frac{d\theta}{2\pi} \to 0$$

as $\epsilon, \epsilon' \to 0$. Thus there exist $c_{m,l,k} = \lim_{\epsilon \to 0} c^\epsilon_{m,l,k}(0)$. Therefore $c^\epsilon_{m,l,k}(\theta)$ converges pointwise to $\mathrm{e}^{i(l-m)\theta} c_{m,l,k}$ and hence

$$c_{m,l,k}(\theta) = \mathrm{e}^{i(l-m)\theta} c_{m,l,k}. \tag{4.7}$$

From (4.6)

$$\sum_k \sum_m |c_{m,l,k}|^2 = 1, \quad l \in \Lambda.$$

Defining

$$\psi_{l,k} = \sum_{m \in \Lambda} c_{m,l,k} |m),$$

we have

$$\sum_k \|\psi_{l,k}\|^2 = 1, \quad l \in \Lambda. \tag{4.8}$$

The operators

$$A_k = \sum_{l \in \Lambda} |\psi_{l,k})(l| \tag{4.9}$$

are then defined on $\mathscr{D}$ and satisfy (4.3b). Moreover, for $\psi\in\mathscr{D}$

$$A_k(\theta)\psi=V_\theta A_k V_\theta^*\psi=\sum_{m,l\in\Lambda} c_{m,l,k}(\theta)|m)(l|\psi)=(V\psi)_k(\theta)$$

by (4.4). This proves (4.2). Note that the functions $A_k(\theta)\psi$ are norm continuous in θ for $\psi\in\mathscr{D}$.□

Let us consider two important particular cases: the minimum uncertainty instruments, and the instruments corresponding to an approximate von Neumann-type measurement process. The probability distribution of the instrument (4.2) is

$$\mu_\psi(d\theta)=\sum_k \|A_k V_\theta\psi\|^2\,\frac{d\theta}{2\pi}.$$

Let us look for the instruments minimizing the angular uncertainty[10,16]

$$\Delta_\psi(n)=n^{-2}[|\hat{\mu}_\psi(n)|^{-2}-1],$$

where

$$\hat{\mu}_\psi(n)=\int_0^{2\pi} e^{i\theta n}\mu_\psi(d\theta)=\sum_l\sum_k (\psi|l)(l+n|\psi)(\psi_{l,k}|\psi_{l+n,k})$$

is the characteristic function of the measurement probability distribution. Assuming $(l|\psi)\geqslant 0$, which can always be achieved via a gauge transformation choosing appropriate phase factors for elements of the basis $|l)$, the maximum under the condition (4.8) is attained if $\psi_{l,k}\equiv\Psi_k$, whence

$$A_k(\theta)=\sqrt{2\pi}V_\theta|\Psi_k)(\theta|,$$

where $(\theta|$ are unbounded nonclosable linear forms on $\mathscr{D}$, algebraically conjugate to antilinear forms

$$|\theta)=\frac{1}{\sqrt{2\pi}}\sum_{l\in\Lambda} e^{-i\theta l}|l). \tag{4.10}$$

This leads to a kind of instrument considered in Example 1. Its probability distribution is

$$\mu_\psi(d\theta)=\left|\sum_{l\in\Lambda} e^{i\theta l}(l|\psi)\right|^2\frac{d\theta}{2\pi}=|(\theta|\psi)|^2\,d\theta=\mathrm{Tr}\,S_\psi M(d\theta),$$

where

$$M(d\theta)=|\theta)(\theta|\,d\theta \tag{4.11}$$

is the p.o.m. representing the canonical angle observable.[10,16] The relations below involving integration with respect to such p.o.m.'s should be understood in the form sense. Notice that

$$(\theta|\theta')=\frac{1}{2\pi}\sum_{l\in\Lambda} e^{il(\theta-\theta')},$$

therefore this p.o.m. is the usual spectral measure only if $\Lambda=Z$, when this is a periodic delta-function.

We now develop another useful representation for the operators $A_k(0)$. For any $m\in Z$ and $l\in\Lambda$ we define

$$a_{m,k}(l)=\begin{cases}(l-m|\psi_{l,k}), & \text{if } l-m\in\Lambda,\\ 0, & \text{otherwise.}\end{cases} \tag{4.12}$$

If we represent $A_k(0)$ as a matrix in the basis $|l)$, these are the elements standing on the intersection of lth column with mth diagonal, $m>0$ corresponding to upper diagonals. Then

$$\psi_{l,k}=\sum_m a_{m,k}(l)|l-m).$$

Relation (4.8) implies

$$\sum_k \sum_m |a_{m,k}(l)|^2=1, \quad l\in\Lambda. \tag{4.13}$$

For any $m\in Z$ we define the partial isometry

$$P_m=\sum_{l:l\in\Lambda,l-m\in\Lambda} |l-m)(l|=\int_0^{2\pi} \mathrm{e}^{-i\theta m}M(d\theta), \tag{4.14}$$

satisfying the generalized canonical commutation relation (CCR):

$$V_\theta P_m=\mathrm{e}^{i\theta m}P_m V_\theta.$$

Then $\psi_{l,k}=\Sigma_m a_{m,k}(l)P_m|l)$ and

$$A_k(\theta)=\sum_m \mathrm{e}^{i\theta m}P_m a_{m,k}(L), \tag{4.15}$$

where $a_{m,k}(L)=\Sigma_{l\in\Lambda}a_{m,k}(l)|l)(l|$ is an essentially normal operator on the domain $\mathscr{D}$. The relation (4.15) gives the desired alternative form of the operators $A_k(\theta)$. It corresponds to special ordering when the action of L precedes the action of the canonically conjugated observable.

Example 2: Assume $\Lambda=Z$ (this is the case for measurements of the angle of rotation with infinite spin, neglecting inessential complications due to possible degeneracy of spectrum). In this case isometries (4.14) form the unitary group

$$P_m=U^m, \quad m\in \mathrm{Z}, \tag{4.16}$$

where $U=\Sigma_l|l-1)(l|$. By Stone's theorem,

$$U=\exp(-i\Theta), \tag{4.17}$$

where Θ is the self-adjoint operator

$$\Theta=\int_0^{2\pi}\theta M(d\theta), \tag{4.18}$$

with the spectral measure $M(d\theta)$ given by (4.11). The functions $A_k(\theta)$ in the representation (4.2) of the covariant instrument take the form

$$A_k(\theta)=\sum_m \mathrm{e}^{i\theta m}U^m a_{m,k}(L)=\sum_m \mathrm{e}^{i(\theta-\Theta)m}a_{m,k}(L). \tag{4.19}$$

Let us consider the particular case where $a_{m,k}(l)$ *do not depend on* l:

$$a_{m,k}(l)=a_{m,k}. \tag{4.20}$$

Then

$$A_k(\theta)=\tilde{a}_k(\theta-\Theta), \tag{4.21}$$

where

$$\tilde{a}_k(\theta)=\sum_m \mathrm{e}^{i\theta m}a_{m,k} \tag{4.22}$$

are the functions satisfying

$$\int_0^{2\pi}\sum_k |\tilde{a}_k(\theta)|^2\,\frac{d\theta}{2\pi}=1. \tag{4.23}$$

This structure of $A_k(\theta)$ is characteristic for approximate measurement of Θ, admitting a von Neumann type of realization described below (see Refs. 9 and 11 for the case $\theta\in\mathrm{R}$). Let $\mathscr{H}_0$ be a copy of the Hilbert space $\mathscr{H}$ spanned by the basis $\{|l)_0; l=0,1,...\}$. In $\mathscr{H}_0$ we can define the operator L_0 and the spectral measure $M_0(d\theta)=|\theta)_{0\ 0}(\theta|d\theta$ in a way similar to L and $M(d\theta)$. The space $\mathscr{H}_0$ will describe the measurement apparatus, the initial state of which is

$$S_0=\sum_k |a_k)_{0\ 0}(a_k|, \tag{4.24}$$

where

$$|a_k)_0=\sum_m a_{m,k}|m)_0. \tag{4.25}$$

Interaction between the system and the apparatus is given by the unitary operator

$$\mathscr{U}=\exp(-i\Theta\otimes L_0)=\sum_m U^m\otimes|m)_{0\ 0}(m|, \tag{4.26}$$

and the measured observable of the apparatus is

$$\Theta_0=\int_0^{2\pi}\theta M\ (d\theta).$$

From (4.19), (4.18), (4.25), and (4.26) it follows that

$$A_k(\theta)={}_0(\theta|\mathscr{U}|a_k)_0, \tag{4.27}$$

hence the described measurement process is a realization of the covariant instrument $\mathscr{M}(d\theta)$:

$$\sum_k (A_k(\theta)\psi|XA_k(\theta)\psi)\,\frac{d\theta}{2\pi}=\sum_k (\psi|\otimes{}_0(a_k|\mathscr{U}^*(X\otimes M_0(d\theta))\mathscr{U}|a_k)_0\otimes|\psi), \quad \psi\in\mathscr{D}. \tag{4.28}$$

Example 3: Let us try to apply a similar argument to the measurement of the phase of a harmonic oscillator. In this case $V_\phi=\mathrm{e}^{i\phi N}$, where N is the number operator. Thus $N=-L$ and $\Lambda=\{0,-1,...\}$. We have

$$P_m=\begin{cases}P^{*m}, & m=0,1,...,\\ P^{-m}, & m=-1,...,\end{cases}$$

where

$$P=\sum_{l=-\infty}^{-1}|l+1)(l|, \quad P^*=\sum_{l=-\infty}^{0}|l-1)(l|.$$

Here $|l)$ is the eigenvector of N such that $N|l)=-l|l)$ (in usual terms, it is the n-quanta vector with $n=-l$). Operator P^* is isometric, while P is co-isometric. Defining the p.o.m. $M(d\theta)$ by the relation (4.11), we have

$$\left.\begin{matrix} P \\ P^* \end{matrix}\right\} = \int_0^{2\pi} \mathrm{e}^{\pm i\theta} M(d\theta)$$

(see Refs. 10 and 16 for more detail).

A covariant instrument, corresponding to measurement of phase, has the form (4.2) with

$$A_k(\theta) = \sum_{m=0}^{\infty} \mathrm{e}^{im\theta} P^{*m} a_{m,k}(-N) + \sum_{m=-\infty}^{-1} \mathrm{e}^{im\theta} P^{-m} a_{m,k}(-N).$$

The ansatz (4.20) is compatible with the condition $a_{m,k}(l)=0$ for $l-m>0$ only if $a_{m,k}=0$ for $m<0$. Assuming this, we have

$$A_k(\theta) = \sum_{m=0}^{\infty} \mathrm{e}^{im\theta} P^{*m} a_{m,k} = \int_0^{2\pi} \tilde{a}_k(\theta-\theta') M(d\theta'), \tag{4.29}$$

where $\tilde{a}_k(\theta)$ are given by (4.22) with m running from 0 to ∞.

Let $\mathscr{H}_0$ be the Hilbert space spanned by the basis $\{|m)_0; m=0,1,...\}$. In this space we define the p.o.m. $M_0(d\theta)=|\theta)_{0\ 0}(\theta|d\theta$, where $|\theta)_0$ are determined by the relation (4.10) with l running from 0 to ∞. Defining also the isometric operator

$$\mathscr{U} = \sum_{m=0}^{\infty} P^{*m} \otimes |m)_{0\ 0}(m|,$$

in $\mathscr{H}\otimes\mathscr{H}_0$, we have the relation (4.27) for operators (4.29), and hence an analog of the realization (4.28) for the phase; however, this realization is not quite satisfactory because $\mathscr{U}$ is not unitary and $M_0(d\theta)$ is not a spectral measure.

If the spectrum Λ of L is bounded (e.g., in the case of measurement of the angle of rotation for a system with finite spin), the ansatz (4.20) is not compatible with the requirement $a_{m,k}(l)=0$ for $l-m\notin\Lambda$, and the construction of Example 2 cannot be directly transferred to this case.

V. THE CASE $G=\mathbf{R}$

Here we describe instruments, covariant with respect to one-dimensional shifts (which can be shifts in position, momentum, time, etc.) For definiteness, we speak of ''position''; at the end we shall make specific comments concerning time. In that case $G=\mathrm{R}$ and $x\to V_x$, $x\in\mathrm{R}$, is a continuous unitary representation of R in $\mathscr{H}$. The intertwining equation (3.7) for the isometry $V:\mathscr{H}\to\mathscr{H}\otimes L^2_h(\mathrm{R})$ reads

$$V_x(VV_x^*\psi)_k(\xi) = (V\psi)_k(\xi+x), \quad x\in\mathrm{R}. \tag{5.1}$$

By Stone's theorem, $V_x=\mathrm{e}^{-ixP}$, where P is a self-adjoint operator with the spectrum $\Lambda\subset\mathrm{R}$. In what follows we restrict ourselves to the case where P has a *purely Lebesgue spectrum with multiplicity one*. The restriction to the Lebesgue spectrum is necessary, since otherwise a covariant p.o.m. representing ''position'' observable does not exist (see Ref. 16); generalization to arbitrary multiplicity can be made easily with the help of techniques developed in Ref. 15. Under this assumption P is represented as multiplication by p in the space $\mathscr{H}=L^2(\Lambda)$ with generalized eigenvectors $|p)$, which are antilinear forms algebraically conjugate to the linear forms $(p|$, defined as $(p|\psi)=\psi(p)$ on the dense domain

$$\mathscr{D} = \{\psi\in\mathscr{H}: \text{the function } p\to(p|\psi) \text{is in } \mathrm{C}_0(\Lambda)\}. \tag{5.2}$$

Defining the family of partial isometries

$$U_y = \int_{\Lambda\cap(\Lambda+y)} |p-y)(p|\, dp, \quad y\in\mathrm{R}, \tag{5.3}$$

one can check the relation

$$V_x U_y = \mathrm{e}^{ixy} U_y V_x, \tag{5.4}$$

which has the form of the Weyl CCR.

By introducing the p.o.m.

$$M(dx) = |x)(x|\, dx,$$

where

$$|x) = \frac{l}{\sqrt{2\pi}} \int_\Lambda \mathrm{e}^{-ixp} |p)\, dp$$

are nonclosable antilinear forms defined on $\mathscr{D}$, we have

$$U_y = \int \mathrm{e}^{iyx} M(dx).$$

An operator with the domain $\mathscr{D}$ will be called a *kernel operator* if there is a measurable function $A(p,p')$ such that

$$\int_\Lambda \int_K |A(p,p')|^2\, dp\, dp < \infty$$

for any compact $K \subset \Lambda$ and

$$(p|A\psi) = \int A(p,p')(p'|\psi) dp'$$

for $\psi \in \mathscr{D}$. We denote

$$(p|\Lambda|p') \triangleq A(p,p')$$

and call it the *kernel* of the operator A.

Theorem 3: An arbitrary c.p. instrument, covariant with respect to the representation $x \to V_x$, has the form

$$(\psi|\mathscr{M}(dx)[X]\psi) = \sum_k (A_k(x)\psi|XA_k(x)\psi)\frac{dx}{2\pi}, \quad \psi \in \mathscr{D}, \tag{5.5}$$

where $A_k(x) = V_x A_k V_x^*$ and A_k are kernel operators satisfying

$$\int_\Lambda \sum_k |(p|A_k|p')|^2\, dp = l, \quad p' \in \Lambda. \tag{5.6}$$

Proof: In solving the equation (5.1), we shall use the method which is a modification of the one developed in Ref. 18 for characterization of covariant quantum Markovian evolutions.

Let B be a bounded Borel subset of R. Then the operator $\Delta_{B,k}: \mathscr{H} \to \mathscr{H} \otimes L_h^2(\mathrm{R})$, defined as

$$(\Delta_{B,k}\psi)_j(x) = \delta_{jk} \int_B \mathrm{e}^{ixy} U_y \psi\, dy,$$

is bounded and satisfies the same equation (5.1) as the operator V. This follows from the CCR (5.4). Hence the bounded operator $\Delta_{B,k}^* V: \mathscr{H} \to \mathscr{H}$ commutes with the representation $x \to V_x$, and therefore there exists a bounded Borel function $p \to A_{B,k}(p)$, $p \in \Lambda$, such that

$$\Delta_{B,k}^* V \psi = A_{B,k}(P)\psi, \quad \psi \in \mathscr{H}.$$

By writing this in P-representation, we have

$$A_{B,k}(p)(p|\psi)=\frac{1}{2\pi}\int_B\left\{\int \mathrm{c}^{-ixy}(p|U_y^*(V\psi)_k(x))dx\right\}dy.$$

This means that for all k and almost all $p\in\Lambda$, $B\to A_{B,k}(p)$ is a σ-finite complex measure, absolutely continuous with respect to the Lebesgue measure dy. Denoting the Radon–Nikodym derivative by $a_{y,k}(p)$, we have

$$a_{y,k}(p)(p|\psi)=\frac{1}{2\pi}\int \mathrm{e}^{-ixy}(p|U_y^*(V\psi)_k(x))dx=\frac{1}{2\pi}\int \mathrm{e}^{-ixy}(p-y|(V\psi)_k(x))\,dx, \quad (5.7)$$

if $p-y\in\Lambda$, and zero if $p-y\notin\Lambda$. It follows that

$$a_{y,k}(p+y)(p+y|\psi)=\frac{1}{2\pi}\int \mathrm{e}^{-ixy}(p|(V\psi)_k(x))\,dx, \quad (5.8)$$

if $p+y\in\Lambda$. Let us show that the right-hand side is equal to zero if $p+y\notin\Lambda$. Fixing such p and y, consider the function

$$f(x')=\frac{1}{2\pi}\int \mathrm{e}^{-ixy}(p|(VV_{x'}^*\psi)_k(x))\,dx.$$

By using Eq. (5.1), we see that $f(x')=\mathrm{e}^{ix'(p+y)}f(0)$, and we have to show that $f(0)=0$. Let us take

$$g(x)=\int_{\mathrm{R}\setminus\Lambda}\mathrm{e}^{ixp}\tilde{g}(p)dp,$$

where $\tilde{g}$ is an arbitrary C^∞-function with compact support in $\mathrm{R}\setminus\Lambda$. Then

$$\int \overline{g(x')}V_{x'}^*\psi\,dx'=0$$

and hence $\int\overline{g(x')}f(x')dx'=0$. However, this means that $\overline{\tilde{g}(p+y)}f(0)=0$, whence $f(0)=0$.

Since $V:\mathscr{H}\to\mathscr{H}\otimes L_h^2(R)$ is isometric,

$$\frac{1}{2\pi}\int\int_\Lambda\sum_k|(p|(V\psi)_k(x))|^2\,dp\,dx=\|\psi\|^2$$

[we take $dg=dx/2\pi$ in $L_h^2(\mathrm{R})$]. Hence by isometry of the Fourier transform (5.8),

$$\int\int_{\Lambda-y}\sum_k|a_{y,k}(p+y)|^2|(p+y|\psi)|^2\,dp\,dy=\|\psi\|^2.$$

By changing variables y,p to $y,p'=p+y$, we deduce

$$\int\sum_k|a_{y,k}(p)|^2\,dy=1,\quad p\in\Lambda. \quad (5.9)$$

It follows that the operators

$$a_{y,k}(P)=\int a_{y,k}(p)|p)(p|\,dp$$

are defined on $\mathscr{D}$. The relation (5.8) can then be written as

$$(p|U_y a_{y,k}(P)\psi)=\frac{1}{2\pi}\int \mathrm{e}^{-ixy}(p|(V\psi)_k(x))\,dx$$

for $\psi\in\mathscr{D}$. By inverting the Fourier transform, we obtain

$$(p|(V\psi)_k(x))=\int \mathrm{e}^{ixy}(p|U_y a_{y,k}(P)\psi)\,dy. \tag{5.10}$$

Let us show that for every $x\in\mathrm{R}$ the right-hand side defines the linear operator $A_k(x):\mathscr{D}\to\mathscr{H}$,

$$A_k(x)\psi=\int \mathrm{e}^{ixy}U_y a_{y,k}(P)\psi\,dy, \tag{5.11}$$

where the integral converges strongly for every $\psi\in\mathscr{D}$. This follows from the estimate

$$\begin{aligned}\|A_k(x)\psi\|^2&=\int\left|\int \mathrm{e}^{ixy}(p|U_y a_{y,k}(P)\psi)dy\right|^2 dp\\&=\int\left|\int \mathrm{e}^{ixy}a_{y,k}(p+y)\psi(p+y)dy\right|^2 dp\\&\leqslant\int\left\{\int |a_{y,k}(p+y)1_K(p+y)|^2\,dy\cdot\int|\psi(p+y)|^2\,dy\right\}dp\\&\leqslant\|\psi\|^2\int\int|a_{y,k}(p')|^2 1_K(p')dy\,dp'=\|\psi\|^2\mathrm{mes}\,K,\end{aligned}$$

where K is the support of ψ. A similar estimate shows that the function $x\to A_k(x)\psi$ is norm continuous.

The relation (5.11) is a continuous analog of (4.15). By CCR (5.4), $A_k(x)=V_xA_k(0)V_x^*$. From (5.11) it follows

$$(p|A_k(0)\psi)=\int_{\Lambda-y}a_{y,k}(p+y)(p+y|\psi)\,dy=\int_\Lambda a_{p'-p,k}(p')(p'|\psi)dp'.$$

Hence $A_k(0)$ is the kernel operator with the kernel

$$(p|A_k(0)|p')=a_{p'-p,k}(p').$$

Conversely,

$$a_{y,k}(p)=\begin{cases}(p-y|A_k(0)|p), & p-y\in\Lambda,\\ 0, & p-y\notin\Lambda.\end{cases} \tag{5.12}$$

The normalization condition (5.6) follows from (5.9). □

As in the case $G=\mathrm{T}$, there are two interesting particular cases corresponding to the minimum uncertainty and to von Neumann-type approximate measurements. The probability distribution of the instrument (5.5) is

$$\mu_\psi(dx)=\sum_k\|A_kV_x\psi\|^2\frac{dx}{2\pi}.$$

The variance of this probability distribution can be expressed in terms of the characteristic function as

$$\mathsf{D}_\psi=-\frac{1}{2}\frac{d^2}{dy^2}\bigg|_{y=0}\left|\int \mathrm{e}^{ixy}\mu_\psi(dx)\right|^2. \tag{5.13}$$

The modulus of the characteristic function is equal to

$$\left|\int\sum_k(\psi|p)(p+y|\psi)(\psi_{p,k}|\psi_{p+y,k})dp\right|,$$

where

$$\psi_{p,k} = \int |p')(p'|A_k(0)|p)dp',$$

and is uniformly (with respect to y) maximized by the choice $\psi_{p,k} \equiv \psi_k$ [provided $(p|\psi) \geqslant 0$]. It follows that this choice minimizes the variance (5.13). This gives

$$A_k(x) = \sqrt{2\pi} V_x |\psi_k)(x|,$$

which is a generalization of Example 1. From the inversion formula (5.12), $a_y(p) = (p-y|\Psi_0)$ in this case.

If $\Lambda = \mathrm{R}$, then $U_y = \mathrm{e}^{-iyQ}$, where $Q = \int xM(dx)$ is the self-adjoint position operator with the generalized eigenvectors $|x)$. This case is a continuous analog of the Example 2. In the case where $a_{y,k}(p) = a_{y,k}$ does not depend on p, we have $A_k(x) = \widetilde{a}_k(x-Q)$, where $\widetilde{a}_k(x) = \int \mathrm{e}^{ixy} a_{y,k}\, dy$. The instrument (5.5) then describes approximate measurement of Q admiting von Neumann-type realization, described in Ref. 11.

The case of the representation $t \to V_t = \mathrm{e}^{itH}$, $t \in \mathrm{R}$, where the self-adjoint operator $L = -H$ has nonpositive Lebesgue spectrum $\Lambda = (-\infty, 0)$, is an apparent continuous analog of Example 3. The instruments covariant with respect to this representation correspond to measurements of ''arrival time.'' The construction of Example 3 can be transformed to this case by using the description of the time observable given in Refs. 10 and 16, and by taking into account the multiplicity of the representation V_t.

ACKNOWLEDGMENTS

The author is grateful to M. Ozawa (Nagoya University), O. Hirota (Tamagawa University), O. Melsheimer (University of Marburg), and L. Lanz (University of Milan) for their hospitality and stimulating discussions. The work was partially supported by RFBR grant No. 96-01-01709, DAAD, and INFN.

[1] E. B. Davies and J. T. Lewis, Commun. Math. Phys. **17**, 239 (1970).
[2] M. Ozawa, J. Math. Phys. **25**, 79 (1984).
[3] M. D. Srinivas, J. Math. Phys. **18**, 2138 (1977); M. D. Srinivas and E. B. Davies, Opt. Acta **28**, 981 (1981); A. S. Holevo, Sov. Math. **26**, 1 (1982).
[4] H. P. Yuen, Phys. Rev. Lett. **51**, 719 (1983); M. Ozawa, Phys. Rev. Lett. **60**, 383 (1988); M. Ozawa, in *Squeezed and Nonclassical Light*, edited by P. Tombesi and E. R. Pike (Plenum, New York, 1989), p. 263.
[5] A. Barchielli, L. Lanz, and G. M. Prosperi, Found. Phys. **13**, 779 (1983); A. Barchielli and G.-C. Lupieri, J. Math. Phys. **26**, 2222 (1985).
[6] A. S. Holevo, Proc. ICM-86, Berkeley, Calif. USA, 1011 (1987); A. S. Holevo, in *Quantum Aspects of Optical Communications*, edited by C. Benjaballah, O. Hirota, and S. Reynaud, Lect. Notes Phys. Vol. 378 (Springer-Verlag, Berlin-Heidelberg, 1991), p. 127.
[7] *Foundations of Quantum Mechanics and Ordered Linear Spaces* edited by A. Hartkämper and H. Neumann, Lect. Notes Phys. Vol. **29** (Springer-Verlag, Berlin, 1974); P. Busch, P. J. Lahti, and P. Mittelstaedt, *The Quantum Theory of Measurement* (Springer-Verlag, Berlin, 1991).
[8] R. Haag and D. Kastler, J. Math. Phys. **5**, 848 (1964).
[9] E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[10] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North Holland, Amsterdam, 1982).
[11] M. Ozawa, J. Math. Phys. **34**, 5596 (1993); **26**, 1948 (1985); **27**, 759 (1986).
[12] M. Ozawa, in *Quantum Communication and Measurement*, edited by V. P. Belavkin, O. Hirota, and R. L. Hudson (Plenum, New York, 1995), p. 109.
[13] J.-P. Gordon and W. Louisell, in *Physics of Quantum Electronics*, edited by P. L. Kelley, B. Lax, and P. E. Tannenwald (McGraw–Hill, New York, 1966), p. 833.
[14] L. V. Denisov, ''On the Stinespring-type theorem for covariant instruments in noncommutative probability theory,'' Preprint No. 20, Steklov Mathematical Institute (1990).
[15] A. S. Holevo, Sov. Math. (Iz. VUZ) **27**, 53 (1983); Lect. Notes Math. **1396**, 229 (1984); Math. USSR Izvestiya **28**, 175 (1987).
[16] A. S. Holevo, Rep. Math. Phys. **13**, 379 (1978); **16**, 289 (1980); Lect. Notes Math. **1021**, 243 (1983).
[17] D. T. Pegg and S. M. Barnett, Phys. Rev. A **39**, 1665 (1989).
[18] A. S. Holevo, Izv. Math. **59**, 427 (1995).

p-adic stochastic hidden variable model

Andrew Khrennikov[a)]
Department of Mathematics, Statistics and Computer Sciences, University of Växjö, 35195, Växjö, Sweden

(Received 28 February 1996; accepted for publication 5 November 1997)

We propose stochastic hidden variables model in which hidden variables have a p-adic probability distribution $\rho(\lambda)$ and at the same time conditional probabilistic distributions $\mathbf{P}(U,\lambda)$, $U=A,A',B,B'$, are ordinary probabilities defined on the basis of the Kolmogorov measure-theoretical axiomatics. A frequency definition of p-adic probability is quite similar to the ordinary frequency definition of probability. p-adic frequency probability is defined as the limit of relative frequencies ν_n but in the p-adic metric. We study a model with p-adic stochastics on the level of the hidden variables description. But, of course, responses of macroapparatuses have to be described by ordinary stochastics. Thus our model describes a mixture of p-adic stochastics of the microworld and ordinary stochastics of macroapparatuses. In this model probabilities for physical observables are the ordinary probabilities. At the same time Bell's inequality is violated. © *1998 American Institute of Physics.* [S0022-2488(98)02802-3]

I. INTRODUCTION

Bell's inequality[1] and its numerous generalizations[2–5] provide very important predictions about properties of physical reality. It is quite surprising that these predictions have been obtained on the basis of a simple mathematical assumption. In fact, to prove Bell's inequality, we use only the assumption that *there exists a probability distribution[2] of hidden variables (HVs)*. (Of course, it is assumed that this probability distribution is defined in Kolmogorov's measure-theoretical framework.[4]) As a consequence we obtain, in particular, predictions about locality of the space–time.[5,6] It seems that it should be impossible to obtain such predictions, starting with this restricted initial information about a physical system.

It should be more natural to assume that information about the probability distribution of HVs can only give us new information about this probability distribution. In particular, all experiments for testing Bell's inequality can only give us information about the probability distribution of HVs. Thus, the violations of Bell's-type inequalities can only show us that the statement

HVs have a probability distribution described by usual theory of probability (i.e., by the system of Kolmogorov's axioms[6])

was not true. This idea is not new. There were many attempts to use this approach to understand the real consequences of the violations of Bell's inequality.[7–11] de Broglie[12] was one of the first who discussed the idea that ''classical'' and ''quantum'' probabilities had essentially different properties. Lochak[7,9] applied these considerations to the Einstein–Podolsky–Rosen[13] (EPR) paradox.

From the mathematical point of view, ''quantum'' behavior of probabilities implies that Kolmogorov's axiomatics must be changed to describe quantum phenomena. Kolmogorov's axiomatics contain the following axioms: *(1) probability yields values in the segment [0,1] of the field of real numbers* **R**; *(2) is a finite additive measure normalized by 1*. To use Lebesgue's integration theory, Kolmogorov proposed a stronger axiom: (2σ) *probability is a σ-additive measure*. In fact, Kolmogorov's axiomatics contain Bayes' formula for conditional probabilities as the additional axiom. [Kolmogorov motivated axioms (1) and (2) by using the properties of relative frequencies; see also the frequency probability theory of von Mises.[14] Bayes' formula can be motivated in the same way. On the other hand, σ additivity of probability has no frequency motivations. This is just

[a)]On leave from the Moscow Institute of Electronic Engineering.

 © 1998 American Institute of Physics

the technical mathematical condition; see the corresponding discussion in Kolmogorov's book.[4]]

On the one hand, Kolmogorov's axiomatics strongly stimulate the development of probability theory, because it provides the rigorous mathematical foundations for this theory. On the other hand, this axiomatics essentially restrict the domain of probability theory, because it reduces all probability theories to the formalism of real-valued measures.

As we have already noticed, quantum theory implies that we need to change the probabilistic formalism. There are two possibilities to do this. The first possibility is to change formally one of Kolmogorov's axioms, i.e., to consider a theory in which, for example, probability is nonadditive, or additive, but not σ additive, or Bayes' formula can be violated or probability can yield values that do not belong to the segment $[0,1]$ of the field of real numbers. Another possibility is to start with one of the old probability formalisms (which has been used before Kolmogorov's axiomatics) and develop a new probability formalism. In this paper we will realize the latter possibility.

The idea of the formal change of Kolmogorov's axiomatics for solving the problem of quantum probabilities has been used by many authors. Feynman[15] proposed that the behavior of probabilities be described in the two slit experiment by using nonadditive probabilities. Accardi[8,9] investigated the same problem (and the EPR paradox) by using a probability formalism that does not contain Bayes' formula.

In fact, the same conclusion (i.e., Bayes' formula is the root of the EPR paradox) has been obtained by Bub.[16] He provided a careful analysis of probabilistic roots of the EPR paradox. His conclusion was that violations of Bell's inequality was not related in any way to locality of the space–time. These violations show that the standard probability formalism is violated in quantum theory. There by some (still unclear reasons) we cannot use the standard rule $\mathbf{P}(A_2)\mathbf{P}(A_1/A_2) = \mathbf{P}(A_1 \cap A_2)$ in quantum theory. For the compatible observables, x_j, $j=1,2$, i.e., $A_j = \{x_j \in S_j\}$, where S_j are the Borel subsets of R, this statement is the standard consequence of the quantum formalism. This is just because the probability $\mathbf{P}(A_1 \cap A_2)$ does not exist. However, Bub has shown that Bayes' formula could not be used for some compatible observables x_j, $j=1,2$. Here the probability $\mathbf{P}(A_1 \cap A_2)$ is well defined. Such compatible observables appear in the EPR experiments. This is the main probabilistic implication of the EPR experiment. Of course, Bub's investigation was just a chain of physical considerations. As we remarked, the idea that we have to consider a new (non-Kolmogorov) probability axiomatics for the description of this physical phenomenon has been proposed by Accardi.[8,9]

Dirac[17] has introduced negative probability distributions for solving a problem of the relativistic quantization of photons. Then negative probability distributions have been used in connection with the EPR paradox (see Refs. 10 and 11). But, as negative probabilities have not any reasonable interpretation, this solution of the EPR paradox has not been considered as a physical solution. (See, for example, Weizsäcker:[11] ''An indefinite measure on the Hilbert space has been considered frequently and is used concretely e.g., in quantum electrodynamics by the method of Bleuler and Gupta; also by Heisenberg's spinor-field theory. But since it has turned out impossible, to conceive negative probability in the visualizable sense, the indefinite measure was no longer understood as probability. If probability is defined, e.g., as the prediction of a relative frequency, clearly it has to be non-negative.'') In this paper we will provide a fruitful frequency interpretation of negative probabilities. Finally, we refer to papers of Pitowsky[18] in which nonmeasurable ''random variables'' are used for the description of the EPR experiment (see also Ref. 19).

Relative frequencies provide the unique connection between theoretical probabilities and the physical reality. We will also approximate negative (and other extended) probabilities by relative frequencies. The basis of such a theory is so called p-adic *frequency* theory of probability[20,21] (see Sec. II about fields of p-adic numbers $\mathbf{Q}_p$ and p-adic probabilities). From the first point of view these are quite exotic probabilities. But their definition is similar to the standard definition of frequency probabilities. It is evident that relative frequencies $\nu_n = n/N$ always belong to the field of rational numbers Q. Therefore their asymptotic can be studied, not only with respect to a real metric $\rho_\infty(x,y)$ on Q, but also with respect to other metrics $\rho(x,y)$ on Q, and, in particular, with respect to the p-adic metric $\rho_p(x,y)$. The limit with respect to a metric $\rho(x,y)$ (if it exists) is said to be a ρ probability. In particular, we can get probabilities belonging to $\mathbf{Q}_p$. Moreover, every p-adic number can be realized as such a frequency probability. As Q is a subset of $\mathbf{Q}_p$, then, in principle, every rational number can be a p-adic probability. Thus negative probability distribu-

tions can be interpreted as frequency objects. This frequency probability theory is a natural generalization of the theory of von Mises.[14]

p-adic probabilities (and, in particular, negative probabilities) can be also obtained by the extension of the standard proportional definition of probability. Let Ω be a finite population. Then for any subset (event) $A \subset \Omega$, probability is defined by

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}, \tag{1}$$

where for the set B, its cardinality is denoted by $|B|$. The measure-theoretical approach is in some sense a generalization of the proportional theory. However, the formula (1) has not been preserved there. p-adic numbers give a possibility to extend the proportional definition of probability to infinite populations Ω and at the same time to preserve the basic formula (1). However, we have to pay for this by p-adic and, in particular, negative probabilities.

Deterministic HV models with p-adic probability distributions have been presented in Ref. 22. In such models Bell's inequality can be violated.

However, the results of quantum measurements are described by usual statistics (based on Kolmogorov's axiomatics). How can we connect the p-adic probability description of HVs and the ordinary probability description of quantum measurements? The present paper is devoted to this problem. [Following Bohm, we understand that ''the so-called 'observables' are not properties belonging to the observed system alone, but instead potentialities whose precise development depends just as much on the observing apparatus as on the observed system.'' p-adic statistics of the (microworld) HV distribution is perturbated by the usual statistics of apparatus.]

The distribution $\rho(\lambda)$ of HVs can be a p-adic distribution corresponding to the properties of the (microworld) internal structure of a system, but responses of the macroequipment can have the ordinary Kolmogorov distributions $\mathbf{P}(A,\lambda)$ for every value λ of HV. These HV models have the mixed (real-p-adic) probability structure (mixed stochastic HV models). We will construct a mixed stochastic HV model in which a combination of p-adic stochastics of HVs and ordinary stochastics of apparatuses gives ordinary probabilities for all physical observables A, A', B, B' and their physical combinations AB, AB', $A'B$, $A'B'$, but Bell's-type inequalities are violated.

II. p-ADIC PROBABILITIES

The field of real numbers R is constructed as a completion of the field of rational numbers Q with respect to the metric $\rho(x,y)=|x-y|$, where $|\cdot|$ is the usual absolute value, norm. Fields of p-adic numbers $\mathbf{Q}_p$ are constructed in the same way. There is an infinite sequence of p-adic number fields; there is its own field for every prime number $p=2,3,5,\ldots$. A p-adic norm $|\cdot|_p$ is defined in the following way. At first, we define it for natural numbers. Every natural number n can be represented as the product of powers of prime numbers: $n=2^{r_2}3^{r_3}\cdots p^{r_p}\cdots$. Then $|n|_p = p^{-r_p}$, by the definition $|0|_p=0$, $|-n|_p=|n|_p$, and $|n/m|_p=|n|_p/|m|_p$. The completion of Q with respect to the metric $\rho_p(x,y)=|x-y|_p$ is a locally compact field $\mathbf{Q}_p$.

And it is an intrinsic fact of the theory of numbers[23] that there is the only possibility to introduce a norm on Q to use a real one $|\cdot|$ or one of p-adic valuations.

For any $x \in \mathbf{Q}_p$ we have a unique canonical expansion (converging in the $|\cdot|_p$ norm) of the form

$$x = a_{-s}/p^s + \cdots a_0 + \cdots + a_k p^k + \cdots = \cdots a_k \cdots a_0, a_{-1} \cdots a_{-s}, \tag{2}$$

where $a_j = 0,1,\ldots,p-1$, are the ''digits'' of the p-adic expansion. It is sufficiently natural (see Ref. 24) to interpret p-adic numbers x that have the infinite number of nonzero digits as infinitely large quantities.

As usual in metric space, we introduce balls in $\mathbf{Q}_p$: $U_r(a)=\{x \in \mathbf{Q}_p : |x-a|_p \leqslant r\}$, $r=p^n$, $n=0,\pm 1,\pm 2,\ldots$. These sets are closed and open at the same time (''clopen''). The unit ball $U_1(0)$ is denoted by $\mathbf{Z}_p$. It plays an important role in p-adic analysis. This is a ring. Elements of $\mathbf{Z}_p$ are called p-adic integers, because canonical expansions of elements of $\mathbf{Z}_p$ do not contain negative powers of p:

$$x=a_0+a_1p+\cdots+a_np^n+\cdots=\cdots a_n\cdots a_1a_0.$$

Elements that have infinite number of nonzero digits can be interpreted as infinitely large natural numbers.

As is well known, the frequency definition of probability proposed by von Mises in 1919 played an important role in the construction of the foundations of modern probability theory. This definition exerted a strong influence on the measure-theoretical axiomatic of Kolmogorov,[4] 1933. As the motivation of his axioms, Kolmogorov used the properties of limits of relative frequencies (see Ref. 4). We are interested in the manner in which one of Kolmogorov's axioms arose; in accordance with this axiom, the probability of any event E is a non-negative real number. In Ref. 4, Kolmogorov considers von Mises' definition[14] of probability as the limit of relative frequencies of occurrence of the event E. Further, since the relative frequencies are rational numbers that lie between zero and unity, their limits with respect to the real metric are real numbers between zero and unity.

It is probably necessary to recall in its general features the main propositions of von Mises' theory of probability.[14]

A basic object of Mises' theory is a *collective*. Let $\mathscr{S}$ be a random experiment and $T=\{\alpha_1,\ldots,\alpha_m\}$ be the set of all possible realizations of $\mathscr{S}$ (in the simplest case $T=\{0,1\}$). An infinite sequence of realizations of $\mathscr{S}$,

$$x=(x_1,\ldots,x_n,\ldots),\quad x_i\in T, \tag{3}$$

is said to be a collective if for every label $\alpha\in T$ there exists a limit of relative frequencies $\nu_n(\alpha)=k(\alpha)/n$: $\mathbf{P}(\alpha)=\lim_{n\to\infty}\nu_n(\alpha)$. This limit is said to be a probability of α.

Now a sequence (3) is said to be a p-adic collective[20,21] if a limit of relative frequencies exists with respect to the p-adic metric for every label $\alpha\in A$. This limit $\mathbf{P}(\alpha)=\lim_{n\to\infty}\nu_n(\alpha)$ is called a p-adic probability. In Refs. 20 and 21, p-adic collectives were considered from the theoretical and computer points of view.

A measure theoretical definition of p-adic probability is generated by the frequency definition. p-adic probability distributions are defined as $\mathbf{Q}_p$-valued measures with the normalization condition, $P(\Omega)=1$ (about $\mathbf{Q}_p$-valued measures, see Schikhov[23] and the Appendix).

Our interpretation of p-adic numbers,

$$N=l_0+l_1p+\cdots+l_sp^s+\cdots, \tag{4}$$

where $l_s=0,1,\ldots,p-1$, with an infinite number of nonzero digits l_s as infinitely large numbers gives the possibility of considering numerous actual infinities. Therefore we can study proportional probabilities on populations of an infinite volume. We shall study a population $\Omega=\Omega_N$ that has volume N, where N is the p-adic integer (4). If N is finite then Ω is the ordinary finite population; if N is infinite then Ω has essentially a p-adic structure. Consider a sequence of populations M_s having volumes l_sp^s, $s=0,1,\ldots$. Set $\Omega=\cup_{s=0}^{\infty}M_s$. Then $|\Omega|=N$.

We may imagine a population Ω as being the population of a tower $T=T_\Omega$, which has an infinite number of floors with the following distribution of population through floors: population of the sth floor is M_s. Set $T_k=\cup_{s=0}^{k}M_s$. This is population of the first $k+1$ floors.

Let $A\subset\Omega$ and let there exist

$$n(A)=\lim_{k\to\infty}n_k(A),\quad \text{where } n_k(A)=|A\cap T_k|. \tag{5}$$

We define the probability of A by the standard proportional relation:

$$\mathbf{P}(A)\equiv\mathbf{P}_\Omega(A)=\frac{n(A)}{N}. \tag{6}$$

Denote the family of all $A\subset\Omega$, for which (6) exists, by $\mathscr{F}=\mathscr{F}_\Omega$. The sets $A\in\mathscr{F}$ are said to be events. Later, we shall study some properties of the family of events. First we consider the set algebra F that consists of all finite subsets and their complements.

Proposition 2.1: $F\subset\mathscr{F}$.

Proof: Let A be a finite set. Then $n(A)=|A|$ and (6) has the form $\mathbf{P}(A)=|A|/|\Omega|$. Now let $B=\underline{A}$. Then $|B\cap T_k|=|T_k|-|A\cap \underline{T}_k|$. Hence there exists $\lim_{k\to\infty}|B\cap T_k|=N-|A|$. This equality implies the standard formula: $\mathbf{P}(\underline{A})=1-\mathbf{P}(A)$.

In particular, we have $\mathbf{P}(\Omega)=1$.

Proposition 2.2: Let $A_1,A_2\in\mathscr{F}$ *and* $A_1\cap A_2=\varnothing$. *Then* $A_1\cup A_2\in\mathscr{F}$ *and* $\mathbf{P}(A_1\cup A_2)=\mathbf{P}(A_1)+\mathbf{P}(A_2)$.

Proposition 2.3: Let $A_1,A_2\in\mathscr{F}$. *The following conditions are equivalent:*

$$(1)\ A_1\cup A_2\in\mathscr{F};\quad (2)\ A_1\cap A_2\in\mathscr{F};$$
$$(3)\ A_1\backslash A_2\in\mathscr{F};\quad (4)\ A_2\backslash A_1\in\mathscr{F}.$$

There are standard formulas:

$$\mathbf{P}(A_1\cup A_2)=\mathbf{P}(A_1)+\mathbf{P}(A_2)-\mathbf{P}(A_1\cap A_2); \tag{7}$$

$$\mathbf{P}(A_1\backslash A_2)=\mathbf{P}(A_1)-\mathbf{P}(A_1\cap A_2). \tag{8}$$

Proof: We have $n_k(A_1\cup A_2)=n_k(A_1)+n_k(A_2)-n_k(A_1\cap A_2)$. Therefore, if, for example, $A_1\cap A_2\in\mathscr{F}$ then there exists a limit of the right-hand side. It implies $A_1\cup A_2\in\mathscr{F}$ and (7) holds. Other implications are proved in the same way.

In general, $A_1,A_2\in\mathscr{F}$ does not imply $A_1\cup A_2\in\mathscr{F}$. To show this, by Proposition 2.3 it suffices to find $A_1,A_2\in\mathscr{F}$ such that $A_1\cap A_2\notin\mathscr{F}$. It is easy to do: let $A_1,A_2\in\mathscr{F}$ are such that $|A_1\cap A_2\cap M_s|=1$ for nonempty M_s (there is only one element $x\in A_1\cap A_2$ on each nonempty floor). If N is infinite then $\lim_{k\to\infty} n_k(A_1\cap A_2)$ does not exist. Therefore $\mathscr{F}$ is not a set algebra. It is closed only with respect to a finite unions of sets that have empty intersections. However, $\mathscr{F}$ is not closed with respect to countable unions of such sets: in general ($A_j\in\mathscr{F}$, $j=1,2,...$, $A_i\cap A_j=\varnothing$, $i\neq j$,) does not imply $U_{j=1}^{\infty}A_j\in\mathscr{F}$. The natural additional assumptions (a) $\Sigma_{j=1}^{\infty}\mathbf{P}(A_j)$ converges in $\mathbf{Q}_p$ or (a stronger assumption), (b) $\Sigma_{j=1}^{\infty}|\mathbf{P}(A_j)|_p<\infty$, also do not imply $A\in\mathscr{F}$.

Example 2.1: Let $m=2$, $N=-1=1+2+2^2+\cdots+2^n+\cdots$. Suppose that the sets A_j have the following structure: $|A_j\cap M_{3(j-1)}|=1$, $|A_j\cap M_{3j-1}|=2^{3j-1}-1$ and $A_j\cap M_i=\varnothing$, $i\neq 3(j-1)$, $3j-1$, i.e., the set A_j is located on two floors of the tower T. In particular, $A_i\cap A_j=\varnothing$, $i\neq j$. As $A_j\in F$, then $A_j\in\mathscr{F}$; the probability $\mathbf{P}(A_j)=-2^{3j-1}$, $j=1,2...$. The series $\Sigma_{j=1}^{\infty}|\mathbf{P}(A_j)|_2<\infty$. We show that $A=U_{j=1}^{\infty}A_j\notin\mathscr{F}$ We have

$$n_{3(j-1)}(A)=|A_j\cap T_{3(j-1)}|+|U_{s=1}^{j-1}A_s\cap T_{3(j-1)}|=1+\gamma,$$

where $|\gamma|_2<1$. Thus $|n_{3(j-1)}(A)|_2=1$. But $|n_{3j-1}(A)|_2<1$.

We note the following useful formula for computing probabilities: $\mathbf{P}(A)=\Sigma_{j=0}^{\infty}\mathbf{P}(A\cap M_j)$.

Definition 2.1: *The system* $\mathscr{P}=(\Omega,\mathscr{F},\mathbf{P})$ *is called p-adic proportional probability space for the population of the volume* $|\Omega|=N$.

In fact, any proportional probability space $\mathscr{P}$ can be approximated by proportional probability spaces $\mathscr{P}_k$ having populations of finite volumes. Set

$$N_k=l_0+l_1p+\cdots+l_kp^k,$$

for N, which has the expansion (4). Let l_s be the first nonzero digit in (4). Consider populations Ω_{N_k}, $k=s,s+1,...$, and proportional probability spaces $\mathscr{P}_{N_k}=(\Omega_{N_k},\mathscr{F}_{N_k},\mathbf{P}_{N_k})$. There $\mathscr{F}_{N_k}$ coincides with the algebra F_{N_k} of all subsets of Ω_{N_k} and

$$\mathbf{P}_{N_k}(A)=\frac{|A|}{|\Omega_{N_k}|},\quad A\in F_{N_k}. \tag{9}$$

Let us identify Ω_{N_k} with the population of the first $k+1$ floors of the tower T_{Ω}.

Proposition 2.4: Let $A\in\mathscr{F}_{\Omega}$. *Then*

$$\mathbf{P}(A)=\lim_{k\to\infty}\mathbf{P}_{N_k}(A\cap\Omega_{N_k}). \tag{10}$$

To prove (10) we have only used that $\mathbf{Q}_p$ is a topological group. This proposition has the natural frequency interpretation. In fact, the probabilities (10) are relative frequencies after a finite number N_k of trials. Therefore (10) is nothing other than the approximation of proportional probabilities by relative frequencies. The main difference from the usual results of such a kind is that this approximation depends essentially on the rule of a measurement, which is defined by the sequence $\{N_k\}$ corresponding to the volume of population. In principle, the change of this rule may change the limiting result. In this case the relative frequencies would not approximate the theoretical proportional probabilities. However, typically the limit is very stable with respect to the rule of a measurement; see Ref. 20 for the details.

Proposition 2.5: (The image of proportional probability). *The probability* $\mathbf{P}$ *maps* $\mathscr{F}_\Omega$ *into the ball* $U_{r_\Omega}(0)$, *where* $r_\Omega = 1/|N|_p$.

To study conditional probabilities we have to extend the notion of the p-adic proportional probability to consider more general populations.

Let Ω be the population of the tower T_Ω with an infinite number of floors M_j, $j=0,1,...$, and the following distribution of population: there are m_j elements on the jth floor, $m_j \in \mathbf{N}$ and the series $\Sigma_{j=1}^{\infty} m_j$ converges in $\mathbf{Z}_p$ to a nonzero number $N=|\Omega|$. We define the p-adic proportional probability of a set $A \subset \Omega$ by (5), (6); $\mathscr{F}$ is the corresponding family of events. It is easy to check that Propositions 2.1–2.5 hold for this more general proportional probability.

Let $A \in \mathscr{F}_\Omega$ and $\mathbf{P}(A) \neq 0$. We can consider elements of A as a new population and introduce the corresponding family of events $\mathscr{F}_A$.

Proposition 2.6 (conditional probability): Let $A \in \mathscr{F}_\Omega$, $\mathbf{P}(A) \neq 0$, *and* $B \in \mathscr{F}_A$. *Then* $B \in \mathscr{F}_\Omega$ *and Bayes' formula,*

$$\mathbf{P}_A(B) = \frac{\mathbf{P}_\Omega(B)}{\mathbf{P}_\Omega(A)}, \tag{11}$$

holds.

Proof: Set $M_{Aj} = A \cap M_j$, $j=0,1,...$, where M_j is population of the jth floor in the tower T_Ω. Then the tower T_A has the following population structure: there are M_{Aj} elements on the jth floor. In particular, $T_{Ak} = T_k \cap A$. Thus

$$n_{Ak}(B) = |B \cap T_{Ak}| = |B \cap T_k| = n_k(B), \tag{12}$$

for each $B \subset A$. Hence the existence of $n_A(B) = \lim_{k\to\infty} n_{Ak}(B)$ implies the existence of $n_\Omega(B) = \lim_{k\to\infty} n_k(B)$. Moreover, $n_\Omega(B) = n_A(B)$. Therefore,

$$\mathbf{P}_A(B) = \frac{n_A(B)}{n_\Omega(A)} = \frac{n_A(B)/|\Omega|}{n_\Omega(A)/|\Omega|}.$$

■

By (12) we obtain the following consequence.

Corollary 2.1: Let A, $B \in \mathscr{F}_\Omega$, $\mathbf{P}(A) \neq 0$, *and* $B \subset A$. *Then* $B \in \mathscr{F}_A$.

Thus, $\mathscr{F}_A = \{B \in \mathscr{F}_\Omega : B \subset A\}$.

Let A, B, $A \cap B \in \mathscr{F}_\Omega$, $\mathbf{P}(A) \neq 0$. We set by definition $\mathbf{P}_A(B) = \mathbf{P}_A(A \cap B)$. Then

$$\mathbf{P}_A(B) = \frac{\mathbf{P}_\Omega(B \cap A)}{\mathbf{P}_\Omega(A)}. \tag{13}$$

If we set $\mathbf{P}_A(B) = \mathbf{P}(B/A)$ and omit the index Ω for the probabilities for a population Ω then we obtain Bayes' formula.

Remark 2.1: We now discuss the domain of applications of Bayes' formula. This question has the exact and simple mathematical answer in the p-adic proportional probability theory. We can use Bayes' formula for events A and B iff $A \cap B$ is also the event, i.e., $A \cap B \in \mathscr{F}_\Omega$. Therefore in this formalism there are no restrictions for realizing incompatible physical observables X and Y as two random variables on the same proportional probability space. There are some events $A = \{X = t\}$ and $B = \{Y = s\}$ such that the event $A \cap B = \{X = t, Y = s\} \notin \mathscr{F}_\Omega$. Therefore the probability of this event is not defined.

III. *p*-ADIC PROBABILITY DISTRIBUTION OF HIDDEN VARIABLES

We study the simplest case of a discrete set of HVs, $\Lambda=\{\lambda_0,\ldots,\lambda_n,\ldots\}$. Suppose that weight of HV, λ_n, increases exponentially when n approaches infinity. For example, $\rho(\lambda_n)=2^n$, $n=0,1,2,\ldots$.

Remark 3.1: Note that this type of increasing will generate 2-adic structure of the corresponding probability distribution.

Thus, HVs with large numbers dominate very strongly over HVs with small numbers.

First we study this distribution on the basis of the ordinary theory of probability. Set $\Lambda_m=\{\lambda_0,\ldots,\lambda_m\}$. It is a finite approximation of HV set Λ. Compute the sum of all weights for Λ_m:

$$N_m=\rho(\lambda_0)+\cdots+\rho(\lambda_m)=2^{m+1}-1,$$

and probabilities

$$\mathbf{P}_m(\lambda_k)=\rho(\lambda_k)/N_m=2^k/(2^{m+1}-1),\quad k=0,1,\ldots,m.$$

But if $m\to\infty$, all these probabilities approach zero: $\lim_{m\to\infty}\mathbf{P}_m(\lambda_k)=0$, $k=0,1,\ldots,m$, in R. Thus, we cannot define the ordinary Kolmogorov distribution on the countable set Λ, corresponding to these quickly increasing sequence of weights.

But this model contains a natural 2-adic probability structure. We can normalize weights $\rho(\lambda_n)$ in Q_2:

$$N_\infty=\sum_{n=0}^{\infty}\rho(\lambda_n)=-1 \tag{14}$$

and

$$\mathbf{P}(\lambda_n)=\rho(\lambda_n)/N_\infty=-2^n. \tag{15}$$

There (-1) is the only 2-adic representation of an infinite number (14).

How would it be possible to image this 2-adic probability distribution in an intuitive way? The structure of HV distribution is the following one. For every HV, $\lambda=\lambda_k$, there exists HV $\lambda'=\lambda_{k+m}$ that dominates extremely over λ. A probability that a particle is described by HV, λ, is practically zero to compare with a probability that the particle is described by HV, λ'. But there exists another HV $\lambda''=\lambda_{k+m+l}$ with the same property with respect to λ'. By the Appendix about rules for working with proportional p-adic probabilities each probability $\mathbf{P}(\lambda_n)$ is infinitely small quantity with respect to the ''ordinary probabilities.'' Thus, in fact, we work on the set that has zero probability, but we can distinguish probabilities of elements of this set.

Remark 3.2: As it has been already remarked, we cannot define the ordinary Kolmogorov distribution on the countable set Λ corresponding the weights $\rho(\lambda_n)=2^n$, $n=0,1,2,\ldots$, because $\mathbf{P}_m(\lambda_k)\to 0$, $m\to\infty$. However, there exist the ordinary probability distributions with similar properties on the real line (for example, the Lebesgue measure on $[0,1]$). Therefore, the (possible) structure of the set Λ of hidden variables plays an important role in our example. In some sense we provide predictions on the topological structure of the set of hidden variables where violations of Bell's inequality are possible.

Remark 3.3: As we noted in the Introduction, the idea to use negative probability distributions to describe the violations of Bell's inequality is quite old. What is the main advantage of our approach? Why do we prefer to use p-adic realization of negative probabilities? The p-adic framework justifies the consideration of measures with nonpositive values as probabilities. We have the corresponding frequency and proportional interpretations. If we use the ordinary real-valued signed measures, we could not (at least at the moment) present any reason to regard these mathematical objects as probabilities.

IV. STOCHASTIC HIDDEN VARIABLE MODEL

As usual, we relax the requirement that each HV, λ, determines a unique measurement response (for example, $+1$ or -1), in favor of the idea that λ only determines the probability of

response. So we replace the response functions of determinism by the probability functions $\mathbf{P}(S,\lambda)$, where $\mathbf{P}(S,\lambda)$ gives the probability that observable S, measured in state λ, yields value $+1$. We have already discussed that it would be sufficiently natural to consider the situation where the distribution $\rho(\lambda)$ of HV is the p-adic one and responses of macroapparatus $\mathbf{P}(S,\lambda)$ are the ordinary Kolmogorov probabilities. We consider the typical EPR situation where S can be replaced by any of A, A', B, B' in $\mathbf{P}(S,\lambda)$ and also by complements $\bar{A},\ldots$. Then we require that

$$\mathbf{P}(S)=\int_{\Lambda}\mathbf{P}(S,\lambda)\rho(\lambda)d\lambda. \tag{16}$$

The same requirement is for all pairs of commuting variables $ST=AB$, $A'B$, AB', $A'B'$:

$$\mathbf{P}(ST)=\int_{\Lambda}\mathbf{P}(ST,\lambda)\rho(\lambda)d\lambda. \tag{17}$$

We also consider factorizable stochastic models: $\mathbf{P}(ST,\lambda)=\mathbf{P}(S,\lambda)\mathbf{P}(T,\lambda)$.

Now we use p-adic probability HV distribution of the previous section as the basis of the mixed stochastic HV model. It is a discrete model and it is convenient to use symbols $\mathbf{P}(S,n)$ and $\mathbf{P}(ST,n)$, $n=0,1,2,...$, instead of $\mathbf{P}(S,\lambda_n)$ and $\mathbf{P}(ST,\lambda_n)$. We have, instead of (16) and (17),

$$\mathbf{P}(S)=\sum_{n=0}^{\infty}\mathbf{P}(S,n)P(n), \tag{18}$$

where $\mathbf{P}(n)$ are the probabilities (15) and

$$\mathbf{P}(ST)=\sum_{n=0}^{\infty}\mathbf{P}(S,n)\mathbf{P}(T,n)\mathbf{P}(n), \tag{19}$$

as $\mathbf{P}(ST,n)=\mathbf{P}(S,n)\mathbf{P}(T,n)$.

We can always consider rational ordinary probabilities (in the Kolmogorov sense) in all physical applications. Thus, suppose that all probabilities of responses $\mathbf{P}(S,n)$ are rational numbers belonging to the segment $[0,1]$. We wish to prove that it is possible to choose these probabilities in such a way that all probabilities (18) and (19) are ordinary probabilities, rational numbers belonging to the segment $[0,1]$. But Bell's-type inequalities are not valid.

Suppose that all condition probabilities $\mathbf{P}(S,n)$ for the first k numbers n are the same, $\mathbf{P}(A,n)=\alpha$, $\mathbf{P}(B,n)=\beta$, $\mathbf{P}(A',n)=\alpha'$, and $\mathbf{P}(B',n)=\beta'$, $n=0,1,...,k$. In the same way, suppose that all condition probabilities for numbers $n>k$ are also equal and they differ from the first probabilities for $\epsilon>0$: $\mathbf{P}(A,n)=\alpha+\epsilon$, $\mathbf{P}(B,n)=\beta+\epsilon$, $\mathbf{P}(A')=\alpha'+\epsilon$, $\mathbf{P}(B',n)=\beta'+\epsilon$, $n=k+1,...$.

Remark 4.1: It is easy to generalize our considerations for different ϵ.

We consider the probability form of Bell's-type inequalities; see Clauser and Horne.[3] Set

$$\Delta=\mathbf{P}(AB)+\mathbf{P}(AB')+\mathbf{P}(A'B')-\mathbf{P}(A'B)-\mathbf{P}(A)-\mathbf{P}(B'). \tag{20}$$

We study the inequality

$$-1\leqslant\Delta\leqslant 0. \tag{21}$$

In our discrete case,

$$\Delta=\sum_{n=0}^{\infty}[\mathbf{P}(AB,n)+\mathbf{P}(AB',n)+\mathbf{P}(A'B',n)-\mathbf{P}(A'B,n)-\mathbf{P}(A,n)-\mathbf{P}(B',n)]\mathbf{P}(n).$$

Set

$$\Delta_1=\alpha\beta+\alpha\beta'+\alpha'\beta'-\alpha'\beta-\alpha-\beta',$$

and get also Δ_2 to add ϵ to all probabilities in Δ_1. Then we have $\Delta=(-2^{k+1}+1)\Delta_1+2^{k+1}\Delta_2 = 2^{k+1}(\Delta_2-\Delta_1)+\Delta_1$. If we set $\alpha=\beta'=0$, then

$$\Delta=-2^{k+2}\epsilon(1-\epsilon)+\Delta_1. \tag{22}$$

We can also consider the case $\alpha'=\beta=1-\delta$. Then $\Delta_1=-1+\delta_1$, where $\delta_1=2\delta-\delta^2$. Choose $\epsilon=g2^{-(k+2)}$. At the moment g is a rational parameter, $0<g<1$. Now

$$\Delta=-g[1-g2^{-(k+2)}]-1+\delta_1.$$

It is easy to find such g and δ that (21) is violated. We cannot discuss here the general solution of this problem. For example, we can choose $\delta=1/2$ and $g=7/8$. Here $\alpha'=\beta=1/2$ and $\Delta_1=-1/4$ and $\Delta=-9/8+(49/64)2^{-(k+2)}$. Now we can choose k so large that $\Delta<-1$.

But we have

$$\mathbf{P}(A)=\sum_{n=0}^{k}(-2^n)\alpha+\sum_{n=k+1}^{\infty}(-2^n)(\alpha+\epsilon)=\alpha+2^{k+1}(\alpha+\epsilon)=2^{k+1}\epsilon=7/16.$$

Remark 4.2: Roughly speaking, the result is based on the fact that Δ contains $2^{k+2}\epsilon$, (22), and probabilities contain $2^{k+1}\epsilon$.

Further, $\mathbf{P}(B')=7/16$, $\mathbf{P}(A')=\mathbf{P}(B)=15/16$ and

$$\mathbf{P}(AB)=\sum_{n=0}^{k}(-2^n)\alpha\beta+\sum_{n=k+1}(-2^n)(\alpha+\epsilon)(\beta+\epsilon)=2^{k+1}\epsilon(\beta+\epsilon)=7/32+49/2^{k+9}\approx 7/32,$$

for large k. As a consequence of $\beta=\alpha'$, we also have $\mathbf{P}(A'B')=\mathbf{P}(AA')=\mathbf{P}(AB)$. If k is large, then $\mathbf{P}(A'B)\approx 11/16$.

Of course, observables A, A', B, B' are not all mutually compatible in the usual EPR experiments. But we have obtained ordinary Kolmogorov probabilities also for these incompatible observables. It is impossible to compute probabilities $\mathbf{P}(AA')$ and $\mathbf{P}(BB')$ in EPR experiments, but these probabilities permit a statistical testing in our model. Probably we can interpret this property of our model as the possibility of a joint nonideal measurement of incompatible pairs of observables; see, for example, the review in Ref. 25. I think also that it is the only mathematical problem to construct such mixed stochastic HV models, where probabilities of incompatible pairs are non-Kolmogorovean and probabilities of compatible pairs are ordinary probabilities.

V. ON QUADRIVARIATE JOINT PROBABILITIES

Now it is well known that the existence of Bell's-type inequalities is equivalent to an existence of a quadrivariate joint probability distribution $\mathbf{P}(A,A',B,B')$; see Fine[10] and Rastall.[10] What about our model? Of course, there exists a p-adic quadrivariate distribution. But it cannot be an ordinary Kolmogorov distribution:

$$\mathbf{P}(\overline{A},\overline{A'},\overline{BB'})=(-2^{k+1}+1)/4+2^{k+1}(1-\epsilon)^2(1/2-\epsilon)^2\approx 1/4-(3/2)2^{k+1}\epsilon=-13/32,$$

for large k. Fine's paper[10] also contains the proof that a quadrivariate joint distribution can be constructed with the aid of joint distributions of three observables and two observables. This is why our model must also contain non-Kolmogorov probabilities for three observables, and it is so: $\mathbf{P}(A,\overline{B},B')\approx -3/8$ for large k.

VI. HYPOTHESIS OF REPRODUCIBILITY?

Now we try to consider our model from the experimental point of view. What are experimental consequences of p-adic HVs internal structure? Of course, we cannot realize an infinite HV space Λ in an experiment. That is why we must consider again finite approximations $\Lambda_m=\{\lambda_0,\ldots,\lambda_m\}$ of Λ and approximate probabilities $\mathbf{P}_m(\lambda_k)$. These probabilities are concentrated very quickly near λ_m. If m is large then $\mathbf{P}(\lambda_m)\approx 1/2$ and $\mathbf{P}(\lambda_{m-1})\approx 1/4$. Hence, if we consider

two finite sets of HV, Λ_m and Λ_s, where $m<s$ but not by very much, then the sets of HV that dominate in Λ_m, $\{\lambda_m,\lambda_{m-1},\lambda_{m-2}\}$, and in Λ_s, $\{\lambda_s,\lambda_{s-1},\lambda_{s-2}\}$, have the empty intersection.

Using this property of our mixed stochastic HV model, we can try to present some speculations on a hypothesis of reproducibility as a possible explanation of violations of a Bell's-type inequality in EPR experiments. It was discussed in papers[26,25] that to define quadrivariate distribution,

$$\mathbf{P}(A,A',B,B')=\int_\Lambda \mathbf{P}(A,\lambda)\mathbf{P}(A',\lambda)\mathbf{P}(B,\lambda)\mathbf{P}(B',\lambda)\rho(\lambda)d\lambda,$$

it is necessary to assume that the same value of λ can be prepared as an initial HV in different experiments. But it would be really impossible for our p-adic probability distribution of HVs because we can get different Λ_m in different experiments. Different HV, λ, strongly dominate in these experiments. Of course, it is only one of the possible interpretations.

VII. p-ADIC BELL'S-TYPE INEQUALITIES

Let Λ be a set of hidden variables λ distributed with a probability (usual Kolmogorov) density $\rho(\lambda)$. Then from the mathematical point of view all problems are concentrated in the integral

$$\langle AB\rangle=\int_\Lambda A(\lambda)B(\lambda)d\nu(\lambda), \tag{23}$$

where $A(\lambda)$ and $B(\lambda)$ are random variables with values $+1$ or -1 and $d\nu(\lambda)=\rho(\lambda)d\lambda$. We can easily estimate this integral:

$$|\langle AB\rangle|\leqslant\int_\Lambda |A(\lambda)B(\lambda)|d\nu(\lambda)\leqslant\int_\Lambda d\nu(\lambda)=1. \tag{24}$$

Can we get a p-adic-type Bell's inequality? Can we get some kind of Bell's estimate of the mean value (23) in the p-adically distributed case? Using properties of $\mathbf{Q}_p$-valued measures, we have

$$\left|\int_\Lambda A(\lambda)B(\lambda)\nu(d\lambda)\right|_p\leqslant\|\nu\|, \tag{25}$$

where $\|\nu\|$ is the norm (variation) of the probability measure ν. But this variation may be arbitrarily large also in the case of the normalized measure $\nu(\Lambda)=1$. Therefore, we need information about the variation of the p-adic probability distribution ν.

It does not seem surprising from the physical point of view. To prepare an experiment, we must have some initial information that will be tested during the experiment. The EPR experiments to verify Bell's inequality do not use any information about the probability distribution of hidden variables. In fact, these are not statistical experiments where some statistical hypothesis H about a form of a probability distribution must be verified. The only statistical hypothesis, which investigators try to verify during the EPR experiments is

$$H=\{\text{hidden variables have a Kolmogorov probability distribution}\}.$$

The result of numerous experiments is that $\bar{H}$ is right with a large probability.

VIII. CONCLUSIONS

(1) We have built the stochastic HV model where the probability distribution of HVs is sufficiently exotic from the point of view of the usual theory of probability. However, only the mathematical description is exotic (using the p-adic metric). We can create these distributions in real experiments.

(2) Despite the p-adic distribution of HVs, all probabilities that might be measured are usual Kolmogorov probabilities. It is a consequence of the stochastic structure of our HV model (responses of equipment destroy a p-adic probability structure of HVs).

(3) There is no Bell's inequality for this HV model. To say more rigorously, any p-adic ''Bell's-type inequality'' must be based on the information about the properties of the concrete distribution of hidden variables. However, it should be not natural to regard such types of inequalities as analogs of the Bell inequality, since the main reason of Bell's activity was to get inequality for the probability distribution of hidden variables that was not connected with any information about the concrete properties of the distribution of hidden variables.

(4) Our HV model might be realized as a real physical experiment with classical objects (for instance, colored balls). We should see that the ''essentially quantum'' structure of EPR-type experiments could have a classical internal structure. Only a probability distribution (for instance, of colors of the balls) must not be classical. Hence, all effects of the EPR experiment might be generated only by the unusual probability structure of HVs.

(5) Of course, we cannot justify our theoretical construction by statistical data of EPR-type experiments, because there are no statistics of HVs. However, we obtained some predictions on the behavior of HV. Probably, these predictions would be tested in some indirect measurements (without direct statistics of HVs). The main prediction is about a very unstable structure of HV distributions for the stochastic HV models, where Bell's inequality is violated (see the previous section).

(6) We hope that it will be possible to propose some physical experiments (not as complicated as EPR experiments), where we shall be able to test this behavior of HVs. The simplest situation is the case of two HVs, λ_1 and λ_2, which have a probability distribution of HVs such that relative frequencies of realizations of λ_j, $\nu(\lambda_j)$, oscillate with respect to the usual real metric.

(7) We understood that our HV model is probably only a toy model. The p-adic probabilities are the simplest non-Kolmogorovean probabilities. It would be possible to study another probability distributions with non-Kolmogorovean behavior. The p-adic model is very attractive from the analytic point of view. Here we have realized computations (in particular, the violation of Bell's inequality) that were not evident for other kinds of stochastics.

ACKNOWLEDGMENTS

This research was realized on the basis of the Alexander von Humboldt–Stiftung, Volkswagen and Rikkyo University Fellowships and SFB-237, Bochum University.

APPENDIX A: p-ADIC MEASURES

Here we present the simplest variant of p-adic integration. Let X be compact open subset of $\mathbf{Q}_p$. Denote by $\Phi(X)$ the algebra of all clopen (i.e., at the same time open and closed) subsets of X. A measure is defined as a finite-additive set function $\mu:\Phi(X)\to\mathbf{Q}_p$, such that

$$\|\mu\|(A)=\sup\{|\mu(B)|_p : B\subset A, B\in\Phi(X)\}<\infty,$$

for each $A\in\Phi(X)$. In particular, the norm $\|\mu\|=\|\mu\|(X)$ is a finite number. Each measure defines a continuous $\mathbf{Q}_p$-linear functional $\mu:C(X)\to\mathbf{Q}_p$, where $C(X)$ is the space of continuous $\mathbf{Q}_p$-valued functions. Thus, the set of measures coincides with the dual space $C'(X)$. This functional (integral) is defined as a limit of Riemann sums. $S_u(f)=\Sigma_{i=1}^{k} f(a_i)\mu(U_i)$, where $u=(U_i)$, $U_i\in\Phi(X)$, $U_i\cap U_j=\emptyset$, $i\neq j$, is the cover of X. Then this integral is extended to a larger class of integrable functions.

APPENDIX B: RULES FOR WORKING WITH p-ADIC PROBABILITIES

One of the main tools of the ordinary theory of probability is based on the order structure on the field of real numbers $\mathbf{R}$. It gives the possibility of comparing probabilities of different events; events E with probabilities $\mathbf{P}(E)\ll 1$ are considered as negligible and events E with probabilities $\mathbf{P}(E)\approx 1$ are considered as practically certain. However, the use of these relations in concrete applications is essentially based on our (real) probability intuition. What is a large probability? What is a small probability? Moreover, it is not easy to compare two arbitrary probabilities. For

instance, do you prefer to win with the probability $\mathbf{P}(E_1)=11/17$ or $\mathbf{P}(E_2)=13/19$. Formally, because $\mathbf{P}(E_1)<\mathbf{P}(E_2)$, it would be better to choose E_2. But in practice this choice does not give many advantages. Thus, ordinary probability intuition is based more on centuries of human experiment than on exact mathematical theory.

If we want to work with p-adic probabilities we have to develop some kind of a p-adic probability intuition. However, there arises a mathematical problem that does not give the possibility of generalizing the real scheme directly. This is the absence of an order structure on $\mathbf{Q}_p$. Of course, we can also do something without an order structure. For example, we can classify (split) different events with the aid of their p-adic probabilities. For instance, it works sufficiently successfully in the frequency probability theory. If there are two ''random sequences'' x and y that are not collectives in the ordinary von Mises' frequency theory, then we could not split properties of x and y. Both these sequences seem to be totally chaotic from the real point of view. However, if they are p-adic collectives, then it would be possible to classify them with respect to p-adic probability distributions, $\mathbf{P}_x(s_i),\mathbf{P}_y(s_i)$. In the proportional approach different p-adic probabilities, $\mathbf{P}(E_1)\neq\mathbf{P}(E_2)$, mean that the events E_1 and E_2 have different volumes of population.

However, we could do much more with p-adic probabilities by using the partial order structure that exists on the ring of p-adic integers. Let $x=\cdots x_n\cdots x_1x_0$ and $y=\cdots y_n\cdots y_1y_0$ be the canonical expansions of two p-adic integers x, $y\in\mathbf{Z}_p$. We set $x<y$ if there exists n such that $x_n<y_n$ and $x_k\leqslant y_k$ for all $k>n$. This partial order structure on $\mathbf{Z}_p$ is the natural extension of the standard order structure on the set of natural numbers $\mathbf{N}$. It is easy to see that $x<y$ for any $x\in\mathbf{N}$ and $y\in\mathbf{Z}_p\backslash\mathbf{N}$, i.e., any finite natural number is less than any infinite number. But we could not compare any two infinite numbers.

Example 10.2: Let $x=\ldots0101....0101$, $z=\ldots.1010...1010$ and $y=\ldots1111...1000$, then $x<y$ and $z<y$, but the numbers x and z are incompatible.

It is important to remark that there exists the maximal number $N_{\max}\in\mathbf{Z}_p$. It is easy to see that

$$N_{\max}=-1=(p-1)+(p-1)p+\cdots+(p-1)p^n+\cdots .$$

Therefore the population Ω_{-1} is the largest population that can be considered in the p-adic framework.

Remark 10.3: It seems to be natural to suppose that the volume of the population increases with the increase of p, i.e., $|\Omega^p_{-1}|<|\Omega^q_{-1}|$, $p<q$.

Proposition 10.7: Let $N\neq0$ have the form (4). Then $\Omega_N\in\mathcal{F}_{\Omega_{-1}}$ and

$$\mathbf{P}_{\Omega_{-1}}(\Omega_N)=\frac{|\Omega_N|}{|\Omega_{-1}|}=-N. \tag{26}$$

Corollary 10.3: Let $N\in\mathbf{Z}_p$, $N\neq0$. Then $\mathcal{F}_{\Omega_N}\subset\mathcal{F}_{\Omega_{-1}}$ and probabilities $\mathbf{P}_{\Omega_N}(A)$ are calculated as conditional probabilities with respect to the subpopulation Ω_N of population Ω.

But $A\in\mathcal{F}_{\Omega_{-1}}$ does not imply $A\cap\Omega_N\in\mathcal{F}_{\Omega_N}$.

By Corollary 10.3 we can, in fact, restrict our considerations to the case of the maximal population Ω_{-1}. Therefore we shall study this case $\Omega\equiv\Omega_{-1}$.

Let $A,B\in\mathcal{F}$. By definition, $\mathbf{P}(A)<\mathbf{P}(B)$ iff $n(A)<n(B)$. Further, we study the properties of probabilities.

(1) As we have only a partial order structure we cannot compare probabilities of arbitrary two events A and B.

(2) As $x\leqslant-1$ for any $x\in\mathbf{Z}_p$, we have $\mathbf{P}(A)\leqslant1=\mathbf{P}(\Omega)$ for any $A\in\mathcal{F}$.

(3) As $0\leqslant x$ for any $x\in\mathbf{Z}_p$, we have $0\leqslant\mathbf{P}(A)$ for any $A\in\mathcal{F}$.

To illustrate further properties of p-adics we shall use the usual real order structure. In this case we shall say an ''r increase'' or an ''r decrease.'' We shall use the symbols $[a,b],\ldots,(a,b)$ for corresponding intervals of the real line.

(4) Set $F^f=\{A\in\mathcal{F}:n(A)\in\mathbf{N}\}$ (in particular, F^f contains all finite subsets of Ω). Because the partial order structure on $\mathbf{N}$ coincides with the standard order structure, we have the standard order relation for $n(A),n(B)$, $A,B\in F^f$. But $\mathbf{P}:F^f\to(-\infty,0)\cap\mathbf{Z}$ and $\mathbf{P}(A)$ is increasing to 1 if $\mathbf{P}(A)$ is r decreasing to $-\infty$. Therefore events with probabilities $\mathbf{P}(A)=-1$ or -3 are practically negligible with respect to events with probabilities $\mathbf{P}(B)=-100$ or -300 or with respect to the total

population Ω. An event A with probability $\mathbf{P}(A)=1000$ seems to be practically certain. Of course, the levels of negligible probabilities depend on the concrete model. An event with probability $\mathbf{P}(A)=-100$ may be negligible in one situation and not negligible in another. Further p-adic probability intuition must be developed from the results of experimental applications.

(5) Set $\overline{F^f}=\{B=\overline{A}:A\in F^f\}$ (in particular, $\overline{F^f}$ contains complements of all finite subsets of Ω). Then $\mathbf{P}:\overline{F^f}\to(1,\infty)\cap\mathbf{Z}$ and $\mathbf{P}(B)$ is decreasing to 0 if $\mathbf{P}(B)$ is r increasing to ∞. Therefore, events with probabilities $\mathbf{P}(E)=100$ or 200 are practically negligible with respect to events with probabilities $\mathbf{P}(C)=1$ or 2. An event C with probability 2 (or 3) seems to be practically certain. Of course, the level of negligible probabilities depends again on the concrete model.

We can use these rules for conditional probabilities. For example, let $\mathbf{P}(B)=100$, $\mathbf{P}(B')=200$, $\mathbf{P}(A)=2$, and $B,B'\subset A$. Then $\mathbf{P}(B/A)=50>\mathbf{P}(B'/A)=100$.

On the other hand, any event A with a probability $x\in(-\infty,0)\cap\mathbf{Z}$ is practically negligible with respect to any event B with a probability $y\in(1,\infty)\cap\mathbf{Z}$. For example, an event with probability -1000 is negligible with respect to an event with the probability 2.

By (4) and (5) we can work with probabilities belonging to $F^f\cup\overline{F^f}$.

(6) Now consider events $A\notin F^f\cup\overline{F^f}$. We can develop our intuition only by examples.

Example 10.3: Let $p=2$. Let $|A\cap M_{2k}|=2^{2k}$ and $A\cap M_{2k+1}=\emptyset$, $k=0,1,\ldots$. Then $n(A)=-1/3$ and $\mathbf{P}(A)=1/3$. Let $B\subset A$ and $B\cap M_{4k}=A\cap M_{4k}$, $B\cap M_j=\emptyset$, $j\neq 4k$. Then $n(B)=-1/15$ and $\mathbf{P}(B)=1/15$. It is evident that $-1/15<-1/4$ in $\mathbf{Z}_2$. Hence, $\mathbf{P}(B)=1/15<\mathbf{P}(A)=1/3$.

Thus it seems to be that the probabilistic relation on the set $[0,1]\cap\mathbf{Q}$ coincides with the standard real order. Moreover, it seems to be reasonable to use this relation also in the case where the numbers $n(A)$ and $n(B)$ are incompatible in $\mathbf{Z}_2$.

Example 10.4: Let p and A be the same as above. Let $|C\cap M_{2k+1}|=2^{2k+1}$, $C\cap M_{2k}=\emptyset$, $k=0,1,\ldots$. Then $n(C)=-2/3$ and $\mathbf{P}(C)=2/3$. The numbers $n(A)=-1/3$ and $n(C)=-2/3$ are incompatible in $\mathbf{Z}_2$. But heuristically it seems to be evident that we can use the r-order structure on $[0,1]$ to compare the probabilities of the events A and C. Therefore the probability of $\omega\in C$ is two times larger than the probability $\omega\in A$. These heuristic reasons were also confirmed by some frequency statistical models; see Ref. 20 for the details.

Further, we have that any set E with a probability $x\in(-\infty,0)\cap\mathbf{Z}$ is practically negligible with respect to any set A with a probability in $(0,1]\cap\mathbf{Q}$. For example, an event with probability -1000 is negligible with respect to an event with probability 1/3.

We can also use these rules for conditional probabilities. For example, let $\mathbf{P}(B)=1/15<\mathbf{P}(B')=2/15$, $\mathbf{P}(A)=1/5$, and $B,B'\subset A$. Then $\mathbf{P}(B/A)=1/3<\mathbf{P}(B'/A)=2/3$. Moreover, for example, let $\mathbf{P}(B)=-1<\mathbf{P}(B')=-5$, $\mathbf{P}(A)=-100$, and $B,B'\subset A$. Then $\mathbf{P}(B/A)=1/100<\mathbf{P}(B'/A)=1/20$. Thus the r-order structure on $(0,1]\cap\mathbf{Q}$ reproduces the rule (4).

[1] J. S. Bell, Physics (N.Y.) **1**, 195 (1964).
[2] J. F. Clauser, M. A. Horne, A. Schimony, and R. A. Holt, Phys. Rev. Lett. **23**, 888 (1969).
[3] J. F. Clauser and M. A. Horne, Phys. Rev. D **10**, 526 (1974).
[4] A. N. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin 1933. English translation by N. Morrison (1950).
[5] J. F. Clauser and A. Shimony, Rep. Prog. Phys. **41**, 1883 (1978).
[6] D. Home and F. Selleri, Nuovo Cimento Riv. **14** (1991).
[7] G. Lochak, Found. Phys. **6**, 173 (1976).
[8] L. Accardi, Phys. Rep. **77**, 169 (1981).
[9] *''The wave-particle dualism,'' A Tribute to Louis de Broiglie on his 90th Birthday*, edited by S. Diner, D. Fargue, G. Lochak, and F. Selleri (Reidel, Dordrecht, 1970); G. Lochak, ''De Broglie's initial conception of de Broiglie waves,'' pp. 1–25; L. Accardi, ''The probabilistic roots of the quantum mechanical paradoxes,'' pp. 297–330.
[10] A. Fine, Phys. Rev. Lett. **48**, 291 (1982); P. Rastall, Found. Phys. **13**, 555 (1983).
[11] W. Muckenheim, Phys. Rep. **133**, 338 (1986).
[12] L. De Broglie, *La Physique Quantique: Restera-t-elle Indeterministe* (Gauthier Villars, Paris, 1953).
[13] A. Einstein, B. Podolsky, and N. Rosen, Phys. Rev. **47**, 777 (1935).
[14] R. von Mises, Grundlagen Wahrsch. Math. Z. **5**, 52 (1919); *The Mathematical Theory of Probability and Statistics* (Academic, London, 1964).
[15] R. P. Feynman, ''Negative probability,'' in *Quantum Implications*, Essays in Honour of David Bohm, edited by B. J. Hiley and F. D. Peat (Routledge and Kegan Paul, London, 1987), pp. 234–248; Int. J. Theor. Phys. **21**, 467 (1982).
[16] J. Bub, Found. Phys. **3**, 29 (1973).
[17] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **180**, 1 (1942).
[18] I. Pitowsky, Phys. Rev. D **27**, 2316 (1983).

[19] S. P. Gudder, J. Math. Phys. **25**, 2397 (1984); A. Stoll, ''Must hidden variables follow a probability law?,'' preprint Fakult. Math., Ruhr-Univ. Bochum, 1987, No. 101.
[20] A. Yu. Khrennikov, *p-adic Valued Distributions in Mathematical Physics*, Kluwer Academic, Dordrecht, 1994.
[21] A. Yu. Khrennikov, Dokl. Akad. Nauk SSSR **322**, 1075–1079 (1992); **338**, 313–316 (1994); J. Math. Phys. **32**, 932 (1991); Teor. Mat. Fiz. **97**, 348 (1993); R. Cianci and A. Yu. Khrennikov, Phys. Lett. B **328**, 109–112 (1994).
[22] A. Yu. Khrennikov, Phys. Lett. A **200**, 119–223 (1995).
[23] K. Mahler, ''Introduction to p-adic numbers and their functions, *Cambridge* Tract. in Mathematics (Cambridge University Press, Cambridge, 1973), p. 64; W. Schikhof, *Ultrametric Calculus*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 1984), Vol. 4; Z. I. Borevich and I. R. Schafarevich, Number theory, Academic Press, New York, 1966.
[24] A. Yu. Khrennikov, Found. Phys. **26**, 1033 (1996); S. Albeverio and A. Yu. Khrennikov, J. Phys. A **29**, 5515 (1996); Int. J. Mod. Phys. A **10**, 1665 (1996); S. Albeverio and A. Yu. Khrennikov, ''A regularization of quantum field Hamiltonians with the of p-adic numbers,'' preprint Bochum University, 1995, SFB 237, No. 274.
[25] W. De Baere, Nuovo Cimento A **40**, 488 (1984).
[26] W. E. Muynck, W. E. Baere, and H. Martens, Found. Phys. **24**, 1598 (1994).

Physical vertex operators in the Ramond sector

Seichi Naito
Department of Physics, Osaka City University, Sumiyoshiku, Osaka, Japan

(Received 17 September 1997; accepted for publication 1 December 1997)

Introducing *physical* string coordinate's modes as well as *physical* string coordino's modes, we derive the generating functional of *physical* vertex operators (GFPVO) of fermionic particles (in superstring), which is just equal to the one already proposed by using superconformal mapping in the Ramond sector. Inversely solving the superconformal mapping problem, we can derive general formulas, which give physical vertex operators of various fermionic particles up to arbitrarily excited mass levels. As an example, we explicitly derive some (Gliozzi–Scherk–Olive-allowed) low-lying physical vertex operators. © *1998 American Institute of Physics.* [S0022-2488(98)02703-0]

I. INTRODUCTION AND PRELIMINARIES

There exist two formulations of superstring: Ramond–Neveu–Schwarz (RNS) superstring[1,2] and Green–Schwarz superstring.[3] In our previous papers,[4,5] we have investigated RNS superstring, along the line proposed by Friedan, Martinec and Schenker (FMS).[6] The problem is how to construct all physical Ramond as well as Neveu–Schwarz vertex operators by using the common operators to both the R and NS sector. This problem has been solved, suggested by our previous paper[7] which has succeeded in constructing any physical vertex operator in the bosonic string. The strategy is using two different kinds of superconformal mappings, in constructing the generating functional of the physical vertex operator (GFPVO) for each R and NS sector. However, in Ref. 7, we have shown in the bosonic string case that GFPVO can be constructed by two methods. One method is using certain conformal mapping, while the other one is using DDF operators.[8] Furthermore, the equivalence of two methods has been explicitly shown. The former one is useful in explicitly obtaining all physical vertex operators, while the latter one is useful in clarifying the structure of the (thus constructed) physical Hilbert space. On the other hand, we have in Refs. 4 and 5 used only the former method in a RNS superstring.

A RNS superstring is described by the string coordinate $X^{\pm j}$ and the string coordino $\Psi^{\pm j} \equiv :\exp(\pm\phi^j):$ (for $j=0-4$). We construct physical string coordinate's modes $A^{\pm j}_{-m}$ and physical coordino's modes $\mathscr{D}^{\pm j}_{-m}$ [for $j=1-4$], which are, respectively, physical components of $X^{\pm j}$ and $\Psi^{\pm j}$ in the R sector. We prove that these physical operators satisfy the spectrum generating algebra proposed by Green and Schwarz.[3] With the help of these physical operators $A^{\pm j}_{-m}$ and $\mathscr{D}^{\pm j}_{-m}$, we construct GFPVO in the R sector, which is subsequently proved to be just equal to the previously proposed GFPVO in Ref. 4. Therefore we summarize the main results in Ref. 4 hereafter.

We introduce the light-cone operator $P_F^{\pm j\dagger}(w,\xi)$ $(j=1-4)$ in the R sector defined by

$$P_F^{\pm j^\dagger}(w,\xi) \equiv \frac{\xi}{w}\left(p^{\pm j}+\sum_{n=1}^{\infty} w^{-n}\alpha_{-n}^{\pm j}\right)+\frac{1}{\sqrt{w}}\left(\sum_{n=1}^{\infty} w^{-n} d_{-n}^{\pm j}\right) \tag{1.1}$$

$$(\equiv \xi\Pi_\alpha^{\pm j^\dagger}(w)+\pi_d^{\pm j^\dagger}(w)), \tag{1.1'}$$

where we have used light-cone (mode's) operators (for $j=1-4$), i.e., $p^{\pm j}$ (total energy-momentum operator), $\alpha_m^{\pm j}$'s, $\alpha_{-n}^{\pm j}$'s, $d_m^{\pm j}$'s, and $d_{-n}^{\pm j}$'s, which satisfy the following (anti-) commutation relations among themselves;

$$[\alpha_m^{\pm j}, \quad \alpha_{-n}^{\mp k}] = m\delta_{mn}\delta_{jk} \tag{1.2}$$

 © 1998 American Institute of Physics

and

$$\{d_m^{\pm j},\quad d_{-n}^{\mp k}\}=\delta_{mn}\delta_{jk}, \tag{1.3}$$

all other (anti-) commutators being zero.

In the R sector of superstring [i.e., in the *fermionic* sector $s=F$], there exists FMS spinor[6] $S_{\{h\}}(z_0)$ at z_0, which is given by

$$S_{\{h\}}(z_0)\equiv :\exp(\tfrac{1}{2}\phi^0(z_0)):_\phi\times\prod_{j=1}^{4}:\exp(\epsilon_h^j\phi^j(z_0)):_\phi\quad(\text{with}\ \ \epsilon_\pm^j=\pm\tfrac{1}{2}). \tag{1.4}$$

Then any non-N operator $G(z)$ composed of $X^{\pm j}$ and $\Psi^{\pm j}$ is N-ized[4] into the N operator $G_N(z)$ by

$$\mathbf{R}\cdot G(z)S_{\{h\}}(z_0)\equiv G_N(z)\cdot S_{\{h\}}(z_0). \tag{1.5}$$

[Inversely, we state that the N operator G_N is changed into the non-N operator G.] For an example, we have that

$$\begin{aligned}\mathbf{R}\cdot :\exp(\pm\phi^j(z)):_\phi S_{\{h\}}(z_0)&=(z-z_0)^{\pm\epsilon_h^j}:\exp(\pm\phi^j(z))\cdot S_{\{h\}}(z_0):_\phi\\&=(z-z_0)^{\pm\epsilon_h^j}:\exp(\pm\phi_{Nh}^j(z)):_\phi S_{\{h\}}(z_0),\end{aligned} \tag{1.6}$$

$\mathbf{R}$ being radial ordering. In Eq. (1.6), $\phi_{Nh}^j(z)$ is the operator which has already been normally ordered with respect to FMS spinor $S_{\{h\}}(z_0)$. From Eq. (1.6), we have found that

$$\Psi^{\pm j}(z)\equiv :\exp(\pm\phi^j(z)):_\phi\overset{N}{\to}\Psi_N^{\pm j}(z)\equiv(z-z_0)^{\pm\epsilon_h^j}:\exp(\pm\phi_{Nh}^j(z)):_\phi, \tag{1.7}$$

$$\Psi_h^{\pm j}(z)\equiv(z-z_0)^{\mp\epsilon_h^j}:\exp(\pm\phi^j(z)):_\phi\overset{N}{\to}\Psi_{Nh}^{\pm j}(z)\equiv:\exp(\pm\phi_{Nh}^j(z)):_\phi, \tag{1.8}$$

and

$$X^{\pm j}(z,\theta)\equiv X^{\pm j}(z)+\theta\cdot\Psi^{\pm j}(z)\overset{N}{\to}X_N^{\pm j}(z,\theta)\equiv X^{\pm j}(z)+\theta\cdot(z-z_0)^{\pm\epsilon_h^j}\Psi_{Nh}^{\pm j}(z). \tag{1.9}$$

Then GFPVO $\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ (in the N operator's form) can be constructed by

$$\begin{aligned}\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)(\equiv\Phi_{N\{h\}}^{F0}(z_0;P_F^\dagger)+\theta_0\cdot\Phi_{N\{h\}}^{F1}(z_0;P_F^\dagger))\equiv:\exp(\rho_{N+}^{0F}(z_0,\theta_0)+\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)\\+\rho_{N\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)):_R.\end{aligned} \tag{1.10}$$

In Eq. (1.10), we have defined that

$$\rho_{N+}^{0F}(z_0,\theta_0)\equiv -ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)-ip^{-0}X^{+0}(z_0) \tag{1.11}$$

$$(\text{with}\ \ Y_{N+}^{-0}(z_0,\theta_0)\equiv X^{-0}(z_0)+\theta_0\Psi_{N+}^{-0}(z_0)), \tag{1.11$'$}$$

$$\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)\equiv-\sum_{\pm}\sum_{j=1}^{4}\ i\oint\frac{dw\ d\xi}{2\pi i}\ P_F^{\pm j^\dagger}(w,\xi)X_N^{\mp j}(z_{NF}(w,\xi),\theta_{NF}(w,\xi)), \tag{1.12}$$

with

$$X_N^{\mp j}(z_{NF}(w,\xi),\theta_{NF}(w,\xi))\equiv X^{\mp j}(z_{NF}(w,\xi))+\theta_{NF}(w,\xi)(z_{NF}(w,\xi)-z_0)^{\mp\epsilon_h^j}\cdot\Psi_{Nh}^{\mp j}(z_{NF}(w,\xi)), \tag{1.12$'$}$$

and

$$\rho^{2F}_{N\{h\}}(z_0,\theta_0;P^\dagger_F)$$

$$\equiv -\sum_{j=1}^{4}\oint\frac{dw\ d\xi}{2\pi i}P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'\ d\xi'}{2\pi i}P_F^{+j^\dagger}(w',\xi')$$

$$\times\log\frac{z_{NF}(w,\xi)-z_{NF}(w',\xi')-([z_{NF}(w,\xi)-z_0]/[z_{NF}(w',\xi')-z_0])^{\epsilon^j_h}\theta_{NF}(w,\xi)\,\theta_{NF}(w',\xi')}{w-w'-(w/w')^{\epsilon^j_h}\cdot\xi\xi'},\tag{1.13}$$

where ϵ^j_h [i.e., $\epsilon^j_\pm\equiv\pm\frac{1}{2}$] is the constant used in $S_{\{h\}}(z_0)$ (1.4). In Eq. (1.10), $:\ :_R$ is the normal ordering in the R sector which is defined by

$$\mathbf{R}\cdot X^{\pm j}(z')X^{\mp k}(z)=\delta_{jk}\log(z'-z)+:X^{\pm j}(z')X^{\mp k}(z):_R\tag{1.14}$$

and

$$\mathbf{R}\cdot\Psi^{\pm j}_{Nh'}(z')\Psi^{\mp k}_{Nh}(z)=\frac{\delta_{jk}}{z'-z}+:\Psi^{\pm j}_{Nh'}(z')\Psi^{\mp k}_{Nh}(z):_R,\tag{1.15}$$

$\mathbf{R}$ being radial ordering. N operators $\theta_{NF}(w,\xi)$ and $z_{NF}(w,\xi)$ in Eqs. (1.12) and (1.13) can be expressed by

$$\theta_{NF}(w,\xi)\equiv J_{NF}(w)+\xi K_{NF}(w)\tag{1.16}$$

and

$$z_{NF}(w,\xi)-z_0\equiv L_{NF}(w)+\xi K_{NF}(w)J_{NF}(w),\tag{1.17}$$

and they are determined by inversely solving the following superconformal mapping in the R sector $(z_{NF},\theta_{NF})\to(w,\xi)$ [in the N operator's form];

$$z_{NF}-z_0-\sqrt{z_{NF}-z_0}\,\theta_{NF}\theta_0=w:\exp\left(\frac{X^{+0}_N(z_{NF},\theta_{NF})}{ip^{+0}}-\frac{X^{+0}(z_0)}{ip^{+0}}\right):_R\tag{1.18}$$

with the N operator $X^{+0}_N(z,\theta)$ being defined by Eq. (1.9), and

$$\xi(z_{NF},\theta_{NF})=(\partial_{z_{NF}}w(z_{NF},\theta_{NF}))^{-1/2}D_{\theta_{NF}}w(z_{NF},\theta_{NF}),\tag{1.19}$$

with $D_{\theta_{NF}}\equiv\partial_{\theta_{NF}}+\theta_{NF}\partial_{z_{NF}}$. It should be noticed that we have used in Eq. (1.18) only the string coordinate $X^{+0}(z_0)$ and *not* any string coordino $\Psi^{+0}_{N+}(z_0)$, while we have used $X^{+0}_N(z_{NF},\theta_{NF})$ (1.9). The inversely solved results $z_{NF}(w,\xi)$ and $\theta_{NF}(w,\xi)$ can be obtained as follows. First we consider operators $A^\pm_N(z)$ and $B^\pm_N(z)$, both of which are arbitrary functions of z analytic at $z=z_0$. With respect to $A^\pm_N(z_{NF})$ and $B^\pm_N(z_{NF})$, the function $[\quad]_{\xi=0}$ (which is the function of w) is defined by

$$[A^+_N(z_{NF})+\theta_{NF}\sqrt{z_{NF}-z_0}B^+_N(z_{NF})]_{\xi=0}$$

$$\equiv A^+_N(z_0)+\sum_{n=1}^{\infty}w^n\left(\sum_{k=1}^{n}\frac{1}{(k-1)!}R^{(n)}_{n-k}(z_0)\partial^{k-1}_{z_0}\left(\frac{1}{n}\,\partial_{z_0}A^+_N(z_0)+\theta_0B^+_N(z_0)\right)\right.$$

$$\left.+\sum_{k=2}^{n}\frac{1}{(k-2)!}R^{(n)}_{n-k}(z_0)\partial^{k-2}_{z_0}(V_N(z_0)B^+_N(z_0))\right)\tag{1.20}$$

and

$$
\begin{aligned}
&[\sqrt{z_{NF}-z_0}A_N^-(z_{NF})+\theta_{NF}B_N^-(z_{NF})]_{\xi=0}\\
&\quad\equiv\sum_{n=0}^{\infty}w^{n+1/2}\Bigg(\sum_{k=0}^{n}\frac{1}{k!}R_{n-k}^{(n+1/2)}(z_0)\partial_{z_0}^k\left(\frac{k+1/2}{n+1/2}A_N^-(z_0)+\theta_0B_N^-(z_0)\right)\\
&\qquad+\sum_{k=1}^{n}\frac{1}{(k-1)!}R_{n-k}^{(n+1/2)}(z_0)\partial_{z_0}^{k-1}(V_N(z_0)B_N^-(z_0))\Bigg),
\end{aligned}
\tag{1.21}
$$

where

$$
I(z_0)\equiv\frac{X^{+0}(z_0)}{ip^{+0}},
\tag{1.22}
$$

$$
R_m^{(n)}(z_0)\equiv\frac{1}{m!}\exp(-nI(z_0))\cdot\partial_{z_0}^m\exp(nI(z_0)),
\tag{1.23}
$$

and

$$
V_N(z_0)\equiv\frac{\Psi_{N+}^{+0}(z_0)}{ip^{+0}}.
\tag{1.24}
$$

Equipped with these preliminaries (1.20)–(1.24), we can state the theorem proved in Ref. 4.

Theorem: Superconformal mapping in the R sector in the N operator's form

N operators $F_N^{\pm}(z_{NF},\theta_{NF})$, given by

$$
F_N^+(z_{NF},\theta_{NF})\equiv A_N^+(z_{NF})+\theta_{NF}\sqrt{z_{NF}-z_0}B_N^+(z_{NF})
\tag{1.25}
$$

and

$$
F_N^-(z_{NF},\theta_{NF})\equiv\sqrt{z_{NF}-z_0}A_N^-(z_{NF})+\theta_{NF}B_N^-(z_{NF})
\tag{1.26}
$$

can be expressed as functions of w and ξ by the following formula:

$$
F_N^{\pm}(z_{NF},\theta_{NF})=[F_N^{\pm}(z_{NF},\theta_{NF})]_{\xi=0}+\xi K_{NF}(w)[D_{\theta_{NF}}F_N^{\pm}(z_{NF},\theta_{NF})]_{\xi=0},
\tag{1.27}
$$

with

$$
K_{NF}(w)=(\partial_w[z_{NF}-z_0]_{\xi=0}-(\partial_w[\theta_{NF}]_{\xi=0})[\theta_{NF}]_{\xi=0})^{1/2}.
\tag{1.28}
$$

By using the above theorem, the N operators $J_{NF}(w)$ in Eq. (1.16) and $L_{NF}(w)$ in Eq. (1.17) can be calculated by

$$
J_{NF}(w)=[\theta_{NF}]_{\xi=0}
\tag{1.29}
$$

and

$$
L_{NF}(w)=[z_{NF}-z_0]_{\xi=0},
\tag{1.30}
$$

as we shall see in this paper. [When Eq. (a.b) expressed in terms of non-N operators is N-ized into Eq. (a.b′) expressed in terms of N operators, we state that Eq. (a.b′) in the N operator's form is changed into Eq. (a.b) in non-N operator's form.]

In Sec. II, we introduce the *physical* string coordinate's mode $A_n^{\pm j}$ (for any integer n) as well as the *physical* string coordino's mode $\mathscr{D}_n^{\pm j}$ (for any integer n) in the R sector, and we prove that they generate *physical* Ramond states in superstring. Therefore we can construct the functional which generates all physical Ramond states. Thus the constructed functional is shown to be just equal to GFPVO $\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ [given by Eqs. (1.4) and (1.10)]. Furthermore, our previous result[5] on OPE among GFPVO and the superstressino operator is shown to be easily derived by the new method in Sec. II. In Sec. III we derive various formulas (obtained by

superconformal mapping) with the help of which N operators ρ_N^{1F} (1.12) and $\rho_{N\{h\}}^{2F}$ (1.13) can be calculated up to an arbitrarily required number of terms. As an example, some (GSO-allowed[10]) physical vertex operators in the R sector of the superstring are explicitly given for some low-lying mass levels. With the help of the obtained vertex operators, we construct physical Ramond states in 1/2 picture. In Sec. IV, various results obtained in the R sector are summarized. On the other hand, we give various results in the NS sector without any proof, since these can be derived in a much simpler way. In Appendix A, we prove (anti-) commutation relations satisfied by $A_n^{\pm j}$ and $\mathscr{D}_n^{\pm j}$, which are used in Sec. II. In Appendix B, some terms of ρ_N^{1F} (1.12) and $\rho_{N\{h\}}^{2F}$ (1.13) are given explicitly.

II. NEW DERIVATION OF THE FORMULA (1.10) IN THE *R* SECTOR

In this section we shall derive the formula (1.10) in the R sector by using a new method different from our previous one;[5] For this purpose, we shall introduce [in the R sector] the *physical* string coordinate's mode $A_{-n}^{\pm j}$ (for any integer n) defined by

$$A_{-n}^{\pm j} \equiv -i \oint_0 \frac{dz\ d\theta}{2\pi i} D_\theta X^{\pm j}(z,\theta) \exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}\right), \tag{2.1}$$

as well as the *physical* string coordino's mode $\mathscr{D}_{-n}^{\pm j}$ (for any integer n) defined by

$$\mathscr{D}_{-n}^{\pm j} \equiv -i \oint_0 \frac{dz\ d\theta}{2\pi i} D_\theta X^{\pm j}(z,\theta) \frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}])^{1/2}} \cdot \exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}\right),$$
$$\text{for } n=0,\pm 1,\pm 2,\ldots, \quad \text{and } j=1-4. \tag{2.2}$$

In Appendix A, operators (2.1) and (2.2) are shown to satisfy the following (anti-) commutation relations (for $n,m=0,1,2,\ldots$, and $j,k=1-4$):

$$[A_n^{\pm j},\ A_{-m}^{\mp k}]=n\delta_{jk}\delta_{nm}, \quad [A_n^{(\ )},\ \mathscr{D}_{-m}^{(\ )}]=0, \quad \text{and } \{\mathscr{D}_n^{\pm j},\ \mathscr{D}_{-m}^{\mp k}\}=\delta_{jk}\delta_{nm}, \tag{2.3}$$

[other (anti-) commutators being equal to zero] and that any of them (anti-) commutes with any of Virasoro generators $L_m^F(z_0)$ and $\mathscr{T}_m^F(z_0)$, which are defined by Eqs. (A4) and (A5);

$$[L_m^F(z_0),\ A_{-n}^{\pm j}]=[L_m^F(z_0),\ \mathscr{D}_{-n}^{\pm j}]=[\mathscr{T}_m^F(z_0),\ A_{-n}^{\pm j}]=\{\mathscr{T}_m^F(z_0),\mathscr{D}_{-n}^{\pm j}\}=0. \tag{2.4}$$

(Anti-) commutation relations (2.3) and (2.4) show that physical operators $A_{-n}^{\pm j}$ (2.1) and $\mathscr{D}_{-n}^{\pm j}$ (2.2) satisfy the spectrum generating algebra proposed by Green and Schwarz.[3]

We shall introduce the *basic* spinorial $\Sigma_{\{h\}}(z_0|\theta_0)$ defined by

$$\Sigma_{\{h\}}(z_0|\theta_0) \equiv :\exp\Bigg(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)-ip^{-0}X^{+0}(z_0)$$
$$-\sum_{j=1}^{4} ip^{hj}Y_{Nh}^{-hj}(z_0,\theta_0)-\sum_{j=1}^{4} ip^{-hj}X^{+hj}(z_0)\Bigg) S_{\{h\}}(z_0):_R, \tag{2.5}$$

where we have used the FMS spinor (1.4) and Eq. (1.11′), together with

$$Y_{Nh}^{-hj}(z_0,\theta_0) \equiv X^{-hj}(z_0)+\theta_0\Psi_{Nh}^{-hj}(z_0). \tag{2.5′}$$

[Incidentally, $\Psi_{Nh}^{+hj}(z_0)$'s (for $j=0-4$) do not exist in Eq. (2.5).] Then we find with the help of contractions in the R sector[4] that

$$\mathbf{R}\cdot A_n^{\pm hj}\cdot\Sigma_{\{h\}}(z_0|\theta_0)\equiv -i\oint_{|z|>|z_0|}\frac{dz\ d\theta}{2\pi i}:\Big(D_\theta X_N^{\pm hj}(z,\theta)-ip^{\pm hj}D_\theta \times\log\Big(z-z_0-\begin{pmatrix}1\\0\end{pmatrix}\sqrt{z-z_0}\theta\theta_0\Big)\Big) \times\exp\left(\frac{-nX_N^{+0}(z,\theta)}{ip^{+0}}\right)(z-z_0-\sqrt{z-z_0}\theta\theta_0)^n\Sigma_{\{h\}}(z_0|\theta_0):_R \tag{2.6}$$

and

$$\mathbf{R}\cdot\mathscr{D}_n^{\pm hj}\cdot\Sigma_{\{h\}}(z_0|\theta_0)\equiv -i\oint_{|z|>|z_0|}\frac{dz\ d\theta}{2\pi i}:\Big(D_\theta X_N^{\pm hj}(z,\theta)-ip^{\pm hj}D_\theta\log\Big(z-z_0 -\begin{pmatrix}1\\0\end{pmatrix}\sqrt{z-z_0}\theta\theta_0\Big)\Big)\exp\left(\frac{-nX_N^{+0}(z,\theta)}{ip^{+0}}\right)(z-z_0\sqrt{z-z_0}\theta\theta_0)^n \times\left(\frac{D_\theta[X_N^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial_z[X_N^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right)\Sigma_{\{h\}} \times(z_0|\theta_0):_R, \tag{2.7}$$

which leads (for $n=1,2,...$) to

$$0=\lim_{z_0\to 0}\mathbf{R}\cdot A_n^{\pm hj}\cdot\Sigma_{\{h\}}(z_0|\theta_0)\cdot|0\rangle=\lim_{z_0\to 0}\mathbf{R}\cdot\mathscr{D}_n^{\pm hj}\cdot\Sigma_{\{h\}}(z_0|\theta_0)\cdot|0\rangle. \tag{2.8}$$

In Eqs. (2.6), (2.7), and hereafter, the integration contour in Eqs. (2.1) and (2.2) has been and will be deformed to satisfy the condition $|z|>|z_0|$ [z_0 being the one in Eq. (2.5)], so that we can use contractions in the R sector.[4] Therefore a complete set of *physical* ket states can be constructed by

$$|\{e_{\not}(n,\pm j)\},\{\tilde{e}_{\not}(n,\pm j)\},\{p\};\theta_0\rangle(\equiv|\not;\theta_0\rangle) \equiv\lim_{z_0\to 0}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(iA_{-n}^{\pm j}/\sqrt{n})^{e_{\not}(n,\pm j)}}{\sqrt{e_{\not}(n,\pm j)!}}(i\mathscr{D}_{-n}^{\pm j})^{\tilde{e}_{\not}(n,\pm j)}\right)\Sigma_{\{h\}}(z_0|\theta_0)\cdot|0\rangle, \tag{2.9}$$

$e_{\not}(n,\pm j)$ being some non-negative integer and $\tilde{e}_{\not}(n,\pm j)$ is either 0 or 1, giving quantum numbers of $|\not;\theta_0\rangle$. Hereafter, $\{p\}$ in Eq. (2.5) is assumed to satisfy the *on-shell* condition

$$p^{+0}p^{-0}+\sum_{j=1}^{4}p^{+j}p^{-j}=0, \tag{2.10}$$

leading to the fact that the *basic* spinorial $\Sigma_{\{h\}}(z_0|0)$ in Eq. (2.9) is the primary operator of conformal weight 5/8. In the following, we shall prove the *theorem*

$$\mathbf{R}\cdot\exp\left(\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4}\left(\frac{1}{n}\alpha_{-n}^{\pm j}\cdot A_{-n}^{\pm j}-d_{-n}^{\mp j}\cdot\mathscr{D}_{-n}^{\pm j}\right)\right)\cdot\Sigma_{\{h\}}(z_0|\theta_0) =:\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)-i\mathscr{P}^{-0}X^{+0}(z_0))\exp(\rho_N^{1F}(z_0,\theta_0;P_F^\dagger) +\rho_{N\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)):_R S_{\{h\}}(z_0)(\equiv\tilde{\Phi}_{N\{h\}}^{F}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)). \tag{2.11}$$

In Eq. (2.11), we have used $\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)$ (1.12) and $\rho_{N\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)$ (1.13), while we have introduced the *operator* $\mathscr{P}^{-0}$ satisfying

$$\mathscr{P}^{-0}|\not{k};\theta_0\rangle=\mathscr{P}_{\not{k}}^{-0}|\not{k};\theta_0\rangle, \tag{2.12}$$

where $|\not{k};\theta_0\rangle$ is the state given by Eq. (2.9) and $\mathscr{P}_{\not{k}}^{-0}$ is the eigenvalue given by $\mathscr{P}_{\not{k}}^{-0}\equiv p^{-0}+N_{\not{k}}(p^{+0})^{-1}$ with

$$N_{\not{k}}\equiv\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4} n(e_{\not{k}}(n,\pm j)+\tilde{e}_{\not{k}}(n,\pm j)). \tag{2.12'}$$

Therefore we have the following mass-shell condition for the state $|\not{k};\theta_0\rangle$ (2.9);

$$p^{+0}\mathscr{P}_{\not{k}}^{-0}+\sum_{j=1}^{4} p^{+j}p^{-j}=N_{\not{k}}. \tag{2.12''}$$

Proof: The left-hand side of Eq. (2.11) can be calculated by

$$\begin{aligned}
&\mathbf{R}\cdot\exp\left(\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4}\frac{1}{n}\alpha_{-n}^{\mp j}\cdot A_{-n}^{\pm j}\exp\left(\frac{-nX^{+0}(z_0)}{ip^{+0}}\right)\right.\\
&\left.-\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4} d_{-n}^{\mp j}\cdot\mathscr{D}_{-n}^{\pm j}\exp\left(\frac{-nX^{+0}(z_0)}{ip^{+0}}\right)\right)\cdot S_{\{h\}}(z_0|\theta_0)\\
&=\mathbf{R}\cdot\exp\left(\sum_{\pm}\sum_{j=1}^{4} i\oint_0\frac{dw\,d\xi}{2\pi i}P_F^{\mp j\dagger}(w,\xi)\oint\frac{dz\,d\theta}{2\pi i}D_\theta X^{\pm j}(z,\theta)\cdot U_F(w,\xi;z,\theta)\right)\\
&\times S_{\{h\}}(z_0|\theta_0).
\end{aligned} \tag{2.13}$$

In the last step of Eq. (2.13) we have introduced the following non-N operator U_F defined by

$$\begin{aligned}
U_F(w,\xi;z,\theta)\equiv&-\sum_{n=1}^{\infty}\frac{w^n}{n}\cdot\exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}-\frac{nX^{+0}(z_0)}{ip^{+0}}\right)\\
&-\xi\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial[X^{+0}(z,\theta)/ip^{+0}])^{1/2}}\left(\sum_{n=1}^{\infty}w^{n-1/2}\exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}-\frac{nX^{+0}(z_0)}{ip^{+0}}\right)\right)
\end{aligned} \tag{2.13'}$$

and the *excited* spinorial $S_{\{h\}}(z_0|\theta_0)$ defined by

$$S_{\{h\}}(z_0|\theta_0)\equiv:\exp(\sigma_{N+}^{0F}(z_0,\theta_0)):_R S_{\{h\}}(z_0), \tag{2.14}$$

with

$$\begin{aligned}
\sigma_{N+}^{0F}(z_0,\theta_0)\equiv&-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)-i\mathscr{P}^{-0}X^{+0}(z_0)-\sum_{j=1}^{4}ip^{hj}Y_{Nh}^{-hj}(z_0,\theta_0)-\sum_{j=1}^{4}ip^{-hj}X^{+hj}(z_0)\\
=&\tilde{\rho}_{N+}^{0F}(z_0,\theta_0)-\sum_{j=1}^{4}ip^{hj}Y_{Nh}^{-hj}(z_0,\theta_0)-\sum_{j=1}^{4}ip^{-hj}X^{+hj}(z_0).
\end{aligned} \tag{2.14'}$$

At this stage, we superconfomally map the c number (w,ξ) into the q number (z_F,θ_F) by using Eqs. (1.18) and (1.19) [changed into the non-N operator's form]. Then we have the radial ordered product among any operator $X^{+0}(z,\theta)$ in $U_F(w,\xi;z,\theta)$ (2.13') and any $Y_{N+}^{-0}(z_0,\theta_0)$ (1.11') in $S_{\{h\}}(z_0|\theta_0)$ (2.14). [Since $w=w(z_F,\theta_F)$ is the c number, any operator originating from $w=w(z_F,\theta_F)$ can be treated as the operator, which has already been normally ordered with respect to $Y_{N+}^{-0}(z_0,\theta_0)$.] We express this radial ordered product in terms of the normal ordered product in the R sector. The result is given by

$$\mathbf{R}\cdot U_F(w,\xi;z,\theta)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0))=:\widetilde{U}_F(w,\xi;z,\theta)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R \tag{2.15}$$

with

$$\begin{aligned}\widetilde{U}_F(w,\xi;z,\theta)\equiv&-\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{z_F-z_0-\sqrt{z_F-z_0}\theta_F\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)^n\cdot\exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}-\frac{nX^{+0}(z_F,\theta_F)}{ip^{+0}}\right)\\&-\xi\left(\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right)\\&\times\left(\sum_{n=1}^{\infty}w^{-1/2}\left(\frac{z_F-z_0-\sqrt{z_F-z_0}\theta_F\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)^n\exp\left(\frac{nX^{+0}(z,\theta)}{ip^{+0}}-\frac{nX^{+0}(z_F,\theta_F)}{ip^{+0}}\right)\right).\end{aligned} \tag{2.16}$$

Therefore the radial ordered product [in Eq. (2.13)] among $X^{\pm hj}(z,\theta),U_F(w,\xi;z,\theta)$ (2.13′), and any $\sigma_{N+}^{0F}(z_0,\theta_0)$ (2.14′) (in $S_{\{h\}}(z_0|\theta_0)$ (2.14)) is expressed [in terms of the normal ordered product] by

$$\begin{aligned}&\mathbf{R}\cdot D_\theta X^{\pm hj}(z,\theta)\cdot U_F(w,\xi;z,\theta):\exp(\sigma_{N+}^{0F}(z_0,\theta_0)):_R\\&=:\left(D_\theta X^{\pm hj}(z,\theta)-ip^{\pm hj}D_\theta\log\left(z-z_0-\begin{pmatrix}1\\0\end{pmatrix}\sqrt{z-z_0}\theta\theta_0\right)\right)\\&\quad\times\widetilde{U}_F(w,\xi;z,\theta)\exp(\sigma_{N+}^{0F}(z_0,\theta_0)):_R.\end{aligned} \tag{2.17}$$

Finally, applying N-izing given by Eq. (1.5), we find that

$$\begin{aligned}&\mathbf{R}\cdot D_\theta X^{\pm hj}(z,\theta)\cdot U_F(w,\xi;z,\theta)S_{\{h\}}(z_0|\theta_0)\\&=:\left(D_\theta X_N^{\pm hj}(z,\theta)-ip^{\pm hj}D_\theta\log\left(z-z_0-\begin{pmatrix}1\\0\end{pmatrix}\sqrt{z-z_0}\theta\theta_0\right)\right)\\&\quad\times\widetilde{U}_{NF}(w,\xi;z,\theta)):_RS_{\{h\}}(z_0|\theta_0),\end{aligned} \tag{2.18}$$

where the N operator $\widetilde{U}_{NF}(w,\xi;z,\theta)$ is N-ized from the non-N operator $\widetilde{U}_F(w,\xi;z,\theta)$ (2.16). In Eq. (2.18) we shall use

$$p^{\pm j}\left(\frac{1}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)^n D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)=-\frac{1}{n}p^{\pm j}D_\theta\left(\frac{1}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)^n \tag{2.19}$$

and

$$p^{\pm j}\left(\frac{1}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)^n D_\theta\log(z-z_0)=-\frac{1}{n}p^{\pm j}\theta\partial_z\left(\frac{1}{z-z_0}\right)^n, \tag{2.20}$$

and then we partially integrate Eq. (2.13), which is thus found out to be given by

$$\begin{aligned}:\Bigg(\mathbf{R}\cdot\exp\Bigg(\sum_{\pm}\sum_{j=1}^{4}i\oint_0\frac{dw\,d\xi}{2\pi i}P_F^{\mp j^\dagger}(w,\xi)\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\widetilde{H}_{NF}^{\pm j}(z,\theta)\Bigg)\\\times\exp(\sigma_{N+}^{0F}(z_0,\theta_0))\Bigg):_RS_{\{h\}}(z_0),\end{aligned} \tag{2.21}$$

where the N operators $\widetilde{H}_{NF}^{\pm hj}(z,\theta)$'s are given, respectively, by

$$\widetilde{H}_{NF}^{+hj}(z,\theta)=\left(D_\theta\left(X_N^{+hj}(z,\theta)-\frac{p^{+hj}}{p^{+0}}X_N^{+0}(z,\theta)\right)\right)\log(\widetilde{u}_{NF}(w,\xi;z,\theta))$$
$$+ip^{+hj}\xi w^{-1/2}D_\theta\left(\frac{D_\theta[X_N^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial_z[X_N^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right)$$
$$\times\log(\widetilde{v}_{NF}(w,\xi;z,\theta)) \tag{2.22}$$

and

$$\widetilde{H}_{NF}^{-hj}(z,\theta)=\left(D_\theta X_N^{-hj}(z,\theta)-\frac{p^{-hj}}{p^{+0}}\cdot\theta\partial_z X_N^{+0}(z,\theta)\right)\log(\widetilde{u}_{NF}(w,\xi;z,\theta))+ip^{-hj}\cdot\xi w^{-1/2}\theta$$
$$\cdot\partial_z\left(\frac{D_\theta[X_N^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial_z[X_N^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right)$$
$$\times\log(\widetilde{v}_{NF}(w,\xi;z,\theta)). \tag{2.23}$$

[We should notice that Eq. (2.21) has the radial ordering among N operators $\widetilde{H}_{NF}^{\pm hj}(z,\theta)$'s, while we have the normal ordered product in the R sector (denoted by $:\ :_R$) among any $\widetilde{H}_{NF}^{\pm hj}(z,\theta)$ and $S_{\{h\}}(z_0|\theta_0)$ (2.14).] In Eqs. (2.22) and (2.23), $\widetilde{u}_{NF}(w,\xi;z,\theta)$ and $\widetilde{v}_{NF}(w,\xi;z,\theta)$ are N operators obtained by N-izing [see Eq. (1.5)] the following non-N operators $\widetilde{u}_F(w,\xi;z,\theta)$ and $\widetilde{v}_F(w,\xi;z,\theta)$, respectively;

$$\widetilde{u}_F(w,\xi;z,\theta)\equiv 1-\frac{z_F-z_0-\sqrt{z_F-z_0}\theta_F\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\exp\left(\frac{X^{+0}(z,\theta)}{ip^{+0}}-\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}\right)$$
$$-\xi w^{-1/2}\left(\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial[X^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right)$$
$$\times\exp\left(\frac{X^{+0}(z,\theta)}{ip^{+0}}-\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}\right) \tag{2.24}$$

and

$$\widetilde{v}_F(w,\xi;z,\theta)\equiv 1-\frac{z_F-z_0-\sqrt{z_F-z_0}\theta_F\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\cdot\exp\left(\frac{X^{+0}(z,\theta)}{ip^{+0}}-\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}\right). \tag{2.25}$$

[Incidentally, (z,θ) is just the *c number* integration variable, while the q number (z_F,θ_F) is the function of (w,ξ).] First, we find in the neighborhood of $z=z_0$ and $\theta=\theta_0$ that

$$\log(\widetilde{u}_F(w,\xi;z,\theta))=-\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)+\mathrm{O}(1) \tag{2.26}$$

and

$$\log(\widetilde{v}_F(w,\xi;z,\theta))=-\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)+\mathrm{O}(1). \tag{2.27}$$

Furthermore, we find from Eq. (1.18) [changed into non-N operator's form] that

$$f^F(z_F,\theta_F|z_0,\theta_0)\equiv\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}-\log(z_F-z_0-\sqrt{z_F-z_0}\theta_F\theta_0)=-\log w+\frac{X^{+0}(z_0)}{ip^{+0}}. \tag{2.28}$$

By using

$$D_{\theta_F}=\frac{1}{D_\xi\theta_F}D_\xi \tag{2.29}$$

we find successively the following formulas:

$$D_{\theta_F} f^F(z_F,\theta_F|z_0,\theta_0) = -\frac{\xi}{wD_\xi\theta_F} = -\frac{\xi}{wK_F(w)}, \tag{2.30}$$

$$\partial_{z_F} f^F(z_F,\theta_F|z_0,\theta_0) = -\frac{1}{w(D_\xi\theta_F)^2} - \frac{\xi\partial_w\theta_F}{w(D_\xi\theta_F)^3} = -\frac{1}{wK_F^2(w)} + \frac{\xi\partial_w J_F(w)}{wK_F^3(w)}, \tag{2.31}$$

and

$$\begin{aligned}\partial_{z_F} D_{\theta_F} f^F(z_F,\theta_F|z_0,\theta_0) &= \frac{\partial_w\theta_F}{w(D_\xi\theta_F)^4} + \xi\left(\frac{\partial_w\theta_F}{w^2(D_\xi\theta_F)^3} + \frac{\partial_w D_\xi\theta_F}{w(D_\xi\theta_F)^4}\right)\\ &= \frac{\partial_w J_F(w)}{wK_F^4(w)} + \xi\left(\frac{2\partial_w K_F(w)}{wK_F^4(w)} + \frac{1}{w^2K_F^3(w)}\right).\end{aligned} \tag{2.32}$$

Then by using the Taylor expansion at $z=z_F$ and $\theta=\theta_F$ [see Eq. (3.4) in Ref. 4], the first factor on the right-hand side of Eq. (2.24) is found to be given by

$$\begin{aligned}\tilde{v}_F(w,\xi;z,\theta) &= (\theta-\theta_F)\frac{\xi}{wK_F(w)} + (z-z_F-\theta\theta_F)\left(\frac{1}{wK_F^2(w)} - \frac{\xi\partial_w J_F(w)}{wK_F^3(w)}\right)\\ &\quad + \mathrm{O}((\theta-\theta_F)(z-z_F-\theta\theta_F)).\end{aligned} \tag{2.33}$$

On the other hand, the second one in Eq. (2.24) is found to be given by

$$-(\theta-\theta_F)\frac{\xi}{wK_F(w)} - (z-z_F-\theta\theta_F)\left(\frac{\xi\partial_w J_F(w)}{wK_F^3(w)}\right) + \mathrm{O}((\theta-\theta_F)(z-z_F-\theta\theta_F)). \tag{2.34}$$

Summing Eqs. (2.33) and (2.34), we finally obtain that

$$\tilde{u}_F(w,\xi;z,\theta) = (z-z_F-\theta\theta_F)\left(\frac{1}{wK_F^2(w)} - \frac{2\xi\partial_w J_F(w)}{wK_F^3(w)}\right) + \mathrm{O}((\theta-\theta_F)(z-z_F-\theta\theta_F)). \tag{2.35}$$

Substituting N-ized forms of Eqs. (2.26), (2.27), (2.33), and (2.35) into the N operator $\tilde{H}_{NF}^{+hj}(z,\theta)$ (2.22), we find by the partial integration that

$$\begin{aligned}\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\,\tilde{H}_{NF}^{+hj}(z,\theta) &= -\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\,(X^{+hj}(z)+\theta\sqrt{z-z_0}\Psi_{Nh}^{+hj}(z))\left(\frac{\theta-\theta_{NF}}{z-z_{NF}-\theta\theta_{NF}}\right.\\ &\quad\left.-\frac{\theta-\sqrt{z-z_0}\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right) + \frac{p^{+hj}}{p^{+0}}\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\,(X^{+0}(z)\\ &\quad + \theta\sqrt{z-z_0}\Psi_{N+}^{+0}(z))\left(\frac{\theta-\theta_{NF}}{z-z_{NF}-\theta\theta_{NF}} - \frac{\theta-\sqrt{z-z_0}\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)\\ &= -X_N^{+hj}(z_{NF},\theta_{NF}) + X^{+hj}(z_0) + \frac{p^{+hj}}{p^{+0}}\,(X_N^{+0}(z_{NF},\theta_{NF}) - X^{+0}(z_0)).\end{aligned} \tag{2.36}$$

By similar calculations, we find from the N operator $\tilde{H}_{NF}^{-hj}(z,\theta)$ (2.23) that

$$\oint_{|z|>|z_0|}\frac{dz\ d\theta}{2\pi i}\,\widetilde{H}_{NF}^{-hj}(z,\theta)=-\oint_{|z|>|z_0|}\frac{dz\ d\theta}{2\pi i}\left(X^{-hj}(z)+\theta\,\frac{\Psi_{Nh}^{-hj}(z)}{\sqrt{z-z_0}}\right)\left(\frac{\theta-\theta_{NF}}{z-z_{NF}-\theta\theta_{NF}}\right.$$
$$\left.-\frac{\theta-\sqrt{z-z_0}\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}\right)+\frac{p^{-hj}}{p^{+0}}\oint_{|z|>|z_0|}\frac{dz\ d\theta}{2\pi i}$$
$$\times(X^{+0}(z)+\theta\sqrt{z-z_0}\Psi_{N+}^{+0}(z))\left(\frac{\theta}{z-z_{NF}}-\frac{\theta}{z-z_0}\right)$$
$$=-X_N^{-hj}(z_{NF},\theta_{NF})+Y_{Nh}^{-hj}(z_0,\theta_0)+\frac{p^{-hj}}{p^{+0}}\,(X^{+0}(z_{NF})-X^{+0}(z_0)),\tag{2.37}$$

where we have noticed from Eqs. (1.19) [changed into non-N operator's form] and (2.16) that

$$\left[\frac{D[X^{+0}(z,\theta)/ip^{+0}]-D_\theta\log(z-z_0-\sqrt{z-z_0}\theta\theta_0)}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}]+\partial_z\log(z-z_0-\sqrt{z-z_0}\theta\theta_0))^{1/2}}\right]_{z=z_F,\,\theta=\theta_F}\propto\xi,\tag{2.38}$$

together with

$$D_\theta\log(\widetilde{v}_F(w,\xi;z,\theta))=-\frac{\theta-\sqrt{z-z_0}\theta_0}{z-z_0-\sqrt{z-z_0}\theta\theta_0}+D_\theta\frac{X^{+0}(z,\theta)}{ip^{+0}}-\frac{\theta-\sqrt{z-z_0}\theta_0}{w}$$
$$+\mathrm{O}(z-z_0-\sqrt{z-z_0}\theta\theta_0).\tag{2.39}$$

With the help of the decomposition [into zero and nonzero modes]

$$P_F^{\mp j\dagger}(w,\xi)\equiv\frac{p^{\mp j}\xi}{w}+\mathscr{P}_F^{\mp j\dagger}(w,\xi)\tag{2.40}$$

we find [with respect to the integrand given by the right-hand side of Eq. (2.37)] that

$$-\sum_{j=1}^{4}i\oint_0\frac{dw\ d\xi}{2\pi i}\,P_F^{+hj\dagger}(w,\xi)\left(X_N^{-hj}(z_{NF}(w,\xi))-X^{-hj}(z_0)-\frac{p^{-hj}}{p^{+0}}\,(X^{+0}(z_{NF}(w,\xi))\right.$$
$$\left.-X^{+0}(z_0))+\theta_{NF}(w,\xi)\,\frac{\Psi_{Nh}^{-hj}(z_{NF}(w,\xi))}{(z_{NF}(w,\xi)-z_0)^{1/2}}-\theta_0\Psi_{Nh}^{-hj}(z_0)\right)$$
$$\equiv\sum_{j=1}^{4}\left(D_{F1}^{+hj}(\mathscr{P}_F^{+hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{-hj})-\frac{p^{-hj}}{p^{+0}}\,D_{F1}^{+hj}(\mathscr{P}_F^{+hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{+0})\right)$$
$$+\sum_{j=1}^{4}D_{F2}^{+hj}(\mathscr{P}_F^{+hj\dagger};X^{+0},\Psi_{N+}^{+0}|\Psi_{Nh}^{-hj})\tag{2.41}$$

and we find [with respect to the integrand given by the right-hand side of Eq. (2.36)] that

$$
\begin{aligned}
&-\sum_{j=1}^{4} i \oint_0 \frac{dw\ d\xi}{2\pi i} P_F^{-hj\dagger}(w,\xi)\Big(X^{+hj}(z_{NF}(w,\xi))-X^{+hj}(z_0)-\frac{p^{+hj}}{p^{+0}}\,(X^{+0}(z_{NF}(w,\xi))-X^{+0}(z_0)) \\
&\quad+\theta_{NF}(w,\xi)(z_{NF}(w,\xi)-z_0)^{1/2}\Psi_{Nh}^{+hj}(z_{NF}(w,\xi)) \\
&\quad-\frac{p^{+hj}}{p^{+0}}\cdot\theta_{NF}(w,\xi)(z_{NF}(w,\xi)-z_0)^{1/2}\Psi_{N+}^{+0}(z_{NF}(w,\xi))\Big) \\
&\equiv\sum_{j=1}^{4}\left(D_{F1}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{+hj})-\frac{p^{+hj}}{p^{+0}}\,D_{F1}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{+0})\right) \\
&\quad+\sum_{j=1}^{4}\left(D_{F3}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|\Psi_{Nh}^{+hj})-\frac{p^{+hj}}{p^{+0}}\,D_{F3}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|\Psi_{N+}^{+0})\right). \qquad (2.42)
\end{aligned}
$$

[Thus we find from Eqs. (2.41) and (2.42) that

$$
\begin{aligned}
\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)&\equiv-\sum_{j=1}^{4} i\oint_0\frac{dw\ d\xi}{2\pi i}P_F^{\mp hj\dagger}(w,\xi)X_N^{\pm j}(z_{NF}(w,\xi),\theta_{NF}(w,\xi)) \\
&=-\sum_{j=1}^{4} ip^{+hj}Y_{Nh}^{-hj}(z_0,\theta_0)-\sum_{j=1}^{4} ip^{-hj}X^{+hj}(z_0) \\
&\quad+\sum_{\pm}\sum_{j=1}^{4} D_{F1}^{\pm hj}(\mathscr{P}_F^{\pm hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{\mp hj})+\sum_{j=1}^{4} D_{F2}^{+hj}(\mathscr{P}_F^{+hj\dagger};X^{+0}\Psi_{N+}^{+0}|\Psi_{Nh}^{-hj}) \\
&\quad+\sum_{j=1}^{4} D_{F3}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|\Psi_{Nh}^{+hj}). \qquad (2.43)
\end{aligned}
$$

Therefore with the help of Eqs. (2.13), (2.21), (2.36), (2.37), and (2.41)–(2.43), we find that Eq. (2.13) is reduced to be given by

$$
:(\mathbf{R}\cdot\ \exp(\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)))\exp(\tilde{\rho}_{N+}^{0F}(z_0,\theta_0)+\sigma_N^{1F}(z_0,\theta_0;P_F^\dagger)):_R S_{\{h\}}(z_0), \qquad (2.44)
$$

with

$$
\begin{aligned}
\sigma_N^{1F}(z_0,\theta_0;P_F^\dagger)&\equiv-\sum_{\pm}\ \frac{p^{\mp hj}}{p^{+0}}\,D_{F1}^{\pm hj}(\mathscr{P}_F^{\pm hj\dagger};X^{+0},\Psi_{N+}^{+0}|X^{+0}) \\
&\quad-\frac{p^{+hj}}{p^{+0}}\,D_{F3}^{-hj}(\mathscr{P}_F^{-hj\dagger};X^{+0},\Psi_{N+}^{+0}|\Psi_{N+}^{+0}). \qquad (2.45)
\end{aligned}
$$

$D_{F(i)}^{\mp j}$'s (for $i=1-3$) in Eq. (2.43) are bilinear functionals of $\mathscr{P}_F^{\mp j\dagger}$ and $X^{\pm j}(\Psi_{Nh}^{\pm j})$. Therefore in Eq. (2.44), we have still the radial ordered product among N operators $\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)$ (2.43). Hereafter, we express Eq. (2.44) in terms of overally normal-ordered products, by taking contractions in the R sector among $(X^{\pm j},\Psi_{Nh}^{\pm j})$. We denote thus obtained contracted terms by $E_{NF}(\mathscr{P}_F^{-j\dagger},\mathscr{P}_F^{+j\dagger}|X^{+0},\Psi_{N+}^{+0})$ [in N operator's form], which are bilinear terms in $\mathscr{P}_F^{-j\dagger}$ and $\mathscr{P}_F^{+j\dagger}$ (not including $p^{\mp j}$). [See the corresponding Eq. (2.11) in Ref. 7.] With the help of $\rho_{N\{h\}}^{2F}$ (1.13) and ρ_N^{1F} (2.43), we can show that E_{NF} is nothing but the one existing in the following formula:

$$
\begin{aligned}
&\rho^{2F}_{N\{h\}}(z_0,\theta_0;P^{\dagger}_F)\\
&\equiv -\sum_{j=1}^{4}\oint\frac{dw\ d\xi}{2\pi i}\left(\frac{p^{-j}\xi}{w}+\mathscr{P}_F^{-j^{\dagger}}(w,\xi)\right)\oint\frac{dw'\ d\xi'}{2\pi i}\left(\frac{p^{+j}\xi'}{w'}+\mathscr{P}_F^{+j^{\dagger}}(w',\xi')\right)\\
&\quad\times\log\frac{z_{NF}(w,\xi)-z_{NF}(w',\xi')-([z_{NF}(w,\xi)-z_0]/[z_{NF}(w',\xi')-z_0])^{\epsilon_h^j}\theta_{NF}(w,\xi)\theta_{NF}(w',\xi')}{w-w'-(w/w')^{\epsilon_h^j}\xi\xi'}\\
&=-\sum_{j=1}^{4}p^{+hj}\oint\frac{dw\ d\xi}{2\pi i}\mathscr{P}_F^{-hj^{\dagger}}(w,\xi)\cdot\log\frac{z_{NF}(w,\xi)-z_0-(z_{NF}(w,\xi)-z_0)^{1/2}\theta_{NF}(w,\xi)\theta_0}{w}\\
&\quad-\sum_{j=1}^{4}p^{-hj}\oint\frac{dw\ d\xi}{2\pi i}\mathscr{P}_F^{+hj^{\dagger}}(w,\xi)\cdot\log\frac{z_{NF}(w,\xi)-z_0}{w}\\
&\quad-\sum_{j=1}^{4}\oint\frac{dw\ d\xi}{2\pi i}\mathscr{P}_F^{-j^{\dagger}}(w,\xi)\oint\frac{dw'\ d\xi'}{2\pi i}\mathscr{P}_F^{+j^{\dagger}}(w',\xi')\log\Bigg(z_{NF}(w,\xi)-z_{NF}(w',\xi')\\
&\quad-\left(\frac{z_{NF}(w,\xi)-z_0}{z_{NF}(w',\xi')-z_0}\right)^{\epsilon_h^j}\theta_{NF}(w,\xi)\theta_{NF}(w',\xi')\Bigg)\\
&=\sigma_N^{1F}(z_0,\theta_0;P^{\dagger}_F)+\sum_{j=1}^{4}E_{NF}(\mathscr{P}_F^{-hj^{\dagger}},\mathscr{P}_F^{+hj^{\dagger}}|X^{+0},\Psi^{+0}_{N+}),
\end{aligned}
\tag{2.46}
$$

$\sigma_N^{1F}(z_0,\theta_0;P_F^{\dagger})$ being given by Eq. (2.45). In obtaining the last N operator E_{NF} in Eq. (2.46), we have used the following contraction in the R sector;

$$
\begin{aligned}
&\overbrace{\qquad\qquad R\qquad\qquad}\\
&X_N^{+j}(z_{NF}(w,\xi),\theta_{NF}(w,\xi))X_N^{-j}(z_{NF}(w',\xi'),\theta_{NF}(w',\xi'))\\
&\quad=\log\left(z_{NF}(w,\xi)-z_{NF}(w',\xi')-\left(\frac{z_{NF}(w,\xi)-z_0}{z_{NF}(w',\xi')-z_0}\right)^{\varepsilon_h^j}\theta_{NF}(w,\xi)\theta_{NF}(w',\xi')\right).
\end{aligned}
\tag{2.47}
$$

[Incidentally, we have also used in Eq. (2.46) that

$$
\log\frac{z_{NF}(w,\xi)-z_0}{w}=\frac{X^{+0}(z_{NF}(w,\xi))}{ip^{+0}}-\frac{X^{+0}(z_0)}{ip^{+0}},
\tag{2.48}
$$

which will be shown to be valid by using various formulas to be derived in the next section. First, we have from Eq. (1.17) that

$$
\log\frac{z_{NF}(w,\xi)-z_0}{w}=\log\frac{L_{NF}(w)}{w}+\xi K_{NF}(w)\frac{J_{NF}(w)}{L_{NF}(w)}.
\tag{2.49}
$$

Furthermore, comparing Eqs. (3.15) with (B17) (for $j=0$) leads to

$$
\log\frac{L_{NF}(w)}{w}=\left[\frac{X^{+0}(z_{NF})}{ip^{+0}}\right]_{\xi=0}-\frac{X^{+0}(z_0)}{ip^{+0}},
\tag{2.50}
$$

while comparing Eqs. (3.5), (3.19), and (3.20) with (B18) (for $j=0$) leads to

$$
\frac{J_{NF}(w)}{L_{NF}(w)}=\Gamma_N^{F2}(0,w)=\left[\theta_{NF}\partial_{z_{NF}}\frac{X^{+0}(z_{NF})}{ip^{+0}}\right]_{\xi=0}.
\tag{2.51}
$$

Therefore, Eqs. (2.49)–(2.51) together with Eq. (1.15) finally gives Eq. (2.48).] Therefore we find from Eqs. (2.44) and (2.46) that

$$
\begin{aligned}
&:(\mathbf{R}\cdot\exp(\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)))\exp(\sigma_N^{1F}(z_0,\theta_0;P_F')):_R \\
&=:\exp\left(\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)+\sigma_N^{1F}(z_0,\theta_0;P_F^\dagger)+\sum_{j=1}^{4} E_{NF}(\mathscr{P}_F^{+hj^\dagger},\mathscr{P}_F^{-hj^\dagger}|X^{+0},\Psi_{N+}^{+0})\right):_R \\
&=:\exp(\rho_N^{1F}(z_0,\theta_0;P_F^\dagger)+\rho_{N\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)):_R . \qquad (2.52)
\end{aligned}
$$

In conclusion, we find from Eqs. (2.44) and (2.52) that the left-hand side of Eq. (2.11) is given (in overally normal-ordered form) by $\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ [i.e., Eqs. (1.10) and (1.14)] in the case of ''on-shell,'' provided we notice the fact that $\mathscr{P}^{-0}$ in $\tilde{\rho}_{N+}^{0F}$ [see Eq. (2.14′)] is just equal to p^{-0} in Eq. (1.11). [See also the on-shell condition (2.12″).] Without imposing the on-shell condition (2.10), we can easily prove Eq. (2.11) even in the off-shell case. (Q.E.D.)]

In the following, we shall prove OPE of GFPVO $\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ with respect to the superstressino operator $(T_{NF},\mathscr{T}_{NF})$ (in the off-shell case); With respect to the stress operator $T_{NF}(z';z_0)$ which is N-ized [see Eq. (1.5)] from

$$
T_F(z')\equiv\sum_{j=0}^{4}\sum_{\pm}:\left(\frac{1}{2}\partial_{z'}X^{\pm j}(z')\partial_{z'}X^{\mp j}(z')+\frac{1}{2}\partial_{z'}\Psi^{\pm j}(z')\cdot\Psi^{\mp j}(z')\right):_R , \qquad (2.53)
$$

we easily find from Eq. (2.5) that

$$
\begin{aligned}
\mathbf{R}\cdot T_{NF}(z';z_0)\sum_{\{h\}}(z_0|\theta_0)&=\frac{-(p^{+0}p^{-0}+\Sigma_{j=1}^{4}p^{+j}p^{-j})}{(z'-z_0)^2}\cdot\sum_{\{h\}}(z_0|\theta_0)+\frac{1}{z'-z_0}\partial_{z_0}\sum_{\{h\}} \\
&\quad\times(z_0|\theta_0)+\mathrm{O}(1). \qquad (2.54)
\end{aligned}
$$

Applying OPE (2.54) to the left-hand side of Eq. (2.11), and taking account of mass relations (2.12)–(2.12″) and (anti-) commutation relations (2.4), we find that

$$
\begin{aligned}
&\mathbf{R}\cdot T_{NF}(z';z_0)\tilde{\Phi}_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0) \\
&=\frac{1}{(z'-z_0)^2}\left(-\left(p^{+0}\mathscr{P}^{-0}+\sum_{j=1}^{4}p^{+j}p^{-j}\right)\tilde{\Phi}_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\right. \\
&\quad\left.+[N_F,\tilde{\Phi}_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)]\right) \\
&\quad+\frac{1}{z'-z_0}\partial_{z_0}(\tilde{\Phi}_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0))+O(1), \qquad (2.55)
\end{aligned}
$$

where we have used the light-cone number operator N_F defined by

$$
N_F\equiv\sum_{n=1}^{\infty}\sum_{\pm}\sum_{k=1}^{4}(\alpha_{-n}^{\mp k}\alpha_n^{\pm k}+nd_{-n}^{\mp k}d_n^{\pm k}). \qquad (2.56)
$$

On the other hand, with respect to the stressino operator $\mathscr{T}_{NF}(z';z_0)$ which is N-ized [see Eq. (1.5)] from

$$
\mathscr{T}_F(z')\equiv\sum_{j=0}^{4}\sum_{\pm}\partial_{z'}X^{\pm j}(z')\Psi^{\mp j}(z'), \qquad (2.57)
$$

we can easily find from Eq. (2.5) that

$$
\mathbf{R}\cdot\mathscr{T}_{NF}(z';z_0)\Sigma_{\{h\}}^0(z_0)=\frac{1}{(\sqrt{z'-z_0})^3}\Sigma_{\{h\}}^1(z_0)+O\left(\frac{1}{\sqrt{z'-z_0}}\right) \qquad (2.58)
$$

and

$$\mathbf{R}\cdot\mathscr{T}_{NF}(z';z_0)\Sigma^1_{\{h\}}(z_0)=\frac{-(p^{+0}p^{-0}+\Sigma_{j=1}^4 p^{+j}p^{-j})}{(\sqrt{z'-z_0})^3}\cdot\Sigma^0_{\{h\}}(z_0)+O\left(\frac{1}{\sqrt{z'-z_0}}\right), \quad (2.59)$$

where we have introduced

$$\Sigma_{\{h\}}(z_0|\theta_0)\equiv\Sigma^0_{\{h\}}(z_0)+\theta_0\Sigma^1_{\{h\}}(z_0). \quad (2.60)$$

On the other hand, Eq. (2.11) is rewritten into

$$\begin{aligned}&\mathbf{R}\cdot\ \exp\left(\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4}\left(\frac{1}{n}\,\alpha_{-n}^{\mp j}\cdot A_{-n}^{\pm j}-d_{-n}^{\mp j}\cdot\mathscr{D}_{-n}^{\pm j}\right)\right)\cdot\Sigma^m_{\{h\}}(z_0)\\&\quad=\widetilde{\Phi}^{Fm}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)\\&\qquad[\text{given by Eqs. (1.4) and (2.11), i.e., Eq. (1.10) with } \sim][\text{for } m=0,1].\end{aligned} \quad (2.61)$$

Applying OPE's (2.58) and (2.59) to the left-hand side of Eq. (2.61) and taking account of mass relations (2.12)–(2.12″) and (anti-) commutation relations (2.4), we find that

$$\mathbf{R}\cdot\mathscr{T}_{NF}(z';z_0)\widetilde{\Phi}^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)=\frac{1}{(\sqrt{z'-z_0})^3}\,\widetilde{\Phi}^{F1}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)+O\left(\frac{1}{\sqrt{z'-z_0}}\right) \quad (2.62)$$

and

$$\begin{aligned}&\mathbf{R}\cdot\mathscr{T}_{NF}(z';z_0)\widetilde{\Phi}^{F1}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)\\&\quad=\frac{1}{(\sqrt{z'-z_0})^3}\left(-\left(p^{+0}\mathscr{P}^{-0}+\sum_{j=1}^{4}p^{+j}p^{-j}\right)\widetilde{\Phi}^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)\right.\\&\qquad\left.+[N_F,\widetilde{\Phi}^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)]\right)+O\left(\frac{1}{\sqrt{z'-z_0}}\right).\end{aligned} \quad (2.63)$$

[It should be noticed that the notation $\mathscr{P}^{-0}$ used in this section is just equal to the notation p^{-0} in Sec. I.]

Although we have exclusively investigated the R sector in this paper, the NS sector can be also analyzed in much the same way.

III. PHYSICAL VERTEX OPERATORS IN THE *R* SECTOR

In this section, we explicitly carry out the loop integrations over $w(w')$ and $\xi(\xi')$ in N operators ρ_N^{1F} (1.12) and $\rho_{N\{h\}}^{2F}$ (1.13), and we will obtain concrete results. With the help of formulas (1.16) and (1.17), we find that

$$\begin{aligned}&G_N^F(w,\xi;w',\xi')\\&\equiv\log\frac{z_{NF}(w,\xi)-z_{NF}(w',\xi')-([z_{NF}(w,\xi)-z_0]/[z_{NF}(w',\xi')-z_0])^{1/2}\theta_{NF}(w,\xi)\,\theta_{NF}(w',\xi')}{w-w'-(w/w')^{1/2}\xi\xi'}\end{aligned} \quad (3.1)$$

$$\begin{aligned}&=A_N^F(w,w')+\xi K_{NF}(w)\Gamma_N^{F1}(w,w')\\&\quad+\xi' K_{NF}(w')\Gamma_N^{F2}(w,w')-\xi\xi'\left(\frac{w}{w'}\right)^{1/2}C_N^F(w,w'),\end{aligned} \quad (3.2)$$

where

$$A_N^F(w,w')\equiv\log\frac{L_{NF}(w)-L_{NF}(w')-(L_{NF}(w)/L_{NF}(w'))^{1/2}J_{NF}(w)J_{NF}(w')}{w-w'} \tag{3.3}$$

$$\equiv\sum_{n,m=0}^{\infty} w^n w'^m A_{N,n,m}^F, \tag{3.3'}$$

$$\Gamma_N^{F1}(w,w')\equiv\sqrt{L_{NF}(w)}\,\frac{[J_{NF}(w)/\sqrt{L_{NF}(w)}]-[J_{NF}(w')/\sqrt{L_{NF}(w')}]}{L_{NF}(w)-L_{NF}(w')} \tag{3.4}$$

$$\equiv\sum_{n=0}^{\infty} w^{n+1/2}\left(\sum_{m=0}^{\infty} w'^m\Gamma_{N,n+1/2,m}^{F1}\right) \tag{3.4'}$$

and

$$\Gamma_N^{F2}(w,w')\equiv\frac{1}{\sqrt{L_{NF}(w')}}\cdot\frac{\sqrt{L_{NF}(w')}J_{NF}(w')-\sqrt{L_{NF}(w)}J_{NF}(w)}{L_{NF}(w')-L_{NF}(w)} \tag{3.5}$$

$$\equiv\sum_{m=-1}^{\infty} w'^{m+1/2}\left(\sum_{n=0}^{\infty} w^n\Gamma_{N,n,m+1/2}^{F2}\right). \tag{3.5'}$$

[We notice that general forms (3.3′)–(3.5′) are concluded from Eqs. (3.3)–(3.5), respectively.] On the other hand, we can calculate $C_N^F(w,w')$ by using

$$C_N^F(w,w')\equiv\frac{\exp(-(w-w')A_N^{F1}(w,w'))\exp((w-w')^2A_N^{F2}(w,w'))-1}{w-w'}. \tag{3.6}$$

In Eq. (3.6), $A_N^{F1(2)}(w,w')$ are functions obtained by

$$A_N^{F1}(w,w')\equiv\frac{A_N^F(w,w')-\log[L_{NF}(w)/w]-A_N^F(w',w)+\log[L_{NF}(w')/w']}{2(w-w')}$$

$$\equiv\sum_{n,m=0}^{\infty} w^n w'^m A_{N,n,m}^{F1} \tag{3.7}$$

and

$$A_N^{F2}(w,w')\equiv\frac{A_N^F(w,w)-A_N^F(w,w')-A_N^F(w',w)+A_N^F(w',w')}{2(w-w')^2}\equiv\sum_{n,m=0}^{\infty} w^n w'^m A_{N,n,m}^{F2}, \tag{3.8}$$

both of which are even under $w\leftrightarrow w'$. In Eq. (3.8), we have used the useful fact that Eqs. (1.28) and (3.3) lead to

$$K_{NF}(w)=\lim_{w'\to w}\left(\frac{L_{NF}(w)-L_{NF}(w')-J_{NF}(w)J_{NF}(w')}{w-w'}\right)^{1/2}=\exp\left(\frac{1}{2}A_N^F(w,w)\right). \tag{3.9}$$

In the following, we shall explain the method of how to calculate Eqs. (3.3)–(3.9) with the help of Eqs. (1.20)–(1.30).

First, Eq. (3.1) is rewritten into

$$G_N^F(w,\xi;w',\xi')=\log\frac{z_{NF}-z'_{NF}}{w-w'-(w/w')^{1/2}\xi\xi'}-\left(\frac{w}{w'}\right)^{1/2}\frac{\theta_{NF}\theta'_{NF}}{z_{NF}-z'_{NF}}\exp\left(-\frac{1}{2}I(z'_{NF})+\frac{1}{2}I(z_{NF})\right) \tag{3.10}$$

$$\begin{aligned}&=-\log\left(w-w'-\left(\frac{w}{w'}\right)^{1/2}\xi\xi'\right)+[\log(z_{NF}-z'_{NF})]_{\xi'=0}+\xi' K_{NF}(w')\\&\times\left[\frac{\theta'_{NF}}{z'_{NF}-z_{NF}}\right]_{\xi'=0}-\left(\frac{w}{w'}\right)^{1/2}\left[\frac{\theta'_{NF}\theta_{NF}}{z'_{NF}-z_{NF}}\exp\left(-\frac{1}{2}I(z'_{NF})+\frac{1}{2}I(z_{NF})\right)\right]_{\xi'=0}\\&-\left(\frac{w}{w'}\right)^{1/2}\xi' K_{NF}(w')\left[\frac{\theta_{NF}}{z'_{NF}-z_{NF}}\exp\left(-\frac{1}{2}I(z'_{NF})+\frac{1}{2}I(z_{NF})\right)\right]_{\xi'=0},\end{aligned} \tag{3.10$'$}$$

where we have used Eqs. (1.17), (1.27), and shorthand notations

$$z_{NF}\equiv z_{NF}(w,\xi),\quad z'_{NF}\equiv z_{NF}(w',\xi'),$$

and

$$\theta_{NF}\equiv\theta_{NF}(w,\xi),\quad \theta'_{NF}\equiv\theta_{NF}(w',\xi'). \tag{3.11}$$

In the case when $|w|>|w'|$, Eq. (3.10$'$) can be rewritten into

$$\begin{aligned}G_N^F(w,\xi;w',\xi')&=[\log(z_{NF}-z'_{NF})]_{\xi'=0}-\left(\frac{w}{w'}\right)^{1/2}\left[\exp\left(-\frac{1}{2}I(z'_{NF})+\frac{1}{2}I(z_{NF})\right)\frac{\theta'_{NF}\theta_{NF}}{z'_{NF}-z_{NF}}\right]_{\xi'=0}\\&-\log w+\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{w'}{w}\right)^n+\xi' K_{NF}(w')\left[\frac{\theta'_{NF}}{z'_{NF}-z_{NF}}\right]_{\xi'=0}-\xi' K_{NF}(w')\\&\times\left(\frac{w}{w'}\right)^{1/2}\left[\frac{\theta_{NF}}{z'_{NF}-z_{NF}}\exp\left(-\frac{1}{2}I(z'_{NF})+\frac{1}{2}I(z_{NF})\right)\right]_{\xi'=0}\\&+\xi\xi'\frac{1}{\sqrt{w'w}}\sum_{n=0}^{\infty}\left(\frac{w'}{w}\right)^n.\end{aligned} \tag{3.12}$$

First, we calculate Eq. (3.12) with the help of Eqs. (1.20), (1.21), and (1.27), using the shorthand notations (1.23) (for $k=0,1/2,1,...$) and

$$V_{k+1}\equiv\frac{1}{k!}\,\partial_{z_0}^k\frac{\Psi_{N+}^{+0}(z_0)}{ip^{+0}}. \tag{3.13}$$

The obtained result (in the zeroth order of both ξ and ξ') is compared with the general form (3.3$'$), leading to

$$\begin{aligned}&\sum_{n,m=0}^{\infty}w^n w'^m A_{N,n,m}^F-\log\frac{L_{NF}(w)}{w}\\&=\sum_{m=1}^{\infty}\frac{w'^m}{m}\left(\sum_{k=1}^{m}R_{m-k}^{(m)}(z_0)\frac{\partial_{z_0}^k I(z_0)}{(k-1)!}\right)+\sum_{m=1}^{\infty}\frac{w'^m}{m}\sum_{n=1}^{\infty}\frac{w^n}{n}\left(\sum_{k=1}^{m}kR_{m-k}^{(m)}(z_0)R_{n+k}^{(n)}(z_0)\right)\\&-\sum_{m=0}^{\infty}w'^m\sum_{n=0}^{\infty}w^n\left(\sum_{k=0}^{m}R_{m-k}^{(m)}(z_0)\left(\sum_{j=1}^{n+k}R_{n+k-j}^{(n)}(z_0)V_j\theta_0\right)\right)+\sum_{m=1}^{\infty}w'^m\sum_{n=0}^{\infty}w^n\\&\times\left(\sum_{k=1}^{m}R_{m-k}^{(m)}(z_0)\left(\sum_{J=1}^{k}R_{n+J-1}^{(n)}(z_0)V_{k-J+1}\theta_0+\sum_{J=2}^{n}R_{n-J}^{(n)}(z_0)\left(\sum_{j=1}^{k}V_{k-j+1}V_{J+j-1}\right)\right)\right)\end{aligned} \tag{3.14}$$

and

$$\log \frac{L_{NF}(w)}{w} = \sum_{n=1}^{\infty} \frac{w^n}{n} R_n^{(n)}(z_0). \tag{3.15}$$

The last term on the right-hand side of Eq. (3.14) has been derived from

$$\sum_{m=1}^{\infty} w'^m \sum_{n=0}^{\infty} w^n \Bigg(\sum_{k=1}^{m} R_{m-k}^{(m+1/2)}(z_0) \Bigg(\sum_{l=1}^{k} V_{k-l+1} \Bigg(\sum_{h=1}^{l} R_{l-h}^{(-1/2)}(z_0) \Bigg(R_{n+h-1}^{(n)}(z_0)\theta_0 + \sum_{j=1}^{n+h-1} R_{n+h-1-j}^{(n)}(z_0) V_j \Bigg) \Bigg) \Bigg) \Bigg), \tag{3.16}$$

which is subsequently rewritten by changing the order of summations in such a way that

$$\sum_{l=1}^{k} \sum_{h=1}^{l} = \sum_{h=1}^{k} \sum_{l=h}^{k}, \quad \text{etc.} \tag{3.17}$$

and by using the formulas such that

$$\sum_{k=l}^{m} R_{m-k}^{(p)}(z_0) R_{k-l}^{(q)}(z_0) = R_{m-l}^{(p+q)}(z_0). \tag{3.18}$$

Furthermore, comparing the coefficient of $\xi' K_{NF}(w')$ on both right-hand sides of Eqs. (3.2) and (3.10′), we find that

$$\Gamma_{N,n,-1/2}^{F2} = R_n^{(n)}(z_0)\theta_0 + \sum_{j=1}^{n} R_{n-j}^{(n)}(z_0) V_j \tag{3.19}$$

and

$$\begin{aligned} \Gamma_{N,n,m+1/2}^{F2} \quad (\text{for } m = 0,1,2,\ldots) = \sum_{k=0}^{m} R_{m-k}^{(m+1/2)}(z_0) \Bigg(& \delta_{n,0} \Bigg(\frac{\theta_0}{k!} \partial_{z_0}^{k+1} I(z_0) + \sum_{l=1}^{k} \frac{V_{k-l+1}}{(l-1)!} \partial_{z_0}^{l} I(z_0) \Bigg) \\ & + (1-\delta_{n,0}) \Bigg(\frac{k+1}{n} R_{n+k+1}^{(n)}(z_0)\theta_0 + \sum_{l=1}^{k} \frac{l}{n} V_{k-l+1} R_{n+l}^{(n)}(z_0) \Bigg) \Bigg) \\ & + \sum_{l=0}^{m+1} \frac{2l-1}{2m+1} \cdot R_{m+1-l}^{(m+1/2)}(z_0) \Bigg(R_{n+l}^{(n)}(z_0)\theta_0 \\ & + \sum_{j=1}^{n+l} R_{n+l-j}^{(n)}(z_0) V_j \Bigg), \end{aligned} \tag{3.20}$$

where we have used

$$\begin{aligned} \sum_{k=1}^{m+1} k R_{m+1-k}^{(m+1)}(z_0) \Bigg(\sum_{l=0}^{k} R_{k-l}^{(-1/2)}(z_0) f_l(z_0) \Bigg) = & \sum_{l=0}^{m+1} \frac{1}{2} \Bigg(\frac{-m-3/2}{m+1/2} (m+1-l) + m + 1 + l \Bigg) \\ & \times R_{m+1-l}^{(m+1/2)}(z_0) f_l(z_0). \end{aligned} \tag{3.21}$$

Instead of Eq. (3.10′), we first expand $z_{NF}(w,\xi)$ and then $z_{NF}(w',\xi')$ in $G_N^F(w,\xi;w',\xi')$ (3.1). Then, comparing the coefficient of $\xi K_{NF}(w)$ on both sides of Eqs. (3.1) and (3.2), we find that

$$\Gamma^{F1}_{N,1/2,m}=\left(\delta_{m,0}\partial_{z_0}I(z_0)+(1-\delta_{m,0})\,\frac{m+1}{m}\,R^{(m)}_{m+1}(z_0)\right)\theta_0+\sum_{j=1}^{m+1}R^{(m)}_{m+1-j}(z_0)V_j \tag{3.22}$$

and

$$\begin{aligned}\Gamma^{F1}_{N,n+1/2,m}\quad(\text{for } n=1,2,...)=&\sum_{k=0}^{n}R^{(n+1/2)}_{n-k}(z_0)\Bigg(\delta_{m,0}\Bigg(\frac{\theta_0}{k!}\,\partial^{k+1}_{z_0}I(z_0)+\sum_{l=1}^{k}\frac{V_{k-l+1}}{(l-1)!}\,\partial^{l}_{z_0}I(z_0)\Bigg)\\&+(1-\delta_{m,0})\Bigg(\frac{k+1}{m}\,R^{(m)}_{m+k+1}(z_0)\theta_0+\sum_{l=1}^{k}\frac{l}{m}\,V_{k-l+1}R^{(m)}_{m+l}(z_0)\Bigg)\Bigg)\\&+\sum_{l=0}^{n}\frac{2l+1}{2n+1}\,R^{(n+1/2)}_{n-l}(z_0)\Bigg(R^{(m)}_{m+l+1}(z_0)\,\theta_0\\&+\sum_{j=1}^{m+l+1}R^{(m)}_{m+l+1-j}(z_0)V_j\Bigg),\end{aligned} \tag{3.23}$$

where we have used

$$\sum_{k=1}^{n}kR^{(n)}_{n-k}(z_0)\Bigg(\sum_{l=0}^{k}R^{(1/2)}_{k-l}(z_0)f_l(z_0)\Bigg)=\sum_{l=0}^{n}\frac{1}{2}\left(\frac{-n+1/2}{n+1/2}\,(n-l)+n+l\right)R^{(n+1/2)}_{n-l}(z_0)f_l(z_0). \tag{3.24}$$

In the above formulas, we need various $R^{(n)}_m(z_0)$ (B1) and $R^{(n+1/2)}_m(z_0)$ (B1′), explicit expressions of which can be obtained (in Appendix B) in terms of I_m and I^n_m defined by Eq. (B2).

On the other hand, N operators ρ^{1F}_N (1.12) is found out to be given by

$$\rho^{1F}_N(z_0,\theta_0;P^\dagger_F)=\sum_{j=1}^{4}\rho^{1Fj}_N(z_0,\theta_0;P^\dagger_F), \tag{3.25}$$

where

$$\rho^{1Fj}_N(z_0,\theta_0;P^\dagger_F)=R^j_X+R^j_{\alpha Nh}+R^j_{dNh}. \tag{3.26}$$

In Eq. (3.26), we have defined that

$$\begin{aligned}R^j_X=&-\sum_{\pm}\,i\oint\frac{dw}{2\pi i}\,\Pi^{\pm j\dagger}_\alpha(w)[X^{\mp j}(z_{NF})]_{\xi=0}\\&+\sum_{\pm}\,i\oint\frac{dw}{2\pi i}\,\pi^{\pm j^\dagger}_d(w)K_{NF}(w)[\theta_{NF}\partial_{z_{NF}}X^{\mp j}(z_{NF})]_{\xi=0},\end{aligned} \tag{3.27}$$

$$\begin{aligned}R^j_{\alpha Nh}=&-i\oint\frac{dw}{2\pi i}\,\Pi^{-hj^\dagger}_\alpha(w)[\theta_{NF}\sqrt{z_{NF}-z_0}\Psi^{hj}_{Nh}(z_{NF})]_{\xi=0}\\&-i\oint\frac{dw}{2\pi i}\,\Pi^{hj^\dagger}_\alpha(w)\left[\theta_{NF}\,\frac{\Psi^{-hj}_{Nh}(z_{NF})}{\sqrt{z_{NF}-z_0}}\right]_{\xi=0}\end{aligned} \tag{3.28}$$

and

$$\begin{aligned}R^j_{dNh}=&\,i\oint\frac{dw}{2\pi i}\,\pi^{-hj^\dagger}_d(w)K_{NF}(w)[\sqrt{z_{NF}-z_0}\Psi^{hj}_{Nh}(z_{NF})]_{\xi=0}\\&+i\oint\frac{dw}{2\pi i}\,\pi^{hj\dagger}_d(w)K_{NF}(w)\left[\frac{\Psi^{-hj}_{Nh}(z_{NF})}{\sqrt{z_{NF}-z_0}}\right]_{\xi=0},\end{aligned} \tag{3.29}$$

where ''h'' is the same as the one in ϵ_h^j of $S_{\{h\}}(z_0)$ (1.4) and we have used the following light-cone operators:

$$\Pi_\alpha^{\pm j^\dagger}(w)=\frac{1}{w}\left(p^{\pm j}+\sum_{n=1}^{\infty}\alpha_{-n}^{\pm j}w^{-n}\right) \tag{3.30}$$

and

$$\pi_d^{\pm j^\dagger}(w)=\frac{1}{\sqrt{w}}\left(d_0^{\pm j}+\sum_{n=1}^{\infty}d_{-n}^{\pm j}w^{-n}\right). \tag{3.31}$$

Furthermore, substituting Eq. (3.2) into N operator $\rho_{N\{h\}}^{2F}$ (1.13), we find that

$$\rho_{N\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)\equiv\sum_{j=1}^{4}\rho_{Nh}^{2Fj}(z_0,\theta_0;P_F^\dagger), \tag{3.32}$$

where

$$\rho_{Nh}^{2Fj}(z_0,\theta_0;P_F^\dagger)=R_{\alpha\alpha Nh}^j+R_{\alpha dNh}^j+R_{d\alpha Nh}^j+R_{ddNh}^j. \tag{3.33}$$

In Eq. (3.33), we have defined that

$$R_{\alpha\alpha Nh}^j\equiv-\oint\frac{dw}{2\pi i}\,\Pi_\alpha^{-hj^\dagger}(w)\oint\frac{dw'}{2\pi i}\,\Pi_\alpha^{hj\dagger}(w')A_N^F(w,w'), \tag{3.34}$$

$$R_{\alpha dNh}^j\equiv\oint\frac{dw}{2\pi i}\,\Pi_\alpha^{-hj^\dagger}(w)\oint\frac{dw'}{2\pi i}\,\pi_d^{hj^\dagger}(w')K_{NF}(w')\Gamma_N^{F2}(w,w'), \tag{3.35}$$

$$R_{d\alpha Nh}^j\equiv\oint\frac{dw}{2\pi i}\,\pi_d^{-hj^\dagger}(w)\oint\frac{dw'}{2\pi i}\,\Pi_\alpha^{hj^\dagger}(w')K_{NF}(w)\Gamma_N^{F1}(w,w'), \tag{3.36}$$

and

$$R_{ddNh}^j\equiv\oint\frac{dw}{2\pi i}\,\pi_d^{-hj^\dagger}(w)\oint\frac{dw'}{2\pi i}\,\pi_d^{hj^\dagger}(w')C_N^F(w,w')\left(\frac{w}{w'}\right)^{1/2}, \tag{3.37}$$

where ''h'' is the same as the one in ϵ_h^j of $S_{\{h\}}(z_0)$ (1.4) and we have used $\Pi_\alpha^{\pm j^\dagger}$ (3.30) and $\pi_d^{\pm j^\dagger}$ (3.31). Finally, we have from Eq. (1.11) that

$$\rho_{N+}^{0F}(z_0,\theta_0)=\sum_{\pm}-ip^{\pm 0}X^{\mp 0}(z_0)-\theta_0 ip^{+0}\Psi_{N+}^{-0}(z_0). \tag{3.38}$$

Substituting ρ_N^{1F} (3.25), $\rho_{N\{h\}}^{2F}$ (3.32), and ρ_{N+}^{0F} (3.38) into the right-hand side of Eq. (1.10), we find that

$$\begin{aligned}&\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)(\equiv\Phi_{N\{h\}}^{F0}(z_0;P_F^\dagger)+\theta_0\Phi_{N\{h\}}^{F1}(z_0;P_F^\dagger))\\&\quad=:\exp\left(\sum_{j=1}^{4}Z_{Nh}^{Fj}(z_0,\theta_0;P_F^\dagger)\right)(1-\theta_0 ip^{+0}\Psi_{0,N+}^{-0})\exp\left(-\sum_{\pm}\sum_{j=0}^{4}ip^{\pm j}X^{\mp j}(z_0)\right):_R,\end{aligned} \tag{3.39}$$

with

$$Z^{Fj}_{Nh}(z_0,\theta_0;P^\dagger_F)=\rho^{1Fj}_N(z_0,\theta_0;P^\dagger_F)+\rho^{2Fj}_{Nh}(z_0,\theta_0;P^\dagger_F) \quad [\text{see Eqs. (3.26) and (3.33)}]. \tag{3.39$'$}$$

In giving formulas of Z^{Fj}_{Nh} in Eq. (3.39), we shall introduce the following shorthand notations (for $j=1-4$, $k=0,1,2,...$):

$$X^{\pm j}_k \equiv \frac{1}{k!}\,\partial^k_{z_0} X^{\pm j}(z_0), \tag{3.40}$$

$$\Psi^{\pm j}_{k+1,N\pm} \equiv \frac{1}{k!}\,\partial^k_{z_0}\Psi^{\pm j}_{N\pm}(z_0), \tag{3.41}$$

and

$$\Psi^{\mp j}_{k,N\pm} \equiv \frac{1}{k!}\,\partial^k_{z_0}\Psi^{\mp j}_{N\pm}(z_0), \tag{3.42}$$

together with the following notations $Q^{\pm hj}$ $(Q=\alpha,d,p,X,\Psi)$;

$$Q^{hj}=\begin{cases} Q^{+j} & \text{for } h=+ \\ Q^{-j} & \text{for } h=- \end{cases} \quad \text{and} \quad Q^{-hj}=\begin{cases} Q^{-j} & \text{for } h+ \\ Q^{+j} & \text{for } h=-, \end{cases} \tag{3.43}$$

where ''h'' is the same as the one in ϵ^j_h in $S_{\{h\}}(z_0)$ (1.4). Then we have from Z^{Fj}_{Nh} (3.39$'$) that

$$\sum_{j=1}^4 Z^{Fj}_{Nh}(z_0,\theta_0;P^\dagger_F)\equiv E_{N\{h\}}(z_0;P^\dagger_F)+\theta_0\mathscr{E}_{N\{h\}}(z_0;P^\dagger_F), \tag{3.44}$$

where

$$\begin{aligned}
E_{N\{h\}}(z_0;P^\dagger_F)=&-\sum_{j=1}^4 i\alpha^{hj}_{-1}(X^{-hj}_1-ip^{-hj}I_1+V_1\Psi^{-hj}_{0,Nh})-\sum_{j=1}^4 i\alpha^{-hj}_{-1}(X^{hj}_1-ip^{hj}I_1)\\
&+\sum_{j=1}^4 id^{hj}_{-1}(\Psi^{-hj}_{1,Nh}+\tfrac{1}{2}I_1\Psi^{-hj}_{0,Nh}-ip^{-hj}V_1)+\sum_{j=1}^4 id^{-hj}_{-1}(\Psi^{hj}_{1,Nh}-ip^{hj}V_1)\\
&-\sum_{j=1}^4 i\alpha^{hj}_{-2}(X^{-hj}_2+I_1X^{-hj}_1-ip^{-hj}(I_1^2+\tfrac{1}{2}I_2)+V_1\Psi^{-hj}_{1,NH}+(V_2+2I_1V_1)\Psi^{-hj}_{0,Nh})\\
&-\sum_{j=1}^4 i\alpha^{-hj}_{-2}(X^{hj}_2+I_1X^{hj}_1-ip^{hj}(I_1^2+\tfrac{1}{2}I_2)+V_1\Psi^{hj}_{1,Nh})\\
&+\sum_{j=1}^4 id^{hj}_{-2}(\Psi^{-hj}_{2,Nh}+\tfrac{3}{2}I_1\Psi^{-hj}_{1,Nh}+(\tfrac{7}{8}I_1^2+\tfrac{1}{2}I_2)\Psi^{-hj}_{0,Nh}-V_1X^{-hj}_1-ip^{-hj}(\tfrac{3}{2}I_1V_1+V_2))\\
&+\sum_{j=1}^4 id^{-hj}_{-2}(\Psi^{hj}_{2,Nh}+\tfrac{3}{2}I_1\Psi^{hj}_{1,Nh}-V_1X^{hj}_1-ip^{hj}(\tfrac{5}{2}I_1V_1+V_2))\\
&-\sum_{j=1}^4 \alpha^{-hj}_{-1}\alpha^{hj}_{-1}(\tfrac{1}{2}I_1^2+\tfrac{1}{2}I_2)+\sum_{j=1}^4 d^{-hj}_{-1}d^{hj}_{-1}\cdot\tfrac{1}{8}I_1^2+\sum_{j=1}^4 \alpha^{-hj}_{-1}d^{hj}_{-1}(\tfrac{3}{2}I_1V_1+V_2)\\
&+\sum_{j=1}^4 \alpha^{hj}_{-1}d^{-hj}_{-1}(I_1V_1+V_2)+\cdots
\end{aligned} \tag{3.45}$$

and

$$\mathscr{E}_{N\{h\}}(z_0;P_F^\dagger)=-\sum_{j=1}^{4} ip^{hj}\Psi_{0,Nh}^{-hj}-\sum_{j=1}^{4} i\alpha_{-1}^{hj}(\Psi_{1,Nh}^{-hj}+I_1\Psi_{0,Nh}^{-hj})-\sum_{j=1}^{4} i\alpha_{-1}^{-hj}(\Psi_{1,Nh}^{hj}-ip^{hj}V_1)$$
$$-\sum_{j=1}^{4} id_{-1}^{hj}(X_1^{-hj}+\tfrac{1}{2}V_1\Psi_{0,Nh}^{-hj}-\tfrac{3}{2}ip^{-hj}I_1)-\sum_{j=1}^{4} id_{-1}^{-hj}(X_1^{hj}-ip^{hj}I_1)$$
$$-\sum_{j=1}^{4} i\alpha_{-2}^{hj}(\Psi_{2,Nh}^{-hj}+2I_1\Psi_{1,Nh}^{-hj}+(2I_1^2+I_2)\Psi_{0,Nh}^{-hj})-\sum_{j=1}^{4} i\alpha_{-2}^{-hj}(\Psi_{2,Nh}^{hj}+2I_1\Psi_{1,Nh}^{hj}$$
$$-(2I_1V_1+V_2)ip^{hj})-\sum_{j=1}^{4} id_{-2}^{hj}(-2X_2^{-hj}-\tfrac{5}{2}I_1X_1^{-hj}-ip^{-hj}(\tfrac{11}{4}I_1^2+\tfrac{7}{4}I_2)$$
$$+\tfrac{1}{2}V_1\Psi_{1,Nh}^{-hj}+(\tfrac{7}{4}I_1V_1+V_2)\Psi_{0,Nh}^{-hj})-\sum_{j=1}^{4} id_{-2}^{-hj}(-2X_2^{hj}-\tfrac{5}{2}I_1X_1^{hj}+\tfrac{1}{2}V_1\Psi_{1,Nh}^{hj}$$
$$-ip^{hj}(\tfrac{5}{2}I_1^2+I_2))-\sum_{j=1}^{4} \alpha_{-1}^{-hj}\alpha_{-1}^{hj}(I_1V_1+V_2)-\sum_{j=1}^{4} d_{-1}^{-hj}d_{-1}^{hj}(\tfrac{5}{4}I_1V_1+V_2)$$
$$-\sum_{j=1}^{4} \alpha_{-1}^{-hj}d_{-1}^{hj}(\tfrac{3}{2}I_1^2+I_2)-\sum_{j=1}^{4} \alpha_{-1}^{hj}d_{-1}^{-hj}(I_1^2+I_2)+\cdots . \tag{3.46}$$

Then we obtain from Eqs. (3.39) and (3.44)–(3.46) that

$$\Phi_{N\{h\}}^{F0}(z_0;P_F^\dagger)=\exp(-E_{N\{h\}}(z_0;P_F^\dagger))-\exp\left(-\sum_{\pm}\sum_{j=0}^{4} ip^{\pm j}X^{\mp j}(z_0)\right) \tag{3.47}$$

and

$$\Phi_{N\{h\}}^{F1}(z_0;P_F^\dagger)=(\mathscr{E}_{N\{h\}}(z_0;P_F^\dagger)-ip^{+0}\Psi_{0,N+}^{-0})\exp(E_{N\{h\}}(z_0;P_F^\dagger))\exp\left(-\sum_{\pm}\sum_{j=0}^{4} ip^{\pm j}X^{\mp j}(z_0)\right). \tag{3.48}$$

Then we can expand $\Phi_{N\{h\}}^{Fm}(z_0;P_F^\dagger)$ (3.47) (for $m=0$) and (3.48) (for $m=1$) as

$$\Phi_{N\{h\}}^{Fm}(z_0;P_F^\dagger)S_{\{h\}}(z_0)\equiv\sum_{\not}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(-i\alpha_{-n}^{\pm j}/\sqrt{n})^{e_{\not}(n,\pm j)}}{\sqrt{e_{\not}(n,\pm j)!}}(id_{-n}^{\pm j})^{\tilde{e}_{\not}(n,\pm j)}\right)$$
$$\times V^{m\not}(z_0)S_{\{h\}}(z_0). \tag{3.49}$$

Comparing the right-hand side of Eq. (3.49) with the left-hand side of Eq. (2.11), we find that $V^{m\not}(z_0)$ (for $m=0,1$) in the expansion (3.49) is just equal to

$$V^{m\not}(z_0)S_{\{h\}}(z_0)=(-)^{\not}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(iA_{-n}^{\pm j}/\sqrt{n})^{e_{\not}(n,\pm j)}}{\sqrt{e_{\not}(n,\pm j)!}}(i\mathscr{D}_{-n}^{\pm j})^{\tilde{e}_{\not}(n,\pm j)}\right)\Sigma_{\{h\}}^{m}(z_0), \tag{3.50}$$

where the $\pm$ sign factor $(-)^{\not}$ is fixed for each $\not$. Then the Ramond state $|\not;\theta_0\rangle$ (2.9) is given by

$$|\not;\theta_0\rangle=(-)^{\not}(V^{0\not}(z_0)+\theta_0V^{1\not}(z_0))S_{\{h\}}(z_0)|0\rangle , \tag{3.51}$$

where p^{-0} is given by $\mathscr{P}_{\not}^{-0}(2.12)$. We can show that the BRST-invariant physical Ramond state[6] in 1/2 picture is given by

$$:\exp(+\sigma(0)):\exp(-1/2\phi(0)):V^{1/}(0)S_{\{h\}}(0)|0\rangle, \tag{3.52}$$

in the on-shell limit;

$$-\sum_{j=0}^{4} p^{+j}p^{-j}+N_{/}\to 0. \tag{3.53}$$

[Incidentally, the BRST-invariant physical Ramond state in $-1/2$ picture is given by

$$:\exp(+\sigma(0)):\exp(-3/2\phi(0)):V^{0/}(0)S_{\{h\}}(0)|0\rangle, \tag{3.54}$$

in the on-shell limit.] In Eq. (3.53), we have used the eigenvalue $N_{/}$ of the light-cone number operator (2.56) N_F, which is given by Eq. (2.12′). In Eq. (3.52), $:\exp(+\sigma(0)):$ is Grassman odd, while $:\exp(-1/2\phi(0)):$ is Grassman even. The GSO-allowed[10] physical vertex operator $V^{1/}$ is such that Eq. (3.52) is Grassman even in total, so that $V^{1/}(0)S_{\{h\}}(0)$ should be Grassman odd. Finally, we remark that the FMS spinor $S_{\{h\}}(0)$ is Grassman even or odd, depending on the values ϵ_h^j; The operator $:\exp(\frac{1}{2}\phi^0(z_0)):$ in Eq. (1.4) is taken to be Grassman even, and it will be called *spinon*. On the other hand, the other factor $\Pi_{j=1}^4:\exp(\epsilon_h^j\phi^0(z_0)):$ in Eq. (1.4) for the *odd* (*even*) value of $\Sigma_{j=1}^4\epsilon_h^j$ will be called *octino* (*octon*), and it is taken to be Grassman *odd* (*even*).

IV. CONCLUSIONS AND DISCUSSIONS

In this paper, we have derived GFPVO and OPE's (among GFPVO and the superstressino operator) by using a completely different method from the one presented in our previous papers:[4,5] In Sec. II, we have constructed both a *physical* string coordinate's mode $A_{-n}^{\pm j}$ (2.1) (for any integer n) and a *physical* string coordino's mode $\mathscr{D}_{-n}^{\pm j}$ (2.2) (for any integer n), in the R sector. We have proved that these operators satisfy (anti-)commutation relations (2.3) and (2.4) (i.e., the spectrum generating algebra[3]). On the other hand, in the NS sector, we can use the (Schwarz) physical string coordino's mode $\mathscr{B}_{-n+1/2}^{\pm j}$ defined by[9]

$$\mathscr{B}_{-n+1/2}^{\pm j}\equiv -i\oint_0 \frac{dz\,d\theta}{2\pi i}D_\theta X^{\pm j}(z,\theta)\,\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}])^{1/2}}$$
$$\cdot\ \exp\left(\frac{(n-1/2)X^{+0}(z,\theta)}{ip^{+0}}\right)\quad[\text{for } j=1-4 \text{ and } n=0,\pm 1,\pm 2,...]. \tag{4.1}$$

We point out that our $(A_m^{\pm j},\mathscr{D}_m^{\pm j})$ in the R sector and $(A_m^{\pm j},\mathscr{B}_{m-1/2}^{\pm j})$[9] in the NS sector are, respectively, representations of Green and Schwarz's (V_m^j,D_m^j) and (V_m^j,B_r^j), which satisfy the spectrum generating algebra.[3] [We remark that our operators $(A_m^{\pm j},\mathscr{D}_m^{\pm j})$ and $(A_m^{\pm j},\mathscr{B}_{m-1/2}^{\pm j})$ are constructed by using the string coordinate $X^{\pm j}$ and string coordino $\Psi^{\pm j}\equiv:\exp(\pm\phi^j):$ which are common to both the R and NS sectors.] With the help of these operators, we can construct a complete set of physical states $|/;\theta_0\rangle$ (2.9), which are generated by the left-hand side of Eq. (2.11). One of main results in this paper is that we have proved in Sec. II that the left-hand side of Eq. (2.11) is just equal to the right-hand side of Eq. (2.11), which is nothing but the GFPVO which has already been proposed[4,5] by using the superconformal mapping in the R sector. In proving Eq. (2.11), we extensively used various formulas [related with the superconformal mapping (in the R sector) given by Eqs. (1.18) and (1.19)] which have been derived and used in Refs. 4 and 5. Furthermore, we have succeeded in deriving OPE's (2.55), (2.62), and (2.63), respectively, from simple OPE's (2.54), (2.58), and (2.59), since the physical string coordinate's mode $A_m^{\pm j}$ (2.1) and physical string coordino's mode $\mathscr{D}_m^{\pm j}$ (2.2) are invariant under the super-Virasoro algebra. [See Eq. (2.4).] Compared with the previous derivation of these OPE's in Ref. 5, the new derivation in this paper is quite simple.

In Sec. III, we have calculated GFPVO (1.10) by using superconformal mapping in the R sector. Some results are given by $\Phi^{Fm}_{N\{h\}}(z_0;P^\dagger_F)$ (3.47) [for $m=0$] and (3.48) [for $m=1$], substituted with $E_{N\{h\}}(z_0;P^\dagger_F)$ (3.45) and $\mathscr{E}_{N\{h\}}(z_0;P^\dagger_F)$ (3.46). In these formulas, we have used the following notations;

$$I(z_0)\equiv\frac{X^{+0}(z_0)}{ip^{+0}},\quad I_m\equiv\partial^m_{z_0}I(z_0),\quad I^n_m\equiv(I_m)^n, \tag{4.2}$$

$$V_{k+1}\equiv\frac{1}{k!}\,\partial^k_{z_0}\,\frac{\Psi^{+0}_{N+}(z_0)}{ip^{+0}}, \tag{4.3}$$

$$X^{\pm j}_k\equiv\frac{1}{k!}\,\partial^k_{z_0}X^{\pm j}(z_0)\quad[\text{for } j=1-4], \tag{4.4}$$

and

$$\Psi^{hj}_{k+1,Nh}\equiv\frac{1}{k!}\,\partial^k_{z_0}\Psi^{hj}_{Nh}(z_0),\quad \Psi^{-hj}_{k,Nh}\equiv\frac{1}{k!}\,\partial^k_{z_0}\Psi^{-hj}_{Nh}(z_0)\quad[\text{for } j=1-4]. \tag{4.5}$$

How these formulas have been derived in Sec. III will be briefly summaried hereafter. N operators ρ^{1F}_N (1.12) and $\rho^{2F}_{N\{h\}}$ (1.13) can be calculated up to an arbitrarily required number of terms, by using ρ^{1F}_N (3.25)–(3.29) and $\rho^{2F}_{N\{h\}}$ (3.32)–(3.37), where we have used A^F_N (3.3) [given by Eqs. (3.14) and (3.15)], Γ^{F1}_N (3.4) [given by Eqs. (3.22) and (3.23)], Γ^{F2}_N (3.5) [given by Eqs. (3.19) and (3.20)] and C^F_N (3.6) [given by A^{F1} (3.7) and A^{F2} (3.8)].

The *physical* Ramond states (in 1/2 picture) are proposed to be constructed by Eq. (3.52), which are shown by OPE (2.63) to be annihilated by the Dirac operator $\mathscr{F}^F_0$, in the on-shell limit $L^F_0\to 0$. (Incidentally, this state can be shown to be BRST invariant.) In Appendix A we have proved (anti-)commutation relations (2.3) and (2.4) by using the normal ordering and the contraction in the R sector. In Appendix B, first some terms of ρ^{1F}_N (3.25) and $\rho^{2F}_{N\{h\}}$ (3.32) are obtained explicitly by using Eqs. (B27)–(B29) and (B23)–(B26), respectively.

In calculating physical scattering amplitudes among external physical states including more than six Ramond particles, we will find from FMS's work that that we need to picture-change[6] some physical Ramond states in 1/2 picture [later given by Eq. (5.12)] into those in 3/2 picture. These picture-changing calculations can be carried out by using the N operator $\hat\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)$ which we have in the following OPE (which has been derived in Ref. 5).

$$\begin{aligned}
&\mathbf{R}\cdot\mathscr{F}_F(z')\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)\\
&\quad=-\frac{\theta_0}{(\sqrt{z'-z_0})^3}\cdot\Bigg(-\Bigg(\sum_{j=0}^{4}p^{+j}p^{-j}\Bigg)\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)\\
&\qquad+[N_F,\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)]\Bigg)+\frac{1}{(\sqrt{z'-z_0})^3}\cdot\partial_{\theta_0}\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)\\
&\qquad-\frac{\theta_0}{\sqrt{z'-z_0}}\cdot\partial_{z_0}(\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0))+\frac{1}{\sqrt{z'-z_0}}\,\hat\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)\\
&\qquad+O(\sqrt{z'-z_0}),
\end{aligned}\tag{4.6}$$

where the N operator $\hat\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)$ is given by

$$\hat\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)(\equiv\hat\Phi^{F3}_{N\{h\}}(z_0;P^\dagger_F)+\theta_0\hat\Phi^{F4}_{N\{h\}}(z_0;P^\dagger_F))=\sum_{i=1}^{3}\hat\Phi^F_{N\{h\},i}(z_0,\theta_0;P^\dagger_F). \tag{4.7}$$

In Eq. (4.7), $\hat{\Phi}^F_{N\{h\},1}(z_0,\theta_0;P_F^\dagger)$ is given by

$$\begin{aligned}\hat{\Phi}^F_{N\{h\},1}(z_0,\theta_0;P_F^\dagger)=&-:\frac{\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\left(\left(-\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)\right.\\&\left.+[N_F,\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)]\right):_R-:\left(\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right)\partial_{\theta_0}\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R\\&-:(\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0)+\theta_0\partial_{z_0}ip^{-0}X^{+0}(z_0)+\partial_{z_0}X^{+0}(z_0)\\&\cdot\Psi^{-0}_{N+}(z_0))\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R\\&-\sum_{\pm}\sum_{j=1}^{4}i\oint\frac{dw\,d\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi):(\mathscr{H}_{NF}(w,\xi)X_N^{\mp j}(z_{NF},\theta_{NF}))\Phi^F_{N\{h\}}\\&\times(z_0,\theta_0;P_F^\dagger):_R-\sum_{j=1}^{4}\oint\frac{dw\,d\xi}{2\pi i}P_F^{-hj^\dagger}(w,\xi)\oint\frac{dw'\,d\xi'}{2\pi i}P_F^{+hj^\dagger}(w',\xi'):\\&\times\left((\mathscr{H}_{NF}(w,\xi)+\mathscr{H}_{NF}(w',\xi'))\log\left(z_{NF}-z'_{NF}-\left(\frac{z_{NF}-z_0}{z'_{NF}-z_0}\right)^{1/2}\theta_{NF}\theta'_{NF}\right)\right)\\&\times\Phi^F_{N\{h\}},\theta_0;P_F^\dagger):_R+O(\sqrt{z'-z_0}),\end{aligned}\tag{4.8}$$

where we have from Eq. (3.1) that

$$\log\left(z_{NF}-z'_{NF}-\left(\frac{z_{NF}-z_0}{z'_{NF}-z_0}\right)^{1/2}\theta_{NF}\theta'_{NF}\right)=G^F_N(w,\xi;w',\xi')+\log\left(w-w'-\left(\frac{w}{w'}\right)^{1/2}\xi\xi'\right).\tag{4.9}$$

In Eq. (4.8), $\mathscr{H}_{NF}(w,\xi)$ is the N operator given by

$$\begin{aligned}\mathscr{H}_{NF}(w,\xi)=&-\frac{J_{NF}(w)+\xi K_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\left(\frac{1}{2}\xi D_\xi+w\partial_w\right)+\frac{wK_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\cdot\partial_\xi\\&+\xi\left(\frac{\frac{1}{2}w\partial_W J_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}+\frac{3}{4}wK^2_{NF}(w)\frac{J_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^5}\right)\partial_\xi.\end{aligned}\tag{4.10}$$

Furthermore, $\hat{\Phi}^F_{N\{h\},i}(z_0,\theta_0;P_F^\dagger)$'s [for $i=2,3$] in Eq. (4.7) are given, respectively, by

$$\begin{aligned}\hat{\Phi}^F_{N\{h\},2}(z_0,\theta_0;P_F^\dagger)=:&\left(1/2\Psi^{+hj}_{Nh}(z_0)\Psi^{-hj}_{Nh}(z_0)+\frac{1}{2}i\oint\frac{dw\,d\xi}{2\pi i}P_F^{-hj^\dagger}(w,\xi)\right.\\&\times\frac{\theta_{NF}(w,\xi)}{\sqrt{z_{NF}(w,\xi)-z_0}}\cdot\Psi^{+hj}_{Nh}(z_0)\\&\left.-\frac{1}{2}i\oint\frac{dw\,d\xi}{2\pi i}P_F^{+hj^\dagger}(w,\xi)\frac{\theta_{NF}(w,\xi)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\cdot\Psi^{-hj}_{Nh}(z_0)\right)\theta_0\\&\cdot\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R+:\left(\frac{\frac{1}{2}\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\right)\theta_0\partial_{\theta_0}\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R\end{aligned}\tag{4.11}$$

and

$$\hat{\Phi}^F_{N\{h\},3}(z_0,\theta_0;P^\dagger_F)\equiv\sum_{j=1}^{4}:\left(\Psi^{-hj}_{Nh}(z_0)-i\oint\frac{dw\ d\xi}{2\pi i}P_F^{-hj^\dagger}(w,\xi)\frac{\theta_{NF}(w,\xi)}{\sqrt{z_{NF}(w,\xi)-z_0}}\right)\left(\partial_{z_0}X^{+hj}(z_0)\right.$$
$$\left.+i\oint\frac{dw\ d\xi}{2\pi i}P_F^{+hj^\dagger}(w,\xi)\frac{1}{z_{NF}(w,\xi)-z_0}\right)\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F):_R. \tag{4.12}$$

In calculating N operator $\hat{\Phi}^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)$ (4.7) by using Eqs. (4.8)–(4.12), we can use the various formulas in terms of N operators. Incidentally, we can use that

$$z_{NF}(w,\xi)-z_0=w\cdot\exp(G^F_N(0,0;w,\xi))\quad[\text{from Eq. (3.1)}] \tag{4.13}$$

and

$$J_{NF}(w)=\sum_{n=0}^{\infty}w^{n+1/2}\left(\theta_0R^{(n+1/2)}_n(z_0)+\sum_{k=1}^{\infty}R^{(n+1/2)}_{n-k}(z_0)V_k\right)\quad[\text{from Eqs. (1.16) and (1.21)}]. \tag{4.14}$$

In the mass shell limit [when the first term on the right-hand side of OPE (4.6) vanishes], we find that

$$\lim_{z'\to z_0}\sqrt{z'-z_0}\mathscr{F}(z')\Phi^{F1}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0)=\partial_{z_0}(\Phi^{F0}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0))-\hat{\Phi}^{F4}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0), \tag{4.15}$$

which can be used in picture-changing[5] Ramond states in 1/2 picture into those in 3/2 picture.

Finally, we can similarly analyze the *NS* sector. Therefore we only sketch various useful results in the *NS* sector; A complete and orthonormal infinite set of *physical NS* states can be constructed by

$$|\{e_{\mathscr{A}}(n,\pm j)\},\{\tilde{e}_{\mathscr{A}}(n-\tfrac{1}{2},\pm j)\},\{p\};\theta_0\rangle(\equiv|\mathscr{A};\theta_0\rangle)$$
$$\equiv\lim_{z_0\to0}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(iA^{\pm j}_{-n}/\sqrt{n})^{e_{\mathscr{A}}(n,\pm j)}}{(e_{\mathscr{A}}(n,\pm j)!)^{1/2}}(i\mathscr{B}^{\pm j}_{-n+1/2})^{\tilde{e}_{\mathscr{A}}(n-1/2,\pm j)}\right)t(z_0|\theta_0)|0\rangle, \tag{4.16}$$

where we have introduced the *basic* tachynorial operator $t(z_0|\theta_0)$ defined by

$$t(z_0|\theta_0)\equiv:\exp\left(-\sum_{\pm}\sum_{j=0}^{4}ip^{\pm j}X^{\mp j}(z_0,\theta_0)\right):_{NS}. \tag{4.17}$$

In Eq. (4.16), $e_{\mathscr{A}}(n,\pm j)$ is some non-negative integer and $\tilde{e}_{\mathscr{A}}(n-\frac{1}{2},\pm j)$ is either 0 or 1. Hereafter, $p^{\pm j}$ [in Eq. (4.17)] is assumed to satisfy the *on-shell* condition

$$p^{+0}p^{-0}+\sum_{j=1}^{4}p^{+j}p^{-j}=-\frac{1}{2}, \tag{4.18}$$

leading to the fact that the basic tachynorial operator $t(z_0|0)$ (4.17) is the primary operator of conformal weight 1/2. On the other hand, we shall introduce the *operator* $\mathscr{P}^{-0}$ satisfying

$$\mathscr{P}^{-0}|\mathscr{A};\theta_0\rangle\equiv\mathscr{P}^{-0}_{\mathscr{A}}|\mathscr{A};\theta_0\rangle, \tag{4.19}$$

where $|\mathscr{A};\theta_0\rangle$ is the state given by Eq. (4.16) and $\mathscr{P}^{-0}_{\mathscr{A}}$ is the eigenvalue given by $\mathscr{P}^{-0}_{\mathscr{A}}\equiv p^{-0}+N_{\mathscr{A}}(p^{+0})^{-1}$ with

$$N_{\mathscr{E}} \equiv \sum_{n=1}^{\infty} \sum_{\pm} \sum_{j=1}^{4} (n \cdot e_{\mathscr{E}}(n, \pm j) + (n - \tfrac{1}{2}) \cdot \widetilde{e}_{\mathscr{E}}(n - \tfrac{1}{2}, \pm j)). \tag{4.20}$$

Therefore, we have the following mass-shell condition for the state $|\mathscr{E}; \theta_0\rangle$ (4.16);

$$p^{+0} \mathscr{P}_{\mathscr{E}}^{-0} + \sum_{j=1}^{4} p^{+j} p^{-j} = N_{\mathscr{E}} - \tfrac{1}{2}. \tag{4.21}$$

In much the same way as we have proved Eq. (2.11) in the R sector, we can prove the following *theorem* in the NS sector;

Theorem:

$$\mathbf{R} \cdot \exp\left(\sum_{n=1}^{\infty} \sum_{\pm} \sum_{j=1}^{4} \left(\frac{1}{n} \alpha_{-n}^{\mp j} \cdot A_{-n}^{\pm j} - b_{-n+1/2}^{\mp j} \cdot \mathscr{B}_{-n+1/2}^{\pm j} \right) \right) t(z_0 | \theta_0) = \widetilde{\Phi}^B(z_0, \theta_0; P_B^{\dagger}). \tag{4.22}$$

In Eq. (4.22), GFPVO $\widetilde{\Phi}^B(z_0, \theta_0; P_B^{\dagger})$ can be defined as follows: We introduce the light-cone operator $P_B^{\pm j\dagger}(w, \xi)$ in the NS sector by

$$P_B^{\pm j\dagger}(w, \xi) \equiv \frac{\xi}{w} \left(p^{\pm j} + \sum_{n=1}^{\infty} w^{-n} \alpha_{-n}^{\pm j} \right) + \frac{1}{\sqrt{w}} \left(\sum_{n=1}^{\infty} w^{-n+1/2} b_{-n+1/2}^{\pm j} \right), \tag{4.23}$$

where we have used light-cone (mode's) operators (for $j = 1-4$), i.e., $p^{\pm j}$ (total energy-momentum operator), $\alpha_m^{\pm j}$'s, $\alpha_{-n}^{\pm j}$'s, $b_{m-1/2}^{\pm j}$'s, and $b_{-n+1/2}^{\pm j}$'s, which satisfy the following (anti-)commutation relations among themselves;

$$[\alpha_m^{\pm j}, \alpha_{-n}^{\mp k}] = m \delta_{mn} \delta_{jk} \tag{4.24}$$

and

$$\{b_{m-1/2}^{\pm j}, b_{-n+1/2}^{\mp k}\} = \delta_{mn} \delta_{jk}, \tag{4.25}$$

all other (anti-)commutators being zero. Then we can define that

$$\widetilde{\Phi}^B(z_0, \theta_0; P_B^{\dagger}) \equiv :\exp(\widetilde{\rho}^{0B}(z_0, \theta_0) + \rho^{1B}(z_0, \theta_0; P_B^{\dagger}) + \rho^{2B}(z_0, \theta_0; P_B^{\dagger})):_{NS} \tag{4.26}$$

$$(\equiv \widetilde{\Phi}^{B0}(z_0; P_B^{\dagger}) + \theta_0 \widetilde{\Phi}^{B1}(z_0; P_B^{\dagger})). \tag{4.26'}$$

In Eq. (4.26), the normal ordering in the NS sector (denoted by $:\ :_{NS}$) is defined by

$$\mathbf{R} \cdot X^{\pm j}(z') X^{\mp k}(z) = \delta_{jk} \log(z' - z) + :X^{\pm j}(z') X^{\mp k}(z):_{NS} \tag{4.27}$$

and

$$\mathbf{R} \cdot \Psi^{\pm j}(z') \Psi^{\mp k}(z) = \frac{\delta_{jk}}{z' - z} + :\Psi^{\pm j}(z') \Psi^{\mp k}(z):_{NS}, \tag{4.28}$$

$\mathbf{R}$ being radial ordering. Furthermore, $\rho^{\iota B}$'s ($\iota = 0,1,2$) in Eq. (4.26) are given by

$$\widetilde{\rho}^{0B}(z_0, \theta_0) \equiv -ip^{+0} X^{-0}(z_0, \theta_0) - i\mathscr{P}^{-0} X^{+0}(z_0, \theta_0), \tag{4.29}$$

$$\rho^{1B}(z_0, \theta_0; P_B^{\dagger}) \equiv -\sum_{\pm} \sum_{j=1}^{4} i \oint \frac{dw\, d\xi}{2\pi i} P_B^{\pm j\dagger}(w, \xi) X^{\mp j}(z_B(w, \xi), \theta_B(w, \xi)) \tag{4.30}$$

and

$$\rho^{2B}(z_0,\theta_0;P_B^\dagger)\equiv-\sum_{j=1}^{4}\oint\frac{dw\ d\xi}{2\pi i}P_B^{+j^\dagger}(w,\xi)\oint\frac{dw'\ d\xi'}{2\pi i}P_B^{-j^\dagger}(w',\xi')$$
$$\times\log\frac{z_B(w,\xi)-z_B(w',\xi')-\theta_B(w,\xi)\theta_B(w',\xi')}{w-w'-\xi\xi'}. \tag{4.31}$$

In Eqs. (4.30) and (4.31), we have used

$$\theta_B(w,\xi)-\theta_0\equiv J_B(w)+\xi K_B(w) \tag{4.32}$$

and

$$z_B(w,\xi)-z_0-\theta_B(w,\xi)\theta_0\equiv L_B(w)+\xi K_B(w)J_B(w), \tag{4.33}$$

which are those to be determined by inversely solving the following superconformal mapping in the NS sector $(z_B,\theta_B)\to(w,\xi)$;

$$z_B-z_0-\theta_B\theta_0=w{:}\exp\left(\frac{X^{+0}(z_B,\theta_B)}{ip^{+0}}-\frac{X^{+0}(z_0,\theta_0)}{ip^{+0}}\right){:}_{NS} \tag{4.34}$$

and

$$\xi(z_B,\theta_B)=(\partial_{z_B}w(z_B,\theta_B))^{-1/2}D_{\theta_B}w(z_B,\theta_B), \tag{4.35}$$

with $D_{\theta_B}\equiv\partial_{\theta_B}+\theta_B\partial_{z_B}$. With the help of

$$t(z_0|\theta_0)\equiv t^0(z_0)+\theta_0 t^1(z_0) \tag{4.36}$$

Eq. (4.22) is rewritten into

$$\mathbf{R}\cdot\exp\left(\sum_{n=1}^{\infty}\sum_{\pm}\sum_{j=1}^{4}\left(\frac{1}{n}\alpha_{-n}^{\mp j}\cdot A_{-n}^{\pm j}-b_{-n+1/2}^{\mp j}\cdot\mathscr{B}_{-n+1/2}^{\pm j}\right)\right)t^m(z_0)=\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)\quad[\text{for } m=0,1]. \tag{4.37}$$

Then, we can expand $\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)$ in Eq. (4.26′) as

$$\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)\equiv\sum_{\mathscr{E}}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(-i\alpha_{-n}^{\pm j}/\sqrt{n})^{e_{\mathscr{E}}(n,\pm j)}}{\sqrt{e_{\mathscr{E}}(n,\pm j)!}}(ib_{-n+1/2}^{\pm j})^{\tilde{e}_{\mathscr{E}}(n-1/2,\pm j)}\right)V^{m\mathscr{E}}(z_0)$$
$$[\text{for } m=0,1]. \tag{4.38}$$

Comparing the right-hand side of Eq. (4.38) with the left-hand side of Eq. (4.22), we find that $V^{m\mathscr{E}}(z_0)$ (for $m=0,1$) in the expansion (4.38) is just equal to

$$V^{m\mathscr{E}}(z_0)=(-)^{\mathscr{E}}\left(\prod_{n=1}^{\infty}\prod_{\pm}\prod_{j=1}^{4}\frac{(iA_{-n}^{\pm j}/\sqrt{n})^{e_{\mathscr{E}}(n,\pm j)}}{\sqrt{e_{\mathscr{E}}(n,\pm j)!}}(i.\mathscr{B}_{-n+1/2}^{\pm j})^{\tilde{e}_{\mathscr{E}}(n-1/2,\pm j)}\right)t^m(z_0), \tag{4.39}$$

where the $\pm$ sign factor $(-)^{\mathscr{E}}$ is fixed for each $|\mathscr{E};\theta_0\rangle$. Then the Neveu–Schwarz $|\mathscr{E};\theta_0\rangle$ (4.16) is given by

$$|\mathscr{E};\theta_0\rangle=(-)^{\mathscr{E}}(V^{0\mathscr{E}}(z_0)+\theta_0V^{1\mathscr{E}}(z_0))|0\rangle, \tag{4.40}$$

where $\mathscr{P}^{-0}$ is to be replaced with $\mathscr{P}_{\mathscr{E}}^{-0}$ (4.19). GSO-allowed[10] physical vertex operators $V^{0\mathscr{E}}$ in 0 picture ($V^{1\mathscr{E}}$ in 1 picture) are Grassman odd ones (Grassman even ones).

We consider the superstressino operator defined by

$$\mathscr{T}_B(z',\theta')\equiv\sum_{j=0}^{4}\sum_{\pm}:\frac{1}{2}\partial_{z'}X^{\pm j}(z',\theta')\cdot D_{\theta'}X^{\mp j}(z',\theta'):_{NS} \tag{4.41}$$

$$\equiv\tfrac{1}{2}\mathscr{T}_B(z')+\theta' T_B(z'). \tag{4.41'}$$

[The notation $\mathscr{T}_B(z',\theta')$ denoting the superstressino operator in the NS sector will be used instead of the notation $T_B(z',\theta')$ in Ref. 4.] From Eqs. (4.17) and (4.36), we can easily find the following OPE valid even in the off-shell case; With respect to the stress operator $T_B(z')$ (4.41′), we can easily find that

$$\mathbf{R}\cdot T_B(z')\cdot t^m(z_0)=\frac{\frac{1}{2}\delta_{m1}-(p^{+0}p^{-0}+\Sigma_{j=1}^{4}p^{+j}p^{-j})}{(z'-z_0)^2}\cdot t^m(z_0)+\frac{1}{z'-z_0}\partial_{z_0}t^m(z_0)+\mathrm{O}(1)$$

$$[\text{for } m=0,1]. \tag{4.42}$$

Applying OPE's (4.42) to the left-hand side of (4.37), and taking account of the mass relation (4.21) and invariance of $A_{-n}^{\pm j}$ (2.1) and $\mathscr{B}_{-n+1/2}^{\pm j}$ (4.1) under the super-Virasoro algebra (in the NS sector), we find that

$$\mathbf{R}\cdot T_B(z')\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)=\frac{1}{(z'-z_0)^2}\left(\left(\frac{1}{2}\delta_{m1}-p^{+0}\mathscr{P}^{-0}-\sum_{j=1}^{4}p^{+j}p^{-j}\right)\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)\right.$$

$$\left.+[N_B,\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)]\right)+\frac{1}{z'-z_0}\partial_{z_0}\widetilde{\Phi}^{Bm}(z_0;P_B^\dagger)+(\text{terms regular at } z'$$

$$=z_0)\quad[\text{for } m=0,1]. \tag{4.43}$$

In Eq. (4.43), we have used the light-cone number operator N_B defined by

$$N_B\equiv\sum_{n=1}^{\infty}\sum_{\pm}\sum_{k=1}^{4}(\alpha_{-n}^{\mp k}\alpha_n^{\pm k}+(n-\tfrac{1}{2})b_{-n+1/2}^{\mp k}b_{n-1/2}^{\pm k}) \tag{4.44}$$

and $\widetilde{\Phi}^{B0(1)}(z_0;P_B^\dagger)$ is the one defined by Eq. (4.26′). On the other hand, we find with respect to the stressino operator $\mathscr{T}_B(z')$ (4.41′) that

$$\mathbf{R}\cdot\mathscr{T}_B(z')\cdot t^0(z_0)=\frac{1}{z'-z_0}\partial_{z_0}t^1(z_0)+O(1) \tag{4.45}$$

and

$$\mathbf{R}\cdot\mathscr{T}_B(z')\cdot t^1(z_0)=\frac{-2(p^{+0}p^{-0}+\Sigma_{j=1}^{4}p^{+j}p^{-j})}{(z'-z_0)^2}\cdot t^0(z_0)+\frac{1}{z'-z_0}\partial_{z_0}t^0(z_0)+O(1). \tag{4.46}$$

Applying OPE's (4.45) and (4.46) to the left-hand side of (4.22), and taking account of the mass relation (4.21) and invariance of $A_{-n}^{\pm j}$ (2.1) and $\mathscr{B}_{-n+1/2}^{\pm j}$ (4.1) under the super-Virasoro algebra (in the NS sector), we find that

$$\mathbf{R}\cdot\mathscr{T}_B(z')\widetilde{\Phi}^{B0}(z_0;P_B^\dagger)=\frac{1}{z'-z_0}\widetilde{\Phi}^{B1}(z_0;P_B^\dagger)+(\text{terms regular at } z'=z_0) \tag{4.47}$$

and

$$\mathbf{R}\cdot\mathscr{T}_B(z')\widetilde{\Phi}^{B1}(z_0;P_B^\dagger)=\frac{2}{(z'-z_0)^2}\left(\left(-p^{+0}\mathscr{P}^{-0}-\sum_{j=1}^{4}p^{+j}p^{-j}\right)\widetilde{\Phi}^{B0}(z_0;P_B^\dagger)\right.$$
$$\left.+[N_B,\widetilde{\Phi}^{B0}(z_0;P_B^\dagger)]\right)+\frac{1}{z'-z_0}\partial_{z_0}\widetilde{\Phi}^{B0}(z_0;P_B^\dagger)$$
$$+(\text{terms regular at } z'=z_0). \tag{4.48}$$

APPENDIX A: PROOF OF (ANTI-)COMMUTATION RELATIONS (2.3) AND (2.4)

The *physical* operator $A_{-n}^{\pm j}$ (2.1) for any integer n is similar to the one (in the NS sector) introduced by Schwarz,[9] while we introduce in the R sector the *physical* operator $\mathscr{D}_{-n}^{\pm j}$ (2.2) for any integer n. [We notice that Schwarz introduced in the NS sector the *physical* operator $\mathscr{B}_{-n+1/2}^{\pm j}$ (4.1) for any integer n.] In the following, we shall prove (anti-)commutation relations (2.3) and (2.4) in the R sector by using the normal ordering and contractions in the R sector.[4]

We shall prove the last anti-commutation relation in Eq. (2.3) as an example. With the help of the contraction in the R sector given by[4]

$$\overbrace{X^{\pm j}(z,\theta)X^{\mp k}(z',\theta')}^{R}=\delta_{jk}\left(\log(z-z')-\frac{\theta\theta'}{z-z'}\left(\frac{z'-z_0}{z-z_0}\right)^{\mp\varepsilon_h^j}\right),$$

we find that

$$\overbrace{D_\theta^m X^{\pm j}(z,\theta)X^{\mp k}(z',\theta')}^{R}=\delta_{jk}D_\theta^m\log(z-z'-\theta\theta')$$
$$+\ (\text{terms regular at } z=z')\quad[\text{for } m=1,2]. \tag{A1}$$

Therefore we obtain that

$$\{\mathscr{D}_n^{\pm j},\mathscr{D}_{-m}^{\mp k}\}=\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\oint_z\frac{dz'\,d\theta'}{2\pi i}\overbrace{D_\theta X^{\pm j}(z,\theta)D_{\theta'}X^{\mp k}(z',\theta')}^{R}$$
$$\times\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}])^{1/2}}\cdot\exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)$$
$$\times\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{1/2}}\cdot\exp\left(\frac{mX^{+0}(z',\theta')}{ip^{+0}}\right)$$
$$=\delta_{jk}\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}\oint_z\frac{dz'\,d\theta'}{2\pi i}\frac{-1}{z'-z-\theta'\theta}$$
$$\times\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}])^{1/2}}\cdot\exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)$$
$$\times(\theta'-\theta)D_\theta\left(\frac{D_\theta[X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z[X^{+0}(z,\theta)/ip^{+0}])^{1/2}}\cdot\exp\left(\frac{mX^{+0}(z,\theta)}{ip^{+0}}\right)\right)$$
$$=-\delta_{jk}\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}D_\theta\frac{X^{+0}(z,\theta)}{ip^{+0}}\cdot\exp\left(\frac{(-n+m)X^{+0}(z,\theta)}{ip^{+0}}\right)$$
$$=\begin{cases}-\dfrac{1}{-n+m}\delta_{jk}\displaystyle\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}D_\theta\exp\left(\frac{(-n+m)X^{+0}(z,\theta)}{ip^{+0}}\right)=0 & [\text{for } -n+m\neq 0]\\ -\delta_{jk}\displaystyle\oint_{|z|>|z_0|}\frac{dz\,d\theta}{2\pi i}D_\theta\frac{X^{+0}(z,\theta)}{ip^{+0}}=\delta_{jk} & [\text{for } -\text{n}+\text{m}=0]\end{cases} \tag{A2}$$

(Q.E.D.)

As for Eq. (2.4), we first notice that the superstressino operator $\mathscr{T}_F(z',\theta')$ (normally ordered in the R sector) is given by

$$\mathscr{T}_F(z',\theta')\left(\equiv\frac{1}{2}\,\mathscr{T}_F(z')+\theta' T_F(z')\right)=\sum_{m=1}^{2}\sum_{j=0}^{4}:\frac{1}{2}\,D_{\theta'}^{3-m}X^{-j}(z',\theta')\cdot D_{\theta'}^{m}X^{+j}(z',\theta'):_R \tag{A3}$$

with

$$T_F(z')\equiv\frac{1}{(z'-z_0)z'}\left(\sum_{n=-\infty}^{\infty}L_n^F(z_0)z'^{-n}\right) \tag{A4}$$

and

$$\mathscr{T}_F(z')\equiv\frac{1}{(z'-z_0)^{1/2}z'}\left(\sum_{n=-\infty}^{\infty}\mathscr{T}_n^F(z_0)z'^{-n}\right). \tag{A5}$$

[Hereafter, we shall use the notation $\mathscr{T}_F(z',\theta')$ (instead of the notation $T_F(z',\theta')$ used in Ref. 4), in order to indicate that it is Grassman odd.] In the following, we shall prove that

$$[A_n^{+j},\mathscr{T}_F(z',\theta')]=0 \tag{A6}$$

and

$$\{\mathscr{D}_n^{+j}(z_0),\mathscr{T}_F(z',\theta')\}=0. \tag{A7}$$

Proof: First, we notice that

$$\begin{aligned}
&\overbrace{\sum_{m=1}^{2}X^{+j}(z,\theta)(D_{\theta'}^{3-m}X^{-j}(z',\theta')\cdot D_{\theta'}^{m}X^{+j}(z',\theta'))}^{R}\\
&=\overbrace{X^{+j}(z,\theta)(\partial_{z'}X^{-j}(z',\theta')D_{\theta'}X^{+j}(z',\theta'))}^{R}\\
&\quad+\overbrace{X^{+j}(z,\theta)(D_{\theta'}X^{-j}(z',\theta')\partial_{z'}X^{+j}(z',\theta'))}^{R}\\
&=\partial_{z'}\log\left(z-z'-\theta\theta'\left(\frac{z-z_0}{z'-z_0}\right)^{+\varepsilon_h^j}\right)\cdot D_{\theta'}X^{+j}(z',\theta')\\
&\quad+D_{\theta'}\log\left(z-z'-\theta\theta'\left(\frac{z-z_0}{z'-z_0}\right)^{+\varepsilon_h^j}\right)\cdot\partial_{z'}X^{+j}(z',\theta')\\
&\equiv\sum_{m=1}^{2}D_{\theta'}^{3-m}\log\left(z-z'-\theta\theta'\left(\frac{z-z_0}{z'-z_0}\right)^{+\varepsilon_h^j}\right)\cdot D_{\theta'}^{m}X^{+j}(z',\theta').
\end{aligned} \tag{A8}$$

Then, we find that

$$
\begin{aligned}
&i2[A_n^{+j}, \mathscr{F}_F(z',\theta')] \\
&= \sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : D_\theta X^{+j}(z,\theta) \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right) \\
&\qquad \overbrace{\qquad\qquad R \qquad\qquad} \\
&\quad \times \left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right) D_{\theta'}^{3-m} X^{-0}(z',\theta') \cdot D_{\theta'}^{m} X^{+0}(z',\theta') :_R \\
&\quad - \sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : D_\theta X^{+0}(z,\theta) \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right) \\
&\qquad \overbrace{\qquad\qquad R \qquad\qquad} \\
&\quad \times \left(\frac{-nX^{+j}(z,\theta)}{ip^{+0}}\right) D_{\theta'}^{3-m} X^{-j}(z',\theta') \cdot D_{\theta'}^{m} X^{+j}(z',\theta') :_R = 0,
\end{aligned}
\tag{A9}
$$

where we have used Eq. (A1). On the other hand, we notice that

$$
i2\{\mathscr{D}_n^{+j}, \mathscr{F}_F(z',\theta')\} = \sum_{m=1}^{4} H_m, \tag{A10}
$$

where H_m's $(m=1-4)$ are, respectively, given by

$$
\begin{aligned}
H_1 = \sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : &\left(D_\theta X^{+j}(z,\theta) \frac{D_\theta [X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z [X^{+0}(z,\theta)/ip^{+0}])^{1/2}} \right. \\
&\overbrace{\qquad\qquad R \qquad\qquad} \\
&\left. \times \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)\right) \left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right) D_{\theta'}^{3-m} X^{-0}(z',\theta') \cdot D_{\theta'}^{m} X^{+0}(z',\theta') :_R ,
\end{aligned}
\tag{A11}
$$

$$
\begin{aligned}
H_2 = \sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : &\left(D_\theta X^{+j}(z,\theta) \frac{1}{(-\partial_z [X^{+0}(z,\theta)/ip^{+0}])^{1/2}} \right. \\
&\overbrace{\qquad\qquad R \qquad\qquad} \\
&\left. \times \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)\right) D_\theta \frac{X^{+0}(z,\theta)}{ip^{+0}} \cdot D_{\theta'}^{3-m} X^{-0}(z',\theta') \cdot D_{\theta'}^{m} X^{+0}(z',\theta') :_R ,
\end{aligned}
\tag{A12}
$$

$$
\begin{aligned}
H_3 = \sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : &\left(D_\theta X^{+j}(z,\theta) \frac{\frac{1}{2} D_\theta [X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z [X^{+0}(z,\theta)/ip^{+0}])^{3/2}} \right. \\
&\overbrace{\qquad\qquad R \qquad\qquad} \\
&\left. \cdot \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)\right) \partial_z \frac{X^{+0}(z,\theta)}{ip^{+0}} \cdot D_{\theta'}^{3-m} X^{-0}(z',\theta') \cdot D_{\theta'}^{m} X^{+0}(z',\theta') :_R
\end{aligned}
\tag{A13}
$$

and

$$
\begin{aligned}
H_4 = -\sum_{m=1}^{2} \oint_{z'} \frac{dz\, d\theta}{2\pi i} : &\left(\frac{D_\theta [X^{+0}(z,\theta)/ip^{+0}]}{(-\partial_z [X^{+0}(z,\theta)/ip^{+0}])^{1/2}} \right. \\
&\overbrace{\qquad\qquad R \qquad\qquad} \\
&\left. \times \exp\left(\frac{-nX^{+0}(z,\theta)}{ip^{+0}}\right)\right) D_\theta X^{+j}(z,\theta) D_{\theta'}^{3-m} X^{-j}(z',\theta') \cdot D_{\theta'}^{m} X^{+j}(z',\theta') :_R ,
\end{aligned}
\tag{A14}
$$

Then we obtain that

$$H_1=\sum_{m=1}^{2}-n(-)^m:D_{\theta'}^{2-m}\Bigg(D_{\theta'}\,\frac{X^{+j}(z',\theta')}{ip^{+0}}\cdot\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{1/2}}\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)D_{\theta'}^{m}X^{+0}(z',\theta'):_R \tag{A15}$$

$$H_2=\sum_{m=1}^{2}(-)^m:D_{\theta'}^{3-m}\Bigg(\frac{D_{\theta'}[X^{+j}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{1/2}}\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)D_{\theta'}^{m}X^{+0}(z',\theta'):_R \tag{A16}$$

$$H_3=-\sum_{m=1}^{3}(-)^m:D_{\theta'}^{4-m}\Bigg(D_{\theta'}\,\frac{X^{+j}(z',\theta')}{ip^{+0}}\cdot\frac{\frac{1}{2}D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)D_{\theta'}^{m}X^{+0}(z',\theta'):_R \tag{A17}$$

and

$$H_4=-\sum_{m=1}^{2}(-)^m:D_{\theta'}^{3-m}\times\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{1/2}}\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)D_{\theta'}^{m}X^{+j}(z',\theta'):_R. \tag{A18}$$

By straightforward calculations, we find that

$$\sum_{m=1}^{4}H_m=0 \tag{A19}$$

with the help of the following formula:

$$\begin{aligned}
&\partial_{z'}D_{\theta'}\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\cdot\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)\cdot D_{\theta'}\,\frac{X^{+0}(z',\theta')}{ip^{+0}}\\
&=\partial_{z'}D_{\theta'}\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\cdot\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\cdot D_{\theta'}\,\frac{X^{+0}(z',\theta')}{ip^{+0}}\Bigg)\\
&\quad-\partial_{z'}\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\cdot\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)\cdot\partial_{z'}\,\frac{X^{+0}(z',\theta')}{ip^{+0}}\\
&\quad+D_{\theta'}\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\cdot\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)\cdot\partial_{z'}D_{\theta'}\,\frac{X^{+0}(z',\theta')}{ip^{+0}}\\
&\quad-\Bigg(\frac{D_{\theta'}[X^{+0}(z',\theta')/ip^{+0}]}{(-\partial_{z'}[X^{+0}(z',\theta')/ip^{+0}])^{3/2}}\cdot\exp\Bigg(\frac{-nX^{+0}(z',\theta')}{ip^{+0}}\Bigg)\Bigg)\cdot\partial_{z'}^{2}\,\frac{X^{+0}(z',\theta')}{ip^{+0}}.
\end{aligned}\tag{A20}$$

It should be noticed that the first term on the right-hand side of Eq. (A20) vanishes identically. (Q.E.D.)

APPENDIX B: SOME EXPLICIT FORMULAS IN THE *R* SECTOR

Functions $R_m^{(n)}(z_0)$'s and $R_m^{(n+1/2)}(z_0)$'s in Eq. (1.23) can be calculated, respectively, by using recursive relations

$$R_0^{(n)}(z_0)=1 \quad \text{and} \quad R_{m+1}^{(n)}(z_0)=\frac{1}{m+1}\,(\partial_{z_0}+nI_1)R_m^{(n)}(z_0), \tag{B1}$$

and

$$R_0^{(n+1/2)}(z_0)=1 \quad \text{and} \quad R_{m+1}^{(n+1/2)}(z_0)=\frac{1}{m+1}\,(\partial_{z_0}+(n+1/2)I_1)R_m^{(n+1/2)}(z_0), \tag{B1$'$}$$

in terms of

$$I_m\equiv\partial_{z_0}^m I(z_0) \quad \text{and} \quad I_m^n\equiv(I_m)^n \quad \text{with Eq. (1.22).} \tag{B2}$$

With the help of Eq. (B1), we find that

$$\begin{pmatrix} R_1^{(1)} \\ R_1^{(2)} \\ R_1^{(3)} \end{pmatrix}=\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} I_1, \tag{B3}$$

$$\begin{pmatrix} R_2^{(1)} \\ R_2^{(2)} \\ R_2^{(3)} \end{pmatrix}=\begin{pmatrix} \frac{1}{2} \\ 2 \\ \frac{9}{2} \end{pmatrix} I_1^2+\begin{pmatrix} \frac{1}{2} \\ 1 \\ \frac{3}{2} \end{pmatrix} I_2, \tag{B4}$$

$$\begin{pmatrix} R_3^{(1)} \\ R_3^{(2)} \\ R_3^{(3)} \end{pmatrix}=\begin{pmatrix} \frac{1}{6} \\ \frac{4}{3} \\ \frac{9}{2} \end{pmatrix} I_1^3+\begin{pmatrix} \frac{1}{2} \\ 2 \\ \frac{9}{2} \end{pmatrix} I_1I_2+\begin{pmatrix} \frac{1}{6} \\ \frac{1}{3} \\ \frac{1}{2} \end{pmatrix} I_3, \tag{B5}$$

and

$$\begin{pmatrix} R_4^{(1)} \\ R_4^{(2)} \\ R_4^{(3)} \end{pmatrix}=\begin{pmatrix} \frac{1}{24} \\ \frac{2}{3} \\ \frac{27}{8} \end{pmatrix} I_1^4+\begin{pmatrix} \frac{1}{4} \\ 2 \\ \frac{27}{4} \end{pmatrix} I_1^2I_2+\begin{pmatrix} \frac{1}{6} \\ \frac{2}{3} \\ \frac{3}{2} \end{pmatrix} I_1I_3+\begin{pmatrix} \frac{1}{8} \\ \frac{1}{2} \\ \frac{9}{8} \end{pmatrix} I_2^2+\begin{pmatrix} \frac{1}{24} \\ \frac{1}{12} \\ \frac{1}{8} \end{pmatrix} I_4, \tag{B6}$$

while we have from Eq. (B1$'$) that

$$\begin{pmatrix} R_1^{(1/2)} \\ R_1^{(1/2+1)} \\ R_2^{(1/2+2)} \end{pmatrix}=\begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \\ \frac{5}{2} \end{pmatrix} I_1 \tag{B7}$$

and

$$\begin{pmatrix} R_2^{(1/2)} \\ R_2^{(1/2+1)} \\ R_2^{(1/2+2)} \end{pmatrix}=\begin{pmatrix} \frac{1}{8} \\ \frac{9}{8} \\ \frac{25}{8} \end{pmatrix} I_1^2+\begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \\ \frac{5}{4} \end{pmatrix} I_2. \tag{B8}$$

Substituting Eqs. (B3)–(B8) into Eqs. (3.14) and (3.15) we find that

$$\begin{aligned} A_N^F(w,w')=&\,w(I_1+\theta_0V_1)+w'I_1+w^2(I_1^2+\tfrac{1}{2}I_2+\theta_0(V_2+2I_1V_1))+w'^2(I_1^2+\tfrac{1}{2}I_2)+ww'(\tfrac{1}{2}I_1^2+\tfrac{1}{2}I_2 \\ &+\theta_0(V_2+I_1V_1))+w^3(\tfrac{3}{2}I_1^3+\tfrac{3}{2}I_1I_2+\tfrac{1}{6}I_3+\theta_0(V_3+3I_1V_2+(\tfrac{9}{2}I_1^2+\tfrac{3}{2}I_2)V_1)) \\ &+w^2w'(\tfrac{2}{3}I_1^3+I_1I_2+\tfrac{1}{6}I_3+V_1V_2+\theta_0(V_3+3I_1V_2+2I_1^2V_1))+ww'^2(\tfrac{2}{3}I_1^3+I_1I_2+\tfrac{1}{6}I_3 \\ &+\theta_0(V_3+2I_1V_2+(2I_1^2+I_2)V_1))+w'^3(\tfrac{3}{2}I_1^3+\tfrac{3}{2}I_1I_2+\tfrac{1}{6}I_3)+\cdots \end{aligned} \tag{B9}$$

and

$$\log \frac{L_{NF}(w)}{w} = wI_1 + w^2(I_1^2 + \tfrac{1}{2}I_2) + w^3(\tfrac{3}{2}I_1^3 + \tfrac{3}{2}I_1I_2 + \tfrac{1}{6}I_3) + \cdots . \tag{B10}$$

Furthermore, Eqs. (3.19) and (3.20) together with Eqs. (B3)–(B8) give

$$\begin{aligned}\Gamma_N^{F2}(w,w') &= w'^{-1/2}\theta_0 + ww'^{-1/2}(V_1 + \theta_0 I_1) + w'^{-1/2+1}(V_1 + \tfrac{1}{2}\theta_0 I_1) + w^2 w'^{-1/2}(V_2 + 2I_1V_1 \\ &\quad + \theta_0(2I_1^2 + 2I_2)) + ww'^{-1/2+1}(V_2 + \tfrac{1}{2}I_1V_1 + \theta_0(\tfrac{1}{2}I_1^2 + I_2)) \\ &\quad + w^3 w'^{-1/2}(V_3 + 3I_1V_2 + (\tfrac{9}{2}I_1^2 + \tfrac{3}{2}I_2)V_1 + \theta_0(\tfrac{9}{2}I_1^3 + \tfrac{9}{2}I_1I_2 + \tfrac{1}{2}I_3)) \\ &\quad + w^2 w'^{-1/2+1}(V_3 + \tfrac{3}{2}I_1V_2 + (I_1^2 + I_2)V_1 + \theta_0(I_1^3 + \tfrac{5}{2}I_1I_2 + \tfrac{1}{2}I_3)) \\ &\quad + ww'^{-1/2+2}(V_3 + \tfrac{3}{2}I_1V_2 + (\tfrac{9}{8}I_1^2 + \tfrac{1}{4}I_2)V_1 + \theta_0(\tfrac{9}{4}I_1^3 + \tfrac{9}{4}I_1I_2 + \tfrac{1}{2}I_3)) \\ &\quad + w'^{-1/2+3}(V_3 + \tfrac{5}{2}I_1V_2 + (\tfrac{25}{8}I_1^2 + \tfrac{5}{4}I_2)V_1 + \theta_0(\tfrac{128}{48}I_1^3 + \tfrac{75}{24}I_1I_2 + \tfrac{5}{12}I_3)) + \cdots ,\end{aligned} \tag{B11}$$

while Eqs. (3.22) and (3.23) together with Eqs. (B3)–(B8) give

$$\begin{aligned}\Gamma_N^{F1}(w,w') &= w^{-1/2+1}(V_1 + \theta_0 I_1) + w^{-1/2+1}w'(V_2 + I_1V_1 + \theta_0(I_1^2 + I_2)) + w^{-1/2+2}(V_2 + \tfrac{3}{2}I_1V_1 \\ &\quad + \theta_0(\tfrac{3}{2}I_1^2 + I_2)) + w^{-1/2+1}w'^2(V_3 + 2I_1V_2 + (2I_1^2 + I_2)V_1 + \theta_0(2I_1^3 + 3I_1I_2 + \tfrac{1}{2}I_3)) \\ &\quad + w^{-1/2+2}w'(V_3 + \tfrac{3}{2}I_1V_2 + (I_1^2 + \tfrac{1}{2}I_2)V_1 + \theta_0(\tfrac{5}{12}I_1^3 + \tfrac{3}{4}I_1I_2 + \tfrac{1}{6}I_3)) \\ &\quad + w^{-1/2+3}(V_3 + \tfrac{5}{2}I_1V_2 + (\tfrac{29}{8}I_1^2 + \tfrac{5}{4}I_2)V_1 + \theta_0(\tfrac{25}{8}I_1^3 + \tfrac{15}{4}I_1I_2 + \tfrac{1}{2}I_3)) + \cdots .\end{aligned} \tag{B12}$$

Then we have from Eqs. (3.6)–(3.8), (B9), and (B10) that

$$\begin{aligned}\left(\frac{w}{w'}\right)^{1/2} C_N^F(w,w') &= w^{-1/2+1}w'^{-1/2}(\tfrac{1}{2}I_1 - \tfrac{1}{2}\theta_0 V_1) + w^{-1/2+2}w'^{-1/2}((\tfrac{7}{8}I_1^2 + \tfrac{1}{2}I_2) - \tfrac{3}{4}\theta_0 I_1V_1) \\ &\quad + w^{-1/2+1}w'^{-1/2+1}(\tfrac{1}{8}I_1^2 - \theta_0(V_2 + \tfrac{5}{4}I_1V_1)) + w^{-1/2+3}w'^{-1/2}(\tfrac{5}{6}I_1^3 + \tfrac{15}{8}I_1I_2 \\ &\quad + \tfrac{1}{4}I_3 + \tfrac{1}{2}V_1V_2 + \theta_0(\tfrac{1}{2}V_3 + \tfrac{3}{4}I_1V_2 - (I_1^2 + \tfrac{3}{8}I_2)V_1)) + w^{-1/2+2}w'^{-1/2+1} \\ &\quad \times(\tfrac{3}{4}I_1^3 + \tfrac{3}{4}I_1I_2 + \tfrac{1}{12}I_3 - \tfrac{1}{2}V_1V_2 - \theta_0(\tfrac{1}{2}V_3 + 2I_1V_2 + (\tfrac{9}{4}I_1^2 + \tfrac{1}{4}I_2)V_1)) \\ &\quad + w^{-1/2+1}w'^{-1/2+2}(-\tfrac{1}{6}I_1^3 - \tfrac{3}{8}I_1I_2 - \tfrac{1}{12}I_3 - \tfrac{1}{2}V_1V_2 - \theta_0(\tfrac{3}{2}V_3 + \tfrac{15}{4}I_1V_2 \\ &\quad + (\tfrac{7}{2}I_1^2 + \tfrac{9}{8}I_2)V_1)) + \cdots .\end{aligned} \tag{B13}$$

Then the function $K_{NF}(w)$ is evaluated by Eqs. (3.9) and (B9) as

$$\begin{aligned}K_{NF}(w) &= 1 + w(I_1 + \tfrac{1}{2}\theta_0 V_1) + w^2(\tfrac{7}{4}I_1^2 + \tfrac{3}{4}I_2 + \theta_0(V_2 + 2I_1V_1)) + w^3(\tfrac{43}{12}I_1^3 + \tfrac{13}{4}I_1I_2 + \tfrac{1}{3}I_3 + \tfrac{1}{2}V_1V_2 \\ &\quad + \theta_0(\tfrac{3}{2}V_3 + 5I_1V_2 + (\tfrac{53}{8}I_1^2 + \tfrac{13}{8}I_2)V_1)) + \cdots .\end{aligned} \tag{B14}$$

We find from Eqs. (B11) and (B14) that

$$\begin{aligned}\Gamma_N^{F2}(w,w')K_{NF}(w') &= w'^{-1/2}\theta_0 + ww'^{-1/2}(V_1 + \theta_0 I_1) + w'^{-1/2+1}(V_1 + \tfrac{3}{2}\theta_0 I_1) + w^2 w'^{-1/2}(V_2 \\ &\quad + 2I_1V_1 + \theta_0(2I_1^2 + 2I_2)) + ww'^{-1/2+1}(V_2 + \tfrac{3}{2}I_1V_1 + \theta_0(\tfrac{3}{2}I_1^2 + I_2)) \\ &\quad + w'^{-1/2+2}(V_2 + \tfrac{3}{2}I_1V_1 + \theta_0(\tfrac{11}{4}I_1^2 + \tfrac{7}{4}I_2)) + w^3 w'^{-1/2}(V_3 + 3I_1V_2 + (\tfrac{9}{2}I_1^2 \\ &\quad + \tfrac{3}{2}I_2)V_1 + \theta_0(\tfrac{9}{2}I_1^3 + \tfrac{9}{2}I_1I_2 + \tfrac{1}{2}I_3)) + w^2 w'^{-1/2+1}(V_3 + \tfrac{5}{2}I_1V_2 + (I_1^2 + I_2)V_1\end{aligned}$$

$$+\theta_0(\tfrac{5}{2}I_1^3+\tfrac{9}{2}I_1I_2+\tfrac{1}{2}I_3+\tfrac{1}{2}N_1N_2))+ww'^{-1/2+2}(V_3+\tfrac{3}{2}I_1V_2+(\tfrac{9}{8}I_1^2+\tfrac{1}{4}I_2)V_1$$
$$+\theta_0(\tfrac{9}{4}I_1^3+\tfrac{9}{4}I_1I_2+\tfrac{1}{2}I_3))+w'^{-1/2+3}(V_3+\tfrac{7}{2}I_1V_2+(\tfrac{39}{8}I_1^2+2I_2)V_1$$
$$+\theta_0(\tfrac{121}{16}I_1^3+\tfrac{31}{4}I_1I_2+\tfrac{1}{3}I_3))+\cdots, \tag{B15}$$

while Eqs. (B12) and (B14) lead to

$$\Gamma_N^{F1}(w,w')K_{NF}(w)=w^{-1/2+1}(V_1+\theta_0I_1)+w^{-1/2+1}w'(V_2+I_1V_1+\theta_0(I_1^2+I_2))+w^{-1/2+2}(V_2$$
$$+\tfrac{5}{2}I_1V_1+\theta_0(\tfrac{5}{2}I_1^2+I_2))+w^{-1/2+1}w'^2(V_3+2I_1V_2+(2I_1^2+I_2)V_1+\theta_0(2I_1^3$$
$$+3I_1I_2+\tfrac{1}{2}I_3))+w^{-1/2+2}w'(V_3+\tfrac{5}{2}I_1V_2+(2I_1^2+\tfrac{1}{2}I_2)V_1$$
$$+\theta_0(\tfrac{17}{12}I_1^3+\tfrac{7}{4}I_1I_2+\tfrac{1}{6}I_3+\tfrac{1}{2}V_1V_2))+w^{-1/2+3}(V_3+\tfrac{7}{2}I_1V_2+(\tfrac{63}{4}I_1^2+2I_2)V_1$$
$$+\theta_0(\tfrac{121}{16}I_1^3+\tfrac{31}{4}I_1I_2+\tfrac{1}{3}I_3))+\cdots. \tag{B16}$$

Moreover, with the help of Eq. (B14), we find that

$$[X^{\mp j}(z_{NF})]_{\xi=0}=X^{\mp j}(z_0)+\sum_{n=1}^{\infty}w^n\left(\sum_{k=0}^{n}\frac{k}{n}\cdot R_{n-k}^{(n)}(z_0)X_k^{\mp j}\right)$$
$$=X_0^{\mp j}+wX_1^{\mp j}+w^2(X_2^{\mp j}+I_1X_1^{\mp j})+\cdots, \tag{B17}$$

$$K_{NF}(w)[\theta_{NF}\partial_{z_{NF}}X^{\mp j}(z_{NF})]_{\xi=0}=K_{NF}(w)\sum_{n=0}^{\infty}w^{n+1/2}\left(\theta_0\left(\sum_{k=0}^{n}(k+1)R_{n-k}^{(n+1/2)}(z_0)X_{k+1}^{\mp j}\right)\right.$$
$$\left.+\sum_{k=1}^{n}R_{n-k}^{(n+1/2)}(z_0)\left(\sum_{l=1}^{k}l\cdot V_{k-l+1}X_l^{\mp j}\right)\right)$$
$$=w^{-1/2+1}\theta_0X_1^{\mp j}+w^{-1/2+2}(V_1X_1^{\mp j}+\theta_0(2X_2^{\mp j}+\tfrac{5}{2}I_1X_1^{\mp j}))$$
$$+w^{-1/2+3}(2V_1X_2^{\mp j}+(V_2+\tfrac{7}{2}I_1V_1)X_1^{\mp j}+\theta_0(3X_3^{\mp j}+7I_1X_2^{\mp j}$$
$$+(\tfrac{51}{8}I_1^2+2I_2)X_1^{\mp j}))+\cdots, \tag{B18}$$

$$[\theta_{NF}\sqrt{z_{NF}-z_0}\Psi_{N\pm}^{\pm J}(z_{NF})]_{\xi=0}=\sum_{n=1}^{\infty}w^n\left(\theta_0\left(\sum_{k=1}^{n}R_{n-k}^{(n)}(z_0)\Psi_{k,N\pm}^{\pm j}\right)+\sum_{k=2}^{n}R_{n-k}^{(n)}(z_0)\right.$$
$$\left.\times\left(\sum_{l=2}^{k}V_{k-l+1}\Psi_{l-1,N\pm}^{\pm j}\right)\right)$$
$$=w\theta_0\Psi_{1,N\pm}^{\pm j}+w^2(V_1\Psi_{1,N\pm}^{\pm j}+\theta_0(\Psi_{2,N\pm}^{\pm j}+2I_1\Psi_{1,N\pm}^{\pm j}))$$
$$+w^3(V_1\Psi_{2,N\pm}^{\pm j}+(V_2+3I_1V_1)\Psi_{1,N\pm}^{\pm j}+\theta_0(\Psi_{3,N\pm}^{\pm j}+3I_1\Psi_{2,N\pm}^{\pm j}$$
$$+(\tfrac{9}{2}I_1^2+\tfrac{3}{2}I_2)\Psi_{1,N\pm}^{\pm j}))+\cdots, \tag{B19}$$

$$K_{NF}(w)[\sqrt{z_{NF}-z_0}\Psi_{N\pm}^{\pm j}(z_{NF})]_{\xi=0}=K_{NF}(w)\sum_{n=0}^{\infty}w^{n+1/2}\left(\sum_{k=0}^{n}\frac{k+1/2}{n+1/2}\cdot R_{n-k}^{(n+1/2)}(z_0)\Psi_{k+1,N\pm}^{\pm j}\right)$$
$$=w^{-1/2+1}\Psi_{1,N\pm}^{\pm j}+w^{-1/2+2}(\Psi_{2,N\pm}^{\pm j}+\tfrac{3}{2}I_1\Psi_{1,N\pm}^{\pm j}+\tfrac{1}{2}\theta_0V_1\Psi_{1,N\pm}^{\pm j})$$
$$+w^{-1/2+3}(\Psi_{3,N\pm}^{\pm j}+\tfrac{5}{2}I_1\Psi_{2,N\pm}^{\pm j}+(\tfrac{23}{8}I_1^2+I_2)\Psi_{1,N\pm}^{\pm j}$$
$$+\theta_0(\tfrac{1}{2}V_1\Psi_{2,N\pm}^{\pm j}+(V_2+\tfrac{9}{4}I_1V_1)\Psi_{1,N\pm}^{\pm j}))+\cdots, \tag{B20}$$

$$\left[\theta_{NF}\frac{\Psi^{\mp j}_{N\pm}(z_{NF})}{\sqrt{z_{NF}-z_0}}\right]_{\xi=0}=\sum_{n=1}^{\infty}w^{n-1}\Bigg(\theta_0\Bigg(\sum_{k=1}^{n}R^{(n-1)}_{n-k}(z_0)\Psi^{\mp j}_{k-1,N\pm}\Bigg)+\sum_{k=2}^{n}R^{(n-1)}_{n-k}(z_0)\times\Bigg(\sum_{l=2}^{k}V_{k-l+1}\Psi^{\mp j}_{l-2,N\pm}\Bigg)\Bigg)$$
$$=\theta_0\Psi^{\mp j}_{0,N\pm}+w(V_1\Psi^{\mp j}_{0,N\pm}+\theta_0(\Psi^{\mp j}_{1,N\pm}+I_1\Psi^{\mp j}_{0,N\pm}))+w^2(V_1\Psi^{\mp}_{1,N\pm}+(V_2+2I_1V_1)\Psi^{\mp j}_{0,N\pm}+\theta_0(\Psi^{\mp j}_{2,N\pm}+2I_1\Psi^{\mp j}_{1,N\pm}+(2I_1^2+I_2)\Psi^{\mp j}_{0,N\pm}))+\cdots \tag{B21}$$

and

$$K_{NF}(w)\left[\frac{\Psi^{\mp j}_{N\pm}(z_{NF})}{\sqrt{z_{NF}-z_0}}\right]_{\xi=0}=K_{NF}(w)\sum_{n=0}^{\infty}w^{n-1/2}\Bigg(\sum_{l=0}^{n}\frac{2l-1}{2n-1}\cdot R^{(n-l/2)}_{n-l}(z_0)\Psi^{\mp j}_{l,N\pm}\Bigg)$$
$$=w^{-1/2}\Psi^{\mp j}_{0,N\pm}+w^{-1/2+1}(\Psi^{\mp j}_{1,N\pm}+\tfrac{1}{2}I_1\Psi^{\mp j}_{0,N\pm}+\tfrac{1}{2}\theta_0V_1\Psi^{\mp j}_{0,N\pm})+w^{-1/2+2}(\Psi^{\mp j}_{2,N\pm}+\tfrac{3}{2}I_1\Psi^{\mp j}_{1,N\pm}+(\tfrac{7}{8}I_1^2+\tfrac{1}{2}I_2)\Psi^{\mp j}_{0,N\pm}+\theta_0(\tfrac{1}{2}V_1\Psi^{\mp j}_{1,N\pm}+(V_2+\tfrac{7}{4}I_1V_1)\Psi^{\mp j}_{0,N\pm}))+\cdots . \tag{B22}$$

Finally, various R's defined by Eqs. (3.27)–(3.29) and (3.34)–(3.37) can be calculated as follows: We find from Eq. (3.34) substituted with Eqs. (3.30) and (B9) that

$$R^j_{\alpha\alpha Nh}=-\alpha^{-hj}_{-1}p^{hj}(I_1+\theta_0V_1)-p^{-hj}\alpha^{hj}_{-1}I_1-\alpha^{-hj}_{-2}p^{hj}(I_1^2+\tfrac{1}{2}I_2+\theta_0(V_2+2I_1V_1))-p^{-hj}\alpha^{hj}_{-2}(I_1^2+\tfrac{1}{2}I_2)-\alpha^{-hj}_{-1}\alpha^{hj}_{-1}(\tfrac{1}{2}I_1^2+\tfrac{1}{2}I_2+\theta_0(V_2+I_1V_1))-\alpha^{-hj}_{-3}p^{hj}(\tfrac{3}{2}I_1^3+\tfrac{3}{2}I_1I_2+\tfrac{1}{6}I_3+\theta_0(V_3+3I_1V_2+(\tfrac{9}{2}I_1^2+\tfrac{3}{2}I_2)V_1))-\alpha^{-hj}_{-2}\alpha^{hj}_{-1}(\tfrac{2}{3}I_1^3+I_1I_2+\tfrac{1}{6}I_3+V_1V_2+\theta_0(V_3+3I_1V_2+2I_1^2V_1))-\alpha^{-hj}_{-1}\alpha^{hj}_{-2}(\tfrac{2}{3}I_1^3+I_1I_2+\tfrac{1}{6}I_3+\theta_0(V_3+2I_1V_2+(2I_1^2+I_2)V_1))-p^{-hj}\alpha^{hj}_{-3}(\tfrac{3}{2}I_1^3+\tfrac{3}{2}I_1I_2+\tfrac{1}{6}I_3)+\cdots . \tag{B23}$$

On both sides of Eq. (B23), ''h'' is the one in ϵ^j_h in $S_{\{h\}}(z_0)$ (1.4). (We will use this notational convention hereafter.) With the help of Eq. (3.35) substituted with Eqs. (3.30), (3.31), and (B15), we find that

$$R^j_{\alpha dNh}=p^{-hj}d^{hj}_0\theta_0+\alpha^{-hj}_{-1}d^{hj}_0(V_1+\theta_0I_1)+p^{-hj}d^{hj}_{-1}(V_1+\tfrac{3}{2}\theta_0I_1)+\alpha^{-hj}_{-2}d^{hj}_0(V_2+2I_1V_1+\theta_0(2I_1^2+2I_2))+\alpha^{-hj}_{-1}d^{hj}_{-1}(V_2+\tfrac{3}{2}I_1V_1+\theta_0(\tfrac{3}{2}I_1^2+I_2))+p^{-hj}d^{hj}_{-2}(V_2+\tfrac{3}{2}I_1V_1+\theta_0(\tfrac{11}{4}I_1^2+\tfrac{7}{4}I_2))+\alpha^{-hj}_{-3}d^{hj}_0(V_3+3I_1V_2+(\tfrac{9}{2}I_1^2+\tfrac{3}{2}I_2)V_1+\theta_0(\tfrac{9}{2}I_1^3+\tfrac{9}{2}I_1I_2+\tfrac{1}{2}I_3))+\alpha^{-hj}_{-2}d^{hj}_{-1}(V_3+\tfrac{5}{2}I_1V_2+(I_1^2+I_2)V_1+\theta_0(\tfrac{5}{2}I_1^3+\tfrac{9}{2}I_1I_2+\tfrac{1}{2}I_3+\tfrac{1}{2}V_1V_2))+\alpha^{-hj}_{-1}d^{hj}_{-2}(V_3+\tfrac{3}{2}I_1V_2+(\tfrac{9}{8}I_1^2+\tfrac{1}{4}I_2)V_1+\theta_0(\tfrac{9}{4}I_1^3+\tfrac{9}{4}I_1I_2+\tfrac{1}{2}I_3))+p^{-hj}d^{hj}_{-3}(V_3+\tfrac{7}{2}I_1V_2+(\tfrac{39}{8}I_1^2+2I_2)V_1+\theta_0(\tfrac{121}{16}I_1^3+\tfrac{31}{4}I_1I_2+\tfrac{1}{3}I_3))+\cdots \tag{B24}$$

while Eq. (3.36) substituted with Eqs. (3.30), (3.31), and (B16) give

$$R^j_{d\alpha Nh}=d^{-hj}_{-1}p^{hj}(V_1+\theta_0 I_1)+d^{-hj}_{-1}\alpha^{hj}_{-1}(V_2+I_1V_1+\theta_0(I_1^2+I_2))+d^{-hj}_{-2}p^{hj}(V_2+\tfrac{5}{2}I_1V_1+\theta_0(\tfrac{5}{2}I_1^2+I_2))+d^{-hj}_{-1}\alpha^{hj}_{-2}(V_3+2I_1V_2+(2I_1^2+I_2)V_1+\theta_0(2I_1^3+3I_1I_2+\tfrac{1}{2}I_3))+d^{-hj}_{-2}\alpha^{hj}_{-1}(V_3+\tfrac{5}{2}I_1V_2+(2I_1^2+\tfrac{1}{2}I_2)V_1+\theta_0(\tfrac{17}{12}I_1^3+\tfrac{7}{4}I_1I_2+\tfrac{1}{6}I_3+\tfrac{1}{2}V_1V_2))+d^{-hj}_{-3}p^{hj}(V_3+\tfrac{7}{2}I_1V_2+(\tfrac{63}{4}I_1^2+2I_2)V_1+\theta_0(\tfrac{121}{16}I_1^3+\tfrac{31}{4}I_1I_2+\tfrac{1}{3}I_3))+\cdots . \tag{B25}$$

Furthermore, we have from Eq. (3.37) substituted with Eqs. (3.31) and (B13) that

$$R^j_{ddNh}=d^{-hj}_{-1}d^{hj}_{0}(\tfrac{1}{2}I_1-\tfrac{1}{2}\theta_0V_1)+d^{-hj}_{-2}d^{hj}_{0}(\tfrac{7}{8}I_1^2+\tfrac{1}{2}I_2-\tfrac{3}{4}\theta_0I_1V_1)+d^{-hj}_{-1}d^{hj}_{-1}(\tfrac{1}{8}I_1^2-\theta_0(V_2+\tfrac{5}{4}I_1V_1))+d^{-hj}_{-3}d^{hj}_{0}(\tfrac{5}{6}I_1^3+\tfrac{15}{8}I_1I_2+\tfrac{1}{4}I_3+\tfrac{1}{2}V_1V_2+\theta_0(\tfrac{1}{2}V_3+\tfrac{3}{4}I_1V_2-(I_1^2+\tfrac{3}{8}I_2)V_1))+d^{-hj}_{-2}d^{hj}_{-1}(\tfrac{3}{4}I_1^3+\tfrac{3}{4}I_1I_2+\tfrac{1}{12}I_3-\tfrac{1}{2}V_1V_2-\theta_0(\tfrac{1}{2}V_3+2I_1V_2+(\tfrac{9}{4}I_1^2+\tfrac{1}{4}I_2)V_1))+d^{-hj}_{-1}d^{hj}_{-2}(-\tfrac{1}{6}I_1^3-\tfrac{3}{8}I_1I_2-\tfrac{1}{12}I_3-\tfrac{1}{2}V_1V_2-\theta_0(\tfrac{3}{2}V_3+\tfrac{15}{4}I_1V_2+(\tfrac{7}{2}I_1^2+\tfrac{9}{8}I_2)V_1))+\cdots . \tag{B26}$$

On the other hand, Eq. (3.27) substituted with Eqs. (3.30), (3.31), (B17), and (B18) leads to

$$R^j_X=-\sum_{\pm} ip^{\pm j}X_0^{\mp j}-\sum_{\pm} i\alpha^{\pm j}_{-1}X_1^{\mp j}-\sum_{\pm} i\alpha^{\pm j}_{-2}(X_2^{\mp j}+I_1X_1^{\mp j})+\cdots+\sum_{\pm} id^{\pm j}_{-1}\theta_0X_1^{\mp j}+\sum_{\pm} id^{\pm j}_{-2}(V_1X_1^{\mp j}+\theta_0(2X_2^{\mp j}+\tfrac{5}{2}I_1X_1^{\mp j}))+\sum_{\pm} id^{\pm j}_{-3}(2V_1X_2^{\mp j}+(V_2+\tfrac{7}{2}I_1V_1)X_1^{\mp j}+\theta_0(3X_3^{\mp j}+7I_1X_2^{\mp j}+(\tfrac{51}{8}I_1^2+2I_2)X_1^{\mp j}))+\cdots . \tag{B27}$$

Furthermore, Eq. (3.28) substituted with Eqs. (3.30), (B19), and (B21) gives

$$R^j_{\alpha Nh}=-i\alpha^{-hj}_{-1}\theta_0\Psi^{hj}_{1,Nh}-i\alpha^{-hj}_{-2}(V_1\Psi^{hj}_{1,Nh}+\theta_0(\Psi^{hj}_{2,Nh}+2I_1\Psi^{hj}_{1,Nh}))-i\alpha^{-hj}_{-3}(V_1\Psi^{hj}_{2,Nh}+(V_2+3I_1V_1)\Psi^{hj}_{1,Nh}+\theta_0(\Psi^{hj}_{3,Nh}+3I_1\Psi^{hj}_{2,Nh}+(\tfrac{9}{2}I_1^2+\tfrac{3}{2}I_2)\Psi^{hj}_{1,Nh}))+\cdots-\theta_0ip^{hj}\Psi^{-hj}_{0,Nh}-i\alpha^{hj}_{-1}(V_1\Psi^{-hj}_{0,Nh}+\theta_0(\Psi^{-hj}_{1,Nh}+I_1\Psi^{-hj}_{0,Nh}))-i\alpha^{hj}_{-2}(V_1\Psi^{-hj}_{1,Nh}+(V_2+2I_1V_1)\Psi^{-hj}_{0,Nh}+\theta_0(\Psi^{-hj}_{2,Nh}+2I_1\Psi^{-hj}_{1,Nh}+(2I_1^2+I_2)\Psi^{-hj}_{0,Nh}))+\cdots . \tag{B28}$$

Finally, Eq. (3.29) substituted with Eqs. (3.31), (B20), and (B22) gives

$$R^j_{dNh}=+id^{-hj}_{-1}\Psi^{hj}_{1,Nh}+id^{-hj}_{-2}(\Psi^{hj}_{2,Nh}+\tfrac{3}{2}I_1\Psi^{hj}_{1,Nh}+\tfrac{1}{2}\theta_0V_1\Psi^{hj}_{1,Nh})+id^{-hj}_{-3}(\Psi^{hj}_{3,Nh}+\tfrac{5}{2}I_1\Psi^{hj}_{2,Nh}+(\tfrac{23}{8}I_1^2+I_2)\Psi^{hj}_{1,Nh}+\theta_0(\tfrac{1}{2}V_1\Psi^{hj}_{2,Nh}+(V_2+\tfrac{9}{4}I_1V_1)\Psi^{hj}_{1,Nh})+id^{hj}_{0}\Psi^{-hj}_{0,Nh}+id^{hj}_{-1}(\Psi^{-hj}_{1,Nh}+\tfrac{1}{2}I_1\Psi^{-hj}_{0,Nh}+\tfrac{1}{2}\theta_0V_1\Psi^{-hj}_{0,Nh})+id^{hj}_{-2}(\Psi^{-hj}_{2,Nh}+\tfrac{3}{2}I_1\Psi^{-hj}_{1,Nh}+(\tfrac{7}{8}I_1^2+\tfrac{1}{2}I_2)\Psi^{-hj}_{0,Nh}+\theta_0(\tfrac{1}{2}V_1\Psi^{-hj}_{1,Nh}+(V_2+\tfrac{7}{4}I_1V_1)\Psi^{-hj}_{0,Nh}))+\cdots . \tag{B29}$$

In *physical* Ramond states in $\frac{1}{2}$ picture (3.52), operators $d_0^{\pm j}$'s existing in Eqs. (B24), (B26), and (B29) vanish since light-cone states in the R sector are those satisfying[4]

$$d_0^{\pm j}|\{\hat{p}\};\{h\}\rangle_F=0 \quad (\text{for } \epsilon^j_h=\pm\tfrac{1}{2}). \tag{B30}$$

[1] P. Ramond, Phys. Rev. D **3**, 2415 (1971).
[2] A. Neveu and J. H. Schwarz, Nucl. Phys. B **31**, 86 (1971); Phys. Rev. D **4**, 1109 (1971). For reviews of RNS (as well as GS) superstring theory, see, M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, New York, 1987), Vols. 1 and 2; M. Kaku, *Introduction to Superstring* (Springer-Verlag, New York, 1988).
[3] M. B. Green and J. H. Schwarz, Nucl. Phys. B **181**, 502 (1981); **189**, 252 (1982); **198**, 441 (1982); Phys. Lett. B **109**, 444 (1982). As for quantum superstring field theory (in GS superstring), see M. B. Green and J. H. Schwarz, Nucl. Phys. B **219**, 437 (1983); **243**, 475 (1984) and various papers cited therein.
[4] S. Naito, J. Math. Phys. **38**, 4980 (1997).

[5] S. Naito, ''Operator product expansions and super-stress operator,'' J. Math. Phys. (submitted).
[6] D. Friedan, E. Martinec, and S. Schenker, Nucl. Phys. B **271**, 93 (1986).
[7] S. Naito, J. Math. Phys. **34**, 2166 (1993).
[8] E. Del Giudice, P. Di Vecchia, and S. Fubini, Ann. Phys. (N.Y.) **70**, 378 (1972).
[9] J. H. Schwarz, Nucl. Phys. B **46**, 61 (1972); L. Brink and J. O. Winnberg, *ibid.* **103**, 455 (1976); K. Hornfeck, *ibid.* **293**, 189 (1987).
[10] F. Gliozzi, J. Scherk, and D. Olive, Nucl. Phys. B **122**, 253 (1977).

On topology and quantum mechanics

Karl-Georg Schlesinger
Department of Mathematics, University of Wuppertal, D-42097 Wuppertal, Germany

(Received 23 May 1997; accepted for publication 4 August 1997)

The relationship between topology and quantum mechanics is considered in two different ways. We suggest regarding quantum mechanics as a first-order approximation to topology theory on a fixed set. In this spirit, the approach of quantizing the lattice $\tau(X)$ of topologies on a set X [C. J. Isham, ''Quantum topology and quantization on the lattice of topologies,'' Class. Quantum. Grav. **6**, 1509–1534 (1989)] is shown to be easily applicable to a quantum mechanical lattice. In the main part of the paper we deal with the question of the quantization of the whole category of topological spaces and continuous injective maps. Ideas of the modern categorical approach to quantization, as known from topological quantum field theory, are invoked in this part. The final paragraphs are devoted to the problem of extending quantization, even to the level of the underlying set theory. © *1998 American Institute of Physics.* [S0022-2488(98)02302-0]

I. INTRODUCTION

In recent years, there have been several contributions to the question of the interplay between topology and quantum mechanics. Most prominent is certainly the use of quantum field theory to construct topological invariant of manifolds. On the other hand, the idea that topology itself should be quantized has been around for quite some time. Originating from Ref. 1, it has been first put into a rigorous mathematical form in Ref. 2 (see also Ref. 3). In this paper, we will deal with the relationship of topology and quantum mechanics in two different ways: First, we suggest that quantum mechanics may in some sense be regarded as a first-order version of the theory of topology (full topology theory is second order since it involves quantification over variables ranging over the power set of a given set). This is the content of Sec. II. As a consequence, it is tempting to quantize quantum mechanics along the lines Isham in Ref. 2 quantizes the lattice $\tau(X)$ of topologies on a given set X. We do this in Sec. III. At the end of this section, we shortly compare the result to the idea of second quantization. Second, in Sec. IV we generalize the idea of quantizing topology to the case of underlying sets of varying cardinality [while in the quantization of $\tau(X)$ the underlying set X is fixed]. We arrive at something like a kinematical structure for the general case by mixing the approach of Ref. 2 with some ideas of quantization by categorification (see Refs. 4–7). We compare the results to the mathematical structure known from topological quantum field theory (see the literature just cited above). Section V contains some final remarks. Parts of this work (especially Sec. II) are motivational in character and because of this the arguments are sometimes of a heuristic type.

Since the approach presented in Ref. 2 influenced some sections of this paper, we first mention briefly the concepts and results needed from there. For a fixed set X let $\tau(X)$ be the set of all topologies on X. $\tau(X)$ gains a lattice structure from the partial ordering $\leqslant$ defined by $\tau_1 \leqslant \tau_2$ iff the topology τ_2 is finer than τ_1. The lattice operations $\wedge$ and $\vee$ are then defined as the intersection and the topological span of τ_1 and τ_2, respectively. Configuration variables

$$R_\tau : \tau(X) \to \{0;1\}$$

are defined for $\tau \in \tau(X)$ by

$$R_\tau(\tau') = \begin{cases} 1, & \text{if } \tau \leqslant \tau', \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

 © 1998 American Institute of Physics

satisfying $R_{\tau_1}R_{\tau_2}=R_{\tau_1\vee\tau_2}$. The quantum theory of $\tau(X)$ then involves a representation of the algebra, generated by these functions, in terms of linear operators $\hat{R}_\tau$, $\tau\in\tau(X)$, on some suitable Hilbert space. While naively one might suspect the appropriate Hilbert space to consist of wave functions on $\tau(X)$, the spectral theory of the $\hat{R}_\tau$ shows that the domain space of the wave functions is actually not $\tau(X)$ itself but the set of ideals of the lattice $\tau(X)$ [in which $\tau(X)$ is naturally embedded]. Isham interprets the extra elements of the domain space as the analog of distributions appearing in the quantization of conventional field theories. Complementary variables M_τ with

$$M_\tau M_{\tau'}=M_{\tau\wedge\tau'}$$

are introduced by requiring that $\hat{M}_\tau$ should act on wave functions according to

$$(M_\tau\psi)(\tau')=\psi(\tau\wedge\tau') \tag{2}$$

and a suitable generalization to general ideals of $\tau(X)$.

II. QUANTUM MECHANICS AS FIRST-ORDER THEORY OF TOPOLOGY

For X as above, let $\alpha(X)$ be the sublattice of $\tau(X)$ containing the topologies generated by the singleton subsets of X, i.e., an element τ of $\alpha(X)$ is generated by a family

$$\{x_i\}_{i\in I},$$

with I an index set, $x_i\in X$, where the $\{x_i\}$ are open with respect to τ.

Lemma 1: *Let X be an arbitrary set, $\alpha(X)$ as above. There exists a pre-Hilbert space H with $\alpha(X)$ isomorphic to the lattice of subspaces of H generated by a fixed algebraic base of H.*

Proof: Let $\tau\in\alpha(X)$. Then τ is uniquely determined by a family of the form

$$\{x_i\}_{i\in I}, \tag{3}$$

$x_i\in X$. Define H to be the free complex vector space over X and a scalar product on H by

$$(x,y)=\begin{cases}1, & \text{if } x=y,\\ 0, & \text{if } x\neq y,\end{cases}$$

for x, $y\in X$ with the obvious extension to H. Then $\alpha(X)$ is isomorphic to the lattice of subspaces of H generated by X as a fixed base of H. □

Since H uniquely determines a Hilbert space $\overline{H}$ by completion, this means that $\alpha(X)$ determines an extension to a quantum-mechanical lattice of closed linear subspaces of a Hilbert space. Topologies in $\alpha(X)$ are determined by giving a family of points of X while for a general topology in $\tau(X)$ one has to give a family of subsets of X. So, $\alpha(X)$ may be considered as the first-order part of topology theory on X. It is tempting to speculate that including quantum superpositions of the elements of $\alpha(X)$ (which is mainly what happens when going to the subspace lattice of $\overline{H}$) might give some kind of first-order approximation to the full lattice $\tau(X)$. We will try to make this idea more precise now.

Let M be a smooth manifold (connected, Hausdorff, paracompact) with underlying set X. It is known that the C^∞-scalar functions on M generate the topology of M in the following sense (see, e.g., Ref. 8): Every open set of M has the form

$$\{y\in M\,|\,f(y)\neq 0\},$$

for some scalar function f. Since we can generate the same topology by restricting to scalar functions of compact support, we can generate the topology of M by the C^∞ functions, which are square integrable with respect to some suitable measure, too, i.e., by a very simple model of quantum mechanics on M (of course, on a topologically nontrivial manifold M, ''true'' quantum mechanics is not as simple as that).

Generalized wave functions in the form of Dirac delta distributions then generate the topologies belonging to $\alpha(X)$. Now, a square integrable wave function can be represented as a formal linear combination of delta distributions. So, the formal span of the delta distributions contains all the manifold topologies one can have on X. On the other hand, the relationship of the span of the delta distributions to $\alpha(X)$ is similar to the one of the subspace lattice of $\bar{H}$ to $\alpha(X)$ (The formal argument used here can be made precise by considering formal sums as words and going to a free vector space again.)

But the formal span of the delta distributions does not only contain the wave functions representing manifold topologies on X. It is also known that the C^* algebra of continuous complex-valued functions on a compact Hausdorff topological space allows for the reconstruction of the topology. So the argument above can be extended to compact nonmanifold topologies.

It is this sense—mediated by reinterpreting the lattice $\alpha(X)$ as representing the delta distributions on some manifold M—in which we regard the quantum-mechanical lattices as giving a first-order approximation (at least suitable for compact Hausdorff and manifold topologies) to topology theory, or put the other way around: $\tau(X)$ as the higher-order generalization of quantum mechanics. One may ask the question if the success with which quantum field theory can be applied to questions of topology can in the long run be understood as a consequence of a relationship like this between the two theories. Seen from this point of view, the idea of quantizing the lattice $\tau(X)$ itself, as followed in Ref. 2, appears to be very natural, reminding one of higher or iterative quantization. It is then tempting to apply the techniques, used in Ref. 2 to quantize $\tau(X)$, to a quantum-mechanical lattice. This is the content of the next section.

III. QUANTIZING QUANTUM MECHANICS

In a quantum-mechanical lattice L (i.e., a lattice of closed linear subspaces of a Hilbert space) the partial ordering $\leqslant$ is given by inclusion of subspaces and the lattice operations $\wedge$ and $\vee$ mean set–theoretic intersection and the closure of the linear span, respectively. We can again introduce functions,

$$R_\tau : L \to \{0;1\},$$

for $\tau \in L$ using (1). Again,

$$R_{\tau_1} \cdot R_{\tau_2} = R_{\tau_1 \vee \tau_2},$$

and therefore

$$R_\tau^2 = R_\tau.$$

As in Ref. 2, the spectral theory of the operators $\hat{R}_\tau$ shows that the domain space of the wave functions is the set $I(L)$ of ideals of the lattice L. So, the Hilbert space representing the kinematical structure of the quantum theory of L should be of the form

$$\mathscr{L}^2(I(L), d\mu),$$

for some measure μ. Remember that an ideal of L is a subset J of L satisfying the following:

(i) for $\tau_1, \tau_2 \in J$ we have $\tau_1 \vee \tau_2 \in J$;

(ii) if $\tau_1 \leqslant \tau_2$ and $\tau_2 \in J$ then $\tau_1 \in J$, too.

Since for $\tau \in L$,

$$\downarrow(\tau) := \{\tau' \in L \,|\, \tau' \leqslant \tau\}$$

is an ideal, L is naturally embedded in $I(L)$.

Again, complementary operators $\hat{M}_\tau$ satisfying

$$\hat{M}_{\tau_1} \hat{M}_{\tau_2} = \hat{M}_{\tau_1 \wedge \tau_2},$$

can be introduced by their action on wave functions ψ on $I(L)$ as

$$(\hat{M}_\tau \psi)(J) = \psi(\downarrow(\tau) \cap J), \tag{4}$$

for $J \in I(L)$. (4) is just the generalization of (2) to the ideals of L. In Ref. 2 the operation of $\hat{M}_\tau$ on some eigenstate $|J\rangle$ of the $\hat{R}_{\tau'}$ operators [the eigenstates of the $\hat{R}_{\tau'}$ operators have the elements of $I(L)$ as labels] is calculated to be

$$\hat{M}_\tau |J\rangle = \left(\frac{1}{\mu(J)}\right)^{1/2} \sum_{\{J' | \downarrow(\tau) \cap J' = J\}} (\mu(J'))^{1/2} |J'\rangle \tag{5}$$

[where the sum is understood to be zero if the condition on the sum cannot be fulfilled for any $J' \in I(L)$] for the quantization of $\tau(X)$, and X a finite set. Since a quantum-mechanical lattice L is never finite, the sum has to be replaced by an integral in this case. But the qualitative features can already be seen from (5): The operators $\hat{M}_\tau$ applied to a state $|J\rangle$ lead to a mixed state where any state $|J'\rangle$ compatible with the intersection requirement

$$\downarrow(\tau) \cap J' = J, \tag{6}$$

is present. For simplicity, consider the case where J' and J represent just elements of L, too. Then (6) can be reformulated as

$$\tau \cap J' = J.$$

This means that any subspace J' is present in the sum where the part belonging to τ is identical to J. Given the subspaces τ and J, the application of $\hat{M}_\tau$ therefore in some sense involves a maximal loss of information.

So, while the kinematical structure is in close analogy to conventional schemes of second quantization (we arrive at a Hilbert space of generalizations of wave functions on L), the operators involved in the quantization of a quantum-mechanical lattice allow for additional statistical effects beyond the usual quantum ones. This might give a suitable framework for generalizations of quantum field theory involving transitions between pure and mixed states.

IV. QUANTIZING THE CATEGORY OF TOPOLOGICAL SPACES AND CONTINUOUS INJECTIVE MAPS

In the last two sections, we have tried to motivate the following standpoint: We have suggested viewing quantum mechanics as a first-order approximation to topology theory on a set X and consequently argued to regard the quantization of $\tau(X)$ as a very natural step, related to second quantization. The ease with which the techniques of Ref. 2 for quantizing $\tau(X)$ translate to the case of a quantum-mechanical lattice L, in our opinion, give additional support to this view.

But the fact that the cardinality of the set X, underlying the topological spaces considered, remains fixed appears to be artificial to us. If we argue that quantum effects in general relativity might cause the topology to change, why should they preserve the identity of individual points? This is especially questionable since the points only get an identity in the physical sense after choosing a special solution of general relativity (see, e.g., Ref. 9), whereas quantization has to be applied to the whole space of solutions. In this section, we will present some ideas on quantizing the category of topological spaces, thereby getting rid of the restriction on the cardinality of X.

Determining a category means determining a class of objects and a class of morphisms. So, actually besides giving the objects (topological spaces) we should also make precise which maps between topological spaces we want to include. Since we want to generalize the structure of $\tau(X)$—where refinement of topology is the natural ordering relation—taking arbitrary continuous maps does not seem to be appropriate. In Ref. 2 the following partial ordering is suggested for the class of topological spaces:

$$(X_1, \tau_1) \prec (X_2, \tau_2),$$

iff there exists a continuous injection from (X_2, τ_2) to (X_1, τ_1). This clearly generalizes the case of $\tau(X)$. So, we take as category the category **Topi** of topological spaces and continuous injective maps (the ''*i*'' in **Topi** standing for ''injective'').

Of course, the totality of all topological spaces forms a proper class. In the sequel we will handle this class as just another mathematical object. At first sight this may cause some unease, but one should not forget that this is common practice in category theory. To stand on set-theoretically firm foundations, it is always possible to invoke type theory on classes (see Ref. 10) or alternatively to assume that what we call classes are actually sets of some inaccessible cardinality.

We start by introducing an algebra of functions from the class of objects of **Topi** to $\{0;1\}$. The spectral theory of this algebra should suggest the kinematical structure for the quantum theory of **Topi**. For (X,τ) a topological space, we define again

$$R_{(X,\tau)}((X',\tau'))=\begin{cases}1, & \text{if } (X,\tau)\prec(X',\tau'),\\ 0, & \text{otherwise.}\end{cases} \tag{7}$$

Observe that the functions $R_{(X,\tau)}$ naturally involve the morphism structure of **Topi** in their definition, too.

Since the functions $R_{(X,\tau)}$ take values only in the set $\{0;1\}$, it follows that

$$R^2_{(X,\tau)}=R_{(X,\tau)}. \tag{8}$$

Now, suppose $(X_1,\tau_1)\prec(X_2,\tau_2)$. Then

$$R_{(X_1,\tau_1)}R_{(X_2,\tau_2)}=R_{(X_2,\tau_2)}, \tag{9}$$

because

$$R_{(X_1,\tau_1)}((X',\tau'))R_{(X_2,\tau_2)}((X',\tau'))=1,$$

iff $(X_1,\tau_1)\prec(X',\tau')$ and $(X_2,\tau_2)\prec(X',\tau')$. But this is precisely the case if

$$(X_2,\tau_2)\prec(X',\tau').$$

Equation (9) shows that though the class of all topological spaces does not involve a lattice structure, the functions $R_{(X,\tau)}$ possess a natural absorption property with respect to the partial ordering $\prec$. This has to be expected if the operator analogs of these functions are projection operators on eigenstates of topology, as (8) suggests. The spectral theory of the operators $\hat{R}_{(X,\tau)}$ is what has to be considered next.

The operators $\hat{R}_{(X,\tau)}$ are commuting projection operators, so we should have a set of complete simultaneous eigenvectors $|h\rangle$ with

$$\hat{R}_{(X,\tau)}|h\rangle=h(X,\tau)|h\rangle,$$

and h taking values in $\{0;1\}$. It follows from (9) that the functions h have an absorption property, too:

$$h(X_1,\tau_1)h(X_2,\tau_2)=h(X_2,\tau_2), \tag{10}$$

for $(X_1,\tau_1)\prec(X_2,\tau_2)$. Let

$$J_h:=\{(X,\tau)|h(X,\tau)=1\}.$$

If again $(X_1,\tau_1)\prec(X_2,\tau_2)$ and $(X_2,\tau_2)\in J_h$, Eq. (10) implies

$$(X_1,\tau_1)\in J_h.$$

Conversely, let J be a subclass of the class of all topological spaces having the property that $(X_2,\tau_2)\in J$ implies $(X_1,\tau_1)\in J$ for every $(X_1,\tau_1)\prec(X_2,\tau_2)$. (We will call subclasses of this kind *downward directed*.) Then

$$h_J(X,\tau)=\begin{cases}1, & \text{if } (X,\tau)\in J,\\ 0, & \text{if } (X,\tau)\notin J,\end{cases}$$

has the absorption property (10). By definition,

$$h_{J_h}=h,$$

i.e., there is a one to one correspondence between the functions h and downward directed subclasses J of the class of all topological spaces. Again we can identify the space (X,τ) with

$$\downarrow(X,\tau):=\{(X',\tau')\,|\,(X',\tau')\prec(X,\tau)\}.$$

Clearly, $\downarrow(X,\tau)$ is a downward directed class.

We handled the spectral theory of the algebra generated by the operators $\hat{R}_{(X,\tau)}$ in a merely heuristic manner again. Our main aim was to get an idea what the kinematical structure of the quantum theory of **Topi** might look like. Having achieved this, we will now proceed in a category theoretic manner which we feel is the appropriate language for dealing with the quantization of **Topi**.

On the level of categories, the domain space of the quantum theory should involve all subcategories of **Topi** that are determined by a downward directed subclass of the class of all topological spaces. So, we take as domain space the 2-category **ST** of all full subcategories of **Topi**, the class of objects of which is a downward directed class J. The 1-morphisms are the functors, of course, and the 2-morphisms the natural transformations on these (for the definition of a 2-category, see e.g., Ref. 4).

Taking the downward directed classes J as the domain space of the quantum theory has a natural interpretation. It means that the quantum theory does not distinguish between a topological space and an extension thereof. This is plausible from a physical standpoint: In physics it does not seem to make much sense to distinguish between a space as such and the same space as embedded into a larger one. Since for topological spaces—in contrast to connected, Hausdorff, paracompact manifolds—there is no notion of maximal extension, this directly leads to the introduction of proper classes into the theory. Remember that the downward directed classes J are the analogs of the ideals of the lattice $\tau(X)$ or the distributional fields to be introduced in the quantization of conventional field theories. We regard it as interesting that the process of quantization automatically brings the notion of topological space (which is ultimately a model for the physical universe) close to the notion of mathematical universe (proper classes and categories as universe of discourse).

Now, the kinematical structure of the quantum theory should be given in terms of wave functions on the domain space. What are wave functions on ST? The generalization of the function concept to categories is the notion of functor, of course. On the level of 2-categories, we get 2-functors. At least on the level of the categorical analog of finite-dimensional linear algebra (see Refs. 4 and 7), the role of $\mathbb{C}$ is taken by the category **Vect** of finite-dimensional vector spaces. Since we have no reason to restrict to finite dimensions at the moment, we should replace this by the category **V** of vector spaces and linear maps. A module category over **V** is a category M with a symmetric monoidal structure $\oplus$ and a functor,

$$\otimes:\mathbf{V}\times M\to M,$$

such that $\oplus$ and $\otimes$ satisfy the axioms of a module up to natural isomorphisms that are subject to the usual coherence laws. Actually, we should introduce the categorical analog of a Hilbert space structure on M (this is possible; see Ref. 4), but for simplicity—in order to get the flavor of the general ideas involved—we restrict to the linear structure in this paper. The module categories over **V** and the functors preserving the module structure form a category **2-V**. Including structure preserving natural transformations, we get a 2-category (the ''2'' in the symbol **2-V** just points at the fact that we are dealing with the 2-categorical analog of the category of vector spaces).

So, wave functions on **ST** should be the 2-functors from **ST** to **2-V**. We have refrained from discussing the question of square integrability of wave functions, so far. It is not clear if an analog of measure theory can be introduced on **ST**, but it might well be that this is not necessary at all.

In ordinary quantum mechanics we need square integrability because we need normalized wave functions in order to give a probalistic interpretation to them. But there is a different possibility for interpreting a 2-functor,

$$\mathscr{F}:\mathbf{ST}\rightarrow\mathbf{2\text{-}V},$$

namely as a generalization of a topological quantum field theory. Remember that we can understand the functors $\mathscr{F}$ as resulting from a generalization of the ''naive quantization'' of **Topi**, where wave functions would just have been functors from **Topi** to **V**. But a functor from **Topi** to **V** generalizes the concept of topological quantum field theory (for the definition see the literature cited at the beginning) in a double sense: First, the domain category is generalized from differentiable manifolds to arbitrary topological spaces. Second, the codomain category is generalized from finite-dimensional vector spaces to the infinite-dimensional case.

Observe that the 2-category $\mathbf{2\text{-}V}^{\mathbf{ST}}$ of 2-functors from **ST** to **2-V** naturally inherits the structure of a module 2-category (the 2-categorical analog of a module category) from **2-V**. For the quantum theory of **Topi**, $\mathbf{2\text{-}V}^{\mathbf{ST}}$ has to be regarded as the analog of the Hilbert space of conventional quantum mechanics. So the state space of this theory appears to be a space of generalizations of topological quantum field theories. Compare this to recent results suggesting that the states of quantum gravity should be related to topological quantum field theories (see Refs. 5, 6, and 11).

Remark: A single topological space (X,τ) can itself be considered as a category, the objects being the open sets and the morphisms given by inclusion. Remembering our considerations of Sec. II and the role played by quantum field theoretic methods in constructing topological invariants, it is tempting to ask what a space of quantum theories on (X,τ) might look like. In a very general sense, giving a quantum (field) theory on (X,τ) should mean attaching Hilbert spaces to local regions of (X,τ), subject to the existence of inclusion maps between spaces for the inclusion of local regions. This means we have a functor from (X,τ) to **H** (the category of Hilbert spaces and linear maps). This is similar to the concept of presheaf on (X,τ) that abstractly formulated (see Ref. 12) is a functor from the opposite category of (X,τ) (i.e., all arrows reversed) to the category **Set** of sets and functions. Since reversing the arrows is not essential from a conceptual standpoint, we are dealing with something like *quantum presheaves* on (X,τ). A candidate for the space of quantum (field) theories on (X,τ) is then given by

$$\mathbf{H}^{(X,\tau)},$$

the category of functors from (X,τ) to **H**.

Observe the analogy between $\mathbf{H}^{(X,\tau)}$ and $\mathbf{2\text{-}V}^{\mathbf{ST}}$. On passing from a single topological space to the category **Topi**, we get a similar object one level of categorification higher (i.e., 2-categories instead of 1-categories). This is mainly due to the fact that we had to pass to a larger domain space (namely downward directed classes) when considering the quantum theory.

This level structure with respect to categorification is again known from (extended) topological quantum field theories, where it appears on passing to the cobordism category (i.e., the domain category of the topological quantum field theory) of the next higher dimension (see again Refs. 4, 5 and 6). □

The next step would be to introduce operators and commutation relations on $\mathbf{2\text{-}V}^{\mathbf{ST}}$. In the class of all topological spaces there is no natural candidate for variables complementary to the $R_{(X,\tau)}$. Besides this, the fact that the elements of $\mathbf{2\text{-}V}^{\mathbf{ST}}$ may have an interpretation (in terms of generalizations of topological quantum field theories) quite different from that of the state vectors of ordinary quantum mechanics, suggests that we should not necessarily expect sensible operators on $\mathbf{2\text{-}V}^{\mathbf{ST}}$ to reflect the properties of a classically given algebra. We plan to deal with these questions in detail in a later publication.

Consider $\mathbf{2\text{-}V}^{\mathbf{ST}}$ as a category for the moment (forgetting about the 2-morphism level of structure). Then the following lemma, needed in the sequel, holds.

Lemma 2: The category $\mathbf{2\text{-}V}^{\mathbf{ST}}$ *has pullbacks.*

Proof: Suppose for $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}$ functors from **ST** to **2-V** and ν, η natural transformations, a diagram of the following form is given:

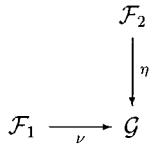

We define a new functor $\mathcal{F}'$ from **ST** to **2-V**. For C an object of **ST**, let $\mathcal{F}'(C)$ be the subcategory of $\mathcal{F}_1(C) \oplus \mathcal{F}_2(C)$ determined by the requirement that on $\mathcal{F}'(C)$, the functor identity

$$\nu_C \circ pr_1 = \eta_C \circ pr_2$$

holds, with pr_1 and pr_2 the projection functors of the Cartesian product (i.e., the direct sum). For a morphism f of **ST**, define

$$\mathcal{F}'(f) := \mathcal{F}_1(f) \oplus \mathcal{F}_2(f)|_{\mathcal{F}'(C)}.$$

Then the projection functors for each C together define a natural transformation from $\mathcal{F}'$ to $\mathcal{F}_1$ and from $\mathcal{F}'$ to $\mathcal{F}_2$, respectively. By construction, the diagram

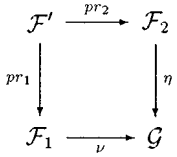

commutes. The universality of $\mathcal{F}'$ is checked by calculation using the properties of the Cartesian product in **2-V**. □

There is one point in the quantization of **Topi** as presented above that remains somewhat unsatisfactory. We will deal with this in the rest of this section now.

When quantizing **Topi**, we quantized the topologies but we kept the underlying set theory in which the category **Topi** is formulated in the classical form. We remarked already above that in general relativity the set-theoretic level only gets physical meaning after the choice of a special solution while quantization is to be applied to the whole space of solutions. This might be taken as a hint that even the underlying set theory cannot remain in its classical form in a fully satisfactory quantization of Einstein's theory. The question on which level to apply quantization has been discussed by Isham several times (see, e.g., Refs. 13 and 14). To stop quantization at the level of topology and to leave set theory unchanged causes special unease because of the close interplay between classical topology and set theory.

Facing these problems, the question arises as to what the appropriate quantum set theory should look like. We take again the standpoint of the category theorist, interpreting this question as meaning what the appropriate category of sets replacing **Set** should be. The candidate first coming to mind is surely the category of Hilbert spaces. But there is another choice naturally suggesting itself that we prefer to take, namely the category $\mathbf{2\text{-}V}^{\mathbf{ST}}$.

We mentioned already that $\mathbf{2\text{-}V}^{\mathbf{ST}}$ is something like a category of quantum presheaves on **ST**. Classical categories of presheaves are known to be topoi (see Ref. 12), and therefore generalizations of set theory to the case of intuitionistic logic. Besides the interpretation as a space of generalizations of topological quantum field theories, $\mathbf{2\text{-}V}^{\mathbf{ST}}$ may thus be seen as a generalization of set theory, possibly involving elements of quantum and intuitionistic logic. We will deal with the details of set theory in $\mathbf{2\text{-}V}^{\mathbf{ST}}$ in a publication to follow and will for now restrict to the question of defining categories in $\mathbf{2\text{-}V}^{\mathbf{ST}}$. This is the first step toward defining an analog of **Topi** (or **ST**) in $\mathbf{2\text{-}V}^{\mathbf{ST}}$ and placing the whole quantization scheme in the context of our new set theory.

That it is possible to define categories in $\mathbf{2\text{-}V}^{\mathbf{ST}}$ is shown by Lemma 2. In a category $\mathcal{C}$ with pullbacks we can define categories in $\mathcal{C}$ by using the usual definition of category that can easily be formulated in arrow language in **Set** (see Ref. 12 for the details).

To understand what a category in **2-V**$^{\mathbf{ST}}$ looks like, consider first the case of a category in **2-V** (which has pullbacks too). First of all, **2-V** is a category of categories. So, a category in **2-V** surely has the structure of a double category. (A double category is defined to be a category in **Cat**, the category of (small) categories. It is the generalization of the 2-category notion to the case of rectangular arrow diagrams, whereas the arrow diagrams of 2-categories are spherical in nature. See Ref. 15 for an introduction.) Besides this, a category in **2-V** inherits a module structure from the objects of **2-V** that represent its object class and morphism class, respectively. So, a category in **2-V** is a module double category.

Now, let us return to the case of a category in **2-V**$^{\mathbf{ST}}$. A category in **2-V**$^{\mathbf{ST}}$ looks like a category in **2-V** varying with the objects of **ST**, where the variation is subject to certain compatibility relations. Observe that *every* category in **2-V**$^{\mathbf{ST}}$ has some topological flavor, it even resembles a manifold in the sense that it is locally a generalization of a module structure.

Placing the whole scheme of the quantization of **Topi** in the category **2-V**$^{\mathbf{ST}}$ now means that we have to consider wave functions in **2-V**$^{\mathbf{ST}}$ on some special category in **2-V**$^{\mathbf{ST}}$. In concluding this section, we will roughly sketch what this should look like.

Observe that a category in **V** (the category of vector spaces and linear maps) is a module category. Now, we argued above that wave functions on a module category $\mathscr{C}$ should be functors from $\mathscr{C}$ to **H** or—if we consider the linear structure only, again—from $\mathscr{C}$ to **V**. It is therefore suggestive to consider as wave functions on a category in **2-V**$^{\mathbf{ST}}$ something like functors from that category to **2-V**$^{\mathbf{ST}}$ itself. We say ''something like'' because at first sight functors are only defined between two different categories in **2-V**$^{\mathbf{ST}}$ and **2-V**$^{\mathbf{ST}}$ need not be representable as a category object in itself. This problem can be circumvented the following way: For the category **Set**, functors from a category in **Set** (which is just an ordinary category) to **Set** can be understood as split discrete opfibrations in **Set** (for details and the equivalence to **Set**-valued functors, see Ref. 12), a notion that readily generalizes to general categories with pullbacks. A split discrete opfibration (sometimes also called split discrete cofibration) in **2-V**$^{\mathbf{ST}}$ is a functor in **2-V**$^{\mathbf{ST}}$ between categories (C_0,C_1) and (D_0,D_1) in **2-V**$^{\mathbf{ST}}$ (where C_0 and D_0 are representing the classes of objects and C_1 and D_1 the classes of morphisms) having the property that

$$\begin{array}{ccc} D_1 & \longrightarrow & D_0 \\ {\scriptstyle \varphi_1}\downarrow & & \downarrow{\scriptstyle \varphi_0} \\ C_1 & \xrightarrow[d^\circ]{} & C_0 \end{array}$$

is a pullback square. Here, φ_0 and φ_1 denote the object and the morphism map of the functors, respectively, and d° is the domain map (assigning the domain to each morphism). If we want to be more precise, we talk about a split discrete opfibration over the category $\mathscr{C}=(C_0,C_1)$ in **2-V**$^{\mathbf{ST}}$.

In conclusion, we suggest that the wave functions on a category $\mathscr{C}$ in **2-V**$^{\mathbf{ST}}$ are the split discrete opfibrations over $\mathscr{C}$. We will deal with the implications of this suggestion in future work on set theory in **2-V**$^{\mathbf{ST}}$.

V. CONCLUSION

Starting from the idea that quantization of the lattice $\tau(X)$ of topologies on a fixed set X can be understood as a generalization of second quantization (which we take as a tiny hint that the abstract approach followed in this paper could have some physically interesting content), we went on to consider the problem of quantizing the whole category of topological spaces and continuous injective maps (getting rid of the assumption of a fixed set X). We found that our approach naturally suggested a further generalization, involving quantization, even of the underlying set theory.

One of the most interesting observations is surely that the approach developed in Ref. 2, which shows the need to enlarge the classical configuration space when considering the quantum theory, fits in with the hierarchy of categorical structures (1- and 2-categories) as known from topological quantum field theory.

ACKNOWLEDGMENTS

I would like to thank Professor Dr. M. Reeken for helpful discussions and I. Ackermann for an excellent job of typing.

[1] C. J. Isham, ''Quantum topology and quantization on the lattice of topologies,'' Class. Quantum Grav. **6**, 1509–1534 (1989).
[2] J. A. Wheeler, ''Geometrodynamics and the issue of the final state,'' in C. and B. De Witt, *Relativity, Groups and Topology*, Les Houches Summer School, edited by C. De Witt and B. De Witt (Gordon and Breach, New York, 1964).
[3] C. J. Isham, ''Canonical groups and the quantization of geometry and topology,'' in *Conceptual Problems of Quantum Gravity*, edited by A. Ashtekar and J. Stachel (Birkhäuser, Boston, 1991).
[4] J. C. Baez and J. Dolan, ''Higher-dimensional algebra and topological quantum field theory,'' J. Math. Phys. **36**, 6073–6105 (1995).
[5] L. Crane, ''Topological field theory as the key to quantum gravity,'' *Knots and Quantum Gravity*, edited by J. Baez (Oxford University Press, Oxford, 1994), Vol. 121–132.
[6] L. Crane, ''Clock and category: Is quantum gravity algebraic?,'' J. Math. Phys. **36**, 6180–6193 (1995).
[7] M. M. Kapranov and V. A. Voevodsky, ''2-categories and Zamolodchikov tetrahedra equations,'' *Algebraic Groups and their Generalizations*, Proceedings of the Symposia in Pure Mathematics, edited by W. J. Haboush and B. J. Parshall (American Mathematical Society, Providence 1994), Vol. 56 Part 2, pp. 177–259.
[8] R. Penrose and W. Rindler, *Spinors and Space–Time* (Cambridge University Press, Cambridge, 1984), Vol. 1.
[9] J. Stachel, ''Einstein and quantum mechanics,'' in A. Ashtekar, J. Stachel, *Conceptual Problems of Quantum Gravity*, edited by A. Ashtekar and J. Stachel (Birkhäuser, Boston, 1991), pp. 13–42.
[10] G. H. Müller, *Sets and Classes* (North-Holland, Amsterdam, 1976).
[11] L. Smolin, ''Linking topological quantum field theory and nonperturbative quantum gravity,'' J. Math. Phys. **36**, 6417–6455 (1995).
[12] M. Barr and C. Wells, *Toposes, Triples and Theories* (Springer, New York, 1985).
[13] C. J. Isham, ''An introduction to quantum gravity,'' in *Quantum Gravity*, edited by C. J. Isham, R. Penrose, and D. W. Sciama (Oxford University Press, Oxford, 1975), pp. 1–77.
[14] C. J. Isham, ''Conceptual and geometrical problems in quantum gravity,'' ''*Recent Aspects of Quantum Fields*,'' Lecture Notes in Physics, edited by H. Mitter and H. Gausterer, (Springer, Berlin, 1991), Vol. 396, pp. 123–229.
[15] J. W. Gray, ''Formal category theory,'' *Lecture Notes in Mathematics* (Springer, Berlin, 1974), p. 391.

Similarity and intertwining of Dirac operators

Barton L. Willis
Department of Mathematics and Statistics, University of Nebraska at Kearney, Kearney, Nebraska 68849-1110

(Received 26 August 1997; accepted for publication 19 November 1997)

Similarity and symmetry of quantum mechanical operators are the basic theoretical means for systematizing quantum dynamics. The classical counterparts of these concepts are canonical transformations and conservation laws. For operators, a concept called intertwining encompasses both similarity and symmetry. In this paper, we construct pairs of Dirac operators with noncentral potentials that are intertwined by multiplication operators. © *1998 American Institute of Physics.* [S0022-2488(98)01303-6]

I. INTRODUCTION

Operators H_1 and H_2 are said to be *intertwined* by Q if[1]

$$QH_1 - H_2Q = 0. \tag{1}$$

The operator Q need not be one-to-one; however, if it is, then we say that H_1 and H_2 are *similar*. Additionally, if $H_1 = H_2$, then Q commutes with H_1 and we say that Q is a *symmetry* of H_1. Thus the concept of intertwining encompasses similarity and symmetry. An important application of intertwining follows from the fact that if H_1 and H_2 are intertwined by Q and $\Psi \in \text{null}(H_1)$, then $Q\Psi \in \text{null}(H_2)$. To apply this result, the null-space (or at least one element of the null-space) of H_1 must be known; consequently, the aim is to intertwine H_1 and H_2 with H_1 being ''simple.''

Anderson[2] examined the intertwining of Dirac operators for *scalar* potentials. In this article, we investigate this problem for nonscalar potentials. Specifically, we consider four-vector potentials of the form

$$\boldsymbol{A} = (U, \sin(\theta)D, -\cos(\theta)D, 0), \tag{2}$$

where U and D are functions of the cylindrical coordinates r and z only. The electric and magnetic fields for vector potentials of this form are axially symmetric and their theta components are zero. In general, fields of this type have nonzero electric and magnetic dipole moments. We refer to the Dirac operator with mass m and vector potential given by Eq. (2) as $H_{[U,D,m]}$.

We offer three examples of how potentials of the form (2) arise in various physical problems. The first is the potential of a Penny trap with an electric quadrapole field and a constant magnetic field.[3] The second example is any atom whose nucleus has a nonvanishing electric dipole or quadrapole moment. The last example is the ''self-field quantum electrodynamics'' of Barut *et al.*[4] In this theory, the Dirac and Maxwell equations are coupled with the Dirac current serving as the electric current. Assuming this system has axially symmetric solutions, one is led to studying Dirac equations whose potential has the form given in Eq. (2).

For potentials of the form given by Eq. (2), the corresponding Dirac operator commutes with the z component of the angular momentum, but it does not in general commute with any other component. Further, the vector potential is time independent, so any Dirac spinor in the null-space of $H_{[U,D,m]}$ can be assumed to have the form

 © 1998 American Institute of Physics

$$\Psi = e^{i\omega t} e^{-in\theta}/\sqrt{r} \begin{pmatrix} e^{-i\theta} F_0 \\ F_1 \\ i e^{-i\theta} F_2 \\ i F_3 \end{pmatrix}, \tag{3}$$

where n is an integer and $F_0, \ldots, F_3$ are functions of r and z only. Using this form for the wave function, we find that the spinor (F_0, F_1, F_2, F_3) is in the null space of the ''reduced'' Dirac operator $\hat{H}_{[U,D,m]}$, where

$$\hat{H}_{[U,D,m]} = \partial_r + M\partial_z + (U-\omega)C_1 + \left(D - \frac{2n+1}{2r}\right) C_2 + mC_3, \tag{4}$$

and where the matrices M, C_1, C_2, and C_3 are given in 2×2 block form as

$$M = \begin{pmatrix} -i\sigma_y & 0 \\ 0 & -i\sigma_y \end{pmatrix}, \quad C_1 = \begin{pmatrix} 0 & \sigma_x \\ -\sigma_x & 0 \end{pmatrix},$$

$$C_2 = \begin{pmatrix} -\sigma_z & 0 \\ 0 & -\sigma_z \end{pmatrix}, \quad C_3 = \begin{pmatrix} 0 & \sigma_x \\ \sigma_x & 0 \end{pmatrix}.$$

These matrices pairwise anticommute and they satisfy $C_2^2 = C_3^2 = I, M^2 = C_1^2 = -I$, where I is the 4×4 identity matrix. The abbreviations

$$v_1 = U - \omega, \quad v_2 = D - (2n+1)/2r, \tag{5}$$

allow us to express the reduced Dirac operator as

$$\partial_r + M\partial_z + V_{[v_1, v_2, m]}, \tag{6}$$

where $V_{[v_1, v_2, m]}$ is the matrix (again, in 2×2 block form)

$$V_{[v_1, v_2, m]} = \begin{pmatrix} -v_2\sigma_z & (m+v_1)\sigma_x \\ (m-v_1)\sigma_x & -v_2\sigma_z \end{pmatrix}. \tag{7}$$

In Sec. II, we find all Dirac operators that are intertwined by a multiplication operator, and in Sec. III we present two examples.

II. MULTIPLICATION OPERATORS

Let Q be a multiplication operator and let H_1 and H_2 be the reduced Dirac operators

$$H_1 = \partial_r + M\partial_z + V_{[v_1, v_2, m]}, \tag{8}$$

$$H_2 = \partial_r + M\partial_z + V_{[w_1, w_2, \mu]}. \tag{9}$$

For the remainder of this paper, we use the notation

$$V_{[v1, v2, m]} = V, \quad V_{[w_1, w_2, \mu]} = W.$$

By commuting the derivative terms of the operator QH_1 to the right and by equating coefficients of like derivative operators, it follows that $QH_1 = H_2 Q$ if

$$[Q, M] = 0, \tag{10}$$

$$Q_r + MQ_z = QV - WQ, \tag{11}$$

where the subscripts indicate partial derivatives. To analyze these equations, we initially determine all matrices that commute with M. We find that Q commutes with M if and only if Q has the form

$$Q=\begin{pmatrix} q_{1,1} & -q_{1,0} & q_{1,3} & -q_{1,2} \\ q_{1,0} & q_{1,1} & q_{1,2} & q_{1,3} \\ q_{3,1} & -q_{3,0} & q_{3,3} & -q_{3,2} \\ q_{3,0} & q_{3,1} & q_{3,2} & q_{3,3} \end{pmatrix}, \tag{12}$$

where the q_{ij}'s are eight arbitrary functions of r and z.

We can derive one more algebraic condition on Q by using Eqs. (10) and (11) to show that M commutes with $QV-WQ$. To prove this, we need to remember that M is a constant matrix and that Q commutes with M. Imposing this condition on Q implies that the eight independent elements of Q satisfy eight homogeneous linear equations. These linear equations have a nontrivial solution if and only if either $v_2=w_2$ or if

$$0=v_2^2-v_1^2+m^2-w_2^2+w_1^2-\mu^2. \tag{13}$$

It can be shown that the condition $v_2=w_2$ gives no results that are not found from using the solvability condition given in Eq. (13).

The algebraic condition imposed by Eq. (13) can always be locally satisfied by a choice of gauge. Specifically, we can choose a gauge for which $0=m^2-v_1^2+v_2^2$ (and similarly for the other potential). In four-vector language, this choice corresponds to a gauge that specifies the value of $\boldsymbol{A}^{\mu}\boldsymbol{A}_{\mu}$.

We now have eight homogeneous linear equations and the solvability condition [Eq. (13)]. In solving these equations, we assume that $w_1\neq\mu$; a detailed calculation shows that if $w_1=\mu$, leads to inconsistent equations. Assuming $w_1\neq\mu$, we find

$$q_{1,0}=\frac{q_{3,1}v_2-q_{3,2}v_1+q_{3,2}m-q_{3,1}w_2}{w_1-\mu}, \tag{14}$$

$$q_{1,1}=\frac{q_{3,0}v_2+q_{3,3}v_1-q_{3,3}m+q_{3,0}w_2}{w_1-\mu}, \tag{15}$$

$$q_{1,2}=\frac{q_{3,3}v_2+q_{3,0}v_1+q_{3,0}m-q_{3,3}w_2}{w_1-\mu}, \tag{16}$$

$$q_{1,3}=\frac{q_{3,2}v_2-q_{3,1}v_1-q_{3,1}m+q_{3,2}w_2}{w_1-\mu}, \tag{17}$$

where q_{30},q_{31},q_{32}, and q_{33} are arbitrary. Making these substitutions for $q_{1,i}, 0\leqslant i\leqslant 3$ in Q we discover that $QV-WQ=0$. Thus the algebraic conditions on Q imply that Q intertwines V and W. This fact[5] reduces Eq. (11) to

$$Q_r+MQ_z=0. \tag{18}$$

The 16 components of Eq. (18) form four pairs of decoupled equations and each pair is a Cauchy–Riemann equation couple. We conclude from this that there are analytic functions F, G, J, and K for which

$$\begin{aligned} F&=q_{3,1}+iq_{3,0}, \quad G=q_{3,3}+iq_{3,2}, \\ J&=q_{1,1}+iq_{1,0}, \quad K=q_{1,3}+iq_{1,2}, \end{aligned} \tag{19}$$

where the complex variable is $r+iz$, where $q_{1,0},q_{1,1},q_{1,2}$, and $q_{1,3}$ are given by Eqs. (14)–(17). Defining a scalar ψ and a 2×2 matrix T by

$$\psi = -\frac{iv_2}{w_1-\mu}, \tag{20}$$

$$T = \frac{1}{w_1-\mu}\begin{pmatrix} iv_2 & -(v_1+\mu) \\ v_1-\mu & iv_2 \end{pmatrix}, \tag{21}$$

and defining column vectors $\boldsymbol{K}=(K,J)$ and $\boldsymbol{G}=(G,F)$, we can eliminate the eight $q'_{ij}s$ from Eqs. (14)–(17) and (19) to derive

$$\boldsymbol{K} = \psi\boldsymbol{G} + T\boldsymbol{G}^\star, \tag{22}$$

where the star is the complex conjugate. By supplementing Eq. (22) with its complex conjugate, we obtain (in 2×2 block form)

$$\begin{pmatrix} \boldsymbol{K} \\ \boldsymbol{G} \end{pmatrix} = \begin{pmatrix} \psi T & (1+\psi^2)T \\ T & \psi T \end{pmatrix}\begin{pmatrix} \boldsymbol{K} \\ \boldsymbol{G} \end{pmatrix}^\star. \tag{23}$$

Since $F,G,J,$ and K are analytic, Eq. (23) imposes onerous conditions on the analytic functions $F,G,J,$ and K. We offer two techniques to analyze these equations.

Our first technique is to find differential equations that the potentials must satisfy to guarantee that the analytic functions $F,G,J,$ and K exist. We start by representing Eq. (22) as $\boldsymbol{H}=\Phi\boldsymbol{H}^\star$, where Φ is a 4×4 matrix and $\boldsymbol{H}$ is a column vector of analytic functions. If $\boldsymbol{H}$ is analytic, then we can use the Cauchy–Riemann equations to derive a differential equation for $\boldsymbol{H}$

$$\partial_s\boldsymbol{H} + \Phi(\partial_{s^\star}\Phi)^\star\boldsymbol{H} = 0, \tag{24}$$

where the complex variable is $s=r+iz$ and $\partial_s^* = (\partial_r + i\partial_z)/2$. On the other hand, if the matrix $\Phi(\partial_{s^\star}\Phi)^\star$ is analytic, then Eq. (24) has a nontrivial analytic solution and further, $\boldsymbol{H}=\Phi\boldsymbol{H}^\star$. Thus our differential condition on the potentials is that $\Phi(\partial_{s^\star}\Phi)^\star$ must be analytic. This condition also can be stated as

$$0 = \partial_{s^*}(\Phi(\partial_{s^\star}\Phi)^\star).$$

We can reduce Eq. (24) to an algebraic problem by factoring the matrix Φ as $S^{-1}S^\star$, where S is analytic. Expressed in terms of S, the general solution to Eq. (24) is $\boldsymbol{H}=S^{-1}\boldsymbol{H}_o$, where $\boldsymbol{H}_o$ is an arbitrary constant vector. Thus the existence of a matrix S is an alternative condition for the existence of analytic functions $F,G,J,$ and K.

A second technique is to view Eqs. (22) as linear equations for the potentials. We have two equations for the four potentials (v_1,v_2,w_1,w_2); however, the potentials are real, so again we supplement Eq. (22) with its complex conjugate to obtain four *linear* equations for the potentials. The linear equations

$$\begin{pmatrix} -G^\star & -iF^\star & J & iF \\ -G & iF & J^\star & -iF^\star \\ F^\star & -iG^\star & K & iG \\ F & iG & K^\star & -iG^\star \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \mu J - mG^\star \\ \mu J^\star - mG \\ \mu K - mF^\star \\ \mu K^\star - mF \end{pmatrix} \tag{25}$$

must be solved subject to the solvability condition

$$m^2 - v_1^2 + v_2^2 = \mu^2 - w_1^2 + w_2^2. \tag{26}$$

If no restrictions are placed on the functions $F,G,J,$ and K (other than analyticity), then the only solution to these equations is the unacceptable $w_1=\mu$. To find other solutions, we demand that the determinant of the coefficient matrix in Eq. (25) be zero. This implies

$$F^2 + G^2 = (F^2 + G^2)^\star$$

or

$$FK-GJ=(FK-GJ)^{\star}.$$

Assuming $G\neq 0$, the second condition implies there is a *real* constant α and an analytic function P for which

$$F=PG+\alpha/K,\quad J=PK+\alpha/G.$$

Making these substitutions for F and J, we can find solutions to Eqs. (13) and (25). For example, if $\alpha=0$, $m=0$, and $\mu=1$, then we find

$$v_1=4\frac{\mathrm{Re}(GK^{\star})\ \mathrm{Im}(P)}{\mathrm{Im}((K^2+G^2)(1+P^2))}, \tag{27}$$

$$v_2=2\frac{\mathrm{Re}(GK^{\star})(|P|^2+1)}{\mathrm{Im}((K^2+G^2)(1+P^2))}, \tag{28}$$

$$w_1=\frac{\mathrm{Im}((K^2-G^2)(1+P^2))}{\mathrm{Im}((K^2+G^2)(1+P^2))}, \tag{29}$$

$$w_2=2\frac{\mathrm{Re}(GK(1+P^2))}{\mathrm{Im}((K^2+G^2)(1+P^2))}, \tag{30}$$

where Re and Im are the real and imaginary parts of a complex number, respectively. Setting $P=1$, we find solutions with $v_1=0$

$$v_1=0,\quad v_2=2\frac{\mathrm{Re}(GK^{\star})}{\mathrm{Im}(K^2+G^2)}, \tag{31}$$

$$w_1=\frac{\mathrm{Im}(K^2-G^2)}{\mathrm{Im}(K^2+G^2)},\quad w_2=2\frac{\mathrm{Re}(GK)}{\mathrm{Im}(K^2+G^2)}. \tag{32}$$

(Recall that it is desirable for v_1 and v_2 to be ''simple.'')

We now present two examples of these results.

III. EXAMPLES

For our first example we set

$$m=1,\quad \mu=1,\quad P=1,\quad G=iz+r,\quad K=i.$$

The potentials are given by

$$v_1=0,\quad v_2=(-z^2-2z+r^2-1)/2rz,$$

$$w_1=-(z+1)/z,\quad w_1=(z^2+2z+r^2+1)/2rz.$$

The intertwining matrix is

$$Q=\begin{pmatrix} 0 & -1 & 0 & -1 \\ 1 & 0 & 1 & 0 \\ r & -z & r & -z \\ z & r & z & r \end{pmatrix}, \tag{33}$$

and the matrix V is

$$V=\begin{pmatrix} -v_2 & 0 & 0 & 1+v_1 \\ 0 & v_2 & 1+v_1 & 0 \\ 0 & 1-v_1 & -v_2 & 0 \\ 1-v_1 & 0 & 0 & v_2 \end{pmatrix} \tag{34}$$

and similarly for W. For this example, the matrix Q is not one-to-one, so the Dirac operators are intertwined, but we cannot conclude that they are similar.

For the second example, we set

$$\alpha=-1, \quad \mu=0, \quad m=1, \quad P=0, \quad K=iz+r, \quad G=iz+r.$$

The potentials are

$$v_1=0, \quad w_1=0,$$

$$v_2=-(z^2+r^2)/2rz, \quad w_1=(z-r)(z+r)/2rz \tag{35}$$

and the intertwining matrix is

$$Q=\begin{pmatrix} r/\rho^2 & z/\rho^2 & r & -z \\ -z/\rho^2 & r/\rho^2 & z & r \\ -r/\rho^2 & -z/\rho^2 & r & -z \\ z/\rho^2 & -r/\rho^2 & z & r \end{pmatrix}, \tag{36}$$

where $\rho^2=z^2+r^2$. For this example, the intertwining matrix is one-to-one (its determinant is four) and consequently, the Dirac operators are similar.

IV. CONCLUSION

For a class of noncentral potentials, we have shown how to construct pairs of intertwined Dirac operators. We also determined a differential equation that the potentials of intertwined Dirac operators must satisfy.

In this paper, we considered intertwining by multiplication operators; however, it can be any combination of differential, integral, multiplication, or composition operators. Unless the intertwining operator is relatively simple, then it will not enable one to generate exactly solvable Dirac equations.

We finish by noting that even the concept of intertwining can be generalized. If there are operators Q and D for which $QH_1=H_2D$, then we still get a useful null-space result for H_1 and H_2. This avenue was not explored in the present study, because we can show that for multiplication operators we must have $D=Q$. For more complex operators, we may expect results with $Q \neq D$.

ACKNOWLEDGMENTS

This work was supported in part by a University of Nebraska at Kearney Research Services Council Summer Grant. This paper contains results obtained using the computer algebra system Macsyma.[6] I am grateful to the staff at Macsyma, Inc. for technical assistance.

[1] Normally, the domain of $QH_1 - H_2Q$ is not the whole space, so we should write the intertwining condition as $QH_1 - H_2Q \subset 0$. We do not use this notation, so all equalities in this paper should be interpreted as ''equal on a nonempty set.''
[2] A. Anderson, Phys. Rev. D **37**, 536 (1989); A. Anderson and R. Camporesi, Commun. Math. Phys. **130**, 61 (1990); A. Anderson, Phys. Rev. A **43**, 4602 (1991).
[3] W. Paul, Rev. Mod. Phys. **62**, 531 (1990).
[4] A. O. Barut, Phys. Rev. A **45**, 7740 (1992).
[5] Matrices A and B are intertwined if and only if they have at least one eigenvalue in common. Equation (13) can be shown to impose this condition on V and W.
[6] Macsyma is a registered trademark of Macsyma, Inc., 20 Academy Street, Arlington, MA 02174.

Radiation principle for inhomogeneous media

D. Eidus
School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

(Received 12 May 1997; accepted for publication 10 October 1997)

The classical Sommerfeld radiation principle, well known for stationary oscillations in homogeneous media, is extended to a more general case where a smooth density $\rho=\rho(x)$, $x\in\mathbb{R}^N$, depends both on $r=|x|$ and on angular coordinates of the point x. We assume that the density behaves like a power r^m, $m>-2$, at infinity, and, moreover, that there exists a one-parameter family of closed asymptotic wave fronts. A Sommerfeld–Rellich-type radiation condition is obtained and a corresponding theorem of unique solvability of the stationary oscillation problem is proved under some conditions imposed on the asymptotic phase. Some examples are given which include the case of elliptic wave fronts. © *1998 American Institute of Physics.* [S0022-2488(98)00602-1]

I. INTRODUCTION

We are concerned with the extension of the classical Sommerfeld radiation principle to a general inhomogeneous medium. We assume that the free oscillation equation in the medium has the form

$$\rho(x)w_{tt}-\Delta w=0, \tag{1.1}$$

where $w=w(x,t)$, $t>0$, $x=(x^{(1)},\ldots,x^{(N)})$, is a point in the Euclidean space $\mathbb{R}^N$, $N\geqslant 2$, and $\rho=\rho(x)$ is a smooth positive function (density) defined on $\mathbb{R}^N$. The separation of the time variable t leads to the equation of stationary oscillations

$$\Delta u+k^2\rho(x)u=0, \tag{1.2}$$

where $k>0$ is the frequency. The corresponding equation of forced oscillations has the form

$$\Delta u+k^2\rho(x)u=f(x). \tag{1.3}$$

First let us assume for simplicity that $f(x)$ vanishes outside a sphere $|x|=r_1$ (i.e., it has a compact support). The case where

$$\rho(x)=\rho_0>0, \quad \rho_0=\text{const}, \tag{1.4}$$

outside a sphere was studied by Sommerfeld. It is known that for any $k>0$ there exists one and only one solution of (1.3) satisfying the following conditions at infinity:

$$u=O(r^{-(N-1)/2}), \tag{1.5}$$

$$u_r-ik\sqrt{\rho_0}u=o(r^{-(N-1)/2}) \tag{1.6}$$

as $r\to\infty$, where $r=|x|$, and u_r is a derivative with respect to r. This statement is called the radiation principle (RP) and the conditions (1.5) and (1.6) are the Sommerfeld radiation conditions (RC). Solutions of (1.3) satisfying RC have the following asymptotics at infinity:

$$u\sim\exp(ik\sqrt{\rho_0}r)g(r,\hat{x}), \quad r\to\infty, \tag{1.7}$$

where $\hat{x}=x/r$ and $g(r,\hat{x})$ behaves as $p(\hat{x})r^{-(N-1)/2}$ at infinity. These solutions are called outgoing because the functions

 © 1998 American Institute of Physics

$$w=e^{-ikt}u$$

are off-going waves described by the equation

$$\rho w_{tt}-\Delta w+e^{-ikt}f=0.$$

Vekua[1] has proved that it suffices to impose the condition (1.6) without (1.5), in order to guarantee the uniqueness of the solution u. Rellich[2] has shown that RC can be written in the following integral form:

$$\int |u_r-ik\sqrt{\rho_0}u|^2\,dx<\infty, \tag{1.8}$$

where we integrate over the whole $\mathbb{R}^N$. The uniqueness under RC (1.8) implies that the point $\lambda=k^2>0$ is not an eigenvalue of the equation (1.2) in H, where H is the space L^2 with the weight function ρ. So the existence of eigenvalues is not compatible with the RP in the above-defined form.

Much literature is devoted to the RP for general high-order equations with coefficients stabilizing fast enough at infinity. The strongest results were obtained by Vainberg (see Ref. 3 and references given there). There are also papers dealing with the RP for different boundary value problems in domains with noncompact boundaries (see Refs. 4 and 5 and the references in Ref. 3). We have to notice that RC can depend both on the equation and on the domain where the boundary value problem is treated.

There are also some results for the equation (1.3) where (1.4) is not valid. For the case of two homogeneous media with an infinite interface between them we can refer to Refs. 6–9. Many papers deal with the special case of stratified media (see Refs. 10–13). Some of these results were generalized recently by Jäger and Saito.[14,15]

Here we extend the RP to the case where the density $\rho(x)$ has the following asymptotics at infinity:

$$\rho(x)\sim\rho_0(\hat{x})r^{-2\alpha},\quad r\to\infty, \tag{1.9}$$

where $\alpha<1$ and $\rho_0(\hat{x})>0$ on the sphere

$$S^{N-1}:=\{x:x\in\mathbb{R}^N,|x|=1\}.$$

The restriction $\alpha<1$ is natural, because for $\alpha>1$ the spectrum of (1.2) in H (defined above) is discrete[16]—which is in contradiction to the RP.

Even for an exact equality (1.9) the variables in (1.2) could be separated only in the following two special cases: $\alpha=1$, which we have eliminated, and the case where

$$\rho_0(\hat{x})=\text{const.} \tag{1.10}$$

In the latter situation we have the exact solution of (1.2) depending on r only:

$$v(r)=r^{-(N-2)/2}H_p^{(1)}(c_0r^{1-\alpha}), \tag{1.11}$$

where c_0 and p are some positive constants. This function has the following asymptotics as $r\to\infty$:

$$v\sim c_1r^{-(N-1-\alpha)/2}\exp(ic_0r^{1-\alpha}) \tag{1.12}$$

[cf. with (1.7)]. This means that $v(r)$ describes an off-going wave and that the following RC is valid:

$$v_r-ik\sqrt{\rho}v=O(r^{-(N+1-\alpha)/2}),\quad r\to\infty. \tag{1.13}$$

Due to $\alpha<1$ the solution $v(r)$ satisfies RC (1.8).

Our first objective is to find a proper RC for densities ρ satisfying (1.9). To this end we apply the following heuristics. By the Dubye procedure[3] one can construct a formal asymptotic solution of (1.2) as $r\to\infty$. The dominant term of the corresponding asymptotic expansion has the following form:

$$\chi=g(r,\hat{x})\ \exp\ (ik\phi(x)), \tag{1.14}$$

where g behaves like $r^{-(N-1-\alpha)/2}$ as $r\to\infty$ and ϕ is a solution of the eikonal equation

$$|\nabla\phi|^2=\rho. \tag{1.15}$$

It is clear that (1.11) is a special case of (1.14). Set

$$\nu=|\nabla\phi|^{-1}\nabla\phi, \quad \chi_\nu=\partial\nu\partial\chi.$$

The function χ satisfies the RC

$$\int\ |\chi_\nu-ik\sqrt{\rho}\chi|^2\ dx<\infty, \tag{1.16}$$

which will be the basis of the RP for ρ satisfying (1.9). We leave out all standard calculations leading to (1.14) because this procedure is purely heuristic. The final results will be formulated in Sec. II and proved in Secs. III–V without using the Dubye method. Their concise description follows.

We assume the existence of an asymptotic solution ϕ (asymptotic phase) of the eikonal equation (1.15), which has a family of closed level surfaces (asymptotic wave fronts) in a neighborhood of infinity, and is increasing at infinity as $r^{1-\alpha}$, where α is the same number as in (1.9). Under these (and some additional) conditions imposed on ϕ, ρ, and f we prove that there exists one and only one solution (1.3) satisfying the RC

$$\int\ (1+r)^{\gamma-1}|u_\nu-ik\sqrt{\rho}u|^2\ dx<\infty, \tag{1.17}$$

where $\gamma=\text{const}$ and $\gamma\in(\alpha,1)$.

The imposed condition, that wave fronts are closed, is essential. For instance in the case where $\rho=\text{const}$, there exist nontrivial plane waves for which

$$u_\nu-ik\sqrt{\rho}u=0$$

in $\mathbb{R}^N$, where ν here is a normal to plane wave fronts.

We can see from the asymptotics (1.14) that the condition $\phi\to\infty$ as $r\to\infty$ in physical terms means that the wave is off-going.

Some of our results are similar to those obtained by Vogelsang in his paper[17] for a general self-adjoint elliptic operator of second order with variable coefficients [instead of Δ in (1.1)–(1.3)] and $\rho=\text{const}$. It might be interesting to combine them.

II. MAIN RESULTS AND EXAMPLES

Let us introduce some notation. For $r_2>r_1>0$ set

$$S(r_1)=\{x:x\in\mathbb{R}^N,|x|=r_1\},$$

$$\Omega(r_1)=\{x:x\in\mathbb{R}^N,|x|<r_1\},$$

$$\Omega_{r_1}=\{x:x\in\mathbb{R}^N,|x|>r_1\},$$

$$\Omega(r_1,r_2)=\{x:x\in\mathbb{R}^N,r_1<|x|<r_2\}.$$

By c and c_j, $j=0,1,...$, we denote positive constants. For $\mathbb{R}_+=\{\tau:\tau\in\mathbb{R},\tau>0\}$ and functions $a:\mathbb{R}^N\to\mathbb{R}$ and $b:\mathbb{R}^N\to\mathbb{R}_+$ we write $a=O(b)$, $r\to\infty$, if $|a(x)|\leqslant c_0 b(x)$ on some Ω_{r_1}, and $a\sim b$, $r\to\infty$, if $ab^{-1}\to 1$ as $r\to\infty$. By $(\xi^{(1)},\dots,\xi^{(N)})$ we denote coordinates of a vector $\xi\in\mathbb{C}^N$ in some orthogonal basis in $\mathbb{C}^N$, by (ξ,η) the inner product of vectors ξ,η in $\mathbb{C}^N$, by $\widehat{(\xi,\eta)}$ the angle between ξ, η, and by $|\xi|$ the norm of ξ in $\mathbb{C}^N$. Throughout the paper we use the tensor rule of summation over repeating indices.

By $\int g\,dx$ we denote the integration over $\mathbb{R}^N$. For derivatives we shall use the following notation:

$$u_j=\partial x_j\partial u,\quad \nabla u=(u_1,\dots,u_N),\quad u_r=\partial r\partial u,\quad r=|x|,$$

$$u_{jl}=\partial x_j\partial x_l\partial^2 u,\quad D^n u=\partial x_1^{n_1}\cdots\partial x_N^{n_N}\partial^{|n|},$$

where $n=(n_1,\dots,n_N)$ and $|n|=n_1+\cdots+n_N$. Denote by $L^2(\omega)$ and L^2_{loc} the sets of complex-valued functions square integrable over $\omega\subset\mathbb{R}^N$ and over every $\Omega(R)$, $R>0$, respectively; by $H^m(\omega)$; H^m_{loc}, $m=0,1,...$, we denote the sets of functions from $L^2(\omega)$, L^2_{loc} which have all distributional derivatives up to the order m belonging to $L^2(\omega)$, L^2_{loc}, respectively. Besides, $H^0=L^2$. Introduce the norm in $H^m(\omega)$:

$$\|g\|^2_{m,\omega}=\int\sum_{|n|=0,\dots,m}|D^n g|^2\,dx,\quad \|g\|_{0,\omega}=\|g\|_{L^2(\omega)}.$$

We say that the sequence $g_n\in H^m_{\mathrm{loc}}$ converges to $g\in H^m_{\mathrm{loc}}$ in H^m_{loc} if $g_n\to g$ in $H^m(\Omega(R))$ for any $R>0$. Let ρ be a given function $\rho:\mathbb{R}^N\to\mathbb{R}_+$. Let H be the Hilbert space of functions $u:\mathbb{R}^N\to\mathbb{C}$ with the norm

$$\|g\|^2=\int\rho|u|^2\,dx.$$

Set

$$H^1_\rho=\{g:g\in H^1_{\mathrm{loc}}\cap H,|\nabla g|\in L^2(\mathbb{R}^N)\},\quad l=-\rho^{-1}\Delta. \tag{2.1}$$

Let C_0^∞ be the set of functions from $C^\infty(\mathbb{R}^N)$ having compact supports, $\mathscr{L}$ be the Friedrichs extension in H of the operator l, initially defined on C_0^∞, and $\mathscr{D}(\mathscr{L})$ be the domain of $\mathscr{L}$. Then $\mathscr{L}$ is a self-adjoint operator in H,

$$\mathscr{D}(\mathscr{L})\subset H^1_\rho. \tag{2.2}$$

We shall assume that on $\mathbb{R}^N$

$$\rho(x)\geqslant c(1+r)^{-2\alpha},\quad \alpha<1. \tag{2.3}$$

Then[18]

$$\mathscr{D}(\mathscr{L})=\{u:u\in H^2_{\mathrm{loc}}\cap H,\ lu\in H\}. \tag{2.4}$$

The equation (1.3) can be written in the form

$$lu-\lambda u=F, \tag{2.5}$$

where $\lambda=k^2>0$, and

$$F=-\rho^{-1}f. \tag{2.6}$$

Definition 2.1: A function $u:\mathbb{R}^N\to\mathbb{C}$ is a solution of the equation (2.5) [or (1.3)] if $u\in H^2_{\mathrm{loc}}$ and (2.5) is valid a.e. on $\mathbb{R}^N$.

If $\lambda \in \mathbb{C}$ is a regular point of the self-adjoint operator $\mathscr{S}$, then for any $F \in H$ there exists one and only one solution of (2.5) belonging to $\mathscr{D}(\mathscr{S})$. So the equation (2.5) has a unique solution in $\mathscr{D}(\mathscr{S})$ for any $F \in H$ if $\lambda < 0$ or $\lambda = k^2 + \varepsilon i$, $\varepsilon \neq 0$.

Definition 2.2: A function $\phi: \Omega_{r_0} \to \mathbb{R}$, $\phi \in C^1$, belongs to the class U_τ, $\tau > 0$, if on some domain Ω_{r_1}, $r_1 \geqslant r_0$, we have $|\nabla \phi| > 0$,

$$0 < c_1 < r^{-\tau} \phi < c_2, \tag{2.7}$$

and there exists such a constant $c_0 > 0$ that for any $c > c_0$ the set

$$\Gamma(c) := \{x : x \in \mathbb{R}^N, \phi(x) = c\} \tag{2.8}$$

is a $(N-1)$-dimensional closed surface.

Note 2.1: Suppose $\phi \in U_\tau$, $\tau > 0$. Then for any $c > c_0$ the set

$$\Gamma_c := \{x : x \in \mathbb{R}^N, \phi(x) > c\} \tag{2.9}$$

is the exterior of the closed surface $\Gamma(c)$ and $\Gamma(c) \to \infty$ as $c \to +\infty$. Moreover, the vector

$$\nu = \nabla \phi / |\nabla \phi| \tag{2.10}$$

is determined at any point $x \in \Omega_{r_1}$ and coincides with the external normal to the corresponding surface $\Gamma(c)$ at the same point x.

Definition 2.3: Let $\rho: \mathbb{R}^N \to \mathbb{R}_+$. A function $\phi: \Omega_{r_0} \to \mathbb{R}$, $r_0 > 0$, $\phi \in C^1(\Omega_{r_0})$, is said to be an asymptotic phase for ρ if

$$|\nabla \phi|^2 \sim \rho, \quad r \to \infty,$$

and a δ-asymptotic phase, $\delta > 0$, if

$$\rho^{-1} |\nabla \phi|^2 = 1 + O(r^{-\delta}), \quad r \to \infty. \tag{2.11}$$

The corresponding surfaces (2.8) are said to be asymptotic (δ-asymptotic) wave fronts (cf. Refs. 19 and 20).

Note 2.2: If ϕ is an asymptotic phase for ρ, and on Ω_{r_0}, $r_0 > 0$,

$$0 < c_1 < r^{2\alpha} |\nabla \phi|^2 < c_2,$$

then on some Ω_{r_1}, $r_1 \geqslant r_0$,

$$0 < c_3 < r^{2\alpha} \rho < c_4.$$

Theorem 2.1: *Suppose* $\rho: \mathbb{R}^N \to \mathbb{R}_+$, $\rho \in C^1(\mathbb{R}^N)$ *and there exist such constants* $r_0 > 0$ *and* $\alpha < 1$ *that*

$$\rho r^{2\alpha} \geqslant c_0 > 0, \quad x \in \Omega_{r_0}, \tag{2.12}$$

$$Z := \lim_{R \to \infty} \inf_{r > R} (r(\log(r^{2\alpha} \rho))_r) > \alpha - 1. \tag{2.13}$$

Let $\phi \in C^1(\Omega_{r_1})$, $r_1 \geqslant r_0$, *be a corresponding asymptotic phase belonging to* U_τ, $\tau > 0$. *Then any solution* $u = u(x)$ *of (1.2) on* $\mathbb{R}^N$ *satisfying the RC*

$$\int_{\Omega_{R_1}} r^{\gamma - 1} |u_\nu - ik\sqrt{\rho} u|^2 \, dx < \infty, \tag{2.14}$$

where $\gamma = const$,

$$\gamma > \max\{\alpha, \alpha - Z\}, \tag{2.15}$$

and ν is defined by (2.10), vanishes a.e. on $\mathbb{R}^N$.

Corollary 2.1: Let u_1 and u_2 be solutions of (1.2) which satisfy the RC (2.14) with $\gamma=\gamma_1$ and $\gamma=\gamma_2$, respectively. Then $u_1=u_2$ a.e. on $\mathbb{R}^N$.

Note 2.3: The condition (2.13) actually coincides with the virial condition introduced by Agmon[21] (see also Ref. 17) for the Schrödinger equation.

Theorem 2.2: *Let $\rho=\rho(x)$ be a function satisfying the same condition as in Theorem 2.1 and, in addition, on Ω_{r_0},*

$$|\nabla\rho|\leqslant cr^{-2\alpha-1}. \tag{2.16}$$

Suppose that there exists a corresponding δ-asymptotic phase $\phi\in C^3(\Omega_{r_0})$, such that $\phi\in U_\tau$, $\tau=1-\alpha$, and

$$\delta>\max\{1-\alpha,1-\alpha-\tfrac{1}{2}Z\}, \tag{2.17}$$

$$|D^j\phi|\leqslant c_j r^{1-\alpha-|j|},\quad x\in\Omega_{r_0},\quad |j|=1,2,3. \tag{2.18}$$

Assume that there exists such a number γ that

$$\max\{\alpha,\alpha-Z\}<\gamma\leqslant\min\{1,2\delta+3\alpha-2\} \tag{2.19}$$

and for all pairs of vectors $(\xi,x)\in R^N\times\Omega_{r_0}$ the following inequality is valid:

$$G(x,\xi,\gamma)\geqslant c_0 r^{\gamma-1}|\xi|^2, \tag{2.20}$$

where $c_0>0$ is independent of x and *ξ,*

$$G(x,\xi,\gamma)=\phi^m(\phi_{lj}\xi^{(\ell)}\xi^{(j)}+m\phi^{-1}|\nabla\phi|^2((\xi,\nu)^2-\tfrac{1}{2}|\xi|^2)), \tag{2.21}$$

$$m=(\alpha+\gamma)/(1-\alpha). \tag{2.22}$$

Then for any $f\in V$, where

$$V=\left\{v: v\in L^2_{\mathrm{loc}}, \int (1+r)^{1+\gamma}|v|^2\,dx<\infty\right\}, \tag{2.23}$$

there exists one and only one solution of the problem (1.3) and (2.14).

Note 2.4: It follows from the conditions of Theorem 2.2 that in some Ω_{r_1}, $r_1>0$, we have

$$0<c_1\leqslant\rho r^{2\alpha}\leqslant c_2.$$

It is not easy, in general, to verify the condition (2.20). First let us deal with the simplest case where $\phi=\phi(r)$. Then

$$G(x,\xi,\gamma)=\phi^m(M_1|\xi|^2+M_2(|\xi|^2-(\xi,\hat{x})^2)),$$

where

$$M_1=\frac{m}{2}\,\phi^{-1}(\phi_r)^2+\phi_{rr},$$

$$M_2=r^{-1}\phi_r-\phi_{rr}-m\phi^{-1}(\phi_r)^2.$$

Thus the conditions $M_1>0$ and $M_2\geqslant 0$ are sufficient for (2.20).

Example 2.1: Let on Ω_{r_0}, $r_0>0$,

$$\phi(r)=(1-\alpha)^{-1}r^{1-\alpha},\quad \rho(x)=r^{-2\alpha}+\varphi(x),$$

where $\alpha<1$, and on Ω_{r_0}, for $|j|=0,1$,

$$|D^j\varphi|\leqslant cr^{-2\alpha-|j|-\delta} \tag{2.24}$$

and $\delta>1-\alpha$. Then

$$M_1=\tfrac{1}{2}(\gamma-\alpha)r^{-1-\alpha},\quad M_2=(1-\gamma)r^{-1-\alpha}.$$

By (2.24) ϕ is a δ-asymptotic phase for ρ and $Z=0$. Moreover (2.12), (2.13), and (2.16) hold, as does (2.20) for $\gamma\in(\alpha,1)$. So the assertion of Theorem 2.2 is valid for any $\gamma\in(\alpha,1)$.

Let us proceed to an example where ρ is weakly oscillating.

Example 2.2: Let on Ω_{r_0}, $r_0>0$,

$$\phi=(1-\alpha)^{-1}r^{1-\alpha}(A+\sin\log r),\quad \alpha<1,\quad A>0,$$

$$\rho=r^{-2\alpha}\left(A+\sin\log r+\frac{\cos\log r}{1-\alpha}\right)^2+\varphi(x),$$

where (2.24) holds with $\delta>\frac{3}{2}(1-\alpha)$. It is easy to see that ϕ is a δ-asymptotic phase for ρ, and for $A\to+\infty$

$$|\nabla\phi|^2=A^2r^{-2\alpha}(1+O(A^{-1})),$$

$$M_1=\tfrac{1}{2}A(\gamma-\alpha)r^{-1-\alpha}(1+O(A^{-1})),$$

$$M_2=A(1-\gamma)r^{-1-\alpha}(1+O(A^{-1})),$$

$$Z=O(A^{-1}).$$

Choose now some $\varepsilon\in(0,1-\alpha)$. Then for all sufficiently large A the inequality (2.12) is valid on some Ω_{r_1}, $r_1\geqslant r_0$, and

$$0<-Z<\varepsilon$$

[so (2.13) is also valid]. Now we can choose some $\gamma\in(\alpha+\varepsilon,1)\subset(\alpha-Z,1)$. Due to the choice of δ our γ satisfies (2.19) for all $A>A_0$, $A_0=A_0(\alpha)$. Now we can find such a number $A_1\geqslant A_0$, $A_1=A_1(\gamma,\alpha)$, that for $A>A_1$ the inequality (2.20) is valid. So the statement of Theorem 2.2 holds for all sufficiently large A.

Now we shall go on to another special case where the function ϕ is homogeneous of degree $1-\alpha$, $\alpha<1$. Here we obtain some geometric criteria for (2.20) imposed on wave fronts, which are quite verifiable. So let us assume that an asymptotic phase ϕ can be written in the form

$$\phi(x)=r^{1-\alpha}\hat{\phi}(\hat{x}),\quad \alpha<1, \tag{2.25}$$

where as above $r=|x|$, $\hat{x}=r^{-1}x$, and $\hat{\phi}:S^{N-1}\to\mathbb{R}_+$, and that $\phi\in C^3(\Omega_{r_0})$. By (2.25) the function ϕ is extended to the whole $\mathbb{R}^N$. The equation of each surface $\Gamma(c)$, $c>0$, can be written as follows:

$$r=c_1J(\hat{x}), \tag{2.26}$$

where $J>0$ on S^{N-1} and $c_1=c^{1/1-\alpha}$. The Euler identity for homogeneous functions

$$(x,\nabla\phi)=(1-\alpha)\phi \tag{2.27}$$

implies that for $x\neq0$,

$$|\nabla\phi|>0, \tag{2.28}$$

$$s:=\cos\psi>0, \tag{2.29}$$

where $\psi=(\widehat{x,\nu})$. By (2.25), (2.26), and (2.28) we can conclude that

$$\phi\in U_\tau,\quad \tau=1-\alpha. \tag{2.30}$$

All the functions $D^j\phi$, $|j|=1,2,3$, are homogeneous of degree $1-\alpha-|j|$ and could be written in the form analogous to (2.25). Hence the vector ν and the angle ψ depend only on $\hat{x}$ and do not depend on r. From (2.28) we have

$$0<c_1\leqslant r^{2\alpha}|\nabla\phi|^2\leqslant c_2,\quad x\in\Omega_{r_0},\quad r_0>0. \tag{2.31}$$

Set

$$\begin{gathered} s_0=\min_{\hat{x}\in S^{N-1}} s(\hat{x}),\\ \sigma=2s^2-1,\quad \sigma_0=2s_0^2-1. \end{gathered} \tag{2.32}$$

From (2.29)

$$s_0>0. \tag{2.33}$$

Definition 2.4: A smooth closed $(N-1)$-dimensional surface is said to be strongly convex if its second Gaussian quadratic form (with respect to the external normal) is negative definite at any point of the surface.

If the representation (2.25) for ϕ is valid, then all wave fronts $\Gamma(c)$, $c>0$, are homothetic. Therefore if one $\Gamma(c_1)$, $c_1>0$, is strongly convex, then all others $\Gamma(c)$, $c>0$, are too. Suppose it is the case and (g) is an $(N-1)$-dimensional tangential plane for a surface $\Gamma(c)$ at a point $x\in\Gamma(c)$. The Meunier formula for $\Gamma(c)$ can be written in the form

$$\phi_{ij}(x)g^{(i)}g^{(j)}=|\nabla\phi(x)|K(x,g), \tag{2.34}$$

where $g=(g^{(1)},\ldots,g^{(N)})\in(g),|g|=1$, and $K(x,g)$ is a positive curvature of the curve, which is the normal section of $\Gamma(c)$ at the point x by a two-dimensional plane of vectors ν, g. Due to (2.34)

$$K(x,g)=r^{-1}K_1(\hat{x},g), \tag{2.35}$$

where $K_1(\hat{x},g)$ is independent of r and will be called a normalized curvature for the direction $\hat{x}$ and the normal section defined by $g\in(g)$. Set

$$\beta=(\widehat{g,x}).$$

Then

$$\sin\beta\in[s,\,1]. \tag{2.36}$$

For $\sigma_0>0$ let us introduce

$$M=\min_{(\hat{x},g)\in Y}\frac{s(\hat{x})\sigma(\hat{x})K_1(\hat{x},g)}{\sigma(\hat{x})+\cos^2\beta}, \tag{2.37}$$

where $\beta=\beta(\hat{x},g)$ and Y is the following set of pairs $(\hat{x},g)$:

$$Y=\{(\hat{x},g):\hat{x}\in S^{N-1},g\in(g),|g|=1\}. \tag{2.38}$$

Theorem 2.3: *Let $\rho:\mathbb{R}^N\to\mathbb{R}_+$, $\rho\in C^1(\mathbb{R}^N)$ and (2.13) and (2.16) be satisfied where $\alpha<1$, $Z\geqslant 0$. Suppose that ϕ is a δ-asymptotic phase for ρ, its wave fronts are strongly convex, $\phi\in C^3(\Omega_{r_0})$, (2.25) is valid where $\hat{\phi}>0$ on S^{N-1}, $\alpha<1$, and*

$$2\delta>3(1-\alpha). \tag{2.39}$$

Then the following hold.

(i) *For any* $\alpha<0$ *there exists such* $\gamma\in(\alpha,1]$ *that for each* $f\in V$ *[see (2.23)], the problem (1.3) and (2.14) has one and only one solution.*

(ii) *For any* $\alpha\geqslant 0$ *such* γ *exists under the following additional conditions:*

$$\sigma_0>0, \tag{2.40}$$

$$\alpha<M\sigma_0(1+\sqrt{1-\sigma_0^2})^{-1}. \tag{2.41}$$

Corollary 2.2: Suppose that ϕ *satisfies the conditions of Theorem 2.3,* $\rho:\mathbb{R}^N\to\mathbb{R}_+$, $\rho\in C^1(\mathbb{R}^N)$, *and in* Ω_{r_0} *the function*

$$\varphi=\rho-|\nabla\phi|^2 \tag{2.42}$$

satisfies (2.24), where $|j|=0,1$, $2\delta>3(1-\alpha)$. *Then the statement of Theorem 2.3 holds.*

Note 2.5: The conditions (2.40) and (2.41) can be easily verified if

$$\phi=\phi_0^{1-\alpha},$$

where ϕ_0 is a homogeneous function of degree 1 and ϕ_0 does not depend on α. Then wave fronts and, consequently, σ_0 and M do not depend on α.

For instance, we can see, returning to Example 2.1, that $\Gamma(c)$ is a sphere $r=c$ and $\psi=0$, $s=1$, $\sigma=1$, $K_1=1$, and $\beta=\pi/2$ for all $\hat{x}$, g. Hence $M=1$ and conditions (2.40) and (2.41) are satisfied for any $\alpha\in[0,1)$.

Example 2.3: Let $N=2$, $\alpha<1$,

$$\phi(x)=(a_1^{-2}x_1^2+a_2^{-2}x_2^2)^{(1-\alpha)/2},$$

$\alpha,a_1,a_2=$const, and $a_1\geqslant a_2>0$. Then the wave fronts are ellipses:

$$a_1^{-2}x_1^2+a_2^{-2}x_2^2=c,\quad c>0. \tag{2.43}$$

Set $t=a_2^2a_1^{-2}$. It is easy to obtain that

$$\sigma_0=(1+t)^{-2}(6t-1-t^2), \tag{2.44}$$

$$M=(8t)^{-1}(6t-1-t^2)\quad\text{for}\quad t\in(3-2\sqrt{2},\tfrac{1}{3}), \tag{2.45}$$

$$M=t\quad\text{for}\quad t\in[\tfrac{1}{3},1]. \tag{2.46}$$

Thus $\sigma_0>0$ for $a_1^{-1}a_2\in(\sqrt{2}-1,1]$. Suppose that ρ satisfies the condition of Corollary 2.2. Then the RP is valid for $\alpha=0$ if $a_1^{-1}a_2\in(\sqrt{2}-1,1]$ and for $\alpha>0$ under the additional condition (2.41) where σ_0 and M are given by (2.44)–(2.46).

III. PROOF OF THEOREM 2.1

Lemma 3.1: Let $\varkappa,q\in\mathbb{R}$, $\rho:\mathbb{R}^N\to\mathbb{R}_+$, $u:\Omega_{r_0}\to\mathbb{C}$, $r_0>0$,

$$\int_{\Omega_{r_0}} r^{\varkappa}\rho|u|^2\,dx<\infty, \tag{3.1}$$

$$\rho r^{2\alpha}\geqslant c>0,\quad x\in\Omega_{r_0}. \tag{3.2}$$

(i) *If* $q\leqslant\min\{2-N,\varkappa+2(1-\alpha)\}$, *then*

$$\liminf_{R\to\infty} R^q\int_{S(R)}(\bar{u}u_r+u\bar{u}_r)\,ds\leqslant 0. \tag{3.3}$$

(ii) *If* $q>2-N$, $\ell\geqslant 2\alpha-N$, *then (3.3) holds.*

These statements can be obtained from the identity

$$\frac{d}{dr}\int_{S^{N-1}}|u|^2\,ds=\int_{S^{N-1}}(\overline{u}u_r+u\overline{u_r})ds. \tag{3.4}$$

Lemma 3.2: Let u *be a solution of the equation (1.2) on* Ω_{r_0}, $r_0>0$, *where (3.1) and (3.2) are satisfied with* $\alpha\leqslant 1$, $\ell=\gamma-1$, and $\gamma>0$. *Then*

$$\int_{\Omega_{r_0}} r^{\gamma-1}|\nabla u|^2\,dx<\infty. \tag{3.5}$$

Proof: By Lemma 3.1 we obtain (3.3) with $q=\gamma-1$. Multiply (1.2) by $r^{\gamma-1}\overline{u}$, pass to real parts, and integrate over $\Omega(r_0,R)$, $R>r_0$. The obtained equality, together with (3.3) where $q=\gamma-1$, implies (3.5). □

Lemma 3.3: Let u *be a solution of the equation (1.2) where* ρ: $\mathbb{R}^N\to\mathbb{R}_+$, $\rho\in C^1(\mathbb{R}^N)$. *Conditions (2.12) and (2.13) hold with* $\alpha<1$,

$$\int (1+r)^{\gamma-1}(\rho|u|^2+|\nabla u|^2)\,dx<\infty, \tag{3.6}$$

where $\gamma\in\mathbb{R}$, and γ *satisfies (2.15).*

Then $u=0$ *a.e. on* $\mathbb{R}^N$.

Proof: Without loss of generality we can assume that Im $u=0$ and, from (2.13) and (2.15), that $\gamma\leqslant 1$.

Multiply (1.2) by $r^{\gamma}u_r$ and integrate over Ω_{R_0}, $R_0>r_0>0$. Integrating by parts we can obtain by (3.6) the following equality:

$$\begin{aligned}
&R_0^{\gamma}\int_{S(R_0)}(u_r^2+1/2(k^2\rho u^2-|\nabla u|^2))\,dS+(1-\gamma)\\
&\quad\times\int_{\Omega_{R_0}} r^{\gamma-1}(|\nabla u|^2-u_r^2)\,dx+1/2(N-1-\alpha)\\
&\quad\times\int_{\Omega_{R_0}} r^{\gamma-1}(k^2\rho u^2-|\nabla u|^2)\,dx+k^2/2\\
&\quad\times\int_{\Omega_{R_0}} r^{\gamma-2\alpha}(r^{2\alpha}\rho)_r u^2\,dx+(\gamma-\alpha)/2\\
&\quad\times\int_{\Omega_{R_0}} r^{\gamma-1}(k^2\rho u^2+|\nabla u|^2)dx=0.
\end{aligned} \tag{3.7}$$

By (1.2), (2.12), (2.15), and (3.7) we obtain the inequality

$$\begin{aligned}
&\int_{\Omega_r} r^{\gamma-1}(|\nabla u|^2+k^2\rho u^2)\,dx\\
&\quad\leqslant c_1 r^{\gamma}\int_{S(r)}(|\nabla u|^2-k^2\rho u^2)\,dS-c_2 r^{\gamma-1}\int_{S(r)} uu_r\,dS,\quad c_1,c_2>0,
\end{aligned} \tag{3.8}$$

which is valid for all $r\geqslant r_1>0$ where r_1 is sufficiently large. Multiplying (3.8) by $r^{1-2\alpha}$ and integrating with respect to r from a fixed $r>r_1$ to $R>r$ we obtain

$$\int_r^R r^{1-2\alpha}\,dr\int_{\Omega_r} r^{\gamma-1}(|\nabla u|^2+k^2\rho u^2)\,dx$$

$$\leqslant c_1\int_{\Omega(r,R)} r^{1+\gamma-2\alpha}(|\nabla u|^2-k^2\rho u^2)\,dx-c_2\int_{\Omega(r,R)} r^{\gamma-2\alpha}uu_r\,dx. \tag{3.9}$$

After integration by parts we obtain by Lemma 3.1 that

$$I(r):=\int_{\Omega_r} r^{1+\gamma-2\alpha}(k^2\rho u^2+|\nabla u|^2)\,dx<\infty \tag{3.10}$$

and

$$I(r)\leqslant r^{2-2\alpha}\int_{\Omega_r} r^{\gamma-1}(k^2\rho u^2+\nabla u|^2)\,dx+c_3\int_{\Omega_r} r^{\gamma-1-2\alpha}u^2\,dx+c_4r^{\gamma-2\alpha}$$

$$\times\int_{S(r)} u^2\,dS-c_5r^{1+\gamma-2\alpha}\int_{S(r)} uu_r\,dS. \tag{3.11}$$

Multiply (3.11) by $r^{1-2\alpha}$ and integrate again with respect to r as above. It results in

$$\int_{\Omega_r} r^{\gamma-1+2p(1-\alpha)}(k^2\rho u^2+|\nabla u|^2)\,dx<\infty \tag{3.12}$$

where $p=2$. Repeating this process p times we obtain that (3.12) holds for any integer p.

Set

$$v(x,m)=r^m u,\quad m>0. \tag{3.13}$$

By (1.2) on Ω_{r_0}, $r_0>0$,

$$\Delta v-\frac{2m}{r}\,v_r+\left(k^2\rho+\frac{m(m-N+2)}{r^2}\right)v=0. \tag{3.14}$$

From (3.12) $\forall m>0$ we have

$$\int r^m(|\nabla v|^2+\rho v^2)\,dx<\infty. \tag{3.15}$$

Multiply (3.14) by rv_r and integrate over Ω_r. After integration by parts we obtain by (3.15)

$$I_1(r):=r\int_{S(r)}\left(v_r^2+\frac{1}{2}\left(\left(k^2\rho+\frac{m(m-N+2)}{r^2}\right)v^2-|\nabla v|^2\right)+\frac{N-2}{2}\left(\frac{m}{r^2}\,v^2-\frac{vv_r}{r}\right)\right)dS$$

$$=-\int_{\Omega_r}\left(2m\left(v_r^2-\frac{(N-2)^2}{4}\,\frac{v^2}{r^2}\right)+k^2\left(\rho+\frac{r\rho_r}{2}\right)v^2\right)dx. \tag{3.16}$$

By the Hardy inequality we have

$$\int_{\Omega_r}\left(v_r^2-\frac{(N-2)^2}{4}\,\frac{v^2}{r^2}\right)dx\geqslant 0. \tag{3.17}$$

The inequalities (3.16) and (3.17) together with (2.13) imply that for $r\geqslant r_2>0$, $\forall m>0$

$$I_1(r)\leqslant 0, \tag{3.18}$$

where r_2 is a sufficiently large number. Returning by (3.13) to the initial function u we can write $I_1(r)$ in the form

$$I_1(r)=r^{2m+1}(Am^2+Bm+C), \tag{3.19}$$

where $A=A(r)$, $B=B(r)$, and $C=C(r)$ are independent of m, and

$$A=r^{-2}\int_{S(r)} u^2\, dS. \tag{3.20}$$

Now (3.18)–(3.20) imply that $u=0$ a.e. on Ω_{r_2}, which in turn yields that $u=0$ a.e. on $\mathbb{R}^N$.[22] □

Note 3.1: Lemma 3.3 follows for some α, γ from a Rellich-type[2] lower estimate of solutions of the Schroedinger equation, given by Agmon.[21] For instance, suppose $Z=0$ (as for $\rho=r^{-2\alpha}$, $r>r_0>0$), $\alpha\in(0,1)$. Then the mentioned estimate together with (2.12) and (2.13) implies that Lemma 3.3 is valid for $\gamma>3\alpha$, which is more restrictive than (2.15).

Proof of Theorem 2.1: Integrate the identity

$$ik\phi^{-q}(\overline{u}u_\nu-u\overline{u_\nu})=\phi^{-q}\phi_\nu^{-1}(|u_\nu-ik\phi_\nu u|^2-|u_\nu|^2-k^2\phi_\nu^2|u|^2), \tag{3.21}$$

where

$$q=(1-\gamma)\tau^{-1}, \tag{3.22}$$

over a closed surface $\Gamma(c)$, $c\geqslant c_0$, where ν, defined by (2.10), is an external normal for $\Gamma(c)$ and ϕ_ν and u_ν are the corresponding derivatives. By the equation (1.2) and the definition of $\Gamma(c)$ we obtain

$$\int_{\Gamma(c)}\frac{|u_\nu|^2+k^2\phi_\nu^2|u|^2}{\phi_\nu\phi^q}\,dS=\int_{\Gamma(c)}\frac{|u_\nu-ik\phi_\nu u|^2}{\phi_\nu\phi^q}\,dS. \tag{3.23}$$

Integrate this identity with respect to c over the interval $c_1\leqslant c\leqslant c_2$, $c_1\geqslant c_0$. Setting

$$\Omega'(c_1,c_2)=\{x:x\in\mathbb{R}^N, c_1\leqslant\phi(x)\leqslant c_2\}$$

we obtain by (2.7) and

$$dx=\phi_\nu^{-1}ds\; dc$$

the following inequality:

$$\int_{\Omega'(c_1,c_2)} r^{\gamma-1}\phi_\nu^2|u|^2\, dx\leqslant b_0\int_{\Omega'(c_1,c_2)} r^{\gamma-1}|u_\nu-ik\phi_\nu u|^2\, dx, \tag{3.24}$$

where $b_0>0$ does not depend on c_1,c_2. Due to Definition 2.3 and Note 2.1 there exists such $c_3>c_0$ for any $\varepsilon>0$ that outside $\Gamma(c_3)$ we have

$$|\phi_\nu-\sqrt{\rho}|<\varepsilon\sqrt{\rho}. \tag{3.25}$$

Choose $\varepsilon>0$ such that

$$2b_0\varepsilon^2(1-\varepsilon)^{-2}<1. \tag{3.26}$$

Then (3.24) yields for $c_2>c_1>c_3$

$$\int_{\Omega'(c_1,c_2)} r^{\gamma-1}\rho|u|^2\, dx\leqslant b_1\int_{\Omega'(c_1,c_2)} r^{\gamma-1}|u_\nu-ik\sqrt{\rho}u|^2\, dx,$$

where b_1 does not depend c_2, $c_2>c_1$. Now (2.14) together with Lemmas 2.2 and 2.3 implies that $u=0$ a.e. on $\mathbb{R}^N$. □

IV. PROOF OF THEOREM 2.2

In order to prove the theorem we shall make use of the following limiting absorption method:[3] first we deal with the perturbed equation

$$\Delta u+(k^2+\varepsilon i)\rho u=f, \tag{4.1}$$

where the imaginary absorption term $i\varepsilon\rho u$, $\varepsilon=\text{const}>0$, is introduced. This equation can be written in the form (2.5) and (2.6); it was stated in Sec. II that, for $F\in H$, the equation (2.5) has a unique solution $u_\varepsilon\in\mathscr{D}(\mathscr{S})$. Moreover, from (2.2) we have

$$u_\varepsilon\in H^1_\rho. \tag{4.2}$$

Under condition (2.3) the set $\mathscr{D}(\mathscr{S})$ can be defined by (2.4). Imposing on F a condition more restrictive than $F\in H$, we shall prove that u_ε converges as $\varepsilon\to 0$, in a less restrictive space than H, and its limit u is a solution of the problem (1.3) and (2.14). To this end we have to obtain some estimates for u_ε analogous with those proved by Jäger.[23] Put

$$u_\varepsilon=e^{ik\phi(x)}v_\varepsilon. \tag{4.3}$$

From now on we write v instead of v_ε. Definition 2.3 together with (4.2) implies

$$v\in H^1_\rho. \tag{4.4}$$

By (4.1) a.e. on $\mathbb{R}^N$

$$\Delta v+2ki(\nabla v,\nabla\phi)+i(k\Delta\phi+\varepsilon\rho)v+k^2(\rho-|\nabla\phi|^2)v=e^{-ik\phi}f. \tag{4.5}$$

In particular (4.5) implies that $v\in\mathscr{D}(\mathscr{S})$. Choose some numbers δ_1, R_0, and R_1 so that $r_0<R_0<R_1$ and $0<\delta_1<\frac{1}{2}$, and on Ω_{R_0}

$$\begin{gathered}\phi(x)\geqslant 1,\\ 1-\delta_1\leqslant\rho/|\nabla\phi|^2\leqslant 1+\delta_1.\end{gathered} \tag{4.6}$$

Introduce a smooth cut-function $\theta:\mathbb{R}^N\to\mathbb{R}$, $0\leqslant\theta\leqslant 1$,

$$\theta=0 \quad \text{for } x\in\Omega(R_0), \tag{4.7}$$

$$\theta=1 \quad \text{for } x\in\Omega_{R_1}. \tag{4.8}$$

Now multiply (4.5) by $\theta\phi^s\overline{v}$, where $s=\text{const}$,

$$s<0, \tag{4.9}$$

pass to imaginary parts, and integrate over $\mathbb{R}^N$. After integration by parts we obtain

$$\begin{aligned}&\varepsilon\int\theta\phi^s\rho|v|^2\,dx-ks\int\theta\phi^{s-1}|\nabla\phi|^2|v|^2\,dx-\text{Im}\,s\int\theta\phi^{s-1}\overline{v}(\nabla v,\nabla\phi)\,dx\\ &\quad-\int\phi^s(|v|^2\,\text{Re}(\nabla\theta,\nabla\phi)+\text{Im}((\nabla v,\nabla\theta)\overline{v}))\,dx=\text{Im}\int\theta\phi^s f\overline{v}e^{-ik\phi}\,dx.\end{aligned} \tag{4.10}$$

Putting $s=1+\mu(1-\alpha)^{-1}$ into (4.10), where

$$\mu<-(1-\alpha), \tag{4.11}$$

we obtain that for μ satisfying (4.11)

$$\varepsilon\int (1+r)^{1+\mu-\alpha}\rho|v|^2\,dx+\int (1+r)^{\mu}\rho|v|^2\,dx$$
$$\leqslant c(\mu)\left(\int_{\Omega(R_1)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{\mu}|\nabla v|^2\,dx+\int (1+r)^{\mu+2}|f|^2\,dx\right), \tag{4.12}$$

where $c(\mu)$ is independent of ε, f.

Let $b:H_\rho\times\mathbb{R}_+\times V\to\mathbb{R}$, $B:\mathbb{R}_+\times V\to\mathbb{R}$, where $B(\varepsilon,f)=b(v,\varepsilon,f)\ \forall(\varepsilon,f)\in\mathbb{R}_+\times V$ where $v=v(\epsilon,f)\in H_\rho$ as above.

Condition Q: The functional B satisfies the condition Q if there exists such a constant $\delta_1>0$ and a function $c(\delta_0)>0$, $\delta_0>0$, which are independent of ε, f, that for any $\delta_0>0$

$$|B|\leqslant\delta_0\int (1+r)^{\gamma-1}|\nabla v|^2\,dx+c(\delta_0)$$
$$\times\left(\int (1+r)^{\gamma-1-\delta_1}|\nabla v|^2\,dx+\int_{\Omega(R_1)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\right). \tag{4.13}$$

Now multiply (4.5) by

$$\theta\phi^m((\nabla\phi,\nabla v)-(2k)^{-1}(k\Delta\phi+\varepsilon\rho)\overline{v}),$$

integrate over $\mathbb{R}^N$, and pass to real parts. After integration by parts we obtain

$$\sum_{j=0}^{11} I_j=0, \tag{4.14}$$

where I_j are the following integrals which are absolutely convergent because of $\alpha<1$, $\gamma\leqslant 1$, and (4.4):

$$I_0=\mathrm{Re}\int \phi^m\left((\nabla v,\nabla\theta)(\nabla\phi,\nabla v)-\frac{1}{2}(\nabla\theta,\nabla\phi)|\nabla v|^2+\frac{1}{2k}(\nabla\theta,\nabla v)(k\Delta\phi+\epsilon\rho)v\right)dx,$$

$$I_1=\int \theta(G(x,\nabla(\mathrm{Re}\ v),\gamma)+G(x,\nabla(\mathrm{Im}\ v),\gamma))\,dx,$$

$$I_2=\frac{m}{2}\,\mathrm{Re}\,i\int \phi^{m-1}\theta v(\nabla\phi,\nabla v)\Delta\phi\,dx,$$

$$I_3=\frac{\varepsilon m}{2k}\,\mathrm{Re}\int \theta\phi^{m-1}\rho v(\nabla\phi,\nabla v)\,dx,$$

$$I_4=\frac{1}{2}\,\mathrm{Re}\int \theta\phi^m v(\nabla(\Delta\phi),\nabla v)\,dx,$$

$$I_5=\frac{\varepsilon}{2k}\,\mathrm{Re}\int \theta\phi^m v(\nabla\rho,\nabla v)\,dx,$$

$$I_6=k^2\,\mathrm{Re}\int \theta\phi^m(|\nabla\phi|^2-\rho)(\nabla\phi,\nabla v)v\,dx,$$

$$I_7=\frac{k^2}{2}\,\mathrm{Re}\int \theta\phi^m(|\nabla\phi|^2-\rho)|v|^2\Delta\phi\,dx,$$

$$I_8=\frac{k\epsilon}{2}\int \theta\phi^m(|\nabla\phi|^2-\rho)\rho|v|^2\,dx,$$

$$I_9=\mathrm{Re}\int \theta\phi^m e^{-ik\phi}f(\nabla\phi,\nabla v)\,dx,$$

$$I_{10}=\frac{1}{2}\,\mathrm{Re}\int \theta\phi^m e^{-ik\phi}\overline{vf}\Delta\phi\,dx,$$

$$I_{11}=\frac{\varepsilon}{2k}\,\mathrm{Re}\int \theta\phi^m e^{-ik\phi}\rho\overline{vf}\,dx.$$

From (4.7) and (4.8) the integral I_0 satisfies the condition Q. Due to (2.20)

$$I_1\geqslant c\int_{\Omega_{R_1}} r^{\gamma-1}|\nabla v|^2\,dx. \tag{4.15}$$

By $\alpha<1$, $\gamma\leqslant 1$, $\phi\in U_{1-\alpha}$, and (2.18) we obtain

$$|I_2|\leqslant \delta_0\int (1+r)^{\gamma-1}|\nabla v|^2\,dx+c(\delta_0)\int (1+r)^{\mu}\rho|v|^2\,dx, \tag{4.16}$$

where $\delta_0>0$ is arbitrarily small, $c(\delta_0)$ does not depend on ε and f, and

$$\mu=\gamma-1+2(\alpha-1).$$

Since $\mu<\alpha-1$ we can apply the inequality (4.12) to the second integral in (4.16). So, we reach the conclusion that I_2 satisfies the condition Q. It is easy to see that

$$|I_3|\leqslant \delta_0\int (1+r)^{\gamma-1}|\nabla v|^2\,dx+\varepsilon^2c(\delta_0)\int (1+r)^{\gamma-1}\rho|v|^2\,dx. \tag{4.17}$$

By (4.1) we have

$$\varepsilon^2\int \rho|v|^2\,dx\leqslant\int \rho^{-1}|f|^2\,dx. \tag{4.18}$$

Now (4.17) and (4.18) imply that I_3 satisfies the condition Q. The terms I_4 and I_5 can be estimated in the same way as I_2 and I_3, respectively, so they also satisfy the condition Q.

By Definition 2.3 we obtain

$$|I_6|\leqslant \delta_0\int (1+r)^{\gamma-1}|\nabla v|^2\,dx+c(\delta_0)\int (1+r)^{\mu_1}\rho|v|^2\,dx, \tag{4.19}$$

where

$$\mu_1=1+\gamma-2\alpha-2\delta.$$

From (2.19)

$$\mu_1<\alpha-1,$$

and we can apply (4.12) to the last term in (4.19). As a result we obtain that I_6 satisfies the condition Q. Analogously we obtain the same for I_7, I_8, and I_9. Now

$$|I_{10}|\leqslant c\left(\int (1+r)^{\alpha-1-\delta_0}\rho|v|^2\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\right)$$

where

$$\delta_0=2-\alpha-\gamma>0.$$

It follows from (4.12) that I_{10} satisfies Q. From (4.18) the same holds for I_{11}. So because of (4.14) I_1 satisfies the condition Q, i.e., the inequality (4.13) is valid for $B=I_1$. This and (4.15) imply

$$c_1\int (1+r)^{\gamma-1}|\nabla v|^2\,dx\leqslant\delta_0\int (1+r)^{\gamma-1}|\nabla v|^2\,dx+c(\delta_0)\Bigg(\int (1+r)^{\gamma-1-\delta_1}|\nabla v|^2\,dx + \int_{\Omega(R_1)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\Bigg). \tag{4.20}$$

Choose here a sufficiently small number $\delta_0>0$. Then

$$\int (1+r)^{\gamma-1}|\nabla v|^2\,dx\leqslant c_2\Bigg(\int (1+r)^{\gamma-1-\delta_1}|\nabla v|^2\,dx+\int_{\Omega(R_1)}(|v|^2+|\nabla v|^2)\,dx + \int (1+r)^{1+\gamma}|f|^2\,dx\Bigg), \tag{4.21}$$

where c_j do not depend on ε and f. From (4.21) we can choose a sufficiently large number $R_2>R_1$ and obtain the inequality

$$\int (1+r)^{\gamma-1}|\nabla v|^2\,dx\leqslant c\Bigg(\int_{\Omega(R_2)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\Bigg), \tag{4.22}$$

where c is independent of ε and f. By (4.12) and (4.22) we have

$$\int (1+r)^{\alpha-1-\delta_0}\rho|v|^2\,dx\leqslant c(\delta_0)\Bigg(\int_{\Omega(R_2)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\Bigg), \tag{4.23}$$

where δ_0 is an arbitrarily small positive number. Let us introduce the norm

$$\|g\|^2_{\delta_0}=\int ((1+r)^{\alpha-1-\delta_0}\rho|g|^2+(1+r)^{\gamma-1}|\nabla g|^2)\,dx.$$

Combining (4.22) and (4.23) we obtain

$$\|v_\epsilon\|^2_{\delta_0}\leqslant c(\delta_0)\Bigg(\int_{\Omega(R_2)}(|v|^2+|\nabla v|^2)\,dx+\int (1+r)^{1+\gamma}|f|^2\,dx\Bigg), \tag{4.24}$$

where $v_\varepsilon=v$ and $c(\delta_0)$ is independent of $(\varepsilon,f)\in(0,1)\times V$.

Now we shall prove that for fixed $\delta_0,k>0,f\in V$, the following inequality holds:

$$\|v_\varepsilon\|_{\delta_0}\leqslant c(\delta_0), \tag{4.25}$$

where $c(\delta_0)$ does not depend on $\varepsilon\in(0,1)$. Without loss of generality we can assume that $\delta_0<2(1-\alpha)$. Suppose that (4.25) is incorrect for some $(\delta_0,f)\in(0,2(1-\alpha))\times V$. Then there exists such a sequence $\varepsilon_n\to0$ as $n\to\infty$ that

$$N_n:=\|v_{\varepsilon_n}\|_{\delta_0}\to\infty. \tag{4.26}$$

Set $v_{\varepsilon_n}=v^{(n)}$,

$$w_n=v^{(n)}N_n^{-1},$$

$$U_n = w_n \exp(ik\phi). \tag{4.27}$$

Then

$$\Delta U_n + (k^2 + \varepsilon_n i)\rho U_n = f N_n^{-1}, \tag{4.28}$$

$$\|w_n\|_{\delta_0} = 1. \tag{4.29}$$

We shall make use of the following well-known inequality (see, e.g., Ref. 24):

$$\|g\|_{2,\Omega(R_1)} \leqslant c(R_1, R_2)(\|g\|_{0,\Omega(R_2)} + \|\Delta g\|_{0,\Omega(R_2)}), \tag{4.30}$$

where $0 < R_1 < R_2$, $g \in H^2_{\text{loc}}$, and $c(R_1, R_2)$ does not depend on g. From (4.28)–(4.30) we have

$$\|U_n\|_{2,\Omega(R)} \leqslant c(R), \tag{4.31}$$

which is valid for any $R > 0$ and $c(R)$ does not depend on n. It follows from (4.30) and (4.31) that there exists a subsequence U_{n_m} which converges, as $m \to \infty$, in H^2_{loc}, to a solution U of the equation

$$\Delta U + k^2 \rho U = 0 \tag{4.32}$$

on $\mathbb{R}^N$. In this connection we have from (4.27) that $w_{n_m} \to w$ in H^2_{loc} and

$$U = w \exp(ik\phi), \tag{4.33}$$

$$\|w\|_{\delta_0} \leqslant 1. \tag{4.34}$$

From (4.33)

$$U_\nu - ik\sqrt{\rho}U = e^{ik\varphi}(w_\nu + ikw(|\nabla\phi| - \sqrt{\rho})). \tag{4.35}$$

Due to the definition of δ-asymptotic phase

$$\big||\nabla\phi| - \sqrt{\rho}\big| \leqslant c r^{-\delta}\sqrt{\rho}, \quad x \in \Omega_{r_0},$$

hence

$$\int (1+r)^{\gamma-1} |U_\nu - ik\sqrt{\rho}U|^2 \, dx \leqslant c \int ((1+r)^{\gamma-1-2\delta}\rho|w|^2 + (1+r)^{\gamma-1}|\nabla w|^2) \, dx. \tag{4.36}$$

By (2.19) $\gamma - 1 - 2\delta \leqslant 3(\alpha - 1)$, and because of our choice of δ_0

$$\gamma - 1 - 2\delta < \alpha - 1 - \delta_0.$$

From (4.34) and (4.36) we have

$$\int (1+r)^{\gamma-1} |U_\nu - ik\sqrt{\rho}U|^2 \, dx \leqslant c\|w\|^2_{\delta_0} < \infty. \tag{4.37}$$

Theorem 2.1 together with (4.32) and (4.37) implies that $U = 0$ a.e. on $\mathbb{R}^N$. So $w_{n_m} \to 0$, as $m \to \infty$, in H^2_{loc}. On the other hand, from (4.24) we have

$$1 = \|w_n\|^2_{\delta_0} \leqslant c\left(\int_{\Omega(R_2)} (|w_n|^2 + |\nabla w_n|^2) \, dx + N_n^{-2} \int (1+r)^{1+\gamma} |f|^2 \, dx \right). \tag{4.38}$$

For $n = n_m$ the right side of (4.38) tends to zero as $m \to \infty$, which contradicts (4.38). So the inequality (4.25) is proved.

Now (4.1), where $u=u_\varepsilon$, and (4.25) and (4.30) yield

$$\|u_\varepsilon\|_{2,\Omega(R)}\leqslant c(R), \tag{4.39}$$

where $c(R)$ is independent of ε.

Therefore as above we can choose a sequence u_{ε_m} which converges in H^2_{loc}, and $\varepsilon_m\to 0$ as $m\to\infty$. The limit function u is a solution of (1.3). Moreover the sequence

$$v_{\varepsilon_m}=u_{\varepsilon_m}\exp(-ik\phi)$$

also has a limit v in H^2_{loc}. From (4.25) we obtain, as above for U, that the function u satisfies the radiation condition (2.14). □

V. PROOF OF THEOREM 2.3

Throughout this section we assume that the conditions of Theorem 2.3 are valid.

Lemma 5.1: Suppose $\sigma_0>0$ and M is defined by (2.37). Then $M\leqslant 1$.

Proof: Suppose $M>1$. By (2.37)

$$\begin{gathered}K_1(\hat{x},g)\geqslant(\sigma s)^{-1}M(\sigma+\cos^2\beta)\geqslant M,\\ K(x,q)=r^{-1}K_1\geqslant r^{-1}M.\end{gathered} \tag{5.1}$$

Let $c>0$ be a fixed number and $\gamma(c)$ be a one-dimensional curve on $\Gamma(c)$, which is an intersection of $\Gamma(c)$ and some two-dimensional plane (σ) passing through the origin $x=0$. Denote by ℓ the arc length on $\gamma(c)$, by $g_1(x)$ the unit tangent vector, and by $\kappa(x)$ the curvature of $\gamma(c)$ at the point $x\in\gamma(c)$. It is easy to see that

$$2\pi=\int_{\gamma(c)}\kappa\,d\ell\geqslant\int_{\gamma(c)}K(x,g_1(x))\,d\ell. \tag{5.2}$$

Passing to polar coordinates r, φ on the plane (σ) we obtain by (5.1)

$$\int_{\gamma(c)}K\,d\ell=\int_0^{2\pi}K\sqrt{r^2+r_\varphi^2}\,d\varphi\geqslant 2\pi M.$$

This together with (5.2) is in contradiction with our assumption. □

Proof of Theorem 2.3: Let us notice that because of (2.31) and Note 2.2 the condition (2.12) is valid. Therefore all the conditions of Theorem 2.1 are satisfied. The conditions (2.16) and (2.18) of Theorem 2.2 are also satisfied, and we have to verify its condition (2.20) only. For $Z\geqslant 0$ and (2.39) the interval (2.19) coincides with $(\alpha,1]$.

Let $x\in\mathbb{R}^N$ be a point belonging to some wave front $\Gamma(c)$, $c>0$. The vector ν and the plane (g) are defined at the point x as in Sec. II. Consider an arbitrary vector $\xi\in\mathbb{R}^N$ and expand it as follows:

$$\xi=\xi_x+\xi_g, \tag{5.3}$$

where $\xi_x=b_1x$, $\xi_g=b_2g$, $b_j\in\mathbb{R}$, $j=1,2$, and $g\in(g)$, $|g|=1$. For any two vectors ξ_1, $\xi_2\in\mathbb{R}^N$ put

$$\widetilde{\phi}(\xi_1,\xi_2)=\phi_{ij}(x)\xi_1^{(i)}\xi_2^{(j)},\quad \widetilde{\phi}(\xi)=\widetilde{\phi}(\xi,\xi).$$

From (2.34)

$$\widetilde{\phi}(\xi_g)=r^{-1}K_1(\hat{x},g)|\nabla\phi||\xi_g|^2. \tag{5.4}$$

The functions ϕ_i and ϕ_{ij} are homogeneous of degree $-\alpha$ and $-1-\alpha$, respectively. Therefore

$$\widetilde{\phi}(\xi_x) = -\alpha(1-\alpha)r^{-2}\phi|\xi_x|^2,$$

$$\widetilde{\phi}(\xi_x,\xi_g) = 0,$$

$$\widetilde{\phi}(\xi) = r^{-1}|\nabla\phi|K_1(\hat{x},g)|\xi_g|^2 - \alpha(1-\alpha)r^{-2}\phi|\xi_x|^2. \tag{5.5}$$

From (2.27) and (5.5)

$$G(x,\xi,\gamma) = \widetilde{A}_1|\xi_x|^2 + 2\widetilde{A}_2|\xi_x||\xi_g| + \widetilde{A}_3|\xi_g|^2. \tag{5.6}$$

where $\widetilde{A}_j = \chi A_j$,

$$\chi = \tfrac{1}{2}\phi^m|\nabla\phi|r^{-1}s^{-1},$$

$$A_1 = \sigma\gamma - \alpha,$$

$$A_2 = -(\alpha+\gamma)\cos\beta,$$

$$A_3 = 2K_1 s - \alpha - \gamma.$$

We have $\chi > 0$, hence it is sufficient to find $\gamma \in (\alpha,1]$ such that for any $(\hat{x},g) \in Y$ [see (2.38)]

$$A_1 > 0, \tag{5.7}$$

$$A_1A_3 - A_2^2 > 0. \tag{5.8}$$

These inequalities can be written as follows:

$$\sigma\gamma - \alpha > 0, \tag{5.9}$$

$$(\sigma+\cos^2\beta)\gamma^2 - 2\gamma(s\sigma K_1 + \alpha(\sin^2\beta - s^2)) + \alpha(2sK_1 - \alpha\sin^2\beta) < 0. \tag{5.10}$$

It is obvious that if $\alpha < 0$, then both (5.9) and (5.10) hold for all sufficiently small $|\gamma|$. If $\alpha = 0$, then it is sufficient for (5.9) and (5.10) that (2.40) and $\gamma \in (0,2M)$ are valid. Let us go on to a more complicated case $\alpha > 0$. By (2.41)

$$d := M^2 - 2\alpha\sigma_0^{-1}M + \alpha^2 > 0. \tag{5.11}$$

Consider the inequality

$$\gamma^2 - 2M\gamma + \alpha(2\sigma_0^{-1}M - \alpha) < 0, \tag{5.12}$$

which holds for all γ from the interval $(\gamma) := (\gamma_-,\gamma_+)$ where

$$\gamma_\pm = M \pm \sqrt{d}. \tag{5.13}$$

We shall prove that for these γ the inequalities (5.9) and (5.10) hold. First let us show that

$$\gamma_- \geqslant \alpha\sigma_0^{-1} \geqslant \alpha. \tag{5.14}$$

By (2.41) we have $M > \alpha\sigma_0^{-1}$. This and

$$(M - \alpha\sigma_0^{-1})^2 - d = \alpha^2(\sigma_0^{-2} - 1) \geqslant 0$$

imply (5.14), which in turn yields (5.9) for all $\gamma \in (\gamma)$, $\hat{x} \in S^{N-1}$. Rewrite (5.12) in the form

$$\gamma^2 - 2M(\gamma - \alpha\sigma_0^{-1}) - \alpha^2 < 0. \tag{5.15}$$

By (5.14), (2.36), and (2.37) we obtain from (5.15) that for all $\gamma \in (\gamma)$ and $(\hat{x},g) \in Y$ the following inequality holds:

$$\gamma^2-\frac{2s\sigma K_1}{\sigma+\cos^2\beta}(\gamma-\alpha\sigma_0^{-1})-\frac{\alpha^2\sin^2\beta}{\sigma+\cos^2\beta}-\frac{2\gamma\alpha(\sin^2\beta-s^2)}{\sigma+\cos^2\beta}<0,$$

which immediately implies (5.10). From (5.13) and Lemma 5.1 we have $\gamma_-<1$. This together with (5.14) leads to the conclusion that there exist points $\gamma\in(\alpha,1)$ such that (5.7) and (5.8) are valid for all $(\hat{x},g)\in Y$. So, the form $G(x,\xi,\gamma)$ is positive definite for these γ and $x\in\Omega_{r_0}$. Now by a continuity argument and by the obvious inequality

$$r^{1-\gamma}\chi\geqslant c_0>0,\quad x\in\Omega_{r_0}$$

we obtain the validity of (2.20) for chosen γ. □

[1] I. N. Vekua, ''On metaharmonic functions,'' Trudy Mat. Inst. Akad. Nauk GSSR **12**, 105–174 (1943) (in Russian).
[2] F. Rellich, ''Ueber das asymptotische Verhalten der Lösungen von $\Delta u+\lambda u=0$ in unendlichen Gebieten,'' J. Deutsch. Math. Verein. **53**, 57–65 (1943).
[3] B. R. Vainberg, *Asymptotic Methods in Equations of Mathematical Physics* (Gordon and Breach, New York, 1989).
[4] D. Eidus and A. Vinnik, ''On radiation conditions for domains with infinite boundaries,'' Sov. Math. Dokl. **15**, 12–15 (1974).
[5] V. Vogelsang, ''Das Ausstrahlungsproblem für elliptische Differentialgleichungen in Gebieten mit unbeschränktem Rand,'' Math. Z. **144**, 101–124 (1975).
[6] D. Eidus, ''The limiting absorption and amplitude principles for the diffraction problem with two unbounded media,'' Commun. Math. Phys. **107**, 29–38 (1986).
[7] Y. Saitō, ''A remark on the limiting absorption principle for the reduced wave equation with two unbounded media,'' Pac. J. Math. **136**, 183–208 (1989).
[8] G. Roach and B. Zhang, ''On Sommerfeld radiation conditions for the diffraction problem with two unbounded media,'' Proc. R. Soc. Edinburgh Sect. A **121**, 149–161 (1992).
[9] G. Roach and B. Zhang, ''A transmission problem for the reduced wave equation in inhomogeneous media with an infinite interface,'' Proc. R. Soc. London, Ser. A **436**, 121–140 (1992).
[10] S. De Biévre and D. W. Pravica, ''Spectral analysis for optical fibres and stratified fluids I: The limiting absorption principle,'' J. Funct. Anal. **98**, 406–436 (1991).
[11] S. De Biévre and D. W. Pravica, ''Spectral analysis for optical fibres and stratified fluids II: Absence of eigenvalue,'' Commun. Part. Diff. Eq. **17**, 69–97 (1992).
[12] C. Wilcox, *Sound Propagation in Stratified Fluids* (Springer-Verlag, New York, 1984).
[13] R. Weder, *Spectral and Scattering Theory for Wave Propagation in Perturbed Stratified Media* (Springer-Verlag, Berlin, 1991).
[14] W. Jäger and Y. Saitō, ''On the spectrum of reduced wave operator with cylindrical discontinuity,'' Forum Math. **9**, 29–60 (1997).
[15] W. Jäger and Y. Saitō, ''The reduced wave equation in layered materials,'' Osaka J. Math. (to appear).
[16] L. T. Lewis, ''Singular elliptic operators of second order with purely discrete spectra,'' Trans. Am. Math. Soc. **271**, 653–666 (1982).
[17] V. Vogelsang, ''Die absolute Stetigkeit des positiven Spektrums der Schwingungsgleichung mit oszillierendem Hauptteil,'' Math. Z. **181**, 201–213 (1982).
[18] D. Eidus, ''The perturbed Laplace operator in a weighted L^2-space,'' J. Funct. Anal. **100**, 400–410 (1991).
[19] Y. Saitō, ''Schrödinger operators with nonspherical radiation condition,'' Pac. J. Math. **126**, 331–359 (1987).
[20] V. Vogelsang, ''Die Rellichsche Abschätzung für die Schwingungsgleichung mit oszillivendem Hauptteil,'' J. Reine Augew. Math. **338**, 109–120 (1983).
[21] S. Agmon, ''Lower bounds for solutions of Schroedinger equations,'' J. Anal. Math. **23**, 1–25 (1970).
[22] E. Heinz, ''Ueber die Eindeutigkeit beim Cauchyschen Anfangswertproblem einer elliptischen Differentialgleichung,'' Nachr. Akad. Wiss. Gött. II: Math.-Phys. Kl. **1**, 1–12 (1955).
[23] W. Jäger, ''Zur Theorie der Schwingungsgleichung mit variablen Koeffizienten in Aussengebieten,'' Math. Z. **102**, 62–86 (1967).
[24] L. Nirenberg, ''Estimates and existence of solutions of elliptic equations,'' Commun. Pure Appl. Math. **9**, 509–529 (1956).

On the electrodynamic group: The relativistic radiation reaction force

Angel Prieto[a)]
Physics Department, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England

(Received 5 September 1996; accepted for publication 12 December 1997)

The relativistic force exerted by a classical electromagnetic field on an electron is examined from the point of view that associates the differential geometry of the space–time trajectory with a continuous group of transformations. The group relevant to this particular case is the group generated by the electromagnetic field tensor (electrodynamic group). This group is studied in some detail by treating separately the two parts of the Liénard–Wiechert expression for the tensor. The intrinsic differential equations of the space–time trajectories associated with the electrodynamic group are obtained by the Cartan method of the moving frame. These equations show that the curvature of the trajectory is associated only with the generalized Coulomb field and the hypertorsion with the radiation field. According to these equations, the relativistic force may have two components along orthogonal spacelike directions: The first, an apparent force, reduces to the Coulomb force in the nonrelativistic limit. The other is orthogonal to the four-momentum and to the apparent force; this is the radiation reaction force. In the nonrelativistic limit it manifests itself as a frictional force reducing the momentum. The approach also offers a point of view of the conventional treatment of radiation reaction in which the nonlocality in time emerges from the geometry of the problem. The implications pertaining to other areas of physics of the causal connection found between the hypertorsion and the radiation field are briefly considered. © *1998 American Institute of Physics.* [S0022-2488(98)03603-2]

I. INTRODUCTION

An important identity in the theory of relativity[1] relates the velocity vector, v^μ, of a free particle of mass m, with its energy-momentum vector, p^μ, by

$$p^\mu = mc^2 v^\mu, \tag{1}$$

where

$$v^\mu = \gamma\left(1, \frac{\mathbf{v}}{c}\right), \quad v^2 = 1, \tag{2}$$

$$p^\mu = (E, \mathbf{p}c). \tag{3}$$

In the rest frame of the particle, $(\mathbf{v}=0)$, (1) reduces to Einstein famous energy-mass[2] formula: $E_0 = mc^2$.

When the particle is subject to an external force, the relativistic dynamics involved requires one to consider the relativistic force, f^μ, which equals the (proper time) rate of change of p^μ,

$$f^\mu \equiv \dot{p}^\mu. \tag{4}$$

Equations (1) and (4) lead to

a)Present address: Departamento de Física, Facultad de Ciencias, Universidad Nacional Autónoma de México, Apdo. 40-147, México 06140, D. F. México; Electronic mail: apr@graef.fciencias.unam.mx

 © 1998 American Institute of Physics

$$f^{\mu}=mc^{2}a^{\mu}, \tag{5}$$

where the acceleration a^{μ} is universally taken as

$$a^{\mu}\equiv\dot{v}^{\mu}=\frac{\gamma^{3}}{c^{3}}\,(\mathbf{v}\cdot\mathbf{a})v^{\mu}+\frac{\gamma^{2}}{c^{2}}\,(0,\mathbf{a}). \tag{6}$$

The operation used to obtain the right-hand side of (6) is considered in this paper from a novel point of view. We know that there is a continuous group of transformations invariably associated with a continuous curve, whatever the space in which the curve is embedded. This association determines the intrinsic differential equations of the curve in that space. If the curve represents the trajectory or world line of a point particle, the association, which in particular determines the rate of change of the tangent, by (1) and (4), determines also the relativistic force. In the case considered in this paper, the group and its relation with the curve, absolutely, cannot be ignored.

Now, let the particle be an electron and let the external force be produced by an electromagnetic field. The force (4), as described by (5) and (6), in this important case, can be shown to be incomplete. Specifically, it fails to account for the dynamical effects produced on the charge by the radiated field.[3,4] In fact, the conventional method of calculating the relativistic Larmor rate of radiated energy,[5] leads to

$$\dot{p}^{\mu}=\frac{2}{3}\,\frac{e^{2}}{4\pi}\,|a|^{2}v^{\mu}. \tag{7}$$

This expression, even if it appears as a rate of change of the momentum, cannot be accepted as a legitimate force because it is not orthogonal to the velocity. Following the conventional approach, the Schott[6] term, involving the third time derivative of the position, is added to this expression to produce the Abraham–Lorentz–Dirac force. The orthogonality problem is then corrected but new ones are introduced.

The operation employed to calculate the rate of change of the velocity (6) is based on the Euclidean group of rotations: The differentials with respect to the arclength are obtained in terms of the differentials with respect to ordinary time. This operation is only partially correct. There is, however, a group of space–time rotations which, if used instead, in addition to that result, produces also the radiation reaction force. This group is associated with the trajectories of electrons.

The relation between the curve followed in space-time by a (point) electron and this other group of rotations can be understood very easily. Let the initial position and momentum of the electron in a homogeneous electromagnetic field F be known. Denote by p the initial momentum vector and by p' the momentum vector after the electron has been subject for a time θ to the electromagnetic field. Both vectors p and p' are, necessarily, timelike, of norm $(mc^{2})^{2}$ and future pointing. Consequently, there is a restricted Lorentz transformation $\Lambda(\theta)$ such that: $p'(\theta)=\Lambda(\theta)p$. Clearly, this transformation depends also on the electromagnetic field. In fact, the transformation $\Lambda(F,\theta)=e^{\theta F}$ has already been studied.[7] Let Λ operate on p for a continuous range of values of the parameter θ, starting from zero. Then, p' describes a continuous path on the unit pseudosphere (a hyperboloid) centered at the origin of p. The curve associated with this rotation is fully defined if we also consider the translation of the base point of p' such that c times the length of the arc covered by the translation equals the proper time of the particle and such that the successive orientations of p' defined by θ, coincide in each point with the direction of the translation.

Lanczos, starting from a different point of view, arrives also at the similar conclusion that the Lorentz–Minkowski force implies that the electromagnetic field tensor is the infinitesimal generator of a Lorentz transformation.[8] Although these facts are known since long ago, nobody, as far as is known to the present author, has investigated the properties of the group generated by this tensor nor has used it to derive the intrinsic differential equations of the trajectories associated with it. This paper is devoted to the study of this particular group and the differential geometry of the curves associated with it, in Minkowski space. The findings that result from this study, however, because of their invariant nature, transcend the realm of electrodynamics.

The theory that studies the association between a given group and a specific differential geometry is of crucial importance to understand the point of view developed in this paper. As a reference, the method, due mainly to Cartan,[9] is reviewed succinctly in Appendix A.

II. THE ELECTRODYNAMIC GROUP

In this section a group of transformations of a continuous parameter is produced by considering the electromagnetic field tensor as the infinitesimal generator of the group. This group, defined in the space–time of special relativity (Lorentz space, also Minkowski space, $M^{3,1}$), has specific and unique characteristics. Some of these characteristics are investigated here.

Consider the transformations of the type[10]

$$\Lambda = I + \frac{\theta F}{1!} + \frac{(\theta F)^2}{2!} + \frac{(\theta F)^3}{3!} + \cdots = e^{\theta F}. \tag{8}$$

In (8), F is the electromagnetic field tensor and θ a continuous parameter. As F is antisymmetric, it is traceless. Therefore Λ is a Lorentz transformation: $\det \Lambda = e^{\theta \operatorname{tr} F} = 1$.

If θ is a real number and F is constant, the set represented by all Λ, for all possible values of θ has the structure of a group. I will refer to this group as the electrodynamic group.

To find the properties of the group, it is necessary to express F more specifically. We can express F using the electric and magnetic vector fields, **E** and **B**. However, this expression of F is not appropriate to investigate the differential geometry associated with the group because it is not possible to discriminate between the Coulomb and the radiation fields, as it is desirable. The expression of F using the retarded Liénard–Wiechert potentials allows for this discrimination very naturally. Of course, the results obtained using either of the two expressions of F are equivalent.

The well-known[11] expression for the electromagnetic field F produced by a trajectory $z(s)$ at the point x of space–time, calculated from the retarded Liénard–Wiechert potentials, is

$$F(x)^{\mu}{}_{\nu} = k\left(\frac{1}{\rho^3}\, d^{[\mu}\dot{z}_{\nu]}\right)_{\text{Coul}} + k\left[-\frac{(d^{\alpha}\ddot{z}_{\alpha})}{\rho^3}\, d^{[\mu}\dot{z}_{\nu]} + \frac{1}{\rho^2}\, d^{[\mu}\ddot{z}_{\nu]}\right]_{\text{rad}}. \tag{9}$$

The tensor F at the field point x depends exclusively on the retarded point of the trajectory, $z(s_0)$, that makes the distance $d = (x - z)$, a null distance: $d^2 = 0$. Here $\rho = d^{\alpha}\dot{z}_{\alpha}(s_0)$ and k is a positive constant. The first parentheses in (9) enclose the velocity fields, also known as the generalized Coulomb fields. The square brackets display the radiation or acceleration fields. They are the infinitesimal generators of the transformations of the group that we are developing.

To further investigate the properties of the electrodynamic group, we first insert (9) in (8). We find that, thanks to some recurrence relations in the powers of F, it is possible to express (8) in a more meaningful way. In fact, if we multiply F by itself n times, we easily find[12]

$$F^{2n} = \left(\frac{k}{\rho^2}\right)^{2(n-1)} F^2, \quad n = 1,2,3,... \tag{10}$$

and

$$F^{(2n+1)} = \left(\frac{k}{\rho^2}\right)^{2n} F, \quad n = 0,1,2,... . \tag{11}$$

Taking advantage of these relations we express the series expansion of the exponential, (8), as

$$\Lambda(x,\theta) = \left\{I - \frac{\rho^4}{k^2}\, F^2(x)\right\} + \left[\cosh\left(\frac{\theta k}{\rho^2}\right)\right]\left\{\frac{\rho^4}{k^2}\, F^2(x)\right\} + \left[\sinh\left(\frac{\theta k}{\rho^2}\right)\right]\left\{\frac{\rho^2}{k}\, F(x)\right\}. \tag{12}$$

The transformation $\Lambda(x,\theta)$, for a fixed value of θ, can now be understood as the simultaneous action of the three operators enclosed in braces. Each operator has a well-defined function in the transformation of a vector at the point x (the initial momentum, for example). The first produces the component of the transformed vector that remains unaffected by the transformation. The

second produces the component along the projection of the original vector on a three-plane. The third produces the component along a transversal direction on the same three-plane. We will investigate these transformations in more detail.

Λ operates on a vector having x as its point of origin. For a finite value of θ, Λ transforms this vector onto another vector of the same norm and of the same space. In other words, the transformation is a space–time rotation. In this respect, it may appear confusing when we operate below with Λ on a vector parallel to $\dot{z}(s_0)$. The fact that this vector has its origin at the point x is considered by the use of the null distance d that corresponds to x in the operators of Λ. The transformed vector, by the same reason, belongs also to this space. Nevertheless, this circumstance is recalled when necessarily changing s_0 to x, writing instead of $\dot{z}(s_0)$, $\dot{z}(x)$.

Let us now consider the Coulomb transformation by itself. The infinitesimal generator is

$$F_{\mathrm{Coul}}(x)=\frac{k}{\rho^3}\,(d^\mu \dot{z}_\nu - d_\nu \dot{z}^\mu). \tag{13}$$

This tensor produces the same recurrence relations as those found before, (10) and (11). The Coulomb transformation is then

$$\Lambda_{\mathrm{Coul}}(x,\theta)=e^{\theta F_{\mathrm{Coul}}(x)}=\left\{I-\frac{\rho^4}{k^2}\,F^2_{\mathrm{Coul}}(x)\right\}+\left[\cosh\left(\frac{\theta k}{\rho^2}\right)\right]\left\{\frac{\rho^4}{k^2}\,F^2_{\mathrm{Coul}}(x)\right\}$$
$$+\left[\sinh\left(\frac{\theta k}{\rho^2}\right)\right]\left\{\frac{\rho^2}{k}\,F_{\mathrm{Coul}}(x)\right\}. \tag{14}$$

It is convenient to express the three operators in braces in (14) with the aid of the retarded distance introduced by Synge[13] and defined, in his words, as follows: ''...from the point x drop the perpendicular x-b on the line of $\dot{z}(s_0)$. Then x-b is space-like, and has a length, say ρ. Let us call ρ the retarded distance of x from $z(s_0)$... .'' In the following, ρ will always be different from zero and from infinity. Let $\hat{r}$ denote a unit vector along the segment x-b, from the point b, on the line of $\dot{z}(s_0)$, to the field point x. The null vector d can now be expressed as

$$d^\mu=\rho(\hat{r}^\mu+\dot{z}^\mu), \tag{15}$$

with

$$\hat{r}^\alpha \dot{z}_\alpha=0, \quad \hat{r}^2=-1. \tag{16}$$

From (15) and (16) we have

$$d^\alpha \dot{z}_\alpha=\rho, \tag{17}$$

which justifies the notation used in (9) and,

$$d^\alpha \ddot{z}_\alpha=\rho(\hat{r}^\alpha \ddot{z}_\alpha)\equiv\rho\phi. \tag{18}$$

Using (15), the operators in the Coulomb transformation, (14), have the simple form

$$\frac{\rho^2}{k}\,F_{\mathrm{Coul}}(x)=(\hat{r}^\mu \dot{z}_\nu-\hat{r}_\nu \dot{z}^\mu), \tag{19}$$

and

$$\frac{\rho^4}{k^2}\,F^2_{\mathrm{Coul}}(x)=(-\hat{r}^\mu \hat{r}_\nu+\dot{z}^\mu \dot{z}_\nu). \tag{20}$$

It is clear from (19) and (20) that the transformation Λ_{Coul}, for a fixed and finite value of θ, produces a rotation on the $(\dot{z}^\mu(x)-\hat{r}^\mu)$-plane. This plane is timelike and cuts the light cone with vertex at x in the two lightlike directions: $\dot{z}^\mu(x)\pm\hat{r}^\mu$. The transformation leaves invariant the spacelike plane orthogonal to the plane of rotation (dual plane).

Consider at x a unit timelike vector, $\hat{v}$, on the plane of rotation

$$\hat{v}^{\mu}=a\hat{r}^{\mu}+b\dot{z}^{\mu}, \quad b^2-a^2=1, \quad b>0. \tag{21}$$

From (16), (19), and (20),

$$\left[I-\frac{\rho^4}{k^2}\,F^2_{\mathrm{Coul}}(x)\right]\hat{v}=0, \tag{22}$$

$$\frac{\rho^4}{k^2}\,F^2_{\mathrm{Coul}}(x)\hat{v}=\hat{v}, \tag{23}$$

and

$$\left[\frac{\rho^2}{k}\,F_{\mathrm{Coul}}(x)\right]\hat{v}=b\hat{r}^{\mu}+a\dot{z}^{\mu}. \tag{24}$$

We can now analyze the action of the operator Λ_{Coul} on the vector $\hat{v}$, as the sum of the actions of its three components, (22), (23), and (24). The operator in (23) projects the vector onto itself because $\hat{v}$ belongs to the $(\dot{z}^{\mu}-\hat{r}^{\mu})$-plane. The orientation of the transformed vector is the same as that of the original vector. Contrarily, the operator in (24) transforms timelike directions, as that of $\hat{v}$, into transversal directions. In fact, the transformed vector is spacelike: $(b\hat{r}^{\mu}+a\dot{z}^{\mu})^2=-1$. For a fixed and finite value of θ, $\Lambda_{\mathrm{Coul}}(x)$ rotates $\hat{v}$ from its original position towards the transversal direction by a hyperbolic angle that depends on the value of θ. For a continuous range of values of θ starting from zero, the end point or tip of the transformed vector describes a segment of a hyperbola drawn on the rotation plane, approaching asymptotically one of the two future null directions. In case the vector $\hat{v}$ has a component on the dual plane defined by the transformation, this component would remain unaffected, as seen from (22).

The radiation transformation seems to be more interesting, partly because it describes the dynamical process involved with the radiation, and partly because it deals with a peculiar type of Lorentz transformations. The infinitesimal generator is

$$F_{\mathrm{rad}}(x)=-\frac{k(d^{\alpha}\ddot{z}_{\alpha})}{\rho^3}\,(d^{\mu}\dot{z}_{\nu}-d_{\nu}\ddot{z}^{\mu})+\frac{k}{\rho^2}\,(d^{\mu}\ddot{z}_{\nu}-d_{\nu}\dot{z}^{\mu}). \tag{25}$$

The radiation generator is a nilpotent tensor of index three,[14]

$$F^3_{\mathrm{rad}}=0. \tag{26}$$

This property characterizes the uncommon type of Lorentz transformations known as intermediate.[15] The radiation transformation, using (18), is

$$\Lambda_{\mathrm{rad}}(x,\theta)\equiv e^{\theta F_{\mathrm{rad}}(x)}=I+\frac{\theta k}{\rho^2}\,d^{[\mu}(\ddot{z}-\phi\dot{z})_{\nu]}-\frac{\theta^2k^2}{2\rho^4}\,(\ddot{z}^2+\phi^2)d^{\mu}d_{\nu}. \tag{27}$$

This truncated series is a Lorentz transformation. To see this, one can take four orthogonal unit vectors at x and apply to them the transformation Λ_{rad}. It is easy to verify that the norm of each vector is preserved. Let us take, for example, the timelike vector of the tetrad to be parallel to $\dot{z}^{\mu}$. Then, using (15), (17), and (18),

$$[\Lambda_{\mathrm{rad}}(x,\theta)]\dot{z}^{\nu}(x)=\dot{z}^{\mu}-\frac{\theta k}{\rho}\,(\phi\hat{r}+\ddot{z})^{\mu}-\frac{\theta^2k^2}{2\rho^3}\,(\ddot{z}^2+\phi^2)d^{\mu}. \tag{28}$$

The square of this expression is one; therefore the norm is preserved. Similar results, except for the sign, are obtained for the two spatial unit vectors evidenced by (28), namely, $\hat{r}^{\mu}$ and $n(\phi\hat{r}+\ddot{z})^{\mu}$, ($n$ is a normalizing factor). The third spatial direction of the tetrad, being orthogonal to all the vectors that enter into the above expression, remains clearly invariant.

The locus of the tip of the transformed vector of $\dot{z}^\mu(x)$ by $\Lambda_{\rm rad}(x,\theta)$, as θ increases from zero, has two components, as shown by (28). The first is on a straight line that starts at the initial position of the $\dot{z}^\mu(x)$ tip, and runs parallel to the direction of the vector $-n(\phi\hat{r}+\ddot{z})^\mu$. Eventually, this component crosses the light cone with vertex at x. The second component, starting at the same initial point, is on a segment of parabola drawn on the timelike $(\dot{z}^\mu-\hat{r}^\mu)$-plane; it approaches the future direction defined by the null vector d^μ for large values of θ; it points to the future because, as is well known, the quantity $(-\ddot{z}^2-\phi^2)$ is positive definite.[16] In the composite curve described by the tip of the transformed vector, the first component dominates for small values of θ evolving towards the parabola for large values of θ. The curve is tridimensional, one of the dimensions being timelike.

Now we will find out how the individual transformations $\Lambda_{\rm Coul}$ and $\Lambda_{\rm rad}$ combine to produce $\Lambda_{\rm tot}$. Comparing $\Lambda_{\rm tot}$, (12), with $\Lambda_{\rm Coul}$, (14), we observe that the functional dependence on the parameter θ is the same for both transformations. From this we infer that θ, in the radiation transformation, has the fixed value ρ^2/k. In fact, it is easy to verify the relation

$$\Lambda_{\rm total}=e^{\theta F}\equiv e^{\theta(F_{\rm Coul}+F_{\rm rad})}=\Lambda_{\rm rad}^{-1}\Lambda_{\rm Coul}\Lambda_{\rm rad}, \tag{29}$$

where

$$\Lambda_{\rm rad}\equiv I+\frac{\rho^2}{k}F_{\rm rad}+\frac{1}{2}\frac{\rho^4}{k^2}F_{\rm rad}^2. \tag{30}$$

There are several ways to verify (29). One way is to carry out the operations indicated in the right-hand side using the commutator

$$[F_{\rm Coul},F_{\rm rad}]\equiv F_{\rm Coul}F_{\rm rad}-F_{\rm rad}F_{\rm Coul}=\frac{k}{\rho^2}F_{\rm rad}. \tag{31}$$

An easier way is by using the well-known relation between regular square matrices of the same order, A and B,

$$Be^AB^{-1}=e^{BAB^{-1}}. \tag{32}$$

Equation (29) helps us to understand how the transformations $\Lambda_{\rm Coul}$ and $\Lambda_{\rm rad}$ combine to produce $\Lambda_{\rm tot}$: While the Coulomb transformation produces the rotation on the timelike $(\dot{z}^\mu-\hat{r}^\mu)$-plane, the radiation transformation rotates this plane by a fixed amount around the axis given by the unit vector $\hat{e}^\mu_{(3)}$, towards the transversal direction given by minus $\hat{e}^\mu_{(2)}$, where

$$\hat{e}^\mu_{(2)}=n(\phi\hat{r}^\mu+\ddot{z}^\mu), \tag{33}$$

and

$$\hat{e}^\mu_{(3)}=\epsilon^{\mu\alpha\beta\gamma}\dot{z}_\alpha\hat{r}_\beta\hat{e}_{(2)\gamma}. \tag{34}$$

Finally, we will see that the electrodynamic group belongs to the restricted inhomogeneous Lorentz group, although this is already obvious. Let d be the null vector used above; $u(s_0)$ and $v(x)$ timelike vectors such that

$$u^\mu=d^\mu+v^\mu. \tag{35}$$

The null vector d is defined by the choice of the field point x and the given world line $z(s)$ through the retarded condition $d^2=0$. The point x represents any one of the field points. Then, the tips of the vectors u and d may be placed at any two points of the field. The norm of the difference, $u-d$ is, by (35), the norm of the vector $v(x)$, which is preserved because $\Lambda(x)$ is a Lorentz transformation. Therefore the electrodynamic group, in its Liénard–Wiechert potentials expression, belongs to the inhomogeneous Lorentz group. This is the group of transformations that preserves the distance between any two points of the space–time.

The transformations of the electrodynamic group preserve also the orientation of the vectors defined at x. Consider the transformation Λ_{tot}, (12). For the value zero of the parameter θ, the transformation is the identity; this clearly preserves the orientation. We can go from the identity to any other transformation of the group by varying continuously the parameter θ. Then, evidently, all the transformations of the group preserve the orientation of the vector space at x. The electrodynamic group, then, belongs to the restricted inhomogeneous Lorentz group.

III. THE INTRINSIC DIFFERENTIAL EQUATIONS OF THE CURVES ASSOCIATED WITH THE ELECTRODYNAMIC GROUP

Here we obtain the intrinsic differential equations of the curves associated with the electrodynamic group (characteristic curves) at an arbitrary point x of the space–time. They will be referred to the most representative tetrad (orthonormal set of four four-vectors) appearing in the expression of the electromagnetic tensor at that point. This tetrad qualifies as a tetrad of the group, as described below. Later, the group and the intrinsic differential equations will be expressed with reference to a tetrad attached intrinsically to the characteristic curve.

The electrodynamic group, as developed in the previous section, is a one parameter group. Consequently, the structure theorem reviewed in Appendix A applies to this group. According to this theorem, the components of the infinitesimal displacement of a tetrad of the group will depend on one parameter whatsoever. One can choose this parameter to be the arclength (proper time) of a curve associated with the group and going through the point x: One should remind, however, that the condition $d^2=0$ links the infinitesimal displacement of the point x to the corresponding one on the trajectory at the point $z(s_0)$.

If we observe the vectors appearing in the expression of the electromagnetic field tensor at x, (19) and (28), we identify the following basic tetrad

$$\begin{aligned}
\hat{e}^{\mu}_{(0)} &= \dot{z}^{\mu}, \\
\hat{e}^{\mu}_{(1)} &= \hat{r}^{\mu}, \\
\hat{e}^{\mu}_{(2)} &= n(\phi\hat{r}+\ddot{z})^{\mu}, \\
\hat{e}^{\mu}_{(3)} &= \epsilon^{\mu\alpha\beta\gamma}\hat{e}_{(0)\alpha}\ \hat{e}_{(1)\beta}\ \hat{e}_{(2)\gamma}.
\end{aligned} \tag{36}$$

The value n is a normalizing factor at the point x. The tetrad above is a tetrad of the group, according to this definition: A tetrad $T_0 \equiv (\hat{e}_{(0)}, \hat{e}_{(1)}, \hat{e}_{(2)}, \hat{e}_{(3)})$ is a tetrad of the group if for every transformation Λ_a of the group, with $\Lambda_a \neq I$, we have $\Lambda_a T_0 \neq T_0$.

Let us denote by T_0 the above tetrad of the group, (36). The components of the infinitesimal displacement of the tetrad, accounted from the field point x, are obtained immediately using (19) and (25): $dT_0 = (F_{\text{Coul}} + F_{\text{rad}})T_0 d\theta$, or, explicitly,

$$\frac{dx}{d\theta} \equiv \hat{e}_{(0)}, \tag{37}$$

$$\frac{d\hat{e}_{(0)}}{kd\theta} = \frac{1}{\rho^2}\,\hat{e}_{(1)} - \frac{\sqrt{-\ddot{z}^2-\phi^2}}{\rho}\,\hat{e}_{(2)}, \tag{38}$$

$$\frac{d\hat{e}_{(1)}}{kd\theta} = \frac{1}{\rho^2}\,\hat{e}_{(0)} + \frac{\sqrt{-\ddot{z}^2-\phi^2}}{\rho}\,\hat{e}_{(2)}, \tag{39}$$

$$\frac{d\hat{e}_{(2)}}{kd\theta} = -\frac{\sqrt{-\ddot{z}^2-\phi^2}}{\rho}\,\hat{e}_{(0)} - \frac{\sqrt{-\ddot{z}^2-\phi^2}}{\rho}\,\hat{e}_{(1)}, \tag{40}$$

$$\frac{d\hat{e}_{(3)}}{kd\theta} = 0. \tag{41}$$

This is the system of intrinsic differential equations that the characteristic curves of the electromagnetic field satisfy. Their relation to the electrodynamic group of rotations is essentially the same as that of the Frenet–Serret equations to the Euclidean group of rotations. The invariant factors may be expressed in different forms than those appearing above. Their expression depends on the tetrad of reference. At a given point x of the field, each invariant represents the norm of a vector at that point. We may denote these invariant factors, the curvature and the hypertorsion, by σ and η, respectively,

$$\sigma(x)=\frac{k}{\rho^2},$$
$$\eta(x)=\frac{k\sqrt{-\ddot{z}^2-\phi^2}}{\rho}. \tag{42}$$

Together with (37), these are the natural equations of the characteristic curves.

Using (37) and (42), Eqs. (38)–(41) can be recast into the more general form

$$\frac{d\hat{e}_{(0)}}{d\theta}= \quad \sigma\hat{e}_{(1)} \quad -\eta\hat{e}_{(2)}, \tag{43}$$
$$\frac{d\hat{e}_{(1)}}{d\theta}=\sigma\hat{e}_{(0)} \quad +\eta\hat{e}_{(2)}, \tag{44}$$
$$\frac{d\hat{e}_{(2)}}{d\theta}=-\eta\hat{e}_{(0)} \quad -\eta\hat{e}_{(1)}, \tag{45}$$
$$\frac{d\hat{e}_{(3)}}{d\theta}=0. \tag{46}$$

The expression of the invariants σ and η, depend on the specific realization of the group. The intrinsic equations, (43)–(46), which now refer to the relevant tetrad of the specific realization, are the same for all the realizations. These equations enclose all the essential information about the space–time curves associated with the electrodynamic group. As in the case of ordinary differential geometry, given the natural equations (42), with $\eta \geqslant 0$, there is, by the fundamental condition of equality,[17] one and only one space–time curve having these functions as its curvature and hypertorsion and having its tetrad oriented in a given form at one space–time point (satisfying the initial conditions). According to (46), the infinitesimal displacement of the tetrad along the curve is such that the vector $\hat{e}_{(3)}$, remains invariant. That is, the infinitesimal displacement of the tetrad is a rotation around the axis that contains this vector. Considering in (43) and (44) only those terms that contain σ, we see that σ represents the infinitesimal angle of a rotation on the timelike plane generated by the first two vectors of the tetrad. Finally, considering in (43), (44), and (45) only the terms that contain η, we find the infinitesimal characteristics of an intermediate transformation: The infinitesimal rotation of angle $(-\eta)$ on the timelike plane generated by the vectors $\hat{e}_{(0)}$ and $\hat{e}_{(2)}$ combined with the infinitesimal rotation of angle η on the space-like plane generated by the vectors $\hat{e}_{(1)}$ and $\hat{e}_{(2)}$.

Our next objective is to duplicate, or to realize, the electrodynamic group using a proper tetrad. A proper tetrad is the space–time version of the Frenet triad.[18] It is a tetrad attached intrinsically to a four-dimensional curve by means of a geometrical construction. Let us consider a curve parametrized by the arclength: $x=x(s)$. At a generic point of the curve, the first two vectors of the proper tetrad can be chosen along the tangent and along the first normal defined by the Euclidean group of rotations, (6), with their usual geometrical constructions,

$$\hat{e}^{\mu}_{(0)}\equiv v^{\mu}\equiv\frac{dx^{\mu}}{ds}, \tag{47}$$

$$\hat{e}^{\mu}_{(1)} \equiv \hat{a}^{\mu} = \frac{a^{\mu}}{|a|}. \tag{48}$$

As s is the arclength, the vector v^{μ} is of unit length and orthogonal to a^{μ}. Going back to the tetrad of the group, we see by (46) that the fourth vector of the tetrad remains invariant along the curve. In particular it remains invariant under the transformation that takes the vector $l=(1,0,0,0)$ onto the velocity vector v^{μ}, and vice versa. Then, the last two vectors of the proper tetrad can be defined as

$$\hat{e}^{\mu}_{(3)} \equiv s^{\mu} \equiv \epsilon^{\mu\nu\alpha\beta} l_{\nu} \hat{e}_{(1)\alpha} \hat{e}_{(0)\beta}, \tag{49}$$

$$\hat{e}^{\mu}_{(2)} \equiv t^{\mu} \equiv \epsilon^{\mu\nu\alpha\beta} \hat{e}_{(3)\nu} \hat{e}_{(1)\alpha} \hat{e}_{(0)\beta}. \tag{50}$$

It is easy to verify that the vector s remains invariant under the Lorentz transformations, $\underset{l\to v}{\Lambda}$ and $\underset{v\to l}{\Lambda}$.

Let α be the angle measured at the laboratory between the velocity **v** and the acceleration **a**, and v the norm of **v**. Then, the vectors s and t, (49) and (50), are given by

$$t^{\mu} = \frac{\gamma}{(\sin\alpha)\sqrt{\gamma^2\cos^2\alpha + \sin^2\alpha}} \left\{ \frac{v}{c}\sin^2\alpha, \hat{\mathbf{v}} - \hat{\mathbf{a}}\cos\alpha \right\}, \tag{51}$$

$$s^{\mu} = \frac{1}{\sqrt{a^2 v^2 \sin^2\alpha}} \{0, \mathbf{v}\times\mathbf{a}\} = \frac{1}{\sin\alpha} \{0, \hat{\mathbf{v}}\times\hat{\mathbf{a}}\}. \tag{52}$$

Let us denote by Ω the infinitesimal generators of the electrodynamic group when they are represented with elements of the proper tetrad $(v,\hat{a},t,s)$. From (19), (25), and (42) we have

$$\Omega_{\text{Coul}} = \sigma(\theta) \hat{a}^{[\mu} v_{\nu]}, \tag{53}$$

$$\Omega_{\text{rad}} = \eta(\sigma,\theta)(v+\hat{a})^{[\mu} t_{\nu]}. \tag{54}$$

Using the above infinitesimal generators in the place of F_{Coul} and F_{rad} as in Sec. II, it is possible to reproduce all the properties of the electrodynamic group. They reproduce also the differential equations (43)–(46). According to the fundamental theorem of equality, the curve described by means of the tetrad of the group (36) at the point x will be the same curve as that described by means of the proper tetrad, (47)–(50), if, having the same initial orientation the two tetrads ($\theta = s = 0$), their respective components of the infinitesimal displacements, σ and η, have the same values. These values are given in terms of the field by Eqs. (42). The meaning of the invariants with respect to the proper tetrad will be given in the next section.

The tetrad of the group (36), and the proper tetrad, (47)–(50), both belong to the same characteristic curve of the field. A proper tetrad can also be attached to an electron located in the field. If the initial velocity of the electron coincides with the timelike vector of the proper tetrad of the characteristic curve, then the trajectory of the electron will be the characteristic curve. If not, only the components of the proper tetrad of the electron lying on the rotation planes defined by the field will obey the differential equations.

IV. THE RELATIVISTIC FORCE ON THE ELECTRON

The rate of change of the velocity, that is, the acceleration in the inertial system of reference where the electromagnetic tensor is defined, is

$$\ddot{x} = \frac{d\hat{e}_{(0)}}{d\theta} = \sigma\hat{e}_{(1)} - \eta\hat{e}_{(2)}. \tag{43$'$}$$

This acceleration has two terms. Employing the terminology of classical kinematics, let us refer to the first term as the apparent acceleration and to the second as the radiation reaction acceleration. This suggests that the radiation reaction term is a kind of fictitious acceleration which can be cancelled out by a transformation of coordinates. This fact is expressed by (29). The length of the apparent acceleration is $-\sigma$. The invariant σ, referred to the tetrad of the group T_0, (36), is given by (42). If (43) refers to the proper tetrad of the characteristic curve at the same point in which we considered the tetrad T_0, then σ is also proportional to the inverse radius of curvature of the hyperbola of curvature as described by Minkowski in his famous address ''Space and Time.'' [19] The radius of curvature and the acceleration have the same direction and sense. Certainly, the value of σ, being an invariant, is the same referred to either one of the two tetrads. As the apparent acceleration is the acceleration that we observe at the laboratory in the absence of radiation, we may denote it by a^μ.

The radiation reaction acceleration, of length $-\eta$, arises from the relativistic considerations of the Liénard–Wiechert field acting on the tetrad of the group T_0, Sec. III. For small values of θ, it has the direction of a straight line (28): Its length can be approximated relating the hyperbola described by the tip of the velocity vector with the hyperbola described by the tip of the radiation acceleration vector as θ varies. We will consider this approximation later.

Referred to the proper tetrad, the radiation acceleration has the direction of the vector t. The direction of t and the magnitude $-\eta$ characterize this fictitious acceleration completely, however, it is convenient to obtain an expression for it in terms of successive differentials of the curve with respect to the arclength. We first notice that

$$\dot{x}^\mu = \hat{e}^\mu_{(0)} \tag{55}$$

and

$$a^\mu = \sigma e^\mu_{(1)}. \tag{56}$$

To replace the vector $\hat{e}_{(2)}$ in (43′) with differentials of the curve, we differentiate (43′) with respect to the arclength, using (44) and (45) and taking σ and η as constants. We find

$$\ddot{x}^\mu = \sigma(\sigma\hat{e}_{(0)} + \eta\hat{e}_{(2)}) - \eta(-\eta\hat{e}_{(0)} - \eta\hat{e}_{(1)}) = (\sigma^2+\eta^2)\dot{x}^\mu + \frac{\eta^2}{\sigma}\, a^\mu + \sigma\eta\hat{e}^\mu_{(2)}. \tag{57}$$

Using this result in (43′),

$$\frac{d\dot{x}}{d\theta} = \left(1+\frac{\eta^2}{\sigma^2}\right)a - \frac{1}{\sigma}\dddot{x} + \left(\frac{\sigma^2+\eta^2}{\sigma}\right)\dot{x}. \tag{58}$$

Comparing with (43′), we first observe that the apparent acceleration is now η^2/σ^2 larger. The second and third terms considered together, resemble the Abraham–Lorentz–Dirac (ALD) acceleration. We find the third-order derivative of the position and the necessary component along the velocity. Noting from (57) that $\ddot{x}^\alpha\dot{x}_\alpha = \sigma^2+\eta^2$ we find that, in fact, it is orthogonal to the velocity, a requirement fulfilled by an acceleration in general and by the ALD acceleration in particular. Except for the coefficients, they are equal.

The relativistic force, consequently, may be expressed in two equivalent forms, using either (43) or (58). If we use (1), (4), (43), and (50)

$$f = mc^2\sigma\hat{e}_{(1)} - mc^2\eta\hat{e}_{(2)} = mc^2 a - mc^2\eta t = f_a - \eta' t. \tag{59}$$

The first term is the apparent force f_a. The second term is the radiation reaction force. There is more than a simple similarity between the last force and the so-called fictitious forces. Fictitious forces, such as the Coriolis force, the centrifugal force, etc., are produced when the observer frame of reference is rotated with respect to a stationary frame where the force is initially considered. Both the fictitious forces and the radiation reaction force are produced by transformations of the Lorentz group. While the Coriolis force, for example, is produced by a rotation of the space, the radiation reaction force is produced by an intermediate transformation. They have a similar origin.

Let us now consider the equations of motion. Let an electron of known initial position (x) and velocity ($\dot{x}$) be subject to an external electromagnetic field whatsoever, F. The relativistic force on the electron is given by the Lorentz–Minkowski force:

$$f^{\mu}=eF^{\mu}{}_{\alpha}\dot{x}^{\alpha}. \tag{60}$$

The equation of motion, using the expression (59) for the force, is then

$$f_a{}^{\mu}-\eta' t^{\mu}=eF^{\mu}{}_{\alpha}\dot{x}^{\alpha}. \tag{61}$$

The field F at the space–time point x gives the values of invariants σ and η, that is, the natural equations of the curve: The trajectory is therefore defined from the initial conditions and the equation of motion is satisfactory in this sense. If we set the external field F equal to zero in (61), we have $a=\eta t=0$, and, by (43), we find that the momentum is a constant.

If we use the acceleration in the form (58) instead, we obtain the force

$$f=\left(1+\frac{\eta^2}{\sigma^2}\right)f_a-\frac{mc^2}{\sigma}\dddot{x}+mc^2\left(\frac{\sigma^2+\eta^2}{\sigma}\right)\dot{x}, \tag{62}$$

and the equation of motion

$$mc^2\left(1+\frac{\eta^2}{\sigma^2}\right)a-\frac{mc^2}{\sigma}\dddot{x}+mc^2\left(\frac{\sigma^2+\eta^2}{\sigma}\right)\dot{x}=eF\dot{x}. \tag{63}$$

Equations (61) and (63) are different expressions of the same equation of motion. Suppose now that we set $a=\ddot{x}$ in (63), introducing an error according to (43). Then (63) becomes the ALD equation, except for the coefficients. For an external field $F=0$, we also find self-accelerated solutions; imposing asymptotic conditions, we find preacceleration. In short, we find the nonlocality in time, characteristic of this equation of motion. Although the situation may seem to be more of the same, we have now some new information at our disposal: The apparent acceleration a is $\ddot{x}$ only if $\eta=0$. Then, the error that we introduced by making $a=\ddot{x}$ depends only on the radiation field and it is of the same order of magnitude as the radiation reaction force.

Now, consider (29). From it see that it is mathematically acceptable to approximate the real trajectory beginning from the trajectory due to the generalized Coulomb field alone, then, correcting it by switching on and off the radiation transformation for fixed values of the parameter. The concept introduced by Caldirola,[20] that the momentum of the electron, under the action of an external force, can only vary after a fixed interval of its proper time has elapsed finds here additional support: He calls this fundamental interval the chronon. This concept, which has been refined over the years,[21] is expressed relativistically by means of a differential-difference equation. This is in agreement with (30), where it is shown that the radiation transformation depends on a fixed value of the parameter θ.

We will now obtain an approximation of the length of the radiation reaction vector, the hypertorsion, in terms of the curvature. To avoid a long digression about the question of units, which was saved with the use of the constant k in the formulas, it will only be mentioned that the natural units of the Liénard–Wiechert field tensor (i.e., deleting all the proportionality constants) are, in our metric, those of an inverse squared length. In contrast, it has the dimensions of force per unit charge in the Lorentz–Minkowski force (60) (Heaviside–Lorentz units). If in (63) we express F in its natural units and absorb in the proportionality constant the factor mc^2 that is on the left, this constant becomes the classical electron radius. In that case the equation of motion (61) would be

$$a^{\mu}-\eta t^{\mu}=r_0F^{\mu}{}_{\alpha}\dot{x}^{\alpha}. \tag{64}$$

Using this proportionality constant in the tensor field to obtain the energy momentum tensor, instead of (7) one obtains

$$\dot{p}^{\mu}=r_0^2\left(\frac{8\pi}{3}\right)|\ddot{x}|^2\dot{x}^{\mu}. \tag{65}$$

The invariant η can be approximated with a hyperbola as follows: Let the direction of the radius of curvature be $-t$. Under the continuous action of the radiation field, the velocity vector $\dot{x}=\gamma(1,\mathbf{v}/c)$ will evolve towards the vector $l=(1,0,0,0)$. It will be translated and rotated. Refer the rotation to the base point of the velocity vector. The tip of the velocity vector for increasing values of the proper time will describe a hyperbola in the interior of the light cone with vertex at the base point of the vector. At the same time, the vector t, preserving its orthogonality with respect to the velocity vector, will describe an external hyperbola. The rotation of the velocity vector corresponds to the transformation

$$\Lambda\dot{x}=\dot{x}\ \cosh\ \vartheta-t\ \sinh\ \vartheta. \tag{66}$$

In particular, for the values

$$\cosh\ \vartheta=\gamma,$$

$$\sinh\ \vartheta=\gamma\,\frac{v}{c}, \tag{67}$$

Λ in (66) transforms the velocity vector $\dot{x}$ onto the laboratory vector l. Considering terms up to the second order in the series expansion of (66), we have

$$\Lambda\dot{x}\approx\dot{x}\left(1+\frac{1}{2!}\,\vartheta^2+\cdots\right)-t(\vartheta+\cdots). \tag{68}$$

We now identify the second term in the first parentheses of (68) with (65). This is not arbitrary because the energy momentum tensor is the projection operator in the transformation of the total electromagnetic field, (12). In particular, the energy momentum tensor corresponding to the so-called ''wave zone,'' is the last term in (27). This is the term that is used in the conventional treatment of radiation reaction to obtain (7).[22] Based on the commutator (31), but referred now to the proper tetrad, we set $\eta=|\ddot{x}|\,\vartheta$ and find

$$\eta=4\,\sqrt{\frac{2\,\pi}{3}}\,r_0|\ddot{x}|^2. \tag{69}$$

Let us now consider the nonrelativistic limit. From (63), with the modification that considers the apparent acceleration as $\ddot{x}$, one is lead to an equation of motion in ordinary space with the radiation reaction term $f_{\text{react}}\propto -d^3\mathbf{x}/dt^3$.

Contrarily, the radiation reaction force given by Eq. (61) reduces in the nonrelativistic limit to a typical frictional force. Consider a nonrelativistic electron moving along a characteristic curve such that $\mathbf{v}\cdot\mathbf{a}=0$. Then, from (51) we find

$$-t^\mu \xrightarrow[v\ll c]{} -\hat{\mathbf{v}}. \tag{70}$$

To end this section, (61) will be reduced to the case of a Coulomb force using the Liénard–Wiechert expression of F, (9) and (19). In this case: $\dot{z}=\dot{x}=(1,0,0,0)$ and $\ddot{z}=0$. First, we notice that the vectors s and t, by (49) and (50), are zero. Then we are left with

$$mc^2\ddot{x}^\mu=\frac{e^2}{4\,\pi}\,\frac{1}{\rho^2}\,(\hat{r}^\mu\dot{z}_\alpha-\hat{r}_\alpha\dot{z}^\mu)\dot{x}^\alpha. \tag{71}$$

In this case we find that $\hat{r}^\mu=(0,\hat{\mathbf{r}})$ and that the invariant ρ is the distance between the two electrons. From (6) and (16) we find the Coulomb law,

$$m\mathbf{a}=\frac{e^2}{4\,\pi r^2}\,\hat{\mathbf{r}}. \tag{72}$$

The Coulomb law expresses physically that the invariant σ given in terms of the tetrad of the group by (42) has the same value as the invariant given in terms of the proper tetrad, that is, as the norm of the ordinary acceleration.

V. CONCLUSIONS

The point of view of electrodynamics introduced in this paper leads one to consider the radiation reaction force as a fictitious force: as such, it can be canceled out by a transformation of the coordinates. From the mathematical viewpoint, there is a similarity with the case dealt with in the Larmor theorem, where the magnetic Lorentz force due to a constant magnetic field can be canceled out by the Coriolis force produced by a rotation of the space. The Abraham force or ALD force appears, except for the coefficients, as a constituent of this fictitious force: The ordinary force, however, becomes an apparent force and is modified when the force of radiation is also considered. It follows from this point of view that the source of the problems faced by the ALD equation resides not in the ALD force but in considering the apparent force as the total force.

The mathematical viewpoint employed in the paper asserts that the world-line curve, the group of transformations, and the intrinsic differential equations are different aspects of the same object: This also has an interesting consequence in the relativistic dynamics of an uncharged body. The electrodynamic group is contained in the restricted inhomogeneous Lorentz group and the characteristic curves associated with the electrodynamic group are embedded in the Minkowski space. We found that the hypertorsion of a space–time trajectory is accompanied by a radiation field and by a reduction of the momentum. These three characteristics occur simultaneously because they are manifestations of one and the same mathematical fact. This statement, which depends on the geometry and not on the charge, necessarily applies also to any space–time trajectory (the curve followed by any physical body, charged or uncharged). If this were not so, it would be possible to show that there is a contradiction between the laws that define the geometry associated with the electrodynamic group in Minkowski space, as the part contained in the whole, and the laws that define the geometry associated with the restricted inhomogeneous Lorentz group also in Minkowski space, as the whole that contains the part. Physically, this implies that if we constrain the motion of a massive, neutral body, to follow a trajectory in which the trajectory of its parts have a hypertorsion different from zero, we should expect, first, that a radiation field is produced and, second, the consequent reduction of the momentum. In other words, any body following a space–time trajectory with hypertorsion radiates.

ACKNOWLEDGMENTS

I want to thank Dr. Basil J. Hiley for his help and sensible guidance. I also thank Professor Clive W. Kilmister for valuable suggestions to a previous version of this paper. I am indebted to Dr. Luis de la Peña for accurate comments that improved the paper in general and, also, for his encouragement and help to proceed with the investigation introduced here. I acknowledge, with thanks, an observation made by the referee on a deficiency, now corrected, that motivated the rewriting of Sec. IV with considerable improvement.

APPENDIX A: THE CARTAN METHOD OF THE MOVING FRAME (REF. 23)

In Euclidean space, consider a set of three orthogonal ordered unit length vectors with common origin o. We will call any set like this a triad. Let $\mathbf{m}$ be a fixed point in the frame of reference defined by the triad and of coordinates (a,b,c). Now let the triad move continuously, but otherwise arbitrarily, from a given initial position in the frame $(\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3)$, and let its motion depend on n parameters, s_n. The point $\mathbf{m}$ moves fixed to the triad, i.e., its coordinates remain constant in that frame. The motion of the triad defines in the space a continuum of triads: the family E of triads. The most general infinitesimal variation in the coordinates of the point $\mathbf{m}$ due to an infinitesimal increment in all the parameters, referred this variation to the initial position of the triad, is given by the equations

$$
\begin{aligned}
d\mathbf{m} &= d\mathbf{o} + a\,d\hat{\mathbf{e}}_1 + b\,d\hat{\mathbf{e}}_2 + c\,d\hat{\mathbf{e}}_3,\\
d\mathbf{o} &= \omega_1\hat{\mathbf{e}}_1 + \omega_2\hat{\mathbf{e}}_2 + \omega_3\hat{\mathbf{e}}_3,\\
d\hat{\mathbf{e}}_1 &= \omega_{12}\hat{\mathbf{e}}_2 + \omega_{13}\hat{\mathbf{e}}_3,\\
d\hat{\mathbf{e}}_2 &= -\omega_{12}\hat{\mathbf{e}}_1 + \omega_{23}\hat{\mathbf{e}}_3,\\
d\hat{\mathbf{e}}_3 &= -\omega_{13}\hat{\mathbf{e}}_1 - \omega_{23}\hat{\mathbf{e}}_2.
\end{aligned}
\tag{A1}
$$

The ω's are known as connection forms. To get a quick idea of the meaning of these equations, suppose that all the forms are null except ω_1. In this case, Eqs. (A1) express that the infinitesimal displacement suffered by the point $\mathbf{m}$ is a translation parallel to $\hat{\mathbf{e}}_1$ and of infinitesimal magnitude ω_1. Suppose now that all the connection forms are null except ω_{12}. In this case, the motion of $\mathbf{m}$ is a rotation around the $\hat{\mathbf{e}}_3$ axis and of infinitesimal angle ω_{12}. Equations (A1) express, in the most general case, that the infinitesimal displacement of the point $\mathbf{m}$, due to infinitesimal increments in the n parameters, is a superposition of three infinitesimal translations parallel to the axes $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$, respectively, and three infinitesimal rotations around those same axes.

Let the point $\mathbf{m}$ coincide with the origin of the moving triad; Eqs. (A1) describe, in this case, the infinitesimal displacement of the triad itself.

Consider now two families of triads E and E' depending on the same number of parameters. Suppose that we can define a bijective mapping between the two families. This mapping defines homologous triads in E and E'. The fundamental condition of equality between the two families is expressed as follows: It is possible to superpose all homologous triads with the same transformation, if and only if, the bijective mapping produces components ω' of the infinitesimal change of a triad in E' equal to the components ω of the infinitesimal change of its homologous triad in E.

The problem simplifies substantially when the motion of the triad that generates the family is ruled by only one parameter. In this case we find the structure theorem: The components ω of the infinitesimal displacement of the moving triad, if its motion depends on only one parameter, are subject to no conditions of structure. In other words, the components ω depend on one variable whatsoever. This case applies directly to the study of the Euclidean space curves.

Let C be an oriented curve parametrized by the arclength, $\mathbf{x}=\mathbf{x}(s)$, with no inflexion points. At each point of the curve choose the triad satisfying the following: the origin of the triad coincides with that point of the curve, $\mathbf{o}=\mathbf{x}$; the first vector $\hat{\mathbf{e}}_1$ lies on the positive side of the tangent; the second vector $\hat{\mathbf{e}}_2$ lies on the principal normal pointing to the interior of the curve; the third vector $\hat{\mathbf{e}}_3$, the binormal, is the cross product $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2$. This triad is known as the Frenet triad corresponding to that point of the curve. As the origin of the Frenet triad moves continuously along the positive direction of the arc, starting from an arbitrary initial point, the Frenet triad generates a family of triads.

The translation $\omega_1 = ds$, $\omega_2 = \omega_3 = 0$, gives in (A1) the usual definition of the tangent and of the arclength. The rotation $\omega_{13}=0$ gives the orientation of the principal normal. The other two rotations have the form: $\omega_{12} = \rho\, ds$, and $\omega_{23} = \kappa\, ds$. With these conditions Eqs. (A1) give the Frenet–Serret formulas. It follows from (A1) and the structure theorem that the curvature ρ and the torsion κ are positive functions of the parameter s.

The Frenet–Serret formulas can be put in the form

$$
d\mathbf{x} = ds\,\hat{\mathbf{e}}_1, \tag{A2}
$$

$$
d\hat{\mathbf{e}}_i = ds\,\Omega\hat{\mathbf{e}}_i, \tag{A3}
$$

$$
d\hat{\mathbf{e}}_{\mathrm{rot}} = ds\,\Omega\hat{\mathbf{e}}_{\mathrm{rot}} = ds\,\hat{\mathbf{e}}_{\mathrm{rot}}. \tag{A4}
$$

The operator Ω produces an infinitesimal rotation of the triad around the axis $\hat{\mathbf{e}}_{\mathrm{rot}}$, as shown by (A4). The velocity of the point $\mathbf{x}$ is defined by the translation (A2). The rate of change of the velocity is given by the infinitesimal rotation $\Omega\hat{\mathbf{e}}_1$. This is a rotation of the Euclidean group.

APPENDIX B: *F* IN TERMS OF THE LIÉNARD–WIECHERT POTENTIALS (REF. 24)

This is an outline of how to obtain the Liénard–Wiechert expression of the electromagnetic field tensor F. A field point is any point of the space–time, except those laying on the electron world line which is the source of the field. The coordinates of a field point will be represented by x^μ, or by x. The coordinates of the world line will be represented by z^μ, or by z. We are given, in parametric form, the four equations that define the world line of the point-electron, $z^\mu = z^\mu(s)$. The parameter (s) is, essentially, the proper time of the electron. The solution of the inhomogeneous wave equation, using this world line as the independent term, leads to the Liénard–Wiechert four potentials

$$A^\mu(x) = k \left. \frac{\dot{z}^\mu}{(x-z)^\alpha \dot{z}_\alpha} \right|_{(x-z)^2=0}. \tag{B1}$$

k is a positive constant. The characteristic property of these potentials is contained in the condition that defines the null vector

$$d^\mu \equiv (x^\mu - z^\mu), \quad d^2 = 0. \tag{B2}$$

The null distance d^μ is a function of x and z. To each field point x, there are two points on the world line $z(s)$ such that the condition $d^2=0$ is satisfied. Of the two points on the world line, we will take only the point $z(s_0)$ corresponding to a positive time component of d. With respect to the observation field point x, the potential from $z(s_0)$ is known as the retarded or causal solution. The condition $d^2=0$ allows us to express the time at the point x in terms of the time at $z(s_0)$ and the spatial distance between x and $z(s_0)$. The invariant in the denominator of (B1) has only one value for the retarded solution; It will be denoted by ρ. We have

$$A^\mu(x) = k \frac{\dot{z}^\mu(s_0)}{\rho}, \tag{B3}$$

with

$$\rho \equiv (x - z(s_0))^\alpha \dot{z}_\alpha(s_0). \tag{B4}$$

The condition $d^2=0$ embodies an essential property of the field. From this property it follows that to an infinitesimal increase in the coordinates of the field point x, it corresponds an infinitesimal increase in the coordinates on the world line z, and, consequently, also in (s). Then, to differentiate the potentials A^μ with respect to the field variables x^μ, we have to preserve the condition $d^2=0$. From (B2),

$$\partial^\mu d_\alpha d^\alpha = 2 d^\alpha (\delta^\mu_\alpha - \dot{z}_\alpha \partial^\mu s) = 0. \tag{B5}$$

Then

$$\partial^\mu s = \frac{d^\mu}{\rho}. \tag{B6}$$

The tensor F is defined by

$$F^\mu{}_\nu \equiv \partial^\mu A_\nu - \partial_\nu A^\mu \equiv \partial^{[\mu} A_{\nu]}. \tag{B7}$$

Using (B1), (B6), and (B7) it is easy to obtain

$$F(x)^\mu{}_\nu = k \left(\frac{1}{\rho^3} d^{[\mu} \dot{z}_{\nu]} \right)_{\text{Coul}} + k \left(-\frac{\Phi}{\rho^3} d^{[\mu} \dot{z}_{\nu]} + \frac{1}{\rho^2} d^{[\mu} \ddot{z}_{\nu]} \right)_{\text{rad}}, \tag{B8}$$

where

$$\Phi \equiv d^\alpha \ddot{z}_\alpha. \tag{B9}$$

APPENDIX C: THE RECURRENCE IN THE POWERS OF *F*

Calculating the square of (B8), using $d^\alpha \dot{z}_\alpha = \rho$, $d^\alpha \ddot{z}_\alpha = \Phi$, and $\ddot{z}^\alpha \dot{z}_\alpha = 0$, we have

$$\frac{F^2}{k^2} = \left[\frac{(1-\Phi)}{\rho^3}(d^\mu \dot{z}^\alpha - d^\alpha \dot{z}^\mu) + \frac{1}{\rho^2}(d^\mu \ddot{z}^\alpha - d^\alpha \ddot{z}^\mu)\right]\left[\frac{1-\Phi}{\rho^3}(d_\alpha \dot{z}_\nu - d_\nu \dot{z}_\alpha) + \frac{1}{\rho^2}(d_\alpha \ddot{z}_\nu - d_\nu \ddot{z}_\alpha)\right]$$

$$= \frac{(1-\Phi)^2}{\rho^6}(\rho d^\mu \dot{z}_\nu - d^\mu d_\nu + \rho d_\nu \dot{z}^\mu) + \frac{(1-\Phi)}{\rho^5}(\rho d^\mu \ddot{z}_\nu + \Phi d_\nu \dot{z}^\mu) + \frac{(1-\Phi)}{\rho^5}(\Phi d^\mu \dot{z}_\nu + d_\nu \ddot{z}^\mu)$$

$$+ \frac{1}{\rho^4}(\Phi d^\mu \ddot{z}_\nu - \ddot{z}^2 d^\mu d_\nu + \Phi \ddot{z}^\mu d_\nu) = \frac{(1-\Phi)^2}{\rho^5} d^{(\mu}\dot{z}_{\nu)} + \frac{(1-\Phi)\Phi}{\rho^5} d^{(\mu}\dot{z}_{\nu)} + \frac{(1-\Phi)}{\rho^4} d^{(\mu}\ddot{z}_{\nu)}$$

$$+ \frac{\Phi}{\rho^4} d^{(\mu}\ddot{z}_{\nu)} - \frac{1}{\rho^6}[(1-\Phi)^2 + \rho^2 \ddot{z}^2] d^\mu d_\nu .$$

Reducing, we arrive at the result,

$$\frac{F^2}{k^2} = \frac{1-\Phi}{\rho^5} d^{(\mu}\dot{z}_{\nu)} + \frac{1}{\rho^4} d^{(\mu}\ddot{z}_{\nu)} - \frac{1}{\rho^6}[(1-\Phi)^2 + \rho^2 \ddot{z}^2] d^\mu d_\nu . \tag{C1}$$

Repeating the process,

$$\frac{F^3}{k^3} = \left[\frac{1-\Phi}{\rho^5}(d^\mu \dot{z}^\alpha + d^\alpha \dot{z}^\mu) + \frac{1}{\rho^4}(d^\mu \ddot{z}^\alpha + d^\alpha \ddot{z}^\mu) - \frac{(1-\Phi)^2 + \rho^2 \ddot{z}^2}{\rho^6} d^\mu d^\alpha\right]\left[\frac{1-\Phi}{\rho^3}(d_\alpha \dot{z}_\nu - d_\nu \dot{z}_\alpha)\right.$$

$$\left. + \frac{1}{\rho^2}(d_\alpha \ddot{z}_\nu - d_\nu \ddot{z}_\alpha)\right] = \frac{(1-\Phi)^2}{\rho^5}(\rho d^\mu \dot{z}_\nu - d^\mu d_\nu - \rho \dot{z}^\mu d_\nu) + \frac{1-\Phi}{\rho^7}(\rho d^\mu \ddot{z}_\nu - \Phi \dot{z}^\mu d_\nu)$$

$$+ \frac{1-\Phi}{\rho^7}(\Phi d^\mu \dot{z}_\nu - \rho \ddot{z}^\mu d_\nu) + \frac{1}{\rho^6}(\Phi d^\mu \ddot{z}_\nu - \ddot{z}^2 d^\mu d_\nu - \Phi \ddot{z}^\mu d_\nu)$$

$$+ \frac{[(1-\Phi)^2 + \rho^2 \ddot{z}^2](1-\Phi)}{\rho^9} \rho d^\mu d_\nu + \frac{(1-\Phi)^2 + \rho^2 \ddot{z}^2}{\rho^8} \Phi d^\mu d_\nu = \frac{1}{\rho^7}[(1-\Phi)^2 + \Phi(1$$

$$- \Phi)] d^{[\mu}\dot{z}_{\nu]} + \frac{1}{\rho^6}[\Phi + (1-\Phi)] d^{[\mu}\ddot{z}_{\nu]} + \frac{1}{\rho^8}\{-(1-\Phi)^2 - \rho^2 \ddot{z}^2 + [(1-\Phi)^2 + \rho^2 \ddot{z}^2](1-\Phi)$$

$$+ \Phi[(1-\Phi)^2 + \rho^2 \ddot{z}^2]\} d^\mu d_\nu = \frac{1-\Phi}{\rho^7} d^{[\mu}\dot{z}_{\nu]} + \frac{1}{\rho^6} d^{[\mu}\ddot{z}_{\nu]} .$$

Comparing with (9), we find

$$F^3 = \left(\frac{k}{\rho^2}\right)^2 F. \tag{C2}$$

Multiplying both sides of (C2) by F, n times, using again (C2) and separating even from odd exponents, we arrive at (10) and (11).

APPENDIX D: CALCULATION OF F^3_{rad}

Eliminating the Coulomb terms from the process that leads to (C1) above, we find[25]

$$\frac{F^2_{\text{rad}}}{k^2} = -\frac{1}{\rho^4}\left(\ddot{z}^2 + \frac{\Phi^2}{\rho^2}\right) d^\mu d_\nu .$$

Using F_{rad} from (B8),

$$\frac{F^3_{\rm rad}}{k^3}=-\frac{1}{\rho^4}\left(\ddot{z}^2+\frac{\Phi^2}{\rho^2}\right)d^\mu d_\alpha\left[-\frac{\Phi}{\rho^3}\,(d^\alpha\dot{z}_\nu-d_\nu\dot{z}^\alpha)+\frac{1}{\rho^2}\,(d^\alpha\ddot{z}_\nu-d_\nu\ddot{z}^\alpha)\right]$$

$$=-\frac{1}{\rho^4}\left(\ddot{z}^2+\frac{\Phi^2}{\rho^2}\right)d^\mu\left(\frac{\rho\Phi}{\rho^3}\,d_\nu-\frac{1}{\rho^2}\,\Phi d_\nu\right)=0.$$

[1] Metric: $(+,-,-,-)$. Fundamental length: $ds^2=c^2dt^2-dx^2-dy^2-dz^2$. Dots refer to differentiation with respect to this length. Strictly, this length is proper time by a constant. An additional factor of c is produced in each differentiation. Matrix form of 4×4 tensors: $A=A^\mu{}_\nu$. In this form square of tensors equals matrix multiplication: $A^2=A^\mu{}_\alpha A^\alpha{}_\nu$. Four-vectors are called vectors throughout and denoted by a or a^μ. Components of vectors: $a=(a^\circ,\mathbf{a})$. Antisymmetric bivector $a^\mu b_\nu-a_\nu b^\mu=a^{[\mu}b_{\nu]}$. Symmetric bivector: $a^\mu b_\nu+a_\nu b^\mu=a^{(\mu}b_{\nu)}$. Planes generated by the two vectors a^μ and b^μ: $(a\text{-}b)$-plane. Totally antisymmetric tensor $e^{\alpha\beta\gamma\delta}=-e_{\alpha\beta\gamma\delta}$ and the Dirac delta δ with usual definitions. Formulas are given in Heaviside–Lorentz units.

[2] L. B. Okun', ''The concept of mass (mass, energy, relativity),'' Sov. Phys. Usp. **32**, 629–638 (1989).
[3] F. Rohrlich, *Classical Charged Particles* (Addison-Wesley, Reading, MA, 1965).
[4] A. O. Barut, *Electrodynamics and Classical Theory of Fields and Particles* (Macmillan, New York, 1964).
[5] A. Schild, ''On the radiation emitted by an accelerated point charge,'' J. Math. Anal. Appl. **1**, 127–131 (1960).
[6] G. A. Schott, *Electromagnetic Radiation* (Cambridge University Press, Cambridge, England, 1912), p. 178.
[7] A. H. Taub, ''Orbits of charged particles in constant fields,'' Phys. Rev. **73**, 786–798 (1948).
[8] C. Lanczos, *The Variational Principles of Mechanics*, 2nd ed. (University of Toronto Press, Toronto, 1962), p. 327.
[9] E. Cartan, *La Théorie des Groupes Finis et Continus et la Geometrie Differentielle Traitées par la Méthode du Repère Mobile* (Gauthier-Villars, Paris, 1951).
[10] H. Weyl, *Space-Time-Matter* (Dover, New York, 1952), p. 143.
[11] A. O. Barut (Ref. 4), p. 168; F. Rohrlich (Ref. 3), p. 84. See also Appendix B.
[12] Appendix C. A detailed derivation of most of the results given in this paper can be found in A. Prieto R., Ph.D. thesis, University of London, Birkbeck College, 1994.
[13] J. L. Synge, ''Point-particles and energy tensors in Special Relativity,'' Ann. Mat. Pur. Appl. **84**, 33–59 (1970).
[14] Appendix D.
[15] H. Bacry, *Lectures on Group Theory and Particle Theory* (Gordon and Breach, New York, 1977), p. 323.
[16] A. O. Barut (Ref. 4), p. 180.
[17] Appendix A.
[18] Appendix A.
[19] H. Minkowski, in *The Principle of Relativity* (Dover, New York, 1923), p. 86.
[20] P. Caldirola, ''Sull'equazione del moto dell'elettrone nell'elettrodinamica classica,'' Nuovo Cimento **10**, 1747–1752 (1953).
[21] P. Caldirola, ''A relativistic theory of the classical electron,'' Riv. Nuovo Cimento **2**, 1–49 (1979).
[22] A. O. Barut (Ref. 4), p. 183.
[23] Extracted from E. Cartan (Ref. 9).
[24] F. Rohrlich (Ref. 3), p. 84 and in A. Barut (Ref. 4), p. 168.
[25] A. O. Barut (Ref. 4), p. 179.

Deriving the Hamilton equations of motion for a nonconservative system using a variational principle

Frank Thomas Tveter[a)]
Vidarsvei 21, N-1342 JAR, Norway

(Received 9 June 1997; accepted for publication 12 November 1997)

The classical derivation of the canonical transformation theory [H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, Reading, 1981)] is based on Hamilton's principle which is only valid for conservative systems. This paper avoids this principle by using an approach that is basically reversed compared to the classical derivation. The Lagrange equations of motion are formulated in the undefined and *general* variable set $\{Q,P\}$, and the general Hamilton equations of motion are derived from the Lagrange equations by using a variational principle. The undefined *general* variables $\{Q,P\}$ are defined through a transformation to a *special* (defined) variable set $\{q,p\}$. The transformation equations connecting the two sets are derived by using the invariants property of the value of the Lagrangian. This approach results in a more general interpretation of the generator function. © *1998 American Institute of Physics.* [S0022-2488(98)00803-2]

I. INTRODUCTION

The classical Hamilton formalism focuses on solving conservative problems. This formalism has also been used as a starting point for a number of perturbation theories designed to solve nonconservative (or complicated conservative) problems. In the early 1960s, Hori and Brouwer[1] developed a perturbation theory which they used to solve the problem of a satellite orbiting a planet under the influence of air drag. They formulated an early version of the general Hamilton equations of motion (in Cartesian coordinates) and derived a canonical transformation theory which allowed transformations of the general equations from one set of canonical variables to another. The equations of motion were solved in an appropriate variable set by using a series approach.

Hori and Brouwer did not attempt to derive the general Hamilton equations of motion. Such an attempt was made by Tveter.[2] The derivation was based on direct manipulation of the differential equations, and resulted in long and complicated intermediate expressions. In a subsequent study[3] a more compact and more general derivation was presented. That derivation was also based on direct manipulation of the differential equations.

The derivation presented in this paper is based on a simple variational principle and an unfrequented property of the Lagrange equations of motion. The Lagrange equations of motion may be formulated in a redundant set of variables, giving a linearly dependent set of equations. This is referred to as the *redundancy property of the Lagrange equations of motion*. This property is used to provide the starting point for the derivation in this paper. The Lagrange equations of motion are formulated in the *general* variable set $\{Q,P\}$ (where the position vector may depend on the full set of variables, i.e., $\mathbf{r}(Q,P,t)$), before these variables have been defined. The Hamilton equations of motion are derived from the Lagrange equations by using a variational principle. The definitions introduced in the transformation from the Lagrange equations to the Hamilton equations also provide the general equations for defining the momenta $\{P\}$. These equations may be reduced to the classical definition of the momenta in the *special* variable set $\{q,p\}$ (where the position vector only depends on the coordinates $\{q\}$, i.e., $\mathbf{r}(q,t)$) when certain assumptions are made. The definitions introduced in the derivation of the Hamilton equations, together with the invariants of the Lagrangian, are used to derive the canonical transformation theory. The *general* variable set $\{Q,P\}$ is finally defined through a transformation from the *special* variable set $\{q,p\}$.

a)Electronic mail: f.t.tveter@astro.uio.no, f.t.tveter@dnmi.no

 © 1998 American Institute of Physics

The approach presented in this paper is basically reversed compared to the classical derivation e.g. Ref. 4, where the starting point is the Lagrange equations of motion formulated in a *special* variable set $\{q\}$ (where the position vector only depends on the coordinates, i.e., $\mathbf{r}(q,t)$). In the classical derivation, the Hamilton equations of motion are first derived in the *special* variable set. Starting with these equations, a canonical transformation theory is developed which gives the Hamilton equations of motion for a *general* variable set $\{Q,P\}$ (where the position vector may depend on the full set of variables, i.e., $\mathbf{r}(Q,P,t)$). The classical derivation of the transformation theory is only valid for conservative systems.

There are several differences between the classical derivation and the derivation presented in this paper. Here, the starting point is the Lagrange equations of motion formulated in a *general* and redundant variable set, before the variables have been defined. This is achieved by using the redundancy property of the Lagrange equations of motion. The variational principle is used to derive the Hamilton equations from the Lagrange equations, and not in the derivation of the transformation equations as it is in the classical derivation. The generator function, which is introduced with the variational principle, is therefore related to a variable set and not a transformation. In this paper the transformation theory is based on the invariants property of the value of the Lagrangian, and not Hamilton's principle (a variational principle that only is valid for conservative systems).

II. THE HAMILTON EQUATIONS OF MOTION FOR NONCONSERVATIVE PROBLEMS

The investigation is limited to one mass under influence of two forces (however the derivation may easily be extended to a system of particles). The first force is a conservative force $\mathbf{F}_0$ for which the force potential U_0 is known. The second is a force $\mathbf{F}$ for which there may not be a force potential. The force $\mathbf{F}$ may be nonconservative.

Let us first consider the motion of a mass under the influence of the conservative force $\mathbf{F}_0$ (referred to as *the original problem*). The Hamilton equations of motion in the canonical variable set $\{Q,P\}$ for that problem are

$$\dot{P}_i=-\left(\frac{\partial K}{\partial Q_i}\right)_{QPt}, \quad \dot{Q}_i=\left(\frac{\partial K}{\partial P_i}\right)_{QPt}, \tag{1}$$

where K is the Hamilton. The subscript in the partial derivative expressions indicates the functional dependency of K, and the nonindexed variables indicate the full subset of variables.

If the mass comes under the influence of a force $\mathbf{F}$ *in addition* to $\mathbf{F}_0$, the equations of motion may be written as[3]

$$\dot{P}_i=-\left(\frac{\partial K}{\partial Q_i}\right)_{QPt}+\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial Q_i}\right)_{QPt},$$

$$\dot{Q}_i=\left(\frac{\partial K}{\partial P_i}\right)_{QPt}-\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial P_i}\right)_{QPt}, \tag{2}$$

where $\mathbf{r}$ is the position vector and the definitions of the functions $K(Q,P,t)$, $\mathbf{r}(Q,P,t)$, and $\dot{\mathbf{r}}(Q,P,t)$ are the same as in the original problem.

III. THE LAGRANGE EQUATIONS OF MOTION

The Lagrange formalism is capable of handling nonconservative problems.

Consider a mass under the influence of the force $\mathbf{F}_0$ with the potential U_0 and the nonconservative force $\mathbf{F}$. The Lagrange equations of motion for that system are

$$\frac{d}{dt}\left[\left(\frac{\partial L}{\partial \dot{u}_i}\right)_{u\dot{u}t}\right]-\left(\frac{\partial L}{\partial u_i}\right)_{u\dot{u}t}=\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial u_i}\right)_{ut}, \tag{3}$$

where $\{u\}$ represents the variable set and L is the Lagrangian, defined as

$$L=\tfrac{1}{2}m\dot{\mathbf{r}}^2-U_0, \tag{4}$$

where m is the mass. Some variable sets allow the velocity vector to depend on $\{u,\dot{u}\}$ in several different ways. The velocity vector is defined as

$$\dot{\mathbf{r}}=\left(\frac{\partial \mathbf{r}}{\partial u_i}\right)_{ut}\dot{u}+\left(\frac{\partial \mathbf{r}}{\partial t}\right)_{ut} \tag{5}$$

in the Lagrange equations of motion. The force potential U_0 is found by solving the equation

$$\frac{d}{dt}\left[\left(\frac{\partial U_0}{\partial \dot{\mathbf{r}}}\right)_{\mathbf{r}\dot{\mathbf{r}}t}\right]-\left(\frac{\partial U_0}{\partial \mathbf{r}}\right)_{\mathbf{r}\dot{\mathbf{r}}t}=\mathbf{F}_0. \tag{6}$$

The only property required from the variable set $\{u\}$ in (3) is that it can be used to determine the position vector. The variable set may therefore be redundant. In other words, even though only three variables are needed to describe the position of a mass in free space, the Lagrange equations of motion may be formulated using six (or more) variables.

In the redundant variable set $\{u\}=\{Q,P\}$ *the Lagrange equations of motion* are

$$\begin{aligned}\frac{d}{dt}\left[\left(\frac{\partial L}{\partial \dot{Q}_i}\right)_{QP\dot{Q}\dot{P}t}\right]-\left(\frac{\partial L}{\partial Q_i}\right)_{QP\dot{Q}\dot{P}t}&=\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial Q_i}\right)_{QPt},\\ \frac{d}{dt}\left[\left(\frac{\partial L}{\partial \dot{P}_i}\right)_{QP\dot{Q}\dot{P}t}\right]-\left(\frac{\partial L}{\partial P_i}\right)_{QP\dot{Q}\dot{P}t}&=\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial P_i}\right)_{QPt}.\end{aligned} \tag{7}$$

This set of (six) equations is linearly dependent and contains only three independent equations. The variable set $\{Q,P\}$ is not defined at this point.

The definition of the potential U_0 does not involve any variables explicitly, only the physical variables $\{\mathbf{r},\dot{\mathbf{r}},t\}$. The value of U_0 is therefore the same in all variable sets, i.e., it is invariant to variable transformations. The Lagrangian, L, is defined using U_0 and $\dot{\mathbf{r}}$, and its value is therefore also invariant to variable transformations. This implies that the Lagrangian is transformed by simple variable substitution. For instance,

$$L(q,\dot{q},t)=L(q(Q,P,t),\dot{q}(Q,P,\dot{Q},\dot{P},t),t)=L(Q,P,\dot{Q},\dot{P},t), \tag{8}$$

transforms L from the variable set $\{q\}$ to the redundant variable set $\{Q,P\}$.

IV. THE HAMILTON EQUATIONS OF MOTION

Let us now introduce the auxiliary function f, the Hamilton $K(Q,P,t)$ and the generator function S. The auxiliary function f (and the Hamilton K) is defined according to the equation

$$f(Q,P,\dot{Q},\dot{P},t)=L(Q,P,\dot{Q},\dot{P},t)-P_i\dot{Q}_i+K(Q,P,t)-\frac{dS}{dt}\equiv 0. \tag{9}$$

The generator function may depend on any list of variables (including variables from other sets). It is important to note that the generator function here is associated with the variable set $\{Q,P\}$, and not with a specific transformation between two variable sets, as it is in the classical approach.

From Eq. (9) four sets of equations may be derived. The first two are derived using the variational principle

$$\delta\int f(Q,P,\dot{Q},\dot{P},t)dt=\delta\int 0\,dt=0, \tag{10}$$

which is equivalent to

$$\frac{d}{dt}\left[\left(\frac{\partial f}{\partial \dot{Q}_i}\right)_{QP\dot{Q}\dot{P}t}\right]-\left(\frac{\partial f}{\partial Q_i}\right)_{QP\dot{Q}\dot{P}t}=0,$$

$$\frac{d}{dt}\left[\left(\frac{\partial f}{\partial \dot{P}_i}\right)_{QP\dot{Q}\dot{P}t}\right]-\left(\frac{\partial f}{\partial P_i}\right)_{QP\dot{Q}\dot{P}t}=0. \tag{11}$$

Using (9) we find

$$\frac{d}{dt}\left[\left(\frac{\partial L}{\partial \dot{Q}_i}\right)_{QP\dot{Q}\dot{P}t}-P_i\right]-\left(\frac{\partial L}{\partial Q_i}\right)_{QP\dot{Q}\dot{P}t}-\left(\frac{\partial K}{\partial Q_i}\right)_{QPt}=0,$$

$$\frac{d}{dt}\left[\left(\frac{\partial L}{\partial \dot{P}_i}\right)_{QP\dot{Q}\dot{P}t}\right]-\left(\frac{\partial L}{\partial P_i}\right)_{QP\dot{Q}\dot{P}t}+\dot{Q}_i-\left(\frac{\partial K}{\partial P_i}\right)_{QPt}=0. \tag{12}$$

By using (7), these equations can easily be reduced to

$$\dot{P}_i=-\left(\frac{\partial K}{\partial Q_i}\right)_{QPt}+\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial Q_i}\right)_{QPt},$$

$$\dot{Q}_i=\left(\frac{\partial K}{\partial P_i}\right)_{QPt}-\mathbf{F}\cdot\left(\frac{\partial \mathbf{r}}{\partial P_i}\right)_{QPt}. \tag{13}$$

These are *the general Hamilton equations of motion*. Observe that the generator function, S, does not appear in these equations.

The next two sets of equations are found by taking the partial derivatives of (9) with respect to $\dot{Q}_i$ and $\dot{P}_i$. This gives

$$\left(\frac{\partial L}{\partial \dot{Q}_i}\right)_{QP\dot{Q}\dot{P}t}-P_i-\left(\frac{\partial}{\partial \dot{Q}_i}\left(\frac{dS}{dt}\right)\right)_{QP\dot{Q}\dot{P}t}=0,$$

$$\left(\frac{\partial L}{\partial \dot{P}_i}\right)_{QP\dot{Q}\dot{P}t}-\left(\frac{\partial}{\partial \dot{P}_i}\left(\frac{dS}{dt}\right)\right)_{QP\dot{Q}\dot{P}t}=0. \tag{14}$$

These are the general equations for defining the momenta. The direct use of Eqs. (14) should be the focus of a future study. In this paper these equations are solved in the special variable set $\{q,p\}$ where $(dS/dt)=0$, i.e., S is a constant S_0. This special variable set is then transformed to the general variable set $\{Q,P\}$.

For a variable set $\{q,p\}$ where $S=S_0$, Eqs. (14) are reduced to

$$\left(\frac{\partial L}{\partial \dot{q}_i}\right)_{qp\dot{q}\dot{p}t}-p_i=0,$$

$$\left(\frac{\partial L}{\partial \dot{p}_i}\right)_{qp\dot{q}\dot{p}t}=0. \tag{15}$$

The last set of equations implies that the Lagrangian cannot depend on $\dot{p}$. The Lagrangian is defined using the velocity vector (4), so the velocity vector in (5) may not depend on $\dot{p}$ either. This is only possible if the position vector itself is independent of p. If the position vector and the velocity vector are independent of p and $\dot{p}$, then the Lagrangian must be independent of them as well. This reduces (15) to the classical definition of the momenta

$$p_i=\left(\frac{\partial L}{\partial \dot{q}_i}\right)_{q\dot{q}t}. \tag{16}$$

If the coordinates q are defined so that they can be used to determine the position vector $\mathbf{r}(q,t)$, the Lagrangian $L(q,\dot{q},t)$ may be found, and the momenta p can be found by using (16). The Hamilton $H(q,p,t)$, which is the equivalent to $K(Q,P,t)$ for $S=S_0$, may then be found by using (9) (with $(dS/dt)=0$).

V. TRANSFORMATION PROPERTIES

The starting point for finding the transformation equations linking the variable set $\{q,p\}$ (where $S=S_0$) with the variable set $\{Q,P\}$ is the definition of the auxiliary function f in (9). This equation, formulated for the two variable sets, gives

$$L(q,\dot{q},t)-p_i\dot{q}_i+H(q,p,t)\equiv 0, \tag{17}$$

and

$$L(Q,P,\dot{Q},\dot{P},t)-P_i\dot{Q}_i+K(Q,P,t)-\frac{dS}{dt}\equiv 0. \tag{18}$$

The discussion which led to (8) concluded that the value of the Lagrangian is invariant to variable transformations. Eliminating the (invariant) Lagrangian from (17) and (18) gives

$$p_i\dot{q}_i-H(q,p,t)=P_i\dot{Q}_i-K(Q,P,t)+\frac{d(S-S_0)}{dt}, \tag{19}$$

where $S-S_0$ is the difference between the generator functions in the two variable sets. Equation (19) is the same equation as the one we used to find the transformation equations in the classical Hamilton formalism. In the classical formalism, the function $\Delta S=S-S_0$ is referred to *as the generator function of the transformation*. For instance, if this function is defined as $\Delta S=\Delta S(q,Q,t)$, (19) becomes

$$p_i\dot{q}_i-H(q,p,t)=P_i\dot{Q}_i-K(Q,P,t)+\left(\frac{\partial \Delta S}{\partial q_i}\right)_{qQt}\dot{q}_i+\left(\frac{\partial \Delta S}{\partial Q_i}\right)_{qQt}\dot{Q}_i+\left(\frac{\partial \Delta S}{\partial t}\right)_{qQt} \tag{20}$$

with a solution described by the following transformation equations:

$$p_i=\left(\frac{\partial \Delta S}{\partial q_i}\right)_{qQt}, \tag{21}$$

$$P_i=-\left(\frac{\partial \Delta S}{\partial Q_i}\right)_{qQt}, \tag{22}$$

$$K=H+\left(\frac{\partial \Delta S}{\partial t}\right)_{qQt}. \tag{23}$$

Notice that the perturbing force **F** appears only in the equations of motion, and that all other definitions are unaffected by the presence of this extra force.

VI. CONCLUSIONS

The main achievement of this study is the derivation of the Hamilton equations of motion for nonconservative systems using a variational principle. The study demonstrates how the Lagrange equations of motion may be formulated in a redundant variable set, and how the Hamilton equations of motion are directly linked to the Lagrange equations of motion. The study also shows how the invariants property of the value of the Lagrangian can be used to derive the canonical trans-

formation theory. This approach associates the generator function to the variable set, and not to a specific canonical transformation. The generator function is a constant for variable sets where the position vector is independent of the momenta.

ACKNOWLEDGMENTS

The redundancy property of the Lagrange equations of motion was first brought to the author's attention by Professor Jacques Henrard (Electronic mail: jhenrard@math.fundp.ac.be). The author is also grateful to Professor Egil Leer (Electronic mail: egil.leer@astro.uio.no) and Professor Jan Trulsen (Electronic mail: jan.trulsen@astro.uio.no) for their suggestions and criticism.

[1] G. Hori and D. Brouwer, ''Theoretical evaluation of atmospheric drag effects in the motion of an artificial satellite,'' Astron. J. **66**, 193–225 (1961).
[2] F. Tveter, ''Analytisk baneteori for ERS-1,'' Can. Sci. thesis, University of Oslo, 1992.
[3] F. Tveter, ''Hamilton's equations of motion for nonconservative systems,'' Celest. Mech. Dyn. Astron. **60**, 409–419 (1994).
[4] H. Goldstein, *Classical Mechanics,* 2nd ed. (Addison–Wesley, Reading, 1981).

A classical large *N* hierarchical vector model in three dimensions: A nonzero fixed point and canonical decay of correlation functions

P. A. Faria da Veiga[a),b)]
ICMSC—Universidade de São Paulo, Caixa Postal 668, 13560 São Carlos SP, Brazil

M. O'Carroll[a),c)] and R. Schor[a)]
Departamento de Física—ICEX, Universidade Federal de Minas Gerais, 31270 Belo Horizonte MG, Brazil

(Received 4 September 1997; accepted for publication 2 December 1997)

We consider a hierarchical N-component classical vector model on a three-dimensional lattice $\mathbb{Z}^3$, for large N. The model differs from the usual one in that the kernel of the inverse Laplace operator is nontranslational invariant but has matrix elements which are positive and exhibit the same falloff as the inverse Laplacian in $\mathbb{Z}^3$. We introduce a renormalization group transformation and for $N=\infty$, corresponding to the leading order of the $1/N$ expansion, we construct explicitly a nonzero fixed point for this transformation and also obtain some correlation functions. The two-point function has canonical decay. For $1 \ll N < \infty$, we obtain the fixed point and the two-point function in the first $1/N$ approximation. Canonical decay is still verified, in contrast to what is reported for the full model. © *1998 American Institute of Physics.* [S0022-2488(98)04203-0]

I. INTRODUCTION

Hierarchical models are formulated to mimic systems of interest in statistical mechanics and euclidean quantum field theories. In particular, they have proved to be a good laboratory to test ideas and to develop tools that allow the rigorous analysis of the continuum and the thermodynamic limits in such systems.

In the case of boson lattice field theories, the hierarchical versions are defined by replacing, in the Gaussian (free) measure, the Laplace operator Δ in the inverse covariance by a hierarchical long-range Laplacian Δ_h. In general, although $(\Delta_h)^{-1}$ is not translational invariant on the lattice, it has matrix elements with the same canonical decay properties as for Δ^{-1}.

The main simplification involved in hierarchical models is of a technical nature and has to do with renormalization group transformations (RGT), which preserve locality of the interacting potential.

In Ref. 1, a given hierarchical version for the lattice real boson vector model, with a large number $N \in \mathbb{N}$ of components and in three euclidean dimensions, was considered that led to a recursion of the RGT which corresponds to the Wilson's (approximate) recursion when the value of the associated block-spin length, $L \in \mathbb{N}, L>1$, in the lattice $\Lambda \subset \mathbb{Z}^3$, equals two for each RGT step. In dimension $d \in \mathbb{N}$, the corresponding RGT is given by

$$e^{-\mathscr{T}V(\boldsymbol{\phi})} = \text{const} \int e^{-(1/2)L^d[V(L^{-(d-2)/2}\boldsymbol{\phi}+\mathbf{z})+V(L^{-(d-2)/2}\boldsymbol{\phi}-\mathbf{z})]} d\nu(\mathbf{z}), \tag{I.1}$$

where $\boldsymbol{\phi}=(\phi_1,\phi_2,\ldots,\phi_N) \in \mathbb{R}^N$ denotes the N-component field (a random process), V is the interacting potential, $\mathbf{z}$ describes the fluctuation field (another random process) and where

a)Partially supported by CNPq.
b)Electronic mail: veiga@icmsc.sc.usp.br
c)Electronic mail: ocarroll@oraculo.lcc.ufmg.br

 © 1998 American Institute of Physics

$$d\nu(\mathbf{z}) = \prod_{x \in \Lambda} e^{-(1/2)|\mathbf{z}(x)|^2} d\mathbf{z}(x)$$

for $d\mathbf{z}(x)$, at each lattice site x, being the product of Lesbesgue measures $dz_1(x)\cdots dz_N(x)$; is a normalized and mean zero Gaussian measure with unit covariance on $\Lambda \subset \mathbb{Z}^d$. In (I.1), the const factor is a normalization and $\mathscr{T}V$ describes the effective potential generated by this RGT.

In Ref. 1, a nonzero fixed point solution of the RGT (I.1) was constructed first for the $N = \infty$ regime, given by the leading order in the $1/N$ expansion for the model, and was then extended to the case $1 \ll N < \infty$. Physical stability, asymptotics and other properties of this solution have also been analyzed.

Recently, in a companion paper,[2] we reconsidered the previous analysis of Ref. 1 for the finite N fixed point and improved our understanding of this model by constructing the thermodynamic limit of some correlations. Even though the fixed point is nonzero, we have shown that the two-point function has canonical (nonanomalous) decay at large distances. The canonical decay is also checked for other correlations studied there. This behavior differs from what is reported to occur in the full model (see, e.g., Refs. 3, 4).

In this paper, we consider another hierarchical version for the same problem for which, as will be shown in the sequel, instead of the Wilson's transformation, the RGT in dimension d is given by (see Ref. 5)

$$e^{-\mathscr{T}V(\boldsymbol{\phi})} = \text{const} \int e^{-L^d[V(L^{-(d-2)/2}\boldsymbol{\phi}+\mathbf{z})]} d\nu(\mathbf{z}). \tag{I.2}$$

Due to the nonsymmetrical form of this transformation, in contrast with the Wilson's case given in Eq. (I.1), we will refer to the present model as the *nonsymmetrical* hierarchical model hereafter.

The point we will be focusing on is the decay of correlations at large distances. Precisely, we want to determine whether or not the correlation functions of the nonsymmetrical hierarchical model also have canonical decay.

The reason that justifies our motivation in performing this analysis lays in the fact that, while sharing the ''simplicity'' of a hierarchical model, it presents desirable features not found in the Wilson's symmetrical model (I.1), namely,

—the associated inverse hierarchical Laplacian ${\Delta_h}^{-1}$, on the lattice, has positive matrix elements as for the Laplacian Δ^{-1} in the full case;

—for $N=1$, as it will be shown in Sec. I C, in the case of a quadratic perturbation $V(\phi) = \mu\phi^2/2, \mu \geqslant 0$, associated with a nonzero Gaussian fixed point, noncanonical or anomalous behavior of the two-point function appears discontinuously in μ, at $\mu = 0$;

—working with the nonsymmetrical hierarchical classical vector model in dimension $d=3$ or higher, the Goldstone picture has been confirmed in Refs. 6, 7 by showing the change from canonical to noncanonical decay of the two-point function, depending on the considered field direction.

These facts make the nonsymmetrical model a better candidate to mimic the full model than Wilson's. Next, we give a detailed description of this model.

A. The model

Formally, the partition function for the N-component vector model on a lattice $\mathbb{Z}^d$ is taken as

$$Z = \int e^{-(1/2)\langle \boldsymbol{\phi}, \Delta\boldsymbol{\phi}\rangle - V(\boldsymbol{\phi})} \prod_x d\boldsymbol{\phi}(x), \tag{I.3}$$

where $\Delta = -\Delta_L$, Δ_L being the lattice Laplacian, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N) \in \mathbb{R}^N, x \in \mathbb{Z}^d$.

We now define what we call the nonsymmetrical hierarchical model. With a view to our renormalization group (RG) analysis, it is convenient to introduce a sequence of unit lattices $\{\Lambda_j, j = 0,1,\ldots,l\}$, where

$$\Lambda_j = \left[-\frac{L^{l-j}}{2}, \frac{L^{l-j}}{2} \right]^d \cap \mathbb{Z}^d, \quad L > 1 \ \text{ odd}.$$

We also introduce N-component vector fields $\boldsymbol{\phi}^{(j)}(x^{(j)}) \in \mathbb{R}^N, x^{(j)} \in \Lambda_j$ and Gaussian measures $d\mu_j(\boldsymbol{\phi}^{(j)})$. $d\mu_j$ are normalized and mean-zero measures and have covariances G_j which are taken to be diagonal in the components of the vector field, being defined by

$$G_j(x^{(j)} = Lx^{(j+1)} + u, y^{(j)} = Ly^{(j+1)} + \bar{u}) = L^{-(d-2)} G_{j+1}(x^{(j+1)}, y^{(j+1)}) + \delta_{j+1}(x^{(j+1)}, y^{(j+1)}), \tag{I.4}$$

where the symbol δ denotes the Kronecker delta function, $G_l \equiv G_l(x^{(l)}, y^{(l)}) = [1 - L^{-(d-2)}]^{-1}$ and u, $\bar{u} \in B_0$, B_0 being the lattice block $[-\frac{L}{2}, \frac{L}{2}]^d \cap \mathbb{Z}^d$ containing L^d points of Λ_j.

Letting the site 0 denote the origin for the coordinate system on the lattice and e_1 the lattice unit vector for any given direction, it can be checked that, upon iteration, we have

$$G_0(0, L^k e_1) = \frac{1}{(1 - L^{-(d-2)})} \frac{1}{L^{k(d-2)}}; \quad k = 0,1,...,l-1, \tag{I.5}$$

that is, canonical decay for the covariances.

The partition function for the nonsymmetrical hierarchical vector model in the finite volume $\Lambda_0 \equiv \Lambda$ is taken as in (I.3) with the Gaussian part being given by $d\mu_0$, and for $x \equiv x^{(0)}$ and $\boldsymbol{\phi}(x) \equiv \boldsymbol{\phi}^{(0)}(x^{(0)})$, we have $V(\boldsymbol{\phi}) = \Sigma_{x^{(0)} \in \Lambda_0} V(\boldsymbol{\phi}^{(0)}(x^{(0)}))$. It reads

$$Z_0 \equiv Z_0(l) = \int e^{-V(\boldsymbol{\phi})} d\mu_0(\boldsymbol{\phi}). \tag{I.6}$$

The definition of the normalized p-point correlation functions follows in a similar way and is given by

$$\langle \phi_{i_1}(x_1) \phi_{i_2}(x_2) \cdots \phi_{i_p}(x_p) \rangle_l = (Z_0)^{-1} \int \phi_{i_1}(x_1) \phi_{i_2}(x_2) \cdots \phi_{i_p}(x_p) e^{-V(\boldsymbol{\phi})} d\mu_0(\boldsymbol{\phi}), \tag{I.7}$$

with $i_1, \ldots, i_p = 1,...,N$, and with Z_0 given by (I.6).

B. The nonsymmetrical RGT

In the context of the RG method, we now present some identities for the partition function, introduce the RGT, effective actions and give representations for correlation functions in terms of effective actions and effective spin (field) variables.

Corresponding to the decomposition (I.4) of the covariance, we introduce a decomposition for the fields given by

$$\boldsymbol{\phi}^{(j)}(x^{(j)} = Lx^{(j+1)} + u) = L^{-(d-2)/2} \boldsymbol{\phi}^{(j+1)}(x^{(j+1)}) + \boldsymbol{\eta}^{(j+1)}(x^{(j+1)}),$$

which yields to the fundamental identity, for any given function f_x,

$$\int \prod_{x \in \Lambda_j} f_x(\boldsymbol{\phi}^{(j)}) d\mu_j(\boldsymbol{\phi}^{(j)}) = \int \prod_{x \in \Lambda_{j+1}} \overline{f_x}(\boldsymbol{\phi}^{(j+1)}) d\mu_{j+1}(\boldsymbol{\phi}^{(j+1)}), \tag{I.8}$$

with

$$\overline{f_x}(\boldsymbol{\phi}) = \int \prod_{u \in B_0} f_{Lx+u}(L^{-(d-2)/2} \boldsymbol{\phi} + \boldsymbol{\zeta}) e^{-\xi^2/2} d\boldsymbol{\zeta},$$

where we adopt the convention that the $\boldsymbol{\zeta}$ integral includes the normalization factor $(2\pi)^{-N/2}$.

The fundamental identity (I.8) follows from the decomposition identity for the generating function of Gaussian measures. As it can be seen, repeated applications of (I.8) to the partition function $Z_0 \equiv Z_0(l)$ gives

$$Z_0 \equiv \int \prod_{x^{(0)} \in \Lambda_0} e^{-V(\boldsymbol{\phi}^{(0)}(x^{(0)}))} d\mu_0 = \int \prod_{x^{(1)} \in \Lambda_1} e^{-RV(\boldsymbol{\phi}^{(1)}(x^{(1)}))} d\mu_1$$

$$= \cdots = \int e^{-R^l V(\boldsymbol{\phi}^{(l)}(x^{(l)}))} d\mu_l \equiv Z_l ,$$

where, up to a normalization, the RGT is defined, in agreement with (I.2), by

$$e^{-RW(\boldsymbol{\phi})} = \int e^{-L^d W(L^{-(d-2)/2}\boldsymbol{\phi} + \boldsymbol{\eta})} e^{-\boldsymbol{\eta}^2/2} d\boldsymbol{\eta}. \tag{I.9}$$

Thus, it is clear that the effective actions remain local under this transformation.

Applying the fundamental identity (I.8) to both, the numerator and the denominator of the expression (I.7), we can represent the p-point functions in terms of effective spin variables. Doing this for the two-point function, and letting $I_i(\boldsymbol{\phi}) = \phi_i$, we get

$$\langle \boldsymbol{\phi}_i(0) \boldsymbol{\phi}_j(L^k e_1) \rangle_l$$

$$= (Z_1)^{-1} \int M_V I_i(\boldsymbol{\phi}^{(1)}(0)) M_V I_j(\boldsymbol{\phi}^{(1)}(L^{k-1} e_1)) \prod_{x^{(1)} \in \Lambda_1} e^{-RV(\boldsymbol{\phi}^{(1)}(x^{(1)}))} d\mu_1$$

$$= \cdots = (Z_k)^{-1} \int [(M_{R^{k-1}V} \cdots M_V I_i)(\boldsymbol{\phi}^{(k)}(0))][(M_{R^{k-1}V} \cdots M_V I_j)(\boldsymbol{\phi}^{(k)}(e_1))]$$

$$\times \prod_{x^{(k)} \in \Lambda_k} e^{-R^k V(\boldsymbol{\phi}^{(k)}(x^{(k)}))} d\mu_k$$

$$= \cdots = (Z_{k+1})^{-1} \int [M_{R^k V}[(M_{R^{k-1}V} \cdots M_V I_i)(M_{R^{k-1}V}$$

$$\times \cdots M_V I_j)]](\boldsymbol{\phi}^{(k+1)}(0)) \prod_{x^{(k+1)} \in \Lambda_{k+1}} e^{-R^{k+1} V(\boldsymbol{\phi}^{(k+1)}(x^{(k+1)}))} d\mu_{k+1}$$

$$= \cdots = (Z_l)^{-1} \int [M_{R^{l-1}V} \cdots M_{R^k V}[(M_{R^{k-1}V} \cdots M_V I_i)(M_{R^{k-1}V} \cdots M_V I_j)]](\boldsymbol{\phi}^{(l)})$$

$$\times e^{-R^{l-1} V(\boldsymbol{\phi}^{(l)})} d\mu_l , \tag{I.10}$$

where all the above partition functions carry an explicit dependence on l and M_W is the linear operator

$$[M_W f](\boldsymbol{\phi}) = \frac{\int f(L^{-(d-2)/2}\boldsymbol{\phi} + \boldsymbol{\eta}) e^{-W(L^{-(d-2)/2}\boldsymbol{\phi} + \boldsymbol{\eta})} e^{-(1/2)\boldsymbol{\eta}^2} d\boldsymbol{\eta}}{\int e^{-W(L^{-(d-2)/2}\boldsymbol{\phi} + \boldsymbol{\eta})} e^{-(1/2)\boldsymbol{\eta}^2} d\boldsymbol{\eta}}. \tag{I.11}$$

Remarks:

(1) In (I.10), after $(k-1)$ iterations, the points 0 and $L^k e_1$ belong to the same block, thus reducing the two corresponding fields to the same variable. After this, the action of M is global, on both fields.
(2) We can similarly derive expressions like (I.10) for other correlations.
(3) We note that in (I.11), apart from a normalization, when acting on a field $\boldsymbol{\phi}$, M_W gives the associated effective spin (field) corresponding obtained by integration over the corresponding fluctuation against an interacting potential W.
(4) Also, it is clear from Eq. (I.11) that, at a given point W, the operator M_W is L^{-d} times the linearized RGT associated with RW at this point. This observation is helpful in understanding the scaling behavior of the effective spin variables.

The above formalism appears in Refs. 6–8, and was used in Ref. 2 to compute the $N=\infty$ correlation functions. When $N=\infty$, at the nonzero fixed point, the spectral properties of the operator M_V turn out to be completely known for the symmetrical case.

The same method is, in principle, successful to treat any case whenever the spectral resolution of M_V can be obtained. Below, we illustrate how it works in another model. For $d>2$, we consider the case of a nonsymmetrical hierarchical scalar (one-component) model with a quadratic perturbation. This consists of an example of how an anomalous decay does appear within a hierarchical nonsymmetrical model.

C. A nonzero fixed point and anomalous decay: A nonsymmetrical hierarchical model with a quadratic potential

We are going to consider the case of a quadratic perturbation $V(\phi)=\frac{1}{2}\mu\phi^2$, $\mu\geqslant 0$, of a nonsymmetrical hierarchical model with a one-component field in dimension $d>2$. The RGT here reduces to [see Eq. (1.9)]

$$e^{-RV(\phi)}=\int e^{-(1/2)\mu L^d(L^{-(d-2)/2}\phi+\eta)^2}e^{-\eta^2/2}d\eta. \tag{I.12}$$

Let us first assume $\mu>0$. Simply by computing the above Gaussian integral we see that, under (I.12), $V(\phi)\mapsto V_1(\phi)\equiv[RV](\phi)=\frac{1}{2}(\mu L^2/(1+\mu L^d))\phi^2\equiv\mu L^2/(1+\mu L^d)$. Obviously, iteration of the RGT (I.12) gives

$$V_n(\phi)\equiv[R^nV](\phi)=\frac{1}{2}\,\mu_n\phi^2=\frac{1}{2}\,\frac{\mu_{n-1}L^2}{1+\mu_{n-1}L^d}\,\phi^2,$$

with

$$\mu_j\equiv\frac{\mu_{j-1}L^2}{(1+\mu_{j-1}L^d)},\quad j=1,\ldots,n-1.$$

Hence, for $\mu\equiv\mu_0>0$, we have the nonzero fixed point solution

$$\widetilde{V}(\phi)\equiv\lim_{n\to\infty}V_n(\phi)=\tfrac{1}{2}\widetilde{\mu}\phi^2,$$

where

$$\widetilde{\mu}=\frac{L^2-1}{L^d} \tag{I.13}$$

corresponds to the $n\to\infty$ limit of

$$\mu_n=\frac{L^{2n}\mu_0\widetilde{\mu}}{\widetilde{\mu}+(L^{2n}-1)\mu_0}. \tag{I.14}$$

Now, we will compute the infinite volume two-point function

$$\langle\phi(0)\phi(L^ke_1)\rangle_\infty\equiv\lim_{l\to\infty}\langle\phi(0)\phi(L^ke_1)\rangle_l.$$

This will be done using (I.10). To compute the action of the M-operator strings appearing in (I.10), taking the two points to the same block, we will use the formalism developed in Ref. 6. There it is shown that the action of each of these operator strings can be expressed *uniquely* in terms of the last step effective action. For the generic N-component case, we have (compare with equation (1.13) of Ref. 6)

$$[M_{R^{n-1}V}\cdots M_VI_k]_l(\boldsymbol{\phi})=L^{-n[(d-2)/2]}\phi_k-a_n\,\frac{\partial R^nV(\boldsymbol{\phi})}{\partial\phi_k}, \tag{I.15}$$

where the constant $a_n \equiv a_n(d)$ is given by

$$a_n = L^{-n((d+2)/2)}(L^{2n}-1)\tilde{\mu}^{-1}. \tag{I.16}$$

When $N=1$, (I.15) amounts to

$$[M_{R^{n-1}V}\cdots M_V\, I]_l(\phi) = L^{-n(d-2)/2}\phi - a_n \frac{\partial R^n V(\phi)}{\partial \phi},$$

for $I(\phi)=\phi$, reducing to $[(L^{-n[(d-2)/2]}-a_n\mu_n)\phi]$ after evaluation on R^nV. Applying this result to (I.10), setting $\phi^{(l)} \equiv \phi$, we then obtain

$$\begin{aligned}
\langle \phi(0)\phi(L^k e_1)\rangle_l &= (Z_l)^{-1}\int [M_{R^{l-1}V}\cdots M_{R^kV}(M_{R^{k-1}V}\cdots M_V I)(M_{R^{k-1}V}\cdots M_V I)](\phi) \\
&\quad \times e^{-R^{l-1}V(\phi)} d\mu_l \\
&= (Z_l)^{-1}\int M_{R^{l-1}V}\cdots M_{R^kV}[(L^{-k(d-2)/2}-a_k\mu_k)I]^2(\phi) e^{-R^{l-1}V(\phi)} d\mu_l \\
&= (Z_l)^{-1}\int [M_{R^{l-1}V}\cdots M_{R^kV}] \\
&\quad \times \left[\left(L^{-k(d+2)/2} + (L^{-k(d+2)/2}-L^{-k(d-2)/2})\left(\frac{\mu_k}{\tilde{\mu}}-1\right)\right) I\right]^2(\phi) \\
&\quad \times e^{-R^{l-1}V(\phi)} d\mu_l,
\end{aligned} \tag{I.17}$$

where, at each step, the partition function $Z_l \equiv Z_l(l)$ is given by the same integral with $M_{R^{l-1}V}\cdots M_{R^kV}[\ ]^2(\phi)$ replaced by one.

On the other hand, we can easily compute the action of the remaining M-string in (I.17) obtaining

$$(M_{R^{l-1}V}\cdots M_{R^kV})\phi^2 = \gamma_{l-k} + \xi_{l-k}\phi^2, \tag{I.18}$$

where

$$\gamma_{l-k} = L^{-2}\sum_{j=k}^{l-1} \xi_j \frac{\mu_{j-1}}{\mu_j} \quad ; \quad \xi_{l-k} = L^{-(d+2)(l-2k)}\left(\frac{\mu_{l-k}}{\mu_k}\right)^2. \tag{I.19}$$

With this, using (I.14) and performing the last integrals in (I.17), we get

$$\begin{aligned}
\langle \phi(0)\phi(L^k e_1)\rangle_l &= \left[L^{-[(d+2)/2]k} + L^{-[(d-4)/2]k}(1-L^{-3k})\frac{\tilde{\mu}-\mu_0}{\tilde{\mu}+(L^{2k}-1)\mu_0}\right]^2 \\
&\quad \times \left(\gamma_{l-k} + \xi_{l-k}\frac{1}{1-L^{-(d-2)}}\right).
\end{aligned} \tag{I.20}$$

As, from (I.13), $0<\tilde{\mu}<1$ and, according to (I.14),

$$\tilde{\mu} < \frac{(L^{2n}-1)\mu_0\tilde{\mu}}{\tilde{\mu}+(L^{2n}-1)\mu_0} < \mu_n < 2\tilde{\mu},$$

the behavior of both γ_{l-k} and ξ_{l-k} in (I.19) are controlled by the $L^{-(d+2)(l-2k)}$ factor. For this reason, only the constant term of (I.18) matters in, both, the thermodynamic limit at fixed k and the double limit $k,l\to\infty$ with $k\leqslant l$, the later corresponding to the limit of infinite separation.

Letting γ_∞ denote the limit $\lim_{l\to\infty}\gamma_{l-k}$ [see (I.19)], then the last observation applied to (I.20) gives, after taking the thermodynamic limit,

$$\langle \phi(0)\phi(L^k e_1)\rangle_\infty = \gamma_\infty \left[L^{-[(d+2)/2]k} + L^{-[(d-4)/2]k}(1-L^{-3k}) \frac{\tilde{\mu}-\mu_0}{\tilde{\mu}+(L^{2k}-1)\mu_0} \right]^2 \tag{I.21}$$

whenever $\mu_0 > 0$. As, for $\mu_0 = 0$ [check below (I.4)], corresponding to a starting model with no interacting potential, the model has a zero fixed point and the two-point function has canonical decay (scaling $L^{-(d-2)}$), we see in (I.21) that the two-point function for $\mu_0 > 0$ has a noncanonical decay as given by the scaling factor $L^{(d+2)/2}$. The appearance of this effect has thus a discontinuity at $\mu = 0$ and is associated with a change from one to γ_∞ in the wave-function renormalization (coefficient).

D. The large *N* nonsymmetrical hierarchical vector model in three dimensions: Definition and results

Going back to the N-component vector model, hereafter we set $d = 3$ and assume that the initial action is rotational invariant, so that the system has an $O(N)$ symmetry. In order to analyze this model when N is a finite but large integer, it is suitable to introduce the reduced variable $\varphi \in \mathbb{R}$, and fluctuation variables $s \in \mathbb{R}$ and $u \geqslant 0$ verifying

$$\boldsymbol{\phi} = N^{1/2}(\varphi, 0, ..., 0) \tag{I.22}$$

and

$$\mathbf{z} = N^{1/2}(s, \zeta_1, ..., \zeta_{N-1}), \quad {\zeta_1}^2 + \cdots + {\zeta_{N-1}}^2 = u. \tag{I.23}$$

Then, $d\mathbf{z} = u^{(N-3)/2} du\, ds\, d\Omega_{N-1}$. Further upon making the substitutions $V(\boldsymbol{\phi}) = v(\boldsymbol{\phi}^2) \to (N/2)v(\varphi^2)$ and $RV(\boldsymbol{\phi}) = Rv(\boldsymbol{\phi}^2) \to (N/2)v_1(\varphi^2)$, the RGT for the nonsymmetrical model (I.9) can be written as

$$e^{-(N/2)v_1} = N\Omega_{N-1} \int_{-\infty}^{\infty} ds \int_0^\infty du \;\; e^{-L^3(N/2)v(\psi)} e^{-(N/2)(s^2+u-\ln u)} e^{-(3/2)\ln u}, \tag{I.24}$$

where $\psi \equiv \psi(\varphi, s, u) = (L^{-1/2}\varphi + s)^2 + u$ and Ω_{N-1} is the $(N-1)$-dimensional solid angle.

Corresponding to the leading order of the (formal) $1/N$-expansion for (I.24), and letting $v(\varphi^2) \equiv v_{N=\infty}(\varphi^2)$ be independent of N, we define the $N = \infty$ RGT as

$$v_1 = \inf_{s,u} \; [L^3 v(\psi) + s^2 + u - \ln u] \equiv L^3 v(\psi_0) + s_0^2 + u_0 - \ln u_0 \tag{I.25}$$

for $s_0 \equiv s_0(\varphi^2)$ and $u_0 \equiv u_0(\varphi^2)$ being solutions of the variational equations

$$L^3 v'(\psi)(s + L^{-1/2}\varphi) + s = 0,$$

$$L^3 v'(\psi) + 1 - 1/u = 0, \tag{I.26}$$

where the prime denotes a derivative with respect to φ^2, and $\psi_0 \equiv \psi(\varphi, s_0, u_0)$.

Our first result describes a nonzero fixed point solution to (I.25). We have:

Theorem 1: *Let $\tau = v'$ and $t_0 = L/(L-1) \equiv (\varphi_0)^2$. Then, when $N = \infty$, there exists a fixed point $\tilde{v}_{N=\infty}$ of (I.25) given, in terms of t, the inverse function of $\tilde{v}'$, by*

$$t(\tau) = t_0 + t_0 L \tau \sum_{n=0}^{\infty} \frac{1}{L^n} \left[\frac{1-L^{-2}}{(1-L^{-2}) - (1-L^{-2n})L\tau} \right]^2 \left[\frac{1-L^{-2}}{(1-L^{-2}) - (1-L^{-2(n+1)})L\tau} \right] \tag{I.27}$$

for $\tau \neq \tau_n \equiv (1-L^{-2})/(L(1-L^{-2n}))$, $n = 0,1,...$. Thus, $t(\tau)$ extends to an analytic function in a disk centered at τ_0. Moreover, the potential $\tilde{v}_{N=\infty}(\varphi^2)$ is stable whenever φ is real, that is $t'(\tau) > 0$ for $\tau < (1-L^{-2})/L$, and $\tilde{v}'_{N=\infty}(\varphi^2)$ is analytic there. The fixed point $\tilde{v}_{N=\infty}$ is nonzero and has a unique minimum at $\varphi^2 = (\varphi_0)^2$, with $\tilde{v}''_{N=\infty}((\varphi_0)^2) = (2L\varphi_0^2)^{-1}$. Finally, normalizing $\tilde{v}_{N=\infty}$ so that $\tilde{v}_{N=\infty}(\varphi^2) = 0$ the following global bounds hold

$$a'+b'|\varphi|\leqslant\tilde{v}_{N=\infty}(\varphi^2)\leqslant a+b\varphi^2,\quad \varphi\in\mathbb{R} \tag{I.28}$$

for some real a and a' and positive b and b'.

Remark: The derivatives of the critical potential $\tilde{v}_{N=\infty}(\varphi^2)$ can be all computed, as an application of the analytic inverse function theorem. An integral representation for the inverse function is given in Ref. 9, and in Ref. 10 explicit formulas for the Taylor coefficients are obtained.

Concerning the two-point function, we note that because of the $O(N)$ symmetry of the model,

$$\langle\phi_i(0)\phi_j(L^k e_1)\rangle_l=\frac{\delta_{ij}}{N}\langle\boldsymbol{\phi}(0)\cdot\boldsymbol{\phi}(L^k e_1)\rangle_l, \tag{I.29}$$

the dot denoting the Euclidean inner product in $\mathbb{R}^n$.

Using in (I.29) the definition (I.10), and applying the relations $\Sigma_{i=1}^N\phi_i(\partial V/\partial\phi_i)=2N\varphi^2v'(\varphi^2)$ and $\nabla V\cdot\nabla V=4N\varphi^2[v'(\varphi^2)]^2$, to pass to the reduced variables (I.22) and (I.23), we obtain, for $I(\varphi^2)=\varphi^2$,

$$\begin{aligned}\langle\phi_i(0)\phi_j(L^k e_1)\rangle_l=\delta_{ij}(Z_l)^{-1}\int&[m_{l-k}[(L^{-k/2}I-a_kv'(I)I)^2(\varphi^2)]]\\&\times e^{-(N/2)v(\varphi^2)}e^{-(N/2)(1-L^{-1})\varphi^2}e^{(N-1)\ln\varphi}d\varphi,\end{aligned} \tag{I.30}$$

where, as in (I.11), for rotational invariant $f(\varphi^2)$, we define

$$[mf](\varphi^2)=\frac{\int_{-\infty}^{\infty}ds\int_0^{\infty}du\,f(\psi)e^{-(N/2)(s^2+u-\ln u)-(3/2)\ln u}e^{-(N/2)L^3v(\psi)}}{\int_{-\infty}^{\infty}ds\int_0^{\infty}du\,e^{-(N/2)(s^2+u-\ln u)-(3/2)\ln u}e^{-(N/2)L^3v(\psi)}}. \tag{I.31}$$

Note that, for $N=\infty$, we define the action of $m_\infty\equiv m_{N=\infty}$ of the operator m in (I.31), according to the Laplace method, as

$$[m_\infty f](\varphi^2)=f(\psi_0\equiv(L^{-1/2}\varphi+s_0)^2+u_0). \tag{I.32}$$

We can now state our second result, regarding the $N=\infty$ two-point functions

Theorem 2: *For $N=\infty$, the thermodynamic limit $l\to\infty$ of the two-point function exists and we have*

$$\langle\phi_i(0)\phi_j(L^k e_1)\rangle_{N=\infty,\infty}\equiv\lim_{l\to\infty}\langle\phi_i(0)\phi_j(L^k e_1)\rangle_{N=\infty,l}=\delta_{ij}L^{-k}\varphi_0^2. \tag{I.33}$$

Remarks:

(1) As it will become clear in the sequel, our techniques apply as well to other correlation functions. In particular, for the ''partially truncated'' (P.T.) functions defined below, we have,

$$\begin{aligned}&\frac{1}{N^{n+j}}\,\langle[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(0)]^n[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(L^k e_1)]^j\rangle_{N=\infty,\infty}^{\text{P.T.}}\\&\quad\equiv\lim_{l\to\infty}\Bigg(\frac{1}{N^{n+j}}\,\langle[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(0)]^n[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(L^k e_1)]^j\rangle_{N=\infty,l}\\&\qquad-\frac{1}{N^{n+j}}\,\langle[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(0)]^n\rangle_{N=\infty,l}\langle[(\boldsymbol{\phi}\cdot\boldsymbol{\phi})(L^k e_1)]^j\rangle_{N=\infty,l}\Bigg)=L^{-2k(n+j)}(\varphi_0)^{2(n+j)}.\end{aligned}$$

Though the coefficient φ_0^2 in (I.33) shows a change in the wave-function renormalization, as compared to the free case, all the above correlations have canonical decay, as $L^{-k}=|x|^{-1}$.

We do not know if Wick's theorem holds for expectations of all monomials.

Let us now consider the $1 \ll N < \infty$ regime and the model as described by the $1/N$ expansion. In this case, by extending the preceding analysis to the first $1/N$ correction, using the Laplace method to write an expansion for the integrals in (I.10), we can state our last result:

Theorem 3: *Let* $1 \ll N < \infty$. *Up to order* $1/N$:

(1) *The fixed point solution to (I.24) is given by*

$$\tilde{v} = \tilde{v}_{N=\infty} + \frac{1}{N}\tilde{r}, \tag{I.34}$$

where $\tilde{v}_{N=\infty}$ *is the* $N=\infty$ *nonzero fixed point solution of Theorem 1 and* $\tilde{r}$ *is given in Eq. (III.4) below, so that* $\tilde{v} \neq 0$ *for sufficiently large* N.

(2) *The (finite-volume) two-point function for extremely separated points 0 and* $L^{l-1}e_1$ *in the lattice* Λ_l, *for fixed* l *and at the fixed point* $\tilde{v}$ *of (I.34), has canonical decay and is given by*

$$\langle \phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l = \delta_{ij} L^{-(l-1)}(\varphi_0)^2\left(1+\frac{c}{N}\right). \tag{I.35}$$

for some finite real constant c, *independent of* N *and* l.

Remark: The extreme separation condition should not be taken as an *a priori* limitation of the last result for the two-point function. Besides the fact c does not depend on l, as we saw in the case of the quadratic perturbation, accompanied by a change in the wave function renormalization, the two-point function has there the same noncanonical decay, no matter whether the thermodynamic limit is taken first, followed by the limit of infinite separation, or the separation taken to infinity, together with the volume.

In Sec. II, we treat the $N=\infty$ case and prove Theorems 1 and 2. Theorem 3 will be proven in Sec. III. We leave some final remarks for Sec. IV and, for completeness regarding the nonsymmetrical model, the spectral properties of the inverse Laplace operator $(\Delta_h)^{-1}$, as well as the falloff of the matrix elements of Δ_h, are presented in the Appendix.

II. THE $N=\infty$ MODEL: FIXED POINT AND CORRELATION FUNCTIONS

In this section, we consider the $N=\infty$ model. We derive and solve a functional equation for t appearing in Theorem 1, the inverse function of $\tilde{v}'_{N=\infty} \equiv d\tilde{v}_{N=\infty}/d\varphi^2$, where $\tilde{v}_{N=\infty} \equiv \tilde{v}_{N=\infty}(\varphi^2)$ is the fixed point solution of (I.25). Furthermore, we give bounds on the globally defined $\tilde{v}$, and also obtain correlation functions.

From the variational equations (I.26), we have

$$s_0 = (u_0 - 1)L^{-1/2}\varphi \tag{II.1}$$

and taking the derivative of (I.25) gives

$$v'_1 = h \cdot (v' \circ \mathscr{S}),$$

where the dot here means multiplication, ψ_0 is given below (I.26) and where we define

$$h = L^2 + \frac{L^{5/2}s_0}{\varphi}; \quad \mathscr{S}(\varphi^2) = \psi_0.$$

At the fixed point

$$v' = h \cdot (v' \circ \mathscr{S}). \tag{II.2}$$

Multiplying (II.2) by $1/h$, applying t, using (II.1) and (I.26), we obtain

$$t(v'/h) = L^{-1}u_0{}^2 + u_0. \tag{II.3}$$

We now introduce the parameter $\tau \equiv v'$. From (I.26) and (II.1), we have $u_0 = 1 - L\tau$ which, upon substituting in (II.3) yields

$$t(\tau)-\frac{L}{(1-L\tau)^2}\,t\left(\frac{\tau}{L^2(1-L\tau)}\right)=-\frac{L}{1-L\tau}. \tag{II.4}$$

From (II.4), we note that $t(0)=\varphi_0^2\equiv L/(L-1)$, so that introducing the linear operator $\mathscr{M}$ by

$$(\mathscr{M}f\,)(\tau)=\frac{L}{(1-L\tau)^2}\,f\left(\frac{\tau}{L^2(1-L\tau)}\right),$$

and, after subtracting $t(0)$, we rewrite (II.4) as

$$[(1-\mathscr{M})(t-t(0))](\tau)=\frac{L^2\tau}{(L-1)(1-L\tau)}. \tag{II.5}$$

Using in (II.5) the Neumann series for $(1-\mathscr{M})^{-1}$, we then obtain

$$t(\tau)=\frac{L}{L-1}\,\frac{1}{1-L\tau}+\frac{L^2}{L-1}\sum_{n=1}^{\infty}\left[\mathscr{M}^n\left(\frac{\cdot}{1-L\cdot}\right)\right](\tau). \tag{II.6}$$

It can be easily verified that

$$\left[\mathscr{M}^n\left(\frac{\cdot}{1-L\cdot}\right)\right](\tau)=\frac{\tau}{L^n}\,\frac{1}{[1-L(1+L^{-2}+\cdots+L^{-2(n-1)})]^2}\,\frac{1}{[1-L(1+L^{-2}+\cdots+L^{-2n})]} \tag{II.7}$$

which establishes (I.27). Thus, from (II.6) and (II.7), we see that $t(\tau)$ is meromorphic with poles at $\tau_n=L^{-1}[1+L^{-2}+\cdots+L^{-2n}]^{-1}$, $n=0,1,\ldots$, with an accumulation point at $\tau_\infty\equiv L^{-1}(1-L^{-2})$. Thus, the analytic inverse function theorem can be used to show analyticity of $\tilde{v}'$, as well as to compute $\tilde{v}''(\varphi_0^2)$.

It is clear from representation (II.6) that $t'\equiv dt(\tau)/d\tau$ is positive for $\tau\in[0,\tau_\infty)$, but we also have

Lemma 1: *The function t' satisfies $t'(\tau)>0$, for $\tau\in(-\infty,\tau_\infty)$.*

Proof: Notice that for

$$A_0\equiv 1;\quad A_n(\tau)\equiv\frac{\sigma}{\sigma-(1-L^{-2n}L\tau)},\quad n=1,\ldots,$$

with $\sigma=1-L^{-2}$, and with t_0 of Theorem 1, (II.6) can be written as

$$t(\tau)=t_0+t_0L\tau\sum_{n\geqslant 0}\frac{1}{L^n}\,[A_n(\tau)]^2A_{n+1}(\tau). \tag{II.8}$$

Next, using

$$\tau A_n'(\tau)=\frac{(1-L^{-2n})L\tau}{\sigma-(1-L^{-2n})L\tau}\,A_n(\tau),$$

we have from (II.8), by direct computation,

$$\frac{t'(\tau)}{t_0L}=[A_1(\tau)]^2+\sum_{n\geqslant 1}\frac{[A_n(\tau)]^2A_{n+1}(\tau)}{L^n}\,[2A_n(\tau)+A_{n+1}(\tau)-2]. \tag{II.9}$$

Then, we notice that A_n, $n>0$, has simple poles at p_n/L, where $\{p_n\equiv\sigma/(1-L^{-2n}),n\in\mathbb{N}\}$ is a decreasing sequence converging to $0<\sigma<1$. This ensures that the sequence $\{A_n,n\in\mathbb{N}\}=\{1/[1-p_n^{-1}],n\in\mathbb{N}\}$ converges to $A_\infty\equiv(\sigma/\sigma-L\tau)$, monotonically, being decreasing for $\tau<0$ and increasing for $0<\tau<\sigma/2$. This result can be used to get lower bounds on (II.9), taking the first or the last term of the sum, depending on the sign, and achieves the proof of Lemma 1. □

By Lemma 1, the analytic inverse function theorem, and the asymptotic behavior $\lim_{\tau\to-\infty}t(\tau)=-L$ and $\lim_{\tau\to\infty}t(\tau)=\infty$, there exists $v'(\varphi^2)$, monotone increasing and analytic in $(-L,\infty)$ with $v'(\varphi_0^2)=0$ and $\lim_{\varphi^2\to\infty} v'(\varphi^2)=(1-L^{-2})/L$. Normalizing v such that $v(\varphi_0^2)=0$, then $v(\varphi^2)$ has a Taylor expansion about φ_0^2 starting with $[1/(2L\varphi_0^2)](\varphi^2-\varphi_0^2)^2$. Moreover, we also have global upper and lower bounds for v given in:

Proposition 1: *Set* $a\equiv\varphi_0^2(2-L^{-2})\sigma^{-1}$ *and let* $\Phi^2=\varphi^2-\varphi_0^2$. *Then*

$$\frac{\sigma}{L}\left[\Phi^2-\frac{3a^{1/3}}{2}\left((\varphi^2)^{2/3}-(\varphi_0^2)^{2/3}\right)\right]\leqslant\tilde{v}_{N=\infty}(\varphi^2)\leqslant\frac{1}{L}\left[\Phi^2-\varphi_0^2\ln\frac{\varphi^2}{\varphi_0^2}\right]. \tag{II.10}$$

Proof: We note that if $f(\tau)>t(\tau)$ then, setting $h\equiv t^{-1}$, $v'(\varphi^2)<h(\varphi^2)$ which upon integration gives $v(\varphi^2)<\int_{\varphi_0^2}^{\varphi^2}h(\phi)d\phi$ and similarly for the lower bound. The upper bound in (II.10) follows by taking for f the first term in (II.6). For the lower bound, we note that the first term of $t(\tau)$ in (II.6) is $\varphi_0^2(1-L\tau)^{-1}$ and that the sum is bounded by $(\varphi^2\sigma)^{-1}\tau(1-\tau_\infty^{-1}\tau)^{-1}$. For f, we take $f=a(1-\tau_\infty^{-1}\tau)^{-3}$.

This concludes Proposition 1 and, as (II.10) gives (I.28), also the proof of Theorem 1. □

We now turn to the correlation functions. Iterating (II.2) p times gives

$$v'=H_p v'\circ\mathscr{G}^{(1)}\circ\cdots\circ\mathscr{G}^{(p)}, \tag{II.11}$$

where the upper indices count the number of compositions, and

$$H_p=\prod_{j=0}^{p-1}h_j$$

with $h_0\equiv h$ and

$$h_j(\varphi^2)=(h\circ\mathscr{G}^{(1)}\circ\cdots\circ\mathscr{G}^{(j)})(\varphi^2).$$

Multiplying (II.11) by $1/H_p$ and applying t yields

$$t\left(\frac{1}{H_p}\right)v'=\mathscr{G}^{(1)}\circ\cdots\circ\mathscr{G}^{(p)}. \tag{II.12}$$

Next, the linear operator m_∞ [see (I.32)] acting on $f(\varphi^2)$ gives

$$(m_\infty f)(\varphi^2)=f(\psi_0)\equiv(f\circ\mathscr{G})(\varphi^2) \tag{II.13}$$

so that $(m_\infty I)(\varphi^2)=\mathscr{G}(\varphi^2)$ and $m_\infty^n I=\mathscr{G}^{(1)}\circ\cdots\circ\mathscr{G}^{(n+1)}$. Thus, using (II.12) we get

$$m_\infty^n I=t\left(\frac{1}{H_{n+1}}v'\right)$$

and then, going back to (I.30), we obtain

$$\begin{aligned}\langle\phi_i(0)\phi_j(L^k e_1)\rangle_{N=\infty,\infty}&\equiv\lim_{l\to\infty}\langle\phi_i(0)\phi_j(L^k e_1)\rangle_{N=\infty,l}\\&=\delta_{ij}L^{-k}\lim_{n\to\infty}\left[t\left(\frac{1}{H_{n+1}}v'(\varphi^2)\right)\right]\\&=\delta_{ij}L^{-k}t(0)=\delta_{ij}L^{-k}\varphi_0^2,\end{aligned} \tag{II.14}$$

which corresponds to (I.33). Taking more general f's in (II.13) allows us to compute other correlation functions. We have now concluded the proof of Theorem 2. □

III. THE FIRST 1/*N* CORRECTION

In this section, we extend some of the $N=\infty$ results of Sec. II up to the first $1/N$ correction and prove Theorem 3. In the framework of the $1/N$ expansion, we write

$$v(\varphi^2)=v_{N=\infty}(\varphi^2)+\frac{1}{N}\,r(\varphi^2), \tag{III.1}$$

where $v_{N=\infty}$ is the N-independent leading order potential of Sec. II [see above Eq. (I.25)]. Doing the analogous decomposition for v_1 in (I.24), letting $w\equiv(s,u)\in\mathbb{R}\times[0,\infty)$ and for ψ defined below (I.24), we have

$$e^{-(N/2)(v_{N=\infty,1}+(1/N)r_1)}=N\Omega_{N-1}\int dw e^{-Nf(w)}e^{-g(w)}, \tag{III.2}$$

where, with an abuse of notation, we have

$$\begin{aligned} f(w)\equiv f(s,u)&=\tfrac{1}{2}L^3 v_{N=\infty}(\psi)+\tfrac{1}{2}(s^2+u-\ln u),\\ g(w)\equiv g(s,u)&=\tfrac{1}{2}L^3 r(\psi)+\tfrac{3}{2}\ln u. \end{aligned} \tag{III.3}$$

Applying the steepest descent method, we can compute an expansion for the integral in (III.2) about $w_0=(s_0,u_0)$, s_0 and u_0 being solutions of (I.26). Comparing terms of the same $1/N$ order gives back (I.25) and shows that, under the RGT (I.24), r transforms according to

$$r_1\varphi^2)=L^3 r(\psi_0)+3\ln u_0+\ln[(\det D^2 f\ )_0]$$

with ψ_0 defined below (I.24) and where $(\det D^2 f)_0$ is the Hessian of f at w_0. At the fixed point, we thus have

$$\tilde{r}(\varphi^2)=L^3\tilde{r}(\psi_0)+3\ln u_0+\ln[(\det D^2 f\ )_0] \tag{III.4}$$

which can already be used to calculate any derivative of $\tilde{r}$ at φ_0^2, assuming $\tilde{r}$ analytic. From (III.4), we also see that the $1/N$ order fixed point solution

$$\tilde{v}(\varphi^2)=\tilde{v}_{N=\infty}(\varphi^2)+\frac{1}{N}\,\tilde{r}(\varphi^2)$$

is nonzero for large enough N.

This shows part one of Theorem 3. To prove the second part, we go back to the two-point function (I.30) and, for technical simplification, we assume the points to have an extreme separation in the lattice. This is enough for our purposes here. The general case can be treated as well, following what is presented below.

Similar to f and g in (III.3), we define the functions $\hat{f}$ and $\hat{g}$ by

$$\begin{aligned} \hat{f}(\hat{w})\equiv\hat{f}(s,u,\varphi)&=f(w)+\frac{1-L^{-2}}{2}\,\varphi^2-\frac{1}{2}\ln\varphi^2,\\ \hat{g}(\hat{w})\equiv\hat{g}(s,u,\varphi)&=g(w)+\ln\varphi^2, \end{aligned} \tag{III.5}$$

with $\hat{w}\equiv(s,u,\varphi)\in\mathbb{R}\times[0,\infty)\times\mathbb{R}$. Now, letting

$$h(\hat{w})=\psi\left[L^{-l/2}\varphi-a_l\,\frac{dv(\psi)}{d\psi}\right]^2 \tag{III.6}$$

the two-point function (I.30), for extremely separated points 0 and $x=L^{l-1}e_1$ can be expressed as

$$\langle\phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l=\delta_{ij}(Z_l)^{-1}\int d\hat{w}h(\hat{w})e^{-N\hat{f}(\hat{w})}e^{-\hat{g}(\hat{w})} \tag{III.7}$$

with Z_l given by the same integral with $h(\hat{w})$ replaced by one.

Direct calculation shows that the function $\hat{f}$ has a minimum at $\hat{w}_0=(s_0,u_0,\varphi_0)$ with the same s_0 and u_0 of (I.26), and $\varphi_0=\sqrt{L/(L-1)}$ given in Theorem 1. Hence, applying the steepest descent method to the integrals in (III.7), with $y=\sqrt{N}(\hat{w}-\hat{w}_0)$, we obtain

$$\begin{aligned}\langle \phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l = \delta_{ij}h_0 + \frac{h_0'}{\sqrt{N}}\int y\, d\rho(y) - \frac{h_0'}{\sqrt{N}}\left(\int y\, d\rho(y)\right)\\ \times\left(\int\left[(T_0^{(3)}\hat{f}\,) + \frac{\hat{g}_0'}{\sqrt{N}}\,y\right]d\rho(y)\right)\\ + \int\left[-\frac{h_0'}{\sqrt{N}}\,y\left[(T_0^{(3)}\hat{f}\,) + \frac{\hat{g}_0'}{\sqrt{N}}\,y\right] + \frac{h_0''}{2N}\,y^2\right]d\rho(y) + \mathscr{O}(N^{-3/2}),\end{aligned} \tag{III.8}$$

where the zero subscript means evaluation at $\hat{w}=[\hat{w}_0+y/\sqrt{N}]_{y=0}=\hat{w}_0$, and $T^{(3)}$ is the third order Taylor operator, so that when acting on $\hat{f}$ produces an $\mathscr{O}(N^{-1/2})$ term. All the primes indicate derivatives within the multivariable context regarding $\hat{w}=(s,u,\varphi)$ entering in y, and $d\rho(y)$ is a normalized measure with Gaussian density $\exp(-\hat{f}_0''y^2/2-\hat{g}_0'y/\sqrt{N})$.

Applying a parity argument, all the linear terms in $s-s_0$ and $\varphi-\varphi_0$ in the $\int y\, d\rho(y)$ integrals vanish. Paying with an exponentially small factor $\mathscr{O}(e^{-\epsilon N})$, $\epsilon>0$, the same is verified for the integral of the linear $u-u_0$ term. This also ensures that only diagonal or product of diagonal terms (regarding the components of $\hat{w}$ and denoted below by $\mathscr{D}$) survive in the integrals of even powers of y so that (III.8) yields

$$\langle \phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l = \delta_{ij}h_0 - \frac{\delta_{ij}}{N}\int\left[(h_0'y)[\hat{g}_0'y+(T_0^{(3)}\hat{f}\,)y^3] + \frac{h_0''}{2}\,y^2\right]_{\mathscr{D}} d\rho(y) + \mathscr{O}(N^{-3/2}). \tag{III.9}$$

Using (III.1), (III.5) and (III.6), and for a_n in (I.16), putting $b_n\equiv L^{n/2}a_n = L^3(1-L^{-2l})(L^2-1)^{-1}$, we have

$$\begin{aligned}h_0 &= \varphi_0^2 L^{-l}\left(1-\frac{2b_l}{N}\left(\frac{dr(\psi)}{d\psi}\right)_0\right)\\ h_0' &= L^{-l}\left(\frac{\partial\psi}{\partial y}\right)_0[1-2\varphi_0^2 b_l v_\infty''(\varphi_0^2)]\\ h_0'' &= L^{-l}\left\{\left(\frac{\partial^2\psi}{\partial y^2}\right)_0[1-4b_l v_\infty''(\varphi_0^2)-2\varphi_0^2 b_l v_\infty''(\varphi_0^2)] + 2\varphi_0^2 b_l\left(\frac{\partial\psi}{\partial y}\right)_0^2[b_l(v_\infty''(\varphi_0^2))^2 - v_\infty'''(\varphi_0^2)]\right\}.\end{aligned} \tag{III.10}$$

As $\hat{g}_0'$ and $\hat{f}_0'''$ do not depend on l, we see that the two-point function (III.9) presents the structure

$$\langle \phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l = \delta_{ij}L^{-l}\varphi_0^2\left[1+\frac{1}{N}(K_1 b_l + K_2 b_l^2) + \mathscr{O}(N^{-3/2})\right]$$

for some constants K_1 and K_2, fixed essentially by Eqs. (III.10) and being independent of l and N. Hence, in the thermodynamic limit $l\to\infty$, we obtain

$$L^l\langle \phi_i(0)\phi_j(L^{l-1}e_1)\rangle_l \to \delta_{ij}\varphi_0^2\left[1+\frac{K}{N}+\mathscr{O}(N^{-3/2})\right]$$

for some finite K, uniform in N.

This shows (I.35) and concludes the proof of Theorem 3. □

Remark: With more work, the above analysis can be used to treat a generic two-point function and may be extended to compute other $2p$-point functions. It can also be used to obtain higher orders in $1/N$. It is likely that only even powers of $\sqrt{1/N}$ appear in the expansion for these functions.

IV. FINAL REMARKS

In the case of finite N, we have considered the fixed point and correlation functions only to order $1/N$, obtaining canonical decay for the two-point function. We expect this result to hold also without this restriction. The reason for this is that, due to the good properties of the $N=\infty$ fixed point solution, such as global stability and analyticity, we do not see any obstacles to carrying out a rigorous construction of a fixed point for N large, as in Ref. 1, and obtain the correlation functions as done in Ref. 2, for the symmetrical model.

There is the question of whether or not there exists some finite value N_0 of N where anomalous behavior sets in. For $N=1$, the existence of a nonzero fixed point in a hierarchical model is shown in Ref. 11, but the behavior of correlation functions is not determined. As our results show absence of anomalous decay (even though the fixed points are nonzero and the wave function renormalization is different from one) in both the symmetrical and the nonsymmetrical $O(N)$ vector hierarchical models, contrarily to what happens in the case of a one-component quadratic perturbation of Sec. I C and positive μ, it is certainly interesting to understand the full model from the point of view of configurations in a real space renormalization group approach.

APPENDIX: INFINITE LATTICE COVARIANCE

Here we obtain the spectral decomposition of the infinite lattice covariance operator G which is used to determine the long-range behavior of G^{-1}. G satisfies the recursion relation [see (I.4) and (I.5)], with $x,y\in\mathbb{Z}^d$, u_α, $\overline{u}_\alpha\in[0,L)$, $\alpha=1,\ldots,d$,

$$G(Lx+u,Ly+\overline{u})=L^{-(d-2)}G(x,y)+\delta(x,y).$$

We define the normalized block average operator $C:l_2(\mathbb{Z}^d)\to l_2(\mathbb{Z}^d)$ by

$$(Cf)(x)=L^{-d/2}\sum_{0\leqslant u_\alpha<L} f(Lx+u),$$

with Σ_u meaning the multiple sum over the d components of u. We find, for the adjoint C^*,

$$(C^*g)(Lx+u)=L^{-d/2}g(x),\quad 0\leqslant u_\alpha<L$$

and easily check that $CC^*=I$.

Now let, with $C_0=I$ and $C_n\equiv C^n$, $n\geqslant 1$,

$$\Gamma_n=C_n^*C_n-C_{n+1}^*C_{n+1}. \tag{A1}$$

We verify that

$$\Gamma_n\Gamma_m=\delta_{nm}\Gamma_n,\quad \Gamma_n^*=\Gamma_n$$

$$C_n^*\Gamma_nC_n=\Gamma_{n+1} \tag{A2}$$

$$C_{n+1}\Gamma_n=0. \tag{A3}$$

From (A1), we see that the functions in the range $R(\Gamma_n)$ of Γ_n are constant on the L^n blocks and, from (A3), we see that the average of $f\in R(\Gamma_n)$ over a L^{n+1} block is zero. Furthermore, $\{\Gamma_n\}$ furnishes a resolution of the identity, i.e., $I=\Sigma_{n=0}^{\infty}\Gamma_n$, and will be used in the spectral decomposition of G in $R(\Gamma_0)^{\perp}$ given by

Proposition 2: Let $\lambda_n=L^dL^{2(n-1)}(1-L^{-2n})/(1-L^{-2})$. *Then*

$$G=\sum_{n=0}^{\infty}\lambda_n\Gamma_n \tag{A4}$$

in the subspace $R(\Gamma_0)^{\perp}$. *Each* λ_n *has infinite multiplicity.*

Proof: It is easily shown that

$$G=L^2C^*GC+L^dC^*C. \tag{A5}$$

Write $G=\Sigma_{n=0}^{\infty}\beta_n\Gamma_n$, substitute in (A5) with $L^dC^*C=L^d[-\Gamma_0+\Sigma_{n=0}^{\infty}\Gamma_n]$ and use (A2). Equating coefficients of Γ_n we have

$$\beta_n=L^2\beta_{n-1}+L^d. \tag{A6}$$

Iterating (A6) with $\beta_0=0$ gives

$$\beta_n=L^d\sum_{j=0}^{n-1}L^{2j}=L^dL^{2(n-1)}\frac{1-L^{2n}}{1-L^2},$$

from which (A4) follows. □

From Proposition 2, the hierarchical Laplacian $\Delta_h=G^{-1}$ has a positive point spectrum $\{\lambda_n^{-1}\}$ with zero as the only point of accumulation.

We now determine the long-range behavior of

$$\Delta_h(0,L^ke)=\sum_n\lambda_n^{-1}\Gamma_n(0,L^ke), \tag{A7}$$

where we recall that $\Gamma_n=C_n^*C_n-C_{n+1}^*C_{n+1}$. Let $f(x)=\delta(x,0)$ and $g(y)=\delta(y,L^ke)$. Since $C_nf(x)=L^{-nd/2}\Sigma_{0\leqslant u<L^n}f(L^nx+u)$, $C_nf(x)$ only has support at $x=0$, and thus we only need $C_ng(0)=L^{-nd/2}\chi(k<n)$, where χ is the characteristic function. Substituting in (A7) gives

$$\Delta_h(0,L^ke)=\frac{L^d}{L^2-1}\sum_n(L^{2n}-1)[L^{-nd}\chi(k<n)-L^{-(n+1)d}\chi(k<(n+1))]$$

which, for large k, has a dominant const $L^{-k(d+2)}$ behavior.

The spectral decomposition of G can also be used to calculate directly the matrix elements of $(\Delta_h+\mu I)^{-1}=(G^{-1}+\mu I)^{-1}$, i.e., the two-point function of the hierarchical model with quadratic perturbation.

Note added in proof:

After the submission of this paper, we found that K. Gawedzki and A. Kupiainen [see ''Non-Gaussian scaling limits. Hierarchical model approximation,'' J. Stat. Phys. **35**, 267–284 (1984)] had already proven the main result of our companion work.[2] Neither did their method make use of our basic equations (I.10) and (I.15), nor did they solve the spectral properties of the linearized RGT at $N=\infty$.

[1] K. Gawedzki and A. Kupiainen, ''Non-Gaussian fixed points of the block spin transformation. Hierarchical model approximation,'' Commun. Math. Phys. **89**, 191–220 (1983).
[2] P. A. Faria da Veiga, M. O'Carroll, and R. Schor, ''Correlations functions for the large *N* classical vector hierarchical model in three dimensions'' (preprint).
[3] S. K. Ma, *Modern Theory of Critical Phenomena* (Benjamin, New York, 1976). See also S. K. Ma, ''The renormalization group and the large *n* limit,'' J. Math. Phys. **15**, 1866–1891 (1974).
[4] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Oxford University Press, Oxford, 1993).
[5] K. Gawedzki and A. Kupiainen, ''Asymptotic freedom and beyond perturbation theory,'' in *Phenomènes Critiques, Théories de Jauge et Systèmes Aléatoires*, edited by K. Osterwalder and R. Stora, Lès Houches Summer School, 1984 Session (Elsevier, Amsterdam, 1986).
[6] R. Schor and M. O'Carroll, ''Correlation functions and the Goldstone picture for the hierarchical classical vector model at low temperatures in three or more dimensions,'' J. Stat. Phys. **64**, 163–191 (1991).
[7] R. Schor and M. O'Carroll, ''Low temperature properties of the hierarchical classical vector model,'' Commun. Math. Phys. **138**, 487–505 (1991).

[8] K. Gawedzki and A. Kupiainen, ''Rigorous renormalization group and asymptotic freedom,'' in *Scaling and Self-Similarity in Physics* (J. Fröhlich, Birkhäuser, 1983).
[9] E. Hille, *Analytic Function Theory* (Ginn, Boston, 1962), 2 vols.
[10] R. Schor and M. O'Carroll, ''On the mass spectrum of the $2+1$ gauge-Higgs lattice quantum field theory,'' Commun. Math. Phys. **103**, 569–597 (1986).
[11] H. Koch and H. Wittwer, ''A non-Gaussian renormalization group fixed point for hierarchical scalar lattice field theories,'' Commun. Math. Phys. **106**, 495–532 (1986); ''On the renormalization group transformation for scalar hierarchical models,'' *ibid.* **138**, 537–568 (1991).

Theory of nonequilibrium first-order phase transitions for stochastic dynamics

Bernard Gaveau
Université Paris 6, Géométrie des équations aux dérivées partielles et physique mathématique, Case courrier 172, Tour 46, 5ème étage, 4, Place Jussieu, 75252 Paris Cédex 05, France

L. S. Schulman
Physics Department, Clarkson University, Potsdam, New York 13699-5820

(Received 5 May 1997; accepted for publication 6 November 1997)

A dynamic definition of a first-order phase transition is given. It is based on a master equation description of the time evolution of a system. When the operator generating that time evolution has an isolated near degeneracy there is a first-order phase transition. Conversely, when phenomena describable as first-order phase transitions occur in a system, the corresponding operator has near degeneracy. Estimates relating degree of degeneracy and degree of phase separation are given. This approach harks back to early ideas on phase transitions and degeneracy, but now enjoys greater generality because it involves an operator present in a wide variety of systems. Our definition is applicable to what have intuitively been considered phase transitions in nonequilibrium systems and to problematic near equilibrium cases, such as metastability. © *1998 American Institute of Physics.* [S0022-2488(98)01803-9]

I. INTRODUCTION

For systems out of equilibrium, much of the formal structure of thermodynamics and statistical mechanics is inapplicable. Numerous authors have addressed this problem, and our own efforts are found in Refs. 1 and 2. In the present article, we give a general theory of first-order phase transitions that holds in the nonequilibrium context, including systems that do not satisfy detailed balance. Our proposal is a natural offshoot of our approach, although we did find that some of the ''obvious'' assertions were not so easy to prove, even at the level of finite matrices. As could have been anticipated, the concept of analyticity of the partition function is of no use to us: there is in general no partition function. Rather, we focus on eigenvalue degeneracy, advocated by Kac among others. The limitation of this idea has been that many systems are not associated with linear operators. By going to dynamics we are able to use a nearly ubiquitous operator, the generator of the master equation, allowing application to a wide variety of equilibrium and nonequilibrium systems. In equilibrium models, eigenvalue degeneracy of this operator has been studied[3] in conjunction with a corresponding analytic structure. However, in that same article[3] an example of a natural boundary at a first-order phase transition was provided. In our present view, the simplicity of the eigenvalue degeneracy idea compares well with the variety of analytic structures that can occur at first-order transitions, coupled with the fact that even for so basic an example as the two-dimensional Ising model, no one knows what singularity appears at the first-order transition. (What is known about the singularity[4] provides evidence against the possibility of using analyticity in definitions of metastability; we comment further below.) In addition to analyticity there is another common idealization that we avoid, namely the thermodynamic limit. We discuss this below but we mention now that there are physical reasons why this concept should fail out of equilibrium, with metastability again providing a relevant example.

In our approach we assume that the dynamics can be formulated as stochastic dynamics. The system is described by states x,y in a state space X. The dynamics are defined by a transition probability R_{xy}, the probability that the system state is in x at time $t+1$, given that it was in y at time t. This context excludes some situations, but we expect that aside from mathematical niceties the important physical exclusions are systems where quantum interference plays a role and sys-

 © 1998 American Institute of Physics

tems with infinite memory. By conservation of probability, R has an eigenvalue 1 and its associated right eigenvector p_0 is its stationary distribution. For simplicity, we assume that R is irreducible (otherwise split X and work with a reducible R) so that it has a unique stationary distribution. [Of course p_0 may not tell the whole story. Often though, a small modification of the dynamics can restore the focus on the stationary state. An example is above-threshold directed percolation, where the trivial absorbing state can be avoided with a vanishingly small (as size $\to\infty$) escape probability.]

Because of the close relation between our master equation approach and the theory of stochastic processes, our results may be of interest in that area as well. For example, issues related to those we raise here are discussed in Ref. 5.

In this article we establish the following two-way relation. (1) If the eigenvalue 1 of R is nearly degenerate and if other eigenvalues are distant (in a sense to be defined), then R is *nearly* reducible, that is, the support of its stationary state breaks into two pieces with hardly any transitions between them. (2) If two transition matrices, R and R' are close (i.e., $V\equiv R'-R$ is small in a certain sense), but have stationary states with nearly disjoint support, then the eigenvalue 1 for R is nearly degenerate. (This can describe critical points as well as first order transitions.) Furthermore, because the degenerate eigenvalue is a dynamic one, its (near-) degeneracy allows for the possibility of long-term phase coexistence. These results generalize to higher-order state multiplicities and eigenvalue degeneracies. Alternatively one could show the converse of item (1) above, namely, if the stationary measure breaks into pieces with a long transit time between them, then there is eigenvalue degeneracy. Running our proofs backward (in particular, the perturbation arguments of Sec. V) could establish this result, but we have instead chosen to focus on the instability aspect of a phase transition: small changes in external conditions lead to big changes in the system. Note that the notion of phase transition presented here implicitly involves the subtleties of asymptotic analysis. We speak of ''near degeneracy'' and ''near disjointness,'' etc. This is given practical meaning in terms of relations, such as Eq. (3.2) below, that allow the definition of time scales during which the transition is a recognizable physical phenomenon. Such an idea is certainly not new to physics, nor even to phase transitions,[6] but it is not the usual perception of an equilibrium phase transition. In particular this view is at odds with the notion of thermodynamic limit, for many a concept without which one cannot even contemplate a phase transition. However, as emphasized in Ref. 6, the relatively mild manifestation of nonequilibrium behavior in metastability arises precisely because volumes are finite; there are *two* large quantities: volume and droplet nucleation time.

As is implicit above, our *primitive* notion of a first-order phase transition is that there is a big change in the system for a small change in its dynamics. Our results establish a relation of this primitive notion to eigenvalue degeneracy for a dynamical operator that occurs in a wide variety of situations, equilibrium and nonequilibrium. As remarked, the significance of eigenvalue degeneracy was noted early in the study of phase transitions, [7–10] but given the frequent absence of a linear operator the idea was limited. Furthermore, results on analyticity breakdown in the thermodynamic limit made the latter concept appear to be the cleanest mathematical criterion for a phase transition. Notwithstanding the general satisfaction among workers in this field with notions of analyticity, phenomena such as metastability[11–17] never fit well with this picture.[18,6] In some of the referenced works on metastability a dynamical approach was taken and indeed the approach here may be considered an outgrowth of Refs. 3, 19, and 20. Moreover, metastability may be considered the gentlest of nonequilibrium phenomena. For the general nonequilibrium situation one can expect no help from ideas of analyticity, so that a dynamical approach is particularly suitable.

Another fundamental intuition for phase transitions is the coexistence of phases. As mentioned above, since the degenerate eigenvalue is dynamical, and since we show that the instability (''small change⇒big change'') implies degeneracy, we have as a corollary that there can be phase coexistence, not on an infinite time scale, but in the same asymptotic sense that underlies our other work.[21]

In Sec. II, we introduce our notation and general framework. Following that we detail assumptions germane to the problem at hand. In Sec. IV we establish basic identities and inequalities. In Sec. V this information is used to establish properties of individual phases, namely a measure of the degree of stability of each phase. This is continued in the following section, where exit times are calculated. In Sec. VII we use a method of two-time conditioning (similar to Ref.

19). In the next section we take up the converse: if there is a phase transition, does this imply degeneracy. We show (with technical assumptions) that it does. Finally in Sec. IX we deal with more than two phases.

We close with a comment on the issues of mathematical rigor and technical assumptions. We believe some of our assumptions are too restrictive and our results have greater generality. For example, Eq. (3.10) below probably assumes more than is needed.[22] But we also believe that for some systems and for certain limits the notions we propose really do break down, for physical reasons. Thus there are systems that are first-order phase transitions under our definition and in the physical sense as well (e.g., metastability) that lose definition in both senses for sufficiently large volume. We also mention that we sometimes do not provide full rigorous proofs, mainly because we did not consider it important at this stage, physically or mathematically, to define the additional hypotheses necessary for such proofs. An example is at the end of Sec. V. Finally, here and in our previous work[1,2] we have not addressed limitations due to assuming a finite, discrete state space and a discrete time. Although we do expect the basic idea to hold, for example, the Fokker–Planck equation for motion in a double well with a high barrier will show degeneracy, we also expect that the level of mathematical agony will increase considerably. And there may be cases where it really will make a difference.

II. NOTATION

Let R be a stochastic matrix on a space X. We assume it has only one eigenvalue of unit norm (to be called λ_0) and order the eigenvalues by decreasing modulus. For simplicity we assume that λ_1 is real. Thus the eigenvalues satisfy $\lambda_0 \equiv 1 > \lambda_1 \geqslant |\lambda_2| \geqslant \cdots$. The corresponding right and left eigenvectors are, respectively, p_k and A_k $(k=0,1,...)$, and satisfy

$$Rp_k = \lambda_k p_k, \quad A_k R = \lambda_k A_k.$$

We normalize

$$\langle A_k | p_l \rangle = \delta_{kl}, \tag{2.1}$$

p_0 is naturally normalized by

$$\sum p_0(x) = 1,$$

which fixes $A_0(x)=1$ for all x. (Thus $A_0 = \chi_X$, the characteristic function.) In the following we assume

$$p_0(x) > 0 \quad \text{for all } x \in X. \tag{2.2}$$

This corresponds to irreducibility and ergodicity assumptions on R. We normalize the left eigenstates by

$$\sup_{x \in X} |A_k(x)| = 1 \quad \text{for all } k. \tag{2.3}$$

For $t>0$ an integer, we denote by R^t the tth power of R; its eigenvalues satisfy $\lambda_0^t = 1 > \lambda_1^t \geqslant |\lambda_2^t| \geqslant \cdots$. We define two norms.

(1°) On left vectors, A,

$$\|A\|_L = \sup_{x \in X} |A(x)|.$$

(2°) On right vectors, q,

$$\|q\|_R = \sum_{x \in X} |q(x)|.$$

Accordingly, we define two norms on matrices, B,

$$\|B\|_R = \sup_{\|q\|_R \leqslant 1} \|Bq\|_R ,$$
$$\|B\|_L = \sup_{\|A\|_L \leqslant 1} \|AB\|_L , \tag{2.4}$$

III. ASSUMPTIONS CONCERNING NEAR DEGENERACY (2 PHASES)

We assume

$$1-\lambda_1 \ll 1-|\lambda_2| \tag{3.1}$$

and consider an intermediate time t such that

$$1-\lambda_1^t \sim \epsilon_1 ,$$
$$\lambda_2^t \quad \text{is small} \quad [\text{as specified in Eq. (3.10)}]. \tag{3.2}$$

We write

$$R^t = |p_0\rangle\langle A_0| + \lambda_1^t |p_1\rangle\langle A_1| + |\lambda_2^t| B^t, \tag{3.3}$$

where B^t is the remainder matrix

$$B^t = \sum_{k \geqslant 2} \frac{\lambda_k^t}{|\lambda_2^t|} |p_k\rangle\langle A_k|. \tag{3.4}$$

Although R need not be diagonalizable, the representation (3.3) is guaranteed by the nondegeneracy of λ_0 and λ_1. (Although for eigenvalues smaller than λ_1 a Jordan form may be needed, this would not affect our arguments below, since all we use is the orthogonality of the subspaces associated with those eigenvalues.) Because $\Sigma_x A_1(x) p_0(x) = 0$ and $p_0 > 0$, we deduce that

$$A_M \equiv \max_{x \in X} A_1(x) > 0 > A_m \equiv \min_{x \in X} A_1(x) \tag{3.5}$$

and we define accordingly the two states, which will correspond to the two phases of the system,

$$p_M^t(x) = p_0(x) + \lambda_1^t p_1(x) A_M + |\lambda_2^t| B_{xM}^t ,$$
$$p_m^t(x) = p_0(x) + \lambda_1^t p_1(x) A_m + |\lambda_2^t| B_{xm}^t . \tag{3.6}$$

Here B_{xM}^t (respectively B_{xm}^t) is the value of B_{xy}^t for some point $y(\equiv y_M)$ (respectively, y_m) such that $A_1(y) = A_M$ (respectively, A_m). From (3.6) we have

$$p_0(x) = \frac{A_M p_m^t(x) - A_m p_M^t(x)}{A_M - A_m} + |\lambda_2^t| \frac{A_m B_{xM}^t - A_M B_{xm}^t}{A_M - A_m},$$
$$p_1(x) = \frac{p_M^t(x) - p_m^t(x)}{\lambda_1^t (A_M - A_m)} + \frac{|\lambda_2^t|}{\lambda_1^t} \frac{B_{xm}^t - B_{xM}^t}{A_M - A_m}. \tag{3.7}$$

More generally, for any y we can define

$$p_y^t(x) = p_0(x) + \lambda_1^t p_1(x) A_1(y) + |\lambda_2^t| B_{xy}^t \equiv R_{xy}^t \tag{3.8}$$

so that, from Eqs. (3.7) and (3.8), we have

$$p_y^t(x) = C_M(y) p_M^t(x) + C_m(y) p_m^t(y) + q_{xy}^t \tag{3.9}$$

with

$$C_M(y)=\frac{\lambda_1^t A_1(y)-A_m}{A_M-A_m},\quad C_m(y)=\frac{A_M-\lambda_1^t A_m}{A_M-A_m},$$

and $q_{xy}=|\lambda_2^t|\times$ (combination of $B_{xM}^t, B_{xm}^t, B_{xy}^t$). Note that $C_M(y)$ and $C_m(y)$ are positive.

Our principal hypothesis is the following. We assume that B^t satisfies

$$|\lambda_2^t|\left|\sum_{y\in Y} B_{yx}^t\right|\leqslant \epsilon_2 \tag{3.10}$$

uniformly for any subset $Y\subset X$ and any point $x\in X$. In terms of the norms (2.4), this means that

$$|\lambda_2^t|\|B^t\|_R\leqslant \epsilon_2 .$$

For many interesting physical applications this assumption will not hold, although we expect our two-phase construction to be valid for systems more general than those obeying Eq. (3.10). Thus we consider the hypothesis to be needed for technical reasons.

IV. BASIC IDENTITIES AND INEQUALITIES

We take the scalar product of p_M^t and p_m^t with A_1. We have

$$\sum_x A_1(x)B_{xy}^t=0.$$

This follows from[23]

$$B^t=\sum_{k\geqslant 2}\frac{\lambda_k^t}{|\lambda_2^t|}|p_k\rangle\langle A_k|$$

and therefore

$$\sum_x A_1(x)p_M^t(x)=\lambda_1^t A_M,\quad \sum_x A_1(x)p_m^t(x)=\lambda_1^t A_m . \tag{4.1}$$

From (4.1) and the fact that $\Sigma p_M^t(x)=\Sigma p_m^t(x)=1$ we deduce the basic identity

$$1-\lambda_1^t=\sum_x p_M^t(x)\left(1-\frac{A_1(x)}{A_M}\right)=\sum_x p_m^t(x)\left(1-\frac{A_1(x)}{A_m}\right). \tag{4.2}$$

The foregoing identity is a fundamental statement about the structure of the two phases. It does not depend on the hypothesis of Eq. (3.10).

Let us now fix $a>0$ and consider the two subsets

$$\begin{aligned} I_M(a)&=\left\{x\in X\;\middle|\;\frac{A_M-A_1(x)}{A_M}<a\right\},\\ I_m(a)&=\left\{x\in X\;\middle|\;\frac{A_m-A_1(x)}{A_m}<a\right\}. \end{aligned} \tag{4.3}$$

These subsets will in fact turn out to be the two phases. We also define $E_M(a)\equiv\mathscr{C}I_M(a)$, $E_m(a)\equiv\mathscr{C}I_m(a)$ (where $\mathscr{C}$ indicates the complement of a set). We take $a<1$ and note that

$$a<1\Rightarrow I_M(a) \text{ and } I_m(a) \text{ disjoint.}$$

From the definition (4.3) of $I_M(a), E_M(a)$, and the identity (4.2), we deduce

$$\sum_{x\in I_M(a)} p_M^t(x) > 1-\frac{1-\lambda_1^t}{a},$$
$$\sum_{x\in E_M(a)} p_M^t(x) < \frac{1-\lambda_1^t}{a}, \tag{4.4}$$

with analogous inequalities for M replaced by m.

We next estimate the total weight of $I_M(a)\cup I_m(a)$ for p_0. From Eq. (3.7)

$$\sum_{x\in \mathcal{C}(I_M\cup I_m)} p_0(x) = \sum_{x\in E_M\cap E_m} p_0(x) \leqslant \frac{1}{A_M-A_m}\left[A_M\sum_{x\in E_m} p_m^t(x) - A_m\sum_{x\in E_M} p_M^t(x)\right]$$
$$+\frac{|\lambda_2^t|}{A_M-A_m}\sum_{x\in E_M\cap E_m}(A_mB_{xM}-A_MB_{xM}),$$

where $A_M>0$ and $A_m<0$ have been used. Thus from Eq. (4.4)

$$\sum_{x\in \mathcal{C}(I_M\cup I_m)} p_0(x) \leqslant \frac{1-\lambda_1^t}{a}+\frac{|\lambda_2^t|}{A_M-A_m}\sum_{x\in E_M\cap E_m}(A_mB_{xM}-A_MB_{xm}). \tag{4.5}$$

Now it may happen that in addition to the conditions we have already placed on ϵ_1 and ϵ_2, they further satisfy $\epsilon_2 \ll \epsilon_1$. In that case it follows from the basic hypothesis (3.10) that up to terms much less than ϵ_1

$$\sum_{x\in \mathcal{C}(I_M\cup I_m)} p_0(x) \leqslant \frac{(1-\lambda_1^t)}{a}, \tag{4.6}$$

which means that almost all the weight of the stationary state is on the union of the two phase regions. Note that for Eq. (4.6) to hold the quantity ''a'' used in it must be slightly smaller $[\mathrm{O}(\epsilon_2)]$ than the quantity by the same name in Eq. (4.5). In other words, we have eliminated the last term in Eq. (4.5), using the hypotheses of Eq. (3.2). On the other hand, if $\epsilon_2 \sim \epsilon_1$ or $\epsilon_2 > \epsilon_1$, then the bound on the sum in Eq. (4.6) may depend on both ϵ_1 and ϵ_2, and may also involve slight changes in ''a''.

A. p_M^t and p_m^t as approximate eigenstates of R

By definition

$$Rp_M^t = p_M^{t+1} = p_M^t + (p_M^{t+1}-p_M^t).$$

Now

$$p_M^{t+1}(x)-p_M^t(x) = (\lambda_1-1)\lambda_1^t p_1(x)A_M + \lambda_2^{t+1}B_{xM}^{t+1} - \lambda_2^t B_{xm}^t.$$

It follows that in the R-norm

$$\|p_M^{t+1}(x)-p_M^t(x)\| = \sum_x |(\lambda_1-1)\lambda_1^t p_1(x)A_M + \lambda_2^{t+1}B_{xM}^{t+1} - \lambda_2^t B_{xm}^t| \leqslant c(1-\lambda_1).$$

Thus $Rp_M^t = p_M^t + \mathrm{O}(1-\lambda_1)$. In fact, for any integer s of the same general size as t we also have

$$R^s p_M^t = p_M^t + \mathrm{O}(1-\lambda_1^s), \tag{4.7}$$

so that p_M^t is substantially unchanged by the action of R^s (for s of order t).

B. Estimation of p_1

We have seen that

$$p_1(x)=\frac{p_M^t(x)-p_m^t(x)}{\lambda_1^t(A_M-A_m)}+r(x), \tag{4.8}$$

where $r(x)=\lambda_2^t\times$ (linear combination of $B_{xM}^t, B_{x,m}^t$). From the basic identity (4.2), we deduce

$$\begin{aligned}\sum_{x\in I_M} p_1(x)&\geqslant \frac{1}{a\lambda_1^t(A_M-A_m)}+\epsilon,\\ \sum_{x\in I_m} p_1(x)&\leqslant \frac{-1}{a\lambda_1^t(A_M-A_m)}+\epsilon',\end{aligned} \tag{4.9}$$

where ϵ, ϵ' are $O(1-\lambda_1^t)$.

V. ALMOST DEGENERATE PERTURBATION THEORY

The phases I_M and I_m are disjoint, and their union, $I_m\cup I_M$, fills almost all X, as measured by p_0. We can define a new stochastic process that is the same as the original one within $I_m\cup I_M$, but which is killed if it steps outside this union. In other words, we consider the substochastic matrix $\widetilde{R}$ defined by

$$\widetilde{R}_{xy}=\begin{cases}R_{xy} & \text{if both } x \text{ and } y \text{ are in } I_M \text{ or both } x \text{ and } y \text{ are in } I_m\\ 0 & \text{otherwise}\end{cases}. \tag{5.1}$$

Because of the disjointness, $\widetilde{R}$ is reducible and will have two distinct non-negative eigenvectors. Each of these will be associated with the positive eigenvalue of the largest norm for vectors having support on the corresponding subspace. Thus we will have two distinct solutions to

$$\widetilde{R}\widetilde{p}=\widetilde{\lambda}_0\widetilde{p}_0 \tag{5.2}$$

with (possibly) different $\widetilde{\lambda}_0$s. Clearly,

$$\widetilde{p}_0(x)=0 \quad \text{for} \quad x\notin I_M\cup I_m.$$

We seek solutions to Eq. (5.2) by perturbing around the largest eigenvalue of R. We write

$$\widetilde{R}_{xy}=R_{xy}-T_{xy}, \tag{5.3}$$

so that

$$T_{xy}=\begin{cases}0 & \text{if both } x \text{ and } y \text{ are in } I_M \text{ or both } x \text{ and } y \text{ are in } I_m\\ R_{xy} & \text{otherwise}\end{cases}.$$

We use degenerate perturbation theory to solve (5.2), treating T as a perturbation. Namely, we write

$$\widetilde{\lambda}_0=1-\mu_1, \quad \widetilde{p}=p_0+\alpha_1 p_1+q,$$

where $q=\Sigma_{k\geqslant 2}\ \alpha_k p_k$. We formally consider μ_1, q, and T to be small. Equation (5.2) becomes

$$(R-T)(p_0+\alpha_1 p_1+q)=(1-\mu_1)(p_0+\alpha_1 p_1+q).$$

After rearrangement this gives the exact equation

$$(R-I)q=(T-\mu_1)p_0+\alpha_1(1-\lambda_1)p_1+\alpha_1(T-\mu_1)p_1+(T-\mu_1)q. \tag{5.4}$$

Our hypothesis (which must be checked later) is that $(T-\mu_1)_q$ is of second order, leading to the approximate equation

$$(R-I)q=(T-\mu_1)p_0+\alpha_1(1-\lambda_1)p_1+\alpha_1(T-\mu_1)p_1. \tag{5.5}$$

We want to solve Eq. (5.5) with $q=\Sigma_{k\geqslant 2}\ \alpha_k p_k$ so that the second member must be orthogonal to A_0 and A_1. Taking the scalar product of the second member of Eq. (5.5) with A_0 and A_1 gives

$$\langle A_0|Tp_0\rangle+\alpha_1\langle A_0|Tp_1\rangle=\mu_1, \tag{5.6}$$

$$\langle A_1|Tp_0\rangle+\alpha_1(1-\lambda_1)+\alpha_1\langle A_1|Tp_1\rangle-\alpha_1\mu_1=0. \tag{5.7}$$

Eliminating μ_1 between (5.6) and (5.7) yields an equation for α_1

$$-\alpha_1^2\langle A_0|Tp_1\rangle+\alpha_1[-\langle A_0|Tp_0\rangle+\langle A_1|Tp_1\rangle+1-\lambda_1]+\langle A_1|Tp_0\rangle=0. \tag{5.8}$$

Call $Q(\alpha_1)$ the trinomial on the left. It can be rewritten as

$$Q(\alpha_1)\equiv-\langle\alpha_1A_0-A_1|T(p_0+\alpha_1p_1)\rangle+\alpha_1(1-\lambda_1).$$

But

$$\alpha_1(1-\lambda_1)=\alpha_1\langle A_0|Rp_0\rangle-\alpha_1\langle A_1|Rp_1\rangle=\langle\alpha_1A_0-A_1|R(p_0+\alpha_1p_1)\rangle,$$

so that finally

$$Q(\alpha_1)=-\langle\alpha_1A_0-A_1|\widetilde{R}(p_0+\alpha_1p_1)\rangle. \tag{5.9}$$

A. Calculation of α_1

We use Eq. (3.7) so that

$$p_0(x)+\alpha_1p_1(x)=\frac{\alpha_1-\lambda_1^tA_m}{\lambda_1^t(A_M-A_m)}\,p_M^t(x)+\frac{\lambda_1^tA_M-\alpha_1}{\lambda_1^t(A_M-A_m)}\,p_m^t(x)+r(x),$$

where $r(x)=\lambda_2^t\times$ (combination of $B_{x,m},B_{x,M},\alpha_1$).

B. Solution of Eq. (5.9), $Q(\alpha_1)=0$

We first analyze the coefficients of the trinomial in Eq. (5.8) and prove that each coefficient is $O(1-\lambda_1^t)$. In the matrix elements of T, we have the sum only on x, y such that $T_{xy}\neq 0$, that is, x and y are not in the same phase. The sum therefore takes pairs from three disjoint subsets:

$$(x\in I_M,y\notin I_M),\quad (x\in I_m,y\notin I_m),\quad (x\notin I_M\cup I_m,y\in X).$$

Thus we examine in turn $\langle A_0|Tp_0\rangle$ [Eq. (5.10)], $\langle A_0|Tp_1\rangle$ [Eqs. (5.11) and (5.12)], and $\langle A_1|Tp_1\rangle$ [Eq. (5.13)].

$$\langle A_0|Tp_0\rangle=\sum_{x\in I_M}\left(\sum_{y\notin I_M}R_{xy}p_0(y)\right)+\sum_{x\in I_m}\left(\sum_{y\notin I_m}R_{xy}p_0(y)\right)+\sum_{x\notin I_M\cup I_m}\sum_{y\in X}R_{xy}p_0(y)$$

$$\leqslant\sum_{y\notin I_M}p_0(y)+\sum_{y\notin I_m}p_0(y)+\sum_{x\notin I_M\cup I_m}p_0(x)=O(1-\lambda_1^t), \tag{5.10}$$

$$\langle A_0|Tp_1\rangle=\sum_{x\notin I_M}\sum_{y\in I_M}R_{xy}p_1(y)+\sum_{x\notin I_m}\sum_{y\in I_m}R_{xy}p_1(y)+\sum_{x\in X}\sum_{y\notin I_M\cup I_m}R_{xy}p_1(y). \tag{5.11}$$

We replace $p_1(y)=(p_M^t(y)-p_m^t(y))/(A_M-A_m)$,

$$\sum_{x\notin I_M}\sum_{y\in I_M} R_{xy}p_m^t(y)\leqslant\sum_x\sum_{y\in I_M} R_{xy}p_m^t(y)=\sum_{y\in I_M} p_m^t(y)=\mathrm{O}(1-\lambda_1^t),$$

$$\sum_{x\notin I_M}\sum_{y\in I_M} R_{xy}p_M^t(y)\leqslant\sum_{x\notin I_M}\sum_y R_{xy}p_M^t(y)=\sum_{x\notin I_M} p_M^{t+1}(y)=\mathrm{O}(1-\lambda_1^t),$$

$$\left|\sum_{x\in X}\sum_{y\notin I_M\cup I_m} R_{xy}p_1(y)\right|\leqslant\sum_{y\notin I_M\cup I_m}|p_1(y)|=\mathrm{O}(1-\lambda_1^t)$$

so

$$\langle A_0|Tp_1\rangle=\mathrm{O}(1-\lambda_1^t). \tag{5.12}$$

Finally we examine the case $\Sigma_{x,y}A_1(x)T_{xy}p_1(y)$,

$$\left|\sum_{x,y} A_1(x)T_{xy}p_1(y)\right|\leqslant\max|A_1|\sum_{x,y} T_{xy}|p_1(y)|\leqslant\frac{\max|A_1|}{A_M-A_m}\sum_{x,y} T_{xy}(p_M^t(y)+p_m^t(y))+\mathrm{O}(1-\lambda_1^t), \tag{5.13}$$

where we have used the fact that

$$p_1=\frac{p_M^t-p_m^t}{\lambda_1^t(A_M-A_m)}+\cdots .$$

The same kind of reasoning as above for the estimate of $\langle A_0|Tp_1\rangle$ proves that

$$\sum_{x,y} T_{xy}(p_M^t(y)+p_m^t(y))=\mathrm{O}(1-\lambda_1^t)$$

and so

$$\langle A_1|Tp_1\rangle,\langle A_1|Tp_0\rangle=\mathrm{O}(1-\lambda_1^t). \tag{5.14}$$

As a consequence, all the coefficients of the trinomial $Q(\alpha_1)$ in Eq. (4.8) are $\mathrm{O}(1-\lambda_1^t)$ and one can solve $Q(\alpha_1)=0$ with two roots α_1',α_1'' which are order unity.

C. Calculation of μ_1

Equation (5.6) gives two possible values, μ_1' and μ_1'', one for each root α_1' and α_1''. Because of Eqs. (5.10)–(5.13) and (5.6) one obtains

$$\mu_1',\mu_1''=\mathrm{O}(1-\lambda_1^t). \tag{5.15}$$

D. Calculation of q

The equation for q is given by

$$Rq-q=V+(T-\mu_1)q, \tag{5.16}$$

[see Eq. (5.4)] where V is given by

$$V=(T-\mu_1)p_0+\alpha_1(1-\lambda_1)p_1+\alpha_1(T-\mu_1)p_1. \tag{5.17}$$

We approximate Eq. (5.16) by

$$Rq-q=V. \tag{5.18}$$

The conditions (5.6) and (5.7), which fix α_1 and μ_1, imply $\langle A_0|V\rangle=0$ and $\langle A_1|V\rangle=0$. Thus

$$V=\sum_{k\geqslant 2} v_k p_k, \quad v_k=\langle A_k|V\rangle$$

and

$$q=\sum_{k\geqslant 2} \frac{v_k}{\lambda_k-1}\, p_k.$$

On the other hand, it is easy to see that

$$|v_k|=|\langle A_k|V\rangle|\leqslant C(1-\lambda_t^1), \tag{5.19}$$

where C does not depend on k (for the normalization, $\max|A_k|=1$). The proof of (5.19) is exactly the same as the proof of Eqs. (5.10)–(5.13), because for $k\geqslant 2$ $\langle A_k|V\rangle=\langle A_k|Tp_0\rangle+\alpha_1\langle A_k|Tp_1\rangle$. One then uses the estimate for Tp_0, Tp_1, as before in the case $k=1$ above.

Remark: We have proved that $q=O(1-\lambda_1^t)$. To have a rigorous perturbation theory one would like to prove that $(T-\mu_1)q\ll O(1-\lambda_1^t)$. We do have control over $\mu_1 q$, since we know it to be $O((1-\lambda_1^t)^2)$. However, we do not have rigorous bounds for Tq.

Conclusion: We have obtained for $\tilde{R}$ two stationary states with eigenvalues $1-\mu_1'$ and $1-\mu_1''$, with μ_1' and $\mu_1''=O(1-\lambda_1^t)$. These have corresponding eigenstates $p_0+\alpha_1' p_1$ and $p_0+\alpha_1'' p_1$, where α_1', α_1'' are $O(1)$.

Remark: We expect the eigenstates of largest eigenvalue to look substantially like p_M and p_m and the corresponding left eigenvectors to look like the characteristic functions of I_M and I_m, except near the boundary.[24]

VI. EXIT TIME FROM ONE OF THE PHASE REGIONS

Let $T_M(x)$ be the average of the first exit time, starting from $x\in I_M$, of the phase region I_M

$$T_M(x)=E\{\inf\{t|X_t\notin I_M\}|X_0=x\},$$

where E is the expectation for the R process. It is easy to see that

$$T_M(x)=\begin{cases} 1+\sum_{y\in X} T_M(y)R_{yx} & \text{for } x\in I_M \\ 0 & \text{for } x\notin I_M \end{cases}. \tag{6.1}$$

The first line is derived by considering a single time step. The advantage of looking at the quantity $T_M(x)$ is that it is fairly constant inside I_M, except near the boundary of I_M, where it drops quickly to zero. One can also define a function $T_m(x)$ relative to the other phase I_m. One can then expand T_M, on the left eigenstates $\tilde{A}_{M,k}$ (eigenvalue $\tilde{\lambda}_{M,k}$) of $\tilde{R}$ relative to the phase I_M

$$T_M=\sum_{k\geqslant 0} \theta_{M,k}\tilde{A}_{M,k}, \tag{6.2}$$

where $\theta_{M,k}=\langle T_M|\tilde{p}_{M,k}\rangle$ and $\tilde{p}_{M,k}$ is the right eigenstate of $\tilde{R}$ relative to the phase I_M. In particular $\theta_{M,0}=\langle T_M|\tilde{p}_{M,0}\rangle>0$.

From Eq. (6.1), we see that

$$\theta_{M,0}\sim \frac{C}{1-\tilde{\lambda}_{M,0}}. \tag{6.3}$$

On the other hand, we have seen in Sec. IV, that

$$\tilde{\lambda}_{M,0}\sim 1-O(1-\lambda_1^t). \tag{6.4}$$

This means that

$$T_M \sim \mathrm{O}\left(\frac{1}{1-\lambda_1^t}\right). \tag{6.5}$$

Remark: Note that this is not a uniform estimate. T_M is in fact a function of $x \in X$ and for x near the boundary[24] we expect the otherwise almost constant $\tilde{A}(x)$ to vary, in particular to be small.

VII. DOUBLE CONDITIONING

In a previous publication[19] we found that a metastable phase could become fully well-defined by a method using two-time conditioning. In this method, the dynamical persistence of the metastable state, its characteristic feature, was the essential property that was used to define the concept of metastability. (This contrasts to analyticity oriented approaches to this problem). Inasmuch as our fundamental method in the present work (which follows Refs. 1 and 2) is dynamical, it is reasonable to expect the two-time conditioning approach to be applicable to the definition of phases of our R-degenerate system. We now show this to be the case.

For $x \in I_M$ define

$$\begin{aligned}\Theta_M(x,T) &= \mathrm{Prob}(X_T = x | [X_t \in I_M \text{ for all } 0 \leqslant t \leqslant 2T] \text{ and } [X_0 = M]) \\ &= \frac{\mathrm{Prob}(X_T = x,\ X_t \in I_M \text{ for all } 0 \leqslant t \leqslant 2T | X_0 = M)}{\mathrm{Prob}(X_t \in I_M \text{ for all } 0 \leqslant t \leqslant 2T | X_0 = M)}.\end{aligned} \tag{7.1}$$

The numerator is

$$\begin{aligned}&\mathrm{Prob}(X_T = x \text{ and } X_t \in I_M \forall t \leqslant T | X_0 = M) \\ &\quad \times \sum_{y \in I_M} \mathrm{Prob}(X_{2T} = y,\ X_t \in I_M \text{ for } T < t \leqslant 2T | X_T = x),\end{aligned}$$

so that

$$\Theta(x,T) = \frac{\tilde{R}^T_{xM} \Sigma_y \tilde{R}^T_{yx}}{\Sigma_z \tilde{R}^{2T}_{zM}}. \tag{7.2}$$

But for points in the M phase, we can expand

$$\tilde{R}_{xy} = \tilde{\lambda}_{M,0}\, \tilde{p}_{M,0}(x)\tilde{A}_{M,0}(y) + \cdots + \tilde{\lambda}_{M,k}\, \tilde{p}_{M,k}(x)\tilde{A}_{M,k}(y) + \cdots .$$

For $T \to \infty$, we expect that

$$\Theta(x,T) \sim \tilde{p}_{M,0}(x)\tilde{A}_{M,0}(x). \tag{7.3}$$

In particular, since $\Sigma \tilde{p}_{M,0}(x)\tilde{A}_{M,0}(x) = 1$, $\theta(x,T)$ is normalized.

Although we do not supply a rigorous proof of Eq. (7.3) we expect it to be true because the largest ''λ''s (excluding the single one for each phase near unity) will correspond to the relaxation times within the phases, namely, what we have been calling λ_2, and eigenvalues of yet smaller norm. As usual, in appropriate (or perhaps, inappropriate) departures from the intended physical situation (as in the thermodynamic limit described in Ref. 6) unphysical modes may be introduced by mathematical idealizations.

VIII. CONVERSE: PHASE TRANSITION IMPLIES DEGENERACY

We have shown that when a stochastic matrix is nearly degenerate (and satisfies certain other conditions) it will, to a good approximation, describe a system with two distinct stationary states. These stationary states are in many ways like the phases on either side of a first-order phase transition, but (unlike analyticity-based descriptions of phase transitions) have a natural place for the metastable phase. In the present section we deal with the converse: Suppose there is what one

might call, literally and by analogy, a phase transition. This would occur for two stochastic matrices that differ only slightly but whose respective invariant states are very different. What we show next is that this implies that for both these stochastic matrices the largest eigenvalue is nearly degenerate. This form of converse is not specific to first order phase transitions, and may apply to certain parameter variations about critical points. [Note that we only show near degeneracy of the largest eigenvalues, not that successive eigenvalues are distant in the sense of Eq. (3.1).]

Two stochastic matrices, R and R', are given, with

$$R' = R + V \tag{8.1}$$

so that $\Sigma_x V_{xy} = 0$. We denote by p_0 and p_0' the stationary states of R and R', and assume that both $p_0 > 0$ and $p_0' > 0$ everywhere. By the irreducibility of both R and R' we still have $A_0 = \chi_X = A_0'$. We denote by $(|)_{p_0}$ the scalar product

$$(w|v)_{p_0} = \sum_x p_0(x) w(x) v(x) \tag{8.2}$$

and by $(|)_{p_0'}$ the scalar product with respect to p_0'.

A. Hypotheses

(1) We assume that there exist disjoint subsets S and S' of X such that

$$\sum_{x \in S} p_0(x) = 1 - \mu, \quad \sum_{x \in S'} p_0'(x) = 1 - \mu' \tag{8.3}$$

and $S \cup S' = X$. It follows that

$$\sum_{x \in S} p_0'(x) = \mu', \quad \sum_{x \in S'} p_0(x) = \mu,$$

where μ and μ' are small.

(2) To say that a quantity (vector, operator) is small means that the norms, $\| \|_{p_0}$ and $\| \|_{p_0'}$, should be small. We assume that

$$\|V\|_{p_0}, \|V\|_{p_0'} \leqslant \delta, \tag{8.4}$$

where $\|V\|_{p_0}$ is the norm of the V matrix for the scalar product defined by p_0.

We want to prove that under these hypotheses R and R' have a near degeneracy in the sense that the eigenvalue of first excited state of R (respectively R') is λ_1 (respectively λ_1') with

$$1 - \lambda_1, 1 - \lambda_1' = \mathrm{O}(\mu, \mu', \delta).$$

We shall prove this for R'. We have to show that there exists a left vector $\hat{A}'$ such that

$$\hat{A}' \text{ is } \mathrm{O}(1) \text{ in the norm } \|\|_{p_0}, \tag{8.5}$$

$$\langle \hat{A}' | p_0' \rangle = 0, \tag{8.6}$$

and

$$\|\hat{A}'(1 - R')\|_{p_0}, \|\hat{A}'(1 - R')\|_{p_0'} = \mathrm{O}(\mu, \mu', \delta). \tag{8.7}$$

Before establishing the relations (8.5)–(8.7), we show that they in turn imply that R' has an eigenvalue (besides unity) whose difference from 1 is $\mathrm{O}(\mu, \mu', \delta)$.

More generally, let us assume that in a vector space E there are two vectors $x, y \in E$ and a matrix M such that

$$(M-I)x=y,$$

$$(M-I) \quad \text{is invertible,}$$

$$\|x\|=\mathrm{O}(1), \quad \|y\|=\mathrm{O}(\epsilon).$$

Then $x=(M-I)^{-1}y$ and it follows that the norm of $(M-I)^{-1}$ is of order $1/\epsilon$. By taking a Jordan expansion of M, we see that M has an eigenvalue within ϵ of 1.

In our situation, we choose for the vector space E the space of vectors A such that $\langle A|p_0\rangle=0$ and $M={}^t(R')$ (thereby satisfying the second condition above on E, using also the irreducibility of R). We take for ''x'' and ''y'' the vectors A and $(M-I)A$, respectively. Note too with respect to our upcoming application that it is sufficient to establish these conditions with respect to any particular norm, and for this purpose we have in mind the p_0 norm.

We now turn to establishing the conditions (8.5)–(8.7). Choose

$$\hat{A}'=\mu'\chi_{S'}-(1-\mu')\chi_S \tag{8.8}$$

so that Eq. (8.6) is satisfied and condition (8.5) is satisfied [for example, $\|\hat{A}\|_{p_0}=\mathrm{O}(1)$]. Now

$$-\hat{A}'(1-R')=-\mu'(\chi_{S'}-\chi_{S'}R')+(1-\mu')(\chi_S-\chi_SR)-(1-\mu')\chi_SV. \tag{8.9}$$

We also notice that $\chi_S+\chi_S{}'\equiv\chi_X\equiv A_0$ so that

$$\chi_S(1-R)=-\chi_{S'}(I-R). \tag{8.10}$$

B. Proof that $\|\chi_S-\chi_SR\|_{p_0}^2=\mathrm{O}(\mu)$

Using Eq. (8.10) we obtain

$$\langle\chi_S-\chi_SR|\chi_S-\chi_SR\rangle_{p_0}=-\langle\chi_{S'}(I-R)|\chi_S\rangle_{p_0}+\langle\chi_{S'}(I-R)|\chi_SR\rangle_{p_0}. \tag{8.11}$$

(1°) $\langle\chi_{S'}(I-R)|\chi_S\rangle_{p_0}=\mathrm{O}(\mu)$.

This is

$$\sum_{\substack{y\in S'\\ x\in S}}(\delta_{yx}-R_{yx})p_0(x)=-\sum_{\substack{y\in S'\\ x\in S}}R_{yx}p_0(x).$$

But

$$\sum_{\substack{y\in S'\\ x\in S}}R_{yx}p_0(x)\leqslant\sum_{\substack{y\in S'\\ x\in X}}R_{yx}p_0(x)=\sum_{y\in S'}p_0(y)=\mathrm{O}(\mu).$$

(2°) $\langle\chi_{S'}(I-R)|\chi_SR\rangle_{p_0}=\mathrm{O}(\mu)$.

This is

$$\sum_{y,z}p_0(y)\chi_S(z)R_{zy}\left(\chi_{S'}(y)-\sum_x\chi_{S'}(x)R_{xy}\right).$$

Now

$$\sum_{y,z}p_0(y)\chi_S(z)R_{zy}\chi_{S'}(y)\leqslant\sum_{z,y}R_{zy}\chi_{S'}(y)p_0(y)=\sum_{y\in S'}p_0(y)=\mathrm{O}(\mu),$$

$$\sum_{y,z,x}p_0(y)\chi_S(z)R_{zy}\chi_{S'}(x)R_{xy}\leqslant\sum p_0(y)R_{zy}\chi_{S'}(x)R_{xy}$$

$$=\sum p_0(y)\chi_{S'}(x)R_{xy}=\sum\chi_{S'}(x)p_0(x)=\mathrm{O}(\mu).$$

From (8.11) we deduce

$$\|\chi_S - \chi_S R\|_{p_0}^2 = O(\mu). \tag{8.12}$$

In the same way, we would obtain

$$\|\chi_{S'} - \chi_{S'} R'\|_{p_0'}^2 = O(\mu'). \tag{8.13}$$

C. Proof that $\|\chi_S - \chi_S R\|_{p_0'}^2 = O(\mu', \delta)$.

We use Eq. (8.10)

$$\|\chi_S - \chi_S R\|_{p_0'} = \|\chi_{S'} - \chi_{S'} R\|_{p_0'}.$$

Then we use $R' = R + V$.

$$\langle \chi_{S'} - \chi_{S'} R | \chi_{S'} - \chi_{S'} R \rangle_{p_0'} = \|\chi_{S'} - \chi_{S'} R'\|_{p_0'}^2 - 2\langle \chi_{S'} V | \chi_{S'}(I - R') \rangle_{p_0'} + \|\chi_{S'} V\|_{p_0'}^2.$$

Now our hypothesis on V says that

$$\|\chi_{S'} V\|_{p_0'}^2 \leqslant \|V\|_{p_0'}^2 \|\chi_{S'}\|_{p_0'}^2 = O(\delta^2). \tag{8.14}$$

Then

$$|\langle \chi_{S'} V | \chi_{S'}(I - R') \rangle_{p_0'}| \leqslant \|\chi_{S'} V\|_{p_0'} \|\chi_{S'}(I - R')\|_{p_0'} \leqslant C\delta\sqrt{\mu'}. \tag{8.15}$$

[We used Eqs. (8.13) and (8.14).] So from Eqs. (8.13), (8.15), and (8.14), we deduce

$$\|\chi_S - \chi_S R\|_{p_0'} = O(\mu', \delta). \tag{8.16}$$

D. Analysis of $\hat{A}'(I - R')$

We come back to Eq. (8.9) and study both norms $\|\hat{A}'(I - R')\|_{p_0}$ and $\|\hat{A}'(I - R')\|_{p_0'}$. The first term of Eq. (8.9) is already of order μ'. The second term is small because of Eq. (8.12) or (8.16). The last term is $\|\chi_S V\|_{p_0}$ or $\|\chi_S V\|_{p_0'}$, which is small by our hypotheses on V. This proves Eq. (8.7).

IX. SEVERAL PHASES

We now assume that the first p eigenvalues (after $\lambda_0 = 1$), $\lambda_1, \ldots, \lambda_p$, of R are real and satisfy

$$\lambda_1^t, \ldots, \lambda_p^t = O(1 - \epsilon_1)$$

but

$$\lambda_{p+1}^t \text{``} B^t \text{''} < \epsilon_2 \quad [\text{as in Eq. (3.10)}].$$

We further assume that $1 - \lambda_k^t$, $(1 \leqslant k \leqslant p)$, are of the same order of magnitude. The tth power of R can now be written [cf. Eq. (3.3)]

$$\begin{aligned} R^t &= |p_0\rangle\langle A_0| + \sum_{k=1}^{p} \lambda_k^t |p_k\rangle\langle A_k| + |\lambda_{p+1}^t| B^t, \\ B^t &= \sum_{k \geqslant p+1} \frac{\lambda_k^t}{|\lambda_{p+1}^t|} |p_k\rangle\langle A_k|. \end{aligned} \tag{9.1}$$

For any $y \in X$, $x \in X$ we define

$$p_y^t(x) = R_{xy}^t \tag{9.2}$$

so that p_y^t is a probability on x. For the two-phase case, the quantities A_M and A_m [cf. Eq. (3.5)] played an important role in our estimates. We must now find the corresponding quantities in the multiphase case. We represent the situation geometrically in $\mathbf{R}^p$: for any point $y \in X$ we consider the vector $(A_1(y),\ldots,A_p(y)) \in \mathbf{R}^p$ so that we obtain a set $\mathscr{A}$ of N vectors in $\mathbf{R}^p$, where N is the cardinality of X. We consider the convex envelope $\widehat{\mathscr{A}}$ of $\mathscr{A}$.

The first remark is that the vector (0,0,...,0) *is in this convex envelope* $\widehat{\mathscr{A}}$.

In fact we know that $\langle A_k | p_0 \rangle = 0$ for $1 \leqslant k \leqslant p$, so this means

$$\sum_{y \in X} p_0(y)(A_1(y),A_2(y),\ldots,A_p(y)) = (0,0,...,0)$$

and (0,...,0) is a convex combination of the vectors of $\mathscr{A}$. As a consequence, one can find $p+1$ points y_l^*, $1 \leqslant l \leqslant p+1$, which are extremal points of the convex set $\mathscr{A}$ and such that (0,0,...,0) is a convex combination of the vectors E_l,

$$E_l \equiv (A_1(y_l^*),A_2(y_l^*),\ldots,A_p(y_l^*)). \tag{9.3}$$

There may be several ways to choose these points y_l^*, but we shall prove that, in fact, the resulting vectors E_l $(1 \leqslant l \leqslant p+1)$ are uniquely defined. In particular, we can find μ_l, $1 \leqslant l \leqslant p+1$, with $0 \leqslant \mu_l \leqslant 1$, $\Sigma_{l=1}^{p+1} \mu_l = 1$ such that

$$\Sigma \mu_l E_l = 0 \quad \text{in} \quad \mathbf{R}^p. \tag{9.4}$$

Now, by definition, we have

$$\lambda_k^t A_k(y_l^*) = \langle A_k | p_{y_l^*}^t \rangle \tag{9.5}$$

so that for all l we have

$$(\lambda_1^t A_1(y_l^*),\lambda_2^t A_2(y_l^*),\ldots,\lambda_p^t A_p(y_l^*)) = \sum_{y \in X} p_{y_l^*}^t(y)(A_1(y),A_2(y),\ldots,A_p(y)). \tag{9.6}$$

Because $0 \leqslant \lambda_k < 1$, the vector on the left side of Eq. (9.6) is in the convex set $\widehat{\mathscr{A}}$; moreover, its distance to the extremal vector $E_l = (A_1(y_l^*,\ldots,A_p(y_l^*))$ is less than $1-\lambda_p^t$. Equation (9.6) says that this vector is a convex combination of the vectors of the set $\mathscr{A}$. Because $\Sigma_y p_{y_l^*}^t(y) = 1$, we have from Eq. (9.6)

$$E_l - (\lambda_1^t A_1(y_l^*),\lambda_2^t A_2(y_l^*),\ldots,\lambda_p^t A_p(y_l^*)) = \sum_y p_{y_l^*}^t(y)[E_l - (A_1(y),\ldots,A_p(y))]. \tag{9.7}$$

Now, consider the translated set $E_l - \mathscr{A}$ in $\mathbf{R}^p$, i.e., the set of vectors $E_l - (A_1(y),\ldots,A_p(y))$, $y \in X$. Then, the vector 0 is extremal for the convex set $E_l - \widehat{\mathscr{A}}$. We can transform our problem in the following way:

(i) We have a finite set of vectors $w(y)$ in $\mathbf{R}^p$ $(y \in X)$ such that the vector 0 is an extremal point of the convex envelope of this set.

(ii) We have a vector w_0 of norm $\|w_0\| < \epsilon$ which lies in the convex envelope. Here

$$w(y) = E_l - (A_1(y),\ldots,A_p(y)),$$

$$w_0 = E_l - (\lambda_1^t A_1(y),\ldots,\lambda_p^t A_p(y)),$$

$$\epsilon = 1 - \lambda_p^t.$$

We have a convex relation, with $0 \leqslant q(y) \leqslant 1$, $\Sigma q(y) = 1$

$$w_0=\sum_{y\in X} q(y)w(y). \tag{9.8}$$

Because 0 is extremal for the set of $w(y)$, there exist p linear forms $h_1,\ldots,h_p$ such that $h_j(w(y))>0$, except for y such that $w(y)=0$, and with coefficients $\leqslant 1$. Define

$$I(a)=\{y\in X | h_j(w(y))<a \quad \text{for all} \quad j=1,...,p\}. \tag{9.9}$$

Then from (9.8) we have

$$\sum_{y\notin I(a)} q(y)<\frac{p\epsilon}{a}.$$

In fact from (9.8) and the fact that h_j has coefficients $\leqslant 1$

$$\epsilon>h_j(w_0)>\sum_{y\in X} q(y)h_j(w(y))\geqslant a\sum_{\{y\in X|h_j(w(y))>a\}} q(y).$$

Thus

$$\sum_{y\notin I(a)} q(y)\leqslant\sum_{j=1}^{p}\left(\sum_{\{y\in X|h_j(w(y))>a\}} q(y)\right)\leqslant\frac{p\epsilon}{a}.$$

Coming back to our previous notation, we see that for any l, $1\leqslant l\leqslant p+1$, there exist p linear forms $h_{l,j}$ with coefficients $\leqslant 1$, such that if we define

$$I_l(a)=\{y\in X | h_{l,j}(E_l-(A_1(y),\ldots,A_p(y)))<a, j=1,...,p\} \tag{9.10}$$

then

$$\sum_{y\in I_l(a)} p^t_{y^*_l}(y)\geqslant 1-\frac{p}{a}(1-\lambda^t_p). \tag{9.11}$$

Moreover, we had certain $0\leqslant\mu_l\leqslant 1$, with $\Sigma\mu_l=1$, such that

$$\Sigma\mu_l E_l=1 \quad [\text{see Eq. (9.4)}].$$

This means that by (9.1), (9.2), and the definition of E_l

$$\sum_{l=1}^{p+1}\mu_l p^t_{y^*_l}(x)=p_0(x)+|\lambda^t_{p+1}|\sum_{l=1}^{p+1}\mu_l B^t_{xy^*_l}. \tag{9.12}$$

Now we assume that the B terms are small, i.e., for any $Y\subset X$ and all $x\in X$

$$|\lambda^t_{p+1}|\left|\sum_{y\in Y} B^t_{yx}\right|\ll 1-\lambda^t_p. \tag{9.13}$$

Then Eq. (9.12), and inequalities (9.11), (9.13) say that for some constant c, we have

$$\sum_{y\in\cup I_l(a)} p_0(y)\geqslant 1-c(1-\lambda^t_p). \tag{9.14}$$

The inequalities (9.11) and (9.14) say that the whole state space X breaks into subsets $I_l(a)$ which fill almost everything (with respect to p_0). This means that we have $p+1$ phases.

ACKNOWLEDGMENTS

We wish to thank Phil Pechukas, Vladimir Privman, and Leonard Schulman for helpful discussions. L.S.S. thanks the University of Paris VI and Columbia University for their kind hospitality.

This work was supported in part by the United States National Science Foundation Grant No. PHY 93 16681.

[1] B. Gaveau and L. S. Schulman, ''Master equation based formulation of non-equilibrium statistical mechanics,'' J. Math. Phys. **37**, 3897–3932 (1996).
[2] B. Gaveau and L. S. Schulman, ''A general framework for non-equilibrium phenomena: The master equation and its formal consequences,'' Phys. Lett. A **229**, 347–353 (1997).
[3] G. Roepstorff and L. S. Schulman, ''Metastability and Analyticity in a Droplet-like model,'' J. Stat. Phys. **34**, 35–56 (1984).
[4] S. N. Isakov, ''Nonanalytic Features of the First Order Phase Transition in the Ising Model,'' Commun. Math. Phys. **95**, 427–443 (1984).
[5] U. Vazirani, ''Rapidly Mixing Markov Chains,'' in Proc. Symposia in Appl. Math. (American Mathematical Society, Providence, 1991), Vol. 44, pp. 99–121.
[6] L. S. Schulman, ''System-size Effects in Metastability,'' in *Finite Size Scaling and Numerical Simulation of Statistical Systems*, edited by V. Privman (World Scientific, Singapore, 1990), pp. 490–518.
[7] M. Kac, ''Toward a Unified View on Mathematical Theories of Phase Transitions,'' in *Fundamental Problems in Statistical Mechanics II*, edited by E. G. D. Cohen (North-Holland, Amsterdam, 1968), pp. 71–105.
[8] M. Kac, ''Mathematical Mechanisms of Phase Transitions,'' in *Statistical Physics: Phase Transitions and Superfluidity*, edited by M. Chretien, E. P. Gross, and S. Deser, Brandeis Summer Institute in Theoretical Physics, 1966, Vol. 1 (Gordon and Breach, New York, 1966), pp. 242–305.
[9] C. J. Thompson and M. Kac, ''Phase Transition and Eigenvalue Degeneracy of a One Dimensional Anharmonic Oscillator,'' Stud. Appl. Math. **48**, 257–264 (1969).
[10] J. Ashkin and W. E. Lamb, Jr., Phys. Rev. **64**, 159 (1943).
[11] M. Cassandro, A. Galves, E. Olivieri, and M. E. Vares, ''Metastable Behavior of Stochastic Dynamics: A Pathwise Approach,'' J. Stat. Phys. **35**, 603–634 (1984).
[12] M. Kalos, J. L. Lebowitz, O. Penrose, and A. Sur, ''Clusters, Metastability, and Nucleation: Kinetics of First-Order Phase Transitions,'' J. Stat. Phys. **18**, 39–52 (1978).
[13] J. L. Lebowitz and O. Penrose, ''Rigorous Treatment of the Van Der Waals–Maxwell Theory of the Liquid-Vapor Transition,'' J. Math. Phys. **7**, 98–113 (1966).
[14] F. Martinelli, E. Olivieri, and E. Scoppola, ''Metastability and Exponential Approach to Equilibrium for Low Temperature Stochastic Ising Models,'' J. Stat. Phys. **61**, 1105–1119 (1990).
[15] C. M. Newman and L. S. Schulman, ''Metastability and the Analytic Continuation of Eigenvalues,'' J. Math. Phys. **18**, 23–30 (1977).
[16] C. M. Newman and L. S. Schulman, ''Complex Free Energies and Metastable Lifetimes,'' J. Stat. Phys. **23**, 131–148 (1980).
[17] O. Penrose and J. L. Lebowitz, ''Towards a Rigorous Molecular Theory of Metastability,'' in *Fluctuation Phenomena*, edited by L. Montroll and J. Lebowitz (North-Holland, Amsterdam, 1987), pp. 323–375.
[18] B. Gaveau and L. S. Schulman, ''Metastable Decay Rates and Analytic Continuation,'' Lett. Math. Phys. **18**, 201–208 (1989).
[19] B. Gaveau and L. S. Schulman, ''Dynamical Metastability,'' J. Phys. A **20**, 2865–2873 (1987).
[20] B. Gaveau and L. S. Schulman, ''Finite-size scaling for mean-field percolation,'' J. Stat. Phys. **70**, 613–634 (1993).
[21] The ''fundamental intuition'' of phase coexistence has been incorporated in other definitions of phase transition, some of them stemming from the C* algebra approach (Refs. 25 and 26). In this language a phase transition is expressed by the nonuniqueness of the infinite volume Gibbs measure. As indicated, we do not know how to bring infinite volume (the thermodynamic limit) into so mild a form of nonequilibrium as metastability, nor do we think that such a limit best reflects the physics (cf. Ref. 6).
[22] The ''remainder'' term in Eq. (3.10) (''B^t'') must be small in some sense for our phases to be well-defined, but the measure of smallness in that equation can probably be replaced by something weaker.
[23] The form for B^t in this equation is adopted for convenience only. In general B^t need *not* be writable as a sum of one-dimensional projections and a Jordan form would be required. See the remark following Eq. (3.4).
[24] By ''boundary'' of (say) I_M, we mean points $x \in I_M$ such that their distance, in terms of R, from any point in $\mathcal{C}I_M$ is small.
[25] D. Ruelle, *Statistical Mechanics: Rigorous Results* (Benjamin, New York, 1969).
[26] D. Ruelle, ''The C*-algebra approach to statistical mechanics,'' in *Statistical Mechanics at the Turn of the Decade*, edited by E. G. D. Cohen (Marcel Dekker, New York, 1971), p. 67–79.

A stochastic method for Brownian-like optical transport calculations in anisotropic biosuspensions and blood

Steven Miller[a)]
Bioengineering Unit, University of Strathclyde, Glasgow G4ONW, Scotland

(Received 27 January 1997; accepted for publication 6 October 1997)

A generic stochastic method is presented that rapidly evaluates numerical bulk flux solutions to the one-dimensional integrodifferential radiative transport equation, for coherent irradiance of optically anisotropic suspensions of nonspheroidal bioparticles, such as blood. As Fermat rays or geodesics enter the suspension, they evolve into a bundle of random paths or trajectories due to scattering by the suspended bioparticles. Overall, this can be interpreted as a bundle of Markov trajectories traced out by a ''gas'' of Brownian-like point photons being scattered and absorbed by the homogeneous distribution of uncorrelated cells in suspension. By considering the cumulative vectorial intersections of a statistical bundle of random trajectories through sets of interior data planes in the space containing the medium, the effective equivalent information content and behavior of the (generally unknown) analytical flux solutions of the radiative transfer equation rapidly emerges. The fluxes match the analytical diffuse flux solutions in the diffusion limit, which verifies the accuracy of the algorithm. The method is not constrained by the diffusion limit and gives correct solutions for conditions where diffuse solutions are not viable. Unlike conventional Monte Carlo and numerical techniques adapted from neutron transport or nuclear reactor problems that compute scalar quantities, this vectorial technique is fast, easily implemented, adaptable, and viable for a wide class of biophotonic scenarios. By comparison, other analytical or numerical techniques generally become unwieldy, lack viability, or are more difficult to utilize and adapt. Illustrative calculations are presented for blood medias at monochromatic wavelengths in the visible spectrum. © *1998 American Institute of Physics.* [S0022-2488(98)00102-9]

I. INTRODUCTION

In this paper we present an approach to the mathematical problem of the scattering of a coherent light beam incident upon a slab or half-plane comprised of scattering–absorbing bioparticles in liquid suspension such as blood. Finding adequate mathematical solutions to this problem is relevant to a variety of technical biophysical and laser biology applications.[1–6] Part of the general mathematical difficulty arises from the complex optical and geometric properties of such medias. The shape, size, and number density of a homogeneous suspension of bioparticles, such as red blood cells, affects the rate of scattering and diffusion of the incident beam, the interior scattered flux profiles, and the scattered reflectance and transmittance. Blood is essentially a suspension of biconcave (dielectric) discoid cells or erythrocytes, each of which contains differing proportions or ratios of haemoglobins. These are haemoglobin and oxyhaemoglobin (Hb and HbO_2, respectively). The number density of suspended cells, their spatial geometry, and their inherent absorptive characteristics are therefore variables that can alter the optical properties in a complex way (red cells completely dominate the optical properties). Increasing the number of cells in suspension increases scattering since more scattering centers are now presented to an incident beam, but optical absorbancy is simultaneously increased since the relative amount of absorbing haemoglobins present is also increased. Furthermore, haemoglobin and oxyhaemoglobin have different optical absorbancy characteristics at the same optical wavelength.[7–9]

[a)]Correspondence address: Steven Miller, Ph.D., 142 Cameron Drive, Kilmarnock KA3 7PL, Scotland, United Kingdom, electronic mail: 101551.1243@Compuserve.com

 © 1998 American Institute of Physics

In most problems of multiple light scattering and diffusion in suspensions of particulate medias, the wavelength of the light λ, the scatter mean-free path l, and the thickness d of the sample obey the inequality $\lambda \ll l \ll d$. Also, if L is the effective dimension of the scatterer (e.g., diameter), then we will have either $L \ll \lambda, L \sim \lambda$, or $L \gg \lambda$. Multiple scattering of scalar waves in such particulate medias can be described by the following levels of theoretical description.

(i) The macroscopic approach[10] resides in solving an effective (Helmholtz) diffusion equation, which gives spatiotemporal variations in the scalar diffuse intensity or fluence $\phi(\mathbf{x},t)$ at points $\mathbf{x}$ at time t, or $\phi(\mathbf{x})$ in the steady state. This approximation is excellent in the bulk of the medium on length scales much larger than the mean-free path l, and when the particles scatter strongly and isotropically with little or no absorption. However, it is invalid in dealing with subsurface or ''quasidiffuse'' light scattered to low-order, strong anisotropy, coherent sources and strong absorption and combinations of these.

(ii) The mesoscopic approach[10–12] utilized in astrophysics and nuclear engineering, is based on radiative transfer theory (RTT). Spatiotemporal variations in radiance or intensity $\psi(\mathbf{x},\hat{\boldsymbol{\omega}},t)$ at all points $\mathbf{x}$ within the medium from propagation direction $\hat{\boldsymbol{\omega}}$, are described by an integrodifferential balance equation, structurally similar to the Boltzmann equation of kinetic theory. The approach is essentially heuristic and deals directly with the transport of scattered energy or power. The formulation is flexible and can be applied to a wide variety of physical and technological problems.[10] However, the general solutions are not known, even in one dimension; most approaches to the equation are therefore a compromise between mathematical and computational tractability and accuracy. The macroscopic diffusion approach is recovered in the limit of length scales larger than l when the particles are highly scattering.

(iii) The microscopic or analytical approach begins with Maxwell's equations or wave equations and the scattering–absorption characteristics of particles. Appropriate differential or integral equations are developed for statistical quantities such as correlation functions and field variances. This approach has a mathematically rigorous structure and can, in principle, include all possible wave related phenomenology.[10] However, the various theories that can provide solutions are all approximate, each being viable within specific ranges of parameters. The diagram method, the Dyson and Bethe–Salpeter equations, and Twersky's theory are theories of this type.[10] The approach is compatible with (ii) and can recover the RTE.

Each of these viewpoints represents a qualitative improvement on the previous one and are applicable to biosuspensions and blood medias. Twersky[13] developed a mathematical formalism as in (iii) to describe scattering of plane electromagnetic waves by large, low refracting, and absorbing particles. Anderson and Sekelj[14] demonstrated the application of this theory to a blood cell suspension. Twersky[15] later applied his own model to biological suspensions. However, the inherent complexity of the theory precludes the possibility of any pragmatic solutions for a biophysical application. It is possible to obtain an optical density formula for scattering in whole blood, but the coefficients within the equation have to empirically fitted from experimental data.[13] The more tractable alternative for blood and biosuspension is via the radiative transfer or diffusion approaches of (i) and (ii). However, again we have the difficulty that general solutions of the transport equation in closed form are unknown, even in one-dimensional geometry.[10,11] Analytical and numerical solutions are possible in the diffusion approximation of (ii) when the medium is isotropic and thick and the suspended particles are assumed to be more scattering than absorbing, however, these conditions do not always generally hold for biosuspensions or blood.

Cohen and Longini[16] initially applied similar principles of semiconductor electron diffusion to the analogous optical diffusion problem in a turbid blood medium. Johnson,[17] Reynolds,[18] and Turnbull and Barbenel,[19,20] developed the same diffusion theory for boundaries and geometries relevant to their biophysical experiments, leading to complex expressions with no analytical solutions. Mathematical techniques included the SDD model (surface distributed dipole source) and the IDCT models (internal distributed cylindrical column of twin pole sources). A concise review is presented by Barbenel.[20] Recent mathematical work in bio-optics and biophotonics has focused on numerically solving or simulating the transport and diffusion equations of approaches (i) and (ii) directly borrowing techniques utilized in nuclear engineering and astrophysics.[1] These include discrete ordinates, Gauss quadrature, higher-order spherical harmonic P_n approximations, finite

element diffusion methods, and standard Monte Carlo simulation methods adapted from neutron diffusion theory.[1,21–24] However, while these techniques are viable and effective for many problems, all such methods generally require sophisticated application software and extensive computational or supercomputer facilities. This is often superfluous and detrimental when one essentially requires simple and fast but accurate quantitative flux or dosimetry solutions that are relevant to a variety of experimental or biological contexts.

In this paper, I develop a simple, yet effective and practical generic stochastic based mathematical approach that encompasses the regimes described by (i) and (ii), for the case of homogeneous suspensions of uncorrelated, anisotropic scattering and absorbing bioparticles. The method is specifically demonstrated for the case of whole blood and red cell suspensions. Given the basic scattering and absorbing properties of the cells (which are known), the technique rapidly ascertains the steady-state numerical solutions for scattered radiant flux, arising from wide beam coherent irradiance of the sample. The information content is equivalent to that which would be provided by the general analytical solutions of the RTE, which are unknown. In particular, the generation and decay of anisotropic multiply scattered light, arising from coherent beam decay, can be studied in detail for all possible optical conditions and for thin samples or films where the diffusive limit is generally not viable. Unlike conventional numerical and Monte Carlo algorithms for anisotropic transport,[22] this alternative formalism can be incorporated into very compact and simple computer codes, with many accurate and practical computations being completed rapidly. For the class of biophysical problems considered, the method has some advantages over all other methods.

II. CROSS SECTIONS AND OPTICAL COEFFICIENTS FOR BIOPARTICLES AND BLOOD CELLS

We consider a homogeneous and isotropic liquid suspension of N red blood cells or erythrocytes, which are uncorrelated and randomly oriented. The cells can be viewed as nonspheroidal or biconcave dielectric discoids, which scatter light due a refractive index continuity at their surface, separating the interior from the surrounding (transparent) plasma. The suspension is irradiated with a wide coherent monochromatic beam in the visible spectrum. Since $L \sim 8\ \mu$m and λ is normally 700–660 nm, $L \gg \lambda$. We first consider the scattering and absorption characteristics of a single erythrocyte, then consider the cumulative macroscopic or bulk optical properties arising statistically from many cells in suspension. The generic properties of scattering particles have been exhaustively presented in the literature.[10,25–27]

To make this paper self-contained and as a necessary prerequisite for the development of the method, we briefly summarize and define what is directly relevant for generic bioparticles and blood cells. If a linearly polarized, monochromatic, electromagnetic wave propagates in a medium with dielectric constant ϵ_0, then the electric field is $\mathbf{E}_i(\mathbf{x}) = \hat{\mathbf{e}}_i \exp(ik\hat{\boldsymbol{\omega}},\mathbf{x})$, where $\hat{\boldsymbol{\omega}}$ is a unit vector in the direction of propagation, $\hat{\mathbf{e}}_i$ is a unit vector in the direction of polarization, and $k = 2\pi\lambda^{-1}$, where λ is the wavelength. The wave is incident upon a particle or cell suspended within the translucent liquid, with relative dielectric constant $\epsilon_{\text{rel}}(\mathbf{x}) = \epsilon(\mathbf{x})/\epsilon_0$, where $\epsilon_{\text{rel}}(\mathbf{x}) \in \mathbb{C}$. The field at a distance R, from a reference point $\mathbf{x}$ within the cell, in direction $\hat{\boldsymbol{\omega}}$, is comprised of the incident field $\mathbf{E}_i$ and the scattered field $\mathbf{E}_s$. Within $R < L^2\lambda^{-1}$, the (near) field $\mathbf{E}_s$ has complicated amplitude and phase variations due to interference between contributions from different spatial regions of the cell. However, in the farfield $R > L^2\lambda^{-1}$, $\mathbf{E}_s$ behaves as a spherical wave and

$$\mathbf{E}_s(\mathbf{x}) = R^{-1} f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}') \exp(ikR), \tag{2.1}$$

where the function $f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}') \in \mathbb{C}$ represents the phase and amplitude of the scattered wave in the farfield in direction $\hat{\boldsymbol{\omega}}'$, when the cell is illuminated by a plane wave initially in along $\hat{\boldsymbol{\omega}}$. If $\mathcal{F}_s$ is the scattered power flux density at R from the particle along unit vector $\hat{\boldsymbol{\omega}}'$, due to an incident flux density $\mathcal{F}_i$, then the differential scattering cross section of the cell is

$$\chi_{\text{diff}}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}') = \lim_{R \to \infty} [\mathcal{F}_i^{-1}(R^2 \mathcal{F}_s)] = |f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2 = (\chi_t/4\pi)\mathcal{P}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}'), \tag{2.2}$$

where $\mathscr{P}(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}') \in [0,1] \subset \mathbb{R}$ is the phase function and χ_t is the total cross section (to be defined below). The total power scattered into all solid angles around the cell is related to the scattering cross section χ_s:

$$\chi_s = \int_{4\pi} \chi_{\text{diff}}\, d\omega = \int_{4\pi} |f(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}')^2\, d\omega = (\chi_t/4\pi) \int_{4\pi} \mathscr{P}(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}') d\omega. \tag{2.3}$$

The ratio $a(0)=(\chi_s/\chi_t)$ is the albedo of a single particle, so clearly $a(0) = \chi_t \int |f(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}')|^2\, d\omega = (4\pi)^{-1} \int \mathscr{P}(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}') d\omega$. If the scatterer is lossless then no power is absorbed and the total power loss is due to scattering with $\chi_t = \chi_s$. This would be the case for clear cells such as platelets. For blood cells, the haemoglobin molecules within the cell can absorb energy from the incident wave that is quantified by χ_a, the absorption cross section. Hence, $\chi_t = \chi_s + \chi_a$. The optical theorem[10] then defines the total χ_t to the imaginary part of the amplitude $f(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}})$ as $\chi_t = (4\pi/k)\text{Im}[f(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}})] \cdot \hat{\mathbf{e}}_i$. If the cellular geometry is spherical, the exact solutions for amplitude f, χ_a, and χ_s for electromagnetic waves incident on a dielectric sphere are derived from Mie theory.[10,27] For more complex biconcave and ellipsoidal geometries, relevant to biological particles, the method is still tractable.[27,28] For red cells, values of $\chi_s(\lambda)$ and $\chi_a(\lambda)$ have been derived for various optical wavelengths λ, using both experimental and theoretical methods.[18,29]

Given the microscopic optical properties of a single cell, the macroscopic bulk optical properties of a homogeneous suspension of a large number N of such red cells in a volume V can be ascertained. If $\eta = (N/V)$ is the number density, then the haematrocrit is $h = \eta \nu_e^{-1}$, where ν_e is the volume of a single erythrocyte and $h \in [0,1]$. The bulk macroscopic absorption coefficient $a(h) = \eta \chi_a$ of the suspension (in cm^{-1}) has a linear functional dependence on h:

$$a(h) = \eta \chi_a = h \nu_e^{-1} \chi_a. \tag{2.4}$$

However, since erythrocytes contain a binary mixture of Hb and HbO_2, the cross section χ_a also varies with the ratio of the molecular concentrations within the cell, since Hb and HBO_2 have different absorption spectra (except at isosbestic points at around $\lambda = 805$ nm). If $c[\text{Hb}]$ is the molecular concentration of Hb and $c[\text{HbO}_2]$ that of HbO_2 (in moles/cm^3), then the haemoglobin saturation $\mathscr{S}$ is defined as $\mathscr{S} = c[\text{HbO}_2]/[c[\text{HbO}_2] + c[\text{Hb}]]$ with $\mathscr{S} \in [0,1]$. As $\mathscr{S} \to 1$, the cells are fully oxygen saturated, and as $\mathscr{S} \to 0$, they are deoxygenated. In any suspension, all N cells will generally have equal $\mathscr{S}$. The coefficient $a(h,\mathscr{S})$ for any $0 < h < 1$ and $0 < \mathscr{S} < 1$ then has a linear functional dependence on h and $\mathscr{S}$:

$$a(h,\mathscr{S}) = h \nu_e^{-1} [\mathscr{S} \chi_{ao} + (1-\mathscr{S}) \chi_{ar}], \tag{2.5}$$

where χ_{ao} is the Mie absorption cross section for fully oxygenated cells ($\mathscr{S} = 1$) and χ_{ar} that for reduced cells with $\mathscr{S} = 0$ at a specific monochromatic wavelength λ. The function $a(h,\mathscr{S})$ is essentially the inverse probability per unit length (cm^{-1}) for light being absorbed along a rectilinear path. $a(h,\mathscr{S})$ can also be written as $a(h,\mathscr{S}) = [\mathscr{S} q_1 + (\mathscr{S}-1) q_2] h$, where the wavelength-dependent constants, q_1 and q_2, can be determined empirically.[9]

The macroscopic scattering coefficient $b = b(h)$ in cm^{-1} is functionally dependent only on h. For small h ($h <= 0.2$), we have $b(h) = \eta \chi_s = h \nu_e^{-1} \chi_s$, but as h increases the cells become more densely packed. As $h \to 1$ (the dense packing limit) the cells conglomerate and refractive index discontinuities $\epsilon_{\text{rel}}(x)$ between cell and plasma are gradually eliminated. Optically, the medium is then equivalent to a dense solution of haemoglobin with no scattering and the farfield approximation $R > L^2 \lambda^{-1}$ becomes invalid as $R \to 0$ as $h \to 1$. However, for all $h > 0$ and $h < 0.7 = 0.9$, the condition $R > L^2 \lambda^{-1}$ is satisfied. Since $b(h) \to 0$ as $h \to 0$ or 1, the scattering coefficient has a parabolic functional dependence on h:

$$b(h) = \chi_s \nu_e^{-1} h(1-h) = \eta \chi_s. \tag{2.6}$$

A correction has been presented as $b(h) = \chi_s \nu_e^{-1} h(1-h)(1.4-h)$,[3] but the parabolic term $h(1-h)$ is the essential feature.

Since $L \gg \lambda$, the coefficient $b(h)$ is not very dependent on λ. A useful formula on the full optical spectrum is $b(h) = (54.1 - 38.97h)h(1-h)$, which has been empirically determined.[9]

The dense packing limit as $h \to 1$ is only possible due to the elastic deformation properties of cells. For a suspension of hard spheres of volume ν, h cannot exceed 0.64[30] and $b(h) = h\chi_s \nu^{-1} \zeta(h)$, where $\zeta(h)$ is a monotone decreasing function of h with $\zeta(0) = 0$ and $\zeta(0.64) = 0$. These generic formulas are viable for all types of biosuspensions with a given number density and Mie cross sections.

III. BIOSCATTERING AMPLITUDES AND ANISOTROPIC PHASE FUNCTIONS

Since the red cell geometry is biconcave, this deviation from spherical symmetry suggests that the cells will not scatter flux isotropically. Deviations from spherically symmetric or isotropic scatter will be quantified by the complex bioscattering amplitude $f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')$ defined in Eq. (3.1), and the optical theorem for a single cell. Consider a specific intensity or radiance $\psi(\mathbf{x}, \hat{\boldsymbol{\omega}})$ incident upon a cylindrical volume of the suspension of cells of saturation $\mathscr{S}$, with unit cross section and length $d\xi$. The volume $d\xi$ contains $\eta\, d\xi$ cells, where η is the homogeneous number density of cells in the suspension. Each cell in the volume will then absorb the power $[\chi_{ao} \mathscr{S} + (\mathscr{S} - 1)\chi_{ar}]\psi(\mathbf{x}, \hat{\boldsymbol{\omega}})$ and scatter the power $\chi_s \psi(\mathbf{x}, \hat{\boldsymbol{\omega}})$, so the differential variational decrease $d\psi$ within the volume $d\xi$ is

$$d\psi(\mathbf{x}, \hat{\boldsymbol{\omega}}) = -d\xi[\chi_s + \chi_{ao} \mathscr{S} + (1 - \mathscr{S})\chi_{ar}]\psi(\mathbf{x}, \hat{\boldsymbol{\omega}}) = -\eta\, d\xi\, \chi_t . \tag{3.1}$$

Radiance $\psi(\mathbf{x}, \hat{\boldsymbol{\omega}})$ also increases due to contributions scattered into this direction from all possible directions $\hat{\boldsymbol{\omega}}'$. For a wave incident in direction $\hat{\boldsymbol{\omega}}'$ on a cell, the incident flux density through a small solid angle is $\mathscr{F}_i = \psi(\mathbf{x}, \hat{\boldsymbol{\omega}}')d\omega$. The power flux density $\mathscr{F}_s$ scattered by a cell in the direction $\hat{\boldsymbol{\omega}}$ at distance R from the cell is $\mathscr{F}_s = [|f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2 R^{-2}]\mathscr{F}_i$. The scattered intensity in direction $\hat{\boldsymbol{\omega}}$ due to $\mathscr{F}_i$ is therefore $\mathscr{F}_s R^2 = |f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2 \mathscr{F}_1 = |f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2 \psi(\mathbf{x}, \hat{\boldsymbol{\omega}}')d\omega'$. Summing over all contributions from all directions $\hat{\boldsymbol{\omega}}'$, the radiance scattered in direction $\hat{\boldsymbol{\omega}}$ by $\eta\, d\xi$ cells in volume $d\xi$ is

$$\int_{4\pi} \eta\, d\xi |f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2 \psi(\mathbf{x}, \hat{\boldsymbol{\omega}}')d\boldsymbol{\omega}'. \tag{3.2}$$

This equation can be expressed using the real cellular phase function $\mathscr{P}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')$ of (2.3):

$$\mathscr{P}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}') = (4\pi/\chi_t)|f(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')|^2. \tag{3.3}$$

If the vector orientation $\hat{\boldsymbol{\omega}}$ is given by (Θ, φ) with $\Theta \in [0, \pi]$, $\varphi \in [0, 2\pi]$ and that of $\hat{\boldsymbol{\omega}}'$ given by (Θ', φ'), then $\mathscr{P}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}') = \mathscr{P}(\Theta, \varphi | \Theta', \varphi')$, where $\cos\Theta = \hat{\boldsymbol{\omega}}.\hat{\boldsymbol{\omega}}'$, with respect to the incident beam axis along $+z$ or $\Theta_0 = \pi/2$. If scattering is symmetric about the incident wave, then the phase function is dependent only on Θ and can be expanded in Legendre polynomials $\mathscr{L}_n(\cos\Theta)$[10,31,32] for anisotropic scattering:

$$\mathscr{P}(\Theta, \varphi | \Theta', \varphi') = \mathscr{P}(\cos\Theta) = \sum_{n=0}^{\infty} A(n) \mathscr{L}_n(\cos\Theta), \tag{3.4}$$

where $A(0) = 1$ is the albedo of the cell. Given $\mathscr{P}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}')$, the effects of anisotropy or weighting of the scatter direction are embodied in the anisotropy parameter $g = \langle \cos\Theta \rangle = \langle \hat{\boldsymbol{\omega}}.\hat{\boldsymbol{\omega}}' \rangle$, defined as

$$g = \langle \cos\Theta \rangle = \left(\int_0^{\pi} \oint \mathscr{P}(\cos\Theta) \cos\Theta' \sin\Theta\ \delta\Theta'\ \delta\varphi' \right) \times \left(\int_0^{\pi} \oint \mathscr{P}(\cos\Theta) \sin\Theta' \delta\Theta' \delta\varphi' \right)^{-1}. \tag{3.5}$$

For isotropic scattering, $g = 0$ and $\mathscr{P}(\cos\Theta) = 1$; anisotropy increases as $g \to 1$. Possibilities for one-parameter scattering functions for blood cell suspensions[29] include the binomial law $\mathscr{P}(\cos\Theta) = \pi^{-1}[(i+1)/2^{i+1}][1 + \cos\Theta]$ and the ellipsoidal law $\mathscr{P}(\cos\Theta) = j[2\pi(1 - j\cos\Theta)]^{-1}[\mathrm{Ln}(1+j) - \mathrm{Ln}(1+j)]^{-1}$, where $i = \text{const}$ and $0 < j < 1$. Theoretically each of these

is viable. However, the experiment has demonstrated that the astrophysical Henyey–Greenstein function is the most convenient one-parameter function,[6,19,33,34] and is a very acceptable approximation to (3.4) for generic anisotropic scattering. Although initially applied to diffusion of radiation in the galaxy,[33] it also provides an adequate description of blood, particulate suspensions, and most other random medias. The function deviates from Mie theory calculations for high anisotropy ($g>0.9$ and $g<1$) with very small scatter angles.[34] However, this can be corrected.[34] For the purposes of this paper, it can be utilized as the paradigmatic example of a viable one-parameter phase function $\mathscr{P}(\Theta)$ for the bulk scattering properties of generic, anisotropic, optical systems comprised of discrete particulate constituents. It can be expressed as

$$\mathscr{P}(\cos\,\Theta)=\frac{1}{2\pi}\,\beta_1[\beta_2-\cos\,\Theta]^{-3/2}. \tag{3.6}$$

Here $\beta_1=(1+g^2)/2g$, $\beta_2=(1-g^2)/2g$. A valuable property of this function is that it is easily integrated. The scattering Mie cross sections of the cells can be related to $\mathscr{P}(\cos\,\Theta)$.[29] If χ_s is the total Mie scattering cross section, then the purely forward and backward scattering Mie cross sections $\chi_s(+)$ and $\chi_s(-)$ can be computed with good accuracy from:[29]

$$\chi_s(+)\sim\chi_s\int_0^1\beta_1(\beta_2-\cos\,\Theta)^{-3/2}d(\cos\,\Theta), \tag{3.7a}$$

$$\chi_s(-)\sim\chi_s\int_{-1}^0\beta_1(\beta_2-\cos\,\Theta)^{-3/2}d(\cos\,\Theta). \tag{3.7b}$$

The ratio $|\chi_s(+)/\chi_s(-)|$ is determined by the geometric shape of the cell, so that $|\chi_s(+)/\chi_s(-)|>1$ for $g>0$ and $|\chi(+)/\chi(-)|=1$ for $g=0$.

IV. STATEMENT OF THE PROBLEM: TRANSPORT THEORY AND PATH INTEGRALS

Adding the equations (3.1) (3.2) and using (3.3), we obtain the integro-differential radiative transfer equation for beam propagation within the blood cell suspension:

$$\frac{d\psi(\mathbf{x},\hat{\boldsymbol{\omega}})}{d\xi}=-\eta[\chi_s+\mathscr{S}\chi_{ao}+(1-\mathscr{S})\chi_{ar}]\psi(\mathbf{x},\hat{\boldsymbol{\omega}})$$
$$+\left(\frac{\eta[\chi_s+\mathscr{S}\chi_{ao}+(1-\mathscr{S})\chi_{ar}]}{4\pi}\right)\int_{4\pi}\mathscr{P}(\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\omega}}')\psi(\mathbf{x},\hat{\boldsymbol{\omega}}')d\boldsymbol{\omega}'. \tag{4.1}$$

The left side can be written as a gradient or divergence operator since $d\psi(\mathbf{x},\hat{\boldsymbol{\omega}})/d\xi=\hat{\boldsymbol{\omega}}\cdot\boldsymbol{\nabla}\psi(\mathbf{x},\hat{\boldsymbol{\omega}})=\hat{\boldsymbol{\omega}}\cdot[\psi(\mathbf{x},\hat{\boldsymbol{\omega}})\hat{\boldsymbol{\omega}}]$ and $\hat{\boldsymbol{\omega}}$ is solenoidal in that $\boldsymbol{\nabla}\cdot\hat{\boldsymbol{\omega}}=0$. Writing $\hat{\boldsymbol{\omega}}=(\Theta,\varphi)$ with $\Theta\in[0,\pi]$ and $\varphi[0,2\pi]$ and using the macroscopic definitions (2.5) and (2.6):

$$[\hat{\boldsymbol{\omega}},\boldsymbol{\nabla}+(a(h,\mathscr{S})+b(h))]\psi(\mathbf{x},\Theta,\varphi)$$
$$=\left(\frac{a(h,\mathscr{S})+b(h)}{4\pi}\right)\int_0^{\pi}\oint\mathscr{P}(\Theta,\varphi|\Theta',\varphi')\psi(\mathbf{x},\Theta',\varphi')\sin\,\Theta'\;\delta\Theta'\;\delta\varphi'. \tag{4.2}$$

We consider flat wide beam coherent irradiance of a finite slab of thickness d, bounded by planes $z=0$ and $z=d$ (this geometry is adequate for realistic samples within thin cuvettes and laser biology problems). However, the transport equation has no general solution, even in this geometry. Since the beam is coherent, the radiance is the linear superposition of coherent and scattered components $\psi=\psi_c+\psi_s$. The coherent intensity ψ_c decays according to $d\psi_c(z,\hat{\boldsymbol{\omega}})/dz=-[a(h,\mathscr{S})+b(h)]\psi_c(z,\hat{\boldsymbol{\omega}})$, where $\hat{\boldsymbol{\omega}}$ is along the beam axis and perpendicular to $z=0$. The solution is simply the Beer–Lambert law:

$$\psi(z,\hat{\boldsymbol{\omega}})=\psi_0\,\exp[-(a(h,\mathscr{S})+b(h))z]. \tag{4.3}$$

The scattered component ψ_s satisfies

$$\hat{\boldsymbol{\omega}}\cdot\nabla\psi_s(\mathbf{x},\Theta,\varphi)=-[a(h,\mathscr{S})+b(h)]\psi_s(\mathbf{x},\Theta,\varphi)+\left(\frac{a(h,\mathscr{S})+b(h)}{4\pi}\right)\int_0^{\pi}\oint\mathscr{P}(\Theta,\varphi|\Theta',\varphi')\psi_s(\mathbf{x},\Theta',\varphi')+\psi_c(\mathbf{x},\Theta',\varphi'))\sin\Theta'\,\delta\Theta'\,\delta\varphi'. \tag{4.4}$$

Exponential decay of ψ_c acts as an additional source term for the generation of ψ_s and ψ_s must be zero at the surface planes $z=0$ or $z=d$ when $\hat{\boldsymbol{\omega}}$ is pointed into the suspension. Let $\mathscr{E}$ be a plane or surface within the medium (in the mathematical sense) and let $d\mathscr{E}$ be an element of the surface with unit normal $\hat{\boldsymbol{\omega}}_0$. The net forward scattered flux intersecting $\mathscr{E}$ is given by a solid angle integral over ψ_s in the forward range $(0<=\Theta<=\pi/2)$: when the blood is described by $(h,\mathscr{S})$, it can be written as $\mathscr{F}(+,\mathscr{E}:h,\mathscr{S})d\mathscr{E}$. The net backscattered radiant flux in $(\pi/2<\Theta<=0)$ is similarly $\mathscr{F}(-,\mathscr{E}:h,\mathscr{S})d\mathscr{E}$, where the forward and back flux densities are

$$\mathscr{F}(+,\mathscr{E}:h,\mathscr{S})=\int_0^{\pi/2}\oint\psi_s(\mathscr{E},\Theta,\varphi)\cos\Theta\,\sin\Theta\,\delta\Theta\,\delta\varphi, \tag{4.5a}$$

$$\mathscr{F}(-,\mathscr{E}:h,\mathscr{S})=\int_{\pi/2}^{0}\oint\psi_s(\mathscr{E},\Theta,\varphi)\cos\Theta\sin\Theta\,\delta\Theta\,\delta\varphi, \tag{4.5b}$$

where $\cos\Theta=\hat{\boldsymbol{\omega}}\cdot\hat{\boldsymbol{\omega}}_0$. The coherent radiant flux density follows directly from (4.3) since $\hat{\boldsymbol{\omega}}$ is perpendicular to $z=\mathscr{E}$, so that $\mathscr{F}_c(+,\mathscr{E};h,\mathscr{S})=\mathscr{F}_0\exp[-(a(h,\mathscr{S})+b(h))\mathscr{E}]$, where $\mathscr{F}_0$ is the incident flux density at plane $z=0$.

The rotational symmetry of the RTE is broken when a solution is constructed on the basis of initial data, as is the case, for example, with slab geometry. The initial data are the radiance $\psi(z=0;\hat{\boldsymbol{\omega}}_0)=\psi\ (z=0;\Theta_0,\varphi_0)$ at boundary $z=0$, where the unit vector $\hat{\boldsymbol{\omega}}_0$ is normal to plane $z=0$ for incident coherent or Fermat light $(\Theta_0=\pi/2)$ and points into the medium.

In Refs. 35 and 36, a solution was constructed for the initial value problem of deriving the forward scattered radiance ψ_s at any plane $z=$const below the initial data plane at $z=0$. The radiance solution is expressed in terms of a path integral kernal over a measure spanning all paths that scattered light can take from the initial angular direction $\hat{\boldsymbol{\omega}}_0$ at data plane $z=0$, to z at direction $\hat{\boldsymbol{\omega}}$. However, the path kernal cannot take into account backscattering and is restricted to very small angle forward scatterings and low absorption. The path kernal cannot be evaluated analytically and requires a Gaussian approximation. Nevertheless, this interpretation of mesoscopic RTT is of considerable efficacy.

Given initial data at $\psi(\mathbf{x}',\Theta',\varphi')$ at $\mathbf{x}=\mathbf{x}'$, the radiance $\psi(\mathscr{E};\Theta,\varphi)$ at an interior data plane $z=\mathscr{E}$ in direction (Θ,φ) is given by Huygen's principle in terms of the point spread Green's function or kernal $G(\mathscr{E},\Theta,\varphi|0,\Theta_0,\varphi_0)$:

$$\psi(\mathscr{E},\Theta,\varphi)=\int\delta\mathbf{x}'\int\delta\Theta'\int\delta\varphi'G(\mathscr{E},\Theta,\varphi|\mathbf{x}',\Theta',\varphi')\psi(\mathbf{x}',\Theta',\varphi'). \tag{4.6}$$

The radiance $\psi(\mathscr{E},\Theta,\varphi)$ at surface $\mathscr{E}$ is due to all optical paths generated by multiple scattering, intersecting $\mathscr{E}$ in direction (Θ,φ) having begun from $z=0$:

$$\psi(\mathscr{E},\Theta,\varphi)=\int\int\int\cdots\prod_{n>0}\delta\mathbf{x}_n\,\delta\Theta_n\,\delta\varphi_nG(\mathbf{x}_{n+1},\Theta_{n+1},\varphi_{n+1}|\mathbf{x}_n,\Theta_n,\varphi_n)\psi(\mathbf{x}=0,\Theta_0\varphi_0). \tag{4.7}$$

This is essentially a path integral over all possible optical paths that can contribute to $\psi(\mathscr{E},\Theta,\varphi)$, and G satisfies Kolmogorov equations:

$$G(\mathbf{x}_{n+1},\Theta_{n+1},\varphi_{n+1}|\mathbf{x}_{n-1},\Theta_{n-1},\varphi_{n-1})$$

$$=\int\int\int G(\mathbf{x}_{n-1},\Theta_{n+1},\varphi_{n+1}|\mathbf{x}_n,\Theta_n,\varphi_n)$$

$$\times G(\mathbf{x}_n,\Theta_n,\varphi_n|\mathbf{x}_{n-1},\Theta_{n-1},\varphi_{n-1})\,\delta x_n\,\sin\,\Theta_n\,\delta\Theta_n\,\delta\varphi_n. \tag{4.8}$$

The forward scattered radiant flux $\mathscr{F}(+,\mathscr{E}{:}h,\mathscr{S})$ through $\mathscr{E}$ is then explicitly expressed as

$$\mathscr{F}(+,\mathscr{E}{:}h,\mathscr{S})=\int_0^{\pi/2}\oint\Bigg[\cdots\int\delta\mathbf{x}_{n+1}\int\delta\Theta_{n+1}\int\delta\varphi_{n+1}\int\delta\mathbf{x}_n\int\delta\Theta_n\int\delta\varphi_n\cdots$$

$$\times\int\delta\mathbf{x}_2\int\delta\Theta_2\int\delta\varphi_2\int\delta\mathbf{x}_1\int\delta\Theta_1\int\delta\varphi_1\cdots G(\mathbf{x}_{n+1},\Theta_{n+1},\varphi_{n+1}|\mathbf{x}_n,\Theta_n,\varphi_n)$$

$$\times G(\mathbf{x}_n\Theta_n\varphi_n|\mathbf{x}_{n-1},\theta_{n-1},\varphi_{n-1})\cdots G(\mathbf{x}_2,\Theta_2,\varphi_2|\mathbf{x}_1,\Theta_1,\varphi_1)$$

$$\times G(\mathbf{x}_1,\Theta_1,\varphi_1|\mathbf{x}_2,\theta_1,\varphi_1)$$

$$\times G(\mathbf{x}_1,\Theta_1,\varphi_1|0,\Theta_0,\varphi_0)\psi(0,\Theta_0,\varphi_0)\Bigg]\cos\,\Theta\,\sin\Theta\,\delta\Theta\,\delta\varphi, \tag{4.9}$$

and similarly for the back flux $\mathscr{F}(-,\mathscr{E}{:}h,\mathscr{S})$. $\mathscr{F}(\pm,\mathscr{E};h,\mathscr{S})$ as a path integral, can be interpreted as a consequence of all possible vectorial intersections of all optical trajectories through surface $\mathscr{E}$, in the forward half-space. These are essentially all possible optical trajectories $\{\gamma\}$ that can evolve within the suspension, in biological state $(h,\mathscr{S})$ due to scattering, from Fermat paths or rays entering normal ($\Theta_0=\pi/2$) to the initial data plane at $z=0$. The rays propagate in the translucent plasma in rectilinear steps $\|\delta\mathbf{x}_n\|$ from cell to cell ($\mathbf{x}_{n-1}\rightarrow\mathbf{x}_n\rightarrow\mathbf{x}_{n+1}$) with $\delta\mathbf{x}_n=\mathbf{x}_{n+1}-\mathbf{x}_n$, where they experience either scattering and deflection or absorption. Overall, each optical trajectory within the biosuspension is constructed from such sequences of rectilinear steps or vectors $\delta\mathbf{x}_n=\mathbf{x}_{n+1}-\mathbf{x}_n$, each pointing in a random direction due to refraction from cellular index discontinuities [in (4.9) the integration measure $|\Pi_n\,\delta\mathbf{x}_n$ is over all possible $\delta\mathbf{x}_n$]. A statistical bundle of such trajectories $\{\gamma\}$ vectorially intersecting any interior surface $\mathscr{E}$, effectively gives rise to a stable scattered flux density $\mathscr{F}(\pm,\mathscr{E})$ through the surface.

The probability of any path γ vectorially intersecting $\mathscr{E}$ in direction Θ, φ, from scattering at $\mathbf{x}_n$ is not dependent on the previous random history of the path that resulted in its eventual arrival at $\mathbf{x}_n$. Any optical migration or trajectory γ is then essentially a Markov chain, since it has no memory of its past for a given present,[36] as exemplified by the Kolmogorov equation (4.8). At the level of description, (ii), as described in Sec. I, radiative transfer can be interpreted statistically as a bundle of such optical Markov trajectories being randomly traced out by a ''gas'' or ensemble of Brownian-like point photons being scattered rectilinearly from cell to cell and/or being absorbed in the cells. A single optical Markov migration or chain γ, from the ensemble defined by the path sum in (4.9), can be expressed as (dropping integrals for a single optical path).

$$G(\mathscr{E},\Theta,\varphi|\Theta,\theta_0,\varphi_0)=G(\mathscr{E},\Theta,\varphi|\mathbf{x}_n,\Theta_n,\varphi_n)\times G(\mathbf{x}_n,\Theta_n,\varphi_n|\mathbf{x}_{n-1},\Theta_{n-1}\varphi_{n-1})\times\cdots$$

$$\times G(\mathbf{x}_1,\Theta_1,\varphi_1|0,\Theta_0\varphi_0), \tag{4.10}$$

where $G(\mathbf{x}_{n+1},\Theta_{n+1},\varphi_{n+1}|\mathbf{x}_n,\Theta_n,\varphi_n)$ is the transition probability or the Green's function for being scattered to the nth order from $\mathbf{x}_n$ with incident direction (Θ_n,φ_n) to $\mathbf{x}_{n+1}$ in direction $(\Theta_{n+1},\varphi_{n+1})$. Since G cannot be ascertained (outside the Gaussian or diffusion approximation), any single optical migration or random path γ can be represented as a dataset of stochastic variables:

$$\left[\left(0,\frac{\pi}{2},0\right),(R_1,\Theta_1,\varphi_1),\langle R_2,\Theta_2,\varphi_2\rangle,\ldots,(R_{n-1},\Theta_{n-1},\varphi_{n-1}),(R_n,\Theta_n,\varphi_n)\ldots\right], \tag{4.11}$$

where $R_n \equiv \|\delta \mathbf{x}_n\| = \|\mathbf{x}_{n+1} - \mathbf{x}_n\|$, and $(z, \Theta_0, \varphi_0) = (0, \pi/2, 0)$ is the initial dataset for all Fermat rays entering the suspension at the initial data plane $z = 0$. The total length of the path is $\Sigma_n \|\delta \mathbf{x}_n\| = \Sigma_n R_n^{(s)} + R^{(a)}$.

V. THE METHOD

The path integral flux expression in (4.9) is statistically approximated by an ensemble or bundle of u Brownian-like or Markov optical trajectories:

$$[\gamma(1), \gamma(2), \ldots \gamma(p) \ldots \gamma(u)].$$

The algorithm generates u random stochastic datasets of the form (4.11), from a bundle of u Fermat rays entering at $z = 0$, all with initial specification $(z, \Theta_0, \varphi_0) = (0, \pi/2, 0)$ on the data surface $z = 0$. Given any $h \in (1,0)$, $\mathscr{S} \in [1,0]$ and $g \in [0,1)$, for the suspension, the probability densities that constrain the allowed stochastic variable sets $\{(R_n, \Theta_n, \varphi_n)\}$, are easily determined, and are functionally dependent on variations in h, $\mathscr{S}$, or g. By sampling these densities, the stochastic datasets $\{(R, \Theta, \varphi)\}$ describing a random optical migration γ, can be generated. The probability of a rectilinear propagation distance $R^{(s)}$ in the translucent plasma between cells, without being scattered by a cell is $\mathscr{P}(R^{(s)})$. Similarly, the probability of rectilinear propagation $R^{(a)}$ in plasma without being absorbed by a cell is $\mathscr{P}(R^{(a)})$. We must have $\mathscr{P}: R \to [0,1]$:

$$\mathscr{P}(R^{(s)}) = b(h) \exp[-b(h) R^{(s)}], \tag{5.1a}$$

$$\mathscr{P}(R^{(a)}) = a(h, \mathscr{S}) \exp[-a(h, \mathscr{S}) R^{(a)}]. \tag{5.1b}$$

Clearly, $\mathscr{P}(R^{(s)}) \to 0$ and $\mathscr{P}(R^{(a)}) \to 0$ as $R^{(s)}$ and $R^{(a)}$ increase. $R^{(s)} \to 0$ as $b(h) \to 0$ as $h \to 1$. However, the ''farfield'' limit $R^{(s)} > L^2 \lambda^{-1}$ generally holds for $h < 0.7 - 0.9$ (above these values scattering can be ignored anyway). These densities are necessarily normalized as

$$\int_0^\infty \mathscr{P}(R^{(s)}) dR^{(s)} = 1, \tag{5.2a}$$

$$\int_0^\infty \mathscr{P}(R^{(a)}) dR^{(a)} = 1, \tag{5.2b}$$

so that for all finite interaction steps $R^{(s)}$ and $R^{(a)}$:

$$\int_0^{R^{(s)}} \mathscr{P}(R^{(s)}) dR^{(s)} = \zeta^{(1)} < 1, \tag{5.3a}$$

$$\int_0^{R^{(a)}} \mathscr{P}(R^{(a)}) dR^{(a)} = \zeta^{(2)} < 1. \tag{5.3b}$$

Integrating and solving for $R^{(s)}$ and $R^{(a)}$ gives the generator formulas:

$$R^{(s)} = -(b(h))^{-1} \text{ Ln } \zeta^{(1)}, \tag{5.4a}$$

$$R^{(a)} = -(a(h, \mathscr{S}))^{-1} \text{ Ln } \zeta^{(2)}. \tag{5.4b}$$

If $\zeta^{(1)}$ and $\zeta^{(2)}$ are pure, uncorrelated random numbers on (1,0), the set $(\zeta^{(1)}, \zeta^{(2)})$ generates the stochastic pair $(R^{(s)}, R^{(a)})$. The weighted anisotropic density $\mathscr{P}(\cos \Theta)$ or $\mathscr{P}(\Theta)$ for stochastic scatter angles is given by Eq. (3.6). Using this Greenstein form, we also have the necessary renormalization:

$$\int_0^\pi \oint \beta_1 (\beta_2 - \cos \Theta)^{-3/2} \sin \Theta \, d\Theta \, d\varphi = 1, \tag{5.5a}$$

$$\int_0^{\theta} \oint \beta_1(\beta_2-\cos\Theta)^{-3/2}\sin\Theta\, d\Theta\, d\varphi=\zeta^3<1. \tag{5.5b}$$

The integration can be done and inverted to give

$$\Theta=\frac{1}{2\pi}\left[\frac{g}{2}\left(\beta_1-\beta_2\left(1-g-2g\zeta^3\right)\right]\right]. \tag{5.6}$$

If $\zeta^{(3)}$ is a random number on (1,0), then the generator (5.6) gives a stochastic variable Θ on $[0,\pi]$. The stochastic polar angles φ are trivially sampled by $\varphi=2\pi\zeta^{(4)}$ with $\zeta^{(4)}\in(0,1)$.

Consider the nth-order scattering of any optical Brownian path γ within a cell at $\mathbf{x}_n$, which is then randomly scattered to another cell randomly located $\mathbf{x}_{n+1}$. A dataset of stochastic variables is generated for the nth scatter event:

$$(\zeta_n^{(1)},\zeta_n^{(2)},\zeta_n^{(3)},\zeta_n^{(4)})\rightarrow(R_n^{(s)},R_n^{(a)},\Theta_n,\varphi_n).$$

Since the shortest rectilinear step from the pair $(R_n{}^{(s)},R_n{}^{(a)}$ is the most probable, the following inequalities and operations are considered.

(1) If $R_n{}^{(s)}<R_n{}^{(a)}$, then $R_n{}^{(s)}$ is selected.

The optical migration γ then stochastically evolves by propagating a rectilinear distance $R_n{}^{(s)}$ in the translucent plasma, before encountering another cell at random location $\mathbf{x}_{n+1}$ such that $R_n{}^{(s)}=\|\mathbf{x}_{n+1}-\mathbf{x}_n\|$. Since $R_n{}^{(s)}$ is a stochastic scatter step, the optical path undergoes another Brownian-like scatter from a cell at $\mathbf{x}_{n+1}$.

(2) If $R_n{}^{(a)}<R_n{}^{(s)}$, then $R_n{}^{(a)}$ is selected.

The optical migration γ then evolves by propagating a rectilinear distance $R_n{}^{(a)}$ in the translucent plasma, before encountering another cell at random location $\mathbf{x}_{n+1}$ such that $R_n{}^{(a)}=\|\mathbf{x}_{n+1}-\mathbf{x}_n\|$. Since $R_n{}^{(a)}$ is a stochastic absorption step or terminal step, the optical trajectory terminates at $\mathbf{x}_{n+1}$. A new trajectory is then generated. If the suspension of particles is lossless, then $A=0$, and only scatter steps are relevant (for example, translucent cells); the trajectory can only terminate by leaving the space containing the particles.

The orientation of the resultant vector $\delta\mathbf{x}_n=\mathbf{x}_{n+1}-\mathbf{x}_n$ is determined by the stochastic angles sampled from the integrated phase function in (3.6). If the optical trajectory approached the cell at $\mathbf{x}_n$ in direction $(\Theta_{n-1},\varphi_{n-1})$, then it is scattered in a random direction (Θ_n,φ_n) toward $\mathbf{x}_{n+1}$. Since the phase function is weighted by the factor $g>0$, optical trajectories are preferentially scattered into the forward half-space $(0<\Theta<\pi/2)$, if the incidence direction was in the forward half-space. If $R_n{}^{(s)}<R_n{}^{(a)}$ then a new dataset of stochastic variables is generated at $\mathbf{x}_{n+1}$:

$$(\zeta_{n+1}^{(1)},\zeta_{n+1}^{(2)},\zeta_{n+1}^{(3)},\zeta_{n+1}^{(4)})\rightarrow(R_{n+1}^{(s)},R_{n+1}^{(a)},\Theta_{n+1},\varphi_{n+1}).$$

This iteration continues until an absorption step is sampled or the trajectory leaves the space containing the suspension.

Any generic random optical trajectory γ contributing to the flux path integral in (4.9) can then be expressed as a dataset of sampled stochastic variables:

$$\gamma:[(0,1/2\pi,0),(R_1^{(s)},\Theta_1,\varphi_1),(R_2^{(s)},\Theta_2,\varphi_2),\ldots,(R_n^{(s)},\Theta_n,\varphi_n)\ldots].$$

Sequential vectors $\mathbf{x}_{n+1}$ and $\mathbf{x}_n$ are related by a linear matrix transformation $\Gamma[R_n{}^{(s)},\Theta_n,\varphi_n]$:

$$\mathbf{x}_{n+1}=\Gamma[R_n{}^{(s)},\Theta_n,\varphi_n]\mathbf{x}_n,\quad \text{if } R_n{}^{(s)}<R_n{}^{(a)}, \tag{5.7a}$$

$$\mathbf{x}_{n+1}=\Gamma[R_n{}^{(a)},\Theta_n,\varphi_n]\mathbf{x}_n,\quad \text{if } R_n{}^{(a)}<R_n{}^{(s)}. \tag{5.7b}$$

Using these rules, the bundle of u optical trajectories $\{\gamma(1)\cdots\gamma(u)\}$ representing the flux path integral (4.9) can be generated. The comparison of inequalities $R^{(a)}<R^{(s)}$ and $R^{(s)}<R^{(a)}$ functions as a roulette. The relative statistical probabilities of either outcome is determined by the ratio $a(h,\mathscr{S})b^{-1}(h)$. The $R^{(s)}$ become shorter on average as h increases from zero. If $\mathscr{S}$ is varied on [1,0] then only $a(h,\mathscr{S})$ and $\mathscr{P}(R^{(a)})$ are varied relative to fixed $b(h)$ and $\mathscr{P}(R^{(s)})$. Varying h

on (0,1) and $\mathscr{S}$ on [0,1] in any combination shifts the ratio $\mathscr{P}(R^{(a)})\mathscr{P}^{-1}(R^{(s)})$ and alters the relative probabilities of $R^{(a)}$ being preferentially selected at any scatter event, relative to $R^{(s)}$. The densities $\mathscr{P}(R^{(s)})$ and $\mathscr{P}(R^{(a)})$ illustrate that the shortest of the rectilinear optical interaction steps is always most probable, which is why it is selected from the sampled pair $(R^{(s)},R^{(a)})$.

In neutron transport simulation, mathematical and statistical optimization techniques were developed by Metropolis and Ulam.[23] The neutron path segments are sampled using the total inverse attenuation $(a+b)^{-1}$. All initial particles have a statistical weight w_0 of unity that is reduced by factor $b(a+b)^{-1}$ at each scatter event so that $w_n=w_0b^n(a+b)^{-n}\rightarrow 0$ as $n\rightarrow\infty$. Random migrations are terminated once $w_n<=w_{\text{th}}$, where w_{th} is a threshold weight. The problem is in accurately determining values for the threshold; an incorrect cutoff will skew the statistical acceptability. A correct application does improve statistics for very thick medias and for cases where particles undergo deep penetration with many hundreds or thousands of scattering events, such as in nuclear reactor and radiation shielding problems.[31,32] It is also tailored toward computing scalars like the average diffuse intensity. However, these techniques are superfluous and a hindrance for the optical vector flux calculations in biosuspensions of finite depth, which are also absorbing as well as scattering. For most biological applications, the depth is typically $d=0.5$ mm to 5 cm, so that trajectories in such medias can be generated and terminated more rapidly and efficiently using the alternative method, while maintaining adequate statistics. In this alternative formalism, vectorial fluxes $\mathscr{F}(\pm,\mathscr{E}:h,\mathscr{S})$ are ascertained more rapidly and effectively than a scalar fluence $\phi(\mathscr{E})$. In addition, the absorption effects of cells can be separated out from their scattering effects.

VI. VECTORIAL INTERSECTIONS OF DATA SURFACES

Given the radiance $\psi(0,\pi/2,0)$ on the initial data plane $z=0$, the scattered flux density $\mathscr{F}(+,\mathscr{E}:h,\mathscr{S})$ through data surface $z=\mathscr{E}$ is given by the flux path integral in (4.9). Since the bundle of optical paths $\{\gamma(p)\}$ vectorially intersects the data plane $\mathscr{E}$, one can easily ascertain the radiant flux densities $\mathscr{F}(\pm,\{\mathscr{E}\}:h,\mathscr{S})$ throughout the medium, by defining a complete set of data planes $\{\mathscr{E}\}$. If any optical trajectory scatters from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$, then the vector $\delta\mathbf{x}_n=\mathbf{x}_{n+1}-\mathbf{x}_n$ will vectorially intersect or pierce a subset of these data planes, consecutively and in either the net forward half-space ($0<\Theta_n<\pi/2$ and $0<\varphi_n<2\pi$) or the back half-space ($\pi/2<\Theta_n<0$ and $0<\varphi_n<2\pi$). In generating the stochastic set of optical Markov trajectories $[\gamma(1),\gamma(2),\ldots\gamma(p)\ldots\gamma(u)]$, the algorithm can record the vectorial intersections $\{\delta\mathbf{x}_n^{(p)}\}\cap(\mathscr{E})$, due to all successive scatterings of all $p=1$ to u trajectories with all specified interior data planes $\{\mathscr{E}\}$.

Given the spatial region $0<z<d$ containing the suspension, define a set of data planes as $\{\mathscr{E}\}=\{z(q)\}=\{z(1),z(2),\ldots z(q)\ldots z(v)\}$, for $q=1$ to integer v. These are simply generated by $z(q)=(q\times d\times v^{-1})$ for all $q=1$ to v. They are therefore parallel and equidistantly spaced along the interval $[0,d]$, and normal to the incident Fermat rays at $z=0$ that have $(z,\Theta_0,\varphi_0)=(0,\pi/2,0)$. The density of surfaces is determined by v and $d=z(v)$. Scattered flux densities $\mathscr{F}(\pm,q:h,\mathscr{S})$ exist through all these data surfaces for any given $1>h>0$ and $0<=\mathscr{S}<=1$. To each data plane $z(q)$ in the set $\{z(q)\}$, there is then assigned a set of real matrix elements $\{M(+,q:h,\mathscr{S})\}$, $\{M(-,q:h,\mathscr{S})\}$ and $\{C(+,q:h,\mathscr{S})\}$, for all $q=1$ to v, so that

$$\{M(+,q:h,\mathscr{S})\}=\{M(+,1:h,\mathscr{S}),M(+,2:h,\mathscr{S})\cdots M(+,q:h,\mathscr{S})\cdots M(+,v:h,\mathscr{S})\},$$

$$\{M(-,q:h,\mathscr{S})\}=\{M(-,1:h,\mathscr{S}),M(-,2:h,\mathscr{S})\cdots M(-,q:h,\mathscr{S})\cdots M(-,v:h,\mathscr{S})\},$$

$$\{C(+,q:h,\mathscr{S})\}=\{C(+,1:h,\mathscr{S}),C(-,2:h,\mathscr{S})\cdots C(+,q;h,\mathscr{S})\cdots C(+,v:h,\mathscr{S})\}.$$

The matrix elements $\{M(+,q:h,\mathscr{S})\}$ score all possible vectorial intersections of all stochastic optical trajectories $\{\gamma(p)\}$ for $p=1$ to u, through all $\{z(q)\}$ for $q=1$ to v, in the forward half-space for some combination $h>0$ and $0<=\mathscr{S}<=1$. Matrix elements $\{M(-,q:h,\mathscr{S})\}$ score all vectorial intersections of all stochastic optical trajectories $\{\gamma(p)\}$ for $p=1$ to u, through all $\{z(q)\}$ for $q=1$ to v, in the back half-space. Matrix elements $C(+,q:h,\mathscr{S})$ score all vectorial intersections normal to any $\{z(q)\}$ due to incident Fermat rays, which have $\Theta_0=\pi/2$. Initially, $\{M(+,q:h,\mathscr{S})\}=\{0,0,\ldots\}$, $M(-q:h,\mathscr{S})=\{0,0,\ldots\}$ and $C(+,q:h,\mathscr{S})=\{0,0,\ldots\}$.

To examine how the matrix elements are incremented within the sequence of operations for generating $\{\gamma(1)\cdots\gamma(p)\cdots\gamma(u)\}$, consider the nth-order scattering of the pth optical or Markovian trajectory $\gamma(p)$ in a red cell biosuspension whose biological state is described by h and $\mathscr{S}$. Then $\mathbf{x}_n^{(p)}\rightarrow\mathbf{x}_{n+1}^{(p)}$. The stochastic set $\{R_n^{(s)},R_n^{(a)},\theta_n,\varphi_n\}$ is generated. Suppose $R_n^{(s)}<R_n^{(a)}$, then from (5.7),

$$\mathbf{x}_{n+1}^{(p)}=\Gamma[R_n^{(s)},\theta_n,\varphi_n]\mathbf{x}_n^{(p)}.$$

Let $\mathbf{x}_{n+1}^{(p)}=(x_{n+1}^{(p)},y_{n+1}^{(p)},z_{n+1}^{(p)})$ and $\mathbf{x}_n=(x_n^{(p)},y_n^{(p)},z_n^{(p)})$ and consider the qth data surface $z(q)$. The following inequalities and operations are implemented:

$$\text{if } z_{n+1}^{(p)}>z(q)>z_n^{(p)} \text{ then } M(+,q{:}h,\mathscr{S})=M(+,q{:}h,\mathscr{S})+1;$$

$$\text{if } z_{n+1}^{(p)}<z(q)<z_n^{(p)} \text{ then } M(-,q{:}h,\mathscr{S})=M(-,q{:}h,\mathscr{S})+1;$$

$$\text{if } z_{n+1}^{(p)}>z_n^{(p)}>z(q) \text{ then } M(\pm,q{:}h,\mathscr{S})=M(\pm,q{:}h,\mathscr{S});$$

$$\text{if } z(q)>z_{n+1}^{(p)}>z_n^{(p)} \text{ then } M(\pm,q;h,\mathscr{S})=M(\pm,q{:}h,\mathscr{S});$$

$$\text{if } z_{n+1}^{(p)}<z_n^{(p)}<z(q) \text{ then } M(\pm,q{:}h,\mathscr{S})=M(\pm,q{:}h,\mathscr{S});$$

$$\text{if } z(q)<z_{n+1}^{(p)}<z_n^{(p)} \text{ then } M(\pm,q{:}h,\mathscr{S})=M(\pm,q{:}h,\mathscr{S});$$

$$\text{if } z_1^{(p)}>z(q)>0 \text{ then } C(+,q{:}h,\mathscr{S})=C(+,q{:}h,\mathscr{S})+1, \quad \text{for } n=1 \text{ only.}$$

This sequence is repeated for all data surfaces $\{z(q)\}$ for $q=1$ to v, for all scatterings of all $p=1$ to u optical Markov trajectories (although sequential, the process is inherently parallel). For sufficiently large u, the complete matrix elements $\{M(+,q{:}h,\mathscr{S})\}$ and $\{M(-,q{:}h,\mathscr{S})\}$ contain the equivalent information of the (unknown) flux solutions $\mathscr{F}(+,q{:}h,\mathscr{S})$ and $\mathscr{F}(-,q{:}h,\mathscr{S})$ of the radiative transfer equation. In addition, one has the option of filtering out scattering beyond a given order m, by only scoring vector intersections for $n<=m$. In this way, specific flux contributions of lower-order multiply scattered light can be separated out from the total flux.

If $\mathscr{F}_0$ is the incident flux density at initial data plane $z=0$, the solutions for scattered anisotropic fluxes $\mathscr{F}(\pm,q{:}h,\mathscr{S})$ through any data surface $z(q)$, for any suspension in state $(h,\mathscr{S},g)$, are

$$\begin{aligned}\mathscr{F}(+,q{:}h,\mathscr{S})&=\int_0^{\pi/2}\oint\psi_s(q,\Theta,\varphi)\cos\Theta\,\sin\Theta\,\delta\Theta\,\delta\varphi\\&=u^{-1}\mathscr{F}_0[M(+,q{:}h,\mathscr{S})\pm V(+,q{:}h,\mathscr{S})],\\\mathscr{F}(-,q{:}h,\mathscr{S})&=\int_{\pi/2}^{0}\oint\psi_s(q,\Theta,\varphi)\cos\Theta\,\sin\Theta\,\delta\Theta\,\delta\varphi\\&=u^{-1}\mathscr{F}_0[M(-,q{:}h,\mathscr{S})\pm V(+,q{:}h,\mathscr{S})],\end{aligned}$$

$$\mathscr{F}_c(+,q{:}h,\mathscr{S})=u^{-1}\mathscr{F}_0C(+,q{:}h,\mathscr{S})=\mathscr{F}_0\exp[-(a(h,\mathscr{S})+b(h))z(q)],$$

$$\mathfrak{T}(+,v{:}h,\mathscr{S})=u^{-1}[M(+,v{:}h,\mathscr{S})\pm V(+,v{:}h,\mathscr{S})],$$

$$\mathfrak{R}(-,0{:}h,\mathscr{S})=u^{-1}[M(-,0{:}h,\mathscr{S})\pm V(-,0{:}h,\mathscr{S})],$$

where the matrix elements $C(+,q{:}h,\mathscr{S})$ match the Beer–Lambert extinction law and $\mathfrak{T}(+,v{:}h,\mathscr{S})$ and $\mathfrak{R}(-,0{:}h,\mathscr{S})$ are the scattered transmittance and reflectance, respectively, at $z=z(v)=d$ and $z(0)=0$. The matrix elements $V(\pm,q{:}h,\mathscr{S})$ are the flux statistical variances or fluctuations due to u being finite. These variances can be computed from the positive and negative residuals of $M(+,q{:}h,\mathscr{S})$ and $M(-,q{:}h,\mathscr{S})$, with respect to the means $\langle M(+,q{:}h,\mathscr{S})\rangle$ and $\langle M(-,q{:}h,\mathscr{S})\rangle$ over a number of repeated calculations using the same input parameters. From

these, one can estimate variances. For typical computations they range from 2% to 5% (typically less than the symbol size for graphical representation). The $V(\pm,q{:}h,\mathscr{S})$ only begin to grow unacceptably large for very deep penetrations $z(q)$, which is avoided since d is finite or negligible for realistic problems. The collective ensemble behavior of the individual stochastic elements $\{\gamma(p)\}$ gives rise to stable macroscopic behavior, in this case the quantative flux solutions.

VII. COMPARISON IN THE DIFFUSION LIMIT

Since there are no general analytical solutions for $\mathscr{F}(\pm,q{:}h,\mathscr{S})$, the scattered flux solutions (6.1) are tested in the macroscopic diffusion limit, for a set of conditions where the diffusion limit is valid. This requires $b(h)\gg a(h,\mathscr{S})$ and weak anisotropy. The radiance of the transport equation (4.4) is expanded out as

$$\psi_s(\mathbf{x},\hat{\omega})=\phi(\mathbf{x})+(3/4\pi)D(\mathbf{x})\cdot\hat{\omega}, \tag{7.1}$$

where $\phi(x)$ is the average diffuse scalar intensity and $D(\mathbf{x})$ is the diffuse vectorial flux. When the generic phase function $\mathscr{P}(\theta)$ is a function of the angle θ only, a standard calculation[6,10] will give the linear Helmholtz scalar diffusion equation for $\phi(z)$ (for slab geometry). For an anisotropic blood cell suspension described by $(h,\mathscr{S},g)$, irradiated at plane $z=0$ by a coherent flux density $\mathscr{F}_0$, $\phi(z)$ satisfies

$$\frac{\partial^2\phi(z)}{\partial z^2}-\zeta^2\phi(z)=-\Lambda_0\exp[-(a(h,\mathscr{S})+b(h))z], \tag{7.2}$$

where $\zeta=\sqrt{[a(h,\mathscr{S})\times b(h)]}$ and where

$$\Lambda_0=[3b(h)b(h)+3(1-g)b(h)a(h,\mathscr{S})+3gb(h)(b(h)+a(h,\mathscr{S}))][\mathscr{F}_0/4\pi]. \tag{7.3}$$

The boundary conditions are

$$\phi(z)+\alpha\frac{\partial}{\partial z}\phi(z)+\frac{\Lambda_1(z)}{2\pi}=0,\quad \text{at } z=0, \tag{7.4a}$$

$$\phi(z)-\alpha\frac{\partial}{\partial z}\phi(z)+\frac{\Lambda_1(z)}{2\pi}=0,\quad \text{at } z=d, \tag{7.4b}$$

where $\Lambda_1=[b(\mathscr{S})g]/[a(h,\mathscr{S})+(1-g)b(h))]\mathscr{F}_0\exp[-(a(h,\mathscr{S})+b(h))z]$ and $\alpha=2/[3a(h,\mathscr{S})+3(1-g)b(h)]$. The general scalar solution is the sum of a particular solution $\phi_{\text{par}}(z)$ and a complimentary solution $\phi_{\text{com}}(z)$ so $\phi(z)=\phi_{\text{com}}(z)+\phi_{\text{par}}(z)$. The solution $\phi_{\text{com}}(z)$ must satisfy $d^2\phi_{\text{com}}(z)/dz^2-\zeta^2\phi_{\text{com}}(z)=0$, so that

$$\phi_{\text{com}}(z)=\xi_1\exp(\zeta z)+\xi_2\exp(-\zeta z), \tag{7.5}$$

with ξ_1 and ξ_2 constants. The particular solution $\phi_{\text{par}}(z)$ is found by $\phi_{\text{par}}(z)=\xi_3\exp[-(a(h,\mathscr{S})+b(h))z]$ and substituting back into (7.2) to obtain $\xi_3=-\Lambda_0[(a(h,\mathscr{S})+b(h))^2-\zeta^2]$. The complete solution is then

$$\phi(z)=\xi_1\exp(\zeta z)+\xi_2\exp(-\zeta z)+\xi_3\exp[-(a(h,\mathscr{S})+b(h))z], \tag{7.6}$$

where constants ξ_1, ξ_2, ξ_3 are derived from the boundary conditions at planes $z=0$ and $z=d$. The diffuse fluxes through the same set of data planes $\{z(q)\}$ for all $q=1$ to v, as previously defined, are then

$$\phi(q)=\xi_1\exp(\zeta z(q))+\xi_2\exp(-\zeta z(q))+\xi_3\exp[-(a(h,\mathscr{S})+b(h))z(q)], \tag{7.7a}$$

$$D(+,q{:}h,\mathscr{S})=2\pi[\phi(q)+\Omega(q)], \tag{7.7b}$$

$$D(-,q{:}h,\mathscr{S})=2\pi[\phi(q)-\Omega(q)], \tag{7.7c}$$

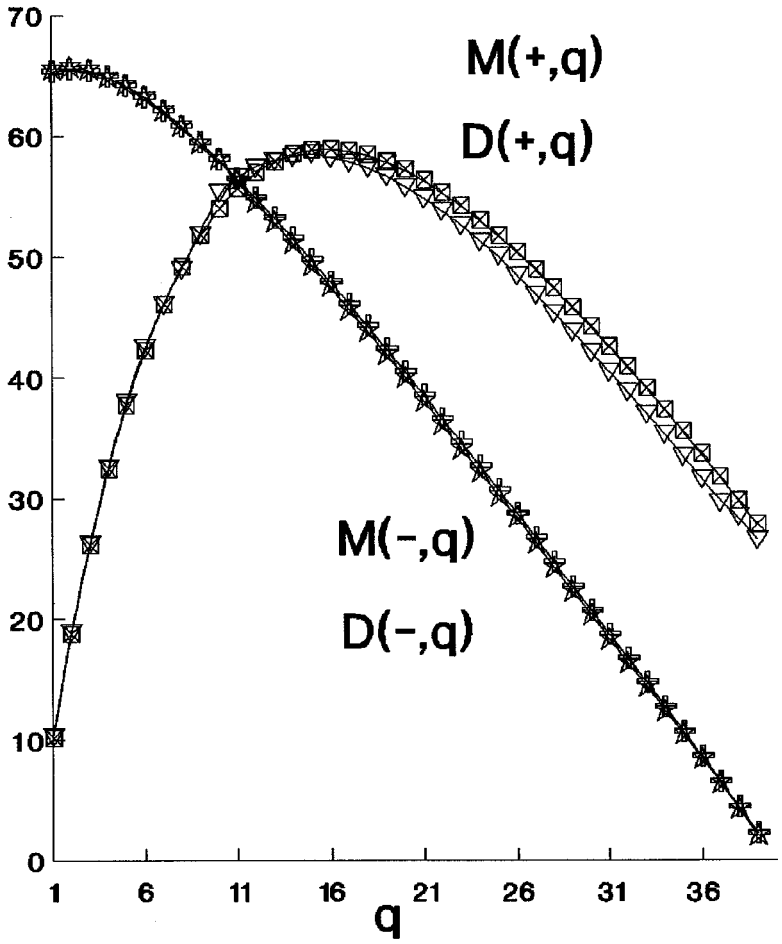

FIG. 1. Comparison of matrix elements $M(\pm,q)$ with diffusion theory solutions $D(\pm,q)$ when the diffusion approximation is strictly satisfied.

where $\Omega(z)=-(1/2\xi_2)d\phi(z)/dz$ and $\Omega(q)$ is the value at data plane $z=z(q)$. In the diffusive limit, the solutions should match the vectorial flux solutions $\mathscr{F}(\pm,q{:}h,\mathscr{S})$ computed from the stochastic algorithm:

$$D(+,q{:}h,\mathscr{S})=u^{-1}\mathscr{F}_0[M(+,q{:}h,\mathscr{S})\pm V(+,q{:}h,\mathscr{S})], \tag{7.8a}$$

$$D(-,q{:}h,\mathscr{S})=u^{-1}\mathscr{F}_0[M(-,q{:}h,\mathscr{S})\pm V(-,q{:}h,\mathscr{S})]. \tag{7.8b}$$

VIII. VERIFICATION AND ILLUSTRATIVE CALCULATIONS

The method was transformed to a compact computer code (using the language *C*) and Eq. (7.8) verified in the diffusive limit. Figure 1 shows a comparison of $D(\pm,q)$ and $M(\pm,q)$ for $a=0.1\ \mathrm{cm}^{-1}$, $b=99\ \mathrm{cm}^{-1}$, $g=0$, and $d=0.04$ cm with $u=10^4$ for $q=1-40$. This is for a set of generic test results where diffusion solutions (7.8) are strictly satisfied. These data were utilized in Ref. 37 to compare optical diffusion theory with a modified discrete ordinates algorithm adapted from nuclear engineering algorithms, for the purpose of computing laser generated thermal source profiles. The comparison is illustrated in Fig. 1, where $\mathscr{F}_0/u=1$. Here, $M(\pm,q)$ give the same solutions with as much accuracy as discrete ordinates, but far more easily, and with greatly reduced computational requirements. Unlike the $D(\pm,q)$ solutions, the $M(\pm,q)$ solutions can be derived for all possible optical and anisotropy conditions. Figures 2 and 3 quickly illustrate typical scattered flux computations into blood medias and erythrocytic suspensions using $a(h,\mathscr{S})$ and

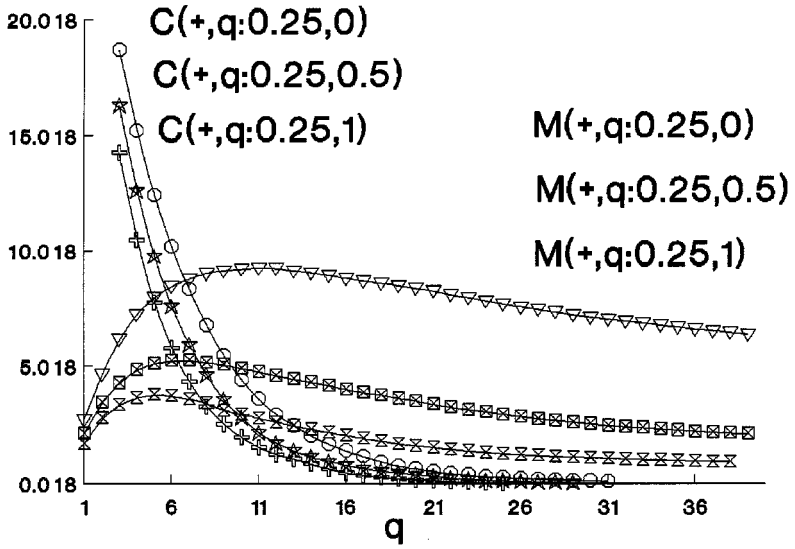

FIG. 2. Penetration of coherent or Fermat fluxes $C(+,q{:}h,\mathscr{S})$ and generation and transport of forward scattered fluxes $M(+,q{:}h,\mathscr{S})$ in a dilute anisotropic red cell suspension.

$b(h)$ data from Ref. 9 and with $g \gg 0$. Fluxes $D((\pm,q{:}h,\mathscr{S}))$ break down for some of these conditions. The effects on $M(+,q{:}h,\mathscr{S})$ of varying h and $\mathscr{S}$ on the intervals (0,1) and [0,1] are illustrated.

Figure 2 illustrates the generic decay of coherent optical beam $C(+,q{:}h,\mathscr{S})$ and generation of scattered fluxes $M(+,q;h,\mathscr{S})$ in a dilute erythrocytic suspension of depth $d=8$ mm, with fixed $h=0.25$, $g=0.9$, $u=3\times10^4$, with $\mathscr{S}=1$, $\frac{1}{2}$, 0 (at 695 nm) for all $q=1$ to $v=40$. As the Fermat beam decays due to scattering and absorption by suspended cells, scattered fluxes $M(+,q;h,\mathscr{S})$ are generated, which are equivalent to the exact solution $\mathscr{T}(+,q;h,\mathscr{S})$ of the radiative transport equation defined on data surfaces $\{z(q)\}$ for $q=1$ to v. The x axis is labeled by q (y-axis units are arbitrary). The effect of cell oxygenation $\mathscr{S}$ does not affect the scattering or

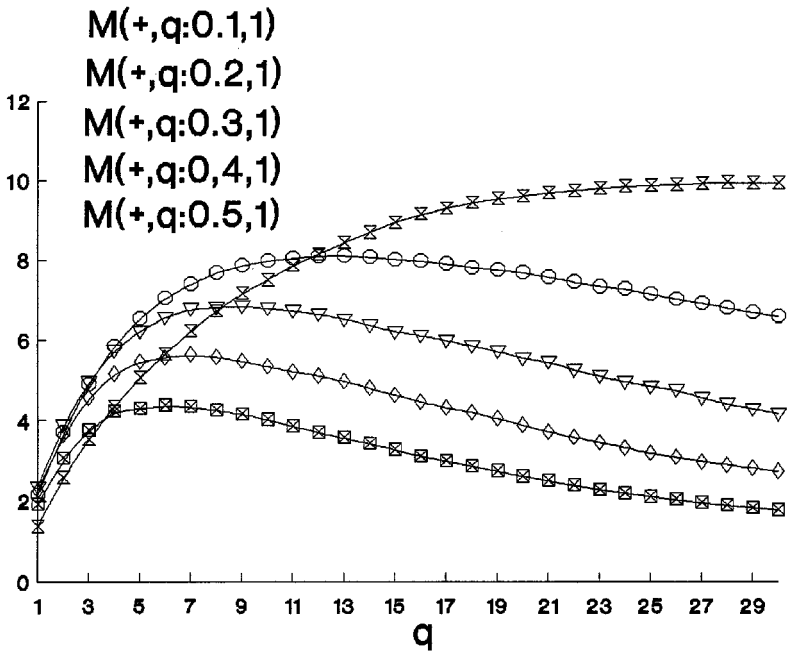

FIG. 3. Generation and transport of scattered fluxes $M(+,q{:}h,\mathscr{S})$ in anisotropic red cell suspension for various h at fixed $\mathscr{S}$.

the rate of generation of the scattered fluxes. The diffusion solution, however, breaks down for these conditions. The scattered beam penetrates through the sample since $g=0.9$ implies high anisotropy. As $\mathscr{S}$ is decreased on [0,1], the fluxes are monotonically reduced along all $z(q)$ for $q=1$ to v. Figure 3 illustrates the effect of varying h on (0,1) with fixed $\mathscr{S}$, for the forward scattered fluxes $M(+,q;h,\mathscr{S})$ for all $q=1$ to 30 along a depth of $d=8$ mm. Here, $\mathscr{S}=1$, $g=0.9$, $u=3\times10^4$, and $h=0.1$, 0.2, 0.3, 0.4, 0.5 at $\lambda=735$ nm. For $h=0.1$, the red cell suspension is dilute and so the scattered forward beam builds up more slowly through the $\{z(q)\}$ since the coherent source beam decays slowly due to the dilute number density of suspended cells. For $h=0.2$, the scattered beam builds up a maximum more rapidly since the density of suspended cellular scatter centers presented to the incident Fermat beam is greater. However, the amount of absorbing HbO_2 present also increases so the suspension is simultaneously more absorptive. Consequently, the scattered flux $M(+,q;h,\mathscr{S})$ decays more rapidly with depth. The effect is increased as h increases ($h=0.4$ corresponds to normal blood). The variances $V(+,q:h,\mathscr{S})$ for these computations are typically less than the graphical symbol size for $u>=10^4$. These conditions and dimensions are typical of many biophysical and biomedical optical/laser applications. Most accurate computations can be done in less than several minutes of cpu time using a typical UNIX mainframe.

IX. CONCLUSIONS

A stochastic-based vectorial method for finding equivalent statistical flux solutions to the radiative transfer equation for beam irradiation of suspensions or thin films of erythrocytes was developed. Although demonstrated explicitly for red blood cells, the general method should be applicable to many scattering problems related to biological or artificial suspensions of scattering–absorbing particles. The method works well for dilute suspensions as $h\rightarrow0.1$ and less, thin films ($d\rightarrow1$) and for calculating the flux contributions of multiscattered or quasi-Fermat light through both films and thick samples. The advantage over the diffusion approximation is that this approximation is not always satisfied for such medias, especially for cells or particles with nonspheroidal geometries and/or strong absorption properties, such as blood cells. The stochastic approach adapts easily to changes in boundary conditions and source beam profiles and is valid for all optical conditions. The method is also valid near the surface or the source where scattered flux begins to be generated. The advantage over Gaussian quadrature and discrete ordinates is in reduced software complexity and computing requirements. The method presented in this paper can be implemented in under 60–70 lines of compact code. Standard Monte Carlo techniques for photon transport adapted from nuclear physics deposit statistical weight in voxels and cut off photon trajectories via a roulette. This conventional method is specifically tailored toward computing scalars like fluence and average intensity in very thick medias, and requires very long computation time for even simple calculations. However, fluxes are usually more relevent. The alternative formalism will perform very many useful calculations with accuracy in typically 1–5 mins. The simple Brownian-like vectorial intersection method also enables a stochastic evaluation of the smooth scattered flux profiles to be deduced more efficiently and rapidly than the scalar fluence. The method begins to become less accurate as d increases, however, most blood vessels, laser biotargets, and realistic samples and cuvettes have $d<4$ cm, for which the method remains fast, accurate and adaptable.

ACKNOWLEDGMENTS

I would like to thank David Miller and Professor Joseph Barbenel and Professor Britton Chance.

[1] SPIE Institute Series on Biomedical Optics, 1990–1996.
[2] E. M. Sevick, J. Weng, J. Maris, and B. Chance, Proc. SPIE **1024**, 481 (1990).
[3] J. M. Steinke and A. P. Shepard, J. Opt. Soc. Am. **5**, 813 (1988).
[4] J. M. Schmitt, J. D. Meindle, and F. G. Mihm, IEEE Trans. Biomed. Eng. **BME-33**, 98 (1986).
[5] L. O. Reynolds, C. Johnson, and A. K. Ishimaru, J. Opt. Soc. Am. **70**, 1206 (1976).
[6] S. A. Prahl, Ph.D. thesis, University of Texas at Austin, 1988.
[7] D. L. Drabkin and J. H. Austin, J. Biol. Chem. **129**, 122 (1935).
[8] F. J. Janssen, Med. Biol. Eng. Comput. **10**, 231 (1972).
[9] N. R. Pisharoty, Ph.D. thesis, University of Carnegie-Mellon, Pittsburgh, 1970.

[10] I. Ishimaru, *Wave Propagation in Random Medias* (Academic, New York, 1978), Vols. I, II.
[11] S. Chandrasekhar, *Radiative Transfer* (Dover, New York, 1960).
[12] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison-Wesley, Reading, MA, 1969).
[13] V. Twersky, J. Opt. Soc. Am. **60**, 908 (1970).
[14] N. M. Anderson and P. Sekelj, Phys. Med. Biol. **12**, 185 (1967).
[15] V. Twersky, J. Opt. Soc. Am. **60**, 1084 (1970).
[16] A. Cohen and R. L. Longini, Med. Biol. Eng. Comp. **9**, 61 (1971).
[17] C. C. Johnson, IEEE Trans. Biomed. Eng. **17**, 129 (1970).
[18] L. O. Reynolds, Ph.D. thesis, University of Washington, Seattle, 1976.
[19] F. W. Turnbull, Ph.D. thesis, University of Strathclyde, Glasgow, 1976.
[20] J. C. Barbenel, F. W. Turnbull, and R. M. Nisbit, Med. Biol. Eng. Comput. **17**, 763 (1979).
[21] S. T. Flock, M. S. Patterson, B. C. Wilson, and D. R. Wyman, IEEE Trans. Biomed. Eng. **36**, 1169 (1989).
[22] B. C. Wilson, D. R. Wyman, and M. S. Patterson, J. Comput. Phys. **81**, 138 (1989).
[23] N. Metropolis and S. Ulam, J. Am. Stat. Assoc. **44**, 335 (1949).
[24] M. Motamedi, S. Rastegar, G. Le Carpentier, and A. J. Welch, Appl. Opt. **13**, 2230 (1989).
[25] R. G. Newton, *Scattering Theories of Waves and Particles* (McGraw-Hill, New York, 1966).
[26] H. C. Van De Hulst, *Light Scattering by Small Particles* (Wiley, New York, 1957).
[27] M. Kerker, *The Scattering of Light* (Academic Press, New York, 1969).
[28] C. Yeh, Phys. Rev. A **135**, 1193 (1964).
[29] G. D. Pederson, N. J. McCormick, and L. O. Reynolds, Biophys. J. **16**, 199 (1976).
[30] C. I. Beard, T. H. Kays, and V. Twersky, IEEE Trans. Antennas Propag. **AP-15. 1. 99**, 118 (1967).
[31] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory* (Van Nostrand-Reinhold, Princeton, NJ, 1970).
[32] B. Davison, *Neutron Transport Theory* (Oxford University Press, London, 1958).
[33] L. G. Henyey and J. L. Greenstein, Astrophys. J. **93**, 70 (1941).
[34] P. Liu, Phys. Med. Biol. **39**, 1025 (1994).
[35] J. Tessendorf, Phys. Rev. A **35**, 872 (1986).
[36] J. Tessendorf, J. Opt. Soc. Am. A **5**, 1410 (1988).
[37] G. Yoon, S. A. Prahl, and A. J. Welch, Appl. Opt. **28**, 12 (1989).

A note on Tauber's expanding universe in conformally flat coordinates

Geoffrey Endean[a)]
University of Durham, Old Shire Hall, Durham, DH1 3HP, England

(Received 16 June 1997; accepted for publication 29 October 1997)

The solutions given by Tauber, based on conformally flat space–time, of Einstein's field equations for the Friedmann–Robertson–Walker, closed topology, dust model are corrected. The corrected solutions are shown to reduce to the standard textbook solutions obtained directly in Friedmann–Robertson–Walker coordinates, with the parameter identified as a component of the transformation to conformally flat space–time coordinates. The general significance of Tauber's work and its relationship to more recent papers by Endean on this subject are discussed. © *1998 American Institute of Physics.* [S0022-2488(98)04103-6]

I. INTRODUCTION

The principles of cosmology in conformally flat space-time (CFS) were laid down by Infeld and Schild.[1] In a recent series of papers, Endean[2–4] has developed these principles and has concluded that the Friedmann–Robertson–Walker (FRW), closed topology, dust model formulated in CFS coordinates provides a good fit with recent observations, as well as resulting in further observational predictions. In a much earlier paper, Tauber[5] solved Einstein's field equations for FRW space–time directly in CFS. This note reconciles Tauber's solutions for the closed topology, dust model with the standard textbook results and establishes the relationship of Tauber's contribution to Endean's more recent work.

II. BASIC EQUATIONS AND NOTATION

In Infeld and Schild's[1] notation, the FRW line element ds for the case of closed three-dimensional space topology may be written

$$ds^2 = d\tau'^2 - R_1^2 d\sigma^2,$$

$$\tau = \int d\tau'/R_1(\tau'),$$

$$R(\tau) = R_1(\tau');$$

$$ds^2 = R^2(\tau)(d\tau^2 - d\sigma^2),$$

$$d\sigma^2 = d\rho'^2 + \sin^2\ \rho'(d\theta^2 + \sin^2\ \theta.d\phi^2), \tag{2.1}$$

where ρ' is FRW radius. The prime has been added to the symbol so as not to confuse it with mass density for which the same symbol is used later in this paper. The CFS line element in Infeld and Schild's[1] notation is

$$ds^2 = \gamma(t,r)\{dt^2 - [dr^2 + r^2(d\theta^2 + \sin^2\ \theta.d\phi^2)]\}. \tag{2.2}$$

The finite transformation from (2.1) to (2.2) given by Infeld and Schild[1] may be expressed as

$$(t \pm r) = \tan[(\tau \pm \rho')/2]. \tag{2.3}$$

[a)]Electronic mail: 100710.2100@compuserve.com

 © 1998 American Institute of Physics

As explained below, apart from a factor of 2 and a change of symbol from τ to S, this agrees with Endean's[2] Eq. (8) combined with his Table I.

It is unfortunate that Infeld and Schild,[1] Tauber,[5] and Endean[2] have used different notations for the CFS metric function. Endean,[2] Eq. (2); Infeld and Schild,[1] Eq. (3.1) and Table III, Type 1; and Tauber,[5] Eqs. (1.5) and (1.6), define the metric function for closed topology, CFS coordinates, respectively, by

$$g^2(r,t)=\gamma(r,t)=F(r,t)=f(\tan S)/w^2, \tag{2.4}$$

where S is the function of the conformally flat radius and time coordinates r and t defined by Endean's[2] Eq. (8) combined with his Table I and

$$w=1-\tfrac{1}{4}(t^2-r^2). \tag{2.5}$$

Tauber[5] agrees with Endean[2] on the definition of the CFS coordinates r and t but, as stated in Tauber's[5] footnote (7), this definition differs by a factor of 2 from that of Infeld and Schild.[1] Also, it should be noted that Endean[2] does not explicitly define the function f.

There are further differences of notation relating to the FRW scale factor R. Endean,[2] Eq. (6), defines

$$\frac{dS(\tau)}{d\tau}=\frac{1}{R(\tau)}, \tag{2.6}$$

whereas Infeld and Schild,[1] Eq. (4.11) and see also Eq. (2.1) above, define

$$\frac{d\tau}{d\tau'}=\frac{1}{R(\tau)}, \tag{2.7}$$

while Tauber,[5] Eq. (1.2), defines

$$\frac{d\tau}{dT}=\frac{1}{R(\tau)}. \tag{2.8}$$

It may be noted that all three assign the same meaning to R, while Endean's[2] S is exactly the same variable as the τ of both Infeld and Schild[1] and of Tauber.[5] In the rest of this paper, S is used to signify this variable in preference to τ, while τ is taken to be the unmodified FRW time as in Endean.[2]

Infeld and Schild's[1] Eq. (4.41) and Tauber's[5] unnumbered equation between Eqs. (1.7a) and (1.7b), are the same basic equation which applies to the closed topology model. It may be expressed

$$R^2=\sec^2(S)f(\tan S). \tag{2.9}$$

Endean[2] does not repeat this result. However, combining this equation with Eq. (2.4) above produces

$$g(r,t)=R\cos S/w, \tag{2.10}$$

while Endean,[2] in Eq. (12), does give the following result for closed topology

$$g(r,t)=R\sin\rho'/r. \tag{2.11}$$

Equations (2.10) and (2.11) are consistent with each other provided that

$$\frac{r}{w}=\frac{\sin\rho'}{\cos S}. \tag{2.12}$$

Remembering that the prime has been added to FRW radius so as not to confuse it with mass density, it is easily confirmed that Eq. (2.12) is indeed an identity by using Eq. (8) of Endean.[2]

III. SOLUTION OF THE FIELD EQUATIONS

Tauber[5] solves Einstein's field equations directly in conformally flat coordinates. He gives, for the solution in the case of the closed topology dust model, his equations (3.12) and (3.12a) for the effective metric function and the density, respectively,

$$f^{1/4}(\sinh\ \theta) = \text{const}\ \cosh(\tfrac{1}{2}\theta)/\cosh\ \theta, \tag{3.1}$$

$$\rho = \rho_0[\cosh\ \theta/(1+\cosh\ \theta)]^3. \tag{3.2}$$

[In these equations, the variable θ is clearly not the same as the polar coordinate θ which appears in Eqs. (2.1) and (2.2).] It can be seen by the inspection of Tauber's[5] Figs. 3 and 4 that these solutions are suspect. Attempts to prove that they are correct by substituting back into his equations shows that they are not. The correct solutions of his equations for this case may be confirmed by checking with the equations to be

$$f^{1/4}(\sinh\ \theta) = \text{const}\ \sinh(\tfrac{1}{2}\theta)/\cosh\ \theta, \tag{3.3}$$

$$\rho = \rho_0[\cosh\ \theta/(\cosh\ \theta - 1)]^3. \tag{3.4}$$

The differences are small in terms of typing, the sort of errors that are easily missed in proofreading. But they are crucial to the reconciliation of Tauber's analysis with standard theory and Endean's work.

Taking the square of Eq. (3.3) and making the substitution $\sinh\theta = \tan S$ (checking through Tauber[5] confirms that this is the correct substitution), so that $\cosh\theta = \sec S$, one obtains

$$f^{1/2}(\tan\ S) = \text{const}(1-\cos\ S)\cos\ S, \tag{3.5}$$

so that, from Eq. (2.9)

$$R = K(1-\cos\ S), \tag{3.6}$$

where K is a constant. Combining this result with Eq. (2.6) gives

$$\tau = K(S-\sin\ S). \tag{3.7}$$

On the same basis, one obtains from Eq. (3.4)

$$\rho = \rho_0/(1-\cos\ S)^3 = \text{const}/R^3, \tag{3.8}$$

where use has been made of Eq. (3.6).

IV. DISCUSSION

Now Eqs. (3.6) and (3.7) are given in standard textbooks[6] but with S identified only as a parameter. Endean[2] made the identification between the CFS transformation variable S and the parameter in the open topology case and continued with the identification for the closed topology case in Endean.[3] Tauber[5] did not make this identification but he might have if his solutions had been correct. It was because he was aware of the identification that Endean was able, without too much difficulty, to spot the errors in Tauber's solutions. Equation (3.8) is also given in standard textbooks.[6]

The presence of the error in the determination of f would also have prevented Tauber from obtaining a correct formula for the closed topology, dust model, metric function in CFS. Finally, in the open topology case, while pointing out that his closed topology analysis could easily be transferred across, Tauber considered a different transformation. Reconciliation of the results based on this different transformation with the standard, open topology transformation and with standard, open topology theory requires further work.

V. CONCLUSION

The paper of Tauber[5] is most valuable, not only because when its solutions have been corrected it provides independent confirmation of standard theory and the basis for Endean's[2–4] work, but also because it applies, not just to the dust model, but to a range of equations of state for which analytical solutions are not available in standard textbooks. While Tauber[5] extends his analysis to calculate the Hubble factor in FRW coordinates, he does not deal with redshift in CFS coordinates. Indeed, the work of Endean[2–4] in identifying the CFS coordinates as the true time and distance coordinates and working out the consequences could not be said to have been anticipated by Tauber[5] at all. In one sense this is disappointing because there was always the possibility that Tauber might have provided more independent confirmation of the correctness of Endean's analysis.

ACKNOWLEDGMENTS

The author thanks the anonymous referee of an earlier paper (Ref. 4) for drawing his attention to the paper of Tauber.[5]

[1] L. Infeld and A. Schild, Phys. Rev. **68**, 250 (1945).
[2] G. Endean, Astrophys. J. **434**, 397 (1994).
[3] G. Endean, Mon. Not. R. Astron. Soc. **277**, 627 (1995).
[4] G. Endean, Astrophys. J. **479**, 405 (1997).
[5] G. E. Tauber, J. Math. Phys. **8**, 118 (1967).
[6] D. J. Raine, *The Isotropic Universe* (Adam-Hilger, Bristol, 1981).

Fast and slow solutions in General Relativity: The initialization procedure

Mirta S. Iriondo, Enzo O. Leguizamón,[a] and Oscar A. Reula[b]
FaMAF, Medina Allende y Haya de la Torre, Ciudad Universitaria, 5000 Córdoba, Argentina

(Received 9 June 1997; accepted for publication 3 November 1997)

We apply recent results in the theory of PDE, specifically in problems with two different time scales, to Einstein's equations near their Newtonian limit. The results imply a justification to post-Newtonian approximations when initialization procedures to different orders are made on the initial data. We determine up to what order initialization is needed in order to detect the contribution to the quadrupole moment due to the slow motion of a massive body as distinct from initial data contributions to fast solutions and prove that such initialization is compatible with the constraint equations. Using the results mentioned, the first post-Newtonian equations and their solutions in terms of Green's functions are presented in order to indicate how to proceed in calculations with this approach. © *1998 American Institute of Physics.* [S0022-2488(98)02902-8]

I. INTRODUCTION

A rigorous justification of the Newtonian limit approximation in General Relativity has been given in recent papers (Refs. 1 and 2). Assuming either symmetric hyperbolic equations for the matter (including appropriate boundary conditions for it), Ref. 1, or Vlasov type matter[2] it has been shown that given a Newtonian solution, there exists a nearby general relativistic solution for a time interval which is independent of the limiting parameter. The proofs of these results rely on ideas pioneered by Kreiss on dynamical systems with different time scales which, in turn, rely on energy estimates for symmetric hyperbolic systems.

The above mentioned results are based on an initialization procedure, that is, the proximity of full relativistic solutions to the Newtonian ones is obtained by choosing the initial data in a very special way, basically ensuring that the time derivatives of the solutions at the initial surface stay bounded in that limit. This initialization procedure is only needed to a finite order on the limiting parameter (which is taken to be one over the speed of light; rigorously it can be taken as $\varepsilon = v/c$, where v is a characteristic velocity of the matter fields, for example, the speed of sound) implying that the fast behaving parts of solutions stay under control order by order along the evolution.

In this paper we advance further into this problem by showing two things: first, that the General Relativity equations can be cast into a form on which the standard theory of different time scales applies (c.f. Refs. 3, 4, and 5). This substantially improves the results of Ref. 1, because it avoids singular terms outside the principal part which had to be dealt with in a very involved way. This gives further information on the behavior of the fast part of the solutions, which is the part that comes from the failure to initialize data to all orders. In particular, it gives the equations that the leading order fast behaving parts of the solutions satisfy and so its bulk behavior. This result should be important in studying the scattering of gravitational waves on slow varying sources.

To be able to apply these standard results (estimates) we shall assume that the sources for Einstein's equations also satisfy symmetric hyperbolic equations which are regular in the limit, and, if boundary conditions are needed to deal with them, that they are of such a nature as to allow for the estimates to hold. Admittedly there may be some problems to grant this to hold for some specific cases, but that is a problem on our understanding of the description of normal matter, and

[a] Fellow of Se.CyT-UNC.
[b] Member of CONICET.

 © 1998 American Institute of Physics

is not very much related to the dynamics of the gravitional degrees of freedom, so we do not address this issue further.

Second, that the initialization procedure can be done to higher enough orders so that the quadrupole formula should follow. The initialization procedure implies some relations between initial data, besides the one implied by the constraint equations, and so the above mentioned results imply that we have to show that certain elliptic systems of equations have solutions with the appropriate asymptotic behavior. The result is only an argument, for the method used does not allow us to have estimates, and so to control evolution, all the way up to some portion of future null infinity, the region where the quadrupole formula should hold. A rigorous result for that issue should follow by studying evolution either along null cones, as considered by Winicour,[6] or along asymptotically null surfaces as studied by Friedrich.[7]

The plan of the paper is as follows: in Section II we introduce the general theory of different time scales systems and quote the relevant theorem. We then follow Ref. 1 and cast Einstein's equations as a symmetric–hyperbolic–elliptic system. By assigning units to the fields appearing in those equations, one can determine how the parameter ε appears in the system and in this way the solutions become a one parameter family of solutions to Einstein's equations. A further rescaling of the fields is needed so that the structure of the equations is in the form assumed in the hypothesis of the mentioned theorem.

In Section III we calculate the flux of energy coming from the slow part of the solution (i.e., the contribution to the flux of energy due to the slow motion of a massive body), and the flux of energy coming from the initial data without sources, that is the contribution from fast solutions. We conclude that one should initialize up to order three in ε, in order to isolate the contribution to the energy flux coming from the slow part of the solution. We further prove a Lemma stating that such initialization is consistent with the constraint equations, and so that solutions with these characteristics exists.

In an Appendix we look at the solution to the initialized data to order ε, and obtain an explicit solution in terms of the shift vector N^a and the source S^{ab}. This is done in order to illustrate how the general setting of the theory of different time scales produces the correct post-Newtonian equations to that order, and how to proceed if one wishes to compute higher order corrections.

II. PRELIMINARIES

The study of systems with different time scales reduces to the study of systems of partial differential equations which are singular in the limit of one of these time scales going to zero. The study of such a limit distinguishes between different classes of solutions according to their limiting behavior. Solutions which behave smoothly on the limit are called slow, they move according to the time scale which remains finite, those which do not have a well defined limit are called fast, they have dependence on the time scale we are setting to zero. As an example we consider the following system:

$$u_t = \frac{1}{\varepsilon}(u_x - v_x), \tag{1}$$

$$v_t = v_x . \tag{2}$$

Since the system is linear we can consider its Fourier modes, with the ansatz

$$\begin{pmatrix} u \\ v \end{pmatrix}(x,t) = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \mathrm{e}^{i(kx-\omega t)},$$

we obtain

$$\begin{pmatrix} u \\ v \end{pmatrix}(x,t) = u_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mathrm{e}^{ik(x+t/\varepsilon)} + v_0 \begin{pmatrix} \dfrac{\varepsilon}{1-\varepsilon} \\ 1 \end{pmatrix} \mathrm{e}^{ik(x+t)}.$$

It is clear that the first term in this expression is a fast solution, while the second term is a slow one. Notice that for a solution to be slow it is necessary that its time derivative stay bounded, in particular at $t=0$. One of the main results of the theory of different time scales is that, under some circumstances, this condition is also sufficient. In particular this allow us to pick slow solutions from initial data requiring that a number of consecutive time derivatives at the initial time are bounded; this procedure is called initialization. Another interesting output of the theory is that it gives precise information on the evolution of the fast and slow parts.

We turn now to the general setting and quote the main results of the theory in the following theorem.

Theorem II.1 (Ref. 8; see also Ref. 9): *Let*

$$A^0{}_{ij}(\varepsilon u^k)\frac{\partial}{\partial t}u^j=\left(\frac{1}{\varepsilon}K^a{}_{ij}+A^a_{1ij}(u^k,\varepsilon)\right)\partial_a u^j+B_i(u^k,\varepsilon), \tag{3}$$

be a symmetric-hyperbolic system containing the small parameter ε; i.e., let the matrices K^a, A^a_1, and A^0 be symmetric and A^0 positive definite. Assume that the matrices A^a_1 and the vector B are continuous in ε uniformly in u for bounded u, and are C^s in u uniformly in ε with $s\geqslant s_0\equiv[n/2]+2$ and that the initial data $u(0,x,\epsilon)=(q_{ab}(x,0),p_{ab}(x,0),r^a{}_{bc}(x,0))$ lies in $H^s(\mathbf{R}^n)$. Then the solution $u(t,x,\epsilon)$ exists for a time T independent of ε. Moreover if the initial data for (3) is

$$u(0,x,\epsilon)=u_0(x)+\varepsilon\ \ u_1(x)+\mathcal{O}(\varepsilon^2),$$

where

$$K^a\partial_a u_0=0;$$

then

$$u(t,x,\epsilon)=u^0(t,x)+\epsilon[v^0(t,x)+\tilde{u}(t,x,\epsilon)]+\mathcal{O}(\epsilon^2),$$

with u^0 and $\tilde{u}$ satisfying

$$\begin{aligned} A^0(0)u^0_t+A^a_1(u^0,0)\partial_a u^0+K^a\partial_a v^0&=B_0,\\ K^a\partial_a u^0&=0,\\ u^0(0,x)&=u_0(x), \end{aligned} \tag{4}$$

and

$$\begin{aligned} [A^0(0)+\varepsilon u^0A^0_u(0)]\tilde{u}_t+A^a_1(u^0,0)\partial_a\tilde{u}+\frac{1}{\varepsilon}K^a\partial_a\tilde{u}&=B_1,\\ \tilde{u}(0,x)&=u_1(x)-v^0(0,x), \end{aligned} \tag{5}$$

respectively,

$$\begin{aligned} B_0&=B(t,x,u^0,0),\\ B_1&=B_\varepsilon(t,x,u^0,0)+(v^0+\tilde{u})B_u(t,x,u^0)\\ &\quad-(v^0+\tilde{u})A^a_u(u^0,0)\partial_a u^0-A^0(0)v^0_t-A^a(u^0,0)\partial_a v^0, \end{aligned}$$

and v^0 chosen in such a way that $K^a\partial_a u^0=0$.

The solutions of (5) are fast solutions depending on the choice of the initial data. There exists further local (in $\mathbf{T}^n$) refinements of these results, that give us information about the behavior of

these fast solutions; see, for example, Refs. 10, 11, and 5. These results should be a useful tool for the understanding of problems like the scattering of gravitational waves by slow moving bodies.

If we wanted the fast solution to appear in the second order of ε, u^0, and u^1 would have to satisfy equations similar to (4) and $\tilde{u}$ one similar to (5) (with different sources), so that u_t and u_{tt} remain bounded when $\varepsilon\to 0$ at $t=0$. Summarizing we can suppress the fast part up to $\mathcal{O}(\varepsilon^n)$, demanding that n-time derivatives stay bounded when $\varepsilon\to 0$.This procedure is called *initialization*; for more details see Refs. 8 and 3. Note that in the case of General Relativity the choice of initial data for the initialization procedure is not trivial for it must be consistent with the constraint equations.

In order to give a formulation of the post-Newtonian limit in General Relativity as an initialization procedure, we use the variables given in Ref. 1, such that the Einstein equations can be put as a symmetric hyperbolic–elliptic system. Defining the lapse function as

$$N\equiv\frac{1}{1-\varepsilon^2 U}, \tag{6}$$

with U being the Newtonian potential and

$$r^{ab}{}_c\equiv\frac{1}{2\varepsilon^3}\left(\partial_c q^{ab}-\frac{1}{2}q^{ab}q_{ed}\partial_c q^{ed}\right)\equiv\frac{1}{2\varepsilon^3\sqrt{q}}\partial_c(\sqrt{q}q^{ab}), \tag{7}$$

$$p^{ab}\equiv\frac{1}{\varepsilon^2}\bar{\pi}^{ab}, \tag{8}$$

where ∂_c is the derivative operator associated to a flat three-metric e^{ab} and q_{ab} is the three-metric induced on the hypersurfaces $t=$const. The evolution equations become

$$\begin{aligned}\dot{q}^{ab}=&-8\varepsilon^2 q^{ab}N\dot{U}+2\varepsilon^2(1-4\varepsilon^2U)N^3q^{1/2}(q^{ab}p-p^{ab})\\&-2\varepsilon^2q^{1/2}N\bar{D}^{(a}N^{b)}+2\varepsilon^2q^{ab}\bar{D}_cN^c,\end{aligned} \tag{9}$$

$$\begin{aligned}\dot{p}^{ab}=&-q^{3/4}N^{1/2}\frac{1}{\varepsilon}(q^{cd}\partial_d r^{ab}{}_c-2q^{c(a}\partial_c r^{b)d}{}_d)-2q^{1/2}\frac{1}{\varepsilon^2}q^{ab}(\Delta U-\rho)\\&+2S^{ab}+\varepsilon^2N^c\partial_c p^{ab}+\varepsilon F^{ab}(\varepsilon,r^{de}{}_c,p^{de},\partial_c U,\partial_c N^d),\end{aligned} \tag{10}$$

$$\begin{aligned}\dot{r}^{ab}{}_c=&-\frac{N^{1/2}}{q^{1/4}}\frac{1}{\varepsilon}(\partial_c p^{ab}-2\delta^{(a}_c\partial_d p^{b)d})+\frac{1}{\varepsilon}\{q^{ab}\partial_c\partial_d N^d-q^{d(b}\partial_c\partial_d N^{a)}\\&+\frac{4N^{1/2}}{q^{1/4}}\frac{1}{\varepsilon}\delta^{(a}_c J^{b)}+\frac{2}{\varepsilon}q^{ab}\partial_c(N\dot{U})+\varepsilon^2N^d\partial_d r^{ab}{}_c-2\varepsilon Nq^{ab}N^d\partial_c\partial_d U\\&+\varepsilon F^{ab}{}_c(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U,\dot{U},\partial_c N^a),\end{aligned} \tag{11}$$

where N^a is the shift, $\bar{D}_a$ is the covariant derivative associated with $\bar{q}_{ab}$, $\Delta\equiv\bar{q}^{ab}\partial_a\partial_b$, $F^{ab}=F^{ab}(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U,\partial_c N^a)$, $F^{ab}{}_c=F^{ab}{}_c(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U,\dot{U},\partial_c N^a)$, and all F's in the equations are smooth pointwise functions of all their arguments and the quadratic terms in $r^{ab}{}_c$ and p^{ab} are the highest powers the factors will appear in and are $\mathcal{O}(\varepsilon^2)$.

In these variables, the constraint equations become

$$\Delta U-\frac{1}{2}\varepsilon\partial_c r^{cd}{}_d=\rho+\varepsilon^2F(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U), \tag{12}$$

$$-2\partial_c p^{ca}=4J^a+\varepsilon^2F^a(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U), \tag{13}$$

$$r^{ab}{}_c=\frac{1}{2\varepsilon^3\sqrt{q}}\partial_c(\sqrt{q}q^{ab}). \tag{14}$$

The system (10)–(11) is symmetric hyperbolic, and has an energy estimate finite for fixed ε . In order to have an energy estimate bounded in the limit when $\varepsilon\to 0$ the authors in Ref. 1, choose a gauge (i.e., a selection of lapse and shift) that makes $r^{ab}{}_b\equiv 0$ and $\dot{r}^{ab}{}_b=0$. As a matter of fact one could relax this gauge choice to $r^{ab}{}_b\equiv\mathcal{O}(\varepsilon^\beta)$ and $\dot{r}^{ab}{}_b=\mathcal{O}(\varepsilon^\beta)$ where β is an arbitrary positive number that guarantees that the energy estimates stay bounded even when $\varepsilon\to 0$. In this paper we choose $\beta=3$, the reason for this choice depends on the initialization procedure and it shall become clear in the next Section.

This choice implies the following equations for N^a:

$$\partial_c D^c N^b-\partial^b D_c N^c-4(J^b+\partial^b N\dot{U})+2\varepsilon^2 q^{cb}N^d\partial_c\partial_d U-\varepsilon^2 G^b=0, \tag{15}$$

$$D_d N^d=-2N\dot{U}; \tag{16}$$

combining equations (15) and (16) we get

$$\partial_c D^c N^b+\partial^b D_c N^c-4J^b+2\varepsilon^2 q^{cb}N^d\partial_c\partial_d U-\varepsilon^2 G^b=0, \tag{17}$$

and the constraint equation (12) becomes

$$\Delta U=\rho+\varepsilon^2 F(\varepsilon,r^{ab}{}_c,p^{ab},\partial_c U). \tag{18}$$

Thus the system given by (10), (11), (13), (14), (17), and (18) constitutes a symmetric hyperbolic–elliptic system; the hyperbolic and elliptic part are given by the equations (10), (11) and (17), (18), respectively.

By inspection one can see that the source $B_i(u^k,\varepsilon)$ in (10) and (11) takes the form $B_i(u^k,\varepsilon)=(C_{ij}/\varepsilon)\,u^j+F_i(u^k,\varepsilon)$, where C_{ij} is a constant matrix. Since we want to apply Theorem II on this system, the source has to be smooth in ϵ, thus we rescale the variables (a similar rescaling is used in Ref. 2) in the following way:

$$\overline{r}^{ab}_c=\varepsilon r^{ab}_c, \quad \overline{p}^{ab}=\varepsilon p^{ab}, \quad \text{and} \quad q^{ab}=e^{ab}+\varepsilon^2\hat{h}^{ab}. \tag{19}$$

Since F^{ab} and F^{ab}_c have quadratic terms in the variables r^{ab}_c and p^{ab} , we obtain a system such as (3) with a source term depending smoothly on ε, namely,

$$\begin{aligned}\dot{q}^{ab}=&-8\varepsilon^2 q^{ab}N\dot{U}+2\varepsilon(1-4\varepsilon^2 U)N^3 q^{1/2}(q^{ab}\overline{p}-\overline{p}^{ab})\\&-2\varepsilon^2 q^{1/2}N\overline{D}^{(a}N^{b)}+2\varepsilon^2 q^{ab}\overline{D}_c N^c,\end{aligned} \tag{20}$$

$$\begin{aligned}\dot{\overline{p}}^{ab}=&-q^{3/4}N^{1/2}\frac{1}{\varepsilon}(q^{cd}\partial_d\overline{r}^{ab}{}_c-2q^{c(a}\partial_c\overline{r}^{b)d}{}_d)-2q^{1/2}\frac{1}{\varepsilon}q^{ab}(\Delta U-\rho)\\&+2\varepsilon S^{ab}+\varepsilon^2 N^c\partial_c\overline{p}^{ab}+\varepsilon F^{ab}(\varepsilon,\overline{r}^{de}{}_c,\overline{p}^{de},\partial_c U,\partial_c N^d),\end{aligned} \tag{21}$$

$$\begin{aligned}\dot{\overline{r}}^{ab}{}_c=&-\frac{N^{1/2}}{q^{1/4}}\frac{1}{\varepsilon}(\partial_c\overline{p}^{ab}-2\,\delta^{(a}_c\partial_d\overline{p}^{b)d})+q^{ab}\partial_c\partial_d N^d-q^{d(b}\partial_c\partial_d N^{a)}\\&+\frac{4N^{1/2}}{q^{1/4}}\delta^{(a}_c J^{b)}+2q^{ab}\partial_c(N\dot{U})+\varepsilon^2 N^d\partial_d\overline{r}^{ab}{}_c-2\varepsilon^2 N q^{ab}N^d\partial_c\partial_d U\\&+\varepsilon^2 F^{ab}{}_c(\varepsilon,\overline{r}^{ab}{}_c,\overline{p}^{ab},\partial_c U,\dot{U},\partial_c N^a).\end{aligned} \tag{22}$$

In the same gauge as before, the constraint equations and the equation for the shift vector become

$$\Delta U = \rho + \varepsilon^2 F(\varepsilon, \overline{r^{ab}}_c, \overline{p^{ab}}, \partial_c U), \tag{23}$$

$$-2\partial_c \overline{p^{ca}} = 4\epsilon J^a + \varepsilon^2 F^a(\varepsilon, \overline{r^{ab}}_c, \overline{p^{ab}}, \partial_c U), \tag{24}$$

$$\overline{r^{ab}}_c = \frac{1}{2\varepsilon^2\sqrt{q}}\partial_c(\sqrt{q}q^{ab}), \tag{25}$$

and

$$\partial_c D^c N^b + \partial^b D_c N^c - 4J^b + 3\varepsilon F^b + \varepsilon^2(4Nq^{cb}N^d\partial_c\partial_d U - F^{ab}{}_b) = 0. \tag{26}$$

From now on, to simplify the notation, we will work without bars on the variables p^{ab} and $r^{ab}{}_c$. As stated in the Introduction we assume that the above system is part of a bigger symmetric hyperbolic system which includes the equations for the sources. These are assumed to be regular in ε and of such a type that global estimates can be obtained.

The existence of a smooth one parameter family of solutions is guaranteed because the energy estimates corresponding to the hyperbolic part of the system stay bounded when $\varepsilon \to 0$ via a particular gauge choice, the elliptic variables satisfy a Gårding estimate in terms of the hyperbolic ones, and the boundedness in time of the solutions is guaranteed via the initialization procedure.

Another interesting problem which can be tackled with these techniques is whether for each given fast post-Newtonian solution (i.e., a fixed slow background solution and a highly oscillating part), there exists a solution to the full Einstein equations in the *same interval of existence* and that it remains close for that whole interval to the first one in some norm. In other words, whether the solution to (3) exists and converges to the solution of the limit equation for as long as the General Relativistic solution exists.

III. THE POST-NEWTONIAN LIMIT AS AN INITIALIZATION PROCEDURE

The Theorem quoted above implies that solutions to Einstein's equations [the system given by (20) to (22)], can be split into slow and fast ones, and this is dictated by the choice of initial data.

In the context of General Relativity, the slow part of the solution contains information about the matter fields that are moving with velocity much lower than c, that is with their own time scale, while the fast part contains information essentially due to the arbitrariness of initial data and so is not related directly to the sources.

In the context of symmetric–hyperbolic systems, this means that on the post-Newtonian approximations the highly oscillatory part (fast part) of the solution should be suppressed up to higher order of ε. For example, if we initialize (choose initial data such that the time derivative of the solution is bounded independently of ε at $t=0$) up to order ε, the slow part appears at zeroth order and the fast one at first order. The degree of initialization needed depends on the problem at hand. In this Section we show that to get the quadrupole formula one needs to initialize up to order ε^3. This guarantees that no contribution from fast solutions will be present in the gravitational radiation flux, to that order. We then show that such an initialization is possible.

To know to which order of ε we have to initialize, we estimate the contribution to the energy flux at $\mathscr{I}^+$ from both sources and vacuum (pure initial data) solutions. Since we are only interested in the order of ε at which these contributions appear, it is enough to estimate them in linearized gravity, for example see Ref. 12.

To this end we keep the physical units in the Einstein equations (i.e., $k = G \neq 1$), and we check the power at which ε appears in the radiation due to the matter (quadrupole contribution, with zero initial data) and in the radiation due to the initial data (without sources).

In this approximation,

$$g_{\mu\nu} = \eta_{\mu\nu} + \delta g_{\mu\nu},$$

and, in the absence of matter, the energy flux in the x^1 direction due to the initial data can be calculated as

$$ct^{01}=\frac{1}{32\pi k\varepsilon^3}\left[\delta\dot{g}_{23}^2+\frac{1}{4}(\delta\dot{g}_{22}-\delta\dot{g}_{33})^2\right],$$

where the dot means a time derivative. In the presence of matter, with initial data $\delta g_{\mu\nu}=\delta\dot{g}_{\mu\nu}=0$, we can use the same formula as above because at a large distance from the bodies the matter waves can be considered as plane waves. By (19) $\delta g_{ij}=\varepsilon^2\hat{h}_{ij}$, the flux of energy becomes

$$ct^{01}=\frac{\varepsilon}{32\pi k}\left[\dot{\hat{h}}_{23}{}^2+\frac{1}{4}(\dot{\hat{h}}_{22}-\dot{\hat{h}}_{33})^2\right],$$

Thus choosing the initial data $\hat{h}_{ij}=\varepsilon^\alpha f_{ij}$, the flux becomes $ct^{01}=\mathcal{O}(\varepsilon^{2\alpha+1})$, while the flux due to the matter is calculated as $ct^{01}=\mathcal{O}(\varepsilon^5)$ (see the Appendix). We conclude that initializing up to order ε^3, the fast part of the solution will contribute at a higher order in ε to the flux. This means that the initial data will be chosen as

$$u|_{t=0}=u_0+\varepsilon u_1+\varepsilon^2u_2+\varepsilon^3u_3+\mathcal{O}(\varepsilon^4).$$

This is equivalent to demanding that the first four time derivatives of the solution stay bounded in $t=0$ when $\varepsilon\to0$ (c.f. Ref. 3).

Besides a hyperbolic system, Eintein's equations also include a constraint system, that is, equations which relate the otherwise arbitrary initial data fields with each other. Thus it must be checked that these constraints are consistent with the initialization procedure. To do that, we add to the equations arising from the initialization the constraint equations, and show that in the resulting system the hierarchy of elliptic systems admits solutions with the correct asymptotic fall off.

Defining the pseudotensor h^{ab},

$$\sqrt{q}q^{ab}=\sqrt{e}e^{ab}+\varepsilon^2h^{ab}, \tag{27}$$

and from the definition of $r^{ab}{}_c$ given by (25), we obtain

$$r^{ab}{}_c=\frac{1}{2\sqrt{q}}\partial_c(h^{ab}).$$

Choosing the gauge condition $r^{ab}{}_b=0$ and $\dot{r}^{ab}{}_b=0$ up to order ε^3, and demanding boundedness of the first four time derivatives, we get that $p_i{}^{ab}$ and $h_i{}^{ab}$, $i=0,1,2,3$ should satisfy at $t=0$ the following differential equations:

$$\begin{aligned}
&\Delta_0h_0{}^{ab}=0,\\
&\partial_cp_0{}^{ab}=0,\\
&\Delta_0h_1{}^{ab}=0,\\
&\Delta_0(p_1{}^{ab}+\partial^{(a}N_0{}^{b)})=0,\\
&\Delta_0(p_2{}^{ab}+\partial^{(a}N_1{}^{b)})=0,\\
&\Delta_0h_2{}^{ab}=2S_0{}^{ab}+F_0{}^{ab}+\partial^{(a}\dot{N}_0{}^{b)}-e^{ab}F_0,\\
&\Delta_0(p_3{}^{ab}+\partial^{(a}N_2{}^{b)})=-2\dot{S}_0{}^{ab}-\dot{F}_0{}^{ab}+2e^{ab}\dot{F}_0-\partial^{(a}\ddot{N}_0{}^{b)}+\partial^dF_0{}^{ab}{}_d\\
&\qquad\qquad-\partial^{(a}F_1{}^{b)}-2e^{ab}\partial^e(N_0{}^d\partial_e\partial_dU_0),\\
&\Delta_0h_3{}^{ab}=2S_1{}^{ab}-F_1{}^{ab}+\partial^{(a}\dot{N}_1{}^{b)}-2e^{ab}F_1,
\end{aligned}\tag{28}$$

where $\Delta_0 = e^{ab}\partial_a\partial_b$. These equations are consistent with the constraint up to order ε^3, for they have been included in the set. Thus we can prove the following

Lemma III.1: The equation system for the initial data given by (28) admits a solution.

Proof: From the first five equations, we can choose

$$p_0{}^{ab}=0,\ r_0{}^{ab}{}_c=0,\ \ r_1{}^{ab}{}_c=0,\ p_1{}^{ab}=-\partial^{(a}N_0{}^{b)}, \tag{29}$$

and

$$p_2{}^{ab}=-\partial^{(a}N_1{}^{b)}. \tag{30}$$

Therefore, in order to guarantee the existence of solutions, it is sufficient to prove that the sources in the three last equations in (28) have the decay $\mathcal{O}(1/r^3)$ (cf. Ref. 13).

Due to the choice made in (29), the functions $F_0{}^{ab}$, F_0, $F_0{}^{ab}{}_c$, F_1, $F_1{}^a$, and $F_1{}^{ab}$ are all linear functions of $\partial_a(U_0N_0{}^b)$. Similarly $\dot{F}_0$ and $\dot{F}_0{}^{ab}$ depend linearly on $\partial_t\partial_a(U_0N_0{}^b)$. But the vector fields $N_0{}^a$ and $N_1{}^a$ and the function U_0 satisfy the following differential equations:

$$\Delta_0 N_0{}^a+\partial^a\partial_c N_0{}^c=4J_0{}^a,$$

$$\Delta_0 N_1{}^a+\partial^a\partial_c N_1{}^c=4J_1{}^a,$$

$$\Delta_0 U_0=\rho_0.$$

Thus, because of the compactness of the support of the sources, the standard theory of elliptic systems in $\mathbf{R}^n$ (cf. Ref. 13) assures that U_0, $N_0{}^a$, and $N_1{}^a$, and their time and spatial derivatives decay as $\mathcal{O}(1/r)$ and $\mathcal{O}(1/r^2)$ respectively. Hence we conclude that the sources decay fast enough as to ensure the existence of solutions for $h_2{}^{ab}$, $h_3{}^{ab}$, and $p_3{}^{ab}+\partial^{(a}N_2{}^{b)}$.

In order to have a well defined $p_3{}^{ab}$, it remains to prove that $N_2{}^a$ exists. Note that the gauge condition implies that $N_2{}^a$ must satisfy

$$\Delta_0 N_2{}^a+\partial^a\partial_c N_2{}^c=4J_2{}^a-3F_1{}^a-4N_0{}^d\partial_d\partial^a U_0+2F_0{}^{ab}{}_b,$$

and that the source in this equation has a fast enough decay. Thereby the existence of $N_2{}^a$ and so of $p_3{}^{ab}$ is assured. □

This Lemma proves that we can initialize at least up to order ε^3, hence by Theorem II.1 we can assure that the fast solution appears in an order higher than ε^3

ACKNOWLEDGMENTS

M. S. Iriondo was supported by STINT, The Swedish Foundation for International Cooperation in Research and Higher Education.

APPENDIX: THE EVOLUTION OF THE INITIALIZED DATA

In order to exemplify the initialization procedure, we initialize up to order ε and solve explicitly the equation for u^0, u^1, and v^0, choosing the initial data u_0 and u_1 as in (29). In fact this choice gives that $\tilde{u}$ becomes a slow solution suppressing the fast one to an higher order of ϵ.

Claim: *Given the initial data*

$$u(0,x)=\varepsilon\begin{pmatrix} p_1^{ab} \\ r_{1c}{}^{ab}\end{pmatrix}(x)+\mathcal{O}(\varepsilon^2)$$

$$=\varepsilon\begin{pmatrix} -\partial^{(a}N^{b)} \\ 0\end{pmatrix}(x)+\mathcal{O}(\varepsilon^2),$$

where $N^a(x)$ is the solution of the elliptic equation (32) for $t=0$, the solution of (21) and (22) can be written as

$$u(t,x)=\varepsilon\begin{pmatrix} p_1^{ab} \\ r_{1c}{}^{ab} \end{pmatrix}+\mathcal{O}(\varepsilon^2)$$

$$=\varepsilon\begin{pmatrix} \tilde{p}^{ab}+v^{ab} \\ \tilde{r}^{ab}{}_c+v^{ab}{}_c \end{pmatrix}+\mathcal{O}(\varepsilon^2),$$

with

$$v=\begin{pmatrix} -\partial^{(a}N^{b)} \\ 0 \end{pmatrix}$$

and

$$\tilde{p}^{ab}=\int_{|\mathbf{x}-\mathbf{x}'|\leqslant t/\varepsilon}\frac{\Gamma_1^{ab}(t-\varepsilon|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d^3x'+\partial^{(a}N^{b)},$$

$$\tilde{r}_c^{ab}=\varepsilon\,\partial_c\int_{|\mathbf{x}-\mathbf{x}'|\leqslant t/\varepsilon}\frac{\Gamma_2^{ab}(t-\varepsilon|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d^3x',$$

where $\Gamma_1^{ab}=-(\Delta\partial^{(a}N^{b)})+2\varepsilon^2\dot{S}^{ab}$ *and* $\Gamma_2^{ab}=\partial^{(a}\dot{N}^{b)}+2S^{ab}$, ρ *and* J^a *have compact support in* R^3. *Furthermore, the asymptotic behavior of* $\tilde{u}$ *is* $\mathcal{O}(1/r)$.

Proof: Since $u_0=0$ and thereby $K^a\partial_a u_0=0$, Theorem II.1 assures that

$$u(t,x,\epsilon)=u^0(t,x)+\epsilon[v(t,x)+\tilde{u}(t,x,\epsilon)]+\mathcal{O}(\epsilon^2).$$

By applying (4) and (5) on the equations system given by (21) and (22), we obtain that u^0, and v will satisfy

$$\begin{aligned} &\dot{p}_0^{ab}+\partial^c v^{ab}{}_c-2\,\partial^{(a}v^{b)d}{}_d=0,\\ &\dot{r}_{0c}{}^{ab}+\partial_c v^{ab}-2\,\delta_c^{(a}\partial_d v^{b)d}=-\partial_c\partial^{(b}N^{a)}+4\,\delta_c^{(a}J^{b)},\\ &K^a\partial_a u^0=0. \end{aligned}\tag{A1}$$

At this order of ε the constraint equations and the gauge become

$$\begin{aligned} &\Delta U=\rho,\\ &-2\,\partial_c p_0^{ca}=0,\\ &r_{0c}{}^{ab}=\frac{1}{2}\partial_c h_0^{ab},\\ &\partial_c\partial^c N^b+\partial^b\partial_c N^c=4J^b, \end{aligned}\tag{A2}$$

where h^{ab} is a pseudotensor defined in (27). Choosing

$$v=\begin{pmatrix} -\partial^{(a}N^{b)} \\ 0 \end{pmatrix},$$

the system (A1) becomes

$$\begin{aligned} &u_t^0=0,\\ &K^a\partial_a u^0=0, \end{aligned}\tag{A3}$$

$$u_0(x)=0,$$

which is trivially satisfied by $u^0(t,x)=0$.

Considering now the equation for $\tilde{u}$, namely

$$\tilde{u}_t+\frac{1}{\varepsilon}K^a\partial_a\tilde{u}=B_1, \tag{A4}$$

$$\tilde{u}(0,x)=u_1(x)-v^0(0,x)=0,$$

with

$$B_1=\begin{pmatrix} 2S^{ab}+\partial^{(a}\dot{N}^{b)} \\ 0 \end{pmatrix},$$

we obtain the following system:

$$\dot{\tilde{p}}^{ab}+\frac{1}{\varepsilon}\partial^c\tilde{r}^{ab}{}_c=2S^{ab}+\partial^{(a}\dot{N}^{b)}, \tag{A5}$$

$$\dot{\tilde{r}}^{ab}{}_c+\frac{1}{\varepsilon}\partial_c\tilde{p}^{ab}=0, \tag{A6}$$

and the constraint and the gauge become

$$\Delta U=\rho,$$

$$\partial_c p_1^{ca}=-2J^a,$$

$$r_{1c}{}^{ab}=\frac{1}{2}\partial_c h_1^{ab},$$

$$\partial_c\partial^c N^b+\partial^b\partial_c N^c=4J^b. \tag{A7}$$

Since $\tilde{p}^{ab}=p_1^{ab}-v^{ab}$ and $\tilde{r}_c{}^{ab}=r_{1c}{}^{ab}$, instead of (A5) we consider the differential equations

$$\dot{p}_1^{ab}+\frac{1}{\varepsilon}\partial^c r_{1c}{}^{ab}=2S^{ab},$$

$$\dot{r}_{1c}{}^{ab}+\frac{1}{\varepsilon}\partial_c p_1^{ab}=-\frac{1}{\varepsilon}\partial_c\partial^{(a}\dot{N}^{b)}. \tag{A8}$$

The equations system (A8) is clearly consistent with the gauge (A7). Taking the time derivative to the system above, we obtain

$$\Box p_1^{ab}=-(\Delta\partial^{(a}N^{b)}+2\varepsilon^2\dot{S}^{ab}),$$

$$\Box h_1^{ab}=2\varepsilon(\partial^{(a}\dot{N}^{b)}+2S^{ab}) \tag{A9}$$

$$:=2\varepsilon\Gamma_2^{ab}$$

and the solution of the second equation in (A9) can be written as

$$h_1^{ab}=2\varepsilon\int_{|\mathbf{x}-\mathbf{x}'|\leqslant t/\varepsilon}\frac{\Gamma_2^{ab}(t-\varepsilon|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d^3x'.$$

Because of the decay of the source, when $|\mathbf{x}| \gg |\mathbf{x}'|$, we write $|\mathbf{x}-\mathbf{x}'| \approx |\mathbf{x}| - \hat{n}\cdot\mathbf{x}'$, where $\hat{n} = \mathbf{x}/|\mathbf{x}|$. Hence in this first approximation we have

$$h_1^{ab} \approx \frac{2\varepsilon}{r} \int_{|\mathbf{x}-\mathbf{x}'| \leqslant t/\varepsilon} \Gamma_2^{ab}(t - \varepsilon r + \varepsilon \hat{n}\cdot\mathbf{x}', \mathbf{x}') d^3x',$$

similarly we calculate p_1^{ab}. □

Following this procedure we can generate the solution that has as initial data (28) and the functions $u^2(t,x)$ and $u^3(t,x)$ obey more similar equations than do $u^0(t,x)$ and $u^1(t,x)$.

Remark III.1: In the calculus of the flux of energy due to the matter fields, we use the pseudotensor field calculated above instead of the spatial components of the perturbation of the metric. The difference between them can be calculated using equations (19) and (25), and because they have the same expansion in ε they can be considered as equal.

[1] S. Fritelli and O. A. Reula, ''On the Newtonian limit of General Relativity,'' Commun. Math. Phys. **166**, 2 (1994).
[2] A. D. Rendall, ''The Newtonian Limit for Asymptotically Flat Solutions of the Vlasov-Einstein system,'' Commun. Math. Phys. **163**, 89–112 (1994).
[3] G. Browning and H. O. Kreiss, ''Problems with different time scales for nonlinear partial differential equations,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **42**, 704–718 (1982).
[4] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, (Springer-Verlag, New York, 1984).
[5] S. Schochet, ''The Fast Singular Limit of Hyperbolic PDEs,'' J. Diff. Eqns. **114**, 476–512 (1994).
[6] J. Winicour, ''Newtonian gravity on the null cone,'' J. Math. Phys. **24**, 1193–1198 (1983).
[7] H. Friedrich, ''Cauchy problems for the conformal vacuum field equations in General Relativity,'' Commun. Math. Phys. **91**, 445–472 (1983).
[8] S. Schochet, ''Asymptotics for symmetric hyperbolic systems with a large parameter,'' J. Diff. Eqns. **75**, 1–26 (1988).
[9] S. Klainerman and A. Majda, ''Compressible and incompressible fluids,'' Commun. Pure Appl. Math. **35**, 629–653 (1982).
[10] J. L. Joly, G. Métivier, and J. Rauch, ''Remarques sur l'optique geométrique nonlinéaire multidimensionelle,'' Séminar, Ecole Polytecnique, Palaiseau, 1990.
[11] J. L. Joly, G. Métivier and J. Rauch, ''Coherent and focusing multidimensional nonlinear geometric optics,'' Centre de Recherche en Mathématiques de Bordeaux, preprint, June 1992.
[12] L. Landau and E. Lifshitz, *The Classical Theory of Fields*, 2nd ed. (Addison-Wesley, New York, 1959).
[13] R. Lockhart and R. McOwen, ''On elliptic systems in $\mathbf{R}^n$,'' Acta Mat. **X**, 125–135 (1983).

Multilinear representations of rotation groups within geometric algebra

M. A. J. Ashdown,[a] S. S. Somaroo,[b] S. F. Gull, C. J. L. Doran, and A. N. Lasenby
MRAO, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 14 July 1997; accepted for publication 7 November 1997)

It is shown that higher-weighted representations of rotation groups can be constructed using multilinear functions in geometric algebra. Methods for obtaining the irreducible representations are found, and applied to the spatial rotation group, SO(3), and the proper Lorentz group, $SO^+(1,3)$. It is also shown that the representations can be generalized to non-linear functions, with applications to relativistic wave equations describing higher-spin particles, such as the Rarita–Schwinger equations. The internal spin degrees of freedom and the external space–time degrees of freedom are handled within the same mathematical structure. © *1998 American Institute of Physics.* [S0022-2488(98)00903-7]

I. INTRODUCTION

The vector and spinor representations of rotation groups are well known. In this paper we show that it is possible to extend these to higher-weighted representations by using multilinear functions of vector variables. Our representations are constructed using the language of geometric algebra, an introduction to which is given in the following subsection.

First we show how to construct the representations of an arbitrary rotation group $SO^+(p,q)$ and demonstrate the methods for finding their irreducible components. The most important method is monogenic decomposition,[1] which can be refined by considering the symmetry between the arguments of the function. A monogenic function belongs to the null space of the vector derivative (the Dirac operator). Analysis of the monogenic properties of functions forms an important part of Clifford analysis.

We use these methods to find the irreducible representations of the rotation group of Euclidean 3-space SO(3) and the proper Lorentz group $SO^+(1,3)$. In both cases we show how to find all the integral- and half-integral-spin representations of the groups. These are useful in constructing wave equations describing higher-spin particles.

In the last section we examine the Dirac equation and its generalizations to higher-spin wave equations, such as the Rarita–Schwinger equations. We show that the multilinear representations can be extended to encompass non-linear functions, and these can be used to construct the wavefunctions. This provides a unified method for describing the internal and external degrees of freedom of a higher-spin particle. In using geometric algebra we are able to overcome one of the deficiencies in the matrix formulation of linear algebra, that it is not easily able to relate linear and non-linear methods. Polynomial projection[1] enables us to immediately construct the non-linear functions which form the basis of the space–time part of the wavefunction from the multilinear functions derived earlier for the spin part of the wavefunction.

A. Geometric algebra

Geometric algebra is based upon Clifford algebra with an emphasis on the geometric interpretation of elements of the algebra rather than their matrix representations. Here we present a brief introduction to geometric algebra and the conventions that are used in this paper. More comprehensive introductions can be found in Hestenes and Sobczyk[2] and Gull, Lasenby and Doran.[3]

[a]Electronic mail: maja1@mrao.cam.ac.uk
[b]Current address: BCMP, Harvard Medical School, 240 Longwood Ave., Boston, MA 02115.

 © 1998 American Institute of Physics

Geometric algebra is a graded algebra in which the fundamental product is the geometric product. Vectors, which have grade 1, will be written in lower case Roman letters (a, b, etc.). The geometric product of vectors ab can be split into its symmetric and antisymmetric parts,

$$ab = a \cdot b + a \wedge b, \tag{1}$$

where

$$\begin{aligned} a \cdot b &\equiv \tfrac{1}{2}(ab + ba), \\ a \wedge b &\equiv \tfrac{1}{2}(ab - ba). \end{aligned} \tag{2}$$

The inner product $a \cdot b$ is scalar-valued (grade 0). The outer product $a \wedge b$ represents a directed plane segment produced by sweeping a along b. This new element is called a bivector and has grade 2. By introducing more vectors, higher-grade elements can be constructed, trivectors (grade 3) $a \wedge b \wedge c$ represent volumes and so on, up to the dimension of the space under consideration. A grade-r element of the algebra is called an r-vector and a general element of the geometric algebra, possibly consisting of many grades, is called a multivector. If we multiply an r-vector A_r by an s-vector B_s the geometric product can be expanded as follows:

$$A_r B_s = \langle A_r B_s \rangle_{r+s} + \langle A_r B_s \rangle_{r+s-2} + \cdots + \langle A_r B_s \rangle_{|r-s|}, \tag{3}$$

where $\langle C \rangle_t$ denotes projecting out the grade-t elements of C. The notation for inner and outer products is retained for the lowest and highest grade elements in this expansion,

$$\begin{aligned} A_r \wedge B_s &\equiv \langle A_r B_s \rangle_{r+s}, \\ A_r \cdot B_s &\equiv \langle A_r B_s \rangle_{|r-s|}. \end{aligned} \tag{4}$$

We use the convention that inner and outer products take precedence over geometric products in expressions.

In addition we define the commutator product, which is equal to half the commutator bracket,

$$A \times B = \tfrac{1}{2}(AB - BA) = \tfrac{1}{2}[A, B]. \tag{5}$$

Reversion is the operation by which the order of vectors in a geometric product is reversed, it is defined by

$$\widetilde{AB} = \widetilde{B}\widetilde{A}, \tag{6}$$

where a vector reverses to give itself, $\widetilde{a} = a$.

Although geometric algebra can be constructed in an entirely basis-free form, in many applications it is useful to introduce a set of basis vectors. The geometric algebra with signature (p,q) which will be denoted as $\mathcal{G}_{p,q}$ has $n = p + q$ orthonormal basis vectors, $e_i, i = 1, \ldots, n$ which obey

$$e_i \cdot e_j = \begin{cases} 1, & i = j = 1, \ldots, p, \\ -1, & i = j = p+1, \ldots, p+q, \\ 0, & i \neq j. \end{cases} \tag{7}$$

The grade-r space in this algebra is denoted $\mathcal{G}^{\,r}_{p,q}$.

The highest grade element that can be formed from this basis has grade n and is called the pseudoscalar. It is defined by

$$I \equiv e_1 \wedge e_2 \wedge \cdots \wedge e_n = e_1 e_2 \cdots e_n. \tag{8}$$

All grade-n elements must be a scalar multiple of the pseudoscalar.

In this paper we make use of some results of geometric calculus, which is based on the vector derivative. If we write a vector in terms of its components, $a=a^i e_i$, then the vector derivative with respect to a is defined as

$$\partial_a \equiv e^i \frac{\partial}{\partial a^i}, \tag{9}$$

where the reciprocal vectors are defined by

$$e^i \equiv (-1)^{i-1}(e_1 \wedge \cdots \wedge \check{e}_i \wedge \cdots \wedge e_n) I^{-1}, \tag{10}$$

and the check on $\check{e}_i$ denotes that it is omitted from the product. The vector derivative, besides being an operator, also inherits the properties of a vector from a. The vector derivative with respect to position x is given the symbol ∇,

$$\nabla \equiv \partial_x . \tag{11}$$

In this paper we use the properties of multilinear functions which are functions in geometric algebra that are linear in all of their arguments. A function,

$$M(a_1, \ldots, a_k), \tag{12}$$

which takes values in any part of the geometric algebra and is linear in its k vector arguments will be called k-linear.

There are two geometric algebras, the Pauli algebra and the space–time algebra, which for reasons of convention have their own notation.

B. The Pauli algebra

The Pauli algebra $\mathcal{G}_{3,0}$ is the geometric algebra of three dimensional Euclidean space. The three basis vectors are σ_1, σ_2 and σ_3, which satisfy

$$\sigma_j \cdot \sigma_k = \delta_{jk} . \tag{13}$$

The pseudoscalar of the algebra is

$$i \equiv \sigma_1 \sigma_2 \sigma_3 . \tag{14}$$

The algebra has eight basis elements,

1	$\{\sigma_k\}$	$\{i\sigma_k\}$	i
1 scalar	3 vectors	3 bivectors	1 pseudoscalar
grade 0	grade 1	grade 2	grade 3

(15)

C. The space–time algebra

The space–time algebra (STA) is the geometric algebra of Minkowski space–time, $\mathcal{G}_{1,3}$. The basis vectors are γ_0, γ_1, γ_2 and γ_3, which satisfy

$$\gamma_\mu \cdot \gamma_\nu = \eta_{\mu\nu} = \mathrm{diag}(+---). \tag{16}$$

We also define the bivectors,

$$\sigma_k = \gamma_k \gamma_0 \quad (k=1,2,3), \tag{17}$$

and the pseudoscalar,

$$i=\gamma_0\gamma_1\gamma_2\gamma_3=\sigma_1\sigma_2\sigma_3. \tag{18}$$

These definitions are compatible with those for the Pauli algebra given above, because the Pauli algebra is a sub-algebra of the STA. The STA has 16 basis elements,

1	$\{\gamma_\mu\}$	$\{\sigma_k, i\sigma_k\}$	$\{i\gamma_\mu\}$	i
1 scalar	4 vectors	6 bivectors	4 trivectors	1 pseudoscalar
grade 0	grade 1	grade 2	grade 3	grade 4

(19)

Vectors in the STA have a projective relation to those in the Pauli algebra,

$$a\gamma_0=a_0+\mathbf{a}, \tag{20}$$

where $a_0=a\cdot\gamma_0$ and $\mathbf{a}=a\wedge\gamma_0$. The bivector $\mathbf{a}$ in the STA corresponds to the vector in the Pauli algebra relative to the frame defined by the unit timelike vector γ_0.

II. REPRESENTATIONS OF ROTATION GROUPS

In this section we shall develop the formalism we need to describe higher-weighted representations of the rotation group $\mathrm{SO}^+(p,q)$ and find its irreducible representations.

A. The representations

Any rotation belonging to the group $\mathrm{SO}^+(p,q)$ leaves the inner product of two vectors in the geometric algebra $\mathcal{G}_{p,q}$ unchanged. Such a rotation is described by a rotor R, an even-graded element of the algebra which satisfies

$$R\tilde{R}=1. \tag{21}$$

Rotors are used to define the two types of representation of the the rotation group. The one-sided representations, which will be denoted as $\hat{\mathrm{U}}_R$, in which the rotor acts one-sidedly on a multivector ψ,

$$\hat{\mathrm{U}}_R[\psi]=R\psi, \tag{22}$$

and the two-sided representations, which will be denoted as $\hat{\mathrm{V}}_R$, in which the rotor R acts two-sidedly on a multivector M,

$$\hat{\mathrm{V}}_R[M]=RM\tilde{R}. \tag{23}$$

The group $\mathrm{Spin}^+(p,q)$ of which the one-sided representations are members is a double covering of the group $\mathrm{SO}^+(p,q)$. This follows from the fact that in the two-sided representations, the rotation represented by a rotor R could equally well be represented by $-R$,

$$\hat{\mathrm{V}}_{(-R)}=\hat{\mathrm{V}}_R,$$

but this does not hold for the one-sided representations.

We extend the carrier spaces of these two types of representations to multilinear functions of vector variables by defining their actions on k-linear functions to be

$$\hat{\mathrm{U}}_R[\psi(a_1,\ldots,a_k)]=R\psi(\tilde{R}a_1R,\ldots,\tilde{R}a_kR), \tag{24}$$

$$\hat{\mathrm{V}}_R[M(a_1,\ldots,a_k)]=RM(\tilde{R}a_1R,\ldots,\tilde{R}a_kR)\tilde{R}. \tag{25}$$

Most rotors can be written in the form

$$R=\exp\left(\frac{1}{2}C\right), \tag{26}$$

where C is a bivector. If C is a simple bivector then it defines the plane in which the rotation occurs. The algebra of the bivectors $\mathcal{G}^2_{p,q}$, which is closed under the commutator product (5), is equivalent to the Lie algebra so(p,q) of the rotation group.[4]

By writing the operator $\hat{\mathrm{U}}_R$ in a form similar to (26) we can define the linear operator $\hat{\mathrm{L}}_C$ which corresponds to C,

$$\hat{\mathrm{U}}_R\equiv\exp\left(\frac{1}{2}\hat{\mathrm{L}}_C\right). \tag{27}$$

Take for example a linear function $\psi(a)$ which transforms under $\hat{\mathrm{U}}_R$. In order to find the action of $\hat{\mathrm{L}}_C$ on $\psi(a)$ we write $C=\alpha B$, where α is a scalar and use (26) to substitute for R,

$$\hat{\mathrm{U}}_R[\psi(a)]=\exp\left(\frac{1}{2}\alpha\hat{\mathrm{L}}_B\right)\psi(a)=e^{(1/2)\,\alpha B}\psi(e^{-(1/2)\,\alpha B}ae^{(1/2)\,\alpha B}). \tag{28}$$

Differentiating with respect to α and setting $\alpha=0$, we find that

$$\tfrac{1}{2}\,\hat{\mathrm{L}}_B\psi(a)=\tfrac{1}{2}B\psi(a)+\psi(a\cdot B). \tag{29}$$

The factors of $\frac{1}{2}$ in this equation arise naturally from the definition of the rotor (26). The set of operators $\hat{\mathrm{L}}_B$ form a representation of the Lie algebra which obeys the commutation relation

$$[\hat{\mathrm{L}}_B,\hat{\mathrm{L}}_C]=\hat{\mathrm{L}}_{[B,C]}, \tag{30}$$

or writing the same expression using the commutator product,

$$\hat{\mathrm{L}}_B\times\hat{\mathrm{L}}_C=\hat{\mathrm{L}}_{B\times C}. \tag{31}$$

The geometric algebra $\mathcal{G}_{p,q}$ has $\frac{1}{2}n(n-1)$ linearly-independent bivectors, thus SO(p,q) has $\frac{1}{2}n(n-1)$ generators.

Extending $\hat{\mathrm{L}}_B$ to act on functions with k arguments is straightforward,

$$\frac{1}{2}\,\hat{\mathrm{L}}_B[\psi(\cdots)]=\frac{1}{2}B\psi(\cdots)+\sum_{i=1}^{k}\psi(\ldots,a_i\cdot B,\ldots); \tag{32}$$

each argument makes a contribution of one term to the expression. A similar procedure gives the action of $\hat{\mathrm{L}}_B$ on functions which transform two-sidedly,

$$\frac{1}{2}\,\hat{\mathrm{L}}_B[M(\cdots)]=B\times M(\cdots)+\sum_{i=1}^{k}M(\ldots,a_i\cdot B,\ldots). \tag{33}$$

B. Invariant subspaces

We must find the irreducible representations of the rotation group. To this end we now seek the invariant subspaces of the representations.

The simplest kind of invariant subspace is one in which elements have a definite symmetry among their arguments. An operator $\hat{\mathrm{T}}_{ij}$ can be defined which exchanges the ith and jth arguments of a function,

$$\hat{\mathrm{T}}_{ij}\psi(\ldots,a_i,\ldots,a_j,\ldots)=\psi(\ldots,a_j,\ldots,a_i,\ldots), \tag{34}$$

which is used to construct operators which symmetrize or antisymmetrize between the ith and jth arguments of a function,

$$\hat{\mathrm{S}}_{ij} \equiv \tfrac{1}{2}(1+\hat{\mathrm{T}}_{ij}),$$
$$\hat{\mathrm{A}}_{ij} \equiv \tfrac{1}{2}(1-\hat{\mathrm{T}}_{ij}). \tag{35}$$

These operators commute with the action of the group, so if a function has symmetric or antisymmetric arguments, the action of the group does not change this property. Combinations of $\hat{\mathrm{S}}_{ij}$ and $\hat{\mathrm{A}}_{ij}$ can be constructed to represent all distinct symmetries between the arguments of a k-linear function. A general product of these operators will be denoted as $\hat{\mathrm{Y}}$. It can be shown that there is a one-to-one correspondence between these functions and the order-k Young tableaux.

A two-sided representation preserves the grade of an object which it transforms. This property can be shown by considering the action of $\hat{\mathrm{V}}_R$ on an r-vector $A_r = a_1 \wedge a_2 \wedge \cdots \wedge a_r$ where $a_1, a_2, \ldots, a_r$ are vectors. Since $R\widetilde{R}=1$,

$$\hat{\mathrm{V}}_R[A_r] = RA_r\widetilde{R} = (Ra_1\widetilde{R}) \wedge (Ra_2\widetilde{R}) \wedge \cdots \wedge (Ra_r\widetilde{R}), \tag{36}$$

which is also be an r-vector, so $\hat{\mathrm{V}}_R$ preserves the grades of all r-vectors, formally,

$$\hat{\mathrm{V}}_R[\mathscr{G}^{\,r}_{p,q}] = \mathscr{G}^{\,r}_{p,q}. \tag{37}$$

A one-sided representation does not preserve the grade of an object, however, since rotors are even-graded elements of the algebra, this representation separates the space of multivectors into even and odd grades, formally,

$$\hat{\mathrm{U}}_R[\mathscr{G}^{+}_{p,q}] = \mathscr{G}^{+}_{p,q},$$
$$\hat{\mathrm{U}}_R[\mathscr{G}^{-}_{p,q}] = \mathscr{G}^{-}_{p,q}. \tag{38}$$

We shall call a one-sided irreducible representation a spinor representation. A spinor representation can also accommodate an ideal structure. A spinor or spinor-valued function ψ can be multiplied on the right by an idempotent element of the algebra $E=E^2$ to project it into a right ideal of the algebra,

$$\psi = \psi E. \tag{39}$$

The ideal structure is preserved by the action of $\hat{\mathrm{U}}_R$ since the rotor acts only on the left.

Another property that is not changed by the action of the group is that of monogenicity. If a function $\mathscr{M}(a_1, \ldots, a_k)$ is monogenic, it has the property

$$\partial_{a_j}\mathscr{M}(a_1, \ldots, a_k) = 0, \tag{40}$$

for all $j=1,2,\ldots,k$. This property is unchanged by the action of a rotor. Take, for example, a function which transforms under $\hat{\mathrm{U}}_R$,

$$\partial_{a_j}\hat{\mathrm{U}}_R[\mathscr{M}(a_1, \ldots, a_k)] = \partial_{a_j} R\mathscr{M}(a'_1, \ldots, a'_k), \tag{41}$$

where $a'_j = \widetilde{R}a_jR$. Using the chain rule to rearrange the transformed function,

$$\partial_{a_j} R\mathscr{M}(a'_1, \ldots, a'_k) = \partial_{a_j}(\widetilde{R}a_jR)\cdot\partial_{a'_j} R\mathscr{M}(a'_1, \ldots, a'_k)$$
$$= R\partial_{a'_j}\mathscr{M}(a'_1, \ldots, a'_k). \tag{42}$$

Thus, under the action of $\hat{U}_R$, the monogenic properties of a function are not changed,

$$\hat{U}_R[\partial_{a_j}\mathscr{M}(a_1,\ldots,a_k)]=R\partial_{a'_j}\mathscr{M}(a'_1,\ldots,a'_k). \tag{43}$$

A similar argument applies to the double-sided transformation (25). From this, it can be shown that monogenicity is also preserved under the action of elements of the Lie algebra. We shall denote any monogenic function by the primary label $\mathscr{M}$.

The grade-invariance property of the two-sided transformation and monogenic invariance do not define mutually exclusive invariant subspaces and so cannot be used simultaneously to find the irreducible representations. We conclude that the irreducible representations of rotation groups consist of basis functions which have distinct symmetry among their arguments. For functions which transform one-sidedly, they correspond to particular terms in the monogenic decomposition and possibly have an ideal structure or, for functions which transform two-sidedly, they possess a definite grade.

C. Monogenic decomposition

In the last section we proved that the monogenic properties of a function are not changed by the action of the rotation group or its Lie algebra. Arbitrary k-linear functions can be decomposed into monogenic and non-monogenic parts in a similar manner to Taylor decomposition of functions in one dimension.

An inner product can be defined between two k-linear functions, ξ and η,

$$\langle\xi|\,\eta\rangle=\langle\xi(\partial_{a_1},\ldots,\partial_{a_k})\,\tilde{\eta}(a_1,\ldots,a_k)\rangle, \tag{44}$$

which is invariant under the action of both types of representation of the rotation group. A complex structure can always be added to this inner product. The functions are orthogonal if $\langle\xi|\,\eta\rangle=0$. For simplicity it will be assumed that at least one of the functions is not null. A function $\psi(a_1,\ldots,a_k)$ can be decomposed into a monogenic part and a part orthogonal to it:

$$\psi(a_1,\ldots,a_k)=\mathscr{M}(a_1,\ldots,a_k)+M(a_1,\ldots,a_k), \tag{45}$$

where

$$M(a_1,\ldots,a_k)=a_1\psi_1(a_2,\ldots,a_k)+a_2\psi_2(a_1,a_3,\ldots,a_k)+\cdots+a_k\psi_k(a_1,\ldots,a_{k-1}). \tag{46}$$

The fact that each term in M is orthogonal to $\mathscr{M}$ follows easily from the definition of the inner product, since for any function $\xi(a_1,\ldots,a_{j-1},a_{j+1},\ldots,a_k)$ which does not depend on a_j,

$$\begin{aligned}&\langle a_j\xi(a_1,\ldots,a_{j-1},a_{j+1},\ldots,a_k)|\mathscr{M}(a_1,\ldots,a_k)\rangle\\&\quad=\langle\xi(a_1,\ldots,a_{j-1},a_{j+1},\ldots,a_k)|\partial_{a_j}\mathscr{M}(a_1,\ldots,a_k)\rangle=0,\end{aligned} \tag{47}$$

by the monogenic properties of $\mathscr{M}$. Recursive use of the partial decomposition produces the complete monogenic decomposition,

$$\psi(a_1,\ldots,a_k)=\sum_{l=0}^{k}M_l(a_1,\ldots,a_k), \tag{48}$$

where

$$\begin{aligned}M_0(a_1,\ldots,a_k)&=\mathscr{M}(a_1,\ldots,a_k),\\M_l(a_1,\ldots,a_k)&=\sum_{\pi}a_{\pi(1)}\ldots a_{\pi(l)}\mathscr{M}_{\pi}(a_{\pi(l+1)},\ldots,a_{\pi(k)}),\end{aligned} \tag{49}$$

where the sum runs over all permutations π of the numbers $1,\ldots,k$. The decomposition is unique and the M_l are orthogonal,

$$\langle M_j|M_l\rangle=0, \quad \text{for } j\neq l. \tag{50}$$

Under the action of $\hat{\mathrm{U}}_R$, the function M_l transforms as

$$\hat{\mathrm{U}}_R[M_l(a_1,\ldots,a_k)]=\sum_{\pi} a_{\pi(1)}\ldots,a_{\pi(l)}\hat{\mathrm{U}}_R[\mathscr{M}_{\pi}(a_{\pi(l+1)},\ldots,a_{\pi(k)})], \tag{51}$$

similarly for $\hat{\mathrm{V}}_R$. This shows that only the monogenic part of the function is transformed, the polynomial of vectors remaining unchanged.

We conclude that any arbitrary k-linear function can be constructed out of monogenic functions as in (48).

The symmetry between the arguments can be used to refine the monogenic decomposition,

$$\psi(a_1,\ldots,a_k)=\sum_{l=0}^{k}\sum_{\hat{\mathrm{Y}}}\hat{\mathrm{Y}}M_l(a_1,\ldots,a_k), \tag{52}$$

where M_l is defined as before (49) and the decomposition is summed over all possible symmetries among the arguments, $\hat{\mathrm{Y}}$. This is known as the Fischer decomposition.[1]

III. REPRESENTATIONS OF SO(3)

In this section we will investigate the multilinear representations of the rotation group of Euclidean space, SO(3).

A. Generators and eigenstates

The Pauli algebra $\mathscr{G}_{3,0}$ has three linearly-independent bivectors, $i\sigma_1$, $i\sigma_2$ and $i\sigma_3$, so SO(3) has three generators, for which the following notation is introduced:

$$\hat{\mathrm{J}}_k\equiv\tfrac{1}{2}\hat{\mathrm{L}}_{i\sigma_k}, \tag{53}$$

for $k=1,2,3$. Since $i\sigma_k^2=-1$ these are compact generators. The $\{\hat{\mathrm{J}}_k\}$ obey the commutation relation

$$[\hat{\mathrm{J}}_k,\hat{\mathrm{J}}_l]=-\epsilon_{klm}\hat{\mathrm{J}}_m. \tag{54}$$

This relation differs from the usual convention because the generators are related to the bivectors of the Pauli algebra, rather than the vectors as in more conventional treatments.

The representations will be classified as eigenstates of the generators. To do this we need to introduce a complex structure, but rather than using the scalar imaginary we shall use an operation within the real geometric algebra to play the same role. The geometric operation that we choose will be denoted by $\hat{\mathrm{J}}$.

The spin of a particle is described by two quantum numbers which give the total spin and the spin in a particular plane. The 'total spin' operator is the Casimir operator of the group,

$$\hat{\mathbf{J}}^2\equiv\hat{\mathrm{J}}_1{}^2+\hat{\mathrm{J}}_2{}^2+\hat{\mathrm{J}}_3{}^2. \tag{55}$$

Denoting this as the square of a vector is somewhat misleading, but we shall retain this notation for the sake of convention. One of the generators must be chosen to complete our commuting set of operators. The conventional choice is $\hat{\mathrm{J}}_3$ which measures the spin in the $i\sigma_3$ plane. The states are described in terms of the eigenstates of these two operators, the eigenvalues of which are s and m, respectively. Acting on an abstract state $|sm\rangle$,

$$\hat{\mathbf{J}}^2|sm\rangle=-s(s+1)|sm\rangle, \tag{56}$$

$$\hat{\mathrm{J}}_3|sm\rangle=\hat{\jmath}m|sm\rangle. \tag{57}$$

The raising and lowering operators are the same as in conventional treatments,

$$\hat{\mathrm{J}}_{\pm}=\hat{\mathrm{J}}_1\pm\hat{\jmath}\hat{\mathrm{J}}_2. \tag{58}$$

The spin-s irreducible representation will be referred to as $D(s)$. Products of the representations obey the familiar Clebsch–Gordon decomposition,

$$D(s_1)\otimes D(s_2)\equiv D(s_1+s_2)\oplus\cdots\oplus D(|s_1-s_2|). \tag{59}$$

B. The spin-$\frac{1}{2}$ representation

The spin-$\frac{1}{2}$ representation is the fundamental spinor representation which could take values in the even or odd sectors of the algebra. It is more convenient to let spinor ψ take values in the even sub-algebra $\mathcal{G}^{+}_{3,0}$, of which the basis elements are

$$\{1,i\sigma_1,i\sigma_2,i\sigma_3\}. \tag{60}$$

Translating the conventional representation into geometric algebra[5,6] it is found that the spin-$\frac{1}{2}$ representation is complexified by right-multiplication by $i\sigma_3$,

$$\hat{\jmath}\psi\equiv\psi i\sigma_3. \tag{61}$$

The even sub-algebra is closed under this operation.

Using the raising and lowering operators (58) the spin-up and spin-down basis states $\chi_{\pm}$ are found to be

$$\begin{aligned}\chi_{+}&=1,\\ \chi_{-}&=i\sigma_1.\end{aligned} \tag{62}$$

Coefficients of these states are defined over the ''complex'' field, $\{1,\hat{\jmath}\}$ so a general state takes values over the whole even sub-algebra. The convention for the basis states used here differs slightly from those of Doran *et al.*[6] since we have adopted a different convention for the generators.

C. The spin-1 representation

The vector space of the Pauli algebra $\mathcal{G}^{1}_{3,0}$ has three basis vectors which could form a basis for the spin-1 representation $D(1)$. We cannot use the same complexification operation as the spin-$\frac{1}{2}$ representation here, but we note that the space of bivectors $\mathcal{G}^{2}_{3,0}$ could also form a suitable basis for $D(1)$. So to complexify this representation we employ an *ad hoc* construction: the pseudoscalar $i=\sigma_1\sigma_2\sigma_3$, which has a negative square, is used to complexify the representation

$$\hat{\jmath}A\equiv iA. \tag{63}$$

A consequence of this is that the $D(1)$ representation is vector- and bivector-valued and has 3 complex or 6 real degrees of freedom. The complexification operations we have defined for the spin-$\frac{1}{2}$ and the spin-1 representations at first sight appear to be incompatible, but later we shall show that they can be reconciled.

Using the raising and lowering operators the three basis states A_{sm} can be constructed,

$$\begin{aligned}A_{1,1}&=\frac{1}{\sqrt{2}}(\sigma_1+i\sigma_2),\\ A_{1,0}&=-i\sigma_3,\end{aligned} \tag{64}$$

$$A_{1,-1}=\frac{1}{\sqrt{2}}(\sigma_1-i\sigma_2),$$

the coefficients of which take values over the ''complex'' field $\{1,\hat{\mathrm{j}}\}$.

D. Relating spin-$\frac{1}{2}$ and spin-1 representations

Here we shall attempt to relate the spin-$\frac{1}{2}$ representations to the spin-1 representations and reconcile the differing complexification operators. This process will be very useful later in constructing higher-spin representations.

We can attempt to construct a spin-1 state from two spin-$\frac{1}{2}$ states, ψ_1 and ψ_2, in the form

$$M=\psi_1\Gamma\tilde{\psi}_2, \tag{65}$$

where Γ is a fixed multivector to be determined. If the spinors are transformed under $\hat{\mathrm{U}}_R$ we find that

$$\psi_1\Gamma\tilde{\psi}_2\overset{R}{\mapsto}\hat{\mathrm{U}}_R[\psi_1]\Gamma\hat{\mathrm{U}}_R[\tilde{\psi}_2]=R\psi_1\Gamma(R\psi_2)\tilde{}=R\psi_1\Gamma\tilde{\psi}_2\tilde{R}=\hat{\mathrm{V}}_R[\psi_1\Gamma\tilde{\psi}_2], \tag{66}$$

so M transforms like an integral-spin representation. Taking the product of the representations, M will transform as

$$D\left(\frac{1}{2}\right)\otimes D\left(\frac{1}{2}\right)=D(1)\oplus D(0). \tag{67}$$

Since Γ is a multivector, M will take values in all possible grades of the algebra. The scalar and pseudoscalar parts can be identified with the spin-0 representation, since these grades are invariant under rotation,

$$\phi=\langle\psi_1\Gamma\tilde{\psi}_2\rangle_{0,3}, \tag{68}$$

and the vector and bivector parts are identified with the spin-1 representation as in the previous section,

$$A=\langle\psi_1\Gamma\tilde{\psi}_2\rangle_{1,2}. \tag{69}$$

Inserting the spin-$\frac{1}{2}$ basis states into (69), we find that Γ is

$$\Gamma\equiv(1+\sigma_3)\sigma_1. \tag{70}$$

The presence of the idempotent $\frac{1}{2}(1+\sigma_3)$ in Γ ensures that the differing complexification operators defined earlier are consistent.

E. Half-integral-spin representations

To find a general half-integral-spin representation, we will be looking for spinor-valued functions which transform under $\hat{\mathrm{U}}_R$. The complexification operator for these functions is the same as that for the spinor representation.

Consider a spinor-valued linear function, $\psi(a)$, which transforms under $\hat{\mathrm{U}}_R$. The monogenic decomposition (48) can be used to analyze the transformation properties of the function,

$$\psi(a)=\mathscr{M}(a)+a\chi, \tag{71}$$

where $\mathscr{M}(a)$ is the monogenic part of the function, and the second term is the residual component which is defined by

$$\chi = \tfrac{1}{3}\partial_a \psi(a), \tag{72}$$

which will take values in the odd-graded part of the algebra $\mathscr{G}_{3,0}$. This function will transform as the

$$D\left(\frac{1}{2}\right)\otimes D(1) = D\left(\frac{3}{2}\right)\oplus D\left(\frac{1}{2}\right) \tag{73}$$

representation. Under the action of $\hat{\mathrm{U}}_R$, $\psi(a)$ transforms to

$$R\psi(\widetilde{R}aR) = R\mathscr{M}(\widetilde{R}aR) + aR\chi. \tag{74}$$

It can be seen that the second term transforms with one rotor like the spin-$\frac{1}{2}$ representation, so the monogenic part of the function can be identified with the spin-$\frac{3}{2}$ representation.

To construct the higher-spin representations we will need some ''complex''-valued functions of a vector variable from which to construct the contributions of the arguments,

$$\begin{aligned}
\phi_1(a) &= \frac{1}{\sqrt{2}}(a\cdot\sigma_1 + a\cdot\sigma_2\hat{\mathrm{j}}),\\
\phi_0(a) &= -a\cdot\sigma_3\hat{\mathrm{j}},\\
\phi_{-1}(a) &= \frac{1}{\sqrt{2}}(a\cdot\sigma_1 - a\cdot\sigma_2\hat{\mathrm{j}}).
\end{aligned} \tag{75}$$

These are closely related to the spin-1 basis states. They have been written using $\hat{\mathrm{j}}$ since they will be useful in constructing both integral and half-integral-spin states.

The spin-$\frac{3}{2}$ multiplet can be constructed from these functions and the spin-$\frac{1}{2}$ basis states $\chi_\pm$ (62),

$$\begin{aligned}
\psi_{3/2,\,3/2}(a) &= \chi_+\phi_1(a),\\
\psi_{3/2,\,1/2}(a) &= \frac{1}{\sqrt{3}}(\chi_-\phi_1(a) + \sqrt{2}\chi_+\phi_0(a)),\\
\psi_{3/2,\,-1/2}(a) &= \frac{1}{\sqrt{3}}(\chi_+\phi_{-1}(a) + \sqrt{2}\chi_-\phi_0(a)),\\
\psi_{3/2,\,-3/2}(a) &= \chi_-\phi_{-1}(a).
\end{aligned} \tag{76}$$

These four basis states are monogenic functions. The associated spin-$\frac{1}{2}$ doublet consists of the residual functions from the monogenic decomposition,

$$\begin{aligned}
\psi_{1/2,\,1/2}(a) &= \frac{1}{\sqrt{3}}(\sqrt{2}\chi_-\phi_1(a) - \chi_+\phi_0(a)) = \frac{1}{\sqrt{3}}ai,\\
\psi_{1/2,\,-1/2}(a) &= \frac{1}{\sqrt{3}}(-\sqrt{2}\chi_+\phi_{-1}(a) + \chi_-\phi_0(a)) = -\frac{1}{\sqrt{3}}a\sigma_1.
\end{aligned} \tag{77}$$

These functions are orthonormal under the functional inner product (44).

Now consider a spinor-valued bilinear function, $\psi(a,b)$, which will transform under $\hat{\mathrm{U}}_R$ as

$$D\left(\frac{1}{2}\right)\otimes D(1)\otimes D(1)=D\left(\frac{5}{2}\right)\oplus 2D\left(\frac{3}{2}\right)\oplus 2D\left(\frac{1}{2}\right). \tag{78}$$

The monogenic decomposition of $\psi(a,b)$ is

$$\psi(a,b)=\mathscr{M}(a,b)+a.\mathscr{M}'(b)+b.\mathscr{M}''(a)+ab\xi_1+ba\xi_2. \tag{79}$$

The symmetry between the arguments need not be taken into account since there cannot be any antisymmetric monogenic functions in $\mathscr{G}_{3,0}$. If there were such a monogenic $\mathscr{M}_A(a,b)$, it could be written as a function of a bivector $\mathscr{M}_A(a\wedge b)$. By linearity,

$$\partial_a.\mathscr{M}_A(a\wedge b)=\partial_a(a\wedge b)\cdot\partial_B.\mathscr{M}_A(B)=0, \tag{80}$$

where the chain rule has been applied. Expanding B over the bivector basis elements, this becomes

$$b\cdot i\sigma_1.\mathscr{M}_A(i\sigma_1)+b\cdot i\sigma_2.\mathscr{M}_A(i\sigma_2)+b\cdot i\sigma_3.\mathscr{M}_A(i\sigma_3)=0, \tag{81}$$

which must hold for all values of b. It is easy to show that this implies that each of the components of $\mathscr{M}_A(a\wedge b)$ vanishes. It follows that in 3 dimensions a k-linear monogenic cannot be antisymmetric in any two of its arguments; it therefore must be totally symmetric.

Under the action of $\hat{\mathrm{U}}_R$, $\psi(a,b)$ transforms to

$$\begin{aligned}R\psi(\widetilde{R}aR,\widetilde{R}bR)=R.\mathscr{M}(\widetilde{R}aR,\widetilde{R}bR)+aR.\mathscr{M}'(\widetilde{R}bR)\\ +bR.\mathscr{M}''(\widetilde{R}aR)+abR\xi_1+baR\xi_2.\end{aligned} \tag{82}$$

Again it can be seen that it is the monogenic part which behaves as the irreducible representation with the highest spin, in this case spin-$\frac{5}{2}$. There are also two spin-$\frac{3}{2}$ and two spin-$\frac{1}{2}$ representations, as predicted in (78). The $m=\frac{5}{2}$ state will be

$$\psi_{5/2\,,\,5/2}(a,b)=\chi_+\phi_1(a)\phi_1(b), \tag{83}$$

and the remaining basis states of the $D(\frac{5}{2})$ representation can be constructed using the lowering operator.

The properties of a function which corresponds to a general half-integral-spin irreducible representation can now be deduced. Consider a spinor-valued k-linear function,

$$\psi(a_1,\ldots,a_k), \tag{84}$$

which transforms under $\hat{\mathrm{U}}_R$ as

$$D\left(\frac{1}{2}\right)\otimes\underbrace{D(1)\otimes\ldots\otimes D(1)}_{k\text{ times}}=D\left(k+\frac{1}{2}\right)\oplus D\left(k-\frac{1}{2}\right)\oplus\ldots\oplus D\left(\frac{1}{2}\right). \tag{85}$$

The monogenic part of this function,

$$\mathscr{M}(a_1,\ldots,a_k), \tag{86}$$

which must be totally symmetric, will correspond to the $D(k+\frac{1}{2})$ irreducible representation. The $m=k+\frac{1}{2}$ basis state will be

$$\psi_{k+(1/2)\,,k+(1/2)}(a_1,\ldots,a_k)=\chi_+\phi_1(a_1)\cdots\phi_1(a_k). \tag{87}$$

The rest of the spin-$(k+\frac{1}{2})$ basis states can be constructed from this using the lowering operator.

F. Integral-spin representations

In the same way as we constructed the spin-1 representation from the spin-$\frac{1}{2}$ representation, we can construct any integral-spin representation from the half-integral-spin representations deduced above. A linear function which transforms under $\hat{\mathrm{V}}_R$ can be constructed from a spinor-valued linear function $\psi_1(a)$ and a spinor ψ_2,

$$M(a)=\psi_1(a)\Gamma\tilde{\psi}_2, \tag{88}$$

where $\Gamma=(1+\sigma_3)\sigma_1$ as before (70). $M(a)$ will be multivector-valued, and it will belong to the

$$D\left(\frac{1}{2}\right)\otimes D(1)\otimes D\left(\frac{1}{2}\right) \tag{89}$$

representation. If $\psi_1(a)$ is restricted to be monogenic $\mathscr{M}(a)$, then the function

$$M'(a)=\mathscr{M}(a)\Gamma\tilde{\psi}_2 \tag{90}$$

belongs to the

$$D\left(\frac{3}{2}\right)\otimes D\left(\frac{1}{2}\right)=D(2)\oplus D(1) \tag{91}$$

representation. This can be restricted further by selecting the vector and bivector part,

$$A(a)=\langle\mathscr{M}(a)\Gamma\tilde{\psi}_2\rangle_{1,2}, \tag{92}$$

which will transform as the $D(2)$ irreducible representation. The scalar and pseudoscalar part of $M'(a)$ will transform as $D(1)$. Inserting the $D(\frac{3}{2})$ and $D(\frac{1}{2})$ basis states from (62) and (76) the basis states of the $D(2)$ irreducible representation are

$$A_{2,2}(a)=A_1\phi_1(a),$$

$$A_{2,1}(a)=\frac{1}{\sqrt{2}}(A_1\phi_0(a)+A_0\phi_1(a)),$$

$$A_{2,0}(a)=\frac{1}{\sqrt{6}}(A_1\phi_{-1}(a)+A_{-1}\phi_1(a)+2A_0\phi_0(a)), \tag{93}$$

$$A_{2,-1}(a)=\frac{1}{\sqrt{2}}(A_{-1}\phi_0(a)+A_0\phi_{-1}(a)),$$

$$A_{2,-2}(a)=A_{-1}\phi_{-1}(a).$$

where A_1, A_0 and A_{-1} are shorthand for the $m=1$, $m=0$ and $m=-1$ basis states of the $D(1)$ representation, respectively.

This procedure can now be generalized to find the properties of a function which corresponds to any integral-spin irreducible representation. Consider a k-linear function which transforms under $\hat{\mathrm{V}}_R$, constructed from a k-linear spinor-valued function $\psi_1(a_1,\ldots,a_k)$ and a spinor ψ_2,

$$M(a_1,\ldots,a_k)=\psi_1(a_1,\ldots,a_k)\Gamma\tilde{\psi}_2. \tag{94}$$

It will transform as

$$D\left(\frac{1}{2}\right)\otimes\underbrace{D(1)\otimes\ldots\otimes D(1)}_{k\text{ times}}\otimes D\left(\frac{1}{2}\right)=D(k+1)\oplus D(k)\oplus\ldots\oplus D(0) \tag{95}$$

If $\psi_1(a_1,\ldots,a_k)$ is restricted to be a monogenic then the function

$$M'(a)=\mathscr{M}(a_1,\ldots,a_k)\Gamma\tilde{\psi}_2 \tag{96}$$

will transform as the

$$D\left(k+\frac{1}{2}\right)\otimes D\left(\frac{1}{2}\right)=D(k+1)\oplus D(k) \tag{97}$$

representation. Again if this is restricted to the vector and bivector parts, the resulting function,

$$A(a_1,\ldots,a_k)=\langle\mathscr{M}(a_1,\ldots,a_k)\Gamma\tilde{\psi}_2\rangle_{1,2}, \tag{98}$$

corresponds to the $D(k+1)$ irreducible representation.

We have now identified the functions which correspond to all of the irreducible representations of $SO(3)$ and in the process discovered many monogenic functions in the Pauli algebra.

IV. REPRESENTATIONS OF THE LORENTZ GROUP

We now turn our attention the the proper Lorentz group $\mathrm{SO}^+(1,3)$. The representations we find in this section will be useful in describing relativistic wave equations.

A. Generators

The space–time algebra has six linearly-independent bivectors, $\{\sigma_k, i\sigma_k\}$, therefore the Lorentz group has six generators, for which we introduce the following shorthand notation:

$$\hat{\mathrm{J}}_k=\frac{1}{2}\hat{\mathrm{L}}_{i\sigma_k},\quad \hat{\mathrm{K}}_k=\frac{1}{2}\hat{\mathrm{L}}_{\sigma_k}, \tag{99}$$

where $k=1,2,3$. These operators obey the commutation relations

$$[\hat{\mathrm{J}}_i,\hat{\mathrm{J}}_j]=-\epsilon_{ijk}\hat{\mathrm{J}}_k,\quad [\hat{\mathrm{K}}_i,\hat{\mathrm{K}}_j]=\epsilon_{ijk}\hat{\mathrm{J}}_k,\quad [\hat{\mathrm{J}}_i,\hat{\mathrm{K}}_j]=\epsilon_{ijk}\hat{\mathrm{K}}_k. \tag{100}$$

The $\{\hat{\mathrm{J}}_k\}$ form a closed sub-algebra, which is identical to the Lie algebra of SO(3) in the previous section. Again these relations differ slightly from conventional treatments.

B. Classification of irreducible representations

We shall introduce a complex structure to the Lorentz group within the real STA, using the operator $\hat{\mathrm{j}}$ to stand for the complexification operation and deducing its action on the representations. The complex structure can be used to define two sets of generators from the generators of the Lorentz group:[7]

$$\hat{\mathrm{A}}_k=\tfrac{1}{2}(\hat{\mathrm{J}}_k+\hat{\mathrm{j}}\hat{\mathrm{K}}_k),\quad \hat{\mathrm{B}}_k=\tfrac{1}{2}(\hat{\mathrm{J}}_k-\hat{\mathrm{j}}\hat{\mathrm{K}}_k), \tag{101}$$

where $k=1,2,3$. These operators have the following commutation relations:

$$[\hat{\mathrm{A}}_k,\hat{\mathrm{A}}_l]=-\epsilon_{klm}\hat{\mathrm{A}}_m,\quad [\hat{\mathrm{B}}_k,\hat{\mathrm{B}}_l]=-\epsilon_{klm}\hat{\mathrm{B}}_m,\quad [\hat{\mathrm{A}}_k,\hat{\mathrm{B}}_l]=0. \tag{102}$$

The operators form two uncoupled sets of generators which individually behave like the generators of SO(3). The group generated by these uncoupled operators will therefore behave as a direct product of two copies of SO(3). If s_a is the spin of the $\hat{\mathrm{A}}_k$ irreducible representation and s_b is the

spin of the $\hat{\mathrm{B}}_k$ irreducible representation then their direct product space, which can be described by a pair of spin weights (s_a, s_b), will have $(2s_a+1)(2s_b+1)$ complex degrees of freedom. An eigenstate will be equivalent to the product of two SO(3) eigenstates,

$$|s_a\, s_b\, m_a\, m_b\rangle \cong |s_a\, m_a\rangle \otimes |s_b\, m_b\rangle. \tag{103}$$

The operators,

$$\hat{\mathbf{A}}^2 = \hat{\mathrm{A}}_1{}^2 + \hat{\mathrm{A}}_2{}^2 + \hat{\mathrm{A}}_3{}^2, \tag{104}$$

$$\hat{\mathbf{B}}^2 = \hat{\mathrm{B}}_1{}^2 + \hat{\mathrm{B}}_2{}^2 + \hat{\mathrm{B}}_3{}^2, \tag{105}$$

are the Casimir operators of the group, where the ''vector'' notation has again been used. These can be written in terms of the $\hat{\mathrm{J}}_k$ and $\hat{\mathrm{K}}_k$ operators by taking the sum and difference,

$$\hat{\mathbf{A}}^2 + \hat{\mathbf{B}}^2 = \tfrac{1}{2}(\hat{\mathbf{J}}^2 - \hat{\mathbf{K}}^2), \tag{106}$$

$$\hat{\mathbf{A}}^2 - \hat{\mathbf{B}}^2 = -\hat{\mathrm{j}}\hat{\mathbf{J}} \cdot \hat{\mathbf{K}}. \tag{107}$$

It will be found that a particular irreducible representation can be distinguished from other representations with the same $s = s_a + s_b$ by its grade or algebraic properties. We shall refer to all such representations as spin-s representations, for example, $(\frac{1}{2}, \frac{1}{2})$ and $(1,0)\oplus(0,1)$ are both spin-1 representations.

Because the representations have a direct product structure, the Lorentz group analogue of the Clebsh–Gordon decomposition is quite simple,

$$(s_{a_1}, s_{b_1}) \otimes (s_{a_2}, s_{b_2}) = \sum_{s_a, s_b} \oplus (s_a, s_b), \tag{108}$$

where the sum runs independently over the values

$$\begin{aligned} s_a &= s_{a_1} + s_{a_2}, s_{a_1} + s_{a_2} - 1, \ldots, |s_{a_1} - s_{a_2}|, \\ s_b &= s_{b_1} + s_{b_2}, s_{b_1} + s_{b_2} - 1, \ldots, |s_{b_1} - s_{b_2}|. \end{aligned} \tag{109}$$

Now that we have identified the irreducible representations, we can classify their basis states. The basis states will be expressed as eigenstates of the operators $\hat{\mathrm{J}}_3$ and $\hat{\mathrm{K}}_3$ with eigenvalues m and n, respectively, because these are related to simple bivector generators. These states will still be eigenstates of $\hat{\mathrm{A}}_3$ and $\hat{\mathrm{B}}_3$; the eigenvalues are related by

$$m = m_a + m_b, \tag{110}$$

$$n = m_a - m_b. \tag{111}$$

The basis states are thus characterized by four spin quantum numbers, s_a, s_b, m and n. Operating on an eigenstate $|s_a s_b m n\rangle$ we find

$$\hat{\mathbf{A}}^2 |s_a s_b m n\rangle = -s_a(s_a+1)|s_a s_b m n\rangle, \tag{112}$$

$$\hat{\mathbf{B}}^2 |s_a s_b m n\rangle = -s_b(s_b+1)|s_a s_b m n\rangle, \tag{113}$$

$$\hat{\mathrm{J}}_3 |s_a s_b m n\rangle = \hat{\mathrm{j}} m |s_a s_b m n\rangle, \tag{114}$$

$$\hat{\mathrm{K}}_3 |s_a s_b m n\rangle = n |s_a s_b m n\rangle. \tag{115}$$

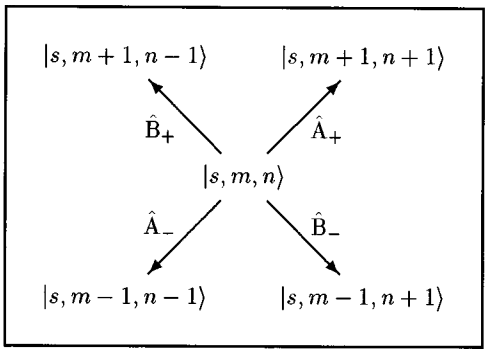

FIG. 1. Raising and lowering operators for spin-s irreducible representation of the Lorentz group.

The raising and lowering operators

$$\hat{\mathrm{A}}_\pm = \hat{\mathrm{A}}_1 \pm \hat{\mathrm{j}}\hat{\mathrm{A}}_2, \qquad \hat{\mathrm{B}}_\pm = \hat{\mathrm{B}}_1 \pm \hat{\mathrm{j}}\hat{\mathrm{B}}_2, \tag{116}$$

which raise and lower the m_a and m_b quantum numbers, respectively, can be used to move between the basis states of the representation, which is shown in terms of the m and n states in Figure 1.

An alternative, incomplete method of classifying the irreducible representations of the Lorentz group is to decompose them into representations of the subgroup SO(3), as eigenstates of $\hat{\mathbf{J}}^2$. The (s_a, s_b) representation contains the

$$D(s_a)\otimes D(s_b) = D(s_a+s_b)\oplus\cdots\oplus D(|s_a-s_b|) \tag{117}$$

representations of SO(3) and the $(s_a, s_b)\oplus(s_b, s_a)$ representation will contain two copies of each of these subgroup representations. The operator $\hat{\mathrm{J}}_3$ will be used to find the eigenstates of these subgroup representations. The alternative quantum numbers which correspond to the eigenstates of $\hat{\mathbf{J}}^2$ and $\hat{\mathrm{J}}_3$ are j and m,

$$(\hat{\mathbf{J}}^2 - \hat{\mathbf{K}}^2)|s_a s_b j m\rangle = -2(s_a(s_a+1)+s_b(s_b+1))|s_a s_b j m\rangle, \tag{118}$$

$$\hat{\mathbf{J}}^2|s_a s_b j m\rangle = -j(j+1)|s_a s_b j m\rangle, \tag{119}$$

$$\hat{\mathrm{J}}_3|s_a s_b j m\rangle = \hat{\mathrm{j}} m|s_a s_b j m\rangle. \tag{120}$$

These are not a complete set of operators because $\hat{\mathbf{J}}^2$ does not commute with $\hat{\mathbf{J}}\cdot\hat{\mathbf{K}}$. Also $\hat{\mathrm{K}}_3$ commutes with $\hat{\mathrm{J}}_3$ but not with $\hat{\mathbf{J}}^2$, so it does not behave as a lowering or raising operator, but mixes the $|s_a s_b j m\rangle$ state with all the other possible j eigenstates leaving the eigenvalue s unchanged (Figure 2).

C. The spin-$\frac{1}{2}$ representation

The Lorentz group has one spin-$\frac{1}{2}$ irreducible representation, $(\frac{1}{2},0)\oplus(0,\frac{1}{2})$. Translating the Dirac spinor into the STA we find it most convenient if it takes values in the even sub-algebra.[5,6] The complexification operation is identical to that of the non-relativistic spin-$\frac{1}{2}$ representation, multiplication on the right by $i\sigma_3$,

$$\hat{\mathrm{j}}\psi \equiv \psi i\sigma_3. \tag{121}$$

The basis states ψ_{mn} of this representation are

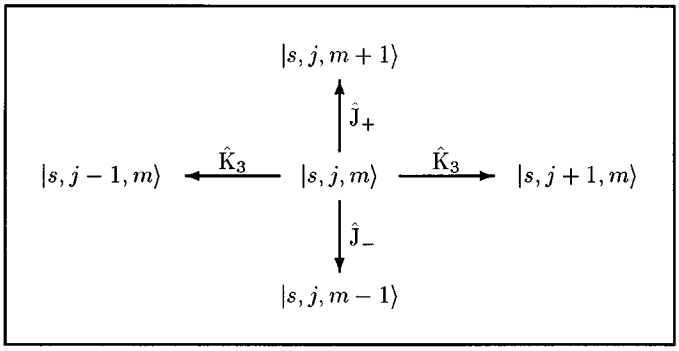

FIG. 2. Alternative raising and lowering operators for spin-s irreducible representation of the Lorentz group.

$$\psi_{1/2,\,1/2}=\frac{1}{2}(1+\sigma_3)\quad \psi_{1/2,\,-1/2}=\frac{1}{2}(1-\sigma_3),$$
$$\psi_{-1/2,\,1/2}=i\sigma_1\frac{1}{2}(1-\sigma_3)\quad \psi_{-1/2,\,-1/2}=i\sigma_1\frac{1}{2}(1+\sigma_3). \tag{122}$$

These basis states have an ideal structure, therefore they correspond most closely to the Weyl representation.

Decomposing the $(\frac{1}{2},0)\oplus(0,\frac{1}{2})$ irreducible representation in terms of the SO(3) subgroup,

$$\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\cong D\left(\frac{1}{2}\right)\oplus D\left(\frac{1}{2}\right), \tag{123}$$

there are two $D(\frac{1}{2})$ doublets which are transformed into each other by the action of $\hat{\mathrm{K}}_3$,

$$\begin{aligned}\chi_+&=1, &&\xi_+=\sigma_3,\\ &&&\overset{\hat{\mathrm{K}}_3}{\leftrightarrow}\\ \chi_-&=i\sigma_1, &&\xi_-=\sigma_2.\end{aligned} \tag{124}$$

The $\chi_\pm$ doublet is identical to the $D(\frac{1}{2})$ representation of $SO(3)$; the $\xi_\pm$ doublet is new, coming from the action of $\hat{\mathrm{K}}_3$ on the $\chi_\pm$ doublet.

D. The spin-1 representations

The Lorentz group has two spin-1 representations, $(\frac{1}{2},\frac{1}{2})$ and $(1,0)\oplus(0,1)$. Like the integral-spin representations of SO(3), we can construct the spin-1 representations from the spinor representation. We use the same form,

$$M=\psi_1\Gamma\tilde{\psi}_2, \tag{125}$$

where Γ is an arbitrary multivector to be chosen later. Since the STA and the Pauli Algebra share the same pseudoscalar we can use the same construction as before to complexify these representations,

$$\hat{\mathrm{j}}M\equiv iM. \tag{126}$$

In general, M will transform under $\hat{\mathrm{V}}_R$ as the

$$\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]=2(0,0)\oplus 2\left(\frac{1}{2},\frac{1}{2}\right)\oplus[(1,0)\oplus(0,1)] \tag{127}$$

representation.

The $(\frac{1}{2},\frac{1}{2})$ irreducible representation is obtained by choosing

$$\Gamma = \gamma_1 + i\gamma_2, \tag{128}$$

which restricts M to be vector- and trivector-valued. Inserting the $(\frac{1}{2},0)\oplus(0,\frac{1}{2})$ basis states from (122) it is found that the basis states A_{mn} of the $(\frac{1}{2},\frac{1}{2})$ representation are

$$\begin{aligned} A_{0,1} &= \gamma_0 + \gamma_3, \quad A_{0,-1} = \gamma_0 - \gamma_3, \\ A_{1,0} &= \gamma_1 + i\gamma_2, \quad A_{1,-1,0} = \gamma_1 - i\gamma_2, \end{aligned} \tag{129}$$

which is identical to the null tetrad $(l,n,m,\overline{m})$ of Penrose and Rindler.[8,9]

If Γ is chosen to be

$$\Gamma = \sigma_1 + i\sigma_2, \tag{130}$$

then M will be even-valued. Further restricting (125) to the symmetric part,

$$F = \tfrac{1}{2}(\psi_1 \Gamma \widetilde{\psi}_2 + \psi_2 \Gamma \widetilde{\psi}_1), \tag{131}$$

then F will be bivector-valued, which corresponds to the $(1,0)\oplus(0,1)$ irreducible representation.

We have chosen the operation (126) to perform the complexification. Multiplication by the pseudoscalar, i, also generates the duality operation, so by this process we only obtain the self-dual set of basis states for F. Inserting the spin-$\frac{1}{2}$ basis states from (122) the basis states F_{mn} are found to be

$$\begin{aligned} F_{1,1} &= \sigma_1 + i\sigma_2, \\ F_{0,0} &= i\sigma_3, \\ F_{-1,-1} &= \sigma_1 - i\sigma_2. \end{aligned} \tag{132}$$

Instead, if we had chosen $\hat{\jmath}M = -iM$ to be the complexification operation, which is consistent with choosing $\Gamma = \sigma_1 - i\sigma_2$, we would obtain the anti-self-dual set of basis states.

The corresponding antisymmetric part of (125),

$$\phi = \tfrac{1}{2}(\psi_1 \Gamma \widetilde{\psi}_2 - \psi_2 \Gamma \widetilde{\psi}_1), \tag{133}$$

is scalar- and pseudoscalar-valued, which is invariant under the action of $\hat{\mathrm{V}}_R$. This corresponds to the scalar (0,0) irreducible representation.

Above, identifying the grades of a multivector with the representations, we have only obtained half of the irreducible representations that should be contained in the decomposition of the complete multivector representation (127). This is a consequence of the minimalist route we have chosen, introducing complexification as a geometric operation.

E. Half-integral-spin representations

We now search for half-integral-spin irreducible representations, which we expect to be multilinear spinor-valued functions. Consider a linear spinor-valued function $\psi(a)$ which transforms under $\hat{\mathrm{U}}_R$. Its monogenic decomposition is

$$\psi(a) = \mathscr{M}(a) + a\chi, \tag{134}$$

where the first term is monogenic and the second is the residual term in which χ is defined by $\chi = \frac{1}{4}\partial_a \psi(a)$. The function $\psi(a)$ belongs to the

$$\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\left(\frac{1}{2},\frac{1}{2}\right)=\left[\left(1,\frac{1}{2}\right)\oplus\left(\frac{1}{2},1\right)\right]\oplus\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right] \tag{135}$$

representation. The second term of the monogenic decomposition (134) transforms like the $(\frac{1}{2},0)\oplus(0,\frac{1}{2})$ irreducible representation so the monogenic component is identified with the $(1,\frac{1}{2})\oplus(\frac{1}{2},1)$ irreducible representation, a ''spin-$\frac{3}{2}$'' representation:

$$\left(1,\frac{1}{2}\right)\oplus\left(\frac{1}{2},1\right)\leftrightarrow\mathscr{M}(a),$$

$$\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\leftrightarrow\chi.$$

The monogenic functions which form the basis states for the $(1,\frac{1}{2})\oplus(\frac{1}{2},1)$ representation can be constructed from the spin-$\frac{1}{2}$ basis states and ''complex''-valued linear functions derived from the $(\frac{1}{2},\frac{1}{2})$ representation. For example,

$$\psi_{1/2\,,\,3/2}(a)=\psi_{1/2\,,\,1/2}\phi_{0,1}(a) \tag{136}$$

is monogenic, where $\phi_{0,1}(a)=a\cdot(\gamma_0+\gamma_3)$.

Now consider a bilinear spinor-valued function $\psi(a,b)$ which transforms under $\hat{\mathrm{U}}_R$. Its monogenic decomposition is

$$\psi(a,b)=\mathscr{M}(a,b)+a.\mathscr{M}'(b)+b.\mathscr{M}''(a)+ab\chi_1+ba\chi_2, \tag{137}$$

which can be further refined by taking into account the symmetry between the arguments:

$$\begin{aligned}\psi(a,b)=&\mathscr{M}_S(a,b)+\mathscr{M}_A(a,b)+a.\mathscr{M}_S'(b)+b.\mathscr{M}_S'(a)\\&+a.\mathscr{M}_A'(b)-b.\mathscr{M}_A'(a)+a\cdot b\chi_S+a\wedge b\chi_A,\end{aligned} \tag{138}$$

where the subscripts S and A denote the symmetric and antisymmetric parts, respectively,

$$\begin{aligned}\mathscr{M}_S(a,b)&=\mathscr{M}_S(b,a),\\ \mathscr{M}_A(a,b)&=-\mathscr{M}_A(b,a).\end{aligned} \tag{139}$$

The function $\psi(a,b)$ belongs to the

$$\begin{aligned}&\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\left(\frac{1}{2},\frac{1}{2}\right)\otimes\left(\frac{1}{2},\frac{1}{2}\right)\\&\quad=\left[\left(\frac{3}{2},1\right)\oplus\left(1,\frac{3}{2}\right)\right]\oplus\left[\left(\frac{3}{2},0\right)\oplus\left(0,\frac{3}{2}\right)\right]\oplus2\left[\left(1,\frac{1}{2}\right)\oplus\left(\frac{1}{2},1\right)\right]\\&\qquad\oplus2\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\end{aligned} \tag{140}$$

representation. The are two new terms in this decomposition compared to that of a linear spinor-valued function, $(\frac{3}{2},1)\oplus(1,\frac{3}{2})$ and $(\frac{3}{2},0)\oplus(0,\frac{3}{2})$. Counting the degrees of freedom, we make the following identifications:

$$\left(\frac{3}{2},1\right)\oplus\left(1,\frac{3}{2}\right)\leftrightarrow\mathscr{M}_S(a,b),$$

$$\left(\frac{3}{2},0\right)\oplus\left(0,\frac{3}{2}\right)\leftrightarrow\mathscr{M}_A(a,b),$$

$$\left(1,\frac{1}{2}\right)\oplus\left(\frac{1}{2},1\right)\leftrightarrow\mathscr{M}'_S(a),\mathscr{M}'_A(a),$$

$$\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\leftrightarrow\chi_S,\chi_A.$$

We can now proceed to find a general half-integral-spin representation. Consider a spinor-valued k-linear function,

$$\psi(a_1,\ldots,a_k), \tag{141}$$

which will transform under the action of the group as

$$\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\underbrace{\left(\frac{1}{2},\frac{1}{2}\right)\otimes\ldots\otimes\left(\frac{1}{2},\frac{1}{2}\right)}_{k\text{ times}}. \tag{142}$$

The highest-spin representation in the decomposition will be a symmetric monogenic function which will correspond to the $((k+1)/2,k/2)\oplus(k/2,(k+1)/2)$ irreducible representation. This is confirmed by counting the degrees of freedom in this function. A symmetric spinor-valued function in 4 dimensions has $8[(k+3)!/3!k!]$ real degrees of freedom. A monogenic version of the same therefore has

$$8\left(\frac{(k+3)!}{3!k!}-\frac{(k+2)!}{3!(k-1)!}\right)=4(k+2)(k+1) \tag{143}$$

real degrees of freedom which is exactly the number contained in the $((k+1)/2,k/2)\oplus(k/2,(k+1)/2)$ irreducible representation.

F. Integral-spin representations

Again the integral-spin representations can be constructed from half-integral-spin representations. Consider a multivector-valued linear function constructed in the form

$$T(a)=\psi_1(a)\Gamma\widetilde{\psi}_2. \tag{144}$$

In general, $T(a)$ will belong to the

$$\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\left(\frac{1}{2},\frac{1}{2}\right)\otimes\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right] \tag{145}$$

representation. If $\psi_1(a)$ is restricted to be monogenic, then the function will belong to the

$$\left[\left(1,\frac{1}{2}\right)\oplus\left(\frac{1}{2},1\right)\right]\otimes\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right] \tag{146}$$

representation which contains the $(1,1)$ and $(\frac{3}{2},\frac{1}{2})\oplus(\frac{1}{2},\frac{3}{2})$ irreducible representations as the highest-spin irreducible representations. Using $\Gamma=\gamma_1+i\gamma_2$ projects out the vector and trivector part which corresponds to the $(1,1)$ representation,

$$A(a)=\psi_1(a)(\gamma_1+i\gamma_2)\widetilde{\psi}_2, \tag{147}$$

and choosing $\Gamma=\sigma_1+i\sigma_2$ and projecting out the bivector part corresponds to the $(\frac{3}{2},\frac{1}{2})\oplus(\frac{1}{2},\frac{3}{2})$ representation,

$$F(a)=\langle\psi_1(a)(\sigma_1+i\sigma_2)\widetilde{\psi}_2\rangle_2. \tag{148}$$

We can now generalize to an arbitrary integral-spin representation,

$$T(a_1,\ldots,a_k)=\psi_1(a_1,\ldots,a_k)\Gamma\widetilde{\psi}_2, \tag{149}$$

which belongs to the

$$\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right]\otimes\underbrace{\left(\frac{1}{2},\frac{1}{2}\right)\otimes\ldots\otimes\left(\frac{1}{2},\frac{1}{2}\right)}_{k\text{ times}}\otimes\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right] \tag{150}$$

representation. If we select the monogenic, totally-symmetric part of the function, we can restrict it to the

$$\left[\left(\frac{k+1}{2},\frac{k}{2}\right)\oplus\left(\frac{k}{2},\frac{k+1}{2}\right)\right]\otimes\left[\left(\frac{1}{2},0\right)\oplus\left(0,\frac{1}{2}\right)\right] \tag{151}$$

representation which contains the $((k+1)/2,(k+1)/2)$ and $((k+2)/2,k/2)\oplus(k/2,(k+2)/2)$ irreducible representations as the highest-spin irreducible representations, which again correspond to the vector–trivector and bivector parts, respectively. These are both ''spin-$(k+1)$'' representations.

We have now found the method for identifying the irreducible representations of the Lorentz group with multilinear functions.

V. WAVE EQUATIONS

So far we have restricted our attention to describing higher-weighted representations of rotation groups with multilinear functions. The wavefunction describing a quantum-mechanical particle depends in an arbitrary non-linear fashion on the position vector x. The method described in this paper can be generalized to include non-linear arguments, and thus handle these ''external'' spatial or space–time degrees of freedom within the same mathematical structure as the ''internal'' spin degrees of freedom.

Consider a function which transforms under $\hat{\mathrm{U}}_R$ which has linear arguments $a_1,\ldots,a_k$ and non-linear arguments $x_1,\ldots,x_l$; we shall separate the linear and non-linear arguments with a semi-colon,

$$\psi(a_1,\ldots,a_k;x_1,\ldots,x_l). \tag{152}$$

The non-linear arguments transform under rotation in the same manner as the linear ones,

$$\hat{\mathrm{U}}_R[\psi(a_1,\ldots,a_k;x_1,\ldots,x_l)]=R\psi(\widetilde{R}a_1R,\ldots,\widetilde{R}a_kR;\widetilde{R}x_1R,\ldots,\widetilde{R}x_lR), \tag{153}$$

but the action of the Lie group is modified,

$$\begin{aligned}\frac{1}{2}\hat{\mathrm{L}}_B[\psi(\ldots;\ldots)]&=\frac{1}{2}B\psi(\ldots;\ldots)+\sum_{i=1}^{k}\psi(\ldots,a_i\cdot B,\ldots;\ldots)\\&\quad-\sum_{i=1}^{l}B\cdot(x_i\wedge\partial_{x_i})\psi(\ldots;\ldots).\end{aligned} \tag{154}$$

Each non-linear argument x_i makes a contribution,

$$-B\cdot(x_i\wedge\partial_{x_i})=(x_i\cdot B)\cdot\partial_{x_i}, \tag{155}$$

to $\frac{1}{2}\hat{\mathrm{L}}_B$. If x_i were a linear argument then this would reduce to the same form as for a linear argument, so this generalization is consistent with our earlier definitions. This operator (155) could also be interpreted as finding the component of angular momentum, $-\hat{\mathrm{j}}(x\wedge\nabla)$, in the plane defined by B.

A similar generalization can be made for functions which transform under $\hat{\mathrm{V}}_R$, only the external transformation properties of the function differ from those transforming under $\hat{\mathrm{U}}_R$.

The Dirac equation written in the STA is

$$\nabla\psi(x)i\sigma_3=m\psi(x)\gamma_0. \tag{156}$$

This has been investigated extensively elsewhere.[10,6] The wavefunction $\psi(x)$ is a spinor function of position x. Some of the functions which form a basis for the Dirac wavefunction can be obtained from multilinear functions by polynomial projection.[1] The l-th order contribution can be made by setting all the arguments of a l-linear function equal to x,

$$\psi_l(x)=\prod_{i=1}^{l}(x\cdot\partial_{a_i})\psi(a_1,\ldots,a_l). \tag{157}$$

Only the symmetric part of the function will survive this procedure; the antisymmetric parts vanish. The complete wavefunction can be expressed as a sum over all orders of these functions.

Using the representations of the Lorentz group we found earlier, we can investigate higher-spin wave equations. The Rarita–Schwinger equations[11] describe half-integral-spin particles. Translating these into geometric algebra, the equation describing a spin-$\frac{3}{2}$ particle becomes

$$\nabla\psi(a;x)i\sigma_3=m\psi(a;x)\gamma_0, \tag{158}$$

with the additional condition

$$\partial_a\psi(a;x)=0, \tag{159}$$

which ensures that the spin part of $\psi(a;x)$ belongs to the $(1,\frac{1}{2})\oplus(\frac{1}{2},1)$ representation. In general, a spinor-valued k-linear wavefunction which is symmetric in its arguments represents a spin-$(k+\frac{1}{2})$ particle; it obeys

$$\nabla\psi_S(a_1,\ldots,a_k;x)i\sigma_3=m\psi_S(a_1,\ldots,a_k;x)\gamma_0, \tag{160}$$

$$\partial_{a_1}\psi_S(a_1,\ldots,a_k;x)=0, \tag{161}$$

where the additional condition restricts the spin part of the wavefunction to the $((k+1)/2,k/2)\oplus(k/2,(k+1)/2)$ representation.

The basis functions of the k-linear wavefunction can also be obtained by polynomial projection. The l-th order contribution to the wavefunction is obtained by setting l arguments of a suitable $(k+l)$-linear function equal to x,

$$\psi_l(a_1,\ldots,a_k;x)=\prod_{i=1}^{l}(x\cdot\partial_{a_{(k+i)}})\psi(a_1,\ldots,a_{k+l}). \tag{162}$$

The symmetry among the first k arguments must be correct for the wavefunction that is being described. Again, the complete wavefunction will be a sum over all orders of these functions.

Different relativistic wave equations with spin corresponding to different representations of the Lorentz group can be explored by selecting the appropriate multilinear wavefunctions. This presents opportunities for further investigation.

ACKNOWLEDGMENTS

MAJA was supported by an EPSRC Research Studentship during the course of this work. SSS is grateful to the Cambridge Livingstone Trust for funding during the course of this work. CJLD is supported by a grant from the Lloyd's of London Tercentenary Foundation.

[1] F. Sommen, ''Clifford tensor calculus,'' in *Proceedings of the 22nd D. G. M. Conference*, Ixtapa, Mexico, 1993.
[2] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).
[3] S. F. Gull, A. N. Lasenby, and C. J. L. Doran, ''Imaginary numbers are not real—the geometric algebra of spacetime,'' Found. Phys. **23**, 1175–1201 (1993).
[4] C. J. L. Doran, D. Hestenes, F. Sommen, and N. Van Acker, ''Lie groups as spin groups,'' J. Math. Phys. **34**, 3642–3669 (1993).
[5] C. J. L. Doran, A. N. Lasenby, and S. F. Gull, ''States and operators in the spacetime algebra,'' Found. Phys. **23**, 1239–1264 (1993).

[6] C. J. L. Doran, A. N. Lasenby, S. F. Gull, A. D. Challinor, and S. S. Somaroo, ''Spacetime algebra and electron physics,'' in *Advances in Imaging and Electron Physics*, edited by P. W. Hawkes (Academic, New York, 1995), Vol. 95.
[7] S. Weinberg, *The Quantum Theory of Fields, Volume 1: Foundations* (Cambridge University Press, Cambridge, 1995).
[8] R. Penrose and W. Rindler, *Spinors and Spacetime* (Cambridge University Press, Cambridge, 1984), Vol. I.
[9] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''2-spinors, twistors and supersymmetry in the spacetime algebra,'' in *Spinors, Twistors, Clifford Algebras and Quantum Deformations*, edited by Z. Oziewicz, A. Borowiec, and B. Jancewicz (Kluwer Academic, Dordrecht, 1993), pp. 233–245.
[10] D. Hestenes, ''Observables, operators and complex numbers in the Dirac theory,'' J. Math. Phys. **16**, 556–572 (1975).
[11] W. Rarita and J. Schwinger, ''On a theory of particles with half-integral spin,'' Phys. Rev. **60**, 61 (1941).

Anomalous dimension in the solution of the Barenblatt's equation

Cheng Chen
Institute of Plasma Physics, Academia Sinica, Hefei, Anhui, 230031, People's Republic of China

G. Cheng
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China, and Fundamental Physics Center, USTC., Hefei, Anhui, 230026, People's Republic of China[a)]

(Received 6 February 1997; accepted for publication 3 December 1997)

A new method is presented to obtain the anomalous dimension in the solution of the Barenblatt's equation. The result is the same as that in the renormalization group (RG) approach. It gives us insight on the perturbative solution of the Barenblatt's equation in the RG approach. Based on this discussion, an improvement is made to take into account, in a more complete way, the nonlinear effect, which is included in the Heaviside function in higher orders. This improved result is better than that in the RG approach. © *1998 American Institute of Physics.* [S0022-2488(98)02903-X]

I. INTRODUCTION

Renormalization and the renormalization group (RG) method are powerful mathematical physics tools in quantum field theory and condensed matter physics.[1,2] Recently, it has been found that there are also important applications in dynamical process, asymptotic analysis, and perturbation theory.[3–5] The fractal and chaos behaviors have been studied with the RG approach.[3] The introduction of the intermediate asymptotic idea is very useful. It can help us to search for the connection between the similarity, fractal, and anomalous dimension of the RG theory.[3] But there are still ambiguities in some aspects. For example, in quantum field theory and in dynamical theory, it is well known that the RG approach can improve the results of the perturbation calculation.[1–5] It can remove the divergence and lead to higher leading-order behavior from a lower-order perturbation calculation. It is also clear that the divergence and the renormalization are physical requirements for some theories. But some of the divergences occur without apparent physical reasons. One question involves the improvement that arises from the RG perturbative calculations. In this paper we study this problem by analyzing a typical example.

Barenblatt's equation has many good characteristics. It has a strong physical background and appears in many different disciplines.[6] It has been studied by applying the method of similarity analysis. It can be solved in terms of the parabolic cylinder functions in the intermediate asymptotic region. Its perturbation solution is divergent and has an anomalous dimension in its asymptotic behavior. It is renormalizable and can be treated with the renormalization and RG approaches.[3,4] But there are still some ambiguities, as mentioned above.

In this paper we study improvements on the perturbation solution of the Barenblatt's equation by renormalization and RG treatment, and attempt to expose some general principles through the discussion of this example. Based on the above analysis, an improved result is obtained by considering the nonlinear effect, which is included in the Heaviside function in higher orders, in a more complete way.

II. ANOMALOUS DIMENSION CALCULATION BY THE RG METHOD

In this section; we repeat some results that are well known, for consistency in notation and discussion style.

We consider the Barenblatt's equation in one space dimension,

---

[a)]Mailing address.

 © 1998 American Institute of Physics

$$\partial_t u(x,t) - \frac{1}{2}\,\partial_x^2 u(x,t) = \frac{\epsilon}{2}\,\theta(-\partial_t u(x,t))\partial_x^2 u(x,t), \tag{1}$$

with a symmetrical initial condition. It describes a nonlinear diffusive process with a nonconservative ''mass.''

Introduce a change of variables,

$$\xi = \xi(x,t) = \frac{x}{\sqrt{t+l^2}} \quad \textit{and} \quad T = T(x,t) = t + l^2. \tag{2}$$

Then Eq. (1) reads as

$$\partial_T u(\xi,T) - \frac{\xi}{2T}\,\partial_\xi u(\xi,T) - \frac{1}{2T}\,\partial_\xi^2 u(\xi,T) = \frac{\epsilon}{2T}\,\theta\left(\left(-\partial_T + \frac{\xi}{2T}\,\partial_\xi\right)u(\xi,T)\right)\partial_\xi^2 u(\xi,T). \tag{3}$$

We use an initial condition of $u(\xi,T)$ as

$$u(\xi,l^2) = \frac{m_0}{\sqrt{2\pi l^2}}\exp\left(-\frac{\xi^2}{2}\right). \tag{4}$$

It corresponds to the initial condition for $u(x,t)$ in the variables x and t as follows:

$$u(x,0) = \frac{m_0}{\sqrt{2\pi l^2}}\exp\left(-\frac{x^2}{2l^2}\right). \tag{5}$$

For a general diffusion equation, there is a finite solution under this initial condition. In this case, the initial condition is singular when l^2 approaches zero as a δ function. m_0 is a finite constant. But it is not so in the case of Barenblatt's equation.[3,4,6] It must be renormalized to get a finite solution. The m_0 is no longer a constant when the l^2 changes. It becomes infinite when the support of the initial state approaches zero. We will discuss that in the end of this section.

Introducing the expansion

$$u(\xi,T) = u_0(\xi,T) + \epsilon u_1(\xi,T) + \cdots, \tag{6}$$

one has a sequence of equations:

$$2T\partial_T u_0(\xi,T) - \xi\partial_\xi u_0(\xi,T) - \partial_\xi^2 u_0(\xi,T) = 0, \tag{7}$$

$$2T\partial_T u_1(\xi,T) - \xi\partial_\xi u_1(\xi,T) - \partial_\xi^2 u_1(\xi,T) = \theta\left(\left(-\partial_T + \frac{\xi}{2T}\,\partial_\xi\right)u_0(\xi,T)\right)\partial_\xi^2 u_0(\xi,T), \tag{8}$$

and so on.

Further, we assume that the initial condition of the $u_0(\xi,T)$ is

$$u_0(\xi,l^2) = \frac{m_0}{\sqrt{2\pi l^2}}\exp\left(-\frac{\xi^2}{2}\right). \tag{9}$$

Then the initial condition of the $u_n(\xi,T)$ is

$$u_n(\xi,l^2) = 0, \quad n \geqslant 1. \tag{10}$$

The solution of the zeroth-order approximation is

$$u_0(\xi,T) = \frac{m_0}{\sqrt{2\pi T}}\exp\left(-\frac{\xi^2}{2}\right). \tag{11}$$

Since one has

$$\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)u_0(\xi,T)=\left(\frac{1}{2T}-\frac{\xi^2}{2T}\right)u_0(\xi,T), \tag{12}$$

one obtains

$$\theta\left(\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)u_0(\xi,T)\right)=\theta(1-|\xi^2|). \tag{13}$$

This is the Heaviside function of the Barenblatt's equation in its lowest order. We can write the first-order equation as

$$2T\partial_T u_1(\xi,T)-\xi\partial_\xi u_1(\xi,T)-\partial_\xi^2 u_1(\xi,T)=\theta(1-|\xi^2|)\partial_\xi^2 u_0(\xi,T). \tag{14}$$

From the similarity analysis we know that the solution of the Barenblatt's equation can be written in a form of variable separation:

$$u(\xi,T)=F(T)G(\xi). \tag{15}$$

It is adequate when we are only interested in the behavior of the solution in the intermediate asymptotic region.

Expanding

$$F(T)=F_0(T)+\epsilon F_1(T)+\ldots \tag{16}$$

and

$$G(\xi)=G_0(\xi)+\epsilon G_1(\xi)+\ldots, \tag{17}$$

we have

$$u_0(\xi,T)=F_0(T)G_0(\xi) \tag{18}$$

and

$$u_1(\xi,T)=F_0(T)G_1(\xi)+F_1(T)G_0(\xi). \tag{19}$$

We now take

$$F_0(T)=\frac{m_0}{\sqrt{2\pi T}}\quad \textit{and}\quad G_0(\xi)=\exp\left(-\frac{\xi^2}{2}\right). \tag{20}$$

The solution in the first-order approximation is

$$u^{(1)}(\xi,T)=F_0(T)G_0(\xi)+\epsilon(F_0(T)G_1(\xi)+F_1(T)G_0(\xi)). \tag{21}$$

It tells us that the approximation solution cannot be expressed in the form of variable separation in finite order of the perturbative theory. But we will see that the RG solution is in this form. Thus, we shall pursue along this line and accept a corresponding approximation, so that we can use the variable separation solution.

Inserting Eq. (19) into Eq. (14) yields

$$2T(F_1'G_0+F_0'G_1)-\xi(F_1G_0'+F_0G_1')-(F_1G_0''+F_0G_1'')=\theta(1-|\xi|)F_0G_0'', \tag{22}$$

where

$$F_i'=\frac{d}{dT}\,F_i(T),\quad G_i'=\frac{d}{d\xi}\,G_i(\xi),\quad \textit{and}\quad G_i''=\frac{d^2}{d\xi^2}\,G_i(\xi),\quad i=0,1. \tag{23}$$

Now, we search for the solution of the $F_1(T)$. Integrate this equation with respect to the ξ from $-\infty$ to $+\infty$ and let

$$\int_{-\infty}^{\infty} G_1(\xi)d\xi = a. \tag{24}$$

After simple integral calculation and the application of the conditions

$$\lim_{\xi\to\pm\infty} \xi G_1(\xi) = 0 \quad \textit{and} \quad \lim_{\xi\to\pm\infty} G_1'(\xi) = 0, \tag{25}$$

one obtains

$$2aF_1(T)TF_0'(T) + 2\sqrt{2\pi}TF_1'(T) + aF_0(T) + \sqrt{2\pi}F_1(T) = -\frac{2}{\sqrt{e}}F_0(T). \tag{26}$$

Substituting $F_0(T) = m_0/\sqrt{2\pi T}$ and $F_0'(T) = -F_0(T)/2T$ into Eq. (26), we get

$$TF_1'(T) + \frac{1}{2}F_1(T) + \frac{1}{\sqrt{2\pi e}}F_0(T) = 0. \tag{27}$$

The solution of Eq. (27) is

$$F_1(T) = -F_0(T)\beta \ln\frac{T}{l^2}, \tag{28}$$

where

$$\beta = \frac{1}{\sqrt{2\pi e}}. \tag{29}$$

For obtaining this solution, we have made use of the initial condition

$$F_n(l^2) = 0, \quad n \geqslant 1. \tag{30}$$

The first-order approximation of $F(T)$ is

$$F^{(1)}(T) = F_0(T)\left(1 - \alpha \ln\frac{T}{l^2}\right), \tag{31}$$

where

$$\alpha = \epsilon\beta = \frac{\epsilon}{\sqrt{2\pi e}}. \tag{32}$$

It is the anomalous dimension and is well known in the solution of the Barenblatt's equation in the intermediate asymptotic region by application of the RG approach.[3,4]

In the following discussion, we keep the $G(\xi)$ in its zeroth-order approximation without considering the higher-order corrections. That is, to put $G_n(\xi) = 0$, $n \geqslant 1$. Then the solution of the first order is

$$u^{(1)}(\xi, T) = F_0(T)G_0(\xi)\left(1 - \alpha \ln\frac{T}{l^2}\right). \tag{33}$$

It has been proved that this approximation does not have any effect on the function $F^{(1)}(T)$ when we make the RG treatment. We will discuss the space behavior in a separate paper.

For getting a finite but nonzero asymptotic solution, the ''mass,''

$$m_0 = \int_{-\infty}^{\infty} dx u(x,t), \tag{34}$$

cannot remain a constant. It is necessary to approach infinity when the support of the initial function approaches zero. To obtain a finite solution of Barenblatt's equation, the $u(x,0)$ must be more singular than the δ function. It is different from the diffusive equation in regular cases. Moreover, the values of the non-zeroth-order terms are not small, but instead, very large in the intermediate asymptotic region. The renormalization is required. It is well known that the renormalization and RG approaches not only remove the divergence but also improve the perturbation expansion solution to higher leading order.[1–5]

Let

$$m_0 = m_R(\mu) Z(\mu, l^2) \tag{35}$$

and

$$Z(\mu, l^2) = 1 + \alpha \ln \frac{\mu}{l^2}. \tag{36}$$

From the work of Goldenfeld *et al.*,[3,4] we know that the result of the renormalization and RG treatment is

$$u_R^{(1)}(\xi, T) = \frac{m_R(l^2)}{\sqrt{2\pi T}} \left(\frac{l^2}{T}\right)^{\alpha} \exp\left(-\frac{\xi^2}{2}\right). \tag{37}$$

It is easy to check that the initial condition, the boundary conditions, and the connection condition are all satisfied when the l^2 is taken in a nonzero value. But we must remember that the renormalization mass will approach the infinite when $l^2 \to 0$, in order to keep the solution finite. Usually we require

$$m_R(l^2) \times l^{2\alpha} = \text{finite}. \tag{38}$$

There is an explanation for the improvement of the RG approach.[7] But the whole case is not very clear yet. We will discuss some aspects of it in the following sections.

III. ANOMALOUS DIMENSION CALCULATION BY SUMMING $F(T)$ TO THE INFINITE ORDER

Now we discuss the improvement to the perturbative solution of the Barenblatt's equation in the RG approach. We do this by using the approximation suggested in the previous section. We are especially interested in the behavior of the intermediate asymptotic region, in which ξ and t are far from the zero point but not infinite. We shall prove that the result of the renormalization and RG approach is equivalent to the infinite sum of the series of $F(T)$.

We assume the solution of Eq. (3) in a form of variable separation in any order. We have known that it is not a suitable form when the approximation order n is finite. But, the higher the approximation order is taken, the better the proposed form is. As the first step, we keep the dependence on space in its lowest order. Thus we set

$$u_n(\xi, T) = f_n(T) u_0(\xi, T). \tag{39}$$

At the same time, we also neglect a part of the nonlinear effects. We keep the Heaviside function in its lowest order as $\theta(1 - |\xi^2|)$. Therefore we have the form of perturbation expansion of the Barenblatt's equation,

$$2T\partial_T u_n(\xi, T) - \xi \partial_\xi u_n(\xi, T) - \partial_\xi^2 u_n(\xi, T) = \theta\left(\left(-\partial_T + \frac{\xi}{2T}\,\partial_\xi\right) u_0(\xi, T)\right) \partial_\xi^2 u_{n-1}(\xi, T), \quad n \geqslant 1. \tag{40}$$

We will prove that the RG anomalous dimension is just the result of summing up the $F(T)$ series to infinite order in the above approximation.

Inserting Eqs. (13) and (39) into Eq. (40) yields

$$2Tu_0(\xi,T)\,\frac{d}{dT}\,f_n(T)=\theta(1-|\xi^2|)f_{n-1}(T)\partial_\xi^2 u_0(\xi,T),\quad n\geqslant 1. \tag{41}$$

The initial condition,

$$u_n(\xi,l^2)=0,\quad n\geqslant 1, \tag{42}$$

becomes

$$f_n(l^2)=0,\quad n\geqslant 1. \tag{43}$$

When $n=1$, one has

$$2Tu_0(\xi,T)\,\frac{d}{dT}\,f_1(T)=\theta(1-|\xi^2|)\partial_\xi^2 u_0(\xi,T). \tag{44}$$

Integrating it with respect to ξ from $-\infty$ to $+\infty$ gives

$$T\,\frac{d}{dT}\,f_1(T)=-\frac{1}{\sqrt{2\pi e}}=-\beta. \tag{45}$$

Using $f_n(l^2)=0$, $n\geqslant 1$, the solution of Eq. (45) is

$$f_1(T)=-\frac{1}{\sqrt{2\pi e}}\ln\frac{T}{l^2}=\ln\left(\frac{T}{l^2}\right)^{-\beta}. \tag{46}$$

When $n=2$, one has

$$2Tu_0(\xi,T)\,\frac{d}{dT}\,f_2(T)=\theta(1-|\xi^2|)f_1(T)\partial_\xi^2 u_0(\xi,T). \tag{47}$$

Doing the same integration as we have done to Eq. (44), we obtain

$$2T\,\frac{m_0}{\sqrt{T}}\,\frac{d}{dT}\,f_2(T)=f_1(T)\int_{-1}^{1}d\xi\partial_\xi^2 u_0(\xi,T). \tag{48}$$

Due to the identity

$$2T\,\frac{m_0}{\sqrt{T}}\,\frac{d}{dT}\,f_1(T)=\int_{-1}^{1}d\xi\partial_\xi^2 u_0(\xi,T), \tag{49}$$

we have

$$\frac{d}{dT}\,f_2(T)=f_1(T)\,\frac{d}{dT}\,f_1(T). \tag{50}$$

Making use of $f_n(l^2)=0$, $n\geqslant 1$, we obtain

$$f_2(T)=\frac{1}{2}\,[f_1(T)]^2. \tag{51}$$

Using the principle of mathematical induction, we now prove that

$$f_m(T)=\frac{1}{m!}[f_1(T)]^m. \tag{52}$$

Suppose that Eq. (52) is true for $m=n-1$, that is to say,

$$f_{n-1}(T)=\frac{1}{(n-1)!}[f_1(T)]^{n-1}. \tag{53}$$

We are going to prove that it is true for $m=n$. Integrating Eq. (41) with respect to ξ from $-\infty$ to $+\infty$ gives

$$2T\frac{m_0}{\sqrt{T}}\frac{d}{dT}f_n(T)=f_{n-1}(T)\int_{-1}^{1}d\xi\partial_\xi^2 u_0(\xi,T). \tag{54}$$

Due to Eq. (49), one has

$$\frac{d}{dT}f_n(T)=f_{n-1}(T)\frac{d}{dT}f_1(T). \tag{55}$$

Substituting Eq. (53) into it gives

$$\frac{d}{dT}f_n(T)=\frac{1}{(n-1)!}[f_1(T)]^{n-1}\frac{d}{dT}f_1(T). \tag{56}$$

Because $f_n(l^2)=0$, the solution of Eq. (56) is

$$f_n(T)=\frac{1}{n!}[f_1(T)]^n. \tag{57}$$

It is true for $m=1$ and $m=2$. Thus we have proved the proposition for any m.

In the infinite limit, we have

$$u(\xi,T)=\left[1+\sum_{n=1}^{\infty}\frac{\epsilon^n}{n!}[f_1(T)]^n\right]u_0(\xi,T). \tag{58}$$

Inserting Eq. (46) into it yields

$$u(\xi,T)=\left[1+\sum_{n=1}^{\infty}\frac{1}{n!}\left[-\epsilon\beta\ln\frac{T}{l^2}\right]^n\right]u_0(\xi,T), \tag{59}$$

that is,

$$u(\xi,T)=\exp\left[-\epsilon\beta\ln\frac{T}{l^2}\right]u_0(\xi,T)=\left[\frac{T}{l^2}\right]^{-\alpha}u_0(\xi,T), \tag{60}$$

where

$$\alpha=\epsilon\beta=\frac{\epsilon}{\sqrt{2\pi e}}, \tag{61}$$

which is just the anomalous dimension given by the renormalization and RG approaches.

IV. AN IMPROVEMENT ON THE CALCULATION OF THE ANOMALOUS DIMENSION

In the previous section, we have kept the Heaviside function and the function $G(\xi)$ in their lowest order in our calculation of the $F(T)$ to higher orders. It has led a partial loss of the nonlinear effect which is included in the Heaviside function. Now we show that the calculation of

the anomalous dimension can be improved by the infinite summing of the function $F(T)$ without the approximation of keeping the Heaviside function in its lowest order to take account of a part of the lost nonlinear effect. Now we still keep the solution in the form of variable separation as Eq. (15), but only let the function $G(\xi)$ in it lowest order. We shall use a trick that makes an analytic treatment possible.

Take the approximation solution obtained in the previous section,

$$u(\xi,T)=\left(\frac{T}{l^2}\right)^{-\alpha} u_0(\xi,T), \tag{62}$$

as an initial solution. After inserting it into the Heaviside function, we solve the $u_n(\xi,T)$ and search for its infinite sum. Thus we obtain the first iterative result.

We put the first iterative solution into the Heaviside function to solve the equation secondly.

In this way, we search for the limit of the iterative solution to infinite. If the limit exists, we can obtain the solution.

An important observation is that the form of the solution remains unchanged in the iterative process,

$$u^{(k)}(\xi,T)=\left(\frac{T}{l^2}\right)^{-\alpha^{(k)}} u_0(\xi,T). \tag{63}$$

The only change is the value of the anomalous dimension $\alpha^{(k)}$. It simplifies the analysis considerably.

We rewrite α as $\alpha^{(0)}$,

$$\alpha=\epsilon\beta=\frac{\epsilon}{\sqrt{2\pi e}}\rightarrow\alpha^{(0)}=\epsilon\beta^{(0)}. \tag{64}$$

The $\alpha^{(k)}=\epsilon\beta^{(k)}$ represents the results after the kth iteration.

We still need to solve the problem by summing up the infinite expansion series, in each step of the iterative process.

Putting Eq. (63) into the Heaviside function of the nth-order equation, we obtain the nth order and $(k+1)$th iteration equation,

$$2Tu_0(\xi,T)\,\frac{d}{dT}\,f_n^{(k+1)}(T)=\theta\left(\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)\left(\left(\frac{T}{l^2}\right)^{-\alpha^{(k)}} u_0(\xi,T)\right)\right) f_{n-1}^{(k+1)}(T)\partial_\xi^2 u_0(\xi,T), \tag{65}$$

where the super-index (k) represents the result of the kth iteration and the subindex n represents the nth order in the perturbation expansion. The initial condition remains in the form of Eq. (43).

When $k=1$ and $n=1$, one has

$$2Tu_0(\xi,T)\,\frac{d}{dT}\,f_1^{(1)}(T)=\theta\left(\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)\left(\left(\frac{T}{l^2}\right)^{-\alpha^{(0)}} u_0(\xi,T)\right)\right)\partial_\xi^2 u_0(\xi,T), \tag{66}$$

where $f_0^{(k)}(T)=1$ has been used. Because of the identity

$$\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)\left(\left(\frac{T}{l^2}\right)^{-\alpha^{(0)}} u_0(\xi,T)\right)=[2\alpha^{(0)}-(\xi^2-1)]\,\frac{1}{2T}\left(\frac{T}{l^2}\right)^{-\alpha^{(0)}} u_0(\xi,T), \tag{67}$$

one has

$$\theta\left(\left(-\partial_T+\frac{\xi}{2T}\,\partial_\xi\right)\left(\left(\frac{T}{l^2}\right)^{-\alpha^{(0)}} u_0(\xi,T)\right)\right)=\theta(2\alpha^{(0)}+1-\xi^2). \tag{68}$$

Substituting it into Eq. (66) and integrating it with respect to ξ from $-\infty$ to $+\infty$ yields

$$2T\frac{m_0}{\sqrt{T}}\frac{d}{dT}f_1^{(1)}(T)=\int_{-\infty}^{\infty}\theta(2\alpha^{(0)}+1-\xi^2)\partial_\xi^2 u_0(\xi,T)$$

$$=\frac{m_0}{\sqrt{2\pi T}}\left[-2\sqrt{(1+2\alpha^{(0)})}\exp\left(-\frac{1+2\alpha^{(0)}}{2}\right)\right], \tag{69}$$

that is,

$$\frac{d}{d\ln T}f_1^{(1)}(T)=\frac{1}{\sqrt{2\pi}}\left[-\sqrt{(1+2\alpha^{(0)})}\exp\left(-\frac{1+2\alpha^{(0)}}{2}\right)\right]. \tag{70}$$

Using $f_1^{(1)}(l^2)=0$, we can obtain the solution of Eq. (70)

$$f_1^{(1)}(T)=\frac{-\sqrt{(1+2\alpha^{(0)})}}{\sqrt{2\pi e^{(1+2\alpha^{(0)})}}}\ln\frac{T}{l^2}=-\beta^{(1)}\ln\frac{T}{l^2}, \tag{71}$$

where

$$\beta^{(1)}=\frac{\sqrt{(1+2\alpha^{(0)})}}{\sqrt{2\pi e^{(1+2\alpha^{(0)})}}}. \tag{72}$$

When $k=1$ and $n=2$, one has

$$2Tu_0(\xi,T)\frac{d}{dT}f_2^{(1)}(T)=\theta(2\alpha^{(0)}+1-\xi^2)f_1^{(1)}(T)\partial_\xi^2 u_0(\xi,T). \tag{73}$$

Integrating it with respect to ξ from $-\infty$ to $+\infty$ yields

$$\frac{d}{dT}f_2^{(1)}(T)=f_1^{(1)}(T)\frac{d}{dT}f_1^{(1)}(T). \tag{74}$$

Making use of $f_n^{(1)}(l^2)=0$, $n\geqslant 2$, the solution of Eq. (74) is

$$f_2^{(1)}(T)=\frac{1}{2}[f_1^{(1)}(T)]^2. \tag{75}$$

Using the principle of mathematical induction, it is easy to prove that

$$f_m^{(1)}(T)=\frac{1}{m!}[f_1^{(1)}(T)]^m. \tag{76}$$

Suppose it is true for $m=n-1$, that is to say

$$f_{n-1}^{(1)}(T)=\frac{1}{(n-1)!}[f_1^{(1)}(T)]^{n-1}. \tag{77}$$

We are going to prove that it is true for $m=n$. We know that

$$2Tu_0(\xi,T)\frac{d}{dT}f_n^{(1)}(T)=\theta(2\alpha^{(0)}+1-\xi^2)f_{n-1}^{(1)}(T)\partial_\xi^2 u_0(\xi,T). \tag{78}$$

Integrating it with respect to ξ from $-\infty$ to $+\infty$ gives

$$\frac{d}{dT}f_n^{(1)}(T)=f_{n-1}^{(1)}(T)\frac{d}{dT}f_1^{(1)}(T). \tag{79}$$

Substituting Eq. (77) into it yields

$$\frac{d}{dT} f_n^{(1)}(T) = \frac{1}{(n-1)!} [f_1^{(1)}(T)]^{n-1} \frac{d}{dT} f_1^{(1)}(T). \tag{80}$$

On account of $f_n^{(1)}(l^2)=0$, one can solve Eq. (80) and obtain

$$f_n^1(T) = \frac{1}{n!} [f_1^{(1)}(T)]^n. \tag{81}$$

We have proved that it is true for $m=1$ and $m=2$. Thus the proposition is true for any m.

Now we sum the infinite series

$$u^{(1)}(\xi,T) = \left[1 + \sum_{n=1}^{\infty} \frac{\epsilon^n}{n!} [f_1^{(1)}(T)]^n \right] u_0(\xi,T). \tag{82}$$

Inserting the expression of $f_1^{(1)}(T)$ into it, we obtain

$$u^{(1)}(\xi,T) = \left[1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left[-\epsilon\beta^{(1)} \ln \frac{T}{l^2} \right]^n \right] u_0(\xi,T). \tag{83}$$

That is,

$$u^{(1)}(\xi,T) = \left[\frac{T}{l^2} \right]^{-\alpha^{(1)}} u_0(\xi,T), \tag{84}$$

where

$$\alpha^{(1)} = \epsilon\beta^{(1)} \quad \textit{and} \quad \beta^{(1)} = \frac{\sqrt{(1+2\alpha^{(0)})}}{\sqrt{2\pi e^{(1+2\alpha^{(0)})}}}. \tag{85}$$

The $\alpha^{(1)}$ is the anomalous dimension corrected to the first order of the $F(T)$. Now, we have finished the first iterative calculation.

When we make the kth iteration, the process is the same as the $(k-1)$th one. The only difference is that we will use $\beta^{(k)}$ in place of the $\beta^{(k-1)}$. Then it is easy to prove that

$$u^{(k)}(\xi,T) = \left[\frac{T}{l^2} \right]^{-\alpha^{(k)}} u_0(\xi,T), \tag{86}$$

where

$$\alpha^{(k)} = \epsilon\beta^{(k)} \quad \textit{and} \quad \beta^{(k)} = \frac{\sqrt{(1+2\alpha^{(k-1)})}}{\sqrt{2\pi e^{(1+2\alpha^{(k-1)})}}}. \tag{87}$$

The $\alpha^{(k)}$ is the anomalous dimension corrected to the kth order of the $F(T)$.

If the iterative process is convergent, the limit exists. Let $\lim_{k\to\infty} \beta^{(k)} = \beta^{(\infty)}$, which leads to

$$\beta^{(\infty)} = \frac{\sqrt{(1+2\epsilon\beta^{(\infty)})}}{\sqrt{2\pi e^{(1+2\epsilon\beta^{(\infty)})}}}. \tag{88}$$

Then one has

$$u^{(\infty)}(\xi,T) = \left[\frac{T}{l^2} \right]^{-\alpha^{(\infty)}} u_0(\xi,T), \tag{89}$$

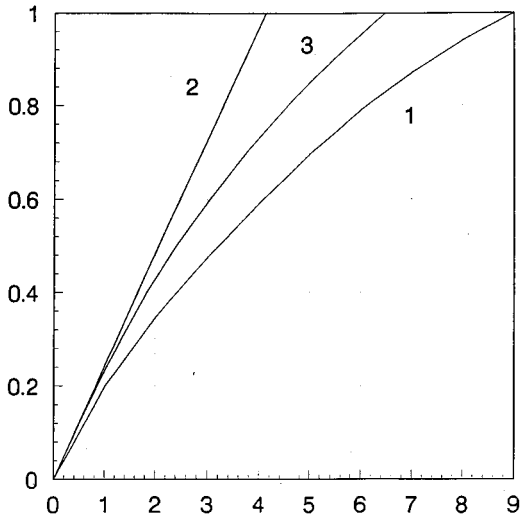

FIG. 1. The function $\alpha(\epsilon)$ vs ϵ, with (1) the numerical result from Barenblatt by similarity analysis; (2) the RG result from Goldenfeld *et al.*; and (3) our result by using an infinite sum of the $F(T)$.

where

$$\alpha^{(\infty)}=\epsilon\beta^{(\infty)} \tag{90}$$

and

$$\alpha^{(\infty)}=\frac{\epsilon\sqrt{(1+2\alpha^{(\infty)})}}{\sqrt{2\pi e^{(1+2\alpha^{(\infty)})}}}. \tag{91}$$

It is obvious that Eq. (91) has a finite solution, which is the anomalous dimension corrected by considering the nonlinear effect of $F(T)$ to infinite order.

The justification of the iterative method and its initial value is that the solution is convergent and satisfies the Barenblatt's equation in our approximation.

We can compare the improved result with the numerical one given by Barenblatt and the result obtained by the RG approach, as shown in Fig. 1. The improved result $\alpha^{(\infty)}$ is better than $\alpha^{(0)}$.

V. CONCLUSION

We have calculated the anomalous dimension by summing the $F(T)$ to infinite order. When we keep the Heaviside function and the space-dependent function $G(\xi)$ in their lowest order, the anomalous dimension is the same as the RG result. When the approximation of the Heaviside function in its lowest order is removed, the value of the anomalous dimension can be improved. It is expressed in Eq. (91) and Fig. 1.

ACKNOWLEDGMENTS

G. Cheng is supported in part by the National Science Foundation in China and by the NDSTPR Foundation in China. The authors would like to thank J. L. Zhou, Dr. Zhang Yang, and Dr. Han Lang for their help in this paper.

[1] N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields*, 3rd ed. (Wiley, New York, 1980); J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Clarendon, Oxford, 1989).

[2] M. Gell-Mann and F. T. Low, Phys. Rev. **95**, 1300 (1964); K. Symanzik, Commun. Math. Phys. **18**, 227 (1970); C. G. Callan, Phys. Rev. D **2**, 1541 (1970); K. G. Wilson, Phys. Rev. B **4**, 3174, 3184 (1971); K. G. Wilson and J. Kogut, Phys. Rep., Phys. Lett. C **12**, 75 (1974).

[3] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, New York, 1992); R. J. Creswick, H. A. Farach, and C. P. Poole, Jr., *Introduction to Renormalization Group Methods in Physics* (Wiley, New York, 1992).
[4] N. Goldenfeld, O. Martin, and Y. Oono, J. Sci. Comput. **4**, 4 (1989); N. Goldenfeld, O. Martin, Y. Oono, and F. Liu, Phys. Rev. Lett. **64**, 1361 (1990); L-Y. Chen, N. Goldenfeld, and Y. Oono, Phys. Rev. A **44**, 6544 (1991); J. Bricmont and A. Kupiainen, Commun. Math. Phys. **150**, 193 (1992).
[5] L-Y. Chen, N. Goldenfeld, and Y. Oono, Phys. Rev. Lett. **73**, 1311 (1994); L-Y. Chen, N. Goldenfeld, Y. Oono, and G. Paquette, Physica A **204**, 111 (1994); G. Paquette, L-Y. Chen, N. Goldenfeld, Y. Oono, Phys. Rev. Lett. **72**, 76 (1994). G. Caginalp, Phys. Rev. E **53**, 66 (1996).
[6] G. I. Barenblatt, *Similarity, Self-Similarity and Intermediate Asymptotics* (Consultants Bureau, New York, 1979).
[7] T. Kunihiro, Prog. Theor. Phys. **97**, 503 (1995).

Demazure modules and vertex models: The $\widehat{\mathrm{sl}}(2)$ case

Omar Foda[a)]
Instituut voor Theoretische Fysica, Universiteit Utrecht, Utrecht 3508 TA, The Netherlands

Kailash C. Misra
Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205

Masato Okado
Department of Mathematical Sciences, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka 560, Japan

(Received 5 September 1996; accepted for publication 20 October 1997)

We characterize, in the case of $\widehat{\mathrm{sl}}(2)$, the crystal base of the Demazure module $E_w(\Lambda)$ in terms of extended Young diagrams or paths for any dominant integral weight Λ and Weyl group element w. Its character is evaluated via two expressions, ''bosonic'' and ''fermionic.'' © *1998 American Institute of Physics.* [S0022-2488(98)00503-9]

I. INTRODUCTION

Our purpose in this work is to characterize the crystal bases of the Demazure modules of $\widehat{\mathrm{sl}}(2)$ in terms of ''paths,'' or extended Young diagrams, and obtain explicit expressions for their full characters. There are two ways to motivate this work: one can motivate it from the viewpoint of certain recent developments in mathematical physics, or from the viewpoint of representation theory.

A. One-point functions in exactly solvable models

Certain physical quantities in exactly solvable two-dimensional lattice models, namely, the so-called ''one-point functions'' can be evaluated using Baxter's corner transfer matrix method. (For an introduction to exactly solvable models, and to the corner transfer matrix method, we refer to Ref. 1.) This method reduces the computation of one-point functions to a computation of a weighted sum. This sum is over combinatorial objects called ''paths,'' one-dimensional configurations, defined on the half-line. They will be discussed in detail in the sequel. The weight of a path is the evaluation, on that path, of an ''energy'' functional defined on the set of all paths. The weighted sum is the generating function of the number of paths of a certain weight.

One way to evaluate this generating function is by solving a recursion relation for the generating function that counts the number of paths defined on a finite segment, of length L, of the half-line. The recursion is with respect to L. The $L\to\infty$ limit of the result, which is the generating function of all paths, turns out to be the character of a highest weight module of an infinite-dimensional algebra: an affine, or Virasoro algebra. (For an introduction to the algebraic approach to exactly solvable models, we refer to Ref. 2.) The result obtained by solving a recursion relation typically has a form that, for physical reasons, is called ''bosonic.'' (For an explanation of the origin of this terminology, we refer to Ref. 3.)

B. Rogers–Ramanujan-type *q*-series identities

More recently, the same physical objects were evaluated in a completely different way, using Bethe Ansatz methods.[3] In that case, the results turn out to have a completely different form, that, for the same physical reasons as above, is referred to as ''fermionic.''

Equating the bosonic and fermionic expressions for the same objects, one obtains Rogers–Ramanujan-type identities between q series. Strictly speaking, such identities are conjectures, since the methods used to obtain them are indirect, and involve certain ''physical assumptions.''

[a)]Permanent address: Department of Mathematics, The University of Melbourne, Parkville, Victoria 3052, Australia.

 © 1998 American Institute of Physics

(What we have in mind are the various assumptions involved in the derivation of the corner transfer matrix method, and the ''string hypothesis.'') What one needs are direct proofs.

C. Schur-type polynomial identities

One method to obtain direct proofs, which dates back to Schur, is to work at the level of the generating functions of all paths that live on a segment of the half-line of length L. These are polynomials that depend on L. Once we can prove the boson-fermion identities for finite L, we take the limit $L\to\infty$, and obtain the original q-series identities.

The following question now arises: If the original infinite q series are characters of highest weight modules of infinite-dimensional algebras, are the finite-L polynomials the characters of anything? Do they have a meaning in mathematics? Or are they just convenient objects that appear in intermediate steps?

The reason why this question arises is that rigorous proofs in exactly solvable models are typically obstructed by the fact that we have to deal with infinite series, and infinite-dimensional quantities. The reason why we typically have to do that, is that only in the limit $L\to\infty$ do these models exhibit invariance under infinite-dimensional algebras, and this invariance is an essential ingredient in solvability. Thus, we have the following problem: we need to work in the $L\to\infty$ limit in order to be able to obtain exact solutions, but in that limit we cannot provide rigorous proofs.

This is the reason why it is very interesting to try to understand as much of the structure of exactly solvable models as possible at the level of finite L. The hope is that we can find algebraic structures at finite L that are sufficiently strong to provide rigorous proofs and solutions. A typical object to investigate are the L-restricted generating functions.

What we find in this work, is that these objects do have a mathematical significance, or are directly related to objects that do. We find that the L-restricted generating functions are closely related to the characters of Demazure modules.

D. Demazure modules

From the viewpoint of representation theory, we can motivate our work in the evaluation of the characters of Demazure modules. The Demazure module is characterized by its highest weight Λ and Weyl group element w, and is denoted by $E_w(\Lambda)$. In Refs. 4–5, a formula for computing the Demazure character $\mathrm{ch}\, E_w(\Lambda)=\Sigma_{\mu\in P}\dim(E_w(\Lambda))_\mu e^\mu$ is given as follows. For $\mu\in P, i=0,1$, define the operator $D_i:\mathbb{Z}[P]\to\mathbb{Z}[P]$ by

$$D_i(e^\mu)=\frac{e^{\mu+\rho}-e^{r_i(\mu+\rho)}}{1-e^{-\alpha_i}}\,e^{-\rho}.$$

Let $w=r_{i_n}\cdots r_{i_2}r_{i_1}\in W$ be a reduced expression. For the terminologies of $\widehat{\mathrm{sl}}(2)$ see Sec. II A. Then, the Demazure character formula states that

$$\mathrm{ch}\, E_w(\Lambda)=D_{i_n}\cdots D_{i_2}D_{i_1}(e^\Lambda).$$

As elegant as this character formula is, it is anything but combinatorial. Hence, using this formula it is very difficult to get any information about specific weight spaces. In Ref. 6, Sanderson used Littelmann's path model[7] and gave nice expressions for the ''real'' characters of $E_w(\Lambda)$ for all Λ and the ''principal'' characters of $E_w(\Lambda)$ for $\Lambda=s\Lambda_0$. We point out that the real (resp., principal) character of $E_w(\Lambda)$ is the specialization of $e^{-\Lambda}\,\mathrm{ch}\, E_w(\Lambda)$, where $e^{-\alpha_0}=q$, $e^{-\delta}=1$ (resp., $e^{-\alpha_0}=q$, $e^{-\delta}=q^2$).

E. Crystal bases

In 1990, Kashiwara brought a notion of crystal base in the representation theory of the quantum group.[8] This notion is quite powerful in the combinatorial aspect of representation theory. As an example, we find Ref. 9, in which the Littlewood–Richardson-type rules for the classical Lie algebras are given in a purely combinatorial way. Kashiwara also obtained the crystal

bases of the Demazure modules as subsets of those of the corresponding highest weight modules.[10] His algorithm is given in a recursive way. Consequently he got a new proof of the Demazure character formula.

Since we know that in the case of affine Lie algebra the crystal base of the highest weight module is described in terms of paths,[11,12] it is natural to try to characterize the crystal base of the Demazure module in this language. For the affine algebra of type A we have yet another notion to describe the crystal base, that is, an extended Young diagram.[13,14]

F. Main results

What we have done in our work is the characterization of Demazure crystal bases in terms of extended Young diagrams (Theorem 2). One of the advantages to do so is that we can make the most of earlier efforts toward the evaluations of so-called one-dimensional configuration sums.[15,16] Recent results on fermionic expressions are also available. We would like to note here that after we had obtained the fermionic expression (Theorem 3), we came to know Schilling also had the same result.[17] However, we have included our proof since it is different from that of Schilling. As a corollary of the characterization theorem, we evaluate the characters of Demazure modules (Theorems 4 and 5).

G. Outline of paper

In Sec. II, we define the basic combinatorial objects in this work: paths and extended Young diagrams. We then describe the crystal structure of the integrable highest weight module of $U_q(\hat{\mathfrak{sl}}(2))$. In Sec. III, we introduce the Demazure modules, and characterize their crystals via paths or extended Young diagrams. In Sec. IV, we review one-dimensional configuration sums and rewrite them into fermionic forms. Using them we can express Demazure characters. We also discuss some specializations and compare them with Sanderson's results. Section V contains a short discussion.

II. PATHS AND EXTENDED YOUNG DIAGRAMS

In this section we recall the basic combinatorial objects used in this work: the paths and extended Young diagrams. Next, we outline the realization of the crystal base of an integrable $\hat{\mathfrak{sl}}(2)$ module in terms of these objects. We also list some properties of the crystal base that we will need in the sequel.

A. Preliminaries

Consider the affine Lie algebra $\hat{\mathfrak{sl}}(2)$.[18] Let $\{\alpha_0,\alpha_1\}$, $\{h_0,h_1\}$, and $\{\Lambda_0,\Lambda_1\}$ be the simple roots, simple coroots, and fundamental weights, respectively. They satisfy $\Lambda_i(h_j)=\delta_{ij}$ and $\alpha_i(h_j)=2(-1)^{1+\delta_{ij}}$ for $i,j=0,1$. $\delta=\alpha_0+\alpha_1$, $c=h_0+h_1$ are the null root and canonical central element, respectively. The sets $P=\mathbb{Z}\Lambda_0\oplus\mathbb{Z}\Lambda_1\oplus\mathbb{Z}\delta$, and $Q=\mathbb{Z}\alpha_0\oplus\mathbb{Z}\alpha_1$ are the weight and root lattices, respectively. Let $\rho\in\mathfrak{h}^*$ be such that $\rho(h_i)=1$ for $i=0,1$.

Let $U_q(\hat{\mathfrak{sl}}(2))$ denote the quantized universal enveloping algebra associated with $\hat{\mathfrak{sl}}(2)$. For its precise definition and Hopf algebra structure, we refer to Ref. 14. For $\Lambda\in P,\Lambda(c)$ is the level of Λ. The set $P^+=\{\Lambda\in P|\Lambda(h_i)\geqslant 0,i=0,1\}$ is the set of dominant weights. For $\Lambda\in P^+$, let $V(\Lambda)$ denote the unique (up to isomorphism) integrable highest weight $U_q(\hat{\mathfrak{sl}}(2))$ module. Since $V(\Lambda+k\delta)\simeq V(\Lambda)\otimes V(k\delta)$ and $\dim V(k\delta)=1$, it suffices to assume that $\Lambda=s\Lambda_0+t\Lambda_1$ for some $s,t\in\mathbb{Z}_{\geqslant 0}$. Let $\Lambda(c)=s+t=k$ be the level of Λ.

B. Paths

For convenience, we extend the subscript i of Λ_i to $i\in\mathbb{Z}$, by setting $\Lambda_i=\Lambda_{i'}$ for $i\equiv i'\pmod 2$. We also set $\hat{i}=\Lambda_{i+1}-\Lambda_i(i=0,1)$. For $k\in\mathbb{Z}_{\geqslant 0}$, we define a set of level k weights P_k by $P_k=\{a_0\Lambda_0+a_1\Lambda_1|a_0,a_1\in\mathbb{Z},a_0+a_1=k\}$.

Definition 1 [path on P_k]: Fixing $k,L\in\mathbb{Z}_{\geqslant 0}$ we define a path p of length L as a sequence $p=(p_0,\ldots,p_L,p_{L+1})$, with all $p_i\in P_k$ and $p_{i+1}-p_i\in\{k\hat{0},(k-1)\hat{0}+\hat{1},\ldots,k\hat{1}\}$.

Definition 2 [set of paths $\mathscr{P}_L(\Lambda)$]: For $L\in\mathbb{Z}_{\geqslant 0}$ and a dominant weight $\Lambda=s\Lambda_0+t\Lambda_1(s+t=k)$, we set

$$\mathscr{P}_L(\Lambda)=\left\{p=(p_0,\ldots,p_L,p_{L+1})\,\middle|\,\begin{array}{l}p \ \text{is a path on} \ P_k,\\ p_i=s\Lambda_i+t\Lambda_{i+1}, \quad \text{for} \ \ i=L,L+1\end{array}\right\}.$$

To the set $\mathscr{P}_L(\Lambda)$, we associate a special path $\overline{p}$, called the ground-state path, defined as follows:

Definition 3 [ground-state path $\overline{p}$]:

$$\overline{p}=(\overline{p_0}\ldots,\overline{p_L},\overline{p_{L+1}}), \quad \overline{p_i}=s\Lambda_i+t\Lambda_{i+1}, \quad \text{for all} \ \ i.$$

The relevance of the ground-state path will become clear in the sequel.

For $k\in\mathbb{Z}_{\geqslant 0}$ set $\mathscr{A}_k^+=\{a_0\epsilon_0+a_1\epsilon_1|a_0,a_1\in\mathbb{Z}_{\geqslant 0},a_0+a_1=k\}$. Here we consider ϵ_0 and ϵ_1 as symbols. Just as in the case of Λ_i, we extend the subscript i of ϵ_i to $i\in\mathbb{Z}$ by setting $\epsilon_i=\epsilon_{i'}$ for $i\equiv i'(\mathrm{mod}\ 2)$. We encode a path in terms of a sequence of elements in $\mathscr{A}_k^+$ as follows.

Definition 4 [sequence of elements $\iota(p)$]: Let $\Lambda=s\Lambda_0+t\Lambda_1(s+t=k)$. For a path $p=(p_0,\ldots,p_L,p_{L+1})\in\mathscr{P}_L(\Lambda)$ we define a sequence $\iota(p)=(\iota(p)_0,\ldots,\iota(p)_L)$ by $\iota(p)_i=m_i\epsilon_0+(k-m_i)\epsilon_1\in\mathscr{A}_k^+$, where for each i, m_i is determined by $p_{i+1}-p_i=m_i\hat{0}+(k-m_i)\hat{1}$.

For the ground-state path $\overline{p}$, $\iota(\overline{p})$ is given by $\iota(\overline{p})_i=s\epsilon_i+t\epsilon_{i+1}$. We remark that the data $\iota(p)$ uniquely determine $p\in\mathscr{P}_L(\Lambda)$.

Define a function $\mu:\mathscr{A}_k^+\to\{0,1,\ldots,k\}$ by

$$\mu(m\epsilon_0+(k-m)\epsilon_1)=m.$$

Using μ, we define the energy, and the weight of a path p as follows:

Definition 5 [energy of a path $E(p)$]:

$$E(p)=\sum_{j=1}^{L} j(H(\iota(p)_{j-1},\iota(p)_j)-H(\iota(\overline{p})_{j-1},\iota(\overline{p})_j)),$$

where

$$H(\epsilon,\epsilon')=\max(k-\mu(\epsilon),\mu(\epsilon')), \quad \text{for} \ \ \epsilon,\epsilon'\in\mathscr{A}_k^+. \tag{1}$$

Definition 6 [weight of a path wt *p]:*

$$\mathrm{wt}\ p=p_0-E(p)\delta.$$

We note that for the ground-state path $\overline{p}\in\mathscr{P}_L(\Lambda)$ we have $E(\overline{p})=0$ and wt $\overline{p}=\Lambda$. We will need the following lemma in the sequel.

Lemma 1: Let $\Lambda=s\Lambda_0+t\Lambda_1(s+t=k)$. For the ground-state path $\overline{p}\in\mathscr{P}_L(\Lambda)$, we have

$$\sum_{j=1}^{L} jH(\iota(\overline{p}_{j-1}),\iota(\overline{p}_j))=\left(\frac{L+\epsilon^{(L)}}{2}\right)^2 k+(-1)^{\epsilon^{(L)}}\frac{L+\epsilon^{(L)}}{2}s.$$

Here $\epsilon^{(L)}$ is defined by

$$\epsilon^{(L)}=\begin{cases}0 & (L:\textit{even}),\\ 1 & (L:\textit{odd}).\end{cases} \tag{2}$$

Proof: Note that $\iota(\overline{p}_j)=s\epsilon_j+t\epsilon_{j+1}$. From (1) we obtain the following expression to calculate:

$$\sum_{\substack{j=1\\ j:\mathrm{odd}}}^{L} jt+\sum_{\substack{j=1\\ j:\mathrm{even}}}^{L} js,$$

which is evaluated easily. ■

So far, we fixed the length of paths L. Next, we consider two lengths L and $L'(L\leqslant L')$. Let $\Lambda=s\Lambda_0+t\Lambda_1$. Then there is an injection given by

$$\mathscr{P}_L(\Lambda)\to\mathscr{P}_{L'}(\Lambda)$$

$$p=(p_0,\ldots,p_L,p_{L+1})\mapsto p'=(p'_0,\ldots,p'_{L'},p'_{L'+1}),$$

where $p'_i=p_i(0\leqslant i\leqslant L+1),=s\Lambda_i+t\Lambda_{i+1}(L+1\leqslant i\leqslant L'+1)$. Let $\overline{p'}$ be the ground-state path of $\mathscr{P}_{L'}(\Lambda)$. The image of $\mathscr{P}_L(\Lambda)$ in $\mathscr{P}_{L'}(\Lambda)$ is characterized by the paths $\{p'\}$ such that $p'_i=\overline{p'_i}(i\geqslant L)$. Note that the energy and weight of a path are unchanged by this injection. We have the following inductive system:

$$\mathscr{P}_0(\Lambda)\to\mathscr{P}_1(\Lambda)\to\cdots\to\mathscr{P}_L(\Lambda)\to\cdots .$$

We define the set of all paths $\mathscr{P}(\Lambda)$ as follows.

Definition 7 [$\mathscr{P}(\Lambda)$]:

$$\mathscr{P}(\Lambda)=\lim_{\to}\mathscr{P}_L(\Lambda).$$

C. Extended Young diagrams

An extended Young diagram Y is an infinite sequence $(y_j)_{j\geqslant 0}$ of integers, such that $y_j\leqslant y_{j+1}$ for all j and $y_j=y_\infty$ (a fixed integer) for sufficiently large j.[15,13,14] In other words, the sequence stabilizes after a finite (though arbitrarily large) number of elements. The number y_∞ is called the ''charge'' of Y. In this paper, we only consider extended Young diagrams of charge $\gamma=0$ or 1.

We can view an extended Young diagram $Y=(y_j)_{j\geqslant 0}$ as a diagram drawn on the lattice in the right-half-plane with sites $\{(m,n)\in\mathbb{Z}\times\mathbb{Z}|m\geqslant 0\}$, where y_j denotes the depth of the jth column. If $y_j\neq y_{j+1}$, for some j, then there will be concave (''⌈''), and convex (''⌋'') corners. A corner located at site (m,n) has diagonal number $d=m+n$.

1. A coloring scheme

We assign each corner of $Y=(y_j)_{j\geqslant 0}$ one of two colors: 0=''white,'' or 1=''black'' as follows: A d-diagonal corner is white (resp., black), if d is even (resp., odd). A white (resp., black) corner will also be called a 0-corner (resp., 1-corner).

Thus, we can also view an extended Young diagram $Y=(y_j)_{j\geqslant 0}$ of charge 0 (resp., 1) as a usual Young diagram, but with its nodes alternately colored ''white'' and ''black,'' such that the top leftmost node has color ''white'' (resp., ''black'').

We define the weight wt Y of an extended Young diagram Y of charge i $(=0$ or $1)$ to be

$$\text{wt } Y=\Lambda_i-k_0\alpha_0-k_1\alpha_1,$$

where k_0 is the number of white nodes in Y and k_1 is the number of of black nodes. We also define the ''width of Y,'' $|Y|$, to be the number of nodes in the first row of Y. Equivalently, $|Y|=L$ if and only if $y_L=y_\infty$ (=charge of Y) but $y_{L-1}<y_L$.

2. Example

Consider the extended Young diagram $Y=(y_j)_{j\geqslant 0}=(-2,-1,-1,0,0,1,1,1,\ldots)$ of charge 1. Y has convex 1-corners at sites $(1,-2)$ and $(5,0)$, a convex 0-corner at site $(3,-1)$, a concave 1-corner at site $(3,0)$, and a concave 0-corner at site $(1,-1)$. Also note that wt $Y=\Lambda_1-4\alpha_0-5\alpha_1$ and $|Y|=5$.

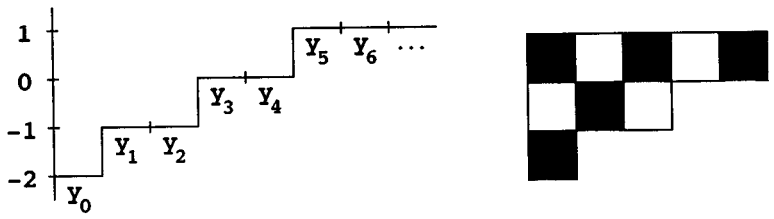

D. From paths to extended Young diagrams

Let us define a ''pattern'' to be a map,

$$t:\mathbb{Z}\times\mathbb{Z}_{\geqslant 0}\to\mathbb{Z},$$

$$(i,j)\mapsto t_{ij},$$

such that (i) for all i, $(t_{ij})_{j\geqslant 0}$ is an extended Young diagram; (ii) $t_{ij}\leqslant t_{i+1j}$ for all i and j; (iii) $t_{i+k\,j}=t_{ij}+2$ for all i and j.

We say the pattern t is normalized if $0\leqslant\gamma_1\leqslant\cdots\leqslant\gamma_k<2$, where $\gamma_i=t_{i\infty}$ is the charge of $(t_{ij})_{j\geqslant 0}$. We call $\gamma=(\gamma_1,\ldots,\gamma_k)$ the charge of t. Viewing condition (iii) we can identify the pattern t with a k-tuple $\mathbf{Y}=(Y_1,\ldots,Y_k)$ of extended Young diagrams $Y_i=(t_{ij})_{j\geqslant 0}$. In terms of extended Young diagrams, (ii) is equivalent to the following inclusion rule:

$$Y_1\supset Y_2\supset\cdots\supset Y_k\supset Y_1[2], \tag{3}$$

where $Y_1[2]$ denotes the extended Young diagram that is obtained by shifting Y_1 upward two units on the lattice in the right-half-plane. Let $\Lambda=s\Lambda_0+t\Lambda_1(s+t=k)$. Denote the set of patterns with charge $\gamma=(\underbrace{0,\ldots,0}_{s},\underbrace{1,\ldots,1}_{t})$ by $\mathscr{T}(\Lambda)$. Then we have a map,

$$\pi:\mathscr{T}(\Lambda)\to\mathscr{P}(\Lambda),$$

$$t=(t_{ij})\mapsto p,$$

where p is determined from $\iota(p)_j=\epsilon_{t_{1j}+j}+\cdots+\epsilon_{t_{kj}+j}$. For a path $p\in\mathscr{P}(\Lambda)$ we say t is a ''lift'' of p if $t\in\pi^{-1}(p)$. Among lifts of $p\in\mathscr{P}(\Lambda)$ there exists a unique normalized lift $t=(t_{ij})\in\mathscr{T}(\Lambda)$ such that $t_{ij}\geqslant t'_{ij}$ for all i,j for any $t'=(t'_{ij})\in\pi^{-1}(p)$ (Proposition 3.4 in Ref. 14). $t(p)$ is called the ''highest lift'' of p. Let $\mathbf{Y}=(Y_1,\ldots,Y_k)$ be the k-tuple of extended Young diagrams corresponding to the highest lift of p. Then we have (Theorem 5.7 in Ref. 15)

$$\text{wt } p=\text{wt } Y_1+\cdots+\text{wt } Y_k.$$

E. The crystal $B(\Lambda)$

Let $(L(\lambda),B(\Lambda))$ be the crystal base of $V(\lambda)$. In this section we give the combinatorial rules used to construct the crystal $B(\Lambda)$, following the work of Refs. 14 and 13.

1. The rules

Let $\Lambda=s\Lambda_0+t\Lambda_1\in P^+$ of level $k=s+t$. We order the elements of $\mathbb{Z}\times\{1,2,\ldots,k\}$ as follows. For (d,j), $(d',j')\in\mathbb{Z}\times\{1,2,\ldots,k\}$, we say

$$(d,j)>(d',j') \text{ if and only if } d>d' \text{ or } d=d' \text{ and } j<j'.$$

Let $\mathscr{Y}(\Lambda)$ denote the set of k-tuples of extended Young diagrams $\mathbf{Y}=(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)$, such that Y_j has charge 0 (resp., 1) for $1\leqslant j\leqslant s$ (resp., $s+1\leqslant j\leqslant k$). For $\mathbf{Y}\in\mathscr{Y}(\Lambda)$ and fixed i $(=0$ or $1)$, we define the i signature of $\mathbf{Y}$ to be the sequence $\varepsilon=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m)$ (Note that we distinguish ε_j from ϵ_j, which has appeared in Definition 5), such that we see the following:

(A1) $\Sigma_{j=1}^{k}\sharp\{i\text{-corners of } Y_j\}=m$;
(A2) each ε_r is either 0 or 1;
(A3) we can define $j(r)(1\leqslant j(r)\leqslant k)$ and $d(r)$ in such a way that $Y_{j(r)}$ has a $d(r)$-diagonal i-corner;
(A4) if $\varepsilon_r=0$ (resp., 1) then $Y_{j(r)}$ has a $d(r)$-diagonal concave (resp., convex) i-corner;
(A5) if $r_1<r_2$, then $(d(r_1),j(r_1))>(d(r_2),j(r_2))$.

For fixed i signature $\varepsilon=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m)$ we partition the set $\{1,2,...,m\}=J\sqcup K_1\sqcup\cdots\sqcup K_r$ into disjoint subsets by the following procedure:
(B1) if there is no j such that $(\varepsilon_j,\varepsilon_{j+1})=(0,1)$ define $J=\{1,2,...,m\}$,
(B2) if there is some j such that $(\varepsilon_j,\varepsilon_{j+1})=(0,1)$ define $K_1=\{j,j+1\}$,
(B3) apply (B1) and (B2) above to $\{1,2,...,m\}\backslash K_1$ to choose J or K_2 and repeat this as necessary to choose J and $K_1,\ldots,K_r$.

Let $\varepsilon_J=(\varepsilon_{j_1},\ldots,\varepsilon_{j_r})$, where $J=\{j_1,\ldots,j_r\}$ and $j_1<\cdots<j_r$. We call a 0 or 1 in the i-signature ''relevant'' if and only if it is in ε_J.

For fixed $i\in\{0,1\}$ and $\mathbf{Y}$, $\mathbf{Y}'\in\mathscr{Y}(\Lambda)$ suppose that the following conditions hold: (C1) $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_m)$ and $\varepsilon'=(\varepsilon'_1,\ldots,\varepsilon'_m)$ are the i-signatures of $\mathbf{Y}$ and $\mathbf{Y}'$ respectively; (C2) the partition $\{1,2,...,m\}=J\sqcup K_1\sqcup\cdots\sqcup K_r$ is the same for both ε and ε'; and (C3) there exists $l\in J$ such that $\varepsilon_l=0$, $\varepsilon'_l=1$, and $\varepsilon_j=\varepsilon'_j=1$ (resp., 0) if $j\in J$ and $j<l$ (resp., $j>l$).

We define $\tilde{f}_i\mathbf{Y}=\mathbf{Y}'$ and $\tilde{e}_i\mathbf{Y}'=\mathbf{Y}$ if and only if the above conditions hold. If there exists no such $\mathbf{Y}'$ (resp., $\mathbf{Y}$) for $\mathbf{Y}$ (resp., $\mathbf{Y}'$), we define $\tilde{f}_i\mathbf{Y}=0$ (resp., $\tilde{e}_i\mathbf{Y}'=0$).

2. Example

Let $\Lambda=\Lambda_0+\Lambda_1$ and take $\mathbf{Y}=(Y_1,Y_2)\in\mathscr{Y}(\Lambda)$, where $Y_1=(-2,-2,-1,-1,0,0,...)$ is of charge 0 and $Y_2=(-1,0,0,1,1,...)$ is of charge 1.

$$\mathbf{Y}=(Y_1,Y_2)=\Big(\ ,\ \Big).$$

The 0-signature of $\mathbf{Y}$ is $\varepsilon=(0,0,1,1,0)$. So using rule B, $\varepsilon_J=(0)$. Hence by using rule C we have

$$\tilde{f}_0\mathbf{Y}=\mathbf{Y}'=\Big(\ ,\ \Big)\quad\text{and}\quad\tilde{e}_0\mathbf{Y}=0.$$

However, the 1-signature of $\mathbf{Y}$ is $\varepsilon=(1,1,0,0,0)$ and by rule B, $\varepsilon_J=\varepsilon$. So by rule C we have

$$\tilde{f}_1\mathbf{Y}=\Big(\ ,\ \Big)\quad\text{and}\quad\tilde{e}_1\mathbf{Y}=\Big(\ ,\ \Big).$$

Thus, in general, $\tilde{e}_i,\tilde{f}_i(i=0,1)$ are well-defined maps from $\mathscr{Y}(\Lambda)$ to $\mathscr{Y}(\Lambda)\sqcup\{0\}$. Now let $\Phi=(\phi_1,\ldots,\phi_s,\phi_{s+1},\ldots,\phi_k)\in\mathscr{Y}(\Lambda)$ be the k-tuple of empty extended Young diagrams, where ϕ_j has charge 0 (resp., 1) for $1\leqslant j\leqslant s$ (resp., $s+1\leqslant j\leqslant k$). Note that wt $\Phi=\Lambda=s\Lambda_0+t\Lambda_1$, and $\tilde{e}_i\Phi=0$ for $i=0,1$. We call Φ the highest weight vector (or vacuum) of weight Λ.

We now define the set $B(\Lambda)\subset\mathscr{Y}(\Lambda)$ as

$$B(\Lambda)=\{\tilde{f}_{i_1}\tilde{f}_{i_2}\cdots\tilde{f}_{i_k}\Phi\,|\,k\geqslant 0,i_j=0,1\}\backslash\{0\}.$$

For $\mathbf{Y}\in B(\Lambda)$, $i=0,1$, define

$$\varepsilon_i(\mathbf{Y})=\max\{m\geqslant 0\,|\,\tilde{e}_i^m\mathbf{Y}\in B(\Lambda)\},$$

$$\varphi_i(\mathbf{Y})=\max\{m\geqslant 0\,|\,\tilde{f}_i^m\mathbf{Y}\in B(\Lambda)\}.$$

Thus, $\varepsilon_i, \varphi_i : B(\Lambda) \to \mathbb{Z}$. (We use this new definition of ε_i only here. We hope this will not cause any confusion.) We also have the map $\mathrm{wt}: B(\Lambda) \to P$ given by $\mathbf{Y} \mapsto \mathrm{wt}\,\mathbf{Y}$. The set $B(\Lambda)$ equipped with these maps $\tilde{e}_i$, $\tilde{f}_i$, ε_i, φ_i, and wt turns out to be the ''crystal'' (see Refs. 19 and 20) associated with the integrable highest weight module $V(\Lambda)$. In particular, for $i=0,1$ we have

$$\tilde{e}_i(B(\Lambda)) \subseteq B(\Lambda) \sqcup \{0\} \quad \text{and} \quad \tilde{f}_i(B(\Lambda)) \subseteq B(\Lambda) \sqcup \{0\}.$$

Moreover, $B(\Lambda)$ turns out to be the set of k-tuples of extended Young diagrams corresponding to the highest lifts of $\mathscr{P}(\Lambda)$ (Proposition 3.12 in Ref. 14).

The crystal $B(\Lambda)$ has the structure of an oriented colored graph with elements of $B(\Lambda)$ as the set of vertices and for $\mathbf{Y}, \mathbf{Y}' \in B(\Lambda), \mathbf{Y} \overset{i}{\to} \mathbf{Y}' (i=0,1)$ if and only if $\tilde{f}_i \mathbf{Y} = \mathbf{Y}'$. As a graph, $B(\Lambda)$ is connected. In particular, for $\mathbf{Y} \in B(\Lambda)$ we have (see Proposition 3.11 in 14)

$$\mathbf{Y} \neq \Phi \Rightarrow \tilde{e}_0 \mathbf{Y} \neq 0 \quad \text{or} \quad \tilde{e}_1 \mathbf{Y} \neq 0. \tag{4}$$

As a consequence of (3), we have

$$|Y_1| \geqslant \cdots \geqslant |Y_s| \quad \text{and} \quad |Y_{s+1}| \geqslant \cdots \geqslant |Y_k|.$$

Now we have the following simple but important observations.

Proposition 1: Let $\mathbf{Y} = (Y_1, \ldots, Y_s, Y_{s+1}, \ldots, Y_k) \in B(\Lambda)$ *and* $\mathbf{Y} \neq \Phi$. *Then* $|Y_1| \neq |Y_{s+1}|$.

Proof: Suppose $|Y_1| = |Y_{s+1}|$. Since Y_1 has charge 0 and Y_{s+1} has charge 1, the last nodes in the first rows of Y_1 and Y_{s+1} will have opposite color. Also by (3), the second rows of Y_1 and Y_{s+1} have at least one node less than their first rows. Consequently, we have two convex corners of opposite colors corresponding to the last nodes in the first rows of Y_1 and Y_{s+1}. One of these convex corners will contribute a 1 to the 0-signature of $\mathbf{Y}$ and the other will contribute a 1 to the 1-signature of $\mathbf{Y}$. However, in both cases this 1 will not be relevant as it will be preceded by at least one 0 in each signature. Therefore, by applying $\tilde{e}_0$ and $\tilde{e}_1$ successively we will obtain $\mathbf{Y}' = (Y_1', \ldots, Y_s', Y_{s+1}', \ldots, Y_k') \in B(\Lambda)$, such that

$$\mathbf{Y}' = \hat{e}_{i_m} \tilde{e}_{i_{m-1}} \cdots \tilde{e}_{i_1} \mathbf{Y}, \quad \text{for some } m \quad \text{and} \quad i_j \in \{0,1\},$$

$$\tilde{e}_0 \mathbf{Y}' = 0, \quad \tilde{e}_1 \mathbf{Y}' = 0,$$

$$|Y_1'| = |Y_1| = |Y_{s+1}| = |Y_{s+1}'|.$$

Since $\mathbf{Y} \neq \Phi$ and $|Y_1'| = |Y_1|$, we also have $\mathbf{Y}' \neq \Phi$. So by (4), either $\tilde{e}_0 \mathbf{Y}' \neq 0$ or $\tilde{e}_1 \mathbf{Y}' \neq 0$, which is a contradiction. Hence $|Y_1| \neq |Y_{s+1}|$. ■

Proposition 2: Let $\mathbf{Y} = (Y_1, \ldots, Y_s, Y_{s+1}, \ldots, Y_k) \in \mathscr{Y}(\Lambda)$. *Suppose that* $\tilde{f}_i^n \mathbf{Y} = \mathbf{Y}' = (Y_1', \ldots, Y_s', Y_{s+1}', \ldots, Y_k')$ *for some* $n > 0$ *and* $i = 0$ *or 1. Then* $|Y_j'| \leqslant |Y_j| + 1$ *for all* $1 \leqslant j \leqslant k$.

Proof: Let $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ be the i-signature of $\mathbf{Y}$ and $\varepsilon_J = (\varepsilon_{j_1}, \ldots, \varepsilon_{j_r})$ be the relevant part of ε. By definition of $\tilde{f}_i$ action (see rule C), $\tilde{f}_i^n \mathbf{Y} = \mathbf{Y}' \neq 0$ implies that $\varepsilon_{j_{r-n+1}} = \varepsilon_{j_{r-n+2}} = \cdots = \varepsilon_{j_r} = 0$. Also recall that each relevant 0 in ε_J corresponds to a unique concave i-corner in some Y_j, $1 \leqslant j \leqslant k$, and each application of $\tilde{f}_i$ changes only one relevant 0 to a relevant 1 without affecting the partition of ε. Hence the first row of each Y_j' can have at most one more i-color node than that of Y_j for $1 \leqslant j \leqslant k$. Therefore, the proposition follows. ■

III. CRYSTALS OF DEMAZURE MODULES

A. Demazure modules

The affine Lie algebra $\hat{\mathfrak{sl}}(2)$ has a triangular decomposition $\mathfrak{n}^- \oplus \mathfrak{h} \oplus \mathfrak{n}^+$, where $\mathfrak{h} = \mathrm{span}\{h_0, h_1, d\}$ is the Cartan subalgebra and $\mathfrak{n}^+$ (resp., $\mathfrak{n}^-$) denote the sum of the positive (resp., negative) root spaces (see Ref. 18). The subalgebra $\mathfrak{b} = \mathfrak{h} + \mathfrak{n}^+$ is the Borel subalgebra of $\hat{\mathfrak{sl}}(2)$. Let r_0, r_1 denote the simple reflections corresponding to the simple roots α_0, α_1, respectively. Recall that $r_i \mu = \mu - \mu(h_i)\alpha_i$ for all $\mu \in P$. Let W denote the Weyl group of $\hat{\mathfrak{sl}}(2)$ generated by r_0, r_1. For $w \in W$ let $l(w)$ denote the length of w and let $\prec$ denote the Bruhat order on W. For Λ

$=s\Lambda_0+t\Lambda_1\in P^+$, $k=s+t$, as before we consider the integrable highest weight $U_q(\widehat{\mathrm{sl}}(2))$-module $V(\Lambda)$. It is known that for $w\in W$, the extremal weight space $V(\Lambda)_{w\Lambda}$ is one dimensional. Let $E_w(\Lambda)$ denote the $U_q(\mathfrak{b})$-module generated by $V(\Lambda)_{w\Lambda}$. These modules $E_w(\Lambda)$, $w\in W$ are called the Demazure modules. They are finite-dimensional subspaces of $V(\Lambda)$ and have the following property: For $w,w'\in W, w\preceq w'$, we have $E_w(\Lambda)\subseteq E_{w'}(\Lambda)$, and $\cup_{w\in W}E_w(\Lambda)=V(\Lambda)$.

B. Demazure crystals

It is known that for $\widehat{\mathrm{sl}}(2)$, the Weyl group is

$$W=\{(r_1r_0)^m,r_0(r_1r_0)^m,(r_0r_1)^m,r_1(r_0r_1)^m|m\in\mathbb{Z}_{\geqslant 0}\}.$$

In particular, for each integer $L>0$, the Weyl group W has two distinct elements of length L:

$$w_L^+=\underbrace{\cdots r_0r_1r_0}_{L}\quad\text{and}\quad w_L^-=\underbrace{\cdots r_1r_0r_1}_{L}.$$

Let $W^+=\{1,w_L^+|L>0\}$ and $W^-=\{1,w_L^-|L>0\}$. Note that on each set W^+ and W^- the Bruhat order is a total order. Also note that $W^+\cup W^-=W$ and $W^+\cap W^-=\{1\}$. For $\Lambda=s\Lambda_0+t\Lambda_1\in P^+$, Kashiwara[10] has defined the crystals $B_w(\Lambda)$ for the Demazure modules $E_w(\Lambda)$ as a suitable subset of the crystal $B(\Lambda)$. For our purpose it suffices to recall the following recursive property of $B_w(\Lambda)$: If $r_iw<w$, then

$$B_w(\Lambda)=\{\tilde{f}_i^m b|m\geqslant 0,b\in B_{r_iw}(\Lambda),\tilde{e}_ib=0\}\backslash\{0\},\tag{5}$$

which can be used to construct $B_w(\Lambda)$ starting from the vacuum $\Phi\in B(\Lambda)$.

For example, if $\Lambda=2\Lambda_0$ and $w=r_1r_0$, the graph of the crystal $B_w(\Lambda)$ can be easily seen in Fig. 1.

Since $w_{L-1}^+<w_L^+, w_{L-1}^-<w_L^-, w_{L-1}^+<w_{L-1}^+r_1=w_L^-$ and $w_{L-1}^-<w_{L-1}^-r_0=w_L^+$, the following result is an immediate consequence of Proposition 3.2.4 in Ref. 10.

Proposition 3: For $L>0$ and $\Lambda\in P^+$ we have

(i) $B_{w_{L-1}^+}(\Lambda)\subseteq B_{w_L^+}(\Lambda)$,
(ii) $B_{w_{L-1}^-}(\Lambda)\subseteq B_{w_L^-}(\Lambda)$,
(iii) $B_{w_{L-1}^+}(\Lambda)\subseteq B_{w_L^-}(\Lambda)$,
(iv) $B_{w_{L-1}^-}(\Lambda)\subseteq B_{w_L^+}(\Lambda)$.

C. Realizations of $B_{w_L^+}(\Lambda)$ and $B_{w_L^-}(\Lambda)$

We define an extended Young diagram $Y=(y_j)_{j\geqslant 0}$ of charge γ ($=0$ or 1) to be *maximal of width L* if

$$|Y|=L\quad\text{and}\quad y_{j+1}=y_j+1\quad\text{for } 0\leqslant j\leqslant L-1.$$

We also define, for $\Lambda=s\Lambda_0+t\Lambda_1\in P^+$, $k=s+t$,

$$B_L(\Lambda)=\{\mathbf{Y}=(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)||Y_j|\leqslant L,1\leqslant j\leqslant k\}.$$

Noting that the weight multiplicities of $w_L^+\Lambda$ and $w_L^-\Lambda$ are one, we define extremal vectors as follows:

Definition 8 [extremal vectors]: $b_{w_L^+\Lambda}$ (resp., $b_{w_L^-\Lambda}$) is defined as the element of $B(\Lambda)$ that has the weight $w_L^+\Lambda$ (resp., $w_L^-\Lambda$).

The following theorem characterizes the extremal vectors.

Theorem 1: *Let $\Lambda=s\Lambda_0+t\Lambda_1$, $k=s+t$ and $L>0$. Then we have the following.*

(i) $b_{w_L^+\Lambda}=(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)$ *where $Y_1=\cdots=Y_s$ are maximal of width L and charge 0, and $Y_{s+1}=\cdots=Y_k$ are maximal of width $L-1$ and charge 1.*

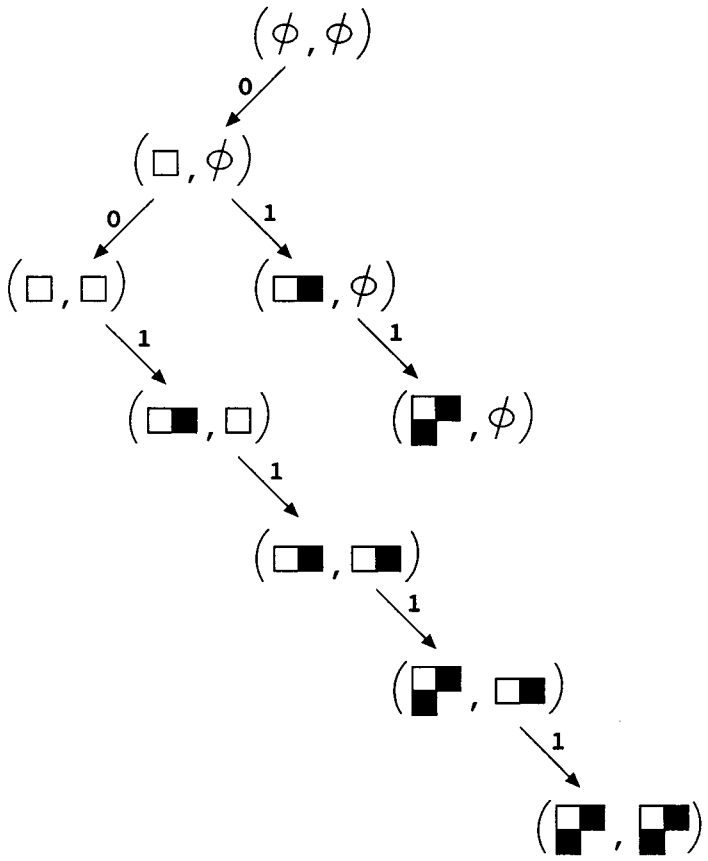

FIG. 1. The crystal graph of $B_{r_1 r_0}(2\Lambda_0)$.

(ii) $b_{w_L^- \Lambda} = (Y_1, \ldots, Y_s, Y_{s+1}, \ldots, Y_k) \in B(\Lambda)$ *where* $Y_1 = \cdots = Y_s$ *are maximal of width* $L-1$ *and charge 0, and* $Y_{s+1} = \cdots = Y_k$ *are maximal of width* L *and charge 1.*

Proof: We use induction on L. If $L=1$, then $w_L^+ = r_0$ and $w_L^+ \Lambda = r_0(s\Lambda_0 + t\Lambda_1) = \Lambda - s\alpha_0$. Also, it is easy to see that the 0-signature of the vacuum $\Phi \in B(\Lambda)$ is $\epsilon = (\epsilon_1, \ldots, \epsilon_s)$, where $\epsilon_j = 0$ for $1 \leqslant j \leqslant s$. Hence

$$b_{r_0 \Lambda} = \tilde{f}_0^s \Phi = (\square, \ldots, \square, \phi, \ldots, \phi) \in B(\Lambda),$$

and (i) holds for $L=1$. Now assume (i) for $L-1$. Observe that $b_{w_{L-1}^+ \Lambda}$ has convex and concave corners of opposite colors. Furthermore, $b_{w_{L-1}^+ \Lambda}$ has exactly $sL + t(L-1)$ numbers of concave i-corners where $L-1 \equiv i \pmod 2$. Also,

$$w_L^+ \Lambda = r_i(w_{L-1}^+ \Lambda) = w_{L-1}^+ \Lambda - (w_{L-1}^+ \Lambda)(h_i)\alpha_i = w_{L-1}^+ \Lambda - (sL + t(L-1))\alpha_i .$$

Since applying $\tilde{f}_i$ to $b_{w_{L-1}^+ \Lambda}(sL + t(L-1))$ times we add one i-color box to each row of each diagram in $b_{w_{L-1}^+ \Lambda}$, we have

$$b_{w_L^+ \Lambda} = \tilde{f}_i^{sL + t(L-1)} b_{w_{L-1}^+ \Lambda} = (Y_1, \ldots, Y_s, Y_{s+1}, \ldots, Y_k) \in B(\Lambda),$$

where $Y_1 = \cdots = Y_s$ are maximal of width L and $Y_{s+1} = \cdots = Y_k$ are maximal of width $L-1$. This proves (i). Proof of (ii) is similar to that of (i). ■

The following theorem characterizes the crystals of the Demazure modules.

Theorem 2: *Let* $\Lambda = s\Lambda_0 + t\Lambda_1$, $k = s+t$ *and* $L > 0$. *Then we have*

$$B_{w_L^+}(\Lambda)=\{(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)\,||Y_1|\leqslant L,|Y_{s+1}|\leqslant L-1\},$$

$$B_{w_L^-}(\Lambda)=\{(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)\,||Y_1|\leqslant L-1,|Y_{s+1}|\leqslant L\}.$$

Proof: Let

$$S_L=\{(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)\,||Y_1|\leqslant L,|Y_{s+1}|\leqslant L-1\},$$

$$T_L=\{(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in B(\Lambda)\,||Y_1|\leqslant L-1,|Y_{s+1}|\leqslant L\}.$$

We use induction on L. First we will show that $B_{w_L^+}(\Lambda)\subseteq S_L$ and $B_{w_L^-}(\Lambda)\subseteq T_L$. This is clear for $L=1$, since by (5), $B_{r_0}(\Lambda)=\{\tilde{f}_0^m\Phi|0\leqslant m\leqslant s\}$ and $B_{r_1}(\Lambda)=\{\tilde{f}_1^m\Phi|0\leqslant m\leqslant t\}$. Assume that $B_{w_{L-1}^+}(\Lambda)\subseteq S_{L-1}$ and $B_{w_{L-1}^-}(\Lambda)\subseteq T_{L-1}$. Since $w_L^+=r_i w_{L-1}^+$ where $L-1\equiv i(\mathrm{mod}\,2)$ by (5), we have

$$B_{w_L^+}(\Lambda)=\{\tilde{f}_i^m b|m\geqslant 0,b\in B_{w_{L-1}^+}(\Lambda),\tilde{e}_i b=0\}\backslash\{0\}.$$

Since $B_{w_{L-1}^+}(\Lambda)\subseteq S_{L-1}$, by Proposition 2 we have $B_{w_L^+}(\Lambda)\subseteq S_L$. Similarly $B_{w_L^-}(\Lambda)\subseteq T_L$.

Now we proceed to show that $S_L\subseteq B_{w_L^+}(\Lambda)$ and $T_L\subseteq B_{w_L^-}(\Lambda)$. As before, this is clear for $L=1$. Assume that $S_{L-1}\subseteq B_{w_{L-1}^+}(\Lambda)$ and $T_{L-1}\subseteq B_{w_{L-1}^-}(\Lambda)$. Let $\mathbf{Y}=(Y_1,\ldots,Y_s,Y_{s+1},\ldots,Y_k)\in S_L$. By Proposition 1, $|Y_1|\neq|Y_{s+1}|$. If $|Y_1|\leqslant L-1$ and $|Y_{s+1}|\leqslant L-2$, then $\mathbf{Y}\in S_{L-1}\subseteq B_{w_{L-1}^+}(\Lambda)\subseteq B_{w_L^+}(\Lambda)$ by (i) of Proposition 3. If $|Y_1|\leqslant L-2$ and $|Y_{s+1}|\leqslant L-1$, then $\mathbf{Y}\in T_{L-1}\subseteq B_{w_{L-1}^-}(\Lambda)\subseteq B_{w_L^+}(\Lambda)$ by (iv) of Proposition 3.

Finally, if $|Y_1|=L$ and $|Y_{s+1}|\leqslant L-1$, then Y_1 will have a convex i-corner corresponding to the last node of the first row of Y_1, where $L-1\equiv i(\mathrm{mod}\,2)$. Furthermore, this convex i-corner will contribute a relevant 1 to the i-signature of $\mathbf{Y}$. Therefore, $\tilde{e}_i\mathbf{Y}\neq 0$. Now, we can choose m such that $\tilde{e}_i^m\mathbf{Y}=\mathbf{Y}'=(Y_1',\ldots,Y_s',Y_{s+1}',\ldots,Y_k')\in B(\Lambda)$ and $\tilde{e}_i\mathbf{Y}'=0$. Then $|Y_1'|=L-1$, and $|Y_{s+1}'|\leqslant L-2$ by Proposition 1. Thus $\mathbf{Y}'\in S_{L-1}\subseteq B_{w_{L-1}^+}(\Lambda)$. Hence by (5), $\mathbf{Y}=\tilde{f}_i^m\mathbf{Y}'\in B_{r_i w_{L-1}^+}(\Lambda)=B_{w_L^+}(\Lambda)$. This proves that $S_L\subseteq B_{w_L^+}(\Lambda)$. Similarly $T_L\subseteq B_{w_L^-}(\Lambda)$ and the theorem holds. ■

The following result is an immediate consequence of Theorem 2.

Corollary 1: For $\Lambda=s\Lambda_0+t\Lambda_1$, we have

$$(i)\quad B_{w_L^+}(\Lambda)\cup B_{w_L^-}(\Lambda)=B_L(\Lambda),$$

$$(ii)\quad B_{w_L^+}(\Lambda)\cap B_{w_L^-}(\Lambda)=B_{L-1}(\Lambda).$$

Let us consider the extreme cases $s=0$ or $t=0$. Assume $\Lambda=k\Lambda_0$. Noting $w_L^-\Lambda=w_{L-1}^+\Lambda$, we have $B_{w_L^-}(\Lambda)=B_{w_{L-1}^+}(\Lambda)$. Similarly, if $\Lambda=k\Lambda_1$, we have $B_{w_L^+}(\Lambda)=B_{w_{L-1}^-}(\Lambda)$. Combining Proposition 3 and the previous corollary, we have the following.

Corollary 2:

$$B_L(\Lambda)=\begin{cases}B_{w_L^+}(\Lambda), & \text{if } \Lambda=k\Lambda_0,\\ B_{w_L^-}(\Lambda), & \text{if } \Lambda=k\Lambda_1.\end{cases}$$

IV. CHARACTERS

A. One-dimensional configuration sum

We summarize the one-dimensional (1-D) configuration sum that have already appeared in the study of exactly solvable SOS models.[15] It will be used in the evaluation of the character for $\mathscr{P}_L(\Lambda)$ in the following section.

Fixing a positive integer k, let us call a pair of integers (b,c) *weakly admissible* if $b-c=-k,-k+2,...,k$. For a weakly admissible pair (b,c) and $L\in\mathbb{Z}_{\geqslant 0}$, let $f_L^{(k)}(b,c)$ denote the unique solution to the following linear difference equation:

$$f_L^{(k)}(b,c)=\sum_d{}' \; f_{L-1}^{(k)}(d,b)q^{L|d-c|/4}, \tag{6}$$

$$f_0^{(k)}(b,c)=\delta_{b0}. \tag{7}$$

Here the sum Σ' is taken over d such that the pair (d,b) is weakly admissible. If (b,c) is not weakly admissible, $f_L^{(k)}(b,c)$ is set to be 0. Note that N,m in Ref. 15 are replaced with k,L, respectively. $f_L^{(k)}(b,c)$ enjoys the following properties: Reflection symmetry:

$$f_L^{(k)}(b,c)=f_L^{(k)}(-b,-c);$$

support property:

$$f_L^{(k)}(b,c)=0, \quad \text{unless} \quad |b|\leqslant Lk, |c|\leqslant (L+1)k \quad \text{and} \quad b\equiv Lk \bmod 2.$$

Given an integer μ, set

$$R_\mu=\{b\in\mathbb{Z}|(\mu-1)k\leqslant b\leqslant(\mu+1)k\},$$

where the left (resp., right) equality sign is taken if $\mu-1\leqslant 0$ (resp., $\mu+1\geqslant 0$). We also set $R_{\mu,\nu}=R_\mu\times R_\nu$. For a weakly admissible pair (b,c) and integer $L(\geqslant 1)$, let $\mu,\nu(=\mu\pm 1)$ be integers such that

$$(b,c)\in R_{\mu,\nu}, \quad \mu\equiv L+1, \nu\equiv L+2 \bmod 2.$$

The above equation uniquely determines μ and ν, except when $b=0$ $(\mu=\pm 1)$ or $c=0$ $(\nu=\pm 1)$, in which case either choice is allowed.

Let us recall standard notations in q analysis (see Ref. 21, for example). We set

$$(z;q)_m=\begin{cases}\displaystyle\prod_{j=1}^{m}(1-zq^{j-1}) & (m\geqslant 1),\\ 1 & (m=0).\end{cases}$$

If there is no danger of confusion, we abbreviate it as $(z)_m$. The *q-multinomial coefficient* is defined as

$$\begin{bmatrix} M \\ m_1\ldots m_n\end{bmatrix}_q=\begin{cases}\dfrac{(q)_M}{\Pi_{j=1}^n(q)_{m_j}}, & \text{if } m_1,\ldots,m_n\geqslant 0 \quad \text{and} \quad m_1+\cdots+m_n=M,\\ 0, & \text{otherwise.}\end{cases} \tag{8}$$

$\left[\begin{smallmatrix} M \\ m_1m_2\end{smallmatrix}\right]_q$ is often written as $\left[\begin{smallmatrix} M \\ m_1\end{smallmatrix}\right]_q$ and called the *Gaussian polynomial*. In both cases, if the subscript is omitted, we understand it as q.

For later use we extend the definition of the Gaussian polynomial $\left[\begin{smallmatrix} M \\ i\end{smallmatrix}\right]$ with $M\geqslant 0, i\in\mathbb{Z}$ to the case of $M,i\in\mathbb{Z}$, by setting

$$\begin{bmatrix} M \\ i\end{bmatrix}=\begin{cases}\dfrac{(q^{M-i+1})_i}{(q)_i}, & \text{if } i\geqslant 0,\\ 0, & \text{otherwise.}\end{cases} \tag{9}$$

Similarly, for $M<0$ we define $(z)_M$ to be $(q^Mz)_{-M}^{-1}$.

Lemma 2: For $M,N,n\in\mathbb{Z}$ we have

$$(z)_M=\sum_i (-z)^i q^{i(i-1)/2}\begin{bmatrix} M \\ i \end{bmatrix}, \tag{10}$$

$$\begin{bmatrix} M+N \\ n \end{bmatrix}=\sum_i q^{(n-i)(M-i)}\begin{bmatrix} M \\ i \end{bmatrix}\begin{bmatrix} N \\ n-i \end{bmatrix}, \tag{11}$$

where the sums are taken over all integers.

Proof: The first equation reduces to (3.3.6) or (3.3.7) in Ref. 21 depending on whether $M \geqslant 0$ or $M<0$.

Next, note that the following equation holds for $M,N\in\mathbb{Z}$:

$$(z)_{M+N}=(z)_M(q^M z)_N. \tag{12}$$

The second equation can be proved by using (10) on both sides of (12) and comparing the coefficients of z^n. ■

We call the following expression bosonic.

Proposition 4 (Theorem 4.1.1 in Ref. 15): For any weakly admissible pair $(b,c)\in R_{\mu,\nu}$ and integers $L,k\geqslant 1$,

$$(q)_{L-1}f_L^{(k)}(b,c)=\left(\sum_{\substack{i>(L+\nu)/2\\ j\leqslant(L+\mu-1)/2}}-\sum_{\substack{i<(L+\nu)/2-1\\ j\geqslant(L+\mu+1)/2}}\right)(-1)^{i+j}q^{Q_{i,j}^{(L)}(b,c)}\begin{bmatrix} L-1 \\ i \end{bmatrix}\begin{bmatrix} L \\ j \end{bmatrix}, \tag{13}$$

$$Q_{i,j}^{(L)}(b,c)=\frac{1}{2}(i-j)(i-j+1)-\left(i-\frac{L-1}{2}\right)\left(j-\frac{L}{2}\right)k+\frac{b}{2}\left(i-\frac{L-1}{2}\right)+\frac{c}{2}\left(j-\frac{L}{2}\right).$$

B. Fermionic expression

In this section we rewrite the bosonic expression to the fermionic expression, which involves q multinomials. At the first step we show the following.

Proposition 5: For any weakly admissible pair $(b,c)\in R_{\mu,\nu}$ satisfying $b\geqslant -Lk$ and integers $L,k\geqslant 1$,

$$\begin{aligned} q^{[L(L-1)k+(L-1)b+Lc]/4}f_L^{(k)}(b,c)=&\sum_{j\in\mathbb{Z}}(-1)^j q^{(1/2)j(j-1)+[(L-1)/2]jk+(c/2)j}\begin{bmatrix} L \\ j \end{bmatrix}\\ &\times\begin{bmatrix} L-j-1-\left(j-\frac{L}{2}\right)k+\frac{b}{2} \\ L-1 \end{bmatrix}\\ &-\sum_{j\in\mathbb{Z}}(-1)^j q^{(1/2)j(j+1)+(L/2)jk+(b/2)j}\begin{bmatrix} L-1 \\ j \end{bmatrix}\\ &\times\begin{bmatrix} L-j-2-\left(j-\frac{L-1}{2}\right)k+\frac{c}{2} \\ L-1 \end{bmatrix}(1-q^{L-j-1-[j-(L-1)/2]k+c/2}). \end{aligned}$$

Proof: Rewrite the sum

$$\left(\sum_{\substack{i>(L+\nu)/2\\ j\leqslant(L+\mu-1)/2}}-\sum_{\substack{i<(L+\nu)/2-1\\ j\geqslant(L+\mu+1)/2}}\right)$$

in (13) as

$$\left(\sum_{\substack{i\in\mathbb{Z}\\ j\leqslant(L+\mu-1)/2}}-\sum_{\substack{i\leqslant(L+\nu)/2-1\\ j\in\mathbb{Z}}}\right).$$

Applying (10) for the sum over $i \in \mathbb{Z}$ or $j \in \mathbb{Z}$, we obtain

$$q^{[L(L-1)k+(L-1)b+Lc]/4} f_L^{(k)}(b,c) = \sum_{j \leqslant (L+\mu-1)/2} (-1)^j q^{(1/2)j(j-1)+[(L-1)/2]jk+(c/2)j} \times \begin{bmatrix} L \\ j \end{bmatrix} \frac{(q^{-j+1-(j-L/2)k+b/2})_{L-1}}{(q)_{L-1}} - \sum_{i \leqslant (L+\nu)/2-1} (-1)^i q^{(1/2)i(i+1)+(L/2)ik+(b/2)i} \times \begin{bmatrix} L-1 \\ i \end{bmatrix} \frac{(q^{-i-[i-(L-1)/2]k+c/2})_L}{(q)_{L-1}}.$$

Recall that we have assumed $b \geqslant -Lk$. Noticing $(\mu-1)k \leqslant b \leqslant (\mu+1)k$, $(\nu-1)k \leqslant c \leqslant (\nu+1)k$ with $\mu \geqslant 1-L$, $\nu \geqslant -L$, we get $L-j-1-(j-L/2)k+b/2 < m-1$ and $L-i-2-[i-(L-1)/2]k+c/2 < m-1$ if $j > (L+\mu-1)/2$ and $i > (L+\nu)/2-1$. Therefore, recalling the definition (9), we have

$$\begin{bmatrix} L-j-1-\left(j-\frac{L}{2}\right)k+\frac{b}{2} \\ L-1 \end{bmatrix} = \begin{cases} \dfrac{(q^{-j+1-(j-L/2)k+b/2})_{L-1}}{(q)_{L-1}}, & \text{if } j \leqslant \dfrac{L+\mu-1}{2} \\ 0, & \text{otherwise,} \end{cases}$$

$$\begin{bmatrix} L-i-2-\left(i-\frac{L-1}{2}\right)k+\frac{c}{2} \\ L-1 \end{bmatrix} = \begin{cases} \dfrac{(q^{-i-[i-(L-1)/2]k+c/2})_{L-1}}{(q)_{L-1}}, & \text{if } i \leqslant \dfrac{L+\nu}{2}-1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus we arrive at the desired result. ■

Proposition 6: For any weakly admissible pair $(b,c) \in R_{\mu,\nu}$ *satisfying* $b \geqslant 0$, $c \neq b+k$ *and integers* $L \geqslant 0$, $k \geqslant 2$, *we have*

$$f_L^{(k)}(b,c) = \sum_{0 \leqslant i \leqslant L} q^{L(L-1)/4-[(k-1)i^2+(2L+b+c-1)i]/4} \times \begin{bmatrix} L \\ i \end{bmatrix} f_{L-i}^{(k-1)}(b+(k+1)i-L, c+(k+1)i-L+1).$$

Proof: If $L=0$, the formula is shown easily from the initial condition (7). We assume $L \geqslant 1$.

Notice that from the assumption $f_{L-i}^{(k-1)}(b+(k+1)i-L, c+(k+1)i-L+1)=0$ when $i=L$. Since the right-hand side of the formula in Proposition 5 is formally zero when $L=0$, we can use it for convenience of our proof. Check that the assumption allows us to use Proposition 5. Setting

$$a_L^{(k)}(b,c;j) = q^{-[L(L-1)k+(L-1)b+Lc]/4+[(L-1)/2]jk+(c/2)j} \begin{bmatrix} L \\ j \end{bmatrix} \begin{bmatrix} L-j-1-\left(j-\frac{L}{2}\right)k+\frac{b}{2} \\ L-1 \end{bmatrix},$$

$$b_L^{(k)}(b,c;j) = q^{-[L(L-1)k+(L-1)b+Lc]/4+(L/2)jk+(b/2)j} \begin{bmatrix} L-1 \\ j \end{bmatrix} \begin{bmatrix} L-j-2-\left(j-\frac{L-1}{2}\right)k+\frac{c}{2} \\ L-1 \end{bmatrix} \times (1-q^{L-j-1-[j-(L-1)/2]k+c/2}),$$

we have to show

$$\sum_{j\in\mathbb{Z}}(-1)^j(q^{(1/2)j(j-1)}a_L^{(k)}(b,c;j)-q^{(1/2)j(j+1)}b_L^{(k)}(b,c;j))$$

$$=\sum_{i=0}^{L}q^{L(L-1)/4-[(k-1)i^2+(2L+b+c-1)i]/4}\begin{bmatrix}L\\ i\end{bmatrix}$$

$$\times\sum_{j\in\mathbb{Z}}(-1)^j(q^{(1/2)j(j-1)}a_{L-i}^{(k-1)}(b',c';j)-q^{(1/2)j(j+1)}b_{L-i}^{(k-1)}(b',c';j)),$$

where $b'=b+(k+1)i-L$, $c'=c+(k+1)i-L+1$. In fact, we can show that the recurrence relation holds term by term, that is,

$$a_L^{(k)}(b,c;j)=\sum_{i=0}^{L}q^{L(L-1)/4-[(k-1)i^2+(2L+b+c-1)i]/4}\begin{bmatrix}L\\ i\end{bmatrix}$$

$$\times a_{L-i}^{(k-1)}(b+(k+1)i-L,c+(k+1)i-L+1;j), \tag{14}$$

and the same for $b_L^{(k)}(b,c;j)$. After some calculation, (14) reduces to

$$\begin{bmatrix}L\\ j\end{bmatrix}\begin{bmatrix}L-j-1-\left(j-\frac{L}{2}\right)k+\frac{b}{2}\\ L-1\end{bmatrix}=\sum_i q^{(L-i-1)(L-i-j)}\begin{bmatrix}L\\ i\end{bmatrix}\begin{bmatrix}L-i\\ j\end{bmatrix}\begin{bmatrix}-1-\left(j-\frac{L}{2}\right)k+\frac{b}{2}\\ L-i-1\end{bmatrix}.$$

Noting that

$$\begin{bmatrix}L\\ i\end{bmatrix}\begin{bmatrix}L-i\\ j\end{bmatrix}=\begin{bmatrix}L\\ j\end{bmatrix}\begin{bmatrix}L-j\\ i\end{bmatrix},$$

we can cancel the factor $\left[\begin{smallmatrix}L\\ j\end{smallmatrix}\right]$. Thus we arrive at (11) with $M=L-j$, $N=-1-(j-L/2)k+b/2$, $n=L-1$.

Similarly, (14) with $a_L^{(k)}(b,c;j)$ replaced by $b_L^{(k)}(b,c;j)$ reduces to

$$\begin{bmatrix}L-1\\ j\end{bmatrix}\begin{bmatrix}L-j-2-\left(j-\frac{L-1}{2}\right)k+\frac{c}{2}\\ L-1\end{bmatrix}(1-q^{L-j-1-[j-(L-1)/2]k+c/2})$$

$$=\sum_i q^{(L-i)(L-i-j-1)}\begin{bmatrix}L\\ i\end{bmatrix}\begin{bmatrix}L-i-1\\ j\end{bmatrix}\begin{bmatrix}-1-\left(j-\frac{L-1}{2}\right)k+\frac{c}{2}\\ L-i-1\end{bmatrix}$$

$$\times(1-q^{-[j-(L-1)/2]k+c/2}). \tag{15}$$

Set $N=-[j-(L-1)/2]k+c/2$. Noting that

$$\begin{bmatrix}L-1\\ j\end{bmatrix}\begin{bmatrix}L-j+N-2\\ L-1\end{bmatrix}(1-q^{L-j+N-1})=\begin{bmatrix}L\\ j\end{bmatrix}\begin{bmatrix}L-j+N-1\\ L\end{bmatrix}(1-q^{L-j}),$$

$$\begin{bmatrix}L\\ i\end{bmatrix}\begin{bmatrix}L-i-1\\ j\end{bmatrix}\begin{bmatrix}N-1\\ L-i-1\end{bmatrix}(1-q^N)=\begin{bmatrix}L\\ j\end{bmatrix}\begin{bmatrix}N\\ L-i\end{bmatrix}\begin{bmatrix}L-j-1\\ i\end{bmatrix}(1-q^{L-j}),$$

and cancelling the factor $\left[\begin{smallmatrix}L\\ j\end{smallmatrix}\right](1-q^{L-j})$, (15) reduces again to (11) with $M=L-j-1$, $n=L$. ■

This recurrence relation with respect to k leads us to an expression of $f_L^{(k)}(b,c)$ in terms of the q-multinomial coefficient (8). The next theorem is a generalization of Theorem 4.2.5 in Ref. 15.

Theorem 3: *For any weakly admissible pair $(b,c)\in R_{\mu,\nu}$ and integers $L\geqslant 0$, $k\geqslant 1$, we have the following expression for $f_L^{(k)}(b,c)$:*

$$f_L^{(k)}(b,c)=\sum q^{\mathcal{C}\left[\begin{smallmatrix}L\\ x_0\cdots x_k\end{smallmatrix}\right]}, \tag{16}$$

$$\mathcal{Q}=\frac{L^2k-L(b-c+k)+b}{4}-\sum_{0\leqslant a\leqslant a'\leqslant k}(a'-a)x_a x_{a'}+\sum_{(c-b+k)/2\leqslant a\leqslant k}\left(a-\frac{c-b+k}{2}\right)x_a .$$

The sum in (16) is taken over all non-negative integers $x_0,\ldots,x_k$, satisfying

$$\sum_{a=0}^{k}x_a=L,\quad \sum_{a=0}^{k}ax_a=\frac{Lk-b}{2}.$$

Proof: We prove by induction on k. If $k=1$, (16) reduces to

$$f_L^{(1)}(b,c)=q^{bc/4}\begin{bmatrix} L \\ \frac{L+b}{2}\ \frac{L-b}{2}\end{bmatrix}.$$

This trivially satisfies (7). Checking (6) reduces to the following identity of Gaussian polynomials:

$$\begin{bmatrix} M \\ i\end{bmatrix}=\begin{bmatrix} M-1 \\ i-1\end{bmatrix}+q^i\begin{bmatrix} M-1 \\ i\end{bmatrix}=q^{M-i}\begin{bmatrix} M-1 \\ i-1\end{bmatrix}+\begin{bmatrix} M-1 \\ i\end{bmatrix}.$$

Thus, the theorem is correct when $k=1$.

Now assume $k\geqslant 2$. If $c=b+k$, (16) turns out to be equivalent to Theorem 4.2.5 [also see Eq. (4.2.1)] in Ref. 15. Therefore we can assume $c\neq b+k$. First suppose $b\geqslant 0$. From Proposition 6 and the assumption of the induction, we obtain

$$f_L^{(k)}(b,c)=\sum_{x_k=0}^{L}\sum q^{\mathcal{Q}}\begin{bmatrix} L \\ x_k\end{bmatrix}\begin{bmatrix} L-x_k \\ x_0\cdots x_{k-1}\end{bmatrix}, \tag{17}$$

with

$$\mathcal{Q}=\frac{L(L-1)}{4}-\frac{(k-1)x_k^2+(2L+b+c-1)x_k}{4}+\mathcal{Q}',$$

$$\mathcal{Q}'=\frac{(L-x_k)^2(k-1)-(L-x_k)(b'-c'+k-1)+b'}{4}-\sum_{0\leqslant a<a'\leqslant k-1}(a'-a)x_a x_{a'}$$
$$+\sum_{(c'-b'+k-1)/2\leqslant a\leqslant k-1}\left(a-\frac{c'-b'+k-1}{2}\right)x_a ,$$

$$b'=b+(k+1)x_k-L,\quad c'=c+(k+1)x_k-L+1,$$

and the inner sum in (17) is taken over non-negative integers $x_0,\ldots,x_{k-1}$ such that $\Sigma_{a=0}^{k-1}x_a=L-x_k$, $\Sigma_{a=0}^{k-1}ax_a=[(L-x_k)(k-1)-b']/2$. Calculating the power of q, and noting that

$$\begin{bmatrix} L \\ x_k\end{bmatrix}\begin{bmatrix} L-x_k \\ x_0\cdots x_{k-1}\end{bmatrix}=\begin{bmatrix} L \\ x_0\cdots x_k\end{bmatrix},$$

we arrive at (16).

For the proof of the case $b<0$, it suffices to check Reflection symmetry on the right-hand side of (16). The check is easily done by the variable change $x_j\to x_{k-j}$ $(0\leqslant j\leqslant k)$ in the summand of (16). ■

C. Character of $\mathcal{P}_L(\Lambda)$

We define the character for $\mathcal{P}_L(\Lambda)$.

Definition 9 [character of $\mathcal{P}_L(\Lambda)$]:

$$\mathrm{ch}_L(\Lambda)=\mathrm{ch}_L(\Lambda)(z,q)=\sum_{p\in\mathscr{P}_L(\Lambda)} z^{(\Lambda-p_0|\Lambda_1)}q^{E(p)}.$$

Then we have the following expression for $\mathrm{ch}_L(\Lambda)$.

Proposition 7:

$$\mathrm{ch}_L(\Lambda)=\sum_{j\in\mathbb{Z}} z^{-j}q^{j/2}f_L^{(k)}(\epsilon^{(L)}(s-t)-2j,\epsilon^{(L+1)}(s-t)-2j).$$

Here $\epsilon^{(L)}$ is given in (2).

Proof: First, for $a\Lambda_0+b\Lambda_1\in P_k$, we set $\nu(a\Lambda_0+b\Lambda_1)=b$. Then we have

$$\mu(\iota(p)_j)=\frac{\nu(p_{j+1})-\nu(p_j)+k}{2}.$$

From the definition of H (Definition 5), we have

$$H(\iota(p)_{j-1},\iota(p)_j)=\frac{k}{2}+\frac{|\nu(p_{j-1})-\nu(p_{j+1})|}{4}+\frac{\nu(p_{j-1})-2\nu(p_j)+\nu(p_{j+1})}{4},$$

and therefore

$$\sum_{j=1}^{L} jH(\iota(p)_{j-1},\iota(p)_j)=\frac{k}{2}\frac{L(L+1)}{2}+\sum_{j=1}^{L} j\,\frac{|\nu(p_{j-1})-\nu(p_{j+1})|}{4}+\frac{\nu(p_0)-(L+1)\nu(p_L)+L\nu(p_{L+1})}{4}.$$

Thus, we have

$$E(p)=\frac{\nu(p_0)-\nu(\bar{p_0})}{4}+\sum_{j=1}^{L} j\,\frac{|\nu(p_{j-1})-\nu(p_j)|}{4}.$$

Using this and parametrizing p_0 by $\Lambda+j\alpha_1$ $(j\in\mathbb{Z})$, we can rewrite the character (Definition 9) as

$$\mathrm{ch}_L(\Lambda)=\sum_{j\in\mathbb{Z}} z^{-j}q^{(\nu(p_0)-\nu(\bar{p_0}))/4}K_{L,j},$$

$$K_{L,j}=\sum_{\substack{p\in\mathscr{P}_L(\Lambda)\\ p_0=\Lambda+j\alpha_1}} q^{\sum_{J=1}^{L} j|\nu(p_{j-1})-\nu(p_{J+1})|/4}.$$

Noting that $K_{L,j}$ satisfies the same difference equation as $f_L^{(k)}(\nu(p_L)-\nu(p_0),\nu(p_{L+1})-\nu(p_0))$, we have

$$\mathrm{ch}_L(\Lambda)=\sum_{j\in\mathbb{Z}} z^{-j}q^{(\nu(p_0)-\nu(\bar{p_0}))/4}f_L^{(k)}(\nu(p_L)-\nu(p_0),\nu(p_{L+1})-\nu(p_0)),$$

with $p_0=\Lambda+j\alpha_1$. Evaluating $\nu(p_j)$ $(j=0,L,L+1)$ explicitly, we have the desired result. ■

Now we have the final result on the character of $\mathscr{P}_L(\Lambda)$.

Theorem 4: *Let $\Lambda=s\Lambda_0+t\Lambda_1$ $(s+t=k)$, C be the Cartan matrix of $sl(k)$, that is, $C_{ij}=2\delta_{ij}-\delta_{ij+1}-\delta_{ij-1}$ $(1\leqslant i,j\leqslant k-1)$. Set $x^t=(x_1,\ldots,x_{k-1})$ and let e_i be the $k-1$-dimensional ith unit vector $(e_0=e_k=0)$. Define $F_{j,L}(\Lambda)$ by*

$$\mathrm{ch}_L(\Lambda)=\sum_{j\in\mathbb{Z}} z^{-j}F_{j,L}(\Lambda).$$

Then we have the following expression for $F_{j,L}(\Lambda)$. *If* L *is even,*

$$F_{j,L}(\Lambda)=\sum_{\Sigma_{i=1}^{k-1} i x_i \in j+k\mathbb{Z}} q^{x^t C^{-1} x - x^t C^{-1} e_s + j(j+t)/k} \begin{bmatrix} L \\ x_0 \cdots x_k \end{bmatrix},$$

where

$$x_0=\frac{L}{2}-\frac{1}{k}\left(\sum_{i=1}^{k-1}(k-i)x_i+j\right), \quad x_k=\frac{L}{2}-\frac{1}{k}\left(\sum_{i=1}^{k-1} i x_i - j\right).$$

If L *is odd,*

$$F_{j,L}(\Lambda)=\sum_{\Sigma_{i=1}^{k-1} i x_i \in j+t+k\mathbb{Z}} q^{x^t C^{-1} x - x^t C^{-1} e_t + J(j+t)/k} \begin{bmatrix} L \\ x_0 \cdots x_k \end{bmatrix},$$

where

$$x_0=\frac{L+1}{2}-\frac{1}{k}\left(\sum_{i=1}^{k-1}(k-i)x_i+j+t\right), \quad x_k=\frac{L-1}{2}-\frac{1}{k}\left(\sum_{i=1}^{k-1} i x_i - j - t\right).$$

Proof: Suppose L is even, from Theorem 3 and Proposition 7; we obtain

$$F_{j,L}(\Lambda)=\sum_{\Sigma_{i=0}^{k} i x_i = Lk/2+j} q^{\frac{L^2 k}{4}-\frac{Lt}{2}-\Sigma_{0\leqslant i\leqslant i'\leqslant k}(i'-i)x_i x_{i'}+\Sigma_{s\leqslant i\leqslant k}(i-s)x_i} \begin{bmatrix} L \\ x_0 \cdots x_k \end{bmatrix}.$$

Solving $\Sigma_{i=0}^{k} x_i = L$ and $\Sigma_{i=0}^{k} i x_i = Lk/2+j$ for x_0 and x_k, we get $x_0=L/2-(1/k)(\Sigma_{i=1}^{k-1}(k-i)x_i+j)$, $x_k=L/2-(1/k)(\Sigma_{i=1}^{k-1} i x_i - j)$. Substitute these expressions into the power of q, and note that $(C^{-1})_{ij}=i(k-j)/k$ (for $i\leqslant j$), $j(k-i)/k$ (for $i>j$). Noticing that the sum restriction changes into $\Sigma_{i=1}^{k-1} i x_i \in j+k\mathbb{Z}$, we obtain the desired expression.

The case of L odd is similar. ■

D. Character of Demazure module

As we have seen in Sec. III, crystals of Demazure modules can be realized as subsets of the crystal $B(\Lambda)$ or equivalently $\mathscr{P}(\Lambda)$. Now it is easy to see that the characters of Demazure modules admit the following definition.

Definition 10:

$$\mathrm{ch}_L^{\pm}(\Lambda)=\mathrm{ch}_L^{\pm}(\Lambda)(z,q)=\sum_{p\in B_{w_L^{\pm}}(\Lambda)} z^{(\Lambda - p_0|\Lambda_1)} q^{E(p)}.$$

A natural question to ask is the following: Is it possible to express $\mathrm{ch}_L^{\pm}(\Lambda)$ in terms of $\mathrm{ch}_L(\Lambda)$? The answer is given by the following theorem.

Theorem 5: Let $\Lambda=s\Lambda_0+t\Lambda_1$, $s+t=k$. *We have*

$$\mathrm{ch}_L^{+}(\Lambda)=\sum_{i=0}^{s} z^{-\epsilon^{(L)}(s-i)} q^{[(L+\epsilon^{(L)})/2](s-i)} \mathrm{ch}_{L-1}(\Lambda^{(i)}),$$

$$\mathrm{ch}_L^{-}(\Lambda)=\sum_{i=0}^{t} z^{\epsilon^{(L)}(t-i)} q^{[(L-\epsilon^{(L)})/2](t-i)} \mathrm{ch}_{L-1}(\Lambda^{(k-i)}).$$

Here $\epsilon^{(L)}$ *is given in (2) and* $\Lambda^{(i)}=i\Lambda_0+(k-i)\Lambda_1$.

Proof: First we prove the expression for $\mathrm{ch}_L^{+}(\Lambda)$. Let us suppose L is even. Interpreting Theorem 2 in terms of paths, for any $p\in B_{w_L^{+}}(\Lambda)$ we have $\iota(p)_{L-1}=(k-i)\epsilon_0+i\epsilon_1$, where $i=0,1,\ldots,s$. For each i consider the following subset of $\mathscr{P}_L(\Lambda)$:

$$S^{(i)}=\{p\in\mathscr{P}_L(\Lambda)|p_{L-1}=\Lambda-((k-i)\hat{0}+i\hat{1})\}.$$

Noting that $\Lambda-((k-i)\hat{0}+i\hat{1})-\Lambda^{(k-i)}=\Lambda-\Lambda^{(i)}$, we see that

$$\{(p_j-(\Lambda-\Lambda^{(i)}))_{j=0,\ldots,L}|p\in S^{(i)}\}$$

can be identified with $\mathscr{P}_{L-1}(\Lambda^{(i)})$. Defining $\overline{p^{(i)}}$ as the ground-state path of $\mathscr{P}_{L-1}(\Lambda^{(i)})$, we have

$$\mathrm{ch}_{L-1}(\Lambda^{(i)})=\sum_{\substack{p\in B_{w_L^+}(\Lambda)\\ p_{L-1}=\Lambda-((k-i)\hat{0}+i\hat{1})}} z^{(\Lambda^{(i)}-(p_0-(\Lambda-\Lambda^{(i)}))|\Lambda_1)} \times q^{\sum_{j=1}^{L-1} j(H(\iota(p)_{j-1},\iota(p)_j)-H(\iota(\overline{p^{(i)}})_{j-1},\iota(\overline{p^{(i)}})_j))}.$$

Therefore,

$$\mathrm{ch}_L^+(\Lambda)=\sum_{i=0}^{s} q^{\sum_{j=1}^{L-1} j(H(\iota(\overline{p^{(i)}})_{J-1},\iota(\overline{p^{(i)}})_J)-H(\iota(\overline{p})_{j-1},\iota(\overline{p})_j))} \times q^{L(H(\iota_{L-1}^{(i)},\iota_L^{(i)})-H(\iota_{L-1},\iota_L))}\,\mathrm{ch}_{L-1}(\Lambda^{(i)}).$$

Here $\iota_{L-1}^{(i)}=(k-i)\epsilon_0+i\epsilon_1$, $\iota_{L-1}=\iota_{L-1}^{(s)}$, $\iota_L^{(i)}=\iota_L=s\epsilon_0+t\epsilon_1$. Using Lemma 1 we get the result. The case of L odd is similar.

For $\mathrm{ch}_L^-(\Lambda)$ we utilize the Dynkin diagram symmetry:

$$\mathrm{ch}_L^-(\Lambda)=\mathrm{ch}_L^+(t\Lambda_0+s\Lambda_1)(qz^{-1},q)$$
$$=\sum_{i=0}^{t}(qz^{-1})^{-\epsilon^{(L)}(t-i)}q^{[(L+\epsilon^{(L)})/2](t-i)}\,\mathrm{ch}_{L-1}(\Lambda^{(i)})(qz^{-1},q).$$

Noting that $\mathrm{ch}_{L-1}(\Lambda^{(i)})(qz^{-1},q)=\mathrm{ch}_{L-1}(\Lambda^{(k-i)})(z,q)$, we get the result. ■

E. Some specializations

In this section, we consider two specializations of Demazure characters, and compare with Sanderson's results.[6]

First we consider ''real characters.'' These are obtained by setting $e^{-\alpha_0}=q$, $e^{-\alpha_1}=q^{-1}$. We have the following.

Corollary 3: Let $\Lambda=s\Lambda_0+t\Lambda_1$, $s+t=k$:

$$\mathrm{ch}_L^+(\Lambda)(q^{-1},1)=q^{-(L-\epsilon^{(L)})k/2}[s+1][k+1]^{L-1}.$$

Here $[a]=(1-q^a)/(1-q)$.

Proof: From Theorem 4 we easily get

$$f_L^{(k)}(b,c;1)=\sum_{\sum_{i=0}^{k} ix_i=(Lk-b)/2}\binom{L}{x_0\cdots x_k}.$$

Here $\binom{L}{x_0\cdots x_k}$ denotes the usual multinomial coefficient. Assume L is even. From Proposition 7 we have

$$\begin{aligned}\mathrm{ch}_L(\Lambda)(q^{-1},1)&=\sum_{j\in\mathbb{Z}} q^j \sum_{\Sigma_{i=0}^k i x_i = Lk/2+j} \begin{pmatrix} L \\ x_0\cdots x_k \end{pmatrix} \\ &= q^{-Lk/2} \sum_{x_0,\ldots,x_k} q^{\Sigma_{i=0}^k i x_i} \begin{pmatrix} L \\ x_0\cdots x_k \end{pmatrix} \\ &= q^{-Lk/2}(1+q+\ldots+q^k)^L \\ &= q^{-Lk/2}[k+1]^L.\end{aligned}$$

Similarly, for L odd we have

$$\mathrm{ch}_L(\Lambda)(q^{-1},1)=q^{-(Lk-s+t)/2}[k+1]^L.$$

Using Theorem 5, we obtain the desired result. ■

Next, we consider ''principal characters'' obtained by setting $e^{-\alpha_0}=e^{-\alpha_1}=q$. In considering this case, we restrict ourselves to the extreme case $t=0$.

Corollary 4: Let $\Lambda=k\Lambda_0$:

$$\mathrm{ch}_L^+(\Lambda)(q,q^2)=\sum_{x_0,\ldots,x_k} q^{2x^tC^{-1}x+(k/2)\mathscr{S}(\mathscr{S}+1)} \begin{bmatrix} L \\ x_0\cdots x_k \end{bmatrix}_{q^2},$$

where we have set

$$\mathscr{S}=\frac{1}{k}\sum_{i=0}^{k}(k-2i)x_i. \tag{18}$$

Proof: Note from Corollary 2 that when $\Lambda=k\Lambda_0$ we have $\mathrm{ch}_L^+(\Lambda)=\mathrm{ch}_L(\Lambda)$. Using Theorem 4 we obtain

$$\mathrm{ch}_L^+(\Lambda)(q,q^2)=\sum_{j\in\mathbb{Z}} q^{-j} \sum_{\Sigma_{i=1}^{k-1} i x_i \in j+k\mathbb{Z}} q^{2x^tC^{-1}x+2j^2/k} \begin{bmatrix} L \\ x_0\cdots x_k \end{bmatrix}_{q^2}, \tag{19}$$

where

$$x_0=\left[\frac{L+1}{2}\right]-\frac{1}{k}\left(\sum_{i=1}^{k-1}(k-i)x_i+j\right), \quad x_k=\left[\frac{L}{2}\right]-\frac{1}{k}\left(\sum_{i=1}^{k-1} i x_i-j\right), \tag{20}$$

with [ ] being the Gauss symbol: $[x]$ is the largest integer that does not exceed x. Since the condition $\Sigma_{i=1}^{k-1} i x_i \in j+k\mathbb{Z}$ is equivalent to the condition that all x_k are integers, we can rewrite the expression in (19) as

$$\sum_{x_0,\ldots,x_k} q^{2x^tC^{-1}x+2j^2/k-j} \begin{bmatrix} L \\ x_0\cdots x_k \end{bmatrix}_{q^2}.$$

Solving (20) with respect to j, we get

$$j=\frac{k}{2}\left(\left[\frac{L+1}{2}\right]-\left[\frac{L}{2}\right]-\mathscr{S}\right).$$

Noting that $[(L+1)/2]-[L/2]=\epsilon^{(L)}$ and substituting this into the expression, we arrive at

$$\sum_{x_0,\ldots,x_k} q^{2x^tC^{-1}x+(k/2)\mathscr{S}(\mathscr{S}+1-2\epsilon^{(L)})} \begin{bmatrix} L \\ x_0\cdots x_k \end{bmatrix}_{q^2}.$$

If L is odd, we can make the variable change $x_i\to x_{k-i}$, which amounts to $\mathscr{S}\to-\mathscr{S}$. Thus we get the desired result. ■

In Ref. 6, Sanderson has evaluated the same character involving the q-multinomial coefficient with argument q, as opposed to our expression with q^2. Accordingly, we get the following new q-multinomial identity:

$$\sum_{x_0,\ldots,x_k} q^{2x^tC^{-1}x+(k/2)\mathscr{S}(\mathscr{S}+1)} \begin{bmatrix} L \\ x_0 \cdots x_k \end{bmatrix}_{q^2}$$
$$= \sum_{0\leqslant i_1\leqslant \ldots \leqslant i_k\leqslant L} q^{i_1(i_1+1)/2+\ldots+i_k(i_k+1)/2} \begin{bmatrix} L \\ L-i_ki_k-i_{k-1}\cdots i_2-i_1i_1 \end{bmatrix}_q, \tag{21}$$

where $\mathscr{S}$ is defined in (18). As pointed out by Andrews, changing q to q^{-1} in (21) for $k=1$ reduces to

$$\sum_{x_1=0}^{L} q^{x_1} \begin{bmatrix} L \\ x_1 \end{bmatrix}_{q^2} = \prod_{j=1}^{L} (1+q^j), \tag{22}$$

which is a known classical identity due to Watson. Thus (21) can be thought of as the higher level generalization of Watson's identity (22).

V. DISCUSSION

We characterized all Demazure crystals associated with $\widehat{\mathrm{sl}}(2)$ in terms of paths. We then obtained explicit expressions for their full characters. Our results extend Sanderson's results, and reduce to hers when specialized to certain limits. Our derivations are based on ideas and techniques that originated in computations of certain physical quantities in exactly solvable lattice models, particularly Baxter's corner transfer matrix method, and the combinatorial structures related to it.

Since the very same methods and techniques are available, we expect that our approach can be used for the cases of other affine algebras too. Let us consider the $\widehat{\mathrm{sl}}(n)$ case. In the case we treated in this work, the affine Weyl group W had two subsets W^+ and W^- satisfying $W^+\cup W^-=W$, $W^+\cap W^-=\{1\}$. On each subset the Bruhat order was a total order. But if $n>2$ the Weyl group does not have such a property. Thus, it is very difficult to treat the Demazure modules corresponding to all Weyl group elements. But for some particular type of elements there does exist a correspondence between Demazure crystals and paths, which we would like to report in the near future.

As for one-dimensional configuration sums in the higher rank case, few results are available, except the level 1 case.[16] We would like to push forward these studies. Along this line we note the works by Kuniba–Nakanishi–Suzuki[22] and Georgiev.[23] In Ref. 22 they presented conjectures of fermionic expressions for string functions of vacuum modules $V(k\Lambda_0)$ when the affine algebra is of type $X_r^{(1)}$. Georgiev proved the conjecture in the $A_r^{(1)}$ case. For our purpose we have to ''finitize'' their conjecture or result, since we need generating functions of paths of finite length. This would be a challenging problem.

Our approach can be also extended to types other than $\widehat{\mathrm{sl}}(n)$. In these cases we do not have descriptions of crystals by extended Young diagrams or similar notions yet. We might have to work directly on paths. We have so far discussed the cases of affine algebra modules. There are definitely the cases of coset modules left. We also would like to understand the representation theoretical meaning of the truncated generating functions.

ACKNOWLEDGMENTS

We wish to thank S. Dasmahapatra, A. Kuniba, and S. O. Warnaar for discussions and interest in this work. This work was started while K.M. and M.O. were visiting the University of Melbourne in July, 1995, and continued while O.F. was visiting the University of Utrecht. We thank the Australian Research Council, and the Netherlands Organization for Scientific Research (NWO) for their financial support of these visits. O.F. was supported in part by the Australian Research Council (ARC), and the Netherlands Organization for Scientific Research (NWO). K.C.M. was

supported in part by NSA/MSP Grant No. MDA904-96-1-0013. Masato Okado was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan.

[1] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic, London, 1982).
[2] M. Jimbo and T. Miwa, *Algebraic Analysis of Solvable Lattice Models*, CBMS Regional Conference Series in Mathematics, Vol. 85 (American Mathematical Society, Providence, RI, 1994).
[3] S. Dasmahapatra, R. Kedem, T. R. Klassen, B. M. McCoy, and E. Melzer, ''Quasi-particles, conformal field theory, and q-series,'' in *Yang Baxter Equations in Paris*, edited by J.-M. Maillard (World Scientific, Singapore, 1992).
[4] S. Kumar, ''Demazure character formula in arbitrary Kac-Moody setting,'' Invent. Math. **89**, 395–423 (1987).
[5] O. Mathieu, ''Formula de Demazure-Weyl et généralisation du théoème de Borel–Weil–Bott,'' C. R. Acad. Sci. Paris, Sér. I, **303**, 391–394 (1986).
[6] Y. B. Sanderson, ''On characters of Demazure modules,'' Ph.D. thesis, Rutgers University, 1995.
[7] P. Littelmann, ''A Littlewood-Richardson rule for symmetrizable Kac–Moody algebras,'' Invent. Math. **116**, 329–346 (1994).
[8] M. Kashiwara, ''Crystalizing the q-analogue of universal enveloping algebra,'' Commun. Math. Phys. **133**, 249–260 (1990).
[9] T. Nakashima, ''Crystal base and a generalization of the Littlewood-Richardson rule for the classical Lie algebras,'' Commun. Math. Phys. **154**, 215–243 (1993).
[10] M. Kashiwara, ''The crystal base and Littelmann's refined Demazure character formula,'' Duke Math. J. **71**, 839–858 (1993).
[11] S-J. Kang, M. Kashiwara, K. C. Misra, T. Miwa, T. Nakashima, and A. Nakayashiki, ''Affine crystals and vertex models,'' Int. J. Mod. Phys. A **7**, 449–484 (1992).
[12] S-J. Kang, M. Kashiwara, K. C. Misra, T. Miwa, T. Nakashima, and A. Nakayashiki, ''Perfect crystals of quantum affine Lie algebras,'' Duke Math. J. **68**, 499–607 (1992).
[13] K. C. Misra and T. Miwa, ''Crystal base for the basic representation of $U_q(\hat{sl}(n))$,'' Commun. Math. Phys. **134**, 79–88 (1990).
[14] M. Jimbo, K. C. Misra, T. Miwa, and M. Okado, ''Combinatorics of representations of $U_q(\hat{sl}(n))$ at $q=0$,'' Commun. Math. Phys. **136**, 543–566 (1991).
[15] E. Date, M. Jimbo, A. Kuniba, T. Miwa, and M. Okado, ''Exactly solvable SOS models II: Proof of the star–triangle relation and combinatorial identities,'' Adv. Stud. Pure Math. **16**, 17–122 (1988).
[16] E. Date, M. Jimbo, A. Kuniba, T. Miwa, and M. Okado, ''One dimensional configuration sums in vertex models and affine Lie algebra characters,'' Lett. Math. Phys. **17**, 69–77 (1989).
[17] A. Schilling, ''Multinomials and polynomial bosonic forms for the branching functions of the $\widehat{\mathrm{su}}(2)_M \times \widehat{\mathrm{su}}(2)_N/\widehat{\mathrm{su}}(2)_{M+N}$ conformal coset models,'' Nucl. Phys. B **467**, 247–271 (1996).
[18] V. G. Kac, *Infinite Dimensional Lie Algebras* (Cambridge University Press, Cambridge, 1990).
[19] M. Kashiwara, ''On crystal bases of the q-analogue of universal enveloping algebras,'' Duke Math. J. **63**, 465–516 (1991).
[20] S-J. Kang, M. Kashiwara, and K. C. Misra, ''Crystal bases of Verma modules for quantum affine Lie algebras,'' Compos. Math. **92**, 299–325 (1994).
[21] G. E. Andrews, *The Theory of Partition* (Addison-Wesley, Reading, MA, 1976).
[22] A. Kuniba, T. Nakanishi, and J. Suzuki, ''Characters in conformal field theories from thermodynamic Bethe Ansatz,'' Mod. Phys. Lett. A **8**, 1649–1659 (1993).
[23] G. Georgiev, ''Combinatorial constructions of modules for infinite-dimensional Lie algebras, II. Parafermionic space,'' preprint, Rutgers University, q-alg/9504024.

L operators and Drinfeld's generators

Norifumi Hayaishi and Kei Miki
Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka 560, Japan

(Received 27 May 1997; accepted for publication 29 October 1997)

Utilizing the multiplicative formula of universal R matrices, the correspondence between the L operators and Drinfeld's generators is explicitly calculated for quantum group $U_q(g)$ with $g=A_l^{(1)},B_l^{(1)},C_l^{(1)},D_l^{(1)}$. © *1998 American Institute of Physics.* [S0022-2488(98)00103-0]

I. INTRODUCTION

By Drinfeld[1,2] and Jimbo,[3] a quantum group was invented by quantizing the universal enveloping algebras of Kac–Moody algebras. It was defined by deforming the defining relations for Chevally generators. Later Drinfeld found another realization of quantum group for affine Kac–Moody algebras in Ref. 4. Finally, Faddeev, Reshetikhin and Takhtatjan[5] defined quantum group by the $RLL=LLR$ relations, which deforms the matrix realization of finite dimensional simple Lie algebras. This was later generalized to the affine case by Reshetikhin and Semenov-Tian-Shansky.[6] Up to now, the relations among these three definitions are clarified as follows. The universal R matrix by Drinfeld[2] gives the map from the third formulation to the first.[5,6] In the original paper,[4] the isomorphism from the first definition to the second is given. (See Ref. 7 for the explicit verification.) Later in the nontwisted affine case, the map in the opposite direction is obtained in Ref. 8, without the use of the Braid group, and in Ref. 9 by using the Braid group actions. The correspondence between the second and third definitions is found by Ding and Frenkel[10] for $g=\widehat{gl_n}$. The purpose of this paper is to generalize the last result to $g=A_l^{(1)},B_l^{(1)},C_l^{(1)},D_l^{(1)}$ by utilizing the multiplicative formula of the universal R matrix. [See Ref. 11 for a similar calculation in the case $\widehat{DY(sl_2)}$.] This formula in the nontwisted affine case is already obtained in Ref. 12 without the use of the Braid group. In this paper, we follow the approach by Beck[9,13] and Lusztig.[14] Their results easily give the multiplicative formula, and this formula immediately yields the desired correspondence. The results in this paper would be well known conceptually. We only hope that the explicit expressions here help to clarify the problem.

This paper is organized as follows. In section II, we fix our notations. In section III, we determine the defining relations of quantum group in the L operator formalism for $g=A_l^{(1)}$, $B_l^{(1)}$, $C_l^{(1)}$, $D_l^{(1)}$, $A_{2l}^{(2)}$, $A_{2l-1}^{(2)}$, $D_{l+1}^{(2)}$, since we cannot find the proof in literature. Here we also consider the twisted case since the proof is essentially the same as the one in the nontwisted case. In section IV, after a review of the result by Beck, we write down the multipilcative formula of the universal R matrix for nontwisted affine Kac–Moody algebras. In section V, utilizing the result in section III, the multiplicative formula and the result by Beck, we calculate the correspondence between the L operators and Drinfeld's generators for $g=A_l^{(1)}$, $B_l^{(1)}$, $C_l^{(1)}$, $D_l^{(1)}$. The work by Edwards, announced in Ref. 15, would immediately generalize the results here to the twisted case.

II. NOTATIONS

A. Root data

Let g be an affine Kac–Moody algebra over $\mathbf{Q}$ and $(a_{ij})_{0\leqslant i,j\leqslant l}$ its generalized Cartan matrix. Let further $a_0,\ldots,a_l$ and $a_0^\vee,\ldots,a_l^\vee$ be relatively prime positive integers such that $\Sigma_i a_i^\vee a_{ij}=0$ and $\Sigma_j a_{ij}a_j=0$, respectively. We set $h^\vee=\Sigma_{i=0}^l a_i^\vee$. We follow the enumeration of the vertices of the Dynkin diagram in Ref. 16, Chap. 4 except in the $A_{2l}^{(2)}$ case, in which we enumerate them in reverse order. In our notation, $a_0=1$ for any g and $a_0^\vee=2$ for $g=A_{2l}^{(2)}$, $=1$ otherwise. We denote the finite type Kac–Moody algebra corresponding to the Cartan matrix $(a_{ij})_{1\leqslant i,j\leqslant l}$ by $\overline{g}$.

 © 1998 American Institute of Physics

Let h^* be a $\mathbf{Q}$ vector space with basis $\alpha_0,\ldots,\alpha_l,\Lambda_0$. Let further $d_0,\ldots,d_l$ be relatively prime positive integers such that $d_i a_{ij}=d_j a_{ji}$. We introduce a symmetric bilinear form $(\ |\ ):h^*\times h^*\to\mathbf{Q}$ by $(\alpha_i|\alpha_j)=d_i a_{ij}$, $(\alpha_i|\Lambda_0)=d_i\delta_{i0}$ and $(\Lambda_0|\Lambda_0)=0$. Set $\delta=\Sigma_{i=0}^{l}a_i\alpha_i$ and $\alpha_i^\vee=\alpha_i/d_i$ $(0\leqslant i\leqslant l)$. Define ω_i and $\omega_i^\vee\in h^*$ $(1\leqslant i\leqslant l)$ by $(\alpha_j^\vee|\omega_i)=\delta_{ij}$ $(1\leqslant j\leqslant l)$, $(\delta|\omega_i)=(\Lambda_0|\omega_i)=0$ and $\omega_i^\vee=\omega_i/d_i$. We further define several $\mathbf{Z}$ submodules of h^* by $Q=\oplus_{i=0}^{l}\mathbf{Z}\alpha_i$, $\Sigma=Q\oplus\mathbf{Z}\Lambda_0/d_0$, $\Gamma=\Sigma+p\mathbf{Z}\omega_1$ and $\overline{Q}=\oplus_{i=1}^{l}\mathbf{Z}\alpha_i$. Here and hereafter $p=2$ for $g=A_2^{(2)}$ and $=1$ otherwise. Finally, as usual, we introduce an order relation on Q by $\lambda\geqslant\mu\Leftrightarrow\lambda-\mu\in Q_+=\oplus_{i=0}^{l}\mathbf{Z}_{\geqslant 0}\alpha_i$.

B. F algebra $U_q^\Gamma(g)$

Let q be an indeterminate and set $F=\mathbf{Q}(q)$. Following Refs. 1 and 3, let $\mathscr{U}=U_q^\Gamma(g)$ be the F algebra generated by e_i, f_i $(0\leqslant i\leqslant l)$ and k_λ $(\lambda\in\Gamma)$ with the relations

$$k_\lambda k_\mu=k_{\lambda+\mu},\quad k_0=1, \tag{2.1}$$

$$k_\lambda e_i k_{-\lambda}=q^{(\lambda|\alpha_i)}e_i,\quad k_\lambda f_i k_{-\lambda}=q^{-(\lambda|\alpha_i)}f_i, \tag{2.2}$$

$$[e_i,f_j]=\delta_{ij}\frac{k_i-k_i^{-1}}{q_i-q_i^{-1}}, \tag{2.3}$$

$$\sum_{r=0}^{1-a_{ij}}(-1)^r e_i^{(r)}e_j e_i^{(1-a_{ij}-r)}=0\quad(i\neq j), \tag{2.4}$$

$$\sum_{r=0}^{1-a_{ij}}(-1)^r f_i^{(r)}f_j f_i^{(1-a_{ij}-r)}=0\quad(i\neq j), \tag{2.5}$$

where $k_i^{\pm 1}=k_{\pm\alpha_i}$ and

$$q_i=q^{d_i},\quad x_i^{(n)}=\frac{x_i}{[n]_i!}\ \ (x=e,f),\quad [n]_i=\frac{q_i^n-q_i^{-n}}{q_i-q_i^{-1}},\quad [n]_i!=\prod_{k=1}^{n}[k]_i.$$

This algebra has the following Hopf algebra structure:

$$\Delta(e_i)=e_i\otimes 1+k_i\otimes e_i,\quad \Delta(f_i)=f_i\otimes k_i^{-1}+1\otimes f_i,$$

$$\Delta(k_\lambda)=k_\lambda\otimes k_\lambda,$$

$$S(e_i)=-k_i^{-1}e_i, S(f_i)=-f_i k_i,\quad S(k_\lambda)=k_{-\lambda},$$

$$\epsilon(e_i)=\epsilon(f_i)=0,\quad \epsilon(k_\lambda)=1. \tag{2.6}$$

We define several subalgebras of $\mathscr{U}$ by

$$\mathscr{U}^+=\langle e_i|0\leqslant i\leqslant l\rangle,\quad \mathscr{U}^-=\langle f_i|0\leqslant i\leqslant l\rangle,\quad \mathscr{U}^0=\langle k_\lambda|\lambda\in\Gamma\rangle,$$

$$\mathscr{U}^{\geqslant 0}=\mathscr{U}^0\mathscr{U}^+,\quad \mathscr{U}^{\leqslant 0}=\mathscr{U}^-\mathscr{U}^0.$$

Then $\mathscr{U}^\pm=\oplus_{\mu\in Q_+}\mathscr{U}^\pm_{\pm\mu}$ where

$$\mathscr{U}^\pm_\mu=\{x\in\mathscr{U}^\pm|\ k_\lambda x k_{-\lambda}=q^{(\lambda|\mu)}x\ \ (\forall\lambda\in\Gamma)\}.$$

Moreover the following is known.

Proposition 1:[17,18,14]

(1) *The multiplication map* $\mathscr{U}^-\otimes\mathscr{U}^0\otimes\mathscr{U}^+\to\mathscr{U}$ *induces an isomorphism of vector spaces.*

(2) $(k_\lambda)_{\lambda\in\Gamma}$ *forms a basis of* $\mathscr{U}^0$.

(3) $\dim_F \mathscr{U}^{\pm}_{\pm\mu} = \mathscr{P}(\mu)$, *where* $\mathscr{P}(\mu)$ *is Kostant's partition function for* g.

Similarly we define $U = U_q^{\Sigma}(g)$, $U' = U_q^{Q}(g)$ and $\overline{U} = U_q^{\overline{Q}}(\overline{g})$. We identify U (resp., U') with the subalgebra of $\mathscr{U}$ (resp., U). Then $\mathscr{U}^{\pm} = U^{\pm} = U'^{\pm}$.

C. Finite dimensional $U_q^Q(g)$ module (ρ, V)

Hereafter to the end of the next section, we consider $g = A_l^{(1)}$ $(l \geqslant 1)$, $B_l^{(1)}$ $(l \geqslant 3)$, $C_l^{(1)}$ $(l \geqslant 2)$, $D_l^{(1)}$ $(l \geqslant 4)$, $A_{2l}^{(2)}$ $(l \geqslant 1)$, $A_{2l-1}^{(2)}$ $(l \geqslant 3)$, $D_{l+1}^{(2)}$ $(l \geqslant 2)$. For a $\overline{U}$ module M, we shall call $v \in M$ a weight vector with weight $\mu \in \overline{P} = \oplus_{1 \leqslant i \leqslant l} \mathbf{Z}\omega_i$ if $k_\lambda v = q^{(\lambda|\mu)} v$ for any $\lambda \in \overline{Q}$. Let $V(p\omega_1)$ be the irreducible highest weight module with highest weight $p\omega_1$ of $\overline{U}$. Let further v_i $[1 \leqslant i \leqslant \dim V(p\omega_1)]$ be a basis of this module consisting of weight vectors such that $i < j$ if $\eta_i - \eta_j > 0$, where η_i is the weight of v_i as a $\overline{U}$ module. $V(\omega_1) \oplus F$ in the case $g = D_{l+1}^{(2)}$, and $V(p\omega_1)$ otherwise can be made into representations of U', on which $k_{\pm\delta}$ acts as 1. We shall denote them by (ρ, V) and set $N = \dim V$. In the case $g = D_{l+1}^{(2)}$, by v_N and η_N we denote the basis vector of the trivial representation as a $\overline{U}$ module and its weight 0, respectively. The representations (ρ, V) are the same as the ones used in Ref. 19 and, in the limit $q = 1$, reduce to the ones in Ref. 16, Chapters 7 and 8. Let τ_b $(b \in F^\times)$ be the automorphism of $\mathscr{U}$ determined by $\tau_b(e_i) = b^{\delta_{i0}} e_i$, $\tau_b(f_i) = b^{-\delta_{i0}} f_i$ and $\tau_b(k_\lambda) = k_\lambda$. Put $\rho_b = \rho \circ \tau_b|_{U'}$. Set $M = l + 1$ and $b_i = q^{-2(i-1)}$ $(1 \leqslant i \leqslant M)$ for $g = A_l^{(1)}$. In other cases, set $M = 2$ and $b_1 = 1$, $b_2 = \sigma q_0^{-h^\vee/a_0^\vee}$. Here $\sigma = -1$ for $g = A_{2l}^{(2)}$, $A_{2l-1}^{(2)}$ and $= 1$ otherwise. Then in $(\rho_{b_1} \otimes \cdots \otimes \rho_{b_M}, V^{\otimes M})$, there exists a trivial representation of U'. We denote its basis vector by w. In the case $M = 2$, we define a matrix $J = (J_{ij})_{1 \leqslant i,j \leqslant N}$ by $w = \Sigma_{i,j} J_{ij} v_i \otimes v_j$. See Appendix A for the explicit expressions of N, η_i, (ρ, V) and w. Below we identify $\mathrm{End}V$ with $M_N(F)$ through the basis $v_1, \cdots, v_N$ normalized as in Appendix A.

D. R matrix and L operators

Let Θ be the quasi-universal R matrix of U.[2,20,14] This is an invertible element of (an appropriate completion of) $U \otimes U$ uniquely determined by

$$\Theta = \sum_{\mu \geqslant 0} \Theta_\mu, \quad \Theta_\mu \in U_\mu^+ \otimes U_{-\mu}^-, \quad \Theta_0 = 1 \otimes 1,$$

$$\iota \circ \Delta(x) \Theta = \Theta \Psi \circ \Delta(x) \quad (x \in U). \tag{2.7}$$

Here $\iota(x \otimes y) = y \otimes x$, and Ψ is the automorphism of $U \otimes U$ defined by

$$\Psi(e_i \otimes 1) = e_i \otimes k_i^{-1}, \quad \Psi(f_i \otimes 1) = f_i \otimes k_i,$$

$$\Psi(k_\lambda \otimes k_\mu) = k_\lambda \otimes k_\mu, \quad \Psi = \iota \circ \Psi \circ \iota.$$

This element further satisfies

$$\Theta_{\alpha_i} = -(q_i - q_i^{-1}) e_i \otimes f_i \quad (0 \leqslant i \leqslant l), \tag{2.8}$$

$$(1 \otimes \Delta)\Theta = \Theta_{13} \Psi_{13}(\Theta_{12}), \quad (\Delta \otimes 1)\Theta = \Theta_{13} \Psi_{13}(\Theta_{23}), \tag{2.9}$$

$$\Theta_{12} \Psi_{12}(\Theta_{13}) \Theta_{23} = \Theta_{23} \Psi_{23}(\Theta_{13}) \Theta_{12}, \tag{2.10}$$

where Θ_{13}, for example, denotes $\Sigma_i x_i \otimes 1 \otimes y_i$ for $\Theta = \Sigma_i x_i \otimes y_i$ as usual.

For Θ^{-1}, define $(\Theta^{-1})_\mu$ $(\mu \in Q_+)$ as for Θ. Let z be a formal variable. Let us introduce $\hat{\mathscr{L}}^{\pm}(z) \in (M_N(F) \otimes U^{\mp})[[z^{\pm 1}]]$ and $\mathscr{R}(z) \in (M_N(F) \otimes M_N(F))[[z]]$ by

$$\hat{\mathscr{S}}^{+}(z)=\sum_{\mu\geqslant 0}(\rho\otimes 1)(\Theta_{\mu})\ z^{(\Lambda_0|\mu)/d_0},$$

$$\hat{\mathscr{S}}^{-}(z)=\sum_{\mu\geqslant 0}(\rho\otimes 1)(((\Theta_{21})^{-1})_{\mu})\ z^{-(\Lambda_0|\mu)/d_0},$$

$$\mathscr{R}(z)=(1\otimes\rho)\mathscr{S}^{+}(z). \tag{2.11}$$

We further introduce L operators and an R matrix by

$$\hat{L}^{+}(z)=\hat{\mathscr{S}}^{+}(z)\hat{\mathscr{T}}^{-1},\quad \hat{L}^{-}(z)=\hat{\mathscr{T}}\hat{\mathscr{S}}^{-}(z),$$

$$R(z)=\mathscr{R}(z)T^{-1}=\sum_{n\geqslant 0}\sum_{1\leqslant i_1,i_2,j_1,j_2\leqslant N}R[n]^{i_1j_1}_{i_2j_2}E_{i_1i_2}\otimes E_{j_1j_2}z^{n}, \tag{2.12}$$

where

$$\hat{\mathscr{T}}^{\pm 1}=\sum_{1\leqslant i\leqslant N}E_{ii}\otimes k_{\pm\eta_i},\quad T^{\pm 1}=\sum_{1\leqslant i,j\leqslant N}q^{\pm((\eta_i|\eta_j)-(\eta_1|\eta_1))}E_{ii}\otimes E_{jj},$$

and E_{ij} is a matrix unit.

Finally we state the simple properties of R matrix, which we shall need in the next section. Let $(\dot{\rho},\mathbf{Q}^N)$ denote the representation of $g'=[g,g]$ obtained by taking the limit $q=1$ of $\rho(\cdot)\in M_N(F)$. Let $r(z)\in(\dot{\rho}(g')\otimes_{\mathbf{Q}}\dot{\rho}(g'))[[z]]$ be the trigonometric classical r matrix[21] of type g normalized by the condition that the diagonal part of the z^0 component of $r(z)$ is $\Sigma_{1\leqslant i,j\leqslant N}[(\eta_i|\eta_j)/2]E_{ii}\otimes E_{jj}(=:r_0)$. Let further A be the $\mathbf{Q}$ subalgebra of F consisting of the elements which have no pole at $q=1$.

Lemma 1:

(1) $R[0]^{i_1j_1}_{i_2j_2}\neq 0$ only if $i_1\leqslant i_2$ and $j_1\geqslant j_2$.

(2) $\eta_{i_1}+\eta_{j_1}=\eta_{i_2}+\eta_{j_2}$, $i_1=i_2\Leftrightarrow j_1=j_2$ when $R[n]^{i_1j_1}_{i_2j_2}\neq 0$.

(3) $\mathscr{R}(z)=1-(q-q^{-1})(r(z)-r_0)$ mod $(q-q^{-1})^2(M_N(A)\otimes_A M_N(A))[[z]]$.

Proof: (3) Let r_μ $(\mu>0)$ be the elements of $M_N(\mathbf{Q})\otimes_{\mathbf{Q}}M_N(\mathbf{Q})$ uniquely determined by the following conditions:

$$r_{\alpha_i}=d_i\dot{\rho}(e_i)\otimes\dot{\rho}(f_i)\quad(0\leqslant i\leqslant l),$$

$$[1\otimes\dot{\rho}(e_i),r_\mu]+[\dot{\rho}(e_i)\otimes 1,r_{\mu-\alpha_i}]=0\quad(\mu(\neq\alpha_i)>0,\ 0\leqslant i\leqslant l),$$

$$[\dot{\rho}(f_i)\otimes 1,r_\mu]+[1\otimes\dot{\rho}(f_i),r_{\mu-\alpha_i}]=0\quad(\mu(\neq\alpha_i)>0,\ 0\leqslant i\leqslant l),$$

$$\sum_{k=1}^{M}(r_\mu)_{k\,M+1}\dot{w}\otimes 1_{\mathbf{Q}^N}=0\quad(\mu>0),\quad\text{for } g\neq A^{(2)}_{2l},A^{(2)}_{2l-1},$$

$$(\mathrm{tr}\otimes 1)(r_\mu)=0\quad(\mu>0),\quad\text{for } g=A^{(2)}_{2l},A^{(2)}_{2l-1},$$

where $\dot{w}=w|_{q=1}$. (See Ref. 16, Proposition 8.3.a.) Then $r(z)=\Sigma_{\mu\geqslant 0}r_\mu z^{(\Lambda_0|\mu)/d_0}$. Hence by considering the equations from (2.7), (2.8) and (2.9) in the limit $q=1$, the claim can be easily proven by induction on $\mathrm{ht}\mu$. In the case $g=A^{(2)}_{2l}$, $A^{(2)}_{2l-1}$, we also use (3.14) below with $l_{ij}[-n]=-\hat{\mathscr{S}}^{+}_{ij}[-n]/(q-q^{-1})$. (In the nontwisted case, it is also possible to use the multiplicative formula of Θ below. See Ref. 9, Section 6 and Ref. 22, Section 1.)

III. *L* OPERATOR FORMALISM

A. *F* algebra $\mathscr{A}(g)$

Let g, (ρ,V), N, η_i $(1\leqslant i\leqslant N)$, $R(z)$, M, b_i $(1\leqslant i\leqslant M)$ and w as before. Following Refs. 5, 6, 10, 23, let $\mathscr{A}=\mathscr{A}(g)$ be the F algebra generated by $L_{ij}^{+}[0]$ $(1\leqslant i\leqslant j\leqslant N)$, $L_{ij}^{-}[0]$ $(1\leqslant j\leqslant i\leqslant N)$, $L_{ij}^{\pm}[\mp n]$ $(1\leqslant i,j\leqslant N, n\in\mathbf{Z}_{>0})$, $C^{\pm 1}$ and $D^{\pm 1}$ with the relations

$$L_{ii}^{\pm}[0]L_{ii}^{\mp}[0]=C^{\pm 1}C^{\mp 1}=D^{\pm 1}D^{\mp 1}=1, \tag{3.1}$$

$$L_{ii}^{\pm}[0]=1, \text{ for } i \text{ such that } \eta_i=0, \tag{3.2}$$

$$C^{\pm 1} \text{ central}, \tag{3.3}$$

$$DL_{ij}^{\pm}[\mp n]D^{-1}=q^{\mp n}L_{ij}^{\pm}[\mp n], \tag{3.4}$$

$$R_{12}(Cz_1/z_2)L_1^{+}(z_1)L_2^{-}(z_2)=L_2^{-}(z_2)L_1^{+}(z_1)R_{12}(C^{-1}z_1/z_2), \tag{3.5}$$

$$R_{12}(z_1/z_2)L_1^{\pm}(z_1)L_2^{\pm}(z_2)=L_2^{\pm}(z_2)L_1^{\pm}(z_1)R_{12}(z_1/z_2), \tag{3.6}$$

$$L_1^{\pm}(b_1z)\cdots L_M^{\pm}(b_Mz)w\otimes 1=w\otimes 1, \tag{3.7}$$

$$(G\otimes 1)L^{\pm}(z)(G^{-1}\otimes 1)=L^{\pm}(-z), \quad \text{for } g=D_{l+1}^{(2)}. \tag{3.8}$$

Here $L^{\pm}(z)$ is defined in terms of the generating series $L_{ij}^{\pm}(z)=\Sigma_{n\geqslant 0}L_{ij}^{\pm}[\mp n]z^{\pm n}$ as

$$L^{\pm}(z)=\sum_{1\leqslant i,j\leqslant N}E_{ij}\otimes L_{ij}^{\pm}(z)\in(M_N(F)\otimes\mathscr{A})[[z^{\pm 1}]];$$

$L_i^{\pm}(z)$ denotes $L^{\pm}(z)$ whose matrix part acts on the ith space, z_1 and z_2 are formal variables, and

$$G=\operatorname{diag}(1,\ldots,1,-1)\in M_N(F).$$

It is well known and easy to check that $\mathscr{A}$ has the following Hopf algebra structure:

$$\Delta(L_{ij}^{+}(z))=\sum_{k=1}^{N}L_{kj}^{+}(C_2^{-1}z)\otimes L_{ik}^{+}(z), \quad \Delta(L_{ij}^{-}(z))=\sum_{k=1}^{N}L_{kj}^{-}(z)\otimes L_{ik}^{-}(C_1^{-1}z),$$

$$\Delta(C^{\pm 1})=C^{\pm 1}\otimes C^{\pm 1}, \quad \Delta(D^{\pm 1})=D^{\pm 1}\otimes D^{\pm 1},$$

$$\epsilon(L_{ij}^{\pm}(z))=\delta_{ij}, \quad \epsilon(C^{\pm 1})=\epsilon(D^{\pm 1})=1,$$

$$(1\otimes S)(L^{\pm}(z))={}^{t}(({}^{t}L^{\pm}(Cz))^{-1}), \quad S(C^{\pm 1})=C^{\mp 1}, \quad S(D^{\pm 1})=D^{\mp 1}, \tag{3.9}$$

where $C_1=C\otimes 1$, $C_2=1\otimes C$ and ^{t} denotes the matrix transpose in the first space of the tensor product. It is also well known and easy to check that there is a surjective Hopf algebra homomorphism $\phi:\mathscr{A}\to\mathscr{U}$ defined by

$$L^{\pm}(z)\mapsto\hat{L}^{\pm}(z), \quad C^{\pm 1}\mapsto k_{\pm\delta}, \quad D^{\pm 1}\mapsto k_{\pm\Lambda_0/d_0}.$$

In the next subsection, we shall show the injectivity of ϕ by the standard specialization argument at $q=1$.[18,22] The only subtlety appears in the case $A_{2l}^{(2)}$, $A_{2l-1}^{(2)}$.

B. Proof of the isomorphism $\mathscr{A}(g)\simeq U_q^{\Gamma}(g)$

Set

$$\mathscr{L}^{+}(z)=L^{+}(z)\mathscr{T}, \quad \mathscr{L}^{-}(z)=\mathscr{T}^{-1}L^{-}(z), \quad \mathscr{T}^{\pm 1}=\sum_{1\leqslant i\leqslant N}E_{ii}\otimes L_{ii}^{\mp}[0], \tag{3.10}$$

and define the components $\mathscr{L}_{ij}^{\pm}[\mp n]$ as before. We further define subalgebras of $\mathscr{A}$ by

$$\mathscr{A}^{\pm}=\langle \mathscr{L}_{ij}^{\mp}[\mp n] | 1\leqslant i,j\leqslant N,\ n\geqslant 0\rangle, \quad \mathscr{A}^{0}=\langle L_{ii}^{\pm}[0], C^{\pm 1}, D^{\pm 1} | 1\leqslant i\leqslant N\rangle,$$

and, for $\mu\in Q$, we set

$$\mathscr{A}_{\mu}^{\pm}=\{x\in\mathscr{A}^{\pm} | L_{ii}^{-}[0]xL_{ii}^{+}[0]=q^{(\eta_i|\mu)}x \quad (1\leqslant i\leqslant N), \quad DxD^{-1}=q^{(\Lambda_0|\mu)/d_0}x\}.$$

Lemma 2:
(1) $L_{kk}^{-}[0]L_{ij}^{\pm}[\mp n]=q^{-(\eta_k|\eta_i-\eta_j)}L_{ij}^{\pm}[\mp n]L_{kk}^{-}[0]$.
(2) $\mathscr{A}^{0}$ is commutative.
(3) $\Pi_{i=1}^{N}L_{ii}^{\pm}[0]^{n_i}=1$ if $\Sigma_i n_i\eta_i=0$.
(4) $\mathscr{A}=\mathscr{A}^{-}\mathscr{A}^{0}\mathscr{A}^{+}$.

Proof: (3) follows from (3.2) and the z^0 components of (3.7) and (3.8).

Lemma 3: $\mathscr{A}^{\pm}=\oplus_{\mu\in Q_+}\mathscr{A}_{\pm\mu}^{\pm}$ and $\dim_F\mathscr{A}_{\pm\mu}^{\pm}\leqslant\mathscr{P}(\mu)$.

Proof: This can be proven as in Ref. 22, Proposition 1.13.

Let $\mathscr{N}$ be the F algebra generated by $\mathscr{L}_{ij}[0]$ $(1\leqslant i<j\leqslant N)$ and $\mathscr{L}_{ij}[-n]$ $(1\leqslant i,j\leqslant N,\ n>0)$ with the relations

$$\mathscr{R}_{12}(z_1/z_2)T_{12}^{-1}\mathscr{L}_1(z_1)T_{12}\mathscr{L}_2(z_2)=\mathscr{L}_2(z_2)T_{12}\mathscr{L}_1(z_1)T_{12}^{-1}\mathscr{R}_{12}(z_1/z_2), \tag{3.11}$$

$$\mathscr{L}_1(b_1z)T_{12}\mathscr{L}_2(b_2z)T_{12}^{-1}\cdots T_{1;M}\mathscr{L}_M(b_Mz)T_{1;M}^{-1}w\otimes 1=w\otimes 1, \tag{3.12}$$

$$(G\otimes 1)\mathscr{L}(z)(G^{-1}\otimes 1)=\mathscr{L}(-z), \quad \text{for } g=D_{l+1}^{(2)}, \tag{3.13}$$

where $T_{1;k}=T_{1k}\cdots T_{k-1\,k}$, and $\mathscr{L}(z)$ are defined as $L^{+}(z)$ with $\mathscr{L}_{ii}[0]=1$. Next let $\mathscr{N}_A$ be the A algebra generated by $l_{ij}[0]$ $(1\leqslant i<j\leqslant N)$ and $l_{ij}[-n]$ $(1\leqslant i,j\leqslant N,\ n>0)$ with the following relations. Let $l(z)$ be defined as before with $l_{ii}[0]=0$. First substitute $\mathscr{L}(z)=1-(q-q^{-1})l(z)$ into (3.11)–(3.13) and then divide (3.11) by $(q-q^{-1})^2$ and the rest by $q-q^{-1}$. We take the thus obtained equations as the defining relations of $\mathscr{N}_A$. In the case $g=A_{2l}^{(2)},A_{2l-1}^{(2)}$, we impose another relation,

$$\frac{\operatorname{tr} l(b_2z)-\operatorname{tr} l(b_2^{-1}z)}{q-q^{-1}}=\operatorname{tr}\ \tilde{l}(z)^2-\operatorname{tr} l(z)^2+(q-q^{-1})(\operatorname{tr} l(z)^2\tilde{l}(b_2z)-\operatorname{tr} l(b_2^{-1}z)\tilde{l}(z)^2), \tag{3.14}$$

where $\tilde{l}(z)=\Sigma_{n\geqslant 0}\Sigma_{ij}E_{ij}\otimes(u_i/u_j)\,q^{(\eta_i|\eta_i-\eta_j)}l_{N+1-j\ N+1-i}[-n]z^n$ (see Sec. 3 in Appendix A for the u_i) and tr denotes the trace of the matrix part. Finally let n^{-} be the subalgebra of g generated by the f_i in the notations of Ref. 16 and $U(n^{-})$ its enveloping algebra. $U(n^{-})$ can be defined by the generators $l_{ij}[0]$ $(1\leqslant i<j\leqslant N)$, $l_{ij}[-n]$ $(1\leqslant i,j\leqslant N,\ n>0)$ and the relations

$$\begin{gathered}
[r_{12}(z_1/z_2),l_1(z_1)+l_2(z_2)]+[l_1(z_1),l_2(z_2)]+[(r_0)_{12},l_1(z_1)-l_2(z_2)]=0,\\
l(z)(\dot{J}\otimes 1)+(\dot{J}\otimes 1)\,{}^{t}l(\sigma z)=0, \quad \text{for } g\neq A_l^{(1)},\\
\operatorname{tr}\ l(z)=0, \quad \text{for } g=A_l^{(1)},\ A_{2l}^{(2)},\ A_{2l-1}^{(2)},\\
(G\otimes 1)l(z)(G^{-1}\otimes 1)=l(-z), \quad \text{for } g=D_{l+1}^{(2)},
\end{gathered} \tag{3.15}$$

where $\dot{J}=J|_{q=1}$. (See Ref. 16, Chapters 7 and 8.)

Set $\mathscr{N}_F=\mathscr{N}_A\otimes_A F$ and $\mathscr{N}_{\mathbf{Q}}=\mathscr{N}_A\otimes_A\mathbf{Q}$, where A acts on F naturally and on $\mathbf{Q}$ via $q\mapsto 1$. In the case $g=A_{2l}^{(2)},A_{2l-1}^{(2)}$, (3.14) follows from (3.12) in $\mathscr{N}$. Thanks to Lemma 1 (3), in the limit $q=1$ the relations of $\mathscr{N}_A$ reduce to (3.15); cf. Refs. 6 and 23. Therefore

$$\begin{gathered}
\mathscr{N}_{\mathscr{F}}\xrightarrow{\sim}\mathscr{N} \quad (l_{ij}[-n]\otimes 1\mapsto -\mathscr{L}_{ij}[-n]/(q-q^{-1})),\\
\mathscr{N}_{\mathbf{Q}}\xrightarrow{\sim}U(n^{-}) \quad (l_{ij}[-n]\otimes 1\mapsto l_{ij}[-n]).
\end{gathered} \tag{3.16}$$

As easily checked, $\mathcal{N}_A$ is a Q-graded algebra by assigning $\eta_j-\eta_i-n\delta\in Q$ to $l_{ij}[-n]$ and each homogeneous component $(\mathcal{N}_A)_\mu$ $(\mu\in Q)$ is finitely generated over A. Moreover there is a surjective F algebra homomorphism from $\mathcal{N}$ to $\mathcal{A}^-$ which maps $\mathcal{L}_{ij}[-n]$ to $\mathcal{L}_{ij}^+[-n]$. Therefore, thanks to (3.16) we obtain $\dim_F\mathcal{A}^-_{-\mu}\leqslant\dim_F(\mathcal{N}_A)_{-\mu}\otimes_A F\leqslant\dim_{\mathbf{Q}}(\mathcal{N}_A)_{-\mu}\otimes_A\mathbf{Q}=\mathcal{P}(\mu)$. The case $\mathcal{A}^+$ can be similarly proven.

Theorem 1: *For $g=A_l^{(1)}, B_l^{(1)}, C_l^{(1)}, D_l^{(1)}, A_{2l}^{(2)}, A_{2l-1}^{(2)}, D_{l+1}^{(2)}$, $\phi:\mathcal{A}(g)\to U_q^\Gamma(g)$ is a Hopf algebra isomorphism.*

Proof: Clearly $\phi(\mathcal{A}^0)=\mathcal{U}^0$ and $\phi(\mathcal{A}^\pm_{\pm\mu})=\mathcal{U}^\pm_{\pm\mu}$. Since $\Gamma=\oplus_{i=1}^l\mathbf{Z}\eta_i\oplus\mathbf{Z}\delta\oplus\mathbf{Z}\Lambda_0/d_0$, Proposition 1 (2), Eq. (3.1) and Lemma 2 (2), (3) show that $\phi|_{\mathcal{A}^0}$ is injective. Proposition 1 (3) and Lemma 3 prove that $\phi|_{\mathcal{A}^\pm}:\mathcal{A}^\pm\to\mathcal{U}^\pm$ are isomorphisms. Hence Proposition 1 (1) and Lemma 2 (4) yield the claim.

Remark: In the case $g=A_l^{(1)}$, we can replace (3.7) in the definition of $\mathcal{A}(g)$ by the following well known conditions for quantum determinants,

$$\sum_{\gamma\in\mathcal{S}_{l+1}}(-q)^{l(\gamma)}L^\pm_{1\,\gamma(1)}(b_1z)\cdots L^\pm_{l+1\,\gamma(l+1)}(b_{l+1}z)=1,$$

where $l(\gamma)$ denotes the length function of $\mathcal{S}_{l+1}$. For the above relation is contained in (3.7) and is sufficient to give Lemma 2 (3) and tr $l(z)=0$ in (3.15).

IV. MULTIPLICATIVE FORMULA OF Θ

A. Review of the result by Beck

In this section, we shall consider nontwisted affine Kac–Moody algebra g. In this subsection, for completeness and a later purpose, we review the result by Beck[9] on the isomorphism of U and Drinfeld's realization $\mathcal{D}$ (see below). We slightly change the notations in order to do without the square root of the central element. See Ref. 9 and references therein for further details.

Set $\overline{P}^\vee=\oplus_{i=1}^l\mathbf{Z}\omega_i^\vee$ and $\overline{Q}^\vee=\oplus_{i=1}^l\mathbf{Z}\alpha_i^\vee$. Let $\overline{W}$ and W be the Weyl groups of $\overline{g}$ and g, respectively. Set $\widetilde{W}=\overline{W}\ltimes\overline{P}^\vee$, where $\overline{W}$ acts on $\overline{P}^\vee$ naturally. $\widetilde{W}$ is known to be isomorphic to $\Pi\ltimes W$. Here Π is a subgroup of the automorphism group of the Dynkin diagram of type g, which is isomorphic to $\overline{P}^\vee/\overline{Q}^\vee$, and $\pi\in\Pi$ acts on s_i, a simple reflection with respect to α_i, as $\pi s_i\pi^{-1}=s_{\pi(i)}$. $\widetilde{W}$ acts on Q. Its action is given by $\widetilde{w}(\alpha)=\overline{w}(\alpha)-(\alpha|x)\delta=\pi(w(\alpha))$, where $\widetilde{w}=\overline{w}x=\pi w\in\widetilde{W}$ $(\overline{w}\in\overline{W},x\in\overline{P}^\vee,\pi\in\Pi,w\in W)$, $\alpha\in Q$ and $\pi(\alpha_i)=\alpha_{\pi(i)}$.

Let T_i $(0\leqslant i\leqslant l)$[14] and T_π $(\pi\in\Pi)$ be the automorphisms of U' defined by

$$T_i(e_i)=-f_ik_i,\quad T_i(f_i)=-k_i^{-1}e_i,\quad T_i(k_\lambda)=k_{s_i(\lambda)},$$

$$T_i(e_j)=\mathrm{ad}(e_i^{(-a_{ij})})\cdot e_j,\quad T_i(f_j)=f_j\cdot\widetilde{\mathrm{ad}}(f_i^{(-a_{ij})})\quad(i\neq j),$$

$$T_\pi(e_i)=e_{\pi(i)},\quad T_\pi(f_i)=f_{\pi(i)},\quad T_\pi(k_{\alpha_i})=k_{\alpha_{\pi(i)}}.$$

Here ad (resp., $\widetilde{\mathrm{ad}}$) is a left (resp., right) adjoint action defined by

$$\mathrm{ad}(x)\cdot y=\sum x^{(1)}yS(x^{(2)}),\quad y\cdot\widetilde{\mathrm{ad}}(x)=\sum S(x^{(1)})yx^{(2)},$$

with $x,y\in U$ and $\Delta(x)=\Sigma x^{(1)}\otimes x^{(2)}$. For $\widetilde{w}\in\widetilde{W}$, we say an expression $\widetilde{w}=\pi s_{i_1}\cdots s_{i_n}$ $(\pi\in\Pi)$ is reduced if $s_{i_1}\cdots s_{i_n}$ is so in W. For any reduced expression $\widetilde{w}=\pi s_{i_1}\cdots s_{i_n}$, we can set $T_{\widetilde{w}}=T_\pi T_{i_1}\cdots T_{i_n}$. Then $T_{\widetilde{w}}$ $(\widetilde{w}\in\widetilde{W})$ gives a Braid group action on U'. Let Ω be the $\mathbf{Q}$ algebra anti-automorphism of U' defined by

$$e_i\mapsto f_i,\quad f_i\mapsto e_i,\quad k_\lambda\mapsto k_{-\lambda},\quad q\mapsto q^{-1}.$$

This satisfies $\Omega T_{\widetilde{w}}=T_{\widetilde{w}}\Omega$.

Fix $o(i)\in\{-1,1\}$ $(1\leqslant i\leqslant l)$ such that $o(i)=-o(j)$ if $a_{ij}<0$. Let us define $x_{i,k}^{(\pm)}$, $h_{i,r}\in U$ $(1\leqslant i\leqslant l,\ k\in\mathbf{Z},\ r\in\mathbf{Z}-\{0\})$ by

$$x_{i,k}^{(+)}=o(i)^k T_{\omega_i^\vee}^{-k}(e_i),\quad x_{i,k}^{(-)}=\Omega(x_{i,-k}^{(+)}), \tag{4.1}$$

$$k_i^{\mp 1}\exp\left(\mp(q_i-q_i^{-1})\sum_{r>0}h_{i,\mp r}z^{\pm r}\right)=\sum_{r\geqslant 0}\phi_{i,\mp r}^{\pm}z^{\pm r}, \tag{4.2}$$

where

$$\phi_{i,0}^{-}=k_i,\quad \phi_{i,r}^{-}=(q_i-q_i^{-1})k_{r\delta}\ [e_i,x_{i,r}^{(-)}]\quad (r>0),\quad \phi_{i,-r}^{+}=\Omega(\phi_{i,r}^{-})\quad (r\geqslant 0). \tag{4.3}$$

Let $\mathscr{D}=\mathscr{D}(g)^4$ be the F algebra generated by $x_{i,k}^{(\pm)}$, $h_{i,r}$, $k_i^{\pm 1}$, $C^{\pm 1}$, $D^{\pm 1}$ $(1\leqslant i\leqslant l,\ k\in\mathbf{Z},\ r\in\mathbf{Z}-\{0\})$ with the relations

$$k_i^{\pm 1}k_i^{\mp 1}=C^{\pm 1}C^{\mp 1}=D^{\pm 1}D^{\mp 1}=1, \tag{4.4}$$

$$C^{\pm 1}\ \text{central}, \tag{4.5}$$

$$[k_i,k_j]=[k_i,D]=[k_i,h_{j,r}]=0, \tag{4.6}$$

$$[h_{i,r},h_{j,s}]=\delta_{r+s,0}\frac{[ra_{ij}]_i}{r}\frac{C^r-C^{-r}}{q_j-q_j^{-1}}, \tag{4.7}$$

$$k_i x_{j,k}^{(\pm)}k_i^{-1}=q^{\pm(\alpha_i|\alpha_j)}x_{j,k}^{(\pm)}, \tag{4.8}$$

$$Dx_{i,k}^{(\pm)}D^{-1}=q^k x_{i,k}^{(\pm)},\quad Dh_{i,r}D^{-1}=q^r h_{i,r}, \tag{4.9}$$

$$[h_{i,r},x_{j,k}^{(\pm)}]=\pm\frac{[ra_{ij}]_i}{r}C^{(r\mp|r|)/2}x_{j,r+k}^{(\pm)}, \tag{4.10}$$

$$[x_{i,m}^{(+)},x_{j,n}^{(-)}]=\frac{\delta_{ij}}{q_i-q_i^{-1}}(C^{-n}\phi_{i,m+n}^{-}-C^{-m}\phi_{i,m+n}^{+}), \tag{4.11}$$

$$x_{i,m+1}^{(\pm)}x_{j,n}^{(\pm)}-q^{\pm(\alpha_i|\alpha_j)}x_{j,n}^{(\pm)}x_{i,m+1}^{(\pm)}=q^{\pm(\alpha_i|\alpha_j)}x_{i,m}^{(\pm)}x_{j,n+1}^{(\pm)}-x_{j,n+1}^{(\pm)}x_{i,m}^{(\pm)}, \tag{4.12}$$

$$\mathrm{Sym}_{m_1,\cdots,m_r}\sum_{s=0}^{r}(-1)^s\frac{[r]_i!}{[s]_i![r-s]_i!}x_{i,m_1}^{(\pm)}\cdots x_{i,m_s}^{(\pm)}x_{j,n}^{(\pm)}x_{i,m_{s+1}}^{(\pm)}\cdots x_{i,m_r}^{(\pm)}=0\quad (i\neq j,\ r=1-a_{ij}), \tag{4.13}$$

where $\phi_{i,\mp r}^{\pm}$ $(r\geqslant 0)$ is expressed in terms of $k_i^{\pm 1}$ and $h_{i,r}$'s by (4.2).

Theorem 2:[9] *For nontwisted affine Kac–Moody algebra g, there exists an algebra isomorphism $\psi:\mathscr{D}(g)\to U_q^{\Sigma}(g)$ determined by*

$$x_{i,k}^{(\pm)}\mapsto x_{i,k}^{(\pm)},\quad h_{i,r}\mapsto h_{i,r},\quad k_i^{\pm 1}\mapsto k_i^{\pm 1},\quad C^{\pm 1}\mapsto k_{\pm\delta},\quad D^{\pm 1}\mapsto k_{\pm\Lambda_0/d_0}.$$

B. Multiplicative formula of Θ

In this subsection, we shall write down the multiplicative formula of Θ, following the approach by Beck[9,13] and Lusztig.[14] Most of the necessary results have already been obtained by them.

Firstly we shall describe their results. Fix $x=\omega_{j_K}^\vee\cdots\omega_{j_1}^\vee\in\overline{Q^\vee}\subset W$ such that $(x|\alpha_i)>0$ $(1\leqslant i\leqslant l)$. Let $s_{i_L}\cdots s_{i_1}$ be one of its reduced expressions obtained by the concatenation of reduced presentations of $\omega_{j_k}^\vee$'s and canceling the elements of Π. Let us define β_k $(k\in\mathbf{Z})$ by

$$\beta_k = s_{i_1}\cdots s_{i_{k-1}}(\alpha_{i_k}) \ \ (1\leqslant k\leqslant L), \quad \beta_{k\pm rL} = \pm x^{\mp r}(\beta_k) \ \ (1\leqslant k\leqslant L,\ r>0).$$

Then they are distinct and run over the whole set of positive real roots of g. Moreover β_k is a positive (resp., negative) root of $\overline{g}$ mod $\mathbf{Z}\delta$ if $k>0$ (resp., $k\leqslant 0$). We define $E_k \in U^+_{\beta_k}$ and $F_k \in U^-_{-\beta_k}$ $(k\in\mathbf{Z})$ by

$$E_{k+rL} = T_x^{-r}T_{i_1}^{-1}\cdots T_{i_{k-1}}^{-1}(e_{i_k}) \ \ (1\leqslant k\leqslant L, r\geqslant 0),$$

$$E_{k-rL} = k_{\beta_{k-rL}}T_x^r T_{i_1}^{-1}\cdots T_{i_{k-1}}^{-1}(f_{i_k}) \ \ (1\leqslant k\leqslant L,\ r>0),$$

$$F_k = \Omega(E_k), \tag{4.14}$$

and set

$$E^{\mathbf{m}} = (E_1^{m_1}E_2^{m_2}\cdots)\prod_{\substack{1\leqslant i\leqslant l\\ r>0}} h_{i,r}^{m_{i,r}}(\cdots E_{-1}^{m_{-1}}E_0^{m_0}), \quad F^{\mathbf{m}} = (F_1^{m_1}F_2^{m_2}\cdots)\prod_{\substack{1\leqslant i\leqslant l\\ r>0}} h_{i,-r}^{m_{i,r}}(\cdots F_{-1}^{m_{-1}}F_0^{m_0}). \tag{4.15}$$

Here $h_{i,\pm r}\in U^{\pm}_{\pm r\delta}$ $(r>0)$ is defined in (4.2), and $\mathbf{m}=(m_k, m_{j,r})_{k\in\mathbf{Z},1\leqslant j\leqslant l,r>0}$ is a family of nonnegative integers such that they all vanish except for a finite number of them. For $k\in\mathbf{Z}$ we let $\overline{k}$ denote the integer which is between 1 and L, and equal to k mod L. Finally, let $(\ ,\ ): U^{\geqslant 0}\times U^{\leqslant 0}\to F$ be the Hopf pairing (Ref. 2 and Ref. 20, Proposition 2.1.1) determined by bilinearity and the following:

$$(k_\lambda, k_\mu) = q^{-(\lambda|\mu)}, \quad (e_i, k_\lambda) = (k_\lambda, f_i) = 0, \quad (e_i, f_j) = -\delta_{ij}/(q_i - q_i^{-1}),$$

$$(x, y_1y_2) = (\Delta(x), y_1\otimes y_2), \quad (x_1x_2, y) = (x_2\otimes x_1, \Delta(y)) \tag{4.16}$$

$$(x, x_1, x_2\in U^{\geqslant 0},\ y, y_1, y_2\in U^{\leqslant 0}).$$

Reference 13, Proposition 3 and Ref. 14, Proposition 40.2.4 give the following.

Proposition 2:[13,14]

(1) $(E^{\mathbf{m}})_{\mathbf{m}}$ and $(F^{\mathbf{m}})_{\mathbf{m}}$ form a basis of U^+ and U^-, respectively.

(2) $(E^{\mathbf{m}}, F^{\mathbf{n}}) = (\Pi_{\substack{1\leqslant i\leqslant l\\ r>0}} h_{i,r}^{m_{i,r}}, \Pi_{\substack{1\leqslant i\leqslant l\\ r>0}} h_{i,-r}^{n_{i,r}})\times\Pi_{k\in\mathbf{Z}}\delta_{m_k,n_k}(e_{i_k}^{m_k}, f_{i_k}^{m_k})$.

Therefore we have only to show the following. For $r\in\mathbf{Z}_{>0}$, let $(C_{ij}(r))_{1\leqslant i,j\leqslant l}$ be the inverse matrix of $([ra_{ij}]_i(q-q^{-1})/r(q_j-q_j^{-1}))_{1\leqslant i,j\leqslant l}$ and set $\tilde{h}_{i,-r} = \Sigma_{j=1}^l C_{ij}(r)h_{j,-r}$.

Lemma 4:

$$\left(\prod_{i,r} h_{i,r}^{m_{i,r}}, \prod_{i,r}\tilde{h}_{i,-r}^{n_{i,r}}\right) = \prod_{i,r}\delta_{m_{i,r},n_{i,r}} m_{i,r}!\left(\frac{-1}{q-q^{-1}}\right)^{m_{i,r}}.$$

Proof: Let $\langle\ ,\ \rangle: U\times U\to F(q^{1/2})$ be the Killing form introduced in Ref. 20. This form is bilinear and has the following properties:

$$\langle \mathrm{ad}(u)\cdot v_1, v_2\rangle = \langle v_1, v_2\cdot\widetilde{\mathrm{ad}}(u)\rangle \ \ (u, v_1, v_2\in U), \tag{4.17}$$

$$\langle xk_\lambda, yk_\mu\rangle = (x,y)q^{-(\lambda|\mu)/2} \ \ (x\in U^+, y\in U^-), \tag{4.18}$$

$$\langle U^{\geqslant 0}U^-_{-\mu}, U^{\leqslant 0}\rangle = \{0\} \ \ (\mu>0). \tag{4.19}$$

In particular, (4.18) implies that $\langle x, yk_\lambda\rangle$ is independent of $\lambda\in\Sigma$ if $x\in U^+$ and $y\in U^{\leqslant 0}$. Since $\tilde{h}_{i,-r}\in U^-_{-r\delta}$,[9]

$$\Delta(\tilde{h}_{i,-r}) = 1\otimes\tilde{h}_{i,-r} + \tilde{h}_{i,-r}\otimes k_{-r\delta} \ \ \mathrm{mod} \ \ \oplus_{r\delta>\beta>0}U^-_{-(r\delta-\beta)}\otimes U^-_{-\beta}U^0.$$

Noting the above and (4.7), we obtain from (4.17) with $u=\tilde{h}_{j,-s}$, $v_1=\Pi_{i,r}h_{i,r}^{m_{i,r}}$ and $v_2=\Pi_{i,r}\tilde{h}_{i,-r}^{n_{i,r}}\cdot k_\lambda$,

$$(q^{-(\lambda|s\delta)}-1)\left\{\left(\prod_{i,r}h_{i,r}^{m_{i,r}},\tilde{h}_{j,-s}\prod_{i,r}\tilde{h}_{i,-r}^{n_{i,r}}\right)\right.$$
$$\left.+\frac{m_{j,s}}{q-q^{-1}}\left(h_{j,s}^{m_{j,s}-1}\prod_{(i,r)\neq(j,s)}h_{i,r}^{m_{i,r}},\prod_{i,r}\tilde{h}_{i,-r}^{n_{i,r}}\right)\right\}=\sum_{s\delta>\beta\geqslant 0}q^{-(\lambda|\beta)}x_\beta.$$

Here x_β's are the elements of F independent of λ. Since the above equality holds for any $\lambda\in\Sigma$, the inside of the curly bracket vanishes. From this, the lemma follows.

The following is essentially due to Beck and Lusztig, cf. Refs. 8, 12, 24, 25. Set

$$\Theta^0=\exp\left(-(q-q^{-1})\sum_{\substack{1\leqslant i,j\leqslant l\\ r>0}}C_{ij}(r)h_{i,r}\otimes h_{j,-r}\right),\quad \Theta^+=\Theta_1\Theta_2\cdots,\quad \Theta^-=\cdots\Theta_{-1}\Theta_0,\tag{4.20}$$

where

$$\Theta_k=\sum_{m\geqslant 0}(-1)^m q_i^{-m(m-1)/2}\frac{(q_i-q_i^{-1})^m}{[m]_i!}E_k^m\otimes F_k^m\quad (i=i_{\bar{k}}).$$

Proposition 3:

(1) $\Theta=\Theta^+\Theta^0\Theta^-$.

(2) *Let* $i\in\{1,\ldots,l\}$. *If* $\beta_k=\alpha_i+m\delta$ $(k>0,\ m\geqslant 0)$, *then* $E_k\otimes F_k=x_{i,m}^{(+)}\otimes x_{i,-m}^{(-)}$. *If* $\beta_k=-\alpha_i+m\delta$ $(k\leqslant 0,\ m>0)$, *then* $E_k\otimes F_k=\Psi(x_{i,m}^{(-)}\otimes x_{i,-m}^{(+)})$.

Proof: (1) Since Θ_μ is the canonical element of $(\ ,\ )|_{U_\mu^+\times U_{-\mu}^-}$,[2,20] the claim follows from Proposition 2 and Lemma 4.

(2) Let x_i denote e_i or f_i. When $1\leqslant k\leqslant L$, $T_{i_1}^{-1}\cdots T_{i_{k-1}}^{-1}(x_{i_k})=T_{\omega_{j_1}^\vee}^{-1}\cdots T_{\omega_{j_{r-1}}^\vee}^{-1}\times T_{p_1}^{-1}\cdots T_{p_{s-1}}^{-1}(x_{p_s})$ for some r, s, where $\pi s_{p_t}\cdots s_{p_1}$ is the reduced expression of $\omega_{j_r}^\vee$. Since among the roots $s_{p_1}\cdots s_{p_{s-1}}(\alpha_{p_s})|1\leqslant s\leqslant t$ only $\alpha_{p_1}=\alpha_{j_r}$ has the form $\pm\alpha_i+m\delta$ $(1\leqslant i\leqslant l, m\geqslant 0)$, (2) follows from $T_{\omega_i^\vee}T_{\omega_j^\vee}=T_{\omega_j^\vee}T_{\omega_i^\vee}$ and $T_{\omega_i^\vee}x_j=x_j$ $(1\leqslant i\neq j\leqslant l)$.[9]

V. *L* OPERATORS AND DRINFELD'S GENERATORS

In the rest of the paper, we shall consider only $g=A_l^{(1)},B_l^{(1)},C_l^{(1)},D_l^{(1)}$. Let $L^\pm(z)=L^{\pm,u}(z)L^{\pm,d}(z)L^{\pm,l}(z)$ be the unique decomposition of the L operators such that $L^{\pm,a}(z)\in(M_N(F)\otimes\mathscr{A})[[z^{\pm 1}]]\simeq M_N(\mathscr{A}[[z^{\pm 1}]])$ $(a=u,d,l)$, $L^{\pm,d}(z)$ is diagonal, and $L^{\pm,u}(z)$ [resp., $L^{\pm,l}(z)$] is upper (resp., lower) triangular with identity elements on the diagonal.[10] For Θ^a $(a=\pm,0)$, define $((\Theta^a)^{\pm 1})_\mu$ $(\mu\in Q_+)$ as for Θ. Let us introduce $\hat{L}^{\pm,a}(z)\in(M_N(F)\otimes\mathscr{U})\times[[z^{\pm 1}]]$ $(a=u,d,l)$ by

$$\hat{L}^{+,u}(z)=\hat{\mathscr{L}}^{+,u}(z),\quad \hat{L}^{+,d}(z)=\hat{\mathscr{L}}^{+,d}(z).\hat{\mathscr{T}}^{-1},\quad \hat{L}^{+,l}(z)=.\hat{\mathscr{T}}\hat{\mathscr{L}}^{+,l}(z).\hat{\mathscr{T}}^{-1},$$
$$\hat{L}^{-,u}(z)=.\hat{\mathscr{T}}\hat{\mathscr{L}}^{-,u}(z).\hat{\mathscr{T}}^{-1},\quad \hat{L}^{-,d}(z)=.\hat{\mathscr{T}}\hat{\mathscr{L}}^{-,d}(z),\quad \hat{L}^{-,l}(z)=\hat{\mathscr{L}}^{-,l}(z)\tag{5.1}$$

where

$$\hat{\mathscr{L}}^{+,a}(z)=\sum_{\mu\geqslant 0}(\rho\otimes 1)(\Theta_\mu^{a'})z^{(\Lambda_0|\mu)/d_0},\quad \hat{\mathscr{L}}^{-,a}(z)=\sum_{\mu\geqslant 0}(\rho\otimes 1)(((\Theta_{21}^{a''})^{-1})_\mu)z^{-(\Lambda_0|\mu)/d_0},$$

with $u'=l''=+$, $d'=d''=0$ and $l'=u''=-$. Noting the property of β_k as a root of $\overline{g}$ mod $\mathbf{Z}\delta$ and the weight of $h_{i,r}$, we obtain the following.

Proposition 4: If we identify $\mathscr{A}$ with $\mathscr{U}$ by the map ϕ, then $L^{\pm,a}(z)=\hat{L}^{\pm,a}(z)$ $(a=u,d,l)$.

This, together with Proposition 3 and Theorem 2, gives the relation between the L operators and Drinfeld's generators. Before we state the results, we prepare some notations. For $1 \leqslant i \leqslant l$, set

$$e_i^+(z) = \sum_{m>0} x_{i,-m}^{(+)} z^m, \quad f_i^+(z) = \sum_{m \geqslant 0} x_{i,-m}^{(-)} z^m, \quad \phi_i^+(z) = \sum_{m \geqslant 0} \phi_{i,-m}^+ z^m,$$

$$e_i^-(z) = \sum_{m \geqslant 0} x_{i,m}^{(+)} z^{-m}, \quad f_i^-(z) = \sum_{m>0} x_{i,m}^{(-)} z^{-m}, \quad \phi_i^-(z) = \sum_{m \geqslant 0} \phi_{i,m}^- z^{-m}. \tag{5.2}$$

Let $\rho_{\pm 1}(x_i)$ $(x = e, f)$ denote the expressions $x_i^{\pm}$ in Sec. 2 Appendix A. Let further $\epsilon(i)$ $(1 \leqslant i \leqslant N)$ denote i for $\epsilon = +1$ and $N+1-i$ for $\epsilon = -1$. For $L^{\pm,a}(z)$ $(a = u, l)$, let $\equiv$ denote the equality modulo $\oplus'_{i,j}(FE_{ij} \otimes \mathscr{A})[[z^{\pm 1}]]$, where the sum is taken over i, j such that $\eta_i - \eta_j \neq 0, \epsilon\alpha_k$ $(1 \leqslant k \leqslant l$, $\epsilon = +1$ for $a = u$, $= -1$ for $a = l)$. Finally, define $b \in F^{\times}$ by $-o(l)b = q^{-(l+1)}$ for $g = A_l^{(1)}$, $= x/y \times q_0^{-h^{\vee}/2}$ [see (A3) for x and y] for $g = B_l^{(1)}$ and $= q_0^{-h^{\vee}/2}$ for $g = C_l^{(1)}, D_l^{(1)}$. (b is an inessential constant determined by the correspondence between the e_j and $x_{i,1}^{(-)}$).

By direct calculations, we obtain

Theorem 3: *Let g be $A_l^{(1)}$, $B_l^{(1)}$, $C_l^{(1)}$ or $D_l^{(1)}$. If we identify $\mathscr{D}(g)$ with the subalgebra of $\mathscr{A}(g)$ via $(\tau_b \circ \phi)^{-1}|_{U_q^{\Sigma}(g)} \circ \psi$, the following holds.*

(1) $g = A_l^{(1)}$:

$$\mp(L^{\pm,u}(z) - 1) \equiv (q - q^{-1}) \sum_{1 \leqslant i \leqslant l} \rho(e_i) \otimes f_i^{\pm}(q^i C^{-(1 \mp 1)/2} z),$$

$$\mp(L^{\pm,l}(z) - 1) \equiv (q - q^{-1}) \sum_{1 \leqslant i \leqslant l} \rho(f_i) \otimes e_i^{\pm}(q^i C^{-(1 \pm 1)/2} z),$$

$$L_{ii}^{\pm,d}(z) L_{i+1\,i+1}^{\pm,d}(z)^{-1} = \phi_i^{\pm}(q^i z) \quad (1 \leqslant i \leqslant l).$$

(2) $g = B_l^{(1)}, C_l^{(1)}, D_l^{(1)}$:

$$\mp(L^{\pm,u}(z) - 1) \equiv \sum_{\substack{1 \leqslant i \leqslant l-1 \\ \epsilon = \pm 1}} (q_i - q_i^{-1}) \rho_{\epsilon}(e_i) \otimes f_i^{\pm}(q_i^{\epsilon i} q_0^{-\epsilon h^{\vee}/2} C^{-(1 \mp 1)/2} z) + (q_l - q_l^{-1})$$
$$\times \begin{cases} \sum_{\epsilon = \pm 1} \rho_{\epsilon}(e_l) \otimes f_l^{\pm}(q_l^{\epsilon} C^{-(1 \mp 1)/2} z), & \text{for } g = B_l^{(1)} \\ \rho(e_l) \otimes f_l^{\pm}(C^{-(1 \mp 1)/2} z), & \text{for } g = C_l^{(1)}, D_l^{(1)}, \end{cases}$$

$$\mp(L^{\pm,l}(z) - 1) \equiv \sum_{\substack{1 \leqslant i \leqslant l-1 \\ \epsilon = \pm 1}} (q_i - q_i^{-1}) \rho_{\epsilon}(f_i) \otimes e_i^{\pm}(q_i^{\epsilon i} q_0^{-\epsilon h^{\vee}/2} C^{-(1 \pm 1)/2} z) + (q_l - q_l^{-1})$$
$$\times \begin{cases} \sum_{\epsilon = \pm 1} \rho_{\epsilon}(f_l) \otimes e_l^{\pm}(q_l^{\epsilon} C^{-(1 \pm 1)/2} z), & \text{for } g = B_l^{(1)} \\ \rho(f_l) \otimes e_l^{\pm}(C^{-(1 \pm 1)/2} z), & \text{for } g = C_l^{(1)}, D_l^{(1)}, \end{cases}$$

$$L_{\epsilon(i)\,\epsilon(i)}^{\pm,d}(z)^{\epsilon}\ L_{\epsilon(i+1)\,\epsilon(i+1)}^{\pm,d}(z)^{-\epsilon} = \phi_i^{\pm}(q_i^{\epsilon i} q_0^{-\epsilon h^{\vee}/2} z) \quad (1 \leqslant i \leqslant l-1,\ \epsilon = \pm 1),$$

$$L_{\epsilon(l)\,\epsilon(l)}^{\pm,d}(z)^{\epsilon}\ L_{\epsilon(l+1)\,\epsilon(l+1)}^{\pm,d}(z)^{-\epsilon} = \phi_l^{\pm}(q_l^{\epsilon} z) \quad (\epsilon = \pm 1), \quad \text{for } g = B_l^{(1)},$$

$$L_{ll}^{\pm,d}(z) L_{l+1\,l+1}^{\pm,d}(z)^{-1} = \phi_l^{\pm}(z), \quad \text{for } g = C_l^{(1)},$$

$$L_{l-1\,l-1}^{\pm,d}(z) L_{l+1\,l+1}^{\pm,d}(z)^{-1} = L_{ll}^{\pm,d}(z) L_{l+2\,l+2}^{\pm,d}(z)^{-1} = \phi_l^{\pm}(z) \quad \text{for } g = D_l^{(1)}.$$

Remark: In Ref. 25, another Hopf algebra structure of the quantum group found by Drinfeld is shown to be obtained by the twisting by $\mathscr{F}=(\Theta_{21}^{+})^{-1}$ (in our notation). The above correspondence between the L operators and Drinfeld's generators gives an intuitive explanation of its comultiplication formula as follows. After the twist, $\Theta^{\mathscr{F}}=\Theta^{0}\Theta^{-}\Psi(\Theta_{21}^{+})$. Define $\hat{L}^{\pm,\mathscr{F}}(z)$ for $\Theta^{\mathscr{F}}$ as before. Then $\hat{L}^{+,\mathscr{F}}(z)=\hat{L}^{+,d}(z)\hat{L}^{+,l}(z)\hat{L}^{-,l}(k_{-\delta}z)^{-1}$ and $\hat{L}^{-,\mathscr{F}}(z)=\hat{L}^{+,u}(k_{-\delta}z)^{-1}\times\hat{L}^{-,u}(z)\hat{L}^{-,d}(z)$. These have the lower and upper triangular forms, respectively, and satisfy the comultiplication formula in (3.9) for $\Delta^{\mathscr{F}}(\cdot)=\mathscr{F}\Delta(\cdot)\mathscr{F}^{-1}$. From this, the comultiplication formula for Drinfeld's generators easily follows. The above argument is formal, but the justification would be possible.

APPENDIX A: U' MODULE (ρ,V) AND w

1. N and $\eta=(\eta_1,\ldots,\eta_N)$

(1) $A_l^{(1)}$ $(l\geqslant 1)$: $N=l+1$, $\eta=\left(\epsilon_1,\cdots,\epsilon_l,-\sum_{i=1}^{l}\epsilon_i\right)$.

(2) $B_l^{(1)}$ $(l\geqslant 3)$, $A_{2l}^{(2)}$ $(l\geqslant 1)$: $N=2l+1$, $\eta=(\epsilon_1,\ldots,\epsilon_l,0,-\epsilon_l,\ldots,-\epsilon_1)$.

(3) $C_l^{(1)}$ $(l\geqslant 2)$, $A_{2l-1}^{(2)}$ $(l\geqslant 3)$, $D_l^{(1)}$ $(l\geqslant 4)$: $N=2l$, $\eta=(\epsilon_1,\ldots,\epsilon_l,-\epsilon_l,\ldots,-\epsilon_1)$.

(4) $D_{l+1}^{(2)}$ $(l\geqslant 2)$: $N=2l+2$, $\eta=(\epsilon_1,\ldots,\epsilon_l,0,-\epsilon_l,\ldots,-\epsilon_1,0)$.

Here $\epsilon_i=p\omega_1-(\alpha_1+\cdots\alpha_{i-1})$ $(1\leqslant i\leqslant l)$. (Note that in the case $g=D_l^{(1)}$, v_l and v_{l+1} can be interchanged.)

2. (ρ,V)

We give the matrix representations of e_i and f_i $(0\leqslant i\leqslant l)$ with repect to the basis $v_1,\ldots,v_N$. Below E_{ij} is a matrix unit and ${}^tf_i$ denotes the transpose of f_i. $x_i^{B_l}$ and $x_i^{C_l}$ $(x=e,f,\ 1\leqslant i\leqslant l)$ stand for the expressions x_i given for $B_l^{(1)}$ and $C_l^{(1)}$, respectively.

(1) $A_l^{(1)}$: $e_i=E_{i\,i+1}={}^tf_i$ $(1\leqslant i\leqslant l)$, $e_0=E_{N1}={}^tf_0$. (A1)

(2) $B_l^{(1)}, C_l^{(1)}, D_l^{(1)}$:

(i) $1\leqslant i<l$: $e_i=e_i^{+}+e_i^{-}$, $f_i=f_i^{+}+f_i^{-}$,

$$e_i^{+}=E_{i\,i+1}={}^tf_i^{+},\quad e_i^{-}=-E_{N-i\ N+1-i}={}^tf_i^{-}. \tag{A2}$$

(ii) $i=l,0$:

$B_l^{(1)}$: $e_l=e_l^{+}+e_l^{-}$, $f_l=f_l^{+}+f_l^{-}$, $e_0=E_{N-1\,1}-E_{N\,2}={}^tf_0$,

$$x^{-1}e_l^{+}=E_{l\,l+1}=y^{-1}{}^tf_l^{+},\quad x^{-1}e_l^{-}=-E_{l+1\,l+2}=y^{-1}{}^tf_l^{-},$$

$$x,y\in A,\ \text{such that } xy=[2]_l. \tag{A3}$$

$$C_l^{(1)}:\ e_l=E_{l\,l+1}={}^tf_l,\ e_0=E_{N1}={}^tf_0. \tag{A4}$$

$$D_l^{(1)}:\ e_l=E_{l-1\,l+1}-E_{l\,l+2}={}^tf_l,\ e_0=E_{N-1\,1}-E_{N2}={}^tf_0. \tag{A5}$$

(3) $A_{2l}^{(2)}$: $x_i=Mx_i^{B_l}M^{-1}$ $(x=e,f,\ 1\leqslant i\leqslant l)$, $e_0=E_{N1}={}^tf_0$,

$$M=\mathrm{diag}(\overbrace{1\cdots 1}^{l+1},-1,1,\ldots,(-1)^l). \tag{A6}$$

(4) $A_{2l-1}^{(2)}$: $x_i = M x_i^{C_l} M^{-1}$ $(x=e,f,\ 1\leqslant i\leqslant l)$, $e_0 = E_{N-1\,1} + E_{N2} = {}^t f_0$,

$$M = \mathrm{diag}(\overbrace{1\cdots 1}^{l+1}, -1, 1, \ldots, (-1)^{l-1}). \tag{A7}$$

(5) $D_{l+1}^{(2)}$: $x_i = x_i^{B_l}$ $(x=e,f,\ 1\leqslant i\leqslant l)$, $x^{-1}e_0 = E_{N1} + E_{N-1\,N} = y^{-1}{}^t f_0$,

$$x,y\in A, \text{ such that } xy=[2]_0. \tag{A8}$$

3. w

(1) $g = A_l^{(1)}$: $w = \sum_{\gamma\in\mathcal{J}_{l+1}} (-q)^{l(\gamma)} v_{\gamma(1)}\otimes\cdots\otimes v_{\gamma(l+1)}$ ($l(\gamma)$: length function).

(2) $g\neq A_l^{(1)}$: $w = \sum_{1\leqslant i,j\leqslant N} J_{ij} v_i\otimes v_j$,

$$J_{ij} = u_i\delta_{i+j,N+1} \text{ for } g\neq D_{l+1}^{(2)}, \quad = u_i\delta_{i+j,N} + u_N\delta_{iN}\delta_{jN} \text{ for } g = D_{l+1}^{(2)}.$$

List of $(u_1,\ldots,u_N)$

(i) $B_l^{(1)}$: $(q^{-(2l-1)},\ldots,q^{-3},q^{-1},q,q,q^3,\ldots,q^{2l-1})$

(ii) $C_l^{(1)}$: $(-q^{-l},\ldots,-q^{-2},-q^{-1},q,q^2,\ldots,q^l)$

(iii) $D_l^{(1)}$: $(q^{-(l-1)},\cdots,q^{-1},1,1,q,\cdots,q^{l-1})$

(iv) $A_{2l}^{(2)}$: $((-1)^l q^{-(2l-1)},\ldots,q^{-3},-q^{-1},q,-q,q^3,\ldots,(-1)^l q^{2l-1})$

(v) $A_{2l-1}^{(2)}$: $((-1)^l q^{-l},\ldots,q^{-2},-q^{-1},q,-q^2,\ldots,(-1)^{l-1} q^l)$

(vi) $D_{l+1}^{(2)}$: $(q^{-(2l-1)},\ldots,q^{-3},q^{-1},q,q,q^3,\ldots,q^{2l-1},-q)$.

[1] V. G. Drinfeld, ''Hopf algebras and the quantum Yang–Baxter equation,'' Sov. Math. Dokl. **32**, 254–285 (1985).
[2] V. G. Drinfeld, ''Quantum groups,'' Proceedings of the International Congress of Mathematicians, Berkeley, 1986, edited by A. M. Gleason (American Mathematical Society, Providence, 1987), Vol. 1, pp. 789–820.
[3] M. Jimbo, ''A q- difference analogue of $U(g)$ and the Yang–Baxter equation,'' Lett. Math. Phys. **10**, 63–69 (1985).
[4] V. G. Drinfeld, ''New realizations of Yangian and quantum affine algebra,'' Sov. Math. Dokl. **36**, 212–216 (1988).
[5] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, ''Quantization of Lie groups and Lie algebras,'' Leningrad Math. J. **1**, 178–201 (1989).
[6] N. Yu. Reshetikhin and M. A. Semenov-Tian-Shansky, ''Central extensions of quantum current groups,'' Lett. Math. Phys. **19**, 133–142 (1990).
[7] N. Jing, ''On Drinfeld realization of quantum affine algebras,'' preprint q-alg/9610035.
[8] S. M. Khoroshkin and V. N. Tolstoy, ''On Drinfeld's realization of quantum affine algebras,'' J. Geom. Phys. **11**, 445–452 (1993); S. M. Khoroshskin and V. N. Tolstoy, ''The Cartan–Weyl basis and the universal R matrix for quantum Kac–Moody algebras and superalgebras,'' *Quantum Symmetries*, Proceedings of the International Workshop on Mathematical Physics, Clausthal, 1991, edited by H. D. Duebner and V. K. Dobrev (World Scientific, Singapore, 1993), pp. 336–351.
[9] J. Beck, ''Braid group actions and quantum affine algebras,'' Commun. Math. Phys. **165**, 555–568 (1994).
[10] J. Ding and I. B. Frenkel, ''Isomorphism of two realizations of quantum affine algebras $U_q(\widehat{gl(n)})$, Commun. Math. Phys. **156**, 277–300 (1993).
[11] S. M. Khoroshkin, D. Lebedev, and S. Pakuliak, ''Intertwining operators for the central extension of the Yangian Double,'' preprint q-alg/9602030.
[12] V. N. Tolstoy and S. M. Khoroshkin, ''The universal R matrix for quantum nontwisted affine Kac–Moody algebras,'' Funct. Anal. Appl. **26**, 69–71 (1992); S. M. Khoroshkin and V. N. Tolstoy, ''The uniqueness theorem for the universal R matrix,'' Lett. Math. Phys. **24**, 231–244 (1992).
[13] J. Beck, ''Convex bases of PBW type for quantum affine algebras,'' Commun. Math. Phys. **165**, 193–199 (1994).
[14] G. Lusztig, *Introduction to Quantum Groups* (Birkhäuser, Boston, 1993).
[15] V. Chari and A. Pressley, ''Twisted quantum affine algebras,'' preprint q-alg/9611002.
[16] V. G. Kac, *Infinite dimensional Lie algebras*, 3rd ed. (Cambridge University, Cambridge, 1990).

[17] M. Rosso, ''Finite dimensional representations of the quantum analog of the enveloping algebra of a complex simple Lie algebra,'' Commun. Math. Phys. **117**, 581–593 (1988).
[18] G. Lusztig, ''Quantum deformations of certain simple modules over enveloping algebras,'' Adv. Math. **70**, 237–249 (1988).
[19] M. Jimbo, ''Quantum R for the generalized Toda system,'' Commun. Math. Phys. **102**, 537–548 (1986).
[20] T. Tanisaki, ''Killing forms, Harish–Chandra isomorphisms and universal R-matrices for quantum algebras,'' Int. J. Mod. Phys. A **7**, Suppl. 1B, 941–961 (1992).
[21] A. A. Belavin and V. G. Drinfeld, ''Solutions of the classical Yang–Baxter equations for simple Lie algebra,'' Funct. Anal. Appl. **16**, 159–180 (1982).
[22] G. Lusztig, ''Finite dimensional certain Hopf algebras arising from quantized universal enveloping algebras,'' J. Am. Math. Soc. **3**, 257–296 (1990).
[23] I. B. Frenkel and N. Yu. Reshetikhin, ''Quantum affine algebras and holonomic difference equations,'' Commun. Math. Phys. **146**, 1–60 (1992).
[24] S. Levendoroskii, Y. Soibelman, and V. Stukopin, ''Quantum Weyl groups and universal quantum R matrix for affine Lie algebra $A_1^{(1)}$,'' Lett. Math. Phys. **27**, 253–264 (1993).
[25] S. M. Khoroshkin and V. N. Tolstoy, ''Twisting of quantum (super) algebras. Connection of Drinfeld's and Cartan–Weyl realization for quantum affine algebras,'' preprint hep-th/94004036.

Transformations of strongly causal space–times preserving null geodesics

Wen-ling Huang
Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

(Received 11 September 1997; accepted for publication 11 November 1997)

Let $(\mathscr{M},g)$ be a strongly causal space–time, $\dim\mathscr{M}\geqslant 3$. Let $f:\mathscr{M}\to\mathscr{M}$ be a bijection such that images and pre-images of null geodesics (as point sets) are null geodesics. Then f is a conformal transformation. © *1998 American Institute of Physics.* [S0022-2488(98)02603-6]

I. INTRODUCTION

A conformal diffeomorphism of an n-dimensional, strongly causal space–time $(\mathscr{M},g)$, $n\geqslant 3$, *can be characterized as a homeomorphism of* $(\mathscr{M},g)$ *which takes null geodesic curves (as point sets) to null geodesic curves.*

This theorem has been shown by Hawking.[1] Peleska has generalized this result of Hawking to arbitrary Pseudo-Riemannian manifolds.[2] According to a theorem of Alexandrov,[3,4] the hypothesis of homeomorphy can be omitted in Minkowski space.

Any bijective transformation f of at least three-dimensional Minkowski space M which preserves the Minkowski distance 0 *in both directions is a conformal mapping of Minkowski space.*

It is a natural question whether the homeomorphy condition in Hawking's theorem is necessary or not. Lester has examined Alexandrov-type transformations for Robertson–Walker space–times,[5] for Einstein's cylinder universe,[6] and for de Sitter space–time.[7] In Schwarzschild space–time the following result is known.[8]

Let $n\geqslant 3$. *Let f be a bijective mapping from n-dimensional Schwarzschild space–time to itself such that f and* f^{-1} *preserve inextendible null geodesic curves (as point sets). Then f is an isometry.*

We show that the homeomorphy condition in Hawking's theorem can be omitted if we demand that also the inverse mapping takes null geodesic curves (as point sets) to null geodesic curves.

Theorem: *Let* $(\mathscr{M},g)$ *be a strongly causal space–time,* $\dim\mathscr{M}\geqslant 3$. *Let* $f:\mathscr{M}\to\mathscr{M}$ *be a bijection which satisfies the following condition.*

For any null geodesic curve $\gamma:\mathscr{E}\to\mathscr{M}$ *there are null geodesic curves* $\gamma_1:\mathscr{E}_1\to\mathscr{M}$ *and* $\gamma_2:\mathscr{E}_2\to\mathscr{M}$ *such that*

$$f(\gamma(\mathscr{E}))=\gamma_1(\mathscr{E}_1) \quad \text{and} \quad f^{-1}(\gamma(\mathscr{E}))=\gamma_2(\mathscr{E}_2).$$

Then f is a homeomorphism and, by Hawking's theorem, a conformal transformation.

The physical meaning of the condition used in this theorem to characterize conformal transformations is clear: *Images and pre-images of paths which photons travel between emission and absorption should again be such paths.*

II. PRELIMINARIES

Let $n\in\mathbb{N}$, $n\geqslant 2$. A *space–time* $(\mathscr{M},g)$ is an n-dimensional, time oriented, connected C^∞-Hausdorff manifold with a Lorentz metric g (i.e., the signature of g is $n-2$) together with the Levi-Civita connection of g on $\mathscr{M}$.

In a space–time $(\mathscr{M},g)$ a non-zero tangent vector X at $x\in\mathscr{M}$ is called *timelike*, *null*, or *spacelike* if $g_x(X,X)$ is negative, zero, or positive, respectively. A future directed timelike (non-spacelike) curve is a C^1-curve $\gamma:\mathscr{E}\to\mathscr{M}$ whose tangent vectors are all timelike (non-spacelike) and future directed. Here and henceforth $\mathscr{E}$ is an open, not necessarily bounded interval of $\mathbb{R}$.

 © 1998 American Institute of Physics

Let $U\subseteq\mathscr{M}$ and $x\in\mathscr{M}$. A point $y\in\mathscr{M}$ is said to be in the *chronological future* $I_U^+(x)$ of x relative to U, $x\ll_U y$, if there is a future directed timelike curve section $\gamma:[0,1]\to U$ connecting $x=\gamma(0)$ and $y=\gamma(1)$. If $x\ll_U y$, then x is in the *chronological past* $I_U^-(y)$ of y relative to U. The point y is said to be in the *causal future* $J_U^+(x)$ of x relative to U, and x is in the *causal past* $J_U^-(y)$ of y relative to U, $x\leqslant_U y$, if there is a future directed non-spacelike curve section $\gamma:[0,1]\to U$ connecting $x=\gamma(0)$ and $y=\gamma(1)$. In this paper we write $x\to_U y$ if there is a future directed **null geodesic** in U from $x\in U$ to $y\in U$. (This definition is different from that used in Ref. 9, p. 182.) Let $C_U^+(x):=\{y\in\mathscr{M}|x\to_U y\}$ and $C_U^-(x):=\{y\in\mathscr{M}|y\to_U x\}$. If $U=\mathscr{M}$ we simply write $I^+(x)$, $x\ll y$, etc. A space–time is called *strongly causal*, if the set of all $I^+(x)\cap I^-(y)$, $x,y\in\mathscr{M}$, is a basis of the ordinary topology of $(\mathscr{M},g)$. In a strongly causal space–time $(\mathscr{M},g)$, each point $x\in\mathscr{M}$ possesses a convex normal neighborhood U such that no non-spacelike curve intersects U more than once.

For $x\in\mathscr{M}$, $X\in T_x(M)$ there is a unique inextendible geodesic γ_X with $\gamma_X(0)=x$, $\dot{\gamma}_X(0)=X$. If $\gamma_X(1)$ exists, we write $\exp_x X:=\gamma_X(1)$ (Ref. 9, p. 33). An open neighborhood $N_x\subset T_xM$ of the origin in the tangent space T_xM of $x\in\mathscr{M}$ is said to be *normal* if $\exp_x|_{N_x}$ is defined and is a diffeomorphism onto its image. We call the image $U_x:=\exp_x(N_x)$ a normal neighborhood of x. If a normal neighborhood U_x is a normal neighborhood for any point $y\in U_x$, then we call it a *convex normal* neighborhood. Hence, in a convex normal neighborhood U, any two points $x,y\in U$ can be joined by a unique inextendible geodesic $\gamma:\mathscr{E}\to U$, $\gamma(0)=x$, $\gamma(1)=y$, (Ref. 10, p. 149). In a convex normal neighborhood U we define the *energy function* $\Phi:U\times U\to\mathbb{R}$,

$$\Phi(x,y):=g_x(\exp_x^{-1}y,\exp_x^{-1}y).$$

$\Phi(x,y)$ is negative, zero, or positive if the unique geodesic γ connecting x and y is timelike, lightlike, or spacelike, respectively.

Proposition 1: Let $(\mathscr{M},g)$ be a space–time. Let $U\subseteq\mathscr{M}$ be open and $x,y,z\in U$. Then either $x\ll_U y\leqslant_U z$ or $x\leqslant_U y\ll_U z$ imply $x\ll_U z$ (cf. Ref. 9, p. 183).

Proposition 2: Let $U\subseteq\mathscr{M}$ be a convex normal neighborhood of $x\in\mathscr{M}$. Let $\gamma:(a,c)\to U$ be a future directed null geodesic such that $x\notin\gamma((a,c))$. There is at most one point $y=\gamma(b)$, $b\in(a,c)$ with $\Phi(x,y)=0$, i.e., there is at most one null geodesic in U which contains x and intersects $\gamma((a,c))$. If y exists, then the set of points in $\gamma((a,c))$ which are chronologically related to x is $\gamma((a,b))$ if $y\to_U x$, and $\gamma((b,c))$ if $x\to_U y$.

Proof: Define $\Phi_x:U\to\mathbb{R}$, $\Phi_x(z):=\Phi(x,z)$ for all $z\in U$. $\Phi_x\circ\gamma:(a,c)\to\mathbb{R}$ is a C^∞-function, and $(\Phi_x\circ\gamma)'(t)=2\,g_{\gamma(t)}(\dot{\gamma}(t),\dot{\alpha}(1))$ where $\alpha:[0,1]\to U$ denotes the geodesic section from $x=\alpha(0)$ to $\gamma(t)=\alpha(1)$ (Ref. 11, p. 127). If there is a $b\in(a,c)$ with $\Phi_x\circ\gamma(b)=0$, then

$$(\Phi_x\circ\gamma)'(b)=2\,g_{\gamma(b)}(\dot{\gamma}(b),\dot{\alpha}(1))\begin{cases}<0, & \text{if } \alpha \text{ is future directed,}\\ >0, & \text{if } \alpha \text{ is past directed.}\end{cases}$$

Since there are no null geodesic triangles in a convex normal neighborhood, $\Phi_x\circ\gamma$ is 0 only at b. In the case $(\Phi_x\circ\gamma)'(b)<0$, i.e., $x\to_U y$, we have $\Phi_x\circ\gamma(t)<0$ for all $t\in(b,c)$, and $\Phi_x\circ\gamma(t)>0$ for all $t\in(a,b)$. In the case $(\Phi_x\circ\gamma)'(b)>0$, i.e., $y\to_U x$, we have $\Phi_x\circ\gamma(t)<0$ for all $t\in(a,b)$, and $\Phi_x\circ\gamma(t)>0$ for all $t\in(b,c)$. □

Proposition 3: Let $U\subseteq\mathscr{M}$ be a convex normal neighborhood.Let $x,y\in U$ be two points with $x\to_U y$. Let α,β be two future directed inextendible null geodesics in U with $x=\alpha(0)$ and $y=\beta(0)$ such that y, resp., x is not contained in the image of α resp. β. Then there are two points $\alpha(t_1)$, $\beta(t_2)$ in U, $0<t_1,t_2$, so that $\alpha(t_1)\to_U\beta(t_2)$ and the map $\lambda:(0,t_1)\to(0,t_2)$ defined by $\alpha(t)\to_U\beta(\lambda(t))$ is a bijection (figure 1).

Proof: It follows from Proposition 2 that $\Phi(y,\alpha(t))>0$ and $\Phi(x,\beta(t))<0$ for all $t>0$. Choose arbitrary $t_\alpha,t_\beta>0$ in the domains of α and β. If $\Phi(\alpha(t_\alpha),\beta(t_\beta))=0$ let $t_1:=t_\alpha$ and $t_2:=t_\beta$; then $\alpha(t_1)\to_U\beta(t_\beta)$.

If $\Phi(\alpha(t_\alpha),\beta(t_\beta))>0$ then there is some $t_1\in(0,t_\alpha)$ satisfying $\Phi(\alpha(t_1),\beta(t_\beta))=0$. Since $x\to_U y$ and since α,β are both future directed, we have $\alpha(t_1)\to_U\beta(t_2)$ for $t_2:=t_\beta$. If

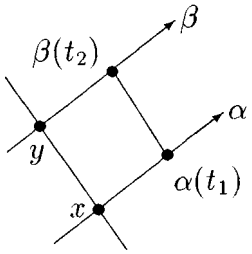

FIG. 1. Proposition 3.

$\Phi(\alpha(t_\alpha),\beta(t_\beta))<0$ then there is some $t_2\in(0,t_\beta)$ satisfying $\Phi(\alpha(t_\alpha),\beta(t_2))=0$, and we obtain $\alpha(t_1)\rightarrow_U\beta(t_2)$ for $t_1:=t_\alpha$. Therefore, we have shown the existence of $t_1,t_2>0$ with $\alpha(t_1)\rightarrow_U\beta(t_2)$.

Let $t\in(0,t_1)$. Since $x\rightarrow_U y$ and $x\rightarrow_U\alpha(t)$, we obtain from Proposition 2 that $\Phi(\alpha(t),\beta(t_2))<0$ and $\Phi(\alpha(t),y=\beta(0))>0$. $\Phi(\alpha(t),\beta(\cdot))$ is continuous, hence there is a $\bar{t}\in(0,t_2)$ with $\Phi(\alpha(t),\beta(\bar{t}))=0$. From Proposition 2, $\bar{t}$ is uniquely determined. Similarly, for $\bar{t}\in(0,t_2)$, we obtain $\Phi(\alpha(t_1),\beta(\bar{t}))>0$ and $\Phi(\alpha(0),\beta(\bar{t}))<0$, and there is a $t\in(0,t_1)$ which satisfies $\Phi(\alpha(t),\beta(\bar{t}))=0$. Hence the mapping $\lambda:(0,t_1)\rightarrow(0,t_2)$ is a bijection. □

Corollary 1: Let $x,z\in U$ be two points in a convex normal neighborhood U with $x\ll_U z$. If there exists some $y\in U$ satisfying $\Phi(x,y)=\Phi(y,z)=0$, then, from Proposition 2, we have $x\rightarrow_U y\rightarrow_U z$.

Proposition 4: Let $(\mathscr{M},g)$ be strongly causal. Let $\gamma:\mathscr{E}\rightarrow\mathscr{M}$ be an inextendible null geodesic, and let $\alpha:\mathscr{E}'\rightarrow\mathscr{M}$ be a null geodesic such that $\alpha(\mathscr{E}')\subset\gamma(\mathscr{E})$. Then there exists an interval $\mathscr{E}_1\subset\mathscr{E}$ with $\gamma(\mathscr{E}_1)=\alpha(\mathscr{E}')$.

Proof: Choose an arbitrary point $x\in\alpha(\mathscr{E}')$. Let U be a sufficiently small normal neighborhood of x such that α and γ intersect U only once, i.e., there are intervals $I\subset\mathscr{E}'$ and $J\subset\mathscr{E}$ with $\alpha(I)=\alpha(\mathscr{E}')\cap U$, $\gamma(J)=\gamma(\mathscr{E})\cap U$, and such that $\alpha(I)=\gamma(J)$. Since U is normal, $\alpha|_J$ and $\gamma|_J$ are the same null geodesics up to an affine reparametrization. Let t,s be affine parameters for α and γ, respectively. Then there are $k_1\neq0$, $k_2\in\mathbb{R}$ so that $k_1t+k_2=s$ and $\alpha(t)=\gamma(k_1t+k_2)$ for all $t\in\mathscr{E}'$. Hence $\alpha(\mathscr{E}')=\gamma(\mathscr{E}_1)$ for $\mathscr{E}_1:=\{k_1t+k_2\mid t\in\mathscr{E}'\}$. □

III. PROOF OF THE THEOREM

(3.1) Let $\alpha:\mathscr{E}_1\rightarrow\mathscr{M}$, $\beta:\mathscr{E}_2\rightarrow\mathscr{M}$ be two inextendible null geodesics with affine parameters t,s, respectively, and satisfying $f\circ\alpha(\mathscr{E}_1)=\beta(\mathscr{E}_2)$. Then the mapping $\mathscr{E}_1\rightarrow\mathscr{E}_2$, $t_0\mapsto s_0$ defined by $f\circ\alpha(t_0)=\beta(s_0)$ is monotone.

Proof: Since $(\mathscr{M},g)$ is strongly causal, α and β are one to one. For $t_1<t_2<t_3$, $t_i\in\mathscr{E}_1$, let $s_i\in\mathscr{E}_2$ satisfy $\beta(s_i)=f\circ\alpha(t_i)$, $i=1,2,3$. Let $I\subset\mathscr{E}_1$ be an open interval containing t_1 and t_2 but not t_3. Since $f\circ\alpha|_I$ is a null geodesic, from Proposition 4 there is an interval $J\subset\mathscr{E}_2$ such that $f\circ\alpha(I)=\beta(J)$. Since β is one to one, we have $s_1,s_2\in J$, $s_3\notin J$. Hence $s_1<s_2<s_3$ or $s_3<s_2<s_1$.

From (3.1) follows immediately (3.2).

(3.2) Let $\alpha:\mathscr{E}_1\rightarrow\mathscr{M}$, $\beta:\mathscr{E}_2\rightarrow\mathscr{M}$ be two future directed inextendible null geodesics with $f\circ\alpha(\mathscr{E}_1)=\beta(\mathscr{E}_2)$. Let $x,y\in\alpha(\mathscr{E}_1)$ be two points, $x\rightarrow y$. Let $z\in\alpha(\mathscr{E}_1)$, $x\rightarrow z$. If $f(x)\rightarrow f(y)$, then $f(x)\rightarrow f(z)$; and $f(z)\rightarrow f(x)$ otherwise.

(3.3) Let $x\in\mathscr{M}$ be a point. Let U, resp. V, be convex normal neighborhoods of x, resp., $f(x)$ such that no non-spacelike curve intersects U, resp. V, more than once. Let $X\in T_x(\mathscr{M})$ be a null vector with $y:=\exp_x X\in U$. If $\exp_{f(x)}Y:=f(y)\in V$, $Y\in T_{f(x)}(\mathscr{M})$, then Y is a null vector and

$$\{f\circ\exp_x(tX)\mid t\in[0,1]\}=\{\exp_{f(x)}(tY)\mid t\in[0,1]\}. \tag{1}$$

Proof: Since f takes null geodesics to null geodesics, there is a null geodesic $\gamma:\mathscr{E}\rightarrow\mathscr{M}$ connecting $f(x)$ and $f(y)$. Let I be an interval with $\gamma(I)=\gamma(\mathscr{E})\cap V$. Since V is convex normal

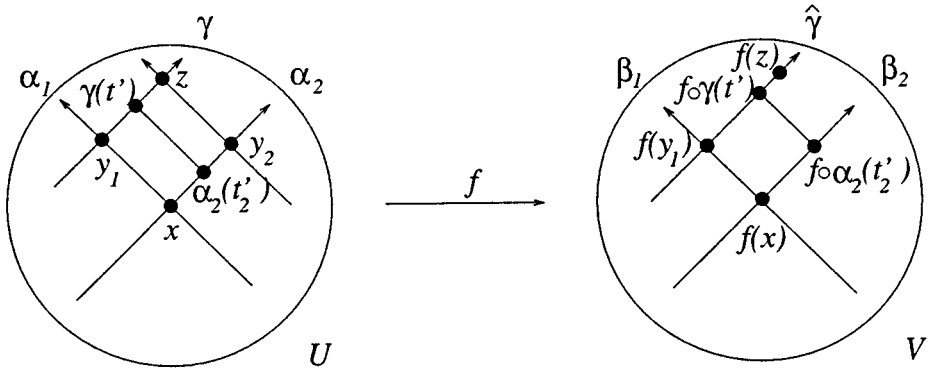

FIG. 2. Proof of (3.4).

we can choose an interval J such that $\gamma|_I: I \to V$ and $\exp_{f(x)}(tY): J \to V$ are the same geodesics, up to a reparametrization. Therefore Y is a null vector. From (3.1) we obtain (1).

(3.4) Let $x \in \mathscr{M}$. Then $\{f(C^+(x)), f(C^-(x))\} = \{C^+(f(x)), C^-(f(x))\}$.

Proof: Choose a convex normal neighborhood U of x such that no non-spacelike curve intersects U more than once.

We take two future directed null geodesics α_1, α_2 with $x = \alpha_1(0) = \alpha_2(0)$ so that $f \circ \alpha_1(t)$, $f \circ \alpha_2(t)$ are both contained in $C^+(f(x))$ or in $C^-(f(x))$ for $t > 0$. Without loss of generality we may assume that $f \circ \alpha_i(t) \in C^+(f(x))$, $i = 1,2$. Let V be a convex normal neighborhood of $f(x)$ such that no non-spacelike curve intersects V more than once. There are intervals I_i with $0 \in I_i$ and $\alpha_i(I_i) \subset U$, $f \circ \alpha_i(I_i) \subset V$, $i = 1,2$. Let $\beta_i: J_i \to V$ be future directed null geodesics with $\beta_i(J_i) \subset V$ and $f \circ \alpha_i(I_i) = \beta_i(J_i)$, $i = 1,2$. Let $0 < t_1 \in I_1$ and $y_1 := \alpha_1(t_1)$. Let γ be a future directed null geodesic from $y_1 = \gamma(0)$ such that $\dot{\gamma}(0)$, $\dot{\alpha}_1(t_1)$ are linearly independent. Then there is an interval I with $\gamma(I) \subset U$ and $f \circ \gamma(I) \subset V$. Let $\hat{\gamma}: J \to U$ denote a future directed null geodesic with $\hat{\gamma}(J) = f \circ \gamma(I)$ and $\hat{\gamma}(0) = y_1$. According to Proposition 3, there are $0 < t_0 \in I$ and $0 < t_2 \in I_2$ such that $z := \gamma(t_0)$ and $y_2 := \alpha_2(t_2)$ are connected by a null geodesic in U. For $0 < t' < t_0$ there is a $0 < t_2' < t_2$ such that $\gamma(t')$, $\alpha_2(t_2')$ are connected by a null geodesic in U. Then $f \circ \gamma(t') =: \hat{\gamma}(s_1)$ and $f \circ \alpha_2(t_2') =: \beta_2(t_2^*)$ are connected by a null geodesic in V (see figure 2).

We claim $f(z) \in C_V^+(f(y_1))$. If not, then $s_2 < s_1 < 0$, where s_2 is defined by $\hat{\gamma}(s_2) := f(z)$. This implies $\Phi(\beta_2(t_2^*), \hat{\gamma}(s_2)) > 0$, $\Phi(\beta_2(t_2^*), \hat{\gamma}(0)) > 0$ and thus $\Phi(\beta_2(t_2^*), \hat{\gamma}(s_1)) > 0$, a contradiction to $\Phi(f \circ \alpha_2(t_2'), f \circ \gamma(t')) = \Phi(\beta_2(t_2^*), \hat{\gamma}(s_1)) = 0$. So $f(z) \in C_V^+(f(y_1)) \subset J_V^+(f(x))$.

Let α_3 be a future directed null geodesic with $\alpha_3(0) = x$ and such that $\dot{\alpha}_3(0)$, $\dot{\alpha}_1(0)$ are linearly independent. Let I_3 be an interval with $x \in \alpha_3(I_3) \subset U$ and $f \circ \alpha_3(I_3) \subset V$. From Proposition 3, we can find $y_3 \in \alpha(I_3)$, $x \to y_3$, and $z^* \in \gamma(I)$, $y_1 \to z^*$, such that $y_3 \to z^*$. From $f(z^*) \in I_V^+(f(x))$ and Corollary 1 follows $f(y_3) \in C_V^+(f(x))$. We obtain from (3.2) that $f(C^+(x)) \subseteq C^+(f(x))$. Since f^{-1} has the same properties as f, we have $f(C^+(x)) = C^+(f(x))$. It is now obvious that $f(C^-(x)) = C^-(f(x))$.

(3.5) Let $x \in \mathscr{M}$ and $\{\epsilon, \epsilon'\} = \{-, +\}$ such that $f(C^+(x)) = C^\epsilon(f(x))$. Then $f(I^+(x)) = I^\epsilon(f(x))$ and $f(I^-(x)) = I^{\epsilon'}(f(x))$.

Proof: Assume $f(C^+(x)) = C^+(f(x))$. For $y \in I^+(x)$ there are $x_0, \ldots, x_k$, $k \geqslant 2$,

$$x = x_0 \to x_1 \to \cdots \to x_k = y,$$

and we can choose x_1, x_2 such that x_0, x_1, x_2 are not joined by a null geodesic. Then

$$f(x) = f(x_0) \to f(x_1) \to \cdots \to f(x_k) = f(y),$$

and $f(x_0), f(x_1), f(x_2)$ are not joined by a null geodesic. $f(x_0)$ and $f(x_2)$ can be joined by a timelike curve (cf. Ref. 9, p. 113). Hence $f(x_0) \ll f(x_2) \leqslant f(y)$. From Proposition 1 we have

$f(y)\in I^+(f(x))$ and thus $f(I^+(x))\subset I^+(f(x))$. Applying the same arguments to f^{-1}, we obtain $f(I^+(x))=I^+(f(x))$. The case $f(C^+(x))=C^-(f(x))$ is treated similarly.

(3.6) Let $x\in\mathscr{M}$ and $\{\epsilon,\epsilon'\}=\{-,+\}$ such that $f(I^+(x))=I^\epsilon(f(x))$. Then $f(I^-(y))=I^{\epsilon'}(f(y))$ for all $y\in J(x)$.

Proof: We have $J(x)=I(x)\cup C(x)\cup\{x\}$ (cf. Ref. 9, p. 184). (3.4) and (3.5) imply $f(J^+(x))=J^\epsilon(f(x))$ and $f(J^-(x))=J^{\epsilon'}(f(x))$. Since $x\in J(y)$, applying (3.4) and (3.5) again yields $f(I^-(y))=I^{\epsilon'}(f(y))$.

(3.7) f is a homeomorphism.

Proof: Let $x,y\in\mathscr{M}$. If $y\notin J(x)$, then $I^+(x)\cap I^-(y)=\varnothing$. If $y\in J(x)$, let $f(I^+(x))=I^\epsilon(f(x))$, $\{\epsilon,\epsilon'\}=\{-,+\}$. Then $f(I^-(y))=I^{\epsilon'}(f(y))$, and $f(I^+(x)\cap f(I^-(y))=I^\epsilon(f(x))\cap I^{\epsilon'}(f(y))$. Thus the basis $\{I^+(x)\cap I^-(y)\mid x,y\in\mathscr{M}\}$ of the topology of $\mathscr{M}$ is preserved under f and f^{-1}. □

ACKNOWLEDGMENT

Financial support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

[1] S. W. Hawking, A. R. King, and P. J. McCarthy, J. Math. Phys. **17**, 174–181 (1976).
[2] J. Pekeska, Aequationes Math. **27**, 20–31 (1984).
[3] A. D. Alexandrov, Usp. Mat. Nauk **37**, 187 (1950).
[4] A. D. Alexandrov, Can. J. Math. **19**, 1119–1128 (1967).
[5] J. A. Lester, Aequationes Math. **25**, 216–232 (1982).
[6] J. A. Lester, C. R. Math. Rep. Acad. Sci. Can. **IV**, 175–178 (1982).
[7] J. A. Lester, Abh. Math. Sem. Univ. Hamburg **53**, 217–224 (1983).
[8] W.-l. Huang, Hamburger Beiträge zur Mathematik aus dem Mathematischen Seminar, 1996, p. 46.
[9] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[10] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963 and 1969), Vol. I.
[11] R. Sachs and H.-H. Wu, *General Relativity for Mathematicians*, Graduate Texts in Mathematics (Springer-Verlag, New York, 1977), p. 48.

The spherical curl transform of a linear force-free magnetic field

Malcolm A. MacLeod
Air Force Research Laboratory, Hanscom Air Force Base, Massachusetts 01731-3010

(Received 18 October 1996; accepted for publication 28 October 1997)

The description of linear force-free magnetic fields in terms of the Moses eigenfunctions of the curl operator begun previously is here completed by the derivation of a general expression for the field's spherical curl transform. This enables the transform space representation of a given field to be determined and compared with that of other fields, assisting the analysis and classification of this type of magnetic field as well as providing a basis for generalization. The result obtained gives the spherical curl transform as a weighted projection of the vector Radon transform of the field on the appropriate curl eigenvector. The process is exemplified by the determination of the transforms of three fields: the simplest force-free magnetic field, and the Lundquist and classical spheromak fields. The latter two are both of interest as models of the magnetic fields of solar magnetic clouds, while the classical spheromak field is relevant to the design of nuclear fusion reactors as well. The use of the transform in generalizing the Lundquist field is briefly discussed. As before, all results apply equally well to the description of the Trkalian subset of Beltrami fields in fluid dynamics. © *1998 American Institute of Physics.* [S0022-2488(98)01103-7]

I. INTRODUCTION

Recent work has led to a new mathematical description of linear force-free magnetic fields (FFMFs), one that provides a unifying scheme for exploring this distinct class of fields as well as providing fresh insight into the solution of practical problems. Both of these characteristics will be illustrated below, but the primary goal of the present work is to present an improved method for determining the spherical curl transform of a given force-free magnetic field.

The essence of the new description is that a general linear FFMF is defined entirely in terms of a scalar function on the unit sphere in transform space, reducing a three-dimensional vector problem in physical space effectively to a two-dimensional scalar problem in transform space [MacLeod,[1] referred to as (I) below]. Of course, to obtain the physical field requires performing a specific vector transform whose mathematical evaluation is typically nontrivial, but the *spherical curl transform* (SCT) defined on the sphere in transform space holds the key to the general linear FFMF. In (I) the transform was considered to be given and a discussion of the various types of resulting fields was sketched. In a subsequent work [MacLeod,[2] referred to as (II) below] the inverse case was considered, resulting in a workman-like, but not entirely satisfactory, determination of the SCT from a given field. The formula presented here for the spherical curl transform completes the new description of these fields by providing a satisfactory general inverse.

The basic idea is to expand the general linear FFMF in terms of the eigenfunctions of the curl operator introduced by Moses,[3] leading to the discovery that such a field requires only one curl eigenfunction whose helicity sense matches the sign of the parameter α appearing in the defining system of the FFMF. The curl eigenfunction expansion of a general vector field is outlined in Sec. II and the required direct and inverse transforms are introduced in order to define and clarify terms. This expansion is specialized to describe linear FFMFs in Sec. III, and the spherical curl transform introduced and defined. The general expression for the SCT is derived in Sec. IV in terms of a weighted projection of the vector Radon transform of the field on the appropriate curl eigenvector, and two alternative expressions for the inverse SCT are shown to be equivalent and to produce the original field. To illustrate the process and its results, the derivation of specific SCTs is carried out in Sec. V for three FFMFs: the simplest FFMF, the Lundquist field, and the

 © 1998 American Institute of Physics

classical spheromak field. While the first of these computations is intended merely to exemplify the process without obscuring it with mathematical details, the latter two computations treat fields of current practical interest. Both the Lundquist field (Burlaga[4]) and the classical spheromak (Ivanov and Harshiladze[5]) have been proposed as models of the magnetic fields of solar magnetic clouds (SMCs), while the latter is still of some interest in the magnetic fusion community (Wright[6]). Moreover, the development of the Lundquist field SCT led directly to a generalization that is being investigated as a more flexible fit to the fields of SMCs (MacLeod[7]). In the final section the main points are recapitulated and several potentially fruitful directions for further work outlined.

II. REPRESENTING A GENERAL VECTOR FIELD BY CURL EIGENFUNCTION EXPANSION

Moses[3] showed that a general (three-dimensional) vector field $\mathbf{F}(\mathbf{x})$ could be represented as an expansion in terms of the three eigenfunctions of the curl operator,

$$\boldsymbol{\chi}_\lambda(\mathbf{x}|\mathbf{k}) \equiv (2\pi)^{-3/2} e^{i\mathbf{k}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\mathbf{k}), \tag{2.1}$$

where the curl eigenvectors $\mathbf{Q}_\lambda(\mathbf{k})$ are given explicitly in Appendix A and the index λ takes on the values $(0,\pm 1)$ reflecting the sense of the helicity of the corresponding eigenfunction. Such an expansion has the form

$$\mathbf{F}(\mathbf{x}) = \sum_\lambda \int d^3k \; \boldsymbol{\chi}_\lambda(\mathbf{x}|\mathbf{k}) f_\lambda(\mathbf{k}), \tag{2.2}$$

which is fundamentally a type of Fourier transform. The vector transform of the original function may be obtained explicitly by multiplying (2.2) by an appropriate exponential and integrating over space:

$$f(\mathbf{k}') \equiv \sum_\lambda \mathbf{Q}_\lambda(\mathbf{k}') f_\lambda(\mathbf{k}') = (2\pi)^{-3/2} \int d^3x \; \mathbf{F}(\mathbf{x}) e^{-i\mathbf{k}'\cdot\mathbf{x}}. \tag{2.3}$$

We consider the spatial field $\mathbf{F}(\mathbf{x})$ to be primary here, so that (2.3) represents the *direct* transform of the field, while (2.2) is the *inverse* transform of the field. So far this is standard Fourier fare. If we now take the scalar product of (2.3) with the conjugate curl eigenvector $\mathbf{Q}^*_\mu(\mathbf{k}')$ and use the orthogonality of the curl eigenvectors, we generate the components of the *curl transform* of $\mathbf{F}(\mathbf{x})$:

$$f_\mu(\mathbf{k}') = \mathbf{Q}^*_\mu(\mathbf{k}') \cdot \mathbf{f}(\mathbf{k}') = (2\pi)^{-3/2} \mathbf{Q}^*_\mu(\mathbf{k}') \cdot \int d^3x \; \mathbf{F}(\mathbf{x}) e^{-i\mathbf{k}'\cdot\mathbf{x}}, \tag{2.4}$$

which in the most general case has three nonzero components corresponding to a field possessing divergence and two nonaligned helicity components ($\mu = 0, +1, -1$). Thus, the μ component of the curl transform is the scalar product of the μ curl eigenvector with the Fourier transform of the field. Note that the curl transform is a *scalar* quantity derived from a *vector* field. The inverse curl transform is (2.2), with three vector components indexed by λ corresponding to the terms of the summation:

$$\mathbf{F}_\lambda(\mathbf{x}) = \int d^3k \; \boldsymbol{\chi}_\lambda(\mathbf{x}|\mathbf{k}) f_\lambda(\mathbf{k}). \tag{2.5}$$

The curl eigenfunction expansion of a linear force-free magnetic field is distinguished by its comparative simplicity: the appearance of only one component of the three required for the general vector field.

III. CURL EIGENFUNCTION EXPANSION OF A LINEAR FORCE-FREE MAGNETIC FIELD

A plasma magnetic field is defined to be force-free if its Lorentz force is zero, implying that the field is parallel to the plasma current density vector. Since the current density is proportional to the field vorticity in the absence of displacement current, the equations defining a FFMF become

$$\nabla \times \mathbf{B} = \alpha \mathbf{B}, \qquad \nabla \cdot \mathbf{B} = 0. \tag{3.1}$$

The proportionality factor α is constant for the linear FFMF case that we consider here (see Sturrock[8] for a discussion of linear versus nonlinear FFMFs). The first of Eqs. (3.1) shows the linear FFMF to be an eigenfunction of the curl operator, and it was shown in (I) that it is represented entirely by the integral over one rotational curl eigenfunction,

$$\mathbf{B}_\lambda(\mathbf{x}) = \int d^3k \ \boldsymbol{\chi}_\lambda(\mathbf{x}|\mathbf{k}) b_\lambda(\mathbf{k}), \tag{3.2}$$

for which the sign of λ matches that of α, and whose curl transform is restricted to the sphere $k = \lambda\alpha$ in transform space,

$$b_\lambda(\mathbf{k}) = \frac{\delta(k - \lambda\alpha)}{k^2} s_\lambda(\mathbf{k}), \tag{3.3}$$

leading to the final expression for a general linear FFMF:

$$\mathbf{B}_\lambda(\mathbf{x}) = (2\pi)^{-3/2} \int d\Omega \ e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}). \tag{3.4}$$

Here the integration is over the transform sphere, $\boldsymbol{\kappa} \equiv \mathbf{k}/k$ defines the unit position vector in transform space, and we have indicated the fact that the curl eigenvector does not depend on the magnitude of $\mathbf{k}$. We call the function s_λ the *spherical curl transform* to distinguish it from the full curl transform $b_\lambda(\mathbf{k})$. The relation (3.4), which defines the inverse curl transform of the field, allows one to calculate the field for given curl transforms, and an introductory exploration of this was carried out in (I). Since there is only one curl eigenfunction component of the linear FFMF, the curl transform of the field, i.e., the inverse of the relation (3.2), is unique and is given by the analog of Eq. (2.4):

$$b_\lambda(\mathbf{k}) = (2\pi)^{-3/2} \mathbf{Q}^*_\lambda(\boldsymbol{\kappa}) \cdot \int d^3x \ \mathbf{B}_\lambda(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} = \int d^3x \ \boldsymbol{\chi}^*_\lambda(\mathbf{x}|\mathbf{k}) \cdot \mathbf{B}_\lambda(\mathbf{x}). \tag{3.5}$$

In words: the curl transform of a linear FFMF is given by the projection of the Fourier transform of the field on the corresponding curl eigenvector, or, in other words, by the integral of the projection of the field on the corresponding curl eigenfunction. This relation enables the transform space behavior of known FFMFs to be computed and compared. As such, it would seem to be potentially more useful from a theoretical, classificatory standpoint than as a practical aid, though we shall point out below a case that belies that conclusion. Note that (3.5), as the analog of the general vector relation (2.4), has only made use of the unique curl eigenfunction decomposition of the linear FFMF, without taking into account the restriction of the curl transform to the sphere in transform space, as given in (3.3). When this is done the determination of the curl transform can be taken one step farther, and this was done in (II). The principal result of that paper was an expression for the curl transform, which, while demonstrated to be workable, lacked clarity of interpretation and was clumsy of execution. Our aim in the present paper is two-fold: to remedy these two features and to exhibit the curl transforms of some nontrivial linear FFMFs.

IV. THE SPHERICAL CURL TRANSFORM OF A LINEAR FORCE-FREE MAGNETIC FIELD

In this section we derive a general expression for the SCT in a form that lends itself to facile interpretation and demonstrate that its inverse leads back to the original field. Additionally we show that the Radon transform formalism produces an alternative representation of this inverse, which in another form was the main result of (I).

A. The direct spherical curl transform

In order to derive an expression for the spherical curl transform of a linear FFMF that makes explicit the restriction of the curl transform to the sphere in transform space, substitute Eq. (3.3) into the formula (3.5), producing the relation (II.4.2),

$$\frac{\delta(k-\lambda\alpha)}{k^2}\, s_\lambda(\mathbf{k}) = (2\pi)^{-3/2}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\int d^3x\ \mathbf{B}_\lambda(\mathbf{x})e^{-i\mathbf{k}\cdot\mathbf{x}} = \int d^3x\ \boldsymbol{\chi}^*_\lambda(\mathbf{x}|\mathbf{k})\cdot\mathbf{B}_\lambda(\mathbf{x}). \quad (4.1)$$

The presence of the delta function on the left-hand side implies that the right-hand side has a latent distributional character. This can be brought out more clearly by experimenting with different test function factors in the realizing integration that must be performed to utilize the delta function. While a complete functional analysis investigation of this relation properly involves a discussion of the function spaces to which it is applicable, we defer this to later investigation and proceed in a purely formal way.

The delta function was in origin a radial delta function; indeed, the fact that the product $\lambda\ \alpha$ must be positive is an essential feature of the curl eigenfunction description of FFMFs (I). Taking k^2 as a test function factor and integrating (4.1) over positive k leads down the path taken in (II):

$$\int_0^\infty k^2\ dk\ \frac{\delta(k-\lambda\alpha)}{k^2}\, s_\lambda(\mathbf{k}) = s_\lambda(\lambda\alpha\mathbf{k}) = (2\pi)^{-3/2}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\int_0^\infty k^2\ dk\int d^3x\mathbf{B}_\lambda(\mathbf{x})e^{-i\mathbf{k}\cdot\mathbf{x}}, \quad (4.2)$$

culminating in the expressions (4.10) and (4.11) of that paper, which see for further details on this approach. The development detailed there lacked elegance, and the resulting expressions lacked the transparency of interpretation that we develop below for an alternative path.

Introducing the function e^{ikp} as a test function factor and integrating both sides of (4.1) over k (with a doubly infinite range) leads to the following expressions for the spherical curl transform:

$$\begin{aligned} s_\lambda(\lambda\alpha\boldsymbol{\kappa}) &= \frac{\alpha^2}{(2\pi)^{3/2}}\, e^{-i\lambda\alpha p}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\int d^3x\ \mathbf{B}_\lambda(\mathbf{x})\int_{-\infty}^{+\infty} dk\ e^{ik(p-\boldsymbol{\kappa}\cdot\mathbf{x})} \\ &= \frac{\alpha^2}{(2\pi)^{1/2}}\, e^{-i\lambda\alpha p}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\int d^3x\ \mathbf{B}_\lambda(\mathbf{x})\,\delta(p-\boldsymbol{\kappa}\cdot\mathbf{x}). \end{aligned} \quad (4.3)$$

These are easily recognized as the (weighted) projection of the vector Radon transform of the FFMF on the corresponding curl eigenvector,

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = \frac{\alpha^2}{(2\pi)^{1/2}}\, e^{-i\lambda\alpha p}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(\mathbf{p}), \quad (4.4)$$

where the vector Radon transform is defined in Appendix B:

$$\mathbf{B}^R_\lambda(\mathbf{p}) = \mathbf{B}^R_\lambda(p,\boldsymbol{\kappa}) = \int d^3x\ \mathbf{B}_\lambda(\mathbf{x})\,\delta(p-\boldsymbol{\kappa}\cdot\mathbf{x}). \quad (4.5)$$

For computational purposes the delta function is typically replaced by its integral exponential form, as in the first of equations (4.3). The final result giving the spherical curl transform of a general linear FFMF in explicit form may then be written as

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = \frac{\alpha^2}{(2\pi)^{3/2}}\, e^{-i\lambda\alpha p}\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\int_{-\infty}^{+\infty} dk\; e^{ikp}\int d^3x\; \mathbf{B}_\lambda(\mathbf{x})e^{-ik\boldsymbol{\kappa}\cdot\mathbf{x}}, \tag{4.6}$$

where the Radon transform is represented as the doubly infinite integral over the radial direction of the FFMF Fourier transform (Deans[9]).

We now calculate the inverse transform to verify the above result.

B. The inverse spherical curl transform

The intimate relation found above between the spherical curl transform and the Radon transform gives a new perspective to the specification of the FFMF field in terms of an integral transform as well as providing an alternative path for deriving the inverse spherical curl transform, i.e., determining the field in terms of its transform. The starting point for such an exercise could be either transform since they are interconvertible using the relation (4.4),

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = \frac{\alpha^2}{(2\pi)^{1/2}}\, e^{-i\lambda\alpha p}\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{B}_\lambda^R(\mathbf{p}), \tag{4.4}$$

and (C2),

$$\mathbf{B}_\lambda^R(p,\boldsymbol{\kappa}) = \frac{(2\pi)^{1/2}}{\alpha^2}\,[e^{i\lambda\alpha p}\mathbf{Q}_\lambda(\boldsymbol{\kappa})s_\lambda(\lambda\alpha\boldsymbol{\kappa}) + e^{-i\lambda\alpha p}\mathbf{Q}_\lambda(-\boldsymbol{\kappa})s_\lambda(-\lambda\alpha\boldsymbol{\kappa})]. \tag{4.7}$$

In the first case we substitute the expression (4.4) in the integral relation (3.4), whereas in the second case we operate on (4.7) with the standard inverse Radon transform operator (Deans[9]),

$$\mathbf{B}(\mathbf{x}) = -\frac{1}{8\pi^2}\,\nabla_x^2\int_S d\Omega\; \mathbf{B}^R(\boldsymbol{\kappa}\cdot\mathbf{x},\boldsymbol{\kappa}). \tag{4.8}$$

The two results must be equivalent, and we show that by reducing the first case to (4.8), then showing that (4.8) reproduces the correct field. Note clearly that the point to carrying out the latter effort is not to verify the general formula for inverting the Radon transform, but to prove that our expression for the Radon transform of the linear FFMF (4.7) gives the correct field when inverted.

Begin by substituting the spherical curl transform expression (4.4) into Eq. (3.4), which inverts the transform to produce the general linear FFMF:

$$\mathbf{B}_\lambda(\mathbf{x}) = \frac{\alpha^2}{4\pi^2}\int d\Omega\; e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{B}_\lambda^R(\mathbf{p}). \tag{4.9}$$

The completeness relation for the curl eigenvectors, (A2b), allows the vector expression occurring here to be written as the dyadic projection,

$$\mathbf{Q}_\lambda\mathbf{Q}_\lambda^*\cdot\mathbf{B}_\lambda^R = \mathfrak{I}\cdot\mathbf{B}_\lambda^R - (\mathbf{Q}_{-\lambda}\mathbf{Q}_{-\lambda}^* + \mathbf{Q}_0\mathbf{Q}_0^*)\cdot\mathbf{B}_\lambda^R, \tag{4.10}$$

where $\mathfrak{I}$ is the unit dyadic. The projections of the $\mathbf{Q}_\lambda$ dyads on the Radon transform can only reproduce the corresponding pieces of (4.7); [$\mathbf{Q}_\lambda(-\boldsymbol{\kappa})$ can be written in terms of $\mathbf{Q}_{-\lambda}(\boldsymbol{\kappa})$ using (A5) and (A3c)], so that (4.10) reduces to

$$\mathbf{Q}_\lambda\mathbf{Q}_\lambda^*\cdot\mathbf{B}_\lambda^R = \mathfrak{I}\cdot\mathbf{B}_\lambda^R - \mathbf{Q}_{-\lambda}\mathbf{Q}_{-\lambda}^*\cdot\mathbf{B}_\lambda^R. \tag{4.11}$$

It is shown in Appendix D that the integrals over the sphere of the two $\mathbf{Q}$ dyadic projections on the Radon transform are equal, so that

$$\int d\Omega\; e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{B}_\lambda^R(\mathbf{p}) = \frac{1}{2}\int d\Omega rs\; e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathfrak{I}\cdot\mathbf{B}_\lambda^R(\mathbf{p}), \tag{4.12}$$

and (4.9) becomes

$$\mathbf{B}_\lambda(\mathbf{x}) = \frac{\alpha^2}{8\pi^2} \int d\Omega\ e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})} \mathbf{B}^R_\lambda(\mathbf{p}). \tag{4.13}$$

By introducing the Laplacian and specializing the magnitude of $\mathbf{p}$ to $\boldsymbol{\kappa}\cdot\mathbf{x}$, we regain the standard expression for the inverse Radon transform (4.8). Thus, we have shown that the inverse spherical curl transform (3.4) reduces to the inverse Radon transform (4.8).

The demonstration that the curl eigenfunction expansion of the FFMF Radon transform (4.7) satisfies (4.8) to reproduce the field now follows easily. Substitute (4.7) into (4.8) to find

$$\begin{aligned}\mathbf{B}_\lambda(\mathbf{x}) &= -\frac{1}{2(2\pi)^{3/2}\alpha^2} \nabla_x^2 \int d\Omega[e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) + e^{-i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(-\boldsymbol{\kappa}) s_\lambda(-\lambda\alpha\boldsymbol{\kappa})] \\ &= \frac{1}{2(2\pi)^{3/2}} \int d\Omega[e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) + e^{-i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(-\boldsymbol{\kappa}) s_\lambda(-\lambda\alpha\boldsymbol{\kappa})]. \end{aligned} \tag{4.14}$$

The second integral is just the first evaluated with the coordinates inverted, whence the expression for the FFMF in terms of the inverse spherical curl transform [(3.4); (I.8)] becomes apparent,

$$\mathbf{B}_\lambda(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}} \int d\Omega\ e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}). \tag{4.15}$$

Thus, the description of the spherical curl transform in terms of the Radon transform is both necessary and sufficient; they are completely equivalent representations.

V. EXAMPLES OF LINEAR FFMF SPHERICAL CURL TRANSFORMS

A significant part of the rationale for introducing the curl eigenfunction expansion of FFMFs was to explore possible advantages in using the transform space representation. The viewpoint of (I) was from the transform space: what physical space fields correspond to given curl transforms. The viewpoint taken here is from the physical space: what curl transforms correspond to given FFMFs. The availability of the curl transform expressions above will ease the development of such representations and the comparison with existing FFMFs, processes that are necessary for the exploration of the representation technique. To illustrate the process we have chosen to derive the curl transforms for three fundamental fields: the simplest FFMF, the Lundquist field, and the classical spheromak field.

The notation below follows that of (I, II); physical space spherical coordinates are denoted by (r,θ,φ), while the corresponding coordinates in transform space are denoted by (k,η,ψ) [see Fig. 1 of (II)].

A. The simplest FFMF

This field, which is simply a particular type of standing vector plane wave, corresponds to a curl transform that is a delta function at a point on the transform sphere. It is represented in physical space by the complex wave [see (I)],

$$\mathbf{B}(\mathbf{x}) = \mathbf{B}_0 e^{i\mathbf{k}_0\cdot\mathbf{x}}, \tag{5.1}$$

whose wave vector must be transverse to satisfy the divergence-free condition (3.1b),

$$\mathbf{k}_0\cdot\mathbf{B}_0 = 0, \tag{5.2}$$

and have magnitude α to satisfy (3.1a):

$$k_0 = \lambda\alpha, \quad \lambda = \pm 1. \tag{5.3}$$

Its Radon transform is

$$\mathbf{B}_\lambda^R(p,\boldsymbol{\kappa}) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dk\ e^{ikp}\int d^3x\ \mathbf{B}_\lambda(\mathbf{x})e^{-ik\boldsymbol{\kappa}\cdot\mathbf{x}}$$

$$= \frac{\mathbf{B}_0}{2\pi}\int_{-\infty}^{+\infty} dk\ e^{ikp}\int d^3x\ e^{i(\mathbf{k}_0-k\boldsymbol{\kappa})\cdot\mathbf{x}}$$

$$= \mathbf{B}_0\int_{-\infty}^{+\infty} dk\ e^{ikp}\delta(\mathbf{k}_0-k\boldsymbol{\kappa})$$

$$= \mathbf{B}_0\int_{-\infty}^{+\infty} dk\ e^{ikp}\frac{\delta(k_0-k)}{k^2}\ \delta(\boldsymbol{\kappa}_0-\boldsymbol{\kappa}) = \frac{\mathbf{B}_0}{k_0^2}\,e^{ik_0p}\delta(\boldsymbol{\kappa}_0-\boldsymbol{\kappa}), \tag{5.4}$$

from which the curl transform is obtained according to (4.4), after substituting the value of k_0 from (5.3):

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = (2\pi)^{-3/2}\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{B}_0\,\delta(\boldsymbol{\kappa}_0-\boldsymbol{\kappa}) = s_{\lambda 0}\,\delta(\boldsymbol{\kappa}-\boldsymbol{\kappa}_0). \tag{5.5}$$

This agrees with (I.9); it is simply a complex amplitude delta function acting at the point $\boldsymbol{\kappa}_0 = \mathbf{k}_0/k_0$ on the unit sphere in transform space, and as such is a representative of the class of point transform fields discussed briefly there.

B. The Lundquist field

The Lundquist field is significant for two reasons: it was the first FFMF (constructed by Lundquist[10] in 1950), and it has been used as a model for the fields of solar magnetic clouds (Burlaga[4]). Its spherical curl transform turns out to be relatively simple, though its derivation is somewhat lengthy.

According to the formula developed above in (4.6), the spherical curl transform of the Lundquist field, $\mathbf{B}_\mathrm{L}$, may be written as the projection on the corresponding curl eigenvector of the doubly infinite radial Fourier integral in transform space of the three-dimensional (3-D) Fourier transform of the field:

$$s_{\mathrm{L}\lambda}(\lambda\alpha\boldsymbol{\kappa}) = \frac{\alpha^2}{(2\pi)^{3/2}}\,e^{-i\lambda\alpha p}\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\int_{-\infty}^{+\infty} dk\ e^{ikp}\int d^3x\ \mathbf{B}_{\mathrm{L}\lambda}(\mathbf{x})e^{-ik\boldsymbol{\kappa}\cdot\mathbf{x}}. \tag{5.6}$$

While the Lundquist field has cylindrical symmetry, it is convenient to express its vector form in Cartesian components to simplify the integration, while retaining its functional form in terms of the circular cylindrical components (r,φ,z):

$$\mathbf{B}_\mathrm{L} = B_0\{-J_1(\alpha r)\sin\varphi,\ J_1(\alpha r)\cos\varphi,\ J_0(\alpha r)\}. \tag{5.7}$$

Recall that the helicity is given by λ (the sign of α), so that it could be explicitly shown as follows:

$$\mathbf{B}_{\mathrm{L}\lambda} = B_{\lambda 0}\{-\lambda J_1(\lambda\alpha r)\sin\varphi,\ \lambda J_1(\lambda\alpha r)\cos\varphi,\ J_0(\lambda\alpha r)\}, \tag{5.8}$$

by taking the parity of the Bessel functions into account and demanding that the Bessel function arguments be positive. To simplify the algebra we carry the form (5.7) through and convert to an explicit helicity form at the end.

The Fourier transform of the Lundquist field is evaluated in Appendix E to be

$$\mathbf{B}_{\mathrm{L}\lambda}^F(\mathbf{k}) = (2\pi)^{+1/2}iB_{\lambda 0}\delta(k_z)\frac{\delta(k-\lambda\alpha)}{\lambda\alpha}\,[\lambda\sin\psi\,\mathbf{i}_x - \lambda\cos\psi\,\mathbf{i}_y - i\mathbf{i}_z]. \tag{5.9}$$

Taking the Fourier transform of this over k simply adds an exponential factor,

$$\int_{-\infty}^{+\infty} dk\ e^{ikp}\mathbf{B}_{\mathrm{L}\lambda}^F(\mathbf{k}) = (2\pi)^{+1/2}iB_{\lambda 0}\delta(k_z)\,\frac{e^{i\lambda\alpha p}}{\lambda\alpha}\,[\lambda\sin\psi\,\mathbf{i}_x - \lambda\cos\psi\,\mathbf{i}_y - i\mathbf{i}_z], \tag{5.10}$$

whose purpose it is to cancel the inverse factor in the spherical curl transform expression (5.6):

$$s_{L\lambda}(\lambda\alpha\boldsymbol{\kappa})=\frac{\alpha i}{2\pi}B_{\lambda 0}\delta(k_z)\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot[\sin\psi\,\mathbf{i}_x-\cos\psi\,\mathbf{i}_y-i\lambda\mathbf{i}_z]. \tag{5.11}$$

The general form for the conjugate curl eigenvector expression as a Cartesian vector in terms of cylindrical components is

$$\mathbf{Q}^*_\lambda=-\frac{\lambda}{\sqrt{2}}\frac{k_r e^{-i\lambda\psi}}{k+k_z}\left(\frac{\mathbf{k}}{k}+\mathbf{i}_z\right)+\frac{\lambda}{\sqrt{2}}(\mathbf{i}_x-i\lambda\mathbf{i}_y), \tag{5.12}$$

but anticipating the effect of the k_z delta function reduces this effectively to

$$\mathbf{Q}^*_\lambda=-\frac{\lambda}{\sqrt{2}}[e^{-i\lambda\psi}(\cos\psi\,\mathbf{i}_x+\sin\psi\,\mathbf{i}_y+\mathbf{i}_z)-(\mathbf{i}_x-i\lambda\mathbf{i}_y)], \tag{5.13}$$

leading to the final expression for the spherical curl transform of the Lundquist field:

$$s_{L\lambda}(\lambda\alpha\boldsymbol{\kappa})=-\frac{\alpha}{\sqrt{2}\pi}B_{\lambda 0}\delta(k_z)e^{-i\lambda\psi}. \tag{5.14}$$

This SCT is a member of the class of line transforms [see (I), Sec. 3. B] and lives on the equatorial circle of the transform sphere with an amplitude that is a simple complex exponential function of the azimuthal angle ψ. It is capable of immediate generalization by associating it with the class of circle transforms sketched in (I) that have constant amplitude over a small circle of the sphere. This has led to a family of fields related to the Lundquist field that has two additional parameters and is expected to find an application in interplanetary magnetohydrodynamics by providing a more flexible model for solar magnetic cloud fields (MacLeod[7]).

C. The classical spheromak field

The classical spheromak field is significant for consideration here because it has found an application both in interplanetary magnetohydrodynamics and in nuclear fusion work. It was first constructed and introduced by Rosenbluth and Bussac[11] in order to investigate the stability properties of a family of fusion devices proposed as alternatives to the tokamak (see Wright[6] for a review of spheromak work and Sheffield[12] to cast this in the perspective of nuclear fusion work in general; Goldenbaum[13] gives a useful presentation of spheromak physics). The magnetic field of the simplest of this family is the classical spheromak, which bears a strong resemblance to Hill's spherical vortex in fluid mechanics. More recently this field has been proposed as a model of the magnetic field of solar magnetic clouds (Ivanov and Harshiladze,[5] Vandas, Fischer, and Geranios;[14] see Farrugia *et al.*[15] for an extensive review and comparison with the Lundquist field model). (This field also serves as a simple introduction to the use of spherical harmonic expansions as a general means for describing FFMFs, since any reasonably behaved function on the transform sphere may be so expanded.)

We adopt the form of the classical spheromak field given by Farrugia *et al.*[15] in spherical coordinates and components as

$$\mathbf{B}_\lambda(r,\theta,\varphi)=B_{\lambda 0}\left\{\frac{2}{\alpha r}j_1(\lambda\alpha r)\cos\theta\,\mathbf{i}_r+\frac{1}{\alpha r}[j_1(\lambda\alpha r)-\sin(\lambda\alpha r)]\sin\theta\,\mathbf{i}_\theta+\lambda j_1(\lambda\alpha r)\sin\theta\,\mathbf{i}_\varphi\right\}, \tag{5.15}$$

where we have replaced the signed parameter α by its magnitude $\lambda\,\alpha$, as well as making use of the parity of the spherical Bessel functions to show the helicity λ explicitly. Now rewrite this for ease of calculation in spherical coordinates and Cartesian components:

$$\mathbf{B}_\lambda(r,\theta,\varphi)=\frac{B_{\lambda 0}}{3}\{[-j_2(\lambda\alpha r)P_2^1(\mu)\cos(\varphi)+3\lambda j_1(\lambda\alpha r)P_1^1(\mu)\sin(\varphi)]\mathbf{i}_x$$
$$-[j_2(\lambda\alpha r)P_2^1(\mu)\sin(\varphi)+3\lambda j_1(\lambda\alpha r)P_1^1(\mu)\cos(\varphi)]\mathbf{i}_y$$
$$+2[j_2(\lambda\alpha r)P_2(\mu)+j_0(\lambda\alpha r)P_0(\mu)]\mathbf{i}_z\}, \tag{5.16}$$

where we have converted to the Legendre function description in terms of the variable $\mu\equiv\cos\theta$. Note that the helicity affects only the sign of the azimuthal component, as is clear from (5.15).

Once again, the first step in calculating the spherical curl transform is to evaluate the Fourier transform of the field. We choose to do this by utilizing the spherical harmonic expansion of the exponential kernel (Stratton[16]):

$$e^{-ik\cdot x}=e^{-ikr\cos\omega}=\sum_{n=0}^{\infty}(-i)^n(2n+1)j_n(kr)[pa_n(\mu,\eta)+pb_n(\mu,\eta,\varphi,\psi)],$$

$$pa_n(\mu,\eta)\equiv P_n(\mu)P_n(\cos\eta), \tag{5.17}$$

$$pb_n(\mu,\eta,\varphi,\psi)\equiv 2\sum_{m=1}^{n}\frac{(n-m)!}{(n+m)!}P_n^m(\mu)P_n^m(\cos\eta)\cos m(\varphi-\psi),$$

which reduces the problem to the evaluation of a small number of elementary integrals; the calculation is outlined in Appendix F.

The resulting expression for the Fourier transform of the classical spheromak field is

$$\mathbf{b}_\lambda(\mathbf{k})=\frac{\lambda}{3}\sqrt{\frac{\pi}{2}}B_{\lambda 0}\frac{\delta(k-\lambda\alpha)}{k\alpha}\{[P_2^1(\mu')\cos\psi-3i\lambda P_1^1(\mu')\sin\psi]\mathbf{i}_x$$
$$+[P_2^1(\mu')\sin\psi+3i\lambda P_1^1(\mu')\cos\psi]\mathbf{i}_y+2[-P_2(\mu')+P_0(\mu')]\mathbf{i}_z\}, \tag{5.18}$$

where $\mu'\equiv\cos(\eta)$. The essence of taking the Fourier transform of this expression over k (i.e., forming the Radon transform of the field) is the evaluation of the integral

$$\int_{-\infty}^{+\infty}dk\ e^{ikp}\frac{\delta(k-\lambda\alpha)}{\lambda k\alpha}=\frac{e^{i\lambda\alpha p}}{\alpha^2}, \tag{5.19}$$

again with the result of annihilating the exponential factor in the spherical curl transform formula (4.6), and the SCT takes the form

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa})=\frac{\sqrt{2\pi}}{6}B_{\lambda 0}\{[P_2^1(\mu')\cos\psi-3i\lambda P_1^1(\mu')\sin\psi]\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{i}_x$$
$$+[P_2^1(\mu')\sin\psi+3i\lambda P_1^1(\mu')\cos\psi]\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{i}_y+2[-P_2(\mu')+P_0(\mu')]\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{i}_z\}. \tag{5.20}$$

The recognition that the Legendre functions can be factored,

$$P_2^1(\mu')=3\mu'P_1^1(\mu')=3P_1(\mu')P_1^1(\mu'), \tag{5.21}$$

and rewritten as

$$-P_2+P_0=\frac{3}{2}(P_1^1)^2, \tag{5.22}$$

reducing the form (5.20) to

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = \frac{\sqrt{2\pi}}{2} B_{\lambda 0} P_1^1(\mu')\{[P_1(\mu')\cos\psi - i\lambda\sin\psi]\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{i}_x$$
$$+[P_1(\mu')\sin\psi + i\lambda\cos\psi]\mathbf{Q}_\lambda^*(\boldsymbol{\kappa})\cdot\mathbf{i}_y + P_1^1(\mu')\mathbf{Q}_\lambda^*(\mathbf{k})\cdot\mathbf{i}_z\} \tag{5.23}$$

simplifies the subsequent algebra. Now evaluating the conjugate curl eigenvector dot products using (A1), we find

$$\mathbf{Q}_\lambda^*\cdot\mathbf{i}_x = -\frac{\lambda}{\sqrt{2}}[(1-P_1)\cos\psi\, e^{-i\lambda\psi} - 1],$$

$$\mathbf{Q}_\lambda^*\cdot\mathbf{i}_y = -\frac{\lambda}{\sqrt{2}}[(1-P_1)\sin\psi\, e^{-i\lambda\psi} + i\lambda], \tag{5.24}$$

$$\mathbf{Q}_\lambda^*\cdot\mathbf{i}_z = \frac{\lambda}{\sqrt{2}} P_1^1 e^{-i\lambda\psi}.$$

Substituting (5.24) in (5.23), the final form for the spherical curl transform of the classical spheromak field reduces to (!!)

$$s_\lambda(\lambda\alpha\boldsymbol{\kappa}) = -\lambda\sqrt{\pi} B_{\lambda 0} P_1^1(\mu') e^{-i\lambda\psi}. \tag{5.25}$$

Thus, this SCT is a member of the class of areal transforms [see (I), Sec. 3.C] and lives on the whole sphere with an amplitude varying with the sine of the polar angle [i.e., $P_1^1(\cos\eta)$] and as a complex exponential in the azimuthal angle.

VI. SUMMARY AND OUTLOOK

The description of the general linear force-free magnetic field in terms of the Moses curl eigenfunctions begun previously by the derivation of an expression giving the physical field in terms of an integral over the transform sphere has been completed by deriving a general expression for the inverse operator that enables the spherical curl transform to be determined in terms of the physical field. This inverse may be represented as a weighted projection of the vector Radon transform of the field on the appropriate curl eigenvector. The process was exemplified by the determination of the transforms of three fields: the simplest force-free magnetic field, and the Lundquist and classical spheromak fields. The use of the SCT in generalizing a given type of FFMF was briefly sketched for the Lundquist field.

The point that the spherical curl transforms found above are considerably simpler in functional form than their corresponding fields in physical space cannot be overemphasized. One way of looking at the effect is to say that the use of the curl eigenvectors to shoulder the responsibility of describing the field geometry in the Fourier transform has simplified the problem considerably. It has, in fact, reduced the vector three-dimensional problem in physical space to a scalar two-dimensional problem on the sphere in transform space. It has, so far, used simpler functions to describe the transform than are used to describe the physical field. Of course, one tends to doubt that the latter is a universal feature of the representation, but the fact is that this has been a prominent feature of the fields considered to date, and, as seen above, these have included some nontrivial fields. To demonstrate this difference in complexity, examples of field/transform pairs are shown in Table I. These have been taken from the three classes of point, line, and areal transforms outlined in (I), including the three pairs discussed above. This suggests that a considerable advantage may be gained in the study of FFMFs by working in transform space, though further techniques are needed to reduce this to practice, such as the development of a Green's function approach to solve specific boundary value problems.

It would be interesting to reconsider the FFMFs found by Richter from symmetry considerations (Richter[17]) from the standpoint of the curl eigenfunction representation to see what light they might shed on the complexity issue. Surely the group symmetries found by Emets' and Kovbasenko[18] and Richter[19] can contribute to the elucidation of this facet of the problem.

TABLE I. The relative simplicity of the expressions for the spherical curl transforms compared to their corresponding fields. [For details on the circle line and constant areal transforms and their parameters see (I).]

Field type/Name	Field expression	Spherical curl transform
Point transform/Simplest:	$\mathbf{B}_0 e^{i\mathbf{k}_0\cdot\mathbf{x}}$	$(2\pi)^{-3/2}\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}_0\delta(\boldsymbol{\kappa}_0-\boldsymbol{\kappa})$
Line transform/Circle:	$\frac{B_{\lambda 0}}{2\pi^{1/2}} e^{iz/L_z}[(\cos\eta_0+1)J_0(r/L_r)\mathbf{i}_\lambda-(\cos\eta_0-1)J_2(r/L_r)e^{i2\lambda\varphi}\mathbf{i}^*_\lambda - \sqrt{2}\lambda i\sin\eta_0 J_1(r/L_r)e^{i\lambda\varphi}\mathbf{i}_z]$	$B_{\lambda 0}\frac{\delta(\eta-\eta_0)}{\sin\eta}$
Line transform/Lundquist:	$B_{\lambda 0}[\lambda J_1(\lambda\alpha r)\mathbf{i}_\varphi+J_0(\lambda\alpha r)\mathbf{i}_z]$	$-\frac{\alpha}{\sqrt{2}\pi}B_{\lambda 0}\delta(k_z)e^{-i\lambda\psi}$
Areal transform/Constant:	$(2\pi)^{-1/2}B_{\lambda 0}\text{-}\varnothing\Big\{[j_0(\alpha r)+ij_1(\alpha r)P_1(\mu)]\mathbf{i}_\lambda - \pi^{-1/2}\lambda i P^1_1(\mu)e^{i\lambda\varphi}j_1(\alpha r)\mathbf{k}-\sqrt{\frac{2}{\pi}}\sum_{n=2}^{\infty}(-i)^n \times(2n+1)j_\pi(\alpha r)\frac{(n-2)!}{(n+2)!}P^2_n(\mu)e^{i2\lambda\varphi}\mathbf{i}^*_\lambda\Big\}$	$B_{\lambda 0}$
Areal transform/Classical spheromak:	$B_{\lambda 0}\Big\{\frac{2}{\alpha r}j_1(\lambda\alpha r)\cos\theta\mathbf{i}_r+\frac{1}{\alpha r}[j_1(\lambda\alpha r)-\sin(\lambda\alpha r)] \times\sin\theta\mathbf{i}_\theta+\lambda j_1(\lambda\alpha r)\sin\theta\mathbf{i}_\varphi\Big\}$	$-\lambda\sqrt{\pi}B_{\lambda 0}P^1_1(\mu)e^{-i\lambda\psi}$

Another path to explore is the use of alternative formulas for the Radon transform. Moses and Prosser[20] showed that it is sufficient to integrate over a ''canonical hemisphere'' for the inverse Radon transform, reducing the well-known double covering of x space to a single covering. (Similar considerations have been discussed previously. The Radon transform literature is extensive; see, e.g., Ludwig[21] and Guillemin[22]). A tantalizing aspect of this is the thought that the Radon transform and the curl eigenfunctions are both aspects of a particular Hopf fibration that could both simplify and generalize our understanding of these concepts and their interrelations by lifting the description to one in four dimensions, in spite of the fact that the curl operator lives in three dimensions.

Finally, it must be emphasized again that, although the entire discussion above has been couched in terms of force-free magnetic fields, the results apply equally well [with the caveat mentioned in (I)] to the description of the Trkalian class of Beltrami flows in fluid dynamics.

ACKNOWLEDGMENTS

Once again, the assistance of the staff of the nonpareil Geophysics Research Library, especially Ms. Susan Benzer, is gratefully acknowledged, as is the continued support of Gary Mullen.

APPENDIX A: PROPERTIES OF THE MOSES CURL EIGENVECTORS

The $\mathbf{Q}_\lambda(\mathbf{k})$ vectors introduced above in the definition of the curl eigenfunctions were defined by Moses[3] in Cartesian coordinates as follows:

$$\mathbf{Q}_0(\mathbf{k})\equiv-\frac{\mathbf{k}}{k},$$

$$\mathbf{Q}_\lambda(\mathbf{k})\equiv-\frac{\lambda}{\sqrt{2}}\left[\frac{k_1(k_1+i\lambda k_2)}{k(k+k_3)}-1,\ \frac{k_2(k_1+i\lambda k_2)}{k(k+k_3)}-i\lambda,\ \frac{k_1+i\lambda k_2}{k}\right]\quad(\lambda=\pm1). \tag{A1}$$

They possess the following useful properties:

$$\mathbf{Q}_\lambda^*(\mathbf{k})\cdot\mathbf{Q}_\mu(\mathbf{k})=\delta_{\lambda\mu}\quad(\lambda,\mu=0,\pm1),$$
$$\sum_\lambda \mathbf{Q}_{\lambda i}^*(\mathbf{k})\cdot\mathbf{Q}_{\lambda j}(\mathbf{k})=\delta_{ij}\quad(i,j=1,2,3), \tag{A2}$$

which provide the foundation for the orthogonality and completeness of the curl eigenfunctions. Three other relations are useful in simplifying expressions:

$$\mathbf{k}\times\mathbf{Q}_\lambda(\mathbf{k})=-i\lambda k\mathbf{Q}_\lambda(\mathbf{k})\quad(\lambda,\mu=0,\pm1),$$
$$\mathbf{Q}_\mu^*(\mathbf{k})\times\mathbf{Q}_\lambda(\mathbf{k})=-i\lambda\ \delta_{\lambda\mu}\mathbf{Q}_0(\mathbf{k})=-i\lambda\delta_{\lambda\mu}\frac{\mathbf{k}}{k}\quad(\lambda=\pm1), \tag{A3}$$
$$\mathbf{Q}_\lambda^*(\mathbf{k})=-\mathbf{Q}_{-\lambda}(\mathbf{k})\quad(\lambda=\pm1).$$

Note that, although the curl eigenvectors are invariant under a positive scale change of their argument,

$$\mathbf{Q}_\lambda(\mathbf{k})=\mathbf{Q}_\lambda(\boldsymbol{\kappa})=\mathbf{Q}_\lambda(\beta\boldsymbol{\kappa}),\quad \beta>0, \tag{A4}$$

a negative scale change introduces a phase factor and transforms a rotational eigenvector to its counterpart with opposite helicity:

$$\mathbf{Q}_\lambda(-\beta\boldsymbol{\kappa})=-\left(\frac{\kappa_1+i\lambda\kappa_2}{\kappa_1-i\lambda\kappa_2}\right)\mathbf{Q}_\lambda^*(\beta\boldsymbol{\kappa})=\left(\frac{\kappa_1+i\lambda\kappa_2}{\kappa_1-i\lambda\kappa_2}\right)\mathbf{Q}_{-\lambda}(\beta\boldsymbol{\kappa})=e^{i2\lambda\psi}\mathbf{Q}_{-\lambda}(\boldsymbol{\kappa}), \tag{A5}$$

while the same operation reverses the sign of $\mathbf{Q}_0(\mathbf{k})$.

N.B.: The k factor on the right-hand side of (A3a) was inadvertently omitted from the corresponding list in both (I) and (II).

APPENDIX B: VECTOR RADON AND FOURIER TRANSFORMS

The three-dimensional scalar Radon transform,

$$f^R(\mathbf{p})\equiv\int d^3x\ f(\mathbf{x})\,\delta(p-\boldsymbol{\kappa}\cdot\mathbf{x}), \tag{B1}$$

and its inverse,

$$f(\mathbf{x})=\frac{1}{8\pi^2}\nabla_x^2\int d\Omega\ f^R(\boldsymbol{\kappa}\cdot\mathbf{x},\boldsymbol{\kappa}) \tag{B2}$$

(Deans[9]), as well as the corresponding Fourier transforms, are easily converted to vector transforms given in terms of the Cartesian components,

$$\mathbf{f}(\mathbf{x})=\sum_{j=1}^{3} f_j\mathbf{i}_j, \tag{B3}$$

by replacing the scalar functions by vector functions:

$$\mathbf{f}^R(\mathbf{p})\equiv\int d^3x\ \mathbf{f}(\mathbf{x})\,\delta(p-\boldsymbol{\kappa}\cdot\mathbf{x}) \tag{B4}$$

(and similarly for the other transforms). Here we have used the linearity of their integral operators and, in the case of the inverse Radon transform, recalled that the vector Laplacian is equivalent to the scalar Laplacian acting on the Cartesian components (Morse and Feshbach[23]).

The relation between the Radon and Fourier transforms depends, however, on the specific definition of the Fourier transform. Since we have adopted the symmetrical definition of Moses

that differs from that which commonly appears in the Radon transform literature (Deans,[9] Helgason[24]), we must rederive the interrelation of the two transforms. Following Deans, p. 97, write the Fourier transform as

$$\mathbf{f}^F(\mathbf{k}) = (2\pi)^{-3/2} \int_{-\infty}^{+\infty} dt \int d^3x \; \mathbf{f}(\mathbf{x}) e^{-it} \delta(t - \mathbf{k}\cdot\mathbf{x}). \tag{B5}$$

Now change variables as follows:

$$\mathbf{k} = s\boldsymbol{\kappa}, \quad t = sp.$$

After substituting these values and rearranging, we find the Fourier transform in terms of the Radon transform; then inverting this relation, the Radon transform in terms of the Fourier transform:

$$\begin{aligned} \mathbf{f}^F(s\boldsymbol{\kappa}) &= (2\pi)^{-3/2} \int_{-\infty}^{+\infty} dp \; e^{-isp} \mathbf{f}^R(p\boldsymbol{\kappa}), \\ \mathbf{f}^R(q\boldsymbol{\kappa}) &= (2\pi)^{1/2} \int_{-\infty}^{+\infty} ds \; e^{isq} \mathbf{f}^F(s\boldsymbol{\kappa}). \end{aligned} \tag{B6}$$

APPENDIX C: THE RADON TRANSFORM OF A LINEAR FORCE-FREE MAGNETIC FIELD

The Radon transform of the FFMF introduced in Sec. IV has a subtlety in its derivation that we explain here. If we set about to develop the Radon transform of the linear FFMF by inserting the expression (3.4) into the vector Radon transform definition (4.5), introducing the integral exponential representation of the delta function and rearranging leads to

$$\begin{aligned} \mathbf{B}^R_\lambda(p,\mathbf{u}) &= \int d^3x \left[(2\pi)^{-3/2} \int d\Omega \; e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \right] \delta(p - \mathbf{u}\cdot\mathbf{x}) \\ &= (2\pi)^{-5/2} \int d^3x \int d\Omega \; e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}} \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \int_{-\infty}^{\infty} dt \; e^{it(p-\mathbf{u}\cdot\mathbf{x})} \\ &= (2\pi)^{-5/2} \int d\Omega \; \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \int_{-\infty}^{\infty} dt \; e^{itp} \int d^3x \; e^{i\mathbf{x}\cdot(\lambda\alpha\boldsymbol{\kappa} - t\mathbf{u})} \\ &= (2\pi)^{1/2} \int d\Omega \; \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \int_{-\infty}^{\infty} dt \; e^{itp} \delta(\lambda\alpha\boldsymbol{\kappa} - t\mathbf{u}). \end{aligned} \tag{C1}$$

Given that $\lambda\alpha > 0$ and that $\mathbf{u}$ has a specific orientation, this delta function contributes to the integral twice: when the sign of $\boldsymbol{\kappa}$ matches that of $\mathbf{u}$ and t is positive, and when the sign of $\boldsymbol{\kappa}$ is oppositely directed to $\mathbf{u}$ and t is negative:

$$\begin{aligned} \mathbf{B}^R_\lambda(p,\mathbf{u}) &= (2\pi)^{1/2} \int d\Omega \; \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \left[\int_0^{\infty} dt \; e^{itp} \delta(\lambda\alpha\boldsymbol{\kappa} - t\mathbf{u}) + \int_0^{\infty} dt \; e^{-itp} \delta(\lambda\alpha\boldsymbol{\kappa} + t\mathbf{u}) \right] \\ &= (2\pi)^{1/2} \left[\int_0^{\infty} dt \; e^{itp} \frac{\delta(t - \lambda\alpha)}{t^2} \int d\Omega \; \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \delta(\boldsymbol{\kappa} - \mathbf{u}) \right. \\ &\quad \left. + \int_0^{\infty} dt \; e^{-itp} \frac{\delta(t + \lambda\alpha)}{t^2} \int d\Omega \; \mathbf{Q}_\lambda(\boldsymbol{\kappa}) s_\lambda(\lambda\alpha\boldsymbol{\kappa}) \delta(\boldsymbol{\kappa} + \mathbf{u}) \right] \\ &= \frac{(2\pi)^{1/2}}{\alpha^2} \left[e^{i\lambda\alpha p} \mathbf{Q}_\lambda(\mathbf{u}) s_\lambda(\lambda\alpha\mathbf{u}) + e^{-i\lambda\alpha p} \mathbf{Q}_\lambda(-\mathbf{u}) s_\lambda(-\lambda\alpha\mathbf{u}) \right]. \end{aligned} \tag{C2}$$

As a check on this result, note that it obviously satisfies the two main symmetries of the Radon transform [which reflect the symmetries of the delta function in the definition (4.5)]:

$$\begin{aligned}\mathbf{B}^R_\lambda(-p,-\mathbf{u})&=\mathbf{B}^R_\lambda(p,\mathbf{u}),\\ \mathbf{B}^R_\lambda(-p,\mathbf{u})&=\mathbf{B}^R_\lambda(p,-\mathbf{u}).\end{aligned}\tag{C3}$$

Thus the Radon transform of a linear FFMF contains within it the essence of both helicities of the field. Comparing (C2) with (3.4), the Radon transform for which $p=\boldsymbol{\kappa}\cdot\mathbf{x}$ is seen to be proportional to the mean of the spherical Fourier transforms of the two helicity components.

APPENDIX D: THE EQUIVALENCE OF CERTAIN INTEGRAL EXPRESSIONS

The essential step in developing (4.12) is to show that the projections of the $\mathbf{Q}_\lambda\mathbf{Q}^*_\lambda$ and $\mathbf{Q}_{-\lambda}\mathbf{Q}^*_{-\lambda}$ dyads on the FFMF Radon transform are equal when integrated over the sphere:

$$\int d\Omega\; e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(\mathbf{p})=\int d\Omega\; \mathbf{e}^{i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_{-\lambda}(\boldsymbol{\kappa})\mathbf{Q}^*_{-\lambda}(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(\mathbf{p}).\tag{D1}$$

Transform the right-hand integral by inverting the coordinates, replacing the resulting negative argument $\mathbf{Q}$ dyad by its positive counterpart using (A5), flip the argument signs of the Radon transform using (C3b), and substitute the expression (C2) for the FFMF Radon transform to obtain sequentially

$$\begin{aligned}\int d\Omega\; e^{i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_{-\lambda}(\boldsymbol{\kappa})\mathbf{Q}^*_{-\lambda}(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(\mathbf{p})&=\int d\Omega\; e^{i\lambda\alpha(p+\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_{-\lambda}(-\boldsymbol{\kappa})\mathbf{Q}^*_{-\lambda}(-\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(p,-\boldsymbol{\kappa})\\ &=\int d\Omega\; e^{i\lambda\alpha(p+\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(p,-\boldsymbol{\kappa})\\ &=\int d\Omega\; e^{i\lambda\alpha(p+\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(-p,\boldsymbol{\kappa})\\ &=\frac{\sqrt{2\pi}}{\alpha^2}\int d\Omega\; e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}}\mathbf{Q}_\lambda(\boldsymbol{\kappa})s_\lambda(\lambda\alpha\boldsymbol{\kappa}).\end{aligned}\tag{D2}$$

Now replace the Radon transform on the left-hand side of (D1) using (C2) to find

$$\int d\Omega\; e^{-i\lambda\alpha(p-\boldsymbol{\kappa}\cdot\mathbf{x})}\mathbf{Q}_\lambda(\boldsymbol{\kappa})\mathbf{Q}^*_\lambda(\boldsymbol{\kappa})\cdot\mathbf{B}^R_\lambda(\mathbf{p})=\frac{\sqrt{2\pi}}{\alpha^2}\int d\Omega\; e^{i\lambda\alpha\boldsymbol{\kappa}\cdot\mathbf{x}}\mathbf{Q}_\lambda(\boldsymbol{\kappa})s_\lambda(\lambda\alpha\boldsymbol{\kappa}),\tag{D3}$$

and the relation (D1) is seen to be true, leading to (4.12).

APPENDIX E. THE FOURIER TRANSFORM OF THE LUNDQUIST FIELD

The Fourier transform of the Lundquist field has the general form

$$\begin{aligned}\mathbf{B}^F_{\mathrm{L}\lambda}(\mathbf{k})=(2\pi)^{-3/2}\int d^3x\,\mathbf{B}_{\mathrm{L}\lambda}(\mathbf{x})e^{-i\mathbf{x}\cdot\mathbf{k}}&=(2\pi)^{-3/2}B_{\lambda 0}\int d^3x[-J_1(\alpha r)\sin\varphi\;\mathbf{i}_x\\ &+J_1(\alpha r)\cos\varphi\;\mathbf{i}_\mathbf{y}+J_0(\alpha r)\mathbf{i}_z]e^{-i\mathbf{x}\cdot\mathbf{k}}.\end{aligned}\tag{E1}$$

The scalar product appearing in the exponential argument is (in circular cylindrical coordinates)

$$\mathbf{k}\cdot\mathbf{x}=k_r r\,\cos(\varphi-\psi)+k_z z,\tag{E2}$$

where (k_r,ψ,k_z) refer to transform space and (r,φ,z) refer to physical space. The integral over z is immediate:

$$\mathbf{B}^F_{\mathrm{L}\lambda}(\mathbf{k}) = (2\pi)^{-1/2} B_{\lambda 0}\delta(k_z)\int r\ dr\ d\varphi[-J_1(\alpha r)\sin\ \varphi\ \mathbf{i}_x + J_1(\alpha r)\cos\ \varphi\ \mathbf{i}_y + J_0(\alpha r)\mathbf{i}_z]e^{-ik_r r\ \cos(\varphi-\psi)}. \tag{E3}$$

The integral over the azimuthal angle requires some care. It is composed of three integrals of the form

$$I_{s/c/1}(\beta,\psi) \equiv \int_0^{2\pi} d\varphi[\sin\ \varphi/\cos\ \varphi/1]e^{i\beta\ \cos(\varphi-\psi)}, \tag{E4}$$

where here $\beta = -k_r r$. These integrals are closely related to Bessel's original definition of the Bessel coefficients and can each be shown to have the form of Hansen's integral formula for the $J_m(z)$ Bessel function (Korn and Korn[25]):

$$J_m(z) = \frac{(-i)^m}{\pi}\int_0^{\pi} dt\ e^{iz\ \cos t}\cos(mt). \tag{E5}$$

Rewriting the angular argument in (E4) as $\varphi - \psi \equiv \varphi'$ changes the limits and expands the trig function factors of $I_{s/c}$ to give both sine and cosine contributions, of which only the cosine pieces survive the integration. The integral values are easily shown to be insensitive to the limit shifts and the surviving integrals evaluated by splitting the integration range and recognizing the Bessel functions from Hansen's formula:

$$I_s(\beta,\psi) = \sin\ \psi\int_0^{2\pi} d\varphi'\ \cos\ \varphi' e^{-i\beta\ \cos\ \varphi'} = \pi i\ \sin\ \psi[J_1(\beta) - J_1(-\beta)] = 2\pi i\ \sin\ \psi\ J_1(\beta), \tag{E6}$$

$$I_c(\beta,\psi) = \cos\ \psi\int_0^{2\pi} d\varphi'\ \cos\ \varphi' e^{-i\beta\ \cos\ \varphi'} = 2\pi i\ \cos\ \psi\ J_1(\beta), \tag{E7}$$

$$I_1(\beta,\psi) = \int_0^{2\pi} d\varphi'\ e^{-i\beta\ \cos\ \varphi'} = 2\pi i J_0(\beta). \tag{E8}$$

Thus, the complete φ integral becomes

$$\begin{aligned}\mathbf{B}^{\mathrm{F}}_{\mathrm{L}\lambda}(\mathbf{k}) &= (2\pi)^{-1/2} B_{\lambda 0}\delta(k_z)\int_0^{\infty} r\ dr[-J_1(\alpha r)I_s\mathbf{i}_x + J_1(\alpha r)I_c\mathbf{i}_y + J_0(\alpha r)I_1\mathbf{i}_z] \\ &= (2\pi)^{+1/2} i B_{\lambda 0}\delta(k_z)\int_0^{\infty} r\ dr[J_1(\alpha r)J_1(kr)\sin\ \psi\ \mathbf{i}_x - J_1(\alpha r)J_1(kr)\cos\ \psi\ \mathbf{i}_y \\ &\quad - iJ_0(\alpha r)J_0(kr)\mathbf{i}_z].\end{aligned} \tag{E9}$$

The remaining integrals are all instances of the delta function according to the familiar formula (Korn and Korn,[25] Eqs. (21.8–22.9):

$$\int_0^{\infty} r\ dr\ J_1(kr)J_1(k'r) = \frac{\delta(k-k')}{k'}, \tag{E10}$$

where $k,\ k' \geqslant 0$. Now making use of the explicit helicity expression analogous to (5.8) and integrating, the Fourier transform of the Lundquist field becomes

$$\mathbf{B}^F_{\mathrm{L}\lambda}(\mathbf{k}) = (2\pi)^{+1/2} i B_{\lambda 0}\delta(k_z)\int_0^\infty r\, dr[\lambda J_1(kr)J_1(\lambda\alpha r)\sin\psi\,\mathbf{i}_x$$
$$-\lambda J_1(kr)J_1(\lambda\alpha r)\cos\psi\,\mathbf{i}_y - iJ_0(kr)J_0(\lambda\alpha r)\mathbf{i}_z]$$
$$= (2\pi)^{+1/2} i B_{\lambda 0}\delta(k_z)\,\frac{\delta(k-\lambda\alpha)}{\lambda\alpha}\,[\lambda\,\sin\psi\,\mathbf{i}_x - \lambda\,\cos\psi\,\mathbf{i}_y - i\mathbf{i}_z]. \tag{E11}$$

APPENDIX F: THE FOURIER TRANSFORM OF THE CLASSICAL SPHEROMAK FIELD

Using the expansion of the exponential kernel outlined in Sec. V C, the Fourier transform of the classical spheromak field takes the form

$$\mathbf{b}_\lambda(\mathbf{k}) = (2\pi)^{-3/2}\sum_{n=0}^{\infty}(-i)^n(2n+1)\int d^3x\; j_n(kr)[pa_n(\mu,\mu') + pb_n(\mu,\mu',\varphi,\psi)]\mathbf{B}_\lambda(\mathbf{x}). \tag{F1}$$

Rewriting (5.16) as

$$B_x(r,\mu,\varphi) = B_1\,\cos\varphi + B_2\,\sin\varphi,$$
$$B_y(r,\mu,\varphi) = B_1\,\sin\varphi - B_2\,\cos\varphi, \tag{F2}$$
$$B_z(r,\mu) = B_3,$$

where we have defined

$$B_1 \equiv -\tfrac{1}{3}\,B_0 j_2(\alpha r)P_2^1(\mu),$$
$$B_2 \equiv B_0 j_1(\alpha r)P_1^1(\mu), \tag{F3}$$
$$B_3 \equiv \tfrac{2}{3}\,B_0[j_2(s)P_2(\mu) + j_0(s)P_0(\mu)],$$

clarifies the integration operations. In integrating (F1) over φ, the orthogonality of the trig functions forces the pa contribution to zero except for that from B_3 with $m=0$, and the pb contribution to zero except for that from B_1 and B_2 with $m=1$:

$$\mathbf{b}_\lambda(\mathbf{x}) = (2\pi)^{-1/2}\int_0^\infty r^2\, dr\sum_{n=0}^{\infty} i^n(2n+1)j_n(kr)$$
$$\times[(I_1\,\cos\psi + I_2\,\sin\psi)\mathbf{i}_x + (I_1\,\sin\psi - I_2\,\cos\psi)\mathbf{i}_y + I_3\mathbf{i}_z]. \tag{F4}$$

Here the integrals over μ have been explicitly indicated by

$$I_1 \equiv \int_{-1}^{+1} d\mu\; pb_n(\mu,\mu')\delta_{m1}B_1,$$
$$I_2 \equiv \int_{-1}^{+1} d\mu\; pb_n(\mu,\mu')\delta_{m1}B_2, \tag{F5}$$
$$I_3 \equiv \int_{-1}^{+1} d\mu\; pa_n(\mu,\mu')\delta_{m0}B_3.$$

Fortunately, the Kronecker deltas force the m values of the pb Legendre functions to coincide with those of the B_i, and the orthogonality of the Legendre functions then collapses the n values to coincide with those of the B_i:(2;1;0,2), so that the Fourier transform takes the form

$$\mathbf{b}_\lambda(\mathbf{k}) = \frac{2}{15}(2\pi)^{-1/2} B_0 \sum_{n=0}^{\infty} (-i)^n (2n+1) \int_0^\infty r^2\, dr\; j_n(kr)\{[-j_2(\alpha r) P_2^1(\mu')\delta_{n2} \cos\psi + 5 j_1(\alpha r) P_n^1(\mu')\delta_{n1} \sin\psi]\mathbf{i}_x - [j_2(\alpha r) P_2^1(\mu')\delta_{n2} \sin\psi + 5 j_1(\alpha r) P_n^1(\mu')\delta_{n1} \cos\psi]\mathbf{i}_y + 2[j_2(\alpha r) P_n(\mu')\delta_{n2} + 5 j_0(\alpha r) P_n(\mu')\delta_{n0}]\mathbf{i}_z\}. \tag{F6}$$

Thanks to the Kronecker deltas, the radial integrals are now all of the form

$$\int_0^\infty r^2\, dr\; j_n(kr) j_n(\alpha r) = \frac{\pi}{2} \frac{\delta(k-\alpha)}{k\alpha}. \tag{F7}$$

Note, however, that this formula assumes that k and α are both positive; since α may be negative, the following modified formula is necessary:

$$\int_0^\infty r^2\, dr\; j_n(kr) j_n(\alpha r) = \lambda^n \int_0^\infty r^2\, dr\; j_n(kr) j_n(\lambda\alpha r) = \lambda^n \frac{\pi}{2} \frac{\delta(k-\lambda\alpha)}{\lambda k \alpha}, \tag{F8}$$

in which we have taken account of the parity of the spherical Bessel functions. Making use of this formula in (F6), we obtain the final form of the classical spheromak Fourier transform:

$$\mathbf{b}_\lambda(\mathbf{k}) = \frac{\lambda}{3}\sqrt{\frac{\pi}{2}} B_0 \frac{\delta(k-\lambda\alpha)}{k\alpha}\{[P_2^1(\mu')\cos\psi - 3i\lambda P_1^1(\mu')\sin\psi]\mathbf{i}_x + [P_2^1(\mu')\sin\psi + 3i\lambda P_1^1(\mu')\cos\psi]\mathbf{i}_y + 2[-P_2(\mu') + P_0(\mu')]\mathbf{i}_z\}. \tag{F9}$$

[1] M. A. MacLeod, J. Math. Phys. **36**, 2951 (1995).
[2] M. A. MacLeod, Inverse Probl. **11**, 1087 (1995).
[3] H. E. Moses, SIAM J. Appl. Math. **21**, 114 (1971).
[4] L. F. Burlaga, *Interplanetary Magnetohydrodynamics* (Oxford University Press, New York, 1995).
[5] K. G. Ivanov and A. F. Harshiladze, Geomagn. Aeron. **25**, 314 (1985).
[6] B. L. Wright, Nucl. Fusion **30**, 1739 (1990).
[7] M. A. MacLeod, EOS Trans. Am. Geophys. Union **77**, S207 (1996).
[8] P. A. Sturrock, *Plasma Physics* (Cambridge University Press, New York, 1994).
[9] S. R. Deans, *The Radon Transform and Some of Its Applications* (Wiley, New York, 1983).
[10] S. Lundquist, Ark. Fys. **2**, 361 (1950).
[11] M. N. Rosenbluth and M. N. Bussac, Nucl. Fusion **19**, 489 (1979).
[12] J. Sheffield, Rev. Mod. Phys. **66**, 1015 (1994).
[13] G. C. Goldenbaum, Phys. Scr. **T2/2**, 359 (1982).
[14] M. Vandas, S. Fischer, and A. Geranios, Planet. Space Sci. **39**, 1147 (1991).
[15] C. J. Farrugia, V. A. Osherovich, and L. F. Burlaga, J. Geophys. Res. **100**, 12 293 (1995).
[16] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).
[17] E. Richter, Z. Phys. **159**, 194 (1960).
[18] Yu. P. Emets' and Yu. P. Kovbasenko, Sov. Phys. Tech. Phys. **28**, 879 (1984).
[19] E. Richter, Z. Naturforsch. A **49**, 902 (1994).
[20] H. E. Moses and R. T. Prosser, Proc. R. Soc. London, Ser. A **422**, 343 (1989).
[21] D. Ludwig, Commun. Pure Appl. Math. **19**, 49 (1966).
[22] V. Guillemin, Adv. Math. **22**, 85 (1976).
[23] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), Vol. I.
[24] S. Helgason, *The Radon Transform* (Birkhäuser, Boston, 1980).
[25] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers* (McGraw-Hill, New York, 1968).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 3 MARCH 1998

The identities of the algebraic invariants of the four-dimensional Riemann tensor. II.

G. E. Sneddon[a)]
School of Computer Science, Mathematics and Physics, James Cook University, Townsville, 4811, Australia

(Received 11 September 1997; accepted for publication 28 October 1997)

This paper makes use of several known results from invariant theory to further investigate the invariants of the Riemann tensor and the identities between them. The work also relies on the idea of expressing the Riemann tensor in terms of two complex matrices in the space of self-dual bivectors. It is shown that all invariants that are of even degree in the trace-free Ricci tensor can be written as polynomial functions of a set of 28 invariants. It is believed that this set is a complete set for this type of invariant. Several identities for matrix polynomials of 3×3 matrices are also obtained. These are used to find a large number of the identities between the 28 invariants. While some of these identities are quite complicated, it is shown how, in the general case, they might be used to obtain all invariants from a knowledge of a smaller set. © *1998 American Institute of Physics.*
[S0022-2488(98)02102-1]

I. INTRODUCTION

Existing classifications of the Riemann tensor depend on separate classifications of the Weyl tensor and the Ricci tensor. In turn, the classification of these tensors depends on the relationships between their algebraic invariants. Any attempt to classify the Riemann tensor as a whole can therefore be expected to depend on an understanding of all of the invariants of the Riemann tensor and the relationships between them. It is known that fourteen of these invariants are independent and it is not too difficult to write down a set of fourteen such invariants. However, as noted by Carminati and McLenaghan,[1] in many situations it is important to have a complete set of invariants. Any other invariant can be written as a polynomial function of the elements of such a set. This will ensure that this set is sufficient to determine any other invariant in any algebraically special case as well as in the general case. The task of finding such a set and the relationships between these invariants has been hampered by a lack of understanding of how these relationships may be obtained and (no doubt) by the considerable algebra involved.

Previous discussions of this problem have used either the tensor representation,[2] or the spinor representation[1,3,4] or a complex bivector representation[5,6] of the Riemann tensor. The latter approach was also taken by Sneddon,[7,8] who made use of the formalism developed by Buchdahl[9] to express the problem in the space of self-dual bivectors. Buchdahl referred to self-dual bivectors as rotors and a brief summary of his procedure is given in Ref. 8. The main advantage of expressing the invariants in terms of self-dual bivectors is that they become essentially the invariants of a pair of complex 3×3 matrices under a given group of transformations. One of these matrices, Ψ, corresponds to the Weyl tensor, and the other, Γ, corresponds to the trace-free part of the Ricci tensor. Sneddon[8] has given a complete solution for the invariants of two symmetric 3×3 matrices under similarity transformations. However, in the present case, the matrix corresponding to the trace-free part of the Ricci tensor is Hermitian and undergoes the transformation $\Gamma\rightarrow S\Gamma\overline{S}^{-1}$. This fact complicates the process of determining the relationships between the invariants quite considerably, as the preliminary investigations indicate.

Much of this work has been carried out without taking full account of known results in invariant theory. Indeed, since some of the results are quite fundamental properties of matrices, it is perhaps not surprising that many of them have appeared in the literature previously. Before pro-

a)Electronic mail: Graeme.Sneddon@jcu.edu.au

 © 1998 American Institute of Physics

ceeding to find a complete set of invariants for the Riemann tensor, it is important to take account of these earlier results. Section II describes some of the results from invariant theory that have particular relevance for the task of describing the invariants of the Riemann tensor. In Sec. III, some of the results presented in Ref. 8 are revisited with these ideas in mind. Also, the relationship between the invariants of non-symmetric matrices is presented. In Sec. IV, a number of identities between matrix products are investigated, and it is shown that there must be relationships between them (identities of identities). This feature is also well known and has important implications for later work. Furthermore, some of these identities can be found, and it is shown that there must be an even higher level of identities (''identities cubed''). The problem of finding a complete set of invariants for the Riemann tensor is addressed in Sec. V. A set containing 28 elements is given which is believed to be a complete set for the invariants of even degree in the Ricci tensor. It can be shown that this set must contain a complete set, but it has yet to be proved that no proper subset can be complete. In Sec. VI, several identities are found for these invariants. A study of these helps to see how the new invariants are related to those used previously.

A number of the identities obtained are identities for matrix polynomials. A matrix product of a set of R matrices, A_i, can be written as $A_{i_1}^{\alpha_1}A_{i_2}^{\alpha_2}\cdots A_{i_e}^{\alpha_e}$, where each of the subscripts lies between 1 and R and each of the indices is a positive integer. The total degree of the matrix product in the matrices A_i is $\alpha=\alpha_1+\alpha_2+\cdots+\alpha_e$. The partial degree of a given matrix, A, in the matrix product is the sum of the indicies α_k for which $A_{i_k}=A$. The degree of a trace can be defined similarly. A matrix polynomial is a sum of terms, each of which is a matrix product multiplied by a coefficient which is either a constant or the product of one or more traces of matrices. (Spencer and Rivlin[10] referred to these as *isotropic* matrix polynomials.) The degree of each such term is the sum of the degrees of the matrix product and its coefficient. The degree of a matrix polynomial is the degree of the term with the largest degree.

Consider a matrix polynomial, P, for which α is the total degree of the matrix product of largest total degree. Then P is reducible if it can be re-expressed as a matrix polynomial, Q, for which each matrix product has total degree less than α. This shall be denoted by $P\approx 0$, and the notation $P_1\approx P_2$ shall indicate that the difference of the two matrix polynomials is reducible. Similarly, the trace, T, of a matrix (or matrix polynomial) is said to be reducible if it can be written as a polynomial whose arguments are each traces of lower total degree than T. Again, this shall be written as $T\approx 0$.

Square brackets will be used to denote the trace of a matrix. Thus $[AB]$ is the trace of the matrix AB. Indices $i,j,\ldots$ will range from 1 to 4, while bivector indices $A,B,\ldots$ will range from 1 to 3.

II. RESULTS FROM INVARIANT THEORY

The study of invariant theory dates back to the middle of the last century. At that time, attention was concentrated on the invariants of homogeneous polynomials and forms. Towards the end of the century, Hilbert produced the main results in this area. These results could also be applied to invariants of sets of vectors and tensors in general. Throughout, there are two fundamental problems, namely (i) the nature of the invariants and (ii) the nature of the relationships between them. Before describing the results, it is important to review various definitions that are needed.

Definition: An *invariant* of a set of tensors in n dimensions is a polynomial function of the components of those tensors that is invariant under some group of transformations of the tensors.

Such an invariant is often referred to as a polynomial invariant. The allowable invariants can depend on the group chosen. Often it will be the full group of linear transformations, $GL(n)$. However, if the Lorentz group is used, then the invariants can automatically involve the metric tensor, η_{ij}. Similarly, if the proper Lorentz group is used, then the Levi-Civita tensor density (ε_{ijkl} or ε^{ijkl}) can also be involved, even though this entity transforms as a tensor density rather than a tensor under $GL(n)$. It is common practice to include invariants formed using this tensor density amongst the invariants of the four dimensional Riemann tensor. This means that the quantities to be considered will be invariant under the proper Lorentz group.

A complete set may be defined as follows.

Definition: A *complete* set of invariants, $\{I_1, \ldots, I_m\}$, is one for which any other invariant I can be written as a polynomial function of $\{I_1, \ldots, I_m\}$, but no element of the set can be written as a polynomial function of the others.

Therefore, with a complete set of invariants, it will always be possible to calculate any other invariant. This will be the case even in algebraically special sub-cases where some invariants may vanish.

Definition: An *independent* set of invariants, $\{I_1, \ldots, I_m\}$, is one for which $f(I_1, \ldots, I_m)=0$ implies that f is the trivial function, $f \equiv 0$.

An independent set can be maximal in the sense that, for any other invariant I, there is a function g such that $g(I, I_1, \ldots, I_m)=0$. Even if this is the case, it may not be possible to calculate all other invariants in some algebraically special cases. For example, a calculation could involve division by an invariant that became zero in the algebraically special case. This feature has been seen in the invariants of the Riemann tensor.[1] On the other hand, a complete set may require such a large number of elements that its usefulness is limited. For this reason, Zakhary and McIntosh[4] and Carminati and McLenaghan[11] have considered other possible requirements for a convenient set to work with. In general, a complete set will not be independent. In those cases where it is not, there must be identities (syzygies) between the invariants in the set. An example of this is the identity given in Ref. 8 connecting the invariants of two symmetric 3×3 matrices.

With these ideas in mind, we can now turn to the main theorems.

(a) **The first fundamental theorem:** Any polynomial invariant of a set of tensors can be expressed as a linear combination of complete contractions of products of those tensors (together with the metric or $\varepsilon_{i_1 \cdots i_n}$ as appropriate).

This result is discussed by Weyl[12] and also by Gurevich.[13] It tells what form the invariants can take. Also, the Hilbert Finiteness Theorem states that the invariants are finitely generated. That is, a complete set will have a finite number of elements. The second fundamental theorem indicates the form of any identities between the invariants.

(b) **The second fundamental theorem:** Any identity between the invariants of a set of tensors in n dimensions can be obtained as a consequence of the fact that skew-symmetrizing over $n+1$ indicies will annihilate any tensor.

This refers to identities that are additional to those that can be obtained from the symmetries of the tensors themselves. The theorem can also be extended to matrix identities. In some cases, the identities themselves might also satisfy certain identities and these identities in turn may be related by further identities, etc. The Hilbert Syzygy theorem[14] states that this process must terminate and that the total number of identities is finite.

Special cases of the second fundamental theorem have been known for some time. In the case of sets of vectors, the proof was given by Weyl.[12] For sets of $n\times n$ matrices, which is the case to be considered here, the result has been obtained by Procesi.[15] The result for general tensors has recently been obtained by Gover.[16] He also allows for the possibility that the tensors themselves may have existing symmetries. He has shown that any identity between the invariants is an algebraic consequence of such symmetries, together with the identities which arise from the result that skew-symmetrizing over $n+1$ indicies annihilates any tensor in n dimensions. These results are obtained for the groups $O(g)$ and $SO(g)$, with the metric g unspecified, so they will be valid for the Lorentz group. The method is quite general and so should also apply to tensors with dotted indicies.

In a series of papers,[10,17,18] Spencer and Rivlin obtained several specific results for 3×3 matrices. They were dealing with stress tensors in isotropic continua and these tensors were expressed in terms of a number of tensors representing the strain. Their work was prompted by the need to simplify the the form of the various products of these strain tensors. They showed how to express products of up to five 3×3 matrices in terms of a comparatively small ''basis set'' of such products. In so doing, they made extensive use of the relationship

$$AB^2+B^2A+BAB-[A]B^2-[B](AB+BA)+\tfrac{1}{2}([B]^2-[B^2])A+([A][B]-[AB])B-([AB^2]$$
$$-[B][AB]+\tfrac{1}{2}([A][B]^2-[A][B^2]))I=0, \tag{1}$$

which follows from

$$X^{[A}{}_AY^B{}_BZ^C{}_C\delta^{D]}{}_E=0 \tag{2}$$

by putting $X=Y=A$, and $Z=B$. They were able to show that any matrix polynomial involving only two matrices could be written in terms of a set of 17 matrix products. (Only nine matrix products are needed if the matrices and matrix polynomial are symmetric.) Equation (2) contains the Cayley–Hamilton equation as a special case, but it can also be deduced from the Cayley-Hamilton theorem by a ''polarization'' process.[10] Of course, it is also an example of where skew-symmetrizing over $n+1$ indicies can lead to matrix identities as well as trace identities. The basis set obtained by Spencer and Rivlin was equivalent to the set

$$\begin{gathered} I \\ A,\ B \\ A^2,\ AB,\ BA,\ B^2 \\ A^2B,\ AB^2,\ BA^2,\ B^2A \\ A^2B^2,\ A^2BA,\ B^2AB,\ B^2A^2 \\ A^2B^2A,\ B^2A^2B. \end{gathered} \tag{3}$$

If A and B are symmetric, then any symmetric matrix polynomial can be written in terms of the matrices

$$\begin{gathered} I \\ A,\ B \\ A^2,\ AB+BA,\ B^2 \\ A^2B+BA^2,\ AB^2+B^2A \\ A^2B^2+B^2A^2. \end{gathered} \tag{4}$$

The sets (3) and (4) shall be referred to as $\mathscr{M}$ and $\mathscr{S}$ respectively. Spencer and Rivlin gave equations that could be used to write any other matrix polynomial in terms of the elements of one of these sets as appropriate. Often the final result was quite lengthy and was not given explicitly. For example, the expression for A^2B^2AB in terms of the set (3) requires 78 terms. This expression is given in Appendix A for the case where both A and B are trace-free. This still requires 22 terms.

Spencer and Rivlin also obtained many of the identities between the traces of matrix products (of up to five 3×3 matrices) that can be obtained by skew-symmetrizing over four indicies. Since the components of the Riemann tensor can be written in terms of the components of 3×3 matrices, their work is particularly relevant. In fact, it will be shown that their results can be used to find a complete set of invariants of the Riemann tensor.

Initially, these results were used to show that the set

$$\begin{array}{cccc} & [A] & [A^2] & [A^3] \\ [B] & [AB] & [A^2B] & \\ [B^2] & [AB^2] & [A^2B^2] & \\ [B^3] & & & \end{array} \tag{5}$$

forms a complete set for the invariants of two symmetric 3×3 matrices under similarity transformations. This result was re-derived by Sneddon,[8] who also obtained the identity $S_0=0$, where

$$S_0=c_3[A^2B^2]^3+c_2[A^2B^2]^2+c_1[A^2B^2]+c_0, \tag{6}$$

and

$$c_3 = -4,$$

$$c_2 = [AB]^2 + 5[A^2][B^2],$$

$$c_1 = 2[B^2][A^2B]^2 + 2[A^2][AB^2]^2 - 2[AB][AB^2][A^2B] - 2[A^2][B^3][A^2B] - 2[A^3][B^2][AB^2] - [A^2][B^2][AB]^2 + 2[A^3][B^3][AB] - 2[A^2]^2[B^2]^2,$$

$$\begin{aligned} c_0 = &-\tfrac{4}{3}([B^3][A^2B]^3 + [A^3][AB^2]^3) + [A^2B]^2[AB^2]^2 \\ &- \tfrac{1}{2}[A^2][B^2]([B^2][A^2B]^2 + [A^2][AB^2]^2) + 2[A^3][B^3][A^2B][AB^2] \\ &- \tfrac{4}{9}[A^3][B^3][AB]^3 + \tfrac{2}{3}[AB]^2([A^3][B^2][AB^2] + [A^2][B^3][A^2B]) \\ &+ \tfrac{1}{3}[AB]([A^2]^2[B^3][AB^2] + [A^3][B^2]^2[A^2B]) \\ &+ \tfrac{2}{3}[A^2][B^2]([A^3][B^2][AB^2] + [A^2][B^3][A^2B]) - \tfrac{2}{3}[A^2][A^3][B^2][B^3][AB] \\ &+ \frac{1}{4}[A^2]^3[B^2]^3 + \tfrac{1}{18}([A^3]^2[B^2]^3 + [A^2]^3[B^3]^2) - \tfrac{1}{3}[A^3]^2[B^3]^2. \end{aligned}$$

This identity is valid for all trace-free, symmetric matrices. The corresponding equation for the case where the matrices are not trace-free can also be obtained. A simple counting argument on the number of independent components remaining after all the coordinate freedom is used up, shows that nine of the invariants in (5) must be independent. This means that the relationship $S_0 = 0$ (or its generalization if the traces are non-zero) is the one relationship between the elements of the complete set for the invariants of a pair of symmetric matrices. The set (5) shall be denoted by $\mathscr{I}$.

If either A or B is not symmetric, $\mathscr{I}$ needs to be augmented with either $[A^2B^2AB]$ or $[B^2A^2BA]$ (or their difference) to form a complete set. Since only ten of these invariants can be independent, there must again be a relationship connecting them. This relationship is obtained in the next section.

Finally, identities such as (2), which rely explicitly on skew-symmetrizing over $n+1$ indicies, have been used in general relativity on other occasions. Lovelock[19] showed that a number of well known identities in general relativity are consequences of this procedure. Also Harvey[2] used this property in his work on invariants, and Edgar and co-workers (see Ref. 20 for example) have used the property to obtain equations for the Lanczos tensor and the Weyl tensor. Lovelock referred to such identities as ''dimensionally dependent'' identities.

III. TRACE IDENTITIES

The results of the previous section raise a number of questions concerning the trace identities of 3×3 matrices. Some of these questions are relevant to the aim of understanding the invariants of the Riemann tensor.

Firstly, the second fundamental theorem states that all trace identities can be obtained by skew-symmetrizing over four indicies. It is clear that this is the case for the identities obtained by Spencer and Rivlin. However, it is not obviously the case for the identity $S_0 = 0$. This identity was obtained in Ref. 8 by two brute force methods, neither of which made use of this theorem. It needs to be confirmed that this identity can also be obtained by skew-symmetrizing over four indicies. Secondly, in the case that A and B are not symmetric, it would be of interest to obtain an identity that connects $[A^2B^2AB]$ with the set $\mathscr{I}$. This would be a relationship between the elements of a complete set for the invariants of two 3×3 matrices under similarity transformations.

It is best to answer the second question first. The answer to the first then follows easily. It has been conjectured[8] that the required identity could possibly be obtained from

$$[C]^3 - 3[C][C^2] + 2[C^3] - 6\det C = 0 \tag{7}$$

with $C=A^2B^2AB$ and $\det C$ replaced by $(\det A\ \det B)^3$. This was partly because a cubic expression for $[A^2B^2AB]$ was expected. However, it was found that the left-hand side (LHS) of (7) actually factorizes and that one of these factors gives the correct identity, which is a quadratic in $[A^2B^2AB]$. It should be possible to obtain this identity as the single solution of some equation.

In view of the second fundamental theorem, it can be expected that this identity will result from an equation such as

$$X^{[A}{}_{A}Y^{B}{}_{B}Z^{C}{}_{C}W^{D]}{}_{D}=0 \tag{8}$$

with an appropriate choice for X,Y,Z and W. Also, it is clear that it will be sufficient to choose these matrices from amongst those in the basis set found by Spencer and Rivlin. In order to find a suitable choice, a multiplication table was set up for the matrices in $\mathscr{M}$ together with A^2B^2AB and B^2A^2BA. It was assumed that both matrices were trace-free to reduce the amount of algebra. The results for the general case could be deduced if needed. Initially, expressions were found for the product of each of these matrices with A and B. Each such product was expressed in terms of the matrices in this set. These calculations were carried out largely by hand, using identities such as (1). Once this was done, a full multiplication table could be produced. An algebraic computing package was needed for this second step as the algebra required for some products was quite considerable, even with the trace-free assumption. (The package actually used was MATHEMATICA.) For example, the expression for $A^2BA\times B^2A^2B$ contains 45 terms. Subsequently, the expressions for A^2B^2AB and B^2A^2BA in terms of $\mathscr{M}$ were found.

With this table available, Eq. (8) can be written out in full and various choices of matrices can be tried. In fact there are many combinations that will give the required identity. Perhaps the simplest choice is to put

$$X=A^2B^2,$$

$$Y=A^2B^2,$$

$$Z=AB,$$

$$W=AB.$$

The resulting identity is

$$\Delta^2=-(c_3[A^2B^2]^3+c_2[A^2B^2]^2+c_1[A^2B^2]+c_0), \tag{9}$$

where $\Delta=[A^2B^2AB]-[B^2A^2BA]$ and the coefficients c_i are the ones that follow Eq. (6). This equation can be generalized to the case where A and B are not trace-free. Thus we have a complete set of invariants for a pair of matrices, and (9) (or its generalization) gives the relationship between them.

It is now straightforward to see how to obtain the identity for symmetric matrices. If A and B are symmetric, then $[A^2B^2AB]=[B^2A^2BA]$ and so $\Delta=0$. Therefore

$$S_0=c_3[A^2B^2]^3+c_2[A^2B^2]^2+c_1[A^2B^2]+c_0=0.$$

Therefore, this equation can be obtained by skew-symmetrizing over four indices as predicted by the second fundamental theorem.

Finally, while this method of deriving this identity may seem more elegant than the previous approaches, it is no less cumbersome. In fact, the calculations required in setting up the multiplication table mean that probably more work is required. It does have the advantage though that, once constructed, the multiplication table can be used for other things.

IV. MATRIX-VALUED IDENTITIES

Identities such as (1) are matrix-valued identities. They can be used to give expressions for products of matrices in terms of the basis set $\mathscr{M}$. For example, Eq. (1) itself can be used to give BAB in terms of AB^2 and B^2A and the other elements of $\mathscr{M}$. Since the coefficient of BAB does

not involve any of the invariants, this equation can always be solved for BAB. There are, however, a number of other matrix identities where the coefficient of each matrix product involves some of the invariants. The solution of these equations for one of its matrix products is not guaranteed, as it may involve division by zero in some cases. Consideration of the basis set $\mathscr{M}$ also suggests that there must be further identities between these matrix products. That is, while there are 17 elements of that set, only nine can be independent as there are only nine independent components of a 3×3 matrix. Similar comments will apply to the elements of $\mathscr{S}$ as there can be only six independent symmetric matrices.

The task then is to find some of these additional matrix identities for the elements of $\mathscr{M}$. This can be done using the equation

$$X^{[A}{}_{A}Y^{B}{}_{B}Z^{C}{}_{C}W^{D]}{}_{E}=0.$$

and replacing X, Y, Z and W with different elements of $\mathscr{M}$ in turn. Equivalently, Eq. (2) could be used, and any identity obtained could be multiplied by a matrix W. In fact, *any* further identity can be obtained in this fashion. Of course, a large number of identities have already been used in the construction of the multiplication tables, and so some of the choices for X, Y, Z and W will simply give the trivial identity. The new identities obtained in this fashion can be characterized by their partial degrees in A and B. Also, for given degrees of A and B, the identities can be separated into a symmetric part and a skew-symmetric part. It appears that this classification is sufficient to identify each of these identities. For example, the symmetric identity of degree six in A and degree four in B is unique to within a constant multiple and the possible addition of multiples of identities of lower total degree. This identity can therefore be referred to as $M^{64}{}_{s}$. The corresponding skew-symmetric identity will be $M^{64}{}_{a}$ and so on.

Now, the only syzygy connecting elements of $\mathscr{T}$ is of sixth degree in A and B, and the elements of $\mathscr{M}$ that are of largest degree are A^2B^2A and B^2A^2B. It follows that any further matrix identity must be of at least the third degree in A and fourth degree in B or vice versa. (Any identity of lower degree would have to be the trivial identity, as the trace of its product with any matrix product would have to be zero.) Choosing $X=B$, $Y=AB$, $Z=A^2B^2$ and $W=I$ gives M^{34}. M^{43} can be obtained similarly. These identities are skew-symmetric, thus removing the need for the subscript. Several more identities can now be obtained, either by different choices of X, Y, Z and W, or by multiplying existing identities such as M^{34} and M^{43} by A or B, etc. From above, it is expected that there would be eight such matrix identities connecting the 17 ''basis'' matrix products. The difficulty now is that, rather than having too many matrix products, there appears to be too many identities between them. Therefore some of these matrix identities must be dependent. In fact, the situation is a little more subtle than that. This will be illustrated by letting A and B be symmetric and considering $M^{nn}{}_{s}$, the identities between symmetric matrix polynomials. These considerations will also assist in understanding the relationships between the invariants of the Riemann tensor.

The basis, $\mathscr{S}$, for symmetric matrix polynomials formed from symmetric matrices is given by Eq. (4). Since there are nine elements in this set, and there are six independent symmetric 3×3 matrices, there should be exactly three independent relations between them. These equations could then be solved to give, say, $A^2B^2+B^2A^2$, A^2B+BA^2 and AB^2+B^2A in terms of the other matrices in $\mathscr{S}$. However, the foregoing procedure actually produces nine identities that appear to be independent. These are

$$\begin{matrix} M^{44}{}_{s} & M^{54}{}_{s} & M^{64}{}_{s} \\ M^{45}{}_{s} & M^{55}{}_{s} & M^{65}{}_{s} \\ M^{46}{}_{s} & M^{56}{}_{s} & M^{66}{}_{s}. \end{matrix} \tag{10}$$

Once again, there seems to be too many identities. We shall look more closely at $M^{44}{}_{s}$, $M^{45}{}_{s}$, $M^{54}{}_{s}$ and $M^{55}{}_{s}$. For brevity, these identities will be written as $S_i=0$. They are given in Appendix A for the case $[A]=[B]=0$. At first glance, they appear to be independent. For example, these identities can be viewed as a system of four equations to determine $A^2B^2+B^2A^2$, A^2B+BA^2, AB^2+B^2A and $AB+BA$ in terms of the other matrices. The determinant of the coefficient matrix can be calculated, and it is a polynomial expression in the invariants $\mathscr{T}$. Surprisingly, this expres-

sion is not identically zero, which suggests that the equations are independent. However, it turns out that the expression for the determinant factorizes, and that one of the factors is the expression S_0 defined in Eq. (6). This cubic is known to be identically zero when A and B are symmetric 3×3 matrices. Therefore, the equations are actually dependent. This must also be the situation for other sets of four equations.

Since these identities are dependent, there should be invariant expressions a_i such that $\Sigma_{i=1}^{4} a_i S_i = 0$. However, the term on the LHS can be zero only as a consequence of the cubic identity $S_0 = 0$. Therefore, the equation must read

$$\sum_{i=1}^{4} a_i S_i = -a_0 S_0, \tag{11}$$

where a_0 is a symmetric matrix polynomial. Such an equation would show clearly the relationship between the syzygies $S_i = 0$. It would be a syzygy of syzgyies. In order to find this equation, the highest power of $[A^2B^2]$ in each of the a_i can be assumed to be two. The equation $S_0 = 0$ could be used to remove any higher powers. Thus the coefficients a_i can be written as $x_i[A^2B^2]^2 + y_i[A^2B^2] + z_i$. The LHS of Eq. (11) can then be expanded, and the coefficients of the powers of $[A^2B^2]$ can be collected. If the equation $S_0 = 0$ is again used to replace any powers higher than the second, the resulting expression must be zero. Furthermore each coefficient of the remaining powers of $[A^2B^2]$ must be identically zero and, within each such coefficient, the coefficients of each of the matrices in $\mathscr{S}$ must be zero. This gives 27 linear homogeneous equations for the 12 unknowns x_i, y_i, z_i, but it turns out that only nine are independent. Therefore, there are three different solutions for (x_i, y_i, z_i) and each of these gives rise to a different solution for a_i. Each of these solutions can be checked by substituting into the LHS of (11) and expanding. This time, the higher powers of $[A^2B^2]$ are not replaced. Instead, the resulting expression should factorize, with S_0 being one of the factors. In this fashion, it can be verified that each of the three choices for a_i gives a solution for (11). It should be noted that some of these calculations were carried out for the special case where $[AB] = [AB^2] = 0$ (as well as $[A] = [B] = 0$) in order to reduce the amount of algebra involved. As has been seen previously,[8] this special case is sufficient to construct the general result.

The fact that there are three solutions of (11) is a surprise at first. It suggests that there are actually three identities between the four relationships $S_i = 0$, when only one was expected. The only possible resolution of this is if these identities themselves also satisfy certain syzygies. These will be syzygies of syzygies of syzygies (or (syzygies)^3!). The possibility of such a hierarchy of syzygies is well known. There is a discussion in the recent publication of Hilbert's lectures on invariant theory.[14] Indeed, Hilbert's Syzygy Theorem states that, under fairly general conditions, such a chain of syzygies must be finite. In the present case, it is not too difficult to see the origin of these ''(syzygies)^3.'' Any particular identity of the type

$$a_1 S_1 + a_2 S_2 + a_3 S_3 + a_4 S_4 = -a_0 S_0$$

can be multiplied by $[A^2B^2]$. Each of the terms $a_i \times [A^2B^2]$ can be expanded and the higher powers of $[A^2B^2]$ can again be replaced. The resulting relationship will differ from the initial one and so must be some combination of the three relationships. This will most probably give one of the (syzygies)^3. Multiplication of the initial identity by $[A^2B^2]^2$ and repeating the process will most likely give the second.

Finally, it is clear that these results are not complete. There will be a hierarchy of syzygies connecting all the identities of (10). The present results simply provide a glimpse of the complexity of the problem. These same complexities arise in considering the invariants of the Riemann tensor. The results just presented will help in the understanding of that case.

V. THE INVARIANTS OF THE RIEMANN TENSOR

The invariants of the Riemann tensor can be expressed in terms of two complex 3×3 matrices. These are a trace-free symmetric matrix, Ψ, and a Hermitian matrix, Γ. (The Ricci scalar is excluded from consideration here.) From Ref. 9, if the components of the Riemann tensor undergo a Lorentz transformation, the elements of the bivector space (the rotors) undergo a complex

orthogonal transformation. That is, $\Psi \to S\Psi S^{-1}$ and $\Gamma \to S\Gamma \overline{S^{-1}}$ where $SS^T = I$. There are two consequences of this when compared with the work in the previous sections, where S could be a general non-singular matrix. Firstly, since the metric tensor is left invariant by the transformation, the transpose of matrices can appear in the invariants. Consequently, invariants such as $[AB^T]$ should also be considered. Secondly, since the Levi-Civita tensor density, ε_{ABC}, is left unchanged, it can also be used in the construction of the invariants. Since the matrices to be considered are symmetric, the first problem won't arise. Also, as discussed below, invariants containing ε_{ABC} will be excluded from consideration.

Two special subsets of the invariants of the Riemann tensor can be identified. They are (i) those invariants in which Γ appears only in the combination $\Gamma\overline{\Gamma}$ and (ii) those invariants which are of even degree in Γ. The first case was dealt with in Ref. 8, and further results for that case have been given in the preceding sections. This section will discuss the invariants of the second type. Clearly this set contains the elements of the first type. The second type of invariant can also be described as those that are also left unchanged by the transformation $\Gamma \to -\Gamma$. It turns out that the invariants that are not contained in this set are those that involve ε_{ABC}. For example, $\varepsilon_{ABC}\varepsilon^{\dot{A}\dot{B}\dot{C}}\Gamma^A{}_{\dot{A}}\Gamma^B{}_{\dot{B}}\Gamma^C{}_{\dot{C}}$ and $\varepsilon_{ABC}\varepsilon^{\dot{A}\dot{B}\dot{C}}\Psi^A{}_D\Gamma^D{}_{\dot{A}}\Gamma^B{}_{\dot{E}}\overline{\Psi}^{\dot{E}}{}_{\dot{B}}\Gamma^C{}_{\dot{C}}$ would be excluded. In the tensor representation it would mean that the invariant $R^i{}_j R^j{}_k R^k{}_i$ was not included. While these are no less important than the other invariants, it is felt that their inclusion at this stage would simply add to the complexity. In any case, the squares of these invariants will be included in this class of invariants, as will be the product of any two of them. In this sense, all invariants can be accounted for to within a minus sign.

Invariants of the first type were investigated by introducing matrices $A = \Psi$ and $B = L$, where L is the trace-free part of $K = \Gamma\overline{\Gamma}$. Even though the results obtained assumed that both matrices were trace-free, they could be used to construct the general results. A means is needed for describing invariants of the second type. They will be formed from the traces of products of matrices such as Ψ, $\overline{\Psi}$, K, $\overline{K}$, Γ, and $\overline{\Gamma}$. In these matrix products, there will need to be either a Γ or a $\overline{\Gamma}$ separating either of the matrices Ψ and K from the matrices $\overline{\Psi}$ and $\overline{K}$. Once again, the Cayley–Hamilton theorem guarantees that matrix powers higher than the second should not be included. These considerations suggest that any such invariant can be expressed in terms of the traces of products of the matrices

$$A = \Psi,$$

$$B = K,$$

$$C = \Gamma\overline{\Psi}\overline{\Gamma},$$

$$D = \Gamma\overline{\Psi}^2\overline{\Gamma}.$$

Note that A is trace-free, but there seems to be no advantage in choosing B to be the trace-free part of K, as the presence of Γ will mean that $[K]$ will appear in the final equations in any case. It is also important to note that each of these matrices is symmetric. Thus the earlier results for symmetric matrices will apply to any pair of these matrices (after allowing for the fact that they are not all trace-free).

Even though these matrices are not fully independent, there is no further simplification that can be made at this stage. For example, $C^2 = \Gamma\overline{\Psi}K\overline{\Psi}\overline{\Gamma}$ and the appropriate modification of Eq. (1) implies $\overline{\Psi}\overline{K}\overline{\Psi} \approx -(\overline{\Psi}^2\overline{K} + \overline{K}\overline{\Psi}^2)$. Therefore $C^2 \approx -(DB + BD)$. However, this does not provide a means to express D simply in terms of C^2. There seems to be no alternative other than to look at the invariants of these four matrices.

Fortunately, much of the work required here has already been done by Spencer and Rivlin.[17] They gave details of complete sets for the invariants of up to five symmetric 3×3 matrices. These results can be summarized as follows:

1. For two symmetric matrices A and B, a complete set is given by $\mathscr{T}$.
2. For three symmetric matrices A, B and C, a complete set consists of

TABLE I. Any invariant of Ψ and Γ that is of even degree in Γ can be written as a polynomial function of these invariants.

Real invariants	Complex invariants
$[B]=[K]$	
$[B^2]=[K^2]$	$[A^2]=[\Psi^2]$
$[B^3]=[K^3]$	$[A^3]=[\Psi^3]$
	$[AB]=[\Psi K]=\overline{[C]}$
	$[AB^2]=[\Psi K^2]=\overline{[BC]}$
	$[A^2B]=[\Psi^2K]=\overline{[D]}$
	$[A^2B^2]=[\Psi^2K^2]=\overline{[BD]}$
$[AC]=[\Psi\Gamma\overline{\Psi}\,\overline{\Gamma}]$	$[AD]=[\Psi\Gamma\overline{\Psi}^2\overline{\Gamma}]$
$[ABC]=[\Psi K\Gamma\Psi\Gamma]$	$=\overline{[A^2C]}$
$[AB^2C]=[\Psi K^2\Gamma\overline{\Psi}\,\overline{\Gamma}]$	$[ABD]=[\Psi K\Gamma\overline{\Psi}^2\overline{\Gamma}]$
$[A^2D]=[\Psi^2\Gamma\overline{\Psi}^2\overline{\Gamma}]$	$=\overline{[A^2BC]}$
$[A^2BD]=[\Psi^2K\Gamma\overline{\Psi}^2\overline{\Gamma}]$	$[AB^2D]=[\Psi K^2\Gamma\overline{\Psi}^2\overline{\Gamma}]$
$[A^2B^2D]=[\Psi^2K^2\Gamma\overline{\Psi}^2\overline{\Gamma}]$	$=\overline{[A^2B^2C]}$
$[A^2BACD]=[\Psi^2K\Psi\Gamma\overline{\Psi}\overline{K}\overline{\Gamma}^2\overline{\Gamma}]$	

(a) all elements of the complete sets of each of the three pairs of matrices that can be selected from A, B and C,
(b) the invariant $[ABC]$,
(c) the invariants $[ABC^2]$, $[BCA^2]$, $[CAB^2]$ and
(d) the invariants $[AB^2C^2]$, $[BC^2A^2]$, $[CA^2B^2]$.

3. For four symmetric matrices A, B, C and D, a complete set consists of

(a) all elements of the complete sets of each of the four sets of three matrices that can be selected from A, B, C and D,
(b) the invariants $[ABCD]$, $[ABDC]$,
(c) the invariants $[ABCD^2]$, $[ACBD^2]$, $[DABC^2]$, $[DBAC^2]$, $[CDAB^2]$, $[CADB^2]$, $[BCDA^2]$, $[BDCA^2]$,
(d) the invariants $[ABC^2D^2]$, $[ACB^2D^2]$, $[ADB^2C^2]$, $[BCA^2D^2]$, $[BDA^2C^2]$, $[CDA^2B^2]$ and
(e) the invariants $[BACDA^2]$, $[CBDAB^2]$, $[DCABC^2]$, $[ADBCD^2]$.

Initially, the task of writing out all these invariants, let alone making sense of them, seems quite formidable. However, there a number of simplifications available. Some of these are as follows:

1. Since $C^2\approx-(BD+DB)$, any invariants involving C^2 can be removed from the list. Similarly, $D^2\approx0$ so any invariant involving D^2 can also be removed. A similar simplification can not be made for CD, but the fact that $CD\approx-DC$ can be used.
2. Sometimes, when an invariant, or its complex conjugate, is written explicitly in terms of Ψ, K and Γ it is obvious that it is reducible. For example, $\overline{[B^2C]}=[K^3\Psi]$ is obviously reducible as it contains K^3.
3. The fact that the complex conjugate of an invariant doesn't appear on the list of Spencer and Rivlin is evidence that the invariant is reducible. $[ACD]$ is not obviously reducible, but its complex conjugate, $[A^2BAC]$, is not on the list given by Spencer and Rivlin. A little manipulation shows that $[A^2BAC]\approx-[ABA^2C]=-[A^2BAC]$. The first step follows from Eq. (1) and the second is true since the matrices are symmetric. Therefore, $2[A^2BAC]\approx0$ and $[ACD]$ can be removed from the list.

Once these simplifications are made, the remaining invariants can be listed. There are 28 of them, some of which appear in complex conjugate pairs. These invariants are given in Table I and shall be denoted by $\mathscr{J}$. The invariants contained in $\mathscr{I}$ are, of course, amongst these, but note that, since $A=\Psi$ which is trace-free, it is understood that, in any reference to $\mathscr{I}$, $[A]$ will be zero.

It is claimed that this set of invariants forms a complete set for the invariants of the Riemann tensor (excluding the Ricci scalar of course) that are of even degree in the trace-free Ricci tensor. Certainly any other invariant can be expressed as a polynomial function of the invariants in this set. However, it is not yet certain that this could not be achieved with a smaller set. The term that seems most likely to be reducible is $[A^2BACD]$. It is possible to obtain a number of syzygies containing this term, and this is discussed in the next section. However, none of these syzygies suggests that it might be possible to find an expression for $[A^2BACD]$. Even if such a relationship did exist, there would still be a large number of invariants in a complete set of invariants of the Riemann tensor. The inclusion of invariants that are of odd degree in the Ricci tensor may improve the situation, but this is unlikely. Whatever the situation, there may be some advantage in relaxing the requirement of a polynomial dependence, and looking for smaller sets of invariants that can still be used to calculate all other invariants in most situations.

It should also be noted that the above considerations are valid even if Ψ and $\overline{\Psi}$ are independent of each other. In some cases, it is best to regard them as independent, and also regard each complex invariant in Table I as being independent of its complex conjugate. Then, if $[\overline{\Psi}^2]$ and $[\overline{\Psi}^3]$ are included amongst the invariants, there will be a total of 28 independent terms representing the invariants in the set. These terms will include $[AB]$, $[C]$, $[ABD]$, $[A^2BC]$ etc. as independent invariants. In this fashion, the real invariants appear only once, and the complex invariants will appear in pairs which are complex conjugates when Ψ and $\overline{\Psi}$ are complex conjugates.

Finally, many of the ''new'' mixed invariants have appeared before. For example, $[AC]$ and $[A^2C]$ are essentially the invariants m_3 and m_5 defined by Carminati and McLenaghan. Their invariant m_4, as expressed by Sneddon,[8] contains the invariants $[ABC]$ and $[AB^2C]$. The expression M_4 defined by Haddow[4] is just $[A^2D]$.

VI. SYZYGIES OF THE INVARIANTS

The next step is to look for syzygies connecting elements of this set. Clearly, there will be a large number of these, and the connections between them can be quite complex. Rather than find all such syzygies, the initial aim will be to find a sufficient number to enable all the invariants in $\mathscr{J}$ to be calculated from those in $\mathscr{I}$ (at least in the general case). As may be expected, the equations needed to do this will be polynomial equations and their solution will not be unique. An alternative approach is to look for a set that is larger than $\mathscr{I}$, but from which the invariants in $\mathscr{J}$ can be calculated uniquely. Either of these two approaches gives some understanding of the relationships between the invariants and the syzygies. In general, even though the syzygies obtained are quite complex, they have a simple structure if the invariants in $\mathscr{I}$ are regarded as known quantities.

One method for finding syzygies would be first to produce a ''basis set'' for matrix polynomials formed from the four matrices A, B, C and D. Then construct a multiplication table for the elements of this basis set and use Eq. (8) to generate the syzygies. This would be a considerable task, especially when, at least in the first instance, only some of the simpler syzygies are required. A better approach would be to start with one of the matrix identities of Sec. IV, multiply this by another matrix such as C or D and take the trace. Provided all the traces can be written in terms of the elements of $\mathscr{J}$, a new syzygy should result. One other method is to start with a known identity and replace one of the matrices, A say, with a ''perturbed'' matrix, $A+\epsilon C$ say. The coefficient of ϵ in this perturbed identity will often be a new identity. This is equivalent to the method of polarization which is well known in invariant theory.[12] Both of these latter methods are used below.

The simplest method for generating syzygies is to start with the matrix identities M^{nn}, multiply by C and take the trace. Each of the trace terms can be expressed in terms of $\mathscr{J}$. Since C is symmetric, the only new identities will come from $M^{nn}{}_s$. Therefore, there will be nine such identities which can be labeled as T_c^{nn}. Initially, they can be regarded as a set of homogeneous, linear equations for the nine invariants

$$
\begin{array}{ccc}
[\overline{\Psi}] & [AC] & [A^2C] \\
[C] & [ABC] & [A^2BC] \\
[BC] & [AB^2C] & [A^2B^2C].
\end{array}
$$

These identities will be subject to the same relationships as the corresponding matrix identities. Thus, only three of them will be independent, and it will be sufficient to consider simply T_c^{44}, T_c^{45} and T_c^{54}. On the other hand, if the invariants $\mathscr{I}$ are known, then $[\overline{\Psi}]$, $[C]$ and $[BC]$ are also known and these identities are linear equations for the six invariants $[AB^mC]$ and $[A^2B^mC]$. The complex conjugates of these equations, denoted by $\overline{T_c^{nn}}$, will be linear equations for $[AB^mC]$ and $[AB^mD]$.

These procedures also follow if we multiply the matrix identities by D and take the trace. The identities T_d^{nn} give equations for $[AB^mD]$ and $[A^2B^mD]$. Their complex conjugates, $\overline{T_d^{nn}}$, are equations for $[A^2B^mC]$ and $[A^2B^mD]$. Altogether, there are twelve equations for twelve unknowns. However, it turns out that only nine of these are independent. Also, none of these equations involves $[A^2BACD]$. This means that at least four more equations are needed.

The next set of identities can be obtained by multiplying the matrix identities by CD and taking the trace. In this case, some work is needed to write some of the traces in terms of the set $\mathscr{I}$, and several of the calculations draw significantly on the work of Spencer and Rivlin. Since $(CD)^T=DC$ and $CD+DC\approx 0$, identities found by this method will effectively come from the skew-symmetric part of the matrix identities, namely $M^{43}{}_a$, $M^{34}{}_a$, $M^{44}{}_a$, etc. These trace identities can be labeled T_{cd}^{43}, T_{cd}^{34} and T_{cd}^{44} and so on. The invariant $[A^2BACD]$ appears linearly in each of these equations and the invariants $[A^nB^mC]$ and $[A^nB^mD]$ appear quadratically. Once again, there will be a chain of identities of identities for these equations, but no attempt has been made to find these.

The other technique that has been successful in obtaining new identities is equivalent to a polarization procedure. Of the several possibilities that were tried, the ''transformation'' $A \rightarrow A+\epsilon C$ seemed to provide the simplest identities. Other choices either did not give new identities, or gave identities of higher degree than have already been found. In carrying out this transformation, the matrices Ψ and $\overline{\Psi}$ were treated as independent, so that C and D were unchanged. The transformation was carried out on each invariant in a given identity and the coefficient of ϵ in the resulting equation then gave a new identity. Also, since A was being transformed into a matrix that was not trace-free, it was neccessary to start with each identity in a form that included $[A]$. $[A]$ could again be set to zero once the new identity had been obtained.

A number of identities have been generated in this fashion. These have been provisionally labeled according the the total degree of Ψ, $K=\Gamma\overline{\Gamma}$ and $\overline{\Psi}$ in the identity. For this purpose, the matrix $C=\Gamma\overline{\Psi}\,\overline{\Gamma}$ was regarded as being of first degree in both K and $\overline{\Psi}$. Thus T^{362} is the name given to the identity generated from T_c^{44} etc. Some care should be taken however, as it is not certain that this notation will identify an identity uniquely in general. Some of the identities obtained from the equations seen so far are

$$
\begin{array}{ll}
T_c^{44}\rightarrow T^{362}, & T_d^{44}\rightarrow T^{363}, \\
T_c^{45}\rightarrow T^{372}, & T_d^{45}\rightarrow T^{373}, \\
T_c^{54}\rightarrow T^{462}, & T_d^{54}\rightarrow T^{463}.
\end{array}
$$

Each of these identities is quadratic in $[A^nB^mC]$ and $[A^nB^mD]$, but none of them contains $[A^2BACD]$. The procedure applied to the equations $\overline{T_c^{nn}}$ gave the null identity, while applied to $\overline{T_d^{nn}}$ it gave identities whose complex conjugates are linear in C and so must be combinations of the identities T_c^{nn}.

Each of these identities could be checked quickly by choosing random numerical values for Ψ and Γ. These were substituted into the LHS of the identity which was then evaluated to check that the result was zero. (This procedure also helped to detect many errors at early stage.)

It remains now to confirm that there are sufficient equations to determine all invariants in $\mathscr{J}$ from a knowledge of $\mathscr{I}$. We start by looking more closely at the structure of some of the identities. Introduce the notation

$$\boldsymbol{u}=\begin{pmatrix}[AC]\\ [ABC]\\ [AB^2C]\end{pmatrix};\quad \boldsymbol{v}=\begin{pmatrix}[A^2C]\\ [A^2BC]\\ [A^2B^2C]\end{pmatrix};\quad \boldsymbol{w}=\begin{pmatrix}[A^2D]\\ [A^2BD]\\ [A^2B^2D]\end{pmatrix},$$

where $\boldsymbol{u}$ and $\boldsymbol{w}$ are real and $\boldsymbol{v}$ is complex. The sets of equations T_c^{nn} and $\overline{T_d^{nn}}$ can then be written as

$$\begin{aligned} M\boldsymbol{u}+N\boldsymbol{v}&=\boldsymbol{s},\\ M\bar{\boldsymbol{v}}+N\boldsymbol{w}&=\boldsymbol{t},\end{aligned} \tag{12}$$

where M and N are matrices whose coefficients are functions of the invariants $\mathscr{I}$, and $\boldsymbol{s}$ and $\boldsymbol{t}$ are vectors of invariants. The expressions for M and N are given in Appendix A. The equations T_{cd}^{nn} have the form

$$d[A^2BACD]+\boldsymbol{v}^TF\bar{\boldsymbol{v}}-\boldsymbol{u}^TF\boldsymbol{w}+\boldsymbol{f}^T\boldsymbol{u}+\boldsymbol{g}^T\boldsymbol{v}+\boldsymbol{h}^T\bar{\boldsymbol{v}}+\boldsymbol{k}^T\boldsymbol{w}=m, \tag{13}$$

where d and m are invariants, F is a matrix and $\boldsymbol{f}$, $\boldsymbol{g}$, $\boldsymbol{h}$ and $\boldsymbol{k}$ are vectors of invariants. Expressions for d and the matrix F for T_{cd}^{34} are given in Appendix A. The equations T^{362} and T^{372} each have the form

$$\boldsymbol{u}^TF\boldsymbol{u}+\boldsymbol{u}^TG\boldsymbol{v}+\boldsymbol{f}^T\boldsymbol{u}+\boldsymbol{g}^T\boldsymbol{v}+\boldsymbol{h}^T\bar{\boldsymbol{v}}=m. \tag{14}$$

The equations in (12) each have about 100 terms. Equations of the type T_{cd}^{nn} and T^{nnn} (at least those that have been calculated) have about 200–300 terms.

Equations (12) show how all the invariants of $\mathscr{J}$ can be determined from only a small number of them. For example, if the three real invariants in $\boldsymbol{u}$ are known, and $\det N\neq 0$, the first equation can be solved to give $\boldsymbol{v}$. Once this is known, the second equation can be solved for $\boldsymbol{w}$. The fact that this value for $\boldsymbol{w}$ must always be real is a further indication that Eqs. (12) must be dependent. Finally, T_{cd}^{34} can be solved to give $[A^2BACD]$.

On the other hand, if only the invariants $\mathscr{I}$ are to be used to find all those in Table I, it will be neccessary to find equations to calculate $\boldsymbol{u}$ first. In principle, this can be done as follows. Equation (12) can be used to write $\boldsymbol{v}=N^{-1}(\boldsymbol{s}-M\boldsymbol{u})$ and $\boldsymbol{w}=N^{-1}\boldsymbol{t}-N^{-1}M\overline{N}^{-1}\overline{\boldsymbol{s}}+N^{-1}M\overline{N}^{-1}\overline{M}\boldsymbol{u}$. These expressions can then be substituted into identities such as T^{362}, T^{372} and T^{462} which are quadratic in $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$. The final result will be a set of three quadratic equations for the three components of $\boldsymbol{w}$. Once these equations are solved, the other invariants can be found. Unfortunately, things are not quite that simple. Trials with specific numerical values for the set $\mathscr{I}$ suggest that the identities T^{362}, T^{372} and T^{462} are actually dependent. The identity T^{363} also does not provide any further information. To obtain an extra equation, $[A^2BACD]$ can be eliminated from T_{cd}^{34} and T_{cd}^{43}. By this means, independent quadratic equations for $[AC]$, $[ABC]$ and $[AB^2C]$ can be obtained. These equations can progressively be reduced to a single eighth degree equation for one of the variables, $[AC]$ say. The eight solutions of this equation can be found numerically.

This procedure was tested by giving random values to the matrices Ψ and Γ and calculating the invariants in $\mathscr{I}$. The above procedures were then followed to calculate the value of $[AC]$ and the other components of $\boldsymbol{u}$ from the values of $\mathscr{I}$ only. The final equation for $[AC]$ will have complex coefficients in general. In those examples tried, typically two or four of the eight solutions of this equation were real—or at least the imaginary part was several orders of magnitude less than the real part. Since $[AC]$ has to be real, these are the only solutions that will be suitable. In each different trial, the ''correct'' value for $[AC]$ was amongst these real solutions. However, in some cases, it wasn't clear that the imaginary part was small enough to deem some of the solutions to be real. It is likely that, in these cases, the set of equations was somewhat ill-conditioned. The situation often improved by increasing the precision of the calculations. Another approach was to use small integer values for the components of Ψ and Γ and then carry out the

calculations using integer arithmetic. Any approximation was delayed until the final step when the eighth degree polynomial was solved. In this fashion, the solutions were either clearly complex, or the order of magnitude of the imaginary part was 10^{-14} times that of the real part. An example of this is given in Appendix B. Despite these numerical difficulties, it is clear that, at least in principle, the invariants $\mathscr{I}$ can be used to determine the invariants in $\boldsymbol{u}$, and then those in Table I, and, consequently, all invariants of even degree in Γ.

One final observation can be made here. The procedures used to solve the equations that give the value of $\boldsymbol{u}$ indicate that the only non-linear equation to be solved is the eighth degree polynomial for $[AC]$. Once this has been solved, the equations needed to calculate $[ABC]$ and $[AB^2C]$ are linear. However, the expressions for these equations in terms of the invariants $\mathscr{I}$ (rather than particular numerical values) will be quite complex.

VII. CONCLUSION

There has been a considerable number of papers directed towards understanding the invariants of the Riemann tensor. It seems that, in some cases at least, these papers have not taken full account of relevant results from invariant theory. One of the purposes of this paper has been to bring some of these results to the attention of other workers in the field. It has been seen that they can be very useful in obtaining a complete set of invariants and the relationships between them.

These results, particularly those of Spencer and Rivlin, have been combined with the idea of expressing the Riemann tensor in terms of matrices in a complex three dimensional space. This has produced a set of invariants of the Riemann tensor which is believed to be complete (at least for invariants that are of even degree in the trace-free Ricci tensor and does not involve the Ricci scalar). At the same time, a large number of syzygies have been found which show how the invariants of this set are related.

The final results are not quite as simple as one might have hoped. It appears that a complete set for the invariants of the Riemann tensor may require 29 elements (if the Ricci scalar is included). This is certainly larger than expected. Also, the syzygies that connect these invariants typically have a few hundred terms. While such equations can be handled easily with modern computer algebra packages, it is still difficult to get any real understanding of them. On the other hand, it appears that the invariants $\mathscr{I}$, together with $\boldsymbol{u}$ and the Ricci scalar will be sufficient to determine all other invariants uniquely in the general case. There are a total of 19 such invariants which is close to the number that Carminati and McLenaghan[11] have estimated will be needed to calculate the other invariants.

ACKNOWLEDGMENTS

This work was carried out while the author was on sabbatical leave at the Queensland University of Technology during the first semester of 1997. I would like to thank Professor Sean McElwain and the School of Mathematics at QUT for their hospitality and the use of their facilities during that period. I am particularly grateful to Rod Gover of QUT for drawing my attention to the relevant work in invariant theory and for a number of illuminating discussions.

APPENDIX A: EXAMPLES OF IDENTITIES

This appendix contains some of the identities that have been found in this paper. This is primarily to give an indication of the nature of these identities, rather than to give a complete list. The full identities are available from the author in an electronic form suitable for input into most computer algebra packages. The identities are written in terms of the determinants of matrices, rather than the trace of the cube. Thus throughout, $[A^3]$ is replaced by the appropriate expression involving $\det A$.

The first equation is the expression for A^2B^2AB in terms of $\mathscr{M}$. It is assumed that $[A]=[B]=0$.

$$
\begin{aligned}
12A^2B^2AB=&\,2B^2AB[A^2]-8AB^2[A^2B]-4B^2A[A^2B]+8A^2B^2[AB]+4B^2A^2[AB]\\
&-4B^2[A^2][AB]+4A^2B[AB^2]-4BA^2[AB^2]\\
&-2A^2BA[B^2]-2A^2[AB][B^2]\\
&+B(2[A^2][AB^2]+2\ \det A[B^2])+AB(8[A^2B^2]-2[A^2][B^2])\\
&+BA(4[A^2B^2]-[A^2][B^2])+A(-2\ \det B[A^2]+4[A^2B][B^2])\\
&+I(4[A^2B^2AB]-8[A^2B^2][AB]+4[A^2B][AB^2]+3[A^2][AB][B^2]).
\end{aligned}
$$

The next group of equations gives four of the identities satisfied by symmetric matrix polynomials. Once again $[A]=[B]=0$. The identity $M^{44}{}_s$ is $S_1=0$ where

$$
\begin{aligned}
6S_1=&\,B^2(-2[A^2B]^2+2[A^2][A^2B^2]-[A^2][AB]^2+6\ \det A[AB^2])\\
&+\tfrac{1}{2}(AB^2+B^2A)(-4\,[A^2B][AB]+2\,[A^2][AB^2]+6\ \det A[B^2])\\
&+\tfrac{1}{2}(A^2B^2+B^2A^2)(-12\,[A^2B^2]+4\,[AB]^2+2\,[A^2][B^2])\\
&+\tfrac{1}{2}(A^2B+BA^2)(6\ \det B[A^2]-4\,[AB][AB^2]+2\,[A^2B][B^2])\\
&+B(-2\ \det B[A^2]^2+2\,[A^2B][A^2B^2]+[A^2][AB][AB^2]\\
&-[A^2][A^2B][B^2]-\det A[AB][B^2])+\tfrac{1}{2}(AB+BA)(-18\ \det A\ \det B\\
&+2\,[A^2B^2][AB]+2\,[A^2B][AB^2]-[A^2][AB][B^2])\\
&+A^2(6\ \det B[A^2B]-2\,[AB^2]^2+2\,[A^2B^2][B^2]-[AB]^2[B^2])\\
&+I(-4\ \det B[A^2][A^2B]+4\,[A^2B^2]^2+6\ \det A\ \det B[AB]\\
&-2\,[A^2B^2][AB]^2+2\,[A^2B][AB][AB^2]-2\,[A^2][A^2B^2][B^2]\\
&+[A^2][AB]^2[B^2]-4\ \det A[AB^2][B^2])+A(-(\det B[A^2][AB])\\
&+2\,[A^2B^2][AB^2]+[A^2B][AB][B^2]-[A^2][AB^2][B^2]-2\ \det A[B^2]^2).
\end{aligned}
$$

The identity $M^{45}{}_s$ is $S_2=0$ where

$$
\begin{aligned}
24\,S_2=&\,\tfrac{1}{2}(A^2B^2+B^2A^2)(-12\ \det B[A^2]+8\,[AB][AB^2]-4\,[A^2B][B^2])\\
&+B^2(4\ \det B[A^2]^2-4\,[A^2B][A^2B^2]-2\,[A^2][AB][AB^2]\\
&+2\,[A^2][A^2B][B^2]+2\ \det A[AB][B^2])+\tfrac{1}{2}(AB^2+B^2A)(36\ \det A\ \det B\\
&-4\,[A^2B^2][AB]-4\,[A^2B][AB^2]+2\,[A^2][AB][B^2])\\
&+B(-4\ \det B[A^2][A^2B]+4\,[A^2B^2]^2-12\ \det A\ \det B[AB]+2\,[A^2][AB^2]^2\\
&-2\,[A^2][A^2B^2][B^2]+2\ \det A[AB^2][B^2])\\
&+\tfrac{1}{2}(AB+BA)(-4\ \det B[A^2][AB]\\
&+8\,[A^2B^2][AB^2]-2\,[A^2][AB^2][B^2]-2\ \det A[B^2]^2)
\end{aligned}
$$

$$
\begin{aligned}
&+\tfrac{1}{2}(A^2B+BA^2)(24\ \det B[A^2B]-8\,[AB^2]^2-4\,[A^2B^2][B^2]+2\,[A^2][B^2]^2)\\
&+A^2(-12\ \det B[A^2B^2]+4\ \det B[AB]^2+8\ \det B[A^2][B^2]\\
&-4\,[AB][AB^2][B^2]\\
&+2\,[A^2B][B^2]^2)+I(-8\ \det B[A^2B]^2+8\ \det B[A^2][A^2B^2]\\
&-12\ \det A\ \det B[AB^2]-4\,[A^2B^2][AB][AB^2]+4\,[A^2B][AB^2]^2\\
&-4\ \det B[A^2]^2[B^2]+4\,[A^2B][A^2B^2][B^2]+2\,[A^2][AB][AB^2][B^2]\\
&-2\,[A^2][A^2B][B^2]^2)+B(-4\ \det B[A^2B][AB]+2\ \det B[A^2][AB^2]\\
&-12\ \det A\ \det B[B^2]+2\,[A^2B^2][AB][B^2]+2\,[A^2B][AB^2][B^2]\\
&-[A^2][AB][B^2]^2).
\end{aligned}
$$

The identity $M^{54}{}_s$ can be deduced from $M^{45}{}_s$ by interchanging A and B. The identity $M^{55}{}_s$ is $S_4=0$, where

$$
\begin{aligned}
12\,S_4=&\tfrac{1}{2}(AB^2+B^2A)(2\ \det B[A^2]^2-8\,[A^2B][A^2B^2]+2\,[A^2][A^2B][B^2]+4\ \det A[AB][B^2])\\
&+\tfrac{1}{2}(A^2B^2+B^2A^2)(-36\ \det A\ \det B+4\,[A^2B^2][AB]+4\,[A^2B][AB^2]\\
&-2\,[A^2][AB][B^2])\\
&+B^2(12\ \det A\ \det B[A^2]-2\,[A^2][A^2B^2][AB]-2\,[A^2][A^2B][AB^2]\\
&+4\ \det A[AB][AB^2]-2\ \det A[A^2B][B^2]+[A^2]^2[AB][B^2])\\
&+B(12\ \det A\ \det B[A^2B]-2\ \det B[A^2]^2[AB]+4\,[A^2][A^2B^2][AB^2]\\
&-4\ \det A[AB^2]^2-2\ \det A[A^2B^2][B^2]-[A^2]^2[AB^2][B^2])\\
&+\tfrac{1}{2}(A^2B+BA^2)(4\ \det B[A^2][AB]-8\,[A^2B^2][AB^2]+2\,[A^2][AB^2][B^2]\\
&+2\ \det A[B^2]^2)+\tfrac{1}{2}(AB+BA)(4\ \det B[A^2][A^2B]+8\,[A^2B^2]^2-24\ \det A\ \det B[AB]\\
&-6\,[A^2][A^2B^2][B^2]+4\ \det A[AB^2][B^2]+[A^2]^2[B^2]^2)+A(-4\ \det B[A^2B]^2\\
&-2\ \det B[A^2][A^2B^2]+12\ \det A\ \det B[AB^2]+4\,[A^2B][A^2B^2][B^2]\\
&-[A^2][A^2B][B^2]^2-2\ \det A[AB][B^2]^2)+A^2(4\ \det B[A^2B][AB]\\
&-2\ \det B[A^2][AB^2]+12\ \det A\ \det B[B^2]-2\,[A^2B^2][AB][B^2]\\
&-2\,[A^2B][AB^2][B^2]+[A^2][AB][B^2]^2)+I(12\ \det A\ \det B[A^2B^2]\\
&-4\ \det B[A^2][A^2B][AB]-4\,[A^2B^2]^2[AB]+8\ \det A\ \det B[AB]^2\\
&+4\,[A^2B][A^2B^2][AB^2]-8\ \det A\ \det B[A^2][B^2]+4\,[A^2][A^2B^2][AB][B^2]\\
&-4\ \det A[AB][AB^2][B^2]-[A^2]^2[AB][B^2]^2).
\end{aligned}
$$

The equations T_c^{nn} and T_d^{nn} are given by Eq. (12). The matrices M and N in those equations are as follows:

$$72M_{11} = -12\det B[A^2][AB] + 24[A^2B^2][AB^2] + 72\det A\ \det B[B] - 24[A^2B^2][AB][B]$$
$$-24[A^2B][AB^2][B] + 12[A^2B][AB][B]^2 + 12[A^2][AB^2][B]^2 - 6[A^2][AB][B]^3$$
$$-12\det A[B]^4 + 12[A^2B][AB][B^2] - 12[A^2][AB^2][B^2] + 6[A^2][AB][B][B^2]$$
$$+36\det A[B]^2[B^2] - 24\det A[B^2]^2,$$

$$72M_{12} = -216\det A\ \det B + 24[A^2B^2][AB] + 24[A^2B][AB^2] - 24[A^2][AB^2][B]$$
$$+12[A^2][AB][B]^2 + 36\det A[B]^3 - 12[A^2][AB][B^2] - 84\det A[B][B^2],$$

$$72M_{13} = -48[A^2B][AB] + 24[A^2][AB^2] - 24\det A[B]^2 + 72\det A[B^2],$$

$$72M_{21} = -12\det B[A^2B][AB] + 6\det B[A^2][AB^2] + 4\det B[A^2][AB][B]$$
$$-8[A^2B^2][AB^2][B] + 24\det A\ \det B[B]^2 + 2[A^2B^2][AB][B]^2$$
$$+2[A^2B][AB^2][B]^2 + 2[A^2B][AB][B]^3 - [A^2][AB^2][B]^3 - [A^2][AB][B]^4$$
$$-2\det A[B]^5 - 36\det A\ \det B[B^2] + 6[A^2B^2][AB][B^2]$$
$$+6[A^2B][AB^2][B^2]$$
$$-10[A^2B][AB][B][B^2] + [A^2][AB^2][B][B^2] + 4[A^2][AB][B]^2[B^2]$$
$$+6\det A[B]^3[B^2] - 3[A^2][AB][B^2]^2 - 4\det A[B][B^2]^2,$$

$$72M_{22} = -12\det B[A^2][AB] + 24[A^2B^2][AB^2] - 72\det A\ \det B[B] - 8[A^2B^2][AB][B]$$
$$-8[A^2B][AB^2][B] + 2[A^2][AB^2][B]^2 + 2[A^2][AB][B]^3 + 6\det A[B]^4$$
$$-6[A^2][AB^2][B^2] - 2[A^2][AB][B][B^2] - 8\det A[B]^2[B^2] - 6\det A[B^2]^2,$$

$$72M_{23} = 108\det A\ \det B - 12[A^2B^2][AB] - 12[A^2B][AB^2] + 16[A^2B][AB][B]$$
$$+4[A^2][AB^2][B] - 6[A^2][AB][B]^2 - 10\det A[B]^3 + 6[A^2][AB][B^2]$$
$$+18\det A[B][B^2],$$

$$72M_{31} = 6\det\ B[A^2][A^2B] + 12[A^2B^2]^2 - 36\det A\ \det B[AB] - 24[A^2B][A^2B^2][B]$$
$$+6[A^2B]^2[B]^2 + 6[A^2][A^2B^2][B]^2 + 12\det A[AB^2][B]^2 - 3[A^2][A^2B][B]^3$$
$$-6\det A[AB][B]^3 + 6[A^2B]^2[B^2] - 6[A^2][A^2B^2][B^2] - 12\det A[AB^2][B^2]$$
$$+3[A^2][A^2B][B][B^2] + 6\det A[AB][B][B^2],$$

$$72M_{32} = -6\det B[A^2]^2 + 24[A^2B][A^2B^2] - 48\det A[AB^2][B] - 6[A^2][A^2B][B]^2$$
$$+36\det A[AB][B]^2 + 3[A^2]^2[B]^3 - 6[A^2][A^2B][B^2] - 12\det A[AB][B^2]$$
$$-3[A^2]^2[B][B^2],$$

$$72M_{33} = -24[A^2B]^2 - 12[A^2][A^2B^2] + 72\det A[AB^2] + 24[A^2][A^2B][B]$$
$$-48\det A[AB][B] - 6[A^2]^2[B]^2 + 6[A^2]^2[B^2].$$

$$72N_{11}=72\det B[A^2B]-24\,[AB^2]^2-48\det B[A^2][B]+48\,[AB][AB^2][B]$$
$$-24\,[A^2B^2][B]^2-12\,[AB]^2[B]^2+12\,[A^2B][B]^3+24\,[A^2B^2][B^2]$$
$$-12\,[AB]^2[B^2]-12\,[A^2B][B][B^2],$$

$$72N_{12}=72\det B[A^2]-48\,[AB][AB^2]+96\,[A^2B^2][B]-72\,[A^2B][B]^2+12\,[A^2][B]^3$$
$$+24\,[A^2B][B^2]-12\,[A^2][B][B^2],$$

$$72N_{13}=-144\,[A^2B^2]+48\,[AB]^2+96\,[A^2B][B]-24\,[A^2][B]^2+24\,[A^2][B^2],$$

$$72N_{21}=-36\det B[A^2B^2]+12\det B[AB]^2+8\,[AB^2]^2[B]-8\det B[A^2][B]^2$$
$$-4\,[AB][AB^2][B]^2+2\,[A^2B^2][B]^3-2\,[AB]^2[B]^3+2\,[A^2B][B]^4$$
$$+24\det B[A^2][B^2]-12\,[AB][AB^2][B^2]$$
$$-2\,[A^2B^2][B][B^2]+10\,[AB]^2[B][B^2]$$
$$-8\,[A^2B][B]^2[B^2]+6\,[A^2B][B^2]^2,$$

$$72N_{22}=72\det B[A^2B]-24\,[AB^2]^2+16\,[AB][AB^2][B]+4\,[A^2B^2][B]^2-12\,[A^2B][B]^3$$
$$+2\,[A^2][B]^4-12\,[A^2B^2][B^2]+28\,[A^2B][B][B^2]-8\,[A^2][B]^2[B^2]+6\,[A^2][B^2]^2,$$

$$72N_{23}=-36\det B[A^2]+24\,[AB][AB^2]-16\,[AB]^2[B]+4\,[A^2B][B]^2+2\,[A^2][B]^3$$
$$-12\,[A^2B][B^2]-2\,[A^2][B][B^2],$$

$$72N_{31}=6\det B[A^2][AB]-12\,[A^2B^2][AB^2]-36\det A\ \det B[B]+12\,[A^2B^2][AB][B]$$
$$+12\,[A^2B][AB^2][B]-6\,[A^2B][AB][B]^2-6\,[A^2][AB^2][B]^2+3\,[A^2][AB][B]^3$$
$$+6\det A[B]^4-6\,[A^2B][AB][B^2]+6\,[A^2][AB^2][B^2]-3\,[A^2][AB][B][B^2]$$
$$-18\det A[B]^2[B^2]+12\det A[B^2]^2,$$

$$72N_{32}=108\det A\ \det B-12\,[A^2B^2][AB]-12\,[A^2B][AB^2]+12\,[A^2][AB^2][B]$$
$$-6\,[A^2][AB][B]^2-18\det A[B]^3+6\,[A^2][AB][B^2]+42\det A[B][B^2],$$

$$72N_{33}=24\,[A^2B][AB]-12\,[A^2][AB^2]+12\det A[B]^2-36\det A[B^2].$$

The equation for T^{34}_{cd} is given by (13). The expressions for d and F in that equation are as follows:

$$54\,d=-54\det B+5\,[B]^3-9\,[B][B^2],$$

$$54F=\begin{pmatrix}[B]^4-5\,[B]^2[B^2]+6\,[B^2]^2, & -[B]^3+3\,[B][B^2], & 3\,[B]^2-9\,[B^2]\\ -[B]^3+3\,[B][B^2], & -3\,[B]^2+9\,[B^2], & 0\\ 3\,[B]^2-9\,[B^2], & 0, & 0\end{pmatrix}.$$

APPENDIX B: AN EXAMPLE FOR CALCULATING $\mathscr{J}$ FROM $\mathscr{I}$

The following is a simple example of the calculation of the invariants $\boldsymbol{u}$ from a knowledge of those in $\mathscr{I}$. The following values were chosen for Ψ and Γ,

$$\Psi=\begin{pmatrix} -2i, & 2, & i \\ 2, & 1-2i, & -i \\ i, & -i, & -1+4i \end{pmatrix};\quad \Gamma=\begin{pmatrix} ; & -2, & i, & 1+2i \\ ; & -i, & 0, & -2-2i \\ & 1-2i, & -2+2i, & -1 \end{pmatrix}.$$

Calculation of the invariants $\mathscr{I}$ gives the following:

$$[\Psi^2]=-18-12i \qquad \det\Psi=21-34i$$
$$[K]=-3 \qquad [\Psi K]=83-12i \qquad [\Psi^2K]=-66+198i$$
$$[K^2]=-195 \quad [\Psi K^2]=-114-254i \quad [\Psi^2K^2]=2979-374i$$
$$\det K=169.$$

The task now is to find the value of $[AC]$ using the values of these invariants, and not the values of Ψ and Γ. The solutions of Eq. (12) can be approximated to give

$$[A^2C]=889.895-544.891i+(0.0612686-0.00125164i)[AB^2C]$$
$$-(0.423424-0.132449i)[ABC]+(2.95822+0.328379i)[AC],$$
$$[A^2BC]=-5473.47+3719.81i+(0.211712-0.0662243i)[AB^2C]$$
$$+(1.37951-0.0107651i)[ABC]+(28.8703-6.90351i)[AC],$$
$$[A^2B^2C]=39493.2+21031.8i-(6.14016-0.514248i)[AB^2C]$$
$$+(5.13792-5.86159i)[ABC]-(384.532-22.5566i)[AC],$$
$$[A^2D]=889.895+544.891i+(0.0612686+0.00125164i)[AB^2C]$$
$$-(0.423424+0.132449i)[ABC]+(2.95822-0.328379i)[AC],$$
$$[A^2BD]=-5473.47-3719.81i+(0.211712+0.0662243i)[AB^2C]$$
$$+(1.37951+0.0107651i)[ABC]+(28.8703+6.90351i)[AC],$$
$$[A^2B^2D]=39493.2-21031.8i-(6.14016+0.514248i)[AB^2C]$$
$$+(5.13792+5.86159i)[ABC]-(384.532+22.5566i)[AC].$$

These can be substituted into T^{34}_{cd}, T^{43}_{cd}, T^{362} and T^{372} to give the following equations for $[AC]$, $[ABC]$ and $[AB^2C]$:

$$0=1.23009\ 10^9-2.99669\ 10^9i+(27126.9-57662.\ i)[AB^2C]$$
$$+(2.40324-1.10915\ i)[AB^2C]^2-(740633.-1.654\ 10^6\ \ i)[ABC]$$
$$+(6.79429-5.6801\ i)[AB^2C][ABC]+(131.37-262.435\ i)[ABC]^2$$
$$+(682518.-1.17136\ 10^6\ \ i)[AC]+(375.736-569.387\ i)[AB^2C][AC]$$
$$+(841.824-1576.45\ i)[ABC][AC]+(12572.-26212.1\ i)[AC]^2,$$
$$0=-1.835\ 10^9-2.75287\ 10^8\ \ i+(132810.+267227.\ i)[AB^2C]$$
$$-(1.40933+0.621754\ i)[AB^2C]^2+(1.79147\ 10^6+382792.\ i)[ABC]$$
$$-(60.9006+75.3397\ i)[AB^2C][ABC]-(439.4-0.988063\ i)[ABC]^2$$

$$+(1.88299\,10^{7}-2.54068\,10^{6}\ i)[AC]-(256.321+234.499\,i)[AB^{2}C][AC]$$
$$-(7861.13-2139.7\,i)[ABC][AC]-(12470.5+20978.4\,i)[AC]^{2},$$
$$0=2.63922\,10^{11}-2.21063\,10^{11}i-(9.04512\,10^{6}-7.57629\,10^{6}\ i)[AB^{2}C]$$
$$+(73.4438-61.5173\,i)[AB^{2}C]^{2}-(2.00194\,10^{8}-1.67685\,10^{8}\ i)[ABC]$$
$$+(3972.51-3327.41\,i)[AB^{2}C][ABC]+(38266.4-32052.3\,i)[ABC]^{2}$$
$$-(6.72171\,10^{8}-5.63018\,10^{8}\ i)[AC]+(30039.6-25161.5\,i)[AB^{2}C][AC]$$
$$+(316964.-265492.\,i)[ABC][AC]+(1.88052\,10^{6}-1.57514\,10^{6}\ i)[AC]^{2}.$$

These equations can be reduced to the following eighth degree polynomial for $[AC]$:

$$0=6.57264-9.68351\,i+(0.286771-0.270973\,i)[AC]$$
$$+(0.00323485-0.00165287\,i)[AC]^{2}-(4.05461\,10^{-6}$$
$$-0.0000112074\,i)[AC]^{3}$$
$$-(2.04805\,10^{-7}-1.05828\,10^{-7}i)[AC]^{4}-(1.50569\,10^{-10}$$
$$+2.43893\,10^{-10}i)[AC]^{5}$$
$$+(6.7192\,10^{-12}-3.54656\,10^{-12}i)[AC]^{6}+(1.85497\,10^{-14}$$
$$-7.37274\,10^{-15}i)[AC]^{7}$$
$$-(1.04906\,10^{-17}+1.30311\,10^{-18}i)[AC]^{8}.$$

The solutions of this polynomial for $[AC]$ are

$$\begin{array}{cc}-187.475-4.76682\,i & -154.602\\ -114.192+40.8714\,i & -89.+2.85274\,10^{-13}\ i\\ -62.7669+21.187\,i & 142.82+62.1616\,i\\ 142.82-62.1616\,i & 1977.78-965.713\,i.\end{array}$$

The only real values are -154.602 and $-89.+2.85274\,10^{-13}\ i$ (which is deemed to be real). The corresponding values for $[ABC]$ and $[AB^{2}C]$ can be calculated from the earlier equations. The value of -89 is the ''correct'' value, and will lead to the correct values for the other invariants.

[1] J. Carminati and R. G. McLenaghan, ''Algebraic invariants of the Riemann tensor in a four-dimensional Lorentzian space,'' J. Math. Phys. **32**, 3135–3140 (1991).
[2] A. Harvey, ''Identities of the scalars of the four-dimensional Riemannian manifold,'' J. Math. Phys. **36**, 356–361 (1995).
[3] C. B. G. McIntosh and E. Zakhary, ''Mixed invariants of the Riemann tensor,'' in *Relativity Today, Proceedings of the Fourth Hungarian Relativity Workshop*, edited by R. P. Kerr and Z. Perjés (Akadémiai Kiadó, Budapest, 1994), pp. 3–12.
[4] B. M. Haddow, ''Geometric interpretation of the mixed invariants of the Riemann spinor,'' Gen. Relativ. Gravit. **28**, 481–509 (1996).
[5] P. J. Greenberg, ''The algebra of the Riemann curvature tensor in general relativity: Preliminaries,'' Stud. Appl. Math. **51**, 277–308 (1972).
[6] G. Sobczyk, ''Space-time algebra approach to curvature,'' J. Math. Phys. **22**, 333–342 (1981).
[7] G. E. Sneddon, ''On the algebraic invariants of the four-dimensional Riemann tensor,'' Class. Quantum Grav. **3**, 1031–1032 (1986).
[8] G. E. Sneddon, ''The identities of the algebraic invariants of the four-dimensional Riemann tensor,'' J. Math. Phys. **37**, 1059–1075 (1996).
[9] H. A. Buchdahl, ''On rotor calculus I,'' J. Aust. Math. Soc. B, Appl. Math. **6**, 402–423 (1966); ''On rotor calculus II,'' *ibid.* **6**, 424–448 (1966).
[10] A. J. M. Spencer and R. S. Rivlin, ''The theory of matrix polynomials and its application to the mechanics of isotropic continua,'' Arch. Ration. Mech. Anal. **2**, 309–336 (1958).

[11] J. Carminati and R. G. McLenaghan (private communication).
[12] H. Weyl, *The Classical Groups* (Princeton University Press, Princeton, 1939).
[13] G. B. Gurevich, *Foundations of the Theory of Algebraic Invariants* (Noordhoff, Groningen, 1964).
[14] D. Hilbert, *Theory of Algebraic Invariants* (Cambridge University Press, Cambridge, 1993).
[15] C. Procesi, ''The invariant theory of $n\times n$ matrices,'' Adv. Math. **19**, 306–381 (1976).
[16] A. R. Gover, ''The algebra of orthogonal group invariants,'' Queensland University of Technology preprint, 1997.
[17] A. J. M. Spencer and R. S. Rivlin, ''Finite integrity bases for five or fewer symmetric 3×3 matrices,'' Arch. Ration. Mech. Anal. **2**, 435–446 (1959).
[18] A. J. M. Spencer and R. S. Rivlin, ''Further results in the theory of matrix polynomials,'' Arch. Ration. Mech. Anal. **4**, 214–230 (1960).
[19] D. Lovelock, ''Dimensionally dependent identities,'' Proc. Cambridge Philos. Soc. **68**, 379–389 (1970).
[20] S. B. Edgar and A. Höglund, ''The Lanczos potential for the Weyl curvature tensor: existence, wave equation and algorithms,'' Proc. R. Soc. London, Ser. A **453**, 835–851 (1997).

On a class of distributions of interest in quantum field theory

Launey J. Thomas, III and Eyvind H. Wichmann
Department of Physics, University of California, Berkeley, California 94720

(Received 13 February 1997; accepted for publication 11 November 1997)

A class of tempered distributions studied by Araki is considered. These distributions are of interest in quantum field theory since they include certain matrix elements of commutators or anticommutators of local quantum field operators. A number of results concerning the null regions in spacetime of such distributions are presented, for the cases when the null regions are causal complements of closed, convex, causally complete subsets of spacetime. Applications to quantum field theory are discussed, and some immediate applications to the theory of the Klein–Gordon equation and the wave equation are noted. © *1998 American Institute of Physics.* [S0022-2488(98)03103-X]

I. INTRODUCTION

In this paper we will discuss a class of tempered distributions (on test functions on Minkowski spacetime) which have been discussed before by Araki.[1] They are of great interest in quantum field theory, and also in the theory of the wave equation and the Klein–Gordon equation. We will discuss applications to quantum field theory in this paper (in Section V), within the framework formulated by Wightman and others.[2–5] In a subsequent paper we will discuss applications to the theory of *local nets* of von Neumann algebras.[5,6] Many of the results in the present paper apply with minor modifications to the theory of local nets.

This work incorporates a number of results from the doctoral thesis of one of us (L. J. T.).[7]

For the formulation and study of the conditions of causality and local independence in quantum field theory the *causal structure* of Minkowski spacetime is of great importance. In an earlier paper,[8] *hereafter referred to as* TW-I, we have discussed the geometry of Minkowski spacetime and its causal structure in considerable detail. For the reader's convenience, and in order to establish notation, we briefly review some basic notions. Minkowski spacetime M is parametrized by the usual Cartesian coordinates $x=(x_1,x_2,x_3,x_4)=(\mathbf{x},t)$, with $t=x_4$ and $\mathbf{x}=(x_1,x_2,x_3)$, with reference to some particular inertial frame. The Minkowski scalar product is denoted $x\cdot y=x_4y_4-\mathbf{x}\cdot\mathbf{y}$, where $\mathbf{x}\cdot\mathbf{y}$ is the usual scalar product on the Euclidean subspace R^3. If R is any subset of M, we denote the closure of R by $\overline{R}$, the interior of R by R^o, the complement of R by $\widetilde{R}$, and the boundary of R by ∂R. The *causal complement* R^c of R is defined as the set of all points of M which are spacelike relative to all points of R. The set $R^{cc}=(R^c)^c$ is called the *causal completion of* R, and we say that a subset R of M is *causally complete* if and only if $R^{cc}=R$. For a detailed discussion of these notions we refer to TW-I.[8] (See, in particular, Section II of the quoted paper.) We here recall some notions which are basic for the present paper.

Definition 1.1: (a) For any $x\in\mathsf{M}$ we denote by $V_+(x)$ the *open forward lightcone* with x as apex, and we denote by $V_-(x)$ the *open backward lightcone* with x as apex.

(b) We denote by $\mathscr{W}$ the set of all open wedge-regions W which are Poincaré-equivalent with the particular wedge-region $W_R=\{x|x_3>|x_4|\}$, i.e., an element $W\in\mathscr{W}$ is any open subset of M bounded below by a characteristic plane π' and bounded above by another characteristic plane π'' not parallel with π'. We write $W(\pi',\pi'')$ for such a wedge in $\mathscr{W}$. We employ the term *causal wedge* for any $W\in\mathscr{W}$ or for its closure $\overline{W}$.

(c) For any two points $x,y\in\mathsf{M}$ such that $y\in V_+(x)$ we define the *closed double cone* $D(x,y)$ with x and y as apices by $D(x,y)=\overline{V}_+(x)\cap\overline{V}_-(y)$, and we denote by $\mathscr{D}$ the set of all such double cones. Note that with this definition every double cone $D\in\mathscr{D}$ has a nonempty interior. The intersection $\partial V_-(y)\cap\partial V_+(x)$ is contained in a unique spacelike plane σ, and it can be regarded

 © 1998 American Institute of Physics

as a sphere. The set $B=\sigma\cap D(x,y)$ is called the *base* of the closed double cone $D(x,y)$, and the set $B_o=\sigma\cap D(x,y)^o$ is the base of the open double cone $D(x,y)^o$.

(d) We define $\mathscr{K}_E$ as the set of all *closed, convex, causally complete* subsets K of M which satisfy the condition that there exist four characteristic planes π_1,π_2,π_3 and π_4 in a *general relative position*, i.e., the normals of these planes are linearly independent, such that K lies *below* π_1 and π_2 and *above* π_3 and π_4.

(e) We define the subset $\mathscr{K}_C$ of $\mathscr{K}_E$ as the set of all *compact* sets $K\in\mathscr{K}_E$, i.e., $\mathscr{K}_C$ is the set of all *compact, convex, causally complete* subsets of M. We regard the empty set $\varnothing$ as an element of $\mathscr{K}_C\subset\mathscr{K}_E$. We have $\mathscr{D}\subset\mathscr{K}_C$.

We say that two wedges W_1 and W_2 are in a *general relative position* if the four boundary planes are in a general relative position. We easily see that the set $\mathscr{K}_E$ can also be described as follows. A closed, convex, and causally complete set K is in $\mathscr{K}_E$ if and only if there exist two wedges $W_1,W_2\in\mathscr{W}$ in a general relative position such that $K\subset W_1\cap W_2$. We note that the sets $\mathscr{K}_E$ and $\mathscr{K}_C$ are closed under intersections (in contrast to what the situation is for the set $\mathscr{D}$ of double cones). Furthermore, if $K\subset\mathsf{M}$ is closed, convex, and causally complete, then $K\cap K_e\in\mathscr{K}_E$ if $K_e\in\mathscr{K}_E$, and $K\cap K_c\in\mathscr{K}_C$ if $K_c\in\mathscr{K}_C$. We also note that the sets $\mathscr{K}_E$ and $\mathscr{K}_C$ include subsets K of M with empty interiors. For a detailed discussion of the sets in $\mathscr{K}_E$ we refer to Section III in TW-I.[8]

In Section II we discuss general features of a class of tempered distributions studied by Araki.[1] We denote this class, specified in Definition 2.1, by $\mathscr{S}'_A(\mathsf{M})$. These distributions appear in quantum field theory in the following context. Let X be a local field operator and let $X(x)=T(x)XT(x)^{-1}$ be the translated operator obtained from X by a conjugation with the unitary operator $T(x)$ representing a translation by x. Let Y be some other operator, and let Φ_1 and Φ_2 be two vectors in the Hilbert space $\mathscr{H}$ of physical states with *compact* support in momentum space. The matrix element $\langle\Phi_1|[Y,X(x)]\Phi_2\rangle=\xi(x)$ is then a function on M which can be regarded as a distribution in the class $\mathscr{S}'_A(\mathsf{M})$. We have here ignored questions of domains for the possibly unbounded operators: in Section V we give a more cautious formulation. The support of the Fourier transform $\hat{\xi}(p)$ of $\xi(x)$ is then contained in the union of some forward lightcone and some backward lightcone in momentum space, and if the operator Y is in some sense ''local,'' then the *null region* $N(\xi)$ of the matrix element $\xi(x)$ is, in general, nonempty. Matrix elements of this kind have been much studied in the past, and we note here the papers by Jost and Lehmann[9] and by Dyson.[10] In our paper we follow the approach of Araki,[1] since our geometrical considerations are similar in spirit to the considerations in his paper. We are concerned with the following kind of problem. Given that the null region of $\xi(x)$ contains a subset S of M, where else does $\xi(x)$ vanish? This problem was also central in Araki's paper, in which a partial solution was given. His main result is here quoted in Lemma 2.7 in Section II. The ultimate aim is to determine the null region of the *commutator* $[Y,X(x)]$, and to get a handle on this problem one considers matrix elements of the above nature. If we consider *vacuum expectation values* of the commutator $[Y,X(x)]$, we are led to a class of *special* Araki distributions $\xi(x)$ for which the support of the Fourier transform $\hat{\xi}(p)$ is contained in the lightcone.

In the case of two fermion operators one considers anticommutators instead of commutators, but the corresponding modifications in the theory are of a trivial nature.

A solution $\psi(x)$ of the Klein–Gordon equation (or the wave equation) can be regarded as an element of $\mathscr{S}'_A(\mathsf{M})$ provided that it satisfies certain growth conditions in the asymptotic region. Every result about the nature of the support of a distribution in $\mathscr{S}'_A(\mathsf{M})$ thus applies to ''asymptotically well-behaved'' solutions of the Klein–Gordon equation. Concerning the null region of an ordinary solution of the Klein–Gordon equation there are certain ''classical'' principles which have been known for a long time, namely the following: I) If $\psi(x)$ vanishes on a neighborhood of a timelike (straight) line segment, then $\psi(x)$ also vanishes on the double cone with the endpoints of the segment as apices. II) If $\psi(x)$ and its gradient vanishes on the base of an open double cone, then $\psi(x)$ also vanishes on the open double cone.[11] In view of these principles it is natural to define an ''envelope'' $\mathscr{E}(S)$ of any open $S\subset\mathsf{M}$ as the smallest subset of M which contains S, and which contains an open double cone $D^o=D(x',x'')^o$ if it contains either the base of D^o or the line segment (x',x''). If $\psi(x)$ is a solution of the Klein–Gordon equation, we thus have the result that $N(\psi)\supset\mathscr{E}(S)$ whenever $N(\psi)\supset S$. The first one of the above ''classical'' principles also applies to

the distributions in $\mathscr{S}'_A(\mathsf{M})$, but the second principle does not apply in general, i.e., it is not true that if S is an arbitrary open subset of M on which a distribution $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$ vanishes, then also $N(\xi) \supset \mathscr{E}(S)$. Araki's theory leads to a set-mapping $S \to a(S)$ of open sets to open sets such that $N(\xi) \supset a(S)$ whenever $N(\xi) \supset S$, and in general the set $a(S)$ is *smaller* than $\mathscr{E}(S)$. For any open subset S of M we thus have three different ''envelopes:'' the *Araki envelope* $a(S)$, the *envelope of uniqueness* $\mathscr{E}(S)$, and the causal completion S^{cc} of S. In Section II we show that we always have $S \subset a(S) \subset \mathscr{E}(S) \subset (S^{cc})^o$. Hence we have $S = a(S) = \mathscr{E}(S)$ if S is *causally complete.*

Much of the discussion in this paper depends on the following general result stated in Theorem 2.11. If $S = \cup\{S_\alpha | \alpha \in \mathscr{T}\}$ is the union of any family of open sets S_α, such that for each α in the index set $\mathscr{T}$ the set S_α consists of all points which lie *below* a characteristic plane π''_α and *above* a characteristic plane π'_α, then $a(S) = \mathscr{E}(S)$. In particular this result applies if S is the union of any family of wedges in $\mathscr{W}$. Section III is devoted to a discussion of this situation. The issue is now under what *additional* premises can we conclude that $\mathscr{E}(S) = (S^{cc})^o$, and hence that $a(S) = (S^{cc})^o$. For a variety of sets $S \subset \mathsf{M}$ it is the case that $\mathscr{E}(S) = (S^{cc})^o$, and these sets are thus of particular interest, but this relation does not hold generally, as we show by examples. Theorem 3.6 and Propositions 3.8 and 3.9 are our main results in Section III. The most important part of Theorem 3.6 says that if $\mathscr{F}$ is any family of wedges in $\mathscr{W}$ which contains at least two wedges in a general relative position, then $a(S) = \mathscr{E}(S) = S^{cc}$, with $S = \cup\{W | W \in \mathscr{F}\}$. Furthermore we have $S^c \in \mathscr{K}_E$. The same conclusions hold in certain other cases, as stated in Theorem 3.6 and Proposition 3.8. These results are optimal, in the sense described in Proposition 4.4 in Section IV. In Proposition 3.9 we state and prove an ''intersection property.'' In particular, if $K_1, K_2 \in \mathscr{K}_E$, then $a(K_1^c \cup K_2^c) = (K_1 \cap K_2)^c$. Note that we have $a(K_1^c) = K_1^c$ and $a(K_2^c) = K_2^c$ since K_1^c and K_2^c are causally complete. The relation $a(K_1^c \cup K_2^c) = (K_1 \cap K_2)^c$ does *not* hold for two arbitrary closed causally complete subsets K_1 and K_2 of M, as we show by Example 4.10 in Section IV.

In Section IV we consider solutions of the Klein–Gordon equation (for an arbitrary mass $m \geqslant 0$), including the special case of the wave equation (with $m = 0$). A solution $\psi(x)$ which does not grow too rapidly as x tends to infinity can be regarded as a distribution in $\mathscr{S}'_A(\mathsf{M})$, and all of our results in Section III are directly applicable to the study of the support of $\psi(x)$. We present a few rather trivial conclusions of this category, but our purpose is not to enrich the theory of the Klein–Gordon equation, but rather to illustrate the general results in Sections II and III. The solutions of this equation provide concrete and tangible examples, as well as counterexamples against unwarranted ''extensions.''

In Proposition 4.4 and Example 4.5 (which concern solutions of the Klein–Gordon equation) we show that the largest *common* null region $\nu(S)$ of *all* distributions $\xi \in \mathscr{S}'_A(\mathsf{M})$ which vanish on an open set S can never be larger than $(S^{cc})^o$. It can, of course, be equal to $(S^{cc})^o$ since $a(S) = \mathscr{E}(S) = S$ if S is causally complete. Example 4.7 shows that for an arbitrary open set S the set $a(S)$ is in general smaller than $\mathscr{E}(S)$.

The solutions $\psi(x)$ of the Klein–Gordon equation in $\mathscr{S}'_A(\mathsf{M})$ have the property, not shared by Araki distributions in general, that $\psi(x) = 0$ on $\mathscr{E}(S)$ if $\psi(x) = 0$ on S. Moreover, a solution $\psi(x)$ is determined everywhere by the initial data on a suitable spacelike surface, and the initial data in two disjoint closed subsets of the surface are independent. Examples 4.18 and 4.19 illustrate some implications of these features for a solution of the Klein–Gordon equation for $m > 0$. We have $\psi(x) = 0$ on certain open sets $S = \mathscr{E}(S)$, and we show that in these cases we have $\psi(x) = 0$ on S^{cc} although $\mathscr{E}(S)$ is strictly smaller than S^{cc}. In Example 4.19, S is the union of two disjoint sets S_1 and S_2 such that $\mathscr{E}(S_1) = S_1$ and $\mathscr{E}(S_2) = S_2$. Both S_1 and S_2 are null regions for some solutions, but their union is not.

The class of special Araki distributions $\xi(x)$ mentioned before, for which the support of $\hat{\xi}(p)$ is contained in the lightcone, are somewhat more akin to solutions of the Klein–Gordon equation than the general Araki distributions. We present some results for these special distributions which do not hold for arbitrary distributions in $\mathscr{S}'_A(\mathsf{M})$. In Lemma 4.12 we consider an Araki distribution $\xi(x)$ such that the support of $\hat{\xi}(p)$ is contained in the *open* lightcone $V(0)$, and we prove that if $\xi(x)$ vanishes on an open ''tube-region'' with an isotropic ''axis,'' then $\xi(x)$ vanishes *identically*. In Lemma 4.14 we show that if $\xi(x)$ vanishes on the region obtained when a bounded portion is removed from the above tube-region, then $\xi(x)$ vanishes identically, *provided* that $\xi(x)$ now satisfies the slightly stronger condition that $\mathrm{supp}(\hat{\xi}(p)) \subset \{p | p \cdot p \geqslant m^2\}$ for some $m > 0$.

In Section V we discuss various applications of the results in Sections II–IV to quantum field theory. Specifically we consider the *weak commutant* $\mathscr{P}_0(R)^w$ of the algebra $\mathscr{P}_0(R)$ of all quantum field operators associated with various regions $R\subset\mathsf{M}$. The set $\mathscr{P}_0(R)^w$ is a weakly closed linear manifold of *bounded* operators B which satisfy the weak commutation relations $X^{**}B\supset BX$ and $X^{**}B^*\supset B^*X$ for all $X\in\mathscr{P}_0(R)$. It is not known at present whether the weak commutants are (nontrivial) von Neumann algebras without additional conditions on the quantum fields, but if they are, they define a local net of von Neumann algebras with which the field operators are affiliated. Results about weak commutants are thus relevant for the theory of local nets, which is one reason why we discuss these objects here.

The necessary and sufficient condition for a bounded operator B to be in $\mathscr{P}_0(R)^w$ is that certain expressions of the form $\langle B^*\Phi'|\varphi_\mu(x)\Phi''\rangle-\langle\varphi_\mu^\dagger(x)\Phi'|B\Phi''\rangle$ must vanish on R^o. Here Φ' and Φ'' are arbitrary vectors of *compact* support in momentum space, and the expressions are then Araki distributions. The results in Sections II and III are thus applicable when we try to determine the largest region R_e such that $B\in\mathscr{P}_0(R_e)^w$ whenever $B\in\mathscr{P}_0(R)^w$, for a given $R\subset\mathsf{M}$. In particular we have $\mathscr{P}_0(S)^w=\mathscr{P}_0(a(S))^w$ for any open S (Lemma 5.6), and hence $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ whenever $\mathscr{E}(S)=a(S)$, and $\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w$ whenever $(S^{cc})^o=a(S)$.

If $K\subset\mathsf{M}$ is a closed causally complete set, it is rather natural to associate the weak commutant $\mathscr{P}_0(K^c)^w$ with K, and hence we write $\mathscr{C}(K)=\mathscr{P}_0(K^c)^w$. In Propositions 5.8 and 5.9 we state some obvious properties of the objects $\mathscr{C}(K)$. We are particularly interested in the case when K is also convex. Theorems 5.10 and 5.11 are main results in Section V. The most important part of Theorem 5.10 states that if $\mathscr{F}$ is any family of wedges in $\mathscr{W}$ which contains at least two wedges in a general relative position, then $\mathscr{C}(K)=\cap\{\mathscr{C}(W^c)|W\in\mathscr{F}\}$, with $K=\cap\{W^c|W\in\mathscr{F}\}$. Or, equivalently stated, $\mathscr{P}_0(S^{cc})^w=\mathscr{P}_0(S)^w$, with $S=\cup\{W|W\in\mathscr{F}\}=K^c$, since we trivially have $\mathscr{P}_0(S)^w=\cap\{\mathscr{P}_0(W)^w|W\in\mathscr{F}\}$. Theorem 5.11 also states an ''intersection property.'' In particular, if $K_1,K_2\in\mathscr{K}_E$, then $K_1\cap K_2\in\mathscr{K}_E$ and $\mathscr{C}(K_1\cap K_2)=\mathscr{C}(K_1)\cap\mathscr{C}(K_2)$. This relation most definitely does not hold for two *arbitrary* closed causally complete subsets of M. We show this by counterexamples within the theory of a single free Hermitian scalar field (see Example 5.13 and Proposition 5.18). The theory of a free scalar field is closely related to the theory of the Klein–Gordon equation, and the examples and counterexamples in Section V are based on the examples and counterexamples in Section IV, in the manner explained in Lemma 5.12.

We also consider vacuum expectation values of commutators, corresponding to the ''special'' Araki distributions discussed in Section IV. Theorem 5.16, Lemma 5.17 and Proposition 5.18 are proved on the basis of properties of these special distributions. These results *cannot* be obtained on the basis of the results in Sections II and III alone: the proofs depend on additional features of a quantum field theory. Proposition 5.18 concerns the weak commutants of the field operators associated with two wedges W_1 and W_2 in $\mathscr{W}$, and it gives a complete description of the cases which can occur.

We conclude Section V by showing, in Example 5.20, that the weak commutant $\mathscr{C}(K)$ for a compact, convex, causally complete set $K\subset\mathsf{M}$ with an *empty* interior can be nontrivial. The relevant example is within the theory of a free scalar field.

In Section VI we make some general remarks about our results, and about the limitations of our methods, which are based on very general features of quantum field theories. Whether such a general theory always satisfies a physically reasonable condition of causality is not known. One might suspect that what is missing is some kind of ''dynamical principle'' which would imply a strong condition of causality.

II. ON A CLASS OF TEMPERED DISTRIBUTIONS DISCUSSED BY ARAKI

We denote by $\mathscr{S}'(\mathsf{M})$ the set of all tempered distributions defined on the set $\mathscr{S}(\mathsf{M})$ of Schwartz functions $f(x)$ on Minkowski spacetime. For any $\xi(x)\in\mathscr{S}'(\mathsf{M})$ we define the Fourier transform $\hat{\xi}(p)$ (which is also a tempered distribution) by

$$\hat{\xi}(p)=\frac{1}{(2\pi)^2}\int_{(\infty)}d^4(x)\,\xi(x)\,\exp(-ip\cdot x). \tag{2.1}$$

Here $p \cdot x$ stands for the Lorentz scalar product of p and x. For the inverse Fourier transform we thus have

$$\xi(x) = \frac{1}{(2\pi)^2} \int_{(\infty)} d^4(p) \hat{\xi}(p) \exp(ix \cdot p).$$

The above formulas, in which we employ the symbolic ''function notation,'' are to be understood in the sense of distribution theory.

We now define a special class of distributions in $\mathscr{S}'(\mathsf{M})$, introduced by Araki.[1] As we said in the Introduction, and as discussed in detail in Araki's paper, we are led to these distributions within quantum field theory when we consider matrix elements of commutators (or anticommutators) of field operators between vector states of compact support in momentum space.

Definition 2.1: The set $\mathscr{S}'_A(\mathsf{M})$ of tempered distributions is the set of all $\xi(x) \in \mathscr{S}'(\mathsf{M})$ for which the Fourier transform satisfies the condition

$$\mathrm{supp}(\hat{\xi}(p)) \subset \overline{V}_+(p'') \cup \overline{V}_-(p')$$

for *some* elements p' and p'' of four-momentum space.

It is well known that the support of a distribution is not independent of the support of its Fourier transform. A principal issue in this paper, as well as in Araki's paper, is the nature of the null regions of the distributions in $\mathscr{S}'_A(\mathsf{M})$. By the *null region* $N(\xi)$ of any $\xi \in \mathscr{S}'_A(\mathsf{M})$ we understand the open complement of the support of ξ (in the sense of distribution theory). Let S be an open subset of M. It can well happen that S is not the null region of any element of $\mathscr{S}'_A(\mathsf{M})$, and in general there will exist an open set $\nu(S)$, strictly larger than S, such that $N(\xi) \supset \nu(S)$ for *all* $\xi \in \mathscr{S}'_A(\mathsf{M})$ for which $N(\xi) \supset S$. (We always have $(S^{cc})^o \supset \nu(S)$: see Proposition 4.4 in Section IV. As an important example of this phenomenon we here state a result originally due to Borchers,[12] in the following form.

Lemma 2.2: Let $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$, and let $N(\xi)$ be the null region of $\xi(x)$. If $N(\xi)$ includes the closed *timelike* line segment $[x', x'']$, then $N(\xi)$ includes the closed double cone $D(x', x'')$.

This result, which follows readily from the general theory of Araki,[1] has turned out to be a very useful ''extension principle.'' We will depend on this principle in our paper, and before we discuss Araki's theory further we will pursue the implications of this lemma to some length. First we present an elaboration which follows readily from the above lemma. See also a similar result by Kraus.[13]

Proposition 2.3: Let $s \to x(s)$ be a continuous mapping of $[-1,1]$ into M such that the vector $x(s) - x(-s)$ is in $\overline{V}_+$ for all $s \in [0,1]$. Let $\xi \in \mathscr{S}'_A(\mathsf{M})$ be such that the curve $C = \{x(s) | s \in [-1,1]\}$ is contained in the null region $N(\xi)$. Then the line segment $[x(-1), x(1)]$ is also contained in $N(\xi)$, and there exists a double cone $D \in \mathscr{D}$, contained in $N(\xi)$, which contains the points $x(-1)$ and $x(1)$.

Proof: (1) We first consider the case when $x(s) - x(-s)$ is *strictly forward timelike* for all $s \in (0,1]$. For any $s \in [0,1]$, let $L(s)$ be the line segment $[x(-s), x(s)]$, which is thus timelike for all $s \in (0,1]$. For any $s \in (0,1]$, let $D(s)$ be the double cone in $\mathscr{D}$ with apices $x(-s)$ and $x(s)$. With reference to Lemma 2.2 we note that if $L(s) \subset N(\xi)$, then $D(s) \subset N(\xi)$.

We have $L(0) = x(0) \in N(\xi)$, and since C is continuous and $N(\xi)$ is open, it follows that there exists an $s_0 \in (0,1]$ such that $L(s) \subset N(\xi)$ for all $s \in [0, s_0)$. Let s_m be the largest such s_0. We then have $(x(-s_m), x(s_m)) \subset \cup \{D(s) | s_m > s > 0\} \subset N(\xi)$, and since $x(-s_m) \in N(\xi)$ and $x(s_m) \in N(\xi)$ by the premises, it follows that $L(s_m) \subset N(\xi)$, and hence also $D(s_m) \subset N(\xi)$. If $s_m = 1$, there is nothing to prove, and we thus consider the possibility that $1 > s_m > 0$. Since $D(s_m)$ is compact and $N(\xi)$ is open, there must exist an open double cone D^o such that $L(s_m) \subset D(s_m) \subset D^o \subset N(\xi)$. For a sufficiently small $\delta > 0$ we obviously have $L(s) \subset D^o \subset N(\xi)$ for all $s \in (s_m, s_m + \delta)$, which contradicts the definition of s_m. Hence $L(1) \subset N(\xi)$.

(2) We now consider the general case. Let v be a forward timelike vector. For any $p > 0$ we define the function $x_p(s)$ on $[-1,1]$ with range in M by $x_p(s) = x(s) + spv$. Since the curve C is compact it follows that $C_p = \{x_p(s) | s \in [-1,1]\} \subset N(\xi)$ for all sufficiently small $p > 0$. Let $p > 0$ be such that $C_p \subset N(\xi)$. We have $x_p(s) - x_p(-s) \in V_+$ for all $s \in (0,1]$, and by the result in step

(1) we conclude that the double cone D with apices $x_p(-1)$ and $x_p(1)$ is contained in $N(\xi)$. By the construction of the function $x_p(s)$, it readily follows that the line segment $[x(-1),x(1)]$ is in D, and hence in $N(\xi)$. This completes the proof.

A special case arises when the above function $x(s)$ is such that $x(s)=x(0)$ for all $s\in[-1,0]$. The points $x_0=x(0)$ and $x_1=x(1)$ are then joined by the continuous curve $C_1=\{x(s)|s\in[0,1]\}$. If $x-x_0\in\overline{V}_+$ for all $x\in C_1$, and if $C_1\subset N(\xi)$, then $[x_0,x_1]\subset N(\xi)$.

We remark here that the principle in Proposition 2.3 is valid for a Minkowski-type spacetime of *arbitrary* dimensionality, because Araki's derivation of Lemma 2.2 (on which the conclusion in the proposition is based) is generally valid for such a space. Concerning the premises in the proposition we note the following. Let x_1 and x_{-1} be two points such that $x_1\in\overline{V}_+(x_{-1})$, and let C be a continuous curve with these points as end points. Suppose that $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$, with a null region $N(\xi)\supset C$. In a *two*-dimensional spacetime one can *always* find a parametrization of the curve C which satisfies the premises in Proposition 2.3, and hence it follows that $[x_{-1},x_1]\subset N(\xi)$. In a spacetime of higher dimensionality it can well happen that *no* parametrization which satisfies the premises in Proposition 2.3 exists for a *given* curve C. In Section IV we will show by an explicit example (in which $\xi(x)$ is a suitable solution of the Klein–Gordon equation) that the conclusion in the proposition can fail in such a situation (see Example 4.11).

We think that Proposition 2.3 is of a certain general interest, with other potential applications than those in this paper. Before we pursue the implications of this proposition we quote the following result from TW-I[8] (see Lemma 4.4 and Proposition 4.3):

Lemma 2.4: Let $l'=\{x'-\lambda u'|\lambda\in\mathsf{R}^1\}$ and $l''=\{x''+\lambda u''|\lambda\in\mathsf{R}^1\}$ be two isotropic lines, where u' and u'' are *forward* isotropic. Let π' be the unique characteristic plane which contains l', and let π'' be the unique characteristic plane which contains l''. We write $x'(\lambda)=x'-\lambda u'$ and $x''(\lambda)=x''+\lambda u''$, and we define S to be the *open* region (possibly empty) bounded *above* by π'' and bounded *below* by π'. Then S is nonempty if and only if there exists a λ_0 such that $x''(\lambda_0)-x'(\lambda_0)$ is *forward timelike,* and if this is the case, then

$$S=\cup\{D(x'(\lambda),x''(\lambda))^o|\lambda\geqslant\lambda_1\}=\mathscr{E}(S) \tag{2.2}$$

for any $\lambda_1\geqslant\lambda_0$.

The meaning of the symbol $\mathscr{E}(S)$ was explained in the Introduction: see also Lemma 2.6 below. For the time being we only need to know that the first two members in (2.2) are equal. As we discussed in TW-I,[8] the above conclusion holds in any Minkowski-type space M_n of dimensionality n. We are interested in the case $\mathsf{M}_4=\mathsf{M}$ (physical spacetime) and in the present section also in the case M_5. On the basis of Proposition 2.3 and Lemma 2.4 we can now obtain the following results.

Proposition 2.5: (a) Let $[x',x'']$ be a *timelike* interval, with $x''\in V_+(x')$. Let u' and u'' be two *forward isotropic* vectors, and let h_+ and h_- be two closed half-lines defined by $h_+=\{x''+\lambda u''|\lambda\geqslant 0\}$ and $h_-=\{x'-\lambda u'|\lambda\geqslant 0\}$. Let π'' be the unique characteristic plane containing h_+, let π' be the unique characteristic plane containing h_-, and let S be the *open* region consisting of all points x which lie *below* π'' and lie *above* π'. Hence $S=W(\pi',\pi'')\in\mathscr{W}$ if u' and u'' are not parallel, and S is an open ''slab-region'' between two characteristic planes if u' and u'' are parallel. Let C be the continuous curve $C=h_-\cup[x',x'']\cup h_+$.

Let $\xi\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)$. If $N(\xi)\supset C$, then $N(\xi)\supset S$.

(b) If u' and u'' are not parallel, then the above conclusion also holds in the case $x'=x''$.

(c) Let $D^o=D(x',x'')^o$ be an open double cone, and let u be a forward isotropic vector. We define the open ''tube-region'' T by $T=\{x+\lambda u|x\in D^o,\lambda\in\mathsf{R}^1\}$. Let π'' be the unique characteristic plane passing through x'', with normal u, and let π' be the plane parallel with π'' passing through x'. Let S be the ''slab-region'' bounded above by π'' and bounded below by π'.

Let $\xi\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)\supset T$. Then $N(\xi)\supset S$.

Proof: (1) We consider part (a). Let $x_0\in(x',x'')$, $C_+=C\cap V_+(x_0)$, and $C_-=C\cap V_-(x_0)$. We have $C=C_-\cup\{x_0\}\cup C_+$, and if $y'\in C_-$ and $y''\in C_+$, then $y''\in V_+(y')$. By an obvious application of Proposition 2.3 we conclude that $D(y',y'')\subset N(\xi)$. Hence with $x'(\lambda)=x'-\lambda u'$ and $x''(\lambda)=x''+\lambda u''$ we have $D(x'(\lambda),x''(\lambda))^o\subset N(\xi)$ for all $\lambda\geqslant 0$, and it then follows from Lemma 2.4 that $S\subset N(\xi)$.

(2) The assertion in part (b) follows trivially from what has been said above. We consider part (c). Let $x'(\lambda)=x'-\lambda u$ and $x''(\lambda)=x''+\lambda u$. With the stated premises we have $x''(\lambda) \in V_+(x'(\lambda))$ and $(x'(\lambda),x''(\lambda))\subset T\subset N(\xi)$, for all $\lambda\geqslant 0$. Hence $D(x'(\lambda),x''(\lambda))^o\subset N(\xi)$ for $\lambda\geqslant 0$, and it then follows from Lemma 2.4 that $S\subset N(\xi)$. This completes the proof.

This proposition provides further examples of the ''enlargement'' of a region contained in a null region $N(\xi)$ on the basis of the principle in Lemma 2.2, and in some cases the extent of the enlargement can be rather surprising.

The principle in Lemma 2.2 is just one aspect of the general problem of determining the common null region of all distributions $\xi\in\mathscr{S}'_A(\mathsf{M})$ which satisfy the condition that $N(\xi)$ includes some particular ''given'' $S\subset\mathsf{M}$. Before we outline Araki's approach to this problem, we consider the analogous problem for the solutions of the Klein–Gordon equation. Let $\phi(x)$ be a C_2-solution of the Klein–Gordon equation for mass $m\geqslant 0$ on an n-dimensional Minkowski-type space M_n (with $n-1\geqslant 1$ spatial dimensions and one temporal dimension), and let $N(\phi)$ be the null region of $\phi(x)$. There are two well-known ''classical'' principles concerning the support of the solution:[11] I) If $N(\phi)$ contains an open timelike segment (x',x''), then $N(\phi)$ contains the open double cone $D(x',x'')^o$. II) If $N(\phi)$ contains the base of an open double cone D^o, then $N(\phi)$ contains D^o. Stated in this form the two principles apply equally well to distribution-valued solutions of the Klein–Gordon equation. According to Lemma 2.2 the Principle I also applies to tempered distributions in the set $\mathscr{S}'_A(\mathsf{M})$, but this is in general *not* the case for Principle II. We note here that a C_2-solution $\phi(x)$ of the Klein–Gordon equation for M with $m\geqslant 0$ can be identified with a distribution in $\mathscr{S}'_A(\mathsf{M})$ provided that $\phi(x)$ satisfies certain growth conditions as x tends to infinity. This implies constraints on the extensions S_e of S in the case of the Araki distributions $\mathscr{S}'_A(\mathsf{M})$: an extension S_e can never be larger than it can be in the case of solutions of the Klein–Gordon equation of the stated kind.

As we discussed in Section IV in TW-I,[8] the above principles motivate the consideration of an *envelope of uniqueness* $\mathscr{E}(S)$ of any open set $S\subset\mathsf{M}$. We here state the relevant facts as a lemma for later reference. (See Definition and Proposition 4.2 in TW-I.[8])

Lemma 2.6: (a) Let S be an *open* subset of Minkowski spacetime M. There then exists a unique *smallest* subset X of M which satisfies the following three conditions: I) $X\supset S$. II) If X contains the base B_0 of an open double cone $D^o\subset\mathsf{M}$, then X contains D^o. III) If X contains the open timelike line segment (x',x''), then X contains the open double cone $D^o=D(x',x'')^o$ with x' and x'' as apices.

Here, and throughout this paper, we denote this smallest set by $\mathscr{E}(S)$, *for any open* $S\subset\mathsf{M}$.

(b) The set $\mathscr{E}(S)$ is *open*, and it satisfies the conditions $\mathscr{E}(S)\subset(S^{cc})^o\subset S^{cc}=\mathscr{E}(S)^{cc}$. If S is *causally complete*, then $\mathscr{E}(S)=S$.

Furthermore, $\mathscr{E}(\mathscr{E}(S))=\mathscr{E}(S)$, and $\mathscr{E}(S_1)\supset\mathscr{E}(S_2)$ if $S_1\supset S_2$.

(c) If the *open* set S contains either the base B of a closed double cone $D=D(x',x'')\in\mathscr{D}$, or else the closed timelike line segment $[x',x'']$, then $\mathscr{E}(S)$ contains D.

(d) If S_1 and S_2 are two open subsets of M such that $\mathscr{E}(S_1)=S_1$ and $\mathscr{E}(S_2)=S_2$, then $\mathscr{E}(S_1\cap S_2)=S_1\cap S_2$. If S_1 and S_2 are *disjoint*, then $\mathscr{E}(S_1\cup S_2)=S_1\cup S_2$.

The properties of the mapping $S\to\mathscr{E}(S)$ were discussed in considerable detail in TW-I,[8] and we will not repeat this discussion here. The mapping $S\to\mathscr{E}(S)$ can obviously be defined in the same manner for any Minkowski space M_n. What was said above about the null region $N(\phi)$ of a C_2-solution of the Klein–Gordon equation can now be rephrased as follows. We have $\mathscr{E}(N(\phi))=N(\phi)$, and if $S\subset N(\phi)$, and S is open, then $\mathscr{E}(S)\subset N(\phi)$. The same conclusion holds for distribution-valued solutions.

We now outline the ideas in Araki's paper.[1] We regard Minkowski spacetime M as corresponding to a timelike hyperplane in a five-dimensional Minkowski space $\mathsf{M}_5=\mathsf{M}\times\mathsf{R}^1$. A point $\hat{x}$ in M_5 is the pair $\hat{x}=(x,s)$, with $x\in\mathsf{M}$ and $s\in\mathsf{R}^1$, and with the ''metric'' on M_5 defined by $\hat{x}\cdot\hat{x}=x\cdot x-s^2$, i.e., the added dimension is spacelike. We denote by Γ the one-one mapping of M into M_5 defined by $\Gamma x=(x,0)$ for any $x\in\mathsf{M}$, and we write $\Gamma X=\{(x,0)|x\in X\}$ for any $X\subset\mathsf{M}$. The inverse mapping is thus

$$\Gamma^{-1}Y=\{x|x\in\mathsf{M},\quad (x,0)\in Y\}$$

for any $Y \subset \mathsf{M}_5$. The four-dimensional Minkowski spacetime is thus in effect identified with the hyperplane $\Gamma\mathsf{M} = \{(x,0) | x \in \mathsf{M}\}$ in M_5, and the image of any $Y \subset \mathsf{M}_5$ under Γ^{-1} depends only on the intersection of Y with this hyperplane.

Let $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$. Somewhat heuristically described the idea is now to associate with $\xi(x)$ a solution $\phi(x,s)$ of the (massless) wave equation on M_5 such that $\phi(x,s) = \phi(x,-s)$ and $\phi(x,0) = \xi(x)$. If X is in the null region $N(\xi)$ of ξ, then ΓX is in the null region $N(\phi)$ of ϕ, and if $Y \subset N(\phi)$, then $\Gamma^{-1} Y \subset N(\xi)$. The two principles concerning the null region apply to ϕ (in M_5), and the second principle thus applies in an *indirect* fashion to the null region of $\xi(x)$. The construction of $\phi(x,s)$ is rather straightforward if the support of the Fourier transform $\hat{\xi}(p)$ of $\xi(x)$ is contained in the lightcone $\overline{V}(0)$ in momentum space (as we will discuss in Section V), but in view of Definition 2.1 this is not necessarily the case, and the above description is not quite right, although the basic ideas are as stated. The *rigorous* approach to this issue involves a number of not entirely trivial questions of detail, and for the discussion and settlement of these we refer to Araki's paper.[1] In our paper we depend on a specific result in Araki's paper, which we state in the form of a lemma.

Lemma 2.7: Let four-dimensional Minkowski spacetime M be related to five-dimensional Minkowski space M_5 as described above. For any *open* $S \subset \mathsf{M}$, let $E(S) \subset \mathsf{M}_5$ be the smallest set in M_5 which contains ΓS and satisfies the condition that it contains the open double cone $D_5^o \subset \mathsf{M}_5$ if it contains *either* the base of D_5^o, *or* the open timelike line segment with the apices of D_5^o as its endpoints. We define the subset $a(S)$ of M by $a(S) = \Gamma^{-1} E(S)$. Then $S \subset a(S)$, and $a(S)$ is *open*.

Let $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$, and let $S \subset \mathsf{M}$ be contained in the null region $N(\xi)$. Then $a(S) \subset N(\xi)$.

In what follows we refer to the set $a(S)$ as the *Araki envelope of* S. We note that, given an arbitrary open region S, it is not at all immediately obvious what the region $a(S)$ ''looks like.'' It is not, however, difficult to see that Lemma 2.2 does follow immediately from Lemma 2.7. Since Lemma 2.2 involves an important principle we wanted to state it explicitly, lest it be obscured by the geometric complications in Lemma 2.7.

For the visualization of the sets $E(S)$ and $a(S)$ it is useful to think about the mapping $S \to E(S)$ as arising in two steps. Let $S \subset \mathsf{M}$ be open. To S corresponds the *relatively open* subset ΓS of the 4-plane $\Gamma\mathsf{M}$ in M_5. We first ''fatten'' ΓS a little bit to obtain an open subset of M_5, and then apply the mapping in Lemma 2.6 to this subset to obtain $E(S)$. It is worthwhile to formulate this idea precisely as follows.

Lemma 2.8: Let the premises and notation be as in Lemma 2.7 above, and let $R \to \mathscr{E}_5(R)$ be the mapping in Lemma 2.6, for the five-dimensional space M_5. For any open $S \subset \mathsf{M}$ we define the region $E_0(S)$ in M_5 by

$$E_0(S) = \cup \{E(D^o) | D \in \mathscr{D},\ D^o \subset S\}. \tag{2.3}$$

Then $E_0(S)$ and $E(S)$ are *open*, and $\Gamma^{-1} E_0(S) = S$. Furthermore,

$$\mathscr{E}_5(E(S)) = E(S) = \mathscr{E}_5(E_0(S)). \tag{2.4}$$

Proof: We consider an open double cone $D^o = D(x',x'')^o \subset \mathsf{M}$. Then $E(D^o)$ is obviously an open double cone $D_5^o \subset \mathsf{M}_5$ with apices $\hat{x}' = (x',0)$ and $\hat{x}'' = (x'',0)$, and we have $\Gamma^{-1} D_5^o = D^o$. It follows that the set $E_0(S)$ in (2.3) is open, and since the open set S is the union of all open double cones which it contains, it follows that $\Gamma^{-1} E_0(S) = S$. We have $E(S) \supset E_0(S) \supset \Gamma S$, and in view of the definition of $E(S)$ we conclude that $E(S) = \mathscr{E}_5(E_0(S))$. This implies that $E(S)$ is open, and since $\mathscr{E}_5(\mathscr{E}_5(R)) = \mathscr{E}_5(R)$ for any open $R \subset \mathsf{M}_5$, we have $\mathscr{E}_5(E(S)) = \mathscr{E}_5(E_0(S)) = E(S)$. This completes the proof.

We note that it follows from this that $a(S) = \Gamma^{-1} E(S)$ is open since $E(S)$ is open.

Before we discuss the determination of the Araki envelope $a(S)$ in particular cases it will be useful to consider another mapping between the spaces M and M_5. We define the set-mapping Λ of subsets of M to subsets of M_5 by $\Lambda X = \{(x,s) | x \in X, s \in \mathsf{R}^1\}$, for any $X \subset \mathsf{M}$. For a singleton we write $\Lambda x = \{(x,s) | s \in \mathsf{R}^1\}$. A point $x \in \mathsf{M}$ is thus mapped onto the line in M_5 which passes through the point $(x,0)$ and is parallel with the 5-axis. We note that $\Gamma^{-1} \Lambda X = X$ for any $X \subset \mathsf{M}$. Every set

ΛX is, of course, mapped onto itself by any translation in the 5-direction, and, conversely, every set $Y \subset \mathsf{M}_5$ invariant under translations in the 5-direction is of the form ΛX, with $X = \Gamma^{-1} Y$, and for such a set we thus have $Y = \Lambda \Gamma^{-1} Y$.

Lemma 2.9: Let Λ be the mapping of subsets of M to subsets of M_5 defined above. Let the mapping $S \to \mathscr{E}(S)$ for M be defined as in Lemma 2.6, and let $R \to \mathscr{E}_5(R)$ be the corresponding mapping for M_5. Then:

(a) If $X \subset \mathsf{M}$, then $\Lambda X \subset \mathsf{M}_5$ is open if and only if X is open.

(b) The mapping Λ commutes with causal complementation and causal completion, in the sense that for any $X \subset \mathsf{M}$,

$$(\Lambda X)^c = \Lambda(X^c), \quad (\Lambda X)^{cc} = \Lambda(X^{cc}). \tag{2.5}$$

(c) If $S \subset \mathsf{M}$ is open, then

$$\mathscr{E}_5(\Lambda S) = \Lambda \mathscr{E}(S). \tag{2.6}$$

Proof: (1) The assertion in part (a) is trivial. We consider part (b). Let $X \subset \mathsf{M}$. The set ΛX is invariant under any translation in the 5-direction, and the set $(\Lambda X)^c \subset \mathsf{M}_5$ obviously has the same property. Hence $(\Lambda X)^c = \Lambda Y$, with $Y = \Gamma^{-1}((\Lambda X)^c) \subset \mathsf{M}$. Let $\hat{x} = (x, s_1) \in \Lambda X$, and let $\hat{y} = (y, s_2) \in \mathsf{M}_5$. We then have $(x,s) \in \Lambda X$ for all s, and the separation between $\hat{y}$ and (x,s) is spacelike (within M_5) for all s if and only if the separation between (x,s_2) and $\hat{y} = (y,s_2)$ is spacelike, but this is the case if and only if the separation between x and y is spacelike (within M). It follows that $Y = X^c$, and hence $(\Lambda X)^c = \Lambda(X^c)$. The second relation in (2.5) then follows readily.

(2) We consider part (c), assuming that $S \subset \mathsf{M}$ is open. Since ΛS is invariant under all translations in the 5-direction it follows trivially that $\mathscr{E}_5(\Lambda S)$ has the same property, and we thus have $\mathscr{E}_5(\Lambda S) = \Lambda S_1$, with $S_1 = \Gamma^{-1} \mathscr{E}_5(\Lambda S) \supset \Gamma^{-1} \Lambda S = S$. We consider the open double cone $D_4^o = D(x,y)^o \subset \mathsf{M}$ and the open double cone $D_5^o = D_5(\hat{x}, \hat{y})^o = E_0(D_4^o) \subset \mathsf{M}_5$, with $\hat{x} = (x,0)$ and $\hat{y} = (y,0)$. Let B_{04} be the base of D_4^o and let B_{05} be the base of D_5^o. It is trivial that: (a) ΛS_1 contains D_5^o if and only if S_1 contains D_4^o, (b) ΛS_1 contains B_{05} if and only if S_1 contains B_{04}, and (c) ΛS_1 contains $(\hat{x}, \hat{y})$ if and only if S_1 contains (x,y). From this we conclude that $\mathscr{E}(S_1) = S_1$, since $\mathscr{E}_5(\Lambda S_1) = \mathscr{E}_5(\mathscr{E}_5(\Lambda S)) = \mathscr{E}_5(\Lambda S) = \Lambda S_1$. Since $S_1 \supset S$ it follows that $\mathscr{E}(S_1) = S_1 \supset \mathscr{E}(S)$, and hence we have $\mathscr{E}_5(\Lambda S) \supset \Lambda \mathscr{E}(S)$.

(3) In the continuation of the proof we will depend on the following rather plausible result from TW-I:[8] If X is a bounded open convex subset of M, then $\mathscr{E}(X) = X^{cc}$. (See Lemma 4.7 in the quoted paper.)

Let now X_5 be a bounded, open convex subset of M_5, and let $X_{5\Lambda}$ be defined by $X_{5\Lambda} = \{\hat{x} + (0,\lambda) | \hat{x} \in X_5, \lambda \in \mathsf{R}^1\}$, and let $X_4 = \Gamma^{-1} X_{5\Lambda}$. We can regard ΓX_4 as the orthogonal projection of X_5 into $\Gamma \mathsf{M}$, and we have $X_{5\Lambda} = \Lambda X_4$. The set $X_{5\Lambda}$ is obviously open and convex, and the set X_4 is open and convex and bounded. Hence we have $\mathscr{E}(X_4) = X_4^{cc}$, and in view of (2.5) we have $\Lambda \mathscr{E}(X_4) = \Lambda(X_4^{cc}) = (\Lambda X_4)^{cc} = X_{5\Lambda}^{cc} \supset \mathscr{E}_5(X_{5\Lambda}) \supset \mathscr{E}_5(X_5)$.

Suppose now that the above set X_5 is included in $\Lambda \mathscr{E}(S)$, and hence $\Lambda X_4 = X_{5\Lambda} \subset \Lambda \mathscr{E}(S)$. This implies that $X_4 \subset \mathscr{E}(S)$, and hence $\mathscr{E}(X_4) \subset \mathscr{E}(S)$. By the above it then follows that $\mathscr{E}_5(X_5) \subset \Lambda \mathscr{E}(X_4) \subset \Lambda \mathscr{E}(S)$.

(4) Let $D = D(\hat{x}', \hat{x}'')$ be a closed double cone in M_5 with base B. Suppose that $B \subset \Lambda \mathscr{E}(S)$. It is obvious that there then exists a bounded open convex set X_5 such that $B \subset X_5 \subset \Lambda \mathscr{E}(S)$, and it then follows from the result in step (3) that $D \subset \Lambda \mathscr{E}(S)$. Suppose instead that $[\hat{x}', \hat{x}''] \subset \Lambda \mathscr{E}(S)$. There then exists a bounded open convex set X_5 such that $[\hat{x}', \hat{x}''] \subset X_5 \subset \Lambda \mathscr{E}(S)$, and it then follows, in this case too, that $D \subset \Lambda \mathscr{E}(S)$. These results imply that $\Lambda \mathscr{E}(S) = \mathscr{E}_5(\Lambda \mathscr{E}(S))$. Since $\Lambda \mathscr{E}(S) \supset \Lambda S$ we have $\mathscr{E}_5(\Lambda \mathscr{E}(S)) \supset \mathscr{E}_5(\Lambda S)$, and in view of the result in step (2) the equality in (2.6) follows. This completes the proof.

For an arbitrary open set $S \subset \mathsf{M}$ the Araki envelope $a(S)$ is in general *smaller* than the envelope $\mathscr{E}(S)$. An example of substantial physical interest is an open set T consisting of all points between two parallel *spacelike* planes. Such a region is called *a time-slice* in quantum field

theory. We trivially have $\mathscr{E}(T)=\mathsf{M}$, but $a(T)=T$, as one easily sees. We will have more to say about this later. At this point we note some simple inclusion relations involving the mappings a and $\mathscr{E}$.

Proposition 2.10: Let the premises and notation be as in Lemma 2.7, and let $S\to\mathscr{E}(S)$ be the mapping in Lemma 2.6. If $S\subset\mathsf{M}$ is *open*, then

$$(S^{cc})^o\supset\mathscr{E}(S)=a(\mathscr{E}(S))=\mathscr{E}(a(S))\supset a(S)\supset S. \tag{2.7}$$

Proof: By Lemma 2.6 we have $(S^{cc})^o\supset\mathscr{E}(S)\supset S$, and it is trivial that $a(S)\supset S$ and that $\mathscr{E}(a(S))\supset a(S)$. We follow the notation in Lemmas 2.8 and 2.9. We trivially have $\Lambda D^o\supset E_0(D^o)$ for any open double cone $D^o\subset\mathsf{M}$, and hence $\Lambda S\supset E_0(S)$ for any open $S\subset\mathsf{M}$. In view of (2.4) and (2.6) we have $\Lambda\mathscr{E}(S)=\mathscr{E}_5(\Lambda S)\supset\mathscr{E}_5(E_0(S))=E(S)$, and hence $\mathscr{E}(S)=\Gamma^{-1}\Lambda\mathscr{E}(S)\supset\Gamma^{-1}E(S)=a(S)$. Forming the $\mathscr{E}$-envelope of all members in the relation $\mathscr{E}(S)\supset a(S)\supset S$ we obtain $\mathscr{E}(S)=\mathscr{E}(a(S))$, since $\mathscr{E}(\mathscr{E}(S))=\mathscr{E}(S)$. Finally we note that $a(\mathscr{E}(S))=\Gamma^{-1}E(\mathscr{E}(S))\subset\Gamma^{-1}\Lambda\mathscr{E}(S)=\mathscr{E}(S)$. Since $a(\mathscr{E}(S))\supset\mathscr{E}(S)$ we conclude that $a(\mathscr{E}(S))=\mathscr{E}(S)$. This completes the proof.

It is worth noting that it follows from (2.7) that if S happens to be such that $\mathscr{E}(S)=S$, then $a(S)=S$. In particular, if S is open and causally complete, then $a(S)=\mathscr{E}(S)=S$.

In this paper we are particularly interested in regions S which are unions of wedges in $\mathscr{W}$. In this special situation we depend on a result which we will now derive from Lemmas 2.4 and 2.7, and from Proposition 2.10.

Theorem 2.11: Let $\{S_\alpha|\alpha\in\mathscr{I}\}$, where $\mathscr{I}$ is some index set, be a family of open regions $S_\alpha\subset\mathsf{M}$ such that S_α is bounded above by the characteristic plane π''_α and below by the characteristic plane π'_α. Let $S=\cup\{S_\alpha|\alpha\in\mathscr{I}\}$. We define the Araki envelope $a(S)$ as in Lemma 2.7, and we define the open set $\mathscr{E}(S)$ as in Lemma 2.6. Then

$$a(S)=\mathscr{E}(S)\subset(S^{cc})^o \tag{2.8}$$

and if $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)\supset S$, then $N(\xi)\supset\mathscr{E}(S)$.

Proof: (1) We follow the notation in Lemmas 2.6–2.9. We first show that for each $\alpha\in\mathscr{I}$ we have

$$E(S_\alpha)=\Lambda S_\alpha, \tag{2.9}$$

and that $E(S_\alpha)$ is the region in M_5 bounded by the characteristic 4-planes

$$\Lambda\pi'_\alpha=\{(x,s)|x\in\pi'_\alpha,s\in\mathsf{R}^1\},\quad \Lambda\pi''_\alpha=\{(x,s)|x\in\pi''_\alpha,s\in\mathsf{R}^1\},$$

such that $E(S_\alpha)$ is below $\Lambda\pi''_\alpha$ and above $\Lambda\pi'_\alpha$. We assume that $S_\alpha\neq\varnothing$, since the relation (2.9) is otherwise trivial. Let $l'=\{x'-\lambda u'|\lambda\in\mathsf{R}^1\}$ be an isotropic line in π'_α, and let $l''=\{x''+\lambda u''|\lambda\in\mathsf{R}^1\}$ be an isotropic line in π''_α, where the vectors u' and u'' are chosen as *forward* isotropic. We write $x'(\lambda)=x'-\lambda u'$ and $x''(\lambda)=x''+\lambda u''$, for $\lambda\in\mathsf{R}^1$. By Lemma 2.4 there then exists a λ_0 such that $x''(\lambda)-x'(\lambda)$ is forward timelike for all $\lambda\geqslant\lambda_0$, and for such a λ we have $(x'(\lambda),x''(\lambda))\subset S_\alpha$.

The sets $\Gamma l'$ and $\Gamma l''$ are isotropic lines in M_5, and it is easily seen that $\Lambda\pi'_\alpha$ is the unique characteristic plane which contains $\Gamma l'$, and that $\Lambda\pi''_\alpha$ is the unique characteristic plane which contains $\Gamma l''$. The region ΛS_α is thus the region bounded below by $\Lambda\pi'_\alpha$ and bounded above by $\Lambda\pi''_\alpha$. For any $\lambda\geqslant\lambda_0$ the open timelike line segment $(\Gamma x'(\lambda),\Gamma x''(\lambda))$ is in ΓS_α, and hence in $E(S_\alpha)$. By the definition of $E(S_\alpha)$ this implies that the open double cone $D(\Gamma x'(\lambda),\Gamma x''(\lambda))^o$ is in $E(S_\alpha)$. Since this holds for all $\lambda\geqslant\lambda_0$, it follows from Lemma 2.4 that $\Lambda S_\alpha\subset E(S_\alpha)$.

It is easily seen that $S_\alpha=\mathscr{E}(S_\alpha)$ if the planes π'_α and π''_α are parallel. If π'_α and π''_α are not parallel, then $S_\alpha=W(\pi'_\alpha,\pi''_\alpha)$, and hence $S_\alpha=\mathscr{E}(S_\alpha)$. We thus always have $S_\alpha=\mathscr{E}(S_\alpha)$, and hence, by (2.6), $\Lambda S_\alpha=\Lambda\mathscr{E}(S_\alpha)=\mathscr{E}_5(\Lambda S_\alpha)$. With the notation in Lemma 2.8 we have $\Lambda S_\alpha\supset E_0(S_\alpha)$, and hence $\Lambda S_\alpha=\mathscr{E}_5(\Lambda S_\alpha)\supset\mathscr{E}_5(E_0(S_\alpha))=E(S_\alpha)$, by Lemma 2.8. In view of what was shown above we conclude that $\Lambda S_\alpha=E(S_\alpha)$, as asserted in (2.9).

(2) By (2.9) we obtain $E(S)\supset\cup\{E(S_\alpha)|\alpha\in\mathscr{I}\}=\cup\{\Lambda S_\alpha|\alpha\in\mathscr{I}\}=\Lambda S$, since $S=\cup\{S_\alpha|\alpha\in\mathscr{I}\}$. By (2.4) we have $E(S)=\mathscr{E}_5(E(S))$, and hence $E(S)=\mathscr{E}_5(E(S))\supset\mathscr{E}_5(\Lambda S)$. By (2.6) we

have $\mathscr{E}_5(\Lambda S)=\Lambda\mathscr{E}(S)$. Hence $E(S)\supset\Lambda\mathscr{E}(S)$, and it follows that $a(S)=\Gamma^{-1}E(S)\supset\Gamma^{-1}\Lambda\mathscr{E}(S)=\mathscr{E}(S)$. By (2.7) we have $a(S)\subset\mathscr{E}(S)\subset(S^{cc})^o$, and the relation (2.8) follows. This completes the proof.

The virtue of this theorem is that it will now not be necessary to make any further excursions into five-dimensional Minkowski space when we want to find the envelope $a(S)$ for a union S of *wedges*. Rather than working with Lemma 2.7 we will depend on Theorem 2.11, in which the five-dimensional space is not mentioned explicitly. For an arbitrary $R\subset\mathsf{M}$ it is in general much easier to find (and visualize) the envelope $\mathscr{E}(R)$ than the Araki envelope $a(R)$. It is important to understand what it is that is so special about a union of *wedges*. The salient point is that the relation (2.9) holds *only* for the kind of set S_α described in the premises, or for a *disjoint* union of such sets, but for *no* other open subset R of M. On the basis of this observation it is not hard to see that the relation $E(S)=\mathscr{E}_5(\Lambda S)$ in step (2) of the proof holds *only* if S is the kind of set described in the premises, which includes the union of any set of wedges in $\mathscr{W}$. This relation, in turn, is crucial for the conclusion in (2.8), which at first sight seems to say that the second principle concerning the null region of a solution of the Klein–Gordon equation, discussed before Lemma 2.6, is applicable after all to the distributions in the set $\mathscr{S}'_A(\mathsf{M})$. Note, however, that to be able to conclude that a double cone D^o is included in $N(\xi)$, given that the base B_0 of the double cone is included in $N(\xi)$, we must know that B_0 is included in the *subset* S of $N(\xi)$, as described in the premises of the theorem. If we are only given that $B_0\subset N(\xi)$, we *cannot* in general conclude that $D^o\subset N(\xi)$.

In the next section we discuss a number of interesting implications of Theorem 2.11. We conclude this section with a discussion of certain properties of the mappings $S\to\mathscr{E}(S)$ and $S\to a(S)$ which are of interest in quantum field theory.

Proposition 2.12: Let $S\subset\mathsf{M}$ be *open*, and let S_n, $n\in\mathsf{Z}_+$, be a sequence of *open* subsets of S such that $S_{n+1}\supset S_n$ for all n, and such that $S=\cup\{S_n|n\in\mathsf{Z}_+\}$. Then $\mathscr{E}(S)=\cup\{\mathscr{E}(S_n)|n\in\mathsf{Z}_+\}$ and $a(S)=\cup\{a(S_n)|n\in\mathsf{Z}_+\}$.

Proof: The relation $\mathscr{E}(S)=\cup\{\mathscr{E}(S_n)|n\in\mathsf{Z}_+\}$ was proved in TW-I.[8] (See Proposition 4.12 in the quoted paper.) It holds for a Minkowski-type space M_n of any dimensionality n.

Let the notation be as in Lemma 2.8. With reference to (2.3) it is immediately obvious that $E_0(S)=\cup\{E_0(S_n)|n\in\mathsf{Z}_+\}$, and we trivially have $E_0(S_{n+1})\supset E_0(S_n)$ for all n. It then follows that $E(S)=\mathscr{E}_5(E_0(S))=\cup\{\mathscr{E}_5(E_0(S_n))|n\in\mathsf{Z}_+\}=\cup\{E(S_n)|n\in\mathsf{Z}_+\}$, and hence $a(S)=\Gamma^{-1}E(S)=\cup\{\Gamma^{-1}E(S_n)|n\in\mathsf{Z}_+\}=\cup\{a(S_n)|n\in\mathsf{Z}_+\}$. This completes the proof.

III. ABOUT ARAKI DISTRIBUTIONS WHICH VANISH ON A UNION OF CAUSAL WEDGES

In this section we will be mostly concerned with the common null region of all distributions in $\mathscr{S}'_A(\mathsf{M})$ which are known to vanish on a union S of wedges in $\mathscr{W}$. It follows from Theorem 2.11 that the common null region also includes the set $\mathscr{E}(S)$. For some configurations of wedges we have $\mathscr{E}(S)=(S^{cc})^o$, and we want to identify these configurations. When S is a union of wedges the causal complement $K=S^c$ of S is a closed, convex, causally complete subset of M, and the converse is *almost* true. In TW-I[8] we discussed the nature of closed, convex, causally complete subsets of M and their causal complements in considerable detail. For convenient reference we here quote as a theorem some results which are relevant and important for the discussion in the present paper. Parts (a), (c), (d), and (e) are verbatim quotes from Theorem 3.2 in TW-I,[8] and the (trivial) statements in part (b) are from Proposition 3.4 in that paper.

Theorem 3.1: (a) Let K be a closed, convex, causally complete subset of M such that K^c has a nonempty interior. Then

$$K=\cap\{\overline{W}|W\in\mathscr{W},\ \overline{W}\supset K\}=\cap\{W^c|W\in\mathscr{W},\ W\subset K^c\},$$

$$(K^c)^o=\cup\{W|W\in\mathscr{W},\ W^c\supset K\}=\cup\{W|W\in\mathscr{W},\ W\subset K^c\},$$

and

$$((K^c)^o)^c=K,\ \ \overline{(K^c)^o=K^c},\quad \overline{(K^c)^o}=(K^c)^o.$$

The set K has an empty interior if and only if it is contained in some characteristic plane.

(b) Every set $K \in \mathscr{K}_E$ (with $\mathscr{K}_E$ as described in Definition 1.1) satisfies the premises in part (a) above, and specifically K^c is nonempty and *open*.

(c) Let $\{K_\alpha | \alpha \in A\}$ be a family of subsets of M which satisfy the premises in part (a). Then $K = \cap\{K_\alpha | \alpha \in A\}$ also satisfies the premises in part (a).

(d) Let $\mathscr{F}$ be any family of wedges in $\mathscr{W}$, and let $S = \cup\{W | W \in \mathscr{F}\}$. Then $K = \cap\{W^c | W \in \mathscr{F}\} = S^c$ satisfies the premises in part (a).

If $\mathscr{F}$ is a *finite* family, then the set K^c is *open*.

(e) Let K be a closed, convex, causally complete subset of M. Then K^c has an empty interior if and only if K contains an isotropic line. If K is a *proper* subset of M containing an isotropic line l, then K is a closed convex union of a set of isotropic lines, all contained in the characteristic plane π containing l (and hence all parallel with l). The set K^c is then the relative (set) complement of K in π, and hence the union of all isotropic lines in π which do not intersect K.

This geometric theorem provides a framework for the discussion in this section, and the statement in part (d) is particularly relevant. In the ''favorable'' cases when we have $\mathscr{E}(S) = (S^{cc})^o$ the Araki envelope $a(S)$ is thus equal to the interior of the causal complement of a closed, convex, causally complete subset K of M, which is a useful and transparent description of $a(S)$.

In view of the discussion in the preceding section the problem of determining the Araki envelope $a(S)$ when S is a union of wedges is essentially a *geometric* problem, involving the determination of the envelope $\mathscr{E}(S)$ defined in Lemma 2.6. The latter problem was discussed in considerable detail in TW-I,[8] and we will depend on various results from that paper. We first quote, as a theorem, some results which will turn out to be useful in the study of unions of wedges, although wedges are not mentioned explicitly. This theorem is also of considerable of interest in its own right.

Theorem 3.2: Let $\Pi' = \{\pi'_\alpha | \alpha \in \mathscr{I}'\}$ and $\Pi'' = \{\pi''_\beta | \beta \in \mathscr{I}''\}$, where $\mathscr{I}'$ and $\mathscr{I}''$ are two index sets, be two families of *characteristic* planes in M. For each $\alpha \in \mathscr{I}'$ and $\beta \in \mathscr{I}''$, let $S_{\alpha\beta}$ be the *open* subset of M (possibly empty) which is bounded below by π'_α and bounded above by π''_β, and let $S = \cup\{S_{\alpha\beta} | \alpha \in \mathscr{I}', \beta \in \mathscr{I}''\}$ and $K = S^c$. The set S is thus the set of all points $x \in \mathsf{M}$ which lie *below* at least one of the planes in Π'' and lie *above* at least one of the planes in Π'. Furthermore we assume that for each $\alpha \in \mathscr{I}'$ there exists a $\beta \in \mathscr{I}''$ such that $S_{\alpha\beta} \neq \varnothing$, and that for each $\beta \in \mathscr{I}''$ there exists an $\alpha \in \mathscr{I}'$ such that $S_{\alpha\beta} \neq \varnothing$. Then:

(a) The set $K = S^c$ is *closed*, *convex*, and *causally complete*.

(b) Let the Araki envelope $a(S)$ be defined as in Lemma 2.7, and let the envelope $\mathscr{E}(S)$ be defined as in Lemma 2.6. Then

$$a(S) = \mathscr{E}(S).$$

If $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi) \supset S$, then $N(\xi) \supset \mathscr{E}(S)$.

(c) The following cases can occur:

Case I: $K \neq \varnothing$, and hence $S^{cc} \neq \mathsf{M}$. We then have

$$\mathscr{E}(S) = (S^{cc})^o = (K^c)^o.$$

Case II: $K = \varnothing$ and $S^{cc} = \mathsf{M} = \mathscr{E}(S)$.

Case IIIa: $K = \varnothing$, $S^{cc} = \mathsf{M}$, and $\mathscr{E}(S)$ is an open half-space bounded below by a characteristic plane π', in which case all the planes in Π' are parallel with π'.

Case IIIb: $K = \varnothing$, $S^{cc} = \mathsf{M}$, and $\mathscr{E}(S)$ is an open half-space bounded above by a characteristic plane π'', in which case all the planes in Π'' are parallel with π''.

Case IV: $K = \varnothing$, $S^{cc} = \mathsf{M}$, and $\mathscr{E}(S)$ is an open ''slab-region'' bounded above by a characteristic plane π'' and bounded below by a characteristic plane π' parallel with π'', in which case all the planes in $\Pi' \cup \Pi''$ are parallel with the planes π' and π''.

(d) If each one of the sets Π' and Π'' contains at least two nonparallel planes, then $\mathscr{E}(S) = S^{cc} = K^c$ and S^{cc} is *open*. If Π' and Π'' satisfy the *additional* condition that the set of normals of the planes in $\Pi' \cup \Pi''$ spans M, then $K \in \mathscr{K}_E$.

Proof: The statements in parts (a), (c), and (d) are quoted verbatim from Theorem 4.11 in TW-I,[8] and we will not repeat the proofs here. The set S satisfies the premises of Theorem 2.11 in the preceding section, and hence the assertions in part (b) follow.

For a discussion of various geometric features of the above configurations we refer to the proof of Theorem 4.11 in TW-I[8] and to Proposition 3.6 in the same paper. We consider a configuration for which there exists a plane $\pi'_\mu \in \Pi'$ such that $S_{\mu\beta}=\emptyset$ for *all* $\beta \in \mathscr{T}''$. The plane π'_μ is then *irrelevant* in the sense that it has no effect on the region S, and without loss of generality we can rule out such configurations, as we have done in the statement of premises. Note that we necessarily have $S\neq\emptyset$ with our premises. The case $S=\emptyset$ is naturally totally uninteresting.

In the next section we will show that the above result is the best possible, in the sense that $\mathscr{E}(S)$ is the largest common null region of *all* Araki distributions which vanish throughout S.

We now consider a result concerning two *intersecting* wedges.

Proposition 3.3: Let $W_1=W(\pi'_1,\pi''_1)$ and $W_2=W(\pi'_2,\pi''_2)$ be two wedges in $\mathscr{W}$ such that $W_1\cap W_2\neq\emptyset$. We write $S=W_1\cup W_2$ and $K=S^c=W_1^c\cap W_2^c$. Let $\mathscr{E}(S)$ be the envelope in Lemma 2.6, and let $a(S)$ be the Araki envelope in Lemma 2.7. Then:

(a) The set $S^{cc}=K^c$ is open, and

$$\mathscr{E}(S)=a(S)=S^{cc}=K^c. \tag{3.1}$$

Hence if $\xi\in\mathscr{S}'_A(\mathsf{M})$ with null region $N(\xi)\supset S$, then $N(\xi)\supset S^{cc}$.

(b) If π'_1 is not parallel with π''_2, then $W(\pi'_1,\pi''_2)\subset\mathscr{E}(S)$, and if π'_2 is not parallel with π''_1, then $W(\pi'_2,\pi''_1)\subset\mathscr{E}(S)$.

(c) If π'_1 is parallel with π''_2, or if π'_2 is parallel with π''_1, then $\mathscr{E}(S)=\mathsf{M}$.

Proof: It was shown in TW-I[8] that $\mathscr{E}(S)=S^{cc}$ with S defined as above, and the assertions in parts (b) and (c) were also proved. (See Proposition 5.6 in that paper.) By Theorem 2.11 we have $a(S)=\mathscr{E}(S)$, and hence the equalities in (3.1) obtain.

As was noted in TW-I,[8] it is *essential* that the wedges W_1 and W_2 have a nonempty intersection. If this is *not* the case, the conclusions can fail to hold. (See Examples 3.12 and 3.14 below.)

The above proposition has an obvious generalization to the case of any family $\mathscr{F}$ of wedges in $\mathscr{W}$ which satisfies the special condition that $W_1\cap W_2\neq\emptyset$ for *all* pairs $\{W_1,W_2\}$ of wedges in $\mathscr{F}$.

Proposition 3.4: Let $\mathscr{F}=\{W_\alpha=W(\pi'_\alpha,\pi''_\alpha)\,|\,\alpha\in\mathscr{T}\}$, where $\mathscr{T}$ is an index set, be a family of wedges in $\mathscr{W}$, such that $W_\alpha\cap W_\beta\neq\emptyset$ for *all* $\alpha,\beta\in\mathscr{T}$. For $\alpha,\beta\in\mathscr{T}$ we denote by $S_{\alpha\beta}$ the open region bounded below by π'_α and bounded above by π''_β. We define the sets S_0, S, and K by $S_0=\cup\{W_\alpha\,|\,\alpha\in\mathscr{T}\}$, $S=\cup\{S_{\alpha\beta}\,|\,\alpha,\beta\in\mathscr{T}\}$, and $K=S^c$. Then:

(a) The families $\Pi'=\{\pi'_\alpha\,|\,\alpha\in\mathscr{T}\}$ and $\Pi''=\{\pi''_\alpha\,|\,\alpha\in\mathscr{T}\}$ satisfy the premises in Theorem 3.2, and all the conclusions concerning S^{cc} and $\mathscr{E}(S)$ in that theorem apply. Furthermore,

$$a(S_0)=a(S)=\mathscr{E}(S_0)=\mathscr{E}(S)\subset(S_0^{cc})^o=(S^{cc})^o,\quad K=S^c=S_0^c,$$

and if $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)\supset S_0$, then $N(\xi)\supset\mathscr{E}(S)$.

(b) In particular we have $a(S_0)=S_0^{cc}$ if the family $\mathscr{F}$ contains at least two wedges with nonparallel upper planes, and at least two wedges with nonparallel lower planes.

Proof: (1) Let $\alpha,\beta\in\mathscr{T}$. By Proposition 3.3 we have $W(\pi'_\alpha,\pi''_\beta)=S_{\alpha\beta}\subset\mathscr{E}(W_\alpha\cup W_\beta)=(W_\alpha\cup W_\beta)^{cc}$ if π'_α is not parallel with π''_β, whereas we have $\mathscr{E}(W_\alpha\cup W_\beta)=(W_\alpha\cup W_\beta)^{cc}=\mathsf{M}$ if π'_α is parallel with π''_β. In both cases we have $S_{\alpha\beta}\subset\mathscr{E}(W_\alpha\cup W_\beta)\subset\mathscr{E}(S_0)$. Since this holds for all $\alpha,\beta\in\mathscr{T}$, it follows that $S\subset\mathscr{E}(S_0)$, and from this it follows that $\mathscr{E}(S)\subset\mathscr{E}(\mathscr{E}(S_0))=\mathscr{E}(S_0)$. We trivially have $\mathscr{E}(S)\supset\mathscr{E}(S_0)$ since $S\supset S_0$, and hence $\mathscr{E}(S_0)=\mathscr{E}(S)$. In view of Theorem 2.11 we then have $a(S_0)=\mathscr{E}(S_0)=\mathscr{E}(S)=a(S)$.

(2) We want to show that $S^c=S_0^c$, which is equivalent to the relation $S^{cc}=S_0^{cc}$. By Lemma 2.6 we have $R^{cc}=\mathscr{E}(R)^{cc}$ for any open $R\subset\mathsf{M}$. Hence $S_0^{cc}=\mathscr{E}(S_0)^{cc}=\mathscr{E}(S)^{cc}=S^{cc}$, and since $\mathscr{E}(S_0)=\mathscr{E}(S)$ we conclude that $S_0^{cc}=S^{cc}$. This completes the proof.

The point of this proposition is that the set S_0 is, in general, *smaller* than the set S, but nevertheless we have $S_0^{cc}=S^{cc}$ and $\mathscr{E}(S_0)=\mathscr{E}(S)$. We emphasize that the condition that $W_\alpha\cap W_\beta\neq\emptyset$, for all $\alpha,\beta\in J$, is *crucial*. If the family $\mathscr{F}$ contains a pair of wedges with an empty intersection, the conclusions may fail to hold.

We now consider a very special class of configurations of wedges to which Proposition 3.4 applies.

Proposition 3.5: Let $\mathscr{F}\subset\mathscr{W}$, and let $S_0=\cup\{W\,|\,W\in\mathscr{F}\}$ and $K=S_0^c=\cap\{W^c\,|\,W\in\mathscr{F}\}$.

(a) Suppose that all the *lower* planes of the wedges $W \in \mathscr{F}$ are *parallel*. Then the premises in Proposition 3.4 are satisfied, and the conclusions in part (a) of the proposition apply. Any one of the Cases I, II, IIIa, or IIIb in Theorem 3.2 can obtain. The set K is *not* in $\mathscr{K}_E$ if $K \neq \emptyset$.

(b) Suppose that all the *upper* planes of the wedges $W \in \mathscr{F}$ are *parallel*. Then the conclusions are the same as in part (a).

Proof: We only consider part (a) since very similar considerations apply to part (b). We first note the obvious fact that if the lower planes of two wedges in $\mathscr{W}$ are parallel, then the two wedges have a nonempty intersection. Since all the lower planes of the wedges in $\mathscr{F}$ are parallel it follows that the configuration is within the purview of Proposition 3.4. It is very easy to find examples which show that any one of the Cases I, II, IIIa, or IIIb can indeed obtain, and we feel that we can omit these details. (See, however, Example 3.10 below for some illustrations.)

We next consider a class of configurations of particular interest in quantum field theory. In the theorem and proposition which follow we depend on some geometric results in Section V of TW-I.[8]

Theorem 3.6: Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$. We write $S = \cup\{W | W \in \mathscr{F}\}$ and $K = \cap\{W^c | W \in \mathscr{F}\}$. Let $\mathscr{E}(S)$ be defined as in Lemma 2.6, and let the Araki envelope $a(S)$ be defined as in Lemma 2.7.

(a) Suppose that $\mathscr{F}$ contains two wedges in a *general relative position*. Then K is an element of the set $\mathscr{K}_E$ described in Definition 1.1, and we have

$$a(S) = \mathscr{E}(S) = S^{cc} = K^c. \tag{3.2}$$

If $\xi \in \mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi) \supset S$, then $N(\xi) \supset S^{cc}$.

(b) Suppose instead that the normals of the boundary planes of the wedges in $\mathscr{F}$ span M, and suppose furthermore that $\mathscr{F}$ contains two wedges such that no two of the boundary planes of these wedges are parallel. Then the conclusions are the same as in part (a).

Proof: Under the stated premises in part (a) or in part (b) we have $\mathscr{E}(S) = S^{cc} = K^c$ and $K \in \mathscr{K}_E$ by Theorem 5.7 in TW-I,[8] and by Theorem 2.11 in the present paper we have $a(S) = \mathscr{E}(S)$. This establishes the conclusions in the theorem.

The configurations in part (a) are obviously just special cases of the configurations in part (b), since the premise in part (a) trivially implies the premises in part (b). The assertions in part (a) are thus ''superfluous,'' but we wanted to state these assertions explicitly since they concern what can be regarded as the *simplest* case. The situation when two ''arbitrary'' wedges are *not* in a general relative position is ''exceptional'' in the sense that the normals of the boundary planes of the two wedges must then be linearly dependent. We might thus regard a family $\mathscr{F}$ of wedges as ''normal'' if it contains some pairs of wedges in a general relative position. Part (a) of the theorem then applies to such a family, and in particular we have $K \in \mathscr{K}_E$. However, it can happen that a family $\mathscr{F}$ contains *no* two wedges in a general relative position, although the family is such that $K \in \mathscr{K}_E$. It is worthwhile to consider a specific example here, similar to an example given in Section III in TW-I.[8]

Example 3.7: Let the index set $\mathscr{I}$ in the above theorem be the set of all unit vectors $\mathbf{u} \in \mathsf{R}^3$. Let $q > 0$. For each unit vector $\mathbf{u}$ we define the wedge $W_{\mathbf{u}}$ by $W_{\mathbf{u}} = \{(\mathbf{x},t) | \mathbf{x}\cdot\mathbf{u} - q > |t|\}$, and we define the family $\mathscr{F}$ by $\mathscr{F} = \{W_{\mathbf{u}} | \mathbf{u} \in \mathscr{I}\}$. We have $K = \cap\{W_{\mathbf{u}}^c | \mathbf{u} \in \mathscr{I}\} = D((\mathbf{0},-q),(\mathbf{0},q)) \in \mathscr{D} \subset \mathscr{K}_C \subset \mathscr{K}_E$, but $\mathscr{F}$ contains no two wedges in a general relative position. The premises in part (b) of the theorem are, however, satisfied, and hence the equalities in (3.2) hold.

This example shows that an ''abnormal'' family need not be particularly ''pathological'' at all. Part (b) of the theorem is intended to deal with some configurations of this kind. The proposition which follows seemingly deals with configurations for which the premises in part (b) of Theorem 3.6 might not be satisfied. However, by Proposition 5.8 in TW-I[8] these premises are *in fact* satisfied. Hence we have the following result.

Proposition 3.8: Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$. We write $S = \cup\{W | W \in \mathscr{F}\}$ and $K = \cap\{W^c | W \in \mathscr{F}\}$, and we assume that $K \neq \emptyset$ and that S is *connected*.

(a) Suppose that $K \in \mathscr{K}_E$. Then the family $\mathscr{F}$ satisfies the premises in part (b) of Theorem 3.6, and hence $a(S) = \mathscr{E}(S) = S^{cc} = K^c$.

(b) Suppose instead that $\mathscr{F}$ contains at least two wedges with nonparallel lower planes, and at least two wedges with nonparallel upper planes. Suppose furthermore that the normals of the

boundary planes of the wedges in $\mathscr{F}$ span M. Then $K \in \mathscr{K}_E$ and the conclusion is as in part (a).

We will not repeat here the proofs of Theorem 5.7 and Proposition 5.8 in TW-I,[8] but some comments about what is involved are in order. The *nontrivial* issue in Theorem 3.6 (and in Proposition 3.8) is the relation $\mathscr{E}(S) = S^{cc}$, since the relation $a(S) = \mathscr{E}(S)$ follows at once from Theorem 2.11. We consider Theorem 3.6. Corresponding to the family $\mathscr{F}$ we define the set $S \subset \mathsf{M}$ by $S = \cup\{W | W \in \mathscr{F}\}$, as in the theorem. We define a new family $\mathscr{F}^*$ of wedges in $\mathscr{W}$ by $\mathscr{F}^* = \{W | W \in \mathscr{W}, W \subset \mathscr{E}(S)\}$ and we write $S_e = \cup\{W | W \in \mathscr{F}^*\}$. We clearly have $\mathscr{F}^* \supset \mathscr{F}$ and $\mathscr{E}(S) \supset S_e \supset S$. It follows readily that $\mathscr{E}(S) = \mathscr{E}(S_e)$ and $\mathscr{E}(S)^{cc} = S_e^{cc} = S^{cc}$. We also easily see that the ''enlarged'' family $\mathscr{F}^*$ satisfies the same premises as the family $\mathscr{F}$. What is not quite so obvious is that $\mathscr{F}^*$ always contains at least two wedges in a general relative position, even in the case when the ''original'' family $\mathscr{F}$ does not have this property. (See the proof of Theorem 5.7 in TW-I[8]).

Given any family $\mathscr{F}$ of wedges, the crux of the matter is really whether the ''enlarged'' family $\mathscr{F}^*$, defined in terms of $\mathscr{F}$ as above, contains two wedges in a general relative position. It is desirable to find criteria for this to be the case in terms of reasonably transparent conditions on the ''primary family'' $\mathscr{F}$. Such conditions are given in part (b) of Theorem 3.6 and in Proposition 3.8.

In Theorem 3.6 and in Proposition 3.8 the family $\mathscr{F}$ is such that the set $K = \cap\{W^c | W \in \mathscr{F}\}$ is an element of the (well-behaved) family $\mathscr{K}_E$ described in Definition 1.1. One aspect of this is that the set of all normals of the boundary planes of the wedges in $\mathscr{F}$ span M. Proposition 3.4 also concerns a family $\mathscr{F}$ of wedges. It is not required in the premises of this proposition that the normals span M, but instead it is required that $W_\alpha \cap W_\beta \neq \varnothing$ for *every* pair $\{W_\alpha, W_\beta\} \subset \mathscr{F}$. It can well happen that the normals of all the upper and lower planes span a *three*-dimensional or *two*-dimensional subspace of M, and in this case K is *not* an element of $\mathscr{K}_E$. The premises are thus quite different in Proposition 3.4 and Theorem 3.6, although it can happen that a particular family $\mathscr{F}$ satisfies the premises of both.

What is, perhaps, not so obvious at this point is to what extent these premises are necessary for the stated conclusions in Theorem 3.6 or in Propositions 3.3 or 3.8 about the null regions of Araki distributions $\xi(x)$, namely that $N(\xi) \supset (S^{cc})^o$ if $N(\xi) \supset S$. This question has two aspects. One is the purely geometric problem of when $\mathscr{E}(S) = (S^{cc})^o$, and the other is within the purview of distribution theory and concerns the question of whether *every* distribution $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$ which vanishes on some open set S might in fact vanish on a *larger* open set than $a(S)$. We will discuss the second question in Section IV. In the present section we will illustrate the *geometric* aspects of the issue with some examples, but before we present these examples we want to draw some simple but important conclusions from Theorem 3.6. The proposition which follows will be of particular interest in our discussion of nets of local algebras in a subsequent paper. In Section V we will consider its implications in quantum field theory.

Proposition 3.9: (a) Let $K_1 \in \mathscr{K}_E$, and let $\mathscr{F}_2 \subset \mathscr{W}$ be any family of wedges. We write $K_2 = \cap\{W^c | W \in \mathscr{F}_2\}$ and $S_2 = \cup\{W | W \in \mathscr{F}_2\}$. Then $K_1 \cap K_2 \in \mathscr{K}_E$, and $a(K_1^c \cup S_2) = \mathscr{E}(K_1^c \cup S_2) = (K_1 \cap K_2)^c = (K_1^c \cup S_2)^{cc}$.

(b) Let $K_1 \in \mathscr{K}_E$, and let K_2 be a closed, convex, causally complete subset of M, such that K_2^c has a nonempty interior $(K_2^c)^o = S_2$. Then $K_1 \cap K_2 \in \mathscr{K}_E$, and $a(K_1^c \cup S_2) = \mathscr{E}(K_1^c \cup S_2) = (K_1^c \cup S_2)^{cc} = (K_1 \cap K_2)^c$.

(c) Let $\mathscr{K}_S \subset \mathscr{K}_E$, $\mathscr{K}_S \neq \varnothing$, and let $S = \cup\{K^c | K \in \mathscr{K}_S\}$ and $K_S = \cap\{K | K \in \mathscr{K}_S\} = S^c$. Then $K_S \in \mathscr{K}_E$ and $a(S) = \mathscr{E}(S) = S^{cc} = K_S^c$.

Proof: (1) We consider part (a), and we write $\mathscr{F}_1 = \{W | W \in \mathscr{W}, W^c \supset K_1\}$ and $\mathscr{F} = \mathscr{F}_1 \cup \mathscr{F}_2$. By Theorem 3.1 we have $K_1 = \cap\{W^c | W \in \mathscr{F}_1\}$ and $K^c = \cup\{W | W \in \mathscr{F}_1\}$. Hence $K_1 \cap K_2 = \cap\{W^c | W \in \mathscr{F}\}$ and $\mathscr{E}(K_1^c \cup K_2^c) \supset \cup\{W | W \in \mathscr{F}\}$. By the definition of $\mathscr{K}_E$ the set $\mathscr{F}_1$ contains two wedges in a general relative position, and hence this is also the case for the set $\mathscr{F}$. The set $\mathscr{F}$ thus satisfies the premise in part (a) of Theorem 3.6, and the conclusion follows at once.

(2) We consider part (b). By Theorem 3.1 we have $S_2^c = ((K_2^c)^o)^c = K_2$. The conclusion in part (b) then follows from the result in part (a) when we select $\mathscr{F}_2 = \{W | W \in \mathscr{W}, W^c \supset K_2\}$, in which case we have $S_2 = \cup\{W | W \in \mathscr{W}, W^c \supset K_2\}$.

(3) To prove part (c) we define the family $\mathscr{F} \subset \mathscr{W}$ by $\mathscr{F} = \cup\{\{W | W \in \mathscr{W}, W \supset K\} | K \in \mathscr{K}_S\}$. Let $K_1 \in \mathscr{K}_S \subset \mathscr{K}_E$. We then have $\{W | W \in \mathscr{W}, W \supset K_1\} \subset \mathscr{F}$, and it follows that $\mathscr{F}$ contains two wedges in a general relative position. The conclusion then follows from Theorem 3.6.

Note that the set K_2 in part (a) is not necessarily in $\mathscr{K}_E$. The simplest case in part (c) is the

case of just two sets $K_1, K_2 \in \mathscr{K}_E$. (Part (b) also applies to this case.) We then have $K = K_1 \cap K_2 \in \mathscr{K}_E$ and $a(S) = a(K_1^c \cup K_2^c) = (K_1^c \cup K_2^c)^{cc}$. If $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi) \supset K_1^c \cup K_2^c$, we thus have $N(\xi) \supset (K_1^c \cup K_2^c)^{cc}$. The significance of this result is that $(K_1^c \cup K_2^c)^{cc}$ is in general *larger* (and in some cases ''very much larger'') than $K_1^c \cup K_2^c$.

We now discuss some examples. At this point the reader may want to go ahead of the discussion and look at Examples 4.8 and 4.9 in the next section.

Example 3.10: Let π' be a characteristic plane, and let Π'' be a family of characteristic planes not parallel with π'. We consider the family $\mathscr{F} \subset \mathscr{W}$ defined by $\mathscr{F} = \{W(\pi', \pi'') \,|\, \pi'' \in \Pi''\}$, and we write $S = \cup\{W \,|\, W \in \mathscr{F}\}$ and $K = S^c$. We then have $K = \cap\{\overline{W}(\pi'', \pi') \,|\, \pi'' \in \Pi''\}$. The set K is *not* in $\mathscr{K}_E$, since $\mathscr{F}$ contains no two wedges with nonparallel lower planes. The span of the normals of the planes in $\Pi'' \cup \{\pi'\}$ can be two-, three-, or four-dimensional. The set S is *connected* but the premises in Theorem 3.6 or in Proposition 3.8 are not satisfied. Proposition 3.5 is obviously applicable. In this kind of example we have just two possibilities: I) $K \neq \varnothing$ and $a(S) = \mathscr{E}(S) = (S^{cc})^o$. This case obtains in particular if the family Π'' is *finite*. II) $K = \varnothing$, $S^{cc} = \mathsf{M}$, but $\mathscr{E}(S)$ is the open half-space bounded below by π'. An example of Case II is the configuration when Π'' consists of *all* characteristic planes which are not parallel with π'.

Similar considerations apply to any family $\mathscr{F}$ of wedges with a common *upper* plane π''. With $S = \cup\{W \,|\, W \in \mathscr{F}\}$ we then have two possibilities: I) $K = S^c \neq \varnothing$ and $a(S) = \mathscr{E}(S) = (S^{cc})^o$. II) $K = S^c = \varnothing$, $S^{cc} = \mathsf{M}$, but $\mathscr{E}(S)$ is the open half-space bounded above by π''.

Example 3.11: Let π' and π'' be two parallel characteristic planes such that π' lies *above* π''. Let $\Pi_u{}'$ and Π_l' be two (nonempty) sets of characteristic planes which are *not* parallel with π'' (and π'). We define the family $\mathscr{F}_u$ of wedges by $\mathscr{F}_u = \{W(\pi', \pi_u'') \,|\, \pi_u'' \in \Pi_u{}'\}$, and similarly we define the family $\mathscr{F}_l$ by $\mathscr{F}_l = \{W(\pi_l', \pi'') \,|\, \pi_l' \in \Pi_l'\}$. We write $S_u = \cup\{W \,|\, W \in \mathscr{F}_u\}$, $K_u = S_u^c$, $S_l = \cup\{W \,|\, W \in \mathscr{F}_l\}$, and $K_l = S_l^c$. Proposition 3.5 applies to each one of the families $\mathscr{F}_u$ and $\mathscr{F}_l$ separately, and the family $\mathscr{F}_u$ is as described in Example 3.10. The set $\mathscr{E}(S_u)$ lies *above* the plane π' and the set $\mathscr{E}(S_l)$ lies *below* the plane π''.

Let $\mathscr{F} = \mathscr{F}_u \cup \mathscr{F}_l$ and $S = \cup\{W \,|\, W \in \mathscr{F}\} = S_u \cup S_l$. The set $K = S^c = K_u \cap K_l$ is contained in the closed ''slab-region'' between the planes π' and π''. The sets S_u and S_l are disjoint, and hence S is not connected. The set S^{cc} is connected if and only if S_u contains a point x'' and S_l contains a point x' such that the interval $[x', x'']$ is timelike, which is the case if and only if S_u is *not* included in S_l^c. If $S_u \subset S_l^c$, then $S^{cc} = S_u^{cc} \cup S_l^{cc}$. The sets $\mathscr{E}(S_u)$ and $\mathscr{E}(S_l)$ are also disjoint, and we have $\mathscr{E}(S) = \mathscr{E}(S_u) \cup \mathscr{E}(S_l)$. Hence we always have $\mathscr{E}(S) \neq (S^{cc})^o$ if S^{cc} is connected.

Proposition 3.4 is not applicable, since $\mathscr{F}$ contains pairs of wedges with an empty intersection. The span of the normals of the boundary planes of the wedges in $\mathscr{F}$ can be two-, three-, or four-dimensional. Every pair of wedges in the family $\mathscr{F}$ is such that one of the boundary planes of one of the wedges is parallel with one of the boundary planes of the other, and in particular $\mathscr{F}$ contains no two wedges in a general relative position. Theorem 3.6 is accordingly not applicable. Proposition 3.8 is not applicable since S is not connected.

As described in Example 3.10 we have two main cases for each one of the sets S_u and S_l. If either $S_u^{cc} = \mathsf{M}$ or $S_l^{cc} = \mathsf{M}$ we have $S^{cc} = \mathsf{M} \neq \mathscr{E}(S)$. If $S_u^{cc} \neq \mathsf{M}$ *and* $S_l^{cc} \neq \mathsf{M}$, we may, or may not, have $\mathscr{E}(S) = (S^{cc})^o$ depending on whether $S_u \subset S_l^c$ or not.

We consider some special configurations within the general class described in Example 3.11.

Example 3.12: (a) Suppose that Π_u'' consists of a single plane π_1'', and that Π_l'' also consists of a single plane π_1', parallel with π_1''. Hence $S_u = W(\pi', \pi_1'') = \mathscr{E}(S_u)$ and $S_l = W(\pi_1', \pi'') = \mathscr{E}(S_l)$. This configuration is ''two-dimensional'' in the sense that the span of the normals of the boundary planes of the wedges in $\mathscr{F}$ is two-dimensional. It is invariant under all translations in a two-dimensional group of spacelike translations. We have $K = S^c = W(\pi', \pi_1'')^c \cap W(\pi_1', \pi'')^c$ and $\mathscr{E}(S) = S$. If π_1'' lies *above* π_1', then $K = \varnothing$ and $S^{cc} = \mathsf{M}$. If π_1'' does *not* lie above π_1', then $K \neq \varnothing$ and $S^{cc} = S = \mathscr{E}(S)$. Proposition 3.3 is not applicable since $W(\pi', \pi_1'') \cap W(\pi_1', \pi'') = \varnothing$.

(b) Suppose that $\Pi_u{}'$ consists of a single plane π_1'', and that $\Pi_l{}'$ also consists of a single plane π_1', *not* parallel with π_1''. This configuration is ''three-dimensional'' in an obvious sense. We have $S = \mathscr{E}(S) = W(\pi', \pi_1'') \cup W(\pi_1', \pi'') \neq S^{cc}$ and $K = W(\pi', \pi_1'')^c \cap W(\pi_1', \pi'')^c \neq \varnothing$. The set S^{cc} is *connected*, and we have $W(\pi_1', \pi_1'') \subset S^{cc}$. Proposition 3.3 is not applicable since $W(\pi', \pi_1'') \cap W(\pi_1', \pi'') = \varnothing$.

Example 3.13: Let the premises and notation be as in Example 3.11. Let $x' \in \pi'$ and x''

$\in \pi''$ be two points, such that $x' \in V_+(x'')$. Hence the closed double cone $D(x'',x')$ lies in the closed ''slab-region'' R between the planes π' and π''. Let Π'_l consist of all characteristic planes which pass through x' and which are not parallel with π' (and π''), and let Π''_u consist of all characteristic planes which pass through x'' and which are not parallel with π'. We then have $K=D(x'',x') \in \mathscr{D} \subset \mathscr{K}_E$, but Proposition 3.8 is not applicable since S is not connected. The set $S^{cc}=K^c=(S^{cc})^o$ is connected, but $\mathscr{E}(S)$ is not connected, and hence $\mathscr{E}(S) \neq (S^{cc})^o$.

We consider a pair of wedges $W_1=W(\pi'_1,\pi''_1)$ and $W_2=W(\pi'_2,\pi''_2)$ in $\mathscr{W}$. Let $u'_j=(\mathbf{u}'_j,1)$ and $u''_j=(\mathbf{u}''_j,1)$ be the normals of the boundary planes in the standard form. Each one of the intersections $W_1 \cap W_2$ and $W_1^c \cap W_2^c$ can be empty or nonempty, and we distinguish between four kinds of configurations which can occur on this basis. It is not hard to show (see Proposition 5.1 in TW-I[8]) that at least one of these intersections is empty if and only if there exist two numbers $\theta', \theta'' \in [0,1]$ such that $\theta' u'_1+(1-\theta')u'_2=\theta'' u''_1+(1-\theta'')u''_2$. If the normals span M, then both intersections are nonempty, and Proposition 3.3 is applicable. Suppose that the normals span a *three*-dimensional space. One easily sees that at least one of the intersections $W_1 \cap W_2$ and $W_1^c \cap W_2^c$ must then be nonempty. In Example 3.12 we have examples of configurations for which $W_1 \cap W_2=\varnothing$ but $W_1^c \cap W_2^c \neq \varnothing$. In these examples the lower plane of one of the wedges is parallel with and lies above the upper plane of the other. It is of interest to present an example in which no two boundary planes are parallel. Such a configuration is presented in Example 4.17 in the next section, but we here give another example which is simpler to analyze.

Example 3.14: We consider four characteristic planes defined by

$$\pi'_1=\{x|t\sqrt{2}=-x_2-x_3+1\}, \quad \pi''_1=\{x|t\sqrt{2}=-x_2+x_3-1\},$$

$$\pi'_2=\{x|t\sqrt{2}=x_2+x_3+1\}, \quad \pi''_2=\{x|t\sqrt{2}=x_2-x_3-1\},$$

and we write $W_1=W(\pi'_1,\pi''_1)$, $W_2=W(\pi'_2,\pi''_2)$, $S=W_1 \cup W_2$, and $K=S^c=W_1^c \cap W_2^c$. It is trivial that $0 \in K$, and hence $K \neq \varnothing$. By the above criteria we then conclude that $W_1 \cap W_2=\varnothing$, as one can also easily see directly by an examination of the relevant inequalities implied by the equations for the four planes. We have $W_1 \cup W_2 \cup W(\pi'_1,\pi''_2) \cup W(\pi'_2,\pi''_1) \subset S^{cc}$, but $\mathscr{E}(S)=S \neq S^{cc}$. The span of the normals of the four planes is trivially three-dimensional. No two of the four planes are parallel. Proposition 3.3 is not applicable since $W_1 \cap W_2=\varnothing$, and hence we cannot conclude that $W(\pi'_1,\pi''_2) \cup W(\pi'_2,\pi''_1) \subset N(\xi)$, given that $N(\xi) \supset W_1 \cup W_2$ for a distribution $\xi(x) \in \mathscr{S}'_A(\mathsf{M})$. As shown in Example 4.17 there exist, in fact, distributions ξ such that $N(\xi) \supset W_1 \cup W_2$ but such that $W(\pi'_1,\pi''_2) \cup W(\pi'_2,\pi''_1)$ is *not* included in $N(\xi)$.

We note that W_1 and W_2 are not separated by any *characteristic* plane. To see this we consider the edges $E_1=\{x|-x_2+x_3-1=t\sqrt{2}=-x_2-x_3+1\}$ and $E_2=\{x|x_2-x_3-1=t\sqrt{2}=x_2+x_3+1\}$ of the wedges W_1, respectively W_2. If a hyperplane π separates W_1 from W_2, then π must contain a linear 2-manifold parallel with E_1, and a linear 2-manifold parallel with E_2. We easily see that π must then contain a *timelike* line, and hence π cannot be a characteristic hyperplane.

We also note the trivial fact that if l is any *isotropic* line, then l necessarily contains a half-line which is *not* in $W_1 \cup W_2$. The same is true for the configuration in Example 3.13.

IV. ABOUT THE KLEIN–GORDON EQUATION AND A CLASS OF SPECIAL ARAKI DISTRIBUTIONS

In this section we consider (among other things) *tempered* distributions $\psi(x) \in \mathscr{S}'(\mathsf{M})$ which satisfy the *Klein–Gordon equation*

$$\frac{\partial^2 \psi(x)}{\partial t^2}-\Delta \psi(x)+m^2\psi(x)=0. \tag{4.1}$$

Here $m \geqslant 0$, and we thus include the case of the *wave equation* corresponding to $m=0$. We denote the set of tempered distributions which satisfy the equation (4.1) for a particular m by Ψ_m. Every such distribution is trivially in the Araki-class $\mathscr{S}'_A(\mathsf{M})$. Every ''ordinary'' C_2-solution $\psi(x)$ of the Klein–Gordon equation which satisfies certain asymptotic growth conditions can also be regarded as a distribution of this kind if we define $\psi[f]$ for any test function $f(x)$ by

$$\psi[f]=\int_{(\infty)} d^4(x)\ f(x)\psi(x). \tag{4.2}$$

Specifically a C_2-solution $\psi(x)$ of (4.1) is in $\mathscr{S}'_A(\mathsf{M})$ if and only if $|\psi(x)|<C(1+|\mathbf{x}|+|t|)^n$ for some constants C and n.

Suppose that the complex-valued functions $g_+(\mathbf{p})$ and $g_-(\mathbf{p})$ on R^3 are elements of the Hilbert space $L_2(\mathsf{R}^3)$. Then the integral at right in

$$\psi(x)=\frac{1}{(2\pi)^{3/2}}\int_{(\infty)}\frac{d^3(\mathbf{p})}{\sqrt{2\omega}}(g_+(\mathbf{p})e^{-ix\cdot p}+g_-(\mathbf{p})e^{ix\cdot p}) \tag{4.3}$$

where $\omega=\omega(\mathbf{p})=\sqrt{m^2+|\mathbf{p}|^2}$ and $p=(\mathbf{p},\omega)$, makes sense and defines a solution $\psi(x)$ of (4.1) in $\mathscr{S}'_A(\mathsf{M})$. Solutions of this kind are of interest in quantum field theory, *and in what follows we denote this class of solutions by* Ψ^H_m. If $\psi(x)$ is a C_2-solution of (4.1) such that $\psi(x)$ and its time derivative $\partial_t\psi(x)$ restricted to the spacelike 3-plane $\sigma_0=\{(\mathbf{x},0)|\mathbf{x}\in\mathsf{R}^3\}$ are functions of compact support, then $\psi(x)\in\Psi^H_m$. If $\psi(x)$ is a C_2-solution of (4.1) such that $\psi(x)$ and its time derivative $\partial_t\psi(x)$ restricted to the spacelike 3-plane σ_0 are in the test function space $\mathscr{S}(\mathsf{R}^3)$, then $\psi(x)\in\Psi^H_m$, and $\psi(x)$ is infinitely differentiable.

All the results in Section III apply as such to the elements of $\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$. Moreover, if $\psi(x)\in\Psi_m$, with null region $N(\psi)$, then $N(\psi)\supset\mathscr{E}(S)$ whenever $N(\psi)\supset S$, and we have $\mathscr{E}(N(\psi))=N(\psi)$. As we remarked earlier, these ''classical'' principles *do not* in general apply to the elements of $\mathscr{S}'_A(\mathsf{M})$. The following is a statement within the framework of our paper of a well-known result concerning the ''region of dependence.''

Proposition 4.1: Let $\psi(x)$ be a C_2-solution of (4.1), with null region $N(\psi)$. Let σ be a spacelike plane, and let K_σ be a closed subset of σ which contains the supports of the restrictions of $\psi(x)$ and its normal derivative to σ. Let S_σ be the relative complement of K_σ in σ. We write $K=K^{cc}_\sigma$ and $S=S^{cc}_\sigma$. Then:

(a) S is *open* and K is *closed*, and we have $S^c=K$ and $N(\psi)\supset S=K^c$.

(b) If $W\in\mathscr{W}$ and $W^c\supset K_\sigma$, then $N(\psi)\supset W$.

About the Proof: The statement that $N(\psi)\supset S^{cc}_\sigma$ is the ''classical'' result.[11] The remaining assertions are of a geometric character, and we omit the simple proofs.

The above can be combined with our Proposition 3.9 as follows.

Proposition 4.2: Let $\mathscr{T}\subset\mathscr{W}$ be a family of wedges. We write $K=\cap\{W^c|W\in\mathscr{T}\}$ and $S=\cup\{W|W\in\mathscr{T}\}$. Let $\psi(x)$ be a C_2-solution of (4.1) such that the supports of the restrictions of $\psi(x)$ and its normal derivative to some spacelike plane σ are *compact*. Suppose furthermore that the null region $N(\psi)$ of $\psi(x)$ contains S. Then $N(\psi)\supset K^c=S^{cc}$.

Proof: Let K_σ be a closed subset of σ which contains the supports of the restrictions of $\psi(x)$ and its normal derivative to σ. There then exists a $K_1\in\mathscr{K}_E$ such that $K_1\supset K_\sigma$ and hence $K_1=K^{cc}_1\supset K^{cc}_\sigma$. By Proposition 4.1 we have $N(\psi)\supset K^c_1$ and it then follows by Proposition 3.9 that $N(\psi)\supset a(K^c_1\cup K^c)=(K^c_1\cup K^c)^{cc}\supset K^c$, which proves the assertion.

In Proposition 4.1 the ''initial-value surface'' σ is a spacelike 3-plane. It is well known that a solution of (4.1) can be uniquely specified by the initial values on other kinds of surfaces, and we here give an example of this.

Proposition 4.3: Let Σ be a hypersurface in M defined by $\Sigma=\{(\mathbf{x},g(\mathbf{x}))|\mathbf{x}\in\mathsf{R}^3\}$, where $g(\mathbf{x})$ is a C_2-function on R^3 which satisfies the condition $|g(\mathbf{x})-g(\mathbf{y})|\leqslant\theta|\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x},\mathbf{y}\in\mathsf{R}^3$, for some fixed $\theta\in(0,1)$. Let K_0 be a closed subset of Σ, and let Σ_0 be its relative complement (within Σ). Let $\psi(x)$ be a C_2-solution of (4.1), such that the supports of the restrictions of $\psi(x)$ and its gradient to Σ are contained in K_0. Then Σ^{cc}_0 is open, and $N(\psi)\supset\Sigma^{cc}_0=K^c_0$, where $N(\psi)$ is the null region of $\psi(x)$.

About the Proof: The statement that $N(\psi)\supset\Sigma^{cc}_0$ is again a ''classical'' result. That Σ^{cc}_0 is open and that $\Sigma^{cc}_0=K^c_0$ was proved in Proposition 2.8 in TW-I.[8]

If $\psi(x)$ is a C_2-solution of the Klein–Gordon equation, and if $S\subset N(\psi)$, then we obviously have $(\overline{S})^o\subset N(\psi)$, and it follows that the null region must satisfy the condition $(\overline{N(\psi)})^o=N(\psi)$. Similar considerations apply to the initial values of $\psi(x)$ on some suitable spacelike surface, and we have, in particular, the following result. Let $W\in\mathscr{W}$, and suppose that the C_2-solution $\psi(x)$ of

(4.1) satisfies the condition that $N(\psi)\supset W\cup(W^c)^o$. Then $\psi(x)=0$ on M. To see this we note that the initial values of $\psi(x)$ vanish identically on any spacelike 3-plane which contains the edge $\overline{W}\cup W^c$ of the wedge W. These properties of C_2-solutions are naturally not shared with more discontinuous distribution solutions of (4.1).

It is not our intention to try to give a complete description of the null regions of solutions of the Klein–Gordon equation here. As far as we know such a description is actually not known at present. As we will see, the matter cannot be settled by the above-mentioned ''classical'' principles alone. We will, however, consider the theory of Sections II and III in the light of known properties of the solutions of equation (4.1). We first note that for any $m>0$ and any point $y\in\mathsf{M}$ there exist solutions whose support (in x-space) is precisely the closed lightcone $\overline{V}(y)$. We consider the ''causal function'' $\Delta_0(x-y;m)$, familiar from quantum field theory, which is proportional to the vacuum expectation value $\langle\Omega|[\varphi(x),\varphi(y)]\Omega\rangle$ of the commutator of two free Hermitian scalar field operators. We regard y as a fixed ''parameter,'' and write $\zeta_y(x)=\Delta_0(x-y;m)$. The ''function'' $\zeta_y(x)$ is a tempered distribution in $\mathscr{S}'_A(\mathsf{M})$ which satisfies equation (4.1). We have the (symbolic) integral representation

$$\zeta_y(x)=\Delta_0(x-y;m)=\frac{1}{(2\pi)^3}\int_{(\infty)}d^3(\mathbf{p})\,\frac{\sin(\omega(t-s))}{\omega}\,\exp(i(\mathbf{x}-\mathbf{y})\cdot\mathbf{p}) \tag{4.4}$$

with $x=(\mathbf{x},t)$, $y=(\mathbf{y},s)$, $p=(\mathbf{p},\omega)$, and $\omega=\sqrt{m^2+|\mathbf{p}|^2}$.

When $m>0$ the support of $\zeta_y(x)$ is precisely the closed lightcone $\overline{V}(y)$, and when $m=0$ the support is precisely $\partial V(y)$. From this we can now draw a simple conclusion, which is worth a formal statement.

Proposition 4.4: (a) Let S be an open subset of M, and let $\nu(S)$ be the *largest* open set such that $N(\xi)\supset\nu(S)$ for *every* $\xi\in\mathscr{S}'_A(\mathsf{M})$ for which $N(\xi)\supset S$. Then $(S^{cc})^o\supset\nu(S)\supset a(S)$.

(b) Under the premises in Theorem 3.6, or in Propositions 3.3 or 3.8, we have $S^{cc}=\nu(S)=\mathscr{E}(S)=a(S)$.

Proof: Let S be an open subset of M, and suppose that $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)\supset S$. For any point $y\in S^c$ it might then happen that the support of $\xi(x)$ is equal to the light cone $\overline{V}(y)$, as would be the case if $\xi(x)=\zeta_y(x)$, with $\zeta_y(x)$ as in (4.4) for some $m>0$. The set $\cup\{\overline{V}(y)|y\in S^c\}$ is the (set) complement of the causal completion S^{cc} of S, and hence the largest possible *common* null region $\nu(S)$ for the set of *all* $\xi(x)$ which vanish on S cannot exceed $(S^{cc})^o$.

With the premises in Theorem 3.6, or in Propositions 3.3 or 3.8, we have the conclusion that $S^{cc}=\mathscr{E}(S)=a(S)$, and hence we have the conclusion in part (b) above.

The results in Theorem 3.6, or in Propositions 3.3 or 3.8, are thus optimal in the sense that S^{cc} is *precisely* the common null region $\nu(S)$ of all $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$ which vanish on S: no ''improvement'' is possible.

Example 4.5: The distribution $\zeta_y(x)$ considered above is not in the class Ψ^H_m. Since it is of interest to see what one can say about the support of functions in this class, we consider the following. Let S be an open subset of M. The causal complement S^c is then closed. We consider the special case when S^c is equal to the closure of its interior. Let σ be a spacelike 3-plane which intersects the interior of S^c, and let $B\subset\sigma\cap(S^c)^o$ be the base of a double cone $D\in\mathscr{D}$. Hence $D\subset S^c$. Let g be a C_∞-function on σ with support equal to B. For any particular $m\geqslant 0$ there then exists a (unique) C_∞-solution $\psi_{g,m}(x)$ of the Klein–Gordon equation (4.1) such that the restriction of $\psi_{g,m}(x)$ to σ agrees with the function g, and such that the normal derivative (relative to σ) of $\psi_{g,m}(x)$ vanishes on σ. This solution is in Ψ^H_m, and it has the representation in (4.3). By the construction we have $N(\psi_{g,m})\supset D^c\supset S^c$, and its support includes B. From what has been said we can then conclude that $(S^{cc})^o$ is the largest possible *common* null region of *all* functions $\psi(x)\in\Psi^H_m$ which vanish on S.

For an open $S\subset\mathsf{M}$ it can happen that $K=S^c$ has an empty interior. The case when S^c is a subset of a characteristic plane with a nonempty relative interior is of interest.

Example 4.6: Let $\pi=\{x|x_3=t\}$, and let K be the subset of this characteristic plane defined by $K=\{x|\tau\geqslant x_3=t\geqslant-\tau, x_1^2+x_2^2\leqslant\rho^2\}$, with $\tau>0$ and $\rho>0$. We then have $K\in\mathscr{K}_C$, i.e., K is compact, convex, and causally complete, and the set $S=K^c$ is *open*. We want to construct a function

$\psi(x) \in \Psi_m^H$, for $m>0$, such that $\psi(x)=0$ on S. The idea is to ''average'' the distribution $\zeta_y(x)$ in (4.4) with a suitable function $f(y)$ on K, and we thus consider the representation of $\psi(x)$ by an integral of the form

$$\psi(x)=\int_K d_3(y)\Delta_0(x-y;m)f_1(s)f_2(y_1,y_2) \tag{4.5}$$

with $y=(y_1,y_2,s,s)$ and $d_3(y)=dy_1dy_2ds$. We select $f_2(y_1,y_2)$ as an infinitely differentiable function on R^2 with support in the closed ball $\{(y_1,y_2)|y_1^2+y_2^2\leqslant\rho^2\}$. To achieve our goal we select $f_1(s)$ as an infinitely differentiable function on R^1 with support in the interval $[-\tau,\tau]$, and with the additional property that the integral of $f_1(s)$ over the interval vanishes. We thus express $f_1(s)$ as the derivative

$$f_1(s)=\frac{df_0(s)}{ds} \tag{4.6}$$

of an infinitely differentiable function $f_0(s)$ on R^1 with support in $[-\tau,\tau]$. Although this is not essential for the construction, we will assume that both $f_2(y_1,y_2)$ and $f_0(s)$ are *real* functions.

For any function $h(\mathbf{x})$ on R^n we define its Fourier transform $\hat{h}(\mathbf{p})$ by

$$\hat{h}(\mathbf{p})=\frac{1}{(2\pi)^{n/2}}\int_{(\infty)} d^n(\mathbf{x})h(\mathbf{x})\ \exp(-i\mathbf{x}\cdot\mathbf{p}). \tag{4.7}$$

By (4.6) we have $\hat{f}_1(q)=iq\hat{f}_0(q)$, and since the functions f_0 and f_2 are real, we have $\hat{f}_0(q)^*=\hat{f}_0(-q)$ and $\hat{f}_2(p_1,p_2)^*=\hat{f}_2(-p_1,-p_2)$. By a simple computation, based on the integral representation in (4.4), we then find that the function $\psi(x)$ in (4.5) is of the form in (4.3) with $g_+(\mathbf{p})$ and $g_-(\mathbf{p})$ given by

$$g_+(\mathbf{p})=\frac{(\omega-p_3)}{\sqrt{2\omega}}\hat{f}_0(p_3-\omega)\hat{f}_2(p_1,p_2), \quad g_-(\mathbf{p})=g_+(\mathbf{p})^*.$$

We have $\hat{f}_0(q)\in\mathscr{S}(\mathsf{R}^1)$ and $\hat{f}_2(p_1,p_2)\in\mathscr{S}(\mathsf{R}^2)$, and it is easily seen that $g_-(\mathbf{p}),g_+(\mathbf{p})\in L_2(\mathsf{R}^3)$. Hence $\psi(x)\in\Psi_m^H$, and the function $\psi(x)$ is *real* as a consequence of the reality of the functions f_0 and f_2. By the construction we have $\psi(x)=0$ on $S=K^c$. This example has implications for quantum field theory, as we will discuss in Section V. It also has implications for the theory of nets of local algebras, as we will discuss in a subsequent paper. It implies that nontrivial algebras of observables can be associated with a set such as K, with an *empty* interior, and such sets can therefore not be dismissed as ''unphysical,'' although one may at first be tempted to do so.

We next consider an example which involves the issues in Lemma 2.7 and Theorem 2.11.

Example 4.7: We want to construct an Araki distribution $\xi(x)$ from a solution $\phi(x,s)$ of the Klein–Gordon equation

$$\frac{\partial^2\phi(x,s)}{\partial t^2}-\sum_{\alpha=1}^{3}\frac{\partial^2\phi(x,s)}{\partial x_\alpha^2}-\frac{\partial^2\phi(x,s)}{\partial s^2}+m^2\phi(x,s)=0 \tag{4.8}$$

on a *five*-dimensional Minkowski-type space M_5. Let M_5 and the four-dimensional Minkowski spacetime M be related as in Lemma 2.7. We will consider a solution $\phi(x,s)$ which satisfies the condition $\phi(x,-s)=\phi(x,s)$, and we will define $\xi(x)$ by $\xi(x)=\phi(x,0)$. In the case when $m=0$ this is kind of a reversal of Araki's construction.[1] We will, however, consider the more general situation of any $m\geqslant 0$.

Let $\tau>\rho>0$, and let $h(\mathbf{x},s)$ be an infinitely differentiable function on R^4 (with Cartesian coordinates $(\mathbf{x},s)=(x_1,x_2,x_3,s)$) with support equal to $K_0=\{(\mathbf{x},s)||\mathbf{x}|^2+(s-\tau)^2\leqslant\rho^2\}\cup\{(\mathbf{x},s)||\mathbf{x}|^2+(s+\tau)^2\leqslant\rho^2\}$, and such that $h(\mathbf{x},-s)=h(\mathbf{x},s)$. We interpret the domain R^4 of this function as the spacelike hyperplane $\sigma_{05}=\{((\mathbf{x},0),s)|\mathbf{x}\in\mathsf{R}^3,s\in\mathsf{R}^1\}$ in M_5. The set K_0 is

thus the union of two closed balls of radii ρ, centered at the points $(\mathbf{0},\tau)$ and $(\mathbf{0},-\tau)$. Since we assumed that $\tau>\rho$, these balls are disjoint and do not intersect the timelike hyperplane $\Gamma\mathsf{M}=\{(x,0)|x\in\mathsf{M}\}$. Let $m\geqslant 0$. There then exists a unique solution $\phi(x,s)$ of the equation in (4.8) such that $\phi((\mathbf{x},0),s)=h(\mathbf{x},s)$, and such that the derivative of $\phi(x,s)$ with respect to t vanishes on σ_{05}, and this solution is a *bounded* C_∞-function. The function $\xi(x)=\phi(x,0)$ is then also a bounded C_∞-function, which we can identify with a tempered distribution in accordance with (4.2). Its Fourier transform $\hat{\xi}(p)$, as given by (2.1), is also a tempered distribution. The support of the Fourier transform $\hat{\phi}(p,q)$ (which is a tempered distribution) is in the set $\{(p,q)|p\cdot p=m^2+q^2, p\in\mathsf{M}, q\in\mathsf{R}^1\}$. It follows that $\mathrm{supp}(\hat{\xi}(p))\subset\{p|p\cdot p\geqslant m^2\}$, and hence $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$. The Araki distribution $\xi(x)$ is, however, ''special'' in the sense that $\mathrm{supp}(\hat{\xi}(p))\subset\overline{V}(0)$.

On the basis of the analog of Proposition 4.1 for a five-dimensional spacetime we find, by a simple computation, that

$$\mathrm{supp}(\xi(x))\subset R=\{x|\rho+|t|\geqslant\sqrt{\tau^2+|\mathbf{x}|^2}\}. \tag{4.9}$$

For a suitable choice of the initial-value function $h(\mathbf{x},s)$ the support of $\xi(x)$ is *precisely* the above closed set R. We write $S=\widetilde{R}$, and hence $N(\xi)\supset S$. We note that S contains the *time-slice* $T=\{x|\tau-\rho>t>-(\tau-\rho)\}$, and hence we have $S^{cc}\supset T^{cc}=\mathsf{M}$, and $\mathscr{E}(S)\supset\mathscr{E}(T)=\mathsf{M}$. The set S is *not* a union of wedges, and Theorem 2.11 is thus *not* applicable. It is not hard to see that the Araki envelope $a(S)$ in Lemma 2.7 is actually equal to S in this case, and hence strictly smaller than $\mathscr{E}(S)$. For a region $S=\widetilde{R}$, with R as given in (4.9), we thus have the result that the largest *common* null region of all distributions in $\mathscr{S}'_A(\mathsf{M})$ which vanish on S is equal to $S=a(S)$. The implications of examples of this kind for quantum field theory have been discussed by Haag and Schroer.[14]

The function $\xi(x)$ constructed above can be regarded as a superposition of solutions of Klein–Gordon equations for different masses ranging from m to infinity. Any solution of a Klein–Gordon equation which vanishes on the time-slice T vanishes identically. The function $\xi(x)$ behaves very differently, and it turns out that the essential feature for this behavior of $\xi(x)$ is that $\mathrm{supp}(\hat{\xi}(p))$ includes arbitrarily large values of $p\cdot p$.

Let $S_0=\cup\{W|W\in\mathscr{W}, W\subset S\}$. We easily see that $S_0=D^c$, where D is the closed double cone with apices $(\mathbf{0},-\rho)$ and $(\mathbf{0},\rho)$. We trivially have $a(S_0)=\mathscr{E}(S_0)=S_0=S_0^{cc}$, and Theorem 2.11 is applicable to the set S_0, but as we said before, not to the strictly larger set S.

We next consider examples in which the null region of an Araki distribution includes a ''slab-region'' between two parallel *characteristic* planes.

Example 4.8: Let $f(s)$ be a twice differentiable bounded function on R^1. The function $\psi(x)=f(t-x_3)$ on M is then a bounded solution of the wave equation (4.1) for $m=0$, and we have $\psi(x)\subset\mathscr{S}'_A(\mathsf{M})$. We consider four special cases.

(a) $\mathrm{supp}(f)=[-\tau,\tau]$, for some $\tau>0$. The support of ψ is thus the closed ''slab-region'' R bounded above by the characteristic plane $\pi'=\{x|t=x_3+\tau\}$, and bounded below by the plane $\pi''=\{x|t=x_3-\tau\}$. Let S_u be the open half-space bounded below by π', and let S_l be the open half-space bounded above by π''. The sets S_u, S_l and $S_u\cup S_l=S_0$ are unions of wedges, and hence Theorem 2.11 is applicable. We have $a(S_0)=\mathscr{E}(S_0)=S_0\neq S_0^{cc}=\mathsf{M}$, and $N(\psi)=S_0$, $\mathrm{supp}(\psi)=\widetilde{S_0}=R$.

(b) $\mathrm{supp}(f)=[-\infty,\tau]$. With the notation as in (a) we have $a(S_u)=\mathscr{E}(S_u)=S_u\neq S_u^{cc}=\mathsf{M}$, and $N(\psi)=S_u$, $\mathrm{supp}(\psi)=\widetilde{S_u}$.

(c) $\mathrm{supp}(f)=[-\tau,+\infty]$. In this case we have $a(S_l)=\mathscr{E}(S_l)=S_l\neq S_l^{cc}=\mathsf{M}$, and $N(\psi)=S_l$, $\mathrm{supp}(\psi)=\widetilde{S_l}$.

(d) $\mathrm{supp}(f)=[-\infty,-\tau]\cup[\tau,+\infty]$. With $S_s=\{x|x_3+\tau>t>x_3-\tau\}$ we have $a(S_s)=\mathscr{E}(S_s)=S_s\neq S_s^{cc}=\mathsf{M}$, and $N(\psi)=S_s$, $\mathrm{supp}(\psi)=\widetilde{S_s}$.

In view of the above examples (b)–(d) we conclude the following. If S is an open half-space bounded above or below by a characteristic plane, then there exist distributions $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$ such that $N(\xi)=S=a(S)=\mathscr{E}(S)$. The same conclusion obtains if S is an open ''slab-region'' bounded by two parallel characteristic planes. In both cases $S^{cc}=\mathsf{M}$ is strictly larger than $a(S)$, and we have trivial examples showing that $N(\xi)\supset S$ does not in general imply that $N(\xi)\supset S^{cc}$. We

can now conclude that under the premises in Theorem 3.2 the largest *common* null region $\nu(S)$ of *all* distributions $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$ which vanish on S is just $\nu(S)=a(S)=\mathscr{E}(S)$.

(e) With $f(s)$ as above we consider the function $\psi_1(x)=f(t-x_3)\exp(mx_1)$ on M, for $m>0$. This function satisfies the Klein–Gordon equation (4.1) for the mass m, and its support is equal to the support of the function $\psi(x)$, but it is *not* in the class $\mathscr{S}'_A(\mathsf{M})$ since it grows exponentially in some directions. As we will see a bit later, there are actually *no* nonzero Araki distributions which satisfy the equation (4.1) for some $m>0$, and have a null region which includes a nonempty slab-region between two parallel characteristic planes.

Example 4.9: The distribution $\eta(x)=\delta(x_4-x_3)$, with Fourier transform $\hat{\eta}(p)=2\pi\delta(p_1)\delta(p_2)\delta(p_3-p_4)$, is an element of $\mathscr{S}'_A(\mathsf{M})$. It can be regarded as a limiting case of the distributions in Example 4.8(a) above, in which the planes π' and π'' coincide and the closed ''slab-region'' R becomes a characteristic plane. We have $N(\xi)=\widetilde{R}=S_u\cup S_l$, with S_u and S_l defined as in Example 4.8.

Let R be the ''slab-region'' as in Example 4.8, or the plane $R=\pi'=\pi''$ in the present example. Let $\xi\subset\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)\supset S_c=S_a\cup S_b$, where S_a lies *above* R and S_b lies *below* R. Our simple examples show that the largest *common* null region $\nu(S_c)$ of *all* distributions $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$ which vanish on S_c *must* be contained in $\widetilde{R}=S_u\cup S_l$. In particular we have $a(S_c)\subset\mathscr{E}(S_c)=\mathscr{E}(S_a)\cup\mathscr{E}(S_b)\subset\nu(S_c)\subset S_u\cup S_l$. This is relevant for the configurations in Examples 3.11–3.13 in Section III. We also see that the conclusion in Proposition 3.8 can fail if the set S is not connected. Similarly we see that the conclusion in Proposition 3.3 can fail if $W_1\cap W_2=\varnothing$.

We next consider two examples which have to do with the issues in Propositions 2.3 and 3.9. The geometric considerations are mildly complicated, and the reader may find it useful to follow the discussion by drawing suitable symbolic diagrams.

Example 4.10: Let $f(s)$ be an infinitely differentiable function on R^1 whose support is the interval $[-\tau,\tau]$, with $\tau>0$. Let $m\geqslant 0$ be fixed in what follows. The function $\psi(x)=f(t-x_3)\exp(mx_1)$ on M is then an infinitely differentiable solution of the equation (4.1) for mass m, and we have $\mathrm{supp}(\psi)=R=\{x|x_3+\tau\geqslant t\geqslant x_3-\tau\}$. We will construct other solutions of the equation (4.1) from the solution $\psi(x)$, specifying the initial values on the spacelike surface $\sigma_0=\{(\mathbf{x},0)|\mathbf{x}\in\mathsf{R}^3\}$. We write $R_0=R\cap\sigma_0=\{(\mathbf{x},0)|\tau\geqslant x_3\geqslant-\tau\}$. Let $u(\mathbf{x})$ be an infinitely differentiable function on R^3 such that $u(\mathbf{x})=1$ when $|\mathbf{x}|\leqslant 3\tau$ and $u(\mathbf{x})=0$ when $|\mathbf{x}|\geqslant 4\tau$. Let $\psi_1(x)$ be the unique solution of the equation (4.1) such that $\psi_1(\mathbf{x},0)=u(\mathbf{x})\psi(\mathbf{x},0)$ and $(\partial_t\psi_1)(\mathbf{x},0)=u(\mathbf{x})\times(\partial_t\psi)(\mathbf{x},0)$, and let $\psi_2(x)$ be the unique solution of the equation (4.1) such that $\psi_2(\mathbf{x},0)=(1-u(\mathbf{x}))\psi(\mathbf{x},0)$ and $(\partial_t\psi_1)(\mathbf{x},0)=(1-u(\mathbf{x}))(\partial_t\psi)(\mathbf{x},0)$. We thus have $\psi(x)=\psi_1(x)+\psi_2(x)$ for all $x\in\mathsf{M}$, and since the initial values of $\psi_1(x)$ on σ_0 are infinitely differentiable and of compact support, it follows that $\psi_1(x)\in\Psi^H_m$. We write $R_{01}=\{(\mathbf{x},0)|\tau\geqslant x_3\geqslant-\tau,|\mathbf{x}|\leqslant 4\tau\}$ and $R_{02}=\{(\mathbf{x},0)|\tau\geqslant x_3\geqslant-\tau,|\mathbf{x}|\geqslant 3\tau\}$. The subset R_{01} of σ_0 thus contains the supports of the initial values of $\psi_1(x)$, and R_{02} contains the supports of the initial values of $\psi_2(x)$. By Proposition 4.1 we then have $N(\psi_1)\supset R^c_{01}$ and $N(\psi_2)\supset R^c_{02}$. Let S_u and S_l be defined as in Example 4.8, i.e., $S_u=\{x|t>x_3+\tau\}$ and $S_l=\{x|t<x_3-\tau\}$. Since $\psi(x)=\psi_1(x)+\psi_2(x)$, and since $N(\psi)=\widetilde{R}=S_u\cup S_l$, it follows that

$$N(\psi_1)\supset R^c_{01}\cup(R^c_{02}\cap(S_u\cup S_l))=R^c_{01}\cup(R^c_{02}\cap S_u)\cup(R^c_{02}\cap S_l)\equiv S_1. \tag{4.10}$$

We consider the four open double cones $D^o_l=D(x'_l,x''_l)^o$, $D^o_i=D(x''_l,x'_u)^o$, $D^o_u=D(x'_u,x''_u)$, and $D^o=D(x'_l,x''_u)^o$, with the apices $x'_l=(\mathbf{0},-3\tau)$, $x''_l=(\mathbf{0},-\tau)$, $x'_u=(\mathbf{0},\tau)$, and $x''_u=(\mathbf{0},3\tau)$. The apices thus all lie on the time axis, and such that D^o_l lies below D^o_i which lies below D^o_u. The two double cones D^o_l and D^o_u are thus *timelike* separated, and we have $D^o_l\cup D^o_i\cup D^o_u\subset D^o$ and $D^o=(D^o_l\cup D^o_u)^{cc}$. By the above definitions of R_{02} and $\psi_2(x)$ we have $D^o\subset R^c_{02}$, and hence $\psi_2(x)=0$ and $\psi_1(x)=\psi(x)$ for all $x\in D^o$. In view of the nature of the support R of $\psi(x)$ it then follows that

$$\mathrm{supp}(\psi_1)\supset\overline{D^o_i},\quad N(\psi_1)\supset D^o_l\cup D^o_u. \tag{4.11}$$

We can now draw some interesting and important conclusions from the above.

Proposition 3.9 says (among other things) that if K_1 and K_2 are two subsets of M in the set $\mathscr{K}_E$, and if $\xi(x)\in\mathscr{S}'_A(\mathsf{M})$, with null region $N(\xi)$, then $N(\xi)\supset(K_1\cap K_2)^c$ whenever $N(\xi)\supset K_1^c\cup K_2^c$. We can think about this as expressing an ''intersection property'' of the causal complements of certain sets contained in the null region of $\xi(x)$. Now it is important to understand that this intersection property is definitely *not* valid for two *arbitrary* closed, causally complete subsets K_l and K_u of M. Let $K_l=(D_l^o)^c$ and $K_u=(D_u^o)^c$. These sets are closed and causally complete, and we have $(K_l\cap K_u)^c=(D_l^o\cup D_u^o)^{cc}=D^o\supset\overline{D_i^o}$. In view of (4.11) the set $(K_l\cap K_u)^c$ is *not* contained in $N(\psi_1)$, whereas $N(\psi_1)\supset D_l^o\cup D_u^o=K_l^c\cup K_u^c$. For the consideration of this issue in quantum field theory, or in the theory of local nets, it is of importance that the example involves distributions in $\mathscr{S}'_A(\mathsf{M})$ which are functions $\psi_1(x)$ in Ψ_m^H, for an arbitrary $m\geqslant 0$. Note that we can clearly find *real-valued* functions of this nature if we like.

Example 4.11: We continue the above discussion, and we consider the two timelike separated points $z'=(\mathbf{0},-2\tau)\in D_l^o\subset S_1$ and $z''=(\mathbf{0},2\tau)\in D_u^o\subset S_1$, with $S_1\subset N(\psi_1)$ as defined in (4.10). We thus have $z',z''\in N(\psi_1)$, but $[z',z'']$ is *not* in $N(\psi_1)$ since $[x_l'',x_u']\subset[z',z'']$ and since $\mathrm{supp}(\psi_1)\supset[x_l',x_{u\prime}]$ by (4.11). We will now give a particular example of a continuous curve $C\subset S_1\subset N(\psi_1)$ with z' and z'' as endpoints. The curve C is composed of five arcs, $C=C_1\cup C_2\cup C_3\cup C_4\cup C_5$ as follows. $C_1=[z',z_1]$ with $z_1=(0,0,2\tau,0)$, $C_2=[z_1,z_2]$ with $z_2=(0,0,5\tau,0)$, $C_3=\{(0,5\tau\sin\theta,5\tau\cos\theta,0)\,|\,\theta\in[0,\pi]\}$, $C_4=[z_3,z_4]$ with $z_3=(0,0,-5\tau,0)$ and $z_4=(0,0,-2\tau,0)$, and $C_5=[z_4,z'']$. We easily see that $C_1\subset D^o\cap S_l\subset R_{02}^c\cap S_l$; $C_5\subset D^o\cap S_u\subset R_{02}^c\cap S_u$; $C_2\subset R_{02}^c\cap S_l$; $C_4\subset R_{02}^c\cap S_u$; and $C_3\subset R_{01}^c$, and in view of (4.10) it follows that $C\subset S_1\subset N(\psi_1)$. The curve C does not satisfy the conditions of the curve C in Proposition 2.3, and as our example shows, the conclusion in the proposition can then fail to hold. We remark that the set S_1 is, in fact, connected, but we need not prove this here.

With the premises in Theorem 3.6, or in Propositions 3.3 or 3.8 we have $N(\xi)\supset S^{cc}$ if $N(\xi)\supset S$, but this conclusion depends on the very special nature of S in these propositions, and it is definitely not valid for arbitrary subsets S of M, as our many examples clearly show. Concerning the results in Theorem 3.2 about the null regions of distributions in $\mathscr{S}'_A(\mathsf{M})$, our Example 4.8 shows that these results can in general not be ''improved'' in the sense that we could find a *larger* region than $\mathscr{E}(S)$ which is always included in the null region $N(\xi)$ of such a distribution ξ, given that $N(\xi)\supset S$. There is, however, a physically interesting *subset* of $\mathscr{S}'_A(\mathsf{M})$ for which we *can* obtain stronger results, and we will now consider this subset.

Lemma 4.12: Let u be a forward isotropic vector, and let $S\subset\mathsf{M}$ be a nonempty open set which is invariant under all translations in the direction u, i.e., we have $S=\{x+\tau u\,|\,x\in S,\tau\in\mathsf{R}^1\}$. Let $\eta(x)\in\mathscr{S}'(\mathsf{M})$, with null region $N(\eta)\supset S$, be such that the support of the Fourier transform of $\eta(x)$ is contained in the open lightcone, i.e., $\mathrm{supp}(\hat{\eta}(p))\subset V(0)$. Then $\eta=0$.

In particular this applies to the case when S is a ''slab-region'' bounded by two parallel characteristic planes, or when S is a half-space bounded by a characteristic plane.

Proof: (1) Let C be a compact subset of S with a nonempty interior. There then exists a nonempty open set S_1 such that the translated set $C_y=\{x+y\,|\,x\in C\}$ is included in S for all $y\in S_1$. We then have $C_{y+\tau u}\subset S$ for all $y\in S_1$ and all $\tau\in\mathsf{R}^1$.

Let $g(x)$ be an infinitely differentiable function on M with support in C, with Fourier transform $\hat{g}(p)$. For any $y\in\mathsf{M}$ we define the function $g_y(x)$ (of x in M) by $g_y(x)=g(x-y)$, and we define the function $f(y)$ by

$$f(y)=\eta[g_y]=\int_{(\infty)}d^4(x)\,\eta(x)g_y(x)=\int_{(\infty)}d^4(p)\,\hat{\eta}(p)\hat{g}(-p)\exp(iy\cdot p).$$

Since $\eta(x)=0$ on S, it follows from the above constructions that $f(y+\tau u)=0$ for all $y\in S_1$, and all $\tau\in\mathsf{R}^1$.

(2) Since $\mathrm{supp}(\hat{\eta}(p))\subset V(0)$ we can write $f(y)$ as the sum $f(y)=f_+(y)+f_-(y)$, with

$$f_+(y)=\int_{V_+(0)}d^4(p)\,\hat{\eta}(p)\hat{g}(-p)\exp(iy\cdot p),\quad f_-(y)=\int_{V_-(0)}d^4(p)\,\hat{\eta}(p)\hat{g}(-p)\exp(iy\cdot p).\tag{4.12}$$

Both functions are infinitely differentiable functions of y, and since η is a *tempered* distribution we have

$$|f_+(y)|<C\cdot(1+|\mathbf{y}|+|y_4|)^{n-1},\quad |f_-(y)|<C\cdot(1+|\mathbf{y}|+|y_4|)^{n-1},$$

for some positive integer n independent of g, and some constant C.

With $y\in S_1$ temporarily fixed we consider the functions $h_+(\tau)=f_+(y+\tau u)$ and $h_-(\tau)=f_-(y+\tau u)$, as functions of τ on R^1. We have $h_+(\tau)+h_-(\tau)=f(y+\tau u)=0$ for all real τ. It follows at once from (4.12) that the supports of the Fourier transforms of these two functions (regarded as distributions in $\mathscr{S}'(\mathsf{R}^1)$) have only the point 0 in common, and hence $h_+(\tau)=h_-(\tau)$ is a polynomial in τ, of a degree not exceeding $n-1$. Hence $\partial_\tau^n h_+(\tau)=0=\partial_\tau^n h_-(\tau)$, and we thus conclude that $(u\cdot\partial)^n f_+(y)=0$ and $(u\cdot\partial)^n f_-(y)=0$ for all $y\in S_1$. Here ∂ denotes the gradient operator.

(3) We define the *open forward imaginary tube* $\mathscr{T}_{i+}$ in complex Minkowski space by $\mathscr{T}_{i+}=\{y'+iy''|y'\in\mathsf{M},y''\in V_+(0)\}$. By inspection of the expression for $f_+(y)$ in (4.12) we see that the integral defines a (jointly) continuous function of the complex variable y on the closed forward imaginary tube $\overline{\mathscr{T}}_{i+}$, which is analytic on the open tube. Since $(u\cdot\partial)^n f_+(y)$ vanishes on the nonempty open set $S_1\subset\mathsf{M}$, it follows that $(u\cdot\partial)^n f_+(y)$ vanishes on $\overline{\mathscr{T}}_{i+}$, and in particular for all real $y\in\mathsf{M}$. In a similar fashion we conclude that $(u\cdot\partial)^n f_-(y)$ vanishes on M, and hence $(u\cdot\partial)^n\eta[g_y]=(u\cdot\partial)^n f(y)=0$ for all $y\in\mathsf{M}$. Since this holds for all $g(x)\in\mathscr{S}(\mathsf{M})$ with support in S_1, and for all $y\in\mathsf{M}$, it follows that $(u\cdot p)^n\hat{\eta}(p)=0$. Since $u\cdot p\neq 0$ on supp$(\hat{\eta})$, it follows that $\eta=0$. This completes the proof.

The considerations in Example 4.8 show that the conclusion in the lemma can fail for distributions in $\mathscr{S}'_A(\mathsf{M})$ which do not satisfy the condition that supp$(\hat{\eta}(p))\subset V(0)$, as well as for solutions of the Klein–Gordon equation for $m>0$ which are not in $\mathscr{S}'_A(\mathsf{M})$.

We recall here part (c) of Proposition 2.5, according to which $\xi(x)=0$ on a ''slab-region'' if $\xi(x)=0$ on a ''tube-region'' of the form $S=\{x+\tau u|x\in D^o,\tau\in\mathsf{R}^1\}$, where D^o is an open double cone. *This* conclusion holds for *all* distributions in $\mathscr{S}'_A(\mathsf{M})$, and does not depend on the assumption in the present lemma that supp$(\hat{\eta}(p))\subset V(0)$.

The above result leads to an immediate ''improvement'' of Theorem 3.2 for the Araki distributions which satisfy the special condition that supp$(\hat{\eta}(p))\subset V(0)$, as follows.

Proposition 4.13: Let the premises and notation be as in Theorem 3.2. Let $\eta(x)\in\mathscr{S}'(\mathsf{M})$ be such that supp$(\hat{\eta}(p))\subset V(0)$. Let $N(\eta)$ be the null region of η, and suppose that $N(\eta)\supset S$. Then $N(\eta)\supset(S^{cc})^o$.

Proof: We have $\eta(x)\in\mathscr{S}'_A(\mathsf{M})$. By Theorem 3.2 we either have $\mathscr{E}(S)=(S^{cc})^o$, or else $S^{cc}=\mathsf{M}$ *and* $\mathscr{E}(S)$ contains an open slab-region bounded by two parallel characteristic planes. In the latter case we have $N(\eta)=\mathsf{M}$ by Lemma 4.12 above, and hence $N(\eta)\supset(S^{cc})^o$ in all cases.

As we have said, we are led to the Araki distributions in $\mathscr{S}'_A(\mathsf{M})$ if we consider matrix elements of commutators of field operators between two states of compact support in momentum space. If we consider *vacuum expectation values* of commutators of field operators in a theory with no massless particles, we are led to the distributions $\eta(x)\in\mathscr{S}'_A(\mathsf{M})$ which satisfy the additional condition that supp$(\hat{\eta}(p))\subset V(0)$.

Lemma 4.12 and Proposition 4.13 apply in particular to those solutions of the Klein–Gordon equation for $m>0$ which can be regarded as elements in $\mathscr{S}'_A(\mathsf{M})$. For C_2-solutions this means that the solution can grow no faster than some power of $(|\mathbf{x}|+|t|)$ as $x=(\mathbf{x},t)$ tends to infinity. As we see from Example 4.8, the conclusions can fail for C_2-solutions which grow exponentially as x tends to infinity in some direction. The conclusions can also fail in the case of the *wave equation*, i.e., the Klein–Gordon equation for $m=0$.

We next consider a stronger result than in Lemma 4.12, but with a slightly stronger condition on the support of $\hat{\eta}(p)$. It is now given that the distribution vanishes throughout a region obtained when we remove a portion ''in the middle'' from a ''tube-region'' of the kind described in Proposition 2.5.

Lemma 4.14: Let $\rho>0$, and let D^o_ρ be the open double cone with apices $x'_\rho=(\mathbf{0},-\rho)$ and $x''_\rho=(\mathbf{0},\rho)$. We define the open ''tube-region'' $T(\rho)$ in Minkowski spacetime by

$$T(\rho)=\{(x_1,x_2,x_3+s,x_4+s)|x\in D_\rho^o, s\in \mathsf{R}^1\}.$$

Let $\tau>0$, and let $S=S(\rho,\tau)$ be the region

$$S(\rho,\tau)=\{x|x\in T(\rho),\ |x_3+x_4|>\tau\}. \tag{4.13}$$

Let $\eta\in\mathscr{S}'(\mathsf{M})$, with null region $N(\eta)\supset S$, and with a Fourier transform $\hat{\eta}$ such that $\text{supp}(\hat{\eta}(p))\subset\{p|p\in\mathsf{M}, p\cdot p\geqslant m^2\}$ for some $m>0$. Then $\eta=0$.

Proof: (1) Let $D=D(x',x'')$ be the double cone with apices $x'=(\mathbf{0},-\rho/2)$ and $x''=(\mathbf{0},\rho/2)$, and let $g(x)\in\mathscr{S}(\mathsf{M})$, with $\text{supp}(g(x))\subset D(x',x'')$. We consider the convolution of η with g,

$$\zeta(x)=(\eta\star g)(x)=\int_{(\infty)} d^4(y)\ \ \eta(x-y)g(y). \tag{4.14}$$

The object $\zeta(x)$ is a C_∞-function on M, which we can also regard as a distribution in $\mathscr{S}'(\mathsf{M})$. We thus have, for all $x\in\mathsf{M}$,

$$|\zeta(x)|<k\cdot(1+|\mathbf{x}|+|x_4|)^{n'} \tag{4.15}$$

for some positive integer n' and some constant k.

Since the Fourier transform of $g(x)$ is an entire function of p, we have $\text{supp}(\hat{\zeta}(p))=\text{supp}(\hat{\eta}(p))\subset\{p|p\in\mathsf{M}, p\cdot p\geqslant m^2\}$. Considering the supports of $\eta(x)$ and $g(x)$ we easily see that $N(\zeta)\supset S(\rho_0,\tau_0)$ with $\rho_0=\rho/2$ and $\tau_0=\tau+\rho/2$. The distribution ζ thus satisfies the same kind of conditions as the original distribution η. Our strategy of proof will be to show that $N(\eta)\supset T(\rho_0)$, and it will then follow from Lemma 4.12 that $\eta=0$. Now $\eta=0$ on $T(\rho_0)$ if and only if $\zeta(0,0,s,s)=0$ for all C_∞-functions $g(x)$ with support in $D(x',x'')$ and for all $s\in\mathsf{R}^1$. In order to show that ζ has this property we will show that

$$\int_{-\tau_0}^{\tau_0} ds\,\zeta(0,0,s/2,s/2)s^l=0 \tag{4.16}$$

for every non-negative integer l and for all g. In view of the nature of the support of $\zeta(x)$ discussed above we have $\zeta(0,0,s/2,s/2)=0$ when $|s|\geqslant\tau_0$. Since $\zeta(0,0,s/2,s/2)$ is continuous, and since the powers of s form a total set in $L^2([-\tau_0,\tau_0])$, it follows from (4.16) that we also have $\zeta(0,0,s/2,s/2)=0$ when $|s|\leqslant\tau_0$, and hence $\zeta(0,0,s/2,s/2)=0$ for all real s.

(2) Let $g(x)$ be a C_∞-function with support in $D(x',x'')$, and let l be a non-negative integer. In what follows we regard $g(x)$ and l as fixed. With $\zeta(x)$ defined as in (4.14) we define the function $A(r,s)$ of the two real variables r and s by

$$A(r,s)=\zeta\left(0,0,\frac{s-r}{2},\frac{s+r}{2}\right)s^l. \tag{4.17}$$

The function $A(r,s)$ is an infinitely differentiable function on R^2, and it follows from (4.15) that

$$|A(r,s)|<k_0(1+\rho_0 m^2(|s|+|r|))^n \tag{4.18}$$

for some constant k_0 and some positive integer n. (The optional multiplier $\rho_0 m^2>0$ of $|s|+|r|$ is inserted for convenience, in anticipation of what will follow.) Since $N(\zeta)\supset S(\rho_0,\tau_0)$, it follows that

$$A(r,s)=0,\quad \text{if } |r|\leqslant\rho_0 \quad\text{and } |s|\geqslant\tau_0. \tag{4.19}$$

The function $A(r,s)$ can be regarded as a tempered distribution in $\mathscr{S}'(R^2)$, and we consider its Fourier transform $\hat{A}(q_1,q_2)$ defined by

$$\hat{A}(q_1,q_2)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dsA(r,s)\ \exp(-irq_1-isq_2).$$

Since $\mathrm{supp}(\hat{\zeta}(p))\subset\{p|p\in\mathsf{M},p\cdot p\geqslant m^2\}$ we conclude, by a simple computation, that

$$\hat{A}(q_1,q_2)=0,\quad \text{if}\ \ 4q_1q_2<m^2. \tag{4.20}$$

(3) Let $f_1(s),f_2(s)\in\mathscr{S}(\mathsf{R}^1)$, with Fourier transforms $\hat{f}_1(q)$ and $\hat{f}_2(q)$ defined as in (4.7), be such that

$$\int_{-\infty}^{\infty}dsf_1(s)=1,\quad f_2(0)=1,\quad \mathrm{supp}(\hat{f}_1(q))=\mathrm{supp}(\hat{f}_2(q))\subset(-1/2,1/2). \tag{4.21}$$

For $\lambda>0$ we define $h_\lambda(r,s)\in\mathscr{S}(\mathsf{R}^2)$ by

$$h_\lambda(r,s)=\lambda f_1(\lambda r)f_2(m^2s/\lambda), \tag{4.22}$$

and we then have

$$\begin{aligned}A[h_\lambda]&=\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dsA(r,s)h_\lambda(r,s)\\&=\frac{\lambda}{m^2}\int_{-\infty}^{\infty}dq_1\int_{-\infty}^{\infty}dq_2\hat{A}(q_1,q_2)\hat{f}_1(-q_1/\lambda)\hat{f}_2(-\lambda q_2/m^2).\end{aligned}$$

By (4.21) we have $\hat{f}_1(-q_1/\lambda)\hat{f}_2(-\lambda q_2/m^2)=0$ *unless* $|q_1|<\lambda/2$ *and* $|q_2|<m^2/(2\lambda)$, and hence *unless* $|q_1q_2|<m^2/4$. In view of (4.20) the supports of the distribution $\hat{A}(q_1,q_2)$ and the test function $\hat{f}_1(-q_1/\lambda)\hat{f}_2(-\lambda q_2/m^2)$ are thus *disjoint*, and it follows that $A[h_\lambda]=0$ for all $\lambda>0$.

(4) We next note that

$$\lim_{\lambda\to+\infty}\int_{-\rho_0}^{\rho_0}dr\int_{-\infty}^{\infty}dsA(r,s)h_\lambda(r,s)=\int_{-\tau_0}^{\tau_0}ds\zeta(0,0,s/2,s/2)s^l. \tag{4.23}$$

The range of the s-integration in the expression at left is actually the interval $[-\tau_0,\tau_0]$ in view of (4.19), and the result in (4.23) then follows readily in view of (4.17), (4.21), and (4.22).

We write

$$A_0[h_\lambda]=\int_{-\infty}^{-\rho_0}dr\int_{-\infty}^{\infty}dsA(r,s)h_\lambda(r,s)+\int_{\rho_0}^{\infty}dr\int_{-\infty}^{\infty}dsA(r,s)h_\lambda(r,s). \tag{4.24}$$

We will show that $A_0(h_\lambda)$ tends to zero as λ tends to $+\infty$, and since $A[h_\lambda]=0$ by the result in step (3), it then follows that the right member in (4.23) equals zero, as we promised to show.

For the test functions $f_1(s)$ and $f_2(s)$ we have estimates of the form

$$|f_1(r)|<\frac{k_1}{(1+|r|)^{n+6}},\quad |f_2(s)|<\frac{k_2}{(1+|s|)^{n+4}}, \tag{4.25}$$

for some constants k_1 and k_2. Suppose now that $|r|\geqslant\rho_0$ and that $\lambda>\rho_0m^2$. We then have $(1+\lambda|r|)(1+m^2|s|/\lambda)\geqslant1+\lambda|r|+m^2|rs|>1+\rho_0m^2(|r|+|s|)$, and, trivially, $(1+\lambda|r|)\geqslant1+\lambda\rho_0$. From this, and from the estimates in (4.18) and (4.25), we then obtain the inequality

$$|A(r,s)h_\lambda(r,s)|<\frac{\lambda k_0k_1k_2}{(1+\lambda\rho_0)^2(1+\rho_0m^2(|r|+|s|))^4},$$

which holds for all s and all r such that $|r| \geqslant \rho_0$, provided only that $\lambda > \rho_0 m^2$. On the basis of this estimate we obtain the inequality $|A_0[h_\lambda]| < k_3/\lambda$, where k_3 is some constant, for the quantity in (4.24) when $\lambda > \rho_0 m^2$, and it follows that $A_0(h_\lambda)$ tends to zero as λ tends to $+\infty$. This completes the proof.

This lemma is immediately applicable to solutions of the Klein–Gordon equation for mass $m > 0$. If such a solution $\psi(x)$ vanishes throughout the region $S = S(\rho, \tau)$ defined in the lemma, and if the solution can be regarded as an element of $\mathscr{S}'(\mathsf{M})$, i.e., if it grows no faster than a polynomial in x for large x, then it vanishes on M. We note that we obviously have $\mathscr{E}(S) \neq \mathsf{M}$, and the conclusion that $\psi(x)$ vanishes identically cannot be drawn on the basis of the two ''classical'' principles alone. We also note, with reference to the discussion in Example 4.8, that the conclusion can fail in the case of solutions which grow exponentially for large x. It can also fail for the wave equation, with $m = 0$.

On the basis of Lemma 4.14 we obtain the following interesting (and plausible) result.

Proposition 4.15: Let $W_1, W_2 \in \mathscr{W}$ be such that $W_1^c \cap W_2^c = \varnothing$. Let $\eta(x) \in \mathscr{S}'(\mathsf{M})$, with null region $N(\eta) \supset W_1 \cup W_2$, have a Fourier transform which satisfies the condition that $\mathrm{supp}(\hat{\eta}(p)) \subset \{p | p \in \mathsf{M},\ p \cdot p \geqslant m^2\}$, for some $m > 0$. Then $\eta = 0$.

Proof: If $W_1 \cap W_2 \neq \varnothing$, the conclusion follows from Proposition 3.3. We thus consider the case when $W_1 \cap W_2 = \varnothing$. The upper plane of either wedge is then parallel with the lower plane of the other wedge, and the upper plane of one of the wedges must lie *above* the lower plane of the other. Suppose that the upper plane π_2'' of the wedge $W_2 = W(\pi_2', \pi_2'')$ lies above the lower plane π_1' of the wedge $W_1 = W(\pi_1', \pi_1'')$. The ''slab-region'' R bounded above by π_2'' and bounded below by π_1' then contains a ''cut'' tube-region S, which is Poincaré equivalent to the region $S(\rho, \tau)$ in (4.13), and such that $N(\eta) \supset S$. On the basis of Lemma 4.14 we then conclude that $\eta = 0$. This completes the proof.

We recall again Proposition 3.3, which played an important role in our discussion in Section III. Let $W_1 = W(\pi_1', \pi_1'')$ and $W_2 = W(\pi_2', \pi_2'')$ be two wedges in $\mathscr{W}$, and let $S = W_1 \cup W_2$. Suppose that $W_1 \cap W_2 \neq \varnothing$. The proposition asserts that we then have $a(S) = \mathscr{E}(S) = S^{cc}$, and hence, if $\xi(x) \in \mathscr{S}_A'(\mathsf{M})$, with null region $N(\xi) \supset S$, then $N(\xi) \supset S^{cc}$. In particular, if π_1' is not parallel with π_2'' we have $W(\pi_1', \pi_2'') \subset S^{cc}$, and hence $N(\xi) \supset W(\pi_1', \pi_2'')$. Suppose now that $W_1 \cap W_2 = \varnothing$. We then have $a(S) = \mathscr{E}(S) = S = W_1 \cup W_2$, but S^{cc} is strictly larger than $\mathscr{E}(S)$ unless $W_1 \subset W_2^c$, and the question arises whether it might nevertheless be true that $N(\xi) \supset S^{cc}$ if $N(\xi) \supset S$. This is an important question in view of the nature of our reasoning in the proofs of the propositions and theorems in Section III, in which it was a constant concern whether the configurations of wedges contained a ''sufficiently large'' set of pairs of *intersecting* wedges.

If $W_1 \cap W_2 = \varnothing$, then the normals of the boundary planes of the wedges span either a three-dimensional space, or else a two-dimensional space. In the latter case the lower plane of W_1 is parallel with the upper plane of W_2, and the upper plane of W_1 is parallel with the lower plane of W_2. If $W_1^c \cap W_2^c \neq \varnothing$, then $a(S) = \mathscr{E}(S) = S = S^{cc}$, and by the discussion in Example 4.8, S is the largest common null region $\nu(S)$ of all $\xi \in \mathscr{S}_A'(\mathsf{M})$ for which $N(\xi) \supset S$. If $W_1^c \cap W_2^c = \varnothing$, then $S^{cc} = \mathsf{M}$. As shown in Examples 4.8 and 4.9 we can easily find nonzero distributions $\xi(x) \in \mathscr{S}_A'(\mathsf{M})$ for which $N(\xi) \supset W_1 \cup W_2$, but for the *restricted* class of distributions $\eta(x)$ considered in Proposition 4.15 we do have $N(\eta) = \mathsf{M}$ if $N(\eta) \supset S$.

We consider the situation when $W_1 \cap W_2 = \varnothing$ and when the normals of the boundary planes of the wedges span a *three*-dimensional space. Without any essential loss of generality we can then assume that the plane π_1' is *not* parallel with the plane π_2''. We will now present an example of a solution $\psi_1(x)$ of the Klein–Gordon equation (for any mass $m \geqslant 0$) such that $N(\psi_1)$ includes $S = W_1 \cup W_2$, but *not* $W(\pi_1', \pi_2'')$. Since we have $W(\pi_1', \pi_2'') \subset S^{cc}$ this example shows conclusively that if $\xi(x) \in \mathscr{S}_A'(\mathsf{M})$ and $N(\xi) \supset S = W_1 \cup W_2$, then it does not follow that $N(\xi) \supset S^{cc}$, even for $\xi(x)$ in the restricted class in Proposition 4.15. This is in marked contrast to the situation for the configuration of wedges in that proposition.

Example 4.16: (1) Let $h(s)$ be a real-valued infinitely differentiable function on R^1 with $\mathrm{supp}(h) = [-\tau, \tau]$, with $\tau > 0$. Let $m \geqslant 0$, $\rho_2 > \rho_1 > 0$, and let $g(q)$ be a real-valued infinitely differentiable function on R^1 with $\mathrm{supp}(g) = [\rho_1, \rho_2]$. We define the function $f(x_1, x_2)$ on R^2 by

$$f(x_1, x_2) = \int_{\rho_1}^{\rho_2} dq\, g(q) \exp(x_2 \sqrt{q^2 + m^2}) \cos(q x_1). \tag{4.26}$$

The function $\psi(x)$ on M defined by

$$\psi(x)=f(x_1,x_2)h(x_3-t)$$

is then a *real-valued* C_∞-solution of the Klein–Gordon equation (4.1), for mass m.

We define H_+ as the open half-space bounded below by the characteristic plane $\pi_2'=\{x|t=x_3+\tau\}$, and we define H_- as the open half-space bounded above by the characteristic plane $\pi_1''=\{x|t=x_3-\tau\}$. The integral in (4.26) defines an entire analytic function of the variables x_1 and x_2, and hence $\mathrm{supp}(\psi(x))=\{x|x\in\mathsf{M},x_3+\tau\geqslant t\geqslant x_3-\tau\}=\widetilde{H}_+\cap\widetilde{H}_-$.

(2) Let $c(s)$ be a real-valued infinitely differentiable function such that $c(s)=0$ for $s\leqslant 0$ and $c(s)=1$ for $s\geqslant s_0>0$, with s_0 a constant. Let $\psi_1(x)$ and $\psi_2(x)$ be the unique solutions of the Klein–Gordon equation for mass m which satisfy the initial conditions

$$\psi_1(\mathbf{x},0)=(1-c(x_2))\psi(\mathbf{x},0),\quad (\partial_t\psi_1)(\mathbf{x},0)=(1-c(x_2))(\partial_t\psi)(\mathbf{x},0),$$

$$\psi_2(\mathbf{x},0)=c(x_2)\psi(\mathbf{x},0),\quad (\partial_t\psi_2)(\mathbf{x},0)=c(x_2)(\partial_t\psi)(\mathbf{x},0),$$

on the spacelike 3-plane $\sigma_0=\{(\mathbf{x},0)|\mathbf{x}\in R^3\}$. With this construction we have the following obvious properties of the solutions ψ_1 and ψ_2. (a) $\psi_1(x)+\psi_2(x)=\psi(x)$, for all $x\in\mathsf{M}$. (b) The supports of the initial values $\psi_1(\mathbf{x},0)$ and $(\partial_t\psi_1)(\mathbf{x},0)$ of $\psi_1(x)$ are contained in the subset $\sigma_{0,1}=\{(\mathbf{x},0)|x_2\leqslant s_0,|x_3|\leqslant\tau\}$ of the plane σ_0, and the supports of the initial values $\psi_2(\mathbf{x},0)$ and $(\partial_t\psi_2)(\mathbf{x},0)$ of $\psi_2(x)$ are contained in the subset $\sigma_{0,2}=\{(\mathbf{x},0)|x_2\geqslant 0,|x_3|\leqslant\tau\}$ of the plane σ_0. (c) The initial values $\psi_1(\mathbf{x},0)=(1-c(x_2))f(x_1,x_2)h(x_3)$ and $(\partial_t\psi_1)(\mathbf{x},0)=-(1-c(x_2))f(x_1,x_2)h'(x_3)$ are in the space of functions $\mathscr{S}(\mathsf{R}^3)$, and hence $\psi_1(x)$ is a real-valued infinitely differentiable solution of the Klein–Gordon equation in the set $\Psi_m^H\subset\mathscr{S}_A'(\mathsf{M})$. (The solution $\psi_2(x)$ is obviously *not* in $\mathscr{S}_A'(\mathsf{M})$ since it grows exponentially in the positive 2-direction.)

(3) We define the two characteristic planes π_1' and π_2'' by $\pi_1'=\{(\mathbf{x},t)|t=x_2\}$ and $\pi_2''=\{(\mathbf{x},t)|t=-x_2\}$. We then have $\sigma_{0,2}\subset\overline{W}(\pi_1'',\pi_1')=W(\pi_1',\pi_1'')^c$, and it follows that $\psi_2(x)=0$ on $W(\pi_1',\pi_1'')$. Now $W(\pi_1',\pi_1'')\subset H_-$, and since $\psi_1(x)+\psi_2(x)=\psi(x)=0$ on H_-, we conclude that $\psi_1(x)=0$ on $W(\pi_1',\pi_1'')=W_1$. Similarly we conclude that $\psi_1(x)=0$ on $W(\pi_2',\pi_2'')=W_2$.

We have $\sigma_{0,2}\subset\overline{W}(\pi_2'',\pi_1')=W(\pi_1',\pi_2'')^c$, and hence $\psi_2(x)=0$ on $W(\pi_1',\pi_2'')$, which implies that $\psi_1(x)=\psi(x)$ on $W(\pi_1',\pi_2'')$. We easily see that $W(\pi_1',\pi_2'')\cap\mathrm{supp}(\psi)=W(\pi_1',\pi_2'')\cap\mathrm{supp}(\psi_1)$ is *nonempty*. In particular this set contains the open double cone $D^o=D(x',x'')^o$ with apices $x'=(0,-\tau,0,-\tau)$ and $x''=(0,-\tau,0,\tau)$.

(4) In summary we have the following situation. The *real-valued* solution $\psi_1(x)\in\Psi_m^H$ vanishes throughout $W_1\cup W_2$. The upper plane π_1'' of the wedge $W_1=W(\pi_1',\pi_1'')$ is parallel with the lower plane π_2' of the wedge $W_2=W(\pi_2',\pi_2'')$ and lies *below* π_2'. Hence $W_1\cap W_2=\varnothing$. The function $\psi_1(x)$ *does not* vanish throughout the wedge $W(\pi_1',\pi_2'')$, and hence does not vanish throughout $(W_1\cup W_2)^{cc}$. This is the example we promised above.

Example 4.17: In the above example the upper plane of the wedge W_1 is parallel with the lower plane of the wedge W_2. It is of interest to find an example in which no two boundary planes are parallel, and we can do this with a minor modification of Example 4.16. Let the premises and notation be as in that example. We define the two characteristic planes π_3' and π_3'' by $\pi_3'=\{x|t\sqrt{2}=\tau+s_0-x_2-x_3\}$ and $\pi_3''=\{x|t\sqrt{2}=x_2+x_3-\tau-s_0\}$. We easily see that $\sigma_{0,1}\subset\overline{W}(\pi_3'',\pi_3')=W(\pi_3',\pi_3'')^c$, and hence we have $\psi_1(x)=0$ on $W_3=W(\pi_3',\pi_3'')$. We next note that the point $x=(0,-s_0,2\tau+2s_0,0)$ lies *above* the planes π_1' and π_3', and *below* the planes π_1'' and π_3''. Hence $W_1\cap W_3\neq\varnothing$, and by Proposition 3.3 we conclude that $\psi_1(x)=0$ on $W(\pi_1',\pi_3'')$. The situation is thus as follows. No two boundary planes of the wedges $W(\pi_1',\pi_3'')$ and $W(\pi_2',\pi_2'')$ are parallel. We have $\psi_1(x)=0$ on $W(\pi_1',\pi_3'')\cup W(\pi_2',\pi_2'')$, but $\psi_1(x)$ does *not* vanish throughout the subset $W(\pi_1',\pi_2'')$ of $(W(\pi_1',\pi_3'')\cup W(\pi_2',\pi_2''))^{cc}$. Note that we necessarily have $W(\pi_1',\pi_3'')\cap W(\pi_2',\pi_2'')=\varnothing$, because otherwise we would have $N(\psi_1)\supset W(\pi_1',\pi_2'')$, by Proposition 3.3. It is not hard to show by a direct computation that the intersection is indeed empty. The configuration of the two wedges $W(\pi_1',\pi_3'')$ and $W(\pi_2',\pi_2'')$ is of the same nature as the configuration in Example 3.14 in Section III, and the remarks about that example apply here, too.

We conclude this section with two simple examples which show that the common null region of all solutions of the Klein–Gordon equation in Ψ_m, with $m>0$, which vanish on certain open sets S can be *larger* than $\mathscr{E}(S)$ (but, of course, *never* larger than $(S^{cc})^o$, in view of the discussion in Proposition 4.4). The salient point in these examples is that the solutions can be expressed in terms of the initial values on any (suitable) spacelike hypersurface.

Example 4.18: Let $\tau>0$, and let S be the open subset of M defined by $S=\{x|t>-x_3, x_3>t>x_3-\tau\}$. The region S thus consists of all points in the slab-region S_0 between the two characteristic planes $\pi''=\{x|t=x_3\}$ and $\pi'=\{x|t=x_3-\tau\}$ which lie *above* the characteristic plane $\pi=\{x|t=-x_3\}$. We then have $\mathscr{E}(S)=S$, but $S^{cc}=W_R=W(\pi,\pi'')=\{x|x_3>|t|\}$. Let $m>0$. Let $\psi(x)$ be a C_∞-solution of the Klein–Gordon equation for mass m in the space $\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$, and suppose that $N(\psi)\supset S$. We will show that we then have $N(\psi)\supset S^{cc}=W_R$.

We consider the initial values of $\psi(x)$ on the spacelike 3-plane $\sigma_0=\{(\mathbf{x},0)|\mathbf{x}\in\mathsf{R}^3\}$, and we note that these vanish on the ''strip'' $\{(\mathbf{x},0)|\tau>x_3>0\}$. Let $\psi_R(x)$ be the (unique) solution of the Klein–Gordon equation for mass m determined by the initial conditions $\psi_R(\mathbf{x},0)=\psi(\mathbf{x},0)$ and $(\partial_t\psi_R)(\mathbf{x},0)=(\partial_t\psi)(\mathbf{x},0)$ when $x_3\geqslant\tau$, and $\psi_R(\mathbf{x},0)=0=(\partial_t\psi_R)(\mathbf{x},0)$ when $x_3<\tau$. We write $\psi_L(x)=\psi(x)-\psi_R(x)$, and we then have $\psi_L(\mathbf{x},0)=0=(\partial_t\psi_L)(\mathbf{x},0)$ when $x_3>0$. We easily see that $\psi_R(x),\psi_L(x)\in\Psi_m$. In view of the supports of the initial values we then have $N(\psi_L)\supset W_R$ and $N(\psi_R)\supset W_{L,\tau}=\{x|x_3<\tau-|t|\}$. Since $S\subset W_R$ and since $\psi(x)=\psi_L(x)+\psi_R(x)=0$ on S it follows that $\psi_R(x)=0$ on S. Now $S_0\subset S\cup W_{L,\tau}$, and hence $\psi_R(x)=0$ on S_0. By Lemma 4.12 we then have $\psi_R(x)=0$ identically, and hence $\psi(x)=\psi_L(x)$ for all x, which establishes that $N(\psi)\supset S^{cc}=W_R$.

We now consider the case when $\eta(x)$ is an *arbitrary* solution of the Klein–Gordon equation in the set $\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$ which satisfies the condition that $N(\eta)\supset S$. We consider the convolution $\psi(x)=(\eta\star c)(x)$ of $\eta(x)$ with a C_∞-function $c(x)$ with support in the double cone $D_\lambda=D((\mathbf{0},-\lambda),(\mathbf{0},\lambda))$, for a ''small'' $\lambda>0$. The function $\psi(x)$ then satisfies the premises in the example (but with a ''slightly smaller'' null region than the null region of $\eta(x)$). Since λ can be selected as small as we like, we easily arrive at the above result $N(\psi)\supset S^{cc}=W_R$ for the distribution $\eta(x)$.

Example 4.19: Let $\tau>\rho>0$, and let $S=S_u\cup S_l$ be the open subset of M defined by $S_u=\{x|\rho>|t-\tau|+|x_3-\tau|\}$ and $S_l=\{x|\rho>|t+\tau|+|x_3+\tau|\}$. We have $S_u\cap S_l=\varnothing$ and we easily see that $\mathscr{E}(S_u)=S_u$, $\mathscr{E}(S_l)=S_l$, and $\mathscr{E}(S)=S$. However, $S^{cc}=\{x|x_3>|t+\tau|-\tau-\rho, x_3<\tau+\rho-|t-\tau|\}$ is strictly larger than $\mathscr{E}(S)$. We first consider the case when $\psi(x)$ is a C_∞-solution of the Klein–Gordon equation for mass $m>0$ in the space $\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$, such that $N(\psi)\supset S$. We will show that we then have $N(\psi)\supset S^{cc}$. The reasoning has many similarities with the reasoning in Example 4.18, but we now consider *two* ''initial-value planes,'' $\sigma_+=\{(\mathbf{x},\tau)|\mathbf{x}\in\mathsf{R}^3\}$ and $\sigma_-=\{(\mathbf{x},-\tau)|\mathbf{x}\in\mathsf{R}^3\}$. We note that the initial values of $\psi(x)$ on σ_+ vanish when $|x_3-\tau|<\rho$, and that the initial values of $\psi(x)$ on σ_- vanish when $|x_3+\tau|<\rho$. Let $\psi_R(x)$ be the unique solution of the Klein–Gordon equation for m such that the initial values of $\psi_R(x)$ and $\psi(x)$ on σ_+ are equal for $x_3\geqslant\tau-\rho$, and such that the initial values of $\psi_R(x)$ equal zero for $x_3<\tau-\rho$. Let $\psi_L(x)$ be the unique solution of the Klein–Gordon equation for m such that the initial values of $\psi_L(x)$ and $\psi(x)$ on σ_- are equal for $x_3<-\tau+\rho$, and such that the initial values of $\psi_L(x)$ equal zero for $x_3>-\tau+\rho$. We then have $\psi_R(x)=\psi(x)$ and $\psi_L(x)=0$ on $W_+=\{x|x_3>\tau-\rho+|t-\tau|\}$ and $\psi_L(x)=\psi(x)$ and $\psi_R(x)=0$ on $W_-=\{x|x_3<-\tau+\rho-|t+\tau|\}$. We write $\psi_0(x)=\psi(x)-\psi_R(x)-\psi_L(x)$, and by the above we then have $\psi_0(x)=0$ on $W_+\cup W_-$. We easily see that $\psi_0(x)\in\Psi_m$, and it then follows by Proposition 4.15 that $\psi_0(x)=0$ throughout M. Since $\psi_L(x)=0$ for $x_3>|t+\tau|-\tau-\rho$, and since $\psi_R(x)=0$ for $x_3<\tau+\rho-|t-\tau|$, we conclude that $\psi(x)=0$ on S^{cc}.

By the same kind of reasoning as in Example 4.18 we conclude that $N(\eta)\supset S^{cc}$ for an *arbitrary* $\eta(x)\in\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$ which satisfies the condition that $N(\eta)\supset S$.

We note that the conclusions in the above two examples can fail for the *wave equation* ($m=0$), and that they can also fail for solutions of the Klein–Gordon equation for mass $m>0$ which are *not* in the space $\Psi_m\subset\mathscr{S}'_A(\mathsf{M})$.

V. LOCAL ALGEBRAS OF QUANTUM FIELD OPERATORS, AND THEIR WEAK COMMUTANTS

In this section we will discuss the implications of the results in Sections III and IV for quantum field theory. We follow the formulation of field theory in the monograph by Streater and Wightman:[2] see in particular Chapter 3. We denote the field operators by $\varphi_\mu(x)$, where μ is in

some index set $\mathscr{I}$, and we denote a field operator averaged with the test function $f \in \mathscr{S}(\mathsf{M})$ by $\varphi_\mu[f]$. These field operators thus satisfy all the general conditions in Ref. 2. Specifically we assume that the field operators form an irreducible system. We regard the domain D_1 (see Sections 3.1 and 3.2 in Ref. 2) as the domain of the operators $\varphi_\mu[f]$, and the Hermitian conjugate $\varphi_\mu[f]^\dagger$ of the operator $\varphi_\mu[f]$ is the restriction of the adjoint of the operator to D_1. The domain D_1 is dense in the Hilbert space $\mathscr{H}$, it is Poincaré invariant, it contains the (normalized) vacuum vector Ω, and it is mapped into itself by the operators $\varphi_\mu[f]$ and $\varphi_\mu[f]^\dagger$. Let R be any subset of M. We denote by $\mathscr{P}_0(R)$ the smallest unital algebra which contains $\varphi_\mu[f]$ and $\varphi_\mu[f]^\dagger$ for all $\mu \in \mathscr{I}$, and all test functions $f(x)$ with support in R. We thus have $X\mathsf{D}_1 \subset \mathsf{D}_1$ for all $X \in \mathscr{P}_0(R)$. Every $X \in \mathscr{P}_0(R)$ is closable, and we denote by $X^\dagger$ the restriction of X^* to D_1. The algebra $\mathscr{P}_0(R)$ is thus a *-algebra, with Hermitian conjugation ''†'' as the *-operation. The condition that the field operators form an irreducible system is equivalent to the condition that the vacuum vector Ω is cyclic for the algebra $\mathscr{P}_0(\mathsf{M})$, and if this is the case then it follows from the Reeh–Schlieder Principle that Ω is cyclic for $\mathscr{P}_0(R)$ for every $R \subset \mathsf{M}$ with a nonempty interior.

With the above definition of $\mathscr{P}_0(R)$ we have $\mathscr{P}_0(R) = \mathscr{P}_0(R^o)$ for any $R \subset \mathsf{M}$. This follows from the fact that for any closed $K \subset \mathsf{M}$ the space of test functions in $\mathscr{S}(\mathsf{M})$ with support in K is equal to the closure, in the test-function space topology, of the space of all test functions with support in the interior K^o of K. With our definition we thus have $\mathscr{P}_0(R) = \mathsf{C}^1 \cdot I$ if R has an empty interior.

For later reference we here state as a lemma a well-known important property of the domain D_1.

Lemma 5.1: Let $g(p) \in \mathscr{S}(\mathsf{M})$, and let P denote the four-momentum operator. Then $g(P)\mathsf{D}_1 \subset \mathsf{D}_1$, and the linear manifold $\mathsf{D}_c = \mathrm{span}\{g(P)\phi | g \in \mathscr{D}(\mathsf{M}), \phi \in \mathsf{D}_1\}$ is a core for the closure X^{**} of every $X \in \mathscr{P}_0(\mathsf{M})$. Here $\mathscr{D}(\mathsf{M})$ denotes the space of all functions in $\mathscr{S}(\mathsf{M})$ of compact support.

We also have the following result due to Borchers and Zimmermann.[15]

Lemma 5.2: Let S be any nonempty open subset of M. Then the linear manifold $\mathscr{P}_0(S)\Omega$ is a core for the closure X^{**} of every $X \in \mathscr{P}_0(\mathsf{M})$.

The *condition of locality* in a quantum field theory is usually stated in terms of the vanishing of commutators *or* anticommutators of two fields with spacelike separated arguments, depending on whether the fields are bose fields or fermi fields. We here assume what is known as the ''normal connection between spin and statistics.'' Following Bisognano and Wichmann[16] we can express all the locality conditions with reference to *commutators*, as follows. We denote the unitary operator which describes a rotation by angle 2π about any axis by U_0, and we define the unitary operator Z by $Z = (I + iU_0)/(1+i)$. We then have $U_0^2 = I$, $Z^2 = U_o$, $U_0\Omega = \Omega = Z\Omega$, and $U_0\mathsf{D}_1 = \mathsf{D}_1 = Z\mathsf{D}_1$. For any operator Q, and any region $R \in \mathsf{M}$ we write $Q^z = ZQZ^{-1}$, $\mathscr{P}_0(R)^z = \{X^z | X \in \mathscr{P}_0(R)\}$, and the condition of locality for the quantum fields can then be stated as follows. If $X \in \mathscr{P}_0(R_1)$ and $Y \in \mathscr{P}_0(R_2)$, and if R_1 and R_2 are *spacelike separated*, then $[X, Y^z] = 0$ *on the domain* D_1. We refer to operators which commute with the involution U_0 as *boson operators*, and we refer to operators which anticommute with U_0 as *fermion operators*. If X is a boson operator and Y is a fermion operator, we thus have $ZXZ^{-1} = X$ and $ZYZ^{-1} = iU_0Y$. For any $R \subset \mathsf{M}$, every $X \in \mathscr{P}_0(R)$ has a unique representation as a sum of a boson operator X_b and a fermion operator X_f, where $X_b, X_f \in \mathscr{P}_0(R)$.

Definition 5.3: We define the *weak commutant* $\mathscr{P}_0(R)^w$ of the algebra $\mathscr{P}_0(R)$ as the set of all *bounded* operators B which satisfy the condition

$$\langle B^*\Phi' | X\Phi'' \rangle = \langle X^\dagger \Phi' | B\Phi'' \rangle \tag{5.1}$$

for all $\Phi', \Phi'' \in \mathsf{D}_1$ and all $X \in \mathscr{P}_0(R)$. Equivalently stated: $\mathscr{P}_0(R)^w$ is the set of all bounded operators B such that $X^*B \supset BX^\dagger$ for all $X \in \mathscr{P}_0(R)$.

The weak commutant is a weakly and strongly closed linear manifold of operators, closed under the adjoint-operation. Under special *additional* conditions on the system of fields it is a von Neumann algebra,[16] but it is not known whether it is an algebra in general. In this paper we make no assumptions about this point. We will depend on the following simple (and well-known) fact about the weak commutant.

Lemma 5.4: A bounded operator B is in $\mathscr{P}_0(R)^w$ if and only if

$$\langle B^*\Phi'|\varphi_\mu[f]\Phi''\rangle=\langle\varphi_\mu[f]^\dagger\Phi'|B\Phi''\rangle \tag{5.2}$$

for all $\Phi',\Phi''\in\mathsf{D}_c$, all $\mu\in\mathscr{T}$, and all $f(x)\in\mathscr{S}(\mathsf{M})$ with support in R.

We omit the trivial proof: the conclusion follows readily from Lemma 5.1 and the definition of the algebra $\mathscr{P}_0(R)$. The condition in (5.2) is seemingly much weaker than the condition in (5.1) since we only demand that the relation in (5.1) holds for all Φ' and Φ'' in the *smaller* domain D_c, and for the operators in $\mathscr{P}_0(R)$ which are *linear* in the field operators, but the condition in the lemma actually implies the condition in the definition.

With $\Phi',\Phi''\in\mathsf{D}_c$ fixed, and with B a fixed bounded operator, the (linear) mappings $f\to\langle B^*\Phi'|\varphi_\mu[f]\Phi''\rangle$ and $f\to\langle\varphi_\mu[f]^\dagger\Phi'|B\Phi''\rangle$ of $\mathscr{S}(\mathsf{M})$ into C^1 are, for all $\mu\in\mathscr{T}$, tempered distributions. Employing the generalized function notation we can thus define the distribution $A_{\mu;\Phi',\Phi''}(B;x)$ by

$$A_{\mu;\Phi',\Phi''}(B;x)=\langle B^*\Phi'|\varphi_\mu(x)\Phi''\rangle-\langle\varphi_\mu^\dagger(x)\Phi'|B\Phi''\rangle \tag{5.3}$$

for any $\Phi',\Phi''\in\mathsf{D}_c$, any $\mu\in\mathscr{T}$, and any bounded operator B. This distribution is then in the set $\mathscr{S}'_A(\mathsf{M})$, since the vectors Φ' and Φ'' are of compact support in momentum space.[1] *Heuristically* we have $T_{\mu;\Phi'\Phi'}(B;x)=\langle\Phi'|[B,\varphi_\mu(x)]\Phi''\rangle$, but the matrix element of the commutator is not well defined (as it is written) unless $B\Phi'\in\mathsf{D}_1$, which we do not want to assume is necessarily the case.

We can now paraphrase Lemma 5.4 as follows.

Lemma 5.5: A bounded operator B is in $\mathscr{P}_0(R)^w$ if and only if $A_{\mu;\Phi'\Phi''}(B;x)=0$ throughout the interior R^o of R for all $\mu\in\mathscr{T}$, and all $\Phi',\Phi''\in\mathsf{D}_c$.

Looking at the weak commutants from this point of view the results in Sections II and III are immediately applicable to the study of these objects. We first state a lemma which is essentially Araki's result,[1] although here stated in a somewhat different context.

Lemma 5.6: Let $S\subset\mathsf{M}$ be *open*. Let $a(S)$ be the Araki envelope defined in Lemma 2.7. Then

$$\mathscr{P}_0(S)^w=\mathscr{P}_0(a(S))^w. \tag{5.4}$$

Proof: We clearly have $\mathscr{P}_0(S)^w\supset\mathscr{P}_0(a(S))^w$ since $S\subset a(S)$. Let $B\in\mathscr{P}_0(S)^w$. With $A_{\mu;\Phi',\Phi''}(B;x)$ defined as in (5.3) we then have, $A_{\mu;\Phi',\Phi''}(B;x)=0$ on S, and hence, by Lemma 2.7, $A_{\mu;\Phi',\Phi''}(B;x)=0$ on $a(S)$, for all $\Phi',\Phi''\in\mathsf{D}_c$ and all $\mu\in\mathscr{T}$. This is precisely the condition that $B\in\mathscr{P}_0(a(S))^w$. We thus have $\mathscr{P}_0(S)^w\subset\mathscr{P}_0(a(S))^w$, and the identity in (5.4) follows.

We also note the following trivial fact. Let $\{S_\alpha|\alpha\in\mathscr{T}\}$ be any family of *open* subsets of M, and let $S=\cup\{S_\alpha|\alpha\in\mathscr{T}\}$. It follows immediately from Lemma 5.5 that we then have $\mathscr{P}_0(S)^w=\cap\{\mathscr{P}_0(S_\alpha)^w|\alpha\in\mathscr{T}\}$.

On the basis of the important Theorem 2.11 we obtain the following general result.

Theorem 5.7: Let $\{S_\alpha|\alpha\in\mathscr{T}\}$, where $\mathscr{T}$ is some index set, be a family of open regions $S_\alpha\subset\mathsf{M}$ such that S_α is bounded above by the characteristic plane π''_α and below by the characteristic plane π'_α. Let $S=\cup\{S_\alpha|\alpha\in\mathscr{T}\}$. We define the open set $\mathscr{E}(S)$ as in Lemma 2.6. Then

$$\mathscr{P}_0(S)^w=\mathscr{P}_0(a(S))^w=\mathscr{P}_0(\mathscr{E}(S))^w\supset\mathscr{P}_0(S^{cc})^w. \tag{5.5}$$

Proof: By Theorem 2.11 we have $a(S)=\mathscr{E}(S)$ under the stated premises, and in view of (5.4) the first three members in (5.5) are equal. Since $\mathscr{E}(S)\subset(S^{cc})^o$ it obviously follows that $\mathscr{P}_0(\mathscr{E}(S))^w\supset\mathscr{P}_0(S^{cc})^w$. This completes the proof.

We *always* have $\mathscr{P}_0(S)^w=\mathscr{P}_0(a(S))^w\supset\mathscr{P}_0(\mathscr{E}(S))^w\supset\mathscr{P}_0(S^{cc})^w$, since $a(S)\subset\mathscr{E}(S)\subset(S^{cc})^o$, for any open $S\subset\mathsf{M}$. There now emerge two interesting questions. Under what conditions is it the case that $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$, and under what conditions is it the case that $\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w$? The equality $\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w$ trivially implies the equality $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ (for a particular open $S\subset\mathsf{M}$), but the converse is not necessarily true if $\mathscr{E}(S)\neq(S^{cc})^o$, as we will see by an explicit example. We will also see, through examples, that for a given open set S there does not exist any open set R *larger* than $(S^{cc})^o$ such that $\mathscr{P}_0(S)^w=\mathscr{P}_0(R)^w$. Theorem 5.7 states a *sufficient* condition for the equality $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ to hold, and in particular this equality holds if S is a union of wedges in $\mathscr{W}$. The question of for what other sets S this equality may hold is more delicate, and we will return to this issue later.

If the (open) region R is *causally complete*, it is rather natural to regard the operators in $\mathscr{P}_0(R)^w$ as ''locally associated'' with the (closed) causal complement R^c of R. We pursue this idea as follows.

Proposition 5.8: Let K be a closed causally complete subset of M. We define the linear manifold of bounded operators $\mathscr{C}(K)$ by

$$\mathscr{C}(K)=\mathscr{P}_0(K^c)^w. \tag{5.6}$$

(a) The above mapping $K\to\mathscr{C}(K)$ satisfies the condition of *isotony*, i.e., if $K_1,K_2\subset\mathsf{M}$ are closed and causally complete, then $\mathscr{C}(K_1)\supset\mathscr{C}(K_2)$ whenever $K_1\supset K_2$.

(b) The mapping is *Poincaré covariant*, i.e., $U(\lambda)\mathscr{C}(K)U(\lambda)^{-1}=\mathscr{C}(K_{(\lambda)})$, where $U(\lambda)$ is the unitary operator which represents the element λ in the universal covering group of the Poincaré group, and where $K_{(\lambda)}$ is the image of K under the action of λ on Minkowski spacetime.

We omit the trivial proof. A closely related object, denoted $\mathscr{C}_w(K)$, was considered by Bisognano and Wichmann[16] (see Section V in the quoted paper). This object is related to our $\mathscr{C}(K)$ by $\mathscr{C}_w(K)=Z\mathscr{C}(K)Z^{-1}=\mathscr{C}(K)^z$, where Z is defined as above. If the $\mathscr{C}(K)$ happen to be von Neumann algebras (which we do not assume here), then the algebras $\mathscr{C}_w(K)$ are natural objects to consider, because (at least for some subsets K) they can be regarded as ''generated'' by the field operators associated with K. We will not pursue these issues here. We felt that the introduction of the operator Z might just obscure the topics we want to discuss, and hence we consider the objects defined in (5.6). $\mathscr{C}(K)$ is thus simply the weak commutant of the operators in $\mathscr{P}_0(K^c)$.

The following results in TW-I[8] are of interest in connection with the definition of $\mathscr{C}(K)$ in (5.6) (see Proposition 2.6 in the quoted paper). If K is a closed causally complete *proper* subset of M, then K^c is not necessarily open, but if K^c has a *nonempty interior*, then $(K^c)^o$ is ''very close'' to K^c in the sense that $((K^c)^o)^c=K$, and $\overline{(K^c)^o}=\overline{K^c}$. It can also happen that K^c has an empty interior, but this case is rather ''pathological'' in the sense that both K and K^c are then confined to some characteristic plane. This case can only occur if K contains an isotropic line.

We are particularly interested in the objects $\mathscr{C}(K)$ when K is a closed, convex, causally complete set in the family $\mathscr{K}_E$ (which includes all compact, convex, causally complete subsets of M). For *these* sets we have certain *intersection properties*, which we will now discuss. We first state a rather trivial result.

Proposition 5.9: Let $K\in\mathscr{K}_E$, with $\mathscr{K}_E$ as described in Definition 1.1. Then $\mathscr{C}(K)=\cap\{\mathscr{C}(W^c)|W\in\mathscr{W},W^c\supset K\}$.

Proof: By Proposition 3.1 we have $K^c=\cup\{W|W\in\mathscr{W},W^c\supset K\}$, and hence $\mathscr{C}(K)=\mathscr{P}_0(K^c)^w=\cap\{\mathscr{P}_0(W)^w|W\in\mathscr{W},W^c\supset K\}=\cap\{\mathscr{C}(W^c)|W\in\mathscr{W},W^c\supset K\}$, which proves the assertion.

On the basis of Theorem 3.6 and Proposition 3.8 we obtain the following result.

Theorem 5.10: Let $\mathscr{C}(K)$ be defined in terms of K as in Proposition 5.8. Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$. We write $S=\cup\{W|W\in\mathscr{F}\}$ and $K=\cap\{W^c|W\in\mathscr{F}\}=S^c$.

(a) Suppose that the normals of the boundary planes of the wedges in $\mathscr{F}$ span M, and suppose furthermore that $\mathscr{F}$ contains two wedges such that no two of the boundary planes of these wedges are parallel. Then K is an element of the set $\mathscr{K}_E$ as described in Definition 1.1, and we have

$$\mathscr{C}(K)=\cap\{\mathscr{C}(W^c)|W\in\mathscr{F}\}. \tag{5.7}$$

(b) In particular the premises and conclusions in part (a) obtain if $\mathscr{F}$ contains two wedges in a *general relative position*.

(c) Suppose instead that the family $\mathscr{F}$ satisfies the conditions in part (a) or part (b) of Proposition 3.8. Then the relation (5.7) holds.

Proof: By Theorem 5.7 we have $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$, and by Theorem 3.6 we have $\mathscr{E}(S)=S^{cc}$ under the premises in parts (a) or (b), and by Proposition 3.8 we have $\mathscr{E}(S)=S^{cc}$ under the premises in part (c). Hence $\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w=\mathscr{C}(K)$. Since $S=\cup\{W|W\in\mathscr{F}\}$ we have $\mathscr{P}_0(S)^w=\cap\{\mathscr{P}_0(W)^w|W\in\mathscr{F}\}=\cap\{\mathscr{C}(W^c)|W\in\mathscr{F}\}$, and the relation (5.7) follows. This completes the proof.

Note carefully the difference between this theorem and Proposition 5.9. We write $\mathscr{F}_K=\{W|W\in\mathscr{W},W^c\supset K\}$, and by Proposition 5.9 we then have $\mathscr{C}(K)=\cap\{\mathscr{C}(W^c)|W\in\mathscr{F}_K\}$. The

same conclusion follows by Theorem 5.10, since the family $\mathscr{F}_K$ satisfies the premises in part (a) of this theorem. However, for a particular $K\in\mathscr{K}_E$ the family $\mathscr{F}$ in Theorem 5.10 might be *smaller* than the ''maximal family'' $\mathscr{F}_K$, and Theorem 5.10 is therefore much stronger than the rather trivial Proposition 5.9.

One of the goals of this paper is the proof of the following *intersection property* for certain pairs of weak commutants.

Theorem 5.11: Let $\mathscr{C}(K)$ be defined in terms of K as in Proposition 5.8. Let $K_1\in\mathscr{K}_E$, with $\mathscr{K}_E$ as described in Definition 1.1, and let K_2 be closed, convex and causally complete, and such that K_2^c has a nonempty interior. Then $K_1\cap K_2\in\mathscr{K}_E$ and

$$\mathscr{C}(K_1)\cap\mathscr{C}(K_2)=\mathscr{C}(K_1\cap K_2). \tag{5.8}$$

In particular (5.8) holds for any two $K_1,K_2\in\mathscr{K}_E$.

Proof: Let $S_2=(K_2^c)^o$. By Proposition 3.9 we have $K_1\cap K_2\in\mathscr{K}_E$ and $\mathscr{E}(K_1^c\cup S_2)=(K_1^c\cup S_2)^{cc}=(K_1\cap K_2)^c$. Taking into account (5.5) we then have $\mathscr{C}(K_1)\cap\mathscr{C}(K_2)=\mathscr{P}_0(K_1^c)^w\cap\mathscr{P}_0(S_2)^w=\mathscr{P}_0(K_1^c\cup S_2)^w=\mathscr{P}_0(\mathscr{E}(K_1^c\cup S_2))^w=\mathscr{P}_0((K_1\cap K_2)^c)^w=\mathscr{C}(K_1\cap K_2)$. This completes the proof.

We here note the following. Let $\mathscr{T}_1=\{W|W\in\mathscr{W},W^c\supset K_1\}$, $\mathscr{T}_2=\{W|W\in\mathscr{W},W^c\supset K_2\}$, and $\mathscr{T}=\{W|W\in\mathscr{W},W^c\supset K_1\cap K_2\}$. The family $\mathscr{T}$ is in general *larger* than the family $\mathscr{T}_1\cup\mathscr{T}_2$, and hence the relation (5.8) *does not* follow directly from Proposition 5.9, as one might, perhaps, think at first.

This theorem has an obvious generalization to the case of an arbitrary family $\{K_\alpha|\alpha\in\mathscr{I}\}$ of closed, convex, causally complete sets K_α, such that K_α^c has a nonempty interior, and such that at least one of these sets is an element of $\mathscr{K}_E$.

The intersection property in (5.8) *does not* hold generally for two *arbitrary* closed and causally complete sets K_1 and K_2. We show this by a counterexample after the next lemma. This lemma is a useful tool in the manufacture of examples and counterexamples. It is within the theory of free fields, which is our only available source for such things.

Lemma 5.12: We consider the quantum field theory of a single free Hermitian scalar field $\varphi(x)$, associated with a scalar particle of mass $m>0$. Let $\psi(x)$ be a *real* solution of the Klein–Gordon equation (4.1) with the integral representation in (4.3), such that $g_+(\mathbf{p})=g_-(\mathbf{p})^*=g(\mathbf{p})\in L_2(\mathsf{R}^3)$. There then exists a bounded operator B, not equal to a multiple of the identity, such that $B\in\mathscr{P}_0(S)^w$ for an *open* $S\subset\mathsf{M}$ if and only if $S\subset N(\psi)$.

Proof: (1) We have the familiar symbolic representation

$$\varphi(x)=\frac{1}{(2\pi)^{3/2}}\int_{(\infty)}\frac{d^3(\mathbf{p})}{\sqrt{2\omega(\mathbf{p})}}(a^\dagger(\mathbf{p})e^{ix\cdot p}+a(\mathbf{p})e^{-ix\cdot p})$$

of the field, where $p=(\mathbf{p},\omega)$, $\omega=\omega(\mathbf{p})=\sqrt{|\mathbf{p}|^2+m^2}$. The creation and destruction operators $a^\dagger(\mathbf{p})$ and $a(\mathbf{p})$, regarded as operator-valued tempered distributions, satisfy the usual canonical commutation relations on D_1. In discussing the free field it is convenient to regard the field operators as defined on a larger domain than D_1, namely the domain D_N which is the intersection of the domains of all positive powers of the number operator N. For any $g(\mathbf{p})\in L_2(\mathsf{R}^3)$ we write

$$a^\dagger[g]=\int_{(\infty)}d^3(\mathbf{p})a^\dagger(\mathbf{p})g(\mathbf{p}),\quad a[g]=\int_{(\infty)}d^3(\mathbf{p})a(\mathbf{p})g(\mathbf{p})^*=(a^\dagger[g])^\dagger.$$

These symbolic integrals make good sense, and the operators $a^\dagger[g]$ and $a[g]$ are closable operators on D_N which map D_N into D_N in a well-known fashion. The operator $Q[g]=i(a^\dagger[g]-a[g])$ is Hermitian, and we have

$$[iQ[g],\varphi(x)]=\frac{1}{(2\pi)^{3/2}}\int_{(\infty)}\frac{d^3(\mathbf{p})}{\sqrt{2\omega(\mathbf{p})}}(g(\mathbf{p})^*e^{ix\cdot p}+g(\mathbf{p})e^{-ix\cdot p})\cdot I\equiv\psi(x)\cdot I$$

on the domain D_N. The right member is thus a solution $\psi(x)$ of the Klein–Gordon equation in the space Ψ_m^H, times the identity operator I on $\mathscr{H}$.

(2) The operator Q is essentially self-adjoint on D_N (and on D_1), and we can define the unitary Weyl operator B by $B=\exp(iQ[g]^{**})$, where $Q[g]^{**}$ denotes the self-adjoint closure of $Q[g]$. It then follows that $A_{\Phi',\Phi''}(B;x)=\langle B^*\Phi'|\varphi(x)\Phi''\rangle-\langle\varphi(x)\Phi'|B\Phi''\rangle=\psi(x)\langle\Phi'|B\Phi''\rangle$ for any $\Phi',\Phi''\in\mathsf{D}_c$, which means that $B\in\mathscr{P}_0(N(\psi))^w$.

If $\Phi'=\Phi''=\Omega$, we have $A_{\Omega,\Omega}(B;x)=\psi(x)\langle\Omega|B\Omega\rangle$, and since $\langle\Omega|B\Omega\rangle\neq 0$ it follows that $N(A_{\Omega,\Omega}(B;x))=N(\psi(x))$, which implies that if $B\in\mathscr{P}_0(S)^w$ for some open $S\subset\mathsf{M}$, then $S\subset N(\psi)$. This completes the proof.

We now return to the issue of the intersection property in (5.8).

Example 5.13: We consider the particular solution $\psi_1(x)$ of the Klein–Gordon equation constructed in Example 4.10 in Section IV. This solution is of the form in (4.3), with $g_-(\mathbf{p})^*=g_+(\mathbf{p})=g(\mathbf{p})\in\mathscr{S}(\mathsf{R}^3)$, and hence Lemma 5.12 is applicable. The function $\psi_1(x)$ vanishes on $D_u^o\cup D_l^o$, with the notation employed in Example 4.10, but it does not vanish throughout $(D_u^o\cup D_l^o)^{cc}$. With $K_1=(D_u^o)^c$ and $K_2=(D_l^o)^c$ the relation in (5.8) thus fails to hold. In this case the sets K_1 and K_2 are not compact, but this is not an essential feature of the counterexample, as we can show as follows, with reference to Example 4.10. Let $D_4=D(z_4',z_4'')$ be the closed double cone with apices $z_4'=(\mathbf{0},-4\tau)$ and $z_4''=(\mathbf{0},4\tau)$. By the construction of the solution $\psi_1(x)$ in Example 4.10 we then have $\psi_1(x)=0$ on D_4^c, and hence $\psi_1(x)=0$ on $S_1\cup S_2$, with $S_1=D_4^c\cup D_u^o$ and $S_2=D_4^c\cup D_l^o$. The sets $S_1^c=D_4\cap(D_u^o)^c$ and $S_2^c=D_4\cap(D_l^o)^c$ are then *compact* and causally complete. The solution $\psi_1(x)$ does not vanish throughout $(S_1\cup S_2)^{cc}$ since $(S_1\cup S_2)^{cc}\supset\overline{D_l^o}$, and hence we have $\mathscr{C}(S_1^c)\cap\mathscr{C}(S_2^c)\neq\mathscr{C}(S_1^c\cap S_2^c)$. This does not contradict Theorem 5.11, since the sets S_1^c and S_2^c are not in $\mathscr{K}_E$.

We will now consider Araki distributions which are essentially vacuum expectation values of commutators. Let $X\in\mathscr{P}_0(\mathsf{M})$ and let B be a bounded operator. We denote by $T(x)$ the unitary operator which describes a translation by x. We write $X(x)=T(x)XT(x)^{-1}$ for the operator obtained from X by the translation by x, and we consider the expression

$$\eta(X,B;x)=\langle B^*\Omega|X(x)\Omega\rangle-\langle X^\dagger(x)\Omega|B\Omega\rangle=\langle B^*\Omega|T(x)X\Omega\rangle-\langle X^\dagger\Omega|T(-x)B\Omega\rangle. \quad (5.9)$$

The function $\eta(x)=\eta(X,B;x)$ is obviously an infinitely differentiable bounded function of x on M, such that all its partial derivatives are bounded. It can be regarded as an Araki distribution, and we have $\mathrm{supp}(\hat{\eta})\subset\overline{V}(0)$. It is in many respects ''better behaved'' than an arbitrary element of $\mathscr{S}_A'(\mathsf{M})$. One aspect of this is that we have a trivial association of $\eta(x)$ with a solution $\phi(x,s)$ of the wave equation on the five-dimensional space M_5, corresponding to Araki's construction discussed in Section II. This solution is given by

$$\phi(x,s)=\langle B^*\Omega|\cos(s\sqrt{P\cdot P})T(x)X\Omega\rangle-\langle X^\dagger\Omega|\cos(s\sqrt{P\cdot P})T(-x)B\Omega\rangle,$$

where P is the four-momentum operator. The function $\phi(x,s)$ is bounded and infinitely differentiable on M_5, and all its partial derivatives are bounded. We trivially have $\phi(x,-s)=\phi(x,s)$ and $\phi(x,0)=\eta(x)$.

It is well known[17,18] that $T(x)=U(x,I)$ converges weakly to the vacuum projection $|\Omega\rangle\langle\Omega|$ as x tends to infinity along any direction, i.e.,

$$w-\lim_{\lambda\to\infty}T(\lambda u)=|\Omega\rangle\langle\Omega| \quad (5.10)$$

for any nonzero four-vector u. From this it follows that the function $\eta(X,B;x)$ in (5.9) tends to zero as x tends to infinity along any direction, and that $\langle B^*\Omega|X(x)\Omega\rangle$ and $\langle X^\dagger(x)\Omega|B\Omega\rangle$ both tend to $\langle\Omega|B\Omega\rangle\langle\Omega|X\Omega\rangle$.

For a bounded operator B the condition that $\langle B^*\Phi'|X\Phi''\rangle-\langle X^\dagger\Phi'|B\Phi''\rangle=0$ for all $\Phi',\Phi''\in\mathsf{D}_1$ for a *particular* operator X is, of course, much stronger than the condition that $\langle B^*\Omega|X\Omega\rangle-\langle X^\dagger\Omega|B\Omega\rangle=0$. We note, however, the following rather trivial fact.

Proposition 5.14: Let R be a subset of M, and let S_0 be a *nonempty open* subset of R. Let B be a bounded operator such that $B\in\mathscr{P}_0(S_0)^w$ and such that

$$\langle B^*\Omega|X\Omega\rangle-\langle X^\dagger\Omega|B\Omega\rangle=0 \quad (5.11)$$

for all $X \in \mathscr{P}_0(R)$. Then $B \in \mathscr{P}_0(R)^w$.

Proof: Let $X \in \mathscr{P}_0(R)$, and let $Y_1, Y_2 \in \mathscr{P}_0(S_0)$. We then have $Y_1^\dagger X Y_2 \in \mathscr{P}_0(R)$, and hence, by (5.11), $\langle B^*\Omega | Y_1^\dagger X Y_2 \Omega\rangle = \langle Y_2^\dagger X^\dagger Y_1 \Omega | B\Omega\rangle$. Since $B \in \mathscr{P}_0(S_0)^w$ we have $\langle B^*\Omega | Y_1^\dagger X Y_2\Omega\rangle = \langle B^* Y_1\Omega | X Y_2\Omega\rangle$ and $\langle Y_2^\dagger X^\dagger Y_1\Omega | B\Omega\rangle = \langle X^\dagger Y_1\Omega | B Y_2\Omega\rangle$, and it follows that $\langle B^* Y_1\Omega | X Y_2\Omega\rangle = \langle X^\dagger Y_1\Omega | B Y_2\Omega\rangle$, for all $X \in \mathscr{P}_0(R)$ and all $Y_1, Y_2 \in \mathscr{P}_0(S_0)$. By Lemma 5.2 the set $\{Y\Omega | Y \in \mathscr{P}_0(S_0)\}$ is a core for the closure of X, and hence $X^* B^* \supset B^* X^\dagger$ for all $X \in \mathscr{P}_0(R)$, which means that $B \in \mathscr{P}_0(R)^w$, as asserted.

We can think about the relation in (5.11) as saying that the vacuum expectation value of the commutator of B with every operator $X \in \mathscr{P}_0(R)$ vanishes. This by itself does *not* imply that the commutator of B with every operator $X \in \mathscr{P}_0(R)$ vanishes, but this conclusion *does* follow if B satisfies the additional condition that it commutes with all the local operators in $\mathscr{P}_0(S_0)$.

We remark here that all the conclusions in this paper about quantum fields and their weak commutants could have been obtained through a study of the *special* Araki distributions in (5.9). Proposition 5.14 is a key element in such an approach.[7] In addition we need the result that if S is a union of wedges in $\mathscr{W}$, and if $\eta(x)$ is a distribution of the form in (5.9), then $N(\eta) \supset \mathscr{E}(S)$ whenever $N(\eta) \supset S$. One can regard this as a simplification, since the validity of Araki's prescription in Lemma 2.7 is completely trivial for the distributions in (5.9). We felt, however, that the results in Sections II and III, which concern general Araki distributions, are of interest as such. Furthermore the proofs of Theorems 5.7, 5.10, and 5.11 are particularly simple on the basis of the results in Section III.

As we remarked the distributions in (5.9) have special properties which are not shared by arbitrary distributions in $\mathscr{S}'_A(\mathsf{M})$. This permits some interesting conclusions about weak commutants which go beyond what can be obtained on the basis of the results in Section III alone. We first present a result which we believe is widely known and part of the folklore of quantum field theory.

Lemma 5.15: Let S_0 be a nonempty open subset of M, and let u be a forward isotropic vector. We define the ''tube-like'' (open) set S by $S = \{x + su | x \in S_0, s \in \mathsf{R}^1\}$. Then $\mathscr{P}_0(S)^w = \mathsf{C}^1 \cdot I$.

Proof: (1) Let $B \in \mathscr{P}_0(S)^w$ and let $X \in \mathscr{P}_0(S)$. We introduce $B_0 \in \mathscr{P}_0(S)^w$ by $B_0 = B - I \cdot \langle \Omega | B\Omega\rangle$, and hence $\langle \Omega | B_0\Omega\rangle = 0$. We denote $X_u(s) = T(su) X T(su)^{-1}$, where $T(x)$ is the unitary translation operator. Let $f_1(s) = \langle B_0^*\Omega | X_u(s)\Omega\rangle = \langle B_0^*\Omega | T(su) X\Omega\rangle$ and $f_2(s) = \langle X_u(s)^\dagger \Omega | B_0\Omega\rangle = \langle X^\dagger\Omega | T(-su) B_0\Omega\rangle$. We then have $f_1(s) = f_2(s)$ for all real s, since $X(s) \in \mathscr{P}_0(S)$ for all real s. Furthermore it follows from (5.10) that

$$\lim_{|s|\to\infty} f_1(s) = \lim_{|s|\to\infty} f_2(s) = \langle \Omega | B_0\Omega\rangle \langle \Omega | X\Omega\rangle = 0.$$

(2) The functions $f_1(s)$ and $f_2(s)$ are bounded and continuous on R^1, and we can regard them as tempered distributions in $\mathscr{S}'(\mathsf{R}^1)$. It follows from the spectrum condition on the translation group that the support of the Fourier transform of $f_1(s)$ has only the point 0 in common with the support of the Fourier transform of $f_2(s)$, and hence the bounded functions $f_1(s)$ and $f_2(s)$ are independent of s. We thus have $f_1(s) = 0 = f_2(s)$ for all real s, and hence we have

$$\langle B_0^*\Omega | X\Omega\rangle = 0 = \langle X^\dagger\Omega | B_0\Omega\rangle$$

for all $X \in \mathscr{P}_0(S)$. Setting $X = X_1^\dagger X_2$, with $X_1, X_2 \in \mathscr{P}_0(S)$, we obtain $\langle X_1\Omega | B_0 X_2\Omega\rangle = \langle \Omega | B_0 X_1^\dagger X_2\Omega\rangle = 0$, and since this holds for all $X_1, X_2 \in \mathscr{P}_0(S)$ it follows that $B_0 = 0$. Hence B is a complex multiple of the identity. This completes the proof.

This lemma is obviously closely related to Lemma 4.12, but we note the following. In the premises of Lemma 4.12 it was assumed that the support of the Fourier transform $\hat{\eta}(p)$ of $\eta(x)$ is contained in $V(0)$. No such assumption was made about the function $\eta(B, X; x)$ in the present lemma, but instead we depended on the fact in (5.10).

With this lemma we can now obtain a definitive result for the configurations described in Theorem 3.2.

Theorem 5.16: Let $\Pi' = \{\pi'_\alpha | \alpha \in \mathscr{T}'\}$ and $\Pi'' = \{\pi''_\beta | \beta \in \mathscr{T}''\}$, where $\mathscr{T}'$ and $\mathscr{T}''$ are two index sets, be two families of *characteristic* planes in M. For each $\alpha \in \mathscr{T}'$ and $\beta \in \mathscr{T}''$, let $S_{\alpha\beta}$ be the *open* subset of M (possibly empty) which is bounded below by π'_α and bounded above by π''_β,

and let $S=\cup\{S_{\alpha\beta}|\alpha\in\mathscr{T}',\beta\in\mathscr{T}''\}$ and $K=S^c$. The set S is thus the set of all points $x\in\mathsf{M}$ which lie *below* at least one of the planes in Π'' and lie *above* at least one of the planes in Π'. Then

$$\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w=\mathscr{C}(K) \tag{5.12}$$

and the set $K=S^c$ is *closed, convex, and causally complete*.

Proof: (1) If S is empty the conclusion is trivial, and we thus assume that S is nonempty. If an $\alpha\in\mathscr{T}'$ is such that $S_{\alpha\beta}=\varnothing$ for all $\beta\in\mathscr{T}''$, then the index α can be omitted from $\mathscr{T}'$ since it is irrelevant for S. Similarly we can omit any index $\beta\in\mathscr{T}''$ such that $S_{\alpha\beta}=\varnothing$ for all $\alpha\in\mathscr{T}'$, and without any loss of generality we can thus assume that the configuration of characteristic planes is precisely as in Theorem 3.2. According to Theorem 3.2 we either have $\mathscr{E}(S)=(S^{cc})^o$ or else $S^{cc}=\mathsf{M}\neq\mathscr{E}(S)$ and $\mathscr{E}(S)$ contains an open slab-region bounded by two parallel characteristic planes.

(2) By Theorem 5.7 we have $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$. If $\mathscr{E}(S)=(S^{cc})^o$, then the conclusion in (5.12) follows at once. We thus consider the case when $S^{cc}=\mathsf{M}\neq\mathscr{E}(S)$. The set $\mathscr{E}(S)$ then contains a ''tube-region'' S_T of the kind considered in Lemma 5.15, and we have $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w\supset\mathscr{P}_0(S_T)^w=\mathsf{C}^1\cdot I$. Since $K=\varnothing$ we have $\mathscr{C}(K)=\mathscr{C}(\varnothing)=\mathsf{C}^1\cdot I$, and the conclusion in (5.12) holds in this case too. This completes the proof.

This theorem is obviously related to Lemma 4.12, and the remark following the proof of Lemma 5.15 is relevant here too.

We can now state a definitive result for all possible configurations of *two* wedges. In preparation for this we first consider a special case.

Lemma 5.17: Let $W_1,W_2\in\mathscr{W}$ be such that $W_1\cap W_2=\varnothing=W_1^c\cap W_2^c$. Then $\mathscr{C}(W_1^c)\cap\mathscr{C}(W_2^c)=\mathsf{C}^1\cdot I$.

Proof: (1) Under the stated premises the edges of the two wedges must be parallel, and the upper plane of one of the wedges must lie *above* the lower plane of the other. Without any loss of generality we assume that the upper plane of W_1 lies above the lower plane of W_2. The lower plane of W_1 then either coincides with the upper plane of W_2, or else lies above the upper plane of W_2. With a suitable choice of standard coordinates the two wedges can then be described by

$$W_1=\{x|x_3+\rho>|x_4-\tau|\},\quad W_2=\{x|x_3-\rho<-|x_4+\tau|\}$$

for some $\tau>0$, with $\rho=\tau$ if the lower plane of W_1 coincides with the upper plane of W_2, and $\rho=0$ if the lower plane of W_1 lies above the upper plane of W_2. Let $D^o=D(x',x'')^o$ be the open double cone with apices $x'=(\mathbf{0},-\tau/2)$, $x''=(\mathbf{0},\tau/2)$, and let D^o_{-x} be the double cone obtained from D^o by the translation by $-x$. We easily see that $D^o_{-x}\subset W_1\cup W_2$ for all $x\in W_3\cup W_4$, where W_3 and W_4 are the wedges

$$W_3=\{x|x_3>|x_4-\tau|+\tau/2\},\quad W_4=\{x|x_3<-|x_4+\tau|-\tau/2\}.$$

Let $X\subset\mathscr{P}_0(D^o)$ and $B\in\mathscr{P}_0(W_1\cup W_2)^w$. Let $\eta(X,B;x)=\langle B^*\Omega|T(x)X\Omega\rangle-\langle X^\dagger\Omega|T(-x)B\Omega\rangle$, as in (5.9). We then have $\eta(X,B;x)=0$ for all $x\in W_3\cup W_4$. We have $W_3^c\cap W_4^c=\varnothing=W_3\cap W_4$, and we would naturally now like to apply Proposition 4.15, but it is not directly applicable since the support of the Fourier transform $\hat{\eta}(p)$ of $\eta(x)=\eta(X,B;x)$ might *not* be included in any set of the form $\{p|p\in\mathsf{M},p\cdot p\geqslant m^2\}$ for some $m>0$ (whereas it *is* contained in $\overline{V}(0)$). We overcome this difficulty as follows.

(2) Let $m>0$ and let $g(s)\in\mathscr{S}(\mathsf{R}^1)$ be such that the support of the Fourier transform $\hat{g}(q)$ of $g(s)$ *does not* intersect the interval $[-m,m]$. We consider the function $\eta_g(x)$ on M defined by

$$\eta_g(x_1,x_2,x_3,x_4)=\int_{-\infty}^{\infty}ds\,g(s)\,\eta(s,x_2,x_3,x_4). \tag{5.13}$$

Both functions $\eta(x)$ and $\eta_g(x)$ are bounded and infinitely differentiable, and can thus be regarded as distributions in $\mathscr{S}'(\mathsf{M})$. We trivially have $\eta_g(x)=0$ for all $x\in W_3\cup W_4$. Note that

$\eta_g(x)$ is actually independent of x_1. Since $\mathrm{supp}(\hat{\eta}) \subset \overline{V}(0)$ it follows from the definition in (5.13) that $\mathrm{supp}(\hat{\eta}_g) \subset \{p | p_4^2 - p_3^2 - p_2^2 \geq m^2, p_1 = 0\}$, and hence $\mathrm{supp}(\hat{\eta}_g) \subset \{p | p \cdot p \geq m^2\}$. Proposition 4.15 is now applicable to η_g, and we conclude that $\eta_g(x) = 0$ for all x.

(3) Let $h(x_2, x_3, x_4) \in \mathscr{S}(R^3)$, and let

$$\xi(s) = \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{\infty} dx_3 \int_{-\infty}^{\infty} dx_4 \, h(x_2, x_3, x_4) \, \eta(s, x_2, x_3, x_4).$$

The function $\xi(s)$ is bounded and continuous, and can be interpreted as a distribution in $\mathscr{S}'(R^1)$. By the result in step (2) we have $\xi[g] = 0$ for every test function $g(s)$ such that $\hat{g}(q) = 0$ when $q \in [-m, m]$. Since $m > 0$ is arbitrary, it follows that $\mathrm{supp}(\hat{\xi}(q)) = \{0\}$. It follows from (5.10) that the function $\xi(s)$ tends to zero as s tends to infinity, and we can then conclude that $\xi(s) = 0$ for all s, which in turn implies that $\eta(x) = \eta(X, B; x) = 0$ for all $x \in \mathsf{M}$. We have thus shown that $\langle B^* \Omega | X(x) \Omega \rangle - \langle X^\dagger(x) \Omega | B \Omega \rangle = 0$ for all $x \in \mathsf{M}$ and all $X \in \mathscr{P}_0(D^o)$, but this implies that $B \in \mathscr{P}_0(\mathsf{M})^w$, and hence we have $B = c \cdot I$ for some complex c. Since this holds for all $B \in \mathscr{P}_0(W_1 \cup W_2)^w$, we conclude that $\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) = \mathscr{P}_0(W_1 \cup W_2)^w = \mathsf{C}^1 \cdot I$. This completes the proof.

Proposition 5.18: Let $W_1, W_2 \in \mathscr{W}$.

(a) If $W_1 \cap W_2 \neq \varnothing$, then

$$\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) = \mathscr{P}_0(W_1 \cup W_2)^w = \mathscr{P}_0((W_1 \cup W_2)^{cc})^w = \mathscr{C}(W_1^c \cap W_2^c). \tag{5.14}$$

(b) If $W_1^c \cap W_2^c = \varnothing$, then the relations in (5.14) also hold, with $\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) = \mathscr{P}_0(W_1 \cup W_2)^w = \mathsf{C}^1 \cdot I$.

(c) If the span of the normals of the boundary planes of the wedges is either *four-dimensional* or *two-dimensional*, then the relations in (5.14) hold.

(d) If the span of the normals of the boundary planes of the wedges is *three-dimensional*, and if $W_1 \cap W_2 = \varnothing$, then $W_1^c \cap W_2^c \neq \varnothing$, and the intersection property $\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) = \mathscr{C}(W_1^c \cap W_2^c)$ *does not* hold generally.

Proof: (1) We consider part (a), and we thus assume that $W_1 \cap W_2 \neq \varnothing$. By Proposition 3.3 we then have $\mathscr{C}(W_1 \cup W_2) = (W_1 \cup W_2)^{cc} = (W_1^c \cap W_2^c)^c$, and hence, by Theorem 5.7, $\mathscr{P}_0(W_1 \cup W_2)^w = \mathscr{P}_0(\mathscr{C}(W_1 \cup W_2))^w = \mathscr{P}_0((W_1 \cup W_2)^{cc})^w = \mathscr{P}_0((W_1^c \cap W_2^c)^c)^w$, which proves the equalities in (5.14) in this case.

(2) If the span of the normals of the boundary planes is four-dimensional, then the premise and conclusion in part (a) obtain. If the span is three-dimensional, then at least one of the sets $W_1 \cap W_2$ or $W_1^c \cap W_2^c$ is nonempty. Hence, if $W_1^c \cap W_2^c = \varnothing$, then the result in part (a) applies.

Suppose now that the span of the normals is three-dimensional, and that $W_1 \cap W_2 = \varnothing$. In Examples 4.16 and 4.17 we considered such a configuration of wedges, and we gave an example of a real-valued solution $\psi_1(x) \in \Psi_m^H$ of the Klein–Gordon equation for mass $m \geq 0$, with the property that $\psi_1(x) = 0$ on $W_1 \cup W_2$ but *not* throughout $(W_1 \cup W_2)^{cc}$. We consider a theory of a single free Hermitian scalar field of mass m. By Lemma 5.12 there then exists a bounded operator B, not equal to a multiple of the identity, such that $B \in \mathscr{P}_0(W_1 \cup W_2)^w$ but $B \notin \mathscr{P}_0((W_1 \cup W_2)^{cc})^w$. Hence $\mathscr{P}_0(W_1 \cup W_2)^w \neq \mathscr{P}_0((W_1 \cup W_2)^{cc})^w$, or, equivalently, $\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) \neq \mathscr{C}(W_1^c \cap W_2^c)$. We have thus proved the assertions in part (d), and we have also accounted for all the configurations for which the span of the normals is either four- or three-dimensional.

(3) It remains to consider the cases when the span of the normals is two-dimensional *and* $W_1 \cap W_2 = \varnothing$. The upper plane of W_1 is then parallel with the lower plane of W_2, and the upper plane of W_2 is parallel with the lower plane of W_1. We first consider the case when $W_1^c \cap W_2^c \neq \varnothing$. We then have $\mathscr{C}(W_1 \cup W_2) = (W_1 \cup W_2)^{cc} = W_1 \cup W_2$, and the equalities in (5.14) follow readily.

(4) We finally consider the case when $W_1^c \cap W_2^c = \varnothing$ (and also $W_1 \cap W_2 = \varnothing$). By Lemma 5.17 we then have $\mathscr{C}(W_1^c) \cap \mathscr{C}(W_2^c) = \mathsf{C}^1 \cdot I$. Since $(W_1 \cup W_2)^{cc} = \mathsf{M}$, we have $\mathscr{P}_0(W_1 \cup W_2)^w = \mathscr{P}_0((W_1 \cup W_2)^c)^w = \mathscr{P}_0(\mathsf{M})^w = \mathsf{C}^1 \cdot I$, and the equalities in (5.14) obtain. This completes the proof.

The above result that $\mathscr{C}(W_1^c)\cap\mathscr{C}(W_2^c)=C_1\cdot I$ if $W_1^c\cap W_2^c=\varnothing$ is eminently plausible. We might also expect that the conclusion is the same if W_1^c and W_2^c only have a common *edge*, and this is indeed the case. See also the related discussion by Driessler and Summers.[19]

Proposition 5.19: (a) Let $W_1, W_2\in\mathscr{W}$ be such that $\overline{W_2}=W_1^c$, and hence $\overline{W_1}=W_2^c$, i.e., the two wedges W_1 and W_2 have the common edge $E=\overline{W_1}\cap W_1^c=\overline{W_2}\cap W_2^c$. Then $\mathscr{C}(E)=\mathscr{P}_0(W_1\cup W_2)^w=\mathsf{C}^1\cdot I$.

(b) Let $K\subset\mathsf{M}$ be closed and causally complete, and such that K is contained in a *spacelike* 2-plane. Then $\mathscr{C}(K)=\mathsf{C}^1\cdot I$.

Proof: The pair of wedges $\{W_1, W_2\}$ is Poincaré equivalent with the pair $\{W_R, W_L\}$, with $W_R=\{x|x_3>|t|\}$ and $W_L=\{x|x_3<-|t|\}$, and we thus prove the assertion in part (a) for the latter pair. We now depend on some results of Bisognano and Wichmann.[16] (See, in particular, part (c) of Lemma 11, and note that the set $\mathscr{C}_w(W_R)$ in the quoted paper is equal to the set $Z\mathscr{C}(W_R)Z^{-1}$ in the present paper.) For any $B\in\mathscr{C}(W_R)=\mathscr{P}_0(W_L)^w$ we thus have

$$V(i\pi)ZB\Omega=JZB^*\Omega. \tag{5.15}$$

Here $V(i\pi)$ is a self-adjoint operator obtained by an analytic continuation of the operators in the Abelian group $\{V(\lambda)|\lambda\in\mathsf{R}^1\}$ of all velocity transformations in the 3-direction, and J is the antiunitary involution $J=ZU_0\theta_0$ where θ_0 is the *PCT*-operator and U_0 is the operator which represents a rotation by angle 2π about any axis. The operator Z is given by $Z=(I+iU_0)/(1+i)$.

For any $B_L\in\mathscr{C}(W_L)=\mathscr{P}_0(W_R)^w$ we have

$$V(-i\pi)ZB_L\Omega=U_0JZB_L^*\Omega. \tag{5.16}$$

Suppose that $B\in\mathscr{C}(W_R)\cap\mathscr{C}(W_L)$. Then $B=B_L$ satisfies *both* the relations (5.15) and (5.16), and since Z commutes with $V(\lambda)$ and U_0 we have $V(-i\pi)B\Omega=U_0V(i\pi)B\Omega$, which implies that $B\Omega$ is an eigenvector of $U_0V(2\pi i)$ (with eigenvalue $+1$). Since Ω is the only eigenvector of $U_0V(2\pi i)$ it follows that $B\Omega=c\Omega$, with $c=\langle\Omega|B\Omega\rangle$. For any $X_1, X_2\in\mathscr{P}_0(W_L)$ we then have $\langle X_1\Omega|BX_2\Omega\rangle=\langle X_2^\dagger X_1\Omega|B\Omega\rangle=c\langle X_1\Omega|X_2\Omega\rangle$, and we thus conclude that $B=cI$. This proves the assertion in part (a). The conclusion in part (b) follows from the trivial fact that a spacelike 2-plane is always the edge of some wedge.

It can well happen that $\mathscr{C}(K)\neq\mathsf{C}^1\cdot I$ for a closed causally complete set K with an empty interior. Such a set K is necessarily a subset of some characteristic plane if it is also convex. We conclude this section with an example of this kind.

Example 5.20: As in Example 4.6 we define K as the subset $K=\{x|\tau\geqslant x_3=t\geqslant-\tau, x_1^2+x_2^2\leqslant\rho^2\}$ of the characteristic plane $\pi=\{x|x_3=t\}$, with $\tau>0$, $\rho>0$. We then have $K\in\mathscr{K}_C$ and K has an empty interior. Let $\psi(x)$ be the solution of the Klein–Gordon equation, for some $m>0$, constructed in Example 4.6, and let B be the unitary operator related to $\psi(x)$ as in Lemma 5.12. We then have $B\in\mathscr{P}_0(K^c)^w$ since $\psi(x)=0$ on K^c, and the operator B is not a multiple of the identity.

VI. CONCLUDING REMARKS

In our discussion of quantum field theories in Section V two basic questions emerge, as we noted before. Given an open set $S\subset\mathsf{M}$, when do we have the equality $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ and when do we have the equality $\mathscr{P}_0(S)^w=\mathscr{P}_0(S^{cc})^w$? The second equality trivially implies the first, since $S\subset\mathscr{E}(S)\subset(S^{cc})^o$, but as was demonstrated conclusively in Example 5.13, the first equality does not imply the second, and the second equality *cannot* be generally true. As we have also seen, the *intersection property* $\mathscr{C}(K_1)\cap\mathscr{C}(K_2)=\mathscr{C}(K_1\cap K_2)$ for the weak commutants, which might seem plausible at first sight, *cannot* be generally true, for two arbitrary closed causally complete subsets K_1, K_2 of M. In this paper we have shown that one has a rather satisfactory theory for the weak commutants $\mathscr{C}(K)$ when K is in the set $\mathscr{K}_E$, and thus equal to an intersection of closed causal wedges. This is related to the fact that open regions S which are unions of wedges have special features which are not shared by ''arbitrary'' open sets. In particular the relations $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ and $a(S)=\mathscr{E}(S)$ always hold for these special sets. Our demonstration of these facts was based on Araki's theory,[1] which in turn is based on the spectrum condition for the

translation operators, and on the commutation or anticommutation relations expressing the locality condition in quantum field theory. These conditions are very general, and one might regard the input of our theory as rather ''meager.''

We have shown that the intersection property $\mathscr{E}(K_1)\cap\mathscr{E}(K_2)=\mathscr{E}(K_1\cap K_2)$ does hold for compact, convex, causally complete sets K_1 and K_2 (and also a bit more generally), but it is hard to believe that *convexity* is the real essence of the matter. The convexity is, however, essential for our methods, and without additional input we do not expect that one can do much better.

The status of the relation $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ is presently clouded. With the weak commutant defined as in our paper the counterexample of Haag and Schroer[14] shows conclusively that the relation cannot hold generally, since the *time-slice condition* does not hold in their example. The example involves a single generalized free field, and the considerations are closely related to the considerations in our Example 4.7. However, if we enlarge the set of field operators to include *all* (generalized free) fields in the Borchers class of the ''original'' generalized free field, then the counterexample evaporates, and the relation $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ now holds generally, for all open sets S. In view of this one could argue that our definition of the weak commutants is defective: we should first enlarge the set of quantum fields to include all fields in the Borchers class of the ''given'' fields, and then define the weak commutants with respect to this enlarged set. There is some reason to believe that this would not affect the weak commutants $\mathscr{E}(K)$ associated with sets $K\in\mathscr{K}_C$, but that it would in general lead to *smaller* weak commutants $\mathscr{P}_0(R)^w$ for *bounded* sets R. To our knowledge there does not exist, at this time, any convincing counterexample ruling out the general validity of the relation $\mathscr{P}_0(S)^w=\mathscr{P}_0(\mathscr{E}(S))^w$ for such an extended field algebra. This relation is obviously of the nature of a *causality condition*, and one might argue that it ought to be valid in a physically satisfactory theory. It is thus of interest to see how our results would be improved if we *assume* that the equality holds.

Let K_1 and K_2 be two closed, causally complete sets such that K_1^c and K_2^c are open. We then have $\mathscr{P}_0(K_1^c)^w\cap\mathscr{P}_0(K_2^c)^w=\mathscr{P}_0(K_1^c\cup K_2^c)^w$ and $\mathscr{P}_0(K_1^c\cup K_2^c)^w=\mathscr{P}_0(\mathscr{E}(K_1^c\cup K_2^c))^w$ with our additional assumption. As we discussed in Section IV in TW-I,[8] it is the case that $\mathscr{E}(S)=(S^{cc})^o$ for a large class of ''reasonable'' open sets S, although this relation most definitely does not hold generally. We gave the necessary and sufficient conditions on S for the relation to be valid, and we also gave a number of examples for which it does not hold. An obvious absolute requirement is that $\mathscr{E}(S)$ must be *timelike convex*, which means that $\mathscr{E}(S)$ contains the timelike (or isotropic) interval $[x',x'']$ whenever $\mathscr{E}(S)$ contains the endpoints x' and x''. It is easily seen that $\mathscr{E}(S)$ is timelike convex if S is timelike convex. Note that the counterexample in Example 5.13 does involve a case when neither $K_1^c\cup K_2^c$ nor $\mathscr{E}(K_1^c\cup K_2^c)$ is timelike convex. Timelike convexity by itself is not sufficient for the conclusion that $\mathscr{E}(S)=(S^{cc})^o$, but if S is also bounded, then the conclusion follows. (See Lemma 4.7 and Theorem 4.8 in TW-I[8]). Suppose now that the above sets K_1 and K_2 are such that $\mathscr{E}(K_1^c\cup K_2^c)=(K_1^c\cup K_2^c)^{cc}=(K_1\cap K_2)^c$. The intersection property $\mathscr{P}_0(K_1^c)^w\cap\mathscr{P}_0(K_2^c)^w=\mathscr{P}_0((K_1\cap K_2)^c)^w$ then holds for the weak commutants associated with the sets K_1 and K_2. The intersection property would thus hold for all pairs K_1 and K_2 which satisfy the condition that $\mathscr{E}(K_1^c\cup K_2^c)=(K_1^c\cup K_2^c)^{cc}$, which is a very much stronger result than our result in Theorem 5.11. In particular there is no requirement that the sets K_1 or K_2 be convex, nor is there any requirement that these sets be connected.

In the preceding section we have considered *bounded* operators B which commute in the weak sense of Definition 5.3 with the operators in $\mathscr{P}_0(R)$, for some R. It should be clear that very similar conclusions can be drawn about *unbounded* operators Q which commute in some ''weak'' sense with the operators in $\mathscr{P}_0(R)$. If Q is unbounded, attention has to be given to questions of domains. One possibility is that we are given that Q is defined and closable on the domain D_1, in which case we would consider expressions of the form $\langle Q^*\Phi'|X\Phi''\rangle-\langle X^\dagger\Phi'|Q\Phi''\rangle$, or of the form $A_{\mu;\Phi',\Phi''}(Q;x)=\langle Q^*\Phi'|\varphi_\mu(x)\Phi''\rangle-\langle\varphi_\mu^\dagger(x)\Phi'|Q\Phi''\rangle$, as corresponding to matrix elements of commutators. It is not required that Q and $Q^\dagger$ map D_1 into D_1. If we consider operators Q which satisfy the additional conditions that $Q\mathsf{D}_1\subset\mathsf{D}_1$ and $Q^\dagger\mathsf{D}_1\subset\mathsf{D}_1$, then the commutators $[X,Q]$ and $[\varphi_\mu(x),Q]$ make sense on D_1, and the set of all operators Q which commute with all $X\in\mathscr{P}_0(R)$ on D_1 for some $R\subset\mathsf{M}$ form a unital *-algebra. Other possibilities can be considered, with Q and the field operators defined on some ''natural'' domain *larger* than D_1. If the reasoning depends on features such as those described in Lemmas 5.1 or 5.2, one must check that analogous principles are actually valid in the case at hand. We did not want to burden the discussion with

domain questions, and hence we discussed weak commutants consisting of bounded operators. Furthermore weak commutants of this kind are of particular interest in connection with the question of whether the local field operators may have closed extensions affiliated with von Neumann algebras in a local net. In the most favorable case the weak commutants would *equal* these algebras.

[1] H. Araki, ''A Generalization of Borchers' Theorem,'' Helv. Phys. Acta **36**, 132–139 (1963).
[2] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Addison–Wesley, Reading, MA, 1989).
[3] R. Jost, *The General Theory of Quantized Fields* (American Mathematical Society, Providence, RI, 1965).
[4] N. N. Bogolubov, A. A. Logunov and I. T. Todorov, *Introduction to Axiomatic Quantum Field Theory* (Benjamin Cummings, Reading, MA, 1975).
[5] R. Haag, *Local Quantum Physics* (Springer-Verlag, Berlin, 1992).
[6] S. S. Horuzhy, *Introduction to Algebraic Quantum Field Theory* (Kluwer Academic, Dordrecht, Netherlands, 1990).
[7] L. J. Thomas, ''About the Geometry of Minkowski Spacetime and Systems of Local von Neumann Algebras,'' University of California, Berkeley, 1989. Submitted in partial fulfillment of the requirements for the degree of the Ph.D.
[8] L. J. Thomas and E. H. Wichmann, ''On the causal structure of Minkowski spacetime,'' J. Math. Phys. **38**, 5044–5086 (1997).
[9] R. Jost and H. Lehmann, ''Integral-Darstellung kausaler Kommutatoren,'' Nuovo Cimento **V**, 1598–1610 (1957).
[10] F. J. Dyson, ''Integral Representation of Causal Commutators,'' Phys. Rev. **110**, 1460–1464 (1958).
[11] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Vol. II. See Sections 6 and 17 of Chap. VI.
[12] H. J. Borchers, ''Über die Vollständigkeit lorentzinvarianter Felder in einer zeitartigen Röhre,'' Nuovo Cimento **19**, 787–793 (1961).
[13] K. Kraus, ''General quantum field theories and strict locality,'' Z. Phys. **181**, 1–12 (1964).
[14] R. Haag and B. Schroer, ''Postulates of quantum field theory,'' J. Math. Phys. **3**, 248–256 (1962).
[15] H. J. Borchers and W. Zimmermann, ''On the Self-Adjointness of Field Operators,'' Nuovo Cimento **31**, 1047–1059 (1963).
[16] J. J. Bisognano and E. H. Wichmann, ''On the duality condition for quantum fields,'' J. Math. Phys. **17**, 303–321 (1976).
[17] H. J. Borchers ''On Structure of the Algebra of Field Operators,'' Nuovo Cimento **24**, 214–236 (1962).
[18] H. Araki, ''On the Algebra of All Local Observables,'' Prog. Theor. Phys. **32**, 844–854 (1964).
[19] W. Driessler and S. J. Summers, ''Nonexistence of Quantum Fields Associated with Two-Dimensional Spacelike Planes,'' Commun. Math. Phys. **89**, 221–226 (1983).

Erratum: "Covariant algebraic method for calculation of the low-energy heat kernel" [J. Math. Phys. 36, 5055–5070 (1995)]

I. G. Avramidi[a)]
Department of Mathematics, University of Greifswald, Jahnstr. 15a, 17489 Greifswald, Germany

(Received 17 October 1997; accepted for publication 17 October 1997)

[S0022-2488(98)00303-X]

Equation (3.31) in our paper should read

$$a_k=\sum_{1\leqslant N\leqslant k;}\ \sum_{\substack{k_1,\dots,k_N\geqslant 1;\\ k_1+\cdots+k_N=k}}(-1)^{[(3k_1+1)/2]+\cdots+[(3k_N+1)/2]}\frac{k!}{N!k_1!\cdots k_N!}\,b_{k_1}\cdots b_{k_N}\,. \quad (3.31)$$

This misprint does not influence any further equations and results in that paper.

a)Electronic mail: avramidi@rz.uni-greifswald.de

 © 1998 American Institute of Physics

Gauge fixing in the partition function for generalized quantum dynamics

Stephen L. Adler
Institute for Advanced Study, Princeton, New Jersey 08540

(Received 29 September 1997; accepted for publication 10 November 1997)

We discuss the problem of gauge fixing for the partition function in generalized quantum (or trace) dynamics, deriving analogs of the De Witt–Faddeev–Popov procedure and of the BRST invariance familiar in the functional integral context.
© *1998 American Institute of Physics.* [S0022-2488(98)00704-X]

I. INTRODUCTION, AND THE GAUGE THEORY AXIAL GAUGE PARTITION FUNCTION

Generalized quantum (or trace) dynamics[1] is a generalization of classical mechanics, in which the Lagrangian and Hamiltonian are constructed as the trace of multinomials in noncommutative operator or matrix variables. This new form of analytic mechanics has a number of interesting properties. When the fermions are realized in the conventional manner as Grassmann matrices, generalized quantum dynamics incorporates and generalizes all rigid supersymmetry theories, such as[2] the Wess–Zumino and supersymmetric Yang–Mills theories, and also[3] the recently much discussed ''matrix model for M theory.'' In addition, one can formulate a statistical mechanics for generalized quantum dynamics, in both the canonical[4] and microcanonical[5] ensembles. Recent work suggests[4,6] that this statistical mechanics can behave as a prequantum mechanics, with the Heisenberg commutation relations holding for statistical averages over the canonical ensemble of the underlying operator variables.

This prior work on the statistical mechanics of generalized quantum dynamics has assumed either that one is dealing with an unconstrained system, or that the constraints have been explicitly integrated out. In the generic case for an unconstrained system, the canonical ensemble takes the form[4–6]

$$\rho = Z^{-1}\exp(-\tau\mathbf{H}-\mathrm{Tr}\lambda\widetilde{C}),$$

$$Z=\int d\mu\ \exp(-\tau\mathbf{H}-\mathrm{Tr}\lambda\widetilde{C}), \tag{1a}$$

with $d\mu$ the invariant matrix (or operator) phase space measure provided[4] by Liouville's theorem, with $\mathbf{H}$ the conserved total trace Hamiltonian, and with $\widetilde{C}$ the conserved operator

$$\widetilde{C}=\sum_r{}_{\text{bosons}}[q_r,p_r]-\sum_r{}_{\text{fermions}}\{q_r,p_r\}. \tag{1b}$$

In order to apply Eq. (1a) directly to a constrained system, one must first explicitly integrate out the constraints. A simple example where this is possible, that forms the focus of this paper, is provided by the generalized quantum dynamics form of non-Abelian gauge theory,[1] which has a full operator valued gauge invariance group. Let A_ν be an anti-self-adjoint operator or matrix valued gauge potential, let $F_{\mu\nu}$ be the corresponding gauge field strength defined by

$$F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu+[A_\mu,A_\nu], \tag{2a}$$

and let $\mathbf{L}$ be the total trace Lagrangian defined by

$$\mathbf{L}=\frac{1}{4}\mathrm{Tr}F_{\mu\nu}F^{\mu\nu}. \tag{2b}$$

 © 1998 American Institute of Physics

As is well known, Eqs. (2a) and (2b) define a constrained dynamical system. Introducing the canonical momenta

$$p_{A_l} = -F_{0l}, \quad l = 1,2,3, \tag{3a}$$

and the covariant derivative

$$D_\mu X = \partial_\mu X + [A_\mu, X], \tag{3b}$$

the constraint equation for the system of Eqs. (2a) and (2b) takes the form

$$\sum_{l=1}^{3} D_l p_{A_l} = 0. \tag{3c}$$

Since in axial gauge, where $A_3 = 0$, the covariant derivative D_3 simplifies to $D_3 = \partial_3$, the constraint of Eq. (3c) is readily integrated out, giving as the explicit expression for the total trace Hamiltonian,

$$\begin{aligned} \mathbf{H} &= \mathrm{Tr}\left[-\frac{1}{2}\sum_{l=1}^{3} p_{A_l}^2 - \frac{1}{4}\sum_{l,m=1}^{3} F_{lm}^2 \right] \\ &= \int d^3x \frac{1}{2}\mathrm{Tr}\left[-\sum_{l=1}^{2} p_{A_l}^2 - F_{03}^2 - F_{12}^2 - \sum_{l=1}^{2} (\partial_3 A_l)^2 \right], \end{aligned} \tag{4a}$$

with F_{03} given by the line integral

$$F_{03} = \int dz' \frac{1}{2} \frac{z - z'}{|z - z'|} \left(\sum_{l=1}^{2} D_l p_{A_l} \right) \Bigg|_{z'}. \tag{4b}$$

The conserved operator $\tilde{C}$ takes the form, in a general gauge,

$$\tilde{C} = \int d^3x \sum_{l=1}^{3} [A_l, p_{A_l}] = \sum_{l=1}^{3} \left(\int d^3x D_l p_{A_l} - \int dS_l p_{A_l} \right), \tag{5a}$$

with dS_l the surface element of the sphere at infinity. To specialize Eq. (5a) to axial gauge, one sets A_3 equal to zero and evaluates p_{A_3} using Eqs. (3a) and (4b). The axial gauge partition function is then given by Eq. (1a), with the phase space measure $d\mu$ given by

$$d\mu_{\mathrm{axial}} = \prod_{\vec{x}} \prod_{l=1}^{2} dA_l(\vec{x}) dp_{A_l(\vec{x})}. \tag{5b}$$

The problem we wish to address in this paper is how to generalize the axial gauge partition function to other gauges in which it may not be possible to explicitly integrate out the constraint. This is a familiar problem in the theory of path integrals, and we shall use methods similar to the ones employed there to give a solution. However, since the partition function singles out a Lorentz frame, we will have to make a restriction not encountered in the Lorentz scalar path integral case, namely we will consider only nontemporal gauge conditions that do not involve the scalar potential A_0. This will still allow us to consider gauge transformations that rotate the axial gauge axis, or that transform to rotationally invariant gauges such as Coulomb gauge. We will also make the further assumption that the allowed gauge transformations leave invariant the surface integral contribution to $\mathrm{Tr}\tilde{\lambda}\tilde{C}$ arising from the second term on the right hand side of Eq. (5a), which places a restriction on the constant of integration that governs the gauge transformation at the point at infinity. [Although the volume integral in the first term on the right hand side of Eq. (5a) is generically gauge covariant rather than gauge invariant, this causes no problem because its integrand is proportional to the constraint of Eq. (3c).] In Sec. II we develop an analog of the standard

DeWitt–Faddeev–Popov method to write the axial gauge partition function in a general nontemporal gauge, subject to the surface term restriction just stated. Since we have shown[2,3] that trace dynamics incorporates rigid supersymmetry, and since BRST invariance is a particular rigid supersymmetry transformation, it is not surprising that the generalized expression for the partition function, when reexpressed in terms of ghost fermions, admits a BRST invariance, and this is demonstrated in Sec. III. In the analysis that follows, we do not attempt to address the issue of convergence of the partition function, which may well require significant restrictions on the class of trace Hamiltonians being considered.

II. GENERAL NONTEMPORAL GAUGES AND THE DE WITT–FADDEEV–POPOV PROCEDURE

To express the partition function in a general nontemporal gauge, we follow closely the treatment of the De Witt–Faddeev–Popov construction in the familiar functional integral case, as given in the text of Weinberg.[7] Let us consider the integral

$$Z_G = \int d\mu B[f(A_l)]\,\delta(D)\det \mathscr{F}[A_l]\exp(-\pi\mathbf{H} - \mathrm{Tr}\tilde{\lambda}\tilde{C}),$$

$$d\mu = \prod_{\vec{x}} \prod_{l=1}^{3} dA_l(\vec{x}) dp_{A_l(\vec{x})}, \tag{6a}$$

with $\mathbf{H}$ given by the first line of Eq. (4a) [which is valid in a general gauge on the constraint surface $D=0$ selected by the delta function in Eq. (6a)], with $\tilde{C}$ given by Eq. (5a), and with D defined by $D \equiv \Sigma_{l=1}^{3} D_l p_{A_l}$. The delta function of the anti-self-adjoint matrix valued argument D appearing in Eq. (6a) is given, in terms of ordinary delta functions of the real (R) and imaginary (I) parts of the matrix elements, by

$$\delta(D) = \prod_{m<n} \delta((D_R)_{mn}) \prod_{m\leqslant n} \delta((D_I)_{mn}), \tag{6b}$$

and the integration measure over the anti-self-adjoint matrix A_l is defined by

$$dA_l = \prod_{m<n} d(A_{lR})_{mn} \prod_{m\leqslant n} d(A_{lI})_{mn}, \tag{6c}$$

and similarly for dp_{A_l}. The function $B[f]$ is an arbitrary integrable scalar valued function of the matrix valued argument $f(A_l)$, which is used to specify the gauge condition. We shall treat f as a column vector f_α with α a composite index formed from the matrix row and column indices m,n; the argument $\mathscr{F}$ of the De Witt–Faddeev–Popov determinant is then given in terms of f by the expression

$$\mathscr{F}_{\alpha\vec{x},\beta\vec{y}}[A_l] \equiv \left.\frac{\delta f_\alpha(A_l(\vec{x}) + D_l\Lambda(\vec{x}))}{\delta\Lambda_\beta(\vec{y})}\right|_{\Lambda=0}, \tag{6d}$$

where δ is the usual functional derivative and β is the composite of the row and column indices of the infinitesimal gauge transformation matrix Λ.

We now demonstrate two properties of the integral Z_G defined in Eq. (6a): (i) first, we show that when the gauge fixing functions $B[f]$ and $f(A_l)$ are chosen to correspond to the axial gauge condition, then Eq. (6a) reduces (up to an overall constant) to the axial gauge partition function of Sec. I; (ii) second, we show that Z_G is in fact independent of the function $f(A_l)$, and depends on the function $B[f]$ only through an overall constant. These two properties together imply that Z_G gives the wanted extension of the axial gauge partition function to general nontemporal gauges.

To establish property (i), we make the conventional axial gauge choice

$$B[f(A_l)]=\delta(A_3)=\prod_{m<n}\delta((A_{3R})_{mn})\prod_{m\leqslant n}\delta((A_{3I})_{mn}) \tag{7a}$$

so that

$$\int dA_3 B[f(A_l)]=\int dA_3\delta(A_3)=1. \tag{7b}$$

With this gauge choice,

$$D_3 p_{A_3}=\partial_3 p_{A_3}, \tag{7c}$$

which implies that

$$\begin{aligned}\delta(D)&=\delta\left(\partial_3 p_{A_3}+\sum_{l=1}^{2} D_l p_{A_l}\right)\\&=|\partial_3|^{-1}\delta\left(p_{A_3}+\int^z dz'\sum_{l=1}^{2} D_l p_{A_l}\right);\end{aligned} \tag{8a}$$

hence the integral over p_{A_3} in Z_G can be done explicitly, giving (up to an overall constant factor coming from the Jacobian $|\partial_3|^{-1}$) the expression

$$Z_G=\int d\mu_{\text{axial}}\ \exp\ (-\tau\mathbf{H}-\mathrm{Tr}\tilde{\lambda}\tilde{C})|_{A_3=0;p_{A_3}=-\int^z dz'\Sigma_{l=1}^{2}D_l p_{A_l}}, \tag{8b}$$

which agrees with the axial gauge partition function of Sec. I.

To establish property (ii), we first examine the gauge transformation properties of the various factors in the integral defining Z_G. We begin with the integration measure $d\mu$. Under the infinitesimal gauge transformation (with Λ anti-self-adjoint)

$$A_l\rightarrow A_l+D_l\Lambda=A_l+\partial_l\Lambda+[A_l,\Lambda], \tag{9a}$$

the inhomogeneous term $\partial_l\Lambda$ does not contribute to the transformation of the differential dA_l, and so dA_l obeys the homogeneous transformation law $dA_l\rightarrow dA_l+\Delta_l$, with

$$\Delta_l\equiv[dA_l,\Lambda]. \tag{9b}$$

Hence to first order in Λ, the Jacobian of the transformation is

$$\begin{aligned}J&=1+\sum_{m<n}\frac{\partial(\Delta_{lR})_{mn}}{\partial(dA_{lR})_{mn}}+\sum_{m\leqslant n}\frac{\partial(\Delta_{lI})_{mn}}{\partial(dA_{lI})_{mn}}\\&=1+\left(\sum_{m<n}+\sum_{m\leqslant n}\right)[(\Lambda_R)_{nn}-(\Lambda_R)_{mm}]=1,\end{aligned} \tag{9c}$$

since the anti-self-adjointness of Λ implies that $(\Lambda_R)_{nm}=-(\Lambda_R)_{mn}$, and so the diagonal matrix elements $(\Lambda_R)_{nn}$ are all zero. Thus each factor $dA_l(\vec{x})$ in the integration measure is gauge invariant. A similar argument applies to each factor $dp_{A_l(\vec{x})}$ in the integration measure, and also to the factor $\delta(D)$ in the integrand, since D obeys the homogeneous gauge transformation law $D\rightarrow D+[D,\Lambda]$. Turning to the exponential, the terms $\mathrm{Tr}p_{A_l}^2$ and $\mathrm{Tr}F_{lm}^2$ are gauge invariant, and so the trace Hamiltonian $\mathbf{H}$ is gauge invariant. From Eq. (5a), the volume integral term in $\tilde{C}$ vanishes by virtue of the integrand factor $\delta(D)$, and so $\mathrm{Tr}\tilde{\lambda}\tilde{C}$ receives a contribution only from the surface term in Eq. (5a), which by hypothesis is left invariant by the class of gauge transformations under consideration. In sum, we see that the integral Z_G has the form

$$Z_G = \int d\mu \mathcal{G}[A_l] B[f(A_l)] \det \mathcal{F}[A_l], \tag{10a}$$

with the integration measure $d\mu$ and the integrand factor

$$\mathcal{G}[A_l] = \delta(D) \exp(-\tau \mathbf{H} - \mathrm{Tr} \tilde{\lambda} \tilde{C}) \tag{10b}$$

both gauge invariant. Hence Z_G has exactly the form assumed in the discussion of Weinberg,[7] and the proof given there completes the demonstration of property (ii).

III. GHOSTS AND BRST INVARIANCE OF THE GENERALIZED PARTITION FUNCTION

Let us continue to follow the standard path integral analysis, and represent the De Witt–Faddeev–Popov determinant $\det \mathcal{F}[A_l]$ as an integral over fermionic ghosts, by writing

$$\det \mathcal{F}[A_l] = \int d\omega^* d\omega \exp\left(\int d^3x d^3y \omega^*_\alpha(\vec{x}) \mathcal{F}_{\alpha\vec{x},\beta\vec{y}}[A_l] \omega_\beta(\vec{y}) \right), \tag{11a}$$

with

$$d\omega = \prod_{\vec{x}} \prod_{m,n} d\omega_{mn}(\vec{x}), \quad d\omega^* = \prod_{\vec{x}} \prod_{m,n} d\omega^*_{mn}(\vec{x}). \tag{11b}$$

Let us also take for $B[f]$ the usual Gaussian

$$B[f] = \exp\left(-\frac{1}{2\xi} \int d^3x \mathrm{Tr} f(A_l(\vec{x}))^2 \right), \tag{11c}$$

and for $f(A_l)$ the linear gauge condition

$$f(A_l) = \sum_l L^l A_l, \tag{11d}$$

in which L^l can be either a fixed vector (such as δ_{l3} in axial gauge) or a differential operator (such as ∂_l in Coulomb gauge), and a summation of l from 1 to 3 is understood. With this choice of $f(A_l)$, we find from Eq. (6d) that

$$\begin{aligned} \mathcal{F}_{nm\vec{x},pq\vec{y}}[A_l] &= \frac{\delta f_{nm}(A_l(\vec{x}) + D_l \Lambda(\vec{x}))}{\delta \Lambda_{pq}(\vec{y})} \\ &= \sum_l L^l_{\vec{x}} \left(\frac{\partial \delta(\vec{x}-\vec{y})}{\partial x^l} \delta_{np}\delta_{mq} + \delta(\vec{y}-\vec{y})[(A_l)_{np}\delta_{mq} - \delta_{np}(A_l)_{qm}] \right), \end{aligned} \tag{12a}$$

which when substituted into the exponent in Eq. (11a) gives, after integrations by parts,

$$\int d^3x d^3y \omega^*_\alpha(\vec{x}) \mathcal{F}_{\alpha\vec{x},\beta\vec{y}}[A_l] \omega_\beta(\vec{y})) = \int d^3x \mathrm{Tr} \bar{\omega}(\vec{x}) \sum_l L^l D_l \omega(\vec{x}), \tag{12b}$$

where we have defined $\bar{\omega}_{mn} = \omega^*_{nm}$. Hence the expression of Eq. (10a) for Z_G becomes

$$Z_G = \int d\mu d\bar{\omega} d\omega \mathcal{G}[A_l] \exp\left[-\int d^3x \mathrm{Tr} \left(\frac{1}{2\xi} \left(\sum_l L^l A_l \right)^2 - \bar{\omega}(\vec{x}) \sum_l L^l D_l \omega(\vec{x}) \right) \right]. \tag{13a}$$

An alternative way of writing Eq. (13a), that is convenient for exhibiting the BRST invariance, is to introduce an auxilliary self-adjoint matrix field h and to reexpress Eq. (13a) as

$$Z_G=\int d\mu dh d\bar{\omega} d\omega \mathcal{G}[A_l]\exp\left[-\int d^3x \mathrm{Tr}\left(\frac{\xi}{2}h^2+ih\sum_l L^l A_l-\bar{\omega}(\vec{x})\sum_l L^l D_l\omega(\vec{x})\right)\right]. \tag{13b}$$

Starting from Eq. (13b), we can now show that Z_G has a BRST invariance of the familiar form. Let θ be an $\vec{x}$-independent c-number Grassmann parameter (i.e., a 1×1 Grassmann matrix), and consider the variations defined by

$$\begin{aligned}\delta\omega&=\omega^2\theta,\\ \delta A_l&=D_l\omega\theta,\\ \delta\bar{\omega}&=-ih\theta,\\ \delta h&=0.\end{aligned} \tag{14}$$

We begin by showing that Eq. (14) defines a nilpotent transformation, in the sense that the second variations of all quantities are zero. To verify this, we show that the variations of ω^2 and $D_l\omega$ are zero (the variations of h and of 0 are trivially 0), as follows:

$$\begin{aligned}\delta\omega^2&=\{\delta\omega,\omega\}=\{\omega^2\theta,\omega\}=\omega^2\{\omega,\theta\}=0,\\ \delta D_l\omega&=[\delta A_l,\omega]+D_l\delta\omega=[D_l\omega\theta,\omega]+D_l\omega^2\theta=-\{D_l\omega,\omega\}\theta+\{D_l\omega,\omega\}\theta=0.\end{aligned} \tag{15a}$$

To see that Z_G is invariant, we note that the action on A_l of the BRST transformation of Eq. (14) is just a gauge transformation (albeit with a Grassmann valued parameter), and so the gauge invariance analysis of Sec. II shows that the factors $d\mu$ and $\mathcal{G}[A_l]$ are invariant. The measure dh is trivially invariant, and the measure $d\bar{\omega}$ is invariant because $\delta\bar{\omega}$ has no dependence on $\bar{\omega}$. Since

$$\delta(d\omega)=d(\delta\omega)=d(\omega^2\theta)=(\omega d\omega+d\omega\omega)\theta, \tag{15b}$$

we have

$$(\delta(d\omega))_{mn}=(\omega_{mm}d\omega_{mn}+d\omega_{mn}\omega_{nn})\theta+\cdots=d\omega_{mn}(\omega_{nn}-\omega_{mm})\theta+\cdots, \tag{15c}$$

with $\ldots$ denoting terms that contain matrix elements $d\omega_{m'n'}$ with $(m',n')\neq(m,n)$. Hence the Jacobian of transformation for $d\omega$ differs from unity by a term proportional to

$$\sum_{nm}(\omega_{nn}-\omega_{mm})\theta=0, \tag{15d}$$

and so the measure $d\omega$ is also invariant. Hence to complete the demonstration that Z_G is BRST invariant, we have to show that the gauge fixing part of the Hamiltonian,

$$\mathbf{H}_G\equiv\int d^3x \mathrm{Tr}\left(\frac{\xi}{2}h^2+ih\sum_l L^l A_l-\bar{\omega}\sum_l L^l D_l\omega\right), \tag{16a}$$

is BRST invariant. Since we have already seen that $D_l\omega$ is invariant, and since h is trivially invariant, we have only to verify that

$$0=\int d^3x \mathrm{Tr}\left[ih\sum_l L^l\delta A_l-(\delta\bar{\omega})\sum_l L^l D_l\omega\right]=\int d^3x \mathrm{Tr} ih\sum_l L^l D_l\{\omega,\theta\}, \tag{16b}$$

which checks, completing the demonstration of BRST invariance of the generalized partition function.

ACKNOWLEDGMENTS

I wish to thank Lowell Brown for helpful conversations, and to acknowledge the hospitality of the Aspen Center for Physics, where this work was done. This work was supported in part by the Department of Energy under Grant No. DE-FG02-90ER40542.

[1] S. L. Adler, Nucl. Phys. B **415**, 195 (1994); S. L. Adler, *Quaternionic Quantum Mechanics and Quantum Fields*, Secs. 13.5–13.7 (Oxford University Press, New York, 1995).
[2] S. L. Adler, hep-th/9703132, Nucl. Phys. B **499**, 569 (1997).
[3] S. L. Adler, hep-th/9703053, Phys. Lett. B **407**, 229 (1997).
[4] S. L. Adler and A. C. Millard, Nucl. Phys. B **473**, 199 (1996).
[5] S. L. Adler and L. P. Horwitz, J. Math. Phys. **37**, 5429 (1996).
[6] S. L. Adler and A. Kempf, preprint IASSNS-HEP-97/51, DAMTP-97-99, hep-th 9709106.
[7] S. Weinberg, *The Quantum Theory of Fields*, Vol. II, Sec. 15.5 (Cambridge University Press, Cambridge, 1996).

Geometric approach to inverse scattering for hydrogen-like systems in a homogeneous magnetic field

Silke Arians[a)]
Institut für Reine und Angewandte Mathematik, RWTH Aachen, D-52056 Aachen, Germany

(Received 23 June 1997; accepted for publication 8 January 1998)

We consider the Hamiltonian of one quantum particle in a homogeneous magnetic field and a scalar potential in three space dimensions. For a given magnetic field the high velocity limit of the scattering operator uniquely determines the scalar potential if it is of short range. If, in addition, long-range potentials are present, some knowledge of (the far out tail of) the long-range part is needed to define a modified Dollard wave operator and a scattering operator. Again its high velocity limit uniquely determines the total scalar potential for a given magnetic field. We generalize our results to a system of two interacting quantum particles with opposite electric charges. © *1998 American Institute of Physics.* [S0022-2488(98)02704-2]

I. INTRODUCTION

We study a system of one quantum particle in a homogeneous magnetic field in $\nu=3$ dimensions. As an example of such a one-body system we mention a hydrogen atom with an infinitely heavy proton. The interactions are given by a vector potential $\mathbf{A}$ with $\mathbf{rotA}=\mathbf{B}$ and a scalar potential V. We chose the coordinates $(x,y,z)\in\mathbb{R}^3$ such that the magnetic field is pointed along the z-axis, i.e., $\mathbf{B}=(0,0,b)$ for $b>0$ constant. In a transversal gauge a corresponding vector potential is

$$\mathbf{A}(\mathbf{r})=(-by/2,bx/2,0) \text{ for a vector } \mathbf{r}=(x,y,z)\in\mathbb{R}^3. \tag{1.1}$$

A classical particle with mass m, charge e and velocity $\mathbf{v}$ in a constant magnetic field $\mathbf{B}$ moves in helixes with the Larmor radius

$$\mathbf{R}(t)=-\frac{m}{e}\frac{\mathbf{v}(t)\times\mathbf{B}}{\mathbf{B}^2}, \tag{1.2}$$

i.e., $|\mathbf{R}(t)|=(\mathbf{v}_x^2(t)+\mathbf{v}_y^2(t))^{1/2}m|e|/b$ and the radius is proportional to the velocity in the (x,y)-direction. Then in the high velocity limit the radius tends to infinity, such that the particle moves in a straight line. The Larmor frequency is

$$\boldsymbol{\omega}=-\frac{e\mathbf{B}}{m}, \tag{1.3}$$

such that the minimal period of the particle $|T|=|2\pi/\boldsymbol{\omega}|=2\pi m/(|e|b)$ is independent of the velocity. Thus for increasing velocity the particle moves in circles with increasing radius, but needs the same time for each circle. Therefore we have to assume that the particle moves parallel to the magnetic field with velocity $|\mathbf{v}_z|>0$, such that the potential is hit only once.

Let $H_0:=(1/2m)\,\mathbf{p}^2$ be the free Hamiltonian. By

$$H_A:=\frac{1}{2m}(\mathbf{p}-\mathbf{A}(\mathbf{r}))^2 \tag{1.4}$$

a)Electronic mail: arians@iram.rwth-aachen.de

 © 1998 American Institute of Physics

we denote the Hamiltonian with the magnetic field alone and by $H=H_A+V$ we denote the interacting Hamiltonian with mass m, charge $e=1$, momentum $\mathbf{p}=-i\nabla_{\mathbf{r}}$ and position $\mathbf{r}$. We abbreviate the first and second components of the position $\mathbf{r}$ with $\boldsymbol{\rho}$ and of the momentum $\mathbf{p}$ with $\boldsymbol{\pi}$:

$$\boldsymbol{\rho}:=(x,y) \text{ and } \boldsymbol{\pi}:=(\mathbf{p}_x,\mathbf{p}_y). \tag{1.5}$$

It is useful to introduce the following representation (see Ref. 1 for the notation):

$$H_A=H_0+\frac{1}{2m}\left\{\frac{b^2}{4}\left(\frac{\mathbf{B}}{|\mathbf{B}|}\times\mathbf{r}\right)^2-\mathbf{B}\mathbf{L}\right\} \tag{1.6}$$

$$=H_0+\frac{1}{2m}\left\{\frac{b^2}{4}(x^2+y^2)-b\mathbf{L}_z\right\} \tag{1.7}$$

$$=\frac{1}{2m}\mathbf{p}_z^2+H_{osc}-\frac{1}{2m}b\mathbf{L}_z \tag{1.8}$$

with the z-component of the angular momentum $\mathbf{L}_z=x\mathbf{p}_y-y\mathbf{p}_x$ and the harmonic oscillator

$$H_{osc}=\frac{1}{2m}\left\{(\mathbf{p}_x^2+\mathbf{p}_y^2)+\frac{b^2}{4}(x^2+y^2)\right\}. \tag{1.9}$$

We consider H acting on $\mathscr{H}=L^2(X)$ for $X=\mathbb{R}^3$. It is known (see Ref. 2 for details) that $H_{xy}=H_{osc}-(1/2m)\,b\mathbf{L}_z$, acting on the Hilbert space $L^2(\mathbb{R}^2)$, has a complete set of eigenfunctions $\{\Theta_{ln}(x,y)\}_{l=0,\pm1,\ldots;n=0,1,\ldots}$ with corresponding eigenvalues

$$E_{ln}=(2m)^{-1}b(2n+|l|-l+1), \tag{1.10}$$

where E_{ln} and Θ_{ln} are often referred to as Landau energy levels and Landau orbits. As H_{xy} does not act on the z-variable, we have $\exp[-iH_At]=\exp[-iH_{xy}t]\exp[-i\mathbf{p}_z^2/(2m)]$.

The scalar potential can be split into a short-range and a long-range part: $V=V^s+V^l$. V^s is assumed to be short range: $V^s\in\mathscr{V}_{SR}$ with

$$\mathscr{V}_{SR}:=\{V^s\,|\,|V^s(\mathbf{r})|\leqslant\text{const}(1+|\mathbf{r}|)^{-(1+\varepsilon)}\}. \tag{1.11}$$

We mention that our results remain valid for unbounded short-range potentials (see Section VI).

Here V^l is assumed to be long range: $V^l\in\mathscr{V}_{LR}$ with

$$\mathscr{V}_{LR}:=\{V^l\,|\,V^l\in C^1(\mathbb{R}^3), V^l \text{ tends to zero for } |\mathbf{r}|\to\infty \text{ and}$$

$$|\nabla V^l(\mathbf{r})|\leqslant\text{const}(1+|\mathbf{r}|)^{-\gamma} \text{ with } \gamma>3/2\}. \tag{1.12}$$

The splitting into short- and long-range parts is not unique. But without loss of generality we can make it in such a way that in addition

$$|D^\alpha V^l(\mathbf{r})|\leqslant\text{const}(1+|\mathbf{r}|)^{-1-|\alpha|(\delta+1/2)} \tag{1.13}$$

for $1\leqslant|\alpha|\leqslant4$, $0<\delta<1/2$ (see Ref. 3).

We remark that for the proof of the asymptotics of the scattering operators (1.17), (1.24) and (1.25) we only need decay of the potentials in the z-direction. The decay in all directions is needed for the reconstruction of the scalar potentials.

Under these assumptions the interacting Hamiltonian is self-adjoint on its domain $\mathscr{D}(H)=\mathscr{D}(H_A)=W^{2,2}(X)$ in $\mathscr{H}=L^2(X)$.

In the special case, when V is of short range the following wave operators, which measure for a given $\mathbf{A}$ the additional effect on scattering due to V^s,

$$\Omega_\pm:=\Omega_\pm(\mathbf{A}):=\underset{t\to\pm\infty}{\text{s-lim}}\ e^{iHt}e^{-iH_At} \tag{1.14}$$

exist and are complete (see Ref. 2, Chapter 4). The scattering operator S is

$$S:=S(\mathbf{A};V^s):=(\Omega_+)^*\Omega_-. \tag{1.15}$$

As wave functions we chose asymptotic configurations $\Phi_0\in\mathscr{H}=L^2(X)$ with compact momentum support in the ball around the origin with radius $m\eta$, $\eta>0$, and momentum space wave function $\hat{\phi}_0\in C_0^\infty(B_{m\eta}(0))$. These functions are dense in $\mathscr{H}$ and we can find a function $g\in C_0^\infty(\mathbb{R}^3)$ with $\Phi_0=g(\mathbf{p})\Phi_0$ and especially a function $f\in C_0^\infty(\mathbb{R})$ with $\Phi_0=f(\mathbf{p}_z)\Phi_0$.

By $\Phi_\mathbf{v}$ we denote the boosted configuration translated by $m\mathbf{v}$ in momentum space, where $\mathbf{v}=(\mathbf{v}_x,\mathbf{v}_y,\mathbf{v}_z)$ is the velocity:

$$\Phi_\mathbf{v}=e^{im\mathbf{v}\mathbf{r}}\Phi_0\Leftrightarrow\hat{\phi}_\mathbf{v}(\mathbf{p})=\hat{\phi}_0(\mathbf{p}-m\mathbf{v}),\quad \mathbf{v}_z\neq 0, \tag{1.16}$$

where $\Phi_\mathbf{v}$ has compact velocity support in $B_\eta(\mathbf{v})$. Let $v:=|\mathbf{v}|$. We will obtain asymptotics of S for the high velocity limit in an arbitrary direction $\hat{\mathbf{v}}=(\hat{\mathbf{v}}_x,\hat{\mathbf{v}}_y,\hat{\mathbf{v}}_z):=\mathbf{v}/v$, $v\to\infty$ with $\hat{\mathbf{v}}_z\neq 0$.

Regarding S as a mapping from the set of short-range potentials $\mathscr{V}_{SR}$ into the set of bounded operators $L(\mathscr{H})$ we have the following:

Theorem 1: Suppose that $V^s\in\mathscr{V}_{SR}$ and $V^l=0$. *Then the following reconstruction formula holds for all* $\Phi_\mathbf{v},\Psi_\mathbf{v}$ *as in Eq. (1.16):*

$$\lim_{v\to\infty} iv((S-\mathbb{1})\Phi_\mathbf{v},\Psi_\mathbf{v})=\int_{-\infty}^{\infty}(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)d\tau. \tag{1.17}$$

In particular, the scattering map

$$S(\mathbf{A};\cdot):\mathscr{V}_{SR}\to L(\mathscr{H})$$

is injective.

The modified Dollard wave operators

$$\Omega_\pm^D:=\Omega_\pm^D(\mathbf{A}):=\operatorname*{s-lim}_{t\to\pm\infty} e^{iHt}U^D(t) \tag{1.18}$$

with

$$U^D(t):=\exp\left[-iH_At-i\int_0^t V^l(\hat{\boldsymbol{\rho}}_\mathbf{v}(s),s\mathbf{p}_z/m)ds\right] \tag{1.19}$$

where

$$\hat{x}_\mathbf{v}(s):=\omega^{-1}\{\mathbf{v}_x\sin(\omega s)+\mathbf{v}_y(1-\cos(\omega s))\}, \tag{1.20}$$

$$\hat{y}_\mathbf{v}(s):=\omega^{-1}\{\mathbf{v}_y\sin(\omega s)-\mathbf{v}_x(1-\cos(\omega s))\} \tag{1.21}$$

and

$$\hat{\boldsymbol{\rho}}_\mathbf{v}(s):=(\hat{x}_\mathbf{v}(s),\hat{y}_\mathbf{v}(s)) \tag{1.22}$$

with $\omega:=|\boldsymbol{\omega}|=b/m$ exist and are complete (see Ref. 1, Theorem 7.1) for a Dollard correction with $\hat{\boldsymbol{\rho}}_\mathbf{v}=0$. We remark that we introduced a velocity dependent correction (1.22) which corresponds to the classical behavior of a particle in a harmonic oscillator:

$$S^D:=S^D(\mathbf{A},V^l;V^s):=(\Omega_+^D)^*\Omega_-^D \tag{1.23}$$

is the scattering operator S^D. Since the splitting of V into a short- and a long-range part is not unique the Dollard wave operators are not uniquely defined. Nevertheless, the scalar potential is uniquely obtained from *any* of the scattering operators S^D.

Theorem 2: *Suppose that* $V^s \in \mathscr{V}_{SR}$ *and* $V^l \in \mathscr{V}_{LR}$. *Then the following reconstruction formula holds for all* $\Phi_{\mathbf{v}}, \Psi_{\mathbf{v}}$ *as in Eq. (1.16):*

$$\lim_{v\to\infty} iv((S^D-\mathbb{1})\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}) - (\{V^l(\mathbf{r}+\tau\hat{\mathbf{v}}) - V^l(\tau\hat{\mathbf{v}})\}\Phi_0,\Psi_0)$$
$$= \int_{-\infty}^{\infty} (V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)\,d\tau. \tag{1.24}$$

In particular, for a given long-range potential V^l *the scattering map*

$$S^D(\mathbf{A},V^l;\cdot): \mathscr{V}_{SR} \to L(\mathscr{H})$$

is injective.

Moreover, the following formula holds for all $\Phi_{\mathbf{v}}, \Psi_{\mathbf{v}}$ *as in Eq. (1.16):*

$$\lim_{v\to\infty} iv([S^D,\mathbf{p}_z]\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}) = \int_{-\infty}^{\infty} \{(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\mathbf{p}_z\Phi_0,\Psi_0) - (V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\mathbf{p}_z\Psi_0)$$
$$+ i((\partial_z V^l)(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)\}\,d\tau. \tag{1.25}$$

In particular, any of the scattering operators S^D *uniquely determines the total potential* V.

Remark 3: We only need the knowledge of $\mathbf{B}$, *in particular of* b, *and fix a gauge* $\mathbf{A}$ *for the definition of the wave operators* $\Omega_{\pm}^{(D)}$ *and the scattering operator* $S^{(D)}$. *The reconstruction formulae (1.17), (1.24) and (1.25) are independent of* $\mathbf{B}$.

Remark 4: The reconstruction formulae of Theorem 1 and 2 for $\mathbf{B}=0$ *are the same as in the case without magnetic field obtained by Enss and Weder in Ref. 4.*

Remark 5: Our results are also valid for a system of two interacting quantum particles with opposite charges (see Section V).

Our proofs use a geometric, time-dependent method developed by Enss and Weder. This method has been used by them in Ref. 4 to prove Theorems 1 and 2 in the case without magnetic field.

As previous works on inverse scattering theory without magnetic fields we mention the works of Faddeev,[5] Berezanskii[6] and Saito[7] who previously proved the uniqueness result of Theorem 1 in the two-body case with stationary methods. The last author proved it for bounded short-range continuous potentials. Also a stationary proof of uniqueness at fixed energy was given by Ramm[8] and Novikov[9] for potentials of compact support. For exponentially decreasing potentials a stationary proof of uniqueness at fixed energy was given by Novikov.[10] Wang[11] proved uniqueness with stationary methods in the N-body case with regular short-range potentials. Moreover, Isozaki and Kitada[12] studied the two-body case with C^∞ long-range potential and proved that the potential is uniquely determined by a particular scattering operator with time-independent modifier. For additional information about multidimensional inverse scattering see Weder,[13] Eskin and Ralston,[14] Newton,[15] and Chadan and Sabatier.[16]

To our knowledge the case with decaying magnetic fields was previously treated by Novikov and Khenkin in Ref. 17 where they proved that bounded, rapidly decreasing and sufficiently smooth potentials are uniquely obtained from the scattering operator at fixed energy. Eskin and Ralston proved in Ref. 18 that the scattering operator at fixed energy uniquely determines the magnetic field and the scalar potential by stationary phase methods for exponentially decaying $\mathbf{A}$ and V^s. By stationary methods Nicoleau proves in Ref. 19 that the high energy limit uniquely determines the magnetic field and the scalar potential, where the potentials have to be C^∞ - functions with decay assumptions on the potentials and all derivatives. In Refs. 20 and 21 we prove a similar result with less decay assumptions on the scalar potential with the geometric, time-dependent method.

As far as we know, there are no rigorous results on inverse scattering for one-body and many-body (including two-body) systems in a homogeneous magnetic field.

II. PROOF IN THE SHORT-RANGE CASE $V=V^s$

Since $S-\mathbb{1}=(\Omega_+-\Omega_-)^*\Omega_-$ we obtain

$$v((S-1)\Phi_{\mathbf{v}},\Psi_{\mathbf{v}})=\underbrace{v((\Omega_--1)\Phi_{\mathbf{v}},(\Omega_+-\Omega_-)\Psi_{\mathbf{v}})}_{=:R(\mathbf{v})}+\underbrace{v(\Phi_{\mathbf{v}},(\Omega_+-\Omega_-)\Psi_{\mathbf{v}})}_{=:l(\mathbf{v})}. \tag{2.1}$$

In the following we will show that $R(\mathbf{v})=\mathcal{O}(v^{-1})$ and that the high velocity limit $\lim_{v\to\infty}l(\mathbf{v})$ exists. For that purpose we need the following Lemma which is a propagation property expressing the fact that the solutions of the (one-dimensional) free Schrödinger equation have rapid decay away from the classically allowed region. In the following $F(z\in\mathcal{M})$ denotes the multiplication operator with the characteristic function $\chi_{\{z\in\mathcal{M}\}}(z)$.

Lemma 6: For any $f\in C_0^\infty(\mathbb{R}^1)$ with $\mathrm{supp}f\subset B_{m\eta}(0)$, for some $\eta>0$ and any $l=1,2,3,\ldots,$ there is a constant C_l such that

$$\|F(z\in\mathcal{M})e^{-i(1/2m)\mathbf{p}_z^2t}f(\mathbf{p}_z-m\mathbf{v}_z)F(z\in\mathcal{M}')\|\leq C_l(1+s+|t|)^{-l} \tag{2.2}$$

for every $\mathbf{v}_z,t\in\mathbb{R}$ and any measurable sets $\mathcal{M}$ and $\mathcal{M}'$ with distance $s:=\mathrm{dist}(\mathcal{M},\mathcal{M}'+\mathbf{v}_zt)-\eta|t|\geq 0$.

Proof: See Ref. 22, Proposition 2.10.

Proposition 7: Assume that $V^s\in\mathcal{V}_{SR}$. Then for all $\Phi_{\mathbf{v}}$ as in Eq. (1.16)

$$\|(\Omega_\pm-\mathbb{1})\Phi_{\mathbf{v}}\|=\mathcal{O}(v^{-1}). \tag{2.3}$$

Proof: We compute the difference

$$\begin{aligned}(\Omega_\pm-\mathbb{1})\Phi_{\mathbf{v}}&=\int_0^{\pm\infty}dt\,\frac{d}{dt}(e^{iHt}e^{-iH_At})\Phi_{\mathbf{v}}\\&=\int_0^{\pm\infty}dt\,e^{iHt}i(H-H_A)e^{-iH_At}\Phi_{\mathbf{v}}\\&=v^{-1}\int_0^{\pm\infty}d\tau\,e^{iH\tau/v}iV^s(\mathbf{r})e^{-iH_A\tau/v}\Phi_{\mathbf{v}}\end{aligned} \tag{2.4}$$

with the substitution $\tau=vt$. If we can show that the norm of the integrand in (2.4) is bounded by an integrable function $h\in L^1((\pm\infty,0],d\tau)$, we have finished the proof. To show integrability we will use $f(\mathbf{p}_z-m\mathbf{v}_z)\Phi_{\mathbf{v}}=\Phi_{\mathbf{v}}$ and $\|(1+|z|)^2\Phi_{\mathbf{v}}\|\leq\mathrm{const}<\infty$. In addition, we use that the particle moves in the direction parallel to the magnetic field with velocity $|\mathbf{v}_z|=v|\hat{\mathbf{v}}_z|>0$. Then

$$\begin{aligned}&\|V^s(\mathbf{r})e^{-iH_A\tau/v}\Phi_{\mathbf{v}}\|\\&\leq\|V^s(\mathbf{r})e^{-iH_{xy}\tau/v}e^{-i\frac{1}{2m}\mathbf{p}_z^2\tau/v}f(\mathbf{p}_z-m\mathbf{v}_z)(1+|z|)^{-2}\|\|(1+|z|)^2\Phi_{\mathbf{v}}\|\\&\leq\mathrm{const}\{\|F(|z|>|\tau\hat{\mathbf{v}}_z|/4)(1+|z|)^{-2}\|\end{aligned} \tag{2.5}$$

$$+\|F(|z-\tau\hat{\mathbf{v}}_z|\geq|\tau\hat{\mathbf{v}}_z|/2)e^{-i\frac{1}{2m}\mathbf{p}_z^2\tau/v}f(\mathbf{p}_z-m\mathbf{v}_z)F(|z|\leq|\tau\hat{\mathbf{v}}_z|/4)\| \tag{2.6}$$

$$+\|V^s(\mathbf{r})F(|z-\tau\hat{\mathbf{v}}_z|<|\tau\hat{\mathbf{v}}_z|/2)\|\} \tag{2.7}$$

$$\leq h(|\tau\hat{\mathbf{v}}_z|) \tag{2.8}$$

since $(2.5)\leq\mathrm{const}(1+|\tau\hat{\mathbf{v}}_z|)^{-2}\leq h_1(|\tau\hat{\mathbf{v}}_z|)$, $(2.6)\leq h_2(|\tau\hat{\mathbf{v}}_z|)$ for $v\geq 8\eta/|\hat{\mathbf{v}}_z|$ because of Lemma 6 and $(2.7)\leq\|V^s(\mathbf{r})F(|z|\geq|\tau\hat{\mathbf{v}}_z|/2)\|\leq h_3(|\tau\hat{\mathbf{v}}_z|)$ because V^s is of short range. Equation (2.8) follows with $h=h_1+h_2+h_3$. ■

The following Lemma expresses the translation in position and momentum space.

Lemma 8: For all measurable bounded or polynomial functions f on suitable domains and all $s\in\mathbb{R}$ we have

$$e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}f(\mathbf{r})e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}=f(\mathbf{r}+s\hat{\mathbf{v}}) \tag{2.9}$$

and

$$e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}f(\mathbf{p})e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}=f(\mathbf{p}+s(\mathbf{B}\times\hat{\mathbf{v}})/2)). \tag{2.10}$$

Proof: The translation in position space (2.9) can be obtained from

$$\begin{aligned}e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}\mathbf{r}e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}&=e^{i\mathbf{p}\hat{\mathbf{v}}s}e^{-i\mathbf{A}\hat{\mathbf{v}}s}\mathbf{r}e^{-i\mathbf{A}\hat{\mathbf{v}}s}e^{-i\mathbf{p}\hat{\mathbf{v}}s}\\&=e^{i\mathbf{p}\hat{\mathbf{v}}s}\mathbf{r}e^{-i\mathbf{p}\hat{\mathbf{v}}s}=\mathbf{r}+s\hat{\mathbf{v}},\end{aligned} \tag{2.11}$$

where in (2.11) $i[\mathbf{p}\hat{\mathbf{v}},\mathbf{A}\hat{\mathbf{v}}]=(b/2)\hat{\mathbf{v}}_x\hat{\mathbf{v}}_y+(-b/2)\hat{\mathbf{v}}_y\hat{\mathbf{v}}_x=0$ is used.

The translation in momentum space (2.10) can be obtained from

$$e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}\mathbf{p}e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}=e^{-i\mathbf{A}\hat{\mathbf{v}}s}\mathbf{p}e^{i\mathbf{A}\hat{\mathbf{v}}s} \tag{2.12}$$

$$=\mathbf{p}+s(\mathbf{B}\times\hat{\mathbf{v}})/2. \tag{2.13}$$

The last equality can be proven by showing that the derivatives in s of (2.12) and (2.13) are equal. To show this we use $i[\mathbf{p},\mathbf{A}\hat{\mathbf{v}}]=(b\hat{\mathbf{v}}_y/2,-b\hat{\mathbf{v}}_x/2,0)=(\mathbf{B}\times\hat{\mathbf{v}})/2$. ■

To prove the next Proposition, where we compute the high velocity limit of the wave operators, we need the following Lemma:

Lemma 9: Assume that $V^s\in\mathscr{V}_{SR}$. Then for fixed $\tau\in\mathbb{R}$

$$\operatorname*{s-lim}_{v\to\infty}e^{i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau}=e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}=\operatorname*{s-lim}_{v\to\infty}e^{i(H_A/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau}. \tag{2.14}$$

Proof: We show the first equality for H which implies the second with $V^s=0$ and compute the difference

$$\begin{aligned}&e^{i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau}-e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}\\&\quad=\{e^{i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau}e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}-1\}e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}\\&\quad=\int_0^\tau ds\,e^{i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})(\tau-s)}(-iH/v)e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}.\end{aligned}$$

Then we obtain

$$\|\{e^{i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau}-e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}\}\Psi_0\|\leqslant v^{-1}\int_0^\tau ds\|He^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}s}\Psi_0\| \tag{2.15}$$

and, if we can show that the integrand in (2.15) is bounded by a constant, we have finished the proof. Because of Lemma 8 the integrand in (2.15) is bounded by

$$\begin{aligned}&\|\{[\mathbf{p}-s(\mathbf{B}\times\hat{\mathbf{v}})/2-\mathbf{A}(\mathbf{r}-s\hat{\mathbf{v}})]^2/(2m)+V^s(\mathbf{r}-s\hat{\mathbf{v}})\}\Psi_0\|\\&\quad=\|\{[\mathbf{p}-\mathbf{A}(\mathbf{r})]^2/(2m)+V^s(\mathbf{r}-s\hat{\mathbf{v}})\}\Psi_0\|,\end{aligned}$$

where both sides are equal because of the transversal gauge. The second term is bounded by a constant because our dense set of states Ψ_0 are localized in position and momentum space. ■

Proposition 10: Assume that $V^s\in\mathscr{V}_{SR}$. Then for all $\Psi_{\mathbf{v}}$ as in Eq. (1.16)

$$\|(\Omega_+-\Omega_-)\Psi_{\mathbf{v}}\|=\mathscr{O}(v^{-1}) \tag{2.16}$$

and

$$\lim_{v\to\infty} e^{im\mathbf{v}\mathbf{r}} v(\Omega_+ - \Omega_-)\Psi_{\mathbf{v}} = \int_{-\infty}^{\infty} d\tau i V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Psi_0. \tag{2.17}$$

Proof: Similarly to the proof of Proposition 7 we compute the difference

$$(\Omega_+ - \Omega_-)\Psi_{\mathbf{v}} = v^{-1}\int_{-\infty}^{\infty} d\tau e^{iH\tau/v} i V^s(\mathbf{r}) e^{-iH_A\tau/v}\Psi_{\mathbf{v}} \tag{2.18}$$

and use that the norm of the integrand in (2.18) is bounded by an integrable function $h\in L^1((-\infty,\infty),d\tau)$ [see (2.4)–(2.8)]. Hence (2.16) is proven.

Now we calculate the high velocity limit (2.17). In the following we will use

$$e^{-im\mathbf{v}\mathbf{r}} f(\mathbf{p}) e^{im\mathbf{v}\mathbf{r}} = f(\mathbf{p}+m\mathbf{v}) \tag{2.19}$$

for measurable and bounded functions f and Lemma 8. Because of the dominated convergence theorem we can interchange the limit and the integral:

$$\begin{aligned}
&\lim_{v\to\infty} e^{im\mathbf{v}\mathbf{r}} v(\Omega_+ - \Omega_-)\Psi_{\mathbf{v}} \\
&\quad = \lim_{v\to\infty} \int_{-\infty}^{\infty} d\tau e^{i(H+(\mathbf{p}-\mathbf{A})\mathbf{v})\tau/v} i V^s(\mathbf{r}) e^{-i(H_A+(\mathbf{p}-\mathbf{A})\mathbf{v})\tau/v}\Psi_0 \\
&\quad = \int_{-\infty}^{\infty} d\tau e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau} i V^s(\mathbf{r}) e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}\Psi_0 \qquad (2.20)\\
&\quad = \int_{-\infty}^{\infty} d\tau i V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Psi_0, \qquad (2.21)
\end{aligned}$$

where we used Lemma 9 in (2.20) and (2.9) in (2.21). ■

With the help of Propositions 7 and 10 we calculate the following limit:

$$\begin{aligned}
&\lim_{v\to\infty} iv((S-\mathbb{1})\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}) \\
&\quad = \lim_{v\to\infty} iv((\Omega_- - \mathbb{1})\Phi_{\mathbf{v}},(\Omega_+ - \Omega_-)\Psi_{\mathbf{v}}) + iv(\Phi_{\mathbf{v}},(\Omega_+ - \Omega_-)\Psi_{\mathbf{v}}) \\
&\quad = \int_{-\infty}^{\infty} d\tau (V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0),
\end{aligned}$$

which is the desired reconstruction formula in Theorem 1.

We remark that we proved the reconstruction formula for all $\hat{\mathbf{v}}\in S^2$ with $\hat{\mathbf{v}}_z\neq 0$, i.e. for almost all $\hat{\mathbf{v}}\in S^2$, which is enough for the reconstruction of the scalar potential.

The injectivity of the scattering map results from the fact that the reconstruction of the potential V^s can be reduced to the inversion of the Radon transform of a bounded, continuous, square integrable function on $\mathbb{R}^2$ (see Ref. 4, Proof of Theorem 1.1) which uniquely determines the function (see Ref. 23, Chapter I, Theorem 2.17). Hence Theorem 1 is proved. ■

III. PROOF IN THE LONG-RANGE CASE $V = V^s + V^l$

Similarly to (2.1) we obtain

$$v((S^D-\mathbb{1})\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}) = \underbrace{v((\Omega_-^D-\mathbb{1})\Phi_{\mathbf{v}},(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}})}_{=:R^D(\mathbf{v})} + \underbrace{v(\Phi_{\mathbf{v}},(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}})}_{=:I^D(\mathbf{v})}. \tag{3.1}$$

In the following we will show that $R^D(\mathbf{v})=\mathcal{O}(v^{-1})$ and that the high velocity limit $\lim_{v\to\infty} l^D(\mathbf{v})$ exists. For that purpose we need the following Lemma which is a propagation property for the Dollard correction $\tilde{U}(t)$ with

$$\tilde{U}(t)=\exp\left[-i\int_0^t V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(s),s\mathbf{p}_z/m)ds\right] \tag{3.2}$$

and $\hat{\boldsymbol{\rho}}_{\mathbf{v}}(s)$ given in (1.20) – (1.22).

Lemma 11: For any $f\in C_0^\infty(\mathbb{R}^1)$ *with* $\mathrm{supp} f\subset B_{m\eta}(0)$ *and* $\mathbf{v}_z\in\mathbb{R}^1$ *with* $|\mathbf{v}_z|\geqslant 2\eta$ *the following estimates are true:*

$$\|z\tilde{U}(t)f(\mathbf{p}_z-m\mathbf{v}_z)(1+|z|^2)^{-1/2}\|\leqslant C(1+|\mathbf{v}_z|^{-2}|\mathbf{v}_z t|^{2-\gamma})\leqslant C(1+|\mathbf{v}_z t|)^{2-\gamma}, \tag{3.3}$$

$$\|F(|z|\geqslant|\mathbf{v}_z t|/4)\tilde{U}(t)f(\mathbf{p}_z-m\mathbf{v}_z)(1+|z|^2)^{-2}\|\leqslant C(1+|\mathbf{v}_z t|)^{-(2+\varepsilon)} \tag{3.4}$$

for some $\varepsilon>0$.

Proof: The estimates can be proven in the same way as the estimates for $\hat{\boldsymbol{\rho}}_{\mathbf{v}}=0$ in Ref. 4, Proposition 3.1 because $\hat{\boldsymbol{\rho}}_{\mathbf{v}}$ commutes with $\mathbf{p}_z$. ■

To show the boundedness of the norm of $\{\boldsymbol{\rho}-\hat{\boldsymbol{\rho}}_{\mathbf{v}}(t)\}e^{-iH_{xy}t}\Phi_{\mathbf{v}}$ we need an explicit expression of $e^{-iH_{xy}t}\boldsymbol{\rho}e^{-iH_{xy}t}$:

Proposition 12: With $\boldsymbol{\rho}^\perp:=(y,-x)$ *and* $\boldsymbol{\pi}^\perp:=(\mathbf{p}_y,-\mathbf{p}_x)$ *and* $\omega=b/m$ *the following equality holds:*

$$\begin{aligned}e^{-iH_{xy}t}\boldsymbol{\rho}e^{-iH_{xy}t}=&\boldsymbol{\rho}\{1+\cos(\omega t)\}/2+\boldsymbol{\rho}^\perp\sin(\omega t)/2\\&+\boldsymbol{\pi}\omega^{-1}\sin(\omega t)+\boldsymbol{\pi}^\perp\omega^{-1}\{1-\cos(\omega t)\}\end{aligned} \tag{3.5}$$

on the domain $(\boldsymbol{\rho}^2+\boldsymbol{\pi}^2)^{-1/2}\mathscr{H}$.

Proof: Let us denote the left hand side of the equation (3.5) with $f(t)$ and the right hand side with $g(t)$. We calculate derivatives in t of $f(t)$ by computing a series of commutators $i[H_{xy},\boldsymbol{\rho}]$ and of $g(t)$ and observe that $f'''=-\omega^2 f', g'''=-\omega^2 g'$ and $f'(0)=g'(0), f''(0)=g''(0)$, which implies $f'=g'$. In addition we know that $f(0)=g(0)$, which implies $f=g$. ■

Corollary 13: For $\Phi_{\mathbf{v}}$ *as in Eq. (1.16) and* $\hat{\boldsymbol{\rho}}_{\mathbf{v}}(t)$ *given in (1.20)–(1.22) the following expression is uniformly bounded in time* t *and velocity* $\mathbf{v}$*:*

$$\|\{\boldsymbol{\rho}-\hat{\boldsymbol{\rho}}_{\mathbf{v}}(t)\}e^{-iH_{xy}t}\Phi_{\mathbf{v}}\|\leqslant\text{const.} \tag{3.6}$$

Proof: Using Proposition 12 and the localization in position and momentum space we obtain the boundedness uniformly in time and velocity. ■

Proposition 14: Assume that $V^s\in\mathscr{V}_{SR}$ *and* $V^l\in\mathscr{V}_{LR}$. *Then for all* $\Phi_{\mathbf{v}}$ *as in Eq. (1.16)*

$$\|(\Omega_\pm^D-\mathbb{1})\Phi_{\mathbf{v}}\|=\mathcal{O}(v^{-1}). \tag{3.7}$$

Proof: We compute the difference

$$(\Omega_{\pm}^{D}-\mathbb{1})\Phi_{\mathbf{v}}=\int_0^{\pm\infty} dt\,\frac{d}{dt}(e^{iHt}U^D(t))\Phi_{\mathbf{v}}$$

$$=\int_0^{\pm\infty} dt e^{iHt} i(H-H_A-V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(t),t\mathbf{p}_z/m))U^D(t)\Phi_{\mathbf{v}}$$

$$=v^{-1}\int_0^{\pm\infty} d\tau e^{iH\tau/v} i\{V^s(\mathbf{r})+V^l(\mathbf{r})-V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(\tau/v),\tau\mathbf{p}_z/(vm))\}$$

$$\times U^D(\tau/v)\Phi_{\mathbf{v}}. \tag{3.8}$$

If we can show that the norm of the integrand in (3.8) is bounded by an integrable function $h \in L^1((\pm\infty,0],d\tau)$, we have finished the proof.

Since $i[\mathbf{p}_z,(\mathbf{p}-\mathbf{A})^2]=0$, the components in U^D commute and $U^D(t)=\exp[-iH_At]\widetilde{U}(t)$. Let $\overline{f}\in C_0^\infty(\mathbb{R})$ with $\overline{f}\equiv 1$ on the support of f. We use that the particle moves in the direction parallel to the magnetic field with velocity $|\mathbf{v}_z|=v|\hat{\mathbf{v}}_z|>0$. We show the integrability for the term with V^s and then for the term with V^l:

$$\|V^s(\mathbf{r})e^{-iH_A\tau/v}\widetilde{U}(\tau/v)\Phi_{\mathbf{v}}\| \tag{3.9}$$

$$=\|V^s(\mathbf{r})e^{-i\,(1/2m)\,\mathbf{p}_z^2\tau/v}e^{-iH_{xy}\tau/v}\widetilde{U}(\tau/v)\Phi_{\mathbf{v}}\|$$

$$\leqslant \text{const}\{\|F(|z|>|\tau\hat{\mathbf{v}}_z|/4)\widetilde{U}(\tau/v)f(\mathbf{p}_z-m\mathbf{v}_z)(1+|z|^2)^{-2}\| \tag{3.10}$$

$$+\|F(|z-\tau\hat{\mathbf{v}}_z|\geqslant|\tau\hat{\mathbf{v}}_z|/2)e^{-i\,(1/2m)\,\mathbf{p}_z^2\tau/v}\overline{f}(\mathbf{p}_z-m\mathbf{v}_z)F(|z|\leqslant|\tau\hat{\mathbf{v}}_z|/4)\| \tag{3.11}$$

$$+\|V^s(\mathbf{r})F(|z-\tau\hat{\mathbf{v}}_z|<|\tau\hat{\mathbf{v}}_z|/2)\|\}. \tag{3.12}$$

Then (3.10)$\leqslant h(|\tau\hat{\mathbf{v}}_z|)$ with Lemma 11 for $v\geqslant 2\eta/|\hat{\mathbf{v}}_z|$, and (3.11) and (3.12) are the same terms as (2.5) and (2.7), respectively. Thus (3.9) is bounded by an integrable function.

To show integrability for the term with the difference of V^l, we introduce a cutoff function $\chi\in C^\infty(\mathbb{R})$ with $0\leqslant\chi\leqslant 1$, $\chi(\lambda)=1$ for $\lambda\geqslant 1/2$, and $\chi(\lambda)=0$ for $\lambda\leqslant 1/4$, and define $V_\tau^l(\boldsymbol{\rho},z):=V^l(\boldsymbol{\rho},z)\chi(|z|/|\tau\hat{\mathbf{v}}_z|)$. Then it follows from (1.13) that

$$\|\nabla V_\tau^l\|=\sup|\nabla V_\tau^l(\mathbf{r})|\leqslant\text{const}|\tau\hat{\mathbf{v}}_z|^{-\gamma},\quad \gamma>3/2, \tag{3.13}$$

$$\|D^{(0,0,2)}V_\tau^l\|\leqslant\text{const}|\tau\hat{\mathbf{v}}_z|^{-(2+\delta)},\quad \delta>0. \tag{3.14}$$

We remark that $V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(\tau/v),\tau\mathbf{p}_z/(vm))=V_\tau^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(\tau/v),\tau\mathbf{p}_z/(vm))$ on the support of $f(\mathbf{p}_z-m\mathbf{v}_z)$. Then

$$\|\{V^l(\mathbf{r})-V^l_\tau(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau\mathbf{p}_z/(vm))\}e^{-iH_A\tau/v}\tilde{U}(\tau/v)\Phi_\mathbf{v}\| \tag{3.15}$$

$$=\|\{V^l(\mathbf{r})-V^l_\tau(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau\mathbf{p}_z/(vm))\}e^{-i\frac{1}{2m}\mathbf{p}_z^2\tau/v}e^{-iH_{xy}\tau/v}\tilde{U}(\tau/v)\Phi_\mathbf{v}\|$$

$$\leqslant \text{const}\{\|F(|z|>|\tau\hat{\mathbf{v}}_z|/4)\tilde{U}(\tau/v)f(\mathbf{p}_z-m\mathbf{v}_z)(1+|z|^2)^{-2}\| \tag{3.16}$$

$$+\|\{V^l(\mathbf{r})-V^l_\tau(\mathbf{r})\}e^{-iH_{xy}\tau/v}e^{-i(1/2m)\mathbf{p}_z^2\tau/v}\overline{f}(\mathbf{p}_z-m\mathbf{v}_z)F(|z|\leqslant|\tau\hat{\mathbf{v}}_z|/4)\|$$

$$+\|\{V^l_\tau(\mathbf{r})-V^l_\tau(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau\mathbf{p}_z/(vm))\}e^{-i(1/2m)\mathbf{p}_z^2\tau/v}e^{-iH_{xy}\tau/v}\tilde{U}(\tau/v)\Phi_\mathbf{v}\|\}$$

$$\leqslant h_1(|\tau\hat{\mathbf{v}}_z|)+\text{const}\{\|F(|z-\tau\hat{\mathbf{v}}_z|\geqslant|\tau\hat{\mathbf{v}}_z|/2)e^{-i(1/2m)\mathbf{p}_z^2\tau/v}\overline{f}(\mathbf{p}_z-m\mathbf{v}_z)$$

$$\times F(|z|\leqslant|\tau\hat{\mathbf{v}}_z|/4)\| \tag{3.17}$$

$$+\|\{V^l_\tau(\boldsymbol{\rho},z+\tau\mathbf{p}_z/(vm))-V^l_\tau(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau\mathbf{p}_z/(vm))\}$$

$$\times e^{-iH_{xy}\tau/v}\tilde{U}(\tau/v)\Phi_\mathbf{v}\|\}, \tag{3.18}$$

where (3.16)$\leqslant h_1(|\tau\hat{\mathbf{v}}_z|)$ with Lemma 11 for $v\geqslant 2\eta/|\hat{\mathbf{v}}_z|$ and (3.17)$\leqslant h_2(|\tau\hat{\mathbf{v}}_z|)$ with Lemma 6 for $v\geqslant 8\eta/|\hat{\mathbf{v}}_z|$. We used $\exp[i\mathbf{p}_z^2t/(2m)]f(\boldsymbol{\rho},z)\exp[-i\mathbf{p}_z^2t/(2m)]=f(\boldsymbol{\rho},z+t\mathbf{p}_z/m)$ to obtain (3.18). Finally (3.18) can be calculated using the Baker–Campbell–Hausdorff formula

$$\exp[-i\mathbf{q}(\boldsymbol{\rho},z+t\mathbf{p}_z/m)]\exp[i\mathbf{q}(s\boldsymbol{\rho}+(1-s)\hat{\boldsymbol{\rho}}_\mathbf{v}(t),sz+t\mathbf{p}_z/m)]$$

$$=\exp[-i\mathbf{q}(\boldsymbol{\rho}-\hat{\boldsymbol{\rho}}_\mathbf{v}(t),z)(1-s)]\exp[it\mathbf{q}_z^2(1-s)/(2m)]$$

for all $\mathbf{q}\in\mathbb{R}^3$. Then for the difference of V^l_τ in (3.18) we obtain

$$V^l_\tau(\boldsymbol{\rho},z+\tau\mathbf{p}_z/(vm))-V^l_\tau(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau\mathbf{p}_z/(vm))$$

$$=\int_0^1 ds\left\{-(\boldsymbol{\nabla}V^l_\tau)(s\boldsymbol{\rho}+(1-s)\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),sz+\tau\mathbf{p}_z/(vm))[\mathbf{r}-(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),z)]\right.$$

$$\left.+\frac{i}{2m}(D^{(0,0,2)}V^l_\tau)(s\boldsymbol{\rho}+(1-s)\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),sz+\tau\mathbf{p}_z/(vm))\tau/v\right\}.$$

With (3.13), (3.14), Corollary 13 and Lemma 11 integrability of (3.18) is proven. ■

Lemma 15: For $\hat{\boldsymbol{\rho}}_\mathbf{v}(t)$ given in (1.20) – (1.22) and all fixed $\tau\in\mathbb{R}$ the following high velocity limit holds:

$$\text{s-lim}_{v\to\infty}\, V^l(\hat{\boldsymbol{\rho}}_\mathbf{v}(\tau/v),\tau(\mathbf{p}_z+m\mathbf{v}_z)/(mv))=V^l(\tau\hat{\mathbf{v}}). \tag{3.19}$$

Proof: As V^l is continuous, decays at infinity and for fixed τ $\lim_{v\to\infty}\hat{x}_\mathbf{v}(\tau/v)=\tau\hat{\mathbf{v}}_x$, $\lim_{v\to\infty}\hat{y}_\mathbf{v}(\tau/v)=\tau\hat{\mathbf{v}}_y$ and $\lim_{v\to\infty}\mathbf{p}_z\tau/(mv)=0$ in the strong resolvent sense we obtain (3.19). ■

Proposition 16: Assume that $V^s\in\mathscr{K}_{SR}$ and $V^l\in\mathscr{K}_{LR}$. Then for all $\Psi_\mathbf{v}$ as in Eq. (1.16)

$$\|(\Omega^D_+-\Omega^D_-)\Psi_\mathbf{v}\|=\mathscr{O}(v^{-1}) \tag{3.20}$$

and

$$\lim_{v\to\infty}e^{im\mathbf{v}\mathbf{r}}v(\Omega^D_+-\Omega^D_-)\Psi_\mathbf{v}=\int_{-\infty}^{\infty}d\tau i\{V^s(\mathbf{r}+\tau\hat{\mathbf{v}})$$

$$+V^l(\mathbf{r}+\tau\hat{\mathbf{v}})-V^l(\tau\hat{\mathbf{v}})\}\Psi_0. \tag{3.21}$$

Proof: Similarly to the proof of Proposition 14 we compute the difference

$$(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}}=v^{-1}\int_{-\infty}^{\infty}d\tau e^{iH\tau/v}i\{V^s(\mathbf{r})+V^l(\mathbf{r})-V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(\tau/v),\tau\mathbf{p}_z/(vm))\}\times U^D(\tau/v)\Psi_{\mathbf{v}} \tag{3.22}$$

and can show that the norm of the integrand in (3.22) is bounded by an integrable function $h\in L^1((-\infty,\infty),d\tau)$. Hence (3.20) is proven.

Now we calculate the high velocity limit (3.21). Because of the dominated convergence theorem we can interchange the limit and the integral and obtain similarly to (2.21):

$$\begin{aligned}
&\lim_{v\to\infty}e^{im\mathbf{vr}}v(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}}\\
&\quad=\lim_{v\to\infty}\int_{-\infty}^{\infty}d\tau e^{i(H+(\mathbf{p}-\mathbf{A})\mathbf{v})\tau/v}i\{V^s(\mathbf{r})+V^l(\mathbf{r})-V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(\tau/v),\tau(\mathbf{p}_z+m\mathbf{v}_z)/(vm))\}\\
&\qquad\times e^{-i(H_A+(\mathbf{p}-\mathbf{A})\mathbf{v})\tau/v-i\int_0^{\tau/v}V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(s),s(\mathbf{p}_z+m\mathbf{v}_z)/m)ds}\Psi_0\\
&\quad=\int_{-\infty}^{\infty}d\tau e^{i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}i\{V^s(\mathbf{r})+V^l(\mathbf{r})-V^l(\tau\hat{\mathbf{v}})\}e^{-i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau}\Psi_0\\
&\quad=\int_{-\infty}^{\infty}d\tau i\{V^s(\mathbf{r}+\tau\hat{\mathbf{v}})+V^l(\mathbf{r}+\tau\hat{\mathbf{v}})-V^l(\tau\hat{\mathbf{v}})\}\Psi_0,
\end{aligned}\tag{3.23}$$

where in (3.23) $\text{s-lim}_{v\to\infty}\exp[i(H/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau]=\exp[i(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}}\tau]=\text{s-lim}_{v\to\infty}\exp[i(H_A/v+(\mathbf{p}-\mathbf{A})\hat{\mathbf{v}})\tau+i\int_0^{\tau/v}V^l(\hat{\boldsymbol{\rho}}_{\mathbf{v}}(s),s(\mathbf{p}_z+m\mathbf{v}_z)/m)ds]$ for fixed τ is used. This can similarly be shown as the equality in the short-range case in Lemma 9. ■

With the help of Propositions 14 and 16 we calculate the following limit:

$$\begin{aligned}
&\lim_{v\to\infty}iv((S^D-\mathbb{1})\Phi_{\mathbf{v}},\Psi_{\mathbf{v}})\\
&\quad=\lim_{v\to\infty}iv((\Omega_-^D-\mathbb{1})\Phi_{\mathbf{v}},(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}})+iv(\Phi_{\mathbf{v}},(\Omega_+^D-\Omega_-^D)\Psi_{\mathbf{v}})\\
&\quad=\int_{-\infty}^{\infty}d\tau(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)+\int_{-\infty}^{\infty}d\tau(\{V^l(\mathbf{r}+\tau\hat{\mathbf{v}})-V^l(\tau\hat{\mathbf{v}})\}\Phi_0,\Psi_0),
\end{aligned}\tag{3.24}$$

which is the reconstruction formula (1.24) in Theorem 2. The injectivity has already been shown in the short-range case above.

A short computation for the commutator of S^D with the third component of the momentum $\mathbf{p}_z$ gives (1.25). Then the total potential V can be computed from (1.25) (see Ref. 4, Proof of Theorem 1.2). Hence Theorem 2 is proved. ■

IV. ERROR BOUNDS

All formulae hold in the high velocity limit. For large, but fixed, v we obtain an error term of order $\mathcal{O}(|\mathbf{v}_z|^{-1})$. We remark that the order is not uniformly bounded in $\mathbf{v}$ if we consider all velocities $\mathbf{v}$ with $\hat{\mathbf{v}}\in S^2$ and $\hat{\mathbf{v}}_z\neq 0$ because the order $|\mathbf{v}_z|^{-1}=|\hat{\mathbf{v}}_z|^{-1}v^{-1}$ tends to infinity for $\hat{\mathbf{v}}_z\to 0$. But if the scalar potential is L^1 and compactly supported we only have to know the x-ray transform for infinitely many directions $\hat{\mathbf{v}}$, but not for all, to uniquely determine the potential from its x-ray transform (see Ref. 24, Theorem 5.2). Then, e.g., velocities with $|\hat{\mathbf{v}}_z|\geqslant 1/2$ are sufficient for the reconstruction and the error terms $|R(\mathbf{v})|$ and $|R^D(\mathbf{v})|$ can be calculated from (2.1) and (3.1), respectively, and are of order $\mathcal{O}(|\mathbf{v}|^{-1})$.

V. THE TWO-BODY CASE

Now we study a system of two interacting quantum particles with opposite charges in a homogeneous magnetic field. As the most important example of such a system we note the hydrogen atom (see Ref. 1 and references there). Let $H_0 := \Sigma_{i=1,2} (1/2m_i)\, \mathbf{p}_i^2$ be the free Hamiltonian. By

$$H_A := \sum_{i=1,2} \frac{1}{2m_i} (\mathbf{p}_i - e_i \mathbf{A}(\mathbf{r}_i))^2 \tag{5.1}$$

we denote the Hamiltonian with the magnetic field alone and by $H = H_A + V$ the interacting Hamiltonian with mass m_i, momentum $\mathbf{p}_i = -i\nabla_{\mathbf{r}_i}$, electric charge e_i and position $\mathbf{r}_i$ of the ith particle, where we will assume that $e_1 = -e_2$. As in the one-body case it is useful to introduce the following representation (see Ref. 1 for the notation):

$$H_A = H_0 + \sum_{i=1,2} \frac{1}{2m_i} \left\{ \frac{b^2}{4} \left(\frac{\mathbf{B}}{|\mathbf{B}|} \times \mathbf{r}_i \right)^2 - e_i \mathbf{B} \mathbf{L}_i \right\} \tag{5.2}$$

$$= H_0 + \sum_{i=1,2} \frac{1}{2m_i} \left\{ \frac{b^2}{4} (x_i^2 + y_i^2) - e_i b \mathbf{L}_{z,i} \right\} \tag{5.3}$$

$$= \sum_{i=1,2} \left\{ \frac{1}{2m_i} \mathbf{p}_{z,i}^2 + H_{osc,i} - \frac{1}{2m_i} e_i b \mathbf{L}_{z,i} \right\} \tag{5.4}$$

with the z-component of the angular momentum $\mathbf{L}_{z,i} = x_i \mathbf{p}_{y,i} - y_i \mathbf{p}_{x,i}$ of the ith particle and

$$H_{osc,i} = \frac{1}{2m_i} \left\{ (\mathbf{p}_{x,i}^2 + \mathbf{p}_{y,i}^2) + \frac{b^2}{4} (x_i^2 + y_i^2) \right\}. \tag{5.5}$$

As $H_{osc,i}$ and $\mathbf{L}_{z,i}$ do not act on the z-variable, we can separate the center of mass motion in the z-direction as usual. For that purpose we consider H acting on $\mathscr{H} = L^2(X)$ with

$$X = \{(\mathbf{r}_1, \mathbf{r}_2) \subset \mathbb{R}^6 | m_1 z_1 + m_2 z_2 = 0\}$$

instead of $L^2(\mathbb{R}^6)$, and write $z = z_1 - z_2$. The momentum space is

$$\hat{X} = \{(\mathbf{u}_1, \mathbf{u}_2) \subset \mathbb{R}^6 | \mathbf{u}_{z,1} + \mathbf{u}_{z,2} = 0\}.$$

It is known (see Ref. 2 for details) that $H_{osc,i} - (1/2m_i)\, e_i b \mathbf{L}_{z,i}$, acting on the Hilbert space $L^2(\mathbb{R}^2)$, has a complete set of eigenfunctions $\{\Theta_{ln}(x_i, y_i)\}_{l=0,\pm 1,\ldots;n=0,1,\ldots}$ with corresponding eigenvalues

$$E_{ln,i} = (2m_i)^{-1} b (2n + |l| - e_i l + 1), \tag{5.6}$$

where $E_{ln,i}$ and Θ_{ln} are often referred to as Landau energy levels and Landau orbits. Therefore the spectrum of $H_{xy} = \Sigma_{i=1,2} H_{osc,i} - (1/2m_i)\, e_i b \mathbf{L}_{z,i}$ acting on $L^2(\mathbb{R}^4)$ consists of infinitely degenerated eigenvalues $\{E_\beta\}_{\beta=(l,n,l',n')}$, $l, l' = 0, \pm 1, \ldots, n, n' = 0, 1, \ldots$, with

$$E_\beta = \frac{b}{2m_1} (2n + |l| - e_1 l + 1) + \frac{b}{2m_2} (2n' + |l'| - e_2 l' + 1) \tag{5.7}$$

and corresponding eigenfunctions $\theta_\beta((x_1, y_1), (x_2, y_2)) = \theta_{ln}(x_1, y_1) \times \theta_{l'n'}(x_2, y_2)$. From now on we assume that $e_1 = -e_2 = 1$.

We will use some notation similar to that commonly used in many-body scattering theory:

$$\mathbf{r} := \mathbf{r}_1 - \mathbf{r}_2, \quad \mathbf{p} := (\mathbf{p}_1/m_1 - \mathbf{p}_2/m_2) m, \tag{5.8}$$

with reduced mass $m=m_1m_2/(m_1+m_2)$, where $\mathbf{r}$ denotes the relative coordinate of $\mathbf{r}_1$ and $\mathbf{r}_2$ and $\mathbf{p}$ its conjugate momentum. As again H_{xy} does not act on the z-variable, we have $\exp[-iH_At]=\exp[-iH_{xy}t]\exp[-i\mathbf{p}_z^2/(2m)]$.

Then the theorems for the two-body case of Theorems 1 and 2 can be proven in the same way as in the one-body case where we use that the propagation properties of Lemma 6 and 11 are also true in the two-body case and that all other statements can be proven as before now using the new relative coordinates $\mathbf{r}$ and $\mathbf{p}$ [see (5.8)]. Finally we obtain the following Theorems for the two-body case:

Theorem 17: Two-Body Case with $V=V^s$: *Suppose that $V^s\in\mathscr{V}_{SR}$ and $V^l=0$. Then the following reconstruction formula holds for all $\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}$ as in Eq. (1.16):*

$$\lim_{v\to\infty} iv((S-1)\Phi_{\mathbf{v}},\Psi_{\mathbf{v}})=\int_{-\infty}^{\infty}(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)d\tau. \tag{5.9}$$

In particular, the scattering map

$$S(\mathbf{A};\cdot):\mathscr{V}_{SR}\to L(\mathscr{H})$$

is injective.

Theorem 18: Two-Body Case with $V=V^s+V^l$: *Suppose that $V^s\in\mathscr{V}_{SR}$ and $V^l\in\mathscr{V}_{LR}$. Then the following reconstruction formula holds for all $\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}$ as in Eq. (1.16):*

$$\begin{aligned}\lim_{v\to\infty} iv((S^D-1)\Phi_{\mathbf{v}},\Psi_{\mathbf{v}})&-(\{V^l(\mathbf{r}+\tau\hat{\mathbf{v}})-V^l(\tau\hat{\mathbf{v}})\}\Phi_0,\Psi_0)\\&=\int_{-\infty}^{\infty}(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)d\tau.\end{aligned} \tag{5.10}$$

In particular, for a given long-range potential V^l the scattering map

$$S^D(\mathbf{A},V^l;\cdot):\mathscr{V}_{SR}\to L(\mathscr{H})$$

is injective.

Moreover, the following formula holds for all $\Phi_{\mathbf{v}},\Psi_{\mathbf{v}}$ as in Eq. (1.16):

$$\begin{aligned}\lim_{v\to\infty} iv([S^D,\mathbf{p}_z]\Phi_{\mathbf{v}},\Psi_{\mathbf{v}})&\\&=\int_{-\infty}^{\infty}\{(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\mathbf{p}_z\Phi_0,\Psi_0)-(V^s(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\mathbf{p}_z\Psi_0)\\&\quad+i((\partial_zV^l)(\mathbf{r}+\tau\hat{\mathbf{v}})\Phi_0,\Psi_0)\}d\tau.\end{aligned} \tag{5.11}$$

In particular, any of the scattering operators S^D uniquely determines the total potential V.

VI. GENERALIZATION

It is possible to generalize the results such that unbounded short-range V^s (i.e., V^s is only Kato-small and unbounded in the direction parallel to the field) are included. The decay assumption (1.11) can be weakened such that

$$\|V^s(\mathbf{r})(\mathbf{p}_z^2+1_z)^{-1}F(|\mathbf{r}|\geqslant R)\|\in L^1([0,\infty),dR) \tag{6.1}$$

is sufficient for the reconstruction of V^s in Theorems 1 and 17 and of $V=V^s+V^l$ in Theorems 2 and 18 because V^s can be regularized with $f(\mathbf{p}_z-m\mathbf{v}_z)$ which commutes with e^{-iH_At} and $U^D(t)$.

ACKNOWLEDGMENTS

I thank Professor V. Enss and Professor R. Weder for their encouragement and helpful conversations. As this paper was written during my visit at the Institute for Advanced Study in

Princeton, I would like to thank the Institute for their hospitality. This research was supported by a grant from Deutsche Forschungsgemeinschaft, hereby gratefully acknowledged, and is in partial fulfillment of the requirements for a Ph.D. degree at the RWTH Aachen.

[1] I. Laba, ''Scattering for Hydrogen-like Systems in a Constant Magnetic Field,'' Commun. Partial Differen. Eq. **20**, 741–762 (1995).
[2] J. Avron, I. Herbst, and B. Simon, ''Schrödinger Operators with Magnetic Fields. I. General Interactions,'' Duke Math. J. **45**, 847–883 (1978).
[3] L. Hörmander, ''The Existence of Wave Operators in Scattering Theory,'' Math. Z. **146**, 69–91 (1976).
[4] V. Enss and R. Weder, ''The geometrical approach to multidimensional inverse scattering,'' J. Math. Phys. **36**, 3902–3921 (1995).
[5] L. D. Faddeev, ''Uniqueness of the Solution of the Inverse Scattering Problem,'' Vestnik Leningr. Univ. **11**(7), 126–130 (1956).
[6] Yu M. Berezanskii, ''The Uniqueness Theorem in the Inverse Problem of Spectral Analysis for the Schrödinger Equation,'' Trudy Moscov. Math. Obshch. **7**, 3–62 (1958); English translation in Am. Math. Soc. Trans. Ser. 2, **35**, 167–235 (1964).
[7] Y. Saito, ''An asymptotic behavior of the S-matrix in the inverse scattering problem,'' J. Math. Phys. **25**, 3105–3111 (1984); ''An approximation formula in the inverse scattering problem,'' J. Math. Phys. **27**, 1145–1153 (1986).
[8] A. G. Ramm, ''Completeness of the Products of Solutions to PDE and Uniqueness Theorems in Inverse Scattering,'' Inverse Probl. **3**, L77–L82 (1987).
[9] R. G. Novikov, ''Multidimensional Inverse Spectral Problem for the Equation $-\triangle\psi+\ (v(x)-Eu(x))\psi=0$,'' Funkt. Anal. Pril. **22**(4), 11–22 (1988); English translation Funct. Anal. Appl. **22**, 263–272 (1988).
[10] R. G. Novikov, ''The Inverse Scattering Problem at Fixed Energy for the Three Dimensional Schrödinger Equation with an Exponentially Decreasing Potential,'' Commun. Math. Phys. **161**, 569–595 (1994).
[11] X. P. Wang, ''On the Uniqueness of Inverse Scattering for N-Body Systems,'' Inverse Probl. **10**, 765–784 (1994).
[12] H. Isozaki and H. Kitada, ''Scattering Matrices for Two-Body Schr'ódinger Operators,'' Sci. Papers College Arts Sci. Univ. Tokyo **35**, 81–107 (1986).
[13] R. Weder, ''Multidimensional Inverse Scattering: The Reconstruction Problem,'' Inverse Probl. **6**, 267–298 (1990).
[14] G. Eskin and J. Ralson, ''Inverse Backscattering in Two Dimensions,'' Commun. Math. Phys. **138**, 451–486 (1991).
[15] R. G. Newton, *Inverse Schrödinger Scattering in Three Dimensions* (Springer, Berlin, 1989).
[16] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer, Berlin, 1989).
[17] R. G. Novokov and G. M. Khenkin, ''The $\bar{\partial}$-Equation in the Multidimensional Inverse Scattering Problem,'' Usp. Mat. Nauk **42(3)**, 93–152 (1987). English translation in Russ. Math. Surv. **42**(3), 109–180 (1987).
[18] G. Eskin and J. Ralston, ''Inverse Scattering Problems for the Schrödinger Equation with Magnetic Potential at a Fixed Energy,'' Commun. Math. Phys. **173**, 199–224 (1995).
[19] F. Nicoleau, ''A Stationary Approach to Inverse Scattering for Schrödinger Operators with First Order Perturbations,'' Comm. Partial Different. Eq. **22**, 527–553 (1997).
[20] S. Arians, ''Geometric approach to inverse scattering for the Schrödinger equation with magnetic and electric potentials,'' J. Math. Phys. **38**, 2761–2773 (1997).
[21] S. Arians, ''Suitable Gauges for the Geometric Approach to Inverse Scattering for the Schr'ódinger Equation with Magnetic and Electric Potentials,'' in preparation.
[22] V. Enss, ''Propagation Properties of Quantum Scattering States,'' J. Funct. Anal. **52**, 219–251 (1983).
[23] S. Helgason, *Groups and Geometric Analysis* (Academic, Orlando, 1984).
[24] F. Kleinert, ''Inversion of k-plane Transforms and Applications in Computer Tomography,'' SIAM (Soc. Ind. Appl. Math.) Rev. **31(2)**, 273–298 (1989).

Phase-space representation of quantum state vectors

Masashi Ban
Advanced Research Laboratory, Hitachi, Ltd., Akanuma 2520, Hatoyama, Saitama 350-0395, Japan

(Received 22 October 1997; accepted for publication 12 January 1998)

Phase-space representation of quantum state vectors is obtained within the framework of the relative-state formulation. For this purpose, the Hilbert space of a quantum system is enlarged by introducing an auxiliary quantum system. Relative-position state and relative-momentum state are defined in the extended Hilbert space of the composite quantum system and expressions of basic operators such as canonical position and momentum operators, acting on these states, are obtained. Phase-space functions which represent a state vector of the relevant quantum system are obtained in terms of the relative-position states and the relative-momentum states. The absolute-square of the phase-space function represents the probability distribution of the phase-space variables. Time-evolution of a quantum system is investigated in terms of the phase-space functions. The relations to the phase-space representations formulated by the other methods are obtained. © *1998 American Institute of Physics.* [S0022-2488(98)02804-7]

I. INTRODUCTION

Classical mechanics and classical statistical mechanics can use the phase space to describe dynamical and statistical properties of physical systems.[1,2] When we consider an n-dimensional system, the phase space of a physical system becomes a $2n$-dimensional space. An arbitrary point in the phase space that represents a state of the system is given by $(\mathbf{x},\mathbf{p})$, where $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ and $\mathbf{p}=(p_1,p_2,\ldots,p_n)$ are vectors of the canonical position and momentum of the system. Time-evolution of the system can be described by the trajectory $\mathbf{l}(t)=(\mathbf{x}_t,\mathbf{p}_t)$ in the phase space, where $\mathbf{x}_t=\mathbf{x}(t)$ and $\mathbf{p}_t=\mathbf{p}(t)$ are solutions of the Hamiltonian equations of motion.[1] Statistical properties of the system are determined by distribution of the phase-space points, where the probability distribution $P_c(\mathbf{x},\mathbf{p})$ completely characterizes the statistical properties of the system.[2] Therefore the phase space is useful for describing the properties of physical systems in classical mechanics and classical statistical mechanics. However it is well known that there is not such a phase-space description of physical systems in quantum mechanics. Because of the uncertainty relation of canonical position and momentum, physical systems cannot simultaneously take the definite values of position and momentum. Thus individual points in the phase space cannot have the physical meaning in quantum mechanics. This fact makes it impossible to define the phase-space probability distribution of the system.

There are the several phase-space functions which describe the properties of physical systems in quantum mechanics. The most important are the Wigner function,[3,4] the Glauber-Sudarshan function,[5–7] and the Husimi function.[8,9] Although the Wigner function $W(x,p)$ has the correct marginal distributions of position and momentum variables of a quantum system, it does not become positive definite.[10,11] Hence the Wigner function cannot be considered a phase-space probability distribution. The Husimi function defined by $Q(x,p)=\pi^{-1}\langle x,p|\hat{\rho}|x,p\rangle$ with a coherent state $|x,p\rangle=|\alpha\rangle$ $(\alpha=(x+ip)/\sqrt{2})$ is suitable for describing the heterodyne detection and the simultaneous measurement of position and momentum.[11–16] The Glauber-Sudarshan function becomes the most singular function for nonclassical states such as the photon-number state. In quantum optics, the degree of the singularity of this function is used to characterize the nonclassicality, such as the nonclassical depth, of quantum states.[17] The generalization of these phase-space functions has been considered by several authors.[18–22] The operational phase-space probability distributions have been also introduced,[23–27] which are closely related to quantum measurement performed on physical systems. The underlying idea of the operational phase-space probability distribution is equivalent to that of the fuzzy-space formulation of quantum

 © 1998 American Institute of Physics

mechanics.[28–31] The fussy or unsharp observables include the additional fluctuation due to the finite accuracy of a measurement apparatus.[32] Such a fluctuation is indispensable for introducing the phase-space probability distribution in quantum mechanics.

To investigate the properties of physical and chemical systems, Torres-Vega and Frederick[33] and Harriman[34] have recently formulated the phase-space representations of quantum state vectors that make it possible to define the phase-space probability distribution. Using their results, the quantum-classical correspondence[35] and the relation to the coherent state representation[36] have been investigated. Furthermore the numerical analysis for physical and chemical systems has been performed by means of the phase-space functions.[37–39] Their methods for formulating the phase-space representations are useful and convenient, but rather abstract. In their methods, some state vector and transformation kernel have been introduced without proving their existence. The present author has recently formulated the relative-state formulation[40,41] to introduce a phase operator in quantum optical systems.[42–47] Therefore, in this paper we will formulate the phase-space representation of quantum state vectors by means of the relative-state formulation. Furthermore it will be shown that under certain conditions, our phase-space representation becomes equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman.[34] Hence it is considered that the results obtained in this paper give some physical and mathematical basis for their phase-space representations.

This paper is organized as follows. In Sec. II, we briefly summarize the phase-space representations of quantum state vectors that have been formulated by Torres-Vega and Frederick[33] and by Harriman.[34] The method by Torres-Vega and Frederick introduces a complete orthonormal system which depends on the phase-space variables while the method by Harriman uses a linear transformation between the phase space and the position (momentum) space. In Sec. III, we introduce the relative-state formulation of quantum systems in a convenient way for obtaining the phase-space representation of quantum state vectors. Using the relative-state formulation, we define relative-position states and relative-momentum states and then we obtain the expressions of basic operators, such as position and momentum operators, acting on these state vectors. In Sec. IV, we obtain the phase-space representation of quantum state vectors by means of the relative-state formulation. It is shown that under certain conditions, our phase-space representation becomes equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman.[34] The quantum dynamics in the phase-space representation is also considered. In Sec. V, we summarize our results.

II. PHASE-SPACE REPRESENTATION OF QUANTUM STATE VECTORS

In this section, we briefly summarize the two phase-space representations of quantum state vectors of physical systems. One obtained by Torres-Vega and Frederick[33] uses the Hilbert-space-vector approach and the other by Harriman[34] uses the linear transformation approach. The two phase-space representations become equivalent under certain conditions.

A. Hilbert-space-vector approach

We now consider the phase-space representation of quantum state vectors by means of the Hilbert-space-vector approach. According to Torres-Vega and Frederick,[33] we first introduce a state vector $|\Gamma\rangle$ which is the eigenstate of a Hermitian phase-space operator $\hat{\Gamma}$ with eigenvalue Γ. It is assumed that the set of the state vectors becomes a complete orthonormal system in the Hilbert space $\mathscr{H}$ of a quantum system,[48]

$$\langle\Gamma|\Gamma'\rangle=\delta(\Gamma-\Gamma'),\quad \int_{\mathscr{C}} d\Gamma|\Gamma\rangle\langle\Gamma|=\hat{I}, \tag{2.1}$$

where $\mathscr{C}$ stands for the set of all possible values of Γ and $\hat{I}$ is an identity operator defined on the Hilbert space $\mathscr{H}$. Here it is not necessary to consider which observable quantity the Hermitian operator $\hat{\Gamma}$ represents. Using the completeness relation, an arbitrary state vector $|\psi\rangle$ in the Hilbert space $\mathscr{H}$ is expanded as $|\psi\rangle=\int_{\mathscr{C}} d\Gamma\Psi(\Gamma)|\Gamma\rangle$, where $\Psi(\Gamma)=\langle\Gamma|\psi\rangle$ is a square-integrable function normalized as $\int_{\mathscr{C}} d\Gamma|\Psi(\Gamma)|^2=1$. In the formulation by Torres-Vega and Frederick,[33] it is assumed that the eigenvalue Γ represents a point in a two-dimensional phase space. In this case,

we have $\mathscr{C}=\mathbf{R}^2$, where $\mathbf{R}$ is the set of all real numbers. Thus we can set $\Gamma=(x,p)$ and $\int_{\mathscr{C}} d\Gamma F(\Gamma)=\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dp F(x,p)$ for any function F of Γ, and Eq. (2.1) becomes

$$\langle \Gamma|\Gamma'\rangle=\delta(x-x')\,\delta(p-p'), \quad \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dp\,|\Gamma\rangle\langle\Gamma|=\hat{I}. \tag{2.2}$$

To describe the quantum mechanical properties of physical systems in terms of the state vector $|\Gamma\rangle$, we have to specify the properties of the state vector $|\Gamma\rangle$. To this end, recall that the basic operators of a quantum system are canonical position and momentum operators, $\hat{x}$ and $\hat{p}$, which satisfy the canonical commutation relation $[\hat{x},\hat{p}]=i$. We set $\hbar=1$ throughout this paper. Any physical quantity $\hat{\mathscr{O}}$ can be expressed in terms of these operators, for example, $\hat{\mathscr{O}}=\Sigma_m \Sigma_n \mathscr{O}(m,n)\hat{x}^m \hat{p}^n$. This fact indicates that to describe a quantum system in terms of the state vectors $|\Gamma\rangle$, we have to know how the canonical operators $\hat{x}$ and $\hat{p}$ act on this state vector. According to Torres-Vega and Frederick,[33] we choose the following representation of the position and momentum operators,

$$\langle \Gamma|\hat{x}|\psi\rangle=\left(\alpha x+i\beta\frac{\partial}{\partial p}\right)\Psi(x,p), \tag{2.3}$$

$$\langle \Gamma|\hat{p}|\psi\rangle=\left(\gamma p+i\delta\frac{\partial}{\partial x}\right)\Psi(x,p), \tag{2.4}$$

where $|\psi\rangle$ is an arbitrary state vector of the quantum system and $\Psi(x,p)=\langle\Gamma|\psi\rangle$ is the phase-space function of the state vector $|\psi\rangle$. The real parameters α, β, γ and δ must satisfy the relation $\beta\gamma-\alpha\delta=1$ for Eqs. (2.3) and (2.4) to be consistent with the canonical commutation relation. The relations given by Eqs. (2.2)–(2.4) make it possible to describe the quantum mechanical properties of physical systems in terms of the state vector $|\Gamma\rangle$. The description of physical systems in terms of the phase-space function $\Psi(x,p)$ and the differential operators $\alpha x+i\beta\partial/\partial p$ and $\gamma p+i\delta\partial/\partial x$ is called the phase-space representation of quantum state vectors.

We next consider the relation between the phase-space representation and the usual wave-function representation of quantum state vectors. In the position (momentum) space, a quantum state vector $|\psi\rangle$ is represented by the wave function $\psi(x)=\langle x|\psi\rangle$ [$\tilde{\psi}(p)=\langle p|\psi\rangle$], where $|x\rangle$ and $|p\rangle$ are the eigenstates of the position and momentum operators. Since $\{|x\rangle|x\in\mathbf{R}\}$ and $\{|\Gamma\rangle|\Gamma\in\mathbf{R}^2\}$ are complete orthonormal systems in the Hilbert space $\mathscr{H}$, the relation between the phase-space function $\Psi(x,p)$ and the wave function $\psi(x)$ is given by the linear transformation,

$$\Psi(x,p)=\int_{-\infty}^{\infty} dx'\,\mathscr{K}(x,p|x')\,\psi(x'), \tag{2.5}$$

$$\psi(x)=\int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dp'\,\mathscr{K}^*(x',p'|x)\Psi(x',p'), \tag{2.6}$$

where the transformation kernel $\mathscr{K}(x,p|x')=\langle\Gamma|x'\rangle$ satisfies the differential equation,

$$\left(\alpha x+i\beta\frac{\partial}{\partial p}\right)\mathscr{K}(x,p|x')=x'\,\mathscr{K}(x,p|x'), \tag{2.7}$$

which yields the solution,

$$\mathscr{K}(x,p|x')=\exp\left[-i\frac{p}{\beta}(x'-\alpha x)\right]\mathscr{G}(x|x'). \tag{2.8}$$

The function $\mathscr{G}(x|x')$ of x and x' cannot be determined unless we impose further conditions on the transformation kernels.

In the same way, we can obtain the linear transformation between the phase-space function $\Psi(x,p)$ and the wave function $\tilde{\psi}(p)$ in the momentum space,

$$\Psi(x,p)=\int_{-\infty}^{\infty}dp'.\tilde{\mathscr{K}}(x,p|p')\tilde{\psi}(p'), \tag{2.9}$$

$$\tilde{\psi}(p)=\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dp'.\tilde{\mathscr{K}}^{*}(x',p'|p)\psi(x',p'), \tag{2.10}$$

where the transformation kernel $.\tilde{\mathscr{K}}(x,p|p')=\langle\Gamma|p'\rangle$ is given by

$$.\tilde{\mathscr{K}}(x,p|p')=\exp\left[-i\frac{x}{\delta}(p'-\gamma p)\right]\tilde{\mathscr{G}}(p|p'). \tag{2.11}$$

Here the function $\tilde{\mathscr{G}}(p|p')$ does not depend on x. Since the position eigenstate $|x\rangle$ is obtained from the momentum eigenstate $|p\rangle$ by the Fourier transformation, the transformation kernels $.\mathscr{K}(x,p|x')$ and $.\tilde{\mathscr{K}}(x,p|p')$ are also related by the Fourier transformation,

$$.\mathscr{K}(x,p|x')=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dp'\exp(-ip'x').\tilde{\mathscr{K}}(x,p|p'). \tag{2.12}$$

Substituting Eqs. (2.8) and (2.11) into this equation, we obtain the relation between the functions $\mathscr{G}(x|x')$ and $\tilde{\mathscr{G}}(p|p')$,

$$\mathscr{G}(x|x')=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dp'\exp\left[-i\left(p'-\frac{p}{\beta}\right)\left(x'+\frac{x}{\delta}\right)\right]\tilde{\mathscr{G}}(p|p'), \tag{2.13}$$

where we have used the relation $\beta\gamma-\alpha\delta=1$. When $\tilde{\mathscr{G}}(p|p')$ is a function of $p'-p/\beta$, that is, $\tilde{\mathscr{G}}(p|p')=\tilde{\mathscr{G}}(p'-p/\beta)$, we obtain from Eq. (2.13),

$$\mathscr{G}(x|x')=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dp\ \exp\left[-ip\left(x'+\frac{x}{\delta}\right)\right]\tilde{\mathscr{G}}(p)=\mathscr{G}(x'+x/\delta), \tag{2.14}$$

where the function $\mathscr{G}(x)$ is the Fourier transformation of the function $\tilde{\mathscr{G}}(p)$. In this case, the transformation kernels given by Eqs. (2.8) and (2.11) become

$$.\mathscr{K}(x,p|x')=\mathscr{G}(x'+x/\delta)\exp\left[-i\frac{p}{\beta}(x'-\alpha x)\right], \tag{2.15}$$

$$.\tilde{\mathscr{K}}(x,p|p')=\tilde{\mathscr{G}}(p'-p/\beta)\exp\left[-i\frac{x}{\delta}(p'-\gamma p)\right]. \tag{2.16}$$

These results indicate that when we set the function $\mathscr{G}(x|x')$ $[\tilde{\mathscr{G}}(p|p')]$ and fix the real parameters α, β, γ and δ, there is the one-to-one correspondence between the phase-space function and the wave function in the position (momentum) space. The detailed properties and applications of the phase-space representation have been investigated in Refs. 33–39.

B. Linear transformation approach

We have introduced the state vector $|\Gamma\rangle$ in the Hilbert space to formulate the phase-space representation of quantum state vectors. However, there is another method to introduce the phase-space representation. The method obtained by Harriman[34] uses the linear transformation between the position (momentum) space and the phase space. According to Harriman,[34] let us consider the following linear transformations,

$$\Phi(x,p)=\int_{-\infty}^{\infty}dx'\,\mathscr{U}(x,p|x')\,\psi(x')=\int_{-\infty}^{\infty}dp'\,\mathscr{T}(x,p|p')\,\tilde{\psi}(p'),\tag{2.17}$$

where $\psi(x)=\langle x|\psi\rangle$ and $\tilde{\psi}(p)=\langle p|\psi\rangle$ are the wave functions of a quantum state vector $|\psi\rangle$ in the position and momentum representations. It should be noted that the phase-space function $\Phi(x,p)$ is defined by these transformations. Thus the transformation kernels $\mathscr{U}(x,p|x')$ and $\mathscr{T}(x,p|p')$ determine the phase-space representation of a quantum state vector $|\psi\rangle$. Here it is assumed that the transformation kernels $\mathscr{U}(x,p|x')$ and $\mathscr{T}(x,p|p')$ satisfy the orthogonality relations,

$$\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dp\,\mathscr{U}(x,p|x_1)\,\mathscr{U}^*(x,p|x_2)=\delta(x_1-x_2),\tag{2.18}$$

$$\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dp\,\mathscr{T}(x,p|p_1)\,\mathscr{T}^*(x,p|p_2)=\delta(p_1-p_2).\tag{2.19}$$

Since the wave function $\psi(x)$ is obtained from $\tilde{\psi}(p)$ by the Fourier transformation, the transformation kernel $\mathscr{T}(x,p|p')$ is connected to $\mathscr{U}(x,p|x')$ by the Fourier transformation,

$$\mathscr{T}(x,p|p')=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dx'\exp(ip'x')\,\mathscr{U}(x,p|x').\tag{2.20}$$

The relations given by Eqs. (2.17)–(2.19) are not sufficient to determine the phase-space representation of the quantum state vector $|\psi\rangle$. Thus Harriman imposed the following conditions on the transformation kernel $\mathscr{U}(x,p|x')$,

$$\left(\tilde{\alpha}x+i\tilde{\beta}\frac{\partial}{\partial p}\right)\mathscr{U}(x,p|z)=z\,\mathscr{U}(x,p|z),\tag{2.21}$$

$$\left(\tilde{\alpha}p-i\tilde{\beta}\frac{\partial}{\partial x}\right)\mathscr{U}(x,p|z)=i\frac{\partial}{\partial z}\mathscr{U}(x,p|z),\tag{2.22}$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are real parameters satisfying the relation $\tilde{\alpha}\tilde{\beta}=\frac{1}{2}$ which ensures the canonical commutation relation. The conditions on the transformation kernel $\mathscr{T}(x,p|z)$ are derived from Eqs. (2.21) and (2.22) by Eq. (2.20). Using Eqs. (2.17)–(2.22), we can obtain the phase-space representation of quantum state vectors of physical systems.[34]

From Eqs. (2.20)–(2.22), we can express the transformation kernels $\mathscr{U}(x,p|z)$ and $\mathscr{T}(x,p|z)$ in the following forms,[34]

$$\mathscr{U}(x,p|z)=\mathscr{W}(z-x/\tilde{\beta})\exp\left(-i\frac{p}{\tilde{\beta}}z+i\frac{px}{2\tilde{\beta}^2}\right),\tag{2.23}$$

$$\mathscr{T}(x,p|z)=\tilde{\mathscr{W}}(z-p/\tilde{\beta})\exp\left(i\frac{x}{\tilde{\beta}}z-i\frac{px}{2\tilde{\beta}^2}\right),\tag{2.24}$$

where the functions $\mathscr{W}(x)$ and $\tilde{\mathscr{W}}(x)$ are related by the Fourier transformation. Comparing Eqs. (2.15) and (2.16) with these equations, we find that the phase-space representation obtained by Torres-Vega and Frederick[33] becomes equal to that obtained by Harriman[34] when we set $\alpha=1/(2\tilde{\beta})$, $\beta=\tilde{\beta}$, $\gamma=1/(2\tilde{\beta})$ and $\delta=-\tilde{\beta}$. If we require the equality $\mathscr{W}(x)=\tilde{\mathscr{W}}(x)$, the transformation kernels become Gaussian functions,

$$\mathscr{U}(x,p|z)=\frac{1}{\tilde{\beta}\sqrt{2\pi}}\mathscr{W}_0(x/\tilde{\beta},p/\tilde{\beta};z),\tag{2.25}$$

$$\mathscr{T}(x,p|z)=\frac{1}{\tilde{\beta}\sqrt{2\pi}}\mathscr{W}_0(p/\tilde{\beta},-x/\tilde{\beta};z), \tag{2.26}$$

where the function $\mathscr{W}_0(x,y;z)$ is given by

$$\mathscr{W}_0(x,y;z)=\pi^{-1/4}\exp[-\tfrac{1}{4}(z-x)^2-iyz+\tfrac{1}{2}ixy]. \tag{2.27}$$

The detailed properties of the phase-space representation by means of the linear transformations have been investigated in Ref. 34.

III. RELATIVE-STATE FORMULATION OF QUANTUM SYSTEMS

We have reviewed the phase-space representations of quantum state vectors formulated by Torres-Vega and Frederick[33] and by Harriman.[34] The state vector $|\Gamma\rangle$ in the Hilbert-space-vector approach and the transformation kernels $\mathscr{U}(x,p|z)$ and $\mathscr{T}(x,p|z)$ in the linear transformation approach are essential for obtaining the phase-space representations. However their existence has been assumed without proof. In this section, we present the relative-state formulation in a suitable way for introducing the phase-space representation. The relative-state formulation gives some physical and mathematical basis for their phase-space representations.

A. Relevant and reference quantum systems

We consider the relative-state formulation of quantum systems that was originally obtained for quantum systems characterized by discrete quantum numbers.[40] The relative-state formulation makes it possible to introduce a phase operator in quantum optical systems. We first suppose that a quantum state of physical system is characterized by means of single real parameter $z\in\mathbf{R}$. We denote a basic state vector as $|z\rangle$ which is an element of the Hilbert space $\mathscr{H}$ of the quantum system.[48] Furthermore we assume that the set $\mathscr{S}=\{|z\rangle|z\in\mathbf{R}\}$ becomes a complete orthonormal system in the Hilbert space $\mathscr{H}$, which satisfies $\langle z_1|z_2\rangle=\delta(z_1-z_2)$ and $\int_{-\infty}^{\infty}dz|z\rangle\langle z|=\hat{I}$, where $\hat{I}$ is an identity operator defined on the Hilbert space $\mathscr{H}_r$. Thus any state vector $|\psi\rangle$ of the quantum system can be expressed as $|\psi\rangle=\int_{-\infty}^{\infty}dzf(z)|z\rangle$, where the expansion coefficient $f(z)=\langle z|\psi\rangle$ is a square-integrable function normalized as $\int_{-\infty}^{\infty}dz|f(z)|^2=1$. We next introduce an auxiliary quantum system to construct the relative-state formulation of the quantum system. We assume that any state vector of the auxiliary quantum system can be expressed in terms of state vectors $|z\rangle_r$ and that the set $\mathscr{S}_r=\{|z\rangle_r|z\in\mathbf{R}\}$ becomes a complete orthonormal system in the Hilbert space $\mathscr{H}_r$ of the auxiliary quantum system, which satisfies ${}_r\langle z_1|z_2\rangle_r=\delta(z_1-z_2)$ and $\int_{-\infty}^{\infty}dz|z\rangle_{rr}\langle z|=\hat{I}_r$, where $\hat{I}_r$ is an identity operator defined on the Hilbert space $\mathscr{H}_r$ of the auxiliary system. An arbitrary state vector $|\psi\rangle_r$ is expressed as $|\psi\rangle_r=\int_{-\infty}^{\infty}dzg(z)|z\rangle_r$, where the function $g(z)={}_r\langle z|\psi\rangle_r$ is normalized as $\int_{-\infty}^{\infty}dz|g(z)|^2=1$. In the following, we refer to the original and auxiliary quantum systems as the relevant and reference quantum systems, and we denote quantities of the reference quantum system by adding subscript r such as $|\psi\rangle_r$ and $\hat{x}_r$.

When we consider the composite system of the relevant and reference quantum systems, we have to treat quantum state vectors in the extended Hilbert space $\hat{\mathscr{H}}=\mathscr{H}\otimes\mathscr{H}_r$. The complete orthonormal system of the extended Hilbert space $\hat{\mathscr{H}}$ is given by

$$\hat{\mathscr{S}}=\{|z,z'\rangle\rangle=|z\rangle\otimes|z'\rangle_r||z\rangle\in\mathscr{S},|z'\rangle_r\in\mathscr{S}_r\}, \tag{3.1}$$

which satisfies the relations,

$$\langle\langle z_1,z_1'|z_2,z_2\rangle\rangle=\delta(z_1-z_2)\,\delta(z_1'-z_2'), \tag{3.2}$$

$$\int_{-\infty}^{\infty}dz\int_{-\infty}^{\infty}dz'|z,z'\rangle\rangle\langle\langle z,z'|=\hat{I}\otimes\hat{I}_r. \tag{3.3}$$

We denote a state vector in the extended Hilbert space $\hat{\mathscr{H}}$ by using ''double bra'' and ''double ket'' such as $\langle\langle\Phi|$ and $|\Psi\rangle\rangle$ to emphasize that it belongs to the extended Hilbert space $\hat{\mathscr{H}}$. Thus an arbitrary state vector $|\Psi\rangle\rangle$ of the relevant and reference quantum systems can be expressed in terms of the basic vectors $|z,z'\rangle\rangle$ in the Hilbert space $\hat{\mathscr{H}}$,

$$|\Psi\rangle\rangle=\int_{-\infty}^{\infty}dz\int_{-\infty}^{\infty}dz'F(z,z')|z,z'\rangle\rangle, \tag{3.4}$$

where we set $F(z,z')=\langle\langle z,z'|\psi\rangle\rangle$ which is normalized as $\int_{-\infty}^{\infty}dz\int_{-\infty}^{\infty}dz'|F(z,z')|^2=1$. The function $F(z,z')$ characterizes the composite state of the relevant and reference quantum systems in terms of the individual parameters, z and z', of the respective quantum systems.

We now introduce a state vector $|\omega(z_r,z;s)\rangle\rangle$ in the extended Hilbert space $\hat{\mathscr{H}}$,

$$|\omega(z_r,z;s)\rangle\rangle=|z+\tfrac{1}{2}(1+s)z_r\rangle\otimes|z-\tfrac{1}{2}(1-s)z_r\rangle_r, \tag{3.5}$$

where s is a real parameter. When we set $|\omega(z_r,z;s)\rangle\rangle=|u\rangle\otimes|v\rangle_r$ with $u=z+\frac{1}{2}(1+s)z_r$ and $v=z-\frac{1}{2}(1-s)z_r$, we obtain the relations,

$$z_r=u-v,\quad z=\tfrac{1}{2}(1-s)u+\tfrac{1}{2}(1+s)v. \tag{3.6}$$

Thus it is seen that in the state vector $|\omega(z_r,z;s)\rangle\rangle$ the parameter z_r represents the difference between the individual system-parameters and the parameter z represents the weighted sum of the individual system-parameters. Furthermore the set $\hat{\mathscr{S}}_r=\{|\omega(z_r,z;s)\rangle\rangle|z_r,z\in\mathbf{R}\}$ becomes a complete orthonormal system in the extended Hilbert space $\hat{\mathscr{H}}$,

$$\langle\langle\omega(z_r,z;s)|\omega(z_r',z';s)\rangle\rangle=\delta(z_r-z_r')\,\delta(z-z'), \tag{3.7}$$

$$\int_{-\infty}^{\infty}dz_r\int_{-\infty}^{\infty}dz|\omega(z_r,z;s)\rangle\rangle\langle\langle\omega(z_r,z;s)|=\hat{I}\otimes\hat{I}_r. \tag{3.8}$$

Thus an arbitrary state vector $|\Psi\rangle\rangle$ of the relevant and reference quantum systems can be expressed in terms of the state vectors $|\omega(z_r,z;s)\rangle\rangle$,

$$|\Psi\rangle\rangle=\int_{-\infty}^{\infty}dz_r\int_{-\infty}^{\infty}dz\mathscr{F}(z_r,z)|\omega(z_r,z;s)\rangle\rangle, \tag{3.9}$$

where the function $\mathscr{F}(z_r,z)=\langle\langle\omega(z_r,z;s)$ is normalized as $\int_{-\infty}^{\infty}dz_r\int_{-\infty}^{\infty}dz|\mathscr{F}(z_r,z)|^2=1$. The function $\mathscr{F}_s(z_r,z)$ characterizes the composite state of the relevant and reference quantum systems in terms of the parameter-difference z_r and the weighted parameter-sum z of the relevant and reference quantum systems. Therefore it is reasonable that we refer to the vector $|\omega(z_r,z;s)\rangle\rangle$ as the relative-parameter state.

B. Relative-position state and its properties

Suppose that a quantum system is described in terms of the position eigenstates $|x\rangle$. Then the reference quantum system is assumed to be described in terms of the position eigenstates $|x\rangle_r$. In this case we obtain a relative-position state $|\omega(r,x;s)\rangle\rangle$ of the relevant and reference quantum systems,

$$|\omega(r,x;s)\rangle\rangle=|x+\tfrac{1}{2}(1+s)r\rangle\otimes|x-\tfrac{1}{2}(1-s)r\rangle_r. \tag{3.10}$$

The set $\{|\omega(r,x;s)\rangle\rangle|r,x\in\mathbf{R}\}$ becomes a complete orthonormal system in the extended Hilbert space $\hat{\mathscr{H}}=\mathscr{H}\otimes\mathscr{H}_r$. The same relations as those given by Eqs. (3.7)–(3.9) are obtained for the relative-position state by replacing the parameters z_r and z with r and x. Note that the relative-position state $|\omega(r,x;s)\rangle\rangle$ satisfies the eigenvalue equations,

$$(\hat{x}-\hat{x}_r)|\omega(r,x;s)\rangle\rangle = r|\omega(r,x;s)\rangle\rangle, \tag{3.11}$$

$$\tfrac{1}{2}(\hat{x}+\hat{x}_r)|\omega(r,x;s)\rangle\rangle = (x+\tfrac{1}{2}sr)|\omega(r,x;s)\rangle\rangle. \tag{3.12}$$

It is found from Eq. (3.11) that the parameter r in the state vector $|\omega(r,x;s)\rangle\rangle$ represents the relative position between the relevant and reference quantum systems. Furthermore, since we obtain the relations,

$$\hat{x}|\omega(r,x;s)\rangle\rangle = [x+\tfrac{1}{2}(1+s)r]|\omega(r,x;s)\rangle\rangle, \tag{3.13}$$

$$\hat{x}_r|\omega(r,x;s)\rangle\rangle = [x-\tfrac{1}{2}(1-s)r]|\omega(r,x;s)\rangle\rangle, \tag{3.14}$$

it is easy to see that if we set the parameter $s=1$, 0, or -1, the parameter x in the state vector $|\omega(r,x;s)\rangle\rangle$ respectively represents the position of the reference quantum system, the center between the relevant and reference quantum systems, or the position of the relevant quantum system.

We next introduce a state vector $|\omega(r,k;s)\rangle\rangle$ by the Fourier transformation of the relative-position state $|\omega(r,x;s)\rangle\rangle$ with respect to the parameter x,

$$|\omega(r,k;s)\rangle\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dx|\omega(r,x;s)\rangle\rangle \exp(ikx), \tag{3.15}$$

which is also referred to as the relative-position state. This state vector plays an important role in formulating the phase-space representation of quantum state vectors. Since the Fourier transformation is unitary, the set $\{|\omega(r,k;s)\rangle\rangle|r,k\in\mathbf{R}\}$ becomes a complete orthonormal system in the extended Hilbert space $\hat{\mathscr{H}}$,

$$\langle\langle\omega(r,k;s)|\omega(r',x';s)\rangle\rangle = \delta(r-r')\,\delta(x-x'), \tag{3.16}$$

$$\int_{-\infty}^{\infty} dr\int_{-\infty}^{\infty} dk|\omega(r,k;s)\rangle\rangle\langle\langle\omega(r,k;s)| = \hat{I}\otimes\hat{I}_r. \tag{3.17}$$

Furthermore it is found from Eqs. (3.10) and (3.15) that the transformed relative-position state $|\omega(r,k;s)\rangle\rangle$ is the simultaneous eigenstate of the relative-position operator $\hat{x}-\hat{x}_r$ and the momentum-sum operator $\hat{p}+\hat{p}_r$, with eigenvalues r and k,

$$(\hat{x}-\hat{x}_r)|\omega(r,k;s)\rangle\rangle = r|\omega(r,k;s)\rangle\rangle, \tag{3.18}$$

$$(\hat{p}+\hat{p}_r)|\omega(r,k;s)\rangle\rangle = k|\omega(r,k;s)\rangle\rangle. \tag{3.19}$$

Here, $\hat{p}$ and $\hat{p}_r$ are the canonical momentum operators of the relevant and reference quantum systems. The operators $\hat{x}-\hat{x}_r$ and $\hat{p}+\hat{p}_r$ commute with each other, $[\hat{x}-\hat{x}_r,\hat{p}+\hat{p}_r]=0$. It is important to note that the dependence of the transformed relative-position state $|\omega(r,k;s)\rangle\rangle$ on the parameter s appears only in its phase factor. In fact, substituting Eq. (3.10) into Eq. (3.15), we obtain

$$|\omega(r,k;s)\rangle\rangle = |\omega(r,k)\rangle\rangle \exp[\tfrac{1}{2}i(1-s)kr], \tag{3.20}$$

where the state vector $|\omega(r,k)\rangle\rangle$ which does not depend on the parameter s is given by

$$|\omega(r,k)\rangle\rangle = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dx|x+r\rangle\otimes|x\rangle_r \exp(ikr). \tag{3.21}$$

Finally, we express the relative-position state $|\omega(r,k;s)\rangle\rangle$ in terms of bosonic annihilation and creation operators of the relevant and reference quantum systems. These operators are defined in terms of the position and momentum operators,

$$\hat{a}=\frac{\hat{x}+i\hat{p}}{\sqrt{2}},\quad \hat{a}^{\dagger}=\frac{\hat{x}-i\hat{p}}{\sqrt{2}},\quad \hat{a}_r=\frac{\hat{x}_r+i\hat{p}_r}{\sqrt{2}},\quad \hat{a}_r^{\dagger}=\frac{\hat{x}_r-i\hat{p}_r}{\sqrt{2}}. \tag{3.22}$$

To obtain the expression, recall the fact that the position eigenstate $|x\rangle$ is expressed as

$$|x\rangle=\pi^{-(1/4)}\exp[-\tfrac{1}{2}x^2+\sqrt{2}x\hat{a}^{\dagger}-\tfrac{1}{2}(\hat{a}^{\dagger})^2]|0\rangle, \tag{3.23}$$

where $|0\rangle$ is the vacuum state defined by the relation $\hat{a}|0\rangle=0$. Substituting Eq. (3.23) into Eq. (3.21) and performing the integration, we can obtain

$$|\omega(r,k;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}|\mu|^2+\mu\hat{a}^{\dagger}-\mu^{*}\hat{a}_r^{\dagger}+\hat{a}^{\dagger}\hat{a}_r^{\dagger}-\frac{1}{2}iskr\right)|0,0\rangle, \tag{3.24}$$

where we set $|0,0\rangle\rangle=|0\rangle\otimes|0\rangle_r$ and the complex parameter μ is given by

$$\mu=\frac{r+ik}{\sqrt{2}}. \tag{3.25}$$

The expression given by Eq. (3.24) is referred to as the Fock-space representation of the relative-position state $|\omega(r,k;s)\rangle\rangle$. It should be noted that by setting $s=0$ in Eq. (3.24), the Fock-state representation becomes equivalent to that obtained by Fan Hong-yi and Klauder[49] in connection with the EPR problem[50]

C. Relative-momentum state and its properties

We have introduced the relative-position states $|\omega(r,x;s)\rangle\rangle$ and $|\omega(r,k;s)\rangle\rangle$ to describe an arbitrary state of the relevant and reference quantum systems. We now consider another state vector obtained by replacing the parameters z_r and z in Eq. (3.5) with momentum variables. In this case, we obtain a relative-momentum state of the relevant and reference quantum systems,

$$|\tilde{\omega}(k,p;s)\rangle\rangle=|p+\tfrac{1}{2}(1+s)k\rangle\otimes|p-\tfrac{1}{2}(1-s)k\rangle_r, \tag{3.26}$$

where $|p\rangle$ and $|p\rangle_r$ are the momentum eigenstates of the relevant and reference quantum systems. It is easy to see that the set $\{|\tilde{\omega}(k,p;s)\rangle\rangle|k,p\in\mathbf{R}\}$ is a complete orthonormal system in the extended Hilbert space $\hat{\mathscr{H}}$. The same relations as those given by Eqs. (3.7)–(3.9) are also established for the relative-momentum states by replacing the parameters z_r and z with k and p. Note that the relative-momentum state $|\omega(k,p;s)\rangle\rangle$ is the simultaneous eigenstate of $\hat{p}-\hat{p}_r$ and $\hat{p}+\hat{p}_r$,

$$(\hat{p}-\hat{p}_r)|\tilde{\omega}(k,p;s)\rangle\rangle=k|\tilde{\omega}(k,p;s)\rangle\rangle, \tag{3.27}$$

$$\tfrac{1}{2}(\hat{p}+\hat{p}_r)|\tilde{\omega}(k,p;s)\rangle\rangle=(p+\tfrac{1}{2}sk)|\tilde{\omega}(k,p;s)\rangle\rangle. \tag{3.28}$$

The parameter k in $|\tilde{\omega}(k,p;s)\rangle\rangle$ represents the relative momentum between the relevant and reference quantum systems, and for $s=1$, 0, or -1, the parameter p in $|\tilde{\omega}(k,p;s)\rangle\rangle$ respectively represents the momentum of the reference quantum system, the arithmetic mean of the momenta of the relevant and reference quantum systems, or the momentum of the relevant quantum system.

As we have done for the relative-position state $|\omega(r,x;s)\rangle\rangle$, we introduce another relative-momentum state $|\tilde{\omega}(k,r;s)\rangle\rangle$ by the Fourier transformation of the state vector $|\tilde{\omega}(k,p;s)\rangle\rangle$ with respect to the parameter p,

$$|\widetilde{\omega}(k,r;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dp|\widetilde{\omega}(k,p;s)\rangle\rangle\exp(-ipr), \tag{3.29}$$

which is the simultaneous eigenstate of the position-sum operator $\hat{x}+\hat{x}_r$ and the momentum-difference operator $\hat{p}-\hat{p}_r$,

$$(\hat{p}-\hat{p}_r)|\widetilde{\omega}(k,r;s)\rangle\rangle=k|\widetilde{\omega}(k,r;s)\rangle\rangle, \tag{3.30}$$

$$(\hat{x}+\hat{x}_r)|\widetilde{\omega}(k,r;s)\rangle\rangle=r|\widetilde{\omega}(k,r;s)\rangle\rangle. \tag{3.31}$$

Because of the unitarity of the Fourier transformation, the set $\{|\widetilde{\omega}(k,r;s)\rangle\rangle|k,r\in\mathbf{R}\}$, becomes a complete orthonormal system in the extended Hilbert space $\hat{\mathscr{H}}$,

$$\langle\langle\widetilde{\omega}(k,r;s)|\widetilde{\omega}(k',r';s)\rangle\rangle=\delta(r-r')\,\delta(k-k'), \tag{3.32}$$

$$\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk|\widetilde{\omega}(k,r;s)\rangle\rangle\langle\langle\widetilde{\omega}(k,r;s)|=\hat{I}\otimes\hat{I}_r. \tag{3.33}$$

It is found that the dependence of the transformed relative-momentum state $|\widetilde{\omega}(k,r;s)\rangle\rangle$ on the parameter s appears only in its phase factor and the following relation is obtained

$$|\widetilde{\omega}(k,r;s)\rangle\rangle=|\widetilde{\omega}(k,r)\rangle\rangle\exp[-\tfrac{1}{2}i(1-s)kr], \tag{3.34}$$

where the state vector $|\widetilde{\omega}(k,r)\rangle\rangle$ which does not depend on the parameter s is given by

$$|\widetilde{\omega}(k,r)\rangle\rangle=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dp|p+k\rangle\otimes|p\rangle_r\,\exp(-ipr). \tag{3.35}$$

Furthermore, using the fact that the momentum eigenstate $|p\rangle$ is expressed as

$$|p\rangle=\pi^{-(1/4)}\exp[-\tfrac{1}{2}p^2+i\sqrt{2}p\hat{a}^\dagger+\tfrac{1}{2}(\hat{a}^\dagger)^2]|0\rangle, \tag{3.36}$$

we can obtain the Fock-space representation of the transformed relative-momentum state,

$$|\widetilde{\omega}(k,r;s)\rangle\rangle=\frac{1}{\sqrt{2\pi}}\,\exp\left(-\frac{1}{2}|\mu|^2+\mu\hat{a}^\dagger+\mu^*\hat{a}_r-\hat{a}^\dagger\hat{a}_r^\dagger+\frac{1}{2}iskr\right)|0,0\rangle, \tag{3.37}$$

where the complex parameter μ is given by Eq. (3.25).

D. Expressions for basic operators

Before closing this section, we obtain the expressions of the position, momentum, annihilation and creation operators when they act on the relative-position state $|\omega(r,k;s)\rangle\rangle$ and the relative-momentum state $|\widetilde{\omega}(k,r;s)\rangle\rangle$. For an arbitrary state vector $|\Psi\rangle\rangle$ of the relevant and reference quantum systems, we define the following functions,

$$\mathscr{F}_s(r,k)=\langle\langle\omega(r,k;s)|\Psi\rangle\rangle,\quad \widetilde{\mathscr{F}}_s(r,k)=\langle\langle\widetilde{\omega}(k,r;s)|\Psi\rangle\rangle, \tag{3.38}$$

which are referred to as the relative-state representation of the quantum state vector $|\Psi\rangle\rangle$. For the relative-position states $|\omega(r,k;s)\rangle\rangle$, the expressions of the position and momentum operators are obtained from Eqs. (3.10) and (3.15),

$$\langle\langle\omega(r,k;s)|\hat{x}|\psi\rangle\rangle=\left[\frac{1}{2}(1+s)r+i\frac{\partial}{\partial k}\right]\mathscr{F}_s(r,k)=\hat{\mathscr{X}}_s(r;k)\mathscr{F}_s(r,k), \tag{3.39}$$

$$\langle\langle \omega(r,k;s)|\hat{p}|\psi\rangle\rangle = \left[\frac{1}{2}(1-s)k - i\frac{\partial}{\partial r}\right].\mathscr{F}_s(r,k) = .\hat{\mathscr{X}}^*_{-s}(k;r).\mathscr{F}_s(r,k), \tag{3.40}$$

$$\langle\langle \omega(r,k;s)|\hat{x}_r|\psi\rangle\rangle = -\left[\frac{1}{2}(1-s)r - i\frac{\partial}{\partial k}\right].\mathscr{F}_s(r,k) = -.\hat{\mathscr{X}}^*_{-s}(r;k).\mathscr{F}_s(r,k), \tag{3.41}$$

$$\langle\langle \omega(r,k;s)|\hat{p}_r|\psi\rangle\rangle = \left[\frac{1}{2}(1+s)k + i\frac{\partial}{\partial r}\right].\mathscr{F}_s(r,k) = .\hat{\mathscr{X}}_s(k;r).\mathscr{F}_s(r,k), \tag{3.42}$$

where $.\hat{\mathscr{X}}_s(x;y) = \frac{1}{2}(1+s)x + i\partial/\partial y$. Using the complex parameter μ defined by Eq. (3.25), we can obtain the expressions of the annihilation and creation operators,

$$\langle\langle \omega(r,k;s)|\hat{a}|\psi\rangle\rangle = \left[\frac{1}{2}(\mu + s\mu^*) + \frac{\partial}{\partial \mu^*}\right].\mathscr{F}_s(\mu) = .\hat{\mathscr{M}}^{(+)}_s(\mu).\mathscr{F}_s(\mu), \tag{3.43}$$

$$\langle\langle \omega(r,k;s)|\hat{a}^\dagger|\psi\rangle\rangle = \left[\frac{1}{2}(\mu^* + s\mu) - \frac{\partial}{\partial \mu}\right].\mathscr{F}_s(\mu) = .\hat{\mathscr{M}}^{(-)}_s(\mu^*).\mathscr{F}_s(\mu), \tag{3.44}$$

$$\langle\langle \omega(r,k;s)|\hat{a}_r|\psi\rangle\rangle = -\left[\frac{1}{2}(\mu^* - s\mu) + \frac{\partial}{\partial \mu}\right].\mathscr{F}_s(\mu) = -.\hat{\mathscr{M}}^{(+)}_{-s}(\mu^*).\mathscr{F}_s(\mu), \tag{3.45}$$

$$\langle\langle \omega(r,k;s)|\hat{a}^\dagger_r|\psi\rangle\rangle = -\left[\frac{1}{2}(\mu - s\mu^*) - \frac{\partial}{\partial \mu^*}\right].\mathscr{F}_s(\mu) = -.\hat{\mathscr{M}}^{(-)}_{-s}(\mu).\mathscr{F}_s(\mu), \tag{3.46}$$

where $\mathscr{F}_s(\mu) = \mathscr{F}_s(r,k)$ and $.\hat{\mathscr{M}}^{(\pm)}_s(\mu) = \frac{1}{2}(\mu + s\mu^*) \pm \partial/\partial\mu^*$.

For the relative-momentum state $|\widetilde{\omega}(k,r;s)\rangle\rangle$, we can obtain the expressions of the position and momentum operators from Eqs. (3.26) and (3.29),

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{x}|\psi\rangle\rangle = \left[\frac{1}{2}(1-s)r + i\frac{\partial}{\partial k}\right].\widetilde{\mathscr{F}}_s(r,k) = .\hat{\mathscr{X}}_{-s}(r;k).\widetilde{\mathscr{F}}_s(r,k), \tag{3.47}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{p}|\psi\rangle\rangle = \left[\frac{1}{2}(1+s)k - i\frac{\partial}{\partial r}\right].\widetilde{\mathscr{F}}_s(r,k) = .\hat{\mathscr{X}}^*_s(k;r).\widetilde{\mathscr{F}}_s(r,k), \tag{3.48}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{x}_r|\psi\rangle\rangle = \left[\frac{1}{2}(1+s)r - i\frac{\partial}{\partial k}\right].\widetilde{\mathscr{F}}_s(r,k) = .\hat{\mathscr{X}}^*_s(r;k).\widetilde{\mathscr{F}}_s(r,k), \tag{3.49}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{p}_r|\psi\rangle\rangle = -\left[\frac{1}{2}(1-s)k + i\frac{\partial}{\partial r}\right].\widetilde{\mathscr{F}}_s(r,k) = -.\hat{\mathscr{X}}_{-s}(k;r).\widetilde{\mathscr{F}}_s(r,k). \tag{3.50}$$

Furthermore the annihilation and creation operators are expressed in the following forms:

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{a}|\psi\rangle\rangle = \left[\frac{1}{2}(\mu - s\mu^*) + \frac{\partial}{\partial \mu^*}\right].\widetilde{\mathscr{F}}_s(\mu) = .\hat{\mathscr{M}}^{(+)}_{-s}(\mu).\widetilde{\mathscr{F}}_s(\mu), \tag{3.51}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{a}^\dagger|\psi\rangle\rangle = \left[\frac{1}{2}(\mu^* - s\mu) - \frac{\partial}{\partial \mu}\right].\widetilde{\mathscr{F}}_s(\mu) = .\hat{\mathscr{M}}^{(-)}_{-s}(\mu^*).\widetilde{\mathscr{F}}_s(\mu), \tag{3.52}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{a}_r|\psi\rangle\rangle = \left[\frac{1}{2}(\mu^* + s\mu) + \frac{\partial}{\partial \mu}\right].\widetilde{\mathscr{F}}_s(\mu) = .\hat{\mathscr{M}}^{(+)}_s(\mu^*).\widetilde{\mathscr{F}}_s(\mu), \tag{3.53}$$

$$\langle\langle \widetilde{\omega}(k,r;s)|\hat{a}^\dagger_r|\psi\rangle\rangle = \left[\frac{1}{2}(\mu + s\mu^*) - \frac{\partial}{\partial \mu^*}\right].\widetilde{\mathscr{F}}_s(\mu) = .\hat{\mathscr{M}}^{(-)}_s(\mu).\widetilde{\mathscr{F}}_s(\mu), \tag{3.54}$$

where we set $\tilde{\mathscr{F}}_s(\mu)=\tilde{\mathscr{F}}_s(r,k)$. Using these expressions of the basic operators, we can investigate the quantum mechanical properties of physical systems in terms of the relative-position and relative-momentum states.

IV. PHASES-PACE REPRESENTATION IN THE RELATIVE-STATE FORMULATION

In this section, we obtain the phase-space representation of quantum state vectors of the relevant system by means of the relative-state formulation. It is shown that under certain conditions, our phase-space representation becomes equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman.[34] Using the results, we investigate the time-evolution of a free particle and harmonic oscillator in the phase space.

A. Reduction of the Hilbert space

We have introduced the relative-state formulation of quantum systems and obtained the expressions of the basic operators acting on the relative-position state $|\omega(r,k;s)\rangle\rangle$ and the relative-momentum state $|\tilde{\omega}(k,r;s)\rangle\rangle$. In the formulation, we have extended the Hilbert space as $\mathscr{H}\rightarrow\hat{\mathscr{H}}=\mathscr{H}\otimes\mathscr{H}_r$ by introducing the reference quantum system. However, when we investigate the properties of the relevant quantum system, we need only the description of the relevant quantum system. To obtain the reduced description, we assume that the state vector of the reference quantum system remains unchanged and there is no correlation between the relevant and reference quantum systems. We denote such a state vector of the reference quantum system as $|\phi\rangle_r$. Fixing the state vector of the reference quantum system corresponds to projecting a state vector in the extended Hilbert space $\hat{\mathscr{H}}=\mathscr{H}\otimes\mathscr{H}_r$ into the subspace $\hat{\mathscr{H}}'=\mathscr{H}\otimes\{|\phi\rangle_r\}\subset\hat{\mathscr{H}}$. In this reduced description, the state vectors $|\omega(r,k;s)\rangle\rangle={}_r\langle\phi|\omega(r,k;s)\rangle\rangle$ and $|\tilde{\omega}(k,r;s)\rangle\rangle={}_r\langle\phi|\tilde{\omega}(k,r;s)\rangle\rangle$ in the Hilbert space $\mathscr{H}$ are considered the relative-position state and the relative-momentum state of the relevant quantum system. The reduced description depends on the state vector $|\phi\rangle_r$ of the reference quantum system. Therefore, we have arbitrariness in the choice of the state vector $|\phi\rangle_r$ of the reference quantum system. However it should be noted that such arbitrariness also appears in the phase-space representations by Torres-Vega and Frederick[33] and by Harriman.[34] In fact, the functions $\mathscr{G}(z)$ and $\tilde{\mathscr{G}}(z)$ in Eqs. (2.15) and (2.16) and the functions $\mathscr{W}(z)$ and $\tilde{\mathscr{W}}(z)$ in Eqs. (2.23) and (2.24) are arbitrary unless we impose further conditions on them.

From Eqs. (3.15) and (3.29), we obtain the reduced relative-position state $|\omega(r,k;s)\rangle$ and the reduced relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$ of the relevant quantum system,

$$|\omega(r,k;s)\rangle=\frac{1}{\sqrt{2\pi}}e^{-i(1+s)kr/2}\int_{-\infty}^{\infty}dx|x\rangle\phi^*(x-r)e^{ikx}, \tag{4.1}$$

$$|\tilde{\omega}(k,r;s)\rangle=\frac{1}{\sqrt{2\pi}}e^{i(1+s)kr/2}\int_{-\infty}^{\infty}dp|p\rangle\tilde{\phi}^*(p-k)e^{-ipr}, \tag{4.2}$$

where $\phi(x)={}_r\langle x|\phi\rangle_r$ and $\tilde{\phi}(p)={}_r\langle p|\phi\rangle_r$ are the wave functions of the reference quantum system in the position and momentum spaces. The sets $\{|\omega(r,k;s)\rangle|r,k\in\mathbf{R}\}$ and $\{|\tilde{\omega}(k,r;s)\rangle|r,k\in\mathbf{R}\}$ are overcomplete systems in the Hilbert space $\mathscr{H}$,

$$\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk|\omega(r,k;s)\rangle\langle\omega(r,k;s)|=\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk|\tilde{\omega}(k,r;s)\rangle\langle\tilde{\omega}(k,r;s)|=\hat{I}, \tag{4.3}$$

$$\langle\omega(r,k;s)|\omega(r',k';s)\rangle=\frac{1}{2\pi}e^{i(1+s)(kr-k'r')/2}\int_{-\infty}^{\infty}dx\,\phi(x-r)\phi^*(x-r')e^{-i(k-k')x}, \tag{4.4}$$

$$\langle\tilde{\omega}(k,r;s)|\tilde{\omega}(k',r';s)\rangle=\frac{1}{2\pi}e^{-i(1+s)(kr-k'r')/2}\int_{-\infty}^{\infty}dp\,\tilde{\phi}(p-k)\tilde{\phi}^*(p-k')e^{ip(r-r')}. \tag{4.5}$$

The normalization conditions for the reduced states are given by $\langle\omega(r,k;s)|\omega(r,k;s)\rangle=\langle\tilde{\omega}(k,r;s)|\tilde{\omega}(k,r;s)\rangle=(2\pi)^{-1}$. The nonorthogonality of these state vectors is due to the restriction of the extended Hilbert space $\hat{\mathscr{H}}$. The degree of the nonorthogonality of the reduced states $|\omega(r,k;s)\rangle$ and $|\tilde{\omega}(k,r;s)\rangle$ depends on the wave function of the reference quantum system. When we set $k=k'$ or $r=r'$ in Eq. (4.4), we obtain the relations,

$$\langle\omega(r,k;s)|\omega(r',k;s)\rangle=\frac{1}{2\pi}e^{i(1+s)k(r-r')/2}\int_{-\infty}^{\infty}dx\,\phi(x-r)\,\phi^*(x-r'), \tag{4.6}$$

$$\langle\omega(r,k;s)|\omega(r,k';s)\rangle=\frac{1}{2\pi}e^{-i(1-s)(k-k')r/2}\int_{-\infty}^{\infty}dp\,\tilde{\phi}(p-k')\,\tilde{\phi}^*(p-k). \tag{4.7}$$

Let ℓ_x and ℓ_p be the spreads of the wave functions $\phi(x)$ and $\tilde{\phi}(p)$ in the position and momentum spaces, where the inequality $\ell_x\ell_p\geqslant 1/2$ should be satisfied because of the Heisenberg uncertainty relation. Thus we find from Eqs. (4.6) and (4.7),

$$\langle\omega(r,k;s)|\omega(r',k;s)\rangle=0 \quad \text{for } |r-r'|>\ell_x,$$

$$\langle\omega(r,k;s)|\omega(r,k';s)\rangle=0 \quad \text{for } |k-k'|>\ell_p.$$

The same result can be obtained for the reduced relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$. The completeness relation given by Eq. (4.3) is essential for obtaining the phase-space representation of state vectors of the relevant quantum system. For example, when the state vector $|\phi\rangle_r$ is a Glauber coherent state $|\alpha^*\rangle_r$ with complex amplitude α^*, the reduced relative-position and relative-momentum states become

$$|\omega(r,k;s)\rangle=\frac{1}{\sqrt{2\pi}}e^{i[\mathrm{Im}(\alpha^*\mu)-skr/2]}|\mu+\alpha\rangle, \tag{4.8}$$

$$|\tilde{\omega}(k,r;s)\rangle=\frac{1}{\sqrt{2\pi}}e^{-i[\mathrm{Im}(\alpha^*\mu)-skr/2]}|\mu-\alpha\rangle, \tag{4.9}$$

where $\mu=(r+ik)/\sqrt{2}$. In this case, the reduced relative-position and relative-momentum states are Glauber coherent states, except for the normalization.

In the reduced description of the relevant quantum system, the position, momentum, annihilation and creation operators acting on the reduced state vectors $|\omega(r,k;s)\rangle$ and $|\tilde{\omega}(k,r;s)\rangle$ are expressed in the same forms as those obtained in Sec. III D. In fact, we obtain the expressions of the position and momentum operators,

$$\langle\omega(r,k;s)|\hat{x}|\psi\rangle=\left[\frac{1}{2}(1+s)r+i\frac{\partial}{\partial k}\right]\psi_s(r,k)=\hat{\mathscr{X}}_s(r;k)\,\psi_s(r,k), \tag{4.10}$$

$$\langle\omega(k,r;s)|\hat{p}|\psi\rangle=\left[\frac{1}{2}(1-s)k-i\frac{\partial}{\partial r}\right]\psi_s(r,k)=\hat{\mathscr{X}}^*_{-s}(k;r)\,\psi_s(r,k), \tag{4.11}$$

$$\langle\tilde{\omega}(k,r;s)|\hat{x}|\psi\rangle=\left[\frac{1}{2}(1-s)r+i\frac{\partial}{\partial k}\right]\tilde{\psi}_s(r,k)=\hat{\mathscr{X}}_{-s}(r;k)\,\psi_s(r,k), \tag{4.12}$$

$$\langle\tilde{\omega}(k,r;s)|\hat{p}|\psi\rangle=\left[\frac{1}{2}(1+s)k-i\frac{\partial}{\partial r}\right]\tilde{\psi}_s(r,k)=\hat{\mathscr{X}}^*_s(k;r)\,\psi_s(r,k), \tag{4.13}$$

where we set $\psi_s(r,k)=\langle\omega(r,k;s)|\psi\rangle$ and $\tilde{\psi}_s(r,k)=\langle\tilde{\omega}(k,r;s)|\psi\rangle$ which are referred to as the phase-space functions of the quantum state vector $|\psi\rangle$ of the relevant quantum system. We can describe the properties of the relevant quantum system in terms of the phase-space function

$\psi_s(r,k)$ $[\tilde{\psi}_s(r,k)]$ and the differential operators $\hat{\mathscr{X}}_s(r;k)$ and $\hat{\mathscr{X}}^*_{-s}(k;r)$ $[\hat{\mathscr{X}}_{-s}(r;k)$ and $\hat{\mathscr{X}}^*_s(k;r)]$. This description is the phase-space representation of quantum state vectors of the relevant system by means of the relative-state formulation.

Before proceeding further, we consider the relation between the reduced relative-position state $|\omega(r,k;s)\rangle$ and the reduced relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$. Since the wave function $\tilde{\phi}(p)$ is derived from the wave function $\phi(x)$ by the Fourier transformation, the reduced relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$ given by Eq. (4.2) is expressed as

$$|\tilde{\omega}(k,r;s)\rangle = \frac{1}{\sqrt{2\pi}} e^{-i(1-s)kr/2} \int_{-\infty}^{\infty} dx |x\rangle \phi^*(r-x) e^{ikx}. \tag{4.14}$$

Comparing this expression with Eq. (4.1), we see that if the wave function $\phi(x)$ is even with respect to x, the following relation is established

$$|\omega(r,k;s)\rangle = |\tilde{\omega}(k,r;-s)\rangle. \tag{4.15}$$

For example, when the reference quantum system is in the vacuum state $|0\rangle_r$, we obtain this relation since the wave function of the vacuum state is $\phi(x)=\pi^{-(1/4)}\exp(-x^2/2)$. In fact, from Eqs. (4.8) and (4.9), we obtain the reduced states $|\omega(r,k;s)\rangle=(2\pi)^{-(1/2)}e^{-iskr/2}|\mu\rangle$ and $|\tilde{\omega}(k,r;s)\rangle=(2\pi)^{-(1/2)}e^{iskr/2}|\mu\rangle$ which satisfy Eq. (4.15).

B. Phase-space function of quantum system vector

We now consider the phase-space functions which represent a state vector of the relevant quantum system and investigate their properties in detail. We also discuss the relations to the phase-space functions obtained by Torres-Vega and Frederick[33] and by Harriman.[34] Suppose that a quantum state of the relevant system is described by state vector $|\psi\rangle$. Using the completeness relations of the reduced relative-position state $|\omega(r,k;s)\rangle$ and the relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$ in the Hilbert space $\mathscr{H}$, we can expand the state vector $|\psi\rangle$ in the following form:

$$|\psi\rangle = \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk \psi_s(r,k) |\omega(r,k;s)\rangle \tag{4.16}$$

$$= \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk \tilde{\psi}_s(r,k) |\tilde{\omega}_s(k,r;s)\rangle, \tag{4.17}$$

where $\psi_s(r,k)=\langle\omega(r,k;s)|\psi\rangle$ and $\tilde{\psi}_s(r,k)=\langle\tilde{\omega}(k,r;s)|\psi\rangle$ are the phase-space functions which represent the state vector $|\psi\rangle$ of the relevant quantum system. Conversely, if a phase-space function $\psi_s(r,k)$ or $\tilde{\psi}_s(r,k)$ is given, the corresponding state vector of the relevant quantum system is obtained by Eq. (4.16) or (4.17). The phase-space functions, $\psi_s(r,k)$ and $\tilde{\psi}_s(r,k)$, are normalized as

$$\int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk |\psi_s(r,k)|^2 = \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk |\tilde{\psi}_s(r,k)|^2 = 1. \tag{4.18}$$

Thus, the functions $|\psi_s(r,k)|^2$ and $|\tilde{\psi}_s(r,k)|^2$ can be interpreted as the probability distributions of the phase-space variables r and k.

To consider the meaning of the phase-space probability distributions $|\psi_s(r,k)|^2$ and $|\tilde{\psi}_s(r,k)|^2$, let us calculate the average values of the phase-space variables r and k,

$$\langle r^m k^n\rangle_\psi \equiv \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk r^m k^n |\psi_s(r,k)|^2 = \langle (\hat{x}-\hat{x}_r)^m (\hat{p}+\hat{p}_r)^n\rangle, \tag{4.19}$$

$$\langle r^m k^n\rangle_{\tilde{\psi}} \equiv \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk r^m k^n |\tilde{\psi}_s(r,k)|^2 = \langle (\hat{x}+\hat{x}_r)^m (\hat{p}-\hat{p}_r)^n\rangle, \tag{4.20}$$

where we set $\langle \hat{A}\rangle = \langle \Psi|\hat{A}|\Psi\rangle\rangle$ with $|\Psi\rangle\rangle = |\psi\rangle\otimes|\phi\rangle_r$. Thus the functions $|\psi_s(r,k)|^2$ and $|\tilde{\psi}_s(r,k)|^2$ represent the probability distributions of the eigenvalues of the operators $\hat{x}\mp\hat{x}_r$ and $\hat{p}\pm\hat{p}_r$. For example, when the state vector $|\phi\rangle_r$ of the reference quantum system is the vacuum state $|0\rangle_r$, we find that

$$\langle r\rangle_\psi = \langle r\rangle_{\tilde{\psi}} = \langle \psi|\hat{x}|\psi\rangle, \quad \langle (\Delta r)^2\rangle_\psi = \langle (\Delta r)^2\rangle_{\tilde{\psi}} = \langle \psi|(\Delta\hat{x})^2|\psi\rangle + \tfrac{1}{2}, \tag{4.21}$$

$$\langle k\rangle_\psi = \langle k\rangle_{\tilde{\psi}} = \langle \psi|\hat{p}|\psi\rangle, \quad \langle (\Delta k)^2\rangle_\psi = \langle (\Delta k)^2\rangle_{\tilde{\psi}} = \langle \psi|(\Delta\hat{p})^2|\psi\rangle + \tfrac{1}{2}, \tag{4.22}$$

where we set $\langle (\Delta r)^2\rangle_\psi = \langle r^2\rangle_\psi - \langle r\rangle_\psi^2$, $\langle \psi|(\Delta\hat{x})^2|\psi\rangle = \langle \psi|\hat{x}^2|\psi\rangle - \langle \psi|\hat{x}|\psi\rangle^2$ and so on. In general, the fluctuations of the phase-space variables satisfy the inequality,

$$\Delta_\psi r \Delta_\psi k \geqslant 1, \quad \Delta_{\tilde{\psi}} r \Delta_{\tilde{\psi}} k \geqslant 1, \tag{4.23}$$

where we set $\Delta_\psi r = \sqrt{\langle (\Delta r)^2\rangle_\psi}$, $\Delta_{\tilde{\psi}} r = \sqrt{\langle (\Delta r)^2\rangle_{\tilde{\psi}}}$ and so on. These inequalities are called the operational uncertainty relation.[51]

When we have a state vector $|\psi\rangle$ of the relevant quantum system, we obtain the wave functions $\psi(x) = \langle x|\psi\rangle$ and $\tilde{\psi}(p) = \langle p|\psi\rangle$ in the position and momentum spaces, $|\psi(x)|^2$ and $|\tilde{\psi}(p)|^2$ being the probability distributions of the position and momentum of the quantum system. We first investigate the relation between the phase-space function $\psi_s(r,k)$ and the wave function $\psi(x)$. Since the sets $\{|\omega(r,k;s)\rangle | r,k\in\mathbf{R}\}$ and $\{|x\rangle | x\in\mathbf{R}\}$ are complete in the Hilbert space $\mathscr{H}$, the phase-space function $\psi_s(r,k)$ and the wave function $\psi(x)$ which represent the same state vector $|\psi\rangle$ are connected by the linear transformation,

$$\psi_s(r,k) = \int_{-\infty}^{\infty} dx . \mathscr{T}_s(r,k|x)\psi(x), \tag{4.24}$$

$$\psi(x) = \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk . \mathscr{T}_s^*(r,k|x)\psi_s(r,k), \tag{4.25}$$

where the transformation kernel $\mathscr{T}_s(r,k|x)$ is given by

$$\mathscr{T}_s(r,k|x) = \langle \omega(r,k;s)|x\rangle = \frac{1}{\sqrt{2\pi}}\phi(x-r)e^{-ik[x-(1+s)r/2]}, \tag{4.26}$$

which satisfies the relations,

$$\int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk . \mathscr{T}_s(r,k|x) . \mathscr{T}_s^*(r,k|x') = \delta(x-x'), \tag{4.27}$$

$$\int_{-\infty}^{\infty} dx . \mathscr{T}_s(r,k|x) . \mathscr{T}_s^*(r',k'|x) = \langle \omega(r,k;s)|\omega(r',k';s)\rangle. \tag{4.28}$$

The right-hand side of the last equation is given by Eq. (4.4).

We consider the relation of the transformation kernel $\mathscr{T}_s(r,k|x)$ to those obtained by Harriman and by Torres-Vega and Frederick. Recall that in the phase-space representation by Harriman,[34] the transformation kernel [see Eq. (2.23)] is given by

$$\mathscr{U}(r,k|x) = \mathscr{W}(x-r/\tilde{\beta})\exp\left[-i\frac{k}{\tilde{\beta}}\left(x-\frac{r}{2\tilde{\beta}}\right)\right]. \tag{4.29}$$

Comparing this equation with Eq. (4.26), we find that when we set the parameter $s=0$ in Eq. (4.26) and $\tilde{\beta}=1$ in Eq. (4.29), the transformation kernels $\mathscr{T}_s(r,k|x)$ and $\mathscr{U}(r,k|x)$ become equivalent. In this case, the function $\mathscr{W}(x)$ corresponds to the wave function $\phi(x)={}_r\langle x|\phi\rangle_r$ of the reference quantum system. On the other hand, in the phase-space representation by Torres-Vega and Frederick,[33] the transformation kernel [see Eq. (2.15)] is given by

$$\mathscr{K}(r,k|x)=\mathscr{G}(x+r/\delta)\exp\left[-i\frac{k}{\beta}(x-\alpha r)\right]. \tag{4.30}$$

It is easy to see from Eqs. (4.26) and (4.30) that when we set the parameters $\alpha=\frac{1}{2}(1+s)$, $\beta=1$, $\gamma=\frac{1}{2}(1-s)$ and $\delta=-1$ $(\beta\gamma-\alpha\delta=1)$ in this equation, the transformation kernel $\mathscr{T}_s(r,k|x)$ is equivalent to $\mathscr{K}(r,k|x)$. Therefore when we appropriately choose the parameters, the phase-space representation by means of the relative-state formulation becomes equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman.[34]

In the same way, the phase-space function $\psi_s(r,k)$ and the wave function $\tilde{\psi}(p)$ in the momentum space are connected by the linear transformation,

$$\psi_s(r,k)=\int_{-\infty}^{\infty}dp.\tilde{\mathscr{T}}_s(r,k|p)\tilde{\psi}(p), \tag{4.31}$$

$$\tilde{\psi}(p)=\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\tilde{\mathscr{T}}_s^*(r,k|p)\psi_s(r,k), \tag{4.32}$$

where the transformation kernel $\tilde{\mathscr{T}}_s(r,k|p)$ is given by

$$\tilde{\mathscr{T}}_s(r,k|x)=\langle\omega(r,k;s)|p\rangle=\frac{1}{\sqrt{2\pi}}\tilde{\phi}(k-p)e^{i[p-(1-s)k/2]r}, \tag{4.33}$$

which satisfies the relations,

$$\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\tilde{\mathscr{T}}_s(r,k|p).\tilde{\mathscr{T}}_s^*(r,k|p')=\delta(p-p'), \tag{4.34}$$

$$\int_{-\infty}^{\infty}dp.\tilde{\mathscr{T}}_s(r,k|p).\tilde{\mathscr{T}}_s^*(r',k'|p)=\langle\omega(r,k;s)|\omega(r',k';s)\rangle. \tag{4.35}$$

Comparing Eq. (2.24) with Eq. (4.33), we find the equivalence of the transformation kernels $\tilde{\mathscr{T}}_s(r,k|p)$ and $\mathscr{V}(r,k|p)$ when we set the parameters $s=0$ and $\tilde{\beta}=1$. Furthermore it is found from Eqs. (2.16) and (4.33) that the equivalence of the transformation kernels $\tilde{\mathscr{T}}_s(r,k|p)$ and $\mathscr{K}(r,k|x)$ is established when we set the parameters $\alpha=\frac{1}{2}(1+s)$, $\beta=1$, $\gamma=\frac{1}{2}(1-s)$ and $\delta=-1$ in Eq. (2.16). For example, if the state vector $|\phi\rangle_r$ of the reference quantum system is the vacuum state $|0\rangle_r$ which yields the wave function $\phi(x)=\pi^{-(1/4)}\exp(-x^2/2)$, the transformation kernels $\mathscr{T}_s(r,k|x)$ and $\tilde{\mathscr{T}}_s(r,k|p)$ becomes Gaussian,

$$\mathscr{T}_s(r,k|x)=\frac{1}{\sqrt{2\pi^{3/2}}}\exp\left[-\frac{1}{2}(x-r)^2-ikx+\frac{1}{2}i(1+s)kr\right], \tag{4.36}$$

$$\tilde{\mathscr{T}}_s(r,k|p)=\frac{1}{\sqrt{2\pi^{3/2}}}\exp\left[-\frac{1}{2}(p-k)^2+ipr-\frac{1}{2}i(1-s)kr\right], \tag{4.37}$$

which are equal to Eqs. (2.25) and (2.26) if we set the parameters $s=0$ and $\tilde{\beta}=1$.

Next we consider the relation between the phase-space function $\tilde{\psi}_s(r,k)$ and the wave function $\psi(x)$ $[\tilde{\psi}(p)]$. Using the completeness relations of the state vectors, we can obtain the linear transformations,

$$\tilde{\psi}_s(r,k)=\int_{-\infty}^{\infty}dx.\mathscr{R}_s(r,k|x)\psi(x)=\int_{-\infty}^{\infty}dp.\tilde{\mathscr{R}}_s(r,k|p)\tilde{\phi}(p), \tag{4.38}$$

$$\psi(x)=\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\mathscr{R}_s^*(r,k|x)\tilde{\psi}_s(r,k), \tag{4.39}$$

$$\tilde{\psi}(p)=\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\tilde{\mathscr{R}}_s^*(r,k|p)\tilde{\psi}_s(r,k), \tag{4.40}$$

where the transformation kernels $.\mathscr{R}_s(r,k|x)$ and $.\tilde{\mathscr{R}}_s(r,k|p)$ are given by

$$.\mathscr{R}_s(r,k|x)=\langle\tilde{\omega}(k,r;s)|x\rangle=\frac{1}{\sqrt{2\pi}}\phi(r-x)e^{-ik[x-(1-s)r/2]}, \tag{4.41}$$

$$.\tilde{\mathscr{R}}_s(r,k|p)=\langle\tilde{\omega}(k,r;s)|p\rangle=\frac{1}{\sqrt{2\pi}}\tilde{\phi}(p-k)e^{i[p-(1+s)k/2]r}, \tag{4.42}$$

which satisfy the relations,

$$\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\mathscr{R}_s(r,k|x).\mathscr{R}_s^*(r,k|x')=\delta(x-x'), \tag{4.43}$$

$$\int_{-\infty}^{\infty}dr\int_{-\infty}^{\infty}dk.\tilde{\mathscr{R}}_s(r,k|p).\tilde{\mathscr{R}}_s^*(r,k|p')=\delta(p-p'), \tag{4.44}$$

$$\int_{-\infty}^{\infty}dx.\mathscr{R}_s(r,k|x).\mathscr{R}_s^*(r',k'|x)=\langle\tilde{\omega}(k,r;s)|\tilde{\omega}(k',r';s)\rangle, \tag{4.45}$$

$$\int_{-\infty}^{\infty}dp.\tilde{\mathscr{R}}_s(r,k|p).\tilde{\mathscr{R}}_s^*(r',k'|p)=\langle\tilde{\omega}(k,r;s)|\tilde{\omega}(k',r';s)\rangle. \tag{4.46}$$

From Eqs. (2.23) and (2.24), the equivalence of $.\mathscr{R}_s(r,k|x)$ and $\mathscr{U}(r,k|x)$ and the equivalence of $.\tilde{\mathscr{R}}_s(r,k|p)$ and $\tilde{\mathscr{V}}(r,k|p)$ are found if we set the parameters $s=0$ and $\tilde{\beta}=1$. Furthermore, from Eqs. (2.15) and (2.16), the equivalence of $.\mathscr{R}_s(r,k|x)$ and $.\mathscr{K}(r,k|x)$ and the equivalence of $.\tilde{\mathscr{R}}_s(r,k|p)$ and $.\tilde{\mathscr{K}}(r,k|p)$ are also found if we set the parameters $\alpha=\frac{1}{2}(1-s)$, $\beta=1$, $\gamma=\frac{1}{2}(1+s)$ and $\delta=-1$. In particular, when the reference quantum system is in the vacuum state, the transformation kernels $.\mathscr{R}_s(r,k|x)$ and $.\tilde{\mathscr{R}}_s(r,k|p)$ are calculated to be

$$.\mathscr{R}_s(r,k|x)=\frac{1}{\sqrt{2\pi^{3/2}}}\exp\left[-\frac{1}{2}(x-r)^2-ikx+\frac{1}{2}i(1-s)kr\right], \tag{4.47}$$

$$.\tilde{\mathscr{R}}_s(r,k|p)=\frac{1}{\sqrt{2\pi^{3/2}}}\exp\left[-\frac{1}{2}(p-k)^2+ipr-\frac{1}{2}i(1+s)kr\right], \tag{4.48}$$

which are equal to Eqs. (2.25) and (2.26) if we set the parameters $s=0$ and $\tilde{\beta}=1$.

We summarize the relations among the transformation kernels in the phase-space representations. When we set the parameters $s=0$ and $\tilde{\beta}=1$ and the functions $\mathscr{W}(x)=(2\pi)^{-(1/2)}\phi(x)$ and $\tilde{\mathscr{W}}(p)=(2\pi)^{-(1/2)}\tilde{\phi}(p)$, we obtain the relations,

$$\mathscr{T}_s(r,k|x)=\mathscr{R}_s(-r,-k|-x)=\mathscr{U}(r,k|x), \tag{4.49}$$

$$\tilde{\mathscr{T}}_s(-r,-k|-x)=\tilde{\mathscr{R}}_s(r,k|x)=\mathscr{V}(r,k|x). \tag{4.50}$$

On the other hand, if we set the parameters $\alpha=\frac{1}{2}(1+s)$, $\beta=1$, $\gamma=\frac{1}{2}(1-s)$ and $\delta=-1$, and the functions $\mathscr{G}(x)=(2\pi)^{-(1/2)}\phi(x)$, we obtain the relations,

$$\mathscr{T}_s(r,k|x)=\mathscr{K}(r,k|x), \quad \tilde{\mathscr{T}}_s(r,k|p)=\tilde{\mathscr{K}}(-r,-k|-p), \tag{4.51}$$

and if we set the parameters $\alpha=\frac{1}{2}(1-s)$, $\beta=1$, $\gamma=\frac{1}{2}(1+s)$ and $\delta=-1$, and the functions $\tilde{\mathscr{G}}(p)=(2\pi)^{-(1/2)}\tilde{\phi}(p)$, we obtain the relations,

$$\mathscr{R}_s(r,k|x)=\mathscr{K}(-r,-k|-x), \quad \tilde{\mathscr{R}}_s(r,k|p)=\tilde{\mathscr{K}}(r,k|p). \tag{4.52}$$

These results indicate that the phase-space representation of quantum state vectors by means of the relative-state formulation is equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman.[34]

We finally consider the relations among our transformation kernels. It is easily found from Eqs. (4.26), (4.33), (4.41) and (4.42) that the following relations are established

$$\mathscr{T}_s(r,k|x)=\mathscr{R}_{-s}(-r,-k|-x), \quad \tilde{\mathscr{T}}_s(r,k|-p)=\tilde{\mathscr{R}}_{-s}(-r,-k|-p). \tag{4.53}$$

Let us suppose that the wave function $\phi(x)={}_r\langle x|\phi\rangle_r$ of the reference quantum system is an even function of x. In this case, since there is the relation between the reduced relative-position state $|\omega(r,k;s)\rangle$ and the reduced relative-momentum state $|\tilde{\omega}(k,r;s)\rangle$ [see Eq. (4.15)], we can obtain the following relations among the transformation kernels,

$$\mathscr{R}_s(r,k|x)=\mathscr{T}_{-s}(r,k|x)=e^{-iskr}\mathscr{T}_s(r,k|x), \tag{4.54}$$

$$\tilde{\mathscr{R}}_s(r,k|p)=\tilde{\mathscr{T}}_{-s}(r,k|p)=e^{-iskr}\tilde{\mathscr{T}}_s(r,k|p). \tag{4.55}$$

In fact, it is easy to check that the transformation kernels given by Eqs. (4.36), (4.37), (4.47) and (4.48) satisfy these relations.

C. Time-evolution of phase-space function

We investigate the time-evolution of the phase-space function $\psi_s(r,k;t)$ which represents a state vector of the relevant quantum system. Let $H(\hat{x},\hat{p})$ be a Hamiltonian of the system and let $|\psi(t)\rangle$ be a state vector of the quantum system at time t. Then the equation of motion for the state vector $|\psi(t)\rangle$ is given by the Schrödinger equation ($\hbar=1$),

$$\frac{\partial}{\partial t}|\psi(t)\rangle=-iH(\hat{x},\hat{p})|\psi(t)\rangle. \tag{4.56}$$

Using Eqs. (4.10) and (4.11), we obtain the equation of motion for the phase-space function $\psi_s(r,k;t)=\langle\omega(r,k;s)|\psi(t)\rangle$,

$$\frac{\partial}{\partial t}\psi_s(r,k;t)=-iH\left[\frac{1}{2}(1+s)r+i\frac{\partial}{\partial k},\frac{1}{2}(1-s)k-i\frac{\partial}{\partial r}\right]\psi_s(r,k;t). \tag{4.57}$$

It is important to note that the initial condition for the phase-space function $\psi_s(r,k;0)=\psi_s(r,k)$ depends not only on the initial state vector of the relevant quantum system but also on the state vector $|\phi\rangle_r$ of the reference quantum system. When we obtain the solution $\psi_s(r,k;t)$ of the equation of motion given by Eq. (4.57), we can calculate the average value of any physical quantity $A(\hat{x},\hat{p})$ at time t in the phase space,

$$\langle A(t)\rangle=\langle\psi(t)|A(\hat{x},\hat{p})|\psi(t)\rangle$$

$$= \int_{-\infty}^{\infty} dr \int_{-\infty}^{\infty} dk \psi_s^*(r,k;t) A\left[\frac{1}{2}(1+s)r + i\frac{\partial}{\partial k}, \frac{1}{2}(1-s)k - i\frac{\partial}{\partial r}\right]\psi_s(r,k;t). \quad (4.58)$$

In the following, we will consider the properties of the phase-space function $\psi_s(r,k;t)$ for a free particle and harmonic oscillator in one-dimensional space.

1. Free particle

We first consider the time-evolution of the phase-space function which represents a quantum state vector of a free particle with unit mass in one-dimensional space. We suppose that the initial state of the particle is given by the Gaussian wave packet, $\psi(x) = \pi^{-(1/4)}\exp(-x^2/2)$. Since we have to fix the state vector of the reference quantum system to obtain the phase-space representation, we assume that the reference quantum system is in the vacuum state $|0\rangle_r$, for the sake of simplicity. In this case, we have $\phi(x) = \pi^{-(1/4)}\exp(-x^2/2)$. Thus the phase-space function of the initial state of the free particle is given by

$$\psi_s(r,k;0) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{4}(r^2+k^2) + \frac{1}{2}iskr\right]. \quad (4.59)$$

Since the Hamiltonian of the free particle is given by $H(\hat{p}) = \hat{p}^2/2$, the equation of motion for the phase-space function $\psi_s(r,k)$ becomes

$$\frac{\partial}{\partial t}\psi_s(r,k;t) = -\frac{1}{2}i\left[\frac{1}{2}(1-s)k - i\frac{\partial}{\partial r}\right]^2\psi_s(r,k;t). \quad (4.60)$$

Solving this equation of motion under the initial condition (4.59), we can obtain the phase-space function at time t,

$$\psi_s(r,k;t) = \frac{1}{\sqrt{\pi(2+it)}}\exp\left[-\frac{r^2+(1+it)k^2+2i(1+it)kr}{2(2+it)} + i\frac{1+s}{2}kr\right]. \quad (4.61)$$

Then the phase-space probability distribution is given by

$$|\psi_s(r,k;t)|^2 = \frac{1}{\pi\sqrt{4+t^2}}\exp\left[-\frac{2}{4+t^2}\left(r-\frac{1}{2}kt\right)^2 - \frac{1}{2}k^2\right]. \quad (4.62)$$

Thus the probability distribution of the phase-space variables r and k is the Gaussian distribution. In the position representation, the wave function at time t is obtained

$$\psi(x;t) = \frac{1}{\sqrt{\pi^{1/2}(1+it)}}\exp\left[-\frac{x^2}{2(1+it)}\right], \quad (4.63)$$

and the probability distribution of the position of the free particle becomes

$$|\psi(x;t)|^2 = \frac{1}{\sqrt{\pi(1+t^2)}}\exp\left(-\frac{x^2}{1+t^2}\right). \quad (4.64)$$

It is easy to check that the phase-space function $\psi_s(r,k;t)$ and the wave function $\psi(x;t)$ are connected by the linear transformation with the kernel function given by Eq. (4.36).

Using the results, we calculate the average values and fluctuations of the phase-space variables r and k,

$$\langle r\rangle_t = \langle k\rangle_t = 0, \quad (\Delta r)_t^2 = \tfrac{1}{2}(2+t^2), \quad (\Delta k)_t^2 = 1, \quad (4.65)$$

while we obtain the average values and fluctuations of the position and momentum variables,

$$\langle \hat{x} \rangle_t = \langle \hat{p} \rangle_t = 0, \quad (\Delta x)_t^2 = \tfrac{1}{2}(1+t^2), \quad (\Delta p)_t^2 = \tfrac{1}{2}. \tag{4.66}$$

The enhancement of the fluctuations in the phase space is due to the vacuum-fluctuation of the reference quantum system.

2. Harmonic oscillator

We next consider the phase-space function which represents a quantum state of a harmonic oscillator with unit mass and unit frequency in one-dimensional space. We here assume that the reference quantum system is in the vacuum state, $|\phi\rangle_r = |0\rangle_r$. The Hamiltonian of the harmonic oscillator is given by $\hat{H} = \frac{1}{2}(\hat{x}^2 + \hat{p}^2) = \hat{a}^\dagger \hat{a} + \frac{1}{2}$. Before considering the time-evolution of the phase-space function, let us investigate the property of the phase-space function of the energy eigenstate. Since the energy eigenstate of the harmonic oscillator is given by $|\psi_n\rangle = (1/n!)\hat{a}^{\dagger n}|0\rangle$, where $\hat{H}|\psi_n\rangle = (n+\frac{1}{2})|\psi_n\rangle$ $(n=0,1,2,\ldots)$, the corresponding phase-space function $\psi_s(r,k|n)$ becomes

$$\psi_s(r,k|n) = \frac{1}{\sqrt{2\pi n!}} \mu^{*n} \exp\left[-\frac{1}{2}|\mu|^2 + \frac{1}{4}s(\mu^2 - \mu^{*2}) \right], \tag{4.67}$$

where the complex parameter μ is given by $\mu = (r+ik)/\sqrt{2}$. In deriving Eq. (4.67), we have assumed the reference quantum system to be in the vacuum state. Thus the phase-space probability distribution of the energy eigenstate becomes

$$|\psi_s(r,k|n)|^2 = \frac{1}{2\pi n!}\left[\frac{1}{2}(r^2+k^2) \right]^n \exp\left[-\frac{1}{2}(r^2+k^2) \right]. \tag{4.68}$$

Since the classical Hamiltonian of the harmonic oscillator is given by $H_{\rm cl}(x,p) = \frac{1}{2}(x^2+p^2)$, we can express the phase-space probability distribution in the following form:

$$|\psi_s(r,k|n)|^2 = \frac{1}{2\pi n!}[H_{\rm cl}(r,k)]^n \exp[-H_{\rm cl}(r,k)]. \tag{4.69}$$

This result indicates that the most probable path of the energy eigenstate in the phase space is obtained by solving the equation $H_{\rm cl}(r,k) = n$ which corresponds to the condition of the semiclassical quantization of the harmonic oscillator.

Let us now consider the time-evolution of the phase-space function of the harmonic oscillator, where we assume that the initial state of the harmonic oscillator is a coherent state $|\alpha\rangle$ with complex amplitude $\alpha = (q+ip)/\sqrt{2}$. Since we have $|\alpha(t)\rangle = \exp(-i\hat{H}t)|\alpha\rangle = |\alpha e^{-it}\rangle$ for the harmonic oscillator Hamiltonian $\hat{H}$, the phase-space function $\psi_s(r,k;t) = \langle \psi_s(r,k;s)|\alpha(t)\rangle$ is given by

$$\psi_s(r,k;t) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{4}(r-q_t)^2 - \frac{1}{4}(k-p_t)^2 + \frac{1}{2}i(p_t r - q_t k + skr) \right], \tag{4.70}$$

where the functions q_t and p_t are the solutions of the classical equations of motion for the harmonic oscillator, that is, $q_t = q\cos t + p\sin t$ and $p_t = p\cos t - q\sin t$. The probability distribution of the phase-space variables of the harmonic oscillator becomes

$$|\psi_s(r,k;t)|^2 = \frac{1}{2\pi}\exp\left[-\frac{1}{2}(r-q_t)^2 - \frac{1}{2}(k-p_t)^2 \right], \tag{4.71}$$

which means that the most probable path of the coherent state is identical with the classical trajectory $\mathbf{l}(t) = (q_t, p_t)$ in the phase space of the harmonic oscillator. The detailed investigations of the phase-space functions have been done by the Hilbert-space-vector approach[33] and by the linear transformation approach.[34]

V. SUMMARY

We have obtained the phase-space representation of quantum state vectors by means of the relative-state formulation.[40,41] The phase-space representation of quantum state vectors is useful for investigating the properties of physical and chemical systems.[33–39] In the relative-state formulation, the Hilbert space is extended by introducing the reference quantum system. The relative-position and relative-momentum states are defined in the extended Hilbert space. The basic operators such as position and momentum operators are expressed as the differential operators acting on these states. After obtaining the relative-position state, the relative momentum state and the expressions of the basic operators, we restrict the extended Hilbert space to the original Hilbert space of the relevant quantum system since we need only the description of the relevant quantum system. Hence we obtain the phase-space representation of quantum state vectors of the relevant system within the framework of the relative-state formulation. Furthermore we have found the linear transformations between the phase-space function and the wave function in the position (momentum) space, which are equivalent to those obtained by Torres-Vega and Frederick[33] and by Harriman[34] under certain conditions. Although in this paper we have confined ourselves to considering the phase-space representation of pure quantum states, it is straightforward to generalize the results in the case that a physical system is in a mixed quantum state described by a statistical operator.

[1] L. D. Landau and E. M. Lifshitz, *Mechanics* (Pergamon, Oxford, 1976).
[2] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, Oxford, 1980).
[3] E. P. Wigner, Phys. Rev. **40**, 749 (1932).
[4] M. Hillery, R. F. O'Connel, M. O. Scully, and E. P. Wigner, Phys. Rep. **106**, 121 (1984).
[5] R. J. Glauber, Phys. Rev. **130**, 2529 (1963).
[6] R. J. Glauber, Phys. Rev. **131**, 2766 (1963).
[7] E. C. G. Sudarshan, Phys. Rev. Lett. **10**, 277 (1963).
[8] K. Husimi, Proc. Phys. Chem. Soc. Jpn. **22**, 264 (1940).
[9] Y. Kano , J. Math. Phys. **6**, 1913 (1965).
[10] D. F. Walls and G. J. Milburn, *Quantum Optics* (Springer, Berlin, 1994).
[11] U. Leonhardt, *Measuring the Quantum State of Light* (Cambridge University Press, Cambridge, 1997).
[12] E. Arthurs and J. K. Kelly, Jr., Bell Syst. Tech. J. **44**, 725 (1965).
[13] J. H. Shapiro and S. S. Wagner, IEEE Trans. Quantum Electron. **QE-20**, 803 (1984).
[14] N. G. Walker, J. Mod. Opt. **34**, 15 (1987).
[15] Y. Lai and H. A. Haus, Quantum Opt. **1**, 99 (1989).
[16] S. Stenholm, Ann. Phys. (N.Y.) **218**, 233 (1992).
[17] C. T. Lee, Phys. Rev. A **44**, 2775 (1991).
[18] K. E. Cahill and R. J. Glauber, Phys. Rev. **177**, 1857 (1969).
[19] K. E. Cahill and R. J. Glauber, Phys. Rev. **177**, 182 (1969).
[20] G. S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2161 (1970).
[21] G. S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2187 (1970).
[22] G. S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2206 (1970).
[23] K. Wódkiewicz, Phys. Rev. Lett. **52**, 1064 (1984).
[24] K. Wódkiewicz, Phys. Lett. A **115**, 304 (1986).
[25] V. Bužek, C. H. Keitel, and P. L. Knight, Phys. Rev. A **51**, 2575 (1995).
[26] V. Bužek, C. H. Keitel, and P. L. Knight, Phys. Rev. A **51**, 2594 (1995).
[27] A. Wünsche and V. Bužek, Quantum Semiclassic. Opt. **9**, 631 (1997).
[28] E. Prugovečki, J. Math. Phys. **17**, 517 (1976).
[29] E. Prugovečki, J. Math. Phys. **17**, 1673 (1976).
[30] S. Twareque Ali and E. Prugovečki, J. Math. Phys. **18**, 219 (1977).
[31] E. Prugovečki, Ann. Phys. (N.Y.) **110**, 102 (1978).
[32] P. Busch, M. Grabowski, and P. J. Lahti, *Operational Quantum Physics* (Springer, Berlin, 1995).
[33] Go. Torres-Vega and J. H. Frederick, J. Chem. Phys. **98**, 3103 (1993).
[34] J. E. Harriman, J. Chem. Phys. **100**, 3651 (1994).
[35] Go. Torres-Vega and J. H. Frederick, J. Chem. Phys. **93**, 8862 (1990).
[36] K. B. Moller, T. G. Jorgensen, and Go. Torres-Vega, J. Chem. Phys. **106**, 7228 (1997).
[37] Go. Torres-Vega, J. Chem. Phys. **98**, 7040 (1993).
[38] Go. Torres-Vega, J. Chem. Phys. **99**, 1824 (1993).
[39] Go. Torres-Vega and J. D. Morales-Guzmán, J. Chem. Phys. **101**, 5847 (1994).
[40] M. Ban, J. Math. Phys. **32**, 3077 (1991).
[41] M. Ban, Phys. Rev. A **48**, 3452 (1993).
[42] M. Ban, J. Opt. Soc. Am. B **9**, 1189 (1992).
[43] M. Ban, Opt. Commun. **94**, 231 (1992).
[44] M. Ban, Phys. Lett. A **176**, 47 (1993).
[45] M. Ban, Phys. Rev. A **50**, 2785 (1994).
[46] M. Ban, Phys. Rev. A **51**, 2469 (1995).

[47] M. Ban, Phys. Lett. A **1995**, 275 (1995).
[48] In mathematical sense, a state vector $|\chi\rangle$ which is an eigenstate of any unbounded operator with continuous spectrum, such as canonical position and momentum operators, cannot be an element of the Hilbert space $\mathscr{H}$. In fact, the state $|\chi\rangle$ obeys the δ-function normalization and cannot have the finite value of norm. To define such a state vector rigorously, we have to introduce the Gel'fand triplets or the rigged Hilbert space [see A. Böhm, *The Rigged Hilbert Space and Quantum Mechanics* (Springer, Berlin, 1978); A. Böhm and M. Gadella, *Dirac Kets, Gamov Vectors, and Gel'fand triplets* (Springer, Berlin, 1989)]. In our treatment, however, there is no problem even if the state vector $|\chi\rangle$ is assumed to be an element of the usual Hilbert space $\mathscr{H}$.
[49] Fan Hong-yi and J. R. Klauder, Phys. Rev. A **49**, 704 (1994).
[50] A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer, Dordrecht, 1993).
[51] K. Wódkiewicz, Phys. Lett. A **124**, 207 (1987).

Feynman integral in regularized nonrelativistic quantum electrodynamics

Z. Haba
Institute of Theoretical Physics, University of Wroclaw, Wroclaw, Poland

(Received 8 March 1996; accepted for publication 15 October 1997)

We express the unitary time evolution in nonrelativistic regularized quantum electrodynamics at zero and positive temperature by a Feynman integral defined in terms of a complex Brownian motion. An average over the quantum electromagnetic field determines the form of the quantum mechanics in an environment of a quantum black body radiation. In this nonperturbative formulation of quantum electrodynamics we prove the existence of the classical limit $\hbar \to 0$. We estimate an error to some approximations commonly applied in quantum radiation theory. © *1998 American Institute of Physics.* [S0022-2488(98)02003-9]

I. INTRODUCTION

Quantum electrodynamics (nonrelativistic as well as relativistic) has been formulated in terms of a path integral by Feynman himself.[1] Starting from the Feynman integral, the conventional perturbation series in powers of the fine structure constant $\alpha = e^2(\hbar c)^{-1}$ is derived. It is a common attitude among physicists to treat the Feynman integral just as a tool for the generation of the perturbation series. However, it is well known that the imaginary time (Euclidean) version of the Feynman integral has a mathematical meaning beyond the perturbation theory. In fact, the Euclidean two-dimensional quantum electrodynamics (QED) has been constructed by means of the path integral.[2,3] The path integral for Euclidean regularized QED at finite temperature in any dimension has been discussed in Ref. 4. Although, in principle, the Euclidean correlation functions can be continued analytically to the Minkowski space,[5] in practice, this is an inconvenient approach to dynamics.

Perturbative QED agrees perfectly well with experiments. Its nonperturbative effects have not been discovered yet. However, the photons form an inevitable environment for all events in nature at least in the form of the heat or (much weaker) cosmic microwave background. The permanent influence of radiation may have a profound effect on mesoscopic phenomena. A lot of theoretical studies has been done investigating the effect of an infinite environment on a finite microscopic system (see, e.g., Refs. 6–8). However, the mathematical model of the heat bath could not be identified with any physical environment. The quantum electromagnetic field constitutes such a physical environment. There are suggestions[9] that Leggett's model approximates the effect of the quantum electromagnetic field. However, we think that this is quite a rough approximation, completely out of control. The approximations commonly applied[10–16] neglect the spatial dependence of the electromagnetic field. The neglect of the spatial dependence is hard to justify in QED where the electromagnetic field has singular vacuum correlation functions.

In this paper we apply to QED our earlier mathematical formulation of the Feynman integral[17,18] (see also Refs. 19 and 20). Such a formulation makes it possible to discuss the dynamics directly in the real time. Working with physical time is indispensable in problems concerning finite time dynamics e.g., decoherence[6,21] and radiation reaction. We consider the Dirac model[22] of a finite number of particles interacting by means of the Coulomb forces and additionally with the quantum electromagnetic field. This is the first approximation to the relativistic quantum field theory of interacting quantum electromagnetic field (in the radiation gauge) and matter fields. As pointed out already by Feynman,[1] a minor modification of the path integral defines the complete relativistic model. For this reason the nonrelativistic theory has a similar perturbative expansion in powers of the fine structure constant α. At the moment we are mainly interested in an application of the quantum electromagnetic field as a model of an environment to the conventional nonrelativistic quantum mechanics. For this reason we include a discussion of

 © 1998 American Institute of Physics

mixed states corresponding to a finite temperature. The main virtue of a rigorous formulation of the Feynman integral is an opportunity to estimate an error to formal calculations. In this paper we discuss the radiation damping and quasiclassical expansion. In the perturbative approach to QED we encounter a serious difficulty with the classical limit (see Ref. 23), because α tends to infinity when $\hbar \to 0$. So, a nonperturbative formulation of this problem is required.

The paper is organized as follows. In Sec. II we define the model in terms of the Hamiltonian in a finite mode approximation for the quantum electromagnetic field. In Sec. III we derive the Feynman formula for the time evolution operator in this model. The finite temperature QED is expressed by the functional integral in Sec. IV. In Sec. V we integrate over the quantum electromagnetic field in order to describe the quantum mechanical particles in equilibrium with a radiation background. We consider the limit of an infinite volume and infinite number of modes in Sec. VI. We define the Feynman kernel for the evolution operator in Sec. VII. We prove semiclassical asymptotics for the Feynman kernel in Sec. VIII. In Sec. IX we discuss the approximation of a slowly varying quantum electromagnetic field (the dipole approximation).[10,11,13,14] Finally, Sec. X contains a summary and perspectives for further research.

We put more emphasis on the precise description of quantum dynamics than on the details of the proofs. This is so because the method of the probabilistic representation of the quantum dynamics has been described in our earlier papers[17,18] together with detailed proofs. Our approach in this paper, starting with a finite mode approximation, is a repetition of methods in our earlier work.

II. THE HAMILTONIAN

We consider the electromagnetic field in a finite volume V. We expand the vector potential A in eigenmodes $g(k,\mathbf{x})$:

$$\omega(k)^2 g(k,\mathbf{x}) = c^2 \Delta g(k,\mathbf{x}).$$

In this formula $\omega(k)$ denotes the eigenfrequencies and k numerates the eigenmodes (so k is a vector if V is a parallelepiped; this will be tacitly assumed further on, but we write some formulas in a way suggesting cavities of a general shape). We assume that $g(k,x)$ are real (see, e.g., Refs. 24 and 25 for such a choice of eigenmodes), normalized as follows:

$$\int_V g(k,\mathbf{x}) g(k',\mathbf{x})\, dx = \delta(k,k') V.$$

The δ-function is equal to 1 for the same g and equal to 0 if the eigenmodes are different. The expansion in eigenmodes takes the form

$$A_j(x) = \sum_{k,\nu} \sqrt{\frac{2\pi\hbar}{\omega(k)V}}\, c f(k) \mathscr{E}_j(k,\nu)(a(k,\nu) + a(k,\nu)^+) g(k,x). \tag{1}$$

In Eq. (1) $\mathscr{E}$ is the polarization transversal to k,

$$\sum_{\nu=1,2} \mathscr{E}_j(k,\nu)\mathscr{E}_l(k,\nu) \overset{\text{def}}{=} \delta^{tr}_{jl} = \delta_{jl} - k_j k_l |\mathbf{k}|^{-2},$$

$a(k,\nu)$ and $a(k,\nu)^+$ are the amplitudes fulfilling the commutation relations for creation and annihilation operators, and $f(k)$ is a formfactor regularizing the electromagnetic field. Such a formfactor has been introduced already by Pauli and Fierz[26] [we choose $f(k)$ real; with no regularization $f(k)=1$].

The free Hamiltonian is

$$H_R = \sum_{k,\nu} \hbar\omega(k) a(k,\nu)^+ a(k,\nu) = \sum_{k,\nu} \frac{1}{2\mu(k)} p(k,\nu)^2 + \frac{1}{2}\mu(k)\omega(k)^2 y(k,\nu)^2, \tag{2}$$

where we defined

$$a(k,\nu)=\sqrt{\frac{\mu(k)\omega(k)}{2\hbar}}\,y(k,\nu)+\sqrt{\frac{1}{2\mu(k)\omega(k)\hbar}}\,ip(k,\nu)$$

with

$$\mu(k)=\frac{4\pi f(k)^2}{\omega(k)^2 V}.$$

In this way we have expressed the radiation Hamiltonian as a sum of the Hamiltonians for harmonic oscillators. We consider the system's Hamiltonian describing an interaction of the rth particle with an external potential u_r, the interaction with photons (the coupling e_r/c, where e_r denotes an electric charge and c the light velocity), and the Coulomb interaction ϕ between the particles. The Coulomb interaction is related to the quantum radiation by the gauge invariance

$$\phi(\mathbf{x}_r,\mathbf{x}_s)=4\pi\sum_k\ f(k)^2c^2\omega(k)^{-2}g(k,\mathbf{x}_r)g(k,\mathbf{x}_s).$$

Now, the system's Hamiltonian reads

$$H_P=\sum_r\ \frac{1}{2m_r}\left(\mathbf{p}_r+\frac{e_r}{c}\,\mathbf{A}(\mathbf{x}_r)\right)^2+\sum_{r,s}\ e_re_s\phi(\mathbf{x}_r,\mathbf{x}_s)+\sum_r\ u_r(\mathbf{x}_r). \tag{3}$$

When the number of modes is finite, then the Hamiltonian H_R+H_P is defined as a self-adjoint operator in a tensor product of a finite number of $L^2(dy)$ Hilbert spaces.

III. PROBABILISTIC REPRESENTATION OF QUANTUM DYNAMICS

If we restrict ourselves to a finite number of modes and quantize the Hamiltonian $H=H_R+H_P$ canonically, then we obtain a second-order [in $\mathbf{x}_r$ and $y(k,\nu)$] differential operator. A probabilistic representation of the unitary group generated by H has been discussed in our earlier papers.[17]

Let us consider first a single harmonic oscillator. Denote its Hamiltonian by H_k. The ground state χ_k of H_k is

$$\chi_k(y)=\exp\left(-\frac{\mu(k)\omega(k)y^2}{2\hbar}\right). \tag{4}$$

Let us define the Brownian motion as the Gaussian process with the covariance

$$E[b_tb_s]=\min(t,s).$$

Let $q(t,y)$ be the solution of the stochastic equation

$$dq=-i\omega(k)q\,dt+\lambda\sigma_k db \tag{5}$$

with the initial condition y. Here

$$\lambda=\frac{1}{\sqrt{2}}\,(1+i)$$

and

$$\sigma_k=\sqrt{\frac{\hbar}{\mu(k)}}.$$

Then, the Schrödinger time evolution of the state $\psi=\chi_k\Phi$ can be expressed in the form

$$\psi_t(y)\equiv\left(\exp\left(-iH_k\frac{t}{\hbar}\right)\psi\right)(y)=\chi_k(y)E[\Phi(q(t,y))]. \tag{6}$$

The free electromagnetic field is just a collection of harmonic oscillators. Let χ be the ground state of H_R, i.e., $H_R\chi=0$. Let again $\psi=\chi\Phi$. Then

$$\psi_t(\mathbf{A})\equiv\left(\exp\left(-iH_R\frac{t}{\hbar}\right)\psi\right)(\mathbf{A})=\chi(\mathbf{A})E[\Phi(\mathscr{A}(t,\mathbf{A}))], \tag{7}$$

where

$$\mathbf{A}=\sum_{k,\nu}\sqrt{4\pi\mu(k)}cf(k)\mathscr{E}(k,\nu)y(k,\nu)g(k,\mathbf{x})$$

and

$$\mathscr{A}(t,\mathbf{A})=\sum_{k,\nu}\sqrt{4\pi\mu(k)}cf(k)\mathscr{E}(k,\nu)q(t;y,k,\nu)g(k,\mathbf{x}), \tag{8}$$

where

$$q(t;y,k,\nu)=\exp(-i\omega(k)t)y(k,\nu)+\lambda\sigma_k\int_0^t\exp(-i\omega(k)(t-\tau))\,db(\tau;k,\nu)$$

is the solution of Eq. (5) with the initial condition $y(k,\nu)$ (now each Brownian motion has an index numerating the eigenmodes of the electromagnetic field; the Brownian motions corresponding to different indices are independent).

In order to derive a probabilistic representation of the Schrödinger evolution for H_R+H_P we need some additional cutoffs. So, let us assume at the beginning that the number of modes is finite. Next, we need to bound the behavior of $\mathbf{A}(\mathbf{x})$ as well as $\phi(\mathbf{x}_r,\mathbf{x}_s)$ for complex $\mathbf{x}$. In Ref. 17 we have bounded $\mathbf{A}$ and ϕ in the Feynman integral, multiplying these functions by a Gaussian damping factor depending on the Brownian motion. We introduce here an alternative regularization which is useful in QED. We regularize the modes $g(k,\mathbf{x})$ as follows:

$$g(k,\mathbf{x})\to g_\delta(k,\mathbf{x})=\exp(-\delta^2|\mathbf{x}|^8)g(k,\mathbf{x}). \tag{9}$$

Here $\mathbf{A}$ with the g_δ modes replacing g does not fulfill the sourceless Maxwell equations. For this reason the replacement $g\to g_\delta$ is treated as a regularization which subsequently is to be removed. The regularization (9) can be related to the one of our earlier papers[17] by means of the Laplace transform

$$\exp(-\delta^2|\mathbf{x}|^8)=\int_0^\infty ds h(s)\exp(-s\sqrt{\delta}|\mathbf{x}|^2), \tag{10}$$

where $h(s)$ is determined by the inverse Laplace transform [we made use of $(iy)^4=y^4$]

$$h(s)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dy\,\exp(-y^4+iys).$$

It follows from Eq. (10) that h is a real, square integrable, and bounded function on $\mathbb{R}$.

We consider an analytic wave function $\psi(\lambda\mathbf{x})$ at $\lambda\mathbf{x}\in\mathbb{C}^3$, where $\lambda=\sqrt{i}$. In such a case $(\lambda(\mathbf{x}+\sigma\mathbf{b}))^8=(\mathbf{x}+\sigma\mathbf{b})^8$. Hence, in the representation (10) if $\mathbf{x}\to\lambda(\mathbf{x}+\sigma\mathbf{b})$, then on the rhs of Eq. (10) we obtain $\exp(-s\sqrt{\delta}(\mathbf{x}+\sigma\mathbf{b})^2)$. Then, the regularization (9) is equivalent to the one of our earlier paper.[17] Applying the methods of that paper we now obtain the following Feynman formula.

Theorem 1: Consider the Hamiltonian H_R+H_P [Eqs. (2) and (3)], where the number of modes is finite and $g(k,\mathbf{x})$ are replaced by $g_\delta(k,\mathbf{x})$ (this Hamiltonian is denoted H_δ). Assume that $g_\delta(k,\mathbf{x})$ can be analytically extended to the region V_C,

$$V_C=(\mathbf{z}\in\mathbb{C}^3:\mathbf{z}=\mathbf{x}+(1+i)\mathbf{y},\mathbf{x}\in V,\mathbf{y}\in\mathbb{R}^3),$$

and remain bounded analytic functions in V_C. Assume that the external potentials u_r are also analytic and bounded in V_C. Let us consider the initial state of the form $\chi\Phi$ where χ is the ground state of H_R and Φ is square integrable. Then, the unitary semigroup U_t^δ $(t\geqslant 0)$ generating the Schrödinger time evolution has for a sufficiently small time t the probabilistic representation

$$\begin{aligned}(U_t^\delta\psi)(\mathbf{A},\lambda\mathbf{x})\equiv(\exp(-iH_\delta t/\hbar)\psi)(\mathbf{A},\lambda\mathbf{x})=\chi(\mathbf{A})E\Bigg[\exp\Bigg(i\sum_r\frac{e_r}{\hbar c}\int_0^t\mathscr{A}_\delta(\tau,\lambda\mathbf{x}_r\\ +\lambda\sigma_r\mathbf{b}_r(\tau))\lambda\sigma_r d\mathbf{b}_r(\tau)\Bigg)\exp\Bigg(-\frac{i}{\hbar}\sum_{r,s}e_re_s\int_0^t\phi_\delta(\lambda\mathbf{x}_r+\lambda\sigma_r\mathbf{b}_r(\tau),\lambda\mathbf{x}_s\\ +\lambda\sigma_s\mathbf{b}_s(\tau))\,d\tau-\frac{i}{\hbar}\sum_r\int_0^t u_r(\lambda\mathbf{x}_r+\lambda\sigma_r\mathbf{b}_r(\tau))\,d\tau\Bigg)\Phi(\mathscr{A}_\delta(t,\mathbf{A}),\lambda\mathbf{x}+\lambda\sigma\mathbf{b}(t))\Bigg].\end{aligned}\tag{11}$$

Remarks:

(1) The index δ at the potentials $\mathscr{A}$ and ϕ means that in their definitions we made the replacement $g\to g_\delta$.
(2) It does not matter whether the stochastic integral in Eq. (11) is in the Stratonovitch or Ito sense[27] because both integrals are equal for transverse vector potentials.
(3) If V is a parallelepiped, then $\cos(\mathbf{kx})$ and $\sin(\mathbf{kx})$ constitute the eigenmodes, hence $g_\delta(k,\mathbf{x})$ are analytic and bounded in V_C.
(4) The wave function in the theorem describes the quantum electromagnetic field interacting with the n-particle system. We denoted the coordinates in the argument of ψ and Φ collectively by $\mathbf{x}$ and $\mathbf{b}$.
(5) We may write $U_t=(U_{t/n})^n$, where the natural number n is chosen large enough. In such a case we can obtain a probabilistic representation for arbitrarily large time through a repetition of the formula (11).
(6) We consider here only $t\geqslant 0$. The evolution for $t\leqslant 0$ can be obtained by time reflection as discussed in Ref. 17.

Proof: The proof is based on general methods of the construction of semigroups.[28] First, the formula (11) holds true if there is no interaction between the particles and the photons (i.e., when $u_r=0$ and $e_r=0$) as shown in Ref. 17. Then, the factor in the expectation value in Eq. (11) is an example of a nonanticipating multiplicative functional. Such a multiplicative functional defines a semigroup. There remains to prove that the multiplicative functional is integrable [so that the expression (11) is finite] and that the generator of the semigroup is $(i/\hbar)H_\delta$. The integrability results from our assumptions implying that ϕ_δ and $\mathscr{A}_\delta$ are bounded. Then, for each $\mathscr{A}_\delta$,

$$\exp\Bigg(z\int_0^t\mathscr{A}_\delta(\tau,\lambda\mathbf{x}+\lambda\sigma\mathbf{b}(\tau))\lambda\sigma\,d\mathbf{b}(\tau)\Bigg)$$

is integrable for any complex z on the basis of an estimate proved in Ref. 29 and discussed in our earlier paper.[17] There remains the Gaussian integral over $\mathscr{A}_\delta$. Owing to the δ-regularization the covariance of $\mathscr{A}_\delta$ is a bounded continuous function. Hence, the $\mathscr{A}_\delta$-integral is bounded by an exponential of a continuous bilinear form in the Brownian motion $\mathbf{b}$ (this bilinear form is discussed in more detail in Sec. V). The expectation value of such an exponential is finite for a sufficiently small time (see Refs. 17 and 27). We can refer now to the Fubini theorem in order to perform first the $\mathscr{A}_\delta$-integral and to conclude that the expectation value in Eq. (11) is finite. Finally, the computation of the generator involves the limit $t\to 0$. In such a case the calculations are reduced to an elementary stochastic calculus[27] which does not depend on whether the poten-

tials are real or complex. Let us only explain in more detail the quantum electromagnetic part which constitutes a new ingredient in comparison to Ref. 17. In the differentiation of $\mathscr{A}_\delta(t,\mathbf{x})$ over t we expand $\mathscr{A}_\delta$ in $q(t;k,\nu)$. Then, the differentiation of q is determined by Eq. (5). As a result of the $q(t;k,\nu)$ differentiation we obtain the photon part of H:

$$\chi^{-1}H_R\chi=-\sum_{k,\nu}\left(\frac{\hbar^2}{2\mu(k)}\frac{\partial^2}{\partial y(k,\nu)\partial y(k,\nu)}+\hbar\omega(k)y(k,\nu)\frac{\partial}{\partial y(k,\nu)}\right).$$

The remaining terms result from differentiation in time in the same way as in the conventional form of the Feynman–Kac formula.[27] □

Let us mention the imaginary time (Euclidean) version of Eqs. (5) and (11) (the Feynman–Kac formula). The Feynman–Kac formula can be applied to more singular potentials. However, there are stronger conditions on the possible decrease of the scalar potential to minus infinity. The imaginary time version can be useful in a study of the spectrum, but is unsuitable for an investigation of dynamics. So, the imaginary time version of Eq. (5) reads

$$dq(t;k,\nu)=-\omega(k)q(t;k,\nu)dt+\sigma_k db(t;k,\nu), \tag{12}$$

whereas the imaginary time version of Eq. (11) takes the form

$$\begin{aligned}\left(\exp\left(-H\frac{t}{\hbar}\right)\psi\right)(\mathbf{A},\mathbf{x})=\chi(\mathbf{A})E\Bigg[&\exp\left(i\sum_r\frac{e_r}{\hbar c}\int_0^t\mathscr{A}(\tau,\mathbf{x}_r+\sigma_r\mathbf{b}_r(\tau))\sigma_r\,d\mathbf{b}_r(\tau)\right)\\&\times\exp\left(-\frac{1}{\hbar}\sum_{r,s}e_re_s\int_0^t\phi(\mathbf{x}_r+\sigma_r\mathbf{b}_r(\tau),\mathbf{x}_s+\sigma_s\mathbf{b}_s(\tau))\,d\tau\right.\\&\left.-\frac{1}{\hbar}\sum_r\int_0^t u_r(\mathbf{x}_r+\sigma_r\mathbf{b}_r(\tau))d\tau\right)\Phi(\mathscr{A}(t,\mathbf{A}),\mathbf{x}+\sigma\mathbf{b}(t))\Bigg],\end{aligned} \tag{13}$$

where $\mathscr{A}(t,\mathbf{x})$ is expressed by $q(t;k,\nu)$ in the same way as in the real time case [Eq. (8)]. Equation (13) holds true under the assumption that the number of modes is finite and $g(k,\mathbf{x})$ are bounded functions on $\mathbb{R}^3$ (in such case no δ-regularization is needed). It is easy to take the limit of an infinite number of modes if the formfactors $f(k)$ decay sufficiently fast for large k (e.g., $|f(k)|\leqslant\text{const}|\mathbf{k}|^{-2}$).

IV. PROBABILISTIC DESCRIPTION OF QUANTUM DYNAMICS AT FINITE TEMPERATURE

A mixed state can be described as a random pure state. Let us begin with one degree of freedom. Assume that the (classical) probability ρ_n of a state $|n\rangle$ is given. We introduce a random variable w such that

$$E[\overline{w^n}w^m]=\delta_{mn}n!\rho_n.$$

Then, the density matrix can be expressed in the form

$$\rho=E[|\chi\rangle\langle\chi|], \tag{14}$$

where

$$|\chi\rangle=\sum_{n=0}^{\infty}\frac{1}{\sqrt{n!}}w^n|n\rangle. \tag{15}$$

Let us consider the quantum harmonic oscillator with the frequency $\omega(k)$ at temperature T. If $|n\rangle$ is the nth excited energy eigenstate of the oscillator, then the Gibbs distribution is

$$\rho_n=(1-\exp(-\beta\hbar\omega(k)))\,\exp(-\beta\hbar\omega(k)n),$$

where $\beta=(KT)^{-1}$ and K is the Boltzman constant. It is easy to see that the expectation value (14) can be realized by the Gaussian distribution

$$\mathrm{Tr}(\rho F)=\mathrm{Tr}(E[|\chi\rangle\langle\chi|F])=(1-\exp(-\beta\hbar\omega(k)))\mathrm{Tr}\left(\int d\rho_k(w)F|\chi\rangle\langle\chi|\right),$$

where $d\rho_k$ is the Gaussian probability measure

$$d\rho_k(w)=\pi^{-1}dwd\bar{w}\exp(\beta\hbar\omega(k))\exp(-|w|^2\exp(\beta\hbar\omega(k))).$$

We can sum up the series (15). In the coordinate representation we obtain an expression for the coherent state

$$\langle y|\chi_k(z)\rangle=\left(\frac{\eta_k}{\pi}\right)^{1/4}\exp\left(-\frac{1}{2}\eta_k y^2-\frac{1}{4\eta_k}(z^2+|z|^2)+zy\right),$$

where

$$\eta_k=\frac{\mu(k)\omega(k)}{\hbar}$$

and we defined

$$z=\sqrt{2\eta_k}w.$$

Under the time evolution

$$\chi_k(t,y)\equiv\langle y|\exp\left(-iH_k\frac{t}{\hbar}\right)|\chi(z)\rangle=\langle y|\chi(\exp(-i\omega(k)t)z)\rangle. \tag{16}$$

According to our general scheme[30] the time evolution of the wave function $|\chi\rangle$ [Eq. (15)] determines the correlation functions in the mixed state ρ. In more detail, we define first the stochastic process $(0\leqslant\tau\leqslant t)$

$$dq(\tau)=\frac{i\hbar}{m}\chi(t-\tau,q)^{-1}\nabla\chi(t-\tau,q)d\tau+\lambda\sigma\, db(\tau) \tag{17}$$

where m is the mass of a particle and

$$\sigma=\sqrt{\frac{\hbar}{m}}.$$

Then, the correlation functions of Heisenberg operators can be expressed by the solution $q(t,y)$ of the stochastic equation with the initial condition y, e.g.,

$$\mathrm{Tr}(y(t)y(0)\rho)=E\left[\int dy|\chi_t(y)|^2 y q_t(y)\right]. \tag{18}$$

In the case of the coherent state χ_k, Eq. (17) reads

$$dq(\tau)=-i\omega(k)qd\tau+\frac{i\hbar}{\mu(k)}z\exp(-i\omega(k)(t-\tau))d\tau+\lambda\sigma_k\, db(\tau). \tag{19}$$

Its solution entering formula (18) is

$$q(t;y,z)=\exp(-i\omega(k)t)y+\frac{i\hbar}{\mu(k)\omega(k)}z\exp(-i\omega(k)t)\sin(\omega(k)t)$$
$$+\lambda\sigma_k\int_0^t\exp(-i\omega(k)(t-\tau))db(\tau). \tag{20}$$

The correlation functions in Eq. (18) can easily be computed from Eqs. (16) and (20):

$$\mathrm{Tr}(y(t)y(0)\rho)=(1-\exp(-\beta\hbar\omega(k)))\int d\rho_k(w)dy|\langle y|\chi_k(\exp(-i\omega(k)t)w)\rangle|^2$$
$$\times\left(\exp(-i\omega(k)t)y^2+i\sqrt{\frac{2}{\eta_k}}\sin(\omega(k)t)\exp(-i\omega(k)t)wy\right)$$
$$=\frac{\hbar}{2\mu(k)\omega(k)}\left(\mathrm{cth}\left(\frac{1}{2}\beta\hbar\omega(k)\right)\cos(\omega(k)t)-i\sin(\omega(k)t)\right). \tag{21}$$

According to formulas (2) and (8), the free quantum electromagnetic field is a sum of oscillators. The density matrix is a product of the density matrices with different masses and frequencies. Hence, we obtain directly from Eqs. (8) and (21)

$$\mathrm{Tr}\,(A_j(t,\mathbf{x})A_l(0,\mathbf{x}')\rho)\equiv G_\beta(\mathbf{x},\mathbf{x}';t)_{jl}=\frac{2\pi\hbar c}{V}\sum_k\frac{f(k)^2}{|\mathbf{k}|}\delta^{tr}_{jl}g(k,\mathbf{x})g(k,\mathbf{x}')$$
$$\times\left(\mathrm{cth}\left(\frac{1}{2}\beta\hbar\omega(k)\right)\cos(\omega(k)t)-i\sin(\omega(k)t)\right). \tag{22}$$

The expectation value in the ground state χ of the free electromagnetic field in a cavity is a special case of Eq. (22) corresponding to the limit $\beta\to\infty$ $(t\geqslant 0)$

$$D_{jl}(\mathbf{x},\mathbf{x}';t)\equiv\langle\chi|A_j(t,\mathbf{x})A_l(0,\mathbf{x}')|\chi\rangle=\frac{2\pi\hbar c}{V}\sum_k\frac{f(k)^2}{|\mathbf{k}|}\delta^{tr}_{jl}g(k,\mathbf{x})g(k,\mathbf{x}')\exp(-i\omega(k)t). \tag{23}$$

The expression for correlation functions of an n-particle system in equilibrium with the black body radiation at temperature T described by the Hamiltonian H_R+H_P (2) and (3) follow directly from our description (14) of a mixed state as a random pure state. In this way we obtain directly from Theorem 1

$$\mathrm{Tr}(U_t^+\Phi_1(\mathbf{A},\lambda\mathbf{x})U_t\Phi_2(\mathbf{A},\lambda\mathbf{x})\rho)=\lambda^3\int\prod_{k,\nu}dy(k,\nu)d\mathbf{x}\;\Phi_1(\mathbf{A},\lambda\mathbf{x})E\Big[\Phi_2(\mathscr{A}^z_\delta(t,\mathbf{A}),\lambda\mathbf{x}$$
$$+\lambda\sigma\mathbf{b}(t))\exp\left(i\sum_r\frac{e_r}{\hbar c}\int_0^t\mathscr{A}^z_\delta(\tau,\lambda\mathbf{x}_r\right.$$
$$\left.+\lambda\sigma_r\mathbf{b}_r(\tau))\lambda\sigma_r\,d\mathbf{b}_r(\tau)\right)\exp\left(-\frac{i}{\hbar}\sum_{r,s}e_re_s\int_0^t\phi_\delta(\lambda\mathbf{x}_r\right.$$
$$+\lambda\sigma_r\mathbf{b}_r(\tau),\lambda\mathbf{x}_s+\lambda\sigma_s\mathbf{b}_s(\tau))\,d\tau-\frac{i}{\hbar}\sum_r\int_0^t u_r(\lambda\mathbf{x}_r$$
$$\left.+\lambda\sigma_r\mathbf{b}_r(\tau))d\tau\right)\Big]. \tag{24}$$

In Eq. (24) $\mathscr{A}^z$ means that in the expression (8) for $\mathscr{A}$ we insert the solution (20), depending on the coherent state random variable z_k as well as on the initial condition $y(k,\nu)$. Then, in Eq. (24) the average is taken over the Brownian motion the random variables z_k as well as over the initial values $y(k,\nu)$ defining $\mathbf{A}$. Note that the left-hand side of Eq. (24) is simple if $\Phi_1=1$, whereas the right-hand side is quite complicated even if $\Phi_1=1$. This is so because the unitarity condition applied at the left-hand side when expressed by a path integral takes a complicated form (a condition of a time invariance of the expectation value is an analog of unitarity).

V. QUANTUM MECHANICS IN QED ENVIRONMENT

In quantum mechanics we usually consider idealized systems consisting of a finite number of particles. However, real systems are open and interact in an uncontrollable way with an infinite

environment. It is believed that if the interaction with the environment is weak, then it is negligible. A model suggested by Caldeira and Leggett[6] indicates that this may be not the case concerning processes on the mesoscopic scale. In particular, the disappearance of the quantum interference may be a result of an interaction with an environment (besides the Caldeira and Leggett papers see also Refs. 21, 31, and 32. However, the model of an environment discussed in the above-mentioned papers does not come from any concrete physical environment (the claims in Refs. 9 and 33 that such an environment can be considered as an approximation to QED are rather unconvincing). There appeared some suggestions[34,35] that QED (e.g., in the form of the black body radiation) constitutes the environment enforcing the appearance of classical properties of macroscopic systems. Unfortunately, the quantum electrodynamics discussed in these papers has been considered in such a rough approximation that a conclusion concerning its effect was rather unfounded.

A restriction to observables associated with a system of particles means that we take an average over all states of the quantum radiation field. Let us note first that the random electromagnetic field (8) consists of two parts: the $\hbar$-independent part (which we call $\mathscr{A}^{\mathrm{vac}}$) depending on the initial values $y(k,\nu)$,

$$\mathscr{A}^{\mathrm{vac}}(t,\mathbf{x})=\sum_{k,\nu}\ \sqrt{4\pi\mu(k)}cf(k)\mathscr{E}(k,\nu)\exp(-i\omega(k)t)y(k,\nu)g(k,\mathbf{x}), \tag{25}$$

and the rest

$$\mathscr{A}^{Q}\equiv\mathscr{A}-\mathscr{A}^{\mathrm{vac}}.$$

An easy computation gives

$$\begin{aligned}E[\mathscr{A}_j^Q(\tau,\mathbf{x})\mathscr{A}_l^Q(\tau',\mathbf{x}')]&=\frac{2\pi\hbar c}{V}\sum_k\ \frac{f(k)^2}{|\mathbf{k}|}\,\delta_{jl}^{tr}g(k,\mathbf{x})g(k,\mathbf{x}')(\exp(-i\omega(k)|\tau'-\tau|)\\&\quad-\exp(-i\omega(k)(\tau'+\tau)))\\&\equiv D_{jl}(\mathbf{x},\mathbf{x}';|\tau'-\tau|)-D_{jl}^{\mathrm{vac}}(\mathbf{x},\mathbf{x}';\tau'+\tau),\end{aligned} \tag{26}$$

where we divided the correlation function (26) into a time-translation invariant part and the rest. If the δ-regularization $g\to g_\delta$ is inserted in Eq. (26), then the corresponding correlation function is denoted D^δ.

We can now calculate transition amplitudes in quantum mechanics in a QED environment. Let $\chi(\mathbf{A})$ be the Fock vacuum for the radiation field. Assume that Φ_1 and Φ_2 are entire analytic functions and depend only on the particle positions. In such a case we can rotate the $\mathbf{x}$ integral into the complex hyperplane $\mathbb{C}^3$. Then, the transition amplitudes can be expressed in the form

$$\begin{aligned}(\Phi_1\chi,U_t\Phi_2\chi)&=\exp\left(\frac{3i\pi}{4}\right)\int d\mathbf{x}\overline{\Phi}_1(\lambda\mathbf{x})E\Big[\Phi_2(\lambda\mathbf{x}+\lambda\sigma\mathbf{b}(t))\\&\quad\times\exp\Big(-i\sum_{rs}\ \frac{e_r\sigma_re_s\sigma_s}{2\hbar^2c^2}\int_0^t\int_0^t d\mathbf{b}_r(\tau)d\mathbf{b}_s(\tau')D^\delta(\lambda\mathbf{x}_r\\&\quad+\lambda\sigma_r\mathbf{b}_r(\tau),\lambda\mathbf{x}_s+\lambda\sigma_s\mathbf{b}_s(\tau');|\tau-\tau'|)\Big)\exp\Big(-\frac{i}{\hbar}\sum_{r,s}\ e_re_s\int_0^t\phi_\delta(\lambda\mathbf{x}_r\\&\quad+\lambda\sigma_r\mathbf{b}_r(\tau),\lambda\mathbf{x}_s+\lambda\sigma_s\mathbf{b}_s(\tau))\,d\tau-\frac{i}{\hbar}\sum_r\ \int_0^t u_r(\lambda\mathbf{x}_r+\lambda\sigma_r\mathbf{b}_r(\tau))d\tau\Big)\Big],\end{aligned} \tag{27}$$

where $d\mathbf{b}d\mathbf{b}D$ means a sum over vector indices of D and vector indices of the Brownian motions of individual particles.

Remarks:

(1) The conditions for applicability of the Fubini theorem are fulfilled in the regularized theory for a sufficiently small time (cf. Theorem 1). Double stochastic integrals are discussed in Refs. 36 and 37.

(2) It is remarkable that the ''vacuum part'' in the exponential of the expectation value (27) is absent. The ''vacuum part'' of Eq. (26) cancelled with the contribution from the expectation value of $\mathscr{A}^{\mathrm{vac}}$ and what remains is the time translation invariant part of the expression (26).

The corresponding formula at finite temperature reads

$$\begin{aligned}\mathrm{Tr}(U_t^+\Phi_1(\lambda\mathbf{x})U_t\Phi_2(\lambda\mathbf{x})\rho)=\exp\left(\frac{3i\pi}{4}\right)\int d\mathbf{x}\ \Phi_1(\lambda\mathbf{x})E\Bigg[\Phi_2(\lambda\mathbf{x}+\lambda\sigma\mathbf{b}(t))\\ \times\exp\Bigg(-i\sum_{rs}\ \frac{e_r\sigma_r e_s\sigma_s}{2\hbar^2c^2}\int_0^t\int_0^t d\mathbf{b}_r(\tau)d\mathbf{b}_s(\tau')G_\beta^\delta(\lambda\mathbf{x}_r\\ +\lambda\sigma_r\mathbf{b}(\tau),\lambda\mathbf{x}_s+\lambda\sigma_s\mathbf{b}_s(\tau');|\tau-\tau'|)\Bigg)\\ \times\exp\Bigg(-\frac{i}{\hbar}\sum_{r,s}\ e_re_s\int_0^t\phi_\delta(\lambda\mathbf{x}_r+\lambda\sigma_r\mathbf{b}_r(\tau),\lambda\mathbf{x}_s\\ +\lambda\sigma_s\mathbf{b}_s(\tau))d\tau-\frac{i}{\hbar}\sum_r\ \int_0^t u_r(\lambda\mathbf{x}_r+\lambda\sigma_r\mathbf{b}_r(\tau))d\tau\Bigg)\Bigg],\end{aligned}\tag{28}$$

where G_β^δ is the finite temperature correlation function (22) with $g\to g_\delta$.

We can see from Eq. (27) that the quantum electromagnetic environment contributes an additional nonlocal interaction (described by D) between particles of the system. One could say that the particles interact through an exchange of photons from the environment. As D (at zero temperature) is proportional to $\hbar$ the additional interaction is of the order of the fine structure constant α. From Eq. (22) we can see that at high temperature the interaction becomes proportional to the temperature. However, because of the light velocity in the denominator, this interaction seems negligible at the room temperatures. Nevertheless, it should be realized that such a comparison of the strength of interactions makes sense only for interactions of the same type. The interaction induced by the QED environment is of different type than the Coulomb interaction ϕ and the external potential u. It depends on velocities and is nonlocal. Such interactions may be relevant on the large time and space scale.

VI. REMOVAL OF CUTOFFS AND REGULARIZATIONS

We restricted ourselves until now to a finite number of modes (an ultraviolet cutoff) in a finite volume V and we introduced additionally the δ-regularization which is demanded by the functional integral representation in the real time (it would be superfluous in the Feynman–Kac formula). Together with a finite number of modes we considered also the formfactor $f(k)$ which can describe a finite distribution of the charge. We can allow the number of modes to increase to infinity if the formfactor decreases sufficiently fast (in k). In such a case the formfactor fulfills the role of the ultraviolet cutoff. We restrict ourselves from now on to the case when V is a cube. Then, $g(k,\mathbf{x})$ consist of pairs $(\cos(\mathbf{kx}),\sin(\mathbf{kx}))$ (with a certain restriction on $\mathbf{k}$ ensuring that the basis is orthogonal, see Refs. 24 and 25). Hence,

$$g(k,\mathbf{x})g(k,\mathbf{x}')\to\cos(\mathbf{kx})\cos(\mathbf{kx}')+\sin(\mathbf{kx})\sin(\mathbf{kx}')=\cos(\mathbf{kx}-\mathbf{kx}')\tag{29}$$

and

$$\omega(k)=c|\mathbf{k}|\,.$$

It can be seen from Eqs. (26) and (27) that in a cube (with periodic boundary conditions) the interaction is invariant under space and time translations. Moreover, if the temperature is equal to

zero, the cube tends to $\mathbb{R}^3$ and the ultraviolet regularization is removed $[f(k)=1]$, then the interaction becomes relativistic invariant (in such a case D coincides with the conventional D_+). It is well known that in the relativistic theory D_+ is singular on the light cone. The origin of this singularity can be seen from Eqs. (26) and (29); the oscillating factor becomes a constant on the light cone, hence the sum over wave vectors tends to infinity. In such a case the double stochastic integral in Eq. (27) is infinite. The ultraviolet problem is unavoidable in QED and presumably can only be solved perturbatively through an expansion of the exponential in power series (which is a power series in α). We consider in this section the remaining regularizations.

Theorem 2: Let V be a cube. Assume that the number of modes is finite. Assume also that the external potentials u_r are bounded on V_C (defined in Theorem 1). Then, the limit $\delta\to 0$ of $\exp(-iH_\delta t/\hbar)\psi(\mathbf{A},\lambda\mathbf{x})$ exists for each $\mathbf{A}$ and $\mathbf{x}$ and sufficiently small t. The limit $\delta\to 0$ defines a unitary evolution solving the Schrödinger equation

$$i\hbar\,\partial_t\psi_t=(H_R+H_P)\psi_t$$

with the initial condition ψ. As a consequence the limits $\delta\to 0$ of the transition amplitude (27) and the finite temperature correlation functions (28) also exist for a sufficiently small time.

Proof: We apply the representation (9) and (10) to the δ-regularization. We expand the exponentials in Eq. (11) in power series in e. Then, it follows from the results of Ref. 17 (Theorem 3) and from the Lebesgue dominated convergence theorem (because $\mathscr{A}_\delta$ and ϕ_δ are bounded) that for each $\mathscr{A}^Q$ the limit $\delta\to 0$ is exchangeable with the sum. There remains the Gaussian integral over $\mathscr{A}^Q$ in the power series. From estimates of the combinatorial factor in front of the term $(\mathscr{A}^Q)^n$ done in Ref. 17 (based on Ref. 29) it follows that it is bounded by $1/\sqrt{n!}$. Hence, after the Gaussian integration of the term $(\mathscr{A}^Q)^n$ (which contributes $\sqrt{n!}$) this term becomes bounded by $c(t)^n$ where $c(t)$ is proportional to t because the upper limit in the stochastic integration is equal to t. Hence, if t is sufficiently small, then the power series is bounded by a convergent geometric series. The proof of the convergence of the power series for the transition amplitude (27) and the correlation functions (28) is similar. In the proof we obtain the Gaussian integral with the covariance (26) instead of the Gaussian integral with the covariance (23). □

Corollary 3: Assume that there exist $a>0$ and $R(a)>0$ such that

$$|f(k)|\leqslant R(a)\exp(-a|\mathbf{k}|^2). \tag{30}$$

(i) Let N be the number of modes. Then, the limit $N\to\infty$ of the transition amplitude (27) and the finite temperature correlation functions (28) exists uniformly in δ for a sufficiently small time t. The limits $N\to\infty$ and $\delta\to 0$ are exchangeable and can be taken simultaneously.

(ii) If time t is sufficiently small, then after $N\to\infty$ the limit $V\to\mathbb{R}^3$ of the transition amplitude and the same limit of the finite temperature correlation functions can be performed according to the rule

$$\frac{1}{V}\sum_k \to \frac{1}{(2\pi)^3}\int d^3k.$$

The limit $V\to\mathbb{R}^3$ is uniform in δ. Hence, the limits $V\to\mathbb{R}^3$ and $\delta\to 0$ are exchangeable.

Remark: We could define directly the evolution operator U_t (11) for an infinite number of modes and a sufficiently small time by the same argument which we apply in the proof below. In principle, an operator can be recovered from the transition amplitude.

Proof: It is sufficient to note that with the bound on the formfactor $f(k)$ the limits considered in the Corollary can be taken at each order of the perturbation series in e. This is so, because according to the results of Ref. 17 (proven under a condition equivalent to the bound on the formfactor assumed here) and Theorem 2 the perturbation series is convergent. Then, the existence of the limit $N\to\infty$ is easy to show at each order of the perturbation expansion. The infinite volume limit and the convergence of the sum over k to an integral is a consequence of Riemann's definition of an integral. □

We approached QED starting from a well-defined Hamiltonian consisting of a finite number of modes of the quantum electromagnetic field in a finite volume. This is a model often studied in quantum optics. However, we could consider the final formula (27) where the limits $\delta\to 0$, N

$\to\infty$ and $V\to\mathbb{R}^3$ are performed directly under the expectation value on the right-hand side in the exponential factor. We would obtain an expression defined by the stochastic integrals of D and ϕ. This could be considered as a definition of QED. In fact, if we first integrated over $\mathscr{A}$ in the (imaginary time) Feynman–Kac formula and afterwards performed in a formal way the analytic continuation on the path space, then we would arrive at the formula (27). In Theorem 2 and Corollary 3 we have shown that the limits on the right-hand side exist for a sufficiently small time and define a limit of an evolution operator. If the limits are performed under the expectation value, then the expectation value may be infinite or, even if it is finite, we may lose a connection with the Hamiltonian formalism. In the rest of this section we investigate the double stochastic integral of the formula (27), treating the right-hand side as a definition of the Feynman integral representation of the left-hand side. We emphasize the following result of Ref. 37.

Lemma 4: The sufficient conditions required for a definition of the double stochastic integral in Eq. (27) as an almost surely finite random variable read

$$E\left[\int_0^t d\tau \int_0^\tau d\tau' \ |D(\mathbf{b}(\tau),\mathbf{b}(\tau');\tau-\tau')|^2\right]<\infty \tag{31}$$

and

$$E\left[\int_0^t d\tau \ |D(\mathbf{b}(\tau),\mathbf{b}(\tau);0)|^2\right]<\infty.$$

Applying Lemma 4 we can show the following.

Lemma 5: Under the assumption (30) the double stochastic integral is defined as an almost surely finite random variable.

Proof: Under our assumptions the correlation function D for equal times is a finite constant. It remains to be proven that the integral (31) is finite. The expectation value (31) is equal to

$$\int_0^t d\tau \int_0^\tau d\tau' \ d^3x \ d^3y \ d^3k \ d^3k' (2\pi)^{-3} (\tau'(\tau-\tau'))^{-3/2}$$
$$\times \exp\left(-\frac{\mathbf{x}^2}{2\tau'}-\frac{(\mathbf{y}-\mathbf{x})^2}{2(\tau-\tau')}\right)(|\mathbf{k}||\mathbf{k}'|)^{-1} \ f(k)^2$$
$$\times \exp(-ic(|\mathbf{k}|+|\mathbf{k}'|)(\tau-\tau')+\lambda\sigma(\mathbf{k}\mathbf{y}-\mathbf{k}\mathbf{x})+\overline{\lambda}\sigma(\mathbf{k}'\mathbf{y}-\mathbf{k}'\mathbf{x})).$$

We change variables $(\mathbf{x},\mathbf{y})\to(\mathbf{x},\mathbf{z})$ where

$$\mathbf{x}-\mathbf{y}=\sqrt{\tau-\tau'}\,\mathbf{z}.$$

After such a change of variables and a calculation of the $\mathbf{x}$-integral the integrand becomes regular in τ and τ' whereas the integrability at large k and k' is ensured by the assumption (30). □

Finally, we consider an explicit example

$$f(k)=\exp\left(-\frac{a}{2}|\mathbf{k}|\right). \tag{32}$$

(where $a>0$) which does not fulfill the assumption (30). Nevertheless, we can show the following.

Proposition 6: (i) If $f(k)$ is defined by Eq. (32), then the function D (the case of an infinite volume) has an analytic continuation to the region $V_C(0)=(\mathbf{z}\in\mathbb{C}^3:\mathbf{z}=(1+i)\mathbf{y},\mathbf{y}\in\mathbb{R}^3)$. Moreover, D is bounded in this region.

(ii) Assume that the external potential u is bounded by a quadratic form in the region V_C of Theorem 1. Then, the functional integral (27) with the formfactor (32) is finite for a sufficiently small time.

Proof: A direct calculation from Eq. (23) [see also Eq. (26)] gives

$$D_{jl}(\mathbf{x},\mathbf{x}';\tau-\tau')=\frac{\delta^{tr}_{jl}}{4\pi}\,((a+ic|\tau-\tau'|)^2+(\mathbf{x}-\mathbf{x}')^2)^{-1}.$$

For a single particle the function D in the double stochastic integral of Eq. (27) takes the form

$$D_{jl}(\lambda\mathbf{x}+\lambda\sigma\mathbf{b}(\tau),\lambda\mathbf{x}'+\lambda\sigma\mathbf{b}(\tau');\tau-\tau')$$
$$\equiv\delta^{tr}_{jl}D=\frac{1}{4\pi}\,\delta^{tr}_{jl}(a^2-c^2(\tau-\tau')^2+2iac|\tau-\tau'|+i\sigma^2(\mathbf{x}-\mathbf{x}'+\mathbf{b}(\tau)-\mathbf{b}(\tau'))^2)^{-1}.$$

Here δ^{tr} is a bounded operator. So, it is sufficient to discuss the scalar function D behind δ^{tr} (a more detailed demonstration would involve an application of the Ito formula; we express the stochastic integral with respect to the longitudinal modes by an ordinary integral which would be bounded owing to the regularization). It is easy to see that D is bounded because, if

$$|\tau-\tau'|c\leqslant\frac{a}{2},$$

then $|D|\leqslant(3\pi a^2)^{-1}$ because the real part of the denominator of D is different from zero. If

$$|\tau-\tau'|c\geqslant\frac{a}{2},$$

then $|D|\leqslant(3\pi a^2)^{-1}$, because the imaginary part of the denominator of D is different from zero.

When D is a bounded continuous function and the potential u is quadratically bounded, then the integrand (27) can be bounded by a Gaussian factor. In such a case the well-known theorems (see, e.g., Ref. 27; **b** can be expressed by independent Gaussian random variables) show the integrability proving (ii). □

Even if D is bounded it may be not easy to prove integrability in Eq. (27) for arbitrarily large time. We have to show that the negative part of D decreases sufficiently fast when $|\tau-\tau'|$ grows. If this is not the case we have to continue the formula (27) beyond the time interval where it loses its meaning. We encounter this problem already for a harmonic oscillator. A solution of the problem is well known in this case. For an anharmonic oscillator it is usually discussed in a semiclassical approximation (Maslov indices).

VII. THE FEYNMAN KERNEL OF THE EVOLUTION OPERATOR

We wish to derive an explicit Feynman integral formula for the evolution kernel of the operator (11). We mainly discuss in this section a finite number of modes. However, in principle there is no difficulty in defining an evolution kernel in an infinite number of dimensions. There is no Lebesgue measure in an infinite number of dimensions. For this reason we define a kernel K as a transformation of a set of measures $\mathscr{G}$ into a set of functions $L^2(d\mu)$ integrable with respect to a certain measure μ. In QED we choose for μ a measure defined on vector fields which gives a realization of the Fock space of the transversal electromagnetic free field, i.e., μ is the Gaussian measure with the covariance

$$\hbar\,\delta^{tr}_{jl}(-\Delta)^{-1}.$$

If Φ_1 and Φ_2 belong to $L^2(d\mu)$, then we define the kernel K_t of the unitary evolution U_t by the formula

$$(\Phi_1,U_t\Phi_2)=(\Phi_1,K_t\Phi_2 d\mu)\equiv\int d^3x\; d\mu(\mathbf{C})\overline{\Phi_1}(\mathbf{C},\mathbf{x})K(t,\mathbf{C},\mathbf{x};0,\mathbf{B},\mathbf{x}')\Phi_2(\mathbf{B},\mathbf{x}')d\mu(\mathbf{B})d^3x'.$$

We have obtained a Feynman formula for the kernel of a quantum mechanical Hamiltonian evolution in the case without the quantum electromagnetic field in Ref. 18. In order to extend this formula to the quantum electromagnetic field let us start with a definition of the Gaussian process $0\leqslant\tau\leqslant t$,

$$Q^t(\tau;k,\nu)=\alpha^t(\tau;k,\nu)+\lambda\sigma_k\gamma^t(\tau;k,\nu), \tag{33}$$

where α^t (we shall sometimes conceal some indices which are inessential at the moment) is the solution of the oscillator equation

$$\left(\frac{\partial^2}{\partial\tau^2}+\omega(k)^2\right)\alpha^t(\tau)=0. \tag{34}$$

with the boundary conditions $\alpha^t(0)=X$ and $\alpha^t(t)=Y$. Explicitly

$$\alpha^t(\tau)=\frac{\sin(\omega(k)(t-\tau))}{\sin(\omega(k)t)}X+\frac{\sin(\omega(k)\tau)}{\sin(\omega(k)t)}Y.$$

Here γ^t is the Gaussian process fulfilling the zero boundary conditions $\gamma^t(0)=\gamma^t(t)=0$ with the mean zero and the covariance

$$E[\gamma^t(\tau)\gamma^t(\tau')]=\mathscr{M}(\tau,\tau'), \tag{35}$$

where

$$\mathscr{L}_\tau\mathscr{M}(\tau,\tau')\equiv\left(-\frac{\partial^2}{\partial\tau^2}-\omega(k)^2\right)\mathscr{M}(\tau,\tau')=\delta(\tau-\tau'), \tag{36}$$

i.e., $\mathscr{L}\mathscr{M}=1$. Here $\mathscr{M}$ is the Green's function of the operator $\mathscr{L}$ with the Dirichlet boundary conditions on the interval $[0,t]$. For $t<\pi\omega(k)$, the kernel $\mathscr{M}$ is positive definite, because for such a t the operator $\mathscr{L}$ is positive definite. This is so because the lowest eigenvalue of $-\partial^2$ is $(\pi/t)^2$. In such a case the random variable γ^t is real. The explicit formula for $\mathscr{M}$ reads

$$\begin{aligned}\mathscr{M}(\tau,\tau')=&\frac{1}{\omega(k)}\,\theta(\tau'-\tau)(\sin(\omega(k)(t-\tau'))\cos(\omega(k)(t-\tau))\\&-\operatorname{ctg}(\omega(k)t)\sin(\omega(k)(t-\tau))\sin(\omega(k)(t-\tau')))+\frac{1}{\omega(k)}\,\theta(\tau-\tau')\\&\times(\sin(\omega(k)(t-\tau))\cos(\omega(k)(t-\tau'))\\&-\operatorname{ctg}(\omega(k)t)\sin(\omega(k)(t-\tau))\sin(\omega(k)(t-\tau'))),\end{aligned}$$

where θ is the Heaviside step function.

We can express the process γ^t by the Brownian motion

$$\gamma^t(\tau)=\sin(\omega(k)(t-\tau))\int_0^\tau(\sin(\omega(k)(t-u)))^{-1}db(u). \tag{37}$$

When $\tau<\pi/\omega$, then the integral (37) is well defined, proving that the kernel $\mathscr{M}$ is positive definite.

We insert Q^t (33) into the expansion (8) of the electromagnetic field $\mathscr{A}$ into eigenmodes. We obtain

$$\mathscr{A}(\tau,\mathbf{x};t,\mathbf{B},\mathbf{C})=\mathscr{A}^{cl}(\tau,\mathbf{x};t,\mathbf{B},\mathbf{C})+\lambda\mathscr{H}^t(\tau,\mathbf{x}), \tag{38}$$

where $\mathscr{A}^{cl}$ is the solution of the wave equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}-\Delta\right)\mathscr{A}^{cl}=0.$$

with the two-point boundary condition $\mathscr{A}^{cl}(0,\mathbf{x};t,\mathbf{B},\mathbf{C})=\mathbf{B}(\mathbf{x})$ and $\mathscr{A}^{cl}(t,\mathbf{x};t,\mathbf{B},\mathbf{C})=\mathbf{C}(\mathbf{x})$. $\mathscr{H}^t$ is a quantum fluctuation. Note that the two-point boundary value problem (34) has the unique solution only if $t<\pi/\omega(k)$ for every k. In terms of the initial condition of Eq. (34) **B** has the form

$$\mathbf{B}=\sum_{k,\nu}\sqrt{4\pi\mu(k)}cf(k)\mathscr{E}(k,\nu)X(k,\nu)g(k,\mathbf{x}).$$

We have a similar formula for $\mathbf{C}$ in terms of $Y(k,\nu)$. The decomposition of the quantum field $\mathscr{A}$ (38) into a solution of a classical boundary value problem and a random fluctuation resembles Boyer's stochastic electrodynamics.[38] However, our random fluctuation $\mathscr{H}$ is multiplied by a complex number λ.

The construction of the random electromagnetic field (38) fulfilling the two-point boundary condition is a generalization of the construction of the Brownian bridge v.[27] This is the Gaussian process defined on the interval $[0,1]$ with zero two-point boundary condition and the covariance

$$E[v_j(\tau)v_l(\tau')]=\delta_{jl}(\theta(\tau-\tau')\tau'(1-\tau)+\theta(\tau'-\tau)\tau(1-\tau')).$$

The bridge connecting the points $\lambda\mathbf{x}$ and $\lambda\mathbf{x}'$ which we apply in the Feynman formula has the form

$$\mathbf{p}(\tau;t,\mathbf{x},\mathbf{x}')=\lambda\mathbf{x}'\ \frac{\tau}{t}+\lambda\ \frac{t-\tau}{t}\ \mathbf{x}+\lambda\sigma\sqrt{t}\mathbf{v}\left(\frac{\tau}{t}\right). \tag{39}$$

We can now express the evolution kernel by the bridge as follows.

Theorem 7: (i) Let $0\leqslant t<\min_k(\pi/\omega(k))$. Then the integral kernel of the unitary semigroup U_t of Theorem 1 (under the assumptions of this theorem) has the representation

$$\begin{aligned}K^{\delta}(t,\mathbf{C},\lambda\mathbf{x};0,\mathbf{B},\lambda\mathbf{x}')&=K^{\text{free}}(t,\mathbf{C};0,\mathbf{B})E\left[\exp\left(i\sum_r\ \frac{e_r}{\hbar c}\ \int_0^t\mathscr{A}_{\delta}(\tau,\mathbf{p}_r(\tau);t,\mathbf{B},\mathbf{C})d\mathbf{p}_r(\tau)\right)\right.\\&\left.\times\exp\left(-\frac{i}{\hbar}\sum_{r,s}\ e_re_s\int_0^t\phi_{\delta}(\mathbf{p}_r(\tau),\mathbf{p}_s(\tau))d\tau-\frac{i}{\hbar}\sum_r\ \int_0^t u_r(\mathbf{p}_r(\tau))d\tau\right)\right]\\&\equiv K^{\text{free}}\mathscr{R},\end{aligned} \tag{40}$$

where K^{free} is the Feynman kernel of $\exp(-iH_R\,t/\hbar)$:

$$\begin{aligned}K^{\text{free}}(t,\mathbf{C};0,\mathbf{B})=\prod_k\ (2\pi i\sigma_k^2\ \sin(\omega(k)t))^{-1}\prod_{k,\nu}\ \exp\left(\frac{i}{2\sigma_k^2}\ (X(k,\nu)ctg(\omega(k)t)X(k,\nu)\right.\\\left.+Y(k,\nu)ctg(\omega(k)t)Y(k,\nu)-2X(k,\nu)(\sin(\omega(k)t))^{-1}Y(k,\nu))\right).\end{aligned}$$

(ii) Assume that (i) holds true. Then under the assumptions of Theorem 2 there exists the limit $\delta\to 0$ of $K^{\delta}(t,\mathbf{C},\lambda\mathbf{x};0,\mathbf{B},\lambda\mathbf{x}')$ for each $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{B}$, and $\mathbf{C}$ if t is sufficiently small.

(iii) The limit $V\to\mathbb{R}^3$ exists for a sufficiently small time if $f(k)=0$ for $|\mathbf{k}|>\kappa$ with a certain ultraviolet cutoff κ.

Remarks:

(1) The allowed time in (ii) and (iii) may be smaller than in (i) because we additionally require a convergence of the perturbation series in powers of the electric charge.
(2) The formula for arbitrary time can in principle be derived by a composition of kernels for a small time.

Proof: In (i) we can repeat the argument of Ref. 18 in order to represent the kernel of the evolution operator of Theorem 1 [Eq. (11)] in terms of the corresponding bridge joining the initial and the final points. Here, for the electromagnetic field we apply the bridge (33) corresponding to the harmonic oscillator instead of the v-bridge corresponding to a free particle used in our earlier paper[18] (the imaginary time version of the oscillator bridge is discussed in Ref. 39). In part (ii) we repeat the proof of Theorem 2, showing that the perturbation series is convergent uniformly in δ and that the δ-regularization can be removed term by term. In part (iii) we apply the same argument as in (ii) of Corollary 3 to show that a sum over wave vectors $\mathbf{k}$ converges to an integral

when $V \to \mathbb{R}^3$. This is possible only if the denominator in the bridge Q^t (33) is different from zero. For this reason we assumed the sharp ultraviolet cutoff in (iii). □

It seems that Eq. (40) remains true for any $t \neq \pi/\omega(k)$. A rough argument is as follows: it could be proved in an independent way under our assumptions that the evolution kernel of $H_R + H_P$ has a convergent expansion around the evolution kernel of H_R (see Ref. 40 for a closely related theorem). On the other hand, Eq. (40) defines such an expansion for any $t \neq \pi/\omega(k)$. Hence, it is sufficient to compare the expansion coefficients. These expansion coefficients must be the same because both kernels fulfill the same equation. Another argument which could be applied comes from the fact that an imaginary time version of Eq. (40) [cf. Eq. (13)] holds true for arbitrarily large time. For an operator bounded from below an analytic continuation in time exists (see a discussion in Ref. 41). At points where the right-hand side of Eq. (40) is regular it must coincide with the integral kernel of the analytic continuation of the semigroup $\exp(-Ht/\hbar)$.

VIII. THE SEMICLASSICAL EXPANSION

A nonperturbative construction of QED is necessary for a demonstration of a classical limit of QED. This is so because the conventional perturbative QED involves an expansion in the fine structure constant α, where $\alpha \to \infty$ when $\hbar \to 0$ (see a discussion of some problems concerning this limit in Ref. 23). We could obtain the conventional perturbation expansion expanding Eq. (27) in powers of the electric charge [see Ref. 1 for a relativistic generalization of Eq. (27)]. In this section we obtain a semiclassical form of the Feynman kernel (40) through an expansion in $\hbar$ (the semiclassical expansion of the Feynman kernel in quantum mechanics with a rigorous version of the Feynman integral is discussed in Refs. 18, 42, and 43; for a version without the Feynman integral see Ref. 44). In such an approach the role of the two-point boundary value problem and its quantum correction (38) become visible. After a derivation of the quasiclassical form of the Feynman kernel other versions of the classical limit can be obtained through an expression of the time evolution by the Feynman kernel. In particular, one can see more explicitly the difference between QED and Boyer's stochastic electrodynamics.[38] In the derivation of the semiclassical expansion we follow our earlier paper.[18] The first step is a shift of variables $\gamma \to \gamma + \zeta\beta$ and $\mathbf{v} \to \mathbf{v} + z\mathbf{w}$, where z and ζ are complex numbers.

Lemma 8: Let $\beta(\tau)$ and $\mathbf{w}(\tau)$ be differentiable and their derivatives square integrable on the intervals $[0,t]$ and $[0,1]$, respectively. Moreover, assume that β and $\mathbf{w}$ are equal to zero on the ends of the respective intervals. Then, with the assumptions and notation of (i) of Theorem 7 we have the following identity:

$$\begin{aligned}\mathscr{R} = E\Bigg[\exp\Bigg(&-\frac{1}{2}\int_0^t \sum_{k,\nu} \zeta(k,\nu)^2 \beta(\tau;k,\nu)\mathscr{L}\beta(\tau;k,\nu)\,d\tau - \int_0^t \sum_{k,\nu} \zeta(k,\nu)\mathscr{L}\beta(\tau;k,\nu)\gamma(\tau;k,\nu)\,d\tau \\ &- \int_0^1 \sum_r z_r \frac{d\mathbf{w}_r}{d\tau}\,d\mathbf{v}_r(\tau) - \frac{1}{2}\int_0^1 \sum_r z_r^2 \frac{d\mathbf{w}_r}{d\tau}\frac{d\mathbf{w}_r}{d\tau}\,d\tau + i\sum_r \frac{e_r}{\hbar c}\int_0^t \mathscr{A}_\delta\Big(\tau, \mathbf{p}_r(\tau) \\ &+ \lambda\sqrt{t}\sigma_r z_r \mathbf{w}_r\Big(\frac{\tau}{t}\Big); t, \mathbf{B}+\mathbf{F}, \mathbf{C}\Big) d\Big(\mathbf{p}_r(\tau) + \lambda\sqrt{t}\sigma_r z_r \mathbf{w}_r\Big(\frac{\tau}{t}\Big)\Big)\Bigg) \exp\Bigg(-\frac{i}{\hbar}\sum_{r,s} e_r e_s \int_0^t \phi_\delta\Big(\mathbf{p}_r(\tau) \\ &+ \lambda\sqrt{t}\sigma_r z_r \mathbf{w}_r\Big(\frac{\tau}{t}\Big), \mathbf{p}_s(\tau) + \lambda\sqrt{t}\sigma_s z_s \mathbf{w}_s\Big(\frac{\tau}{t}\Big)\Big) d\tau - \frac{i}{\hbar}\sum_r \int_0^t u_r\Big(\mathbf{p}_r(\tau) + \lambda\sqrt{t}\sigma_r z_r \mathbf{w}_r\Big(\frac{\tau}{t}\Big)\,d\tau\Big)\Bigg],\end{aligned} \tag{41}$$

where $\mathscr{L}$ is defined in Eq. (36) and

$$\mathscr{A}_\delta(\tau;\mathbf{x};t,\mathbf{B}+\mathbf{F},\mathbf{C}) \equiv \mathscr{A}_\delta(\tau;\mathbf{x};t,\mathbf{B},\mathbf{C}) + \mathscr{F}_\delta(\tau,\mathbf{x})$$

with

$$\mathscr{F}_\delta(\tau,\mathbf{x}) = \sum_{k,\nu} \sqrt{4\pi\mu(k)}\, c f(k) \mathscr{E}(k,\nu) \lambda \sigma_k \zeta(k,\nu) \beta(\tau;k,\nu) g_\delta(k,\mathbf{x}).$$

Proof: If z and ζ are real, then Eq. (41) [cf. the definition of $\mathscr{R}$ in Eq. (40)] is a version of the Cameron–Martin formula expressing the transformation of the Gaussian integral under a shift of variables. The validity of the formula (41) for complex z and ζ follows by analytic continuation; this has been discussed first in Ref. 43 and then in Ref. 18. □

Now, we choose $z_r=(\lambda\sigma_r)^{-1}$ and $\zeta(k,\nu)=(\lambda\sigma_k)^{-1}$. We study the limit $\hbar\to 0$ of the wave function $(U_l^{\delta}\psi)(\lambda\mathbf{x})$ of Theorem 1 defined on the complex hyperplane in $\mathbb{C}^3$. When $\lambda\mathbf{x}\to\mathbf{x}$, then we obtain the conventional semiclassical limit with real fields and real trajectories. Let us introduce a field [see Eq. (38) for a definition of $\mathscr{A}^{cl}$]

$$\mathscr{A}^I(\tau,\mathbf{x})\equiv\mathscr{A}^{cl}(\tau,\mathbf{x};t,\mathbf{B},\mathbf{C})+\mathscr{F}(\tau,\mathbf{x}) \tag{42}$$

and a trajectory (it becomes real if $\lambda\mathbf{x}\to\mathbf{x}$ and $\lambda\mathbf{x}'\to\mathbf{x}'$)

$$\mathbf{x}(\tau)=\lambda\mathbf{x}'\,\frac{\tau}{t}+\lambda\,\frac{t-\tau}{t}\,\mathbf{x}+\sqrt{t}\mathbf{w}\left(\frac{\tau}{t}\right).$$

Let us choose $\mathbf{w}_r$ in such a way that $\mathbf{x}_r(\tau)$ is the solution of the Lorentz equation

$$m_r\frac{d^2\mathbf{x}_r}{d\tau^2}=-e_r\boldsymbol{\nabla}_r\sum_s\, e_s\phi(\mathbf{x}_r,\mathbf{x}_s)-\boldsymbol{\nabla}_r u_r(\mathbf{x}_r)-\frac{1}{c}\,\partial_\tau\mathscr{A}^I(\tau,\mathbf{x}_r)+e_r\,\frac{d\mathbf{x}_r}{d\tau}\times\boldsymbol{\nabla}\times\mathscr{A}^I(\tau,\mathbf{x}_r) \tag{43}$$

with the two-point boundary condition $\mathbf{x}(0)=\lambda\mathbf{x}$ and $\mathbf{x}(t)=\lambda\mathbf{x}'$. For a sufficiently small time t (what we assume here) the two-point boundary value problem (43) has the unique solution.[45]

The field $\mathscr{A}^I$ is constructed from the amplitudes $\beta(\tau;k,\nu)$. Let us define

$$\xi(\tau;k,\nu)=\alpha(\tau;k,\nu)+\beta(\tau;k,\nu), \tag{44}$$

where α is defined below Eq. (34). Then, $\mathscr{A}^I$ of Eq. (42) is defined by ξ:

$$\mathscr{A}^I(\tau,\mathbf{x})=\sum_{k,\nu}\,\sqrt{4\pi\mu(k)}cf(k)\mathscr{E}(k,\nu)\xi(\tau;k,\nu)g(k,\mathbf{x}).$$

Now, we choose $\beta(\tau;k,\nu)$ in such a way that $\xi(\tau;k,\nu)$ is the solution of the equation

$$(\partial_\tau^2+\omega(k)^2)\xi(\tau;k,\nu)=\sqrt{\frac{4\pi}{\mu(k)}}\,f(k)\sum_r\, e_r\mathscr{E}(k,\nu)\,\frac{d\mathbf{x}_r(\tau)}{d\tau}\,g(k,\mathbf{x}(\tau)) \tag{45}$$

with the two-point boundary condition $\xi(0;k,\nu)=X(k;\nu)$ and $\xi(t;k,\nu)=Y(k;\nu)$. Equation (45) implies the usual equation for a vector potential of a given particle current

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}-\Delta\right)\mathscr{A}^I(\mathbf{x})=\frac{4\pi}{c}\sum_r\, e_r\,\frac{d\mathbf{x}_r}{d\tau}\,\delta_f(\mathbf{x}_r(\tau),\mathbf{x}), \tag{46}$$

where

$$\delta_f(\mathbf{x},\mathbf{y})=\frac{1}{V}\sum_k\, f(k)^2g(k,\mathbf{x})g(k,\mathbf{y}).$$

Equation (46) is solved with the two-point boundary condition $\mathscr{A}^I(\tau=0,\mathbf{x})=\mathbf{B}(\mathbf{x})$ and $\mathscr{A}^I(\tau=t,\mathbf{x})=\mathbf{C}(\mathbf{x})$. When $f(k)\to 1$, then we obtain the classical electrodynamics of point charges. We wrote Eqs. (43) and (46) without the δ-regularization (9). As a first step we insert again the δ-regularization and prove the semiclassical expansion. Subsequently, we show that the δ-regularization can be removed.

Theorem 9: Let us assume that V is a cube and the number of modes is finite. Define the classical action [where x and ξ solve the classical equations of motion (43) and (45)]

$$S_\delta=\sum_r \int_0^t\left(\frac{m_r 1}{2}\left(\frac{d\mathbf{x}_r}{d\tau}\right)^2-u_r(\mathbf{x}_r(\tau))\right)d\tau-\sum_{r,s}\int_0^t e_r e_s \phi_\delta(\mathbf{x}_r,\mathbf{x}_s)d\tau$$
$$+\sum_{k,\nu}\int_0^t\left(\frac{\mu(k)}{2}\left(\frac{d\xi(k,\nu)}{d\tau}\right)^2-\frac{1}{2}\mu(k)\omega(k)^2\xi(k,\nu)^2\right)d\tau$$
$$+\sum_r \frac{e_r}{c}\int_0^t \mathscr{A}^I_\delta(\tau,\mathbf{x}_r(\tau))d\mathbf{x}_r(\tau). \tag{47}$$

Then (i) for a sufficiently small time t

$$K_\delta(t,\mathbf{C},\lambda\mathbf{x};0,\mathbf{B},\lambda\mathbf{x}')\exp\left(-i\frac{S_\delta}{\hbar}\right)\equiv\Omega_\delta(t,\hbar) \tag{48}$$

is bounded in $\hbar$.

(ii) The limit $S_\delta\to S$ as $\delta\to 0$ exists if t is small enough.

(iii) For a sufficiently small time Ω_δ can be expanded in a power series in the electric charge and the limit $\delta\to 0$ of each term exists uniformly in $\hbar$. Hence, the limits $\delta\to 0$ and $\hbar\to 0$ in $\Omega_\delta(t,\hbar)$ exist and are exchangeable.

Proof: We expand K_δ in powers of $\sqrt{\hbar}$ according to Eq. (41). S_δ of Eq. (47) constitutes the singular part of the expansion. Equations of motion (43) and (45) result from the condition of the cancellation of the terms of order $(\sqrt{\hbar})^{-1}$. There remains to bound the remainder. In the cube $g(k,\mathbf{x})$ are the sine and cosine functions. In such a case the problem of the estimate of the remainder is very similar to the estimate of the remainder for trigonometric interactions done in Ref. 18 (Theorem 6.2). So, after a subtraction of the term S_δ from K_δ we expand the exponentials in Eq. (41) in power series in the electric charge. The series is convergent for a sufficiently small time. We show that the singular terms of order $(\sqrt{\hbar})^{-1}$ cancel identically in the series if the classical equations of motion (43) and (45) are fulfilled. The remaining terms in the series are already bounded in $\hbar$ and δ. As a consequence, the series is convergent uniformly in $\hbar$ and δ. Hence, the limit $\delta\to 0$ can be performed term by term in the convergent power series. The part (ii) of the theorem is an easy consequence of the continuity of solutions of differential equations with respect to variation of parameters if the fields are regular and the time is small enough. The part (iii) follows from our discussion of the expansion of Ω_δ in the power series. □

We can solve the linear equation (45) and insert the solution into Eq. (43). In such a case we obtain an integro-differential equation describing particles interacting by Coulomb forces and additionally under the interaction of the retarded magnetic field produced by their movements. We would obtain this equation if we performed the classical limit in Eq. (27) where the functional integral over the quantum electromagnetic field has been calculated. The double stochastic integral in the exponential of Eq. (27) describes the quantum radiation reaction. Equations (43) and (45) lead to the classical expression for the radiation reaction.[46,47] We have shown in Theorem 9 that the quantum radiation reaction has a classical limit for a sufficiently small time. It is well known that we encounter serious acausal effects with both the classical and quantum radiation reaction (see a clarification of some difficulties in Ref. 48). However, these problems involving a large time behavior are beyond the scope of this paper.

IX. SOME APPROXIMATIONS IN QED

It is quite difficult to obtain nonpertubative results in QED without further approximations. It is often argued (see, e.g., Refs. 11 and 22) that because the electron in an atom is localized around the nucleus and the wavelength involved in the light–atom interaction is large in comparison to the atom size, then in quantum radiation theory we may neglect the spatial dependence of quantum electromagnetic field (such an assumption is also applied in mathematical physics papers[13,14]). Although this argument is convincing classically it would be very difficult to justify it in a complete relativistic quantum theory of the electromagnetic field. In a relativistic theory the propagator D in Eq. (27) is singular on the light cone. Hence, an expansion in coordinates of this

propagator in Eq. (27) could not be justified. We could consider the ultraviolet cutoff in D as an expression of the belief that the high-frequency atom–photon interactions are not relevant to atomic physics.

We intend to formulate this short-range approximation in a particular case. Let us consider the formfactor

$$f(k)^2=\int_0^\infty ds\varsigma(s)\exp(-\epsilon s|\mathbf{k}|). \tag{49}$$

We wish to compare the transition amplitude (27) with the one where the spatial dependence of D is neglected. Let us note that (see the propagator D of Proposition 6)

$$D(\mathbf{x},\mathbf{x}';|\tau-\tau'|)-D(0,0;|\tau-\tau'|)=\frac{1}{4\pi}\,\delta^{tr}\int_0^\infty \varsigma(s)(\mathbf{x}-\mathbf{x}')^2(\epsilon s+ic|\tau-\tau'|)^{-2}$$
$$\times((\epsilon s+ic|\tau-\tau'|)^2+(\mathbf{x}-\mathbf{x}')^2)^{-1}. \tag{50}$$

Hence, with a proper choice of ς this difference can be a bounded function of $\mathbf{x}-\mathbf{x}'$.

Proposition 10: Consider the transition amplitude (27) for a single particle in $\mathbb{R}^3$. Assume that $\int_0^\infty s^{-2}\varsigma(s)ds<\infty$. Then, for a sufficiently small time t,

$$|(\Phi_1\chi,U_t\Phi_2\chi)-(\Phi_1\chi,U_t\Phi_2\chi)_0|\leqslant \text{const}\,\frac{e^2}{\hbar c}\,t, \tag{51}$$

where $(,)_0$ denotes the transition amplitude computed with the propagator $D(0,0;|\tau-\tau'|)$ where the spatial dependence is neglected.

Proof: We note that under the assumptions of the Proposition the difference of the propagators $D(\lambda\sigma\mathbf{b}(\tau),\lambda\sigma\mathbf{b}(\tau');|\tau-\tau'|)-D(0,0;|\tau-\tau'|)$ is a bounded function of $\mathbf{b}$ and $\tau-\tau'$. Then, in Eq. (27) the exponential of the double stochastic integral $\int db\ db(D-D_0)$ (where the propagator without the spatial dependence is denoted D_0) can be expanded in power series, which is convergent for a sufficiently small time. The first term $\int db\ db(D-D_0)$ in the series (51) is estimated by the rhs of Eq. (51). Then, the estimate of the Proposition is a consequence of the convergence of the series. □

The proposition says that under our assumptions an error of the neglect of the spatial dependence is of the order of the fine structure constant. Such a result can be considered as a first step of a more systematic investigation. The substantial result should hold true for arbitrarily large time. However, this would be quite difficult to prove. The difficulty results from the problem of integrability for arbitrarily large time of the exponential of the double stochastic integral. The approximations to the double stochastic integral are interesting from the point of view of the quantum theory of radiation reaction. The well-known formula for classical radiation reaction of a point charge[46] confirmed by approximate quantum calculations[48,49] can be obtained by a neglect of spatial dependence. In order to investigate the problem from the point of view of the Feynman integral (for a Hamiltonian formulation see Ref. 50) let us introduce the decomposition $D=D^0+D^1$ for the zero temperature propagator and $G=G^0+G^1$ for the finite temperature propagator (22), where the index zero means that the spatial coordinates are set equal to zero. In the infinite volume we have (the factor $\frac{2}{3}$ comes from the replacement of δ^{tr} by δ)

$$D^0_{jl}(t)=\frac{2}{3}\,\delta_{jl}\,\frac{\hbar c}{\pi}\int_0^\infty dk\ kf(k)^2\ \exp(-ickt)$$

and

$$G^0(t)_{jl}=\frac{2}{3}\,\delta_{jl}\,\frac{\hbar c}{\pi}\int_0^\infty dk\ kf(k)^2\left(cth\left(\frac{1}{2}\,\beta\hbar ck\right)\cos(ckt)-i\,\sin(ckt)\right).$$

When $f(k)\to 1$, then $\mathrm{Im}\,D^0(t)_{jl}\to(2/3c)\hbar\,\delta_{jl}\delta'(t)$ (where Im denotes the imaginary part). Inserting this result into Eq. (27) we obtain the standard expression for the radiation reaction [strictly

speaking the integral (27) becomes divergent and requires a renormalization]. We can take the radiation reaction into account explicitly by means of a transformation of paths. Let us introduce an operator in $L^2([0,t])$,

$$\mathscr{K}=1+\frac{e^2}{\hbar c^2 m}\,\mathrm{Im}\,D^0,$$

where by $\mathrm{Im}\,D^0$ we denoted the integral operator defined by the kernel $\mathrm{Im}\,D^0$. A transformation

$$d\mathbf{b}\to d\mathbf{b}\equiv\mathscr{K}^{1/2}d\mathbf{b}$$

expresses the functional integral (27) in terms of a new Brownian motion which contains the effect of the radiation reaction. Such a transformation is mathematically correct if $\mathrm{Im}\,D^0$ is a Hilbert–Schmidt operator and moreover if $\mathscr{K}$ is positive definite. For a positive temperature it is useful to note that $G-D$ is real. So, the heat bath does not contribute to the classical radiation reaction. In order to see some effects of the temperature we perform another transformation upon the functional integral. For any functional Φ of the Brownian motion defined on the interval $[0,t]$ we have the Cameron–Martin identity

$$E[\Phi(\mathbf{b})]=E\left[\exp\left(-\int_0^t\frac{d\mathbf{f}}{d\tau}\,d\mathbf{b}(\tau)-\frac{1}{2}\int_0^t\left(\frac{d\mathbf{f}}{d\tau}\right)^2 d\tau\right)\Phi(\mathbf{b}+\mathbf{f})\right]$$

for any $\mathbf{f}$ with a square integrable derivative. We choose now

$$\frac{d\mathbf{f}}{d\tau}=\frac{ie}{\hbar c}\,\lambda\sigma\mathscr{A}^z(0,\tau)$$

in Eq. (24) (for simplicity restricted to a single particle). Then in Eq. (24) $\Phi_2(\mathbf{x}+\lambda\sigma\mathbf{b}(t))\to\Phi_2(\mathbf{q}_t(\mathbf{x}))$ where $\mathbf{q}_t(\mathbf{x})$ is the solution of the equation

$$d\mathbf{q}=-\frac{e}{m}\,\mathscr{A}^z(0,\tau)d\tau+\lambda\sigma d\mathbf{b}(\tau). \tag{52}$$

Using Eqs. (52) and (21) we can compute the correlation functions of q. In this way, for a large temperature and $f(k)=1$ we obtain

$$E\left[\frac{dq_j}{d\tau}\,\frac{dq_l}{d\tau'}\right]=\frac{e^2}{m^2}\,KT\delta_{jl}\delta(\tau-\tau')$$

plus less singular terms. So, the charged particle in equilibrium with a black body radiation behaves like a Brownian particle in a fluid of temperature T. This result is approximate, because we neglected the change of the covariance of the electromagnetic field caused by the Cameron–Martin transformation. We could approach in this way in the framework of QED the model discussed first by Einstein and Hopf[51] and later on in Refs. 52 and 53. Let us note that the behavior of the correlation functions of the quantum particle in equilibrium with the radiation is in qualitative agreement with the requirements of the ''quantum Brownian motion'' due to Benguria and Kac[54] (see also Ref. 6).

X. DISCUSSION AND OUTLOOK

In this paper we concentrated mainly on the precise formulation of regularized QED dynamics. In some cases we were able to show that the expressions were finite only for a small time. Such a requirement resulted from a demand of a convergence of the perturbation series. In principle, we can prolong the dynamics to a larger time using the group property of the time evolution. In practice, the multiple integration involved in such a procedure may be inconvenient. The results of Sec. VII may indicate a way out of this difficulty. The formula for the propagation kernel of a harmonic oscillator holds true for arbitrarily large time (except of the moments equal to the classical periods) in spite of the fact that the functional integral loses its sense already when

t tends to the first period. In another approach to the harmonic oscillator we have obtained in Sec. III the dynamics for an arbitrarily large time through an application of stochastic differential equations [see Eq. (5)]. We have discussed the method of stochastic differential equations for nonlinear quantum systems in Refs. 17 and 55. This method could be applied to QED. The first step has been done in Eq. (52). A nonlinear transformation (discussed for quantum mechanics in Ref. 55) can be applied to formulate exactly the dynamics in QED as a random dynamics. We intend to discuss this method in a subsequent publication.

The dependence on the formfactor (ultraviolet cutoff) constitutes another unsatisfactory aspect of the theory. A nonperturbative removal of the cutoff is a difficult problem. However, some results may be insensitive to the cutoff. One may believe that phenomena on a large time and space scale involving low energies should not depend on the ultraviolet cutoff. We consider the formulation of this paper as a substantial step towards a control over approximations made in QED, quantum optics, and quantum radiation theory. In principle, in Sec. IX we were able to estimate an error involed in a dipole approximation. In Sec. VIII we are able to estimate subsequent terms in the semiclassical expansion.

Note added in proof: After the submission of this paper we have learned about the work of L. H. Ford[56] studying the effect of QED on quantum coherence and the paper of F. Hiroshima[57] who constructs the path integral in QED.

[1] R. P. Feynman, Phys. Rev. **80**, 440 (1950).
[2] J. Challifour and D. Weingarten, Ann. Phys. (N.Y.) **123**, 61 (1979).
[3] D. C. Brydges, J. Fröhlich, and E. Seiler, Commun. Math. Phys. **79**, 353 (1981).
[4] G. Roepstorff, Ann. Phys. (N.Y.) **242**, 386 (1995).
[5] K. Osterwalder and R. Schrader, Commun. Math. Phys. **31**, 83 (1973).
[6] A. O. Caldeira and A. J. Leggett, Physica A **121**, 587 (1983).
[7] G. W. Ford, J. T. Lewis, and R. F. O'Connell, Phys. Rev. A **37**, 4419 (1988).
[8] A. O. Caldeira and A. J. Leggett, Phys. Rev. Lett. **46**, 211 (1981).
[9] P. M. V. B. Barone and A. O. Caldeira, Phys. Rev. A **43**, 57 (1991).
[10] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Photons and Atoms* (Wiley, New York, 1989).
[11] W. H. Louisell, *Quantum Statistical Properties of Radiation* (Wiley, New York, 1973).
[12] R. P. Feynman and F. L. Vernon, Ann. Phys. (N.Y.) **24**, 118 (1963).
[13] V. Bach, J. Fröhlich, and I. M. Sigal, Lett. Math. Phys. **34**, 183 (1995).
[14] M. Hübner and H. Spohn, Rev. Math. Phys. **6** (1994).
[15] Yu. V. Nazarov, Phys. Rev. B **43**, 6220 (1991).
[16] M. Ueda and T. Ando, Phys. Rev. Lett. **72**, 1726 (1994).
[17] Z. Haba, J. Math. Phys. **35**, 6344 (1994).
[18] Z. Haba, J. Phys. A **27**, 6457 (1994).
[19] R. H. Cameron, J. Math. and Phys. **39**, 126 (1960).
[20] H. Doss, Commun. Math. Phys. **73**, 247 (1980).
[21] W. G. Unruh and W. H. Zurek, Phys. Rev. D **40**, 1071 (1989); J. P. Paz, S. Habib, and W. H. Zurek, Phys. Rev. D **47**, 488 (1993).
[22] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Clarendon, Oxford, 1958).
[23] I. Bialynicki-Birula, Acta Phys. Austriaca, Suppl. **XVIII**, 111 (1977).
[24] N. Kroll, in *Quantum Optics and Electronics, Les Houches, 1964*, edited by C. de Witt, A. Blandin, and C. Cohen-Tannoudji (Gordon and Breach, New York, 1965).
[25] L. J. Landau, Phys. Rev. A **37**, 4449 (1988).
[26] W. Pauli and M. Fierz, Nuovo Cimento **15**, 167 (1938).
[27] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[28] E. B. Dynkin, *Markov Processes* (Springer, New York, 1965).
[29] E. Carlen and P. Kree, Ann. Prob. **19**, 354 (1991).
[30] Z. Haba, Phys. Lett. A **206**, 290 (1995).
[31] M. Gell-Mann and J. B. Hartle, in *Complexity, Entropy and the Physics of Information*, edited by W. H. Zurek (Addison–Wesley, Reading, MA, 1990).
[32] Z. Haba, Int. J. Theor. Phys. **35**, 2461 (1996).
[33] P. Ullersma, Physica (Amsterdam) **32**, 27 (1966); N. G. van Kampen, Dan. Mat. Fys. Medd. **26**, 15 (1951).
[34] E. Joos and H. D. Zeh, Z. Phys. B **59**, 223 (1985).
[35] E. Santos, Phys. Lett. A **188**, 198 (1994).
[36] E. Nelson, in *Analysis in Function Space*, edited by W. T. Martin and I. Segal (MIT, Cambridge, 1964).
[37] M. A. Berger and V. J. Mizel, Trans. Am. Math. Soc. **252**, 249 (1979); J. Int. Eqs. **2**, 187 (1980).
[38] T. H. Boyer, Phys. Rev. D **11**, 790 (1975), **11**, 809 (1975).
[39] I. Simao, Stoch. Anal. Appl. **9**, 85 (1991).
[40] A. Weinstein and S. Zelditch, Bull. Am. Math. Soc. **6**, 449 (1982); S. Zelditch, Commun. Math. Phys. **90**, 1 (1983).
[41] J. Feldman, Trans. Am. Math. Soc. **108**, 251 (1963).
[42] S. Albeverio, Ph. Blanchard, and R. Hoegh-Krohn, Commun. Math. Phys. **83**, 49 (1982).
[43] R. Azencott and H. Doss, in *Lecture Notes in Mathematics*, No. 1109 (Springer, New York, 1985).
[44] D. Fujiwara, J. Anal. Math. **35**, 41 (1979).

[45] P. B. Bailey, L. F. Shampine, and P. E. Waltman, *Non-linear Two-Point Boundary Value Problems* (Academic, New York, 1968).
[46] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962).
[47] F. Rohrlich, *Classical Charged Particles-Foundations of Their Theory* (Addison–Wesley, Reading, MA, 1965).
[48] E. J. Moniz and D. H. Sharp, Phys. Rev. D **10**, 1133 (1974); **15**, 2850 (1977).
[49] J. Dalibard, J. Dupont-Roc, and C. Cohen-Tannoudji, J. Phys. (Paris) **43**, 1617 (1982).
[50] A. Arai, J. Math. Phys. **22**, 534 (1981); **24**, 1896 (1983).
[51] A. Einstein and L. Hopf, Ann. Phys. (N.Y.) **33**, 1105 (1910).
[52] T. H. Boyer, Phys. Rev. **182**, 1374 (1969); Phys. Rev. D **19**, 1112 (1979).
[53] P. W. Milonni, Am. J. Phys. **49**, 177 (1981).
[54] R. Benguria and M. Kac, Phys. Rev. Lett. **46**, 1 (1981).
[55] Z. Haba, Lett. Math. Phys. **37**, 223 (1996).
[56] L. H. Ford, Phys. Rev. D **47**, 5571 (1993).
[57] F. Hiroshima, Rev. Math. Phys. **9**, 201 (1997).

Semidirect product of CCR and CAR algebras and asymptotic states in quantum electrodynamics

Andrzej Herdegen[a)]
Institute of Physics, Jagiellonian University, Reymonta 4, 30-059 Cracow, Poland

(Received 3 March 1997; accepted for publication 17 October 1997)

A C^*-algebra containing the CCR and CAR algebras as its subalgebras and naturally described as the semidirect product of these algebras is discussed. A particular example of this structure is considered as a model for the algebra of asymptotic fields in quantum electrodynamics in which Gauss' law is respected. The appearance in this algebra of a phase variable related to electromagnetic potential leads to the universal charge quantization. Translationally covariant representations of this algebra with energy-momentum spectrum in the future lightcone are investigated. It is shown that vacuum representations are necessarily nonregular with respect to total electromagnetic field. However, a class of translationally covariant, irreducible representations is constructed explicitly, which remain as close as possible to the vacuum, but are regular at the same time. The spectrum of energy-momentum fills the whole future lightcone, but there are no vectors with energy-momentum lying on a mass hyperboloid or in the origin. © *1998 American Institute of Physics.* [S0022-2488(98)01703-4]

I. INTRODUCTION AND SUMMARY

In this paper we continue the study of the algebra of asymptotic fields in quantum electrodynamics, in the framework developed earlier in Ref. 1 (and motivated by the classical asymptotic structure discussed in Ref. 2). However, the present work is self-contained: the main results of Ref. 2 are recalled and the construction of Ref. 1 is restated here in a modified form.

The existence of, and the algebraic relations between, the asymptotic (''in'' or ''out'') observables and fields in quantum electrodynamics is a question of great physical interest. In the asymptotic limit, on the one hand, the details and full complication of the dynamics should lose their importance. On the other hand, the consequences of Gauss' law and of the long-range character of the electromagnetic interaction must survive. Suppose a closed algebra of asymptotic fields in quantum electrodynamics may be constructed. This should imply, at least, that (some) states over this algebra are approached in the asymptotic limit by the expectation values of actual fields in physical states. Therefore, the infrared and charged structure of the full theory should be encoded in the asymptotic algebra, and physical insight into this structure may be gained by investigating representations of this algebra.

Investigations into the infrared structure and asymptotic fields of electrodynamics have a long history (see Refs. 3–5 for a review). They have led to the discovery of such structures and effects as superselection sectors of the local observables, the infraparticle problem, and spontaneous breaking of the Lorentz symmetry. However, a clear formulation of a closed algebraic structure of asymptotic fields has not been achieved, although there do exist various partial answers to this problem, with varying balance of mathematical rigor on the one hand and physical concreteness on the other (the asymptotic dynamics of Kulish and Faddeev[6] and Zwanziger,[7] the dressed electron states of Fröhlich,[8] the asymptotic electromagnetic fields[9] and particle weights[10] of Buchholz). The difficulties are twofold. First, we do not have a complete, mathematically sound formulation of QED. Second, the complete asymptotic separation of matter and electromagnetic field may not be expected. These difficulties are not of purely technical nature. The physical factor playing the decisive role in the infrared structure of electrodynamics is the presence of constraints, the Gauss' law. It is not clear at all, in our opinion, to what degree the experience gained by the

[a)]Electronic mail: herdegen@thrisc.if.uj.edu.pl

 © 1998 American Institute of Physics

quantization of simpler, unconstrained, and short-range interactions may be taken over to the formulation of quantum electrodynamics. In fact, the analysis of Refs. 6 and 8 shows that the usual canonical quantization on hypersurfaces of constant, finite time leads to the asymptotic evolution which in the limit takes the states out of the space in which the theory is defined for finite times. This may be understood as an indication that this method of obtaining a quantum theory may not be best suited for electrodynamics. Also, the localization properties of observables which form the base of the axiomatic algebraic approach to quantum field theory problems still need justification in the case of electrodynamics, and may seem somewhat artificial in the version considered there (localization in spacelike cones).

The above remarks should not be understood as an attempt to question *a priori* the standard wisdom on the subject. We would like, rather, to indicate that the problem of the algebraic structure of quantum electrodynamics may still be validly regarded as open. This view is also supported by the failure of the standard electrodynamics to explain such striking and universal physical fact as the observed spectrum of charge.[11]

In the present work we follow the approach of Refs. 1 and 2 to formulate and investigate a concrete C^*-algebra of asymptotic fields in electrodynamics. The basic idea of the approach consists of interchanging the order of ''quantization'' and ''asymptotic limit.'' (Quantization of free electromagnetic field along these lines has been discussed in Refs. 12 and 13. Our approach develops the ideas of Ref. 13 and extends the program to include charged fields.) In Ref. 2 the asymptotic structure of the classical field electrodynamics (the Maxwell–Dirac system) has been discussed. In Ref. 1 then a class of models of the asymptotic algebra has been obtained by the quantization of this classical structure by the correspondence principle. The Gauss' law is naturally implemented on the algebraical level. Now we develop the approach further in the following ways.

(i) We restrict attention to the technically simplest of the models introduced in Ref. 1, the one in which only the Coulomb field is undetachable from a particle [the possibility (b) of Sec. VI in Ref. 1]. While not rejecting other possibilities at this stage, we observe that this model is also the one following most naturally by the quantization of the classical structure. In this case the construction of Ref. 1 may be simplified to yield the asymptotic algebra in the form of a particular instant of a semidirect product of the CCR (canonical commutation relations) and CAR (canonical anticommutation relations) algebras. The definition and the discussion of some mathematical aspects of this structure comprise the self-contained Sec. II of the paper.

(ii) In the course of quantization we bring into play a new factor, which is present at the classical level in Ref. 2, but has not been properly taken into account in the quantization procedure in Ref. 1. It has been shown in Ref. 2 that one of the asymptotic variables has a natural interpretation as a phase variable. The quantization of the classical structure which properly respects the phase character of this variable leads unambiguously to the quantization of the physical charge spectrum in units of elementary charge. As the phase variable is connected with the free electromagnetic potential, this quantization law is universal: it has to be respected by any carriers of charge. (For similar reasoning, but with a different identification of the phase variable, see the works by Staruszkiewicz.[11]) The classical asymptotic structure, its quantization, and the resulting algebra are discussed in Sec. III, including the action of the Poincaré group on the algebra and the identification of observables.

(iii) Section IV contains some general results on physically admissible representations of the asymptotic algebra. It is argued that the representations should be regular with respect to all Weyl operators, as otherwise the Coulomb field part of the total field is lost. In representations satisfying Borchers' criterion (the spectrum of energy-momentum in the future lightcone) the Gauss constraint, which is hidden in the commutation relations on the abstract algebraic level, is shown to be recovered in the functional form.

(iv) More special representations are investigated in Sec. VI with the use of technical tools discussed in Sec. V (some technical material is shifted to the Appendix). All vacuum representations are shown to be nonregular with respect to the Weyl operators with infrared-singular test functions, which is a handicap as explained earlier. Also, superselec-

tion sectors with respect to regular operators are distinguished by the *free* field spacelike asymptotic, which is not what one would expect in a physical state.[14] There exists a Poincaré-invariant vacuum on the field algebra (sectors transform into each other under Lorentz transformations), whose GNS representation space contains charge-one vector states with the energy-momentum on the mass hyperboloid (no infraparticle problem in this vacuum). A class of representations satisfying Borchers' criterion is constructed, which remain as close as possible to this representation, but are regular at the same time. The nonexistence of charged states on the mass hyperboloid follows here from nonexistence of a vacuum vector state. It is not known at present whether there do exist regular Poincaré covariant representations satisfying Borchers' criterion (standard arguments[15,14] do not apply here).

II. SEMIDIRECT PRODUCT OF CCR AND CAR ALGEBRAS

In this section we discuss a generalization of the direct product of CCR and CAR algebras.[33] We also give a class of representations of the resulting C^*-algebra.

Suppose we are given the following constructs:

(i) a complex Hilbert space $\mathscr{K}$ and the *-algebra *CAR generated by the elements $a(f)$ depending antilinearly on $f\in\mathscr{K}$ and by the identity E, according to the CAR relations [(Eqs. (2.2) below];

(ii) an Abelian group X (with the additive notation of the group multiplication) equipped with a symplectic form $X\times X\ni(x,y)\to\{x,y\}\in\mathbf{R}$;

(iii) a representation of the group X in the automorphism group of the *-algebra *CAR,

$$X\ni x\to\beta_x\in \text{Aut } {}^*\text{CAR}, \quad \beta_x\beta_y=\beta_{x+y}.$$

We consider the *-algebra $\mathscr{B}$ generated by the elements $a(f)$ $(f\in\mathscr{K})$, $W(x)$ $(x\in X)$, and the identity E according to the relations:

$$W(x)^*=W(-x), \quad W(0)=E, \quad W(x)W(y)=e^{-(i/2)\{x,y\}}W(x+y), \tag{2.1}$$

$$[a(f),\, a(g)]_+=0, \quad [a(f),\, a(g)^*]_+=(f,g)E, \tag{2.2}$$

$$W(x)C=\beta_x(C)W(x) \quad \forall C\in {}^*\text{CAR}. \tag{2.3}$$

With the use of (2.3) every element of $\mathscr{B}$ may be given the form $\Sigma_{i=1}^n C_i W(x_i)$, where $C_i\in{}^*\text{CAR}$.

Proposition 2.1: (i) If $x_i\neq x_k$ for $i\neq k$ and $C_i\neq 0$ in *CAR, *then* $\Sigma_{i=1}^n C_i W(x_i)\neq 0$. *Every nonzero element of $\mathscr{B}$ is uniquely represented in this way.*

*(ii) The *-algebra $\mathscr{B}$ has a faithful representation by bounded operators in a Hilbert space.*

Proof: Let π_f be a representation of *CAR in a Hilbert space $\mathscr{H}$ (e.g., the Fock representation). Let $\pi(W(y))$ and $\pi(C)$ be operators acting in the direct sum Hilbert space $\oplus_{x\in X}\mathscr{H}_x$, $\mathscr{H}_x=\mathscr{H}$, defined by

$$[\pi(W(y))\psi]_x=e^{(i/2)\{x,y\}}\psi_{x-y}, \quad [\pi(C)\psi]_x=\pi_f(\beta_{-x}C)\psi_x,$$

where $\psi=\oplus_x\psi_x$. One easily shows that these bounded operators satisfy the relations (2.1)–(2.3) and part (i) of the proposition. This implies part (ii). □

This representation defines a C^*-norm on the algebra $\mathscr{B}$. All C^*-norms on $\mathscr{B}$ are jointly bounded, as for each such norm p there is $p(W(x))=1$ and $p(a(f))=\|f\|_{\mathscr{K}}$. Therefore, the set of C^*-norms contains the maximal element defined by $\|A\|:=\sup_p p(A)$.[16]

*Definition 2.1: The field algebra $(\mathscr{F},\|.\|)$ is the C^*-completion of the *-algebra $\mathscr{B}$ in the norm $\|.\|$.*

Remarks: (i) The elements $a(f)$ [resp. $W(x)$] generate a C^*-subalgebra of $\mathscr{F}$ which will be called the CAR (resp. the CCR, or the Weyl) algebra in this article.

(ii) The construction would be slightly more general if the CAR algebra with its automorphisms instead of *CAR and β_x were used. This, however, would not be convenient for our purposes.

Corollary 2.2: (i) Every Hilbert space representation of $\mathscr{B}$ is given in terms of bounded operators and extends to a representation of $\mathscr{F}$.

(ii) Every positive linear functional on $\mathscr{B}$ defines via the GNS construction a Hilbert space representation of $\mathscr{B}$ and extends to a positive linear functional on $\mathscr{F}$.

*(iii) Every *-automorphism of $\mathscr{B}$ extends to a *-automorphism of $\mathscr{F}$.*

We omit a simple proof. The corollary establishes the 1:1 correspondence between the positive functionals and representations of $\mathscr{F}$ on the one hand, and of $\mathscr{B}$ on the other hand. It also gives a class of automorphisms of $\mathscr{F}$ in simple terms. We shall repeatedly (and tacitly, mostly) take advantage of these simplifications in what follows.

An important class of automorphisms of $\mathscr{B}$ is given as follows. Let T be a symplectic additive mapping in X:

$$T(x+y)=Tx+Ty, \quad \{Tx,Ty\}=\{x,y\}. \tag{2.4}$$

Let, further, τ be an automorphism of *CAR satisfying for each $x\in X$

$$\tau\beta_x=\beta_{Tx}\tau. \tag{2.5}$$

Define $\tau(W(x))=W(Tx)$. Then τ extends to an automorphism of $\mathscr{B}$ (and $\mathscr{F}$). This is a Bogoliubov transformation when restricted to the Weyl algebra.

In the following sections the constructions outlined above will be needed in a special case. Both β_x and τ will be Bogoliubov transformations given by

$$\beta_x a(f)=a(S_x f), \quad \tau a(f)=a(Rf), \tag{2.6}$$

where S_x is a unitary representation of X in $\mathscr{K}$ and R is a unitary operator in $\mathscr{K}$. The condition (2.5) now takes the form

$$RS_x=S_{Tx}R. \tag{2.7}$$

In the remaining part of this section we introduce a particular class of representations of $\mathscr{B}$. We pose, namely, the following question. Suppose ω_f and ω_r are states on the CAR and the Weyl algebra, respectively. When does $\omega(CW(x))=\omega_f(C)\omega_r(W(x))$ define a state on $\mathscr{F}$? The hermiticity of ω implies that $(\omega_f(\beta_x C)-\omega_f(C))\omega_r(W(x))=0$ for every $C\in\mathrm{CAR}$ and $x\in X$. This condition is satisfied, in particular, in each of the following two cases:

$$\omega_f(\beta_x C)=\omega_f(C) \quad \text{for all } x\in X, \quad C\in\mathrm{CAR}, \tag{2.8}$$

$$\omega_r(W(x))=0 \quad \text{if } \beta_x\neq \mathrm{id}. \tag{2.9}$$

One easily shows that in each of these cases ω is also positive on $\mathscr{B}$, hence it is a state on $\mathscr{F}$. We generalize this construction and give the explicit prescription for the resulting representation in terms of π_r and π_f in the following theorem.

Proposition 2.3: Suppose π_f and π_r are representations of the CAR algebra and the Weyl *algebra in Hilbert spaces $\mathscr{H}_f$ and $\mathscr{H}_r$, respectively. Define operators $\pi(A)$, acting in the space $\mathscr{H}_f\otimes\mathscr{H}_r$, in the following two cases:*

(i) π_f has a cyclic vector Ω_f, such that the expectation values in that state satisfy (2.8). *Then*

$$\begin{aligned}\pi(C)&:=\pi_f(C)\otimes\mathbf{1}_r,\\ \pi(W(x))[\pi_f(B)\Omega_f\otimes\varphi]&:=\pi_f(\beta_x B)\Omega_f\otimes\pi_r(W(x))\varphi.\end{aligned} \tag{2.10}$$

(ii) π_r has a cyclic vector Ω_r, such that the expectation values in that state satisfy (2.9). *Then*

$$\pi(W(x)) := \mathbf{1}_f \otimes \pi_r(W(x)),$$
$$\pi(C)[\psi \otimes \pi_r(W(y))\Omega_r] := \pi_f(\beta_{-y}C)\psi \otimes \pi_r(W(y))\Omega_r; \tag{2.11}$$

Then π defines a representation of the algebra $\mathscr{B}$ (and $\mathscr{F}$) in each of the two cases.

Proof: Part (i) is immediately verified if one observes that $\pi(W(x)) = U_x \otimes \pi_r(W(x))$, where U_x canonically implement β_x in the representation generated by ω_f. To prove (ii) one has to show that $\pi(C)$ in (2.11) is well defined as a linear operator. Once this is established, the verification of the algebraic properties is a simple exercise, which we omit. Let $\Sigma_{i=1}^n \psi_i \otimes \pi_r(W(x_i))\Omega_r = 0$. Without loss of generality $\beta_{x_i} = \beta_{x_k}$ $(i,k = 1,...,n)$, as otherwise the vectors are orthogonal. Then $\Sigma_i \pi_f(\beta_{-x_i}C)\psi_i \otimes \pi_r(W(x_i))\Omega_r = \pi_f(\beta_{-x_1}C) \otimes \mathbf{1}_r \Sigma_{i=1}^n \psi_i \otimes \pi_r(W(x_i))\Omega_r = 0$, which justifies the definition of $\pi(C)$. □

The following result will be needed later.

Proposition 2.4: Let τ be an automorphism of $\mathscr{F}$ satisfying (2.4) *and* (2.5), *and let π be a representation of $\mathscr{F}$ of the type* (2.10). *The following two conditions are equivalent.*

(i) The symmetry τ is implementable in the representation π by a unitary operator leaving the subspace $\Omega_f \otimes \mathscr{H}_r$ invariant:

$$\pi(\tau A) = U\pi(A)U^* \quad \forall A \in \mathscr{F}, \tag{2.12}$$

$$U(\Omega_f \otimes \mathscr{H}_r) \subset \Omega_f \otimes \mathscr{H}_r. \tag{2.13}$$

(ii) The state ω_f is invariant under τ,

$$\omega_f(\tau C) = \omega_f(C) \quad \forall C \in \mathrm{CAR}, \tag{2.14}$$

and τ is implementable in the representation π_r

$$\pi_r(\tau A) = U_r \pi_r(A) U_r^* \quad \forall A \in \mathrm{CCR}. \tag{2.15}$$

If these conditions are satisfied, there is a 1:1 correspondence between operators U and U_r given by

$$U = U_f \otimes U_r, \quad \textit{where} \quad U_f \pi_f(C)\Omega_f := \pi_f(\tau C)\Omega_f. \tag{2.16}$$

If, moreover, $G \ni g \to \tau_g \in \mathrm{Aut}\,\mathscr{F}$ is a symmetry group, then $U(g)$ is a representation of G iff $U_r(g)$ is.

Proof: (i)⇒(ii). If U satisfies (2.13), then the equation $U(\Omega_f \otimes \varphi) = \Omega_f \otimes U_r\varphi$ defines a unitary operator U_r. Then, by (2.10) and (2.12), $\Omega_f \otimes U_r\pi_r(W(x))\varphi = U\pi(W(x))(\Omega_f \otimes \varphi) = \Omega_f \otimes \pi_r(\tau W(x))U_r\varphi$, which is equivalent to (2.15). Moreover, again by (2.10) and (2.12), $U(\pi_f(C)\Omega_f \otimes \varphi) = U\pi(C)(\Omega_f \otimes \varphi) = \pi_f(\tau C)\Omega_f \otimes U_r\varphi$, hence U has the form (2.16). Finally, $\omega_f(\tau C) = (\Omega_f, \pi_f(\tau C)\Omega_f) = \omega_f(C)$.

(ii)⇒(i). Choose U as in (2.16). Then (2.12) for $A \in \mathrm{CAR}$ is obvious from (2.10). Moreover, by (2.10) and (2.5),

$$\begin{aligned} U\pi(W(x))U^*(\pi_f(C)\Omega_f \otimes \varphi) &= \pi_f(\tau\beta_x\tau^{-1}C)\Omega_f \otimes U_r\pi_r(W(x))U_r^*\varphi \\ &= \pi_f(\beta_{Tx}C)\Omega_f \otimes \pi_r(W(Tx))\varphi = \pi(W(Tx))(\pi_f(C)\Omega_f \otimes \varphi). \end{aligned}$$

If τ_g is a symmetry group, then $U_f(g)$ is a representation of G, which implies the last equivalence of the proposition. □

A similar proposition holds for the representations of the type (2.11).

III. THE ASYMPTOTIC FIELD ALGEBRA

We turn in this section to the proper task of this article, the investigation of an asymptotic algebra of fields for charged particles in interaction with an electromagnetic field, as outlined in the Introduction. For this purpose it is necessary to review the asymptotic structure of the classical field electrodynamics (the Maxwell–Dirac system) discussed in Ref. 2 (see also Ref. 13). This

paper contains, more precisely, rigorous results for both external field problems and supplies plausibility arguments for the persistence of the resulting structures in the fully interacting theory. (For recent rigorous results on the dynamics of the classical Maxwell–Dirac system obtained by the adaptation of a modified Dollard method[17] see Ref. 18; our approach is different.) Here we only briefly sketch the results without bothering about regularity assumptions or the exact sense of limits, with the purpose of merely identifying the asymptotic variables. For more details we refer the reader to Ref. 2. On the other hand, it should be clear that quantization by the correspondence principle is a heuristic procedure itself, so there is no need for excessive formalization of this point.

Due to the difference in propagation of the electromagnetic field on the one hand, and massive fields on the other, the natural asymptotic directions are also different in the two cases, lightlike in the first and timelike in the second case, respectively. Consider the electromagnetic field first. Let l_a be a null, future-pointing vector ($a,b,$ etc. are space–time indices). Then the leading asymptotic term in this direction of the electromagnetic potential in a Lorentz gauge $A_a(x)$ is given by

$$A_a(x+Rl)\sim\frac{1}{R}\,V_a(x\cdot l,l)\quad\text{for}\quad R\to\infty, \tag{3.1}$$

where x is any space–time point, $V_a(s,l)$ is a real, space–time-vector-valued function of a real variable s and a null vector l, and $x\cdot l$ denotes the scalar product with signature $(+,-,-,-)$. Different gauges yield V's differing by the transformation $V_a(s,l)\to V_a(s,l)+l_a\alpha(s,l)$, which will also be referred to as gauge transformation, but determine the same electromagnetic field asymptotic

$$F_{ab}(x+Rl)\sim\frac{1}{R}\,(l_a\dot{V}_b(x\cdot l,l)-l_b\dot{V}_a(x\cdot l,l))\quad\text{for}\quad R\to\infty, \tag{3.2}$$

where the dot over V denotes differentiation with respect to s. The functions $V_a(s,l)$ are homogeneous of degree -1,

$$V_a(\mu s,\mu l)=\mu^{-1}V_a(s,l),\quad \mu>0, \tag{3.3}$$

satisfy

$$l\cdot V(s,l)=Q, \tag{3.4}$$

where Q is the total charge of the system as measured by the integrated electric flux in spacial infinity, and tend to the limit functions $V_a(\pm\infty,l)$ for $s\to\pm\infty$. The homogeneity implies that $V_a(s,l)$ is determined by its values for l's on a manifold cutting each null direction once. In each Minkowski frame, if l's are scaled to satisfy $t\cdot l=1$ (where t is the timelike basis vector of the frame), then $\dot{V}_a(s,l)$ falls off componentwise for $|s|\to\infty$ at least as $|s|^{-(1+\epsilon)}$, for some $\epsilon>0$. Differentiations with respect to independent variables in the null vector l (i.e., differentiations in directions tangent to the lightcone) are conveniently carried out with the use of the operator $L_{ab}:=l_a\partial_b-l_b\partial_a$, $\partial_a:=\partial/\partial l^a$. With this notation the limit functions $V_a(\pm\infty,l)$ satisfy the differential condition

$$L_{[ab}V_{c]}(\pm\infty,l)=0. \tag{3.5}$$

The limit function $V_a(-\infty,l)$ has a clear physical meaning. For any spacelike vector y there is

$$\lim_{R\to\infty}R^2F_{ab}(x+Ry)=\frac{1}{2\pi}\int(l_aV_b(-\infty,l)-l_bV_a(-\infty,l))\delta'(y\cdot l)d^2l, \tag{3.6}$$

where $\delta'(\cdot)$ is the derivative of the Dirac delta-function and d^2l is the Lorentz-invariant measure on the set of null directions applicable to functions homogeneous in l of degree -2. (If l's are scaled to $t\cdot l=1$, then d^2l is the spherical angle measure on the unit sphere in the hyperplane

orthogonal to t.) Therefore, $l \wedge V(-\infty,l)$ is responsible for the long-range part of the electromagnetic field. The physical content of the property (3.5) of $V(-\infty,l)$ is that the long-range field is of purely electrical type. The physical meaning of the limit function $V(+\infty,l)$ and its property (3.5) will become clear in the sequel.

Consider now the timelike asymptotic of the Dirac field, in the sense of asymptotic behavior for $\lambda \to \infty$ of $\psi(\lambda v)$, where v lies on the future part of the unit hyperboloid. In the Dirac equation choose the electromagnetic potential in a local gauge A^{tr}, related (locally) to a Lorentz potential A by $A^{\mathrm{tr}}_a(x) = A_a(x) - \boldsymbol{\nabla}_a S(x)$, with the condition that for $x^2 \to +\infty$, $x^0 > 0$, the leading term of $S(x)$ is $\ln \sqrt{x^2}\, x \cdot A$. Then the leading asymptotic term of the Dirac field in this gauge, $\psi^{\mathrm{tr}}(x) = e^{-ieS(x)}\psi(x)$, is

$$\psi^{\mathrm{tr}}(\lambda v) \sim -i\lambda^{-3/2} e^{-i(m\lambda + \pi/4)\gamma \cdot v} f(v) \quad \text{for} \quad \lambda \to \infty, \tag{3.7}$$

where γ^a are the Dirac matrices. The function $f(v)$ is a $\mathbf{C}^4$-valued function with finite norm squared $\int \overline{f(v)}\gamma \cdot v f(v)\, d\mu(v)$, where the bar denotes the usual Dirac conjugation and $d\mu(v)$ is the invariant measure d^3v/v^0. The merit of the above asymptotics lies in its simplicity and form-independence of the potential, as long as the latter lies in the distinguished class. This does not contradict the modified Dollard asymptotic dynamics,[6,17,18] as here the asymptotic sequence of hyperboloids rather than spacelike hyperplanes is considered, and a special class of gauges is used.

The asymptotic variables $V_a(s,l)$ and $f(v)$ have well-defined transformation properties under the action of the Poincaré group. The element (x,A) of its universal covering group [x is a space–time vector and $A \in \mathrm{SL}(2,\mathbf{C})$] transforms the Lorentz (covariant) gauge fields by $A_a(y) \to \Lambda(A)_a{}^b A_b(\Lambda^{-1}(y-x))$, $\psi(y) \to S(A)\psi(\Lambda^{-1}(y-x))$, where $S(A)$ is the bispinor representation and $\Lambda(A)$ is the representation of $\mathrm{SL}(2,\mathbf{C})$ in $\mathscr{L}^{\uparrow}_{+}$ [with the notation $(\Lambda y)^a = \Lambda^a{}_b y^b$]. The field $\psi^{\mathrm{tr}}(y)$ transforms noncovariantly, but the noncovariant phase factors are shown to cancel out in the limit. The asymptotics transform then by the representations of the Poincaré group

$$[T_{x,A}V]_a(s,l) = \Lambda(A)_a{}^b V_b(s - x \cdot l, \Lambda^{-1} l), \tag{3.8}$$

$$[R_{x,A} f](v) = e^{imx \cdot v \gamma \cdot v} S(A) f(\Lambda^{-1} v). \tag{3.9}$$

Let us introduce the following structures on the space of asymptotic variables: the symplectic form

$$\{V_1, V_2\} = \frac{1}{4\pi} \int (\dot{V}_1 \cdot V_2 - \dot{V}_2 \cdot V_1)(s,l)\, ds\, d^2 l, \tag{3.10}$$

and the scalar product

$$(f_1, f_2) = \int \overline{f_1(v)}\gamma \cdot v f_2(v)\, d\mu(v) \tag{3.11}$$

(both of them are well defined). Then $T_{x,A}$ is a symplectic transformation, and $R_{x,A}$ a unitary one:

$$\{T_{x,A}V_1, T_{x,A}V_2\} = \{V_1, V_2\}, \quad (R_{x,A}f_1, R_{x,A}f_2) = (f_1, f_2). \tag{3.12}$$

The generators of these transformations defined by

$$T_{x,A} - \mathbf{1} \approx x^a r_a - \frac{1}{2}\omega^{ab} n_{ab}, \quad R_{x,A} - \mathbf{1} \approx i x^a p_a - \frac{i}{2}\omega^{ab} m_{ab},$$

for infinitesimal x^a and ω^{ab}, where $\Lambda^a{}_b \approx g^a_b + \omega^a{}_b$, are

$$(r_a V)_c(s,l) = -l_a \dot{V}_c(s,l), \quad (n_{ab}V)_c(s,l) = -L_{ab}V_c(s,l) - g_{ca}V_b(s,l) + g_{cb}V_a(s,l),$$

$$(p_a f)(v) = m v_a \gamma \cdot v f(v), \quad (m_{ab} f)(v) = \left(v_a i\delta_b - v_b i\delta_a + \frac{i}{4}[\gamma_a, \gamma_b] \right) f(v),$$

where δ_a is the derivative tangent to the hyperboloid, $\delta_a f(v) := (\nabla_a - x_a x^c \nabla_c) f(x)|_{x=v}$, and on the rhs any extension of $f(v)$ to the local neighborhood of the hyperboloid is used.

The discussion of Ref. 2 suggests that the asymptotic variables $V(s,l)$ and $f(v)$ form a causally complete set, in the sense that they determine the state of the system at any space–time point (this has not been proved in the present approach, but cf. Ref. 18). The total energy-momentum and angular momentum are shown to be sums of two terms, the first one describing the respective quantity going out in timelike directions and expressed in terms of $f(v)$ only, and the second one describing the respective quantity going out in lightlike directions and expressed in terms of $V_a(s,l)$ only. (For angular momentum this is, actually, the natural and well-defined way for extending the definition of this quantity to the infrared-singular case; the standard integral over a Cauchy surface, as resulting from Noether theorem, is ill defined.) Explicitly, they may be put into the form

$$P_a = (f, p_a f\ ) + \tfrac{1}{2}\{V, r_a V\}, \tag{3.13}$$

$$M_{ab} = (f, m_{ab} f\ ) + \tfrac{1}{2}\{V, n_{ab} V\}. \tag{3.14}$$

[The first terms in these formulas are those of Eqs. (5.15) and (5.16) in Ref. 2, while the second ones are the tensor forms of the expressions (3.13) and (3.14) in the same reference.] Note that the local gauge freedom of the Dirac field is lost in the asymptotic limit as defined in the present approach—a change of phase of $f(v)$ by a nonconstant function of v spoils the form of the angular momentum. The electromagnetic terms are separately gauge invariant, although the symplectic form (3.10) depends on gauges of $V_{\mathrm{i}a}(-\infty,l) - V_{\mathrm{i}a}(+\infty,l)$, i=1,2, (cf. below).

The conditions of relativistic quantization for the quantum variables V^{q} and f^{q} corresponding to the classical ones are[19]

$$\begin{gathered}[P_a,\ V_c^{\mathrm{q}}] = i(r_a V^{\mathrm{q}})_c, \quad [M_{ab},\ V_c^{\mathrm{q}}] = i(n_{ab} V^{\mathrm{q}})_c, \\ [P_a,\ f^{\mathrm{q}}] = -p_a f^{\mathrm{q}}, \quad [M_{ab},\ f^{\mathrm{q}}] = -m_{ab} f^{\mathrm{q}},\end{gathered} \tag{3.15}$$

where in the generators also the respective quantum variables should be substituted. Suppose, that the variables V^{q} and f^{q} commute, $[V^{\mathrm{q}}, f^{\mathrm{q}}] = 0$, and all fundamental (anti)commutators are c-numbers. This assumption remains in concord with the standard wisdom on canonical quantization in local gauges (on which the derivation of both variables is based) and fixes the quantization rules uniquely:

$$[\{V_1, V^{\mathrm{q}}\}, \{V_2, V^{\mathrm{q}}\}] = i\{V_1, V_2\}, \tag{3.16}$$

$$[(f_1, f^{\mathrm{q}}), (f_2, f^{\mathrm{q}})]_+ = 0, \quad [(f_1, f^{\mathrm{q}}), (f_2, f^{\mathrm{q}})^*]_+ = (f_1, f_2). \tag{3.17}$$

The above quantization relations must be considered as merely the first step towards our aim, as up to now we have not taken into account the constraints between the asymptotic variables. To remedy this deficiency we return to the discussion of the classical structure. The symplectic form (3.10) is invariant under the constant gauge transformation $V_a(s,l) \to V_a(s,l) + l_a \alpha(l)$. One shows that with the appropriate choice of this gauge there is

$$V_a(+\infty, l) = \int n(v) V_a^e(v,l)\ d\mu(v). \tag{3.18}$$

where $n(v) = \overline{f(v)}\, \gamma \cdot v f(v)$ is the asymptotic density of particles moving with velocity v and $V_a^e(v,l) = e v_a / v \cdot l$ is the null asymptotic (3.1) of the Lorentz potential of the Coulomb field surrounding a particle with charge e, moving with constant velocity v. Therefore, the above relation is the implementation of the Gauss constraint on the space of classical asymptotic variables. The relation (3.5) for this limit function is now seen to be satisfied identically. Let, furthermore, $A_a^{\mathrm{adv}}(x)$ be the advanced field of the sources. It turns out that the asymptotic (3.1) of this field is given by $V_a^{\mathrm{adv}}(s,l) = V_a(+\infty, l)$. Hence, the asymptotic of the free outgoing field potential, standardly defined by $A^{\mathrm{out}} = A - A^{\mathrm{adv}}$, is determined by $V_a^{\mathrm{out}}(s,l) = V_a(s,l) - V_a(+\infty, l)$. The ''out'' field is recovered from its asymptotic by the formula

$$A_b^{\text{out}}(x) = -\frac{1}{2\pi}\int \dot{V}_b^{\text{out}}(x\cdot l, l)\; d^2 l. \tag{3.19}$$

The connection of this formula with the Fourier representation

$$A_b^{\text{out}}(x) = \frac{1}{\pi}\int a_b(k)\,\delta(k^2)\,\epsilon(k^0) e^{-ix\cdot k}\; d^4 k \tag{3.20}$$

is supplied by the relation $a_b(\omega l) = -\omega^{-1}\tilde{V}_b^{\text{out}}(\omega, l)$, where the following one-dimensional Fourier transformation has been introduced:

$$\tilde{h}(\omega, l) = \frac{1}{2\pi}\int h(s,l) e^{i\omega s}\; ds. \tag{3.21}$$

The limit function $V^{\text{out}}(-\infty, l) = -2\pi\tilde{V}^{\text{out}}(0,l)$ describes the long-range (infrared-singular) part of A^{out}, the limit function at $s\to\infty$ vanishes, and the charge in formula (3.4) is zero.

The infrared characteristic $V^{\text{out}}(-\infty, l)$ has a simple representation, to become of importance below. Equations (3.3), (3.4) (with $Q=0$), and (3.5) satisfied by $V^{\text{out}}(-\infty, l)$ imply that there exists a homogeneous of degree 0 function $\Phi(l)$ such that

$$l_a V_b^{\text{out}}(-\infty, l) - l_b V_a^{\text{out}}(-\infty, l) = L_{ab}\Phi(l). \tag{3.22}$$

The function $\Phi(l)$ is determined up to an additive constant, but one of the solutions is distinguished by being determined linearly and Lorentz-covariantly by $V_a^{\text{out}}(-\infty, l)$:

$$\Phi_{V^{\text{out}}}(l) = \frac{1}{4\pi}\int \frac{l\cdot V^{\text{out}}(-\infty, l')}{l\cdot l'}\; d^2 l'. \tag{3.23}$$

[This explicit formula appears here for the first time; it may be obtained by a technique similar to that used in Appendix to prove (A8).] This new variable transforms by an addition of a constant with the gauge transformation of the potential: if $V_a^{\text{out}}(s,l) \to V_a^{\text{out}}(s,l) + \alpha(s,l) l_a$, then $\Phi_{V^{\text{out}}}(l) \to \Phi_{V^{\text{out}}}(l) + (1/4\pi)\int \alpha(-\infty, l')\; d^2 l'$. The solution (3.23) is the only one which satisfies (as shown by a simple calculation)

$$\int \frac{\Phi_{V^{\text{out}}}(l)}{(v\cdot l)^2}\; d^2 l = \int \frac{v\cdot V^{\text{out}}(-\infty, l)}{v\cdot l}\; d^2 l \tag{3.24}$$

for any velocity v.

Next, we want to determine the outgoing Dirac field which may be regarded as independent of V^{out} from the point of view of Poincaré generators. To this end put $V(s,l) = V^{\text{out}}(s,l) + V(+\infty, l)$ into (3.13) and (3.14). One finds $\frac{1}{2}\{V, r_a V\} = \frac{1}{2}\{V^{\text{out}}, r_a V^{\text{out}}\}$, but $\frac{1}{2}\{V, n_{ab} V\} = \frac{1}{2}\{V^{\text{out}}, n_{ab} V^{\text{out}}\} + \{V^{\text{out}}, n_{ab} V(+\infty, .)\}$. Substituting (3.18) for $V(+\infty, l)$ and using the identity $(n_{ab} V^e)_c(v,l) = (v_a\delta_b - v_b\delta_a) V_c^e(v,l)$ we obtain $\{V^{\text{out}}, n_{ab} V(+\infty, .)\} = \int n(v)(v_a i\delta_b - v_b i\delta_a) i\{V^e(v, .), V^{\text{out}}\} d\mu(v)$. Finally, introducing a new variable

$$g(v) = e^{i\{V^e(v,.), V^{\text{out}}\}} f(v), \tag{3.25}$$

we bring the generators to the form

$$P_a = (g, p_a g) + \tfrac{1}{2}\{V^{\text{out}}, r_a V^{\text{out}}\}, \tag{3.26}$$

$$M_{ab} = (g, m_{ab} g) + \tfrac{1}{2}\{V^{\text{out}}, n_{ab} V^{\text{out}}\}. \tag{3.27}$$

Now, define the free outgoing Dirac field by

$$\psi^{\text{out}}(x) = \left(\frac{m}{2\pi}\right)^{3/2}\int e^{-imx\cdot v\,\gamma\cdot v}\gamma\cdot v\, g(v)\; d\mu(v), \tag{3.28}$$

which is a special, concise form of the Fourier representation and which implies the asymptotic of the form (3.7) with $f(v)$ replaced by $g(v)$. Then the generators (3.26) and (3.27) turn out to be the sums of the conserved quantities for free fields $\psi^{\text{out}}(x)$ and $F^{\text{out}}_{ab}(x)$. Therefore, $\psi^{\text{out}}(x)$ should be interpreted as the field describing free particles *together with their Coulomb fields.* We have seen that the new separation of variables (3.26) and (3.27) forced the explicit appearance of a gauge-dependent quantity $\{V^e(v,.),V^{\text{out}}\}$, but only as a phase transformation. With the use of (3.24) the phase factor in (3.25) takes the form

$$e^{i\{V^e(v,.),V^{\text{out}}\}}=\exp\left(\frac{ie}{4\pi}\int\frac{\Phi_{V^{\text{out}}}(l)}{(v\cdot l)^2}\,d^2l\right), \tag{3.29}$$

and this is the only way in which a gauge-dependent quantity appears in the classical asymptotic structure. It is natural and economic, therefore, to assume that the additive constant in $e\Phi_{V^{\text{out}}}(l)$ is a phase variable. Consequently, we put into one class gauges V^{out}_1 and V^{out}_2 such that $l\wedge V^{\text{out}}_1(s,l)=l\wedge V^{\text{out}}_2(s,l)$ and $\Phi_{V^{\text{out}}_2}(l)-\Phi_{V^{\text{out}}_1}(l)=n2\pi/e,\ n\in\mathbf{Z}$.

With the above knowledge of the classical structure we can now return to the problem of taking into account the Gauss constraint on the quantum level. The form (3.18) of this constraint is not suited for the translation to an abstract algebraic level. However, the physical interpretation prompts an indirect solution. Instead of either the pair $(V,f\,)$ or (V^{out},g) it is natural to work with the pair of variables having the direct physical meaning: the asymptotic total electromagnetic field V and the asymptotic field of charged particles, with their Coulomb fields included, g. The commutation relations (3.16) and (3.17) will be now reformulated in terms of these variables with the use of the relation

$$g(v)=e^{i\{V^e(v,.),V\}}f(v), \tag{3.30}$$

which is equivalent to (3.25) by $\{V^e(v,.),V(+\infty,.)\}=0$. Two circumstances have to be taken into account. First, on the classical level the Gauss constraint has been completely solved, so for the electromagnetic test field asymptotic the free field part only should be taken. Second, the quantization has to be consistent with our identifying the gauges differing by $n2\pi$ in $e\Phi(l)$. This problem is solved, as is easily seen from (3.16) and (3.17), by using the electromagnetic variable in the form $e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}$ only. Then, the relations (3.16) take the form

$$e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}e^{-i\{V^{\text{out}}_2,V^{\text{q}}\}}=e^{-i(1/2)\{V^{\text{out}}_1,V^{\text{out}}_2\}}e^{-i\{V^{\text{out}}_1+V^{\text{out}}_2,V^{\text{q}}\}}, \tag{3.31}$$

the relations (3.17) remain true for g^{q}, but now the two variables do not commute:

$$e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}g^{\text{q}}(v)=e^{+i\{V^{\text{out}}_1,V^e(v,.)\}}g^{\text{q}}e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}. \tag{3.32}$$

The last relation has an obvious physical interpretation: the element $g^{\text{q}}(v)$, besides its fermionic role, annihilates the Coulomb field with the asymptotic $V^e(v,l)$. This is, clearly, the implementation of Gauss' law on the quantum level. Note, also, that by (3.29) the relation is indeed consistent with our identification of gauge classes. Observe, furthermore, that the element $e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}$ creates the field with the asymptotic $V^{\text{out}}_1(s,l)$. The use of only free fields as test fields reflects the fact that the Coulomb field is fastened to a particle, which is a neat confirmation of the consistency of our scheme. Nevertheless, the element $e^{-i\{V^{\text{out}}_1,V^{\text{q}}\}}$ is a functional of the *total* field V^{q}, as assumed in the construction. This is seen from (3.32), and also from

$$\{V^{\text{out}}_1,V\}=\{V^{\text{out}}_1,V^{\text{out}}\}-\frac{1}{4\pi}\int V^{\text{out}}_1(-\infty,l)\cdot V(+\infty,l)d^2l. \tag{3.33}$$

It is clear from both arguments that in order to ''catch'' the whole field, it is absolutely necessary that all free test fields are admitted, also those infrared-singular, for which $V^{\text{out}}(-\infty,l)\neq 0$. This fact is to become of crucial importance for the interpretation of further results.

The quantum structure thus obtained will be now given an unobjectionable algebraic form, formulated in terms of elements heuristically identified by

$$W(V)=e^{-i\{V,V^{\mathrm{q}}\}},\quad B(g)=(g,g^{\mathrm{q}}),$$

where from now on all the test fields $V(s,l)$ are free fields, so we omit the superscript ''out.'' We have to specify the scope of the test functions. Let $\mathscr{K}$ be the Hilbert space of (equivalence classes of) $\mathbf{C}^4$-valued functions $g(v)$ on the hyperboloid $v^2=1$, $v^0>0$, with the scalar product (3.11). Let $\mathscr{V}$ be the linear space of homogeneous of degree -1 [Eq. (3.3)] functions $V_a(s,l)$, infinitely differentiable in both variables outside $l=0$ (differentiations with respect to l in the sense of the action of the operator L_{ab}) and satisfying the conditions

$$l\cdot V(s,l)=0, \tag{3.34}$$

$$|L_{b_1c_1}\cdots L_{b_kc_k}\dot{V}_a(s,l)|<\frac{\mathrm{const}(k)}{(t\cdot l)^2(1+|s|/t\cdot l)^{1+\epsilon}},\quad k=0,1,..., \tag{3.35}$$

$$V_a(+\infty,l)=0, \tag{3.36}$$

$$L_{[ab}V_{c]}(-\infty,l)=0, \tag{3.37}$$

where the second condition holds for some (V- and k-dependent) $\epsilon>0$ and for an arbitrarily chosen unit timelike, future-pointing vector t; the bounds are then true for any other such vector [with some other constants const(k)]. These bounds guarantee the existence of the limit functions as infinitely differentiable, homogeneous functions of degree -1. Let L be the Abelian additive group of elements (V) defined as pairs

$$(V)=(l\wedge V(s,l),\Phi_V(l)\bmod 2\pi/e), \tag{3.38}$$

where $V\in\mathscr{V}$ and Φ_V is defined by (3.23). In other words, L is the quotient of the additive group $\mathscr{V}$ through the equivalence relation $\sim$, $L=\mathscr{V}/\sim$, where

$$V_1\sim V_2\quad\text{iff}\quad V_2(s,l)-V_1(s,l)=l\alpha(s,l)\quad\text{and}\quad\frac{1}{4\pi}\int\alpha(-\infty,l)d^2l=n\,\frac{2\pi}{e}. \tag{3.39}$$

The group L inherits from $\mathscr{V}$ the symplectic form (3.10). Denote, also, for later use, $\mathscr{V}_0:=\{V\in\mathscr{V}\,|\,l\wedge V(-\infty,l)=0\}$ and $L_0:=\mathscr{V}_0/\sim$.

The symplectic group L and the Hilbert space $\mathscr{K}$ supply the test fields for the elements $W(V)$, $(V)\in L$ [the parenthesis in the symbol (V) appearing as an argument of W or of the symplectic form will be omitted], and $B(f)$, $f\in\mathscr{K}$, which generate a particular *-algebra $\mathscr{B}$ of Sec. II according to the relations

$$W(V)^*=W(-V),\quad W(0)=E,\quad W(V_1)W(V_2)=e^{-(i/2)\{V_1,V_2\}}W(V_1+V_2), \tag{3.40}$$

$$[B(g_1),\,B(g_2)]_+=0,\quad [B(g_1),\,B(g_2)^*]_+=(g_1,g_2)E, \tag{3.41}$$

$$W(V)B(g)=\beta_{\Phi_V}(B(g))W(V), \tag{3.42}$$

where

$$\beta_\Phi(B(g))=B(S_\Phi g),\quad (S_\Phi g)(v)=\exp\left(i\,\frac{e}{4\pi}\int\frac{\Phi(l)}{(v\cdot l)^2}\,d^2l\right)g(v). \tag{3.43}$$

Definition 3.1: The asymptotic field algebra is the C-algebra* $(\mathscr{F},\|.\|)$ *obtained from the above *-algebra* $\mathscr{B}$ *according to Definition 2.1.*

Let us consider the role of elements $W((0,c \bmod 2\pi/e))$, which form an Abelian one-parameter group $W((0,c_1 \bmod 2\pi/e))W((0,c_2 \bmod 2\pi/e))=W((0,c_1+c_2 \bmod 2\pi/e))$. The relations (3.4) and (3.33) suggest that in any representation in which $\pi(W((0,c \bmod 2\pi/e)))$ are strongly continuous and written as e^{icQ_π}, the operator Q_π has the interpretation of the charge operator. This interpretation is confirmed by the action of the automorphism γ_c of $\mathscr{F}$,

$$A \to \gamma_c(A) := W((0,c \bmod 2\pi/e))AW((0,c \bmod 2\pi/e))^* \tag{3.44}$$

on the basic elements:

$$\gamma_c(W(V)) = W(V), \quad \gamma_c(B(g)) = e^{-iec}B(g).$$

Now, as $2\pi/e = 0 (\mathrm{mod}\, 2\pi/e)$, there is $e^{i2\pi Q_\pi/e} = \mathbf{1}$, which implies that the spectrum of charge is contained in the set $\{ne | n \in \mathbf{Z}\}$. The variable $e\Phi(l)$ is connected with the free electromagnetic field, so bringing into play other carriers of charge should respect the phase character of the additive constant in this function. This means that the assumption in the following corollary is well founded.

Corollary 3.1: If a C-algebra of asymptotic fields contains the subalgebra generated by elements* $W((0,c \bmod 2\pi/e))$, *then the charge is quantized in units of* e.

In the following definition particular elements of $\mathscr{F}$ are distinguished as observables in the obvious way.

Definition 3.2: The algebra of observables $\mathscr{A}$ *is the C*-subalgebra of* $\mathscr{F}$ *of elements invariant under the gauge transformation (3.44).*

One has to stress at this point that all the Weyl elements are therefore (functions of) observables. Denying the elements with infrared-singular test functions $[V(-\infty, l) \neq 0]$ the status of observables one would deprive the total electromagnetic field of its Coulomb part, as discussed earlier. We shall return to this important point when discussing representations of our algebra. Also, all elements $B(f)^*B(g)$ are in $\mathscr{A}$.

The restricted Poincaré group (or rather its covering group) is represented in the group of automorphisms of the field algebra $\mathscr{F}$. One easily shows that the operators (3.8) and (3.9) [the variable $g(v)$ undergoes the same transformations as $f(v)$] satisfy the consistency condition (2.7), namely,

$$R_{x,A}S_\Phi = S_{T_{x,A}\Phi}R_{x,A},$$

where the transformation $[T_{x,A}\Phi](l) = \Phi(\Lambda^{-1}l)$ is implied by (3.8) and (3.23). Therefore, the action of the Poincaré group on $\mathscr{F}$ may be consistently defined by

$$\alpha_{x,A}(W(V)) = W(T_{x,A}V), \quad \alpha_{x,A}(B(g)) = B(R_{x,A}g).$$

We end this section with the demonstration that for free test fields the symplectic structure discussed above is an extension of the structure used in more traditional algebraic formulations. To see this we find the connection with the work of Roepstorff.[20] This author uses the electromagnetic test fields of the form $F_{ab}(x) = 4\pi \int D(x-y)(\nabla_a \varphi_b(y) - \nabla_b \varphi_a(y))\, d^4y$, where $D(x)$ is the Pauli–Jordan function and $\varphi_a(x) = \nabla^b \varphi_{ab}(x)$ for some antisymmetric test function of compact support $\varphi_{ab}(x)$. With the use of representation $D(x) = (1/8\pi^2)\int \delta'(x \cdot l)\, d^2l$ a Lorentz gauge potential for this field takes the form (3.19), with

$$V_a(s,l) = \int \delta(s - y \cdot l)\varphi_a(y)\, d^4y. \tag{3.45}$$

Substituting two functions of this form in (3.10) one has

$$\{V_1, V_2\} = 4\pi \int \varphi_1^a(x) D(x-y) \varphi_{2a}(y)\, d^4x\, d^4y. \tag{3.46}$$

The rhs is the symplectic form used by Roepstorff (up to multiplicative constants due to electromagnetic conventions). However, the space of test fields is smaller in this formulation. It is obvious from (3.45) that $V_a(-\infty, l) = 0$, so all these fields are infrared-regular (the spacelike asymptotic of F_{ab} has no $1/r^2$ term). In fact, even a stronger regularity property holds. There is $V_a(s, l) = \dot{J}_a(s, l)$, where $J_a(s, l)$ is a smooth function vanishing outside a compact region (for a fixed scaling $t \cdot l = 1$), given by $J_a(s, l) = \int \delta(s - y \cdot l) l^b \varphi_{ab}(y)\, d^4y$. It will prove convenient to reformulate the above formulas in the Fourier-transformed version. It is shown in Sec. V below that

$$\{V_1,V_2\}=i\mathrm{P}\int \overline{\tilde{V}_1}\cdot\tilde{V}_2(\omega,l)\,\frac{d\omega}{\omega}\,d^2l, \tag{3.47}$$

where P denotes the principal value. If $\tilde{V}_a(\omega,l)$ vanishes for $\omega\to 0$ sufficiently fast (e.g., as $|\omega|^{\epsilon}$), then the principal value sign may be omitted. This is true, in particular, for V given by (3.45). In this case $\tilde{V}_a(\omega,l)=\hat{\varphi}_a(\omega l)$, where $\hat{\varphi}(p)=(1/2\pi)\int\varphi(x)e^{ip\cdot x}\,d^4x$, and then $\tilde{V}_a(\omega,l)=-i\omega\hat{\varphi}_a(\omega l)$. On the other hand, the last equation shows that our general field satisfies the condition of Roepstorff's space L_1.

IV. EXISTENCE OF CHARGE AND ENERGY-MOMENTUM, AND THE REGULARITY OF REPRESENTATIONS

In the present section we investigate the consequences of putting some physical restrictions on representations. The following definitions, the second of which is standard,[5] will simplify the formulation of propositions.

Definition 4.1: (i) A representation π of $\mathscr{T}$ will be called a charge-representation of $\mathscr{T}$ iff the one-parameter group $\mathbf{R}\ni c\to\pi(W((0,c \bmod 2\pi/e)))$ is strongly continuous.

(ii) A representation π of the algebra $\mathscr{T}$ acting in the Hilbert space $\mathscr{H}$ is said to satisfy Borchers' criterion iff there exists a unitary, strongly continuous representation of the translation group $U(x)$ acting in $\mathscr{H}$, with the spectrum contained in the closed forward lightcone, Spec $U(x)\subset\overline{V_+}$, *and implementing translations of all $A\in\mathscr{T}$*:

$$\pi(\alpha_x A)=U(x)\,\pi(A)\,U(-x). \tag{4.1}$$

(iii) A representation π of the algebra $\mathscr{T}$ will be said to satisfy Borchers' criterion with respect to $\mathscr{A}$ (wrt observables) iff the restriction of π to the subalgebra $\mathscr{A}$ satisfies Borchers' criterion.

The defining properties of (i) and (iii) are necessary requirements for a physically admissible representation. We analyze their implications.

Proposition 4.1: If π is a charge-representation of $\mathscr{T}$, then it has one of the following properties (or is a direct sum of these three types):

(i) The charge takes on all the values ke, $k\in\mathbf{Z}$, and each charge eigenspace is cyclic.

(ii) [resp. (iii)] The subspace of vectors satisfying $\pi(B(g))\psi=0\ \ \forall g\in\mathscr{K}$ [resp. $\pi(B(g)^*)\psi=0\ \ \forall g\in\mathscr{K}$] is cyclic.

Proof: Given $\pi(\mathscr{T})$ on $\mathscr{H}$ let $\mathscr{H}_k\subset\mathscr{H}$ be the subspace of all charge eigenvectors to the eigenvalue ke for a given $k\in\mathbf{Z}$, and let $\mathscr{H}'_k=[\pi(\mathscr{T})\mathscr{H}_k]$ [the closed linear subspace spanned by vectors in $\pi(\mathscr{T})\mathscr{H}_k$]. Then $\mathscr{H}=\mathscr{H}'_k\oplus\mathscr{H}'^{\perp}_k$, where both subspaces are invariant. Moreover, if we split $\mathscr{H}'^{\perp}_k=\mathscr{H}'_{k+}\oplus\mathscr{H}'_{k-}$, where the spectrum of charge goes from $(k+1)e$ upwards on $\mathscr{H}'_{k+}$ and from $(k-1)e$ downwards on $\mathscr{H}'_{k-}$, then both subspaces are separately invariant. This occurs because the generating elements $\pi((W(V)))$, $\pi(B(g))$ and $\pi(B(g)^*)$ carry charge 0 or $\pm e$, so they cannot match the gap between $\mathscr{H}'_{k+}$ and $\mathscr{H}'_{k-}$.

The representations $\pi(\mathscr{T})|_{\mathscr{H}'_{k+}}$ and $\pi(\mathscr{T})|_{\mathscr{H}'_{k-}}$ have the properties (ii) and (iii), respectively. To see this, let $\mathscr{H}''_{k+}$ be the subspace of $\mathscr{H}'_{k+}$ of vectors satisfying the equation in (ii). Let ψ be an element of the invariant subspace $\mathscr{H}'_{k+}\cap[\pi(\mathscr{T})\mathscr{H}''_{k+}]^{\perp}$, with a bounded spectral content of charge [these vectors form a dense subspace as $e^{icQ_\pi}\in\pi(\mathscr{T})$]. If $\psi\neq 0$, then there exists such $f\in\mathscr{K}$ that $\pi(B(f))\psi\neq 0$. Continuing in this way one can lower the charge spectral content of the vector unlimitedly. This contradicts the charge content of $\mathscr{H}'_{k+}$. A similar proof holds for $\pi(\mathscr{T})|_{\mathscr{H}'_{k-}}$.

The rest of the proof is simple inductive reasoning. Let the set of integers $\mathbf{Z}$ be organized into a sequence $\{k_n\}$. Suppose that for $\pi(\mathscr{T})$ on $\mathscr{H}(n)$ the charge eigensubspaces $\mathscr{H}_{k_1},\ldots,\mathscr{H}_{k_n}$ are cyclic. Take the next charge eigenspace $\mathscr{H}_{k_{n+1}}$ and decompose $\mathscr{H}(n)$ according to the above prescription: $\mathscr{H}(n)=\mathscr{H}'_{k_{n+1}}\oplus\mathscr{H}'_{k_{n+1}+}\oplus\mathscr{H}'_{k_{n+1}-}$. The representations $\pi(\mathscr{T})|\mathscr{H}'_{k_{n+1}\pm}$ satisfy the properties (ii) and (iii), respectively. By construction $\mathscr{H}_{k_{n+1}}$ is cyclic for $\pi(\mathscr{T})$ on $\mathscr{H}(n+1):=\mathscr{H}'_{k_{n+1}}$. As each of the subspaces $\mathscr{H}'_{k_{n+1}}$, $\mathscr{H}'_{k_{n+1}+}$, and $\mathscr{H}'_{k_{n+1}-}$ is invariant, in particular with respect to charge operator, so $\mathscr{H}_{k_i}=(\mathscr{H}_{k_i}\cap\mathscr{H}'_{k_{n+1}})\oplus(\mathscr{H}_{k_i}\cap\mathscr{H}'_{k_{n+1}+})$

$\oplus(\mathscr{H}_{k_i}\cap\mathscr{H}'_{k_{n+1}-})$, and the spaces $\mathscr{H}_{k_i}\cap\mathscr{H}(n+1)$ are cyclic in $\mathscr{H}(n+1)$ for $i=1,...,n$. Continuing in this way we obtain a direct sum of representations satisfying the properties (ii) or (iii) and a representation on $\mathscr{H}=\cap_{n=1}^{\infty}\mathscr{H}(n)$, which satisfies the property (i) by construction. □

If, in addition to the charge, physical energy-momentum observables are assumed to exist, then one has the following result.

Proposition 4.2: The charge spectrum consists of all values ke, $k\in\mathbf{Z}$, and eigenspace to each charge value is cyclic for a charge-representation of $\mathscr{F}$ satisfying Borchers' criterion with respect to observables.

The proof of the proposition will be based on the following observation.

Lemma 4.3: If the representation π of $\mathscr{F}$ has the property (ii) of Proposition 4.1, then it is unitarily equivalent to a representation of type (2.10), where π_f is the Fock representation based on the cyclic vector Ω_f satisfying $\pi_f(B(g))\Omega_f=0\ \forall g\in\mathscr{K}$.

An analogous result holds for representations satisfying the property (iii) of Proposition 4.1. [In that case π_f is the Fock representation based on the cyclic vector satisfying $\pi_f(B(g)^*)\Omega_f=0\ \forall g\in\mathscr{K}$. We use the term Fock representation in the wider sense, referring to any of the representations differing by a Bogoliubov transformation from the one appearing in the lemma.]

Proof: Let the representation π act in $\mathscr{H}$ and denote the respective cyclic subspace by $\mathscr{H}_r$. If $C\in{}^*\mathrm{CAR}$, then it may be represented as $C=\omega_F(C)E+C'$, where ω_F is the Fock state and C' is a sum of elements having $B(f)$ on the right and/or $B(g)^*$ on the left. Therefore for $\varphi,\ \psi\in\mathscr{H}_r$ there is $(\varphi,\pi(C)\psi)=\omega_F(C)(\varphi,\psi)$. $\mathscr{H}_r$ is invariant under $\pi(W(V))$, hence the vectors $\Sigma_{k=1}^{N}\pi(C_k)\varphi_k$, $\varphi_k\in\mathscr{H}_r$, are dense in $\mathscr{H}$. Let π_r be the restriction of π to the Weyl algebra and to the space $\mathscr{H}_r$. We map $\mathscr{H}$ onto $\mathscr{H}_F\otimes\mathscr{H}_r$ ($\mathscr{H}_F$ is the Hilbert space of the Fock representation π_F) by the rule $\Sigma_k\pi(C_k)\varphi_k\to\Sigma_k\pi_F(C_k)\Omega_F\otimes\varphi_k$. It is now easy to show this is a unitary map providing the claimed equivalence of representations. □

Proof of Proposition 4.2: In view of the result of Proposition 4.1 one has to show that a representation of the type described in the lemma cannot be a subrepresentation of a representation satisfying the assumptions of Proposition 4.2. By a general theorem by Borchers[21] (see also Ref. 5) the representation $U(x)$ may be chosen to lie in $\pi(\mathscr{A})''$, hence it suffices to show that for the representation of the lemma itself there is no $U(x)$ in $\pi(\mathscr{A})''$ implementing translations of observables and satisfying the spectral condition. Suppose the converse is true. Then for any $C\in\mathrm{CAR}\cap\mathscr{A}$ there is $U(x)\pi(C)U(-x)=\pi(\alpha_x C)=\pi_F(\alpha_x C)\otimes\mathbf{1}_r=[U_F(x)\otimes\mathbf{1}_r][\pi_F(C)\otimes\mathbf{1}_r]\times[U_F(-x)\otimes\mathbf{1}_r]$, where $U_F(x)$ is the unitary representation of translations in the Fock representation. This means that $U(x)[U_F(-x)\otimes\mathbf{1}_r]\equiv R(x)$ commutes with $\pi_F(C)\otimes\mathbf{1}_r$ for all $C\in\mathrm{CAR}\cap\mathscr{A}$. The representation $\pi_F(\mathrm{CAR}\cap\mathscr{A})$ acts irreducibly on each of the subspaces $\mathscr{H}_n\subset\mathscr{H}_F$ spanned by all vectors $\pi_F(B(g_1)^*\cdots B(g_n)^*)\Omega_F$, and all spaces $\mathscr{H}_n\otimes\mathscr{H}_r$ are invariant with respect to $R(x)$ [as they are invariant with respect to $\pi(\mathscr{A})$]. Hence, $U(x)|_{\mathscr{H}_n\otimes\mathscr{H}_r}=U_F(x)|_{\mathscr{H}_n}\otimes U_{r,n}(x)$, where $U_{r,n}$ is a strongly continuous representation of translations in $\mathscr{H}_r$. However, the Fock representation π_F appearing in this construction is not ''the right'' Fock representation of the free Dirac field [$B(g)$ contains both positive and negative frequencies], and the energy spectrum of $U_F(x)|_{\mathscr{H}_n}$ is not bounded from below for $n\geqslant 1$, which contradicts the assumption. □

The existence of a charge operator is a necessary, but by far not a sufficient condition for the operators $\pi(W(V))$ to have a clear physical interpretation. We recall that, as explained in the previous section, in our algebra all these Weyl operators should be understood as observables, more precisely, as exponentials of (unbounded) observable electromagnetic field operators. The test function set L is an Abelian group rather than a vector space, so a direct multiplication of V by a parameter is not possible. However, it is sufficient to find a map $\mathbf{R}\ni\lambda\to(V^\lambda)\in L$, such that $(V^0)=(0)$, $(V^1)=(V)$, $(V^\lambda)+(V^\nu)=(V^{\lambda+\nu})$, and $-(V^\lambda)=(V^{-\lambda})$. Then $W(V^\lambda)$ is a one-parameter group and the above condition on the representation may be formulated as the strong continuity of $\pi(W(V^\lambda))$ in λ. Let $(V)=(l\wedge V,\Phi \bmod 2\pi/e)$. Then it is easily shown that for each choice of the representant Φ from the class $\Phi \bmod 2\pi/e$ the map

$$\mathbf{R}\ni\lambda\to(V^\lambda)_\Phi:=(l\wedge\lambda V,\lambda\Phi \bmod 2\pi/e) \tag{4.2}$$

satisfies the listed requirements.

Definition 4.2: Representation π of the algebra $\mathscr{T}$ (or of a subalgebra of $\mathscr{T}$) will be called regular iff all one-parameter groups $\mathbf{R}\ni\lambda\to\pi(W((V^\lambda)_\Phi))$ are strongly continuous.

Remarks: (i) If π is regular, then it is a charge representation. This follows by the special choice $(V)=(0,c \bmod 2\pi/e)$.

(ii) The generators of the groups $\pi(W((V^\lambda)_{\Phi'}))$ and $\pi(W((V^\lambda)_\Phi))$, where $\Phi' \in \Phi \bmod 2\pi/e$, differ by a multiple of the charge operator. This follows from $W((V^\lambda)_{\Phi'}) = W((V^\lambda)_\Phi)W((0,\lambda 2k\pi/e \bmod 2\pi/e))$ for $\Phi'=\Phi+2k\pi/e$.

We are now prepared to partly characterize the representations satisfying the condition of Definition 4.1 (ii). For the positive energy Fock representation π_F of the algebra CAR let us denote $\pi_F(B(g))=\int \overline{g(v)}\gamma\cdot v b(v)\, d\mu(v)$ and let :...: denote the standard normal ordering.

Theorem 4.4: *(i) A representation π of the algebra $\mathscr{T}$ satisfies Borchers' criterion if, and only if, it is unitarily equivalent to a representation of type (2.10), where $\pi_f=\pi_F$ is the positive energy Fock representation of* CAR *and π_r is a representation of the Weyl algebra satisfying Borchers' criterion.*

For representations of this type, the Gauss constraint in the form (3.18) is recovered in the von Neumann algebra $\pi(\mathscr{A})''$:

$$\pi(W(V))=\exp\left(-i\int :\overline{b(v)}\gamma\cdot v b(v):\{V,V^e(v,\cdot)\}d\mu(v)\right)\otimes\pi_r(W(V)). \tag{4.3}$$

(ii) π is irreducible iff π_r is irreducible.

(iii) π is regular iff π_r is regular.

Remark: For $\pi(\mathscr{T})$ in the form given by (i) the operator $\mathbf{1}\otimes\pi_r(W(V))$ has the interpretation of the free field exponent, and one sees once more the essentiality of the regularity assumption.

Proof: If the unitary equivalence is proved, then (4.3) follows by simple calculation. This formula then implies (ii) by irreducibility of the Fock representation [and the identity[22] $(\mathscr{L}(\mathscr{H}_F)\otimes\pi_r(\mathrm{CCR}))'=\mathbf{C}\mathbf{1}_F\otimes\pi_r(\mathrm{CCR})'$] and (iii) by $\{(V^\lambda)_\Phi,V^e(v,\cdot)\}=-\lambda(e/4\pi)\times\int(\Phi(l)/(v\cdot l)^2)\, d^2l$. The equivalence of representations is proved by adapting to the present case the idea of Ref. 23. Let $g\in\mathscr{K}$. The positive and negative frequency parts of g are easily extracted from g by $g_\pm=P_\pm g$, where $P_\pm$ are projection operators in $\mathscr{K}$ defined by $(P_\pm g)(v)=P_\pm(v)g(v)$, $P_\pm(v)=\frac{1}{2}(1\pm\gamma\cdot v)$. Let π satisfy Borchers' criterion. From translational covariance one shows in standard way that $\pi(B(g_+))$ and $\pi(B(g_-))^*$ [resp. $\pi(B(g_+))^*$ and $\pi(B(g_-))$] lower (resp. raise) the energy content of a vector by at least m (the mass of the fermion). Let $\mathscr{H}_r$ be the subspace of the representation space formed by all vectors φ satisfying $\pi(B(g_+))\varphi=\pi(B(g_-))^*\varphi=0\,\forall g\in\mathscr{K}$. The subspace $\mathscr{H}_r$ is invariant under U and the subspaces $[\pi(\mathscr{T})\mathscr{H}_r]$ and $[\pi(\mathscr{T})\mathscr{H}_r]^\perp$ are invariant both under π and U. Let $\psi\neq 0$ lie in the dense subspace of finite energy vectors in $[\pi(\mathscr{T})\mathscr{H}_r]^\perp$. Then for some $g\in\mathscr{K}$ there is $\pi(B(g_+))\psi\neq 0$ or $\pi(B(g_-))^*\psi\neq 0$. By recurrence, this gives the way for reaching negative energies, which contradicts the assumption. Hence $\mathscr{H}_r$ is cyclic for $\pi(\mathscr{T})$. The unitary equivalence to a representation of type (2.10) follows now as in the proof of Lemma 4.3. However, the Fock representation appearing in the construction is in the present case ''the right,'' positive energy Fock representation. The use of Proposition 2.4 finishes the proof. From this proposition also the ''if'' statement of (i) follows easily. □

Our next objectives are the characterization of vacuum states on our algebra, and the construction of a class of physically meaningful regular representations. The first problem is solved by the adaptation to the present situation of standard analytical methods. We shall have to discuss some properties of the symplectic space of test functions. For $V\in\mathscr{V}$ let us denote by $[\dot V]$ the respective *field* test function, i.e., the equivalence class of $\dot V$ with respect to the equivalence

$$\dot V'\approx\dot V \ \text{ iff } \ l\wedge\dot V'=l\wedge\dot V. \tag{4.4}$$

The vector space of these classes will be denoted by L^f, and its subspace of classes $[\dot V]$ with $l\wedge V(-\infty,l)=0$ by L^f_0. The space L^f inherits the symplectic form from $\mathscr{V}$, $\{[\dot V_1],[\dot V_2]\}=\{V_1,V_2\}$. Now the analytical properties of this space needed for the characterization of vacuum

states will also play a role for its extension $\hat{L}^f$ to be used in the construction of regular representations. Therefore, in the next section we introduce this auxiliary symplectic space and formulate the necessary properties in this wider setting.

V. EXTENSION AND ANALYTICAL PROPERTIES OF SYMPLECTIC STRUCTURE

The first step towards extension of the symplectic space of test fields L^f will be generalizing the falloff condition (3.35) for $k=0$. We do it first for scalar functions. Consider the real Hilbert space $L^2_{\epsilon,t}$ of (equivalence classes of) real, measurable functions $f(s,l)$, homogeneous of degree -2, with finite norm

$$\|f\|^2_{\epsilon,t}=\int f^2(\tau t\cdot l,l)(|\tau|+1)^{1+\epsilon}(t\cdot l)^2\, d\tau\, d^2l, \tag{5.1}$$

where $\epsilon>0$ and t is a unit future-pointing vector. If $\tilde{t}$ is another future-pointing vector and $c_{\tilde{t}\cdot t}\equiv\tilde{t}\cdot t+\sqrt{(\tilde{t}\cdot t)^2-1}$, then for every null vector l there is $c^{-1}_{\tilde{t}\cdot t}\leqslant\tilde{t}\cdot l/t\cdot l\leqslant c_{\tilde{t}\cdot t}$. Using these bounds one shows that $c^{-2-\epsilon}_{\tilde{t}\cdot t}\|f\|^2_{\epsilon,t}\leqslant\|f\|^2_{\epsilon,\tilde{t}}\leqslant c^{2+\epsilon}_{\tilde{t}\cdot t}\|f\|^2_{\epsilon,t}$. Therefore $L^2_{\epsilon,t}$ does not depend on t when considered as a linear topological space. Components of smooth functions satisfying (3.35) obviously lie in this space.

If $f\in L^2_{\epsilon,t}$ then for $\delta<\epsilon$

$$\begin{aligned}&\int |f(\tau t\cdot l,l)|(|\tau|+1)^{\delta/2}|f(st\cdot l,l)|(|s|+1)^{\delta/2}(t\cdot l)^2 d\tau\, ds\, d^2l\\ &\leqslant\left(\int\left[\int|f(\tau t\cdot l,l)|^2(t\cdot l)^2d^2l\right]^{1/2}(|\tau|+1)^{\delta/2}\,d\tau\right)^2\leqslant\frac{2}{\epsilon-\delta}\|f\|^2_{\epsilon,t}.\end{aligned}$$

Consequently, the integral $\int|f(\tau t\cdot l,l)|(|\tau|+1)^{\delta/2}\,d\tau\, t\cdot l$ is finite d^2l-almost everywhere and as a function of l (homogeneous of degree -1) is square integrable with respect to d^2l. In particular, the Fourier transform of $f(s,l)$ with respect to s is defined by the integral (3.21), satisfies

$$\int\left|\tilde{f}\left(\frac{\omega}{t\cdot l},l\right)\right|^2 d^2l\leqslant\frac{1}{2\pi^2\epsilon}\|f\|^2_{\epsilon,t}, \tag{5.2}$$

and

$$\int\left|\tilde{f}\left(\frac{\omega}{t\cdot l},l\right)\right|^2 d\omega=\frac{1}{2\pi}\int f^2(\tau t\cdot l,l)d\tau(t\cdot l)^2. \tag{5.3}$$

The function $\tilde{f}(\omega/t\cdot l,l)$ depends continuously on ω, both pointwise in l and as an element of the Hilbert space $L^2(d^2l)$. In fact, even stronger conditions are satisfied. With the use of the bound $|e^{ix}-1|<2|x|^\alpha$ for $0\leqslant\alpha<1$, one finds for any $\delta\in\langle 0,\min\{2,\epsilon\}\rangle$

$$\left|\tilde{f}\left(\frac{\omega'}{t\cdot l},l\right)-\tilde{f}\left(\frac{\omega}{t\cdot l},l\right)\right|\leqslant\frac{1}{\pi}\,|\omega'-\omega|^{\delta/2}\int|\tau|^{\delta/2}|f(\tau t\cdot l,l)|d\tau\,t\cdot l, \tag{5.4}$$

and then

$$\int\left|\tilde{f}\left(\frac{\omega'}{t\cdot l},l\right)-\tilde{f}\left(\frac{\omega}{t\cdot l},l\right)\right|^2 d^2l\leqslant\frac{2}{(\epsilon-\delta)\pi^2}\,\|f\|^2_{\epsilon,t}|\omega'-\omega|^{\delta}. \tag{5.5}$$

This implies, with the use of (5.2),

$$\left|\int\left|\tilde{f}\left(\frac{\omega'}{t\cdot l},l\right)\right|^2 d^2l-\int\left|\tilde{f}\left(\frac{\omega}{t\cdot l},l\right)\right|^2 d^2l\right|\leqslant\frac{2}{\pi^2\sqrt{\epsilon(\epsilon-\delta)}}\,\|f\|^2_{\epsilon,t}|\omega'-\omega|^{\delta/2}. \tag{5.6}$$

Next, we derive some integral identities. Let $\tilde{K}_{\alpha,\beta}(\omega)=\omega^{-1}\chi_{\alpha,\beta}(\omega)$, where $\chi_{\alpha,\beta}$ is the characteristic function of the set $\langle-\beta,-\alpha\rangle\cup\langle\alpha,\beta\rangle$, and let $f,g\in L^2_{\epsilon,t}$. For almost all l the function $(1/(2\pi)^2)f(\tau,l)\tilde{K}_{\alpha,\beta}(\omega)e^{-i\omega(\tau-s)}g(s,l)$ is absolutely integrable with respect to $d\tau\,ds\,d\omega$, so the iterated integrals are equal, i.e.,

$$\frac{1}{(2\pi)^2}\int f(\tau,l)K_{\alpha,\beta}(\tau-s)g(s,l)d\tau\,ds=\int_{|\omega|\in\langle\alpha,\beta\rangle}\overline{\tilde{f}(\omega,l)}\tilde{g}(\omega,l)\,\frac{d\omega}{\omega}. \tag{5.7}$$

$K_{\alpha,\beta}(s)$ is uniformly bounded and $\lim_{\alpha\to0}\lim_{\beta\to\infty}K_{\alpha,\beta}(s)=-i\pi\epsilon(s)$, so

$$\frac{1}{4\pi}\int f(\tau,l)\epsilon(\tau-s)g(s,l)d\tau\,ds=i\mathrm{P}\int\overline{\tilde{f}(\omega,l)}\tilde{g}(\omega,l)\,\frac{d\omega}{\omega}, \tag{5.8}$$

where P denotes the principal value operation. Integrating over d^2l we obtain

$$\frac{1}{4\pi}\int d^2l\int f(\tau,l)\epsilon(\tau-s)g(s,l)d\tau\,ds=i\int d^2l\,\mathrm{P}\int\overline{\tilde{f}(\omega,l)}\tilde{g}(\omega,l)\,\frac{d\omega}{\omega}. \tag{5.9}$$

Another form of the Fourier representation of this integral will be useful. We integrate $f(\tau' t\cdot l,l)\tilde{K}_{\alpha,\beta}(\omega)e^{-i\omega(\tau'-s')}g(s' t\cdot l,l)(t\cdot l)^2$ over $ds'\,d\tau'\,d^2l d\omega$, and then calculate the limits (in the same order as before). Changing the variables τ',s' back to τ,s we have

$$\frac{1}{4\pi}\int d^2l\int f(\tau,l)\epsilon(\tau-s)g(s,l)d\tau\,ds=i\mathrm{P}\int\frac{d\omega'}{\omega'}\int\overline{\tilde{f}\left(\frac{\omega'}{t\cdot l},l\right)}\tilde{g}\left(\frac{\omega'}{t\cdot l},l\right)d^2l. \tag{5.10}$$

Consider now the linear space of vector functions $u_a(\tau,l)$ such that $l\cdot u(\tau,l)=0$ and each component of u_a (in any Minkowski frame) is an element of the space $L^2_{\epsilon,t}$. Divide this space into classes with respect to the equivalence relation (4.4): $[u_a]=u_a(\tau,l)$ mod $\alpha(s,l)l_a$. We denote this factor space by $\mathscr{T}_\epsilon$. For decreasing ϵ the spaces $\mathscr{T}_\epsilon$ form an increasing family of vector spaces. Their union $\mathscr{T}:=\cup_{\epsilon>0}\mathscr{T}_\epsilon$ is therefore again a vector space. We now list some properties of $\mathscr{T}$ and of some structures on it. Simple proofs based on the preceding discussion are omitted.

$\mathscr{T}$ becomes a symplectic space with the form

$$\begin{aligned}\{[u_1],[u_2]\}&=\frac{1}{4\pi}\int d^2l\int u_{1a}(\tau,l)\epsilon(\tau-s)u_2^a(s,l)d\tau\,ds\\&=-i\mathrm{P}\int\frac{d\omega'}{\omega'}\int\left[-\overline{\tilde{u}_{1a}\left(\frac{\omega'}{t\cdot l},l\right)}\tilde{u}^a_2\left(\frac{\omega'}{t\cdot l},l\right)\right]d^2l.\end{aligned} \tag{5.11}$$

The form is nondegenerate on $\mathscr{T}$. Consider, further, the linear space of real, measurable vector functions $f_a(l)$ on the cone, homogeneous of degree -1, orthogonal to $l, l\cdot f(l)=0$, and such that each

component is an element of the Hilbert space $L^2(d^2l)$. Divide this space into equivalence classes $[f_a](l):=f_a(l)\text{mod }\beta(l)l_a$. This factor space is a real Hilbert space, denoted by $\mathscr{H}_0$, with the scalar product

$$([f_1],[f_2])_0=\int [-f_{1a}(l)f_2^a(l)]d^2l. \tag{5.12}$$

Now, one easily shows with the use of (5.2) that if $[u]\in\mathscr{T}$, then $[\tilde{u}(0,\cdot)]\in\mathscr{H}_0$. The map

$$\mathscr{T}\ni[u]\to p([u])=[\tilde{u}(0,\cdot)]\in\mathscr{H}_0 \tag{5.13}$$

is linear and onto, $p(\mathscr{T})=\mathscr{H}_0$ {for a given $[f]\in\mathscr{H}_0$ put $u_a(s,l)=f_a(l)(t\cdot l)^{-1}h(s/t\cdot l)$, with h of compact support and $\tilde{h}(0)=1$; then $p([u])=[f]$}. If at least one of the pair of functions $[u_i]\in\mathscr{T}$, $i=1,2$, satisfies $p([u_i])=0$, then, by (5.4), the following integral is well defined:

$$F([u_1],[u_2])=\int_{\omega\geqslant 0}[-\overline{\tilde{u}_{1a}(\omega,l)}\tilde{u}_2^a(\omega,l)]\,\frac{d\omega}{\omega}\,d^2l, \tag{5.14}$$

and the symplectic form may be expressed by

$$\{[u_1],[u_2]\}=2\,\text{Im}\,F([u_1],[u_2]), \tag{5.15}$$

where the hermiticity of F has been used:

$$\overline{F([u_1],[u_2])}=F([u_2],[u_1]). \tag{5.16}$$

For almost all l the integral $U_a(s,l)=-\int_s^{+\infty}u_a(\tau,l)\,d\tau$ is well defined for all s, and for almost all s there is $(\partial/\partial_s)U_a(s,l)=u_a(s,l)$. It is now easy to see that the functions $U_a(s,l)$ generalize the test functions $V_a(s,l)$, and the symplectic form (5.11) extends the form (3.10). The limit values at $-\infty$ may be expressed by $U_a(-\infty,l)=-2\pi\tilde{u}_a(0,l)$. The only essential property of V_a which has not been taken into account yet is (3.22), $l_{[a}\tilde{V}_{b]}(0,l)=-(1/2\pi)l_{[a}\partial_{b]}\Phi(l)$. In order to formulate its generalization for U_a consider the Hilbert subspace $\mathscr{H}_{IR}$ of $\mathscr{H}_0$ obtained as the closure of the linear subspace of elements $[\partial_a\phi]$, where ϕ is a real smooth homogeneous function of degree 0 {different extensions of $\phi(l)$ outside the cone yield $\partial_a\phi(l)$ on the cone in one class $[\partial_a\phi]$},

$$\mathscr{H}_{IR}:=\overline{\{[\partial_a\phi]\in\mathscr{H}_0|\,\phi\in\mathscr{C}^\infty\}}^{\mathscr{H}_0}. \tag{5.17}$$

For $\phi\in\mathscr{C}^\infty$ the class $[\partial_a\phi]$ determines ϕ uniquely up to a constant. With the notation $[\phi]\equiv\phi$ mod const, the map $[\partial_a\phi]\to[\phi]$ is linear and injective. We show in the Appendix that this map extends canonically to an injective map j of $\mathscr{H}_{IR}$ into the space of classes $[\phi]$ with ϕ square-integrable with respect to any (and all) of the measures $d^2l/(t\cdot l)^2$. In this way every element of $\mathscr{H}_{IR}$ corresponds uniquely to some $[\phi]\in j(\mathscr{H}_{IR})$, and may be written in this generalized sense as $[\partial_a\phi]$.

The extended symplectic space is now chosen as the subspace of $\mathscr{T}$ given by $\hat{L}^f:=p^{-1}(\mathscr{H}_{IR})$, with p defined in (5.13), and equipped with the symplectic form (5.11). The elements of $\hat{L}^f$ satisfying $p([u])=0$ form a linear subspace, denoted by $\hat{L}_0^f$. These structures extend the field test functions spaces $L^f\subset\hat{L}^f$ and $L_0^f=L^f\cap\hat{L}_0^f$.

The following properties of the form $F(.,.)$ will be needed for the characterization of vacuum states.

Lemma 5.1: (i) For each pair $[u_i]\in\hat{L}^f$, $i=1,2$, *the function* $x\to F([u_1],[T_xu_2-u_2])$ *is the boundary value for* $\text{Im}\,z=0$ *of the function*

$$\mathbf{R}^4+i\overline{V_+}\ni z\to F([u_1],[T_zu_2-u_2]):=-\int_{\omega\geqslant 0}(e^{i\omega l\cdot z}-1)\overline{\tilde{u}_1}\cdot\tilde{u}_2(\omega,l)\,\frac{d\omega}{\omega}\,d^2l,$$

which is continuous on its domain and analytic on $\mathbf{R}^4+iV_+$.
The following bounds (ii)–(iv) hold on the whole domain $z=x+iy\in\mathbf{R}^4+i\overline{V_+}$.

(ii) For any fixed, unit, future-pointing vector t,

$$|e^{F([u_1],[T_zu_2-u_2])}|\leqslant C(t,[u_1],[u_2])[(1+y^0+|\vec{y}|)^2+(|x^0|+|\vec{x}|)^2]^{(1/2)\int \tilde{u}_1\cdot\tilde{u}_2(0,l)\,\theta(\tilde{u}_1\cdot\tilde{u}_2(0,l))d^2l},$$

where $y^0\equiv y\cdot t$, $|\vec{y}|\equiv\sqrt{(y\cdot t)^2-y^2}$ *(and the same for x), and* $\theta(\cdot)$ *is the Heaviside step function.*

(iii) If $\tilde{u}_1\cdot\tilde{u}_2(0,l)\leqslant 0$ d^2l*-almost everywhere, then*

$$|e^{F([u_1],[T_zu_2-u_2])}|\leqslant C(t,[u_1],[u_2])[(1+y^0-|\vec{y}|)^2+\theta(x\cdot x)(|x^0|-|\vec{x}|)^2]^{-(1/2)\int|\tilde{u}_1\cdot\tilde{u}_2(0,l)|d^2l}.$$

(iv) If $\tilde{u}_1\cdot\tilde{u}_2(0,l)\geqslant 0$ d^2l*-almost everywhere, then*

$$|e^{-F([u_1],[T_zu_2-u_2])}|\leqslant C(t,[u_1],[u_2]).$$

(v) If $p([u_1+u_2])=0$, *then for* $x\in\mathbf{R}^4$

$$F([u_1+T_xu_2],[u_1+T_xu_2])=F([u_1],[T_xu_2-u_2])+F([T_xu_2-u_2],[u_1])+F([u_1+u_2],[u_1+u_2]).$$

(vi) The Fourier representation (5.14) of $F(.,.)$ *is the usual one-photon scalar product when restricted to* $\hat{L}_0^f$, *which yields a dense subspace of the one-photon Hilbert space, and*

$$F([T_xu_1],[T_xu_2])=F([u_1],[u_2]).$$

Proof: If $z\in\mathbf{R}^4+iV_+$, then for any $k\in\mathbf{R}^4$ the function $\mathbf{C}\ni\xi\to F([u_1],[T_{z+\xi k}u_2-u_2])$ is analytic in some neighborhood of $\xi=0$, which implies (i) by Hartog's theorem. Properties (ii)–(iv) follow easily from the bound

$$\left|F([u_1],[T_zu_2-u_2])-\int\tilde{u}_1\cdot\tilde{u}_2(0,l)\ln\left|1-i\,\frac{z\cdot l}{t\cdot l}\right|d^2l\right|\leqslant\text{const}(t,[u_1],[u_2]),$$

and the inequalities $1\leqslant(1+y^0-|\vec{y}|)^2+\theta(x\cdot x)(|x^0|-|\vec{x}|)^2\leqslant|1-iz\cdot l/t\cdot l|^2\leqslant(1+y^0+|\vec{y}|)^2+(|x^0|+|\vec{x}|)^2$. To prove the bound we split F into two parts

$$F([u_1],[T_zu_2-u_2])=-\int_{\omega\geqslant 0}(e^{i\omega l\cdot z}-1)[\overline{\tilde{u}_1}\cdot\tilde{u}_2(\omega,l)-\tilde{u}_1\cdot\tilde{u}_2(0,l)e^{-\omega t\cdot l}]\,\frac{d\omega}{\omega}\,d^2l$$
$$-\int\left[\tilde{u}_1\cdot\tilde{u}_2(0,l)\int_0^\infty\left(\exp\left(i\omega'\,\frac{z\cdot l}{t\cdot l}\right)-1\right)e^{-\omega'}\,\frac{d\omega'}{\omega'}\right]d^2l.$$

The first term is absolutely bounded by $2\int|\overline{\tilde{u}_1}\cdot\tilde{u}_2(\omega,l)-\tilde{u}_1\cdot\tilde{u}_2(0,l)e^{-\omega t\cdot l}|(d\omega/\omega)d^2l<\infty$. The second term is equal to $\int\tilde{u}_1\cdot\tilde{u}_2(0,l)\ln(1-iz\cdot l/t\cdot l)d^2l$. The imaginary part of $\ln(1-iz\cdot l/t\cdot l)$ yields a term bounded in z, which ends the proof of the bound. Property (v) is easily proved by straightforward calculation in the special case $u_1=-u_2$, and then the general case follows from $u_1+T_xu_2=(u_1+u_2)+(T_xu_2-u_2)$. Statement (vi) follows from our discussion of the relation of the present formulation with the traditional one, ending Sec. III. □

VI. VACUUM VERSUS REGULAR, POSITIVE ENERGY REPRESENTATIONS

Now we can take up the study of physical representations of the asymptotic algebra. First of all, vacuum states have to be characterized.

Theorem 6.1: *(i) If a cyclic representation* π *of the algebra* $\mathscr{T}$ *satisfies Borchers' criterion with respect to the Weyl algebra of the electromagnetic field, and* $U(x)\Omega=\Omega$ *for some choice of the pertinent representation of translations* $U(x)$ *and of the cyclic vector* Ω, *then for each* $C\in\text{CAR}$

$$(\Omega,\pi(CW(V))\Omega)=0\quad if\quad l\wedge V(-\infty,l)\neq 0. \tag{6.1}$$

(ii) A cyclic representation π of the algebra $\mathscr{F}$ satisfies Borchers' criterion and $U(x)\Omega=\Omega$ for some choice of the pertinent representation of translations $U(x)$ and of the cyclic vector Ω if, and only if, there is $(\Omega,\pi(CW(V))\Omega)=\omega_F(C)\omega_r(W(V))$, where ω_F is the positive energy Fock state on CAR *and ω_r is the state on* CCR *given by*

$$\omega_r(W(V))=\begin{cases}0, & if \quad l\wedge V(-\infty,l)\neq 0,\\ f(V)e^{-(1/2)F([\dot V],[\dot V])}, & if \quad l\wedge V(-\infty,l)=0,\end{cases} \tag{6.2}$$

where $f:L_0\to\mathbf{C}$ is a function of positive type, satisfying the condition

$$f(V_1+(T_x-1)V_2)=f(V_1) \quad \forall (V_1)\in L_0,(V_2)\in L,x\in M. \tag{6.3}$$

Proof: (i) Using the algebraic relations (3.40) and (3.42) and the relation (5.15) one finds

$$e^{F([\dot V_1],[T_x\dot V_2-\dot V_2])}CW(V_1)W(T_xV_2)=e^{-i\{V_1,V_2\}}\overline{e^{F([\dot V_1],[T_x\dot V_2-\dot V_2])}}W(T_xV_2)\beta_{-\Phi_2}(C)W(V_1).$$

It follows now from the invariance of Ω under $U(x)$ and from the spectral properties of $U(x)$ that the value of $\omega(.):=(\Omega,\pi(.)\Omega)$ on the lhs of this identity is the boundary value for Im $z=0$ of the function

$$R^4+i\overline{V_+}\ni z\to e^{F([\dot V_1],[T_z\dot V_2-\dot V_2])}(\Omega,\pi(CW(V_1))U(z)\pi(W(V_2))\Omega),$$

continuous on its domain and analytic inside. By Lemma 5.1(ii) this function is polynomially bounded. The expectation value of the rhs of the above identity has similar properties in the time-reflected region $\mathbf{R}^4-i\overline{V_+}$. By the edge of the wedge theorem there is an open region containing $\mathbf{R}^4+i\overline{V_+\cup V_-}$ and a function analytic in this region which is the analytic continuation of both these functions. This function is polynomially bounded on $\mathbf{R}^4+i\overline{V_+\cup V_-}$, hence it is a polynomial. To prove this implication let $h(z)$ be a function with the stated properties, and choose a basis of Minkowski vector space consisting of four unit future-pointing vectors t_i, $i=1,...,4$. For fixed real α^2, α^3, α^4 the function $\mathbf{C}\ni\xi\to h(\xi t_1+\alpha^2t_2+\alpha^3t_3+\alpha^4t_4)$ is analytic and polynomially bounded, hence by Cauchy inequality for analytic functions it is a polynomial in ξ. Therefore, by repeated use of the implication, $h(x)$ is a polynomial, hence $h(z)$ is the same polynomial for complex z. We have thus proved that

$$e^{F([\dot V_1],[T_x\dot V_2-\dot V_2])}\omega(CW(V_1)W(T_xV_2))=P_{C,V_1,V_2}(x), \tag{6.4}$$

where P_{C,V_1,V_2} is a polynomial. If $p([\dot V_1])=p([\dot V_2])\neq 0$, then by Lemma 5.1(iii) $P_{C,V_1,V_2}(x)=0$ (take $|\vec x|=$const and $|x^0|\to+\infty$). For $x=0$, $[\dot V_1]=[\dot V_2]=\frac{1}{2}[\dot V]$, $p([\dot V])\neq 0$ we get $\omega(CW(V))=0$, which ends the proof if (i).

(ii) If $U(x)$ implement translations and $U(x)\Omega=\Omega$ for cyclic Ω, then writing each $C\in{}^*\mathrm{CAR}$ in the form $C=\omega_F(C)E+C'$, where ω_F is the positive energy Fock state and C' is a sum of elements having $B(g_+)$, $B(g_-)^*$ on the right and/or $B(g_+)^*$, $B(g_-)$ on the left, one finds that $\omega(CW(V))=\omega_F(C)\omega(W(V))$. By (3.40), (5.15), and Lemma 5.1(v), for $p([\dot V_1+\dot V_2])=0$ Eq. (6.4) is now equivalent to

$$f(V_1+T_xV_2)=P'_{V_1,V_2}(x), \tag{6.5}$$

where P'_{V_1,V_2} is a polynomial and f is a function on L_0 defined by

$$f(V)=\omega(W(V))e^{(1/2)F([\dot V],[\dot V])}. \tag{6.6}$$

By Lemma 5.1(vi) $f(V_1+T_xV_2)$ is absolutely bounded as a function of x for fixed (V_1), $(V_2)\in L_0$. Hence, by (6.5),

$$f(V_1+T_xV_2)=f(V_1+V_2) \quad \text{for } (V_1),(V_2)\in L_0. \tag{6.7}$$

The positivity of $\omega(A^*A)$ for $A=\Sigma_{i=1}^{n}\alpha_i W(V_i)$, for $(V_i)\in L_0$ and all sequences of complex numbers $\{\alpha_i\}_{i=1}^{n}$, is equivalent to the condition $\Sigma_{i,k=1}^{n}\overline{\beta_i}\beta_k e^{F([V_i],[V_k])}f(V_k-V_i)\geqslant 0$ for all sequences of complex numbers $\{\beta_i\}_{i=1}^{n}$. Following Ref. 24 we replace all V_i in this condition by $(1/N)\Sigma_{k=0}^{N-1}\underline{T_x^k V_i}$ and take the limit $N\to\infty$. Then by (6.7), Lemma 5.1(vi), and the ergodic theorem one has $\Sigma\overline{\beta_i}\beta_k f(V_k-V_i)\geqslant 0$, hence f is of positive type [the remaining conditions, $f(0)=1$ and $\overline{f(V)}=f(-V)$, are obviously satisfied]. Consequently, f is bounded and, by (6.5), (6.7) is satisfied for all (V_1), $(V_2)\in L$ with $(V_1+V_2)\in L_0$. This is equivalent to (6.3), which ends the proof that the state ω has the form given in (ii).

Conversely, let ω_r be a linear functional of the form given by (ii). Then it is a state (one uses the fact that if A and B are two Hermitian positive matrices $n\times n$, then the matrix C defined by $C_{ik}=A_{ik}B_{ik}$ is also positive). Translations are implemented in π_r by the canonically defined representation of translations $U_r(x)$ (ω_r is translationally invariant). This representation is strongly continuous and satisfies the spectrum condition. To show this, it is sufficient to demonstrate that for any pair (V_1), $(V_2)\in L$ the function $x\to\omega_r(W(V_1)W(T_xV_2))$ is continuous, and its Fourier transform has support in $\overline{V_+}$. If $(V_1+V_2)\notin L_0$, then this function vanishes identically. For $(V_1+V_2)\in L_0$, we have by (5.15) and the Lemma 5.1(v)

$$\begin{aligned}\omega_r(W(V_1)W(T_xV_2))&=e^{-(i/2)\{V_1,V_2\}}e^{-(i/2)\{V_1,T_xV_2-V_2\}}\omega_r(W(V_1+T_xV_2))\\&=\omega_r(W(V_1)W(V_2))e^{-F([\dot V_1],[T_x\dot V_2-\dot V_2])}.\end{aligned}$$

The function $\mathbf{R}^4+i\overline{V_+}\ni z\to e^{-F([u_1],[T_zu_2-u_2])}$ is analytic and, by Lemma 5.1(iv), bounded on its domain. Therefore it is the Laplace transform of a distribution with support in $\overline{V_+}$.[25] Now, the Fock state ω_F on CAR satisfies (2.8), so the state on $\mathscr{T}$ defined by $\omega(CW(V))=\omega_F(C)\omega_r(W(V))$ generates a representation unitarily equivalent to a representation of the type (2.10). Translations are implemented in this representation by $U_F(x)\otimes U_r(x)$, where $U_F(x)$ is the representation canonically correlated to ω_F, which ends the proof. □

Let us denote by $\mathscr{T}_0$ the subalgebra of $\mathscr{T}$ generated by CAR and by elements $W(V)$ with $(V)\in L_0$. The representation space $\mathscr{H}$ of a vacuum representation, as implied by the general result (i) of the above theorem, is easily seen to be the uncountable direct sum of the space $[\pi(\mathscr{T}_0)\Omega]$ and spaces derived from it by the action of operators $\pi(W(V))$; if $l\wedge V_1(-\infty,l)=l\wedge V_2(-\infty,l)$ the respective spaces are equal, in other cases they are orthogonal. In consequence, vacuum representations are nonregular with respect to the Weyl operators with infrared-singular test functions $[l\wedge V(-\infty,l)\neq 0]$. Now, the derivation of our asymptotic algebra has led unambiguously to its interpretation, as explained in Secs. III and IV. From the point of view of this interpretation the above structure does not seem a physically justified idealization, for two reasons:

(i) The infrared singular Weyl operators, being degraded to operators intertwining between different representations of $\mathscr{T}_0$, no longer describe the electromagnetic field observable. However, the regular Weyl operators, as discussed earlier, are functions of the free outgoing field only, and the Coulomb field is lost. If the vacuum is of the form (ii) of the theorem, one can separately *define* it by the first term in (4.3), but this is done ''by hand,'' and the information on the unique way in which the Coulomb field and the ''out'' field add to form the total field is lost.

(ii) The superselection sectors with respect to regular observables $\mathscr{A}_0:=\mathscr{T}_0\cap\mathscr{A}$ are labeled by the spacelike asymptotic of the *free* field (and, of course, by total charge, if it exists), so that even the standard wisdom does not apply here. The spacelike asymptotic of electromagnetic field according to Buchholz[14] yields in the subspace $[\pi(W(V))\pi(\mathscr{T}_0)\Omega]$ the field (3.6), which is a free field for free field asymptotic V. Moreover, $\mathscr{A}_0$ should not be interpreted as the algebra of local observables: creation or annihilation of a charged particle together with its Coulomb field is a nonlocal operation, so $B(g)^*B(f)\in\mathscr{A}_0$ is a nonlocal observable.

Consider the particular, Poincaré-invariant vacuum state as given by Theorem 6.1(ii) with $f\equiv 1$. In this representation there is no infraparticle problem: all one-particle states $\pi(B(g_+)^*)\Omega$, $\pi(B(g_-))\Omega$ have energy-momentum on the mass hyperboloid. We want to construct representations which remain as close in their structure to this vacuum representation, but which are regular

at the same time. The obvious idea how to do it is to try to integrate the superselection sectors of this vacuum into a direct integral Hilbert space. As the representation is of the form determined by Theorem 4.4, it is sufficient to confine attention to the electromagnetic part of the representation, π_r. However, for measure-theoretical reasons one has to extend the scope of sectors which are to be integrated. It is now that the extension of the symplectic space introduced in Sec. V will be needed. This extension allows us to consider the Weyl algebra $\widehat{\mathrm{CCR}}$ generated uniquely (due to nondegeneracy of the symplectic form[26]) by elements $\hat{W}([u])$, $[u]\in\hat{L}^f$,

$$\hat{W}([u_1])\hat{W}([u_2])=e^{-(i/2)\{[u_1],[u_2]\}}\hat{W}([u_1+u_2]).$$

The elements $\hat{W}([u])$ with $[u]$ from the subspace $\hat{L}^f_0$ generate a C^*-subalgebra, which we denote $\widehat{\mathrm{CCR}}_0$. The Poincaré transformations act on the algebra $\widehat{\mathrm{CCR}}$ by

$$\alpha_{x,A}\hat{W}([u])=\hat{W}([T_{x,A}u]),$$

$$(T_{x,A}u)_a(s,l)=\Lambda(A)_a{}^b u_b(s-x\cdot l,\Lambda^{-1}l).$$

The Poincaré-invariant vacuum state is easily obtained:

$$\hat{\omega}(\hat{W}([u]))=\begin{cases}0, & \text{if } p([u])\neq 0,\\ e^{-(1/2)F([u],[u])}, & \text{if } p([u])=0.\end{cases}$$

The representation space $\hat{\mathscr{H}}$ of the representation $\hat{\pi}$ canonically obtained from $\hat{\omega}$ is the uncountable direct sum $\hat{\mathscr{H}}=\oplus_{[f]\in\mathscr{H}_{IR}}\hat{\mathscr{H}}_{[f]}$, where $\hat{\mathscr{H}}_{[f]}$ is the subspace spanned by the vectors $\hat{\pi}(\hat{W}([u]))\hat{\Omega}$ with $p([u])=[f]$. The restriction of the representation $\hat{\pi}$ to the subalgebra $\widehat{\mathrm{CCR}}_0$ and to the subspace $\hat{\mathscr{H}}_{[f]}$ is a coherent state representation,[20] $\hat{\pi}_{[f]}(\widehat{\mathrm{CCR}}_0):=\hat{\pi}(\widehat{\mathrm{CCR}}_0)|_{\hat{\mathscr{H}}_{[f]}}$. Two representations $\hat{\pi}_{[f_i]}$, $i=1,2$, are disjoint for $[f_1]\neq[f_2]$.[20] In particular, $\hat{\pi}_{[0]}(\widehat{\mathrm{CCR}}_0)$ is the (positive energy) Fock representation. Each subspace $\hat{\mathscr{H}}_{[f]}$ is invariant under the action of translations $\hat{U}(x)$, thus $\hat{U}(x)=\oplus_{[f]\in\mathscr{H}_{IR}}\hat{U}_{[f]}(x)$, each component representation being a positive energy representation, Spec $\hat{U}_{[f]}(x)\subset\overline{V_+}$.

Suppose we are given a cylindrical σ-additive measure μ on the Hilbert space $\mathscr{H}_{IR}$ (5.17).[27] Let $\psi([f])$ vary over the set of all measurable functions $\mathscr{H}_{IR}\to\hat{\mathscr{H}}_{[0]}$ [i.e., the set of functions for which $(\varphi,\psi([f]))_{\hat{\mathscr{H}}_{[0]}}$ is measurable for all $\varphi\in\hat{\mathscr{H}}_{[0]}$ (Ref. 22)]. Let further $h(s,l)$ vary over the set of smooth homogeneous functions of degree -1, satisfying for some $\epsilon>0$ and all $k=0,1,\ldots$ the bounds

$$|L_{b_1c_1}\cdots L_{b_kc_k}h(s,l)|\leqslant\frac{\mathrm{const}(k)}{(|s|+1)^{1+\epsilon}}\tag{6.8}$$

(for any gauge $t\cdot l=1$) and such that $\tilde{h}(0,l)=1$. Then $[fh]\in\hat{L}^f$ and $p([fh])=[f]$ for any $[f]\in\mathscr{H}_{IR}$. Consider the set Γ of all functions of the form

$$\mathscr{H}_{IR}\ni[f]\to\Psi([f])=\hat{\pi}(\hat{W}([fh]))\psi([f])\in\hat{\mathscr{H}}_{[f]}.\tag{6.9}$$

Lemma 6.2: For any fixed h every function in Γ has a unique representation (6.9).

Proof: If h' is another function satisfying the same conditions as h, then

$$\hat{\pi}(\hat{W}([fh']))\psi([f])=\hat{\pi}(\hat{W}([fh]))e^{(i/2)\{[fh],[fh']\}}\hat{\pi}_{[0]}(\hat{W}([f(h'-h)]))\psi([f]).$$

It has to be shown that if $\psi([f])$ is measurable, then $\psi'([f])=e^{(i/2)\{[fh],[fg]\}}\times\hat{\pi}_{[0]}(\hat{W}([fg]))\psi([f])$, where $g\equiv h'-h$, is also measurable. It is sufficient to show that $(e_i,\psi'([f]))$ are measurable for a basis $\{e_i\}$ of $\hat{\mathscr{H}}_{[0]}$. If one chooses for $\{e_i\}$ the finite ''particle'' number basis, then these products have the following form: $e^{(i/2)\{[fh],[fg]\}-(1/2)F([fg],[fg])}\sum_{k=1}^{\infty}(e_k,\psi([f]))C_k([f])$, where $C_k([f])$ are linear combinations of products of expressions $F([fg],[\chi])$ and their complex conjugations, where $[\chi]\in\hat{L}^f_0$ are profiles

of the photons in the basis vectors. Now, the expressions $F([fg],[\chi])$, $\{[fh],[fg]\}$, and $F([fg],[fg])$ are easily seen to be absolutely bounded as functions of $[f]$, the first by const$\|[f]\|_{\mathscr{K}_{IR}}$, the other two by const$\|[f]\|^2_{\mathscr{K}_{IR}}$. Hence, the first may be written as $([f],[k])$, $[k] \in \mathscr{K}_{IR}$, and the other two as $([f],B[f])_{\mathscr{K}_{IR}}$, where B is a bounded operator on $\mathscr{K}_{IR}$. Both expressions are measurable functions of $[f]$, which ends the proof. □

The set Γ thus has the natural structure of a vector space, and the pair $(\{\hat{\mathscr{H}}_{[f]}\}_{[f]\in\mathscr{K}_{IR}},\Gamma)$ is easily seen to form a μ-measurable family of Hilbert spaces.[22] This family determines the direct integral Hilbert space $\mathscr{H}=\int_{\oplus}\hat{\mathscr{H}}_{[f]}d\mu([f])$, with its elements denoted by $\Psi = \int_{\oplus}\Psi([f])\, d\mu([f])$. It is necessary for our purposes to assume that μ is quasi-invariant with respect to translations by smooth elements of $\mathscr{K}_{IR}$. This set, more exactly the set of elements in $\mathscr{K}_{IR}$ having smooth representants, will be denoted by $\mathscr{C}^{\infty}_{IR}$. Hence, we demand that

$$\mu_{[k]}(B) := \mu(B-[k]) = 0 \Leftrightarrow \mu(B) = 0 \quad \forall \text{ measurable } B,\ [k] \in \mathscr{C}^{\infty}_{IR}. \tag{6.10}$$

(This was the reason for the need to extend L^f to $\hat{L}^f$, and consequently $\mathscr{C}^{\infty}_{IR}$ to $\mathscr{K}_{IR}$: there are no measures on function spaces quasi-invariant with respect to all translations.[27]) Then, by Radon–Nikodym theorem, $d\mu_{[k]}([f]) = (d\mu_{[k]}/d\mu)([f])d\mu([f])$, where $(d\mu_{[k]}/d\mu)([f])$ and $[(d\mu_{[k]}/d\mu)([f])]^{-1}$ are integrable, non-negative functions.

Formula (6.9) may be used to define, for each fixed h, an isomorphism of the space $\mathscr{H}$ with the tensor product space $L^2(\mathscr{K}_{IR},\mu)\otimes\hat{\mathscr{H}}_{[0]}$. Namely, we put by definition

$$\mathscr{U}_h : L^2(\mathscr{K}_{IR},\mu)\otimes\hat{\mathscr{H}}_{[0]} \to \mathscr{H},$$

$$\chi\otimes\psi \to \Psi = \int_{\oplus}\Psi([f])\, d\mu([f]), \tag{6.11}$$

$$\Psi([f]) = \hat{\pi}(\hat{W}([fh]))\chi([f])\psi.$$

This is a noncanonical isomorphism, as it depends on the choice of the function h, which has no intrinsic meaning. The use of the isomorphism will be restricted to technical purposes only.

We now define in $\mathscr{H}$ a new representation of CCR. Note that if V is a test function of $W(V)$, then by (3.35) $\tilde{V}(\omega,l)$ is a smooth function of l for each ω. Therefore $p([\dot{V}])$ is a smooth element of $\mathscr{K}_{IR}$ [has a smooth representant $\tilde{V}(0,l)$].

Theorem 6.3: *The linear operators on $\mathscr{H}$ introduced by*

$$[\pi_r(W(V))\Psi]([f]) := \left(\frac{d\mu_{p([\dot{V}])}}{d\mu}([f])\right)^{1/2}\hat{\pi}(\hat{W}([\dot{V}]))\Psi([f]-p([\dot{V}])) \tag{6.12}$$

define a representation π_r of CCR *satisfying Borchers' criterion. The unitary representation of translations defined by*

$$[U_r(x)\Psi]([f]) := \hat{U}_{[f]}(x)\Psi([f]) \tag{6.13}$$

implements translations in the representation π_r and satisfies Spec $U(x) = \overline{V_+}$.

Proof: A straightforward calculation shows that π_r and U_r satisfy the algebraic conditions of representations and that $U_r(x)$ implement translations in the representation π_r. From the strong continuity and spectral properties of each of $\hat{U}_{[f]}(x)$ it follows that also $U_r(x)$ is strongly continuous and Spec $U_r(x)\subset\overline{V_+}$. The proof will be now completed by showing that Spec $\hat{U}_{[0]}(x)$ $\subset$ Spec $U_r(x)$ ($\hat{U}_{[0]}$ is the standard transformation of the Fock representation, hence Spec $\hat{U}_{[0]}(x) = \overline{V_+}$). Take $\Psi = \mathscr{U}_h(\chi\otimes\psi)$, with $\|\chi\|_{L^2(\mathscr{K}_{IR},\mu)} = 1$, $\|\psi\|_{\hat{\mathscr{H}}_{[0]}} = 1$. Then

$$(\Psi, U_r(x)\Psi) = \int |\chi([f])|^2 e^{(i/2)\{[fh],[fT_xh]\}}(\psi, \hat{\pi}_{[0]}(\hat{W}([f(T_xh-h)]))\hat{U}_{[0]}(x)\psi)\, d\mu([f]).$$

Let $\Psi_k=\mathscr{U}_{h_k}(\chi\otimes\psi)$ $(k=1,2,...)$, with $h_k(s,l)=k^{-1}h(k^{-1}s,l)$. Then

$$\lim_{k\to\infty} F([fh_k],[f(T_xh_k-h_k)])=0, \quad \lim_{k\to\infty} F([f(T_xh_k-h_k)],[f(T_xh_k-h_k)])=0$$

and

$$\mathrm{w}-\lim_{k\to\infty} \hat{\pi}_{[0]}(\hat{W}([f(T_xh_k-h_k)]))=1.$$

By Lebesgue's theorem

$$\lim_{k\to\infty}(\Psi_k, U_r(x)\Psi_k)=(\psi,\hat{U}_{[0]}(x)\psi).$$

Hence Spec $\hat{U}_{[0]}(x)\subset$Spec $U_r(x)$. □

Further properties of the representation π_r are discussed after finding its unitarily equivalent form $\mathscr{U}_h^{-1}\pi_r\mathscr{U}_h$ by (6.11). Let $V_a(s,l)$ be any test function of $W(V)$. The integral

$$-\frac{1}{4\pi}\int h(s,l)\epsilon(s-\tau)\dot{V}_a(\tau,l)ds\ d\tau=-i\mathrm{P}\int \overline{\tilde{h}(\omega,l)}\tilde{V}_a(\omega,l)\ \frac{d\omega}{\omega}$$

is a real smooth vector function orthogonal to l, and a representant of an element of $\mathscr{H}_0$ [defined before (5.12)]. The orthogonal projection of this element to the subspace $\mathscr{H}_{IR}$ of $\mathscr{H}_0$ will be denoted by $r_h([\dot{V}])$. We show in the Appendix that the orthogonal projection to $\mathscr{H}_{IR}$ of a smooth element of $\mathscr{H}_0$ is smooth, hence $r_h([\dot{V}])\in\mathscr{C}^\infty_{IR}$. Let ccr be the Weyl algebra over the vector space $\mathscr{C}^\infty_{IR}\oplus\mathscr{C}^\infty_{IR}$ with the symplectic form $\{[g_1]\oplus[k_1],[g_2]\oplus[k_2]\}_{IR}:=([g_1],[k_2])_{\mathscr{H}_{IR}}-([k_1],[g_2])_{\mathscr{H}_{IR}}$. This algebra is generated by elements $w([g]\oplus[k])$ satisfying

$$\begin{gathered} w([g_1]\oplus[k_1])w([g_2]\oplus[k_2])=e^{-(i/2)\{[g_1]\oplus[k_1],[g_2]\oplus[k_2]\}_{IR}}w([g_1+g_2]\oplus[k_1+k_2]),\\ w([g]\oplus[k])^*=w(-([g]\oplus[k])). \end{gathered} \tag{6.14}$$

(This algebra is unambiguously defined as the symplectic form is nondegenerate.[26])

Theorem 6.4: *(i) The following definition determines a representation π_μ of the algebra* ccr *on the Hilbert space $L^2(\mathscr{H}_{IR},\mu)$:*

$$(\pi_\mu(w([g]\oplus[k]))\chi)([f]):=\left(\frac{d\mu_{[g]}}{d\mu}([f])\right)^{1/2}e^{i([f]-(1/2)[g],[k])_{\mathscr{H}_{IR}}}\chi([f-g]). \tag{6.15}$$

The vector $\Omega_\mu([f])=1$ is cyclic in $L^2(\mathscr{H}_{IR},\mu)$ for the representation π_μ restricted to the subalgebra generated by elements $w([0]\oplus[k])$.

(ii) The representation π_r goes over under the unitary transformation $\mathscr{U}_h$ of representation space to

$$\mathscr{U}_h^{-1}\pi_r(W(V))\mathscr{U}_h=\pi_\mu(w(p([\dot{V}])\oplus r_h([\dot{V}])))\oplus\hat{\pi}_{[0]}(\hat{W}([\dot{V}-\tilde{\dot{V}}(0,.)h])). \tag{6.16}$$

(iii) $(\mathscr{U}_h^{-1}\pi_r(\mathrm{CCR})\mathscr{U}_h)''=\pi_\mu(\mathrm{ccr})''\otimes\mathscr{L}(\hat{\mathscr{H}}_{[0]})$ (the von Neumann tensor product), where $\mathscr{L}(\hat{\mathscr{H}}_{[0]})$ is the algebra of bounded operators on $\hat{\mathscr{H}}_{[0]}$. Therefore, π_r is irreducible iff π_μ is irreducible.

(iv) π_r is regular iff π_μ is regular.

Proof: (i) The algebraic properties needed for π_μ to be a representation of ccr are checked by direct calculation. For any $\chi\in L^2(\mathscr{H}_{IR},\mu)$ there is

$$(\chi,\pi_\mu(w([0]\oplus[k]))\Omega_\mu)=\int\overline{\chi([f])}e^{i([f],[k])}\,d\mu([f]),$$

If this integral vanishes for all smooth $[k]$, then by continuity in $[k]$ it vanishes for all $[k]\in\mathscr{H}_{IR}$, and then $\chi=0$, which shows that Ω_μ is cyclic.

(ii) A direct calculation gives

$$[\pi_r(W(V))\mathscr{U}_h(\chi\otimes\psi)]([f])=\hat{\pi}(\hat{W}([fh]))\left(\frac{d\mu_{p([\dot V])}}{d\mu}([f])\right)^{1/2}e^{i\{[(f-(1/2)\tilde{\dot V}(0,.))h],[\dot V]\}}$$
$$\times\chi([f]-p([\dot V]))\hat{\pi}_{[0]}(\hat{W}([\dot V-\tilde{\dot V}(0,.)h]))\psi.$$

Denote $f'\equiv f-\frac{1}{2}\tilde{\dot V}(0,.)$. The symplectic form in the exponent of the last formula is evaluated by

$$\{[f'h],[\dot V]\}=-\int f'_a(l)\left(-i\mathrm{P}\int\overline{\tilde h(\omega,l)}\tilde{\dot V}^a(\omega,l)\,\frac{d\omega}{\omega}\right)d^2l$$
$$=([f'],r_h([\dot V]))_{\mathscr{H}_{IR}}=([f]-\tfrac{1}{2}p([\dot V]),r_h([\dot V]))_{\mathscr{H}_{IR}}.$$

Transforming back with $\mathscr{U}_h^{-1}$ one obtains (6.16).

(iii) Let $(V')\in L_0$. Further, let $f_a(l)$ be a smooth function representing an element $[f]\in\mathscr{H}_{IR}$ and put $\dot V_a(s,l)=f_a(l)h(s,l)+\dot V'_a(s,l)$. Then $(V)\in L$, $p([\dot V])=[f]$, and

$$\mathscr{U}_h^{-1}\pi_r(W(V))\mathscr{U}_h=\pi_\mu(w([f]\oplus r_h([\dot V'])))\otimes\hat{\pi}_{[0]}(\hat{W}([\dot V'])). \tag{6.17}$$

The functions $\tilde{\dot V}'$ are dense in the Hilbert space of test functions of the Fock representation, which is irreducible. Therefore, to prove the equality in (iii) it is sufficient to show that all operators $\pi_\mu(w([f]\oplus[g]))\otimes\mathbf{1}_{[0]}$ are in $(\mathscr{U}_h^{-1}\pi_r(\mathrm{CCR})\mathscr{U}_h)''$. The second statement of (iii) then follows from $(\mathscr{U}_h^{-1}\pi_r(\mathrm{CCR})\mathscr{U}_h)'=\pi_\mu(\mathrm{ccr})'\otimes\mathbf{C1}_{[0]}$ (see Ref. 22).

To fill the missing step consider for $\beta\in(0,\epsilon)$ a one-parameter family of functions $\tilde N_\beta(\omega,l)=\tilde\kappa_\beta(\omega t\cdot l)\tilde h(\omega,l)$, where $\tilde\kappa_\beta(\omega)=ie^{-|\omega|}|\omega|^\beta\,\mathrm{sgn}(\omega)$. Then $N_\beta(s,l)$ are real homogeneous functions of degree -1 given by

$$N_\beta(s,l)=\frac{1}{2\pi t\cdot l}\int\kappa_\beta\left(\frac{s-\tau}{t\cdot l}\right)h(\tau,l)\,d\tau,$$

with $\kappa_\beta(s)=2\Gamma(1+\beta)(s^2+1)^{-(1+\beta)/2}\sin((1+\beta)\mathrm{arctg}\,s)$. With the use of bounds (6.8) one shows that $N_\beta(s,l)$ are smooth and also satisfy bounds of the form (6.8), with ϵ replaced by β. Consider the smooth homogeneous function of degree 0 given by the integral

$$c_\beta(l)=-\frac{1}{4\pi}\int h(\tau,l)\epsilon(\tau-s)N_\beta(s,l)\,d\tau\,ds=2\int_0^\infty\left|\tilde h\left(\frac{\omega'}{t\cdot l},l\right)\right|^2e^{-\omega'}\omega'^{\beta-1}\,d\omega',$$

the last equality by (5.8). From the bound of the form (5.4) satisfied by $\tilde h$ and the condition $\tilde h(0,l)=1$ we know that there is a positive u, such that $|\tilde h(\omega'/t\cdot l,l)|^2e^{-\omega'}>\frac{1}{2}$ for $\omega'\in\langle 0,u\rangle$. On the other hand, the function $|\tilde h(\omega'/t\cdot l,l)|^2$ is bounded by a constant from above. Hence, $(u^\beta/\beta)<c_\beta(l)<\mathrm{const}\,\Gamma(\beta)$. Therefore, the new auxiliary function defined by $n_\beta(s,l)=N_\beta(s,l)/c_\beta(l)$ is smooth, has all the properties listed above for $N_\beta(s,l)$, and in addition satisfies

$$-\frac{1}{4\pi}\int h(\tau,l)\epsilon(\tau-s)n_\beta(s,l)\,d\tau\,ds=1. \tag{6.18}$$

Now, choose a smooth element $[g]$ in $\mathscr{H}_{IR}$ and put $\dot V'_{\beta a}(s,l)=g_a(l)n_\beta(s,l)$. These V'_β satisfy the conditions for a test function V' in (6.17) and, by (6.18), yield $r_h([\dot V'_\beta])=[g]$. For $\dot V_{\beta a}(s,l)=f_a(l)h(s,l)+\dot V'_{\beta a}(s,l)$ Eq. (6.17) reads

$$\mathscr{U}_h^{-1}\pi_r(W(V_\beta))\mathscr{U}_h=\pi_\mu(w([f]\oplus[g]))\otimes\hat{\pi}_{[0]}(\hat{W}([\dot V'_\beta])). \tag{6.19}$$

By a straightforward calculation one obtains

$$F([\dot{V}'_\beta],[\dot{V}'_\beta]) = \int d^2l \frac{-g^2(l)}{c^2_\beta(l)} \int_0^\infty \left| \tilde{h}\left(\frac{\omega'}{t\cdot l}, l\right)\right|^2 e^{-2\omega'} \omega'^{2\beta-1}\, d\omega'$$

$$< \int (-g^2(l)) \frac{c_{2\beta}(l)}{2c^2_\beta(l)}\, d^2l$$

$$< \text{const}\, \frac{\beta^2\Gamma(2\beta)}{u^{2\beta}} \int (-g^2(l)) d^2l = \text{const}\, \frac{\Gamma(1+2\beta)}{u^{2\beta}}\, \beta.$$

Hence $\lim_{\beta\to 0} F([\dot{V}'_\beta],[\dot{V}'_\beta]) = 0$, and then $\mathrm{w}-\lim_{\beta\to 0} \hat{\pi}_{[0]}(\hat{W}([\dot{V}'_\beta])) = \mathbf{1}$. Then, by (6.19), $\mathrm{w}-\lim_{\beta\to 0} \mathscr{U}_h^{-1} \pi_r(W(V_\beta)) \mathscr{U}_h = \pi_\mu(w([f]\oplus[g])) \otimes \mathbf{1}$, which ends the proof of (iii).

(iv) This is obvious by regularity of the Fock representation and by

$$\mathscr{U}_h^{-1} \pi_r(W((V^\lambda)_\Phi)) \mathscr{U}_h = \pi_\mu(\omega(\lambda(p([\dot{V}]) \oplus r_h([\dot{V}])))) \otimes \hat{\pi}_{[0]}(\hat{W}(\lambda[\dot{V} - \tilde{\dot{V}}(0,.)h])). \qquad \square$$

Remarks: (i) Theorems 4.4, 6.3, and 6.4 together characterize a class of regular, irreducible representations of the algebra $\mathscr{F}$ satisfying Borchers' criterion. When restricted to CCR$\cap\mathscr{F}_0$ the representations decompose into direct integral of coherent state representations and in this respect resemble the scattering representations considered in Ref. 15. However, here the infrared clouds are independent of the charged particles (they are there even if there are no such particles present). In particular, the arguments of this reference for the Lorentz symmetry breaking do not apply here.

(ii) The vacuum vector is replaced here by ''infravacua,'' and states with finite charged particle number are obtained by the action of creation operators on any such state. (The ''infravacua'' are not of the KPR type,[28] which does not lead to coherent states.) There are no vectors with the energy-momentum on mass hyperboloid, but the arguments of Ref. 14 do not apply here either: the asymptotic of electromagnetic field according to Buchholz catches only the free field part, and does not characterize states by classical distribution of electric flux.

(iii) The operators $\pi(W(V))$ with $l\wedge V(-\infty,l)\neq 0$ representing the exponentials of total electromagnetic field do not commute with $\pi(B(g))$, which reflects the fact that creation or annihilation of a charged particle together with its Coulomb field is a nonlocal operation. They do not commute with nonlocal observables $\pi(B(g)^*B(f))$ either.

A particular representation in the class thus characterized is given whenever a measure μ is chosen, such that the condition (6.10) is satisfied and the representation π_μ is regular and irreducible. Explicit characterization of such measures may be given in the subclass of Gaussian measures. For any positive, trace-class operator B in the Hilbert space $\mathscr{H}_{IR}$ the characteristic function

$$\int e^{i([f],[g])}\, d\mu_B([f]) = e^{-(1/2)([g],B[g])}$$

defines a cylindrical, σ-additive measure, a Gaussian measure with covariance B.[29] The following proposition is obtained by the application of general standard results.

Proposition 6.5: (i) A Gaussian measure with covariance B satisfies the quasi-invariance condition (6.10) iff

$$\mathscr{C}^\infty_{IR} \subset B^{1/2}.\mathscr{H}_{IR} \tag{6.20}$$

(*hence* Ker $B^{1/2}=\{0\}$, *as* $\overline{\mathscr{C}^\infty_{IR}}^{\mathscr{H}_{IR}} = \mathscr{H}_{IR}$). *Then the representation* π_μ *is regular and it is unitarily equivalent to the GNS representation of the quasi-free state*

$$\omega_\mu(w([g]\oplus[k])) = e^{-(1/4)s([g]\oplus[k],[g]\oplus[k])}, \tag{6.21}$$

where s is a bilinear, positive definite form on $\mathscr{C}^\infty_{IR}\oplus\mathscr{C}^\infty_{IR}$ *satisfying the defining condition of a quasi-free state*

$$|\{[g_1]\oplus[k_1],[g_2]\oplus[k_2]\}_{IR}|^2 \leqslant s([g_1]\oplus[k_1],[g_1]\oplus[k_1])\, s([g_2]\oplus[k_2],[g_2]\oplus[k_2]) \tag{6.22}$$

and given explicitly by

$$s([g_1]\oplus[k_1],[g_2]\oplus[k_2])=\tfrac{1}{2}(B^{-1/2}[g_1],B^{-1/2}[g_2])+2(B^{1/2}[k_1],B^{1/2}[k_2]). \tag{6.23}$$

(ii) If (6.20) is satisfied, then π_μ is irreducible iff $\overline{B^{-1/2}\mathscr{C}^\infty_{IR}}^{\mathscr{H}_{IR}}=\mathscr{H}_{IR}$. The state (6.21) then yields a Fock representation. Otherwise π_μ is nonfactor.

Proof: (i) A Gaussian measure on a Hilbert space $\mathscr{H}_{IR}$ and with covariance B is equivalent to its translation by an element $[g]\in\mathscr{H}_{IR}$ if, and only if, $[g]\in B^{1/2}\mathscr{H}_{IR}$ (see Ref. 29), which proves the first statement of (i). If this is the case, then

$$\frac{d\mu_{[g]}}{d\mu}([f])=e^{-(1/2)\|B^{1/2}[g]\|^2+([f],B^{-1}[g])},$$

where $([f],B^{-1}[g])$ is to be understood in the sense of a measurable linear functional on $\mathscr{H}_{IR}$.[29] As the vector Ω_μ is cyclic for π_μ, this representation is unitarily equivalent to the GNS representation obtained from the state $\omega_\mu(A):=(\Omega_\mu,\pi_\mu(A)\Omega_\mu)$. The relations (6.21)–(6.23) are now verified by calculation. By the results of Ref. 30 ω_μ is a quasi-free state, which always yields a regular representation.

(ii) By the results of Ref. 30 the state ω_μ is primary iff the extension $\{.,.\}^s_{IR}$ of the symplectic form $\{.,.\}_{IR}$ to the completion of $\mathscr{C}^\infty_{IR}\oplus\mathscr{C}^\infty_{IR}$ with respect to s is nondegenerate. This completed space is a real Hilbert space $\mathscr{H}_s$. In our case $\mathscr{H}_s=\mathscr{H}_-\oplus\mathscr{H}_+$, where $\mathscr{H}_-$ (resp. $\mathscr{H}_+$) is the completion of $\mathscr{C}^\infty_{IR}$ with respect to the norm $\|[f]\|_-:=\|B^{-1/2}[f]\|$ (resp. $\|[f]\|_+:=\|B^{1/2}[f]\|$), and s is extended to $(x_1\oplus y_1,x_2\oplus y_2)=\frac{1}{2}(x_1,x_2)_-+2(y_1,y_2)_+$. The linear operator U_- (resp. U_+) on $\mathscr{C}^\infty_{IR}$ defined by $U_-[f]:=B^{-1/2}[f]$ (resp. by $U_+[f]:=B^{1/2}[f]$) maps $\mathscr{C}^\infty_{IR}$ as a subspace of $\mathscr{H}_-$ (resp. of $\mathscr{H}_+$) isometrically into $\mathscr{H}_{IR}$. By continuity $U_\mp$ extend to isometric operators $U_\mp:\mathscr{H}_\mp\to\mathscr{H}_{IR}$. By restricting arguments in the following equation to smooth elements one sees that the extension of $\{.,.\}_{IR}$ to $\{.,.\}^s_{IR}$ is given by $\{x_1\oplus y_1,x_2\oplus y_2\}^s_{IR}=(U_-x_1,U_+y_2)_{IR}-(U_-x_2,U_+y_1)_{IR}$. This form is nondegenerate iff $U_-\mathscr{H}_-=U_+\mathscr{H}_+$. However, $U_+\mathscr{H}_+=\overline{B^{1/2}\mathscr{C}^\infty_{IR}}^{\mathscr{H}_{IR}}=\mathscr{H}_{IR}$ (as Ker $B^{1/2}=\{0\}$). Therefore, ω_μ is primary iff $U_-\mathscr{H}_-\equiv\overline{B^{-1/2}\mathscr{C}^\infty_{IR}}^{\mathscr{H}_{IR}}=\mathscr{H}_{IR}$. If this is the case, then it is checked by direct calculation that $(x_1\oplus y_1,x_2\oplus y_2)_s=\{x_1\oplus y_1,A(x_2\oplus y_2)\}^s_{IR}$, where $A(x\oplus y):=(-2U_-^{-1}U_+y)\oplus(\frac{1}{2}U_+^{-1}U_-x)$. The operator A satisfies the equation $A^2=-1$, which is a necessary and sufficient condition for ω_μ to be pure.[30] □

Concrete examples of trace-class operators B satisfying the conditions of Proposition 6.5 (i) and (ii) are most easily constructed in the unitarily equivalent version of the space $\mathscr{H}_{IR}$, the Hilbert space $\mathscr{H}_{\partial^2}$ discussed in the Appendix. Let $\widetilde{B}_\alpha=[((t\cdot l)^2\partial^2)^{-1}]^{1+\alpha}$, $\alpha>0$, where $((t\cdot l)^2\partial^2)^{-1}$ is the positive operator on $\mathscr{H}_{\partial^2}$ defined in Appendix [Eq. (A10) and the following discussion]. It follows from the spectral properties of this operator that each of the operators $\widetilde{B}_\alpha$ may serve as an example of the transformed covariance operator.

ACKNOWLEDGMENTS

I am grateful to Professor Andrzej Staruszkiewicz for discussions and support, to Professor Detlev Buchholz for helpful correspondence and to Walter Kunhardt for discussions and critical reading of an early version of the manuscript.

APPENDIX: HOMOGENEOUS FUNCTIONS ON THE LIGHTCONE

In this Appendix we briefly discuss some structures and operations on spaces of homogeneous functions on the future lightcone. Let, first, $\phi(l)$, $\phi_1(l)$, and $\phi_2(l)$ be smooth ($\mathscr{C}^\infty$ in the sense of differentiation by $L_{ab}=l_a\partial_b-l_b\partial_a$) functions, homogeneous of degree 0. Take, for the sake of differentiation, extensions of these functions which remain homogeneous in some neighborhood of the cone. Straightforward calculation then gives on the cone

$$L_a{}^bL_{cb}\phi=*L_c{}^b*L_{ab}\phi=-L_{ac}\phi+l_al_c\partial^2\phi, \tag{A1}$$

$$*L_a{}^bL_{cb}\phi=-L_c{}^b*L_{ab}\phi=-*L_{ac}\phi, \tag{A2}$$

$$L_a{}^b\phi_1 L_{cb}\phi_2 = {}^*L_c{}^b\phi_1 {}^*L_{ab}\phi_2 = l_a l_c \partial\phi_1\cdot\partial\phi_2, \tag{A3}$$

where ${}^*L_{ab}=\frac{1}{2}e_{abcd}L^{cd}$ and $\partial^2=\partial_a\partial^a$. As the action of L_{ab} is extension independent (this is the tangent derivative), these formulas give extension-independent meanings of $\partial^2\phi$ and $\partial\phi_1\cdot\partial\phi_2$, which in this form were calculated for a homogeneous (but otherwise arbitrary) extensions. Contracting (A1) and (A3) with $t^a t^c$, where t is any unit future-pointing vector, one obtains

$$\partial^2\phi=(t\cdot l)^{-2}{}^*L_0{}^b{}^*L_{0b}\phi, \tag{A4}$$

$$\partial\phi_1\cdot\partial\phi_2=(t\cdot l)^{-2}{}^*L_0{}^b\phi_1{}^*L_{0b}\phi_2, \tag{A5}$$

where ${}^*L_{0b}=t^a{}^*L_{ab}$. If ψ is any differentiable function, homogeneous of degree -2, then

$$\int L_{ab}\psi(l)d^2l=0. \tag{A6}$$

Using this identity one obtains from (A4) and (A5) by integration by parts (and taking into account ${}^*L_{0b}t\cdot l=0$)

$$\int \phi_1\partial^2\phi_2\ d^2l=-\int \partial\phi_1\cdot\partial\phi_2\ d^2l. \tag{A7}$$

Therefore, $\int\phi\ \partial^2\phi\ d^2l\geqslant 0$ and $\int\phi\ \partial^2\phi\ d^2l=0$ iff $\phi=$const. Thus $\partial^2\phi=0$ iff $\phi=$const. This positivity of ∂^2 is also seen from the identity (A4), which says that ∂^2, when applied to a homogeneous function of degree 0, is the ''orbital angular momentum squared'' in each Minkowski frame.

The action of ∂^2 may be explicitly reversed. For each smooth, homogeneous of degree 0 function ϕ the function $\psi=\partial^2\phi$ is smooth, homogeneous of degree -2, and satisfies $\int\psi\ d^2l=0$. Conversely, if ψ is any function with these properties, then the formula

$$\phi_t(l)=-\frac{1}{4\pi}\int \ln\frac{l\cdot l'}{t\cdot l'}\ \psi(l')d^2l' \tag{A8}$$

gives the unique such smooth function that $\partial^2\phi_t=\psi$, and with the additive constant chosen such that

$$\int \frac{\phi_t(l)}{(t\cdot l)^2}\ d^2l=0. \tag{A9}$$

Smoothness of $\phi_t(l)$ is proved by showing that for $\epsilon\searrow 0$ the functions

$$\phi_{t\epsilon}(l)=-\frac{1}{4\pi}\int \ln\frac{l\cdot l'+\epsilon t\cdot l t\cdot l'}{t\cdot l'}\ \psi(l')\ d^2l'$$

converge uniformly to $\phi_t(l)$, and $L_{ab}\phi_{t\epsilon}(l)$ converge uniformly to

$$-\frac{1}{4\pi}\int \ln\frac{l\cdot l'}{t\cdot l'}\ L'_{ab}\psi(l')d^2l'+\frac{1}{4\pi}\int\frac{l'_a t_b-l'_b t_a}{t\cdot l'}\ \psi(l')d^2l'.$$

Then it remains to show that

$$-\frac{1}{4\pi(t\cdot l)^2}\int \ln\frac{l\cdot l'+\epsilon t\cdot l t\cdot l'}{t\cdot l'}\ {}^*L'_0{}^b{}^*L'_{0b}\psi(l')d^2l'=\frac{2\epsilon+\epsilon^2}{4\pi}\int\left(\frac{t\cdot l'}{l\cdot l'+\epsilon t\cdot l t\cdot l'}\right)^2\psi(l')d^2l'$$

converges to $\psi(l)$ pointwise, which is an easy exercise.

Consider now the linear space $[\mathscr{C}^\infty]$ of equivalence classes $[\phi]=\{\phi'\,|\,\phi'=\phi+\text{const}\}$ of smooth, homogeneous of degree 0, functions ϕ. On this space the operators $[\phi]\to[(t\cdot l)^2\partial^2\phi]$ and $[\phi]\to[((t\cdot l)^2\partial^2)^{-1}\phi_t]$, where ϕ_t is the representant satisfying (A9), and

$$(((t\cdot l)^2\partial^2)^{-1}\phi_t)(l)=-\frac{1}{4\pi}\int \ln\frac{l\cdot l'}{t\cdot l'}\,\frac{\phi_t(l')}{(t\cdot l')^2}\,d^2l', \tag{A10}$$

are bijective, mutually inverse maps. Moreover, it is easily seen that the operator $((t\cdot l)^2\partial^2)^{-1}$ is bounded with respect to the norm of the scalar product on $[\mathscr{C}^\infty]$ defined by

$$([\phi_1],[\phi_2])_t=\int \frac{\phi_{1t}\phi_{2t}(l)}{(t\cdot l)^2}\,d^2l, \tag{A11}$$

that is

$$\|[((t\cdot l)^2\partial^2)^{-1}\phi_t]\|_t\leqslant c\|[\phi_t]\|_t. \tag{A12}$$

The expression (A7) introduces another scalar product on $[\mathscr{C}^\infty]$:

$$([\phi_1],[\phi_2])_{\partial^2}=\int \phi_1\partial^2\phi_2(l)d^2l. \tag{A13}$$

By (A11)–(A13) we have

$$\begin{aligned}\|[\phi]\|_t^2&=([\phi],[((t\cdot l)^2\partial^2)^{-1}\phi_t])_{\partial^2}\leqslant\|[\phi]\|_{\partial^2}\|[((t\cdot l)^2\partial^2)^{-1}\phi_t]\|_{\partial^2}\\&=\|[\phi]\|_{\partial^2}\sqrt{([\phi],[((t\cdot l)^2\partial^2)^{-1}\phi_t])_t}\leqslant\sqrt{c}\|[\phi]\|_{\partial^2}\|[\phi]\|_t,\end{aligned}$$

hence

$$\|[\phi]\|_t\leqslant\sqrt{c}\|[\phi]\|_{\partial^2}, \tag{A14}$$

and

$$\|[((t\cdot l)^2\partial^2)^{-1}\phi_t]\|_{\partial^2}\leqslant c\|[\phi_t]\|_{\partial^2}, \tag{A15}$$

that is, the operator (A10) is bounded with respect to this norm as well. Furthermore, for any sequence $[\phi_n]\in[\mathscr{C}^\infty]$ there is

$$\|[\phi_n]\|_{\partial^2}^2\leqslant\|[\phi_n-\phi_m]\|_{\partial^2}^2+2([\phi_n],[\phi_m])_{\partial^2}\leqslant\|[\phi_n-\phi_m]\|_{\partial^2}^2+2\|[\phi_n]\|_t\|[(t\cdot l)^2\partial^2\phi_m]\|_t.$$

Suppose $\lim_{m,n\to\infty}\|[\phi_m-\phi_n]\|_{\partial^2}=0$ and $\lim_{n\to\infty}\|[\phi_n]\|_t=0$. For $\epsilon>0$ let $\|[\phi_m-\phi_n]\|<\epsilon$ for all $m,n\geqslant N$. Put $m=N$ and let $N'\geqslant N$ be such that for all $n\geqslant N'$ there is $2\|[\phi_n]\|_t\|[(t\cdot l)^2\partial^2\phi_N]\|<\epsilon^2$. Then for all $n\geqslant N'$ there is $\|[\phi_n]\|_{\partial^2}^2<2\epsilon^2$, hence $\lim_{n\to\infty}\|[\phi_n]\|_{\partial^2}=0$. Summing up, the norm $\|.\|_{\partial^2}$ is stronger than $\|.\|_t$ and these norms are compatible.[31] This implies that the Hilbert space $\mathscr{H}_{\partial^2}$ obtained from $[\mathscr{C}^\infty]$ by completion with respect to $\|.\|_{\partial^2}$ is (may be canonically identified with) a subspace of the completion of $[\mathscr{C}^\infty]$ with respect to $\|.\|_t$. The latter space is the Hilbert space L_t^2 of equivalence classes of homogeneous of degree 0, measurable functions modulo constant for which $\|[\phi]\|_t<\infty$. This space does not depend on t when considered as a vector space. All elements of $\mathscr{H}_{\partial^2}$ are therefore equivalence classes $[\phi]$ of such functions.

The operator (A10) extends to a bounded, self-adjoint, positive operator on $\mathscr{H}_{\partial^2}$. In view of the comment following (A7) this is a compact operator with eigenvalues $(j(j+1))^{-1}$ of multiplicity $2j+1$, where $j=1,2,...$, and eigenspaces $\mathscr{H}_j$ contained in $[\mathscr{C}^\infty]$.

With the use of the Hilbert space $\mathscr{H}_{\partial^2}$ the operation of projecting $\mathscr{H}_0$ to $\mathscr{H}_{IR}$ needed in Sec. V may be described more explicitly. Let, first, $f_a(l)$ be a smooth function, homogeneous of degree -1 and orthogonal to l^a, which means that $[f_a]$ is a smooth element of $\mathscr{H}_0$. For every such function there are smooth, homogeneous of degree 0 functions $\phi(l)$ and $\psi(l)$, unique up to additive constants, such that

$$l_af_b(l)-l_bf_a(l)=L_{ab}\phi(l)-{}^*L_{ab}\psi(l). \tag{A16}$$

This is shown most easily with the use of the spinor formalism, which we do not intend to discuss here and refer the reader to Ref. 32, and to Ref. 2 for application to related problems. Within this formalism it is a simple result that $o^{A'}f_{A'A}(l)=(\partial/\partial o^A)\chi(l)$, where o^A is the spinor of the null

vector l^a and $\chi(l)$ is a smooth complex function, homogeneous of degree 0, determined by this equation up to an additive constant. The equivalent form of this equation in the tensor language is ${}^-(l_a f_b - l_b f_a) = {}^-L_{ab}\chi$, where ${}^-G_{ab}$ for an antisymmetric tensor G_{ab} denotes its anti-self-dual part ${}^-G_{ab} = \frac{1}{2}(G_{ab} + i{}^*G_{ab})$. Solving this equation for $l_a f_b - l_b f_a$ one obtains (A16), with $\phi = \mathrm{Re}\,\chi$ and $\psi = \mathrm{Im}\,\chi$. Now, let f_1 and f_2 be two such smooth functions represented as in (A16). Then by (A16) and the first equality in (A3)

$$f_1 \cdot f_2(l) = (t\cdot l)^{-2} t^a t^c (l_a f_{1b} - l_b f_{1a})(l_c f_2^b - l^b f_{2c})$$
$$= (t\cdot l)^{-2}({}^*L_{0b}\phi_1 {}^*L_0{}^b\phi_2 + {}^*L_{0b}\psi_1 {}^*L_0{}^b\psi_2 - {}^*L_{0b}\psi_1 L_0{}^b\phi_2 - L_{0b}\phi_1 {}^*L_0{}^b\psi_2).$$

Now integrate this identity with respect to $d^2 l$, integrate one ${}^*L$ in each term by parts and use (A1) and (A2) to obtain

$$([f_1],[f_2])_0 = ([\phi_1],[\phi_2])_{\partial^2} + ([\psi_1],[\psi_2])_{\partial^2}. \tag{A17}$$

Therefore, the map $(\phi,\psi)\to f$ given by (A16) for smooth elements extends to a unitary map $\mathscr{H}_{\partial^2}\oplus\mathscr{H}_{\partial^2}\to\mathscr{H}_0$. The space $\mathscr{H}_{IR}$ is the image of the subspace $\mathscr{H}_{\partial^2}\oplus 0$ in this map.

[1] A. Herdegen, J. Math. Phys. **37**, 100 (1996).
[2] A. Herdegen, J. Math. Phys. **36**, 4044 (1995).
[3] J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Springer, Berlin, 1976).
[4] G. Morchio and F. Strocchi, ''Infrared problem, Higgs phenomenon and long range interactions,'' in *Fundamental Problems in Gauge Field Theory*, edited by G. Velo and A. S. Wightman (Plenum, New York, 1986).
[5] R. Haag, *Local Quantum Physics* (Springer, Berlin, 1992).
[6] P. P. Kulish and L. D. Faddeev, Teor. Mat. Fiz. **4**, 153 (1970) [Theor. Math. Phys. **4**, 745 (1971)].
[7] D. Zwanziger, Phys. Rev. D **14**, 2570 (1976).
[8] J. Fröhlich, Ann. Inst. Henri Poincaré, Sect. A **19**, 1 (1973).
[9] D. Buchholz, Commun. Math. Phys. **85**, 49 (1982).
[10] D. Buchholz, M. Porrmann, and U. Stein, Phys. Lett. B **267**, 377 (1991).
[11] A. Staruszkiewicz, Ann. Phys. (N.Y.) **190**, 354 (1989); Acta Phys. Pol. B **23**, 591 (1992).
[12] B. D. Bramson, ''Physics in cone space,'' in *Asymptotic Structure of Space-Time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977), pp. 273–359.
[13] A. Ashtekar, *Asymptotic Quantization* (Bibliopolis, Napoli, 1986).
[14] D. Buchholz, Phys. Lett. B **174**, 331 (1986).
[15] J. Fröhlich, G. Morchio, and F. Strocchi, Ann. Phys. (N.Y.) **119**, 241 (1979).
[16] J. Dixmier, *Les C*-algèbres et leurs représentations* (Gauthier-Villars, Paris, 1969); English translation: *C*-Algebras* (North-Holland, Amsterdam, 1977).
[17] J. D. Dollard, J. Math. Phys. **5**, 729 (1964).
[18] M. Flato, J. C. H. Simon, and E. Taflin, Rev. Math. Phys. **6**, 1071 (1994); ''Asymptotic completeness, global existence and the infrared problem for the Maxwell-Dirac equations,'' (1995) preprint, Paper #9502061 in the hep-th electronic archive.
[19] K. Symanzik, *Lectures on Lagrangian Quantum Field Theory*, DESY-report T-71/1.
[20] G. Roepstorff, Commun. Math. Phys. **19**, 301 (1970).
[21] H. J. Borchers, Commun. Math. Phys. **2**, 49 (1966).
[22] J. Dixmier, *Les algèbres d'operateurs dans l 'espace Hilbertien* (Gauthier-Villars, Paris, 1969); English translation: *Von Neumann Algebras* (North-Holland, Amsterdam, 1981).
[23] H. J. Borchers, R. Haag, and B. Schroer, Nuovo Cimento **29**, 148 (1963).
[24] F. Rocca, M. Sirugue, and D. Testard, Commun. Math. Phys. **19**, 119 (1970).
[25] M. Reed and B. Simon, *Methods of Mathematical Physics, Vol. II.: Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).
[26] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, Commun. Math. Phys. **32**, 231 (1973).
[27] I. M. Gel'fand and N. Ya. Vilenkin, *Obobshchennyie Funkcii, vyp. 4*. (GIFML, Moskva, 1961); English translation: *Generalized Functions, Vol. IV* (Academic, New York, 1964).
[28] K. Kraus, L. Polley, and G. Reents, Ann. Inst. Henri Poincaré, Sect. A **26**, 109 (1977).
[29] Yu. L. Daleckij and S. V. Fomin, *Miery i Differencialnye Uravnienia v Beskonietchnomiernych Prostranstvach* (Nauka, Moskva, 1983); English translation: *Measures and Differential Equations in Infinite-Dimensional Space* (Kluwer Academic, Dodrecht, 1991).
[30] J. Manuceau and A. Verbeure, Commun. Math. Phys. **9**, 293 (1968).
[31] I. M. Gel'fand and G. Ye. Shilov, *Obobshchennyie Funkcii, vyp.2*. (GIFML, Moskva, 1958); English translation: *Generalized Functions, Vol. II*. (Academic, New York).
[32] R. Penrose and W. Rindler, *Spinors and Spacetime, Vol. I*. (Cambridge U. P., Cambridge, 1984).
[33] The resulting algebra, and some of the statements in Sec. II may be obtained, as pointed out to me by H. Grundling, by the application of the general theory of twisted crossed products of *C**-algebras by groups, cf.: J. A. Packer and I. Raeburn, Math. Proc. Camb. Philos. Soc. **106**, 293 (1989).

Continuous time and consistent histories

C. J. Isham[a)]
Blackett Laboratory, Imperial College, South Kensington, London SW7 2BZ, United Kingdom

N. Linden[b)]
D.A.M.T.P., University of Cambridge, Cambridge CB3 9EW, United Kingdom

K. Savvidou[c)] and S. Schreckenberg[d)]
Blackett Laboratory, Imperial College, South Kensington, London SW7 2BZ, United Kingdom

(Received 8 December 1997; accepted for publication 6 January 1998)

We discuss the use of histories labelled by a continuous time in the approach to consistent-histories quantum theory in which propositions about the history of the system are represented by projection operators on a Hilbert space. This extends earlier work by two of us [C. J. Isham and N. Linden, J. Math. Phys. **36**, 5392–5408 (1995)] where we showed how a continuous time parameter leads to a history algebra that is isomorphic to the canonical algebra of a quantum *field* theory. We describe how the appropriate representation of the history algebra may be chosen by requiring the existence of projection operators that represent propositions about the time average of the energy. We also show that the history description of quantum mechanics contains an operator corresponding to velocity that is quite distinct from the momentum operator. Finally, the discussion is extended to give a preliminary account of quantum field theory in this approach to the consistent histories formalism. © *1998 American Institute of Physics.* [S0022-2488(98)03304-0]

I. INTRODUCTION

The consistent histories approach to quantum theory can be formulated in several different ways. In the original scheme,[1–3] the crucial object is the decoherence function written as

$$d(\alpha,\beta)=\mathrm{tr}(\widetilde{C}_\alpha^\dagger \rho \widetilde{C}_\beta), \tag{1}$$

where ρ is the initial density-matrix, and where the *class operator* $\widetilde{C}_\alpha$ is defined in terms of the standard Schrödinger-picture projection operators α_{t_i} as

$$\widetilde{C}_\alpha := U(t_0,t_1)\alpha_{t_1}U(t_1,t_2)\alpha_{t_2}\ldots U(t_{n-1},t_n)\alpha_{t_n}U(t_n,t_0), \tag{2}$$

where $U(t,t')=e^{-i(t-t')H/\hbar}$ is the unitary time-evolution operator from time t to t'. Each projection operator α_{t_i} represents a proposition about the system at time t_i, and the class operator $\widetilde{C}_\alpha$ represents the composite history proposition ''α_{t_1} is true at time t_1, and then α_{t_2} is true at time t_2, and then $\ldots$, and then α_{t_n} is true at time t_n.''

As a product of (generically, non-commuting) projection operators, the class-operator $\widetilde{C}_\alpha$ is not itself a projector. This difference between the representation of propositions in standard quantum mechanics and in the history theory is avoided in the alternative approach[4,5] to the latter in which the history proposition ''α_{t_1} is true at time t_1, and then α_{t_2} is true at time t_2, and

a)Electronic mail: c.isham@ic.ac.uk
b)Electronic mail: n.linden@newton.cam.ac.uk
c)Electronic mail: k.savvidou@ic.ac.uk
d)Electronic mail: s.schreckenberg@ic.ac.uk

 © 1998 American Institute of Physics

then . . . , and then α_{t_n} is true at time t_n'' is represented by the *tensor product* $\alpha_{t_1}\otimes\alpha_{t_2}\otimes\cdots\otimes\alpha_{t_n}$ which, unlike $\widetilde{C}_\alpha$, *is* a genuine projection operator. In this 'history projection operator' (HPO) scheme, the decoherence function can be written as

$$d(\alpha,\beta)=\mathrm{tr}_{\mathscr{V}\otimes\mathscr{V}}(\alpha\otimes\beta X) \tag{3}$$

for a suitable operator X (independent of α and β) defined on $\mathscr{V}\otimes\mathscr{V}$.[6] Here, $\mathscr{V}$ denotes the 'history space' $\mathscr{H}_{t_1}\otimes\mathscr{H}_{t_2}\otimes\cdots\otimes\mathscr{H}_{t_n}$ on which the history-proposition projectors α and β are defined, where, for each t_i, $\mathscr{H}_{t_i}$ is a copy of the Hilbert space $\mathscr{H}$ of the standard quantum theory. The representation of history propositions with projection operators has clear advantages when discussing logical operations; in our case, quantum temporal logic. It could also be applicable in quantum gravity situations where there is no background concept of time and where, therefore, the construction of a class operator is particularly problematic: in this case, history propositions could still be represented by projection operators, but on a Hilbert space that does not arise as a temporal tensor product.

The introduction of a *continuous* time clearly poses difficulties for both approaches to the consistent history theory: in the class-operator scheme one has to define continuous products of projection operators; in the HPO approach, the problem is to define a continuous *tensor* product of projection operators.

In an earlier paper by two of us,[7] the latter problem was tackled by exploiting the well-known existence of continuous tensor products of coherent states. However, several interesting issues were sidestepped in the process. For example, the natural history propositions in this scheme are represented by continuous tensor products of projectors onto coherent states, and these do not have a transparent physical interpretation. On the other hand, the formalism as given was not well-equipped to handle the more physically-motivated history propositions about continuous time averages.

In the present paper we take a fresh look at the question of continuous time in the HPO formalism. As in our earlier work, the starting point is the *history group*: a history analogue of the canonical group used in standard quantum mechanics. The key idea is that a unitary representation of the history group leads to a self-adjoint representation of its Lie algebra, the spectral projectors of which are to be interpreted as propositions about the histories of the system. Thus we employ a history group whose associated projection operators represent propositions about continuous-time histories. It transpires that the history algebra for one-dimensional quantum mechanics is *infinite* dimensional—in fact, it is isomorphic to the canonical commutation algebra of a standard quantum *field* theory in one spatial dimension. This suggests that it might be profitable to study the history theory using tools that are normally employed in quantum field theory. This we shall do; in particular, we show that the physically appropriate representation of the history algebra can be selected by requiring the existence of operators that represent propositions about the time-averaged values of the energy. The Fock space thus constructed is related to the notion of a continuous tensor product as used in our earlier paper, thus establishing the link with the idea of continuous temporal logic. We also introduce the continuous-time history analogue of the Heisenberg picture, and we discuss the role of velocity in the history theory. Finally, the discussion is extended to include the case of a free relativistic quantum field.

In what follows we have deliberately adopted a 'physicist's approach' to the analytical problems that arise in the theory; for example, we frequently use unsmeared commutation relations, and domains of operators are not discussed. This enables us to present the essential physical ideas without getting lost in mathematical detail. However, nothing of real importance is hidden thereby since only Fock space representations of the history algebra are used, and the full mathematical theory of these is well-known from normal quantum field theory and poses no major problems.

II. THE HISTORY SPACE

A. The history group

We start by considering the HPO version of the quantum theory of a particle moving on the real line $\mathbb{R}$. As explained above, the history proposition ''α_{t_1} is true at time t_1, and then α_{t_2} is true

at time t_2, and then ... , and then α_{t_n} is true at time t_n'' is represented by the projection operator $\alpha_{t_1}\otimes\alpha_{t_2}\otimes\cdots\otimes\alpha_{t_n}$ on the n-fold tensor product $\mathscr{V}_n=\mathscr{H}_{t_1}\otimes\mathscr{H}_{t_2}\otimes\cdots\otimes\mathscr{H}_{t_n}$ of n-copies of the Hilbert-space $\mathscr{H}$ of the canonical theory. Since $\mathscr{H}$ carries a representation of the Heisenberg–Weyl group with Lie algebra

$$[x,p]=i\hbar, \tag{4}$$

the Hilbert space $\mathscr{V}_n$ carries a unitary representation of the n-fold product group whose generators satisfy

$$[x_k,x_m]=0, \tag{5}$$

$$[p_k,p_m]=0, \tag{6}$$

$$[x_k,p_m]=i\hbar\,\delta_{km}, \tag{7}$$

with $k,m=1,2,\ldots,n$. Thus the Hilbert space $\mathscr{V}_n$ carries a representation of the 'history group' whose Lie algebra is defined to be that of Eqs. (5)–(7). However, we can also turn the argument around and *define* the history version of n-time quantum mechanics by starting with Eqs. (5)–(7). In this approach, $\mathscr{V}_n$ arises as a representation space for Eqs. (5)–(7), and tensor products $\alpha_{t_1}\otimes\alpha_{t_2}\otimes\cdots\otimes\alpha_{t_n}$ that correspond to sequential histories about the values of position or momentum (or linear combinations of them) are then elements of the spectral representations of this Lie algebra.

We shall employ this approach to discuss continuous-time histories. Thus, motivated by Eqs. (5)–(7), we start with the history-group whose Lie algebra (referred to in what follows as the 'canonical history algebra', or CHA for short) is[7]

$$[x_{t_1},x_{t_2}]=0, \tag{8}$$

$$[p_{t_1},p_{t_2}]=0, \tag{9}$$

$$[x_{t_1},p_{t_2}]=i\hbar\,\delta(t_1-t_2), \tag{10}$$

where $-\infty\leqslant t_1,t_2\leqslant\infty$. Note that these operators are in the *Schrödinger* picture: they must not be confused with the Heisenberg-picture operators $x(t),p(t)$ of normal quantum theory.

An important observation is that Eqs. (8)–(10) are mathematically the same as the canonical commutation relations of a quantum *field* theory in one space dimension:

$$[\phi(x_1),\phi(x_2)]=0, \tag{11}$$

$$[\pi(x_1),\pi(x_2)]=0, \tag{12}$$

$$[\phi(x_1),\pi(x_2)]=i\hbar\,\delta(x_1-x_2). \tag{13}$$

This analogy will be exploited fully in the present paper. For example, the following two issues arise immediately. First—to be mathematically well-defined—equations of the type Eqs. (8)–(10) must be smeared with test functions to give

$$[x_f,x_g]=0, \tag{14}$$

$$[p_f,p_g]=0, \tag{15}$$

$$[x_f,p_g]=i\hbar\int_{-\infty}^{\infty}f(t)g(t)dt, \tag{16}$$

which leads at once to the question of which class τ of test functions to use. The minimal requirement for the right hand side of Eq. (16) to make sense is that τ must be a linear subspace of the space $L^2(\mathbb{R},dt)$ of square integrable functions on $\mathbb{R}$. For the moment we shall leave τ unspecified beyond this.

The second issue is concerned with finding the physically appropriate representation of the CHA Eqs. (14)–(16), bearing in mind that infinitely many unitarily inequivalent representations are known to exist in the analogous case of Eqs. (11)–(13). Note that this problem does not arise in standard quantum mechanics, or in the history version of quantum mechanics with propositions defined at a finite number of times, since—by the Stone–von Neumann theorem—there is a unique representation of the corresponding algebra up to unitarily equivalence.

Of course, from the perspective of the history theory the physically appropriate representation is expected to involve some type of continuous tensor product; this was the path followed in our earlier work.[7] On the other hand, in standard quantum field theory there is a folk lore, going back at least to a famous paper by Araki,[8] to the effect that requiring the Hamiltonian to exist as a proper self-adjoint operator is sufficient to select a unique representation; for example, the representations appropriate for a free boson field with different masses are unitarily inequivalent. In our case, this suggests that the appropriate representation of the algebra Eqs. (14)–(16) should be chosen by requiring the existence of operators that represent history propositions about (time-averaged) values of the energy. As we shall see, this is indeed the case.

B. The Hamiltonian algebra

We start with the ubiquitous example of the one-dimensional, simple harmonic oscillator with Hamiltonian

$$H=\frac{p^2}{2m}+\frac{m\omega^2}{2}x^2. \tag{17}$$

The naïve idea behind the HPO theory is that to each time t there is associated a Hilbert space $\mathscr{H}_t$ that carries propositions appropriate to that time (the 'naïvety' refers to the fact that, in a *continuous* tensor product $\otimes_{t\in\mathbb{R}}\mathscr{H}_t$, the individual Hilbert spaces $\mathscr{H}_t$ do not strictly exist as subspaces; this is related to the need to smear operators). Thus we expect to have a one-parameter family of operators

$$H_t:=\frac{p_t^2}{2m}+\frac{m\omega^2}{2}x_t^2 \tag{18}$$

that represent the energy at time t.

As it stands, the right hand side of Eq. (18) is not well-defined, just as in normal canonical quantum field theory it is not possible to define products of field operators at the same spatial point. However, the commutators of H_t with the generators of the CHA can be computed formally as

$$[H_t,x_s]=-\frac{i\hbar}{m}\delta(t-s)p_s, \tag{19}$$

$$[H_t,p_s]=i\hbar m\omega^2\delta(t-s)x_s, \tag{20}$$

$$[H_t,H_s]=0, \tag{21}$$

and are the continuous-time, history analogues of the familiar result in standard quantum theory:

$$[H,x]=-\frac{i\hbar}{m}p, \tag{22}$$

$$[H,p]=i\hbar m\omega^2 x. \tag{23}$$

In standard quantum theory, the spectrum of the Hamiltonian operator can be computed directly from the algebra of Eqs. (22) and (23) augmented with the requirement that the underlying representation of the canonical commutation relations Eq. (4) is irreducible. This suggests that we try to *define* the history theory by requiring the existence of a family of operators H_t that satisfy the relations Eqs. (19)–(21) and where the representation of the canonical history algebra Eqs. (8)–(10) is irreducible. More precisely, we augment the CHA with the algebra (in semi-smeared form)

$$[H(\chi),x_t]=-\frac{i\hbar}{m}\chi(t)p_t, \tag{24}$$

$$[H(\chi),p_t]=i\hbar m\omega^2\chi(t)x_t, \tag{25}$$

$$[H(\chi_1),H(\chi_2)]=0, \tag{26}$$

where $H(\chi)$ is the history energy-operator, time averaged with the function χ; heuristically, $H(\chi)=\int_{-\infty}^{\infty}dt\chi(t)H_t$.

It is useful to integrate these equations in the following sense. *If* self-adjoint operators $H(\chi)$ exist satisfying Eqs. (24)–(26), we can form the unitary operators $e^{iH(\chi)/\hbar}$, and these satisfy

$$e^{iH(\chi)/\hbar}x_te^{-iH(\chi)/\hbar}=\cos[\omega\chi(t)]x_t+\frac{1}{m\omega}\sin[\omega\chi(t)]p_t, \tag{27}$$

$$e^{iH(\chi)/\hbar}p_te^{-iH(\chi)/\hbar}=-m\omega\sin[\omega\chi(t)]x_t+\cos[\omega\chi(t)]p_t. \tag{28}$$

However, it is clear that the right hand sides of Eqs. (27) and (28) defines an *automorphism* of the canonical history algebra Eqs. (8)–(10). Thus the task at hand can be rephrased as that of finding an irreducible representation of the CHA in which these automorphisms are unitarily implementable: the self-adjoint generators of the corresponding unitary operators will then be the desired time-averaged energy operators $H(\chi)$ [strictly speaking, weak continuity is also necessary, but this poses no additional problems in the cases of interest here].

C. The Fock representation

It is natural to contemplate the use of a Fock representation of the CHA since this plays such a central role in the analogue of a free quantum field in one spatial dimension. To this end, we start by defining the 'annihilation operator'

$$b_t:=\sqrt{\frac{m\omega}{2\hbar}}x_t+i\sqrt{\frac{1}{2m\omega\hbar}}p_t \tag{29}$$

in terms of which the CHA (8)–(10) becomes

$$[b_t,b_s]=0, \tag{30}$$

$$[b_t,b_s^\dagger]=\delta(t-s). \tag{31}$$

Note that

$$\hbar\omega b_t^\dagger b_s=\frac{1}{2m}p_tp_s+\frac{m\omega^2}{2}x_tx_s-\frac{\hbar\omega}{2}\delta(t-s), \tag{32}$$

which suggests that there exists an additively renormalized version of the operator H_t in Eq. (18) of the form $\hbar\omega b_t^\dagger b_t$. In turn, this suggests strongly that a Fock space based on Eq. (29) should provide the operators we seek.

To make this explicit we recall that the bosonic Fock space $\mathscr{F}[\mathscr{H}]$ associated with a Hilbert space $\mathscr{H}$ is defined as

$$\mathscr{F}[\mathscr{H}]:=\mathbb{C}\oplus\mathscr{H}\oplus(\mathscr{H}\otimes_S\mathscr{H})\oplus\cdots, \tag{33}$$

where $\mathscr{H}\otimes_S\mathscr{H}$ denotes the symmetrized tensor product of $\mathscr{H}$ with itself. Any unitary operator U on the 'one-particle' space $\mathscr{H}$ gives a unitary operator $\Gamma(U)$ on $\mathscr{F}[\mathscr{H}]$ defined by

$$\Gamma(U):=1\oplus U\oplus(U\otimes U)\oplus\cdots. \tag{34}$$

Furthermore, if $U=e^{iA}$ for some self-adjoint operator A on $\mathscr{H}$, then $\Gamma(U)=e^{id\Gamma(A)}$ where

$$d\Gamma(A):=0\oplus A\oplus(A\otimes 1+1\otimes A)\oplus\cdots. \tag{35}$$

The implications for us of these well-known constructions are as follows. Consider the Fock space $\mathscr{F}[L^2(\mathbb{R},dt)]$ that is associated with the Hilbert space $L^2(\mathbb{R},dt)$ via the annihilation operator b_t defined in Eq. (29); i.e., the space built by acting with (suitably smeared) operators $b_t^\dagger$ on the 'vacuum state' $|0\rangle$ that satisfies $b_t|0\rangle=0$ for all $t\in\mathbb{R}$. The equations (27) and (28) show that, *if* it exists, the operator $e^{iH(\chi)/\hbar}$ acts on the putative annihilation operator b_t as

$$e^{iH(\chi)/\hbar}b_te^{-iH(\chi)/\hbar}=e^{-i\omega\chi(t)}b_t. \tag{36}$$

However, thought of as an action on $L^2(\mathbb{R},dt)$, the operator $U(\chi)$ defined by

$$(U(\chi)\psi)(t):=e^{-i\omega\chi(t)}\psi(t) \tag{37}$$

is unitary for any measurable function χ. Hence, using the result mentioned above, it follows that in this particular Fock representation of the CHA the automorphism on the right hand side of Eq. (36) *is* unitarily implementable, and hence the desired self-adjoint operators exist. Note that $H(\chi)=\hbar\omega d\Gamma(\hat{\chi})$, where the self-adjoint operator $\hat{\chi}$ is defined on $L^2(\mathbb{R},dt)$ as

$$(\hat{\chi}\psi)(t):=\chi(t)\psi(t). \tag{38}$$

In summary, we have shown that the Fock representation of the CHA Eqs. (8)–(10) associated with the annihilation operator b_t of Eq. (29) is such that there exists a family of self-adjoint operators $H(\chi)$ for which the algebra Eqs. (24)–(26) is satisfied. This Fock space is the desired carrier of the history propositions in our theory. Note that, in this case, the natural choice for the test function space $\tau\subseteq L^2(\mathbb{R},dt)$ used in Eqs. (14)–(16) is simply $L^2(\mathbb{R},dt)$ itself.

The position history variable x_t can be written in terms of b_t and $b_t^\dagger$ as

$$x_t=\sqrt{\frac{\hbar}{2m\omega}}(b_t+b_t^\dagger) \tag{39}$$

and has the correlation function

$$\langle 0|x_tx_s|0\rangle=\frac{\hbar}{2m\omega}\,\delta(t-s). \tag{40}$$

Thus the carrier space of our history theory is Gaussian white noise.

Finally, we note that the formalism discussed above can be extended to a wide class of Hamiltonian operators by employing the types of argument used by axiomatic field theorists to construct Euclidean quantum field theories in one space plus one time dimension.[9] In particular, the underlying Gaussian stochastic process of our history theory of the simple harmonic oscillator can be successfully perturbed by the addition to H_t of a wide class of polynomial functions of the configuration variable x_t.

D. The '*n*-particle' history propositions

The Fock space construction produces a natural collection of history propositions: namely, those represented by the projection operators onto what, in a normal quantum field theory, would be called the 'n-particle states.' To see what these correspond to physically in our case we note first that a δ-function normalized basis for $\mathscr{F}[L^2(\mathbb{R},dt)]$ is given by the vectors $|0\rangle$, $|t_1\rangle$, $|t_1,t_2\rangle$,

. . . , where $|t_1\rangle := b^\dagger_{t_1}|0\rangle$, $|t_1,t_2\rangle := b^\dagger_{t_1}b^\dagger_{t_2}|0\rangle$, etc. (of course, properly normalized vectors are of the form $|\phi\rangle := b^\dagger_\phi|0\rangle$, etc., for suitable smearing function ϕ). The physical meaning of the projection operators of the form $|t\rangle\langle t|$ (or, more rigorously, $|\phi\rangle\langle\phi|$), $|t_1,t_2\rangle\langle t_1,t_2|$, etc., can be seen by studying the equations

$$H(\chi)|0\rangle = 0, \tag{41}$$

$$H(\chi)|t\rangle = \hbar\omega\chi(t)|t\rangle, \tag{42}$$

$$H(\chi)|t_1,t_2\rangle = \hbar\omega[\chi(t_1)+\chi(t_2)]|t_1,t_2\rangle, \tag{43}$$

or, in totally unsmeared form,

$$H_t|0\rangle = 0, \tag{44}$$

$$H_t|t_1\rangle = \hbar\omega\,\delta(t-t_1)|t_1\rangle, \tag{45}$$

$$H_t|t_1,t_2\rangle = \hbar\omega[\,\delta(t-t_1)+\delta(t-t_2)]|t_1,t_2\rangle. \tag{46}$$

It is clear from the above that, for example, the projector $|t_1,t_2\rangle\langle t_1,t_2|$ represents the proposition that there is a unit of energy $\hbar\omega$ concentrated at the time point t_1 and another unit concentrated at the time point t_2. Note that $H(\chi)|t,t\rangle = 2\hbar\omega\chi(t)|t,t\rangle$, and hence $|t,t\rangle\langle t,t|$ represents the proposition that there are two units of energy concentrated at the *single* time point t (thus exploiting the bose-structure of the canonical history algebra!). This interpretation of projectors like $|t_1,t_2\rangle\langle t_1,t_2|$ is substantiated by noting that the time-averaged energy obtained by choosing the averaging function χ to be 1 acts on these vectors as

$$\int_{-\infty}^{\infty} ds H_s|t\rangle = \hbar\omega|t\rangle, \tag{47}$$

$$\int_{-\infty}^{\infty} ds H_s|t_1,t_2\rangle = 2\hbar\omega|t_1,t_2\rangle, \tag{48}$$

and so on. This is the way in which the HPO account of the simple harmonic oscillator recovers the integer-spaced energy spectrum of standard quantum theory.

Finally, we note in passing that

$$\frac{1}{\hbar\omega}\int_{-\infty}^{\infty} ds\,s H_s|t_1,t_2,\ldots,t_n\rangle = (t_1+t_2+\cdots+t_n)|t_1,t_2,\ldots,t_n\rangle \tag{49}$$

so that $(1/\hbar\omega)\int_{-\infty}^{\infty} ds\,s H_s$ acts as a 'total time' or 'center-of-time' operator.

E. The Heisenberg picture

It is interesting to investigate the analogue of the Heisenberg picture in our continuous-time HPO theory. In standard quantum theory, the Heisenberg picture version of an operator A is defined with respect to a time origin $t=0$ as

$$A_H(s) := e^{isH/\hbar}Ae^{-isH/\hbar}. \tag{50}$$

In particular, for the simple harmonic oscillator we have

$$x(s) = \cos[\omega s]x + \frac{1}{m\omega}\sin[\omega s]p, \tag{51}$$

$$p(s) = -m\omega\sin[\omega s]x + \cos[\omega s]p. \tag{52}$$

The Heisenberg picture operator $x(s)$ satisfies the classical equation of motion

$$\frac{d^2x(s)}{ds^2}+\omega^2x(s)=0, \tag{53}$$

and the commutator of these operators is

$$[x(s_1),x(s_2)]=\frac{i\hbar}{m\omega}\sin[\omega(s_1-s_2)], \tag{54}$$

which, on using the equation of motion

$$p:=m\left.\frac{dx(s)}{ds}\right|_{s=0}, \tag{55}$$

reproduces the familiar canonical commutation relation Eq. (4).

In trying to repeat this construction for the history theory we might be tempted to define the Heisenberg picture analogue of, say, x_t as

$$x_{H,t}(s):=e^{isH_t/\hbar}x_te^{-isH_t/\hbar}. \tag{56}$$

However, this expression is not well-defined since it corresponds to choosing the test-function in Eq. (27) as $\chi(t'):=s\delta(t-t')$, which leads to ill-defined products of $\delta(t-t')$.

What is naturally suggested instead is to define 'time-averaged' Heisenberg quantities

$$x_{\kappa,t}:=e^{iH(\kappa)/\hbar}x_te^{-iH(\kappa)/\hbar}=\cos[\omega\kappa(t)]x_t+\frac{1}{m\omega}\sin[\omega\kappa(t)]p_t \tag{57}$$

for suitable test functions κ. The analogue of the equation of motion (53) is the functional differential equation

$$\frac{\delta^2x_{\kappa,t}}{\delta\kappa(s_1)\delta\kappa(s_2)}+\delta(t-s_1)\delta(t-s_2)\omega^2x_{\kappa,t}=0, \tag{58}$$

while the history analogue of Eq. (55) is

$$\delta(t-s)p_t=m\left.\frac{\delta x_{\kappa,t}}{\delta\kappa(s)}\right|_{\kappa=0}, \tag{59}$$

and the analogue of the 'covariant commutator' Eq. (54) is

$$[x_{\kappa_1,t_1},x_{\kappa_2,t_2}]=\frac{i\hbar}{m\omega}\delta(t_1-t_2)\sin[\omega(\kappa_1(t_1)-\kappa_2(t_2)], \tag{60}$$

which correctly reproduces the canonical history algebra.

It is worth remarking that we could have proceeded in a slightly different way by starting with a set of operators $x_t(s)$ that satisfy a postulated history version of the covariant commutator Eq. (54):

$$[x_{t_1}(s_1),x_{t_2}(s_2)]=\frac{i\hbar}{m\omega}\delta(t_1-t_2)\sin[\omega(s_1-s_2)] \tag{61}$$

and with the standard Schrödinger canonical history operators then being *defined* as $x_t:=x_t(0)$ and $p_t:=m[dx_t(s)/ds]|_{s=0}$. However, in fact, this is just a special case of the first scheme with the test function κ chosen to be $\kappa(t):=s$ for all t; i.e., the 'Heisenberg picture' operators $x_t(s)$ are generated by the time-averaged energy in the form

$$x_t(s):=e^{is\int drH_r/\hbar}x_te^{-is\int drH_r/\hbar}. \tag{62}$$

In our HPO formalism, the Heisenberg picture operators—unlike those in the Schrödinger picture—have no obvious direct physical interpretation and their main use is likely to be mathematical. Therefore, there is no *a priori* reason for rejecting the simple time-averaged quantity in Eq. (62). What *is* clear however is that—whichever version is used—*two* different time labels appear: the 'external' label t that specifies the time at which a proposition is asserted, and the 'internal' label s that specifies the time parameter in the Heisenberg picture associated with the copy of standard quantum theory on the Hilbert space $\mathscr{H}_t$.

F. The relation with continuous temporal logic

Relating the construction above to the idea of 'continuous temporal logic' involves showing that

$$\otimes_{t\in\mathbb{R}} L^2_t(\mathbb{R},dx) \simeq \mathscr{F}[L^2(\mathbb{R},dt)], \tag{63}$$

where $L^2(\mathbb{R},dx)$ is the Hilbert space of the standard quantum theory of a particle moving in one dimension. For completeness we shall summarize here the discussion of this relation given in our earlier paper.[7]

At a heuristic level, the inner product between two continuous tensor product vectors $|\otimes_{t\in\mathbb{R}} u_t\rangle$ and $|\otimes_{t\in\mathbb{R}} v_t\rangle$ is required to be

$$\langle \otimes_{t\in\mathbb{R}} u_t | \otimes_{t\in\mathbb{R}} v_t\rangle_{\otimes_t \mathscr{H}_t} = \prod_{t\in\mathbb{R}} \langle u_t | v_t\rangle_{\mathscr{H}_t} := \exp\int_{-\infty}^{\infty} dt \log\langle u_t | v_t\rangle_{\mathscr{H}_t}. \tag{64}$$

Since $\langle u_t | v_t\rangle$ is a complex number the logarithm on the right hand side of Eq. (64) is generally not well-defined. However, consider the special case of trying to construct the continuous tensor product of copies of a *Fock* space $\mathscr{F}[K]$ for some Hilbert space K. In particular, consider the *coherent states* $|\exp\phi\rangle$, $|\phi\rangle \in K$, defined as

$$|\exp\phi\rangle := \oplus_{n=0}^{\infty} \frac{1}{n!} (\otimes |\phi\rangle)^n, \tag{65}$$

where $(\otimes|\phi\rangle)^n$ denotes the tensor product of $|\phi\rangle$ with itself n times (and with the convention that $(|\phi\rangle)^0 := 1$). Then

$$\langle \exp\phi | \exp\psi\rangle_{\mathscr{F}[K]} = \exp\langle \phi | \psi\rangle_K \tag{66}$$

and hence, using Eq. (64),

$$\langle \otimes_t \exp\phi_t | \otimes_t \exp\psi_t\rangle_{\otimes_t \mathscr{F}[K]_t} = \exp\int_{-\infty}^{\infty} dt \langle \phi_t | \psi_t\rangle_{K_t} \tag{67}$$

which *is* well-defined. But the exponent on the right hand side of Eq. (67) is just the inner product in the direct integral $\int^{\oplus} K_t dt$ of the Hilbert spaces K_t, and hence we arrive at the basic isomorphism

$$\otimes_{t\in\mathbb{R}} \mathscr{F}[K]_t \simeq \mathscr{F}\left[\int^{\oplus} K_t dt\right], \quad \otimes_t |\exp\phi_t\rangle \mapsto |\exp\phi(\cdot)\rangle. \tag{68}$$

However, the single-time Hilbert space of our theory—$L^2(\mathbb{R},dx)$—can be written as the Fock space for the one-dimensional Hilbert space $\mathbb{C}$ via the isomorphism

$$\mathscr{F}[\mathbb{C}] \simeq L^2(\mathbb{R},dx), \tag{69}$$

$$|\exp z\rangle \mapsto \langle x | \exp z\rangle := (2\pi)^{-1/4} e^{zx - (1/2)z^2 - (1/4)x^2}, \tag{70}$$

where the right hand side involves the familiar coherent states in $L^2(\mathbb{R},dx)$. Thus the isomorphism Eq. (68) becomes

$$\otimes_{t\in\mathbb{R}} L^2_t(\mathbb{R},dx) \simeq \mathscr{F}\left[\int^{\oplus} \mathbb{C}_t dt\right]. \tag{71}$$

But the direct integral $\int^{\oplus}\mathbb{C}_t dt$ is isomorphic to $L^2(\mathbb{R},dt)$ via the map that takes the parametrized family of complex numbers λ_t to the *function* $\lambda(\cdot)$ in $L^2(\mathbb{R},dt)$ [i.e., $\lambda(t):=\lambda_t$]. Hence we arrive at the desired isomorphism in Eq. (63).

G. The extension to three dimensions

The extension of the formalism above to a particle moving in three spatial dimensions appears at first sight to be unproblematic. The analogue of the history algebra Eqs. (8)–(10) is

$$[x^i_{t_1}, x^j_{t_2}] = 0, \tag{72}$$

$$[p^i_{t_1}, p^j_{t_2}] = 0, \tag{73}$$

$$[x^i_{t_1}, p^j_{t_2}] = i\hbar\,\delta^{ij}\delta(t_1 - t_2), \tag{74}$$

$i,j=1,2,3$, while the formal expression Eq. (18) for the energy at time t becomes

$$H_t := \frac{\underline{p}_t\cdot\underline{p}_t}{2m} + \frac{m\omega^2}{2}\underline{x}_t\cdot\underline{x}_t. \tag{75}$$

It is straightforward to generalize the discussion above to this situation and, in particular, to find a Fock representation of Eqs. (72)–(74) in which the rigorous analogues of Eq. (75) exist as *bona fide* self-adjoint operators. However, an interesting issue then arises that has no analogue in one-dimensional quantum theory. Namely, we expect to have angular-momentum operators whose formal expression is

$$L^i_t := \epsilon^i{}_{jk} x^j_t p^k_t \tag{76}$$

and whose commutators can be computed heuristically as

$$[L^i_t, L^j_s] = i\hbar\,\epsilon^{ij}{}_k \delta(t-s) L^k_t. \tag{77}$$

Such operators L^i_t can be constructed rigorously using, for example, the method employed for the energy operators H_t: viz., compute the automorphisms of the canonical history algebra that are formally induced by the angular-momentum operators and then see if these automorphisms can be unitarily implemented in the given Fock representation. However, the interesting observation is that, even if this can be done (which is the case, see below), this does not guarantee *a priori* that the commutators in Eq. (77) will be reproduced: in particular, it is necessary to check directly if a c-number *central extension* is present since we know from other branches of theoretical physics that algebras of the type in Eq. (77) are prone to such anomalies.

An obvious technique for evaluating such a commutator would be to define the angular momentum operators by point-splitting in the form

$$L^i_{t,\epsilon} := i\hbar\,\epsilon^i{}_{jk}(b^j_t)^\dagger b^k_{t+\epsilon} \tag{78}$$

so that the commutator in Eq. (77) is the analogue of an equal-time commutator in standard quantum field theory, and the point splitting is the analogue of spatial point splitting. It is then straightforward to compute the commutators of these point-split operators and take the limit $\epsilon\to 0$. The result is the anticipated algebra Eq. (77).

However, in standard quantum field theory it is known that the limit of the commutator has to be considered at unequal times (i.e., using Heisenberg picture operators), and that there is a subtle relation between the two limits of the times becoming equal and the spatial point splitting tending to zero.[10] Therefore, in order to calculate correctly the commutator in our case it seems appropriate to consider the analogue of an unequal time commutator, namely,

$$[L^i_{\chi,t,\epsilon}, L^j_{0,s,\epsilon}], \tag{79}$$

where

$$L^i_{\chi,t,\epsilon} := i\hbar\,\epsilon^i{}_{jk}(b^j_{\chi,t})^\dagger b^k_{\chi,t+\epsilon}, \tag{80}$$

and where

$$b^k_{\chi,t} := e^{iH(\chi)} b^k_t e^{-iH(\chi)} = e^{-i\omega\chi(t)} b^k_t \tag{81}$$

is a time-averaged Heisenberg picture operator of the type defined earlier.

It is not difficult to show that

$$[L^i_{\chi,t,\epsilon}, L^j_{0,s,\epsilon}] = -\hbar^2 e^{i\omega(\chi(t)-\chi(t+\epsilon))}[\,\delta(t-s+\epsilon)((b^j_t)^\dagger b^i_{t+2\epsilon} - \delta^{ij}(b^m_t)^\dagger b^m_{t+2\epsilon})$$
$$-\delta(t-s-\epsilon)((b^i_{t-\epsilon})^\dagger b^j_{t+\epsilon} - \delta^{ij}(b^m_{t-\epsilon})^\dagger b^m_{t+\epsilon})], \tag{82}$$

and then, by evaluating the matrix element of the commutator in the vacuum state, one sees that there is no central extension in this case. Furthermore, by considering the matrix element of the commutator in general coherent states, one can check that the limits of $\epsilon \to 0$ and $\chi \to 0$ are straightforward, and that as long as the test functions are smooth, the angular momentum generators do indeed satisfy the heuristic commutator Eq. (77) in the limit.

H. The role of the velocity operator

The HPO approach to the consistent histories theory has the striking feature that, formally, there exists an operator that corresponds to propositions about the *velocity* of the system: namely, $\dot{x}_t := (d/dt)\,x_t$. More rigorously, we can adopt the procedure familiar from standard quantum field theory and define

$$\dot{x}_f := -x_{\dot{f}}, \tag{83}$$

which is meaningful provided that (i) the test-function f is differentiable; and (ii) f 'vanishes at infinity' so that the implicit integration by parts used in Eq. (83) is allowed; i.e., heuristically, $x_f = \int_{-\infty}^{\infty} dt\, x_t f_t$.

The rigorous existence of $\dot{x}_t$ depends on the precise choice of test-function space used in the smeared form of the CHA in Eqs. (14)–(16). In the analogous situation in normal quantum field theory, the test-functions are chosen so that the spatial derivatives of the quantum field exist, this being necessary to define the Hamiltonian operator. In our case, the situation is somewhat different since the energy operator H_t [see Eq. (18)] does not depend on $\dot{x}_t$ and hence there is no *a priori* requirement for $\dot{x}_t$ to exist. However, what *is* clear from Eq. (8) is that *if* $\dot{x}_t$ exists, then

$$[x_t, \dot{x}_s] = 0 \tag{84}$$

and hence our theory allows for history propositions that include assertions about the position of the particle and its velocity at the same time; in particular, the velocity $\dot{x}_t$ and momentum p_t are not related. In this context it should be emphasized once more that x_t, $t \in \mathbb{R}$, is a one-parameter family of *Schrödinger* picture operators—it is *not* a Heisenberg picture operator, and the equations of motion do not enter at this level.

The existence of a velocity operator that commutes with position is a striking property of the HPO approach to consistent histories and raises some intriguing questions. For example, a classic paper by Park and Margenau[11] contains an interesting discussion of the uncertainty relations, including a claim that it *is* possible to measure position and momentum simultaneously provided the latter is defined using time-of-flight measurements. The existence in our formalism of the vanishing commutator Eq. (84) throws some new light on this old discussion. Also relevant in this respect is Hartle's discussion of the operational meaning of momentum in a history theory.[12] In particular, he emphasizes that an accurate measurement of momentum requires a long time-of-

flight, whereas—on the other hand—our definition of velocity as the time-derivative of the history variable x_t clearly involves a vanishingly small time interval. Presumably this is the operational difference between momentum and velocity in the HPO approach to consistent histories.

The potential existence of $\dot{x}_t$ also raises the interesting possibility of defining a 'velocity-extended' version of the energy H_t as

$$H_t := \frac{p_t^2}{2m} + \frac{m}{2}(\omega^2 x_t^2 + \lambda \dot{x}_t^2) \tag{85}$$

for some real parameter $\lambda \geqslant 0$. Note that in the one-dimensional quantum field theory analogue in which x_t and p_t are replaced by $\phi(x)$ and $\pi(x)$, respectively, for an appropriate choice of λ the expression in Eq. (85) becomes the usual Hamiltonian density $H(x)$ for a massive scalar field (the $\lambda = 0$ case of Eq. (18) then corresponds to an *ultralocal* quantum field). Guided by this observation, we can try to repeat our earlier analysis and look for a representation of the canonical history algebra in which a suitably smeared version of this new energy exists as a proper operator. The quantum field theory analogue suggests that the appropriate replacement for Eq. (29) is

$$c_t := \left(\frac{m\sqrt{\omega^2 - \lambda D^2}}{2\hbar}\right)^{1/2} x_t + i\left(\frac{1}{2m\hbar\sqrt{\omega^2 - \lambda D^2}}\right)^{1/2} p_t, \tag{86}$$

where D denotes the differential operator d/dt. Note that $-D^2$ is a positive semi-definite operator on $L^2(\mathbb{R}, dt)$, and hence the square-root of $\omega^2 - \lambda D^2$ is well-defined; of course, to make all this rigorous a suitably-smeared form of Eq. (86) should be used.

We note that

$$[c_t, c_s^\dagger] = \delta(t-s) \tag{87}$$

and that, *if* it existed, the smeared form $H(\chi)$ of the velocity-extended Hamiltonian would generate the CHA automorphism

$$e^{iH(\chi)/\hbar} c_t e^{-iH(\chi)/\hbar} = \exp[-i\chi(t)\sqrt{\omega^2 - \lambda D^2}]c_t. \tag{88}$$

For the special case in which χ is a *constant*, the right hand side corresponds to a unitary transformation on the 'one-particle' space $L^2(\mathbb{R}, dt)$, and hence the time-averaged energy $\int_{-\infty}^{\infty} dt H_t$ exists as a genuine operator on the associated Fock space. Of course, the eigenvalues and eigenvectors of this operator are well-known for the quantum field theory analogue of the total Hamiltonian operator $\int dx H(x)$; in the history theory, the associated projection operators correspond to appropriate propositions about the value of the time average of the energy.

Note that the relation between x_t and the new annihilation and creation operators is non-local in time:

$$x_t = \left(\frac{\hbar}{2m\sqrt{\omega^2 - \lambda D^2}}\right)^{1/2} (c_t + c_t^\dagger). \tag{89}$$

In particular, the correlation function is

$$\langle 0 | x_t x_s | 0 \rangle = \left(\frac{\hbar}{2m\sqrt{\omega^2 - \lambda D^2}}\right) \delta(t-s), \tag{90}$$

where the non-local quantity on the right hand side is the Green's function of the elliptic partial differential operator $\sqrt{\omega^2 - \lambda D^2}$. Thus we still have a Gaussian stochastic process but it is 'softer' than the one constructed earlier whose correlation function was Eq. (40). Of course, existence of non-local terms is a common occurrence in normal relativistic quantum field theory, but there the non-locality is in space, not time. It remains to be seen whether the 'velocity-extended' Hamiltonian in Eq. (85) has any real physical application in the consistent histories theory.

III. QUANTUM FIELD THEORY

A. The canonical history algebra

We wish now to extend the discussion to the HPO theory of a free scalar field. Hartle[13] proposed a consistent histories approach to quantum field theory based on path integrals, and Blencowe[14] gave a careful analysis of the use of class operators. However, almost nothing has been said about the HPO scheme in this context, and we shall now briefly present the necessary developments. The resemblance of the history version of quantum mechanics ('field theory in zero spatial dimensions') to a canonical field theory in one spatial dimension suggests that the history version of quantum field theory in three spatial dimensions should resemble canonical quantum field theory in *four* spatial dimensions. We shall see that this expectation is fully justified.

The first step in constructing a HPO version of quantum field theory is to foliate four-dimensional Minkowski space–time with the aid of a timelike vector n^μ that is normalized by $\eta_{\mu\nu}n^\mu n^\nu=1$, where the signature of the Minkowski metric $\eta_{\mu\nu}$ has been chosen as $(+,-,-,-)$. The canonical commutation relations for a standard bosonic quantum field theory (the analogue of Eq. (4)) in three spatial dimensions are

$$[\phi(\underline{x}_1),\phi(\underline{x}_2)]=0, \tag{91}$$

$$[\pi(\underline{x}_1),\pi(\underline{x}_2)]=0, \tag{92}$$

$$[\phi(\underline{x}_1),\pi(\underline{x}_2)]=i\hbar\,\delta^3(\underline{x}_1-\underline{x}_2), \tag{93}$$

where x_1 and x_2 are three-vectors that are spatial with respect to the foliation vector n. In constructing the associated HPO theory we shall assume that the passage from the canonical algebra Eq. (4) to the history algebra Eqs. (8)–(10) is reflected in the field theory case by passing from Eqs. (91)–(93) to

$$[\phi_{t_1}(\underline{x}_1),\phi_{t_2}(\underline{x}_2)]=0, \tag{94}$$

$$[\pi_{t_1}(\underline{x}_1),\pi_{t_2}(\underline{x}_2)]=0, \tag{95}$$

$$[\phi_{t_1}(\underline{x}_1),\pi_{t_2}(\underline{x}_2)]=i\hbar\,\delta(t_1-t_2)\,\delta^3(\underline{x}_1-\underline{x}_2), \tag{96}$$

where, for each $t\in\mathbb{R}$, the fields $\phi_t(x)$ and $\pi_t(x)$ are associated with the spacelike hypersurface (n,t) whose normal vector is n and whose foliation parameter is t; in particular, the three-vector x in $\phi_t(x)$ or $\pi_t(x)$ denotes a vector in this space.

In using this algebra, we have in mind a representation that is some type of continuous tensor product $\otimes_{t\in\mathbb{R}}\mathscr{H}_t$ where each $\mathscr{H}_t$ carries a representation of the standard canonical commutation relations Eqs. (91)–(93) for a scalar field theory associated with the given space–time foliation. However, to emphasize the underlying space–time picture it is convenient to rewrite Eqs. (94)–(96) in terms of four-vectors X and Y as

$$[\phi(X),\phi(Y)]=0, \tag{97}$$

$$[\pi(X),\pi(Y)]=0, \tag{98}$$

$$[\phi(X),\pi(Y)]=i\hbar\,\delta^4(X-Y). \tag{99}$$

In relating these expressions to those in Eqs. (94)–(96) the three-vector x may be equated with a four-vector x_n that satisfies $n\cdot x_n=0$ (the dot product is taken with respect to the Minkowski metric $\eta_{\mu\nu}$) so that the pair $(t,x)\in\mathbb{R}\times\mathbb{R}^3$ is associated with the space–time point $X=tn+x_n$ (in particular, $t=n\cdot X$). Note, however, that the covariant-looking nature of these expressions is deceptive and it is not correct to assume *a priori* that the fields $\phi(X)$ and $\pi(Y)$ transform as space–time scalars under the action of some 'external' space–time Poincaré group that acts on the X and Y labels—as things stand there is an implicit n label on both ϕ and π. We shall return to this question later.

B. The Hamiltonian algebra

The key idea of our HPO approach to quantum field theory is that the physically-relevant representation of the canonical history algebra Eqs. (94)–(96) [or, equivalently, Eqs. (97)–(99)] is to be selected by requiring the existence of operators that represent history propositions about temporal averages of the energy defined with respect to the chosen space–time foliation. Thus, for a fixed foliation vector n, we seek a family of 'internal' Hamiltonians $H_{n,t}$, $t\in\mathbb{R}$, whose explicit formal form (i.e., the analogue of Eq. (18)) can be deduced from the standard quantum field theory expression to be

$$H_{n,t}:=\frac{1}{2}\int d^4X\{\pi(X)^2+(n^\mu n^\nu-\eta^{\mu\nu})\partial_\mu\phi(X)\partial_\nu\phi(X)+m^2\phi(X)^2\}\delta(t-n\cdot X). \quad (100)$$

The analogous, temporally-averaged object is

$$\begin{aligned} H_n(\chi):&=\int_{-\infty}^{\infty}dt\chi(t)H_{n,t}\\ &=\frac{1}{2}\int d^4X\{\pi(X)^2+(n^\mu n^\nu-\eta^{\mu\nu})\partial_\mu\phi(X)\partial_\nu\phi(X)+m^2\phi(X)^2\}\chi(n\cdot X) \end{aligned} \quad (101)$$

where χ is a real-valued test function.

As in the discussion above of the simple harmonic oscillator, the next step is to consider the commutator algebra that would be satisfied by the operators $H_n(\chi)$ *if* they existed. These field-theoretic analogues of Eqs. (24)–(26) are readily computed as

$$[H_n(\chi),\phi(X)]=-i\hbar\chi(n\cdot X)\pi(X), \quad (102)$$

$$[H_n(\chi),\pi(X)]=i\hbar\chi(n\cdot X)K_n\phi(X), \quad (103)$$

$$[H_n(\chi_1),H_n(\chi_2)]=0, \quad (104)$$

where K_n denotes the partial differential operator

$$(K_nf)(X):=[(\eta^{\mu\nu}-n^\mu n^\nu)\partial_\mu\partial_\nu+m^2]f(X). \quad (105)$$

The exponentiated form of Eqs. (102)–(103) is

$$e^{iH_n(\chi)/\hbar}\phi(X)e^{-iH_n(\chi)/\hbar}=\cos[\chi(n\cdot X)\sqrt{K_n}]\phi(X)+\frac{1}{\sqrt{K_n}}\sin[\chi(n\cdot X)\sqrt{K_n}]\pi(X), \quad (106)$$

$$e^{iH_n(\chi)/\hbar}\pi(X)e^{-iH_n(\chi)/\hbar}=-\sqrt{K_n}\sin[\chi(n\cdot X)\sqrt{K_n}]\phi(X)+\cos[\chi(n\cdot X)\sqrt{K_n}]\pi(X), \quad (107)$$

where the square-root operator $\sqrt{K_n}$, and functions thereof, can be defined rigorously using the spectral theory of the self-adjoint, partial differential operator K_n on the Hilbert space $L^2(\mathbb{R}^4,d^4X)$. Note that the expression $\chi(n\cdot X)\sqrt{K_n}$ is unambiguous since, viewed as an operator on $L^2(\mathbb{R}^4,d^4X)$, multiplication by $\chi(n\cdot X)$ commutes with K_n.

C. The Fock space representation

The right hand side of Eqs. (106) and (107) defines an automorphism of the CHA Eqs. (97)–(99) and the task is to find a representation of the latter in which these automorphisms are unitarily implemented. To this end, define new operators

$$q(X):=K_n^{1/4}\phi(X), \quad (108)$$

$$p(X):=K_n^{-1/4}\pi(X), \quad (109)$$

and

$$b(X) := \frac{1}{\sqrt{2}}(q(X)+ip(X)) = \frac{1}{\sqrt{2}}(K_n^{1/4}\phi(X)+iK_n^{-1/4}\pi(X)), \tag{110}$$

which satisfy

$$[b(X),b(Y)]=0, \tag{111}$$

$$[b^\dagger(X),b^\dagger(Y)]=0, \tag{112}$$

$$[b(X),b^\dagger(Y)]=\hbar\,\delta^4(X-Y). \tag{113}$$

Then

$$e^{iH_n(\chi)/\hbar}q(X)e^{-iH_n(\chi)/\hbar} = \cos[\chi(n\cdot X)\sqrt{K_n}]q(X)+\sin[\chi(n\cdot X)\sqrt{K_n}]p(X), \tag{114}$$

$$e^{iH_n(\chi)/\hbar}p(X)e^{-iH_n(\chi)/\hbar} = -\sin[\chi(n\cdot X)\sqrt{K_n}]q(X)+\cos[\chi(n\cdot X)\sqrt{K_n}]p(X), \tag{115}$$

and so

$$e^{iH_n(\chi)/\hbar}b(X)e^{-iH_n(\chi)/\hbar} = e^{-i\chi(n\cdot X)\sqrt{K_n}}b(X). \tag{116}$$

However, the operator defined on $L^2(\mathbb{R}^4)$ by

$$(U(\chi)\psi)(X) := e^{-i\chi(n\cdot X)\sqrt{K_n}}\psi(X) \tag{117}$$

is unitary, and hence—using the same type of argument invoked earlier for the simple harmonic oscillator—we conclude that the desired quantities $H_n(\chi)$ exist as self-adjoint operators on the Fock space $\mathscr{F}[L^2(\mathbb{R}^4,d^4X)]$ associated with the creation and annihilation operators $b^\dagger(X)$ and $b(X)$. The spectral projectors of these operators then represent propositions about the time-averaged value of the energy in the spacetime foliation determined by n.

D. The question of external Lorentz invariance

An important part of standard quantum field theory is a proof of invariance under the Poincaré group—something that, in the canonical formalism, is not totally trivial since the Schrödinger picture fields depend on the reference frame (i.e., the space–time foliation). The key ingredient is a construction of the generators of the Poincaré group as explicit functions of the canonical field variables; in practice, the first step is often to construct the Heisenberg picture fields with the aid of the Hamiltonian, and then to demonstrate manifest Poincaré covariance within that framework. The canonical fields associated with any spacelike surface in a particular Lorentz frame can then be obtained by restricting the Heisenberg fields (and their normal derivatives) to the surface.

When considering the role of the Poincaré group in the HPO picture of consistent histories, the starting point is the observation that, heuristically speaking, for a given foliation vector n—and for each value of the associated time t—there will be a Hilbert space $\mathscr{H}_t$ carrying an independent copy of the standard quantum field theory. In particular, therefore, for fixed n, there will be a representation of the Poincaré group associated with each spacelike slice (n,t), $t\in\mathbb{R}$. Thus if A_a, $a=1,2,\ldots,10$, denote the generators of the Poincaré group, there should exist a family of operators A_t^a which, for each $t\in\mathbb{R}$, generate the '*internal*' Poincaré group $\mathscr{P}_{n,t}$ associated with the slice (n,t). These operators will satisfy a 'temporally gauged' version of the Poincaré algebra. More precisely, if $C^{ab}{}_c$ are the structure constants of the Poincaré group, so that

$$[A^a,A^b]=iC^{ab}{}_cA^c, \tag{118}$$

then the algebra satisfied by the history theory operators A_t^a is

$$[A_t^a,A_s^b]=i\,\delta(t-s)C^{ab}{}_cA_t^c, \tag{119}$$

which, of course, reflects the way in which the canonical commutation relations Eqs. (91)–(93) are replaced by Eqs. (94)–(96) in the history theory.

As always in quantum theory, the energy operator is of particular importance, and in the present case we have a family of Hamiltonian operators $H_{n,t}$, $t\in\mathbb{R}$, which are related to the generators $P^{\mu}_{n,t}$ of translations for the quantum field theory associated with the hypersurface (n,t) by

$$H_{n,t}=n_{\mu}P^{\mu}_{n,t}. \tag{120}$$

In fact, it is straightforward to show that

$$P^{\mu}_{n,t}=n^{\mu}H_{n,t}+\int d^4X\,\delta(t-n\cdot X)(n^{\mu}n\cdot\partial\phi-\partial^{\mu}\phi)\pi, \tag{121}$$

which suggests that, as would be expected, the components of $P^{\mu}_{n,t}$ normal to n are the generators of spatial translations in the hypersurface (n,t). Indeed, Eq. (102) generalizes to

$$[P^{\mu}_{n}(\chi),\phi(X)]=-i\hbar\chi(n\cdot X)\{n_{\mu}\pi(X)+(\partial_{\mu}\phi(X)-n_{\mu}n\cdot\partial\phi(X))\}. \tag{122}$$

Similarly, the 'temporally gauged' Lorentz generators satisfy

$$[J^{\mu\nu}_{n,t},\phi(X)]=i\hbar\,\delta(t-n\cdot X)\{X^{\mu}(\partial^{\nu}\phi-n^{\nu}n\cdot\partial\phi)-X^{\nu}(\partial^{\mu}\phi-n^{\mu}n\cdot\partial\phi)-(X^{\mu}n^{\nu}-X^{\nu}n^{\mu})\pi\}. \tag{123}$$

As emphasized above, each generator of the group $\mathscr{P}_{n,t}$ acts 'internally' in the Hilbert space $\mathscr{H}_t$; in particular, this is true of the Hamiltonian, which (modulo the need to smear in t) generates translations along an 'internal' time label s that is to be associated with each leaf (n,t) of the foliation. It is important to note that $H_{n,t}$ does *not* generate translations along the 'exernal' time parameter t that appears in the CHA Eqs. (94)–(96) and which labels the spacelike surface (of course, there is an analogous statement for the Hamiltonians H_t in the HPO model of the simple harmonic oscillator considered earlier). The existence of these internal Poincaré groups is sufficient to guarantee covariance of physical quantities, such as transition amplitudes, that can be calculated in the class operator version of the theory.

However, the HPO formalism admits an additional type of Poincaré group—what we shall call the '*external*' Poincaré group—which is defined to act on the pair of labels (x,t) that appear in the CHA Eqs. (94)–(96). Thus these labels include the 'external' time parameter t that specifies the leaf (n,t) of the foliation associated with the timelike vector n. In the context of the covariant-looking version Eqs. (97)–(99) of the CHA, the main question is whether the fields $\phi(X)$ and $\pi(X)$ transform in a covariant way under this external group.

As far as the field $\phi(X)$ is concerned it seems reasonable to consider the possibility that this may be an external scalar in the sense that there exists a unitary representation $U(\Lambda)$ of the external Lorentz group $U(\Lambda)$ such that

$$U(\Lambda)\phi(X)U(\Lambda)^{-1}=\phi(\Lambda X). \tag{124}$$

The spectral projectors of the (suitably smeared) operators $\phi(X)$ then represent propositions about the values of the space–time field in a covariant way.

However, the situation for the field momentum $\pi(X)$ is different since this is intrinsically associated with the timelike vector n. Indeed, the natural thing would be to require the existence of a *family* of operators $\pi_n(X)$ where n lies in the hyperboloid of all timelike (future-pointing) vectors, and such that

$$U(\Lambda)\pi_n(X)U(\Lambda)^{-1}=\pi_{\Lambda n}(\Lambda X). \tag{125}$$

The next step in demonstrating external Poincaré covariance would be to extend the algebra (97)–(99) to include the n parameter on the π field; in particular, one would need to specify the commutator $[\pi_n(X),\pi_m(Y)]$, but it is not obvious *a priori* what this should be.

Another possibility would be to try to combine the Heisenberg picture—and its associated 'internal' time s—with the external time parameter t of the space–time foliation to give some scheme that was manifestly covariant in the context of a five-dimensional space with signature $(+++,--)$ associated with the variables (x,t,s). However, we do not know if this is possible and the demonstration of external Poincaré covariance, if it exists, remains the subject for future research.

IV. CONCLUSIONS

We have discussed the introduction of continuous-time histories within the 'HPO' version of the consistent histories formalism in which propositions about histories of the system are represented by projection operators on a 'history' Hilbert space. The history algebra (whose representations specify this space) for a particle moving in one dimension is isomorphic to the canonical commutation relations for a one-dimensional quantum field theory, thus allowing the history theory to be studied using techniques drawn from quantum field theory. In particular, we have shown how the problem of the existence of infinitely many inequivalent representations of the history algebra can be solved by requiring the existence of operators whose spectral projectors represent propositions about time-averages of the energy.

We have shown how the Heisenberg picture is changed in the HPO formalism in such a way that the familiar, partial-differential equations of motion are replaced by functional differential equations; these operators are used in the proof that the angular momentum operators of the three-dimensional theory are anomaly free. The question of potential anomalies in the history algebra is rather intriguing and it would be interesting to study a theory in which such things might be expected, such as a fermionic system.

A striking property of the HPO formalism is the potential existence of a velocity operator that commutes with the position operators, thus opening up a new perspective on the old debate about the operational meaning of the Heisenberg uncertainty relations between position and momentum. The introduction of the velocity operator suggests a number of topics for future work: for example, it would be most interesting to see if some thing like the action functional of the classical action has a natural role to play in the HPO theory.

Finally, we have shown how the HPO scheme can be extended to the history version of canonical quantum field theory. We discussed the difference between the 'internal' and 'external' Poincaré groups and indicated how the former are implemented in the formalism. A major challenge for future research is to construct an HPO quantum field theory which is manifestly covariant under this external symmetry group.

ACKNOWLEDGMENTS

Noah Linden is grateful to the Leverhulme and Newton Trusts for their continued support. Dina Savvidou thanks the NATO Science Fellowship Program for their support.

[1] R. B. Griffiths ''Consistent histories and the interpretation of quantum mechanics,''J. Stat. Phys. **36**, 219–272 (1984).
[2] R. Omnès, ''Logical reformulation of quantum mechanics. I. Foundations,'' J. Stat. Phys. **53**, 893–932 (1988).
[3] M. Gell-Mann and J. Hartle, ''Quantum mechanics in the light of quantum cosmology,'' in *Complexity, Entropy and the Physics of Information, SFI Studies in the Science of Complexity, Vol. VIII*, edited by W. Zurek (Addison-Wesley, Reading, MA, 1990), pp. 425–458.
[4] C. J. Isham ''Quantum logic and the histories approach to quantum theory,'' J. Math. Phys. **35**, 2157–2185 (1994).
[5] C. J. Isham and N. Linden, ''Quantum temporal logic and decoherence functionals in the histories approach to generalized quantum theory,'' J. Math. Phys. **35**, 5452–5476 (1994).
[6] C. J. Isham, N. Linden, and S. Schreckenberg, ''The classification of decoherence functionals: an analogue of Gleason's theorem,'' J. Math. Phys. **35**, 6360–6370 (1994).
[7] C. J. Isham and N. Linden, ''Continuous histories and the history group in generalized quantum theory,'' J. Math. Phys. **36**, 5392–5408 (1995).
[8] H. Araki, ''Hamiltonian formalism and the canonical commutation relations in quantum field theory,'' J. Math. Phys. **1**, 492–504 (1960).
[9] B. Simon, *The $P(\phi)_2$ Euclidean (Quantum) Field Theory* (Princeton U. P., Princeton, NJ, 1974).
[10] R. Jackiw, ''Field theoretic investigations in current algebra,'' in *Current Algebra and Anomalies*, edited by S. B. Treiman, R. Jackiw, B. Zumino, and E. Witten (World Scientific, Singapore, 1985).
[11] J. Park and H. Margenau, ''Simultaneous measurement in quantum theory,'' Int. J. Theor. Phys. **1**, 211–283 (1968).
[12] J. Hartle, ''Spacetime grainings in nonrelativistic quantum mechanics'' Phys. Rev. D **44**, 3173–3195 (1991).
[13] J. Hartle, ''Spacetime quantum mechanics and the quantum mechanics of spacetime,'' in *Proceedings on the 1992 Les Houches School, Gravitation and Quantisation* (North-Holland, Amsterdam, 1993).
[14] M. Blencowe, ''The consistent histories interpretation of quantum fields in curved space-time,'' Ann. Phys. (N.Y.) **211**, 87–111 (1991).

Classical limits for quantum maps on the torus

Andrzej Lesniewski, Ron Rubin, and Nathan Salwen
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138-2901

(Received 11 March 1997; accepted for publication 26 March 1997)

We provide a rigorous canonical quantization for the following toral automorphisms: cat maps, generalized kicked maps, and generalized Harper maps. For each of these systems we construct a unitary propagator and prove the existence of a well-defined classical limit. We also provide an alternative derivation of Hannay and Berry results for the cat map propagator on the plane. © *1998 American Institute of Physics.* [S0022-2488(98)02404-9]

I. INTRODUCTION

The burgeoning field of ''quantum chaos'' or, more precisely, the quantum mechanics of classically chaotic dynamics has opened the door to several issues in quantization (see e.g., Ref. 1). One of these is the quantization of classical maps.[2–5]

Classical maps arise naturally in the study of chaos. In the Poincaré surface of section maps (see Ref. 6), for instance, the flow on a d-dimensional phase space is ''reduced'' to a $(d-1)$-dimensional map by considering the discrete evolution of points of intersection of the trajectory with a co-dimension of one hypersurface. Chaotic properties of the flow typically leave tell-tale signs on the surface of section. It is clear that the two problems are intimately connected, and numerical results indicate this. Another typical way in which classical flows lead to maps is via a coarse graining of Hamilton's equation,

$$p'=p+f(x)\Delta t, \quad x'=x+p',$$

where $f(x)=-\nabla_x u(x)$ is the classical force function for the Hamiltonian $H=p^2/2+u(x)$, and where in the second equation p' is used rather than p to ensure the map is area preserving. For maps corresponding to integrable flows, the parameter Δt is a type of ''order parameter'' for the transition to chaos in the corresponding family of maps. Typically for large Δt, the system is completely chaotic, while for small Δt, the system mimics its continuous time progenitor and remains on regular trajectories for all time. (Examples abound. See, e.g., Ref. 6.)

The quantization of classically chaotic maps has recently become an interesting issue in its own right. Hannay and Berry[3] provide a quantization of the cat map

$$(x',p')=\beta(x,p) \mod 1, \quad \beta\in SL(2,\mathbb{Z}), \tag{1}$$

using a semiclassical argument. First they impose the integrality condition of Planck's constant $h=1/N$. Then argue that on the torus the Hilbert space has dimension N. This is semiclassical in that it corresponds to the quantization condition

$$\oint p\,dx=nh. \tag{2}$$

For the unit torus, the left-hand side of (2) is unity, and the integrality condition for Planck's constant follows. In fact, this is also the integrality condition of geometric quantization which requires the symplectic form divided by Planck's constant to define a deRham cohomology class with integer coefficients.[7] With the Hannay–Berry prescription for quantization, a further restriction was necessary: only a subset of $SL(2,\mathbb{Z})$, the so-called ''checkerboard'' matrices

$$\begin{pmatrix} odd & even \\ even & odd \end{pmatrix}, \quad \begin{pmatrix} even & odd \\ odd & even \end{pmatrix} \tag{3}$$

 © 1998 American Institute of Physics

could be quantized.

In Refs. 8, 4, and 5, a different approach toward quantizing the cat maps was introduced, in which the dynamics is described in the Heisenberg picture as the evolution of a noncommutative algebra of observables. The Hilbert space remains infinite dimensional, and no restrictions on Planck's constant or the allowed classical maps are made. The quantum torus is constructed by deforming the classical algebra of observables. We let

$$U=Q_\hbar(e^{2\pi ix}), \quad V=Q_\hbar(e^{2\pi ip}),$$

where $Q_\hbar(\sigma)$ is an operator on the Hilbert space $\mathscr{H}$ corresponding to the classical symbol σ, and the algebra of observables $\mathfrak{A}_\hbar$ is then generated by $\{U,V,U^\dagger,V^\dagger\}$. Technically, the algebra is given with an appropriate closure. For example, taking $\mathfrak{A}_\hbar=\{U,V,U^\dagger,V^\dagger\}''$, where R'' is the bicommutant of R, the algebra generated is a von Neumann algebra. The quantum dynamics is implemented as a discrete automorphism group of $\mathfrak{A}_\hbar$ generated by a unitary operator (called the propagator) F such that as $\hbar\to 0$, $F^\dagger Q_\hbar(\sigma)F\to Q_\hbar(\sigma\circ\beta)$ in a suitable topology.

A quantization scheme of the kind just described was given in Refs. 4 and 5 for the cat map, along with the connection to the original quantization proposed by Hannay and Berry. In this paper, we extend the results to three classes of maps.

We divide the remainder of the paper into three sections, corresponding to the different maps. In Sec. II, we review the relevant results of Refs. 4 and 5 for the cat maps, and also provide a canonical derivation of the propagator *on the plane* obtained by Ref. 3 using a ''discrete time path integral.'' In Sec. III, we provide a quantization scheme for ''generalized kick maps'' $\kappa(x,p)=(x',p')$, with

$$p'=p+\epsilon f(x), \quad x'=x+p'. \tag{4}$$

Here ϵ is a real constant, and f is a continuous real function (the force function) on the unit circle with Fourier coefficients f_k obeying the bound

$$\sum_{k\in\mathbb{Z}} k^2|f_k|<\infty. \tag{5}$$

In Sec. IV, we quantize ''generalized Harper maps'': $\eta(x,p)=(x',p')$, with

$$p'=p+\epsilon_1 f(x), \quad x'=x+\epsilon_2 g(p'), \tag{6}$$

where ϵ_1 and ϵ_2 are real constants, and f and g are real functions on the unit circle with Fourier coefficients satisfying the bound:

$$\sum_j j^2|f_j|<\infty, \quad \sum_k k^2|g_k|\exp\left(2\pi\epsilon_1 k\left(\sum_j |f_j|\right)\right)<\infty.$$

Note for generic f in kick maps, and generic f and g in Harper maps, the coefficients ϵ and ϵ_1, ϵ_2 are the parameters which induce a transition to chaos as they are increased in magnitude.

II. THE CAT MAP REVISITED

The cat map (1) has become a paradigm for the study of quantum chaos (see, e.g., Ref. 9). The first quantization scheme for the cat map was given by Hannay and Berry[3] in which a path integral argument from continuous time linear dynamics was used to write down the propagator for quantum evolution. In Ref. 4, a canonical quantization was proposed using Toeplitz operators in Bargmann space, and an explicit integral kernel for the unitary one-step evolution on the entire complex plane was given. Recall that Bargmann space $\mathscr{H}^2(\mathbb{C},d\mu_\hbar)\subset L^2(\mathbb{C},d\mu_\hbar)$ is the Hilbert space of entire functions on $\mathbb{C}$ which are square integrable with respect to the measure $d\mu_\hbar(z)=(\pi\hbar)^{-1}\exp(-|z|^2/\hbar)d^2z$. Formally, Toeplitz quantization $T_\hbar$ of a symbol over phase space is the operator of multiplication on $L^2(\mathbb{C},d\mu_\hbar)$ by this symbol followed by a projection onto $\mathscr{H}^2(\mathbb{C},d\mu_\hbar)$. For calculations, we must simply note that this gives an anti-Wick ordering of the corresponding quantum operator: $T_\hbar=(z^n\overline{z^m})=T_\hbar(\overline{z^m})T_\hbar(z^n)$. The quantum torus is then con-

structed by restricting the algebra of observables to be the Toeplitz quantization of Fourier series with classical generators $\exp(2\pi ix)$ and $\exp(2\pi ip)$. As in Refs. 4 and 5, we define the unitary operators U and V on $\mathscr{H}^2(\mathbb{C},d\mu_\hbar)$ to be the quantization of classical generators for $f\in C(\mathbb{T}^2)$, i.e.,

$$U=\exp(2\pi i\hat{x})=\exp(\pi^2\hbar)T_\hbar(\exp(2\pi ix)),$$

$$V=\exp(2\pi i\hat{p})=\exp(\pi^2\hbar)T_\hbar(\exp(2\pi ip)),$$

where $\hat{x}=T_\hbar(x)$, $\hat{p}=T_\hbar(p)$, and $z=(1/\sqrt{2})(x-ip)$. Observe also that the canonical position and momentum variables on the plane satisfy the usual Heisenberg relation $[\hat{x},\hat{p}]=i\hbar$, and so

$$UV=e^{-4\pi^2\hbar i}VU. \tag{7}$$

Bargmann space is natural in that a phase space dynamics is given a phase space quantization. However, it is clear from the isomorphism between $\mathscr{H}^2(\mathbb{C},d\mu_\hbar)$ and $L^2(\mathbb{R},dx)$ (see Ref. 10) that we should be able to reformulate the problem in $L^2(\mathbb{R})$. Following Ref. 4, to the classical cat map β defined in Eq. (1) we assign a corresponding quantum propagator C. For linear dynamics *on the torus* we have the following result.

Theorem 1: *There exists a unique (up to a phase) unitary quantum propagator C satisfying the following properties*:

(1)

$$C^\dagger\hat{x}C=a\hat{x}+b\hat{p}=\hat{x}', \quad C^\dagger\hat{p}C=c\hat{x}+d\hat{p}=\hat{p}', \tag{8}$$

where $\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)\in SL(2,\mathbb{Z})$, *such that C maps the harmonic oscillator ground state into the unique ground state of the transformed coordinates.*

(2) The mapping $U\to U'=C^\dagger UC$, *and* $V\to V'=C^\dagger VC$ *extends to an automorphism of the von Neumann algebra* $\mathfrak{A}_\hbar$.

(3) For $\sigma\in C(\mathbb{T}^2)$,

$$\|C^\dagger T_\hbar(\sigma)C-T_\hbar(\sigma\circ\beta)\|\to 0 \quad \text{as} \quad \hbar\to 0.$$

(4) Acting on a state $\Phi\in L^2(\mathbb{R})$, *C can be written explicitly as*

$$C\Phi(x)=\left(\frac{1}{hb}\right)^{1/2}\int_{\mathbb{R}} e^{i(ay^2-2yx+dx^2)/2b}\Phi(y)dy. \tag{9}$$

Remark 1: $C(x,y)=(1/hb)^{1/2}e^{i(ay^2-2yx+dx^2)/2b}$ *is precisely the integral kernel proposed in Ref. 3 using a semiclassical path integral argument.*

Proof: The proofs of (1), (2), and (3) are given in Ref. 4.

We prove statement (4). Following Ref. 10, the isomorphism between $\mathscr{H}^2(\mathbb{C},d\mu_\hbar)$ and $L^2(\mathbb{R})$ is implemented by the Bargmann transform. For $x\in\mathbb{R}$, and $z\in\mathbb{C}$, the integral kernel $B:L^2(\mathbb{R})\to\mathscr{H}^2(\mathbb{C},d\mu_\hbar)$ is given by

$$B(z,x)=\left(\frac{1}{\pi\hbar}\right)^{1/4}\exp(\sqrt{2}xz/\hbar-x^2/2\hbar-z^2/2\hbar).$$

Note also that

$$B^{-1}(z,x)=B(\overline{z},x).$$

The integral kernel for the linear dynamics on the plane in Bargmann representation was found in Ref. 4 to be

$$F(z,w)=\frac{1}{\sqrt{|\alpha|}}\exp\left\{-\frac{\overline{\beta}z^2}{2\hbar\alpha}+\frac{\overline{w}z}{\hbar\alpha}+\frac{\beta\overline{w}^2}{2\hbar\alpha}\right\},$$

where

$$\alpha=\tfrac{1}{2}(a+d+i(b-c)),\quad \beta=\tfrac{1}{2}(a-d+i(b+c)).$$

Then

$$C(x,y)=\int B(\overline{z},x)F(z,w)B(w,y)d\mu_\hbar(w)d\mu_\hbar(z)$$

$$=\frac{1}{\sqrt{\pi\hbar|\alpha|}}\int \exp(\sqrt{2}x\overline{z}/\hbar-x^2/2\hbar-\overline{z}^2/2\hbar-\overline{\beta}\overline{z}^2/2\hbar\alpha+\overline{w}z/\hbar\alpha+\beta\overline{w}^2/2\hbar\alpha$$
$$+\sqrt{2}yw/\hbar-y^2/2\hbar-w^2/2\hbar-|z|^2/\hbar-|w|^2/\hbar)d^2zd^2w.$$

This integral is a straightforward but tedious Gaussian integral. It can be calculated directly with the following result. For complex numbers A, B, C, D, E such that $\mathrm{Re}\,A>0$ and $(\mathrm{Re}\,A)^2>(\mathrm{Re}\,B+\mathrm{Re}\,C)^2+(\mathrm{Im}\,B-\mathrm{Im}\,C)^2$, we find

$$\int_C \exp(-A|z|^2+Bz^2+C\overline{z}^2+Dz+E\overline{z})d^2z$$

$$=\frac{\pi}{\sqrt{A^2-4BC}}\exp\frac{(D+E)^2(A^2-4BC)-(A(D-E)+2(EB-CD))^2}{4(A-B-C)(A^2-4BC)}.$$

Applying this formula twice to the above integral gives the desired result. ■

III. QUANTUM KICK MAPS

We provide here a procedure for quantizing another class of automorphisms of the torus described in Sec. I. Exponentiating (4), we have for the classical maps on the torus

$$e^{2\pi ix'}=e^{2\pi i(x+p')},\quad e^{2\pi ip'}=e^{2\pi i(p+\epsilon f(x))}. \tag{10}$$

In fact, we let (10) also give the quantum evolution, that is, we find a quantum propagator such that the one-step dynamics is given by (4) exactly, with x and p replaced by the Toeplitz operators $\hat{x}$ and $\hat{p}$. However, it is not immediately clear that such a map will give a well-defined automorphism of the quantum algebra $\mathfrak{A}$ on the torus. We now demonstrate this.

Lemma 2: Let f be a real valued function satisfying (5). Then the sum $\Sigma_{k\in Z}\widetilde{f}_kU^k$, where

$$\widetilde{f}_k=\left(\frac{1-e^{-i\lambda k}}{i\lambda k}\right)f_k \tag{11}$$

defines a bounded self-adjoint operator.

Proof: This is an immediate consequence of the estimate

$$\sum_k \|\widetilde{f}_kU^k\|\leqslant\sum_k|\widetilde{f}_k|\leqslant\sum_k|f_k|<\infty.$$

The second inequality follows from (19), while the last bound on the Fourier coefficients is by assumption. ■

We now define the quantum map corresponding to (10),

$$U'=e^{-i\lambda/2}V'U,\quad V'=V\exp\left(2\pi i\epsilon\sum_{k\in\mathbb{Z}}\widetilde{f}_kU^k\right), \tag{12}$$

where

$$\lambda=4\pi^2\hbar.$$

In the following we use the standard functional calculus for self-adjoint operators to define functions of the operators $\hat{x}$ and $\hat{p}$.

Lemma 3: With the above definitions,

$$U' = \exp(2\pi i(\hat{x} + \hat{p}')), \quad V' = \exp 2\pi i(\hat{p} + \epsilon f(\hat{x})). \tag{13}$$

Proof: Using Trotter's formula (see, e.g., Ref. 11), for T an integer,

$$\exp 2\pi i(\hat{p} + \epsilon f(\hat{x})) = \lim_{T\to\infty} (\exp[2\pi i\hat{p}/T]\exp[2\pi i\epsilon f(\hat{x})/T])^T, \tag{14}$$

where the limit is meant in the strong operator topology. Using

$$e^{-2\pi i\hat{p}/T}\hat{x}e^{2\pi i\hat{p}/T} = \hat{x} - 2\pi\hbar/T,$$

we see that the right-hand side of (14) reduces to

$$\begin{aligned}
&\exp(2\pi i\hat{p}) \lim_{T\to\infty} \exp \frac{2\pi i\epsilon}{T} \sum_{j=0}^{T-1} \sum_k f_k \exp[2\pi ik(\hat{x} - 2\pi j\hbar/T)] \\
&= \exp(2\pi i\hat{p}) \lim_{T\to\infty} \exp\left(\frac{2\pi i\epsilon}{T} \sum_k f_k \exp[2\pi ik\hat{x}] \frac{1 - \exp(-4\pi^2\hbar ik)}{1 - \exp(4\pi^2\hbar ik/T)} \right) \\
&= \exp(2\pi i\hat{p}) \exp\left(2\pi i\epsilon \sum_k \widetilde{f}_k \exp(2\pi ik\hat{x}) \right) = V',
\end{aligned}$$

where $\widetilde{f}_k$ is defined in (11). We also find

$$\exp(2\pi i(\hat{x} + \hat{p}')) = \exp(2\pi i\hat{p}')\exp(2\pi i\hat{x})\exp(-2\pi^2[\hat{x}, \hat{p}']) = V'UE^{-i\lambda/2} = U',$$

and the claim follows. ∎

Theorem 4: *Let f be a real valued function satisfying (5).*

(1) There exists a unitary operator (the quantum propagator) which implements (12), i.e.,

$$U' = K^\dagger UK, \quad V' = K^\dagger VK.$$

Explicitly, this operator is given by

$$K = e^{-i\hat{p}^2/2\hbar} e^{i\epsilon u(\hat{x})/\hbar}, \tag{15}$$

where u is a real differentiable function on the unit circle defined (up to a real constant) by $du(x)/dx = f(x)$.

(2) The map $U \to U'$, $V \to V'$ defined above extends to an automorphism of the von Neumann algebra of observables on the quantized torus. We call this automorphism the quantized kick map.

(3) The quantum dynamics has a well-defined classical limit in the norm topology. For $\sigma \in C(\mathbb{T}^2)$, we have

$$\|K^{-1}T_\hbar(\sigma)K - T_\hbar(\sigma\circ\kappa)\| \to 0, \text{ as } \hbar \to 0,$$

where $\kappa(x,p) = (p',x')$ is defined in Eq. (4).

Proof: (1) This is a straightforward calculation. Observe for K as in (15),

$$\begin{aligned}
K^\dagger VK &= \exp(2\pi iK^\dagger\hat{p}K) = \exp[2\pi i\{\exp(-i\epsilon u(\hat{x})/\hbar)\hat{p}\ \exp(i\epsilon u(\hat{x})/\hbar)\}] \\
&= \exp\left[2\pi i\left(\hat{p} + \epsilon \frac{du}{dx}(\hat{x}) \right) \right] = V',
\end{aligned}$$

where in the last step we have used Lemma 3. Likewise,

$$K^\dagger UK = \exp[2\pi i(\exp(-i\epsilon u(\hat{x})/\hbar)\exp(i\hat{p}^2/2\hbar)\hat{x}\ \exp(-i\hat{p}^2/2\hbar)\exp(i\epsilon u(\hat{x})/\hbar))]$$
$$= \exp[2\pi i(\hat{p}' + \hat{x})] = U'.$$

(2) It follows easily from part (1) that U' and V' satisfy (7). Since U and V can be expressed in terms of $U'$and V', namely,

$$U = e^{i\lambda} V'^\dagger U', \quad V = V'\ \exp\left(-2\pi i\epsilon \sum_{k\in\mathbb{Z}} \tilde{f}_k (e^{i\lambda} V'^\dagger U')^k\right),$$

this implies that U' and V' generate $\mathfrak{A}_\hbar$. The map is thus an automorphism of the algebra $\mathfrak{A}_\hbar$.

(3) As the proof of (3) is somewhat involved, we divide it into steps.

Step 1. We first calculate the terms in (3) for the case of the pure harmonic

$$\sigma_{mn}(x,p) := \exp(2\pi i(mx+np)).$$

Applying the Toeplitz quantization to the symbol σ_{mn} and using the Baker–Hausdorff–Campbell relation in the last step, we get

$$T_\hbar(\sigma_{mn}) = T_\hbar(\exp(\sqrt{2}\pi i(m-ni)\overline{z})\exp(\sqrt{2}\pi i(m+ni)z))$$
$$= \exp(\pi i(m-ni)(\hat{x}+i\hat{p}))\exp(\pi i(m+ni)(\hat{x}-i\hat{p}))$$
$$= V^n U^m\ \exp(-\pi^2\hbar(n^2+m^2+2inm)). \tag{16}$$

Thus, using (14) once again, we find

$$K^\dagger T_\hbar(\sigma_{mn}) K = (V')^n (U')^m\ \exp(-\pi^2\hbar(n^2+m^2+2inm))$$
$$= (V')^{n+m} U^m\ \exp(-i\lambda m/2)\exp(-i\lambda[m(m-1)]/2)\exp(-\pi^2\hbar(n^2+m^2+2inm))$$
$$= V^{n+m} U^m\ \exp(-\lambda(n^2+m^2+2inm+2im^2)/4)$$
$$\times \exp\left(2\pi i\epsilon \sum_k\ (m+n)\tilde{f}_k(m+n) U^k\right), \tag{17}$$

where, for any integer n,

$$\tilde{f}_k(n) = \left(\frac{1-\exp(-i\lambda kn)}{i\lambda kn}\right) f_k. \tag{18}$$

Expanding the right-hand side of (17), we have

$$K^\dagger T_\hbar(\sigma_{mn})K = V^{m+n}U^m \exp(-\lambda(n^2+m^2+2inm+2im^2)/4) \times \sum_{l=0}^{\infty} \frac{(2\pi i\epsilon)^l (m+n)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} \tilde{f}_{k_j}(m+n)U^{k_j}.$$

We want to compare this last expression with

$$T_\hbar(\sigma_{mn}\circ\kappa(x,p)) = T_\hbar(\exp(2\pi i[n(p+\epsilon f(x))+m(x+p+\epsilon f(x))])) = T_\hbar\left(\exp 2\pi i((m+n)p+mx)\sum_{l=0}^{\infty} \frac{[2\pi i(m+n)\epsilon(\Sigma_{k\in\mathbb{Z}} f_k e^{2\pi ikx})]^l}{l!}\right).$$

Using (16) to expand the right-hand side of the equation above, we find

$$T_\hbar(\sigma_{mn}\circ\kappa(x,p)) = V^{m+n}U^m \exp\left(-\frac{\lambda}{4}[(m+n)^2+2i(m^2+mn)]\right) \times \sum_{l=0}^{\infty} \frac{(2\pi i\epsilon)^l (m+n)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} f_{k_j}U^{k_j} \times \exp\left(-\frac{\lambda}{4}[(m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l)]\right).$$

Step 2. Next we prove (3) for the case of a pure harmonic. Since $\tilde{f}_{k_j}(n)$ approaches f_k as $\hbar \to 0$, it is clear that term by term (in powers of V^pU^q) the difference $K^\dagger T_\hbar(\sigma_{mn})K$ and $T_\hbar(\sigma_{mn}\circ\kappa)$ vanishes as $\hbar\to 0$. To show the entire sum vanishes requires a little more work.

We shall see that this indeed happens provided the force function f satisfies our assumption. To this end, we define the following norm of f:

$$\|f\|_2 := \sum_k (1+|k|)^2|f_k| < \infty.$$

Notice first of all the difference

$$\|K^\dagger T_\hbar(\sigma_{mn})K - T_\hbar(\sigma_{mn}\circ\kappa)\| \leq \Bigg\| \exp\left(-\frac{\lambda}{4}(n^2+m^2+2inm+2im^2)\right) \times \sum_{l=0}^{\infty} \frac{(2\pi i\epsilon)^l (m+n)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} \tilde{f}_{k_j}(m+n)U^{k_j} - \exp\left(-\frac{\lambda}{4}[(m+n)^2+2i(m^2+mn)]\right) \times \sum_{l=0}^{\infty} \frac{(2\pi i\epsilon)^l (m+n)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} f_{k_j}U^{k_j} \times \exp -\frac{\lambda}{4}[(m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l)]\Bigg\|,$$

where we have used the fact that $\|U\| = \|V\| = 1$. We next use a simple argument to eliminate the exponentials preceding the sums in both terms on the right-hand side of the above expression. We see that

$$\begin{aligned}\|K^\dagger T_\hbar(\sigma_{mn})K - T_\hbar(\sigma_{mn}\circ\kappa)\| \leqslant & \left\| \left(1-\exp-\frac{\lambda}{4}(n^2+m^2+2inm-2im^2)\right)\right. \\ & \left.\times\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}\prod_{j=1}^{l}\widetilde{f}_{k_j}(m+n)U^{k_j}\right\| \\ & +\left\|\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}\prod_{j=1}^{l}U^{k_j}\left(\sum_{j=1}^{l}\widetilde{f}_{k_j}(m+n)-\prod_{j=1}^{l}f_{k_j}\right.\right. \\ & \left.\left.\times\exp-\frac{\lambda}{4}[(m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l)]\right)\right\| \\ & +\left\|\left(1-\exp-\frac{\lambda}{4}((m+n)^2+2i(m^2+mn))\right)\right. \\ & \times\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}\prod_{j=1}^{l}f_{k_j}U^{k_j} \\ & \left.\times\exp-\frac{\lambda}{4}((m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l))\right\|.\end{aligned}$$

The first and third terms on the right-hand side are both of the form $\|(1-\exp-a\lambda)F\|$, where F is a bounded operator as $\lambda\to 0$ and $\mathrm{Re}(a)>0$. From the mean value theorem, it follows that

$$|1-\exp(-a\lambda)|\leqslant|a|\lambda, \tag{19}$$

so that these two terms are both $O(\lambda)$. Thus

$$\begin{aligned}&\|K^\dagger T_\hbar(\sigma_{mn})K - T_\hbar(\sigma_{mn}\circ\kappa)\| \\ &\quad\leqslant O(\lambda)+\left\|\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}U^{k_j}\left(\prod_{j=1}^{l}\widetilde{f}_{k_j}(m+n)-\prod_{j=1}^{l}f_{k_j}\right.\right. \\ &\qquad\left.\left.\times\exp-\frac{\lambda}{4}((m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l))\right)\right\|.\end{aligned}$$

An only slightly more complicated telescoping argument eliminates the exponentials inside the remaining sums. We see

$$\begin{aligned}&\|K^\dagger T_\hbar(\sigma_{mn})K - T_\hbar(\sigma_{mn}\circ\kappa)\| \\ &\quad\leqslant O(\lambda)+\left\|\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}U^{k_j}\left(\prod_{j=1}^{l}\widetilde{f}_{k_j}(m+n)-\prod_{j=1}^{l}f_{k_j}\right)\right\| \\ &\qquad+\left\|\sum_{l=0}^{\infty}\frac{(2\pi i\epsilon)^l(m+n)^l}{l!}\sum_{k_1,\ldots,k_l}\prod_{j=1}^{l}U^{k_j}f_{k_j}\right. \\ &\qquad\left.\times\left[1-\exp-\frac{\lambda}{4}((m+k_1+\cdots+k_l)^2+2i(m+n)(k_1+\cdots+k_l))\right]\right\|.\end{aligned}$$

Consider the last set of terms only in the above expression. We have, using (19) once again

$$\begin{aligned}
&\left\| \sum_{l=0}^{\infty} \frac{(2\pi i\epsilon)^l (m+n)^l}{l!} \right. \\
&\qquad \left. \times \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} U^{k_j} f_{k_j} \left[1 - \exp - \frac{\lambda}{4} \left((m+k_1+\cdots+k_l)^2 + 2i|m+n|(k_1+\cdots+k_l) \right) \right] \right\| \\
&\leqslant \frac{\lambda}{4} \sum_{l=0}^{\infty} \frac{(2\pi\epsilon)^l (|n|+|m|)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} |f_{k_j}| \\
&\qquad \times \left| (m+k_1+\cdots+k_l)^2 + 2(|m|+|n|)(k_1+\cdots+k_l) \right| \\
&\leqslant \frac{\lambda}{4} \sum_{l=0}^{\infty} \frac{(2\pi\epsilon)^l (|n|+|m|)^l}{l!} \sum_{k_1,\ldots,k_l} \prod_{j=1}^{l} |f_{k_j}| \\
&\qquad \times (m^2 + l(k_1^2+\cdots+k_l^2) + (4|m|+2|n|)(k_1+\cdots+k_l)) \\
&\leqslant \frac{\lambda}{4} \sum_{l=0}^{\infty} \frac{(2\pi\epsilon)^l (|n|+|m|)^l}{l!} (m^2 + 4|m| + 2|n| + l) \|f\|_2^l \\
&\leqslant \frac{\lambda}{4} (m^2 + 4|m| + 2|n| + 2\pi\epsilon(|n|+|m|)\|f\|_2) \exp 2\pi\epsilon(|n|+|m|)\|f\|_2 = O(\lambda).
\end{aligned}$$

Combining all of this, we have reduced our estimate to

$$\begin{aligned}
&\|K^\dagger T_\hbar(\sigma_{mn}) K - T_\hbar(\sigma_{mn} \circ \kappa)\| \\
&\qquad \leqslant O(\lambda) + \sum_{l=0}^{\infty} \frac{(2\pi\epsilon)^l (|n|+|m|)^l}{l!} \\
&\qquad\qquad \times \sum_{k_1,\ldots,k_l} \left| 1 - \left(\prod_{j=1}^{l} \frac{1 - \exp(-i\lambda k_j (m+n))}{i\lambda k_j (m+n)} \right) \right| |f_{k_1}| \cdots |f_{k_l}|. \qquad (20)
\end{aligned}$$

By Taylor's theorem with remainder, we know that for $\alpha \in \mathbb{R}$,

$$|1 - i\alpha - e^{-i\alpha}| = \left| \frac{\alpha^2}{2} (\cos\beta_1 - i\cos\beta_2) \right| \leqslant |\alpha^2|, \qquad (21)$$

where β_1 and β_2 are between 0 and α. Similarly, $|1 - e^{-i\alpha}| \leqslant |\alpha|$. Using these two bounds, we see that

$$\begin{aligned}
\left| 1 - \prod_{j=1}^{l} \frac{1 - \exp(-i\lambda k_j (m+n))}{i\lambda k_j (m+n)} \right| &\leqslant \left| 1 - \frac{1 - \exp(-i\lambda k_1 (m+n))}{i\lambda k_1 (m+n)} \right| + \left| \frac{1 - \exp(-i\lambda k_1 (m+n))}{i\lambda k_1 (m+n)} \right| \\
&\quad \times \left| 1 - \frac{1 - \exp(-i\lambda k_2 (m+n))}{i\lambda k_2 (m+n)} \right| + \cdots \\
&\quad + \left| \frac{1 - \exp(-i\lambda k_1 (m+n))}{i\lambda k_1 (m+n)} \right| \cdots \left| 1 - \frac{1 - \exp(-i\lambda k_l (m+n))}{i\lambda k l (m+n)} \right| \\
&\leqslant \lambda (m+n)(|k_1| + \cdots + |k_l|).
\end{aligned}$$

Putting this back in (20) yields

$$\|K^\dagger T_\hbar(\sigma_{mn}) K - T_\hbar(\sigma_{mn} \circ \kappa)\| \leqslant O(\lambda) + \lambda (2\pi\epsilon)(|n|+|m|)^2 \|f\|_2 e^{2\pi\epsilon(|n|+|m|)\|f\|_2} = O(\lambda).$$

The right-hand side vanishes as $\lambda \to 0$. This completes the proof for a pure harmonic σ_{mn}.

Step 3. We can now complete the proof for any classical symbol $\sigma \in C(\mathbb{T}^2)$ with a simple application of the Stone–Weierstrass theorem. For such σ there exists a trigonometric polynomial P such that given $\delta > 0$,

$$\|\sigma - P\|_\infty \leqslant \delta/3,$$

where $\|\sigma\|_\infty = \sup_z |\sigma(z)|$ is the sup-norm. Now, since the operator norm of a Toeplitz operator does not exceed the sup-norm of the symbol, i.e., $\|T_\hbar(\sigma)\| \leqslant \|\sigma\|_\infty$ (see Ref. 10), we find

$$\|T_\hbar(\sigma) - T_\hbar(P)\| \leqslant \delta/3.$$

Since P is a linear combination of finitely many simple harmonics, by part Step 2, we can find γ (depending on P) such that for $\hbar < \gamma$

$$\|K^\dagger T_\hbar(P) K - T_\hbar(P \circ \kappa)\| \leqslant \delta/3.$$

Thus

$$\begin{aligned}\|K^\dagger T_\hbar(\sigma) K - T_\hbar(\sigma \circ \kappa)\| &\leqslant \|K^\dagger T_\hbar(\sigma) K - K^\dagger T_\hbar(P) K\| + \|K^\dagger T_\hbar(P) K - T_\hbar(P \circ \kappa)\| + \|T_\hbar(P \circ \kappa) \\ &\quad - T_\hbar(\sigma \circ \kappa)\| \leqslant \|T_\hbar(\sigma) - T_\hbar(P)\| + \delta/3 + \|(P - \sigma) \circ \kappa\|_\infty \leqslant \delta.\end{aligned}$$

This completes the proof of the theorem. ■

IV. GENERALIZED HARPER MAPS

We complete the class of quantizations of toral automorphisms with the generalized Harper maps. Exponentiating (6), we find

$$e^{2\pi i x'} = e^{2\pi i (x + \epsilon_2 g(p'))}, \quad e^{2\pi i p'} = e^{2\pi i (p + \epsilon_1 f(x))}. \tag{22}$$

As before, we let (22) give the quantum evolution. That is, we are led to the following definition of the quantum Harper dynamics:

$$U' = \exp\left(2\pi i \epsilon_2 \sum_{k \in \mathbb{Z}} (V')^k \widetilde{g}_k \right) U, \quad V' = V \exp\left(2\pi i \epsilon_1 \sum_{k \in \mathbb{Z}} \widetilde{f}_k U^k \right), \tag{23}$$

where

$$\widetilde{g}_k = \left(\frac{1 - \exp(-i\lambda k)}{i\lambda k} \right) g_k, \quad \widetilde{f}_k = \left(\frac{1 - \exp(-i\lambda k)}{i\lambda k} \right) f_k,$$

and

$$g(\hat{p}) = \frac{dv}{dp}(\hat{p}) = \sum_{k \in \mathbb{Z}} g_k V^k, \quad f(\hat{x}) = \frac{du}{dx}(\hat{x}) = \sum_{k \in \mathbb{Z}} f_k U^k.$$

Calculations similar to the kick map case show that

$$U' = e^{2\pi i (\hat{x} + \epsilon_2 g(\hat{p} + \epsilon_1 f(\hat{x})))}, \quad V' = e^{2\pi i (\hat{p} + \epsilon_1 f(\hat{x}))}.$$

We have the following theorem, which shows that the map above indeed yields a well-defined quantum dynamics.

Theorem 5: *(1) There exists a unitary quantum propagator which implements (23), namely,*

$$H=\exp(i\epsilon_2 v(\hat{p})/\hbar)\exp(-i\epsilon_1 u(\hat{x})/\hbar).$$

(2) The quantum Harper map defined by (23) extends to an automorphism of the von Neumann algebra $\mathfrak{A}_\hbar$ of observables on the quantized torus.

(3) For maps with f and g satisfying

$$\sum_j j^2|f_j|<\infty, \quad \sum_k k^2|g_k|\exp\left(2\pi\epsilon_1 k\sum_j |f_j|\right)<\infty,$$

the quantum propagator has a well-defined classical limit:

$$\|H^\dagger T_\hbar(\sigma)H-T_\hbar(\sigma\circ\eta)\|\to 0 \quad \text{as} \quad \hbar\to 0,$$

where $\eta(p,x)=(p',x')$ is the classical evolution defined in (6).

Proof: (1) Observe

$$H^\dagger UH=\exp(i\epsilon_l u(\hat{x})/\hbar)\exp(-i\epsilon_2 v(\hat{p})/\hbar)U\ \exp(i\epsilon_2 v(\hat{p})/\hbar)\exp(-i\epsilon_1 u(\hat{x})/\hbar)$$

$$=\exp(i\epsilon_1 u(\hat{x})/\hbar)\exp(2\pi i(\hat{x}+\epsilon_2 g(\hat{p})))\exp(-i\epsilon_1 u(\hat{x})/\hbar)=\exp(2\pi i(\hat{x}+\epsilon_2 g(\hat{p}'))).$$

Similarly we see $H^\dagger VH=V'$.

(2) The proof that the mapping is an automorphism of the algebra $\mathfrak{A}_\hbar$ follows that of the kick map. We omit the details here.

(3) We break the proof of (3) into the same steps as in the proof for the kick map.

Step 1. First we calculate $H^\dagger T_\hbar(\sigma_{mn})H=(V')^n(U')^m\exp-\pi^2\hbar(m^2+n^2+2imn)$. Then using Trotter's formula, we get

$$(V')^n=V^n\ \exp 2\pi i\epsilon_1 n\sum_{k\in\mathbb{Z}}\tilde{f}_k(n)U^k,$$

and

$$(U')^m=\exp\left(2\pi i\epsilon_2 m\sum_{k\in\mathbb{Z}}\tilde{g}_k(m)(V')^k\right)U^m,$$

where $\tilde{f}_k(n)$ and $\tilde{g}_k(n)$ are defined as in (18). Thus

$$\begin{aligned}H^\dagger T_\hbar(\sigma_{mn})H=&\sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\frac{(2\pi i\epsilon_2 m)^l}{l!}\tilde{g}_{k_1}(m)\cdots\tilde{g}_{k_l}(m)V')^{n+k_1+\cdots+k_l}U^m\\&\times\exp(-\pi^2\hbar(m^2+n^2+2imn))\\=&\sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\frac{(2\pi i\epsilon_2 m)^l}{l!}\tilde{g}_{k_1}(m)\cdots\tilde{g}_{k_l}(m)V^{n+k_1+\cdots+k_l}\\&\times\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q}\frac{(2\pi i\epsilon_1(n+k_1+\cdots+k_l))^q}{q!}\left(\prod_{i=1}^{q}\tilde{f}_{j_i}(n+k_1+\cdots+k_l)\right)\\&\times U^{m+j_1+\cdots+j_q}\exp(-\pi^2\hbar(m^2+n^2+2imn)).\end{aligned}$$

Next, we calculate

$$
\begin{aligned}
T_\hbar(\sigma_{mn}\circ\eta) &= T_\hbar(\exp[2\pi i(mx'+np')]) \\
&= \sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q} T_\hbar\Bigg(\frac{(2\pi i\epsilon_2 m)^l}{l!}\, g_{k_1}\cdots g_{k_l} \\
&\quad \times\frac{(2\pi i\epsilon_1(n+k_1+\cdots+k_l))^q}{q!}\, f_{j_1}\cdots f_{j_q} \\
&\quad \times\exp\{2\pi i((m+j_1+\cdots+j_q)x+(n+k_1+\cdots+k_l)p)\}\Bigg) \\
&= \sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q} \frac{(2\pi i\epsilon_2 m)^l}{l!}\, g_{k_1}\cdots g_{k_l} \\
&\quad \times\frac{(2\pi i\epsilon_1(n+k_1+\cdots+k_l))^q}{q!}\, f_{j_1}\cdots f_{j_q} V^{n+k_1+\cdots+k_l}U^{m+j_1+\cdots-j_q} \\
&\quad \times\exp[(n+k_1+\cdots+k_l)^2+(m+j_1+\cdots+j_q)^2 \\
&\quad +2i(n+k_1+\cdots+k_l)(m+j_1+\cdots+j_q)].
\end{aligned}
$$

It is obvious that term by term (in powers of U^pV^q) that the previous two expressions approach each other as $\hbar\to 0$. The proof that the sums converge is very similar to the analogous proof for the kick map in the previous section. An outline of the results is as follows. We find that

$$
\|H^\dagger T_\hbar(\sigma_{mn})H - T_\hbar(\sigma_{mn}\circ\kappa)\| \leqslant D_1+D_2+D_3,
$$

where (writing $\overline{n}=n+k_1+\cdots+k_l$, $\overline{m}=m+j_1+\cdots+j_q$)

$$
\begin{aligned}
D_1 &= \Bigg\| \sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q} \frac{(2\pi i\epsilon_2 m)^l(2\pi i\epsilon_1\overline{n})^q}{l!q!}\, V^{\overline{n}}U^{\overline{m}} \\
&\quad \times g_{k_1}\cdots g_{k_l} f_{j_1}\cdots f_{j_q}(e^{-\pi^2\hbar(\overline{n}^2+\overline{m}^2+2i\overline{nm})}-1)\Bigg\|,
\end{aligned}
$$

$$
\begin{aligned}
D_2 &= \Bigg\| \sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q} \frac{(2\pi i\epsilon_2 m)^l(2\pi i\epsilon_1\overline{n})^q}{l!q!}\, V^{\overline{n}}U^{\overline{m}} g_{k_1}\cdots g_{k_l} f_{j_1}\cdots f_{j_q} \\
&\quad -\widetilde{g}_{k_1}(m)\cdots\widetilde{g}_{k_l}(m)\widetilde{f}_{j_1}(\overline{n})\cdots\widetilde{f}_{j_q}(\overline{n})\Bigg\|,
\end{aligned}
$$

$$
\begin{aligned}
D_3 &= \Bigg\| \sum_{l=0}^{\infty}\sum_{k_1,\ldots,k_l}\sum_{q=0}^{\infty}\sum_{j_1,\ldots,j_q} \frac{(2\pi i\epsilon_2 m)^l(2\pi i\epsilon_1\overline{n})^q}{l!q!}\, V^{\overline{n}}U^{\overline{m}}\widetilde{g}_{k_1}(m)\cdots\widetilde{g}_{k_l}\widetilde{f}_{j_1}(\overline{n})\cdots\widetilde{f}_{j_q}(\overline{n}) \\
&\quad \times(e^{-\pi^2\hbar(m^2+n^2+2inm)}-1)\Bigg\|.
\end{aligned}
$$

Using (19) once again, we see that D_3 is $O(\lambda)$. To show that D_1 and D_2 go to zero we set

$$
\|g\|_{2,f} := \sum_{k\in\mathbb{Z}} (1+|k|)^2|g_k|\exp(2\pi\epsilon_1 k\|f\|_2) < \infty,
$$

and with these bounds,

$$
\begin{aligned}
D_1 \leqslant & \frac{\lambda}{4} \sum_{l=0}^{\infty} \sum_{k_1,\ldots,k_l} \sum_{q=0}^{\infty} \sum_{j_1,\ldots,j_q} \frac{(2\pi\epsilon_2|m|)^l (2\pi\epsilon_1|\overline{n}|)^q}{l!q!} |g_{k_1}|\cdots|g_{k_l}||f_{j_1}|\cdots|f_{j_q}| \\
& \times (\overline{n}^2 + m^2 + 2m(j_1+\cdots+j_q) + q(j_1^2+\cdots+j_q^2) + 2\overline{n}(m+j_1+\cdots+j_q)) \\
\leqslant & \frac{\lambda}{4} \sum_{l=0}^{\infty} \sum_{k_1,\ldots,k_l} \frac{(2\pi\epsilon_2|m|)^l}{l!} |g_{k_1}|\cdots|g_{k_6}| \exp(2\pi\epsilon_1|\overline{n}|\|f\|_2) \\
& \times (\overline{n}^2 + m^2 + 2|m| + 2\pi\epsilon_1\overline{n}\|f\|_2 + 2|\overline{n}|(|m|+1)) \\
\leqslant & \frac{\lambda}{4} \exp(2\pi\epsilon_1|n|\|f\|_2) \exp(2\pi\epsilon_2 m\|g\|_{2,f}) ((|m|+|n|)^2 + 4(|n|+|m|) + 2\pi\epsilon_2|m|\|g\|_{2,f} \\
& + 2\pi\epsilon_1\|f\|_2(|n|+1)+2) = O(\lambda),
\end{aligned}
$$

while using (21),

$$
\begin{aligned}
D_2 \leqslant & \sum_{l=0}^{\infty} \sum_{k_1,\ldots,k_l} \sum_{q=0}^{\infty} \sum_{j_1,\ldots,j_q} \frac{(2\pi\epsilon_2|m|)^l (2\pi\epsilon_1|\overline{n}|)^q}{l!q!} |g_{k_1}|\cdots|g_{k_l}||f_{j_1}|\cdots|f_{j_q}| \\
& \times \left(\left| 1 - \prod_{r=1}^{l} \left| \frac{\widetilde{g}_{k_r}(m)}{g_{k_r}} \right| \right| + \left| 1 - \prod_{r=1}^{l} \left| \frac{\widetilde{g}_{k_r}(m)}{g_{k_r}} \right| \right| \left| 1 - \prod_{r=1}^{l} \left| \frac{\widetilde{f}_{j_s}(\overline{n})}{f_{j_s}} \right| \right| + \left| 1 - \prod_{r=1}^{l} \left| \frac{\widetilde{f}_{j_s}(\overline{n})}{f_{j_s}} \right| \right| \right) \\
\leqslant & \sum_{l=0}^{\infty} \sum_{k_1,\ldots,k_l} \sum_{q=0}^{\infty} \sum_{j_1,\ldots,j_q} \frac{(2\pi\epsilon_2|m|)^l (2\pi\epsilon_1|\overline{n}|)^q}{l!q!} |g_{k_1}|\cdots|g_{k_l}||f_{j_1}|\cdots|f_{j_q}| (\lambda|m|(|k_1|+\cdots \\
& + |k_l|) + \lambda|\overline{n}|(|j_1|+\cdots+|j_q|)\lambda m(|k_1|+\cdots+|k_l|) + \lambda|\overline{n}|(|j_1|+\cdots+|j_q|)) \\
= & \lambda \sum_{l=0}^{\infty} \sum_{k_1,\ldots,k_l} \frac{(2\pi\epsilon_2|m|)^l}{l!} |g_{k_1}|\cdots|g_{k_l}| \exp(2\pi\epsilon_1\overline{n}\|f\|_2)(|m|(|k_1|+\cdots+|k_l|) \\
& + \lambda|\overline{n}|m(|k_1|+\cdots+|k_l|)) + |\overline{n}| \\
= & \lambda \exp(2\pi\epsilon_1|n|\|f\|_2) \exp(2\pi\epsilon_2|m|\|g\|_{2,f})(|m| + \lambda|m|(|n| + 2\pi\epsilon_2 m\|g\|_{2,f}) \\
& + (|n|+1)) = O(\lambda).
\end{aligned}
$$

We have thus shown that $\|H^\dagger T_\hbar(\sigma_{mn})H - T_\hbar(\sigma_{mn}\circ\eta)\| = O(\lambda)$. The proof is completed for a general symbol $\sigma \in C(\mathbb{T}^2)$ by referring to Step 3 in the proof of the analogous theorem for the kick map. ■

ACKNOWLEDGMENTS

The authors acknowledge the support of the National Science Foundation and Harvard University. A.L. is supported by NSF Grant No. DMS-9424344. R.R. is supported by a NSF Graduate Research Fellowship. N.S. is supported by a Harvard University Graduate Teaching Fellowship.

[1] M. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer, Berlin, 1990)
[2] M. Berry, N. Balazs, M. Tabor, and A. Voros, ''Quantum Maps,'' Ann. Phys. (N.Y.) **122**, 26–63 (1979).
[3] J. H. Hannay and M. Berry, ''Quantization of Linear Maps on a Torus—Fresnel Diffraction of a Periodic Grating,'' Physica D **1**, 267–290 (1980).
[4] S. Klimek and A. Lesniewski, ''Quantizes Chaotic Dynamics and non-commutative KS entropy,'' Ann. Phys. (N.Y.) **248**, 173–198 (1996).
[5] S. Klimek, A. Lesniewski, N. Maitra, and R. Rubin, ''Ergodic Properties of quantized Toral Automorphisms,'' Harvard University, preprint (1995)
[6] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, (1993)
[7] V. Guillemin and S. Sternberg, ''Geometric Asymptotics,'' Math. Surveys **14** (1977).
[8] D. Esposti, ''Quantization of Orientation Preserving Automorphisms of the Torus,'' Ann. Inst. Henri Poincaré Phys. Theor. **58**, 323–341 (1993).
[9] R. Adler and B. Weiss, ''Similarity Automorphisms of the Torus,'' AMS Mem. **98** (1970).
[10] G. Folland, *Harmonic Analysis in Phase Space* (Princeton University Press, Princeton, 1988).
[11] M. Reed and B. Simon, *Functional Analysis* (Academic, New York, 1980), Vol. 1.

Operator product expansions and stress operators

Seichi Naito
Department of Physics, Osaka City University, Sumiyoshiku, Osaka, Japan

(Received 10 February 1997; accepted for publication 29 December 1997)

We explain in detail how to derive the operator product expansions (OPEs) among generating functionals of physical vertex operators (GFPVO) of fermionic particles (i.e., GFPVO in the Ramond sector of superstring) and the stress operators. As for OPEs among GFPVO of bosonic particles (i.e., in the Neveu–Schwarz sector of superstring) and the stress operators, they are those simply obtained by super-generalizing OPEs among GFPVO (in bosonic string) and the stress operator. © *1998 American Institute of Physics.* [S0022-2488(98)01704-6]

I. INTRODUCTION AND PRELIMINARIES

In the preceding paper,[1] we have explained in detail how to construct the generating functionals of physical vertex operators (GFPVO) of both fermionic and bosonic particles (i.e., in both Ramond[2] and Neveu–Schwarz[3] sectors of superstring). However, operator product expansions (OPEs) among GFPVO and the superstressino operator have been given without any proof. Since these OPEs play very important roles in calculating physical scattering amplitudes,[4] these OPEs should be proved explicitly. In this paper, we explain in detail how we can derive these OPEs [i.e., (4.18)–(4.21) and (5.26) in Ref. 1].

In the following, we shall briefly review how GFPVO in the R (NS) sector can be constructed. (See Ref. 1 for more details.) We introduce the light-cone operator $P_F^{\pm j\dagger}(w,\xi)$ $(j=1-4)$ in the R sector defined by

$$P_F^{\pm j\dagger}(w,\xi)\equiv \frac{\xi}{w}\left(p^{\pm j}+\sum_{n=1}^{\infty} w^{-n}\alpha_{-n}^{\pm j}\right)+\frac{1}{\sqrt{w}}\left(d_0^{\pm j}+\sum_{n=1}^{\infty} w^{-n}d_{-n}^{\pm j}\right) \tag{1.1}$$

$$\times(\equiv \xi\Pi_{\alpha}^{\pm j\dagger}(w)+\pi_d^{\pm j\dagger}(w)), \tag{1.1'}$$

where we have used light-cone (mode's) operators (for $j=1-4$) i.e., $p^{\pm j}$ (total energy–momentum operator), $\alpha_m^{\pm j}$'s, $\alpha_{-n}^{\pm j}$'s, $d_m^{\pm j}$'s, and $d_{-n}^{\pm j}$'s, which satisfy the following (anti) commutation relations among themselves;

$$[\alpha_m^{\pm j},\alpha_{-n}^{\mp k}]=m\,\delta_{mn}\delta_{jk} \tag{1.2}$$

and

$$\{d_m^{\pm j},d_{-n}^{\mp k}\}=\delta_{mn}\delta_{jk}, \tag{1.3}$$

all other (anti) commutators being zero.

In the R sector of superstring (i.e., in the *fermionic* sector $s=F$), there exists the Friedan, Martinee, and Schenker (FMS) spinor[5] $S_{\{h\}}(z_0)$ at z_0, which is given by

$$S_{\{h\}}(z_0)\equiv :\exp(1/2\phi^0(z_0)):_{\phi}\times\prod_{j=1}^{4} :\exp(\epsilon_h^j\phi^j(z_0)):_{\phi}, \tag{1.4}$$

$$(\text{with } \epsilon_{\pm}^j=\pm 1/2)$$

Then, any non-N operator $G(z)$ composed of $X^{\pm j}$ and $\psi^{\pm j}$ is N-ized into the N operator $G_N(z)$ by

$$\vec{R}\cdot G(z)S_{\{h\}}(z_0)\equiv G_N(z)\cdot S_{\{h\}}(z_0). \tag{1.5}$$

 © 1998 American Institute of Physics

(Inversely, we state that the N operator G_N is changed into the non-N operator G.) For an example, we have that

$$\begin{aligned}\vec{R}\cdot :\exp(\pm\phi^j(z)):_\phi S_{\{h\}}(z_0) &= (z-z_0)^{\pm\epsilon_h^j} :\exp(\pm\phi^j(z))\cdot S_{\{h\}}(z_0):_\phi \\ &= (z-z_0)^{\pm\epsilon_h^j} :\exp(\pm\phi_{Nh}^j(z)):_\phi S_{\{h\}}(z_0),\end{aligned} \tag{1.6}$$

with $\vec{R}$ being radial ordering. In Eq. (1.6), $\phi_{Nh}^j(z)$ is the operator which has already been normally ordered with respect to the FMS spinor $S_{\{h\}}(z_0)$. From Eq. (1.6), we have found that

$$\Psi^{\pm j}(z) \equiv :\exp(\pm\phi^j(z)):_\phi \overset{N}{\to} \Psi_N^{\pm j}(z) \equiv (z-z_0)^{\pm\epsilon_h^j} :\exp(\pm\phi_{Nh}^j(z)):_\phi , \tag{1.7}$$

$$\Psi_h^{\pm j}(z) \equiv (z-z_0)^{\mp\epsilon_h^j} :\exp(\pm\phi^j(z)):_\phi \overset{N}{\to} \Psi_{Nh}^{\pm j}(z) \equiv :\exp(\pm\phi_{Nh}^j(z)):_\phi , \tag{1.8}$$

and

$$X^{\pm j}(z,\theta) \equiv X^{\pm j}(z) + \theta\cdot\Psi^{\pm j}(z) \overset{N}{\to} X_N^{\pm j}(z,\theta) \equiv X^{\pm j}(z) + \theta\cdot(z-z_0)^{\pm\epsilon_h^j}\cdot\Psi_{Nh}^{\pm j}(z). \tag{1.9}$$

The normal ordering in the R sector is denoted by $: :_R$ and it is defined by

$$\vec{R}\cdot X^{\pm j}(z')X^{\mp k}(z) = \delta_{jk}\log(z'-z) + :X^{\pm j}(z')X^{\mp k}(z):_R \tag{1.10}$$

and

$$\vec{R}\cdot\Psi_{Nh'}^{\pm j}(z')\Psi_{Nh}^{\mp k}(z) = \frac{\delta_{jk}}{z'-z} + :\Psi_{Nh'}^{\pm j}(z')\Psi_{Nh}^{\mp k}(z):_R , \tag{1.11}$$

$\vec{R}$ being radial ordering. In the R sector of superstring (i.e., in fermionic sector $s=F$), GFPVO $\vec{R}\cdot\Phi_{\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ (in non-N operator's form) was proposed to be constructed by

$$\Phi_{\{h\}}^F(z_0,\theta_0;P_F^\dagger) \equiv :\exp(\rho_{N+}^{0F}(z_0,\theta_0) + \rho^{1F}(z_0,\theta_0;P_F^\dagger) + \rho_{\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger)):_R \tag{1.12}$$

$$(\equiv \Phi_{\{h\}}^{F0}(z_0;P_F^\dagger) + \theta_0\Phi_{\{h\}}^{F1}(z_0;P_F^\dagger)). \tag{1.12'}$$

In Eq. (1.12), $: :_R$ is the normal ordering in the R sector and we have used

$$\rho_{N+}^{0F}(z_0,\theta_0) \equiv -ip^{+0}Y_{N+}^{-0}(z_0,\theta_0) - ip^{-0}X^{+0}(z_0), \tag{1.13}$$

$$(\text{with}^1\ \ Y_{N+}^{-0}(z_0,\theta_0) \equiv X^{-0}(z_0) + \theta_0\Psi_{N+}^{-0}(z_0)), \tag{1.13'}$$

$$\rho^{1F}(z_0,\theta_0;P_F^\dagger) \equiv -\sum_{\pm}\sum_{j=1}^{4} i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi)X^{\mp j}(z_F(w,\xi),\theta_F(w,\xi)) \tag{1.14}$$

and

$$\rho^{2F}_{\{h\}}(z_0,\theta_0;P^{\dagger}_F)\equiv-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_F^{-j^{\dagger}}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j^{\dagger}}(w',\xi')$$
$$\times\log\frac{z_F(w,\xi)-z_F(w',\xi')-\left(\frac{z_F(w,\xi)-z_0}{z_F(w',\xi')-z_0}\right)^{\epsilon^j_h}\theta_F(w,\xi)\theta_F(w',\xi')}{w-w'-\left(\frac{w}{w'}\right)^{\epsilon^j_h}\xi\xi'},\tag{1.15}$$

where ϵ^j_h is the constant (i.e., $\epsilon^j_+\equiv+\frac{1}{2}$ and $\epsilon^j_-\equiv-\frac{1}{2}$) in the FMS spinor $S_{\{h\}}(z_0)$ defined by Eq. (1.4). In Eqs. (1.14) and (1.15), we have used various non-N operators

$$\theta_F(w,\xi)\equiv J_F(w)+\xi K_F(w)\tag{1.16}$$

and

$$z_F(w,\xi)-z_0\equiv L_F(w)+\xi K_F(w)J_F(w),\tag{1.17}$$

which are those determined by inversely solving the following superconformal mapping in the R sector $(z_F,\theta_F)\to(w,\xi)$;

$$z_F-z_0-\sqrt{z_F-z_0}\cdot\theta_F\theta_0=w{:}\exp\left(\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}-\frac{X^{+0}(z_0)}{ip^{+0}}\right){:}_R\tag{1.18}$$

and

$$\xi(z_F,\theta_F)=(\partial_{z_F}w(z_F,\theta_F))^{-1/2}D_{\theta_F}w(z_F,\theta_F),\tag{1.19}$$

with $D_{\theta_F}\equiv\partial_{\theta_F}+\theta_F\partial_{z_F}$. Various functions in Eqs. (1.16) and (1.17) can be obtained by using the following theorem.

Theorem: (superconformal mapping in the R sector)

In the case when non-N operators $A^{\pm}(z)$ and $B^{\pm}(z)$ are analytic at $z=z_0$, the functions $F^{\pm}(z_F,\theta_F)$ given by

$$F^{+}(z_F,\theta_F)\equiv A^{+}(z_F)+\theta_F\sqrt{z_F-z_0}B^{+}(z_F)\tag{1.20}$$

and

$$F^{-}(z_F,\theta_F)\equiv\sqrt{z_F-z_0}A^{-}(z_F)+\theta_F B^{-}(z_F)\tag{1.21}$$

can be expressed as the functions of w and ξ by the following formulas:

$$F^{\pm}(z_F,\theta_F)=[F^{\pm}(z_F,\theta_F)]_{\xi=0}+\xi K_F(w)[D_{\theta_F}F^{\pm}(z_F,\theta_F)]_{\xi=0},\tag{1.22}$$

with

$$K_F(w)=\left(\partial_w[z_F-z_0]_{\xi=0}-(\partial_w[\theta_F]_{\xi=0})[\theta_F]_{\xi=0}\right)^{1/2}.\tag{1.23}$$

In Eqs. (1.22) and (1.23), we have used the functions $[\ ]_{\xi=0}$ defined by

$$[A^{+}(z_F)+\theta_F\sqrt{z_F-z_0}B^{+}(z_F)]_{\xi=0}\equiv A^{+}(z_0)+\sum_{n=1}^{\infty}\frac{w^n}{(n-1)!}e^{-nI(z_0)}\lim_{z\to z_0}\left[\partial_z^{n-1}\left(e^{nI(z)}\left(\frac{1}{n}\partial_z A^{+}(z)\right.\right.\right.$$
$$\left.\left.\left.+\left((z-z_0)\frac{\Psi^{+0}_{+}(z;z_0)}{ip^{+0}}+\theta_0\right)B^{+}(z)\right)\right)\right]\tag{1.24}$$

and

$$[\sqrt{z_F-z_0}A^-(z_F)+\theta_F B^-(z_F)]_{\xi=0}$$

$$\equiv\sum_{n=0}^{\infty}\frac{w^{n+1/2}}{n!}e^{-(n+1/2)I(z_0)}\lim_{z\to z_0}\Bigg[\partial_z^n\Bigg(e^{(n+1/2)I(z)}\Bigg(\times\sqrt{z-z_0}\partial_z\Bigg(\sqrt{z-z_0}\,\frac{A^-(z)}{n+1/2}\Bigg)+\Bigg((z-z_0)\,\frac{\Psi_+^{+0}(z;z_0)}{ip^{+0}}+\theta_0\Bigg)B^-(z)\Bigg)\Bigg)\Bigg],\tag{1.25}$$

with the help of shorthand notations

$$I(z)\equiv\frac{X^{+0}(z)}{ip^{+0}},\quad\text{and}\quad\Psi_+^{+0}(z;z_0)\equiv\frac{\Psi^{+0}(z)}{\sqrt{z-z_0}}.\tag{1.26}$$

On the other hand, we introduce the light-cone operator $P_B^{\pm j\dagger}(w,\xi)$ in the NS sector by

$$P_B^{\pm j\dagger}(w,\xi)\equiv\frac{\xi}{w}\left(p^{\pm j}+\sum_{n=1}^{\infty}w^{-n}\alpha_{-n}^{\pm j}\right)+\frac{1}{\sqrt{w}}\left(\sum_{n=1}^{\infty}w^{-n+1/2}b_{-n+1/2}^{\pm j}\right)\tag{1.27}$$

$$(\equiv\xi\Pi_\alpha^{\pm j\dagger}(w)+\pi_b^{\pm j\dagger}(w)).\tag{1.27$'$}$$

where we have used light-cone (mode's) operators (for $j=1-4$), i.e., $p^{\pm j}$ (total energy–momentum operator), $\alpha_m^{\pm j}$'s, $\alpha_{-n}^{\pm j}$'s, $b_{m-1/2}^{\pm j}$'s, and $b_{-n+1/2}^{\pm j}$'s, which satisfy the following (anti)-commutation relations among themselves:

$$[\alpha_m^{\pm j},\alpha_{-n}^{\mp k}]=m\delta_{mn}\delta_{jk}\tag{1.28}$$

and

$$\{b_{m-1/2}^{\pm j},b_{-n+1/2}^{\mp k}\}=\delta_{mn}\delta_{jk},\tag{1.29}$$

all other (anti)commutators being zero. In the NS sector of superstring (i.e., in bosonic sector $s=B$), GFPVO $\Phi^B(z_0,\theta_0;P_B^\dagger)$ was proposed to be constructed by

$$\Phi^B(z_0,\theta_0;P_B^\dagger)\equiv:\exp(\rho^{0B}(z_0,\theta_0)+\rho^{1B}(z_0,\theta_0;P_B^\dagger)+\rho^{2B}(z_0,\theta_0;P_B^\dagger)):_{\rm NS}\tag{1.30}$$

$$(\equiv\Phi^{B0}(z_0;P_B^\dagger)+\theta_0\Phi^{B1}(z_0;P_B^\dagger)).\tag{1.30$'$}$$

In Eq. (1.30), $:\ :_{\rm NS}$ is the normal ordering in the NS sector defined by

$$\vec{R}\cdot X^{\pm j}(z')X^{\mp k}(z)=\delta_{jk}\log(z'-z)+:X^{\pm j}(z')X^{\mp k}(z):_{\rm NS}\tag{1.31}$$

and

$$\vec{R}\cdot\Psi^{\pm j}(z')\Psi^{\mp k}(z)=\frac{\delta_{jk}}{z'-z}+:\Psi^{\pm j}(z')\Psi^{\mp k}(z):_{\rm NS},\tag{1.32}$$

$\vec{R}$ being radial ordering. Furthermore, $\rho^{\iota B}$'s ($\iota=0,1,2$) in Eq. (1.30) are given by

$$\rho^{0B}(z_0,\theta_0)\equiv\sum_{\pm}-ip^{\pm 0}X^{\mp 0}(z_0,\theta_0),\tag{1.33}$$

$$\rho^{1B}(z_0,\theta_0;P_B^\dagger)\equiv-\sum_{\pm}\sum_{j=1}^{4}i\oint\frac{dwd\xi}{2\pi i}P_B^{\pm j\dagger}(w,\xi)X^{\mp j}(z_B(w,\xi),\theta_B(w,\xi))\tag{1.34}$$

and

$$\rho^{2B}(z_0,\theta_0;P_B^\dagger)\equiv-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_B^{+j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_B^{-j^\dagger}(w',\xi')$$
$$\times\log\frac{z_B(w,\xi)-z_B(w',\xi')-\theta_B(w,\xi)\theta_B(w',\xi')}{w-w'-\xi\xi'}. \tag{1.35}$$

In Eqs. (1.34) and (1.35), we have used

$$\theta_B(w,\xi)-\theta_0\equiv J_B(w)+\xi K_B(w) \tag{1.36}$$

and

$$z_B(w,\xi)-z_0-\theta_B(w,\xi)\theta_0\equiv L_B(w)+\xi K_B(w)J_B(w), \tag{1.37}$$

which are those to be determined by inversely solving the following superconformal mapping in the NS sector $(z_B,\theta_B)\to(w,\xi)$:

$$z_B-z_0-\theta_B\theta_0=w{:}\exp\left(\frac{X^{+0}(z_B,\theta_B)}{ip^{+0}}-\frac{X^{+0}(z_0,\theta_0)}{ip^{+0}}\right){:}_{\rm NS} \tag{1.38}$$

and

$$\xi(z_B,\theta_B)=(\partial_{z_B}w(z_B,\theta_B))^{-1/2}D_{\theta_B}w(z_B,\theta_B), \tag{1.39}$$

with $D_{\theta_B}\equiv\partial_{\theta_B}+\theta_B\partial_{z_B}$. {Various function in Eqs. (1.36) and (1.37) can be obtained by using the theorem [superconformal mapping in the NS sector] which has been proved in Ref. 1.}

In Sec. II, we introduce the modified superconformal mapping in the R sector, with the help of which we can derive the contractions (in the R sector) of non-N operators $\theta_F(w,\xi)$ (1.16) and $z_F(w,\xi)$ (1.17) with respect to non-N-operator $\lambda\cdot ip^{+0}Y_+^{-0}(z',\theta';z_0)$. Furthermore, we derive z_0 and θ_0 dependence of non-N operators θ_F (1.16) and z_F (1.17). With the help of various formulas derived in Sec. II, we can derive in Sec. III the operator product expansions (OPEs) among $\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ and the superstressino operator. Especially in Sec. IV, we derive (in the R sector) the next leading terms in OPE among GFPVO and stressino operator. In Sec. V, we discuss our obtained results in the R sector, while derivation of OPEs among GFPVO in the NS sector [i.e., $\Phi^B(z_0,\theta_0;P_B^\dagger)$ (1.30)] and the superstressino operator is only briefly commented upon, since this can be done in a much simpler way. In Appendices A–E, we give various formulas which are necessary in obtaining the final results on OPEs among $\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ (in N-operator's form) and the stress operators.

II. CONTRACTIONS AND z_0 AND θ_0 DEPENDENCE OF z_F AND θ_F

In order to derive the operator product expansions (OPEs) among GFPVO $\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ and the stress operators, we shall need explicit expressions (in the non-N operator's form) for the following contractions in the R sector with respect to $Y_+^{-0}(z',\theta';z_0)$ $(\equiv X^{-0}(z')+\theta'\sqrt{z'-z_0}\Psi^{-0}(z'))$:

$$\lambda\cdot ip^{+0}\overbrace{Y_+^{-0}(z',\theta';z_0)z_F(w,\xi)}^{R}\equiv\delta z_F(w,\xi) \tag{2.1}$$

and

$$\lambda\cdot ip^{+0}\overbrace{Y_+^{-0}(z',\theta';z_0)\theta_F(w,\xi)}^{R}\equiv\delta\theta_F(w,\xi). \tag{2.2}$$

Furthermore, we shall need the following z_0 and θ_0 dependence of non-N operators:

$$\partial_{z_0}z_F-\partial_{z_0}\theta_F\cdot\theta_F,\quad \partial_{z_0}\theta_F \tag{2.3}$$

and

$$D_{\theta_0} z_F - D_{\theta_0}\theta_F \cdot \theta_F, \quad D_{\theta_0}\theta_F, \tag{2.4}$$

with $D_{\theta_0} \equiv \partial_{\theta_0} + \theta_0 \partial_{z_0}$. In the following, we obtain explicit formulas for Eqs. (2.1) and (2.2): First, we have the following contractions (in the R sector) among non-N operators $X^{+0}(z,\theta)$, $X^{+0}(z_0)$, and $\lambda \cdot ip^{+0} Y_+^{-0}(z',\theta';z_0)$:

$$\lambda \cdot ip^{+0} \overbrace{Y_+^{-0}(z',\theta';z_0) X^{+0}(z,\theta)}^{R} \equiv \delta X^{+0}(z,\theta) \tag{2.5}$$

$$= \lambda ip^{+0} \log(z' - z - \sqrt{z - z_0}\,\theta'\theta) \tag{2.5'}$$

and

$$\lambda \cdot ip^{+0} \overbrace{Y_+^{-0}(z',\theta';z_0) X^{+0}(z_0)}^{R} \equiv \delta X^{+0}(z_0) \tag{2.6}$$

$$= \lambda ip^{+0} \log(z' - z_0), \tag{2.6'}$$

provided that any of (z',θ'), (z,θ), and z_0 is c number. Therefore, we modify the superconformal mapping in the R sector (1.18) by the following replacements:

$$X^{+0}(z_F,\theta_F) \to X^{+0}(z_F,\theta_F) + \delta X^{+0}(z_F,\theta_F) \tag{2.7}$$

and

$$X^{+0}(z_0) \to X^{+0}(z_0) + \delta X^{+0}(z_0). \tag{2.8}$$

[In Eq. (2.7), $\delta X^{+0}(z_F,\theta_F)$ is the one obtained from Eq. (2.5′) by replacing c number (z,θ) with q number (z_F,θ_F).] Thus, we are lead to consider the following modified superconformal mapping in the R sector:

$$\begin{aligned} & z_F(w,\xi;\lambda) - z_0 - \sqrt{z_F(w,\xi;\lambda) - z_0}\cdot \theta_F(w,\xi;\lambda)\,\theta_0 \\ & = w : \exp\left(\frac{X^{+0}(z_F(w,\xi;\lambda),\theta_F(w,\xi;\lambda))}{ip^{+0}} - \frac{X^{+0}(z_0)}{ip^{+0}} \right. \\ & \qquad \left. + \lambda \log \frac{z' - z_F(w,\xi;\lambda) - \sqrt{z_F(w,\xi;\lambda) - z_0}\cdot \theta'\,\theta_F(w,\xi;\lambda)}{z' - z_0} \right) :_R, \end{aligned} \tag{2.9}$$

while the superpartner $\xi(z_F,\theta_F;\lambda)$ for $w(z_F,\theta_F;\lambda)$ (2.9) should be determined by

$$\xi = (\partial_{z_F} w(z_F,\theta_F;\lambda))^{-1/2} D_{\theta_F} w(z_F,\theta_f;\lambda). \tag{2.10}$$

In the limit $\lambda \to 0$, Eqs. (2.9) and (2.10) are reduced to Eqs. (1.18) and (1.19), respectively. In this $\lambda = 0$ limit, we have proved[1] the theorem (superconformal mapping in the R sector) i.e., Eqs. (1.20)–(1.25). Similarly, we we can prove the following generalized theorem, which is valid even for $\lambda \neq 0$.

Theorem: (modified superconformal mapping in the R sector)

Non-N operators $F^{\pm}(z_F(w,\xi;\lambda),\theta_F(w,\xi;\lambda))$ can be expressed as the function of w and ξ by using the same functions (1.22) and (1.23), provided that we modify the factors $I(z)$, $I(z_0)$ and $\Psi_+^{+0}(z;z_0)/ip^{+0}$ [which are to be used on the right-hand sides of Eqs. (1.24) and (1.25)] by the following replacements;

$$I(z) \to I(z) + \lambda \log(z' - z), \tag{2.11}$$

$$I(z_0) \to I(z_0) + \lambda \log(z' - z_0), \tag{2.11'}$$

and

$$\frac{\Psi_+^{+0}(z;z_0)}{ip^{+0}} \to \frac{\Psi_+^{+0}(z;z_0)}{ip^{+0}} + \frac{\lambda\,\theta'}{z'-z_0}. \tag{2.12}$$

Expanding thus obtained non-N operators $z_F(w,\xi;\lambda)$ and $\theta_F(w,\xi;\lambda)$ up to the first-order (of λ) term, these operators are expressed by

$$z_F(w,\xi;\lambda) = z_F(w,\xi;0) + \delta z_F(w,\xi) + O(\lambda^2)(\equiv z_F + \delta z_F + O(\lambda^2)) \tag{2.13}$$

and

$$\theta_F(w,\xi;\lambda) = \theta_F(w,\xi;0) + \delta\theta_F(w,\xi) + O(\lambda^2)(\equiv \theta_F + \delta\theta_F + O(\lambda^2)). \tag{2.14}$$

It is remarkable that the first-order (in λ) term δz_F (2.13) and $\delta\theta_F$ (2.14) are found to be equal to those given by Eqs. (2.1) and (2.2), respectively. [This can be seen from the above theorem, especially from Eqs. (2.11), (2.11′), and (2.12).] Furthermore, $(\delta z_F, \delta\theta_F)$ can be calculated explicitly as shown hereafter; First, we notice that

$$\begin{aligned}\delta G &\equiv G(z_F(w,\xi;\lambda),\theta_F(w,\xi;\lambda);\lambda) - G(z_F(w,\xi;0),\theta_F(w,\xi;0);0)\\ &= (\delta\theta_F) D_{\theta_F} G(z_F,\theta_F;0) + (\delta z_F - \delta\theta_F\cdot\theta_F)\partial_{z_F} G(z_F,\theta_F;0)\\ &\quad + \lambda[\partial_\lambda G(z_F(w,\xi;0),\theta_F(w,\xi;0);\lambda)]_{\lambda=0} + O(\lambda^2).\end{aligned} \tag{2.15}$$

Expressing $\log w$ and ξ in terms of z_F, θ_F, z_0, θ_0, and λ with the help of $w \equiv w(z_F,\theta_F;\lambda)$ (2.9), $\xi \equiv \xi(z_F,\theta_F;\lambda)$ (2.10), we apply Eq. (2.15) to $G=\log w$ and $G=\xi$, by noticing that

$$\delta \log w = \delta\xi = 0, \tag{2.16}$$

which result from the fact that both w and ξ are just c numbers, so that they do not have any contraction with $\lambda\cdot ip^{+0}Y_+^{-0}(z',\theta';z_0)$. Then, Eq. (2.16) leads to both

$$(\delta\theta_F) D_{\theta_F} \log w + (\delta z_F - \delta\theta_F\cdot\theta_F)\partial_{z_F} \log w = \lambda \log \frac{z'-z_F-(z_F-z_0)^{1/2}\theta'\,\theta_F}{z'-z_0} \tag{2.17}$$

and

$$\begin{aligned}(\delta\theta_F) D_{\theta_F}\xi + (\delta z_F - \delta\theta_F\cdot\theta_F)\partial_{z_F}\xi &= \lambda w(\partial_{z_F} w)^{-1/2} D_{\theta_F} \log(z'-z_F-(z_F-z_0)^{1/2}\theta'\,\theta_F)\\ &\quad + \lambda\xi\left(\begin{array}{l}\dfrac{1}{2}\log\dfrac{z'-z_F-(z_F-z_0)^{1/2}\theta'\,\theta_F}{z'-z_0}\\ \qquad -\dfrac{\frac{1}{2}w}{\partial_{z_F} w}\,\partial_{z_F}\log(z'-z_F-(z_F-z_0)^{1/2}\theta'\,\theta_F)\end{array}\right).\end{aligned} \tag{2.18}$$

On the other hand, applying

$$D_{\theta_F} = \frac{1}{D_\xi\theta_F}\, D_\xi, \tag{2.19}$$

we find that

$$D_{\theta_F}\log w = \frac{\xi/w}{D_\xi\theta_F}, \quad \partial_{z_F}\log w = \frac{1}{w(D_\xi\theta_F)^2} + \frac{\dfrac{\xi}{w}\,\partial_w\theta_F}{(D_\xi\theta_F)^3},$$

$$D_{\theta_F}\xi=\frac{1}{D_\xi\theta_F},\quad \partial_{z_F}\xi=-\frac{\partial_w\theta_F}{(D_\xi\theta_F)^3},\tag{2.20}$$

and

$$(\partial_{z_F}w)^{-1/2}=D_\xi\theta_F-\tfrac{1}{2}\xi\partial_w\theta_F.\tag{2.20'}$$

Therefore, we have from Eqs. (2.20) and (2.20′) the following identity:

$$(\delta z_F-\delta\theta_F\cdot\theta_F)\frac{1}{w(D_\xi\theta_F)^2}=((\delta\theta_F)D_{\theta_F}\log w+(\delta z_F-\delta\theta_F\cdot\theta_F)\partial_{z_F}\log w)+\frac{\xi}{w}((\delta\theta_F)D_{\theta_F}\xi+(\delta z_F-\delta\theta_F\cdot\theta_F)\partial_{z_F}\xi).\tag{2.21}$$

Substituting Eqs. (2.17) and (2.18) into the right-hand side of Eq. (2.21), we find that

$$\frac{1}{\lambda}(\delta z_F-\delta\theta_F\cdot\theta_F)=w(D_\xi\theta_F)^2\log\frac{z'-z_F-(z_F-z_0)^{1/2}\theta'\theta_F}{z'-z_0}+\xi wK_F^3(w)D_{\theta_F}\log(z'-z_F-(z_F-z_0)^{1/2}\theta'\theta_F),\tag{2.22}$$

where we have used Eq. (1.16). Finally, substituting Eq. (2.22) into the left-hand side of Eq. (2.18), we find that

$$\begin{aligned}\frac{1}{\lambda}(\delta\theta_F)=&(\tfrac{1}{2}\xi D_\xi\theta_F+w\partial_w\theta_F)\log\frac{z'-z_F-(z_F-z_0)^{1/2}\theta'\theta_F}{z'-z_0}\\&+(wK_F^2(w)+\tfrac{1}{2}\xi wK_F(w)\partial_wJ_F(w))D_{\theta_F}\log(z'-z_F-(z_F-z_0)^{1/2}\theta'\theta_F)\\&-\tfrac{1}{2}\xi wK_F^3(w)\partial_{z_F}\log(z'-z_F-(z_F-z_0)^{1/2}\theta'\theta_F).\end{aligned}\tag{2.23}$$

Next, we calculate Eqs. (2.3) and (2.4), which show z_0 and θ_0 dependence of non-N-operators z_F and θ_F; we express non-N-operators θ_F (1.16) and z_F (1.17) by $\theta_F(w,\xi;z_0,\theta_0)$ and $z_F(w,\xi;z_0,\theta_0)$, respectively. If we define

$$\Delta\theta_F\equiv\theta_F(w,\xi;z_0+\Delta z_0,\theta_0+\Delta\theta_0)-\theta_F(w,\xi;z_0,\theta_0)\tag{2.24}$$

and

$$\Delta z_F\equiv z_F(w,\xi;z_0+\Delta z_0,\theta_0+\Delta\theta_0)-z_F(w,\xi;z_0,\theta_0),\tag{2.25}$$

we obtain

$$\begin{aligned}\Delta F(z_F,\theta_F,z_0,\theta_0)\equiv&F(z_F+\Delta z_F,\theta_F+\Delta\theta_F,z_0+\Delta z_0,\theta_0+\Delta\theta_0)-F(z_F,\theta_F,z_0,\theta_0)\\=&(\Delta z_0-\Delta\theta_0\cdot\theta_0)((\partial_{z_0}z_F-\partial_{z_0}\theta_F\cdot\theta_F)\partial_{z_F}+(\partial_{z_0}\theta_F)D_{\theta_F}+\partial_{z_0})F(z_F,\theta_F,z_0,\theta_0)\\&+\Delta\theta_0((D_{\theta_0}z_F-D_{\theta_0}\theta_F\cdot\theta_F)\partial_{z_F}+(D_{\theta_0}\theta_F)D_{\theta_F}+D_{\theta_0})F(z_F,\theta_F,z_0,\theta_0)\\&+O(\Delta^2).\end{aligned}\tag{2.26}$$

Expressing $\log w$ and ξ in terms of z_F, θ_F, z_0, and θ_0 with the help of $w\equiv w(z_F,\theta_F,z_0,\theta_0)$ (1.18) and $\xi\equiv\xi(z_F,\theta_F;z_0,\theta_0)$ (1.19), we can apply Eq. (2.26) to $G=\log w$ and $G=\xi$, by noticing that

$$\Delta\log w=\Delta\xi=0,\tag{2.27}$$

which result from the fact that w and ξ are *independent of* both z_0 and θ_0. In much the same way as conditions (2.16) lead to explicit results (2.22) and (2.23), conditions (2.27) are shown to lead to the explicit results [for Eqs. (2.3) and (2.4)] by the following steps: (First, we should notice that z_0 dependence of non-N operators is *different* from that of N operators.) With the help of w (1.18), ξ (1.19) (given in terms of non-N operators), we calculate $\Delta \log w$ and $\Delta\xi$ by using Eq. (2.26). Then we find from Eq. (2.27) the following four formulas in terms of non-N operators; As for ∂_{z_0}, we have

$$(\partial_{z_0} z_F - \partial_{z_0}\theta_F \cdot \theta_F)\partial_{z_F} \log w + (\partial_{z_0}\theta_F) D_{\theta_F} \log w = \frac{1}{z_F - z_0} + \frac{\frac{1}{2}\theta_F\theta_0}{(\sqrt{z_F - z_0})^3} - \partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}, \tag{2.28}$$

and

$$(\partial_{z_0} z_F - \partial_{z_0}\theta_F \cdot \theta_F)\partial_{z_F}\xi + (\partial_{z_0}\theta_F) D_{\theta_F}\xi = \frac{w}{\sqrt{\partial_{z_F} w}}\left(-\frac{\theta_F}{(z_F - z_0)^2} + \frac{\frac{1}{2}\theta_0}{(\sqrt{z_F - z_0})^3}\right) + \frac{1}{2}\,\xi\left(\begin{array}{l} \dfrac{1}{z_F - z_0} + \dfrac{\frac{1}{2}\,\theta_F\theta_0}{(\sqrt{z_F - z_0})^3} - \partial_{z_0}\dfrac{X^{+0}(z_0)}{ip^{+0}} \\ + wK_F^2(w)\left(\dfrac{1}{(z_F - z_0)^2} + \dfrac{\frac{3}{4}\,\theta_F\theta_0}{(\sqrt{z_F - z_0})^5}\right) \end{array}\right). \tag{2.29}$$

Furthermore, as for D_{θ_0}, we have

$$(D_{\theta_0} z_F - D_{\theta_0}\theta_F \cdot \theta_F)\partial_{z_F} \log w + (D_{\theta_0}\theta_F) D_{\theta_F} \log w = -\frac{\theta_F}{\sqrt{z_F - z_0}} + \frac{\theta_0}{z_F - z_0} - \theta_0\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}} \tag{2.30}$$

and

$$(D_{\theta_0} z_F - D_{\theta_0}\theta_F \cdot \theta_F)\partial_{z_F}\xi + (D_{\theta_0}\theta_F) D_{\theta_F}\xi = \frac{w}{\sqrt{\partial_{z_F} w}}\left(\frac{1}{\sqrt{z_F - z_0}} + \frac{\theta_F\theta_0}{(z_F - z_0)^2}\right) + \frac{1}{2}\,\xi\left(\begin{array}{l} \dfrac{\theta_F}{\sqrt{z_F - z_0}} - \dfrac{\theta_0}{z_F - z_0} + \theta_0\partial_{z_0}\dfrac{X^{+0}(z_0)}{ip^{+0}} \\ + wK_F^2(w)\left(\dfrac{\frac{1}{2}\,\theta_F}{(\sqrt{z_F - z_0})^3} - \dfrac{\theta_0}{(z_F - z_0)^2}\right) \end{array}\right). \tag{2.31}$$

In Eqs. (2.28)–(2.31), we have used Eq. (1.16). First, solving Eqs. (2.28) and (2.29) leads to

$$\partial_{z_0} z_F - \partial_{z_0}\theta_F \cdot \theta_F = w(D_\xi\theta_F)^2\left(\frac{1}{z_F - z_0} + \frac{\frac{1}{2}\theta_F\theta_0}{(\sqrt{z_F - z_0})^3} - \partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right) - \xi w K_F^3(w)\left(\frac{\theta_F}{(z_F - z_0)^2} - \frac{\frac{1}{2}\theta_0}{(\sqrt{z_F - z_0})^3}\right) \tag{2.32}$$

and

$$\partial_{z_0}\theta_F=\left(\frac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\right)\left(\frac{1}{z_F-z_0}+\frac{\frac{1}{2}\theta_F\theta_0}{(\sqrt{z_F-z_0})^3}-\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right)$$
$$+\left(wK_F^2(w)+\frac{1}{2}\,\xi wK_F(w)\partial_w J_F(w)\right)\left(-\frac{\theta_F}{(z_F-z_0)^2}+\frac{\frac{1}{2}\theta_0}{(\sqrt{z_F-z_0})^3}\right)$$
$$+\frac{1}{2}\,\xi wK_F^3(w)\left(\frac{1}{(z_F-z_0)^2}+\frac{\frac{3}{4}\theta_F\theta_0}{(\sqrt{z_F-z_0})^5}\right). \tag{2.33}$$

Finally, Eqs. (2.30) and (2.31) lead to

$$D_{\theta_0}z_F-D_{\theta_0}\theta_F\cdot\theta_F=w(D_\xi\theta_F)^2\left(\frac{\theta_0}{z_F-z_0}-\frac{\theta_F}{\sqrt{z_F-z_0}}-\theta_0\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right)$$
$$-\xi wK_F^3(w)\left(\frac{\theta_F\theta_0}{(z_F-z_0)^2}+\frac{1}{\sqrt{z_F-z_0}}\right) \tag{2.34}$$

and

$$D_{\theta_0}\theta_F=\left(\frac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\right)\left(-\frac{\theta_0}{z_F-z_0}+\frac{\theta_F}{\sqrt{z_F-z_0}}+\theta_0\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right)$$
$$+\left(wK_F^2(w)+\frac{1}{2}\,\xi wK_F(w)\partial_w J_F(w)\right)\left(\frac{\theta_F\theta_0}{(z_F-z_0)^2}+\frac{1}{\sqrt{z_F-z_0}}\right)$$
$$+\frac{1}{2}\,\xi wK_F^3(w)\left(-\frac{\theta_0}{(z_F-z_0)^2}+\frac{\frac{1}{2}\theta_F}{(\sqrt{z_F-z_0})^3}\right). \tag{2.35}$$

III. OPERATOR PRODUCT EXPANSIONS (OPEs) IN THE RAMOND SECTOR

In order to derive the operator product expansions (OPEs) among GFPVO $\vec{R}\cdot\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ and the stress operators, we shall not use the jth component $\mathscr{F}^{jF}(z',\theta')$ of the superstressino operator defined by

$$\mathscr{F}^{jF}(z',\theta')\equiv\tfrac{1}{2}\mathscr{F}^{jF}(z')+\theta' T^{jF}(z') \tag{3.1}$$

$$=\sum_{\pm}1/2{:}\partial_{z'}X^{\pm j}(z',\theta')D_{\theta'}X^{\mp j}(z',\theta'){:}_R\,, \tag{3.1$'$}$$

since the double-valued stressino operator $\mathscr{F}^{jF}(z')$ is not easily treated. [We shall hereafter use the notation $\mathscr{F}^{jF}(z',\theta')$ instead of $T^{jF}(z',\theta')$ in Ref. 1.] We shall use the following (single-valued) new operator $\tau^j(z',\theta';z_0)$ defined by

$$\tau^j(z',\theta';z_0)\equiv\theta' T_F^j(z')+\frac{1}{2}\left(\sqrt{z'-z_0}+\frac{1}{\sqrt{z'-z_0}}\right)\frac{1}{2}\,\mathscr{F}_F^j(z'). \tag{3.2}$$

N operator $\tau_N^j(z',\theta';z_0)$ is obtained from non-N operator $\tau^j(z',\theta';z_0)$ (3.2) by replacing the non-N operator

$$X^{\pm j}(z,\theta)\equiv X^{\pm j}(z)+\theta{:}\exp(\pm\phi^j(z)){:}_\phi \tag{3.3}$$

with the (N-ized) N operator

$$X_N^{\pm j}(z,\theta;z_0)\equiv X^{\pm j}(z)+\theta(z-z_0)^{\pm\epsilon_h^j}{:}\exp(\pm\phi_{Nh}^j(z)){:}_\phi\,. \tag{3.3$'$}$$

This process was called N-izing in Ref. 1. [See Eq. (1.5).]; Non-N operator $\tau^j(z',\theta';z_0)$ is N-ized into N operator $\tau^j_N(z',\theta';z_0)$ by N-izing $X^{\pm j}(z',\theta')$ (3.3) into $X^{\pm j}_N(z',\theta';z_0)$ (3.3′).

First, we can show that

$$\tau^j(z',\theta';z_0)=\tau^j_+(z',\theta';z_0)=\tau^j_-(z',\theta';z_0), \tag{3.4}$$

where the new non-N operator $\tau^j_h(z',\theta';z_0)$ is the one defined by

$$\begin{aligned}\tau^j_h(z',\theta';z_0)&(\equiv \tau^j_h(z',\theta';z_0)_1+\tau^j_h(z',\theta';z_0)_2+\tau^j_h(z',\theta';z_0)_3)\\ &\equiv\sum_{\pm}\ :\tfrac{1}{2}\,\partial_{z'}Y^{\pm j}_h(z',\theta';z_0)D_{\theta'}Y^{\mp j}_h(z',\theta';z_0):_R+\epsilon^j_h\\ &\quad\cdot\frac{\theta'}{z'-z_0}:\Psi^{+j}_h(z';z_0)\Psi^{-j}_h(z';z_0):_R+\sum_{\pm}\ \tfrac{1}{4}\,((z'-z_0)^{\mp 2\epsilon^j_h}-1)\partial_{z'}X^{\pm j}(z')\\ &\quad\cdot\Psi^{\mp j}_h(z';z_0),\quad \text{for } j=0-4 \text{ and } h=\pm.\end{aligned} \tag{3.5}$$

Non-N operator $\tau^j_h(z',\theta';z_0)$ is N-ized [see Eq. (1.5)] into N operator $\tau^j_{Nh}(z',\theta';z_0)$ (for $h=+,-$) by using the following N-izing denoted by $\overset{N}{\rightarrow}$:

$$\begin{aligned}&\text{non-}N\ \text{operators} && \overset{N}{\rightarrow}\ N\ \text{operators}\\ &\Psi^{\pm j}_h(z;z_0)\equiv(z-z_0)^{\mp\epsilon^j_h}:\exp(\pm\phi^j(z)):_\phi && \overset{N}{\rightarrow}\Psi^{\pm j}_{Nh}(z)\equiv:\exp(\pm\phi^j_{Nh}(z)):_\phi\\ &Y^{\pm j}_h(z,\theta;z_0)\equiv X^{\pm j}(z)+\theta\Psi^{\pm j}_h(z;z_0) && \overset{N}{\rightarrow}Y^{\pm j}_{Nh}(z,\theta)\equiv X^{\pm j}(z)+\theta\Psi^{\pm j}_{Nh}(z).\end{aligned} \tag{3.6}$$

Then, we can easily see that τ^j_{Nh} has no branch point at $z'=z_0$. Similar to Eq. (3.4), we can also show that

$$\tau^j_N(z',\theta';z_0)=\tau^j_{N+}(z',\theta';z_0)=\tau^j_{N-}(z',\theta';z_0). \tag{3.7}$$

Since non-N operators $\tau^j_\pm(z',\theta';z_0)$ are single valued, we hereafter investigate radially ordered products among the non-N operator $\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)$ (1.12), FMS spinor $S_{\{h\}}(z_0)$ (1.4), and the (single-valued) non-N-operator $\tau^j_\pm$ (3.5), where τ^j_h should be used when $S_{\{h\}}(z_0)$ has the factor $:\exp(\epsilon^j_h\phi^j(z_0)):_\phi$. By using Wick's theorem, the radially ordered product can be rewritten in terms of the normal ordered products in the R sector (denoted by $:\ :_R$), with the help of the following contractions in the R sector (for $j,k=1-4$):

$$\begin{aligned}&ip^{+0}\overbrace{Y^{-0}_+(z',\theta';z_0)X^{\pm j}(z_F,\theta_F)}^{R}+ip^{+0}\overbrace{Y^{-0}_+(z',\theta';z_0)X^{\pm j}(z_F,\theta_F)}^{R}\\ &=\frac{1}{\lambda}(\delta z_F-\delta\theta_F\cdot\theta_F)\partial_{Z_F}X^{\pm j}(z_F,\theta_F)+\frac{1}{\lambda}(\delta\theta_F)D_{\theta_F}X^{\pm j}(z_F,\theta_F),\end{aligned} \tag{3.8}$$

$$\begin{aligned}&\overbrace{Y^{+0}_+(z',\theta';z_0)Y^{-0}_{N+}(z_0,\theta_0)}^{R}=\log(z'-z_0-\theta'\theta_0),\\ &\overbrace{Y^{-0}_+(z',\theta';z_0)X^{+0}(z_0)}^{R}=\log(z'-z_0)\end{aligned} \tag{3.9}$$

and

$$\overbrace{Y^{\pm j}_h(z',\theta';z_0)X^{\mp k}(z_F,\theta_F)}^{R}=\delta^{jk}\log(z'-z_F-(z_F-z_0)^{\mp\epsilon^j_{h'}}\cdot\theta'\theta_F), \tag{3.10}$$

which are valid when (z',θ') is a c number. [On the right-hand side of Eq. (3.8), we can use Eqs. (2.22) and (2.23).]

In Appendix A, we find the following OPE:

$$\begin{aligned}\vec{R}\cdot\tau_+^0(z',\theta';z_0):&\exp(\rho_{N+}^{0F}(z_0,\theta_0))\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)):_R\\ =:&\left(\tau_+^0(z',\theta';z_0)+G_+^F(z_0,\theta_0|z',\theta')-\sum_{\pm}\sum_{j=1}^{4} i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j\dagger}(w,\xi)G^{\mp jF}(w,\xi|z',\theta')\right)\\ &\times\exp(\rho_{N+}^{0F}(z_0,\theta_0))\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)):_R\end{aligned}\tag{3.11}$$

and those obtained from Eq. (3.11) by

$$\left.\begin{aligned}\tau_+^0(z',\theta';z_0)&\to\tau_+^0(z',\theta';z_0)_\iota,\\ G_+^F(z_0,\theta_0|z',\theta')&\to G_+^F(z_0,\theta_0|z',\theta')_\iota\\ G^{\mp jF}(w,\xi|z',\theta')&\to G^{\mp jF}(w,\xi|z',\theta')_\iota,\end{aligned}\right\}[\text{for }\ \iota=1-3],\tag{3.11$'$}$$

so that we have from Eq. (3.5)

$$\begin{aligned}G_+^F(z_0,\theta_0|z',\theta')&=\sum_{\iota=1}^{3}G_+^F(z_0,\theta_0|z',\theta')_\iota,\\ G^{\mp jF}(w,\xi|z',\theta')&=\sum_{\iota=1}^{3}G^{\mp jF}(w,\xi|z',\theta')_\iota.\end{aligned}\tag{3.11$''$}$$

In Eq. (3.11$'$), non-N operators $G_+^F(\ )_\iota$ and $G^{\mp jF}(\ )_\iota$ (for $\iota=1-3$) are calculated by Eqs. (A1)–(A3) and (A4)–(A6), respectively. Furthermore, non-N operator $G^{\mp jF}$ in OPE (3.11) is given by

$$G^{\mp jF}(w,\xi|z',\theta')=(G^{F1/2}(w,\xi|z',\theta')\partial_{z_F}+G^{F1}(w,\xi|z',\theta')D_{\theta_F})X^{\mp j}(z_F,\theta_F),$$

$$\text{with }\ z_F\equiv z_F(w,\xi)\quad\text{and}\quad\theta_F\equiv\theta_F(w,\xi)\tag{3.12}$$

and we have those obtained from Eq. (3.12) by

$$\left.\begin{aligned}G^{\mp jF}(w,\xi|z',\theta')&\to G^{\mp jF}(w,\xi|z',\theta')_\iota,\\ G^{F(1/2)(1)}(w,\xi|z',\theta')&\to G^{F(1/2)(1)}(w,\xi|z',\theta')_\iota,\end{aligned}\right\}[\text{for }\ \iota=1-3],\tag{3.12$'$}$$

so that we have from Eq. (3.11$''$)

$$G^{F(1/2)(1)}(w,\xi|z',\theta')=\sum_{\iota=1}^{3}G^{F(1/2)(1)}(w,\xi|z',\theta')_\iota.\tag{3.12$''$}$$

In Eq. (3.12$'$), non-N operators $G^{F1/2}(\ )_\iota$ and $G^{F1}(\ )_\iota$ (for $\iota=1-3$) are given by Eqs. (A8)–(A10) and (A11)–(A13), respectively.

In Appendix B, we find that

$$\vec{R}\cdot\tau_h^j(z',\theta';z_0):\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)):_R$$

$$=:\Bigg(\tau_h^j(z',\theta';z_0)-\sum_{\pm}\sum_{j=1}^{4} i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi)g_h^{\mp jF}(w,\xi|z',\theta')$$

$$-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j}(w',\xi')g_h^{jjF}(w,\xi;w',\xi'|z',\theta')\Bigg)$$

$$\times\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)):_R \tag{3.13}$$

and those obtained from Eq. (3.13) by

$$\left.\begin{aligned}\tau_h^j(z',\theta';z_0)&\to\tau_h^j(z',\theta';z_0)_\iota,\\ g_h^{\mp jF}(w,\xi|z',\theta')&\to g_h^{\mp jF}(w,\xi|z',\theta')_\iota\\ g_h^{jjF}(w,\xi;w',\xi'|z',\theta')&\to g_h^{jjF}(w,\xi;w',\xi'|z',\theta')_\iota\end{aligned}\right\}[\text{for }\ \iota=1-3], \tag{3.13$'$}$$

so that we have from Eq. (3.5)

$$g_h^{\mp jF}(w,\xi|z',\theta')=\sum_{\iota=1}^{3}g_h^{\mp jF}(w,\xi|z',\theta')_\iota,$$

$$g_h^{jjF}(w,\xi;w',\xi'|z',\theta')=\sum_{\iota=1}^{3}g_h^{jjF}(w,\xi;w',\xi'|z',\theta')_\iota. \tag{3.13$''$}$$

In Eq. (3.13), non-N operator $g_h^{\mp jF}(\ )_\iota$ [for $\iota=1-3$] is given by Eqs. (B1)–(B3), while $g_h^{jjF}(\ )_\iota$ [for $\iota=1-3$] is given by Eqs. (B4)–(B6).

In Appendix C, we find the following OPE:

$$\vec{R}\cdot\tau_+^0(z',\theta';z_0):\exp(\rho_{\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger))\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R$$

$$=:\Bigg(\tau_+^0(z',\theta';z_0)-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j^\dagger}(w',\xi')$$

$$\times G_h^{jjF}(w,\xi;w',\xi'|z',\theta')\Bigg)\exp(\rho_{\{h\}}^{2F}(z_0,\theta_0;P_F^\dagger))\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R, \tag{3.14}$$

where non-N operator G_h^{jjF} is given by Eq. (C2).

Previously obtained OPEs (3.11), (3.13), and (3.14) will be summarized into OPE among GFPVO $\vec{R}\cdot\Phi_{\{h\}}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$ and

$$\tau(z',\theta';z_0)\equiv\sum_{j=0}^{4}\tau^j(z',\theta';z_0). \tag{3.15}$$

Thus obtained OPE (in non-N operator's form) is given as follows:

$$\vec{R}\cdot\tau(z',\theta';z_0)\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)$$

$$=\vec{R}:\Bigg(\tau(z',\theta';z_0)+G^F_+(z_0,\theta_0|z',\theta')-\sum_{\pm}\sum_{j=1}^{4} i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi)H_h^{\mp jF}(w,\xi|z',\theta')$$

$$-\sum_{j=1}^{4}\oint\frac{dwd\xi}{\pi i}P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j^\dagger}(w',\xi')H_h^{jjF}(w,\xi;w',\xi'|z',\theta')\Bigg)$$

$$\times\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F):_R S_{\{h\}}(z_0), \tag{3.16}$$

where

$$H_h^{\mp jF}=G^{\mp jF}+g_h^{\mp jF}$$

$$H_h^{jjF}\equiv G_h^{jjF}+g_h^{jjF}. \tag{3.17}$$

In Eqs. (3.16) and (3.17), we have used G^F_+ and $G^{\mp jF}$ in OPE (3.11), $g_h^{\mp jF}$ and g_h^{jjF} in OPE (3.13), and G_h^{jjF} in OPE (3.14). On the right-hand side of OPE (3.16), there exist two types of singularities, i.e., type (A): those at $z'=z_F,z'_F$ and type (B): those at $z'=z_0$. For the time being, we shall investigate singularies of type (A); Singular terms of type (A) in F will be denoted by $(F)_{\text{sing.}}$, hereafter. Then, non-N operator $(g_h^{\mp j}(\ ))_{\text{sing.}}$ in OPE (3.13) is expanded as

$$(g_h^{\mp jF}(w,\xi|z',\theta'))_{\text{sing.}}\equiv((g_h^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}\partial_{z_F}+(g_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}D_{\theta_F})X^{\mp j}(z_F,\theta_F) \tag{3.18}$$

and we have also those obtained from Eq. (3.18) by

$$\left.\begin{aligned}(g_h^{\mp jF}(w,\xi|z',\theta'))_{\text{sing.}}&\to(g_h^{\mp jF}(w,\xi|z',\theta')_\iota)_{\text{sing.}}\\ (g_h^{F1(1/2)}(w,\xi|z',\theta'))_{\text{sing.}}&\to(g_h^{F1(1/2)}(w,\xi|z',\theta')_\iota)_{\text{sing.}}\end{aligned}\right\}\ [\text{for}\ \ \iota=1-3]. \tag{3.18$'$}$$

Therefore, we have from Eq. (3.13″)

$$(g_h^{F1(1/2)}(w,\xi|z',\theta'))_{\text{sing.}}=\sum_{i=1}^{3}(g_h^{F1(1/2)}(w,\xi|z',\theta')_\iota)_{\text{sing.}}. \tag{3.18$''$}$$

Non-N operator $(g_h^{F1(1/2)}(\ )_\iota)_{\text{sing.}}$ [for $\iota=1-3$] in Eq. (3.18′) is given by Eqs. (D4)–(D6) [(D7) and (D8)], so that non-N operator $(g_h^{F1(1/2)}(\ ))_{\text{sing.}}$ in Eq. (3.18″) is found out to be given by Eq. (D10) [(D9)]. On the other hand, non-N operator $(g_h^{jjF}(\ )_\iota)_{\text{sing.}}$ (for $\iota=1-3$) in Eq. (3.13′) is given by Eqs. (D12)–(D14), so that $(g_h^{jjF}(\ ))_{\text{sing.}}$ in Eq. (3.13″) is given by Eq. (D15). Therefore, with the help of Eqs. (D9), (D10), and (D15), we find the important relation

$$\begin{aligned}(g_h^{jjF}(w,\xi;w',\xi'|z',\theta'))_{\text{sing.}}=&((g_h^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}\partial_{z_F}+(g_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}D_{\theta_F}\\&+(g_h^{F1/2}(w',\xi'|z',\theta'))_{\text{sing.}}\partial_{z'_F}\\&+(g_h^{F1}(w',\xi'|z',\theta'))_{\text{sing.}}D_{\theta'_F})\log\left(z_F-z'_F-\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\theta_F\theta'_F\right).\end{aligned} \tag{3.19}$$

Therefore, Eq. (3.17) together with Eqs. (3.12) and (3.18) gives

$$(H_h^{\mp jF}(w,\xi|z',\theta'))_{\text{sing.}}\equiv((H_h^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}\partial_{z_F}+(H_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}D_{\theta_F})X^{\mp j}(z_F,\theta_F), \tag{3.20}$$

while Eq. (3.17) together with Eqs. (C2) and (3.19) gives

$$
\begin{aligned}
&(H_h^{jjF}(w,\xi;w',\xi'|z',\theta'))_{\text{sing.}}\\
&\quad=((H_h^{F\,1/2}(w,\xi|z',\theta'))_{\text{sing.}}\partial_{z_F}+(H_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}D_{\theta_F}+(H_h^{F1/2}(w',\xi'|z',\theta'))_{\text{sing.}}\partial_{z'_F}\\
&\qquad+(H_h^{F1}(w',\xi'|z',\theta'))_{\text{sing.}}D_{\theta'_F})\log\left(z_F-z'_F-\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\theta_F\theta'_F\right),
\end{aligned}
\tag{3.21}
$$

where we have

$$
(H_h^{F1(1/2)}(\ ))_{\text{sing.}}\equiv(G^{F1(1/2)}(\ ))_{\text{sing.}}+(g_h^{F1(1/2)}(\ ))_{\text{sing.}}. \tag{3.22}
$$

In Appendix E, we can prove Eqs. (E20) and (E21) among $(G^{F\,1/2(1)}(\ ))_{\text{sing.}}$ and $(g_h^{F\,1/2(1)}(\ ))_{\text{sing.}}$, which lead to the essential fact that $(H_h^{F\,1/2(1)}(\ ))_{\text{sing.}}$ in Eqs. (3.20) and (3.21) is just equal to zero. This fact together with

$$
\theta' T^{0F}(z')=\tau^0(z',\theta';z_0)-\tau^0(z',0;z_0) \tag{3.23}
$$

and

$$
\mathscr{T}^{0F}(z')=4\left(\sqrt{z'-z_0}+\frac{1}{\sqrt{z'-z_0}}\right)^{-1}(1-\theta'\partial_{\theta'})\tau^0(z',\theta';z_0) \tag{3.24}
$$

shows that the radially ordered product among $T_F(\mathscr{T}_F)$, $\Phi^F_{\{h\}}$, and $S_{\{h\}}$ does not have any singularity at $z'=z_F,z'_F$, but it has singularities (*only*) at $z'=z_0$.

Hereafter, we shall investigate the singularities of type (B) (i.e., those at $z'=z_0$) on the right-hand side of the following OPE [which is obtained from (3.11) and (3.14)];

$$
\begin{aligned}
&\vec{R}\cdot\tau_+^0(z',\theta';z_0)\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\\
&\quad=\vec{R}\cdot:\Bigg(\tau_+^0(z',\theta';z_0)+G^F_{N+}(z_0,\theta_0|z',\theta')\\
&\qquad-\sum_{\pm}\sum_{j=1}^{4} i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi)G^{\mp jF}(w,\xi|z',\theta')\\
&\qquad-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j^\dagger}\\
&\qquad\times(w',\xi')G_h^{jjF}(w,\xi;w',\xi'|z',\theta')\Bigg)\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger):_R S_{\{h\}}(z_0).
\end{aligned}
\tag{3.25}
$$

{In OPE (3.25), we have used N operator G_{N+}, which is N-ized [see Eq. (1.5)] from non-N operator G_+ in OPE (3.11).} Then, singular terms at $z'=z_0$ in G^F_{N+} and $G^{F1(1/2)}$ can be obtained after tedious but straightforward calculations, as follows: in the neighborhood of $z'=z_0$, $G^F_{N+}(\ )$ is easily expanded at $z'=z_0$. We can express G^F_{N+} in Eq. (3.25) by

$$
G^F_{N+}(z_0,\theta_0|z',\theta')=\theta'\cdot M^F_{N+}(z_0,\theta_0|z')+N^F_{N+}(z_0,\theta_0|z'). \tag{3.26}
$$

In the neighborhood of $z'=z_0$, the singular terms in M^F_{N+} in Eq. (3.26) are found to be

$$
M^F_{N+}(z_0,\theta_0|z')=-\frac{p^{+0}p^{-0}}{(z'-z_0)^2}+\frac{1}{z'-z_0}(\partial_{z_0}\rho^{0F}_{N+}(z_0,\theta_0)+\tfrac{1}{2}\theta_0\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0))+O(1). \tag{3.27}
$$

Similarly, in the neighborhood of $z'=z_0$, the singular terms in N^F_{N+} in Eq. (3.26) are found out to be

$$4\left(\sqrt{z'-z_0}+\frac{1}{\sqrt{z'-z_0}}\right)^{-1} N^F_{N+}(z_0,\theta_0|z')(\equiv \mathscr{N}^F_{N+}(z_0,\theta_0|z'))$$

$$=\frac{1}{(\sqrt{z'-z_0})^3}\,(p^{+0}p^{-0}\theta_0+\partial_{\theta_0}\rho^{0F}_{N+}(z_0,\theta_0))+\frac{1}{\sqrt{z'-z_0}}$$
$$\times\left(\frac{\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\,p^{+0}p^{-0}-\theta_0\partial_{z_0}\rho^{0F}_{N+}(z_0,\theta_0)-\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0)-\theta_0\partial_{z_0}ip^{-0}X^{+0}(z_0)\right)$$
$$+O(\sqrt{z'-z_0}). \tag{3.28}$$

On the other hand, in order to analyze singular terms (at $z'=z_0$) of non-N operators $G^{F(1/2)(1)}$ {calculated by summing up Eqs. (A8)–(A10) [(A11)–(A13)]}, non-N operator $f_+(z',\theta'|z_0,\theta_0)$ in Eqs. (A8)–(A13) must be N-ized [see Eq. (1.5)] into N operator $f_{N+}(z',\theta'|z_0,\theta_0)$ (A7′), which can be expanded at $z'=z_0$. Thus, we obtain

$$G^{F\,1/2}(w,\xi|z',\theta')\partial_{z_F}+G^{F1}(w,\xi|z',\theta')D_{\theta_F}=\theta'(M^{F\,1/2}(w,\xi|z')\partial_{z_F}+M^{F1}(w,\xi|z')D_{\theta_F})$$
$$+N^{F\,1/2}(w,\xi|z')\partial_{z_F}+N^{F1}(w,\xi|z')D_{\theta_F}, \tag{3.29}$$

where we have found that

$$M^{F\,1/2}(w,\xi|z')\partial_{z_F}+M^{F1}(w,\xi|z')D_{\theta_F}=\frac{1}{(z'-z_0)^2}\,(\tfrac{1}{2}\xi D_\xi+w\partial_w)$$
$$+\frac{1}{z'-z_0}\left(\partial_{z_0}-\frac{\frac{1}{2}\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\,\partial_{\theta_0}\right)+O(1) \tag{3.30}$$

and

$$4\left(\sqrt{z'-z_0}+\frac{1}{\sqrt{z'-z_0}}\right)^{-1}(N^{F\,1/2}(w,\xi|z')\partial_{z_F}+N^{F1}(w,\xi|z')D_{\theta_F})(\equiv \mathscr{N}^{F\,1/2}(w,\xi|z')\partial_{z_F}$$
$$+\mathscr{N}^{F1}(w,\xi|z')D_{\theta_F})$$
$$=\frac{1}{(\sqrt{z'-z_0})^3}\,(-\theta_0(\tfrac{1}{2}\xi D_\xi+w\partial_w)+\partial_{\theta_0})$$
$$+\frac{1}{\sqrt{z'-z_0}}\left(-\frac{\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\,(\tfrac{1}{2}\xi D_\xi+w\partial_w)-\left(\theta_0\partial_{z_0}+\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\,\partial_{\theta_0}\right)\right.$$
$$\left.-\oint_{z_0}\frac{dz}{2\pi i}\,\frac{ip^{+0}}{(z-z_0)^2}\,\Psi^{-0}_{N+}(z)\otimes\right)+O(\sqrt{z'-z_0}). \tag{3.31}$$

Our results have been rewritten into final expressions of Eqs. (3.30) and (3.31), by using the following identities:

$$(D_{\theta_0}z_F-D_{\theta_0}\theta_F\cdot\theta_F)\partial_{z_F}F(z_F,\theta_F)+(D_{\theta_0}\theta_F)D_{\theta_F}F(z_F,\theta_F)=D_{\theta_0}F(z_F,\theta_F) \tag{3.32}$$

and

$$(\partial_{z_0}z_F-\partial_{z_0}\theta_F\cdot\theta_F)\partial_{z_F}F(z_F,\theta_F)+(\partial_{z_0}\theta_F)D_{\theta_F}F(z_F,\theta_F)=\partial_{z_0}F(z_F,\theta_F). \tag{3.33}$$

We remark that D_{θ_0} and ∂_{z_0} on the left-hand sides of Eqs. (3.32) and (3.33) have appeared by using relations (2.32)–(2.35). Furthermore, we have used the identity

$$w(D_\xi\theta_F)^2\partial_{z_F}F(z_F,\theta_F)+(\tfrac{1}{2}\xi D_\xi\theta_F+w\partial_w\theta_F)D_{\theta_F}F(z_F,\theta_F)=(\tfrac{1}{2}\xi D_\xi+w\partial_w)F(z_F,\theta_F), \tag{3.34}$$

together with the following formulas in non-N operator's form:

$$\begin{aligned}\partial_{\theta_0}z_F-\partial_{\theta_0}\theta_F\cdot\theta_F&=(D_{\theta_0}z_F-D_{\theta_0}\theta_F\cdot\theta_F)-\theta_0(\partial_{z_0}z_F-\partial_{z_0}\theta_F\cdot\theta_F)\\&=-w(D_\xi\theta_F)^2\,\frac{\theta_F}{\sqrt{z_F-z_0}}-\frac{\xi wK_F^3(w)}{\sqrt{z_F-z_0}}\end{aligned} \tag{3.35}$$

and

$$\begin{aligned}\partial_{\theta_0}\theta_F&=D_{\theta_0}\theta_F-\theta_0(\partial_{z_0}\theta_F)=(\tfrac{1}{2}\xi D_\xi\theta_F+w\partial_w\theta_F)\,\frac{\theta_F}{\sqrt{z_F-z_0}}+(wK_F^2(w)\\&\quad+\tfrac{1}{2}\xi wK_F(w)\partial_wJ_F(w))\,\frac{1}{\sqrt{z_F-z_0}}+\frac{1}{4}\,\xi wK_F^3(w)\,\frac{\theta_F}{(\sqrt{z_F-z_0})^3},\end{aligned} \tag{3.36}$$

which result from Eqs. (2.32) to (2.35). Moreover, we have defined the operator $A\otimes$ which operates on adjacent operator B as

$$A\otimes B\equiv\overbrace{A\quad B}^{R}, \tag{3.37}$$

so that we have

$$\begin{aligned}&-\oint_{z_0}\frac{dz}{2\pi i}\frac{ip^{+0}}{(z-z_0)^2}\;\Psi_{N+}^{-0}(z)\otimes X^{\mp j}(z_F,\theta_F)\\&\equiv-\oint_{z_0}\frac{dz}{2\pi i}\frac{ip^{+0}}{(z-z_0)^2}\,\overbrace{\Psi_{N+}^{-0}(z)X^{\mp j}}^{R}(z_F,\theta_F)\\&=\left(-w(D_\xi\theta_F)^2\,\frac{\theta_F}{(\sqrt{z_F-z_0})^3}-\xi\,\frac{wK_F^3(w)}{(\sqrt{z_F-z_0})^3}\right)\partial_{z_F}X^{\mp j}(z_F,\theta_F)\\&\quad+\left((\tfrac{1}{2}\xi D_\xi\theta_F+w\partial_w\theta_F)\,\frac{\theta_F}{(\sqrt{z_F-z_0})^3}+(wK_F^2(w)+\tfrac{1}{2}\,\xi wK_F(w)\partial_wJ_F(w))\,\frac{1}{(\sqrt{z_F-z_0})^3}\right.\\&\quad\left.+\frac{3}{4}\,\xi wK_F^3(w)\,\frac{\theta_F}{(\sqrt{z_F-z_0})^5}\right)D_{\theta_F}X^{\mp j}(z_F,\theta_F)\end{aligned} \tag{3.38}$$

resulting from contractions in the R sector (2.22) and (2.23).

At this stage, it should be noticed that light-cone operators α's and d's in $P_F^{\pm j}$ (1.1) satisfy the following (anti) commutation relations:

$$[\alpha_m^{\pm j},\alpha_{-n}^{\mp k}]=m\,\delta_{mn}\delta^{jk}\quad\text{for }\ j,k=1-4\quad\text{and }\ n,m=1,2,3,\ldots. \tag{3.39}$$

and

$$\{d_m^{\pm j},d_{-n}^{\mp k}\}=\delta_{mn}\delta^{jk}\quad\text{for }\ j,k=1-4\quad\text{and }\ n,m=0,1,2,\ldots. \tag{3.40}$$

Then, the light-cone number operator N_F defined by

$$N_F \equiv \sum_{n=1}^{\infty} \sum_{\pm} \sum_{k=1}^{4} \left(\alpha_{-n}^{\mp k} \alpha_n^{\pm k} + n d_{-n}^{\mp k} d_n^{\pm k} \right) \tag{3.41}$$

satisfies the following commutation relation:

$$[N_F, P_F^{\pm j\dagger}(w,\xi)] = -\partial_w (w P_F^{\pm j\dagger}(w,\xi)) + \tfrac{1}{2} \partial_\xi (\xi P_F^{\pm j\dagger}(w,\xi)). \tag{3.42}$$

Furthermore, we have the identity

$$1 = (\tfrac{1}{2}\, \xi \partial_\xi + w \partial_w + \tfrac{1}{2}\, \xi' \partial_{\xi'} + w' \partial_{w'}) \log \left(w - w' - \left(\frac{w'}{w} \right)^{\pm 1/2} \xi \xi' \right). \tag{3.43}$$

With the help of Eqs. (3.42) and (3.43), we obtain the following relations:

$$-\sum_{\pm} \sum_{j=1}^{4} i \oint \frac{dw d\xi}{2\pi i} P_F^{\pm j\dagger}(w,\xi) (\tfrac{1}{2}\, \xi \partial_\xi + w \partial_w) X^{\mp j}(z_F, \theta_F) = [N_F, \rho^{1F}(z_0, \theta_0; P_F^\dagger)] \tag{3.44}$$

and

$$\begin{aligned} &-\sum_{j=1}^{4} \oint \frac{dw d\xi}{2\pi i} P_F^{-j\dagger}(w,\xi) \oint \frac{dw' d\xi'}{2\pi i} P_F^{+j\dagger}(w',\xi') \\ &\quad \times (\tfrac{1}{2}\, \xi \partial_\xi + w \partial_w + \tfrac{1}{2}\, \xi' \partial_{\xi'} + w' \partial_{w'}) \log \left(z_F - z_F' - \left(\frac{z_F - z_0}{z_F' - z_0} \right)^{\epsilon_h^j} \theta_F \theta_F' \right) \\ &\quad = [N_F, \rho_{\{h\}}^{2F}(z_0, \theta_0; P_F^\dagger)] - \left(\sum_{j=1}^{4} p^{+j} p^{-j} \right). \end{aligned} \tag{3.45}$$

Since we have $T^{OF}(z')$ by using Eq. (3.23), we finally obtain [from Eq. (3.25) substituted with both Eq. (3.27) in Eq. (3.26) and Eq. (3.30) in Eq. (3.29)] the following operator product expansion (OPE):

$$\begin{aligned} \vec{R} \cdot T^{0F}(z') \Phi_{\{h\}}^F(z_0, \theta_0; P_F^\dagger) S_{\{h\}}(z_0) &= \frac{1}{(z' - z_0)^2} \vec{R} \cdot \left(-\left(\sum_{j=0}^{4} p^{+j} p^{-j} \right) \Phi_{\{h\}}^F(z_0, \theta_0; P_F^\dagger) S_{\{h\}}(z_0) \right. \\ &\quad \left. + [N_F, \Phi_{\{h\}}^F(z_0, \theta_0; P_F^\dagger) S_{\{h\}}(z_0)] \right) + \frac{1}{z' - z_0} \vec{R} \\ &\quad \cdot (\partial_{z_0} \Phi_{\{h\}}^F(z_0, \theta_0; P_F^\dagger)) S_{\{h\}}(z_0) + \frac{1}{z' - z_0} \vec{R} \\ &\quad \cdot : \left(\tfrac{1}{2} \Psi_{N+}^{+0}(z_0) \Psi_{N+}^{-0}(z_0) + \tfrac{1}{2} \theta_0 (\partial_{z_0} i p^{+0} \Psi_{N+}^{-0}(z_0)) \right. \\ &\quad \left. - \frac{\tfrac{1}{2}\, \Psi_{N+}^{+0}(z_0)}{i p^{+0}} \partial_{\theta_0} (\rho^{1F}(z_0, \theta_0; P_F^\dagger) + \rho_{\{h\}}^{2F}(z_0, \theta_0; P_F^\dagger)) \right) \\ &\quad \times \Phi_{\{h\}}^F(z_0, \theta_0; P_F^\dagger) :_R S_{\{h\}}(z_0) + O(1), \end{aligned} \tag{3.46}$$

where $S_{\{h\}}(z_0)$ is the FMS spinor (1.4). In the following, we shall prove that the last term on the right-hand side of Eq. (3.46) can be rewritten into the term on the left-hand side of the following Eq. (3.48):

Proof: Since we have

$$:\exp(-\phi^0_{N+}(z_0)):_\phi :(\partial_{z_0}\phi^0(z_0)\exp(\tfrac{1}{2}\phi^0(z_0))):_\phi$$

$$=:(\exp(-\phi^0(z_0))\partial_{z_0}\phi^0(z_0)\exp(\tfrac{1}{2}\phi^0(z_0))):_\phi$$

$$=:(\exp(-\phi^0_{N+}(z_0))\partial_{z_0}\phi^0_{N+}(z_0)):_\phi :\exp(\tfrac{1}{2}\phi^0(z_0)):_\phi , \tag{3.47}$$

we find that

$$\frac{1}{z'-z_0}\vec{R}\cdot :\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)(\partial_{z_0}:\exp(\tfrac{1}{2}\phi^0(z_0)):_\phi)\left(\prod_{j=1}^{4}:\exp(\epsilon_h^j\phi^j(z_0)):_\phi\right)$$

$$=\frac{1}{z'-z_0}\vec{R}\cdot :(\exp(\rho^{0F}_{N+}(z_0,\theta_0))\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)):_\phi :(\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)$$
$$+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))):_R S_{\{h\}}(z_0), \tag{3.48}$$

the right-hand side of which can be calculated as follows. First, we notice that

$$\vec{R}\cdot :(\exp(\rho^{0F}_{N+}(z_0,\theta_0))\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)):_\phi :(\exp(\rho^{1F}(z_0,\theta_0;P_F^\dagger)+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))):_R$$
$$=:(\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)\exp(\rho^{0F}_{N+}(z_0,\theta_0)+\rho^{1F}(z_0,\theta_0;P_F^\dagger)+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))):_R+:((\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)$$
$$\otimes(\rho^{1F}(z_0,\theta_0;P_F^\dagger)+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger)))\exp(\rho^{0F}_{N+}(z_0,\theta_0)+\rho^{1F}(z_0,\theta_0;P_F^\dagger)$$
$$+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))):_R , \tag{3.49}$$

where we have used the operation $\otimes$ defined by Eq. (3.37). With the help of

$$:\Psi^{+j}_{Nh}(z')\Psi^{-j}_{Nh}(z'):_R=\partial_{z'}\phi^j_{Nh}(z') \quad \text{for } j=0-4, \tag{3.50}$$

we find in the second term on the right-hand side of Eq. (3.49) that

$$\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)\otimes(\rho^{1F}(z_0,\theta_0;P_F^\dagger)+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))$$

$$=:\tfrac{1}{2}\Psi^{+0}_{N+}(z_0)\left(\overbrace{\Psi^{-0}_{N+}(z_0)0\rho^{1F}}^{R}(z_0,\theta_0;P_F^\dagger)+\overbrace{\Psi^{-0}_{N+}(z_0)\rho^{2F}_{\{h\}}}^{R}(z_0,\theta_0;P_F^\dagger)\right):_R . \tag{3.51}$$

By calculating the contractions in the R sector [i.e., $ip^{+0}\Psi^{-0}_{N+}(z_0)\otimes$ in Eq. (3.51)] with the help of Eqs. (2.22) and (2.23), Eq. (3.51) is reduced to

$$-:\left(\frac{\tfrac{1}{2}\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\,\partial_{\theta_0}(\rho^{1F}(z_0,\theta_0;P_F^\dagger)+\rho^{2F}_{\{h\}}(z_0,\theta_0;P_F^\dagger))\right):_R , \tag{3.51$'$}$$

where we have used the partial derivative ∂_{θ_0} given by Eqs. (3.35) and (3.36). Furthermore, we have from Eqs. (1.13) and (3.47)

$$:(\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)\exp(\rho^{0F}_{N+}(z_0,\theta_0))):_R=:\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)(1-\theta_0 ip^{+0}\exp(-\phi^0_{N+}(z_0)))$$
$$\times\exp\left(-\sum_{\pm} ip^{\pm 0}X^{\mp 0}(z_0)\right):_R$$
$$=:\exp\left(-\sum_{\pm} ip^{\pm 0}X^{\mp 0}(z_0)\right)$$
$$\times(\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0)+\tfrac{1}{2}\theta_0\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0)):_R. \tag{3.52}$$

On the other hand, we have from Eq. (1.3)

$$:\exp(\rho^{0F}_{N+}(z_0,\theta_0))\cdot\tfrac{1}{2}\Psi^{+0}_{N+}(z_0)\Psi^{-0}_{N+}(z_0):_R=:\exp\left(-\sum_{\pm} ip^{\pm 0}X^{\mp 0}(z_0)\right)\tfrac{1}{2}\Psi^{+0}_{N+}(z_0)\Psi^{-0}_{N+}(z_0):_R$$
$$=:\exp\left(-\sum_{\pm} ip^{\pm 0}X^{\mp 0}(z_0)\right)\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0):_R, \tag{3.53}$$

which is just equal to the first term on the right-hand side of Eq. (3.52). Therefore, we find from Eqs. (3.52) and (3.53)

$$:\exp(\rho^{0F}_{N+}(z_0,\theta_0))\tfrac{1}{2}\partial_{z_0}\phi^0_{N+}(z_0):_R=:\exp(\rho^{0F}_{N+}(z_0,\theta_0))(\tfrac{1}{2}\Psi^{+0}_{N+}(z_0)\Psi^{-0}_{N+}(z_0)$$
$$+\tfrac{1}{2}\theta_0\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0)):_R. \tag{3.54}$$

Substituting Eq. (3.51′) [(3.54)] into the second (first) term on the right-hand side of Eq. (3.49), we find that the right-hand side of Eq. (3.48) is just equal to the last term on the right-hand side of Eq. (3.46). (Q.E.D.)

On the other hand, we find

$$\vec{R}\cdot\left(\sum_{j=1}^{4} T^{jF}(z')\right)\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)$$
$$=\frac{1}{z'-z_0}\vec{R}\cdot\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F):\exp(\tfrac{1}{2}\phi^0(z_0)):_\phi\partial_{z_0}$$
$$\times\left(\prod_{j=1}^{4}:\exp(\epsilon^j_h\phi^j(z_0)):_\phi\right)+O(1). \tag{3.55}$$

In summing Eqs. (3.46) and (3.55), the following factor appears:

$$\vec{R}\cdot(\partial_{z_0}\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F))S_{\{h\}}(z_0)+\vec{R}\cdot\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)(\partial_{z_0}S_{\{h\}}(z_0))$$
$$=\partial_{z_0}(\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)). \tag{3.56}$$

We should notice that we must use non-N operator $\Phi^F_{\{h\}}(z_0,\theta_0;P^\dagger_F)$ in each term on the left-hand side of Eq. (3.56), while we can use N operator $\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)$ on the right-hand side of Eq. (3.56). With respect to $T_{NF}(z';z_0)$ which is N-ized [see Eq. (1.5)] from

$$T_F(z')\equiv\sum_{j=0}^{4} T^{jF}(z') \tag{3.57}$$

[see Eq. (3.1)], we finally find the following OPE in N operator's form:

$$\vec{R}\cdot T_{NF}(z';z_0)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)=\frac{1}{(z'-z_0)^2}\left(-\left(\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\right.$$
$$\left.+[N_F,\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)]\right)$$
$$+\frac{1}{z'-z_0}\,\partial_{z_0}(\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0))+O(1). \tag{3.58}$$

With the help of Eq. (3.24), the following OPE [with respect to the stressino operator $\mathscr{F}^{0F}(z')$] is obtained (in non-N operator's form) by substituting both N^F_{N+} (3.28) (in Eq. (3.26)) and $N^{F1/2(1)}$ (3.31) [in Eq. (3.30)] into OPE (3.26);

$$\vec{R}\cdot\mathscr{F}^{0F}(z')\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)$$
$$=-\frac{\theta_0}{(\sqrt{z'-z_0})^3}\,\vec{R}$$
$$\cdot\left(-\left(\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)+[N_F,\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)]\right)$$
$$+\frac{1}{(\sqrt{z'-z_0})^3}\,\vec{R}\cdot\partial_{\theta_0}\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)+O\left(\frac{1}{\sqrt{z'-z_0}}\right). \tag{3.59}$$

On the other hand, we have

$$\vec{R}\cdot\left(\sum_{j=1}^{4}\mathscr{F}^{jF}\right)\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)=O\left(\frac{1}{\sqrt{z'-z_0}}\right). \tag{3.60}$$

With respect to $\mathscr{F}_{NF}(z';z_0)$ which is N-ized [see Eq. (1.5)] from

$$\mathscr{F}_F(z')\equiv\sum_{j=0}^{4}\mathscr{F}^{jF}(z') \tag{3.61}$$

[see Eq. (3.1)], we finally find the following OPE in N operator's form;

$$\vec{R}\cdot\mathscr{F}_{NF}(z';z_0)\Phi^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)=\frac{1}{(\sqrt{z'-z_0})^3}\,\Phi^{F1}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)+O\left(\frac{1}{\sqrt{z'-z_0}}\right) \tag{3.62}$$

and

$$\vec{R}\cdot\mathscr{F}_{NF}(z';z_0)\Phi^{F1}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)=\frac{1}{(\sqrt{z'-z_0})^3}\left(-\left(\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)\right.$$
$$\left.+[N_F,\Phi^{F0}_{N\{h\}}(z_0;P_F^\dagger)S_{\{h\}}(z_0)]\right)+O\left(\frac{1}{\sqrt{z'-z_0}}\right). \tag{3.63}$$

IV. NEXT LEADING TERMS IN OPEs (3.59)–(3.63)

In calculating physical scattering amplitudes among external physical states including more than six Ramond particles, we will find from FMS's work that that we need to picture change[5] some physical Ramond states in 1/2 picture [later given by Eq. (5.12)] into those in 3/2 picture. These picture-changing calculations can be carried out by using the N operator $\hat{\Phi}^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)$ which we have in the following OPE:

$$\begin{aligned}\vec{R}\cdot\mathscr{F}_F(z')\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)=&-\frac{\theta_0}{(\sqrt{z'-z_0})^3}\left(-\left(\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\right.\\&\left.+[N_F,\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)]\right)+\frac{1}{(\sqrt{z'-z_0})^3}\\&\cdot\partial_{\theta_0}\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)-\frac{\theta_0}{\sqrt{z'-z_0}}\\&\cdot\partial_{z_0}(\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0))\\&+\frac{1}{\sqrt{z'-z_0}}\hat{\Phi}^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)+O(\sqrt{z'-z_0}).\end{aligned}\tag{4.1}$$

In the following, the explicit formula of the N operator $\hat{\Phi}^F_{N\{h\}}$ [in OPE (4.1)] will be derived by using various formulas given in Sec. III: We have investigated the OPE among GFPVO and the jth stressino operator $\mathscr{F}^{jF}(z')$ defined by

$$\mathscr{F}^{jF}(z')\equiv\sum_{\pm}\partial_{z'}X^{\pm j}(z')\cdot\Psi^{\mp j}(z')\quad(\text{for } j=0-4).\tag{4.2}$$

Then, we find from Eqs. (3.24) to (3.26), (3.12), and (C2) that

$$\begin{aligned}&\vec{R}\cdot\mathscr{F}^{0F}(z')\cdot\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\\&=\vec{R}\cdot\left(\mathscr{F}^{0F}(z')+\mathscr{N}^F_{N+}(z_0,\theta_0|z')-\sum_{\pm}\sum_{j=1}^{4}i\oint\frac{dwd\xi}{2\pi i}P_F^{\pm j^\dagger}(w,\xi)\right.\\&\times((\mathscr{N}^{F\,1/2}(w,\xi|z')\partial_{z_F}+\mathscr{N}^{F1}(w,\xi|z')D_{\theta_F})X^{\mp j}(z_F,\theta_F))-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}P_F^{-j^\dagger}(w,\xi)\\&\times\oint\frac{dw'd\xi'}{2\pi i}P_F^{+j^\dagger}(w',\xi')(\mathscr{N}^{F\,1/2}(w,\xi|z')\partial_{z_F}+\mathscr{N}^{F1}(w,\xi|z')D_{\theta_F}+\mathscr{N}^{F\,1/2}(w',\xi'|z')\partial_{z'_F}\\&\left.+\mathscr{N}^{F1}(w',\xi'|z')D_{\theta'_F})\log\left(z_F-z'_F-\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\cdot\theta_F\theta'_F\right)\right)\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0).\end{aligned}\tag{4.3}$$

In the neighborhood of $z'=z_0$, the singular terms in $\mathscr{N}^F_{N+}$ are found from Eq. (3.28), while the singular terms in $\mathscr{N}^{F\,1/2}$ and $\mathscr{N}^{F1}$ in OPE (4.3) are found from Eq. (3.31). In Eq. (4.3), we have used the operator $\Psi^{-0}_{N+}(z)\otimes$ which operates on adjacent operator B as Eq. (3.37) and we have found Eq. (3.38).

On the other hand, we find from Eqs. (2.19) and (1.16) that

$$D_{\theta_F}=\left(\frac{1}{K_F(w)}-\frac{\xi\partial_w J_F(w)}{K_F^2(w)}\right)D_\xi\tag{4.4}$$

and

$$\partial_{z_F}=D_{\theta_F}\cdot D_{\theta_F}=-\frac{1}{K_F^3(w)}\,(\partial_w J_F(w)+\xi\partial_w K_F(w))D_\xi+\frac{1}{K_F^2(w)}\left(1-\frac{2\xi\partial_w J_F(w)}{K_F(w)}\right)\partial_w\,. \tag{4.5}$$

Substituting Eqs. (4.4), (4.5), and (1.16) into Eq. (3.38), we find

$$\begin{aligned}
&-\oint_{z_0}\frac{dz}{2\pi i}\,\frac{ip^{+0}}{(z-z_0)^2}\,\Psi_{N+}^{-0}(z)\otimes(\equiv\mathscr{H}_F(w,\xi))\\
&\quad=-\frac{\theta_F}{(\sqrt{z_F-z_0})^3}\,(\tfrac{1}{2}\xi D_\xi+w\partial_w)+\frac{wK_F(w)}{(\sqrt{z_F-z_0})^3}\,\partial_\xi\\
&\qquad+\xi\left(\tfrac{1}{2}w\partial_w J_F(w)\cdot\frac{1}{(\sqrt{z_F-z_0})^3}+\frac{3}{4}\,wK_F^2(w)\,\frac{\theta_F}{(\sqrt{z_F-z_0})^5}\right)\partial_\xi\,.
\end{aligned}\tag{4.6}$$

Then, OPE (4.3) substituted with Eqs. (3.28), (3.31), and (4.6) leads to

$$\begin{aligned}
\vec{R}\cdot\mathscr{F}^{0F}(z')\cdot\Phi_{\{h\}}^F(z_0\theta_0;P_F^\dagger)S_{\{h\}}(z_0)=&\frac{1}{(\sqrt{z'-z_0})^3}\,(\cdots)-\frac{\theta_0}{\sqrt{z'-z_0}}\cdot\vec{R}\cdot(\partial_{z_0}\Phi_{\{h\}}^F(z_0,\theta_0;P_F^\dagger))S_{\{h\}}\\
&\times(z_0)+\frac{1}{\sqrt{z'-z_0}}\,\hat{\Phi}_{N\{h\},1}^F(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)\\
&+O(\sqrt{z'-z_0}),
\end{aligned}\tag{4.7}$$

with

$$\begin{aligned}
&\hat{\Phi}_{N\{h\},1}^F(z_e,\theta_0;P_F^\dagger)\\
&=-:\frac{\Psi_{N+}^{+0}(z_0)}{ip^{+0}}\left(\left(-\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)+[N_F,\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger)]\right):_R\\
&\quad-:\left(\partial_{z_0}\frac{X^{+0}(z_0)}{ip^{+0}}\right)\partial_{\theta_0}\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger):_R-:(\partial_{z_0}ip^{+0}\Psi_{N+}^{-0}(z_0)\\
&\quad+\theta_0\partial_{z_0}ip^{-0}X^{+0}(z_0)+\partial_{z_0}X^{+0}(z_0)\cdot\Psi_{N+}^{-0}(z_0))\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger):_R\\
&\quad-\sum_{\pm}\sum_{j=1}^{4}i\oint\frac{dwd\xi}{2\pi i}\,P_F^{\pm j^\dagger}(w,\xi):(\mathscr{H}_{NF}(w,\xi)X_N^{\mp j}(z_{NF},\theta_{NF}))\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger):_R\\
&\quad-\sum_{j=1}^{4}\oint\frac{dwd\xi}{2\pi i}\,P_F^{-j^\dagger}(w,\xi)\oint\frac{dw'd\xi'}{2\pi i}\,P_F^{+j^\dagger}(w',\xi'):\\
&\quad\times\left((\mathscr{H}_{NF}(w,\xi)+\mathscr{H}_{NF}(w',\xi'))\log\left(z_{NF}-z'_{NF}-\left(\frac{z_{NF}-z_0}{z'_{NF}-z_0}\right)^{\epsilon_h^j}\cdot\theta_{NF}\theta'_{NF}\right)\right)\\
&\quad\times\Phi_{N\{h\}}^F(z_0,\theta_0;P_F^\dagger):_R+O(\sqrt{z'-z_0}).
\end{aligned}\tag{4.8}$$

In Eq. (4.8), $\mathscr{H}_{NF}(w,\xi)$ is the N operator which is N-ized [see Eq. (1.5)] from $\mathscr{H}_F(w,\xi)$ (4.6), so that it is given by

$$\mathscr{H}_{NF}(w,\xi)=-\frac{J_{NF}(w)+\xi K_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\,(\tfrac{1}{2}\xi D_\xi+w\partial_w)+\frac{wK_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\cdot\partial_\xi$$
$$+\xi\left(\frac{\frac{1}{2}w\partial_w J_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}+\frac{3}{4}\,wK_{NF}^2(w)\,\frac{J_{NF}(w)}{(\sqrt{z_{NF}(w,\xi)-z_0})^5}\right)\partial_\xi. \tag{4.9}$$

We should notice that the right-hand side of OPE (4.7) still has the non-N operator $\vec{R}\cdot(\partial_{z_0}\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger))S_{\{h\}}(z_0)$, which can be expressed in terms of N operators as follows: First, we find from Eqs. (3.47) to (3.56) [together with Eq. (1.13)] that

$$\vec{R}\cdot(\partial_{z_0}\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger))S_{\{h\}}(z_0)=-\vec{R}$$
$$\cdot(\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger):\exp(\tfrac{1}{2}\phi^0(z_0)):)\partial_{z_0}\left(\prod_{j=1}^4:\exp(\epsilon_h^j\phi^j(z_0)):\right)$$
$$-:\Bigg(\tfrac{1}{2}\theta_0(\partial_{z_0}ip^{+0}\Psi^{-0}_{N+}(z_0))$$
$$-\frac{\frac{1}{2}\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\,\partial_{\theta_0}\Bigg)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R S_{\{h\}}(z_0)$$
$$+\partial_{z_0}(\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)\cdot S_{\{h\}}(z_0)) \tag{4.10}$$

In order to express the first term [which is the only one non-N operator on the right-hand side of Eq. (4.10)] in terms of N operators, it should be noticed that we have derived Eq. (3.55) as

$$\vec{R}:(\partial_{z'}\Psi^{+hj}(z')\cdot\Psi^{-hj}(z')+\partial_{z'}\Psi^{-hj}(z')\cdot\Psi^{+hj}(z')):_R:\exp(\epsilon_h^j\phi^j(z_0)):$$
$$=:\Bigg(\partial_{z'}(\sqrt{z'-z_0}\Psi^{+hj}_{Nh}(z'))\cdot\frac{1}{\sqrt{z'-z_0}}\,\Psi^{-hj}_{Nh}(z')+\partial_{z'}\left(\frac{1}{\sqrt{z'-z_0}}\,\Psi^{-hj}_{Nh}(z')\right)$$
$$\cdot\sqrt{z'-z_0}\Psi^{+hj}_{Nh}(z')\Bigg):_R:\exp(\epsilon_h^j\phi^j(z_0)):$$
$$=\frac{1}{z'-z_0}:\Psi^{+hj}_{Nh}(z_0)\Psi^{-hj}_{Nh}(z_0):_R:\exp(\epsilon_h^j\phi^j(z_0)):+O(1), \tag{4.11}$$

which is reduced to

$$\frac{1}{z'-z_0}\,2\epsilon_h^j\cdot\partial_{z_0}\phi_{Nh}(z_0):\exp(\epsilon_h^j\phi^j(z_0)):+O(1), \tag{4.12}$$

by using Eq. (3.50). Therefore, we find

$$\vec{R}\cdot(\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger):\exp(\tfrac{1}{2}\phi^0(z_0)):)\partial_{z_0}\left(\prod_{j=1}^4:\exp(\epsilon_h^j\phi^j(z_0)):\right)$$
$$=:\Bigg(1/2\Psi^{+hj}_{Nh}(z_0)\Psi^{-hj}_{Nh}(z_0)+\tfrac{1}{2}i\oint\frac{dwd\xi}{2\pi i}\,P_F^{-hj^\dagger}(w,\xi)\,\frac{\theta_{NF}(w,\xi)}{\sqrt{z_{NF}(w,\xi)-z_0}}\cdot\Psi^{+hj}_{Nh}(z_0)$$
$$-\tfrac{1}{2}i\oint\frac{dwd\xi}{2\pi i}\,P_F^{+hj^\dagger}(w,\xi)\,\frac{\theta_{NF}(w,\xi)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\cdot\Psi^{-hj}_{Nh}(z_0)\Bigg)\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger):_R S_{\{h\}}(z_0), \tag{4.13}$$

the right-hand side of which is given in terms of N operators. Finally, we obtain from Eqs. (4.10) and (4.13)

$$-\vec{R}\cdot(\theta_0\partial_{z_0}\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger))S_{\{h\}}(z_0)$$
$$=-\partial_{z_0}{:}(\theta_0\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)\cdot S_{\{h\}}(z_0)){:}_R+\hat{\Phi}^F_{N\{h\},2}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0), \tag{4.14}$$

with

$$\hat{\Phi}^F_{N\{h\},2}(z_0,\theta_0;P_F^\dagger)={:}\Bigg(1/2\Psi^{+hj}_{Nh}(z_0)\Psi^{-hj}_{Nh}(z_0)+\tfrac{1}{2}i$$
$$\times\oint\frac{dwd\xi}{2\pi i}P_F^{-hj^\dagger}(w,\xi)\frac{\theta_{NF}(w,\xi)}{\sqrt{z_{NF}(w,\xi)-z_0}}\cdot\Psi^{+hj}_{Nh}(z_0)-\tfrac{1}{2}i$$
$$\times\oint\frac{dwd\xi}{2\pi i}P_F^{+hj^\dagger}(w,\xi)\frac{\theta_{NF}(w,\xi)}{(\sqrt{z_{NF}(w,\xi)-z_0})^3}\cdot\Psi^{-hj}_{Nh}(z_0)\Bigg)$$
$$\times\theta_0\cdot\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger){:}_R$$
$$+{:}\left(\frac{\frac{1}{2}\Psi^{+0}_{N+}(z_0)}{ip^{+0}}\right)\theta_0\partial_{\theta_0}\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger){:}_R. \tag{4.15}$$

On the other hand, from the following OPE:

$$\vec{R}\cdot\mathscr{F}^{jF}(z'){:}\exp(\epsilon^j_h\phi^j(z_0)){:}=\frac{1}{\sqrt{z'-z_0}}\partial_{z_0}X^{+hj}(z_0)\cdot\Psi^{-hj}_{Nh}(z_0){:}\exp(\epsilon^j_h\phi^j(z_0)){:}+O(\sqrt{z'-z_0}), \tag{4.16}$$

we can derive that

$$\sum_{j=1}^4\vec{R}\cdot\mathscr{F}^{jF}(z')\cdot\Phi^F_{\{h\}}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)=\frac{1}{\sqrt{z'-z_0}}\hat{\Phi}^F_{N\{h\},3}(z_0,\theta_0;P_F^\dagger)S_{\{h\}}(z_0)+O(\sqrt{z'-z_0}), \tag{4.17}$$

with

$$\hat{\Phi}^F_{N\{h\},3}(z_0,\theta_0;P_F^\dagger)\equiv\sum_{j=1}^4{:}\Bigg(\Psi^{-hj}_{Nh}(z_0)-i\oint\frac{dwd\xi}{2\pi i}P_F^{-hj^\dagger}(w,\xi)\frac{\theta_{NF}(w,\xi)}{\sqrt{z_{NF}(w,\xi)-z_0}}\Bigg)$$
$$\times\Bigg(\partial_{z_0}X^{+hj}(z_0)$$
$$+i\oint\frac{dwd\xi}{2\pi i}P_F^{+hj^\dagger}(w,\xi)\frac{1}{z_{NF}(w,\xi)-z_0}\Bigg)$$
$$\times\Phi^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger){:}_R. \tag{4.18}$$

With the help of OPE (4.7), Eq. (4.14), and OPE (4.17), we finally find that the N operator $\hat{\Phi}^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)$ in OPE (4.1) is given by

$$\hat{\Phi}^F_{N\{h\}}(z_0,\theta_0;P_F^\dagger)(\equiv\hat{\Phi}^{F3}_{N\{h\}}(z_0;P_F^\dagger)+\theta_0\hat{\Phi}^{F4}_{N\{h\}}(z_0;P_F^\dagger))=\sum_{i=1}^3\hat{\Phi}^F_{N\{h\},i}(z_0,\theta_0;P_F^\dagger). \tag{4.19}$$

with $\hat{\Phi}^F_{N\{h\},i}(z_0,\theta_0;P_F^\dagger)$ (4.8) [for $i=1$], (4.15) [for $i=2$], and (4.18) [for $i=3$]. In the mass shell limit [when the first term on the right-hand side of OPE (4.1) vanishes], we find

$$\lim_{z'\to z_0}\sqrt{z'-z_0}\mathscr{T}(z')\Phi^{F1}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0)=\partial_{z_0}(\Phi^{F0}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0))-\hat{\Phi}^{F4}_{N\{h\}}(z_0;P^\dagger_F)S_{\{h\}}(z_0), \tag{4.20}$$

which can be used in picture-changing Ramond states in 1/2 picture into those in 3/2 picture.

V. CONCLUSIONS AND DISCUSSIONS

In Ref. 1, we have proposed GFPVO $\Phi^B(z_0,\theta_0;P^\dagger_B)$ (1.30) in the NS sector of superstring, which can be easily obtained by simply supergeneralizing GFPVO in bosonic string theory [i.e., Eq. (2.6) in Ref. 4]. It is easily *imagined* that GFPVO $\Phi^B(z_0,\theta_0;P^\dagger_B)$ might satisfy our proposed OPE (5.26) in Ref. 1], since it is the one obtained by supergeneralizing OPE (3.10) in Ref. 4. On the other hand, we have proposed[1] GFPVO $\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)$ (1.12) in the R sector, which is quite unfamiliar and typical of superstring theory. The fact that our $\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)S_{\{h\}}(z_0)$ [given by Eqs. (1.12) and (1.4)] actually satisfy our proposed OPEs (4.18)–(4.20) in Ref. 1 might not be easily imagined. Therefore, in this paper, we have explained in detail how to derive OPEs (3.58) and (4.1), giving so many new materials necessary in proving these OPEs. The proof is straightforward but extremely tedious; In Sec. II, the contractions (in the R sector) of $\theta_F(w,\xi)$ (1.16) and $z_F(w,\xi)$ (1.17) with respect to $\lambda\cdot ip^{+0}Y^{-0}_+(z',\theta';z_0)$ are denoted by $\delta\theta_F(w,\xi)$ (2.2) and $\delta z_F(w,\xi)$ (2.1), respectively. These quantities are calculated with the help of the modified superconformal mapping in the R sector given by Eqs. (2.9) and (2.10); Since (w,ξ) is just c number, there is not any contraction with $\lambda\cdot ip^{+0}Y^{-0}_+(z',\theta';z_0)$, so that we have Eq. (2.16), which leads to explicit formulas (2.22) and (2.23). Furthermore, we can derive z_0 and θ_0 dependence of θ_F (1.16) and z_F (1.17). Since (w,ξ) is independent of z_0 and θ_0, we have Eq. (2.27), which leads to explicit formulas (2.32)–(2.35). With the help of various formulas already obtained in Sec. II, we investigate in Sec. III the OPE among GFPVO in the R sector and the superstressino operator. For this purpose, it is not helpful to use double-valued $\mathscr{T}^{jF}(z',\theta')$ (3.1). On the other hand, $\tau^j(z',\theta';z_0)$ defined by Eq. (3.2) is single valued. Therefore, we investigate OPE among GFPVO and $\tau(z',\theta';z_0)$ (3.15), thus obtaining OPE (3.16) by straightforward calculations. Denoting singular terms of F (in the neighborhood of $z'=z_F,z'_F$) by $(F)_{\text{sing.}}$, we find $(H^{\mp j}_h)_{\text{sing.}}$ (3.20) and $(H^{jjF}_h)_{\text{sing.}}$ (3.21). It is essential that $(H^{F1/2(1)}_h)_{\text{sing.}}$ (3.22) is shown to be zero, by proving Eqs. (E20) and (E21). Thus, we have only to investigate singularities at $z'=z_0$ on the right-hand side of OPE (3.16). Then, with the help of identities (3.23) [(3.24)], OPE (3.25) leads to OPE (3.46) [(3.59)] among GFPVO and $T^{0F}(\mathscr{T}^{0F})$. Finally, summing OPE (3.46) [(3.59)] and OPE (3.55) [(3.60)] gives OPE (3.58) [(3.62) and (3.63)] among GFPVO in the R sector and $T_{NF}(\mathscr{T}_{NF})$. However, the term proportional to $(z'-z_0)^{-1/2}$ in OPEs (3.60)–(3.63) are not obtained explicitly. On the other hand, in Sec. IV, the next leading term $\hat{\Phi}^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)$ in OPE (4.1) is found out to be given by Eq. (4.19). This term is related with the picture changing, as seen from Eq. (4.20). In Appendices A–E, we derive various formulas needed in proving OPEs among GFPVO $\Phi^F_{N\{h\}}(z_0,\theta_0;P^\dagger_F)\cdot S_{\{h\}}(z_0)$ and the stress operators; In Appendix A, operators G^F_+ and $G^{\mp jF}$ in OPE (3.11) are explicitly derived. In Appendix B, operators $g^{\mp jF}_h$ and g^{jjF}_h in OPE (3.13) are explicitly derived. In Appendix C, operators G^{jjF}_h in OPE (3.14) are explicitly derived. In Appendix D, singularities at $z'=z_F,z'_F$ in $g^{\mp jF}_h$ and g^{jjF}_h in OPE (3.13) are analyzed, and results are given by Eqs. (D4)–(D10) and (D12)–(D15), respectively. In Appendix E, both $H^{F1/2}_h$ and H^{F1}_h in Eqs. (3.20) and (3.21) are shown to be free of singularities at $z'=z_F,z'_F$, by proving Eqs. (E20) and (E21).

In the following, we investigate the operator product expansion (OPE) among GFPVO $\Phi^B(z_0,\theta_0;P^\dagger_B)$ (1.30) and the superstressino operator $\mathscr{T}_B(z',\theta')$ in the NS sector, which is defined by

$$\mathscr{T}_B(z',\theta')\equiv\sum_{\pm}\sum_{j=0}^{4}:\tfrac{1}{2}\partial_{z'}X^{\pm j}(z',\theta')D_{\theta'}X^{\mp j}(z',\theta'):_{\text{NS}} \tag{5.1}$$

$$(\equiv\tfrac{1}{2}\mathscr{T}_B(z')+\theta' T_B(z')). \tag{5.1'}$$

[Hereafter, we shall use the notation $\mathscr{T}_B(z',\theta')$ instead of the notation $T_B(z',\theta')$ in Ref. 1.] In Eq. (5.1), $:\ :_{\mathrm{NS}}$ is the normal ordering in the NS sector[1] and the supervector field is given by

$$X^{\pm j}(z',\theta')\equiv X^{\pm j}(z')+\theta'\Psi^{\pm j}(z'). \tag{5.2}$$

Then, in a much simpler way than the one described (in detail) in Sec. III, we can derive the following OPE, by using the techniques given in Sec. III:

$$\begin{aligned}\vec{R}\cdot\mathscr{T}_B(z',\theta')\Phi^B(z_0,\theta_0;P_B^\dagger)=&\frac{\theta'-\theta_0}{(z'-z_0-\theta'\theta_0)^2}\left(-\left(\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^B(z_0,\theta_0;P_B^\dagger)\right.\\&\left.+[N_B,\Phi^B(z_0,\theta_0;P_B^\dagger)]\right)+\frac{1/2}{z'-z_0-\theta'\theta_0}D_{\theta_0}\Phi^B(z_0,\theta_0;P_B^\dagger)\\&+\frac{\theta'-\theta_0}{z'-z_0-\theta'\theta_0}\partial_{z_0}\Phi^B(z_0,\theta_0;P_B^\dagger)+(\text{terms regular at } z'=z_0),\end{aligned} \tag{5.3}$$

where we have used the light-cone number operator N_B defined by

$$N_B\equiv\sum_{n=1}^{\infty}\sum_{\pm}\sum_{k=1}^{4}(\alpha_{-n}^{\mp k}\alpha_n^{\pm k}+(n-1/2)b_{-n+1/2}^{\mp k}b_{n-1/2}^{\pm k}). \tag{5.4}$$

With the help of Eqs. (1.30′) and (5.1′), OPE (5.3) can be rewritten as

$$\begin{aligned}\vec{R}\cdot T_B(z')\Phi^{Bm}(z_0;P_B^\dagger)=&\frac{1}{(z'-z_0)^2}\left(\left(\tfrac{1}{2}\delta_{m1}-\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^{Bm}(z_0;P_B^\dagger)+[N_B,\Phi^{Bm}(z_0;P_B^\dagger)]\right)\\&+\frac{1}{z'-z_0}\partial_{z_0}\Phi^{Bm}(z_0;P_B^\dagger)+(\text{terms regular at } z'=z_0),\\&[\text{for } m=0,1],\end{aligned} \tag{5.5}$$

$$\vec{R}\cdot\mathscr{T}_B(z')\Phi^{B0}(z_0;P_B^\dagger)=\frac{1}{z'-z_0}\Phi^{B1}(z_0;P_B^\dagger)+(\text{terms regular at } z'=z_0) \tag{5.6}$$

and

$$\begin{aligned}\vec{R}\cdot\mathscr{T}_B(z')\Phi^{B1}(z_0;P_B^\dagger)=&\frac{2}{(z'-z_0)^2}\left(\left(-\sum_{j=0}^{4}p^{+j}p^{-j}\right)\Phi^{B0}(z_0;P_B^\dagger)+[N_B,\Phi^{B0}(z_0;P_B^\dagger)]\right)\\&+\frac{1}{z'-z_0}\partial_{z_0}\Phi^{B0}(z_0;P_B^\dagger)+(\text{terms regular at } z'=z_0).\end{aligned} \tag{5.7}$$

Finally, we discuss how our results can be used in quantum superstring field theory. $\Phi^{Fm}_{N\{h\}}(z_0;P_F^\dagger)$ [for $m=0,1$] given by Eqs. (1.2′) can be expanded as sums of following terms:

$$\sum\ \text{const.}\,\alpha\cdots\alpha d\cdots d\times V^m_{\alpha\cdots\alpha d\cdots d}(z_0)\quad(\text{for } m=0,1). \tag{5.8}$$

Hereafter, the physical vertex operator $V^m_{\alpha\cdots\alpha d\cdots d}(z_0)$ will be simply denoted by $V^{m\not{}}(z_0)$, $\not{}$ representing quantum numbers. [How to fix ''const.'' in Eq. (5.8) and how to select out GSO-allowed[6] physical vertex operators will be discussed in our succeeding paper.[7]] Then, we can construct the state given by $V^{0\not{}}(0)S_{\{h\}}(0)|0\rangle$, which constitutes the complete system, in the following on-shell limit of the state $|\not{}\rangle$:

$$-\sum_{j=0}^{4}p^{+j}p^{-j}+N_{\not{}}\to 0\quad(\text{i.e., } L_0^F\to 0), \tag{5.9}$$

$N_{\not}$ being the eigenvalue [of the light-cone number operator N_F (3.41)] determined by

$$[N_F, \alpha\cdots\alpha d\cdots d] = N_{\not}\alpha\cdots\alpha d\cdots d. \tag{5.10}$$

Unfortunately, this state does not satisfy the Dirac equation, since we find from OPE (3.62) that

$$\mathscr{F}^F_{N0}(V^{0\not}(0)S_{\{h\}}(0)|0\rangle) = V^{1\not}(0)S_{\{h\}}(0)|0\rangle \neq 0, \tag{5.11}$$

even in the on-shell limit, $\mathscr{F}^F_{N0}$ being the Dirac operator. Using the technique used in calculating the norm given by Eq. (3.29) of Ref. 8, we can find the following state constructed by

$$V^{1\not}(0)S_{\{h\}}(0)|0\rangle \tag{5.12}$$

has the nonvanishing norm, where we have used $V^{1\not}$ given by Eq. (5.8).

Similarly, $\Phi^{Bm}(z_0; P_B^\dagger)$ [for $m=0,1$] given by Eq. (1.30′) can be expressed as sums of following terms:

$$\sum \text{ const.}\, \alpha\cdots\alpha b\cdots b \times V^m_{\alpha\cdots\alpha b\cdots b}(z_0) \quad (\text{for } m=0,1). \tag{5.13}$$

Hereafter, the physical vertex operator $V^m_{\alpha\cdots\alpha b\cdots b}(z_0)$ will be simply denoted by $V^{m\not\ell}(z_0)$, $\not\ell$ representing quantum numbers. Then, the *physical* Neveu–Schwarz state $|\not\ell\rangle$ is given by $V^{0\not\ell}(0)|0\rangle$, which constitutes the orthonormal complete system, in the following on-shell limit of the state $|\not\ell\rangle$:

$$-\sum_{j=0}^{4} p^{+j}p^{-j} + N_{\not\ell} \longrightarrow 1/2 \quad (\text{i.e., } L_0^B \to 1/2). \tag{5.14}$$

$N_{\not\ell}$ being the eigenvalue $N_{\not\ell}$ [for the light-cone number operator N_B (4.4)] determined by

$$[N_B, \alpha\cdots\alpha b\cdots b] = N_{\not\ell}\alpha\cdots\alpha b\cdots b. \tag{5.15}$$

[How to fix ''const.'' in Eq. (5.13) and how to select out GSO-allowed[6] physical vertex operator will be discussed in our succeeding paper.[7]] On the other hand, we find from OPEs (5.6) and (5.7) that

$$\mathscr{F}^B_{-1/2}\cdot V^{0\not\ell}(0)|0\rangle = V^{1\not\ell}(0)|0\rangle \tag{5.16}$$

and

$$\mathscr{F}^B_{1/2}\cdot V^{1\not\ell}(0)|0\rangle = V^{0\not\ell}(0)|0\rangle, \tag{5.17}$$

with

$$\{\mathscr{F}^B_{1/2}, \mathscr{F}^B_{-1/2}\} = 2L_0^B. \tag{5.18}$$

In RNS superstring theory, Friedan, Martinec, and Schenker[5] have originally succeeded in constructing physical vertex operators for massless fermions and bosons, with the help of conformal field theoretical techniques.[9] For this purpose, they had to introduce various ghost fields (which we did not discuss in this paper). They have also given the picture changing operator X. FMS have shown that the operator X plays important roles in obtaining physical scattering amplitudes in RNS superstring theory. Based on their ideas, quantum superstring field theory has been proposed by Witten,[10] where there exist interactions among BRST invariant R states (in 1/2 picture) given by

$$\exp(+\sigma(0))\exp(-\tfrac{1}{2}\phi(0))\cdot(V^{1\not}(0)S_{\{h\}}(0)|0\rangle), \tag{5.19}$$

and BRST invariant NS states (in 0 picture) given by

$$\exp(+\sigma(0))\exp(-\phi(0))\cdot V^{0\not{\prime}}(0)|0\rangle. \tag{5.20}$$

However, Wendt[11] has pointed out that Witten's quantum superstring field theory is plagued with divergences induced by colliding of two *local* picture changing operators. To overcome this difficulty, Preitshopf, Thorn, and Jost[12] have investigated quantum superstring field theory using BRST invariant NS states in one picture, which are created by

$$(\exp(+\sigma(0))\cdot V^{1\not{\prime}}(0)-\tfrac{1}{2}\gamma(0)\cdot V^{0\not{\prime}}(0))|0\rangle, \tag{5.21}$$

γ being the ghost operator of conformal weight $-1/2$, introduced and discussed by FMS.[5] We should notice that there exist physical vertex operators $V^{1\not{\prime}}$ and $V^{1\not{\prime}}$ in Eqs. (5.8) and (5.13), which are generated by $\Phi^{F1}(z_0;P_F^\dagger)$ and $\Phi^{B1}(z_0;P_B^\dagger)$, respectively.

In Ref. 13, we have introduced various techniques in quantum bosonic string field theory, which make it possible to calculate physical scattering amplitudes, by using the generating functional for physical vertex operators (GFPVO). Physical scattering amplitudes among N excited particles can be obtained by Eq. (7.3) [with Eqs. (7.4)–(7.7) and (7.28) in Ref. 13.] It should be noticed that Eq. (7.3) can be calculated by using c number conformal mapping (7.30) (in Ref. 13), which is obtained from q number conformal mapping (7.10). In Eq. (7.30), we have the effective c number field (7.31), which is induced by the presence of external particles at various punctures. (See Ref. 13, for more details.) Thus, GFPVO [given by Eq. (7.8) in Ref. 13] in bosonic string theory has played important roles in calculating physical particles. Therefore, we can imagine that similar situations occur also in superstring theory. However, we have two types of formalisms in superstring theory. Historically, Ramond[2]–Neveu–Schwarz[3] superstring was at first formulated and then Green–Schwarz[7,14] superstring was formulated. In this paper, we only investigated RNS superstring, where we have only string coordinate $X^{\pm j}$ and string coordino $\Psi^{\pm j}$, both of which are vectors in ten-dimensional space-time. Therefore, it is extremely important in RNS superstring to know how we can construct (all) fermionic Ramond states (described by spinor fields). We have shown in this paper that physical R states can be constructed by Eq. (5.12). {Incidentally, physical NS states in 0 (1) picture can be constructed by Eq. (5.20) [(5.21)]}. On the other hand, Green and Schwarz[8] have analyzed physical fermionic and bosonic states in GS superstring, by using the spectrum generating algebras (SGA) for each fermionic sector and bosonic sector. We shall see in our succeeding papers[7] that our physical states states (5.12), (5.20), and (5.21) can be actually generated by physical string coordinate modes and physical string coordino modes which satisfy SGA. Thus, our approach gives new insight into the relation between RNS superstring and GS superstring. [Incidentally, this relation has been discussed by Berkovits.[15]]

APPENDIX A: NON-N OPERATORS G_+^F AND $G^{\mp jF}$ IN EQ. (3.11)

In this Appendix, explicit expressions for non-N operators $G_+^F(\ )_\iota$ and $G^{\mp jF}(\ )_\iota$ in OPE (3.11′) are given. Applying contractions in the R sector (3.9) to ρ_{N+}^{0F} (1.3), we find that the non-N operator $G_+^F(\ )_\iota$ (for $\iota=1-3$) in OPE (3.11′) is obtained as follows:

$$\begin{aligned}
G_+^F(z_0,\theta_0|z',\theta')_1=&\frac{\theta'-\theta_0}{(z'-z_0-\theta'\theta_0)^2}\,(-p^{+0}p^{-0})+\frac{1/2}{z'-z_0-\theta'\theta_0}\,D_{\theta'}\\
&\times\left(\sum_{\pm}\,-ip^{\pm 0}Y_+^{\mp 0}(z',\theta';z_0)\right)+\frac{\frac{1}{2}(\theta'-\theta_0)}{z'-z_0-\theta'\theta_0}\,\partial_{z'}\\
&\times\left(\sum_{\pm}\,-ip^{\pm 0}Y_+^{\mp 0}(z',\theta';z_0)\right)+p^{+0}p^{-0}\left(-\frac{\frac{1}{2}\theta'\theta_0}{(z'-z_0)^2}\cdot\frac{\Psi_+^{+0}(z';z_0)}{ip^{+0}}\right.\\
&\left.-\frac{\frac{1}{2}\theta_0}{(z'-z_0)^2}+\frac{\frac{1}{2}\theta_0}{z'-z_0}\,\partial_{z'}\,\frac{Y_+^{+0}(z',\theta';z_0)}{ip^{+0}}\right),
\end{aligned}\tag{A1}$$

$$G_+^F(z_0,\theta_0|z',\theta')_2=\frac{\frac{1}{2}\theta'\theta_0}{(z'-z_0)^2}\,ip^{+0}\Psi_+^{-0}(z';z_0), \tag{A2}$$

and

$$G_+^F(z_0,\theta_0|z',\theta')_3=p^{+0}p^{-0}\frac{\theta_0}{(z'-z_0)^2}\left(\sum_{\pm}\frac{1}{4}((z'-z_0)^{\pm 1}-1)\right)$$
$$-\frac{\theta_0}{z'-z_0}\left(\sum_{\pm}\frac{1}{4}((z'-z_0)^{\pm 1}-1)\partial_{z'}(-ip^{\pm 0}X^{\mp 0}(z'))\right)$$
$$+\frac{1}{z'-z_0}\left(\sum_{\pm}\frac{1}{4}((z'-z_0)^{\mp 1}-1)(-ip^{\pm 0}\Psi_+^{\mp 0}(z';z_0))\right)$$
$$+\frac{p^{+0}p^{-0}}{4}\left(\frac{1}{z'-z_0}-1\right)\left(-\frac{\theta_0}{(z'-z_0)^2}+\frac{\theta_0}{z'-z_0}\,\partial_{z'}\,\frac{X^{+0}(z')}{ip^{+0}}\right). \quad \text{(A3)}$$

Furthermore, non-N operator $G^{\mp jF}(\ )_\iota$ [for $\iota=1-3$] in Eq. (3.11′) is found by applying contraction in the R sector (3.8) to ρ^{1F} (1.14);

$$G^{\mp jF}(w,\xi|z',\theta')_1=\tfrac{1}{2}(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0)\partial_{z'}+\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0)D_{\theta'})\frac{1}{\lambda}((\delta z_F-\delta\theta_F\cdot\theta_F)\partial_{z_F}$$
$$+\delta\theta_F D_{\theta_F})X^{\mp j}(z_F,\theta_F), \quad \text{(A4)}$$

$$G^{\mp jF}(w,\xi|z',\theta')_2=\frac{\frac{1}{2}\theta'}{z'-z_0}(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}\frac{1}{\lambda}((D_{\theta'}(\delta z_F-\delta\theta_F\cdot\theta_F))_{\theta'=0}\partial_{z_F}$$
$$+(D_{\theta'}\delta\theta_F)_{\theta'=0}D_{\theta_F})X^{\mp j}(z_F,\theta_F) \quad \text{(A5)}$$

and

$$G^{\mp jF}(w,\xi|z',\theta')_3=\frac{1}{4}(z'-z_0-1)(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}\frac{1}{\lambda}((\partial_{z'}(\delta z_F-\delta\theta_F\cdot\theta_F))_{\theta'=0}\partial_{z_F}$$
$$+(\partial_{z'}\delta\theta_F)_{\theta'=0}D_{\theta_F})X^{\mp j}(z_F,\theta_F)+\frac{1}{4}\left(\frac{1}{z'-z_0}-1\right)$$
$$\times(\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}\frac{1}{\lambda}((D_{\theta'}(\delta z_F-\delta\theta_F\cdot\theta_F))_{\theta'=0}\partial_{z_F}$$
$$+(D_{\theta'}\delta\theta_F)_{\theta'=0}D_{\theta_F})X^{\mp j}(z_F,\theta_F). \quad \text{(A6)}$$

The second term on the right-hand side of the following non-N operator:

$$f_+^F(z',\theta'|z_0,\theta_0)\equiv\frac{Y_+^{+0}(z',\theta';z_0)}{ip^{+0}}-\log(z'-z_0-\theta'\theta_0) \quad \text{(A7)}$$

[existing in Eqs. (A4)–(A6),) has arisen from the contraction (3.9) in the R sector. [Incidentally, Eq. (A7) is N-ized into

$$f_{N+}^F(z',\theta'|z_0,\theta_0)\equiv\frac{Y_{N+}^{+0}(z',\theta')}{ip^{+0}}-\log(z'-z_0-\theta'\theta_0), \quad \text{(A7′)}$$

which will be useful in analyzing singularities at $z'=z_0$.]

Substituting Eqs. (2.22) and (2.23) into Eqs. (A4)–(A6), we shall find the following explicit formulas for non-N operator functions $G^{F1/2}(\ )_\iota$'s and $G^{F1}(\ )_\iota$'s in Eq. (3.12′). As for $G^{F1/2}(\ )_\iota$'s (for $\iota=1-3$) in Eq. (3.12′), we have

$$G^{F1/2}(w,\xi|z',\theta')_1=\tfrac{1}{2}D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0)\Bigg(w(D_\xi\theta_F)^2\Bigg(\frac{1}{z'-z_F-\sqrt{z_F-z_0}\,\theta'\theta_F}-\frac{1}{z'-z_0}\Bigg)$$
$$+\xi wK_F^3(w)\,\frac{\theta_F-\sqrt{z_F-z_0}\,\theta'}{(z_F-z')^2}\Bigg)+\tfrac{1}{2}\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0)$$
$$\times\Bigg(w(D_\xi\theta_F)^2\Bigg(\frac{\theta'-\sqrt{z_F-z_0}\,\theta_F}{z'-z_F}-\frac{\theta'}{z'-z_0}\Bigg)$$
$$+\xi wK_F^3(w)\Bigg(\frac{\theta_F\theta'}{(z_F-z')^2}+\frac{\sqrt{z_F-z_0}}{z_F-z'}\Bigg)\Bigg),\tag{A8}$$

$$G^{F1/2}(w,\xi|z',\theta')_2=\frac{\frac{1}{2}\theta'}{z'-z_0}\,(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}\Bigg(-w(D_\xi\theta_F)^2\cdot\,\frac{\sqrt{z_F-z_0}\,\theta_F}{z'-z_F}$$
$$-\xi wK_F^3(w)\,\frac{\sqrt{z_F-z_0}}{z'-z_F}\Bigg)\tag{A9}$$

and

$$G^{F\,1/2}(w,\xi|z',\theta')_3=\frac{1}{4}\,(z'-z_0-1)(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}\Bigg(w(D_\xi\theta_F)^2\Bigg(\frac{1}{z'-z_F}-\frac{1}{z'-z_0}\Bigg)$$
$$+\xi wK_F^3(w)\,\frac{\theta_F}{(z_F-z')^2}\Bigg)+\frac{1}{4}\,\Bigg(\frac{1}{z'-z_0}-1\Bigg)(\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}$$
$$\times\Bigg(-w(D_\xi\theta_F)^2\cdot\,\frac{\sqrt{z_F-z_0}}{z'-z_F}\,\theta_F-\xi wK_F^3(w)\,\frac{\sqrt{z_F-z_0}}{z'-z_F}\Bigg).\tag{A10}$$

Furthermore, as for $G^{F1}(\ )_\iota$ (for $\iota=1-3$) in Eq. (3.12′), we have

$$G^{F1}(w,\xi|z',\theta')_1=\ \tfrac{1}{2}D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0)$$
$$\times\Bigg(\begin{array}{l}\Bigg(\dfrac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\Bigg)\Bigg(\dfrac{1}{z'-z_F-\sqrt{z_F-z_0}\,\theta'\theta_F}-\dfrac{1}{z'-z_0}\Bigg)\\ +\Bigg(wK_F^2(w)+\dfrac{1}{2}\,\xi wK_F(w)\partial_wJ_F(w)\Bigg)\,\dfrac{\theta_F-\sqrt{z_F-z_0}\,\theta'}{(z_F-z')^2}\\ +\dfrac{1}{2}\,\xi wK_F^3(w)\Bigg(\begin{array}{c}\dfrac{2\sqrt{z_F-z_0}\,\theta'\theta_F}{(z_F-z')^3}\\ -\dfrac{1}{(z_F-z')^2}-\dfrac{\frac{1}{2}\theta'\theta_F}{(z_F-z')^2\sqrt{z_F-z_0}}\end{array}\Bigg)\end{array}\Bigg)$$
$$+\frac{1}{2}\,\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0)$$
$$\times\Bigg(\begin{array}{c}\Bigg(\dfrac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\Bigg)\Bigg(-\dfrac{\theta'-\sqrt{z_F-z_0}\,\theta_F}{z'-z_F}+\dfrac{\theta'}{z'-z_0}\Bigg)\\ +\Bigg(wK_F^2(w)+\dfrac{1}{2}\,\xi wK_F(w)\partial_wJ_F(w)\Bigg)\Bigg(\dfrac{\theta'\theta_F}{(z'-z_F)^2}+\dfrac{\sqrt{z_F-z_0}}{z'-z_F}\Bigg)\\ +\dfrac{1}{2}\,\xi wK_F^3(w)\Bigg(\dfrac{\frac{1}{2}\theta_F}{(z_F-z')\sqrt{z_F-z_0}}+\dfrac{\theta'-\sqrt{z_F-z_0}\,\theta_F}{(z'-z_F)^2}\Bigg)\end{array}\Bigg),\tag{A11}$$

$$G^{F1}(w,\xi|z',\theta')_2=\frac{\frac{1}{2}\theta'}{z'-z_0}\,(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))$$
$$\times\begin{pmatrix}\left(\frac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\right)\frac{\sqrt{z_F-z_0}\theta_F}{z'-z_F}\\ +\left(wK_F^2(w)+\frac{1}{2}\,\xi wK_F(w)\partial_w J_F(w)\right)\frac{\sqrt{z_F-z_0}}{z'-z_F}\\ +\frac{1}{2}\,\xi wK_F^3(w)\left(\frac{\frac{1}{2}\theta_F}{(z_F-z')\sqrt{z_F-z_0}}-\frac{\sqrt{z_F-z_0}\theta_F}{(z_F-z')^2}\right)\end{pmatrix},\tag{A12}$$

and

$$G^{F1}(w,\xi|z',\theta')_3=\frac{1}{4}\,(z'-z_0-1)(D_{\theta'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}$$
$$\times\begin{pmatrix}\left(\frac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\right)\left(\frac{1}{z'-z_F}-\frac{1}{z'-z_0}\right)\\ +\left(wK_F^2(w)+\frac{1}{2}\,\xi wK_F(w)\partial_w J_F(w)\right)\frac{\theta_F}{(z_F-z')^2}\\ -\frac{1}{2}\,\xi wK_F^3(w)\,\frac{1}{(z_F-z')^2}\end{pmatrix}+\frac{1}{4}\left(\frac{1}{z'-z_0}-1\right)$$
$$\times(\partial_{z'}f_+^F(z',\theta'|z_0,\theta_0))_{\theta'=0}$$
$$\times\begin{pmatrix}\left(\frac{1}{2}\,\xi D_\xi\theta_F+w\partial_w\theta_F\right)\frac{\sqrt{z_F-z_0}\theta_F}{z'-z_F}\\ +\left(wK_F^2(w)+\frac{1}{2}\,\xi wK_F(w)\partial_w J_F(w)\right)\frac{\sqrt{z_F-z_0}}{z'-z_F}\\ +\frac{1}{2}\,\xi wK_F^3(w)\left(\frac{\frac{1}{2}\theta_F}{(z_F-z')\sqrt{z_F-z_0}}-\frac{\sqrt{z_F-z_0}\theta_F}{(z_F-z')^2}\right)\end{pmatrix}.\tag{A13}$$

APPENDIX B: NON-N OPERATORS $g_h^{\mp jF}$ AND g_h^{jjF} IN EQ. (3.13)

In this Appendix, we shall give explicit expressions for non-N operators $g_h^{\mp jF}(\ )_\iota$'s and $g_h^{jjF}(\ )_\iota$'s in Eq. (3.13′). (We shall notice soon that they are dependent on ϵ_h^j.) Applying the contraction in the R sector (3.10) to ρ^{1F} (1.4), we obtain after straightforward calculations

$$g_h^{\mp jF}(w,\xi|z',\theta')_1=\frac{1}{2}\,(D_{\theta'}Y_h^{\mp j}(z',\theta';z_0)\partial_{z'}+\partial_{z'}Y_h^{\mp j}(z',\theta';z_0)D_{\theta'})$$
$$\times\log(z'-z_F-\theta'\theta_F(z_F-z_0)^{\mp\epsilon_h^j}),\tag{B1}$$

$$g_h^{\mp jF}(w,\xi|z',\theta')_2=\mp\frac{\epsilon_h^j\theta'\theta_F(z_F-z_0)^{\mp\epsilon_h^j}}{(z'-z_0)(z'-z_F)}\,\Psi_h^{\mp j}(z';z_0),\tag{B2}$$

and

$$g_h^{\mp jF}(w,\xi|z',\theta')_3=-\frac{1}{4}\,((z'-z_0)^{\pm 2\epsilon_h^j}-1)\,\frac{\theta_F(z_F-z_0)^{\mp\epsilon_h^j}}{z'-z_F}\,\partial_{z'}X^{\mp j}(z')$$
$$+\frac{1}{4}\,((z'-z_0)^{\mp 2\epsilon_h^j}-1)\,\frac{1}{z'-z_F}\,\Psi_h^{\mp j}(z';z_0).\tag{B3}$$

On the other hand, non-N operator $g_h^{jjF}(\ )_\iota$ (for $\iota=1-3$) is found out to be given as follows:

$$g_h^{jjF}(w,\xi;w',\theta'|z',\theta')_1$$

$$=\frac{1}{2}D_{\theta'}\log(z'-z'_F-\theta'\theta'_F(z'_F-z_0)^{-\epsilon_h^j})\partial_{z'}\log(z'-z_F-\theta'\theta_F(z_F-z_0)^{+\epsilon_h^j})$$

$$+\frac{1}{2}\partial_{z'}\log(z'-z'_F-\theta'\theta'_F(z'_F-z_0)^{-\epsilon_h^j})D_{\theta'}\log(z'-z_F-\theta'\theta_F(z_F-z_0)^{+\epsilon_h^j}), \quad \text{(B4)}$$

$$g_h^{jjF}(w,\xi;w',\xi'|z',\theta')_2=\frac{\epsilon_h^j\theta'\theta'_F\theta_F(z'_F-z_0)^{-\epsilon_h^j}(z_F-z_0)^{+\epsilon_h^j}}{(z'-z_0)(z'-z'_F)(z'-z_F)}, \quad \text{(B5)}$$

and

$$g_h^{jjF}(w,\xi;w',\xi'|z',\theta')_3=-\frac{1}{4}((z'-z_0)^{+2\epsilon_h^j}-1)\frac{\theta'_F(z'_F-z_0)^{-\epsilon_h^j}}{(z'-z'_F)(z'-z_F)}$$

$$-\frac{1}{4}((z'-z_0)^{-2\epsilon_h^j}-1)\frac{\theta_F(z_F-z_0)^{+\epsilon_h^j}}{(z'-z'_F)(z'-z_F)}. \quad \text{(B6)}$$

In Eqs. (B1)–(B6), we have used shorthand notations z_F and θ_F in Eq. (3.12), together with

$$z'_F\equiv z_F(w',\xi'), \quad \theta'_F\equiv\theta_F(w',\xi'). \quad \text{(B7)}$$

APPENDIX C: NON-N OPERATOR G_h^{jjF} IN EQ. (3.14)

In this Appendix, explicit formulas for non-N operators G_h^{jjF} in Eq. (3.14) are given. Using Eqs. (3.11) and (3.12), we find that

$$\vec{R}\cdot\tau_+^0(z',\theta';z_0):\overbrace{X^{+j}(z_F,\theta_F)X^{-j}}^{R}(z'_F,\theta'_F)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R$$

$$=:(G^{F1/2}(w,\xi|z',\theta')\partial_{z_F}+G^{F1}(w,\xi|z',\theta')D_{\theta_F})$$

$$\times\overbrace{X^{+j}(z_F,\theta_F)X^{-j}}^{R}(z'_F,\theta'_F)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R$$

$$+:(G^{F1/2}(w',\xi'|z',\theta')\partial_{z'_F}+G^{F1}(w',\xi'|z',\theta')D_{\theta'_F})$$

$$\times\overbrace{X^{-j}(z'_F,\theta'_F)X^{+j}}^{R}(z_F,\theta_F)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R$$

$$+:\tau_+^0(z',\theta';z_0)\overbrace{X^{+j}(z_F,\theta_F)X^{-j}}^{R}(z'_F,\theta'_F)\exp(-ip^{+0}Y_{N+}^{-0}(z_0,\theta_0)):_R, \quad \text{(C1)}$$

which leads to

$$G_h^{jjF}(w,\xi;w',\xi'|z',\theta')=\begin{pmatrix}G^{F\,1/2}(w,\xi|z',\theta')\partial_{z_F}+G^{F1}(w,\xi|z',\theta')D_{\theta_F}\\ +G^{F\,1/2}(w',\xi'|z',\theta')\partial_{z'_F}+G^{F1}(w',\xi'|z',\theta')D_{\theta'_F}\end{pmatrix}$$

$$\times\log\left(z_F-z'_F-\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\cdot\theta_F\theta'_F\right). \quad \text{(C2)}$$

As we shall see from Eq. (C2), $G_h^{jjF}(\ )$ is dependent on ϵ_h^j, while $G^{F\,1/2}$ and G^{F1} on the right-hand side of Eq. (C2) do *not* depend on ϵ_h^j.

APPENDIX D: EXPANSIONS IN THE NEIGHBORHOOD OF $z'=z_F$

Non-N operators $g_h^{\mp jF}(w,\xi|z',\theta')$ [, given by Eqs. (B1) and (B4)–(B6),] have singularities at $z'=z_F(w,\xi)$, and they can be analyzed by expanding non-N operators $Y_h^{\pm j}(z',\theta';z_0)$ at $z'=z_F(w,\xi)(\equiv z_F)$ as

$$D_{\theta'}Y_h^{\pm j}(z',\theta';z_0)=(1+(\theta'-\theta_F)D_{\theta_F}-\theta'\theta_F\partial_{z_F})D_{\theta_F}Y_h^{\pm j}(z_F,\theta_F;z_0)+(z'-z_F)\times(1+(\theta'-\theta_F)D_{\theta_F}-\theta'\theta_F\partial_{z_F})\partial_{z_F}D_{\theta_F}Y_h^{\pm j}(z_F,\theta_F;z_0)+O((z'-z_F)^2) \tag{D1}$$

and

$$\partial_{z'}Y_h^{\pm j}(z',\theta';z_0)=(1+(\theta'-\theta_F)D_{\theta_F}-\theta'\theta_F\partial_{z_F})\partial_{z_F}Y_h^{\pm j}(z_F,\theta_F;z_0)+O((z'-z_F)). \tag{D2}$$

With the help of

$$Y_h^{\pm j}(z_F,\theta_F;z_0)=X^{\pm j}(z_F,\theta_F)+\theta_F((z_F-z_0)^{\mp\epsilon_h^j}-1)D_{\theta_F}X^{\pm j}(z_F,\theta_F), \tag{D3}$$

together with Eqs. (D1) and (D2), we find from Eqs. (B1) to (B3) that $g_h^{F1/2(1)}(\ )_\iota$ in Eq. (3.15′) is given as follows: As for $(g_h^{F1}(\ )_\iota)_{\text{sing.}}$ (for $\iota=1-3$), we find

$$(g_h^{F1}(w,\xi|z',\theta')_1)_{\text{sing.}}=\frac{\frac{1}{2}\theta'\theta_F}{(z'-z_F)^2}+\frac{1/2}{z'-z_F}(z_F-z_0)^{\pm\epsilon_h^j}\pm\frac{\epsilon_h^j\theta'\theta_F}{(z'-z_F)(z_F-z_0)}, \tag{D4}$$

$$(g_h^{F1}(w,\xi|z',\theta')_2)_{\text{sing.}}=\mp\frac{\epsilon_h^j\theta'\theta_F}{(z'-z_F)(z_F-z_0)} \tag{D5}$$

and

$$(g_h^{F1}(w,\xi|z',\theta')_3)_{\text{sing.}}=\mp\frac{1/2\epsilon_h^j}{z'-z_F}\left(\sqrt{z_F-z_0}-\frac{1}{\sqrt{z_F-z_0}}\right). \tag{D6}$$

On the other hand, as for $(g_h^{F1/2}(\ )_\iota)_{\text{sing.}}$ (for $\iota=1-3$), we find

$$(g_h^{F1/2}(w,\xi|z',\theta')_1)_{\text{sing.}}=\frac{\theta'}{z'-z_F}-\frac{1/2\theta_F}{z'-z_F}\left(\sqrt{z_F-z_0}+\frac{1}{\sqrt{z_F-z_0}}\right) \tag{D7}$$

and

$$(g_h^{F1/2}(w,\xi|z',\theta')_2)_{\text{sing.}}=(g_h^{F1/2}(w,\xi|z',\theta')_3)_{\text{sing.}}=0. \tag{D8}$$

Therefore, we obtain from Eqs. (3.15″), (D7), and (D8) that

$$(g_h^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}=\frac{\theta'}{z'-z_F}-\frac{\frac{1}{2}\theta}{z'-z_F}\left(\sqrt{z_F-z_0}+\frac{1}{\sqrt{z_F-z_0}}\right), \tag{D9}$$

the right-hand side of which should be noticed to be *independent* of ϵ_h^j. On the other hand Eqs. (3.15″) and (D4)–(D6) give

$$(g_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}=\frac{\frac{1}{2}\theta'\theta_F}{(z'-z_F)^2}+\frac{1/4}{z'-z_F}\left(\sqrt{z_F-z_0}+\frac{1}{\sqrt{z_F-z_0}}\right), \tag{D10}$$

the right-hand side of which should be noticed to be *independent* of ϵ_h^j. Furthermore, with the help of formulas

$$\frac{1}{(z'-z_F)(z'-z'_F)}=\frac{1}{z_F-z'_F}\left(\frac{1}{z'-z_F}-\frac{1}{z'-z'_F}\right) \tag{D11}$$

and

$$\frac{1}{(z'-z_F)^2(z'-z'_F)}=\frac{-1}{(z_F-z'_F)^2}\left(\frac{1}{z'-z_F}-\frac{1}{z'-z'_F}\right)+\frac{1}{(z_F-z'_F)(z'-z_F)^2}, \tag{D11$'$}$$

we find from Eqs. (B4) to (B6) that non-N operator $(g_h^{jjF}(\ )_\iota)_{\text{sing.}}$ (for $\iota=1-3$) is given as follows:

$$\begin{aligned}(g_h^{jjF}(w,\xi;w',\theta'|z',\theta')_1)_{\text{sing.}}=&\frac{\theta'-1/2(z_F-z_0)^{\epsilon_h^j}\theta_F-1/2(z'_F-z_0)^{-\epsilon_h^j}\theta'_F}{(z'-z_F)(z_F-z'_F)}\\&+\frac{\theta'\theta_F\theta'_F}{(z'-z_F)(z_F-z'_F)^2}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\\&-\frac{\frac{1}{2}\theta'\theta_F\theta'_F}{(z'-z_F)^2(z_F-z'_F)}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j},\end{aligned} \tag{D12}$$

$$\begin{aligned}(g_h^{jjF}(w,\xi;w',\theta'|z',\theta')_2)_{\text{sing.}}=&-\frac{\theta'\theta_F\theta'_F\epsilon_h^j}{(z'-z_F)(z_F-z'_F)(z_F-z_0)}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\\&+(\text{terms obtained by } z_F\leftrightarrow z'_F,\ \theta_F\leftrightarrow\theta'_F \text{ and } \epsilon_h^j\rightarrow-\epsilon_h^j)\end{aligned} \tag{D13}$$

and

$$\begin{aligned}(g_h^{jjF}(w,\xi;w',\theta'|z',\theta')_3)_{\text{sing.}}=&\frac{1/2(z_F-z_0)^{\epsilon_h^j}\theta_F+1/2(z'_F-z_0)^{-\epsilon_h^j}\theta'_F}{(z'-z_F)(z_F-z'_F)}\\&-\frac{1/4}{(z'-z_F)(z_F-z'_F)}\left(\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\theta'_F+\theta_F\right)\\&\times\left(\sqrt{z_F-z_0}+\frac{1}{\sqrt{z_F-z_0}}\right)\\&+(\text{terms obtained by } z_F\leftrightarrow z'_F,\ \theta_F\leftrightarrow\theta'_F \text{ and } \epsilon_h^j\rightarrow-\epsilon_h^j).\end{aligned} \tag{D14}$$

Summing up Eqs. (D12)–(D14), we find from Eq. (3.13″) that

$$\begin{aligned}&(g_h^{jjF}(w,\xi;w',\theta'|z',\theta'))_{\text{sing.}}\\&=\frac{\theta'}{z'-z_F}\left(\frac{1}{z_F-z'_F}\left(1-\frac{\epsilon_h^j\theta_F\theta'_F}{z_F-z_0}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\right)+\frac{\theta_F\theta'_F}{(z_F-z'_F)^2}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\right)\\&-\frac{1/4}{(z'-z_F)(z_F-z'_F)}\left(\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\theta'_F+\theta_F\right)\left(\sqrt{z_F-z_0}+\frac{1}{\sqrt{z_F-z_0}}\right)\\&-\frac{1/2\theta'\theta_F\theta'_F}{(z'-z_F)^2(z_F-z'_F)}\left(\frac{z_F-z_0}{z'_F-z_0}\right)^{\epsilon_h^j}\\&+(\text{terms obtained by } z_F\leftrightarrow z'_F,\ \theta_F\leftrightarrow\theta'_F \text{ and } \epsilon_h^j\rightarrow-\epsilon_h^j).\end{aligned} \tag{D15}$$

It is remarkable that Eq. (D15) can be shown to be rewritten into Eq. (3.16), on the right-hand side of which the singular terms $(g_h^{F1/2}(\ ))_{\text{sing.}}$ (D9) and $(g_h^{F1}(\ ))_{\text{sing.}}$ (D10) have been found to be *independent* of ϵ_h^j.

APPENDIX E: ABSENCE OF SINGULARITIES AT $z'=z_F(w,\xi)$

In this Appendix, we shall prove that both $(H^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}$ $(H^{F1}(w,\xi|z',\theta'))_{\text{sing.}}$ in Eq. (3.18) are zero. First, we notice that non-N operator $f_+^F(z',\theta'|z_0,\theta_0)$ (A10) can be expanded as

$$\begin{aligned} f_+^F(z',\theta'|z_0,\theta_0)=&(1+(\theta'-\theta_F)D_{\theta_F}-\theta'\theta_F\partial_{z_F})f_+^F(z_F,\theta_F|z_0,\theta_0)+(z'-z_F)(1+(\theta'-\theta_F)D_{\theta_F}\\ &-\theta'\theta_F\partial_{z_F})\partial_{z_F}f_+^F(z_F,\theta_F|z_0,\theta_0))+\frac{(z'-z_F)^2}{2}(1+(\theta'-\theta_F)D_{\theta_F}\\ &-\theta'\theta_F\partial_{z_F})\partial_{z_F}^2f_+^N(z_F,\theta_F|z_0,\theta_0))+O((z'-z_F)^3). \end{aligned} \tag{E1}$$

Furthermore, with the help of

$$f^F(z_F,\theta_F|z_0,\theta_0)\equiv\frac{X^{+0}(z_F,\theta_F)}{ip^{+0}}-\log(z_F-z_0-\sqrt{z_F-z_0}\,\theta_F\theta_0), \tag{E2}$$

we find that

$$\begin{aligned} &f^F(z_F,\theta_F|z_0,\theta_0)-\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)\\ &=\frac{Y_+^{+0}(z_F,\theta_F;z_0)}{ip^{+0}}-\log(z_F-z_0-\theta_F\theta_0)(\equiv f_+^F(z_F,\theta_F|z_0,\theta_0)), \end{aligned} \tag{E3}$$

on the left-hand side of which we have from Eqs. (E2) and (1.18) the following relation.

$$f^F(z_F,\theta_F|z_0,\theta_0)=-\log w+\frac{X^{+0}(z_0)}{ip^{+0}}. \tag{E4}$$

We can derive from Eq. (E3)

$$D_{\theta_F}f_+^F(z_F,\theta_F|z_0,\theta_0)=\frac{1}{\sqrt{z_F-z_0}}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)+\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)\partial_{z_F}f^F(z_F,\theta_F|z_0,\theta_0), \tag{E5}$$

$$\begin{aligned} \partial_{z_F}f_+^F(z_F,\theta_F|z_0,\theta_0)=&\partial_{z_F}f^F(z_F,\theta_F|z_0,\theta_0)\\ &-\frac{\frac{1}{2}\theta_F}{(\sqrt{z_F-z_0})^3}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)\\ &-\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)\partial_{z_F}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0), \end{aligned} \tag{E6}$$

$$\begin{aligned} &\partial_{z_F}D_{\theta_F}f_+^F(z_F,\theta_F|z_0,\theta_0)\\ &=-\frac{1/2}{(\sqrt{z_F-z_0})^3}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)+\frac{1}{\sqrt{z_F-z_0}}\partial_{z_F}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)\\ &+\frac{\frac{1}{2}\theta_F}{(\sqrt{z_F-z_0})^3}\partial_{z_F}f^F(z_F,\theta_F|z_0,\theta_0)+\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)\partial_{z_F}^2f^F(z_F,\theta_F|z_0,\theta_0), \end{aligned} \tag{E7}$$

$$\partial_{z_F}^2 f_+^F(z_F,\theta_F|z_0,\theta_0)=\partial_{z_F}^2 f^F(z_F,\theta_F|z_0,\theta_0)+\frac{\frac{3}{4}\theta_F}{(\sqrt{z_F-z_0})^5}\,D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)$$
$$-\frac{\theta_F}{(\sqrt{z_F-z_0})^3}\,\partial_{z_F}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)$$
$$-\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)\partial_{z_F}^2 D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0), \tag{E8}$$

and

$$\partial_{z_F}^2 D_{\theta_F}f_+^F(z_F,\theta_F|z_0,\theta_0)$$
$$=\frac{3/4}{(\sqrt{z_F-z_0})^5}\,D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)-\frac{1}{(\sqrt{z_F-z_0})^3}\,\partial_{z_F}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)$$
$$+\frac{1}{\sqrt{z_F-z_0}}\,\partial_{z_F}^2 D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)-\frac{3/4\theta_F}{(\sqrt{z_F-z_0})^5}\,\partial_{z_F}f^F(z_F,\theta_F|z_0,\theta_0)$$
$$+\frac{\theta_F}{(\sqrt{z_F-z_0})^3}\,\partial_{z_F}^2 f^F(z_F,\theta_F|z_0,\theta_0)+\left(\theta_F-\frac{\theta_F}{\sqrt{z_F-z_0}}\right)\partial_{z_F}^3 f^F(z_F,\theta_F|z_0,\theta_0). \tag{E9}$$

On the right-hand sides of Eqs. (E5)–(E9), we can use the following expressions which can be derived from Eqs. (2.19) and (E4) [together with Eq. (1.6)]:

$$D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)=-\frac{\xi}{wD_\xi\theta_F}=-\frac{\xi}{wK_F(w)}, \tag{E10}$$

$$\partial_{z_F}f^F(z_F,\theta_F|z_0,\theta_0)=-\frac{1}{w(D_\xi\theta_F)^2}-\frac{\xi\partial_w\theta_F}{w(D_\xi\theta_F)^3}=-\frac{1}{wK_F^2(w)}+\frac{\xi\partial_w J_F(w)}{wK_F^3(w)}, \tag{E11}$$

$$\partial_{z_F}D_{\theta_F}f^F(z_F,\theta_F|z_0,\theta_0)=\frac{\partial_w\theta_F}{w(D_\xi\theta_F)^4}+\xi\left(\frac{\partial_w\theta_F}{w^2(D_\xi\theta_F)^3}+\frac{\partial_w D_\xi\theta_F}{w(D_\xi\theta_F)^4}\right)$$
$$=\frac{\partial_w J_F(w)}{wK_F^4(w)}+\xi\left(\frac{2\partial_w K_F(w)}{wK_F^4(w)}+\frac{1}{w^2K_F^3(w)}\right), \tag{E12}$$

and

$$\partial_{z_F}^2 f^F(z_F,\theta_F|z_0,\theta_0)$$
$$=\frac{1}{w^2(D_\xi\theta_F)^4}+\frac{2\partial_w D_\xi\theta_F}{w(D_\xi\theta_F)^5}+\xi\left(\frac{2\partial_w\theta_F}{w^2(D_\xi\theta_F)^5}-\frac{\partial_w^2\theta_F}{w(D_\xi\theta_F)^5}+\frac{4\partial_w\theta_F\partial_w D_\xi\theta_F}{w(D_\xi\theta_F)^6}\right)$$
$$=\frac{1}{w^2K_F^4(w)}+\frac{2\partial_w K_F(w)}{wK_F^5(w)}+\xi\left(-\frac{2\partial_w J_F(w)}{w^2K_F^5(w)}+\frac{\partial_w^2 J_F(w)}{wK_F^5(w)}-\frac{6\partial_w J_F(w)\partial_w K_F(w)}{wK_F^6(w)}\right). \tag{E13}$$

Non-N-operator $(G^{F(1/2)(1)}(w,\xi|z',\theta')_\iota)_{\text{sing.}}$ is calculated by using Eqs. (A8)–(A10) [(A11)–(A13)], with the help of Eq. (E1) together with Eqs. (E5)–(E13). After extremely tedious but straightforward calculations, we obtain the following final results:

$$(G^{F1/2}(w,\xi|z',\theta')_1)_{\text{sing.}}$$

$$=\frac{1}{z'-z_F}\begin{pmatrix} -\theta'+1/2\theta_F\left(\sqrt{z_F-z_0}+\dfrac{1}{\sqrt{z_F-z_0}}\right)-\xi J_F(w)K_F(w)\dfrac{\theta'}{z_F-z_0} \\ +1/2\xi K_F(w)\left(\sqrt{z_F-z_0}-\dfrac{1}{\sqrt{z_F-z_0}}\right)\end{pmatrix}, \tag{E14}$$

$$(G^{F1/2}(w,\xi|z',\theta')_2)_{\text{sing.}}=\frac{1}{z'-z_F}\left(\xi J_F(w)K_F(w)\frac{\theta'}{z_F-z_0}\right) \tag{E15}$$

and

$$(G^{F1/2}(w,\xi|z',\theta')_3)_{\text{sing.}}=\frac{-1}{z'-z_F}\left(1/2\xi K_F(w)\left(\sqrt{z_F-z_0}-\frac{1}{\sqrt{z_F-z_0}}\right)\right). \tag{E16}$$

Furthermore, we find that

$$(G^{F1}(w,\xi|z',\theta')_1)_{\text{sing.}}=\frac{1}{(z'-z_F)^2}\left(-\frac{1}{2}\theta'\theta_F+\frac{1}{4}\xi J_F(w)K_F(w)\left(\sqrt{z_F-z_0}-\frac{1}{\sqrt{z_F-z_0}}\right)\right)$$

$$+\frac{1}{z'-z_F}\begin{pmatrix} \dfrac{1}{8}\xi J_F(w)K_F(w)\dfrac{1}{(\sqrt{z_F-z_0})^3} \\ +\dfrac{\frac{1}{2}\theta'}{z_F-z_0}J_F(w)\left(-1+\dfrac{\xi\partial_w J_F(w)}{K_F(w)}\right) \\ +\dfrac{1}{8}\xi J_F(w)K_F(w)\dfrac{1}{\sqrt{z_F-z_0}} \\ +\dfrac{1}{4}\left(-1+\dfrac{\xi\partial_w J_F(w)}{K_F(w)}\right)\left(\sqrt{z_F-z_0}-\dfrac{1}{\sqrt{z_F-z_0}}\right) \\ -\dfrac{1}{4}\left(\sqrt{z_F-z_0}+\dfrac{1}{\sqrt{z_F-z_0}}\right)\end{pmatrix} \tag{E17}$$

$$(G^{F1}(w,\xi|z',\theta')_2)_{\text{sing.}}=-\frac{\frac{1}{2}\theta'}{(z'-z_F)(z_F-z_0)}J_F(w)\left(-1+\frac{\xi\partial_w J_F(w)}{K_F(w)}\right) \tag{E18}$$

and

$$(G^{F1}(w,\xi|z',\theta')_3)_{\text{sing.}}=-\frac{1/4}{(z'-z_F)^2}\xi J_F(w)K_F(w)\left(\sqrt{z_F-z_0}-\frac{1}{\sqrt{z_F-z_0}}\right)$$

$$+\frac{1}{z'-z_F}\begin{pmatrix} -\dfrac{1}{8}\xi J_F(w)K_F(w)\dfrac{1}{(\sqrt{z_F-z_0})^3} \\ -\dfrac{1}{8}\xi J_F(w)K_F(w)\dfrac{1}{\sqrt{z_F-z_0}} \\ -\dfrac{1}{4}\left(-1+\dfrac{\xi\partial_w J_F(w)}{K_F(w)}\right)\left(\sqrt{z_F-z_0}-\dfrac{1}{\sqrt{z_F-z_0}}\right)\end{pmatrix}. \tag{E19}$$

Therefore, we find that

$$(G^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}=-(g_h^{F1/2}(w,\xi|z',\theta'))_{\text{sing.}}, \tag{E20}$$

the right- (left-) hand side of which is given by Eq. (D10) [Eq. (3.12′) with Eqs. (E14)–(E16)]. Furthermore, we find that

$$(G^{F1}(w,\xi|z',\theta'))_{\text{sing.}}=-(g_h^{F1}(w,\xi|z',\theta'))_{\text{sing.}}, \tag{E21}$$

the right- (left-) hand side of which is given by Eq. (D11) [Eq. (3.12′) with Eqs. (E17)–(E19)].

[1] S. Naito, J. Math. Phys. **38**, 4980 (1997).
[2] P. Ramond, Phys. Rev. D **3**, 2415 (1971).
[3] A. Neveu and J. H. Schwarz, Nucl. Phys. B **31**, 86 (1971); Phys. Rev. D **4**, 1109 (1971).
[4] S. Naito, J. Math. Phys. **34**, 2166 (1993).
[5] D. Friedan, E. Martinec, and S. Schenker, Nucl. Phys. B **271**, 93 (1986).
[6] F. Gliozzi, J. Scherk, and D. Olive, Nucl. Phys. B **122**, 253 (1977).
[7] S. Naito, J. Math. Phys. **39**, 1402 (1998).
[8] M. B. Green and J. H. Schwarz, Nucl. Phys. B **181**, 502 (1981).
[9] A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, Nucl. Phys. B **241**, 333 (1983).
[10] E. Witten, Nucl. Phys. B **276**, 291 (1986).
[11] C. Wendt, Nucl. Phys. B **314**, 209 (1989).
[12] C. R. Preitschopf, C. B. Thorn, and S. Yost, Nucl. Phys. B **337**, 363 (1990).
[13] S. Naito, J. Math. Phys. **38**, 1413 (1997).
[14] M. B. Green and J. H. Schwarz, Nucl. Phys. B **189**, 252 (1982); **198**, 441 (1982); Phys. Lett. **109B**, 444 (1982). As for quantum superstring field theory (in GS superstring), see M. B. Green and J. H. Schwarz, Nucl. Phys. B **219**, 437 (1983); **243**, 475 (1984), and various papers cited therein.
[15] N. Berkovits, Nucl. Phys. B **420**, 332 (1994).

Lie-series approach to the evolution of resonant and nonresonant anharmonic oscillators in quantum mechanics

A. W. Sáenz
Department of Physics, Catholic University, Washington, DC 20064
and Naval Research Laboratory, Washington, DC 20375

(Received 30 April 1996; accepted for publication 4 September 1997)

A version of a normal-form method of Kummer and Gompa, analogous to the Lie-series methods of classical mechanics, is used to rigorously construct uniform long-time approximations to the quantum evolution of a system of resonant or nonresonant anharmonic oscillators. The Hamiltonian H_ε governing the system is a self-adjoint operator in the Hilbert space $\mathscr{H}=L^2(\mathbf{R}^\nu)$ $(\nu\geqslant 2)$, given formally by $H_0+\varepsilon V$. Here, H_0 denotes the Hamiltonian of ν one-dimensional simple harmonic oscillators and εV their anharmonic coupling, $\varepsilon\geqslant 0$ being a small parameter and V an operator of multiplication by a smooth function of polynomial growth at infinity. We treat the case in which an arbitrary number $\rho\geqslant 1$ of the frequencies of these oscillators are rationally independent, imposing a standard diophantine condition on the independent frequencies if $\rho\geqslant 2$. Under these assumptions, stated precisely in the paper, an Nth-order approximant $\psi_N(t,\varepsilon)$ to the exact solution $\psi(t,\varepsilon)$ of the Schrödinger equation $id\psi(t,\varepsilon)/dt=H_\varepsilon\psi(t,\varepsilon)$, satisfying the initial condition $\psi(0,\varepsilon)=\psi_0$, is constructed inductively, where the initial state ψ_0 belongs to a family of smooth functions dense in $\mathscr{H}$. The main result is that $\psi_N(t,\varepsilon)$ differs in $\mathscr{H}$-norm from $\psi(t,\varepsilon)$ by $\leqslant \text{const}\,\varepsilon^{N+1}|t|$ for $t\in\mathbf{R}$ and ε in an arbitrary compact interval $[0,\varepsilon_0]$. This result is obtained by an approach simpler and quite different from that of Kummer and Gompa, and extends their work to oscillators with nonpolynomial couplings, under very general resonance conditions. © *1998 American Institute of Physics.* [S0022-2488(98)01203-1]

I. INTRODUCTION

Two normal-form methods for obtaining uniform long-time approximations to the quantum evolution of a system of harmonic oscillators, coupled by a potential which is a polynomial in coordinates and momenta, have been proposed by Kummer and Gompa.[1,2] The first method is close in spirit to the perturbation theory discussed by Born and Jordan[3] in the early days of quantum mechanics. The second, and presumably more accurate method, is based on a Lie-series approach by which the Hamiltonian of the coupled system is reduced to the pertinent normal form to any desired order by a suitable product of exponential operators.[4] This approach is analogous to those used in mathematical studies of vector fields and classical Hamiltonian systems,[5–8] and in related physical applications.[9] It is elegantly formulated in the setting of Bargmann's space of entire analytic functions.[10] Both procedures have the merit of avoiding the secular terms which plague the usual time-dependent perturbation theory in quantum mechanics (see also Refs. 11 and 12), thus limiting its validity to relatively short time intervals. However, the restrictions on the allowed initial states imposed in these procedures severely limit their usefulness for dealing with resonant cases.[13]

In a recent paper,[14] we developed a rigorous procedure for approximating uniformly over long times the quantum evolution of a wide class of systems of resonant and nonresonant harmonic oscillators coupled by potentials with polynomial growth at infinity, without necessarily being polynomials. The approach in that paper was a rigorous version of the first normal-form method in Refs. 1 and 2, but our methods of proof were quite different from the ones used in these references.[15]

In the present paper, we use a modification of the second method to approximate the long-time

 © 1998 American Institute of Physics

evolution of such systems. The time-development of interest is defined by the initial-value problem (IVP)[16]

$$i\,\frac{d\psi(t,\varepsilon)}{dt}=H_\varepsilon\psi(t,\varepsilon), \tag{1.1a}$$

$$\psi(0,\varepsilon)=\psi_0. \tag{1.1b}$$

Here, H_ε denotes a self-adjoint operator acting in the Hilbert space $\mathscr{H}=L^2(\mathbf{R}^\nu)$ $(\nu\geqslant 2)$ and is given formally by

$$H_\varepsilon=H_0+\varepsilon V, \tag{1.2}$$

and d/dt denotes the strong t-derivative in $\mathscr{H}$. In (1.2), H_0 stands for the Hamiltonian of ν uncoupled harmonic oscillators, or more exactly, for the unique self-adjoint operator on $\mathscr{H}$ which acts as

$$H_0=\frac{1}{2}\sum_{i=1}^{\nu}\left(-\frac{\partial^2}{\partial x_i^2}+\omega_i^2x_i^2\right) \tag{1.3}$$

on $C_0^\infty(\mathbf{R}^\nu)$. Moreover, $\varepsilon\geqslant 0$ is a small parameter and V an operator of multiplication by a smooth function which together with its derivatives has polynomial growth at infinity, again without necessarily being a polynomial. The frequencies ω_i are assumed to be fixed positive numbers, an arbitrary number ρ of them being rationally independent. If $\rho=1$, i.e., if there is a positive real ω_0 such that $\omega_i=k_i\omega_0$ $(i=1,...,\nu)$, the k_i being positive integers, then one has the analog of the classical $k_1:k_2:\ldots:k_\nu$ resonance. In this case the k_i are allowed to be completely arbitrary. If $\rho\geqslant 2$, we subject the rationally independent frequencies to a standard diophantine condition. No other requirements are imposed on the frequencies. Finally, the initial state ψ_0 is assumed to lie in a suitable dense subset of $\mathscr{H}$, a condition which is not unduly restrictive physically.

The Lie-series formalism discussed in this paper is an inductive procedure for reducing H_ε to the desired normal form to $O(\varepsilon^N)$ $(N\geqslant 1)$ in a sense explained heuristically in Sec. II and rigorously in Secs. III and IV. We define an Nth-order approximant $\psi_N(t,\varepsilon)$ $(N\geqslant 1)$ to the exact solution $\psi(t,\varepsilon)$ of the IVP (1.1) in terms of this normal form and prove that it satisfies the uniform error estimate

$$\|\psi(t,\varepsilon)-\psi_N(t,\varepsilon)\|\leqslant \text{const}\ \varepsilon^{N+1}|t| \tag{1.4}$$

for $t\in\mathbf{R}$, $0\leqslant\varepsilon\leqslant\varepsilon_0$, where ε_0 is an arbitrary positive constant and $\|.\|$ denotes the usual $L^2(\mathbf{R}^\nu)$-norm. This estimate is our main result. It improves the corresponding Nth-order error estimate $O(\varepsilon^{N+1}(|t|+1))$ obtained in Ref. 14 by using a modification of the first normal-form method. The proof of (1.4) is lengthy, but the basic ideas underlying it are simple, as will be clear from its heuristic proof given near the end of Sec. II.

Reference 14 and the present work are the first to have discussed long-time approximations to the evolution of systems of resonant and nonresonant harmonic oscillators with nonpolynomial couplings under such general conditions on the frequencies. The methods of proof used in this paper are quite different and conceptually simpler than those in Refs. 1 and 2. In the first place, our discussions are carried out in the framework of the L^2-space $\mathscr{H}$, rather than in the corresponding Bargmann space, which is unsuitable for dealing with a Hamiltonian of type (1.2) equipped with a local potential V of the present degree of generality. Second, a key role in the paper is played by an associative algebra $\mathscr{M}$ of infinite matrices and a set M of generally unbounded operators, which are those represented by these matrices in a suitable basis, and to which H_0, V, and most of the other operators of the theory belong. The operators in M have the important property of leaving invariant a common dense subset of $\mathscr{H}$ containing the initial state ψ_0.[17] Moreover, when restricted to this subset they constitute an algebra M_0 isomorphic to $\mathscr{M}$. This isomorphism is not a mere mathematical curiosity, but plays an essential role in allowing one to prove simply and rigorously that certain key relations satisfied by operators in M hold. We note that the simple proof given in this paper of the essential self-adjointness of the truncated normal

form $K^{[N]}(\varepsilon)$ of interest here is based on the fact that its unique self-adjoint extension is an operator associated with a suitable matrix in $\mathscr{M}$. The existence of such an extension is needed to define the approximant $\psi_N(t,\varepsilon)$ in a satisfactory manner, as well as for other reasons.

The paper is organized as follows. Leaving rigor aside, Sec. II is devoted to a discussion of the normal-form method used in the present study, as well as to a heuristic construction of $\psi_N(t,\varepsilon)$. These considerations motivate the precise mathematical approach followed in Secs. III and IV. In the former section, we state our hypotheses on V and the frequencies ω_i, specify the set $\mathscr{M}$ and define key operators of the theory in terms of the matrices therein, give a precise definition of $\psi_N(t,\varepsilon)$, and state Theorem 1, which asserts that (1.4) holds at the above mentioned t,ε -values under these hypotheses. It is proved in Sec. IV on the basis of the technical Lemmas 1–9. The elementary Lemma 1 serves to control the small denominators which would otherwise defeat our normal-form approach, while Lemma 2 entails that V lies in M. Both of these results were proved in Ref. 14 and are essential first steps in the proof of Theorem 1. The invariance of the previously mentioned dense subset under M is stated in Lemma 4 and the above isomorphism result is contained in Lemma 5, whose main relevance to the paper lies in the corollary to these two lemmas which is included in this section. The corollary asserts that if a finite number of matrices in $\mathscr{M}$ satisfy an equation of suitable type, then the operators in M which they represent obey a similar equation on this dense subset. We illustrate this corollary by a simple example and then apply it to prove Lemma 6, which states that the key operator relations alluded to in the previous paragraph hold. In a first reading of the paper, the short proof of Theorem 1, placed near the end of Sec. IV, could be immediately perused after reading the statement of Lemma 9, since it is the only lemma in the section which is directly involved in the proof.

In addition to leading to this theorem, our Lie-series approach promises to be an effective vehicle for generalizing the exact quantization results of Degli Esposti, Graffi, and Herczinski[4] to resonant oscillator systems of the type considered here.

II. FORMAL CONSTRUCTION OF THE *N*TH-ORDER APPROXIMANT

This section is devoted to a heuristic review of a version of the second normal-form method alluded to in Sec. I.

We begin by introducing some notation. $\mathbf{N}$ denotes the positive integers and $\mathbf{Z}_+^p$ $(p\in\mathbf{N})$ stands for the set of all vectors $r=(r_i,\ldots,r_p)$ whose components r_i are non-negative integers. We denote by $\Phi=\{\Phi_m\}$ $(m\in\mathbf{Z}_+^\nu)$ the complete orthonormal set of eigenfunctions of H_0 of the form

$$\Phi_m(x_1,\ldots,x_\nu)=\prod_{i=1}^{\nu}\varphi_{m_i}^{(i)}(x_i), \tag{2.1a}$$

where $\varphi_{m_i}^{(i)}(x_i)$ is the usual (real) eigenfunction of $\frac{1}{2}(-(d^2/dx_i^2)+\omega_i^2x_i^2)$ $(i=1,...,\nu)$ having unit $L^2(\mathbf{R})$-norm. Hence for all $m\in\mathbf{Z}_+^\nu$,

$$H_0\Phi_m=\lambda_m\Phi_m, \tag{2.1b}$$

where

$$\lambda_m=\sum_{i=1}^{\nu}\omega_i(m_i+1/2). \tag{2.1c}$$

Since some of the frequencies $\omega_1,\ldots,\omega_\nu$ may be rationally dependent, the eigenvalues λ_m of H_0 are generally degenerate.

Let $\mathscr{L}(\mathscr{H})$ be the set of linear operators on $\mathscr{H}$, i.e., those whose domains and ranges are in $\mathscr{H}$. For every $A\in\mathscr{L}(\mathscr{H})$, we define an operator ad A which acts on each $B\in\mathscr{L}(\mathscr{H})$ by ad $A(B)=$ad $AB=[A,B]=AB-BA$. In this section, we will use the splitting

$$A=\widetilde{A}+\hat{A}, \tag{2.2}$$

for operators $A\in\mathscr{L}(\mathscr{H})$, ignoring domain questions. Here $\widetilde{A}$ and $\hat{A}$ are operators in $\mathscr{L}(\mathscr{H})$ with matrix elements $\widetilde{A}_{mn}=\langle\Phi_m,\widetilde{A}\Phi_n\rangle$, $\hat{A}_{mn}=\langle\Phi_m,\hat{A}\Phi_n\rangle$ given by

$$\tilde{A}_{mn}=(1-\delta_{\lambda_m,\lambda_n})A_{mn}, \quad \hat{A}_{mn}=\delta_{\lambda_m,\lambda_n}A_{mn}, \tag{2.3a}$$

where $A_{mn}=\langle\Phi_m,A\Phi_n\rangle$.[18] Hence, the splitting (2.2) induces the decomposition.

$$A_{mn}=\tilde{A}_{mn}+\hat{A}_{mn} \tag{2.3b}$$

of each of these matrix elements. We define X (resp., Y) as the set composed of operators $A \in \mathscr{L}(\mathscr{H})$ for which $\tilde{A}=0$ (resp., $\hat{A}=0$). Roughly speaking, X and Y are the image and kernel of ad H_0, but to make this statement exact would require refinements of the relevant definitions which would be out of place in this section.

We will denote by M^∞ the set of formal power series

$$B(\varepsilon)=\sum_{i=0}^{\infty}\varepsilon^i B^{(i)} \tag{2.4}$$

whose coefficients $B^{(i)}$ are in $\mathscr{L}(\mathscr{H})$. On the other hand, N^∞ will denote the set of such formal series with coefficients which are linear operators (superoperators) mapping $\mathscr{L}(\mathscr{H})$ into itself. In both cases, we will write $B^{[i]}(\varepsilon)$ for the ith-order truncation of $B(\varepsilon)$

$$B^{[i]}(\varepsilon)=\sum_{j=0}^{i}\varepsilon^j B^{(j)}. \tag{2.5}$$

The normal-form method advocated in the present paper can be succinctly described as a procedure for finding operators $S^{(i)}\in X$ and $K^{(i)}\in Y$ $(i=1,...,N)$, such that H_ε is reduced to normal form to order ε^N for some $N\geqslant 1$. By this we mean that the equation

$$\begin{aligned}&\exp(\varepsilon^N S^{(N)})\cdots\exp(\varepsilon S^{(1)})H_\varepsilon\ \exp(-\varepsilon S^{(1)})\cdots\exp(-\varepsilon^N S^{(N)})\\ &\quad=K^{[N]}(\varepsilon)+O(\varepsilon^{N+1})\end{aligned} \tag{2.6}$$

holds for this N in the sense of formal power series. Here

$$K^{[N]}(\varepsilon)=\sum_{i=0}^{N}\varepsilon^i K^{(i)} \tag{2.7}$$

is the Nth order truncation of the pertinent normal form in M^∞ and $O(\varepsilon^{N+1})$ denotes a series in M^∞ in which the coefficients of ε^i for $i\leqslant N$ are the zero operator. Later in this section, we will consider briefly the relation between this normal form and Rayleigh–Schrödinger perturbation theory.

We begin by discussing the method in the simplest case: $N=1$. Setting $N=1$ in (2.6) and equating the coefficients of ε^0, ε^1 on both sides of this equation, we obtain

$$K^{(0)}=H_0, \tag{2.8a}$$

$$\text{ad } H_0 S^{(1)}=V-K^{(1)}. \tag{2.8b}$$

We require that $S^{(1)}$, besides belonging to X, should satisfy

$$\text{ad } H_0 S^{(1)}=\tilde{V}. \tag{2.8c}$$

This equation and the condition $S^{(1)}\in X$ determine $S^{(1)}$ uniquely, at least in a formal sense, since they completely determine the matrix elements of this operator in the Φ basis:

$$\langle\Phi_m,S^{(1)}\Phi_n\rangle=\begin{cases}\langle\Phi_m,V\Phi_n\rangle/(\lambda_m-\lambda_n) & (\lambda_m\neq\lambda_n),\\ 0 & (\lambda_m=\lambda_n).\end{cases} \tag{2.9}$$

By (2.2), (2.8b), and (2.8c),

$$K^{(1)}=\hat{V}. \tag{2.8d}$$

We next consider the present method in the general case $N\geqslant 1$. We first rewrite (2.6) in the form

$$H_\varepsilon=R_N(\varepsilon)K^{[N]}(\varepsilon)+\varepsilon^{N+1}Z_{N+1}(\varepsilon), \tag{2.10}$$

where

$$R_N(\varepsilon)=\exp(-\varepsilon\ \mathrm{ad}\ S^{(1)})\cdots\exp(-\varepsilon^N\ \mathrm{ad}\ S^{(N)}) \tag{2.11}$$

and $Z_{N+1}(\varepsilon)\in M^\infty$.[19] To arrive at (2.10), we have used the relation[20]

$$\exp(\alpha\ \mathrm{ad}\ B)A=\exp(\alpha B)A\ \exp(-\alpha B)\quad(\alpha\in\mathbf{R};A,B\in\mathscr{L}(\mathscr{H})). \tag{2.12}$$

The requirement (2.10) can be replaced by two conditions. The first includes all terms of $O(\varepsilon^k)$ in (2.10) with $k\leqslant 2N$ and the second all those with $k>2N$. The first condition can be expressed in the form

$$H_\varepsilon=R_N^{[N]}(\varepsilon)K^{[N]}(\varepsilon)+\varepsilon^{N+1}W_{N+1}(\varepsilon), \tag{2.13}$$

where $W_{N+1}(\varepsilon)=Z_{N+1}^{[N-1]}(\varepsilon)$ and where, of course, $R_N^{[N]}(\varepsilon)$ denotes the truncation of the formal power series in N^∞ symbolized by (2.11) after the term of order ε^N. As we will show in a moment, (2.13) completely determines a truncated normal form $K^{[N]}(\varepsilon)$ with the desired properties. The second condition merely determines the coefficients $Z_{N+1}^{(k)}$ with $k\geqslant N$ in the formal series $Z_{N+1}(\varepsilon)$ in terms of the $K^{(i)}$ with $i\leqslant N$, and will be of no further interest here. We write

$$W_{N+1}(\varepsilon)=\sum_{i=0}^{N-1}\varepsilon^i W_{N+1}^{(i)}, \tag{2.14}$$

the operators $W_{N+1}^{(i)}$ being, of course, ε-independent.

We claim that (2.13) holds if the polynomials $R_N^{[N]}(\varepsilon),K^{[N]}(\varepsilon),W_{N+1}(\varepsilon)$ in ε are constructed in terms of the operators $S^{(i)},K^{(i)},W_{i+1}^{(j)}$ which will now be specified. $S^{(i)}$ is defined as the unique solution in X of the equation

$$\mathrm{ad}\ H_0S^{(i)}=V_i-K^{(i)}\quad(i\geqslant 1), \tag{2.15a}$$

K_0 is given by (2.8b), and

$$K^{(i)}=\hat{V}_i\quad(i\geqslant 1), \tag{2.15b}$$

whence $K^{(i)}\in Y$ for $i\geqslant 0$. In (2.15a) and (2.15b),

$$V_1=V, \tag{2.16a}$$

$V^{(i)}(i\geqslant 2)$ is defined recursively in terms of $S^{(j)},K^{(k)}(1\leqslant j\leqslant i-1,0\leqslant k\leqslant i-1)$ by

$$V_i=-\sum_{j=1}^{i}R_{i-1}^{(j)}K^{(i-j)}\quad(i\geqslant 2) \tag{2.16b}$$

and we define the operators $W_{i+1}^{(j)}\in\mathscr{L}(\mathscr{H})$ by

$$W_{i+1}^{(j)}=-\sum_{k=j+1}^{i}R_i^{(k)}K^{(i+j-k+1)}\quad(0\leqslant j\leqslant i-1,i\geqslant 1). \tag{2.15c}$$

Here, $R_j^{(k)}$ denotes the coefficient of ε^k $(k\geqslant 0)$ in the formal-power-series representation of the operator $R_j(\varepsilon)\in N^\infty$, and is given by

$$R_j^{(0)}A=A \quad (j\geqslant 1,\ A\in\mathscr{L}(\mathscr{H})), \tag{2.17a}$$

$$R_j^{(k)}A=\sum_{n\in\mathbf{Z}_+^j,\,s(n)=k}\frac{(-1)^{|n|}}{n!}\,(\mathrm{ad}\,S^{(1)})^{n_1}\cdots(\mathrm{ad}\,S^{(j)})^{n_j}A \quad (j\geqslant 1,k\geqslant 1,A\in\mathscr{L}(\mathscr{H})), \tag{2.17b}$$

where $|n|=\Sigma_{r=1}^{j}n_r$, $n!=n_1!\cdots n_j!$ and $s(n)=\Sigma_{r=1}^{j}rn_r$ for each $n=(n_1,\ldots,n_j)\in\mathbf{Z}_+^j$, and where, as usual, $(\mathrm{ad}\,S)^0A=A,(\mathrm{ad}\,S)^kA=(\mathrm{ad}\,S)^{k-1}(\mathrm{ad}\,SA)(k\in\mathbf{N},A\in\mathscr{L}(\mathscr{H}))$.

To prove the above claim, we begin by operating on the truncated normal form $K^{[N]}(\varepsilon)$ with the truncated superoperator $R_N^{[N]}(\varepsilon)$ in a purely formal way, thus obtaining

$$R_N^{[N]}(\varepsilon)K^{[N]}(\varepsilon)=\sum_{i=0}^{N}\varepsilon^i\sum_{j=0}^{i}R_N^{(j)}K^{(i-j)}+\varepsilon^{N+1}\sum_{i=0}^{N-1}\varepsilon^i\sum_{j=i+1}^{N}R_N^{(j)}K^{(i-j+N+1)}, \tag{2.18}$$

and use (2.18) and (2.14) to express the rhs of (2.13) as a polynomial of order $2N$ in ε. We then equate coefficients of ε^i in both sides of (2.13) for $0\leqslant i\leqslant 2N$. If $N=1$, our previous calculations for $i=0$, 1, when supplemented by a trivial calculation for $i=2$ show that the claim is true. Henceforth, until further notice, we will assume that $N\geqslant 2$. In this case, it is only necessary to equate coefficients of ε^i in the subcases $2\leqslant i\leqslant N$ and $N+1\leqslant i\leqslant 2N$, since the results of doing this for $i=0$, 1 merely reproduce those obtained in the case $N=1$. Equating coefficients of ε^i in the first subcase yields the condition

$$\sum_{j=0}^{i}R_N^{(j)}K^{(i-j)}=0 \quad (2\leqslant i\leqslant N). \tag{2.19}$$

To show that (2.19) is satisfied by the $S^{(i)}$ and K_i specified in the preceding paragraph, we note that (2.8a), (2.15a), (2.16b), (2.17a), and the relations

$$R_i^{(j)}A=R_k^{(J)}A \quad (1\leqslant j\leqslant i\leqslant k,A\in\mathscr{L}(\mathscr{H})), \tag{2.20a}$$

$$R_i^{(i)}A=-\,\mathrm{ad}\,S^{(i)}A+R_{i-1}^{(i)}A \quad (i\geqslant 2,\ A\in\mathscr{L}(\mathscr{H})) \tag{2.20b}$$

(following directly from (2.17b)), entail that definition (2.16b) can be written in the form (2.19) for arbitrary $N\geqslant 2$.

Finally, it is trivial to verify that the expression for $W_{N+1}^{(j)}$ obtained by equating coefficients of ε^i for $N+1\leqslant i\leqslant 2N$ $(N\geqslant 2)$, agrees with definition (2.15c). Since the conditions on the coefficients of ε^i have been shown to be satisfied for all $N\geqslant 1$ by the $S^{(i)},K^{(i)},W_i^{(j)}$ specified in the penultimate paragraph, the proof of the claim is complete.

This concludes the formal construction of the relevant operators $S^{(i)}\in X$, $K^{(i)}\in Y$, $W_{N+1}^{(j)}$ satisfying (2.13) for arbitrary $N\geqslant 1$, an inequality which we suppose to hold henceforth. We note that the matrices $(K_{mn}^{(i)})$ $(i\geqslant 0)$, and therefore $(K_{mn}^{[N]}(\varepsilon))$, representing $K^{(i)}$ and $K^{[N]}(\varepsilon)$ in the Φ-basis, are Hermitian. The hermiticity property follows from the corresponding property possessed by the matrices representing H_0 and V, the relevant definitions, and a simple inductive argument, which will be stated precisely in Sec. IV. (The fact that these matrices are real in the Φ-basis is immaterial to our purposes.)

To illustrate these general considerations, we record the following formulas for the matrix elements of $K^{(1)},K^{(2)}$ in this basis, obtained from (2.15b), (2.16), and (2.17):

$$K_{mn}^{(1)}=\delta_{\lambda_m,\lambda_n}V_{mn},$$

$$K_{mn}^{(2)}=\delta_{\lambda_m,\lambda_n}\sum_{r\lambda_r\neq\lambda_m}\frac{V_{mr}V_{rn}}{\lambda_m-\lambda_r}.$$

Using these formulas, one can show that for $k=1,2$ the eigenvalues of $K^{[k]}(\varepsilon)$ coincide to within $O(\varepsilon^k)$ with those of H_ε formally given by kth order degenerate Rayleigh–Schrödinger perturbation theory. We conjecture that this is also the case for general $k\geqslant 1$, but will not pursue the matter

any further here, since it would involve us in questions which are not strictly relevant to our present goals. Nevertheless, it would be interesting to pursue them in a separate investigation.

The Nth-order approximant to the exact solution $\psi(t,\varepsilon)$ of the IVP (1.1) is defined by[1]

$$\psi_N(t,\varepsilon)=R_N^{[N]}(\varepsilon)(U_N(t,\varepsilon))\psi_0, \tag{2.21}$$

where the parenthesis around $U_N(t,\varepsilon)$ emphasizes that $R_N^{[N]}(\varepsilon)$ acts on it and where

$$U_N(t,\varepsilon)=\exp(-itK^{[N]}(\varepsilon)). \tag{2.22}$$

Since we require that $U_N(t,\varepsilon)$ be a well-defined unitary operator and want to avoid mathematical subtleties with which we cannot cope in this section, we shall assume for the moment that $K^{[N]}(\varepsilon)$ is self-adjoint. In Secs. III and IV we shall replace $K^{[N]}(\varepsilon)$ by its self-adjoint extension in (2.22), thus avoiding this assumption.

More explicitly, (2.21), (2.22), and (2.17) yield for $N=1,2,3$:[1]

$$\psi_1(t,\varepsilon)=(I-\varepsilon\tau_1)U_1(t,\varepsilon),$$

$$\psi_2(t,\varepsilon)=\{I-\varepsilon\tau_1+\varepsilon^2[\tfrac{1}{2}\tau_1^2-\tau_2]\}U_2(t,\varepsilon),$$

$$\psi_3(t,\varepsilon)=\{I-\varepsilon\tau_1+\varepsilon^2[\tfrac{1}{2}\tau_1^2-\tau_2]+\varepsilon^3[-\tfrac{1}{6}\tau_1^3+\tau_1\tau_2-\tau_3]\}U_3(t,\varepsilon),$$

where I is the unit superoperator on $\mathscr{L}(\mathscr{H})$ and $\tau_i=\mathrm{ad}\,S^{(i)}$. It is worth remarking that $\psi_N(t,\varepsilon)$ could be calculated, in a practical numerical sense, for sufficiently small ν and N and ''simple'' enough initial states ψ_0 (e.g., finite linear combinations of Φ_k's). In particular, note that for any given $m\in\mathbf{Z}_+^\nu$, the matrix elements $\langle\Phi_m,U_N(t,\varepsilon)\Phi_n\rangle$ are nonvanishing for at most a finite number of n-values, namely, those with $\lambda_m\neq\lambda_n$. Hence the calculation of $U_N(t,\varepsilon)\Phi_k$ for any $k\in\mathbf{Z}_+^\nu$ can be reduced to an elementary diagonalization procedure in a finite-dimensional space. This is also the case for the more satisfactorily defined operator $U_N(t,\varepsilon)$ considered in Secs. III and IV.

Our next goal is to show that $\psi_N(t,\varepsilon)$ satisfies a differential equation which only differs from (1.1a) by terms of order ε^{N+1}. We find by formally differentiating (2.21) and then proceeding as described below

$$\begin{aligned} i\frac{d\psi_N(t,\varepsilon)}{dt}&=R_N^{[N]}(\varepsilon)(K^{[N]}(\varepsilon)U_N(t,\varepsilon))\psi_0\\ &=\sum_{i=0}^{N}\varepsilon^i R_N^{(i)}(K^{[N]}(\varepsilon)U_N(t,\varepsilon))\psi_0\\ &=\sum_{i=0}^{N}\varepsilon^i\sum_{j=0}^{i}R_N^{(j)}K^{[N]}(\varepsilon)\cdot R_N^{(i-j)}(U_N(t,\varepsilon))\psi_0\\ &=R_N^{[N]}(\varepsilon)K^{[N]}(\varepsilon)\cdot R_N^{[N]}(\varepsilon)(U_N(t,\varepsilon))\psi_0+\varepsilon^{N+1}\rho_N(t,\varepsilon), \end{aligned} \tag{2.23}$$

where

$$\rho_N(t,\varepsilon)=-\sum_{i=0}^{N-1}\varepsilon^i\sum_{j=i+1}^{N}R_N^{(j)}K^{[N]}(\varepsilon)\cdot R_N^{(i-j+N+1)}(U_N(t,\varepsilon))\psi_0 \tag{2.24}$$

and where we have carried out obvious formal manipulations, analogous to those employed to arrive at (2.18), and have invoked the relation[21]

$$R_j^{(i)}(AB)=\sum_{k=0}^{i}R_j^{(k)}A\cdot R_j^{(i-k)}B. \tag{2.25}$$

By (2.23), (2.21), and (2.13), we see that $\psi_N(t,\varepsilon)$ satisfies the differential equation

$$i\,\frac{d\psi_N(t,\varepsilon)}{dt}=H_\varepsilon\psi_N(t,\varepsilon)+\varepsilon^{N+1}\varphi_N(t,\varepsilon), \tag{2.26a}$$

where

$$\varphi_N(t,\varepsilon)=-W_{N+1}(\varepsilon)\psi_N(t,\varepsilon)+\rho_N(t,\varepsilon). \tag{2.27}$$

Furthermore, by (2.21), (2.22), and (2.17),

$$\psi_N(0,\varepsilon)=\psi_0. \tag{2.26b}$$

Hence $\psi_N(t,\varepsilon)$ satisfies the same Schrödinger evolution equation (1.1a) as $\psi(t,\varepsilon)$ to within terms of $O(\varepsilon^N)$ and fulfills exactly the same initial condition (1.1b) as $\psi(t,\varepsilon)$. The last fact plays an essential role in the heuristic derivation of the error estimate (1.4) in the next paragraph, as well as in its rigorous counterpart in Sec. IV.

Integrating (2.26a) by the method of variation of constants and using (2.26b), we obtain

$$\psi_N(t,\varepsilon)=\exp(-itH_\varepsilon)\psi_0-i\varepsilon^{N+1}\int_0^t\exp[-i(t-s)H_\varepsilon]\varphi_N(s,\varepsilon)ds. \tag{2.28}$$

By (2.28), $\psi(t,\varepsilon)=\exp(-itH_\varepsilon)\psi_0$, the unitarity of $\exp(-itH_\varepsilon)$, and the boundedness of $\|\varphi_N(t,\varepsilon)\|$ for $t\in\mathbf{R}$, $0\leqslant\varepsilon\leqslant\varepsilon_0<\infty$ (see (4.19)), implied by our conditions on V and the frequencies ω_i which are stated in Sec. III, we infer that

$$\|\psi(t,\varepsilon)-\psi_N(t,\varepsilon)\|\leqslant\varepsilon^{N+1}\left|\int_0^t\|\varphi_N(s,\varepsilon)\|ds\right|\leqslant\text{ const }\varepsilon^{N+1}|t| \tag{2.29}$$

at each of the latter t,ε-values. We emphasize that these conditions are satisfied by a very wide class of resonant and nonresonant oscillators.

III. ASSUMPTIONS, DEFINITIONS, AND STATEMENT OF MAIN RESULT

As a first step in making precise the heuristic discussions in Sec. II, we state our assumptions on V and the ω_i. Fixing $\nu\geqslant 2$, we will assume henceforth that V possesses the following properties:

(I) V is a maximal operator of multiplication in $\mathscr{H}=L^2(\mathbf{R}^\nu)$ by a real-valued function in $C^\infty(\mathbf{R}^\nu)$ which will also be denoted by V.

(II) For each $r=(r_1,\ldots,r_\nu)\in\mathbf{Z}_+^\nu$, there exist non-negative constants Γ_r,μ_r such that

$$\left|\frac{\partial^{|r|}V(x)}{\partial x_1^{r_1}\cdots\partial x_\nu^{r_\nu}}\right|\leqslant\Gamma_r(|x|+1)^{\mu},$$

at each $x=(x_1,\ldots,x_\nu)\in\mathbf{R}^\nu$, where $|r|=r_1+\cdots+r_\nu$ and $|x|=\Sigma_{i=1}^\nu|x_i|$.

(III) The differential operator

$$\frac{1}{2}\sum_{i=1}^{\nu}\left(-\frac{\partial^2}{\partial x_i^2}+\omega_i^2x_i^2\right)+\varepsilon V(x_1,\ldots,x_\nu)$$

on $C_0^\infty(\mathbf{R}^\nu)$ is essentially self-adjoint in $\mathscr{H}$ for each $\varepsilon\geqslant 0$, its unique self-adjoint extension being denoted by H_ε at each such ε. The frequencies ω_i are fixed positive numbers.

Remark: Property (III) is entailed by (I) and (II) if $\frac{1}{2}\Sigma_{i=1}^\nu\omega_i^2x_i^2+\varepsilon V(x)$ is bounded below on $\mathbf{R}^\nu$ for each $\varepsilon\geqslant 0$.[22]

Letting ρ be the maximum number of rationally independent frequencies ω_i, we will number then as $\omega_1,\ldots,\omega_\rho$ w.l.g. Writing $|j|_\rho=\Sigma_{i=1}^\rho|j_i|$ for the norm of vectors $j\in\mathbf{R}^\rho$, we suppose that the following diophantine condition is satisfied:

(IV) If $\rho\geqslant 2$, there exist positive constants γ, η such that for all $0\neq k=(k_1,\ldots,k_\rho)\in\mathbf{Z}^\rho$,

$$\left|\sum_{i=1}^{\rho} \omega_i k_i\right| \geqslant \gamma |k|_\rho^{-\eta}, \tag{3.1}$$

where $\mathbf{Z}^\rho$ is the set of all vectors $(j_1,\ldots,j_\rho)$ with integer components.

Remarks: (1) Condition (IV) is imposed to avoid convergence difficulties in sums involving differences $\lambda_m-\lambda_n$ in denominators. These difficulties are analogous to ones occurring in classical canonical perturbation theory.[23] One only needs to assume that (IV) is satisfied for $\rho\geqslant 2$. Indeed, if $\rho=1$, the ω_i are multiples of some $\omega_0>0$, and thus $|\lambda_m-\lambda_n|$ is bounded away from zero for each fixed $m\in\mathbf{Z}_+^\nu$ as n ranges over this set. Hence (3.1) is automatically satisfied in this case.

(2) It is well known that if $\rho\geqslant 2$ and $\eta\geqslant\nu$, then the set of vectors $(\omega_1,\ldots,\omega_\rho)$ with positive components for which there exists no $\gamma>0$ such that the condition (3.1) holds for all $0\neq k\in\mathbf{Z}^\rho$ has Lebesgue measure zero.[24]

In order to state our main result rigorously, we need to define the Nth-order approximant $\psi_N(t,\varepsilon)$ more precisely than in Sec. II. The following definitions and remarks are made for this purpose, as well as to prepare the way for the developments in Sec. IV.

Definition 1: If A is an operator on $\mathscr{H}$, then we write[25]

$$C^\infty(A)=\bigcap_{j=1}^{\infty} D(A^j).$$

Remarks: (1) If $A\in\mathscr{L}(\mathscr{H})$ is self-adjoint, then $C^\infty(A)$ is dense in $\mathscr{H}$.[26] In particular, $C^\infty(H_0)$ is dense in $\mathscr{H}$ and contains the Schwartz space $S(\mathbf{R}^\nu)$ of functions of fast decrease.

(2) $C^\infty(H_0)$ is invariant under H_0, V, and the other operators of the theory.

Our last assumption is motivated by these properties of $C^\infty(H_0)$, as well as by the desire to avoid unessential complications which would ensue by admitting ε-dependent initial states.

(V) The initial state $\psi_0=\psi(0,\varepsilon)$ has the following properties:

$$\psi_0=\psi(0,\varepsilon)\in C^\infty(H_0), \tag{3.2a}$$

$$\psi_0 \text{ is } \varepsilon\text{-independent.} \tag{3.2b}$$

Remark: Conditions (III) and (V) entail that $\psi_0\in D(H_\varepsilon)$ holds for each $\varepsilon\geqslant 0$, and thus that the IVP (1.1) has the unique solution $\psi(t,\varepsilon)=\exp(-itH_\varepsilon)\psi_0$ for all $t\in\mathbf{R}$, $\varepsilon\geqslant 0$. Indeed, at all such ε, $(H_0+\varepsilon V)|C^\infty(H_0)$ is a symmetric extension of $(H_0+\varepsilon V)|C_0^\infty(\mathbf{R}^\nu)$ having H_ε as its unique self-adjoint extension (by (III) and $C_0^\infty(\mathbf{R}^\nu)\subset C^\infty(H_0)$).[25] Thus under these conditions, (1.2) holds on $C^\infty(H_0)$ and $\psi_0\in C^\infty(H_0)\subset D(H_\varepsilon)$.

The next definition makes precise the sense in which the infinite sums occurring in the sequel are to be understood. Henceforth, it is convenient to set $\Omega=\mathbf{Z}_+^\nu$.

Definition 2: Let $\{\Delta_k\}(k=(k_1,k_2,\ldots,k_\nu)\in\Omega)$ be a sequence of subsets $\Delta_k=[0,k_1]\times[0,k_2]\times\cdots\times[0,k_\nu]$. A sum of the form $\Sigma_{m\in\Omega}f_m$, where each $f_m\in\mathbf{C}$, is defined as $\lim_{k_1,\ldots,k_\nu\to\infty}\Sigma_{m\in\Delta_k}f_m$ if this limit exists in $\mathbf{C}$. Analogously, a sum of the form $\Sigma_{m\in\Omega}g_m\Phi_m$, where each $g_m\in\mathbf{C}$, is defined as the strong limit $\mathrm{s}-\lim_{k_1,\ldots,k_\nu\to\infty}\Sigma_{m\in\Delta_k}g_m\Phi_m$ in $\mathscr{H}$ if this limit exists.

Given matrices $\mathscr{A}=(A_{mn})$, $\mathscr{B}=(B_{mn})(m,n\in\Omega)$, we naturally define their product $\mathscr{A}\mathscr{B}$ as the matrix $(\Sigma_{r\in\Omega}A_{mr}B_{rn})$ provided that this sum exists in the sense of Definition 2 for each of the latter m,n.[27]

Moreover, we will denote the adjoint of an operator A by A^*, the hermitian adjoint of a matrix $\mathscr{A}$ by $\mathscr{A}^*$, and the complex conjugate of $z\in\mathbf{C}$ by $\overline{z}$.

Definition 3: Denote the inner product in $\mathscr{H}$ by $\langle .,. \rangle$, assuming that it is antilinear in the first argument and linear in the second. Let $\mathscr{A}=(A_{mn})$ $(m,n\in\Omega)$ be a matrix of complex numbers which is square-summable by rows and columns, i.e., $\Sigma_{m\in\Omega}|A_{mn}|^2<\infty$ $(n\in\Omega)$ and $\Sigma_{n\in\Omega}|A_{mn}|^2<\infty$ $(m\in\Omega)$, respectively. An operator $A\in\mathscr{L}(\mathscr{H})$ is said to be *associated* with $\mathscr{A}$, or $A\leftrightarrow\mathscr{A}$ in symbols, if

$$D(A)=\left\{ f\in\mathscr{H}\colon \sum_{m\in\Omega}\left|\sum_{n\in\Omega} A_{mn}\langle\Phi_n,f\rangle\right|^2<\infty\right\}, \tag{3.3a}$$

and

$$Af=\sum_{m\in\Omega}\left(\sum_{n\in\Omega} A_{mn}\langle\Phi_n,f\rangle\right)\Phi_m \quad (f\in D(A)). \tag{3.3b}$$

Remarks: (1) Stone[28] has considered linear operators associated with infinite matrices in a more general sense than that of Definition 3. In general, infinite matrices have various pathological properties which make them unsuitable for representing unbounded linear operators.[29] In particular, they may not be associative under matrix multiplication.[30] However, the infinite matrices considered in this paper are well enough behaved to be useful tools for establishing its conclusions.

(2) The sums over n in (3.3a) and (3.3b) converge absolutely, because of the square-summability of $\mathscr{A}$ by columns, Schwarz's inequality in $\ell^2(\Omega)$, and Parseval's identity.

(3) The relation $A\leftrightarrow\mathscr{A}$ between the set J of operators A specified by Definition 3 and the set $\mathscr{J}$ of matrices $\mathscr{A}=(A_{mn})$ with the properties stated in this definition is a bijective map from J onto $\mathscr{J}$. Indeed, by the square summability of each $\mathscr{A}\in\mathscr{J}$ by rows and the orthonormality of the Φ_k, one sees that the entries of each such matrix $\mathscr{A}$ are uniquely determined by the operator $A\leftrightarrow\mathscr{A}$, namely, $A_{mn}=\langle\Phi_m,A\Phi_n\rangle$.

We proceed to define the matrices and operators of the theory. In contrast with the heuristic ''definitions'' given in Sec. II, most of the operators of interest will now be defined as those associated with infinite matrices of the type specified next.

Definition 4: We denote by $\mathscr{M}$ the set of matrices $\mathscr{A}=(A_{mn})$ $(m,n\in\Omega)$ over $\mathbf{C}$ such that the inequalities

$$|A_{mn}|\leqslant \text{const}\ \lambda_m^{a(p)}\lambda_n^{-p}, \tag{3.4a}$$

$$|A_{mn}|\leqslant \text{const}\ \lambda_n^{a(p)}\lambda_m^{-p} \tag{3.4b}$$

hold for all $m,n\in\Omega$ and all $p\in\mathbf{N}$, where const and the $a(p)$ are real constants independent of m,n. The $a(p)$ are called *exponents* of $\mathscr{A}$. The set whose members are those operators in $\mathscr{L}(\mathscr{H})$ associated with matrices in $\mathscr{M}$ is denoted by M.[31] That composed of operators in M restricted to $C^\infty(H_0)$ is called M_0.

Remark: By Lemma 3(2) (Sec. IV),[32] all matrices in $\mathscr{M}$ are square-summable by rows and columns. Hence the definition of $\underline{M}$ makes sense (recall Definition 3).

Analogously to the notation $\tilde{A},\hat{A}$ for operators used in Sec. II, for any matrix $\mathscr{B}=(B_{mn})$ we will denote by $\tilde{\mathscr{B}}$, $\hat{\mathscr{B}}$ the matrices with entries $(1-\delta_{\lambda_m,\lambda_n})B_{mn}$, $\delta_{\lambda_m,\lambda_n}B_{mn}$, respectively. Moreover, for each matrix $\mathscr{B}\in\mathscr{M}$ we denote by ad $\mathscr{B}$ the linear operator from $\mathscr{M}$ to $\mathscr{M}$ acting by ad $\mathscr{B}\mathscr{C}=\mathscr{B}\mathscr{C}-\mathscr{C}\mathscr{B}$. This last definition is legitimate because $\mathscr{M}$ is an associative matrix algebra over $\mathbf{C}$ when equipped with the usual definition of matrix identity, with the usual operations of multiplication by scalars and matrix summation, and with matrix multiplication as defined after Definition 2 (see Lemma 3(1)).

The key matrices and operators acting on matrices specified by Definitions 5(a) and 5(b) below have the same ''structure'' as do the analogous expressions in terms of the operators in $\mathscr{L}(\mathscr{H})$ and the operators acting on them which were introduced formally in Sec. II. Indeed, these definitions anticipate the isomorphism existing between $\mathscr{M}$ and M_0 when viewed as associative algebras, which is stated in Lemma 5.

Definition 5(a): The infinite matrices $\mathscr{S}^{(i)}=(S_{mn}^{(i)})$, $\mathscr{V}_i=((V_i)_{mn})$, $\mathscr{K}^{(i)}=(K_{mn}^{(i)})$ $(i\geqslant 1)$ and $\mathscr{W}_{i+1}^{(j)}=((\mathscr{W}_{i+1}^{(j)})_{mn})$ $(0\leqslant j\leqslant i-1,\ i\geqslant 1)$ are defined inductively as follows:

(1) $\mathscr{S}^{(i)},\mathscr{K}^{(i)}$ are the unique matrices with $(\mathscr{S}^{(i)})\hat{}=0$, $(\mathscr{K}^{(i)})\tilde{}=0$ which satisfy

$$\text{ad}\,\mathscr{K}^{(0)}\mathscr{S}^{(i)}=\mathscr{V}_i-\mathscr{K}^{(i)} \quad (i\geqslant 1), \tag{3.5}$$

or, equivalently,

$$\text{ad}\,\mathscr{K}^{(0)}\mathscr{S}^{(i)}=(\mathscr{V}_i)^{\sim}\quad(i\geqslant 1),\tag{3.6a}$$

$$\mathscr{K}^{(i)}=(\mathscr{V}_i)^{\wedge}\quad(i\geqslant 1).\tag{3.6b}$$

Here, $\mathscr{K}^{(0)}$ is the diagonal matrix with which H_0 is associated:

$$\mathscr{K}^{(0)}=(\lambda_m\delta_{m,n}).\tag{3.7}$$

(2) One defines

$$\mathscr{V}_1=(\langle\Phi_m,V\Phi_n\rangle)=(V_{mn}),\tag{3.8a}$$

$$\mathscr{V}_i=-\sum_{j=1}^{i}\mathscr{R}_{i-1}^{(j)}\mathscr{K}^{(i-j)}\quad(i\geqslant 2),\tag{3.8b}$$

$$\mathscr{W}_{i+1}^{(j)}=-\sum_{k=j+1}^{i}\mathscr{R}_i^{(k)}\mathscr{K}^{(i+j-k+1)}\quad(0\geqslant j\geqslant i-1,i\geqslant 1),\tag{3.8c}$$

where $\mathscr{R}_j^{(k)}$ $(j\geqslant 1,\ k\geqslant 0)$ denotes a linear operator from $\mathscr{M}$ into $\mathscr{M}$ acting by

$$\mathscr{R}_j^{(0)}\mathscr{A}=\mathscr{A}\quad(j\geqslant 1,\mathscr{A}\in\mathscr{M}),\tag{3.9a}$$

$$\mathscr{R}_j^{(k)}\mathscr{A}=\sum_{n\in\mathbf{Z}_+^j,\,s(n)=k}\frac{(-1)^{|n|}}{n!}(\text{ad}\,\mathscr{S}^{(1)})^{n_1}\cdots(\text{ad}\,\mathscr{S}^{(j)})^{n_j}\mathscr{A},$$

$$(j\geqslant 1,k\geqslant 1,\mathscr{A}\in\mathscr{M}).\tag{3.9b}$$

Remarks: (1) Note that each term in the sums in (3.8b) and (3.8c) is a linear combination of finite products of matrices in $\mathscr{M}$ [see (3.9b) and the proof of Lemma 6(1)], and is thus well defined since $\mathscr{M}$ is an associative algebra over **C** under the stated operations.

(2) Obviously, Definition 5(a) implies that for each $i\geqslant 1$,

$$S_{mn}^{(i)}=\begin{cases}(V_i)_{mn}/(\lambda_m-\lambda_n) & (\lambda_m\neq\lambda_n),\\ 0 & (\lambda_m=\lambda_n),\end{cases}\tag{3.10}$$

Definition 5(b): (1) $S^{(i)},V_i,K^{(i)}$ $(i\geqslant 1)$ and $W_{i+1}^{(j)}$ $(0\leqslant j\leqslant i-1,\ i\geqslant 1)$ are the operators in $\mathscr{L}(\mathscr{H})$ which are associated with the respective matrices $\mathscr{S}^{(i)},\hat{\mathscr{V}}_i,\mathscr{K}^{(i)},\mathscr{W}_{i+1}^{(j)}$. Moreover, we set $K^{(0)}=H_0$.

(2) For each $i\geqslant 1$, $\varepsilon\geqslant 0$, the operator $K^{[i]}(\varepsilon)$ is defined as the operator sum

$$K^{[i]}(\varepsilon)=\sum_{j=0}^{i}\varepsilon^jK^{(j)},\tag{3.11}$$

while the operators $K_i(\varepsilon)$ and $W_{i+1}(\varepsilon)$ are those associated with the respective matrices

$$\mathscr{K}^{[i]}(\varepsilon)=\sum_{j=0}^{i}\varepsilon^j\mathscr{K}^{(j)},\tag{3.12}$$

$$\mathscr{W}_{i+1}(\varepsilon)=\sum_{j=0}^{i-1}\varepsilon^j\mathscr{W}_{i+1}^{(j)}.\tag{3.13}$$

Remarks: (1) Part (1) of this definition makes sense because the matrices mentioned therein are in $\mathscr{M}$, thus being square-summable by rows and columns.

(2) By Lemma 7, $K_i(\varepsilon)$ in Part (2) of Definition 5(b) is the unique self-adjoint extension of $K^{[i]}(\varepsilon)$ in (3.11) for $i\geqslant 1$, $\varepsilon\geqslant 0$.

Definition 5(c): For each $i,N\geqslant 1$, $R_i^{[N]}(\varepsilon)$ denotes the superoperator acting by $R_i^{[N]}A=\Sigma_{j=0}^{N}\varepsilon^j R_i^{(j)}A$ on each $A\in\mathscr{L}(\mathscr{H})$, where the $R_i^{(j)}$ are defined by (2.17), the operators $S^{(k)}$ in (2.17b) being those specified by Definition 5(b).

Definition 5(d): The Nth-order approximant to the solution of the IVP (1.1) is defined by

$$\psi_N(t,\varepsilon)=R_N^{[N]}(\varepsilon)(U_N(t,\varepsilon))\psi_0 \tag{3.14}$$

for all $N\geqslant 1$, $t\in\mathbf{R}$, $0\leqslant\varepsilon\leqslant\varepsilon_0$. Here,

$$U_N(t,\varepsilon)=\exp(-itK_N(\varepsilon)), \tag{3.15}$$

$K_N(\varepsilon)$ being the self-adjoint operator in Part (2) of Definition 5(b) and ψ_0 satisfying conditions (3.2).

Remark: Definition (3.14) makes sense by properties (a) and (b) in the proof of Lemma 8. Note that in (3.15) the self-adjoint operator $K_N(\varepsilon)$ replaces the operator $K^{[N]}(\varepsilon)$ in (2.22), which was loosely defined in Sec. II and is now precisely specified by Part (2) of Definition 5(b). This replacement has been made because $K^{[N]}(\varepsilon)$ is not generally self-adjoint, and therefore $\exp(-itK^{[N]}(\varepsilon))$ is not generally well-defined. If it were not, a breakdown of our analysis would occur.

Our principal result is as follows.

Theorem 1: Let H_0 and V obey conditions (I)–(III), let (IV) hold if $\rho\geqslant 2$, and let the initial state ψ_0 satisfy (V). Then for each $N\geqslant 1$, $0\leqslant\varepsilon_0<\infty$ we have the uniform error estimate

$$\|\psi(t,\varepsilon)-\psi_N(t,\varepsilon)\|\leqslant C_N\varepsilon^{N+1}|t| \tag{3.16}$$

if $t\in\mathbf{R}$, $0\leqslant\varepsilon\leqslant\varepsilon_0$, where $\|.\|$ is the $\mathscr{H}$-norm and $C_N=C_N(\psi_0,\varepsilon_0)$ a positive constant.

Remarks: (1) This theorem is the first rigorous error estimate for the present normal-form method of this degree of generality, applying to cases in which V is not necessarily a polynomial and/or H_0 has degenerate eigenvalues.

(2) Under certain nonresonance conditions, an error estimate of type (3.16) was obtained in Refs. 1 and 2 for the case of a system of simple harmonic oscillators coupled by a potential which was a polynomial in the coordinates and momenta.

(3) From Theorem 1, one immediately infers that for any $\delta\in(0,N+1]$ and any pair T,ε_0 of positive numbers, the uniform error estimate

$$\|\psi(t,\varepsilon)-\psi_N(t,\varepsilon)\|\leqslant C'_{N,\delta,T}\varepsilon^{N+1-\delta}$$

holds for all $|t|\leqslant T\varepsilon^{-\delta}$, $\varepsilon\in[0,\varepsilon_0]$, where $C'_{N,\delta,T}=C'_{N,\delta,T}(\psi_0,\varepsilon_0)$ is a positive constant.

IV. AUXILIARY RESULTS AND PROOF OF THEOREM 1

The proof of Theorem 1 is based on Lemmas 1–9 below. Once Lemma 9 has been proved, the theorem will follow readily.

Assertions (1) and (2) of Lemma 1 are immediate consequences of Lemma 1(1) and 1(2) of Ref. 14, respectively. The proof of the latter lemma involves the use of Condition (IV) in an essential way.

Lemma 1: (1) If $m,n\in\mathbf{Z}_+^\nu$ and $\lambda_m\neq\lambda_n$, the inequality

$$|\lambda_m-\lambda_n|\geqslant\text{const }\lambda_m^{-\eta}\lambda_n^{-\eta} \tag{4.1a}$$

holds, where η is the positive constant in Condition (IV) and const is independent of m,n.

(2) For each $m\in\mathbf{Z}_+^\nu$, the estimate

$$g_m\leqslant\text{const }\lambda_m^{\nu-\rho} \tag{4.1b}$$

holds for the degeneracy g_m of each eigenvalue λ_m, where const is independent of m.[33]

The two results asserted in the next lemma are contained in the respective Lemmas 2 and 5 of Ref. 14, and are proved in detail there. The first of these results is indispensable in this section because it constitutes the initial step of the inductive process by which we will prove that the matrices entering into the present theory belong to the set $\mathscr{M}$ (Definition 4).

Lemma 2: The matrix $\mathscr{V}_1=(V_{mn})=(\langle\Phi_m,V\Phi_n\rangle)$ is in $\mathscr{M}$ and the operator $V_1\leftrightarrow\mathscr{V}_1$ has the property[25]

$$V_1=V \quad \text{on} \quad C^\infty(H_0). \tag{4.2}$$

The next three lemmas are basic building blocks for subsequent developments in this section.

Lemma 3: The set of matrices $\mathscr{M}$ has the following properties:

(1) Under the standard definition of matrix equality and when equipped with the usual operations of multiplication by complex numbers and matrix summation, and with matrix multiplication as specified after Definition 2, $\mathscr{M}$ is an associative algebra over **C**.

(2) Each matrix in $\mathscr{M}$ is square-summable by rows and columns.

Remark: (1) Once it has been proved that $\mathscr{M}$ is closed under such matrix multiplication, the associativity of this operation follows immediately from the facts that multiplication of operators belonging to $\mathscr{L}(\mathscr{H})$ is associative and that there is a bijection $\varphi:\mathscr{M}\to M_0$ having the property $\varphi(\mathscr{A}\mathscr{B})=\varphi(\mathscr{A})\varphi(\mathscr{B})$ for all $\mathscr{A},\mathscr{B}\in\mathscr{M}$ (see the proof of Lemma 5). The associativity of this matrix multiplication is established directly in the proof of the present lemma.

(2) As we have already remarked in Sec. III, assertion (2) of this lemma entails that the definition of M as the set of operators in $\mathscr{H}$ associated with matrices in $\mathscr{M}$ makes sense.

Proof: (1) We will prove that $\mathscr{M}$ has the following three properties:

$$\text{(a)}\quad \mathscr{A},\mathscr{B}\in\mathscr{M}\Rightarrow\mathscr{A}+\mathscr{B}\in\mathscr{M};$$

$$\text{(b)}\quad \mathscr{A},\mathscr{B}\in\mathscr{M}\Rightarrow\mathscr{A}\mathscr{B}\in\mathscr{M};$$

$$\text{(c)}\quad \mathscr{A},\mathscr{B},\mathscr{C}\in\mathscr{M}\Rightarrow\mathscr{A}(\mathscr{B}\mathscr{C})=(\mathscr{A}\mathscr{B})\mathscr{C}.$$

Given these properties, it is a trivial matter to show that assertion (1) of the lemma holds.

Let $\mathscr{A}=(A_{mn})$, $\mathscr{B}=(B_{mn})$, $\mathscr{C}=(C_{mn})$ be in $\mathscr{M}$. To prove (a), note that

$$|A_{mn}+B_{mn}|\leqslant|A_{mn}|+|B_{mn}|\leqslant\text{const}\ (\lambda_m^{a(p)}\lambda_n^{-p}+\lambda_m^{b(p)}\lambda_n^{-p})\leqslant\text{const}\ \lambda_m^{\alpha(p)}\lambda_n^{-p} \tag{4.3}$$

for all $p\in\mathbf{N}$, where const is independent of m,n; $a(p),b(p)$ are exponents of $\mathscr{A}$, $\mathscr{B}$, respectively; and $\alpha(p)=\max\{a(p),b(p)\}$. Here, we have invoked the fact each A_{mn} and B_{mn} satisfies an inequality of the form (3.4a). Since they also obey inequalities of the form (3.4b), (4.3) obtains with m,n interchanged on the rhs.[34] Hence (a) follows.

As to (b), note that[35]

$$\left|\sum_r A_{mr}B_{rn}\right|\leqslant\sum_r|A_{mr}|.|B_{rn}|\leqslant\text{const}\ \lambda_m^{a(q)}\lambda_n^{-p}\sum_r\lambda_r^{b(p)}\lambda_r^{-q}\leqslant\text{const}\ \lambda_m^{a(q)}\lambda_n^{-p} \tag{4.4}$$

for all $p\in\mathbf{N}$ if one chooses a positive integer $q=q(p)>b(p)+\nu$. Here, we have used (3.4a), (3.4b), and the estimate[36]

$$\sum_{r\in\mathbf{Z}_+^\nu}\lambda_r^{-\mu}<\infty \quad (\mu>\nu). \tag{4.5}$$

Similar arguments show that (4.4) is true with m,n interchanged in the expression after the second inequality, and hence we have proved (b).

Statement (c) follows from the absolute convergence of $\Sigma_{r,s}A_{mr}B_{rs}C_{sn}$. To prove this property, one uses (3.4a) and (4.5), thus obtaining

$$\sum_{r,s}|A_{mr}|.|B_{rs}|.|C_{sn}|\leqslant\text{const}\ \lambda_m^{a(p)}\lambda_n^{-t}\left(\sum_r\lambda_r^{-p+b(q)}\right)\left(\sum_s\lambda_s^{-q+c(t)}\right)<\infty$$

for each $t\in\mathbf{N}$ if for every such t one first chooses a positive integer $q=q(t)>c(t)+\nu$ and then a positive integer $p=p(q(t))>b(q(t))+\nu$. Of course, here $c(t)$ denotes an exponent of $\mathscr{C}$.

(2) Let $\mathscr{A}$ be as in part (1) of the proof. Its square-summability by rows follows from

$$\sum_m |A_{mn}|^2 \leqslant \text{const}\ \lambda_n^{2a(p)} \sum_m \lambda_m^{-2p} < \infty \quad (p > \nu/2),$$

where $a(p)$ denotes an exponent of $\mathscr{A}$ and where we have once more used (3.4a) and (4.5). The square-summability of $\mathscr{A}$ by columns can be proved similarly by invoking (3.4b) and (4.5). □

Lemma 4: If $A \in M$, then A and A^* have domains containing the set $C^\infty(H_0)$ and leave it invariant.

Proof: (1) An operator $A \in \mathscr{L}(\mathscr{H})$ associated with a matrix (A_{mn}) of the type specified in Definition 3 has the properties $D(A) \supset C^\infty(H_0)$, $AC^\infty(H_0) \subset C^\infty(H_0)$ iff the condition

$$\sum_m \lambda_m^{2p} \left| \sum_n A_{mn} \langle \Phi_n, f \rangle \right|^2 < \infty \quad (p \in \mathbf{N}, f \in C^\infty(H_0)),$$

is satisfied, as follows from the spectral theorem for H_0^p and Definition 3. By the positivity of the λ_m and Schwarz's inequality in $\mathscr{C}^2(\Omega)$, this condition holds if

$$\sum_m \lambda_m^{p} \left| \sum_n A_{mn} \langle \Phi_n, f \rangle \right| < \infty \quad (p \in \mathbf{N}, f \in C^\infty(H_0)). \tag{4.6}$$

To prove that (4.6) holds when $(A_{mn}) \in \mathscr{M}$, we will use the inequality

$$|\langle \Phi_n, f \rangle| \leqslant \lambda_n^{-q} \|H_0^q f\| \quad (q \in \mathbf{N}, f \in C^\infty(H_0)), \tag{4.7}$$

which follows by the fact that H_0 is invertible (because its spectrum lies at a positive distance from $0 \in \mathbf{R}$), Schwarz's inequality, and the normalization $\|\Phi_k\| = 1$ $(k \in \Omega)$. By (3.4b), (4.7), and (4.5), we have

$$\sum_{m,n} \lambda_m^p |A_{mn}| . |\langle \Phi_n, f \rangle| \leqslant \text{const} \sum_m \lambda_m^{p-q} \sum_n \lambda_n^{a(q)-r} < \infty$$

for each p and f in (4.6) if we choose positive integers $q = q(p) > p + \nu$ and $r > a(q) + \nu$, where the $a(q)$ are exponents of A.

(2) Let $A \in M$. Since $D(A)$ is dense in $\mathscr{H}$, A^* exists. For reasons of the same kind as those adduced in an analogous connection in the proof of Lemma 4(2) of Ref. 14, A^* leaves $C^\infty(H_0)$ invariant if the matrix (A_{mn}) with which A is associated has the properties

$$\sum_{m,n} \lambda_m^p |A_{nm}| . |\langle \Phi_n, f \rangle| < \infty \quad (p \in \mathbf{N}, f \in C^\infty(H_0)), \tag{4.8a}$$

$$\sum_{m,n} |A_{mn}| . |\langle \Phi_n, f \rangle| . |\langle \Phi_m, g \rangle| < \infty \quad (f, g \in C^\infty(H_0)). \tag{4.8b}$$

Statement (4.8a) (resp., (4.8b)) can be proved similarly to (4.6), by using (3.4a), (4.7), and (4.5) (resp., (3.4b), (4.7), and (4.5)). □

The considerable relevance of the next lemma to this paper was already mentioned in Sec. I and will be discussed more precisely after its proof below. In the proof of this lemma and hereafter, we will denote the restriction of an operator $A \in \mathscr{L}(\mathscr{H})$ to $C^\infty(H_0)$ by $(A)_0$.

Lemma 5: When equipped with the usual operations of operator addition and multiplication, and multiplication by scalars, M_0 is an algebra over $\mathbf{C}$ which is isomorphic to the matrix algebra $\mathscr{M}$.[37]

Proof: The mapping $\varphi: \mathscr{M} \to M_0$ of each element $\mathscr{A} \in \mathscr{M}$ into the element $(A)_0 \in M_0$, where A is the operator associated with $\mathscr{A}$, is a bijection, as follows for reasons similar to those mentioned in Remark (3) to Definition 3 in a similar connection. We will show that the relations

$$\varphi(\mathscr{A} + \mathscr{B}) = \varphi(\mathscr{A}) + \varphi(\mathscr{B}), \tag{4.9a}$$

$$\varphi(\mathscr{A}\mathscr{B}) = \varphi(\mathscr{A})\varphi(\mathscr{B}), \tag{4.9b}$$

$$\varphi(\alpha.\mathscr{A})=\alpha\varphi(\mathscr{A}) \tag{4.9c}$$

hold for all $\mathscr{A},\mathscr{B}\in\mathscr{M}$ and all $\alpha\in\mathbf{C}$. Relations (4.9a) and (4.9b) entail that M_0 is closed under operator summation and multiplication, and it is obviously closed under multiplication by complex numbers. Moreover, it is well-known and easily provable that the set of all operators in $\mathscr{L}(\mathscr{H})$, restricted to a linear manifold $\mathscr{D}\subset\mathscr{H}$ which they leave invariant, is an operator algebra over $\mathbf{C}$ if it is closed under each of these operations. Hence M_0 is such an algebra. By the properties (4.9a)–(4.9c) of the bijection φ, the algebras $\mathscr{M}$ and M_0 are isomorphic.

Fix such a triple $\mathscr{A},\mathscr{B},\alpha$ and let $A\leftrightarrow\mathscr{A},B\leftrightarrow\mathscr{B}$. We first prove (4.9a). By Lemma 3 and Definition 3, it is legitimate to define the operator $C\leftrightarrow\mathscr{A}+\mathscr{B}$, and Definition 3 and Schwarz's inequality in $\mathscr{L}^2(\Omega)$ entail that $C\supset A+B$. Since the domains of A,B,C contain $C^\infty(H_0)$ (Lemma 4), it follows that $(C)_0=(A+B)_0=(A)_0+(B)_0$, and thus that (4.9a) holds. As to (4.9c), it is an immediate consequence of $\alpha A\leftrightarrow\alpha.\mathscr{A}$, which in turn follows from Definition 3 and the fact that the mapping of $\mathscr{M}$ onto M prescribed in that definition is bijective.

To prove (4.9b), we proceed as follows. Let $G\leftrightarrow\mathscr{A}\mathscr{B}$, which is a legitimate definition by Lemma 3 and Definition 3. We will show that

$$G=AB \quad \text{on} \quad C^\infty(H_0). \tag{4.10}$$

Since A,B leave $C^\infty(H_0)$ invariant, (4.9b) follows immediately by (4.10). To prove (4.10), we fix $f\in C^\infty(H_0)$ and note that the repeated use of Definition 3 yields

$$ABf=\sum_m\left(\sum_n A_{mn}\sum_r B_{nr}\langle\Phi_r,f\rangle\right)\Phi_m. \tag{4.11}$$

If it were allowable to interchange the order of the n and r summations, (4.11) would yield

$$ABf=\sum_m\left(\sum_r\left(\sum_n A_{mn}B_{nr}\right)\langle\Phi_r,f\rangle\right)\Phi_m=Gf,$$

and we would be done. To justify this interchange, we note that

$$\begin{aligned}\left|\sum_n A_{mn}\sum_r B_{nr}\langle\Phi_r,f\rangle\right| &\leqslant\sum_n|A_{mn}|\sum_r|B_{nr}|.|\langle\Phi_r,f\rangle| \\ &\leqslant\|f\|\sum_{n,r}|A_{mn}|.|B_{nr}| \\ &\leqslant\text{const}\,\lambda_m^{a(p)}\left(\sum_n\lambda_n^{-p+b(q)}\right)\left(\sum_r\lambda_r^{-q}\right)<\infty\end{aligned} \tag{4.12}$$

for each $q\in\mathbf{N}$ greater than ν, provided that we choose $p=p(q)\in\mathbf{N}$ such that $p(q)>b(q)+\nu$. Here $a(p),b(q)$ denote exponents of A,B, respectively, and we have invoked (3.4a) and (4.5). The absolute convergence of the repeated sum over n,r in the lhs of the inequality in the first line of (4.12) legitimizes the above interchange. □

The main reason why Lemma 5 is of interest in this paper is that it and Lemma 4 have the following corollary.

Corollary: Let $\mathscr{A}_1,\ldots,\mathscr{A}_r,\mathscr{B}_1,\ldots,\mathscr{B}_s$ $(r,s<\infty)$ denote matrices in $\mathscr{M}$ which satisfy an equation of the form

$$g(\mathscr{A}_1,\ldots,\mathscr{A}_r)=h(\mathscr{B}_1,\ldots,\mathscr{B}_s), \tag{4.13}$$

where g (resp., h) denote finite linear combinations of products of $\mathscr{A}_i$'s (resp., $\mathscr{B}_j$'s) with coefficients in $\mathbf{C}$. Then the linear operators $A_1,\ldots,A_r$, $B_1,\ldots,B_s$ in M associated with these respective matrices satisfy the operator equation

$$g(A_1,\ldots,A_r)=h(B_1,\ldots,B_s) \quad \text{on} \quad C^\infty(H_0). \tag{4.13′}$$

Remark: This result allows one to recover in a rigorous form certain operator equations stated loosely in Sec. II, which play key roles in the present section. Here is a simple illustration. By (3.5), the equation

$$\text{ad } \mathscr{K}^{(0)}\mathscr{S}^{(1)}=\mathscr{V}_1-\mathscr{K}^{(1)}$$

holds. Since $H_0,S^{(1)},V_1,K^{(1)}$ are the operators associated with the respective matrices $\mathscr{K}^{(0)},\mathscr{S}^{(1)},\mathscr{V}_1,\mathscr{K}^{(1)}$ in $\mathscr{M}$ (by the line preceding (3.7), Definition 5(b), and the proof of Lemma 6), it follows from the corollary that the equation

$$\text{ad } H_0S^{(1)}=V_1-K^{(1)}$$

holds on $C^\infty(H_0)$. More complete information on operator equations obtaining on this dense set which were proved in this manner is provided by Lemma 6.

Proof: By (4.13) and the isomorphism between the algebras M_0 and $\mathscr{M}$, the equation $g((A_1)_0,\ldots,(A_r)_0)=h((B_1)_0,\ldots,(B_s)_0)$ emerges, where we recall that $(A_i)_0,(B_j)_0$ denote the restrictions of the respective operators A_i,B_j to $C^\infty(H_0)$. Since these restricted operators leave $C^\infty(H_0)$ invariant by Lemma 4, the last equation is equivalent to (4.13′). □

Lemma 6: (1) The operators $S^{(i)},V_i$ $(i\geqslant 1)$ and $K^{(i)}$ $(i\geqslant 0)$, together with $W_{i+1}^{(j)}$ $(0\leqslant j\leqslant i-1,\ i\geqslant 1)$ and $K_i(\varepsilon),W_{i+1}(\varepsilon)$ $(i\geqslant 1,\ \varepsilon\geqslant 0)$ are in M. (2) Equations (2.14), (2.15a), (2.15c), (2.16a), (2.16b) hold on $C^\infty(H_0)$.

Proof: (1) In view of Definitions 4 and 5(a)–5(c), $\mathscr{K}^{(0)}\in\mathscr{M}$, and the fact that $\mathscr{M}$ is an associative algebra, it suffices to show that $\mathscr{S}^{(i)},\mathscr{V}_i,\mathscr{K}_i\in\mathscr{M}$ for $i\geqslant 1$ in order to prove assertion (1) of the lemma. Note in particular that these properties of $\mathscr{S}^{(i)},\mathscr{K}^{(i)}(i\geqslant 1)$ and $\mathscr{M}$, together with (3.8c) and (3.9b), entail that $\mathscr{W}_{i+1}^{(j)}\in\mathscr{M}$ $(0\leqslant j\leqslant i-1,i\geqslant 1)$. Actually, it suffices to show that $\mathscr{V}_i\in\mathscr{M}$ for $i\geqslant 1$ to prove assertion (1), as can be seen from the following simple argument.

Suppose that $\mathscr{V}_k\in\mathscr{M}$ for some fixed $k\geqslant 1$, and hence that the following inequalities are fulfilled for all $p\in\mathbf{N}$, for suitable exponents $\alpha(k,p)$:

$$|(V_k)_{mn}|\leqslant\text{const }\lambda_m^{\alpha(k,p)}\lambda_n^{-p},\tag{4.14a}$$

$$|(V_k)_{mn}|\leqslant\text{const }\lambda_n^{\alpha(k,p)}\lambda_m^{-p}.\tag{4.14b}$$

By (3.10), (4.14a), and (4.1a), one has

$$|S_{mn}^{(k)}|\leqslant\text{const }\lambda_m^{\alpha(k,p)+\eta}\lambda_n^{-p+\eta}\leqslant\text{const }\lambda_m^{\alpha(k,p)+\eta}\lambda_n^{-p+[\eta]+1}\quad(p\in\mathbf{N}),$$

where $[x]$ denotes the integer part of $x\in\mathbf{R}$ and where we have used the fact that $\lambda_n^{\eta}\lambda_n^{-([\eta]+1)}$ is bounded for $n_1,\ldots,n_\nu\to\infty$, where $n=(n_1,\ldots,n_\nu)\in\Omega$. Therefore, choosing $q=p-[\eta]-1$ for each positive integer $p>[\eta]+1$, one sees that

$$|S_{mn}^{(k)}|\leqslant\text{const }\lambda_m^{\beta(k,q)}\lambda_n^{-q}\quad(q\in\mathbf{N}),\tag{4.15}$$

where $\beta(k,q)=\alpha(k,q+[\eta]+1)+\eta$. A similar proof using (4.14b) shows that (4.15) holds with m and n interchanged on its rhs, and with $\beta(k,q)$ possibly replaced by a larger exponent. We have therefore proved that $\mathscr{V}_k\in\mathscr{M}$ entails $\mathscr{S}^{(k)}\in\mathscr{M}$. On the other hand, $\mathscr{V}_k\in\mathscr{M}$ entails $\mathscr{K}^{(k)}\in\mathscr{M}$ by (3.6b). Next, we recall that $\mathscr{V}_1\in\mathscr{M}$ by Lemma 2 and we assume inductively that $\mathscr{V}_i\in\mathscr{M}$ for $1\leqslant i\leqslant\ell$, where ℓ is some positive integer. By combining

$$\mathscr{V}_{l+1}=-\sum_{j=1}^{l+1}\mathscr{K}_l^{(j)}\mathscr{K}^{(l+1-j)}\tag{4.16}$$

[recall definition (3.8b)] with the consequence $\mathscr{S}^{(i)},\mathscr{K}^{(i)}\in\mathscr{M}(1\leqslant i\leqslant\ell)$ of the inductive hypothesis and with $\mathscr{K}^{(0)}\in\mathscr{M}$, (3.9b), and Lemma 3(1), it follows that $\mathscr{V}_{\ell+1}\in\mathscr{M}$, and thus that $\mathscr{V}_i\in\mathscr{M}$ $(i\geqslant 1)$. Whence $\mathscr{S}^{(i)},\mathscr{K}^{(i)}\in\mathscr{M}$ at each such i. We have thus proved assertion (1).

(2) To prove assertion (2) of the lemma, we remark that each of the matrix Eqs. (3.13), (3.5), (3.8c), (3.8a), (3.8b) can be written as a relation of the form (4.13) involving only matrices in $\mathscr{M}$.

Furthermore, these equations formally yield the respective operator equations Eqs. (2.14), (2.15a), (2.15c), (2.16a), (2.16b) if they are subjected to the following substitutions. First, one replaces the matrices appearing explicitly in each of these equations by the operators with which they are associated according to Definition 5(b). Second, one replaces each operator $\mathscr{R}_{i-1}^{(j)}$ (resp., $\mathscr{R}_i^{(k)}$) occurring in (3.8b) (resp., (3.8c)) by the operator $R_{i-1}^{(j)}$ (resp., $R_i^{(k)}$), understood in the sense of Definition 5(c). In view of this remark and the above corollary, the operator equations mentioned in the first sentence of this paragraph are true on $C^\infty(H_0)$. □

The next lemma entails that $U_N(t,\varepsilon)$ in (3.14) is a well-defined unitary operator.

Lemma 7: For each $i\geqslant 1$, $\varepsilon\geqslant 0$ the operator $K_i(\varepsilon)$ is the unique self-adjoint extension of $K^{[i]}(\varepsilon)$ (Definition 5(b)).[38]

Proof: We fix $\varepsilon\geqslant 0$ and omit the ε-dependence of the relevant matrices and operators. We divide the proof into three parts. In the first, we show that each matrix $\mathscr{V}_i$ ($i\geqslant 1$) is Hermitian, a property which together with (3.6b), (3.7), and (3.12) immediately entails that every $\mathscr{K}^{[i]}$ ($i\geqslant 1$) is Hermitian. This last result is a basic ingredient in the proof of the self-adjointness of the K_i given in the second step. In the third step, we sketch a proof of the essential self-adjointness of the $K^{[i]}$.

(i) We prove that $\mathscr{V}_i^*=\mathscr{V}_i (i\geqslant 1)$ inductively. Note first that $\mathscr{V}_1^*=\mathscr{V}_1$ by (3.8a) and $V_{mn}=\overline{V_{nm}}$. Note next that if $\mathscr{V}_i^*=\mathscr{V}_i$ at any given $i\geqslant 1$, then $\mathscr{K}^{(i)}=\mathscr{K}^{(i)*}$ by (3.6b) and $\mathscr{S}^{(i)*}=-\mathscr{S}^{(i)}$ by (3.10) at this i. We now assume inductively that $\mathscr{V}_i=\mathscr{V}_i^*$ for $1\leqslant i\leqslant \ell$, where ℓ is some positive integer. By this assumption, together with $\mathscr{K}^{(0)}=\mathscr{K}^{(0)*}$ and (3.9b), each summand on the rhs of (4.16) is a Hermitian matrix, since it is a linear combination with real coefficients of finite products of factors ad $S^{(r)}$ ($1\leqslant r\leqslant \ell$) acting on operators $\mathscr{K}^{(s)}$ ($0\leqslant s\leqslant \ell$).[39] Thus $\mathscr{V}_{\ell+1}^*=\mathscr{V}_{\ell+1}$.

(ii) We fix $i\geqslant 1$ henceforth in the proof. In part (2) of Definition 5(b), we defined $K^{[i]}$ as the operator sum (3.11) and K_i as the operator associated with the matrix $\mathscr{K}^{[i]}$ in (3.12). From these definitions, Definition 3, and Schwarz's inequality in $\ell^2(\Omega)$, it follows that $K_i(\varepsilon)\supset K^{[i]}(\varepsilon)$. Since $\mathscr{K}^{(0)}$ is diagonal and the entries $\langle\Phi_m,K^{(j)}\Phi_n\rangle$ of $\mathscr{K}^{(j)}$ for $j\geqslant 1$ vanish if $\lambda_m\neq\lambda_n$, it follows that the quantities $\langle\Phi_m,K_i\Phi_n\rangle$ have the same vanishing property. Let $\{\Lambda_\sigma\}(\sigma\in N)$ be the set of all distinct eigenvalues Λ_σ of H_0, numbered in a fixed but arbitrary manner, and for each $\sigma\in\mathbf{N}$ let $\mathscr{J}_\sigma$ be the set of all λ_k's equal to Λ_σ. Since the degeneracy of each λ_k is finite, it follows that the $\mathscr{J}_\sigma$'s have the same properties attributed to the J_σ's in the Corollary to Lemma A.4, Ref. 14, in the special case $\mathscr{X}=\mathscr{H}$, $\Gamma=\Omega$, $\Psi=\Phi$, $\mathscr{A}=\mathscr{K}^{[i]}$. By this fact and the Hermiticity of $\mathscr{K}^{[i]}$, we may apply this corollary to infer that K_i is self-adjoint.

(iii) Since $K^{[i]}$ is a densely defined restriction of K_i, it is symmetric. That K_i is the unique self-adjoint extension of $K^{[i]}$ follows by standard arguments of the type invoked in step (4) of the proof of Lemma 6 in Ref. 14, including the use of the fact that $C^\infty(H_0)$ is invariant under $U_N(t)$ (see the proof of Lemma 8). □

Henceforth in this section, we fix $N\leqslant 1$.

Lemma 8: The Nth-order approximant $\psi_N(t,\varepsilon)$ is a well-defined element of $C^\infty(H_0)$ (Definition 5(d)).

Proof: We again fix $\varepsilon\geqslant 0$ and omit it in the relevant symbols. Definitions (3.14) and (3.15) make sense because: (a) $\psi_0\in C^\infty(H_0)$ and $U_N(t)$ is a well-defined unitary operator on $\mathscr{H}$ which leaves $C^\infty(H_0)$ invariant; (b) $R_N^{[N]}$ maps each operator in $\mathscr{L}(\mathscr{H})$ which leaves this set invariant into one having the same property. As to (a), in order for $U_N(t)C^\infty(H_0)\subset C^\infty(H_0)$ to hold, it suffices that the condition (4.6) be satisfied with A_{mn} replaced by $\langle\Phi_m,U_N(t)\Phi_n\rangle$ (see the first paragraph of the proof of Lemma 4). That it is satisfied follows by using the inequality

$$|\langle\Phi_m,U_N(t)\Phi_n\rangle|\leqslant\delta_{\lambda_m,\lambda_n} \tag{4.17}$$

(to which we will return in a moment), together with (4.7), (4.1b), and (4.5), to conclude that

$$\sum_{m,n}\lambda_m^p|\langle\Phi_m,U_N(t)\Phi_n\rangle|\cdot|\langle\Phi_n,f\rangle|\leqslant\text{const}\sum_m g_m\lambda_m^{p-q}\|H_0^q f\|\leqslant\text{const}\sum_m\lambda_m^{p-q+\nu-\rho}$$

$$<\infty\quad(p\in\mathbf{N},f\in C^\infty(H_0)),$$

if for each such p we choose $q \in \mathbf{N}$ large enough. Here, g_m denotes the degeneracy of λ_m. As to property (b), $(R_N^{[N]}A)_0$ (and in fact any $(R_j^{[i]}A)_0$) can be expressed as a finite linear combination of finite products of $(A)_0$ and $(S^{(i)})_0$'s if the operator A leaves $C^\infty(H_0)$ invariant [see Definition 5(c), (2.17), and the sentence containing (4.21)]. Since each $S^{(i)} \in M$, the $(S^{(i)})_0$'s leave the latter set invariant, whence (b) follows. To prove (4.17), we first note that since $\langle \Phi_m, K_N \Phi_n \rangle$ vanishes for $\lambda_m \neq \lambda_n$, the self-adjoint operators H_0 and K_N commute in the usual technical sense.[40] This and (3.15) entail that $\langle \Phi_m, U_N(t)\Phi_n \rangle$ has the same vanishing property. The proof of (4.17) is completed by invoking the unitarity of $U_N(t)$, Schwarz's inequality, and the normalization of the Φ_k. □

We now state the main lemma of this section.

Lemma 9: The following statements hold:

(1) At all $\varepsilon \geqslant 0$, $t \in \mathbf{R}$, $\psi_N(t,\varepsilon)$ satisfies the differential equation

$$i \frac{d\psi_N(t,\varepsilon)}{dt} = H_\varepsilon \psi_N(t,\varepsilon) + \varepsilon^{N+1} \varphi_N(t,\varepsilon), \tag{4.18a}$$

and at each such ε

$$\psi_N(0,\varepsilon) = \psi_0, \tag{4.18b}$$

where d/dt denotes the strong derivative in $\mathscr{H}$ and $\varphi_N(t,\varepsilon)$ is defined by (2.27) where $W_{N+1}(\varepsilon)$ and $\psi_N(t,\varepsilon)$ are as in Definitions 5(b) and 5(d), respectively, and where $\rho_N(t,\varepsilon)$ is given by (2.24), in terms of the relevant operators and superoperators specified by Definitions 5(b)–5(d).

(2) At each fixed $\varepsilon \geqslant 0$, $\varphi_N(t,\varepsilon)$ is strongly continuous in t at all $t \in \mathbf{R}$ and

$$\sup_{(t,\varepsilon) \in \mathbf{R} \times [0,\varepsilon_0]} \|\varphi_N(t,\varepsilon)\| < \infty \tag{4.19}$$

for each compact interval $[0,\varepsilon_0]$.

Proof: (1) We will prove assertion (1) of the lemma in steps (i) and (ii) below, but step (ii) also contains results which will be used to prove assertion (2).

(i) In this step, we will show that (2.13) holds on $C^\infty(H_0)$. To do this, we will prove that the rhs and lhs of (2.13), when restricted to this dense set, can be written as polynomials in ε and that the operator coefficients of the powers of ε involved agree on $C^\infty(H_0)$. As to the lhs of (2.13), note that $H_\varepsilon = H_0 + \varepsilon V_1$ on this set by the Remark to Condition V and by Eq. (4.2). The rhs of (2.13) can be expressed rigorously on $C^\infty(H_0)$ as a polynomial in ε, because of the statement

$$R_N^{[N]}(\varepsilon) K^{[N]}(\varepsilon) = \sum_{i,j=0}^{N} \varepsilon^{i+j} R_N^{(i)} K^{(j)} \quad \text{on} \quad C^\infty(H_0) \tag{4.20}$$

and since (2.14) also holds on this dense set [Lemma 6(2)]. Note that (4.20) is equivalent to (2.18) holding on the latter set. To prove (4.20), we will need the formally obvious fact that

$$(\text{ad } S^{(1)})^{n_1} \cdots (\text{ad } S^{(k)})^{n_k} \left(\sum_{j=0}^{r} \varepsilon^j A_j \right) = \sum_{j=0}^{r} \varepsilon^j (\text{ad } S^{(1)})^{n_1} \cdots (\text{ad } S^{(k)})^{n_k} A_j \quad \text{on} \quad C^\infty(H_0)$$

$$(n_1, \ldots, n_k \geqslant 0,\ k \geqslant 1), \tag{4.21}$$

where the A_j are operators in $\mathscr{L}(\mathscr{H})$ leaving $C^\infty(H_0)$ invariant. Indeed, (4.20) follows by Definition 5(c), (2.17), (3.11), the invariance of $C^\infty(H_0)$ under the $K^{(j)}$ $(j \geqslant 1)$ [Lemmas 4 and 6(1)], and (4.21), specialized to the case $A_j = K^{(j)}$ $(0 \leqslant j \leqslant r = N)$. To prove (4.21), we first note that ad $S^{(1)}(\Sigma_{j=0}^r \varepsilon^j A_j) = \Sigma_{j=0}^r \varepsilon^j \text{ad } S^{(1)} A_j$ on $C^\infty(H_0)$, since $(S^{(1)})_0$ and the $(A_i)_0$ belong to the complex algebra of all operators in $\mathscr{L}(\mathscr{H})$ having this invariance property. Second, we note that ad $S^{(1)}B$ leaves $C^\infty(H_0)$ invariant if $B \in \mathscr{L}(\mathscr{H})$ does. These two remarks and an easy inductive argument show that (4.21) holds for $k=1$, $n_1 \geqslant 1$. The proof of (4.21) can be completed by arguments of similar type which invoke the invariance of $C^\infty(H_0)$ under the $S^{(i)}$ and A_j.

Expressing the rhs and lhs of (2.13) on $C^\infty(H_0)$ as polynomials and using (2.17a), (2.17b), (2.20a), and (2.20b), together with the facts that (2.15a), (2.15c), (2.16a), and (2.16b) hold on this dense set by Lemma 6(2), it follows that the coefficients of each ε^i $(0\leqslant i\leqslant 2N)$ on the rhs and lhs of (2.13) are equal on $C^\infty(H_0)$.

(ii) The only nontrivial claim in assertion (1) of the lemma is (4.18a), since (4.18b) follows immediately by (3.14), (3.15), and (2.17b). The heart of our proof of (4.18a) is the justification of the relation

$$i\,\frac{d\psi_N(t)}{dt}=R_N^{[N]}(K_N U_N(t))\psi_0, \tag{4.22}$$

where here and henceforth, until further notice, we once more fix $\varepsilon\geqslant 0$ and suppress the ε-dependence of the vectors and operators considered. The remainder of the proof of (4.18a) is easy. If the operators $S^{(i)}$ in $R_N^{[N]}$ were bounded (which they are not in general), (4.22) would legitimately follow by differentiating (3.14) strongly wrt t and interchanging the strong differentiation operator d/dt with $R_N^{[N]}$ on the rhs. We will prove (4.22) by the following uniform convergence argument in which no hypotheses on the boundedness of these operators are made.

Note that if $A\in\mathscr{L}(\mathscr{H})$ leaves $C^\infty(H_0)$ invariant, then

$$R_N^{[N]}A=\sum_{\alpha\in I_N} c_\alpha T_\alpha A T'_\alpha \quad \text{on} \quad C^\infty(H_0) \tag{4.23}$$

holds, where the T_α and T'_α denote finite products of operators in M, namely, finite products of $S^{(i)}$'s or the unit operator in $\mathscr{H}$; the $c_\alpha=c_\alpha(\varepsilon)$ are monomials in ε with real coefficients; and the sum over α ranges over a finite index set $I_N\subset\mathbf{N}$. The N-dependence of the T_α, T'_α, and c_α has been suppressed for simplicity. To prove (4.23), one uses Definition 5(c), (2.17), and an inductive argument analogous to that used to arrive at (4.21). By (3.14) and (4.23), and since $\psi_0\in C^\infty(H_0)$ and $U_N(t)$ leaves $C^\infty(H_0)$ invariant (see the proof of Lemma 8),

$$\psi_N(t)=\sum_{\alpha\in I_N} c_\alpha T_\alpha U_N(t) T'_\alpha\psi_0. \tag{4.24}$$

This equation and the relation $T'_\alpha\psi_0\in C^\infty(H_0)$ motivate us to consider a time-dependent vector $TU_N(t)u\in\mathscr{H}$. Here $T=A_1\cdots A_r$ for some $r\geqslant 1$, with $A_i\in M$ $(i=1,...,r)$; $u\in C^\infty(H_0)$; and the A_i and u are ε- and t-independent. We write[35]

$$TU_N(t)u=\sum_m F_m(t)\Phi_m, \tag{4.25}$$

where

$$F_m(t)=\langle\Phi_m, TU_N(t)u\rangle=\langle A_r^*\cdots A_1^*\Phi_m, U_N(t)u\rangle. \tag{4.26}$$

Here, we have used the facts that $D(A_i^*)\supset C^\infty(H_0)\supset\Phi$ for each i and that each A_i^* leaves $C^\infty(H_0)$ invariant, which follow from the assumption that the A_i are in M (Lemma 4). Because $u\in C^\infty(H_0)\subset D(K_N)$ [Lemmas 4 and 6(1)], $U_N(t)u$ has a strongly continuous t-derivative $d(U_N(t)u)/dt=-iK_N U_N(t)u$ on $\mathbf{R}$. Thus $F_m\in C^1(\mathbf{R})$ by (4.26). In addition, each A_i leaves $\mathbf{C}^\infty(H_0)$ invariant, and hence

$$\frac{dF_m(t)}{dt}=-i\langle A_r^*\cdots A_1^*\Phi_m, K_N U_N(t)u\rangle=-i\langle\Phi_m, TK_N U_N(t)u\rangle=-i\langle\Phi_m, TU_N(t)K_N u\rangle. \tag{4.27}$$

The series (4.25) and the series

$$\sum_m \frac{dF_m(t)}{dt}\,\Phi_m \tag{4.28}$$

obtained from it by termwise strong differentiation converge uniformly on $\mathbf{R}$ wrt t in the strong $L^2(\mathbf{R}^\nu)$-sense. This follows from the respective estimates

$$\sum_m \sup_{t\in\mathbf{R}}|F_m(t)|<\infty, \tag{4.29a}$$

$$\sum_m \sup_{t\in\mathbf{R}}\left|\frac{dF_m(t)}{dt}\right|<\infty, \tag{4.29b}$$

which will be proved shortly. Since each F_m is continous on $\mathbf{R}$, and since the series (4.25) has the above uniform convergence property, we conclude that its sum $TU_N(t)u$ is strongly continous on $\mathbf{R}$, and by (4.29a) it is also bounded thereon:

$$\sup_{t\in\mathbf{R}}\|TU_N(t)u\|<\infty. \tag{4.30}$$

Similarly, by the strong convergence of the series (4.25) on $\mathbf{R}$, the strong uniform convergence of the series (4.28) thereon, and the continuity of the derivatives dF_m/dt on $\mathbf{R}$, strong termwise differentiation of the series (4.25) is legitimate at all $t\in\mathbf{R}$, and thus

$$\frac{d}{dt}\,TU_N(t)u=\sum_m\ \frac{dF_m(t)}{dt}\,\Phi_m=-iTU_N(t)K_Nu, \tag{4.31}$$

where we have used (4.27).

The desired result (4.22) is proved by using (4.24), (4.31), and (4.23), keeping in mind that each T_α (resp., $T'_\alpha\psi_0$) is a special case of T (resp., u), that $C^\infty(H_0)$ is invariant under K_N [Lemmas 4 and 6(1)] and $U_N(t)$, and that the sums over α in (4.23) and (4.24) contain only a finite number of terms.

Estimates (4.29) can be established in a way which closely parallels that used to obtain estimate (4.48) in Ref. 14. However, since these estimates play such a crucial role in the proof of the present lemma, we will now prove (4.29b), the proof of (4.29a) being almost identical. Setting $w=K_Nu$ in the expression after the third equality sign in (4.27), we obtain

$$\begin{aligned}\left|\frac{dF_m(t)}{dt}\right|&=|\langle\Phi_m,TU_N(t)w\rangle|=\left|\sum_r\ T_{mr}\sum_s\ \langle\Phi_r,U_N(t)\Phi_s\rangle\langle\Phi_s,w\rangle\right|\\&\leqslant\sum_{r,s}\ |T_{mr}|\,\delta_{\lambda_r,\lambda_s}|\langle\Phi_s,w\rangle|\\&\leqslant\text{const}\sum_r\ \lambda_r^{t(p)}\lambda_m^{-p}g_r\lambda_r^{-q}\|H_0^qw\|\\&\leqslant\text{const}\,\lambda_m^{-p}\sum_r\ \lambda_r^{t(p)+\upsilon-\rho-q}\leqslant\text{const}\,\lambda_m^{-p}\quad(p\in\mathbf{N}),\end{aligned} \tag{4.32}$$

if for each such p we choose $q>2\nu+t(p)-\rho$. Here, $t(p)$ denotes an exponent of the matrix $(T_{mn})=\langle\Phi_m,T\Phi_n\rangle\in\mathscr{M}$. To arrive at (4.32), we also used the following results. To derive the second equality in (4.32), we applied Lemma A2 of Ref. 14 [corrected according to part (3) of Ref. 15 and with $r=2,\mathscr{X}=\mathscr{H}$, $\Gamma=\Omega$, $\Psi=\Phi,\mathscr{N}=C^\infty(H_0)$, $A^{(1)}=T,A^{(2)}=U_N(t)$]. This application was legitimate because T and $U_N(t)$ leave $C^\infty(H_0)$ invariant, $\Phi\subset C^\infty(H_0)\subset D(T^*)$, $D(U_N(t))=\mathscr{H}$, and $w\in C^\infty(H_0)$.[41] To prove the inequalities in (4.32), we used the fact that each T_{mn} satisfies an inequality of the form (3.4b) and that (4.17), (4.1b), (4.7), and (4.5) hold. Summing over all $m\in\Omega$ in (4.32) and choosing $p>\nu$, (4.29b) follows by (4.5).

In order to complete the proof of (4.18a), we will show briefly that the four equalities in (2.23) hold rigorously if the operators and superoperators in (2.23) are understood in the present sense. This, together with (3.14), $\psi_N(t)\in C^\infty(H_0)$ (Lemma 8), the fact that (2.13) holds on $C^\infty(H_0)$, and definition (2.27) (understood as stated in the present lemma), implies trivially that (4.18a) is true.

As to the first equality (2.23), we remark that the invariance of $C^\infty(H_0)$ under K_N and $U_N(t)$ and the fact that $(K_N)_0=(K^{[N]})_0$ entail that $K_N U_N(t)$ and $K^{[N]}U_N(t)$ are equal on $C^\infty(H_0)$ and that they both leave this set invariant. By this, together with (4.23), the fact that T'_α has this invariance property, and $\psi_0 \in C^\infty(H_0)$, it follows that (4.22) holds with K_N replaced by $K^{[N]}$, and thus that the first equality (2.23) obtains. The second equality (2.23) follows trivially from Definition 5(c). The third follows by applying relation (2.25), with $A=K^{[N]}$ and $B=U_N(t)$, which is legitimate because of $\psi_0 \in C^\infty(H_0)$ and the invariance of $C^\infty(H_0)$ under $K^{[N]}$, $U_N(t)$, and the $S^{(i)}$.[21] Finally, the fourth can be proved rigorously by a procedure similar to that used to establish (4.20).

(2) To prove assertion (2) of the lemma, it is convenient to exhibit the dependence on ε of the vectors and operators considered. A key element in the proof is the formula

$$\varphi_N(t,\varepsilon)=\sum_k \varepsilon^k \sum_\beta V_{k\beta} U_N(t,\varepsilon) V'_{k\beta}\psi_0, \tag{4.33}$$

where the number of terms in each of the two sums is finite, the $V_{k\beta}$ and $V'_{k\beta}$ are finite products of t-independent operators in M (their N-dependence has been omitted for simplicity), and the only ε-dependence is that shown explicitly. Hence, each term in the β-sum is a special case of the vector $TU_N(t,\varepsilon)u$ defined immediately after (4.24). Equation (4.33) follows from (2.27), the facts that (2.14) holds on $C^\infty(H_0)$ [Lemma 6(2)] and that the $W^{(i)}_{N+1}$ leave $C^\infty(H_0)$ invariant [Lemmas 4 and 6(1)], (4.24) and (2.24), and arguments of the type adduced to prove (4.20). We also note the boundedness property

$$\sup_{(t,\varepsilon)\in \mathbf{R}\times[0,\infty]} \|TU_N(t,\varepsilon)u\|<\infty. \tag{4.30$'$}$$

In view of the stated properties of the rhs of (4.33), together with the strong continuity of $TU_N(t,\varepsilon)u$ in t on $\mathbf{R}$ at each $\varepsilon \geqslant 0$, (4.30′), and the compactness of the interval $[0,\varepsilon_0]$, assertion (2) follows. To prove (4.30′), one observes that steps analogous to those in (4.32) show that $|F_m(t,\varepsilon)| \leqslant \text{const}\, \lambda_m^p (p \in \mathbf{N})$, where we now exhibit the ε-dependence of $F_m(t)$ and where the constant is m- and ε-independent. By this result and proceeding in a way analogous to that in which (4.29b) was derived using (4.32), (4.30′) follows. □

We proceed to use use Lemma 9 to give a short rigorous proof of Theorem 1.

Proof of Theorem 1: We will need the following simple proposition.[42]

Let $A \in \mathscr{L}(\mathscr{H})$ be self-adjoint, let $u:\mathbf{R}\to\mathscr{H}$ be strongly differentiable on $\mathbf{R}$, with $u(t) \in D(A) (t \in \mathbf{R})$, and let

$$i\,\frac{du(t)}{dt}=Au(t)+f(t) \quad (t\in\mathbf{R}),$$

$$u(0)=u_0,$$

where $f:\mathbf{R}\to\mathscr{H}$ is strongly continuous. Then

$$\|u(t)\| \leqslant \|u_0\| + \left| \int_0^t \|f(s)\| ds \right|. \tag{4.34}$$

We fix $\varepsilon \geqslant 0$ for the moment and consider the case when $u(t)=\psi(t,\varepsilon)-\psi_N(t,\varepsilon)$, $A=H_\varepsilon$, and $f(t)=\varepsilon^{N+1}\varphi_N(t,\varepsilon)$, whence $u_0=0$. The hypotheses of the proposition are fulfilled in this case, since H_ε is self-adjoint, $\psi(t,\varepsilon)$ and $\psi_N(t,\varepsilon)$ are in $D(H_\varepsilon)$ for $t\in\mathbf{R}$, and Lemma 9 holds. Specializing (4.34) to the case in question and taking into account the boundedness property (4.19), we recover (2.29) for $t\in\mathbf{R}$, $0\leqslant\varepsilon\leqslant\varepsilon_0$. This completes the proof of Theorem 1. □

ACKNOWLEDGMENTS

The author is indebted to Professor A. Vogt of the Mathematics Dept., Georgetown University and the referee for their constructive remarks. This work was partially supported by the U.S. Office of Naval Research. This paper is dedicated to the memory of Professor Martin Kummer.

[1] M. Kummer and R. Gompa, J. Math. Phys. **29**, 1405 (1988).
[2] R. Gompa, ''Approximations to the quantum mechanical time evolution,'' Ph.D. thesis, University of Toledo, Toledo, OH, 1987.
[3] M. Born and P. Jordan, *Elementare Quantenmechanik* (Springer, Berlin, 1930), Secs. 38, 41.
[4] This normal form approach differs from that used, e.g., by J. Bellisard and M. Vittot, Ann. Inst. Henri Poincaré **52**, 175 (1990) and by M. Degli Esposti, S. Graffi, and J. Herczinski, Ann. Phys. (N.Y.) **209**, 364 (1991). These authors use a transformation based on a single exponential operator, rather than on a product of such operators, to reduce the pertinent Hamiltonian operators to normal form.
[5] J. Moser, Mem. Am. Math. Soc. **81** (1968).
[6] R. Churchill, M. Kummer, and D. L. Rod, J. Diff. Eqns. **49**, 359 (1983).
[7] A. Baider and R. Churchill, Proc. R. Soc. Edinburgh Sect. A **108**, 27 (1988).
[8] P. Lochak and C. Meunier, *Multiphase Averaging for Classical Systems* (Springer-Verlag, New York, 1988), especially Appendix 7.
[9] J. M. Finn, in *Local and Global Methods of Nonlinear Dynamics*, Lecture Notes in Physics **252**, edited by A. W. Sáenz, W. W. Zachary, and R. Cawley (Springer-Verlag, Berlin, 1986), p. 63.
[10] V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961); Proc. Natl. Acad. Sci. USA **48**, 199 (1962); Commun. Pure Appl. Math. **26**, 1 (1969).
[11] M. Kummer, Nuovo Cimento B **1**, 123 (1971).
[12] N. D. Antoniou, ''On the convergence of perturbation theories that avoid secular terms in quantum mechanics,'' Ph.D. thesis, University of Toledo, Toledo, OH, 1972.
[13] See, however, the interesting article by M. Kummer, in *Essays in Classical and Quantum Mechanics* (Festschrift for A. W. Sáenz), edited by J. A. Ellison and H. Überall (Gordon and Breach, Reading, PA, 1991), p. 139.
[14] A. W. Sáenz, J. Math. Phys. **37**, 2182 (1996).
[15] See the Erratum for Ref. 14 in J. Math. Phys. **38**, 4398 (1997). In addition, the following corrections should be made in Ref. 14. (1) Definition 2 on p. 2189 of Ref. 14 should be replaced by a slightly more general version of Definition 2 of the present paper. More precisely the following changes should be made on p. 2189 of Ref. 14: (a) in lines 2 and 3, replace ''and $\{\Delta_n\}$ be a sequence of finite subsets $\Delta_n \subset \Gamma$ such that Δ_n eventually contains any given finite subset of Γ.'' by ''and $\{\Delta_n\}$ $(n=(n_1,\ldots,n_p)\in\Gamma)$ be a sequence of subsets of $\mathbf{R}^p$ of the form $\Delta_n=[0,n_1]\times\cdots\times[0,n_p]$.''; (b) in line 3 (resp., lines 4 and 5), replace $\lim_{n\to\infty}$ (resp., $\mathrm{s}-\lim_{n\to\infty}$) by $\lim_{n_1,\ldots,n_p\to\infty}$ (resp., $\mathrm{s}-\lim_{n_1,\ldots,n_p\to\infty}$). (2) In the second line of the displayed relations for $\|\psi(t,\varepsilon)-\psi_N(t,\varepsilon)\|$ in p. 2187 and in Eq. (4.52) of p. 2202, replace $\int_0^{|t|}\|g_N(s,\varepsilon)\|ds$ by $|\int_0^t\|g_N(s,\varepsilon)\|ds|$. (3) In the Appendix, p. 2202, omit the final comma in line 3; replace $A_{\alpha\beta}=0(\alpha\in\Gamma)(\alpha\in\Gamma)$ at the beginning of line 4 by $(\alpha\in\Gamma)$; and in line 5 replace ''matrix element'' by ''matrix element $\langle\Psi_\alpha,A\Psi_\beta\rangle$.'' In Lemma A2, p. 2203, the condition $f\in\mathscr{M}$ should read $f\in\mathscr{N}$.
[16] Here and henceforth, all equations involving t (resp., ε) hold for $t\in\mathbf{R}$ (resp., $\varepsilon\geqslant 0$), in the absence of an explicit statement to the contrary.
[17] The relevant definitions are given in Sec. III.
[18] Here and henceforth, equations involving unsummed subscripts m,n which label matrix elements of linear operators in the Φ-basis are intended to hold for all $m,n\in\mathbf{Z}_+^\nu$.
[19] The obvious formal power series defining the exponentials in (2.11) are analogous to the Lie series occurring in classical mechanics. See, e.g., Refs. 8 and 9.
[20] We mention in passing that, strictly speaking, A in Eq. (2.12) stands for a formal power series in which the coefficient of ε^0 is A and in which the higher-order coefficients are the zero operator. In general, (2.12) is not rigorously true as a formal-operator-series identity, but does hold strictly if the equality sign is replaced by $\supset$, signifying that the operator coefficient of each ε^i $(i\geqslant 0)$ in the formal series on the lhs is an extension of the corresponding coefficient on the rhs. For a review of formal power series with coefficients in $\mathscr{L}(\mathscr{H})$, based on the properties of the pseudo-algebra of unbounded linear operators, see Ref. 2, Chap. 1, Sec. 1.1. The basic rules of calculation with unbounded operators in Hilbert space are given, e.g., by F. Riesz B. Sz. Nagy, *Functional Analysis* (F. Ungar Publishing, New York, 1955), p. 299. In the rigorous discussions in Sec. IV, we will avoid using formal power series.
[21] According to Theorem 1.13 in p. 15 of Ref. 2, generally (2.25) does not hold as an operator identity, but is true on any linear manifold in $\mathscr{H}$ which is invariant under A, B, and the $S^{(i)}$.
[22] M. Reed and B. Simon, *Methods of Mathematical Physics*, Vol. II (Academic, New York, 1975), Example 3, p. 266.
[23] See, for example, V. I. Arnold, Russ. Math. Surveys **18**, 85 (1963) and G. Gallavotti, *The Elements of Mechanics* (Springer-Verlag, New York, 1983), Sec. 5.12.
[24] C. L. Siegel and J. K. Moser, *Lectures in Celestial Mechanics* (Springer-Verlag, Berlin, 1971), pp. 260, 261.
[25] If A is an operator on $\mathscr{H}$, its domain is denoted by $D(A)$ and its restriction to a subset $\mathscr{D}\subset D(A)$ of $\mathscr{H}$ by $A|\mathscr{D}$. If A,B are operators on $\mathscr{H}$, then we write $A=B$ iff A and B have the same domain and range; we say that $A=B$ on $\mathscr{D}\subset\mathscr{H}$ if $\mathscr{D}\subset D(A)\cap D(B)$ and $A|\mathscr{D}=B|\mathscr{D}$; and $A\subset B$ (resp., $A\supset B$) means that A is a restriction (resp., an extension) of B.
[26] See the above volume of Reed and Simon,[22] p. 201.
[27] The range of the rows and columns of matrices denoted by $\mathscr{A}=(A_{mn})$, $\mathscr{B}=(B_{mn})$, etc., will always be Ω, even if this is not stated explicitly.
[28] M. H. Stone, *Linear Transformations in Hilbert Space*, A.M.S. Colloq. Pub., Vol. XV (A.M.S., New York, 1932), pp. 88–93.
[29] J. von Neumann, Math. Ann. **102**, 49 (1929); J. f. Math. **161**, 208 (1929).
[30] See, for example, Ref. 28, p. 88 and R. G. Cooke, *Infinite Matrices and Sequence Spaces* (Macmillan and Co., Ltd., London, 1950), pp. 8–9.
[31] The meaning of $\mathscr{M}$ in this paper is different from the one it has in Ref. 14, where it denotes the set of operators in $\mathscr{L}(\mathscr{H})$ which leave $C^\infty(H_0)$ invariant.
[32] The notations Lemma 2(2), Lemma 4(1), etc., denote assertion (2) of Lemma 2, assertion (1) of Lemma 4, etc. The

lemmas cited in the text are those stated in Sec. IV of the present paper, unless a statement to the contrary is made.

[33] The symbol ''const'' appearing in a relation involving the unsummed indices m or n should be understood to be a constant which is independent of these indices.

[34] Whenever inequalities (3.4a) or (3.4b) are invoked in an argument in the text, allusion is being made to the fact that the entries of the matrix in $\mathscr{M}$ considered in the argument satisfy an inequality of one of these respective forms.

[35] Symbols of the type $\Sigma_r f_r$, $\Sigma_s g_s$,..., where the $f_r, g_s, \ldots$ are complex numbers or vectors in $\mathscr{H}$ whose indices $r,s,\ldots$ label matrix elements, are abbreviations for sums $\Sigma_{r\in\Omega} f_r$, $\Sigma_{s\in\Omega} g_s$,..., respectively, understood in the sense of Definition 2. Similarly, ''double'' sums $\Sigma_{r,s} f_{rs}$ of complex numbers f_{rs} labeled by such indices are, naturally, defined as $\lim_{k_1,\ldots k_v, l_1,\ldots,l_v\to\infty} \Sigma_{r\in\Delta_k}\Sigma_{s\in\Delta_l} f_{rs}$ if this limit exists. Here $\Delta_k, \Delta_l (k=(k_1,\ldots,k_v)\in\Omega,\ l=(l_1,\ldots,l_v)\in\Omega)$ are as specified in Definition 2.

[36] This estimate follows readily by using, in particular, Cauchy's integral test for the convergence of multiple infinite series, together with (2.1c) and the positivity of the ω_i.

[37] Recall that two algebras A, A' over the same field F are isomorphic if there exists a bijective map $h: A\to A'$ such that $h(ab)=h(a)h(b)$, $h(a+b)=h(a)+h(b)$, and $h(\alpha a)=\alpha h(a)$ for all $a,\ b\in A$ and all $\alpha\in F$.

[38] Of course, arguments analogous to ones used in the proof of this lemma show that each $K^{(i)}$ $(i\geqslant 1)$ is self-adjoint.

[39] We recall an elementary fact. Let $\mathscr{F}$ be a set of matrices of the form $\mathscr{A}=(A_{mn})$ $(m,n\in\Omega)$ with entries in $\mathbf{C}$ which is an associative algebra over $\mathbf{C}$ when equipped with the usual operations of matrix summation and multiplication by scalars, and with matrix multiplication in the sense specified in Sec. III. Then, if $\mathscr{A}_r=-\mathscr{A}_r^*$ $(r=1,...,s<\infty)$ and $\mathscr{B}=\mathscr{B}^*$ are in $\mathscr{F}$, it follows that $(\mathrm{ad}\,\mathscr{A}_1\ldots\mathrm{ad}\,\mathscr{A}_s)\mathscr{B}$ is hermitian.

[40] See, for example, M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. I, (Academic, New York, 1972), pp. 270, 271.

[41] To show that $(T_{mn})\in\mathscr{M}$ and that the operators T and T^* have the properties ascribed to them in this sentence of the text, we proceed as follows. T leaves $C^\infty(H_0)$ invariant since $T=A_1\ldots A_r$ and since each $A_i\in M$. Define $T'\leftrightarrow\mathscr{A}_1\ldots\mathscr{A}_r$, where $\mathscr{A}_i$ is the matrix in $\mathscr{M}$ with which A_i is associated. Hence $T'\in M$ and $T'\supset T$ [by Definition 3 and Schwarz's inequality in $\ell^2(\Omega)$]. Therefore, $D(T'^*)\supset C^\infty(H_0)$ (Lemma 4). But $T'\supset T$ entails $T^*\supset T'^*$, and hence $D(T^*)\supset C^\infty(H_0)\supset\Phi$. Since $T'\supset T$ and $D(T)\supset\Phi$, it follows that $(T_{mn})=(\langle\Phi_m, T\Phi_n\rangle)=(\langle\Phi_m, T'\Phi_n\rangle)\in\mathscr{M}$.

[42] This result is a slight variant of Proposition 10.6, p. 186 of V. P. Maslov and M. V. Fedoriuk, *Semi-Classical Approximation in Quantum Mechanics* (Reidel, Dordrecht, 1981).

Generalized Pöschl–Teller, Toda, Morse potentials and $SL(2,R)$ group

M. Sezgin, A. Y. Verdiyev,[a] and Y. A. Verdiyev[b]
Trakya University, Department of Mathematics, 22030, Edirne, Turkey

(Received 28 July 1997; accepted for publication 19 December 1997)

Quantum integrable systems related with $SL(2,R)$ group manifold in various coordinate systems defined by the various decompositions: KHK, HAH, HKH, NHN, NHK of the group $SL(2,R)$ on one parameter subgroups K,H,A,N are considered. The explicit expressions for waves functions, spectra, and S-matrices are given.
© *1998 American Institute of Physics.* [S0022-2488(98)00204-7]

I. INTRODUCTION

The one dimensional integrable quantum systems related to Lie algebra and symmetric spaces are considered in Refs. 1–3. The key point of the relation between the theory of homogeneous spaces and quantum systems is the existence of a simple transformation of the radial part Laplace–Beltrami operators on homogeneous spaces to some Hamiltonian quantum systems. Here we consider the quantum systems with generalized Pöschl–Teller, Toda, and Morse potentials:

$$V(x)=\begin{cases}\dfrac{m^2+n^2-2mn\cosh x-1/4}{\sinh^2 x}, & m,n=0,\pm1,\pm2,\ldots\\ \dfrac{-1/16+m_2^2}{\sinh^2(x/2)}+\dfrac{1/16-m_1^2}{\cosh^2(x/2)} & \\ -\dfrac{\mu^2+\nu^2-2\mu\nu\cosh x+1/4}{\sinh^2 x}, & \dfrac{\mu^2+\nu^2-2\mu\nu\cos x+1/4}{\sin^2 x}\\ -\lambda\mu e^{-x}, \quad \lambda^2e^{-2x}-2n\lambda e^{-x}, & -\infty<\lambda,\mu,\nu<\infty, \quad n=0,\pm1,\pm2,\ldots\end{cases} \tag{1.1}$$

related to various coordinate systems on $SL(2,R)$ group manifold.

In the special case, when **m** or **n=0** (Pöschl–Teller potential), the quantum systems have only scattering states,[1,2] but in the case when $mn>0$ the quantum systems have also bound states.

This report is a generalization of the results given in Ref. 2 and carries on the research reviewed in Ref. 4.

The special linear group $SL(2,R)$ in two dimensional space over the real field consist of matrices:

$$g=\begin{bmatrix}ab\\cd\end{bmatrix}, \quad a,b,c,d\in R, \quad \det g=ad-bc=1. \tag{1.2}$$

They satisfy the condition $glg^t=I$, where $I=\left[\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right]$. The metric in the space of the matrices X given by $[x_1,x_2]=\mathrm{tr}(Ix_1Ix_2^t)$, $x_1,x_2\in X$ define the $SL(2,R)$ group manifold $[g,g]=2$, $g\in SL(2,R)$ in this space. The metric matrix in this manifold is given by the formula:

$$h_{ij}=2\,\mathrm{tr}(I\dot{g}_{t_i}I\dot{g}^t_{t_j}), \quad i,j=1,2,3, \tag{1.3}$$

[a]TÜBITAK, Marmara Research Center, Research Institute for Basic Sciences, Department of Physics, P.O. Box 21, 41470, Gebze, Turkey, Electronic-mail: verdieva@hamlin.cc.boun.edu.tr
[b]On leave of absence from Physical Institute of Academy Sciences of Azerbaijan Republic.

 © 1998 American Institute of Physics

where $\dot{g}_{t_i}=dg_{t_i}/dt_i$, t_i are the coordinates of the manifold. The coordinate systems on the manifold are defined by the decompositions of the group $SL(2,R)$ on one parameter subgroups K,H,A,N.

In this study, we consider the decompositions:[5] $g=k(\varphi)h(\theta)k(\psi)$, $g=h(\varphi)a(\theta)h(\psi)$, $g=h(\varphi)k(\theta)h(\psi)$, $g=n(t)h(\theta)n^t(q)$, $g=n(t)h(\theta)k(\varphi)$, $g\in SL(2,R)$, $h\in H$, $k\in K$, $a\in A$, $n\in N$, *where*

$$n(t)=\begin{pmatrix}1&0\\t&1\end{pmatrix},\quad h(\theta)=\mathrm{diag}(e^{\theta/2},e^{-\theta/2}),\quad a(\theta)=\begin{pmatrix}\cosh\,\theta/2&\sinh\,\theta/2\\\sinh\,\theta/2&\cosh\,\theta/2\end{pmatrix},$$

$$k(\theta)=\begin{pmatrix}\cos\,\theta/2&\sinh\,\theta/2\\-\sin\,\theta/2&\cos\,\theta/2\end{pmatrix}. \tag{1.4}$$

The one dimensional quantum systems related with the above decompositions are presented in the Sec. II. Here we give explicit expressions for wave functions, spectra, and S-matrices.

II. THE QUANTUM SYSTEMS RELATED WITH $SL(2,R)$ GROUP MANIFOLD

Let us consider the various decompositions of the elements g of the group $SL(2,R)$:

$$g=k(\varphi)h(\theta)k(\psi),\quad -\infty<\theta<\infty,-2\pi<\varphi,\psi<2\pi. \tag{2.1}$$

Using Eq. (1.3) for this factorization we have the metric

$$ds^2=d\theta^2-d\varphi^2-d\psi^2-2\cosh\,\theta d\varphi d\psi \tag{2.2}$$

and the Laplace–Beltrami operator

$$\Delta_{L,B}=\frac{1}{\sinh\,\theta}\frac{\partial}{\partial\theta}\sinh\,\theta\frac{\partial}{\partial\theta}+\frac{1}{\sinh^2\,\theta}\left(\frac{\partial^2}{\partial\varphi^2}+\frac{\partial^2}{\partial\psi^2}-2\cosh\,\theta\frac{\partial^2}{\partial\psi\partial\varphi}\right). \tag{2.3}$$

The free motion on the $SL(2,R)$ group manifold is defined by the equation

$$\Delta_{L,B}f(\varphi,\theta,\psi)=\sigma(\sigma+1)f(\varphi,\theta,\psi). \tag{2.4}$$

The solutions of Eq. (2.4) are the matrix elements of the representations of the group $SL(2,R)$ [or the group $SO(1,2)$] in the basis associated with compact subgroup $\mathbf{K}$.[4,5]

After the substitution $f(\varphi,\theta,\psi)=u(\theta)e^{-im\varphi}e^{-in\psi}$ we have

$$\Delta_\theta u(\theta)=\sigma(\sigma+1)u(\theta), \tag{2.5}$$

where the radial part of the Laplace–Beltrami operator Δ_θ has the form

$$\Delta_\theta=\frac{d^2}{d\theta^2}+\tanh\,\theta\frac{d}{d\theta}-\left(\frac{m^2+n^2-2mn\,\cosh\,\theta}{\sinh^2\,\theta}\right).$$

Using $dg=\sqrt{h}d\theta d\varphi d\psi$, $h=|\det(h_{ij})|=\sinh^2\,\theta$, and making the transformation

$$u(\theta)=\frac{1}{\sqrt[4]{h(\theta)}}\Psi_E,\quad \sqrt[4]{h}\Delta_\theta\frac{1}{\sqrt[4]{h}}=H+1/4, \tag{2.6}$$

Eq. (2.5) is reduced to the one dimensional Schrödinger equation

$$\frac{d^2\Psi_E(\theta)}{d\theta^2}+(E-V(\theta))\Psi_E(\theta)=0 \tag{2.7}$$

with the generalized Pöschl–Teller potential

$$V(\theta)=\frac{m^2+n^2-2mn\ \cosh\ \theta-1/4}{\sinh^2\ \theta} \tag{2.8}$$

and the energy spectrum $E=-(\sigma+1/2)^2$.

The regular solution of the Schrödinger equation (2.7) at $\theta=0$ is given by

$$\Psi(\theta)=c\left(\cosh\frac{\theta}{2}\right)^{|-m-n|+1/2}\left(\sinh\frac{\theta}{2}\right)^{|n-m|+1/2}F\left(\sigma-|m|+1,-\sigma-|m|;|n-m|+1;-\sinh^2\frac{\theta}{2}\right). \tag{2.9}$$

It follows that the quantum system has the bound and scattering states. The square integrable normalized wave functions $\Psi_E(\theta)$:

$$\int_0^\infty|\Psi_E(\theta)|^2d\theta=1 \tag{2.10}$$

for $E=-(l+1/2)^2$ with $|m|=l+1+k,\ l,k=0,1,2,...$ have the form:

$$\Psi_E(\theta)=\frac{\Gamma(l-|m|+1)}{\Gamma(l-|n|+1)(|n|-|m|)!}\left(\cosh\frac{\theta}{2}\right)^{|-m-n|+1/2}\left(\sinh\frac{\theta}{2}\right)^{|n-m|+1/2}$$
$$\times F\left(l-|m|+1,-l-|m|;|n-m|+1;-\sinh^2\frac{\theta}{2}\right). \tag{2.11}$$

For the continuous spectrum $\sigma=-1/2+i\rho,\ E=\rho^2>0$ using the analytical continuation formula for the hypergeometric function 2.10.(2)[6] we find the following asymptotic expression:

$$\underset{\theta\to\infty}{\Psi(\theta,\rho)}=c_1[B(m,n,\rho)e^{-i\rho\theta}+\hat{B}(m,n,\rho)e^{i\rho\theta}], \tag{2.12}$$

where

$$B(m,n,\rho)=\frac{\Gamma(|n|-|m|+1)\Gamma(-2i\rho)}{\Gamma(1/2-|m|-i\rho)\Gamma(1/2+|n|-i\rho)}.$$

The solutions of the Schrödinger equation (2.7) satisfying the condition

$$\int_0^\infty\Psi_E(\theta)\overline{\Psi_{E'}(\theta)}d\theta=\delta(\rho-\rho') \tag{2.13}$$

have the form:

$$\Psi_E(\theta)=c_1\left(\cosh\frac{\theta}{2}\right)^{|-m-n|+1/2}\left(\sinh\frac{\theta}{2}\right)^{|n-m|+1/2}$$
$$\times F\left(\frac{1}{2}-|m|+i\sqrt{E},\ \frac{1}{2}-|m|-i\sqrt{E};|n-m|+1;-\sinh^2\frac{\theta}{2}\right), \tag{2.14}$$

where $|c_1|=1/2\sqrt{\pi}|B|$.

The S-matrix is found to be given by

$$S=\frac{B}{\bar{B}}=\frac{\Gamma(-2i\sqrt{E})\Gamma(\frac{1}{2}-|m|+i\sqrt{E})\Gamma(\frac{1}{2}+|n|+i\sqrt{E})}{\Gamma(2i\sqrt{E})\Gamma(\frac{1}{2}-|m|-i\sqrt{E})\Gamma(\frac{1}{2}+|n|-i\sqrt{E})}. \tag{2.15}$$

The decomposition $g=k(\varphi)h(\theta)k(\psi)$ corresponds to the spherical coordinate system on $SL(2,R)$ group manifold. The substitution $\varphi\to\varphi_1+\varphi_2,\ \psi\to\varphi_1-\varphi_2$ define the bispherical coordinate system on this manifold.

Then we have the metric

$$ds^2 = d\theta^2 - 4\cosh^2\ \theta/2 d\varphi_1^2 + 4\sinh^2\ \theta/2 d\varphi_2^2. \tag{2.16}$$

The solutions of the related quantum systems with potential:

$$V(\theta) = \frac{-1/16 + m_2^2}{\sinh^2\ \theta/2} + \frac{1/16 - m_1^2}{\cosh^2\ \theta/2} \tag{2.17}$$

are obtained from the solutions of Eq. (2.9) by the substitutions $m \to m_1 + m_2$, $n \to m_1 - m_2$. For the factorization

$$g = h(\varphi)a(\theta)h(\psi), \quad \theta, \varphi, \psi \in (-\infty, \infty); \tag{2.18}$$

we have the metric

$$ds^2 = d\theta^2 + d\varphi^2 + d\psi^2 + 2\cosh\ \theta d\varphi d\psi \tag{2.19}$$

and Laplace–Beltrami operator

$$\Delta_{L,B} = \frac{1}{\sinh\ \theta} \frac{\partial}{\partial\theta} \sinh\ \theta \frac{\partial}{\partial\theta} - \frac{1}{\sinh^2\ \theta} \left(\frac{\partial^2}{\partial\varphi^2} + \frac{\partial^2}{\partial\psi^2} - 2\cosh\ \theta \frac{\partial^2}{\partial\psi\partial\varphi} \right). \tag{2.20}$$

The solutions

$$f(\varphi, \theta, \psi) = u(\theta) e^{-i\mu\varphi} e^{-i\nu\psi}, \quad \mu, \nu \in (-\infty, \infty) \tag{2.21}$$

of the equation:

$$\Delta_{L,B} f(\varphi, \theta, \psi) = \sigma(\sigma + 1) f(\varphi, \theta, \psi) \tag{2.22}$$

are the matrix elements of the representations of the group $SL(2,R)$ in the basis associated with noncompact subgroup H. These solutions are calculated in Ref. 5, but they have not symmetry with respect to $\sigma \to -1 - \sigma$. The symmetrization (unitarization) of these solutions by the intertwining operator of the representation of the group $SO(1,2)$ has been done in Ref. 4.

Related quantum system has the potential:

$$V(\theta) = -\frac{\mu^2 + \nu^2 - 2\mu\nu\ \cosh\ \theta + 1/4}{\sinh^2\ \theta} \tag{2.23}$$

and energy spectrum $E = -(\sigma + 1/2)^2$.

The solutions of the Schrödinger equation (2.7) with the condition

$$\int_{-\infty}^{\infty} \Psi_E(\theta) \overline{\Psi_{E'}(\theta)} d\theta = \frac{1}{(\sigma + 1/2)\tanh(\sigma + 1/2)\pi} \delta(\sigma - \sigma'), \quad \text{for} \quad \sigma = -1/2 + i\rho \tag{2.24}$$

are given in Ref. 4. They have the form

$$\Psi_E(\theta) = (\cosh\ \theta/2)^{1/2} (\sinh\theta/2)^{1/2} \times [q(\mu, \nu, \sigma) P_\sigma^{-i\mu, -i\nu}(\cosh\ \theta) + q(-\mu, -\nu, \sigma) P_\sigma^{i\mu, i\nu}(\cosh\ \theta)], \tag{2.25}$$

where

$$P_\sigma^{\pm i\mu, \pm i\nu}(\cosh\ \theta) = \frac{1}{\Gamma(1 \pm i\nu \mp i\mu)} (\cosh\ \theta/2)^{\mp i(\mu - \nu)} (\sinh\ \theta/2)^{\mp i(\mu + \nu)} \times F(\sigma + 1 \mp i\mu, -\sigma \mp i\mu, 1 \pm i\nu \mp i\mu, -\sinh^2\ \theta/2),$$

and

$$q(\mu,\nu;\sigma)=q_{\mu,\nu}=\frac{1}{\sinh[(i\mu-i\nu)/2]\pi}\sqrt{\Gamma(1+\sigma-i\mu)\Gamma(-\sigma-i\mu)\Gamma(1+\sigma+i\nu)\Gamma(-\sigma+i\nu)}$$
$$\times\sqrt{\sin\frac{\sigma+i\mu}{2}\pi\cos\frac{\sigma-i\mu}{2}\pi\cos\frac{\sigma+i\nu}{2}\pi\sin\frac{\sigma-i\nu}{2}\pi}.$$

Using the analytical continuation formula for the hypergeometric functions 2.10.(2)[6] we find the asymptotic expression of Eq. (2.25) for $\sigma=-1/2+i\rho,\ \rho=\sqrt{E}$ as

$$\Psi^{E}_{\theta\to\infty}(\theta)=[q_{-\mu,-\nu}B(\mu,\nu,\rho)+q_{\mu,\nu}B(-\mu,-\nu,\rho)]e^{-i\rho\theta}$$
$$+[q_{-\mu,-\nu}B(\mu,\nu,-\rho)q_{\mu,\nu}B(-\mu,-\nu,\rho)]e^{i\rho\theta},\tag{2.26}$$

where

$$B(\mu,\nu,\sigma)=\frac{(i\nu-i\mu)\Gamma(-2i\rho)}{\Gamma(1/2-i\rho-i\mu)\Gamma(1/2-i\rho+i\nu)}.$$

The S-matrix is found to be given by

$$S=\frac{q_{-\mu,-\nu}B(\mu,\nu,\rho)+q_{\mu,\nu}B(-\mu,-\nu,\rho)}{q_{-\mu,-\nu}B(\mu,\nu,-\rho)+q_{\mu,\nu}B(-\mu,-\nu,-\rho)}.\tag{2.27}$$

The square integrable solution of Schrödinger equation is found from (2.25) by setting $\sigma=l=0,1,2,\ldots$.

$$\Psi_E(\theta)=q(-\mu,-\nu,l)(\cosh\theta/2)^{1/2}(\sinh\theta/2)^{1/2}\Big\{P_l^{i\mu,i\nu}(\cosh\theta)$$
$$\times\frac{\Gamma(1+l+i\nu)}{\Gamma(1+l-i\nu)}\frac{\Gamma(1+l-i\mu)}{\Gamma(1+l+i\mu)}P_l^{-i\mu,-i\nu}(\cosh\theta)\Big\}.\tag{2.28}$$

Using the analytical continuation formula for hypergeometric functions 2.10.(4)[6] we have

$$\Psi_E(\theta)=\dot{q}(-\mu,-\nu,l)(\cosh\theta/2)^{1/2}(\sinh\theta/2)^{1/2}Q_l^{i\mu,i\nu}(\cosh\theta),$$

where

$$\dot{q}(-\mu,-\nu,l)=\sqrt{\Gamma(1+l+i\mu)\Gamma(-l+i\mu)\Gamma(1+l-i\nu)\Gamma(-l-i\nu)}\times\sqrt{\sinh\nu\pi\sinh\mu\pi}$$

and

$$Q_l^{i\mu,i\theta}(\cosh\theta)=\frac{\Gamma(1+l-i\mu)\Gamma(1+l+i\nu)}{2\Gamma(2l+2)}\left(\frac{\cosh\theta-1}{2}\right)^{(i\nu-i\mu)/2}\left(\frac{\cosh\theta+1}{2}\right)^{-l-1-(i\nu-i\mu)/2}$$
$$\times F\left(1+l-i\mu,1+l+i\nu;2l+2;\frac{2}{1+\cosh\theta}\right).\tag{2.29}$$

These functions are square integrable and the orthogonality condition

$$\int_{\infty}^{\infty}\Psi_E(\theta)\overline{\Psi_{E'}(\theta)}d\theta=\frac{1}{E}\delta_{E,E'}\tag{2.30}$$

follows from the condition Eq. (2.24) by analytical continuation.

The quantum system related with the decomposition:

$$g=h(\varphi)k(\theta)h(\psi),\quad \varphi,\psi\in(-\infty,\infty),\quad -\pi/2<\theta<\pi/2$$

has the potential

$$V(\theta)=\frac{\mu^2+\nu^2-2\mu\nu\ \cosh\ \theta+1/4}{\sin^2\ \theta}. \tag{2.31}$$

The solutions of this quantum system are obtained from Eq. (2.29) by the analytical continuation $\theta\rightarrow i\theta$. For the factorization

$$g=n(t)h(\theta)n^t(q), \quad -\infty<t,\theta,q<\infty;$$

we have the metric

$$ds^2=d\theta^2-4e^{\theta}dtdq \tag{2.32}$$

and Laplace–Beltrami operator

$$\Delta_{L,B}=\frac{\partial^2}{\partial\theta^2}+\frac{\partial}{\partial\theta}-e^{-\theta}\frac{\partial^2}{\partial t\partial q}. \tag{2.33}$$

The solutions

$$f(t,\theta,q)=u(\theta)e^{-i\lambda t}e^{-i\mu q}, \quad \lambda,\mu\in(-\infty,\infty) \tag{2.34}$$

of the equation:

$$\Delta_{L,B}f(t,\theta,q)=\sigma(\sigma+1)f(t,\theta,q) \tag{2.35}$$

are the matrix elements of the representations of the group $SL(2,R)$ [or the group $SO(1,2)$] in the basis associated with noncompact subgroup N.[5] The related quantum system has the potential

$$V(\theta)=-\lambda\mu e^{-\theta} \quad \text{(generalized Toda potential)} \tag{2.36}$$

and energy spectrum $E=-(\sigma+1/2)^2$. This quantum system has only scattering states when $\lambda\mu<0$ and the wave functions $\Psi_E(\theta)$ are given by the Macdonald K-functions:

$$\Psi_E(\theta)=\tilde{c}_3e^{\theta/2}K_{2i\sqrt{E}}(2\sqrt{-\lambda\mu}e^{-\theta/2}). \tag{2.37}$$

In order to calculate S-matrix and $\tilde{c}_3$-factor we use the definition of the K-function 7.2.2.(12) and (13) Ref. 6;

$$K_\nu(z)=\frac{\pi}{2\sin\ \nu\pi}\,[I_{-\nu}(z)-I_\nu(z)], \tag{2.38}$$

where

$$I_\nu(z)=\frac{(z/2)^\nu}{\Gamma(\nu+1)}\,{}_0F_1\left(\nu+1;\frac{z^2}{4}\right).$$

It follows that

$$I_\nu\underset{z\rightarrow 0}{}(z)\rightarrow\frac{z^\nu}{2^\nu\Gamma(1+\nu)}. \tag{2.39}$$

Thus we have:

$$\Psi_E\underset{\theta\rightarrow\infty}{}(\theta)=\tilde{c}_3[De^{2i\rho\theta}+\bar{D}e^{-2i\rho\theta}], \quad \rho=\sqrt{E}; \tag{2.40}$$

where

$$D=\frac{(\sqrt{\lambda\mu})^{-2i\rho}}{\Gamma(1-2i\rho)}.$$

So, the S-matrix to be found is given by

$$S=\frac{D}{\bar{D}}=\frac{(\sqrt{\lambda\mu})^{-i4\sqrt{E}}\Gamma(1+2i\sqrt{E})}{\Gamma(1-2i\sqrt{E})}. \tag{2.41}$$

From the condition

$$\int_{-\infty}^{\infty}\Psi_E(\theta)\overline{\Psi_{E'}(\theta)}d\theta=\delta(\rho-\rho') \tag{2.42}$$

it follows that $|\tilde{c}_3|=1/2\sqrt{\pi}|D|$.

When $\lambda\mu>0$ in the quantum system Eq. (2.36) has bound and scattering states. For the bound states the wave functions are $\Psi_E(\theta)$, $E=-(l+1/2)^2$ with the condition

$$\int_0^{\infty}|\Psi_E(\theta)|^2d\theta=1 \tag{2.43}$$

has the form:

$$\Psi_E(\theta)=\frac{1}{2\sqrt{l+1/2}}\,e^{\theta/2}J_{2l+1}(2\sqrt{\lambda\mu}e^{-\theta/2}). \tag{2.44}$$

The wave functions $\Psi_E(\theta)$, $E=\rho^2$ for the scattering states are obtained from Eq. (2.38) by analytical continuation: $\sqrt{-\lambda\mu}\to i\sqrt{\lambda\mu}$, $\lambda\mu>0$.

Finally we consider decomposition:

$$g=n(t)h(\theta)k(\varphi),\quad -\infty<t,\theta<\infty,0\leqslant\varphi<2\pi. \tag{2.45}$$

For this factorization we have the metric

$$ds^2=d\theta^2-d\varphi^2+2e^{\theta}\,dtd\varphi \tag{2.46}$$

and Laplace–Beltrami operator

$$\Delta_{L,B}=\frac{\partial^2}{\partial\theta^2}+\frac{\partial}{\partial\theta}+e^{-2\theta}\,\frac{\partial^2}{\partial t^2}+2e^{-\theta}\,\frac{\partial^2}{\partial t\partial\varphi}. \tag{2.47}$$

The solutions

$$f(t,\theta,\varphi)=u(\theta)e^{i\lambda t}e^{-in\varphi} \tag{2.48}$$

of the equation:

$$\Delta_{L,B}f(t,\theta,\varphi)=\sigma(\sigma+1)f(t,\theta,\varphi) \tag{2.49}$$

are the matrix elements of the representations of the group $SL(2,R)$ in the mixed basis associated with noncompact **N** and compact **K** subgroups. This matrix elements are calculated in Refs. 5 and 7. Here we give the orthogonality conditions and calculate normalization factors in continuous $\sigma=-1/2+i\rho$, $0<\rho<\infty$ and discrete $\sigma=l$, $l=0,1,2,...$ cases. The related quantum system has the potential

$$V(\theta)=\frac{\lambda^2}{e^{2\theta}}-\frac{2n\lambda}{e^{\theta}}\quad\text{(generalized Morse potential)} \tag{2.50}$$

and energy spectrum $E=-(\sigma+1/2)^2$.

The solutions of this quantum system which tend to zero $\Psi \to 0$ when $\theta \to -\infty$ are given by the Whittaker functions.

$$\Psi_E(\theta) = c_2 e^{\theta/2} W = c_2 e^{\theta/2} \left\{ \frac{\Gamma(-2i\rho)}{\Gamma(1/2 - i\rho - n\alpha)} M_{n\alpha, i\rho}(2|\lambda| e^{-\theta}) + \frac{\Gamma(2i\rho)}{\Gamma(1/2 + i\rho - n\alpha)} M_{n\alpha, -i\rho}(2|\lambda| e^{-\theta}) \right\}, \tag{2.51}$$

where $\alpha = \text{sign}\,\lambda$, $M_{n\alpha, \pm i\rho}(2|\lambda|z) = (2|\lambda|z)^{1/2 \pm i\rho} e^{-|\lambda| z}\, {}_1F_1(\pm i\rho - n\alpha + 1/2, \pm 2i\rho + 1; 2|\lambda|z)$, $z = e^{-\theta}$. For the scattering states with spectrum $E = -(\sigma + 1/2)^2 = \rho^2$ the asymptotic expression of the $\Psi_E(\theta)$-function when $\theta \to \infty$ has the form:

$$\Psi_E(\theta) \underset{\theta \to \infty}{=} c_2 [A(\rho, n\alpha) e^{-i\rho\theta} + \overline{A}(\rho, n\alpha) e^{i\rho\theta}], \tag{2.52}$$

where

$$A(\rho, n\alpha) = \frac{\Gamma(-2i\rho)(2|\lambda|)^{i\rho + 1/2}}{\Gamma(1/2 - i\rho\text{-}n\alpha)}.$$

The normalization factor c_2 is defined from the condition

$$\int_{-\infty}^{\infty} \Psi_E(\theta) \overline{\Psi_{E'}(\theta)}\, d\theta = \delta(\rho - \rho'). \tag{2.53}$$

It follows that

$$|c_2| = \frac{1}{2\sqrt{\pi}|A|}.$$

The S-matrix is found to be given by

$$S = \frac{A}{\overline{A}} = \frac{(2|\lambda|)^{2i\rho} \Gamma(-2i\sqrt{E}) \Gamma(1/2 + i\sqrt{E} - n\alpha)}{\Gamma(2i\sqrt{E}) \Gamma(1/2 - i\sqrt{E} - n\alpha)}. \tag{2.54}$$

From the expression of the potential Eq. (2.50) it follows that for $n\lambda > 0$ and $n\lambda > \lambda^2$ quantum systems have also bound states with spectrum $E = -(l + 1/2)^2$. For the bound states the solutions have the form:

$$\Psi_E(\theta) = \chi z^{l + 1/2} e^{-|\lambda| z}\, {}_1F_1(l + 1 - n\alpha, 2l + 2; 2|\lambda|z), \quad z = e^{-\theta}, \tag{2.55}$$

where $l + 1 - n\alpha = -k$, $k = 0, 1, 2, \ldots$.

Using the formula 6.4.(14)[6] for the degenerate hypergeometric functions from the condition

$$\int_{-\infty}^{\infty} |\Psi_E(\theta)|^2 d\theta = 1 \tag{2.56}$$

for the normalization factor χ, we obtain

$$\chi = \sqrt{\frac{(2l + 2)_k (2|\lambda|)^{2l + 1}}{(2l)!(n\alpha - l - 1)!}}. \tag{2.57}$$

In conclusion, we see one dimensional integrable quantum systems are the results of the broken symmetry of the free one particle motion by the geodesic paths.

ACKNOWLEDGMENTS

We would like to thank I. Hakki Duru and G. A. Kerimov for useful discussions. The authors, A. Y. Verdiyev and Y. A. Verdiyev, are grateful to TÜBİTAK for supporting the research.

[1] M. A. Olshanetsky and A. M. Perelomov, ''Quantum integrable systems related to Lie algebra,'' Phys. Rep. **94**, 313–404 (1993).
[2] C. Dane and Y. A. Verdiyev, ''Integrable systems of group SO(1,2) and Green's functions,'' J. Math. Phys. **37**, 39–60 (1996).
[3] Y. A. Verdiyev, ''Quantum integrable systems related with symmetric spaces of the groups U(1,2) and Sp(1,2) and Green's functions on these spaces,'' J. Math. Phys. **36**, 3320–3331 (1995).
[4] Y. A. Verdiyev, *Harmonic Analysis on Homogeneous Spaces of SO(1,2)* (Hadronic, MA, 1988).
[5] N. Ya. Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions* (Kluwer Academic, London, 1991), Vols. I, II, III.
[6] *Bateman Manuscript*, edited by A. Erdelyi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vols. I, II.
[7] E. G. Kalnins, ''Mixed basis matrix elements for the subgroup reduction of SO(2,1),'' J. Math. Phys. **14**, 654—657 (1973).

Duality, partial supersymmetry, and arithmetic number theory

Donald Spector[a)]
Department of Physics, Eaton Hall, Hobart and William Smith Colleges, Geneva, New York 14456

(Received 13 October 1997; accepted for publication 8 December 1997)

We find examples of duality among quantum theories that are related to arithmetic functions by identifying distinct Hamiltonians that have identical partition functions at suitably related coupling constants or temperatures. We are led to this after first developing the notion of partial supersymmetry, in which some, but not all, of the operators of a theory have superpartners, and using it to construct fermionic and parafermionic thermal partition functions, and to derive some number theoretic identities. In the process, we also find a bosonic analog of the Witten index, and use this, too, to obtain some number theoretic results related to the Riemann zeta function. © *1998 American Institute of Physics.* [S0022-2488(98)03303-9]

I. INTRODUCTION

Arithmetic quantum theories, theories for which the partition function is related to the Riemann zeta function or other Dirichlet series, have yielded compelling results in mathematical physics.[1,2] In addition to providing physical derivations and interpretations of various number theoretic results, such as an understanding of the Möbius inversion formula based on supersymmetry, these theories possess, as does string theory, a Hagedorn temperature, and so these theories may serve as toy models for certain aspects of string theory.

In this paper, I find further evidence of the usefulness of arithmetic quantum theories in these areas. Most importantly, I find examples of duality in arithmetic theories, that is, of distinct arithmetic theories at different couplings that in fact have the same partition function. Such dualities are analogous to the coupling constant dualities of string theory. See Sec. VII for a fuller discussion of this comparison. In addition, I find physical underpinnings for additional number theoretic results. The most interesting examples here rest on ''partial supersymmetry,'' a property identified in this paper, in which some, but not all, of the creation and annihilation operators have superpartners. This leads not only to a physical basis for some number theoretic results, but also to a formulation of parafermions of noninteger order and to the identification of an analog for the Witten index based on a bosonic grading.

II. REVIEW OF NOTATION AND IDEAS

We here review the notation and way of thinking that allow a connection to be drawn between arithmetic number theory and quantum field theory.

Consider some noninteracting quantum field theories defined in terms of their creation and annihilation operators. Let $b_k^\dagger$ and b_k denote bosonic creation and annihilation operators, and let $f_k^\dagger$ and f_k denote fermionic creation and annihilation operators. In addition, we label the prime numbers p_k in order of increasing magnitude, so that $p_1=2$, $p_2=3$, $p_3=5$, etc.

For use throughout this paper, we define the bosonic Hamiltonian

$$H_B=\omega\sum_{k=1}^{\infty}\log(p_k)b_k^\dagger b_k, \tag{2.1}$$

and the fermionic Hamiltonian

a)Electronic mail: spector@hws.edu

 © 1998 American Institute of Physics

$$H_F=\omega\sum_{k=1}^{\infty}\ \log(p_k)f_k^\dagger f_k\,. \tag{2.2}$$

We label the states of such theories using a Gödel numbering. Let us denote the vacuum state $|1\rangle$. For the bosonic theory, we label the states

$$(b_{i_1}^\dagger)^{r_1}(b_{i_2}^\dagger)^{r_2}\cdots(b_{i_n}^\dagger)^{r}|1\rangle=|N\rangle, \tag{2.3}$$

where $N=(p_1)^{r_1}(p_2)^{r_2}\cdots(p_n)^{r_n}$. For simplicity, we use the atypical convention that the creation and annihilation operators produce normalized states. We may use a similar labeling for the states of the fermionic theory. Note that with this labeling, the Fock space of states for the bosonic field theory is labeled by the positive integers, while the Fock space of states for the fermionic field theory is labeled by the square-free positive integers. One sees[1] that

$$\text{tr}\ e^{-\beta H_B}=\sum_{m=1}^{\infty}\ \frac{1}{m^{\beta\omega}}=\zeta(\beta\omega) \tag{2.4}$$

and

$$\text{tr}(-1)^F e^{-\beta H_F}=\sum_{m=1}^{\infty}\ \frac{\mu(m)}{m^{\beta\omega}}, \tag{2.5}$$

where $\mu(m)$ is the Möbius inversion function.[3] Supersymmetry tells us, via the Witten index,[4] that $\text{tr}[(-1)^F e^{-\beta(H_B+H_F)}]=1$, and so

$$\sum_{m=1}^{\infty}\ \frac{\mu(m)}{m^{\beta\omega}}=\text{tr}[(-1)^F e^{-\beta H_F}]=\frac{1}{\zeta(\beta\omega)}. \tag{2.6}$$

III. THERMAL PARTITION FUNCTIONS

In the preceding work, we have identified the bosonic thermal partition function with the Riemann zeta function, but only the graded fermionic partition function with a number theoretic function. On the other hand, the physics of a theory is governed by the genuine thermal partition function, and so to understand the thermodynamics of the fermionic field theory of H_F, we need to evaluate the partition function

$$Z_F(\beta,\omega)=\text{tr}\ e^{-\beta H_F}. \tag{3.1}$$

To calculate this partition function, we will introduce a notion of partial supersymmetry. This will then give us a physical and algebraic understanding for a particular zeta function identity. This partition function, like the parafermionic one of the next section, was obtained through number theoretic arguments in Ref. 2. Thus it is our method in this section that is new, not the result.

Our approach is to start with the bosonic partition function and to cancel out just the right states so that we are left with the fermionic partition function. Suppose, then, that we define operators $q_k=(b_k)^2$ and $q_k^\dagger=(b_k^\dagger)^2$, and in addition we define operators c_k and $c_k^\dagger$ which have the same effect as b_k and $b_k^\dagger$, respectively, except that c_k and $c_k^\dagger$ both square to zero. It is easy to see, then, that we can write the bosonic Hamiltonian

$$H_B=\omega\sum_k\ \log(p_k)c_k^\dagger c_k+\omega\sum_k\ 2\log(p_k)q_k^\dagger q_k\,. \tag{3.2}$$

The two portions of this decomposition of H_B commute, and so

$$\text{tr}\ e^{-\beta H_B}=\text{tr}[e^{-\beta\omega\Sigma_k\ \log(p_k)c_k^\dagger c_k}]\text{tr}[e^{-\beta\omega\Sigma_k 2\log(p_k)q_k^\dagger q_k}]. \tag{3.3}$$

Because the c_k and $c_k^\dagger$ square to zero, the first term in this product is exactly the fermionic partition function.

Notice that the $q_k^\dagger$ and q_k operators just appear as ordinary creation and annihilation operators, albeit associated with energy $2\log(p_k)$. (Due to our nonconventional normalization convention, there is no problem with how the resulting states are normalized.) Adding to the original Hamiltonian the expression $\omega\Sigma_{k=1}^{\infty} 2\log(p_k) f_k^\dagger f_k$, we are in a position to cancel out all but the c_k and $c_k^\dagger$ terms by taking a graded sum, by means of a Witten index type of result. It is this situation in which we have fermionic superpartners for some of the bosonic creation and annihilation operators (the $q_k^\dagger$ and q_k) but not others (the $c_k^\dagger$ and c_k) that we term partial supersymmetry. We thus find that

$$Z_f(\beta,\omega) = \mathrm{tr}[e^{-\beta\omega\Sigma_k \log(p_k) c_k^\dagger c_k}] = \mathrm{tr}[(-1)^F e^{-\beta\omega\Sigma_k(\log(p_k) c_k^\dagger c_k + 2\log(p_k) q_k^\dagger q_k + 2\log(p_k) f_k^\dagger f_k)}]$$
$$= \mathrm{tr}[(-1)^F e^{-\beta(H_B + 2\beta H_F)}]. \tag{3.4}$$

The bosonic and fermionic pieces in the final expression in (3.4) factorize into pieces that are easy to evaluate (see Ref. 1), leading to the number theoretic result

$$Z_f(\beta,\omega) = \sum_{m=1}^{\infty} \frac{|\mu(m)|}{m^s} = \frac{\zeta(s)}{\zeta(2s)}, \tag{3.5}$$

where $s = \beta\omega$. Notice that we have obtained this number theoretic identity using a partial supersymmetry of a quantum mechanical theory, rather than the conventional explicit calculations of number theory using infinite products.

Our use of partial supersymmetry to reduce a bosonic to a fermionic theory appears calculationally much like a BRST[5] cancellation. However, there is no gauge symmetry that is enforcing these cancellations, and so we stick to the term ''partial supersymmetry.'' If one could identify a gauge invariance underlying this result, then one could perhaps recast these results as proper BRST cancellations. Such a result would undoubtedly provide further deep insights into the structure of these number theoretic quantum models.

IV. PARAFERMIONS OF ARBITRARY ORDER

A natural generalization of the above results is to consider

$$Z_r(\beta,\omega) = \mathrm{tr}[(-1)^F e^{-\beta(H_B + rH_F)}]. \tag{4.1}$$

As we know, $Z_1 = 1$, as the H_F terms cancel all the H_B terms, yielding the Witten index of theory,while, as we just saw, $Z_2 = Z_F$, because the fermionic terms cancel all quadratic and higher powers of the original bosonic creation operators.

For higher integers r, we can identify a partial supersymmetry as well. Let us define $\chi_k^\dagger$ and χ_k such that their rth powers, but no lower powers, vanish as operators, and let $r_k^\dagger = (b_k^\dagger)^r$ and $r_k = b_k^r$. Thus

$$H_B = \omega\sum_{k=1}^{\infty} (\log(p_k)\chi_k^\dagger\chi_k + \log(p_k^r) r_k^\dagger r_k). \tag{4.2}$$

Then the fermionic terms in the expression for Z_r (4.1) cancel the $r_k^\dagger$ and r_k terms in H_B, and so

$$Z_r(\beta,\omega) = \mathrm{tr}[(-1)^F e^{-\beta(H_B + rH_F)}] = \mathrm{tr}\ e^{-\beta H_r}, \tag{4.3}$$

where $H_r = \omega\Sigma_{k=1}^{\infty} \log(p_k)\chi_k^\dagger\chi_k$ is the Hamiltonian for order r parafermions for the logarithmic spectrum we are studying. We can evaluate this expression, as the bosonic and fermionic pieces of (4.1) are noninteracting and hence factorize. One sees that the partition function for parafermions of order r, i.e., objects that are subject to an exclusion principle that states that no more than $r-1$ parafermions can have the same quantum numbers, is given by

$$Z_r(\beta,\omega)=\frac{\zeta(\beta\omega)}{\zeta(r\beta\omega)}. \tag{4.4}$$

What is compelling about this result is that we have found it using partial supersymmetry.

One knows, of course, that the bosons are the infinite order limit of parafermions. It is interesting to see how that appears in this context. As we take r to infinity, the energies of states with fermionic excitations go to infinity, and so the fermionic creation and annihilation operators decouple,[6] leaving behind the purely bosonic partition function. Thus the recovery of the bosonic theory at infinite r is, in our language, a decoupling effect.

One of the most intriguing features of the partition function Z_r is that it is well-defined for noninteger values of r. We do not simply mean that noninteger values of r can be plugged into the zeta function expression (4.4). Rather, we can put noninteger values of r into (4.1) quite sensibly and naturally—after all, why should the parameters of a Hamiltonian be integral—and this yields, from an expression with ordinary bosons and fermions, a partition function for exotic parafermions, parafermions of nonintegral order. Thus although there is no simple exclusion principle that one could state for such exotic parafermions, nor a simple Hamiltonian directly in terms of creation and annihilation operators for such exotic parafermions, we can obtain the thermal partition function quite easily from

$$Z_r(\beta,\omega)=\mathrm{tr}[(-1)^F e^{-\beta(H_B+rH_F)}]. \tag{4.5}$$

What is the meaning of such theories? On the one hand, the partition function tells us what the theory is, so any quest for further interpretations might be viewed as superfluous. On the other hand, understanding what nonintegral parafermions mean in a practical sense is physically intriguing. It is worth observing that the ability to vary the parameter r continuously in (4.3) means that we have a way to vary continuously from conventional parafermions of one integer order to another. This should enable one to connect results in the different conventional parafermion theories, presumably through analyticity or topological arguments.

Note that for parafermions of order r real and less than 1, there is a temperature less than the Hagedorn temperature[1,7] where the thermal partition function vanishes, indicating that the physical sense of these theories is limited to $r>1$. This is no surprise. At $r=1$, the corresponding order 1 parafermions obey an exclusion principle such that states with one or more excitations are forbidden, which means that only the vacuum state is physical. (Indeed, $r=1$ is simply the case of the Witten index calculation, in which all states but the vacuum state are canceled out.) Phrased in this way, parafermions of order $r<1$ seem clearly to be physically untenable, as well as mathematically troublesome.

V. SUPERSYMMETRY AND BOSON NUMBER

Recall from Ref. 1 that the mathematical interpretation of the supersymmetry Witten index result

$$\mathrm{tr}\, e^{-\beta H_B}\ \mathrm{tr}[(-1)^F e^{-\beta H_F}]=1 \tag{5.1}$$

is

$$\sum_{m=1}^{\infty}\ \frac{\mu(m)}{m^s}=\frac{1}{\zeta(s)}, \tag{5.2}$$

while we also know from (5.1) that the graded fermionic partition function is the inverse of the thermal bosonic partition function. Now that we have just calculated the thermal fermionic partition function, we can ask what its inverse is. We will find the result physically, and give its corresponding mathematical interpretation.

Supersymmetry, we know, pairs bosonic and fermionic states. Consider the Hamiltonian $H_S=H_B+H_F$. Supersymmetry pairs each b_k with f_k and each $b_k^\dagger$ with $f_k^\dagger$. Let us define $(-1)^{N_B}$ to be the operator that has value $+1$ for Fock space states with an even number of bosonic creation operators, and -1 for Fock space states with an odd number of bosonic creation

operators. Note that $(-1)^{N_B}$ commutes with $f_k^\dagger$ and f_k. Because the supercharges of H_S involve terms like $b_k^\dagger f_k$ and $f_k^\dagger b_k$, all states with nonzero energy come in superpartnerships with an equal number of even and odd N_B states. Only the zero energy states need not be paired in this way, as they are the supersymmetry singlet states. Consequently, for the theory with Hamiltonian $H_S = H_B + H_F$, we see

$$\mathrm{tr}[(-1)^{N_B} e^{-\beta(H_B+H_F)}] = 1, \tag{5.3}$$

which is analogous to but distinct from the familiar Witten index result for $\mathrm{tr}[(-1)^F e^{-\beta H_s}]$. Since H_B and H_F commute with each other, we see that

$$\mathrm{tr} e^{-\beta H_F} = \frac{1}{\mathrm{tr}[(-1)^{N_B} e^{-\beta H_B}]}. \tag{5.4}$$

The inverse of the thermal fermionic partition function is the graded bosonic partition function.

The mathematical translation of this is as follows. Using the Gödel numbering described above, we can represent $(-1)^{N_B}$ by the arithmetic function $\psi(m)$, where $\psi(m)$ is totally multiplicative [i.e., $\psi(m)\psi(n) = \psi(mn)$ for all natural numbers m and n] and with $\psi(p^r) = (-1)^r$ for any prime number p. Thus, using the connection with Dirichlet convolution described in Ref. 1, we see that

$$\psi * |\mu|(m) = \mathbf{1}(m), \tag{5.5}$$

where $\mathbf{1}$ is the identity function for Dirichlet convolution, $\mathbf{1}(m) = \delta_{m,0}$, or, equivalently,

$$\sum_{m=1}^{\infty} \frac{\psi(m)}{m^s} = \frac{1}{\sum_{m=1}^{\infty}[|\mu|(m)/m^s]} = \frac{\zeta(2s)}{\zeta(s)}. \tag{5.6}$$

Thus we have used our thinking about number theory to recognize that, generically, we can introduce a bosonic grading quite analogous to the fermionic grading, and this has led us physically to another number theoretic identity, namely (5.6). We note that the bosonically graded index is topological in nature, unaffected by deformations of the parameters of a theory as long as the supercharges remain odd under $(-1)^{N_B}$, and thus it is reliably calculable even by approximation methods. Supersymmetry is unbroken if this bosonic index is nonzero. Note, too, that $(-1)^F$ corresponds to the arithmetic function μ, and $(-1)^{N_B}$ to the arithmetic function ψ. One learns from supersymmetry that the inverses of these functions under Dirichlet convolution obey a nice symmetry: $\mu^{-1} = |\psi|$, while $\psi^{-1} = |\mu|$. It would be interesting to find other pairs of functions with this type of partnering when taking the inverse, and to see if there is a supersymmetric reason underlying such a relationship.

VI. GRADED PARAFERMION PARTITION FUNCTIONS

Our discussion of graded partition functions leads us naturally to consider graded parafermion partition functions. Here we present some results, but leave the detailed calculation to the readers, and reserve speculation on these results for elsewhere.

Consider the parafermions of order r discussed previously. Define the operator $\hat{\eta}$ such that a one-parafermion state has $\hat{\eta}$-eigenvalue η, where $\eta^r = 1$. Then

$$\mathrm{tr}[\hat{\eta} e^{-\beta H_r}] = \mathrm{tr}[\hat{\eta} e^{-\beta H_B}]\mathrm{tr}[(-1)^F e^{-\beta r H_F}] \tag{6.1}$$

using our representation of parafermions (4.3) and partial supersymmetry. Note that the bosonic creation operators also have $\hat{\eta}$-eigenvalue η, while the fermionic creation operators are $\hat{\eta}$-neutral.

There is an alternative graded partition function that one can naturally define. Let $\hat{\xi}$ be an operator such that a one-parafermion state has $\hat{\xi}$-eigenvalue ξ where $\xi^r = -1$. Then using our method of partial supersymmetry and the notion of the bosonic index (5.3), we find that

$$\mathrm{tr}[\hat{\xi} e^{-\beta H_r}] = \mathrm{tr}[\hat{\xi} e^{-\beta H_B}]\mathrm{tr}[e^{-\beta r H_F}]. \tag{6.2}$$

The convention, again, is that the bosonic creation operators have $\hat{\xi}$-eigenvalue ξ, while the fermionic creation operators are $\hat{\xi}$-neutral.

The $\hat{\eta}$- and $\hat{\xi}$-graded partition functions can be written in these factorized forms involving products of purely bosonic and purely fermionic terms because of the existence of a partial supersymmetry in these expressions, which leads to cancellations along the lines of the Witten index in (6.1) and of the bosonically graded index (5.3) in (6.2). Other gradings would not admit such a factorization.

VII. DUALITY

The identities we have obtained above provide relationships for partition functions (thermal and graded) at different temperatures. This offers an interesting possibility, akin to duality in supersymmetric and string theories: that we can identify two distinct theories at distinct temperatures that nonetheless give rise to the same thermodynamics. We will in fact realize this possibility, with an example in which the Euler totient function plays a central role.

Before proceeding further, it is worth commenting on our use of the term ''duality.'' In string theory, this term is most conventionally used to refer to relationships that directly relate the weak and strong coupling regimes (roughly speaking, replacing a coupling constant g by $1/g$). Here we use the term in a somewhat broader sense to refer to identities among theories at different coupling constants in general. Although the relationships we identify do not simply involve inverting a coupling constant, they can be iterated to related weak and strong couplings. Furthermore, even without such iteration, the identities we find between arithmetic quantum theories can clearly play a role analogous to the role played by weak–strong coupling dualities in string theory, establishing the equivalence of superficially distinct theories, and thus allowing the use of one formulation of a theory to analyze another. Hence we employ the term ''duality'' to characterize the identities among arithmetic quantum theories that we establish below.

To begin our analysis, then, consider the thermal partition functions associated with the original Hamiltonians H_B and H_F, which we will denote as Z_B and Z_F, respectively. Then (3.5) can be interpreted as

$$Z_F(\beta)Z_B(2\beta)=Z_B(\beta). \tag{7.1}$$

Alternatively, if we denote the fermionic partition function graded with $(-1)^F$ as Δ_F, we have

$$\Delta_F(2\beta)=Z_F(\beta)\Delta_F(\beta). \tag{7.2}$$

What is interesting about these expressions is that they relate partition functions at different temperatures. There are two interpretations possible here:

(a) In (7.1), we have a noninteracting mixture of two systems, each at its own temperature, equivalent to a third system at its own temperature.

(b) In (7.2), we have a mixture of two systems thermalized with each other at one temperature equivalent to a different system at a different temperature.

Note that in (b) not all the partition functions are thermal (some are fermionically graded), while in (a) we have an unthermalized mixture of two systems. Ideally, we would like a thermalized system at one temperature equivalent to a thermalized system at a different temperature. It turns out that this is possible to achieve.

Consider the Euler totient function, defined by $\varphi(m)=\Sigma_{d:(d,m)=1}1$, i.e., it is the number of positive integers less than m that are relatively prime to m. Then it is a standard result in number theory that

$$\zeta(s-1)=\zeta(s)\sum_{m=1}^{\infty}\frac{\varphi(m)}{m^s}. \tag{7.3}$$

We can interpret this physically as follows. Let us imagine two theories, one is governed by the original Hamiltonian H_B from (2.1) above. The other has the same spectrum as H_B, but has degeneracies given by the totient function; it has $\varphi(m)$ states with energy $\omega\log(m)$. We call the Hamiltonian for this theory H_φ and its partition function Z_φ.

Then we see that, in terms of the dimensionless parameter $s=\beta\omega$,

$$Z_B(s-1)=Z_B(s)Z_\varphi(s). \tag{7.4}$$

In other words, a thermalized mixture of two ideal gases, governed by the Hamiltonians H_B and H_φ, is equivalent simply to a higher-temperature gas governed by H_B. Alternatively, we can phrase this as a coupling constant duality. The Hamiltonian H_B at coupling constant $\omega-1/\beta$ and the Hamiltonian H_B+H_φ at coupling constant ω have the same partition function at temperature $T=\beta^{-1}$. The presence of this duality indicates an important additional similarity connecting arithmetic quantum theories and string theories.

The duality described above is clearly intriguing. There are two especially interesting directions to consider.

(1) These expressions give a way to shift temperature via coupling new systems. Let us call a system governed by Hamiltonian H_B a Riemann gas, and one governed by H_φ a totient gas. Then what we have found is that if we start with a Riemann gas at one temperature, and then mix it with a totient gas at the same temperature, this is simply equivalent to taking the Riemann gas and increasing its temperature a certain amount. Note that we can iterate this process by mixing in additional totient gases, and thereby shift the temperature further. This gives us an alternative to standard methods such as the renormalization group for thinking about how to shift temperatures; relating such distinct methods is potentially powerful. The most interesting outcome would be to determine a way to generalize this notion of obtaining a temperature shift by mixing in a new quantum system, finding ways to apply it to theories not so closely tied to the Riemann zeta function.

(2) We have now identified, among the class of arithmetically based quantum models, theories that exhibit duality behavior in their partition functions, with the best example being the duality (7.4) occurring for the theory with Hamiltonian H_B. We also know that this same theory exhibits a Hagedorn temperature. Both these traits, existence of duality and existence of a Hagedorn temperature, are distinguishing characteristics of string theory. This adds further support to the notion that the H_B theory and its relatives are related in a meaningful way to string theory, and thus that exploration of these arithmetic theories will prepare us to understand better the nature of string theory.

We can expand on this idea a bit further. It is well-known that in string theory, the duality between large and small radii effectively means there is a minimum distance in string theory. On the other hand, the Hagedorn temperature means there is a maximum temperature in a theory. Perhaps, in some similar fashion, the existence of a maximum temperature is associated with a temperature duality (recall that the duality we found could be understood as a temperature or as a coupling constant duality) in and of itself. These number theoretic field theories offer us an arena other than string theory in which to explore this question.

VIII. MORE WITH PARTIAL SUPERSYMMETRY

In earlier sections, we found that we could evaluate certain physically interesting partition functions and simultaneously obtain some interesting number theoretic identities, using partial supersymmetry, in which some but not all of the bosonic creation operators had fermionic superpartners. There are other number theoretic results which appear as if they should have an interpretation via partial supersymmetry. We give here two particular examples to show the kinds of results that it appears ought to be possible to derive via partial supersymmetry, although to date no such derivation exists.

As a first example, consider the function

$$\sigma_\alpha(m)=\sum_{d|m} d^\alpha, \tag{8.1}$$

where the summation is over all positive integers d that are factors of the positive integer m. The Dirichlet series of this function is related to the Riemann zeta function by

$$\zeta(s-\alpha)=\frac{1}{\zeta(s)}\sum_{m=1}^{\infty}\frac{\sigma_\alpha(m)}{m^s}. \tag{8.2}$$

Let us define H_{σ_α} to be the Hamiltonian that has as its spectrum $\sigma_\alpha(m)$ states of energy $\omega\ \log(m)$. Then

$$\mathrm{tr}[e^{-(\beta-\alpha/\omega)H_B}]=\mathrm{tr}[(-1)^F e^{-\beta(H_{\sigma_\alpha}+H_F)}]. \tag{8.3}$$

The right side involves a partial supersymmetric cancellation (note that none of the degeneracies go to zero) at temperature β^{-1}, resulting in a theory a different temperature.

A second representative case comes from rewriting the identity (7.3) of the preceding section as

$$Z_\varphi(\beta,\omega)=\frac{\zeta(s-1)}{\zeta(s)}=\mathrm{tr}[e^{-(\beta-1/\omega)H_B}]\mathrm{tr}[(-1)^F e^{-\beta H_F}]. \tag{8.4}$$

In this context, the Euler totient function appears as the degeneracy factor resulting after a cancellation between bosons and fermions at different temperatures. This is a variation on the partial supersymmetry used above; here, if the bosons and fermions were at the same temperature, the cancellation would be complete. Mathematically, the right side of (8.4) resembles a partition function for a supersymmetric Hamiltonian with nonsupersymmetric boundary conditions.

Needless to say, these two examples could both be written as identities involving a single temperature but different coupling constants.

The importance of highlighting these identities is to demonstrate what partial supersymmetry can offer. If these and a host of similar identities involving ratios of zeta functions and other Dirichlet series can be understood in this way, partial supersymmetry promises to be a valuable new tool for enriching our understanding of arithmetic number theory.

IX. CONCLUSIONS

Arithmetic quantum theories have provided some of the most aesthetically pleasing examples of connections between mathematics and physics, and in this paper, we have found that these connections are even more compelling and worthy of study.

Most importantly, we have found dualities among the partition functions for arithmetic quantum theories, dualities that can be interpreted as coupling constant or temperature dualities, demonstrating the equivalence of apparently distinct Hamiltonians. Our work has suggested that there may actually be a reason to expect a connection between some types of dualities and the existence of a Hagedorn temperature, an area that warrants further investigation. With these dualities, it is possible to shift the temperature of a certain theory by mixing in additional quantum excitations at the original temperature; relating this interpretation to more familiar methods such as the renormalization groups should also provide useful insights into, or constraints on, the behavior of these theories. We are interested, too, in finding more useful applications of the bosonically graded topological index.

On the technical side, the most significant result of this paper is the identification of the technique of partial supersymmetry. This has given a physical basis for a variety of number theoretic results, and we have identified further types of results that should yield to such an analysis. Indeed, one would expect this method to be useful beyond the realm of arithmetic quantum theories.

This technique of partial supersymmetry has given us a physically based formalism for interpolating from one Dirichlet series to another using ordinary bosons and fermions. Indeed, it is this idea that underlies our formulation of noninteger order parafermions. This is a potentially powerful idea, as it may allow the use of topological insights to study arithmetic number theory, by using continuity arguments to connect the behavior of different Dirichlet series.

It is our hope that the ideas introduced in this paper will be of value not only in understanding arithmetic quantum theories and directly related examples (physical ones such as string theory, mathematical ones such as Dirichlet series), but may also be extended to a wider range of theories. There is no reason to think that the notion of partial supersymmetry or our particular approach to

duality and its implications should not, in some fashion, be extended to relativistic, interacting theories with a variety of spectra (and not simply those theories so closely tied to the Riemann zeta function and other Dirichlet series), nor that we have exhausted the physical insights into number theories that consideration of quantum theories will provide. We hope further work will shed light on these possibilities.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant No. PHY-9509991.

[1] B. Julia, J. Phys. (France) **50**, 1371 (1989); D. Spector, Phys. Lett. A **140**, 311 (1989); B. Julia, ''Statistical theory of numbers,'' in *Number Theory and Physics*, edited by J. M. Luck, P. Moussa, and M. Waldschmidt; *Proceedings in Physics 47* (Springer-Verlag, Berlin, 1990), p. 276; D. Spector, Commun. Math. Phys. **127**, 239 (1990).
[2] I. Bakas and M. Bowick, J. Math. Phys. **32**, 1881 (1991); J.-B. Bost and A. Connes, Selecta Mathematica **1**, 411 (1995); A. Connes, C. R. Acad. Sci., Ser. I: Math. **323**, 1231 (1996); P. Contucci and A. Knauf, J. Math. Phys. **37**, 5458 (1996).
[3] T. M. Apostol, *An Introduction to Analytic Number Theory* (Springer-Verlag, New York, 1976); see also M. Aignier, *Combinatorial Theory* (Springer-Verlag, New York, 1979).
[4] E. Witten, Nucl. Phys. B **202**, 253 (1982).
[5] C. Becchi, A. Rouet, and R. Stora, Phys. Lett. **52B**, 344 (1974); I. V. Tyupin, Lebedev preprint FIAN No. 39 (1975); C. Becchi, A. Rouet, and R. Stora, Ann. Phys. **98**, 287 (1976).
[6] T. Appelquist and J. Carazone, Phys. Rev. D **11**, 2856 (1975).
[7] R. Hagedorn, Nuovo Cimento A **56**, 1027 (1968).

Symmetry and symmetry breaking in quantum-chromo (flavor)-dynamics

A. H. Völkel
Institut für Mathematik I, Freie Universität Berlin, 14195 Berlin, Germany

(Received 22 January 1997; accepted for publication 3 September 1997)

Symmetries of flavored quantum-chromo(flavor)-dynamics (QCD) and their breaking are reinvestigated within the Lagrange–Hamilton frame as well as on the basis of axiomatic gauge quantum field theory aiming at the resolution of some shortcomings by which QCD is still burdened despite all its undisputable successes: The Noether energy momentum tensors arising from those general representations (substitution rules) of the translation group that leave the standard SU(3) local color gauge and global flavor invariant chiral QCD-Lagrangian (and hence the Euler–Lagrange field equations) form invariant are shown to contain, besides the standard color and flavor symmetric canonical part, additional terms breaking the internal SU(3) symmetries of the Lagrangian down to chiral SU(2)×U(1) flavor symmetry, and in case of color confinement also the color group down to U(1)×SU(2) color symmetry only. With these general representations (covariance conditions) of the translation group as an input, a strictly Poincaré covariant gauge-quantum-field theory with color confinement is formulated and flavor Goldstone states are shown to occur with *finite masses* due to relative Einstein causality, the spectrum condition in color neutral sectors, translational covariance, and current conservation. Identifying the finite mass Goldstone states with pseudoscalar mesons, the axial vector part of the chiral SU(3) flavor symmetry is spontaneously broken, leaving only the chiral SU(2)×U(1) subgroup as a strict flavor symmetry of QCD. © *1998 American Institute of Physics.* [S0022-2488(98)02203-8]

I. INTRODUCTION

In spite of its indisputable success all unitary symmetry schemes from SU(3) to SU(?) of flavored quantum-chromo(flavor)-dynamics (QCD) showed right from the beginning some deficiencies leaving the state of affairs a little bit artificially flavored. Some authors of textbooks state it this way: ''Put more concisely, the theory would be more convincing if the symmetry breaking was 'intrinsic' to the scheme, in the sense that it emerges automatically, without it having to be put in 'by hand' '' [1] (Part C, Chap. 18). ''Considering the fact that m_π is much smaller than the mass differences $m_K - m_\pi$ and $m_\eta - m_\pi$, it seems quite strange that H_{asym} can be treated as a perturbation'' [2] (Chap. 12).

In the hierarchy of flavor interactions, the strongest symmetry conserving part of the Hamiltonian is invariant under all SU(3) transformations, whereas the ''medium strong'' one is assumed to break the symmetry and the ''rest symmetry group'' is SU(2)×U(1), with SU(2) being the isospin group. In view of the rather large mass splittings between the isotopic submultiplets within a SU(3) multiplet, it is not possible to assert the effect of the ''medium strong'' interaction to be a small one. But then the question arises at once as to the significance of the SU(3) [respectively, SU(?)] group. Is its role merely to specify rather roughly the multiplets or is there something more in the unitary symmetry scheme still hidden in the underground?

Indeed, right at the outset the unitary scheme met the flaw, that in contrast to the baryon multiplets the masses in the splitting relations for the pseudoscalar meson octet had—again by hand—to be replaced by their squares in order to fit the experimental data[1–7] to an acceptable accuracy. Critically seen, however, even then three pion masses squared are still missing. Concerning the higher multiplets when grouped into octets in such a way that the mass differences are all small compared to a medium mass value of the multiplet, the linear mass formulas fit the experimental values, at least with the same accuracy as the quadratic ones and apart from the

 © 1998 American Institute of Physics

lowest lying pseudovector octet 1^{--} ($\omega-\phi$ mixing) with a variation of less than 2.5%. Moreover, the chiral flavor symmetry meets the further flaw that the conservation of its axial vector current in combination with either parity conservation in a strong interaction[2] (Chap. 24.1–3) or the finite meson-decay constants in leptonic weak decays[4] (Chap. 2.8) dictates in both cases zero masses for the pseudoscalar mesons, which on the one hand seems compatible with the role of Goldstone states for the spontaneous breaking of the chiral symmetry ascribed to them but on the other hand contradicts the finiteness of their experimental masses. Hence the pseudoscalar meson octet seems to play an exceptional role for ''some unknown weaker mechanism,'' [5] breaking the unitary flavor symmetry of the strong interaction, which still hides itself behind all plausible explanations.

In Sec. II the conserved energy–momentum tensors and hence the Hamilton densities are derived that arise via Noether's Theorem from the form invariance of the usual flavored QCD-Lagrangian with respect to space–time translations, if, besides the familiar standard representation of the translation group for the basic quark and gluon fields, more general one-dimensional ones affiliated to the flavor and color degrees of freedom are admitted. The result is a continuum of admitted energy–momentum tensors parametrized by five arbitrary four-vectors, one of which is affiliated to the color degrees of freedom and the remaining four to the flavor ones. The Hamilton density corresponding to each of them except the standard one is a sum of the standard canonical one, the conserved Noether current generating the color charge, and last but not least the ''medium strong'' and unitary symmetry breaking conserved flavor current, usually added ''by hand.'' On the basis of these Noether currents the symmetry breaking is reconsidered in Sec. III. The global SU(3) flavor symmetry is, on the algebraic level, seen to be broken in the sense that the flavored CD-Lagrangian is form invariant with respect to the whole SU(3) group of flavor transformations, but depending on the representation chosen for the translation group its infinitesimal generators, in general, do not commute with the Hamilton density, that is with the generator of the time translation. However, for a subclass of representations of the translation group selected from the continuum by taking one of the four flavor vectors to be a fixed multiple of a certain other one of them, the global flavor $SU(2)\times U(1)$ subgroup remains a strict symmetry with its infinitesimal generators commuting with those of the space–time translation.

On the other hand, it is proved in Sec. VI that in the representations of the translation group with nonzero flavor (and/or color) parameters, Goldstone states occur with *nonzero* masses. Taking the pseudoscalar mesons with their finite experimental masses as Goldstone particles affiliated to the conserved axial vector currents of the chiral SU(3) flavor group, then the flavor parameters of representations of the translation group and hence of Hamiltonians admitted by the chiral QCD-Lagrangian are restricted to certain Lorentz orbits within the subclass stated above, in which the chiral $SU(2)\times U(1)$ flavor group remains a strict symmetry group commuting with the space–time translations.

In Sec. IV a Poincaré covariant general gauge quantum field theory of flavored chromodynamics with color-charge confinement, which mirrors the basic structures (especially of the color confining and internal symmetry breaking representations of the subgroup of translation) obtained in Secs. II and III within the Lagrange–Hamilton frame, is formulated in a Hilbert space equipped with an additional indefinite metric. Poincaré invariance requires that this indefinite metric space is the direct integral of such spaces attached to every point of a Lorentz orbit in the space of the flavor parameters and the color vector describing color confinement. The direct integral is taken with respect to the invariant measure of these orbits.

In Sec. V the spectrum condition is formulated and the construction of one particle state is sketched. By means of them and the direct integral structure of the underlying indefinite metric space, it is rather obvious that the finite meson-decay constants together with the conservation of the axial vector current do not imply anymore that the masses of the pseudoscalar mesons vanish.

In Sec. VI, finally, the generalized Goldstone Theorems are proved by means of relative locality, and it is shown that the conserved vector currents of the SU(3) flavor group generate a representation of this group that is isometric with respect to the indefinite metric. Hence, it lifts to a unitary one in the physical Hilbert space obtained in the familiar way via factorization and completion of some subspaces. But only its restriction to the subgroup $SU(2)\times U(1)$ commutes with the translation operators. The corresponding chiral flavor group, however, is spontaneously broken, since some axial charges do not exist due to the presence of the pseudoscalar K mesons with finite masses.

II. NOETHER CURRENTS OF LCD

The Lagrangian of (quantum-)chromodynamics is built up from two SU(3)-color triplets, respectively, SU(n)-flavor multiplets ($n\geqslant 3$) $\varphi^{3,n}_{a,f}(x)$ and $\varphi^{3^*,n^*}_{a,f}(x)$ of Dirac spinors called quark, respectively, antiquark fields and a SU(3)-color octet (flavor singlet) $\varphi^{8,1}_{a}(x)$ of Lorentz vectors called gluon (gauge) fields.[1–4] Here we will restrict the flavor group $G(3)_{\mathscr{F}}$ to be SU(3), respectively, some subgroup of it [isomorphic to SU(2)×U(1) or U(1)×U(1)], which means consider a system with three flavors (up, down, strange). However, all considerations may easily be extended to a flavor group of higher degree in order to include charmed, bottom, top, and so on quarks. The two ''different'' SU(3) groups will be marked by a suffix $\mathscr{C}$ for color, respectively, $\mathscr{F}$ for flavor whenever it seems necessary to avoid confusion.

For the sake of notational shortness, the unitary irreducible matrix representations $D^{(p,q)}$, $(p,q)\in\mathbb{N}_0\times\mathbb{N}_0$ of SU(3) with dimension $d=1/2[(1+p)^2(1+q)+(1+p)(1+q)^2]$ will be denoted by $R^d=D^{(p,q)}$ for $p\geqslant q$, respectively, $R^{d^*}=D^{(p,q)}$ for $p\leqslant q$. Then *conventionally* for any triple (α,u,v) from the direct product $\mathrm{SL}(2,\mathbb{C})\times\mathrm{SU}(3)_{\mathscr{C}}\times G(3)_{\mathscr{F}}$ the fields are supposed to transform according to

$$\varphi^{m,n}_{\mu,a,f}(x)\to(U(\alpha,u,v)\varphi)^{m,n}_{\mu,a,f}(x):=S^{(m,n)}(\alpha)^{\nu}_{\mu}R^{m}_{\mathscr{C}}(u^{-1})^{b}_{a}R^{n}_{\mathscr{F}}(v^{-1})^{g}_{f}\varphi^{m,n}_{\nu,b,g}(\Lambda(\alpha)x), \tag{1}$$

with

$$S^{(m,n)}(\alpha)^{\mu}_{\nu}:=\begin{cases} S^{1/2}(\alpha^{-1})_{\nu\mu}, & \text{for } (m,n)=(3,3),\\ \Lambda(\alpha^{-1})^{\mu}_{\nu}, & \text{for } (m,n)=(8,1),\\ S^{1/2}(\alpha)_{\mu\nu}, & \text{for } (m,n)=(3^*,3^*),\end{cases} \tag{2}$$

and

$$S^{1/2}(\alpha)^{-1}\gamma^{\mu}S^{1/2}(\alpha)=\Lambda(\alpha)^{\mu}_{\nu}\gamma^{\nu};\quad \gamma^0S^{1/2}(\alpha)^{+}\gamma^0=S^{1/2}(\alpha)^{-1};$$

$$\Lambda(\alpha)^{\mu}_{\nu}g_{\mu\sigma}\Lambda(\alpha)^{\sigma}_{\kappa}=g_{\nu\kappa}. \tag{3}$$

The Dirac matrices are assumed to have the hermiticity properties $(\gamma^{\mu})^{+}=(\overline{\gamma^{\mu}})^{t}=g^{\mu\mu}\gamma^{\mu}$, $(\gamma^0)^t=\gamma^0$. With respect to the group SU(3) and its representations, we adopt the conventions and (apart from minor changes) the notations of de Swart.[5–7] The color/flavor indices are triples $a=(I,I^3,Y)$ running over the index set Γ_d defined by the Okubo rule:[6]

$$\Gamma_d(p,q):=\{(I,I^3,Y)\,|\,I=1/2(\lambda-\rho);\quad I^3=-I,\ldots,I;\quad Y=\lambda+\rho-1/3(2p+4q);\quad \lambda,\rho\in\mathbb{N}_0;$$

$$0\leqslant\rho\leqslant q\}. \tag{4}$$

$I\ldots$ is interpreted to be the total ''color/flavor isospin,'' $I^3\ldots$ its third component, and $Y\ldots$ the ''color/flavor hypercharge.'' An immediate consequence of the definition of Γ_d is

$$(I,I^3,Y)\in\Gamma_d(\Gamma_{d^*})\Leftrightarrow(I,-I^3,-Y)\in\Gamma_{d^*}(\Gamma_d). \tag{5}$$

In accordance with this relation the negative of an index triple $a=(I,I^3,Y)$ is by definition the triple $-a:=(I,-I^3,-Y)$. Hence the ''color/flavor charge,''

$$Q(a)\ldots=Q(I,I^3,Y)\ldots:=I^3\ldots+\frac{1}{2}\,Y\ldots, \tag{6}$$

satisfies $Q(-a)\ldots=-Q(a)\ldots$. Finally, there exists an important connection between the complex conjugate representation $\overline{R^d}(\overline{R^{d^*}})$ and the representation $R^{d^*}(R^d)$. Adopting the notational convention $d^{**}=(d^*)^*=d$ it reads for all d and d^*,

$$R^{d^*}(u)^{b}_{a}=(-1)^{Q(b)-Q(a)}\overline{R^d(u)}^{-b}_{-a};\quad a,b\in\Gamma_{d^*},\quad u\in\mathrm{SU}(3). \tag{7}$$

In case of the adjoint (octet) representation as well as all other (''diagonal'') representations with $R^d = D^{(p,p)} = R^{d*}$ the condition (7) is an equation between different elements of the same matrix, which is called *self-conjugateness*. As an immediate consequence the gluon fields form a self-conjugate color octet:

$$\overline{\varphi^{8,1}_{\mu,a}(x)} = (-1)^{Q(a)_{\mathscr{C}}} \varphi^{8,1}_{\mu,-a}(x). \tag{8}$$

Dirac indices will often be dropped by the usual conventions and $\varphi^{3,3}_{a,f}$ is considered to be a column, whereas $\varphi^{3^*,3^*}_{a,f}$ a row. Then the Dirac adjoint of $\varphi^{3,3}_{a,f}$ is connected to $\varphi^{3^*,3^*}_{a,f}$ by

$$\widetilde{\varphi}^{3,3}_{a,f}(x) := \varphi^{3,3}_{a,f}(x)^+ \gamma^0 = (-1)^{-Q(a)_{\mathscr{C}} - Q(f)_{\mathscr{F}} - 2/3} \varphi^{3^*,3^*}_{-a,-f}(x). \tag{9}$$

In terms of these fields the Lagrangian of chromodynamics reads as

$$\begin{aligned} L(\varphi(x), B(x)) = &-1/4(-1)^{Q(a)_{\mathscr{C}}} G_{\mu\nu,-a}(x) G^{\mu\nu,a}(x) + \widetilde{\varphi}^{3,3}_{a,f}(x)[(i/2)\gamma^\mu \{\delta^b_a \overleftrightarrow{\partial}_\mu + ig(\tau_c)^b_a \varphi^{8,1}_{\mu,c}(x)\} \\ &- \delta^b_a \varrho_f] \delta^e_f \varphi^{3,3}_{b,e}(x) - (-1)^{Q(a)_{\mathscr{C}}} [\varphi^{8,1}_{\mu,-a}(x) \partial^\mu B^{8,1}_a(x) - (\alpha/2) B^{8,1}_{-a}(x) B^{8,1}_a(x)], \end{aligned} \tag{10}$$

with the gluon field tensor

$$G_{\mu\nu,a}(x) := \partial_\mu \varphi^{8,1}_{\nu,a}(x) - \partial_\nu \varphi^{8,1}_{\mu,a}(x) + ig\sqrt{3} \begin{pmatrix} 8 & 8 & 8_A \\ b & c & a \end{pmatrix} \varphi^{8,1}_{\mu,b}(x) \varphi^{8,1}_{\nu,c}(x), \tag{11}$$

being obviously a self-conjugate color octet (flavor singlet).

τ_a, $a \in \Gamma_8$ denote the traceless (3,3) matrices of the self-conjugate fundamental representation of the (complex) Lie algebra su(3) satisfying the relations[5–7]

$$[\tau_a, \tau_b]_- = -2\sqrt{3} \begin{pmatrix} 8 & 8 & 8_A \\ a & b & c \end{pmatrix} \tau_c, \quad \mathrm{tr}(\tau_a \cdot \tau_b^+) = 2\delta_{ab}, \quad \tau_a^+ = (-1)^{Q(a)} \tau_{-a}. \tag{12}$$

The gauge fixing field $B^{8,1}_a$ is related to the gluon octet via the Euler–Lagrange field equation $\partial^\nu \varphi^{8,1}_{\nu,a}(x) + \alpha B^{8,1}_a(x) = 0$. Usually, there is added a further term to the Lagrangian describing FP ghosts.[8] In order to keep the explicit formulas as short as possible, we dropped it, since it would only add some further terms to the explicit expression of the Noether currents in terms of the basic fields without being in any way tangent to the general structure of these currents that we are going to use in our considerations. The members of the set $\{\varphi^{3,3}_{\mu,a,f}, \widetilde{\varphi}^{3,3}_{\mu,a,f}, \varphi^{8,1}_{k,b}, B^{8,1}_b | \mu = 0,...,3; k = 1,2,3; a,f \in \Gamma_3; b \in \Gamma_8\}$ are taken as an independent canonical variable.

By construction, the Lagrangian is form invariant under the global color group $\mathrm{SU}(3)_{\mathscr{C}}$ and the corresponding eight conserved Noether currents read ($c \in \Gamma_8$) as

$$\begin{aligned} j^{8,\mu}_c(x)_{\mathscr{C}} := &-\frac{1}{2} \left\{ \widetilde{\varphi}^{3,3}_{a,f}(x) \gamma^\mu (\tau_c)^b_a \varphi^{3,3}_{b,f}(x) + i2\sqrt{3}(-1)^{Q(a)_{\mathscr{C}}} \begin{pmatrix} 8 & 8 & 8_A \\ c & b & a \end{pmatrix} [G^{\mu\epsilon}_{-a}(x) \varphi^{8,1}_{\epsilon,b}(x) \right. \\ &\left. + \varphi^{8,1}_{\lambda,-a}(x) g^{\mu\lambda} B^{8,1}_b(x)] \right\} \\ = &\frac{1}{g} (-1)^{Q(a)_{\mathscr{C}}} \left\{ \delta^a_c \partial_\lambda G^{\mu\lambda}_{-a}(x) + g^{\mu\lambda} \left[\delta^a_c \partial_\lambda + ig\sqrt{3} \begin{pmatrix} 8 & 8 & 8_A \\ c & d & a \end{pmatrix} \varphi^{8,1}_{\lambda,d}(x) \right] B^{8,1}_{-a}(x) \right\}. \end{aligned} \tag{13}$$

In the last row the Euler–Lagrange equations have been used.

Here we are especially interested in the combination $j^{8,\mu}_{(1,0,0)}(x) + (1/\sqrt{3}) j^{8,\mu}_{(0,0,0)}(x)$ because it will reoccur in the Noether energy momentum tensor below. From the explicit connection between the real self-conjugate traceless matrices τ_c, $c \in \Gamma_8$ and the Hermitian traceless matrices λ_j, $j = 1,...,8$ of Gell-Mann,

$$\tau_{(0,0,0)}=\lambda_8, \quad \tau_{(1,0,0)}=\lambda_3, \quad \tau_{(1,1,0)}=-\frac{1}{\sqrt{2}}(\lambda_1+i\lambda_2)$$

$$\tau(1/2,1/2,1)=-\frac{1}{\sqrt{2}}(\lambda_4+i\lambda_5), \quad \tau_{(1/2,-1/2,1)}=-\frac{1}{\sqrt{2}}(\lambda_6+i\lambda_7),$$

respectively, multiplying the commutation relations (12) by $(\tau_e)^+$ and taking the trace it follows at once:

$$\left(\tau_{(1,0,0)}+\frac{1}{\sqrt{3}}\tau_{(0,0,0)}\right)^b_a=2Q(a)\delta^b_a,$$

$$2\sqrt{3}\left[\begin{pmatrix} 8 & 8 & 8_A \\ (1,0,0) & b & a \end{pmatrix}+\frac{1}{\sqrt{3}}\begin{pmatrix} 8 & 8 & 8_A \\ (0,0,0) & b & a \end{pmatrix}\right]=2Q(a)\delta_{ba},$$

and therefore

$$j^{8,\mu}_{(1,0,0)}(x)_{\mathscr{C}}+\frac{1}{\sqrt{3}}j^{8,\mu}_{(0,0,0)}(x)_{\mathscr{C}}=-\{Q(a)_{\mathscr{C}}\tilde{\varphi}^{3,3}_{a,f}(x)\gamma^\mu\varphi^{3,3}_{a,f}(x)+i(-1)^{Q(a)_{\mathscr{C}}}Q(a)_{\mathscr{C}}[G^{\mu,\epsilon}_{-a}(x)\varphi^{8,1}_{\epsilon,a}(x)$$
$$+g^{\mu\epsilon}\varphi^{8,1}_{\epsilon,-a}(x)B^{8,1}_a(x)]\}. \tag{14}$$

Concerning the flavor transformations the subgroup of SU(3) with respect to which the CD Lagrangian is form invariant, is determined alone by the ''mass'' matrix of the quarks,

$$\varrho=\begin{pmatrix} \varrho_u & 0 & 0 \\ 0 & \varrho_d & 0 \\ 0 & 0 & \varrho_s \end{pmatrix}=\begin{pmatrix} \varrho_{(1/2,1/2,1/3)} & 0 & 0 \\ 0 & \varrho_{(1/2,-1/2,1/3)} & 0 \\ 0 & 0 & \varrho_{(0,0,-2/3)} \end{pmatrix},$$

which is not observable, but can only be estimated indirectly. If all three quark ''masses'' are equal $G(3)_{\mathscr{F}}$ is of course again SU(3). If $\varrho_u=\varrho_d\neq\varrho_s$ form invariance holds only for the subgroup

$$G(3)_{\mathscr{F}}=\left\{\begin{pmatrix} & & 0 \\ & g & \\ & & 0 \\ 0 & 0 & 1 \end{pmatrix}\exp\left[i\frac{t}{2\sqrt{3}}\tau_{(0,0,0)}\right]\middle| g\in \mathrm{SU}(2)\wedge t\in\mathbb{R}\right\}$$

isomorphic to SU(2)×U(1). Finally, if all three ''masses'' are pairwise different their remains only the subgroup,

$$G(3)_{\mathscr{F}}=\left\{\exp i/2\left(\tau_{(1,0,0)}\rho+\frac{1}{\sqrt{3}}\tau_{(0,0,0)}t\right)\middle|\rho,t\in\mathbb{R}\right\},$$

isomorphic to U(1)×U(1).

Corresponding to these three cases the Noether Theorem delivers the existence of eight, four, respectively, two locally conserved flavor currents:

$$j^{8,\mu}_c(x)_{\mathscr{F}}=-1/2\{\tilde{\varphi}^{3,3}_{a,f}(x)\gamma^\mu(\tau_c)^e_f\varphi^{3,3}_{a,e}(x)\}, \tag{15}$$

with

$$c \in \cdot \begin{cases} \Gamma_8, & \text{if} \quad G(3)_{\mathscr{F}} = \mathrm{SU}(3), \\ \{(0,0,0),(1,0,0),(1,1,0),(1,-1,0)\}, & \text{if} \quad G(3)_{\mathscr{F}} \sim \mathrm{SU}(2)\times\mathrm{U}(1), \\ \{(0,0,0),(1,0,0)\}, & \text{if} \quad G(3)_{\mathscr{F}} \sim \mathrm{U}(1)\times\mathrm{U}(1). \end{cases}$$

It is generally accepted[1–4] that $\varrho_d \approx \varrho_u \approx 0$ and $\varrho_u + 100\ \mathrm{MeV} \leqslant \varrho_s \leqslant \varrho_u + 200\ \mathrm{MeV}$, which reflects the mass splitting in the hadron spectrum with the splitting within an isospin submultiplit due to electromagnetism. In the last two cases, where the Noether theorem delivers only four or two instead of eight conserved currents, one can define four, respectively, six further currents—the missing ones with respect to the full SU(3) symmetry—by the right-hand side of Eq. (18). Of course they are not conserved. Important for the following is that the two currents $j^{8,\mu}_{(1,0,0)}(x)_{\mathscr{F}}$ and $j^{8,\mu}_{(0,0,0)}(x)_{\mathscr{F}}$ that will reoccur in the Noether energy–momentum tensor are locally conserved in all three cases above.

Besides the irreducible octet operator formed by the eight generators τ_a, $a \in \Gamma_8$, there exists a second linear independent one:[7]

$$\sigma_a : 2\sqrt{5}/3 \begin{pmatrix} 8 & 8 & 8_S \\ b & c & a \end{pmatrix} \tau_b \tau_c .$$

In terms of the total isospin operator $\mathscr{T}^2$, the hypercharge $\mathscr{Y}$, and the second-order Casimir operator $\mathscr{C}^2$ [with eigenvalues $I(a)(I(a)+1)$, $Y(a)$, and $K(d)$, respectively] its (0,0,0) and (1,0,0) components read explicitly as

$$\begin{aligned} \sigma_{(0,0,0)} &= 2[\mathscr{T}^2 - 1/4\mathscr{Y}^2 - 1/3\mathscr{C}^2], \\ \sigma_{(1,0,0)} &= 10/(3\sqrt{3})\,\tau_{(1,0,0)}. \end{aligned} \tag{16}$$

Unless the quark ''masses'' are pairwise different, the Lagrangian is form invariant with respect to the substitutions

$$\begin{aligned} \varphi^{3,3}_{a,f}(x) &\to \varphi^{3,3\wedge}_{a,f}(x) = (\exp\, i/2\sigma_{(r,0,0)}t)^e_f \varphi^{3,3}_{a,e}(x), \\ \widetilde{\varphi}^{3,3}_{a,f}(x) &\to \widetilde{\varphi}^{3,3\wedge}_{a,f}(x) = (\exp - i/2\sigma_{(r,0,0)}t)^e_f \widetilde{\varphi}^{3,3}_{a,e}(x), \end{aligned} \tag{17}$$

with $r = 0,1$ and the gluon, respectively, gauge fields unchanged. But then the Noether Theorem delivers two further conserved flavor currents:

$$j^{8_S,\mu}_{(r,0,0)}(x)_{\mathscr{F}} = -\tfrac{1}{2}\widetilde{\varphi}^{3,3}_{a,f}(x)\,\gamma^\mu(\sigma_{(r,0,0)})^e_f \varphi^{3,3}_{a,e}(x), \tag{18}$$

with $r = 0,1$, which also are going to reoccur in the Noether energy momentum tensor.

Next we consider the translation group $(\mathbb{R}_4, +)$. For any fixed complex four-vector $\delta \in \mathbb{C}_4$ and any fixed real four-vectors (κ, ω, M, N), a representation $T^{\delta,M,N}_{\kappa,w}$ of the translation group may be defined by

$$\begin{aligned} (T^{\delta,M,N}_{\kappa,\omega}(y)\varphi)^{n,d}_{\mu,a,f}(x) :&= \exp[Q(a)_{\mathscr{C}}(\delta\cdot y)]R^d(\exp\, i/2[\sigma_{(1,0,0)}(N\cdot y) \\ &\quad + \sigma_{(0,0,0)}(M\cdot y)])^b_f R^d \\ &\quad \times (\exp\, i/2[\tau_{(1,0,0)}(\kappa\cdot y) + (1/\sqrt{3})\,\tau_{(0,0,0)}(\omega\cdot y)])^e_b \varphi^{n,d}_{\mu,a,e}(x+y), \end{aligned} \tag{19}$$

with $(n,d) \in \{(3,3),(3^*,3^*),(8,1)\}$, and $B^{8,1}_a$ transforming in the same way as the gluon field $\varphi^{8,1}_{\ldots,a}$ (i.e., color octet and flavor singlet). By means of the equation (7) and (9), it follows for the Dirac adjoint spinor $\widetilde{\varphi}^{3,3}$,

$$(T^{\delta,M,N}_{\kappa,\omega}(y)\widetilde{\varphi})^{3,3}_{a,f}(x)=(-1)^{-Q(a)_{\mathscr{C}}-Q(f)_{\mathscr{F}}-2/3}(T^{\delta,M,N}_{\kappa,\omega}(y)\varphi)^{3^*,3^*}_{-a,-f}(x)$$
$$=\exp[-Q(a)_{\mathscr{C}}(\delta\cdot y)](\exp(-i/2)[\sigma_{(1,0,0)}(N\cdot y)+\sigma_{(0,0,0)}(M\cdot y)])^b_f$$
$$\times(\exp(-i/2)[\tau_{(1,0,0)}(\kappa\cdot y)+(1/\sqrt{3})\tau_{(0,0,0)}(\omega\cdot y)])^e_b\widetilde{\varphi}^{3,3}_{a,e}(x+y). \quad (20)$$

The equations (19) are nothing else than the definition of a set of operations $\{T^{\cdots}_{\cdots}(y)\,|\,y\in\mathbb{R}_4\}$ satisfying the (Abelian) composition law $T^{\cdots}_{\cdots}(x)\circ T^{\cdots}_{\cdots}(y)=T^{\cdots}_{\cdots}(x+y)$ of the translation group and therefore generating a representation of it. (For $M=N=m=\kappa=\omega=0$ it is just the standard ''transformation law'' of fields under translations well known from any textbook.) The essential point comes now.

The *CD-Lagrangian as given by the equations (10)–(11) is form invariant; that means* $L(\hat{\varphi}(\hat{x}),\hat{B}(\hat{x}))=L(\varphi(x),B(x))$, *under the following group of substitutions:*

$$x\to\hat{x}=x+y,\quad y\in\mathbb{R}_4,$$
$$\varphi^{n,d}_{\mu,a,f}(x)\to\hat{\varphi}^{n,d}_{\mu,a,f}(\hat{x}):=(T^{\delta,M}_{\kappa,\omega}(-y)\varphi)^{n,d}_{\mu,a,f}(x+y),(n,d)\in\{(3,3),(3^*,3^*),(8,1)\}, \quad (21)$$
$$B^{8,1}_a(x)\to\hat{B}^{8,1}_a(\hat{x}):=(T^{\delta,M}_{\kappa,\omega}(-y)B)^{8,1}_a(x+y).$$

The proof for the flavor part is obvious since the Lagrangian is diagonal in the flavor indices. The proof of the color part is given in the appendix of Ref. 9 and consists mainly in showing that $(\tau_c)^b_a=0$ unless $Q(a)=Q(b)+Q(c)$.

Due to the form invariance of the Lagrangian under the (four-parameter) group of translations above Noether's Theorem[10] dictates the existence of four locally conserved currents $\Theta^{\delta,M,N}_{\kappa,\omega}(x)^\mu_\nu[\nu=0,1,2,3,\partial_\mu\Theta^{\delta,M,N}_{\kappa,\omega}(x)^\mu_\nu=0]$, which are considered as the components of the energy–momentum density tensor, provided they possess the necessary reality properties of an observable. In the case at hand they read explicitly as

$$\Theta^{\delta,M,N}_{\kappa,\omega}(x)^\mu_\nu=\Theta(x)^\mu_\nu+i\,\delta_\nu Q(a)_{\mathscr{C}}\widetilde{\varphi}^{3,3}_{a,f}(x)\,\gamma^\mu\varphi^{3,3}_{a,f}(x)-\frac{1}{2}\,\widetilde{\varphi}^{3,3}_{a,f}(x)\,\gamma^\mu\Big[\kappa_\nu(\tau_{(1,0,0)})^e_f$$
$$+\frac{1}{\sqrt{3}}\,\omega_\nu(\tau_{(0,0,0)})^e_f\Big]\varphi^{3,3}_{a,e}(x)-\frac{1}{2}\,\widetilde{\varphi}^{3,3}_{a,f}(x)\,\gamma^\mu(N_\nu\sigma_{(1,0,0)}+M_\nu\sigma_{(0,0,0)})^e_f\varphi^{3,3}_{a,e}(x)$$
$$-Q(a)_{\mathscr{C}}(-1)^{Q(a)_{\mathscr{C}}}[G^{\mu\epsilon}_{-a}(x)\varphi^{8,1}_{\epsilon,a}(x)+g^{\mu\epsilon}\varphi^{8,1}_{\epsilon,-a}(x)B^{8,1}_a(x)]\delta_\nu. \quad (22)$$

Here $\Theta(x)^\mu_\nu=\Theta^{0,0,0}_{0,0}(x)^\mu_\nu$ is the (standard) *canonical* energy–momentum density tensor,

$$\Theta(x)^\mu_\nu=-\delta^\mu_\nu L(\varphi(x),B(x))+\frac{i}{2}\,\widetilde{\varphi}^{3,3}_{a,f}(x)\,\gamma^\mu\overleftrightarrow{\partial}_\nu\varphi^{3,3}_{a,f}(x)-(-1)^{Q(a)}[G^{\mu\epsilon}_{-a}(x)\partial_\nu\varphi^{8,1}_{\epsilon,a}(x)$$
$$+g^{\mu\epsilon}\varphi^{8,1}_{\epsilon,-a}(x)\partial_\nu B^{8,1}_a(x)].$$

Especially $\Theta(x)^0_0$ is just the *canonical* Hamiltonian density.

Comparing the right-hand side of the equation (22) with the equations (14) and (15), we see that due to the relation (14) the terms occurring in addition to the canonical part are just special components of the locally conserved flavor, respectively, color currents. Hence the final result for the locally conserved Noether tensor obtained from translational invariance reads as

$$\Theta^{\delta,M,N}_{\kappa,\omega}(x)^\mu_\nu=\Theta(x)^\mu_\nu-i\,\delta_\nu\Big[j^{8,\mu}_{(1,0,0)}(x)_{\mathscr{C}}+\frac{1}{\sqrt{3}}\,j^{8,\mu}_{(0,0,0)}(x)_{\mathscr{C}}\Big]+\kappa_\nu j^{8,\mu}_{(1,0,0)}(x)_{\mathscr{F}}+\frac{1}{\sqrt{3}}\,\omega_\nu j^{8,\mu}_{(0,0,0)}(x)_{\mathscr{F}}$$
$$+N_\nu j^{8_S,\mu}_{(1,0,0)}(x)_{\mathscr{F}}+M_\nu j^{8_S,\mu}_{(0,0,0)}(x)_{\mathscr{F}}, \quad (23)$$

with $\delta\in\mathbb{C}_4$ and $M,N,\kappa,\omega\in\mathbb{R}_4$.

All further considerations and results will be based on this formula without any reference to the dependence of the currents on the basic quark and gluon fields with the exception of their properties under translations originating from them. They follow directly from the equations (15), (16), and (18)–(21) for $\rho=8,8_s$:

$$(T^{\delta,M,N}_{\kappa,\omega}(y)j)^{\rho,\mu}_a(x)_{\mathscr{F}}=R^8(\exp\cdot[i/2(y\cdot\wp)])^b_a j^{\rho,\mu}_b(x+y)_{\mathscr{F}},$$

$$\wp:=\tau_{(1,0,0)}\kappa+1/\sqrt{3}\,\tau_{(0,0,0)}\omega+\sigma_{(1,0,0)}N+\sigma_{(0,0,0)}M, \tag{24}$$

which in turn leads to

$$(T^{\delta,M,N}_{\kappa,\omega}(y)j)^{\rho,\mu}_a(x)_{\mathscr{F}}=\exp\cdot[i(p(8,a)\cdot y)]j^{\rho,\mu}_a(x+y)_{\mathscr{F}},$$

$$p(8,a)=\Sigma_{b\in\Gamma_3}\Sigma_{c\in\Gamma_{3^*}}\begin{pmatrix}3 & 3^* & 8\\ b & c & a\end{pmatrix}^2(p(3,b)-p(3,-c)),$$

$$p(d,a)_\mu:=-i\partial_\mu R^d(\exp[i/2(\wp\cdot y)])^a_{a_{|y=0}}. \tag{25}$$

In the remainder of this section the dependence of the fields, respectively, currents on the parameters $(\kappa,\omega,\delta,M,N)$ is not written out explicitly, but should be kept in mind (especially in connection with relativistic covariance). Its consequences for the total action, the Euler–Lagrange field equations, the Noether currents, and their total charges are considered at the end of Sec. III after the region over which these variables vary have been determined by the Goldstone theorem.

In classical field theories as well as in quantized ones using a noncovariant gauge (Coulomb gauge, for instance), which admits a (positive definite) Hilbert space $(\mathscr{H},(\cdot|\cdot))$ spanned by physical states the components $\Theta^{\delta,M,N}_{\kappa,\omega}(x)^0_\nu$, $\nu=0,...,3$ are interpreted as the energy momentum density. Hence $\Theta^{\delta,M,N}_{\kappa,\omega}(x)^0_\nu$ has to be a real, respectively, Hermitian operator. Therefore κ, ω, M, N have to be real and δ_ν purely imaginary. In other words, all representations of the translation group that do not meet these conditions would be excluded via physical interpretation. Besides losing explicit Lorentz covariance and Einstein causality (locality) in the intermediate steps within these gauges, the fact that only color singlets are realized in nature and hence admitted in the physical Hilbert space has to be put in as an additional condition ''by hand.''

The situation is different (for the color coupling δ_ν) in a covariant gauge of the Gupta–Bleuler type and especially in the case of QCD at hand with colored basic fields and only color singlets admitted to the physical Hilbert space. In such a gauge,[11–22] the physical Hilbert space $(\mathscr{H}_{\text{phys}},\langle\cdot,\cdot\rangle)$ is detached from a Hilbert space $(\mathscr{H},(\cdot|\cdot))$ in which the basic fields, symmetry transformations, currents, etc. act as operators. It is obtained via factorization H/H_0 of two subspaces $H_0\subset H\subset\mathscr{H}$ generated by an indefinite metric $\langle\cdot,\cdot\rangle=(\cdot|\eta\cdot)$ on $\mathscr{H}$ that is positive semidefinite on H and zero on $\mathscr{H}_0\supset H_0=\mathscr{H}_0\cap H$. Moreover, the scalar product on $\mathscr{H}_{\text{phys}}=\overline{H/H_0}$ is again given by the metric $\langle\cdot,\cdot\rangle$ and, of course, is positive definite on this space. The local gauge invariance of the theory is reflected in the requirement[13,21]

$$\langle\Psi,(j^{8,\mu}_a(x)_{\mathscr{C}}-\partial_\nu G^{\mu\nu}_a(x))\Phi\rangle=0, \tag{26}$$

for all Φ from a dense subspace of H and all $\Psi\in H$.

This is the situation in case $\delta=0$. For $\delta\neq 0$, Poincaré covariance is again lost for the intermediate steps in the unphysical Hilbert space $(\mathscr{H},(\cdot|\cdot))$, but invariance is restored via color confinement in the physical Hilbert space $\mathscr{H}_{\text{phys}}$, provided δ is real.[9,11]

On the other hand, the generally accepted physical picture behind the confinement mechanism, namely, that the forces between the quarks grow in strength with the distance between them[1–4] demands the translation group to be represented by nonisometric operators in the nonphysical Hilbert space $(\mathscr{H},(\cdot|\cdot))$, because otherwise there would arise a contradiction to the cluster properties of the Wightman functions.[22] However, it is well known[23] (Theorem 4.2) that in gauge quantum field theories with the vacuum being an eigenstate of the metric operator an

η-isometric representation of the translation group is nonunitary in $(\mathscr{H},(\cdot|\cdot))$ if and only if Re $\delta\neq 0$. (We note in passing that the vacuum being an eigenstate of the metric operator is only a sufficient condition, but *not* a necessary one).

As will be shown in Sec. IV in order to achieve both color confinement and besides Poincaré invariance in the physical Hilbert space also Poincaré covariance everywhere in the intermediate steps the Hilbert space $(\mathscr{H},(\cdot|\cdot))$ with metric operator η, respectively, indefinite metric $\langle\cdot,\cdot\rangle$ and corresponding subspaces H and H_0 have to be replaced by a set of such objects $(\mathscr{H}(m)$, $(\cdot|\cdot)m)$, $\eta m,\langle\cdot,\cdot\rangle m, H(m), H_0(m)$, with m running through some Lorentz orbit in the parameter space $\{(\delta,\kappa,w,M,N)\}$ of the representations of the translation group. The physical Hilbert space $\mathscr{H}_{\text{phys}}$ is then given by the direct integral of the completed factor spaces $(\overline{H(m)/H_0(m)},\langle\cdot,\cdot\rangle_m)$ with respect to the invariant measure of this Lorentz orbit.

In contrast to the imaginary color coupling $i\delta$, the flavor couplings have to be real, because otherwise the same mechanism as in the case of color confinement would ban all charged flavor states from $\mathscr{H}_{\text{phys}}$.

In accordance with the discussion above in the following $\delta_\nu\neq 0$ will be taken real (for the covariant gauges). Then the energy density $E^{\delta,M,N}_{\kappa,\omega}(x):=\Theta^{\delta,M,N}_{\kappa,\omega}(x)^0_0$ reads as

$$E^{\delta,M,N}_{\kappa,\omega}(x)=H(x)-i\delta_0\left[j^{8,0}_{(1,0,0)}(x)_{\mathscr{C}}+\frac{1}{\sqrt{3}}\,j^{8,0}_{(0,0,0)}(x)_{\mathscr{C}}\right]+\kappa_0 j^{8,0}_{(1,0,0)}(x)_{\mathscr{F}}+\frac{\omega_0}{\sqrt{3}}\,j^{8,0}_{(0,0,0)}(x)_{\mathscr{F}}$$
$$+N_0 j^{8_S,0}_{(1,0,0)}(x)_{\mathscr{F}}+M_0 j^{8_S,0}_{(0,0,0)}(x)_{\mathscr{F}}, \tag{27}$$

with $\delta_0,\kappa_0,\omega_0,M_0,N_0\in\mathbb{R}$, and $H(x)$ the *canonical* Hamiltonian density.

III. SYMMETRY BREAKING

Of particular interest are the four flavor SU(3)—octet terms that occur on the right-hand side of (27), in addition to the flavor singlet canonical Hamiltonian and the flavor singlet color terms in the first row. They are just identical with the flavor symmetry breaking ''medium strong interaction Hamiltonian'' ($j^{8,0}_{(0,0,0)}$ and $j^{8_S,0}_{(0,0,0)}$), respectively, the electromagnetic interaction $j^{8,0}_{(1,0,0)}$ and $j^{8_S,0}_{(1,0,0)}$, which have been successfully used to calculate the mass splitting within hadron multiplets,[1–7] and especially derive the famous Gell-Mann–Okubo mass formulas.[24,25] Up to now either these terms have to be added by hand to the flavor singlet canonical Hamiltonian or the first few terms have to be selected again by hand from the expansion of a mass operator into ''SU(3) multipoles'' that preserve the selection rules for isospin and hypercharge. The medium strong flavor terms $j^{\cdots}_{(0,0,0)}(x)_{\mathscr{F}}$ do not touch total isospin and hypercharge conservation, and they are responsible for the mass splitting between the different isospin multiplets within a flavor multiplet, whereas $j^{\cdots}_{(1,0,0)}$ leaves only the third component of the isospin besides the hypercharge conserved and could cause the splitting within an isospin multiplet.

However, it seems rather curious that the electromagnetic mass splitting within an isospin multiplet could be achieved without an electromagnetic interaction term showing up in the Lagrangian. Indeed, it will be shown below that the Goldstone Theorem with the pseudoscalar mesons as Goldstone particles implies, $\kappa=-16/\sqrt{27}N$. This, in turn, enforces [via Eq. (16)] the mutually compensation of the isospin breaking terms ($\sim\tau_{(1,0,0)}$) in the two flavor currents $j^{8,\mu}_{(1,0,0)}$ and $j^{8_S,\mu}_{(1,0,0)}$ on the right-hand side of the equations (23) and (26). Hence the mass splitting within an isospin submultiplet is not due to a genuine Goldstone mechanism.

The first main new result of our investigation is that the form invariance of the CD-Lagrangian with respect to (nonstandard) representations of the translation group produces exactly these mass splitting (flavor symmetry breaking) terms in the corresponding conserved Noether current (energy momentum density tensor); in short, they are derived from translational invariance. Hence the ''unknown medium strong interaction Hamiltonian'' has come to be known (to those who want to know it).

Since the representations of the (Abelian) translation group occurring in Eq. (19) are the most general one-dimensional ones that leave the CD-Lagrangian form invariant, our result explains why the octet model i.e., the assumption that all the higher flavor tensor operators $j^{27}_{(0,0,0)}$,

$j^{64}_{(0,0,0)}\ldots$, respectively, $j^{27}_{(1,0,0)}$, $j^{64}_{(1,0,0)},\ldots$, do not contribute to the hadron mass splitting, works so well. In other words, the symmetry breaking is intrinsic to the scheme, in the sense that it emerges automatically, without having to be put in by hand.

One may even turn the argument around and take the hadron mass splitting as an indication that nature prefers to select (for the strong interaction) a representation with $(\kappa_0,\ \omega_0,\ M_0,\ N_0)$ unequal to zero instead of the standard one. Because whenever one adds the flavor symmetry breaking interaction Hamiltonian—by hand or otherwise—to the strong symmetry conserving canonical one, the resulting states of the system cannot transform any more with the standard representation of the translation group.

From the equation (27) all the hadron mass formulas known in the literature[1–7,24,25] can easily be reproduced by means of the usual and generally accepted rules of the game to be found in any textbook on this topic. They consist in taking the expectation value for the various one-particle hadron states and adding the following two assumptions: (i) The flavor SU(3) breaking interaction Hamiltonian, which means in the present scheme the four flavor currents in Eq. (27), are members of an irreducible octet tensor operator of the SU(3) flavor group. (ii) The one-particle hadron states may be grouped together into and approximately treated as exact SU(3) flavor multiplets. For instance, the Gell-Mann–Okubo mass formula for the barion sector is directly obtained from Eq. (27) with $\kappa=N=0$.

Now a second result within our scheme is that the property (i) must not be put in as an additional assumption but follows via Noether's Theorem, provided all three quark ''masses'' in the CD-Lagrangian (10) are equal $(\rho_u=\rho_d=\rho_s)$. However, similarly to the case of the physical hadrons above, the observed splitting of the phenomenological quark masses is obtained via Eq. (27), starting with equal and even zero quark mass parameters in the CD-Lagrangian and taking the expectation values for the (color and flavor) quark triplet with the (positive definite) scalar product $(\cdot|\cdot)_m$ in the unphysical Hilbert space $\mathscr{H}(m)$. [In every representation with Im $\delta_0\neq 0$ there emerges also a contribution from the color charge density $j^{8,0}_{(1,0,0)}(x)_{\mathscr{C}}+(1/\sqrt{3}),j^{8,0}_{(0,0,0)}(x)_{\mathscr{C}}$.] Hence there exists no physical reason that can prevent one from starting with a CD Lagrangian, which is SU(3) form invariant with respect to both the color and the flavor group and even with vanishing quark mass parameters $(\rho_u=\rho_d=\rho_s=0)$.

The physical hadron mass splitting as well as the phenomenological quark masses then emerge via a *(algebraic) symmetry breaking of the global flavor group, in the sense that the* CD-Lagrangian *and hence the* Euler–Lagrange *field equations are form invariant with respect to these groups, but their infinitesimal generators* τ_a *with* $a\neq(1,0,0),(0,0,0)$ *may not commute with the energy operator (i.e., the total Hamiltonian consisting of the canonical plus the symmetry breaking one).* Since they do not commute with the generator of the time translations, the corresponding charges may not be conserved in time, if they exist at all. For $\delta\neq 0$ the equation (27) implies a symmetry breaking in the same sense (at least on the Lagrange–Hamilton level) for the global subgroup of the SU(3) local gauge color group.

Hence, in general, the form invariance of the CD-Lagrangian under the substitution rules of the basic fields with respect to SU(3) transformations do not evoke the corresponding SU(3) symmetry of the physical system. Depending critically on the representation $T^{\delta,M,N}_{\kappa,\omega}$ of the translation group selected out of all those that themselves leave the Lagrangian, respectively, the Euler–Lagrange field equations form invariant the SU(3) symmetry is broken in the sense above, with only a rest symmetry with respect to some subgroup being left. For instance in all representations with $\kappa=-10/\sqrt{27}N$ due to mutual compensation, all terms containing $\tau_{(1,0,0)}$ disappear from the exponentials of the flavor matrices $R\cdots$ in the substitution rules (19), (20), respectively, (24), (25), and hence from the Hamilton density (27). Therefore all generators of the SU(2) subgroup commute with the translations, respectively, the Hamiltonian. Furthermore, in view of Eq. (16), the generator $\tau_{(0,0,0)}$ (hypercharge) of the unitary subgroup U(1) obviously commutes with the translations and, of course, with all generators of SU(2). Therefore in all representations with $\kappa=-10/\sqrt{27}N$, the subgroup SU(2)×U(1) remains a symmetry group on the Lagrange–Hamilton or field algebra level. Of course in the standard representation the full flavor SU(3) symmetry holds.

The (algebraic) symmetry breaking discussed so far should not be confused with the spontaneous symmetry breaking via a Goldstone mechanism. This eventually occurs in the next step, when the representation of the field algebra in an inner product space or Hilbert space is taken into

account, giving rise to a further reduction of the symmetry of the system. As will be shown below, the existence of the pseudoscalar K mesons with their finite physical masses evokes the spontaneous symmetry breaking of the *chiral* flavor symmetry via a genuine Goldstone mechanism and at the same time ruling out all representations $T^{\delta,M,N}_{\kappa,\omega}$ of the translation group with $\kappa \neq -10/\sqrt{27}N$ from quantum flavor dynamics.

Postponing the Goldstone problem for a moment, let us shortly consider the chiral flavor symmetry. With all ρ_f, $f=u$, d, s being zero the quark triplet may be decomposed into a left-handed $\varphi^{3,3,L}_{\ldots}:=(1+\gamma_5)\varphi^{3,3}_{\ldots}$ and a right-handed $\varphi^{3,3,R}_{\ldots}:=(1-\gamma_5)\varphi^{3,3}_{\ldots}$ one ($\gamma_5:=i\gamma^0\gamma^1\gamma^2\gamma^3$). Then the Lagrangian (10) reads as

$$L(\varphi(x),B(x))=-1/4(-1)^{Q(a)}G_{\mu,\nu-a}(x)G^{\mu\nu}{}_{,a}(x)+\sum_{h=L,R}\widetilde{\varphi}^{3,3,h}_{a,f}(x)\,\frac{i}{2}\,\gamma^\mu\{\delta^b_a\overleftrightarrow{\partial}_\mu + ig(\tau_c)^b_a\varphi^{8,1}_{\mu,c}(x)\}\varphi^{3,3,h}_{b,f}(x)-(-1)^{Q(a)}\left[\varphi^{8,1}_{\mu,-a}(x)\partial^\mu B^{8,1}_a-\frac{\alpha}{2}\,B^{8,1}_{-a}(x)B^{8,1}_a(x)\right], \tag{28}$$

and is now form invariant under the chiral flavor group $\mathrm{SU}(3)_{\mathscr{F},L}\times\mathrm{SU}(3)_{\mathscr{F},R}$.

Performing everywhere the replacements

$$\varphi^{3,3}_{\ldots}(x)\to\varphi^{3,3,h}_{\ldots}(x),\quad j^{8,\mu}_{\ldots}(x)_{\mathscr{F}}\longrightarrow j^{8,h,\mu}_{\ldots}(x)_{\mathscr{F}},\quad j^{8_S,\mu}_{\ldots}(x)_{\mathscr{F}}\longrightarrow j^{8_S,h,\mu}_{\ldots}(x)_{\mathscr{F}}, \tag{29}$$

and summing on the right-hand side of the equations (13), (15), (22)–(24), and (27) all terms carrying the new chiral index over $h=L$, R, everything that has been said before remains true also in the chiral case with the groups SU(3), respectively, SU(2)×U(1) replaced by their chiral counterparts.

The implicit assumption that left- and right-handed quark fields transform under translations with the same representation ($\kappa_L=\kappa_R,\ldots$) is enforced by parity conservation, since in this way only vector currents $j^{\ldots}_{\ldots,\nu}:=j^{\ldots,L}_{\ldots}+j^{\ldots,R}_{\ldots}$ occur in the Noether energy momentum tensor (23), respectively (27), but no axial ones $j^{\ldots}_{\ldots,A}:=j^{\ldots,L}_{\ldots}-j^{\ldots,R}_{\ldots}$. A second reason is that only in this case the vector and axial vectors transform covariantly with respect to translations.

The spontaneous symmetry breaking mechanism and especially the Goldstone problem deserves a detailed and rigorous reinvestigation, since some changes concerning the mass of the Goldstone particles are to be expected due to the difference in the representations of the translation group considered here compared to the usually applied (standard) one. In the standard case $\delta=w=\kappa=M=N=0$; the generalizations of the Goldstone theorem from classical (nongauge) quantum field theories to gauge theories with an indefinite metric (see, e.g., Refs. 26–29, respectively,[13,21] and references therein) state that the spontaneous breaking of an algebraic (global) symmetry group whose generators are associated to conserved currents is necessarily accompanied by the occurrence of $\delta(p^2)$ singularities (Goldstone particles), which are created by the corresponding conserved currents from the vacuum and are pushed via the local gauge condition (26) into the unphysical sections of the Hilbert space in case the (global) symmetry group is the subgroup with constant elements of a local gauge group like the color group in QCD.[13,21] In case the spontaneous broken (global) symmetry group is not considered to be a subgroup of a local gauge group like the flavor group in QCD, the massless Goldstone states may be lying either in the unphysical sections or in the physical one $H\backslash H_0$ and then occur as real particles in $\mathscr{H}_{\mathrm{phys}}$ (genuine Goldstone mechanism).

However, as far as the zero mass of the Goldstone states is concerned, these results may not hold true anymore in the case at hand, since they are strongly tied to the standard representation of the translation group (with $\omega=\kappa=\delta=M=N=0$) and the four-momentum spectrum of the states allowed in both the physical and the unphysical sections of the total Hilbert space $\mathscr{H}$ (for details see Refs. 13, 21, 28). Obviously, the exponential factors in the representations (19) and (20), respectively (24), which make the difference to the standard one will displace and even deform the support of the spectral functions in the Jost–Lehmann–Dyson representation of the commutator matrix elements involved, thereby moving the light cone of the Goldstone particles, and even edge them away from the light cone.

Indeed our second main result is that the Goldstone particles accompanying the spontaneous symmetry breaking of the chiral flavor group must possess the squared mass $p(8,a)^2$, $a \in \Gamma_8$. The proof follows in almost the classical way[13,21,27–29] from a generalized Jost–Lehmann–Dyson representation for the commutator matrix element of the current $j^{8,0}_{a,h}(x)_{\mathscr{F}}$ involved and an element $\phi(f)$ from the polynomial + algebra generated by the basic fields that incorporates automatically besides locality (Einstein causality) and the spectrum condition also the current conservation. The proof and the precise definition of the current generated charges,

$$Q^{\cdots}_{a}(y)^{\kappa N}_{\delta,M,\omega} = '' \int d^3\mathbf{x} (T^{\delta,M,N}_{\kappa,\omega}(y) j)^{8,0}_{a,\ldots}(0,\mathbf{x})_{\mathscr{F}}'',$$

are given in Sec. VI below.

From the equations (25) we obtain

$$\begin{aligned} p(8,(1/2,\pm 1/2,1)) &= -p(8,(1/2,\pm 1/2,-1)) \\ &= I(1/2,\pm 1/2,1/3)[I(1/2,\pm 1/2,1/3)+1]M + 1/2Y(1,\pm,1) \\ &\quad + 1/2M + 2I^3(1/2,\pm 1/2,1/3)(\kappa + 10/\sqrt{27}N), \end{aligned} \tag{30}$$

$$p(8,(1,\pm 1,0)) = 2I^3(1,\pm 1,0)(\kappa + 10/\sqrt{27}N), \tag{31}$$

$$p(8,(1,0,0)) = p(8,(0,0,0)) = 0. \tag{32}$$

Hence there is no representation $T^{\delta,M,N}_{\kappa,\omega}$ of the translation group for which the (finite mass) pseudoscalar π_0 or η meson is a Goldstone particle.

In every representation with the properties

$$m(K_+)^2 = p(8,(1/2,1/2,1))^2 = p(8,1/2,-1/2,-1))^2 = m(K_-)^2,$$

$$m(K_0)^2 = p(8,(1/2,-1/2,1)) = p(8,(1/2,1/2,-1))^2 = m(\bar{K_0})^2,$$

which means satisfying the Goldstone conditions:

$$4(\kappa + 10/\sqrt{27}N)^2 + (\omega + 5/3M)^2 = 2m(K_+)^2 + 2m(K_0)^2,$$

$$2(\kappa + 10/\sqrt{27}N)(\omega + 5/3M) = m(K_+)^2 - m(K_0)^2, \tag{33}$$

the (finite mass) pseudoscalar K-meson doublets are Goldstone particles. Therefore the chiral U-spin subgroup is spontaneously broken by the charged $K_{+/-}$ mesons and the V-spin subgroup by the neutral ones $K_{0/\bar{0}}$, because current generated axial charges carrying nonzero hypercharge $Q^A_a(y)^{\kappa,N}_{\delta,M,\omega}$ with $a \in \{(1/2,1/2,\pm 1),(1/2,-1/2,\pm 1)\}$ do not exist. On the other hand, the remaining axial charges with $a \in \{(1,\pm 1,0),(1,0,0),(0,0,0)\}$ together with all vector charges $Q^V_a(y)^{\kappa,N}_{\delta,M,\omega}$, $a \in \Gamma_8$ exist as *bona fide* operators destroying the vacuum. In general, not all of them commute with the translation operators. They are translational covariant but not invariant. In other words, they depend on the space–time coordinates y that simply reflects the algebraic symmetry breaking on the Lagrangian level just discussed before. On the basis of a mathematically consistent and careful formulation of the content of the commutation relations between the Noether currents and the basic fields of the chiral Lagrangian that follow on the formal Lagrangian level by blindly applying equal time canonical commutation relations, it follows (Sec. VI) that the left- and right-handed charges,

$$\{Q^{L/R}_a(0)^{\cdots}_{\cdots} := 1/2(Q^V_a(0)^{\cdots}_{\cdots} \pm Q^A_a(0)^{\cdots}_{\cdots}) | a \in \{(1,\pm 1,0),(1,0,0),(0,0,0)\}\}, \tag{34}$$

form a self-conjugate representation of the (complexified) chiral subalgebra $(\mathrm{su}(2)\oplus\mathrm{u}(1))_L \oplus (\mathrm{su}(2)\oplus\mathrm{u}(1))_R$ and the vector charges $Q^V_a(0)^{\cdots}_{\cdots} a \in \Gamma_8$ separately a representation of su(3) both

generating via exponentiation an η_m-isometric representation of the corresponding Lie Groups $(\mathrm{SU}(2)\times\mathrm{U}(1))_L\times(\mathrm{SU}(2)\times\mathrm{U}(1))_R$, respectively, SU(3). Furthermore, the charges $Q_a^{\cdots}(y)_{\cdots}^{\cdots}$ commute with the translation operators if and only if

$$[\wp,\tau_a]_-=(2I^3(a)(\kappa+10/\sqrt{27N})+Y(a)(\omega+5/3M))\tau_a=0, \tag{35}$$

with $\wp$ given in Eq. (24). Therefore the η_m-isometric representation of the chiral SU(2)×U(1) subgroup emerging from the two octets of conserved chiral currents via the Goldstone theorem is a strict symmetry if and only if

$$\kappa=-10/\sqrt{27N}, \tag{36}$$

which means only in the representations of the translation group satisfying, besides the Goldstone conditions (33), the condition (36). But this in turn demands equal masses of charged and uncharged K mesons. Hence the mass splitting within isospin submultiplets may only be covered by the introduction of the usual electromagnetic interaction term $j_{el}^{\mu}\cdot A_{\mu}$ into the Lagrangian.

By the same reasoning, the condition (35) implies only the restriction of the η_m-isometric representation of SU(3) emerging from the octet of conserved vector currents to the subgroup SU(2)×U(1) isomorphic to the diagonal subgroup of the chiral group $(\mathrm{SU}(2)\times\mathrm{U}(1))_L\times(\mathrm{SU}(2)\times\mathrm{U}(1))_R$ to be a strict symmetry.

Summing up: *In every representation of the translation group satisfying the conditions (33) and (36) the (finite mass) K mesons are Goldstone particles, and the two octets of conserved chiral currents generate an η_m-isometric representation of the chiral group* $(\mathrm{SU}(2)\times\mathrm{U}(1))_L\times(\mathrm{SU}(2)\times\mathrm{U}(1))_R$ *commuting with the translations, provided the masses of charged and uncharged K mesons are considered equal. The octet of conserved vector currents generates an η_m-isometric representation of SU(3), but only its restriction to the subgroup* SU(2)×U(1) *commutes with the translations.*

Similarly, due to Eq. (31) in every representation, with the properties

$$m(\pi_+)^2=p(8,(1,1,0))^2=p(8,(1,-1,0))^2=m(\pi_-)^2,$$

which means satisfying the Goldstone condition,

$$4(k+10/\sqrt{27N})^2=m(\pi_{\pm})^2, \tag{37}$$

the charged pions are Goldstone particles breaking spontaneously the chiral I-spin subgroup, since the axial charges $Q^A_{(1,\pm1,0)}(y)_{\cdots}^{\cdots}$ do not exist. In the same way as above the chiral charges $Q^{L/R}_{(1,0,0)}(0)_{\cdots}^{\cdots}$, $Q^{L}_{(0,0,0)}(0)_{\cdots}^{\cdots}$ together with $Q^{L/R}_{(1/2,-1/2,1)}(0)_{\cdots}^{\cdots}$, $Q^{L/R}_{(1/2,1/2,-1)}(0)_{\cdots}^{\cdots}$ generate an η_m-isometric representation of the chiral U-spin subgroup, together with $Q^{L/R}_{(1/2,1/2,1)}(0)_{\cdots}^{\cdots}$, $Q^{L/R}_{(1/2,-1/2,-1)}(0)_{\cdots}^{\cdots}$ an η_m-isomeric representation of the chiral V-spin subgroup and the vector charges $Q^V_a(0)_{\cdots}^{\cdots}$, $a\in\Gamma_8$ separately one of SU(3). However, none of them is a strict symmetry since the condition (35) for commutativity with the translation operators requires, in contradiction to the Goldstone condition (37), $\kappa=-10/\sqrt{27N}$ and $\omega=-5/3M$, unless $I^3(a)=Y(a)=0$. Hence only the chiral U(1)×U(1) subgroup is left as a strict symmetry.

Since the Goldstone conditions (33) and (37) together admit no solution unless the masses of the charged pions are taken to be zero, only representations $T^{\delta,M,N}_{\kappa,\omega}$ of the translation group with

$$\kappa=-10/\sqrt{27N}\wedge(\omega+5/3M)^2=4m(K)^2, \tag{38}$$

seem to be selected by nature for flavored chiral QCD. Every solution of the equations (38) fixes an admissible representation of the translation group, and hence an admissible Hamiltonian for chiral flavor dynamics with chiral SU(3) symmetry of the Lagrangian being spontaneously broken by the K mesons down to chiral SU(2)×U(1) symmetry. Moreover, the success of the Gell–Mann–Okulo mass formula in the barion sector seems to indicate that the representations with $N=\kappa=0$ are the most relevant ones for flavored QCD.

To classify all Lorentz orbits and their invariant measures is beyond the scope of this paper. We state them only for the special cases (which in our opinion cover the most relevant ones):

$\omega=(2/\alpha-5/3)M$, $N=\beta M$, $\delta=\gamma M$ with $\alpha\in\mathbb{R}\backslash\{0\}$ and β, $\gamma\in\mathbb{R}$ just to demonstrate the existence of Lorentz orbits and invariant measures covering the Goldstone conditions (38). In these cases the Lorentz orbits are just the hyperboloids $\mathcal{O}_{\mathscr{L}}=\mathcal{O}_{\alpha}^{\pm}=\{M\in\mathbb{R}_4|M^2=\alpha^2 m(K)^2, M_0\gtrless 0\}$ with the invariant measures

$$d\mu_{\alpha}^{\pm}(M)=\Theta(\pm M_0)\,\delta(M^2-\alpha^2 m(K)^2)d^4M.$$

Due to the occurrence of the conserved color currents $i\delta_0(j^{8,0}_{(1,0,0)}(x)_{\mathscr{C}}+1/\sqrt{3}j^{8,0}_{(0,0,0)}(x)_{\mathscr{C}})$ in the Hamilton density (27) in case of confinement ($\delta\neq 0$) also the SU(3)-color symmetry is broken on the algebraic level and the remaining color symmetry group is the subgroup

$$G(3)_C=\left\{\exp\cdot[it/2(\tau_{(1,0,0)}+1/\sqrt{3}\,\tau_{(0,0,0)})]\begin{pmatrix}1&0&0\\0&&\\&&g\\0&&\end{pmatrix}\,\middle|\, t\in\mathbb{R}, g\in \mathrm{SU}(2)\right\}, \tag{39}$$

isomorphic to $\mathrm{U}(1)\times\mathrm{SU}(2)$. However, due to color confinement (Sec. IV), there exists no physical reason that the generators of the color group must commute with the Hamiltonian as long as δ is real. It will be shown in Sec. VI that the color symmetry is not spontaneously broken unless there exist states of discrete mass $\kappa(a)=0$ created by the conserved color current $j_a^{8,0}(x)_{\mathscr{C}}$ from the vacuum. However, due to the local gauge condition (26) they are confined in the unphysical section of the Hilbert space.

Before proceeding to the general quantum theory, some remarks should be made concerning the total action from which, via the principles of stationary or least action, the Euler–Lagrange field equations (Hamilton's principle), respectively, the conserved currents (Noether Theorem) are derived. Due to the dependence of the fields on the ''flavor'' vectors $\mathscr{M}=(M,N,\omega)$ and the ''color'' vector δ the total action contains, in addition to the integration over the space–time variables x, an integration with the invariant measure $d\mu_{\mathscr{L}}(\delta,\mathscr{M})$ over the Lorentz orbit $\mathcal{O}_{\mathscr{L}}$, respectively, parts of them;

$$A:=\int d^4x\; d\mu_{\mathscr{L}}(\delta,\mathscr{M})L(\varphi(x)_{(\delta,\mathscr{M})},B(x)_{(\delta,\mathscr{M})}).$$

Even the standard case may be viewed to belong to the (exceptional) Lorentz orbits consisting of only one point $\delta=0$, respectively, $M=N=\omega=0$ with the Dirac δ^4-functions as the invariant measures. Hence the additional integrals disappear in this case.

Since the Lagrangians (10), respectively (28), contain *no* derivative of the fields with respect to any of the variables $(\delta,\mathscr{M})$, it is easily seen that the Euler–Lagrange field equations are the same for all Lorentz orbits. Because, in addition, for the internal symmetry transformations as well as for the translations, the variables $(\delta,\mathscr{M})$ are kept fixed—the corresponding variations are zero!—the Noether currents of the internal symmetries are again the same for all Lorentz orbits, and the energy–momentum density differs only by terms due to the different substitution rule (19) for the fields. Hence, in the calculations before the additional variables $(\delta,\mathscr{M})$ could be considered as fixed parameters and the integration over them be dropped. It is needless to mention that in the step from the densities to the total charges (for instance the total Hamiltonian) the integrations over the Lorentz orbits are to be included.

The situation changes if Lorentz transformations are considered, since then the new vector variables change in the same way as the space–time vector. Thus their variations do not vanish any more and in the Noether spin-density tensor contributions due to the internal geometry of the Lorentz orbits could be expected. However, the detailed calculation shows that these contributions compensate each other in the Noether conservation law. Thus the Noether spin-angular–momentum tensor is the same for all orbits (up to the dependence of the fields on the color and flavor variables). The reason once again is the nonoccurrence of derivatives with respect to these variables in the Langrangian and the independence of the Euler–Lagrange field equations from the orbits.

Finally, on the basis of these remarks the direct integral structure of the quantum theoretical Hilbert space $\mathscr{H}$ with its components $\mathscr{H}(\delta,\mathscr{M})$ being clued together via the Lorentz group is perhaps rather obvious.

IV. POINCARÉ COVARIANCE AND CONFINEMENT

In view of the considerations and results of the foregoing sections, the transformation laws (1) of the basic fields that are clued to the standard representation ($w=M=N=\delta=0$) of the translation group [that is to a (Noether) Hamiltonian without symmetry breaking and confinement] have to be reinvestigated and modified according to what is prescribed by the symmetry breaking dynamics of QCD mirrored in the choice of the representation for the (time) translation group, or equivalently the Hamiltonian.

Furthermore, as already stated in Sec. III in the general representation (19), the basic quark and gluon fields depend in addition to the space time variable x on the ''color vector'' δ and the ''flavor vector'' triple (M,N,w) specifying the representation. For the sake of notational shortness it has been dropped before but gains decisive importance in the construction of a Poincaré covariant quantum theory. According to the equations (19), (24), (25), and (32), the dependence on the flavor triple $\mathscr{M}:=(M,N,w)$ comes via the matrix

$$\wp(\mathscr{M}):=\sigma_{(0,0,0)}M+(\sigma_{(1,0,0)}-10/\sqrt{27}\tau_{(1,0,0)})N+1/\sqrt{3}\,\tau_{(0,0,0)}w, \tag{40}$$

with eigenvalues of its representatives in the representation R^d:

$$p_{\mathscr{M}}(d,a)_\mu=-i\partial_\mu R^d(\exp[i/2(\wp(\mathscr{M})\cdot x)])^a_{a_{|x=0}}. \tag{41}$$

For every fixed quartet $m=(\delta,\mathscr{M})=(\delta,M,N,w)$ of four vectors from the Lorentz orbit $\mathscr{O}_{\mathscr{L}}$ besides the smeared fields,

$$\varphi^{n,d}_{\nu,a,b}(f)_m:=\int d^4x f(x)\varphi^{n,d}_{\nu,a,b}(x)_m, \tag{42}$$

and conserved currents (vector and axial vector indices are dropped in the following):

$$j^{8,\nu}_a(f)_{m_{\mathscr{F}/\mathscr{C}}}:=\int d^4x f(x) j^{8,\nu}_a(x)_{m,\mathscr{F}/\mathscr{C}}, \tag{43}$$

two- and three-fold composite quark fields [from which the one particle meson—respectively barion—states are constructed], are defined by

$$\varphi^{3^*3,3^*3}_{\nu_1\nu_2,a_1a_2,b_1b_2}(f)_m:=\int d^4x_1 d^4x_2 f(x_1,x_2)(-1)^{-Q(a_1)-Q(b_1)-2/3}\varphi^{3^*,3^*}_{\rho,-a_1,-b_1}\times(x_1)_m(\gamma^0)^\rho_{\nu_1}\varphi^{3,3}_{\nu_2,a_2,b_2}(x_2)_m, \tag{44}$$

respectively, for $n\in\{3,3^*\}$:

$$\varphi^{nnn,nnn}_{\nu_1\nu_2\nu_3,a_1a_2a_3,b_1b_2b_3}(f)_m:=\int d^4x_1 d^4x_2 d^4x_3 f(x_1,x_2,x_3)\varphi^{n,n}_{\nu_1,a_1,b_1}(x_1)_m\ldots\varphi^{n,n}_{\nu_3,a_3,b_3}(x_2)_m. \tag{45}$$

Here f is from the countably normed space $S(\mathbb{R}_{4k})$, $k=1,2,3$ of strongly decreasing $\mathscr{C}^\infty$ functions in all variables. With the abbreviations $|x|^2_k:=\Sigma^k_{i=1}\Sigma^3_{j=0}(x^j_i)^2$ and $\partial^{|r|}_x$ for the partial differential operator in the components x^j_i of total order $|r|=\Sigma^k_{i=1}\Sigma^r_{j=0}r^j_i$, $r^j_i\in\mathbb{N}_0$ the norms $\|f\|_n=\sqrt{[f,f]_n}$, $n\in\mathbb{N}_0$ are defined via the scalar products[30] (Chap. I, Sec. 3.6):

$$[f,h]_n=\int d^4x_1\cdots d^4x_k(1+|x|_k)^{2n}\sum_{\substack{r\in\mathbb{N}_0^{4k}\\ |r|\leqslant n}}(\overline{\partial^{|r|}_x f(x)})\partial^{|r|}_x h(x). \tag{46}$$

For every fixed $m \in \mathcal{O}_{\mathcal{L}}$ the smeared fields are closable linear operators with a common dense subdomain in a Hilbert space $(\mathcal{H}(m),(\cdot|\cdot)_m,\eta_m)$ with η_m a bounded symmetric operator generating the nondegenerate indefinite metric $\langle\cdot,\cdot\rangle_m := (\cdot|\eta_m\cdot)_m$, which is non-negative on a proper maximal subspace $H(m)$ and zero on the subset $\mathcal{H}_0(m)$. η_m maps the common dense subdomain of all fields into itself. The counterpart to the complex conjugate $\overline{\varphi_m}$ and Dirac adjoint field $\widetilde{\varphi}_m$ of the classical level now is the η_m-adjoint field operator $\varphi^{\cdots\,+}_{\cdots}(f)_m := \varphi^{\cdots}_{\cdots}(\hat{f})^{+}_m$, respectively, $\widetilde{\varphi}^{\cdots}_{\cdots}(f)_m = \varphi^{\cdots\,+}_{\cdots}(f)_m\gamma^0$ with $\hat{f}(x_1\ldots x_k) := \overline{f(x_k,\ldots,x_1)}$. Accordingly, the self-conjugateness (8) of the gluon fields and conserved currents, respectively, the relation (9) between the (anti-)quark fields read on the operator level with * denoting the Hilbert adjoint operator,

$$\varphi^{8,1*}_{\mu,c}(f)_m\eta_m = \eta_m(-1)^{Q(c)}\varphi^{8,1}_{\mu,-c}(f)_m, \quad j^{8,\mu*}_{a,\ldots}(f)_m\eta_m = \eta_m(-1)^{Q(a)}j^{8,\mu}_{-a,\ldots}(f)_m, \tag{47}$$

$$\varphi^{3^*3,3^*3*}_{\mu_1\mu_2,c_1c_2,a_1a_2}(f)_m\eta_m = \eta_m\varphi^{3^*3,3^*3}_{\mu_2\mu_1,c_2c_1,a_2a_1}(f)_m, \tag{48}$$

$$\begin{aligned}\eta_m\widetilde{\varphi}^{nnn,nnn}_{\mu_1\ldots\mu_k,c_1\ldots c_k,a_1\ldots a_k}(f)_m &= \varphi^{nnn,nnn*}_{\nu_1\ldots\nu_k,c_1\ldots c_k,a_1\ldots a_k}(f)_m\Pi_{i=1}^k(\gamma^0)^{\nu_i}_{\mu_i}\eta_m \\ &= \eta_m(\Pi_{i=1}^k(-1)^{-Q(c_i)-Q(a_i)-2/3})\varphi^{n^*n^*n^*,n^*n^*n^*}_{\mu_1\ldots\mu_k,-c_1\ldots-c_k,-a_1\ldots-a_k}(f)_m,\end{aligned} \tag{49}$$

with $n \in \{3,3^*\}, k \in \{1,3\}$.

In the subset $H(m)\backslash H_0(m)$ with $H_0(m) := H(m)\cap\mathcal{H}_0$ exists a normalized eigenvector $\Psi_0(m)$ of η_m with $\langle\Psi_0(m),\Psi_0(m)\rangle_m = 1$ that is cyclic for the polynomial + algebra $\pi(\varphi_m,j_m)$ spanned by the basic (including the conserved currents) and the composite field operators; that means the subspace $\mathcal{D}(m) := \pi(\varphi_m,j_m)\Psi_0(m)$ is dense in $(\mathcal{H}(m),(\cdot|\cdot)_m)$. Moreover, the subspace $\mathcal{D}_H(m) := \mathcal{D}(m)\cap H(m)$ is assumed to be dense in $H(m)$. Last but not least, the matrix elements of the field operators with all states from $\mathcal{D}(m)$ are continuous linear functionals on $S(\mathbb{R}_{4k})$, $k = 1,2,3$.

On $\mathcal{D}(m)$ a representation of the translation group is obtained by a linear extension of the operators $\mathcal{T}_m(y)$ defined successively via the following covariance equations on all states created by products of the field operators from $\Psi_0(m)$:

$$\mathcal{T}_m(y)\Psi_0(m) := \Psi_0(m), \tag{50}$$

$$\begin{aligned}&\mathcal{T}_m(y)\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\nu_1\ldots\nu_k,c_1\ldots c_k,a_1\ldots a_k}(f)_m\mathcal{T}_m(-y) \\ &\quad := \int d^{4k}x f(x_1,\ldots,x_k)\Pi_{i=1}^k(T^{\delta,M,N}_{-10/\sqrt{27}N,w}(y)\varphi)^{n_i,d_i}_{\nu_i,c_i,a_i}(x_i)_m \\ &\quad = (\Pi_{i=1}^k\exp[Q(c_i)(\delta\cdot y)])(\Pi_{j=1}^k R^{dj}(\exp[i/2(\wp(\mathcal{M})\cdot y)])^{b_j}_{a_j}) \\ &\qquad\times\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\nu_1\ldots\nu_k,c_1\ldots c_k,b_1\ldots b_k}(f_{(1,y)})_m,\end{aligned} \tag{51}$$

with $k \in \{1,2,3\}$ and $f_{(\Lambda(\alpha),y)}$, defined by

$$f_{(\Lambda(\alpha),y)}(x_1,\ldots,x_k) := f(\Lambda(\alpha)^{-1}(x_1-y),\ldots,\Lambda(\alpha)^{-1}(x_k-y)). \tag{52}$$

Here and in the following the covariance equations for the currents are not written down explicitly because they are obtained by obvious replacements from those of the gluon fields.

This representation is certainly not isometric with respect to the Hilbert scalar product $(\cdot|\cdot)_m$[9,11,23] and by many authors[13–22] claimed to be even unbounded also in the standard case $[\mathcal{M}=(0,0,0),\ \delta=0]$. However, it has to be isometric with respect to the indefinite metric $\langle\cdot,\cdot\rangle_m$ in order that it can be lifted to a unitary representation in the physical Hilbert space.[11] In other words, the metric operator η_m has to be constructed in such a way that this so-called η-isometry does hold.

In the same way as before, a representation of the internal symmetry group $G(3)_{\mathcal{C}}\times G(3)_{\mathcal{F}}$ that commutes with the translation operators $\mathcal{T}_m(y)$ is obtained from the covariance equations:

$$I_m(u,v)\Psi_0(m) := \Psi_0(m) \tag{53}$$

$$I_m(u,v)\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\nu_1\ldots\nu_k,c_1\ldots c_k,e_1\ldots e_k}(f)_m I_m(u^{-1},v^{-1})$$

$$:= (\Pi_{i=1}^k R^{n_i}(u^{-1})^{a_i}_{c_i} R^{d_i}(v^{-1})^{b_i}_{e_i})\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\nu_1\ldots\nu_k,a_1\ldots a_k,b_1\ldots b_k}(f)_m . \tag{54}$$

As will be shown in Sec. VI, this representation is the η_m-*isometric* integral or exponential of the representation of the Lie algebra $\mathscr{G}(3)_{\mathscr{C}}\oplus\mathscr{G}(3)_{\mathscr{F}}$, which is given by the space integral over the zeroth component of the η_m-self-conjugate conserved color currents $j^{8,\mu}_{\pm}(f)_{m,\mathscr{C}} = \pm 1/2[j^{8,\mu}_{(1,0,0)}(f)_{m,\mathscr{C}} \pm \sqrt{3}^{\mp 1} j^{8,\mu}_{(0,0,0)}(f)_{m,\mathscr{C}}], j^{8,\mu}_{(1/2,1/2,-1)}(f)_{m,\mathscr{C}}, j^{8,\mu}_{(1/2,-1/2,1)}(f)_{m,\mathscr{C}}$, respectively, the flavor (vector) currents $j^{8,\mu}_{a,V}(f)_{m,\mathscr{F}}, a\in\{(1,0,0),(1,\pm 1,0),(0,0,0)\}$. (In the cases where the algebraic color symmetry or part of it is spontaneously broken, this is still true for the remaining unbroken subgroup and in any case for the flavor subgroup $I^{\mathscr{F}}_m(v)=I_m(1,v),\ v\in G(3)_{\mathscr{F}}$).

It should be remarked in passing that the irreducible matrix representations R^d of SU(3) become completely reducible when restricted to the subgroup $G(3)\sim \mathrm{SU}(2)\times\mathrm{U}(1)$. They decompose into a direct sum of irreducible representation D^I of SU(2) labeled by the total isospin I and each term carrying a specific hyperchage Y. Explicitly we have

$$R^d(r,t)^{(I,I^3,Y)}_{(\tilde{I},\tilde{I}^3,\tilde{Y})} = \delta_{I\tilde{I}}\delta_{\tilde{Y}Y} D^I(r)^{I^3}_{\tilde{I}^3}\exp(iYt), \tag{55}$$

with $(I,Y)\in\{(1/2,1/3),(0,-2/3)\}$ for $d=3$ and $(I,Y)\in\{(1/2,1),(1,0),(0,0),(1/2,-1)\}$ for $d=8$, etc.

Concerning the Lorentz- and even more the Poincaré-group matters are more involved. It is known[23] (Secs. VII, VIII) that there does not exist a representation of the Poincaré group in $\mathscr{H}(m)$ resembling the substitution rule (1) of Sec. II for Lorentz transformations, together with the representation $\mathscr{T}_m$, $m\neq 0$ of the subgroup of translations. However, for every Lorentz transformation $\Lambda(\alpha)$, $\alpha\in\mathrm{SL}(2,\mathbb{C})$, a closable linear operator $\mathscr{L}(\alpha)_m$ from $\mathscr{D}(m)\subset\mathscr{H}(m)$ onto $\mathscr{D}(\Lambda(\alpha)m)\subset\mathscr{H}(\Lambda(\alpha)m)$ is obtained in the same way as before (the successive definition on product states and linear extension) by

$$\mathscr{L}(\alpha)_m\Psi_0(m) := \Psi_0(\Lambda(\alpha)m), \tag{56}$$

$$\mathscr{L}(\alpha)_m\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\nu_1\ldots\nu_k,\ldots,\ldots}(f)_m\mathscr{L}(\alpha)^{-1}_m := (\Pi_{i=1}^k S^{(n_i,d_i)}(\alpha)^{\rho_i}_{\nu_i})\varphi^{n_1\ldots n_k,d_1\ldots d_k}_{\rho_1\ldots\rho_k,\ldots,\ldots}(f_{(\Lambda(\alpha),0)})_{\Lambda(\alpha)m}, \tag{57}$$

with $f_{(\Lambda(\alpha),0)}$ defined in Eq. (52) and the matrices $S^{(n,d)}(\alpha)$ given by the equations (2), (3) of Sec. II. The operators $\mathscr{L}(\alpha)_m$ obviously satisfy the composition law $\mathscr{L}(\alpha_1)_{\Lambda(\alpha_2)m}\mathscr{L}(\alpha_2)_m = \mathscr{L}(\alpha_1\alpha_2)_m$. η-isometry for the Lorentz operators means for all $\Psi,\Phi\in\mathscr{D}(m)$:

$$\langle\mathscr{L}(\alpha)_m\Psi,\mathscr{L}(\alpha)_m\Phi\rangle_{\Lambda(\alpha)m} = \langle\Psi,\Phi\rangle_m \quad \text{or} \quad \mathscr{L}(\alpha)^*_m\eta_{\Lambda(\alpha)m}\mathscr{L}(\alpha)_m\Phi = \eta_m\Phi. \tag{58}$$

For all $y\in\mathbb{R}_4$, $\alpha\in\mathrm{SL}(2,\mathbb{C})$, the linear operators $\mathscr{T}_m(y)$ and $\mathscr{L}(\alpha)_m$ satisfy the commutation relations $\mathscr{T}_{\Lambda(\alpha)m}(y)\mathscr{L}(\alpha)_m\Psi(m) = \mathscr{L}(\alpha)_m\mathscr{T}_m(\Lambda(\alpha^{-1})y)\Psi(m)$, and therefore

$$W(\alpha,y)_m:\mathscr{D}(m)\to\mathscr{D}(\Lambda(\alpha)m), \Psi(m)\to W(\alpha,y)_m\Psi(m) := \mathscr{T}_{\Lambda(\alpha)m}(y)\mathscr{L}(\alpha)_m\Psi(m) \tag{59}$$

defines linear operators satisfying the composition law of the Poincaré group, and, in addition, η-isometry if the Lorentz operators $\mathscr{L}(\alpha)_m$ do so. The question of if the η-isometry of $\mathscr{L}(\alpha)_m$ can be derived from the presence of a conserved tensor current (corresponding to the Noether spin tensor) as in the case of the internal flavor, respectively, color symmetry is a problem of its own and will be treated together with the spin of Goldstone particles in a separate note.

Due to the exponential factor in the representations (51) of the translation group, the problem of cluing these representations and the Lorentz operators $\mathscr{L}$ together to form a unitary representation of the Poincaré group in a suitable Hilbert space, is closely connected with the color confinement. The key to its solution is the following theorem, which states that no colored state

exists in the ''physical section'' $H(m)\backslash H_0(m)$ for any m. They all have a nonpositive η_m norm.

Theorem (color confinement): (a) *Every state* $\Psi^{n_1\ldots n_r}_{c_1\ldots c_r}(m)\in H(m)$ *with total color charge* $Q_{\mathscr{C}}=\Sigma_{j=1}^{r}Q(c_j)\neq 0$ *belongs to* $H_0(m)=H(m)\cap\mathscr{H}_0(m)$, *that is,*

$$\langle\Psi^{n_1\ldots n_r}_{c_1\ldots c_r}(m),\Psi^{n_1\ldots n_r}_{c_1,\ldots c_r}(m)\rangle_m=0.$$

(b) $\langle\Pi_{j=1}^{r}\varphi^{n_j,d_J}_{\ldots,c_j,a_j}(f_j)_m\Psi_0(m),\Pi_{i=1}^{t}\varphi^{q_i,l_i}_{\ldots,e_i,b_i}(h_i)_m\Psi_0(m)\rangle_m=0$ *unless* $\Sigma_{j=1}^{r}Q(c_j)=-\Sigma_{i=1}^{t}Q(e_i)$. (*Hence all these colored states are from* $\mathscr{H}_0$!)

Proof: Since $\mathscr{D}_H(m)$ is dense in $H(m)$ and the metric $\langle\cdot,\cdot\rangle_m$ is semipositive on $H(m)$ it suffices to prove the validity only for all states from $\mathscr{D}_H(m)$ that are created from $\Psi_0(m)$ by pure or mixed products of the basic (again including the conserved currents) and composite field operators. A typical example reads as

$$\Psi^{n_1\ldots n_r,d_1\ldots d_r}_{\ldots,c_1\ldots c_r,a_1\ldots a_r}(m)=\Pi_{i=1}^{r}\,\varphi^{n_i,d_i}_{\ldots,c_i,a_i}(f_i)_m\Psi_0(m).$$

Denoting by $|\eta_m|$ the operator norm of the metric operator η_m, it follows by means of the η_m-isometry of $\mathscr{T}_m$ for all $y\in R_4$:

$$\begin{aligned}
&|\langle\Pi_{k=1}^{r}\,\varphi^{n_k,d_k}_{\ldots,c_k,a_k}(f_k)_m\Psi_0(m),\Pi_{j=1}^{r}\,\varphi^{n_j,d_j}_{\ldots,c_j,a_j}(f_j)_m\Psi_0(m)\rangle_m|\\
&\quad=(\Pi_{i=1}^{r}\,\exp[2Q(c_i)(\delta\cdot y)])|(\Pi_{k=1}^{r}\,\varphi^{n_k,d_k}_{\ldots,c_k,b_k}((f^{d_k}_{k(1,y)})^{b_k}_{a_k})_m\Psi_0(m)\\
&\quad\times|\,\eta_m\Pi_{j=1}^{r}\,\varphi^{n_j,d_j}_{c_j,b_j}((f^{d_j}_{j(1,y)})^{b_j}_{a_j})_m\Psi_0(m))_m|\\
&\quad\leqslant|\,\eta_m|\exp[2Q_{\mathscr{C}}(\delta\cdot y)]\\
&\qquad\times\sum_{b_1,e_1\in\Gamma_{d_1}}\cdots\sum_{b_r,e_r\in\Gamma_{d_r}}\|\Pi_{k=r}^{1}\,\varphi^{n_k,d_k*}_{\ldots,c_k,e_k}((\hat{f}^{d_k}_{k(1,y)})^{e_k}_{a_k})_m\Psi_0(m)\|_m\\
&\qquad\times\|\Pi_{j=1}^{r}\,\varphi^{n_j,d_J}_{\ldots,c_j,b_j}((f^{d_j}_{j(1,y)})^{b_j}_{a_j})_m\Psi_0(m)\|_m,
\end{aligned}$$

with $f^{d_k}_{k(1,y)}:=R^{d_k}(\exp[i/2(\wp(\mathscr{M})\cdot y)])f_{k(1,y)}$. Since every norm in the $2r$-fold sum on the right-hand side is a continuous linear functional on the $2r$-fold tensor product of the countably normed nuclear space $S(\mathbb{R}_4)$, it is of finite order $n_0(e_1\ldots e_r,b_1\ldots b_r)$ (see Ref. 31 Vol. 2). Therefore, there exists a positive, real number A and a natural number $n_0=\max\{n_0(e_1\ldots e_r,b_1\ldots b_r)|e_i,b_i\in\Gamma_{d_i},\ i=1,\ldots,r\}$, such that

$$\begin{aligned}
&|\langle\Pi_{k=1}^{r}\,\varphi^{n_k,d_k}_{\ldots,c_k,d_k}(f_k)_m\Psi_0(m),\Pi_{j=1}^{r}\,\varphi^{n_j,d_j}_{\ldots,c_j,a_j}(f_j)_m\Psi_0(m)\rangle_m|\\
&\quad\leqslant A\cdot|\,\eta_m|\exp[2Q_{\mathscr{C}}(\delta\cdot y)]\sum_{b_1,e_1\in\Gamma_{d_1}}\cdots\sum_{b_r,e_r\in\Gamma_{d_r}}\\
&\qquad\times\|[\Pi_{k=1}^{r}\,(\hat{f}^{d_k}_{k(1,y)})^{e_k}_{a_k}]\Pi_{j=1}^{r}\,(f^{d_J}_{j(1,y)})^{b_J}_{a_j}\|_{n_0}\leqslant A\cdot|\,\eta_m|\exp[2Q_{\mathscr{C}}(\delta\cdot y)]\\
&\qquad\times\sum_{b_1,e_1\in\Gamma_{d_1}}\cdots\sum_{b_r,e_r\in\Gamma_{d_r}}\Pi_{j=1}^{r}\,\|(f^{d_j}_{j(1,y)})^{e_j}_{a_j}\|_{n_0}\|(f^{d_j}_{j(1,y)})^{b_J}_{a_J}\|_{n_0}.
\end{aligned}$$

The elements of the diagonal unitary matrix $R^d(\exp[i/2(\wp\cdot y)])^a_b=\delta^a_b\exp[i(p_{\mathscr{M}}(a,d)\cdot y)]$ with $\wp$ and $p_{\mathscr{M}}(a,d)$ given in the equations (40)–(41), together with all derivatives with respect to the components of $\mathscr{M}=(M,N,w)$, are bounded by polynomials in y. Therefore the equations (52) and (46) imply the whole $2r$-fold sum on the right-hand side of the last inequality to be bounded by a polynomial $\Pi_m(y)$ in y. Thus we arrive at the bound:

$$|\langle\Pi_{k=1}^{r}\,\varphi^{n_k,d_k}_{\ldots,c_k,a_k}(f_k)_m\Psi_0(m),\Pi_{j=1}^{r}\,\varphi^{n_j,d_j}_{\ldots,c_j,a_j}(f_j)_m\Psi_0(m)\rangle_m|\leqslant A|\,\eta_m|\Pi_m\,(y)\exp[2Q_{\mathscr{C}}(\delta\cdot y)],$$

which proves the theorem, since for every $\delta \neq 0$ and $Q_{\mathscr{C}} \neq 0$ there is a direction in $\mathbb{R}_4$ in which the exponential goes to zero for y going to infinity. Part (b) follows in the same way, since in case of these states positive semidefiniteness need not be applied. □

Now it is rather obvious how to proceed in order to construct a carrier space for an isometric representation of the Poincaré group resembling the color confining representations of the translation group as well as the Lorentz transformations $\mathscr{L}$ given by the equations (50)–(52), respectively, (56)–(58). Let $d\mu_{\mathscr{L}}(m)$ denote the invariant measures of the Lorentz orbit $\mathscr{O}_{\mathscr{L}}$ and $\mathscr{V}_H$ the *vector space* of vector-valued mappings,

$$\psi: \mathscr{O}_{\mathscr{L}} \to \cup_{m \in \mathscr{O}_{\mathscr{L}}} \mathscr{D}_H(m), \quad m \to \Psi(m) \in \mathscr{D}_H(m) \subset H(m), \tag{60}$$

with the properties $\langle \Psi(m), e_\nu(m) \rangle_m$, is $\mu_{\cdots}^{\cdots}$ measurable for every member of some orthonormal basis $(e_\nu(m))_{\nu \in \mathbb{N}}$ in $H(m)$ and

$$\int_{\mathscr{O}_{\mathscr{L}}} d\mu_{\mathscr{L}}(m) \langle \Psi(m), \Psi(m) \rangle_m < +\infty.$$

In $\mathscr{V}_H$, vector addition and multiplication by complex numbers are defined in the usual way:

$$\psi + \phi: \mathscr{O}_{\mathscr{L}} \to \cup_{m \in \mathscr{O}_{\mathscr{L}}} \mathscr{D}_H(m), \quad m \to (\Psi + \Phi)(m) = \Psi(m) + \Phi(m),$$

$$\alpha\psi: \mathscr{O}_{\mathscr{L}} \to \cup_{m \in \mathscr{O}_{\mathscr{L}}} \mathscr{D}_H(m), \quad m \to (\alpha\Psi)(m) = \alpha\Psi(m).$$

The *vector space* $\mathscr{V}_H$ is just a subspace of the Cartesian product $\times_{m \in \mathscr{O}_{\mathscr{L}}} H(m)$. Since the metric $\langle \cdot, \cdot \rangle_m$ is positive semidefinite on $H(m)$ and hence also on $\mathscr{D}_H(m)$, a positive semidefinite inner product on $\mathscr{V}_H$ is given by

$$\langle \psi, \phi \rangle := \int_{\mathscr{O}_{\mathscr{L}}} d\mu_{\mathscr{L}}(m) \langle \Psi(m), \Phi(m) \rangle_m. \tag{61}$$

Moreover, the invariance of the measure $\mu_{\mathscr{L}}$ implies $\langle \psi, \psi \rangle = 0$ if and only if $\langle \Psi(m), \Psi(m) \rangle_m = 0$ almost every where on $\mathscr{O}_{\mathscr{L}}$. Hence an isometric representation $\mathscr{P}: ISL(2,\mathbb{C}) \to \mathrm{Aut}\, \mathscr{V}_H$, $(\alpha, y) \to \mathscr{P}(\alpha, y)$ of the Poincaré group on the inner product space $(\mathscr{V}_H, \langle \cdot, \cdot \rangle)$, which leaves the subspace $\mathscr{V}_H^0 := \{\psi \in \mathscr{V}_H | \langle \psi, \psi \rangle = 0\}$ invariant, is given by

$$\mathscr{P}(\alpha, y): \mathscr{V}_H \to \mathscr{V}_H, \quad \psi \to \mathscr{P}(\alpha, y)\psi,$$

$$\mathscr{P}(\alpha, y)\psi: \mathscr{O}_{\mathscr{L}} \to \cup_{m \in \mathscr{O}_{\mathscr{L}}} \mathscr{D}_H(m),$$

$$m \to (\mathscr{P}(\alpha, y)\psi)(m) := \mathscr{T}_{\Lambda(\alpha)m}(y) \mathscr{L}(\alpha)_m \Psi(m) = \mathscr{L}(\alpha)_m \mathscr{T}_m(\Lambda(\alpha^{-1})y) \Psi(m), \tag{62}$$

where $\mathscr{T}_m(y)$ and $\mathscr{L}(\alpha)_m$ are the η-isomeric operators defined via the equations (50)–(52), respectively, (56)–(57).

In the same way, an isometric representation of the internal symmetry group $G(3)_{\mathscr{C}} \times G(3)_{\mathscr{F}}$ and therefore also of its direct product with the Poincaré group, is obtained by

$$\mathscr{I}(u, v): \mathscr{V}_H \to \mathscr{V}_H, \quad \psi \to \mathscr{I}(u, v)\psi,$$

$$\mathscr{I}(u, v)\psi: \mathscr{O}_{\mathscr{L}} \to \cup_{m \in \mathscr{O}_{\mathscr{L}}} \mathscr{D}_H(m), \quad m \to (\mathscr{I}(u, v)\psi)(m) = I_m(u, v)\Psi(m), \tag{63}$$

with the η-isometric operators $I_m(u, v)$ defined via the equations (53)–(55).

It is the triple $(\mathscr{V}_H, \langle \cdot, \cdot \rangle, \mathscr{V}_H^0)$ from which the physical Hilbert space is obtained via factorization and completion $(\mathscr{H}_{\text{phys}}, \langle \cdot, \cdot \rangle) = (\overline{\mathscr{V}_H / \mathscr{V}_H^0}, \langle \cdot, \cdot \rangle)$ and the isometric representations on $\mathscr{V}_H$ [respectively, η-isometric ones in $\mathscr{H}(m)$] are lifted to unitary representations in $\mathscr{H}_{\text{phys}}$.

A direct way to the physical Hilbert space and the unitary representations of the (rest) symmetry groups in it which shortcuts the introduction of the vector space $\mathscr{V}_H$ is to take the direct integral of the completed factor spaces $(\overline{H(m)/H_0(m)}, \langle \cdot, \cdot \rangle_m)$ over the orbits $\mathscr{O}_{\mathscr{L}}$:

$$\mathscr{H}'_{\text{phys}} = \int_{\otimes} d\mu_{\mathscr{L}}(m)\overline{H(m)/H_0(m)},$$

with its scalar product again given by the equation (61) [where $\Psi(m)$ and $\Phi(m)$ now denote classes from $\overline{H(m)/H_0(m)}$]. $\mathscr{H}'_{\text{phys}}$ is unitary equivalent to $\mathscr{H}_{\text{phys}}$ as it follows from Theorem 3.3 in Ref. 23.

In the case of the Lorentz orbits $\mathscr{O}_{\mathscr{L}} = \mathscr{O}_{\alpha}^{\pm}$, given explicitly in Sec. III due to the direct integral structure required by Poincaré invariance, there are only three real parameters α, β, γ left free in the theory: β may be fixed to be zero by the barion mass spectrum, α interpreted as the coupling constant or strength of the flavor interaction, and γ be considered as the strength of the color confinement.

V. ONE-PARTICLE HADRON STATES

For the remainder of our investigations and especially the construction of one-particle hadron states, a spectrum condition on the fields and translation operators has to be specified that ensures that the physical Hilbert space $\mathscr{H}_{\text{phys}}$, respectively, the ''physical section'' $\mathscr{V}_H \backslash \mathscr{V}_H^0$ of the inner product space $\mathscr{V}_H$, contains only states carrying non-negative energy and non-negative momentum (mass) squares.

Let $\mathscr{D}_0(m) := L.H.\{\Pi_{j=1}^{L} \varphi^{n_j \cdots}_{\ldots, c_j, \ldots}(f_j)_m \Psi_0(m) \in \mathscr{D}(m) | Q_{\mathscr{C}}(L) = \Sigma_{j=1}^{L}\, Q(c_j) = 0, L \in \mathbb{N}_0\}$ be the linear hull of all states from $\mathscr{D}(m)$ with total color charge zero. Then the restriction $\mathscr{T}_m(y)\lceil \mathscr{D}_0(m)$ satisfies the preconditions of Theorem 4.2 in Ref. 9, which states that $\mathscr{T}_m(y)\lceil \mathscr{D}_0(m)$ is isometric with respect to the Hilbert scalar product $(\cdot|\cdot)_m$ for every $y \in \mathbb{R}_4$ and therefore has a unique unitary extension $\mathscr{T}_m^0(y)$ onto the closure $\overline{\mathscr{D}_0(m)}$. Being a unitary operator on the Hilbert subspace $(\overline{\mathscr{D}_0(m)}, (\cdot|\cdot)_m)$, it allows a spectral resolution $\mathscr{T}_m^0(y) = \int_{\mathbb{R}_4} \exp[i(q \cdot y)] dE_m(q)$ with a bounded measure E_m. Now, the *spectrum condition* reads as follows: *For μ-almost all m the support of E_m is contained in the closed forward cone,*

$$\text{supp } E_m \subseteq \overline{V_+^0} = \{q \in \mathbb{R}_4 | q^2 \geqslant 0 \wedge q^0 \geqslant 0\}.$$

The spectrum condition implies

$$\int_{\mathbb{R}_4} dy h(y) (\Psi(m) \Big| \mathscr{T}_m^0(y)\Phi(m))_m = (2\pi)^{3/2} \int_{\mathbb{R}_4} \tilde{h}(-q) d(\Psi(m) | E_m(q)\Phi(m)_m = 0, \quad (64)$$

for $\Phi(m) \in \overline{\mathscr{D}_0(m)}$, μ-almost all m, and all $h \in S(\mathbb{R}_4)$, whose Fourier transforms,

$$\tilde{h}(q) = (2\pi)^{-3/2} \int d^4y \, \exp[-i(q \cdot y)] h(y), \quad (65)$$

vanish in a neighborhood of the closed backward cone $\overline{V_-^0} = \{q \in \mathbb{R}_4 | q^2 \geqslant 0 \wedge q^0 \leqslant 0\}$.

In quantum field theories with unitary translation operators, the states of discrete mass κ arise from a δ function of the form $\delta(q^2 - \kappa^2)$ in the spectral support of the translation operators. For the translation operators at hand, the occurrence of a state ψ_κ with discrete mass κ may be characterized via the existence of a sufficiently small neighborhood $\mathscr{N}_\kappa$ of the branch of the hyperboloid $q^2 = \kappa^2$ in the backward cone V_-^0, such that for all $\Phi(m) \in \mathscr{H}(m)$ and almost all $m \in \mathscr{O}_{\mathscr{L}}$,

$$\int d^4y \langle \Phi(m), \mathscr{T}_m(y)\Psi_\kappa(m)\rangle_m (\Box_y + \kappa^2) h_\kappa(y) = 0, \quad (66)$$

for all $h_\kappa \in S(\mathbb{R}_4)$ with supp $\tilde{h}_\kappa \subseteq \mathscr{N}_\kappa$, but

$$\int d^4y h_\kappa(y) \langle \Psi_\kappa(m), \mathscr{T}_m(y)\Psi_\kappa(m)\rangle_m \neq 0, \quad (67)$$

for at least one $h_\kappa \in S(\mathbb{R}_4)$ with supp $\widetilde{h}_\kappa \subseteq \mathscr{N}_\kappa$.

Once the existence of discrete mass states is established by the spectrum condition, the explicit construction of the corresponding one-particle states in $\mathscr{V}_H \backslash \mathscr{V}_H^0$, which means the relativistic color singlet states with definite spin, (flavor-)isospin I and (flavor-)hypercharge Y satisfying the Klein–Gordon, respectively, Dirac equation, etc., with mass parameter $\kappa(I,Y)$ is easily performed, applying suitable linear combinations of the composite fields (44), respectively, (45) or the currents (43) with suitably restricted test functions to the cyclic state $\Psi_0(m)$ in $H(m)$. Since this construction is well known from the formation of so-called ''one-particle excitation operators'' in the Haag–Ruelle scattering theory of classical Wightman quantum field theories (see, e.g., Refs. 29, 32 and references therein) we may drop all details. Furthermore, we restrict ourselves to the meson states especially (axial-)vector and (pseudo)scalar states. The barion states are obtained from the three-fold products of quark fields (45) in the same way as the meson states from the two-fold ones (44).

(Pseudo)scalar mesons: In order to obtain (pseudo)scalar meson multiples we take for $\mu_i \in \{0,...,3\}$, $c_i \in \Gamma_3$, $a_i \in \Gamma_3$, $\alpha \in \Gamma_d$ and $d \in \{1,8\}$ the special functions

$$h^{[\mu_1\mu_2,c_1,c_2,a_1,a_2]\lambda}_{p_{\mathscr{M}}(d,\lambda),0\pm,d}(x_1,x_2) = (-1)^{Q(c_1)+Q(a_1)-Q(\lambda)} \begin{pmatrix} 3^* & 3 & 1 \\ -c_1 & c_2 & 0 \end{pmatrix} \times \begin{pmatrix} 3^* & 3 & d \\ -a_1 & a_2 & -\lambda \end{pmatrix} E^{\mu_1\mu_2}_{0\pm} h^{0\pm,d}_{\kappa(I(\lambda),Y(\lambda))}(x_1,x_2,p_{\mathscr{M}}(d,\lambda)), \quad (68)$$

with E_{0+} the unit matrix $E_{0-} = \gamma_5$ and $h^{0\pm,d}_{\kappa(\ldots)}(x_1,x_2,q)$ a Lorentz-invariant function from $S(\mathbb{R}_{12})$ and consider the states created in $\mathscr{D}_H(m)$ from the cyclic state $\Psi_0(m)$ by the translated twofold composite field (44) smeared with these special functions:

$$\begin{aligned} &\Psi^\lambda(x,m,h^{0\pm,d}_{\kappa(\ldots)}) \\ &:= \mathscr{T}_m(x)\varphi^{3^*3,3^*3}_{\mu_1\mu_2,c_1c_2,a_1a_2}(h^{[\mu_1\mu_2,c_1c_2,a_1a_2]\lambda}_{p_{\mathscr{M}}(d,\lambda),0\pm,d})_m \Psi_0(m) \\ &= (-1)^{Q(c_1)+Q(a_1)-Q(\lambda)} \begin{pmatrix} 3^* & 3 & 1 \\ -c_1 & c_2 & 0 \end{pmatrix} \begin{pmatrix} 3^* & 3 & d \\ -a_1 & a_2 & -\lambda \end{pmatrix} \\ &\quad \times \int d^4x_1 d^4x_2 h^{0\pm,d}_{\kappa(\ldots)}(x_1-x,x_2-x,p_{\mathscr{M}(d,\lambda)}) \exp[-i(x\cdot p_{\mathscr{M}}(d,\lambda))] \widetilde{\varphi}^{3,3}_{\mu_1,c_1,a_1}(x_1)_m \\ &\quad \times E^{\mu_1\mu_2}_{0\pm} \varphi^{3,3}_{\mu_2,c_2,a_2}(x_2)_m \Psi_0(m). \end{aligned} \quad (69)$$

For the second equality the transformation properties (50)–(52) and the relation

$$R^{3^*}(v)^{b_1}_{a_1} R^3(v)^{b_2}_{a_2} \begin{pmatrix} 3^*3d \\ a_1a_2\lambda \end{pmatrix} = R^d(v)^\rho_\lambda \begin{pmatrix} 3^*3d \\ b_1b_2\rho \end{pmatrix} \quad (70)$$

have been used together with the fact that $R^d(\exp -i/2(\wp(\mathscr{M})\cdot x))$ is a diagonal matrix. The color exponential of Eq. (51) has disappeared due to the first Clebsch–Gordan coefficient in Eq. (68), which at the same time ensures that these states are color singlets and thus escape the fate of confinement.

By straightforward, however, tedious Clebsch–Gordan calculations, it follows from the transformation properties of the quark fields, the Lorentz invariance of $h^{\cdots}_{\cdots}$ together with the equation (70) and the special structure (55) of the representation R^d restricted to the subgroup SU(2)×U(1) that the states $\Psi^\lambda(x,m,h^{\cdots}_{\cdots})$ transform under the QCD (rest) symmetry group $ISL(2,\mathbb{C}) \times G(3)_{\mathscr{C}} \times G(3)_{\mathscr{F}}$, according to

$$\begin{aligned}&\mathscr{U}((\alpha,y),u,v)_m\Psi^{\lambda}(x,m,h^{0\pm,d}_{\kappa(I(\lambda),Y(\lambda))})\\&\quad:=\mathscr{T}_{\Lambda(\alpha)m}(y)\mathscr{L}(\alpha)_{m'}\mathscr{T}_m(u,v)\Psi^{\lambda}(x,m,h^{0\pm,d}_{\kappa(I(\lambda),Y(\lambda))})\\&\quad=R^d(v)^{\lambda}_{\rho}\Psi^{\rho}(\Lambda(\alpha)x+y,\Lambda(\alpha)m,h^{0\pm,d}_{\kappa(I(\rho),Y(\rho))}).\end{aligned}\tag{71}$$

Finally, taking into account only Lorentz-invariant functions $h^{0\pm,d}_{\kappa(\ldots)}(x_1,x_2,q)$ such that the support of their Fourier transform,

$$\tilde{h}^{\cdots}_{\cdots}(k_1,k_2,q)=(2\pi)^{-3}\int d^4x_1d^4x_2\ \exp[-i(x_1k_1+x_2k_2)]h^{\cdots}_{\cdots}(x_1,x_2,q)$$

in the variable k_1+k_2+q is contained in the neighborhood $\mathscr{N}_{\kappa(I(\lambda),Y(\lambda))}$ of the backward branch of the mass shell $(k_1+k_2+q)^2=\kappa(I(\lambda),Y(\lambda))^2$ specified in the spectrum condition before, then the resulting states $\Psi^{\lambda}(x,m,h^{0\pm,d}_{\kappa(\ldots)})$ are positive frequency solutions of the Klein–Gordon equation,

$$(\Box_x+\kappa(I(\lambda),Y(\lambda))^2)\Psi^{\lambda}(x,m,h^{0\pm,d}_{\kappa(I(\lambda),Y(\lambda))})=0.$$

This is most easily seen by rewriting the equation (69) in terms of Fourier transforms with respect to the space–time variables and using the spectrum condition (66). But then, also, the states generated in the inner product space $(\mathscr{V}_H,\langle\cdot,\cdot\rangle)$ by the above ones (in $\mathscr{D}_H(m)$):

$$\Psi^{0\pm,d;\lambda}_{k(I(\lambda),Y(\lambda))}(x):\mathscr{O}_{\mathscr{L}}\rightarrow\cup_{m\in\mathscr{O}_{\mathscr{L}}}\mathscr{D}_H(m),m\rightarrow\Psi^{\lambda}(x,m,h^{0\pm,d}_{\kappa(I(\lambda),Y(\lambda))})\tag{72}$$

are positive frequency solutions of the Klein–Gordon equation with mass parameter $\kappa(I(\lambda),Y(\lambda))$, transforming under the QCD (rest) symmetry group according to

$$\mathscr{U}((\alpha,y),u,v)\psi^{0\pm,d;\lambda}_{\kappa(I(\lambda),Y(\lambda))}(x)=R^d(v)^{\lambda}_{\rho}\psi^{0\pm,d;\rho}_{\kappa(I(\lambda),Y(\lambda))}(\Lambda(\alpha)x+y),\tag{73}$$

with R^d the d-dimensional reducible representation of the flavor subgroup SU(2)×U(1) of SU(3) explicitly given in Eq. (55). Hence, they describe (pseudo)scalar, color singlet and SU(2)×U(1) flavor one-particle multiplets with isospin $I(\lambda)$, hypercharge $Y(\lambda)$, and mass $\kappa(I(\lambda),Y(\lambda))$.

(Axial-) vector mesons: The (axial-)vector mesons could be obtained in the same way as before from the two-fold composite quark field by special choices of (pseudo-)vectorial test functions. However, since it is rather complicated to build in the tranversality condition, it is much simpler to use the conserved flavor (axial-)vector currents $j^{8,\mu,V(A)}_{a,\mathscr{F}}(f)_m$ of Eq. (43) for the flavor octets, which *a priori* satisfy this condition due to current conservation.

Let $h_{\kappa(I(\lambda),Y(\lambda))}(x,q)$ be a Lorentz-invariant function from $S(\mathbb{R}_8)$ and the support of its Fourier transform $\tilde{h}(k,p_{\mathscr{M}}(8,\lambda))$ in the variable $k+p_{\mathscr{M}}(8,\lambda)$ be contained in the neighborhood $\mathscr{N}_{\kappa(I(\lambda),Y(\lambda))}$ of the backward branch of the mass shell $(k+p_{\mathscr{M}}(8,\lambda))^2=\kappa(I(\lambda),Y(\lambda))^2$. Then in exactly the same way as before, it follows that the states,

$$\begin{aligned}\Psi^{\nu,\lambda}_{8,V(A)}(x,m,h_{\kappa(I(\lambda),Y(\lambda))}):&=\mathscr{T}_m(x)j^{8,\nu,V(A)+}_{\lambda,\mathscr{F}}(h_{\kappa(I(\lambda),Y(\lambda))})_m\Psi_0(m)\\&=(-1)^{Q(\lambda)}\int d^4x h_{\kappa(I(\lambda),Y(\lambda))}(x,p_{\mathscr{M}}(8,\lambda))\\&\quad\times\exp[-i(x\cdot p_{\mathscr{M}}(8,\lambda)]j^{8,\nu,V(A)}_{-\lambda}(x)_{m,\mathscr{F}}\Psi_0(m),\end{aligned}\tag{74}$$

are positive frequency solutions of the Klein–Gordon equation with mass parameter $\kappa(I(\lambda),Y(\lambda))$ satisfying the transversality condition

$$\partial_\nu\Psi^{\nu,\lambda}_{8,V(A)}(x,m,h_{\kappa(I(\lambda),Y(\lambda))})=0,$$

and transforming under the QCD (rest) symmetry group according to

$$\mathscr{U}((\alpha,y),u,v)_m \Psi^{\nu,\lambda}_{8,V(A)}(x,m,h_{\kappa(I(\lambda),Y(\lambda))}) = R^8(v)^\lambda_\rho \Lambda(\alpha^{-1})^\nu_\mu \Psi^{\mu,\rho}_{8,V(A)}(\Lambda(\alpha)x + y, \Lambda(\alpha)m, h_{\kappa(I(\lambda),Y(\lambda))}). \tag{75}$$

But then the states $\psi^{8,V(A);\mu,\lambda}_{\kappa(I(\lambda),Y(\lambda))}(x)$ arising from them in $\mathscr{V}_H \backslash \mathscr{V}^0_H$ share the analogous properties. Hence they describe colorless (pseudo)vector SU(2)×U(1) flavor one-particle octets with spin 1, isospin $I(\lambda)$, hypercharge $Y(\lambda)$, and mass $\kappa(I(\lambda),Y(\lambda))$.

VI. GENERALIZED GOLDSTONE THEOREMS

In this section we prove the generalization of the Goldstone Theorem for the spontaneously broken symmetries to the general representations of the translation group. The proof rests on a Jost–Lehmann–Dyson representation for the commutator of the conserved currents $j^{8,\mu}_{a,\dots}(f)_{m,\dots}$ with the elements of the polynomial field algebra $\pi(\varphi_m, j_m)$ that incorporates locality, spectrum condition, and current conservation. Besides the ingredient that the flavor currents create from $\Psi_0(m)$ only states in $H(m)$ and thus due to the spectrum condition (64) only states with non-negative energy and non-negative four momentum squares, a locality or causality condition is needed. Each current is assumed to commute with all basic fields and currents (inclusively itself) at space-like separated points (*relative Einstein causality*):

$$[j^{8,\mu}_{a,\dots}(x)_{m,\dots}\varphi^{\dots}_{\dots}(y)_m]_- = [j^{8,\mu}_{a,\dots}(x)_{m,\dots}, j^{\dots}_{\dots}(y)_{m,\dots}]_- = 0, \quad \text{for} \quad (x-y)^2 < 0. \tag{76}$$

Finally, let $f_R(|\mathbf{x}|)$ be from $S(\mathbb{R}_3)$, where $f_R(|\mathbf{x}|)$, besides being equal to 1 for $0 \leqslant |\mathbf{x}| \leqslant R$ and zero for $|\mathbf{x}| > R + \Delta R$, satisfies the bound $|Rd/dr f_R(r)| \leqslant c < +\infty$ for all $R > 0$. Then the Goldstone Theorem[13,21,26–29] reads as follows.

(Generalized Flavor-) Goldstone Theorem: The flavor symmetry generated by the conserved current $j^{8,\mu}_{a,\dots}(h)_{m,\mathscr{F}}$ is not spontaneously broken, that is

$$\left\langle \Psi_0(m), \int d^3\mathbf{x} f_R(|\mathbf{x}|)[j^{8,0}_{a,\dots}(x)_{m,\mathscr{F}}\phi(f)]_- \Psi_0(m) \right\rangle_m = 0, \tag{77}$$

for all $|x^0| < \epsilon, \phi(f) \in \pi(\varphi_m, j_m)$ with test functions f of compact support and all $R > R(\epsilon, f)$, unless there exists a particle of mass $\kappa(a)^2 = p_{\mathscr{M}}(8,a)^2$ created from $\Psi_0(m)$ by $j^{8,0}_{a,\dots}(f)_{m,\mathscr{F}}$, that is, a state $\Psi_{\kappa(a)}(m) \in H(m)$ of discrete mass $\kappa(a)$ with $\langle \Psi_{\kappa(a)}(m), j^{8,0}_{a,\dots}(0,f_R)_{m,\mathscr{F}} \Psi_0(m) \rangle_m \neq 0$ for almost all $m \in \mathscr{O}_{\mathscr{L}}$.

Proof: Following the pattern of Wightman,[33] Vladimirov,[34] and especially the version of Araki, Hepp, and Ruelle,[32] it can be shown that in consequence of Einstein causality, spectrum condition, and translational covariance, the vacuum expectation value of the commutator may be represented by the following Jost–Lehmann–Dyson representation:

$$\langle \Psi_0(m), [j^{8,\mu}_{a,\dots}(x)_{m,\mathscr{F}}, \phi(f)]_- \Psi_0(m) \rangle_m = \exp\{-i(p(a)x)\} \int d^3\mathbf{u} \int_0^\infty ds [\rho^\mu_a(\mathbf{u},s,m)_1 \times \Delta_s(x^0, \mathbf{x}-\mathbf{u}) + \rho^\mu_a(\mathbf{u},s,m)_2 (\partial/\partial x^0) \Delta_s(x^0, \mathbf{x}-\mathbf{u})], \tag{78}$$

with $p(a) := p_{\mathscr{M}}(8,a)$ and

$$\Delta_s(x) = \frac{1}{i(2\pi)^3} \int d^4k \, \exp\{-i(kx)\} \epsilon(k^0) \delta(k^2 - s) \tag{79}$$

being the solution of the initial value problem

$$(\Box_x + s)\Delta_s(x) = 0, \quad \Delta_s(x)_{|x^0=0} = 0, \quad \frac{\partial}{\partial x^0} \Delta_s(x)_{|x^0=0} = -\delta(\mathbf{x}), \quad s \geqslant 0. \tag{80}$$

The structure of this representation is determined by the spectrum condition (77) via a solution of a five-dimensional wave equation in configuration space. The spectral functions $\rho^{\dots}_{\dots}(\mathbf{u},s,m)\dots$ are

strongly decreasing distributions in the variable s and $\mathscr{C}^\infty$-functions in $\mathbf{u}$ with compact support as a consequence of Einstein causality. The exponential factor in Eq. (78) arises from translational covariance.

Remark: The unitarity of the translation operator in $\mathscr{H}(m)$ used in Ref. 32 is not a necessary precondition to establish the structure of the representation (78). It can be dropped for the price of some gymnastics in (tempered) distribution theory along the lines of Ref. 35 and the exploitation of the unitarity of $\mathscr{T}_m^0$ on $(\overline{\mathscr{D}_0(m)},(\cdot|\cdot)_m)$.

The most important point for the following is now the application of current conservation $\partial_\mu j^{8,\mu}_{a,\ldots}(x)_{m,\mathscr{T}}=0$. By straightforward calculations it leads to the following two relations between the spectral functions:

$$\rho_a^0(\mathbf{u},s,m)_1=ip_0(a)\rho_a^0(\mathbf{u},s,m)_2+\left(\frac{\partial}{\partial u^j}+ip_j(a)\right)\rho_a^j(\mathbf{u},s,m)_2,$$

$$\left(\frac{\partial}{\partial u^j}+ip_j(a)\right)[\rho_a^j(\mathbf{u},s,m)_1+ip_0(a)\rho_a^j(\mathbf{u},s,m)_2]=(\Delta\mathbf{u}-s+p_0(a)^2)\rho_a^0(\mathbf{u},s,m)_2. \tag{81}$$

Here and in the following, summation over the Latin double index j from 1 to 3 is understood. Introducing instead of $\rho_a^j(\mathbf{u},s,m)_1$ the new spectral function $\sigma_a^j(\mathbf{u},s,m)$ via

$$\rho_a^j(\mathbf{u},s,m)_1+ip_0(a)\rho_a^j(\mathbf{u},s,m)_2=-\left(\frac{\partial}{\partial u^j}-ip^j(a)\right)\rho_a^0(\mathbf{u},s,m)_2+\sigma_a^j(\mathbf{u},s,m), \tag{82}$$

the second of the equations (81) reduces to the much simpler one,

$$(s-p(a)^2)\rho_a^0(\mathbf{u},s,m)_2=\left(\frac{\partial}{\partial u^j}+ip_j(a)\right)\sigma_a^j(\mathbf{u},s,m), \tag{83}$$

with the most general solution,[36]

$$\rho_a^0(\mathbf{u},s,m)_2=\left(\frac{\partial}{\partial u^j}+ip_j(a)\right)(s-p(a)^2)^{-1}\sigma_a^j(\mathbf{u},s,m)+\delta(s-p(a)^2)E_a(\mathbf{u},m). \tag{84}$$

$E_a(\mathbf{u})$ is a $\mathscr{C}^\infty$-function with compact support since all other spectral functions share these properties. Feeding the equations (82) and (84) together with the first one of the system (81) back into the representation (78) and integrating by parts over the variables u_j, we end up with the following representation for the zeroth component of the commutator matrix element:

$$\begin{aligned}&\langle\Psi_0(m),[j^{8,0}_{a,\ldots}(x)_{m,\mathscr{T}},\phi(f\ )]_-\Psi_0(m)\rangle_m\\&\quad=\exp\{-i(p(a)x)\}\int d^3\mathbf{u}\Bigg\{E_a(\mathbf{u},m)\left(ip_0(a)+\frac{\partial}{\partial x^0}\right)\Delta_{p(a)^2}(x^0,\mathbf{x}-\mathbf{u})\\&\qquad-\int_0^\infty ds\left[\sigma_a^j(\mathbf{u},s,m)_1+\sigma_a^j(\mathbf{u},s,m)_2\,\frac{\partial}{\partial x^0}\right]\left(\frac{\partial}{\partial x^j}-ip_j(a)\right)\Delta_s(x^0,\mathbf{x}-\mathbf{u})\Bigg\}.\end{aligned} \tag{85}$$

Here the abbreviations,

$$\sigma_a^j(\mathbf{u},s,m)_1:=\rho_a^j(\mathbf{u},s,m)^2+ip_0(a)(s-p(a)^2)^{-1}\sigma_a^j(\mathbf{u},s,m),$$

$$\sigma_a^j(\mathbf{u},s,m)_2:=(s-p(a)^2)^{-1}\sigma_a^j(\mathbf{u},s,m), \tag{86}$$

have been used and for the partial integration the compact support of the spectral functions. By means of a further partial integration we get, from Eq. (85),

$$\left\langle \Psi_0(m), \int d^3\mathbf{x} f_R(|\mathbf{x}|)[j^{8,0}_{a,\ldots}(x)_{m,\mathscr{F}}, \phi(f\ )]_- \Psi_0(m) \right\rangle_m - \int d^3\mathbf{u} E_a(\mathbf{u},m) \int d^3\mathbf{x} f_R(|\mathbf{x}|)$$

$$\times \exp\{-i(p(a)\cdot x)\}\left(ip_0(a) + \frac{\partial}{\partial x^0}\right)\Delta_{p(a)^2}(x^0, \mathbf{x}-\mathbf{u})$$

$$= \int d^3\mathbf{x}\left(\frac{\partial}{\partial x^j} f_R(|\mathbf{x}|)\right)\exp\{-i(p(a)\cdot x)\}$$

$$\times \int d^3\mathbf{u} \int_0^\infty ds \left[\sigma^j_a(\mathbf{u},s,m)_1 + \sigma^j_a(\mathbf{u},s,m)_2 \frac{\partial}{\partial x^0}\right]\Delta_s(x^0, \mathbf{x}-\mathbf{u}). \tag{87}$$

Now, the further arguments are the same as in the standard case. For all $|x^0|$ smaller than some $\epsilon > 0$ the right-hand side vanishes for all R larger than some fixed $R_0(\epsilon)$ as a consequence of the support properties of the spectral functions and the commutator function $\Delta_s(x)$, since the derivative of f_R is unequal to zero only in the ring region between R and $R+\Delta R$. By the same reasons, the second term on the left-hand side becomes independent from f_R for $R > R_0(\epsilon)$, that is, only the ball $|\mathbf{x}| \leqslant R_0(\epsilon)$ contributes to the x integrations. Hence, the only contribution to the left-hand side of (77) comes from the one-particle intermediate state in the commutator with the internal quantum numbers $a=(I,I^3,Y)$ of the current and the mass $p(a)^2$. □

From the initial values (80) we obtain for $x^0=0$ and all $R > R_0(\epsilon)$,

$$\left\langle \Psi_0(m), \int d^3\mathbf{x} f_R(|\mathbf{x}|)[j^{8,0}_{a,\ldots}(x)_{m,\mathscr{F}}, \phi(f\ )]_- \Psi_0(m) \right\rangle_{m|x^0=0}$$

$$= -\int d^3\mathbf{u} E_a(\mathbf{u},m)\exp[i(\mathbf{p}(a)\cdot\mathbf{u})] = -\widetilde{E}_a(\mathbf{p}_{\mathscr{M}}(a,8),m). \tag{88}$$

The remaining considerations rest essentially on the identity

$$\langle \Psi_0(m), [\mathscr{T}_m(y) j^{8,0}_{a,\ldots}(x^0, f_R)_{m,\mathscr{F}} \mathscr{T}_m(-y), \phi(f\ )]_- \Psi_0(m)\rangle_m$$

$$= \langle \Psi_0(m), [j^{8,0}_{a,\ldots}(x^0, f_R)_{m,\mathscr{F}}, \mathscr{T}_m(y)\phi(f\ )\mathscr{T}_m(-y)]_- \Psi_0(m)\rangle_m, \tag{89}$$

which is easily obtained from the η_m isometry of the translation operators and the invariance of $\Psi_0(m)$ under the translations [Eq. (59)].

Let $\pi(\varphi_m, j_m)_{SL}$ denote the subalgebra of $\pi(\varphi_m, j_m)$ defined by taking only test functions of compact support in the smeared fields and $\mathscr{D}(m)_{SL} \subset \mathscr{D}(m)$ the corresponding dense subspace of $\mathscr{H}(m)$. Then, in the same way as in the case of a Wightman theory[27,37,38] for every $y \in \mathbb{R}_4$, an octet of current generated charges may be defined on $\mathscr{D}(m)_{SL}$ by

$$\mathbf{Q}^{\cdots}_a(y)_{m,\mathscr{F}}\phi(f\ )\Psi_0(m) := [\mathscr{T}_m(y) j^{8,0}_{a,\ldots}(f_\epsilon, f_R)_{m,\mathscr{F}} \mathscr{T}_m(-y), \phi(f\ )]_{-|R\rangle R(\epsilon,f\ )}\Psi_0(m). \tag{90}$$

Here f_ϵ is an arbitrary $\mathscr{C}^\infty$-function with support contained in $[-\epsilon,\epsilon]$ and normalized according to $\int dx^0 f_\epsilon(x^0) = 1$.

Due to the Goldstone Theorem, the invariance of $\Psi_0(m)$ under η_m, the η_m-isometry of the translation operators and the relations (47)–(49), it follows by literally the same arguments as in Ref. 27.

(I) $\mathbf{Q}^{\cdots}_a(y)_{m,\mathscr{F}}$ is a well-defined linear operator on $\mathscr{D}(m)_{SL}$ independent of the details of f_ϵ and f_R if and only if the equation (77) of the Goldstone Theorem holds. Therefore, due to the equations (30)–(32) and (36) the vector charges $\mathbf{Q}^V_a(y)_{m,\mathscr{F}}$, $a \in \Gamma_8$ and (unless the π and η mesons have zero mass) the axial charges $\mathbf{Q}^A_a(y)_{m,\mathscr{F}}$ with $a \in \{(0,0,0),(1,0,0),(1,+/-1,0)\}$ exist as *bona fide* linear operators, since no discrete scalar particles, respectively, no massless pseudoscalar particles, are known in nature, whereas the axial vector charges $\mathbf{Q}^A_a(y)_{m,\mathscr{F}}$ with $a \in \{(1/2,1/2,+/-1),(1/2,-1/2,+/-1)\}$ do not exist as linear operators due to the presence of the finite mass pseudoscalar K mesons. Hence the chiral SU(3) symmetry is spontaneously broken.

(II) The charge operators $\mathbf{Q}^{\cdots}_a(y)_{m,\mathscr{F}}$ are closable, η_m-self-conjugate and translational covariant operators mapping $\mathscr{D}(m)_{SL}$ into itself,

$$\mathbf{Q}_a^{\cdots}(y)^*_{m,\mathscr{F}}\eta_m\Psi=\eta_m(-1)^{Q(a)}\mathbf{Q}^{\cdots}_{-a}(y)_{m,\mathscr{F}}\Psi,$$

$$\mathscr{T}_m(y)\mathbf{Q}_a^{\cdots}(x)_{m,\mathscr{F}}\mathscr{T}_m(-y)=\mathbf{Q}_a^{\cdots}(x+y)_{m,\mathscr{F}},$$

$$\mathbf{Q}_a^{\cdots}(y)_{m,\mathscr{F}}\mathscr{D}(m)_{SL}\subseteq\mathscr{D}(m)_{SL}. \tag{91}$$

Up to this point, besides the spectrum condition, relative Einstein causality, and translational covariance, only current conservation and the nonexistence of discrete scalar particles, respectively, *massless* pseudoscalar particles have been exploited. In order to proceed we have to tighten the bonds between the chiral currents $j_a^{8,\mu,h}$, $h\in\{L,R\}$ and the Noether currents arising from the flavor transformations in Lagrangian field theory (Sec. II) by imposing the commutation relations,[27,28] which in Lagrangian field theory define the infinitesimal flavor transformations.

For $a\in\hat{\Gamma}_8:=\{a\in\Gamma_8|Q(a)\geqslant 0\}$, $h\in\{L,R\}$ let $j^{\cdots h}_{a+/-}:=1/2i^{1/2+/-1/2}(j_a^{\cdots h}+/-j^{\cdots h}_{-a})$ if $Q(a)=1$ and $j^{\cdots h}_{a+/-}:=1/2i^{1/2-/+1/2}(j_a^{\cdots h}+/-j^{\cdots h}_{-a})$ if $Q(a)=0$ denote the η_m-symmetric chiral flavor currents and $\tau^a_{+/-}$ the same Hermitian combinations of the traceless self-conjugate τ matrices. Then the following commutation relations are assumed to hold for $a\in\hat{\Gamma}_8$, h, $h'\in\{L,R\}$, $\Psi\in\mathscr{D}(m)_{SL}$ and supp f compact:

$$[j^{8,0,h}_{c+/-}(f_\epsilon,f_R)_{m,\mathscr{F}},\varphi^{n,d,h'}_{\cdots,b}(f)_m]_{-|R\rangle R(\epsilon,f)}\Psi=-i\,\delta_{hh'}\,\frac{d}{dt}\,R^d(\exp[i/2t\tau^c_{+/-}])^b_a\varphi^{n,d,h}_{\cdots,b}(f)_{m|t=0}\Psi, \tag{92}$$

and the same ones with $\varphi^{\cdots,h}_{\cdots}$ replaced by the chiral flavor currents $[(n,d)=(1,8)]$, respectively, the color currents $[(n,d)=(8,1)]$ or gluon fields $\varphi^{8,1}_{\cdots}$.

The equations (92) are a mathematically consistent and careful formulation of the content of the commutation relations that follow on the formal Lagrangian level by blindly applying equal-time canonical commutation relations. The corresponding commutation relations with the composite fields (44), (45) are easily derived from the above ones.

Now it follows by a straightforward calculation from the equations (50)–(52) and (92) that the charges $\mathbf{Q}^h_a(x)_{m,\mathscr{F}}$ commute with the translations if and only if $[\wp(\mathscr{M}),\tau_a]_-=0$, with $\wp(\mathscr{M})$ explicitly given in Eq. (40), that is, for $a\in\{(1,\pm 1,0),(1,0,0),(0,0,0)\}$. In other words, the charges representing the Lie generators of the $\mathrm{SU}(2)_{\mathscr{F}}\times\mathrm{U}(1)_{\mathscr{F}}$ chiral subgroup and only these are constant in space and time:

$$\mathbf{Q}^h_a(0)_{m,\mathscr{F}}=\mathbf{Q}^h_a(x)_{m,\mathscr{F}},\quad \text{for } a\in\{(1,\pm 1,0),(1,0,0),(0,0,0)\},h\in\{L,R\}. \tag{93}$$

Due to the commutation relations (92) and the analogous ones for the currents and composite fields, the charges operate on $\mathscr{D}(m)_{SL}$ via finite-dimensional matrices acting on the space of indices. Therefore $\mathscr{D}(m)_{SL}[\eta_m\mathscr{D}(m)_{SL}]$ is a dense set of analytic vectors for the charge operators $\mathbf{Q}^{\cdots}_{a+/-}(0)_{m,\mathscr{F}}[\mathbf{Q}^{\cdots}_{a+/-}(0)^*_{m,\mathscr{F}}]$, since the exponential power series of them (respectively, of the finite-dimensional matrices corresponding to them) converge on $\mathscr{D}(m)_{SL}$ in the Hilbert norm of $\mathscr{H}(m)$,[27,39] giving rise to one-parameter groups of η_m-isometric linear operators that map $\mathscr{D}(m)_{SL}[\eta_m\mathscr{D}(m)_{SL}]$ into itself and can be lifted to unitary ones on $\overline{H(m)/H_0(m)}$:

$$I_{a+/-}(t)_{m,\mathscr{F}}\phi(f)\Psi_0(m):=\sum_{n=0}^{\infty}\frac{1}{n!}\,(it\mathbf{Q}^{\cdots}_{a+/-}(0)_{m,\mathscr{F}})^n\phi(f)\Psi_0(m),$$

$$I^*_{a+/-}(t)_{m,\mathscr{F}}\eta_m\phi(f)\Psi_0(m)=I_{a+/-}(-t)^*_{m,\mathscr{F}}\eta_m\phi(f)\Psi_0(m)=\eta_m I_{a+/-}(t)_{m,\mathscr{F}}\phi(f)\Psi_0(m),$$

$$t\in\mathbb{R},\phi(f)\in\pi(\varphi_m,j_m)_{SL}. \tag{94}$$

In view of the commutation relations (92), the flavor charge operators $\mathbf{Q}^{\cdots}_{a+/-}(0)_{m,\mathscr{F}}$ form a representation $\mathscr{Q}_m$ of the Lie algebra $\mathscr{G}(3)_{\mathscr{F}}=(\mathrm{su}(2)\oplus\mathrm{u}(1))_L\oplus(\mathrm{su}(2)\oplus\mathrm{u}(1))_R$ or $\mathscr{G}(3)_{\mathscr{F}}=\mathrm{su}(3)$ of the chiral flavor group $G(3)_{\mathscr{F}}=(\mathrm{SU}(2)\times\mathrm{U}(1))_L\times(\mathrm{SU}(2)\times\mathrm{U}(1))_R$, respectively, $G(3)_{\mathscr{F}}=\mathrm{SU}(3)$ on $\mathscr{D}(m)_{SL}$. The final step in our program is the integration or exponentiation of this

representation, which means to establish the existence of a η_m-isometric representation $I_m^{\mathscr{F}}$ of the flavor group $G(3)_{\mathscr{F}}$ on $\mathscr{D}(m)\supset\mathscr{D}(m)_{SL}$, such that the restriction of its differential $\partial I_m^{\mathscr{F}}$ to $\mathscr{D}(m)_{SL}$ is equal to $\mathscr{Q}_m$,

$$\partial I_m^{\mathscr{F}}(\tau_{+/-}^a)\lceil\mathscr{D}(m)_{SL}=\mathbf{Q}_{a+/-}^{\cdots}(0)_{m,\mathscr{F}}. \tag{95}$$

The details are given only for the vector charges $\mathbf{Q}_{a+/-}^V(0)_{m,\mathscr{F}}$ and $G(3)_{\mathscr{F}}=\mathrm{SU}(3)$. The construction for the chiral case runs literally the same way. If η_m commutes with $\mathbf{Q}_a^V(0)_{m,\mathscr{F}}$, then the one-parameter groups (94) consist of strongly continuous operators on $\mathscr{D}(m)_{SL}$ and hence admit a unique unitary extension onto $\mathscr{H}(m)$. According to criterions of Simon[40] and Jorgensen,[41] this establishes the integrability of $\mathscr{Q}_m$. The difficulties arising in the general case rest in the fact that the charge operators $\mathbf{Q}_{a+/-}^V(0)_{m,\mathscr{F}}$ are not symmetric with respect to the Hilbert scalar product $(\cdot|\cdot)$ and therefore do not establish the strong continuity of the operators $I_{a+/-}^{(*)}(t)_{m,\mathscr{F}}$. Hence, none of the classical integrability criterions (known to the author) is directly applicable.

However, due to the commutation relations (92), the representation $\mathscr{Q}_m$ on $\mathscr{D}(m)_{SL}$ is directly connected with a representation of $su(3)$ in the underlying countably normed test function space. For the domain, $\mathscr{D}(m)$ is the linear hull of states created from $\Psi_0(m)$ by polynomials of the Wightman–Gårding fields[11] with respect to the flavor group:

$$\phi_{\mu,c}^{n,d}(\mathbf{f})_m:=\sum_{b\in\Gamma_d}\phi_{\mu,c,b}^{n,d}(f^b)_m, \tag{96}$$

with $\phi_{\cdots}^{\cdots}$ denoting any one of the basic fields or currents and $f^b\in S(\mathbb{R}_4)$. The corresponding expressions for the composite fields are obvious. For the sake of shortness we do not write them down explicitly, neither here nor in the following. Now the equations (90)–(92) imply for all f_i^b with compact support and $a\in\hat{\Gamma}_8$,

$$\begin{aligned}&\mathbf{Q}_{a+/-}^V(0)_{m,\mathscr{F}}\prod_{i=1}^k\phi_{\mu_i,c_i}^{n_i,d_i}(f_i)_m\Psi_0(m)\\&\quad=\sum_{j=1}^k\left(\prod_{i=1}^{j-1}\phi_{\mu_i,c_i}^{n_i,d_i}(\mathbf{f}_i)_m\right)\phi_{\mu_j,c_j}^{n_j,d_j}(\delta^{d_j}(\tau_{+-/}^a)\mathbf{f}_j\Bigg)_m\prod_{\ell=j+1}^k\phi_{\mu_\ell,c_\ell}^{n_\ell,d_\ell}(\mathbf{f}_\ell)_m\Psi_0(m),\end{aligned} \tag{97}$$

where δ^d is defined for all $\mathbf{f}=(f^1,\ldots,f^d)\in S(\mathbb{R}_4,\mathbb{C}_d)$ by

$$(\delta^d(\tau_{+/-}^a)\mathbf{f})^e:=\sum_{b\in\Gamma_d}i\frac{d}{dt}R^d(\exp-i/2\tau_{+/-}^at)_b^ef_{|t=0}^b. \tag{98}$$

Obviously the $\delta^d(\tau_{+/-}^a)$ as defined in Eq. (98) make up a basis for a d-dimensional representation Δ^d of su(3) on the test function space $S(\mathbb{R}_4,\mathbb{C}_d)$ equipped with the countable set of norms $\|\mathbf{f}\|_n^d$, $n\in\mathbb{N}_0$ derived from the scalar products $[\mathbf{f},\mathbf{f}]_n^d=\Sigma_{a\in\Gamma_d}[f^a,f^a]_n$. The representatives once again act via d-dimensional matrices only on the space of indices. Therefore every element of $S(\mathbb{R}_4,\mathbb{C}_d)$ is an analytic vector (with respect to every norm) for the representation Δ^d. In addition, $\delta^d(\tau_{+/-}^a)$ is skew symmetric with respect to every scalar product $[\mathbf{f},\mathbf{f}]_n^d$, $n\in\mathbb{N}_0$. But then due to the general theory of unbounded representations of symmetry groups in gauge quantum field theories,[23,24] there exists a unique strongly continuous representation $R^d:\mathrm{SU}(3)\to\mathrm{Aut}\,S(\mathbb{R}_4,\mathbb{C}_d)$, $v\to R^d(v)$ on the countably normed space $S(\mathbb{R}_4,\mathbb{C}_d)$, the differential ∂R^d of which is equal to Δ^d (Ref. 42, Sec. IV, Theorem 4.1) and moreover an η_m-isometric representation $I_m^{\mathscr{F}}:\mathrm{SU}(3)\to\mathrm{Aut}\,\mathscr{D}(m)$, $v\to I_m^{\mathscr{F}}(v)$ on $\mathscr{D}(m)$ with $I_m^{\mathscr{F}}(v)=I_m(1,v)$ explicitly given in the equations (53), (54) and satisfying the condition (95) (Ref. 42, Theorem 3.1 and 5.1).

Summing up we have arrived at the following result.

Theorem (Flavor Symmetry): *The 16 conserved chiral flavor currents* $j_a^{S,\mu,h}(f)_{m,\mathscr{F}}$, $a\in\Gamma_8$, $h\in\{L,R\}$ *generate on* $\mathscr{D}(m)$ *a* η_m*-isometric representation [unitary one on* $(H(m)/H_0(m),\langle\cdot,\cdot\rangle_m)$, *respectively,* $\mathscr{H}_{\mathrm{phys}}$*] of the chiral (sub)group* $(\mathrm{SU}(2)\times\mathrm{U}(1))_L\times(\mathrm{SU}(2)\times\mathrm{U}(1))_R$ *that commutes with the translation operators and a further* η_m*-isometric one of*

the flavor group SU(3) for which only its restriction to the subgroup $\mathrm{SU}(2)\times\mathrm{U}(1)$ *(isomorphic to the diagonal subgroup of the chiral one) commutes with the translations.*

In order to extend the investigation above to the color group, it has to be minded that the spectrum condition (64) holds only in the closure $\overline{\mathscr{D}_0(m)}$ of the subspace of uncolored states. Therefore the Jost–Lehmann–Dyson representation (78) cannot be established anymore for the colored currents $j_c^{8,\mu}(f)_{m,\mathscr{C}}$ with $Q(c)\neq 0$. But the four color currents affiliated to the (algebraically unbroken) color group $G(3)_{\mathscr{C}}\sim\mathrm{U}(1)\times\mathrm{SU}(2)$ carry the color charge $Q(c)=0$. Then the generalized Goldstone Theorem and its consequences for color transformations are obtained in literally the same way as above just by replacing everywhere the flavor current by the color current and $p(a)=p_{\mathscr{M}}(8,a)$ by zero.

(Generalized Color-) Goldstone Theorem: The color symmetry generated by the conserved current $j_a^{8,\mu}(h)_{m,\mathscr{C}}$ *with* $Q(a)=0$ *is not spontaneously broken, that is,*

$$\left\langle \Psi_0(m), \int d^3\mathbf{x} f_R(|\mathbf{x}|)[j_a^{8,0}(x)_{m,\mathscr{C}}, \phi(f)]_- \Psi_0(m)\right\rangle_m = 0, \tag{99}$$

for all $|x^0|<\epsilon$, $\phi(f)\in\pi(\varphi_m, j_m)_{SL}$ *and all* $R>R(\epsilon, f)$, *unless there exists a state* $\Psi_{\kappa(a)}(m)\in\mathscr{H}(m)$ *with mass* $\kappa(a)=0$ *and* $\langle\Psi_{\kappa(a)}(m), j_a^{8,0}(f)_{m,\mathscr{C}}\Psi_0(m)\rangle_m\neq 0$ *for almost all* $m\in\mathscr{O}_{\mathscr{D}}$.

In other words, in all representations, uncolored Goldstone states with zero mass can exist. If uncolored ones occur at all, they may appear in any part of $\mathscr{H}(m)$. If, however, a local gauge condition of the type (26) holds in $\mathscr{H}(m)$, then also the uncolored Goldstone states are confined in the unphysical section $(\mathscr{H}(m)\backslash H(m))\cup H_0(m)$.

Similarly to Wightman theories there may be further restrictions concerning the spin- or helicity-quantum numbers of the Goldstone states.[28,43,44] However, in view of the indefinite metric and the direct integral structure of the representation of the Lorentz group this is a problem of its own beyond the present note.

[1] J. F. Cornwell, *Group Theory in Physics* (Academic, London, 1984), Vol. II.
[2] T. D. Lee, *Particle Physics and Introduction to Field Theory* (Harwood Akademic, New York, 1988).
[3] F. J. Yndurian, *The Theory of Quark and Gluon Interactions* (Springer-Verlag, Berlin, 1993).
[4] D. Flamm and F. Schröberl, *Introduction to the Quark Model of Elementary Particles* (Gordon and Breach, London, 1982).
[5] J. J. de Swart, Rev. Mod. Phys. **35**, 916 (1963).
[6] J. J. de Swart, Rev. Mod. Phys. **37**, 326 (1965).
[7] J. J. de Swart, *Symmetries of Strong Interactions*, in 1966 CERN School of Physics, Noordwijk-aan-Zee, The Netherlands, 1966.
[8] T. Kugo and I. Ojima, Prog. Theor. Phys. **66**, 1 (1979).
[9] A. H. Völkel, J. Math. Phys. **26**, 1057 (1985).
[10] E. L. Hill, Rev. Mod. Phys. **23**, 253 (1951).
[11] A. S. Wightman and L. Gårding, Ark. Fys. **28**, 129 (1964).
[12] F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974).
[13] F. Strocchi, Commun. Math. Phys. **56**, 57 (1977).
[14] P. J. M. Bongaarts, J. Math. Phys. **18**, 1510 (1977).
[15] P. J. M. Bongaarts, J. Math. Phys. **23**, 1881 (1982).
[16] G. Morchio and F. Strocchi, Ann. Inst. Henri Poincar´e **XXIII**, 251 (1980).
[17] R. Ferrari, L. E. Picasso, and F. Strocchi, Commun. Math. Phys. **35**, 25 (1974).
[18] J. P. Antoine, Acta Univ. Wratislav **519**, 226 (1978).
[19] J. P. Antoine and C. Malou, in *Proceedings of the 6th International Colloquium on Group Theoretical Methods in Physics*, edited by P. Kramer and A. Rieckers (Springer, Berlin, 1978).
[20] M. Minchev and E. d'Emilio, J. Math. Phys. **22**, 1267 (1981).
[21] M. Kh. Minchev and I. T. Todorov, Sov. J. Part. Nucl. **16**, 26 (1985).
[22] F. Strocchi, Phys. Rev. D **17**, 2010 (1978).
[23] A. H. Völkel, J. Math. Phys. **26**, 2956 (1985).
[24] M. Gell-Mann, Phys. Rev. **125**, 1067 (1962).
[25] S. Okubo, Prog. Theor. Phys. **27**, 949 (1962).
[26] D. Robinson, ''Symmetries, broken symmetries, currents and charges,'' in *Symmetry Principles and Fundamental Particles*, edited by B. Kursunoglu and P. Perlmuttei (Freeman, San Francisco, 1967).
[27] J. A. Swieca, ''Goldstone theorem and related topics,'' in *Gargese Lectures in Physics*, edited by D. Kastler (Gordon and Breach, New York, 1970), Vol. 4.
[28] H. Reeh, ''Symmetries, currents and infinitesimal generators,'' in *Statistical Mechanics and Field Theory*, edited by R. N. Sen and C. Weil (Halsted, New York, 1972).
[29] A. H. Völkel, *Fields, Particles and Currents*, Lecture Notes in Physics (Springer, Berlin, 1977), Vol. 66.
[30] I. M. Gelfand and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1964), Vol. 4.
[31] I. M. Gelfand and G. E. Shilow, *Generalized Functions* (Academic, New York, 1967/68), Vols. 1–3.

[32] K. Hepp, ''On the connection between Wightman and LSZ quantum field theory,'' in *Particle Symmetries and Axiomatic Field Theory*, edited by M. Chretien and S. Deser (Gordon and Breach, New York, 1965), Vol. 1.
[33] A. S. Wightman, ''Analytic functions of several complex variables,'' in *Relations de Dispersion et Particules Elementaires*, Les Houches Lectures 1960, edited by C. DeWitt and R. Omnes (Hermann, Paris, 1960).
[34] V. S. Vladimirov, *Methods of the Theory of Functions of Many Complex Variables* (MIT Press, Cambridge, MA, 1966).
[35] A. H. Völkel, Phys. Rev. D **1**, 3377 (1970); and ''Cluster-Theorem and collision states in QCD'' (preprint to Berlin, Mathematik·serie A, in print).
[36] L. Schwartz, *Theorie des Distributions* (Hermann, Paris, 1957/5–9), Vols. I and II.
[37] W. D. Garber and H. Reeh, J. Math. Phys. **19**, 59 (1978).
[38] W. D. Garber and H. Reeh, J. Math. Phys. **19**, 985 (1978).
[39] J. A. Swieca, Docent thesis, University of Sao Paulo, 1967.
[40] J. Simon, Commun. Math. Phys. **28**, 39 (1972).
[41] P. E. T. Jorgensen, Exp. Math. **4**, 289 (1983).
[42] A. H. Völkel, J. Math. Phys. **27**, 1113 (1986).
[43] D. Maison and H. Reeh, Commun. Math. Phys. **24**, 67 (1971).
[44] R. N. Sen and C. Weil, Nuovo Cimento A **6**, 581 (1971).

Wave scattering in one dimension with absorption

Tuncay Aktosun
Department of Mathematics, North Dakota State University, Fargo, North Dakota 58105

Martin Klaus
Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Cornelis van der Mee
Department of Mathematics, University of Cagliari, Cagliari, Italy

(Received 13 January 1997; accepted for publication 19 November 1997)

Wave scattering is analyzed in a one-dimensional nonconservative medium governed by the generalized Schrödinger equation $d^2\psi/dx^2+k^2\psi=[ikP(x)+Q(x)]\psi$, where $P(x)$ and $Q(x)$ are real, integrable potentials with finite first moments. Various properties of the scattering solutions are obtained. The corresponding scattering matrix is analyzed, and its small-k and large-k asymptotics are established. The bound states, which correspond to the poles of the transmission coefficient in the upper-half complex plane, are studied in detail. When the medium is not purely absorptive, i.e., unless $P(x)\leqslant 0$, it is shown that there may be bound states at complex energies, degenerate bound states, and singularities of the transmission coefficient imbedded in the continuous spectrum. Some explicit examples are provided illustrating the theory. © *1998 American Institute of Physics.* [S0022-2488(98)01503-5]

I. INTRODUCTION

Wave propagation in a one-dimensional nonconservative medium is described, in the frequency domain, by the generalized Schrödinger equation

$$\psi^{+\prime\prime}(k,x)+k^2\psi^+(k,x)=[ikP(x)+Q(x)]\psi^+(k,x), \quad x\in\mathbf{R}, \tag{1.1}$$

where $\mathbf{R}$ is the real line, the prime denotes the derivative with respect to the spatial coordinate x, k is the wave number (also known as the momentum), k^2 is the energy, $P(x)$ describes the combined effect of energy absorption and energy generation, and $Q(x)$ denotes the restoring force density. In the time domain, (1.1) corresponds to a wave equation of the form

$$\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial t^2}-P(x)\,\frac{\partial u}{\partial t}=Q(x)u, \quad t,x\in\mathbf{R},$$

where the wave speed is equal to 1. When $P(x)\leqslant 0$, there is net absorption; however, unless otherwise stated we will not put any restriction on the sign of $P(x)$. In the sequel a significant role will be played by the associated equation

$$\psi^{-\prime\prime}(k,x)+k^2\psi^-(k,x)=[-ikP(x)+Q(x)]\psi^-(k,x), \quad x\in\mathbf{R}, \tag{1.2}$$

where the sign of $P(x)$ in (1.1) has been changed.

Let $L^p_q(I)$ denote the measurable functions $f(x)$ such that $\int_I dx\ (1+|x|)^q|f(x)|^p<+\infty$, and let $L^p(I)=L^p_0(I)$. Throughout the paper we will use $\|f\|_1$ and $\|f\|_{1,1}$ to denote the $L^1(\mathbf{R})$ and $L^1_1(\mathbf{R})$ norms, $\int_{-\infty}^{\infty}dx\ |f(x)|$ and $\int_{-\infty}^{\infty}dx\ [1+|x|]|f(x)|$, respectively. All the results given in this paper are valid if we assume that P and Q are real valued and belong to $L^1_1(\mathbf{R})$. The existence of the first moment is needed for certain results that involve the limit $k\to 0$. Some results, however, will be proved under the weaker conditions $P,Q\in L^1(\mathbf{R})$.

 © 1998 American Institute of Physics

The scattering solutions of (1.1) and (1.2) are those behaving like e^{ikx} or e^{-ikx} as $x\to\pm\infty$, and such solutions occur when $k^2>0$. Among the scattering solutions are the Jost solution from the left $f_l^{\pm}(k,x)$ and the Jost solution from the right $f_r^{\pm}(k,x)$ satisfying the boundary conditions

$$f_l^{\pm}(k,x)=\begin{cases} e^{ikx}+\mathrm{o}(1), & x\to+\infty,\\ \dfrac{1}{T^{\pm}(k)}\,e^{ikx}+\dfrac{L^{\pm}(k)}{T^{\pm}(k)}\,e^{-ikx}+\mathrm{o}(1), & x\to-\infty,\end{cases} \tag{1.3}$$

$$f_r^{\pm}(k,x)=\begin{cases} \dfrac{1}{T^{\pm}(k)}\,e^{-ikx}+\dfrac{R^{\pm}(k)}{T^{\pm}(k)}\,e^{ikx}+\mathrm{o}(1), & x\to+\infty,\\ e^{-ikx}+\mathrm{o}(1), & x\to-\infty,\end{cases} \tag{1.4}$$

where $T^{\pm}(k)$ are the transmission coefficients, and $R^{\pm}(k)$ and $L^{\pm}(k)$ are the reflection coefficients from the right and from the left, respectively. The scattering matrices $\mathbf{S}^{+}(k)$ associated with (1.1) and $\mathbf{S}^{-}(k)$ with (1.2) are given by

$$\mathbf{S}^{\pm}(k)=\begin{bmatrix} T^{\pm}(k) & R^{\pm}(k)\\ L^{\pm}(k) & T^{\pm}(k)\end{bmatrix}.$$

When $P(x)\leqslant 0$, it will be seen that $\mathbf{S}^{+}(k)$ exists for all $k\in\mathbf{R}$; however, when $P(x)\geqslant 0$ or when $P(x)$ has mixed sign, we will see that $\mathbf{S}^{+}(k)$ may not exist at $k=0$ or at some other real values of k.

In this paper we analyze the direct scattering problem in preparation of a more detailed study of various inverse scattering problems for (1.1). One of these inverse problems consists of the recovery of $P(x)$ and $Q(x)$ from an appropriate set of scattering data. In the radial case, when there are no bound states, Jaulent and Jean[1] presented an inversion method with real $Q(x)$ and imaginary $P(x)$. They[2,3] also extended their method to solve the full-line one-dimensional inverse problem for real $Q(x)$ and imaginary $P(x)$. In this method, using the scattering data $\{R^{+}(k),R^{-}(k)\}$, a pair of two coupled Marchenko integral equations is solved and these solutions are used in a first-order ordinary differential equation whose solution leads to $P(x)$. Jaulent[4] also extended this method to the case when $P(x)$ is real although complete details and proofs were not given. When $P(x)$ is purely imaginary and $\int_{-\infty}^{\infty}dz\ P(z)=0$, Sattinger and Szmigielski[5] showed that one can simplify the method of Jaulent and Jean and recover $P(x)$ by solving an algebraic equation rather than a differential equation.

We should also mention the study by Kaup[6] on the direct and inverse scattering problem for

$$\phi''+\left[k^2+\frac{1}{4\beta^2}\right]\phi=[ikP(x)+Q(x)]\phi, \tag{1.5}$$

where β is a nonzero constant and $P,Q\in L_1^1(\mathbf{R})$. Under certain additional assumptions on $P(x)$, Tsutsumi[7] studied the scattering problem for (1.5) with $\beta=\frac{1}{2}$ using a 2×2 matrix analog of (1.1) with k replaced by $\sqrt{k^2+1}$. When P and Q are in the Schwartz space, $\int_{-\infty}^{\infty}dz\ P(z)=0$, and $\beta=\frac{1}{2}$, Sattinger and Szmigielski[8] also studied the inverse scattering problem for (1.5). In Refs. 6 and 8 the inverse scattering problem is analyzed by studying a Riemann–Hilbert problem on a particular Riemann surface.

When $P(x)$ is purely imaginary, the methods available for self-adjoint differential operators can be employed to analyze the inverse scattering problem for (1.1); furthermore, in this case[2] the scattering matrices $\mathbf{S}^{\pm}(k)$ are unitary, and hence the reflection coefficients cannot exceed 1 in absolute value. However, when $P(x)$ is real valued, the differential operator pertaining to (1.1) is no longer self-adjoint and the scattering matrices $\mathbf{S}^{\pm}(k)$ are no longer unitary. Consequently, the analysis of the direct and inverse scattering problems with real $P(x)$ is different and more difficult than with imaginary $P(x)$. The standard proof[9,10] of the absence of singularities of the transmission coefficient for $k\in\mathbf{R}$, which relies heavily on the self-adjointness of the differential operator, breaks down. Since the reflection coefficients may be larger than one in absolute value, the standard proof of the unique solvability of the Marchenko integral equations is no longer valid. Fortunately, when $P(x)\leqslant 0$, some of the usual properties of the one-dimensional Schrödinger

equation given in (2.6), such as the simplicity of the poles of the transmission coefficient, the confinement of these poles to the imaginary axis in the upper-half complex plane $\mathbf{C}^+$, and the absence of singularities of the transmission coefficient when $k \in \mathbf{R}$ are still valid for (1.1), and the proofs of such properties are obtained by a variation of the arguments used for (2.6).

This paper is organized as follows. In Secs. II and III, relying on techniques established in a variety of papers,[10–12] we study the analyticity properties of the Jost solutions of (1.1) and analyze their small-k and large-k asymptotics. In Secs. IV–VI we analyze various properties of the scattering matrices $\mathbf{S}^{\pm}(k)$. In Sec. VII we study the change in the scattering coefficients when $P(x)$ and $Q(x)$ are perturbed. In Sec. VIII we analyze the relation between the poles of $T^+(k)$ in $\mathbf{C}^+$ and the bound states; we also study multiple poles of $T^+(k)$ in terms of Jordan chains of the differential operator given in (8.3). Recall that the bound state solutions of (1.1) and (1.2) are those nontrivial solutions belonging to $L^2(\mathbf{R})$. In the radial case when $P(x) \leqslant 0$, under certain additional conditions on $P'(x)$, using the theory of abstract operator polynomials, Pivovarchik has shown that[13] the number of bound states is independent of $P(x)$ and that[14–16] the bound states are simple and can only occur when k is located on the positive imaginary axis. In Sec. IX we study the bound states for (1.1) further and show that the poles of $T^+(k)$ in $\mathbf{C}^+$ can only occur in a certain region in $\mathbf{C}^+$ determined by $P(x)$ and $Q(x)$. When $P(x) \leqslant 0$, we derive Pivovarchik's results in an elementary way without using the theory of abstract operator polynomials and without assuming the differentiability of $P(x)$; we also show that the bound states can only occur at certain negative energies and obtain some lower and upper bounds for these energies. In Sec. IX we also obtain a Levinson theorem relating the number of bound states to the change in the argument of $T^+(k)$, and we show that the number of bound states is unchanged under certain small perturbations of P and Q. In Sec. X we analyze the zeros of the Jost solutions and obtain various results concerning the number and location of these zeros and their relationship to the bound states; we also show that the number of bound states of (1.1) with real energies is greater than or equal to the number of bound states with $P(x) = 0$. In Sec. XI, we show by examples that there may be bound states with complex energies and that the multiplicity of a bound state (in the sense of the order to which $1/T^+(k)$ vanishes) may be larger than one. Finally, in the Appendix we obtain various small-k estimates that are needed in the proof of Theorem 5.2.

II. ANALYTICITY AND SMALL-k ASYMPTOTICS OF JOST SOLUTIONS

Let us define the Faddeev functions from the left, $m_l^{\pm}(k,x)$, and from the right, $m_r^{\pm}(k,x)$, by

$$m_l^{\pm}(k,x) = e^{-ikx} f_l^{\pm}(k,x), \quad m_r^{\pm}(k,x) = e^{ikx} f_r^{\pm}(k,x).$$

Then $m_l^{\pm}(k,x)$ satisfies

$$m_l^{\pm}(k,x) = 1 + \frac{1}{2ik} \int_x^{\infty} dy \; [e^{2ik(y-x)} - 1][\pm ikP(y) + Q(y)] m_l^{\pm}(k,y), \tag{2.1}$$

$$m_l^{\pm\prime}(k,x) = -\int_x^{\infty} dy \; e^{2ik(y-x)} [\pm ikP(y) + Q(y)] m_l^{\pm}(k,y). \tag{2.2}$$

By $\overline{\mathbf{C}^+}$ we denote $\mathbf{C}^+ \cup \mathbf{R}$. In the following theorem and throughout the paper we will use C to denote a generic constant (independent of x and k) that does not necessarily assume the same value at each appearance.

Theorem 2.1: (i) Assume $P, Q \in L^1(\mathbf{R})$. Then, for each $x \in \mathbf{R}$, the functions $m_l^{\pm}(k,x)$, $m_r^{\pm}(k,x)$, $m_l^{\pm\prime}(k,x)$, and $m_r^{\pm\prime}(k,x)$ are analytic in $\mathbf{C}^+$ and continuous in $\overline{\mathbf{C}^+} \backslash \{0\}$. Consequently, for each $x \in \mathbf{R}$ the Jost solutions $f_l^{\pm}(k,x), f_r^{\pm}(k,x)$ and their derivatives $f_l^{\pm\prime}(k,x), f_r^{\pm\prime}(k,x)$ are analytic in $\mathbf{C}^+$ and continuous in $\overline{\mathbf{C}^+} \backslash \{0\}$. Moreover, for each $k \in \overline{\mathbf{C}^+} \backslash \{0\}$ we have

$$|m_l^{\pm}(k,x)| \leqslant Ce^{C/|k|}, \quad |m_r^{\pm}(k,x)| \leqslant Ce^{C/|k|}, \tag{2.3}$$

$$m_l^{\pm}(k,x) = 1 + o(1), \quad m_l^{\pm\prime}(k,x) = o(1), \quad x \to +\infty,$$

$$m_r^{\pm}(k,x) = 1 + o(1), \quad m_r^{\pm\prime}(k,x) = o(1), \quad x \to -\infty.$$

(ii) Assume $P \in L^1(\mathbf{R})$ and $Q \in L_1^1(\mathbf{R})$. Then the continuity of the functions in (i) extends to $\overline{\mathbf{C}^+}$. Moreover, for $k \in \overline{\mathbf{C}^+}$ we have

$$|m_l^{\pm}(k,x)| \leqslant C[1 + \max\{0,-x\}], \quad |m_r^{\pm}(k,x)| \leqslant C[1 + \max\{0,x\}], \tag{2.4}$$

$$m_l^{\pm\prime}(k,x) = o(1/x), \quad x \to +\infty; \quad m_r^{\pm\prime}(k,x) = o(1/x), \quad x \to -\infty.$$

Proof: The proof is obtained from (2.1), (2.2), and similar equations for $m_r^{\pm}(k,x)$ in a manner analogous[10] to the case with $P(x) = 0$. In the proof of (ii), one also uses the estimates

$$|1 - e^{2ik(y-x)}| \leqslant 2, \quad |1 - e^{2ik(y-x)}| \leqslant 2|k|(y-x),$$

for $k \in \overline{\mathbf{C}^+}$ and $y \geqslant x$. ∎

When $P \in L^1(\mathbf{R})$ and $Q \in L_1^1(\mathbf{R})$, using (2.1), (2.2), and Theorem 2.1 (ii), we obtain

$$m_l^{\pm}(k,x) = m_l^{\pm}(0,x) + o(1), \quad m_l^{\pm\prime}(k,x) = m_l^{\pm\prime}(0,x) + o(1), \quad k \to 0 \text{ in } \overline{\mathbf{C}^+}, \tag{2.5}$$

uniformly on compact x-intervals. Let us consider the Schrödinger equation obtained from (1.1) and (1.2) by setting $P(x) = 0$, namely

$$\psi^{[0]\prime\prime}(k,x) + k^2\psi^{[0]}(k,x) = Q(x)\psi^{[0]}(k,x), \quad x \in \mathbf{R}. \tag{2.6}$$

Let $f_l^{[0]}(k,x)$ and $f_r^{[0]}(k,x)$ denote the Jost solutions of (2.6) from the left and from the right, respectively. From (1.1) and the corresponding boundary conditions we see that $f_l^{\pm}(0,x), f_r^{\pm}(0,x)$, and their derivatives are determined by $Q(x)$ alone and

$$m_l^{\pm}(0,x) = f_l^{\pm}(0,x) = f_l^{[0]}(0,x), \quad m_l^{\pm\prime}(0,x) = f_l^{\pm\prime}(0,x) = f_l^{[0]\prime}(0,x), \tag{2.7}$$

As seen from (2.6) and (2.7) we have

$$Q(x) = \frac{f_l^{\pm\prime\prime}(0,x)}{f_l^{\pm}(0,x)} = \frac{f_l^{[0]\prime\prime}(0,x)}{f_l^{[0]}(0,x)}.$$

Let $\mathbf{S}^{[0]}(k)$ denote the scattering matrix associated with (2.6):

$$\mathbf{S}^{[0]}(k) = \begin{bmatrix} T^{[0]}(k) & R^{[0]}(k) \\ L^{[0]}(k) & T^{[0]}(k) \end{bmatrix},$$

where $T^{[0]}(k)$ is the transmission coefficient and $R^{[0]}(k)$ and $L^{[0]}(k)$ are the reflection coefficients from the right and from the left, respectively. Generically $f_l^{\pm}(0,x)$ and $f_r^{\pm}(0,x)$ are linearly independent, but in the so-called exceptional case these two functions are linearly dependent. We have[9,10]

$$[f_l^{[0]}(0,x); f_r^{[0]}(0,x)] = \int_{-\infty}^{\infty} dy\, Q(y) f_l^{[0]}(0,y) = \int_{-\infty}^{\infty} dy\, Q(y) f_r^{[0]}(0,y) = \lim_{k \to 0} \frac{-2ik}{T^{[0]}(k)}, \tag{2.8}$$

where $[f;g] = fg' - f'g$ denotes the Wronskian. Thus $T^{[0]}(0) = 0$ generically and $T^{[0]}(0) \neq 0$ in the exceptional case. In the exceptional case, let us define

$$\gamma = \frac{f_l^{[0]}(0,x)}{f_r^{[0]}(0,x)}. \tag{2.9}$$

Then γ is a nonzero real constant determined by $Q(x)$ alone, and we have $\gamma=f_l^{[0]}(0,-\infty)$.

III. LARGE-k ASYMPTOTICS OF JOST SOLUTIONS

In this section we analyze the large-k asymptotics of the Jost solutions. We assume that $P\in L^1(\mathbf{R})$ and $Q\in L^1_1(\mathbf{R})$. The results given here will be used in Sec. VI to obtain the large-k asymptotics of the scattering matrix $\mathbf{S}^+(k)$.

Let us define

$$\eta_l^{\pm}(k,x)=e^{\pm\zeta}m_l^{\pm}(k,x)=e^{-ikx\pm\zeta}f_l^{\pm}(k,x),$$
$$\eta_r^{\pm}(k,x)=e^{\pm p\mp\zeta}m_r^{\pm}(k,x)=e^{ikx\pm p\mp\zeta}f_r^{\pm}(k,x), \tag{3.1}$$

where

$$\zeta=\zeta(x)=\frac{1}{2}\int_x^{\infty}dz\ P(z),\quad p=\frac{1}{2}\int_{-\infty}^{\infty}dz\ P(z), \tag{3.2}$$

so that $\int_{-\infty}^{x}dz\ P(z)/2=p-\zeta$. Thus

$$f_l^{\pm}(k,x)=e^{ikx\mp\zeta}\eta_l^{\pm}(k,x),$$
$$f_l^{\pm\prime}(k,x)=e^{ikx\mp\zeta}[(ik\pm P/2)\,\eta_l^{\pm}(k,x)+\eta_l^{\pm\prime}(k,x)], \tag{3.3}$$

$$f_r^{\pm}(k,x)=e^{-ikx\mp p\pm\zeta}\eta_r^{\pm}(k,x),$$
$$f_r^{\pm\prime}(k,x)=e^{-ikx\mp p\pm\zeta}[(-ik\mp P/2)\,\eta_r^{\pm}(k,x)+\eta_r^{\pm\prime}(k,x)]. \tag{3.4}$$

Theorem 3.1: Assume $P,Q\in L^1(\mathbf{R})$. Then, for each $x\in\mathbf{R}$, the functions $\eta_l^{\pm}(k,x)$ and $\eta_r^{\pm}(k,x)$ are analytic in $\mathbf{C}^+$, continuous in $\overline{\mathbf{C}^+}\backslash\{0\}$, and we have

$$|\eta_l^{\pm}(k,x)|\leq Ce^{C/|k|},\quad |\eta_r^{\pm}(k,x)|\leq Ce^{C/|k|},\quad k\in\overline{\mathbf{C}^+}\backslash\{0\}. \tag{3.5}$$

If $P\in L^1(\mathbf{R})$ and $Q\in L^1_1(\mathbf{R})$, then, for each $x\in\mathbf{R}$, the functions $\eta_l^{\pm}(k,x)$ and $\eta_r^{\pm}(k,x)$ are continuous in $\overline{\mathbf{C}^+}$, and we have

$$|\eta_l^{\pm}(k,x)|\leq C[1+\max\{0,-x\}],\quad |\eta_r^{\pm}(k,x)|\leq C[1+\max\{0,x\}],\quad k\in\overline{\mathbf{C}^+}. \tag{3.6}$$

Moreover, if $P,Q\in L^1(\mathbf{R})$, then

$$\eta_l^{\pm}(k,x)=1+o(1),\quad \eta_r^{\pm}(k,x)=1+o(1),\quad k\to\infty\ \text{ in }\ \overline{\mathbf{C}^+}, \tag{3.7}$$

$$\eta_l^{\pm\prime}(k,x)=o(k),\quad \eta_r^{\pm\prime}(k,x)=o(k),\quad k\to\infty\ \text{ in }\ \overline{\mathbf{C}^+}. \tag{3.8}$$

Proof: The analyticity in $\mathbf{C}^+$, the continuity in $\overline{\mathbf{C}^+}\backslash\{0\}$, and (3.5) and (3.6) follow from (3.1) and Theorem 2.1 (i). To prove (3.7) we use (3.1) in (2.1). Letting $z(k,x)=\eta_l^+(k,x)-1$, after some simplifications, we obtain

$$z(k,x)=z_0(k,x)+\frac{1}{2ik}\int_x^{\infty}dy\ [e^{2ik(y-x)}-1]e^{\zeta(x)-\zeta(y)}[ikP(y)+Q(y)]z(k,y), \tag{3.9}$$

with

$$z_0(k,x)=\frac{1}{2}\int_x^{\infty}dy\ e^{2ik(y-x)}e^{\zeta(x)-\zeta(y)}P(y)+\frac{1}{2ik}\int_x^{\infty}dy\ [e^{2ik(y-x)}-1]e^{\zeta(x)-\zeta(y)}Q(y). \tag{3.10}$$

Using (2.4) in (3.10) we get

$$|z_0(k,x)|\leqslant C_1\int_x^\infty dy\ [|P(y)|+|Q(y)|/|k|],\quad k\in\overline{\mathbf{C}^+}\backslash\{0\},$$

where C_1 is a suitable constant. For each $x\in\mathbf{R}$, let $s_x(k)=\sup_{t\geqslant x}|z_0(k,t)|$. Solving (3.9) by iteration we obtain

$$|z(k,x)|\leqslant s_x(k)e^{C_1\int_x^\infty dy\ [|P(y)|+|Q(y)|/|k|]}. \tag{3.11}$$

Applying the Riemann–Lebesgue lemma to (3.10), we get

$$s_x(k)=\mathrm{o}(1),\quad k\to\pm\infty. \tag{3.12}$$

Using (3.11) and (3.12) we see that $z(k,x)$ is uniformly bounded in $\overline{\mathbf{C}^+}$ for $|k|\geqslant a>0$ for each $x\in\mathbf{R}$ and $a>0$. Hence, in view of (3.12) and a Phragmén–Lindelöf theorem[17] we conclude that $z(k,x)\to 0$ as $k\to\infty$ in $\overline{\mathbf{C}^+}$. The proof of (3.7) for $\eta_l^-(k,x)$ and $\eta_r^\pm(k,x)$ is similar. To prove (3.8) for $\eta_l^{+\prime}(k,x)$ we introduce the functions

$$\begin{aligned}\xi_l^\pm(k,x)&=\frac{1}{2ik}\,[\pm P(x)\,\eta_l^\pm(k,x)+2\,\eta_l^{\pm\prime}(k,x)],\\ \xi_r^\pm(k,x)&=\frac{1}{2ik}\,[\mp P(x)\,\eta_r^\pm(k,x)+2\,\eta_r^{\pm\prime}(k,x)].\end{aligned} \tag{3.13}$$

Since $\xi_l^\pm(k,x)=(1/ik)\ m_l^{\pm\prime}(k,x)e^{\pm\zeta}$, it follows from (3.1) and (3.13) that

$$\xi_l^+(k,x)=-\int_x^\infty dy\ [P(y)-iQ(y)/k]e^{2ik(y-x)}e^{\zeta(x)-\zeta(y)}\eta_l^+(k,y). \tag{3.14}$$

Thus, using (3.5), we see that the integrand on the right-hand side of (3.14) is bounded by the integrable function $C_a[|P(y)|+|Q(y)|/a]$, uniformly in x and $k\in\overline{\mathbf{C}^+}$ for $|k|\geqslant a>0$ and each $a>0$, where the constant C_a only depends on a but not on x and k. By a variant of the Riemann–Lebesgue lemma, we conclude that the right-hand side of (3.14) is o(1) as $k\to\pm\infty$, so that by a Phragmén–Lindelöf theorem,[17] we see that the left-hand side of (3.14) is o(1) as $k\to\infty$ in $\overline{\mathbf{C}^+}$. Consequently, $\xi_l^+(k,x)=\mathrm{o}(1)$ as $k\to\infty$ in $\overline{\mathbf{C}^+}$, and so (3.8) for $\eta_l^{+\prime}(k,x)$ follows by using (3.13). The proof of (3.8) for $\eta_l^{-\prime}(k,x)$ and $\eta_r^{\pm\prime}(k,x)$ is similar. ■

Let us also mention that it is possible to study the large k-behavior of the solutions of (1.1) by converting it into a system of two coupled, first-order differential equations. We will not give the details here but refer the interested reader to Ref. 18.

IV. SCATTERING COEFFICIENTS

In this section we summarize some basic facts about the scattering coefficients. In particular, we are concerned with symmetries, Wronskian relations, analyticity, and continuity properties. Applications and refinements of these results will be given in the subsequent sections. Note that, since $P(x)$ is assumed to be real, $\mathbf{S}^\pm(k)$ is not unitary. However, there are certain relations (Proposition 4.3) involving the scattering coefficients which reduce to the usual unitarity relations when $P(x)=0$.

We begin with the observation that when $k\in\mathbf{R}$, the quantities $f_l^-(-k,x)$ and $f_r^-(-k,x)$ are also solutions of (1.1) and hence can be expressed as linear combinations of $f_l^+(k,x)$ and $f_r^+(k,x)$, unless the latter functions are linearly dependent. Using (1.3) and (1.4) we obtain the two vector equations

$$\begin{bmatrix}f_l^\mp(-k,x)\\ f_r^\mp(-k,x)\end{bmatrix}=\begin{bmatrix}T^\pm(k) & -R^\pm(k)\\ -L^\pm(k) & T^\pm(k)\end{bmatrix}\begin{bmatrix}f_r^\pm(k,x)\\ f_l^\pm(k,x)\end{bmatrix},\quad k\in\mathbf{R}. \tag{4.1}$$

In general, $f_l^\mp(-k,x)$ and $f_r^\mp(-k,x)$ cannot be continued to $\mathbf{C}^+$ as functions of k because $f_l^\pm(k,x)$ and $f_r^\pm(k,x)$ usually cannot be extended to the lower-half complex plane $\mathbf{C}^-$.

Using (1.3) and (1.4), the scattering coefficients can be expressed[2] in terms of Wronskians:

$$[f_l^{\pm}(k,x);f_r^{\pm}(k,x)]=-\frac{2ik}{T^{\pm}(k)},\quad k\in\overline{\mathbf{C}^+},\tag{4.2}$$

$$[f_l^{\pm}(k,x);f_r^{\mp}(-k,x)]=\frac{2ikL^{\pm}(k)}{T^{\pm}(k)}=-\frac{2ikR^{\mp}(-k)}{T^{\mp}(-k)},\tag{4.3}$$

$$[f_r^{\pm}(k,x);f_l^{\mp}(-k,x)]=-\frac{2ikR^{\pm}(k)}{T^{\pm}(k)}=\frac{2ikL^{\mp}(-k)}{T^{\mp}(-k)},\tag{4.4}$$

$$[f_l^{\pm}(k,x);f_l^{\mp}(-k,x)]=-2ik=-2ik\,\frac{1-L^{\pm}(k)L^{\mp}(-k)}{T^{\pm}(k)T^{\mp}(-k)},\tag{4.5}$$

$$[f_r^{\pm}(k,x);f_r^{\mp}(-k,x)]=2ik=2ik\,\frac{1-R^{\pm}(k)R^{\mp}(-k)}{T^{\pm}(k)T^{\mp}(-k)}.\tag{4.6}$$

Let an overline denote complex conjugation. It is already known[2] that

$$f_l^{\pm}(-\overline{k},x)=\overline{f_l^{\pm}(k,x)},\quad f_r^{\pm}(-\overline{k},x)=\overline{f_r^{\pm}(k,x)},\quad k\in\overline{\mathbf{C}^+},\tag{4.7}$$

$$f_l^{\pm}(-k,x)=\overline{f_l^{\pm}(k,x)},\quad f_r^{\pm}(-k,x)=\overline{f_r^{\pm}(k,x)},\quad k\in\mathbf{R},\tag{4.8}$$

$$\frac{1}{T^{\pm}(-\overline{k})}=\frac{1}{\overline{T^{\pm}(k)}},\quad k\in\overline{\mathbf{C}^+}\backslash\{0\},\tag{4.9}$$

$$\mathbf{S}^{\pm}(-k)=\overline{\mathbf{S}^{\pm}(k)},\quad \mathbf{S}^{\pm}(k)\mathbf{S}^{\mp}(-k)^t=\mathbf{I},\quad k\in\mathbf{R},\tag{4.10}$$

where $\mathbf{I}$ is the 2×2 unit matrix and the superscript t denotes the matrix transpose. It is understood that (4.10) holds only at the points where the scattering coefficients are defined; we will see later that $\mathbf{S}^{\pm}(k)$ may not be defined for certain real values of k.

From (4.10) we see that

$$T^{\mp}(k)=\frac{T^{\pm}(-k)}{T^{\pm}(-k)^2-L^{\pm}(-k)R^{\pm}(-k)},$$

and therefore

$$\det\mathbf{S}^{\pm}(k)=T^{\pm}(k)^2-L^{\pm}(k)R^{\pm}(k)=\frac{T^{\pm}(k)}{T^{\mp}(-k)},\quad k\in\mathbf{R}.\tag{4.11}$$

It follows from (4.9) that the zeros of $1/T^{\pm}(k)$ either lie on or are symmetrically located with respect to the imaginary axis in $\overline{\mathbf{C}^+}$; in particular, $1/T^{+}(k_0)=0$ for some $k_0\in\mathbf{R}\backslash\{0\}$ implies that $1/T^{+}(-k_0)=0$.

Proposition 4.1: Assume $P,Q\in L^1(\mathbf{R})$. Then: (i) $1/T^{\pm}(k)$ are analytic in $\mathbf{C}^+$, continuous in $\overline{\mathbf{C}^+}\backslash\{0\}$, and bounded in the sector $\{k\in\overline{\mathbf{C}^+}:|k|\geqslant a>0\}$ for every $a>0$. If, in addition, $Q\in L^1_1(\mathbf{R})$, then $k/[(k+i)T^{\pm}(k)]$ are continuous and bounded in $\overline{\mathbf{C}^+}$, and $kL^{\pm}(k)/T^{\pm}(k)$ and $kR^{\pm}(k)/T^{\pm}(k)$ are continuous and bounded in $\mathbf{R}$. Consequently, $T^{\pm}(k)$ cannot have any zeros in $\overline{\mathbf{C}^+}\backslash\{0\}$. (ii) The zeros of $k/[(k+i)T^{\pm}(k)]$ in $\mathbf{C}^+$ are all isolated, and their accumulation points, if any, must lie on the real axis or at infinity.

Proof: All the assertions follow from (4.3), (4.4), (4.9), and Theorem 2.1. ■

Note that contrary to the case where $P(x)$ is either zero or purely imaginary, we cannot *a priori* rule out possible singularities of the scattering coefficients $T^{\pm}(k)$, $R^{\pm}(k)$, and $L^{\pm}(k)$ on the real axis. In Examples 11.2 and 11.4, we show that $1/T^{+}(k)$ may have zeros on $\mathbf{R}$ or off the imaginary axis in $\mathbf{C}^+$.

Proposition 4.2: Assume $P,Q\in L^1(\mathbf{R})$. Then, for any $k_0\in\mathbf{R}\backslash\{0\}$, the quantities $1/T^+(k_0)$ and $R^\pm(k_0)/T^\pm(k_0)$ cannot be zero simultaneously; similarly, $1/T^+(k_0)$ and $L^\pm(k_0)/T^\pm(k_0)$ cannot be zero simultaneously. If $1/T^+(k_0)=0$ for some $k_0\in\mathbf{R}\backslash\{0\}$, then none of the eight quantities $R^\pm(k_0)/T^\pm(k_0)$, $L^\pm(k_0)/T^\pm(k_0)$, $R^\pm(-k_0)/T^\pm(-k_0)$, $L^\pm(-k_0)/T^\pm(-k_0)$ are zero. Moreover, $R^+(k)$ is continuous for $k\in\mathbf{R}\backslash\{0\}$ if and only if $1/T^+(k)$ does not vanish for $k\in\mathbf{R}\backslash\{0\}$. Similarly, $L^+(k)$ is continuous for $k\in\mathbf{R}\backslash\{0\}$ if and only if $1/T^+(k)$ does not vanish for $k\in\mathbf{R}\backslash\{0\}$.

Proof: By Proposition 4.1 we know that $1/T^\pm(k)$, $R^\pm(k)/T^\pm(k)$, $L^\pm(k)/T^\pm(k)$ are continuous when $k\in\mathbf{R}\backslash\{0\}$. Moreover, the right-hand sides in (4.5) and (4.6) cannot be zero when $k\neq 0$. ∎

Note that, if $1/T^+(k_0)=0$ for some $k_0\in\mathbf{R}\backslash\{0\}$, from (4.2) we see that $f_l^+(k_0,x)$ and $f_r^+(k_0,x)$ are linearly dependent, and from (1.3) and (1.4) we obtain

$$\frac{f_l^+(k_0,x)}{f_r^+(k_0,x)}=\frac{L^+(k_0)}{T^+(k_0)}=\frac{T^+(k_0)}{R^+(k_0)}.$$

Moreover, if $1/T^+(k)=c(k-k_0)^m+\mathrm{o}(k-k_0)^m$ as $k\to k_0$, for some nonzero constant c and positive integer m, then $R(k)=d(k-k_0)^{-m}+\mathrm{o}(k-k_0)^{-m}$ as $k\to k_0$ with $d\neq 0$. Thus, we see that $T^+(k)$ is continuous for $k\in\mathbf{R}\backslash\{0\}$ if and only if $R^+(k)$ [or $L^+(k)$] is continuous.

Proposition 4.3: Assume $P,Q\in L^1(\mathbf{R})$. The scattering coefficients satisfy

$$\frac{1}{|T^\pm(k)|^2}=1+\left|\frac{L^\pm(k)}{T^\pm(k)}\right|^2\mp\int_{-\infty}^{\infty}dx\ |f_l^\pm(k,x)|^2P(x),\quad k\in\mathbf{R}\backslash\{0\},\tag{4.12}$$

$$\frac{1}{|T^\pm(k)|^2}=1+\left|\frac{R^\pm(k)}{T^\pm(k)}\right|^2\mp\int_{-\infty}^{\infty}dx\ |f_r^\pm(k,x)|^2P(x),\quad k\in\mathbf{R}\backslash\{0\}.\tag{4.13}$$

Hence, if $P(x)\leqslant 0$, then $1/T^+(k)$ cannot have any zeros for $k\in\mathbf{R}$, and we have

$$|T^+(k)|^2+|L^+(k)|^2\leqslant 1,\quad |T^+(k)|^2+|R^+(k)|^2\leqslant 1,\quad k\in\mathbf{R}\backslash\{0\}.\tag{4.14}$$

If $1/T^+(k)$ does not have any zeros for $k\in\mathbf{R}\backslash\{0\}$ and $P(x)\geqslant 0$, then we have

$$|T^+(k)|^2+|L^+(k)|^2\geqslant 1,\quad |T^+(k)|^2+|R^+(k)|^2\geqslant 1,\quad k\in\mathbf{R}\backslash\{0\}.\tag{4.15}$$

Moreover, if $P(x)\leqslant 0$, then $1/|T^+(k)|\geqslant 1/|T^-(k)|$ for $k\in\mathbf{R}\backslash\{0\}$.

Proof: From (1.1) and (1.2) we obtain

$$\frac{d}{dx}[f_l^\pm(-k,x);f_l^\pm(k,x)]=\pm 2ikf_l^\pm(-k,x)f_l^\pm(k,x)P(x),\tag{4.16}$$

$$\frac{d}{dx}[f_r^\pm(-k,x);f_r^\pm(k,x)]=\pm 2ikf_r^\pm(-k,x)f_r^\pm(k,x)P(x).\tag{4.17}$$

Hence, using (1.3), (1.4), (4.8), (4.10) in (4.16) and (4.17), for $k\in\mathbf{R}\backslash\{0\}$, we obtain (4.12) and (4.13), which imply (4.14) and (4.15). The last inequality follows from (4.10) by subtracting (4.12) from (4.13). ∎

V. SMALL-*k* ANALYSIS OF SCATTERING COEFFICIENTS

In this section we analyze the small-k asymptotics of $\mathbf{S}^\pm(k)$. Our results will depend on whether we are in the generic or the exceptional case.

Proposition 5.1: Assume $P\in L^1(\mathbf{R})$ and $Q\in L^1_1(\mathbf{R})$ and suppose that we are in the generic case. Then $R^\pm(0)=L^\pm(0)=-1$, $T^\pm(k)$ vanish linearly as $k\to 0$ in $\overline{\mathbf{C}^+}$, and

$$\lim_{k\to 0}\frac{2ik}{T^+(k)}=\lim_{k\to 0}\frac{2ik}{T^-(k)}=\lim_{k\to 0}\frac{2ik}{T^{[0]}(k)}.\tag{5.1}$$

Furthermore, $\det \mathbf{S}^{\pm}(0)=-1$, and

$$T^{\pm}(k)=\frac{-2ik}{\int_{-\infty}^{\infty}dy\ Q(y)f_l^{[0]}(0,y)}+o(k),\quad k\to 0\ \text{ in }\ \overline{\mathbf{C}^+}. \tag{5.2}$$

Proof: (5.1) follows from (4.2) and Theorem 2.1 (ii). Since generically $R^{[0]}(0)=L^{[0]}(0)=-1$, from (4.3) and (4.4) in the limit $k\to 0$ we get $R^{\pm}(0)=L^{\pm}(0)=-1$. Then, using (4.11) we obtain $\det S^{\pm}(0)=-1$. From (2.8) and (5.1) we obtain (5.2). ■

Under stronger conditions on P and Q, the small-k asymptotics of the scattering coefficients in the generic case were known [see, e.g., Ref. 2 in the case with imaginary $P(x)$]. Next we study the small-k asymptotics of the scattering coefficients in the exceptional case, and we show that $S^{\pm}(0)$ is affected by $P(x)$.

Theorem 5.2: In the exceptional case, under the assumptions $P,Q\in L_1^1(\mathbf{R})$, we have

$$T^{\pm}(0)=\frac{2\gamma}{\gamma^2+1\mp\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2}, \tag{5.3}$$

$$L^{\pm}(0)=\frac{\gamma^2-1\pm\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,\ x)^2}{\gamma^2+1\mp\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2}, \tag{5.4}$$

$$R^{\pm}(0)=\frac{1-\gamma^2\pm\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2}{\gamma^2+1\mp\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2}, \tag{5.5}$$

where γ is the constant defined in (2.9).

Proof: The technical details of the proof are given in the Appendix. As in the case $P(x)=0$ the difficulty is to prove that the transmission coefficients are continuous at $k=0$. This is only straightforward if $P,Q\in L_2^1(\mathbf{R})$ as[10] in the case $P(x)=0$, but not if $P,Q\in L_1^1(\mathbf{R})$. We obtain $T^{+}(0)$ in (5.3) by using (A18) and (A19) in (A17) and also using (4.2). The value of $T^{-}(0)$ in (5.3) is obtained by changing the sign of $P(x)$. In order to obtain $L^{+}(0)$ in (5.4), as in the displayed equation following (A32) of Ref. 12, we first derive

$$\begin{aligned}
&f_l(0,0)[f_l^{+}(k,x);f_r^{-}(-k,x)]\\
&\quad=f_r^{-}(-k,0)\left[-ikf_l(0,0)+f_l'(0,0)+\int_0^{\infty}dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)\right]\\
&\qquad-f_l^{+}(k,0)\left[-ikf_l(0,0)+f_l'(0,0)-\int_{-\infty}^{0}dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)\right],
\end{aligned} \tag{5.6}$$

where $\widetilde{\psi}(k,x)$ is the function in (A2). Estimating various terms in (5.6) as in the proof of Proposition A.4 in connection with (A17), and also by using (4.3), we obtain $L^{+}(0)$ in (5.4). Similarly we obtain $R^{+}(0)$ in (5.5). The values of $L^{-}(0)$ and $R^{-}(0)$ are obtained from $L^{+}(0)$ and $R^{+}(0)$ by changing the sign of $P(x)$. ■

In the exceptional case, it is possible that $\mathbf{S}^{\pm}(k)$ is discontinuous at $k=0$, and as seen from (5.3)–(5.5) this happens if and only if $\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2=\pm(\gamma^2+1)$. For example, if $Q(x)=0$, we have $f_l^{[0]}(0,x)=f_r^{[0]}(0,x)=1$, and hence $\gamma=1$. Thus

$$\frac{1}{T^{\pm}(0)}=1\mp\frac{1}{2}\int_{-\infty}^{\infty}dx\ P(x). \tag{5.7}$$

Hence, if $\int_{-\infty}^{\infty}dx\ P(x)=\pm 2$, then $\mathbf{S}^{\pm}(0)$ is undefined.

We remark that in the exceptional case, when $\mathbf{S}^{\pm}(k)$ is continuous at $k=0$, we can obtain (5.3)–(5.5) also as follows. In (4.1) let $k\to 0$ and use (2.7) and (2.9) to get

$$\gamma=\frac{f_l^{\pm}(0,x)}{f_r^{\pm}(0,x)}=\frac{T^{\pm}(0)}{1+R^{\pm}(0)}=\frac{T^{[0]}(0)}{1+R^{[0]}0}=\frac{1+L^{[0]}(0)}{T^{[0]}(0)}=\frac{1+L^{\pm}(0)}{T^{\pm}(0)}. \tag{5.8}$$

Combining (5.8) with (4.12) for $k=0$ and eliminating $L^{\pm}(0)/T^{\pm}(0)$, we get

$$1+\gamma^2-\frac{2\gamma}{T^{\pm}(0)}=\pm\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2,$$

which gives us (5.3). Then, using (5.3) in (5.8) we obtain (5.4) and (5.5).

In the exceptional case, since $\mathbf{S}^{\pm}(0)$ is real valued, from (5.8) it is seen that $\mathbf{S}^{\pm}(0)$ is a unitary matrix if and only if $R^{\pm}(0)=-L^{\pm}(0)$. By (5.4) and (5.5), this occurs when the integral $\int_{-\infty}^{\infty}dx\ P(x)f_l^{[0]}(0,x)^2$ vanishes in which case we have $\mathbf{S}^{\pm}(0)=\mathbf{S}^{[0]}(0)$ and $\det\mathbf{S}^{\pm}(0)=1$.

From the discussion above and Theorem 5.2 we obtain:

Proposition 5.3: Assume that $P\in L^1(\mathbf{R})$ and $Q\in L_1^1(\mathbf{R})$ in the generic case and that $P,Q\in L_1^1(\mathbf{R})$ in the exceptional case. If any one of the quantities $1/T^{\pm}(k)$, $R^{\pm}(k)/T^{\pm}(k)$, $L^{\pm}(k)/T^{\pm}(k)$ is discontinuous at $k=0$, then all six are discontinuous at $k=0$ and we are in the generic case. If any one of the quantities $1/T^{\pm}(k)$, $R^{\pm}(k)/T^{\pm}(k)$, $L^{\pm}(k)/T^{\pm}(k)$ is continuous at $k=0$, then all six are continuous at $k=0$ and we are in the exceptional case. Moreover, the three quantities $T^{+}(k)$, $R^{+}(k)$, $L^{+}(k)$ [or $T^{-}(k)$, $R^{-}(k)$, $L^{-}(k)$] are all continuous on $\mathbf{R}$, or they are all discontinuous on $\mathbf{R}$.

VI. LARGE-*k* ANALYSIS OF SCATTERING COEFFICIENTS

In this section we analyze the large-k asymptotics of $\mathbf{S}^{\pm}(k)$. Similar results were obtained under stronger conditions on $P(x)$ and $Q(x)$ [see, e.g., Ref. 2 in the case with imaginary $P(x)$].

Theorem 6.1: Assume $P,Q\in L^1(\mathbf{R})$. Then

$$\frac{1}{T^{\pm}(k)}\,e^{\pm p}=1+\mathrm{o}(1),\quad k\to\infty\ \text{ in }\ \overline{\mathbf{C}^{+}}, \tag{6.1}$$

$$\frac{R^{\pm}(k)}{T^{\pm}(k)}=\mathrm{o}(1),\quad \frac{L^{\pm}(k)}{T^{\pm}(k)}=\mathrm{o}(1),\quad k\to\pm\infty, \tag{6.2}$$

where p is the constant defined in (3.2).

Proof: From (4.2) using (3.3) and (3.4), we obtain

$$\frac{2ik}{T^{\pm}(k)}\,e^{\pm p}=[2ik\pm P(x)]\,\eta_l^{\pm}(k,x)\,\eta_r^{\pm}(k,x)+\eta_l^{\pm\prime}(k,x)\,\eta_r^{\pm}(k,x)-\eta_l^{\pm}(k,x)\,\eta_r^{\pm\prime}(k,x). \tag{6.3}$$

Now (6.1) follows from (3.7), (3.8), and (6.3). Similarly, using (3.3), (3.4), and (4.4), we have

$$-\frac{2ikR^{\pm}(k)}{T^{\pm}(k)}=e^{-2ikx\mp p\pm 2\zeta}[\,\eta_r^{\pm}(k,x);\eta_l^{\mp}(-k,x)\,],\quad k\in\mathbf{R}, \tag{6.4}$$

and the first relation in (6.2) follows by using (3.7) and (3.8) in (6.4). The proof of the second relation in (6.2) is analogous, using (4.3). ■

Note that from (6.1) it follows that e^{p} is known when either of $T^{\pm}(k)$ is known. From Proposition 4.1 and Theorem 6.1, we have:

Corollary 6.2: Assume $P,Q\in L^1(\mathbf{R})$. If $1/T^{\pm}(k)$ does not vanish for $k\in\mathbf{R}$, then its number of zeros in $\mathbf{C}^{+}$ is finite.

VII. PERTURBATION OF SCATTERING COEFFICIENTS

In this section we establish the stability of $1/T^+(k)$, $R^+(k)/T^+(k)$, and $L^+(k)/T^+(k)$ under small perturbations of P and Q in the norm of either $L^1(\mathbf{R})$ or $L^1_1(\mathbf{R})$. We will not state the results for $R^+(k)/T^+(k)$, since they are identical to those for $L^+(k)/T^+(k)$. An application of these results will be given in Sec. IX.

Given two sets of potentials $P_j(x)$ and $Q_j(x)$ with $j=1,2$, we consider the generalized Schrödinger equations

$$\psi_j^{+\prime\prime}(k,x)+k^2\psi_j^+(k,x)=[ikP_j(x)+Q_j(x)]\psi_j^+(k,x), \quad x\in\mathbf{R}, \tag{7.1}$$

and denote their corresponding Faddeev solutions by $m^+_{l;j}(k,x)$ and $m^+_{r;j}(k,x)$, their transmission coefficients by $T_j^+(k)$, and their reflection coefficients from the right and from the left by $R_j^+(k)$ and $L_j^+(k)$, respectively.

Proposition 7.1: Assume $P_j, Q_j\in L^1(\mathbf{R})$ for $j=1,2$. Then for $k\in\overline{\mathbf{C}^+}$ with $|k|\geq 1$, we have

$$\left|\frac{1}{T_1^+(k)}-\frac{1}{T_2^+(k)}\right|\leq C(\|P_1-P_2\|_1+\|Q_1-Q_2\|_1), \tag{7.2}$$

and for $k\in\mathbf{R}$ with $|k|\geq 1$, we have

$$\left|\frac{L_1^+(k)}{T_1^+(k)}-\frac{L_2^+(k)}{T_2^+(k)}\right|\leq C(\|P_1-P_2\|_1+\|Q_1-Q_2\|_1). \tag{7.3}$$

Proof: First, by iterating (2.1) and using (2.3), we obtain

$$|m^+_{l;1}(k,x)-m^+_{l;2}(k,x)|\leq C_1(\|P_1-P_2\|_1+\|Q_1-Q_2\|_1), \tag{7.4}$$

for some constant C_1. Furthermore, from (1.3) and (2.1) we obtain

$$\frac{1}{T^\pm(k)}=1-\frac{1}{2ik}\int_{-\infty}^{\infty}dy\ [\pm ikP(y)+Q(y)]m_l^\pm(k,y), \tag{7.5}$$

$$\frac{L^\pm(k)}{T^\pm(k)}=\frac{1}{2ik}\int_{-\infty}^{\infty}dy\ e^{-2iky}[\pm ikP(y)+Q(y)]m_l^\pm(k,y). \tag{7.6}$$

Now using (7.4) in (7.5) we get (7.2). The proof of (7.3) is similarly obtained by using (7.4) and (7.6). ■

Proposition 7.2: Assume $P_j\in L^1(\mathbf{R})$ and $Q_j\in L^1_1(\mathbf{R})$ for $j=1,2$. Then, for $k\in\overline{\mathbf{C}^+}$ with $|k|\leq 1$, we have

$$\left|\left(\frac{k}{T_1^+(k)}-\frac{k}{T_2^+(k)}\right)\right|\leq C(\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}), \tag{7.7}$$

and for $k\in\mathbf{R}$ with $|k|\leq 1$, we have

$$\left|\left(\frac{kL_1^+(k)}{T_1^+(k)}-\frac{kL_2^+(k)}{T_2^+(k)}\right)\right|\leq C(\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}). \tag{7.8}$$

Proof: Iterating (2.1) with the help of (2.4) and the inequality

$$\frac{1+\max\{0,-y\}}{1+\max\{0,-x\}}(y-x)\leq 1+|y|, \quad y\geq x, \tag{7.9}$$

we obtain

$$|km^+_{l;j}(k,x)|\leq C_1, \quad |k|\leq 1 \tag{7.10}$$

$$|m^+_{l;1}(k,x)-m^+_{l;2}(k,x)|\leq C_2[1+\max\{0,-x\}](\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}), \tag{7.11}$$

$$|km^+_{l;1}(k,x)-km^+_{l;2}(k,x)|\leq C_3(\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}), \quad |k|\leq 1, \tag{7.12}$$

for some constants C_1, C_2, and C_3. Then (7.7) and (7.8) follow from using (2.4), (7.10)–(7.12) in (7.5) and (7.6). ■

Proposition 7.3: Assume $P_j, Q_j \in L^1_1(\mathbf{R})$ for $j=1,2$ and $Q_1(x)=Q_2(x)$. Then, we have

$$\left|\frac{1}{T^+_1(k)}-\frac{1}{T^+_2(k)}\right|\leq C\|P_1-P_2\|_{1,1}, \quad k\in\overline{\mathbf{C}^+}, \tag{7.13}$$

$$\left|\frac{L^+_1(k)}{T^+_1(k)}-\frac{L^+_2(k)}{T^+_2(k)}\right|\leq C\|P_1-P_2\|_{1,1}, \quad k\in\mathbf{R}. \tag{7.14}$$

Proof: First from (2.4), (7.9), and (2.1) we get

$$|m^+_{l;1}(k,x)-m^+_{l;2}(k,x)|\leq C_1|k|[1+\max\{0,-x\}]\|P_1-P_2\|_{1,1}, \tag{7.15}$$

for some constant C_1. Using (2.4), (7.11), and (7.15) in (7.5) and (7.6) we obtain (7.13) and (7.14). ■

VIII. BOUND STATES AND JORDAN CHAINS

Recall that the bound states of (1.1) are its nontrivial solutions belonging to $L^2(\mathbf{R})$. In this section we show that the zeros of $1/T^+(k)$ in $\mathbf{C}^+$ correspond to the bound states of (1.1). We also analyze the order of each zero of $1/T^+(k)$ in $\mathbf{C}^+$ in terms of Jordan chains of the differential operator $\mathbf{W}(k)$ defined in (8.3).

Proposition 8.1: Assume $P, Q \in L^1(\mathbf{R})$. A point $k_0 \in \mathbf{C}^+$ corresponds to a bound state of (1.1) if and only if $1/T^+(k_0)=0$. If $k_0 \in \mathbf{R}\backslash\{0\}$ and $1/T^+(k_0)=0$, then k_0 does not correspond to a bound state of (1.1); if we further assume $Q \in L^1_1(\mathbf{R})$, then $k=0$ cannot correspond to a bound state even when $1/T^+(0)=0$.

Proof: The first assertion follows from (4.2) and Theorem 2.1 (i). For $k_0 \in \mathbf{R}\backslash\{0\}$ every nontrivial solution of (1.1) has the asymptotic form $c_+e^{ik_0x}+c_-e^{-ik_0x}+\mathrm{o}(1)$ as $x\to+\infty$, with the constants c_+ and c_- not both equal to zero, and hence cannot be in $L^2(\mathbf{R})$. If $Q \in L^1_1(\mathbf{R})$, then $k_0=0$ cannot be a bound state because any nontrivial solution of (1.1) for $k=0$ has the asymptotic form $c_1x[1+o(1)]+c_2+\mathrm{o}(1)$ as $x\to+\infty$, with c_1 and c_2 not both equal to zero. Hence, the proof is complete. ■

Next we analyze multiple poles of $T^+(k)$. Let us differentiate (1.1) with $\psi=f^+_l(k,x)$ or $\psi=f^+_r(k,x)$ with respect to k repeatedly. Defining

$$g^+_{l,n}(k,x)=\frac{1}{n!}\left(\frac{\partial}{\partial k}\right)^n f^+_l(k,x), \quad g^+_{r,n}(k,x)=\frac{1}{n!}\left(\frac{\partial}{\partial k}\right)^n f^+_r(k,x), \quad n=0,1,2,..., \tag{8.1}$$

and $g^+_{l,n}(k,x)=g^+_{r,n}(k,x)=0$ for $n=-1,-2,...$, we obtain the coupled system of differential equations

$$\begin{aligned} &g^{+\prime\prime}_{l,n}(k,x)+k^2g^+_{l,n}(k,x)+2kg^+_{l,n-1}(k,x)+g^+_{l,n-2}(k,x)\\ &\quad=[ikP(x)+Q(x)]g^+_{l,n}(k,x)+iP(x)g^+_{l,n-1}(k,x),\\ &g^{+\prime\prime}_{r,n}(k,x)+k^2g^+_{r,n}(k,x)+2kg^+_{r,n-1}(k,x)+g^+_{r,n-2}(k,x)\\ &\quad=[ikP(x)+Q(x)]g^+_{r,n}(k,x)+iP(x)g^+_{r,n-1}(k,x). \end{aligned} \tag{8.2}$$

Defining the differential operator

$$\mathbf{W}(k)=-k^2-\frac{d^2}{dx^2}+ikP+Q, \tag{8.3}$$

so that $\dot{\mathbf{W}}(k)=-2k+iP$ and $\ddot{\mathbf{W}}(k)=-2\mathbf{I}$, we obtain the system of linear equations

$$\mathbf{T}(k)\mathbf{g}_l^+(k,x)=\mathbf{0}, \tag{8.4}$$

where $\mathbf{0}$ is the zero column vector of m components, $\mathbf{g}_l^+(k,x)$ is the column vector

$$[g_{l,m-1}^+(k,x),\ldots,g_{l,0}^+(k,x)]^t,$$

and $\mathbf{T}(k)$ is the $m\times m$ Toeplitz matrix given by

$$\mathbf{T}(k)=\begin{bmatrix} \mathbf{W}(k) & \dot{\mathbf{W}}(k) & \frac{1}{2}\ddot{\mathbf{W}}(k) & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mathbf{W}(k) & \dot{\mathbf{W}}(k) & \frac{1}{2}\ddot{\mathbf{W}}(k) & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mathbf{W}(k) & \dot{\mathbf{W}}(k) & \frac{1}{2}\ddot{\mathbf{W}}(k) & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \mathbf{W}(k) & \dot{\mathbf{W}}(k) \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \mathbf{W}(k) \end{bmatrix}.$$

Using the Leibnitz formula for repeated derivatives of products, we find from (4.2)

$$\frac{1}{n!}\left(\frac{d}{dk}\right)^n\frac{-2ik}{T^+(k)}=\sum_{j=0}^{n}[g_{l,j}^+(k,x);g_{r;n-j}^+(k,x)]. \tag{8.5}$$

We call $k_0\in\mathbf{C}^+$ an eigenvalue of $\mathbf{W}(k)$ if there exists a nontrivial $\phi\in L^2(\mathbf{R})$ such that $\mathbf{W}(k_0)\phi=0$. Because of Proposition 8.1, this is equivalent to $1/T^+(k_0)=0$. Further, ϕ is called an eigenfunction of $\mathbf{W}(k)$ corresponding to the eigenvalue k_0. More generally,[19] if k_0 is an eigenvalue of $\mathbf{W}(k)$, then the string of functions $\phi_0,\ldots,\phi_{m-1}$ in $L^2(\mathbf{R})$ is called a Jordan chain of length m corresponding to the eigenvalue k_0 if $\phi_0\neq 0$ and (8.4) holds with $[g_{l,m-1}^+(k_0,x),\ldots,g_{l,0}^+(k_0,x)]^t$ replaced by the column vector $[\phi_{m-1}(k_0,x),\ldots,\phi_0(k_0,x)]^t$.

Proposition 8.2: Assume $P,Q\in L^1(\mathbf{R})$ and let $k_0\in\mathbf{C}^+$ be an eigenvalue of $\mathbf{W}(k)$. If $\{g_{l,j}^+(k_0,\cdot)\}_{j=0}^{m-1}$ is a Jordan chain of $\mathbf{W}(k)$ of length m at the eigenvalue k_0, then for $n=0,1,\ldots,m-1$, we have as $x\to\pm\infty$

$$g_{l,n}^+(k_0,x)=\mathrm{O}[(1+|x|)^n e^{-|x|\mathrm{Im}\,k_0}],\quad g_{r,n}^+(k_0,x)=\mathrm{O}[(1+|x|)^n e^{-|x|\mathrm{Im}\,k_0}], \tag{8.6}$$

$$g_{l,n}^{+\prime}(k_0,x)=\mathrm{O}[(1+|x|)^n e^{-|x|\mathrm{Im}\,k_0}],\quad g_{r,n}^{+\prime}(k_0,x)=\mathrm{O}[(1+|x|)^n e^{-|x|\mathrm{Im}\,k_0}]. \tag{8.7}$$

Proof: If $k_0\in\mathbf{C}^+$ is an eigenvalue of $\mathbf{W}(k)$, by Proposition 8.1 we have $1/T^+(k_0)=0$. Hence $f_l^+(k_0,x)$ and $f_r^+(k_0,x)$ are linearly dependent, and thus we need to prove (8.6) and (8.7) only for $g_{l,n}^+(k_0,x)$ and $g_{l,n}^{+\prime}(k_0,x)$. For $n=0$ these follow from Theorem 2.1 (i). The following argument used to prove (8.6) and (8.7) for $n=m-1$ can be used recursively for $n=1,\ldots,m-2$. Note that, for each $k\in\mathbf{C}^+$, (1.1) has[20] an unbounded solution $X(k,x)$ such that

$$X(k,x)=\mathrm{O}(e^{|x|\mathrm{Im}\,k}),\quad X'(k,x)=\mathrm{O}(e^{|x|\mathrm{Im}\,k}),\quad x\to\pm\infty. \tag{8.8}$$

Let us choose $X(k_0,x)$ such that $[f_l^+(k_0,x);X(k_0,x)]=1$. Let us consider (8.2) as a second-order linear, nonhomogeneous differential equation for $g_{l,m-1}^+(k_0,x)$ and solve it by variation of parameters using the linearly independent solutions $f_l^+(k_0,x)$ and $X(k_0,x)$ of (1.1). We obtain

$$g_{l,m-1}^+(k_0,x)=a_{m-1}f_l^+(k_0,x)+b_{m-1}X(k_0,x)-\int_0^x dy([iP(y)-2k_0]g_{l,m-2}^+(k_0,y)$$
$$-g_{l,m-3}^+(k_0,y))[f_l^+(k_0,x)X(k_0,y)-f_l^+(k_0,y)X(k_0,x)], \tag{8.9}$$

with arbitrary constants a_{m-1} and b_{m-1}. Since $g_{l,n}(k_0,\cdot)\in L^2(\mathbf{R})$ for $n=0,1,\ldots, m-2$ and $X(k_0,x)$ is unbounded as $x\to\pm\infty$, the term proportional to $X(k_0,x)$ in (8.9) must vanish. Thus we must have

$$b_{m-1}+\int_0^\infty dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y)=0,$$

$$b_{m-1}-\int_{-\infty}^\infty dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y)=0,$$

and hence

$$\int_{-\infty}^\infty dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y)=0. \tag{8.10}$$

Using (8.10), we can write (8.9) as

$$g^+_{l,m-1}(k_0,x)=A(k_0,x)X(k_0,x)+f^+_l(k_0,x)\times\left[a_{m-1}-\int_0^x dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))X(k_0,y)\right], \tag{8.11}$$

where we have

$$A(k_0,x)=\begin{cases}\displaystyle\int_{-\infty}^x dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y), & x\leqslant 0,\\ \displaystyle\int_x^\infty dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y), & x\geqslant 0.\end{cases} \tag{8.12}$$

Using (8.6) for $n=0,1,\ldots, m-2$, (8.8), and (8.12), we obtain (8.6) for $g^+_{l,m-1}(k_0,x)$. Differentiating (8.11) and using (8.10), we obtain

$$g^{+\prime}_{l,m-1}(k_0,x)=[a_{m-1}-I_1(x)]f^{+\prime}_l(k_0,x)+X'(k_0,x)I_2(x),$$

where

$$I_1(x)=\int_0^x dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))X(k_0,y),$$

$$I_2(x)=\int_{-\infty}^x dy([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))f^+_l(k_0,y).$$

Finally, we obtain (8.7) for $g^{+\prime}_{l,m-1}(k_0,x)$ by using (8.6) for $n=0,1,\ldots, m-2$, (8.8), and (8.12).■

Proposition 8.3: Assume $P,Q\in L^1(\mathbf{R})$ and let $k_0\in\mathbf{C}^+$ be an eigenvalue of $\mathbf{W}(k)$. Then for $n=0,1,\ldots, m-1$, we have

$$\int_{-\infty}^\infty dy([iP(y)-2k_0]g^+_{l,n-1}(k_0,y)-g^+_{l,n-2}(k_0,y))f^+_l(k_0,y)=0 \tag{8.13}$$

if and only if $\{g^+_{l,j}(k_0,\cdot)\}_{j=0}^{m-1}$ is a Jordan chain of $\mathbf{W}(k)$ of length m corresponding to the eigenvalue k_0.

Proof: If (8.13) holds for $n=1$, then we must have $f^+_l(k_0,\cdot)\in L^2(\mathbf{R})$; from Proposition 8.2 and its proof it is seen that $f^+_l(k_0,\cdot)\in L^2(\mathbf{R})$ only if $g^+_{l,0}(k_0,x)=f^+_l(k_0,x)$ is an eigenvector of $\mathbf{W}(k)$. Recursively, we can show that $g^+_{l,n}$ given in (8.1) satisfies (8.4), (8.6), and (8.7), and hence

$g^+_{l,n}(k_0,\cdot)\in L^2(\mathbf{R})$ for $n=1,...,m-1$. Thus $\{g^+_{l,j}(k_0,\cdot)\}^{m-1}_{j=0}$ is a Jordan chain of $\mathbf{W}(k)$ of length m corresponding to the eigenvalue k_0. The converse is proved by proceeding recursively as in the proof of Proposition 8.2 leading to (8.10). ■

Theorem 8.4: Assume $P,Q\in L^1(\mathbf{R})$ and let $k_0\in\mathbf{C}^+$. Then the following four statements are equivalent:

(a) $\mathbf{W}(k)$ has a Jordan chain of length m corresponding to the eigenvalue k_0.
(b) $\{g^+_{l,j}(k_0,\cdot)\}^{m-1}_{j=0}$ is a Jordan chain of $\mathbf{W}(k)$ of length m corresponding to the eigenvalue k_0.
(c) $\{g^+_{r,j}(k_0,\cdot)\}^{m-1}_{j=0}$ is a Jordan chain of $\mathbf{W}(k)$ of length m corresponding to the eigenvalue k_0.
(d) $1/T^+(k)$ has a zero at k_0 of order at least m.

Proof: Clearly (b) implies (a). Now assume (a) holds and let $\{\phi_j\}^{m-1}_{j=0}$ be a Jordan chain of $\mathbf{W}(k)$ of length m at the eigenvalue k_0. Then ϕ_0 must be proportional to $f^+_l(k_0,x)$ and $f^+_r(k_0,x)$ because the latter two are linearly dependent and $\phi_0(k_0,x)$ is a solution of (1.1) for $k=k_0$. Thus we have $\mathbf{W}(k_0)g^+_{l,0}(k_0,x)=0$ and consequently (8.2) is satisfied for $n=0,1,...,m-1$. Hence (b) holds.

Note that (b) and (c) are equivalent because, if $\mathbf{W}(k_0)g^+_{l,0}(k_0,x)=0$ for some $k_0\in\mathbf{C}^+$ and $g^+_{l,0}(k_0,\cdot)\in L^2(\mathbf{R})$, by Proposition 8.1 we must have $1/T^+(k_0)=0$ and hence $g^+_{r,0}(k_0,x)$ must be a constant multiple of $g^+_{l,0}(k_0,x)$.

If (b) holds, then (8.6) and (8.7) must hold for $n=0,1,...,m-1$ because of Proposition 8.2. Then, for $n=0,1,...,m-1$, by evaluating the right-hand side of (8.5) at $x=-\infty$ or at $x=+\infty$, we find that its left-hand side must be zero and thus (d) holds. Now assume that (d) holds and let us show that (b) is true. By Proposition 8.3 it is sufficient to show that (8.13) is satisfied for $n=0,1,...,m-1$. We will do this recursively. First notice that (8.13) trivially holds for $n=0$ because $g^+_{l,-1}(k_0,x)=g^+_{l,-2}(k_0,x)=0$ and that (8.6) and (8.7) hold for $g^+_{l,0}$ and $g^{+\prime}_{l,0}$, respectively, because $1/T^+(k_0)=0$ and thus $f^+_l(k_0,x)$ is exponentially decaying as $|x|\to\infty$. For $n=1,...,m-2$, the proofs of (8.13) and of (8.6) and (8.7) for $g^+_{l,n}$ and $g^{+\prime}_{l,n}$, respectively, are similar to the case when $n=m-1$. Thus, it suffices to give the proofs for $n=m-1$ by assuming that these equations hold for $n=1,...,m-2$. Using (1.1) for $g^+_{r,0}(k_0,x)$ and (8.2) for $g^+_{l,m-1}(k_0,x)$, we obtain the Wronskian relation

$$-\frac{d}{dx}[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)]=([iP(x)-2k_0]g^+_{l,m-2}(k_0,x)-g^+_{l,m-3}(k_0,x))g^+_{r,0}(k_0,x). \tag{8.14}$$

In a similar way, we obtain

$$\frac{d}{dx}[g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]=([iP(x)-2k_0]g^+_{r,m-2}(k_0,x)-g^+_{r,m-3}(k_0,x))g^+_{l,0}(k_0,x). \tag{8.15}$$

Integrating (8.14) and (8.15) we get

$$\int_{-\infty}^{\infty}dy\ ([iP(y)-2k_0]g^+_{l,m-2}(k_0,y)-g^+_{l,m-3}(k_0,y))g^+_{r,0}(k_0,y)$$
$$=\lim_{x\to-\infty}[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)]-\lim_{x\to+\infty}[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)], \tag{8.16}$$

$$\int_{-\infty}^{\infty}dy\ ([iP(y)-2k_0]g^+_{r,m-2}(k_0,y)-g^+_{r,m-3}(k_0,y))g^+_{l,0}(k_0,y)$$
$$=\lim_{x\to+\infty}[g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]-\lim_{x\to-\infty}[g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]. \tag{8.17}$$

Because (8.6) and (8.7) hold for $n=0,1,...,m-2$, we have

$$\lim_{x\to\pm\infty}\sum_{j=1}^{m-2}[g^+_{l,j}(k_0,x);g^+_{r;m-j}(k_0,x)]=0. \tag{8.18}$$

Since $1/T^+(k_0)$ is assumed to have a zero of order at least m, using (8.5) for $n=m-1$ and (8.18), we obtain

$$\lim_{x\to\pm\infty}([g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]+[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)])=0. \tag{8.19}$$

Using the linear dependence of $g^+_{l,0}(k_0,x)$ and $g^+_{r,0}(k_0,x)$ and the exponential decay of $g^+_{l,m-1}(k_0,x)$ as $x\to+\infty$ and of $g^+_{r,m-1}(k_0,x)$ as $x\to-\infty$, we first conclude that

$$\lim_{x\to+\infty}[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)]=0,\quad \lim_{x\to-\infty}[g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]=0. \tag{8.20}$$

Now, from (8.19) and (8.20) we see that

$$\lim_{x\to-\infty}[g^+_{l,m-1}(k_0,x);g^+_{r,0}(k_0,x)]=0,\quad \lim_{x\to+\infty}[g^+_{l,0}(k_0,x);g^+_{r,m-1}(k_0,x)]=0, \tag{8.21}$$

and thus using (8.20) and (8.21) in (8.16) and (8.17), we get

$$\int_{-\infty}^{\infty}dy\ ([iP(y)-2k_0]g^+_{l,m-1}(k_0,y)-g^+_{l,m-2}(k_0,y))g^+_{r,0}(k_0,y)=0.$$

Thus (8.13) is proved for $n=m-1$, and hence (b) holds. ■

IX. BOUND STATES AND POLES OF $T^+(k)$

In this section we further analyze the poles of $T^+(k)$ in $\mathbf{C}^+$. We show that such poles cannot occur in certain regions in $\mathbf{C}^+$ determined in terms of the constants defined in (9.1). When $P(x)\leqslant 0$, we show that such poles are confined to a certain interval on the positive imaginary axis. We analyze the change in the number of bound states when P and Q are perturbed. In the generic case we find that the number of bound states is unchanged under small perturbations of P and Q; in the exceptional case we find that the number of bound states is unchanged under small perturbations of $P(x)$. When $P(x)\leqslant 0$ we show that the number of bound states is independent of $P(x)$. We also present a Levinson theorem relating the number of bound states to the change in the argument of $T^+(k)$.

Next we obtain some simple conditions on $P(x)$ and $Q(x)$ guaranteeing that there are no bound states outside certain k-regions in $\mathbf{C}^+$ determined by the following parameters:

$$P_{\min}=\operatorname{ess}\inf_{x\in\mathbf{R}}P(x),\quad P_{\max}=\operatorname{ess}\sup_{x\in\mathbf{R}}P(x),\quad Q_{\min}=\operatorname{ess}\inf_{x\in\mathbf{R}}Q(x). \tag{9.1}$$

Let us also define $\beta^*=P_{\max}/2+\sqrt{P^2_{\max}/4-Q_{\min}}$. Note that if $P,Q\in L^1(\mathbf{R})$, then it follows that $P_{\max}\geqslant 0$ with the equality holding if and only if $P(x)\leqslant 0$, that $Q_{\min}\leqslant 0$ with the equality holding if and only if $Q(x)\geqslant 0$, and that $P_{\min}\leqslant 0$ with the equality holding if and only if $P(x)\geqslant 0$. Furthermore, $\beta^*\geqslant P_{\max}$ with the equality holding if and only if $Q(x)\geqslant 0$. Note also that $\beta^*\geqslant 0$ with the equality holding if and only if $P(x)=Q(x)=0$; hence, the case $\beta^*=0$ is trivial.

Theorem 9.1: Assume $P,Q\in L^1(\mathbf{R})$, $P(x)\not\equiv 0$, and $P_{\max}$ is finite. Then for $P_{\max}/2\leqslant \operatorname{Im}k<\beta^*$ the zeros of $1/T^+(k)$ can only occur on the imaginary axis, and all such zeros are simple. If, in addition, $Q_{\min}$ is finite, then there are no zeros of $1/T^+(k)$ in the region $\{k\in\mathbf{C}^+:(\operatorname{Im}k)^2-(\operatorname{Re}k)^2-(\operatorname{Im}k)P_{\max}\geqslant -Q_{\min}\}$. Consequently, $1/T^+(k)$ has no zeros in $\mathbf{C}^+$ satisfying $\operatorname{Im}k\geqslant\beta^*$.

Proof: From (1.1), after using (4.7), we obtain

$$\frac{d}{dx}\{f^+_l(-\overline{k_0},x)f^{+\prime}_l(k_0,x)\}=|f^{+\prime}_l(k_0,x)|^2+[-k_0^2+ik_0P(x)+Q(x)]|f^+_l(k_0,x)|^2. \tag{9.2}$$

If $k_0 \in \mathbf{C}^+$ is a zero of $1/T^+(k)$, then by integrating (9.2) and using (4.7) and Theorem 2.1 (i), we obtain

$$\int_{-\infty}^{\infty} dx\ |f_l^+{}'(k_0,x)|^2 = \int_{-\infty}^{\infty} dx\ [k_0^2 - ik_0 P(x) - Q(x)]|f_l^+(k_0,x)|^2. \tag{9.3}$$

Letting $k_0 = \alpha + i\beta$ and separating the real and imaginary parts in (9.3), we obtain

$$i\alpha \int_{-\infty}^{\infty} dx\ [2\beta - P(x)]|f_l^+(k_0,x)|^2 = 0, \tag{9.4}$$

$$\int_{-\infty}^{\infty} dx\ |f_l^+{}'(k_0,x)|^2 = \int_{-\infty}^{\infty} dx\ [\alpha^2 - \beta^2 + \beta P(x) - Q(x)]|f_l^+(k_0,x)|^2. \tag{9.5}$$

From (9.4) we see that we must have $\alpha = 0$ when $P_{\max} \leqslant 2\beta$, and hence any zero of $1/T^+(k)$ with $\operatorname{Im} k \geqslant P_{\max}/2$ can only occur on the positive imaginary axis. All such zeros are simple; otherwise, a zero of order two or higher would imply (8.13) with $n=1$, i.e., $\int_{-\infty}^{\infty} dx\ [P(x) - 2\operatorname{Im} k_0]|f_l^+(k_0,x)|^2 = 0$, which cannot happen if $\operatorname{Im} k_0 \geqslant P_{\max}/2$. From (9.5) we see that we cannot have $\alpha^2 - \beta^2 + \beta P(x) - Q(x) \leqslant 0$. Hence there are no zeros of $1/T^+(k)$ in $\{\alpha + i\beta \in \mathbf{C}^+ : \beta^2 - \alpha^2 - \beta P_{\max} \geqslant -Q_{\min}\}$. The analysis of the corresponding region in the $\alpha\beta$-plane indicates that there cannot be any zeros of $1/T^+(k)$ on the imaginary axis when $\operatorname{Im} k \geqslant \beta^*$, and hence there cannot be any zeros of $1/T^+(k)$ either on or off the imaginary axis when $\operatorname{Im} k \geqslant \beta^*$. ■

When $P(x) \leqslant 0$, from Theorem 9.1 we obtain the following corollary.

Corollary 9.2: Assume $P(x) \leqslant 0$ and $P, Q \in L^1(\mathbf{R})$. Then, the poles of $T^+(k)$ in $\mathbf{C}^+$ are all purely imaginary and simple. In addition, assume that $Q_{\min}$ defined in (9.1) is finite; then there are no zeros of $1/T^+(k)$ in $\mathbf{C}^+$ for $\operatorname{Im} k \geqslant \sqrt{-Q_{\min}}$.

Theorem 9.3: Assume $Q(x) \equiv 0$ and $P \in L_1^1(\mathbf{R})$. If $\int_{-\infty}^{\infty} dx\ P(x) > 2$, then (1.1) has at least one bound state at $k = i\beta$ for some positive β. If $\int_{-\infty}^{\infty} dx\ |P(x)| \leqslant 2$, then $1/T^+(k)$ has no zeros in $\mathbf{C}^+$.

Proof: When $\int_{-\infty}^{\infty} dx\ P(x) > 2$, from (5.7) we see that $1/T^+(i\beta)$ is negative at $\beta = 0$ and from (6.1) we see that it is positive as $\beta \to +\infty$. Being a real-valued, continuous function of β, $1/T^+(i\beta)$ must have a zero for some positive β. Now let us prove the second statement. Assume $k \in \mathbf{C}^+$ corresponds to a bound state; we can transform (1.1) into

$$\varphi(k,x) = \int_{-\infty}^{\infty} dy\ \mathscr{B}(k;x,y)\varphi(k,y), \tag{9.6}$$

where we have defined

$$\varphi(k,x) = |P(x)|^{1/2}\psi^+(k,x), \quad \mathscr{B}(k;x,y) = \tfrac{1}{2}e^{ik|x-y|}|P(x)|^{1/2}P(y)/|P(y)|^{1/2}.$$

When $P \in L^1(\mathbf{R})$ and $k \in \mathbf{C}^+$, the integral operator in (9.6) is Hilbert–Schmidt with the Hilbert–Schmidt norm

$$\|\mathscr{B}\|_{\mathrm{HS}}^2 = \tfrac{1}{4}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy |P(x)| e^{-2|x-y|\operatorname{Im} k} |P(y)|,$$

and hence, if $k \in \mathbf{C}^+$ and $\int_{-\infty}^{\infty} dx\ |P(x)| \leqslant 2$, we have $\|\mathscr{B}\|_{\mathrm{HS}} < 1$. Thus the operator norm of that integral operator is also strictly less than 1 and hence $\varphi = 0$, implying that there cannot be any bound states of (1.1) for $k \in \mathbf{C}^+$. ■

It is already known[9] that if $Q \in L_1^1(\mathbf{R})$, then the number of bound states for (2.6) is finite; let us denote that number by $\mathscr{N}$, and let $i\kappa_1, \ldots, i\kappa_{\mathscr{N}}$ with $0 < \kappa_1 < \ldots < \kappa_{\mathscr{N}}$ denote the zeros of $1/T^{[0]}(k)$ in $\mathbf{C}^+$. In the following, we generalize the second result of Theorem 9.3 to the case $Q(x) \not\equiv 0$.

Theorem 9.4: Assume $P,Q\in L^1_1(\mathbf{R})$. The k-values in $\mathbf{C}^+$ satisfying $\int_{-\infty}^{\infty}dx\ |ikP(x)+Q(x)|\leqslant 2|k|$ cannot be zeros of $1/T^+(k)$. Moreover, there are no zeros of $1/T^+(k)$ in $\mathbf{C}^+\backslash\{i\kappa_1,\ldots,i\kappa_{\mathcal{N}}\}$ satisfying $|T^{[0]}(k)|\|P\|_{1,1}<2e^{-\|Q\|_{1,1}}$.

Proof: Let $k\in\mathbf{C}^+$ correspond to a bound state of (1.1). We can transform (1.1) into (9.6) with

$$\varphi(k,x)=|ikP(x)+Q(x)|^{1/2}\psi^+(k,x),$$

$$\mathscr{B}(k;x,y)=\frac{1}{2ik}\,e^{ik|x-y|}|ikP(x)+Q(x)|^{1/2}\,\frac{[ikP(y)+Q(y)]}{|ikP(y)+Q(y)|^{1/2}}.$$

A sufficient condition for the absence of bound states is $\|\mathscr{B}\|_{\mathrm{HS}}<1$. Proceeding as in the proof of Theorem 9.3, for $k\in\mathbf{C}^+$ we obtain $\|\mathscr{B}\|_{\mathrm{HS}}<\int_{-\infty}^{\infty}dx\ |ikP(x)+Q(x)|/(2|k|)$, and hence there are no zeros of $1/T^+(k)$ at the k-values in $\mathbf{C}^+$ satisfying $\int_{-\infty}^{\infty}dx\ |ikP(x)+Q(x)|\leqslant 2|k|$. In the special case $P(x)\equiv 0$, this implies that there are no bound states when $|k|>\int_{-\infty}^{\infty}dx\ |Q(x)|/2$. To prove the second part of the theorem, we note that the kernel of the resolvent of the operator $[-d^2/dx^2+Q(x)-k^2]^{-1}$ is given by[20]

$$\mathscr{R}(k;x,y)=\frac{\theta(y-x)f_r^{[0]}(k,x)f_l^{[0]}(k,y)+\theta(x-y)f_l^{[0]}(k,x)f_r^{[0]}(k,y)}{[f_l^{[0]}(k,\cdot);f_r^{[0]}(k,\cdot)]},\tag{9.7}$$

where $f_l^{[0]}(k,x)$ and $f_r^{[0]}(k,x)$ are the Jost solutions of (2.6) and $\theta(x)$ is the Heaviside function. As seen from (2.8), the Wronskian in (9.7) is equal to $-2ik/T^{[0]}(k)$, and hence we get

$$\|ik|P(x)|^{1/2}\mathscr{R}(k;x,y)P(y)/|P(y)|^{1/2}\|^2_{\mathrm{HS}}=\tfrac{1}{4}|T^{[0]}(k)|^2C(k),\tag{9.8}$$

where we have defined

$$C(k)=\int_{-\infty}^{\infty}dx\ |P(x)||f_l^{[0]}(k,x)|^2\int_{-\infty}^{x}dy\ |f_r^{[0]}(k,y)|^2|P(y)|$$
$$+\int_{-\infty}^{\infty}dx\ |P(x)||f_r^{[0]}(k,x)|^2\int_{x}^{\infty}dy\ |f_l^{[0]}(k,y)|^2|P(y)|.\tag{9.9}$$

Using (2.1) with $P(x)=0$, we deduce that

$$|f_l^{[0]}(k,x)|\leqslant(1+\max\{0,-x\})e^{-x\ \mathrm{Im}\,k}e^{\int_x^{\infty}dy\ (1+|y|)|Q(y)|},\quad k\in\overline{\mathbf{C}^+},\tag{9.10}$$

$$|f_r^{[0]}(k,x)|\leqslant(1+\max\{0,x\})e^{x\ \mathrm{Im}\,k}e^{\int_{-\infty}^{x}dy\ (1+|y|)|Q(y)|},\quad k\in\overline{\mathbf{C}^+}.\tag{9.11}$$

In (9.9), using (9.10), (9.11), and the estimates $1+\max\{0,\pm x\}\leqslant 1+|x|$ and $e^{\mp 2(x-y)\mathrm{Im}\,k}\leqslant 1$ for $\pm(x-y)\geqslant 0$, we obtain $|C(k)|\leqslant\|P\|^2_{1,1}e^{2\|Q\|_{1,1}}$. Hence the Hilbert–Schmidt norm on the left-hand side of (9.8) is strictly less than 1, provided $|T^{[0]}(k)|<2e^{-\|Q\|_{1,1}}/\|P\|_{1,1}$. Under this condition on k, there is no bound state corresponding to that $k\in\mathbf{C}^+$. ■

Let us denote the number of bound states of (1.1), i.e., the number of zeros of $1/T^+(k)$ in $\mathbf{C}^+$ (including multiplicities) by $N(P,Q)$. In the next two propositions we obtain some stability results for $N(P,Q)$ under certain perturbations of P and Q. As in Sec. VII, we let $T_j^+(k)$ denote the transmission coefficient corresponding to (7.1) for $j=1,2$.

Proposition 9.5: Assume $P_1,P_2\in L^1(\mathbf{R})$, $Q_1,Q_2\in L^1_1(\mathbf{R})$, $1/T_1^+(k)$ does not have any real zeros, and $Q_1(x)$ is a generic potential. If $\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}$ is small, i.e., if (9.13) is satisfied, then

(a) $1/T_2^+(k)$ does not have any real zeros.

(b) $N(P_2,Q_2)=N(P_1,Q_1)$.

(c) If all zeros of $1/T_1^+(k)$ are simple and purely imaginary, so are those of $1/T_2^+(k)$.

Proof: For $a>0$ let Γ_a be the positively oriented contour consisting of the interval $[-a,a]$ and the semicircle $\{k\in\overline{\mathbf{C}^+}:|k|=a\}$, and let us choose a large enough so that all zeros of $1/T_1^+(k)$ in $\mathbf{C}^+$ have an absolute value less than a. Putting $F_j(k)=k/[(k+i)T_j^+(k)]$ for $j=1,2$, from Propositions 7.1 and 7.2 we get

$$\left|\frac{F_1(k)-F_2(k)}{F_1(k)}\right| \leqslant \frac{C}{|F_1(k)|}\,(\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}). \tag{9.12}$$

If $1/T_1^+(k)$ does not have any real zeros, then in the generic case $F_1(k)$ cannot have any real zeros; thus, $F_1(k)$ does not vanish on Γ_a. Moreover, by Proposition 4.1, $F_1(k)$ is continuous and bounded in $\overline{\mathbf{C}^+}$. Hence $\min_{k\in\Gamma_a}|F_1(k)|>0$. Now choosing

$$\|P_1-P_2\|_1+\|Q_1-Q_2\|_{1,1}<\frac{1}{C}\min_{k\in\Gamma_a}\left|\frac{k}{(k+i)T_1^+(k)}\right|, \tag{9.13}$$

from (9.12) we obtain

$$\left|\frac{F_1(k)-F_2(k)}{F_1(k)}\right|<1, \quad k\in\Gamma_a. \tag{9.14}$$

Hence from (9.14) we see $F_2(k)$ cannot vanish on Γ_a, which implies (a). Part (b) follows from (9.14) with the use of Rouché's theorem. Part (c) follows by replacing Γ_a with the union of $N(P_1,Q_1)$ small, positively oriented circles centered at the zeros of $1/T_1^+(k)$ and by applying Rouché's theorem. ■

Proposition 9.6: Assume $Q_1=Q_2=Q$ in (7.1), $1/T_1^+(k)$ does not have any real zeros, $Q(x)$ is an exceptional potential, and $P_1,P_2,Q\in L_1^1(\mathbf{R})$. If $\|P_1-P_2\|_{1,1}$ is small, i.e., if (9.15) is satisfied, then we have

(a) $1/T_2^+(k)$ does not have any real zeros.

(b) $N(P_2,Q)=N(P_1,Q)$.

(c) If all zeros of $1/T_1^+(k)$ are simple and purely imaginary, so are those of $1/T_2^+(k)$.

Proof: We will proceed as in the proof of Proposition 9.5. Let us choose Γ_a as in that proof but define $F_j(k)=1/T_j^+(k)$ for $j=1,2$, instead. Note that $F_1(k)$ is bounded, continuous, and nonzero on Γ_a. From (7.13) we have

$$\left|\frac{F_1(k)-F_2(k)}{F_1(k)}\right|\leqslant C\|P_1-P_2\|_{1,1}.$$

From (9.8) we get

$$\left|\frac{F_1(k)-F_2(k)}{F_1(k)}\right|\leqslant \frac{C}{|F_1(k)|}\,\|P_1-P_2\|_{1,1},$$

and hence by choosing

$$\|P_1-P_2\|_{1,1}<\frac{1}{C}\min_{k\in\Gamma_a}\left|\frac{1}{T_1^+(k)}\right|, \tag{9.15}$$

and proceeding as in the proof of Proposition 9.5, we complete the proof. ■

Theorem 9.7: Assume $P,Q\in L^1(\mathbf{R})$ and $P(x)\leqslant 0$. Then, either $N(P,Q)$ and $N(0,Q)$ are both infinite, or they are both finite and $N(P,Q)=N(0,Q)$. Thus the number of bound states of (1.1) coincides with the number of bound states of (2.6).

Proof: Since $P(x)\leqslant 0$, by Corollary 9.2 we know that the bound states of (1.1) can only occur when k is on the positive imaginary axis. Let us write (1.1) with $k=i\beta$ as two simultaneous equations:

$$-\psi''+V(\beta,x)\psi=E(\beta)\psi, \tag{9.16}$$

$$E(\beta)=-\beta^2, \tag{9.17}$$

where β is considered to be a parameter in the potential $V(\beta,x)=Q(x)-\beta P(x)$ of the Schrödinger equation (9.16), and $E(\beta)$ denotes the corresponding energy for each β. Each bound-state energy $-\kappa_j^2$ of (2.6) gives rise to an eigenvalue branch $E_j(\beta)$. From (9.16) we have

$$E'(\beta)=\frac{\langle \psi,-P\psi\rangle}{\langle \psi,\psi\rangle}, \tag{9.18}$$

where $\langle\cdot,\cdot\rangle$ denotes the usual scalar product on $L^2(\mathbf{R})$. If $P(x)\leqslant 0$, from (9.18) we see that $E'(\beta)\geqslant 0$ and hence each $E_j(\beta)$ is a nondecreasing function of β. Therefore for $\beta\geqslant 0$, the graph of $E_j(\beta)$ must intersect the parabola $E=-\beta^2$ at exactly one point, say $(\beta_j,-\beta_j^2)$, and each $E_j(\beta)$ gives rise to exactly one solution of (9.17). The number $N(P,Q)$ is equal to the number of intersections of the eigenvalue branches $E_j(\beta)$ with the parabola given in (9.17) for $j\geqslant 1$. Since each of the $N(0,Q)$ branches is responsible for exactly one intersection, we conclude that $N(P,Q)=N(0,Q)$. Note that if $Q\in L^1(\mathbf{R})$ but $Q\notin L_1^1(\mathbf{R})$, it is possible that $N(0,Q)=+\infty$, but then we also have $N(P,Q)=+\infty$. ■

If $P,Q\in L^1(\mathbf{R})$ and $P(x)\leqslant 0$, then either $N(P,Q)=+\infty$, in which case the set of bound-state energies of (1.1) consists of a strictly decreasing sequence of negative numbers converging to 0, or $N(P,Q)$ is finite and equal to $\mathcal{N}$, in which case we let $k=i\beta_j$ for $j=1,\ldots,\mathcal{N}$, with $0<\beta_1<\ldots<\beta_{\mathcal{N}}$ denote the zeros of $1/T^+(k)$ in $\mathbf{C}^+$. Since the condition $Q\in L_1^1(\mathbf{R})$ guarantees the finiteness of $N(0,Q)$, from Theorem 9.1 we obtain the following:

Corollary 9.8: Assume $P(x)\leqslant 0$, $Q(x)\geqslant 0$, $P\in L^1(\mathbf{R})$, and $Q\in L_1^1(\mathbf{R})$. Then, there are no zeros of $1/T^+(k)$ in $\overline{\mathbf{C}^+}$.

In the next theorem, when $P(x)\leqslant 0$, using the constant $P_{\min}$ defined in (9.1), we obtain some upper and lower bounds on each bound-state energy of (1.1).

Theorem 9.9: Assume $N(0,Q)$ is finite and nonzero, $P_{\min}$ is finite, $P(x)\leqslant 0$, and $P,Q\in L^1(\mathbf{R})$; let $k=i\kappa_j$ correspond to the bound states of (2.6) for $j=1,\ldots,\mathcal{N}$. Then, the zeros of $1/T^+(k)$ in $\mathbf{C}^+$ occur at $k=i\beta_j$ satisfying $\beta_*\leqslant\beta_j\leqslant\kappa_j$ for $j=1,\ldots,\mathcal{N}$, where $\beta_*=P_{\min}/2+\sqrt{P_{\min}^2/4+\kappa_1^2}$. In particular, $\beta_1\geqslant\beta_*$ and $\beta_{\mathcal{N}}\leqslant\kappa_{\mathcal{N}}$, with the equalities holding if and only if $P(x)=0$.

Proof: At a bound state with $k=i\beta_j$ of (1.1), replacing k_0 in (9.5) by $0+i\beta_j$, we get

$$\int_{-\infty}^{\infty}dx\ f_l^{+\prime}(i\beta_j,x)^2=\int_{-\infty}^{\infty}dx\ [-\beta_j^2+\beta_j P(x)-Q(x)]f_l^+(i\beta_j,x)^2. \tag{9.19}$$

On the other hand, since $-\kappa_{\mathcal{N}}^2$ is the lowest bound-state energy for (2.6), we have

$$-\kappa_{\mathcal{N}}^2\leqslant\frac{\int_{-\infty}^{\infty}dx\ [f_l^{+\prime}(i\beta_j,x)^2+Q(x)f_l^+(i\beta_j,x)^2]}{\int_{-\infty}^{\infty}dx\ f_l^+(i\beta_j,x)^2}, \tag{9.20}$$

with the equality holding if and only if $f_l^{[0]}(i\kappa_{\mathcal{N}},x)$ and $f_l^+(i\beta_j,x)$ are linearly dependent. From (9.19) and (9.20) we obtain

$$-\kappa_{\mathcal{N}}^2\leqslant-\beta_j^2+\beta_j\frac{\int_{-\infty}^{\infty}dx\ P(x)f_l^+(i\beta_j,x)^2}{\int_{-\infty}^{\infty}dx\ f_l^+(i\beta_j,x)^2}. \tag{9.21}$$

Since $P(x)\leqslant 0$, from (9.21) we see that $\beta_j\leqslant\kappa_{\mathcal{N}}$ with the equality holding if and only if $P(x)=0$ and $j=\mathcal{N}$. Thus $\beta_j\in(0,\kappa_{\mathcal{N}}]$ for $j=1,\ldots,\mathcal{N}$. Now let us improve the bounds on β_j. From the proof of Theorem 9.7, recall that each eigenvalue branch $E_j(\beta)$ gives rise to exactly one solution of (9.17) starting with $-\kappa_j^2$ at $\beta=0$ and ending with $-\beta_j^2$ at $\beta=\beta_j$. Since $E_j(\beta)$ is an increasing function of β, we get $-\kappa_j^2=E_j(0)\leqslant E_j(\beta_j)=-\beta_j^2$, and hence $\beta_j\leqslant\kappa_j$. Now consider $E_1(\beta)$, the eigenvalue branch corresponding to β_1. From (9.18) using $P_{\min}\leqslant P(x)$, we obtain $0\leqslant E_1'(\beta)\leqslant -P_{\min}$; more specifically, $0<E_1'(\beta)<-P_{\min}$ unless $P(x)=0$. Since $E_1(0)=-\kappa_1^2$ and $E_1(\beta)$ is nondecreasing, we get $E_1(\beta)\leqslant-\beta P_{\min}-\kappa_1^2$. Thus from the inequality $-\beta_1^2\leqslant-\beta_1P_{\min}-\kappa_1^2$, we get $\beta_1\geqslant\beta_*$. Note that the equality in $\beta_1\geqslant\beta_*$ holds if and only if $P(x)=0$ because $\beta_*\geqslant\kappa_1$ with the equality holding if and only if $P(x)\geqslant 0$. ■

From the proof of Theorem 9.9 we get the following corollary that holds even when (2.6) has infinitely many bound states.

Corollary 9.10: Assume $P,Q\in L^1(\mathbf{R})$ and $P(x)\leqslant 0$, and let $\{\mathscr{E}_j\}$ and $\{\mathscr{E}_j^{[0]}\}$ for $j\geqslant 1$ denote the bound-state energies of (1.1) and (2.6), respectively, ordered such that $\mathscr{E}_j<\mathscr{E}_{j+1}$ and $\mathscr{E}_j^{[0]}<\mathscr{E}_{j+1}^{[0]}$. Then we have $\mathscr{E}_j^{[0]}\leqslant\mathscr{E}_j<0$ for $j\geqslant 1$, and hence the bound-state energies of (1.1) cannot occur below the lowest bound-state energy of (2.6).

Recall that the Levinson theorem[21] relates the number of bound states for the usual Schrödinger equation to the change in the phase of the transmission coefficient. Next we generalize the Levinson theorem to (1.1).

Theorem 9.11: Assume that $P\in L^1(\mathbf{R})$ in the generic case and $P\in L_1^1(\mathbf{R})$ in the exceptional case and that $Q\in L_1^1(\mathbf{R})$, and suppose $1/T^+(k)$ does not have any real zeros. Then the number of bound states of (1.1) is given by

$$\arg\, T^+(0+)=\pi\left[N(P,Q)-\frac{d}{2}\right], \tag{9.22}$$

where $d=0$ in the exceptional case and $d=1$ in the generic case, and $\arg\, T^+(k)$ denotes the continuous branch of the argument of $T^+(k)$ normalized such that $\arg\, T^+(+\infty)=0$.

Proof: For $b>a>0$, let $\Gamma_{a,b}$ be the positively oriented contour consisting of the circular arcs $\{k\in\overline{\mathbf{C}^+}:|k|=a\}$ and $\{k\in\overline{\mathbf{C}^+}:|k|=b\}$ and the segments $[-b,-a]$ and $[a,b]$. Let us choose a and b so that all zeros of $1/T^+(k)$ in $\mathbf{C}^+$ are enclosed by $\Gamma_{a,b}$. By the argument principle we have

$$N(P,Q)=-\frac{1}{2\pi i}\int_{\Gamma_{a,b}}dk\,\frac{\dot{T}^+(k)}{T^+(k)}=-\frac{1}{2\pi}\,\Delta_{\Gamma_{a,b}}[\arg\, T^+(k)], \tag{9.23}$$

where $\Delta_{\Gamma_{a,b}}[\arg\, T^+(k)]$ indicates the change in the argument of $T^+(k)$ when $\Gamma_{a,b}$ is traversed once. This change is independent of a and b, and hence we evaluate it by letting $a\to 0$ and $b\to+\infty$. By Theorem 6.1 the contribution to that change from the large semicircle $\{k\in\overline{\mathbf{C}^+}:|k|=b\}$ vanishes as $b\to+\infty$. In view of (5.2) and (5.3), we see that the contribution from the small semicircle $\{k\in\overline{\mathbf{C}^+}:|k|=a\}$ in the limit $a\to 0$ is equal to 0 in the exceptional case and $-\pi$ in the generic case. Thus the contribution from the interval $(0,+\infty)$ is given by $\arg\, T^+(+\infty)-\arg\, T^+(0+)=-\arg\, T^+(0+)$. By the first equality in (4.10), the contribution from the interval $(-\infty,0)$ is the same. Hence, the right-hand side in (9.23) is equal to $(1/\pi)\arg\, T^+(0+)$ in the exceptional case and $(1/2)+(1/\pi)\arg\, T^+(0+)$ in the generic case, which gives us (9.22). ■

Finally, let us show that in the special case when P and Q have support in a half-line, we can relate the poles of the transmission coefficient to the poles of a reflection coefficient. Since there is no loss of generality in choosing our half-lines as $\mathbf{R}^+=(0,+\infty)$ or $\mathbf{R}^-=(-\infty,0)$ instead of $(a,+\infty)$ or $(-\infty,b)$, respectively, for some constants a and b, we will state the following proposition using $\mathbf{R}^\pm$.

Proposition 9.12: Assume $P(x)=Q(x)=0$ for $x\in\mathbf{R}^-$ and $P,Q\in L^1(\mathbf{R}^+)$. Then $L^+(k)$ is meromorphic in $\mathbf{C}^+$ having poles coinciding with the poles of $T^+(k)$. Furthermore, none of the zeros of $L^+(k)$ coincide with the poles of $T^+(k)$ in $\mathbf{C}^+$. These assertions remain valid if $\mathbf{R}^-$ and $L^+(k)$ are replaced by $\mathbf{R}^+$ and $R^+(k)$, respectively.

Proof: If $P(x)=Q(x)=0$ for $x\in\mathbf{R}^-$, from Theorem 2.1 (i), we see that $f_l^+(k,0)$ and $f_l^{+\prime}(k,0)$ are analytic in $\mathbf{C}^+$. Hence using (1.3) we can conclude that $L^+(k)/T^+(k)$ is analytic in $\mathbf{C}^+$, allowing us to conclude that the poles of $L^+(k)$ and $T^+(k)$ must coincide in $\mathbf{C}^+$. Since $f_l^+(k,0)$ and $f_l^{+\prime}(k,0)$ cannot vanish simultaneously, it follows that $1/T^+(k)$ and $L^+(k)/T^+(k)$ cannot vanish simultaneously in $\mathbf{C}^+$, and hence the zeros of $L^+(k)$ and the poles of $T^+(k)$ cannot coincide in $\mathbf{C}^+$. The proof when P and Q have support in $\mathbf{R}^-$ is obtained in a similar manner. ■

X. EIGENVALUE CURVES AND ZEROS OF JOST SOLUTIONS

In this section we study the zeros of the Jost solutions of (1.1) for a fixed $k\in\overline{\mathbf{C}^+}$ and analyze the number of such zeros in relation to the bound states of (1.1) and (2.6). As in Sec. IX, we let $N(P,Q)$ denote the number of bound states of (1.1). When $P(x)\leqslant 0$ we show that the number of

zeros of the Jost solutions of (1.1) is related to $N(P,Q)$ in a simple manner, and we present some examples showing that this relation does not hold in general. We establish the connection between the results of Sec. VIII on Jordan chains and certain zeros of the Jost solutions of (1.1). This connection uses the eigenvalue branches introduced in the proof of Theorem 9.7. We also show that the number of bound states of (1.1) with real energies is greater than or equal to $N(0,Q)$.

In the first proposition, we collect some results about the oscillation properties of solutions of generalized Schrödinger equations related by inequalities involving the coefficients. Although the methods for proving such results are familiar,[20,22] we include a proof for the convenience of the reader.

Consider the pair of generalized Schrödinger equations

$$\chi_j''(\mu,x)-\mu^2\chi_j(\mu,x)=V_j(\mu,x)\chi_j(\mu,x),\quad \mu\geqslant 0,\quad j=1,2. \tag{10.1}$$

Note that if we let $V_j(\mu,x)=-\mu P(x)+Q(x)$ in (10.1), we get (1.1) for $k=i\mu$.

Proposition 10.1: Assume $V_j(\mu,\cdot)\in L^1(\mathbf{R})$ if $\mu>0$, $V_j(0,\cdot)\in L^1_1(\mathbf{R})$, and $V_1(\mu_2,x)\leqslant V_2(\mu_1,x)$ if $0\leqslant\mu_1\leqslant\mu_2$. Let $\chi_1(\mu_1,x)$ and $\chi_2(\mu_2,x)$ denote two nontrivial solutions of (10.1) with the corresponding coefficients $V_1(\mu_1,x)$ and $V_2(\mu_2,x)$, respectively. Then:

(i) Suppose $\chi_2(\mu_2,x)$ has two successive zeros a and b with $a<b$. If $0\leqslant\mu_1<\mu_2$, then $\chi_1(\mu_1,x)$ has at least one zero in (a,b). If $0\leqslant\mu_1=\mu_2$ and $V_1(\mu_1,x)\not\equiv V_2(\mu_1,x)$ on (a,b), then $\chi_1(\mu_1,x)$ has at least one zero in (a,b). If $0\leqslant\mu_1=\mu_2$, $V_1(\mu_1,x)\equiv V_2(\mu_1,x)$ on (a,b), and $\chi_1(\mu_1,x)$ and $\chi_2(\mu_1,x)$ are linearly independent in (a,b), then $\chi_1(\mu_1,x)$ has exactly one zero in (a,b).

(ii) Suppose $\chi_2(\mu_2,x)$ remains bounded as $x\to+\infty$. Let a denote the largest zero of $\chi_2(\mu_2,x)$, and set $b=+\infty$. Then the assertions of (i) remain true if we replace the interval (a,b) by $(a,+\infty)$.

(iii) If $0\leqslant\mu_1\leqslant\mu_2$, $\chi_2(\mu_2,x)$ is bounded as $x\to+\infty$, and $\chi_1(\mu_1,x)$ has no zeros in $\mathbf{R}$, then $\chi_2(\mu_2,x)$ has no zeros in $\mathbf{R}$ either.

(iv) If $\chi_2(\mu_2,x)$ is bounded as $x\to-\infty$ and a is the smallest zero of $\chi_2(\mu_2,x)$, then the assertion of (iii) holds, and the assertions in (i) remain true if we replace the interval (a,b) by $(-\infty,a)$.

Proof: We will omit the proof of (i) because on a finite interval such results are known (e.g., Theorem 1.1 on p. 208 of Ref. 20). Moreover, our proof of (ii) is easily modified to prove (i). The proof of (iv) is analogous to the proofs of (ii) and (iii), and hence we will only prove (ii) and (iii).

(ii) The proof can be given using contradiction. Without loss of generality we may assume that $\chi_1(\mu_1,x)$ and $\chi_2(\mu_2,x)$ are strictly positive in $(a,+\infty)$. When $b>a$, where a is the largest zero of $\chi_2(\mu_2,x)$, from (10.1) we get

$$\begin{aligned}&\chi_2(\mu_2,b)\chi_1'(\mu_1,b)-\chi_2'(\mu_2,b)\chi_1(\mu_1,b)+\chi_2'(\mu_2,a)\chi_1(\mu_1,a)\\&\quad=\int_a^b dx\,[V_1(\mu_1,x)-V_2(\mu_2,x)+\mu_1^2-\mu_2^2]\chi_1(\mu_1,x)\chi_2(\mu_2,x).\end{aligned} \tag{10.2}$$

Note that, by the asymptotic properties of the solutions and their assumed positivity in the interval $(a,+\infty)$, we have for $\mu_2>0$ and some $c_2>0$

$$\chi_2(\mu_2,x)=c_2e^{-\mu_2x}+o(e^{-\mu_2x})\quad \chi_2'(\mu_2,x)=-c_2\mu_2e^{-\mu_2x}+o(e^{-\mu_2x}),\quad x\to+\infty. \tag{10.3}$$

Furthermore, if $\chi_1(\mu_1,x)$ is unbounded as $x\to+\infty$, then for some $c_1>0$ we have

$$\chi_1(\mu_1,x)=c_1e^{\mu_1x}+o(e^{\mu_1x}),\quad \chi_1'(\mu_1,x)=c_1\mu_1e^{\mu_1x}+o(e^{\mu_1x}),\quad x\to+\infty. \tag{10.4}$$

If $\mu_1=0$ and $\chi_1(0,x)$ is unbounded as $x\to+\infty$, then for some $\tilde{c}_1>0$

$$\chi_1(0,x)=\tilde{c}_1x+o(x),\quad \chi_1'(0,x)=\tilde{c}_1+o(1),\quad x\to+\infty. \tag{10.5}$$

If $\chi_2(0,x)$ is bounded as $x\to+\infty$, then for some $\tilde{c}_2>0$ we have

$$\chi_2(0,x)=\tilde{c}_2+o(1), \quad \chi_2'(0,x)=o(1/x), \quad x\to+\infty. \tag{10.6}$$

Using (10.3)–(10.6) we will let $b\to+\infty$ in (10.2). When $0\leqslant\mu_1\leqslant\mu_2$, the limit as $b\to+\infty$ of the right-hand side of (10.2) exists and is nonpositive; it is equal to zero precisely when $\mu_1=\mu_2$ and $V_1(\mu_1,x)\equiv V(\mu_2,x)$ on $(a,+\infty)$. If $0\leqslant\mu_1<\mu_2$, the limit of the left-hand side of (10.2) is equal to $\chi_2'(\mu_2,a)\chi_1(\mu_1,a)$, which is nonnegative. Hence we have a contradiction, and thus $\chi_1(\mu_1,x)$ must have a zero in $(a,+\infty)$. If $0<\mu_1=\mu_2$ and $\chi_1(\mu_1,x)$ is unbounded, then the limit on the left-hand side of (10.2) is equal to $2c_1c_2\mu_1+\chi_2'(\mu_1,a)\chi_1(\mu_1,a)$, which is strictly positive; the right-hand side is nonpositive and so again we have a contradiction. If $0<\mu_1=\mu_2$ and $\chi_1(\mu_1,x)$ is bounded, then the limit of the left-hand side is equal to $\chi_2'(\mu_2,a)\chi_1(\mu_1,a)$, which is nonnegative. If also $V_1(\mu_1,x)\not\equiv V_2(\mu_2,x)$ on $(a,+\infty)$, then the right-hand side is strictly negative and we have a contradiction. If $V_1(\mu_1,x)\equiv V_2(\mu_2,x)$ on $(a,+\infty)$, then $\chi_1(\mu_1,a)>0$ due to the linear independence of $\chi_1(\mu_1,x)$ and $\chi_2(\mu_2,x)$, and so the left-hand side of (10.2) is strictly positive while its right-hand side is zero. In this case, by (i), there can only be one zero of $\chi_1(\mu_1,x)$ in $(a,+\infty)$. If $0=\mu_1=\mu_2$ and $\chi_1(0,x)$ is unbounded, then because of (10.6) the left-hand side of (10.2) approaches $\tilde{c}_1\tilde{c}_2+\chi_2'(0,a)\chi_1(0,a)$, which is again strictly positive. If $\mu_2=\mu_1=0$ and $\chi_1(0,x)$ is bounded, then the limit of the left-hand side of (10.2) is $\chi_2'(0,a)\chi_1(0,a)$, which is nonnegative. If $V_1(0,x)\not\equiv V_2(0,x)$ on $(a,+\infty)$, then the right-hand side of (10.2) is strictly negative, while if $V_1(0,x)\equiv V_2(0,x)$ on $(a,+\infty)$, then its right-hand side is zero and its left-hand side is strictly positive due to the linear independence of $\chi_1(0,x)$ and $\chi_2(0,x)$. In both cases we arrive at a contradiction. As in (i), if $V_1(0,x)\equiv V_2(0,x)$ on $(a,+\infty)$, we conclude that there is exactly one zero of $\chi_1(0,x)$ in $(a,+\infty)$.

(iii) Suppose $\chi_2(\mu_2,x)$ does have some zeros, the largest of which is a. Then, under the assumptions made in (i) and (ii), it follows that $\chi_1(\mu_1,x)$ has a zero to the right of a, contradicting the assumptions of (iii). The only situation not covered by (i) and (ii) is when $\mu_2=\mu_1$, $V_1(\mu_1,x)\equiv V_2(\mu_2,x)$ on $(a,+\infty)$, and $\chi_1(\mu_1,x)$ and $\chi_2(\mu_1,x)$ are linearly dependent in $(a,+\infty)$, but then $\chi_2(\mu_1,a)=0$ implies $\chi_1(\mu_1,a)=0$, which is again a contradiction. ■

From Proposition 10.1 we obtain the following:

Corollary 10.2: Assume $P,Q\in L^1(\mathbf{R})$ and let $\beta>0$. Suppose that $f_l^+(i\beta,x)$ and $f_r^+(i\beta,x)$ are linearly independent. Then $f_l^+(i\beta,x)$ and $f_r^+(i\beta,x)$ have the same number of zeros and their zeros are separated, i.e., between two successive zeros of $f_l^+(i\beta,x)$ there is a zero of $f_r^+(i\beta,x)$ and vice versa. Moreover, to the right of the largest zero of $f_l^+(i\beta,x)$ there is a zero of $f_r^+(i\beta,x)$, and to the left of the smallest zero of $f_r^+(i\beta,x)$ there is a zero of $f_l^+(i\beta,x)$.

Our next result concerns the zeros of the Jost solutions of (2.6). Since some theorems of this type have already been proved elsewhere (see, e.g., Theorem 14.10 of Ref. 22 or Theorem XIII.8 on p. 90 of Ref. 23), we only comment on certain details that may not be obvious from those references. Recall that $N(0,Q)$ denotes the number of bound states of (2.6).

Proposition 10.3: (i) Suppose $Q\in L^1(\mathbf{R})$ and assume $\beta>0$. Then the number of zeros of $f_l^{[0]}(i\beta,x)$ is equal to the number of bound states of (2.6) with energies contained in the interval $(-\infty,-\beta^2)$.

(ii) Suppose further that $Q\in L_1^1(\mathbf{R})$. Then the number of zeros of $f_l^{[0]}(0,x)$ is equal to $N(0,Q)$.

Proof: (i) Since we only assume $Q\in L^1(\mathbf{R})$, there may be infinitely many bound states of (2.6) with energies accumulating at zero. All such energies are negative, and let us denote them by $-\gamma_j^2$ with $\gamma_j>\gamma_{j+1}>0$ for $j\geqslant 1$. It is known (Theorem 14.10 of Ref. 22) that $f_l^{[0]}(i\gamma_j,x)$ has exactly $(j-1)$ zeros. Hence we only need to consider the zeros of $f_l^{[0]}(i\beta,x)$ when β is not equal to any γ_j. If $\beta>\gamma_1$, then from Proposition 10.1 (iii) with $V_1=V_2=Q$, $\mu_1=\gamma_1$, $\mu_2=\beta$, $\chi_1(\mu_1,x)=f_l^{[0]}(i\gamma_1,x)$, and $\chi_2(\mu_2,x)=f_l^{[0]}(i\beta,x)$, it follows that $f_l^{[0]}(i\beta,x)$ has no zeros. If $\beta\in(\gamma_{j+1},\gamma_j)$, then, by Proposition 10.1 (i) and (ii) with $\mu_1=\beta$ and $\mu_2=\gamma_j$, we observe that $f_l^{[0]}(i\beta,x)$ has at least j zeros. On the other hand, using Proposition 10.1, we can conclude that $f_l^{[0]}(i\beta,x)$ cannot have more than j zeros because the number of its zeros is nondecreasing as β decreases and $f_l^{[0]}(i\gamma_{j+1},x)$ has exactly j zeros. Thus $f_l^{[0]}(i\beta,x)$ has exactly j zeros when $\beta\in(\gamma_{j+1},\gamma_j)$. This proves (i) when $N(0,Q)=+\infty$ because the bound states of (2.6) can only occur when k is on the positive imaginary axis. If $N(0,Q)$ is finite and is denoted by $\mathcal{N}$, then we must still consider the case when $\beta\in(0,\gamma_{\mathcal{N}})$. Then using Lemma 1 on p. 91 of Ref. 23 we conclude

that $f_l^{[0]}(i\beta,x)$ has exactly $\mathcal{N}$ zeros because if it had more than $\mathcal{N}$ zeros one could find a subspace of dimension at least $\mathcal{N}+1$ on which the expectation value of $(-d^2/dx^2+Q-\beta P)$ is less than or equal to $-\beta^2$, and this would imply the existence of at least $(\mathcal{N}+1)$ eigenvalues less than or equal to $-\beta^2$.

(ii) In this case the condition $Q\in L_1^1(\mathbf{R})$ guarantees that $N(0,Q)$ is finite. It only remains to consider the case $\beta=0$. Note that $f_l^{[0]}(0,x)$ cannot have more than $\mathcal{N}$ zeros; this is because $f_l^{[0]}(i\beta,x)$ has exactly $\mathcal{N}$ zeros when β is sufficiently small and by (2.5) we see that as $\beta\to 0$ we have $f_l^{[0]}(i\beta,x)\to f_l^{[0]}(0,x)$ uniformly on compact x-intervals. On the other hand, in (10.1) by setting $\mu_1=0$, $\mu_2=\beta$, $V_1(\mu_1,x)=Q(x)$, $V_2(\mu_2,x)=Q(x)-\beta P(x)$, $V_1(\mu_1,x)=Q(x)$, and $\chi_2(\mu_2,x)=f_l^{[0]}(\beta,x)$, and using Proposition 10.1, we see that $f_l^{[0]}(0,x)$ has at least $\mathcal{N}$ zeros. Hence $f_l^{[0]}(0,x)$ must have exactly $\mathcal{N}$ zeros. ■

If $Q\in L_1^1(\mathbf{R})$, then $N(0,Q)$ is finite and as in Sec. IX we let $k=i\kappa_j$ for $j=1,...,\mathcal{N}$ denote the bound states of (2.6). From Theorems 9.7 and 9.9, when $P\in L^1(\mathbf{R})$, $Q\in L_1^1(\mathbf{R})$, and $P(x)\leqslant 0$, we already know that the bound states of (1.1) occur at $k=i\beta_j$ satisfying $\beta_j\leqslant\kappa_j$ for $j=1,...,\mathcal{N}$. In the next theorem, we extend Proposition 10.3 to (1.1) and analyze the number of zeros of the Jost solutions of (1.1) when k is on the positive imaginary axis.

Theorem 10.4: Assume that $P\in L^1(\mathbf{R})$, $Q\in L_1^1(\mathbf{R})$, and $P(x)\leqslant 0$. Then, for each $\beta\geqslant 0$, the functions $f_l^+(i\beta,x)$ and $f_r^+(i\beta,x)$ have the same number of zeros, and this number is equal to the number of bound states of (1.1) with energies contained in the interval $(-\infty,-\beta^2)$.

Proof: From Proposition 10.1 (i) and (ii), we see that $f_l^+(i\beta,x)$ and $f_r^+(i\beta,x)$ have the same number of zeros. Since $P(x)\leqslant 0$, from the proof of Theorem 9.7 it follows that, for any fixed $\beta>0$, the number of eigenvalues of the operator $(-d^2/dx^2+Q-\beta P)$ below $-\beta^2$ is equal to the number of $E_j(\beta)$-values that lie below $-\beta^2$. Note that if $\beta\in[\beta_j,\beta_{j+1})$ for $j=1,...,\mathcal{N}-1$, then the $(\mathcal{N}-j)$ values $E_{\mathcal{N}}(\beta),E_{\mathcal{N}-1}(\beta),\ldots,E_{j+1}(\beta)$ lie strictly below $-\beta^2$; if $\beta\in[\beta_{\mathcal{N}},+\infty)$ then there are no eigenvalues below $-\beta^2$, and if $\beta\in[0,\beta_1)$ then exactly $\mathcal{N}$ eigenvalues lie below $-\beta^2$. Using Proposition 10.3 when the potential $Q(x)$ in (2.6) is replaced by $Q(x)-\beta P(x)$, we conclude that $f_l^+(i\beta,x)$ has no zeros for $\beta\in[\beta_{\mathcal{N}},+\infty)$, $\mathcal{N}$ zeros for $\beta\in[0,\beta_1)$, and $(\mathcal{N}-j)$ zeros for $\beta\in[\beta_j,\beta_{j+1})$ for $j=1,...,\mathcal{N}-1$. ■

If we weaken the condition $P(x)\leqslant 0$ in Theorem 10.4, the number of bound states may not be easily related to the number of zeros of the Jost solutions, as we will see in Example 11.2. From Proposition 10.1 and Theorem 10.4 we have the following:

Corollary 10.5: Assume that $P,Q\in L^1(\mathbf{R})$ and that $P(x)\leqslant 0$. The zeros of $f_l^+(i\beta_j,x)$ separate the zeros of $f_l^+(i\beta_{j+1},x)$, i.e., between two consecutive zeros of $f_l^+(i\beta_j,x)$ there is exactly one zero of $f_l^+(i\beta_{j+1},x)$, where $k=i\beta_j$ for $j=1,...,\mathcal{N}$ correspond to the bound states of (1.1). Similarly, the zeros of $f_r^+(i\beta_j,x)$ separate the zeros of $f_r^+(i\beta_{j+1},x)$.

When we no longer have $P(x)\leqslant 0$, then there may be bound states of (1.1) with complex energies and $N(P,Q)$ may be larger than $N(0,Q)$. We refer the reader to Examples 11.2 and 11.4. In the next theorem, when $P\in L^1(\mathbf{R})$ and $Q\in L_1^1(\mathbf{R})$, we analyze the bound states of (1.1) when k is on the positive imaginary axis, establish the connection between Theorem 8.4 and the zeros of $f_r^+(i\beta,x)$, and also consider multiple zeros of $1/T^+(k)$ on the positive imaginary axis.

Theorem 10.6: Suppose $P\in L^1(\mathbf{R})$ and $Q\in L_1^1(\mathbf{R})$. Then:

(i) If (2.6) has $\mathcal{N}$ bound states with $\mathcal{N}\geqslant 1$, then (1.1) has at least $\mathcal{N}$ bound states with real (negative) energies.

(ii) $1/T^+(i\beta)$ has a zero of order m at some positive β_0 if and only if the function $E_0(\beta)+\beta^2$ has a zero of order m at β_0, where $E_0(\beta)$ denotes the unique eigenvalue branch of the operator $(-d^2/dx^2+Q-\beta P)$ satisfying $E_0(\beta)\to-\beta_0^2$ as $\beta\to\beta_0$. If $m=1$, then the graph of $E_0(\beta)$ and the graph of the parabola $E=-\beta^2$ intersect with different slopes at β_0. If $m\geqslant 2$ and m is even, then the graphs touch at β_0 but do not cross each other. If $m\geqslant 3$ and m is odd, then the graphs cross smoothly such that at the point of intersection they have the same slope.

(iii) The lowest eigenvalue branch $E_{\mathcal{N}}(\beta)$ satisfies $E''_{\mathcal{N}}(\beta)<0$ for $\beta>0$ unless $P(x)\equiv 0$, in which case $E_{\mathcal{N}}(\beta)$ is a constant. Hence the graph of the lowest eigenvalue branch is concave down if $P(x)\not\equiv 0$.

(iv) The number of zeros of $f_l^+(i\beta,x)$ behaves in the following manner as β is increased from $\beta_0-\epsilon$ to $\beta_0+\epsilon$ when ϵ is sufficiently small: If m is even, then the number of zeros is either constant throughout the interval $(\beta_0-\epsilon,\beta_0+\epsilon)$ or it is constant in $(\beta_0-\epsilon,\beta_0)\cup(\beta_0,\beta_0+\epsilon)$ but

one less at β_0. If m is odd, then the number of zeros either increases or decreases by one as β crosses β_0. The number of zeros of $f_l^+(i\beta,x)$ can only change at β-values corresponding to bound states of (1.1).

Proof: (i) Since we are only interested in the bound states corresponding to $k=i\beta$ for $\beta>0$, we first derive a lower bound for $E_{\mathscr{N}}(\beta)$, which will show that $E_{\mathscr{N}}(\beta)=o(\beta^2)$ as $\beta\to+\infty$. Let us indicate the Fourier transform by a caret:

$$\hat{\psi}(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dx\ e^{iqx}\psi(x),\quad \psi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dq\ e^{-iqx}\hat{\psi}(q).$$

Then, letting $\|\cdot\|_2$ denote the norm on $L^2(\mathbf{R})$, for $a>0$ we get (cf. Theorem IX.28 of Ref. 24)

$$|\psi(x)|\leqslant\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dq\ |\hat{\psi}(q)|\leqslant\frac{1}{\sqrt{2a}}\left[\int_{-\infty}^{\infty}dq\ (q^2+a^2)|\hat{\psi}(q)|^2\right]^{1/2}=\frac{1}{\sqrt{2a}}\,(\|\psi'\|_2^2+a^2\|\psi\|_2^2)^{1/2},\tag{10.7}$$

where we have used the Schwarz inequality, $\|\psi\|_2=\|\hat{\psi}\|_2$, $\int_{-\infty}^{\infty}dq/(a^2+q^2)=\pi/a$, and $\hat{\psi}'=-iq\hat{\psi}$. Next we use (10.7) to estimate the quadratic forms $\langle Q\psi,\psi\rangle$ and $\langle P\psi,\psi\rangle$. From (10.7) we obtain

$$\int_{-\infty}^{\infty}dx\ |Q(x)||\psi(x)|^2\leqslant\frac{1}{2a}\,(\|\psi'\|_2^2+a^2\|\psi\|_2^2)\left(\int_{-\infty}^{\infty}dx|Q(x)|\right).\tag{10.8}$$

If $P(x)\equiv0$, then (1.1) and (2.6) become identical, and in this trivial case both equations have exactly $\mathscr{N}$ bound states. Thus there is no loss of generality in assuming $P(x)\not\equiv0$. In order to estimate the integral $\int_{-\infty}^{\infty}dx\ |P(x)||\psi(x)|^2$, we split it into two parts: one over the region $\{x:|P(x)|>M\}$ and the other over the region $\{x:|P(x)|\leqslant M\}$, where the constant $M\geqslant0$ is arbitrary for the moment but will be fixed later. Then, for any $b>0$, (10.7) implies

$$\int_{-\infty}^{\infty}dx\ |P(x)||\psi(x)|^2\leqslant\frac{1}{2b}\,(\|\psi'\|_2^2+b^2\|\psi\|_2^2)\left(\int_{\{|P(x)|>M\}}dx\ |P(x)|\right)+M\|\psi\|_2^2.\tag{10.9}$$

Combining (10.8) and (10.9) we get

$$\langle-\psi''+Q\psi-\beta P\psi,\psi\rangle\geqslant J_1\|\psi'\|_2^2-J_2\|\psi\|_2^2,$$

where

$$J_1=1-\frac{1}{2a}\int_{-\infty}^{\infty}dx\ |Q(x)|-\frac{\beta}{2b}\int_{\{|P(x)|>M\}}dx\ |P(x)|,$$

$$J_2=\frac{a}{2}\int_{-\infty}^{\infty}dx\ |Q(x)|+\frac{\beta b}{2}\int_{\{|P(x)|>M\}}dx\ |P(x)|+\beta M.$$

We now set

$$a=\int_{-\infty}^{\infty}dx\ |Q(x)|,\quad b=\beta\int_{\{|P(x)|>M\}}dx\ |P(x)|,$$

and assume that ψ is a normalized eigenfunction corresponding to the eigenvalue $E_{\mathscr{N}}(\beta)$. Then the left-hand side of (10.9) is equal to $E_{\mathscr{N}}(\beta)$ and hence

$$E_{\mathscr{N}}(\beta)\geqslant-\frac{1}{2}\left(\int_{-\infty}^{\infty}dx\ |Q(x)|\right)^2-\frac{\beta^2}{2}\left(\int_{\{|P(x)|>M\}}dx\ |P(x)|\right)^2-\beta M.\tag{10.10}$$

Since by choosing M large enough we can make the second term on the right-hand side of (10.10) as small as we please, it follows that $E_{\mathscr{N}}(\beta)=o(\beta^2)$ as $\beta\to+\infty$. Thus $E_{\mathscr{N}}(\beta)>-\beta^2$ for β sufficiently large, while $E_{\mathscr{N}}(0)=-\kappa_{\mathscr{N}}^2<0$. Hence by the intermediate value theorem, the equation $E_{\mathscr{N}}(\beta)=-\beta^2$ has at least one solution. A similar argument shows that each of the remaining eigenvalue branches $E_j(\beta)$ for $j=1,...,\mathscr{N}-1$ must intersect the parabola $E=-\beta^2$ at least once. Since each intersection increases the number of negative-energy bound states of (1.1), the proof of (i) is complete. Note that if an eigenvalue branch $E_j(\beta)$ touches or intersects the parabola $E=-\beta^2$ at other points, such additional points are also responsible for additional negative-energy bound states of (1.1). Moreover, there may be other eigenvalue branches $E(\beta)$ starting at $(\widetilde{\beta},0)$ for some $\widetilde{\beta}>0$ and intersecting or touching the parabola $E=-\beta^2$ at one or more points; again, each of such points also increases the number of negative-energy bound states of (1.1).

(ii) If $P(x)\equiv 0$, each eigenvalue branch $E_0(\beta)$ becomes the horizontal line $E_0(\beta)=-\beta_0^2$ for $\beta\geqslant 0$, and hence $E_0''(\beta)=0$ for $\beta>0$. Thus in the rest of the analysis we can assume that $P(x)\not\equiv 0$. Associated with the eigenvalue $E_0(\beta)$ there exists[25] a real-valued, analytic eigenvector $\psi(\beta,x)$. Near $\beta=\beta_0$ we have the convergent expansions

$$E_0(\beta)=\sum_{n=0}^{\infty} a_n(\beta-\beta_0)^n, \quad \psi(\beta,x)=\sum_{n=0}^{\infty}\psi_n(x)(\beta-\beta_0)^n, \tag{10.11}$$

with $\psi_n\in L^2(\mathbf{R})$ for $n\geqslant 0$. Substituting (10.11) in (1.1) we get the following set of equations (see pp. 333 and 334 of Ref. 26) for $n\geqslant 0$:

$$\psi_n''(x)-\beta_0^2\psi_n(x)+a_1\psi_{n-1}(x)+a_2\psi_{n-2}(x)=[-\beta_0 P(x)+Q(x)]\psi_n(x)-P(x)\psi_n(x) - \sum_{j=3}^{n} a_j\psi_{n-j}(x), \tag{10.12}$$

where it is assumed that $a_{-n}=\psi_{-n}(x)=0$ if $n\geqslant 1$. From (9.17) and (9.18) we see that

$$a_0=E_0(\beta_0), \quad a_1=-\frac{1}{\|\psi_0\|_2^2}\int_{-\infty}^{\infty} dx\ P(x)\psi_0(x)^2. \tag{10.13}$$

We may choose $\psi_0(x)=f_l^+(i\beta_0,x)$. It suffices to prove that $1/T^+(i\beta)$ has a zero of order at least m at β_0 if and only if $E_0(\beta)+\beta^2$ has a zero of order at least m at β_0. From Proposition 8.1 we know that this is true when $m=1$. If β_0 is a zero of $E_0(\beta)+\beta^2$ of order m for some $m\geqslant 2$, then the coefficients a_n in (10.3) are determined for $n=0,1,...,m-1$ by expanding $E_0(\beta)+\beta^2$ about β_0. Thus for $m=2$ we get $a_0=-\beta_0^2$, $a_1=-2\beta_0$; for $m=3$ we get $a_0=-\beta_0^2$, $a_1=-2\beta_0$, $a_2=-1$; for $m\geqslant 4$ we get $a_0=-\beta_0^2$, $a_1=-2\beta_0$, $a_2=-1$, and $a_3=\cdots=a_{m-1}=0$. Then, comparing (10.12) and (8.2) and using the fact that the functions $g_{l,n}^+(i\beta_0,x)$ are uniquely determined as solutions of (8.2) by the requirement that $g_{l,n}^+(i\beta_0,\cdot)\in L^2(\mathbf{R})$, we obtain

$$\psi_n(x)=i^n g_{l,n}^+(i\beta_0,x), \quad n=0,...,m-1. \tag{10.14}$$

Thus, by Theorem 8.4, we see that $1/T^+(i\beta)$ has a zero of order at least m at β_0. Conversely, suppose $1/T^+(i\beta)$ has a zero of order at least m at β_0. From (10.12) one can derive (see p. 334 of Ref. 26) the following recursion formula for the coefficients a_n:

$$a_n=-\frac{1}{\|\psi_0\|_2^2}\int_{-\infty}^{\infty} dx\ \psi_0(x)\left(P(x)\psi_{n-1}(x)+\sum_{j=1}^{n-1} a_j\psi_{n-j}(x)\right), \quad n\geqslant 2. \tag{10.15}$$

Now assume $m\geqslant 2$. Since the functions $\{g_{l,n}^+(i\beta_0,\cdot)\}_{n=0}^{m-1}$ form a Jordan chain of length m, using (8.13) with $n=1$ and (10.13) we get $a_1=-2\beta_0$. Hence $E_0(\beta)+\beta^2$ has a zero of order at least 2 at β_0. If $m\geqslant 3$, then using (8.13) with $n=2$, (10.14), and (10.15) we obtain $a_2=-1$ and this, in turn, implies that $E_0(\beta)+\beta^2$ has a zero of order at least 3 at β_0. If $m\geqslant 4$, then (8.13) and (10.15) give $a_3=0$ and then $a_j=0$ for all $j=3,...,m-1$. As a result, $E_0(\beta)+\beta^2$ has a zero of order at

least m at β_0. If m is even, then $E_0(\beta)$ touches the parabola $E=-\beta^2$ at β_0 but stays either above or below the parabola; if m is odd, then $E_0(\beta)$ intersects the parabola by crossing from one side to the other.

(iii) For the lowest eigenvalue, we need to show that a_2 in (10.11) is negative for any $\beta_0>0$. We have (see p. 334 of Ref. 26)

$$\psi_1(x)=-\psi_0(x)\int_{x_0}^{x}\frac{dt}{\psi_0(t)^2}\int_{-\infty}^{t}ds\ \psi_0(s)[P(s)\psi_0(s)+a_1\psi_0(s)], \tag{10.16}$$

where the constant x_0 is arbitrary; however, since changing x_0 amounts to adding a constant multiple of $\psi_0(x)$ to $\psi_1(x)$, with the help of (10.6) one can show that the value of a_n given in (10.15) is independent of x_0. Using (10.16) in (10.15) with $n=2$ and the positivity of $\psi_0(x)$, after performing an integration by parts, we get

$$a_2=-\frac{1}{\|\psi_0\|_2^2}\int_{-\infty}^{\infty}\frac{dx}{\psi_0(x)^2}\left(\int_{-\infty}^{x}dt\ \psi_0(t)^2[P(t)+a_1]\right)^2<0.$$

Thus $a_2<0$ in (10.11) for any $\beta_0>0$, and hence we have $E_{\mathcal{N}}''(\beta)<0$ for any $\beta>0$.

(iv) Let us consider the number of zeros of $f_l^+(i\beta,x)$ in relation to the behavior of the eigenvalue branch $E_0(\beta)$ near β_0. From Proposition 10.3 (i), when $Q(x)$ in (2.6) is replaced by $Q(x)-\beta P(x)$, we know that the number of zeros of $f_l^+(i\beta,x)$ is equal to the number of eigenvalue branches lying below $-\beta^2$. Let I_{β_0} denote the interval $(\beta_0-\epsilon,\beta_0+\epsilon)$ and let J_{β_0} denote $(\beta_0-\epsilon,\beta_0)\cup(\beta_0,\beta_0+\epsilon)$ for sufficiently small $\epsilon>0$, and let us consider the number of eigenvalue branches below $-\beta^2$ when $\beta\in I_{\beta_0}$. If m is even, then $E_0(\beta)$ touches the parabola $E=-\beta^2$ at β_0 but stays either above or below that parabola; in the former case $E_0(\beta)>-\beta^2$ for $\beta\in J_{\beta_0}$ and hence the number of zeros of $f_l^+(i\beta,x)$ remains unchanged for $\beta\in I_{\beta_0}$; in the latter case $E_0(\beta)<-\beta^2$ for $\beta\in J_{\beta_0}$ and hence the number of zeros of $f_l^+(i\beta,x)$ for $\beta\in J_{\beta_0}$ is exactly one more than the number of zeros of $f_l^+(i\beta_0,x)$. If m is odd, then $E_0(\beta)$ intersects the parabola $E=-\beta^2$ by crossing from one side to the other of that parabola; if $E_0(\beta)<-\beta^2$ on $(\beta_0-\epsilon,\beta_0)$, then the number of zeros of $f_l^+(i\beta,x)$ decreases by one as β increases through β_0; if $E_0(\beta)>-\beta^2$ on $(\beta_0-\epsilon,\beta_0)$, then the number of zeros increases by one as β increases through β_0. In order to prove that the number of zeros of $f_l^+(i\beta,x)$ can only change if β corresponds to a bound state of (1.1) with real (negative) energy, we can proceed as follows. If β_1 and β_2 with $\beta_1<\beta_2$ correspond to two consecutive real bound-state energies of (1.1), then no eigenvalue branch can intersect the parabola $E=-\beta^2$ for $\beta\in(\beta_1,\beta_2)$. Hence the number of eigenvalue branches that lie below $-\beta^2$ is constant for $\beta\in(\beta_1,\beta_2)$, or equivalently, the number of zeros of $f_l^+(i\beta,x)$ is constant for $\beta\in(\beta_1,\beta_2)$. ■

We remark that statement (iii) is a familiar result that also follows from the min-max principle. An example illustrating Theorem 10.6 is given in the next section (Example 11.3).

Proposition 10.7: Assume $P,Q\in L^1(\mathbf{R})$ and let $k_0=\alpha+i\beta$ for some $\alpha\neq 0$ and $\beta>0$. If $P(x)\leqslant 2\beta$, then $f_l^+(k_0,x)$ and $f_r^+(k_0,x)$ cannot vanish for any $x\in\mathbf{R}$.

Proof: Using (4.7) in (4.16) we obtain

$$\frac{d}{dx}[f_l^+(-\overline{k_0},x);f_l^+(k_0,x)]=2i\alpha[P(x)-2\beta]|f_l^+(k_0,x)|^2. \tag{10.17}$$

Suppose $f_l^+(k_0,x)$ has at least one zero and let d be the right-most zero of $f_l^+(k_0,x)$. Note that, as seen from (4.7), the zeros of $f_l^+(-\overline{k_0},x)$ and $f_l^+(k_0,x)$ coincide. Integrating (10.17) over $(d,+\infty)$ and using (8.1) and (8.2), we obtain

$$2i\alpha\int_{d}^{\infty}dx\ [P(x)-2\beta]|f_l^+(k_0,x)|^2=0.$$

This is impossible if $\alpha\neq 0$ and $P(x)\leqslant 2\beta$; note that $P(x)=2\beta$ on a semi-infinite interval would contradict $P\in L^1(\mathbf{R})$. Hence $f_l^+(k_0,x)$ cannot vanish for any $x\in\mathbf{R}$. The proof for $f_r^+(k_0,x)$ is analogous. ■

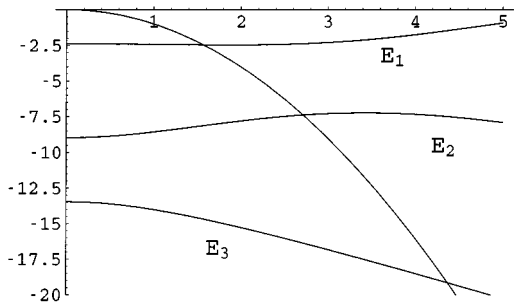

FIG. 1. Eigenvalue curves $E_1(\beta)$, $E_2(\beta)$, $E_3(\beta)$ intersecting the parabola $E=-\beta^2$.

XI. EXAMPLES

In this section we illustrate our results on the number and location of the zeros of $1/T^+(k)$ and on the zeros of the Jost solutions with some explicit examples. Example 11.1 shows the typical behavior of some eigenvalues branches as it was discussed in Secs. IX and X. Example 11.2 exhibits the possibility of real zeros of $1/T^+(k)$ and of complex zeros off the imaginary axis in $\mathbf{C}^+$. In Example 11.3 we consider the zeros on the imaginary axis and the corresponding eigenvalue branches. We also demonstrate the possibility of a double zero on the positive imaginary axis. While Examples 11.1–11.3 involve simple step potentials, Example 11.4 concerns potentials that decay exponentially.

Example 11.1: For real parameters $a_\pm$ and $b_\pm$, let

$$P(x)=\begin{cases} b_+, & x\in(0,1),\\ b_-, & x\in(-1,0),\\ 0, & \text{elsewhere,}\end{cases} \qquad Q(x)=\begin{cases} a_+, & x\in(0,1),\\ a_-, & x\in(-1,0),\\ 0, & \text{elsewhere.}\end{cases} \tag{11.1}$$

Then by straightforward calculations one obtains

$$\frac{e^{-2ik}}{T^+(k)}=\cos s_+ \cos s_- + F_+ \sin s_+ \cos s_- + F_- \cos s_+ \sin s_- - G \sin s_+ \sin s_-, \tag{11.2}$$

where

$$s_\pm=\sqrt{k^2-ikb_\pm-a_\pm}, \quad F_\pm=\frac{k^2+s_\pm^2}{2iks_\pm}, \quad G=\frac{s_+^2+s_-^2}{2s_+s_-}.$$

For $a_\pm=-15$, $b_+=2$, and $b_-=-2$ we get three purely imaginary bound states at $k=1.568\overline{i}$, $k=2.717i$, and $k=4.376\overline{i}$. Here and below we use an overline on the last digit to indicate round-off. These are all the bound states because a plot of $\arg T^+(k)$ for $k\in\mathbf{R}^+$ reveals that $\arg T^+(0+)=5\pi/2$, so that by (9.22) we have $N(P,Q)=3$; note that we are in the generic case. Figure 1 shows the eigenvalue curves associated with the potential $V(\beta,x)=Q(x)-\beta P(x)$ [cf. (9.16)]. The analysis shows that the branches $E_2(\beta)$ and $E_1(\beta)$ are not concave, only $E_3(\beta)$ is concave down. This, in particular, illustrates part (iii) of Theorem 10.6.

Example 11.2: In this example we demonstrate the existence of nonreal zeros of $1/T^+(k)$. Putting $a_-=0$, $a_+=a$, $b_-=0$, and $b_+=b$ in (11.1) and (11.2), we get

$$\frac{1}{T^+(k)}=e^{ik}\left[\cos s+\frac{k^2+s^2}{2iks}\sin s\right], \tag{11.3}$$

where $s=\sqrt{k^2-ibk-a}$. By Theorem 9.3, if $a=0$ and $b>2$, then we must have a bound state at $k=i\beta$ for some positive β. Indeed, if $a=0$ and $b=21/10$, we obtain a bound state at $k=0.15\overline{i}$ and this is the only bound state. When $a=-9$ and $b=5$, we find bound states at $k_1=5.619i$, $k_2^\pm$

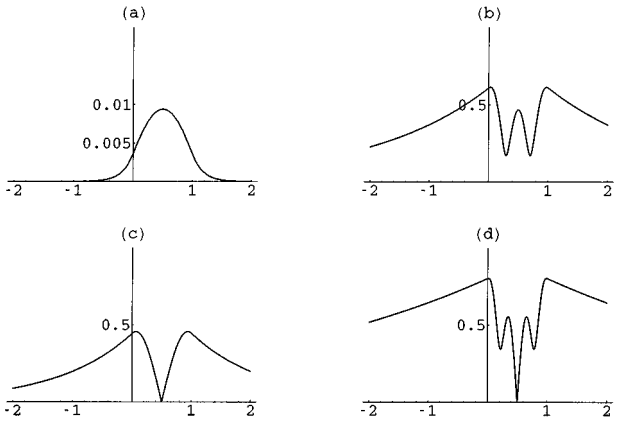

FIG. 2. (a) $f_l^+(k_1,x)$, (b) $|f_l^+(k_2^+,x)|$, (c) $|f_l^+(k_3^+,x)|$, and (d) $|f_l^+(k_4^+,x)|$.

$=\pm 6.28\overline{8}+0.49\overline{5}i$, $k_3^{\pm}=\pm 0.83\overline{8}+0.8\overline{1}i$, and $k_4^{\pm}=\pm 9.89\overline{8}+0.219\overline{2}i$. As in Example 11.1 it follows that these are all the bound states. In Fig. 2 we plot $f_l^+(k_1,x)$, $|f_l^+(k_2^+,x)|$, $|f_l^+(k_3^+,x)|$, and $|f_l^+(k_4^+,x)|$ on the interval $x\in(-2,2)$. Let us remark that the functions $f_l^+(k_1,x)$ and $f_l^+(k_2^{\pm},x)$ are even with respect to $x=1/2$, and the functions $f_l^+(k_3^{\pm},x)$ and $f_l^+(k_4^{\pm},x)$ are odd with respect to $x=1/2$. When $a=-9.2738$ and $b=3.9708$, we obtain a zero of $1/T^+(k)$ at $k=1$. In the special case when $a=0$ we have

$$\frac{1}{T^+(0)}=1-\frac{b}{2},\quad \frac{L^+(0)}{T^+(0)}=\frac{R^+(0)}{T^+(0)}=\frac{b}{2}.$$

Hence $1/T^+(0)=0$ if and only if $b=2$. Note that the bound states may occur even when $Q(x)\geqslant 0$. For example, when $a=1$ and $b=10$, we obtain over 200 bound states, four of which correspond to k values on the positive imaginary axis with $k=i\beta_j$, where

$$\beta_1=0.1\overline{3},\quad \beta_2=2.5\overline{0},\quad \beta_3=5.6\overline{3},\quad \beta_4=9.1\overline{6}.$$

When $a=0$ and $b=100$, we obtain 31 bound states on the positive imaginary axis with $k=i\beta_j$, where

$$\beta_1=0.1\overline{0},\quad \beta_2=0.4\overline{1},\quad \beta_3=0.9\overline{3},\quad \beta_4=1.6\overline{7},\quad \beta_5=2.6\overline{4},\quad \beta_6=3.8\overline{5},\quad \beta_7=5.3\overline{3},$$

$$\beta_8=7.0\overline{9},\quad \beta_9=9.1\overline{9},\quad \beta_{10}=11.6\overline{9},\quad \beta_{11}=14.6\overline{3},\quad \beta_{12}=18.2\overline{0},\quad \beta_{13}=22.6\overline{1},$$

$$\beta_{14}=28.4\overline{3},\quad \beta_{15}=37.6\overline{3},\quad \beta_{16}=60.4\overline{1},\quad \beta_{17}=69.6\overline{9},\quad \beta_{18}=75.6\overline{0},\quad \beta_{19}=80.1\overline{1},$$

$$\beta_{20}=83.7\overline{7},\quad \beta_{21}=86.8\overline{3},\quad \beta_{22}=89.4\overline{2},\quad \beta_{23}=91.6\overline{3},\quad \beta_{24}=93.5\overline{2},\quad \beta_{25}=95.1\overline{2},$$

$$\beta_{26}=96.4\overline{6},\quad \beta_{27}=97.5\overline{7},\quad \beta_{28}=98.4\overline{6},\quad \beta_{29}=99.1\overline{4},\quad \beta_{30}=99.6\overline{2},\quad \text{and}\quad \beta_{31}=99.9\overline{1},$$

and there are also many more bound states corresponding to k-values off the imaginary axis in $\mathbf{C}^+$. In this case, one finds that $f_l^+(i\beta,x)$ has no zeros for $\beta=0$, no zeros for $\beta\in(\beta_{31},+\infty)$, one zero for $\beta\in(0,\beta_1)$, and one zero for $\beta\in(\beta_{30},\beta_{31})$, j zeros for $\beta\in(\beta_{j-1},\beta_j)$ and j zeros for $\beta\in(\beta_{31-j},\beta_{32-j})$ with $j=2,3,...,15$, and 16 zeros for $\beta\in(\beta_{16},\beta_{17})$.

Example 11.3: (a) Let $a=0$ and $b=10$ in Example 11.2, and hence $Q(x)=0$ and $P(x)\geqslant 0$. Note that $N(0,Q)=0$. From (9.16) we obtain [cf. (11.3)]

$$2\Delta\sqrt{-E}\cos\Delta=(\Delta^2+E)\sin\Delta, \tag{11.4}$$

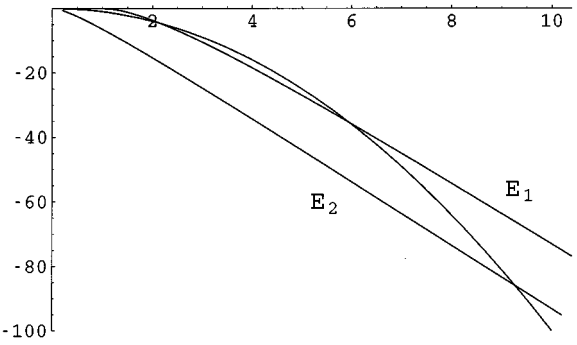

FIG. 3. The parabola $E=-\beta^2$ intersecting the eigenvalue curves $E_1(\beta)$ and $E_2(\beta)$.

where $E=E(\beta)$ is the energy in (9.16) and $\Delta=\sqrt{\beta b-a+E}$. Using the half-angle formula for the tangent function, we can write (11.4) as a pair of equations determining the eigenvalue curves:

$$\tan\left(\frac{\Delta}{2}\right)=\frac{\sqrt{-E}}{\Delta},\quad \tan\left(\frac{\Delta}{2}\right)=-\frac{\Delta}{\sqrt{-E}}. \tag{11.5}$$

Recall from the proof of Theorem 9.7 that the bound states of (1.1) with real energies correspond to the β-values where the eigenvalue curves intersect the parabola $E=-\beta^2$. When $a=0$ and $b=10$, from (11.5) we obtain two eigenvalue branches intersecting the parabola $E=-\beta^2$. Let $E_2(\beta)$ denote the eigenvalue branch responsible for the lowest real bound-state energy. We see that $E_2(\beta)$ emerges from (0,0) and intersects the parabola $E=-\beta^2$ at $\beta_3=9.273\overline{3}$. The second eigenvalue branch, $E_1(\beta)$, emerges from zero at $\beta=\pi^2/10$ and then intersects the parabola $E=-\beta^2$ at $\beta_1=2.114$ and at $\beta_2=5.963$. These eigenvalue branches and the parabola $E=-\beta^2$ are plotted in Fig. 3. The values β_1, β_2, and β_3 correspond to simple zeros of $1/T^+(i\beta)$. If $\beta\geqslant\beta_3$, then $f_l^+(i\beta,x)$ has no zeros. If $\beta\in[\beta_2,\beta_3)$, then $f_l^+(i\beta,x)$ has one zero because $E_2(\beta)$ is the only eigenvalue below $-\beta^2$. If $\beta\in(\beta_1,\beta_2)$, then $f_l^+(i\beta,x)$ has two zeros because both $E_2(\beta)$ and $E_1(\beta)$ lie below $-\beta^2$. If $\beta\in(0,\beta_1]$ then $f_l^+(i\beta,x)$ has one zero, and if $\beta=0$ then $f_l^+(0,x)$ has no zeros because (2.6) with $Q(x)=0$ has no bound states.

(b) When $a=0$, one can choose the parameter b such that the branch $E_1(\beta)$ just touches the parabola $E=-\beta^2$ at β_1. Then the slope of the eigenvalue curve at β_1 must be equal to $-2\beta_1$, and this happens when

$$\tan\left(\frac{\sqrt{b^2-4}}{4}\right)=-\frac{\sqrt{b+2}}{\sqrt{b-2}},$$

from which we get $b=9.206\overline{6}$, leading to $\beta_1=3.60\overline{3}$, and β_1 corresponds to a double zero of $1/T^+(i\beta)$. The eigenvalue branch $E_2(\beta)$ intersects the parabola $E=-\beta^2$ at β_2; we have $\beta_2=8.433$, which is responsible for the lowest real bound-state energy. In this case, $f_l^+(i\beta,x)$ has no zeros for $\beta=0$, one zero for $\beta\in(0,\beta_2)$, and no zeros for $\beta\in[\beta_2,+\infty)$. We show the two eigenvalue branches and the parabola $E=-\beta^2$ in Fig. 4.

(c) When $b=10$ we can find a such that the lowest real bound-state energy corresponds to a double zero of $1/T^+(i\beta)$. Proceeding as in (b), we get

$$a=\frac{10\beta_1(\beta_1-4)}{\beta_1-3},$$

where β_1 is obtained by solving

$$\tan\left(\frac{\sqrt{10\beta+3\beta^2-\beta^3}}{2\sqrt{\beta-3}}\right)=\frac{\beta\sqrt{\beta-3}}{\sqrt{10\beta+3\beta^2-\beta^3}}. \tag{11.6}$$

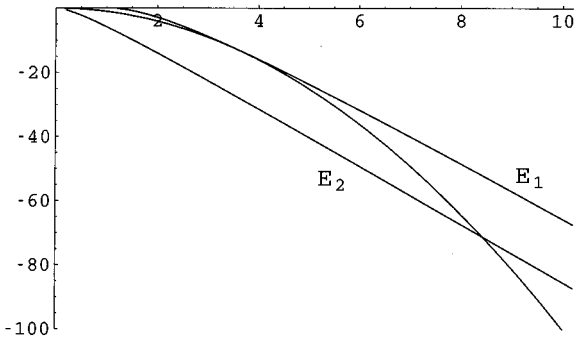

FIG. 4. The parabola $E=-\beta^2$ touching $E_1(\beta)$ and intersecting $E_2(\beta)$.

From (11.6) we get $\beta_1=4.724\overline{}$ and hence $a=19.852\overline{}$. In this case $f_l^+(i\beta,x)$ has no zeros for any $\beta\geqslant 0$. The eigenvalue curve $E_1(\beta)$ and the parabola $E=-\beta^2$ are plotted in Fig. 5, and it is seen that there are no other real bound-state energies besides $-\beta_1^2$.

Example 11.4: Let

$$\eta_l^+(k,x)=1-\frac{2i(1+b)\epsilon}{k+i\epsilon}\,\frac{ce^{-2\epsilon x}}{1+ce^{-2\epsilon x}},\quad x\geqslant 0, \tag{11.7}$$

$$\eta_r^+(k,x)=1-\frac{2i(1+b)\epsilon}{k+i\epsilon}\,\frac{ce^{2\epsilon x}}{1+ce^{2\epsilon x}},\quad x\geqslant 0, \tag{11.8}$$

where c, ϵ are positive parameters and b is a real parameter. Using (1.1), (1.2), (3.3), (3.4), (11.7), and (11.8), we obtain

$$P(x)=\frac{4b\epsilon ce^{-2\epsilon|x|}}{1+ce^{-2\epsilon|x|}},$$

$$Q(x)=\frac{4\epsilon^2 ce^{-2\epsilon|x|}[-3b-2+b^2ce^{-2\epsilon|x|}]}{(1+ce^{-2\epsilon|x|})^2}.$$

Hence $e^p=(1+c)^{2b}$ and using (6.3) we get

$$T^+(k)=\frac{k(k+i\epsilon)^2e^p}{(k-k_0)(k-k_+)(k-k_-)},$$

where we have defined

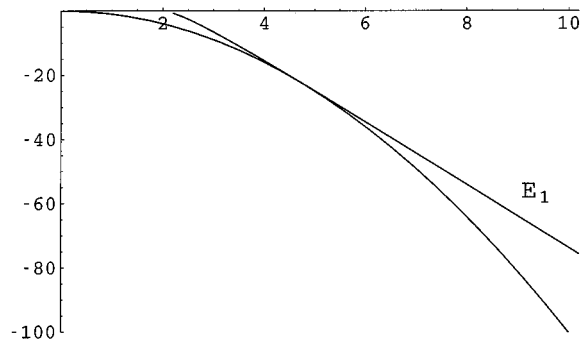

FIG. 5. The parabola $E=-\beta^2$ touching the eigenvalue curve $E_1(\beta)$.

$$k_0 = i\,\frac{\epsilon}{1+c}\,[-1+c+2bc],$$

$$k_\pm = \frac{i\epsilon}{2(1+c)}\,[(-1+c+4bc) \pm \sqrt{1+c^2+14c+16bc}]. \tag{11.9}$$

Let us now analyze the poles of $T^+(k)$. First note that ϵ simply acts as a scaling parameter for the location of the poles; thus the relevant parameters are b and c. We can divide the half-plane $\{(c,b): c>0\}$ into four separate regions by using the three nonintersecting curves $b=\Gamma_+(c)$, $b=\Gamma_0(c)$, and $b=\Gamma_-(c)$, where

$$\Gamma_0(c) = \frac{1-c}{2c}, \quad \Gamma_\pm(c) = -\frac{c-3}{4c} \pm \sqrt{\frac{(c-3)^2}{16c^2} + \frac{1}{c}}.$$

On these three curves the exceptional case occurs; note that $k_-=0$ on Γ_+, $k_+=0$ on Γ_-, and $k_0=0$ on Γ_0. The number of bound states changes by one as we cross each of these three curves; otherwise, we are in the generic case. Note that:

(i) If $b>\Gamma_+(c)$, then k_0, k_+, and k_- all lie on the positive imaginary axis, and hence we have three bound states.

(ii) If $\Gamma_0(c)<b\leqslant\Gamma_+(c)$, then k_0 and k_+ lie on the positive imaginary axis, but $k_-\notin \mathbf{C}^+$; hence we have two bound states.

(iii) If $\Gamma_-(c)<b\leqslant\Gamma_0(c)$, then there is exactly one bound state because k_+ lies on the positive imaginary axis but k_0 and k_- are not in $\mathbf{C}^+$.

(iv) There are no bound states when $b\leqslant\Gamma_-(c)$ because none of k_0, k_+, and k_- lie in $\mathbf{C}^+$. In this case, k_0 is always located on the imaginary axis; k_+ and k_- lie on the imaginary axis when $b\geqslant -(c+14+1/c)/16$ and they lie in $\mathbf{C}^-$ symmetrically located with respect to the imaginary axis.

This example can also be specialized to show the occurrence of a double zero of $1/T^+(k)$. Indeed, choose $b=-(c+14+1/c)/16$ and $c\in(-1,-5+2\sqrt{5})$. Then, from (11.9) we see that $k_+=k_-=-i\epsilon(c^2+10c+5)/[8(1+c)]$, and hence $T^+(k)$ has a double pole on the positive imaginary axis in $\mathbf{C}^+$ for any $c\in(-1,-5+2\sqrt{5})$. Note also that when $b=(1-c)/(4c)$ and $c\in(-1,-5+2\sqrt{5})$, although k_0 is located on the negative imaginary axis, k_+ and k_- are symmetrically located on the real axis; thus, in this case $T^+(k)$ has poles on the real axis. When $b=-(5+\sqrt{5})/10$ and $c=-5+2\sqrt{5}$, both k_+ and k_- vanish, and hence we get a simple pole for $T^+(k)$ at $k=0$; this illustrates the exceptional case when the denominator in (5.3) vanishes.

APPENDIX: SMALL-k ESTIMATES

In this Appendix, proceeding as in Refs. 11 and 12, we obtain various small-k estimates that are needed in the proof of Theorem 5.2.

In the exceptional case, let $\tilde{\psi}(k,x)$ be the solution of (1.1) satisfying the initial conditions

$$\tilde{\psi}(k,0) = f_l(0,0), \quad \tilde{\psi}'(k,0) = f_l'(0,0), \quad k\in\mathbf{R}. \tag{A1}$$

Note that $\tilde{\psi}(0,x)=f_l(0,x)$, and hence $\tilde{\psi}(0,x)$ is bounded in such a way that $\tilde{\psi}(0,+\infty)=1$ and $\tilde{\psi}(0,-\infty)=\gamma$, where γ is the constant defined in (2.9). We have

$$\tilde{\psi}(k,x) = f_l(0,0)\cos kx + f_l'(0,0)\,\frac{\sin kx}{k} + \frac{1}{k}\int_0^x dy\,\sin k(x-y)[ikP(y)+Q(y)]\tilde{\psi}(k,y). \tag{A2}$$

Let $\psi_1(k,x)$ denote the solution of (1.1) with $P(x)=0$ and satisfying (A1). We have

$$\psi_1(k,x) = f_l(0,0)\cos kx + f_l'(0,0)\,\frac{\sin kx}{k} + \frac{1}{k}\int_0^x dy\,\sin k(x-y)Q(y)\psi_1(k,y). \tag{A3}$$

Note that $\tilde{\psi}(0,x)=\psi_1(0,x)$.

Proposition A.1: Assume $Q\in L^1_1(\mathbf{R})$. For $x,k\in\mathbf{R}$, we have

$$|\psi_1(k,x)-\psi_1(0,x)|\leq C\left(\frac{|kx|}{1+|kx|}\right)^2, \quad |\psi_1(k,x)|\leq C(1+|k|). \tag{A4}$$

Proof: Note that $\psi_1(0,x)=f_l(0,x)$ and hence it is uniformly bounded for $x\in\mathbf{R}$. Furthermore,

$$\psi_1(0,x)=f_l(0,0)+xf_l'(0,0)+\int_0^x dy\ (x-y)Q(y)\psi_1(0,y). \tag{A5}$$

Subtracting (A5) from (A3) and iterating the resulting integral equation as in the proof of Proposition A.1 of Ref. 12 (using $H(x)\equiv H_+=1$ there) we obtain the first inequality in (A4). Using that inequality and the boundedness of $\psi_1(0,x)$, we obtain the second inequality in (A4). ■

Let us choose a second linearly independent solution of (1.1) with $P(x)=0$ such that the Wronskian $[\psi_1(k,x);\psi_2(k,x)]$ is equal to 1. For example, we can choose $\psi_2(k,x)$ so that it satisfies the initial conditions $\psi_2(k,0)=0$ and $\psi_2'(k,0)=1/f_l(0,0)$; note that there is no loss of generality in assuming $f_l(0,0)\neq 0$, since the case $f_l(0,0)=0$ can be handled by a shift of the origin. We have

$$\psi_2(k,x)=\frac{\sin kx}{kf_l(0,0)}+\frac{1}{k}\int_0^x dy\ \sin k(x-y)Q(y)\psi_2(k,y). \tag{A6}$$

Proposition A.2: Assume $Q\in L^1_1(\mathbf{R})$. Then, for $x,k\in\mathbf{R}$ we have

$$|\psi_2(k,x)|\leq\frac{C|x|}{1+|kx|}, \quad |\psi_2(k,x)-\psi_2(0,x)|\leq C|x|\left(\frac{|kx|}{1+|kx|}\right)^2. \tag{A7}$$

Proof: Iterating (A6) as in the proof of Proposition A.1 of Ref. 12, we obtain the first inequality in (A7). Note that from (A6) we have

$$\psi_2(0,x)=\frac{x}{f_l(0,0)}+\int_0^x dy\ (x-y)Q(y)\psi_2(0,y). \tag{A8}$$

Subtracting (A8) from (A6) and iterating the resulting integral equation, we obtain the second inequality in (A7). ■

Proposition A.3: Assume $P,Q\in L^1_1(\mathbf{R})$. Then, for $x\in\mathbf{R}$ and as $k\to 0$ in $\mathbf{R}$, we have

$$\begin{aligned}\tilde{\psi}(k,x)-\psi_1(k,x)=&-ik\psi_1(0,x)\int_0^x dz\ \psi_2(0,z)P(z)\psi_1(0,z)\\&+ik\psi_2(0,x)\int_0^x dz\ P(z)\psi_1(0,z)^2+\mathrm{O}\left(|kx|\left[\frac{|kx|}{1+|kx|}\right]^2\right).\end{aligned} \tag{A9}$$

Proof: Recall that $\psi_1(k,x)$ and $\psi_2(k,x)$ are two linearly independent solutions of (1.1) when $P(x)=0$. Using variation of parameters on (1.1), we obtain (assuming $x\geq 0$)

$$\begin{aligned}\tilde{\psi}(k,x)-\psi_1(k,x)=&-ik\psi_1(k,x)\int_0^x dz\ \psi_2(k,z)P(z)\tilde{\psi}(k,z)\\&+ik\psi_2(k,x)\int_0^x dz\ \psi_1(k,z)P(z)\tilde{\psi}(k,z).\end{aligned} \tag{A10}$$

Let us write (A10) as

$$\tilde{\psi}(k,x)-\psi_1(k,x)=A_1+A_2+A_3+A_4+A_5+A_6+A_7+B_1+B_2, \tag{A11}$$

where we have defined

$$B_1=-ik\psi_1(k,x)\int_0^x dz\ \psi_2(k,z)P(z)[\widetilde{\psi}(k,z)-\psi_1(k,z)],$$

$$B_2=ik\psi_2(k,x)\int_0^x dz\ \psi_1(k,z)P(z)[\widetilde{\psi}(k,z)-\psi_1(k,z)],$$

$$A_1=-ik\psi_1(0,x)\int_0^x dz\ \psi_2(0,z)P(z)\psi_1(0,z),$$

$$A_2=-ik[\psi_1(k,x)-\psi_1(0,x)]\int_0^x dz\ \psi_2(k,z)P(z)\psi_1(k,z),$$

$$A_3=-ik\psi_1(0,x)\int_0^x dz\ [\psi_2(k,z)-\psi_2(0,z)]P(z)\psi_1(k,z),$$

$$A_4=-ik\psi_1(0,x)\int_0^x dz\ \psi_2(0,z)P(z)[\psi_1(k,z)-\psi_1(0,z)],$$

$$A_5=ik\psi_2(0,x)\int_0^x dz\ P(z)\psi_1(0,z)^2,$$

$$A_6=ik[\psi_2(k,x)-\psi_2(0,x)]\int_0^x dz\ P(z)\psi_1(k,z)^2,$$

$$A_7=ik\psi_2(0,x)\int_0^x dz\ [\psi_1(k,z)-\psi_1(0,z)][\psi_1(k,z)+\psi_1(0,z)]P(z).$$

Using the estimates in (A4) and (A7), we obtain

$$|A_2|\leq C|k|(1+|k|)\left(\frac{|kx|}{1+|kx|}\right)^2\int_0^x dz\ z|P(z)|, \tag{A12}$$

$$|A_3|\leq C|k|(1+|k|)\left(\frac{|kx|}{1+|kx|}\right)^2\int_0^x dz\ z|P(z)|, \tag{A13}$$

$$|A_4|\leq C|k|\left(\frac{|kx|}{1+|kx|}\right)^2\int_0^x dz\ z|P(z)|, \tag{A14}$$

$$|A_6|\leq C|kx|(1+|k|)^2\left(\frac{|kx|}{1+|kx|}\right)^2\int_0^x dz\ |P(z)|, \tag{A15}$$

$$|A_7|\leq C|kx|(1+|k|)\left(\frac{|kx|}{1+|kx|}\right)^2\int_0^x dz\ |P(z)|. \tag{A16}$$

For $x<0$ analogous estimates hold. Iterating the integral equation for $\widetilde{\psi}(k,x)-\psi_1(k,x)$ given in (A11) and using (A12)–(A16), we obtain (A9). ■

In order to estimate the small-k asymptotics of $T^+(k)$, we will use (4.2). Note that as in (A24) of Ref. 12 we have

$$f_l(0,0)[f_l^+(k,x);f_r^+(k,x)]=f_r^+(k,0)\left[-ikf_l(0,0)+f_l'(0,0)+\int_0^\infty dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)\right]$$
$$-f_l^+(k,0)\left[ikf_l(0,0)+f_l'(0,0)-\int_{-\infty}^0 dz\ e^{-ikz}[ikP(z)\right.$$
$$\left.+Q(z)]\widetilde{\psi}(k,z)\right]. \tag{A17}$$

Proposition A.4: Assume $P,Q\in L_1^1(\mathbf{R})$. Then, as $k\to 0$ we have

$$\int_0^\infty dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)=-f_l'(0,0)+ikf_l(0,0)-ik+ik\int_0^\infty dz\ P(z)f_l(0,z)^2+o(|k|), \tag{A18}$$

$$\int_{-\infty}^0 dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)=f_l'(0,0)-ik\gamma+ikf_l(0,0)$$
$$+(ik/\gamma)\int_{-\infty}^0 dz\ P(z)f_l(0,z)^2+o(|k|), \tag{A19}$$

where γ is the constant defined in (2.9).

Proof: Let us write

$$\int_0^\infty dz\ e^{ikz}[ikP(z)+Q(z)]\widetilde{\psi}(k,z)=J_1+J_2+J_3+J_4+J_5+J_6, \tag{A20}$$

where

$$J_1=\int_0^\infty dz\ Q(z)\widetilde{\psi}(0,z),\quad J_2=\int_0^\infty dz\ Q(z)[e^{ikz}-1]\widetilde{\psi}(0,z),$$
$$J_3=ik\int_0^\infty dz\ e^{ikz}P(z)\widetilde{\psi}(0,z),$$
$$J_4=\int_0^\infty dz\ e^{ikz}Q(z)[\psi_1(k,z)-\widetilde{\psi}(0,z)],$$
$$J_5=ik\int_0^\infty dz\ e^{ikz}P(z)[\widetilde{\psi}(k,z)-\widetilde{\psi}(0,z)],$$
$$J_6=\int_0^\infty dz\ e^{ikz}Q(z)[\widetilde{\psi}(k,z)-\psi_1(k,z)].$$

As in (A25) and (A26) of Ref. 12 we have $J_1=-f_l'(0,0)$ and

$$J_2=ik[f_l(0,0)-1]+o(|k|),\quad k\to 0.$$

As $k\to 0$, using (A4) we obtain $J_4=o(|k|)$ and

$$J_3=ik\int_0^\infty dz\ P(z)\widetilde{\psi}(0,z)+o(|k|),$$

and using (A9) we have $J_5=o(|k|)$ and

$$J_6=ik\int_0^\infty dz\ Q(z)\left[-\psi_1(0,z)\int_0^z dt\ \psi_2(0,t)P(t)\psi_1(0,t)+\psi_2(0,z)\int_0^z dt\ P(t)\psi_1(0,t)^2\right]$$
$$+o(|k|). \tag{A21}$$

Notice that $[\psi_1(k,x);\psi_2(k,x)]=1$ and $Q(z)\psi_s(0,z)=\psi_s''(0,z)$ for $s=1,2$. Therefore, using $\psi_2'(0,+\infty)=1$, $\psi_1'(0,z)=o(z^{-1})$ as $z\to+\infty$, $\psi_1(0,z)=\widetilde{\psi}(0,z)$, and integration by parts twice in (A21), we obtain

$$J_6 = -ik \int_0^\infty dz\ P(z)\overline{\psi}(0,z) + ik \int_0^\infty dz\ P(z)\overline{\psi}(0,z)^2 + o(|k|), \quad k \to 0.$$

Thus, from (A20) we obtain (A18). Similarly, using $\psi_1'(0,-\infty)=0$ and $\psi_1(0,-\infty)=\gamma$, we get (A19).

ACKNOWLEDGMENTS

The research leading to this article was supported in part by the National Science Foundation under Grant No. DMS-9501053. One of the authors (CvdM) is indebted to the Mathematics Departments at Virginia Tech and NDSU for their hospitality and support and to CNR-GNFM and MURST for support.

[1] M. Jaulent and C. Jean, ''The inverse *s*-wave scattering problem for a class of potentials depending on energy,'' Commun. Math. Phys. **28**, 177–220 (1972).
[2] M. Jaulent and C. Jean, ''The inverse problem for the one-dimensional Schrödinger equation with an energy-dependent potential. I,'' Ann. Inst. Henri Poincaré, Sect. A **25**, 105–118 (1976).
[3] M. Jaulent and C. Jean, ''The inverse problem for the one-dimensional Schrödinger equation with an energy-dependent potential. II,'' Ann. Inst. Henri Poincaré, Sect. A **25**, 119–137 (1976).
[4] M. Jaulent, ''Inverse scattering problems in absorbing media,'' J. Math. Phys. **17**, 1351–1360 (1976).
[5] D. H. Sattinger and J. Szmigielski, ''Energy dependent scattering theory,'' Differ. Integral Eqs. **8**, 945–959 (1995).
[6] D. J. Kaup, ''A higher-order water-wave equation and the method for solving it,'' Prog. Theor. Phys. **54**, 396–408 (1975).
[7] M. Tsutsumi, ''On the inverse scattering problem for the one-dimensional Schrödinger equation with an energy dependent potential,'' J. Math. Anal. Appl. **83**, 316–350 (1981).
[8] D. H. Sattinger and J. Szmigielski, ''A Riemann-Hilbert problem for an energy dependent Schrödinger operator,'' Inverse Probl. **12**, 1003–1025 (1996).
[9] L. D. Faddeev, ''Properties of the *S*-matrix of the one-dimensional Schrödinger equation,'' Am. Math. Soc. Trans. **2**, 139–166 (1964) [Trudy Mat. Inst. Steklova **73**, 314–336 (1964) (Russian)].
[10] P. Deift and E. Trubowitz, ''Inverse scattering on the line,'' Commun. Pure Appl. Math. **32**, 121–251 (1979).
[11] M. Klaus, ''Low-energy behaviour of the scattering matrix for the Schrödinger equation on the line,'' Inverse Probl. **4**, 505–512 (1988).
[12] T. Aktosun, M. Klaus, and C. van der Mee, ''Inverse wave scattering with discontinuous wave speed,'' J. Math. Phys. **36**, 2880–2928 (1995).
[13] V. N. Pivovarchik, ''On the discrete spectrum of a boundary value problem,'' Sib. Math. J. **31**, 853–856 (1990) [Sibirsk. Mat. Zh. **31**, 182–186 (1990) (Russian)].
[14] V. N. Pivovarchik, ''Eigenvalues of a certain quadratic pencil of operators,'' Funct. Anal. Appl. **23**, 70–72 (1989) [Funkt. Anal. i ego Priloz. **23**, 80–81 (1989) (Russian)].
[15] V. N. Pivovarchik, ''On the total algebraic multiplicity of the spectrum in the right-half plane for a quadratic operator pencil,'' St. Petersburg Math. J. **3**, 447–454 (1992) [Algebra i Analiz **3**, 223–230 (1991) (Russian)].
[16] V. N. Pivovarchik, ''On positive spectra of one class of polynomial operator pencils,'' Integr. Eqs. Oper. Th. **19**, 314–326 (1994).
[17] R. P. Boas, Entire Functions (Academic, New York, 1954).
[18] D. B. Hinton, M. Klaus, and J. K. Shaw, ''Levinson's theorem and Titchmarsh-Weyl $m(\lambda)$ theory for Dirac systems,'' Proc. R. Soc. Edinburgh Sect. A **109**, 173–186 (1988).
[19] I. C. Gohberg and E. I. Sigal, ''An operator generalization of the logarithmic residue theorem and the theorem of Rouché,'' Math. USSR Sbornik **13**, 603–625 (1971) [Mat. Sbornik **84**, 607–629 (1971) (Russian)].
[20] E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations (McGraw-Hill, New York, 1955).
[21] N. Levinson, ''On the uniqueness of the potential in a Schrödinger equation for a given asymptotic phase,'' Danske Vid. Selsk. Mat.-Fys. Medd. **25**, 1–29 (1949).
[22] J. Weidmann, Spectral Theory of Ordinary Differential Operators, Lecture Notes in Mathematics, 1258 (Springer, Berlin, 1987).
[23] M. Reed and B. Simon, Methods of Modern Mathematical Physics. IV. Analysis of Operators (Academic, New York, 1978).
[24] M. Reed and B. Simon, Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-adjointness (Academic, New York, 1975).
[25] T. Kato, Perturbation Theory for Linear Operators, 2nd ed. (Springer, Berlin, 1976).
[26] C. M. Bender and S. A. Orszag, Advanced Mathematical Methods for Scientists and Engineers (McGraw-Hill, New York, 1978).

Geodesic flow on an hyperboloid model and the two-body motion under repulsive Coulomb forces

Alejandro M. Mesón[a] and Fernando Vericat[b]
Instituto de Física de Líquidos y Sistemas Biológicos (IFLYSIB)-UNLP-CONICET, cc. 565-(1900) La Plata, Argentina and Grupo de Aplicaciones Matemáticas y Estadísticas de la Facultad de Ingeniería, (GAMEFI), Departamento de Fisicomatemática, Facultad de Ingeniería, Universidad Nacional de La Plata, Argentina

(Received 19 November 1997; accepted for publication 12 January 1998)

We show that (in n dimension) the flow defined by the Hamiltonian system for two charged particles of the same sign, is mapped into the geodesic flow over the non-Euclidean space: $\mathscr{H}^{n+} \sqcup_{\partial} \mathscr{H}^{n-} - \{P\}$, say the hyperboloid with the axis along $\hat{e}_0$ which besides having its ''boundaries at infinite'' identified, it is also punctured at the point $P \equiv (\xi_0, \xi_1, \ldots, \xi_n) = (-1,0,...,0)$ (corresponding to the collision states) and whose metric is of negative curvature. © *1998 American Institute of Physics.* [S0022-2488(98)02104-5]

I. INTRODUCTION

The problem of the movement of two interacting particles (the two-body problem) has drawn for centuries the attention of physicists and mathematicians and its general solution is well known.[1] This is particularly true for the so-called Kepler's problem that considers the classical dynamics of two massive particles which interact via the attractive force of Newton or two particles with charges of the opposite sign so that between them an attractive electrostatic Coulomb force acts. The problem is best described in the reduced form in which the center of mass is at rest, and we have to study the motion of a single mass point only. In n dimensions the position and the momentum of the mass point are given by points of $\mathscr{R}^n$: $q=(q_1,q_2,\ldots,q_n)$ and $p=(p_1,p_2,\ldots,p_n)$, respectively. Defining the modulus $|k|^2=(k/k)=\Sigma_{i=1}^n k_i^2$ and after an appropriate choice of the mass (and charge) unit, we can write Kepler's Hamiltonian system as $\dot{q}=\partial H_K/\partial p$, $\dot{p}=-\partial H_K/\partial q$, where $H_K=\frac{1}{2}|p|^2-1/|q|$ is the corresponding Hamilton function.

An interesting property of Kepler's problem was demonstrated some time ago by Moser. He showed the ''equivalence'' between the motion of a particle in an attractive Newtonian (or Coulombian) field of forces and that of a particle constrained to freely move on a sphere surface. Specifically he proved that the surfaces of constant negative energy for Kepler's problem can be mapped one to one into the unit tangent bundle of the n-dimensional sphere $\mathscr{S}^n$. Then the flow associated to H_K is mapped into the geodesic flow on $\mathscr{S}^n - \{P\}$, the sphere punctured at one point (the ''north pole'' corresponding to the collision states).[2] With relation to the equivalence between Kepler's problem and the geodesic flow in spaces of constant curvature see also Refs. 3 and 4. In particular, the case of constant positive energies is treated by Belbruno in Ref. 3.

Here we report the corresponding property for a Hamiltonian system that describes the motion of two particles with equal sign charges or, in reduced form, a point charge interacting with a repulsive central field of Coulomb forces.

II. GEODESIC FLOW ON AN HYPERBOLOID AND THE MOTION OF TWO CHARGES OF THE SAME SIGN

In the units we are considering, the Hamitonian system for the movement of a charged particle in a repulsive Coulomb field reads $\dot{q}=\partial H_C/\partial p$, $\dot{p}=-\partial H_C/\partial q$, where the dot denotes the derivative with respect to the time and

[a]Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.
[b]Electronic mail: vericat@iflysibl.unlp.edu.ar

 © 1998 American Institute of Physics

$$H_C=\frac{1}{2}|p|^2+\frac{1}{|q|} \tag{1}$$

is the Hamilton function.

The property in which we are interested here is the ''equivalence'' between the flow described by the Hamilton function (1) and the motion of a particle constrained to freely move on the surface of the hyperboloid whose axis is along the unit vector $\hat{e}_0$, whose ''boundaries at infinite'' are identified and that is punctured at the point $P\equiv(\xi_0,\xi_1,\ldots,\xi_n)=(-1,0,\ldots,0)$, the metric being of negative curvature. Denoting this hyperboloid $\mathscr{H}^{n\pm}_{-P}\equiv\mathscr{H}^{n+}\sqcup_{\partial}\mathscr{H}^{n-}-\{P\}$, the previous statement can be put into more precise terms:

Theorem: For a positive constant h, the energy surface $H_C=h$ can be mapped one to one into the unit tangent bundle of $\mathscr{H}^{n\pm}_{-P}$ so that, after a change of variables, the flow defined by H_C is mapped into the geodesic flow on $\mathscr{H}^{n\pm}_{-P}$.

The demonstration closely follows that of Moser for the Kepler's problem. We begin describing the geodesic flow on $\mathscr{H}^{n\pm}_{-P}$ and then, by introducing suitable mappings, we transform it into the two-body problem for charges of the same sign.

If $\xi=(\xi_0,\xi_1,\ldots,\xi_n)\in\mathscr{R}^{n+1}$ and define the inner product $\langle\xi^{(a)},\xi^{(b)}\rangle=\sum_{i=1}^n\xi_i^{(a)}\xi_i^{(b)}-\xi_0^{(a)}\xi_0^{(b)}$ then the hyperbolic spaces $\mathscr{H}^{n+}$ and $\mathscr{H}^{n-}$ are given by $\mathscr{H}^{n+}=\{\xi:\langle\xi,\xi\rangle=-1;\xi_0>0\}$ and $\mathscr{H}^{n-}=\{\xi:\langle\xi,\xi\rangle=-1;\xi_0<0\}$, respectively, with the metric given by $d(\xi^{(a)},\xi^{(b)})=\inf_\gamma\{L(\gamma)\}$. Here $L(\gamma)=\int\langle\gamma',\gamma'\rangle ds$, the inf is taken over all the curves $\gamma(s)$ joining the points $\xi^{(a)}$ and $\xi_{(b)}$ and $\gamma'\equiv d\gamma/ds$. The union of both spaces will be simply denoted by $\mathscr{H}^n$. Considering $\xi=\xi(s)$, then $\eta=\xi'=d\xi/ds$ so that $(\xi,\eta)\in T(\mathscr{H}^n)$, the tangent bundle of $\mathscr{H}^n$. With this notation, the geodesic flow can be described by the variational problem

$$\delta\int\langle\xi',\xi'\rangle ds=0 \tag{2}$$

with the condition

$$\langle\xi,\xi\rangle=-1. \tag{3}$$

This is equivalent to consider the Hamiltonian system

$$\frac{\partial\Phi(\xi,\eta)}{\partial\eta}=\xi'\Omega;\quad\frac{\partial\Phi(\xi,\eta)}{\partial\xi}=-\eta'\Omega, \tag{4}$$

where the form is defined by the $(n+1)\times(n+1)$ matrix Ω whose entries are zero except the diagonal elements: $\Omega_{00}=-1$; $\Omega_{ii}=1$ $(i=1,2,\ldots,n)$ and

$$\Phi(\xi,\eta)=-\tfrac{1}{2}\langle\xi,\xi\rangle\langle\eta,\eta\rangle \tag{5}$$

is its Hamilton function.

To describe this flow in $\mathscr{R}^n$, we use the symplectic transformations $(i=1,2,\ldots,n)$:

$$x_i=\frac{\xi_i}{1+\xi_0} \tag{6}$$

and

$$y_i=\eta_i(1+\xi_0)-\xi_i\eta_0. \tag{7}$$

In the Euclidean space we define the scalar product $(x^{(a)}/x^{(b)})=\sum_{i=1}^n x_i^{(a)}x_i^{(b)}$ so that $|x|^2=(x/x)=(\xi_0-1)/(\xi_0+1)$ and the inverse relations of Eqs. (6) and (7) read $\xi_i=2x_i/(1-|x|^2)$; $\eta_i=x_i(x/y)+y_i(1-|x|^2)/2$ for $i=1,2,\ldots,n$ and $\xi_0=(1+|x|^2)/(1-|x|^2)$; $\eta_0=(x/y)$. We observe that Eq. (6) transforms $\mathscr{H}^{n+}$ into the disc $\mathscr{D}^n=\{x:|x|\leqslant 1\}$ and $\mathscr{H}^{n-}$ into the set $\{x:|x|\geqslant 1\}$. In particular the point $(1,0,0,\ldots,0)$ of $\mathscr{H}^{n+}$ gives the center of $\mathscr{D}^n$ and its ideal border, say $\partial^+=\partial(\mathscr{H}^{n+})=\{\xi:\xi\in\mathscr{H}^{n+};\xi_0=\infty\}$, gives the border of $\mathscr{D}^n$: $\partial(\mathscr{D}^n)=\{x:|x|=1\}$. On the other

hand, the ideal border of $\mathscr{H}^{n-}$, $\partial^- = \partial(\mathscr{H}^{n-}) = \{\xi : \xi \in \mathscr{H}^{n-}; \xi_0 = -\infty\}$, yields $\partial(\mathscr{D}^n)$ too, whereas the ''image'' of the point $P \equiv (-1,0,0,...,0)$ is interpreted as the compactification point of $\mathscr{R}^n$ and we will, in principle, omit it. Thus, to have a one to one correspondence we identify ∂^+ and ∂^- and call both of them simply ∂. In this way, we are considering a one to one mapping between the punctured hyperboloid $\mathscr{H}^{n+} \sqcup_\partial \mathscr{H}^{n-} - \{P\} \subset \mathscr{R}^{n+1}$ and $\mathscr{R}^n$.

With the previous transformations we have $\langle \eta, \eta \rangle = [(1-|x|^2)/2]^2 |y|^2$ and Eq. (5) yields, using Eq. (3),

$$\Phi(\xi, \eta) = F(x,y) = \tfrac{1}{8}(1-|x|^2)^2 |y|^2. \tag{8}$$

From this Hamilton function we have the transformed motion equations

$$x' = \frac{\partial F}{\partial y}; \quad y' = -\frac{\partial F}{\partial x}. \tag{9}$$

Here and in any case in which Euclidean space is involved, the canonical symplectic form is considered.

To prove the mapping of the Hamiltonian system Eqs. (8) and (9) into repulsive two-body Coulomb problem on positive energy surfaces $H_C = h$, it is enough to consider the particular case $h = \frac{1}{2}$ since the general case of an arbitrary positive energy surface reduces to that case by a simple scaling of the variables.

Defining the function $G(x,y) = \sqrt{2F(x,y)} - 1$, we see that the system given by Eq. (9) on the surface $F(x,y) = \frac{1}{2}$ goes over into the Hamiltonian system

$$x' = \frac{\partial G}{\partial y}; \quad y' = -\frac{\partial G}{\partial x} \tag{10}$$

on the surface $G = 0$. Changing the independent variable s to $t = \int |y(s)| ds$, the system (10) yields, always on $G = 0$:

$$\dot{x} = |y|^{-1} \frac{\partial G}{\partial y}; \quad \dot{y} = -|y|^{-1} \frac{\partial G}{\partial x}, \tag{11}$$

where $\dot{z} \equiv dz/dt$.

Therefore, on $H = \frac{1}{2}$ we have the Hamiltonian system

$$\dot{x} = \frac{\partial H}{\partial y}; \quad \dot{y} = -\frac{\partial H}{\partial x} \tag{12}$$

with Hamilton function given by

$$H = \frac{1}{2} - |y|^{-1} G = \frac{1}{2} - |y|^{-1} (\sqrt{2F(x,y)} - 1) = \frac{1}{2}|x|^2 + \frac{1}{|y|}. \tag{13}$$

So that, doing $p = -x$ and $q = y$, the Hamiltonian system

$$\dot{q} = \frac{\partial H_C}{\partial p}; \quad \dot{p} = -\frac{\partial H_C}{\partial q} \tag{14}$$

for the function of Hamilton (1) on the energy surface $H_C = \frac{1}{2}$ is obtained. The generalization to arbitrary positive energy surfaces $H_C = \frac{1}{2}\alpha^2$ reduces to the previous particular case by simply scaling the variables: $q \to \alpha^2 q$; $.p \to p/\alpha$ and $t \to \alpha^3 t$.

Till now, we have omitted the point $P \equiv (-1,0,0,...,0)$ of the hyperboloid and the energy surface can be compactified by restoring it. In this way the regularization of the flow is obtained, the geodesics through the point P corresponding to the collision states $|p| = \infty$; $q = 0$.

Using the transformation given by Eq. (7), we can easily obtain the center of mass orbits for two charged particles of the same sign corresponding to the surface energy $H_C=\frac{1}{2}$. Since the orbits are planar, we consider $n=2$. Then geodesics in the hyperboloid $\mathscr{H}^{2+}\sqcup_{\partial}\mathscr{H}^{2-}\subset\mathscr{R}^3$ are given by the intersection of

$$-\xi_0^2+\xi_1^2+\xi_2^2=-1 \tag{15}$$

with the plane in which the geodesic lies, say β.

Choosing the coordinate system $(\hat{e}_0,\hat{e}_1,\hat{e}_2)$ in such a way that $\hat{e}_0$ is along the hyperboloid axis and the geodesic plane β contains $\hat{e}_2$, then

$$\xi_1=\xi_0\tan(\theta), \tag{16}$$

where θ is the dihedral angle between β and the plane defined by $\hat{e}_1$ and $\hat{e}_2$. Moreover, ξ_0, ξ_1 and ξ_2 can be parameterized in the form:

$$\xi_0=\frac{1}{\sqrt{1-\tan^2(\theta)}}\cosh(z);\quad \xi_1=\sinh(z);\quad \xi_2=\frac{\tan(\theta)}{\sqrt{1-\tan^2(\theta)}}\cosh(z) \tag{17}$$

and since, from Eq. (7),

$$y_i=\xi_i'(1+\xi_0)-\xi_i\xi_0';\quad i=1,2, \tag{18}$$

then the orbit for a particle in a repulsive Coulomb central field is obtained in the well known form:[1]

$$q_1=y_1=\sqrt{e^2-1}\,\sinh(z), \tag{19}$$

$$q_2=y_2=\cosh(z)+e, \tag{20}$$

where $e=1/\sqrt{1-\tan^2(\theta)}$ is the eccentricity of the hyperbola on the plane.

It has therefore been shown that the Eqs. (6), (7) and $t=\int|y(s)|ds$, map the unit tangent bundle $T(\mathscr{H}^n)=\{(\xi,\eta):\xi\in\mathscr{H}^n;\eta=d\xi(s)/ds\}$ into the $2n$-dimensional phase space $\mathscr{R}^{2n}=\{(q,p):q\in\mathscr{R}^n;p=dq(t)/dt\}$ and that the geodesic orbits on the hyperboloid map into the hyperbolas for a charged particle in a repulsive Coulomb field, as the theorem states.

ACKNOWLEDGMENTS

We wish to thank Professor J. E. Solomín for helpful comments on symplectic forms. Support of this work by UNLP, CONICET and FONCYT (PICT0311) of Argentina is greatly appreciated. F.V. is a member of CONICET.

[1] L. D. Landau and E. M. Lifshitz, *Mecánica* (Editorial Reverté, S.A, Barcelona, 1965).
[2] J. Moser, Commun. Pure Appl. Math. **23**, 609 (1970).
[3] E. A. Belbruno, Celest. Mech. **15**, 467 (1977).
[4] Y. S. Osipov, Celest. Mech. **16**, 191 (1977).

Diagrams in classical and semiclassical perturbation theory

Panayotis Panayotaros
Department of Physics, The University of Texas at Austin, Austin, Texas 78712-1081

(Received 21 May 1996; accepted for publication 24 November 1997)

We describe a diagrammatic method for the Poincare–Birkhoff normal forms algorithm of classical mechanics, and indicate the use of the diagrams with an example from hydrodynamics. We also present a generalization of the diagrammatic method to quantum mechanics. The quantum diagrams can be used in the semiclassical version of normal forms developed by Graffi, Paul, and others. © *1998 American Institute of Physics.* [S0022-2488(98)02503-1]

I. INTRODUCTION

The dynamics of a Hamiltonian system near an elliptic fixed point is one of the basic problems of classical physics, arising in the stability of motion in mechanics, in weakly nonlinear waves in fluids, plasmas, and other media, and in many other situations. A powerful tool to analyze such dynamical systems is the Poincare–Birkhoff normal forms method of eliminating nonresonant terms in the Hamiltonian via canonical transformations (see, e.g., Ref. 1). Normal forms also appear in generalizations of the Bohr–Sommerfeld quantization condition to near-integrable systems. A recent approach to this question, developed by in Refs. 2–4 and others, led to a theory of semi-classical Poincare–Birkhoff normal forms that can be applied to many problems of quantum perturbation theory.

In most recent applications of classical normal forms, canonical transformations are constructed by the so-called Lie series method (see, e.g., Refs. 5–7). An analog of the Lie series is also used in semiclassical normal forms to construct unitary transformations (see, e.g., Refs. 3 and 8). A computational issue in using Lie series and its semiclassical analog is the efficient evaluation of repeated Poisson brackets (see, e.g., Ref. 9). In this note we point out that these operations have an elementary diagrammatic interpretation and that using the diagrams speeds up the calculations significantly. We indicate the use of the diagrams with an example of a normal form for an infinite-dimensional Hamiltonian system arising in hydrodynamics. We also show that diagrams and the operations between them have a natural ''quantization,'' and that in the $\hbar \to 0$ limit we obtain the classical diagrams. Quantum diagrams can be applied to semi-classical normal forms calculations, very much along the lines of the classical case.

II. CLASSICAL DIAGRAMS

The classical diagrammatic method consists of a correspondence between diagrams and functions on the phase space M, and a rule for assigning a new graph to the Poisson bracket of two graphs.

The classical phase space M will be $\mathbf{R}^{2d}$ with the canonical symplectic structure and corresponding Poisson bracket. Given a harmonic oscillator Hamiltonian with frequencies $\omega = [\omega_1,\ldots,\omega_d]$ we define the complex variables a_i, a_i^* by

$$a_i = \frac{1}{\sqrt{2}}\left(\sqrt{\omega_i}q_i + i\,\frac{p_i}{\sqrt{\omega_i}}\right), \quad a_i^* = \frac{1}{\sqrt{2}}\left(n\sqrt{\omega_i}q_i - i\,\frac{p_i}{\sqrt{\omega_i}}\right). \tag{2.1}$$

The map $a(q,p)$ embeds M canonically to $\mathbf{C}^{2d}$ with the Poisson bracket J given by

$$J(f,g) = \{f,g\} = i\sum_{i=1}^{d}\left(\frac{\partial f}{\partial a_i}\frac{\partial g}{\partial a_i^*} - \frac{\partial g}{\partial a_i}\frac{\partial f}{\partial a_i^*}\right). \tag{2.2}$$

 © 1998 American Institute of Physics

We may also consider infinite-dimensional analogs of the above phase space, e.g., with the index i running over $\mathbf{Z}^n$. What follows applies to the infinite-dimensional case as well. Observables will be assumed to be real analytic (i.e., polynomials in a_i, a_j^* for the purposes of formal calculations), vanishing at the origin. We now describe the diagrammatic rules.

(1) The set of graphs will be the set D of simply connected directed trees with indices at each leg (branch). (Legs will meet at vertices.) Legs connecting two vertices will be referred to as internal legs, while legs with a free end will be referred to as external legs. The one-vertex trees obtained by cutting the internal legs of a multi-vertex tree t will be called the one-vertex subtrees of t. External legs will carry one index. Internal legs will carry two indices, one for each vertex being connected, so that the one-vertex subtrees will be in D.

(2) To describe the correspondence between trees and polynomials, we start with one-vertex trees. We assign to each homogeneous polynomial term

$$\sum_{\kappa_1,\ldots,\kappa_n,\lambda_1,\ldots,\lambda_m} I_{\kappa_1\ldots\kappa_n\lambda_1\ldots\lambda_m} a_{\kappa_1}\cdots a_{\kappa_r} a^*_{\lambda_1}\cdots a^*_{\lambda_m}$$

a vertex with $n+m$ legs (branches), one leg corresponding to each of the a_{κ_i} and $a^*_{\lambda_j}$. On each leg there will be an arrow, pointing into the vertex for legs representing the $a^*_{\lambda_j}$ and outwards from the vertex for legs representing the a_{κ_i}. Also, each leg will carry the index of the a_{κ_i} or $a^*_{\lambda_j}$ it represents. The dots in the vertices represent the coefficients $\omega_{\kappa\lambda}$, $A_{\kappa\lambda\mu}$, etc., and the summation over the indices. The dots in the vertices should be represented by different symbols, e.g., boxes, triangles letters, so that we can distinguish between graphs with the same number of in-going and out-going legs but different coefficients.

(3) A multi-vertex tree represents a homogeneous polynomial in the variables indexed by the external legs of the tree. The out-going and in-going arrows represent the a_{κ_i} and $a^*_{\lambda_j}$, respectively. The coefficient of this polynomial will be the product of the coefficients of all the one-vertex subtrees of the multi-vertex tree, summed over all the indices appearing in external legs and contracted over the pairs of indices representing legs that connect vertices. Also, the polynomial represented by multi-vertex tree may be alternatively represented by a one-vertex tree with an appropriate coefficient.

Proposition 2.1: With the above conventions, the Poisson bracket $\{f,g\}$ of two trees $f,g \in D$ is the sum of the graphs obtained by joining an out-going arrow of f to an in-going arrow of g and the trees obtained by joining an in-going arrow of f to an out-going arrow of g. The trees obtained this way will also have signs: $(-i)$ for trees obtained by joining an out-going arrow of f to an in-going arrow of g, and $(+i)$ for trees formed by joining an in-going arrow of f to an out-going arrow of g.

Proof: Let

$$F=\sum_{c_1,\ldots,c_n,c_{n+1},\ldots,c_{n+m}} I_{c_1\ldots c_n c_{n+1}\ldots c_{n+m}} a_{c_1}\cdots a_{c_n} a^*_{c_{n+1}}\cdots a^*_{c_{n+m}}$$

and F' same as F with the I, n, and m primed. Using the bilinearity and the derivation property of the bracket we have

$$\{F,F'\}=\sum_{i,j}\sum_{\substack{c_1,\ldots,c_n,c_{n+1},\cdots,c_{n+m}\\ c_{1'},\ldots,c_{n'},c_{n'+1},\ldots,c_{n'+m'}}} I_{c_1\ldots c_n c_{n+1}\ldots c_{n+m}} I'_{c_{1'}\ldots c'_n c_{n'+1}\ldots c_{n'+m'}} \times\{f_{c_i},f_{c_j}\}\prod_{\alpha=1,\alpha\neq i}^{n+m} f_{c_\alpha}\prod_{\beta=1,\beta\neq j}^{n'+m'} f_{c_\beta}$$

with $f_{c_\alpha}=a_{c_\alpha}$ if $\alpha<n$, $f_{c_\alpha}=a^*_{c_\alpha}$ otherwise, and similarly for f_{c_β}. From $\{a_\kappa,a^*_\lambda\}=-i\delta_{\kappa\lambda}$, $\{a_\kappa,a_\lambda\}=\{a^*_\kappa,a^*_\lambda\}=0$, and the correspondence between trees and polynomials we obtain the rule. ■

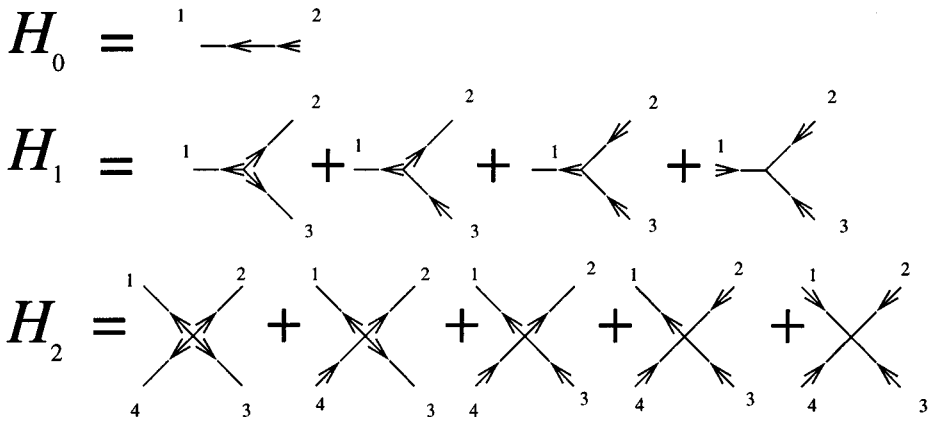

FIG. 1. The quartic water wave Hamiltonian.

The rules are essentially the same if we use the variables (q_γ, p_γ) or the well-known action-angle variables. Also, the diagrammatic notation can be compressed by considering diagrams without arrows or indices. One can calculate Poisson brackets with such trees and add arrows and indices afterwards.

III. AN APPLICATION TO FLUID MECHANICS

We indicate the use of diagrams with a normal form calculation for a Hamiltonian model of waves on the surface of a fluid layer surrounding a gravitating sphere. A detailed description of the model can be found in Ref. 10.

The canonical variables are the wave amplitude q and the surface hydrodynamic potential p, both functions on the two-sphere. The Hamiltonian is the total energy of the system. We denote the ith spherical harmonic coefficient of q,p by q_i, p_i, respectively. The index i runs over the pairs (l,m) with l a positive integer, $m=-l,-l+1,...,l$. Using the frequencies of the linearized problem we also define the complex variables a_i and a_i^* as in (2.1). The Poisson bracket is (after a change of coordinates) as in (2.2). The Hamiltonian H can be written as $H=H_0+\epsilon H_1+\epsilon^2 H_2+\cdots$, with H_m homogeneous polynomials of degree $m+2$ in a_i, a_i^* and ϵ a small dimensionless parameter. Physically, ϵ is the ratio of the typical wave amplitude over the thickness of the fluid layer. The other dimensionless parameter of the problem, the ratio β of depth over radius, is set here to $O(1)$. Graphically H is as in Fig. 1.

The normal form calculation consists of eliminating the cubic and part of the quartic terms of H by canonical transformations. It is convenient to use the multi-index notation $z^k \bar{z}^{\bar{k}} a_1^{k_1}\cdots a_d^{k_d}\cdots (a_1^*)^{\bar{k}_1}\cdots (a_d^*)^{\bar{k}_d}\cdots$, with $|k|=k_1+k_2+\cdots$, $|\bar{k}|=\bar{k}_1+\bar{k}_2+\cdots$, and write $H_m=\Sigma_{|\kappa|+|\bar{\kappa}|=m+2} A^m_{|\kappa|,|\bar{\kappa}|} z^k \bar{z}^{\bar{k}}$. Following the Lie series method (in the version of Ref. 6), we seek a function S_1 for which $\exp \epsilon \mathrm{Ad}_{S_1} H$ has no cubic terms. We are led to the ''homological equation'' $\{S_1,H_0\}+H_1=0$, which is solved by

$$S_1=\sum_{\kappa+\bar{\kappa}=3} \frac{A^1_{\kappa,\bar{\kappa}}}{\Sigma_i \omega_i (k_i-\bar{k}_i)} z^k \bar{z}^{\bar{k}},$$

i.e., the resonance condition is $\Sigma_i \omega_i (k_i-\bar{k}_i)=0$. In Ref. 10 we have shown that when the resonance condition is satisfied for some $k,\bar{k}$, the corresponding $A^1_{\kappa,\bar{\kappa}}$ must vanish, and thus ϵH_1 is formally eliminated. Next we eliminate the nonresonant terms of $\tilde{H}_2$, the quartic terms of $\exp \epsilon \mathrm{Ad}_{S_1} H$. We have that $\tilde{H}_2=H_2+\frac{1}{2}\{S_1,H_1\}$. Writing S_1 as in Fig. 2, we obtain $\tilde{H}_2$ graphically. We have shown that the resonant part of $\tilde{H}_2$ (denoted by $\bar{H}_2$) can only consist of terms corresponding to diagrams with two out-going and two in-going legs. $\bar{H}_2$ is represented in Fig. 3.

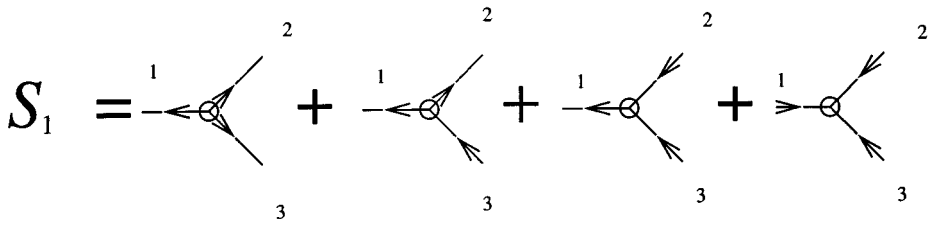

FIG. 2. The function S_1.

The normal form Hamiltonian $H_0+\overline{H}_2$ has several interesting properties. We can easily identify families of periodic and quasiperiodic orbits (the former representing traveling and standing waves) and finite-dimensional invariant manifolds. Also finite-dimensional truncations of $H_0+\overline{H}_2$, restricted to axisymmetric solutions, are completely integrable.

IV. SEMI-CLASSICAL DIAGRAMS

The rules for classical graphs have an interesting quantization that we now describe. As with the classical graphs, the natural application of the quantum diagrams is in semiclassical normal forms algorithm, for which we refer to Ref. 3.

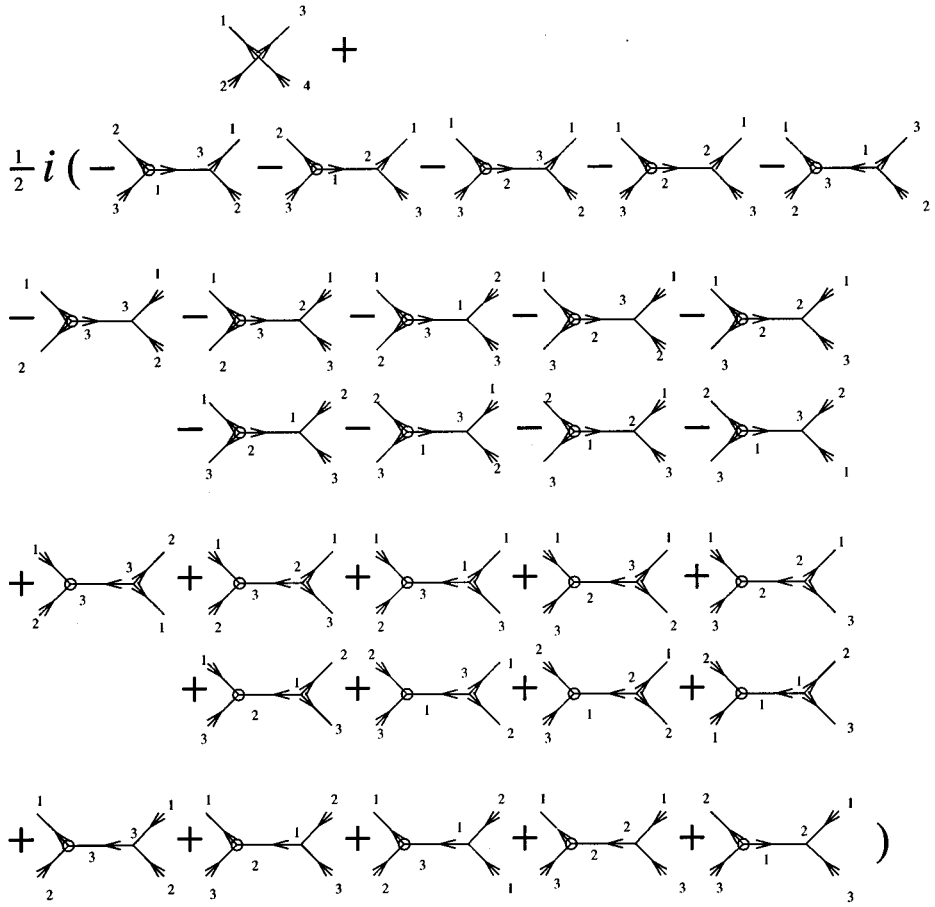

FIG. 3. The quartic part of the normal form Hamiltonian.

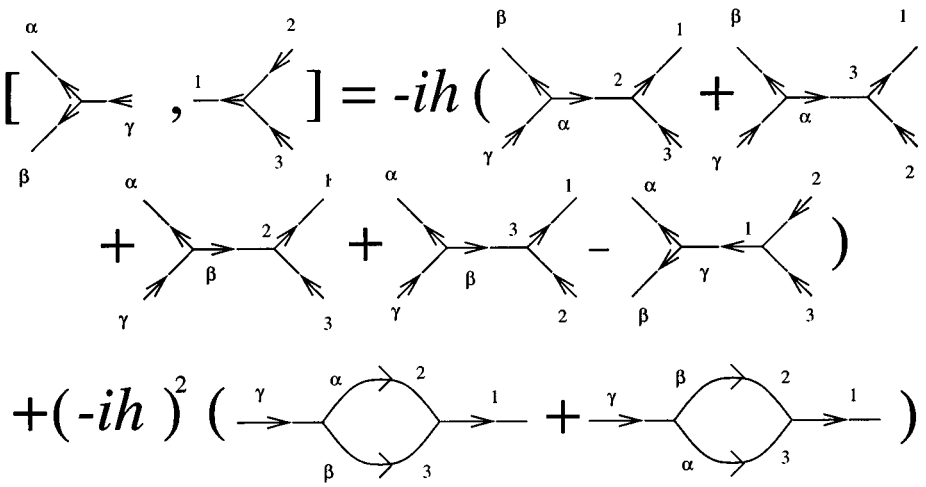

FIG. 4. Commutator of two quantum diagrams.

To describe the quantum diagrams, it is sufficient to consider the associative algebra W generated by the vectors $(a_j^*)\hat{}, \hat{a}_i$, $i,j=1,2,...,d$, that also satisfy the commutation relations

$$[\hat{a}_i,\ (a_j^*)\hat{}\,]=-i\hbar\,\delta_{ij},\quad [\hat{a}_i,\ \hat{a}_j]=[(a_i^*)\hat{},\ (a_j^*)\hat{}\,]=0,\quad \forall i,j=1,2...,d, \tag{4.1}$$

with $[f,g]=fg-gf$, $f,g\in W$. Informally, the $(a_j^*)\hat{}, \hat{a}_i$ are the ''creation–annihilation'' operators.

The graphical rules assign graphs to elements (polynomials) of W and also to the commutator of two graphs. Without loss of generality, we will assign graphs only to elements $\hat{w}$ of W that are Wick-ordered, i.e., with the $(a_j^*)\hat{}$ preceding the $\hat{a}_i$ in each monomial of $\hat{w}$. The rules are as follows:

(1) The set of graphs we consider will now be denoted by D_q. An element of D_q will be a directed tree with indices, constructed recursively in the following way: we take κ one-vertex elements t_i of D (as in Sec. II) and start by joining s_1 out-going (in-going) legs of t_1 to s in-going(out-going) legs of t_2. Denote this by $t_1\cup t_2$. At the ith step we join s_i out-going (in-going) legs of the $t_1\cup\cdots\cup t_i$ to s_i in-going(out-going) legs of t_{i+1}, until $i=\kappa-1$. The s_i are arbitrary. The external legs will carry one index, while internal legs will carry two indices, one for each vertex being joined. Note that $D\subset D_q$.

(2) We first discuss one-vertex trees. To a Wick-ordered homogeneous polynomial

$$\sum_{\kappa_1,...,\kappa_n,\lambda_1,...,\lambda_m} I_{\kappa_1...\kappa_n,\lambda_1...\lambda_m}(a_{\lambda_1}^*)\hat{}\cdots(a_{\lambda_m}^*)\hat{}\,\hat{a}_{\kappa_1}\cdots\hat{a}_{\kappa_n}$$

we assign a vertex with $n+m$ indexed legs with arrows, one out-going (in-going) leg for each $a_{k_i}(a_{k_j}^*)$. [This rule follows the classical one verbatim, with the $\hat{a}_i,(a_j^*)\hat{}$ playing the role of the a_i,a_j^*, respectively.]

The rule for assigning polynomials to multi-vertex graphs is similar to the one for the classical graphs. A multi-vertex tree represents a homogeneous polynomial in the variables indexed by the external legs of the tree. The out-going and in-going arrows represent the a_{κ_i} and $a_{\lambda_j}^*$, respectively. The coefficient of this polynomial will be the product of the coefficients of all the one-vertex subtrees of the multi-vertex tree, summed over all the indices appearing in external legs and contracted over the pairs of indices representing legs that connect vertices. Note that now multi-vertex graphs can be multiply connected. An example is in Fig. 4.

Proposition 4.1: The commutator $[F,G]$ of two graphs F,G is the sum of all graphs obtained by joining n in-going arrows of F to n out-going arrows of G and n out-going arrows of F to n in-going arrows of G, where $n=1,2,3,...$. The graphs will be multiplied by signs and powers of $\hbar$:

$(-i\hbar)^n$ for graphs obtained by joining n out-going legs from F to n in-going of G, and $-(-i\hbar)^n$ for graphs obtained by joining n in-going legs from F to n out-going from G.

Proof: We consider the commutator of two operators

$$\hat{F}=\sum A_{\overline{\kappa}_1,\ldots,\overline{\kappa}_{\overline{\mu}},\kappa_1,\ldots,\kappa_\mu}(a^*_{\overline{\kappa}_1})\hat{}\cdots(a^*_{\overline{\kappa}_{\overline{\mu}}})\hat{}\,\hat{a}_{\kappa_1}\cdots\hat{a}_{\kappa_\mu}$$

and

$$\hat{G}=\sum B_{\overline{\lambda}_1,\ldots,\overline{\lambda}_{\overline{\nu}},\lambda_1,\ldots,\lambda_\nu}(a^*_{\overline{\lambda}_1})\hat{}\cdots(a^*_{\overline{\lambda}_{\overline{\nu}}})\hat{}\,\hat{a}_{\lambda_1}\cdots\hat{a}_{\lambda_\nu}.$$

We first Wick-order $\hat{F}\hat{G}$ by passing the $\hat{a}_{\kappa_i}$ to the right of the $(a^*_{\overline{\lambda}_j})$ in a systematic way. We order the pairs of the subindices (i,j) of the $\hat{a}_{\kappa_i},(a^*_{\overline{\lambda}_j})\hat{}$ lexicographically, i.e., $(1,1)<(1,2)<\cdots<(\mu,\overline{\nu})$. Also we use the multi-index $\gamma=(i,j)$ and for $\gamma=(a,b)$ let $I(\gamma)=a$, $J(\gamma)=b$. Also $\Gamma_0=\{(i,j)|i=1,\ldots,\mu,j=1,\ldots,\overline{\nu}\}$. Define $L_{ij}\hat{F}\hat{G}$ to be $\hat{F}\hat{G}$ with $\hat{a}_{\kappa_i}$ and $(a^*_{\overline{\lambda}_j})\hat{}$ interchanged, and $R_{ij}\hat{F}\hat{G}$ to be $\hat{F}\hat{G}$ with $\hat{a}_{\kappa_i}$ and $(a^*_{\overline{\lambda}_j})\hat{}$ replaced by $[\hat{a}_{\kappa_i},(a_{\lambda_j})\hat{}]$. We write

$$\hat{F}\hat{G}=L_{1,1}\hat{F}\hat{G}+R_{1,1}\hat{F}\hat{G}=L_{1,2}L_{1,1}\hat{F}\hat{G}+R_{1,2}L_{1,1}\hat{F}\hat{G}+R_{1,1}\hat{F}=\cdots,$$

i.e., apply $L_{i,j}$ repeatedly, in the order induced by the (i,j), each time producing a commutator, and so that at most one commutator appears in each term. The procedure terminates after μ steps and we have $\hat{F}\hat{G}=S_1\hat{F}\hat{G}$ with

$$S_1\hat{F}\hat{G}=\sum_{\gamma_1\in\Gamma_1}\widetilde{R}_{\gamma_1}\prod_{\tau_1\in T_1}L_\tau\hat{F}\hat{G},$$

where $\Gamma_1=\Gamma_0$ and $\widetilde{R}_{\gamma_1}=L_{\gamma_1}$ if $\gamma_1=\max\{\gamma\in\Gamma_1\}$, R_{γ_1} otherwise, and $T_1\equiv T(\gamma_1)=\{\tau\in\Gamma_1|\tau<\gamma_1\}$. We repeat the procedure to terms of $S_1\hat{F}\hat{G}$ that are not Wick-ordered, so as to have at most two commutators. We get $S_2\hat{F}\hat{G}$, and similarly $\hat{F}\hat{G}=S_1\hat{F}\hat{G}=\cdots=S_n\hat{F}\hat{G}=\cdots$ with

$$S_n\hat{F}\hat{G}=\sum_{\gamma_1\in\Gamma_1}\left(\cdots\left(\sum_{\gamma_{n-1}\in\Gamma_{n-1}}\left(\sum_{\gamma_n\in\Gamma_n}\widetilde{R}_{\gamma_n}\prod_{\tau_n\in T_n}L_{\tau_n}\right)\widetilde{R}_{\gamma_{n-1}}\prod_{\tau_{n-1}\in T_{n-1}}L_{\tau_{n-1}}\right)\cdots\cdots\right)$$
$$\times\widetilde{R}_{\gamma_1}\prod_{\tau_1\in T_1}L_{\tau_1}\hat{F}\hat{G},$$

where $\Gamma_p=\{\gamma\in\Gamma_0|I(\gamma)>I(\gamma_{p-1}),J(\gamma)\neq J(\gamma_{p-1}),\ldots,J(\gamma_1)\}$, $T_p\equiv T(\gamma_p)=\{\tau\in\Gamma_p|\tau<\gamma_p\}$ and $\widetilde{R}_{\gamma_p}=L_{\gamma_p}$ if $\gamma_p=\max\{\gamma\in\Gamma_p\}$, R_{γ_p} otherwise.

Then note that, first in $S_n\hat{F}\hat{G}$ we encounter all the possible products of commutators $[\hat{a}_{\kappa_{\sigma_1}},(a^*_{\overline{\lambda}_{\rho_1}})\hat{}]\cdots[\hat{a}_{\kappa_{\sigma_n}},(a^*_{\overline{\lambda}_{\rho_n}})\hat{}]$ with $\sigma_1<\sigma_2<\cdots<\sigma_n$ and $\rho_j\neq\{\rho_1,\ldots,\rho_{j-1}\}$, $j=1,\ldots,n$, each such product appearing once as a summand of $S_n\hat{F}\hat{G}$. Moreover, in $S_{n+1}\hat{F}\hat{G}$ all these terms are Wick-ordered and therefore, from the diagrammatic interpretation at $S_{n+1}\hat{F}\hat{G}$, we have all the diagrams with m out-going legs of $\hat{F}$ joined to m in-going legs of $\hat{G}$, with $m=1,2,\ldots,n$, plus $\hat{F}\hat{G}$ Wick-ordered.

Applying the algorithm to $\hat{G}\hat{F}$ we similarly obtain the diagrams with m in-going legs of $\hat{F}$ joined to m out-going legs of $\hat{G}$, with $m=1,2,\ldots,n$, plus $\hat{F}\hat{G}$ Wick-ordered. Considering $\hat{F}\hat{G}-\hat{G}\hat{F}$, we have the proposition.

■

Remark 3.1: Comparing the last rule with its classical analog we have, graphically, the classical limit (see, e.g., Ref. 11)

$$\lim_{\hbar \to 0} \frac{i}{\hbar} [\hat{f}, \hat{g}] = \{f, g\}.$$

In discussing the quantum graphs, quantization, i.e., the map from functions on the phase space to operators in a Hilbert space (see Ref. 11), enters only indirectly, through the commutation relations. In applying the graphs to the semi-classical normal form algorithm of Ref. 2, it is assumed that the quantization step is already taken and that the Hamiltonian operator has been Wick-ordered. Thus we only deal with Wick-ordered operators.

In conclusion, we have presented simple diagrammatic rules for classical and semi-classical normal forms. The rules highlight the combinatorial nature of the computations and can also be implemented in the computer.

ACKNOWLEDGMENTS

I would like to thank Professor de la Llave for encouraging me to write this note and for several suggestions. This work was partially supported by NSF and TARP grants.

[1] V. I. Arnold, *Mathematical methods in the theory of ordinary differential equations* (Springer-Verlag, New York, 1983).
[2] S. Graffi and T. Paul, Commun. Math. Phys. **108**, 71 (1987).
[3] M. degli Esposti, S. Graffi, and J. Herczynski, Ann. Phys. (N.Y.) **209**(2), 364 (1991).
[4] J. Bellisard and M. Vittot, Ann. Inst. Henri Poincaré: Phys. Teor. **52**, 175 (1990).
[5] A. Deprit, Celest. Mech. **1**, 12 (1969).
[6] A. J. Dragt and J. M. Finn, J. Math. Phys. **17**, 2215 (1976).
[7] J. R. Cary, Phys. Rep. **79**(2), 129 (1981).
[8] D. Bambusi, Nonlinearity **8**, 93 (1995).
[9] R. Grossman and R. Larson, J. Symbol. Comp. **13**, 511 (1992).
[10] R. de la Llave and P. Panayotaros, J. Nonlinear Sci. **6**, 147 (1996).
[11] F. A. Berezin and M. A. Shubin, *The Schrödinger equation* (Kluwer Academic, Dordrecht, 1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 4 APRIL 1998

Exact solutions for nonconservative two-velocity discrete Boltzmann models

H. Cornille
Service de Physique théorique, CE Saclay, F-91191 Gif-sur-Yvette, France

(Received 31 January 1997; accepted for publication 29 December 1997)

Extended discrete kinetic theory (which we call nonconservative), including sources, sinks, creation and annihilation of test particles, inelastic scattering,... added to the elastic collisions, was recently introduced. For instance, the mass conservation law, by adding polynomials of the mass, becomes nonconservative. To the classical two-velocity elastic collision models, we associate two nonconservative models with polynomials of the first and second degree. We obtain exact stationary, similarity, $(1+1)$-dimensional and periodic solutions. For their stability we study the corresponding nodes and saddles. © *1998 American Institute of Physics.* [S0022-2488(98)01304-8]

I. INTRODUCTION

Recently,[1–3] extended discrete kinetic theory (nonconservative) became a popular field of research. The reason being, as it is written in the literature, that the *study of a gas with only elastic scattering between the molecules seems to be too idealized for application to the real world*. So they add a background medium, external sources and sinks, effects of absorption and generation due to inelastic scattering... . A great difference, with conservative discrete kinetic theory,[4] is that the conservation laws are modified by including polynomials in the mass variable. While for conservative models the determination of exact solutions is well understood,[5] for *nonconservative models exact solutions are missing*. One advantage of these nonconservative models is that important difficulties of the conservative ones disappear. For instance, the models with two-velocity violating momentum conservation relation, those with only one speed for which mass and energy conservation laws cannot be distinguished, the problem of appearance of spurious conservation laws for which ternary, quaternary collisions (not present in the Boltzmann equation) must be introduced... . However, I recall that in the instance of shock waves, the existence of conservation laws[4] is useful for the study of supersonic and subsonic inequalities and applications of the Whitham and Lax criteria. There remains, in both cases, the problem of the link between discrete and continuous theory (see Cercignani[4]).

On the other hand, for the exact solutions of conservative models, the existence of linear conservation laws is crucial for the understanding of these solutions. Let us summarize: for the similarity solutions let us first consider the models where in addition to the linear conservation laws, only one nonlinear quadratic equation remains. When we introduce the properties coming from the linear differential conservation law, the nonlinear one is reduced to a scalar Riccati[6] equation that we can integrate. Second consider the models with two, three,... remaining nonlinear quadratic equations. Integrating the linear conservation laws leaves only two, three... independent densities satisfying a 2×2, 3×3,... Riccati coupled system. Then, with the compatibility condition, we can still obtain scalar Riccati-type of solutions. However, other solutions exist[5] which are not of the scalar Riccati-type. Let us now consider $(1+1)$, $(2+1)$ dimensional solutions which are sums $f=\Sigma f_i$ of two, three similarity solutions. From the superposition principle they are automatically satisfied for the linear differential conservation laws. For the remaining nonlinear equations we obtain compatibility conditions between the loss and gain collision terms. For quadratic, cubic... collision terms we have constraints coming from the products $f_i f_j$, $f_i f_j f_k$,... .

Now clearly when the linear conservation laws do not exist, they cannot be obtained in this context; our goal is to try to obtain a new characterization of the exact solutions. For simplicity we limit our study to elastic collisions with two velocities—the Illner[7] models which include the

 © 1998 American Institute of Physics

McKean, Carleman, and dual McKean models. The addition of only constants in the linear conservation laws, previously studied,[1–5] will not be considered.

For a two-velocity model with densities N_i, $i=1,2$ associated to the velocities ± 1 let us consider the mass conservation law for the mass $\rho=N_1+N_2$ and the momentum $j=N_1-N_2$. For the nonconservative models the rhs becomes a polynomial of degree k in ρ,

$$\rho_t+j_x=0\rightarrow P_k(\rho). \tag{1.1}$$

It remains a nonlinear equation which, for the conservative models, depends only on the elastic collision term. Here we consider the Illner model with collision terms for N_i,

$$\mathrm{Col}_{N_1}=-\mathrm{Col}_{N_2}=aN_1^2-(a+c)N_1N_2+cN_2^2,\ \ a\leqslant 0,\ \ c\geqslant 0,$$

called McKean when $a=0$, $c=1$, Carleman for $-a=c=1$, and dual McKean for $a=-1$, $c=0$. For the nonconservative model, Col_{N_i} correspond to the collisions between the test particles and we add another polynomial in ρ, j,

$$\rho_x+j_t-(a+c)j^2-j\rho(a-c)=0\rightarrow Q_k(\rho,j). \tag{1.2}$$

In Sec. II we consider the Platkowski–Spiga[2] linear nonconservative model,

$$P_1=(\eta-\epsilon)\rho+S,\ \ S=S_1-S_2,\ \ Q_1=-j\epsilon+\eta\rho(2\kappa-1)+S_1+S_2, \tag{1.3}$$

with $\epsilon\geqslant 0$ ($\eta\geqslant 0$), for the absorption collision (generation) by the background and $\kappa\in(0,1)$ (or $1-\kappa$) for the secondary particles generated with velocity $+1$ (or -1) with the background. $S_1>0$, ($S_2>0$) are associated to the external source (sink). We determine exact solutions and study their stability properties. The equilibrium states for the mass ρ is given by $P_1=0$ with only one state for this model. Up to now we have not found that traveling satisfies the positivity constraints. Let us consider stationary solutions for the associated conservative model. We get $j=c_1$, $\rho_x=c_2+c_3\rho$ ($c_i=$const.) and exponential type solutions which are not bounded when $|x|=\infty$. The nonconservative models could be interesting if we obtain solutions bounded when $|x|=\infty$. For stationary solutions the (1.1)–(2)–(3) equations give, for the momentum $j=j_0+j_1J(x)$, a second-order nonlinear differential equation satisfying the necessary Painlevé condition for the nonexistence of movable critical singularities except poles.[8] We want to give examples of Riccati solutions either bounded or not. For a first class (Figs. 1) with only j_0 as asymptotic state, Sec. II A, we apply the Bureau[8] transform giving Riccati equations with or without x-dependent coefficients. We obtain either solutions written in terms of the modified Bessel function $I_0(e^{Cx/2})$ not bounded when $|x|=\infty$ and we require a specular reflection boundary condition, or solutions[9] $j(x)=j_0+j_1/(x+d)$, infinite for $x+d=0$, $d>0$ or <0 but there exist positive $N_i(x)$ solutions. For a second class with j_0, j_0+j_1 asymptotic states, Sec. II B, we apply another transform, still reducing to Riccati equations with either solutions of the modified Bessel-type or solutions $j(x)=j_0+j_1/(1+de^{\gamma x})$ bounded when $|x|=\infty$ (Figs. 2). For the two classes we must satisfy the same parameters $S(a,c,\epsilon,\eta,\kappa)$ condition. For the elementary solutions with d an arbitrary integration constant, contrary to solutions with inverse powers, those with inverse exponentials and $d>0$ are not infinite for a particular x value. For these solutions, we find two cases for the positivity of the densities: either only solutions with $d>0$, (finite for any x value) or for both $d>0$ and $d<0$ (infinite for one x value). We study their stability properties and determine the saddles and nodes. We consider two classes of models; either with only absorption $\epsilon\neq 0$, $\eta=0$ and source, sink S_i or adding generation $\eta\neq 0$, $\kappa\neq 0$. With $S_1\neq 0$ we find either sinks $S_2>0$ or $S_2=0$. For the Illner model, we obtain positive solutions for $a/c\in(0,-0.1)$ and for the dual McKean model new solutions with $S_1<0$ (sink) and $S_2<0$ (source).

In Sec. III we consider the Piechòr–Platkowski[3] quadratic nonconservative model where the nonlinear (including Illner term) equations are written for the densities N_i, $i=1,2$,

$$\begin{gathered}P_2=(\alpha-\beta)\rho^2+(\eta-\epsilon)\rho+\sigma,\ \ \sigma=S_1+S_2+\zeta,\ \ L_i\equiv N_{i,t}+(-1)^{i+1}N_{i,x}\equiv R_i,\\ R_i\equiv\mathrm{Col}_{N_i}+(2-i+(-1)^i\kappa)[\alpha\rho^2+\eta\rho+\zeta]-N_i(\beta\rho+\epsilon)+S_i,\end{gathered} \tag{1.4}$$

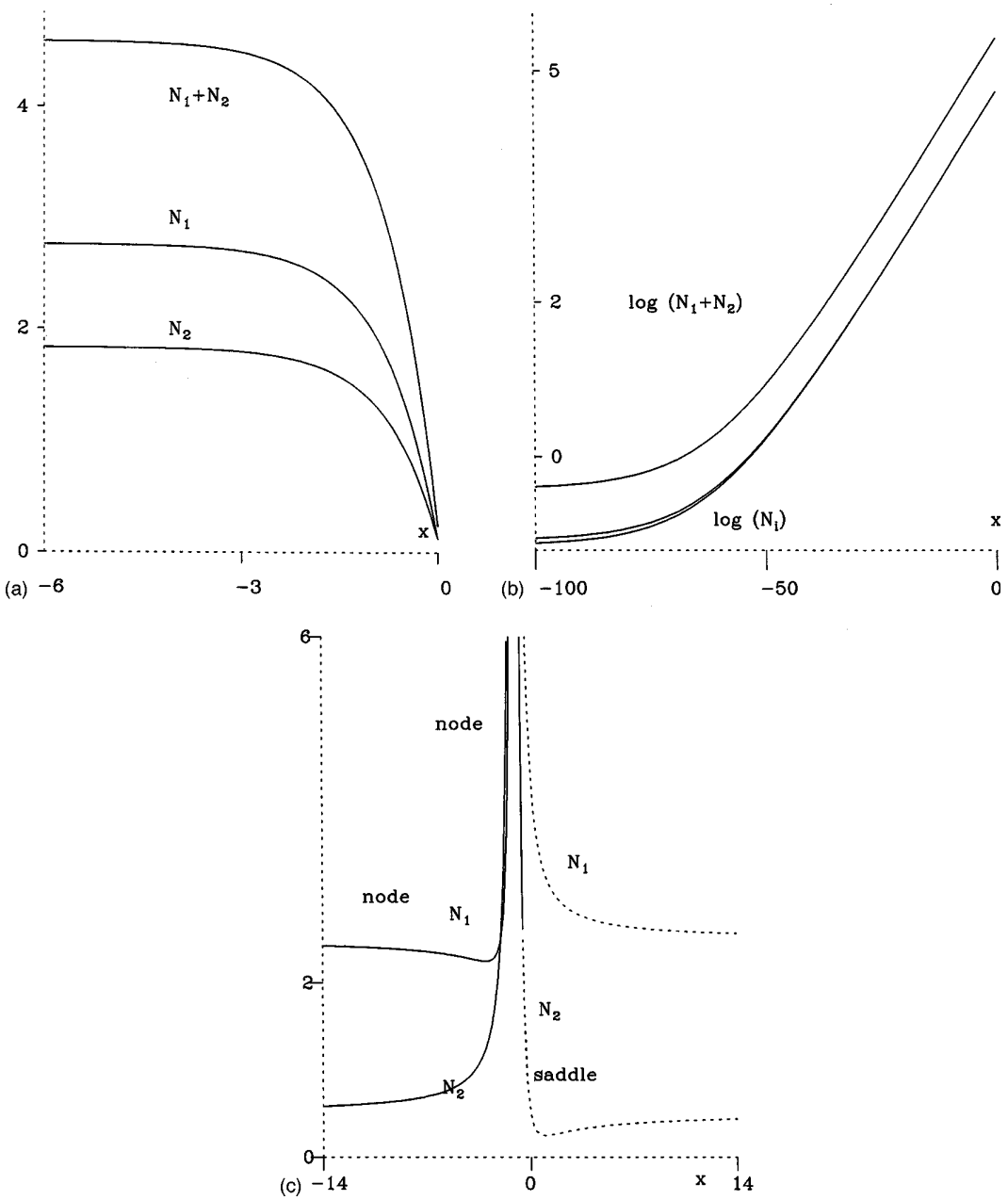

FIG. 1. Linear nonconservative model. (a), (b) Modified Bessel stationary solutions with specular reflection at $x=0$. (a) $a=-0.4$, $c=0.2$ S_1 source, S_2 source and (b) $a=-0.95$, $c=1$ S_1 sink, S_2 sink. (c) Inverse power-like stationary solution for McKean $a=0$, $c=1$, $\epsilon=1$ model with S_1 source, S_2 sink.

with $\beta\geqslant 0$, $\epsilon\geqslant 0$ ($\alpha\geqslant 0,\eta\geqslant 0,\zeta\geqslant 0$) for the annihilation (creation) of test particles as a result of inelastic collisions, $\kappa=0$ or 1 represents the fraction of secondary particles generated with velocity $+1$ or -1 and $\sigma>0$ is essentially the external bulk source.

Here, contrary to the previous model, the equilibrium states for the mass ρ are the two different roots of $P_2=0$. We are more interested in similarity, $(1+1)$-dimensional and periodic solutions. Firstly we show that the similarity (stationary) solutions with the variable $z=x-\xi t(z=x)$ can be obtained from a compatibility condition between two Riccati equations with constant coefficients or equivalently between two linear first-order differential equations. These solutions are rational functions of an exponential variable $de^{\gamma z}$. We study the stability at the two asymptotic states. Second, for the $(1+1)$-dimensional solutions (which are not sum of two similarity solutions) we find the same structure, the exponential variable being replaced by a sum of an arbitrary

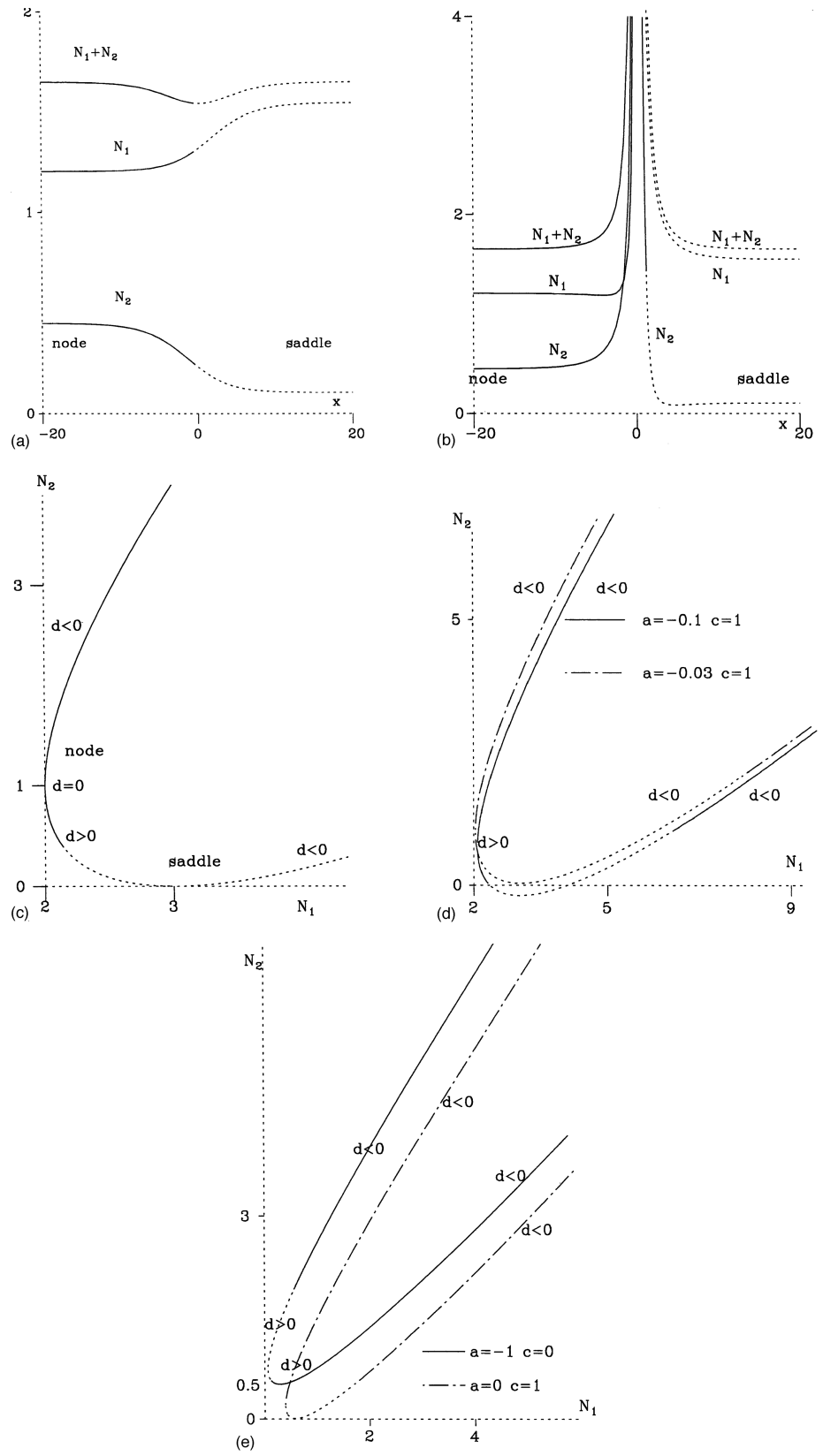

FIG. 2. Linear nonconservative model. Stationary solutions. Only absorption and McKean model $a=0$, $c=1$. $d>0$ in (a) $d<0$ in (b) and N_1, N_2 curve for $d>0$, <0 in (c). (d) Stationary solutions. Only absorption and Illner model, $c=1$ and N_1, N_2 curves for $d>0$, <0. For $a=-0.03$ both $N_i>0$ while for $a=-1$ part of $N_2<0$ for $d<0$. (e) Stationary solution. Absorption, generation, and positive N_1, N_2 curve for $d>0$, <0. Both McKean model with $a=0$, $c=1$, S_i source, sink and dual McKean model with $a=-1$, $c=0$, S_i sink, source.

number of $d_i e^{\gamma i(x-\xi_i t)}$ terms, one density being constant and only for the McKean and dual McKean models. In Sec. III A for the model with $\kappa=0$ we study the similarity solutions for both the Illner, McKean, Carleman, and dual McKean models. In Sec. III B for the McKean and its dual models we get both $(1+1)$-dimensional and periodic solutions.

II. EXACT SOLUTIONS FOR THE MODELS WITH A LINEAR NONCONSERVATIVE RELATION

For the Platkowski–Spiga models with mass ρ and momentum j we rewrite both (1.1)–(2)–(3),

$$\rho_t+j_x=(\eta-\epsilon)\rho+S,\ \ S_1>0,\ \ S=S_1-S_2>0,\ \ \epsilon>0,\ \ \eta>0, \tag{2.1}$$

$$\rho_x+j_t=j^2(a+c)+j(\rho(a-c)-\epsilon)+\rho\eta(2\kappa-1)+S_1+S_2,\quad \kappa\in(0,1). \tag{2.2}$$

We have found both similarity (functions of $x-\xi t$), $(1+1)$ and stationary solutions but only the last ones are positive. For these stationary N_i solutions we mainly study those with second-order movable poles in the variable $w=de^{\gamma x}$ (d being arbitrary). What is new, compared with previous similar exact solutions, is that there exist solutions $N_i(x,d)$ which remain positive for both $d>0$ and <0 values.

For the stationary solutions we rewrite (2.1)–(2), define ($'=d/dx$), the constant asymptotic values $\rho_{\rm as}=\rho_0$, $G(j_{\rm as})=0$ when $j'=0$,

$$j(x)=j_0+j_1J(x),\ \ \rho(x)=\rho_0+j'/(\eta-\epsilon),\ \ \rho_0=S/(\epsilon-\eta)=\rho_{\rm as}, \tag{2.1$'$}$$

$$[J'-J^2(a-c)j_1/2-J(j_0(a-c)+\eta(2\kappa-1)]'\equiv\overline{G}(J,j_0,j_1)+G(j_0)/j_1, \tag{2.2$'$}$$

$$\overline{G}=(\eta-\epsilon)[J^2j_1(a+c)+J[2j_0(c+a)+\rho_0(a-c)-\epsilon)]],$$

$$G(j_{\rm as})/(\eta-\epsilon)=j_{\rm as}^2(c+a)+j_{\rm as}[\rho_0(a-c)-\epsilon]+S_1+S_2+\rho_0\eta(2\kappa-1). \tag{2.3}$$

No J'^2 term is present in (2.2$'$) which belongs, in the Painlevé classification,[8] to a well-known class. We seek Riccati-type of solutions. Different cases occur with $G(j_0)=0,\neq 0$, $G(j_0+j_1)=0,\neq 0$... . As we shall see $j(x)$, for $N_i(x)>0$ solutions, can be not bounded, in particular when $|x|=\infty$, or bounded for all x.

A. Bureau transform with only $G(j_0)=0$

In Appendix A1 we give a transform[8] $W(x)=J'(x)+J^2(x)$, $W'=CW$ leading to Riccati-type of equations for $W\neq 0,=0$. First, in (A11) we get $J(x)=\partial_x \log I_0(e^{Cx/2})$ (I_0 modified Bessel function) with momentum $j(x)$ not bounded when $e^{Cx/2}\to\infty$. We introduce a specular reflection boundary condition $N_1(0)=N_2(0)$ and $N_i(x)$ along a half-axis $Cx\leqslant 0$. In Fig. 1(a), (b) we get S_i sources (sinks). For $(a,c,\eta,\epsilon,\kappa,S_1,S_2,S,C)$ we have in Fig. 1(a) $(-0.4,0.2,0.85<2.7,0.91,0.2,-8.31,>0,1.26)$ and in Fig. 1(b), $(-0.95,1,2.65>0.67,0.51,-0.7,0.62,<0,0.1)$.

Second,[9] in (A13) we consider a class $J(x)(x+d)=1$ of power-like N_i solutions, infinite when $x=-d$ ($d>0$ or <0), but positive [see (A3)–(A3$'$)]. In Fig. 1(c) we present a McKean solution with $c=\epsilon=d=1$, $a=\eta=\kappa=0$, $S_i=3.5,0.5$ (source and sink).

B. Transform associated with $G(j_0)=G(j_0+j_1)=0$

In Appendix A2 we give a transform $\gamma:=j_1(a-c)/2$, $W(x)=J'(x)-\gamma J(J-1)$, $W'=CW$ leading to Riccati-type of equations with $W\neq 0,=0$. From $G(j_{\rm as})=0$ we get a j_0,j_1 relation and rewrite $\overline{G}$,

$$j_1+2j_0=[S(a-c)+\epsilon(\eta-\epsilon)]/(a+c)(\eta-\epsilon),\ \ \overline{G}=j_1(a+c)(\eta-\epsilon)J(J-1). \tag{2.4}$$

In the sequel we are restrict to $W(x)=0$ and (2.2$'$)–(4) leads to two Riccati equations,

$$J'=\gamma J(J-1)=-[j_1(a+c)(\eta-\epsilon)]J(J-1)/[(j_1/2+j_0)(a-c)+\eta(2\kappa-1)]. \tag{2.5}$$

TABLE I. Stationary solutions for the linear nonconservative models.

$\delta \doteq [\rho_0(c-a+\epsilon)/2(a+c)]^2-[S_1+S_2+\rho_0\eta(2\kappa-1)]/(a+c)$
$j_0=[\rho_0(c-a)+\epsilon]/2(a+c)\pm\sqrt{\delta},\ j_1=\mp 2\sqrt{\delta}=2\gamma/(a-c),$
$j(x)=j_0\mp 2\sqrt{\delta}/\Delta(x),\ \Delta(x)-1=w(x)=de^{\gamma x},\ \gamma=\pm(c-a)\sqrt{\delta}$
$(\epsilon-\eta)\rho(x)=S+2\delta(a-c)w(x)/\Delta^2(x),\ de^{\gamma x}=-1+j_1/(N_1-N_2-j_0)$
$j_1^2(N_1+N_2-\rho_0)=\rho_1(N_1-N_2-j_0)(j_1-N_1+N_2+j_0)$

McKean model $a=\eta=0,\ c=1,\ S_1=3\epsilon^2+S_2$:
$S_2\neq 0$: $j(x)=2\epsilon\pm\sqrt{\delta}[1-2/\Delta(x)],\ \rho(x)=3\epsilon-2\delta w(x)/\epsilon\Delta^2,$
$\Delta-1=w(x)=de^{\pm\sqrt{\delta}x},$
$\delta=\epsilon^2-2S_2,\ 2n_{1,\mathrm{as}}=5\epsilon\pm\sqrt{\delta}>0,\ 2n_{2,\mathrm{as}}=\epsilon\mp\sqrt{\delta}>0$
$S_2=0$: $S_1=S=3\epsilon^2,\ j/\epsilon=2\pm 1\mp 2/\Delta,\ \rho/\epsilon=3-2w/\Delta^2,$
$\Delta(x)-1=w(x)=de^{\pm\epsilon x}$
$N_1/\epsilon=(5\pm 1)/2-[\pm 1+w(\pm 1+1)]/\Delta^2,$
$N_2/\epsilon=(1\mp 1)/2+[\pm 1+w(\pm 1-1)]/\Delta^2$

The compatibility gives, with (2.4), a relation between the model parameters $S(\epsilon,\eta,\kappa,a,c)$,

$$S(c-a)=[2(\epsilon-\eta)(c+a)/(a-c)]^2-(\epsilon-\eta)[\epsilon+2\eta(1-2\kappa)(c+a)/(c-a)]. \tag{2.6}$$

Writing $J=1/\Delta(x)$ we get $\Delta=1+w,\ w=de^{\gamma x}$ with d arbitrary (see Table I).

In Appendix A21 we study the linear stability around the asymptotic values $|x|=\infty$ and for any finite x. One characteristic value is $\eta-\epsilon$ negative for $\eta=0$, while the other has opposite values for $|x|=\infty$ and vanishes for one finite $w=de^{\gamma x}>0$ and another <0 values. Consequently, for both $d>0$ and $d<0$ solutions, the second characteristic vanishes only one time and we always have both a node (attractive for $\eta=0$) and a saddle.

Illner models with only absorption, $-0.1\leqslant a/c\leqslant 0$ *and* $\epsilon\neq 0$. In Appendix A22 we give the positivity proof and in Table I summarize the McKean results for $S_2\neq 0,=0$.

Numerical calculations for N_i, $\rho=N_1+N_2$ and both the McKean (dual) models [Figs. 2(a)–2(c) and 2(e)] and the Illner model $|a/c|\leqslant 0.1$, Fig. 2(d). The solid and dotted lines correspond, respectively, to attractive nodes and saddle-points,

Figs. 2	a	c	ϵ	S_1	S_2	S	d
a,b,c	0	1	0.55,0.55,1	1	0.092,0.092,0	0.9,0.9,1	$0.9,-0.9$
							$>0<0$
d,e	$-0.1,-0.03$	1	1.6,1.15	4.5,3.6	0.596,0.33	3.9,3.27	$>0<0$

In Fig. 2(a) with $d>0$, $(1+w)^2\neq 0$ and we find N_i monotonic and ρ nonmonotonic. On the contrary $d<0$ in Fig. 2(b), the denominator of ρ has a double root and $\rho>0$ while with a simple root $j>0$ and <0. The positivity conditions, only for N_i,ρ, (infinite for a particular value) are satisfied. In Fig. 2(b) both N_i and ρ are nonmonotonic. Other $d>0$, <0 values lead to similar patterns. The transition value being $d=0$, we choose in Fig. 2(c),

$$N_1(x)=3-(1+2w(x))/\Delta^2,\ \ N_2(x)=1/\Delta^2,\ \ N_1=3+N_2+\zeta 2\sqrt{N_2},\ \ \zeta^2=1, \tag{2.7}$$

and present the N_1,N_2 curve which is d-independent but different for $d>0$ or <0. The two $d>0$, <0 curves are separated; both $N_2<1$, $N_1<3$ for $d>0$, $\zeta=-1$ and either $N_2>1$, $\zeta=-1$ or $N_1>3$ for $d<0$, $\zeta=1$. $d=0$ is at the transition $N_2=1$, $N_1=0$. The change between solid and dotted lines, depending only on a w fixed value, is independent of d and is always the same here. We find for $d>0$, $w=(\sqrt{5}-1)/2$, $N_2=0.38$, $N_1=2.14$ and for $d<0$, $w=-(\sqrt{5}+1)/2$, $N_2=2.6$, $N_1=8.85$ [not present in Fig. 2(c)].

For the Illner models ($a<0$) we can write the N_2,N_1 curve independent of $w(x,d)$ and deduce the sign of $de^{\gamma x}$ (Table I). One N_2 curve with $d<0$ can violate positivity. In Fig. 2(d), $a=-0.1$, $d>0$, $N_i>0$ corresponds to $N_1\in(2.38,2.49)$, $N_2\in(0.04,0.056)$ while for $d<0$ and $N_1\in(2.44,4.12)$ one branch of the N_2 curve is negative. For $a=-0.03$, $N_i>0$ for $d>0$ and <0. Starting with $a=0$ or $-a$ small, $N_i>0$ for $d>0$, <0 exist while when $-a$ increases, after a critical a value, positivity subsists only for $d>0$.

Models with absorption and generation $\eta\neq 0$, $\epsilon\neq 0$. For $\eta<\epsilon$, η sufficiently small all the previous solutions with $\eta=0$, $S_1>0$ (source) and $S_2\geqslant 0$ (sink) can be extended but new $\eta\neq 0$ solutions exist. For the dual McKean model we find for $\kappa\in(0,1)$ positive solutions with $S_1<0$ (sink) and $S_2<0$ (source). In Fig. 2(e) for $\eta=0.01$, $d>0$, <0, for the McKean $a=0$, $c=1$ and the dual McKean $a=-1$, $c=0$ models we, respectively, choose for $(\epsilon,\kappa,S_1,S_2,S)$: (0.2,1,0.11,0,0.11) and (0.237,0,−0.025,−0.182,0.157).

III. EXACT SOLUTIONS FOR MODELS WITH A QUADRATIC NONCONSERVATIVE RELATION

For the Piechor–Platkowski models (1.4), for the mass ρ and one of the two densities $w=2N_2$ we rewrite both the nonconservative equation and the w nonlinear equation, including the Illner models, with non-negative constants α, β, ζ, η, ϵ, σ and $\kappa=0$ or 1,

$$\rho_t+j_x=\rho_t+\rho_x-w_x=(\alpha-\beta)\rho^2+(\eta-\epsilon)\rho+\sigma,\quad \alpha\neq\beta, \tag{3.1}$$

$$w_x-w_t=2a\rho^2+(c+a)w^2-\rho w(3a+c)+w(\beta\rho+\epsilon)-2\kappa(\alpha\rho^2+\rho\eta+\zeta)+2S_2. \tag{3.2}$$

A. Similarity solution $\rho(z=x-\xi t)$, $w(z)$, $N_i(z)$

We define $'=d/dz$, begin with the (1.4) coupled Riccati equations and look at the standard[10] solutions. For the Riccati-projectif, from the N_i' equation without N_j^2 term, we get $a=c=\alpha=0$ not possible. For the conformal Riccati solution and a generalization[5] we get as necessary conditions either $\alpha=\beta$ or $c-a+\beta=0$, both not possible.

We write the asymptotic (3.1)–(3.2) states $\rho_{as}=\rho_0$, $\rho_0+\rho_1$ and $w_{as}=w_0$, w_0+w_1,

$$(\alpha-\beta)\rho_{as}^2+(\eta-\epsilon)\rho_{as}+\sigma=0,\quad \rho_1=-2\rho_0-(\eta-\epsilon)/(\alpha-\beta), \tag{3.3}$$

$$w_{as}^2(a+c)=w_{as}(\rho_{as}(3a+c-\beta)-\epsilon)-2a\rho_{as}^2+2\kappa(\alpha\rho_{as}^2+\eta\rho_{as}+\zeta)-2S_2. \tag{3.4}$$

In Appendix B1 we start with $\rho=\rho_0+R(z)/D(z)$, $w=w_0+W(z)/D(z)$, substitute into (3.1)–(3.2), get a D^{-1} polynomial, assume that the coefficient of the most singular term D^{-2} is zero and deduce $R\propto W$. We go on writing $\rho=\rho_0+\rho_1 Y(z)$, $w=w_0+w_1 Y(z)$. We get two scalar Riccati equations $C_i Y'=Y(Y-1)$ with C_i constants and a compatibility condition $C_1=C_2$ or equivalently two linear first-order differential equations for $\Delta(z)=1/Y$. We integrate, get a variable $de^{\gamma z}$ and deduce γ and the speed wave ξ (see Table II).

For stationary solutions, it is sufficient to require $\xi=0$, leading to a constraint between the model parameters: $\alpha,\beta,\eta,\epsilon,a,c$.

B. Similarity solutions for the models $c\geqslant 0$, $a\leqslant 0$, $\kappa=S_2=0$, and w_{as}, $(c+a)w_{as}^2-w_{as}(\rho_{as}(3a+c-\beta)-\epsilon)+2a\rho_{as}^2=0$

We begin with the Illner models, $a<0$, $c>0$, $a+c>0$ and discuss the positivity conditions for the asymptotic states (3.3)–(3.4). For ρ_{as} we assume $\alpha>\beta$, $\eta<\epsilon$, the conditions in order that the two roots ρ_0, $\rho_0+\rho_1$ are real and they are positive. The product of the two w_{as} roots is negative. Respective to ρ_0, $\rho_0+\rho_1$ we associate the positive w_{as} roots for N_2. For $N_1=\rho-w/2>0$ or $n_{1as}=\rho_{as}-w_{as}/2>0$ we give a rough estimate—more accurate results require heavy analytical studies,

$$-2<3a/c<-1+\beta\rightarrow \text{sign } n_{1as}=\rho_{as}^2(3a+3c)+w_{as}(\epsilon+\rho_{as}(\beta-3a-c)/2)>0.$$

Numerically for $-1\leqslant a\leqslant 0$, $c=1$ we have found restrictions on the other parameters.

TABLE II. Similarity solutions for quadratic nonconservative models.

$\rho=\rho_0+\rho_1/\Delta(z)$, $w=w_0+w_1/\Delta(z)$, $\Delta=1+de^{\gamma(x-\xi t)}$, $\overline{w_1}=w_1/\rho_1$
$N_1=-N_2(1-2\rho_1/w_1)+\rho_0-\rho_1 w_0/w_1=n_{01}+n_1/\Delta$
$X=\overline{w_1}(c+a)+2(a-\alpha\kappa)/\overline{w_1}-3a-c+\beta$, $\gamma=\rho_1 X/(1+\xi)$,
$\xi=(\beta-\alpha+(1-\overline{w_1})X)/(\alpha-\beta+X)$, [$\rho_{as}$, w_{as} in (3.3)–(3.4)]

For the Carleman model with $a+c=0$ we define $Z=\epsilon+\rho_{as}(2c+\beta)>0$ and all n_{ias} are positive; $w_{as}=2\rho_{as}^2 c/Z>0$, $n_{1as}=\rho_{as}(\epsilon+\rho_{as}(c+\beta))/Z>0$.

For the McKean model with $a=0$ we write w_{as} with two different cases,

$$w_{as}=0\rightarrow n_{1as}=\rho_{as}>0,$$

$$cw_{as}=-\epsilon+\rho_{as}(c-\beta)\rightarrow n_{1as}=\rho_{as}/2+(\epsilon+\beta\rho_{as})/2c>0.$$

For $w_{as}>0$, a sufficient condition is given by $(c-\beta)(\epsilon-\eta)/(\alpha-\beta)>2\epsilon$ (satisfied for c sufficiently large), if for ρ_{as} we choose in (3.3) the root with positive discriminant.

For the dual McKean model with $c=0$, $a<0$ the w_{as} are given by a quadratic polynomial with positive product and sum so that if they are real,

$$w_{as}^2+w_{as}(\rho_{as}(-3+\beta/a)+\epsilon/a)+2\rho_{as}^2=0\rightarrow w_{as}>0. \tag{3.5}$$

In Appendix B2 we study the linear stability around the two asymptotic states ρ_{as}, w_{as} and determine the two characteristic values. For the Illner model we find a node and a saddle, for the Carleman model the node is attractive, for the McKean model and the dual McKean model we can have either nodes or saddles or node, saddle.

Numerical calculations for the densities $N_i(z)$, mass $\rho(z)$, and curves N_1, N_2 for the Illner $(a,c)=(-0.1,1)$, Figs. 3(a)–3(b), Carleman $(-1,1)$, Fig. 3(b), and dual McKean models$(-1,0)$, Fig. 3(d). We choose the same model parameters values $\alpha=0.05$, $\beta=0.01$, $\eta=1.81$, $\epsilon=2.1$, $\sigma=0.5$ and deduce $(\xi,\gamma)=(0.75,0.1)$, $(0.65,1.05)$, $(0.5,1.7)$.

The $N_i(z)$ curves are very similar with N_i increasing from a saddle $(z=-\infty)$ to an attractive node $(z=\infty)$ so that for the two last models, for brevity, we only present in Fig. 3(b) the N_1, N_2 curves which are straight lines. However, for the dual McKean model [curve (c) in Fig. 3(b)], the above theoretical prediction was that the pair (attractive node saddle) could change. So we keep all the above parameters values except $\epsilon\in[2.1,3.1]$. We obtain a transition for $\epsilon\sharp 3$, find a repulsive node at $z=-\infty$, and a saddle at $z=\infty$. While the γ sign is the same we have $\xi>1$ and the transition occurs when $\xi=1$.

In Fig. 3(c) for the McKean model with the theoretical predictions that the pair (saddle-node) could change, we choose the same model parameters values $\alpha=0.05$, $\beta=0.01$, $\eta=0.25$, $\sigma=0.012$ except for $\epsilon=0.3$, 0.32 leading to $(\xi,\gamma)=(-0.55,0.045)$, $(0.91,-0.07)$. We find, respectively, saddle, attractive node and saddle, saddle. At the transition $|\xi|\rightarrow\infty$, $\xi\rightarrow 0$ and ξ, γ change of signs, so that the $N_i(z)$ which were decreasing become increasing. For the characteristics at the $z=\infty$ state, the attractive node with $\xi<0$ becomes repulsive (but $\xi>1$) and finally a saddle ($\xi<1$). At the $z=-\infty$ state the saddle ($\xi<1$) becomes a repulsive node (but $\xi>1$) and finally a node ($\xi<1$).

C. (1+1)-dimensional solutions

We introduce similarity variables $\nu_i=d_i e^{\gamma_i(x-\xi_i t)}$. In Appendix B3.1 we first show that the sum of two similarity solutions is not possible. For conservative models such solutions lead to conditions between loss and gain terms while here in (3.1) only ρ^2 is present. Second, starting from solutions which are ratios of first-order ν_i polynomials by second-order ones, we obtain from (3.1), the following possible solutions:

$$\rho=\rho_0+\rho_1/\Delta,\quad w=w_0+w_1/\Delta,\quad \Delta=1+\sum\ \nu_i,\quad \nu_i=d_i e^{\gamma i(z-\xi_i t)}. \tag{3.6}$$

In Appendix B32, starting with (3.6), we obtain, from both (3.1)–(3.2), two possibilities,

$$\text{dual McKean:}\quad c=0,\ \kappa=1,\ \beta+a=0,\ w_1=2\rho_1,\ N_1=w_0-2\rho_0=\text{const.} \tag{3.7a}$$

$$\text{McKean}\quad a=0,\ \kappa=0,\ \beta-c=0,\ w_1=0,\ N_2=w_0/2=\text{const.} \tag{3.7b}$$

(1+1)-dimensional solutions for the McKean (dual) models. We present a more general formalism starting from (1.4) and assuming

$$N_i(x,t)=n_{0i}+n_i N(x,t),\quad R_i(n_{1as},n_{2as})\equiv 0,\quad n_{ias}=n_{0i},n_{0i}+n_i. \tag{3.8}$$

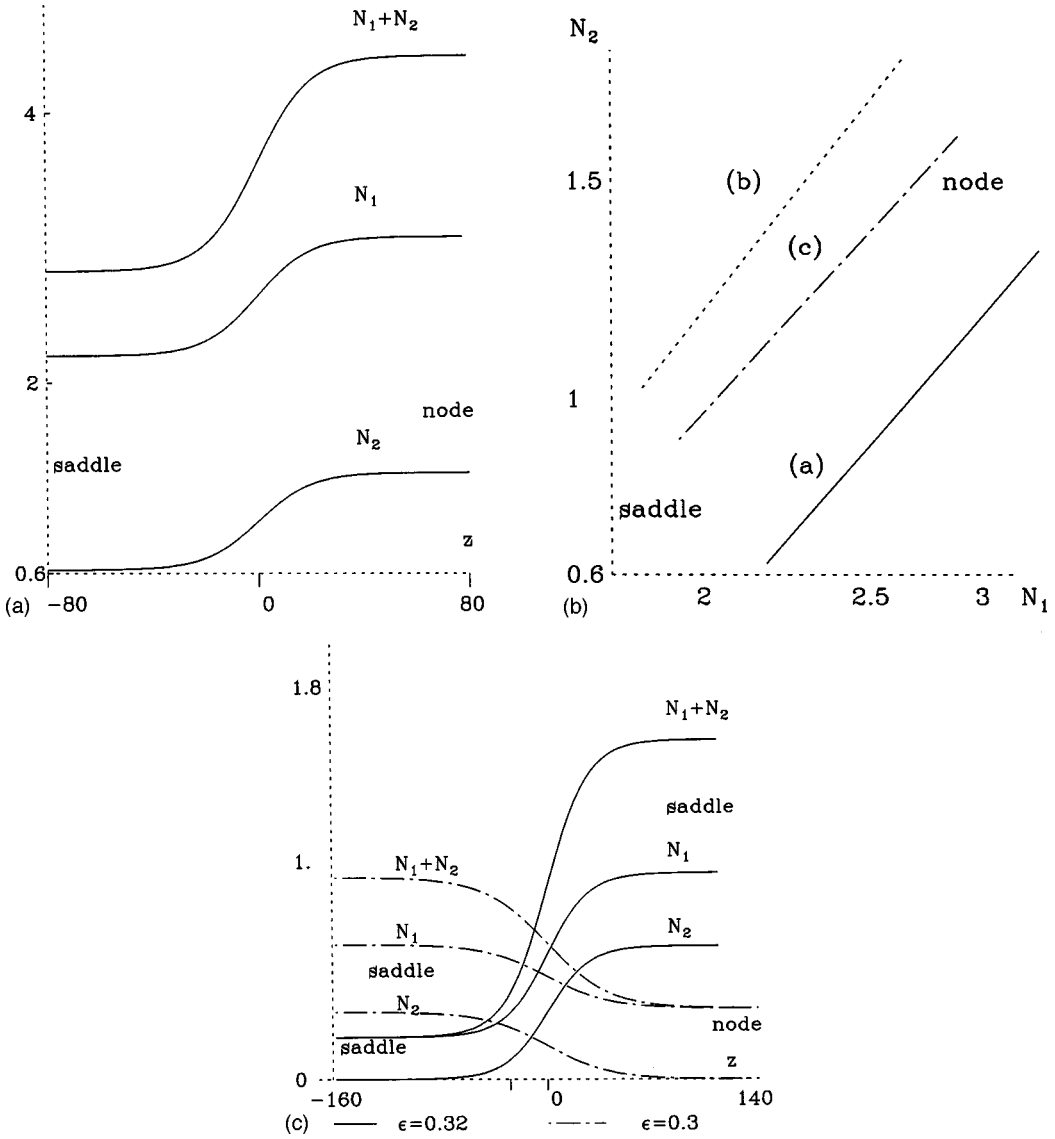

FIG. 3. Quadratic nonconservative model. Similarity solutions. Illner model $a=-0.1$, $c=1$ in (a), McKean model $a=0$, $c=1$ in (c); Illner, dual McKean $a=-1$, $c=0$ plus Carleman $c=1=-a$ models in (b). In (c) when $\epsilon \in (0.3,0.32)$ we observe a transition node, saddle becoming saddle, saddle.

TABLE III. $(1+1)$ solutions: a) dual McKean, b) McKean, $\kappa=1$

(a) $c=\alpha+\beta=0$, $\kappa=1$

$N_1=n_{01}$, $N_2=n_{02}+n_2/\Delta$, $\Delta=1+\Sigma\nu_i$, $\nu_i=d_i e^{\gamma_i(x-\xi_i t)}$,
$2an_{01}^2-n_{01}\epsilon+S_1=0\rightarrow n_{01}$,
$(\alpha+a)\rho_0^2+(\eta-\epsilon)\rho_0+\sigma \quad =0\rightarrow\rho_0=n_{02}+n_{01}\rightarrow n_{02}$
$(\alpha+a)(n_2+2\rho_0)+\eta-\epsilon=0\rightarrow n_2$,
$\gamma_i(1+\xi_i)+(\alpha+a)n_2 \quad =0\rightarrow\gamma_i$
with arbitrary number of ν_i terms and arbitrary ξ_i

(b) $a=0=-c+\beta=0$, $\kappa=0$

$(N_1,N_2,n_{01},n_{02},n_2,a,\xi_i,\gamma_i,S_1)$
$\rightarrow(N_2,N_1,n_{02},n_{01},n_1, \quad -c,-\xi_i,-\gamma_i,S_2)$
$\rho=\rho_0+\rho_1/\Delta$ with $\rho_1=n_1(n_2)$ for the McKean (dual) models.

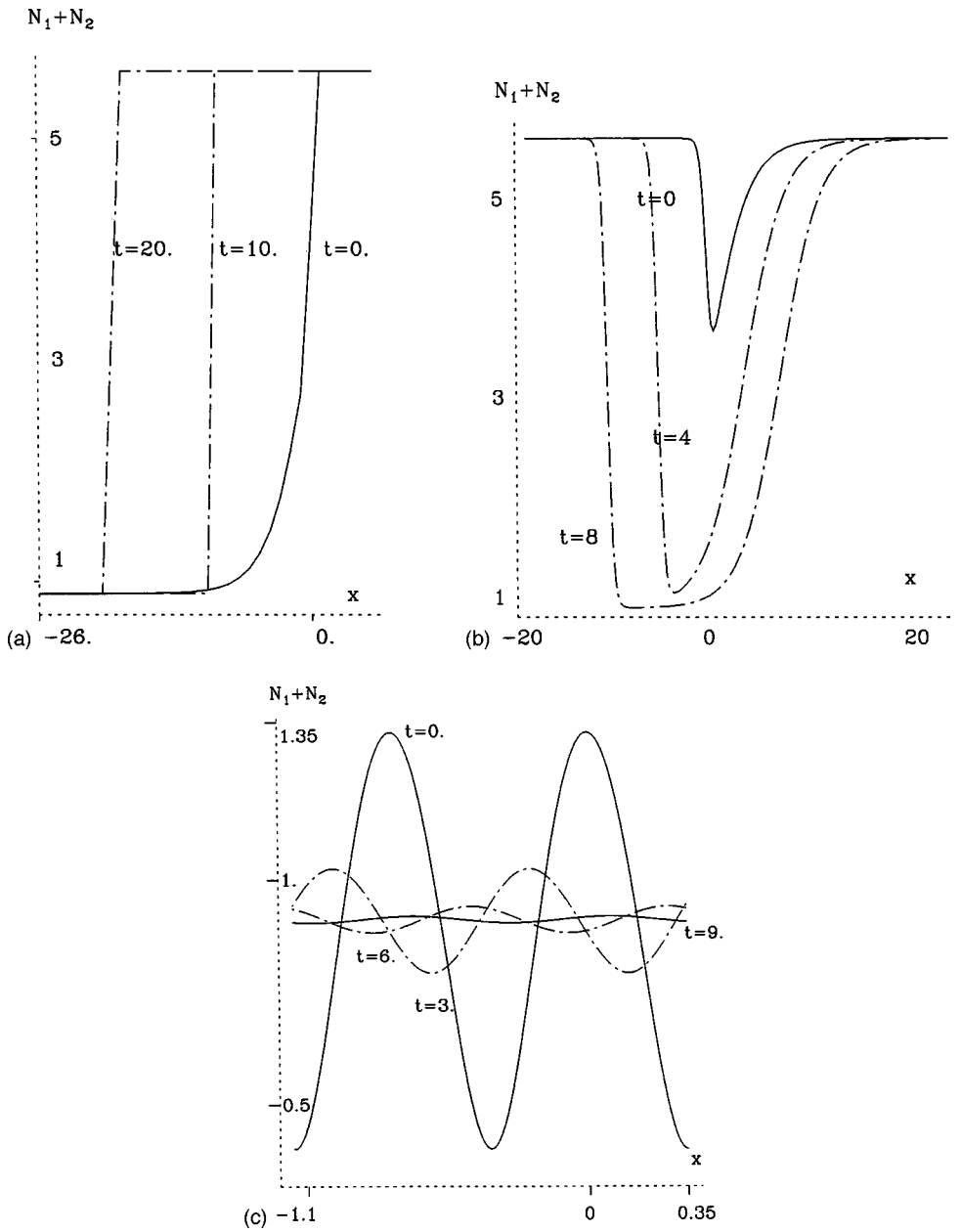

FIG. 4. Quadratic nonconservative model. $(1+1)$-dimensional solutions for the dual McKean model $a=-1$, $c=0$. In (a) we observe two asymptotic states and only one in (b). In (c) we observe a damped periodic solution.

We substitute into (1.4), take into account the asymptotic $n_{i\mathrm{as}}$ states and get

$$(\partial_t+(-1)^{i+1}\partial_x)n_iN=C_iN(N-1)\rightarrow[(\partial_t+\partial_x)n_1C_2-(\partial_t-\partial_x)n_2C_1]N=0, \tag{3.9}$$

with C_i coefficients of N^2, $-N$ written down in Appendix B3.3 and two possibilities for $(1+1)$ solutions; either $n_1=C_1=0$ leading to (3.7a) or $n_2=C_2=0$ leading to (3.7b). We are reduced to only one NLPDE for $N(x,t)$ or one linear PDE for $\Delta(x,t)=1/N$,

$$\text{dual McKean: } (\partial_t-\partial_x)N=(\alpha+a)n_2N(N-1),\ (\partial_t-\partial_x)\Delta(x,t)=(\alpha+a)n_2(\Delta-1), \tag{3.10}$$

$$\text{McKean: } (\partial_t+\partial_x)N=(\alpha-c)n_1N(N-1),\ (\partial_t+\partial_x)\Delta(x,t)=(\alpha-c)n_1(\Delta-1), \tag{3.10$'$}$$

that we integrate easily for $\Delta=1+\Sigma\,\nu_i$ (see Table III).

For the periodic solutions (3.8) and Table III with only two ν_i complex conjugate,

$$1/N(x,t)=\Delta=1+\nu+\nu^*,\ \ \nu=de^{rt+i\gamma_I x},\ \ \tau=\tau_R+i\tau_I,\ \ n_{0i},n,d\ \ \text{real}, \tag{3.11}$$

only γ_i, ξ_i are changed. Extensions with $\nu_{i+1}=\nu_i^*$, $i=1,3,...$ exist,

$$\text{dual McKean:}\ \ \gamma_I=\tau_I,\ \ \tau_R=(\alpha+a)n_2,\ \ \Delta=1+2de^{\tau_R t}\cos\ \gamma_I(x+t), \tag{3.12}$$

$$\text{McKean:}\ \ \gamma_I=-\tau_I,\ \ \tau_R=(\alpha-c)n_1,\ \ \Delta=1+2de^{\tau_R t}\cos\ \gamma_I(x-t), \tag{3.12'}$$

γ_I and d are arbitrary in both cases. The oscillations are propagating with time and if τ_R is negative we have a damping factor when $t\to\infty$ leading to the asymptotic states ρ_0+n_2, $n_{02}+n_2$ for ρ, N_2 (dual McKean) and ρ_0+n_1, $n_{01}+n_1$ for ρ, N_1 (McKean).

Numerical $(1+1)$*-dimensional solutions* with, for brevity, only $\rho(x,t)$, two ν_i, and the dual McKean model. In Figs. 4(a)–4(c) we arbitrarily choose the same values for the model parameters from which we deduce the asymptotic states,

$$\begin{gathered}c=0,\ \ a=-1=-\beta,\ \ \alpha=1.1,\ \ \eta=0.006,\ \ \epsilon=0.7,\ \ \sigma=S_1=0.5,\ \ \zeta=S_2=0,\\ \rho_0=5.61,\ \ \rho_0+\rho_1=0.892,\ \ n_{02}=5.25,\ \ n_2=-4.71,\ \ N_1=n_{01}=0.355.\end{gathered} \tag{3.13}$$

In Fig. 4(a) we present a solution vs x, give the arbitrary parameters ξ_i and deduce γ_i, $d_i=1$, $\xi_1=0.9$, $\xi_2=-0.95$, $\gamma_1=0.5$, $\gamma_2=18.8$. Due to $\gamma_1\gamma_2>0$ we observe for, $|x|$ large, the two asymptotic states ρ_0, $\rho_0+\rho_1$ and roughly speaking a translation of ρ when t is increasing. Other features can occur for $\gamma_1\gamma_2<0$, in Fig. 4(b), with the only change $\xi_2=-1.25$ leading to $\gamma_2=-3.7$, we see only one asymptotic state for t fixed and $|x|$ large. In Fig. 4(c) for $\rho>0$ vs x we present a periodic solution with $n_2/(1+2d)\leqslant\rho(x,0)-\rho_0\leqslant n_2/(1-2d)$ and, from (3.12), $\tau_R=-0.47<0$. For the arbitrary parameters we choose $d=0.05$, $\tau_I=\gamma_I=9$. The oscillations are propagating when t increases and, due to the damping factor, the asymptotic ($t\to\infty$) limit is a constant $\rho_0+\rho_1$.

IV. CONCLUSION

Comparing both conservative and nonconservative Illner, Carleman, and McKean discrete models, we see that the classes of exact solutions are different. Instead of solutions coming from only one Riccati scalar equation, here we also construct solutions compatible with two scalar Riccati equations. For the linear nonconservative models (and only for stationary solutions) we find solutions built-up with elementary functions [Figs. 1(c) and 2(a)–2(e)] or with modified Bessel functions [Figs. 1(a)–1(b)] for which we add boundary conditions. For the quadratic models, the similarity solutions found [Figs. 3(a)–3(c)] still result from a compatibility between two scalar Riccati equations and they are built-up with elementary exponential terms. The $(1+1)$-dimensional solutions are very different in conservative and nonconservative models. The sums of similarity solutions are not possible for the nonconservative present models. The $(1+1)$ solutions [Figs. 4(a)–4(c)] correspond to one of the densities being constant and, due to the fact that the problem is reduced to a linear partial differential equation, they can be sums of an arbitrary number of exponential similarity variables. Such solutions are not possible for the present associated conservative models.

The present study has been extended to the nonconservative Broadwel models, which means compatibility between three Riccati equations. For nonconservative models it will be useful to study exact solutions satisfying boundary conditions[1] other than the simplest one considered here for only one class of solutions, with modified Bessel functions. Very often in the literature it is written that nonelementary analytical solutions including elliptic... functions exist, but without proving that they do or do not satisfy (which is not always trivial[11]) physical properties like positivity and boundedness.

ACKNOWLEDGMENT

I thank T. Platkowski who suggested that I study the exact solutions for the nonconservative models.

APPENDIX A: LINEAR NONCONSERVATIVE MODELS

A1: Bureau[8] transform for stationary solutions. We recall $\rho_0=S/(\epsilon-\eta)$, $(\prime=\partial_x)$ and with four conditions reduce (2.2′) to a linear $J'+J^2$ equation,

$$W=J'+J^2,\ \ W'=CW+G(j_0)/j_1,\ \ C=2(\eta-\epsilon)(c+a)/(c-a),\ \ j_1=2/(c-a),$$

$$j_0=[C+\eta(1-2\kappa)]/(a-c)=[\rho_0(c-a)S+\epsilon]/2(a+c),\rightarrow(2.6)\ \text{relation}, \tag{A1}$$

$$j=j_0+j_1J(x),\ \ \rho(x)=\rho_0+j_1J'(x)/(\eta-\epsilon),\ \ 2N_i(x)=\rho+(-1)^{i+1}j. \tag{A2}$$

A1.1: $G(j_0)=0$, $C\neq0$, $W\neq0$. We get $W=de^{Cx}$, assume $4d=C^2$ and the Riccati equation $J'+J^2=C^2/4e^{Cx}$ gives the modified Bessel $I_0(z)$ for $z=e^{Cx/2}\in(0,\infty)$,

$$J=\partial_x\widetilde{J}/\widetilde{J},\rightarrow\partial_{z^2}\widetilde{J}+z^{-1}\partial_z\widetilde{J}=\widetilde{J}\rightarrow\widetilde{J}=I_0(z),\ \ J(x)=\partial_x\log I_0(e^{Cx/2}),$$

$$f:=[1-(y/z)^2]^{-1/2},\ \ J(x)\propto\int_{-z}^{z}ye^{-y}fdy\Big/\int_{-z}^{z}e^{-y}fdy\rightarrow\infty\ \text{ when }\ z\ \text{ or }\ Cx\rightarrow\infty.$$

We add a specular reflection boundary condition $N_1(0)=N_2(0)$, $j(0)=0$. Notice that $J(0)/C=\Sigma ma_m/\Sigma a_m$ with $a_m=(!m)^{-2}4^{-m}$ is a known number. From a, c, η, κ, $J(0)$ given we deduce $\epsilon=\eta+\eta(2\kappa-1)(c-a)/[2(c-a)(1-2J(0)/C)]$ and C, j_0 from (A1). We assume S_1 arbitrary and from the (2.6) condition deduce S and S_2. Finally from j_0 in both (A1), (2.3) we get a compatibility condition from which a is fixed while $c,\eta\kappa,S_1$ are still arbitrary but j_1,ρ_0 are known. [Notice that if $G(j_0)\neq0$, $C\neq0$, in (2.3), we define $G(j_0)/j_1C=:(C\nu/2)^2$ and get $J(x)=\partial_x\log I_\nu(x)$ still not bounded when $z\rightarrow\infty$.] In the sequel we assume $G(j_0)=0$ but either $C=0$, $W=0,\neq0$ or $W=0$, $C\neq0$.

A1.2:[9] $G(j_0)=C=a+c=0$, $W=\beta$. (A1) gives $j_1c=1$ but $\rho_0=-\epsilon/2c<0$. First $\beta=W=0$, $J'=-1/(x+d)^2$ substituted in (A2) gives $\rho(x)\rightarrow\rho_0<0$ when $|x|\rightarrow\infty$. Second, for $\beta\neq0$ we get two cases: either $J'/\beta=1-[1-\alpha e^{-2x\sqrt{\beta}}/(1+\alpha e^{-2x\sqrt{\beta}})]^2$ with $\beta>0$ and $\rho(x)\rightarrow\rho_0<0$ when $|x|\rightarrow\infty$ or $\beta<0$ and $J'/\beta=1+[(\alpha-tgx\sqrt{-\beta})/(1+\alpha tgx\sqrt{-\beta})]^2$, $\rho(x)\rightarrow\infty$ for an infinite number of $|x|\in(0,\infty)$ values.

A1.3:[9] $G(j_0)=W=0$, $J=1/(x+d)$ with (A1)–(A2) and simple models $\eta=\kappa=0$, $\epsilon\neq0$: McKean $c=1$, $a=0$: $j_1=2$, $j_0=2\epsilon$, $\rho_0=3\epsilon$, $S_i=(13-6i)\epsilon^2/2$

$$4\epsilon(x+d)^2N_i(x)=(17-8i)[\epsilon(x+d)]^2+[\epsilon(x+d)+(-2)^{i+1}]^2\geqslant0,\ \ i=1,2; \tag{A3}$$

dual McKean $a=-1$, $c=0$: $j_1=2$, $j_0=-2\epsilon$, $\rho_0=3\epsilon$, $S_i=(5-6i)\epsilon^2/2$,

$$4\epsilon(x+d)^2N_i(x)=(8-7i)[\epsilon(x+d)]^2+[\epsilon(x+d)+(-2)^{i+1}]^2>0,\ \ i=1,2. \tag{A3′}$$

A2: Transform associated with $G(j_0)=G(j_0+j_1)=0$ and (2.2′)–(2.4). With $\gamma:=j_1(a-c)/2$ we add $J'\gamma$ in both sides and with $W=J'-\gamma J(J-1)$ get $W'=\overline{C}W$ or $W=\overline{d}e^{\overline{C}x}$

$$[J'-J(J-1)\gamma]'=J'[(j_0+j_1/2)(a-c)+\eta(2\kappa-1)]+j_1(a+c)(\eta-\epsilon)J(J-1). \tag{A4}$$

We get $\overline{C}=\eta(2\kappa-1)+(a-c)\epsilon/2(a+c)+S(a-c)^2/2(a+c)(\eta-\epsilon)$ from the rhs $\overline{C}W$ and a compatibility condition which is (2.6). For $W=0$ we get $J(1+e^{\gamma x})=1$. For $W\neq0$ we define $z=e^{\overline{C}x/2}$, $\nu=\gamma\overline{C}/4$, assume $\overline{d}\gamma4/\overline{C}^2=1$. The Riccati equation $J'-\gamma J(J-1)=\overline{d}e^{\overline{C}x}$ still leads to modified Bessel functions, $\gamma J=\partial_z\log[z^{-(\gamma\overline{C})/2}I_\nu(z)]$. In the sequel we consider the II.2, $W=0$ solution (see Table I).

A2.1: Linear stability of the x-dependent stationary solutions $\rho(x,t)=\rho(x)+e^{lt}X_\rho$, $j(x,t)=j(x)+e^{lt}X_j$. We give the characteristic l_i, $i=1,2$,

$$l_1=\eta-\epsilon,\ \ l_2=2j(x)(a+c)+\rho(x)(a-c)-\epsilon=N_1(3a+c)-N_2(3c+a)-\epsilon. \tag{A5}$$

We study the sign of l_2. At the two asymptotic states (see Table I) we get opposite values,

$$l_{2\text{as}}=2j_{\text{as}}(a+c)+\rho_{\text{as}}(a-c)-\epsilon=\pm 2(a+c)\sqrt{\delta},$$

$$w=de^{\gamma x}\to l_2=0 \ \text{ if } \ w^2-1\pm w(c-a)^2\sqrt{\delta}/(\epsilon-\eta)(c+a)=0. \tag{A6}$$

The two w roots have opposite sign ($\pm$ is fixed for a solution), one is associated to $d>0$ and the other to $d<0$ but independent of the d and x values except in the product $de^{\gamma x}$.

A2.2: Positivity for the $-0.1\leqslant a/c\leqslant 0$ models at the asymptotic stationary values $|x|=\infty$. For $\epsilon\neq 0,\ \eta=\kappa=0$, we rewrite (see Table I) $S,\ \rho_0,\ j_{\text{as}},\ j_0,\ \delta$,

$$S=(3c+a)(3a+c)\epsilon^2/(c-a)^3>0 \ \text{ if } \ -1/3<a/c\leqslant 0,\ \ \rho_0=S/\epsilon>0,$$

$$j_{\text{as}}^2-j_{\text{as}}4\epsilon(c+a)/(c-a)^2+(S_1+S_2)/(c+a)=0,\ \ j_0=2\epsilon(a+c)/(c-a)^2\pm\sqrt{\delta}, \tag{A7}$$

$$\text{sign } \delta+2S_2/(c+a)=[7a^3+19a^2c+5ac^2+c^3]>0,$$

with a constraint on $S_2>0$ small such that $\delta>0$. $j_0>0$ is one of the two $j_{\text{as}}>0$ roots. For N_i or $2n_{i,\text{as}}=\rho_0-(-1)^i j_{\text{as}}>0$ for $i=1$ we get positivity conditions for $i=2$,

$$n_{2,\text{as}}^2-n_{2,\text{as}}\epsilon(5a^2+10ac+c^2)/(c-a)^3+(S_2+a(S/\epsilon)^2)/2(c+a)=0,$$
$$n_{2,\text{as}}>0 \ \text{ if } \ S_2+a(S/\epsilon)^2>0 \ \text{ and } \ -1+2/\sqrt{5}\simeq -0.1<a/c\leqslant 0. \tag{A8}$$

APPENDIX B: QUADRATIC NONCONSERVATIVE MODELS

B1: Similarity solutions functions of $z=x-\xi t$. First we substitute $\rho=\rho_0+R(z)/D(z)$, $w=w_0+W(z)/D(z)$ into (3.1)–(3.2), find $\Sigma A_k(z)/D(z)^k$, assume $A_2(z)=0$, define $c_R=R/D'$, $c_W=W/D'$ and get a cubic polynomial with $c_R=$ const.,

$$c_W=c_R[1-\xi+(\alpha-\beta)c_R],$$

$$(1+\xi)c_W+2(a-\alpha\kappa)c_R^2=(3a+c-\beta)c_Wc_R-(c+a)c_W^2. \tag{B1}$$

Second we start with $\rho=\rho_0+\rho_1 Y,\ w=w_0+w_1 Y,\ Y(z)$ unknown while $\rho_0,\rho_0+\rho_1$ and w_0, w_0+w_1 are the asymptotic states (3.3)–(3.4). From (3.1) to (3.2) we get two Riccati equations for $Y(z)$ or two first-order linear differential equations for $\Delta(z)=1/Y$ and X in Table II,

$$Y(1-Y)=Y'(\overline{w}_1-1+\xi)/(\alpha-\beta)\rho_1=Y'(1+\xi)/\rho_1 X,\to\Delta=1+de^{\gamma z}. \tag{B2}$$

B2: Linear stability of the $\kappa=0$ similarity solutions $\rho(z)=\rho_{\text{as}}(1+e^{lz}X_\rho),\ w(z)=w_{\text{as}}(1+e^{lz}X_w)$. We give the two characteristic values $l=l_1,l_2$,

$$l^2A_2+lA_1+A_0=0,\ \ \xi^2<1\to A_2=1-\xi^2>0. \tag{B3}$$

For the Illner model $a<0,\ c>0,\ a+c>0$ we first rewrite the asymptotic states,

$$\rho_{\text{as}}2(\alpha-\beta)+\eta-\epsilon=-\zeta\sqrt{\Delta_\rho},\ \ \zeta=\pm 1,$$

$$2(c+a)w_{\text{as}}=\rho_{\text{as}}(3a+c-\beta)-\epsilon+\zeta_2\sqrt{\Delta_w}>0,\ \ \zeta_2=1,\to\zeta A_0=\sqrt{\Delta_\rho\Delta_w}>0,$$

$$A_1=(1-\xi)\sqrt{\Delta_w}+(1+\xi)\zeta\sqrt{\Delta_\rho}-4a\rho_{\text{as}}+w_{\text{as}}(3a+c-\beta), \tag{B4}$$

sign $l_1l_2=\zeta$ and we have at one asymptotic state a node and a saddle at the other.

For the Carleman model $a+c=0$ we get $\zeta A_0=\sqrt{\Delta_\rho}2c\rho_{\text{as}}^2/w_{\text{as}}>0$ and

$$A_1=(1+\xi)\zeta\sqrt{\Delta_\rho}+2\rho_{\text{as}}c(1-\xi)\rho_{\text{as}}/w_{\text{as}}+2c\rho_{\text{as}}^2/w_{\text{as}}+\epsilon>0 \ \text{ for } \ \zeta=1.$$

At the $\zeta=1$ asymptotic state we have an attractive node and a saddle at $\zeta=-1$.

For the McKean model $a=0$ we get $w_{as}\neq 0$, $\zeta A_0=\sqrt{\Delta_\rho}w_{as}>0$ and $w_{as}=0$, $\zeta A_0=\sqrt{\Delta_\rho}(\rho_{as}(\beta-c)+\epsilon)$. For $w_{as}\neq 0$ we have a node or a saddle depending whether $\zeta=1$ or -1 whereas for $w_{as}=0$, independently of ζ we can have saddle or node.

For the dual Mc Kean model $c=0$ we get $A_0=\zeta\zeta_2\sqrt{\Delta_\rho\Delta_w}$. At each state we can have either a node or a saddle depending whether $\zeta\zeta_2=\pm$.

B3: Possible $(1+1)$-dimensional exact solutions.

B3.1: Constraints coming from (3.1). First, we start with the sum,

$$\rho(x,t)=\rho_0+\Sigma\rho_i/\Delta_i,\ \ w(x,t)=w_0+\Sigma w_i/\Delta_i,\ \ \Delta_i=1+\nu_i,\ \ \nu_i=d_ie^{\gamma_i(x-\xi_it)} \tag{B5}$$

of two similarity solutions which, substituted into (3.1), gives for $1/\Delta_1\Delta_2$ the condition $\rho_1\rho_2=0$ which is not possible. Second we substitute into (3.1) another ansatz,

$$\rho=\rho_0+(\rho_{00}+\Sigma\rho_i\nu_i)/\Delta,\ \ w=\ldots,\ \ \Delta=1+\Sigma\nu_i+\mu\nu_1\nu_2,\ \ \nu_i=d_ie^{\gamma_i(x-\xi_it)}, \tag{B6}$$

with $\mu\neq 1$ because $\mu=1$ corresponds to (B5) and ρ_0 one ρ_{as} (3.3). We get,

$$\Delta^{-2}[a_0+\Sigma\nu_ia_i+b_i\nu_i^2+c_j\nu_i^2\nu_j+a_{12}\nu_1\nu_2]\equiv 0,\ \ a_0=a_i=b_i=\cdots=0. \tag{B7}$$

From a_0, b_i, a_i, a_{12} we get $\rho_{00}=\rho_1=\rho_2$, $w_{00}=w_1=w_2$, $\mu\rho_{00}^2=0$, and $\mu=0$. Third we restrict to $\mu=\rho_{00}=0$ in (B6) and substitute into (3.1),

$$\rho=\rho_0+(\Sigma\rho_i\nu_i)/\Delta,\ \ w=w_0+(\Sigma w_i\nu_i)/\Delta,\ \ \Delta=1+\Sigma\nu_i. \tag{B8}$$

We have terms of the (B7) type and $a_0=c_j=0$, give that $\rho_0+\rho_i=\rho_{as}$, $i=1,2$. Two possibilities follow. (i) $\rho_1=\rho_2$, we deduce $(w_1-w_2)(\gamma_1-\gamma_2)=0$ and either $w_1=w_2$ with (B8) rewritten: $\rho=\rho_0+\rho_1-\rho_1/\Delta$ (similarly for w) or $\gamma_1=\gamma_2$ but (3.2) still leads to $w_1=w_2$. (ii) $\rho_1\neq 0$, $\rho_2=0$, deduce $w_1\neq 0$, $w_2=0$, with (B8) rewritten $\rho=\rho_0+\rho_1/\bar{\Delta}$, $\bar{\Delta}=1+\nu_1^{-1}(1+\nu_2)$. In conclusion, starting with (B7)–(B8), (3.1) leads to $(1+1)$ of the type,

$$\begin{gathered}\rho=\rho_0+\rho_1/\Delta,\ \ w=w_0+w_1/\Delta,\ \ \Delta=1+\Sigma\nu_i,\ \ \rho_{as}=\rho_0,\rho_0+\rho_1,\\ \text{def: }\ \bar{w}_1=w_1/\rho_1,\ \ \gamma_i(1-\xi_i-\bar{w}_1)=\rho_1(\alpha-\beta),\ \ i=1,2.\end{gathered} \tag{B9}$$

B3.2: Constraints coming from Eq. (3.2) with the Illner term. We substitute (B9) into (3.2) with the w_{as} relation (3.4) satisfied by w_0 and get the identity, $\Delta^{-2}[a_0+\Sigma\nu_ia_i]\equiv 0$. $a_0=0$ gives the other $w_{as}=w_0+w_1$, while $a_i=0$ lead to

$$\gamma_i(1+\xi_i)=(c+a)w_1+2\rho_1(a-\alpha\kappa)/\bar{w}_1-\rho_1(3a+c-\beta),\ \ i=1,2. \tag{B10}$$

From (B9) to (B10) we first get $\bar{w}_1=2$, $N_1=w_0-2\rho_0\neq 0$ is a constant and $c+\alpha(1-\kappa)=0$ or $c=0$, $\kappa=1$. From the relations $\rho_{as}=\rho_0+\rho_1$, $w_{as}=w_0+2\rho_1$ we get (B11). Another compatibility for (B9)–(B10) is $w_1=0$, $N_2=w_0/2=$const. in (3.4) with $w_{as}=w_0$, $\rho_{as}=\rho_0,\rho_0+\rho_1$, the coefficients of ρ_1 are zero and we get (B11′),

$$\text{dual McKean: }\ \gamma_i(1+\xi_i)=-(\alpha+a)\rho_1,\ \ w_1=2\rho_1,\ \ c=0,\ \ \kappa=1,\ \ \beta+a=0, \tag{B11}$$

$$\text{McKean: }\ \gamma_i(1-\xi_i)=(\alpha-c)\rho_1,\ \ a=0,\ \ \kappa=0,\ \ \beta=c. \tag{B11′}$$

B3.3: In (1.4) we write $R_i(N)=C_iN(N-1)$ with $N_i=n_{0i}+n_iN$, $R_i(n_{1as},n_{2as})\equiv 0$,

$$\begin{aligned}C_i=&n_1^2((-1)^{i+1}-\beta)+(-1)^i(a+c)+\beta)n_1n_2+(-1)^{i+1}cn_2^2\\ &+(2-i+(-1)^i\kappa)(n_1+n_2)^2\alpha\\ -C_i=&2n_{01}n_1((-1)^{i+1}-\beta)+(-1)^i(a+c)(n_1n_{02}+n_2n_{01})+(-1)^{i+1}c2n_2n_{02}\\ &-n_i(\beta(n_{01}+n_{02})+\epsilon)+(2-i+(-1)^i\kappa)(n_1+n_2)(2(n_{01}+n_{02})\alpha+\eta).\end{aligned} \tag{B12}$$

First, with $n_1=C_1=0$ we get $c=\beta+\alpha=0$, $\kappa=1$ (B11) and with $n_2=C_2=0$ we get $a=\kappa=\beta-c=0$ (B11′).

[1] Survey, G. Spiga, *Lecture Notes in Math.* edited by G. Toscani *et al.* (Springer, Berlin, 1989), Vol 1460, p. 203, with 60 references therein; G. Spiga, Ĝ. Dudek, and V. C. Boffi, *Discrete Kinet. Theo., Lattice Gases, Hydrodyn*, edited by R. Monaco (Singapore, 1989), p. 315; K. Piechor, Arch. Mech. **41**, 95 (1989); R. Monaco and M. Pandolfi-Bianchi, *Adv. Kinet. Theor. Cont. Mech.*, edited by R. Gatignol *et al.* (Springer, Berlin, 1991), p. 169; T. Platkowski and G. Spiga, Eur. J. Mech. B Fluids, **11**, 3, 349 (1992); G. L. Caraffini and G. Spiga, Transp. Theory Stat. Phys. **23**, 9 (1994).
[2] T. Platkowski and G. Spiga, Physica A **210**, 155 (1994).
[3] K. Piechor and T. Platkowski RW 96-03, (1996).
[4] R. Gatignol, *Lecture Notes in Physics* (Springer, Berlin, 1975), Vol. 36; H. Cabannes, *The discrete Boltzmann equation (Theory and application) Lecture notes* (University of California, Berkeley, 1980); T. Platkowski and R. Illner, SIAM (Soc. Ind. Appl. Math.) Rev. **30**, 213 (1988); N. Bellomo and T. Gustafsson, Rev. Math. Phys. **3**, 137 (1991); C. Cercignani, Transp. Theory Stat. Phys. **23**, 1 (1994).
[5] H. Cornille, J. Math. Phys. **28**, 1567 (1987), and references therein; Survey, *Part. Integr. Eqs. Phys.* edited by R. Conte (Kluwer Academic, New York, 1989); H. Ĉornille and T. Platkowski, J. Math. Phys. **33**, 2587 (1992); Transp. Theory Stat. Phys. **23**, 75 (1994); A. V. Bobylev, B̂erichte der Arbeitsg. Technomat. (Kaiselautern) Vol. 99 (1993); H. Cabannes, Tiem, CRAS **318**, 1583 (1994); H. Cornille and A. d'Almeida, J. Math. Phys. **37**, 5476 (1996).
[6] H. Cornille, *Lecture Notes in Math. 1460*, edited by G. Toscani *et al.* (Springer, Berlin, 1989), proof, p. 79; *Rigorous Methods in Particle Physics*, edited by S. Ciulli *et al.* (Springer, Berlin, 1990), Vol. 119, 168, proof, p. 178); Lett. Math. Phys. **19**, 211 (1990).
[7] R. Illner, Math. Methods Appl. Sci. **1**, 187 (1979); M. P. McKean, J. Comb. Theory **2**, 358 (1967).
[8] F. J. Bureau, Differential equation with fixed critical points, Annali di Matematica pura ed applicata **LXIV**, 229–366 (1964), for the Riccati solution p. 269; E. L. Ince, *Ordin. Diff. Eqs.* (Dover, New York, 1956); for the Painlevé equations see Chap. XIV, 317–355.
[9] I thank the referee who suggested that I study these solutions.
[10] W. T. Reid, *Riccati Differential Equations* (Academic, New York, 1972); T. G. Bountis, V. Papageorgiou, and P. Winternitz, J. Math. Phys. **27**, 1215 (1986).
[11] H. Cornille, J. Stat. Phys. **47**, 611 (1986), with the proof that an equianharmonic Weierstrass exact solution, proposed for the Boltzmann equation, was in fact violating positivity.

Infinite moments of the partition function for random walks in a random potential

Lester N. Coyle[a)]
Department of Mathemetical Sciences, Loyola College, Baltimore, Maryland 21210

(Received 9 May 1997; accepted for publication 17 November 1997)

We consider a system of random walks (directed polymers) in a potential which is random in space and time. Investigations into untypical behavior and the inverse of the partition function leads to the result that, regardless of the strength of the potential and in all dimensions, all higher moments of the partition function are unbounded as time goes to infinity. One can interpret this result in a many-body quantum mechanical setting as saying that no matter how small the strength of an attractive potential, in any dimension one can find N (the number of particles) large enough so that there is a bound state. © *1998 American Institute of Physics.* [S0022-2488(98)01904-5]

I. INTRODUCTION

Random walks in a random potential (otherwise referred to as directed polymers in a random potential) have been used to model many physical processes. For example, the interface in two-dimensional Ising models with random exchange interactions can be modeled by a one-dimensional random walk in a random potential (see Ref. 1). While superdiffusive behavior is conjectured for this dimension $d=1$ case, no rigorous results have been obtained. For a background on this subject see Refs. 2 and 3.

Imbrie and Spencer[3] considered the following model of a random walk in a random potential. Let $v(t,x)$ be independent for each t and x, with $v(t,x)=\pm 1$ with probability 1/2 where $x \in \mathbb{Z}^d$, $t \in \mathbb{Z}^+$, and $d=$dimension. Without loss of generality we start all walks at the origin unless otherwise stated. Let $dW_T(X)$ be the probability measure for the symmetric nearest neighbor discrete time random walk (i.e., simple random walk) on $\mathbb{Z}^d$, where $X(t)$ represents the position of the walk at integer time t. For $0<\epsilon<1$, the probability measure for the random walk in a random potential is given by

$$d\mu_T = \prod_{0<t\leqslant T} [1+\epsilon v(t,X(t)]dW_T/Z(T),$$

so that

$$P(X(T)=x) = \int \delta(X(T)-x) \prod_{0<t\leqslant T} [1+\epsilon v(t,X(t)]dW_T/Z(T),$$

where δ is the Kronecker delta, and $Z(T)$ is the partition function, given by

$$Z(T) = \int \prod_{0<t\leqslant T} [1+\epsilon v(t,X(t)]dW_T.$$

The partition function is needed in order to normalize the probability density.

Note that this is a distinctly different model than the random walk in random environment studied by Kesten in Ref. 4. Also note that in Ref. 5 the authors considered random walks in a random potential where the random potential had no time dependence [i.e., the model described above with $v=v(x)$], and proved subdiffusive behavior in all dimensions. In this paper we only consider time-dependent potentials.

[a)]Electronic mail: coyle@fischer.loyola.edu

 © 1998 American Institute of Physics

Let $P_x(\cdot)$ and $E_x[\cdot]$ denote probability and expectation (respectively) with respect to the simple random walk measure dW_T for walks starting at x, $P_v(\cdot)$ and $\langle\cdot\rangle_v$ denote probability and expectation (respectively) with respect to the random potential v, and $E_x^v[\cdot]$ denote expectation with respect to the random walk in a random potential measure $d\mu_T$ for walks starting at x. Note that $\langle Z(T)\rangle_v = 1$.

A. Known rigorous results

Imbrie and Spencer[3] proved the following result:

Theorem I.1: *For small enough* $\epsilon > 0$ *and dimension* $d > 2$,

$$\lim_{T\to\infty}\left(E_0^v\left[\frac{X(T)^2}{T}\right]\right) = 1 \quad \text{almost surely.}$$

Bolthausen[6] used martingales to extend this theorem to arbitrary powers of $X(T)/\sqrt{T}$. Song and Zhou[7] showed that Bolthausen's result holds in dimensions $d \geqslant 3$ for any $\epsilon \in (0,1)$. Sinai[8] considered random potentials having a general distribution and showed, under conditions on the distribution of $v(x,t)$, that theorem I.1 holds. Conlon and Olsen[9] considered a continuous time and space version of the random walk in a random potential and showed that, under the conditions of theorem I.1, $\lim_{T\to\infty} E_0^v[f(X(T)/\sqrt{T})] = E[f(Y)]$ almost surely, where Y is a Gaussian variable and for a wide class of functions f which include polynomials and exponentials, provided the limit is taken exponentially fast. Coyle[10] considered a continuous time version of the random walk in a random potential, and proved a result similar to that of Conlon's and Olsen's without the restriction of an exponentially fast limit.

B. Why study the partition function?

No rigorous results are known for the random walk in a random potential in one or two dimensions though scaling of the form

$$\frac{\langle E_0^v[X(T)^2]\rangle_v}{T^{4/3}} \to 1$$

has been conjectured both numerically and heuristically in both dimensions $d=1$ and $d=2$ (see Refs. 1 and 11–14). In an attempt to understand the $d=1$ case, Fisher and Hwa[2] have studied rare events in the random walk in a random potential system (with $d=1$) by mapping onto Burgers's equation with random forcing. Also, numerical studies (e.g., Ref. 15) indicate the system in dimension $d=1$ is particularly sensitive to the details of the realization of the random potential. In the time-independent random potential case (see Ref. 5), the system is dominated by untypical behavior, namely paths that pass through regions with a large concentration of positive v's. We believe that something similar is occurring in dimensions 1 and 2 when the potential is time dependent, though it is not yet clear just what the crucial untypical behavior is. We believe that, as in Ref. 5, analysis of the partition function will be key in understanding the importance of untypical behavior.

We conjecture that in dimensions $d=1$ and $d=2$ the partition function goes to zero as $T \to \infty$ with probability one. Thus we desire to study

$$\lim_{T\to\infty}\left\langle \frac{1}{Z(T)}\right\rangle_v$$

in all dimensions to see whether there is a drastic change in this quantity as we go from dimension $d=3$ to dimensions $d=1$ and $d=2$. Let $N(T) = \int X(T)^2 \Pi_{0<t\leqslant T}[1+\epsilon v(t,X(t)]dW_T$. Imbrie and Spencer[3] showed that for ϵ sufficiently small in dimension $d>2$

$$\lim_{T\to\infty}\left\langle E_0^v\left[\frac{X(T)^2}{T}\right]\right\rangle_v = \lim_{T\to\infty}\left\langle \frac{N(T)}{TZ(T)}\right\rangle_v = 1.$$

We know that $N(T)/(TZ(T))\to 1$ with probability 1 (in v) and it is easy to see that $\langle N(T)/T\rangle_v = 1$ and $\langle Z(T)\rangle_v = 1$. It would appear that it should be no more difficult to find $\lim_{T\to\infty}\langle 1/Z(T)\rangle_v$. Suppose $\lim_{T\to\infty}\langle 1/Z(T)\rangle_v = C<\infty$. Then for $0<\delta\ll 1$, Chebyshev gives

$$\lim_{T\to\infty} P_v(Z(T)<\delta) = \lim_{T\to\infty} P_v\left(\frac{1}{Z(T)}>\frac{1}{\delta}\right) \leqslant \lim_{T\to\infty} \delta\left\langle \frac{1}{Z(T)}\right\rangle_v = \delta C.$$

However, the best that Imbrie's and Spencer's calculations appear to yield (see the Appendix) is

$$P_v(Z(T)<\delta)<\frac{c}{|\log\ \delta|}.$$

The best upper bounds achieved in Refs. 9 and 10 were of the form $P_v(Z(T)<\delta)\leqslant C|\log(\delta)|^{-\alpha}$ for some $\alpha>0$. It certainly appears that some information is lacking regarding the partition function in all dimensions. If one were to do a Taylor expansion of $\langle 1/Z(T)\rangle_v$ one would need to consider the higher moments of the partition function, and we have proven the following result.

C. Main result

Theorem I.2: *For any ϵ satisfying $0<\epsilon<1$, and in any dimension d, there exists $n_0\in\mathbb{Z}^+$ such that for all $n\geqslant n_0$*

$$\lim_{T\to\infty}\langle Z(T)^n\rangle_v = \infty.$$

In particular, $n_0=2$ in dimensions 1 and 2, and $n_0=C/\epsilon^d$ for $d>2$.

This result verifies that the analysis of Imbrie and Spencer cannot be extended to dimensions 1 and 2, since finiteness of $\langle Z(T)^2\rangle_v$ was essential in their calculations. In the dimension $d>2$ regime, where diffusive behavior has been proven, we have the result that the partition function is very large with a sizable (nonzero) probability. This shows that the random potential really does effect the system in the case of $d>2$ in a manner which is not detected in results such as theorem I.1.

D. Connection to many-body quantum mechanics

Letting $\epsilon' = \tanh(\epsilon)$, we have

$$\prod_{0<t\leqslant T}[1+\epsilon' v(t,X(t)] = (\cosh(\epsilon))^{-T} e^{c\Sigma_{t=1}^{T} v(t,X(t))}.$$

So up to normalization factors, the analysis writing $e^{\epsilon v}$ rather than $1+\epsilon v$ will be the same. With this in mind, we can make the following connection to quantum mechanics.

We have $\langle v(t_1,x_1)v(t_2,x_2)\rangle_v = \delta(t_1-t_2)\delta(x_1-x_2)$. Let $V(x_1,x_2)=\delta(x_1-x_2)$. Going to the continuum system, there exists the following relationship (given by the Hubbard–Stratanovich transformation, see Ref. 16):

$$\langle Z(T)^n\rangle_v = \int dx_1\cdots dx_n e^{-TH_n}(0,...,0|x_1,\ldots,x_n),$$

where $x_k\in\mathbb{R}^d \forall k$ and H_n is the n-body quantum mechanical Hamiltonian with attractive potential $V(x)$. For $n=2$, it is known (see Ref. 17) that if $V(x)$ is a short-range *attractive* potential, then in

dimensions 1 and 2 H_2 has a bound state, and this would indicate that in dimensions 1 and 2, $\lim_{T\to\infty}\langle Z(T)^2\rangle_v=\infty$, which is true. On the other hand theorem I.2 indicates that if $n>C/(\epsilon^d)$, where ϵ is the strength of the potential, then H_n has a bound state in dimension d. Thus we can think of theorem I.2 as being a consequence of the lack of stability of the corresponding quantum mechanical system.

II. PROOF OF INFINITE MOMENTS

To prove theorem I.2 we will use the local central limit theorem (see e.g., Ref. 18); throughout this paper we shall refer to it as the local CLT.

For the simple random walk, one version of the local CLT is

$$P_0(X(n)=x)=2\left(\frac{d}{2\pi n}\right)^{d/2}e^{-d|x|^{2/(2n)}}+O(n^{-(d+2)/2}).$$

We also need:

Proposition II.1: Let $C_M=(-M,M)^d\cap\mathbb{Z}^d$, *for* $M\in\mathbb{R}$, *and let* $N\in\mathbb{Z}^+$. *There exists positive constants C and c such that* $\forall T\in\mathbb{Z}^+$,

$$P_0(X(t)\in C_N,0\leqslant t\leqslant T)\geqslant Ce^{-cT/N^2}. \tag{1}$$

Also, for N sufficiently large, $N^2\leqslant t\leqslant T$, *and* $x\in C_{N/2}$, *there exists constants* K_1 *and* K_2 *such that*

$$\frac{K_1}{N^d}\leqslant P_0(X(t)=x|\text{stay in } C_N \text{ up to time } T)\leqslant\frac{K_2}{N^d}. \tag{2}$$

Note that $P(\cdot|A)$ denotes probability conditioned on the set of events A. We prove proposition II.1 in Sec. III. Intuitively, Eq. (1) makes sense since we know that simple random walks on average travel a distance $\sqrt{T}$ in time T, so if $T<N^2$, there is a good chance of staying in the cube, and if $T>N^2$, it is unlikely that the walk will stay in the cube. Turning our attention to Eq. (2), it is reasonable that given the walk stays in C_N, after N^2 steps there should be no particularly favored sites in the sub-cube $C_{N/2}$.

Note that we can rewrite the partition function as

$$Z(T)=\int\prod_{0<t\leqslant T}\prod_{x\in\mathbb{Z}^d}[1+\epsilon v(t,X(t))\delta(X(t)-x)]dW_T.$$

We first consider the simpler case of dimension $d\leqslant 2$,

$$\begin{aligned}Z(T)^2=&\int\int\prod_{0<s\leqslant T}\prod_{x\in\mathbb{Z}^d}\{[1+\epsilon v(s,X_1(s))\delta(X_1(s)-x)]\\&\times[1+\epsilon v(s,X_2(s))\delta(X_2(s)-x)]\}d\mu_T(X_1)d\mu_T(X_2)\\\Rightarrow\langle Z(T)^2\rangle_v=&\int\int\prod_{0<s\leqslant T}\prod_x[1+\epsilon^2\delta(X_1(s)-x)\delta(X_2(s)-x)]d\mu_T(X_1)d\mu_T(X_2).\end{aligned} \tag{3}$$

Expanding the products over s and x and fixing $x_0=0$, we have

$$\begin{aligned}\langle Z(T)^2\rangle_v=&\sum_{n=1}^{T}\sum_{0=s_0\leqslant s_1\leqslant\cdots\leqslant s_n\leqslant T}\epsilon^{2n}\prod_{i=1}^{n}\left[\sum_{x_i\in\mathbb{Z}^d}P_0(X(s_i-s_{i-1})=x_i-x_{i-1})^2\right]\\\geqslant&\sum_{n=1}^{T}\sum_{s_0\leqslant\leqslant\cdots\leqslant s_n}\epsilon^{2n}\prod_{i=1}^{n}\left[\sum_{\{x_i:|x_i-x_{i-1}|\leqslant\sqrt{s_i-s_i-1}\}}P_0(X(s_i-s_{i-1})=x_i-x_{i-1})^2\right].\end{aligned}$$

By the local CLT $\exists M>0$ and $C_1>0$ such that if $|s_i-s_{i-1}|>M$ and $|x_i-x_{i-1}| \leqslant \sqrt{s_i-s_{i-1}}$, then

$$P_0(X(s_i-s_{i-1})=x_i-x_{i-1}) \geqslant \frac{C_1}{(s_i-s_{i-1})^{d/2}}.$$

Also $\exists C_M>0$ such that if $|s_i-s_{i-1}| \leqslant M$ and $|x_i-x_{i-1}| \leqslant \sqrt{s_i-s_{i-1}}$, then $P_0(X(s_i-s_{i-1})=x_i-x_{i-1}) \geqslant C_M$. Thus

$$\sum_{\{x_i:|x_i-x_{i-1}| \leqslant \sqrt{s_i-s_{i-1}}\}} [P_0(X(S_i-S_{i-1})=x_i-x_{i-1})]^2 \geqslant \begin{cases} \dfrac{C_1^2}{(s_i-s_{i-1})^{d/2}} & \text{if } |s_i-s_{i-1}|>M \\ C_M^2 c M^{d/2}=C_2 & \text{if } |s_i-s_{i-1}| \leqslant M. \end{cases}$$

Thus $\langle Z(T)^2 \rangle_v$ diverges as $T \to \infty$ for $d \leqslant 2$. For $n>2$, it is easy to see that $\langle Z(T)^n \rangle_v$ is greater than the right-hand side of Eq. (3), and the result follows.

Now we turn our attention to dimension $d>2$. The idea of the proof goes as follows. Whenever two of the n walks intersect, this increases $\langle Z(T)^n \rangle_v$ by a factor e^{ϵ} (approximately); however, the probability of intersections is exponentially small. We constrain all walks to remain in the cube C_N, so that there is a strong likelihood of intersections. We choose n and N so that the cost of staying in the cube is outweighed by the benefit of more intersections. Letting X_k denote the walk indexed by k, we have

$$\begin{aligned} \langle Z(T)^n \rangle_v &= E_0\left[\prod_{0 \leqslant s \leqslant T} \prod_{x \in \mathbb{Z}^d} \left\{ 1+\epsilon^2 \sum_{1 \leqslant i<j \leqslant n} \delta(X_i(s)-x)\,\delta(X_j(s)-x)+\cdots \right\} \right] \\ &\geqslant E_0\left[\prod_{0 \leqslant s \leqslant T} \prod_{x \in \mathbb{Z}^d} \left\{ 1+\epsilon^2 \sum_{1 \leqslant i<j \leqslant n} \delta(X_i(s)-x)\,\delta(X_j(s)-x) \right\} \right] \\ &= E_0\left[\exp\left\{ \sum_{0 \leqslant s \leqslant T, x \in \mathbb{Z}^d} \log\left(1+\epsilon^2 \sum_{1 \leqslant i<j \leqslant n} \delta(X_i(s)-x)\,\delta(X_j(s)-x) \right) \right\} \right] \\ &\geqslant E_0\left[\exp\left\{ \sum_{0 \leqslant s \leqslant T, x \in C_N} \log\left(1+\epsilon^2 \sum_{1 \leqslant i<j \leqslant n} \delta(X_i(s)-x)\,\delta(X_j(s)-x) \right) \right\} \right. \\ &\qquad \left. \Big| \text{all paths stay in } C_N \right] P(\text{all paths stay in } C_N \text{ up to time } T). \end{aligned}$$

Writing $E_0^{C_N}[\,\cdot\,]=E_0[\,\cdot\,|$ all paths stay in C_N up to time $T]$, we use Jensen's theorem, the fact that

$$\log\left(1+\epsilon^2 \sum_{i<j} y_i y_j \right) \geqslant c\,\epsilon^2 \sum_{i<j} y_i y_j e^{-\Sigma_k y_k},$$

where $0 \leqslant y_i \leqslant 1$, and proposition II.1 to get that for sufficiently large N

$$\langle Z(T)^n\rangle_v \geqslant \exp\left\{ E_0^{C_N}\left[\sum_{0\leqslant s\leqslant T, x\in C_{N/2}} \log\left\{ 1+\epsilon^2 \sum_{1\leqslant i<j\leqslant n} \delta(X_i(s)-x)\,\delta(X_j(s)-x)\right\}\right]\right\}(Ce^{-cT/N^2})^n$$

$$\geqslant \exp\left\{ E_0^{C_N}\left[\sum_{s,x} c_1\epsilon^2 \sum_{i<j} \delta(X_i(s)-x)\,\delta(X_j(s)-x) e^{-\Sigma_{1\leqslant k\leqslant n}\delta(X_k(s)-x)}\right]\right\} C' e^{-cTn/(N^2)}$$

$$\geqslant \exp\left[\sum_{0\leqslant s\leqslant T, x\in C_{N/2}} C_1\epsilon^2 \sum_{1\leqslant i<j\leqslant n} P_0(X_i(s)=x|\text{stay in } C_N) P_0(X_j(s)=x|\text{stay in } C_N)\right.$$

$$\left. \times \prod_{k\neq i,j} \{1-P_0(X_k(s)=x|\text{stay in } C_N)\}\right] C' e^{-cTn/(N^2)}$$

$$\geqslant \exp\left[(T-N^2)c_2 N^d c_1\epsilon^2 n^2 \frac{K_1}{N^d}\frac{K_1}{N^d}\left(1-\frac{K_2}{N^d}\right)^n\right] C' e^{-cTn/(N^2)}$$

$$\geqslant C'' \exp\left[\frac{T_n}{N^2}\left\{ \frac{C_3\epsilon^2 n}{N^{d-2}}\left(1-\frac{C}{N^d}\right)^n - c\right\}\right],$$

where the constants are independent of T. We need $\epsilon^2 n/N^{d-2} > c'$ which implies $n \geqslant N^{d-2}$ and we need $n \leqslant N^d$ to bound $(1-C/(N^d))^n$, so we choose $n=N^d$. Then we need $c_3\epsilon^2 N^2 - c > 0$, so $N > c/\epsilon$, which implies $n > C/\epsilon^d$. With this choice of n and N, we have

$$\langle Z(T)^n\rangle_v \to \infty \text{ as } T\to\infty. \qquad \blacksquare$$

III. WALKS THAT REMAIN IN A CUBE

The goal of this section is to prove proposition II.1. We first need to prove a number of lemmas concerning simple random walks that are constrained to a cube. In general let $P(x,y,t) = P_x(X(t)=y)$ and $P_D(x,y,t) = P_x(X(t)=y; X(s)\in D, 0\leqslant s\leqslant t)$, where D is some domain in $\mathbb{Z}^d$.

Lemma III.1: Let Δ_L be the d-dimensional lattice Laplacian. Then if φ_0 is an eigenfunction of Δ_L on a domain D, satisfying $\varphi_0(x)=0$, for x on the boundary of $D=\delta D$ (where $D\cap\delta D=\theta$), with eigenvalue λ_0, we have that for $x\in D$,

$$\sum_{y\in D} P_D(x,y,n)\varphi_0(y) = (1+\lambda_0)^n \varphi_0(x).$$

Proof: Let 1_A be the indicator function of a set A (referred to as the characteristic function in analysis), let i_k be the unit vector in the kth direction, and let

$$u(x,n) = E_x[f(X(n))1_{(X(s)\in D, 0\leqslant s\leqslant n)}],$$

for some function f. Then, by the Markov property and defining

$$\overline{u}(x,n) = \begin{cases} u(x,n) & \text{if } x\in D \\ 0 & \text{if } x\in \delta D, \end{cases}$$

we have, for $x\in D$,

$$u(x,n+1)-u(x,n) = \frac{1}{2d}\left\{ \sum_{k=1}^{d} [\overline{u}(x+i_k,n)+\overline{u}(x-i_k,n)-2u(x,n)]\right\}$$

$$= \overline{\Delta}_L u(x,n) \quad (\text{this defines } \overline{\Delta}_L) \Rightarrow u(x,n) = (1+\overline{\Delta}_L)^n u(x,0).$$

For any $y\in D$, let $f_y(X(n)) = 1_{(X(n)=y)}$. Then

$$E_x[f_y(X(n))1_{(X(s)\in D, 0\leqslant s\leqslant n)}] = P_D(x,y,n).$$

If φ_0 is an eigenfunction of Δ_L on D, satisfying $\varphi_0(x)=0$ for $x\in\partial D$, with eigenvalue λ_0, then φ_0 is also an eigenfunction of Δ_L with eigenvalue λ_0 and we have that for $x\in D$,

$$\sum_{y\in D} P_D(x,y,n)\varphi_0(y)=\sum_{y\in D}(1+\bar{\Delta}_L)^n P_D(x,y,0)\varphi(y)=(1+\lambda_0)^n\varphi_0(x). \quad \blacksquare$$

In general, let $y\leftrightarrow t$ denote the assertion that y and t have the same parity [i.e., if $y=(y_1,\ldots,y_d)$, then $y_1+\cdots y_d+t$ is even] and $\lfloor\cdot\rfloor$ is the greatest integer function. We now want to investigate walks on $\mathbb{Z}$ that remain in an interval.

A. One-dimensional walks

Lemma III.2: Suppose $0<\rho<1$, $0<\mu$, $0<\beta$, $x, y\in\mathbb{Z}$, $y\leftrightarrow x$, *and*

$$|x|<\rho\mu R, \quad |y-x|<\left(\frac{1-\rho}{2}\right)\mu R.$$

Let $N=\max\{k\in\mathbb{Z}:k\leqslant\mu R\}$, $D_{\mu R}=(-N,N)\cap\mathbb{Z}$, $m=\min\{r\in\mathbb{Z}:r\geqslant\beta R^2$ and $r\leftrightarrow(y-x)\}$. *Then for* $R>10/(1-\rho)\mu$, *we have*

$$P_{D_{\mu R}}(x,y,m)\geqslant\frac{C(\rho,\mu,\beta)}{R},$$

where $C(\rho,\mu,\beta)$ *is a constant depending on* ρ, μ, *and* β.

Proof: Fix $x\in\mathbb{Z}^d$ and γ so that $0<\gamma\leqslant\beta$. Fix $z\in\mathbb{Z}$ with $|z-x|<\frac{3}{4}(1-\rho)\mu R$. Let $n=\min\{r\in\mathbb{Z}:r>\gamma R^2$ and $r\leftrightarrow(z-x)\}$. Also let

$$\sigma_1(t)=P_x(\text{exit } (-N,N) \text{ at } N \text{ at time } t),$$

and

$$\sigma_2(t)=P_x(\text{exit } (-N,N) \text{ at } -N \text{ at time } t).$$

Then we have

$$P_{D_{\mu R}}(x,z,n)=P(x,z,n)-\sum_{t=1}^{n}\sigma_1(t)P(N,z,n-t)-\sum_{t=1}^{n}\sigma_2(t)P(-N,z,n-t). \tag{4}$$

By the local CLT, we have that for $1\leqslant t\leqslant n-1$,

$$P(N,z,n-t)\leqslant\max_{1\leqslant t\leqslant n-1}\frac{c}{\sqrt{n-t}}\,e^{-c(N-(3+\rho)\mu R/4)^2/(n-t)}\leqslant\frac{C}{\mu R(1-\rho)}.$$

Then using the reflection principle and the local CLT, we have

$$\begin{aligned}\sum_{t=1}^{n}\sigma_1(t)&\leqslant P_x(\text{time to exit } (-N,N)\leqslant n)\\&\leqslant P_0(\max_{0\leqslant t\leqslant n}|X(t)|\geqslant N-x)\\&\leqslant 2P_0(|X(n)|\geqslant(1-\rho)\mu R)\\&\leqslant\sum_{x=\lfloor(1-\rho)\mu R\rfloor}^{\infty}\frac{C}{\sqrt{n}}\exp\left[-\frac{x^2}{2n}\right]\leqslant\frac{C\sqrt{\gamma}}{(1-\rho)\mu}\exp[-(1-\rho)^2\mu^2/(2\gamma)].\end{aligned}$$

Similarly,

$$P(-N,z,n-t)\leqslant\frac{C}{\mu R(1-\rho)},$$

$$\sum_{t=1}^{n} \sigma_2(t) \leqslant \frac{C\sqrt{\gamma}}{(1-\rho)\mu \exp[-(1-\rho)^2\mu^2/(2\gamma)]}.$$

Also, the local CLT gives

$$P(x,z,n) \geqslant \frac{c}{\sqrt{\gamma}R} \exp\left[-\frac{9(1-\rho)^2\mu^2}{32\gamma}\right].$$

Returning to Eq. 4, we have

$$P_{D_{\mu R}}(x,z,n) \geqslant \frac{c}{\sqrt{\gamma}R} \exp\left[-\frac{9(1-\rho)^2\mu^2}{32\gamma}\right] - \frac{C}{\mu R(1-\rho)} \frac{\sqrt{r}}{(1-\rho)\mu} \exp\left[-\frac{(1-\rho)^2\mu^2}{2\gamma}\right].$$

By choosing γ sufficiently small, we get

$$P_{D_{\mu R}}(x,z,n) \geqslant \frac{C(\rho,\mu)}{R}. \tag{5}$$

Let $M = \max\{m \in \mathbb{Z} : m \leqslant \frac{3}{4}(1-\rho)\mu R\}$, and $D_x = [x-M, x+M]$. For fixed $x \in \mathbb{R}$, we have $\varphi_0(z) = \cos(\pi(z-x)/(2M))$ is an eigenfunction of Δ_L on D_x, with eigenvalue $\lambda_0 = -\sin^2(\pi/(4M))$ and $\varphi_0(y) = 0$ if $y \in \partial D_x$. Using Eq. (5), lemma III.1, and $\max_{z \in D_x}\{\varphi_0(z)\} = 1$, we have

$$P_{D_{\mu R}}(x,y,m) \geqslant \sum_{z \in D_x} P_{D_{\mu R}}(z,y,m-n) P_{D_{\mu R}}(x,z,n) \geqslant \frac{C(\rho,\mu)}{R} \sum_{z \in D_x} P_{D_x}(z,y,m-n)\varphi_0(z)$$

$$= \frac{C(\rho,\mu)}{R} (1-\sin^2(\pi/(4M)))^{m-n} \varphi_0(y).$$

Noticing that $\varphi_0(y) \geqslant \cos(\pi/3)$ and

$$(1-\sin^2(\pi/(4M)))^{m-n} \geqslant \exp\left[-\frac{C\beta}{(1-\rho)^2\mu^2}\right],$$

we are done. ■

Lemma III.3: Suppose $0 < \rho < c < 1$, $0 < \mu$, $0 < \beta < c < 1$, $0 < \gamma$, *and* $R > 10/(1-c)\mu$. *Let* $n = \min\{r \in \mathbb{Z} : r \geqslant \gamma R^2$ *and* $r \leftrightarrow (y-x)\}$, *where* $x, y \in \mathbb{Z}$, $x \leftrightarrow y$, $|x| < \rho\mu R$, *and* $|y| < \beta\mu R$. *Letting* $D_{\mu R}$ *be the domain defined in lemma III.2, we have*

$$P_{D_{\mu R}}(x,y,n) \geqslant \frac{C(\rho,\mu,\gamma,\beta)}{R}.$$

Proof: If $|y-x| \leqslant (1-\rho)\mu R/2$, then the lemma holds by lemma III.2. Otherwise, let

$$\operatorname{sgn}(y-x) = \begin{cases} +1 & \text{if } y-x > 0 \\ -1 & \text{if } y-x < 0 \end{cases}$$

and

$$m=\left\lfloor \frac{|y-x|}{\left(\frac{1-\rho}{4}\right)\mu R}\right\rfloor = C(\rho,\beta).$$

Also let $E_1=\{y_1\in\mathbb{Z}:|y_1-x|<(1-\rho)\mu R/2,|y_1-y|<|x-y|-(1-\rho)R\mu/4,\lfloor n/m\rfloor\leftrightarrow(x-y_1)\}$ and for $2\leqslant k\leqslant m-1$, let $E_k=\{y_k\in\mathbb{Z}:|y_k-(x+\mathrm{sgn}(y-x)k(1-\rho)R\mu/4)|\leqslant(1-\rho)R\mu/4,\ |y_k-y|\leqslant|x-y|-k(1-\rho)R\mu/4,\lfloor n/m\rfloor\leftrightarrow(y_k-y_{k-1})\}$. Then by lemma III.2

$$\begin{aligned}P_{D_{\mu R}}(x,y,n)&\geqslant \sum_{y_k\in E_k,1\leqslant k\leqslant m-1} P_{D_{\mu R}}(x,y_1,\lfloor n/m\rfloor)P_{D_{\mu R}}(y_1,y_2,\lfloor n/m\rfloor)\cdots\\ &\quad\times P_{D_{\mu R}}(y_{m-1},y,n-(m-1)\lfloor n/m\rfloor)\\ &\geqslant\left(\frac{(1-\rho)R\mu}{8}\right)^{m-1}\left(\frac{C(\rho,\mu,\gamma)}{R}\right)^m\geqslant\frac{C(\rho,\mu,\gamma,\beta)}{R}.\end{aligned}$$ ■

B. *d*-dimensional walks

*Lemma III.4: For the simple random walk on $\mathbb{Z}^d$, let $S_n^{(k)}$ be the number of steps taken in the (positive or neqative) k*th *direction out of a total of n steps taken on $\mathbb{Z}^d$. Then*

$$P\left(\frac{3n}{4d}\leqslant S_n^{(k)}\leqslant\frac{5n}{4d},k=1,\ldots,d\right)\geqslant 1-\frac{16d(d-1)}{n}.$$

Proof: By Chebyshev,

$$\begin{aligned}P\left(\left|S_n^{(k)}-\frac{n}{d}\right|\geqslant\frac{n}{4d}\right)&\leqslant\frac{16d^2}{n^2}E\left[\left(S_n^{(k)}-\frac{n}{d}\right)^2\right]=\frac{16(d-1)}{n}\\ &\Rightarrow P\left(\left|S_n^{(k)}-\frac{n}{d}\right|<\frac{n}{4d},k=1,\ldots,d\right)\\ &\geqslant 1-\left\{\sum_{k=1}^d P\left(\left|S_n^{(k)}-\frac{n}{d}\right|\geqslant\frac{n}{4d}\right)\right\}\geqslant 1-\frac{16d(d-1)}{n}.\end{aligned}$$ ■

Note that henceforth, N will represent a large positive integer.

Lemma III.5: Let $C_N=[-(N-1),N-1]^d\cap\mathbb{Z}^d$, $\partial C_N=\{[-N,N]^d\cap\mathbb{Z}^d\}\backslash C_N$, $C_{N/2}=(-N/2,N/2)^d\cap\mathbb{Z}^d$, and

$$P_{C_N}(x,y,T)=P_x(X(T)=y;X(t)\in C_N,0\leqslant t\leqslant T).$$

Then there exists a constants C_1 and C_2 such that for all $y\in C_{N/2}$,

$$\frac{C_1}{N^d}\leqslant P_{C_N}(0,y,N^2)\leqslant\frac{C_2}{N^d}.$$

Proof: Letting $n_i(\gamma_i,y_i)=\min\{k\in\mathbb{Z}:k\geqslant\gamma_iN^2$ and $k\leftrightarrow y_i\}$, and $y=(y_1,\ldots,y_d)$, we have by lemmas III.3 and III.4,

$$P_{C_N}(0,y,N^2)\geqslant\inf_{3/(4d)\leqslant\gamma_i\leqslant5/(4d),|y_i|<N/2}\left\{\prod_{i=1}^d P_{D_N}(0,y_i,n_i(\gamma_i,y_i))\right\}\frac{1}{2^d}\left(1-\frac{16d(d-1)}{n}\right)\geqslant\frac{C_1}{N^d}.$$

Also, by the local CLT

$$P_{C_N}(0,y,N^2) \leqslant P_0(X(N^2)=y) \leqslant \frac{C_2}{N^d}. \qquad \blacksquare$$

We shall only use the dimension 1 version of the following lemma. For completeness we have included the d-dimensional version.

Lemma III.6: Suppose R is a sufficiently large positive real number. Let

$$\partial B_R = \{x \in \mathbb{Z}^d : |x| \leqslant R \ \text{ and } \ \exists x' \in \mathbb{Z}^d \text{ s.t.} |x-x'|=1 \ \text{ and } \ |x'|>R\},$$

$$B_R = \{x \in \mathbb{Z}^d : |x| < R \ \text{ and } \ x \notin \partial B_R\}.$$

Then there exists a universal constant C such that, $\forall x \in B_R$ *and* $t \in \{s \in \mathbb{Z} : s \geqslant R^2\}$,

$$\sum_{y \in B_R} P_{B_R}(x,y,t) \leqslant C \sum_{|y|<R/(2\sqrt{d})} P_{B_R}(x,y,t). \tag{6}$$

Proof: Let $n = \min\{s \in \mathbb{Z} : s \geqslant R^2\}$, and suppose the lemma is true for $t=n$. Then by the Markov property, we have, for $t>n$,

$$\begin{aligned}\sum_{y \in B_R} P_{B_R}(x,y,t) &= \sum_{z \in B_R} P_{B_R}(x,z,t-n) \sum_{y \in B_R} P_{B_R}(z,y,n) \\ &\leqslant \sum_{z \in B_R} P_{B_R}(x,z,t-n) C \sum_{|y|<R/(2\sqrt{d})} P_{B_R}(z,y,n) \\ &= C \sum_{|y|<R/(2\sqrt{d})} P_{B_R}(x,y,t).\end{aligned}$$

So we need only consider $t=n$. Suppose $0<\alpha<c_1<1$, for some constant c_1. For any x so that $R/(2\sqrt{d}) < |x| < \alpha(R-1)$, and for $\nu = 1/4d\alpha$, let

$$\begin{aligned}C_{R,x} = [-|x_1| - \lfloor (1-\alpha)(R-1)\nu \rfloor, |x_1| + \lfloor (1-\alpha)(R-1)\nu \rfloor] \times \cdots \times [-|x_d| \\ - \lfloor (1-\alpha)(R-1)\nu \rfloor, |x_d| + \lfloor (1-\alpha)(R-1)\nu \rfloor],\end{aligned}$$

where $x=(x_1,\ldots,x_d)$. Clearly $x \in C_{R,x}$, and if $z \in C_{R,x}$, then

$$|z|^2 \leqslant (|x_1| + (1-\alpha)(R-1)\nu)^2 + \cdots + (|x_d| + (1-\alpha)(R-1)\nu)^2 \leqslant (R-1)^2.$$

If $|x| \leqslant R/(2\sqrt{d})$, let

$$C_{R,x} = [-\lfloor R/\sqrt{d} \rfloor, \lfloor R/\sqrt{d} \rfloor] \times \cdots \times [-\lfloor R/\sqrt{d} \rfloor, \lfloor R/\sqrt{d} \rfloor] \subset B_R.$$

Therefore for all x such that $|x| \leqslant \alpha(R-1)$, $\exists C_{R,x}$ s.t. $x \in C_{R,x}$ and $C_{R,x} \subset B_R$. How close can x be to the side of the box $C_{R,x}$? If $|x| \leqslant R/(2\sqrt{d})$, then $|x_i|/R/\sqrt{d} \leqslant \frac{1}{2}$. If $R/(2\sqrt{d}) < |x| \leqslant \alpha(R-1)$, then

$$\frac{|x_i|}{|x_i| + \lfloor (1-\alpha)(R-1)\nu \rfloor} \leqslant \frac{\alpha}{\alpha + (1-\alpha)/(4d\alpha)} = \rho_1(\alpha) < 1.$$

So if $|x| \leqslant R/(2\sqrt{d})$, let $\rho = 1/2$ and $\mu_i = 1/\sqrt{d}$, for $1 \leqslant i \leqslant d$. Otherwise let $\rho = \rho_1$ and $\mu_i = [|x_i| + (1-\alpha)(R-1)\nu]/(R-1)$. Note that $(1-\alpha)\nu(\alpha) \leqslant \mu_i \leqslant \alpha + (1-\alpha)\nu(\alpha)$. Letting $n_i(\gamma_i, x_i, y_i) = \min\{k \in \mathbb{Z} : k \geqslant \gamma_i R^2 \text{ and } k \leftrightarrow (y_i - x_i)\}$, and $x=(x_1,\ldots,x_d)$, we have by lemmas III.4 and III.3

$$\sum_{|y|<R/(2\sqrt{d})} P_{B_R}(x,y,n) \geqslant \sum_{|y|<R/(2\sqrt{d}),\, y\in C_{R/2,x}} P_{C_{R,x}}(x,y,n)$$

$$\geqslant C \sum_{|y|<R/(2\sqrt{d}),\, y\in C_{R/2,x}} \inf_{3/4d\leqslant\gamma_i\leqslant 5/4d,\, |y_i|<\mu_i R/2} \prod_{i=1}^{d} P_{D_{\mu_i(R-1)}}(x_i,y_i,n_i(\gamma_i,x_i,y_i))$$

$$\geqslant C \sum_{|y|<R/(2\sqrt{d}),\, y\in C_{R/2,x}} \inf_{3/4d\leqslant\gamma_i\leqslant 5/4d} \prod_{i=1}^{d} \frac{C(\rho_i(\alpha),\mu_i(\alpha),\gamma_i)}{R}$$

$$\geqslant \sum_{|y|<R/(2\sqrt{d}),\, y\in C_{R/2,x}} \frac{C(\alpha)}{R^d} = C(\alpha), \tag{7}$$

for $R \geqslant 1/(1-\rho_1(c_1))$, where $C(\alpha)$ is a positive constant depending only on α. Since the left-hand side (lhs) of Eq. (6) is bounded above by 1, Eq. (6) holds for $t=n$ and $\forall x$ satisfying $|x| < \alpha(R-1)$. Note that as $|x|\to R$, both the right-hand side (rhs) and the lhs of Eq. (6) go to zero, so we need to find some estimates of these quantities close to the boundary of B_R.

Let τ be the first time for the random walk $X(t)$ to exit B_R and $E_x[\tau]=u(x)$. Then by Chebyshev

$$\sum_{y\in B_R} P_{B_R}(x,y,n) = P_x(\tau>n) \leqslant \frac{u(x)}{n}. \tag{8}$$

We also have $\Delta_L u(x) = -1$ where $u(x)=0$ if $x\in\partial B_R$. Let $u_1(x)=R^2-r^2$, where $r=|x|$. Then $\Delta_L u_1(x) = -1$ and $u_1(x)=0$ if $|x|=R$. Therefore, $\forall x\in B_R$, $\Delta_L(u(x)-u_1(x))=0$, and $\forall x \in \partial B_R$, $u(x)-u_1(x)\leqslant 0$, and by the maximum principle,

$$u(x) \leqslant R^2 - r^2 \forall x \in B_R. \tag{9}$$

Let $\partial B_{\alpha R} = \{x\in\mathbb{Z}^d : |x| \geqslant \alpha(R-1)$ and $\exists x'$ s.t. $|x-x'|=1$ and $|x'|<\alpha(R-1)\}$, and let $U_{\alpha R}=\{x \in \mathbb{Z}^d : \alpha(R-1)\leqslant|x|\leqslant R$ and $x\notin\partial B_{\alpha R}, x\notin\partial B_R\}$. Then for $x\in U_{\alpha R}$,

$$\begin{aligned}
&\sum_{|y|<R/(2\sqrt{d})} P_{B_R}(x,y,n) \\
&\quad = P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R}; X(t)\in B_R, 0\leqslant t\leqslant n; X(n)<R/(2\sqrt{d})) \\
&\quad \geqslant P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R} \text{ in time } \tau<n/2) \\
&\qquad \times \inf_{y\in\partial B_{\alpha R}, 0<\tau<n/2} P_y\left(X(t)\in B_R, 0<t<n-\tau; |X(n-\tau)|<\frac{R}{2\sqrt{d}} \right) \\
&\quad \geqslant P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R} \text{ in time } \tau<n/2) \\
&\qquad \times \inf_{y\in\partial B_{\alpha R}, n/2\leqslant m\leqslant n} \sum_{|z|<R/(2\sqrt{d})} P_{B_R}(y,z,m) \\
&\quad \geqslant C_\alpha P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R} \text{ in time } \tau<n/2),
\end{aligned} \tag{10}$$

using Eq. (7). Letting τ_U=time to exit $U_{\alpha R}$, Chebyshev gives

$$P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R} \text{ in time } \tau\geqslant n/2) \leqslant \frac{2E_x[\tau_U]}{n}.$$

Therefore

$$P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R} \text{ in time } \tau<n/2)$$

$$\geqslant P_x(X(t) \text{ exits } U_{\alpha R} \text{ through } \partial B_{\alpha R}) - \frac{2E_x[\tau_U]}{n}. \tag{11}$$

Let $v(x) = P_x(X(t)$ exits $U_{\alpha R}$ through $\partial B_{\alpha R})$. We have $\Delta_L v(x) = 0$ with $v(x) = 1$ if $x \in \partial B_{\alpha R}$ and $v(x) = 0$ if $x \in \partial B_R$. Let $v_1(x) = \alpha^d/(1-\alpha^d)((R-1)^d/r^d - 1)$. Taylor's theorem gives

$$\Delta_L v_1(x) = \frac{\alpha^d (R-1)^d}{2d(1-\alpha^d)} \left\{ \frac{2d}{r^{2+d}} + O\left(\frac{1}{r^{4+d}} \right) \right\},$$

and we can see that for $x \in U_{\alpha R}$, with α close to one, and R large, $\Delta_L v_1(x) > 0$. The maximum principle gives

$$v_1(x) \leqslant v(x) \forall x \in U_{\alpha R}. \tag{12}$$

Let $w(x) = E_x[\tau_U]$ and $w_1(x) = R^2 - r^2$. The maximum principle gives

$$w < w_1 \text{ on } U_{\alpha R}. \tag{13}$$

Returning to Eq. (10) we have by Eqs. (11)–(13), that for $x \in U_{\alpha R}$,

$$\sum_{|y| < R/(2\sqrt{d})} P_{B_R}(x,y,n) \geqslant C_\alpha \left\{ \frac{\alpha^d}{1-\alpha^d} \left(\frac{(R-1)^d}{r^d} - 1 \right) - \frac{2(R^2 - r^2)}{n} \right\}.$$

Choosing α sufficiently close to one, there exists a constant C_1 so that for all $x \in U_{\alpha R}$ with $|x| \leqslant R-2$,

$$\sum_{|y| < R/(2\sqrt{d})} P_{B_R}(x,y,n) \geqslant C_1 \frac{2(R^2 - r^2)}{n} \geqslant C_1 \sum_{y \in B_R} P_{B_R}(x,y,n), \tag{14}$$

using Eqs. (8) and (9). If $x \in B_R$ and $|x| > R-2$, then there exists $x' \in \mathbb{Z}^d$ so that $|x'| < R-2$ and it is possible for a walk to get from x to x', staying in B_R in less than $2d$ steps. Then we have, using Eqs. (8) and (9) again,

$$\begin{aligned}
\sum_{|y| < R/(2\sqrt{d})} P_{B_R}(x,y,n) &= \sum_{|y| < R/(2\sqrt{d})} \sum_{z \in B_R} P_{B_R}(z,y,n-2d) P_{B_R}(x,z,2d) \\
&\geqslant \sum_{|y| < R/(2\sqrt{d})} \frac{1}{(2d)^{2d}} P_{B_R}(x',y,n-2d) \\
&\geqslant \frac{C_1}{(2d)^{2d}} \frac{2(R^2 - |x'|^2)}{n-2d} \\
&\geqslant \frac{C_1}{(2d)^{2d}} \frac{2(R^2 - |x'|^2)}{n} \\
&\geqslant \frac{C_1}{(2d)^{2d}} \sum_{y \in B_R} P_{B_R}(x,y,n).
\end{aligned} \tag{15}$$

Letting $1/C = \min\{C_1/(2d)^{2d}, C(\alpha)\}$, the lemma is proved. ■

Corollary III.7: For N a sufficiently large positive integer, there exists a universal constant C such that $\forall x \in CN$, *and* $\forall t \geqslant N^2$,

$$\sum_{y \in C_N} P_{C_N}(x,y,t) \leqslant C \sum_{y \in C_{N/2}} P_{C_N}(x,y,t).$$

Proof: As in lemma III.6, by the Markov property, we need only consider $t=n=N^2$. We can write $S_n^{(k)}=Y_1+\cdots+Y_n$, where the Y_i are independent and identically distributed, $Y_i=1$ with probability $1/d$, and $Y_i=0$ with probability $1-1/d$. Fix θ so that $0<\theta\leqslant 1/2$. Then by the central limit theorem, for n sufficiently large,

$$P\left(\left|\frac{S_n^{(k)}-n/d}{(\sqrt{d-1}/d)\ \sqrt{n}}\right|>n^{\theta}\right)<\frac{4}{\sqrt{2\pi}}\int_{n^{\theta}}^{\infty}e^{-x^2/2}dx$$

$$\Rightarrow P\left(\left|S_n^{(k)}-\frac{n}{d}\right|>n^{\theta+1/2}\right)<\frac{2}{n^{\theta}}\,e^{-n^{2\theta}/2}.$$

We thus have, writing $(P_{C_N}(x,y,n);A)=P_x(X(n)=y;X(t)\in C_N,0\leqslant t\leqslant n;A)$,

$$\sum_{y\in C_N}P_{C_N}(x,y,n)\leqslant\sum_{y\in C_N}(P_{C_N}(x,y,n);|S_n^{(k)}-n/d|\leqslant n^{\theta+1/2}\forall k=1,\ldots,d)+\frac{2}{n^{\theta}}\,e^{-n^{2\theta}/2}. \quad (16)$$

Let $n_i(\gamma_i,y_i)=\min\{p\in\mathbb{Z}:p\geqslant\gamma_iN^2 \text{ and } p\leftrightarrow y_i\}$. Noting $n=N^2$, there exists positive constants c_1, c_2, and N_0 such that $\forall N>N_0$,

$$c_1n\leqslant n/d-n^{\theta+1/2}<n/d+n^{\theta+1/2}\leqslant c_2n.$$

Then, using the dimension 1 version of lemma III.6, where $B_N=[-(N-1),N-1]$,

$$\begin{aligned}\sum_{y\in C_{N/2}}P_{C_N}(x,y,n)&\geqslant\sum_{|y_i|<N/2,c_1\leqslant\gamma_i\leqslant c_2,\Sigma_{i=1}^d n_i=n}\prod_{i=1}^{d}P_{B_N}(x_i,y_i,n_i(\gamma_i,y_i))\\&\geqslant C\sum_{|y_i|<N,c_1\leqslant\gamma_i\leqslant c_2,\Sigma_{i=1}^d n_i=n}\prod_{i=1}^{d}P_{B_N}(x_i,y_i,n_i(\gamma_i,y_i))\\&\geqslant C\sum_{|y_i|<N}(P_{C_N}(x,y,n);c_1n<S_n^{(k)}<c_2n,\forall k=1,\ldots,d)\\&\geqslant\frac{1}{C_1}\sum_{y\in C_N}(P_{C_N}(x,y,n);|S_n^{(k)}-n/d|\leqslant n^{\theta+1/2},\forall k=1,\ldots,d), \quad (17)\end{aligned}$$

where C_1 is a constant. Equation (17) also gives, using Eqs. (7), (14), and (15), and $|x_i|\leqslant N-1$,

$$\sum_{y\in C_{N/2}}P_{C_N}(x,y,n)\geqslant\frac{C}{N^d}.$$

Therefore there exists constants $N_1\in\mathbb{Z}^+$ and C_2 such that (for $N>N_1$)

$$C_2\sum_{y\in C_{N/2}}P_{C_N}(x,y,n)\geqslant\frac{2}{n^{\theta}}\,e^{-n^{2\theta}/2}.$$

Returning to Eq. (16), we have for $N>\max\{N_0,N_1\}$,

$$\sum_{y\in C_N}P_{C_N}(x,y,n)\leqslant(C_1+C_2)\sum_{y\in C_{N/2}}P_{C_N}(x,y,n)=C\sum_{y\in C_{N/2}}P_{C_N}(x,y,n). \quad \blacksquare$$

*Lemma III.8: Let $x\in C_{N/2}$. Then there exists positive constants C_1 and C_2*s.t.*$\forall n\in\mathbb{Z}^+$, and N sufficiently large*

$$C_1(1+\lambda_0)^n\leqslant\sum_{y\in C_N}P_{C_N}(x,y,n)\leqslant C_2(1+\lambda_0)^n,$$

where

$$-\frac{2\pi^2}{16N^2}\leqslant\lambda_0\leqslant-\frac{\pi^2}{16N^2}.$$

Proof: $\varphi_0(x)=\Pi_{k=1}^{d}\cos(\pi x_k/(2N))$ is an eigenfunction of Δ_L on C_N, with eigenvalue $\lambda_0=-2\sin^2(\pi/(4N))$ and $\varphi_0(x)=0$ if $x\in\partial C_N$, and we have

$$-\frac{2\pi^2}{16N^2}\leqslant\lambda_0\leqslant-\frac{\pi^2}{16N^2}.$$

Using lemma III.1, $\sup_{y\in C_N}\varphi_0(y)\leqslant 1$, and $\inf_{x\in C_{N/2}}\varphi_0(x)\geqslant C>0$, we have

$$\sum_{y\in C_N}P_{C_N}(x,y,n)\geqslant C_1(1+\lambda_0)^n.$$

Also, by corollary III.7, and using lemma III.1 again, we have, for $n\geqslant N^2$,

$$\sum_{y\in C_N}P_{C_N}(x,y,n)\leqslant C\sum_{y\in C_{N/2}}P_{C_N}(x,y,n)\varphi_0(y)\leqslant C'\sum_{y\in C_N}P_{C_N}(x,y,n)\varphi_0(y)\leqslant C'(1+\lambda_0)^n.$$

Letting $C''=\inf_{1\leqslant n\leqslant N^2}(1+\lambda_0)^n$, we have, for $n\leqslant N^2$,

$$\sum_{y\in C_N}P_{C_N}(x,y,n)\leqslant 1\leqslant\frac{(1+\lambda_0)^n}{C''}.$$

Choosing $C_2=\max\{C',1/C''\}$, we are done. ■

We are now ready to prove proposition II.1. From lemma III.8, we have

$$\sum_{y\in C_N}P_{C_N}(x,y,T)\geqslant C_1\left(1-2\frac{\pi^2}{16N^2}\right)^T\geqslant Ce^{-cT/N^2}.$$

This proves Eq. (1). Also, using lemma III.5 and corollary III.7, we have

$$\begin{aligned}P_{C_N}(0,x,t)&\geqslant\sum_{y\in C_{N/2}}P_{C_N}(0,y,N^2)P_{C_N}(y,x,t-N^2)\\&\geqslant\frac{C_1}{N^d}\sum_{y\in C_{N/2}}P_{C_N}(y,x,t-N^2)\geqslant\frac{C'}{N^d}\sum_{y\in C_N}P_{C_N}(y,x,t-N^2),\end{aligned}$$

and

$$P_{C_N}(0,x,t)=\sum_{y\in C_N}P_{C_N}(0,y,N^2)P_{C_N}(y,x,t-N^2)\leqslant\frac{C_2}{N^d}\sum_{y\in C_N}P_{C_N}(y,x,t-N^2).$$

We have, using lemma III.8,

$$\begin{aligned}&P_0(X(t)=x\,|\,\text{stay in } C_N \text{ up to time } T)\\&\quad=\frac{P_{C_N}(0,x,t)\Sigma_{y\in C_N}P_{C_N}(x,y,T-t)}{\Sigma_{y\in C_N}P_{C_N}(0,y,T)}\\&\quad\geqslant\frac{C''}{N^d}\frac{(1+\lambda_0)^{t-N^2}(1+\lambda_0)^{T-t}}{(1+\lambda_0)^T}\geqslant\frac{C''}{N^d}\frac{1}{\left(1-\frac{\pi^2}{16N^2}\right)^{N^2}}\geqslant\frac{C_1}{N^d}.\end{aligned}$$

A similar calculation yields the upper bound, and hence Eq. (2) holds. ■

ACKNOWLEDGMENTS

Much of this work constituted part of the author's dissertation at The University of Michigan. The author wishes to thank his dissertation advisor, Joseph Conlon, for his help and inspiration, and Greg Lawler for useful discussions.

APPENDIX: PROBABILITY THAT THE PARTITION FUNCTION IS SMALL

Proposition: Given $\epsilon>0$ and $\delta>0$ both sufficiently small, we have that for all T sufficiently large in dimensions $d>2$,

$$P_v(Z(T)<\delta)\leqslant \frac{c}{|\log(\delta)|},$$

where c is a constant that depends on ϵ.

Proof: Using the notation of Imbrie and Spencer,[3] propositions 2.2 and 3.1 of Ref. 3 (choosing $\eta=0.5$ and $\kappa=1$) imply that for any $k\in\mathbb{Z}^+$, ϵ sufficiently small, letting $s=2^k-1$ and $T>s$,

$$P_v\left(\frac{Z(T)}{Z(s)}\geqslant cs^{-d}(1-c\epsilon^{1/2})\right)\geqslant P_v\Big(\bigcap_{j\geqslant k} E_j^{(k)}\Big)$$

$$\geqslant 1-\sum_{j\geqslant k} P_v(E_j^{(k)c})\geqslant 1-\sum_{j=k+1}^{\infty}\epsilon 2^{-j}=1-\epsilon 2^{-k}.$$

Note that $Z(s)\geqslant(1-\epsilon)^s$. Therefore

$$P_v(Z(T)<c(2^k-1)^{-d}(1-c\epsilon^{1/2})(1-\epsilon)^{2k-1})<\epsilon 2^{-k}$$

$$\Rightarrow P_v(Z(T)<c'2^{-kd}(1-\epsilon)^{2^k})<\epsilon 2^{-k}.$$

Choose k sufficiently large so that $c'(1-\epsilon)^{2^k}2^{-kd}>(1-\epsilon^{1/2})^{2^k}$ and $\delta=(1-\epsilon^{1/2})^{2^k}$. Then for all $T>s$,

$$P_v(Z(T)<\delta)<\frac{c}{|\log\ \delta|}.$$

Note that if $T\leqslant s$,

$$P_v(Z(T)<\delta)=P_v(Z(T)<(1-\epsilon^{1/2})^{2^k})\leqslant P_v(Z(T)<(1-\epsilon)^s)=0.$$ ■

[1] D. A. Huse and C. L. Henley, ''Pinning and Roughening of Domain Walls in Ising Systems Due to Random Impurities,'' Phys. Rev. Lett. **54**, 2708–2711 (1985).
[2] D. Fisher and T. Hwa, ''Anomalous Fluctuations of Directed Polymers in Random Media,'' Phys. Rev. B **49**, 3136–3154 (1994).
[3] J. Z. Imbrie and T. Spencer, ''Diffusion of Directed Polymers in a Random Environment,'' J. Stat. Phys. **52**, 609–626 (1988).
[4] H. Kesten, ''The Limit Distribution of Sinai's Random Walk in Random Environment,'' Physica A **138**, 299–309 (1986).
[5] K. M. Khanin, A. E. Mazel, S. B. Shlosman, and Ya. G. Sinai, ''Loop Condensation Effects in the Behavior of the Random Walks,'' The Dynkin Festschrift, Prog. Probab. **34**, 167–184 (1994).
[6] E. Bolthausen, ''A Note on the Diffusion of Directed Polymers in a Random Environment,'' Commun. Math. Phys. **123**, 529–534 (1989).
[7] R. Song and X. Y. Zhou, ''A Remark on Diffusion of Directed Polymers in Random Environments,'' J. Stat. Phys. **85**, 277–289 (1996).
[8] Ya. G. Sinai, ''A Remark Concerning Random Walks with Random Potentials,'' Fund. Math. **147**, 173–180 (1995).
[9] J. G. Conlon and P. A. Olsen, ''A Brownian Motion Version of the Directed Polymer Problem,'' J. Stat. Phys. **84**, 415–454 (1996).
[10] L. Coyle, ''A Continuous Time Version of Random Walks in a Random Potential,'' Stochastic Process. Appl. **64**, 209–235 (1996).
[11] D. Fisher, D. A. Huse, and C. L. Henley, ''Huse, Henley, and Fisher Respond,'' Phys. Rev. Lett. **55**, 2924 (1985).
[12] M. Kardar, ''Roughening By Impurities At Finite Temperatures,'' Phys. Rev. Lett. **55**, 2923 (1985).
[13] M. Kardar and Y.-C. Zhang, ''Scaling of Directed Polymers in Random Media,'' Phys. Rev. Lett. **58**, 2087–2090 (1987).

[14] D. Fisher and D. Huse, ''Directed Paths in a Random Potential,'' Phys. Rev. B **43**, 10728–10742 (1991).
[15] Y.-C. Zhang, ''Ground State Instability of a Random System,'' Phys. Rev. Lett. **59**, 2125–2128 (1987).
[16] J. Krug and H. Spohn, ''Kinetic Roughening of Growing Surfaces,'' in *Solids Far from Equilibrium*, edited by G. Godreche (Cambridge University Press, Cambridge, 1991), pp. 479–582.
[17] B. Simon, ''The Bound States of Weakly Coupled Schrödinger Operators in One and Two Dimensions,'' Ann. Phys. (N.Y.) **97**, 279–288 (1976).
[18] G. L. Lawler, *Intersections of Random Walks* (Birkhäuser, Boston, 1991).

Algebraic approach to scattering theory of the Fokker–Planck equation

Toru Miyazawa
Department of Physics, Gakushuin University, Mejiro, Toshima-ku, Tokyo 171, Japan

(Received 14 July 1997; accepted for publication 7 January 1998)

The one-variable Fokker–Planck equation is studied in a scattering formalism. Scattering processes are represented by paths in one-dimensional space, and the paths are treated as algebraic objects that constitute infinite-dimensional representations of SL(2, **C**). Various expressions for the scattering coefficients are derived in a systematic way by means of algebraic methods with considerations on symmetries. © *1998 American Institute of Physics.* [S0022-2488(98)02304-4]

I. INTRODUCTION

The Fokker–Planck equation, which is a fundamental equation describing fluctuations in nonlinear systems, has also come to play an important role in quantum mechanics in connection with stochastic approaches to quantization.[1] Fundamental studies of the Fokker–Planck equation are important not only for the application to nonequilibrium phenomena but also for a better understanding of quantum mechanics.

The one-variable Fokker–Planck equation describing the diffusion process in a potential $V(x)$ has the form[2]

$$\frac{\partial}{\partial t}P(x;t)=\frac{\partial^2}{\partial x^2}P(x;t)-2\frac{\partial}{\partial x}[f(x)P(x;t)] \tag{1.1}$$

with

$$f(x)=-\frac{1}{2}\frac{d}{dx}V(x). \tag{1.2}$$

We set $P(x;t)=e^{-k^2t}\phi(x)$, where the parameter k is a complex number. Then we obtain

$$-\frac{d^2}{dx^2}\phi(x)+2\frac{d}{dx}[f(x)\phi(x)]=k^2\phi(x). \tag{1.3}$$

Equation (1.3) is equivalent to the Schrödinger equation

$$-\frac{d^2\psi}{dx^2}+\left(f^2+\frac{df}{dx}\right)\psi=k^2\psi, \tag{1.4}$$

where

$$\psi(x)=e^{V(x)/2}\phi(x). \tag{1.5}$$

Inversely, a Schrödinger equation

$$-\frac{d^2}{dx^2}\psi(x)+V_S(x)\psi(x)=\epsilon\psi(x) \tag{1.6}$$

can be converted into the Fokker–Planck equation (1.3) with

 © 1998 American Institute of Physics

$$\phi(x)=\psi_0(x)\psi(x), \quad f(x)=\frac{(d/dx)\psi_0(x)}{\psi_0(x)}, \quad k^2=\epsilon-\epsilon_0, \tag{1.7}$$

where $\psi_0(x)$ and ϵ_0 are the wave function and the energy eigenvalue of the ground state satisfying

$$-\frac{d^2}{dx^2}\psi_0(x)+V_S(x)\psi_0(x)=\epsilon_0\psi_0(x). \tag{1.8}$$

From the second equation of (1.7) it follows that

$$V(x)=-2\log\psi_0(x)+\text{constant}, \quad V_S(x)=[f(x)]^2+\frac{d}{dx}f(x)+\epsilon_0. \tag{1.9}$$

Thus, with knowledge of the ground state, a problem of quantum mechanics can be described by a Fokker–Planck equation. The merit of the Fokker–Planck equation as opposed to the Schrödinger equation lies in its symmetrical structure. Because of the symmetries the Fokker–Planck equation is more suitable for theoretical studies than the Schrödinger equation, and we can gain a better insight into the the problem by studying the Fokker–Planck equation.

One of the remarkable features of the Fokker–Planck equation is that the high-energy expansion (which corresponds to the Born series) and the low-energy expansion are related by a symmetry transformation. The difference between these two expressions are merely a difference of viewpoint. The Fokker–Planck equation has a structure that does not depend on the viewpoint, and, as it turns out, such a structure is well described in terms of paths in one-dimensional space. (The meaning of the ''path'' will be explained in due course.) In order to deal with the paths it is convenient to use the language of scattering theory, as we will shortly see.

As mentioned above, it is natural to study the Fokker–Planck equation in a formalism of scattering theory.[3–7] The aim of the present paper is to provide a unified view on the scattering formalism from an algebraic standpoint. In particular, our main interest is in the structure of the scattering coefficients (the transmission and reflection coefficients). It should be stressed that the usefulness of the scattering formalism is not limited within so-called scattering problems. Almost everything can be described in the scattering formalism. For example, the Green's function can be expressed in terms of the scattering coefficients, and the asymptotic behavior of solutions can be studied through the analysis of these scattering coefficients. The eigenvalue problem can be formulated in terms of the reflection coefficients, too, as we shall see later in this section.

Incidentally, the study of the scattering coefficients is also significant in relation to theories of integrable systems.[8] The Fokker–Plank equation is equivalent to the linear problem that is used in the inverse scattering method for the modified Korteweg–de-Vries equation.[9] Linear scattering problems associated with other nonlinear integrable equations, such as the nonlinear Schrödinger equation, are also similar in structure. So a systematic theory of the scattering coefficients will also shed light on the study of integrable systems.

In this paper we study the Fokker–Planck equation in the time-independent form (1.3), or equivalently (1.4). We do not assume a specific form of the potential $V(x)$. We can deal with not only potentials that have a finite limit at $x\to\pm\infty$ but also potentials that diverge at either $x\to+\infty$ or $x\to-\infty$ or both, or even the cases where $V(x)$ diverges at some finite x. In particular, three-dimensional problems with a spherically symmetric potential, which reduce to one-dimensional problems for the radial part, are also manageable in our formalism.

We define the transmission and reflection coefficients for an interval (x_a,x_b). (The usual transmission and reflection coefficients, which are defined with regard to the behavior of solutions at infinity, are obtained by taking the limit $x_a\to-\infty, x_b\to+\infty$.) Here we give two equivalent definitions. The first definition is the one commonly used in conventional scattering theories.[10] As shown in Fig. 1, we cut out the potential within the interval (x_a,x_b), and assume that the potential is constant outside this interval. That is to say, we consider the potential

$$\overline{V}(x)\equiv\begin{cases} V(x_a), & (x<x_a),\\ V(x), & (x_a\leqslant x\leqslant x_b),\\ V(x_b), & (x_b<x),\end{cases} \tag{1.10}$$

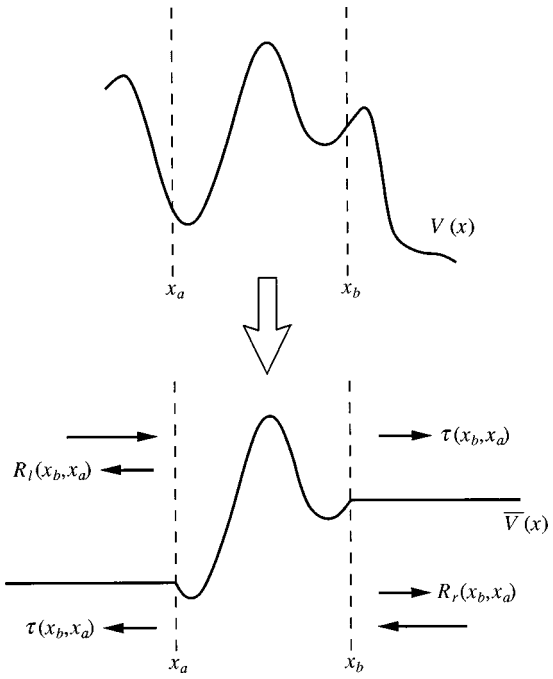

FIG. 1. Definition of the transmission and reflection coefficients for a finite interval.

instead of $V(x)$, and corresponding to (1.4) the Schrödinger equation

$$-\frac{d^2\psi}{dx^2}+\left(\overline{f}^2+\frac{d\overline{f}}{dx}\right)\psi=k^2\psi, \quad \overline{f}(x)=-\frac{1}{2}\frac{d}{dx}\overline{V}(x). \tag{1.11}$$

Since $\overline{f}=0$ outside the interval (x_a,x_b), any solution of Eq. (1.11) can be expressed as a linear combination of e^{ikx} and e^{-ikx} outside (x_a,x_b). In particular, we can construct two solutions ψ_1 and ψ_2 that have the form

$$\psi_1(x)=\begin{cases} e^{ik(x-x_a)}+R_l(x_b,x_a;k)e^{-ik(x-x_a)} & (x<x_a), \\ T(x_b,x_a;k)e^{ik(x-x_a)} & (x_b<x), \end{cases} \tag{1.12a}$$

$$\psi_2(x)=\begin{cases} T(x_b,x_a;k)e^{-ik(x-x_b)} & (x<x_a), \\ e^{-ik(x-x_b)}+R_r(x_b,x_a;k)e^{ik(x-x_b)} & (x_b<x). \end{cases} \tag{1.12b}$$

This defines the transmission coefficient T and the reflection coefficients R_r, R_l for the interval (x_a,x_b). When dealing with a finite interval, it is convenient to define

$$\tau(x_b,x_a)\equiv e^{ik(x_b-x_a)}T(x_b,x_a). \tag{1.13}$$

In this paper we use τ rather than T, and we shall call τ the transmission coefficient.

There is another way of defining the scattering coefficients, which is maybe more essential than the previous one. The transmission of a particle through the interval (x_a,x_b) is a process in which the particle travels from the point x_a to the point x_b undergoing multiple scattering by the potential. It is natural to define the transmission coefficient as the sum over all possible multiple scattering processes. This can be done as follows.[7] We define the transmission coefficient as the sum over all the diagrams such as the one shown in Fig. 2(a) with the following rules: A line element connecting the points x_1 and x_2 $(x_2 \geqslant x_1)$ represents the free propagator of the particle, $e^{ik(x_2-x_1)}$ [Fig. 3(a)]. A single scattering at position x gives a factor $+f(x)$ or $-f(x)$. The sign is plus if the particle comes to x from the right, and minus if it comes from the left [Figs. 3(b), (c)]. (Note that the particle always changes its direction with each single scattering.) The transmission coefficient $\tau(x_b,x_a)$ is defined as the sum over all the diagrams that connect the points x_a and x_b

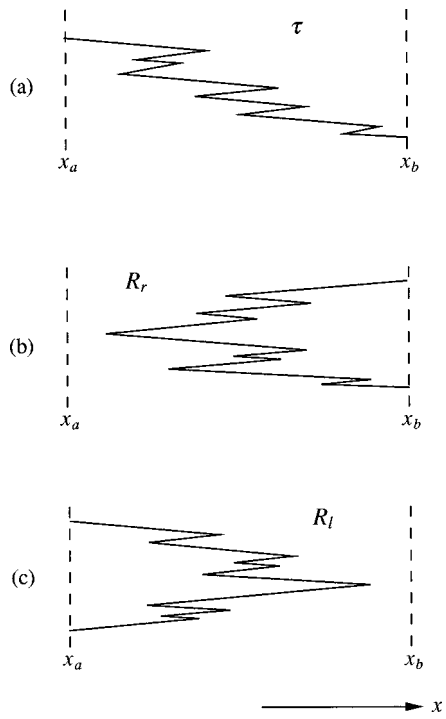

FIG. 2. The transmission and reflection coefficients are obtained as the sum over these diagrams. The rules for interpreting the diagrams are given in Fig. 3. The horizontal direction of this figure corresponds to the space coordinate. The vertical direction has no meaning.

without leaving the interval (x_a, x_b). A diagram can also be regarded as a path in one-dimensional space. We can say that the transmission coefficient $\tau(x_b, x_a)$ is the sum over all configurations of a path that penetrates the interval (x_a, x_b). Similarly, the reflection coefficients can be defined as the sum over all configurations of a path that starts from one of the endpoints and returns to that point. [Figs. 2(b), (c)] These definitions can be written as

$$\tau(x_b, x_a) = \sum_{N=0}^{\infty} (-1)^N \int dx_1 \cdots \int dx_{2N} \left[\prod_{n=1}^{2N} f(x_n) \right] \exp\left(\sum_{n=0}^{2N} ik|x_{n+1} - x_n| \right),$$

$$(x_0 \equiv x_a, x_{2N+1} \equiv x_b), \tag{1.14a}$$

$$R_r(x_b, x_a) = \sum_{N=0}^{\infty} (-1)^N \int dx_1 \cdots \int dx_{2N+1} \left[\prod_{n=1}^{2N+1} f(x_n) \right] \exp\left(\sum_{n=0}^{2N+1} ik|x_{n+1} - x_n| \right),$$

$$(x_0 \equiv x_b, x_{2N+2} \equiv x_b), \tag{1.14b}$$

$$R_l(x_b, x_a) = \sum_{N=0}^{\infty} (-1)^{N+1} \int dx_1 \cdots \int dx_{2N+1} \left[\prod_{n=1}^{2N+1} f(x_n) \right] \exp\left(\sum_{n=0}^{2N+1} ik|x_{n+1} - x_n| \right),$$

$$(x_0 \equiv x_a, x_{2N+2} \equiv x_a). \tag{1.14c}$$

The integral variables x_n $(n = 1,2,...)$ in Eqs. (1.14) correspond to the points where the scattering occurs, or in other words, the turning points of the path. Since the path changes its direction at x_n, the domain of integration for the variable x_n is

$$x_{n-1} \leqslant x_n \leqslant x_b \quad (n \text{ odd}), \quad x_a \leqslant x_n \leqslant x_{n-1} \quad (n \text{ even}) \tag{1.15a}$$

in Eqs. (1.14a) and (1.14c), and

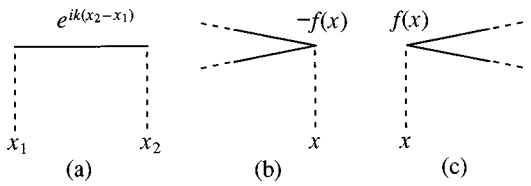

FIG. 3. Rules for the representation of scattering processes in terms of diagrams.

$$x_a \leqslant x_n \leqslant x_{n-1} \quad (n \text{ odd}), \quad x_{n-1} \leqslant x_n \leqslant x_b \quad (n \text{ even}) \tag{1.15b}$$

in Eq. (1.14b). It can be shown[3,4] that the scattering coefficients defined by (1.14) are the same as the ones defined by (1.12). We graphically express these scattering coefficients as shown in Fig. 4. These graphs represent the sum over all possible configurations of the paths in Fig. 2.

As an example of application, let us consider the eigenvalue problem. When $V(x)$ diverges (to either $+\infty$ or $-\infty$) at both $x \to +\infty$ and $x \to -\infty$ sufficiently fast (faster than linear functions of x, as a matter of fact), equation (1.3) can be interpreted as an eigenvalue problem, where k^2 is the eigenvalue of the Fokker–Planck operator $-d^2/dx^2+2(d/dx)f$. Now the quantity k is real. The eigenvalues are determined from the condition[4]

$$R_l(\infty,x_0;k)R_r(x_0,-\infty;k)=1, \tag{1.16}$$

where x_0 is arbitrary. [The value $k=0$ is an exception. Even if $k=0$ satisfies (1.16), it does not correspond to an eigenvalue unless $V(x)$ diverges to plus infinity at both $x \to +\infty$ and $x \to -\infty$.] For example, the reflection coefficients for the parabolic potential $V(x)=x^2$ [which corresponds to the Schrödinger potential $V_S(x)=x^2-1$] can be obtained as[11]

$$R_l(\infty,0;k)=R_r(0,-\infty;k)=\frac{2\Gamma\left(1-\frac{k^2}{4}\right)+ik\Gamma\left(\frac{1}{2}-\frac{k^2}{4}\right)}{2\Gamma\left(1-\frac{k^2}{4}\right)-ik\Gamma\left(\frac{1}{2}-\frac{k^2}{4}\right)}, \tag{1.17}$$

where we have chosen $x_0=0$. From (1.17) and (1.16) we can reproduce the well-known result for the eigenvalues, $k^2=2n$ $(n=0,1,2,...)$. Of course, in general it is not possible to calculate the reflection coefficients exactly. A convenient approximation method for evaluating low-lying eigenvalues is to use power series expressions of the reflection coefficients in terms of k. If we express

$$\begin{aligned} \log R_l(\infty,0;k)&=a_1k+a_3k^3+a_5k^5+\ldots, \\ \log R_r(0,-\infty;k)&=b_1k+b_3k^3+b_5k^5+\ldots \end{aligned} \tag{1.18}$$

[only odd powers of k appear in (1.18); see Sec. XIII], from (1.16) we can obtain the nth eigenvalue as the solution of

$$(a_1+b_1)k+(a_3+b_3)k^3+(a_5+b_5)k^5+\cdots=2n\pi i. \tag{1.19}$$

An example of such evaluation of eigenvalues can be found in Ref. 6. The power series expressions are also useful for analyzing the long time behavior of solutions.[4,5]

Power series expressions for the scattering coefficients such as (1.18) as functionals of the potential $V(x)$ have already been obtained in previous works.[5,6] However, the derivations were rather specific and lacked a general prospect on the problem. In particular, it was not clear why the

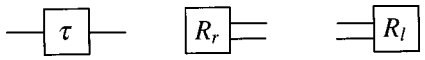

FIG. 4. Graphical representation of the scattering coefficients.

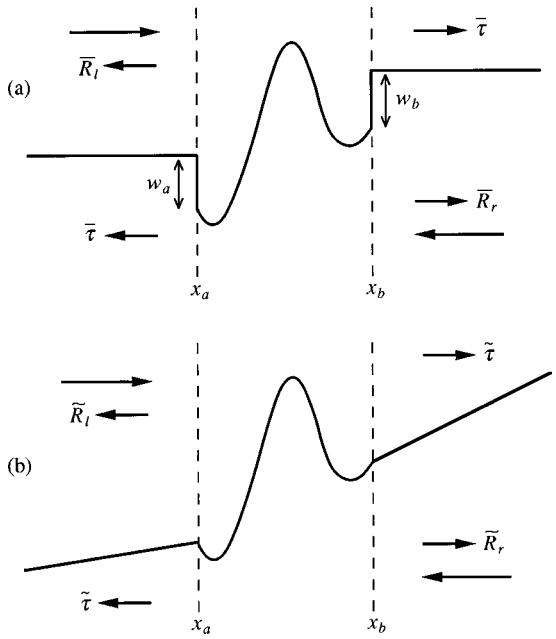

FIG. 5. Generalized scattering coefficients.

resulting expression should have the form it has. It is one of the purposes of the present work to give a general and systematic method for deriving expressions like (1.18) with a clear view on the underlying algebraic structure. This method not only reproduces previously obtained results in a transparent way but also brings forth new useful expressions.

We generalize the definition of the scattering coefficients in Sec. II. These generalized scattering coefficients play an essential role in our analysis. In Sec. III we briefly review some basic matters indispensable for the discussion in the present paper. The evolution operator is introduced here. In Sec. IV we give an intuitive interpretation of the evolution operator in terms of paths. In Sec. V we study a special representation that is related to the reflection coefficients. In Sec. VI we investigate symmetrical structures of the Fokker–Planck equation and set up our theory in a form that makes use of these symmetries. We introduce some mathematical tools in Sec. VII in order to treat the paths algebraically. Using these tools we study the structure of the scattering coefficients in Secs. VIII–XIII.

II. GENERALIZED SCATTERING COEFFICIENTS

There is some arbitrariness in the definition of scattering coefficients for a finite interval. Indeed, we can generalize the definition (1.10)–(1.13) by using, instead of (1.10),

$$\overline{V}(x)\equiv\begin{cases} V(x_a)+w_a & (x<x_a) \\ V(x) & (x_a\leqslant x\leqslant x_b), \\ V(x_b)+w_b & (x_b<x) \end{cases} \tag{2.1}$$

with arbitrary parameters w_a and w_b [Fig. 5(a)]. Since this $\overline{V}(x)$ still gives $\overline{f}=0$ outside the interval (x_a,x_b), by replacing (1.10) with (2.1) we can define a set of scattering coefficients in the same way as we did in (1.11) and (1.12). We write the transmission and reflection coefficients obtained by this generalization as $\overline{\tau}(x_b,x_a;w_b,w_a)$, $\overline{R}_r(x_b,x_a;w_b,w_a)$, and $\overline{R}_l(x_b,x_a;w_b,w_a)$. (We shall omit to write the dependence on k.) Needless to say, we have

$$\tau=\overline{\tau}(w_b=w_a=0), \quad R_r=\overline{R}_r(w_b=w_a=0), \quad R_l=\overline{R}_l(w_b=w_a=0). \tag{2.2}$$

The potential $\overline{V}$ defined by (2.1) has discontinuities at both x_a and x_b (unless $w_a=0$ or $w_b=0$). We can say that $\overline{\tau}$, $\overline{R}_r$, and $\overline{R}_l$ are the scattering coefficients including the effect of these discontinuities. This set of generalized scattering coefficients plays a crucial role in our subsequent studies.

There is still another natural way to generalize the definition of the scattering coefficients. We know that if $f(x)$ is a constant ($\equiv c$), then any solution of Eq. (1.4) can be expressed as a linear combination of $\exp[i(k^2-c^2)^{1/2}]$ and $\exp[-i(k^2-c^2)^{1/2}]$. Therefore, to define the scattering coefficients for the potential $V(x)$ within the interval (x_a,x_b) we can use

$$\overline{V}(x)\equiv\begin{cases} V(x_a)-2c_a(x-x_a) & (x<x_a)\\ V(x) & (x_a\leqslant x\leqslant x_b),\\ V(x_b)-2c_b(x-x_b) & (x_b<x)\end{cases} \tag{2.3}$$

where c_a and c_b are constants [Fig. 5(b)]. This $\overline{V}(x)$ gives $\overline{f}(x)=c_a$ for $x<x_a$ and $\overline{f}(x)=c_b$ for $x_b<x$. So Eq. (1.11) with $V(x)$ given by (2.3) has two independent solutions of the form

$$\psi_1(x)=\begin{cases} e^{ip_a(x-x_a)}+\widetilde{R}_l(x_b,x_a;c_b,c_a)e^{-ip_a(x-x_a)} & (x<x_a)\\ \sqrt{p_a/p_b}\,\widetilde{\tau}(x_b,x_a;c_b,c_a)e^{ip_b(x-x_b)} & (x_b<x),\end{cases} \tag{2.4a}$$

$$\psi_2(x)=\begin{cases} \sqrt{p_b/p_a}\,\widetilde{\tau}(x_b,x_a;c_b,c_a)e^{-ip_a(x-x_a)} & (x<x_a)\\ e^{-ip_b(x-x_b)}+\widetilde{R}_r(x_b,x_a;c_b,c_a)e^{ip_b(x-x_b)} & (x_b<x),\end{cases} \tag{2.4b}$$

with

$$p_a\equiv(k^2-c_a^2)^{1/2},\quad p_b\equiv(k^2-c_b^2)^{1/2}. \tag{2.5}$$

(Here p_a and p_b may not necessarily be real.) The factors $\sqrt{p_a/p_b}$ and $\sqrt{p_b/p_a}$ in (2.4) are now necessary in order to make the transmission coefficients $\widetilde{\tau}$ identical in both expressions (2.4a) and (2.4b). By (2.4) we can define another set of scattering coefficients, $\widetilde{\tau}$, $\widetilde{R}_r$, and $\widetilde{R}_l$, and also

$$\widetilde{T}(x_b,x_a;c_b,c_a)\equiv e^{i(p_ax_a-p_bx_b)}\widetilde{\tau}(x_b,x_a;c_b,c_a). \tag{2.6}$$

These $\widetilde{T}$, $\widetilde{R}_r$, and $\widetilde{R}_l$ are significant in connection with scattering problems (in a narrow sense) for the Schrödinger equation.[10] Let us consider the Schrödinger equation (1.6) with potential $V_S(x)$ that tends to a finite limit at both $x\to+\infty$ and $x\to-\infty$

$$\lim_{x\to-\infty}V_S(x)=V_a,\quad \lim_{x\to+\infty}V_S(x)=V_b. \tag{2.7}$$

If $V(x)$ converges to these limits sufficiently fast, the transmission and reflection coefficients T^S, R_r^S and R_l^S for this Schrödinger equation can be defined from two independent solutions of (1.6) with the asymptotic form

$$\psi_1(x)\sim\begin{cases} e^{ip_ax}+R_l^Se^{-ip_ax} & (x\to-\infty)\\ \sqrt{p_a/p_b}\,T^Se^{ip_bx} & (x\to+\infty),\end{cases}\quad \psi_2(x)\sim\begin{cases} \sqrt{p_b/p_a}\,T^Se^{-ip_ax} & (x\to-\infty)\\ e^{-ip_bx}+R_r^Se^{ip_bx} & (x\to+\infty),\end{cases} \tag{2.8}$$

where $p_a=\sqrt{\epsilon-V_a}$, $p_b=\sqrt{\epsilon-V_b}$. The wave function of the ground state behaves like e^{c_ax} and e^{c_bx} at $x\to-\infty$ and $x\to+\infty$, respectively, where $c_a=\sqrt{V_a-\epsilon_0}$, $c_b=-\sqrt{V_b-\epsilon_0}$. From the second equation of (1.7) we find that $f(x)$ behaves as $f(x)\sim c_a$ at $x\to-\infty$ and $f(x)\sim c_b$ at $x\to+\infty$. Therefore the corresponding Fokker–Planck potential $V(x)$ has the same asymptotic form as $\overline{V}(x)$ given by (2.3). Comparing (2.4) with (2.8) we find

$$T^S=\lim_{\substack{x_a\to-\infty\\ x_b\to+\infty}}\widetilde{T}(x_b,x_a;c_b,c_a), \tag{2.9a}$$

$$R_r^S=\lim_{\substack{x_a\to-\infty\\ x_b\to+\infty}} e^{-2ip_bx_b}\tilde{R}_r(x_b,x_a;c_b,c_a),\quad R_l^S=\lim_{\substack{x_a\to-\infty\\ x_b\to+\infty}} e^{2ip_ax_a}\tilde{R}_l(x_b,x_a;c_b,c_a). \tag{2.9b}$$

It is also possible to replace c_b and c_a in (2.9) by $f(x_b)$ and $f(x_a)$, respectively.

Thus we have defined two sets of generalized scattering coefficients, $(\bar{\tau},\bar{R}_r,\bar{R}_l)$ and $(\tilde{\tau},\tilde{R}_r,\tilde{R}_l)$. We shall see later how they can be described in our formalism. For the time being we work with the ''ordinary'' scattering coefficients τ, R_r, and R_l.

III. PRELIMINARIES

Here we review minimal basic materials that are necessary for our discussion. Details can be found in Refs. 7 and 4. Throughout this paper $V(x)$ denotes the potential for the Fokker–Planck equation, and $f(x)$ is its derivative multiplied by $-1/2$ [Eq. (1.2)].

We define the evolution operator $U(x,x')$, which is an element of the SL(2,**C**) Lie group, as the solution of the first-order differential equation

$$\frac{\partial}{\partial x}U(x,x')=[2ikJ_3-2f(x)J_1]U(x,x') \tag{3.1}$$

with the initial condition $U(x',x')=1$ (identity element). Here J_1 and J_3 are operators satisfying the SU(2) commutation relations

$$[J_1,J_2]=iJ_3,\quad [J_2,J_3]=iJ_1,\quad [J_3,J_1]=iJ_2 \tag{3.2a}$$

or, with the raising and lowering operators $J_\pm\equiv J_1\pm iJ_2$,

$$[J_+,J_-]=2J_3,\quad [J_3,J_\pm]=\pm J_\pm. \tag{3.2b}$$

This $U(x,x')$ is an analog of the time evolution operator in ordinary quantum mechanics, and it satisfies the product rule

$$U(x_c,x_b)U(x_b,x_a)=U(x_c,x_a). \tag{3.3}$$

The Fokker–Planck equation (1.3) is equivalent to Eq. (3.1). To see this, let us consider the fundamental representation given by

$$J_1=\frac{1}{2}\begin{pmatrix}0&-1\\-1&0\end{pmatrix},\quad J_2=\frac{1}{2}\begin{pmatrix}0&-i\\i&0\end{pmatrix},\quad J_3=\frac{1}{2}\begin{pmatrix}-1&0\\0&1\end{pmatrix}. \tag{3.4}$$

In this representation Eq. (3.1) reads

$$\frac{\partial}{\partial x}U(x,x')=\begin{pmatrix}-ik&f(x)\\f(x)&ik\end{pmatrix}U(x,x'), \tag{3.5}$$

and $U(x,x')$ is a 2×2 matrix with unit determinant. We write its elements as

$$U(x,x')=\begin{pmatrix}\alpha^+(x,x')&\beta^-(x,x')\\\beta^+(x,x')&\alpha^-(x,x')\end{pmatrix}. \tag{3.6}$$

If we define

$$\phi^+(x,x')\equiv\alpha^+(x,x')+\beta^+(x,x'),\quad \phi^-(x,x')\equiv\alpha^-(x,x')+\beta^-(x,x'), \tag{3.7}$$

then it is easy to see from (3.5) that ϕ^+ and ϕ^-, as functions of x, are two independent solutions of (1.4). We can easily show that the scattering coefficients are expressed in terms of these matrix elements as

$$\tau(x_b,x_a)=\frac{1}{\alpha^+(x_b,x_a)}, \tag{3.8a}$$

$$R_r(x_b,x_a)=\frac{\beta^+(x_b,x_a)}{\alpha^+(x_b,x_a)}, \quad R_l(x_b,x_a)=-\frac{\beta^-(x_b,x_a)}{\alpha^+(x_b,x_a)}. \tag{3.8b}$$

We can obtain a different picture by writing Eq. (3.1) in a different representation. Let us consider the boson representation[12]

$$J_+=-\tfrac{1}{2}a^\dagger a^\dagger, \quad J_-=\tfrac{1}{2}aa, \quad J_3=\tfrac{1}{4}(aa^\dagger+a^\dagger a)=\tfrac{1}{2}a^\dagger a+\tfrac{1}{4}, \tag{3.9}$$

where a and $a^\dagger$ are boson creation-annihilation operators satisfying $[a,a^\dagger]=1$. Now Eq. (3.1) takes the form

$$\frac{\partial}{\partial x}U(x,x')=\left[ik\left(aa^\dagger+\frac{1}{2}\right)-\frac{1}{2}f(x)(aa-a^\dagger a^\dagger)\right]U(x,x'). \tag{3.10}$$

It is easy to interpret the meaning of Eq. (3.10) by comparing it with Fig. 2 or Fig. 3. In this description the paths in Fig. 2 are interpreted as consisting of propagation lines of the boson, and the scattering is described as pair creation or pair annihilation of bosons. Obviously the scattering coefficients can be written as

$$\tau(x,x')=\frac{\langle 0|aU(x,x')a^\dagger|0\rangle}{\langle 0|U(x,x')|0\rangle}, \tag{3.11a}$$

$$R_r(x,x')=\frac{\langle 0|aaU(x,x')|0\rangle}{\langle 0|U(x,x')|0\rangle}, \quad R_l(x,x')=\frac{\langle 0|U(x,x')a^\dagger a^\dagger|0\rangle}{\langle 0|U(x,x')|0\rangle}, \tag{3.11b}$$

where $|0\rangle$ denotes the vacuum state. The quantity $\langle 0|U|0\rangle$ in the denominator is necessary in order to cancel out disconnected bubble diagrams.

We can use Eq. (3.1) in any other representation. In any representation the evolution operator can be expressed in terms of the scattering coefficients as

$$U(x,x')=e^{-R_r(x,x')J_+}[\tau(x,x')]^{2J_3}e^{R_l(x,x')J_-} \tag{3.12}$$

[where $\tau^{2J_3}\equiv e^{(2\log\tau)J_3}$]. We can also rewrite the expressions (3.8) or (3.11) in a form that holds in any representation. We define an inner product, which we shall write as $(\cdots,\cdots)$, that have the following properties:

$$(u,v)=(v,u)^*, \quad (u,J_3v)=(J_3u,v), \quad (u,J_\pm v)=-(J_\mp u,v). \tag{3.13}$$

Here u and v are arbitrary vectors in the representation space. This inner product need not satisfy the positivity condition $(u,u)\geqslant 0$. We assume that there exists a lowest state vector u_0 in the representation space. It is the state annihilated by the lowering operator:

$$J_-u_0=0. \tag{3.14}$$

Then the reflection coefficients can be expressed in the representation-independent form as

$$R_r=\frac{1}{2}\frac{(u_0,J_-Uu_0)}{(u_0,J_3Uu_0)}, \quad R_l=-\frac{1}{2}\frac{(u_0,UJ_+u_0)}{(u_0,UJ_3u_0)}, \tag{3.15}$$

which is the generalized form of (3.8b) or (3.11b).

For the most part we shall deal with irreducible representations with a definite ''spin,'' where the lowest state is an eigenstate of J_3. Namely, $J_3u_0=-ju_0$, where j is the spin index. If j is not a positive integer or half-odd-integer, the representation space is infinite dimensional. We define $\nu\equiv -2j$, and we shall label the representation by this ν. For example, we write

$$J_3^{(\nu)}u_0^{(\nu)}=\frac{\nu}{2}\,u_0^{(\nu)}, \tag{3.16}$$

and we express the inner product in this representation as $(u,v)^{(\nu)}$. We use ν as the label[13] rather than j, since we will mainly deal with infinite-dimensional representations with $\nu\geqslant 0$.

Now let us explicitly write out expressions for the representation labeled by ν. We choose the basis such that the representation matrices of $J_\pm$ and J_3 have the form

$$J_+^{(\nu)}=\begin{pmatrix} 0 & & & & & \\ -1 & 0 & & & & \\ & -2 & 0 & & & \\ & & -3 & 0 & & \\ & & & \ddots & \ddots & \end{pmatrix}, \qquad J_-^{(\nu)}=\begin{pmatrix} 0 & \nu & & & & \\ & 0 & \nu+1 & & & \\ & & 0 & \nu+2 & & \\ & & & 0 & \ddots & \\ & & & & \ddots & \end{pmatrix}, \tag{3.17a}$$

$$J_3^{(\nu)}=\begin{pmatrix} \frac{\nu}{2} & & & & \\ & \frac{\nu}{2}+1 & & & \\ & & \frac{\nu}{2}+2 & & \\ & & & \ddots & \end{pmatrix}. \tag{3.17b}$$

In this basis the symmetry relation between $J_+^{(\nu)}$ and $J_-^{(\nu)}$ is expressed as

$$M^{(\nu)}J_+^{(\nu)}(M^{(\nu)})^{-1}=-(J_-^{(\nu)})^T, \tag{3.18}$$

where $M^{(\nu)}$ is the diagonal matrix whose (n,n) element is

$$(M^{(\nu)})_{nn}=\frac{\nu(\nu+1)\cdots(\nu+n-1)}{n!}=\binom{\nu+n-1}{n}, \quad (n=0,1,2,\ldots). \tag{3.19}$$

[For $n=0$ we have $\binom{\nu-1}{0}=1$.] The inner product satisfying (3.13) can be defined using this $M^{(\nu)}$ as

$$(v,u)^{(\nu)}\equiv(v^*)^TM^{(\nu)}u, \tag{3.20}$$

where u and v are column vectors, and T denotes transposition. The lowest state vector satisfying (3.14) is the column vector $u_0^{(\nu)}=(1,0,0,\ldots)^T$. The representation space is spanned by the basis $u_n^{(\nu)}$ $(n=0,1,2,\ldots)$, where

$$u_n^{(\nu)}=(-1)^n\,\frac{1}{n!}\,[J_+^{(\nu)}]^nu_0^{(\nu)}=(\overbrace{0,\ldots,0}^{n},1,0,\ldots)^T. \tag{3.21}$$

In this representation, the matrix elements of $U^{(\nu)}=e^{-R_rJ_+^{(\nu)}}\tau^{2J_3^{(\nu)}}e^{R_lJ_-^{(\nu)}}$ [Eq. (3.12)] are obtained as

$$(U^{(\nu)})_{mn}=\frac{1}{\binom{\nu+m-1}{m}}\,(u_m^{(\nu)},U^{(\nu)}u_n^{(\nu)})^{(\nu)}=\sum_{i=0}^{\min(m,n)}\binom{m}{i}\binom{\nu+n-1}{n-i}R_r^{m-i}R_l^{n-i}\tau^{\nu+2i} \tag{3.22}$$

$(m,n=0,1,2,\ldots)$. Note that the binomial coefficients are defined as in Eq. (3.19). In particular, since $(U^{(\nu)})_{00}=\tau^\nu$ we have, for any ν,

$$\tau=[(u_0^{(\nu)},U^{(\nu)}u_0^{(\nu)})^{(\nu)}]^{1/\nu}. \tag{3.23}$$

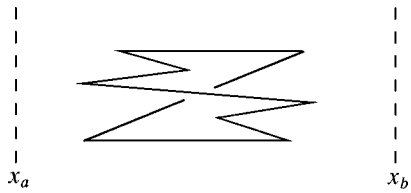

FIG. 6. The quantity $\log Z$ is obtained as the sum over connected loop diagrams like this one.

The representation given by the matrices in (3.4) is the spin-1/2 representation, and hence $\nu=-1$. The matrix U in Eqs. (3.5) and (3.6) is $U^{(-1)}$ in our present notation. From (3.6), (3.19), and (3.20) it is obvious that $(u_0^{(-1)}, U^{(-1)}u_0^{(-1)})^{(-1)}=\alpha^+$, and thus we can see that (3.8a) is obtained from (3.23) by setting $\nu=-1$.

The boson representation given by (3.9) is reducible. The spaces with even number and odd number of bosons form two invariant subspaces, each corresponding to $\nu=1/2$ and $\nu=3/2$, respectively. In fact, if we choose as the basis

$$u_n^{(1/2)}=\frac{1}{n!}\left(\frac{1}{2}\,a^\dagger a^\dagger\right)^n|0\rangle, \quad u_n^{(3/2)}=\frac{1}{n!}\left(\frac{1}{2}\,a^\dagger a^\dagger\right)^n a^\dagger|0\rangle, \tag{3.24}$$

the matrix expression of the operators (3.9) is given by (3.17) with $\nu=1/2$ and $\nu=3/2$, respectively. The inner product (3.20) is equivalent to the ordinary inner product in the boson Fock space. Obviously Eqs. (3.11b) are a special case of (3.15). The relation (3.23) with $\nu=1/2$ and $\nu=3/2$, respectively, can be written in the boson representation as

$$\tau^{1/2}=\langle 0|U|0\rangle, \quad \tau^{3/2}=\langle 0|aUa^\dagger|0\rangle. \tag{3.25}$$

We can reproduce (3.11a) from (3.25).

In the boson representation we define the vacuum amplitude Z as

$$Z(x,x')\equiv e^{-(1/2)ik(x-x')}\langle 0|U(x,x')|0\rangle. \tag{3.26}$$

It is the sum over all bubble diagrams. [The factor $e^{-(1/2)ik(x-x')}$ cancels the evolution caused by the term $\frac{1}{2}ikU$ on the right-hand side of (3.10).] From (3.25) and (3.26) we obtain a relation between the transmission coefficient and the vacuum amplitude,

$$\tau(x,x')=e^{ik(x-x')}[Z(x,x')]^2, \tag{3.27}$$

and therefore

$$\log\,\tau(x,x')=2\,\log\,Z(x,x')+ik(x-x'). \tag{3.28}$$

We remark that $\log Z$ is obtained as the sum over all connected loop diagrams, while Z includes disconnected diagrams. (See Fig. 6. Note that Z and $\log Z$, respectively, correspond to the partition function and the free energy in statistical mechanics.) Thus the quantity $\log \tau(x,x')$, apart from $ik(x-x')$, is obtained as the sum over all configurations of a connected closed path. Because of the factor 2 in front of $\log Z$ on the right-hand side, we regard mirror images as different diagrams. We shall see later that the evolution of $\log \tau$ is described by the $\nu=0$ representation.

IV. DIAGRAMMATIC INTERPRETATION

The expressions given in the last section are valid for any real values of ν, no matter whether it is positive or negative, integer or noninteger. However, in order to describe the scattering coefficients, representations with a positive integer ν are of particular importance. For example, the transmission coefficient τ is the (0,0) element of the evolution matrix for $\nu=1$. So in order to study the transmission coefficient it is natural to deal with the $\nu=1$ representation.

When ν is a positive integer, the evolution operator can be given an intuitive interpretation in terms of diagrams. Let us sketch out the basic idea. The transmission coefficient $\tau(x_2,x_1)$ is the

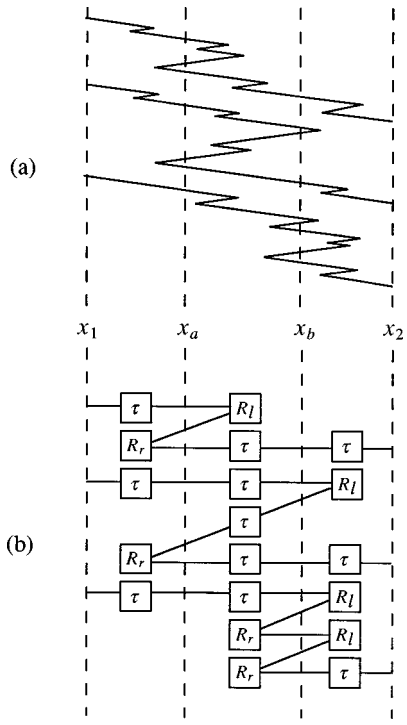

FIG. 7. (a) A diagram contributing to τ^3. (b) The diagram in (a) is included in this graph.

sum over all possible configurations of a path that connects x_1 and x_2. Therefore the quantity $(U^{(\nu)})_{00}=\tau^\nu$ can be understood as the sum over configurations of ν such paths. In other words, it is the sum over all such diagrams as the one shown in Fig. 7(a). Suppose that we dissect such a diagram at two points x_a and x_b $(x_1<x_a<x_b<x_2)$. The matrix elements of $U^{(\nu)}(x_b,x_a)$ correspond to the fragments that appear by the dissection. [See Fig. 7(b).] The precise meaning of this statement will soon become clear.

Before discussing the evolution operator, let us give a graphical interpretation to the basis given by Eq. (3.21) and the inner product defined by (3.20). We express the basis vector $u_n^{(\nu)}$ graphically as shown in Fig. 8. It consists of ν objects with one external leg (open circles) and n objects with two legs (solid circles). Its transpose $(u_n^{(\nu)})^T$ is expressed in a similar way with the legs pointing to the opposite direction. The inner product of two basis vectors $(u_m^{(\nu)},u_n^{(\nu)})^{(\nu)}$ is interpreted as the number of ways in which the legs of $u_n^{(\nu)}$ and $(u_m^{(\nu)})^T$ can be connected without creating a loop. If $m\neq n$, such a connection is impossible since the legs do not match. If $m=n$, there are $\binom{\nu+n-1}{n}$ possible ways of doing so, which is the number of ways to divide the n solid circles among the ν open circles (Fig. 9). Thus we can reproduce

$$(u_m^{(\nu)},u_n^{(\nu)})^{(\nu)}=(u_m^{(\nu)})^T M^{(\nu)} u_n^{(\nu)}=\binom{\nu+n-1}{n}\delta_{mn}, \tag{4.1}$$

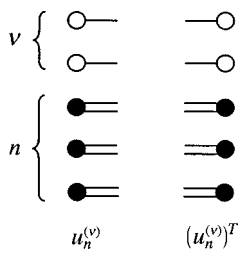

FIG. 8. Graphical representation of $u_n^{(\nu)}$ and its transpose.

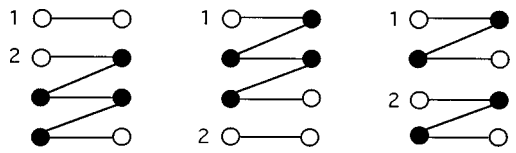

FIG. 9. Graphical interpretation of the inner product $(u_n^{(\nu)},u_n^{(\nu)})^{(\nu)}=\binom{\nu+n-1}{n}$. In this example, $\nu=2$, $n=2$, and $\binom{\nu+n-1}{n}=3$. As shown here, there are three possible ways of connecting the legs of $u_2^{(2)}$ and $(u_2^{(2)})^T$. The open circles on the left side (1 and 2 in this figure) are treated as distinct. The open circles on the right side are not distinguished. (It would be double counting if we distinguished them as well.)

which is the inner product defined by (3.20) with (3.19). As shown in Fig. 10, the matrix $M^{(\nu)}$ stands for the reshuffling of the legs. We define the conjugate of a vector v as

$$v^\dagger \equiv (v^*)^T M^{(\nu)}. \tag{4.2}$$

Namely, $v^\dagger$ includes the rearrangements of the legs. In particular, the conjugate of a basis vector $u_n^{(\nu)}$ is $(u_n^{(\nu)})^\dagger=(M^{(\nu)})_{nn}(u_n^{(\nu)})^T=\binom{\nu+n-1}{n}(u_n^{(\nu)})^T$. As can be seen from Fig. 10, this $(u_n^{(\nu)})^\dagger$ corresponds to $\binom{\nu+n-1}{n}$ graphs with different arrangement of the legs. Generally, graphs representing a column vector have only right legs, and graphs corresponding to a row vector have only left legs. Taking the inner product amounts to connecting the legs so as to create ν paths without a closed loop.

Now we can give a graphical interpretation to the evolution operator $U^{(\nu)}$. Let us first consider the simplest case, $\nu=1$. For $\nu=1$ the matrix $M^{(\nu)}$ is the identity matrix and we have $(u_m^{(1)},u_n^{(1)})^{(1)}=\delta_{mn}$. Namely, there is only one possible way of connecting the legs of $u_n^{(1)}$ and $(u_n^{(1)})^T$, and hence $(u_n^{(1)})^\dagger$ has only one possible arrangement of the legs. The (m,n) component of the matrix $U^{(1)}$ is obtained as

$$(U^{(1)})_{mn}=(u_m^{(1)})^\dagger U^{(1)} u_n^{(1)}. \tag{4.3}$$

Corresponding to the graphical interpretation of the inner product explained above, we can interpret the matrix element $(U^{(1)})_{mn}$ as follows: The matrix element $(U^{(1)})_{mn}$ is the sum of all the diagrams with $2n+1$ left legs and $2m+1$ right legs, such that one connected path is obtained when its left and right legs are connected with the graph of $u_n^{(1)}$ and $(u_m^{(1)})^\dagger$, respectively, as in Fig. 11. Here a diagram means the one like Fig. 2 or Fig. 7(a) with the rules given by Fig. 3. [A ''left leg'' of a diagram means an endpoint on the left side. For example, the diagram shown in

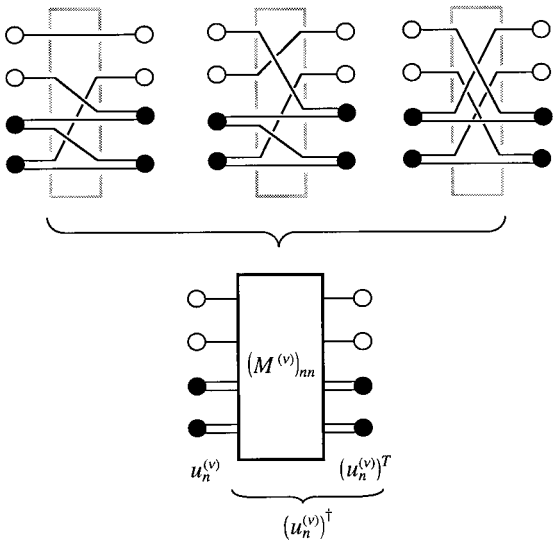

FIG. 10. We reexpress Fig. 9 like this, placing the open circles at the top. Now the matrix element $(M^{(\nu)})_{nn}=\binom{\nu+n-1}{n}$ can be graphically interpreted as representing the $\binom{\nu+n-1}{n}$ different ways of rearranging the legs.

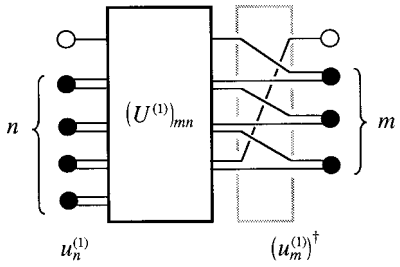

FIG. 11. The matrix element $(U^{(1)})_{mn}$ is the sum of all the diagrams that make a path connecting the two open circles when inserted in the box shown here.

Fig. 7(a) has three left legs and three right legs for the whole interval (x_1,x_2), and the part within the interval (x_a,x_b) has seven left legs and nine right legs.] The endpoints of the resulting path are the two open circles, and there should be no closed loop. An example is shown in Fig. 12.

Let us now consider general positive integer ν. We have

$$(U^{(\nu)})_{mn}=\frac{1}{\binom{\nu+m-1}{m}}\,(u_m^{(\nu)},U^{(\nu)}u_n^{(\nu)})^{(\nu)}=\frac{1}{\binom{\nu+m-1}{m}}\,(u_m^{(\nu)})^\dagger U^{(\nu)}u_n^{(\nu)}. \tag{4.4}$$

The graphical interpretation of $(U^{(\nu)})_{mn}$ given above for $\nu=1$ can be naturally extended to general ν as follows: The matrix element $(U^{(\nu)})_{mn}$ is the sum of all the diagrams with $2n+\nu$ left legs and $2m+\nu$ right legs, such that ν paths are obtained when its left and right legs are connected with $u_n^{(\nu)}$ and $(u_m^{(\nu)})^\dagger$, respectively, just like Fig. 11. Each of the ν paths should connect an open circle on the left side and one on the right side.

The above statement might be somewhat ambiguous for $\nu\geqslant 2$, since $(u_m^{(\nu)})^\dagger$ includes $\binom{\nu+m-1}{m}$ possible ways of arranging the legs. Let us explain this with the example shown in Fig. 13 ($\nu=2$, $n=1$, $m=2$). The graphical expression of $(u_2^{(2)})^\dagger$ has three possible arrangements of the legs (A,B,C in Fig. 13; see also Fig. 10). We arbitrarily choose one of these arrangements, say A. Then there are four distinct ways of making two connected paths, as shown in Fig. 13. Different paths are obtained by replacing A in Fig. 13 by B or C, but we do not count them. [This is what the factor $1/\binom{\nu+m-1}{m}$ in Eq. (4.4) means.] For example, the paths in Fig. 14(a) should not be counted as a different contribution since they are identical to the ones in Fig. 14(b), which are

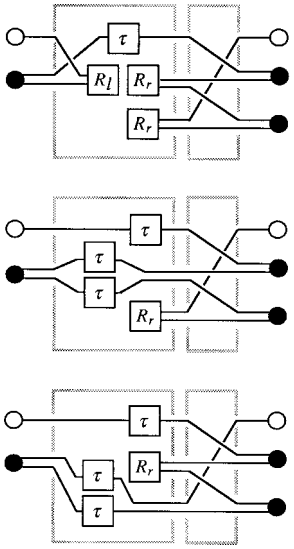

FIG. 12. An example: $\nu=1$, $n=1$, $m=2$. As shown here, there are three different ways of making a connected path. These three paths add up to $\tau R_r^2 R_l$, $\tau^3 R_r$, and $\tau^3 R_r$, respectively. Hence we find $(U^{(1)})_{21}=\tau R_r^2 R_l+2\tau^3 R_r$.

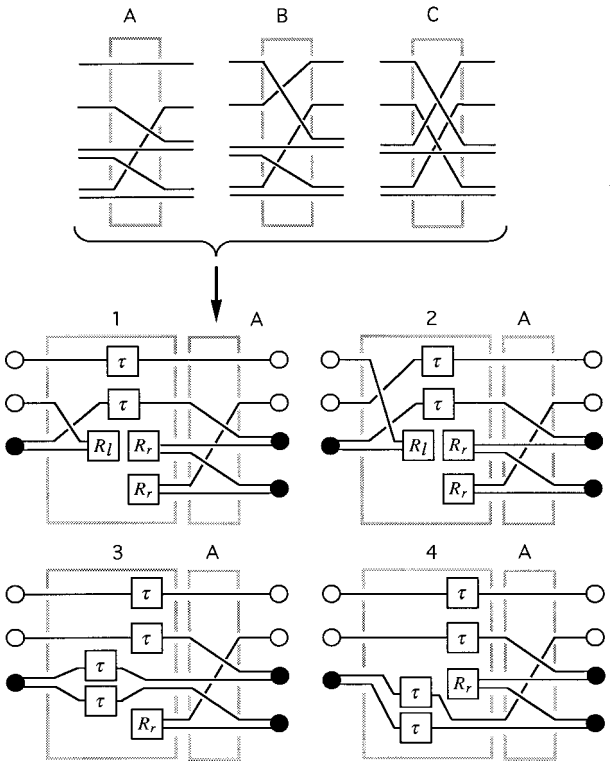

FIG. 13. An example: $\nu=2$, $m=2$, $n=1$. There are four distinct ways of making two connected paths, each adding up to $\tau^2R_r^2R_l$, $\tau^2R_r^2R_l$, τ^4R_r, and τ^4R_r. Hence we find $(U^{(2)})_{21}=2\tau^2R_r^2R_l+2\tau^4R_r$. (Remember that the open circles on the left side are distinguished, so the graphs 1 and 2 in this figure are considered different. See also Figs. 14 and 15.)

obtained from graph 4 of Fig. 13 with arrangement B. Thus, in this example we obtain $(U^{(2)})_{21}=2\tau^2R_r^2R_l+2\tau^4R_r$. It is also possible to derive the general result (3.22) from diagrammatic considerations. Details are given in Appendix A.

We can now graphically interpret the product rule (3.3). As shown in Fig. 15, the multiplication of $U^{(\nu)}(x_c,x_b)$ and $U^{(\nu)}(x_b,x_a)$ is achieved in the graphical representation simply by connecting their legs at x_b. The product rule describes how a path in the interval (x_c,x_a) can be decomposed into fragments in the intervals (x_c,x_b) and (x_b,x_a).

Thus we have found that the representation labeled by ν (positive integer) describes the evolution of ν paths. Paths penetrating a larger interval can be obtained by connecting segments in smaller intervals, and that is what manifests itself algebraically as the property of a group.

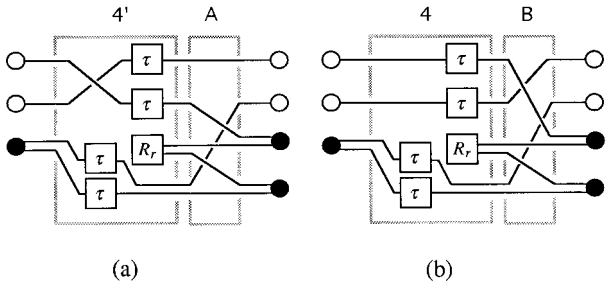

FIG. 14. The paths in (a) and (b) are identical, and so 4 and 4′ should not be regarded as distinct contributions.

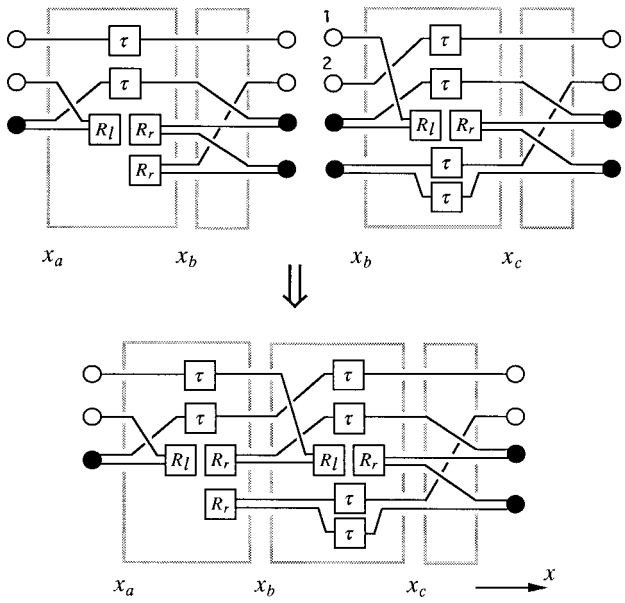

FIG. 15. Graphical explanation of $U^{(\nu)}(x_c,x_a)=U^{(\nu)}(x_c,x_b)U^{(\nu)}(x_b,x_a)$ for $\nu=2$. This figure also explains why it is necessary to distinguish the open circles on the left side. It is obvious that different paths are obtained by interchanging the open circles 1 and 2 in this figure.

V. THE $\nu=0$ REPRESENTATION

Next let us consider the case $\nu=0$. What we mean by ''the $\nu=0$ representation'' is not the trivial scalar representation but the infinite-dimensional representation in which the operators $J_{\pm}$ and J_3 are represented by the infinite-dimensional matrices obtained by setting $\nu=0$ in (3.17). [In fact, this is nothing but the standard representation of the global conformal group, which is given by the operators (7.1) in Sec. VII.] In this representation the evolution operator becomes the matrix

$$U^{(0)}=e^{-R_r J_+^{(0)}}\tau^{2J_3^{(0)}}e^{R_l J_-^{(0)}}=\begin{pmatrix} 1 & 0 & 0 & \cdots \\ R_r & \tau^2 & \tau^2 R_l & \cdots \\ R_r^2 & 2\tau^2 R_r & 2\tau^2 R_r R_l+\tau^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{5.1}$$

Obviously this representation is not irreducible. The vector $u_0^{(0)}=(1,0,0,...)^T$ spans a one-dimensional invariant subspace, which gives the trivial representation. It is a matter of course since the spin-0 irreducible representation is the scalar representation. It is important to consider the infinite-dimensional representation given by (5.1) because the reflection coefficients belong to this representation. The matrix elements of $U^{(0)}$ are

$$(U^{(0)})_{mn}=\sum_{i=0}^{\min(m,n)}\binom{m}{i}\binom{n-1}{n-i}R_r^{m-i}R_l^{n-i}\tau^{2i}, \tag{5.2}$$

and we have $R_r=(U^{(0)})_{10}$. The left reflection coefficient R_l is not obtained as a matrix element of $U^{(0)}$. To deal with R_l we use the matrix $U^{(0)L}$ defined by

$$U^{(0)L}\equiv(M^{(0)})^{-1}U^{(0)}M^{(0)}=\lim_{\epsilon\to 0}(M^{(\epsilon)})^{-1}U^{(\epsilon)}M^{(\epsilon)}. \tag{5.3}$$

Its matrix elements are $(U^{(0)L})_{mn}=(n/m)(U^{(0)})_{mn}$ for $m\neq 0$ and $(U^{(0)L})_{0n}=R_l^n$, so we have $R_l=(U^{(0)L})_{01}$.

For $\nu=0$ the inner product defined by (3.20) seems meaningless, since $(u_n^{(0)},u_n^{(0)})^{(0)}=0$ unless $n=0$. We can define a meaningful inner product for $\nu=0$ by an appropriate change of

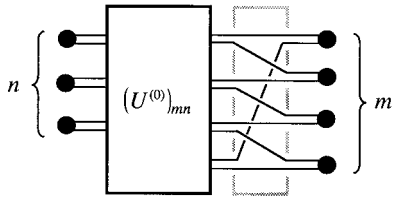

FIG. 16. The matrix element $(U^{(0)})_{mn}$ is the sum of all the diagrams that make one connected loop when inserted in the box shown here.

scale. Let us set $\nu=\epsilon$ and take the limit $\epsilon\to 0$. To the lowest order in ϵ we have $(u_m^{(\epsilon)},u_n^{(\epsilon)})^{(\epsilon)}=(\epsilon/m)\delta_{mn}$. Dividing this by ϵ before taking the limit $\epsilon\to 0$, we can define the rescaled inner product as

$$(u_{m,}^{(0)},u_n^{(0)})^0\equiv\lim_{\epsilon\to 0}\frac{1}{\epsilon}(u_m^{(\epsilon)},u_n^{(\epsilon)})^{(\epsilon)}=\frac{1}{m}\delta_{mn}. \tag{5.4}$$

Then $(u_n^{(0)},u_n^{(0)})^0$ has a finite value unless $n=0$. (It is divergent for $n=0$.) From (4.4) and (5.4) we obtain the expression for the matrix elements as

$$(U^{(0)})_{mn}=m(u_m^{(0)},U^{(0)}u_n^{(0)})^0. \tag{5.5}$$

In particular, we have $R_r=(u_1^{(0)},U^{(0)}u_0^{(0)})^0=(u_0^{(0)},J_-^{(0)}U^{(0)}u_0^{(0)})^0$. In a similar way, the left reflection coefficient can be written as $R_l=-(u_0^{(0)},U^{(0)}J_+^{(0)}u_0^{(0)})^0$.

In the graphical representation the vector $u_n^{(0)}$ and its transpose $(u_n^{(0)})^T$ have the form shown in Fig. 8 without open circles. Obviously it is impossible to connect their legs without making a loop. If we try to make a connected path out of these things, it inevitably becomes a loop. Hence we can guess that this $\nu=0$ representation is related to a closed path, not paths with a beginning and an end. Indeed, the rescaled inner product (5.4) can be interpreted as the operation of making a loop by connecting the legs of $(u_m^{(0)})^T$ and $u_n^{(0)}$.

The graphical interpretation of the matrix elements for $\nu=0$ is as follows. The matrix element $(U^{(0)})_{mn}$ $(m\neq 0)$ is the sum of all the diagrams with $2n$ left legs and $2m$ right legs, such that one closed path (i.e., a loop) is obtained when the left and the right legs are connected with $u_n^{(0)}$ and $(u_m^{(0)})^T$, respectively, as in Fig. 16. Note that the legs of $(u_m^{(0)})^T$ are rearranged so as to make a loop when connected with $u_m^{(0)}$. We choose one of the solid circles on the right side (say the uppermost one) as the reference point, and distinguish the paths keeping track of this point. [This is what the factor m in (5.5) means.] Examples are shown in Fig. 17. Note that A, B, and C in Fig. 17(a) are identical except for the reference point. [So are E and F in Fig. 17(b), too.] If we had chosen one of the solid circles on the left side as the reference point, we would have obtained an element of $U^{(0)L}$ instead of $U^{(0)}$.

It is possible to give a diagrammatic interpretation to the group property for $\nu=0$, too. As shown in Fig. 18, a loop in a larger interval can be made up of loops in smaller intervals. This is what the product rule (3.3) means for $\nu=0$. From Fig. 18 we can also see why it is necessary to specify the reference point; graphs with different positions of the reference point make different loops in a larger interval. The reference point is needed only in the process of making a larger loop. It is no longer necessary once the loop is completed. In order to disregard the reference point, we have to divide the final result by m. This is the meaning of the factor $1/m$ in (5.4).

Thus the $\nu=0$ representation describes the evolution of a closed path. This is consistent with the fact that, as we saw in Sec. III, the quantity $\log\tau$ corresponds to a closed path.

VI. SYMMETRIES OF THE FOKKER–PLANCK EQUATION AND A COVARIANT FORMULATION

The Fokker–Planck equation is characterized by its symmetrical structure. Let us rewrite the evolution equation (3.1) to make its symmetries more transparent. We rescale the space coordinate and the potential as

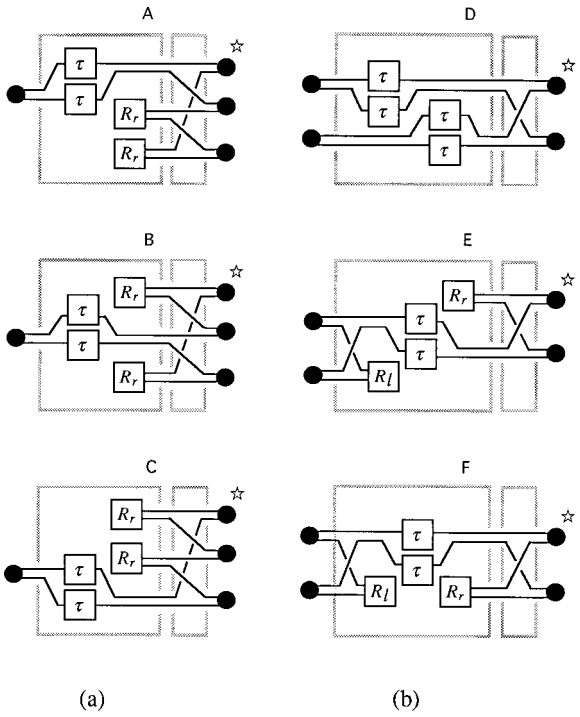

FIG. 17. Examples: (a) $m=3$, $n=1$, (b) $m=2$, $n=2$. The reference point is marked with a star. Hence we have $(U^{(0)})_{31}=3\tau^2R_r^2$ and $(U^{(0)})_{22}=\tau^4+2\tau^2R_rR_l$.

$$X\equiv ikx, \quad Y\equiv \tfrac{1}{2}V(x). \tag{6.1}$$

Then Eq. (3.1) can be written as

$$dU=2[(dX)J_3+(dY)J_1]U. \tag{6.2}$$

It is easy to see that Eq. (6.2) and the commutation relations (3.2) are invariant under the transformation

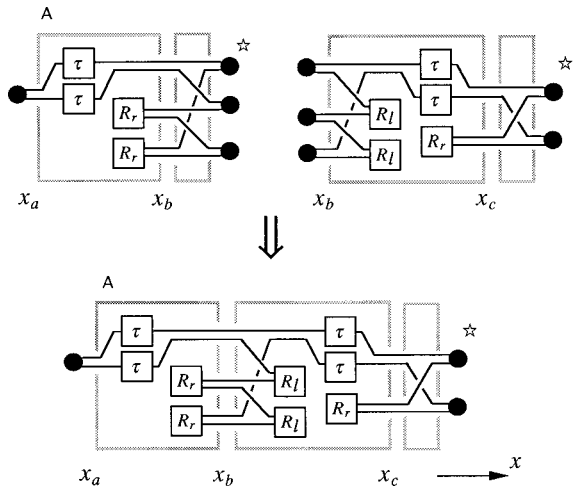

FIG. 18. Graphical interpretation of $U^{(0)}(x_c,x_a)=U^{(0)}(x_c,x_b)U^{(0)}(x_b,x_a)$. A closed loop for a larger interval (x_a,x_c) is obtained by simply connecting the legs of $U^{(0)}$ for smaller intervals (x_a,x_b) and (x_b,x_c). Note that a different loop is obtained if we replace A in this figure by B or C of Fig. 17(a). This is the reason why it is necessary to specify the reference point.

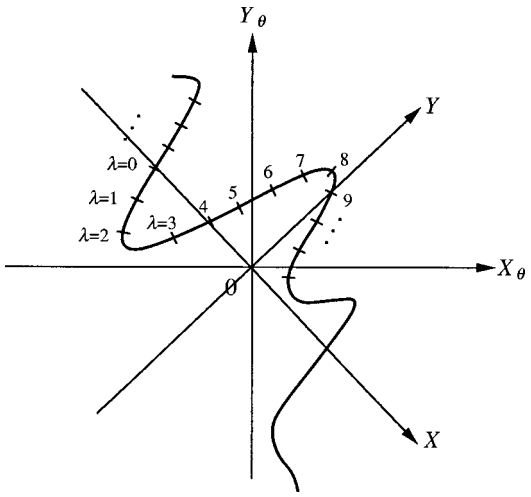

FIG. 19. Rotation of the coordinate axes and parameterization of the potential curve.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \to \begin{pmatrix} X_\theta \\ Y_\theta \end{pmatrix} \equiv \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \tag{6.3a}$$

$$\begin{pmatrix} J_3 \\ J_1 \\ J_2 \end{pmatrix} \to \begin{pmatrix} J_{3,\theta} \\ J_{1,\theta} \\ J_{2,\theta} \end{pmatrix} \equiv \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} J_3 \\ J_1 \\ J_2 \end{pmatrix}, \tag{6.3b}$$

where θ is an arbitrary complex number. If θ is real and k is pure imaginary, the transformation (6.3a) amounts to pictorially rotating the graph of the potential (Fig. 19). (Since θ is not necessarily real, this transformation is not necessarily a rotation. However, for simplicity we will sometimes call it a rotation.) Any linear transformation that leaves both (6.2) and (3.2) invariant is either of the form (6.3) or obtained as a combination of (6.3) and the discrete transformation $(X,Y)\to(-X,Y)$ with $(J_3,J_1,J_2)\to(-J_3,J_1,-J_2)$, which corresponds to the reversal of the sign of k.

We define the operator $P(\theta)\equiv e^{-i\theta J_2}$, which induces the transformation (6.3b) as

$$J_{a,\theta}=P(\theta)J_aP(-\theta) \quad (a=1,2,3). \tag{6.4}$$

Obviously we have $[P(\theta)]^{-1}=P(-\theta)$. We can also express this operator as[14]

$$P(\theta)=e^{-\tan(\theta/2)J_+}[\cos(\theta/2)]^{-2J_3}e^{\tan(\theta/2)J_-}=e^{\tan(\theta/2)J_-}[\cos(\theta/2)]^{2J_3}e^{-\tan(\theta/2)J_+}. \tag{6.5}$$

Since we have $(dX)J_3+(dY)J_1=(dX_\theta)J_{3,\theta}+(dY_\theta)J_{1,\theta}$ with arbitrary θ, multiplying Eq. (6.2) by $P(-\theta)$ from the left and $P(\theta)$ from the right we obtain

$$dU_\theta=2[(dX_\theta)J_3+(dY_\theta)J_1]U_\theta, \tag{6.6}$$

where we have defined

$$U_\theta\equiv P(-\theta)UP(\theta). \tag{6.7}$$

Thus we find that the evolution equation in the rotated coordinates (X_θ,Y_θ) with arbitrary θ has locally the same form as in the original coordinates (X,Y), with U replaced by U_θ.

The relation (6.1) with a given potential $V(x)$ defines a curve in the (X_θ,Y_θ) space. (Restricting the original coordinate x to be real, it is a one-dimensional manifold in $\mathbf{C}^2$.) This curve determines the relation between X_θ and Y_θ. Since Y_θ cannot, in general, be expressed as a

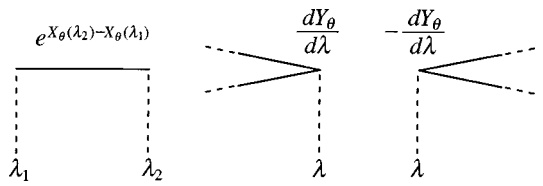

FIG. 20. Diagrammatic rules for general θ.

single-valued function of X_θ, we need to parameterize the curve as $X_\theta(\lambda)$, $Y_\theta(\lambda)$ with some real parameter λ (Fig. 19). This λ is arbitrary as long as it is monotonously increasing (or decreasing) along the curve. Using λ we can write Eq. (6.6) in the form

$$\frac{\partial}{\partial\lambda}U_\theta=2\left(\frac{dX_\theta}{d\lambda}J_3+\frac{dY_\theta}{d\lambda}J_1\right)U_\theta. \tag{6.8}$$

If $V(x)$ is a single-valued function as usually assumed [we can also deal with $V(x)$ that is not single-valued, though], we may use the original space coordinate x in place of λ.

By generalizing the diagrammatic rules explained in Sec. I, we can define the scattering coefficients in the rotated frame (X_θ,Y_θ) with arbitrary θ. Comparing (6.8) with (3.1) we can see that the diagrammatic rules in the rotated frame are obtained by the replacement

$$\frac{d}{dx}\to\frac{d}{d\lambda},\quad ikx=X\to X_\theta,\quad \frac{1}{2}V(x)=Y\to Y_\theta. \tag{6.9}$$

In diagrams for the original coordinate system (Fig. 2), the unperturbed path between x' and x was $\exp[ik(x-x')]$, and each single scattering gave a factor $\pm f$ (see Fig. 3). Now, in the rotated coordinate system, the horizontal axis x in Fig. 2 is replaced by a general parameter λ (though it is possible to use x in place of λ). According to (6.9) we have

$$e^{ik(x-x')}=e^{X(\lambda)-X(\lambda')}\to e^{X_\theta(\lambda)-X_\theta(\lambda')},\quad -f(x)=\frac{dY}{dx}\to\frac{dY_\theta}{d\lambda}. \tag{6.10}$$

Namely, the unperturbed path between points λ' and $\lambda(\lambda\geqslant\lambda')$ is $\exp[X_\theta(\lambda)-X_\theta(\lambda')]$, as shown in Fig. 20, and the factor given by each scattering is

$$\mp\frac{dY_\theta}{d\lambda}=\mp\frac{d}{d\lambda}\left[\frac{1}{2}V(x)\cos\theta-ikx\sin\theta\right]. \tag{6.11}$$

We define the scattering coefficients in the rotated frame, τ_θ, $R_{r,\theta}$, and $R_{l,\theta}$, as the sum over the diagrams similar to the ones in Fig. 2, now with the rules given in Fig. 20. Namely, τ_θ is written explicitly as

$$\tau_\theta(\lambda_b,\lambda_a)=\sum_{N=0}^{\infty}(-1)^N\int\cdots\int_{\substack{\lambda_{n-1}\leqslant\lambda_n\leqslant\lambda_b\ (n\ \text{odd})\\ \lambda_a\leqslant\lambda_n\leqslant\lambda_{n-1}\ (n\ \text{even})}}d\lambda_1\ldots d\lambda_{2N}\left[\prod_{n=1}^{2N}\frac{d}{d\lambda}Y_\theta(\lambda_n)\right]$$
$$\times\exp\left(\sum_{n=0}^{2N}(-1)^n[X_\theta(\lambda_{n+1})-X_\theta(\lambda_n)]\right). \tag{6.12}$$

($\lambda_0\equiv\lambda_a,\lambda_{2N+1}\equiv\lambda_b$.) Of course, this expression reduces to (1.14a) when we set $\theta=0$. In a similar way, the expressions for $R_{r,\theta}$ and $R_{l,\theta}$ can be obtained from (1.14b,c) by the replacement (6.10), with the integral variables x_n replaced by λ_n.

Since Eq. (6.2) is invariant under (6.3), the evolution operator given by the expression (3.12) does not change under the replacement $(\tau,R_r,R_l)\to(\tau_\theta,R_{r,\theta},R_{l,\theta})$ with $(J_3,J_+,J_-)\to(J_{3,\theta},J_{+,\theta},J_{-,\theta})$. (We have defined $J_{\pm,\theta}\equiv J_{1,\theta}\pm iJ_{2,\theta}$.) Therefore, we have

$$U=e^{-R_{r,\theta}J_{+,\theta}}(\tau_\theta)^{2J_{3,\theta}}e^{R_{l,\theta}J_{-,\theta}} \tag{6.13}$$

for arbitrary θ. Applying $P(-\theta)$ to (6.13) from the left and $P(\theta)$ from the right we obtain

$$U_\theta = e^{-R_{r,\theta} J_+} (\tau_\theta)^{2J_3} e^{R_{l,\theta} J_-}. \tag{6.14}$$

We may say that equation (6.8), together with the expression (6.14), is covariant with respect to the rotation, under which the coordinates and the scattering coefficients change as $(X,Y) \to (X_\theta, Y_\theta)$, $(\tau, R_r, R_l) \to (\tau_\theta, R_{r,\theta}, R_{l,\theta})$. The matrix elements of U_θ in the ν representation (see Sec. III) with the basis (3.21) can be obtained from (3.22) by the replacement $U^{(\nu)} \to U_\theta^{(\nu)}$ with $(\tau, R_r, R_l) \to (\tau_\theta, R_{r,\theta}, R_{l,\theta})$.

We can use these expressions for any complex number θ, but particularly important values are $\theta = \pm \pi/2$. We have $X_{\pm\pi/2} = \pm Y = \pm \frac{1}{2} V(x)$, $Y_{\pm\pi/2} = \mp X = \mp ikx$, and thus the roles of the space coordinate and the potential are completely exchanged at $\theta = \pm \pi/2$. If we choose the parameterization as $\lambda = x$, the factor assigned to each scattering vertex becomes $(d/dx) Y_{\pm\pi/2} = \mp ik$. Consequently, the diagrammatic expression with $\theta = \pm \pi/2$ gives a power series expansion in terms of k. We can use this property for studying the behavior of the scattering coefficients at small k.

Finally, let us remark that if the potential $V(x)$ is a linear function of x, say $V(x) = -2cx$, there is no scattering in the coordinates (X_θ, Y_θ) with $\theta = \arctan(-Y/X) = \arctan(-ic/k)$. This is almost obvious from Fig. 19, and indeed from (6.11) we can see $dY_\theta/d\lambda = 0$ for $\theta = \arctan(-ic/k)$. We will see the significance of this fact in Sec. X.

VII. REALIZATION IN TERMS OF DIFFERENTIAL OPERATORS

As shown in Sec. IV, the evolution equation (3.1) describes how to make paths penetrating a larger interval by connecting fragmentary paths in smaller intervals. We wish to describe the operation of connecting the paths as an algebraic manipulation. To this end we use a realization of the SU(2) [or SL(2, **C**)] generators as differential operators. What we are going to do in this section is in its spirit analogous to the path-integral description of boson systems using coherent states,[15] although we do not deal with path integrals explicitly.

It is well-known that the set of the differential operators

$$\mathscr{J}_+ = -\xi^2 \frac{\partial}{\partial \xi}, \quad \mathscr{J}_- = \frac{\partial}{\partial \xi}, \quad \mathscr{J}_3 = \xi \frac{\partial}{\partial \xi} \tag{7.1}$$

satisfies the commutation relations (3.2). (We use the symbol $\mathscr{J}$ for differential operators, instead of J.) The operators $\mathscr{J}_3$, $\mathscr{J}_-$, and $\mathscr{J}_+$ are, respectively, the generator of dilatation, translation, and special conformal transformation.[16] That is, for an arbitrary function $g(\xi)$ and a constant c we have

$$\exp(c\mathscr{J}_3) g(\xi) = g(e^c \xi), \quad \exp(c\mathscr{J}_-) g(\xi) = g(\xi + c), \quad \exp(c\mathscr{J}_+) g(\xi) = g\left(\frac{\xi}{1 + c\xi}\right). \tag{7.2}$$

The representation space is the space of all analytic functions of ξ. The lowest state u_0 satisfying $\mathscr{J}_- u_0 = 0$ is $u_0 = 1$ (constant), and since $\mathscr{J}_3 u_0 = 0$ we find from (3.16) that this representation corresponds to $\nu = 0$.

Let us try to generalize (7.1) for other values of ν. We need to find a set of differential operators $\mathscr{J}_+^{(\nu)}$, $\mathscr{J}_-^{(\nu)}$, and $\mathscr{J}_3^{(\nu)}$ satisfying the commutation relations (3.2), such that the function $u_0^{(\nu)}(\xi)$ satisfying $\mathscr{J}_-^{(\nu)} u_0^{(\nu)} = 0$ also satisfies $\mathscr{J}_3^{(\nu)} u_0^{(\nu)} = (\nu/2) u_0^{(\nu)}$. It is easy to see that the following set of operators fulfills the requirements:

$$\mathscr{J}_+^{(\nu)} = -\gamma^\nu \xi^{-\nu+2} \frac{\partial}{\partial \xi} \xi^\nu \gamma^{-\nu}, \quad \mathscr{J}_-^{(\nu)} = \gamma^\nu \frac{\partial}{\partial \xi} \gamma^{-\nu}, \quad \mathscr{J}_3^{(\nu)} = \gamma^\nu \xi^{-(\nu/2)+1} \frac{\partial}{\partial \xi} \xi^{\nu/2} \gamma^{-\nu}, \tag{7.3}$$

where $\gamma(\xi)$ is an arbitrary function of ξ. (We will specify the explicit form of $\gamma(\xi)$ in the next section.) As a generalization of Eqs. (7.2) we can easily derive

$$\exp(c\mathscr{J}_3^{(\nu)}) \gamma^\nu g(\xi) = \gamma^\nu (e^c)^{\nu/2} g(e^c \xi), \quad \exp(c\mathscr{J}_-^{(\nu)}) \gamma^\nu g(\xi) = \gamma^\nu g(\xi + c),$$

$$\exp(c\mathscr{J}_+^{(\nu)})\gamma^\nu g(\xi)=\gamma^\nu\left(\frac{1}{1+c\xi}\right)^\nu g\left(\frac{\xi}{1+c\xi}\right). \tag{7.4}$$

Now the representation space is the space of functions of the form $[\gamma(\xi)]^\nu g(\xi)$, where g is an analytic function of ξ. The lowest state is $u_0^{(\nu)}(\xi)=[\gamma(\xi)]^\nu$. We choose other base states as

$$u_n^{(\nu)}(\xi)\equiv(-1)^n\frac{1}{n!}[\mathscr{J}_+^{(\nu)}]^n u_0^{(\nu)}(\xi)=\binom{\nu+n-1}{n}\gamma^\nu\xi^n \tag{7.5}$$

[cf. Eq. (3.21)], and then the representation matrices for the operators (7.3) coincide with (3.17). Let us compare the expression (7.5) with Fig. 8. We can see that a γ and a ξ corresponds to an open circle and solid circle, respectively. The factor $\binom{\nu+n-1}{n}$, which corresponds to the rearrangement of the legs, is also included in this expression.

We define the inner product between the base states as

$$(u_m^{(\nu)},u_n^{(\nu)})^{(\nu)}=\binom{\nu+n-1}{n}\delta_{mn}, \tag{7.6}$$

which is equivalent to (3.20). For $\nu\geqslant 1$, such an inner product can be explicitly written as

$$(u,v)^{(\nu)}=\frac{\nu-1}{2\pi i}\int_{|\xi|<1}\frac{(1-\xi^*\xi)^{\nu-2}}{|\gamma(\xi)|^{2\nu}}[u(\xi)]^* v(\xi)d\xi\, d\xi^*, \tag{7.7}$$

where $d\xi\, d\xi^*/(2i)=d(\mathrm{Re}\,\xi)\,d(\mathrm{Im}\,\xi)$, and the domain of integration is the inside of the unit circle on the complex ξ-plane. [For $\nu=1$, the right-hand side of (7.7) should be understood as a limit.] Indeed, using the formula

$$(\nu-1)\int_0^1 t^n(1-t)^{\nu-2}dt=(\nu-1)\frac{\Gamma(n+1)\Gamma(\nu-1)}{\Gamma(\nu+n)}=\frac{1}{\binom{\nu+n-1}{n}}, \tag{7.8}$$

we can readily verify that the inner product given by (7.7) satisfies (7.6). However, in this paper we do not need the expression (7.7). All we need is Eq. (7.6). For arbitrary $v(\xi)$, the inner product with $u_n^{(\nu)}$ singles out the coefficient of ξ^n in $v(\xi)/[\gamma(\xi)]^\nu$:

$$(u_n^{(\nu)},v)^{(\nu)}=\frac{1}{n!}\frac{\partial^n}{\partial\xi^n}\frac{v(\xi)}{[\gamma(\xi)]^\nu}\bigg|_{\xi=0}. \tag{7.9}$$

The completeness of the basis is expressed as

$$v(\xi)=\sum_{n=0}^{\infty}u_n^{(\nu)}\frac{1}{\binom{\nu+n-1}{n}}(u_n^{(\nu)},v)^{(\nu)}. \tag{7.10}$$

The right-hand side of (7.10) is nothing but the Taylor expansion in terms of ξ.

By using the inner product defined above, we can express a differential operator in terms of an integral kernel. Let $\mathscr{A}$ be a differential operator acting on functions of ξ. We define the integral kernel $A(\xi,\eta^*)$ associated with $\mathscr{A}$ as

$$\begin{aligned}A(\xi,\eta^*)&\equiv\sum_{n=0}^{\infty}\mathscr{A}u_n^{(\nu)}(\xi)\frac{1}{\binom{\nu+n-1}{n}}[u_n^{(\nu)}(\eta)]^*\\&=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\binom{\nu+m-1}{m}[\gamma(\xi)\gamma^*(\eta)]^\nu\xi^m\eta^{*n}A_{mn},\end{aligned} \tag{7.11}$$

where

$$A_{mn}=\frac{1}{\binom{\nu+m-1}{m}}\,(u_m^{(\nu)},\mathscr{A}u_n^{(\nu)})^{(\nu)}$$

are the matrix elements of the operator $\mathscr{A}$. This $A(\xi,\eta^*)$ is an integral kernel in the sense that the operation of $\mathscr{A}$ can be expressed as

$$\mathscr{A}v(\xi)=\frac{\nu-1}{2\pi i}\int_{|\eta|<1}\frac{(1-\eta^*\eta)^{\nu-2}}{|\gamma(\eta)|^{2\nu}}A(\xi,\eta^*)v(\eta)d\eta\,d\eta^*, \tag{7.12}$$

although we do not use (7.12) in this paper. The integral kernel can be also interpreted as a generating function for the matrix elements A_{mn}. The matrix elements are obtained from $A(\xi,\eta^*)$ as

$$A_{mn}=\frac{1}{\binom{\nu+m-1}{m}m!n!}\frac{\partial^n}{\partial\xi^n}\frac{\partial^n}{\partial\eta^{*n}}\frac{A(\xi,\eta^*)}{[\gamma(\xi)\gamma^*(\eta)]^\nu}\bigg|_{\xi=0,\eta^*=0}. \tag{7.13}$$

The kernel of the identity operation is

$$I^{(\nu)}(\xi,\eta^*)=\sum_{n=0}^{\infty}u_n^{(\nu)}(\xi)\frac{1}{\binom{\nu+n-1}{n}}[u_n^{(\nu)}(\eta)]^*=\left[\frac{\gamma(\xi)\gamma^*(\eta)}{1-\xi\eta^*}\right]^\nu. \tag{7.14}$$

From (7.11) and (7.14) we can see that the kernel of an arbitrary differential operator $\mathscr{A}$ is obtained as $A(\xi,\eta^*)=\mathscr{A}I^{(\nu)}(\xi,\eta^*)$.

Now let us consider the evolution operator $U^{(\nu)}$. Its form as a differential operator is obtained from Eq. (3.12) by replacing J with $\mathscr{J}$:

$$\mathscr{U}^{(\nu)}(x,x')=e^{-R_r(x,x')\mathscr{J}_+^{(\nu)}}[\tau(x,x')]^{2\mathscr{J}_3^{(\nu)}}e^{R_l(x,x')\mathscr{J}_-^{(\nu)}}. \tag{7.15}$$

We let $U_K^{(\nu)}(x,x';\xi,\eta^*)$ denote the kernel of this operator. (The subscript K is added here to stress that it is a kernel.) Using (7.4) we can calculate it as

$$\begin{aligned}U_K^{(\nu)}(x,x';\xi,\eta^*)&=\mathscr{U}^{(\nu)}(x,x')I^{(\nu)}(\xi,\eta^*)=e^{-R_r\mathscr{J}_+^{(\nu)}}\tau^{2\mathscr{J}_3^{(\nu)}}e^{R_l\mathscr{J}_-^{(\nu)}}\left[\frac{\gamma(\xi)\gamma^*(\eta)}{1-\xi\eta^*}\right]^\nu\\&=e^{-R_r\mathscr{J}_+^{(\nu)}}\tau^{2\mathscr{J}_3^{(\nu)}}\left[\frac{\gamma(\xi)\gamma^*(\eta)}{1-(\xi+R_l)\eta^*}\right]^\nu\\&=e^{-R_r\mathscr{J}_+^{(\nu)}}\left[\frac{\tau\gamma(\xi)\gamma^*(\eta)}{1-(\tau^2\xi+R_l)\eta^*}\right]^\nu=\left[\frac{\tau\gamma(\xi)\gamma^*(\eta)}{1-R_r\xi-R_l\eta^*-(\tau^2-R_rR_l)\xi\eta^*}\right]^\nu.\end{aligned} \tag{7.16}$$

We can write all these expressions in a covariant form with respect to the rotation discussed in Sec. VI. The differential operator corresponding to $P(\theta)$ [Eq. (6.5)] is

$$\mathscr{P}^{(\nu)}(\theta)=e^{-\tan(\theta/2)\mathscr{J}_+^{(\nu)}}[\cos(\theta/2)]^{-2\mathscr{J}_3^{(\nu)}}e^{\tan(\theta/2)\mathscr{J}_-^{(\nu)}}=e^{\tan(\theta/2)\mathscr{J}_-^{(\nu)}}[\cos(\theta/2)]^{2\mathscr{J}_3^{(\nu)}}e^{-\tan(\theta/2)\mathscr{J}_+^{(\nu)}}. \tag{7.17}$$

Applying this operator to $u_n^{(\nu)}$ [Eq (7.5)] we obtain the basis in the rotated frame as

$$u_{n,\theta}^{(\nu)}\equiv\mathscr{P}^{(\nu)}(\theta)u_n^{(\nu)}=\binom{\nu+n-1}{n}\gamma_\theta^\nu\xi_\theta^n, \tag{7.18}$$

where

$$\xi_\theta \equiv \frac{\tan(\theta/2)+\xi}{1-\xi\,\tan(\theta/2)}, \quad \gamma_\theta(\xi) \equiv \frac{\gamma(\xi)}{\cos(\theta/2)-\xi\,\sin(\theta/2)}. \tag{7.19}$$

In deriving the last expression of (7.18), we successively used the three equations of (7.4) as we did in (7.16). In terms of ξ_θ and γ_θ we can write $\mathscr{J}^{(\nu)}_{a,\theta}=\mathscr{R}^{(\nu)}(\theta)\mathscr{J}^{(\nu)}_{a}\mathscr{R}^{(\nu)}(-\theta)$ as

$$\mathscr{J}^{(\nu)}_{+,\theta}=-\gamma_\theta^\nu\xi_\theta^{-\nu+2}\frac{\partial}{\partial\xi_\theta}\xi_\theta^\nu\gamma_\theta^{-\nu}, \quad \mathscr{J}^{(\nu)}_{-,\theta}=\gamma_\theta^\nu\frac{\partial}{\partial\xi_\theta}\gamma_\theta^{-\nu}, \quad \mathscr{J}^{(\nu)}_{3,\theta}=\gamma_\theta^\nu\xi_\theta^{-(\nu/2)+1}\frac{\partial}{\partial\xi_\theta}\xi_\theta^{\nu/2}\gamma_\theta^{-\nu}. \tag{7.20}$$

From (3.13) and (7.17) it follows that $(u,v)=(\mathscr{R}(-\theta^*)u,\mathscr{R}(\theta)v)$ for arbitrary u and v. Therefore the kernel of identity can be rewritten as

$$I^{(\nu)}(\xi,\eta^*)=\sum_{n=0}^{\infty}[\mathscr{R}(\theta)u_n(\xi)]\frac{1}{\binom{\nu+n-1}{n}}[\mathscr{R}(-\theta^*)u_n(\eta)]^*=\left[\frac{\gamma_\theta(\xi)\gamma^*_{-\theta}(\eta)}{1-\xi_\theta\eta^*_{-\theta}}\right]^\nu, \tag{7.21}$$

where, as in (7.19), $\eta^*_{-\theta}=(-\tan(\theta/2)+\eta^*)/(1+\eta^*\tan(\theta/2))$, $\gamma^*_{-\theta}(\eta)=\gamma^*(\eta)/(\cos(\theta/2)+\eta^*\sin(\theta/2))$. (Note that $\eta^*_{-\theta}=(\eta_{-\theta^*})^*$ and $\gamma^*_{-\theta}(\eta)=[\gamma_{-\theta^*}(\eta)]^*$.) Using (6.13) and (7.21) we can generalize the expression (7.16) to arbitrary θ as

$$\begin{aligned}U^{(\nu)}_K(x,x';\xi,\eta^*)&=e^{-R_{r,\theta}\mathscr{J}^{(\nu)}_{+,\theta}}(\tau_\theta)^{2\mathscr{J}^{(\nu)}_{3,\theta}}e^{R_{l,\theta}\mathscr{J}^{(\nu)}_{-,\theta}}\left[\frac{\gamma_\theta(\xi)\gamma^*_{-\theta}(\eta)}{1-\xi_\theta\eta^*_{-\theta}}\right]^\nu\\&=\left[\frac{\tau_\theta\gamma_\theta(\xi)\gamma^*_{-\theta}(\eta)}{1-R_{r,\theta}\xi_\theta-R_{l,\theta}\eta^*_{-\theta}-(\tau_\theta^2-R_{r,\theta}R_{l,\theta})\xi_\theta\eta^*_{-\theta}}\right]^\nu.\end{aligned} \tag{7.22}$$

The kernel of the evolution operator introduced in this section is not merely an abstract tool. We shall see in the next section that $U^{(1)}_K$ can be identified with the generalized transmission coefficient $\bar{\tau}$ defined in Sec. II.

VIII. IDENTIFICATION OF $\bar{\tau}$ AS A GENERATING FUNCTION

In this section we show that $U^{(1)}_K$ is equivalent to $\bar{\tau}$. To simplify the notation, we shall hereafter use ξ' in place of η^* and write the kernel of the evolution operator as $U^{(\nu)}_K(x,x';\xi,\xi')$. The kernel of identity [Eqs. (7.14) and (7.21)] is written as

$$I^{(\nu)}(\xi,\xi')=\left(\frac{\gamma\gamma'}{1-\xi\xi'}\right)^\nu=\left(\frac{\gamma_\theta\gamma'_{-\theta}}{1-\xi_\theta\xi'_{-\theta}}\right)^\nu, \tag{8.1}$$

where $\gamma\equiv\gamma(\xi)$, $\gamma'\equiv\gamma^*(\xi'^*)$, $\gamma_\theta\equiv\gamma_\theta(\xi)$, and $\gamma'_{-\theta}\equiv\gamma^*_{-\theta}(\xi'^*)$. Just like (7.19) we have $\xi'_{-\theta}=(-\tan(\theta/2)+\xi')/(1+\xi'\tan(\theta/2))$ and $\gamma'_{-\theta}=\gamma'/(\cos(\theta/2)+\xi'\sin(\theta/2))$.

First, we remark that $I^{(\nu)}(\xi,\xi')$ can be written as

$$I^{(\nu)}(\xi,\xi')=\left(\frac{\gamma\gamma'}{1-\xi\xi'}\right)^\nu=(\gamma\gamma')^\nu(u_0^{(\nu)},e^{\xi J^{(\nu)}_-}e^{-\xi' J^{(\nu)}_+}u_0^{(\nu)})^{(\nu)}. \tag{8.2}$$

This equation can be verified by using Eqs. (3.16), (3.14), (3.13), and the formula[17]

$$e^{AJ_-}e^{BJ_+}=(1+AB)^{-J_3}e^{BJ_+}e^{AJ_-}(1+AB)^{-J_3}. \tag{8.3}$$

In the last expression of (8.2), the representation space is not the space of functions of ξ discussed in Sec. VII, but some other arbitrary space. The vector $u_0^{(\nu)}$ is not the function of ξ defined by

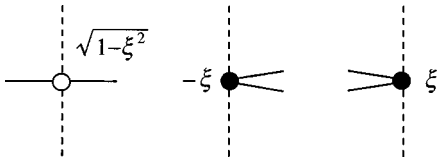

FIG. 21. Graphical representation of transmission and reflection at a potential discontinuity. The vertical dashed line stands for the discontinuity.

(7.5), and the operators J are different from the operators $\mathscr{J}$ defined by (7.3). Here ξ and ξ' are merely parameters. Now let us apply the differential operators (7.3) to (8.2). We can prove the relation

$$\mathscr{J}_a^{(\nu)}\gamma^\nu(u_0^{(\nu)},e^{\xi J_-^{(\nu)}}\cdots)^{(\nu)}=\gamma^\nu(u_0^{(\nu)},e^{\xi J_-^{(\nu)}}J_a^{(\nu)}\cdots)^{(\nu)} \quad (a=1,2,3, \text{ or } \pm), \tag{8.4}$$

where ''...'' is arbitrary. (See Appendix B.) From (7.15), (8.2), and (8.4) we obtain

$$\begin{aligned} U_K^{(\nu)}(x,x';\xi,\xi') &= e^{-R_r\mathscr{J}_+^{(\nu)}}\tau^{2\mathscr{J}_3^{(\nu)}}e^{R_l\mathscr{J}_-^{(\nu)}}(\gamma\gamma')^\nu(u_0^{(\nu)},e^{\xi J_-^{(\nu)}}e^{-\xi' J_+^{(\nu)}}u_0^{(\nu)})^{(\nu)} \\ &= (\gamma\gamma')^\nu(u_0^{(\nu)},e^{\xi J_-^{(\nu)}}e^{-R_r J_+^{(\nu)}}\tau^{2J_3^{(\nu)}}e^{R_l J_-^{(\nu)}}e^{-\xi' J_+^{(\nu)}}u_0^{(\nu)})^{(\nu)} \\ &= (\gamma\gamma')^\nu(u_0^{(\nu)},e^{\xi J_-^{(\nu)}}U^{(\nu)}(x,x')e^{-\xi' J_+^{(\nu)}}u_0^{(\nu)})^{(\nu)}. \end{aligned} \tag{8.5}$$

Next let us think about $\overline{\tau}$. As explained in Sec. II, the quantities $\overline{\tau}$, $\overline{R_r}$, and $\overline{R_l}$ can be interpreted as the scattering coefficients that include the effect of potential discontinuities at the endpoints. Let us fist see the effect of a discontinuity. Suppose that the potential $V(x)$ has a finite discontinuity of magnitude w at position x_0

$$\lim_{\epsilon\to 0+}[V(x_0+\epsilon)-V(x_0-\epsilon)]=w. \tag{8.6}$$

Then the reflection coefficients across the discontinuity, $R_r(x_0+\epsilon,x_0-\epsilon)$ and $R_l(x_0+\epsilon,x_0-\epsilon)$, have nonzero values in the limit $\epsilon\to 0$. It can be shown that[4]

$$\lim_{\epsilon\to 0+} R_r(x_0+\epsilon,x_0-\epsilon)=-\xi, \quad \lim_{\epsilon\to 0+} R_l(x_0+\epsilon,x_0-\epsilon)=\xi$$

$$\lim_{\epsilon\to 0+} \tau(x_0+\epsilon,x_0-\epsilon)=\sqrt{1-\xi^2}, \tag{8.7}$$

where

$$\xi\equiv -\frac{1-e^w}{1+e^w}. \tag{8.8}$$

The coefficients (8.7) represent the scattering caused by the discontinuity, as shown in Fig. 21. (Of course, these coefficients do not depend on the position x_0.) Substituting (8.7) into (3.12) we obtain the evolution operator across the discontinuity as

$$\lim_{\epsilon\to 0} U(x_0+\epsilon,x_0-\epsilon)=D(\xi), \tag{8.9}$$

where

$$D(\xi)\equiv e^{\xi J_+}(1-\xi^2)^{J_3}e^{\xi J_-}. \tag{8.10}$$

From the multiplication rule (3.3) we find

$$U(x_0+\epsilon,z)=D(\xi)U(x_0-\epsilon,z), \quad U(z,x_0-\epsilon)=U(z,x_0+\epsilon)D(\xi) \tag{8.11}$$

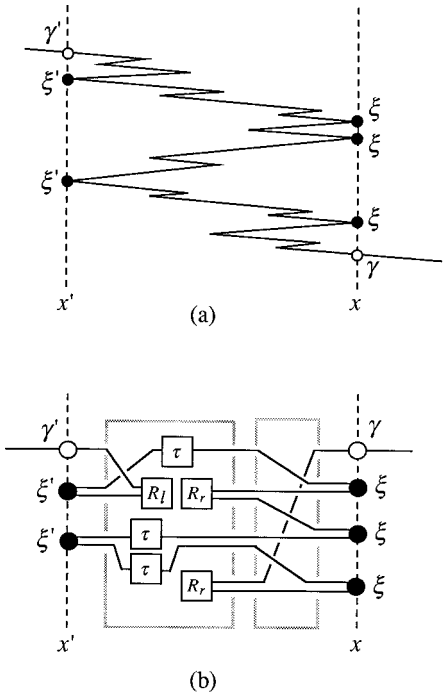

FIG. 22. (a) A diagram contributing to $\bar{\tau}$. This diagram is included in (b).

for arbitrary z and positive infinitesimal ϵ. In other words, the evolution operator is multiplied by $D(\xi)$ when the discontinuity is crossed.

Now let us return to general $V(x)$ and consider $\bar{\tau}$, $\bar{R}_r$, and $\bar{R}_l$ for the interval (x_a, x_b). The $\bar{V}$ given by (2.1) has discontinuities at x_a and x_b. Just like (8.9), the scattering at these discontinuities is described by $D(-\xi_a)$ and $D(\xi_b)$, respectively, where $\xi_a \equiv -(1-e^{w_a})/(1+e^{w_a})$, $\xi_b \equiv -(1-e^{w_b})/(1+e^{w_b})$. [Owing to the definition of the sign of w_a in (2.1), the scattering at x_a corresponds to $D(-\xi_a)$ rather than $D(\xi_a)$.] From (8.11) we can see that the evolution operator including the discontinuities is given by

$$\bar{U}(x_b, x_a; w_b, w_a) \equiv D(\xi_b) U(x_b, x_a) D(-\xi_a). \tag{8.12}$$

It is obvious that $\bar{\tau}$, $\bar{R}_r$, and $\bar{R}_l$ are related to $\bar{U}$ as $\bar{U} = e^{-\bar{R}_r J_+} \bar{\tau}^{2J_3} e^{\bar{R}_l J_-}$, in the same way as (3.12). Relations such as (3.15), (3.22), or (3.23) also hold with the replacement $(U, \tau, R_r, R_l) \to (\bar{U}, \bar{\tau}, \bar{R}_r, \bar{R}_l)$.

From now on we write $(x, x'; w, w')$ in place of $(x_b, x_a; w_b, w_a)$. The transmission coefficient including the discontinuities is obtained from (3.23) as

$$\bar{\tau}(x, x'; w, w') = (u_0^{(1)}, \bar{U}^{(1)} u_0^{(1)})^{(1)} = (u_0^{(1)}, D^{(1)}(\xi) U^{(1)}(x, x') D^{(1)}(-\xi') u_0^{(1)})^{(1)}, \tag{8.13}$$

where ξ is defined by (8.8), and ξ' is given by (8.8) with w replaced by w'. Substituting (8.10) into (8.13) and using $J_- u_0 = 0$ and $J_3^{(\nu)} u_0^{(\nu)} = (\nu/2) u_0^{(\nu)}$, we obtain

$$\bar{\tau}(x, x'; w, w') = [(1-\xi^2)(1-\xi'^2)]^{1/2} (u_0^{(1)}, e^{\xi J_-^{(1)}} U^{(1)}(x, x') e^{-\xi' J_+^{(1)}} u_0^{(1)})^{(1)}. \tag{8.14}$$

Comparing (8.14) with (8.5) we find that $U_K^{(1)}$ is identical with $\bar{\tau}$ if we identify γ and γ' as

$$\gamma = (1-\xi^2)^{1/2}, \quad \gamma' = (1-\xi'^2)^{1/2}. \tag{8.15}$$

Graphically, $\bar{\tau}$ is obtained as the sum over all such diagrams as shown in Fig. 22. A diagram contributing to $\bar{\tau}$ is similar to the one in Fig. 2(a), but now includes the effect of scattering at the endpoints. Each reflection at x and x' gives a factor ξ and ξ', respectively, and we have a factor

γ and γ' corresponding to the transmission through each of the endpoints. Comparing Fig. 22(b) with the definition of the integral kernel [Eq. (7.11) or (7.13)] we can indeed see the exact likeness between $\tau(x,x';w,w')$ and $U_K^{(1)}(x,x';\xi,\xi')$.

Hereafter we assume that γ and γ' are given by (8.15). Now we have $\overline{\tau}=U_K^{(1)}$, and from (7.16) we also find $\overline{\tau^\nu}=U_K^{(\nu)}$. Substituting (8.15) into (7.3) we obtain

$$\mathscr{J}_+^{(\nu)}=-\frac{\nu\xi}{1-\xi^2}-\xi^2\frac{\partial}{\partial\xi},\quad \mathscr{J}_-^{(\nu)}=\frac{\nu\xi}{1-\xi^2}+\frac{\partial}{\partial\xi},\quad \mathscr{J}_3^{(\nu)}=\frac{\nu}{2}\frac{1+\xi^2}{1-\xi^2}+\xi\frac{\partial}{\partial\xi}. \tag{8.16}$$

Equations (8.15), together with the definition (8.8), can be expressed as

$$\xi=\tanh(w/2),\quad \xi'=\tanh(w'/2),\quad \gamma=\frac{1}{\cosh(w/2)},\quad \gamma'=\frac{1}{\cosh(w'/2)}. \tag{8.17}$$

From (7.19) and (8.15) we have

$$\gamma_\theta=[(1-\xi_\theta^2)\cos\,\theta+2\xi_\theta\,\sin\,\theta]^{1/2},\quad \gamma'_{-\theta}=[(1-\xi'^2_{-\theta})\cos\,\theta-2\xi'_{-\theta}\,\sin\,\theta]^{1/2}. \tag{8.18}$$

Let us also remark that $\mathscr{J}_1^{(\nu)}=\frac{1}{2}(\mathscr{J}_+^{(\nu)}+\mathscr{J}_-^{(\nu)})$ is θ-invariant and independent of ν:

$$\mathscr{J}_1^{(\nu)}=\frac{1}{2}\,\gamma^2\frac{\partial}{\partial\xi}=\frac{1}{2}\,\gamma_\theta^2\frac{\partial}{\partial\xi_\theta}. \tag{8.19}$$

To summarize, we have

$$\overline{\tau}=U_K^{(1)}=\frac{\tau\gamma\gamma'}{1-R_r\xi-R_l\xi'-(\tau^2-R_rR_l)\xi\xi'}=\frac{\tau_\theta\gamma_\theta\gamma'_{-\theta}}{1-R_{r,\theta}\xi_\theta-R_{l,\theta}\xi'_{-\theta}-(\tau_\theta^2-R_{r,\theta}R_{l,\theta})\xi_\theta\xi'_{-\theta}}. \tag{8.20}$$

Note that $\overline{\tau}$ is invariant with respect to rotation; it does not depend on θ. Therefore, the diagrammatic interpretation shown in Fig. 22(a) holds in any rotated frame (with the rules given by Fig. 20) if we replace $(\xi,\xi',\gamma,\gamma')$ by $(\xi_\theta,\xi'_{-\theta},\gamma_\theta,\gamma'_{-\theta})$. The quantity $\overline{\tau}$ serves as a generating function for the evolution matrix elements. Since $U_K^{(\nu)}=\overline{\tau^\nu}$, from (7.13) we have

$$(U^{(\nu)})_{mn}=\frac{1}{\binom{\nu+m-1}{m}m!n!}\frac{\partial^m}{\partial\xi^m}\frac{\partial^n}{\partial\xi'^n}\left(\frac{\overline{\tau}}{\gamma\gamma'}\right)^\nu\Bigg|_{\xi=\xi'=0}. \tag{8.21}$$

The matrix elements of $U^{(0)}$ are obtained by taking the limit $\nu\to0$ in (8.21). Since $x^\nu=1+\nu\log x+O(\nu)$ and $\binom{\nu+m-1}{m}=(\nu/m)+O(\nu)$, we find

$$(U^{(0)})_{mn}=\frac{m}{m!n!}\frac{\partial^m}{\partial\xi^m}\frac{\partial^n}{\partial\xi'^n}\log\left(\frac{\overline{\tau}}{\gamma\gamma'}\right)\Bigg|_{\xi=\xi'=0}. \tag{8.22}$$

In particular, the transmission and reflection coefficients are obtained as

$$\tau=\frac{\overline{\tau}}{\gamma\gamma'}\Bigg|_{\xi=\xi'=0},\quad R_r=\frac{\partial}{\partial\xi}\log\left(\frac{\overline{\tau}}{\gamma\gamma'}\right)\Bigg|_{\xi=\xi'=0},\quad R_l=\frac{\partial}{\partial\xi'}\log\left(\frac{\overline{\tau}}{\gamma\gamma'}\right)\Bigg|_{\xi=\xi'=0}. \tag{8.23}$$

The matrix elements of U_θ with arbitrary θ are given by (8.21) or (8.22) with the replacement $(\xi,\xi',\gamma,\gamma')\to(\xi_\theta,\xi'_{-\theta},\gamma_\theta,\gamma'_{-\theta})$. The scattering coefficients in the rotated frame are

$$\tau_\theta=\frac{\overline{\tau}}{\gamma_\theta\gamma'_{-\theta}}\Bigg|_{\xi_\theta=\xi'_{-\theta}=0}, \tag{8.24a}$$

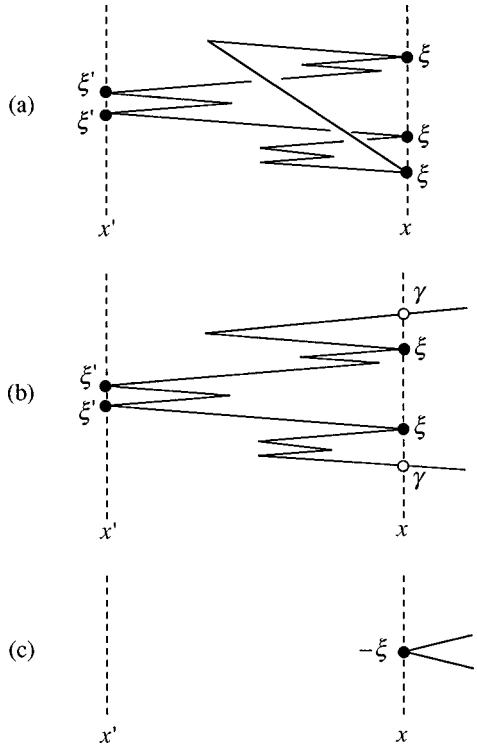

FIG. 23. (a) A diagram contributing to $\log \overline{Z}$. (b) A typical diagram contributing to $\overline{R_r}$. (c) The only contribution to $\overline{R_r}$ that is not included in $2\gamma^2(\partial/\partial\xi)\log \overline{Z}$.

$$R_{r,\theta}=\frac{\partial}{\partial \xi_\theta}\log\left(\frac{\overline{\tau}}{\gamma_\theta \gamma'_{-\theta}}\right)\Bigg|_{\xi_\theta=\xi'_{-\theta}=0}, \quad R_{l,\theta}=\frac{\partial}{\partial \xi'_{-\theta}}\log\left(\frac{\overline{\tau}}{\gamma_\theta \gamma'_{-\theta}}\right)\Bigg|_{\xi_\theta=\xi'_{-\theta}=0}. \tag{8.24b}$$

Using (8.18) and (7.19) we can rewrite (8.24a) as

$$\tau_\theta=\frac{\overline{\tau}}{\cos\,\theta}\Bigg|_{\xi=-\tan(\theta/2),\,\xi'=\tan(\theta/2)}. \tag{8.25}$$

We will see in the next section that (8.24b) can be rewritten in terms of $\overline{R_r}$ and $\overline{R_l}$ just like (8.25).

IX. STRUCTURE OF $\overline{R_r}$ AND $\overline{R_l}$

In this section we study the generalized reflection coefficients $\overline{R_r}$ and $\overline{R_l}$. Before doing that, let us generalize Eq. (3.26) and define

$$\overline{Z}(x,x';\xi,\xi')\equiv e^{-(1/2)ik(x-x')}\langle 0|e^{(1/2)\xi a a}U(x,x')e^{(1/2)\xi' a^\dagger a^\dagger}|0\rangle, \tag{9.1}$$

where U is the operator in the boson representation satisfying (3.10). The operators $e^{(1/2)\xi aa}$ and $e^{(1/2)\xi' a^\dagger a^\dagger}$ in (9.1) stand for the reflection at the endpoints. Just like $\log Z$ discussed in Sec. III, the quantity $\log \overline{Z}$ is obtained by taking the sum over all configurations of a single closed loop, now with reflection at the endpoints. Each reflection at x and x' gives a factor ξ and ξ', respectively [Fig. 23(a)]. Equation (8.5) with $\nu=1/2$ can be written in the boson representation as $\overline{\tau}^{1/2}=(\gamma\gamma')^{1/2}\langle 0|e^{(1/2)\xi aa}Ue^{(1/2)\xi' a^\dagger a^\dagger}|0\rangle$. (Recall that $U_K^{(1/2)}=\overline{\tau}^{1/2}$.) Therefore, we have, as a generalization of (3.27),

$$\overline{\tau}=e^{ik(x-x')}\gamma\gamma'\overline{Z}^2. \tag{9.2}$$

We can obtain $\overline{R_r}$ by taking the sum over diagrams in the same way as we did for R_r. The only difference is that we also take account of the scattering at the endpoints. A typical diagram contributing to $\overline{R_r}$ is shown in Fig. 23(b). We remark that this diagram can be obtained from the diagram of $\log Z$ shown in Fig. 23(a) by replacing a ξ by two γ's. Therefore, the sum over all these diagrams can be expressed as $2\gamma^2(\partial/\partial\xi)\log Z$. (The factor 2 means that we make distinction between mirror images.) The diagram shown in Fig. 23(c) is the only exception. This diagram, which also contributes to $\overline{R_r}$, is not included in $2\gamma^2(\partial/\partial\xi)\log Z$, and so it must be added separately. Thus we have

$$\overline{R_r}=2\gamma^2\frac{\partial}{\partial\xi}\log\overline{Z}-\xi. \tag{9.3}$$

Using (9.2), (8.15) and (8.19) we may write

$$\overline{R_r}=\gamma^2\frac{\partial}{\partial\xi}\log\overline{\tau}=2\mathscr{J}_1\log\overline{\tau}. \tag{9.4}$$

[A different derivation of (9.4) is given in Appendix C.]

Substituting (8.20) into (9.4) we can calculate $\overline{R_r}$ as

$$\begin{aligned}\overline{R_r}&=\gamma^2\frac{\partial}{\partial\xi}\log\gamma-\gamma^2\frac{\partial}{\partial\xi}\log[1-R_r\xi-R_l\xi'-(\tau^2-R_lR_r)\xi\xi']\\&=-\xi+\gamma^2\frac{R_r+(\tau^2-R_lR_r)\xi'}{1-R_r\xi-R_l\xi'-(\tau^2-R_lR_r)\xi\xi'}.\end{aligned} \tag{9.5}$$

We may rewrite (9.5) in the form

$$\overline{R_r}=-\xi+\gamma^2\frac{\rho_r}{1-\rho_r\xi},\quad \rho_r\equiv R_r+\frac{\tau^2\xi'}{1-R_l\xi'}. \tag{9.6}$$

We can clearly see the structure of Eqs. (9.6) by writing them as

$$\overline{R_r}=-\xi+\gamma^2\rho_r\sum_{n=0}^{\infty}(\rho_r\xi)^n,\quad \rho_r=R_r+\tau^2\xi'\sum_{n=0}^{\infty}(R_l\xi')^n. \tag{9.7}$$

These equations are graphically represented as in Fig. 24. As explained before, $\overline{R_r}(x,x';\xi,\xi')$ includes the effect of scattering at both endpoints x and x', while $R_r(x,x')$ is the reflection coefficient without scattering at the endpoints. The quantity $\rho_r(x,x';\xi')$ defined in Eq. (9.6) is the reflection coefficient with scattering at x' but not at x.

It is straightforward to generalize these expressions for the rotated coordinate system with arbitrary θ. The quantity corresponding to $\overline{R_r}$ in the rotated frame can be obtained from (9.6) by the replacement $(\xi,\xi',\gamma)\to(\xi_\theta,\xi'_{-\theta},\gamma_\theta)$ and $(\tau,R_r,R_l)\to(\tau_\theta,R_{r,\theta},R_{l\theta})$:

$$\overline{R_{r,\theta}}=-\xi_\theta+\gamma_\theta^2\frac{\rho_{r,\theta}}{1-\rho_{r,\theta}\xi_\theta},\quad \rho_{r,\theta}\equiv R_{r,\theta}+\frac{\tau_\theta^2\xi'_{-\theta}}{1-R_{l,\theta}\xi'_{-\theta}}. \tag{9.8}$$

On the other hand, since $\mathscr{J}_1$ is invariant with respect to θ [see (8.19)], we can write (9.4) as

$$\overline{R_r}=\gamma_\theta^2\frac{\partial}{\partial\xi_\theta}\log\overline{\tau}, \tag{9.9}$$

and using the last expression of (8.20) we obtain

$$\overline{R_r}=\sin\theta-\xi_\theta\cos\theta+\gamma_\theta^2\frac{\rho_{r,\theta}}{1-\rho_{r,\theta}\xi_\theta}. \tag{9.10}$$

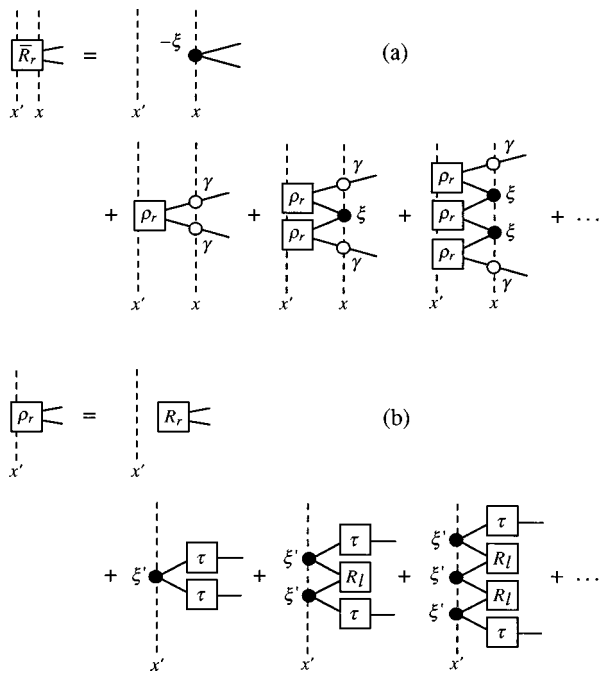

FIG. 24. Graphical representation of Eqs. (9.7).

Comparing (9.10) and (9.8) we find that $\overline{R_r}$ and $\overline{R_{r,\theta}}$ are almost identical; the difference is a term proportional to ξ_θ and a constant term, both independent of the potential $V(x)$:

$$\overline{R_r}=\overline{R_{r,\theta}}+(1-\cos\ \theta)\xi_\theta+\sin\ \theta. \tag{9.11}$$

We can do without $\overline{R_{r,\theta}}$ since it is essentially the same as $\overline{R_r}$. It is enough if we deal with only $\overline{R_r}$. The quantities $\rho_{r,\theta}$ and $R_{r,\theta}$ can be obtained from R_r as

$$\rho_{r,\theta}=\left.\frac{\overline{R_r}-\sin\ \theta}{\cos\ \theta}\right|_{\xi_\theta=0},\quad R_{r,\theta}=\rho_{r,\theta}|_{\xi'_{-\theta}=0}, \tag{9.12}$$

where we have used $\gamma_\theta^2|_{\xi_\theta=0}=\cos\ \theta$ [Eq. (8.18)]. From (9.12) and (7.19) we have

$$R_{r,\theta}=\left.\frac{\overline{R_r}-\sin\ \theta}{\cos\ \theta}\right|_{\xi=-\tan(\theta/2),\xi'=\tan(\theta/2)}. \tag{9.13}$$

This result can also be obtained from (8.24b) with (9.9).

In the same way we can derive expressions for the left reflection coefficient. We have

$$\overline{R_l}=\gamma'^2\frac{\partial}{\partial\xi'}\log\ \overline{\tau}=2\mathscr{J}_1'\log\ \overline{\tau},\quad R_{l,\theta}=\left.\frac{\overline{R_l}+\sin\ \theta}{\cos\ \theta}\right|_{\xi=-\tan(\theta/2),\xi'=\tan(\theta/2)}. \tag{9.14}$$

X. IDENTIFICATION OF $\tilde{\tau}$, $\tilde{R}_r$, AND $\tilde{R}_l$

Here we consider $\widetilde{\tau}$, $\widetilde{R}_r$, and $\widetilde{R}_l$ defined by (2.4). We found in Sec. VI that a linear potential $V(x)=2cx$ can be nullified by the transformation $P(\theta)$ [Eq. (6.5)] with $\theta=\arctan(-ic/k)$. So, in order to deal with the potential $\overline{V}$ given by (2.3) it seems appropriate to assign different rotation angles to both endpoints of the interval. Generalizing (6.7) we define the ''twisted'' evolution operator as

$$U^{(\nu)}_{\theta\theta'}\equiv P^{(\nu)}(-\theta)U^{(\nu)}P^{(\nu)}(\theta')=e^{-R_{r,\theta\theta'}J_+}[\tau_{\theta\theta'}]^{2J_3}e^{R_{l,\theta\theta'}J_-}. \tag{10.1}$$

By the last expression of (10.1) we defined the twisted scattering coefficients $\tau_{\theta\theta'}$, $R_{r,\theta\theta'}$, and $R_{l,\theta\theta'}$. [Cf. Eq. (6.14).] It is not difficult to show that

$$\tau_{\theta\theta'}=\left.\frac{\overline{\tau}}{(\cos\,\theta\,\cos\,\theta')^{1/2}}\right|_{\xi=-\tan(\theta/2),\xi'=\tan(\theta'/2)}, \tag{10.2a}$$

$$R_{r,\theta\theta'}=\left.\frac{\overline{R}_r-\sin\,\theta}{\cos\,\theta}\right|_{\xi=-\tan(\theta/2),\xi'=\tan(\theta'/2)}, \quad R_{l,\theta\theta'}=\left.\frac{\overline{R}_l+\sin\,\theta'}{\cos\,\theta'}\right|_{\xi=-\tan(\theta/2),\xi'=\tan(\theta'/2)}. \tag{10.2b}$$

[Cf. (8.25), (9.13), and (9.14).] Inversely, we can obtain $\overline{\tau}$, $\overline{R}_r$, and $\overline{R}_l$ from $\tau_{\theta\theta'}$, $R_{r,\theta\theta'}$, and $R_{l,\theta\theta'}$ by setting $\theta=-2\arctan\xi$, $\theta'=2\arctan\xi'$. Thus $\overline{\tau}$, $\overline{R}_r$, and $\overline{R}_l$ are essentially equivalent to $\tau_{\theta\theta'}$, $R_{r,\theta\theta'}$, and $R_{l,\theta\theta'}$, respectively. This can be understood in view of the similarity between $P(\theta)$ and $D(\xi)$ [Eqs. (6.5) and (8.10)].

As a matter of fact, $\widetilde{\tau}$, $\widetilde{R}_r$, and $\widetilde{R}_l$ are proportional to these twisted scattering coefficients. Namely, we have

$$\widetilde{\tau}(x_b,x_a;c_b,c_a)=\left(\frac{1-\tan(\theta_b/2)}{1+\tan(\theta_b/2)}\right)^{1/2}\left(\frac{1+\tan(\theta_a/2)}{1-\tan(\theta_a/2)}\right)^{1/2}\tau_{\theta_b\theta_\alpha}(x_b,x_a), \tag{10.3a}$$

$$\widetilde{R}_r=\frac{1-\tan(\theta_b/2)}{1+\tan(\theta_b/2)}\,R_{r,\theta_b\theta_\alpha}, \quad \widetilde{R}_l=\frac{1+\tan(\theta_a/2)}{1-\tan(\theta_a/2)}\,R_{l,\theta_b\theta_a}, \tag{10.3b}$$

where $\theta_a=\arctan(-ic_a/ik)$, $\theta_b=\arctan(-ic_b/ik)$. A derivation of (10.3) is given in Appendix D. Therefore, we find that the set of coefficients $(\widetilde{\tau},\widetilde{R}_r,\widetilde{R}_l)$ is essentially equivalent to $(\overline{\tau},\overline{R}_r,\overline{R}_l)$.

XI. POWER SERIES EXPRESSIONS OF SCATTERING COEFFICIENTS

Equation (3.1) is a general expression that holds for any representation. Any operator U that has the form (3.12) satisfies (3.1). Therefore, the operator $\mathscr{U}$ given by (7.15) also satisfies

$$\frac{\partial}{\partial x}\,\mathscr{U}^{(\nu)}(x,x')=[2ik\mathscr{J}_3^{(\nu)}-2\,f(x)\,\mathscr{J}_1^{(\nu)}]\mathscr{U}^{(\nu)}(x,x'). \tag{11.1}$$

From $\overline{\tau}^{\nu}=\mathscr{U}^{(\nu)}(x,x')I^{(\nu)}(\xi,\xi')$ it is obvious that $\overline{\tau}^{\nu}$ satisfies the same equation:

$$\frac{\partial}{\partial x}\,[\overline{\tau}(x,x';\xi,\xi')]^{\nu}=[2ik\mathscr{J}_3^{(\nu)}-2\,f(x)\,\mathscr{J}_1^{(\nu)}][\overline{\tau}(x,x';\xi,\xi')]^{\nu}. \tag{11.2}$$

Since $\mathscr{U}(x=x')=1$, from (8.1) and (8.17) we obtain the initial condition for (11.2) as

$$[\overline{\tau}(x=x')]^{\nu}=I^{(\nu)}(\xi,\xi')=\left(\frac{\gamma\gamma'}{1-\xi\xi'}\right)^{\nu}=\left[\cosh\left(\frac{w-w'}{2}\right)\right]^{-\nu}. \tag{11.3}$$

The equation for $\log\overline{\tau}$ can be derived by taking the limit $\nu\to0$ in Eq. (11.2). Since

$$\overline{\tau}^{\nu}=1+\nu\,\log\,\overline{\tau}+O(\nu^2), \quad \mathscr{J}_3^{(\nu)}=\mathscr{J}_3^{(0)}+\frac{\nu}{2}\,\frac{1+\xi^2}{1-\xi^2}, \quad \mathscr{J}_1^{(\nu)}=\mathscr{J}_1^{(0)}, \tag{11.4}$$

retaining terms of order ν in Eq. (11.2) we obtain

$$\frac{\partial}{\partial x}\log\,\overline{\tau}(x,x';\xi,\xi')=[2ik\mathscr{J}_3^{(0)}-2\,f(x)\mathscr{J}_1^{(0)}]\log\,\overline{\tau}(x,x';\xi,\xi')+ik\,\frac{1+\xi^2}{1-\xi^2}. \tag{11.5}$$

Note that this equation, unlike Eq. (11.2), contains an inhomogeneous term.

Substituting (8.17) into (8.16) and (8.19) we can write the operators $\mathscr{J}_1^{(\nu)}$ and $\mathscr{J}_3^{(\nu)}$ as

$$\mathscr{J}_1^{(\nu)}=\frac{\partial}{\partial w},\quad \mathscr{J}_3^{(\nu)}=\frac{1}{2}\left[e^{w}\left(\frac{\nu}{2}+\frac{\partial}{\partial w}\right)+e^{-w}\left(\frac{\nu}{2}-\frac{\partial}{\partial w}\right)\right]. \tag{11.6}$$

Therefore, equation (11.2) reads

$$\frac{\partial}{\partial x}\,\overline{\tau}^{\nu}=ik\left[e^{w}\left(\frac{\nu}{2}+\frac{\partial}{\partial w}\right)+e^{-w}\left(\frac{\nu}{2}-\frac{\partial}{\partial w}\right)\right]\overline{\tau}^{\nu}-2\,f\,\frac{\partial}{\partial w}\,\overline{\tau}^{\nu}. \tag{11.7}$$

We can eliminate the last term on the right-hand side of (11.7) by changing the independent variables from (x,x',w,w') [or (x,x',ξ,ξ')] to (x,x',W,W'), where

$$W\equiv V(x)+w,\quad W'\equiv V(x')+w'. \tag{11.8}$$

By this change of variables the partial derivatives change as

$$\frac{\partial}{\partial x}\to\frac{\partial}{\partial x}+\frac{dV}{dx}\,\frac{\partial}{\partial W}=\frac{\partial}{\partial x}-2\,f(x)\,\frac{\partial}{\partial W},\quad \frac{\partial}{\partial w}\to\frac{\partial}{\partial W}, \tag{11.9}$$

and Eq. (11.7) becomes

$$\frac{\partial}{\partial x}\,\overline{\tau}^{\nu}=ik[e^{-V(x)}\hat{\mathscr{J}}_+^{(\nu)}+e^{V(x)}\hat{\mathscr{J}}_-^{(\nu)}]\overline{\tau}^{\nu}, \tag{11.10}$$

where

$$\hat{\mathscr{J}}_{\pm}^{(\nu)}\equiv e^{\pm W}\left(\frac{\nu}{2}\pm\frac{\partial}{\partial W}\right),\quad \hat{\mathscr{J}}_3^{(\nu)}\equiv\frac{\partial}{\partial W}. \tag{11.11}$$

In (11.10), and in the rest of this section, the independent variables are (x,x',W,W'). It is obvious that the operators $\hat{\mathscr{J}}_+$, $\hat{\mathscr{J}}_-$, and $\hat{\mathscr{J}}_3$ satisfy the commutation relations (3.2). Note that we are now working in the rotated frame with $\theta=\pm\pi/2$. Indeed, the operators (11.11) can be expressed in terms of $\mathscr{J}_{a,\theta}$ [Eqs. (7.20)] as

$$\hat{\mathscr{J}}_{\pm}^{(\nu)}=-e^{\pm V(x)}\mathscr{J}_{\pm,(\pi/2)}^{(\nu)}=e^{\pm V(x)}\mathscr{J}_{\mp,-(\pi/2)}^{(\nu)},\quad \hat{\mathscr{J}}_3^{(\nu)}=\mathscr{J}_{3,(\pi/2)}^{(\nu)}=-\mathscr{J}_{3,-(\pi/2)}^{(\nu)}. \tag{11.12}$$

It is also convenient to define, in correspondence with (11.11),

$$\hat{\xi}_+\equiv e^{W},\quad \hat{\xi}_-\equiv-e^{-W},\quad \hat{\xi}'_+\equiv e^{W'},\quad \hat{\xi}'_-\equiv-e^{-W'}. \tag{11.13}$$

From (7.19) and (8.8) we find $\xi_{\pm(\pi/2)}=\pm e^{\pm w}$. So (11.13) means $\hat{\xi}_{\pm}=e^{\pm V(x)}\xi_{\pm(\pi/2)}$. In terms of $\hat{\xi}_+$ or $\hat{\xi}_-$ the operators $\hat{\mathscr{J}}_{\pm}$ and $\hat{\mathscr{J}}_3$ have the form

$$\hat{\mathscr{J}}_{\pm}^{(\nu)}=(\hat{\xi}_+)^{\pm1}\left(\frac{\nu}{2}\pm\hat{\xi}_+\,\frac{\partial}{\partial\hat{\xi}_+}\right)=-(\hat{\xi}_-)^{\mp1}\left(\frac{\nu}{2}\mp\hat{\xi}_-\,\frac{\partial}{\partial\hat{\xi}_-}\right),\quad \hat{\mathscr{J}}_3^{(\nu)}=\hat{\xi}_+\,\frac{\partial}{\partial\xi_+}=-\hat{\xi}_-\,\frac{\partial}{\partial\xi_-}. \tag{11.14}$$

Similarly, after the change of independent variables to (W,W'), equation (11.5) becomes

$$\frac{\partial}{\partial x}\log\,\overline{\tau}=ik[e^{-V(x)}\hat{\mathscr{J}}_+^{(0)}+e^{V(x)}\hat{\mathscr{J}}_-^{(0)}]\log\,\overline{\tau}+ik\,\cosh[W-V(x)], \tag{11.15}$$

where we have used $(1+\xi^2)/(1-\xi^2)=\cosh w$. In much the same way we can also derive

$$\frac{\partial}{\partial x'}\,\overline{\tau^{\nu}}=-ik[e^{-V(x')}\hat{\mathscr{J}}'^{(\nu)}_{+}+e^{V(x')}\hat{\mathscr{J}}'^{(\nu)}_{-}]\hat{\tau}^{\nu}, \tag{11.16}$$

$$\frac{\partial}{\partial x'}\log\,\overline{\tau}=-ik[e^{-V(x')}\hat{\mathscr{J}}'^{(0)}_{+}+e^{V(x')}\hat{\mathscr{J}}'^{(0)}_{-}]\log\,\overline{\tau}-ik\,\cosh[W'-V(x')], \tag{11.17}$$

where $\hat{\mathscr{J}}'_{+}$ and $\hat{\mathscr{J}}'_{-}$ are defined by (11.11) with W replaced by W'. (Note the prime on $\hat{\mathscr{J}}$.)

Equations for $\overline{R_r}$ and $\overline{R_l}$ can be obtained from (11.15) and (11.17). In terms of the variables W and W', equations (9.4) and (9.14) can be expressed simply as

$$\overline{R_r}=2\frac{\partial}{\partial W}\log\,\overline{\tau},\quad \overline{R_l}=2\frac{\partial}{\partial W'}\log\,\overline{\tau}. \tag{11.18}$$

Differentiating Eq. (11.15) [or Eq. (11.17)] with respect to W (or W'), and using the relation $(\partial/\partial W)\hat{\mathscr{J}}^{(0)}_{\pm}=\hat{\mathscr{J}}^{(2)}_{\pm}(\partial/\partial W)$, which can be easily derived from (11.11), we obtain

$$\frac{\partial}{\partial x}\,\overline{R_r}=ik[e^{-V(x)}\hat{\mathscr{J}}^{(2)}_{+}+e^{V(x)}\hat{\mathscr{J}}^{(2)}_{-}]\overline{R_r}+2ik\,\sinh[W-V(x)], \tag{11.19a}$$

$$\frac{\partial}{\partial x'}\,\overline{R_r}=-ik[e^{-V(x')}\hat{\mathscr{J}}'^{(0)}_{+}+e^{V(x')}\hat{\mathscr{J}}'^{(0)}_{-}]\overline{R_r}, \tag{11.19b}$$

$$\frac{\partial}{\partial x}\,\overline{R_l}=ik[e^{-V(x)}\hat{\mathscr{J}}^{(0)}_{+}+e^{V(x)}\hat{\mathscr{J}}^{(0)}_{-}]\overline{R_l}, \tag{11.19c}$$

$$\frac{\partial}{\partial x'}\,\overline{R_l}=-ik[e^{-V(x')}\hat{\mathscr{J}}'^{(2)}_{+}+e^{V(x')}\hat{\mathscr{J}}'^{(2)}_{-}]\overline{R_l}-2ik\,\sinh[W'-V(x')]. \tag{11.19d}$$

It should be noted that $\hat{\mathscr{J}}^{(2)}_{+}$ and $\hat{\mathscr{J}}^{(2)}_{-}$ appear in (11.19a) instead of $\hat{\mathscr{J}}^{(0)}_{+}$ and $\hat{\mathscr{J}}^{(0)}_{-}$. This is quite understandable with the diagrammatic interpretation of $\overline{R_r}$. Each diagram contributing to $\overline{R_r}$ has two right legs and no left legs. That is the reason why the representation is $\nu=2$ in (11.19a) and $\nu=0$ in (11.19b).

The equations for $\overline{\tau}$ are obtained by setting $\nu=1$ in (11.10) and (11.16):

$$\frac{\partial}{\partial x}\,\overline{\tau}=ik[e^{-V(x)}\hat{\mathscr{J}}^{(1)}_{+}+e^{V(x)}\hat{\mathscr{J}}^{(1)}_{-}]\overline{\tau},\quad \frac{\partial}{\partial x'}\,\overline{\tau}=-ik[e^{-V(x')}\hat{\mathscr{J}}'^{(1)}_{+}+e^{V(x')}\hat{\mathscr{J}}'^{(1)}_{-}]\overline{\tau}. \tag{11.20}$$

Equations (11.19) and (11.20) are the basic equations for the generalized scattering coefficients. To solve these equations, we need to know the initial condition. From (11.3) and (11.8) we have

$$\overline{\tau}(x=x')=\left[\cosh\left(\frac{w-w'}{2}\right)\right]^{-1}\Bigg|_{x=x'}=\left[\cosh\left(\frac{W-W'}{2}\right)\right]^{-1}\equiv\tau_0, \tag{11.21a}$$

$$\overline{R_r}(x=x')=2\frac{\partial}{\partial W}\log\,\overline{\tau}(x=x')=-\tanh\left(\frac{W-W'}{2}\right)\equiv r_0, \tag{11.21b}$$

$$\overline{R_l}(x=x')=2\frac{\partial}{\partial W'}\log\,\overline{\tau}(x=x')=\tanh\left(\frac{W-W'}{2}\right)\equiv l_0. \tag{11.21c}$$

We define the operators $\mathscr{L}$ and $\mathscr{L}'$ as

$$\mathscr{L}^{(\nu)}g(x,x';W,W')\equiv\int_{x'}^{x}dz[e^{-V(z)}\hat{\mathscr{J}}^{(\nu)}_{+}+e^{V(z)}\hat{\mathscr{J}}^{(\nu)}_{-}]g(z,x';W,W'), \tag{11.22a}$$

$$\mathscr{L}'^{(\nu)}g(x,x';W,W')\equiv\int_{x'}^{x}dz[e^{-V(z)}\hat{\mathscr{T}}_{+}'^{(\nu)}+e^{V(z)}\hat{\mathscr{T}}_{-}'^{(\nu)}]g(x,z;W,W'),\tag{11.22b}$$

for an arbitrary function g. They are integral operators with respect to x or x', and differential operators with respect to W or W'. The differential equations (11.20) and (11.19) with the initial condition (11.21) are equivalent to

$$\overline{\tau}=\tau_0+ik\mathscr{L}^{(1)}\overline{\tau},\quad \overline{\tau}=\tau_0+ik\mathscr{L}'^{(1)}\overline{\tau},\tag{11.23a}$$

$$\overline{R_r}=r_0+ik\mathscr{L}^{(2)}\overline{R_r}+2ik\int_{x'}^{x}\sinh[W-V(z)]dz,\quad \overline{R_r}=r_0+ik\mathscr{L}'^{(0)}\overline{R_r},\tag{11.23b}$$

$$\overline{R_l}=l_0+ik\mathscr{L}^{(0)}\overline{R_l},\quad \overline{R_l}=l_0+ik\mathscr{L}'^{(2)}\overline{R_l}+2ik\int_{x'}^{x}\sinh[W'-V(z)]dz.\tag{11.23c}$$

Let us derive expressions of scattering coefficients as power series in terms of k. We write

$$\overline{\tau}=\sum_{n=0}^{\infty}\tau_n k^n,\quad \overline{R_r}=\sum_{n=0}^{\infty}r_n k^n,\quad \overline{R_l}=\sum_{n=0}^{\infty}l_n k^n.\tag{11.24}$$

The coefficients of expansion τ_n, r_n, and l_n can be obtained by substituting (11.24) into (11.23). The coefficients for $n=0$ are given by (11.21). Using (11.13) we can also express them in terms of $\hat{\xi}_{\pm}$ and $\hat{\xi}'_{\mp}$. It is convenient to use either $(\hat{\xi}_{+},\hat{\xi}'_{-})$ or $(\hat{\xi}_{-},\hat{\xi}'_{+})$ as a set. We can write

$$\tau_0=\frac{2(-\hat{\xi}_{\pm}\hat{\xi}'_{\mp})^{1/2}}{1-\hat{\xi}_{\pm}\hat{\xi}'_{\mp}}=2i\sum_{m=0}^{\infty}(\hat{\xi}_{\pm}\hat{\xi}'_{\mp})^{m+(1/2)},\tag{11.25a}$$

$$r_0=-l_0=\pm\frac{1+\hat{\xi}_{\pm}\hat{\xi}'_{\mp}}{1-\hat{\xi}_{\pm}\hat{\xi}'_{\mp}}=\pm1\pm2\sum_{m=1}^{\infty}(\hat{\xi}_{\pm}\hat{\xi}'_{\mp})^{m}.\tag{11.25b}$$

To express τ_n, r_n, and $l_n(n\geqslant1)$ in an explicit form, we introduce the notation

$$[s_1,s_2,\ldots,s_n]_{x'}^{x}\equiv\int_{x'}^{x}dz_1\int_{z_1}^{x}dz_2\cdots\int_{z_{n-1}}^{x}dz_n\,\exp\left[\sum_{i=1}^{n}s_iV(z_i)\right],\tag{11.26}$$

where $s_i=\pm1$. The domain of integration in (11.26) is $x'\leqslant z_1\leqslant z_2\leqslant\cdots\leqslant z_n\leqslant x$. From the two equations of (11.23a) we obtain, respectively,

$$\tau_n=i^n(\mathscr{L}^{(1)})^n\tau_0=i^n\sum_{s_1=\pm1}\cdots\sum_{s_n=\pm1}[s_1,\ldots,s_n]_{x'}^{x}\hat{\mathscr{T}}_{-s_n}^{(1)}\cdots\hat{\mathscr{T}}_{-s_1}^{(1)}\tau_0,\tag{11.27a}$$

$$\tau_n=i^n(\mathscr{L}'^{(1)})^n\tau_0=i^n\sum_{s_1=\pm1}\cdots\sum_{s_n=\pm1}[s_1,\ldots,s_n]_{x'}^{x}\hat{\mathscr{T}}'^{(1)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(1)}_{-s_n}\tau_0.\tag{11.27b}$$

Here $\mathscr{T}_{-s}^{(\nu)}$ stands for $\mathscr{T}_{-}^{(\nu)}$ and $\mathscr{T}_{+}^{(\nu)}$ for $s=+1$ and $s=-1$, respectively. (Note that $\mathscr{T}_{-s}^{(\nu)}$ with $s=-1$ is not $\mathscr{T}_{1}^{(\nu)}$ but $\mathscr{T}_{+}^{(\nu)}$.) For R_r we obtain two expressions of different form. The second equation of (11.23b) yields

$$r_n=i^n(\mathscr{L}'^{(0)})^nr_0=i^n\sum_{s_1=\pm1}\cdots\sum_{s_n=\pm1}[s_1,\ldots,s_n]_{x'}^{x}\hat{\mathscr{T}}'^{(0)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(0)}_{-s_n}r_0.\tag{11.28}$$

On the other hand, from the first equation of (11.23b) we obtain

$$r_1=i\mathscr{L}^{(2)}r_0+ie^{W}[-1]_{x'}^{x}-ie^{-W}[+1]_{x'}^{x}=i\sum_{s_1\pm1}[s_1]_{x'}^{x}(\hat{\mathscr{T}}_{-s_1}^{(2)}r_0+\hat{\xi}_{-s_1}),\tag{11.29a}$$

$$r_n = i^{n-1}(\mathscr{L}^{(2)})^{n-1} r_1 = i^n \sum_{s_1=\pm 1} \cdots \sum_{s_n=\pm 1} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}^{(2)}_{-s_n} \cdots \hat{\mathscr{T}}^{(2)}_{-s_2} (\hat{\mathscr{T}}^{(2)}_{-s_1} r_0 + \hat{\xi}_{-s_1}), \tag{11.29b}$$

where $\hat{\xi}_{-s_1}$ means $\hat{\xi}_{\mp}$ for $s_1 = \pm 1$. Similar results for $\overline{R_l}$ can be obtained from (11.23c). To summarize, τ, $\overline{R_r}$, and $\overline{R_l}$ can be expressed as

$$\overline{\tau} = \tau_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}^{(1)}_{-s_n} \cdots \hat{\mathscr{T}}^{(1)}_{-s_1} \tau_0 \tag{11.30a}$$

$$= \tau_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}'^{(1)}_{-s_1} \cdots \hat{\mathscr{T}}'^{(1)}_{-s_n} \tau_0, \tag{11.30b}$$

$$\overline{R_r} = r_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}^{(2)}_{-s_n} \cdots \hat{\mathscr{T}}^{(2)}_{-s_2} (\hat{\mathscr{T}}^{(2)}_{-s_1} r_0 + \hat{\xi}_{-s_1}) \tag{11.30c}$$

$$= r_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}'^{(0)}_{-s_1} \cdots \hat{\mathscr{T}}'^{(0)}_{-s_n} r_0, \tag{11.30d}$$

$$\overline{R_l} = l_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}^{(0)}_{-s_n} \cdots \hat{\mathscr{T}}^{(0)}_{-s_1} l_0 \tag{11.30e}$$

$$= l_0 + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}'^{(2)}_{-s_1} \cdots \hat{\mathscr{T}}'^{(2)}_{-s_{n-1}} (\hat{\mathscr{T}}'^{(2)}_{-s_n} l_0 + \hat{\xi}'_{-s_n}). \tag{11.30f}$$

Here $\Sigma_{\{s_1,\ldots,s_n\}}$ stands for $\Sigma_{s_1=\pm 1} \cdots \Sigma_{s_n=\pm 1}$. Equations (11.30a) and (11.30b) can be generalized to arbitrary ν as

$$\overline{\tau^{\nu}} = \tau_0^{\nu} + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}^{(\nu)}_{-s_n} \cdots \hat{\mathscr{T}}^{(\nu)}_{-s_1} \tau_0^{\nu} \tag{11.31a}$$

$$= \tau_0^{\nu} + \sum_{n=1}^{\infty} i^n k^n \sum_{\{s_1,\ldots,s_n\}} [s_1,\ldots,s_n]^x_{x'} \hat{\mathscr{T}}'^{(\nu)}_{-s_1} \cdots \hat{\mathscr{T}}'^{(\nu)}_{-s_n} \tau_0^{\nu}. \tag{11.31b}$$

Using (11.25) and (11.14) we can write the expressions (11.30) in terms of either $(\hat{\xi}_+, \hat{\xi}'_-)$ or $(\hat{\xi}_-, \hat{\xi}'_+)$. The original scattering coefficients $\tau(x,x')$, $R_r(x,x')$, and $R_l(x,x')$ are obtained from these expressions by setting $\hat{\xi}_+ = e^{V(x)}$ and $\hat{\xi}'_- = -e^{-V(x')}$, or $\hat{\xi}_- = -e^{-V(x)}$ and $\hat{\xi}'_+ = e^{V(x')}$.

XII. SCATTERING COEFFICIENTS FOR $\theta = \pm \pi/2$

In Sec. XI we studied power series expansion of $\overline{\tau}$, $\overline{R_r}$, and $\overline{R_l}$. The coefficients of expansion are written in terms of the integrals of the form (11.26). When the potential $V(x)$ is given, all we need to do is to evaluate such integrals by some means. As a matter of fact, it is not necessary to calculate all these integrals. As we shall see below, it is sufficient if we evaluate only the integrals of the form $[s_1,\ldots,s_n]$ with $\Sigma_{i=1}^n s_i = 0$, 1, and -1. All the terms that appear in the expressions (11.30) can be obtained from them.

As mentioned in Sec. VI, when dealing with power series expressions in terms of k it is natural to work in the rotated frame with $\theta = \pm \pi/2$. In this section we study the scattering coefficients for $\theta = \pm \pi/2$ in order to clarify the structure of the power series expressions. What was said in the above paragraph can also be verified through this analysis.

It is convenient to define

$$\hat{\tau}^{\pm}(x,x') = e^{\mp [V(x) - V(x')]/2} \tau_{\pm(\pi/2)}(x,x'), \tag{12.1a}$$

$$\hat{R}_r^{\pm}(x,x')=e^{\mp V(x)}R_{r,\pm(\pi/2)}(x,x'),\quad \hat{R}_l^{\pm}(x,x')=e^{\pm V(x')}R_{l,\pm(\pi/2)}(x,x'),\tag{12.1b}$$

$$\hat{\rho}_r^{\pm}\equiv e^{\mp V(x)}\rho_{r,\pm(\pi/2)},\quad \hat{\rho}_l^{\pm}\equiv e^{\pm V(x')}\rho_{l,\pm(\pi/2)}.\tag{12.1c}$$

In the following analysis we deal with $\theta=+\pi/2$. The expressions for $\theta=-\pi/2$ can be easily obtained from these results by considering the symmetry. Setting $\theta=\pi/2$ in (8.20) and (9.10), and using (8.18) we obtain

$$\overline{\tau}=\frac{2(-\hat{\xi}_+\hat{\xi}'_-)^{1/2}\hat{\tau}^+}{1-\hat{R}_r^+\hat{\xi}_+-\hat{R}_l^+\hat{\xi}'_--[(\hat{\tau}^+)^2-\hat{R}_r^+\hat{R}_l^+]\hat{\xi}_+\hat{\xi}'_-},\tag{12.2}$$

$$\overline{R}_r=\frac{1+\hat{\rho}_r^+\hat{\xi}_+}{1-\hat{\rho}_r^+\hat{\xi}_+}=1+2\sum_{n=1}^{\infty}(\hat{\rho}_r^+\hat{\xi}_+)^n,\quad \hat{\rho}_r^+=\hat{R}_r^+ +\frac{(\hat{\tau}^+)^2\hat{\xi}'_-}{1-\hat{R}_l^+\hat{\xi}'_-},\tag{12.3}$$

and similarly for $\overline{R_l}$. From (12.3) it follows that

$$(\hat{\rho}_r^+)^n=\frac{1}{2n!}\frac{\partial^n\overline{R}_r}{\partial\hat{\xi}_+^n}\bigg|_{\hat{\xi}_+=0},\quad (\hat{R}_r^+)^n=\frac{1}{2n!}\frac{\partial^n\overline{R}_r}{\partial\hat{\xi}_+^n}\bigg|_{\hat{\xi}_+=\hat{\xi}'_-=0}\tag{12.4}$$

for any positive integer n. Since

$$\frac{1}{2n!}\frac{\partial^n}{\partial\hat{\xi}_+^n}r_0\bigg|_{\hat{\xi}_+=0}=(\hat{\xi}'_-)^n$$

[Eq. (11.25b)], from (12.4) we find that $(\hat{\rho}_r^+)^n$ is obtained by replacing r_0 in Eq. (11.30d) with $(\hat{\xi}'_-)^n$:

$$(\hat{\rho}_r^+)^n=(\hat{\xi}'_-)^n+\sum_{m=1}^{\infty}(ik)^m\sum_{\{s_1,\ldots,s_m\}}[s_1,\ldots,s_m]_{x'}^{x}\hat{\mathscr{T}}'^{(0)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(0)}_{-s_m}(\hat{\xi}'_-)^n.\tag{12.5}$$

We obtain $(\hat{R}_r^+)^n$ by setting $\hat{\xi}'_-=0$ in (12.5). Since the operators $\hat{\mathscr{T}}'^{(0)}_-$ and $\hat{\mathscr{T}}'^{(0)}_+$ raises and lowers, respectively, the power of $\hat{\xi}'_-$ by one, setting $\hat{\xi}'_-=0$ in (12.5) is equivalent to imposing the restriction $\Sigma_{i=1}^m s_i=-n$ on the summation. Thus we have

$$(\hat{R}_r^+)^n=\sum_{m=1}^{\infty}(ik)^m\sum_{\substack{\{s_1,\ldots,s_m\}\\ \Sigma_{i=1}^m s_i=-n}}[s_1,\ldots,s_m]_{x'}^{x}\hat{\mathscr{T}}'^{(0)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(0)}_{-s_m}(\hat{\xi}'_-)^n,\tag{12.6}$$

for arbitrary positive integer n. In (12.6), only even (odd) m's survive in the summation for even (odd) n.

It is also possible to express $\hat{R}_r^+$ in terms of operators in the $\nu=2$ representation. From (12.3) it follows that

$$\overline{R}_r|_{\hat{\xi}'_-=0}=1+2\sum_{n=1}^{\infty}(\hat{R}_r^+\hat{\xi}_+)^n.\tag{12.7}$$

On the other hand, setting $\hat{\xi}'_-=0$ in Eq. (11.30c) we obtain

$$\overline{R}_r|_{\hat{\xi}'_-=0}=1+2\sum_{m=1}^{\infty}(ik)^m\sum_{\{s_2,\ldots,s_m\}}[-1,s_2,\ldots,s_m]_{x'}^{x}\hat{\mathscr{T}}^{(2)}_{-s_m}\cdots\hat{\mathscr{T}}^{(2)}_{-s_2}\hat{\xi}_+.\tag{12.8}$$

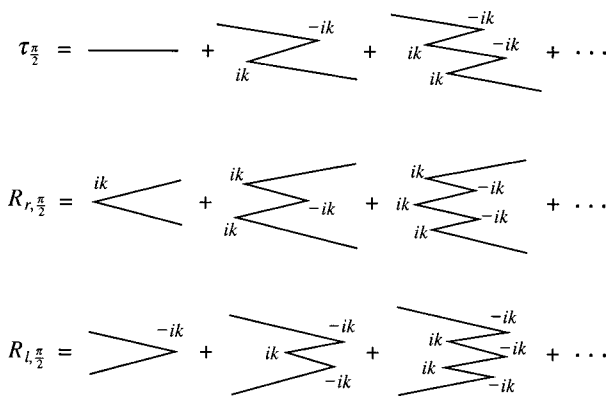

FIG. 25. Diagrammatic representation of the power series expression of the scattering coefficients for $\theta=\pi/2$.

(The term of order $m=1$ in the summation is $2ik[-1]^x_{x'}\hat{\xi}_+$.) The operators $\hat{\mathscr{T}}^{(2)}_+$ and $\hat{\mathscr{T}}^{(2)}_-$ raises and lowers, respectively, the power of $\hat{\xi}_+$ by one. Therefore by comparing (12.7) with (12.8) we find

$$(\hat{R}^+_r\hat{\xi}_+)^n=\sum_{m=1}^{\infty}(ik)^m\sum_{\substack{\{s_2,\ldots,s_m\}\\ \Sigma_{i=2}^m s_i=1-n}}[-1,s_2,\ldots,s_m]^x_{x'}\hat{\mathscr{T}}^{(2)}_{-s_m}\cdots\hat{\mathscr{T}}^{(2)}_{-s_2}\hat{\xi}_+. \tag{12.9}$$

[As can be seen from (12.9), only the terms with $s_1=-1$ survive in the expression (12.6), too. The quantity $\hat{\mathscr{T}}'^{(0)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(0)}_{-s_m}(\hat{\xi}'_-)^n$ vanishes for $s_1=+1$ if $\Sigma_{i=1}^m s_i=-n$.]

It is easy to derive the expressions for $\hat{\tau}^+$. As can be seen from (12.2), we have

$$\hat{\tau}^+=\tfrac{1}{2}(-\hat{\xi}_+\hat{\xi}'_-)^{-1/2}\overline{\tau}\big|_{\hat{\xi}_+=\hat{\xi}'_-=0}. \tag{12.10}$$

From (11.30a), (11.30b), and (11.25a) we obtain

$$\hat{\tau}^+=(\hat{\xi}_+)^{-1/2}\sum_{j=0}^{\infty}(ik)^{2j}\sum_{\substack{\{s_i,\ldots,s_{2j}\}\\ \Sigma_{i=0}^{2j}s_i=0}}[s_1,\ldots,s_{2j}]^x_{x'}\hat{\mathscr{T}}^{(1)}_{-s_{2j}}\cdots\hat{\mathscr{T}}^{(1)}_{-s_1}(\hat{\xi}_+)^{1/2} \tag{12.11a}$$

$$=(\hat{\xi}'_-)^{-1/2}\sum_{j=0}^{\infty}(ik)^{2j}\sum_{\substack{\{s_i,\ldots,s_{2j}\}\\ \Sigma_{i=0}^{2j}s_i=0}}[s_1,\ldots,s_{2j}]^x_{x'}\hat{\mathscr{T}}'^{(1)}_{-s_1}\cdots\hat{\mathscr{T}}'^{(1)}_{-s_{2j}}(\hat{\xi}'_-)^{1/2}. \tag{12.11b}$$

Only even powers of k are present in (12.11) because of the restriction $\Sigma s_i=0$.

To be concrete, let us write down lower order terms of $\hat{\tau}^+$, $\hat{R}^+_r$, and $\hat{R}^+_l$ explicitly. From the above results we have

$$\hat{\tau}^+=1+[--+]k^2+(4[----++]+[--+--+])k^4+\cdots, \tag{12.12a}$$

$$\hat{R}^+_r=[-]ik+2[---+]ik^3+(12[-----++]+4[---+--+])ik^5+\cdots, \tag{12.12b}$$

$$\hat{R}^+_l=-[+]ik-2[-++]ik^3-(12[-----+++]+4[--+--+++])ik^5+\cdots, \tag{12.12c}$$

where $[--+]$, etc., is the shorthand notation for $[-1,-1,+1]^x_{x'}$, etc. The expressions for $\hat{\tau}^-$, $\hat{R}^-_r$, and $\hat{R}^-_l$ are obtained from (12.12) by interchanging the plus and minus signs in $[\cdots]$, and replacing k by $-k$. The quantities $\hat{\tau}^+$, $\hat{R}^+_r$, and $\hat{R}^+_l$ are the scattering coefficients for $\theta=\pi/2$ apart from the factors $e^{-[V(x)-V(x')]/2}$, $e^{-V(x)}$, and $e^{V(x')}$, respectively. So Eqs. (12.12) can be diagrammatically expressed as in Fig. 25. The diagrams in Fig. 25 are to be interpreted according

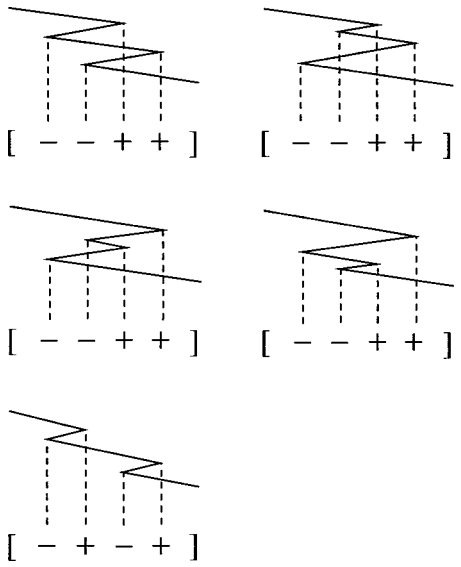

FIG. 26. The correspondence between the diagrams of order k^4 in $\hat{\tau}^+$ and the expression $4[--++]+[-+-+]$.

to the rules in Fig. 20 with $\theta=\pi/2$. Each line element connecting the points x_a and x_b stands for $\exp[X_{\pi/2}(x_b)-X_{\pi/2}(x_a)]=\exp[\frac{1}{2}V(x_b)-\frac{1}{2}V(x_a)]$, and the scattering factor is $\mp ik$, as explained in Sec. VI. As an example, Fig. 26 explains how the diagrams for the term of order k^4 in $\hat{\tau}^+$ correspond to the expression $4[--++]+[-+-+]$ as in Eq. (12.12a).

Thus τ, R_r, and R_l can be expressed in terms of $\hat{\tau}^+$, $\hat{R}_r^+$, and $\hat{R}_l^+$, (or $\hat{\tau}^-$, $\hat{R}_r^-$, and $\hat{R}_l^-$,) which include fewer types of integrals. Since the power series expressions of $\bar{\tau}$, R_r, and R_l can be obtained by substituting (12.12) into (12.2) and (12.3), we can see that the coefficients τ_n, r_n, and l_n in (11.24) can also be expressed in terms of integrals $[s_1,\ldots,s_n]$ where $\Sigma s_i=0$, $+1$, or -1. Therefore in practical problems we have only to evaluate $[s_1,\ldots,s_n]$ with $\Sigma s_i=0$, $+1$, and -1.

XIII. REFLECTION COEFFICIENT FOR SEMI-INFINITE INTERVALS

Until now we have studied the scattering coefficients for the interval (x',x) assuming that both x and x' are finite. However, $R_r(x,x')$ is well-defined even for $x'=-\infty$. [Similarly, $R_l(x,x')$ is well-defined for $x=+\infty$.] Indeed, for a semi-infinite interval that has only the right endpoint, the right-reflection coefficient R_r is the only quantity that is ''experimentally'' measurable from outside. Here let us think about R_r for the semi-infinite interval $(-\infty,x)$.

We first consider the case where the potential converges to a finite limit at minus infinity; $\lim_{z\to-\infty}V(z)=V(-\infty)$, where $V(-\infty)$ is a finite value. In this case, the quantities $\hat{\tau}^+$, $\hat{R}_r^+$, and $\hat{R}_l^+$ do not have a definite value. It is easy to see that all the integrals appearing in (12.12), $[-]_{-\infty}^x$, $[+]_{-\infty}^x$, $[-+]_{-\infty}^x$, etc., are divergent. Nevertheless, the reflection coefficient $R_r(-\infty,x)$ itself has a definite value, and it can be expressed as a power series in terms of k with finite expansion coefficients. Although all the integrals $[s_1,\ldots,s_n]_{-\infty}^x$ in (11.30c) or (11.30d) are divergent, the divergence is canceled. This can be understood diagrammatically. Figure 27 shows diagrams contributing to $R_r(x,x')$ in the rotated frame with $\theta=\pi/2$, where a reflection at the endpoint x' gives a factor $\xi'_{-\pi/2}=e^{V(x')}\hat{\xi}'_-$. From this figure it can be seen that the integrals $[-1,s_2,\ldots,s_n]$ and $[+1,s_2,\ldots,s_n](\hat{\xi}'_-)^2$, which correspond to the two diagrams shown here, always appear as a pair with the same weight and the opposite sign. Therefore, the divergence of these two integrals cancel out if we set $\hat{\xi}'_-=-e^{-V(-\infty)}$ [provided that $V(x)$ converges to the finite limit $V(-\infty)$ sufficiently fast at $x\to-\infty$].

Next we consider the case $V(-\infty)=+\infty$, or more precisely, $\lim_{z\to-\infty}V(z)=+\infty$. The expression for the reflection coefficient becomes much simpler in this case. As $x'=-\infty$ now, we have $\hat{\xi}'_-=e^{-V(-\infty)}\xi'_{-(\pi/2)}=0$. (We are assuming that $\xi'_{-(\pi/2)}=-e^{-w'}$ is finite. Recall that R_r is obtained by setting $w=w'=0$.) Thus R_r becomes independent of w'. From (12.3) we have $\hat{\rho}_r^+$

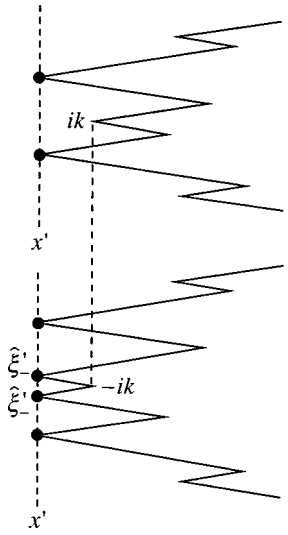

FIG. 27. As shown here, the integrals $[-1,s_2,\ldots,s_n]_{x'}^{x}$ and $[+1,s_2,\ldots,s_n]_{x'}^{x}(\hat{\xi}'_-)^2$ appear as a pair with the same weight and opposite sign in the expression for $R_r(x,x')$ or $\hat{\rho}_r^+(x,x')$.

$=\hat{R}_r^+$, and the expression for $\overline{R_r}$ is given by (12.8). [This holds true not only for semi-infinite intervals but also for the cases where $\lim_{z\downarrow x'}V(z)=+\infty$ at some finite x'. Similarly, if $V(-\infty)=-\infty$ we have $\hat{\xi}'_+=0$ and $\hat{\rho}_r^-=\hat{R}_r^-$.] We find from the first equation of (12.3) and $\hat{\rho}_r^+=\hat{R}_r^+$ that the expression for $\log R_r(x,x')$ has the form

$$\log \overline{R_r}=\log(1+\hat{R}_r^+\hat{\xi}_+)-\log(1-\hat{R}_r^+\hat{\xi}_+)=2\sum_{n=1,3,5,\cdots}\frac{1}{n}(\hat{R}_r^+\hat{\xi}_+)^n. \tag{13.1}$$

The power series expression of $\log \overline{R_r}$ can be obtained by substituting (12.9) into (13.1). After some manipulation we obtain

$$\log \overline{R_r}=2ik[-1]\hat{\xi}_+ +2\sum_{m=3,5,7,\cdots}(ik)^m\sum_{\{s_2,\ldots,s_m\}}[-1,s_2,\ldots,s_{m-1},s_m]\hat{\xi}_{-s_m}\hat{\mathscr{T}}^{(2)}_{-s_{m-1}}\cdots\hat{\mathscr{T}}^{(2)}_{-s_2}\hat{\xi}_+, \tag{13.2}$$

where we have used $\hat{\xi}_-=-(\hat{\xi}_+)^{-1}$. This is the power series expression in (1.18).

XIV. CONCLUDING REMARKS

In the present work we confined ourselves to general features of the scattering formalism, without specifying the form of the potential $V(x)$. The expressions obtained here will serve as a basis on which various approximation methods can be constructed for specific problems. Although we have focused our attention on power series expansion in terms of k, the resulting expressions are not merely approximations for small k. In fact, it is possible to convert the expressions such as (11.30) into large-k expressions through an algebraic manipulation. It is also possible to study various analytic properties in the intermediate-k region on the basis of these expressions.

In this paper we left aside the Green's function. As discussed in a previous paper,[7] the Green's function can be described in terms of a superalgebra, which allows the spin to change at the beginning and the end of a path. The transmission coefficient is also naturally described in terms of this superalgebra. In order to fully understand the scattering processes, it will be necessary to study the supersymmetric structure.

APPENDIX A: DIAGRAMMATIC DERIVATION OF (3.22)

A diagram contributing to $[U^{(\nu)}(x_b,x_a)]_{mn}$ consists of:

(i) ν paths that start from x_a and end at x_b. Each of them amounts to $\tau(x_b,x_a)$ when summed over all possible configurations.
(ii) Additional i pairs of $\tau(x_b,x_a)$'s. [The difference between (i) and (ii) is explained below.] The number i cannot exceed $\min(m,n)$.
(iii) $n-i$ paths that start from x_a and return to x_a, each amounting to $R_l(x_a,x_b)$.
(iv) $m-i$ paths that start from x_b and return to x_b, each amounting to $R_r(x_a,x_b)$.

Let us study the combinatorial factors that arise when we try to connect their left and right legs with $u_n^{(\nu)}$ and $(u_m^{(\nu)})^\dagger$, respectively, so as to make ν connected paths without a loop.

First, we connect the left legs of these elements with $u_n^{(\nu)}$, leaving the right legs disconnected. As in Fig. 10, the ν open circles are distinguished. In addition, we distinguish the i pairs of τ's in (ii). Then there are $\binom{\nu+n-1}{n-i}$ different ways of making such a connection. It is the number of ways to distribute the R_l's among the $\nu+i$ elements in either (i) or (ii). [Each τ belonging to (i) is connected with an open circle of $u_n^{(\nu)}$ either directly or indirectly. The τ's in (ii) are not connected with an open circle of $u_n^{(\nu)}$. Instead, the two τ's of a pair are connected with each other through one or more solid circles of $u_n^{(\nu)}$.] Hence we have the second factor on the right-hand side of (3.22).

Next we connect the right legs with $(u_m^{(\nu)})^\dagger$. Different paths are obtained by reordering of the $m-i$ right-reflection coefficients and the i pairs of τ's [which belong to (ii) above]. The R_r's and the pairs of τ's can be arranged in $\binom{m}{i}$ different ways, and this gives the first factor on the right-hand side of (3.22). (Now the τ-pairs are not distinguished.) Thus we have diagrammatically reproduced the expression (3.22).

APPENDIX B: DERIVATION OF (8.4)

Here we use the formulas

$$C^{J_3}J_\pm = J_\pm C^{J_3\pm 1}, \quad [J_3, e^{CJ_\pm}] = \pm CJ_\pm e^{CJ_\pm}, \tag{B1}$$

$$[J_\pm, e^{CJ_\mp}] = \pm(2CJ_3 \pm C^2 J_\mp)e^{CJ_\mp} = \pm e^{CJ_\mp}(2CJ_3 \mp C^2 J_\mp), \tag{B2}$$

which immediately follow form the commutation relations (3.2). We may calculate

$$\begin{aligned}
\mathscr{J}_3^{(\nu)}\gamma^\nu(u_0^{(\nu)}, e^{\xi J_-^{(\nu)}}\cdots)^{(\nu)} &= \gamma^\nu\left(u_0^{(\nu)}, \left(\frac{\nu}{2} + \xi\frac{\partial}{\partial\xi}\right)e^{\xi J_-^{(\nu)}}\cdots\right)^{(\nu)} \\
&= \gamma^\nu\left(u_0^{(\nu)}, \left(\frac{\nu}{2} + \xi J_-^{(\nu)}\right)e^{\xi J_-^{(\nu)}}\cdots\right)^{(\nu)} \\
&= \gamma^\nu(u_0^{(\nu)}, (J_3^{(\nu)} + \xi J_-^{(\nu)})e^{\xi J_-^{(\nu)}}\cdots)^{(\nu)} \\
&= \gamma^\nu(u_0^{(\nu)}, e^{\xi J_-^{(\nu)}} J_3^{(\nu)}\cdots)^{(\nu)}.
\end{aligned} \tag{B3}$$

The relations for $\mathscr{J}_+$ and $\mathscr{J}_-$ can be derived in a similar way. Let us also note that in terms of the operator $D(\xi)$ defined by (8.10) we can rewrite (8.4) as

$$\mathscr{J}_a^{(\nu)}(u_0^{(\nu)}, D(\xi)\cdots)^{(\nu)} = (u_0^{(\nu)}, D(\xi)J_a^{(\nu)}\cdots)^{(\nu)}. \tag{B4}$$

APPENDIX C: DIFFERENT DERIVATION OF (9.4)

Here we derive (9.4) by direct calculation without using diagrams. We have

$$\log\overline{\tau} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}[(u_0^{(\epsilon)}, DUD'u_0^{(\epsilon)})^{(\epsilon)} - 1], \quad \overline{R_r} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}(u_0^{(\epsilon)}, J_- DUD' u_0^{(\epsilon)})^{(\epsilon)}, \tag{C1}$$

where $D \equiv D(\xi)$ and $D' \equiv D(-\xi')$ with the definition (8.10). By using (B1), (B2), (8.3), and (8.15) we can show that

$$DJ_1 = J_1 D. \tag{C2}$$

From (B4), (C1), and (C2) we can derive $\overline{R_r} = 2\mathscr{J}_1 \log \overline{\tau}$.

APPENDIX D: DERIVATION OF (10.3)

Setting $\nu = -1$ in (10.1) we obtain

$$U^{(-1)}_{\theta\theta'} = \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix} \begin{pmatrix} \alpha^+ & \beta^- \\ \beta^+ & \alpha^- \end{pmatrix} \begin{pmatrix} \cos(\theta'/2) & -\sin(\theta'/2) \\ \sin(\theta'/2) & \cos(\theta'/2) \end{pmatrix} \equiv \begin{pmatrix} \alpha^+_{\theta\theta'} & \beta^-_{\theta\theta'} \\ \beta^+_{\theta\theta'} & \alpha^-_{\theta\theta'} \end{pmatrix}, \tag{D1}$$

where $\alpha^\pm$ and $\beta^\pm$ are the matrix elements of $U^{(-1)}$ defined by (3.6). By comparing (D1) with the last expression of (10.1) we have

$$\tau_{\theta\theta'} = \frac{1}{\alpha^+_{\theta\theta'}}, \quad R_{r,\theta\theta'} = \frac{\beta^+_{\theta\theta'}}{\alpha^+_{\theta\theta'}}, \quad R_{l,\theta\theta'} = -\frac{\beta^-_{\theta\theta'}}{\alpha^+_{\theta\theta'}}. \tag{D2}$$

Now let us consider the Fokker–Planck equation with the potential $\overline{V}$ given by Eq. (2.3). We define $\theta_a \equiv \arctan(-ic_a/k)$ and $\theta_b \equiv \arctan(-ic_b/k)$. As (3.7), two independent solutions of (1.11) are given by

$$\phi^+_{0,\theta_0}(x,x_0) \equiv \alpha^+_{0,\theta_0}(x,x_0) + \beta^+_{0,\theta_0}(x,x_0), \quad \phi^-_{0,\theta_0}(x,x_0) \equiv \alpha^-_{0,\theta_0}(x,x_0) + \beta^-_{0,\theta_0}(x,x_0), \tag{D3}$$

where x_0 and θ_0 are arbitrary. It is easy to see that

$$U^{(-1)}_{\theta_b\theta_b}(x,x_b) = \begin{pmatrix} e^{-ip_b(x-x_b)} & 0 \\ 0 & e^{ip_b(x-x_b)} \end{pmatrix} \quad (x > x_b),$$

$$U^{(-1)}_{\theta_a\theta_a}(x,x_a) = \begin{pmatrix} e^{-ip_a(x-x_a)} & 0 \\ 0 & e^{ip_a(x-x_a)} \end{pmatrix} \quad (x < x_a), \tag{D4}$$

where p_a and p_b are defined by (2.5). From (D1) and the multiplication property of the evolution operator it follows that

$$\begin{aligned} U^{(-1)}_{0,\theta_a}(x,x_a) &= \begin{pmatrix} \cos(\theta_a/2) & -\sin(\theta_a/2) \\ \sin(\theta_a/2) & \cos(\theta_a/2) \end{pmatrix} U^{(-1)}_{\theta_a\theta_a}(x,x_a) \\ &= \begin{pmatrix} \cos(\theta_b/2) & -\sin(\theta_b/2) \\ \sin(\theta_b/2) & \cos(\theta_b/2) \end{pmatrix} U^{(-1)}_{\theta_b\theta_b}(x,x_b) \begin{pmatrix} \alpha^+_{\theta_b\theta_a}(x_b,x_a) & \beta^-_{\theta_b\theta_a}(x_b,x_a) \\ \beta^+_{\theta_b\theta_a}(x_b,x_a) & \alpha^-_{\theta_b\theta_a}(x_b,x_a) \end{pmatrix}. \end{aligned} \tag{D5}$$

Using (D4) and (D5) we find that $\phi^+_{0,\theta_a}(x,x_a)$ behaves as

$$\begin{aligned} \phi^+_{0,\theta_a}(x,x_a) &= \left(\cos\frac{\theta_a}{2} + \sin\frac{\theta_a}{2}\right) e^{-ip_a(x-x_a)} \qquad (x < x_a) \\ &= \left(\cos\frac{\theta_b}{2} + \sin\frac{\theta_b}{2}\right) \alpha^+_{\theta_b\theta_a}(x_b,x_a) e^{-ip_b(x-x_b)} \\ &\quad + \left(\cos\frac{\theta_b}{2} - \sin\frac{\theta_b}{2}\right) \beta^+_{\theta_b\theta_a}(x_b,x_a) e^{ip_b(x-x_b)} \qquad (x > x_b). \end{aligned} \tag{D6}$$

Since $p_a/p_b=(\cos^2(\theta_b/2)-\sin^2(\theta_b/2))/(\cos^2(\theta_a/2)-\sin^2(\theta_a/2))$, from (D6), (D2) and (2.4) we obtain (10.3). (The relation for $\widetilde{R}_l$ can be derived in the same way.)

[1] E. Nelson, Phys. Rev. **150**, 1079 (1966); G. Parisi and Y. Wu, Sci. Sin. **24**, 483 (1981).
[2] H. Risken, *The Fokker-Planck Equation* (Springer, Berlin, 1984).
[3] T. Miyazawa and T. Izuyama, Phys. Rev. A **36**, 5791 (1987).
[4] T. Miyazawa, Phys. Rev. A **39**, 1447 (1989).
[5] T. Miyazawa, J. Math. Phys. **34**, 1587 (1993).
[6] T. Miyazawa, J. Phys. A **25**, 2359 (1992).
[7] T. Miyazawa, J. Math. Phys. **36**, 5643 (1995).
[8] S Novikov, S. V. Manakov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons* (Consultants Bureau, New York, 1984); W. Eckhaus and A. van Harten, *The Inverse Scattering Transformation and the Theory of Solitons* (North-Holland, Amsterdam, 1981).
[9] M. Wadati, J. Phys. Soc. Jpn. **34**, 1289 (1973);
[10] R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966); K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer, New York, 1989).
[11] See Section XII of Ref. 4 for a method of calculation.
[12] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[13] In Ref. 7, representations are labeled by the spin index j.
[14] By using the formula (8.3) shown later and (B1), (B2) in Appendix B, it can be directly checked that the expressions in (6.5) satisfy (6.4).
[15] See, for example, L. D. Faddeev, in *Methods in Field Theory* (Les Houches, Session XXVIII, 1975), edited by R. Balian and J. Zinn-Justin (North-Holland, Amsterdam, 1976), p. 1 and references therein.
[16] J. L. Cardy, in *Phase Transitions and Critical Phenomena*, Vol. 11, edited by C. Domb and J. L. Lebowits (Academic, London, 1987), p. 55; C. Gómez, M. Ruiz-Altaba, and G. Sierra, *Quantum Groups in Two-dimensional Physics* (Cambridge University Press, Cambridge, 1996).
[17] See Appendix B of Ref. 7.

A reformulation of simple liquids theory—Renormalization by one-, two-, and three-particle densities

Ko Okumura
Theoretical Studies, Institute for Molecular Science, Myodaiji, Okazaki, Aichi 444-8585, Japan

(Received 27 May 1997; accepted for publication 6 January 1998)

We reformulate the theory of simple liquids in a field theoretical way by taking into account the triplet potential w_{ijk}, in addition to the external potential u_i and the pair potential v_{ij}. The innovation here is *the inversion method* and *the on-shell expansion* which are the building blocks of a novel use of Legendre transformation developed in field theory. By the inversion method, we renormalize the theory in terms of one-, two-, and three-particle densities, and present a diagrammatic representation for a thermodynamical functional, which is the entropy except for a trivial constant, in terms of renormalized variables. In other words, we present an expression for the entropy in terms of only one-, two-, and three-particle densities: the n-particle density where $n \geqslant 4$ does not appear in the expression. The on-shell condition, which is a starting point of the on-shell expansion, of the thermodynamical functional thus obtained (the entropy) leads to a set of three self-consistent equations for one-, two-, and three-particle densities. Through one of the self-consistent equations, we can systematically improve the Kirkwood's superposition approximation for the three-particle density. The on-shell conditions for other thermodynamical functionals, also obtained in this article, are found to be extentions of various well-known equations in the theory of simple liquids. The formulation presented here is complementary to the conventional re-summation techniques for renormalization of diagrams. In the present formulation, we do not have to care about the topological structure of diagrams, often characterized by the irreducibility of diagrams. Instead, by a perturbative calculation, we can automatically single out the diagrams with the topological structure predicted by the re-summation techniques. © *1998 American Institute of Physics.* [S0022-2488(98)01604-1]

I. INTRODUCTION

In the theory of simple liquids, renormalization of the theory in terms of one-, and two-particle densities is essential. (The n-particle density here is sometimes called the correlation function of the n-th order.) To perform such renormalization, re-summation techniques of diagrams had been extensively developed.[1–4] Such techniques, though elegant and attractive, are sometimes complex and difficult to understand. In such techniques, the topological structure of the diagrams for the renormalized theory is determined by using graphical terms such as irreducibility of diagrams. If one wants to obtain higher order diagrams, we have to single out, by hand, diagrams which have the topological structure thus determined, after considering all possible diagrams of that order. At a higher order, this task can be laborious.

In the present reformulation, however, we can obtain the renormalized diagrams automatically order by order through simple (but sometimes tedious) calculations. We do not have to care about the topological structure at all; the diagrams with the demanded topology are automatically singled out. In this sense, this reformulation is complementary to the conventional re-summation techniques.

The innovation here is a novel use of the Legendre transformation which has been developed in field theory and has been applied to quantum systems.[5] In this article we present the first application to a classical system. This methodology consists of the inversion method and the one-shell expansion, which shall be reviewed briefly below. However, in short, the former is a method by which we can obtain a non-perturbative result through a sort of perturbative calculation, and the latter is a formalism in which we can systematically extract all physical information

 © 1998 American Institute of Physics

concerning a physical variable of interest, starting from an effective action. (The effective action is a kind of thermodynamical functional in a classical system as we see below.)

The systematic expansion of the entropy of simple liquids theory in terms of the correlation function has received considerable attention. Such an expansion for the canonical ensemble was studied in Ref. 6, which followed by the works for the grand canonical ensemble.[7,8] Recently, these expansion schemes have been revisited[9–12] and numerically examined[13,14] by several researchers. In all these works, the entropy is expanded in terms of one-particle density and the n-th order term contains the correlation function up to the n-th order: the entropy is expressed by the n-particle densities where $n=0,1,...,\infty$.

However, as shown below, the entropy functional, which is the effective action except for a trivial term, can be described by the correlation functions of the order *up to the n-th* if we neglect higher-body potential than n-body. Indeed, the entropy given in Ref. 1 is described only by one- and two-particle densities since, as in conventional simple liquids theory, the triplet and higher-body potential are neglected there.

In this article we present an expression for the entropy as functional of one-, two-, and three-particle densities of the grand canonical ensemble. The expansion does not include correlation functions of the order higher than three and thus the theory is renormalized in terms of one, two-, and three-particle densities.

To renormalize the theory in terms of the three-particle density, in the present paper, we first explicitly include a triplet potential into a theory, and then, make a Legendre transformation from the triplet potential to the three particle density. The triplet potential can be set to zero after the renormalization if we like to go back to the (conventional) liquids theory without the triplet potential. Note here that, by first introducing the triplet potential and then by setting the potential at the end, we can extract new information, which can not be obtained if we neglect the potential from the start.

Extention of the liquid theory to include triplet potentials is not new and had been studied, though the above viewpoint for the triplet potential as a tool for the Legendre transformation might have not. For example, for a weak triplet potential, they included the effects of the triplet potential by modifying the pair potential with the aid of diagrams.[15–17] Other important references can be found in Refs. 12, 18.

Furthermore, inclusion of the general m-body potential ($m=1,2,3,4,...$) into the liquid theory is performed in Ref. 19 in a simple and clear way. There, various diagrammatic rules are neatly given and are renormalized up to the two-particle densities. In the present work, we further renormalize the theory by the three-particle densities though our case is limited to the special case where there are no more-than-three-body potentials. (See Sec. VII.)

In the present formalism, the stationary conditions for the entropy or the effective action come out as identities of the Legendre transformation. The set of conditions can be seen as a set of self-consistent equations to be solved for the variables of interest.

In this article, we obtain three kinds of the effective actions, which are closely related to the entropy. As stationary conditions for these actions, we get three sets of self-consistent equations, some of which can be seen as a generalization of well-known equations in simple liquids theory.

Especially, one of the self-consistent equations leads to successive improvement of the Kirkwood's superposition approximation[4] for the three-particle densities. The zero-th order approximation of this theory coincides with the previously known approximation (for example, see Refs. 20, 12), which reduces to the Kirkwood's approximation for the system without the triplet potential. The first order approximation of ours can be seen as an improvement over the well-known approximations.[21–24]

Let us review the inversion method and the on-shell expansion briefly. Consider a physical variable of interest ϕ in a system described by a Hamiltonian H. In a classical system at an equilibrium, a physical variable is a thermal expectation value of some quantity $\hat{\phi}$. (In a quantum system a physical variable is a quantum expectation value of an operator $\hat{\phi}$.) In the inversion method, we modify the Hamiltonian H to $H+J\hat{\phi}$ where J is an artificial external source which should be set to zero at the end of calculation. However, if the original Hamiltonian H already has a term proportional to $\hat{\phi}$, say $j\hat{\phi}$, we merely replace j with an artificial source J and set it back

to j at the end of calculation. After this modification of the Hamiltonian, we calculate ϕ by a perturbative expansion in terms of the expansion parameter, say g, in the presence of the artificial source J. We have an expansion called *the original series* of the form,

$$\phi=\phi(J)=\phi^{(0)}(J)+\phi^{(1)}(J)+\phi^{(2)}(J)+\cdots, \tag{1.1}$$

where $\phi^{(i)}$ is the i-th order in g when we regard J as independent of g. Now we invert Eq. (1.1) in favor of J to get *the inversion series*,

$$J=J(\phi)=J^{(0)}(\phi)+J^{(1)}(\phi)+J^{(2)}(\phi)+\cdots, \tag{1.2}$$

where $J^{(i)}(\phi)$ is the i-th order in g when we regard ϕ as independent of g. In order to obtain an explicit form of $J^{(i)}(\phi)$, we insert Eq. (1.2) into Eq. (1.1):

$$\phi=\phi^{(0)}(J^{(0)}+J^{(1)}+J^{(2)}+\cdots)+\phi^{(1)}(J^{(0)}+J^{(1)}+\cdots)+\phi^{(2)}(J^{(0)}+\cdots)+\cdots, \tag{1.3}$$

or

$$\begin{aligned}\phi=&\phi^{(0)}(J^{(0)})+\phi^{(0)\prime}(J^{(0)})J^{(1)}+\phi^{(1)}(J^{(0)})+\phi^{(0)\prime}(J^{(0)})J^{(2)}+\tfrac{1}{2}\phi^{(0)\prime\prime}(J^{(0)})\cdot(J^{(1)})^2\\&+\phi^{(1)\prime}(J^{(0)})J^{(1)}+\phi^{(2)}(J^{(0)})+\cdots.\end{aligned} \tag{1.4}$$

The trick of the inversion method is to regard ϕ in the left-hand side of this equation as independent of g. Then we get a set of equations called *inversion formulae*:

$$\phi=\phi^{(0)}(J^{(0)}), \tag{1.5}$$

$$\phi^{(0)\prime}(J^{(0)})J^{(1)}+\phi^{(1)}(J^{(0)})=0, \tag{1.6}$$

$$\phi^{(0)\prime}(J^{(0)})J^{(2)}+\frac{1}{2}\,\phi^{(0)\prime\prime}(J^{(0)})\cdot(J^{(1)})^2+\phi^{(1)\prime}(J^{(0)})J^{(1)}+\phi^{(2)}(J^{(0)})=0, \tag{1.7}$$

$$\begin{aligned}&\phi^{(0)\prime}(J^{(0)})J^{(3)}+\phi^{(0)\prime\prime}(J^{(0)})J^{(1)}J^{(2)}+\frac{1}{3!}\,\phi^{(0)\prime\prime\prime}(J^{(0)})\cdot(J^{(1)})^3+\phi^{(1)\prime}(J^{(0)})J^{(2)}\\&+\frac{1}{2}\,\phi^{(1)\prime\prime}(J^{(0)})\cdot(J^{(1)})^2+\phi^{(2)\prime}(J^{(0)})J^{(1)}+\phi^{(3)}(J^{(0)})=0,\end{aligned} \tag{1.8}$$

and so on. In this way we can obtain $J^{(i)}(\phi)$ up to the desired order. Note here that all the derivatives of $\phi^{(i)}$ and $\phi^{(i)}$ itself in the above formulae are evaluated at $J=J^{(0)}$, which is determined by the lowest inversion formula (1.5). Finally, we set the artificially introduced source J back to the original value, say j, to go back to the original theory:

$$J^{(0)}(\phi)+J^{(1)}(\phi)+\cdots=j. \tag{1.9}$$

If we truncate the original series (1.1) at the i-th order, we obtain Eq. (1.9) truncated at the i-th order. However, these two truncated equations are not equivalent and the latter, which can be regarded as an equation determining ϕ, is a more improved equation for ϕ than the former.

The trick of regarding ϕ as independent of g corresponds to making the Legendre transformation; ϕ and g are dealt with as mutually independent in the Legendre transformation. To further clarify this point, we introduce a generating function W which satisfies the relation

$$\phi(J)=\frac{dW(J)}{dJ}. \tag{1.10}$$

In the case of an equilibrium system W is nothing but the thermodynamical potential [$W=\Omega=-(1/\beta)\ln\Xi$: see Eq. (2.8)]. Now we introduce *the effective action* Γ by the Legendre transfor-

mation: $\Gamma = W(J,g) - J\phi(J,g)$ where $\phi(J,g) = \partial W(J,g)/\partial J$. Here, we have emphasized the g dependence. The variation of Γ, *if we assume that J and g are independent variables*, is expressed as

$$d\Gamma = \frac{\partial W(J,g)}{\partial J}\, dJ + \frac{\partial W(J,g)}{\partial g}\, dg - dJ\phi(J,g) - Jd\phi(J,g) = \frac{\partial W(J,g)}{\partial g}\, dg - Jd\phi(J,g). \tag{1.11}$$

Hence, as is well known, we see that the quantity Γ can be regarded as a function of two independent variables ϕ and g—$\Gamma = \Gamma(\phi,g)$. Equation (1.11) seems trivial but is essential in the present discussion. What is implied in Eq. (1.11) is as follows: if we solve (or invert) the relation $\phi(J,g) = \partial W(J,g)/\partial J$ in favor of J, *assuming that two quantities ϕ and g are mutually independent*, to obtain $J = J(\phi,g)$ and then insert this expression of J into all J appearing in $W(J,g) - J\phi$, this quantity $W(J,g) - J\phi$ is automatically written by only two independent variables ϕ and g. In other words, this inversion process of the Legendre transformation is carried out regarding ϕ as independent of g. This is the justification of the trick of the inversion method. Note that once the inversion or the Legendre transformation is performed and after the sources are set to an appropriate value, j for example, the resultant ϕ depends on g. In many quantum cases such as Quantum Electrodynamics (QED),[25] Superconductivity,[26] Hubbard model,[27] and Superfluidity,[5] it is known that this trick of the inversion method reproduces the mean field approximation at the first order (in which we truncate the series at $i = 1$). In this sense, the inversion method offers a way to systematically advance the mean field approximation.

Reciprocal to Eq. (1.10), there is an identity,

$$J = -\frac{d\Gamma(\phi)}{d\phi}. \tag{1.12}$$

This equation is equivalent to the inversion series (1.2) and, at an appropriate value of J, say j, corresponds to Eq. (1.9). In other words, Eq. (1.12) at $J = j$ is an equation which determines ϕ. This equation corresponds to the on-shell condition in field theory. The on-shell expansion is an expansion of the right-hand side of the on-shell condition $j = -d\Gamma(\phi)/d\phi$ at the point $\phi = \phi_0 + \Delta\phi$ around the solution of the on-shell condition $\phi = \phi_0$ [i.e., $j = -d\Gamma(\phi_0)/d\phi_0$]. Through the on-shell expansion, we obtain a set of equations containing all the physical information related to ϕ such as excitations and scatterings. However, since we concentrate on equilibrium properties of liquids in this article, the higher order on-shell equations shall not be examined. Such an important issue is reserved for a future work. In this sense, a real application of the on-shell expansion to a classical system is relegated to a separate work, which shall deal with dynamical properties of simple liquids.

In the above we consider a simple case where ϕ is scalar and position independent. However, the generalization to a case where ϕ is a tensor and position dependent is straightforward [see, for example, Eq. (4.10) below].

Organization of this article is as follows. In Section II, we reformulate the simple liquids theory in a field theoretical way and successively introduce three kinds of effective actions $(\Gamma_1, \Gamma_2, \Gamma_3)$ through the first, the second, and the third Legendre transformation. We then give the on-shell condition for each effective action. It is stressed here that the effective action Γ_3 is the entropy except for a trivial term and it can be written as a functional of the correlation functions of *up to the third order*. In Section III, we present a diagrammatic expression for the thermodynamic potential with explicitly including the triplet potential w_{ijk}. The introduction of *pseudo 2-bonds* is essential for the presentation of the diagrams with the triplet potential. Sections IV to VI deal with renormalization of the theory in terms of one-, two-, and three-particle densities, respectively. After renormalization, we obtain a diagrammatic expression for Γ_1, Γ_2, or Γ_3 (the entropy) and show that the on-shell condition for each Γ_i $(i = 1,2,3)$ can be considered as some extension of various well-known equations in simple liquids theory. In Section VI, we obtain a systematic way of improving the Kirkwood's superposition approximation for the three-particle densities.

II. A REFORMULATION OF THE SIMPLE LIQUIDS THEORY

Consider an N-particle classical system subject to the potential $u(\mathbf{r})$, the pair potential $v(\mathbf{r},\mathbf{r}')$, and the triplet potential $w(\mathbf{r},\mathbf{r}',\mathbf{r}'')$,[28]

$$\mathscr{H}_N \equiv \mathscr{H}_N(\mathbf{p}_1,\ldots,\mathbf{p}_N,\mathbf{q}_1,\ldots,\mathbf{q}_N) = \frac{1}{2m}\sum_{i=1}^{N}|\mathbf{p}_i|^2 + V_N, \tag{2.1}$$

$$V_N \equiv V_N(\mathbf{q}_1,\ldots,\mathbf{q}_N) = \sum_{i=1}^{N} u(\mathbf{q}_i) + \sum_{i<j}^{N} v(\mathbf{q}_i,\mathbf{q}_j) + \sum_{i<j<k}^{N} w(\mathbf{q}_i,\mathbf{q}_j,\mathbf{q}_k). \tag{2.2}$$

Here the pair potential $v(\mathbf{r}_i,\mathbf{r}_j)$ and $w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)$ are symmetric under the interchange of position variables [$v(\mathbf{r}_1,\mathbf{r}_2)=v(\mathbf{r}_2,\mathbf{r}_1)$ etc.]. By introducing the n-particle density for an N-particle system $(n=1,2,3)$ by

$$\rho_N^{(1)}(\mathbf{r}) = \sum_{i=1}^{N}\delta(\mathbf{r}-\mathbf{q}_i), \tag{2.3}$$

$$\rho_N^{(2)}(\mathbf{r},\mathbf{r}') = \sum_{i\neq j}^{N}\delta(\mathbf{r}-\mathbf{q}_i)\,\delta(\mathbf{r}'-\mathbf{q}_j) = \rho_N^{(1)}(\mathbf{r})\rho_N^{(1)}(\mathbf{r}') - \delta(\mathbf{r}-\mathbf{r}')\rho_N^{(1)}(\mathbf{r}), \tag{2.4}$$

$$\rho_N^{(3)}(\mathbf{r},\mathbf{r}',\mathbf{r}'') = \sum_{i\neq j\neq k}^{N}\delta(\mathbf{r}-\mathbf{q}_i)\,\delta(\mathbf{r}'-\mathbf{q}_j)\,\delta(\mathbf{r}''-\mathbf{q}_k), \tag{2.5}$$

the grand potential $V_N-\mu N$ is cast into the form

$$V_N-\mu N = \int d\mathbf{r}_1\,\tilde{u}(\mathbf{r}_1)\rho_N^{(1)}(\mathbf{r}_1) + \frac{1}{2}\int d\mathbf{r}_1\int d\mathbf{r}_2\, v(\mathbf{r}_1,\mathbf{r}_2)\rho_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2) + \frac{1}{3!}\int d\mathbf{r}_1\int d\mathbf{r}_2\int d\mathbf{r}_3\, w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)\rho_N^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3), \tag{2.6}$$

where

$$\tilde{u}(\mathbf{r}) = -\mu + u(\mathbf{r}). \tag{2.7}$$

This expression for the potential plays a key role in our reformulation of the simple liquids theory in a field theoretical way. From this expression, we can identify $(J,\hat{\phi})$ of the inversion method with $(\tilde{u},\rho_N^{(1)})$, $(v,\rho_N^{(2)})$, or $(w,\rho_N^{(3)})$.

The grand partition function Ξ and the thermodynamic potential,

$$\Omega = -\frac{1}{\beta}\ln\Xi, \tag{2.8}$$

is defined by

$$e^{-\beta\Omega} = \Xi = \sum_{N=0}^{\infty}\frac{1}{N!}\frac{1}{h^{3N}}\int d\mathbf{p}^N\int d\mathbf{q}^N e^{-\beta(\mathscr{H}_N-\mu N)}, \tag{2.9}$$

where $\int d\mathbf{p}^N$ and $\int d\mathbf{q}^N$ implies the integrations over $\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_N$ and $\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_N$, respectively.

Regarding quantities Ω and Ξ as functionals of the potentials $\tilde{u}(\mathbf{r})$, $v(\mathbf{r},\mathbf{r}')$, and $w(\mathbf{r},\mathbf{r}',\mathbf{r}'')$, we have the n-particle density $(n=1,2,3)$ by differentiating Ω with respect to $\tilde{u}$, v, or w:[29]

$$\rho^{(1)}(\mathbf{r}) \equiv \frac{d\Omega}{\delta\tilde{u}(\mathbf{r})} = \langle\rho_N^{(1)}(\mathbf{r})\rangle, \tag{2.10}$$

$$\rho^{(2)}(\mathbf{r},\mathbf{r}')\equiv\frac{\delta\Omega}{\delta\,\dfrac{1}{2}\,v(\mathbf{r},\mathbf{r}')}=\langle\rho_N^{(2)}(\mathbf{r},\mathbf{r}')\rangle, \tag{2.11}$$

$$\rho^{(3)}(\mathbf{r},\mathbf{r}',\mathbf{r}'')\equiv\frac{\delta\Omega}{\delta\,\dfrac{1}{3!}\,w(\mathbf{r},\mathbf{r}',\mathbf{r}'')}=\langle\rho_N^{(3)}(\mathbf{r},\mathbf{r}',\mathbf{r}'')\rangle. \tag{2.12}$$

Here, the expectation implies[30]

$$\langle\mathscr{O}_N\rangle=\frac{1}{\Xi}\sum_{N=0}^{\infty}\frac{1}{N!}\frac{1}{h^{3N}}\int d\mathbf{p}^N\int d\mathbf{q}^N\mathscr{O}_N\,e^{-\beta(\mathscr{H}_N-\mu N)}, \tag{2.13}$$

and we have used the relations

$$\frac{\delta\tilde{u}(\mathbf{r})}{\delta\tilde{u}(\mathbf{r}')}=\delta(\mathbf{r}-\mathbf{r}'), \tag{2.14}$$

$$\frac{\delta v(\mathbf{r}_1,\mathbf{r}_2)}{\delta v(\mathbf{r}_{1'},\mathbf{r}_{2'})}=\frac{1}{2!}\left[\delta(\mathbf{r}_1-\mathbf{r}_{1'})\delta(\mathbf{r}_2-\mathbf{r}_{2'})+\delta(\mathbf{r}_1-\mathbf{r}_{2'})\delta(\mathbf{r}_2-\mathbf{r}_{1'})\right], \tag{2.15}$$

$$\frac{\delta w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)}{\delta w(\mathbf{r}_{1'},\mathbf{r}_{2'},\mathbf{r}_{3'})}=\frac{1}{3!}\left[\delta(\mathbf{r}_1-\mathbf{r}_{1'})\delta(\mathbf{r}_2-\mathbf{r}_{2'})\delta(\mathbf{r}_3-\mathbf{r}_{3'})+(3!-1)\text{terms}\right]. \tag{2.16}$$

In order to rewrite the theory in terms of $\rho^{(1)}(\mathbf{r})$, $\rho^{(2)}(\mathbf{r},\mathbf{r}')$, and $\rho^{(3)}(\mathbf{r},\mathbf{r}',\mathbf{r}'')$ instead of $\tilde{u}(\mathbf{r})$, $v(\mathbf{r},\mathbf{r}')$, and $w(\mathbf{r},\mathbf{r}',\mathbf{r}'')$, the Legendre transformations are successively performed. The first, the second, and the third Legendre transformations are, respectively, defined by

$$\Gamma_1[\rho^{(1)},v]\equiv\Omega[\tilde{u},v,w]-\int d\mathbf{r}_1\tilde{u}(\mathbf{r}_1)\rho^{(1)}(\mathbf{r}_1), \tag{2.17}$$

$$\Gamma_2[\rho^{(1)},\rho^{(2)}]\equiv\Gamma_1[\rho^{(1)},v]-\frac{1}{2}\int d\mathbf{r}_1\int d\mathbf{r}_2 v(\mathbf{r}_1,\mathbf{r}_2)\rho^{(2)}(\mathbf{r}_1,\mathbf{r}_2), \tag{2.18}$$

$$\Gamma_3[\rho^{(1)},\rho^{(2)},\rho^{(3)}]\equiv\Gamma_2[\rho^{(1)},\rho^{(2)},w]-\frac{1}{3!}\int d\mathbf{r}_1\int d\mathbf{r}_2\int d\mathbf{r}_3 w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)\rho^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3). \tag{2.19}$$

In these equations, $(\tilde{u},\rho^{(1)})$, $(v,\rho^{(2)})$, or $(w,\rho^{(3)})$ can be regarded as (J,ϕ) of the inversion method. In the present case, the original Hamiltonian already has a term proportional to ϕ: we should set $J=-\mu+u$, v, or w (instead of $J=0$) at the end. (In the conventional simple liquids theory where the triplet potential is neglected, we have to set $J=w=0$ at the end.) By using the relation

$$\begin{aligned}\delta\Omega[\tilde{u},v,w]=&\int d\mathbf{r}_1\,\frac{\delta\Omega}{\delta\tilde{u}(\mathbf{r}_1)}\,\delta\tilde{u}(\mathbf{r}_1)+\int d\mathbf{r}_1\int d\mathbf{r}_2\,\frac{\delta\Omega}{\delta v(\mathbf{r}_1,\mathbf{r}_2)}\,\delta v(\mathbf{r}_1,\mathbf{r}_2)\\&+\int d\mathbf{r}_1\int d\mathbf{r}_2\int d\mathbf{r}_3\,\frac{\delta\Omega}{\delta w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)}\,\delta w(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3),\end{aligned} \tag{2.20}$$

we can explicitly confirm that Γ_1, Γ_2, and Γ_3 are functionals of $(\rho^{(1)},v,w)$, $(\rho^{(1)},\rho^{(2)},w)$, and $(\rho^{(1)},\rho^{(2)},\rho^{(3)})$, respectively, with the identities

$$\frac{\delta\Gamma_1[\rho^{(1)},v,w]}{\delta\rho_1^{(1)}}=-\widetilde{u}_1,\quad \frac{\delta\Gamma_1[\rho^{(1)},v,w]}{\delta\,\frac{1}{2}\,v_{12}}=\rho_{12}^{(2)},\quad \frac{\delta\Gamma_1[\rho^{(1)},v,w]}{\delta\,\frac{1}{3!}\,w_{123}}=\rho_{123}^{(3)}, \tag{2.21}$$

$$\frac{\delta\Gamma_2[\rho^{(1)},\rho^{(2)},w]}{\delta\rho_1^{(1)}}=-\widetilde{u}_1,\quad \frac{\delta\Gamma_2[\rho^{(1)},\rho^{(2)},w]}{\delta\,\frac{1}{2}\,\rho_{12}^{(2)}}=-v_{12},\quad \frac{\delta\Gamma_2[\rho^{(1)},\rho^{(2)},w]}{\delta\,\frac{1}{3!}\,w_{123}}=\rho_{123}^{(3)}, \tag{2.22}$$

$$\frac{\delta\Gamma_3[\rho^{(1)},\rho^{(2)},\rho^{(3)}]}{\delta\rho_1^{(1)}}=-\widetilde{u}_1,\quad \frac{\delta\Gamma_3[\rho^{(1)},\rho^{(2)},\rho^{(3)}]}{\delta\,\frac{1}{2}\,\rho_{12}^{(2)}}=-v_{12},\quad \frac{\delta\Gamma_3[\rho^{(1)},\rho^{(2)},\rho^{(3)}]}{\delta\,\frac{1}{3!}\,\rho_{123}^{(3)}}=-w_{123}. \tag{2.23}$$

Here, we have used simple and self-evident notations. These can be seen as the on-shell conditions for the first, second, and third Legendre transformations, respectively, and shall be shown to correspond to familiar equations of simple liquids theory.

The effective actions introduced above are some thermodynamic functionals and closely related to the entropy. We notice that the entropy functional is given by

$$S/k=-\beta\Gamma_3+3\langle N\rangle/2, \tag{2.24}$$

where k is the Boltzmann constant, because the definition of the entropy S is given by

$$\Omega=\langle E\rangle-TS-\mu\langle N\rangle, \tag{2.25}$$

where T is the temperature, and because there is the identity

$$\langle E-\mu N\rangle=\frac{3}{2}\,\langle N\rangle kT+\int d1\widetilde{u}_1\rho_1^{(1)}+\int d1\int d2v_{12}\rho_{12}^{(2)}+\int d1\int d2\int d3w_{123}\rho_{123}^{(3)}. \tag{2.26}$$

It also follows that, if we neglect the three-body potential, the entropy is expressed as

$$S/k=-\beta\Gamma_2+3\langle N\rangle/2\quad (w=0), \tag{2.27}$$

and, if we further neglect the two-body potential, it is given by

$$S/k=-\beta\Gamma_1+3\langle N\rangle/2\quad (v=w=0). \tag{2.28}$$

We stress here that the stationary condition for the entropy, which leads to the self-consistent equations, automatically comes out as identities of the Legendre transformation. For example, the identities in (2.23) are equivalent to

$$\frac{\partial S/k}{\partial\rho_1^{(1)}}=\beta\widetilde{u}_1+\frac{3}{2},\quad \frac{\partial S/k}{\partial\rho_{12}^{(2)}}=\beta v_{12},\quad \frac{\partial S/k}{\partial\rho_{123}^{(3)}}=\beta w_{123}. \tag{2.29}$$

From the above argument it may be clear that, in general, the entropy of simple liquids theory, where potentials of higher than m-body are neglected, can be written as a functional of the correlation functions of the order up to the m-th. In such a case, higher order correlation functions can be expressed by the correlation functions of up to the m-th order.

By using the results given in this section, we can establish the general statement that n-particle density can be given as an explicit functional of s-particle densities where $s\leqslant n-1$. From the second equality in Eq. (2.21) we readily notice that $\rho^{(2)}$ is given as a functional of $\rho^{(1)}$, v, and w. From the last equality in Eq. (2.22), we notice that $\rho^{(3)}$ is given by $\rho^{(2)}$, $\rho^{(1)}$, and w. In the same way, if we introduce Γ_n by taking into general m-body potentials, we can confirm that Γ_n is a functional of $\rho^{(s)}$ $(s=1,2,...,n)$ as a consequence of the basic property of the Legendre transfor-

mation. Since $\rho^{(n)}$ is given by the derivative of Γ_{n-1} as in Eqs. (2.21) and (2.22), we arrive at the general statement. This statement is not limited to the classical case; the parallel argument in relativistic quantum field theory is developed in Ref. 31.

III. DIAGRAMMATIC EXPRESSION FOR THE THERMODYNAMIC POTENTIAL

The thermodynamic potential can be expressed as

$$e^{-\beta\Omega}=\Xi=\sum_{N=0}^{\infty}\frac{1}{N!}\int d[N]\left[\prod_i z_i^*\right]\cdot\left[\prod_{i<j}(1+f_{ij})\right]\cdot\left[\prod_{i<j<k}(1+t_{ijk})\right], \tag{3.1}$$

where

$$\int d[N]\equiv\int d\mathbf{q}_1\int d\mathbf{q}_2\cdots\int d\mathbf{q}_N, \tag{3.2}$$

$$z_i^*\equiv z^*(\mathbf{q}_i)=ze^{-\beta u(\mathbf{q}_i)}, \tag{3.3}$$

$$f_{ij}\equiv f(\mathbf{q}_i,\mathbf{q}_j)=e^{-\beta v(\mathbf{q}_i,\mathbf{q}_j)}-1, \tag{3.4}$$

and

$$t_{ijk}\equiv t(\mathbf{q}_i,\mathbf{q}_j,\mathbf{q}_k)=e^{-\beta w(\mathbf{q}_i,\mathbf{q}_j,\mathbf{q}_k)}-1. \tag{3.5}$$

Here, we have introduced the activity

$$z=\frac{e^{\beta\mu}}{\Lambda^3}, \tag{3.6}$$

where the de Broglie thermal wavelength $\Lambda=\sqrt{2\pi\beta\hbar^2/m}$ originates from the momentum integration $\int dp e^{-\beta|p|^2/2m}/h=\Lambda^{-1}$.

Diagrammatically $-\beta\Omega$ can be expressed as a sum of connected diagrams built up with black circles, 2-bonds, 3-bonds, and pseudo 2-bonds, where every pair of circles connected by the same 3-bond should be connected by a pseudo 2-bond.[32] Here, the graphical terms are defined as follows. A 2-bond connects a pair of circles but a pair can not be connected directly more than once by 2-bonds. A 3-bond connects a triplet of circles but a triplet can not be connected directly more than once by 3-bonds. A pseudo 2-bond is basically the same as the 2-bond: it connects a pair of circles but a pair can not be connected directly more than once by pseudo 2-bonds. The difference between 2-bonds and pseudo 2-bonds is the following: if a pair is connected by the same 3-bond, the pair should be connected by a pseudo 2-bond; otherwise, simply by a 2-bond. The pseudo 2-bond is found to be very useful for a collective representation of diagrams which contain 3-bonds. Without use of pseudo 2-bonds, the number of diagrams is dramatically increased. We give all diagrams of $-\beta\Omega$ explicitly up to the fourth order in the activity z:

$$-\beta\Omega=-\beta\Omega^{(1)}+-\beta\Omega^{(2)}+\cdots, \tag{3.7}$$

where

$-\beta\Omega^{(1)}=$ (3.8)

$-\beta\Omega^{(2)}=$ 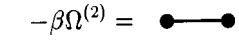 (3.9)

$-\beta\Omega^{(3)}=$ (3.10)

TABLE I. Analytical expressions corresponding to the symbols in the specified parts of the text. The white circles are the same as the black circles but no integrations are associated with them.

Symbol	Eqs. (3.8)–(4.7)	Eqs. (4.12)–(5.8) Appendix A	Eqs. (5.12)–(6.6) Appendix B	Eqs. (6.11)–(6.14)
1	z_1^*	n_1	n_1	n_1
1 2	f_{12}	f_{12}	h_{12}	h_{12}
1 2	$1+f_{12}$	$1+f_{12}$	$1+h_{12}$	$1+h_{12}$
3 1 2	t_{123}	t_{123}	t_{123}	γ_{123}

$-\beta\Omega^{(4)} =$ 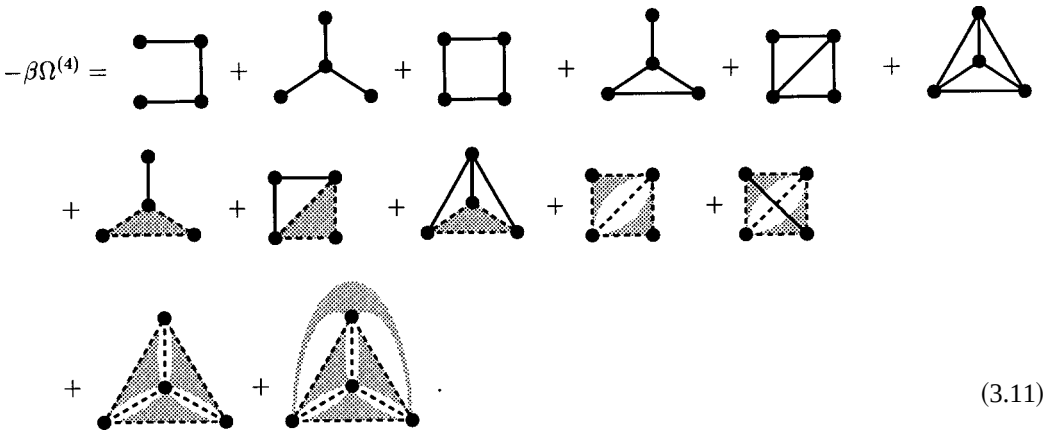 (3.11)

We note here that, as suggested above, every triplet connected by a 3-bond forms a triangle whose sides are pseudo 2-bonds. We can convince ourselves that all the diagrams at the fourth order are written out in Eq. (3.11) if we note the following three points, which, respectively, concern the fourth order diagrams including two, three, and four 3-bonds. (1) If we chose two distinct triplets from four distinct points, two (and only two) points are necessarily shared by the two triplets. (2) If we chose three distinct triplets from four distinct points, one (and only one) point is necessarily shared by the three triplets, while each of the other three points is shared by two (and only two) of the three triplets. (3) If we chose four distinct triplets from four distinct points, every point is necessarily shared by three (and only three) of the four triplets, and every pair of points should be shared by two (and only two) of the four triplets.

The algebraic expression corresponding to a diagram is obtained by the following steps (see Table I).

1. Assign the numbers 1,2,...,n to all the black circles in an arbitrary way. We associate the factors $z_1^*, z_2^*, \ldots, z_n^*$ to the black circles labeled 1,2,...,n, respectively. Here n is the total number of black circles in the diagram.
2. Associate the factor f_{ij} and $1+f_{ij}$ to the 2-bond and the pseudo 2-bond, respectively, whose ends are labeled i and j in the first step.
3. Associate the factor t_{ijk} to the 3-bond whose ends are labeled i, j, and k in the the first step.
4. Integrate the product of all the factors z^*, f, $1+f$ and t over the variables $\mathbf{q}_1, \ldots, \mathbf{q}_n$.
5. Multiply the integral by the inverse of the symmetry number S of the diagram.

The symmetry number is the number of all possible ways of labeling (one of which is done in the first step) topologically equivalent to each other. For example, if a diagram contains j black circles

which play equivalent roles (in the sense any interchange of the labels among those circles does not change the topology of the graph), the set of j circles contributes the symmetry factor $j!$. The symmetry numbers for the diagrams in Eqs. (3.8)–(3.11) are successively given by 1,2, (2,3!,3!), (2,3!,$2\cdot 4$,2!,2!2!,4!,…,4!).

Since the pseudo 2-bonds are not conventional, we present two examples of diagrams which contain the pseudo 2-bonds:

$$= \frac{1}{2}\int d1\int d2\int d3\int d4\; n_1 n_2 n_3 n_4 t_{123} f_{34}(1+f_{12})(1+f_{23})(1+f_{31}) \tag{3.12}$$

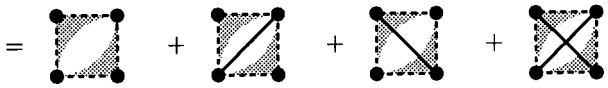
(3.13)

Though Eq. (3.13) is mathematically correct under the above rules 1–5 for analytical expressions, only the second expression (which consists of two diagrams) accords with our convention of distinguishing 2-bonds from pseudo ones. (For example, in the first expression, a pair which is not connected by the same 3-bond is connected by a pseudo 2-bond and thus this expression violates our convention.) Because of this, we employ the second expression in what follows.

IV. INVERSION FROM ONE-BODY POTENTIAL TO ONE-PARTICLE DENSITY

In this section we renormalize the theory in terms of the one-particle density by using the inversion method. To this end we identify ϕ in the inversion formulae with the one-particle density $n_1 \equiv \rho^{(1)}(\mathbf{r}_1)$. The diagram expression for n_1 can be easily obtained from that for $-\beta\Omega$ if we notice that the differentiation by the one-particle potential u_1 changes one of the black circles to a white circle. This is understood from

$$-\beta n_1 = -\beta \frac{\delta\Omega}{\delta u_1}, \tag{4.1}$$

and

$$\frac{\delta z_1^*}{\delta u_{1'}} = -\beta z_1^* \delta_{11'}. \tag{4.2}$$

Here, a white circle should be associated with z_i^* but no integration is performed over the variable i. Except for this rule for a white circle, the rules for obtaining an analytical expression from a diagram are not changed (see Table I).

We give all the diagrams up to the fourth order in the activity z:

$$n_1 = n_1^{(0)}[\tilde{u}\,] + n_1^{(1)}[\tilde{u}\,] + \cdots, \tag{4.3}$$

where

$$n_1^{(0)}[\tilde{u}] = \tag{4.4}$$

$$n_1^{(1)}[\tilde{u}] = \tag{4.5}$$

$$n_1^{(2)}[\tilde{u}] =$$ 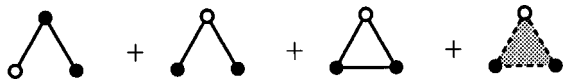 (4.6)

$$n_1^{(3)}[u] =$$ 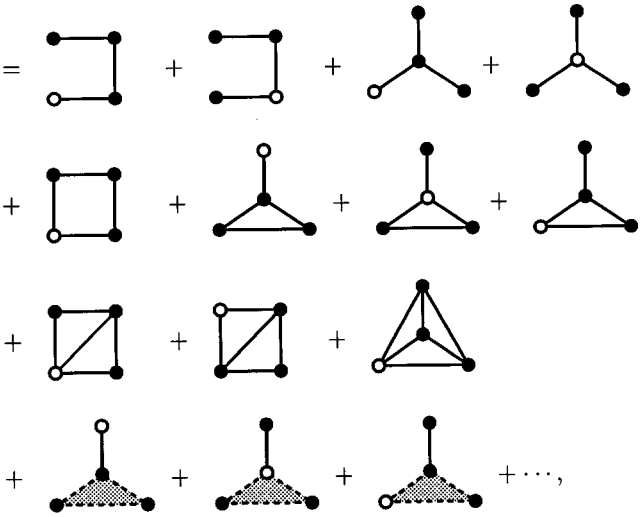 (4.7)

where $\cdots$ represents derivatives of the last six diagrams written in Eq. (3.11).

In the above $n^{(0)}$ is NOT the zero-th order in z and, strictly speaking, we should identify ϕ in the inversion formulae not with n_1 but n_1/z. However, here, we have identified ϕ with n_1 since the result is the same. We have also identified J with $\tilde{u}$ and regarded $\tilde{u}$ as an *artificial* source, though it should be set to the original value at the end. In other words, in the following we shall use the same symbol $\tilde{u}$ for both J and j.

From Eq. (1.5), or $n_1 = n_1^{(0)}[\tilde{u}^{(0)}]$, and from Eq. (4.4), we have

$$n_1 = e^{-\beta \tilde{u}_1^{(0)}}/\Lambda^3 \quad (-\beta \tilde{u}_1^{(0)} = \ln n_1 + 3 \ln \Lambda). \tag{4.8}$$

This relation implies

$$z_1^*|_{\tilde{u}\to\tilde{u}^{(0)}} = n_1. \tag{4.9}$$

This means that the replacement $\tilde{u}\to\tilde{u}^{(0)}$, as implied in the inversion formulae Eqs. (1.5)–(1.8), is equivalent to a change of interpretation of circles: in the new diagram all the white and black circles should be associated not with z_i^* but with n_i.

In the present case where $\phi = n_1$, Eq. (1.6) can be re-written as

$$\int d1' \, \frac{\delta n_1^{(0)}[\tilde{u}^{(0)}]}{\delta \tilde{u}_{1'}^{(0)}} \, \tilde{u}_{1'}^{(1)} + n_1^{(1)}[\tilde{u}^{(0)}] = 0, \tag{4.10}$$

where, from Eq. (4.4),

$$\frac{\delta n_1^{(0)}[\tilde{u}^{(0)}]}{\delta \tilde{u}_{1'}^{(0)}} = -\beta n_1 \delta_{11'}. \tag{4.11}$$

Thus, from Eq. (4.5), Eq. (4.10) is diagrammatically expressed as

$$-\beta n_1 \tilde{u}_1^{(1)} +$$ 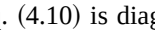 $$= 0,$$ (4.12)

from which we obtain

$$\beta\tilde{u}_1^{(1)} = \quad \text{—●} \quad . \tag{4.13}$$

We stress here in the above two graphs the white and black circles now correspond not to z_i^* but to n_i. This is because these diagrams originate from the inversion formula in which we replace $\tilde{u}_1$ by $\tilde{u}_1^{(0)}$ and, as explained above, this replacement changes the circles in the diagrams from z_1^* to n_i. By the same token, all the circles should be associated with n_i in what follows (see Table I).

The first term in Eq. (1.7) can be written as $-\beta n_1 \tilde{u}_1^{(2)}$ as above. The second term in Eq. (1.7) in this case is written as

$$\frac{1}{2}\phi^{(0)''}(J^{(1)})^2 = \frac{1}{2}\int d1' \int d2' \frac{\delta^2 n_1^{(0)}[\tilde{u}^{(0)}]}{\delta\tilde{u}_{1'}^{(0)}\delta\tilde{u}_{2'}^{(0)}}\tilde{u}_{1'}^{(1)}\tilde{u}_{2'}^{(1)}, \tag{4.14}$$

where, from Eq. (4.4),

$$\frac{\delta^2 n_1^{(0)}[\tilde{u}^{(0)}]}{\delta\tilde{u}_{1'}^{(0)}\tilde{u}_{2'}^{(0)}} = (-\beta)^2\delta_{11'}\delta_{12'} \quad \text{○} \quad . \tag{4.15}$$

From Eq. (4.13), Eq. (4.14) can be expressed as .

Since $\phi^{(1)\prime}$ in Eq. (1.7), in this case, implies [from Eq. (4.5)]

$$\frac{\delta n_1^{(1)}[\tilde{u}^{(0)}]}{\delta\tilde{u}_{1'}^{(0)}} = -\beta \quad \text{○—○}_{1\quad 1'} \quad -\beta\delta_{11'} \quad \text{○—●} \quad , \tag{4.16}$$

the third term in Eq. (1.7) can be expressed as 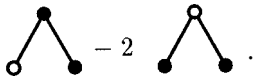 .

Here the factor 2 in front of the second graph appears because the symmetry factor of the graph is 2. (Up to now we have omitted the index 1 for a white circle, if there is only one white circle in the diagram.) Thus, Eq. (1.7) reduces to

$$-\beta n_1\tilde{u}_1^{(2)} + \quad \text{△} \quad -\left(\quad \text{△} \quad +2 \quad \text{△}\quad\right) + n_1^{(2)}[\tilde{u}^{(0)}] = 0. \tag{4.17}$$

In this equation, we see that the three diagrams explicitly written exactly cancel the 1-reducible diagrams appearing in $n_1^{(2)}[\tilde{u}^{(0)}]$. Here, the 1-reducible diagram is a diagram in which there is at least one circle the removal of which leads to a separation of the diagram. Thus we have

$$\beta\tilde{u}_1^{(2)} = \quad \overset{1}{\text{△}} \quad + \quad \overset{1}{\text{▲}} \quad . \tag{4.18}$$

Concerning Eq. (1.8), in a similar manner, we have (see Appendix A)

$$\phi^{(0)\prime}J^{(3)} = -\beta n_1\tilde{u}_1^{(3)}, \tag{4.19}$$

$$\phi^{(0)''}J^{(1)}J^{(2)} \rightarrow \quad \text{△} \quad + \quad \text{▲} \quad , \tag{4.20}$$

$$\frac{1}{3!}\phi^{(0)\prime\prime\prime}(J^{(1)})^3 \to -$$, (4.21)

$$\phi^{(1)\prime}J^{(2)} \to -$$ 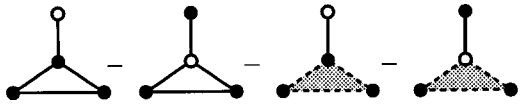 , (4.22)

$$\frac{1}{2}\phi^{(1)\prime\prime}(J^{(1)})^2 \to$$ 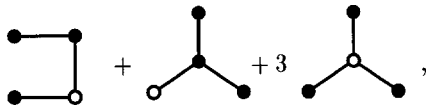 , (4.23)

$$\phi^{(2)\prime}J^{(1)} \to -$$ 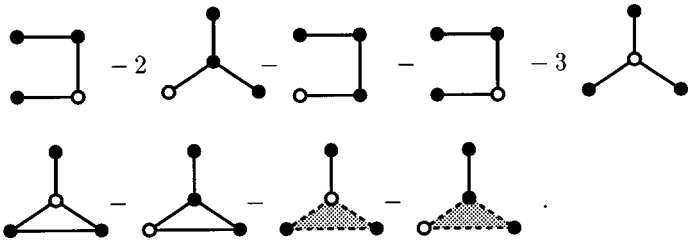 . (4.24)

By substituting these results in Eq. (1.8), we see again that all the 1-reducible diagrams exactly cancel out. As a result we have

$$\beta\tilde{u}_1^{(3)} =$$ 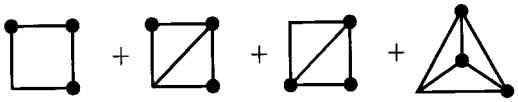 , (4.25)

where $\cdots$ represents derivatives of the last six diagrams written in Eq. (3.11).

We have derived $\tilde{u}_1$ explicitly up to the third order in n_1 (black circle) and in principle we can continue such a calculation up to the desired order. Then the inversion series is given by

$$\beta\tilde{u}_1 = \beta\tilde{u}_1^{(0)} + \beta\tilde{u}_1^{(1)} + \beta\tilde{u}_1^{(2)} + \cdots , \tag{4.26}$$

where $\tilde{u}_1^{(i)}$ $(i=0,1,2,3)$ are given in Eqs. (4.8), (4.13), (4.18), and (4.25). As we see below in Eq. (4.30), this inversion series, obtained by the inversion method, can be regarded as the on-shell condition and as a generalization of a well-known formula in simple liquids theory.

Now we can integrate the expression (4.26) to get a generating functional $\Gamma_1[n,v,w]$ due to the relation $\beta\tilde{u}_1 = -\beta\delta\Gamma_1[n,v,w]/\delta n_1$. The result is given by

$$-\beta\Gamma_1 = \int d1 n_1(1 - \ln n_1 - 3\ln\Lambda) +$$ 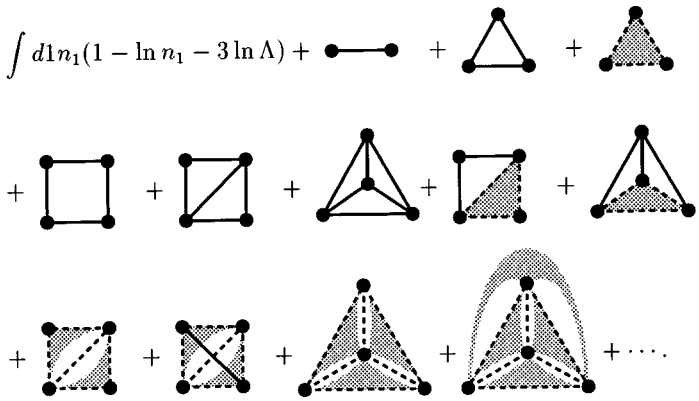 . (4.27)

We stress here that in the above diagram circles are associated with n_i and thus all the diagrams are functional not of z_i^* but of n_i. All the diagrams up to the fourth order of n_i are explicitly shown in Eq. (4.27).

Equation (4.27) can be related to the previously known results for the entropy and the thermodynamic potential. The entropy in the case where $v=w=0$ can be given by $-\beta\Gamma_1+(3/2)\times\langle N\rangle$, as stated before. Indeed, if we further set $u=0$ (the ideal gas), we recover the entropy of the ideal gas, $S/k=Vn(5/2-\ln n-3\ln\Lambda)$, where n and V is the density and the volume of the system, respectively. If we neglect the triplet potential, Eq. (4.27) reduces to the standard expression for the thermodynamic potential renormalized by one-particle density:[1,2]

$$-\beta\Omega=\int d1\, n_1(1-\ln n_1+\ln z-\beta u_1)+\mathscr{K}^{(1)}[n,f], \tag{4.28}$$

where $\mathscr{K}^{(1)}[n,f]$ is the whole class of the 1-irreducible (1IR) diagrams built up with n-circles and f-bonds (f-bonds are functionals of v). Here the 1IR diagram is the one in which deletion of any one of the black circles does not cause a separation of the diagram.

The on-shell condition for the first Legendre transformation [the first equation in (2.21)], or

$$\beta\,\frac{\delta\Gamma_1[n,v,w]}{\delta n_1}=-\beta\tilde{u}_1, \tag{4.29}$$

for Γ_1 in Eq. (4.27), reduces to

$$\beta u_1=-\ln n_1+\ln z+$$ 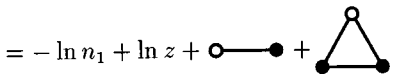 $$+\cdots, \tag{4.30}$$

which is equivalent to the inversion series (4.26), as stated above. This inversion series is a generalization of a familiar equation in the simple liquid theory:[1,4]

$$\ln(n_1/z_1^*)=K_1^{(1)}[n,f], \tag{4.31}$$

where $K_1^{(1)}[n,f]=\delta\mathscr{K}^{(1)}[n,f]/\delta n_1$ is a sum of all 1-irreducible diagrams consisting of one white circle labeled 1, one or more black circles, and f-bonds. If we again neglect the triplet potential and set u to zero (a homogeneous liquid), Eq. (4.30) reduces to this standard result (4.31).

V. INVERSION FROM PAIR POTENTIAL TO TWO-PARTICLE DENSITY

In this section we renormalize the theory in terms of the two-particle density by using the inversion method. To this end we try to identify ϕ in the inversion formulae with the two-particle density $n_{12}\equiv\rho^{(2)}(\mathbf{r}_1,\mathbf{r}_2)$. The diagram expression for n_{12} can be easily obtained from that for Γ_1 if we notice that the differentiation by the two-particle potential v_{12} removes one of the (pseudo) 2-bonds with making black circles of both ends white circles if we multiply the resultant diagram by a factor $1+f_{12}$. This is understood from the relations

$$-\beta n_{12}=-\beta\,\frac{\delta\Omega[u,v,w]}{\delta\,\frac{1}{2}\,v_{12}}=-\beta\,\frac{\delta\Gamma_1[n,v,w]}{\delta\,\frac{1}{2}\,v_{12}}, \tag{5.1}$$

and

$$\frac{\delta f_{12}}{\delta\,\frac{1}{2}\,v_{1'2'}}=-\beta(1+f_{12})(\delta_{11'}\delta_{22'}+\delta_{12'}\delta_{21'}). \tag{5.2}$$

Thus, we have

$$n_{12} = (1 + f_{12})(\circ\ \circ + \text{ } + \cdots) \tag{5.3}$$

We notice here that all the diagrams have two white circles, which stand for n_i. This implies that, if we consider n_i as the expansion parameter of the inversion method, $n_{12}^{(0)}$ vanishes while the inversion method assumes a non-zero $\phi^{(0)}$. To avoid this difficulty we identify ϕ with a new quantity defined by

$$h_{12} \equiv \frac{n_{12}}{n_1 n_2} - 1. \tag{5.4}$$

It is diagrammatically given as

$$h_{12} = h_{12}^{(0)}[v] + h_{12}^{(1)}[v] + \cdots, \tag{5.5}$$

where

$$h_{12}^{(0)}[v] = \text{ } \tag{5.6}$$

$$h_{12}^{(1)}[v] = (1 + f_{12})(\text{ 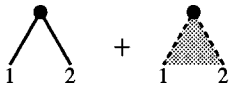 }) \tag{5.7}$$

$$h_{12}^{(2)}[v] = (1 + f_{12})(\text{ 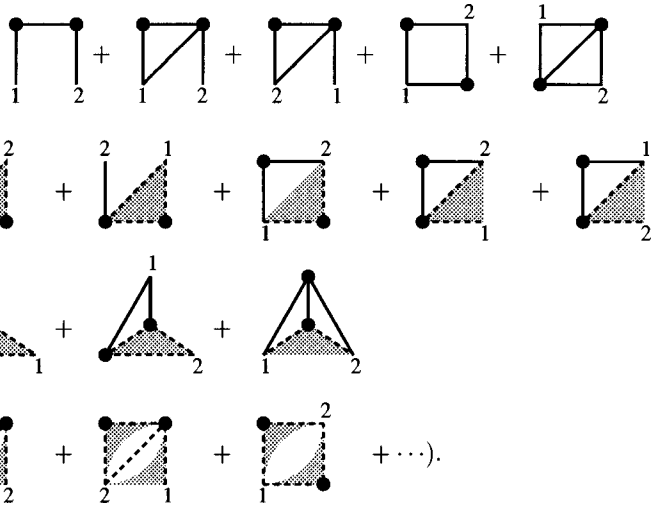 }) \tag{5.8}$$

Here, the terms represented by $\cdots$ are the derivatives of the last three diagrams in Eq. (4.27).

From Eq. (1.5), or $h_{12} = h_{12}^{(0)}[v^{(0)}]$, and from Eq. (5.6), we have

$$h_{12} = e^{-\beta v_{12}^{(0)}} - 1 \quad [-\beta v_{12}^{(0)} = \ln(h_{12} + 1)]. \tag{5.9}$$

This implies

$$f_{12}|_{v \to v(0)} = h_{12}. \tag{5.10}$$

Equation (1.6) in this case can be written as

$$\int d1' \int d2' \, \frac{\delta h_{12}^{(0)}[v^{(0)}]}{\delta v_{1'2'}^{(0)}} \, v_{1'2'}^{(1)} + h_{12}^{(1)}[v^{(0)}] = 0, \tag{5.11}$$

which reduces to (see Appendix B)

$$-\beta(1+h_{12})v_{12}^{(1)} + (1+h_{12})\Big(\quad + \quad \Big) = 0. \tag{5.12}$$

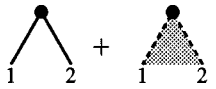

Thus, we have

$$\beta v_{12}^{(1)} = \quad + \quad . \tag{5.13}$$

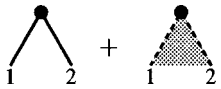

In the above diagrams the (pseudo) 2-bonds are not f-bonds but (pseudo) h-bonds due to the replacement $v \to v^{(0)}$ in the inversion formula (1.6) [see Eq. (5.10)]. Similarly the 2-bonds and the pseudo 2-bonds appearing in the following diagrams should all be associated with h_{ij} and $1+h_{ij}$, respectively (see Table I).

The terms in Eq. (1.7) are calculated as follows (see Appendix B):

$$\phi^{(0)\prime} J^{(2)} = -\beta(1+h_{12})v_{12}^{(2)}, \tag{5.14}$$

$$\frac{1}{2}\phi^{(0)\prime\prime}(J^{(1)})^2 \to (1+h_{12})\Big(\quad + \quad + \quad \Big), \tag{5.15}$$

$$\phi^{(1)\prime} J^{(1)} \to -(1+h_{12})\Big(2\Big(\quad + \quad + \quad \Big)$$

$$+2 \quad + \quad + \quad + \quad + \quad$$

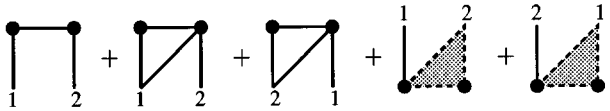

$$+ \quad + \quad + \quad + \quad \Big). \tag{5.16}$$

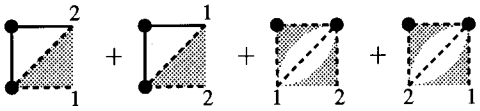

By substituting these results in Eq. (1.7), we see that all the 2-reducible diagrams except

exactly cancel out. Here, the 2-reducible diagrams are defined as follows. First, we make a 0-diagram from the diagram by connecting external points 1,2 by a (pseudo) 2-bond along with n_1, n_2 and then by integrating the diagram over the variables 1, 2. (The 0-diagram, here, is a diagram which has no external points and white circles.) Then the original diagram is 2-reducible if the resultant 0-diagram is the one in which there is at least one pair of circles the deletion of which leads to a separation of the diagram.

From Eqs. (5.14) to (5.16), we have

$$\beta v_{12}^{(2)} = - \quad + \quad + \quad + \quad + \quad + \cdots, \tag{5.17}$$

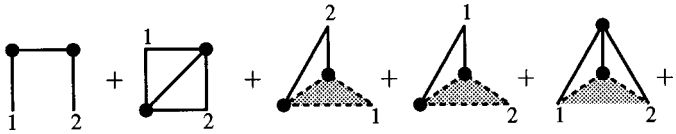

where $\cdots$ represents derivatives of the last three diagrams in Eq. (4.27).

We have derived v_{12} explicitly up to the second order in n_1 (black circle) and in principle we can continue such a calculation up to the desired order. Then the inversion series is given by

$$\beta v_{12} = \beta v_{12}^{(0)} + \beta v_{12}^{(1)} + \beta v_{12}^{(2)} + \cdots, \tag{5.18}$$

where $\beta v_{12}^{(i)}$ $(i=0,1,2)$ are given in Eqs. (5.9), (5.13), and (5.17). As we see below in Eq. (5.35), Eq. (5.18) can be regarded as the on-shell condition and as a generalization of some equations familiar in the theory of simple liquids.

Now we can integrate the inversion series (5.18) to get a generating functional $\Gamma_2[n,h,w]$ due to the relation $\beta v=-2\beta\delta\Gamma_2[n,h,w]/\delta\rho^{(2)}$, or

$$\beta v_{12}=-\beta\frac{2}{n_1n_2}\frac{\delta\Gamma_2[n,h,w]}{\delta h_{12}}, \tag{5.19}$$

and the result is given by

$$-\beta\Gamma_2=\int d1n_1(1-\ln n_1-3\ln\Lambda)+\frac{1}{2}\int d1\int d2n_1n_2[h_{12}-(1+h_{12})\ln(1+h_{12})]$$

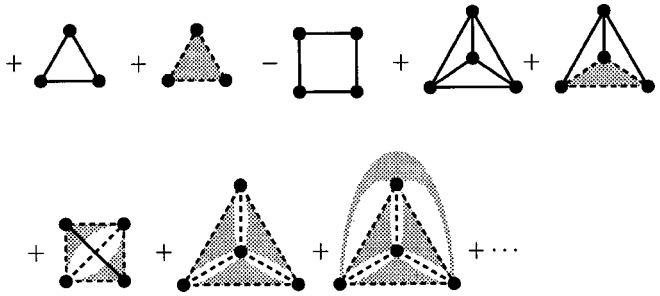

(5.20)

We stress here that in the above diagram circles, 2-bonds, and pseudo 2-bonds are associated with n_i, h_{ij} and $1+h_{ij}$, respectively, and thus all the diagrams are functional of n_i, h_{ij}, and w_{ijk}. All the diagrams up to the fourth order of n_i are explicitly shown in Eq. (5.20). As stated before, from this expression we can readily obtain the entropy for the system without the triplet potential, which coincides with the expression given in Ref. 1.

If we neglect the triplet potential, Eq. (5.20) reduces to the standard expression for the thermodynamic potential renormalized by one-particle density:[1,2]

$$-\beta\Omega=\int d1n_1(1-\ln n_1+\ln z-\beta u_1)-\frac{\beta}{2}\int d1\int d2\rho_{12}^{(2)}v_{12}$$
$$+\frac{1}{2}\int d1\int d2n_1n_2\{h_{12}-[1+h_{12}]\ln[1+h_{12}]\}+\mathcal{N}[n,h]+\mathcal{K}^{(2)}[n,h]. \tag{5.21}$$

$\mathcal{N}[n,h]$ are made up of polygonal diagrams and, if we regard $[nh]_{12}\equiv n_1h_{12}$ is the (1,2) element of the functional matrix $[nh]$, it is given by

$$\mathcal{N}[n,h]=-\sum_{p=3}^{\infty}\frac{(-1)^p}{2p}\mathrm{Tr}[nh]^p=\frac{1}{2}\mathrm{Tr}\left[\ln(1+nh)-nh+\frac{1}{2}(nh)^2\right]. \tag{5.22}$$

$\mathcal{K}^{(2)}[n,h]$ is the whole class of the 2-irreducible (2IR) diagrams built up with n-circles and h-bonds. Here, a 2IR diagram is the one in which the deletion of any pair of the black circles does not cause a separation of the diagram.

In the conventional definition, a 2-irreducible diagram consists of circles and 2-bonds but not of 3-bonds and pseudo 2-bonds. However, in the following, we include such diagrams consisting of 3-bonds and pseudo 2-bonds into the definition of 2IR diagrams, and call the whole class of such 2IR diagrams $\mathcal{K}^{(2)}[n,h,w]$ instead of $\mathcal{K}^{(2)}[n,h]$. Then, Eq. (5.20) is equivalent to Eq. (5.21) where $\mathcal{K}^{(2)}[n,h]$ is replaced by $\mathcal{K}^{(2)}[n,h,w]$.

The on-shell condition for the second Legendre transformation [the first two equations in (2.22)] are equivalent to

$$\frac{\delta\Gamma_2[n,h,w]}{\delta n_1} = -\tilde{u}_1 - \int d2\, n_2(1+h_{12})v_{12}, \tag{5.23}$$

$$\frac{\delta\Gamma_2[n,h,w]}{\delta \frac{1}{2} h_{12}} = -n_1 n_2 v_{12}, \tag{5.24}$$

from which we have

$$\ln z - \ln n_1 + \int d2\, n_2\{h_{12} - (1+h_{12})\ln(1+h_{12})\} + \frac{\delta\mathscr{N}[n,h]}{\delta n_1} + \frac{\delta\mathscr{Z}^{(2)}[n,h,w]}{\delta n_1}$$
$$= \beta u_1 + \beta \int d2\, n_2(1+h_{12})v_{12}, \tag{5.25}$$

$$-\ln(1+h_{12}) + \frac{2}{n_1 n_2}\left(\frac{\delta\mathscr{N}[n,h]}{\delta h_{12}} + \frac{\delta\mathscr{Z}^{(2)}[n,h,w]}{\delta h_{12}}\right) = \beta v_{12}, \tag{5.26}$$

where

$$\frac{\delta\mathscr{N}[n,h]}{\delta n_1} = \frac{1}{2}\left(h\cdot\frac{1}{1+[nh]} - h + h\cdot[nh]\right)_{11}, \tag{5.27}$$

$$\frac{2}{n_1 n_2}\frac{\delta\mathscr{N}[n,h]}{\delta h_{12}} = \left(h - h\,\frac{1}{1+[nh]}\right)_{12}. \tag{5.28}$$

This set of equations, corresponding to Eq. (4.30) in the previous section, is to be solved self-consistently for the two variables n and h. In the above, h and $[nh]$ stand for the functional matrix as before. The diagrammatic expression for the sum of derivatives of $\mathscr{N}[n,h]$ or $\mathscr{Z}[n,h,w]$ in Eq. (5.25) and in Eq. (5.26) can be easily obtained from the diagrams in Eq. (5.20) by removing, in all possible ways, one of the black circles and one of the (pseudo) 2-bonds (along with two black circles at the both ends), respectively.

If we introduce the direct correlation function $c(\mathbf{r}_1,\mathbf{r}_2)=c_{12}$ by the so-called Ornstein–Zernike relation,

$$h_{12} = c_{12} + \int d3\, c_{13} n_3 h_{32}, \tag{5.29}$$

or, in our simplified notation,

$$h = c + c\cdot[nh] = c + c\cdot[nc] + c\cdot[nc]^2 + \cdots, \tag{5.30}$$

we have

$$c = h - c\cdot[nh] = h - h\cdot[nh] + h\cdot[nh]^2 - \cdots = h\,\frac{1}{1+[nh]}. \tag{5.31}$$

Thus we obtain

$$\frac{\delta\mathscr{N}[n,h]}{\delta n_1} = \frac{1}{2}\,(c - h + h\cdot[nh])_{11}, \tag{5.32}$$

$$\frac{2}{n_1 n_2}\frac{\delta\mathscr{N}[n,h]}{\delta h_{12}} = h_{12} - c_{12}, \tag{5.33}$$

from which we have

$$\ln z-\ln n_1+\int d2n_2\{h_{12}-(1+h_{12})[\ln(1+h_{12})+\beta v_{12}]\}+\frac{1}{2}\left(c_{11}-h_{11}+\int d2h_{12}n_2h_{21}\right)$$
$$+\frac{\delta\mathscr{K}^{(2)}[n,h,w]}{\delta n_1}=\delta u_1, \tag{5.34}$$

$$-\ln(1+h_{12})+h_{12}-c_{12}+\frac{2}{n_1n_2}\frac{\delta\mathscr{K}^{(2)}[n,h,w]}{\delta h_{12}}=\beta v_{12}. \tag{5.35}$$

Note here that c_{12} and the derivatives of $\mathscr{K}^{(2)}[n,h,w]$ are explicitly given as functionals of n and h. This set of equations are self-consistent equations for n_i,h_{ij}, which is equivalent to the set of (5.25) and (5.26), and Eq. (5.35) can be regarded as another expression of the inversion series (5.18), as stated above.

The inversion series (5.35), obtained as the on-shell condition, is a generalization of a familiar exact relation,

$$h_{12}-c_{12}-\ln(1+h_{12})=\beta v_{12}-d_{12}, \tag{5.36}$$

which was derived by van Leenwen *et al.*,[4] if d_{12}, originally given by the f-bond expansion, is rewritten as an h-bond expansion.

The hypernetted-chain (HNC) approximation,

$$h_{12}-c_{12}-\ln(1+h_{12})=\beta v_{12}, \tag{5.37}$$

is reproduced if we neglect all the 2IR diagrams in Eq. (5.35).

The Percus–Yevick (PY) approximation can be expressed as

$$\beta v_{12}+\ln(1+h_{12})=\ln(1+h_{12}-c_{12}), \tag{5.38}$$

implying the approximation

$$h_{12}-c_{12}+\frac{2}{n_1n_2}\frac{\delta\mathscr{K}^{(2)}[n,h,w]}{\delta h_{12}}\simeq\ln(1+h_{12}-c_{12}). \tag{5.39}$$

VI. INVERSION FROM TRIPLET POTENTIAL TO THREE-PARTICLE DENSITY

In this section we renormalize the theory in terms of three-particle density by using the inversion method. To this end we try to identify ϕ in the inversion formulae with the two-particle density $n_{12}\equiv\rho^{(2)}(\mathbf{r}_1,\mathbf{r}_2)$. The diagram expression for n_{123} can be easily obtained from that for Γ_2 if we notice that the differentiation by the three-particle potential w_{123} removes one of the 3-bonds with making black circles of the three ends white circles if we multiply the resultant diagram by a factor $1+t_{123}$. This is understood from the relations

$$-\beta n_{123}=-\beta\frac{\delta\Omega[u,v,w]}{\delta\frac{1}{3!}w_{123}}=-\beta\frac{\delta\Gamma_2[n,h,w]}{\delta\frac{1}{3!}w_{123}} \tag{6.1}$$

and

$$\frac{\delta t_{123}}{\delta\frac{1}{3!}w_{1'2'3'}}=-\beta(1+t_{123})[\delta_{11'}\delta_{22'}\delta_{33'}+(3!-1)\text{terms}]. \tag{6.2}$$

Then, we notice that all the diagrams of n_{123} have three white circles. This implies that, if we regard n_i as the expansion parameter of the inversion method, $n_{123}^{(0)}$ vanishes while the inversion

method assumes non-zero $\phi^{(0)}$. In addition, since, in all the diagrams of $\Gamma_2[n,h,w]$, every t_{ijk} is accompanied by a factor $n_i n_j n_k(1+h_{ij})(1+h_{jk})(1+h_{ki})$, it is convenient to identify ϕ not with n_{123} but with γ_{123} defined by

$$\gamma_{123}=\frac{n_{123}}{n_1 n_2 n_3(1+h_{12})(1+h_{23})(1+h_{31})}-1. \tag{6.3}$$

Then this quantity is diagrammatically given as

$$\gamma_{123}=\gamma_{123}^{(0)}+\gamma_{123}^{(1)}+\cdots, \tag{6.4}$$

where

$\gamma_{123}^{(0)} =$ 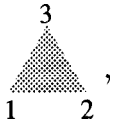

, (6.5)

$\gamma_{123}^{(1)} = (1+t_{123})($ 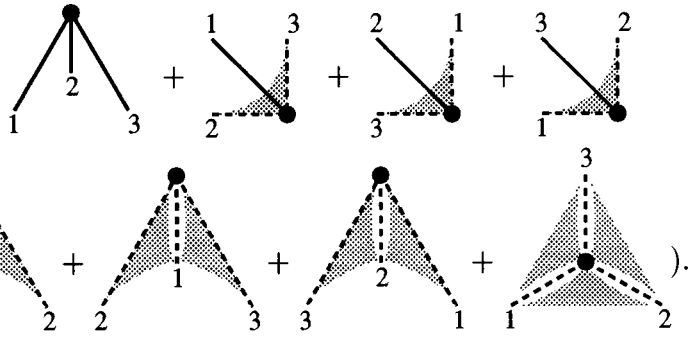

$).$ (6.6)

From Eq. (1.5), or $\gamma_{123}=\gamma_{123}^{(0)}[w^{(0)}]$, and from Eq. (6.5), we have

$$\gamma_{123}=e^{-\beta w_{123}^{(0)}}-1 \quad [-\beta w_{123}^{(0)}=\ln(\gamma_{123}+1)]. \tag{6.7}$$

This implies

$$t_{123}|_{w\to w^{(0)}}=\gamma_{123}. \tag{6.8}$$

Equation (1.6) in this case can be written as

$$\int d1'\int d2'\int d3'\,\frac{\delta\gamma_{123}^{(0)}[w^{(0)}]}{\delta w_{1'2'3'}^{(0)}}\,w_{1'2'3'}^{(1)}+\gamma_{123}^{(1)}[w^{(0)}]=0, \tag{6.9}$$

where

$$\frac{\delta\gamma_{123}^{(0)}[w^{(0)}]}{\delta w_{1'2'3'}^{(0)}}=-\frac{\beta}{3!}\,(1+\gamma_{123})[\,\delta_{11'}\delta_{22'}\delta_{33'}+(3!-1)\text{terms}]. \tag{6.10}$$

Thus, we have

$\beta w_{123}^{(1)} =$ 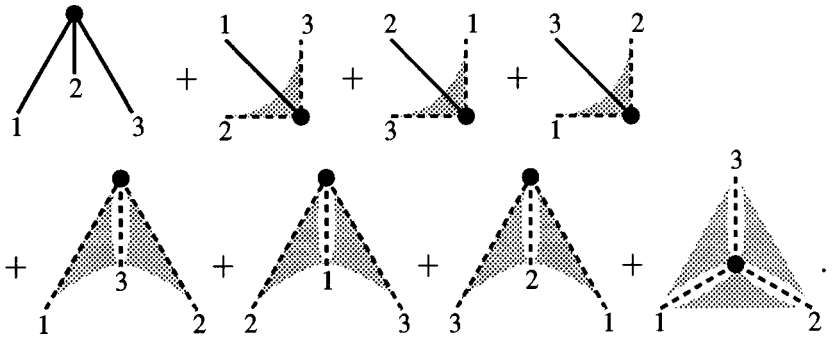

(6.11)

Here, the 3-bonds should be associated not with t_{ijk} but with γ_{ijk} [see Eq. (6.8) and Table I].

We have derived w_{123} explicitly up to the first order in n_1 (black circle) and in principle we can continue such a calculation up to the desired order. Then the inversion series is given as

$$\beta w_{123}=\beta w_{123}^{(0)}+\beta w_{123}^{(1)}+\beta w_{123}^{(2)}+\cdots , \tag{6.12}$$

where $\beta w_{123}^{(i)}$ $(i=0,1)$ are given in Eqs. (6.7) and (6.11). As we see below in Eq. (6.20), this inversion series can be obtained from the on-shell condition and can be seen as a series which systematically improve the Kirkwood's approximation for the three-particle density.

Now we can integrate the inversion series (6.12) to get a generating functional $\Gamma_3[n,h,\gamma]$ due to the relation $\beta w=-3!\beta\delta\Gamma_3[n,h,\gamma]/\delta\rho^{(3)}$, or

$$\beta w_{123}=-\beta\,\frac{3!}{n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})}\,\frac{\delta\Gamma_3[n,h,\gamma]}{\delta\gamma_{123}}, \tag{6.13}$$

and the result is given by

$$-\beta\Gamma_3=\int d1n_1(1-\ln n_1-3\ln\Lambda)+\frac{1}{2}\int d1\int d2n_1n_2[h_{12}-(1+h_{12})\ln(1+h_{12})]$$

$$+\frac{1}{3!}\int d1\int d2\int d3n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})[\gamma_{123}-(1+\gamma_{123})\ln(1+\gamma_{123})]$$

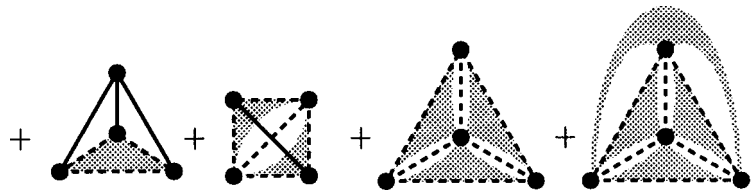 (6.14)

In the above diagrams, the 3-bonds are not t-bonds but γ-bonds due to the replacement $w\to w^{(0)}$ in the inversion formula (1.6) [see Eq. (6.8) and Table 1]. We stress here that, in the above diagram, circles, 2-bonds, pseudo 2-bonds, and 3-bonds are associated with n_i, h_{ij}, $1+h_{ij}$, and γ_{ijk}, respectively, and thus all the diagrams are functional of n_i, h_{ij}, and γ_{ijk}. All the diagrams up to the fourth order of n_i are explicitly shown in Eq. (6.14).

From Eq. (6.14), the renormalized entropy S/k of the system can be easily obtained as a functional of one- to three-particle densities due to the relation $S/k=-\beta\Gamma_3+(3/2)\int d1n_1$, as stated earlier. This result for the entropy should be compared with the previously known result, for example, given in Ref. 8. Up to the third order (in n_i), these two coincide with each other. However, above this order, they look different. For example, while the fourth order of $\Delta S/k$ given in Ref. 8 includes the four-particle density, our expression does not contain the four-particle density and is written in terms of only one- to three-particle densities. This implies that, for the system without higher-body potential than 3-body, the four-particle density can be written in terms of one- to three-particle densities.

The result (6.14) was suggested in Ref. 2, where the diagrammatic part is collectively called $\mathscr{T}[n,h,\gamma]$. In there, $\mathscr{T}[n,h,\gamma]$ is not defined explicitly and any diagrams contained in it are not presented, unfortunately.

The on-shell conditions for the third Legendre transformation [the three equations in (2.23)] are equivalent to

$$\frac{\delta\Gamma_3}{\delta n_1}=-u_1-\int d2n_2(1+h_{12})v_{12}-\frac{1}{2}\int d2d3n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})(1+\gamma_{123})w_{123}, \tag{6.15}$$

$$\frac{\delta\Gamma_3}{\delta h_{12}}=-\frac{1}{2}n_1n_2v_{12}-\frac{1}{2}n_1n_2\int d3n_3(1+h_{23})(1+h_{31})(1+\gamma_{123})w_{123}, \tag{6.16}$$

$$\frac{\delta\Gamma_3}{\delta\gamma_{123}}=-\frac{1}{3!}n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})w_{123}, \tag{6.17}$$

from which we have

$$\ln z-\ln n_1+\int d2n_2[h_{12}-(1+h_{12})\{\ln(1+h_{12})+\beta v_{12}\}]+\frac{1}{2}\int d2d3n_2n_3(1+h_{12})(1+h_{23})$$

$$\times(1+h_{31})[\gamma_{123}-(1+\gamma_{123})\{\ln(1+\gamma_{123})+\beta w_{123}\}]+\frac{\delta\mathscr{T}}{\delta n_1}=\beta u_1, \tag{6.18}$$

$$-\ln(1+h_{12})+\int d3n_3(1+h_{23})(1+h_{31})[\gamma_{123}-(1+\gamma_{123})\{\ln(1+\gamma_{123})+\beta w_{123}\}]$$

$$+\frac{2}{n_1n_2}\frac{\delta\mathscr{T}}{\delta h_{12}}=\beta v_{12}, \tag{6.19}$$

$$-\ln(1+\gamma_{123})+\frac{3!}{n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})}\frac{\delta\mathscr{T}}{\delta\gamma_{123}}=\beta w_{123}. \tag{6.20}$$

This set of equations, corresponding to the set of (5.25) and (5.26) in the previous section, is to be solved for the three variables n, h, and γ.

This set can also be seen as a generalization of Eqs. (6)–(8) in Ref. 12. If we approximate $\mathscr{T}$ by the fifth term in Eq. (5) in Ref. 12, which is originally obtained in Ref. 11, this set of equations here reduce to Eqs. (6)–(8) in Ref. 12. However, we notice that our diagrammatic expression for $\mathscr{T}$ does not have diagrams directly corresponding to the fifth term in Eq. (5) in Ref. 12. This is because this fifth term originates from the last term in Eq. (4) of Ref. 12, and this term, which should be equal to $\mathscr{T}$ for the system without higher order potential than the third, contains correlation functions of up to the infinite order while $\mathscr{T}$ here contains correlation functions up to the third order.

As mentioned above, Eq. (6.20) can be expressed as

$$\beta w_{123}=\beta w^{(0)}+\beta w^{(1)}+\beta w^{(2)}+\cdots, \tag{6.21}$$

where

$$\beta w^{(0)}=-\ln(1+\gamma_{123}), \tag{6.22}$$

$$\beta w^{(1)}=\int d4n_4h_{41}h_{42}h_{23}+\cdots, \tag{6.23}$$

where $\cdots$ represents γ-dependent terms which correspond to the last seven diagrams in Eq. (6.11).

Equation (6.21) or (6.20) can be related to some of the previously known approximations. If we truncate the inversion series (6.21) at $i=0$, we obtain the well-known approximation $\gamma_{123}=e^{-\beta w_{123}}-1$ (for example, see Ref. 12). At $w_{123}=0$, this approximation reduces to the Kirkwood's superposition approximation,

$$n_{123}=n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31}). \tag{6.24}$$

If we truncate the series at $i=1$ and further neglect the γ-dependent terms in $\beta w^{(1)}$ (with $w_{123}=0$), we get the first order approximation given in Ref. 24,

$$n_{123}=n_1n_2n_3(1+h_{12})(1+h_{23})(1+h_{31})\exp\left[\int d4n_4h_{41}h_{42}h_{43}\right]. \tag{6.25}$$

Equation (6.25) with h-bonds in the exponential replaced by f-bonds had also been studied (see, for example, Refs. 21–23).

Thus, in our formulation, the zero-th order approximation reproduces the superposition approximation and the first order approximation, which is the solution of the equation [obtained from Eq. (6.11)],

$$\begin{aligned}\beta w_{123}=&-\ln(1+\gamma_{123})+\int d4n_4h_{41}h_{42}h_{43}+\int d4n_4h_{41}(1+h_{42})(1+h_{43})\gamma_{423}\\&+\int d4n_4h_{42}(1+h_{41})(1+h_{43})\gamma_{413}+\int d4n_4h_{43}(1+h_{41})(1+h_{42})\gamma_{412}\\&+\int d4n_4(1+h_{41})(1+h_{42})(1+h_{43})(\gamma_{412}\gamma_{423}+\gamma_{421}\gamma_{413}+\gamma_{413}\gamma_{423})\\&+\int d4n_4(1+h_{41})(1+h_{42})(1+h_{43})\gamma_{412}\gamma_{423}\gamma_{431},\end{aligned} \tag{6.26}$$

contains more information than the first approximation in Ref. 24. Solving our first order approximation is practically an interesting problem, which is reserved for a future work. Further systematic improvement of the approximation is obtained if we calculate $\beta w^{(2)},\beta w^{(3)},\ldots$, by the inversion method.

VII. DISCUSSION

In this paper, we applied the inversion method to a classical system and renormalized liquid theory in terms of one-, two-, and three-particle densities in the presence of the three-body potential. Our approach here complements the previously known re-summation techniques. We showed that the entropy functional can be expressed by up to three-particle densities and gave lower order diagrams of the entropy explicitly. We also pointed out that the stationary conditions for the entropy leads to a set of self-consistent equations. In a certain case, the self-consistent equation can be a base for a systematic advancement of the Kirkwood's superposition approximation.

Some of the results presented here are mere repetition or only a slight extension of the well-known results as suggested frequently in the text. We could not avoid mixing up known and unknown results to explain our approach explicitly. Here, we appreciate and clarify to what extent these results can be considered original. In Section II, the introduction of Γ_3 and its relation to the entropy seems a rather novel viewpoint. In Sections III–V, all the results are at least implied in the previous literature (see, for example, Ref. 19). However, we pointed out that, if one introduces the pseudo 2-bonds, the diagrams of Ω and Γ_n in the presence of the 3-body potential can be written economically and we gave the diagrams of (up to) the fourth order ($\Omega^{(4)}$, etc.) explicitly; otherwise the number of diagrams is too many to be included in a paper—this may be one reason the fourth order diagrams (in the presence of the 3-body potential) have not been given explicitly in the literature. In addition, the derivation of Γ_n based on the inversion method has originality and complements the previously known ways. The results in Sec. VI are the renormalization in terms

of one-, two-, and three-particle densities. In that the entropy does not include more-than-three-particle densities, this explicit diagrammatic result may be original. The result reduces to some previously known approximations in special cases.

The present work can be related to the work by Stell[19] in the following way. In the second article in Ref. 19, he gave diagrammatic rules for some correlation functions in terms of ρ_1, h_2, and f_s (where $s \geqslant 3$) in his notation. In the special case where there are no more-than-three-body potentials, his result can be regarded as the diagrammatic rule in terms of one- and two-particle densities and the three-body potential. (Notice here that his f_3 is our t_{ijk}.) This corresponds to the description by Γ_2, which is still a functional of the three-body potential w_{123} (or f_3), in our language: the theory is renormalized up to the two-particle densities. In the present work, though it is limited to the special case (without m-body potential where $m \geqslant 4$), we further changed the variable w_{ijk} (or f_3) of the theory to the three-particle density $\rho^{(3)}$ or γ_{ijk}. This is accomplished by use of Γ_3, which is a functional of $\rho^{(3)}$, and then the theory is renormalized up to the three-particle density. For example, $w_3(123)$ in Eq. (3.12) of the second article in Ref. 19 (which is expressed in terms of $\rho^{(1)}$, h_{ij}, and t_{ijk}), coincides with the second term on the left-hand side of Eq. (6.20) of this article (which is expressed in terms of $\rho^{(1)}$, h_{ij}, and γ_{ijk}), when renormalized up to the three-particle densities by changing the variable from t_{ijk} to γ_{ijk}.

ACKNOWLEDGMENT

The author would like to thank Professor Fumio Hirata for an informative discussion.

APPENDIX A: USEFUL FORMULAE FOR THE INVERSION FROM ONE-PARTICLE POTENTIAL TO ONE-PARTICLE DENSITY

In this appendix, we present useful formulae for the inversion process. Below $\phi^{(i)\prime\cdots\prime}$ stands for $\delta^k n_1^{(i)}[u^{(0)}]/\delta\tilde{u}_{1'}^{(0)}\delta\tilde{u}_{2'}^{(0)}\cdots\delta\tilde{u}_{k'}^{(0)}$:

$$\phi^{(0)\prime} \to -\beta\delta_{11'} \quad 1 \tag{A1}$$

$$\phi^{(0)\prime\prime} \to (-\beta)^2\delta_{11'}\delta_{12'} \quad 1 \tag{A2}$$

$$\phi^{(0)\prime\prime\prime} \to (-\beta)^3\delta_{11'}\delta_{12'}\delta_{13'} \quad 1 \tag{A3}$$

$$\phi^{(1)\prime} \to -\beta \quad 1 \text{—} 1' \quad -\beta\delta_{11'} \quad 1 \text{—} \bullet \tag{A4}$$

$$\phi^{(1)\prime\prime} \to (-\beta)^2(\delta_{11'}\delta_{12'} \quad 1 \text{—} \bullet \quad +\delta_{11'} \quad 1 \text{—} 2' \quad +\delta_{12'} \quad 1 \text{—} 1' \quad +\delta_{1'2'} \quad 1 \text{—} 1' \quad) \tag{A5}$$

$$\phi^{(2)\prime} \to (\ \triangle\ )' + (\ \triangle\ )' + (\ \triangle\ )' + (\ \blacktriangle\ )', \tag{A6}$$

where

$$(\ \triangle\ )' = -\beta(\delta_{11'}\ 1\ \triangle\ +\ 1'\ \triangle\ 1\ +\ \triangle\ 1\ 1'\ ) \tag{A7}$$

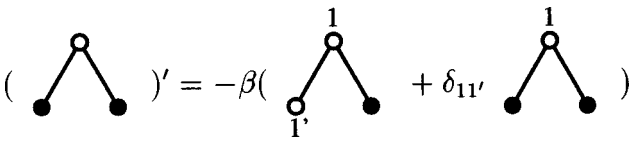
(A8)

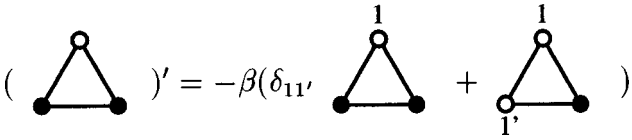
(A9)

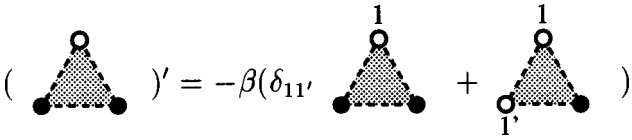
(A10)

APPENDIX B: USEFUL FORMULAE FOR THE INVERSION FROM TWO-PARTICLE POTENTIAL TO TWO-PARTICLE DENSITY

In this appendix, we present useful formulae for the inversion process. Below $\phi^{(i)\prime\cdots\prime}$ stands for $\delta^k h_{12}^{(i)}[v^{(0)}]/\delta v_{1'}^{(0)}\delta v_{2'}^{(0)}\cdots\delta v_{k'}^{(0)}$:

$$\phi^{(0)\prime}\to-\frac{\beta}{2}(1+h_{12})(\delta_{11'}\delta_{22'}+\delta_{12'}\delta_{21'}) \tag{B1}$$

$$\phi^{(0)}\to\left(-\frac{\beta}{2}\right)^2(1+h_{12})(\delta_{11'}\delta_{22'}+\delta_{12'}\delta_{21'})(\delta_{13'}\delta_{24'}+\delta_{14'}\delta_{23'}) \tag{B2}$$

$\phi^{(1)\prime}\to(1+h_{12})'($ 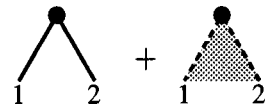 $)+(1+h_{12})($ 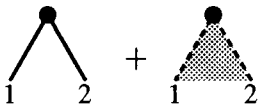 $)'$ (B3)

where

$$(1+h_{12})'=-\frac{\beta}{2}(1+h_{1'2'})(\delta_{11'}\delta_{22'}+\delta_{12'}\delta_{21'}) \tag{B4}$$

$($ 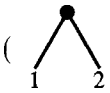 $)'=-\frac{\beta}{2}(1+h_{1'2'})($ $)$ (B5)

$($ $)'=-\frac{\beta}{2}($ 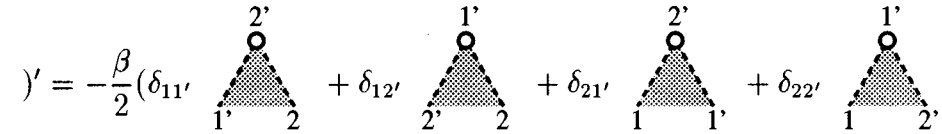 $)$ (B6)

[1] T. Morita and K. Hiroike, Prog. Theor. Phys. **25**, 537 (1961).
[2] C. De Dominicis, J. Math. Phys. **3**, 983 (1962).
[3] G. Stell, Physica **29**, 517 (1963); G. Stell, ''Cluster expansions for classical systems in equilibrium,'' in *The Equilibrium Theory of Classical Fluids*, edited by H. L. Frisch and J. L. Lebowitz (Benjamin, New York, 1964), p. II-171.
[4] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic Press, London, 1991).
[5] R. Fukuda, M. Komachiya, S. Yokojima, Y. Suzuki, K. Okumura, and T. Inagaki, Prog. Theor. Phys. Suppl. **121**, 1 (1995), and references therein.
[6] H. S. Green, *The Molecular Theory of Fluids* (North-Holland, Amsterdam, 1952).

[7] R. E. Nettleton and M. S. Green, J. Chem. Phys. **29**, 1365 (1958).
[8] H. J. Raveché, J. Chem. Phys. **55**, 2242 (1971); **55**, 2250 (1971).
[9] D. C. Wallace, J. Chem. Phys. **87**, 2282 (1987).
[10] A. Baranyai and D. J. Evans, Phys. Rev. A **40**, 3817 (1989).
[11] J. A. Hernando, Mol. Phys. **69**, 319 (1990); **69**, 327 (1990).
[12] B. B. Laird, J. Wang, and A. D. J. Haymet, Phys. Rev. E **47**, 2491 (1993).
[13] A. Baranyai and D. J. Evans, Z. Naturforsch. **46a**, 27 (1991).
[14] B. B. Laird, J. Wang, and A. D. J. Haymet, J. Chem. Phys. **97**, 2153 (1992); Phys. Rev. A **45**, 5680 (1992).
[15] G. Casanova, R. J. Dulla, D. A. Jonah, J. S. Rowlinson, and G. Saville, Mol. Phys. **18**, 589 (1970).
[16] R. J. Dulla, J. S. Rowlinson, and W. R. Smith, Mol. Phys. **21**, 299 (1971).
[17] J. Ram and Y. Singh, J. Chem. Phys. **66**, 924 (1977).
[18] A. D. J. Haymet, J. Phys. (Paris) **46**, C9 (1985).
[19] G. Stell, J. Math. Phys. **6**, 1193 (1965); G. Stell, ''Correlation functions and their generating functionals,'' in *Phase Transitions and Critical Phenomena*, edited by Domb and M. S. Green (Academic, London, 1976), Vol. 5B, p. 205.
[20] A. G. Schlijper and R. Kikuchi, J. Stat. Phys. **61**, 143 (1990).
[21] S. A. Rice and J. Lekner, J. Chem. Phys. **42**, 3559 (1965).
[22] D. A. Young and S. A. Rice, J. Chem. Phys. **46**, 539 (1967).
[23] D. A. Young and S. A. Rice, J. Chem. Phys. **47**, 5061 (1967).
[24] A. D. J. Haymet, S. A. Rice, and W. G. Madden, J. Chem. Phys. **74**, 3033 (1981).
[25] M. Ukita, M. Komachiya, and R. Fukuda, Int. J. Mod. Phys. D **5**, 1789 (1990).
[26] T. Inagaki and R. Fukuda, Phys. Rev. B **46**, 10931 (1992).
[27] K. Okumura, J. Magn. Magn. Mater. **140–144**, 191 (1995); Phys. Rev. B **52**, 13358 (1995); Int. J. Mod. Phys. A **11**, 65 (1996).
[28] If we consider the system without the three-body force, $w(\mathbf{r},\mathbf{r}',\mathbf{r}'')$ here is artificial. We can still extract new information, which otherwise could not be obtained, by setting $w(\mathbf{r},\mathbf{r}',\mathbf{r}'')=0$ at the end (as mentioned before).
[29] The two-particle density can also be obtained by noting the following relation:
$$-(1/\beta)[\,\delta^2\Omega/\delta u(\mathbf{r})\,\delta u(\mathbf{r}')\,]=\langle\rho_N^{(1)}(\mathbf{r})\rho_N^{(1)}(\mathbf{r}')\rangle-\langle\rho_N^{(1)}(\mathbf{r})\rangle\langle\rho_N^{(1)}(\mathbf{r}')\rangle=\rho^{(2)}(\mathbf{r},\mathbf{r}')+\delta(\mathbf{r}-\mathbf{r}')\rho^{(1)}(\mathbf{r})-\rho^{(1)}(\mathbf{r})\rho^{(1)}(\mathbf{r}'). \tag{B7}$$
The second equality can be understood from
$$\rho^{(2)}(\mathbf{r},\mathbf{r}')=\langle\rho_N^{(1)}(\mathbf{r})\rho_N^{(1)}(\mathbf{r}')\rangle=\delta(\mathbf{r}-\mathbf{r}')\rho^{(1)}(\mathbf{r}), \tag{B8}$$
which is obtained by taking the expectation of Eq. (2.4).
[30] We tacitly assume $\rho_N^{(1)}(\mathbf{r})=0$ for $N=0$, $\rho_N^{(2)}(\mathbf{r},\mathbf{r}')=0$ for $N=0$, 1, and $\rho_N^{(3)}(\mathbf{r},\mathbf{r}',\mathbf{r}'')=0$ for $N=0,1,2$.
[31] K. Okumura, Prog. Theor. Phys. **87**, 703 (1992).
[32] The right-hand side of Eq. (3.1) can be expressed by a sum of all possible diagrams built up with the four elements (black circles, 2-bonds, 3-bonds, and pseudo 2 ponds), which includes disconnected diagrams. Here, as suggested in the text, if there is a 3-bond, every pair from the triplet (connected by the 3-bond) should be connected by a pseudo 2-bond. On the other hand, due to the operation of logarithm, $-\beta\Omega$ does not include disconnected diagrams.

Integrable coupled KdV systems

Metin Gürses
Department of Mathematics, Faculty of Sciences, Bilkent University, 06533 Ankara, Turkey

Atalay Karasu
Department of Physics, Faculty of Arts and Sciences, Middle East Technical University, 06531 Ankara, Turkey

(Received 23 September 1997; accepted for publication 5 November 1997)

We give the conditions for a system of N-coupled Korteweg de Vries (KdV) type of equations to be integrable. We find the recursion operators of each subclass and give all examples for $N=2$. © *1998 American Institute of Physics.* [S0022-2488(98)03003-5]

I. INTRODUCTION

In Ref. 1 we gave an extension of the recently proposed Svinolupov Jordan KdV[2,3] systems to a class of integrable multicomponent KdV systems and gave their recursion operators. This class is known as the degenerate subclass of the KdV system. In this work we will extend it to a more general KdV type of system equations containing both the degenerate and nondegenerate cases. This is a major step towards the complete classification of KdV systems. In addition we give a new extension of such a system of equations.

Let us consider a system of N nonlinear equations of the form

$$q_t^i = b_j^i q_{xxx}^j + s_{jk}^i q^j q_x^k, \tag{1}$$

where $i,j,k=1,2,...,N$, q^i are functions depending on the variables x, and t, and b_j^i, and S_{jk}^i are constants. The purpose of this work is to find the conditions on these constants so that the equations in (1) are integrable. In general the existence of infinitely many conserved quantities is admitted as the definition of integrability. This implies the existence of infinitely many generalized symmetries. In this work we assume the following definition for integrability:

Definition: A system of equations is said to be integrable if it admits a recursion operator.

The recursion operator (if it exists) of the system of equations given in (1), in general, may take a very complicated form. Let the highest powers of the operators D and D^{-1} be respectively defined by $m=$degree of R and $n=$order nonlocality of R. In this work we are interested in a subclass of equations admitting a recursion operator with $m=2$ and $n=1$. Namely, it is of the form

$$R_j^i = b_j^i D^2 + a_{jk}^i q^k + c_{jk}^i q_x^k D^{-1}, \tag{2}$$

where D is the total x derivative, D^{-1} is the inverse operator, and a_{jk}^i and c_{jk}^i are constants with

$$s_{jk}^i = a_{kj}^i + c_{jk}^i. \tag{3}$$

Before starting to the classification of (1) we recall a few fundamental properties of the recursion operator. An operator R_j^i is a recursion operator if it satisfies the following equation

$$R_{j,t}^i = F_k'^i R_j^k - R_k^i F_j'^k, \tag{4}$$

where $F_k'^i$ is the Fréchet derivative of the system (1), which is given by

$$\sigma_t^i = F_j'^i \sigma^j, \tag{5}$$

 © 1998 American Institute of Physics

where σ^i's are called the symmetries of the system (1). The condition (3) implies that Eq. (1) itself is assumed to be in the family of the hierarchy of equations (or flows)

$$q^i_{\tau_n}=\sigma^i_n,$$

where for all $n=0,1,...$, σ^i_n denotes the symmetries of the integrable KdV system (1). For instance, for $n=0,1$ we have respectively the classical symmetries $\sigma^i_0=q^i_x$ and $\sigma^i_1=q^i_t$.

Equation (5) is called the symmetry equation of (1) with

$$F'^{\,i}_j=b^i_jD^3+s^i_{jk}q^k_x+s^i_{kj}q^kD. \tag{6}$$

Recursion operators are defined as operators mapping symmetries to symmetries, i.e.,

$$R^i_j\sigma^j=\lambda\sigma^i, \tag{7}$$

where λ is an arbitrary constant. Equations (5) and (7) imply (4). It is the equation (4) which determines the constants a^i_{jk} and c^i_{jk} in terms of b^i_j and s^i_{jk}'s. The same equation (4) brings severe constraints on b^i_j and s^i_{jk}.

We shall obtain a classification of (1) based on the matrix b^i_j,

(i) $\det(b^i_j)=0$,
(ii) $\det(b^i_j)\neq 0$,

and also we divide the classification procedure, for each class, into two parts where $s^i_{jk}=s^i_{kj}$ and $s^i_{jk}\neq s^i_{kj}$. For the system of equations admitting a recursion operator we have the following proposition.

Proposition 1: Let $q^i(t,x)$ be functions of t and x satisfying the N KdV equations (1) and admitting a recursion operator R^i_j in (2). Then the constants b^i_j, s^i_{jk}, a^i_{jk}, and c^i_{jk} satisfy [in addition to the (3)] the following relations:

$$b^k_lc^i_{jk}-b^i_kc^k_{jl}=0, \tag{8}$$

$$b^k_la^i_{jk}-b^i_k(a^i_{jl}+3c^k_{jl}-s^k_{jl})=0, \tag{9}$$

$$b^i_k(3a^k_{jl}+3c^k_{jl}-2s^k_{jl}-s^k_{lj})=0, \tag{10}$$

$$c^i_{jk}s^k_{lm}-s^i_{lk}c^k_{jm}=0, \tag{11}$$

$$c^i_{jk}s^k_{lm}+c^i_{jk}s^k_{ml}-c^k_{jm}s^i_{kl}-c^k_{jl}s^i_{km}=0, \tag{12}$$

$$a^i_{jk}s^k_{lm}-s^i_{km}a^k_{jl}-s^i_{lk}a^k_{jm}-s^i_{lk}c^k_{jm}+s^k_{jl}c^i_{km}+a^i_{kl}s^k_{jm}=0, \tag{13}$$

$$c^i_{km}(s^k_{pj}-s^k_{jp})=0. \tag{14}$$

Now we will discuss the problem of classifying the integrable system of equations (1) for the two exclusive cases depending upon the matrix b^i_j.

II. CLASSIFICATION FOR THE CLASS $\det(b^i_j)=0$

In this subclass we assume the rank of the matrix b^i_j as $N-1$. Investigation of the subclasses for other ranks of matrix b can be done similarly. For this case we may take $b^i_j=\delta^i_j-k^ik_j$ where k_i is a unit vector, $k^ik_i=1$. In this work we use the Einstein convention, i.e., repeated indices are summed up from 1 to N. We find the following solution of (8)–(14) for the parameters a^i_{jk} and c^i_{jk} for all N

Proposition 2: Let k^i be a constant unit vector and $b^i_j=\delta^i_j-k^ik_j$. Then the complete solutions of the equations (8)–(14) are given by

$$a^i_{lj}=\tfrac{2}{3}s^i_{jl}+\tfrac{1}{3}[k^i(k_ja_l-2k_ln_j)+(-ak^i+b^i)k_lk_j],$$
$$c^i_{jl}=\tfrac{1}{3}s^i_{lj}-\tfrac{1}{3}[k^i(k_la_j+k_jn_l)+(-ak^i+b^i)k_lk_j]+k^ik_ln_j, \tag{15}$$

where

$$n_j=k^kk_is^i_{jk}-ak_j, \quad a_j=k^kk_is^i_{kj}-nk_j, \quad b^i=k^lk^ks^i_{lk}-nk^i, \tag{16}$$

$a=k^na_n$, $n=k^in_i$ and s^i_{jk}'s are subject to satisfy the following:

$$c^i_{jk}s^k_{lm}=c^k_{jm}s^i_{lk}, \tag{17}$$

$$s^i_{jk}-s^i_{kj}=k^i[k_j(a_k-n_k)-k_k(a_j-n_j)], \tag{18}$$

$$k_ns^n_{lj}=n_lk_j+k_la_j, \tag{19}$$

$$k^ns^i_{jn}=k^in_j+b^ik_j, \tag{20}$$

$$k^ns^i_{nj}=(n-a)k^ik_j+k^ia_j+b^ik_j, \tag{21}$$

$$n[(a_i-n_i)k_j-(a_j-n_j)k_i]=0, \tag{22}$$

$$a_i=\rho n_i+ak_i, \tag{23}$$

where ρ is a constant. At this point we will discuss the classification procedure with respect to the symbol s^i_{jk} whether it is symmetric or nonsymmetric with respect to its lower indices.

A. The symmetric case, $s^i_{jk}=s^i_{kj}$

Among the constraints listed in Proposition 2 the one given in (22) implies that s^i_{jk}'s are symmetric if and only if $a_i-n_i=\alpha k_i$ where $\alpha=a-n$. There are two distinct cases depending on whether $n=0$ or $n\neq 0$. We shall give these two subcases as corollaries of the previous proposition.

Corollary 1: Let $s^i_{jk}=s^i_{kj}$ and $n=0$. Then we have the following solution for all N:

$$a^i_{kj}=\tfrac{2}{3}s^i_{jk}+\tfrac{1}{3}[k^i(k_ja_k-2k_kn_j)+k_kk_jb^i],$$
$$c^i_{jk}=\tfrac{1}{3}s^i_{jk}-\tfrac{1}{3}[k^i(k_ja_k-2k_kn_j)+k_kk_jb^i], \tag{24}$$

where $a=0$, $\rho=1$, and

$$a_l=n_l, n_l=k_ik^js^i_{lj}, b^i=k^jk^ls^i_{lj}. \tag{25}$$

The vector k^i and s^i_{jk} are not arbitrary; they satisfy the following constraints:

$$s^i_{jk}s^k_{lm}-s^i_{lk}s^k_{jm}=-2(k_jn_l-k_ln_j)(-k^in_m+b^ik_m),$$
$$k_ns^n_{lj}=n_lk_j+k_ln_j, \quad k^ns^i_{jn}=k^in_j+b^ik_j. \tag{26}$$

As an illustration we give an example for this case.[1] A particular solution of the equations (26) for $N=2$ is

$$b^i_j=\delta^i_j-y^iy_j=x^ix_j,$$
$$s^i_{jk}=\frac{3}{2}\alpha_1x^ix_jx_k+\alpha_2x^iy_jy_k+\frac{\alpha_1}{2}y^i(y_jx_k+y_kx_j), \tag{27}$$

where $i,j=1,2$ and

$$x^i=\delta^i_1, y^i=\delta^i_2, \tag{28}$$

and

$$k_i=y_i, \quad n_i=\tfrac{1}{2}\alpha_1 x_i, \quad b^i=\alpha_2 x^i. \tag{29}$$

Constants a^i_{jk} and c^i_{jk} appearing in the recursion operator are given by

$$\begin{aligned} a^i_{jk}&=\alpha_1 x^i x_j x_k+\alpha_2 x^i y_j y_k+\frac{\alpha_1}{2} y^i x_j y_k, \\ c^i_{jk}&=\frac{\alpha_1}{2} x^i x_j x_k+\frac{\alpha_1}{2} y^i x_j y_k. \end{aligned} \tag{30}$$

Taking $\alpha_1=2$ and $\alpha_2=1$ (without loss of generality) we obtain the following coupled system

$$u_t=u_{xxx}+3uu_x+vv_x, \quad v_t=(uv)_x. \tag{31}$$

The above system was first introduced by Ito[4] and the multi-Hamiltonian structure studied by Antonowicz and Fordy[5] and by Olver and Rosenau.[6] The recursion operator of this system is given by

$$R=\begin{pmatrix} D^2+2u+u_xD^{-1} & v \\ v+v_xD^{-1} & 0 \end{pmatrix}. \tag{32}$$

In Ref. 1 we have another example for $N=3$.

For the case $n\neq 0$ for all N we have the following.

Corollary 2: Let $s^i_{jk}=s^i_{kj}$, $n\neq 0$, and $\rho=0$. Then the solution given in Proposition 2 reduces to

$$\begin{aligned} a^i_{lj}&=\tfrac{2}{3}s^i_{jl}+\tfrac{1}{3}(a-2n)k^ik_jk_l, \\ c^i_{jl}&=\tfrac{1}{3}s^i_{lj}-\tfrac{1}{3}(a-2n)k^ik_jk_l, \end{aligned} \tag{33}$$

where

$$n_j=nk_j, \quad a_j=ak_j, \quad b^i=ak^i, \tag{34}$$

and the constraint equations

$$s^i_{kj}s^k_{ml}-s^k_{mj}s^i_{kl}=0, \quad k_ns^n_{lj}=(a+n)k_lk_j, \quad k^ns^i_{jn}=(a+n)k^ik_j. \tag{35}$$

For this case we point out that solution of (33) and (35) gives decoupled systems.

B. The nonsymmetric case, $s^i_{jk}\neq s^i_{kj}$

In this case the constraints in Proposition 1, in particular (22), implies that we must have $n=0$. In this case we have the following expressions for a^i_{lj} and c^i_{lj} for all N:

$$\begin{aligned} a^i_{lj}&=\tfrac{2}{3}s^i_{jl}+\tfrac{1}{3}[k^i(k_ja_l-2k_ln_j)+(-ak^i+b^i)k_lk_j], \\ c^i_{jl}&=\tfrac{1}{3}s^i_{lj}-\tfrac{1}{3}[k^i(k_la_j+k_jn_l)+(-ak^i+b^i)k_lk_j]+k^ik_ln_j, \end{aligned} \tag{36}$$

where

$$\begin{aligned} n_j&=k^kk_is^i_{jk}-ak_j, \\ a_j&=k^kk_is^i_{kj}, \quad b^i=k^lk^ks^i_{lk}, \end{aligned} \tag{37}$$

and the constraint equations among the parameters are

$$c^i_{jk}c^k_{lm}-c^i_{lk}c^k_{jm}=0, \tag{38}$$

$$(a^i_{jk}-c^i_{jk})s^k_{lm}+(c^i_{km}-a^i_{km})s^k_{lj}+(s^i_{mk}-s^i_{km})a^k_{jl}+(s^k_{jm}-s^k_{mj})a^i_{kl}=0. \tag{39}$$

For $N=2$ we will give an example. Constants a^i_{jk} and c^i_{jk} appearing in the recursion operator are given by

$$\begin{aligned} a^i_{jk}&=\alpha_1 x^i x_j x_k+\alpha_2 x^i y_j y_k+\alpha_3 y^i x_j y_k,\\ c^i_{jk}&=\frac{\alpha_1}{2}x^i x_j x_k+\alpha_1 y^i x_j y_k, \end{aligned} \tag{40}$$

where α_1, α_2, and α_3 are arbitrary constants and

$$k_i=y_i, \quad n_i=\alpha_3 x_i, \quad b^i=\alpha_2 x^i, \quad a_i=\alpha_1 x_i. \tag{41}$$

We obtain the following coupled system:

$$\begin{aligned} u_t&=u_{xxx}+3\alpha_1 u u_x,\\ v_t&=\alpha_3 u_x v+\alpha_1 u v_x, \end{aligned} \tag{42}$$

which is equivalent to the symmetrically coupled KdV system[7]

$$\begin{aligned} u_t&=u_{xxx}+v_{xxx}+6uu_x+4uv_x+2u_x v,\\ u_t&=u_{xxx}+v_{xxx}+6vv_x+4vu_x+2v_x u, \end{aligned} \tag{43}$$

and the recursion operator for this integrable system of equations (42) is

$$R=\begin{pmatrix} D^2+\alpha_1(2u+u_xD^{-1}) & 0\\ \alpha_3 v+\alpha_1 v_x D^{-1} & 0 \end{pmatrix}. \tag{44}$$

III. CLASSIFICATION FOR THE CASE det $(b^i_j)\neq 0$

As in the degenerate case det $(b^i_j)=0$, we have two subcases, symmetric and nonsymmetric. Before these we have the following proposition.

Proposition 3: Let det $(b^i_j)\neq 0$. Then the solution of equations given Proposition 1 is given as follows:

$$\begin{aligned} a^i_{jl}&=\tfrac{2}{9}(s^i_{lj}+2s^i_{jl})-\tfrac{1}{9}C^k_l b^i_m K^m_{jk},\\ c^i_{lj}&=\tfrac{1}{9}(7s^i_{lj}-4s^i_{jl})+\tfrac{1}{9}C^k_l b^i_m K^m_{jk}, \end{aligned} \tag{45}$$

where

$$K^i_{lj}=s^i_{lj}-s^i_{jl} \tag{46}$$

and the constraint equations

$$b^k_l c^i_{jk}-b^i_k c^k_{jl}=0, \tag{47}$$

$$5C^m_i K^i_{rj}-C^l_r K^m_{lj}-C^l_j K^m_{rl}=0, \tag{48}$$

$$c^k_{jm}K^i_{lk}+c^k_{jl}K^i_{mk}=0, \tag{49}$$

$$c^i_{jk}s^k_{lm}-c^k_{jm}s^i_{lk}=0, \tag{50}$$

$$K^k_{lj}c^i_{km}=0, \tag{51}$$

$$(a^i_{jk}-c^i_{jk})s^k_{lm}+(c^i_{km}-a^i_{km})s^k_{lj}+K^i_{mk}a^k_{jl}+K^k_{jm}a^i_{kl}=0. \tag{52}$$

where C^i_r is the inverse of b^r_i.

A. The nonsymmetric case, $s^i_{jk}\neq s^i_{kj}$

Equations (47)–(52) define an over-determined system for the components of s^i_{jk}. Any solution of this system leads to the determination of the parameters a^i_{jl} and c^i_{jl} by (45).

As an example we give the following, for $N=2$, coupled system

$$v_t=av_{xxx}+2bvv_x,$$

$$u_t=4au_{xxx}+2bu_xv+buv_x, \tag{53}$$

where a and b are arbitrary constants. This system, under a change of variables, is equivalent to the KdV equation with the time evolution part of its Lax equation.[8] The recursion operator of the system (53) is

$$R=\begin{pmatrix} \frac{4}{3}(3aD^2+bv) & \frac{b}{3}(3u+2u_xD^{-1}) \\ 0 & \frac{1}{3}(3aD^2+4bv+2bv_xD^{-1}) \end{pmatrix}. \tag{54}$$

Hence the KdV equation coupled to time evolution part of its Lax-pair is integrable and its recursion operator is given above. This is the only new example for $N=2$ system.

B. The symmetric case, $s^i_{jk}=s^i_{kj}$

When the symbol s^i_{jk} is symmetric with respect to subindices, the parameters K^i_{jk} vanish. Then the equations (45)–(52) reduce to

$$a^i_{jk}=\tfrac{2}{3}s^i_{jk}, \quad c^i_{jk}=\tfrac{1}{3}s^i_{jk}, \tag{55}$$

where the parameters b^i_k and s^i_{jk} satisfy

$$b^k_l s^i_{jk}-b^i_k s^k_{jl}=0, \tag{56}$$

$$s^i_{jk}s^k_{lm}-s^i_{lk}s^k_{jm}=0. \tag{57}$$

We shall not study this class in detail, because in Ref. 1 some examples of this class are given for $N=2$. Here we give another example which correspond to the perturbation expansion of the KdV equation. Let $q^i=\delta^i u$, where $i=0,1,2,...,N$, and u satisfies the KdV equation $u_t=u_{xxx}+6uu_x$. The q^i's satisfy a system of KdV equations which belong to this subclass:

$$q^0_t=q^0_{xxx}+6q^0q^0_x, \tag{58}$$

$$q^1_t=q^1_{xxx}+6(q^0q^1)_x, \tag{59}$$

$$q^2_t=q^2_{xxx}+6[(q^1)^2+q^0q^2]_x, \tag{60}$$

$$\cdots. \tag{61}$$

$$q^N_t=q^N_{xxx}+3\sum_{i=1}^{N}[\delta^i(q^0)^2]_x. \tag{62}$$

IV. FOKAS–LIU EXTENSION

The classification of the KdV system given in this work with respect to the symmetries can be easily extended to the following simple modification of (1):

$$q_t^i = b_j^i q_{xxx}^j + s_{jk}^i q^j q_x^k + \chi_j^i q_x^j, \tag{63}$$

where χ_j^i's are arbitrary constants. Equation (63) without the last term will be called the principle part of that equation. Hence the equation (1) we have studied so far is the principle part of its modification (63). We assume the existence of a recursion operator corresponding to the above system in the form

$$R_j^i = b_j^i D^2 + a_{jk}^i q^k + c_{jk}^i q_x^k D^{-1} + w_j^i, \tag{64}$$

where w_j^i's are constants. We have the following proposition corresponding to the integrability of the above system.

Proposition 4: The operator given in (64) is the recursion operator of the KdV system (63) if in addition to the equations listed in Proposition 1 [(8)–(14)] the following constraints on the constants χ_j^i and w_j^i are satisfied:

$$\chi_l^k a_{jk}^i - \chi_k^i a_{jl}^k - \chi_k^i c_{jl}^k + \chi_j^k c_{kl}^i - w_j^k s_{kl}^i + w_k^i s_{jl}^k = 0, \tag{65}$$

$$\chi_l^k c_{jk}^i - \chi_k^i c_{jl}^k = 0, \tag{66}$$

$$\chi_k^i w_j^k - \chi_j^k w_k^i = 0, \tag{67}$$

$$(\chi_j^k - w_j^k) b_k^i - (\chi_k^i - w_k^i) b_j^k = 0, \tag{68}$$

$$\chi_j^k a_{kl}^i - \chi_k^i a_{jl}^k + w_k^i s_{lj}^k - w_j^k s_{lk}^i = 0. \tag{69}$$

Since the constraints (8)–(14) are enough to determine the coefficients a_{jk}^i and c_{jk}^i with some constraints on the given constants b_j^i and s_{jk}^i, we have the following corollary of the above proposition.

Corollary 3: The KdV system in (63) is integrable if and only if its principle part is integrable.

The principle part of (63) is obtained by ignoring the last term (the term with χ_j^i). Hence the proof of this corollary follows directly by observing that the constraints on the constants χ_j^i and w_j^i listed in (65)–(69) are independent of the constraints on the constants of the principle part listed in (8)–(14). Before the application of this corollary let us go back to the Proposition 4 and ask the question whether the KdV system (63) admits a recursion operator with $w_j^i = 0$.

Corollary 4: The KdV system (63) admits a recursion operator of the principle part. Then the last term $\chi_j^i q_x^j$ is a symmetry of the principle part. If $w_j^i = 0$, the above equations (65)–(69) reduce to

$$\chi_l^k a_{jk}^i - \chi_k^i a_{jl}^k - \chi_k^i c_{jl}^k + \chi_j^k c_{kl}^i = 0, \tag{70}$$

$$\chi_l^k s_{jk}^i - \chi_k^i s_{jl}^k = 0, \tag{71}$$

$$\chi_j^k b_k^i - \chi_k^i b_j^k = 0, \tag{72}$$

$$\chi_l^k c_{jk}^i - \chi_k^i c_{jl}^k = 0. \tag{73}$$

In order that the term $\chi_j^i q_x^j$ be a symmetry of the principle part, the constants χ_j^i are subject to satisfy the following equations:

$$\chi_l^k s_{jk}^i - \chi_k^i s_{jl}^k = 0, \tag{74}$$

$$\chi_j^k b_k^i - \chi_k^i b_j^k = 0, \tag{75}$$

$$\chi_l^k K_{jk}^i+\chi_j^k K_{lk}^i=0. \tag{76}$$

These equations simply follow from the set of equations (70)–(73), and hence the quantities $\sigma^i=\chi_j^i q_x^j$ are symmetries of the principle part.

By the application of Corollary 3, the full classification of the system (63) with the recursion operator (64) such that q_t^i (i.e., the system of equations themselves) belong to the symmetries of this system is possible. To each subclass given in the previous sections there exists a Fokas–Liu extension such that $w_j^i=\chi_j^i$ with the following constraints:

$$\chi_k^i c_{jl}^k-\chi_l^k c_{jk}^i=0, \tag{77}$$

$$\chi_l^k a_{jk}^i-\chi_j^k a_{lk}^i-\chi_k^i(a_{jl}^k-a_{lj}^k)=0. \tag{78}$$

The above constraints are identically satisfied for the class [$\det(b_j^i)\neq 0$, symmetrical case] when $\chi_j^i=\alpha\delta_j^i+\beta b_j^i$. Hence the Fokas–Liu extension of the nondegenerate symmetrical case is straightforward with this choice of χ_j^i. Here α and β are arbitrary constants.

For the degenerate case the set of equations (77) and (78) must solved for a given principle part, b_j^i and a_{jk}^i. Recently a system of integrable KdV system with $N=2$ has been introduced by Fokas and Liu.[9] This system is a nice example for the application of Corollary 3. We shall give this system in its original form first and then simplify:

$$\begin{aligned} &u_t+v_x+(3\beta_1+2\beta_4)\beta_3 uu_x+(2+\beta_1\beta_4)\beta_3(uv)_x+\beta_1\beta_3 vv_x\\ &\quad+(\beta_1+\beta_4)\beta_2 u_{xxx}+(1+\beta_1\beta_4)\beta_2 v_{xxx}=0, \end{aligned} \tag{79}$$

$$\begin{aligned} &v_t+u_x+(2+3\beta_1\beta_4)\beta_3 vv_x+(\beta_1+2\beta_4)\beta_3(uv)_x+\beta_1\beta_3\beta_4 uu_x\\ &\quad+(\beta_1+\beta_4)\beta_2\beta_4 u_{xxx}+(1+\beta_1\beta_4)\beta_2\beta_4 v_{xxx}=0, \end{aligned} \tag{80}$$

where β_1, β_2, β_3, and β_4 are arbitrary constants. The recursion operator of this system is given in Ref. 8. Consider now a linear transformation

$$u=m_1 r+n_1 s, \quad v=m_2 r+n_2 s, \tag{81}$$

where m_1, m_2, n_1, and n_2 are constants, and s and r are new dynamical variables, $q^i=(s,r)$. Choosing these constants properly, the Fokas–Liu system reduces to a simpler form

$$\begin{aligned} r_t&=(rs)_x+\alpha_1 r_x+\alpha_2 s_x,\\ s_t&=\gamma_1 s_{xxx}+\gamma_2 r r_x+3ss_x+\alpha_3 r_x+\alpha_4 s_x, \end{aligned} \tag{82}$$

where we are not giving the coefficients $\alpha_1, \alpha_2, \alpha_3$, and α_4 in terms of the parameters of the original equation given above, because these expressions are quite lengthy. The only condition on the parameters α_i is given by $\alpha_3=\gamma_2\alpha_2$. This guarantees the integrability of the above system (82). On the other hand, the transformation parameters are given by

$$m_2=-\frac{\beta_1+\beta_4}{1+\beta_1\beta_4}m_1, \quad n_2=\beta_4 n_1, \tag{83}$$

$$n_1=-\frac{1}{\delta\beta_3}, \quad \delta=\beta_1(1+\beta_4^2)+2\beta_4. \tag{84}$$

The principle part of the Fokas–Liu system (82) is exactly the Ito system given in (31) and hence the recursion operator is the sum of the one given in (32) and χ_j^i which are given by $\chi_1^1=\alpha_4$, $\chi_2^1=\alpha_3$, $\chi_1^2=\alpha_2$, $\chi_2^2=\alpha_1$. That is,

$$R=\begin{pmatrix} \gamma_1 D^2+2s+s_x D^{-1}+\alpha_4 & \gamma_2 r+\alpha_3\\ r+r_x D^{-1}+\alpha_2 & \alpha_1 \end{pmatrix}. \tag{85}$$

Another example was given very recently[10] in a very different context for $N=2$:

$$u_t=\tfrac{1}{2}v_{xxx}+2uv_x+u_xv, \quad v_t=3vv_x+2au_x. \tag{86}$$

The principle part of these equations is transformable to the Fuchssteiner system given in (42). Taking $\alpha_1=2$, $\alpha_3=4$, and scaling x and t properly we obtain (without losing any generality)

$$r_t=2r_{xxx}+6rr_x, \quad s_t=2s_xr+4sr_x. \tag{87}$$

The transformation between the principle part ($a=0$) of (86) and (87) is simply given by $u=m_1s+\frac{1}{2}r$, $v=2r$. Then the recursion operator of the system (86) is given by

$$R=\begin{pmatrix} 0 & \frac{1}{2}D^2+2u+u_xD^{-1} \\ 2a & 2v+v_xD^{-1} \end{pmatrix}. \tag{88}$$

V. CONCLUSION

We have given a classification of a system of KdV equations with respect to the existence of a recursion operator. This is indeed a partial classification. Although we have found all conditions for each subclass, we have not presented them explicitly. We obtained three distinct subclasses for all values of N and gave the corresponding recursion operators. We also gave an extension of such systems by adding a linear term containing the first derivative of dynamical variables. We called such systems the Fokas–Liu extensions. We proved that these extended systems of KdV equations are also integrable if and only if their principle parts are integrable. For $N=2$, we have given all subclassess explicitly. Among these the recursion operator of the KdV coupled to the time evolution part of its Lax pair seems to be new. Here we would like to add that when $N=2$ recursion operators, including the Fokas–Liu extensions, are hereditary.[11]

Our classification crucially depends on the form of the recursion operator. The recursion operators used in this work were assumed to have degree two (highest degree of the operator D in R) and nonlocality order one (highest degree of the operator D^{-1} in R). The next work in this program should be the study on the classification problems with respect to the recursion operators with higher degree and higher nonlocalities. For instance, when $N=2$, Hirota–Satsuma, Boussinesq, and Bogoyavlenskii coupled KdV equations admit recursion operators with $m=4$ and $n=1$.[12] Hence these equations do not belong to our classification given in this work.

ACKNOWLEDGMENTS

This work is partially supported by the Scientific and Technical Research Council of Turkey (TUBITAK). M. G. is a member Turkish Academy of Sciences (TUBA).

[1] M. Gürses and A. Karasu, Phys. Lett. A **214**, 21 (1996).
[2] S. L. Svinolupov, Teor. Mat. Fiz. **87**, 391 (1991).
[3] S. I. Svinolupov, Funct. Anal. Appl. **27**, 257 (1993).
[4] M. Ito, Phys. Lett. A **91**, 335 (1982).
[5] M. Antonowicz and A. P. Fordy, Physica D **28**, 345 (1987).
[6] P. J. Olver and P. Rosenau, Phys. Rev. E **53**, 1900 (1996).
[7] B. Fuchssteiner, Prog. Theor. Phys. **65**, 861 (1981).
[8] M. Ablowitz and H. Segur, *Solutions and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[9] A. S. Fokas and Q. M. Liu, Phys. Rev. Lett. **77**, 2347 (1996).
[10] L. Bonora, Q. P. Liu, and C. S. Xiong, Commun. Math. Phys. **175**, 177 (1996).
[11] B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47 (1981).
[12] M. Gürses and A. Karasu, *Coupled KdV Systems: Recursion Operators of Degree Four* (submitted).

KdV extensions with Painlevé property

Sen-yue Lou[a)]
CCAST (World Laboratory), P. O. Box 8730, Beijing 100080, People's Republic of China and Institute of Mathematical Physics, Ningbo University, Ningbo 315211, People's Republic of China[b)]

(Received 24 September 1996; accepted for publication 22 August 1997)

By means of the conformal invariance (Möbious transformation invariance), the well known KdV equation is extended to (i) a (1+1)-dimensional space-time symmetric form; (ii) two (2+1)-dimensional space isotropic forms; and (iii) general (3+1)-dimensional and N-dimensional forms. The extensions are proven to be integrable under the meaning that they possess the Painlevé property. © *1998 American Institute of Physics.* [S0022-2488(98)02303-2]

I. INTRODUCTION

The soliton theory has attracted much attention from a rather diverse group of scientists (e.g., both physicists and mathematicians) because it is widely applied in many physically significant fields (such as plasma physics, fluids, optics, relative field theory, astrophysics and geophysics, etc.) and demonstrates a deep and fundamental relationship with both modern and core Mathematics (e.g., algebraic and differential geometry, analysis, group theory and chaos). However, there is still a long list of important open problems in soliton theory.[1] Here are three of them:

(i) Because of the real physical space is (3+1)-dimensional, one hopes to find some (3+1)-dimensional integrable models. Unfortunately, except for the ''2+2''-dimensional self-dual Yang–Mills (SDYM) equations, there is no one known integrable model in more than (2+1)-dimensions.

(ii) Usually, the physical space is isotropic. That is to say the physical laws should be invariant under the exchange of space variables. However, except for the Davey–Stewartson (DS) and the Nizhnik–Novikov–Veselov (NNV) equations, almost all the known (2+1)-dimensional integrable models are non-invariant under the exchange of the space variables.

(iii) If we consider relativistic physics, the models should be also invariant under the exchange of the space-time. Nevertheless, most of the famous integrable models are non-symmetric with respect to space-time.

In this paper, we would like to treat these problems by means of the Painlevé analysis. The Painlevé tests, as formulated by Ablowitz, Ramani and Segur (ARS)[2] and Weiss, Tabor and Carnevale (WTC),[3] have been proven to be useful criteria for the identification of completely integrable models.

In the next section, according to the fact that the conformal invariance in solution space (Möbious transformations of fields) plays important roles, we propose a (1+1)-dimensional integrable model with space-time exchange symmetry under the meaning that the model can be changed to a variant form with the Painlevé property.

Using the same idea, two space isotropic (2+1)-dimensional models are studied in section III. In section IV, a general (3+1)-dimensional Schwartz KdV extension is proved to be integrable under the meaning that a variant of the model possesses the Painlevé property. Section V is devoted to discussion of the integrability of an N-dimensional extension. The last section is a summary and discussion.

a)Electronic mail: sylou@fudan.ac.cn
b)Mailing address.

 © 1998 American Institute of Physics

II. A SPACE-TIME SYMMETRIC (1+1)-DIMENSIONAL KdV EXTENSION

It is known that the symmetry study is one of the important works both in physics and in mathematics, especially in the study of the soliton systems because every integrable model possesses infinitely many symmetries.[4–6] Recently we found that the conformal invariance (Möbious transformation invariance) of the Schwartz form is quite important.[7] For instance, for the KdV equation

$$U_t+U_{xxx}-6UU_x=0, \tag{1}$$

three sets of infinitely many nonlocal symmetries and one set of local symmetries and the Darbourx transformation can be derived from the conformal invariance of the Schwartz KdV equation

$$\frac{\phi_t}{\phi_x}+\{\phi;x\}=0, \tag{2}$$

where $\{\phi;x\}\equiv(\phi_{xx}/\phi_x)_x-\frac{1}{2}(\phi_{xx}/\phi_x)^2$ is the Schwartz derivative, while the KdV equation (1) and its Schwartz form (2) are related by[8]

$$U=-\frac{1}{2}\left(\frac{\phi_{xx}}{\phi_x}\right)_x+\frac{1}{4}\left(\frac{\phi_{xx}}{\phi_x}\right)^2.$$

This fact hints to us that if we want to construct an integrable model, the conformal invariant equations may be good candidates. In this paper we would like to construct some KdV type of extensions by using this idea.

First, because of the relativity requirement, we extend (2) to a space time symmetric form

$$\frac{\phi_t}{\phi_x}+\{\phi;x\}+c_1\left(\frac{\phi_x}{\phi_t}+\{\phi;t\}\right)=0, \tag{3}$$

where we have inserted a constant c_1 in such that (3) can be reduced to the usual Schwartz KdV (2) when $c_1=0$. Equation (3) is symmetric for the exchange of the space-time $x\leftrightarrow t$ for $c_1=1$. The model is obviously conformal invariant in the solution space ϕ. In other words, (3) is invariant under the Möbious transformation

$$\phi\rightarrow\frac{a+b\phi}{c+d\phi},\quad ad\neq bc. \tag{4}$$

To check whether the model (3) is integrable or not, we make some transformations first. Using the transformation

$$\phi=\exp F, \tag{5}$$

eq. (3) is changed to

$$\frac{F_t}{F_x}+\{F;x\}-\frac{1}{2}F_x^2+c_1\left(\frac{F_x}{F_t}+\{F;t\}-\frac{1}{2}F_t^2\right)=0. \tag{6}$$

The further transformations

$$F_x=u,\quad F_t=v \tag{7}$$

make (6) become an equation system possessing the Painlevé property,

$$2uv(vu_{xx}+c_1uv_{tt})-3(c_1u^2v_t^2+v^2u_x^2)-u^2v^2(u^2+c_1v^2)+2uv(v^2+c_1u^2)=0, \tag{8}$$

$$u_t=v_x, \tag{9}$$

where eq. (9) comes from the compatibility condition of the transformations (7).

To prove the Painlevé property of the system (8) and (9), we can use the standard WTC approach.[3] According to the WTC paper, we say a model has the Painlevé property if its solutions are single valued about an arbitrary movable singular manifold which is given by $\Phi(x_1,x_2,\ldots,x_n,t)=0$. For the system (8) and (9), we can expand u and v as

$$u=\sum_{j=0}^{\infty} u_j\Phi^{j+\alpha}, \quad v=\sum_{j=0}^{\infty} v_j\Phi^{j+\beta}, \tag{10}$$

with $\{u_j,v_j\}$ being arbitrary functions of (x,t) and $\{u_0\neq 0,v_0\neq 0\}$. By using leading order analysis (i.e., substituting $u\sim u_0\Phi^{\alpha}$, $v\sim v_0\Phi^{\beta}$, into (8), (9)), we have

$$\alpha=-1, \quad \beta=-1, \quad u_0^2=\Phi_x^2, \quad v_0=u_0\Phi_t/\Phi_x. \tag{11}$$

Now substituting (10) with (11) into (8) and (9), we can get the recursion relations on the functions u_j and v_j:

$$(j+1)(j-1)(u_0u_j+c_1v_0v_j)=f \quad (u_i,v_i,i\leqslant j-1), \tag{12}$$

$$(j-1)(v_ju_0-u_jv_0)=u_{j-1,t}-v_{j-1,x}, \tag{13}$$

where f is a complicated function of $\{u_i,v_i,i\leqslant j-1\}$ and the derivatives of the singularity manifold Φ. From (12) and (13), the resonance values of j are given by

$$\begin{vmatrix} (j-1)u_0 & -(j-1)v_0 \\ c_1(j^2-1)v_0 & (j^2-1)u_0 \end{vmatrix}=(j+1)(j-1)^2(u_0^2+c_1v_0^2)=0, \tag{14}$$

i.e.,

$$j=-1,1,1. \tag{15}$$

The resonance at $j=-1$ corresponds to the arbitrary singularity manifold Φ. At two other resonances, j=1, there are two compatibility conditions

$$u_0v_0\left(\Phi_{xx}+c_1\Phi_{tt}-\frac{u_{0x}}{u_0}\Phi_x-c_1\frac{v_{0t}}{v_0}\Phi_t\right)=0, \tag{16}$$

$$u_{0t}-v_{0x}=0, \tag{17}$$

that need to be satisfied. It is clear that (16) and (17) are satisfied identically due to (11). The Painlevé property of the system (8) and (9) is proven. So the equation system (8) and (9) is integrable under the meaning that it possesses the Painlevé property. Now we turn to study (2+1)-dimensional examples.

III. TWO SPACE ISOTROPIC (2+1)-DIMENSIONAL EXTENSIONS

The well known KdV equation has some different extensions in (2+1)-dimensions, say, the Kadomtsev–Petviashvili (KP) equation,[9]

$$(u_t+u_{xxx}-6uu_x)_x+3u_{yy}=0, \tag{18}$$

the Boiti–Leon–Manna–Pempinelli (BLMP) equation[10]

$$u_t+u_{xxx}-3(u\partial_y^{-1}u_x)_x=0, \tag{19}$$

and the Nizhnik–Novikov–Veselov (NNV) equation,[11]

$$u_t+u_{xxx}+u_{yyy}-3(u\partial_y^{-1}u_x)_x-3(u\partial_x^{-1}u_y)_y=0, \tag{20}$$

etc. Actually, the NNV equation is only a combination of two BLMP equations because of $u_{xxx}-3(u\partial_y^{-1}u_x)_x$ and $u_{yyy}-3(u\partial_x^{-1}u_y)_y$ being commutative. Correspondingly, the extensions of the Schwartz KdV to Schwartz KP and Schwartz BLMP equations have the forms[12]

$$\left(\frac{\phi_t}{\phi_x}+\{\phi;x\}+\frac{3\phi_y^2}{2\phi_x^2}\right)_x+3\left(\frac{\phi_y}{\phi_x}\right)_y=0, \tag{21}$$

and

$$\left(\partial_y+\frac{\phi_y}{\phi_x}\partial_x-\left(\frac{\phi_y}{\phi_x}\right)_{xy}-\frac{\phi_y}{\phi_x}\left(\frac{\phi_y}{\phi_x}\right)_{xx}+2\left(\frac{\phi_y}{\phi_x}\right)_x^2\right)$$
$$\times\left(\frac{\phi_t}{\phi_x}+\{\phi;x\}\right)_y-3\{\phi;x\}_y\left(\frac{\phi_y}{\phi_x}\right)_x^2=0, \tag{22}$$

respectively. In this section, we would like to propose two new types of (2+1)-dimensional Schwartz KdV extensions with the space exchange, $x\leftrightarrow y$, invariance.

The first one is

$$\frac{\phi_t}{\phi_x}+\{\phi;x\}+\frac{\phi_t}{\phi_y}+\{\phi;y\}=0, \tag{23}$$

which will be reduced back to the Schwartz KdV (2) for $x=y$. The second extension is

$$\left(\frac{\phi_t}{\phi_x}+\{\phi;x\}+\frac{3\phi_y^2}{2\phi_x^2}\right)_x+3\left(\frac{\phi_y}{\phi_x}\right)_y+a\left(\left(\frac{\phi_t}{\phi_y}+\{\phi;y\}+\frac{3\phi_x^2}{2\phi_y^2}\right)_y+3\left(\frac{\phi_x}{\phi_y}\right)_x\right)=0, \tag{24}$$

which will be reduced back to the Schwartz KP (21) for $a=0$ and to the Schwartz KdV (2) for $x=y$.

Similar to the (1+1)-dimensional case, making the transformation

$$\phi=\exp F \tag{25}$$

for both (23) and (24), we obtain

$$\frac{F_t}{F_x}+\{F;x\}+\frac{F_t}{F_y}+\{F;y\}-(F_x^2+F_y^2)=0 \tag{26}$$

for (23) and

$$\left(\frac{F_t}{F_x}+\{F;x\}+\frac{3F_y^2}{2F_x^2}\right)_x+3\left(\frac{F_y}{F_x}\right)_y$$
$$+a\left(\left(\frac{F_t}{F_y}+\{F;y\}+\frac{3F_x^2}{2F_y^2}\right)_y+3\left(\frac{F_x}{F_y}\right)_x\right)-F_xF_{xx}-aF_yF_{yy}=0 \tag{27}$$

for (24).

Furthermore, making the transformations

$$F_x=u,\quad F_y=v,\quad F_t=w, \tag{28}$$

eq. (26) becomes

$$2uvw(u+v)-3(v^2u_x^2+u^2v_y^2)+2uv(vu_{xx}+uv_{yy})-u^2v^2(u^2+v^2)=0, \tag{29}$$

$$u_t = w_x, \tag{30}$$

$$v_t = w_y \tag{31}$$

after multiplying it by u^2v^2 while eq. (27) is changed to

$$u^2v^2(w_xv + aw_yu + vu_{xxx} + auv_{yyy} + 3vv_y + 3auu_x) - 3v^5u_x - 3au^5v_y$$
$$- uv(wv^2u_x + awu^2v_y + 4v^2u_xu_{xx} + 4au^2v_yv_{yy}) + 3u_x^3v^3$$
$$+ 3av_y^3u^3 - u^3v^3(uu_x + avv_y) = 0, \tag{32}$$

$$u_t = w_x, \tag{33}$$

$$v_t = w_y \tag{34}$$

after multiplying it by u^3v^3.

To prove the Painlevé property of (29)–(31), we expand $\{u,v,w\}$ as

$$u = \sum_{j=0}^{\infty} u_j\Phi^{j+\alpha}, \quad v = \sum_{j=0}^{\infty} v_j\Phi^{j+\beta}, \quad w = \sum_{j=0}^{\infty} w_j\Phi^{j+\gamma} \tag{35}$$

with

$$\alpha = \beta = \gamma = -1, \quad w_0^2 = \Phi_t^2, \quad u_0 = w_0\Phi_x/\Phi_t, \quad v_0 = w_0\Phi_y/\Phi_t, \tag{36}$$

where eq. (36) is given by leading order analysis. Substituting (35) with (36) into the system (29)–(31), we have

$$(j+1)(j-1)(u_0u_j + v_0v_j) = f(u_i, v_i, w_i, i \leqslant j-1), \tag{37}$$

$$(j-1)(w_ju_0 - u_jw_0) = u_{j-1,t} - w_{j-1,x}, \tag{38}$$

$$(j-1)(w_jv_0 - v_jw_0) = v_{j-1,t} - w_{j-1,y}, \tag{39}$$

where f is a function of $\{u_i, v_i, w_i, i \leqslant j-1\}$ and the derivatives of the singularity manifold Φ. It is clear that the resonances occur at $j = -1,1,1,1$. From (37)–(39), we can get three resonance conditions which should be satisfied identically and have the forms

$$u_0v_0(\Phi_{xx} + \Phi_{yy}) - u_0\Phi_yv_{0y} - v_0\Phi_xu_{0x} = 0, \tag{40}$$

$$u_{0t} - w_{0x} = 0, \tag{41}$$

$$v_{0t} - w_{0y} = 0. \tag{42}$$

Obviously, (40)–(42) are true because of (36).

Using the leading order analysis to the system (32)–(34) leads to the same Laurent series (35) with (36). Substituting (35) with (36) into (32)–(34) yields

$$(j+1)(j-1)(j-2)u_0^3v_0^3(u_0^2u_j + av_0^2v_j) = f(u_i, v_i, i \leqslant j-1), \tag{43}$$

$$(j-1)(w_ju_0 - u_jw_0) = u_{j-1,t} - w_{j-1,x}, \tag{44}$$

$$(j-1)(w_jv_0 - v_jw_0) = v_{j-1,t} - w_{j-1,y}, \tag{45}$$

where f is again a function of $\{u_i, v_i, w_i, i \leqslant j-1\}$ and the derivatives of Φ. Now the resonances are located at

$$j=-1,1,1,1,\ 2, \tag{46}$$

and the corresponding compatibility conditions are

$$u_0^2v_0^2(2u_0v_0\Phi_x\Phi_{xx}-v_0\Phi_x^2u_{0x}-u_0^2v_0u_{0x})$$
$$-au_0^2v_0^2(2u_0v_0\Phi_y\Phi_{yy}-u_0\Phi_y^2v_{0y}-v_0^2u_0v_{0y})=0, \tag{47}$$

$$u_0v_0^3(u_{0x}^2\Phi_x+u_0^2\Phi_{xxx}+3u_0v_0\Phi_y-u_0u_{0x}\Phi_{xx}-u_0u_{0xx}\Phi_x-3v_0^2\Phi_x)$$
$$-av_0u_0^3(v_{0y}^2\Phi_y+v_0^2\Phi_{yyy}+3u_0v_0\Phi_x-v_0v_{0y}\Phi_{yy}-v_0v_{0yy}\Phi_y-3u_0^2\Phi_y)=0, \tag{48}$$

$$u_{0t}-w_{0x}=0, \tag{49}$$

$$v_{0t}-w_{0y}=0. \tag{50}$$

The compatibility conditions (47) and (48) come from (43) for $j=1$ and $j=2$, respectively. It is also straightforward to see that the conditions (47)–(50) are satisfied identically because of eq. (36).

From the above calculations, we know that both systems (29)–(31) and (32)–(34) possess the Painlevé property. Then two space symmetric (2+1)-dimensional Schwartz KdV extensions (23) and (24) are integrable under the meaning that they can be changed to the variant forms with the Painlevé property.

The excellent property of the (1+1)-dimensional and (2+1)-dimensional Schwartz equations leads us to construct integrable models in higher dimensions.

IV. A GENERALIZED (3+1)-DIMENSIONAL KdV EXTENSION

As mentioned in the Introduction, because the real physical space-time is (3+1)-dimensional, we discuss a (3+1)-dimensional extension of the Schwartz KdV equation before discussing more general higher dimensional extension.

According to discussions of the last two sections, one may guess that the following (3+1)-dimensional system,

$$u^2v^2w^2g^2\left(a_1\frac{u_{xx}}{u}+a_2\frac{v_{yy}}{v}+a_3\frac{w_{zz}}{w}+a_4\frac{g_{tt}}{g}-\frac{3}{2}\left(a_1\frac{u_x^2}{u^2}+a_2\frac{v_y^2}{v^2}+a_3\frac{w_z^2}{w^2}+a_4\frac{g_t^2}{g^2}\right)\right.$$
$$-\frac{1}{2}(a_1u^2+a_2v^2+a_3w^2+a_4g^2)+a_5\frac{g}{u}+a_6\frac{u}{g}+a_7\frac{u}{v}+a_8\frac{v}{u}+a_9$$
$$\left.\times\left(\frac{u}{w}+a_{10}\frac{w}{u}+a_{11}\frac{v}{w}+a_{12}\frac{w}{v}+a_{13}\frac{g}{w}+a_{14}\frac{w}{g}+a_{15}\frac{g}{v}+a_{16}\frac{v}{g}\right)=0, \tag{51}$$

$$u_t=g_x, \tag{52}$$

$$v_t=g_y, \tag{53}$$

$$w_t=g_z, \tag{54}$$

with arbitrary constants a_i, $i=1,2,\ldots,16$, may possess the Painlevé property. The system (51)–(54) can be simplified to a single Schwartz form

$$a_1\{\phi;x\}+a_2\{\phi;y\}+a_3\{\phi;z\}+a_4\{\phi;t\}+a_5\frac{\phi_t}{\phi_x}+a_6\frac{\phi_x}{\phi_t}+a_7\frac{\phi_x}{\phi_y}+a_8\frac{\phi_y}{\phi_x}+a_9\frac{\phi_x}{\phi_z}+a_{10}\frac{\phi_z}{\phi_x}$$
$$+a_{11}\frac{\phi_y}{\phi_z}+a_{12}\frac{\phi_z}{\phi_y}+a_{13}\frac{\phi_t}{\phi_z}+a_{14}\frac{\phi_z}{\phi_t}+a_{15}\frac{\phi_t}{\phi_y}+a_{16}\frac{\phi_y}{\phi_t}=0, \tag{55}$$

with $\phi=\exp F$, $u=F_x$, $v=F_y$, $w=F_z$, and $g=F_t$.

To check the Painlevé property of the system (51)–(54), we use the leading order analysis first again. Substituting

$$\{u,v,w,g\}\sim\{u_0\Phi^{\alpha},v_0\Phi^{\beta},w_0\Phi^{\gamma},g_0\Phi^{\delta}\} \tag{56}$$

to (51)–(54) leads to

$$\alpha=\beta=\gamma=\delta=-1,\quad g_0^2=\Phi_t^2,\quad u_0=g_0\frac{\Phi_x}{\Phi_t},\quad v_0=g_0\frac{\Phi_y}{\Phi_t},\quad w_0=g_0\frac{\Phi_z}{\Phi_t}. \tag{57}$$

Then the Laurent expansions of $\{u,v,w,g\}$ have the forms

$$\begin{aligned} u&=\sum_{j=0}^{\infty}u_j\Phi^{j-1},\quad v=\sum_{j=0}^{\infty}v_j\Phi^{j-1},\\ w&=\sum_{j=0}^{\infty}w_j\Phi^{j-1},\quad g=\sum_{j=0}^{\infty}g_j\Phi^{j-1}. \end{aligned} \tag{58}$$

Using (58) with (57) in (51)–(54), we get

$$(j+1)(j-1)(a_1u_0u_j+a_2v_0v_j+a_3w_0w_j+a_4g_0g_j)=f(u_i,v_i,w_i,g_i,i\leqslant j-1), \tag{59}$$

$$(j-1)(g_ju_0-u_jg_0)=u_{j-1,t}-g_{j-1,x}, \tag{60}$$

$$(j-1)(g_jv_0-v_jg_0)=v_{j-1,t}-g_{j-1,y}, \tag{61}$$

$$(j-1)(g_jw_0-w_jg_0)=w_{j-1,t}-g_{j-1,z} \tag{62}$$

with the resonances

$$j=-1,1,1,1. \tag{63}$$

The remaining discussions are similar to that of the last two sections. Four compatibility conditions

$$\begin{aligned} &u_0v_0w_0g_0(a_1\Phi_{xx}+a_2\Phi_{yy}+a_3\Phi_{zz}+a_4\Phi_{tt})-a_1w_0v_0g_0\Phi_xu_{0x}\\ &-a_2u_0w_0g_0\Phi_yv_{0y}-a_3u_0v_0g_0\Phi_zw_{0z}-a_4u_0v_0w_0\Phi_tg_{0t}=0, \end{aligned} \tag{64}$$

$$u_{0t}-g_{0x}=0, \tag{65}$$

$$v_{0t}-g_{0y}=0, \tag{66}$$

and

$$w_{0t}-g_{0z}=0 \tag{67}$$

are satisfied naturally thanks to (57). So the system (51)–(54) possesses Painlevé property and the generalized (3+1)-dimensional Schwartz system (55) is integrable because it can be transformed to some variants with Painlevé property.

Some special cases of (55) are worth mentioning. (i) When we take the constants $a_i=0$ except $a_1=a_4=a_5=a_6=1$, the model (55) reduces back to the (1+1)-dimensional space-time symmetric model given in section II. (ii) If we take $a_1=a_2=a_5=a_{15}=1$ and all other a_i being zero, we get the (2+1)-dimensional space isotropic model (23). (iii) If we take all the constants as one, then the model (55) becomes a (3+1)-dimensional relativistic integrable model, i.e., the model is space-time exchange invariant.

Actually, one can easily obtain various other types of $(3+1)$-dimensional models by means of the conformal invariant quantities (which are invariant under the Möbious transformation). However, we prefer to give a possible integrable generalization in arbitrary dimensions rather than to discuss this problem further.

V. AN *N*-DIMENSIONAL KdV EXTENSION

Generally, an N-dimensional conformal invariant model may have the form

$$G\left(\{\phi;x_i\},\frac{\phi_{x_i}}{\phi_{x_j}},i,j,=1,2,\ldots,N\right)=0, \tag{68}$$

where G is an arbitrary functions of the conformal invariants $\{\phi;x_i\},\phi_{x_i}/\phi_{x_j},(i,j,=1,2,\ldots,N)=0$ and their derivatives. We believe that there are many types of special function G which may be changed to a variant form with the Painlevé property. In this section, we write down only one special G and change it to a form with the Painlevé property.

Using the same idea as in lower dimensions discussed in sections II–IV, we may prove the following special case of (68),

$$\sum_{k=0}^{N} a_k\{\phi;x_k\}+g\left(\frac{\phi_{x_i}}{\phi_{x_j}},i,j,=1,2,\ldots,N\right)=0, \tag{69}$$

where g is an arbitrary polynomial functions of $\phi_{x_i}/\phi_{x_j}(i,j,=1,2,\ldots,N)$, may be transformed to a form possessing Painlevé property. Taking the transformations

$$\phi=\exp F,\quad u_i=F_{x_i},\quad i=1,2,\ldots N, \tag{70}$$

eq. (69) is changed to

$$\left(\prod_{k=1}^{N} u_k^{m_k}\right)\left[\sum_{k=1}^{N} a_k\left(n;\frac{u_{kx_kx_k}}{u_k}-\frac{3u_{kx_k}^2}{2u_k^2}-\frac{1}{2}u_k^2\right)+g\left(\frac{u_i}{u_j},i,j=1,2,\ldots,N\right)\right]=0, \tag{71}$$

$$u_{1x_i}=u_{ix_i},\quad i=1,2,\ldots,N, \tag{72}$$

where $m_j\geqslant 2$ are positive integers and are fixed such that $(\Pi_{k=1}^{N} u_k^{m_k})g(u_i/u_j,\ i,j=1,2,\ldots,N)$ is only a polynomial function of $u_i,\ i=1,2,\ldots,N$.

To prove the Painlevé property of (71) and (72), we expand u_i as

$$u_i=\sum_{j=0}^{\infty} u_{ij}\Phi^{j+\alpha_i},\quad i=1,2,\ldots,N, \tag{73}$$

where α_i and u_{i0} are fixed by leading order analysis. The result is

$$\alpha_i=-1,\quad u_{i0}=\Phi_{x_i},\quad i=1,2,\ldots,N. \tag{74}$$

Substituting eq. (73) with (74) into (71) and (72), we obtain

$$(j+1)(j-1)\sum_{k=1}^{N} a_k u_{k0}u_{kj}=f(u_{ki},k=1,2,\ldots,N,i\leqslant j-1), \tag{75}$$

$$(j-1)(u_{1j}u_{k0}-u_{kj}u_{10})=u_{k(j-1),t}-u_{1(j-1),x_k},\ \ k=2,3,\ldots,N \tag{76}$$

with $N+1$ resonances,

$$j=-1,\underbrace{1,1,\ldots,1}_{N}, \tag{77}$$

where f is a very complicated function of u_{ki}, $k=1,2,\ldots,N, i\leqslant j-1$. The resonances $j=-1$ corresponds to the arbitrary singular manifold Φ. For other resonances, $j=1$, after finishing some tedious calculations, we get N compatibility conditions

$$\sum_{k=0}^{N} a_k\left(\Phi_{x_kx_k}-\frac{u_{k0,x_k}}{u_{k0}}\Phi_{x_k}\right)=0, \tag{78}$$

$$u_{k0,t}-u_{10,x_k}=0, \quad (k=2,3,\ldots,N). \tag{79}$$

Obviously the resonance conditions (78) and (79) are satisfied identically because of eq. (74). Then equation (69) are integrable under the meaning that its variant (71) and (72) possesses the Painlevé property.

VI. SUMMARY AND DISCUSSION

According to the fact that many integrable models can be transformed to the conformal invariant forms (Schwartz forms) and various integrable properties such as the Darbourx transformation, infinitely many symmetries, etc., can be derived from the conformal invariance, we believe that one may get various integrable models by extending the known integrable Schwartz equations. Starting from this idea, we have constructed some significant KdV extensions. Some of them are isotropic in space and some of them are relativistic (i.e., space-time symmetric). Especially the (Schwartz) KdV equation has been extended to real (3+1)-dimensional physical space. The Painlevé properties of these extensions are all proved by using the standard WTC approach.

Though the extensions are all proven to be integrable under the meanings that they possess the Painlevé property and the Schwartz forms with conformal invariance, there are still many important problems worth further study:

(i) Usually, one would say a model is integrable under some different meanings. For instance, a model is integrable under the meanings that it can be solved by the inverse scattering transformation, it possesses the Painlevé property, it possesses Lax Pairs, it possesses bi-Hamiltonian structure and infinitely many symmetries and conservation laws and so on. Are the models obtained here integrable under other meanings, say, under the meaning that they can be solved by inverse scattering transformation?

(ii) Though the models are obtained only from the consideration of the integrability, we believe that they may be found also from the real physical systems because the real physical space-time is isotropic, relativistic and (3+1)-dimensional. The most possible physical systems which can be used to derive these models are the SDYM equations because of the considerably increasing the validity of Ward's conjecture[13,1]:
''many (and perhaps all?) of the ordinary and partial differential equations that are regarded as being integrable or solvable may be obtained from the self-dual gauge field equations (or its generalizations) by reduction.''
Almost all the known integrable models, say, the KdV, KP, DS, N-wave interaction equations, nonlinear Schrödinger, chiral field equation, Toda equations and sine-Gordon equation, have been obtained from the reduction of the SDYM equations.[1] How do we get the model obtained in this paper from concrete physical systems, say, SDYM equation?

(iii) Generally, one can construct an N-dimensional integrable model (which possesses the Painlevé property) quite freely by means of the conformal invariants. For instance, an arbitrary polynomial function of some conformal invariants has been included in our special example (69). Though not all the equations in the Schwartz form are integrable (can be changed to a form with the Painlevé property), it is still significant to ask: What kind of

conformal invariant form (function G in (68)) may be changed to a form with Painlevé property?

ACKNOWLEDGMENTS

The work was supported by the National Nature Science Foundation of China and the Nature Science foundation of Zhejiang province in China. I thank Professors Q-p Liu, X-b Hu and G-j Ni for their helpful discussions.

[1] M. J. Ablowitz and P. A. Clarkson, *Solitons Nonlinear Evolution Equations and Inverse Scattering*, London Mathematical Society Lecture Note Series 149 (Cambridge U.P., Cambridge, 1991).
[2] M. J. Ablowitz, A. Ramani, and H. Segur, Lett. Nuovo Cimento **23**, 333 (1978).
[3] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[4] P. J. Olver, *Application of Lie Group to differential equations*, Graduate texts in Mathematics (Springer, Berlin, 1986); Q. W. Bluman and S. Kumei, *Symmetries and differential equations* (Springer, Berlin, 1989).
[5] B. Fuchssteiner, Prog. Theor. Phys. **70**, 1508 (1983); P. M. Santini and A. S. Fokas, Commun. Math. Phys. **115**, 375 (1988); A. S. Fokas and P. M. Santini, Commun. Math. Phys. **116**, 449 (1989); J. Math. Phys. **29**, 1606 (1988); H. H. Chen, Y. C. Lee, and J. E. Lin, Phys. Lett. **91**, 381 (1983); Physica D **26**, 165 (1987); D. David, N. Kamran, D. Levi, and P. Winternitz, J. Math. Phys. **27**, 1255 (1986).
[6] S-y Lou, Phys. Lett. B **302**, 261 (1993); Phys. Lett. A **175**, 23 (1993); **181**, 13 (1993); J. Math. Phys. **35**, 2336 (1994); **35**, 2390 (1994); Phys. Rev. Lett. **71**, 4099 (1993); J. Phys. A **26**, 4387 (1993); **27**, 3235 (1994); J. Math. Phys. **35**, 1755 (1994).
[7] S-y Lou, J. Phys. A **30**, 4083 (1997); **30**, 7259 (1997).
[8] M. C. Nucci, J. Phys. A **22**, 2897 (1989).
[9] B. B. Kadomtsev and V. I. Patviashvili, Sov. Phys. Dokl. **15**, 539 (1970).
[10] M. Boiti, J. Jp. Leon, M. Manna, and F. Pempinelli, Inverse Probl. **2**, 271 (1986).
[11] L. P. Nizhnik, Sov. Phys. Dokl. **25**, 707 (1980); A. P. Veselov and S. P. Novikov, Sov. Math. Dolk. **30**, 588 (1984); **30**, 750 (1984); S. P. Novikov and A. P. Veselov, Physica D **18**, 267 (1986).
[12] Q-x Wu and S-y Lou, Commun. Theor. Phys. **27**, 109 (1997).
[13] R. S. Ward, Philos. Trans. R. Soc. London, Ser. A **315**, 451 (1985).

On the simplest (2+1) dimensional integrable spin systems and their equivalent nonlinear Schrödinger equations

R. Myrzakulov[a)]
Physical Technical Institute, National Academy of Sciences, Alma-Ata-480 082, Kazakstan and Center for Nonlinear Problems, P. O. Box 30, 480035 Alma-Ata-35, Kazakstan

S. Vijayalakshmi
Centre for Nonlinear Dynamics, Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, India

R. N. Syzdykova
Center for Nonlinear Problems, P. O. Box 30, 480035 Alma-Ata-35, Kazakstan

M. Lakshmanan[b)]
Centre for Nonlinear Dynamics, Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, India

(Received 29 September 1997; accepted for publication 12 November 1997)

Using a moving space curve formalism, geometrical as well as gauge equivalence between a (2+1) dimensional spin equation (M-I equation) and the (2+1) dimensional nonlinear Schrödinger equation (NLSE) originally discovered by Calogero, discussed then by Zakharov and recently rederived by Strachan, have been established. A compatible set of three linear equations are obtained and integrals of motion are discussed. Through stereographic projection, the M-I equation has been bilinearized and different types of solutions such as line and curved solitons, breaking solitons, induced dromions, and domain wall type solutions are presented. Breaking soliton solutions of (2+1) dimensional NLSE have also been reported. Generalizations of the above spin equation are discussed. © *1998 American Institute of Physics.* [S0022-2488(98)00504-0]

I. INTRODUCTION

The two dimensional sigma-models with nontrivial topological structures are known to play a useful role in modern field theory. This set of models constitute a laboratory for studying two dimensional analogues of elementary particles within the framework of classical field theory.[1] In this context, important case studies concern with the existence of localized coherent structures (dromions, lumps, etc.) and other types of soliton-like solutions of some classical nonlinear field models such as, for example, the Ishimori and the other (2+1) dimensional spin systems.[1–3]

Generally speaking, in (2+1) dimensions we have a number of remarkable properties, which may not exist in (1+1) dimensionsal counterparts. For example,

1. These equations possess the so called localized coherent structures (such as dromions, lumps and so on), which may undergo both elastic and inelastic scattering depending upon the initial conditions.
2. The corresponding spectral parameter (eigenvalue) of the associated linear problem (the Lax representation) can be dependent on the t (time) and y (space) variables and satisfy even nonlinear equations (see, for instance, Refs. 3 and 4). As a consequence, the original soliton equation can have breaking solutions.[4–6] Besides, in this case, for finding solutions one even needs to use nonisospectral Inverse Scattering Transform (IST).[5]
3. Each integrable (1+1) dimensional equation may admit several integrable (and nonintegrable) extensions[2,3] in (2+1) dimensions. For example, the KdV equation has integrable extensions such

[a)]Electronic mail: cnlpmyra@satsun.sci.kz
[b)]Electronic mail: lakshman@bdu.ernet.in

 © 1998 American Institute of Physics

as KP, NNV and breaking soliton equations.[6]

4. Regarding the subject of the present paper, many of the (2+1) dimensional spin equations possess the topological invariant-with the so called topological charge

$$Q=\frac{1}{4\pi}\int dxdy\vec{S}\cdot\vec{S}_x\wedge\vec{S}_y, \tag{1}$$

and their solutions are classified by the integer values of Q_N, $N=0,\pm1,\pm2,\ldots$ Here, $\vec{S}=(S_1,S_2,S_3)$ and $|\vec{S}|^2=S_1^2+S_2^2+S_3^2=1$ and $\vec{S}_x=(\partial\vec{S}/\partial x)$, $\vec{S}_y=(\partial\vec{S}/\partial y)$. This property can be realised, for example, in the (2+1) dimensional analogues of the well known (1+1) dimensional Heisenberg ferromagnet model (or the (1+1) dimensional isotropic Landau–Lifshitz equation (LLE))[7–9]

$$\vec{S}_t=\vec{S}\wedge\vec{S}_{xx}. \tag{2}$$

The Heisenberg ferromagnetic spin equation (2) possesses many useful (2+1) dimensional extensions: Some of them are as follows:

(a) The Ishimori equation[10]

$$\vec{S}_t=\vec{S}\wedge(\vec{S}_{xx}+\epsilon^2\vec{S}_{yy})+u_x\vec{S}_y+u_y\vec{S}_x, \tag{3a}$$

$$u_{xx}-\epsilon^2u_{yy}=-2\epsilon^2\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_y). \tag{3b}$$

(b) The M-I equation[5]

$$\vec{S}_t=(\vec{S}\wedge\vec{S}_y+u\vec{S})_x, \tag{4a}$$

$$u_x=-\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_y). \tag{4b}$$

(c) The (2+1) dimensional isotropic LLE[11]

$$\vec{S}_t=\vec{S}\wedge(\vec{S}_{xx}+\vec{S}_{yy}). \tag{5}$$

Here u is a scalar function and $\epsilon^2=\pm1$. It turns out that Eqs. (3) and (4) are integrable, while Eq. (5) is apparently nonintegrable. All the three Eqs. (3)–(5) describe the nonlinear dynamics of the classical spin systems in the plane and in the (1+1) dimensional case reduce to one and the same Eq. (2). Properties of Eqs. (3) are relatively well studied (see, for example, Refs. 3 and 10). In this paper we wish to concentrate on studying the M-I equation (4). In particular, we wish to identify the geometrically equivalent (2+1) dimensional NLSE for (4) and study the nature of the nonlinear excitations admitted by the spin system.

The paper is organized as follows. In Sec. II we briefly review some necessary informations about (4) relevant to this paper. The geometrical and gauge equivalent counterpart of Eq. (4) is constructed in Sec. III and Sec. IV, respectively. Integrals of motion are discussed in Sec. V. In Sec. VI the Hirota bilinear form of Eq. (4) is derived. The solitons (line and curved), domain walls and dromion-like solutions as well as their breaking analogues are obtained in Sec. VII and breaking solitons and dromions for its equivalent counterpart are obtained in Sec. VIII. In Sec. IX we comment on the possible further extensions and we conclude in Sec. X.

II. THE M-I EQUATION

The Lax representation of Eq. (4) can be shown to take the form[5]

$$\phi_{1x}=U_1\phi_1, \tag{6a}$$

$$\phi_{1t}=V_1\phi_1+\lambda\phi_{1y}, \tag{6b}$$

with

$$U_1=\frac{i\lambda}{2}S, \tag{7a}$$

$$V_1=\frac{\lambda}{4}([S,S_y]+2iuS). \tag{7b}$$

Here

$$S=\begin{pmatrix} S_3 & rS^- \\ rS^+ & -S_3 \end{pmatrix}, \tag{8}$$

$S^{\pm}=S_1\pm iS_2$ and λ is the eigenvalue parameter, which satisfies the following nonlinear equation

$$\lambda_t=\lambda\lambda_y. \tag{9}$$

Hence, using the compatibility of Eqs. (6a) and (6b), we get

$$iS_t=([S,S_y]+2iuS)_x, \tag{10a}$$

$$u_x=-\frac{1}{2i}\mathrm{tr}(SS_xS_y), \tag{10b}$$

where tr denotes the trace of the matrix. This system is obviously the matrix form of Eq. (4).

Thus, for solving Eq. (4), we must use the nonisospectral extension of IST.[5] As a consequence, Eq. (4) admits besides the usual solutions, corresponding to constant solution of Eq. (9), breaking analogues related to the t and y dependence of λ. Note that Eq. (9) itself is the compatibility condition of the following system of the linear equations,

$$f_x=\frac{i\lambda}{2}f, \tag{11a}$$

$$f_t=\lambda f_y. \tag{11b}$$

Another useful form of Eq. (4) can be obtained by using the complex (stereographic) variable $\omega(x,y,t)$ defined through the relations

$$S^+=S_1+iS_2=\frac{2\omega}{1+|\omega|^2}, \quad S_3=\frac{1-|\omega|^2}{1+|\omega|^2}. \tag{12}$$

In this case, Eq. (4) takes the form

$$i(\omega_t-u\omega_x)+\omega_{xy}-\frac{2\omega^*\omega_x\omega_y}{1+|\omega|^2}=0, \tag{13a}$$

$$u_x+\frac{2i(\omega_x\omega_y^*-\omega_x^*\omega_y)}{(1+|\omega|^2)^2}=0. \tag{13b}$$

We will use this form also frequently in our following analysis.

III. L-EQUIVALENT COUNTERPART

It is well known that in (1+1) dimensions there exists geometrical equivalence between spin systems and nonlinear Schrödinger type equations,[7,12] which in Ref. 5 was called the Lakshmanan equivalence or shortly the L-equivalence (see also Refs. 12–14). In Refs. 5,13–15 a (2+1) dimensional generalization of the L-equivalence was presented. In this section we find the L-equivalent counterpart of Eq. (4). For this purpose, we will extend the geometrical method appli-

cable to (1+1) dimensional systems suitably to the (2+1) dimensional case. We now associate a moving space curve parametrised by the arclength x, and endowed with an additional coordinate y, with the spin system.[12,16] Then the Serret–Frenet equation associated with the curve has the form

$$\vec{e}_{ix}=\vec{D}\wedge\vec{e}_i, \tag{14}$$

where

$$\vec{D}=\tau\vec{e}_1+\kappa\vec{e}_3 \tag{15}$$

and $\vec{e}_i$'s, $i=1,2,3$, form the orthogonal trihedral.

Mapping the spin variable on the unit tangent vector

$$\vec{S}(x,y,t)=\vec{e}_1, \tag{16}$$

the curvature and the torsion are given by

$$\kappa(x,y,t)=(\vec{S}_x^2)^{1/2}, \tag{17a}$$

$$\tau(x,y,t)=\kappa^{-2}\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_{xx}). \tag{17b}$$

Due to the orthonormality nature of the trihedral, $\vec{e}_{it}.\vec{e}_i=0$, $\vec{e}_{iy}.\vec{e}_i=0$, $i,j=1,2,3$ and using the compatibility condition $\vec{e}_{ixy}=\vec{e}_{iyx}$, we find the equation for the y-part

$$\vec{e}_{iy}=\vec{\gamma}\wedge\vec{e}_i, \tag{18}$$

where $\vec{\gamma}=(\gamma_1,\gamma_2,\gamma_3)$ and

$$\gamma_1=u+\partial_x^{-1}\tau_y, \tag{19a}$$

$$\gamma_2=\frac{u_x}{\kappa}, \tag{19b}$$

$$\gamma_3=\partial_x^{-1}\left(\kappa_y-\frac{\tau u_x}{\kappa}\right). \tag{19c}$$

Alternatively the trihedral $\vec{e}_i(x,y,t)$, $i=1,2,3$ could be thought of as defining a suitable surface in E^3 (so that Eqs. (14) and (18) represent the Gauss–Weingarten equations in orthogonal coordinates and that the compatibility condition $\vec{e}_{ixy}=\vec{e}_{iyx}$ gives rise to the Codazzi–Mainardi equations), which is then set in motion. Now, from Eq. (4) and using Eqs. (14) and (18), we can easily find the time evolution of the trihedral as

$$\vec{e}_{it}=\vec{\Omega}\wedge\vec{e}_i, \tag{20}$$

with

$$\vec{\Omega}=(\omega_1,\omega_2,\omega_3)=\left(\frac{\kappa_{xy}}{\kappa}-\tau\partial_x^{-1}\tau_y,-\kappa_y,-\kappa\partial_x^{-1}\tau_y\right). \tag{21}$$

Ultimately, the compatibility condition $\vec{e}_{ixt}=\vec{e}_{itx}$, which is also consistent with the relation $\vec{e}_{iyt}=\vec{e}_{ity}$, $i=1,2,3$ yields the following evolution equations for the curvature and torsion,

$$\kappa_t=-(\kappa\tau)_y-\kappa_x\partial_x^{-1}\tau_y, \tag{22a}$$

$$\tau_t=\left[\frac{\kappa_{xy}}{\kappa}-\tau\partial_x^{-1}\tau_y\right]_x+\kappa\kappa_y\,. \tag{22b}$$

On making the complex transformation[7]

$$\psi(x,y,t)=\frac{\kappa(x,y,t)}{2}\exp\left[-i\int_{-\infty}^{x}\tau(x',y,t)dx'\right], \tag{23}$$

the set of Eq. (22) reduces to the following (2+1) dimensional NLSE

$$i\psi_t=\psi_{xy}+r^2V\psi, \tag{24a}$$

$$V_x=2\partial_y|\psi|^2. \tag{24b}$$

Here, $r^2=+1$, that is, we have the attractive type NLSE (the case $r^2=-1$ corresponds to the repulsive case). Equation (24) belongs to the class of equations discovered by Calogero[17] and then discussed by Zakharov[18] and recently rederived by Strachan (for $r^2=+1$).[19] Its Painlevé property and some exact solutions were also obtained.[20] N-soliton solutions of Eq. (24) for both the cases ($r^2=\pm1$) can be found in Ref. 21. Thus, we have proved that Eq. (24) is equivalent to Eq. (4) in the geometrical sense.

Linearization

Introducing now the complex variable corresponding to an orthogonal rotation

$$z_l=\frac{e_{2l}+ie_{3l}}{1-e_{1l}},\quad e_{1l}^2+e_{2l}^2+e_{3l}^2=1,\quad l=1,2,3$$

the spatial and temporal evolution of the trihedral [Eqs. (14), (18) and (20)] can be rewritten as a set of the following three Riccati equations:

$$z_{lx}=-i\tau z_l+\frac{\kappa}{2}[1+z_l^2], \tag{25a}$$

$$z_{ly}=-i\gamma_1 z_l+\frac{1}{2}[\gamma_3+i\gamma_2]z_l^2+\frac{1}{2}[\gamma_3-i\gamma_2], \tag{25b}$$

$$z_{lt}=-i\omega_1 z_l+\frac{1}{2}[\omega_3+i\omega_2]z_l^2+\frac{1}{2}[\omega_3-i\omega_2]. \tag{25c}$$

It is easy to check that Eq. (25a) is equivalent to the Serret–Frenet Eqs. (14), (25b) is equivalent to the y-variation of the trihedral Eq. (18), while the temporal evolution (25c) is equivalent to (20).

Further introducing the transformation

$$z_l=\frac{v_2}{v_1}, \tag{26}$$

Eq. (25) can be written as a system of three coupled two component first order equations,

$$\begin{pmatrix}v_{1x}\\ v_{2x}\end{pmatrix}=\begin{pmatrix}\frac{i\tau}{2} & \frac{-\kappa}{2}\\ \frac{\kappa}{2} & \frac{-i\tau}{2}\end{pmatrix}\begin{pmatrix}v_1\\ v_2\end{pmatrix}, \tag{27a}$$

$$\begin{pmatrix} v_{1y} \\ v_{2y} \end{pmatrix} = \begin{pmatrix} \frac{i\gamma_1}{2} & \frac{-1}{2}(\gamma_3 + i\gamma_2) \\ \frac{1}{2}(\gamma_3 - i\gamma_2) & \frac{-i\gamma_1}{2} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{27b}$$

$$\begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix} = \begin{pmatrix} \frac{i\omega_1}{2} & \frac{-1}{2}(\omega_3 + i\omega_2) \\ \frac{1}{2}(\omega_3 - i\omega_2) & \frac{-i\omega_1}{2} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{27c}$$

Once again one can check that the compatibility of the three sets of Eqs. (27) gives rise to the evolution equation (22).

IV. GAUGE EQUIVALENT COUNTERPART

Next, it is also of interest to note that Eqs. (4) and (12) are gauge equivalent to each other in the sense of Zakharov and Takhtajan.[22] To obtain the gauge equivalent counterpart of Eq. (4), in the usual way we consider the following gauge transformation

$$\phi_1 = g^{-1}\phi_2, \tag{28}$$

where $g(x,y,t)$ and $\phi_2(x,y,t,\lambda)$ are arbitrary (2×2) matrix functions of the type defined on a compact manifold $S^2 = SU(2)/U(1)$. Substituting Eq. (28) into Eq. (6), after some algebra we get the following system of linear equations for ϕ_2,

$$\phi_{2x} = U_2\phi_2, \tag{29a}$$

$$\phi_{2t} = V_2\phi_2 + \lambda\phi_{2y} \tag{29b}$$

with

$$U_2 = \frac{i\lambda}{2}\sigma_3 + G, \quad G = \begin{pmatrix} 0 & \phi \\ -r^2\phi^* & 0 \end{pmatrix}, \quad r^2 = +1, \tag{30a}$$

$$V_2 = -i\sigma_3\left(\frac{VI}{2} + G_y\right), \quad I = \mathrm{diag}(1,1), \tag{30b}$$

$$V = 2\partial_x^{-1}\partial_y(|\psi|^2). \tag{30c}$$

The compatibility condition of Eq. (29) along with (9) becomes (24), that is, Eq. (4) and Eq. (24) are gauge equivalent to each other. The above transformation is in fact reversible and we can similarly prove that Eq. (24) is gauge equivalent to Eq. (4). It is also of interest to note that the set of linear Eqs. (27) can be recast in the form (29) after suitable transformations.

Next, we present some important formulae which are just consequences of the geometrical/gauge equivalence of Eqs. (4) and (24). We have

$$\mathrm{tr}(S_x^2) = 8|\psi|^2 = 2\vec{S}_x^2. \tag{31a}$$

In a similar manner we find that

$$-2i\vec{S}\cdot(\vec{S}_x \wedge \vec{S}_{xx}) = \mathrm{tr}(SS_xS_{xx}) = 4(\psi^*\psi_x - \psi\psi_x^*). \tag{31b}$$

These relations are obviously equivalent to Eq. (23). One notes that these are of the same form as in the case of $(1+1)$ dimensional Heisenberg spin chain.[7,8]

V. INTEGRALS OF MOTION

The spin Eq. (4) allows an infinite number of integrals of motion as a consequence of integrability. These integrals can be consructed using for example the Lax representation (6). However some of the integrals can be constructed using that L-equivalence property. Now from (22a), it follows that

$$(\kappa^2)_t=[-\kappa^2\partial_x^{-1}\tau_y]_x+[-k^2\tau]_y. \tag{32a}$$

Hence we get the first integral

$$K_1=\int \kappa^2 dxdy. \tag{32b}$$

Similarly (22b) leads to

$$K_2=\int \kappa^2\tau dxdy. \tag{33}$$

In terms of the spin vector the corresponding two conservation laws are

$$(\vec{S}_x^2)_t+\partial_x\left[\vec{S}_x^2\partial_x^{-1}\left(\frac{\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}}{\vec{S}_x^2}\right)_y\right]+\partial_y[\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}]=0, \tag{34}$$

$$[\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}]_t+\partial_x\left[\frac{(\vec{S}_x^2)_x(\vec{S}_x^2)_y}{4\vec{S}_x^2}+\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}\partial_x^{-1}\left(\frac{\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}}{\vec{S}_x^2}\right)_y\right]$$
$$+\partial_y\left\{\frac{(\vec{S}_x^2)_x^2}{4\vec{S}_x^2}+\frac{(\vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx})^2}{\vec{S}_x^2}-\frac{(\vec{S}_x^2)_{xx}}{2}-\frac{\vec{S}_x^4}{4}\right\}=0. \tag{35}$$

Note that these integrals have the same forms as in the (1+1) dimensional case. More interesting integrals of purely (2+1) dimensional nature can be obtained from the condition

$$\vec{e}_{jxy}=\vec{e}_{jyx}. \tag{36}$$

Making use of the various relations in Sec. III and after some algebra we have

$$(-\kappa\gamma_2)_t=(-\gamma_{1t})_x+(\tau_t)_y, \tag{37}$$

$$(-\tau\gamma_2)_t=(\gamma_{3t})_x+(-\kappa_t)_y. \tag{38}$$

These equations give two integrals

$$K_3=\int (-\kappa\gamma_2)dxdy, \tag{39}$$

and

$$K_4=\int (-\tau\gamma_2)dxdy. \tag{40}$$

In terms of the spin vector $\vec{S}$ these integrals of motion take the forms

$$K_1=\int (\vec{S}_x^2)dxdy, \tag{41}$$

$$K_2=\int \vec{S}\cdot\vec{S}_x\wedge\vec{S}_{xx}dxdy, \tag{42}$$

$$K_3=\int \vec{S}\cdot(\vec{S}_x\wedge\vec{S}_y)dxdy \tag{43}$$

and

$$K_4=\int \frac{[\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_y)][\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_{xx})]}{(\vec{S}_x^2)^{3/2}}dxdy. \tag{44}$$

Note that K_3 is the topological charge given by Eq. (1) to within a constant. One can proceed to find the other integrals of motion using the eigenvalue problem (29).

VI. HIROTA BILINEAR FORM

In order to solve the Cauchy initial value problem of Eq. (4), it will be of interest to investigate the system in the framework of the inverse spectral transform method, for example, by the d-bar dressing method.[2,3] However, for our present purpose we concentrate on exact analytic solutions of Eq. (4). In doing so, it is convenient to rewrite Eq. (4) or its equivalent stereographic form (13) in the Hirota bilinear form. On writing

$$\omega=\frac{g}{f}, \tag{45}$$

Eq. (4) or Eq. (13) becomes

$$(iD_t-D_xD_y)(f^*\circ g)=0, \tag{46a}$$

$$(iD_t-D_xD_y)(f^*\circ f-g^*\circ g)=0, \tag{46b}$$

$$D_x(f^*\circ f+g^*\circ g)=0, \tag{46c}$$

while the potential u takes the form

$$u(x,y,t)=-\frac{iD_y(f^*\circ f+g^*\circ g)}{f^*\circ f+g^*\circ g}, \tag{46d}$$

where g and f are complex valued functions. Here D_x is the Hirota bilinear operator, defined by

$$D_x^kD_y^mD_t^n(f\circ g)=(\partial_x-\partial_{x'})^k(\partial_y-\partial_{y'})^m(\partial_t-\partial_{t'})^nf(x,y,t)g(x,y,t)\|_{x=x',y=y',t=t'}. \tag{47}$$

Using the above definition of the D-operator, we get from (31d) that

$$u_x=-2i[D_y(f\circ g)D_x(f^*\circ g^*)-c.c]. \tag{48a}$$

In terms of g and f, the spin field takes the form

$$S^+=\frac{2f^*g}{|f|^2+|g|^2},\quad S_3=\frac{|f|^2-|g|^2}{|f|^2+|g|^2}. \tag{48b}$$

Equation (46) represents the starting point to obtain interesting classes of solutions for the spin system (4).The construction of the solutions is standard. One expands the functions g and f as a series

$$g=\epsilon g_1+\epsilon^3g_3+\epsilon^5g_5+\cdots, \tag{49a}$$

$$f=1+\epsilon^2 f_2+\epsilon^4 f_4+\epsilon^6 f_6+\ \ldots\ . \tag{49b}$$

Substituting these expansions into (46 a,b,c) and equating the coefficients of ϵ, one obtains the following system of equations from (46a):

$$\epsilon^1: ig_{1t}+g_{1xy}=0, \tag{50a}$$

$$\epsilon^3: [i\partial_t+\partial_x\partial_y]g_3=[iD_t-D_xD_y](f_2^*.g_1), \tag{50b}$$

$$\ldots$$

$$\ldots$$

$$\ldots$$

$$\epsilon^{2n+1}: [i\partial_t+\partial_x\partial_y]g_{2n+1}=\sum_{k+m=n}[iD_t-D_xD_y](f_{2k}^*.g_{2m+1}), \tag{50c}$$

and from (46b):

$$\epsilon^2: i\partial_t(f_2^*-f_2)-\partial_x\partial_y(f_2^*+f_2)=[iD_t-D_xD_y](g_1^*.g_1), \tag{51a}$$

$$\epsilon^4: i\partial_t(f_4^*-f_4)-\partial_x\partial_y(f_4^*+f_4)=[iD_t-D_xD_y](g_1^*.g_3+g_3^*.g_1-f_2^*.f_2), \tag{51b}$$

$$\ldots$$

$$\ldots$$

$$\ldots$$

$$\epsilon^{2n}: i\partial_t(f_{2n}^*-f_{2n})-\partial_x\partial_y(f_{2n}^*+f_{2n})=(iD_t-D_xD_y)\left(\sum_{n_1+n_2=n-1} g_{2n_1+1}^*.g_{2n_2+1}\right)$$
$$-(iD_t-D_xD_y)\left(\sum_{m_1+m_2=n} f_{2m_1}^*.f_{2m_2}\right). \tag{51c}$$

Further from (46c), we have the following:

$$\epsilon^2: \partial_x(f_2^*-f_2)=-D_x(g_1^*.g_1), \tag{52a}$$

$$\epsilon^4: \partial_x(f_4^*-f_4)=-D_x(g_1^*.g_3+g_3^*.g_1+f_2^*.f_2), \tag{52b}$$

$$\ldots$$

$$\ldots$$

$$\ldots$$

$$\epsilon^{2n}: \partial_x(f_{2n}^*-f_{2n})=-D_x\left[\sum_{n_1+n_2=n-1}\left(g_{2n_1+1}^*.g_{2n_2+1}+\sum_{n_1+n_2=n} f_{2n_1}^*.f_{2n_2}\right)\right]. \tag{52c}$$

Solving recursively the above equations, we obtain many interesting classes of solutions to Eq. (4).

VII. SOLUTIONS OF THE SPIN SYSTEM

Using the results of the previous section, we are in a position to construct many exact solutions such as solitons, domain walls, breaking solitons and induced dromions of Eq. (4). To obtain such solutions, we can use Eqs. (46) as the starting point.

A. The 1-line soliton and curved soliton solutions

On solving Eq. (50a), we obtain

$$g_1=\sum_{j=1}^{N}\exp\chi_j, \quad \chi_j=a_jx+b_j(y,t)+c_j, \tag{53a}$$

where $b_j(y,t)$ is an arbitrary function of (y,t) satisfying the relation

$$b_j(y,t)=b_j(\rho)=b_j(y+ia_jt), \tag{53b}$$

and a_j and c_j are complex constants. In order to construct the one soliton solution, we take the case of $N=1$ in (53a) and substitute it in Eq. (51a). We obtain

$$f_2=\exp(\chi_1+\chi_1^*+\psi). \tag{54a}$$

Using Eqs. (51a) and (52a), it is found that

$$\exp\psi=-\frac{a_1^2}{(a_1+a_1^*)^2}. \tag{54b}$$

If we use the above forms of g_1 and f_2 in Eqs. (50b), (51b) and (52b), we can see that $g_j=0$ for $j\geqslant 3$ and $f_j=0$ for $j\geqslant 4$. By substituting the values of g_1 and f_2 in Eqs. (48a) and (48b), we obtain the expressions for the 1-soliton solution for the spin components and for the potential for example with the choice $\exp c_1=2a_{1R}/a_1^*$ as

$$S^+(x,y,t)=\frac{2a_{1R}}{a_{1R}^2+a_{1I}^2}\exp\, i\chi_{1I}[ia_{1I}-a_{1R}\tanh\chi_{1R}]\operatorname{sech}\chi_{1R}, \tag{55a}$$

$$S_3(x,y,t)=1-\frac{2a_{1R}^2}{a_{1R}^2+a_{1I}^2}\operatorname{sech}^2\chi_{1R}, \tag{55b}$$

and the potential as

$$u(x,y,t)=\frac{2a_{1R}}{a_{1R}^2+a_{1I}^2}(a_{1I}b_{1R}'-a_{1R}b_{1I}')\operatorname{sech}^2\chi_{1R}. \tag{55c}$$

It follows from Eqs. (55) that one can identify two types of solitons.

1. Line solitons

If we choose the function $b_1(y,t)$ in Eq. (53b) as

$$b_1(y,t)=b_1y+ib_1a_1t, \tag{56a}$$

then

$$\chi_1=a_1x+b_1y+ia_1b_1t+c_1, \tag{56b}$$

where b_1 is now a complex constant, and Eqs. (55) correspond to the line solitons. In this case the spin vector $\vec{S}\to(0,0,1)$ for fixed t as x, $y\to\pm\infty$, except along the line

$$\chi_{1R}=a_{1R}x+b_{1R}y-(a_{1R}b_{1I}+a_{1I}b_{1R})t+c_{1R}=0, \tag{57}$$

where it is still bounded.

2. Curved solitons

However for arbitrary form of $b_1(\rho)=b_1(y+ia_1t)$ in Eq. (53b) and for fixed (y,t), it follows from (55) that $\vec{S}\to(0,0,1)$ as $x,y\to\pm\infty$ and the wavefront itself is defined by the equation

$$\chi_{1R}=a_{1R}x+b_{1R}(\rho)+c_{1R}=0. \tag{58}$$

We may call such solitons (which do not decay along the curve (58)) as curved solitons.[23]

B. The 2-soliton and N-soliton solutions

To generate a 2 line or curved soliton solution (2-SS), we take $N=2$ in (53a) and hence g_1 takes the form

$$g_1=\exp\,\chi_1+\exp\,\chi_2. \tag{59}$$

Substituting (59) in (50)–(52), after some calculation we obtain

$$f_2=N_{11}\exp(\chi_1+\chi_1^*)+N_{12}\exp(\chi_1+\chi_2^*)+N_{21}\exp(\chi_1^*+\chi_2)+N_{22}\exp(\chi_2+\chi_2^*), \tag{60a}$$

$$g_3=L_{12}N_{11}N_{12}\exp(\chi_1+\chi_1^*+\chi_2)+L_{12}N_{22}N_{21}\exp(\chi_1+\chi_2+\chi_2^*), \tag{60b}$$

$$f_4=L_{12}L_{12}^*N_{11}N_{12}N_{21}N_{22}\exp(\chi_1+\chi_1^*+\chi_2+\chi_2^*), \tag{60c}$$

where

$$N_{rs}=-\frac{a_r^2}{(a_r+a_s^*)^2},\quad L_{rs}=-\frac{(a_r-a_s)^2}{a_s^2}, \tag{60d}$$

and $g_j=0$ for $j\geqslant 5$ and $f_j=0$ for $j\geqslant 6$. Inserting (60) into (48b) we get

$$S^+(x,y,t)=2\frac{(1+f_2^*+f_4^*)(g_1+g_3)}{|1+f_2+f_4|^2+|g_1+g_3|^2}, \tag{61a}$$

$$S_3(x,y,t)=\frac{|1+f_2+f_4|^2-|g_1+g_3|^2}{|1+f_2+f_4|^2+|g_1+g_3|^2}, \tag{61b}$$

and similarly the expression for the potential can also be obtained from (48a) or (46d).

Finally by taking g_1 as

$$g_1=\sum_{j=1}^{N}\exp\,\chi_j$$

and extending the above procedure, one can obtain the N-SS also.

C. The domain wall type solution

The soliton solutions of (55) and (61) correspond to the boundary condition

$$\vec{S}(x,y,t)=(0,0,1),\quad \text{as}\quad x,y\to\pm\infty. \tag{62}$$

Another class of physically interesting solutions are the domain wall type solutions, which have the asymptotic form

$$\vec{S}(x,y,t)=(0,0,\pm 1),\quad \text{as}\quad x,y\to\pm\infty. \tag{63}$$

In order to obtain domain wall solutions in the present model, we make the choice

$$\omega(x,y,t)=g(x,y,t),\quad f(x,y,t)=1. \tag{64}$$

Then, Eq. (46) reduces to

$$ig_t+g_{xy}=0, \tag{65a}$$

$$g_x^*g_y+g_y^*g_x=0, \tag{65b}$$

$$g_x^*g-g^*g_x=0, \tag{65c}$$

which is consistent with Eq. (13). Alternately, we can use another substitution

$$\omega(x,y,t)=\frac{1}{f(x,y,t)},\quad g(x,y,t)=1. \tag{66}$$

Here also it follows from (46) that

$$if_t^*-f_{xy}^*=0, \tag{67a}$$

$$f_x^*f_y+f_xf_y^*=0, \tag{67b}$$

$$f_x^*f-f^*f_x=0. \tag{67c}$$

Comparing Eqs. (65) and (67), we see that if $\omega(x,y,t)$ is a solution of Eq. (13), so also

$$\omega'(x,y,t)=\pm\frac{1}{\omega(x,y,t)} \tag{68}$$

are solutions of Eq. (13). This is an obvious consequence of the fact that Eq. (13) is invariant under inversion.

Now, we find the simplest nontrivial solutions for example of Eq. (65). Let us take the ansatz

$$g=\exp(ax+iby-abt), \tag{69}$$

where a, b are real constants. The components of the spin vector $\vec{S}$ are given by

$$S^+(x,y,t)=\frac{\exp\ iby}{\cosh[a(x-bt-x_0)]}, \tag{70a}$$

$$S_3(x,y,t)=-\tanh[a(x-bt-x_0)]. \tag{70b}$$

We can also have a more general solution of the form

$$g=\exp[ax+im(y,t)], \tag{71}$$

where a is a real constant and $m(y,t)$ is an arbitrary function of y and t. From Eq. (65a), it follows that

$$m=m(\rho)=m(y+iat),\quad \rho=y+iat. \tag{72}$$

Expressions for the spin components are then given by

$$S^+(x,y,t)=\frac{\exp[i\ \mathrm{Re}(m(\rho))]}{\cosh[ax-\mathrm{Im}(m(\rho))]}, \tag{73a}$$

$$S_3(x,y,t)=-\tanh[ax+\mathrm{Re}(m(\rho))], \tag{73b}$$

and the potential is

$$u(x,y,t)=-2m_R'\{1+\exp[-2(ax-m_I)]\}^{-1}, \tag{74}$$

where the $'$ denotes the differentiation with respect to the real part of the argument. Naturally even more general solutions can be obtained by taking more general forms for $g(x,y,t)$ than (69) or (71).

D. The breaking soliton solution

We have already noted in Sec. II that for the present system (4), we have a nonisospectral problem, as the spectral parameter λ satisfies Eq. (9). The above presented solutions all correspond to the constant solution of Eq. (9), namely, $\lambda=\lambda_1=$ constant. One may consider other interesting solutions of Eq. (9). For example, one can have a special solution

$$\lambda=\lambda_1=\delta(y,t)+i\xi(y,t)=\frac{y+k+i\eta}{q-t}, \tag{75}$$

where q, k and η are real constants. Corresponding to this case, we may call the resulting solutions of Eqs. (4) and (24) as breaking solitons.[4] Using the Hirota method, one can also construct the breaking 1-SS of Eq. (4) associated with (75). For this purpose, we take g_1 in the form

$$g=g_1=\exp\chi, \quad \chi=ax+m+c=\chi_R+i\chi_I, \tag{76}$$

where $a=a(y,t)$, $m=m(y,t)$ and $c=c(t)$ are functions to be determined. Substituting (76) into the first of Eq. (53), we get

$$ia_t+aa_y=0, \quad im_t+am_y=0, \quad iA_t+Aa_y=0, \tag{77}$$

where $A=\exp(c)$. Particular solutions of Eqs. (77) have the forms

$$a=-i\lambda=\frac{\eta-i(y+k)}{q-t}, \quad m=m\left(\frac{y+k+i\eta}{q-t}\right), \quad A=\frac{A_0}{q-t}, \tag{78}$$

where η, k, q and A_0 are some constants. From Eqs. (50)-(52), we obtain

$$f_2=B\exp 2\chi_R, \quad B=\frac{(y+k+i\eta)^2}{4\eta^2}. \tag{79}$$

Now, we can write the breaking 1-SS of Eq. (4) (using equations (48b), (76)–(79)),

$$S^+(x,y,t)=\frac{\exp\left[i\chi_{1I}+\ln\dfrac{2\eta}{y+k+i\eta}\right]\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k+i\eta}{2\eta}\right]}{1+\dfrac{\eta^2}{(y+k)^2+\eta^2}\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k+i\eta}{2\eta}\right]\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k-i\eta}{2\eta}\right]}, \tag{80a}$$

$$S_3(x,y,t)=\frac{1-\dfrac{\eta^2}{(y+k)^2+\eta^2}\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k+i\eta}{2\eta}\right]\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k-i\eta}{2\eta}\right]}{1+\dfrac{\eta^2}{(y+k)^2+\eta^2}\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k+i\eta}{2\eta}\right]\operatorname{sech}\left[\chi_{1R}+\ln\dfrac{y+k-i\eta}{2\eta}\right]}, \tag{80b}$$

where $\chi_1=\chi$ as defined in Eq. (76). We see that the solution (80) corresponds to an algebraically decaying solution for large x, y.

E. Localized coherent structures (dromions)

Next, we present the dromion type localized solutions of Eq. (4), the so-called induced localized structures/or induced dromions[23] for the potential $u(x,y,t)$. This is possible by utilizing the freedom in the choice of the arbitrary functions b_{1R} and b_{1I} of b_1 in Eq. (55c) and (53b). For example, if we make the ansatz

$$b_{1I}(\rho_R)=kb_{1R}(\rho_R)=\tanh(\rho_R), \tag{81}$$

$$u=2\eta(\xi-\eta k)\operatorname{sech}^2\ \rho_R\ \operatorname{sech}[\ \eta x+\tanh\ \rho_R-\eta x_0], \tag{82}$$

where $\rho_R=y-a_{1I}t$ and k is a constant. Similarly, the expressions for the spin can be obtained from Eqs. (55 a,b). The solution (82) for $u(x,y,t)$ decays exponentially in all the directions, even though the spin $\vec{S}$ itself is not fully localized. Analogously we can construct another type of ''induced dromion'' solution with the choice

$$b_{1I}=kb_{1R}=\int \frac{d\rho_R}{(\rho_R+\rho_0)^2+1}+b_0, \tag{83}$$

where ρ_0 and b_0 are constants, so that

$$u(x,y,t)=\frac{2\eta(\xi-ky)}{(\rho_R+\rho_0)^2+1}\operatorname{sech}^2\left[\ \eta x+\int \frac{d\rho_R}{(\rho_R+\rho_0)^2+1}-\eta x_0\right]. \tag{84}$$

Proceeding in this way we can construct even more general solutions and multidromions for the potential.

VIII. SOLUTIONS OF (2+1) DIMENSIONAL NLSE

In this section, we wish to consider briefly the corresponding solutions of the equivalent generalized NLSE Eq. (24). Already this equation has received some attention in the literature. The following types of solutions are available:[23,24]

(a) Line solitons,

(b) Induced dromions.

Now, we can construct the N-breaking soliton solutions of Eq. (24) as well. As an example, let us obtain the 1-breaking soliton solution of Eq. (24). The Hirota bilinear form of Eq. (24) can be obtained by using the transformation.

$$\psi=\frac{h}{\phi} \tag{85}$$

as[20,24]

$$[iD_t+D_xD_y](h\circ\phi)=0, \tag{86a}$$

$$D_x^2(\phi\circ\phi)=2hh^*. \tag{86b}$$

We look for the 1-breaking soliton solution in the following form:

$$h=\exp\ \chi, \tag{87a}$$

$$\phi=1+\phi_2, \tag{87b}$$

where $\chi=b(y,t)x+n(y,t)+c(t)$. Substituting (87) into (86), we get

$$ib_t+bb_y=0, \tag{88a}$$

$$in_t+bn_y=0, \tag{88b}$$

$$iB_t+Bb_y=0, \tag{88c}$$

and

$$\phi_2=\frac{1}{(b+b^*)^2}\exp\ \chi+\chi^*=\exp\ 2(b_Rx+n_R+\chi_0), \tag{89}$$

where $\exp\ 2\chi_0=1/4b_R^2$, $B=\exp\ c(t)$ and $b_R=b_R(t)=\mathrm{Re}(b)$. Now, formula (85) provides us the 1-breaking soliton solution of Eq. (24),

$$\psi(x,y,t)=\frac{b_R(t)\exp\ i[b_I(y,t)x+n_I(y,t)+c_0]}{\cosh[b_Rx+n_R(y,t)+\chi_0]}, \tag{90}$$

where $b(y,t)=b_R+ib_I$, $n(y,t)=n_R+in_I$ and $B(t)$ are the solutions of Eqs. (88). Just as in the case of Eq. (77), if we take the following particular solutions of the system of Eqs. (88);

$$b=-i\lambda=\frac{\eta-i(y+k)}{q-t},\quad n=\frac{y+k+i\eta}{q-t},\quad B=\frac{B_0}{(q-t)}, \tag{91}$$

then the 1-breaking soliton solution of Eq. (24) takes the form

$$\psi(x,y,t)=-\frac{\eta}{q-t}\exp\ i\left[-\frac{y+k}{q-t}x+n_I(y,t)+c_0\right]\mathrm{sech}\ Z, \tag{92}$$

where $Z=(\eta/q-t)\,x+n_R(y,t)+\chi_0$ and c_0, χ_0 are constants.

Similarly, we obtain the breaking N-SS of (24). In this case we can take the ansatz

$$g_1=\sum_{j=1}^{N}\exp\ \chi_j \tag{93}$$

with $\chi_j=b_j(y,t)x+n_j(y,t)+c_j(t)$. Inserting (93) into (92), one is lead to

$$ib_{jt}+b_jb_{jy}=0, \tag{94a}$$

$$in_{jt}+b_jn_{jy}=0, \tag{94b}$$

$$iB_{jt}+B_jb_{jy}=0. \tag{94c}$$

Proceeding as before, one can obtain breaking N-soliton solution.

IX. SIMPLEST INTEGRABLE EXTENSIONS

As mentioned above in Ref. 5 (see also Refs. 25,26) a new class of (2+1) dimensional integrable spin equations was proposed. In particular, Eq. (4) is a particular reduction of the following so called M-III equation (according to the notations of Ref. 5)

$$\vec{S}_t=(\vec{S}\wedge\vec{S}_y+u\vec{S})_x+2l(cl+d)\vec{S}_y-4cV\vec{S}_x+\vec{S}\wedge A\vec{S}, \tag{95a}$$

$$u_x=-\vec{S}\cdot(\vec{S}_x\wedge\vec{S}_y), \tag{95b}$$

$$V_x=\frac{1}{4(2lc+d)^2}(\vec{S}_x^2)_y, \tag{95c}$$

where l, c and d are constants and A is the anisotropy tensor. This equation possesses some integrable reductions. For example, one obtains

(a) the isotropic M-I equation (4), when $c=0$, $A=0$;

(b) the isotropic M-II equation, when $d=0$, $A=0$;

(c) the isotropic M-III equation, when $c\neq0\neq d$, $A=0$;

(d) the anisotropic M-I equation, when $c=0$,

and so on. All these equations are integrable in the sense that each one of such reduction has a Lax representation.[5] As an example, let us present the associated linear problem for the isotropic M-III equation,[5]

$$\psi_{1x}=U_1\psi_1, \tag{96a}$$

$$\psi_{1t}=V_1\psi_1+(2c\lambda^2+2d\lambda)\psi_{1y}, \tag{96b}$$

with

$$U_1=[ic(\lambda^2-l^2)+id(\lambda-l)]S+\frac{c(\lambda-l)}{2cl+d}SS_x, \tag{97a}$$

$$V_1=[2c(\lambda^2-l^2)+2d(\lambda-l)]B+\lambda^2F_2+\lambda F_1+F_0, \tag{97b}$$

where

$$F_2=-4ic^2VS,$$

$$F_1=-4icdVS-\frac{4c^2V}{2cl+d}VSS_x-\frac{ic}{2cl+d}S\{(SS_x)_y-[SS_x,B]\}, \tag{98}$$

$$F_0=-lF_1-l^2F_2.$$

Here

$$B=\frac{1}{4}([S,S_y]+2iuS),\quad S=\vec{S}\cdot\vec{\sigma}.$$

From these equations one can deduce the corresponding Lax representations of the isotropic M-I and M-II equations by choosing $c=0$ and $d=0$, respectively. We note that Eq. (95) for the isotropic M-III equation reduction case is gauge[27] and L-equivalent[14] to the following equation

$$i\phi_t=\phi_{xy}-4ic(V\phi)_x+2d^2V\phi, \tag{99a}$$

$$V_x=(|\phi|^2)_y. \tag{99b}$$

This equation has two integrable cases: (a) if $c=0$, Eq. (99) reduces to Eq. (24); (b) if $d=0$, we get the Strachan equation.[19] The Lax representations corresponding to (99) is obtained as follows.

$$\psi_{2x}=U_2\psi_2, \tag{100a}$$

$$\psi_{2t}=V_2\psi_2+(2c\lambda^2+2d\lambda)\psi_{2y}, \tag{100b}$$

with

$$U_2=i[(c\lambda^2+d\lambda)\sigma_3+(2c\lambda+d)Q], \tag{101a}$$

$$V_2=\lambda^2B_2+\lambda B_1+B_0. \tag{101b}$$

Here

$$B_2=-4ic^2V\sigma_3,$$

$$B_1=-4idcV\sigma_3+2c\sigma_3Q_y-8ic^2VQ, \tag{102}$$

$$B_0=\frac{d}{2c}B_1-\frac{d^2}{4c^2}B_2,$$

$$Q=\begin{pmatrix} 0 & \phi \\ -\phi^* & 0 \end{pmatrix}.$$

X. CONCLUSIONS

In this paper we have obtained many interesting classes of exact solutions of Eq. (4) and its equivalent counterpart Eq. (24), after establishing their L-equivalence and gauge equivalence. The equivalence concept was also extended to more general cases in Sec. IX. Another interesting class of solutions are the periodic solutions which for Eq. (4) are given by

$$S_1(x,y,t)=\frac{CD-AB}{AD-BC},\quad S_2(x,y,t)=-i\frac{CD+AB}{AD-BC},$$

$$S_3(x,y,t)=\frac{AD+BC}{AD-BC}, \tag{103}$$

where

$$A=\theta\left(z-\frac{r}{2}-\frac{\delta}{2},\tau\right),\quad B=\theta\left(z+\frac{r}{2}-\frac{\delta}{2},\tau\right),$$

$$C=\theta\left(z-\frac{r}{2}+\frac{\delta}{2},\tau\right),\quad D=\theta\left(z+\frac{r}{2}+\frac{\delta}{2},\tau\right). \tag{104}$$

Here, $\theta(z,\tau)$ is the Riemann θ functions and r, δ are some constants. More detailed investigation of the periodic solution (103) will be considered elsewhere.

Also we would like to note that the extended version of the L-equivalence method which we used in Sec. III works out for many other (2+1) dimensional spin equations as well, for instance, to the Ishimori equation.[14] In the later case, the unit vectors $\vec{e}_k$, $k=1,2,3$, satisfy the Eqs. (14) and (18), while the vector $\vec{e}_1\equiv\vec{S}$ satisfies the Ishimori equation

$$\vec{e}_{1t}=\vec{e}_1\wedge(\vec{e}_{1xx}+\epsilon^2\vec{e}_{1yy})+u_x\vec{e}_{1y}+u_y\vec{e}_{1x}, \tag{105a}$$

$$u_{xx}-\epsilon^2u_{yy}=-2\epsilon^2\vec{e}_1\cdot(\vec{e}_{1x}\wedge\vec{e}_{1y}). \tag{105b}$$

Proceeding as in Sec. III, we can obtain an evolution equation for the curvature κ and torsion τ in the form

$$\kappa_t=\omega_{3x}+\tau\omega_2, \tag{106a}$$

$$\tau_t=\omega_{1x}-\kappa\omega_2, \tag{106b}$$

where

$$\omega_1=\frac{\tau\omega_3-\omega_{2x}}{\kappa},\quad \omega_2=u_x\gamma_2-\kappa_x-\epsilon^2\gamma_{3y}+\gamma_1\gamma_2,$$

$$\omega_3=-\kappa\tau+\kappa u_y+u_x\gamma_3+\epsilon^2\gamma_{2y}-\gamma_1\gamma_3,\quad \gamma_2=\frac{(u_{xx}-\epsilon^2u_{yy})}{2\epsilon^2\kappa}$$

and the values of γ_1 and γ_3 can be obtained from the relations (19a) and (19c) respectively. Then the complex transformation

$$\phi = a \exp ib, \tag{107}$$

where

$$a = \frac{1}{2}[\kappa^2 + \gamma_2^2 + \gamma_3^2 - 2\kappa\gamma_2]^{1/2}, \tag{108a}$$

$$b = \partial_x^{-1}\left(\tau - \frac{u_y}{2} + \frac{\gamma_2\gamma_{3x} - \gamma_3\gamma_{2x} - \kappa\gamma_{3x}}{\kappa^2 + \gamma_2^2 + \gamma_3^2 - 2\kappa\gamma_2}\right) \tag{108b}$$

satisfies the Davey–Stewartson equation for $\epsilon^2 = -1$,

$$i\phi_t = \phi_{yy} - \phi_{xx} + 2\phi u, \tag{109a}$$

$$u_{yy} + u_{xx} + (|\phi|)^2_{xx} - (|\phi|)^2_{yy} = 0, \tag{109b}$$

where u is a function of x, y and t. Here κ and τ have the forms (17). Thus the geometrical equivalence between (105) and (109) can be established using the generalized transformation (107).

To summarize, in this paper, we have shown the equivalence of the (2+1) dimensional spin equation (4) and the generalized (2+1) dimensional NLSE (24). Besides we have found interesting classes of exact solutions for Eq. (4) and NLSE (24). However, many questions remain open and deserve further investigation. These include the existence of, for example, localized coherent structures (other than the induced dromion-like solution presented here), the determination of the Hamiltonian structure and the finding of possible physical applications of the solutions obtained above. Another interesting problem is the classification of solitons by the values of the topological charge. These questions are being pursued further.

ACKNOWLEDGMENTS

This work of M.L. forms part of a Department of Science and Technology, Government of India sponsored research project. R.M. wishes to thank Bharathidasan University for hospitality during his visits to Tiruchirapalli. S.V. acknowledges the receipt of a Junior Research Fellowship from the Council of Scientific and Industrial Research, India. We would like to thank Professor J. Zagrodzinski for useful discussions which lead us to find the explicit forms of the periodical solution (92).

[1] B. Piette and W. J. Zakrzewski, Chaos Solitons Fractals **5**, 2495 (1995).
[2] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1992).
[3] B. G. Konopelchenko, *Solitons in Multidimensions* (World Scientific, Singapore, 1993).
[4] O. I. Bogoyavlensky, *Breaking Solitons* (Nauka, Moscow, 1991).
[5] R. Myrzakulov, *On some integrable and nonintegrable soliton equations of magnets I–IV* (HEPI, Alma-Ata, 1987).
[6] M. Lakshmanan and R. Radha, Pramana **48**, 163 (1997).
[7] M. Lakshmanan, Phys. Lett. A **61**, 53 (1977).
[8] M. Lakshmanan, Th. W. Ruijgrok, and C. J. Thompson, Physica (Utrecht) **84A**, 577 (1976).
[9] L. A. Takhtajan, Phys. Lett. A **64**, 235 (1977).
[10] Y. Ishimori, Prog. Theor. Phys. **72**, 33 (1984).
[11] M. Daniel, K. Porsezian, and M. Lakshmanan, J. Math. Phys. **35**, 6498 (1994); M. Lakshmanan and M. Daniel, Physica (Utrecht) **107A**, 533 (1981).
[12] M. Lakshmanan, J. Math. Phys. **20**, 1667 (1979).
[13] R. Myrzakulov and M. Lakshmanan (HEPI Preprint, Alma-Ata) (1996).
[14] R. Myrzakulov and A. K.Danlybaeva, *The L-equivalent counterpart of the M-III equation*, Preprint (CNLP, Alma-Ata, 1994).
[15] R. Myrzakulov and R. N. Syzdykova, *On the L-equivalence between the Ishimori equation and the Davey-Stewartson equation*, Preprint (CNLP, Alma-Ata, 1994).
[16] R. Myrzakulov, S. Vijayalakshmi, G. N. Nugmanova, and M. Lakshmanan, Phys. Lett. A **233**, 391 (1997).
[17] F. Calogero, Lett. Nuovo Cimento **14**, 43 (1975).
[18] V. E. Zakharov, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980).
[19] I. A. B. Strachan, J. Math. Phys. **34**, 243 (1993).
[20] R. Radha and M. Lakshmanan, Inverse Probl. **10**, L29 (1994).
[21] R. Myrzakulov, K. N. Bliev, and A. B. Borzykh, Reports NAS RK **5**, 17 (1996).

[22] V. E. Zakharov and L. A. Takhtajan, Theor. Math. Phys. **38**, 17 (1979).
[23] R. Radha and M. Lakshmanan, J. Phys. A **30**, 3229 (1997).
[24] I. A. B. Strachan, Inverse Probl. **8**, L21 (1992).
[25] R. Myrzakulov and G. N. Nugmanova, Izvestya NAN RK. Ser. fiz.-mat. **6**, 32 (1992).
[26] R. Myrzakulov, N. K. Bliev, and G. N. Nugmanova, Reports NAS RK **3**, 9 (1992).
[27] G. N. Nugmanova, ''The Myrzakulov equations: the gauge equivalent counterparts and soliton solutions,'' Ph.D. dissertation (Kazak State University, Alma-Ata, 1992).

Positivity properties of the Bel–Robinson tensor

Göran Bergqvist[a)]
Department of Mathematics, Mälardalen University, S-721 23 Västerås, Sweden

(Received 28 April 1997; accepted for publication 15 December 1997)

The Bel–Robinson tensor is a rank 4 tensor constructed from the Weyl curvature tensor. It is well known that when it is contracted with four future-pointing vectors a non-negative value is always obtained. In this paper we investigate various cases in which the value zero is attained and show that it occurs in some situations where space–time is of an algebraically special type. This is of importance for expressions of gravitational energy related to the Bel–Robinson tensor. Some new invariant ways of classifying the Petrov type of space–times using the Bel–Robinson tensor are also presented. © *1998 American Institute of Physics.* [S0022-2488(98)01204-3]

I. INTRODUCTION

In this paper we present some properties and applications of the Bel–Robinson tensor. This rank 4 tensor was introduced independently in 1958–1959 by Bel[1] and Robinson (unpublished). Let C_{abcd} denote the Weyl curvature tensor for a manifold of any dimension and any signature of the metric and let $*C_{abcd}$ denote the dual of C_{abcd} with respect to the first two indices. Then we define the Bel–Robinson tensor T_{abcd} as

$$T_{abcd}=\tfrac{1}{4}(C_a{}^e{}_b{}^f C_{cedf}+*C_a{}^e{}_b{}^f *C_{cedf}). \tag{1.1}$$

Sometimes the factor $\frac{1}{4}$ is omitted, but we shall follow the notation and conventions of Penrose and Rindler.[2] The main interest for this tensor has been in general relativity. As we shall see, it has certain positivity properties and this has led to speculations about relations to energy. However, a clear physical interpretation is still lacking and the applications have been of a more mathematical nature.

Here we shall use spinors to study the Bel–Robinson tensor and therefore we restrict our attention to space–time manifolds. We assume that the dimension is four and that we have a metric g_{ab} of signature $(+---)$ as is natural when spinors are used. The notation will be the abstract index notation[2] where tensors have indices $a,b,c,\ldots$ and spinors $A,A',B,B',\ldots$ with the usual identification $a=AA'$. The antisymmetric spinor ϵ_{AB} is related to the metric by $g_{ab}=\epsilon_{AB}\overline{\epsilon}_{A'B'}$. Let Ψ_{ABCD} denote the Weyl spinor. The Weyl spinor is totally symmetric so $\Psi_{ABCD}=\Psi_{(ABCD)}$ and trace-free $\Psi_{ABC}{}^C=0$.[2] Furthermore, in vacuum we have $\nabla^{AA'}\Psi_{ABCD}=0$. The Weyl tensor can be expressed as

$$C_{abcd}=\Psi_{ABCD}\overline{\epsilon}_{A'B'}\overline{\epsilon}_{C'D'}+\overline{\Psi}_{A'B'C'D'}\epsilon_{AB}\epsilon_{CD}. \tag{1.2}$$

Using the Weyl spinor, the expression (1.1) can be written as[2]

$$T_{abcd}=T_{AA'BB'CC'DD'}=\Psi_{ABCD}\overline{\Psi}_{A'B'C'D'}, \tag{1.3}$$

which is a much simpler form than (1.1) to work with. The properties of the Weyl spinor translate to

$$T_{abcd}=T_{(abcd)} \tag{1.4}$$

and

a)Electronic mail: gbt@mdh.se

 © 1998 American Institute of Physics

$$T_{abc}{}^{c}=0, \tag{1.5}$$

so T_{abcd} is totally symmetric and trace-free. In vacuum it becomes divergence-free, i.e.,

$$\nabla^{a}T_{abcd}=0. \tag{1.6}$$

If C_{abcd} in (1.1) is substituted by the Riemann curvature tensor R_{abcd} one obtains a tensor called the Bel tensor. Originally, the Bel–Robinson tensor was just considered as the special case of the Bel tensor in vacuum. However, (1.1) has become the standard definition of the Bel–Robinson tensor in any space–time. For more details on the early work on these tensors, see Ref. 3.

The Bel–Robinson tensor has certain positivity properties which we shall study in the following sections. Recall that a vector v^{a} is said to be future-pointing if it lies within the future half of the null cone in the tangent space. It is timelike if $v_{a}v^{a}>0$ and null if $v_{a}v^{a}=0$. The simplest positivity property of T_{abcd} is that

$$T_{abcd}a^{a}b^{b}c^{c}d^{d}\geqslant 0 \tag{1.7}$$

for all future-pointing vectors a^{a}, b^{a}, c^{a}, and d^{a} (we return to this inequality in Sec. II). This property, which is similar to the dominant energy condition

$$T_{ab}a^{a}b^{b}\geqslant 0 \tag{1.8}$$

for all future-pointing vectors a^{a} and b^{a}, where T_{ab} is the energy momentum tensor, has led to attempts to give physical interpretation to the Bel–Robinson tensor. Relations to gravitational energy or superenergy have been discussed in, for instance, Refs. 4–8. It has also been related to definitions of quasilocal energy in situations with small surfaces in vacuum (see Refs. 9 and 10 and the references therein). Note that T_{abcd} has the dimension of energy squared, but the positivity property (1.7) has also led to a definition of a kind of tensorial square root of T_{abcd} with the dimension of energy density (see Ref. 3).

Non-negative quantities are also useful in mathematical proofs and the Bel–Robinson tensor was used by Christodoulou and Klainerman[11] in their work on the stability of Minkowski space–time. Recently Bonilla and Senovilla[12] also used it to study causal propagation of gravity in vacuum. They also used the Bel–Robinson tensor to provide a new way of characterizing Petrov type N space–times. We recall that the Weyl spinor can always be written[2]

$$\Psi_{ABCD}=\alpha_{(A}\beta_{B}\gamma_{C}\delta_{D)}, \tag{1.9}$$

i.e., as the symmetric product of four spinors α_{A}, β_{A}, γ_{A}, and δ_{A} (which are uniquely determined up to scaling and ordering). These are called the principal null directions of the Weyl spinor (or Weyl tensor) or the gravitational principal null directions (GPNDs).[13] Space–times are said to be Petrov type I if they are all distinct (by which we really mean nonparallel), of type II if two are equal, of type D if there are two pairs of equal ones, of type III if three are equal and of type N if all four are equal. All types except type I are called algebraically special. Sometimes one also calls the null vector $k_{a}=\alpha_{A}\overline{\alpha}_{A'}$ a GPND if α_{A} is a GPND. In Sec. III we present some new invariant ways of finding Petrov types using the Bel–Robinson tensor. These results generalize the results of Ref. 12.

Before that, we study in Sec. II some cases when we have equality in Eq. (1.7) and this will also be seen to occur mainly in algebraically special space–times. With relations between the Bel–Robinson tensor and gravitational energy it is important to know when the value zero can be attained.

II. POSITIVITY PROPERTIES

We begin by formulating an elementary result on future-pointing vectors.

Lemma 1: Let k^{a} be a given future-pointing null vector and s^{a} a given future-pointing timelike vector. Then there is a future-pointing null vector n^{a} such that $s^{a}=ck^{a}+n^{a}$, where $c=s_{a}s^{a}/(2s_{a}k^{a})>0$.

Proof: Define $n^a=s^a-ck^a$ with c as above. $c>0$ as both s^a and k^a are future-pointing. Furthermore, $n_an^a=s_as^a-2cs_ak^a+c^2k_ak^a=s_as^a-s_as^a=0$ so n^a is null, and $n_as^a=s_as^a-s_as^a/2=s_as^a/2>0$ so n^a is also future-pointing.

Next we recall a well-known positivity result for the Bel–Robinson tensor (see, e.g., Ref. 2).

Theorem 1: $T_{abcd}a^ab^bc^cd^d\geqslant 0$ for all future-pointing vectors a^a, b^a, c^a, and d^a.

The proof is done most simply using spinors, and we illustrate it here since the ideas will be useful in other proofs in this paper. By Lemma 1, any future-pointing timelike vector can be written as the sum of two future-pointing null vectors. The vectors a^a, b^a, c^a, and d^a can be either timelike or null. Writing those that are timelike as a sum of two null vectors, $T_{abcd}a^ab^bc^cd^d$ can be expressed as the sum of between 1 and 16 terms of the type $T_{abcd}e^af^bg^ch^d$ where e^a, f^a, g^a, and h^a are null. Using spinors to write each null vector as $e^a=E^A\overline{E^{A'}}$ and so on and using (1.3) we have

$$T_{abcd}e^af^bg^ch^d=|\Psi_{ABCD}E^AF^BG^CH^D|^2\geqslant 0, \tag{2.1}$$

and of course the sum of such terms is non-negative.

We also remark that Theorem 1 is more general than proposition 1 of Ref. 12. Given a future-pointing timelike vector u^a, define

$$\tau_a=\tau_a(u,u,u)=T_{abcd}u^bu^cu^d. \tag{2.2}$$

Then proposition 1 of Ref. 12 states that $\tau_a\tau^a\geqslant 0$. This follows immediately from Theorem 1 above, since obviously $\tau_aa^a\geqslant 0$ for any future-pointing vector a^a, and this also implies that τ^a is future-pointing (so $\tau_a\tau^a\geqslant 0$).

We now continue by studying some cases in which $T_{abcd}a^ab^bc^cd^d$ takes the value 0. The results are summarized in the following theorem:

Theorem 2: Assume that $T_{abcd}\neq 0$, k^a is a future-pointing null vector, and s^a, u^a, v^a, and w^a are future-pointing timelike vectors. Then

(i) $T_{abcd}s^au^bv^cw^d>0$.
(ii) $T_{abcd}k^au^bv^cw^d=0\Leftrightarrow T_{abcd}k^a=0\Leftrightarrow$space–time is of type N and k^a is a GPND of multiplicity 4.
(iii) $T_{abcd}k^ak^bv^cw^d=0\Leftrightarrow T_{abcd}k^ak^b=0\Leftrightarrow$space–time is of type III or N and k^a is a GPND of multiplicity at least 3.
(iv) $T_{abcd}k^ak^bk^cw^d=0\Leftrightarrow T_{abcd}k^ak^bk^c=0\Leftrightarrow$space–time is of type II, D, III, or N and k^a is a GPND of multiplicity at least 2.
(v) $T_{abcd}k^ak^bk^ck^d=0\Leftrightarrow k^a$ is a GPND.

Before the proof we make some comments. The proof of statement (v) is indicated in Ref. 2. Statement (ii) generalizes a result of Ref. 12, which says that for τ^a defined by (2.2) we have $\tau_a\tau^a=0$ (as $T_{abcd}\neq 0$) if and only if space–time is of type N. To see this, $\tau_a\tau^a=0$ means that τ^a is a null vector and therefore there is another null vector k^a (proportional to τ^a) such that $\tau_ak^a=0$. This implies that $T_{abcd}k^au^bu^cu^d=0$, which by (ii) is equivalent to that space–time is of type N.

Proof of Theorem 2: For completeness we also give the proof of (v) as it is a natural starting point for the other results. Writing $k^a=K^A\overline{K^{A'}}$ and using (1.3) we have that $T_{abcd}k^ak^bk^ck^d=0$ is equivalent to

$$\Psi_{ABCD}K^AK^BK^CK^D=0. \tag{2.3}$$

Using (1.9) to express Ψ_{ABCD} in terms of the GPNDs we obtain

$$\alpha_{(A}\beta_B\gamma_C\delta_{D)}K^AK^BK^CK^D=0, \tag{2.4}$$

but as

$$\alpha_{(A}\beta_B\gamma_C\delta_{D)}K^AK^BK^CK^D=\alpha_A\beta_B\gamma_C\delta_DK^AK^BK^CK^D=(\alpha_AK^A)(\beta_BK^B)(\gamma_CK^C)(\delta_DK^D) \tag{2.5}$$

we find

$$(\alpha_A K^A)(\beta_B K^B)(\gamma_C K^C)(\delta_D K^D)=0. \tag{2.6}$$

This means that at least one factor is zero, e.g., $\alpha_A K^A=0$. Then K^A is proportional to α^A and thus K^A is a GPND of the Weyl spinor or, on vector form, we say that k^a is a GPND which proves (v).

To prove (iv) we first assume that k^a is a GPND of multiplicity at least 2. Then

$$\Psi_{ABCD}=K_{(A}K_B\gamma_C\delta_{D)}, \tag{2.7}$$

which obviously implies that

$$\Psi_{ABCD}K^AK^BK^C=0 \tag{2.8}$$

as every term will contain some contraction of the form $K_AK^A=0$. The tensor form of (2.8) is $T_{abcd}k^ak^bk^c=0$. This trivially implies that

$$T_{abcd}k^ak^bk^cw^d=0. \tag{2.9}$$

It remains to be proven that (2.9) implies that k^a is a GPND of multiplicity at least 2. To do this we use Lemma 1 to write $w^d=ck^d+n^d$. Then

$$T_{abcd}k^ak^bk^cw^d=cT_{abcd}k^ak^bk^ck^d+T_{abcd}k^ak^bk^cn^d \tag{2.10}$$

with $c>0$. By Theorem 1, both terms on the right-hand side are non-negative so (2.9) implies that both are zero. By (v) we obtain that the first term is zero if k^a is a GPND and therefore $\Psi_{ABCD}=K_{(A}\beta_B\gamma_C\delta_{D)}$. Writing $n^d=N^D\overline{N}^{D'}$ the second term is zero if

$$0=\Psi_{ABCD}K^AK^BK^CN^D=K_{(A}\beta_B\gamma_C\delta_{D)}K^AK^BK^CN^D=(K_AN^A)(\beta_BK^B)(\gamma_CK^C)(\delta_DK^D), \tag{2.11}$$

since all the terms containing $K_AK^A=0$ vanish. Now, as w^d is timelike, n^d is not proportional to k^d. Therefore N^A is not proportional to K^A and $K_AN^A\neq0$. Thus we obtain that, for instance, $\beta_BK^B=0$, so β^A is proportional to K^A and hence K^A is a GPND of multiplicity at least 2 which proves (iv).

We prove (ii) and (iii) in a similar way. In (iii) we first assume that k^a is a GPND of multiplicity at least 3. Then $\Psi_{ABCD}=K_{(A}K_BK_C\delta_{D)}$ and $\Psi_{ABCD}K^AK^B=0$ since all terms contain contractions K_AK^A. Then we have $T_{abcd}k^ak^b=0$, which again trivially implies that

$$T_{abcd}k^ak^bv^cw^d=0. \tag{2.12}$$

To show that (2.12) implies that k^a is a GPND of multiplicity at least 3 we use Lemma 1 to write $v^c=c_1k^c+n_1^c$ and $w^d=c_2k^d+n_2^d$ with $c_1>0$ and $c_2>0$. Substituting this into (2.12) we obtain

$$c_1c_2T_{abcd}k^ak^bk^ck^d+c_1T_{abcd}k^ak^bk^cn_2^d+c_2T_{abcd}k^ak^bk^cn_1^d+T_{abcd}k^ak^bn_1^cn_2^d=0, \tag{2.13}$$

where each term is non-negative and therefore each term has to be zero. By (v) the first term is zero if k^a is a GPND. Thus $\Psi_{ABCD}=K_{(A}\beta_B\gamma_C\delta_{D)}$. As in (2.11) the second and third terms are then zero if k^a is a GPND of multiplicity of at least 2 so $\Psi_{ABCD}=K_{(A}K_B\gamma_C\delta_{D)}$. Writing $n_1^d=N_1^D\overline{N}_1^{D'}$ and $n_2^d=N_2^D\overline{N}_2^{D'}$, where neither N_1^A nor N_2^A is proportional to K^A, the fourth term is zero if

$$0=\Psi_{ABCD}K^AK^BN_1^CN_2^D=K_{(A}K_B\gamma_C\delta_{D)}K^AK^BN_1^CN_2^D=(K_AN_1^A)(K_BN_2^B)(\gamma_CK^C)(\delta_DK^D). \tag{2.14}$$

The two first factors are nonzero so we must have that, e.g., $\gamma_CK^C=0$, which means that γ^A and K^A are proportional and therefore K^A is now a GPND of multiplicity at least 3 which proves (iii).

In (ii) first assume that k^a is a GPND of multiplicity 4 so $\Psi_{ABCD}=K_AK_BK_CK_D$ and $T_{abcd}=k_ak_bk_ck_d$. Then obviously $T_{abcd}k^a=0$ and

$$T_{abcd}k^a u^b v^c w^d = 0. \tag{2.15}$$

Conversely, if (2.15) holds, use Lemma 1 to write $u^b = c_1 k^b + n_1^b$, $v^c = c_2 k^c + n_2^c$, and $w^d = c_3 k^d + n_3^d$. Substituting these into (2.15) we obtain that each of the following expressions must be zero: $T_{abcd}k^a k^b k^c k^d$, $T_{abcd}k^a k^b k^c n_3^d$, $T_{abcd}k^a k^b n_2^c n_3^d$, and $T_{abcd}k^a n_1^b n_2^c n_3^d$. With the same argument as above, the first three are zero if k^a is a GPND of multiplicity at least 3 so $\Psi_{ABCD} = K_{(A}K_B K_C \delta_{D)}$. The fourth expression is zero if

$$0 = \Psi_{ABCD}K^A N_1^B N_2^C N_3^D = (K_A N_1^A)(K_B N_2^B)(K_C N_3^C)(\delta_D K^D). \tag{2.16}$$

As before, the first three factors are nonzero and therefore $\delta_D K^D = 0$ which implies that δ^A and K^A are proportional and hence K^A is a GPND of multiplicity 4 and this proves (ii).

In (i) we know that $T_{abcd}s^a u^b v^c w^d \geqslant 0$ by Theorem 1. Assume that $T_{abcd} \neq 0$ and $T_{abcd}s^a u^b v^c w^d = 0$. Let k^a be an *arbitrary* future-pointing null vector and write $s^a = c_1 k^a + n_1^a$, $u^b = c_2 k^b + n_2^b$, $v^c = c_3 k^c + n_3^c$, and $w^d = c_4 k^d + n_4^d$ with $c_i > 0$ for $i = 1,2,3,4$. Then

$$0 = T_{abcd}s^a u^b v^c w^d = c_1 c_2 c_3 c_4 T_{abcd}k^a k^b k^c k^d + 15 \text{ more terms.} \tag{2.17}$$

All 16 terms in (2.17) are non-negative so all must be zero. Therefore $T_{abcd}k^a k^b k^c k^d = 0$ which by (v) means that k^a is a GPND. However, k^a is an arbitrary future-pointing null vector and as there can be at most four different GPNDs as $T_{abcd} \neq 0$, we get a contradiction. Thus $T_{abcd}s^a u^b v^c w^d > 0$ when $T_{abcd} \neq 0$ and (i) is proved.

This ends the proof of Theorem 2.

Remark: In case there is more than one different null vector contracted with the Bel–Robinson tensor it is more difficult to draw conclusions. As an example, consider a case similar to (iv) of Theorem 2. Suppose that

$$T_{abcd}k^a k^b k^c n^d = 0, \tag{2.18}$$

where both k^a and n^a are future-pointing null vectors but assumed not to be parallel. The corresponding spinor equation is

$$\Psi_{ABCD}K^A K^B K^C N^D = 0. \tag{2.19}$$

Suppose that we have a space–time with four different GPNDs $\alpha_A, \beta_A, \gamma_A = \alpha_A + 2\beta_A$, and $\delta_A = 2\alpha_A + \beta_A$. Then if $K_A = \alpha_A + \beta_A$ and $N_A = \alpha_A - \beta_A$ (thus neither of them is a GPND), we have

$$\begin{aligned}
4\Psi_{ABCD}K^A K^B K^C N^D &= 4\alpha_{(A}\beta_B\gamma_C\delta_{D)}K^A K^B K^C N^D \\
&= 4\alpha_A\beta_B\gamma_C\delta_D K^{(A}K^B K^C N^{D)} \\
&= \alpha_A\beta_B(\alpha_C + 2\beta_C)(2\alpha_D + \beta_D)[(\alpha^A + \beta^A) \\
&\quad \times (\alpha^B + \beta^B)(\alpha^C + \beta^C)(\alpha^D - \beta^D) + (\alpha^A + \beta^A) \\
&\quad \times (\alpha^B + \beta^B)(\alpha^C - \beta^C)(\alpha^D + \beta^D) \\
&\quad + (\alpha^A + \beta^A)(\alpha^B - \beta^B)(\alpha^C + \beta^C)(\alpha^D + \beta^D) \\
&\quad + (\alpha^A - \beta^A)(\alpha^B + \beta^B)(\alpha^C + \beta^C)(\alpha^D + \beta^D)] \\
&= (\alpha_A\beta^A)(-3 + 3 + 1 - 1) = 0.
\end{aligned} \tag{2.20}$$

Hence $T_{abcd}k^a k^b k^c n^d = 0$ in a space–time with four distinct GPNDs (i.e., an algebraically general space–time) and neither k^a nor n^a is a GPND.

III. INVARIANT PETROV CLASSIFICATIONS

We now give some alternative ways of classifying the Petrov type of the Weyl tensor using the Bel–Robinson tensor. As opposed to the results of the previous section, some of the results here could equivalently be written directly in terms of the Weyl spinor.

Theorem 3: Suppose that $T_{abcd}\neq 0$. Then

(i) $T_{abcd}T^{abcd}=0\Leftrightarrow T_{abcd}T^{abce}=0\Leftrightarrow$space–time is of type III or N, and
(ii) $T_{abcd}T^{abef}=0\Leftrightarrow T_{abcd}T^{aefg}=0\Leftrightarrow$space–time is of type N.

Proof: The case with $T_{abcd}T^{abcd}=0$ is proven in Ref. 13. We first note that using spinors we have

$$T_{abcd}T^{abcd}=0\Leftrightarrow\Psi_{ABCD}\Psi^{ABCD}=0, \tag{3.1a}$$

$$T_{abcd}T^{abce}=0\Leftrightarrow\Psi_{ABCD}\Psi^{ABCE}=0, \tag{3.1b}$$

$$T_{abcd}T^{abef}=0\Leftrightarrow\Psi_{ABCD}\Psi^{ABEF}=0, \tag{3.1c}$$

$$T_{abcd}T^{aefg}=0\Leftrightarrow\Psi_{ABCD}\Psi^{AEFG}=0. \tag{3.1d}$$

If space–time is of type N, then $\Psi_{ABCD}=\alpha_A\alpha_B\alpha_C\alpha_D$ and clearly all terms in all expressions in (3.1) contain at least one contraction $\alpha_A\alpha^A=0$. If space–time is of type III, then $\Psi_{ABCD}=\alpha_{(A}\alpha_B\alpha_C\beta_{D)}$ and all terms of the expressions (3.1a) and (3.1b) with three or four contractions must contain at least one contraction of type $\alpha_A\alpha^A=0$. This proves the implications from right to left of the theorem. Note that in (3.1c) and (3.1d) with only one or two contractions there are terms without $\alpha_A\alpha^A$ factors, e.g., in (3.1c) we have $(\alpha_A\beta^A)(\beta_B\alpha^B)\alpha_C\alpha_D\alpha^E\alpha^F\neq 0$.

Conversely, to prove the implications from left to right we begin by noting that a proof of the fact that $\Psi_{ABCD}\Psi^{ABCD}=0$ implies type III or N is given in Ref. 13. Next we use the equality[2]

$$\Psi_{ABCD}\Psi^{ABCE}=\tfrac{1}{2}\Psi_{ABCF}\Psi^{ABCF}\epsilon_D{}^E \tag{3.2}$$

or, on tensor form,

$$T_{abcd}T^{abce}=\tfrac{1}{4}T_{abcf}T^{abcf}g_d{}^e \tag{3.3}$$

to see that $T_{abcd}T^{abce}=0$ is equivalent to $T_{abcd}T^{abcd}=0$. This proves case (i) of the theorem.

Consider now Eq. (3.1d). Express Ψ_{ABCD} in a standard spinor basis (o^A,ι^A) as[2]

$$\begin{aligned}\Psi_{ABCD}=&\Psi_0\iota_A\iota_B\iota_C\iota_D-\Psi_1(\iota_A\iota_B\iota_C o_D+\iota_A\iota_B o_C\iota_D+\iota_A o_B\iota_C\iota_D+o_A\iota_B\iota_C\iota_D)+\Psi_2(\iota_A\iota_B o_C o_D\\&+\iota_A o_B\iota_C o_D+\iota_A o_B o_C\iota_D+o_A\iota_B\iota_C o_D+o_A\iota_B o_C\iota_D+o_A o_B\iota_C\iota_D)-\Psi_3(\iota_A o_B o_C o_D\\&+o_A\iota_B o_C o_D+o_A o_B\iota_C o_D+o_A o_B o_C\iota_D)+\Psi_4 o_A o_B o_C o_D\,.\end{aligned} \tag{3.4}$$

A straightforward calculation of all the independent components of $\Psi_{ABCD}\Psi^{AEFG}=0$ with respect to the basis leads to the following equations:

$$\Psi_0\Psi_2=\Psi_1^2, \tag{3.5a}$$

$$\Psi_0\Psi_3=\Psi_1\Psi_2, \tag{3.5b}$$

$$\Psi_0\Psi_4=\Psi_1\Psi_3, \tag{3.5c}$$

$$\Psi_1\Psi_3=\Psi_2^2, \tag{3.5d}$$

$$\Psi_1\Psi_4=\Psi_2\Psi_3, \tag{3.5e}$$

$$\Psi_2\Psi_4=\Psi_3^2. \tag{3.5f}$$

We may now assume that o^A is one of the GPNDs. Then $\Psi_{ABCD}o^Ao^Bo^Co^D=\Psi_0=0$ which by (3.5a) implies that $\Psi_1=0$. By (3.5d) we then get $\Psi_2=0$, which substituted into (3.5f) gives $\Psi_3=0$ and hence only $\Psi_4\neq 0$. By (3.4)

$$\Psi_{ABCD}=\Psi_4 o_A o_B o_C o_D\,, \tag{3.6}$$

so o^A is a GPND of multiplicity 4 which means that space–time is of type N.

Finally we check (3.1c). A similar calculation of the components of $\Psi_{ABCD}\Psi^{ABEF}$ leads to a set of six equations which are (3.5a), (3.5b), (3.5d)–(3.5f), and $\Psi_0\Psi_4+\Psi_2^2=2\Psi_1\Psi_3$. Subtracting (3.5d) from the last equation gives (3.5c) and therefore this new set of equations is equivalent to (3.5), so again space–time must be of type N. This finishes the proof of the theorem.

There are other possibilities for invariant Petrov classification of the Weyl tensor/spinor using quantities constructed from the Weyl spinor (see Chap. 8, of Ref. 13 for some results).

One can also construct other quantities from the Bel–Robinson tensor which can be used for Petrov classification. As mentioned in Sec. II, for the quantity τ_a defined by (2.2) we have if $T_{abcd}\neq 0$ that $\tau_a\tau^a=0$ if and only if space–time is of type N. Similarly, if we define

$$\tau_{abc}=\tau_{abc}(u)=T_{abcd}u^d, \tag{3.7}$$

then, using (3.3),

$$\tau_{abc}\tau^{abc}=T_{abcd}T^{abc}{}_e u^d u^e=\tfrac{1}{4}T_{abcf}T^{abcf}g_{de}u^d u^e. \tag{3.8}$$

Hence, if u^a is a null vector, then (3.8) always implies $\tau_{abc}\tau^{abc}=0$, while if u^a is timelike, then by Theorem 3 and (3.8) $\tau_{abc}\tau^{abc}=0$ as $T_{abcd}\neq 0$ if and only if space–time is of type III or N. Thus, if u^a is timelike, a space–time is of type III if and only if $\tau_{abc}\tau^{abc}=0\neq\tau_a\tau^a$.

IV. SUMMARY

We have studied various non-negative expressions involving the Bel–Robinson tensor and especially under what circumstances these expressions can be zero. There are two immediate applications of these results. The first is that it provides some new invariant ways of determining the Petrov type of space–time. The other is that one can determine when expressions of gravitational energy built on the Bel–Robinson tensor are zero. Although it is not known which kind of expressions for gravitational energy one can expect to be always well defined or unique, there are (as mentioned in the Introduction) relations between the Bel–Robinson tensor and various proposed energy expressions. More specific implications of the results of this paper to these expressions will be studied in coming work.

We would also like to emphasize the advantages of using spinors when studying the Bel–Robinson tensor. Most results presented here, as well as other well-known properties of the Bel–Robinson tensor, are definitely simpler to prove using spinor techniques. Moreover, even when proofs are done using spinors, most results are easy to express in terms of tensor like, for instance, the classification of Petrov types.

ACKNOWLEDGMENTS

This work was supported financially by The Swedish Natural Science Research Council (NFR).

[1] L. Bel, C. R. Acad. Sci. **247**, 1094 (1958).
[2] R. Penrose and W. Rindler, *Spinors and spacetime* (Cambridge U.P., Cambridge, 1984), Vol. 1.
[3] M. A. G. Bonilla and J. M. M. Senovilla, Gen. Relativ. Gravit. **29**, 91 (1997).
[4] S. Deser, in *Gravitation and geometry*, edited by W. Rindler and A. Trautman (Bibliopolis, Naples, 1987).
[5] J. W. York, in *Gravitation and geometry*, edited by W. Rindler and A. Trautman (Bibliopolis, Naples, 1987).
[6] M. D. Roberts, Gen. Relativ. Gravit. **20**, 775 (1988).
[7] N. Bretón, A. Feinstein, and J. Ibáñez, Gen. Relativ. Gravit. **25**, 267 (1993).
[8] I. Robinson, Class. Quantum Grav. **14**, A331 (1997).
[9] G. Bergqvist, Class. Quantum Grav. **11**, 3013 (1994).
[10] A. J. Dougan, Class. Quantum Grav. **9**, 2461 (1992).
[11] D. Christodoulou and S. Klainerman, *The global nonlinear stability of the Minkowski space* (Princeton U.P., Princeton, NJ, 1993).
[12] M. A. G. Bonilla and J. M. M. Senovilla, Phys. Rev. Lett. **11**, 783 (1997).
[13] R. Penrose and W. Rindler, *Spinors and spacetime* (Cambridge U.P., Cambridge, 1986), Vol. 2.

Rotating dust solutions of Einstein's equations with three-dimensional symmetry groups. III. All Killing fields linearly independent of u^α and w^α

Andrzej Krasiński[a)]
N. Copernicus Astronomical Center and College of Science, Polish Academy of Sciences, Bartycka 18, 00 716 Warszawa, Poland

(Received 18 July 1997; accepted for publication 21 August 1997)

This is the third and last part of a series of three papers. Using the same method and the same coordinates as in papers I and II, rotating dust solutions of Einstein's equations are investigated that possess three-dimensional symmetry groups, under the assumption that each of the Killing vectors is linearly independent of velocity u^α and rotation w^α at every point of the spacetime region under consideration. The Killing fields are found and the Killing equations are solved for the components of the metric tensor in every case that arises. No progress was made with the Einstein equations in any of the cases, and no previously known solutions were identified. A brief overview of literature on solutions with rotating sources is given. © *1998 American Institute of Physics.* [S0022-2488(98)01004-4]

I. SUMMARY OF THE METHOD

This paper is the third and last part of a series of three papers; for papers I and II see Refs. 1 and 2. For convenience to the readers, this section is repeated after Paper II.

This is a concise summary of results that will be used in this paper. For proofs, motivations and references see Paper I.[1]

Every timelike vector field u^α of unit length that has zero acceleration and nonzero rotation defines the functions $\tau(x), \eta(x)$ and $\xi(x)$ such that:

$$u_\alpha = \tau_{,\alpha} + \eta\xi_{,\alpha}. \tag{1.1}$$

These functions are defined up to the transformations:

$$\tau = \tau' - S(\xi', \eta'), \quad \xi = F(\xi', \eta'), \quad \eta = G(\xi', \eta'), \tag{1.2}$$

where the functions F and G obey:

$$F_{,\xi'}G_{,\eta'} - F_{,\eta'}G_{,\xi'} = 1 \tag{1.3}$$

(this guarantees that the Jacobian of the transformation is 1), and S is determined by:

$$S_{,\xi'} = GF_{,\xi'} - \eta', \quad S_{,\eta'} = GF_{,\eta'}. \tag{1.4}$$

If u^α is the velocity field of a fluid whose number of particles is conserved:

$$(\sqrt{-g}nu^\alpha)_{,\alpha} = 0 \tag{1.5}$$

(where g is the determinant of the metric tensor and n is the particle number density), then one more function $\zeta(x)$ exists such that:

$$\sqrt{-g}nu^\alpha = \varepsilon^{\alpha\beta\gamma\delta}\xi_{,\beta}\eta_{,\gamma}\zeta_{,\delta}, \tag{1.6}$$

[a)]Electronic mail: akr@alfa.camk.edu.pl

 © 1998 American Institute of Physics

and it is determined up to the transformations:

$$\zeta=\zeta'+T(\xi',\eta'). \tag{1.7}$$

Note that n is not defined uniquely by (1.5). For example, if $u^\alpha=\delta^\alpha{}_0$ and n obeys (1.5), then $n'=nf(x,y,z)$ (where f is an arbitrary function) will also obey (1.5). This nonuniqueness allows for a greater freedom in the choice of ζ than (1.7), and the freedom will be used in some cases.

The following relations hold:

$$u^\alpha\tau_{,\alpha}=1,\quad u^\beta\xi_{,\beta}=u^\beta\eta_{,\beta}=u^\beta\zeta_{,\beta}=0,$$

$$\frac{\partial(\tau,\eta,\xi,\zeta)}{\partial(x^0,x^1,x^2,x^3)}=\sqrt{-g}n\neq 0. \tag{1.8}$$

The last of (1.8) guarantees that $\{\tau,\xi,\eta,\zeta\}$ can be chosen as coordinates, they will be called Plebański coordinates. Then, with $\{\tau,\xi,\eta,\zeta\}=\{x^0,x^1,x^2,x^3\}=\{t,x,y,z\}$:

$$u^\alpha=\delta^\alpha{}_0,\quad u_\alpha=\delta^0{}_\alpha+y\delta^1{}_\alpha,$$

$$g_{00}=1,\quad g_{01}=y,\quad g_{02}=g_{03}=0,\quad g=\det(g_{\alpha\beta})=-n^{-2},$$

$$w^\alpha=n\delta^\alpha_3,\quad \omega_{\alpha\beta}=-\omega_{\beta\alpha}=(\tfrac{1}{2})\delta^1{}_\alpha\delta^2{}_\beta, \tag{1.9}$$

where w^α is the rotation vector field, and $\omega_{\alpha\beta}$ is the rotation tensor corresponding to the velocity field u_α:

$$\omega_{\alpha\beta}=\tfrac{1}{2}(u_{\alpha,\beta}-u_{\beta,\alpha}-\dot{u}_\alpha u_\beta+\dot{u}_\beta u_\alpha),\quad w^\alpha=-(1/\sqrt{-g})\varepsilon^{\alpha\beta\gamma\delta}u_\beta\omega_{\gamma\delta}. \tag{1.10}$$

If $\omega_{\alpha\beta}\neq 0$ and $\dot{u}^\alpha=0$ (what is assumed throughout), then necessarily the pressure $p=$const and κp may be interpreted as the cosmological constant ($\kappa:=8\pi G/c^4$).

If any Killing vector field exists on a manifold (on which all the assumptions specified so far are fulfilled), then, in the coordinates of (1.9), it must be of the form:

$$k^\alpha=(C+\phi-y\phi_{,y})\delta^\alpha{}_0+\phi_{,y}\delta^\alpha{}_1-\phi_{,x}\delta^\alpha{}_2+\lambda\delta^\alpha{}_3, \tag{1.11}$$

where C is an arbitrary constant and $\phi(x,y)$ and $\lambda(x,y)$ are arbitrary functions of two coordinates. Whenever $\phi_{,\alpha}\neq 0$, a transformation of the class (1.2)–(1.4) can be found that leads to:

$$k^\alpha=\delta^\alpha{}_1. \tag{1.12}$$

The metric then becomes independent of x, and the coordinates preserving (1.12) are determined up to the transformations:

$$t'=t-\int yH_{,y}dy+A,\quad x'=x+H(y),\quad y'=y,\quad z'=z+T(y), \tag{1.13}$$

where A is an arbitrary constant and H,T are arbitrary functions.

The condition $\phi_{,\alpha}\neq 0$ that allows one to fulfill (1.12) means that the Killing vector k^α is linearly independent of the vectors u^α and w^α at every point of the spacetime region under consideration. In Paper I, solutions of the Killing equations and of the Einstein equations were considered under the assumption that there exist three Killing vector fields on the manifold, two of which have $\phi=$ const in (1.11), while the third one has $\phi_{,\alpha}\neq 0$ and can be transformed to the form (1.12). In Paper II, it was assumed that only one Killing field has $\phi=$ const, while two have $\phi_{,\alpha}\neq 0$. In the present paper, all three Killing fields will be assumed to have $\phi_{,\alpha}\neq 0$. One of them ($k_{(1)}$) can be transformed to the simple form (1.12), while the remaining two will have the general form (1.11).

In this Paper III, no progress was made with the Einstein equations in any of the cases. Also, no related results were found in the literature except in case 1.1.2.2 (see the end of Sec. V).

II. THE LIE ALGEBRA OF THE SYMMETRY GROUP

According to the assumptions made in the preceding section, there exist the following three Killing vector fields:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_{1},$$

$$k^{\alpha}_{(2)}=(C_2+\phi-y\phi_{,y})\delta^{\alpha}_{0}+\phi_{,y}\delta^{\alpha}_{1}-\phi_{,x}\delta^{\alpha}_{2}+\lambda_2(x,y)\delta^{\alpha}_{3},$$

$$k^{\alpha}_{(3)}=(C_3+\psi-y\psi_{,y})\delta^{\alpha}_{0}+\psi_{,y}\delta^{\alpha}_{1}-\psi_{,x}\delta^{\alpha}_{2}+\lambda_3(x,y)\delta^{\alpha}_{3}, \tag{2.1}$$

where C_2 and C_3 are arbitrary constants, and ϕ, ψ, λ_2 and λ_3 are unknown functions of (x,y), to be determined from the commutation relations. The coordinates of (2.1) are determined up to (1.13).

The fields $k_{(1)},k_{(2)}$ and $k_{(3)}$ will form a Lie algebra if constants $a,\ldots,j$ exist such that:

$$[k_{(1)},k_{(2)}]=ak_{(1)}+bk_{(2)}+ck_{(3)},$$

$$[k_{(1)},k_{(3)}]=dk_{(1)}+ek_{(2)}+fk_{(3)},$$

$$[k_{(2)},k_{(3)}]=gk_{(1)}+hk_{(2)}+jk_{(3)}. \tag{2.2}$$

Equations (2.2) are equivalent to the following set:

$$\phi_{,x}-y\phi_{,xy}=b(C_2+\phi-y\phi_{,y})+c(C_3+\psi-y\psi_{,y}), \tag{2.3a}$$

$$\phi_{,xy}=a+b\phi_{,y}+c\psi_{,y}, \tag{2.3b}$$

$$\phi_{,xx}=b\phi_{,x}+c\psi_{,x}, \tag{2.3c}$$

$$\lambda_{2,x}=b\lambda_2+c\lambda_3, \tag{2.3d}$$

$$\psi_{,x}-y\psi_{,xy}=e(C_2+\phi-y\phi_{,y})+f(C_3+\psi-y\psi_{,y}), \tag{2.3e}$$

$$\psi_{,xy}=d+e\phi_{,y}+f\psi_{,y}, \tag{2.3f}$$

$$\psi_{,xx}=e\phi_{,x}+f\psi_{,x}, \tag{2.3g}$$

$$\lambda_{3,x}=e\lambda_2+f\lambda_3, \tag{2.3h}$$

$$\begin{aligned}&\phi_{,y}(\psi_{,x}-y\psi_{,xy})+y\phi_{,x}\psi_{,yy}-\psi_{,y}(\phi_{,x}-y\phi_{,xy})-y\psi_{,x}\phi_{,yy}\\&\quad=h(C_2+\phi-y\phi_{,y})+j(C_3+\psi-y\psi_{,y}),\end{aligned} \tag{2.3i}$$

$$\phi_{,y}\psi_{,xy}-\phi_{,x}\psi_{,yy}-\psi_{,y}\phi_{,xy}+\psi_{,x}\phi_{,yy}=g+h\phi_{,y}+j\psi_{,y}, \tag{2.3j}$$

$$-\phi_{,y}\psi_{,xx}+\phi_{,x}\psi_{,xy}+\psi_{,y}\phi_{,xx}-\psi_{,x}\phi_{,xy}=-h\phi_{,x}-j\psi_{,x}, \tag{2.3k}$$

$$\phi_{,y}\lambda_{3,x}-\phi_{,x}\lambda_{3,y}-\psi_{,y}\lambda_{2,x}+\psi_{,x}\lambda_{2,y}=h\lambda_2+j\lambda_3. \tag{2.3l}$$

The Eqs. (2.3a)–(2.3c) are integrated with the result:

$$\phi_{,x}=ay+b\phi+c\psi+bC_2+cC_3; \tag{2.4a}$$

the Eqs. (2.3e)–(2.3g) are integrated with the result:

$$\psi_{,x}=dy+e\phi+f\psi+eC_2+fC_3; \tag{2.4b}$$

and the Eqs. (2.3i)–(2.3k) are integrated with the result:

$$\phi_{,y}\psi_{,x}-\phi_{,x}\psi_{,y}=gy+h\phi+j\psi+hC_2+jC_3. \tag{2.4c}$$

The equations are now sorted as follows. Equations (2.3d) and (2.3h) form a set that determines λ_2 and λ_3, Eqs. (2.4a)–(2.4b) form a set that determines ϕ and ψ, the remaining two equations [(2.4c) and (2.3l)] are consistency conditions to be imposed on the solutions of the two former sets.

Note that the set (2.4a)–(2.4b) and the set {(2.3d), (2.3h)} are of the same general form: each of them is an (ordinary differential) linear vector equation of first order:

$$U_{,x}=AU+W, \tag{2.5}$$

where, for (2.4a)–(2.4b), the constant matrix A and the vectors U and W are:

$$A=\begin{pmatrix} b & c \\ e & f \end{pmatrix}, \quad U=\begin{pmatrix} \phi \\ \psi \end{pmatrix}, \quad W=y\begin{pmatrix} a \\ d \end{pmatrix}+A\begin{pmatrix} C_2 \\ C_3 \end{pmatrix}, \tag{2.6}$$

while the set {(2.3d), (2.3h)} is homogeneous, so $W=0$ in (2.5), the matrix A is the same as in (2.6) and

$$U=\begin{pmatrix} \lambda_2 \\ \lambda_3 \end{pmatrix}.$$

With the constants b, c, e and f being all arbitrary, several cases will have to be considered separately. Just as in Paper II, the cases that arise will be organized into a binary tree and numbered in a positional system that will enable one to quickly identify the complementary part of each alternative (see Fig. 1).

The first alternative appears in solving the characteristic equation for the matrix A. Its eigenvalues are:

$$\alpha_{1,2}=\tfrac{1}{2}(b+f+\varepsilon_{1,2}\sqrt{\Delta}), \tag{2.7}$$

where:

$$\Delta:=(b-f)^2+4ce, \quad \varepsilon_1=1, \quad \varepsilon_2=-1. \tag{2.8}$$

We first consider:

Case 1: $\Delta\neq 0$ *(i.e., A has two distinct eigenvalues).* The second alternative appears immediately in finding the eigenvectors of A: the cases $c\neq 0$ and $c=0$ have to be considered separately.

Case 1.1: $c\neq 0$. The solution of the set {(2.3d), (2.3h)} is then:

$$\lambda_2=2cL_2(y)e^{\alpha_1 x}+2cL_3(y)e^{\alpha_2 x},$$

$$\lambda_3=(f-b+\sqrt{\Delta})L_2(y)e^{\alpha_1 x}+(f-b-\sqrt{\Delta})L_3(y)e^{\alpha_2 x}, \tag{2.9}$$

where $L_2(y)$ and $L_3(y)$ are arbitrary functions. The cases $\Delta>0$ and $\Delta<0$ could be considered together for a large part of the reasoning. When $\Delta<0$, α_1 and α_2 are complex and $\alpha_2=\overline{\alpha_1}$. Then, L_2 and L_3 have to be complex, too, with $L_3=\overline{L_2}$. However, the two cases lead to different sets of Bianchi types, and so it will be convenient to split them here.

Case 1.1.1: $\Delta>0$ *(i.e., both eigenvalues are real).* Then, in solving the set (2.4a)–(2.4b), the cases $\det A\neq 0$ and $\det A=0$ have to be considered separately. In the end, however, the case $\det A=0$ turns out to be empty, i.e., in all subcases that arise in it there exists a linear combination of the Killing vectors $k_{(2)}$ and $k_{(3)}$ with constant coefficients that is spanned on u and w. This means that all these subcases are in the domain of Paper II and need not be considered here. Therefore we will do away with the case $\det A=0$ by only indicating the method of verification of the statement above.

When $\det A=0$, the following is true (from (2.7)–(2.8)):

$$e=bf/c, \quad \Delta=(b+f)^2, \quad \alpha_1=b+f:=\alpha, \quad \alpha_2=0. \tag{2.10}$$

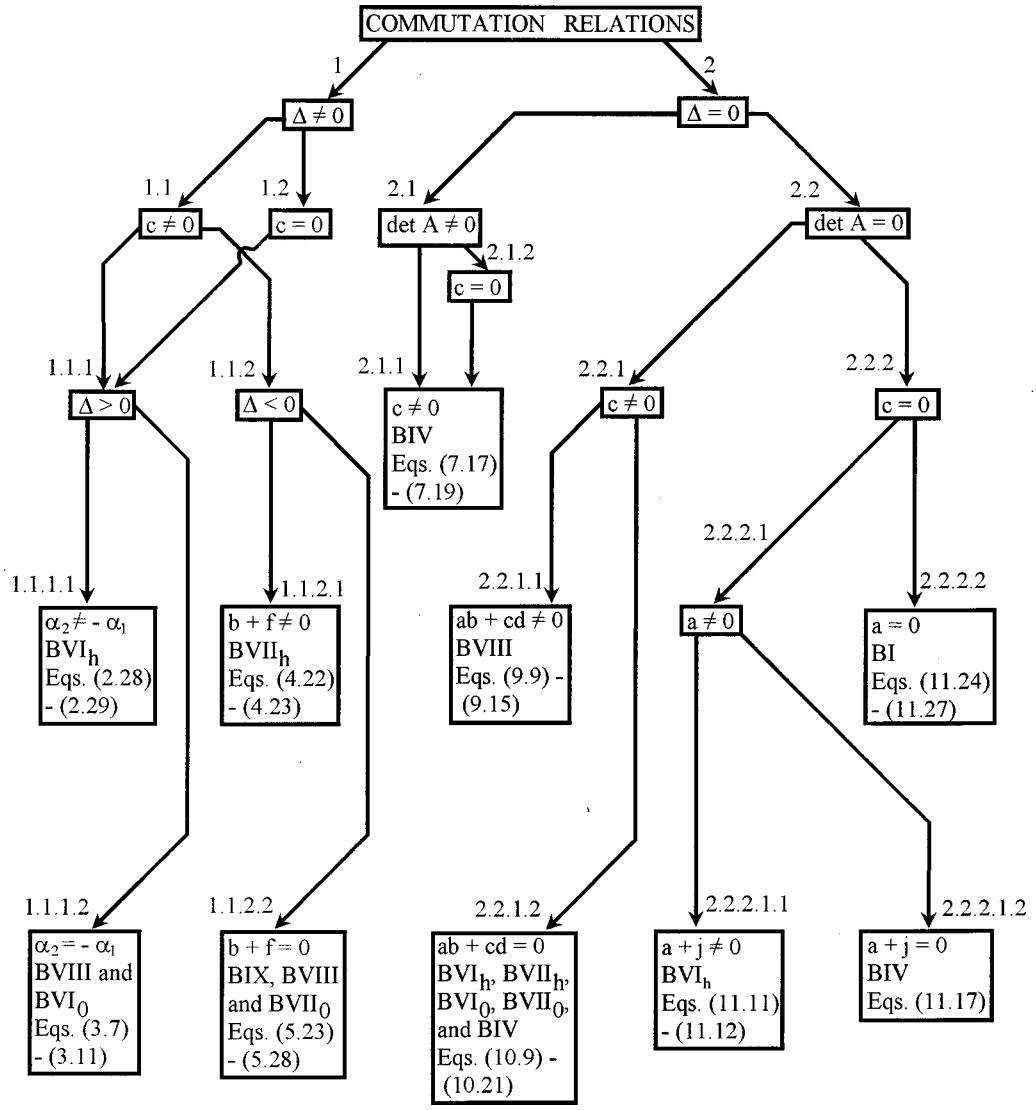

FIG. 1. The classes of metrics considered in the paper. Arrows point from more general classes to subclasses. The numbers at arrows are the case-numbers used in the text. The first entry in each rectangle is the property defining the case; all the symbols are introduced in Eqs. (2.1)–(2.8). The subsequent entries give the following information: 1. The Bianchi type of the corresponding algebra (2.2); 2. The equation-numbers corresponding to the final result in the given case. No progress was made with the Einstein equations in any of the cases. Apart from case 1.1.2.2 for which a subcase was discussed in Ref. 3, none of the cases seem to have appeared in earlier literature (see Sec. XIII). The diagram does not show the links to the entries in the corresponding diagram in Paper II; they are numerous and would obscure the drawing.

Since we are still in case 1 in which $\alpha_2 \neq \alpha_1$ by assumption, we can take it for granted that $\alpha \neq 0$ here. The solutions of (2.3d), (2.3h) and (2.4a)–(2.4b) are:

$$\lambda_2 = 2c[L_2(y) + L_3(y)e^{\alpha x}], \quad \lambda_3 = 2fL_2(y) - 2bL_3(y)e^{\alpha x},$$

$$\phi = F(y)e^{\alpha x} + (af - cd)xy/\alpha + P(y) - (ab + cd)y/\alpha^2 - C_2,$$

$$\psi = (f/c)F(y)e^{\alpha x} - (b/c)(af - cd)xy/\alpha - (b/c)P(y) - (f/c)(ab + cd)y/\alpha^2 - C_3, \quad (2.11)$$

where $F(y)$ and $P(y)$ are other arbitrary functions. The further procedure goes exactly as for the case det $A \neq 0$ presented below, and leads to the results specified above.

From now on, in case 1.1.1 we assume that:

$$\det A = bf - ce \neq 0. \quad (2.12)$$

Then the solution of (2.4)–(2.4b) is:

$$\phi = 2cF(y)e^{\alpha_1 x} + 2cP(y)e^{\alpha_2 x} - \frac{af-cd}{bf-ce}y - C_2,$$

$$\psi = (f-b+\sqrt{\Delta})F(y)e^{\alpha_1 x} + (f-b-\sqrt{\Delta})P(y)e^{\alpha_2 x} - \frac{-ae+bd}{bf-ce}y - C_3. \tag{2.13}$$

The solutions (2.9) and (2.13) must now obey the consistency conditions (2.4c) and (2.3l). Equation (2.4c) is a polynomial in $e^{\alpha_1 x}$ and $e^{\alpha_2 x}$ whose coefficients are functions of y, the polynomial contains $e^{(\alpha_1+\alpha_2)x}, e^{\alpha_1 x}, e^{\alpha_2 x}$ and terms independent of x. The cases $\alpha_2 \neq -\alpha_1$ and $\alpha_2 = -\alpha_1$ require separate consideration. We first consider:

Case 1.1.1.1: $\alpha_2 \neq -\alpha_1$. Comparison of coefficients of $e^{(\alpha_1+\alpha_2)x}$ on both sides of (2.4c) leads to:

$$-\alpha_2 F_{,y} P + \alpha_1 F P_{,y} = 0. \tag{2.14}$$

If $F=0$, then (2.14) is fulfilled identically, and this case has to be considered separately. The result is similar to the case $\det A = 0$. The consideration, parallel to the one that follows below, reveals that with $F=0$ either one of the Killing fields becomes collinear with rotation (and this situation is in the domain of Paper II) or the symmetry group becomes two dimensional (which case is not considered here at all). Consequently, we proceed assuming $F \neq 0$. The solution of (2.14) is then:

$$P = \beta F^{\alpha_2/\alpha_1}, \tag{2.15}$$

where β is an arbitrary constant; $\alpha_1 \neq 0$ because $\det A \neq 0$.

The coefficients of $Fe^{\alpha_1 x}$ on both sides of (2.4c) imply:

$$h = [\alpha_1/(bf-ce)]\left[-\frac{1}{2c}(af-cd)(f-b+\sqrt{\Delta}) - ae + bd\right] - \frac{j}{2c}(f-b+\sqrt{\Delta}). \tag{2.16}$$

In considering the coefficients of $e^{\alpha_2 x}$ in (2.4c) we have to set aside the case $P=0$ for separate consideration because the terms with $e^{\alpha_2 x}$ all vanish identically when $P=0$. However, the case $P=0$ is in fact empty in the same sense as the case $F=0$: either the Killing field $k_{(3)}$ becomes collinear with rotation, and this situation is in the domain of Paper II, or the symmetry group becomes two-dimensional. Hence, we shall follow the case $P \neq 0$ only. Then the coefficients of $Pe^{\alpha_2 x}$ on both sides of (2.4c) imply:

$$j = (bf-ce)^{-1}[-(af-cd)f + (-ae+bd)c], \tag{2.17}$$

and the terms independent of x imply:

$$g = (bf-ce)^{-1}[-(af-cd)d + (-ae+bd)a]. \tag{2.18}$$

Equation (2.3l) is a polynomial in exponential functions of x that involves $e^{2\alpha_1 x}$, $e^{2\alpha_2 x}$, $e^{(\alpha_1+\alpha_2)x}$, $e^{\alpha_1 x}$, and $e^{\alpha_2 x}$. In consequence of (2.9)–(2.18) most of the resulting equations are fulfilled identically, the only one that brings in new information is:

$$-2\alpha_2 F_{,y} L_3 + 2\alpha_1 F L_{3,y} + 2(\beta\alpha_2 F^{\alpha_2/\alpha_1})(L_2 F_{,y}/F - L_{2,y}) = 0. \tag{2.19}$$

This is integrated with the result:

$$L_3 = (1/\alpha_1)F^{\alpha_2/\alpha_1}(\gamma + \beta\alpha_2 L_2/F), \tag{2.20}$$

where γ is a new arbitrary constant.

Finally, the following Killing fields resulted:

$$k^\alpha_{(1)}=\delta^\alpha_1,$$

$$\begin{aligned}
k^\alpha_{(2)}=&\left[2c(F-yF_{,y})e^{\alpha_1 x}+2c\beta F^{\alpha_2/\alpha_1}\left(1-\frac{\alpha_2}{\alpha_1}yF_{,y}/F\right)e^{\alpha_2 x}\right]\delta^\alpha_0\\
&+\left(2cF_{,y}e^{\alpha_1 x}+2c\beta\frac{\alpha_2}{\alpha_1}F^{\alpha_2/\alpha_1-1}F_{,y}e^{\alpha_2 x}-\frac{af-cd}{bf-ce}\right)\delta^\alpha_1\\
&-(2c\alpha_1 Fe^{\alpha_1 x}+2c\alpha_2\beta F^{\alpha_2/\alpha_1}e^{\alpha_2 x})\,\delta^\alpha_2\\
&+\left[2cL_2e^{\alpha_1 x}+2\frac{c}{\alpha_1}F^{\alpha_2/\alpha_1}(\gamma+\beta\alpha_2 L_2/F)e^{\alpha_2 x}\right]\delta^\alpha_3,\\
k^\alpha_{(3)}=&\left[(f-b+\sqrt{\Delta})(F-yF_{,y})e^{\alpha_1 x}+(f-b-\sqrt{\Delta})\beta F^{\alpha_2/\alpha_1}\left(1-\frac{\alpha_2}{\alpha_1}yF_{,y}/F\right)e^{\alpha_2 x}\right]\delta^\alpha_0\\
&+\left[(f-b+\sqrt{\Delta})F_{,y}e^{\alpha_1 x}+(f-b-\sqrt{\Delta})\beta\frac{\alpha_2}{\alpha_1}F^{\alpha_2/\alpha_1-1}F_{,y}e^{\alpha_2 x}-\frac{-ae+bd}{bf-ce}\right]\delta^\alpha_1\\
&-[(f-b+\sqrt{\Delta})\alpha_1 Fe^{\alpha_1 x}+(f-b-\sqrt{\Delta})\beta\alpha_2 F^{\alpha_2/\alpha_1}e^{\alpha_2 x}]\,\delta^\alpha_2\\
&+\left[(f-b+\sqrt{\Delta})L_2e^{\alpha_1 x}+(f-b-\sqrt{\Delta})\frac{1}{\alpha_1}F^{\alpha_2/\alpha_1}(\gamma+\beta\alpha_2 L_2/F)e^{\alpha_2 x}\right]\delta^\alpha_3.
\end{aligned}\tag{2.21}$$

These formulae are simplified by changing the basis in the algebra of the Killing vectors. Taking $k'^\alpha_{(2)}=k^\alpha_{(2)}+[(af-cd)/(bf-ce)]k^\alpha_{(1)}$ and $k'^\alpha_{(3)}=k^\alpha_{(3)}+[(-ae+bd)/(bf-ce)]k^\alpha_{(1)}$ instead of $k^\alpha_{(2)}$ and $k^\alpha_{(3)}$ we obtain the same result as if:

$$a=d=0.\tag{2.22}$$

With (2.22), we take $k'^\alpha_{(3)}=(2\sqrt{\Delta})^{-1}[k^\alpha_{(3)}-(1/2c)(f-b-\sqrt{\Delta})k^\alpha_{(2)}]$ instead of $k^\alpha_{(3)}$ and $k'^\alpha_{(2)}=(2c\beta)^{-1}(k^\alpha_{(2)}-2ck'^\alpha_{(3)})$ instead of $k^\alpha_{(2)}$. Further simplification results from the transformation (1.13) with:

$$H=\alpha_1^{-1}\ln F,\quad T_{,y}=L_2/(\alpha_1 F).\tag{2.23}$$

The Killing field $k^\alpha_{(1)}$ does not change, while the other two become (all primes dropped):

$$k^\alpha_{(2)}=e^{\alpha_2 x}\{\delta^\alpha_0-\alpha_2\delta^\alpha_2+[\gamma/(\beta\alpha_1)]\delta^\alpha_3\},\quad k^\alpha_{(3)}=e^{\alpha_1 x}(\delta^\alpha_0-\alpha_1\delta^\alpha_2).\tag{2.24}$$

Using (2.22) in (2.16)–(2.18) we obtain:

$$g=h=j=0.\tag{2.25}$$

Looking at the commutation relations one can see that in effect we have also achieved $c=e=0$, $\alpha_1=b$, $\alpha_2=f$, even though the initial basis (2.21) was calculated under the assumption $c\neq 0$. This allows us to predict that in the case $c=0$, set aside for separate consideration, we will obtain (2.24) again; see case 1.2 in Sec. VI. In view of the assumptions made earlier, we have:

$$b+f\neq 0\neq b-f,\quad b\neq 0\neq f,\tag{2.26}$$

and consequently the Bianchi type is VI_h, with the free parameter being $a_B=(b+f)/(b-f)$.

The Killing fields are simplified even further after the following transformation that leads out of the Plebański class:

$$t'=(b-f)^{-1}e^{-fx}(bt+y),\quad x'=x,\quad y'=(b-f)^{-1}be^{-bx}(ft+y),$$

$$z'=-\gamma[\beta(b-f)]^{-1}(t+y/b)+z, \tag{2.27}$$

that results in (primes dropped):

$$k^\alpha_{(1)}=-ft\delta^\alpha_0+\delta^\alpha_1-by\delta^\alpha_2,\quad k^\alpha_{(2)}=\delta^\alpha_0,\quad k^\alpha_{(3)}=\delta^\alpha_2,$$

$$u^\alpha=(b-f)^{-1}[be^{-fx}\delta^\alpha_0+bfe^{-bx}\delta^\alpha_2-(\gamma/\beta)\delta^\alpha_3],\quad w^\alpha=n\delta^\alpha_3. \tag{2.28}$$

In the new coordinates, the Killing equations imply for the metric:

$$g_{00}=e^{2fx}[1-(f/b)^2+\gamma^2g_{33}/(b\beta)^2-2f\gamma h_{23}/(b\beta)+f^2h_{22}],$$

$$g_{01}=e^{fx}[\gamma g_{13}/(b\beta)-fh_{12}],\quad g_{02}=e^{(b+f)x}[\gamma h_{23}/(b\beta)-(b-f)/b^2-fh_{22}],$$

$$g_{03}=e^{fx}[\gamma g_{33}/(b\beta)-fh_{23}],\quad g_{12}=e^{bx}h_{12},\quad g_{22}=e^{2bx}h_{22},\quad g_{23}=e^{bx}h_{23}, \tag{2.29}$$

where g_{11}, h_{12}, g_{13}, h_{22}, h_{23} and g_{33} are arbitrary functions of z.

III. CASE 1.1.1.2: $\alpha_2=-\alpha_1$

All equations up to (2.13) still apply, but (2.4c) has to be reconsidered. With $\alpha_2=-\alpha_1$ the following equations hold:

$$f=-b,\quad \Delta=4(b^2+ce),\quad \alpha_1=\sqrt{\Delta}/2=-\alpha_2. \tag{3.1}$$

In considering Eqs. (2.4c) and (2.3l) it may be assumed that $F\neq0\neq P$ because the opposite cases lead, just as before, out of the domain of this paper. With $F\neq0\neq P$, the coefficients of e^{α_1x}, of $e^{-\alpha_1x}$ and the terms independent of x (which now include the coefficients of $e^{(\alpha_1+\alpha_2)x}$) in (2.4c) imply, respectively:

$$h=\Delta^{-1/2}[-(ab/c+d)(-2b+\sqrt{\Delta})-2(-ae+bd)]-j(-2b+\sqrt{\Delta})/(2c),$$

$$j=a,\quad c\Delta FP=(1/2)\{g+\Delta^{-1}[-h(ab+cd)+a(-ae+bd)]\}y^2+B, \tag{3.2}$$

where B is an arbitrary constant.

In Eq. (2.3l) only the terms independent of x provide a piece of new information, the other parts of the equation are fulfilled identically in consequence of (3.1)–(3.2). The new information is:

$$L_3=\gamma/(\alpha_1F)-PL_2/F \tag{3.3}$$

(the integration constant γ was chosen so as to correspond to (2.20)). Like in Sec. II, we first construct the Killing fields by substituting (3.1)–(3.3), (2.9) and (2.13) into (2.1), then simplify the result by changing the basis in the Lie algebra and carrying out the coordinate transformations (1.13) (the difference is that here the coefficient in the formula for $k'^\alpha_{(2)}$ is Δ instead of $(2c\beta)^{-1}$). The result is:

$$a=d=0,\quad k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(3)}=e^{\alpha_1x}(\delta^\alpha_0-\alpha_1\delta^\alpha_2),$$

$$k^\alpha_{(2)}=e^{-\alpha_1x}[(-gy^2+2B)\delta^\alpha_0+2gy\delta^\alpha_1+\alpha_1(gy^2+2B)\delta^\alpha_2+2(c\gamma\Delta/\alpha_1)\delta^\alpha_3]. \tag{3.4}$$

The commutation relations are:

$$[k_{(1)},k_{(2)}]=-\alpha_1k_{(2)},\quad [k_{(2)},k_{(3)}]=2g\alpha_1k_{(1)},\quad [k_{(3)},k_{(1)}]=-\alpha_1k_{(3)}, \tag{3.5}$$

and they correspond to the Bianchi type VIII when $g\neq0$ and type VI_0 when $g=0$ (note that $\alpha_1\neq0$ by the assumption defining case 1.1.1).

The Killing fields are further simplified by the coordinate transformation:

$$t'=t+y/\alpha_1-[2B\alpha_1/(c\gamma\Delta)]z,\quad x'=x,\quad y'=y,\quad z'=[\alpha_1/(2c\gamma\Delta)]z \tag{3.6}$$

that results (with primes dropped) in:

$$k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=e^{-\alpha_1 x}[2gy\delta^\alpha_1+\alpha_1(gy^2+2B)\delta^\alpha_2+\delta^\alpha_3],\quad k^\alpha_{(3)}=-\alpha_1 e^{\alpha_1 x}\delta^\alpha_2,$$

$$u^\alpha=\delta^\alpha_0,\quad w^\alpha=[n\alpha_1/(2c\gamma\Delta)](-4B\delta^\alpha_0+\delta^\alpha_0),$$

$$g_{00}=1,\quad g_{01}=y,\quad g_{02}=-1/\alpha_1,\quad g_{03}=4B, \tag{3.7}$$

the other components of g_{ij} being just unknown functions. The Killing equations for $k^\alpha_{(1)}$ now imply that the metric is independent of x, while those for $k^\alpha_{(3)}$ are solved by:

$$g_{11}=(\alpha_1 y)^2 g_{22}-2\alpha_1 y H_{12}+H_{11},$$

$$g_{12}=-\alpha_1 y g_{22}+H_{12},\quad g_{13}=-\alpha_1 y g_{23}+H_{13}, \tag{3.8}$$

where H_{11}, H_{12}, H_{13}, g_{22}, g_{23} and g_{33} are arbitrary functions of t and z.

In solving the Killing equations for $k^\alpha_{(2)}$, three cases have to be considered separately:

Case I: $gB\neq 0$. In fact, the cases $gB>0$ and $gB<0$ lead to different results, but the formulae for $gB<0$ can be easily reconstructed from those for $gB>0$ by taking real combinations of the complex solutions. Hence, we shall only present the formulae for $gB>0$. The solution of the Killing equations is:

$$\mu:=(2gB)^{1/2}\alpha_1,\quad U:=h_{12}\cos(4\mu z)-k_{12}\sin(4\mu z),$$

$$V:=h_{13}\cos(2\mu z)-k_{13}\sin(2\mu z),$$

$$H_{11}=-(2B\alpha_1^2/\mu)U+h_{11},\quad H_{12}=h_{12}\sin(4\mu z)+k_{12}\cos(4\mu z)-H_{13}/(4B\alpha_1),$$

$$H_{13}=h_{13}\sin(2\mu z)+k_{13}\cos(2\mu z),\quad g_{22}=(g/\mu)U-[g/(2B\alpha_1\mu)]V+h_{22},$$

$$g_{23}=[\mu/(2B\alpha_1^2)]V-h_{33}/(4B\alpha_1),\quad g_{33}=h_{33}, \tag{3.9}$$

where the $h_{ij}(t)$ and $k_{ij}(t)$ are arbitrary functions.

When $gB<0$, μ is imaginary. Then the trigonometric functions go over into the appropriate hyperbolic functions, and h_{12} and h_{13} have to be taken imaginary, too.

Case II: $B=0$. The solution of the Killing equations for $k^\alpha_{(2)}$ is then:

$$H_{11}=h_{33}(\alpha_1 z)^2+2\alpha_1 h_{13}z+h_{11},$$

$$H_{12}=-\alpha_1^2 g h_{33}z^3-3\alpha_1 g h_{13}z^2+(\alpha_1 h_{23}-2g h_{11})z+h_{12},\quad H_{13}=\alpha_1 h_{33}z+h_{13},$$

$$g_{22}=(\alpha_1 g)^2 h_{33}z^4+4\alpha_1 g^2 h_{13}z^3-2g(\alpha_1 h_{23}-2g h_{11})z^2-4g h_{12}z+h_{22},$$

$$g_{23}=-\alpha_1 g h_{33}z^2-2g h_{13}z+h_{13},\quad g_{33}=h_{33}, \tag{3.10}$$

where the $h_{ij}(t)$ are arbitrary functions.

Case III: $g=0$. The Killing equations for $k^\alpha_{(2)}$ imply here:

$$H_{11}=[(4B\alpha_1)^2 h_{22}+8B\alpha_1 h_{23}+h_{33}](\alpha_1 z)^2+(8B\alpha_1 h_{12}+2h_{13})\alpha_1 z+h_{11},$$

$$H_{12}=(4B\alpha_1 h_{22}+h_{23})\alpha_1 z+h_{12},\quad H_{13}=(4B\alpha_1 h_{23}+h_{33})\alpha_1 z+h_{13},$$

$$g_{22}=h_{22},\quad g_{23}=h_{23},\quad g_{33}=h_{33}, \tag{3.11}$$

where the $h_{ij}(t)$ are arbitrary functions.

With this, case 1.1.1 is exhausted and we go back to (2.9) to consider the other branch of the alternative.

IV. CASE 1.1.2: $\Delta<0$ (i.e., THE EIGENVALUES α_1 AND α_2 ARE COMPLEX AND CONJUGATE TO EACH OTHER)

As stated before, it is more convenient to reparametrize (2.9) so that it contains only real quantities, and then repeat the procedure of case 1.1.1 in this new parametrization. We denote:

$$\sqrt{-\Delta}=D,\quad F=G+iJ,\quad P=G-iJ,\quad L_2=M+iN,\quad L_3=M-iN, \tag{4.1}$$

where G, J, M and N are new unknown functions of y, and then:

$$\alpha_{1,2}=\tfrac{1}{2}(b+f)+\tfrac{1}{2}i\varepsilon_{1,2}D,\quad \varepsilon_1=1,\quad \varepsilon_2=-1. \tag{4.2}$$

In this notation, Eqs. (2.9) and (2.13) adapted to the case $\Delta<0$ are:

$$\begin{aligned}
\lambda_2&=4ce^{(b+f)x/2}[M\ \cos(Dx/2)-N\ \sin(Dx/2)],\\
\lambda_3&=\frac{1}{2c}(f-b)\lambda_2-2De^{(b+f)x/2}[M\ \sin(Dx/2)+N\ \cos(Dx/2)],\\
\phi&=4ce^{(b+f)x/2}[G\ \cos(Dx/2)-J\ \sin(Dx/2)]-\frac{af-cd}{bf-ce}y-C_2,\\
\psi&=2(f-b)e^{(b+f)x/2}[G\ \cos(Dx/2)-J\ \sin(Dx/2)]\\
&\quad-2De^{(b+f)x/2}[G\ \sin(Dx/2)+J\ \cos(Dx/2)]-\frac{-ae+bd}{bf-ce}y-C_3.
\end{aligned} \tag{4.3}$$

Just as in case 1.1.1, these expressions must now satisfy the consistency conditions (2.4c) and (2.3l). However, in considering them, the cases $b+f\neq 0$ and $b+f=0$ have to be taken separately.

Case 1.1.2.1: $b+f\neq 0$. Both sides of (2.4c) are then polynomials in $e^{(b+f)x/2}$, and some of their coefficients involve $\cos(Dx/2)$ and $\sin(Dx/2)$. The coefficient of $e^{(b+f)x}$ leads to the equation:

$$4cD[-D(GG_{,y}+JJ_{,y})+(b+f)(GJ_{,y}-G_{,y}J)]=0. \tag{4.4}$$

We are working in the case $c\neq 0\neq D$, so only the expression in square brackets can vanish. Its form suggests the substitution:

$$G=K\ \cos L,\quad J=K\ \sin L, \tag{4.5}$$

where K and L are new functions of y. Equation (4.4) becomes then $-DKK_{,y}+(b+f)K^2L_{,y}=0$, and its solution is:

$$K=Be^{(b+f)L/D}, \tag{4.6}$$

where B is an arbitrary constant; $B\neq 0$ or else we are back in the domain of Paper II. Equations (4.5) and (4.6) provide a parametric representation of G and J in terms of the function L, which is arbitrary at this stage.

The coefficients of $e^{(b+f)x/2}$ in (2.4c) involve sin and cos that always go in fixed pairs. The coefficients of $\{e^{(b+f)x/2}[G\ \cos(Dx/2)-J\ \sin(Dx/2)]\}$ imply:

$$h=[4c(bf-ce)]^{-1}[(af-cd)(b^2-f^2+D^2)+2c(b+f)(-ae+bd)]-j(f-b)/(2c), \tag{4.7}$$

and the coefficients of $\{e^{(b+f)x/2}[G\ \sin(Dx/2)+J\ \cos(Dx/2)]\}$ imply:

$$j=-(bf-ce)^{-1}[f(af-cd)-c(-ae+bd)]. \tag{4.8}$$

Finally, the terms independent of x imply:

$$g=-(bf-ce)^{-1}[h(af-cd)+j(-ae+bd)]. \tag{4.9}$$

With (4.7)–(4.8), Eq. (2.3l) is reduced to:

$$4cDe^{(b+f)x}[-D(GM_{,y}+G_{,y}M+JN_{,y}+J_{,y}N)+(b+f)(GN_{,y}-G_{,y}N-JM_{,y}+J_{,y}M)]=0. \tag{4.10}$$

Only the expression in square brackets can vanish in the case now considered. After (4.5) and (4.6) are substituted into (4.10), the resulting equation integrates to:

$$e^{-(b+f)L/D}\{-M[D\ \cos L+(b+f)\sin L]+N[(b+f)\cos L-D\ \sin L]\}=\gamma B(b+f)=\text{const}. \tag{4.11}$$

Since $b+f\neq 0\neq D$ by assumption, this can be solved for N:

$$N=[(b+f)\cos L-D\ L]^{-1}\{\gamma B(b+f)e^{-(b+f)L/D}+M[D\ \cos L+(b+f)\sin L]\}. \tag{4.12}$$

The resulting Killing fields are:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_{1},$$

$$\begin{aligned}
k^{\alpha}_{(2)}=&\,4ce^{(b+f)x/2}[(G-yG_{,y})\cos(Dx/2)-(J-yJ_{,y})\sin(Dx/2)]\delta^{\alpha}_{0}\\
&+\left\{4ce^{(b+f)x/2}[G_{,y}\cos(Dx/2)-J_{,y}\sin(Dx/2)]-\frac{af-cd}{bf-ce}\right\}\delta^{\alpha}_{1}\\
&-2ce^{(b+f)x/2}\{(b+f)[G\ \cos(Dx/2)-J\ \sin(Dx/2)]-D[G\ \sin(Dx/2)+J\ \cos(Dx/2)]\}\delta^{\alpha}_{2}\\
&+4ce^{(b+f)x/2}[M\ \cos(Dx/2)-N\ \sin(Dx/2)]\delta^{\alpha}_{3},\\
k^{\alpha}_{(3)}=&\,2e^{(b+f)x/2}\{(f-b)[(G-yG_{,y})\cos(Dx/2)-(J-yJ_{,y})\sin(Dx/2)]\\
&-D[(G-yG_{,y})\sin(Dx/2)+(J-yJ_{,y})\cos(Dx/2)]\}\delta^{\alpha}_{0}\\
&+2e^{(b+f)x/2}\left\{(f-b)[G_{,y}\ \cos(Dx/2)-J_{,y}\sin(Dx/2)]-D[G_{,y}\ \sin(Dx/2)+J_{,y}\ \cos(Dx/2)]\right.\\
&\left.-\frac{-ae+bd}{bf-ce}\right\}\delta^{\alpha}_{1}-e^{(b+f)x/2}\{(f^2-b^2-D^2)[G\ \cos(Dx/2)-J(Dx/2)]-2fD[G\ \sin(Dx/2)\\
&+J\ \cos(Dx/2)]\}\delta^{\alpha}_{2}+e^{(b+f)x/2}\{2(f-b)[M\ \cos(Dx/2)-N\ \sin(Dx/2)]-2D[M\ \sin(Dx/2)\\
&+N\ \cos(Dx/2)]\}\delta^{\alpha}_{3}.
\end{aligned} \tag{4.13}$$

Again, the formulae simplify when the basis in the Lie algebra is changed. First, we take $k'^{\alpha}_{(2)}=k^{\alpha}_{(2)}+[(af-cd)/(bf-ce)]k^{\alpha}_{(1)}$ and $k'^{\alpha}_{(3)}=k^{\alpha}_{(3)}+[(-ae+bd)/(bf-ce)]k^{\alpha}_{(1)}$ instead of $k^{\alpha}_{(2)}$ and $k^{\alpha}_{(3)}$, respectively. The result is equivalent to:

$$a=d=0, \tag{4.14}$$

and then (4.7)–(4.9) simplify to:

$$g=h=j=0. \tag{4.15}$$

With (4.14) taken into account, we change the basis again by taking $k'^{\alpha}_{(3)}=(-2BD)^{-1}\{k^{\alpha}_{(3)}-[(f-b)/2c]k^{\alpha}_{(2)}\}$ and $k'^{\alpha}_{(2)}=(4cB)^{-1}k^{\alpha}_{(2)}$ instead of $k^{\alpha}_{(3)}$ and $k^{\alpha}_{(2)}$, and carry out the transformation (1.13) with:

$$H=2L/D,\quad T_{,y}=2[B(b+f)]^{-1}e^{-(b+f)L/D}(M\ \cos L+N\ \sin L). \tag{4.16}$$

The result is equivalent to $L=0=M$, i.e.:

$$G=B,\quad J=0,\quad N=\gamma B. \tag{4.17}$$

The resulting Killing fields are:

$$k^\alpha_{(1)}=\delta^\alpha_1,$$

$$k^\alpha_{(2)}=e^{(b+f)x/2}[\cos(Dx/2)\,\delta^\alpha_0-\tfrac{1}{2}W\delta^\alpha_2-\gamma\,\sin(Dx/2)\,\delta^\alpha_3],$$

$$k^\alpha_{(3)}=e^{(b+f)x/2}[\sin(Dx/2)\,\delta^\alpha_0-\tfrac{1}{2}V\delta^\alpha_2+\gamma\,\cos(Dx/2)\,\delta^\alpha_3], \tag{4.18}$$

where:

$$W:=(b+f)\cos(Dx/2)-D\,\sin(Dx/2),$$

$$V:=D\,\cos(Dx/2)+(b+f)\sin(Dx/2). \tag{4.19}$$

The commutation relations are:

$$[k_1,k_2]=\tfrac{1}{2}(b+f)k_2-\tfrac{1}{2}Dk_3,\quad [k_2,k_3]=0,\quad [k_3,k_1]=-\tfrac{1}{2}Dk_2-\tfrac{1}{2}(b+f)k_3, \tag{4.20}$$

and they correspond to Bianchi type VII_h with the free parameter $a_B=-(b+f)/D$. Since $[k_2,k_3]=0$, coordinates can be adapted to k_2 and k_3 simultaneously. The following transformation does it:

$$t'=-D^{-1}e^{-(b+f)x/2}[Wt+2\cos(Dx/2)y],\quad x'=x,$$

$$y'=e^{-(b+f)x/2}\{D^{-1}Vt-2W^{-1}[1-D^{-1}\,\cos(Dx/2)V]y\},$$

$$z'=(\gamma/D)(b+f)t+2(\gamma/D)y+z, \tag{4.21}$$

but the new coordinates are no longer in the Plebański class. After the transformation, with primes dropped:

$$k^\alpha_{(1)}=\tfrac{1}{2}[-(b+f)t+Dy]\delta^\alpha_0+\delta^\alpha_1-\tfrac{1}{2}[Dt+(b+f)y]\delta^\alpha_2,\quad k^\alpha_{(2)}=\delta^\alpha_2,\quad k^\alpha_{(3)}=\delta^\alpha_0,$$

$$u^\alpha=D^{-1}e^{-(b+f)x/2}(-W\delta^\alpha_0+V\delta^\alpha_2)+(\gamma/D)(b+f)\,\delta^\alpha_3,\quad w^\alpha=n\delta^\alpha_3. \tag{4.22}$$

The formulae for $g_{0\alpha}$ $(\alpha=0,1,2,3)$ in terms of g_{ij} $(i,j=1,2,3)$ in the new coordinates are given in Appendix A.

In the new coordinates, the metric is independent of t and y, while the Killing equations for $k^\alpha_{(1)}$ have the following solution:

$$g_{11}=h_{11},\quad g_{12}=e^{(b+f)x/2}[Wh_{12}-2(\gamma/D^2)(b+f)\cos(Dx/2)h_{13}],\quad g_{13}=h_{13},$$

$$g_{22}=e^{(b+f)x}\{\left[\gamma^2(b+f)^2/D^2+\tfrac{1}{2}D\right]\cos^2(Dx/2)-2(\gamma/D)(b+f)\cos(Dx/2)Wh_{23}+W^2h_{22}\},$$

$$g_{23}=e^{(b+f)x/2}[Wh_{23}-(\gamma/D)(b+f)\cos(Dx/2)h_{33}],\quad g_{33}=h_{33}, \tag{4.23}$$

where the $h_{ij}(z)$ are arbitrary functions. The components $g_{0\alpha}$ in terms of those given above are given in Appendix A.

V. CASE 1.1.2.2: $f=-b$

We go back to Eqs. (4.3) which simplify as follows:

$$\lambda_2=4c[M\,\cos(Dx/2)-N\,\sin(Dx/2)],$$

$$\lambda_3=-(b/c)\lambda_2-2D[M\,\sin(Dx/2)+N\,\cos(Dx/2)], \tag{5.1}$$

$$\phi = 4c[G\ \cos(Dx/2) - J\ \sin(Dx/2)] - \frac{ab+cd}{b^2+ce}y - C_2,$$

$$\psi = -4b[G\ \cos(Dx/2) - J\ \sin(Dx/2)]$$

$$-2D[G\ \sin(Dx/2) + J\ \cos(Dx/2)] + \frac{-ae+bd}{b^2+ce}y - C_3, \tag{5.2}$$

With these new forms of ϕ, ψ, λ_2 and λ_3 we reconsider (2.4c) and (2.3l). Equation (2.4c) is now linear inhomogeneous in $\sin(Dx/2)$ and $\cos(Dx/2)$. The coefficients of $\cos(Dx/2)$, of $\sin(Dx/2)$ and the terms independent of x imply, respectively:

$$\frac{ab+cd}{b^2+ce}D^2G - 2aDJ = 4chG - 4bjG - 2jDJ, \tag{5.3}$$

$$-\frac{ab+cd}{b^2+ce}D^2J - 2aDG = -4chJ + 4bjJ - 2jDG, \tag{5.4}$$

$$-4cD^2(GG_{,y} + JJ_{,y}) = gy + (b^2+ce)^{-1}[-h(ab+cd) + j(-ae+bd)]y. \tag{5.5}$$

From (5.3) and (5.4) it follows that $2(j-a)D(G^2+J^2) = 0$. Since $D \neq 0$ by assumption, and $G = J = 0$ leads to the domain of Paper II, this implies:

$$j = a. \tag{5.6}$$

With (5.6), Eqs. (5.3) and (5.4) reduce to:

$$h = [4c(b^2+ce)]^{-1}(ab+cd)D^2 + ab/c. \tag{5.7}$$

The integral of (5.5) is:

$$-2cD^2(G^2+J^2) = \{g + (b^2+ce)^{-1}[-h(ab+cd) + a(-ae+bd)\}y^2/2 + B, \tag{5.8}$$

where $B = \text{const}$.

Equation (2.3l) leads now to two additional equations, one of which has the solution:

$$GM + JN = \gamma/D = \text{const}, \tag{5.9}$$

and what remains of (2.3l) is then:

$$4aD[M\ \sin(Dx/2) + N\ \cos(Dx/2)] = 0. \tag{5.10}$$

This has two solutions, $a = 0$ and $M = N = 0$, that must be considered separately. However, the case $M = N = 0$ turns out to be included as the subcase $\gamma = 0$ of the formulae below. Hence:

$$a = j = 0, \quad h = \tfrac{1}{4}(b^2+ce)^{-1}dD^2. \tag{5.11}$$

The resulting Killing fields are:

$$k^\alpha_{(1)} = \delta^\alpha_1,$$

$$k^\alpha_{(2)} = 4c[(G - yG_{,y})\cos(Dx/2) - (J - yJ_{,y})\sin(Dx/2)]\delta^\alpha_0$$
$$+ \left\{4c[G_{,y}\ \cos(Dx/2) - J_{,y}\ \sin(Dx/2)] - \frac{cd}{b^2+ce}\right\}\delta^\alpha_1$$
$$+ 2cD[G\ \sin(Dx/2) + J\ \cos(Dx/2)]\delta^\alpha_2 + 4c[M\ \cos(Dx/2) - N\ \sin(Dx/2)]\delta^\alpha_3,$$

$$k^{\alpha}_{(3)}=\{-4b[(G-yG_{,y})\cos(Dx/2)-(J-yJ_{,y})\sin(Dx/2)]$$
$$-2D[(G-yG_{,y})\sin(Dx/2)+(J-yJ_{,y})\cos(Dx/2)]\}\delta^{\alpha}_{0}$$
$$+\left\{-4b[G_{,y}\cos(Dx/2)-J_{,y}\sin(Dx/2)]-2D[G_{,y}\sin(Dx/2)+J_{,y}\cos(Dx/2)]+\frac{bd}{b^2+ce}\right\}\delta^{\alpha}_{1}$$
$$+\{2bD[-G\sin(Dx/2)-J\cos(Dx/2)]+D^2[G\cos(Dx/2)-J\sin(Dx/2)]\}\delta^{\alpha}_{2}$$
$$+\{-4b[M\cos(Dx/2)-N\sin(Dx/2)]-2D[M\sin(Dx/2)+N\cos(Dx/2)]\}\delta^{\alpha}_{3}. \tag{5.12}$$

By changing the basis to $k'^{\alpha}_{(2)}=k^{\alpha}_{(2)}+(b^2+ce)^{-1}cdk^{\alpha}_{(1)}$ and $k'^{\alpha}_{(3)}=k^{\alpha}_{(3)}-(b^2+ce)^{-1}bdk^{\alpha}_{(1)}$ the result equivalent to:

$$d=h=0 \tag{5.13}$$

is achieved, and then, from (5.8):

$$G^2+J^2=-(gy^2+2B)/(4cD^2). \tag{5.14}$$

Equation (5.14) suggests the parametrization:

$$G=K\cos L,\quad J=K\sin L, \tag{5.15}$$

then, from (5.14):

$$K^2=G^2+J^2=-(gy^2+2B)/(4cD^2), \tag{5.16}$$

and L remains arbitrary. With $a=d=0$ we change the basis again to $k''^{\alpha}_{(2)}=(4c)^{-1}k'^{\alpha}_{(2)}$ and $k''^{\alpha}_{(3)}=(-2D)^{-1}[k'^{\alpha}_{(3)}+(b/c)k'^{\alpha}_{(2)}]$ and we carry out the transformation (1.13) with:

$$H=2L/D,\quad T_{,y}=-2(M\sin L-N\cos L)/(DK). \tag{5.17}$$

The result is equivalent to $L=0=N$ which implies:

$$J=0,\ G=K,\quad M=\gamma/(DK), \tag{5.18}$$

and the Killing fields become:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_{1},$$
$$k^{\alpha}_{(2)}=(K-yK_{,y})\cos(Dx/2)\,\delta^{\alpha}_{0}+K_{,y}\cos(Dx/2)\,\delta^{\alpha}_{1}$$
$$+\tfrac{1}{2}DK\sin(Dx/2)\,\delta^{\alpha}_{2}+[\gamma/(DK)]\cos(Dx/2)\,\delta^{\alpha}_{3},$$
$$k^{\alpha}_{(3)}=(K-yK_{,y})\sin(Dx/2)\,\delta^{\alpha}_{0}+K_{,y}\sin(Dx/2)\,\delta^{\alpha}_{1}$$
$$-\tfrac{1}{2}DK\cos(Dx/2)\,\delta^{\alpha}_{2}+[\gamma/(DK)]\sin(Dx/2)\,\delta^{\alpha}_{3}. \tag{5.19}$$

The commutation relations are:

$$[k_1,k_2]=\tfrac{1}{2}Dk_3,\quad [k_2,k_3]=-[g/(8cD)]k_1,\quad [k_3,k_1]=-\tfrac{1}{2}Dk_2. \tag{5.20}$$

The Bianchi type depends on the constant g: for $g/c>0$ it is type IX, for $g/c<0$ it is type VIII and for $g=0$ it is type VII_0. The last case is contained in (4.20)–(4.23) as the subcase $b+f=0$.

The Killing equations for $k^{\alpha}_{(1)}$ imply that the metric tensor is independent of x. Knowing this, one can simplify the Killing equations for $k^{\alpha}_{(2)}$ and $k^{\alpha}_{(3)}$. Since $k^{\alpha}_{(2)}$ and $k^{\alpha}_{(3)}$ are linear in $\sin(Dx/2)$

and $\cos(Dx/2)$, while $g_{\alpha\beta}$ are independent of x, each Killing equation implies two equations: the coefficients of $\cos(Dx/2)$ and of $\sin(Dx/2)$ have to vanish separately. The pair of equations implied by $k^{\alpha}_{(2)}$ is identical to the pair implied by $k^{\alpha}_{(3)}$, and it is:

$$\begin{aligned}(K-yK,_y)g_{\alpha\beta,t}+[\gamma/(DK)]g_{\alpha\beta,z}+\tfrac{1}{4}D^2K\delta^1_\alpha g_{2\beta}\\ +\delta^2_\alpha\{-yK,_{yy}g_{0\beta}+K,_{yy}g_{1\beta}-[\gamma K,_y/(DK^2)]g_{3\beta}\}+\tfrac{1}{4}D^2K\delta^1_\beta g_{\alpha 2}\\ +\delta^2_\beta\{-yK,_{yy}g_{\alpha 0}+K,_{yy}g_{\alpha 1}-[\gamma K,_y/(DK^2)]g_{\alpha 3}\}=0,\end{aligned}\tag{5.21}$$

$$\begin{aligned}Kg_{\alpha\beta,y}+\delta^1_\alpha\{-(K-yK,_y)g_{0\beta}-K,_yg_{1\beta}-[\gamma/(DK)]g_{3\beta}\}+K,_y\delta^2_\alpha g_{2\beta}\\ +\delta^1_\beta\{-(K-yK,_y)g_{\alpha 0}-K,_yg_{\alpha 1}-[\gamma/(DK)]g_{\alpha 3}\}+K,_y\delta^2_\beta g_{\alpha 2}=0.\end{aligned}\tag{5.22}$$

The solution of (5.22) is:

$$g_{11}=y^2+2H_{33}(\gamma K/D)^2\int K^{-3}R(y)dy+2(\gamma/D)K^2H_{13}R(y)+K^2H_{11},$$

$$g_{12}=(\gamma/D)H_{23}R(y)+H_{12},\quad g_{13}=(\gamma/D)H_{33}KR(y)+KH_{13},$$

$$g_{22}=H_{22}/K^2,\quad g_{23}=H_{23}/K,\quad g_{33}=H_{33},\tag{5.23}$$

where the $H_{ij}(t,z)$ are arbitrary functions, and $R(y)$ is:

$$R(y):=\int K^{-3}\mathrm{d}y.\tag{5.24}$$

With K given by (5.16), $R(y)$ and $\int K^{-3}R(y)dy$ can be easily calculated, but the result has to be given separately for $gB\neq 0$, for $g=0$ and for $B=0$, so Eq. (5.24) is the most compact notation (but see below).

Note that γ and B cannot vanish simultaneously; if $\gamma=0=B$, then $K-yK,_y=0=K,_{yy}$, and then Eq. (5.21) implies $g_{12}=g_{22}=g_{23}=0$. Together with $g_{02}=0$ (we are still in the Plebański class) this means that $\det(g_{\alpha\beta})=0$. Also, g and B cannot vanish simultaneously because with $g=0=B$ we are back in the domain of Paper II. With $\gamma^2+B^2\neq 0$, the following new variables can be introduced for solving (5.21):

$$u=2cD\gamma t+Bz,\quad v=-Bt+2cD\gamma z.\tag{5.25}$$

With γ and B running through all possible values, the hypersurfaces $u=\text{const}$ are timelike, null or spacelike. However, the solution of the Killing equations has the same dependence on u and v in every case.

For solving (5.21), the cases $gB\neq 0$ and $gB=0$ have to be separated.

Case I: $gB\neq 0$. (This means we are considering the Bianchi types IX and VIII here, but not VII_0.) In this case:

$$R=-2cD^2y/(BK),\tag{5.26}$$

and the solution of (5.21) (with (5.23)–(5.24) already taken into account) is:

$$\lambda^2:=gB/(8\delta^4D^2),\quad \delta^2:=(B/D)^2+(2c\gamma)^2,\quad U:=h_{12}\sinh(2\lambda v)+k_{12}\cosh(2\lambda v),$$

$$H_{11}=-cD^2U/(2\delta^2\lambda)+h_{11},\quad H_{12}=h_{12}\cosh(2\lambda v)+k_{12}\sinh(2\lambda v),$$

$$H_{13}=-[cD^2/(2\delta^2\lambda)][h_{23}\sinh(\lambda v)+k_{23}\cosh(\lambda v)],$$

$$H_{22}=-2\delta^2\lambda U/(cD^2)-[gB/(2c^2D^6)]h_{11}-[8c\gamma^2/(BD^2)]h_{33},$$

$$H_{23}=h_{23}\cosh(\lambda v)+k_{23}\sinh(\lambda v),H_{33}=h_{33},\tag{5.27}$$

where the $h_{ij}(u)$ and $k_{ij}(u)$ are arbitrary functions.

Equations (5.27) are adapted to the case $gB>0$. When $gB<0$, $\lambda^2<0$, i.e., λ is imaginary. Then k_{12} and k_{23} have to be taken imaginary (and U becomes imaginary in consequence of this), the hyperbolic functions go over into the corresponding trigonometric functions in the well-known way, and the functions H_{ij} remain real.

Case II: $gB=0$. Then $K_{,yy}=0$. The form of $R(y)$ still depends on whether $g\neq 0$ or $g=0$ and $B\neq 0$ or $B=0$. The formulae below apply in each case. The solution of (5.21) (again with (5.23)–(5.24) already taken into account) is here:

$$H_{11}=\frac{(c\gamma gy)^2}{64\delta^8D^2K^2}h_{33}v^4-\frac{c^2\gamma Dgy}{8\delta^6}h_{23}v^3+\left(\frac{c^2D^4}{4\delta^4}h_{22}+\frac{c\gamma gy}{4\delta^4DK}h_{13}\right)v^2-\frac{cD^2}{\delta^2}h_{12}v+h_{11},$$

$$H_{12}=-\frac{c(\gamma gy)^2}{16\delta^6D^4K^2}h_{33}v^3+\frac{3c\gamma gy}{8\delta^4D}h_{23}v^2-\left(\frac{cD^2}{2\delta^2}h_{22}+\frac{\gamma gy}{2\delta^2D^3K}h_{13}\right)v+h_{12},$$

$$H_{13}=\frac{c\gamma gy}{8\delta^4DK}h_{33}v^2-\frac{cD^2}{2\delta^2}Kh_{23}v+h_{13},\quad H_{22}=\frac{(\gamma gy)^2}{4\delta^4D^6K^2}h_{33}v^2-\frac{\gamma gy}{\delta^2D^3}h_{23}v+h_{22},$$

$$H_{23}=-\frac{\gamma gy}{2\delta^2D^3K}h_{33}v+Kh_{23},H_{33}=h_{33}. \tag{5.28}$$

The subcase of (5.28) in which $g=0$ and the hypersurfaces $u=$const are spacelike should have a common subset with the class considered by Demiański and Grishchuk.[3] These authors considered Bianchi type VII_0 models with nonzero rotation, with spacelike orbits of the symmetry group which are flat and with the source being a perfect fluid (the pressure is not constant in their class). A member of the present collection should result when $p=$const. However, Ref. 3 does not contain sufficient information to identify it.

With this, case 1.1 is exhausted. We go back to (2.8) with $\Delta\neq 0$ and consider:

VI. CASE 1.2: $c=0$

Just as it was announced in the paragraph after (2.25), this case brings no new information. Three situations occur here:

(1) The group becomes two-dimensional (because two Killing vectors become collinear); these cases are not considered here.
(2) A linear combination of the Killing fields with constant coefficients is spanned on u^α and w^α; these cases are in the domain of Paper II.
(3) In the case when the group is three-dimensional and none of the Killing fields is spanned on u^α and w^α, the formulae are equivalent to (2.25)–(2.29) and (3.2)–(3.11) (both sets reappear).

The proof consists simply in retracing the whole reasoning from (2.8) on with $c=0$. As seen from (2.8), with $c=0$ necessarily $\Delta\geqslant 0$, and so no analog of case 1.1.2 arises here. The essential steps of the reasoning are described in Appendix B.

Case 1 is exhausted at this point.

VII. CASE 2: $\Delta=0$ (i.e., A HAS ONE DOUBLE EIGENVALUE)

The reasoning from (2.8) on has to be repeated with this new assumption. Then the double eigenvalue is α and:

$$b=f\pm 2\sqrt{-ce},\quad \alpha=f\pm\sqrt{-ce}=(b+f)/2. \tag{7.1}$$

The double sign denotes two different cases, but they will be considered at one go. In finding the solutions of (2.5), the cases $\det A\neq 0$ and $\det A=0$ have to be considered separately. We first consider:

Case 2.1: det $A \neq 0$. This means:

$$b+f \neq 0 \neq \alpha. \tag{7.2}$$

In the next step, the case $c=0$ has to be set aside for separate consideration, so we first assume:

Case 2.1.1: $c \neq 0$. The solutions for $\lambda_{2,3}, \phi$ and ψ are here:

$$\lambda_2 = \{[1+(b-f)x/2]L_2(y) + cxL_3(y)\}e^{(b+f)x/2},$$

$$\lambda_3 = \{-(b-f)^2 xL_2(y)/(4c) + [1+(f-b)x/2]L_3(y)\}e^{(b+f)x/2}, \tag{7.3}$$

$$\phi = \{[1+(b-f)x/2]F(y) + cxP(y)\}e^{(b+f)x/2} + 4(cd-af)y/(b+f)^2 - C_2,$$

$$\psi = \{-(b-f)^2 xF(y)/(4c) + [1+(f-b)x/2]P(y)\}e^{(b+f)x/2}$$

$$-[a(b-f)^2/c + 4bd]y/(b+f)^2 - C_3. \tag{7.4}$$

With such $\lambda_{2,3}$, ϕ and ψ, Eq. (2.4c) involves polynomials of second degree in $e^{(b+f)x/2}$ and also of second degree in x. The equations implied by the coefficients of $x^2 e^{(b+f)x}$ and of $xe^{(b+f)x}$ are fulfilled identically, while the remaining ones are:

$$-(4c)^{-1}(b-f)^2 FF_{,y} + fF_{,y}P - bFP_{,y} - cPP_{,y} = 0, \tag{7.5}$$

$$[(b-f)F/2 + cP][d-h+(a+j)(b-f)/(2c)] = 0, \tag{7.6}$$

$$[(b-f)F/2 + cP]\{[a(b-3f)+4cd]/(b+f) - j\} = 0, \tag{7.7}$$

$$g = \{-4h(cd-af) + j[a(b-f)^2/c + 4bd]\}/(b+f)^2. \tag{7.8}$$

The vanishing of the first factor in (7.6) and (7.7) leads to the relation:

$$k^\alpha_{(3)} + \frac{b-f}{2c}k^\alpha_{(2)} - \frac{a(f-b)-2cd}{c(b+f)}k^\alpha_{(1)} = \left(\frac{b-f}{2c}L_2 + L_3\right)e^{(b+f)x/2}\delta^\alpha_3,$$

which means that the combination on the left is collinear with w^α, and so this case belongs to the domain of Paper II. Hence, (7.6) and (7.7) imply:

$$h = d + (a+j)(b-f)/(2c), \quad j = [a(b-3f)+4cd]/(b+f). \tag{7.9}$$

The solution of (7.5) may be represented parametrically by:

$$F = 2(\mathscr{A}R + cR\ \ln R)/(b+f), \quad P = -(b-f)F/(2c) + R, \tag{7.10}$$

where $\mathscr{A}$ is an arbitrary constant and $R(y)$ is an arbitrary function. It can be assumed that $R \neq 0$ because otherwise we are back in the domain of Paper II.

Equation (2.3l), with the functions given by (7.3)–(7.4) is a polynomial of the same form as (2.4c) and only the coefficients of $e^{(b+f)x}$ provide a new equation:

$$-(4c)^{-1}(b-f)^2(F_{,y}L_2 + FL_{2,y}) + fF_{,y}L_3 - bFL_{3,y} - c(L_{3,y}P + L_3P_{,y})$$

$$-bL_2P_{,y} + FL_{2,y}P = 0. \tag{7.11}$$

Using (7.10), the solution is found again in a parametric form:

$$L_2 = 2S(\mathscr{A} + c\ln R + c)/(b+f) + \gamma R, \quad L_3 = -(b-f)L_2/(2c) + S, \tag{7.12}$$

where γ is another arbitrary constant and S is another arbitrary function. The resulting Killing fields are now calculated from (2.1) using (7.10) and (7.12). Since this procedure has already been performed a few times in this paper, we shall not quote the intermediate results. It turns out that

$k^\alpha_{(2)}$ and $k^\alpha_{(3)}$ contain terms which are constant multiples of $k^\alpha_{(1)}$, these are removed when a new basis, $k'^\alpha_{(2)}$ and $k'^\alpha_{(3)}$ is appropriately defined. Then we change $k^\alpha_{(3)}$ again, to $k''^\alpha_{(3)} = k'^\alpha_{(3)} + [(b-f)/(2c)]k'^\alpha_{(2)}$ and carry out the transformation (1.13) with:

$$H = 2\ln R/(b+f), \quad T = 2(b+f)^{-1}\int (S/R)\mathrm{d}y. \tag{7.13}$$

The Killing fields that result are (with primes dropped):

$$k^\alpha_{(1)} = \delta^\alpha_1,$$

$$k^\alpha_{(2)} = e^{(b+f)x/2}\{[cx + 2\mathscr{A}/(b+f)]\delta^\alpha_0 - [c + \mathscr{A} + (b+f)cx/2]\delta^\alpha_2 + \gamma\delta^\alpha_3\},$$

$$k^\alpha_{(3)} = e^{(b+f)x/2}[\delta^\alpha_0 - \tfrac{1}{2}(b+f)\delta^\alpha_2]. \tag{7.14}$$

It can be assumed now that $\mathscr{A} = 0$ because this is equivalent to changing the basis to $k'^\alpha_{(2)} = k^\alpha_{(2)} - 2\mathscr{A}k^\alpha_{(3)}/(b+f)$. The commutation relations are:

$$[k_{(1)}, k_{(2)}] = \tfrac{1}{2}(b+f)k_{(2)} + ck_{(3)}, \quad [k_{(2)}, k_{(3)}] = 0, \quad [k_{(3)}, k_{(1)}] = -\tfrac{1}{2}(b+f)k_{(3)}, \tag{7.15}$$

and they correspond to Bianchi type IV.

Because of $[k_{(2)}, k_{(3)}] = 0$, coordinates can be adapted to $k_{(2)}$ and $k_{(3)}$ simultaneously. The following transformation does it:

$$t' = e^{-(b+f)x/2}[t + \tfrac{1}{2}(b+f)tx + xy], \quad x' = x,$$

$$y' = -c^{-1}e^{-(b+f)x/2}[\tfrac{1}{2}(b+f)t + y], \quad z' = \gamma(b+f)t/(2c) + \gamma y/c + z, \tag{7.16}$$

and it results in (primes dropped):

$$k^\alpha_{(1)} = -[\tfrac{1}{2}(b+f)t + cy]\delta^\alpha_0 + \delta^\alpha_1 - \tfrac{1}{2}(b+f)y\delta^\alpha_2, \quad k^\alpha_{(2)} = \delta^\alpha_2, \quad k^\alpha_{(3)} = \delta^\alpha_0,$$

$$u^\alpha = e^{-(b+f)x/2}[\delta^\alpha_0 - (2c)^{-1}(b+f)\delta^\alpha_2 + \gamma(2c)^{-1}(b+f)\delta^\alpha_3], \quad w^\alpha = n\delta^\alpha_3. \tag{7.17}$$

In the new coordinates the metric is independent of t and y. After the coordinate transformation (7.16) and after solving the Killing equations for $k^\alpha_{(1)}$ it assumes the form:

$$g_{00} = \tfrac{1}{4}(b+f)^2 e^{(b+f)x}h_{22}, \quad g_{01} = \tfrac{1}{2}(b+f)e^{(b+f)x/2}h_{12},$$

$$g_{02} = e^{(b+f)x}[-2c/(b+f) + \tfrac{1}{2}(b+f)Wh_{22} + \tfrac{1}{2}(b+f)\gamma h_{23}],$$

$$g_{03} = \tfrac{1}{2}(b+f)e^{(b+f)x/2}h_{23}, \quad g_{11} = h_{11},$$

$$g_{12} = e^{(b+f)x/2}(Wh_{12} + \gamma h_{13}), \quad g_{13} = h_{13},$$

$$g_{22} = e^{(b+f)x}\{-8c(b+f)^{-2}W + [2c/(b+f)]^2 + 2bWh_{23} + \gamma^2 h_{33} + W^2 h_{22}\},$$

$$g_{23} = e^{(b+f)x/2}(Wh_{23} + \gamma h_{33}), \quad g_{33} = h_{33}, \tag{7.18}$$

where the $h_{ij}(z)$ are arbitrary functions, and W, not to be confused with the same symbol from Sec. IV, is:

$$W := \tfrac{1}{2}(b+f)cx + c. \tag{7.19}$$

VIII. CASE 2.1.2: $c = 0$

It follows from (7.1) that in this case:

$$b = f = \alpha \neq 0. \tag{8.1}$$

Similar to what happened in Sec. VI, it turns out that this case is included in the case $c\neq 0$. When the procedure of Sec. VII is retraced with (8.1), the Killing fields that result at the stage corresponding to (7.14) are equivalent to (7.14), with $b+f=2b$, e playing the role of c and the roles of $k_{(2)}$ and $k_{(3)}$ interchanged.

IX. CASE 2.2: det $A=0$

With $\Delta=\det A=0$, the following follows from (2.8) and (2.6):

$$bf=ce,\quad f=-b,\quad \alpha_1=\alpha_2=0, \tag{9.1}$$

and the matrix A is nilpotent, $A^2=0$. Again the case $c=0$ has to be set aside for separate consideration, so we first follow:

Case 2.2.1: $c\neq 0$. The solutions of (2.5)–(2.6) are here:

$$\lambda_2=L_2(y)+cxL_3(y),\quad \lambda_3=-(b/c)\lambda_2+L_3(y), \tag{9.2}$$

$$\phi=F(y)+cxP(y)+\tfrac{1}{2}(ab+cd)x^2y,$$

$$\psi=-(b/c)\phi+P(y)+c^{-1}(ab+cd)xy+c^{-1}(-ay-bC_2-cC_3). \tag{9.3}$$

In considering Eqs. (2.4c) and (2.3l), the case $ab+cd=0$ has to be considered separately, so now we follow:

Case 2.2.1.1: $ab+cd\neq 0$. Equation (2.4c) is now a polynomial of second degree in x, and it implies the following equations:

$$j=a,\quad h=-d,\quad (ab+cd)(yF_{,y}+F+C_2)-c^2PP_{,y}=(cg-a^2)y. \tag{9.4}$$

The integral of the last equation is:

$$F=\tfrac{1}{2}(ab+cd)^{-1}[(cP)^2/y+(cg-a^2)y]+\mathscr{A}(ab+cd)/y-C_2. \tag{9.5}$$

Equation (2.3l) reduces to the single equation:

$$-c^2(L_{3,y}P+L_3P_{,y})+(ab+cd)(yL_{2,y}+L_2)=0. \tag{9.6}$$

In solving (9.6) the case $P=0$ has to be considered separately. However, although some of the intermediate steps of the calculation depend on the assumption $P\neq 0$, in the final formulae for the Killing fields $P=0$ is achieved by a transformation of the (1.13) set. The result is identical to the one obtained with $P=0$ from (9.6) on.

With $P\neq 0$, the solution of (9.6) is:

$$L_3=(ab+cd)[yL_2-\gamma(ab+cd)]/(c^2P). \tag{9.7}$$

We change the basis of the resulting Killing fields by taking $k'^{\alpha}_{(3)}=k^{\alpha}_{(3)}+(a/c)k^{\alpha}_{(1)}+(b/c)k^{\alpha}_{(2)}$, then carry out the transformation (1.13) with:

$$H=cP/[y(ab+cd)],\quad T_{,y}=cL_3/[y(ab+cd)], \tag{9.8}$$

and again change the basis to $k'^{\alpha}_{(2)}=(ab+cd)^{-1}[k^{\alpha}_{(2)}-\tfrac{1}{2}(ab+cd)^{-1}(cg-a^2)k^{\alpha}_{(1)}]$, $k''^{\alpha}_{(3)}=c(ab+cd)^{-1}k'^{\alpha}_{(3)}$. The final Killing fields are:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_1,\quad k^{\alpha}_{(2)}=(2\mathscr{A}/y)\,\delta^{\alpha}_0+(-\mathscr{A}/y^2+x^2/2)\,\delta^{\alpha}_1-xy\,\delta^{\alpha}_2+(\gamma/y)\,\delta^{\alpha}_3,$$

$$k^{\alpha}_{(3)}=x\,\delta^{\alpha}_1-y\,\delta^{\alpha}_2. \tag{9.9}$$

The commutation relations are:

$$[k_{(1)},k_{(2)}]=k_{(3)},\quad [k_{(2)},k_{(3)}]=-k_{(2)},\quad [k_{(3)},k_{(1)}]=-k_{(1)}, \tag{9.10}$$

and they correspond to Bianchi type VIII.

The Killing equations for $k_{(1)}$ imply that the metric is independent of x. Those for $k_{(3)}$ are solved by:

$$g_{11}=H_{11}y^2,\quad g_{12}=H_{12},\quad g_{13}=H_{13}y,$$

$$g_{22}=H_{22}/y^2,\quad g_{23}=H_{23}/y,\quad g_{33}=H_{33}, \tag{9.11}$$

where the $H_{ij}(t,z)$ are arbitrary functions (the components $g_{0\alpha}$ are as in (1.9), we are still in the Plebański class in this case). In solving the Killing equations for $k_{(2)}$ it is useful to observe that $\mathscr{A}$ and γ cannot vanish simultaneously: with $\mathscr{A}=\gamma=0$, the Killing equations for $k_{(2)}$ imply that $g_{12}=g_{22}=g_{23}=0$, i.e., $\det(g_{\alpha\beta})=0$. With $\mathscr{A}^2+\gamma^2\neq 0$, the following variables can be introduced:

$$u=\gamma t-2.\mathscr{A}z,\quad v=2.\mathscr{A}t+\gamma z. \tag{9.12}$$

The Killing equations for $k^{\alpha}_{(2)}$ have to be solved separately for $\mathscr{A}\neq 0$ and for $\mathscr{A}=0$. When $\mathscr{A}\neq 0$, we define δ and λ by:

$$\delta^2=4.\mathscr{A}^2+\gamma^2,\quad \lambda^2=-2.\mathscr{A}/\delta^4, \tag{9.13}$$

and then the solution is:

$$U:=h_{12}\sinh(2\lambda v)+k_{12}\cosh(2\lambda v),\quad V:=h_{23}\sinh(\lambda v)+k_{23}\cosh(\lambda v),$$

$$H_{11}=(-2.\mathscr{A})^{-1/2}U+(\gamma/.\mathscr{A})(-2.\mathscr{A})^{-1/2}V+h_{11},$$

$$H_{12}=h_{12}\cosh(2\lambda v)+k_{12}\sinh(2\lambda v)+\gamma H_{23}/(2.\mathscr{A}),$$

$$H_{13}=(-2.\mathscr{A})^{-1/2}V+\gamma h_{33}/(2.\mathscr{A}),\quad H_{22}=(-2.\mathscr{A})^{1/2}U+2.\mathscr{A}h_{11}-\gamma^2 h_{33}/(2.\mathscr{A}),$$

$$H_{23}=h_{23}\cosh(\lambda v)+k_{23}\sinh(\lambda v),\quad H_{33}=h_{33}, \tag{9.14}$$

where the $h_{ij}(u)$ are arbitrary functions. Equations (9.14) are adapted to the case $\lambda^2>0$ ($\mathscr{A}<0$), but the solution for $\lambda^2<0$ can be easily constructed from this one.

When $\mathscr{A}=0$, necessarily $\gamma\neq 0$ and the Killing equations for $k^{\alpha}_{(2)}$ are solved in the original variables (t,z) as follows:

$$H_{11}=h_{33}z^4/(2\gamma)^2+h_{23}z^3/\gamma^2+(h_{22}/\gamma^2+h_{13}/\gamma)z^2+2h_{12}z/\gamma+h_{11},$$

$$H_{12}=h_{33}z^3/(2\gamma)+3h_{23}z^2/(2\gamma)+(h_{22}/\gamma+h_{13})z+h_{12},$$

$$H_{13}=h_{33}z^2/(2\gamma)+h_{23}z/\gamma+h_{13},\quad H_{22}=h_{33}z^2+2h_{23}z+h_{22},$$

$$H_{23}=h_{33}z+h_{23},\quad H_{33}=h_{33}, \tag{9.15}$$

where the $h_{ij}(t)$ are arbitrary functions.

Equations (9.14) and (9.15) are very similar in form to (5.27) and (5.28), respectively. This suggests that case 2.2.1.1 considered here may be included in case 1.1.2.2 of Sec. V as a limit (combined with a coordinate transformation). However, this author was not able to prove or disprove this hypothesis.

We go back now to (9.3) and consider:

X. CASE 2.2.1.2: $ab+cd=0$

Equations (9.2) still apply and Eqs. (9.3) simplify in the obvious way. The difference with Sec. IX is that Eqs. (2.4c) and (2.3l) impose weaker conditions here. The case $P=0$ leads to the domain of Paper 2, so it may be assumed that $P\neq 0$. Then, the consequences of (2.4c) are:

$$h=bj/c, \tag{10.1}$$

$$cPP_{,y}+(j-a)P+(g-aj/c)y=0. \tag{10.2}$$

Equation (10.2) defines P as a function of y. The consequence of Eq. (2.3) is:

$$c(L_3P)_{,y}+(j-a)L_3=0, \tag{10.3}$$

which defines L_3 once P is given. The functions F and L_2 are still arbitrary at this point.

In the resulting Killing fields we take $k'^{\alpha}_{(3)}=k^{\alpha}_{(3)}+(a/c)k^{\alpha}_{(1)}+(b/c)k^{\alpha}_{(2)}$ instead of $k^{\alpha}_{(3)}$, and carry out the transformation (1.13) with:

$$H=(F+C_2)/(cP), \quad T_{,y}=(L_2/c-HL_3)/P. \tag{10.4}$$

The result of the change of basis and of (10.4) is equivalent to:

$$F=-C_2, \quad L_2=a=b=d=0, \quad c=1, \tag{10.5}$$

and the Killing fields become:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_1, \quad k^{\alpha}_{(2)}=x(P-yP_{,y})\delta^{\alpha}_0+xP_{,y}\delta^{\alpha}_1-P\delta^{\alpha}_2+xL_3\delta^{\alpha}_3,$$

$$k^{\alpha}_{(3)}=(P-yP_{,y})\delta^{\alpha}_0+P_{,y}\delta^{\alpha}_1+L_3\delta^{\alpha}_3. \tag{10.6}$$

Note that with (10.5) fulfilled, the subcase $g=0$ implies $P_{,y}=$const, and then $k^{\alpha}_{(3)}-P_{,y}k^{\alpha}_{(1)}$ is spanned on u^{α} and w^{α}, i.e., this case is in the domain of Paper II. Hence, in what follows it will be assumed that $g\neq 0$. For the same reason it will also be assumed that $(P-yP_{,y})\neq 0$.

The commutation relations (with (10.5) already taken into account) are:

$$[k_{(1)},k_{(2)}]=k_{(3)}, \quad [k_{(2)},k_{(3)}]=jk_{(3)}+gk_{(1)}, \quad [k_{(3)},k_{(1)}]=0. \tag{10.7}$$

The collection of Bianchi types contained in (10.7) is as follows. When:

(a) $g<j^2/4$, the Bianchi type is VI_h with the free parameter $a_B=j/(j^2-4g)^{1/2}$; and VI_0 when $j=0$.
(b) $g>j^2/4$, the Bianchi type is VII_h with the free parameter $a_B=j/(4g-j^2)^{1/2}$; and VII_0 when $j=0$.
(c) $g=j^2/4$, the Bianchi type is IV if $j\neq 0$ and II if $j=0$. This last case belongs to the domain of Paper II, as explained above, and will not be presented here.

The Killing equations have to be solved separately for these three cases.

The coordinates are adapted to $k_{(1)}$ and $k_{(3)}$ by the following transformation:

$$t'=t/(P-yP_{,y}), \quad x'=x-tP_{,y}/(P-yP_{,y}), \quad y'=y,$$

$$z'=-L_3t/(P-yP_{,y})+z. \tag{10.8}$$

In the new coordinates, with primes dropped:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_1, \quad k^{\alpha}_{(2)}=(x-jt)\delta^{\alpha}_0-gt\delta^{\alpha}_1-P\delta^{\alpha}_2, \quad k^{\alpha}_{(3)}=\delta^{\alpha}_0,$$

$$u^{\alpha}=(P-yP_{,y})^{-1}(\delta^{\alpha}_0-P_{,y}\delta^{\alpha}_1-L_3\delta^{\alpha}_3), \quad w^{\alpha}=\delta^{\alpha}_3,$$

$$g_{00}=P^2-y^2P_{,y}{}^2+P_{,y}{}^2g_{11}+2L_3P_{,y}g_{13}+L_3{}^2g_{33},$$

$$g_{01}=y(P-yP_{,y})+P_{,y}g_{11}+L_3g_{13},$$

$$g_{02}=P_{,y}g_{12}+L_3g_{23}, \quad g_{03}=P_{,y}g_{13}+L_3g_{33}. \tag{10.9}$$

The Killing equations for $k_{(1)}$ and $k_{(3)}$ imply now that the metric is independent of t and x, while those for $k_{(2)}$ are:

$$-Pg_{11,y}+2y(P-yP_{,y})+2P_{,y}g_{11}+2L_3g_{13}=0,$$

$$-Pg_{12,y}+L_3g_{23}=0,\quad -Pg_{13,y}+P_{,y}g_{13}+L_3g_{33}=0,$$

$$-Pg_{22,y}-2P_{,y}g_{22}=0,\quad -Pg_{23,y}-P_{,y}g_{23}=0,\quad -Pg_{33,y}=0. \tag{10.10}$$

The last three have the same solutions in all three cases:

$$g_{22}=h_{22}/P^2,\quad g_{23}=h_{23}/P,\quad g_{33}=h_{33}, \tag{10.11}$$

where the $h_{ij}(z)$ in the formulae (10.11)–(10.21) are arbitrary functions. The remaining equations in (10.10) have to be solved separately for each case. In each case it is useful to introduce the new variable $Y(y)$ by:

$$Y_{,y}=1/P. \tag{10.12}$$

Then Eq. (10.3) (with (10.5) taken into account) has the solution:

$$L_3=(\gamma/P)e^{-jY}. \tag{10.13}$$

In order to find P one has to differentiate (10.2) (again with (10.5) taken into account) by Y and use $y_{,Y}=P$. The equation becomes:

$$P_{,YY}+jP_{,Y}+gP=0, \tag{10.14}$$

and y is then found from $y=\int PdY$. The solutions of (10.10) and of (10.14) are now as follows:

Case I: $g<j^2/4$.

$$P=Me^{\mu_1 Y}+Ne^{\mu_2 Y},\quad y=(M/\mu_1)e^{\mu_1 Y}+(N/\mu_2)e^{\mu_2 Y}+y_0, \tag{10.15}$$

where M, N and y_0 are arbitrary constants and

$$\mu_{1,2}=-j/2+\varepsilon_{1,2}(j^2/4-g)^{1/2},\quad \varepsilon_1=1,\quad \varepsilon_2=-1. \tag{10.16}$$

It can be assumed that $MN\neq 0$ because in both the cases $M=0$ and $N=0$ we are back in the domain of Paper II (then $P_{,y}=$ const). The solutions of the remaining Killing equations are here:

$$g_{11}=y^2+h_{11}P^2-2\gamma h_{13}e^{\mu_2 Y}P/[M(\mu_1-\mu_2)]+\gamma^2 h_{33}e^{2\mu_2 Y}/[M(\mu_1-\mu_2)]^2,$$

$$g_{12}=h_{12}-\gamma h_{23}e^{\mu_2 Y}/[M(\mu_1-\mu_2)P],\quad g_{13}=h_{13}P-\gamma h_{33}e^{\mu_2 Y}/[M(\mu_1-\mu_2)]. \tag{10.17}$$

Case II: $g>j^2/4$. Here we define:

$$D=(g-j^2/4)^{1/2},\quad U:=M\ \cos(DY)+N\ \sin(DY),$$

$$V:=M\ \sin(DY)-N\ \cos(DY), \tag{10.18}$$

where M and N are arbitrary constants, and then the solutions are:

$$P=e^{-jY/2}U,\quad y=(4D^2+j^2)^{-1}e^{-jY/2}(4DV-2jU)+y_0,$$

$$g_{11}=y^2+h_{11}P^2+2\gamma h_{13}e^{-jY}UV/[D(M^2+N^2)]+\gamma^2 h_{33}e^{-jY}/[D^2(M^2+N^2)],$$

$$g_{12}=h_{12}+\gamma h_{23}V/[D(M^2+N^2)U],\quad g_{13}=h_{13}P+\gamma h_{33}PV/[D(M^2+N^2)U]. \tag{10.19}$$

Case III: $g=j^2/4\neq 0$. With M,N and y_0 being arbitrary constants, the solutions for P and y are here:

$$P=(MY+N)e^{-jY/2},\quad y=-2P/j-4Me^{-jY/2}j^2+y_0. \tag{10.20}$$

We can assume $M\neq 0$ because with $M=0$ again $P_{,y}=$const. Then, the solutions of the Killing equations are as follows:

$$g_{11}=y^2+h_{11}P^2-2\gamma h_{13}e^{-jY/2}P/M+\gamma^2 h_{33}e^{-jY}/M^2,$$

$$g_{12}=h_{12}-\gamma h_{23}/[M(My+N)],\quad g_{13}=h_{13}P-\gamma h_{33}e^{-jY/2}/M. \tag{10.21}$$

With this, case 2.2.1 is exhausted. We go back to (9.1) and consider:

XI. CASE 2.2.2: $c=0$

With $\Delta=\det A=c=0$ the only element of the matrix A that may be nonzero is e. The solutions of (2.5)–(2.6) are here:

$$\lambda_2=L_2(y),\quad \lambda_3=exL_2+L_3(y), \tag{11.1}$$

$$\phi=axy+F(y),\quad \psi=\tfrac{1}{2}aex^2y+(dy+eC_2+eF)x+P(y). \tag{11.2}$$

Equation (2.4c) now implies that either $j=a$ or $a=0$ or $e=0$. However, $j=a$ leads to a result equivalent to (9.9) and $a=0$ leads to a result equivalent to (10.6). Hence:

$$e=0, \tag{11.3}$$

which makes the whole matrix $A=0$. The remaining implications of (2.4c) and (2.3l) are:

$$ah+dj=0,\quad y(dF_{,y}-aP_{,y}-g)=h(F+C_2)+j(P+C_3),$$

$$y(dL_{2,y}-aL_{3,y})=hL_2+jL_3. \tag{11.4}$$

In solving these conditions, again several cases have to be considered separately.

Case 2.2.2.1: $a\neq 0$. Then:

$$h=-dj/a, \tag{11.5}$$

but in solving the second of (11.4) the case $j=-a$ has to be set aside for separate consideration. We first follow:

Case 2.2.2.1.1: $j\neq -a$. Then:

$$P=-gy/(j+a)+By^{-j/a}-C_3+(d/a)(F+C_2),$$

$$L_3=-\mathscr{A}y^{-j/a}+(d/a)L_2, \tag{11.6}$$

where $\mathscr{A}$ and B are arbitrary constants. In the resulting Killing fields we change the basis to $k'^{\alpha}_{(2)}=a^{-1}k^{\alpha}_{(2)}$, $k'^{\alpha}_{(3)}=k^{\alpha}_{(3)}-(d/a)k^{\alpha}_{(2)}+[g/(j+a)]k^{\alpha}_{(1)}$, and then carry out the transformation (1.13) with:

$$H=(C_2+F)/(ay),\quad T_{,y}=L_2/(ay). \tag{11.7}$$

The final Killing fields are:

$$k^{\alpha}_{(1)}=\delta^{\alpha}_1,\quad k^{\alpha}_{(2)}=x\delta^{\alpha}_1-y\delta^{\alpha}_2,$$

$$k^{\alpha}_{(3)}=y^{-j/a}[B(j/a+1)\delta^{\alpha}_0-Bj(ay)^{-1}\delta^{\alpha}_1-\mathscr{A}\delta^{\alpha}_3]. \tag{11.8}$$

The commutation relations are:

$$[k_{(1)},k_{(2)}]=k_{(1)},\quad [k_{(2)},k_{(3)}]=(j/a)k_{(3)},\quad [k_{(3)},k_{(1)}]=0, \tag{11.9}$$

and they correspond to Bianchi type VI_h with the free parameter $a_B=(j-a)/(j+a)$.

The coordinates will be adapted to $k_{(1)}$ and $k_{(3)}$ after the transformation:

$$t'=aty^{j/a}/[B(j+a)],\quad x'=jt/[(j+a)y]+x,\quad y'=y,\quad z'=\mathscr{A}at+B(j+a)z. \tag{11.10}$$

After the transformation:

$$k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=-jt\delta^\alpha_0+ax\delta^\alpha_1-ay\delta^\alpha_2,\quad k^\alpha_{(3)}=\delta^\alpha_0,$$

$$u^\alpha=(j+a)^{-1}[aB^{-1}y^{j/a}\delta^\alpha_0+(j/y)\delta^\alpha_1]+\mathscr{A}a\delta^\alpha_3,\quad w^\alpha=B(j+a)n\delta^\alpha_3,$$

$$g_{00}=(By^{-j/a})^2[1-(j/a)^2+(j/a)^2g_{11}/y^2+2\mathscr{A}j(j/a+1)g_{13}/y+\mathscr{A}^2(j+a)^2g_{33}],$$

$$g_{01}=By^{-j/a}[(j/a+1)y-(j/a)g_{11}/y-\mathscr{A}(j+a)g_{13}],$$

$$g_{02}=-By^{-j/a}[(j/a)g_{12}/y+\mathscr{A}(j+a)g_{23}],\quad g_{03}=-By^{-j/a}[(j/a)g_{13}/y+\mathscr{A}(j+a)g_{33}], \tag{11.11}$$

and the Killing equations imply:

$$g_{11}=h_{11}y^2,\quad g_{12}=h_{12},\quad g_{13}=h_{13}y,$$

$$g_{22}=h_{22}/y^2,\quad g_{23}=h_{23}/y,\quad g_{33}=h_{33}, \tag{11.12}$$

where the $h_{ij}(z)$ are arbitrary functions.

Case 2.2.2.1.2: $j=-a$. The formulae for h and L_3 simply follow from (11.5) and (11.6) while the second of (11.4) now has a different integral:

$$P=(-g/a)y\ \ln y+By-C_3+(d/a)(F+C_2). \tag{11.13}$$

In the resulting Killing fields we change the basis to $k'^\alpha_{(2)}=a^{-1}k^\alpha_{(2)}$, $k'^\alpha_{(3)}=k^\alpha_{(3)}-(d/a)k^\alpha_{(2)}+(g/a-B)k^\alpha_{(1)}$, and then carry out the transformation (1.13) with the same H and T as in (11.7). The final Killing fields $k^\alpha_{(1)}$ and $k^\alpha_{(2)}$ are the same as in (11.8), and:

$$k^\alpha_{(3)}=y\delta^\alpha_0-\ln y\,\delta^\alpha_1-\mathscr{A}y\delta^\alpha_3. \tag{11.14}$$

The commutation relations are:

$$[k_{(1)},k_{(2)}]=k_{(1)},\quad [k_{(2)},k_{(3)}]=k_{(1)}-k_{(3)},\quad [k_{(3)},k_{(1)}]=0, \tag{11.15}$$

and they correspond to Bianchi type IV.

In order to adapt the coordinates to $k^\alpha_{(1)}$ and $k^\alpha_{(2)}$ we now carry out the transformation:

$$t'=t/y,\quad x'=(t/y)\ln y+x,\quad y'=y,\quad z'=\mathscr{A}t+z, \tag{11.16}$$

and it leads to:

$$k^\alpha_{(1)}=\delta^\alpha_1,\quad k^\alpha_{(2)}=t\delta^\alpha_0+(x-t)\delta^\alpha_1-y\delta^\alpha_2,\quad k^\alpha_{(3)}=\delta^\alpha_0,$$

$$u^\alpha=y^{-1}(\delta^\alpha_0+\ln y\,\delta^\alpha_1)+\mathscr{A}\delta^\alpha_3,\quad w^\alpha=n\delta^\alpha_3,$$

$$g_{00}=y^2-2y^2\ \ln y+(\ln y)^2g_{11}+2\mathscr{A}y\ \ln yg_{13}+(\mathscr{A}y)^2g_{33},$$

$$g_{01}=y^2-\ln yg_{11}-\mathscr{A}yg_{13},\quad g_{02}=-\ln yg_{12}-\mathscr{A}yg_{23},$$

$$g_{03}=-\ln yg_{13}-\mathscr{A}yg_{33}, \tag{11.17}$$

and the Killing equations lead to formulae for the other metric components that are identical to (11.12).

Case 2.2.2.2: $a=0$. Then the first of (11.4) implies that either $d=0$ or $j=0$. However, if $d=0$, then either the algebra of the Killing vectors necessarily becomes two-dimensional (this case is not considered here) or

$$j=0 \tag{11.18}$$

follows anyway. Hence only (11.18) will be considered further. Then, with $0\neq d\neq h$, a result equivalent to (11.8) follows, and when $0\neq d=h$, a result equivalent to case 2.2.2.1.2 follows. Therefore, new results are contained only in the case:

$$d=0. \tag{11.19}$$

Then, from (11.4):

$$h(F+C_2)+gy=0. \tag{11.20}$$

If $h\neq 0$, then (11.20) implies that the combination of the Killing fields $k^\alpha_{(2)}+(g/h)k^\alpha_{(1)}$ is collinear with w^α, and this case is in the domain of Paper II. Hence, from (11.20):

$$h=g=0, \tag{11.21}$$

which leaves us with an Abelian algebra (Bianchi type I) and the functions F, P, L_2 and L_3 being all arbitrary. To avoid landing in the domain of Paper II, we have to assume:

$$F_{,y}\neq 0\neq P_y, \quad C_2+F-yF_{,y}\neq 0\neq C_3+P-yP_{,y}. \tag{11.22}$$

We will denote:

$$W_F:=C_2+F-yF_{,y}, \quad W_P:=C_3+P-yP_{,y}, \tag{11.23}$$

and then the Killing fields are:

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=W_F\delta^\alpha_0+F_{,y}\delta^\alpha_1+L_2\delta^\alpha_3, \quad k^\alpha_{(3)}=W_P\delta^\alpha_0+P_{,y}\delta^\alpha_1+L_3\delta^\alpha_3. \tag{11.24}$$

It may also be assumed that:

$$\mathscr{L}:=L_3-L_2W_P/W_F\neq 0 \tag{11.25}$$

because if $\mathscr{L}=0$, then the Killing equations imply that either the symmetry group becomes two-dimensional or the metric is singular. With $\mathscr{L}\neq 0$, the following transformation is permissible:

$$t'=(L_3t-W_Pz)/(\mathscr{L}W_F), \quad x'=(L_2P_{,y}-L_3F_{,y})t/(\mathscr{L}W_F)+x-(P_{,y}-F_{,y}W_P/W_F)z/\mathscr{L},$$

$$y'=y, \quad z'=-L_2t/(\mathscr{L}W_F)+z/\mathscr{L}, \tag{11.26}$$

after which the Killing fields become:

$$k^\alpha_{(1)}=\delta^\alpha_1, \quad k^\alpha_{(2)}=\delta^\alpha_0, \quad k^\alpha_{(3)}=\delta^\alpha_3. \tag{11.27}$$

The resulting metric depends only on y, but the Einstein equations are hopelessly complicated.

This is the end of the collection of solutions of the Killing equations.

XII. PERSPECTIVES

The research for this series of papers was motivated by the desire to find a rotating (exact) perturbation of the Friedmann–Lemaître cosmological models. Several papers have been published whose authors found exact solutions of Einstein's equations with a rotating source (see a brief overview in Sec. XIII), but all except one of them are either stationary from the beginning or become static in the limit of zero rotation. The one exception is the solution 2 by Stephani[4] that has still nonzero expansion in the limit $\omega\to 0$, but it is a rotating perturbation of a degenerate limit of the hyperbolically symmetric Kantowski–Sachs solution[5] (see Ref. 6 for more on this point) and cannot reproduce any Robertson–Walker geometry.

There is no proof available that a rotating perturbation of the F–L models can be spatially homogeneous. However, if it has any spatially homogeneous subcases and has a dust source, then the subcases must be contained among the metrics listed in this series of papers—just because this is a complete list of all hypersurface-homogeneous rotating dust metrics (and, over all three papers, all Bianchi types appeared on the list, some of them more than once). Hence, from now on, instead of testing various metric ansatzes for rotating dust by trial and error, one can choose an ansatz from a limited collection. It is well-known[7] that the Robertson–Walker geometries have the following relation to the spatially homogeneous Bianchi-type geometries:

The spatially flat ($k=0$) R–W geometry is a common subset of Bianchi types I and VII_0.

The R–W geometry with negative spatial curvature ($k=-1$) is a common subset of Bianchi types V and VII_h.

The R–W geometry with positive spatial curvature ($k=+1$) is a subset of Bianchi type IX geometries.

With this information, the following can be concluded:

(1) The metrics from Paper I do not contain any generalization of the F–L models; the Bianchi type I class contained there has timelike symmetry orbits and the velocity field is tangent to the orbits.
(2) The same is true for the Bianchi type I metrics from Paper II (cases 2.1.2.2 and 2.2) and from the present paper (Eqs. (11.22)–(11.27)). This is in agreement with Theorem 3.1 of King and Ellis[8] which says that no tilted Bianchi type I models exist (tilted means that the velocity field is not orthogonal to the symmetry orbits. Bianchi models with rotation obviously must be tilted).
(3) The Bianchi type V metrics from Paper II (cases 1.2.1 and 1.2.2.2) with suitably chosen parameters do have spacelike symmetry orbits and so may harbour some generalization of the $k=-1$ F–L model.
(4) The Bianchi type VII_h metrics of the present paper [Eqs. (4.23) and (10.19)] may have spacelike symmetry orbits at least on some open subsets of the manifold, i.e., they may contain generalizations of the $k=-1$ F–L model.
(5) The same is true for the Bianchi type IX and type VII_0 metrics of the present paper (Eqs. (5.23) with (5.27) or (5.28)); generalizations of the $k=+1$ and $k=0$ F-L models may be contained there.

Hence, the cases listed in points 3, 4 and 5 are most promising from the point of view of cosmology.

XIII. A BRIEF OVERVIEW OF LITERATURE

Partly in order to justify the claim made in the first paragraph of Sec. XII, a brief overview of literature on solutions of Einstein's equations with rotating sources will be presented here. For the period up to 1973, the overview is based on a thorough survey of subject indexes to *Physics Abstracts* starting with the year 1915 (made in connection with Refs. 57–59). For the period 1973–1996 the survey was less thorough—I was looking for only four keywords: ''Bianchi,'' ''homogeneous,'' ''rotation'' and ''spatially homogeneous.'' In both searches the following sections of *Physics Abstracts* were surveyed: Cosmology, General Relativity, Gravitation, and Space-Time Configurations (of course, references in the papers included in the survey were also checked). Ref. 9 is a more extended survey that covers the period up to 1973, and papers in which perturbative methods were used were also listed in it, these are omitted here. Vacuum solutions are omitted, too.

As a rule, each paper is mentioned in one sentence, so this overview does not pretend to represent the contents of the papers; it is only meant to sort the papers by subjects and serve as an introductory guide through the literature.

Lanczos[10] found the historically earliest exact solution with rotating matter (although he may have been unaware of its rotation); it is a dust solution in which the velocity field and the rotation field are collinear with Killing fields. Van Stockum[11] rediscovered the $\Lambda=0$ subcase of the Lanczos solution; Ref. 11 contains in addition important contributions to the techniques of solving

the Einstein equations for stationary axisymmetric metrics (i.e., with a two-dimensional symmetry group) with a rigidly rotating source. Ref. 12 is another rediscovery of the Lanczos solution, and Refs. 13 and 14 contain discussions of its properties.

Gödel's solution[15] has a five-dimensional symmetry group whose orbit is the whole space–time, and consequently the physical scalars in it are all constant. Its source is dust of zero expansion and shear, and constant matter density and rotation scalars. Rediscoveries of the Gödel solution were published in Refs. 16 and 17 (see Ref. 18). Reference 17 contains in addition a stationary axisymmetric perfect fluid solution.

References 19–32 deal with properties of a metric form that is a modest generalization of the Gödel solution and is known in the literature as the ''Gödel-type metric.'' This notion was introduced in Ref. 19, and in Ref. 20 (which in fact preceded Ref. 19) it was shown that the only Gödel-type metric with a perfect fluid source is the Gödel solution itself. However, various Gödel-type solutions with nonperfect fluid sources have been derived and investigated in Refs. 19 and 21–32.

In Refs. 33–40 solutions of Einstein–Maxwell equations with a charged fluid source were discussed, some of them under the additional assumption that the Lorentz force is zero. The latter are fully within the scope of the formalism used in the present series of papers because the charged dust in them moves with zero acceleration. The relation of the results of these papers to those obtained here was described in Paper I. Some of them are generalizations of the Lanczos[10] and Gödel[15] solutions. Those from Refs. 33–39 are stationary cylindrically symmetric, i.e., are closely related to those from Paper I, the one from Ref. 40 has a two-dimensional symmetry group. The solutions in Ref. 39 are coordinate transforms of those from Ref. 38.

A generalization of the $\Lambda=0$ subcase of the Lanczos solution to a mixture of scalar field and dust was found by Santos and Mondaini.[41]

Other generalizations of the Gödel solution were found in Refs. 42–50. Raval and Vaidya[42] found two solutions with anisotropic pressure, one of them nonstationary. Bray[43] found a collection of solutions with the rotating fluid immersed in a magnetic field. Novello and Rebouças,[44] Kitamura[45–47] and Ray[48] found generalizations with heat flow (the second of Kitamura's solutions presented in Ref. 45 is a coordinate transform of the Gödel solution). Rebouças[49] found a generalization with free electromagnetic field; see Ref. 51 and Paper I for an explanation of the relation between the solutions in Refs. 49 and 37. Finally, Panov[50] found a generalization with two scalar fields, anisotropic fluid, null radiation and heat conduction.

Other solutions with rotating charged matter source were obtained in Refs. 52–55. The Einstein–Maxwell equations for stationary axisymmetric charged dust were analyzed by Bonnor.[56]

In addition, several solutions with a perfect fluid source (i.e., with nonconstant pressure) and with the same Bianchi type I symmetry as was considered in Paper I have been published. These include a family of solutions by this author[57–60] in which the velocity and the rotation fields are collinear with the Killing fields. For the solutions considered in Refs. 57 and 59 the proportionality factor between w^α and the Killing field is explicitly given. For the family of metrics from Ref. 60, the factor is an arbitrary function and the family is defined by a differential equation. The solutions by Davidson[61,62] are explicit examples from the family of Ref. 60, possibly they are also coordinate transforms of members of the family from Ref. 57 defined by certain fixed values of parameters, but Refs. 61 and 62 do not contain sufficient information for precise identification.

Nilsson and Uggla[63] did a qualitative analysis (using the theory of dynamical systems) of perfect fluid solutions with the same symmetry that obey the linear barotropic equation of state.

The list above includes papers that are related to Paper I of this series. Results related to those of Paper II are contained in Refs. 64–77.

Among the solutions found by Ellis[64] there are some that directly belong to the collection of Paper II, they are identified and described in Paper II. Also within the scope of Paper II are the results of King[65] who investigated properties of the subcase $\beta=h_{13}=0$ of case 2.1.2.2 (of Paper II) and provided a few examples of explicit solutions. Other explicit solutions in King's class were found by Maitra,[66] Zimmerman[67] (this reference was not given in Paper II) and Vishveshwara and Winicour.[68]

In Refs. 69–73 rotating dust solutions with four-dimensional symmetry groups were found; at least some of them have three-dimensional subgroups and are within the domain of Paper II.

However, as mentioned in Paper II, these papers do not contain sufficient information for a complete identification of all such subcases. The first three of the six solutions given in Ref. 70 are among the case 2.1.2.2 metrics of Paper II.

Davidson[74] found an example of a solution with King's subclass of symmetry (stationary, cylindrically symmetric, differentially rotating) and with a perfect fluid source obeying the linear barotropic equation of state. Two other stationary cylindrically symmetric perfect fluid solutions were found by Garcia and Kramer,[75] the first is differentially rotating and has the symmetry of King's subclass, the other is rigidly rotating and so has the Bianchi type I symmetry of the class considered in Paper I. Nilsson and Uggla[76] analyzed by the method of dynamical systems the Einstein equations for a metric with a Bianchi type II symmetry and a perfect fluid source that obeys the linear barotropic equation of state. Its set of Killing fields is in the subcase $\lambda_3=0$ of case 1.1.2.2 of Paper II. Stewart and Ellis[77] considered perfect fluid generalizations of the Ellis (dust) solutions from Ref. 64, they also considered sources with anisotropic pressure, viscosity and electric charge.

The only paper directly relevant to the present Paper III is Ref. 3, see Sec. V. The remaining part of the present section is a list of papers in which various problems connected with rotating matter were discussed, but which are more remotely related to the present series of papers.

Stephani[4] found a solution that is unique in one more respect in addition to that mentioned in Sec. XII: it is so far the only rotating matter (dust) solution with no symmetry. (Note: solution 3 in Ref. 4 is not a perfect fluid solution, contrary to the paper's statement.)

Stationary axisymmetric (i.e., with two-dimensional symmetry groups) perfect fluid solutions were found in Refs. 78–83. Wahlquist[78] found a rigidly rotating solution with the equation of state $\epsilon=-3p+\text{const}$ that, with specific values of two parameters, can describe the interior of a body with compact outer surface. Herlt[79] found a source of the NUT vacuum solution and Kramer[80] found another solution, both are rigidly rotating (Kramer's solution was rediscovered by Patra and Roy,[81] see Ref. 82). Another class of rigidly rotating stationary axisymmetric solutions was found by Herlt.[83]

Several stationary axisymmetric metrics were devised as nonperfect fluid sources of the Kerr solution. Results in this class that were published up to 1976 are reviewed in Ref. 84. Later, a few more papers on this subject were published, but in those that are known to the present author the source is either a surface distribution of matter or an energy-momentum tensor that does not correspond to any identifiable kind of matter, hence they are not mentioned here.

A spatially homogeneous solution of Bianchi type VI_0 with a rotating perfect fluid source was found by Rosquist;[85] it has the equation of state $\epsilon=3p$ and nonzero expansion. Spatially homogeneous solutions with heat-conducting sources were found in Refs. 86–89. Two other hypersurface-homogeneous perfect fluid solutions were found by Wainwright,[90] one has a three-dimensional symmetry group of unidentified Bianchi type, the other has a four-dimensional multiply transitive symmetry group.

The remaining papers (Refs. 91–116) contain results obtained without finding explicit solutions of Einstein's equations. This part of the survey is likely to be incomplete.

Narlikar[91] proposed a metric form for a rotating dust model in which dust particles move on 3-cylinders (axial symmetry is not assumed) whose orthogonal sections are two-dimensional surfaces of constant curvature. Winicour[98] reduced the Einstein equations with a stationary axisymmetric dust source to a sequence of integrations. This author[101] presented results of partial integration of the Einstein equations for a cylindrically symmetric nonstationary perfect fluid.

Bampi and Cianci[102] investigated spacetimes with an Abelian two-dimensional group of symmetries that has null orbits; the example of exact solution provided is a vacuum. Wils and van den Bergh[104] showed that a stationary axisymmetric differentially rotating charged dust has either a nonvanishing Lorentz force of a nonconstant ratio of charge density to mass density.

In Refs. 105–110 general properties of rotating spatially homogeneous Bianchi type IX models were investigated without attempting to solve the Einstein equations. Of these, Ref. 109 gives a kinetic theory description of sources in such models and Ref. 110 gives a qualitative analysis of rotating mixmaster models.

References 111–113 contain general considerations about rotating matter models, and Refs. 8 and 114 contain more general results, applicable also to nonrotating models, but having consequences for rotating models as well. Collins[111] considered properties of shearfree rotating perfect

fluids. Again Collins[112] reviewed arguments for the hypothesis that shear being zero implies that either rotation or expansion is zero (apart from several specific examples, this is still unproven). Mason and Pooe[113] investigated properties of the Lie derivative of the rotation vector along the velocity vector in rigidly rotating matter. King and Ellis[8] investigated properties of tilted spatially homogeneous models and Nilsson and Uggla[114] investigated hypersurface homogeneous and hypersurface self-similar perfect fluid models by the method of dynamical systems.

Hawking[115] and Collins and Hawking[116] investigated limits set on the rotation parameter in Bianchi-type models by the observations of the CMB radiation.

Finally, there is a paper in the domain of the history of science. Ellis[117] described the influence of Gödel's ideas presented in Refs. 15 and 105 on the development of several concepts and research programs in relativity, such as, among other things, the Bianchi-type models, singularity theorems and causal structure of space–time.

ACKNOWLEDGMENT

Calculations for this paper were done with use of the algebraic computer program ORTOCARTAN.[118,119]

Note added in proof: The Gödel-type metric of Refs. 19–32 was discussed also in Ref. 120. Other rotating solutions of Einstein–Maxwell equations were discussed in Refs. 121 and 122. I am grateful to M. J. Rebouças and A. Georgiou for this information.

APPENDIX A: THE RESULT OF TRANSFORMATION (4.21) ON THE METRIC

The coordinate transformation (4.21) changes the metric tensor from the Plebański form (1.9) to one in which the following relations hold (primes dropped, all the components displayed are expressed in the new coordinates):

$$\begin{aligned}
g_{00} &= e^{(b+f)x}[\sin^2(Dx/2)-(V/W)^2\cos^2(Dx/2)]+(V/W)^2 g_{22}\\
&\quad +2(V/W)\gamma e^{(b+f)x/2}[\cos(Dx/2)+(V/W)\sin(Dx/2)]g_{23}\\
&\quad +\gamma^2 e^{(b+f)x}[\cos(Dx/2)+(V/W)\sin(Dx/2)]^2 g_{33},\\
g_{01} &= (V/W)g_{12}+\gamma e^{(b+f)x/2}[\cos(Dx/2)+(V/W)\sin(Dx/2)]g_{13},\\
g_{02} &= e^{(b+f)x}\cos(Dx/2)[\sin(Dx/2)-(V/W)\cos(Dx/2)]+(V/W)g_{22}\\
&\quad +\gamma e^{(b+f)x/2}[\cos(Dx/2)+(V/W)\sin(Dx/2)]g_{23},\\
g_{03} &= (V/W)g_{23}+\gamma e^{(b+f)x/2}[\cos(Dx/2)+(V/W)\sin(Dx/2)]g_{33}.
\end{aligned}$$

APPENDIX B: THE EQUIVALENCE OF CASE 1.2 AND CASE 1.1.1

These are the main points of the reasoning from (2.8) on in the case $c=0$. When $c=0$, Eqs. (2.7)–(2.8) imply:

$$\alpha_1=b, \quad \alpha_2=f, \tag{B1}$$

and $b\neq f$ since $\Delta\neq 0$. The calculation requires checking several cases separately.

Case 1.2.1: det $A\neq 0$. The formulae corresponding to (2.9) and (2.13) are then:

$$\begin{aligned}
&\lambda_2=L_2(y)e^{bx}, \quad \lambda_3=[e/(b-f)]L_2(y)e^{bx}+L_3(y)e^{fx}, \quad \phi=F(y)e^{bx}-ay/b-C_2,\\
&\psi=[e/(b-f)]F(y)e^{bx}+P(y)e^{fx}-[-ae/(bf)+d/f]y-C_3.
\end{aligned} \tag{B2}$$

In verifying (2.4c) and (2.3l) the cases $F=0$ and $P=0$ have to be considered separately. The results are as follows:

When $F=0\neq P$ and $L_2=0$ the group becomes two-dimensional.

When $F=0\neq P$ and $L_2\neq 0$, the Killing field $[k^\alpha_{(2)}+(a/b)k^\alpha_{(1)}]$ is collinear with w^α, and this case is in the domain of Paper II.

When $P=0$, another linear combination of the Killing vectors with constant coefficients is collinear with w^α.

When $F\neq 0\neq P$, formulae equivalent to (2.24)–(2.29) result.

Case 1.2.2: det $A=0$. Two different (but equivalent) subcases have to be considered here: $b=0$ and $f=0$ ($b\neq f$ because of the assumption $\Delta\neq 0$).

When $f=0$, the functions λ_2,λ_3 and ϕ are given by the limit $f=0$ of (B.2), while ψ is:

$$\psi=(e/b)F(y)e^{bx}+P(y)-aey/b^2+(d-ae/b)xy-C_3. \tag{B3}$$

Equations (2.4c) and (2.3l) show then that there always exists a linear combination of the Killing fields with constant coefficients which is spanned on u^α and w^α.

When $b=0$, the conclusion is the same, only the functions in (B.2) are different at the starting point. λ_2 and λ_3 are the limits $b=0$ of those from (B.2), and:

$$\phi=F(y)+axy-C_2,$$

$$\psi=-(e/f)(\phi+C_2)+P(y)e^{fx}-(d/f+ae/f^2)y-C_3. \tag{B4}$$

[1] A. Krasiński, J. Math. Phys. **39**, 380 (1998) (Paper I).
[2] A. Krasiński, J. Math. Phys. **39**, 401 (1998) (Paper II).
[3] M. Demiański and L. P. Grishchuk, Commun. Math. Phys. **25**, 233 (1972).
[4] H. Stephani, Class. Quantum Grav. **4**, 125 (1987).
[5] R. Kantowski and R. K. Sachs, J. Math. Phys. **7**, 443 (1966).
[6] A. Krasiński, *Inhomogeneous Cosmological Models* (Cambridge University Press, Cambidge, 1997).
[7] L. P. Grishchuk, Astron. Zh. **44**, 1097 (1967) [Sov. Astron. **11**, 881 (1968)].
[8] A. R. King and G. F. R. Ellis, Commun. Math. Phys. **31**, 209 (1973).
[9] A. Krasiński, *Solutions of the Einstein field equations for a rotating perfect fluid; Part 3: A survey of models of rotating perfect fluid or dust* (preprint), Warszawa, 1975.
[10] K. Lanczos, Z. Phys. **21**, 73 (1924); English translation in Gen. Relativ. Gravit. **29**, 363 (1997).
[11] W. J. van Stockum, Proc. R. Soc. Edinburgh **57**, 135 (1937).
[12] J. P. Wright, J. Math. Phys. **6**, 103 (1965).
[13] W. B. Bonnor, J. Phys. A **13**, 2121 (1980).
[14] P. R. C. T. Pereira, N. O. Santos, and A. Z. Wang, Class. Quantum Grav. **13**, 1641 (1996).
[15] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[16] C. Hoenselaers and C. V. Vishveshwara, Gen. Relativ. Gravit. **10**, 43 (1979).
[17] A. H. Khater and M. F. Mourad, Astrophys. Space Sci. **163**, 247 (1990).
[18] S. K. Chakraborty, Gen. Relativ. Gravit. **12**, 925 (1980).
[19] M. Novello and M. J. Rebouças, Phys. Rev. D **19**, 2850 (1979).
[20] F. Bampi and C. Zordan, Gen. Relativ. Gravit. **9**, 393 (1978).
[21] M. J. Rebouças and J. Tiomno, Phys. Rev. D **28**, 1251 (1983).
[22] M. J. Rebouças and J. Tiomno, Nuovo Cimento B **90**, 204 (1985).
[23] A. F. F. Teixeira, M. J. Rebouças and J. E. Åman, Phys. Rev. D **32**, 3309 (1985).
[24] J. D. Oliveira, A. F. F. Teixeira, and J. Tiomno, Phys. Rev. D **34**, 3661 (1986).
[25] M. J. Rebouças, J. E. Åman, and A. F. F. Teixeira, J. Math. Phys. **27**, 1370 (1986).
[26] M. J. Rebouças and A. F. F. Teixeira, Phys. Rev. D **34**, 2985 (1986).
[27] M. J. Rebouças and A. F. F. Teixeira, in *Proceedings of the 4th Marcel Grossman Meeting on General Relativity*, edited by R. Ruffini (North–Holland, Amsterdam, 1986), Vol. 2, p. 1025.
[28] M. J. Rebouças and J. E. Åman, J. Math. Phys. **28**, 888 (1987).
[29] M. O. Calvão, M. J. Rebouças, A. F. F. Teixeira, and W. M. Silva, J. Math. Phys. **29**, 1127 (1988).
[30] V. A. Korotkii and Yu. N. Obukhov, Vestn. Mosk. Univ. Fiz. Astron. **32**, No. 2,3 (1991) [Moscow Univ. Phys. Bull. No. 2, 4 (1991)].
[31] V. F. Panov, Izv. VUZ Fiz. **34**, No. 2, 54 (1991) [Sov. Phys. J. **34**, 130 (1991)].
[32] R. X. Saibatalov, Gen. Relativ. Gravit. **27**, 697 (1997).
[33] J. N. Islam, Proc. R. Soc. London, Ser. A **385**, 189 (1983).
[34] P. Wils and N. van den Bergh, Proc. R. Soc. London, Ser. A **394**, 437 (1984).
[35] A. Gerogiou, Nuovo Cimento B **108**, 69 (1993).
[36] M. M. Som and A. K. Raychaudhuri, Proc. R. Soc. London, Ser. A **304**, 81 (1968).
[37] A. Banerjee and S. Banerji, J. Phys. A **1**, 188 (1968).
[38] N. V. Mitskevič and G. A. Tsalakou, Class. Quantum Grav. **8**, 209 (1991).
[39] A. M. Upornikov, Class. Quantum Grav. **11**, 2085 (1994).
[40] J. N. Islam, Proc. R. Soc. London, Ser. A **353**, 523 (1977).
[41] N. O. Santos and R. P. Mondaini, Nuovo Cimento B **72**, 13 (1982).
[42] H. M. Raval and P. C. Vaidya, Ann. Inst. Henri Poincaré, Sect. A **4**, 21 (1966).
[43] M. Bray, C. R. Hebd. Acad. Sci. (Paris) A **274**, 809 (1972); **274**, 874 (1972).

[44] M. Novello and M. J. Rebouças, Astrophys. J. **225**, 719 (1978).
[45] S. Kitamura, Tensor N. S. **32**, 156 (1978).
[46] S. Kitamura, Math. Japanica **23**, 655 (1979).
[47] S. Kitamura, Tensor N. S. **35**, 183 (1981).
[48] D. Ray, J. Math. Phys. **21**, 2797 (1980).
[49] M. J. Rebouças, Phys. Lett. **70A**, 161 (1979).
[50] V. F. Panov, Izv. VUZ Fiz. **33** (No. 1), 62 (1990) [Sov. Phys. J. **33**, 54 (1990)].
[51] A. K. Raychaudhuri and S. N. Guha Thakurta, Phys. Rev. D **22**, 802 (1980).
[52] J. N. Islam, Proc. R. Soc. London, Ser. A **362**, 329 (1978).
[53] J. N. Islam, Proc. R. Soc. London, Ser. A **367**, 271 (1979).
[54] J. N. Islam, Proc. R. Soc. London, Ser. A **372**, 111 (1980).
[55] M. J. Rebouças and J. B. S. d'Olival, J. Math. Phys. **27**, 417 (1986).
[56] W. B. Bonnor, J. Phys. A **13**, 3465 (1980).
[57] A. Krasiński, Acta Phys. Pol. B **5**, 411 (1974).
[58] A. Krasiński, Acta Phys. Pol. B **6**, 223 (1975).
[59] A. Krasiński, J. Math. Phys. **16**, 125 (1975).
[60] A. Krasiński, Rep. Math. Phys. **14**, 225 (1978).
[61] W. Davidson, Class. Quantum Grav. **13**, 283 (1996).
[62] W. Davidson, Class. Quantum Grav. **14**, 119 (1997).
[63] U. S. Nilsson and C. Uggla, Class. Quantum. Grav. **14**, 2931 (1997).
[64] G. F. R. Ellis, J. Math. Phys. **8**, 1171 (1967).
[65] A. R. King, Commun. Math. Phys. **38**, 157 (1974).
[66] S. C. Maitra, J. Math. Phys. **7**, 1025 (1966).
[67] J. C. Zimmerman, J. Math. Phys. **16**, 2458 (1975).
[68] C. V. Vishveshwara and J. Winicour, J. Math. Phys. **18**, 1280 (1977).
[69] I. Ozsváth and E. Schucking, Nature (London) **193**, 1168 (1962).
[70] I. Ozsváth, J. Math. Phys. **6**, 590 (1965).
[71] D. L. Farnsworth and R. P. Kerr, J. Math. Phys. **7**, 1625 (1966).
[72] I. Ozváth and E. Schucking, Ann. Phys. (N.Y.) **55**, 166 (1969).
[73] I. Ozsváth, J. Math. Phys. **11**, 2871 (1970).
[74] W. Davidson, Class. Quantum Grav. **11**, L129 (1994).
[75] A. Garcia and D. Kramer, Class. Quantum Grav. **14**, 499 (1997).
[76] U. S. Nilsson and C. Uggla, J. Math. Phys. **38**, 2611 (1997).
[77] J. Stewart and G. F. R. Ellis, J. Math. Phys. **9**, 1072 (1968).
[78] H. D. Wahlquist, Phys. Rev. **172**, 1291 (1968).
[79] E. Herlt, Wiss. Z. Friedrich Schiller Univ. Jena, Math. Naturwiss. Reihe **21**, 19 (1972).
[80] D. Kramer, Class. Quantum Grav. **1**, L3 (1984).
[81] A. C. Patra and D. Roy, Class. Quantum Grav. **4**, 195 (1987).
[82] H. Stephani, Class. Quantum Grav. **4**, 1047 (1987).
[83] E. Herlt, Gen. Relativ. Gravit. **20**, 635 (1988).
[84] A. Krasiński, Ann. Phys. (N.Y.) **112**, 22 (1978).
[85] K. Rosquist, Phys. Lett. **97A**, 145 (1983).
[86] M. J. Rebouças and J. A. S. de Lima, J. Math. Phys. **22**, 2699 (1981).
[87] M. J. Rebouças, Nuovo Cimento B **67**, 120 (1982).
[88] J. M. Bradley and E. Sviestins, Gen. Relativ. Gravit. **16**, 1119 (1984).
[89] V. F. Panov, Izv VUZ Fiz. **32** (No. 5), 98 (1989) [Sov. Phys. J. **32**, 403 (1989)].
[90] J. Wainwright, Commun. Math. Phys. **17**, 42 (1970).
[91] J. V. Narlikar, in *Evidence for gravitational theories* in *Proceedings of the International School of Physics ''Enrico Fermi,'' Course 20*, edited by C. Møller (Academic, New York, 1962), p. 222.
[92] W. Kundt and M. Trümper, Z. Phys. **192**, 419 (1966).
[93] M. Trümper, Z. Naturforsch. A **22a**, 1347 (1967).
[94] B. G. Schmidt, Z. Naturforsch. A **22a**, 1351 (1967).
[95] B. K. Harrison, Phys. Rev. D **1**, 2269 (1970).
[96] J. M. Bardeen, Astrophys. J. **162**, 71 (1970).
[97] J. Pachner and M. J. Miketinac, Phys. Rev. D **6**, 1479 (1972).
[98] J. Winicour, J. Math. Phys. **16**, 1806 (1975).
[99] E. N. Glass, J. Math. Phys. **18**, 708 (1977).
[100] E. N. Glass and D. A. Wilkinson, Gen. Relativ. Gravit. **9**, 531 (1978).
[101] A. Krasiński, Acta Cosmologica **7**, 133 (1978).
[102] F. Bampi and R. Cianci, Commun. Math. Phys. **70**, 69 (1979).
[103] G. T. Carlson and J. L. Safko, Ann. Phys. (N.Y.) **128**, 131 (1980).
[104] P. Wils and N. van den Bergh, Class. Quantum Grav. **1**, 399 (1984).
[105] K. Gödel, in *Proceedings of the International Congress of Mathematicians*, edited by L. M. Graves *et al.* (Cambridge University Press, Massachusetts, 1952), Vol. 1, p. 175.
[106] R. A. Matzner, L. C. Shepley, and J. B. Warren, Ann. Phys. (N.Y.) **57**, 401 (1970).
[107] R. A. Matzner, J. Math. Phys. **11**, 2432 (1970).
[108] V. A. Korotkii and Yu. N. Obukhov, Zh. Eksp. Teor. Fiz. **99**, 24 (1991).
[109] R. A. Matzner, Ann. Phys. (N.Y.) **65**, 438 (1971).
[110] R. A. Matzner and D. M. Chitre, Commun. Math. Phys. **22**, 173 (1973).
[111] C. B. Collins, J. Math. Phys. **25**, 995 (1984).
[112] C. B. Collins, Can. J. Phys. **64**, 191 (1986).

[113] D. Mason and C. A. Pooe, J. Math. Phys. **28**, 2705 (1987).
[114] U. S. Nilsson anmd C. Uggla, Class. Quantum Grav. **14**, 1965 (1997).
[115] S. W. Hawking, Mon. Not. R. Astron. Soc. **142**, 129 (1969).
[116] C. B. Collins and S. W. Hawking, Mon. Not. R. Astron. Soc. **162**, 307 (1973).
[117] G. F. R. Ellis, *Contributions of K. Göbel to relativity and cosmology* (preprint), Cape Town, 1996.
[118] A. Krasiński, Gen. Relativ. Gravit. **25**, 165 (1993).
[119] A. Krasiński and M. Perkowski, Gen. Relativ. Gravit. **13**, 67 (1981).
[120] K. Dunn, Gen. Relativ. Gravit. **21**, 137 (1989).
[121] A. Georgiou, Class. Quantum Grav. **11**, 167 (1994).
[122] A. Georgiou, Class. Quantum Grav. **13**, 791 (1996).

Riemannian space-times of Gödel type in five dimensions

M. J. Rebouças[a] and A. F. F. Teixeira[b]
Centro Brasileiro de Pesquisas Físicas, Departamento de Relatividade e Partículas, Rua Dr. Xavier Sigaud 150, 22290-180 Rio de Janeiro-RJ, Brazil

(Received 20 May 1997; accepted for publication 11 December 1997)

The five-dimensional (5D) Riemannian Gödel-type manifolds are examined in light of the equivalence problem techniques, as formulated by Cartan. The necessary and sufficient conditions for local homogeneity of these 5D manifolds are derived. The local equivalence of these homogeneous Riemannian manifolds is studied. It is found that they are characterized by two essential parameters m^2 and ω: identical pairs (m^2,ω) correspond to locally equivalent 5D manifolds. An irreducible set of isometrically nonequivalent 5D locally homogeneous Riemannian Gödel-type metrics are exhibited. A classification of these manifolds based on the essential parameters is presented, and the Killing vector fields as well as the corresponding Lie algebra of each class are determined. It is shown that apart from the $(m^2=4\omega^2,\omega\neq 0)$ and $(m^2\neq 0,\omega=0)$ classes the homogeneous Riemannian Gödel-type manifolds admit a seven-parameter maximal group of isometry (G_7). The special class $(m^2=4\omega^2,\omega\neq 0)$ and the degenerated Gödel-type class $(m^2\neq 0,\omega=0)$ are shown to have a G_9 as maximal group of motion. The breakdown of causality in these classes of homogeneous Gödel-type manifolds are also examined.
© *1998 American Institute of Physics.* [S0022-2488(98)03703-7]

I. INTRODUCTION

Kaluza–Klein-type theories in five and more dimensions have been of notable interest in several contexts. In the framework of gauge theories they have been used in the quest for unification of the fundamental interactions in physics. The idea that the various interactions in nature might be unified by enlarging the dimensionality of the space-time has a long and honorable history that goes back to the works of Nordström, Kaluza, and Klein.[1–3] Its earlier adherents were mainly interested in extending general relativity, but a late increased interest has been apparent in the particle physics community, especially among those investigating supersymmetry.

The possibility that space-time has more than four dimensions has also received much attention regarding its cosmological aspects. Investigation has been focused on attempts to understand, for example, why the universe presently appears to have only four space-time dimensions, and whether it is a higher-dimensional dynamically evolving manifold, the space-time expands while the extra dimensions contract or remain constant. The first cosmological model in which the extra dimension contracts as a result of the cosmological evolution was proposed by Chodos and Detweiler.[4] Since then, a great deal of work has emerged along this line of research, particularly regarding exact solutions of Einstein's equations, entropy production during the contracting process,[5] resolution of the horizon and flatness problems,[6] and the like (see Ref. 7 and references therein).

From a purely technical viewpoint higher-dimensional Kaluza–Klein-type theories have also been used as a way of finding new exact solutions of Einstein's equations in four dimensions.[8–10]

More recently Wesson[11–13] has given a new impetus to the study of (4+1)-gravity by investigating a five-dimensional extension of general relativity with a variable rest mass (see also Refs. 14 and 15). In this theory we have space-time-mass (STM) Riemannian manifolds, and the fifth dimension is a convenient mathematical way of geometrizing the rest mass and of allowing one to study the possibility that it may be variable. The four-dimensional (4D) general relativity theory is recovered when the rate of change of the rest mass is zero. Ever since the foundations of this

[a]Electronic mail: reboucas@cat.cbpf.br
[b]Electronic mail: teixeira@novell.cat.cbpf.br

 © 1998 American Institute of Physics

Kaluza–Klein-type STM theory were laid, there have been investigations of its potentialities and physical consequences, particularly as concerns its consistency with Mach's principle,[16] causality conditions and inflation,[14,17] confrontation between theory and observation,[18,19] and the solutions of its field equations.[15]

In 1949 Gödel found a solution of Einstein's field equations with cosmological constant for incoherent matter with rotation.[20] Owing to its striking properties, the cosmological solution presented by Gödel has a well-recognized importance and has to a large extent motivated the investigations on rotating cosmological space-times within the framework of general relativity. Particularly, the search for rotating Gödel-type space-times has received a good deal of attention in recent years, and the literature on these geometries is fairly large today (see Refs. 21–34 and references therein).

However, the general problem of space-time local homogeneity (ST homogeneity, hereafter[35]) of four-dimensional Riemannian manifolds endowed with a Gödel-type metric was considered for the first time only in 1980 by Raychaudhuri and Thakurta.[36] They found the necessary conditions for ST homogeneity of these manifolds. Three years later, Rebouças and Tiomno[37] proved that Raychaudhuri–Thakurta necessary conditions are also sufficient for ST homogeneity of Gödel-type manifolds. However, in both articles[36,37] the study of ST homogeneity is limited in that only time-independent Killing vector fields were considered.[38] The Raychaudhuri–Thakurta–Rebouças–Tiomno conditions were finally proved to be the necessary and sufficient conditions for a Gödel-type manifold to be ST homogeneous without assuming any such simplifying hypothesis in Ref. 39, where the powerful equivalence problem techniques for Riemannian space-times, as formulated in terms of spinors by Karlhede[40] and embodied in a suite of computer algebra programs called CLASSI[41] written in SHEEP,[42] were used.

In light of the equivalence problem techniques, as formulated by Cartan[43] and using CLASSI, we extend these investigations by examining a class of five-dimensional (5D) Riemannian Gödel-type manifolds. The necessary and sufficient conditions for local homogeneity of these 5D manifolds are derived. The local equivalence of these homogeneous space-time-mass Riemannian manifolds is discussed; they are found to be characterized by two essential parameters m^2 and ω: identical pairs (m^2,ω) correspond to locally equivalent STM manifolds. An irreducible set of isometrically nonequivalent 5D homogeneous Riemannian Gödel-type metrics are exhibited. A classification of these manifolds based on the essential parameters is presented, and the Killing vector fields as well as the corresponding Lie algebra of each class are determined. We show that apart from the $(m^2=4\omega^2,\omega\neq 0)$ and $(m^2\neq 0,\omega=0)$ classes the 5D locally homogeneous Riemannian Gödel-type manifolds have a seven-parameter maximal group of isometry (G_7). The special class $(m^2=4\omega^2,\omega\neq 0)$ and the degenerated Gödel-type class $(m^2\neq 0,\omega=0)$ are shown to admit a G_9 as maximal group of motion. The breakdown of causality in these classes of homogeneous Gödel-type manifolds is also examined and shown to have the same basic features of the corresponding 4D counterparts.

To close this Introduction, we should like to emphasize that although the STM Wesson's theory of gravitation is often referred to, the results of the following sections hold for any five-dimensional Riemannian Gödel-type manifold regardless of the underlying 5D Kaluza–Klein-type theory[44–47] one may be concerned with.

II. HOMOGENEITY, IRREDUCIBLE SET, AND CAUSALITY

The arbitrariness in the choice of coordinates in the geometric theories of gravitation gives rise to the problem of deciding whether or not two Riemannian manifolds whose metrics g and $\tilde{g}$ are given explicitly in terms of coordinates, viz.,

$$ds^2=g_{\mu\nu}dx^\mu dx^\nu \quad \text{and} \quad d\tilde{s}^2=\tilde{g}_{\mu\nu}d\tilde{x}^\mu d\tilde{x}^\nu, \tag{2.1}$$

are locally the same. This is the so-called equivalence problem (see Cartan[43] for the local equivalence of n-dimensional Riemannian manifolds, Karlhede[40] and MacCallum[42,48,49] for the special case $n=4$ of general relativity).

The Cartan solution[43] to the equivalence problem for Riemannian manifolds can be reworded as follows. Two n-dimensional Lorentzian Riemannian manifolds M and $\tilde{M}$ are locally equivalent

if there exist coordinate and generalized n-dimensional Lorentz transformations such that the following *algebraic* equations relating the frame components of the curvature tensor and their covariant derivatives:

$$
\begin{aligned}
R^{A}{}_{BCD} &= \widetilde{R}^{A}{}_{BCD}, \\
R^{A}{}_{BCD;M_1} &= \widetilde{R}^{A}{}_{BCD;M_1}, \\
R^{A}{}_{BCD;M_1M_2} &= \widetilde{R}^{A}{}_{BCD;M_1M_2}, \\
&\vdots \\
R^{A}{}_{BCD;M_1\ldots M_{p+1}} &= \widetilde{R}^{A}{}_{BCD;M_1\ldots M_{p+1}}
\end{aligned}
\tag{2.2}
$$

are compatible as equations in (x^{μ},ξ^{A}). Here and in what follows we use a semicolon to denote covariant derivatives. Note that x^{μ} are coordinates on the manifold M while ξ^{A} parametrize the group of allowed frame transformations [n-dimensional generalized Lorentz group usually denoted[50] by $O(n-1,1)$]. Reciprocally, Eqs. (2.2) imply local equivalence between the n-dimensional manifolds M and $\widetilde{M}$.

In practice, a fixed frame is chosen to perform the calculations so that only coordinates appear in the components of the curvature tensor, i.e., there is no explicit dependence on the parameters ξ^{A} of the generalized Lorentz group.

Another important practical point to be considered, once one wishes to test the local equivalence of two Riemannian manifolds, is that before attempting to solve Eqs. (2.2) one can extract and compare partial pieces of information at each step of differentiation as, for example, the number $\{t_0,t_1,\ldots,t_p\}$ of functionally independent functions of the coordinates x^{μ} contained in the corresponding set

$$
I_p=\{R^{A}{}_{BCD},R^{A}{}_{BCD;M_1},R^{A}{}_{BCD;M_1M_2},\ldots,R^{A}{}_{BCD;M_1M_2\ldots M_p}\},
\tag{2.3}
$$

and the isotropy subgroup $\{H_0,H_1,\ldots,H_p\}$ of the symmetry group G_r under which the corresponding set I_p is invariant. They must be the same for each step $q=0,1,\ldots,p$ if the manifolds are locally equivalent.

In practice it is also important to note that in calculating the curvature and its covariant derivatives, in a chosen frame, one can stop as soon as one reaches a step at which the pth derivatives (say) are algebraically expressible in terms of the previous ones, and the residual isotropy group (residual frame freedom) at that step is the same isotropy group of the previous step, i.e., $H_p=H_{(p-1)}$. In this case further differentiation will not yield any new piece of information. Actually, if $H_p=H_{(p-1)}$ and, in a given frame, the pth derivative is expressible in terms of its predecessors, for any $q>p$ the qth derivatives can all be expressed in terms of the 0th, 1st, $\ldots$,$(p-1)$th derivatives.[43,51,42] As in the worst case we have only one functionally independent function of the coordinates x^{μ} at each step of the differentiation process, and the generalized Lorentz group has $n(n-1)/2$ independent parameters, it follows that for five-dimensional Riemannian manifolds $p+1\leqslant 15$.

Since there are t_p essential coordinates, in 5D clearly $5-t_p$ are ignorable, so the isotropy group will have dimension $s=\dim(H_p)$, and the group of isometries of the metric will have dimension r given by (see Cartan[43])

$$
r=s+5-t_p,
\tag{2.4}
$$

acting on an orbit with dimension

$$
d=r-s=5-t_p.
\tag{2.5}
$$

The line element of the five-dimensional Riemannian Gödel-type manifolds M_5 we are concerned with in this work is given by

$$ds^2 = dt^2 + 2H(x)dtdy - dx^2 - G(x)dy^2 - dz^2 - du^2, \tag{2.6}$$

where $H(x)$ and $G(x)$ are arbitrary real functions of x, and the five STM coordinates clearly are t,x,y,z,u. As a matter of fact, to ensure the local Lorentzian character of (2.6) one has to require that $H^2(x)+G(x)>0$. At an arbitrary point of M_5 one can choose the following set of linearly independent one-forms θ^A:

$$\theta^0 = dt + H(x)dy, \quad \theta^1 = dx, \quad \theta^2 = D(x)dy, \qquad \theta^3 = dz, \quad \theta^4 = du, \tag{2.7}$$

such that the Gödel-type line element (2.6) can be written as

$$ds^2 = \eta_{AB}\theta^A\theta^B = (\theta^0)^2 - (\theta^1)^2 - (\theta^2)^2 - (\theta^3)^2 - (\theta^4)^2, \tag{2.8}$$

where $D^2(x) = G + H^2$. Here and in what follows capital letters are pentad indices (or Lorentz frame indices) and run from 0 to 4; they are raised and lowered with Lorentz matrices $\eta^{AB} = \eta_{AB} = \text{diag}(+1,-1,-1,-1,-1)$, respectively.

Using as input the one-forms (2.7) and the Lorentz frame (2.8), the computer algebra package CLASSI gives the following nonvanishing Lorentz frame components R_{ABCD} of the curvature:

$$R_{0101} = R_{0202} = -\frac{1}{4}\left(\frac{H'}{D}\right)^2, \tag{2.9}$$

$$R_{0112} = \frac{1}{2}\left(\frac{H'}{D}\right)', \tag{2.10}$$

$$R_{1212} = \frac{D''}{D} - \frac{3}{4}\left(\frac{H'}{D}\right)^2, \tag{2.11}$$

where the prime denotes derivative with respect to x.

For STM homogeneity from Eq. (2.5) one must have $t_q = 0$ for $q = 0,1,\ldots,p$, that is, the number of functionally independent functions of x^μ in the set I_p must be zero. Therefore from Eqs. (2.9)–(2.11) we conclude that for STM homogeneity it is necessary that

$$\frac{H'}{D} = \text{const} \equiv -2\omega, \tag{2.12}$$

$$\frac{D''}{D} = \text{const} \equiv m^2. \tag{2.13}$$

We shall now show that the above necessary conditions are also sufficient for STM local homogeneity. Indeed, under these conditions the nonvanishing frame components of the curvature reduce to

$$R_{0101} = R_{0202} = -\omega^2, \tag{2.14}$$

$$R_{1212} = m^2 - 3\omega^2. \tag{2.15}$$

Following Cartan's method for the local equivalence, we next calculate the first covariant derivative of the Riemann tensor. Now one obtains the following non-null covariant derivatives of the curvature:

$$R_{0112;1} = R_{0212;2} = \omega(m^2 - 4\omega^2). \tag{2.16}$$

As the first covariant derivative of the curvature is algebraically expressible in terms of the Riemann tensor and for a given pair (m^2,ω) the isotropy group H_1 is the same as H_0 (see next paragraph), no new covariant derivative of the curvature should be calculated. Clearly one also has $t_0=t_1=0$.

As far as the dimension of the residual isotropy group is concerned we distinguish two different classes of locally homogeneous 5D Gödel-type Riemannian manifolds, according to the relevant parameters m^2 and ω, namely[52]

(i) $m^2\neq 4\omega^2$ and $\omega\neq 0$ for which $\dim(H_0)=\dim(H_1)=2$;

(ii) $m^2=4\omega^2$ with $\omega\neq 0$, and the degenerated Gödel-type manifolds $m^2\neq 0$ and $\omega=0$. Here one has $\dim(H_0)=\dim(H_1)=4$.

Thus from Eqs. (2.4) and (2.5) one finds that the locally homogeneous 5D Riemannian Gödel-type manifolds admit a (local) G_r, with either $r=7$ or $r=9$, acting on an orbit of dimension 5, that is on the whole STM manifold.

The above results can be collected together in the following theorems:

Theorem 1: *The necessary and sufficient conditions for a five-dimensional Riemannian Gödel-type manifold to be locally homogeneous are those given by Eqs. (2.12) and (2.13).*

Theorem 2: *The five-dimensional locally homogeneous Riemannian Gödel-type manifolds admit group of isometry G_r with*

(i) $r=7$ if $m^2\neq 4\omega^2$ and $\omega\neq 0$;

(ii) $r=9$ if $m^2=4\omega^2$ with $\omega\neq 0$, or when $m^2\neq 0$ and $\omega=0$.

Theorem 3: *The five-dimensional homogeneous Riemannian Gödel-type manifolds are locally characterized by two independent parameters m^2 and ω: identical pairs (m^2,ω) specify locally equivalent manifolds.*

We remark that the particular case $m^2=\omega=0$ has not been included in our study inasmuch as, from (2.14) and (2.15), it is clearly the 5D flat manifold.

We shall now be concerned with the irreducible set of isometrically nonequivalent homogeneous Gödel-type metrics. To this end, we distinguish four classes of metrics according to:

Class I: $m^2>0,\omega\neq 0$. For this case, the general solution of (2.12) and (2.13) can be written as

$$D(x)=a_0e^{mx}+a_1e^{-mx} \quad \text{and} \quad H(x)=-\frac{2\omega}{m}(a_0e^{mx}-a_1e^{-mx})+a_2, \tag{2.17}$$

where a_0, a_1, and a_2 are arbitrary constants. According to the above theorem (Theorem 3), the constants a_0, a_1, and a_2 are not essential. In other words they can be eliminated by coordinate transformations. Indeed, if one performs the coordinate transformation $u'=u$ and successively the transformations (3.10), (3.15), and (3.16) of Ref. 37 one finds that the line element for this class of homogeneous Gödel-type manifolds is brought into the form

$$ds^2=[dt+H(r)d\phi]^2-D^2(r)d\phi^2-dr^2-dz^2-du^2 \tag{2.18}$$

in cylindrical coordinates (r,ϕ,z), where

$$H(r)=\frac{2\omega}{m^2}[1-\cosh(mr)] \quad \text{and} \quad D(r)=\frac{1}{m}\sinh(mr). \tag{2.19}$$

Class II: $m^2=0,\omega\neq 0$. For this case, the general solution of (2.12) and (2.13) is

$$D(x)=b_0x-b_1 \quad \text{and} \quad H(x)=-\omega x(b_0x-2b_1)+b_2, \tag{2.20}$$

where b_0,b_1, and b_2 are arbitrary constants. By trivial coordinate transformations the line element for this class can be brought to the form (2.18) but now with

$$H(r)=-\omega r^2 \quad \text{and} \quad D(r)=r, \tag{2.21}$$

where only the essential parameter ω appears ($m^2=0$, for this class).

Class III: $m^2 \equiv -\mu^2 < 0, \omega \neq 0$. Similarly for this class, the integration of the conditions for homogeneity (2.12) and (2.13) leads to

$$D(x) = c_0 \sin(\mu x) + c_1 \cos(\mu x) \quad \text{and} \quad H(x) = \frac{2\omega}{\mu}[c_0 \cos \mu x) - c_1 \sin(\mu x)] + c_2. \tag{2.22}$$

Here again the nonessential constants c_0, c_1, and c_2 can be eliminated by coordinate transformations so that the line element for this class reduces to (2.18) with

$$H(r) = \frac{2\omega}{\mu^2}[\cos(\mu r) - 1] \quad \text{and} \quad D(r) = \frac{1}{\mu} \sin(\mu r). \tag{2.23}$$

Class IV: $m^2 \neq 0, \omega = 0$. We refer to the manifolds of this class as degenerated Gödel-type manifolds, since the cross term in the line element, related to the rotation ω in the Gödel model, vanishes. By a trivial coordinate transformation one can make $H = 0$ with $D(r)$ given, respectively, by (2.19) or (2.23) depending on whether $m^2 > 0$ or $m^2 \equiv -\mu^2 < 0$.

In three out of the above four classes of homogeneous Gödel-type manifolds there are closed timelike curves. Indeed, the Gödel's analysis for 4D manifolds can be extended to the 5D manifolds in a straightforward way to prove this. To this end, we write the line element (2.18) as

$$ds^2 = dt^2 + 2H(r)dtd\phi - dr^2 - G(r)d\phi^2 - dz^2 - du^2, \tag{2.24}$$

where $G(r) = D^2 - H^2$ and (r, θ, ϕ) are cylindrical coordinates.

The existence of closed timelike curves of the Gödel-type depends on the behavior of $G(r)$. Indeed, if $G(r) < 0$ for a certain range of r ($r_1 < r < r_2$, say), Gödel's circles[53] $u, t, z, r = const$ are closed timelike curves.

It is, therefore, easy to show that for the above Class II there is a critical radius r_c, defined by $|\omega| r_c = 1$, such that for $r_c < r < \infty$ the circles $u, t, z = const$ and $r = const > r_c$ are closed timelike curves. So there is breakdown of causality for this class.

For the Class III noncausal Gödel circles occur depending on the root r_n of the equation

$$\sin^2\left[\frac{\mu r_n}{2}\right] = \left[1 + \frac{4\omega^2}{\mu^2}\right]^{-1} \leqslant 1. \tag{2.25}$$

Thus for this class there are alternatively causal and noncausal regions (circular rings) in the surfaces $u, t, z = const$.

For the Class I if $m^2 < 4\omega^2$ the noncausal Gödel circles occur for $r > r_1$, where

$$\sinh^2\left[\frac{m r_1}{2}\right] = \left[1 + \frac{4\omega^2}{m^2}\right]^{-1} \geqslant 1. \tag{2.26}$$

However, for $m^2 \geqslant 4\omega^2$ then $G(r) > 0$ and there is no breakdown of causally of Gödel-type. As a matter of fact, following Calvão's reasoning[53] one can show that: (a) although causal, the $m^2 = 4\omega^2$ manifolds are not stably causal,[50] that is, an arbitrary small perturbation of this metric gives rise to causality violation; (b) the family $m^2 > 4\omega^2$ is stably causal, which means that the causal properties of this family are invariant under arbitrary small perturbation of its metrics. We remark that Calvão's approach holds only if the 5D homogeneous Riemannian Gödel-type manifolds are endowed with the topology of $\mathscr{R}^5$.

Finally, for the degenerated Gödel-type geometries (Class IV), as $G(r) = D^2 > 0$ there is no closed timelike curve of Gödel type.

III. ISOMETRIES

In this section we shall be concerned with the isometries of the homogeneous 5D Riemannian Gödel-type manifolds, whose line element (2.18) can be brought into the Lorentzian form (2.8) but now with the one-forms θ^A given by

$$\theta^0 = dt + H(r)d\phi, \quad \theta^1 = dr, \quad \theta^2 = D(r)d\phi,$$
$$\theta^3 = dz, \quad \theta^4 = du, \tag{3.1}$$

where the functions $H(r)$ and $D(r)$ are given by (2.19), (2.21), or (2.23) depending on the sign of m^2.

Denoting the coordinate components of a generic Killing vector field K by $K^\mu \equiv (Q,R,S,Z,U)$, where Q,R,S,Z, and U are functions of all coordinates t,r,ϕ,z,u, then the 15 Killing equations

$$K_{(A;B)} \equiv K_{A;B} + K_{B;A} = 0 \tag{3.2}$$

can be written in the Lorentz frame (2.8)–(3.1) as

$$T_t = 0, \quad T_z - Z_t = 0, \tag{3.3}$$

$$R_r = 0, \quad Z_r + R_z = 0, \tag{3.4}$$

$$Z_z = 0, \tag{3.5}$$

$$D(T_r - R_t) - H_r P = 0, \tag{3.6}$$

$$DP_z + Z_\phi - HZ_t = 0, \tag{3.7}$$

$$T_\phi + H_r R - DP_t = 0, \tag{3.8}$$

$$R_\phi - HR_t - D_r P + DP_r = 0, \tag{3.9}$$

$$P_\phi - HP_t + D_r R = 0, \tag{3.10}$$

$$T_u - U_t = 0, \tag{3.11}$$

$$R_u + U_r = 0, \tag{3.12}$$

$$DP_u - HU_t + U_\phi = 0, \tag{3.13}$$

$$U_z + Z_u = 0, \tag{3.14}$$

$$U_u = 0, \tag{3.15}$$

where the subscripts denote partial derivatives, and where we have made

$$T \equiv HS + Q \quad \text{and} \quad P \equiv DS \tag{3.16}$$

to make easier the comparison and the use of the results obtained in Ref. 38. To this end we note that in the present form Eqs. (3.3)–(3.10) are formally identical to the Killing Eqs. (4)–(11) of Ref. 38. However, in the above Eqs. (3.3)–(3.10) the functions T,R,P,Z depend additionally on the fifth coordinate u. Taking into account this similitude, the integration of the Killing Eqs. (3.3)–(3.15) can be obtained in two steps as follows. First, by analogy with (4)–(11) of Ref. 38 one integrates (3.3)–(3.10), but at this step instead of the integration constants one has integration functions of the fifth coordinate u. Second, one uses the remaining Eqs. (3.11)–(3.15) to achieve explicit forms for these integration functions and to obtain the last component U of K.

In what follows we shall use the above two-steps procedure to integrate the Killing Eqs. (3.3)–(3.15) for the Class I of Sec. II in detail. However, for the sake of brevity, for the remaining classes we shall only present the Killing vector fields and the corresponding Lie algebras without going into the details of calculation, which can be similarly obtained, or simply verified by using

the computer algebra program KILLNF, written in CLASSI by Åman,[41] with further extensions by MacCallum.[42] As a matter of fact, even the Killing Eqs. (3.3)–(3.15) have been achieved by using KILLNF.

Class I: $m^2>0,\omega\neq 0$. In the integration of the Killing equation for this general class one is led to distinguish two different classes of solutions depending on whether $m^2\neq 4\omega^2$ or $m^2=4\omega^2$. We shall refer to these subclasses as Classes Ia and Ib, respectively.

Class Ia: $m^2>0,m^2\neq 4\omega^2$. According to the above two-steps procedure, from (3.16) and according to Eqs. (36)–(39) of Ref. 38 one finds

$$Q=\alpha+2\omega(\gamma D_r+\delta D\ \sin\ \phi+\epsilon D\ \cos\ \phi)+\frac{H}{D}\,(\gamma\ m^2 D+\delta\ D_r\ \sin\ \phi+\epsilon D_r\ \cos\ \phi), \tag{3.17}$$

$$R=\delta\ \cos\ \phi-\epsilon\ \sin\ \phi, \tag{3.18}$$

$$S=-\gamma\ m^2-\frac{D_r}{D}(\delta\ \sin\ \phi+\epsilon\ \cos\ \phi), \tag{3.19}$$

$$Z=\beta, \tag{3.20}$$

$$U=U(t,r,\phi,z,u), \tag{3.21}$$

where $\alpha,\beta,\gamma,\delta,$ and ϵ are integration functions of the fifth coordinate u.

From (3.14) and (3.15) one finds

$$\beta=\kappa_7 u+\kappa_2, \tag{3.22}$$

$$U=-\kappa_7 z+\zeta(t,r,\phi), \tag{3.23}$$

where κ_2 and κ_7 are arbitrary constants and $\zeta(t,r,\phi)$ is an arbitrary function.

Equation (3.12) can now be used to obtain the integration functions δ and ϵ and to simplify the expression of U. In doing so one finds

$$\delta=\delta_0 u+\kappa_4,\quad \epsilon=\epsilon_0 u+\kappa_5, \tag{3.24}$$

$$U=-\kappa_7 z-\delta_0 r\ \cos\ \phi+\epsilon_0 r\ \sin\ \phi+\nu(t,\phi), \tag{3.25}$$

where $\delta_0,\epsilon_0,\kappa_4,\kappa_5$ are constants and $\nu(t,\phi)$ is an arbitrary function.

Finally, Eqs. (3.11) and (3.12) give the remaining integration functions α, γ and reduce the functions δ, ϵ, and U to

$$\gamma=const\equiv m^{-1}\kappa_3,\quad \alpha=const\equiv\kappa_1, \tag{3.26}$$

$$\delta=\kappa_4,\quad \epsilon=\kappa_5,\quad U=-\kappa_7 z+\kappa_6, \tag{3.27}$$

where κ_6 is a new arbitrary constant.

Inserting (3.22), (3.26), and (3.27) into (3.17)–(3.21) and using (2.19) one obtains

$$Q=\kappa_1+\frac{2\omega}{m}\kappa_3-\frac{H}{D}(\kappa_4\ \sin\ \phi+\kappa_5\ \cos\ \phi), \tag{3.28}$$

$$R=\kappa_4\ \cos\ \phi-\kappa_5\ \sin\ \phi, \tag{3.29}$$

$$S=-\kappa_3 m-\frac{D_r}{D}(\kappa_4\ \sin\ \phi+\kappa_5\ \cos\ \phi), \tag{3.30}$$

$$Z=\kappa_7 u+\kappa_2, \tag{3.31}$$

$$U=-\kappa_7 z+\kappa_6. \tag{3.32}$$

Thus in the coordinate basis in which (2.18) is given, a set of linearly independent Killing vector fields K_N (N is an enumerating index) can be written as

$$K_1=\partial_t, \quad K_2=\partial_z, \quad K_3=\frac{2\omega}{m}\partial_t-m\partial_\phi, \tag{3.33}$$

$$K_4=-\frac{H}{D}\sin\phi\partial_t+\cos\phi\partial_r-\frac{D_r}{D}\sin\phi\partial_\phi, \tag{3.34}$$

$$K_5=-\frac{H}{D}\cos\phi\partial_t-\sin\phi\partial_r-\frac{D_r}{D}\cos\phi\partial_\phi, \tag{3.35}$$

$$K_6=\partial_u, \quad K_7=u\partial_z-z\partial_u. \tag{3.36}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_7]=-K_6, \quad [K_3,K_4]=-m\,K_5, \tag{3.37}$$

$$[K_3,K_5]=m\,K_4, \quad [K_4,K_5]=m\,K_3, \tag{3.38}$$

$$[K_6,K_7]=K_2. \tag{3.39}$$

So, the corresponding algebra is $\mathscr{L}_{Ia}=t^2\subset\!\!\!+\,\mathrm{so}(2)\oplus\tau\oplus\mathrm{so}(2,1)$. The semidirect sum of subalgebras $t^2\subset\!\!\!+\,\mathrm{so}(2)$ corresponds to the translations K_2 and K_6, and the rotation K_7. The symbol τ is associated to the time translation K_1. Finally, the infinitesimal generators of subalgebra so(2,1) are K_3, K_4, and K_5.

Class Ib: $m^2=4\omega^2, \omega\neq 0$. For this class, using the results of Ref. 38 and (3.16) one finds

$$Q=\alpha+\gamma-\frac{H}{D}(\delta\sin\phi+\epsilon\cos\phi)-\frac{H}{D}[\eta\sin(mt+\phi)+\xi\cos(mt+\phi)], \tag{3.40}$$

$$R=\delta\cos\phi-\epsilon\sin\phi+\xi\sin(mt+\phi)-\eta\cos(mt+\phi), \tag{3.41}$$

$$S=-\gamma m-\frac{D_r}{D}(\delta\sin\phi+\epsilon\cos\phi)+\frac{1}{D}[\eta\sin(mt+\phi)+\xi\cos(mt+\phi)], \tag{3.42}$$

$$Z=Z(u), \tag{3.43}$$

$$U=U(t,r,\phi,z,u), \tag{3.44}$$

where $\alpha,\gamma,\delta,\epsilon,\xi$, and η are (new) integration functions of the fifth coordinate u.

Again, from (3.14) and (3.15) one finds

$$Z=k_8u+k_2, \tag{3.45}$$

$$U=-k_8z+\chi(t,r,\phi), \tag{3.46}$$

where k_2 and k_8 are arbitrary constants and $\chi(t,r,\phi)$ is an arbitrary function.

Similarly to the Class Ia, Eq. (3.12) can be used to obtain a first expression for the integration functions ξ, η, δ, ϵ, and to reduce U according to

$$\xi=\xi_0u+k_6, \quad \eta=\eta_0u+k_7, \tag{3.47}$$

$$\delta=\delta_0u+k_4, \quad \epsilon=\epsilon_0u+k_5, \tag{3.48}$$

$$U=-k_8z+r[-\delta_0\cos\phi+\epsilon_0\sin\phi-\xi_0\sin(mt+\phi)+\eta_0\cos(mt+\phi)]+\sigma(t,\phi), \tag{3.49}$$

where $\xi_0,\eta_0,\delta_0,\epsilon_0, k_4,\ldots,k_7$ are constants, and $\sigma(t,\phi)$ is an arbitrary function.

Equations (3.11) and (3.12) give the remaining integration functions α, γ and reduce the functions ξ,η,δ,ϵ, and U to

$$\xi=k_6,\quad \eta=k_7,\quad \gamma=const\equiv k_3,\quad \alpha=const\equiv k_1, \tag{3.50}$$

$$\delta=k_4,\quad \epsilon=k_5,\quad U=-k_8z+k_9, \tag{3.51}$$

where k_9 is a new arbitrary constant.

Inserting (3.45), (3.50), and (3.51) into (3.40)–(3.44) one obtains

$$Q=k_1+k_3-\frac{H}{D}(k_4\sin\phi+k_5\cos\phi)-\frac{H}{D}[k_7\sin(mt+\phi)+k_6\cos(mt+\phi)], \tag{3.52}$$

$$R=k_4\cos\phi-k_5\sin\phi+k_6\sin(mt+\phi)-k_7\cos(mt+\phi), \tag{3.53}$$

$$S=-k_3m-\frac{D_r}{D}(k_4\sin\phi+k_5\cos\phi)+\frac{1}{D}[k_7\sin(mt+\phi)+k_6\cos(mt+\phi)], \tag{3.54}$$

$$Z=k_8u+k_2, \tag{3.55}$$

$$U=-k_8z+k_9, \tag{3.56}$$

which give rise to the Killing vector fields

$$K_1=\partial_t,\quad K_2=\partial_z,\quad K_3=\partial_t-m\partial_\phi, \tag{3.57}$$

$$K_4=-\frac{H}{D}\sin\phi\partial_t+\cos\phi\partial_r-\frac{D_r}{D}\sin\phi\partial_\phi, \tag{3.58}$$

$$K_5=-\frac{H}{D}\cos\phi\partial_t-\sin\phi\partial_r-\frac{D_r}{D}\cos\phi\partial_\phi, \tag{3.59}$$

$$K_6=-\frac{H}{D}\cos(mt+\phi)\partial_t+\sin(mt+\phi)\partial_r+\frac{1}{D}\cos(mt+\phi)\partial_\phi, \tag{3.60}$$

$$K_7=-\frac{H}{D}\sin(mt+\phi)\partial_t-\cos(mt+\phi)\partial_r+\frac{1}{D}\sin(mt+\phi)\partial_\phi, \tag{3.61}$$

$$K_8=u\partial_z-z\partial_u,\quad K_9=\partial_u, \tag{3.62}$$

whose Lie algebra is given by

$$[K_1,K_6]=-mK_7,\quad [K_1,K_7]=mK_6,\quad [K_2,K_8]=-K_9, \tag{3.63}$$

$$[K_3,K_4]=-mK_5,\quad [K_3,K_5]=mK_4,\quad [K_4,K_5]=mK_3, \tag{3.64}$$

$$[K_6,K_7]=mK_1,\quad [K_8,K_9]=-K_2. \tag{3.65}$$

So, the corresponding algebra for this case is $\mathscr{L}_{Ib}=t^2\subset\!\!\!+\,\mathrm{so}(2)\oplus\mathrm{so}(2,1)\oplus\mathrm{so}(2,1)$. The semidirect sum $t^2\subset\!\!\!+\,\mathrm{so}(2)$ comprises the translations K_2 and K_9 and the rotation K_8, while the two subalgebras so(2,1) are generated by the Killing vector fields K_1,K_6,K_7 and K_3,K_4,K_5.

Class II: $m^2=0,\omega\neq0$. Similarly for this class, using (2.21) and the results of Ref. 37 the above two-steps procedure gives rise to the following Killing vector fields:

$$K_1=\partial_t, \quad K_2=\partial_z, \quad K_3=\partial_\phi, \tag{3.66}$$

$$K_4=-\omega r \sin\phi\partial_t-\cos\phi\partial_r+\frac{1}{r}\sin\phi\partial_\phi, \tag{3.67}$$

$$K_5=-\omega r \cos\phi\partial_t+\sin\phi\partial_r+\frac{1}{r}\cos\phi\partial_\phi, \tag{3.68}$$

$$K_6=u\partial_z-z\partial_u, \quad K_7=\partial_u. \tag{3.69}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_6]=-K_7, \quad [K_3,K_4]=K_5, \tag{3.70}$$

$$[K_3,K_5]=-K_4, \quad [K_4,K_5]=2\omega K_1, \tag{3.71}$$

$$[K_6,K_7]=-K_2. \tag{3.72}$$

Therefore the corresponding algebra for this case is $\mathscr{L}_{II}=t^2\subset\!\!\!+\,\mathrm{so}(2)\oplus\mathscr{L}_4$. The semidirect sum of subalgebras corresponds here to the translations K_2 and K_7, and the rotation K_6. The subalgebra $\mathscr{L}_4$ is generated by K_1,K_3,K_4, and K_5. This algebra $\mathscr{L}_4$ is soluble and does not contain Abelian 3D subalgebras; it is classified as type *III* with $q=0$ by Petrov.[54]

Class III: $m^2\equiv-\mu^2<0,\omega\neq0$. Again for this class using (2.23) and Ref. 38, the above outlined two-step procedure furnishes

$$K_1=\partial_t, \quad K_2=\partial_z, \quad K_3=\frac{2\omega}{\mu}\partial_t+\mu\partial_\phi, \tag{3.73}$$

$$K_4=-\frac{H}{D}\sin\phi\partial_t+\cos\phi\partial_r-\frac{D_r}{D}\sin\phi\partial_\phi, \tag{3.74}$$

$$K_5=-\frac{H}{D}\cos\phi\partial_t-\sin\phi\partial_r-\frac{D_r}{D}\cos\phi\partial_\phi, \tag{3.75}$$

$$K_6=u\partial_z-z\partial_u, \quad K_7=\partial_u. \tag{3.76}$$

The Lie algebra has the following nonvanishing commutators:

$$[K_2,K_6]=-K_7, \quad [K_3,K_4]=\mu K_5, \tag{3.77}$$

$$[K_3,K_5]=-\mu K_4, \quad [K_4,K_5]=\mu K_3, \tag{3.78}$$

$$[K_6,K_7]=-K_2. \tag{3.79}$$

Thus the corresponding algebra for this case is $\mathscr{L}_{III}=t^2\subset\!\!\!+\,\mathrm{so}(2)\oplus\tau\oplus\mathrm{so}(3)$. The semidirect sum of subalgebras corresponds again to the translations K_2 and K_7, and the rotation K_6. Here τ is associated to the Killing vector field K_1, whereas to the subalgebra so(3) correspond K_3, K_4, and K_5.

Class IV: $m^2\neq0,\omega=0$. This class corresponds to the so-called degenerated Gödel-type manifolds. By a similar procedure one obtains for this class the following Killing vector fields:

$$K_1=\partial_t, \quad K_2=\partial_z, \quad K_3=z\,\partial_t+t\partial_z, \tag{3.80}$$

$$K_4=\cos\phi\partial_r-\frac{D_r}{D}\sin\phi\partial_\phi, \tag{3.81}$$

$$K_5 = -\sin\,\phi\partial_r - \frac{D_r}{D}\cos\,\phi\partial_\phi, \quad K_6 = \partial_\phi, \tag{3.82}$$

$$K_7 = u\,\partial_z - z\,\partial_u, \quad K_8 = u\,\partial_t + t\partial_u, \quad K_9 = \partial_u, \tag{3.83}$$

where $D(r)=(1/m)\sinh mr$ for $m^2>0$, or $D(r)=(1/\mu)\sin \mu r$ for $m^2\equiv -\mu^2<0$. The Lie algebra has the following nonvanishing commutators:

$$[K_1,K_3]=K_2, \quad [K_1,K_8]=K_9, \quad [K_2,K_3]=K_1, \tag{3.84}$$

$$[K_2,K_7]=-K_9, \quad [K_3,K_7]=-K_8, \quad [K_3,K_8]=-K_7, \tag{3.85}$$

$$[K_4,K_5]=-m^2K_6, \quad [K_4,K_6]=-K_5, \quad [K_5,K_6]=K_4, \tag{3.86}$$

$$[K_7,K_8]=-K_3, \quad [K_7,K_9]=-K_2, \quad [K_8,K_9]=-K_1, \tag{3.87}$$

where one should substitute $-m^2$ by μ^2 if $m^2<0$. So, when $m^2>0$ the corresponding algebra is $\mathscr{L}_{IV}=t^3 \subsetplus \mathrm{so}(2,1)\oplus\mathrm{so}(2,1)$, where to t^3 correspond K_1, K_2, and K_9. The infinitesimal generators of so(2,1) in the semidirect sum are K_3, K_7, and K_8, while to the other so(2,1) correspond K_4, K_5, and K_6. When $m^2<0$ the algebra associated to K_4, K_5, and K_6 is so(3), instead.

It is worth noting that none of the above Lie algebras is semisimple, but some of their subalgebras are. Besides, most of the simple subalgebras are noncompact. The 3D subalgebras so(3) present in the cases $\mathscr{L}_{III}$ and $\mathscr{L}_{IV}$ with $m^2<0$ are compact, though.

We have therefore succeeded in finding the maximal group of motions of all classes of 5D homogeneous Riemannian Gödel-type manifolds we have studied in the previous section.

Equations (3.33)–(3.36), (3.57)–(3.62), (3.66)–(3.69), (3.73)–(3.76), and (3.80)–(3.83) make explicit that the 5D locally homogeneous Riemannian Gödel-type manifolds admit maximal group of isometry G_r with: (a) $r=7$ if $m^2\neq 4\omega^2$ and $\omega\neq 0$; or (b) $r=9$ if $m^2=4\omega^2$ with $\omega\neq 0$, or when $m^2\neq 0$ and $\omega=0$, in agreement with Theorem 2 of the previous section. Actually the integration of the Killing equations constitutes a different way of deriving that theorem. Furthermore, these equations also show that the isotropy subgroup H of G_r is such that $\dim(H)=2$ for $m^2\neq 4\omega^2$ and $\omega\neq 0$, while for $m^2=4\omega^2$ with $\omega\neq 0$ as well as for the degenerated Gödel-type ($m^2\neq 0$ and $\omega=0$) we have $\dim(H)=4$, also in agreement with the previous section.

To conclude, we should like to emphasize again that the results of this work hold for any five-dimensional locally homogeneous Riemannian Gödel-type manifold, regardless of the underlying 5D Kaluza–Klein-type theory of gravitation one may be concerned with. This gives a measure of the generality of our results in the context of these theories, in which the 5D Riemannian manifolds are the underlying arena for the formulation of the physical laws.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial assistance from CNPq.

[1] G. Nordström, Phys. Z. **15**, 504 (1914).
[2] Th. Kaluza, Sitzungsber. K. preuss. Akad. Wiss., Phys. Math. Kl. **LIV**, 966 (1921).
[3] O. Z. Klein, Phys. Z. **37**, 895 (1926).
[4] A. Chodos and S. Detweiler, Phys. Rev. D **21**, 2167 (1980).
[5] E. Alvarez, Phys. Rev. Lett. **51**, 931 (1983).
[6] A. H. Guth, Phys. Rev. D **23**, 347 (1981).
[7] L. Burakovsky and L. P. Horwitz, Gen. Relativ. Gravit. **27**, 1043 (1995).
[8] J. Ibáñez and E. Verdaguer, Phys. Rev. D **34**, 1202 (1986).
[9] R. J. Gleiser and M. C. Diaz, Phys. Rev. D **37**, 3761 (1986). See also references therein.
[10] Y. S. Vladimirov and S. S. Kokarev, Gen. Relativ. Gravit. **29**, 141 (1997).
[11] P. S. Wesson, Astron. Astrophys. **119**, 145 (1983).
[12] P. S. Wesson, Gen. Relativ. Gravit. **16**, 193 (1984).
[13] P. S. Wesson, Astron. Astrophys. **143**, 233 (1985).
[14] Ø. Grøn and H. H. Soleng, Gen. Relativ. Gravit. **20**, 1115 (1988).
[15] P. S. Wesson, Gen. Relativ. Gravit. **22**, 707 (1990).
[16] G.-W. Ma, Phys. Lett. A **146**, 375 (1990).
[17] Ø. Grøn, Astron. Astrophys. **193**, 1 (1988).
[18] P. S. Wesson, Astron. Astrophys. **166**, 1 (1986).

[19] A. Coley, Astron. Astrophys. **233**, 305 (1990).
[20] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[21] M. M. Som and A. K. Raychaudhuri, Proc. R. Soc. London, Ser. A **304**, 81 (1968).
[22] A. Banerjee and S. Banerji, J. Phys. A **1**, 188 (1968).
[23] F. Bampi and C. Zordan, Gen. Relativ. Gravit. **9**, 393 (1978).
[24] M. Novello and M. J. Rebouças, Phys. Rev. D **19**, 2850 (1979).
[25] M. J. Rebouças,Phys. Lett. **70**, 161 (1979).
[26] J. Pfarr, Gen. Relativ. Gravit. **13**, 1073 (1981). See also references therein quoted on the philosophical discussions which came about after Gödel's paper.
[27] S. K. Chakraborty, Gen. Relativ. Gravit. **12**, 925 (1980).
[28] M. J. Rebouças and A. F. F. Teixeira, Phys. Rev. D **34**, 2985 (1986).
[29] M. J. Rebouças, J. E. Åman, and F. F. Teixeira, J. Math. Phys. **27**, 1370 (1986).
[30] F. M. Paiva, M. J. Rebouças, and A. F. F. Teixeira, Phys. Lett. A **126**, 168 (1987).
[31] K. Dunn, Gen. Relativ. Gravit. **21**, 137 (1989).
[32] R. X. Saibatalov, Gen. Relativ. Gravit. **7**, 697 (1995).
[33] T. Singh and A. K. Agrawal, Fortschr. Phys. **42**, 71 (1994). It contains a fair list of references on Gödel-type space-times in general relativity and in the torsion theories of gravitation.
[34] A. Krasiński, *Rotating Dust Solutions of Einstein's Equations with 3-Dimensional Symmetry Groups. Part 1: All Killing Fields Linearly Independent of u^α and ω^α*, gr-qc/9607021, 1997. This reference contains an overview of the literature on rotating models in general relativity.
[35] For a clear distinction between local and global (topological) homogeneity of a manifold see, for example, T. Koike, M. Tanimoto, and A. Hosoya, J. Math. Phys. **35**, 4855 (1994).
[36] A. K. Raychaudhuri and S. N. G. Thakurta, Phys. Rev. D **22**, 802 (1980).
[37] M. J. Rebouças and J. Tiomno, Phys. Rev. D **28**, 1251 (1983).
[38] A. F. F. Teixeira, M. J. Rebouças, and J. E. Åman, Phys. Rev. D **32**, 3309 (1985). Note that in this reference there are two misprints, namely: in Eq. (5) the term $-R_z$ should read $+R_z$, and in Eq. (38) the term $\kappa_3 mD$ should read $\kappa_3 m^2 D$.
[39] M. J. Rebouças and J. E. Åman, J. Math. Phys. **28**, 888 (1987).
[40] A. Karlhede, Gen. Relativ. Gravit. **12**, 693 (1980).
[41] J. E. Åman, ''Manual for CLASSI: Classification Programs for Geometries in General Relativity,'' Institute of Theoretical Physics Technical Report, 1987. Third provisional edition. Distributed with the SHEEP sources.
[42] M. A. H. MacCallum and J. E. F. Skea, ''SHEEP: A Computer Algebra System for General Relativity,'' in *Algebraic Computing in General Relativity, Lecture Notes from the First Brazilian School on Computer Algebra*, Vol. II, edited by M. J. Rebouças and W. L. Roque (Oxford University Press, Oxford, 1994). See also references therein.
[43] E. Cartan, ''Leçons sur la Géométrie des Éspaces de Riemann,'' (Gauthier-Villars, Paris, 1951). English translation by J. Glazebrook (Math Sci, Brookline, 1983).
[44] P. S. Wesson, Astrophys. J. **394**, 19 (1992).
[45] P. S. Wesson, Phys. Lett. B **276**, 299 (1992).
[46] P. S. Wesson and J. Ponce de Leon, J. Math. Phys. **33**, 3883 (1992).
[47] A. Billyard and P. S. Wesson, Gen. Relativ. Gravit. **28**, 129 (1997). See also the references therein cited on a new 5D Kaluza–Klein-type theory of gravitation.
[48] M. A. H. MacCallum, ''Classifying metrics in theory and practice,'' in *Unified Field Theory in More Than 4 Dimensions, Including Exact Solutions*, edited by V. de Sabbata and E. Schmutzer (World Scientific, Singapore, 1983).
[49] M. A. H. MacCallum, ''Computer-aided classification of exact solutions in general relativity,'' in *General Relativity and Gravitational Physics (9th Italian Conference)*, edited by R. Cianci, R. de Ritis, M. Francaviglia, G. Marmo, C. Rubano, and P. Scudellaro (World Scientific, Singapore, 1991).
[50] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge, England, 1973).
[51] T. Y. Thomas, *Differential Invariants of Generalized Spaces* (Cambridge University Press, Cambridge, England, 1934).
[52] In the next section we integrate the Killing equations for the 5D Riemannian Gödel-type manifolds and explicitly show that the isotropy group depends on these special relations between the essential parameters m^2 and ω.
[53] M. O. Calvão, M. J. Rebouças, A. F. F. Teixeira, and W. M. Silva, Jr., J. Math. Phys. **29**, 1127 (1988).
[54] A. Z. Petrov, *Einstein Spaces*, 1st English ed. (Pergamon, New York, 1969), p. 63.

Gauge conditions in modern Kaluza–Klein theory

William N. Sajko[a] and Paul S. Wesson[b]
Department of Physics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

Hongya Liu
Department of Physics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada, and Department of Physics, Zhengzhou University, Zhengzhou, Henan, People's Republic of China

(Received 4 August 1997; accepted for publication 29 December 1997)

In the modern version of 5-D general relativity where the fifth dimension induces matter in 4-D space–time, a choice of coordinates and its restrictions on the geometry (a gauge) dictates the physics of the gravitational, electromagnetic, and scalar fields (spin 2, 1, 0). We extend previous work by concentrating on two related problems. First, we use the lapse and shift formalism to carry out a $(4+1)$ split of Kaluza–Klein theory valid in any gauge. Second, we use a conformal gauge to simplify the 5-D field equations, find a plane-wave solution, and identify the energy and momentum of the particle associated with it. © *1998 American Institute of Physics.* [S0022-2488(98)01504-7]

I. INTRODUCTION

It has been known since 1921 that in vacuum the Kaluza–Klein equations contain the Einstein equations of general relativity and the Maxwell equations of classical electromagnetism without sources.[1,2] It has been known since 1992 that the 15 Kaluza–Klein equations for apparently empty space can be rewritten as 10 Einstein equations *with* sources, 4 conservation equations, and 1 scalar-wave equation (Ref. 3; for a review of Kaluza–Klein, see Ref. 4). The latter result was proved algebraically using a special choice of 4 of the 5 available coordinate degrees of freedom to set 4 of the 15 metric coefficients to zero with consequent simplification of the associated tensors (in short, a gauge condition). That particular gauge has proven useful in showing, among other things, that exact cosmological solutions of the 5-D field equations have induced properties of matter which agree with observation.[5] It also applies to exact one-body solutions which are in agreement with the classical tests of relativity in the solar system.[6] However, while a practical gauge for elucidating the gravitational and scalar (matter) fields, it curtails investigation of the electromagnetic field. Exact solutions for the latter are known,[7] but hitherto there has not been any decomposition done of the general Kaluza–Klein field equations into parts relevant to the three classical fields of physics. In what follows we aim to rectify this situation. We will extend previous work[3,8] and give a $(4+1)$ split of the 5-D field equations that defers gauge conditions as far as possible, leaving the choice of gauge to be applied later and as required by physics. We will also, in what follows, treat a related problem, namely how to apply the induced-matter prescription to particles. We will extend previous work[9,10] and use a conformal gauge to put the field equations into a form where they admit a wave solution whose parameters relate to the energy and momentum of the associated particle.

Before proceeding, we wish to make some comments on the usefulness (or otherwise) of certain gauges in 5-D Kaluza–Klein theory. This theory, which exists in several versions, has as its basic geometrical object a 5×5 symmetric tensor g_{ab} with 15 independent components. There are, in general, five conditions that can be applied to the metric tensor and/or geometrical objects derived from it, corresponding to the freedom to choose five coordinates x^a. The theory is commonly regarded as a unified description of the gravitational, electromagnetic, and scalar fields. That is, it is a classical theory of a spin-2 graviton, a spin-1 photon and a spin-0 boson. However, whether or not these three fields can be disentangled from each other and identified with their 4-D

[a]Electronic mail: wnsajko@astro.uwaterloo.ca
[b]Electronic mail: wesson@astro.uwaterloo.ca

 © 1998 American Institute of Physics

counterparts (Einstein, Maxwell, and Higgs fields) depends crucially on a choice of 5-D gauge. Some of the latter are useful and some (the vast majority) are not. For example, one could use up the five available gauge choices by setting $g_{4\alpha}=0$, $g_{44}=1$ ($a=0-4$, $\alpha=0-3$). This is mathematically appealing, but physically confusing because the gravitational, electromagnetic, and scalar interactions are combined in the 4×4 sector of the theory. A considerably more useful gauge is the ''electromagnetic'' one of the original Kaluza–Klein theory.[1,2] This consists of $g_{\alpha\beta}=g_{\alpha\beta}(x^{\gamma})$, $g_{4\alpha}=A_{\alpha}(x^{\beta})$, and $g_{44}=1$, and enables us to identify the potentials of classical electromagnetism. However, the restriction $\partial g_{ab}/\partial x^4=0$ (the ''cylinder condition'') makes it problematic to describe anything other than radiationlike matter. Hence the ''matter'' gauge of modern Kaluza–Klein theory.[3] This consists of $g_{\alpha\beta}=g_{\alpha\beta}(x^a)$, $g_{4\alpha}=0$, and $g_{44}=g_{44}(x^a)$, and enables us to describe the pressure, density, and equation of state of macroscopic matter. But even this latter gauge is lacking in that it does not enable us to see how electromagnetism interacts with gravity (including its source matter) and the scalar field. It also does not help much in enabling us to identify the four-momenta and mass of microscopic matter.

A natural question, and one we wish to answer before proceeding, concerns the status of the 5-D analog of the 4-D harmonic gauge. The latter is extensively used in general relativity, notably to describe gravitational waves in the weak-field limit of Einstein's theory.[11] The corresponding 5-D harmonic gauge, in terms of the Christoffel symbol and the determinant of the metric, is given by

$$g^{ab}\Gamma^{c}_{ab}=-\frac{1}{g^{1/2}}\frac{\partial}{\partial x^{d}}\,(g^{1/2}g^{cd})=0. \tag{1.1}$$

These are five conditions, but since Γ^{c}_{ab} is not a tensor the conditions (1.1) are not 5-D covariant. This is the classical analog of the situation in quantum field theory involving noncovariant gauges.[12] In physical terms, fields which are split on one hypersurface of the metric will in general be mixed on another hypersurface, implying a problem when comparing 5-D and 4-D physics. By analog with the latter, we assume that the metric is close to flat:

$$g_{ab}=\eta_{ab}+h_{ab}, \quad \eta_{44}=\pm1, \quad |h_{ab}|\ll1. \tag{1.2}$$

Here we should not presume the sign of η_{44}, since exact solutions of the 5-D field equations with good 4-D physical properties exist for both signs.[4] The gauge conditions (1.1) with (1.2) now read

$$\frac{\partial}{\partial x^{a}}\,h^{a}_{b}=\frac{1}{2}\frac{\partial}{\partial x^{b}}\,h^{a}_{a}. \tag{1.3}$$

We note in passing that these conditions are very restrictive, being first-order relations that must be applied to second-order field equations. The latter in the approximation (1.2) become

$$\Box h_{ab}=0, \quad \Box\equiv\eta^{ab}\frac{\partial^2}{\partial x^{a}\partial x^{b}}. \tag{1.4}$$

This 5-D box operator can be written in terms of its 4-D counterpart ${}^{(4)}\Box\equiv\eta^{\alpha\beta}\partial^2/\partial x^{\alpha}\partial x^{\beta}$, so (1.4) then read

$${}^{(4)}\Box h_{ab}=-2\,\eta^{\alpha4}\frac{\partial^2}{\partial x^{\alpha}\partial x^{4}}\,h_{ab}-\eta^{44}\frac{\partial^2}{\partial x^{4}\partial x^{4}}\,h_{ab}. \tag{1.5}$$

That is, free waves in five dimensions become waves with sources in four dimensions. Denoting the right-hand side of (1.5) by S_{ab}, solutions can be written at least in principle in terms of retarded Green's functions:[13]

$$h_{ab}(r,t;x^4)=-\frac{1}{4\pi}\int d^3\bar{r}\,\frac{S_{ab}(\bar{r},t-|r-\bar{r}|;x^4)}{|r-\bar{r}|}, \tag{1.6}$$

where

$$\frac{1}{|r-\bar{r}|}=\frac{1}{r}+\frac{x^{\alpha}\bar{x^{\alpha}}}{r^3}+\frac{1}{2}\frac{x^{\alpha}x^{\beta}}{r^5}(3\bar{x^{\alpha}}\bar{x^{\beta}}-\bar{r^2}\delta^{\alpha\beta})+\cdots .$$

In practice, the signature may present problems that would require an imaginary coordinate transformation (Wick rotation) depending on the sign of the last part of the metric. Our conclusion is that the 5-D harmonic gauge confirms the general property of modern Kaluza–Klein theory, namely that source-free equations in five dimensions become ones with sources in four dimensions, but that there are several problems with it.

We therefore proceed in the next two sections to elucidate what appear to us to be two major issues in Kaluza–Klein theory. In Sec. II we will extend previous work,[3] and give a gauge-deferred decomposition of 5-D to 4-D relations, using the lapse and shift formalisms.[14–16] In Sec. III we will look again at the issue of how field and particle properties relate,[9] and give an identification of energy and momentum using a conformal form of the metric.[10] Section IV is a conclusion.

II. A (4+1) SPLIT USING LAPSE AND SHIFT

The purely geometric field equations of Kaluza–Klein theory in terms of the Ricci tensor are

$$R_{ab}=0 \quad (a,b=0-3,4). \tag{2.1}$$

The field equations of general relativity with sources are the Einstein equations

$$R_{\alpha\beta}-\frac{R}{2}g_{\alpha\beta}-\Lambda g_{\alpha\beta}=T_{\alpha\beta} \quad (\alpha,\beta=0-3). \tag{2.2}$$

Here $R_{\alpha\beta}$ is the 4-D Ricci tensor, R is the 4-D Ricci scalar, Λ is the cosmological constant, and $g_{\alpha\beta}$ is the 4-D part of the 5-D metric tensor whose interval is defined by $dS^2=g_{ab}\,dx^a\,dx^b$. The coordinates x^a we leave unspecified for now, and we absorb physical constants by an appropriate choice of units. As mentioned above, it is known that (2.2) are a ten-component subset of the 15 equations (2.1). However, the algebraic method employed to prove this used the gauge condition $g_{\alpha 4}=0$,[3] and if we wish to investigate (2.1) without the initial imposition of such a gauge it becomes apparent that a simple decomposition of (2.1) is algebraically forbidding. Therefore we will apply the lapse and shift formalism, which is well known in the (3+1) split of Einstein's theory. For Kaluza–Klein theory, we proceed to write out the 10, 4, 1 components of (2.1) in a general way.

The 5-D space–time is chosen to be foliated by 4-D hypersurfaces Σ_{ℓ} which are defined to be surfaces $x^4\equiv\ell=$constants. The normal vectors to these surfaces Σ_{ℓ} satisfy

$$n^a n_a=\varepsilon=\pm 1, \tag{2.3}$$

so that n^a may be either spacelike ($\varepsilon=-1$) or timelike ($\varepsilon=+1$). In addition to a normal vector on Σ_{ℓ} we define a 4-D intrinsic metric tensor

$$ds^2=g_{\alpha\beta}(x^{\Sigma},\ell)\,dx^{\alpha}\,dx^{\beta}, \tag{2.4}$$

which is related to the 5-D metric tensor by using a (4+1) lapse and shift decomposition imported from the (3+1) split of 4-D space–time. Thus the 5-D metric tensor may be written as

$$dS^2=g_{\alpha\beta}(dx^{\alpha}+N^{\alpha}\,d\ell)(dx^{\beta}+N^{\beta}\,d\ell)+\varepsilon\Phi^2\,d\ell^2, \tag{2.5}$$

where $\Phi=\Phi(x^{\Sigma},\ell)$ is the lapse function, and $N^{\alpha}=N^{\alpha}(x^{\Sigma},\ell)$ is the 4-D shift vector (N^{α} transforms as a four-vector on Σ_{ℓ}). In matrix form the 5-D metric and its inverse are given by

$$g_{ab}=\begin{pmatrix} g_{\alpha\beta} & N_{\alpha} \\ N_{\beta} & \varepsilon\Phi^2+N_{\gamma}N^{\gamma}\end{pmatrix}, \quad g^{ab}=\begin{pmatrix} g^{\alpha\beta}+\dfrac{\varepsilon N^{\alpha}N^{\beta}}{\Phi^2} & \dfrac{-\varepsilon N^{\alpha}}{\Phi^2} \\ \dfrac{-\varepsilon N^{\beta}}{\Phi^2} & \dfrac{\varepsilon}{\Phi^2}\end{pmatrix}, \tag{2.6}$$

where 4-D vectors can be raised and lowered by the intrinsic metric on Σ_ℓ ($N_\gamma = g_{\gamma\delta}N^\delta$). The normal vector n_a is given in component form by

$$n_a = (0,0,0,0,\varepsilon\Phi), \tag{2.7}$$

from which we obtain n^a by raising with g^{ab}:

$$n^a = g^{ab}n_a = \left(\frac{-N^\alpha}{\Phi}, \frac{1}{\Phi}\right). \tag{2.8}$$

We will also need a projection tensor, since our goal is to obtain field equations in the 4-D space–time of Σ_ℓ from the 5-D space. This is

$$h_{ab} = g_{ab} - \varepsilon n_a n_b, \tag{2.9}$$

which will project out 4-D tensor components of 5-D tensors. The projection tensor and its inverse are given in matrix form as

$$h_{ab} = \begin{pmatrix} g_{\alpha\beta} & N_\alpha \\ N_\beta & N_\gamma N^\gamma \end{pmatrix}, \quad h^{ab} = \begin{pmatrix} g^{\alpha\beta} & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.10}$$

and in mixed form (useful for calculations) as

$$h_a{}^b = \begin{pmatrix} \delta_\alpha{}^\beta & 0 \\ N^\beta & 0 \end{pmatrix}, \quad h^b{}_a = \begin{pmatrix} \delta^\beta{}_\alpha & N^\beta \\ 0 & 0 \end{pmatrix}. \tag{2.11}$$

From (2.9) and (2.10) the projection tensor has the properties

$$h_{ab}n^a = 0, \quad h_{ab}h^b{}_c = h_{ac}, \quad h^a{}_a = 4, \tag{2.12}$$

and may be thought of as the metric on Σ_ℓ.

Since Σ_ℓ is embedded in a 5-D space, we need to describe the rate of change of Σ_ℓ as it moves in the normal direction. The geometrical object which provides this is the second fundamental form of the surface Σ_ℓ, usually called the extrinsic curvature tensor:

$$K_{ab} = -h_b{}^c\nabla_c n_a. \tag{2.13}$$

This is symmetric and orthogonal to n^a:

$$K_{ab}n^a = 0. \tag{2.14}$$

Using the matrix form of h_b^c and the usual definition of the 5-D covariant derivative operator ∇_c, it is easily shown that the matrix form of K_{ab} is

$$K_{ab} = \begin{pmatrix} K_{\alpha\beta} & N^\gamma K_{\alpha\gamma} \\ N^\gamma K_{\beta\gamma} & 0 \end{pmatrix}, \tag{2.15}$$

$$K_{\alpha\beta} = \frac{1}{2\Phi}(D_\alpha N_\beta + D_\beta N_\alpha - \partial_\ell g_{\alpha\beta}), \tag{2.16}$$

where D_α is the derivative operator defined on Σ_ℓ and $\partial_\ell \equiv \partial/\partial\ell$. The action of D_α on scalar and vector fields is defined by

$$D_\alpha\Phi = \partial_\alpha\Phi, \quad D_\alpha N_\beta = \partial_\alpha N_\beta - \Gamma^\gamma_{\alpha\beta}N_\gamma,$$

where $\Gamma^\gamma_{\alpha\beta}$ is the 4-D Christoffel symbol associated with the metric $h_{\alpha\beta} = g_{\alpha\beta}$.

We can now decompose the 15 5-D field equations $R_{ab}=0$ into $10+4+1$ 4-D equations. These are ten field equations for the 4-D Ricci tensor $R_{\alpha\beta}$, four Gauss–Codazzi equations for the curvature which have the form of conservation laws, and one relation for the scalar field which resembles a wave equation. Explicitly, they are

$$R_{\alpha\beta}=\frac{1}{\Phi}D_\alpha D_\beta\Phi+\frac{\varepsilon}{\Phi}(\mathcal{L}_N K_{\alpha\beta}-\partial_\perp K_{\alpha\beta})+\varepsilon(KK_{\alpha\beta}-2K_{\alpha\gamma}K^\gamma{}_\beta), \tag{2.17}$$

$$0=D_\alpha(K^\alpha{}_\beta-\delta^\alpha{}_\beta K), \tag{2.18}$$

$$\varepsilon\Box\Phi=\partial_\perp K-N^\alpha D_\alpha K-\Phi K^{\alpha\beta}K_{\alpha\beta}. \tag{2.19}$$

Here $\mathcal{L}_N K_{\alpha\beta}$ is the Lie derivative of $K_{\alpha\beta}$ wrt N^α, $K\equiv g^{\alpha\beta}K_{\alpha\beta}$, and $\Box\equiv g^{\alpha\beta}D_\alpha D_\beta$. From (2.17), we can obtain the 4-D Ricci scalar and the 4-D Einstein tensor. The latter defines the 4-D (induced) energy-momentum tensor, whose trace defines the equation of state of the matter. Explicitly, we have

$$R=\varepsilon(K^2-K^{\alpha\beta}K_{\alpha\beta}), \tag{2.20}$$

$$T_{\alpha\beta}\equiv G_{\alpha\beta}=\frac{1}{\Phi}D_\alpha D_\beta\Phi+\frac{\varepsilon}{\Phi}(\mathcal{L}_N K_{\alpha\beta}-\partial_\perp K_{\alpha\beta})$$
$$+\varepsilon(KK_{\alpha\beta}-2K_{\alpha\gamma}K^\gamma{}_\beta)-\frac{\varepsilon}{2}g_{\alpha\beta}(K^2-K^{\alpha\beta}K_{\alpha\beta}), \tag{2.21}$$

$$T=-R=-\varepsilon(K^2-K^{\alpha\beta}K_{\alpha\beta}). \tag{2.22}$$

These are 4-D physical objects derived from 5-D geometrical ones. To amplify the role of the Gauss–Codazzi equations (2.18), we can write them as explicit conservation relations:

$$D_\alpha\left(\frac{1}{\sqrt{-g}}\,\pi^{\alpha\beta}\right)=0,\quad \pi^{\alpha\beta}\equiv\sqrt{-g}(K^{\alpha\beta}-g^{\alpha\beta}K). \tag{2.23}$$

Here $\pi_{\alpha\beta}$ is the momentum canonically conjugate to the metric $g_{\alpha\beta}$. Apart from the dimension and the determinant of the hypersurface, (2.23) is the same as in the $3+1$ split of space–time.[16] We hope to report on a Hamiltonian treatment of $4+1$ space–time in the near future.

The expression for the energy-momentum tensor, the conservation equations, and the scalar equation are in general quite complicated, unless $K_{\alpha\beta}$ possesses some symmetries which condense the algebra.[17] Significant simplification occurs when the N^α are Killing vectors on $\Sigma_\perp$ or

$$D_\alpha N_\beta+D_\beta N_\alpha=0. \tag{2.24}$$

Then the extrinsic curvature (2.16) reduces to

$$K_{\alpha\beta}=-\frac{1}{2\Phi}\partial_\perp g_{\alpha\beta}. \tag{2.25}$$

The gauge choice $N^\alpha=0$ studied by Wesson and Ponce de Leon[3] is a trivial solution of Killing's equation on $\Sigma_\perp$, and hence Eqs. (2.17)–(2.19) with the extrinsic curvature (2.25) exactly reproduce their results. In particular $T_{\alpha\beta}$, the conservation equations, and the scalar equation are

$$T_{\alpha\beta}=\frac{1}{\Phi}D_\alpha D_\beta\Phi-\frac{\varepsilon}{2\phi^2}\left\{\frac{\dot\Phi}{\Phi}\dot g_{\alpha\beta}-\ddot g_{\alpha\beta}+g^{\lambda\mu}\dot g_{\alpha\lambda}\dot g_{\beta\mu}-\frac{1}{2}g^{\mu\nu}\dot g_{\mu\nu}\dot g_{\alpha\beta}\right.$$
$$\left.+\frac{1}{4}g_{\alpha\beta}[\dot g^{\mu\nu}\dot g_{\mu\nu}+(g^{\mu\nu}\dot g_{\mu\nu})^2]\right\}, \tag{2.26}$$

$$0 = D_\alpha \left[\frac{1}{\Phi} \left(g^{\alpha\gamma} \dot{g}_{\gamma\beta} - \delta^\alpha{}_\beta g^{\mu\nu} \dot{g}_{\mu\nu} \right) \right], \tag{2.27}$$

$$\varepsilon \Phi \Box \Phi = \frac{1}{2\Phi} \dot{\Phi} g^{\alpha\beta} \dot{g}_{\alpha\beta} - \frac{1}{2} \left(g^{\alpha\beta} \ddot{g}_{\alpha\beta} + \frac{1}{2} \dot{g}^{\alpha\beta} \dot{g}_{\alpha\beta} \right). \tag{2.28}$$

For the case $\dot{g}_{\alpha\beta} \equiv \partial g_{\alpha\beta} / \partial \ell = 0$, we see that the conservation relations (2.27) drop out, the last relation becomes a wave equation for a massless scalar field, and $T \equiv g^{\alpha\beta} T_{\alpha\beta}$ becomes zero (which implies a radiationlike equation of state if the induced matter is modeled as a perfect fluid). As mentioned previously, these results would hold equally well if the N_α satisfy Killing's equations and $\partial_\ell g_{\alpha\beta} = 0$. If $N_\alpha \neq 0$ and do not satisfy Killing's equation, and $\partial_\ell g_{\alpha\beta} \neq 0$, then the $T_{\alpha\beta}$ will in general be complicated and the solutions to $R_{ab} = 0$ will not describe 4-D matter with a radiationlike equation of state.

III. THE CONFORMAL GAUGE

The ansatz for the metric presented here is a generalization of conformally scaled metrics that have sometimes been used in Kaluza–Klein theory.[4] The 5-D metric is assumed to be independent of the fifth coordinate $x^4 = \ell$ ($\partial_\ell g_{ab} = 0$) and the off-diagonal metric elements are set to zero ($g_{a4} = 0$). The 5-D metric then has the form

$$ds^2 = g_{ab}\, dx^a\, dx^b = g_{\alpha\beta}(x^\gamma)\, dx^\alpha\, dx^\beta + \varepsilon \Phi^2(x^\gamma)\, d\ell^2. \tag{3.1}$$

The field equations induced from the 5-D vacuum equation $R_{ab} = 0$ are

$$R_{\alpha\beta} = \frac{1}{\Phi} D_\beta \Phi_\alpha, \tag{3.2}$$

$$\Box \Phi = 0. \tag{3.3}$$

The above field equations are independent of the sign of the extra dimension ε, which implies that so should be their solutions. This may explain the existence of the ''vacuum wave'' solutions of 5-D cosmology.[18]

In 4-D general relativity a conformal rescaling of the metric transforms each metric component equally. We wish to retain this character in 5-D theory as a hypersurface (ℓ = constant) limit. This suggests that the 5-D metric should be conformally scaled by two factors such that

$$\begin{aligned} g_{\alpha\beta}(x^\gamma) &\to e^{2a\Phi} g_{\alpha\beta}(x^\gamma), \\ g_{44} = \varepsilon \Phi^2(x^\gamma) &\to g_{44} = \varepsilon e^{2b\Phi}, \end{aligned} \tag{3.4}$$

where a and b are constants and $\Phi = \Phi(x^\gamma)$. The 5-D metric in the conformal gauge then has the form

$$dS^2 = e^{2a\Phi} g_{\alpha\beta}(x^\gamma)\, dx^\alpha\, dx^\beta + \varepsilon e^{2b\Phi}\, d\ell^2, \tag{3.5}$$

which on a hypersurface ℓ = constant is the 4-D conformal metric. The ten field equations and one scalar wave equation arising from $R_{ab} = 0$ are now

$$R_{\alpha\beta} = (b^2 - 2a^2 - 2ab) \Phi_\alpha \Phi_\beta + (2a + b) D_\beta \Phi_\alpha, \tag{3.6}$$

$$\Box \Phi = -(b + 2a) \Phi^\gamma \Phi_\gamma, \tag{3.7}$$

where as before $R_{\alpha\beta}$ is the Ricci tensor associated with $g_{\alpha\beta}$, $\Box \Phi \equiv g^{\alpha\beta} D_\alpha D_\beta \Phi$, and $\Phi_\alpha \equiv D_\alpha \Phi = \partial_\alpha \Phi$. The induced 4-D energy-momentum tensor defined by $T_{\alpha\beta} \equiv G_{\alpha\beta}$ is then

$$T_{\alpha\beta} = (b^2 - 2a^2 - 2ab) \Phi_\alpha \Phi_\beta + (2a + b) D_\beta \Phi_\alpha + 3a(b + a) \Phi^\gamma \Phi_\gamma g_{\alpha\beta}. \tag{3.8}$$

To simplify the above equations, it is convenient to make the parameter choice $2a+b=0$, which eliminates the covariant-derivative term $D_\beta \Phi_\alpha$ and reduces the metric, field equations, and energy-momentum tensor to

$$dS^2 = e^{2a\Phi} g_{\alpha\beta}(x^\gamma)\ dx^\alpha\ dx^\beta + \varepsilon e^{-4a\Phi}\ d\ell^2, \tag{3.9}$$

$$R_{\alpha\beta} = 6a^2 \Phi_\alpha \Phi_\beta, \tag{3.10}$$

$$\Box \Phi = 0, \tag{3.11}$$

$$T_{\alpha\beta} = 6a^2 (\Phi_\alpha \Phi_\beta - \tfrac{1}{2} g_{\alpha\beta} \Phi^\gamma \Phi_\gamma). \tag{3.12}$$

The equations (3.10) and (3.11) can be recognized as the field equations for a scalar field coupled to gravity with coupling $6a^2$. We note that the field equations do not constrain the constant a to be real. If we let $a=-i\alpha/\sqrt{6}$, the coupling changes sign ($6a^2=-\alpha^2$) and the metric becomes complex. Such solutions have been considered before, notably in Kaluza–Klein theory where quantities calculated from the metric turn out to be real,[18] and in Euclidean quantum gravity where they allow one to calculate the action.[19] We now proceed to show such a solution, which describes a plane wave of the type used in quantum theory to describe a moving particle.

If we consider a plane wave propagating in the z direction, the 4-D metric in terms of $u=t+z$ and $v=t-z$ can be expressed as

$$ds^2 = g_{\alpha\beta}\ dx^\alpha\ dx^\beta = du\,dv - e^{2\lambda}\ dx^2 - e^{2\mu}\ dy^2, \tag{3.13}$$

where λ and μ are functions of u only. The 5-D metric (with $a=-i\alpha/\sqrt{6}$) then has the form

$$dS^2 = e^{-2i\alpha\Phi/\sqrt{6}}(du\ dv - e^{2\lambda}\ dx^2 - e^{2\mu}\ dy^2) + \varepsilon e^{4i\alpha\Phi/\sqrt{6}}\ d\ell^2. \tag{3.14}$$

Here the scalar function is $\Phi=\Phi(u)$. A simple solution exists when $\mu=-\lambda=\alpha\Phi$, the constant α and the function Φ still being arbitrary. It is

$$dS^2 = e^{-2i\alpha/\sqrt{6}}(du\,dv - e^{2\alpha\Phi} dx^2 - e^{-2\alpha\Phi} dy^2) + \varepsilon e^{4i\alpha\Phi/\sqrt{6}}\ d\ell^2. \tag{3.15}$$

For this, $\Phi^\gamma \Phi_\gamma = 0$ and the energy-momentum tensor is found from (3.12) to be

$$T_{\alpha\beta} = -\alpha^2 \Phi_\alpha \Phi_\beta. \tag{3.16}$$

This is algebraically simple, but implies that the choice $a=-i\alpha/\sqrt{6}$ leads to a negative effective energy density:

$$T_{uu} = -\alpha^2 \left(\frac{\partial \Phi}{\partial u} \right)^2. \tag{3.17}$$

This may physically be interpreted as the energy density of a classical or de Sitter vacuum associated with a negative cosmological constant.[20]

The metric (3.15) can always be put into the form

$$dS^2 = e^{-2i\alpha\Phi/\sqrt{6}}(dt^2 - dz^2 - e^{2\alpha\Phi}\ dx^2 - e^{-2\alpha\Phi}\ dy^2) + \varepsilon e^{4i\alpha\Phi/\sqrt{6}}\ d\ell^2, \tag{3.18}$$

where $\Phi=\Phi(t,z)$. A simple solution to the wave equation (3.11) is now

$$\Phi = Et - pz. \tag{3.19}$$

Here the condition $\Phi^\gamma \Phi_\gamma = 0$ implies $E=\pm|p|$. The trace of the energy-momentum tensor for this specific case of Φ is

$$T = \alpha^2 (E^2 - p^2) = 0. \tag{3.20}$$

This implies by the quantum energy equation $E^2-p^2=m^2$ that we are dealing with a particle of zero rest mass. In fact (3.18) and (3.19) describe a wave propagating in the z direction which has associated with it a particle of energy E, momentum p, and zero rest mass. The energy-momentum tensor (3.8) effectively extends the induced-matter approach from fluids[3] to particles; and while the trace (3.20) is zero for the solution examined here, we expect to report in the future on corresponding solutions for massive particles.

IV. CONCLUSION

In Kaluza–Klein theory, the electromagnetic gauge[1,2] is mathematically restrictive, the matter gauge[3] is inconvenient for the physics of particles, and the harmonic gauge (1.1) and (1.3) is noncovariant. Therefore, it is advisable to use the lapse-and-shift technique[14–16] to provide a gauge-invariant decomposition of the 15 field equations of 5-D Kaluza–Klein theory into the sets of 10, 4, 1 useful in 4-D physics. These are given by (2.17)–(2.19), and their associated induced energy-momentum tensor is given by (2.21). However, if one wishes to apply a gauge, a useful one particularly in regard to wave equations and their associated particles is the conformal gauge (3.5). For this, the 4-D induced energy-momentum tensor in general form is (3.8), and for a special choice of parameters is (3.12). This gauge admits wavelike solutions. A simple plane-wave solution is (3.15) or (3.18). In the latter form, it depends on two constants which physically correspond to the energy and momentum of the particle associated with the wave. However, the trace of the associated energy-momentum tensor is zero (3.20), so the particle has zero rest mass.

While the decomposition from five to four dimensions in Sec. II here is general, what we have done in the conformal gauge in Sec. III can clearly be extended.

ACKNOWLEDGMENTS

We thank B. Mashhoon for comments, and O.G.S. and N.S.E.R.C. for financial support.

[1] T. Kaluza, Sitz. Preuss. Akad. Wiss. **33**, 966 (1921).
[2] O. Klein, Z. Phys. **37**, 895 (1926).
[3] P. S. Wesson and J. Ponce de Leon, J. Math. Phys. **33**, 3883 (1992).
[4] J. M. Overduin and P. S. Wesson, Phys. Rep. **283**, 303 (1997).
[5] P. S. Wesson, Ap. J. **436**, 547 (1994).
[6] D. Kalligas, P. S. Wesson, and C. W. F. Everitt, Astrophys. J. **439**, 548 (1995).
[7] H. Liu and P. Wesson, Phys. Lett. B **381**, 420 (1996).
[8] G. Abolghasem, A. A. Coley, and D. J. McManus, J. Math. Phys. **37**, 1 (1996).
[9] P. S. Wesson and J. Ponce de Leon, Gen. Relativ. Gravit. **26**, 555 (1994).
[10] T. Appelquist and A. Chodos, Phys. Rev. D **28**, 772 (1983).
[11] S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972), pp. 162, 253.
[12] G. Leibbrandt, *Non-Covariant Gauges* (World Scientific, Singapore, 1994).
[13] R. C. Myers and M. J. Perry, Ann. Phys. (N.Y.) **172**, 304 (1986).
[14] C. Misner, K. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), p. 506.
[15] S. W. Hawking and G. F. R. Ellis, *The Large-Scale Structure of Space-Time* (Cambridge U.P., Cambridge, 1973), p. 44.
[16] R. Wald, *General Relativity* (Univ. Chicago, Chicago, 1984), p. 255.
[17] H. Stephani and T. Wolf, in *Galaxies, Axisymmetric Systems and Relativity* (Cambridge U.P., Cambridge, 1985), p. 275.
[18] A. Billyard and P. S. Wesson, Gen. Relativ. Gravit. **28**, 129 (1996).
[19] G. W. Gibbons and S. W. Hawking, *Euclidean Quantum Gravity* (World Scientific, Singapore, 1993).
[20] R. N. Henriksen, A. G. Emslie, and P. S. Wesson, Phys. Rev. D **27**, 1219 (1983).

Geodesics and geodesic deviation for impulsive gravitational waves

R. Steinbauer[a)]
Institute for Theoretical Physics, University of Vienna, Boltzmanng. 5
A-1090 Wien, Austria

(Received 4 November 1997; accepted for publication 19 November 1997)

The geometry of impulsive pp-waves is explored via the analysis of the geodesic and geodesic deviation equation using the distributional form of the metric. The geodesic equation involves formally ill-defined products of distributions due to the nonlinearity of the equations and the presence of the Dirac δ-distribution in the space–time metric. Thus, strictly speaking, it cannot be treated within Schwartz's linear theory of distributions. To cope with this difficulty we proceed by first regularizing the δ-singularity, then solving the regularized equation within classical smooth functions and, finally, obtaining a distributional limit as solution to the original problem. Furthermore, it is shown that this limit is independent of the regularization without requiring any additional condition, thereby confirming earlier results in a mathematically rigorous fashion. We also treat the Jacobi equation which, despite being linear in the deviation vector field, involves even more delicate singular expressions, like the ''square'' of the Dirac δ-distribution. Again the same regularization procedure provides us with a perfectly well behaved smooth regularization and a regularization-independent distributional limit. Hence it is concluded that the geometry of impulsive pp-waves can be described consistently using distributions as long as careful regularization procedures are used to handle the ill-defined products. © *1998 American Institute of Physics.*
[S0022-2488(98)02204-X]

I. INTRODUCTION

Plane fronted gravitational waves with parallel rays (pp-waves) are space–times characterized by the existence of a covariantly constant null vector field, which can be used to write the metric tensor in the form[1]

$$ds^2 = H(u,x,y)du^2 - du\ dv + dx^2 + dy^2, \tag{1}$$

where u, v is a pair of null coordinates ($u=t-z$, $v=t+z$) and x, y are transverse (Cartesian) coordinates. In this paper we shall deal especially with impulsive (in the diction of Penrose[2]) pp-waves where the profile function H is proportional to a δ-distribution, i.e., takes the form $H(u,x,y)=f(x,y)\,\delta(u)$, where we leave the (smooth) function f of the transverse coordinates arbitrary for the moment. This metric is flat everywhere except on the null hypersurface $u=0$, where it has a δ-shaped ''shock.''

Such space–times arise most prominently as ultrarelativistic limits of black hole geometries as first derived by Aichelburg and Sexl for the Schwarzschild case.[3–5] On the other hand, Penrose has given a more intrinsic description of such space–times by his ''scissor and paste'' approach,[6] which essentially consists of glueing together two pieces of Minkowski space–time along the null hypersurface $u=0$ with a shift in the v-direction. A similar idea was used by Dray and t'Hooft,[7] who introduced a coordinate shift along geodesics in Minkowski space–time to rederive the Aichelburg–Sexl geometry as well as Penrose's junction conditions from the field equations.

The philosophy of the present work is somehow complementary. We take the δ-shaped metric literally and try to explore the properties of this geometry via investigation of geodesics and the

a)Electronic mail: stein@doppler.thp.univie.ac.at

 © 1998 American Institute of Physics

geodesic deviation, thereby following the approaches of Ferrari, Pendenza, and Veneziano[8] and Balasin.[9]

The main purpose of this work is to deal with the singular, i.e., distributional quantities in a mathematically rigorous fashion. A thorough analysis of the geodesic equation shows that it involves ill-defined products of the δ-distribution with the step function. This difficulty can be circumvented by a proper regularization procedure providing us with a perfectly well behaved smooth approximation, which has a distributional limit coinciding with the earlier results.[8,9] However, we neither have to impose ''multiplication rules'' like $\delta\theta=(1/2)\delta$, nor to use any additional requirements as the constancy of the norm of the geodesic's tangent vector across the shock. In fact, this property comes out as a result in our approach. The details of this calculation are given in Sec. II.

Further investigations show that the Jacobi equation involves ill-defined terms of an even worse type, such as the ''square'' of the Dirac δ-distribution and the square of the step function times the δ-distribution. However, our regularization strategy again provides us with a smooth approximation and a well behaved distributional limit, even in this case, where it seems to be hopeless to use ''*ad-hoc* extensions'' of Schwartz's linear distribution theory such as special ''multiplication rules.'' We give the details of the rather lengthy calculation in Sec. III. Our distributional ''solution'' of the Jacobi equation fits perfectly well in the (heuristically) expected picture, showing the consistency of our approach. Finally, we make some comments and give an outlook to future work in Sec. IV.

II. GEODESIC EQUATION

We start with an impulsive pp-wave metric of the form

$$ds^2=f(x^i)\,\delta(u)du^2-du\;dv+(dx^i)^2, \tag{2}$$

where $x^i(i=1,2)$ denote the transverse coordinates. It is straightforward to derive the geodesic equation. The nonvanishing Christoffel symbols are

$$\Gamma^v_{uu}=-f\dot{\delta},\quad \Gamma^i_{uu}=-\tfrac{1}{2}\partial_i f\,\delta,\quad \Gamma^v_{ui}=\Gamma^v_{iu}=-\partial_i f\,\delta, \tag{3}$$

where we have denoted the partial derivatives of f by $\partial_i f$ and the derivative of the δ-distribution by $\dot{\delta}$. Hence we get the equations

$$u''=0,\quad v''=f\dot{\delta}+2(\partial_i f x^{i\prime})u'\delta,\quad x^{i\prime\prime}=\tfrac{1}{2}\partial_i f u'^2\delta, \tag{4}$$

where $'$ denotes the derivative with respect to an affine parameter and summation over i is understood. We use the first equation to introduce u as a new affine parameter (there excluding trivial geodesics parallel to the shock hypersurface) to get

$$\begin{aligned}\ddot{v}(u)&=f(x^i(u))\,\dot{\delta}(u)+2(\partial_i f(x^i(u))\dot{x}^i(u))\,\delta(u),\\ \ddot{x}^i(u)&=\tfrac{1}{2}\partial_i f(x^i(u))\,\delta(u),\end{aligned} \tag{5}$$

where $\cdot$ again denotes the derivative with respect to u and we have inserted all the dependences explicitly. Equations (5) form a system of three coupled, nonlinear ODEs of second order in the vectorspace $\mathscr{D}'$ of distributions. For $u\neq 0$ all the right hand sides vanish, which is clear from the form of the metric tensor, and we expect the geodesics to be broken, possibly refracted straight lines. However, if we take a closer look at system (5) we see immediately that the first equation cannot be taken literally in the sense of distributions as the terms $\dot{x}^i\delta$ involve ill-defined products of the δ-distribution with the step function.

To analyze the situation in some more detail we integrate the last two equations using the (distributional) identity $f(x^i(u))\,\delta(u)=f(x^i(0))\,\delta(u)$ to get

$$x^i(u)=\text{initial values}+\tfrac{1}{2}\partial_i f(x^i(0))u_+, \tag{6}$$

where we have denoted the ''kink''-function $u\theta(u)$ by u_+. Note, however, that distributions can only be multiplied by C^∞-functions, whereas it's not clear *a priori* that the solution $x^i(u)$ will be smooth; in fact as suggested by Eq. (6) and as shown later in our calculations the solution will not even be differentiable at $u=0$. If we still try to go on by brute force it comes even worse: inserting (6) into the first equation (5) we see that the term $\dot{x}^i(u)\delta(u)$ gives rise to the ill-defined product $\theta\delta$. (For some further comments on this product see Ref. 10, p. 21.)

To overcome the undefinedness described in detail above, we apply a careful regularization procedure. More precisely we regularize the δ-distribution by a (standard) mollifier or model δ-net, i.e., a net ρ_ϵ, defined as follows. Let ρ be a smooth function with support contained in the interval $[-1,1]$ and $\int\rho=1$; now put $\rho_\epsilon(x)=(1/\epsilon)\rho(x/\epsilon)$, $(\epsilon>0)$. Hence system (5) takes the regularized form

$$\begin{aligned}\ddot{v}_\epsilon(u)&=f(x^i_\epsilon(u))\dot{\rho}_\epsilon(u)+2\partial_i f(x^i_\epsilon(u))\dot{x}^i_\epsilon(u)\rho_\epsilon(u),\\ \ddot{x}^i_\epsilon(u)&=\tfrac{1}{2}\partial_i f(x^i_\epsilon(u))\rho_\epsilon(u).\end{aligned}\tag{7}$$

Next we choose initial conditions in $u=-1$ (i.e., ''long before the shock''), more precisely

$$\begin{aligned}v_\epsilon(-1)&=v_0,\quad x^i_\epsilon(-1)=x^i_0,\\ \dot{v}_\epsilon(-1)&=\dot{v}_0,\quad \dot{x}^i_\epsilon(-1)=\dot{x}^i_0,\end{aligned}\tag{8}$$

for all ϵ. Note that choosing initial conditions in $u=0$ would mean to ''start at the shock'' and one cannot expect to (and indeed does) end up with a regularization independent result in this case.

Due to the regularization procedure we now have to deal with the fully nonlinear character of the last equation of (7). However, as shown in appendix A the special form of the right hand side of the equation guarantees global existence of the solutions for small ϵ. These are given implicitly by

$$\begin{aligned}v_\epsilon(u)&=v_0+\dot{v}_0(1+u)+(\theta*\theta*\ddot{v}_\epsilon)(u),\\ x^i_\epsilon(u)&=x^i_0+\dot{x}^i_0(1+u)+(\theta*\theta*\ddot{x}^i_\epsilon)(u),\end{aligned}\tag{9}$$

where ''$*$'' denotes convolution.

We are going to calculate the (distributional) limits of the solutions (9) as ϵ tends to zero. Since distributions supported in an acute cone form a convolution algebra (where, in particular, convolution is a separately continuous operation) it suffices to calculate the limits of the right hand sides of (7); the distributional limits of the solutions (9) are then computed simply by integration.

We begin with the latter two equations of system (7) and choose a test function φ. We have to calculate the limit of

$$\frac{1}{2}\int_{\mathbb{R}}\partial_i f(x^i_\epsilon(u))\rho_\epsilon(u)\varphi(u)du=\frac{1}{2}\int_{-1}^{1}\partial_i f(x^i_\epsilon(\epsilon u))\rho(u)\varphi(\epsilon u)du.\tag{10}$$

Since (for small ϵ) $x_\epsilon(u)$ is bounded uniformly on compact sets (see the Appendix) we can use Lebesgue's dominated convergence. Hence the only term we have to compute is

$$\begin{aligned}\lim_{\epsilon\to 0}x^i_\epsilon(\epsilon u)&=\lim\left(x^i_0+\dot{x}^i_o(1+\epsilon u)+\frac{1}{2}\int_{-\epsilon}^{\epsilon u}\int_{-\epsilon}^{v}\partial_i f(x^i_\epsilon(s))\rho_\epsilon(s)ds\ dv\right)\\ &=x^i_o+\dot{x}^i_o+\frac{1}{2}\lim\underbrace{\int_{-\epsilon}^{\epsilon u}\int_{-1}^{v/\epsilon}\partial_i f(x^i_\epsilon(\epsilon s))\rho(s)ds\ dv}_{|\int\int\leqslant\|\partial_i f\|\,\|\rho\|\int\int ds\ dv}.\end{aligned}\tag{11}$$

For $\epsilon\to 0$ the last integral gives zero, since in the limit the range of integration only covers a set of zero measure. Hence we get (as expected)

$$\ddot{x}^i_\epsilon(u)=\tfrac{1}{2}\partial_i f(x^i_\epsilon(u))\rho_\epsilon(u)\to\tfrac{1}{2}\partial_i f(x^i_0+\dot{x}^i_0)\,\delta(u), \tag{12}$$

within distributions.

Next we turn to the ''critical'' first equation in (7) and again calculate the limit of the right hand side, i.e., of the expression

$$\begin{aligned}\ddot{v}_\epsilon(u)&=f(x^i_\epsilon(u))\dot{\rho}_\epsilon(u)+2\partial_i f(x^i_\epsilon(u))\dot{x}^i_\epsilon(u)\rho_\epsilon(u)\\&=(f(x^i_\epsilon(u))\rho_\epsilon(u))^{\cdot}+\partial_i f(x^i_\epsilon(u))\dot{x}^i_\epsilon(u)\rho_\epsilon(u).\end{aligned} \tag{13}$$

The limit of the first term is easily seen to be $f(x^i_0+\dot{x}^i_0)\,\dot{\delta}(u)$, and we are only left with the second term which contains the ''critical products.'' Inserting (9) and (7) respectively, we have

$$\begin{aligned}\partial_i f(x^i_\epsilon)\dot{x}^i_\epsilon\rho_\epsilon&=\partial_i f(x^i_\epsilon)(\dot{x}^i_0+\theta*\ddot{x}^i_\epsilon)\rho_\epsilon\\&=\partial_i f(x^i_\epsilon)\dot{x}^i_0\rho_\epsilon+\partial_i f(x^i_\epsilon)(\theta*[\tfrac{1}{2}\partial_i f(x^i_\epsilon)\rho_\epsilon])\rho_\epsilon.\end{aligned} \tag{14}$$

The limit of the first term again is easily seen to give $\dot{x}^i_0\partial_i f(x^i_0+\dot{x}^i_0)\,\delta(u)$, and we are finally left with the task of computing the limit of the convolution term, i.e., the expression ''$\theta\delta$.'' Using

$$\begin{aligned}(\theta*[\partial_i f(x^i_\epsilon)\rho_\epsilon])(u)&=\int_{-\infty}^{u}\partial_i f(x^i_\epsilon(s))\rho_\epsilon(s)ds\\&=\int_{-1}^{u/\epsilon}\partial_i f(x^i_\epsilon(\epsilon s))\rho(s)ds\end{aligned} \tag{15}$$

and again denoting by φ a test function we get

$$\begin{aligned}&\lim_{\epsilon\to 0}\left(\frac{1}{2}\int_{-1}^{1}\partial_i f(x^i_\epsilon(\epsilon u))\left[\int_{-1}^{u}\partial_i f(x^i_\epsilon(\epsilon s))\rho(s)ds\right]\rho(u)\varphi(\epsilon u)du\right)\\&=\frac{1}{2}(\partial_i f(x^i_0+\dot{x}^i_0))^2\varphi(0)\underbrace{\int_{1}^{1}\left[\int_{-1}^{u}\rho(s)ds\right]\rho(u)du}_{1/2}\\&=\frac{1}{4}(\partial_i F(x^i_0+\dot{x}^i_o))^2\varphi(0).\end{aligned} \tag{16}$$

Collecting things together we have

$$\lim_{\epsilon\to 0}\ddot{v}_\epsilon=f(x^i_0+\dot{x}^i_0)\,\dot{\delta}+\dot{x}^i_0\partial_i f(x^i_0+\dot{x}^i_0)\,\delta+\tfrac{1}{4}(\partial_i f(x^i_0+\dot{x}^i_0))^2\delta \tag{17}$$

within distributions.

At the end of this section let us give a summary on what we have done so far. We have regularized the distributionally ill-defined geodesic equations (5) by replacing the δ-distribution by a (generic) class of mollifiers to obtain (7). These equations, although nonlinear, provide us with global solutions v_ϵ and x^i_ϵ for small ϵ (as shown in the Appendix), implicitly given by (9). Using the latter formula we have shown that the smooth solutions have a regularization independent distributional limit given by

$$\begin{aligned}v_\epsilon(u)&\to v_0+\dot{v}_0(1+u)+f(0)\,\theta(u)+\partial_i f(0)(\dot{x}^i_0+\tfrac{1}{4}\partial_i f(0))u_+,\\x^i_\epsilon(u)&\to x^i_0+\dot{x}^i_0(1+u)+\tfrac{1}{2}\partial_i f(0)u_+,\end{aligned} \tag{18}$$

where we have used the abbreviation $f(0)=f(x^i_0+\dot{x}^i_0)$.

Hence, viewed distributionally the geodesics are given by refracted, broken straight lines as expected. Of course, equations (18) coincide with the earlier results.[8,9] However, from the point of view of our approach the (deeper) reason why here the ''rule'' $\theta\delta=(1/2)\,\delta$ used by Ref. 8 [which

in fact coincides with the ''determination of the point value'' $\theta(0)=1/2$ used by Ref. 9] leads to a physically reasonable result is the following: The geodesic equations involve *only one* singular object and hence the δ's as well as the θ's appearing above share the same root: namely the δ-shaped wave profile. Hence, when regularizing the equations (which in fact corresponds to the physical idea of viewing the impulsive wave as an idealized sandwich wave) both factors of the ill-defined product naturally involve the same regularization which immediately leads to the (regularization-independent) result $\rho_\epsilon \int \rho_\epsilon \to (1/2)\,\delta$.

We thus conclude that the geodesic equation can be treated consistently by regularization, leading to a regularization-independent distributional result. This, of course, is only possible at all due to the relatively mild character of the singular terms which allows for a distributional limit of the solutions to the regularized problem. However, we shall see in a moment that even in the considerably more complicated case of the Jacobi equation our strategy can be applied successfully.

III. JACOBI EQUATION

In this section we solve the Jacobi equation for an impulsive pp-wave. To keep formulas more transparent we restrict ourselves to the axisymmetric case. More precisely we restrict the function f of the transverse coordinates $x^i=(x,y)$ in the metric tensor (2) to depend on the two-radius $r=\sqrt{x^2+y^2}$ only and work entirely within the $y=0$-hypersurface (initial conditions $x_0^2\equiv y_0=0=\dot{x}_0^2\equiv\dot{y}_0$). Furthermore, we take initial values $v_0=0=\dot{x}_0$. With these assumptions equations (18) simplify to

$$\begin{aligned} v_\epsilon(u) &\to \dot{v}_0(1+u)+f(x_0)\,\theta(u)+\tfrac{1}{4}f'(x_0)^2 u_+ ,\\ x_\epsilon^1(u)\equiv x_\epsilon(u) &\to x_0+\tfrac{1}{2}f'(x_0)u_+ ,\\ x_\epsilon^2(u)\equiv y_\epsilon(u) &= 0, \end{aligned} \tag{19}$$

where f' denotes the derivative with respect to the single variable $r=x$. Hence we shall deal with a geodesic tangent vector of the form

$$T^a(u)=\begin{pmatrix} 1 \\ \dot{v}_0+f(x_0)\,\delta(u)+\tfrac{1}{4}f'(x_0)^2\,\theta(u) \\ \tfrac{1}{2}f'(x_0)\,\theta(u) \\ 0 \end{pmatrix}, \tag{20}$$

where we are going to use the abbreviations $A:=f(x_0)\,\delta(u)+(1/4)f'(x_0)^2\theta(u)$ and $B:=(1/2)f'(x_0)\,\theta(u)$ for its components.

Our next task is to compute the explicit form of the Jacobi equation $D^2N^a/du=-R^a_{bcd}T^bT^dN^c$ for a vector field $N^a(u)=(N^u(u),N^v(u),N^x(u),N^y(u))$ over the geodesic. After some (tedious) calculations we end up with the following form of the system:

$$\begin{aligned} &\ddot{N}^u=0, \quad \ddot{N}^v=2[N^xf'\delta]^{\cdot}-N^xf'\dot{\delta}+[N^uf\delta]^{\cdot\cdot}-N^uf''B^2\delta-N^uf'\dot{B}\delta,\\ &\ddot{N}^x=[\dot{N}^uf'+\tfrac{1}{2}N^xf'']\delta+\tfrac{1}{2}f'N^u\dot{\delta}, \quad \ddot{N}^y=0, \end{aligned} \tag{21}$$

where we have suppressed the dependence on the parameter u and the variable x. Equations (21) form a system of four coupled ODE's linear in the components of the vector field N^a but nonlinear in the derivatives of the metric. From the fact that B involves the step function we see immediately that (in the second equation) we again have to deal with distributionally ill-defined expressions, but now of even worse type than before. Indeed, the term $\dot{B}\delta$ is proportional to the ''square'' of the Dirac δ-distribution, and the term $B^2\delta$ involves an expression ''$\theta^2\delta$.'' Note, however, that the critical terms arise from the second covariant derivative, where some of the Christoffel symbols get multiplied, and not from the Riemann tensor in which components are just proportional to the δ-distribution. To overcome these problems we apply the same regularization procedure as in the case of the geodesic equation. In particular, we use the regularized geodesic tangent vector

$$T^a_\epsilon = \begin{pmatrix} 1 \\ \dot v_0 + \theta * \ddot v_\epsilon \\ \theta * \ddot x_\epsilon \\ 0 \end{pmatrix}, \tag{22}$$

where $\ddot v_\epsilon(u) = f(x_\epsilon(u))\dot\rho_\epsilon(u) + 2f'(x_\epsilon(u))\dot x_\epsilon(u)\rho_\epsilon(u)$, $\ddot x_\epsilon(u) = (1/2)f'(x_\epsilon(u))\rho_\epsilon(u)$ and from now on we use the new abbreviations $A_\epsilon := \theta * \ddot v_\epsilon$ and $B_\epsilon := \theta * \ddot x_\epsilon$. Denoting the regularized Jacobi field by $N^a_\epsilon(u) = (N^u_\epsilon(u), N^v_\epsilon(u), N^x_\epsilon(u), N^y_\epsilon(u))$, system (21) takes the regularized form

$$\ddot N^u_\epsilon = 0,$$

$$\ddot N^v_\epsilon = 2[N^x_\epsilon f'(x_\epsilon)\rho_\epsilon]^{\cdot} - N^x_\epsilon f'(x_\epsilon)\dot\rho_\epsilon + [N^u_\epsilon f(x_\epsilon)\rho_\epsilon]^{\cdot\cdot} - N^u_\epsilon f''(x_\epsilon)[\theta * \ddot x_\epsilon]^2\rho_\epsilon - N^u_\epsilon f'(x_\epsilon)\ddot x_\epsilon\rho_\epsilon,$$

$$\ddot N^x_\epsilon = [\dot N^u_\epsilon f'(x_\epsilon) + \tfrac{1}{2}N^x_\epsilon f''(x_\epsilon)]\rho_\epsilon + \tfrac{1}{2}f'(x_\epsilon)N^u_\epsilon\dot\rho_\epsilon, \quad \ddot N^y_\epsilon = 0. \tag{23}$$

To maintain transparency of formulas we choose appropriate and simple initial conditions on the Jacobi field N^a_ϵ at $u = -1$, i.e.,

$$N^a_\epsilon(-1) = (0,0,0,0), \quad \dot N^a_\epsilon(-1) = (a,b,0,0), \tag{24}$$

for all ϵ. Note that this corresponds to a focal point at $u = -1$, and that ''nearby'' geodesics have relative initial velocities only in the u- and v-direction.

The first and the last equation of system (23) are easily solved to give

$$N^u_\epsilon(u) = a(1+u), \quad N^y_\epsilon(u) = 0. \tag{25}$$

For the remaining, more complicated equations of system (23) we apply the same strategy as in the preceeding section. Since the equations are linear in the components of the deviation vector field we obtain globally defined (smooth) solutions which, due to the initial conditions (24), are implicitly given by

$$N^v_\epsilon(u) = b(1+u) + [\theta * \theta * \ddot N^v_\epsilon](u), \quad N^x_\epsilon(u) = [\theta * \theta * \ddot N^x_\epsilon](u). \tag{26}$$

Again it suffices to compute the distributional limits of the right hand sides of equation (23), since by continuity of the convolution we immediately get the limits of N^v_ϵ and N^x_ϵ.

We start with the third equation of the system (23). The main problem we have to face here is due to the fact that the unknown function N^x_ϵ (which, in the limit, we cannot even expect to be continuous at $u = 0$) appears on the right hand side. Inserting the initial conditions (24) and the solutions (25) we get

$$\ddot N^x_\epsilon(u) = \underbrace{af'(x_\epsilon(u))\rho_\epsilon(u)}_{I_\epsilon} + \underbrace{\tfrac{1}{2}f'(x_\epsilon(u))a(1+u)\dot\rho_\epsilon(u)}_{II_\epsilon} + \underbrace{\tfrac{1}{2}N^x_\epsilon(u)f''(x_\epsilon(u))\rho_\epsilon(u)}_{III_\epsilon}. \tag{27}$$

The distributional limits of the first two terms in equation (27) are easily seen (using similar methods as in the previous section) to be

$$\lim_{\epsilon\to 0} \mathrm{I}_\epsilon = af'(x_0)\delta,$$

$$\lim_{\epsilon\to 0} \mathrm{II}_\epsilon = \tfrac{1}{2}af'(x_0)(\dot\delta - \delta) - \tfrac{1}{8}af'(x_0)f''(x_0)\delta. \tag{28}$$

However, the last term in equation (27) requires a more detailed analysis. Inserting again $\theta * \theta * \ddot N^x_\epsilon$ for the unknown function N^x_ϵ and denoting by φ a test function we obtain

$$\mathrm{III}_\epsilon(u)=\frac{1}{2}\,\epsilon\int_{-1}^{1}\varphi(\epsilon u)f''(x_\epsilon(\epsilon u))\rho(u)\int_{-1}^{u}\int_{-1}^{s}\Big[af'(x_\epsilon(\epsilon r))$$
$$+\frac{1}{2}\,N^x_\epsilon(\epsilon r)f''(x_\epsilon(\epsilon r))\Big]\rho(r)dr\;ds\;du$$
$$+\frac{1}{4}\,a\int_{-1}^{1}\varphi(\epsilon u)f''(x_\epsilon(\epsilon u))\rho(u)\int_{-1}^{u}\int_{-1}^{s}f'(x_\epsilon(\epsilon r))\dot\rho(r)dr\;ds\;du$$
$$+\frac{1}{4}\,a\epsilon\int_{-1}^{1}\varphi(\epsilon u)f''(x_\epsilon(\epsilon u))\rho(u)\int_{-1}^{u}\int_{-1}^{s}f'(x_\epsilon(\epsilon r))r\dot\rho(r)dr\;ds\;du. \tag{29}$$

The last term in equation (29) goes to zero, since φ, f', f'' are smooth functions and x_ϵ is bounded uniformly on compact sets (see the Appendix). For the same reasons the second summand in equation (29) approaches the limit

$$\tfrac{1}{8}a\varphi(0)f'(x_0)f''(x_0), \tag{30}$$

whereas the first line also vanishes in the limit, since $N^x_\epsilon(\epsilon r)$ is bounded on the compact range of integration. (Note, however, that $\lim N^x_\epsilon$ is not continuous at 0!) In some more detail the latter argument can be seen from

$$N^x_\epsilon(\epsilon u)=\int_{-1}^{u}\int_{-1}^{s}\Big\{af'(x_\epsilon(\epsilon r))\epsilon\rho(r)+\frac{1}{2}\,af'(x_\epsilon(\epsilon r))[\dot\rho(r)+r\epsilon\dot\rho(r)]\Big\}dr\;ds$$
$$+\frac{1}{2}\int_{-\epsilon}^{\epsilon u}\int_{-1}^{s/\epsilon}N^x_\epsilon(\epsilon r)f''(x_\epsilon(\epsilon r))\rho(r)dr\;ds, \tag{31}$$

where we have now used equation (26) for the third time. Hence we have for $u\leqslant T$, where T is some constant, and for fixed ϵ

$$|N^x_\epsilon(\epsilon u)|\leqslant C+C'\int_{-\epsilon}^{\epsilon u}\int_{-1}^{s/\epsilon}|N^x_\epsilon(\epsilon r)||\rho(r)|dr\;ds\leqslant C+2C'\int_{-1}^{u}|N^x_\epsilon(\epsilon r)||\rho(r)|dr, \tag{32}$$

where $C=C(T)$ and C' denote constants. From Gronwall's inequality we gain

$$|N^x_\epsilon(\epsilon u)|\leqslant Ce^{2C'\int_{-1}^{u}|\rho(r)|dr}, \tag{33}$$

which is bounded by a constant for $u\leqslant T$.

Collecting things together [equations (28) and (30)] we find that within distributions

$$\ddot N^x_\epsilon\to\tfrac{1}{2}af'(x_0)(\delta+\dot\delta), \tag{34}$$

which gives us immediately the distributional limit of the regularized solution of the third equation of system (23), i.e.,

$$N^x_\epsilon(u)\to\tfrac{1}{2}af'(x_0)(u_++\theta(u)). \tag{35}$$

Now we face the most complicated equation of system (23) which contains the distributionally ill-defined expressions, namely

$$\ddot N^v_\epsilon(u)=2[N^x_\epsilon(u)f'(x_\epsilon(u))\rho_\epsilon(u)]^{\cdot}+[N^u_\epsilon(u)f(x_\epsilon(u))\rho_\epsilon(u)]^{\cdot\cdot}-N^u_\epsilon(u)f''(x_\epsilon(u))[\theta*\ddot x_\epsilon]^2(u)\rho_\epsilon(u)$$
$$-N^x_\epsilon(u)f'(x_\epsilon(u))\dot\rho_\epsilon(u)-N^u_\epsilon(u)f'(x_\epsilon(u))\ddot x_\epsilon(u)\rho_\epsilon(u), \tag{36}$$

where we have inserted all the dependences explicitly. Note that in addition to the obvious troublesome fifth term which involves the ''square'' of the δ-distribution ($\ddot x_\epsilon\propto\rho_\epsilon$!), also the fourth term contains divergent and regularization-dependent contributions, since the derivatives of

$f(x)$ and N_ϵ^x produce δ- and step functions, respectively. We are going to calculate the distributional limits of the right hand side of equation (36) term by term, which can easily be done for the first three using yet well known techniques.

First, second and third term:

$$\lim_{\epsilon\to 0} 2[N_\epsilon^x(u) f'(x_\epsilon(u))\rho_\epsilon(u)]^{\cdot} = \tfrac{1}{2} a f'(x_0)^2 \dot{\delta},$$

$$\lim_{\epsilon\to 0} [N_\epsilon^u(u) f(x_\epsilon(u))\rho_\epsilon(u)]^{\cdot\cdot} = a f(x_0)\ddot{\delta}, \tag{37}$$

$$\lim_{\epsilon\to 0} N_\epsilon^u(u) f''(x_\epsilon(u))[\theta * \ddot{x}_\epsilon]^2(u)\rho_\epsilon(u) = \tfrac{1}{12} a f'(x_0)^2 f''(x_0)\delta.$$

We now come to the expressions involving the divergent and regularization-dependent factors. We outline the calculations in some detail.

Fourth term: $(\mathrm{iv}) := -N_\epsilon^x(u) f'(x_\epsilon(u))\dot{\rho}_\epsilon(u)$.
By integration by parts we obtain (φ again a test function)

$$-\int_{-\epsilon}^{\epsilon} \varphi(u) N_\epsilon^x(u) f'(x_\epsilon(u))\dot{\rho}_\epsilon(u)\,du = \underbrace{\int_{-1}^{1} \dot{\varphi}(\epsilon u) N_\epsilon^x f'(x_\epsilon(\epsilon u))\rho(u)\,du}_{A}$$

$$+ \underbrace{\int_{-1}^{1} \varphi(\epsilon u) N_\epsilon^x(\epsilon u) f''(x_\epsilon(\epsilon u))\dot{x}_\epsilon(\epsilon u))\rho(u)\,du}_{B} + \underbrace{\int_{-1}^{1} \varphi(\epsilon u)\dot{N}_\epsilon^x(\epsilon u) f'(x_\epsilon(\epsilon u))\rho(u)\,du}_{C}. \tag{38}$$

Inserting $\theta * \theta * \ddot{N}_\epsilon^x$ according to equations (26) and (27) for N_ϵ^x we find for the expressions labeled A and B

$$A \to -\tfrac{1}{4} a f'(x_0)^2 \dot{\delta}, \tag{39}$$

$$B \to \tfrac{1}{12} a f'(x_0)^2 f''(x_0)\delta. \tag{40}$$

Finally, we have for the most complicated expression

$$C = \frac{1}{2} a \underbrace{\int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))\rho(u) \int_{-1}^{u} f'(x_\epsilon(\epsilon s))\dot{\rho}(s)\,ds\,du}_{C1}$$

$$+ \frac{1}{2} a \underbrace{\int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))\rho(u) \int_{-1}^{u} f'(x_\epsilon(\epsilon s)) s\dot{\rho}(s)\,ds\,du}_{C2}$$

$$+ \underbrace{\int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))\rho(u) \int_{-1}^{u} [a f'(x_\epsilon(\epsilon s)) + \frac{1}{2} N_\epsilon^x(\epsilon s) f''(x_\epsilon(\epsilon s))]\rho(s)\,ds\,du}_{C3}. \tag{41}$$

The term labeled $C1$ diverges due to the factor ϵ^{-1}. However, as we shall see later, this expression combined with another one arising from the fifth term of equation (36) will provide us with a finite result.

The term labeled $C2$ involves a regularization-dependent factor, i.e.,

$$C2 \to \frac{1}{2}\, af'(x_0)^2 \delta \int_{-1}^{1} \rho(u) \int_{-1}^{u} s\dot{\rho}(s) ds\ dt. \tag{42}$$

As in the case of the previous term $C1$, this problem will be resolved later on, so that we are now left with the task of computing the limit of $C3$. Again using similar techniques we obtain

$$C3 \to \tfrac{1}{2} af'(x_0)^2 \delta + \tfrac{1}{24} af'(x_0)^2 f''(x_0)\delta. \tag{43}$$

Collecting together the results of equations (39), (40) and (43), we finally have

$$(\mathrm{iv}) - C1 - C2 \to -\tfrac{1}{4} af'(x_0)^2 \dot{\delta} + \tfrac{1}{8} af'(x_0)^2 f''(x_0)\delta + \tfrac{1}{2} af'(x_0)^2 \delta, \tag{44}$$

where now, by an abuse of notation $C1$ and $C2$ denote the respective quantities in equation (41) without integration over the test function.

Now we finally come to the last term in equation (36) which will resolve the problems with the expressions $C1$ and $C2$.

Fifth term: $(\mathrm{v}) := -N_\epsilon^u(u) f'(x_\epsilon(u))\ddot{x}_\epsilon(u)\rho_\epsilon(u)$
Inserting solution (25) we find

$$-\int_{-\epsilon}^{\epsilon} \varphi(u) N_\epsilon^u(u) f'(x_\epsilon(u))\ddot{x}_\epsilon(u)\rho_\epsilon(u) du = \underbrace{-\frac{1}{2\epsilon} a \int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))^2 \rho(u)^2 du}_{(*)}$$
$$\underbrace{-\frac{1}{2} a \int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))^2 u \rho(u)^2 du}_{(**)}. \tag{45}$$

We see now that the expression (*) diverges, but that the sum $C1+(*)$ converges. More precisely, by integration by parts we get

$$C_1 + (*) = -\frac{1}{4}\, a \int_{-1}^{1} \varphi(\epsilon u) f'(x_\epsilon(\epsilon u))\rho(u) \int_{-1}^{u} f''(x_\epsilon(\epsilon s))\rho(s) \int_{-1}^{s} f'(x_\epsilon(\epsilon r))\rho(r) dr\ ds\ du$$
$$\to -\frac{1}{4}\, a\varphi(0) f'(x_0)^2 f''(x_0) \underbrace{\int_{-1}^{1} \rho(u) \int_{-1}^{u} \rho(s) \int_{-1}^{s} \rho(r) dr\ ds\ du}_{1/6}$$
$$= -\frac{1}{24}\, a\varphi(0) f'(x_0)^2 f''(x_0). \tag{46}$$

Hence we have the distributional limit

$$C1 + (*) \to -\tfrac{1}{24}\, af'(x_0)^2 f''(x_0)\delta. \tag{47}$$

On the other hand expression (**) contains a regularization-dependent factor which is proportional to the one contained in the term $C2$. More precisely, from equation (42) and the fact that

$$(**) \to -\tfrac{1}{2}\, af'(x_0)^2 \delta \int_{-1}^{1} u\rho(u)^2\ du \tag{48}$$

we have

$$C2+(**)\to\frac{1}{2}af'(x_0)^2\delta\underbrace{\left[\int_{-1}^{1}\rho(u)\left[\int_{-1}^{u}s\dot{\rho}(s)ds-u\rho(u)\right]du\right]}_{1/2}$$

$$=-\frac{1}{4}af'(x_0)^2\delta. \tag{49}$$

We are now ready to collect together all five terms of equation (36), i.e., of the second equation of the regularized system (23). From equations (37), (44), (47) and (49) we get the following result:

$$\ddot{N}^v_\epsilon\to af(x_0)\ddot{\delta}+\tfrac{1}{4}af'(x_0)^2(\dot{\delta}+\delta), \tag{50}$$

which, using initial condition (24), we readily integrate to

$$N^v_\epsilon(u)\to b(1+u)+a[f(x_0)\delta+\tfrac{1}{4}f'(x_0)^2(\theta(u)+u_+)]. \tag{51}$$

Finally we have for the whole system (23) the following (distributional limits of the) solutions

$$\begin{aligned}
N^u_\epsilon(u)&=a(1+u),\\
N^v_\epsilon(u)&\to b(1+u)+a[f(x_0)\delta+\tfrac{1}{4}f'(x_0)^2(\theta(u)+u_+)],\\
N^x_\epsilon(u)&\to\tfrac{1}{2}af'(x_0)(\theta(u)+u_+),\\
N^y_\epsilon(u)&=0.
\end{aligned} \tag{52}$$

Hence, viewed distributionally, the Jacobi field suffers a kink, a jump and a δ-like pulse in the v-direction as well as a kink and jump in the x-direction overlapping the linear flat space behavior. These effects can be understood heuristically from the corresponding behavior of the geodesics, given by equation (18). The constant factor a, which gives the ''scale'' of all the nonlinear effects, arises from the ''time advance'' of the ''nearby'' geodesics, represented by the initial velocity of the Jacobi field in the u-direction [cf. initial conditions (24)].

Note, however, that this ''time advance'' is not the only effect generically generating deviations from the flat space behavior, but rather arises as an artifact of our initial conditions. One can easily show that different initial conditions on the deviation vector field, even without ''time advance,'' produce kinks and jumps as well. For example, let

$$N^a_\epsilon(-1)=(0,0,0,0)\quad\text{and}\quad N^a_\epsilon(-1)=(0,a,b,0); \tag{53}$$

then a similar (but now even simpler) calculation leads to the following (distributional limits of the) regularized Jacobi field

$$\begin{aligned}
N^u_\epsilon&=0,\\
N^v_\epsilon&\to a(1+u)+\tfrac{1}{4}bf'(x_0)f''(x_0)u_+ +bf'(x_0)\theta(u),\\
N^x_\epsilon&\to b(1+u)+\tfrac{1}{2}bf''(x_0)u_+,\\
N^y_\epsilon&=0.
\end{aligned} \tag{54}$$

The kink of the x-component of the deviation field now arises from the fact that a ''nearby'' geodesic passes the shock at an x-value of x_0+b, hence according to equation (18) suffers a kink of ''strength'' $(1/2)f'(x_0+b)$. Taylor expansion yields $f'(x_0+b)\approx f'(x_0)+f''(x_0)b$, such that

the kink difference of ''nearby'' geodesics is given by $(1/2)f''(x_0)b$, which is exactly the factor given in the third equation of (54). The kink and jump in v-direction can be explained by similar heuristic arguments.

IV. CONCLUSION

We have shown that the geodesic and Jacobi equation of impulsive pp-waves can be treated consistently using the distributional form of the space–time metric (2).

The main problem was to deal with products ill-defined within Schwartz's linear distribution theory. The geodesic equation involves terms proportional to ''$\theta\delta$'' while the Jacobi equation, although linear in the components of the deviation vector field, contains even more singular terms like ''δ^2'' and ''$\theta^2\delta$.'' Our strategy consisted essentially in a careful and general regularization procedure, replacing the δ-distribution by a mollifying sequence ρ_ϵ. While mathematically sound, this procedure corresponds to the physical idea of viewing the impulsive wave as the limiting case of a sandwich wave of ever decreasing support but constant strength. Our approach leads to smooth solutions of the regularized equations which possess a regularization-independent limit within the vector space of distributions. However, this is not obvious (cf., for example, Ref. 11), but rather should strengthen our trust in impulsive waves as reasonable solutions of Einstein's equations. Furthermore our distributional ''solutions'' of the geodesic and Jacobi equation perfectly coincide with physical expectations, showing that the geometry of impulsive pp-waves can be described in a consistent way using the distributional form of the metric, as long as one applies proper regularization methods as opposed to ''ad-hoc strategies'' involving certain ''multiplication rules.'' In fact when dealing with nonlinear operations (of a certain complexity) on singular, i.e., distributional data, reliable results can only be achieved by careful regularization procedures (cf. the comments on ''common errors'' even in the mathematical literature given by Hájek[12]).

However, from an even more formal point of view, the approach taken in this paper suffers from the fact that we do not really have an elaborated solution concept for the regularized equations. Such a notion is provided by the theory of Nonlinear Generalized Functions[10,13,14] due to Colombeau and others, where—loosely speaking—the regularized sequence is viewed as one mathematical object, i.e., a generalized function. Recently, singular, nonlinear ODEs (exactly the type of equations we were dealing with) have been studied in the context of this new mathematical framework,[15,16] providing a powerful tool to handle both, *singular* data and especially this type of *nonlinear* differential equations. Future work will be concerned with the analysis of the geodesic and Jacobi equation for pp-waves from the point of view of this new solution concept.

ACKNOWLEDGMENT

The author owes special thanks to M. Kunzinger for great support with the calculations and to M. Oberguggenberger and M. Grosser for many helpful discussions. This work was supported by Austrian Academy of Science, Ph.D. Programme, Grant No. 338.

APPENDIX

In this appendix we prove (for small ϵ) the global existence and uniform boundedness (in ϵ) of the solutions to the last equation of system (7). For simplicity we only treat the one-dimensional case and write g instead of Df. The main idea is to use $\|\ \|_1$-bounds on the regularizing sequence ρ_ϵ. We start with the following statement.

Proposition: Let $g:\mathbb{R}\to\mathbb{R}$ *smooth and* $\rho\in\mathscr{D}([-1,1])$, $\int\rho=1$, $\rho_\epsilon(x):=(1/\epsilon)\rho(x/\epsilon)$. *For fixed* ϵ *consider the differential equation*

$$\ddot{x}_\epsilon(t)=g(x_\epsilon(t))\rho_\epsilon(t), \tag{A1}$$

with initial conditions $x_\epsilon(-1)=x_0$ *and* $\dot{x}_\epsilon(-1)=\dot{x}_0>0$. *Then there exists a unique solution* $x_\epsilon(t)$ *on the interval* $J_\epsilon:=[-1,-\epsilon+\alpha]$, *where*

$$\alpha:=\min\left\{\frac{b}{\|g\|_{\infty,I}\|\rho\|_1+\dot{x}_0},\frac{1}{2L}\right\}\quad(b>0), \tag{A2}$$

$I := \{x \in \mathbb{R} : |x - x_0| \leq b + \dot{x}_0\}$ *and* L *is a Lipschitz constant for* g *on* I.

The *Proof* is just a modification of the proof of the classical (first order) ODE existence and uniqueness theorem. We work on the (nonempty, closed) subset $X_\epsilon := \{x_\epsilon \in C(J_\epsilon) : |x_\epsilon(t) - x_0| \leq b + \dot{x}_0\}$ of the Banach space of continuous functions on the interval J_ϵ. Since no first order derivative of x_ϵ enters the right hand side of equation (A1) we may define the integration operator A by

$$(Ax_\epsilon)(t) := x_0 + \dot{x}_0(t+1) + \int_{-\epsilon}^{t}\int_{-\epsilon}^{s} g(x_\epsilon(r))\rho_\epsilon(r)\,dr\,ds. \tag{A3}$$

By construction $Ax_\epsilon \in C(J_\epsilon)$, but we even have $Ax_\epsilon \in X_\epsilon$ since

$$\begin{aligned} |[Ax_\epsilon](t) - x_0| &\leq \dot{x}_0(t+1) + \int_{-\epsilon}^{t}\int_{-\epsilon}^{s} |g(x_\epsilon(r))|\,|\rho_\epsilon(r)|\,dr\,ds \\ &\leq \alpha(\dot{x}_0 + \|g\|_{\infty,I}\|\rho\|_1) + \dot{x}_0 \leq b + \dot{x}_0. \end{aligned} \tag{A4}$$

By a similar estimate one shows that A, in fact, is a contraction on X_ϵ. Hence by the fixed point theorem we have a unique solution $x_\epsilon \in X_\epsilon$. □

Since α is independent of ϵ we have for small ϵ: $-\epsilon + \alpha \geq \epsilon$. But at $t = \epsilon$ the right hand side of the differential equation (A1) has already been ''turned off.'' Hence for small ϵ the solutions are defined on the whole real line.

Furthermore, since for all ϵ the solution lies in X_ϵ, x_ϵ is bounded uniformly in ϵ for say $t \leq \alpha/2$ (for ϵ small enough). For larger values of t x_ϵ grows only linearly with $\dot{x}_\epsilon$ which is bounded by $\dot{x}_0 + \|g\|_{\infty,I}\|\rho\|_1$. Hence (for small ϵ) $x_\epsilon(t)$ is bounded uniformly in ϵ on every compact set.

[1] J. Ehlers and W. Kundt, ''Exact solutions of the gravitational field equations,'' in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1964), pp. 49–99.

[2] R. Penrose, ''The geometry of impulsive gravitational waves,'' in *General Relativity, Papers in Honour of J. L. Synge*, edited by L. O'Raifeartaigh (Clarendon, Oxford, 1972), pp. 101–115.

[3] P. C. Aichelburg and R. U. Sexl, ''On the gravitational field of a massless particle,'' J. Gen. Rel. Grav. **2**, 303–312 (1971).

[4] H. Balasin and H. Nachbagauer, ''The ultrarelativistic Kerr-geometry and its energy-momentum tensor,'' Class. Quantum Grav. **12**, 707–713 (1995).

[5] C. O. Loustó and N. Sánchez, ''The ultrarelativistic limit of the Kerr-Newman geometry and particle scattering at the Planck scale,'' Phys. Lett. B **232**, 462–466 (1989).

[6] R. Penrose, ''Structure of space-time,'' in *Battelle Recontres*, edited by C. M. DeWitt and J. A. Wheeler (Benjamin, New York, 1968), pp. 121–235.

[7] T. Dray and G. 't Hooft, ''The gravitational shockwave of a massless particle,'' Nucl. Phys. B **232**, 173–188 (1985).

[8] V. Ferrari, P. Pendenza, and G. Veneziano, ''Beam like gravitational waves and their geodesics,'' J. Gen. Rel. Grav. **20**, 1185–1191 (1988).

[9] H. Balasin, ''Geodesics for impulsive gravitational waves and the multiplication of distributions,'' Class. Quantum Grav. **14**, 455–462 (1997).

[10] M. Oberguggenberger, *Multiplication of Distributions and Applications to Partial Differential Equations* (Longman Scientific and Technical, New York, 1992).

[11] R. Steinbauer, ''The ultrarelativistic Reissner-Nordström field in the Colombeau algebra,'' J. Math. Phys. **38**, 1614–1622 (1997).

[12] O. Hájek, Bull. Am. Math. Soc. **12**, 272–279 (1985).

[13] J. F. Colombeau, *New Generalized Functions and Multiplication of Distributions* (North Holland, Amsterdam, 1984).

[14] J. F. Colombeau, *Multiplication of Distributions* (Springer, Berlin, 1992).

[15] R. Hermann and M. Oberguggenberger, ''Generalized functions, calculus of variations and nonlinear ODEs,'' preprint (1995).

[16] M. Kunzinger, ''Lie transformation groups in Colombeau algebras,'' Ph.D. thesis, University of Vienna, 1996.

The Einstein–Hilbert–Yang–Mills–Higgs action and the Dirac–Yukawa operator

Jürgen Tolksdorf[a)]
Humboldt-Universität zu Berlin Unter den Linden 6, D-10099 Berlin, Germany

(Received 9 September 1997; accepted for publication 23 September 1997)

In this article we show in some detail how the full action functional of the standard model of elementary particle physics can be described within the geometrical setting of generalized Dirac operators. We thereby introduce a new model building kit for (a certain class of) gauge invariant theories which provides a unified geometrical description of Einstein's theory of gravity and Yang–Mills gauge theories on the ''classical'' level. Moreover, when the gauge symmetry is spontaneously broken, the Higgs sector as well has a natural geometrical interpretation. It turns out that the Higgs field is related to the gravitational potential. Since the full action functional of the standard model is derived in one stroke, the appropriate parameters of the model have to satisfy certain relations similar to those in the Connes–Lott approach. Likewise, this may yield some phenomenological consequences, which is illustrated by using the gauge group of the standard model in the case of N generations of leptons and quarks. © *1998 American Institute of Physics.* [S0022-2488(98)02103-3]

I. INTRODUCTION

In this article we propose a certain model building kit which permits derivation of the action functional of the standard model of elementary particle physics with gravity included in terms of generalized Dirac operators. For this we introduce the following geometrical data:

$$(\mathrm{G},\rho,\mathrm{D}), \tag{1}$$

where G denotes a real compact semi-simple Lie group, ρ is its unitary representation on an N-dimensional Hermitian vector space V, and D denotes a Dirac operator acting on sections into a Clifford module bundle $\mathscr{E}:=\mathrm{S}\otimes\mathrm{E}$, such that $\mathrm{E}:=\mathscr{P}\times_\rho V$. Here, S denotes the spinor bundle over a closed compact orientable Riemannian spin manifold $(\mathscr{M},g)$ without boundary and of even dimension $(2n>2)$; $\mathscr{P}$ is a G-principal bundle over $\mathscr{M}$.

Having given the geometrical setting we propose the following functional:

$$\mathscr{T}_\mathrm{D}:=(\psi,i\mathrm{D}\psi)_{\Gamma(\mathscr{E})}+\mathrm{res}_\zeta(\mathrm{D}^{-2n+2}). \tag{2}$$

Here, res denotes the Wodzicki residue, ζ is an element of the commutant defined by (G,ρ), satisfying $[\mathrm{D},\zeta]=[\chi,\zeta]=0$, $\zeta>0$, with χ the involution operator on $\mathscr{E}$, and $(,)_{\Gamma(\mathscr{E})}$ denotes the induced Hermitian product on the $C^\infty(\mathscr{M})$ module $\Gamma(\mathscr{E})$ of sections ψ into $\mathscr{E}$. In our frame this functional serves as a ''general action'' functional.

Using this frame the main result of this paper may be summarized in the following.

Theorem: *There exists a natural generalization (1) of the Dirac–Yukawa operator of the standard model such that the functional (2) is proportional to the full action of the standard model with gravity included.*

As a consequence we obtain certain relations between the parameters involved. Especially, the mass of the Higgs field m_h is a function of the fermion masses. In the most general case the range for the electroweak angle θ_w reads

$$0.25\leqslant\sin^2\theta_w\leqslant 0.45.$$

[a)]Electronic mail: Tolkdorf@euler.math.uni-mannheim.de

 © 1998 American Institute of Physics

In the case where all irreducible subspaces of the fermion representation are equally weighted we obtain the GUT preferred relations

$$\sin^2\theta_w=\tfrac{3}{8},\quad g_{(3)}=g_{(2)},$$

where, respectively, $g_{(3)}$ and $g_{(2)}$ are the strong and weak coupling constants. Of course, all derived relations are expected to be scale dependent and thus the corresponding renormalization flux must be carefully taken into account. We stress that the corresponding hyperchanges of the particles involved in the model are fixed when the electrical charges are assumed to be known.

The model building kit as defined by the generalized Dirac operator (1) and the universal action (2) is motivated by the assumption that the basic objects in nature are fermions and that their dynamics are described by Dirac's equation. Then, the appropriate dynamics of the various fields involved in the definition of the fermionic interactions should be a consequence of the latter and not independent thereof. Mathematically, this may be rephrased as follows: when fermions are geometrically described by sections into a twisted spinor bundle, the most natural operator acting on those objects is a Dirac operator. A Dirac operator, however, is but a Clifford superconnection (i.e., a certain generalization of a connection) defined by (homomorphism-valued) differential forms of *various* degrees. Then, fixing the admissible fermionic interactions geometrically means fixing the admissible differential forms defining a certain superconnection and thus a certain Dirac operator. The idea of the kit, proposed here, is that the functionals leading to the field equations of the differential forms defining the Dirac operator are not arbitrary but determined by this operator. As it turns out, this idea not only permits a geometrical understanding of the Higgs action but also a new geometrical interpretation of the Einstein–Hilbert and Yang–Mills functional $(\mathscr{T}_{\mathrm{EH}},\mathscr{T}_{\mathrm{YM}})$. Indeed, from this point of view the former occurs—in a sense—as a natural ''companion'' of the latter and both are ''consequences'' of the fermionic interaction. In this (classical) description of particle physics the *a priori* assumption of a flat space–time seems artificial. To make this more clear, let us assume we consider a ''free'' fermion. From a geometrical viewpoint, such an object may be considered as a section into a twisted spinor bundle, where now the twisting part is assumed to possess a trivial connection. Nevertheless, the Clifford module bundle may be nonflat. Indeed, the fermion carries energy and thereby produces a nontrivial gravitational field. Hence, the corresponding spin connection is nontrivial in general. Of course, under ''normal'' conditions the energy of an elementary particle is so weak that the appropriate gravitational field cannot be measured. From our point of view this simply means that in the particular case at hand the energy-momentum tensor—defined by the Dirac action—is ignorable with respect to any inertial frame and, therefore, space–time becomes *approximately* flat. Though this is usually accepted by physicists, the point here is that the functional $\mathscr{T}_{\mathrm{D}}$ fixes the energy-momentum tensor. Consequently, in our approach a vanishing Einstein tensor *a priori* makes no sense. This can strictly hold iff there are no fermions in the world (whatever such a world looks like!). Next, let us consider the case when the fermion becomes massive. Then, the interaction with the Higgs field must be taken into account. As it turns out, geometrically, the Higgs field defines a certain connection on the Clifford module bundle where the fermions live. The corresponding curvature is nonzero, even in the case when the Higgs field represents a (classical) nontrivial vaccum. Of course, since the energy-momentum of the Higgs field is nonzero, space–time must be curved as well. Note that this holds true even in the case when the world ''sits in the (classical) vaccum.'' We stress that in our scheme the Higgs field seems intimately related to gravity. Of course, we are only considering classical field theory and one may object that on the level of elementary particles quantum theory has to be taken into account and then gravity may look completely different from that described by Einstein's equation. However, as a ''first step'' towards a real understanding of the interplay between gravity and particle physics it might be useful to have a unified geometrical description of all interactions on the level of the classical field equations known so far.

Mathematically, it is evident that the action (2) is gauge invariant. Hence, our kit provides a general scheme for building (a certain class of) gauge invariant theories. It therefore might be worth remembering the input of a general Yang–Mills model and to compare both building kits, correspondingly. Here, we adopt the notation as given in Ref. 1.

The input of a *general Yang–Mills model* consists of the following data:

(1) a finite-dimensional real, compact Lie group G,

(2) a unitary representation $\rho_f=\rho_{\mathrm{L}}\oplus\rho_{\mathrm{R}}$ on the $\mathbb{Z}_2$-graded Hermitian vector space $V_f=V_{\mathrm{L}}\oplus V_{\mathrm{R}}$ of the left- and right-handed fermions,

(3) a finite set of positive constants $\{g_{(k)}\}$ (the gauge coupling constants), parametrizing the general Killing form of the Lie algebra $\mathscr{G}$ of G [the number of these constants is defined as the number of simple components of $\mathscr{G}$, including u(1) factors],

(4) a unitary representation ρ_h on a Hermitian vector space V_h of the Higgs field,

(5) a G-invariant polynomial (Higgs potential) $V_h\overset{V}{\to}\mathbb{R}$ of order four, which is bounded from below,

(6) one complex constant (the Yukawa coupling constant) g_y for every one-dimensional invariant subspace in the decomposition of the representation

$$(V_L^*\otimes V_R\otimes V_h)\oplus(V_L^*\otimes V_R\otimes V_h^*),$$

(7) and an action functional

$$\mathscr{I}:=\mathscr{I}_{\mathrm{EH}}+\mathscr{I}_{\mathrm{Dirac}}+\mathscr{I}_{\mathrm{Yukawa}}+\mathscr{I}_{\mathrm{YM}}+\mathscr{I}_{\mathrm{Higgs}}.$$

Usually, it is assumed that space–time is flat, so that $\mathscr{I}_{\mathrm{EH}}$ is ignored. In particular, the *standard model* is defined by

$$G:=\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1) \tag{3}$$

with three coupling constants $(g_{(3)},g_{(2)},g_{(1)})$;

$$\mathrm{V_L}:=\bigoplus_1^3[(1,2,-1/2)\oplus(3,2,1/6)], \tag{4}$$

$$\mathrm{V_R}:=\bigoplus_1^3[(1,1,-1)\oplus(3,1,-1/3)\oplus(3,1,2/3)], \tag{5}$$

where (n_3,n_2,n_1) denote the tensor product, respectively, of an n_3-dimensional representation of SU(3), an n_2-dimensional representation of SU(2), and a one-dimensional representation of U(1) with ''hypercharge'' $y:\rho(e^{i\theta}):=e^{iy\theta}$, $y\in\mathbb{Q}$, $\theta\in[0,2\pi[$;

$$V:=\lambda(\varphi\varphi^*)^2-\mu^2\varphi\varphi^*, \tag{6}$$

with $\lambda,\mu>0$. There are 27 Yukawa coupling constants which, however, are not all independent. In fact, the standard model can be parametrized by 18 constants (cf. Ref. 2).

The full action functional (i.e., with gravity included) in the usual description of the standard model ($N=3$) reads

$$\mathscr{I}:=\frac{1}{16\pi G}\int_{\mathscr{M}}{}^*r_{\mathscr{M}} \tag{7}$$

$$+\int_{\mathscr{M}}{}^*(\psi^{(l)*}i\mathrm{D}_{(l)}\psi^{(l)}+\psi^{(q)*}i\mathrm{D}_{(q)}\psi^{(q)}) \tag{8}$$

$$-\int_{\mathscr{M}}{}^*\Bigg(\sum_{i,j=1}^{N}(\mathbf{g}_y^{(l)})_{ij}\overline{\psi}_{\mathrm{L}_i}^{(l)}(\gamma_5\varphi)\psi_{\mathrm{R}_j}^{(l)}+(\mathbf{g}_y^{(q)})_{ij}\overline{\psi}_{\mathrm{L}_i}^{(u)}$$

$$\times(\gamma_5\varphi)\psi_{\mathrm{R}_j}^{(u)}+(\mathbf{g}_y'^{(q)})_{ij}\overline{\psi}_{\mathrm{L}_i}^{(d')}(\gamma_5\overline{\boldsymbol{\epsilon}\varphi})\psi_{\mathrm{R}_j}^{(d')}\Bigg)+\text{comp. conj.} \tag{9}$$

$$+\frac{1}{g_{(3)}^2}\int_{\mathscr{M}}\mathrm{tr}(\mathbf{C}\wedge{}^*\mathbf{C})+\frac{1}{g_{(2)}^2}\int_{\mathscr{M}}\mathrm{tr}(\mathbf{W}\wedge{}^*\mathbf{W})+\frac{1}{2g_{(1)}^2}\int_{\mathscr{M}}B\wedge{}^*B \tag{10}$$

$$+\int_{\mathscr{M}}(\nabla\varphi)^{*}\wedge *(\nabla\varphi)+\int_{\mathscr{M}} *V \tag{11}$$

$$\equiv \mathscr{T}_{\mathrm{EH}}+\mathscr{T}_{\mathrm{Dirac}}+\mathscr{T}_{\mathrm{Yukawa}}+\mathscr{T}_{\mathrm{YM}}+\mathscr{T}_{\mathrm{Higgs}}. \tag{12}$$

The traces in the definition of the Yang–Mills action (10) are taken with respect to the corresponding fundamental representations of SU(3) and SU(2). Note that $\mathscr{M}$ denotes a Riemannian manifold, which explains the occurence of the apparently wrong relative sign in front of the Higgs potential and the occurence of γ_5 in the Yukawa coupling term (9).

In contrast to a general Yang–Mills model our proposed kit (1) and (2) has the following input:

(1) a finite-dimensional real, compact Lie group G,
(2) a unitary representation $\rho=\rho_{\mathrm{L}}\oplus\rho_{\mathrm{R}}$ on the $\mathbb{Z}_2$-graded Hermitian vector space $\mathrm{V}=\mathrm{V}_{\mathrm{L}}\oplus\mathrm{V}_{\mathrm{R}}$ of the left- and right-handed fermions,
(3) a Dirac operator D,
(4) the general functional: $\mathscr{T}_{\mathrm{D}}$.

As in the ''noncommutative approach'' introduced by Connes and Lott (cf. Refs. 3 and 4), the Higgs representation is not arbitrary but has to lie within the fermionic representation, symbolically: $\rho_h\subset\rho_f$. In fact, this has significant consequences with respect to the relations between the various parameters involved in the model. Hence it might not come as a surprise that there are certain similarities between the Connes–Lott model and the model introduced here in this respect (see below). Of course, the mathematical background is quite different. We mention that with respect to our physical interpretation it is quite natural that all fields involved in the model carry the same representation.

Concerning the *standard model*, the Dirac operator $\mathrm{D}\equiv\widetilde{D}_{\phi}$ is defined by a generalization of the *Dirac–Yukawa operator* D_{ϕ}. It is well known that the Yukawa coupling (9) together with the (standard) Dirac operator D defining (8) can be considered as a new Dirac operator D_{ϕ}—the Dirac–Yukawa operator. The main feature, then, is that this Dirac operator is a ''nonstandard'' Dirac operator (i.e., not associated with a Clifford connection, see below). In fact, such Dirac operators will play a key role in our geometrical description of the standard model. Correspondingly, we shall discuss those operators in some detail in the first part of our paper, which is totally concerned with the mathematical frame of our model. Though the larger portion of Part 1 is actually not new and may be found in much more detail, e.g., in Ref. 5, we nevertheless summarize the basic mathematical notions in order for our paper to be self-contained and to permit full understanding of the issue also for those who are not familiar with the notion of nonstandard Dirac operators. Also, we have emphasized the relations between nonstandard Dirac operators and connections on a (general) Clifford module bundle. This is of technical significance and, moreover, explains how the Higgs field yields a certain connection on a Clifford module bundle. In Part 2 we introduce a certain generalization of the Dirac–Yukawa operator and prove our main theorem. Moreover, we also investigate the ''phenomenological'' consequences of our scheme with respect to the standard model. Finally, we mention some similarities to the Connes–Lott model (cf. Refs. 3, 4, and, concerning the new approach, 6–8). We conclude this paper with an outlook.

Before we start to describe the mathematical frame of our model building kit, however, some remarks concerning its similarity to the Connes–Lott approach to the standard model seem appropriate. Obviously, our notion of a generalized Dirac operator, as defined by (1), is similar to Connes' notion of a ''spectral triple.'' Needless to say that the latter notion is more profound, mathematically, since it offers the possibility of ''new mathematics,'' like Connes' noncommutative geometry (cf. Ref. 4). Also, the idea to of *derive* specific functionals, such as in the case of the standard model of particle physics, from a ''universal functional'' must go back to Connes. For example, in the Connes–Lott approach to the standard model the Dixmier trace serves as the general action functional (cf. Refs. 3 and 4). In a sense this trace can be considered as a special case of the more general Wodzicki residue. Unfortunately, using the Dixmier trace as the universal action functional it seems hard to derive both the Einstein–Hilbert and the Yang–Mills (–Higgs)

action in one stroke (for the pure EH functional see Ref. 9). In fact, in the Connes–Lott description of the standard model the geometric information contained in the Dirac operator is lost. But as Connes has remarked, the Wodzicki residue of D^{-2n+2} with D a Standard Dirac operator (see below) becomes proportional to the Einstein–Hilbert action of gravity (cf. Ref. 10). However, this time the geometrical information contained in the Yang–Mills potential is lost. This follows immediately from D being a Standard Dirac operator (see below). As a natural question one may ask whether it is possible to derive the full action functional (EHYMH action) by considering ''nonstandard'' Dirac operators. In Ref. 11 this has been investigated and affirmatively answered in the case of the Einstein–Hilbert–Yang-Mills functional. In the case of the full action of the standard model and with gravity included the above question was investigated in Ref. 12. However, in this work there is a mistake. In fact, the definition of the Dirac–Yukawa operator is wrong and, as a consequence, the derived functional does not coincide with the action functional of the standard model. As we shall show in the paper at hand, however, by using the right definition of the Dirac–Yukawa operator the properly corrected generalized Dirac–Yukawa operator proposed in Ref. 12 is but a special case of the generalized Dirac–Yukawa operator introduced in part two of our paper and which in fact gives rise to the full action of the standard model. Though the basic idea of our kit is already introduced in Ref. 12, and which indeed affirmatively answers the above-mentioned question, the scheme in Ref. 12, however, is still not general enough to discuss physical implications of the proposed model building kit. This is because in Ref. 12 neither the considered Dirac operator—even if properly corrected—nor the proposed universal action functional is general enough. Both have been remedied in this article.

II. PART 1: THE MATHEMATICAL FRAME

The geometrical setting which we propose in order to describe gauge theories is that of a Clifford module bundle $(\mathscr{E},c)$ over a Riemannian manifold $(\mathscr{M},g)$ of even dimension. Within this setting there exists a distinguished class of operators called generalized Dirac operators. [In the following, the term ''generalized'' simply means ''nonstandard,'' though by a generalized Dirac operator we actually mean the triple (G, ρ, D), with D a (nonstandard) Dirac operator.] We therefore start with a brief review on the notion of Clifford modules and generalized Dirac operators. More details of this issue can be found, e.g., in Ref. 5. Afterwards we shall discuss at some length how a given generalized Dirac operator (1) determines a particular functional $\mathscr{T}_D(\nabla^{\mathscr{E}},\psi)$, using (2).

To get started, let us denote by $(\mathscr{M},g)$ a smooth, closed compact *Riemannian* (spin) manifold without boundary and of even dimension: $\dim(\mathscr{M})\equiv m:=2n(>2)$. Moreover, let $\mathscr{E}:=\mathscr{E}^+\oplus\mathscr{E}^-$ be (the total space of) a $\mathbb{Z}_2$-graded Hermitian vector bundle $\mathscr{E}\xrightarrow{\pi}\mathscr{M}$ over $\mathscr{M}$. The corresponding Hermitian product on $\mathscr{E}$ is indicated by $(,)_{\mathscr{E}}$. If $\Gamma(\mathscr{E})$ denotes the $C^\infty(\mathscr{M})$ module of smooth sections into $\mathscr{E}$, then the induced Hermitian product $(,)_{\Gamma(\mathscr{E})}$ on $\Gamma(\mathscr{E})$ is given by $(,)_{\Gamma(\mathscr{E})}:=\int_{\mathscr{M}}*(,)_{\mathscr{E}}$. Here, ''$*$'' means the Hodge map regarding the Riemannian metric g on $\mathscr{M}$.

Definition 1: A *generalized Dirac operator* D is any odd first-order differential operator acting on sections $\psi\in\Gamma(\mathscr{E})$:

$$D:\Gamma(\mathscr{E}^\pm)\to\Gamma(\mathscr{E}^\mp), \tag{13}$$

so that D^2 is a *generalized Laplacian*, i.e., there exists a connection: $\Gamma(\mathscr{E})\xrightarrow{\hat\nabla^{\mathscr{E}}}\Gamma(T^*\mathscr{M}\otimes\mathscr{E})$ on the vector bundle $\mathscr{E}$ and an endomorphism $\mathscr{F}\in\Gamma(\mathrm{End}(\mathscr{E}))$, both uniquely defined by D, such that

$$D^2=\Delta^{\hat\nabla^{\mathscr{E}}}+\mathscr{F}. \tag{14}$$

Here, $\Delta^{\hat\nabla^{\mathscr{E}}}:=-ev_g(\hat\nabla^{T^*\mathscr{M}\otimes\mathscr{E}}\hat\nabla^{\mathscr{E}})$ denotes the horizontal (Bochner) Laplacian associated with the connection $\hat\nabla^{\mathscr{E}}$, and ''ev_g'' means the evaluation map regarding the metric g.

Remark 1: D exists on $\mathscr{E}$ iff $\mathscr{E}$ denotes a *Clifford module* bundle over $(\mathscr{M},g)$, i.e., there exists a graded (left) action on $\mathscr{E}$.

$$c:\mathscr{C}(\mathscr{M})\times\mathscr{E}\to\mathscr{E}. \tag{15}$$

of the Clifford bundle $\mathscr{C}(\mathscr{M})\overset{\tau}{\to}\mathscr{M}$ associated with the metric g. [Here, the fiber $\tau^{-1}(x)$ is isomorphic to the Clifford algebra generated by the elements $u,v\in T_x^*\mathscr{M}$, using the relation: $uv+vu:=-2g_x(u,v)$, $\forall x\in\mathscr{M}$.] This holds because in the case that D denotes a generalized Dirac operator, this operator induces via

$$\mathscr{C}(\mathscr{M})\times\mathscr{E}\to\mathscr{E}(df,\psi)\mapsto c(df)\psi:=[\mathrm{D},f]\psi,\quad f\in C^\infty(\mathscr{M}) \tag{16}$$

a graded left action of the Clifford bundle on $\mathscr{E}$ (cf. Ref. 5). Conversely, if $(\mathscr{E},c)$ denotes a Clifford module bundle over $(\mathscr{M},g)$, then the well-known construction

$$\Gamma(\mathscr{E})\overset{\nabla^{\mathscr{E}}}{\to}\Gamma(T^*\mathscr{M}\otimes\mathscr{E})\hookrightarrow\Gamma(\mathscr{C}(\mathscr{M})\otimes\mathscr{E})\overset{c}{\to}\Gamma(\mathscr{E}) \tag{17}$$

defines an operator $D_{\nabla^{\mathscr{E}}}:=c(\nabla^{\mathscr{E}})$ satisfying (13) and (14) for any connection $\nabla^{\mathscr{E}}$ on $\mathscr{E}$. Moreover, it is easily checked that such a defined operator also fulfills (16).

Hence, from now on $\mathscr{E}\equiv(\mathscr{E},c)$ will always denote a Clifford module bundle over $\mathscr{M}\equiv(\mathscr{M},g)$. As a consequence, the endomorphism bundle End $(\mathscr{E})$ is also a Clifford module and it follows that (cf. Ref. 5)

$$\mathrm{End}(\mathscr{E})\simeq\mathscr{C}(\mathscr{M})\otimes\mathrm{End}_{\mathrm{Cl}}(\mathscr{E}), \tag{18}$$

where $\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})$ denotes the algebra bundle of bundle endomorphisms of $\mathscr{E}$ supercommuting with the action of $\mathscr{C}(\mathscr{M})$, i.e.,

$$\mathrm{End}_{\mathrm{Cl}}(\mathscr{E}):=\{\sigma\in\mathrm{End}(\mathscr{E})\,|\,[\![c(a),\sigma]\!]=0,\forall a\in\mathscr{C}(\mathscr{M})\}. \tag{19}$$

If in addition $\mathscr{M}$ is assumed to be a spin manifold, we have $\mathscr{C}(\mathscr{M})_{\mathrm{C}}\simeq\mathrm{End}(\mathrm{S})$ and correspondingly $\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})\simeq\mathrm{End}(\mathrm{E})$, where S denotes the *spinor* bundle and E a vector bundle over $\mathscr{M}$. In this case $\mathscr{E}$ is called a *twisted spinor bundle* over $\mathscr{M}$ and we have (cf. Ref. 5)

$$\mathscr{E}\simeq\mathrm{S}\otimes\mathrm{E}. \tag{20}$$

We say that the Clifford module bundle $(\mathscr{E},c)$ has a *twisting graduation* if $\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})$ possesses a (nontrivial) $\mathbb{Z}_2$ graduation (cf. Ref. 11). Clearly, in the case of (20) this is equivalent to saying that the vector bundle $\mathrm{E}=\underset{\sim}{\mathrm{E}}_{\mathrm{L}}\oplus\mathrm{E}_{\mathrm{R}}$ is $\mathbb{Z}_2$ graded, as well.

Definition 2: Let $\widetilde{\mathrm{D}}$ be a generalized Dirac operator on $\mathscr{E}$. It is said to be compatible with the Clifford action c if it satisfies the relation (16). Then, let

$$\mathscr{D}(\mathscr{E}):=\{\widetilde{\mathrm{D}}\,|\,[\widetilde{\mathrm{D}},f]=c(df\ ),\forall f\in C^\infty(\mathscr{M})\} \tag{21}$$

be the set of all generalized Dirac operators on $\mathscr{E}$ which are compatible with the Clifford action c. We have

$$\mathscr{D}(\mathscr{E})\simeq\Omega^0(\mathscr{M},\mathrm{End}^-(\mathscr{E})). \tag{22}$$

[Since the bundle $\mathscr{E}=\mathscr{E}^+\oplus\mathscr{E}^-$ is $\mathbb{Z}_2$ graded, so is the associated endomorphism bundle $\mathrm{End}(\mathscr{E})$, i.e., $\mathrm{End}(\mathscr{E})=\mathrm{End}^+(\mathscr{E})\oplus\mathrm{End}^-(\mathscr{E})$, with $\mathrm{End}^+(\mathscr{E}):=\mathrm{End}(\mathscr{E}^+)\oplus\mathrm{End}(\mathscr{E}^-)$ and $\mathrm{End}^-(\mathscr{E}):=Hom(\mathscr{E}^+,\mathscr{E}^-)\oplus Hom(\mathscr{E}^-,\mathscr{E}^+)$.] Also, let us denote by

$$\mathscr{A}(\mathscr{E}):=\{\nabla^{\mathscr{E}}\,|\,\Gamma(\mathscr{E})\overset{\nabla^{\mathscr{E}}}{\to}\Gamma(T^*\mathscr{M}\otimes\mathscr{E})\} \tag{23}$$

the set of all (even) connections on $\mathscr{E}$.

As a consequence of (18), there exists a natural class of connections—called *Clifford connections*—on any Clifford module bundle $\mathscr{E}$:

$$\mathscr{A}_{\mathrm{Cl}}(\mathscr{E}):=\{\nabla^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})\,|\,[\nabla^{\mathscr{E}},c(a)]=c(\nabla^{\mathrm{Cl}}a),\forall a\in\Gamma(\mathscr{C}(\mathscr{M}))\}\subset\mathscr{A}(\mathscr{E}), \tag{24}$$

where ∇^{Cl} is the induced Levi–Cevita connection on $\mathscr{C}(\mathscr{M})$. Note, in the case of a twisted spinor bundle (20) any Clifford connection $\nabla^{\mathscr{E}} \in \mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ takes the form of a tensor product connection (cf. Ref. 5)

$$\nabla^{\mathscr{E}} \equiv \nabla^{\mathrm{S}\otimes\mathrm{E}} := \nabla^{\mathrm{S}} \otimes \mathbf{1}_E + \mathbf{1}_S \otimes \nabla^{\mathrm{E}}, \tag{25}$$

where, respectively, ∇^{S} denotes the spin connection on S and ∇^{E} any connection on the vector bundle E. In general we have

$$\begin{aligned} \mathscr{A}(\mathscr{E}) &\simeq \Omega^1(\mathscr{M}, \mathrm{End}^+(\mathscr{E})), \\ \mathscr{A}_{\mathrm{Cl}}(\mathscr{E}) &\simeq \Omega^1(\mathscr{M}, \mathrm{End}^+_{\mathrm{Cl}}(\mathscr{E})), \end{aligned} \tag{26}$$

where the latter isomorphism follows from (18).

Remark 2: Using the linear isomorphism

$$\begin{aligned} \Lambda T^*\mathscr{M} &\xrightarrow{\mathbf{c}} \mathscr{C}(\mathscr{M}), \\ e^{i_1} \wedge e^{i_2} \cdots \wedge e^{i_k} &\mapsto e^{i_1} e^{i_2} \cdots e^{i_k}, \quad \forall\, 0 \leqslant k \leqslant m, \end{aligned} \tag{27}$$

between the Grassmann and the Clifford bundle, where $\{e^i\}_{1\leqslant i\leqslant m}$ denotes a local orthonormal basis in $T^*\mathscr{M}$ and Clifford multiplication is indicated by juxtaposition, the Clifford action c induces a linear mapping, also denoted by c,

$$\begin{aligned} \Omega^p(\mathscr{M}, \mathrm{End}^{\pm}(\mathscr{E})) &\xrightarrow{c} \Omega^0(\mathscr{M}, \mathrm{End}^{\mp}(\mathscr{E})), \\ \alpha &\mapsto c(\alpha). \end{aligned} \tag{28}$$

Lemma 1: In the case of $p=1$ *the linear mapping (28) has a canonical right inverse defined by*

$$\begin{aligned} \delta_\xi : \Omega^p(\mathscr{M}, \mathrm{End}^{\mp}(\mathscr{E})) &\to \Omega^{p+1}(\mathscr{M}, \mathrm{End}^{\pm}(\mathscr{E})), \\ \alpha &\mapsto \xi \wedge \alpha, \end{aligned} \tag{29}$$

where $\xi \in \Omega^1(\mathscr{M}, \mathrm{End}^-(\mathscr{E}))$ *is locally given by*

$$\xi := -\frac{1}{m}\, g(e_a, e_b) e^a \otimes c(e^b) \otimes \mathbf{1}_{\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})}. \tag{30}$$

(Throughout this paper we adopt Einstein's convention for summation.) The product in (29) simply means $(\sigma\otimes a)\wedge(\sigma'\otimes a') := \sigma\wedge\sigma'\otimes aa'$ for all homogeneous elements $\sigma\otimes a, \sigma'\otimes a' \in \Omega^*(\mathscr{M}, \mathrm{End}(\mathscr{E})) \simeq \Gamma(\Lambda T^*\mathscr{M} \otimes \mathrm{End}(\mathscr{E}))$. Again, $\{e^a\}_{1\leqslant a\leqslant m}$ is a basis in $T^*\mathscr{M}$ and $\{e_a\}_{1\leqslant a\leqslant m}$ is its dual. Note that this ''wedge product'' is not graded commutative and that $\Omega^*(\mathscr{M}, \mathrm{End}(\mathscr{E}))$ is a bigraded algebra. The map (29) is even with respect to the total grading.

Proof: Obviously, this follows by construction.

Remark 3: The form ξ, locally defined by (30), can be characterized via

$$\begin{aligned} \nabla^{T^*\mathscr{M}\otimes\mathrm{End}(\mathscr{E})}\xi &\equiv 0, \quad \forall\, \nabla^{\mathscr{E}} \in \mathscr{A}_{\mathrm{Cl}}(\mathscr{E}), \\ c(\xi) &= \mathbf{1}_{\mathscr{E}}. \end{aligned} \tag{31}$$

Lemma 2: Let c *be the linear mapping (28) restricted to* $\Omega^1(\mathscr{M}, \mathrm{End}^+(\mathscr{E}))$. *Then we have*

$$\mathscr{D}(\mathscr{E}) \simeq \mathscr{A}(\mathscr{E})/\ker(c). \tag{32}$$

Proof: For $p := \delta_\xi c : \Omega^1(\mathscr{M}, \mathrm{End}^+(\mathscr{E})) \to \Omega^1(\mathscr{M}, \mathrm{End}^+(\mathscr{E}))$ we get $p^2 = p$ and thus

$$\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E}))=\mathrm{im}(p)\oplus\mathrm{im}(q), \tag{33}$$

with $q:=\mathbf{1}-p$. Restricting the linear mapping (28) to $\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E}))$ yields the identities

$$pc\equiv p,\quad cp\equiv c. \tag{34}$$

Since δ_ξ is a right inverse of c we have $\ker(c)\subset\mathrm{im}(q)$. Moreover, for all $\alpha\in\mathrm{im}(p)$ with $c(\alpha)=0$ (34) implies $\alpha\equiv 0$. Hence

$$\ker(c)=\mathrm{im}(q). \tag{35}$$

The statement follows from

$$\mathscr{A}(\mathscr{E})\simeq\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E}))\overset{c}{\to}\Omega^0(\mathscr{M},\mathrm{End}^-(\mathscr{E}))\simeq\mathscr{D}(\mathscr{E}) \tag{36}$$

and we are done. Note that actually we have shown that the sequence

$$0\to\ker(c)\to\Omega^1(\mathscr{M},\mathrm{End}^+\ \mathscr{E})\overset{c}{\to}\Omega^0(\mathscr{M},\mathrm{End}^-\ \mathscr{E})\to 0$$

is exact and splits. Also, in this case the kernel of the mapping (28) becomes explicit.

Definition 3: Two connections $\nabla^{\mathscr{E}},\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})$ are defined to be equivalent iff

$$\nabla^{\mathscr{E}}-\widetilde{\nabla}^{\mathscr{E}}\in\ker(c). \tag{37}$$

By the preceding Lemma this is equivalent to

$$\nabla^{\mathscr{E}}\sim\widetilde{\nabla}^{\mathscr{E}}\Leftrightarrow\widetilde{\nabla}^{\mathscr{E}}=\nabla^{\mathscr{E}}+\omega,$$

$$\omega\in\mathrm{im}(q)\subset\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E})). \tag{38}$$

Therefore, any Dirac operator $\mathrm{D}\in\mathscr{D}(\mathscr{E})$ is uniquely associated with an equivalence class of connections $[\nabla^{\mathscr{E}}]$ on $\mathscr{E}$ so that $\mathrm{D}_{\nabla^{\mathscr{E}}}=\mathrm{D}$.

Remark 4: Let $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ be a given Dirac operator on $\mathscr{E}$. Then,

$$\widetilde{\nabla}^{\mathscr{E}}:=\nabla^{\mathscr{E}}+\delta_\xi(\widetilde{\mathrm{D}}-\mathrm{D}_{\nabla^{\mathscr{E}}}) \tag{39}$$

defines a connection on $\mathscr{E}$, so that

$$\mathrm{D}_{\widetilde{\nabla}^{\mathscr{E}}}=\widetilde{\mathrm{D}}, \tag{40}$$

where $\nabla^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})$ denotes any connection on $\mathscr{E}$. Clearly, this ambiguity simply reflects that $\mathscr{D}(\mathscr{E})$ is an affine space and thus a given Dirac operator $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ may be decomposed in infinitely many ways like

$$\widetilde{\mathrm{D}}=\mathrm{D}+\Phi_{\mathrm{D}} \tag{41}$$

with $\Phi_{\mathrm{D}}:=\widetilde{\mathrm{D}}-\mathrm{D}\in\Omega^0(\mathscr{M},\mathrm{End}^-(\mathscr{E}))$.

Definition 4: We call a Dirac operator $\mathrm{D}\in\mathscr{D}(\mathscr{E})$ a *standard Dirac operator* (SDO) if there is a Clifford connection $\nabla^{\mathscr{E}}\in\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ so that

$$\mathrm{D}=\mathrm{D}_{\nabla^{\mathscr{E}}}. \tag{42}$$

Lemma 3: Let $\nabla^{\mathscr{E}},\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ be Clifford connections on the Clifford module bundle $\mathscr{E}$ with $\nabla^{\mathscr{E}}\sim\widetilde{\nabla}^{\mathscr{E}}$. Then we have $\nabla^{\mathscr{E}}\equiv\widetilde{\nabla}^{\mathscr{E}}$.

Proof: Using (38), $\nabla^{\mathscr{E}},\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ implies that there exists a $\omega:=\nabla^{\mathscr{E}}-\widetilde{\nabla}^{\mathscr{E}}=\mathbf{1}_{\mathrm{Cl}}\otimes A$ with $A\in\Omega^1(\mathscr{M},\mathrm{End}^+_{\mathrm{Cl}}(\mathscr{E}))$. By assumption, we have

$$c(\omega)=c(e^\mu)\otimes A_\mu=0, \tag{43}$$

where $\{e^\mu\}_{1\leqslant\mu\leqslant m}$ is a local orthonormal frame. Hence, $A_\mu=0$, $\forall\,\mu=1,...,m$, which proves the lemma.

We therefore have shown that the class of connections defining a SDO on $\mathscr{E}$ admits a canonical representative. In what follows we shall always denote by $\mathrm{D}\in\mathscr{D}(\mathscr{E})$ a SDO and by $\nabla^{\mathscr{E}}\in\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ the appropriate Clifford connection, so that $\mathrm{D}=\mathrm{D}_\nabla$. In contrast, by $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ and $\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})$, respectively, we denote an arbitrary Dirac operator and connection on $\mathscr{E}$.

Let $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ be an arbitrary Dirac operator on the Clifford module bundle $\mathscr{E}$ and let

$$\widetilde{\nabla}^{\mathscr{E}}:=\nabla^{\mathscr{E}}+\delta_\xi(\widetilde{\mathrm{D}}-\mathrm{D}). \tag{44}$$

Then, in Ref. 11 it is shown that (see also Ref. 13)

$$\widetilde{\mathrm{D}}^2=\Delta^{\hat{\nabla}^{\mathscr{E}}}+\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}. \tag{45}$$

Here, respectively, the connection $\hat{\nabla}\in\mathscr{A}(\mathscr{E})$ and the endomorphism $\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}\in\Gamma(\mathrm{End}(\mathscr{E}))$ are defined by

$$\hat{\nabla}^{\mathscr{E}}:=\widetilde{\nabla}^{\mathscr{E}}+\omega_{\widetilde{\nabla}^{\mathscr{E}}} \tag{46}$$

and

$$\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}:=c(\widetilde{\nabla}^{\mathscr{E}\,2})+\mathrm{ev}_g(\widetilde{\nabla}^{T^*\mathscr{M}\otimes\mathrm{End}(\mathscr{E})}\omega_{\widetilde{\nabla}^{\mathscr{E}}}+\omega^2_{\widetilde{\nabla}^{\mathscr{E}}}), \tag{47}$$

where the one-form $\omega_{\widetilde{\nabla}^{\mathscr{E}}}\in\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E}))$ is locally given by

$$\omega_{\widetilde{\nabla}^{\mathscr{E}}}:=-\tfrac{1}{2}g(e_\mu,e_\nu)e^\mu\otimes c(e^\lambda)([\widetilde{\nabla}^{\lambda}_{\mathscr{E}},c(e^\nu)]+\Gamma^{\nu}_{\sigma\lambda}c(e^\sigma)). \tag{48}$$

The Γ's denote the Christoffel symbols defined by the metric g.

Lemma 4: The endomorphism $\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}$ defined in (47) is independent of the representative $\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})$ of the class of connections defining the Dirac operator $\widetilde{\mathrm{D}}$.

Proof: To prove this lemma we introduce the affine mapping

$$\overline{\omega}:\mathscr{A}(\mathscr{E})\to\mathscr{A}(\mathscr{E}),\quad \widetilde{\nabla}^{\mathscr{E}}\mapsto\widetilde{\nabla}^{\mathscr{E}}+\omega_{\widetilde{\nabla}^{\mathscr{E}}} \tag{49}$$

on $\mathscr{A}(\mathscr{E})$ and show that this mapping is well defined on $\mathscr{A}(\mathscr{E})/\ker(c)$. Let $[\widetilde{\nabla}^{\mathscr{E}}]$ be the equivalence class of connections defining the given Dirac operator $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ and denote by $\widetilde{\nabla}^{\mathscr{E}},\widetilde{\nabla}'^{\mathscr{E}}\in[\widetilde{\nabla}^{\mathscr{E}}]$ two representatives of this class. Hence, $\alpha:=\widetilde{\nabla}^{\mathscr{E}}-\widetilde{\nabla}'^{\mathscr{E}}\in\ker(c)$ and with respect to a local orthonormal frame we obtain

$$\hat{\nabla}^{\mathscr{E}}-\hat{\nabla}'^{\mathscr{E}}=\alpha-\tfrac{1}{2}\delta_{\mu\nu}e^\mu\otimes c(e^\lambda)[i_\lambda\alpha,c(e^\nu)] \tag{50}$$

$$=\tfrac{1}{2}\delta_{\mu\nu}e^\mu\otimes[c(\alpha),c(e^\nu)]_+ \tag{51}$$

$$=0, \tag{52}$$

where i_μ is the inner derivative with respect to the local vector field e_μ and $[,]_+$ means the anticommutator. Consequently, the map: $\widetilde{\mathrm{D}}\mapsto\Delta^{\hat{\nabla}^{\mathscr{E}}}$ with $\hat{\nabla}^{\mathscr{E}}$ given by (46), is well defined. Hence the endomorphism $\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}=\widetilde{\mathrm{D}}^2-\Delta^{\hat{\nabla}^{\mathscr{E}}}$ only depends on the Dirac operator $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ which proves the lemma.

Corollary 1: The endomorphism $\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}\in\Gamma(\mathrm{End}(\mathscr{E}))$ does not depend on the decomposition (44). In particular, it does not depend on the chosen Clifford connection $\nabla^{\mathscr{E}}$.

Proof: By Lemma 4 the proof is obvious.

Remark 5: The corresponding linear part of the affine map (49) has a nontrivial kernel; especially

$$\overline{\omega}|_{\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})}=\mathbf{1}_{\mathscr{A}_{\mathrm{Cl}}(\mathscr{E})}. \tag{53}$$

In this case the decomposition of the square of the appropriate SDO is but the usual *Lichnerowicz formula* and the endomorphism $\mathscr{F}$ takes its well-known form

$$\mathscr{F}^{\nabla^{\mathscr{E}}} = c[(\nabla^{\mathscr{E}})^2] \tag{54}$$

$$= \tfrac{1}{4} r_{\mathscr{M}}\, \mathbf{1}_{\mathscr{E}} + c(F^{\mathscr{E}/S}), \tag{55}$$

where $r_{\mathscr{M}}$ is the Ricci scalar curvature on the base manifold $\mathscr{M}$ and $F^{\mathscr{E}/S} := (\nabla^{\mathscr{E}})^2 - (\nabla^{\mathrm{Cl}})^2 \otimes 1_{\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})}$ denotes the relative curvature on the Clifford module bundle $\mathscr{E}$. Since in this particular case the relative curvature only depends on the connection on the twisting part of the Clifford module bundle $\mathscr{E}$, $F^{\mathscr{E}/S}$ is also called the *twisting curvature*.

We now turn to the notion of superconnections which can be considered as a generalization of connections on a $\mathbb{Z}_2$-graded vector bundle. As is well known, superconnections permit the generalization of the one-to-one correspondence between SDO and Clifford connections to arbitrary Dirac operators and Clifford superconnections on a Clifford module bundle. Hence, there is no essential difference between talking about Dirac operators and Clifford superconnections. We therefore call into mind the following.

Definition 5: A *superconnection* on a $\mathbb{Z}_2$-graded vector bundle $\mathscr{E}$ is any odd first-order differential operator ($\pm$ is understood with respect to the *total* grading)

$$\boldsymbol{\nabla}^{\mathscr{E}} : [\Omega^*(\mathscr{M}, \mathscr{E})]^{\pm} \to [\Omega^*(\mathscr{M}, \mathscr{E})]^{\mp}, \tag{56}$$

satisfying the generalized Leibniz rule

$$\boldsymbol{\nabla}^{\mathscr{E}}(\lambda \wedge \alpha) = \mathrm{d}\lambda \wedge \alpha + (-1)^{|\alpha|} \alpha \wedge \boldsymbol{\nabla}^{\mathscr{E}} \alpha, \tag{57}$$

for all $\lambda \in \Omega^*(\mathscr{M})$ and $\alpha \in \Omega^*(\mathscr{M}, \mathscr{E})$. If in addition $\mathscr{E}$ denotes a Clifford module bundle and the superconnection fulfills

$$[\boldsymbol{\nabla}^{\mathscr{E}}, c(a)] = c(\nabla^{\mathrm{Cl}} a), \quad \forall a \in \Gamma(\mathscr{C}(\mathscr{M})), \tag{58}$$

it is called a *Clifford superconnection* (CSC) (cf. Ref. 5). In this case we have (cf. loc. cit.)

$$\boldsymbol{\nabla}^{\mathscr{E}} \mapsto \mathrm{D}_{\boldsymbol{\nabla}} \varepsilon := c(\boldsymbol{\nabla}^{\mathscr{E}}) \tag{59}$$

is one-to-one. [Note, by abuse of notation we suppress the linear isomorphism (27).] Note that because of the generalized Leibniz rule any superconnection locally takes the form

$$\begin{gathered} \boldsymbol{\nabla}^{\mathscr{E}} = d + \sum_{k=0}^{m} \mathscr{A}_{[k]}, \\ \mathscr{A}_{[k]} \in [\Omega^k(\mathscr{M}, \mathrm{End}(\mathscr{E}))]^{-}. \end{gathered} \tag{60}$$

In particular, in the case of a CSC

$$\begin{gathered} \mathscr{A}_{[1]} = \boldsymbol{\omega}^{\mathrm{Cl}} \otimes \mathbf{1}_{\mathrm{End}_{\mathrm{Cl}}(\mathscr{E})} + \mathbf{1}_{\mathrm{Cl}} \otimes A, \\ \mathscr{A}_{[k]} = \sum_{1 \le i_1 < \cdots < i_k \le m} e^{i_1} \wedge \cdots \wedge e^{i_k} \otimes \mathbf{1}_{\mathrm{Cl}} \otimes B_{i_1 \ldots i_k}, \quad \forall k = 0,2,\ldots,m, \end{gathered} \tag{61}$$

where $\boldsymbol{\omega}^{\mathrm{Cl}}$ denotes the induced Levi-Civita form on the Clifford bundle $\mathscr{C}(\mathscr{M})$ and $A, B \in [\Omega^*(\mathscr{M}, \mathrm{End}_{\mathrm{Cl}}(\mathscr{E}))]^{-}$. Again, $\{e^i\}_{1 \le i \le m}$ is a local orthonormal frame in $T^*\mathscr{M}$. Moreover, if $\mathscr{M}$ denotes a spin-manifold, then any CSC is of the form (cf. Ref. 5)

$$\boldsymbol{\nabla}^{\mathscr{E}} = \nabla^S \otimes \mathbf{1}_{\mathrm{E}} + \mathbf{1}_{\mathrm{S}} \otimes \boldsymbol{\nabla}^{\mathrm{E}}. \tag{62}$$

Clearly, the notion of a CSC completely parallels that of a Clifford connection and coincides with the latter iff $B\equiv 0$. However, in general a CSC is not just defined by an element of $\Omega^1(\mathscr{M},\mathrm{End}^+_{\mathrm{Cl}}(\mathscr{E}))$. This will be of crucial importance in what follows. Indeed, in Ref. 11 it was shown how the combined Einstein–Hilbert–Yang–Mills (EHYM) action functional can be derived using non-SDOs. Before we define a particular functional on $\mathscr{D}(\mathscr{E})$ we still give another remark.

Remark 6: Let $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ be a Dirac operator on $\mathscr{E}$ and $\boldsymbol{\nabla}^{\mathscr{E}}$ be the corresponding CSC. Then we have

$$\widetilde{\nabla}'^{\mathscr{E}}:=\nabla^{\mathscr{E}}+\omega \tag{63}$$

with

$$\omega:=\omega_\phi+e^i\otimes\left(\sum_{k=1}^{m-1}\frac{1}{k!}\sum_{i_1,\ldots,i_k=1}^{m}c(e^{i_1})\cdots c(e^{i_k})\otimes B_{ii_1\ldots i_k}\right), \tag{64}$$

where $\omega_\phi:=\delta_\xi\phi$, $\phi\equiv\mathscr{A}_{[0]}\in\Omega^0(\mathscr{M},\mathrm{End}^-(\mathscr{E}))$, so that

$$\mathrm{D}_{\widetilde{\nabla}'^{\mathscr{E}}}=\widetilde{\mathrm{D}}. \tag{65}$$

In particular, $\widetilde{\nabla}'^{\mathscr{E}}\sim\widetilde{\nabla}^{\mathscr{E}}$, where the latter is defined by (44). Note, the local decomposition of the form $\omega\in\Omega^1(\mathscr{M},\mathrm{End}^+(\mathscr{E}))$ may also contain a degree $[k]=1$ form.

After summarizing the notion of Clifford modules and generalized Dirac operators we have also proved some lemmas, which permit an understanding of the relations between Dirac operators and connections on a Clifford module bundle. The reason to clarify these relations mainly is motivated by the following.

Definition 6: Let $(\mathscr{M},g)$ be a closed, compact, orientable Riemannian manifold of even dimension $(m=2n>2)$ and without boundary. Also, let $(\mathscr{E},c)$ be a Clifford module bundle over $\mathscr{M}$ and let $\mathscr{D}(\mathscr{E})$ be the affine space of all (generalized) Dirac operators compatible with the Clifford action c on $\mathscr{E}$. Then, we introduce the functional

$$\begin{aligned}\mathrm{res}:\mathscr{D}(\mathscr{E})&\to\mathbb{C}\\ \widetilde{\mathrm{D}}&\mapsto\mathrm{res}(\widetilde{\mathrm{D}}^{-2n+2}).\end{aligned} \tag{66}$$

Here, res means the Wodzicki residue, which in this case takes the explicit form (cf. Refs. 14–17)

$$\mathrm{res}(\widetilde{\mathrm{D}}^{-2n+2})=\frac{2}{\Gamma(n-1)}\int_{\mathscr{M}}*\mathrm{tr}_{\mathscr{E}}\left[\frac{1}{6}\,r_{\mathscr{M}}-\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}\right]. \tag{67}$$

and where the endomorphism $\mathscr{F}^{\widetilde{\nabla}^{\mathscr{E}}}\in\Gamma(\mathrm{End}(\mathscr{E}))$ is given by (47).

Since there exists a connection $\widetilde{\nabla}^{\mathscr{E}}\in\mathscr{A}(\mathscr{E})$ for every $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$, so that $\widetilde{\mathrm{D}}=\widetilde{\mathrm{D}}_{\widetilde{\nabla}^{\mathscr{E}}}$, this functional may be interpreted, alternatively, as a functional defined on $\mathscr{A}(\mathscr{E})$. However, we are interested less in the functional (66) itself than in the fact that for a *given* Dirac operator $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ the Wodzicki residue of $\widetilde{\mathrm{D}}^{-2n+2}$ can be considered as a functional

$$\mathscr{T}'_{\mathrm{D}}(\widetilde{\nabla}^{\mathscr{E}}):=\mathrm{res}(\widetilde{\mathrm{D}}^{-2n+2}), \tag{68}$$

of all connections $\widetilde{\nabla}\in\mathscr{A}(\mathscr{E})$ so that $c(\widetilde{\nabla}^{\mathscr{E}})=\widetilde{\mathrm{D}}$. In other words, with respect to a given Dirac operator (68) can be considered as a certain functional on the subspace of all (endomorphism valued) differential forms defining this Dirac operator. In the case that $\widetilde{D}$ denotes a SDO the functional (68) is proportional to the Einstein–Hilbert action. We again stress that this was recognized by Connes (cf. Ref. 10) and was proved in Ref. 18. From a more general point of view (see below) this was also discovered in Ref. 17, which in turn was the starting point to deal with non-SDOs in Refs. 19 and 11.

There is still another motivation for (68): the connection of (68) to the heat trace associated with (the square of) a Dirac operator. For this let us remind the reader that there is a natural functional on $\mathscr{D}(\mathscr{E})$:

$$\begin{aligned} \mathscr{D}(\mathscr{E}) &\to \mathbb{C}, \\ \widetilde{\mathrm{D}} &\mapsto \mathrm{Tr}\, e^{-\tau\widetilde{\mathrm{D}}^2}. \end{aligned} \tag{69}$$

Though in general one is not able to calculate this functional, it is well known that it has an asymptotic expansion:

$$\mathrm{Tr}\, e^{-\tau\widetilde{\mathrm{D}}^2} \sim (4\pi\tau)^{-n} \sum_{k\geqslant 0} \tau^{(k-2n)/2} \sigma_k(\widetilde{\mathrm{D}}^2) \quad (\tau \to 0), \tag{70}$$

where the coefficients (Seeley–DeWitt coefficients)

$$\sigma_k(\widetilde{\mathrm{D}}^2) := \int_{\mathscr{M}} * \mathrm{tr}_{\mathscr{E}}\, \sigma_k(x;\widetilde{\mathrm{D}}^2) \tag{71}$$

are known to contain geometric information. In particular, the subleading term $\sigma_2(\widetilde{\mathrm{D}}^2)$ is of the general form (cf. Ref. 20)

$$\sigma_2(\widetilde{\mathrm{D}}^2) = \int_{\mathscr{M}} * \mathrm{tr}_{\mathscr{E}} \left[\frac{1}{6}\, r_{\mathscr{M}} - \mathscr{F} \right], \tag{72}$$

and thus is proportional to (66). Of course, this is by no means accidental. In general, for all $(2n-k)/2 \notin \mathbb{Z}_-$ ($\mathscr{M}$ smooth) one has (cf. Ref. 15)

$$\sigma_k(\widetilde{\mathrm{D}}^2) = \frac{\Gamma((2n-k)/2)}{2}\, \mathrm{res}(\widetilde{\mathrm{D}}^{-2n+k}). \tag{73}$$

From this point of view the statement of our main theorem (as given in the Introduction) may be rephrased as follows:

There exists a Hamiltonian (generalized Laplacian) $\mathscr{H}$ such that the subleading term in the asymptotic expansion of the corresponding heat trace associated with this Hamiltonian is proportional to the classical bosonic action of the standard model with gravity included. Moreover, this Hamiltonian has a square root, $\mathscr{H} = \widetilde{\mathrm{D}}^2$, which gives rise also to the fermionic action of the standard model.

In what follows we assume that $\mathscr{M}$ denotes a *spin* manifold. (Indeed, it is widely believed by physicists that fermions are geometrically described by spinors.) Although this is not necessary, since we are only interested in local objects (densities), it simplifies notation. Consequently, the (total space of the) Clifford module bundle globally takes the form $\mathscr{E} = \mathrm{S} \otimes \mathrm{E}$. To get in touch with gauge theory we assume that E denotes an associated (Hermitian) vector bundle: $\mathrm{E} = \mathscr{P} \times_\rho V$, where $\mathscr{P}$ is a G-principal bundle over $\mathscr{M}$ and $\rho : \mathrm{G} \to V$ is a unitary representation of the (real, compact and semi-simple) Lie group G on a Hermitian vector space V. By $\mathscr{G}$ we denote the corresponding Lie algebra of G and by ρ' we denote the induced representation of $\mathscr{G}$ on $\mathrm{End}(V)$. Consequently, any connection form A on E takes its values in $\rho'(\mathscr{G}) \subset \mathrm{End}(V)$. This offers the possibility of defining the slightly more general functional

$$\mathscr{T}_{\mathrm{D}}(\widetilde{\nabla}^{\mathscr{E}}) := \mathrm{res}_\zeta(\widetilde{\mathrm{D}}^{-2n+2}) \equiv \mathrm{res}((\zeta\widetilde{\mathrm{D}}^2)^{-n+1}). \tag{74}$$

[That this functional actually is more adequate than (68) becomes clear when discussing the physical implications of the proposed model (cf. Sec. III).] Here, $\zeta \in \Gamma(\mathrm{End}(\mathscr{E}))$ denotes an element of the commutant defined by (G,ρ). More precisely, let us recall the simple fact that any section $s \in \Gamma(\mathrm{E})$ in an associated vector bundle $\mathrm{E} = \mathscr{P} \times_\rho V$ uniquely corresponds to an equivariant section $\overline{s} \in \Gamma_{eq}(\mathscr{P}, V)$. Here, equivariant means $\overline{s}(pg) = \rho^{-1}(g)\overline{s}(p)$, $\forall g \in G$, $p \in \mathscr{P}$. We then have the following.

Definition 7: For a given generalized Dirac operator, consisting of the triple

$$(G,\rho,\widetilde{\mathrm{D}}) \tag{75}$$

with $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$, let us denote by $\overline{z}\in\Gamma_{aq}(\mathscr{P},\mathrm{End(V)})$ an element of the *commutant*

$$\mathscr{C}_\rho(\mathrm{G}):=\{a\in\mathrm{End(V)}|[\rho(g),a]=0,\forall g\in G\}, \tag{76}$$

and by $z\in\Gamma(\mathrm{End(E)})$ its corresponding section in the endomorphism bundle associated with E. (In the following we shall not distinguish between s and $\overline{s}$.) Since End(S) is simple, this generalizes to End($\mathscr{E}$) via $\zeta:=\mathbf{1}_\mathrm{S}\otimes z$. We impose the following three conditions on ζ: it is a positive operator ($\zeta>0$) and satisfies $[\widetilde{\mathrm{D}},\zeta]=[\chi,\zeta]=0$, where $\chi\in\Gamma(\mathrm{End}(\mathscr{E}))$ denotes the grading operator on $\mathscr{E}=\mathscr{E}^+\oplus\mathscr{E}^-$.

Lemma 5: As a consequence, ζ^α has a constant spectrum (i.e., it is independent of $x\in\mathscr{M}$) and the operator: $\zeta^\alpha\widetilde{\mathrm{D}}^2$ is elliptic for any power $\alpha\in\mathbb{R}$. Using this, we obtain

$$\mathrm{res}_\zeta(\widetilde{\mathrm{D}}^{-2n+2})=\frac{2}{\Gamma(n-1)}\int_\mathscr{M}*\mathrm{tr}_\mathscr{E}(\zeta\sigma_2(x;\widetilde{\mathrm{D}}^2)). \tag{77}$$

Proof: Let us denote by spec $\zeta(x)=\{\lambda_1\ldots\lambda_k\}|_x$, $x\in\mathscr{M}$, the spectrum of the positive operator $\zeta\in\Gamma(\mathrm{End}(\mathscr{E}))$. Here, k indicates the number of distinct eigenvalues of ζ. We first prove that the spectrum is independent of $x\in\mathscr{M}$.

Since $z\in\mathscr{C}_\rho(G)$ this section is gauge invariant, i.e., for all gauge transformations $f\in\Gamma_{eq}(\mathrm{Aut}(\mathscr{P}))\simeq\Gamma(\mathscr{P}\times_{ad}\mathrm{G})$ of $\mathscr{P}$ we have $f^*z=z$. Hence, it is sufficient to consider a local situation. Using the fact that locally, any Dirac operator $\widetilde{\mathrm{D}}\in\mathscr{D}(\mathscr{E})$ may be written as

$$\widetilde{\mathrm{D}}=c(\mathrm{d})+\sum_{j=0}^{m}c(\mathscr{A}_{[j]}), \tag{78}$$

it follows that $[\widetilde{\mathrm{D}},\zeta]=0$ is equivalent to

$$\begin{aligned}&[c(\mathscr{A}_{[j]},\zeta)]=0,\quad\forall j=0,2,...,m,\\&c(\mathrm{d}\zeta)=-[c(\mathscr{A}_{[1]}),\zeta].\end{aligned} \tag{79}$$

However, the latter implies $\mathrm{d}z=0$. Indeed, $\mathscr{A}_{[1]}=\omega^S\otimes\mathbf{1}_\mathrm{E}+\mathbf{1}_\mathrm{S}\otimes\mathrm{A}$, $\mathrm{A}\in\Omega^1(\mathscr{M},\rho'(\mathscr{G}))$ and thus $[\mathrm{A},z]=0$. Consequently, ζ must be constant, which yields the first assertion.

Since the spectrum of z is constant we obtain

$$\mathrm{E}=\oplus_{j=1}^k\ker(z-\lambda_j)=:\oplus_{j=1}^k\mathrm{E}_j\,. \tag{80}$$

Hence,

$$\begin{aligned}\zeta&=\oplus_{j=1}^k\lambda_j\mathbf{1}_{\mathscr{E}_j},\\\mathrm{D}&=\oplus_{j=1}^k\mathrm{D}_j,\end{aligned} \tag{81}$$

where $\widetilde{\mathrm{D}}_j$ is the restriction of $\widetilde{D}$ to $\mathscr{E}_j:=\mathrm{S}\otimes\mathrm{E}_j$. Consequently, we end up with

$$\zeta\widetilde{\mathrm{D}}^2=\oplus_{j=1}^k\lambda_j\widetilde{\mathrm{D}}_j^2, \tag{82}$$

which implies

$$\mathrm{Tr}\;e^{-\tau\zeta\widetilde{\mathrm{D}}^2}=\sum_{j=1}^k\mathrm{Tr}\;e^{-\tau\lambda_j\widetilde{\mathrm{D}}_j^2}\sim\sum_{l\geqslant0}\sum_{j=1}^k\tau^{(l-m)/2}\lambda_j^{(l-m)/2}\int_\mathscr{M}*\mathrm{tr}_{\mathscr{E}_j}\sigma_l(x,\widetilde{\mathrm{D}}_j^2), \tag{83}$$

and thus proves the lemma, when ζ is replaced by ζ^{-n+1}.

Since we now have fixed the mathematical frame we conclude this part by summarizing the proposed model building kit as follows: Let

$$(\mathrm{G},\rho,\widetilde{\mathrm{D}}) \tag{84}$$

be a given generalized Dirac operator defined on a Clifford module bundle $\mathscr{E}$. Then, the general action functional on $\mathscr{A}(\mathscr{E})\times\Gamma(\mathscr{E})$ is defined as

$$\mathscr{I}_{\widetilde{\mathrm{D}}}\equiv\mathscr{I}_{\mathrm{fermionic}}(\widetilde{\nabla}^{\mathscr{E}},\psi)+\mathscr{I}_{\mathrm{bosonic}}(\widetilde{\nabla}^{\mathscr{E}}):=(\psi,\widetilde{\mathrm{D}}\psi)_{\Gamma(\mathscr{E})}+\mathrm{res}_{\zeta}(\widetilde{\mathrm{D}}^{-2n+2}), \tag{85}$$

where $c(\widetilde{\nabla}^{\mathscr{E}})=\widetilde{\mathrm{D}}$.

III. PART 2: THE STANDARD MODEL

In this part of the paper we are concerned with the application of the kit introduced in Part 1 concerning the standard model of particle physics. We therefore shall introduce in the following section an appropriate generalization of the Dirac–Yukawa operator and prove our main theorem. Moreover, we shall discuss some consequences regarding the various parameters involved in the model.

A. The Dirac–Yukawa operator and the EHYMH action

To get started let us give the following.

Definition 8: Let $\mathscr{E}:=\mathrm{S}\otimes\mathrm{E}$ be a Clifford module bundle with a twisting graduation $(\mathrm{E}=\mathrm{E_L}\oplus\mathrm{E_R})$ and denote by $\chi:=\gamma_5\otimes\chi^E$ the appropriate grading operator: $\chi^2=\mathbf{1}_{\mathscr{E}}$, $\chi^*=\chi$. (Later we shall be mostly interested in the case $n=2$.) A Dirac operator D_ϕ is called a (Euclidean) *Dirac–Yukawa* operator if it takes the form

$$\begin{aligned}\mathrm{D}_\phi&:=\mathrm{D}+i\gamma_5\otimes\begin{pmatrix}0&\widetilde{\phi}\\ \widetilde{\phi}^*&0\end{pmatrix},\\ &\equiv\mathrm{D}+\gamma_5\otimes\phi,\end{aligned} \tag{86}$$

where D is a SDO and $\widetilde{\phi}\in\Gamma(\mathrm{Hom}(\mathrm{E_R},\mathrm{E_L}))$.

Since the Yukawa coupling (9) geometrically can be considered as defining a particular section ϕ (see below), one may try to naturally generalize the operator (86) in such a way that it not only defines the fermionic action (8) and (9), but also yields the bosonic action (10) and (11). Here, ''naturally'' means that the generalization of (86) is determined by only those elements which already determine the Dirac–Yukawa operator, i.e., by (g,ϕ,A).

Theorem 1: *Let $\mathscr{E}=\mathrm{S}\otimes\mathrm{E}$ be a twisted Clifford module over $\mathscr{M}$, with $\mathrm{E}:=\mathrm{E_L}\oplus\mathrm{E_R}$. The functional (74) evaluated with respect to the Dirac operator*

$$\widetilde{\mathrm{D}}_\phi:=\mathrm{D}+a_4\Phi+\mathscr{J}(a_2c(F^{\mathscr{E}/S})+a_3c(\nabla^{\mathrm{End}(\mathscr{E})}\Phi)+a_0\Phi^2) \tag{87}$$

defined on $\widetilde{\mathscr{E}}:=\mathrm{S}\otimes\widetilde{\mathrm{E}}$ yields

$$\begin{aligned}\mathrm{res}_{\zeta}(\widetilde{\mathrm{D}}_\phi^{-2n+2})=-\frac{\mathrm{tr}_{\mathscr{E}}\,\zeta}{3\Gamma(n-1)}\int_{\mathscr{M}}*\{&r_{\mathscr{M}}+a_4'\,\mathrm{tr}_E(z\phi^2)-a_2'\,\mathrm{tr}_E(zF_{\mu\nu}F^{\mu\nu})\\ &+a_3'\,\mathrm{tr}_E(z\nabla_\mu\phi\nabla^\mu\phi)+a_0'\,\mathrm{tr}_E(z\phi^4)\},\end{aligned} \tag{88}$$

with

$$\begin{aligned}a_0'&:=\frac{24n(1-1/2n)}{\mathrm{tr_E}z}\,a_0^2,\quad a_2':=\frac{24(2n-3)}{\mathrm{tr_E}z}\,a_2^2,\\ a_3'&:=\frac{24(n-1)}{\mathrm{tr_E}z}\,a_3^2,\quad a_4':=\frac{12}{\mathrm{tr_E}z}\,a_4^2.\end{aligned} \tag{89}$$

Here, $\Phi := \gamma_5 \otimes \phi$ and $F \in \Omega^2(\mathscr{M}, \rho'(\mathscr{G}))$ is the Yang–Mills curvature, induced by the gauge potential A in the definition of the SDO $D \in \mathscr{D}(\mathscr{E})$. The covariant derivative $\nabla\phi$ is defined with respect to the adjoint representation of G, where the ϕ sits [i.e., the corresponding equivariant section $\phi \in \Gamma_{eq}(\mathscr{P}, \mathrm{End}(V))$ fulfills $\phi(pg) = \rho^*(g)\phi(p)\rho(g)$, $\forall p \in \mathscr{P}$, $g \in G$.] The a's denote arbitrary (complex) constants. The structure group G is assumed to act on $\widetilde{E} := E \oplus E$ via the representation $\widetilde{\rho} := \rho \oplus \rho$, and the automorphism $\widetilde{E} \xrightarrow{\mathscr{J}} \widetilde{E}$ denotes the corresponding ''complex structure'' on $\widetilde{E}$. Moreover, we have used the canonical identification

$$A + \mathscr{J}(B) = \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \tag{90}$$

for all $A, B \in \mathrm{End}(E)$. Note, in what follows we do not distinguish between the Clifford action c on the Clifford module $\mathscr{E}$ and the corresponding action $\widetilde{c} := c \oplus c$ on the (canonically) induced Clifford module $\widetilde{\mathscr{E}}$. Likewise, we do not distinguish between the representation ρ and $\widetilde{\rho}$.

Consequently, if the constant a_3 is purely imaginary and the other constants are real, the functional (74) with respect to the generalized Dirac–Yukawa operator (87) becomes proportional to the Einstein–Hilbert–Yang–Mills–Higgs action (EHYMH) of the standard model. Moreover, if one considers ''diagonal sections,'' $\widetilde{\psi} := (\psi, \psi)$, $\psi \in \Gamma(\mathscr{E})$ only, the fermionic functional in (85) becomes proportional to the Dirac–Yukawa action. Note, in (85) there is still a length scale missing because the fields do not yet have the right dimensions. This will be discussed in the next section where we shall consider some constraints of our approach to the standard model.

Remark 7: The (Euclidean) Dirac–Yukawa operator (86) is uniquely defined by the Clifford superconnection

$$\boldsymbol{\nabla}^{E} := \nabla^{E} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} := i^n * \phi \tag{91}$$

on E, where we have used the fact that the grading operator γ_5 of the spinor bundle S is proportional to the volume form on $\mathscr{M}$. More precisely, we have $c(i^n * 1) = \gamma_5 \otimes \mathbf{1}_E \in \Gamma(\mathrm{End}^+(\mathscr{E}))$.

Hence, the (Euclidean) Dirac–Yukawa operator is determined by a one-form (gauge potential) and the $2n$-form in (91). In four dimensions, however, the most general Dirac operator in addition depends on a zero-form, a two-form, and a three-form. By considering only those forms which already determine the Dirac–Yukawa operator, (91) naturally yields the following ansatz,

$$\boldsymbol{\nabla}^{\widetilde{E}} := \nabla^{E} + a_4\boldsymbol{\epsilon} + \mathscr{J}(a_2 F - a_3 * \nabla^{\mathrm{End}(E)}(*\boldsymbol{\epsilon}) + a_0(*\boldsymbol{\epsilon})^2), \tag{92}$$

on $\widetilde{E}$. Note that the forms within the brackets are even with respect to the total degree and hence $\mathscr{J}(\cdots)$ becomes odd (cf. also the final section). It is easily checked that

$$\widetilde{D}_{\boldsymbol{\nabla}^{\widetilde{\mathscr{E}}}} = \widetilde{D}_{\phi} \tag{93}$$

with $\nabla^{\widetilde{\mathscr{E}}} := \nabla^{S} \otimes \mathbf{1}_E + \mathbf{1}_S \otimes \boldsymbol{\nabla}^{\widetilde{E}}$.

There are two interesting choices for the constants a_k so that (87) takes a particularly nice geometric form: First, the most natural choice is $a_k \equiv 1$, $\forall k = 0,2,...,4$. In this case the Dirac operator (87) reads

$$\widetilde{D}_{\phi} = c(\nabla^{\mathscr{E}} + \Phi) + \mathscr{J}(c(\mathscr{R}_{\widetilde{\boldsymbol{\nabla}}}\tilde{e})) =: c(\boldsymbol{\nabla}^{\mathscr{E}}) + \mathscr{J}(c(\mathscr{R}_{\widetilde{\boldsymbol{\nabla}}}\tilde{e})) = D_{\phi} + \mathscr{J}(c(\mathscr{R}_{\widetilde{\boldsymbol{\nabla}}}\tilde{e})) \equiv D_{\widetilde{\boldsymbol{\nabla}}}\tilde{e} \tag{94}$$

with the ''super relative curvature'' $\mathscr{R}_{\widetilde{\boldsymbol{\nabla}}}\tilde{e} := (\widetilde{\boldsymbol{\nabla}}^{\mathscr{E}})^2 - (\nabla^{S})^2 \otimes \mathbf{1}_E$.

However, there is still another nice choice: $a_0 = a_3 := (1 - 1/2n)$, $a_2 = a_4 := 1$; in this case the generalization (87) of the Dirac–Yukawa operator (86) reads

$$\begin{aligned} \widetilde{D}_{\phi} &= c(\nabla^{\mathscr{E}} + \omega_{\phi}) + \mathscr{J}(c(F^{\mathscr{E}/S} + d^{\nabla^{\mathscr{E}}}\omega_{\phi} + \omega_{\phi} \wedge \omega_f)) \\ &=: c(\widetilde{\boldsymbol{\nabla}}^{\mathscr{E}}) + \mathscr{J}(c(\mathscr{R}_{\widetilde{\nabla}}\tilde{e})) = D_{\phi} + \mathscr{J}(c(\mathscr{R}_{\widetilde{\nabla}}\tilde{e})) \equiv D_{\widetilde{\nabla}}\tilde{e}, \end{aligned} \tag{95}$$

with the ''*Higgs form*'' $\omega_\phi := \delta_\xi \Phi \in \Omega^1(\mathscr{M}, \mathrm{End}^+(\mathscr{E}))$ and the relative curvature $\mathscr{R}_{\tilde{\nabla}^{\mathscr{E}}} := (\tilde{\nabla}^{\mathscr{E}})^2 - (\nabla^S)^2 \otimes \mathbf{1}_E$. Because of

$$D_{\tilde{\mathbf{\nabla}}}\tilde{\mathscr{E}} - D_{\tilde{\nabla}}\tilde{\mathscr{E}} = \frac{1}{2n}\,\mathscr{J}(c(\mathscr{R}_{\tilde{\mathbf{\nabla}}}\mathscr{E} - \mathrm{F}^{\mathscr{E}/S})), \tag{96}$$

the Dirac operators $D_{\tilde{\mathbf{\nabla}}}\tilde{\mathscr{E}}$ and $D_{\tilde{\nabla}}\tilde{\mathscr{E}}$ however, are different. Therefore, strictly speaking, the definition (87) gives a whole class of Dirac operators, parametrized by the constants $a_0, \ldots, a_4$, and which all yield (88).

After this remark let us turn to the proof of the theorem.

Proof: Clearly, to prove this theorem we just have to calculate the endomorphism (47) with respect to the Dirac operator (87). This tedious but straightforward calculation can most easily be achieved using the local form of (47):

$$\begin{aligned}\mathscr{F}^{\tilde{\nabla}^{\tilde{\mathscr{E}}}} = {} & \tfrac{1}{4} r_{\mathscr{M}} \mathbf{1}_{\tilde{\mathscr{E}}} + \tfrac{1}{2}\gamma^{\mu\nu} \otimes \mathrm{F}_{\mu\nu} + \tfrac{1}{2}[\gamma^\mu[\omega_\mu, \gamma^\nu], \omega_\nu] + \gamma^{\mu\nu}({}'\nabla_\mu \omega_\nu) - \tfrac{1}{2}\gamma^\mu[({}'\nabla_\nu \omega_\mu), \gamma^\nu] \\ & + \tfrac{1}{2}\gamma^{\mu\nu}[\omega_\mu, \omega_\nu] + \tfrac{1}{4} g_{\mu\nu}\gamma^\alpha[\omega_\alpha, \gamma^\mu]\gamma^\beta[\omega_\beta, \gamma^\nu],\end{aligned} \tag{97}$$

where $\mathscr{C}(\mathscr{M})_{\mathbb{C}} \xrightarrow{\gamma} \mathrm{End}(S)$ denotes the (chiral) representation of the complexified Clifford algebra and $\gamma^\mu \equiv c\,\gamma(e^\mu)$ with $\{e^\mu\}_{1 \leqslant \mu \leqslant 2n}$ a local frame in $T^*\mathscr{M}$. Also we use the shorthand notation $\gamma^{\mu\nu} := [\gamma^\mu, \gamma^\nu]/2$. Moreover, F is the curvature on E, associated to the connection form A. The connection $\tilde{\nabla}^{\tilde{\mathscr{E}}} \in \mathscr{A}(\mathscr{E})$ denotes any representative of the corresponding class defining the Dirac operator (87) and $\omega := \tilde{\nabla}^{\mathscr{E}} - \nabla^{\tilde{\mathscr{E}}}$, with $\nabla^{\tilde{\mathscr{E}}} \in \mathscr{A}_{\mathrm{Cl}}(\mathscr{E})$ arbitrarily chosen, and ${}'\nabla_\mu$ denotes the covariant derivative induced by this Clifford connection on $\mathrm{End}(\mathscr{E})$.

In particular, we may choose $\tilde{\nabla}^{\tilde{\mathscr{E}}}$ such that

$$\omega := \omega^0 + \omega^2 + \omega^3 + \omega^4$$

with

$$\begin{aligned}\omega^0 := a_0 \delta_\xi \mathscr{J}(\Phi^2), \quad \omega^2 := a_2 e^\mu \otimes \gamma^\nu \otimes \mathscr{J}(F_{\mu\nu}), \\ \omega^3 := a_3 \mathscr{J}(\nabla\Phi), \quad \omega^4 := a_4 \delta_\xi(\Phi).\end{aligned} \tag{98}$$

Note, we have already omitted $\omega^1 := \omega^S \otimes \mathbf{1}_E + \mathbf{1}_S \otimes A$, which would change the Clifford connection $\nabla^{\tilde{\mathscr{E}}}$ only and not the functional (88) by Corollary 1. The main advantage of the local form (97) is that the whole calculation becomes purely algorithmic. Moreover, because of the trace $\mathrm{tr}_{\tilde{\mathscr{E}}}$ in (77), one only has to calculate the last four terms in (97). For the same reason most of the terms to be calculated will drop out. Indeed, using (98), it is easily checked that the two derivative terms in (97), actually, do not contribute to the functional (74). Hence, all that is left to be done is to calculate the last two quadratic terms in (97) up to ''traceless'' contributions. This calculation becomes even more simplified by the following.

Remark 8: Let $\mathscr{E}_j$, $j = 1,2$, be two Clifford modules over $\mathscr{M}$ with the appropriate actions c_j. Then, $(\mathscr{E}, c)$ with $\mathscr{E} := \mathscr{E}_1 \oplus \mathscr{E}_1$ and $c := c_1 \oplus c_2$ is also a Clifford module. The most general Dirac operator on this Clifford module takes the form

$$\tilde{\mathrm{D}} = \begin{pmatrix} \mathrm{D}_1 & A_{12} \\ A_{21} & \tilde{\mathrm{D}}_2 \end{pmatrix}, \tag{99}$$

where, respectively, $\tilde{\mathrm{D}}_j \in \mathscr{D}(\mathscr{E}_j)$, $j = 1,2$, $A_{12} \in \Gamma(\mathrm{Hom}(\mathscr{E}_2, \mathscr{E}_1))$ and $A_{21} \in \Gamma(\mathrm{Hom}(\mathscr{E}_1, \mathscr{E}_2))$. The Wodzicki residue of $\tilde{\mathrm{D}}^{-2n+2}$ is such that there are no terms ''mixing the diagonal with the off-diagonal'' (cf. Ref. 11).

As a consequence, in calculating the corresponding quadratic terms of (97), products of the generic form $\omega^4 \omega^k$, $k = 0,\ldots,3$, may be omitted as well. As a result we end up with

$$\frac{1}{2}\,\gamma^{\mu\nu}[\omega_\mu,\ \omega_\nu]=a_0^2\left(1-\frac{1}{2n}\right)\Phi^4+a_2^2\mathrm{ev}_g(\mathrm{F}^{\mathscr{E}/\mathrm{S}^2})+a_4^2\left(1-\frac{1}{2n}\right)\Phi^2 \tag{100}$$

and

$$\frac{1}{4}\,g_{\mu\nu}\gamma^\alpha[\omega_\alpha,\ \gamma^\mu]\gamma^\beta[\omega_\beta,\ \gamma^\nu]=2na_0^2\left(1-\frac{1}{2n}\right)^2\Phi^4+(5-4n)a_2^2\mathrm{ev}_g(\mathrm{F}^{\mathscr{E}/\mathrm{S}^2})$$

$$+2(n-1)a_3^2\mathrm{ev}_g((\nabla^{\mathrm{End}(\mathscr{E})}\Phi)^2)+\frac{1}{2n}\,a_4^2\Phi^2. \tag{101}$$

Note that both equalities hold up only to traceless terms! Finally, if we put everything together, the theorem is proven.

To summarize: we have introduced a certain Dirac operator (87), generalizing the Dirac–Yukawa operator (86), such that the functional (74) looks like that of the bosonic action of the standard model with gravity including. This result is independent of the specific structure group G and its appropriate (fermionic) representation. To get in touch with physics, however, we also have to specify the pair (G,ρ). In other words, we have to define the generalized Dirac operator of the standard model.

B. The generalized Dirac–Yukawa operator of the standard model

To begin with, we give the following.

Definition 9: Let $\mathscr{E}$ be a twisted spinor bundle, so that $\mathrm{E}=\mathrm{E}_\mathrm{L}\oplus\mathrm{E}_\mathrm{R}$. We call

$$(\mathrm{G},\rho,\mathrm{D}) \tag{102}$$

the *generalized Dirac–Yukawa operator of the standard model* provided the structure group G takes the form

$$\mathrm{G}:=\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1) \tag{103}$$

and has the (fermionic) representation $\rho:=\rho_\mathrm{L}\oplus\rho_\mathrm{R}\colon \mathrm{G}\to\mathrm{Aut}(V)$,

$$\rho_\mathrm{L}(\mathbf{c},\mathbf{w},b):=\begin{pmatrix}\mathbf{c}\otimes\mathbf{1}_\mathrm{N}\otimes\mathbf{w}b_\mathrm{L}^q & 0\\ 0 & \mathbf{1}_\mathrm{N}\otimes\mathbf{w}b_\mathrm{L}^l\end{pmatrix}, \tag{104}$$

$$\rho_R(\mathbf{c},\mathbf{w},b):=\begin{pmatrix}\mathbf{c}\otimes\mathbf{1}_\mathrm{N}\otimes\mathbf{B}_\mathrm{R}^q & 0\\ 0 & \mathbf{B}_\mathrm{R}^l\end{pmatrix}, \tag{105}$$

on the typical fiber

$$V:=V_\mathrm{L}\oplus V_\mathrm{R} \tag{106}$$

$$\simeq[(\mathbb{C}_q^{6\mathrm{N}}\oplus\mathbb{C}_l^{2\mathrm{N}})]_\mathrm{L}\oplus[(\mathbb{C}^{3\mathrm{N}})_q\oplus\mathbb{C}_l^\mathrm{N}]_\mathrm{R} \tag{107}$$

$$\simeq\mathbb{C}_\mathrm{L}^{8\mathrm{N}}\oplus\mathbb{C}_\mathrm{R}^{7\mathrm{N}}. \tag{108}$$

Moreover, the Dirac operator

$$\mathrm{D}:=\widetilde{\mathrm{D}}_\phi \tag{109}$$

is the generalized Dirac–Yukawa operator (87), where the homomorphism $\widetilde{\phi}\in\Gamma(\mathrm{Hom}(\mathrm{E}_\mathrm{R},\mathrm{E}_\mathrm{L}))$ is defined by

$$\widetilde{\phi}:=\begin{pmatrix}\mathbf{1}_3\oplus(\mathbf{g}_y'^q\otimes\varphi,\mathbf{g}_y^q\otimes\boldsymbol{\epsilon}\overline{\varphi}) & 0\\ 0 & \mathbf{g}_y^l\otimes\varphi\end{pmatrix} \tag{110}$$

$$\equiv \begin{pmatrix} \mathbf{1}_3 \otimes \widetilde{\varphi}_q & 0 \\ 0 & \widetilde{\varphi}_l \end{pmatrix}. \tag{111}$$

Here, respectively, $\mathbf{g}_y'^q$, $\mathbf{g}_y^q \in \mathbf{M}_N(\mathbb{C})$ denote the matrices of the Yukawa coupling constants for quarks of electrical charge $-\frac{1}{3}$ and $\frac{2}{3}$ (i.e., of quarks of ''d'' type, and of ''u'' type) and $\mathbf{g}_y^l \in \mathbf{M}_N(\mathbb{C})$ is the matrix of the Yukawa coupling constants for the leptons of charge -1 (i.e., of leptons of ''electron'' type). While $\mathbf{g}_y^q$ and $\mathbf{g}_y^l$ can be assumed to be diagonal and real, the matrix $\mathbf{g}_y'^q$ is related to the Kobayashi–Maskawa matrix and therefore is neither diagonal nor real. The ''weak hypercharges'' for the left- and right-handed quarks (indicated by the subscript ''q'') and leptons (subscript ''l'') are defined by $\rho(b) := e^{iy\theta}$, $b \in \mathrm{U}(1)$, $y \in \mathbb{Q}$, $\theta \in [0, 2\pi[$ (cf. the Introduction). Then the two by two diagonal matrices $\mathbf{B}_R^q$ and $\mathbf{B}_R^l$ in the definition (105) are $\mathbf{B}_R^q := \mathrm{diag}\,(b_R^{d'}, b_R^u)$ and $\mathbf{B}_R^l := b_R^l \mathbf{1}_N$. In (110), φ denotes a section into a rank two subbundle E_h of the vector bundle E and carries the defining representation ρ_h of the electroweak subgroup $G_h := \mathrm{SU}(2) \times \mathrm{U}(1)$ of the structure group G_h i.e., when φ is considered as an element of $\Gamma_{eq}(\mathscr{P}, V)$ it transforms like $\varphi(pg) = \rho_h(g)^* \varphi(p)$ with $g := (\mathbf{w}, b) \in G_h = \mathrm{SU}(2) \times \mathrm{U}(1)$ and $\rho_h(\mathbf{w}, b) := \mathbf{w} e^{iy_h\theta}$. Finally, $\boldsymbol{\epsilon}$ is the antidiagonal matrix $\boldsymbol{\epsilon} := \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $\overline{\varphi}$ here means the complex conjugate of φ.

Hence, the Higgs field φ defines an element Φ in $\Omega^0(\mathscr{M}, \mathrm{End}^-(\mathscr{E}))$ and thus a Dirac–Yukawa operator (86). Of course, the particular form of $\widetilde{\phi}$ in (110) is such that Φ gives the correct Yukawa coupling term (9) in the definition of the fermionic action of the standard model. However, since the section $\phi \in \Gamma(\mathrm{End}^-(E))$ transforms with respect to the (ρ-induced) adjoint representation of G it follows that the hypercharge y_h of the Higgs field and, respectively, the hypercharges y_L^l, y_R^l of the left-and right-handed leptons and the hypercharges $y_L^q, y_R^{d'}, y_R^u$ of the left- and right-handed quarks must satisfy the following well-known relations:

$$y_h = y_L^l - y_R^l = y_L^q - y_R^{d'} = y_R^u - y_L^q. \tag{112}$$

In other words, when the Yukawa coupling is of the form as defined by (9) the corresponding hypercharges are not all independent but have to satisfy the relations (112). Moreover, since we know the electrical charge of the particles, the numerical values of the ''y's'' are fixed:

$$(y_L^q, y_L^l) = (\tfrac{1}{6}, -\tfrac{1}{2}), \tag{113}$$

$$((y_R^{d'}, y_R^u), y_R^l) = ((-\tfrac{1}{3}, \tfrac{2}{3}), -1). \tag{114}$$

This is a consequence of the generalized Gell–Mann–Nishijima relation

$$\mathbf{Q} = \mathbf{T}_3 + \mathbf{Y}, \tag{115}$$

where $\{i\mathbf{T}_3 := \rho'((i\boldsymbol{\tau}_3)/2),\ i\mathbf{Y} := \rho'(i)\}$ is a basis of the maximal Cartan subalgebra of $\rho'(\mathrm{su}(2) \oplus \mathrm{u}(1))$, so that $i\mathbf{Q}$ generates the residual structure group, in more physical terms the ''electromagnetic gauge group,''

$$\mathrm{U}_{elm}(1) \subset \mathrm{SU}(2) \times \mathrm{U}(1) \tag{116}$$

in the fermionic representation ρ after the mechanism of *spontaneous symmetry breaking* is established. [As usual we use $\{\mathbf{E}_k\}_{1 \leqslant k \leqslant 12} := \{(i\boldsymbol{\lambda}_a)/2, (i\boldsymbol{\tau}_b)/2, i\}_{\substack{1 \leqslant a \leqslant 8 \\ 1 \leqslant b \leqslant 3}}$ as a basis of $\mathrm{su}(3) \oplus \mathrm{su}(2) \oplus \mathrm{u}(1)$, where $\boldsymbol{\lambda}_a$, $a = 1, \ldots, 8$, denote the Gell–Mann matrices and $\boldsymbol{\tau}_b$, $b = 1, \ldots, 3$, denote the Pauli matrices.] To make the latter more precise mathematically, let us remember how the notion of spontaneous symmetry breaking can be geometrically rephrased in terms of the reduction of a G-principal bundle (cf. Refs. 21 and 22).

Let therefore $H \subset G$ be a Lie subgroup of G and, respectively, $\mathscr{P}_G$ and $\mathscr{P}_H$ be the (total spaces of the) corresponding principal bundles over the same base manifold $\mathscr{M}$. Then, $\mathscr{P}_H$ is called an H reduction of $\mathscr{P}_G$ iff $\mathscr{P}_H \subset \mathscr{P}_G$ is a submanifold, so that the injection $\mathscr{P}_H \hookrightarrow \mathscr{P}_G$ is a bundle homomorphism. A necessary and sufficient condition for a G-principal bundle $\mathscr{P}$ to be H reducible is that the $\mathscr{P}$-associated fiber bundle $\mathscr{P}_G/H \to \mathscr{M}$, with typical fiber G/H, admits a global section.

Definition 10: A solution (g,A,φ,Ψ) of the Euler–Lagrange equations corresponding to the functional (12) is called a *classical vacuum* iff $A=\Psi=0$, the Higgs field $\varphi=\varphi_0$ minimizes the Higgs potential

$$V:\Gamma(E_h)\to\mathbb{R},\quad V'|_{\varphi=\varphi_0}=0, \tag{117}$$

and the metric g satisfies the corresponding Einstein equation. Here, E_h denotes the (total space of the) bundle where the Higgs field lives (see below) and, in principle, V may be any gauge invariant polynomial of order less than or equal to 4 of the Higgs field φ. Again, it is assumed that V is bounded from below. Of course, concerning the standard model, V has its well-known fashion (6). Note that in the vacuum (g,φ_0) the metric has to fulfill

$$\mathrm{Ric}(g)=\Lambda(\varphi_0)g,$$

where Ric denotes the Ricci curvature associated with the metric g and Λ denotes a constant proportional to $V(\varphi_0)$. Hence, if $\varphi=\varphi_0$, then $(\mathscr{M},g)$ must be a *Einstein manifold* (see also our corresponding remark in the Introduction).

Let us denote by $\Sigma\subset V_h$ the (disjoint union of) orbits of classical vacuums with respect to the representation ρ_h that carries the Higgs field φ of the structure group G_h. In the special case of the standard model this group is identified with the ''electroweak gauge group'' $\mathrm{SU}(2)\times\mathrm{U}(1)\subset\mathrm{G}$ and, correspondingly, $\mathrm{E}_h:=\mathscr{P}\times_{\rho_h}\mathrm{V}_h\subset\mathrm{E}$ denotes the rank two subbundle describing, geometrically, the Higgs sector of the standard model. In the so-called ''minimal version'' of the standard model the typical fiber is $\mathrm{V}_h\simeq\mathbb{C}^2$. Note that in the usual (nongeometrical) description of the standard model the fermion representation ρ, defined by (104)–(108), and the Higgs representation ρ_h are completely independent. Let us denote by $\mathrm{I}(\varphi_0)\subset\rho_h(\mathrm{G}_h)\subset\mathrm{End}(\mathrm{V}_h)$ the isotropy group associated with a chosen classical vacuum $\varphi_0\in\Sigma'\subset\Sigma$, connected. Up to onjugation this isotropy group can be identified with some Lie subgroup H of G_h and we have $\Sigma'\simeq(\rho_h(\mathrm{G}_h)/\rho_h(\mathrm{H}))$. Therefore, φ_0 considered as an element in $\Gamma_{\mathrm{eq}}(\mathscr{P}_h,\mathrm{V}_h)$ uniquely induces a section (also denoted by φ_0) $\mathscr{M}\overset{\varphi_0}{\to}\mathscr{P}_h/\mathrm{H}$. In other words, from a geometrical point of view a necessary condition for a (classical) vacuum to exist is that the G_h bundle must be H reducible, where the Lie subgroup $\mathrm{H}\subset\mathrm{G}_h$ is identified with the isotropy group of some choosen Higgs field φ_0, minimizing the Higgs potential V.

Definition 11: The gauge symmetry is called *spontaneously* broken by the (classical) vacuum, represented by φ_0, iff $\mathrm{H}\subset\mathrm{G}_h$ is a proper subgroup.

Note that the H reduction of $\mathscr{P}_h$ is only necessary but not sufficient for φ_0 to represent a classical vacuum. It is also required that $\mathscr{P}_h$ possesses a flat connection. In the particular case of the (minimal) standard model we obtain $\Sigma=\{\rho_h(\mathbf{w},b)\varphi_0\}\amalg\{0\}$, with $\varphi_0:=(0,v/\sqrt{2})^T$, $v:=\sqrt{\mu^2/\lambda}\in\mathbb{R}$, and $(\mathbf{w},b)\in\mathrm{G}_h=\mathrm{SU}(2)\times\mathrm{U}(1)$. Then, the isotropy group associated with the nontrivial φ_0 is generated by the anti-Hermitian operator $i\mathbf{Q}$, and the corresponding residual structure group H can be identified with $\mathrm{U}_{\mathrm{elm}}(1)$ [cf. (115) and (116)]. Actually, the full residual structure group (''little group'') of the standard model reads

$$\mathrm{H}=\mathrm{SU}(3)\times\mathrm{U}_{\mathrm{elm}}(1), \tag{118}$$

since the structure group G is given by (103).

Consequently, assuming that $\mathscr{P}\equiv\mathscr{P}_{\mathrm{G}}$ is H reducible and φ_0 represents an appropriate (nontrivial) classical vacuum of the standard model the section $\mathscr{D}\in\Omega^0(\mathscr{M},\mathrm{End}^-(\mathrm{E}))$, defined by

$$\mathscr{D}\equiv\phi_0:=\phi|_{\varphi=\varphi_0}=:i\begin{pmatrix}0 & \mathbf{M}\\ \mathbf{M}^* & 0\end{pmatrix}, \tag{119}$$

is H invariant and thus well defined on the reduced bundle. Fixing the gauge so that $\varphi_0=(0,v/\sqrt{2})^T$ (''unitary gauge'') we may further write

$$\mathbf{M}\equiv\begin{pmatrix}\mathbf{1}_3\otimes\mathbf{M}_q & 0\\ 0 & \mathbf{M}_l\end{pmatrix},$$

with

$$\mathbf{M}_q = \begin{pmatrix} 0 & \mathbf{m}^{d\prime} \\ \mathbf{m}^u & 0 \end{pmatrix}, \tag{120}$$

$$\mathbf{M}_l = \begin{pmatrix} 0 \\ \mathbf{m}^l \end{pmatrix}, \tag{121}$$

where, respectively, the matrices $\mathbf{m}^l := (v/\sqrt{2})\mathbf{g}_y^l \in \mathrm{M_N}(\mathbb{C})$ and $\mathbf{m}^u := (v/\sqrt{2})\mathbf{g}_y^q \in \mathrm{M_N}(\mathbb{C})$ denote the ''mass matrices'' of the charged leptons (l) and quarks (q) of ''u type.'' They can be assumed to be diagonal and real. The corresponding N×N matrix $\mathbf{m}^{d'} := (v/\sqrt{2})\mathbf{g}_y'^q$ of ''d-type'' quarks is neither diagonal nor real. It is related to the mass matrix of ''d-type'' quarks $\mathbf{m}^d = \mathrm{diag}(m^{d_1},...,m^{d_N})$, $m^{d_k} \in \mathbb{R}$, $k = 1,...,N$, via the Kobayashi–Maskawa matrix $\mathbf{V} \in \mathrm{U(N)}$: $\mathbf{m}^{d'} = \mathbf{V}\mathbf{m}^d\mathbf{V}^*$.

Obviously, $\mathbf{M}$ denotes the *fermionic mass matrix* and we have recovered the ''internal Dirac operator'' $\mathscr{D}$ of the Connes–Lott approach to the standard model (cf. Ref. 3 and the corresponding references therein). Again, we stress that $\mathscr{D}$ minimizes the Higgs potential V but represents a classical vacuum only if $\mathscr{P}_\mathrm{H}$ (and thus $\mathscr{P}$) possesses a flat connection (cf. Ref. 22).

As we have already mentioned, in the usual approach to the standard model the representation ρ_h of the Higgs field φ is independent of the fermionic representation ρ_f as defined by (104)–(108). However, to be consistent one has to impose the relations (112) to the hypercharges. That these relations are not accidental and, actually, must not be chosen by hand follows from the Dirac–Yukawa operator of the standard model:

$$(\mathrm{G}, \rho, \mathrm{D}_\phi), \tag{122}$$

with $(\mathrm{G}, \rho \equiv \rho_f)$ as in (103)–(108) and D_ϕ defined by (110). In other words, whenever one starts with (122) the representation of the Higgs field must be contained in the fermionic representation and then the relations between the hypercharge of the Higgs field and those of the corresponding fermions are fixed. If this does not hold, the Yukawa coupling (9) would not define a generalized Dirac operator.

We now turn to the consequences as implied by the generalization $(\mathrm{G}, \rho, \widetilde{\mathrm{D}}_\phi)$ of (122). We therefore compare the functional (88), derived from the generalized Dirac–Yukawa operator (87), with the corresponding bosonic action of the standard model. Before we can do this, however, we still have to give the various fields involved in the model their right dimensions, i.e., we first have to introduce an arbitrary length scale ''l.'' Also, we may choose to introduce a second endomorphism ϕ' on E in order to define the ''off-diagonal'' of (87). This additional endomorphism is defined by (110), but with the Yukawa-coupling matrices $\mathbf{g}_y'^q, \mathbf{g}_y^q, \mathbf{g}_y^l$ replaced by arbitrary matrices $\mathbf{\Lambda}'^q, \mathbf{\Lambda}^q, \mathbf{\Lambda}^l$ of the corresponding size. This freedom arises from the fact that the form of the functional (88) does not change by this replacement and that the off-diagonal of $\widetilde{\mathrm{D}}_\phi$ does not act on the fermions Ψ, giving the matrices $\mathbf{g}_y$ their physical interpretation. [At least, when one restricts oneself to diagonal sections, $\widetilde{\Psi} = (\Psi, \Psi) \in \Gamma(\mathscr{E})$. The more general case of nondiagonal sections will be discussed in a future paper, when also further phenomenological consequences of our model and a physical interpretation of the doubling of the ''internal freedoms,'' $E \to \widetilde{E} = E \oplus E$, are considered (cf. our remarks at the end of this paper.)] Note that this freedom is crucial for the definition of the mass matrix of the gauge bosons. Moreover, without loss of generality we may assume that ϕ' is Hermitian, so that all the constants ''a'' are real [cf. (89)]. With these replacements the ''bosonic part'' of the universal functional (85) reads

$$\mathscr{I}_\mathrm{bosonic} = \mathscr{I}_\mathrm{EH} + \alpha_2 \int_\mathscr{M} \mathrm{tr_E}(zF^* \wedge {*F}) \tag{123}$$

$$+ \alpha_3 \int_\mathscr{M} \mathrm{tr_E}(z(\nabla\phi')^* \wedge {*}(\nabla\phi')) \tag{124}$$

$$-\alpha_4 \int_{\mathscr{M}} *\mathrm{tr}_{\mathrm{E}}(\mathrm{z}\phi^*\phi) \tag{125}$$

$$+\alpha_0 \int_{\mathscr{M}} *\mathrm{tr}_{\mathrm{E}}(\mathrm{z}(\phi'^*\phi')^2) \tag{126}$$

with the constants

$$\alpha_0 := \frac{3n(1-1/2n)}{2\pi\ \mathrm{tr}_{\mathrm{E}}\,\mathrm{z}} \left(\frac{l}{l_p}\right)^2 a_0^2, \tag{127}$$

$$\alpha_2 := \frac{3(2n-3)}{\pi\ \mathrm{tr}_{\mathrm{E}}\,\mathrm{z}} \left(\frac{l}{l_p}\right)^2 a_2^2, \tag{128}$$

$$\alpha_3 := \frac{3(n-1)}{2\pi\ \mathrm{tr}_{\mathrm{E}}\,\mathrm{z}} \left(\frac{l}{l_p}\right)^2 a_3^2, \tag{129}$$

$$\alpha_4 := \frac{3}{4\pi\ \mathrm{tr}_{\mathrm{E}}\,\mathrm{z}}\, m_p^2 a_4^2. \tag{130}$$

Here, z is considered as an element of $\Gamma_{\mathrm{eq}}(\mathscr{P},\mathrm{End}(V))$ that lies in the commutant (76) and satisfies

$$[\mathrm{z}, \chi^{\mathrm{E}}]=[\mathrm{z}, \phi]=[\mathrm{z}, \phi']=0, \tag{131}$$

as well as $\mathrm{z}>0$.

Since we use our units so that $c=\hbar=1$ we have identified Newtons gravitational constant G with the (square of the) ''Planck length'' $l_p \equiv m_p^{-1}$. Moreover, we already have normalized the Einstein–Hilbert functional $\mathscr{T}_{\mathrm{EH}}$ so that

$$\mathscr{T}_{\mathrm{EH}} = \frac{1}{16\pi l_p^2} \int_{\mathscr{M}} *r_{\mathscr{M}}. \tag{132}$$

The corresponding normalized fermionic action reads (again, when restricted to diagonal sections)

$$\mathscr{T}_{\mathrm{fermionic}} = \int_{\mathscr{M}} *(\Psi, iD\Psi)_\varepsilon \tag{133}$$

$$+a_4 \int_{\mathscr{M}} *(\Psi, i\Phi\Psi)_\varepsilon, \tag{134}$$

with Φ defined by (110) and D a SDO. Note that only the constant α_4 carries a dimension.

We now can compare the derived functionals (123)–(126) and (133)–(134) with the corresponding bosonic and fermionic action functionals of the standard model

$$\mathscr{T}_{\mathrm{EHYMH}} := \frac{1}{16\pi l_p^2} \int_{\mathscr{M}} *r_{\mathscr{M}} + \int_{\mathscr{M}} *\left\{\frac{1}{2}\,\mathrm{tr}\,(\mathbf{C}_{\mu\nu}\mathbf{C}^{\mu\nu}) + \frac{1}{2}\,\mathrm{tr}\,(\mathbf{W}_{\mu\nu}\mathbf{W}^{\mu\nu}) + \frac{1}{4}\,B_{\mu\nu}\,B^{\mu\nu}\right\} \tag{135}$$

$$+\int_{\mathscr{M}} *(\nabla_\mu\,\varphi)^*(\nabla^\mu\varphi) \tag{136}$$

$$+\int_{\mathscr{M}} *[\lambda(\varphi^*\varphi)^2 - \mu^2\varphi^*\varphi] \tag{137}$$

and

$$\mathcal{I}_{\mathrm{DY}}=\int_{\mathcal{M}}*(\Psi,i\gamma^{\mu}\nabla_{\mu}\,\Psi)_{\mathcal{E}} \tag{138}$$

$$+\int_{\mathcal{M}}*(\Psi,i\Phi\Psi)_{\mathcal{E}}. \tag{139}$$

Here, respectively,

$$\begin{aligned}\mathbf{C}_{\mu\nu}&:=\partial_{\mu}\mathbf{C}_{\nu}-\partial_{\nu}\mathbf{C}_{\mu}+ig_{(3)}[\mathbf{C}_{\mu},\,\mathbf{C}_{\nu}],\\ \mathbf{W}_{\mu\nu}&:=\partial_{\mu}\mathbf{W}_{\nu}-\partial_{\nu}\mathbf{W}_{\mu}+ig_{(2)}[\mathbf{W}_{\mu},\,\mathbf{W}_{\nu}],\\ B_{\mu\nu}&:=\partial_{\mu}B_{\nu}-\partial_{\nu}B_{\mu}\end{aligned} \tag{140}$$

denote the su(3), su(2), and u(1) valued curvatures with respect to a local coordinate system and in the fundamental representation. Moreover, the covariant derivatives, acting on the Higgs field φ and on the fermions Ψ, are locally defined as

$$\nabla_{\mu}\varphi:=\left(\partial_{\mu}+ig_{(2)}W_{\mu}^{b}\,\frac{\boldsymbol{\tau}_{b}}{2}+ig_{(1)}y_{h}B_{\mu}\right)\varphi, \tag{141}$$

$$\nabla_{\mu}\Psi:=(\partial_{\mu}+ig_{(3)}C_{\mu}^{a}\mathbf{1}_{\mathrm{S}}\otimes\mathbf{F}_{a}+ig_{(2)}W_{\mu}^{b}\mathbf{1}_{\mathrm{S}}\otimes\mathbf{T}_{b}+ig_{(1)}B_{\mu}\mathbf{1}_{\mathrm{S}}\otimes\mathbf{Y})\Psi, \tag{142}$$

where again $(g_{(3)},g_{(2)},g_{(1)})$ are the coupling constants and $\{i\mathbf{F}_{a},i\mathbf{T}_{b},i\mathbf{Y}\}_{1\leqslant b\leqslant 3}^{1\leqslant a\leqslant 8}$ denote the generators of $\rho(\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1))$; $\lambda,\mu^{2}>0$ are the positive real constants, parametrizing the classical vacuum. Note that the curvatures (140) are Hermitian.

Lemma 6: The derived functionals (123)–(126) and (133)–(134) are identical with the bosonic and fermionic action functionals of the standard model (135)–(140), iff the following relations hold:

$$\mathrm{N}\alpha_{2}g_{(1)}^{2}=\frac{1}{2(3y_{q}\lambda_{q}+y_{l}\lambda_{l})}, \tag{143}$$

$$\mathrm{N}\alpha_{2}g_{(2)}^{2}=\frac{1}{(3\lambda_{g}+\lambda_{l})}, \tag{144}$$

$$\mathrm{N}\alpha_{2}g_{(3)}^{2}=\frac{1}{4\lambda_{q}}, \tag{145}$$

and

$$2\alpha_{0}[3\lambda_{q}\,\mathrm{tr}\,(\boldsymbol{\Lambda}_{q}^{*}\boldsymbol{\Lambda}_{q})^{2}+\lambda_{l}\,\mathrm{tr}\,(\boldsymbol{\Lambda}_{l}^{*}\boldsymbol{\Lambda}_{l})^{2}]=\lambda, \tag{146}$$

$$2\alpha_{3}[3\lambda_{q}\,\mathrm{tr}\,\boldsymbol{\Lambda}_{q}^{*}\boldsymbol{\Lambda}_{q}+\lambda_{l}\,\mathrm{tr}\,\boldsymbol{\Lambda}_{l}^{*}\boldsymbol{\Lambda}_{l}]=1, \tag{147}$$

$$2\alpha_{4}[3\lambda_{q}\,\mathrm{tr}\,\mathbf{g}_{q}^{*}\mathbf{g}_{q}+\lambda_{l}\,\mathrm{tr}\,\mathbf{g}_{l}^{*}\mathbf{g}_{l}]=\mu^{2}, \tag{148}$$

$$a_{4}=1, \tag{149}$$

with the abbreviations

$$
\begin{aligned}
y_q &:= 2(y_{\mathrm{L}}^q)^2 + (y_{\mathrm{R}}^u)^2 + (y_{\mathrm{R}}^{d'})^2 \in \mathbb{R}_+ ,\\
y_l &:= 2(y_{\mathrm{L}}^l)^2 + (y_{\mathrm{R}}^l)^2 \in \mathbb{R}_+ ,\\
\mathbf{\Lambda}_q &:= (\mathbf{\Lambda}'^q, \mathbf{\Lambda}^q) \in \mathbf{M}_{2\mathrm{N}\times\mathrm{N}}(\mathbb{C}),\\
\mathbf{g}_q &:= (\mathbf{g}'^q, \mathbf{g}^q) \in \mathbf{M}_{2\mathrm{N}\times\mathrm{N}}(\mathbb{C}),\\
\mathbf{\Lambda}_l &:= \mathbf{\Lambda}^l \in \mathbf{M}_{\mathrm{N}}(\mathbb{C}),\\
\mathbf{g}_l &:= \mathbf{g}^l \in \mathbf{M}_{\mathrm{N}}(\mathbb{R}).
\end{aligned}
\tag{150}
$$

Proof: The proof of this lemma mainly consists in determining z of the commutant, yielding

$$
\mathrm{z} = \begin{pmatrix} \mathrm{z}_{\mathrm{L}} & 0 \\ 0 & \mathrm{z}_{\mathrm{R}} \end{pmatrix}, \tag{151}
$$

$$
\mathrm{z}_{\mathrm{L}} = \begin{pmatrix} \lambda_q \mathbf{1}_{6\mathrm{N}} & 0 \\ 0 & \lambda_l \mathbf{1}_{2\mathrm{N}} \end{pmatrix}, \tag{152}
$$

$$
\mathrm{z}_{\mathrm{R}} := \begin{pmatrix} \lambda_q \mathbf{1}_{6\mathrm{N}} & 0 \\ 0 & \lambda_l \mathbf{1}_{\mathrm{N}} \end{pmatrix}, \tag{153}
$$

with λ_q, $\lambda_l \in \mathbb{R}_+$. Then, rewriting the functionals (123)–(126) into the form (135)–(137) gives the desired relations. Clearly, the relation for a_4 simply follows by direct comparison of the corresponding fermionic functionals.

Remark 9: Note that we still have one more free parameter, α say, to introduce in our model kit and to write instead of (85):

$$
\mathscr{T}_{\widetilde{\mathrm{D}}} = (\psi, \widetilde{\mathrm{D}}\psi)_{\Gamma(\mathscr{E})} + \alpha \ \mathrm{res}_\zeta(\widetilde{\mathrm{D}}^{-2n+2}).
$$

Of course, this additional parameter indicates that in the definition of the universal action (2) the fermionic and the bosonic action functionals are considered as independent of each other. Alternatively, one may put $\alpha \equiv 1$ and then rescale both functionals independently (although we only have one overall constant!). This is what we did, actually, to obtain (132)–(134). Note that the right normalization of (124) crucially depends on the relations (112), which the hypercharges have to satisfy. Also note that the reason that each generation of quarks and each generation of leptons is equally weighted by the corresponding two constants λ_q and λ_l, respectively, is a consequence of the arbitrariness of the matrices $\mathbf{\Lambda}$ used in the definition of the endomorphism ϕ'.

The derived relations (143)–(149) can be related to physically measurable parameters. First, we have the following.

Lemma 7: The ''electroweak angle'' θ_{W} has the range

$$
0.25 \leqslant \sin^2\theta_{\mathrm{W}} \leqslant 0.45. \tag{154}
$$

Proof: By definition, the electroweak angle θ_{W} measures the portion of electromagnetism to weak force:

$$
\sin\theta_{\mathrm{W}} := \frac{\|\mathbf{T}_3\|}{\|\mathbf{Q}\|}. \tag{155}
$$

The norm $\|.\|$ used here is defined with respect to the ''su(N) normalization:'' $\kappa(\mathbf{E}_a, \mathbf{E}_b) := (1/2g_{\mathrm{N}}^2)\delta_{ab}$, where $\{\mathbf{E}_a\}_{1\leqslant a\leqslant \dim \mathrm{su}(\mathrm{N})}$ denotes an appropriate basis in the fundamental representation of su(N), $\mathrm{N} \geqslant 2$. To explain this more mathematically, we remind the reader of the fact that the general Killing form $\widetilde{\kappa}$ on a simple Lie algebra $\mathscr{G}$ may be written as $\widetilde{\kappa}(a,b) = (\lambda/g^2)(\rho'(a), \rho'(b)) \forall a, b \in \mathscr{G}$, where (.,.) denotes any ad-invariant scalar product on

$\rho'(\mathcal{G})\subset \mathrm{End}(V)$ with representation $\mathcal{G}\overset{\rho'}{\to}\mathrm{End}(V)$. Here, the constant λ depends on the scalar product and g is an arbitrarily positive constant, parametrizing the scalar product. This holds true also when $\mathcal{G}$ is semi-simple. However, the constant g—the ''coupling constant'' in physical terms—may be chosen differently for each simple component. Without loss of generality we can assume that $\lambda(\rho'(\mathbf{E}_a),\rho'(\mathbf{E}_b))=\delta_{ab}$ to obtain the well-known formula

$$\sin^2\theta_{\mathrm{W}}=\frac{g_{(1)}^2}{g_{(1)}^2+g_{(2)}^2} \tag{156}$$

for the electroweak angle. Hence,

$$\sin^2\theta_{\mathrm{W}}=\frac{\lambda_l+3\lambda_q}{(1+2y_l)\lambda_l+3(1+2y_q)\lambda_q}. \tag{157}$$

The range (154) then follows by the appropriate numerical values (113) and (114) of the hypercharges.

Remark 10: Since all norms proportional to each other give the same $\sin\theta_{\mathrm{W}}$ we may choose, alternatively, $(\rho'(a),\rho'(b)):=\mathrm{tr}(\mathrm{z}\rho'(a)^*,\rho'(b))$ to define the norm in (155). Hence,

$$\sin^2\theta_{\mathrm{W}}=\frac{(\mathbf{T}_3,\mathbf{T}_3)}{(\mathbf{Q},\mathbf{Q})}, \tag{158}$$

which, again, leads to (156). In this form, however, it becomes evident how the range (154) does depend on the generalized Dirac–Yukawa operator of the standard model, namely just by fixing the commutant.

The analog holds true for the ratio $g_{(2)}^2/g_{(3)}^2$ of the weak and strong coupling constants, yielding

$$g_{(2)}^2/g_{(3)}^2=\frac{4\lambda_q}{3\lambda_q+\lambda_l}. \tag{159}$$

Note that when we disregarded the possibility of introducing the element ζ in the definition of the bosonic action functional (74), the relations (157) and (159) were just a consequence of the fermionic representation ρ used in the derived Yang–Mills action. Obviously, the same holds true in the case of $\lambda_q=\lambda_l$, giving the ''GUT-preferred'' numerical values:

$$\begin{aligned}\sin^2\theta_{\mathrm{W}}&=\tfrac{3}{8},\\ g_{(3)}&=g_{(2)}.\end{aligned} \tag{160}$$

On the today energy scale, however, $\lambda_q\ll\lambda_l$ seems to be preferred (cf. Ref. 2). Hence, $g_{(2)}\ll g_{(3)}$, which might be expected, intuitively.

A more model-specific relation may be obtained concerning the ratio m_h/m_w of the ''Higgs mass'' and the ''W-boson mass'' of the electroweak interaction. By definition, having fixed a (classical) vacuum $\varphi_0\in\Sigma$, the masses of the gauge bosons are given by the quadratic form

$$\begin{aligned}\Omega^1(\mathcal{M},\rho_h'(\mathcal{G}))&\overset{\mathbf{M}^2}{\to}C^\infty(\mathcal{M})\\ \alpha&\mapsto \tfrac{1}{2}\,\mathbf{M}_{ab}^2*(\alpha^a\wedge *\alpha^b).\end{aligned} \tag{161}$$

with $\mathbf{M}_{ab}^2:=\varphi_0^*[\rho_h'(\mathbf{E}_a),\,\rho_h'(\mathbf{E}_b)]_+\varphi_0$ and $[.,.]_+$ the anti-commutator; $\{\mathbf{E}_a\}_{1\leqslant a\leqslant\dim\mathcal{G}}$ is a basis in the (semi-simple) Lie algebra $\mathcal{G}$ of the structure group G, so that $\alpha=\alpha^a\otimes\rho_h'(\mathbf{E}_a)$. In the unitary gauge: $\varphi_0=(0,v/\sqrt{2})^T$ we get $m_w^2=g_{(2)}^2v^2/4$.

The mass (matrix) of the Higgs field is defined with respect to the quadratic form:

$$\begin{aligned}\Gamma(E_h) &\xrightarrow{\mathbf{M}_h^2} C^\infty(\mathscr{M})\\ h &\mapsto \tfrac{1}{2}\,(h, V''|_{\varphi=\varphi_0} h)_{\mathrm{E}_h},\end{aligned} \tag{162}$$

with $h:=\varphi-\varphi_0$ and $(.,.)_{\mathrm{E}_h}$ the induced scalar product on the subbundle $\mathrm{E}_h\subset\mathrm{E}$. In the unitary gauge this yields $m_h^2=2\lambda v^2$.

Consequently, using (146) and (148) we end up with

$$\frac{m_h^2}{m_w^2}=\frac{4(2n-1)(2n-3)}{N\pi^2}\left(\frac{l}{l_p}\right)^4\frac{3\lambda_q+\lambda_l}{(4\lambda_q+\lambda_l)^2}\,(3\lambda_q\Lambda_q^4+\lambda_l\Lambda_l^4)(a_0a_2)^2, \tag{163}$$

with the abbreviation $\Lambda^4:=\mathrm{tr}\,(\mathbf{\Lambda}^*\mathbf{\Lambda})^2$. Some further investigations similar to those in Ref. 23 are needed in order to find out whether there are sufficiently relations between the unknowns on the right-hand side of (163) determining a range where the Higgs mass has to lie.

We finish this section by considering the special case of $\mathbf{\Lambda}\equiv\mathbf{g}$, yielding the following.

Lemma 8: Let $\phi'=-i\phi$ as defined by (110). Then the mass squared of the Higgs field and of the W-boson reads

$$m_h^2=\frac{2(2n-1)}{(n-1)}\left(\frac{a_0}{a_3}\right)^2\frac{3M_q^4+M_l^4}{3M_q^2+M_l^2}, \tag{164}$$

$$m_w^2=\frac{(n-1)}{2(2n-3)N}\left(\frac{a_3}{a_2}\right)^2\frac{3M_q^2+M_l^2}{3\lambda_q+\lambda_l}, \tag{165}$$

with $M_q^2:=\mathrm{tr}(\mathbf{\lambda}_q\mathbf{m}_q^*\mathbf{m}_q)$ *and* $M_l^2:=\mathrm{tr}(\mathbf{\lambda}_l\mathbf{m}_l^*\mathbf{m}_l)$. [Note, by abuse of notation $M^4\equiv\mathrm{tr}(\mathbf{\lambda}(\mathbf{m}^*\mathbf{m})^2)$.] Here, we used the abbreviation $\mathbf{m}_q^*\mathbf{m}_q\equiv\mathbf{m}_d^*\mathbf{m}_d+\mathbf{m}_u^*\mathbf{m}_u$ and $\mathbf{\lambda}_q:=\lambda_q\mathbf{1}_N$, $\mathbf{\lambda}_l=\mathrm{diag}(\lambda_{l_1},\ldots,\lambda_{l_N})\in\mathbf{M}_\mathrm{N}(\mathbb{R})$. Correspondingly, $\lambda_q:=\mathrm{tr}\,\mathbf{\lambda}_q/\mathrm{N}$, as before. However, $\lambda_l:=\mathrm{tr}\,\mathbf{\lambda}_l/\mathrm{N}$. Note that by the commutant all irreducible subspaces of the fermionic representation space V are now independently weighted. Because of the Kobayashi–Maskawa matrix the quark sector is considered as irreducible.

Proof: Since $m_h^2=2\lambda v^2$ we have with (146)

$$m_h^2=16\alpha_0\,\frac{3M_q^4+M_l^4}{v^2}; \tag{166}$$

The relation (147) then implies (164). With

$$l^2=\frac{1}{2(2n-1)}\left(\frac{1}{a_0}\right)^2\frac{3M_q^2+M_l^2}{3M_q^4+M_l^4}, \tag{167}$$

and

$$v^2=\frac{(n-1)}{\pi(2n-1)N}\left(\frac{a_3}{a_0}\right)^2\frac{(3M_q^2+M_l^2)^2}{3M_q^4+M_l^4}\,\frac{m_p^2}{4\lambda_q+\lambda_l}, \tag{168}$$

an analogous calculation yields (165).

Note that in the two geometrically distinguished cases where either the Higgs field defines a certain supercurvature or a certain differential form, the Higgs form ω_ϕ, as discussed in (94) and (95), the Higgs mass is determined by the mass of the fermions, and its range depends on the weights of all the irreducible subspaces of the fermionic representation. In this sense the parameter ''Higgs-mass'' in the standard model may be considered as *derived* from the generalized Dirac–Yukawa operator of the standard model. Clearly, the same holds for all other (possible) derivable relations between physical parameters. We stress that the commutant in the definition of the bosonic functional (85) is motivated mainly by the fact that absolute values of the physical

parameters, like the masses or charges of the particles involved, are of no physical significance. In this sense the generalization (74) becomes very plausible. Again, one has to take into account that the parameters defining the commutant are scale dependent in general.

IV. OUTLOOK

To summarize, in this paper we have introduced a particular geometrical model building kit based on the notion of generalized Dirac operators, considered as a triple $(\mathrm{G},\rho,\mathrm{D})$. Correspondingly, the geometrical setting is given by a Clifford module bundle $(\mathscr{E},c)$ over a (closed, compact) Riemannian (spin-) manifold $(\mathscr{M},g)$ of even dimension $2n>2$. Within this geometrical frame we proposed the universal functional

$$\mathscr{T}_{\mathrm{D}} := (\Psi, i\mathrm{D}\Psi)_{\Gamma(\mathscr{E})} + \mathrm{res}_{\zeta}(\mathrm{D}^{-2n+2}) \tag{169}$$

on $\Gamma(\mathscr{E})\times\mathscr{A}(\mathscr{E})$, $\mathrm{D}_{\nabla}=\mathrm{D}$, generalizing the classical action functional of the standard model. Indeed, we have shown how the action functional of the standard model—with the gravity action included—can be derived from a generalization of the Dirac–Yukawa operator. For this we have introduced a certain class of Dirac operators, parametrized by some constants. Two particular choices of these constants are distinguished geometrically. In particular, we have shown that the structure of the Yukawa coupling is such that it naturally defines a certain one-form $\omega_{\phi} := \delta_{\xi}\Phi$ on a twisted Clifford module bundle $\mathscr{E}$ (the Higgs form) and hence a particular connection thereof. Using this connection to define a certain Dirac operator and then evaluate the proposed functional (85) (with respect to diagonal sections) one obtains the full action of the standard model. Having derived the action we have shown how the free parameters of the model are linked to physical quantities like masses and charges of the various fields involved. Depending on the number of independent parameters of the model this may then lead to nontrivial relations between the physical quantities. For a special case we derived a relation between the mass of the Higgs field and the masses of the fermions.

The basic idea is that the fermionic interactions determine the dynamics of *all* the fields involved in the theory. In the case of the standard model the fermionic interaction is defined with respect to the Yukawa coupling (besides the gauge covariant coupling), giving rise to a nonstandard Dirac operator—the Dirac–Yukawa operator. Here, ''nonstandard'' means that the appropriate class of connections, defining the Dirac operator in question, does not contain a Clifford connection as a representative and thus indicates that the basic geometrical setting is that of a Clifford module bundle instead of the particular case of a twisted spinor bundle. Indeed, according to the Higgs form, the tensor product structure of the Clifford module is ignored. In contrast to mathematical applications, where mainly SDOs are of interest, in physics non-SDOs seem to play a dominant role. Let us assume that all particles are massless. In this case the fermionic interaction is described by a gauge potential. Accordingly, the corresponding Dirac operator D is a SDO. However, in our scheme one has to consider not this operator but, instead, $\widetilde{\mathrm{D}} := D + \mathscr{J}(c(F))$, where F is the relative curvature (in this case the twisting curvature) on the Clifford module bundle $\mathscr{E}$. Clearly, this new Dirac operator $\widetilde{\mathrm{D}}$ on $\widetilde{\mathscr{E}}$ is nonstandard. Evaluation of the functional (85) with respect to this operator leads to the Einstein–Hilbert–Yang–Mills action and hence the description of the full dynamics of the physical system under consideration (cf. Ref. 11) therefore, although in the case where the dynamics of the fermions are described by a SDO it seems to be more natural, actually, to consider the non-SDO $\widetilde{\mathrm{D}}$, describing the full dynamics. As a consequence of this scheme, however, one necessarily has to double the ''internal degrees of freedom'' of the fermions: $\mathrm{E}\to\widetilde{\mathrm{E}} := \mathrm{E}\oplus\mathrm{E}$. Moreover, both parts have to carry the same fermionic representation. Note that this doubling may be described, geometrically, as the pullback bundle $\Delta^*(\mathrm{E}\times\mathrm{E})$ of $\mathrm{E}\times\mathrm{E}\to\mathscr{M}\times\mathscr{M}$ with respect to the diagonal map $\mathscr{M}\ni x\overset{\Delta}{\to}(x,x)\in\mathscr{M}\times\mathscr{M}$. The latter may be identified with $\mathscr{M}\times\{\pm 1\}$ and thus, implicitly, also a doubling of space time is involved in our scheme. Since this construction is fundamental we are left to interpret this doubling physically (cf. discussion to follow).

With respect to the standard model, our model building kit somehow parallels the Connes–Lott approach. In fact, both approaches yield analogous results concerning the relations between the various parameters of the models. It therefore might be worth comparing both approaches,

though the general mathematical frame work is quite different and there is no doubt that, from a mathematical point of view, Connes' noncommutative geometry is much more indepth (cf. Refs. 4, 24, and 25 for a good review). Obviously, both kits have in common that the basic building block are Dirac operators. In the Connes–Lott scheme it is the internal Dirac operator $\mathscr{D}$ and in the kit proposed here it is the Dirac–Yukawa operator D_ϕ. As already mentioned in the Introduction, the basic ideas of our model and, especially, the approach of considering the Yukawa coupling as the fundamental input to derive the bosonic action of the standard model with gravity included was already proposed in Ref. 12. [By use of the correct Yukawa coupling (9), the generalized Dirac–Yukawa operator in Ref. 12 is of the form (95) of the paper at hand, and thus gives rise to the EHYMH action.] Indeed, also in the Chamseddine–Connes scheme the Dirac–Yukawa operator is considered as the main ingredient (cf. Refs. 6 and 8). In this scheme the universal action is defined by the heat trace in contrast to the original Connes–Lott description of the standard model, where the action was defined via the Dixmier trace (cf. the corresponding remarks in the Introduction). Concerning the standard model, in the Chamseddine–Connes approach one has to consider all Seeley–deWitt coefficients at least up to order four in the asymptotic expansion of the heat trace since the heat kernel is defined with respect to the Dirac–Yukawa operator D_ϕ. In our scheme, however, the generalized Dirac–Yukawa operator $\widetilde{\mathrm{D}}_\phi$ is considered as the basic Dirac operator. Hence, the full action of the standard model is given by the subleading term of the asymptotic expansion of the corresponding heat trace. Note that the latter yields exactly the action of the standard model, which is not the case in the Chamseddine–Connes scheme. Since in this scheme the derived action obtained by the Dirac–Yukawa operator is more general than the action of the standard model (it also contains the well-known *quadratic* terms in the curvature of the base manifold), one obtains more constraints for the parameters involved in the model (cf. Ref. 8). However, the relations (112) for the weak hypercharges and the relations (157) and (159) for the electroweak angle and the coupling constants, respectively, are merely a consequence of the Yukawa coupling (9) and that all fields involved in the model carry the fermionic representation. Hence, in this respect, the Chamseddine–Connes scheme and our kit yield similar results.

The power ''$-2n+2$'' in the definition of the action $\mathscr{S}_\mathrm{D}$ proposed here is motivated by the fact that there exists a generalized Dirac operator (in our sense) that gives back the *exact* action of the standard model. In fact, σ_2 is the only coefficient in the asymptotic expansion of the heat trace which is linear in the curvature of the base manifold. Also by formal analogy between the asymptotic expansion of the heat trace and the asymptotic expansion of the effective action in quantum field theory one may expect, intuitively, that the ''classical action'' is covered by the subleading term, only. As already mentioned, while the bosonic functional (68) was already considered in Ref. 10, and in Ref. 18 in the case of the pure gravity action, and in Ref. 17 from a somewhat more general perspective, the possibility to also derive the Yang–Mills and Higgs action from (68) was not taken into account.

If the former is considered as ''facts,'' we now turn to some ''fictions.'' As we have already mentioned, the price we have to pay for ''adding even terms'' to Dirac operators and thereby changing the emphasis from SDOs to non-SDOs is the additional structure $(\mathscr{E},\mathscr{J})$, which we have to introduce in our model. Of course, this needs some physical interpretation. A physically satisfying way to understand this additional structure may consist of interpreting the doubling of the internal degrees of fermionic freedoms by introducing the notion of ''antiparticles'' in our scheme. Hence, we have to incorporate the notion of ''charge conjungation'' within our model, which will be done in a forthcoming paper where we shall also investigate in more detail the relation (163).

Another point that we have in mind concerns spectral geometry. Since the whole dynamics of the fields involved in a physical theory should be determined by a single (generalized) Dirac operator, and, moreover, the corresponding fermionic action is proposed—naturally enough—to take its well-known form, one may ask whether the bosonic action actually can be considered as ''generalizing'' the fermionic action. Indeed, it is well known by physicists that, in a sense, the bosonic action can be recovered from the fermionic action by considering the former as a ''one-loop'' correction of the latter. For this, one has to introduce a zeta-regularized determinant of a certain operator. This may be the point where spectral geometry comes in. In other words, if the bosonic action is considered as a modification of the fermionic action, like above, we are left with the mathematical question of how the proposed bosonic action in our model can be expressed as a zeta-regularized determinant of a Dirac operator. Moreover, as we have seen, from a geometrical

point of view the action functional of the standard model is but the subleading term of the asymptotic expansion of the heat trace of a certain Hamiltonian. Therefore, it might be natural to ask whether the higher terms in the expansion permit a physical interpretation as well. Again, so far this is but fiction. However, it might be worth investigating these points more carefully in a future paper.

Still quite another point, of course, is concerned with the fact that we are dealing with Riemannian manifolds instead of Lorentzian manifolds. Hence, the notion of gravity is just formal. However, our model may be flexible enough to work also in the case when Lorentzian manifolds are considered. The point here is that also the kernels of differential operators of ''Huygens-type'' have an asymptotic expansion like the heat trace of an elliptic operator (cf. Ref. 26). Of course, spectral geometry in this case is not so well established, but see, e.g., Ref. 27 and the corresponding references therein. Finally, since the basic geometrical setting of our model kit is a Clifford module bundle and there is a one-to-one correspondence between Dirac operators and Clifford superconnections, it might be possible to incorporate the notion of ''supersymmetry'' in our scheme. Thus, one may put more emphasis on the ''superformalism,'' as developed by Quillen *et al.* (cf. Refs. 5, 28, and 29), in the case of physics. Concerning our model, this point of view was taken, e.g., in Ref. 30.

We finish this paper with a citation, which expresses the feeling of many physicists. We do, however, hope to have convinced the reader that in fact the contrary holds true:

> ''... this prescription [via the Yukawa-coupling] of the fermion masses is one of the least satisfactory aspects of the theory [standard model]. It is an entirely *ad hoc* procedure... .'' [31]

ACKNOWLEDGMENTS

I would like to thank Th. Schücker for explaining to me the beauty of noncommutative geometry and its application to physics. However, I would like to thank him much more for his ''stubbornness.'' Also, I would like to thank Th. Ackermann for many discussions on this subject, yielding three joined works, and E. Binz for his encouragements. Finally, I would like to thank M. Lesch for very helpful discussions and for explaining the calculation of Eq. (80) to me.

[1] B. Iochum and Th. Schücker, ''Yang–Mills versus Connes–Lott,'' Commun. Math. Phys. **178**, 1–26 (1996), hep-th/9501142.
[2] O. Nachtmann, *Elementarteilchen Physik, Phänomene und Konzepte* (Vieweg, 1986).
[3] A. Connes and J. Lott, ''The metric aspect of non-commutative geometry,'' in *The Proceedings of the 1991 Cargese Summer Conference*, edited by J. Fröhlich (Plenum, New York, 1992).
[4] A. Connes, *Non-commutative Geometry* (Academic, New York, 1994).
[5] N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators* (Springer, New York, 1992).
[6] A. H. Chamseddine and A. Connes, ''The spectral action principle,'' Commun. Math. Phys. **186**, 731–750 (1997), hep-th/9606001.
[7] A. H. Chamseddine and A. Connes, ''A universal action formula,'' hep-th/9606056.
[8] B. Iochum, D. Kastler, and Th. Schücker, ''On the uiversal Chamseddine-Connes action; 1. Details of the action computation,'' hep-th/9607158.
[9] A. Chamseddine, J. Fröhlich, and O. Grandjean, ''The gravitational sector in the Connes-Lott formulation of the standard model,'' J. Math. Phys. **36**, 6255 (1995), hep-th/9503093.
[10] A. Connes, ''Noncommutative geometry and physics,'' in *The Proceeding of the 1992 Les Houches Summer School*, edited by B. Julia and J. Zinn-Justin (North–Holland, Amsterdam, 1995).
[11] Th. Ackermann and J. Tolksdorf, ''The generalized Lichnerowicz formula and analysis of Dirac operators,'' J. Reine Angew. Math. **471**, 23–42 (1996).
[12] Th. Ackermann and J. Tolksdorf, ''Unification of gravity and Yang-Mills-Higgs gauge theories,'' Preprint CPT-95/P.3180, Mannheimer Manuskripte 191 (1995), hep-th/9503180.
[13] J. M. Bismut, ''A local index theorem for non Kähler manifolds,'' Math. Ann. **284**, 681–699 (1989).
[14] M. Wodzicki, ''Local invariants of spectral asymmetry,'' Invent. Math. **75**, 143–178 (1984).
[15] M. Wodzicki, ''Non-commutative residue 1, '' Lect. Notes Math. **1289**, 320–399 (1987).
[16] V. Guillemin, ''A new proof of Weyl's formula on the asymptotic distribution of eigenvalues,'' Adv. Math. **55**, 131–160 (1985).
[17] W. Kalau and M. Walze, ''Gravity, non-commutative geometry and the Wodzicki residue,'' J. Geom. Phys. **16**, 327–344 (1995).
[18] D. Kastler, ''The Dirac operator and gravitation,'' Commun. Math. Phys. **166**, 633–643 (1995).
[19] Th. Ackermann and J. Tolksdorf, ''A generalized Lichnerowicz formula, the Wodzicki residue and Gravity,'' J. Geom. Phys. **19**, 143–150 (1996).

[20] P. B. Gilkey, *Invariance Theory, the Heat Equation and the Atiyah-Singer Theorem*, 2nd ed. (CRC, Boca Raton, FL, 1995).
[21] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry; Part 1* (Wiley, New York, 1963).
[22] Y. Choquet-Bruhat and C. DeWitt-Morette, *Analysis, Manifolds and Physics; Part 1* (North Holland, Amsterdam, 1989).
[23] B. Iochum, D. Kastler, and Th. Schücker, ''Fuzzy mass relations in the standard model,'' J. Math. Phys. **36**, 6232–6254 (1995), CPT-95/P.3235, hep-th/9507150.
[24] D. Kastler and M. Mebkhout, *Lectures on Non-Commutative Differential Geometry* (World Scientific, to be published).
[25] J. C. Varilly and J. M. Gracia-Bondia, ''Connes' noncommutative differential geometry and the Standard Model,'' J. Geom. Phys. **12**, 223–301 (1993).
[26] H. Baum, ''The Dirac operator on Lorentzian spin manifolds and the Huygens property,'' SFB 288 Preprint No. 214, 1996.
[27] H. Baum and I. Kath, ''Normally hyperbolic operators, the Huygens principle and conformal geometry,'' Ann. Global Anal. Geom. **14**(4), 315–371 (1996).
[28] Y. Ne'eman and S. Sternberg, *Internal Supersymmetry and Superconnections*, Sympl. Geom. and Math. Physics, edited by P. Donato, C. Durce, J. Elhadan, and G. M. Tuymmann (Birkhäuser, 1991).
[29] R. Häussling, N. A. Papadopoulos, and F. Scheck, ''SU(2|1) symmetry, algebraic superconnections and a generalized theory of electroweak interactions,'' Phys. Lett. B **260(1)**, 2 (1991).
[30] Th. Ackermann, ''Dirac operators and Clifford geometry—New unifying principles in particle physics?,'' hep-th/9605129.
[31] I. J. R. Aitchison and A. J. Hey, *Gauge Theories in Particle Physics* (Hilger, Bristol, 1982), p. 260.

Yang–Mills flow and uniformization theorems

S. P. Braham
Center for Experimental and Constructive Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada

J. Gegenberg
Department of Mathematics and Statistics, University of New Brunswick, Fredericton, New Brunswick E3B 4A3 Canada

(Received 4 August 1997; accepted for publication 29 September 1997)

We consider a paraboliclike system of differential equations involving geometrical quantities to examine uniformization theorems for two- and three-dimensional closed orientable manifolds. We find that in the two-dimensional case there is a simple gauge theoretic flow for a connection built from a Riemannian structure, and that the convergence of the flow to the fixed points is consistent with the Poincare Uniformization Theorem. We construct a similar system for the three-dimensional case. Here the connection is built from a Riemannian geometry, an SO(3) connection and two other 1-form fields which take their values in the SO(3) algebra. The flat connections include the eight homogeneous geometries relevant to the three-dimensional uniformization theorem conjectured by Thurston. The fixed points of the flow include, besides the flat connections (and their local deformations), nonflat solutions of the Yang–Mills equations. These latter ''instanton'' configurations may be relevant to the fact that generic 3-manifolds do not admit one of the homogeneous geometries, but may be decomposed into ''simple 3-manifolds'' which do. © *1998 American Institute of Physics.* [S0022-2488(98)01903-3]

I. INTRODUCTION

The uniformization theorem in two dimensions is a powerful tool in geometry and topology, with applications in physics. In essence, the theorem states that the topology of a closed orientable two-dimensional manifold (a Riemann surface) determines which geometries it admits. In particular, if the manifold has handle number zero, it admits the spherical geometry and its local deformations; handle number one admits the flat geometry and its local deformations; and for handle number two or greater, the admissible geometry is that of the hyperbolic plane and its local deformations. It is important that one cannot deform any of these three geometries to obtain one of the other two.

This theorem was proved finally around the beginning of the twentieth century by Poincare.[1] It is a heroic proof, using the most sophisticated mathematics of the day. Unfortunately, the classical proof depended heavily on results in complex analysis, and the generalization to three or higher dimensions was not obvious. In fact, the three-dimensional analog of this theorem was first consistently formulated only in the late 1970's by Thurston and is called Thurston's Geometrization Conjecture.[2] To date, although no counterexamples have emerged and it has been shown to hold for very large classes of manifolds, the conjecture remains unproved.

Recently, Hamilton and Chow have constructed a new proof of the two-dimensional uniformization theorem using techniques not obviously restricted to that number of dimensions.[3,4] They consider a one parameter family of Riemannian metrics $g_{\mu\nu}$ on an n-dimensional smooth manifold M_n, with the ''flow'' governed by the Ricci curvature tensor $R_{\mu\nu}$:

$$\frac{\partial g_{\mu\nu}}{\partial t} = -2R_{\mu\nu} + \frac{2}{n} r g_{\mu\nu}. \tag{1}$$

In the above, r is given by

 © 1998 American Institute of Physics

TABLE I. The eight homogeneous geometries.

Manifold	Isometry group	Metric
S^3	SO(4)	$\cos^2 y\,dx^2+dy^2+(dz-\sin y\,dx)^2$
E^3	$R^3\times$SO(3)	$dx^2+dy^2+dz^2$
H^3	PSL(2,C)	$dx^2+e^{2x}(dy^2+dz^2)$
$S^2\times E^1$	$(\mathrm{Isom}(S^2)\times\mathrm{Isom}(E^1))^+$	$dx^2+dy^2+\sin^2 y\,dz^2$
$H^2\times E^1$	$(\mathrm{Isom}(H^2)\times\mathrm{Isom}(E^2))^+$	$dx^2+dy^2+e^{2x}dz^2$
$\widetilde{\mathrm{SL}(2,R)}$	$\mathrm{Isom}(H^2)\times R$	$\cosh^2 y\,dx^2+dy^2+(dz+\sinh y\,dx)^2$
Nil	$\mathrm{Isom}(E^2)\times R$	$dx^2+dy^2+(dz-x\,dy)^2$
Sol	$\mathrm{Sol}\times(Z_2)^2$	$dx^2+e^{-2x}\,dy^2+e^{2x}\,dz^2$

$$r:=\left(\int_{M_n} d^n x\sqrt{g}\right)^{-1}\int_{M_n} d^n x\sqrt{g}R, \tag{2}$$

where $d^n x\sqrt{g}$, with $g=\det(g_{\mu\nu})$, is the volume element on M_n. It is easy to show that (i) $\partial r/\partial t=0$ along the flow; and (ii) the fixed points of the flow are the ''Einstein metrics'', satisfying $R_{\mu\nu}=(r/n)g_{\mu\nu}$. In two dimensions, the Einstein metrics have constant curvature, and include all homogeneous geometries; in three dimensions, the Einstein metrics include the constant curvature, but not all the homogeneous geometries; and finally, in four and higher dimensions, the Einstein metrics do not have nearly so clear a geometric significance as they do in lower dimensions.

The Hamilton/Chow proof of the uniformization theorem in two dimensions analyzes the ''Ricci scalar flow''

$$\frac{\partial R}{\partial t}=\Delta R+R(R-r), \tag{3}$$

derived from the flow of the metric above. For the cases where R is nonpositive, it is fairly straightforward to show that the flow converges to the fixed points with nonpositive constant curvature.[3] It took a few more years to analyze the case of $R>0$, since this involves a repulsive fixed point.[4]

Isenberg and Jackson[5] examined the Ricci flow in three dimensions in order to shed light on Thurston's Geometrization Conjecture.[2] Unlike the two-dimensional case, it is not true in three dimensions that all closed orientable manifolds admit one of the constant curvature geometries. Rather, one must first canonically decompose the given manifold M into its prime pieces,[6] obtained by cutting along 2-spheres and gluing 3-balls onto the cuts until one or both pieces is homeomorphic to the 3-ball; then cutting along incompressible T^2 until the pieces are either Seifert fiber spaces or contain no embedded incompressible T^2. Denoting the resulting manifolds by M_i, so that $M=M_1\#M_2\#\ldots$, Thurston then conjectures that the universal covers $\widetilde{M}_i$ of each M_i admits one and only one of the *eight* homogeneous geometries possible in three dimensions. See Table I for a list of these eight geometries.

The problem with the approach of Isenberg and Jackson is that on the one hand, there does not seem to be sufficient structure in the Ricci flow to cope with the necessity of decomposing the given manifold as above; and on the other, the fixed points are the constant curvature geometries, a proper subset of the homogeneous geometries. Hence, one must first cut and paste the manifold, then flow the geometry on each piece; and one must look in general for asymptotics, rather than convergence to fixed points.

What we propose here is a one-parameter family of *connections* whose flow converges to flat connections. In two dimensions, these flat connections are equivalent to the constant curvature geometries. This suggests the possibility of another proof of the two-dimensional uniformization theorem. What is important here is that this flow generalizes to three dimensions such that the fixed points of the flow are the eight homogeneous geometries plus certain ''instanton'' configurations, which describe ''necks'' between three-manifolds. We believe that this is a promising approach to proving Thurston's Geometrization Conjecture.

In Sec. II, motivated by the gauge-theoretic formulation of certain two-dimensional gravity theories, we will recast the structure equations for a constant curvature Riemannian *metric* into the form of a flatness condition on an appropriate *connection*. This sets the stage for constructing a one-parameter family of connections—the Yang–Mills flow. In Sec. III, the properties of the Yang–Mills flow will be explicated. We will show that the fixed points of the flow are Yang–Mills connections, a subclass of which is the set of flat connections, and that the latter describe Riemannian metrics of constant curvature. In Sec. IV, we analyze the behavior of the flow. For the cases of spherical and flat Euclidean topologies, we may conclude that the flow converges to the fixed points. We present arguments to show that with initial conditions consistent with regular torsion-free Riemannian geometries, the flow actually converges to the fixed points associated with flat connections, and hence to constant curvature Riemannian geometries. This leads us to believe that one may prove the two-dimensional uniformization theorem using the Yang–Mills flow. In Sec. V, we outline a numerical integration of the Yang–Mills flow. Finally, in Sec. VI, we will propose a generalization of the Yang–Mills flow to three dimensions. We will conclude by sketching the outline of the form that a proof of the Thurston Geometrization Conjecture, using this three-dimensional flow, might take.

II. GAUGE THEORY FORM OF 2D GRAVITY

One of the most fruitful areas of research in fundamental physics at the moment begins with the reformulation of Einstein's theory of gravity—the general theory of relativity—as a *gauge theory* of the (complexified) rotation group SO(3). This approach, pioneered by Ashtekar and his collaborators,[7] has resulted in some progress in constructing a quantum theory of gravity. This is in part due to the fact that the constraints in the theory become polynomial when expressed in terms of the above connection and its conjugate momentum; and this in turn has allowed for the construction of solutions of the constraints. Even more striking results have been attained by reformulating lower dimensional gravity theories as gauge theories. In particular, three-dimensional general relativity can be formulated as a Chern–Simons gauge theory,[8] and some simple two-dimensional gravity theories can be expressed as topological field theories of the so-called BF type.[9,10]

This suggests that it might be useful to examine the flow of connections constructed from the Riemannian geometry, rather than the flow of the metrics themselves. For one thing, as in the above theories of gravity, when expressed in connection form, we expect that the partial differential equations that describe the flow will be polynomial, unlike the Ricci flow, where the terms involving the curvature tensor are nonpolynomial in the metric. More importantly, we note that the connection formulation of lower dimensional gravity theories are topological field theories, and provide a more direct route to the global issues that must be addressed.

The fixed points of the Ricci flow are the Einstein spaces. In two dimensions these are just the spaces which have constant curvature Riemannian metrics: $R=2k$. About ten years ago, Jackiw and Teitelboim (separately) considered the above as a toy model of gravity in two dimensions.[11] A few years later, at least three groups constructed a gauge theory formulation.[9] In this formulation, the field equations of the theory required that a certain connection over spacetime is flat; this in turn was equivalent to the existence of a constant curvature (pseudo-) Riemannian geometry on the spacetime. In the following, we will construct the connection which has this property.

We now consider Riemannian geometry in the first-order Cartan formalism. Instead of the metric tensor $g_{\mu\nu}$ we consider a frame-field e^a, a set of two 1-form fields on M^2. The indices $a,b,\ldots=1,2$. The metric and frame-fields are related by $g_{\mu\nu}=\delta_{ab}e^a_\mu e^b_\nu$. Instead of the Christoffel symbols, we have the spin-connection $\omega_b{}^a$, also a set of 1-form fields on M^2. The spin-connection is skew-symmetric in the indices a,b, so in two dimensions, there is only one algebraically independent component, which we denote simply by ω, defined by $\omega_b{}^a=-\omega\epsilon_b{}^a$.

We can now define a connection 1-form field A:

$$A:=e^aP_a+\omega J, \tag{4}$$

where $\{P_a,J\}$ generate the Lie algebra:

$$[P_a,P_b]=k\epsilon_{ab}J;[J,P_a]=\epsilon_{ab}\delta^{bc}P_c, \tag{5}$$

with δ^{ab} the Euclidean metric. The algebra generated by $\{P_a, J\}$ is so(3) if $k=+1$, iso(2) (the ''Poincare algebra'') if $k=0$, or so(2,1) if $k=-1$. If the torsion

$$T^a := de^a - \epsilon^a{}_b \omega \wedge e^b = 0, \tag{6}$$

then the connection A determines a Riemannian geometry.

The curvature 2-form corresponding to the connection A is given by

$$F(A) := dA + \tfrac{1}{2}[A,A], \tag{7}$$

$$= T^a P_a + \left(d\omega + \frac{k}{2} \epsilon_{ab} e^a \wedge e^b \right) J, \tag{8}$$

$$= T^a P_a - \tfrac{1}{2}(R-2k) v J. \tag{9}$$

In the above, $v = \frac{1}{2}\epsilon_{ab} e^a \wedge e^b$ is the volume element on the manifold M induced by the Riemannian metric.

If the connection A is flat, then the curvature $F=0$, and hence

$$T^a = 0; \tag{10}$$

$$R = 2k. \tag{11}$$

Hence, a flat connection A, with algebra given by k, determines a Riemannian geometry with constant curvature $2k$, and vice-versa.

III. YANG–MILLS FLOW

In this section, we will describe a one-parameter family of connections, of the form given in the previous section.

We start with a two-dimensional manifold M and an admissible Riemannian metric $\bar{g}_{\mu\nu}$ and a (not necessarily compatible) spin-connection $\bar{\omega}$. We use this structure to define the *duals* of form fields, e.g.,

$$*(b_\mu dx^\mu) := \bar{g}^{1/2} \epsilon_{\mu\nu} \bar{g}^{\nu\sigma} b_\sigma \, dx^\mu, \tag{12}$$

where $\bar{g}$ is the determinant of $\bar{g}_{\mu\nu}$. The algebra Eq. (5) is characterized by the constant k. This is determined from the topological structure of M by the Euler number $\chi(M)$

$$k := \frac{\chi(M)}{|\chi(M)|}, \tag{13}$$

if $\chi(M) \neq 0$, and by

$$k = 0, \tag{14}$$

if $\chi(M) = 0$. In fact, $\chi(M)$ can be computed from $\bar{g}_{\mu\nu}$ by

$$\chi(M) = \frac{1}{4\pi} \int_M d^2x \sqrt{\bar{g}} \bar{g}^{\mu\nu} \bar{R}_{\mu\nu}, \tag{15}$$

where $\bar{R}_{\mu\nu}$ is the Ricci tensor of $\bar{g}_{\mu\nu}$. The Euler number $\chi(M)$ is related to the handle number, or genus $h(M)$ of M by $\chi(M) = 2-2h(M)$.

We now flow the connection $A(t)$ given by

$$A(t) = e^a(t) P_a + \omega(t) J. \tag{16}$$

The initial values are given by

$$A(0)=\overline{e^a}P_a+\overline{\omega}J. \tag{17}$$

The differential equations that determine the flow are

$$\frac{\partial A}{\partial t}=-*D_A*F(A). \tag{18}$$

The dual $*$ and the Lie algebra are determined by the initial fields, as discussed above. Hence $*$ and $\partial/\partial t$ commute.

The paraboliclike structure of the flow is displayed most transparently in the equations for the flow of the *curvature*. Using Eq. (18) and the duality in two dimensions of 2-forms with 0-forms, we arrive at

$$\frac{\partial f}{\partial t}=\Delta_A f, \tag{19}$$

where $f:=*F(A)$ is a Lie algebra valued 0-form equivalent to the curvature; and $\Delta_A:=-(*D_A*D_A+D_A*D_A*)$ is the Laplacian with respect to the connection $A(t)$.

The fixed points of the flow are the Yang–Mills connections A_{ym}, i.e., connections which satisfy the Yang–Mills equations:

$$-*D_{A_{ym}}*F(A_{ym})=0. \tag{20}$$

There are two types of Yang–Mills connections. The first are flat connections, i.e., $F(A)=0$. If the e^a are nondegenerate, these connections are equivalent to the constant curvature Riemannian geometries, as we discussed in the last section. The second type of Yang–Mills connections are ''instantons,'' with $F(A)\neq 0$. In this case, the structure group of the connection is reduced to a subgroup which commutes with $f:=*F$.[12]

In order to prove a two-dimensional uniformization theorem, we must establish the following.

Conjecture: From an initial connection corresponding to a sufficiently smooth nondegenerate Riemannian geometry on a 2-manifold M with Euler number $\chi(M)$, the Yang–Mills flow converges to the flat connection corresponding to a Riemannian geometry with constant curvature having the same sign as $\chi(M)$.

In the following two sections we will provide analytical and numerical evidence for this conjecture.

IV. CONVERGENCE OF THE YANG–MILLS FLOW: ANALYTICAL EVIDENCE

There has been some discussion of the properties of Yang–Mills (and related) flows by mathematicians and mathematical physicists.[13] It is clear from this literature, in particular from the thesis of Rade, that the flow exists and is unique for short times, at least for the case of nonnegative Euler number. Unfortunately, the situation with regard to the question of the convergence of flow as $t\to\infty$ is not clear at the moment. In the following, we will discuss the question of convergence from an analytic (but fairly heuristic) perspective. In the next section, encouraging results from numerical treatment will be presented.

The Yang–Mills flow resembles the heat equation

$$\frac{\partial \phi}{\partial t}=\Delta \phi, \tag{21}$$

where Δ is the Laplacian operator with respect to some Riemannian structure defined on the manifold M upon which the field ϕ takes its values. The existence/uniqueness for short times and the convergence as $t\to\infty$ to the ''average'' initial data is well-known.[14] Indeed, it is easy to see that for initial data infinitesimally close to a fixed point, the Yang–Mills flow is parabolic. In general, the Yang–Mills flow is polynomially nonlinear.

We will now discuss the question of convergence for each of the cases $k=0$ and $k=+1$. The $k=-1$ case is the least well-understood at the moment, and is under investigation by the authors.

For the case $k=0$ the Lie group is ISO(2), which is a semidirect product of the Abelian group SO(2) with the two-parameter group of translations. The Yang–Mills flow itself splits into an SO(2) piece which depends only on the A^2-component of the ISO(2) connection:

$$\frac{\partial A^2}{\partial t} = -*d*dA^2. \tag{22}$$

Although this is not strictly parabolic, the flow of the corresponding dual of the curvature component, $f^2 := *F^2 = *dA^2$ is a parabolic system:

$$\frac{\partial f^2}{\partial t} = \overline{\Delta} f^2, \tag{23}$$

where $\overline{\Delta}$ is the Laplacian with respect to the initial Riemannian geometry. Now the average of the initial curvature component,

$$\overline{F^2} := \int_{M^2} dA^2 = 0, \tag{24}$$

where the last equality follows from the fact the in this case M^2 has Euler number zero. Since f^2 converges to a constant 0-form, it must converge to zero everywhere on M_2.

What we have shown is that the manifold which is topologically T^2, i.e., which has Euler number 0, admits a closed 1-form. This is a sufficient condition that the manifold admits a Riemanian geometry with a compatible spin-connection with zero curvature. To see this, consider the following: Let ω be a closed 1-form. But since for T^2 the space of harmonic 1-forms is two-dimensional, there is a harmonic 1-form β such that $\omega \wedge \beta \neq 0$. We can find a chart in which there is a function $p(x)$, such that $\partial_\mu p(x) = \beta_\mu(x)$. Now define the 1-forms e^0, e^1 in this chart by $\epsilon^0(x) = p(x)\omega(x)$, $e^1(x) = -dp(x)$. In the chart, the volume element $e^0(x) \wedge e^1(x) \neq 0$ by construction, and the compatibility condition $de^a - \epsilon^a_b \omega \wedge e^b = 0$ since $d\omega = 0$.

It remains to address the question of which initial conditions, if any, determine flows which converge to the instantons, with $F(A_\infty) \neq 0$.

In the $k=+1$ case, for which the gauge group is SO(3), we have the results of Rade[13] wherein it is proved that for compact simple gauge groups, e.g., SO(3), the Yang–Mills flow converges with respect to the Sobolev norm H^1 to a Yang–Mills connection. What remains here is the question of the instantons, as well as whether the flow converges under stronger smoothness requirements. (It is easy to show that, analytically, a round sphere with arbitrary radius will exponentially converge to the sphere with ''correct'' radius $\sqrt{2}$, i.e., so that $R=2$.)

V. CONVERGENCE OF THE YANG–MILLS FLOW: EXPERIMENTAL DIFFERENTIAL GEOMETRY

The system of coupled partial differential equations Eq. (18) is comprised of polynomially *nonlinear* PDEs, and is therefore potentially quite complicated. However, the right-hand side of Eq. (18) contains second derivatives in the spatial variables and is therefore much like a diffusive system with strange convective terms. If we effectively restrict consideration to the very high frequency components of Eq. (18) by only retaining terms that contain the highest number of spatial derivatives, we find that

$$\frac{\partial A}{\partial t} \approx \delta\, dA, \tag{25}$$

where $\delta = -*d*$. The right-hand side of Eq. (25) is not the Laplacian on M defined by the initial connection, which would be given by $\Delta = \delta d + d\delta$. Thus the system does not, even in this approximation, represent an exact diffusive evolution. However, under the same conditions for which Eq. (25) is valid, we have

$$\frac{\partial f}{\partial t} \approx \Delta f, \tag{26}$$

which *is* a diffusive equation. It is therefore highly probable that high frequency spatial perturbations in the curvature and torsion induced on M by $A(t)$ are rapidly damped as $t \to \infty$. We would therefore expect a breakdown of our conjectured convergence behavior of the flow only if the low frequency modes do not have the appropriate evolution. These modes are dominated by the nonlinear coupling in Eq. (18).

It seems highly probable that such a system will not lend itself to easy analytic study. Rather, before attempting such an analysis, it seems appropriate that we should verify the conjectured behavior as best we can. To do this, we resort to techniques of *experimental mathematics*, wherein we view the conjectured convergence behavior as a *hypothesis*. We will seek experimental evidence for or against the hypothesis.

We can then study the following two questions:

(1) Does A evolve smoothly under Eq. (18) from smooth initial data?
(2) Does A converge to a fixed point in an appropriate space of connections over M? (In general, we expect convergence within an appropriately normed space, which we would have to describe. However, since the computer experiments use only a fixed number of mesh points to represent the manifold, all norms are effectively equivalent for purposes of this section.) If both questions have affirmative answers, we would also like to answer the concomitant question:
(3) Does $F(A) \to 0$ as $t \to \infty$?

Should the answers to these questions turn out to be yes, for a reasonably wide set of initial conditions and topologies, we would have hope that further study of the flow method ought to be useful in understanding uniformization theorems. We might further hope that our numerical ''experiments'' would enable us to observe useful properties of Eq. (18) that help suggest ways to prove our conjecture. To the contrary, should we find a numerical counterexample to our conjecture, we could rapidly verify that the conjecture would be false. This is the heart of the experimental mathematics concept. Given that we conjecture Eq. (18) to be diffusive and parabolic, we must first choose a numerical method appropriate to such equations. It is then an important part of the experimental process to verify the consistency of the chosen method. The explicit forward Euler method is the most naive numerical integration method appropriate for diffusive parabolic PDEs. It is important to choose carefully. Assume for the moment that the ''convection'' terms were to dominate over the ''diffusion'' terms, producing dynamics for A that are far more like hyperbolic systems. Hyperbolic PDEs are generally unstable when evolved via naive numerical methods, such as the forward Euler method. Furthermore, it would seem improbable that A would converge if Eq. (18) was primarily hyperbolic in nature and if M was compact.

In the Euler method, the iterated value of the connection at time $t_{n+1} > t_n$ is taken to be

$$A_{n+1} = A_n - \kappa(*D_A f(A_n)), \tag{27}$$

where A_n is the value of A at a time $t_n := t_0 + n\kappa$, $\kappa > 0$, and where A_i is given on a mesh of discrete points approximating the manifold M. Such an approximation is shown in Fig. 1, where an initial connection is represented on a torus. The curvature is represented by variations of the torus geometry, and the actual torus generated is only a representation of the state of $A(t)$ and $f(t)$.

We wish to compare Eq. (18) to the evolution of a linear system of the form

$$\frac{\partial u}{\partial t} = Ku(t), \tag{28}$$

where u is a quantity given on M, and K is a linear operator acting on u. The corresponding forward Euler method is

$$u_{n+1} = u_n + \kappa K u_n, \tag{29}$$

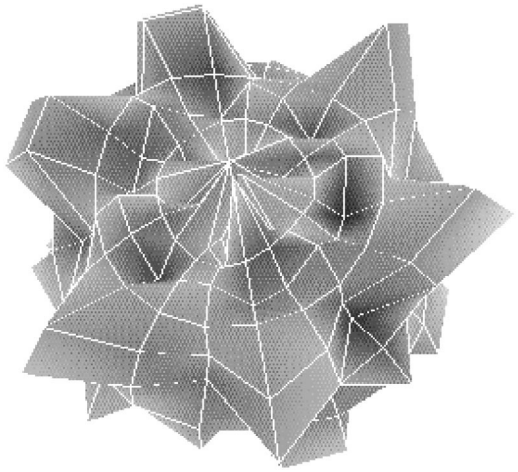

FIG. 1. An initial toroidal mesh.

where everything is expressed on a mesh over M. It is well-known that Eq. (29) fails to produce a valid approximation to the solution of the hyperbolic equation Eq. (28): The numerically produced solution undergoes rapid growth in modes that are high in spatial frequency. The approximation is bad, independently of how small we take κ to be. Diffusive parabolic systems are different in behavior: For κ small enough, the numerical solution obeys Eq. (28).

In view of this comparison, part of our experiment consists of identifying whether Eq. (27) produces a stable approximation (suggesting parabolic behavior) or an unstable one (suggesting hyperbolic behavior).

The experimental procedure is as follows: the symbolic system MAPLE is used to write part of the computer program, converting Eq. (27) into computer code representing evolution on an appropriate mesh with cell size h and time-step κ. We then verify the stability behavior for κ vs h. If we observe the expected stability behavior, we can then analyze the resulting evolution to see if it is consistent with our conjecture.

Happily, Eq. (27) does indeed display, for $k=0$, the precise behavior of a diffusive parabolic PDE. To be precise, for a toroidal rectangular mesh, the numerical stability is observed when κ is smaller than a value that is of order h^2, for a wide range of initial conditions. One particular example of this can be seen in Fig. 2. This is the anticipated signature of a diffusive system, whence we have verified one part of our conjecture, as discussed above. Furthermore, we obtain evidence that, for the $k=0$ case, all three of our earlier questions may be answered in the affirmative. We find that the numerical evolution of Eq. (18) under Eq. (27) produces a convergent A_n, and that $F(A_n)\to 0$, as $n\to\infty$. This is shown in the final state of the mesh; cf. Fig. 3, also drawn on top of the representative torus. We have therefore found support for our conjectured convergence behavior of solutions of Eq. (18), at the experimental level, for $k=0$.

The conjecture can also be experimentally verified for the $k=1$ (spherical topology) case. In the spherical case, we construct an initial connection on a mesh with spherical topology, as shown in Fig. 4. The radius from the center in this image denotes the curvature radius at that point. The flow equations can then be evolved. The connection then flows to one corresponding to a unit 2-sphere, as shown in Fig. 5. We thus have experimental confirmation of the conjecture for $k=0$ and $k=1$.

It remains to verify the conjecture for $k<0$. Unfortunately, Eq. (27) is not very useful in this case. The fact that κ must be smaller than something of order h^2 means that the equations cannot be evolved quickly when we wish to have h small. For complicated topologies that arise for $k<0$ (particularly handlebodies, which do not arise in the other cases), we need many cells, and h *must* be small. Furthermore, it becomes difficult to control the build up of numerical noise with such a naive, conditionally stable, method. For $k<0$ (*and* the 3-D flow described in the next section), more sophisticated techniques will be needed. These techniques will have to deal with the fact that M will generally need to be covered by more than one coordinate patch, more than one gauge patch, will have complicated topology, and even more nonlinear behavior. The best prac-

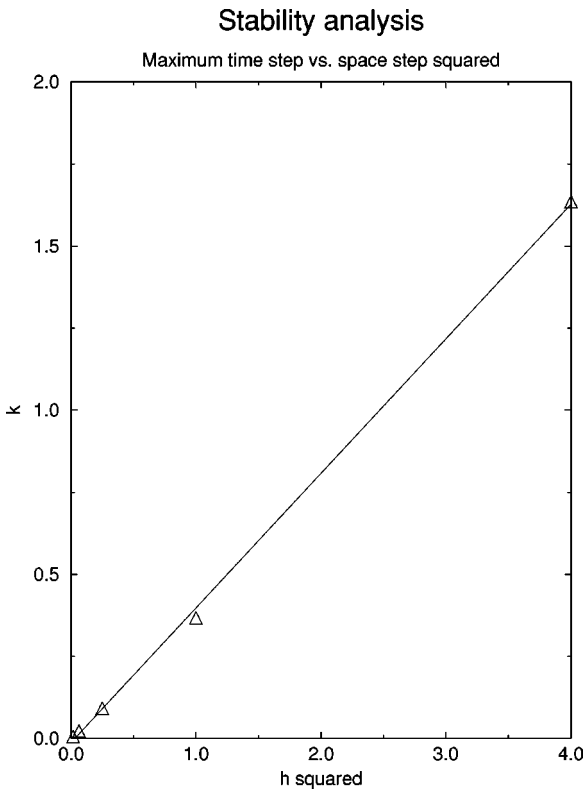

FIG. 2. Stability behavior of the forward Euler method for the flow equations. Line marks linear fit.

tical algorithms seem to be those using multigrid finite element methods. One of us (S.P.B.) is currently developing such techniques.

VI. THE 3-D FLOW

The gauge-theoretic version of a two-dimensional theory of gravity was the starting point for a new approach to proving the two-dimensional uniformization theorem. We emphasize here that we have not completely succeeded in constructing the proof; we have only demonstrated its plausibility. Nevertheless, this suggests that it might be useful to examine gauge-theoretic versions of 3-D gravity theories as possible starting points for a proof of the 3-D uniformization theorem conjectured by Thurston.

Such theories are well-known. Einstein gravity (with or without a cosmological constant) can be formulated as a Chern–Simons gauge theory with gauge group ISO(3) (if the cosmological constant is zero) or SO(4) or SO(3,1) if the cosmological constant is positive or negative.[8] This suggests that we consider a flow of the form

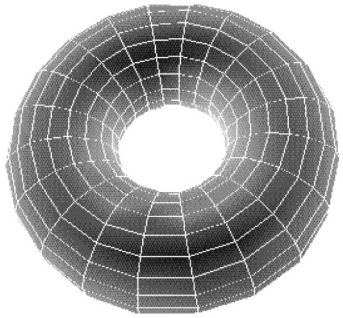

FIG. 3. The final configuration, after numerical evolution.

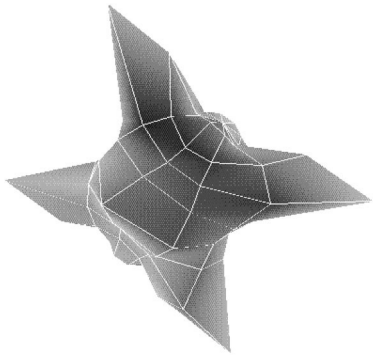

FIG. 4. An initial connection for a spherical topology.

$$\frac{\partial A}{\partial t} = *D_A * F(A), \tag{30}$$

where A is the connection 1-form on some 3-manifold for some gauge group G. One recovers the Riemannian geometry from the connection 1-form via a relation of the form:

$$A = \omega^a G_a + e^a F_a + \cdots , \tag{31}$$

where the G_a, F_n are generators of G, and the ϵ^a, ω^a can be interpreted as a frame field and spin-connection, respectively. The $\cdots$ indicates other terms in any additional generators of the gauge group G. The idea is to choose the gauge group G so that the flat connections (which are a subset of the fixed points of the flow) include at least the *eight* homogeneous geometries that occur in Thurston's conjecture. This is not the case if one chooses one of the groups [ISO(3), etc.] relevant to Einstein gravity. The flat connections for these groups determine frame-fields e^a and compatible spin-connections ω^a which have *constant curvature.*

There is a gauge group whose flat connections include the eight homogeneous three-dimensional geometries. The group is the ''doubly inhomogenized'' group IISO(3). This group is the semidirect product of the ''Poincare group'' ISO(3) with its Lie algebra. It is a 12 parameter noncompact group, whose Lie algebra is

$$[F^a, G^b] = \tfrac{1}{2}\epsilon^{abc} F_c ; \quad [G^a, G^b] = \tfrac{1}{2}\epsilon^{abc} G_c ; \tag{32}$$

$$[G^a, J^b] = \tfrac{1}{2}\epsilon^{abc} J_c ; \quad [G^a, K^b] = \tfrac{1}{2}\epsilon^{abc} K_c ; \tag{33}$$

$$[J^a, K^b] = \tfrac{1}{2}\epsilon^{abc} F_c . \tag{34}$$

The remaining brackets vanish. The G^a generate the SO(3) subgroup, while the remaining generators F^a, J^a, K^a behave like generators of translations, except that the latter two do not commute.

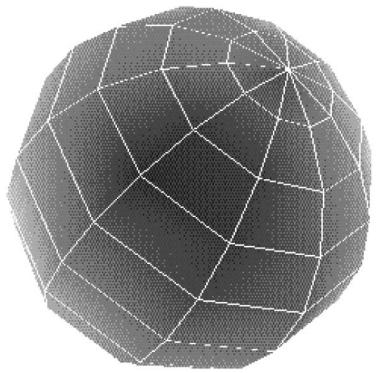

FIG. 5. Final, unit sphere, result of evolution.

TABLE II. Frame fields for homogeneous geometries.

3-manifold	(e^1,e^2,e^3)
S^3	$(\cos y\,dx,dy,dz-\sin y\,dx)$
E^3	(dx,dy,dz)
H^3	$(dx,e^x dy,e^x dz)$
$S^2\times E^1$	$(dx,dy,\sin y\,dz)$
$H^2\times E^1$	$(dx,dy,e^x dz)$
$\mathrm{SL}(\widetilde{2},R)$	$(\cosh y\,dx,dy,dz+\sinh y\,dx)$
Nil	$(dx-z dy,dy,dz)$
Sol	$(dx,e^{-x}dy,e^x dz)$

It was shown in Ref. 15 that the Chern–Simons functional with this gauge group is equivalent to a three-dimensional theory of gravity interacting with topological matter. This is accomplished by constructing the IISO(3) connection as follows:

$$A=\omega^a G_a+e^a F_a+B^a J_a+C^a K_a\,. \tag{35}$$

If A is flat, i.e., $F(A)=0$, then by use of the algebra Eq. (34) it follows that ω^a is a flat SO(3) connection, B^a,C^a are covariantly constant with respect to ω^a and e^a satisfies:

$$D_\omega e^a+\tfrac{1}{2}\epsilon^{abc}B_b\wedge C_c=0, \tag{36}$$

where D_ω is the gauge covariant derivative with respect to the connection ω^a.

One may now construct a spin-connection which is compatible with the frame-field e^a, and hence determines a Riemannian geometry. In particular, one can show that if the e^a are the frame-fields for a homogeneous geometry, then there exist flat connections ω^a and fields B^a,C^a satisfying $D_\omega B^a=0$, $D_\omega C^a=0$ such that ϵ^a satisfies Eq. (36). The explicit expressions are shown in Tables II and III, where we have taken $\omega^a=0$ for simplicity.

It must be remarked here that one may perform an IISO(3) gauge transformation on the flat connection equivalent to the homogeneous geometries. The connections are still flat, of course, but in general the e^a are no longer frame-fields for a homogeneous geometry. It is not clear at this point how general the gauge-transformed e^a are; though it was shown in Ref. 15 that for the case of the 3-manifold topology $S^2\times S^1$, all admissible e^a could be obtained from the ''trivial'' configuration $\omega^a=e^a=B^a=C^a=0$ by a gauge transformation.

If it could be shown that an arbitrary IISO(3) connection on the 3-manifold flowed to one of the flat connections, then it would follow that the 3-manifold would admit the *homogeneous* representative of the gauge orbit. This is what we expect of a proof of the uniformization conjecture for 3-manifolds.

In the two-dimensional Yang–Mills flow, the fixed points of the system are the Yang–Mills connections. However, it seems to be the case that regular initial conditions will flow to the appropriate subclass of flat connections. The reason for this is that in two dimensions, the nonflat

TABLE III. B and C 1-form fields for homogeneous geometries.

3-manifold	(B^1,B^2,B^3)	(C_1,C_2,C_3)
S^3	$(0,dx,0)$	$(d(2\sin y),0,d(2\cos y))$
E^3	$(0,0,0)$	$(0,0,0)$
H^3	$(d(2e^x),0,0)$	$(0,d(-z).dy)$
$S^2\times E^1$	$(d(-2\sin y),0,0)$	$(0,dz,0)$
$H^2\times E^1$	$(d(-2e^x),0,0)$	$(0,dz,0)$
$\mathrm{SL}(\widetilde{2},R)$	$(0,dx,0)$	$(d(-2\sinh y),0,d(2\cosh y))$
Nil	$(0,dy,0)$	$(0,0,d(-2z))$
Sol	$\left(d\left(\frac{z-y}{2}\right),d(e^x),d(-e^{-x})\right)$	$\left(d\left(\frac{z+y}{2}\right),d(2e^x),d(2e^{-x})\right)$

Yang–Mills connections are reducible. This is not the case in three dimensions, since the duals of the curvature 2-forms are 1-forms, and hence *not* gauge parameters. Hence we conjecture that in three dimensions some regular connections, built from nondegenerate Riemannian metrics, will flow to instanton fixed points, not flat fixed points. This might be related to a fundamental difference between the two- and three-dimensional cases. Indeed, as we discussed in Sec. I, in contrast to the two-dimensional case, a given closed orientable 3-manifold does not in general admit a homogeneous Riemannian geometry. Thurston's conjecture is that any 3-manifold admits a canonical decomposition into the connected sum of *simple manifolds*, i.e., prime manifolds which either have no incompressible embedded T^2 or are Seifert fiber spaces; and each simple manifold in turn admits one and only one of the eight homogeneous geometries. If the flow converged to fixed points which were flat connections, then in general it would have to be singular when the manifold was not simple.

However, consider the flow of the curvature as above. The fixed points are the flat connections $F(A)=0$ *and* the Yang–Mills instantons which satisfy $*D_A*F(A)=0$, $F(A)\neq 0$. It would be interesting to examine IISO(3) connections on 3-manifolds to see if the Yang–Mills equations have nonflat solutions when the manifold is simple. If it turns out that simple 3-manifolds admit flat solutions of the Yang–Mills equations, then the next question to examine would be the structure of 3-manifolds which consist of two simple manifolds joined by a ''neck.'' If Thurston's conjecture is correct, then the following scenario should hold: the manifold-with-neck would presumably admit an instanton which was asymptotically flat at the ends of the neck. In general, nonsimple 3-manifolds would *not* admit nontrivial flat IISO(3) connections. Work in this direction is underway by the authors.

ACKNOWLEDGMENTS

The authors thank Eric Woolgar for many useful discussions and for his initial collaboration in this work. They also thank Steve Boyer and Jim Isenberg for helpful discussions.

[1] H. Poincare, Acta Math. **31**, 1 (1907).
[2] W. Thurston, *The Geometry and Topology of Three Manifolds*, Princeton University Lecture Notes (Princeton University Press, Princeton, 1978).
[3] R. Hamilton, in *Mathematics and General Relativity*, edited by J. Isenberg, Contemporary Mathematics, Vol. 71 (American Mathematical Society, Providence, RI, 1988).
[4] B. Chow, J. Diff. Geom. **33**, 325 (1991).
[5] J. Isenberg and M. Jackson, J. Diff. Geom. **35**, 723 (1992).
[6] H. Kneser, Jahr. Deutscher Math. Ver. **38**, 248 (1929); J. Milnor, Am. J. Math. **84**, 1 (1962).
[7] For a review see, A. Ashtekar, *Lectures on Non-Perturbative Canonical Gravity* (World Scientific, Singapore, 1991).
[8] A. Achucarro and P. Townsend, Phys. Lett. B **180**, 89 (1986); E. Witten, Nucl. Phys. A **311**, 46 (1988).
[9] T. Fukuyama and K. Kamimura, Phys. Lett. B**160**, 259 (1985); K. Isler and C. Trugenberger, Phys. Rev. Lett. **63**, 834 (1989); A. Chamseddine and D. Wyler, Phys. Lett. B **228**, 75 (1989).
[10] D. Cangemi and R. Jackiw, Phys. Rev. Lett. **69**, 233 (1992).
[11] R. Jackiw, in *Quantum Theory of Gravity*, edited by S. Christensen (Adam Hilger, Bristol, 1984); C. Teitelboim, in *Quantum Theory of Gravity*, edited by S. Christensen (Adam Hilger, Bristol, 1984).
[12] M. F. Atiyah and R. Bott, Topology **23**, 1 (1984); E. Witten, J. Geom. Phys. **9**, 303 (1992).
[13] For our purposes, the most relevant discussion occurs in J. Rade, Ph.D. thesis, University of Texas at Austin, 1991. A discussion of the four-dimensional case is C.-L. Shen, in *XXI International Conference on Differential Geometric Methods in Theoretical Physics*, edited by C. N. Yang, M. L. Ge, and X. W. Zhou (World Scientific, Singapore, 1993). See also L. A. Sadun, Ph.D. thesis, University of California at Berkeley, 1987; K. Corlette, in *Geometry of Group Representations*, edited by W. M. Goldman and A. R. Magid, Contemporary Mathematics, Vol. 74 (American Mathematical Society, 1988).
[14] See, for example, R. S. Hamilton, *Harmonic Maps of Manifolds with Boundary*, Lecture Notes in Math., Vol. 471 (Springer-Verlag, Berlin, 1975).
[15] S. Carlip and J. Gegenberg, Phys. Rev. D **44**, 424 (1991).

Integral equation formulation of the spinless Salpeter equation

F. Brau[a),b)]
Université de Mons-Hainaut, Place du Parc 20, B-7000 Mons, Belgique

(Received 27 June 1997; accepted for publication 12 November 1997)

The spinless Salpeter equation presents a rather particular differential operator. In this paper we rewrite this equation into integral and integro-differential equations. These kinds of equations are well known and can be more easily handled. We also present some analytical results concerning the spinless Salpeter equation and the action of the square-root operator. © *1998 American Institute of Physics.* [S0022-2488(98)00403-4]

I. INTRODUCTION

The Schrödinger equation is a very well defined partial derivative equation since its differential operator is a Laplacian. Moreover it reduces to a simple differential equation when the interaction is central. This equation has been intensively studied and it is well understood. A simple relativistic version of the Schrödinger equation, the spinless Salpeter equation (SSE), presents more difficulties. This equation is

$$(\sqrt{\mathbf{p}^2+m_1^2}+\sqrt{\mathbf{p}^2+m_2^2})\Psi(\mathbf{r})=(E-V(\mathbf{r}))\Psi(\mathbf{r}), \tag{1}$$

where m_1, m_2 are the masses of the particles, $\mathbf{p}$ is their relative momentum, $V(\mathbf{r})$ is the potential interaction and E the eigenenergy of the stationary state $\Psi(\mathbf{r})$ ($\hbar=c=1$). $\mathbf{p}$ and $\mathbf{r}$ are conjugate variables. Actually, this last equation is not so well defined since the kinetic energy operator is a nonlocal one. Its action on a function $f(\mathbf{r})$ is known only if $f(\mathbf{r})$ is an eigenfunction of the operator $\mathbf{p}^2$. In this case we obtain

$$\sqrt{\mathbf{p}^2+m^2}f(\mathbf{r})=\sqrt{\alpha+m^2}f(\mathbf{r}), \tag{2}$$

where α is the corresponding eigenvalue of $\mathbf{p}^2$. Consequently this operator is difficult to handle. However the SSE is not a marginal equation. The correct description of the bound states of two particles is achieved with the Bethe–Salpeter equation.[1] This last reduces to the SSE (Ref. 2, p. 297) when the following occurs.

- The elimination of any dependences on timelike variables is performed (which leads to the Salpeter equation[3]).
- Any references to the spin degrees of freedom of particles are neglected as well as negative energy solutions.

Despite the presence of a so particular operator, the SSE is often used (see for instance Refs. 4–10) since its numerical resolution is very easy for bound states (see for example Refs. 9,11–13). However, it would be interesting to reformulate this equation in a way to better understand it or even to better use it. The main problem is the nonlocality hidden inside the kinetic energy operator. The idea is thus to extract this nonlocality to put it into evidence. That is why in this work we rewrite the SSE as integral or integro-differential equations where the nonlocality is then explicit. Moreover these kinds of equations are well known and defined. Another integral equation, different from those presented here, was found in Ref. 11. This formalism allows us to define the action of the square-root operator on wave functions and calculate the resulting functions.

[a)]Chercheur I.I.S.N.
[b)]Electronic mail: fabian.brau@umh.ac.be

 © 1998 American Institute of Physics

This paper is organized as follows. In Sec. II, we present the integral and integro-differential equations and we calculate their kernels. In Sec. III, we find some analytical results concerning the action of the square-root operator and solutions of the SSE. At last, we present, in Sec. IV, a summary of this work.

II. INTEGRAL EQUATION FORMULATION

It is possible to rewrite the SSE into different forms. The first one that we present is an integral equation. It is obtained using the Green functions of the kinetic energy operator. The corresponding kernels are different depending on whether the particles are identical or not. The second one is an integro-differential equation. It is obtained taking the square of the square-root operator and is only valid for two identical particles. For these formulations of the SSE only local interactions can be used contrary to the formulation proposed in Ref. 11.

A. First form

In the Introduction we gave the operator expression of the SSE. We are going now to formulate the integral expression of this equation. To obtain it, we use the Green functions of the kinetic energy operator. We first present the equal masses case. Let us consider the following function:

$$G(\Delta)=\frac{1}{(2\pi)^3}\int \frac{\exp(-i\mathbf{p}.\boldsymbol{\Delta})}{2\sqrt{p^2+m^2}}d\mathbf{p}, \tag{3}$$

with $\boldsymbol{\Delta}=\mathbf{r}-\mathbf{r}'$ and $\Delta=|\boldsymbol{\Delta}|$. Then, $G(\Delta)$ is the Green function of the equal masses square-root operator since

$$2\sqrt{\mathbf{p}^2+m^2}G(\Delta)=\frac{1}{(2\pi)^3}\int \exp(-i\mathbf{p}\cdot\boldsymbol{\Delta})d\mathbf{p}=\delta^3(\mathbf{r}-\mathbf{r}'). \tag{4}$$

The solution of the SSE can be written as

$$\Psi(\mathbf{r})=\Psi_0(\mathbf{r})+\int G(\Delta)(E-V(\mathbf{r}'))\Psi(\mathbf{r}')d\mathbf{r}', \tag{5}$$

with $\sqrt{\mathbf{p}^2+m^2}\Psi_0(\mathbf{r})=0$. The solutions of this last equation are $\Psi_0(\mathbf{r})=0$ and $\Psi_0(\mathbf{r})=i_l(mr)Y_{lm}(\hat{r})$, where $i_l(x)=\sqrt{(\pi/2x)}\,I_{l+1/2}(x)$, $I_\nu(x)$ is a modified Bessel function (Ref. 14, p. 952), $Y_{lm}(\hat{r})$ is a spherical harmonic, $r=|\mathbf{r}|$ and $\hat{r}=\mathbf{r}/r$. In Appendix B we show that $\Psi_0(\mathbf{r})=k_l(mr)Y_{lm}(\hat{r})$ is not a solution, the function $k_l(x)$ being defined as equal to $\sqrt{(2/\pi x)}K_{l+1/2}(x)$, with $K_\nu(x)$ a modified Bessel function (Ref. 14, p. 952). Thus only the vanishing solution is relevant for physical problems. One sees immediately that the utilization of nonlocal potentials is impossible if we want to obtain a 1-dimensional integral equation. Calculating the 3-dimensional integral (3) we obtain

$$G(\Delta)=\frac{m}{4\pi^2\Delta}K_1(m\Delta). \tag{6}$$

If we consider only the case of central potentials, the wave function can be written in the form $\Psi(\mathbf{r})=R_l(r)Y_{lm}(\hat{r})$ where $R_l(r)$ is the radial part of the wave function. If, moreover, we place the z axis along $\mathbf{r}$, the angular dependence of Δ is just given by the angle between the vector $\mathbf{r}'$ and the z axis. Then knowing that (Ref. 15, p. 158)

$$Y_{lm}(0,\varphi)=\sqrt{\frac{2l+1}{4\pi}}\delta_{m0}, \tag{7}$$

Eq. (5) is written as

$$R_l(r)\sqrt{\frac{2l+1}{4\pi}}\delta_{m0}=\frac{m}{4\pi^2}\int_0^\infty r'^2(E-V(r'))R_l(r')dr'\int_0^\pi \sin\,\theta'\frac{K_1(m\Delta)}{\Delta}d\theta'\int_0^{2\pi}Y_{lm}(\hat{r}')d\varphi'. \tag{8}$$

We can now perform the integration over the angular variables. We have the well known relations

$$\int_0^{2\pi}Y_{lm}(\hat{r}')d\varphi'=2\pi Y_{lm}(\hat{r}')\delta_{m0}, \tag{9}$$

and (Ref. 15, p. 133)

$$Y_{lm}(\theta,\varphi)=\exp(im\varphi)\sqrt{\frac{2l+1(l-m)!}{4\pi(l+m)!}}P_l^m(\cos\,\theta), \tag{10}$$

where the functions $P_l^m(x)$ are the Legendre functions (Ref. 14, p. 998). Finally, using Eqs. (8)–(10) and the regularized function $u_l(r)=rR_l(r)$, we can write the spinless Salpeter equation on the form of the following integral equation:

$$u_l(r)=\frac{1}{2\pi}\int_0^\infty \mathscr{G}_l(mr,mr')(E-V(r'))u_l(r')dr', \tag{11}$$

with

$$\mathscr{G}_l(mr,mr')=mrr'\int_0^\pi \sin\,\theta'\frac{K_1(m\Delta)}{\Delta}P_l(\cos\,\theta')d\theta', \tag{12}$$

where the functions $P_l(x)$ are the Legendre polynomials (Ref. 14, p. 1025). The functions $\mathscr{G}_l(mr,mr')$ are analytic for each value of l. Performing the transformation $y=\sqrt{r^2+r'^2-2rr'\,\cos\,\theta'}$, we obtain

$$\mathscr{G}_l(mr,mr')=m\int_{|r-r'|}^{r+r'}K_1(my)P_l\left(\frac{r^2+r'^2-y^2}{2rr'}\right)dy. \tag{13}$$

For each value of l, we can find the primitive and thus calculate the kernel. Indeed, we have the following relations (Ref. 16, p. 87):

$$\int x^{\nu+1}K_\nu(x)dx=-x^{\nu+1}K_{\nu+1}(x), \tag{14a}$$

$$\int x^{-\nu+1}K_\nu(x)dx=-x^{-\nu+1}K_{\nu-1}(x). \tag{14b}$$

With these it is easy to show that

$$\int x^\nu K_\mu(x)dx=-x^\nu K_{\mu-1}(x)-(\nu+\mu-1)x^{\nu-1}K_\mu(x)+(\nu+\mu-1)(\nu-\mu-1)\int x^{\nu-2}K_\mu(x)dx. \tag{15}$$

This last relation is useful when one calculates the kernel for $l\geqslant 2$. The modified Bessel functions recursion relation is (Ref. 14, p. 970)

$$K_{\nu+1}(x)=\frac{2\nu}{x}K_\nu(x)+K_{\nu-1}(x). \tag{16}$$

With Eq. (13) we see that the polynomial under the integral is composed only of even powers of the integration variable. Then with relation (15), we can calculate analytically the kernel for each value of l. Another way (more simple) to calculate the kernel is to use the relation derived in Ref. 11,

$$\mathcal{G}_l(x,x')=2^l z^{l+1}\left(\frac{1}{z}\frac{\partial}{\partial z}\right)^l \frac{1}{z}[(y-z)^{l/2}K_l(\sqrt{y-z})-(y+z)^{l/2}K_l(\sqrt{y+z})], \tag{17}$$

with $y=x^2+x'^2, z=2xx'$. The kernel has as an expression for $l=0$ and $l=1$,

$$\mathcal{G}_0(x,x')=K_0(|x-x'|)-K_0(x+x'), \tag{18}$$

$$\mathcal{G}_1(x,x')=K_0(|x-x'|)+K_0(x+x')+\frac{1}{xx'}[(x+x')K_1(x+x')-|x-x'|K_1(|x-x'|)]. \tag{19}$$

The expressions (18) and (19) show that kernels decrease exponentially when arguments are important enough ($(x+x')$ and $|x-x'|\gg 1$) since $K_\nu(x)\sim\sqrt{\pi/(2x)}\exp(-x)$ when $|x|\gg 1$ (Ref. 14, p. 961). This behavior indicates that when the masses of particles are large, the nonlocality becomes negligible, the SSE becomes almost local and its spectrum coincides with the Schrödinger spectrum. However the explicit limit cannot be done in this formalism. One can remark that these kernels present a logarithmic singularity for $r=r'$.

We treat now the unequal masses case. After the previous calculations, it is obvious that the Green function for the unequal masses kinetic energy operator is given by

$$\widetilde{G}(\Delta)=\frac{1}{(2\pi)^3}\int \frac{\exp(-i\mathbf{p}.\mathbf{\Delta})}{\sqrt{p^2+m_1^2}+\sqrt{p^2+m_2^2}}d\mathbf{p}. \tag{20}$$

This equation becomes, if $m_1\neq m_2$,

$$\widetilde{G}(\Delta)=\frac{1}{(2\pi)^3}\frac{1}{m_1^2-m_2^2}\int (\sqrt{p^2+m_1^2}-\sqrt{p^2+m_2^2})\exp(-i\mathbf{p}.\mathbf{\Delta})d\mathbf{p}. \tag{21}$$

The solution of the SSE can be written as a form similar to Eq. (5) and the function $\Psi_0(\mathbf{r})$ is here always null. One can find it easily by acting the operator $(\sqrt{\mathbf{p}^2+m_1^2}-\sqrt{\mathbf{p}^2+m_2^2})$ on the equation $(\sqrt{\mathbf{p}^2+m_1^2}+\sqrt{\mathbf{p}^2+m_2^2})\Psi_0(\mathbf{r})=0$. Extracting the operator $(-\Delta+m^2)$ from Eq. (21), performing the integration over the momenta and using the Green's theorem (see Ref. 11), one can show that

$$\Psi(\mathbf{r})=\frac{m_1}{2\pi^2(m_1^2-m_2^2)}\int\frac{K_1(m_1\Delta)}{\Delta}(-\Delta_{r'}+m_1^2)(E-V(r'))\Psi(\mathbf{r}')d\mathbf{r}'+(m_1\to m_2). \tag{22}$$

Again, placing the z axis along $\mathbf{r}$ and repeating the previous calculation we obtain

$$u_l(r)=\frac{1}{\pi(m_1^2-m_2^2)}\int_0^\infty \widetilde{\mathcal{G}}_l(m_1 r,m_1 r')(E-V(r'))u_l(r')dr'+(m_1\to m_2), \tag{23}$$

with

$$\widetilde{\mathcal{G}}_l(mr,mr')=\mathcal{G}_l(mr,mr')\left[-\frac{d^2}{dr'^2}+\frac{l(l+1)}{r'^2}+m^2\right]. \tag{24}$$

B. Second form

This second rewriting of the SSE leads to an integro-differential equation. By acting the kinetic energy operator on the left of the equal masses SSE we obtain

$$4(\mathbf{p}^2+m^2)\Psi(\mathbf{r})=E(E-V(\mathbf{r}))\Psi(\mathbf{r})-2\sqrt{\mathbf{p}^2+m^2}V(\mathbf{r})\Psi(\mathbf{r}). \tag{25}$$

Thus, we must calculate the action of the square-root operator on the product of functions $V(\mathbf{r})\Psi(\mathbf{r})$. If we consider only central potentials this product gives a function for which the angular dependence is a spherical harmonic. One can show, with the formalism developed in Ref. 11, that the radial part $g(r)$ of the function $g(r)Y_{lm}(\hat{r})$, resulting from the action of the kinetic energy operator on the function $h(r)Y_{lm}(\hat{r})$, is given by

$$g(r)=\frac{1}{\pi r}\int_0^\infty \mathcal{G}_l(mr,mr')\tilde{h}(r')dr', \tag{26}$$

with $\tilde{h}(x)=xh(x)$. When the right-hand side integral is relevant, this rewriting gives meaning to the action of the kinetic energy operator on a central problem wave function. With this result, one obtains from Eq. (25),

$$\left[\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}-m^2+\frac{E(E-V(r))}{4}\right]u_l(r)=\frac{1}{2\pi}\int_0^\infty \mathcal{G}_l(mr,mr')V(r')u_l(r')dr'. \tag{27}$$

This last equation is interesting because this is a kind of nonlocal Klein–Gordon Equation (KGE). This shows how the SSE differs from KGE when a nonvanishing interaction is introduced (the free SSE leads to a the free KGE with a different wave number k). Thus the difference is partially given by the nonlocal right-hand side of Eq. (27).

To show how the square-root operator can be particular, we give a simple example in which we show how the hidden nonlocality of the kinetic energy operator could lead to some problems. Let us consider a square well with a range r_c and a depth $-V_0$. In this case, according to r is less than r_c or not, the right-hand side of equal masses SSE is a constant (E or $E+V_0$) times the wave function. Then one can think that in each area this equation leads to

$$[\Delta+\mathrm{Sgn}(r_c-r)k^2]\Psi(\mathbf{r})=0, \tag{28}$$

with

$$k^2=\frac{(E+V_0)^2}{4}-m^2, \quad \text{if } r<r_c,$$

$$k^2=m^2-\frac{E^2}{4}, \quad \text{if } r>r_c.$$

The continuity conditions, for the interior and exterior solutions, at distance $r=r_c$, fix the wave function and the energy. But this method to resolve the SSE does not take into account the nonlocality of the kinetic operator and we obtain the same kind of solution as the KGE. Actually, the problem is the discontinuity. The potential is not a constant for all values of r and the kinetic operator does not commute with it. Thus the equal masses SSE does not lead to Eq. (28). However, in Eq. (27) the nonlocality is contained in the potential part and we can now write an equation for each area ($r<r_c$ and $r>r_c$). We remark that this equation does not reduce to the free KGE when $r>r_c$ because the right-hand side does not vanish. Thus Eq. (27) leads to different solutions that these obtained with KGE for the same square well.

We conclude this section with some remarks concerning these integral equations. The equation derived in Ref. 11 is, for the equal masses case,

$$[E-V(r)]u_l(r)=\frac{2}{\pi}\int_0^\infty \mathcal{G}_l(mr,mr')u_l(r')dr'. \tag{29}$$

Thus if the potential is just a constant, $V(r)=V$, using this equation and Eq. (27) we obtain

$$\left[\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}-m^2+\frac{(E-V)^2}{4}\right]u_l(r)=0. \tag{30}$$

The nonlocality disappears and we find the corresponding Klein–Gordon equation.

It is easy to show, with the operator expression of the SSE, that the free radial solutions are $R_l(r)\propto j_l(kr)$ where the function $j_l(kr)$ is the regular spherical Bessel function defined as

$$j_l(x)=\sqrt{\frac{\pi}{2x}}J_{l+1/2}(x), \tag{31}$$

where the function $J_\nu(x)$ is the Bessel function of the first kind (Ref. 14, p. 951). The wave number k is given by

$$k^2=\left(\frac{1}{4E^2}(m_1^2-m_2^2+E^2)^2-m_1^2\right). \tag{32}$$

In Appendix A we show that

$$\int_0^\infty \mathcal{G}_l(mr,mr')\gamma r' j_l(\gamma r')dr'=\frac{\pi}{\sqrt{\gamma^2+m^2}}\gamma r j_l(\gamma r). \tag{33}$$

Then one can verify that Eqs. (11), (23) and (27) are true for a vanishing potential when γ is replaced by k. In Appendix B, we show that the situation is rather different if one considers the irregular spherical Bessel functions.

III. ANALYTICAL RESULTS

In this section we present some analytical results concerning both the action of the square-root operator and solutions of the SSE. We first calculate the action of the kinetic energy operator on the functions $r^n \exp(-mr)$ ($n\geqslant -1$ and n integer). The Fourier transform, $\tilde{f}(p)$, of these functions are

$$\tilde{f}(p)=(-)^{n+1}\sqrt{\frac{2}{\pi}}\frac{\partial^{n+1}}{\partial m^{n+1}}\frac{1}{p^2+m^2}. \tag{34}$$

Then for each value of n we can calculate the action of the square-root operator. We have, for $n=-1$ and $n=0$,

$$\sqrt{\mathbf{p}^2+m^2}\ \frac{\exp(-mr)}{r}=\frac{2m}{\pi r}K_1(mr), \tag{35}$$

$$\sqrt{\mathbf{p}^2+m^2}\ \exp(-mr)=\frac{4m}{\pi}K_0(mr). \tag{36}$$

We can get two more relations knowing the following integrals (Ref. 14, p. 429):

$$\int_0^\infty \frac{x^{2b+1}\sin(ax)}{(m^2+x^2)^{n+1/2}}dx=(-)^{b+1}\frac{\sqrt{\pi}}{2^n\, m^n\,\Gamma\left(n+\frac{1}{2}\right)}\frac{\partial^{\,2b+1}}{\partial\, a^{2b+1}}[a^n\, K_n(ma)], \tag{37}$$

with $a>0$, Re $m>0$, $-1\leqslant b\leqslant n$ and

$$\int_0^\infty \frac{x^{2b+1}\sin(ax)}{(m^2+x^2)^{n+1}}dx=(-)^{b+n}\frac{\pi}{2\,n!}\left[\frac{\partial^n}{\partial\,\gamma^n}(\gamma^b\,\exp(-a\sqrt{\gamma}))\right]_{\gamma=m^2}, \tag{38}$$

with $a>0$, $0\leqslant b\leqslant n$, $|\arg(m^2)|<\pi$. Rewriting these integrals as 3-dimensional integrals, extracting the square-root operator and integrating the resulting integrals one obtains the following formulas:

$$\sqrt{\mathbf{p}^2+m^2}\left\{\frac{1}{r}\left[\frac{\partial^n}{\partial\,\gamma^n}(\gamma^b\,\exp(-r\sqrt{\gamma}))\right]_{\gamma=m^2}\right\}=(-)^{n+1}\frac{2\,n!}{\pi\,m^n\,(2n-1)!!\,r}\frac{\partial^{\,2b+1}}{\partial\,r^{2b+1}}[r^n\,K_n(mr)], \tag{39}$$

and

$$\sqrt{\mathbf{p}^2+m^2}\left\{\frac{1}{r}\frac{\partial^{\,2b+1}}{\partial\,r^{2b+1}}(r^{n+1}\,K_{n+1}(mr))\right\}=(-)^{n+1}\frac{\pi\,m^{n+1}(2n+1)!!}{2\,n!\,r}\times\left[\frac{\partial^n}{\partial\,\gamma^n}(\gamma^b\,\exp(-r\sqrt{\gamma}))\right]_{\gamma=m^2}. \tag{40}$$

These two equations are valid for $0\leqslant b\leqslant n$.

With these results we can find easily some analytical solutions of the SSE with a nonlocal interaction. Let us consider the following equation:

$$2\sqrt{\mathbf{p}^{\,2}+m^2}\,R_0(r)=E\,R_0(r)-\int_0^\infty W(r,r')\,R_0(r')\,dr'. \tag{41}$$

If we choose

$$W(r,r')=\left[\alpha\,\exp(-mr)-\frac{8m}{\pi}\,(\beta+m)\,K_0(mr)\right]\exp(-\beta r'), \tag{42}$$

then using Eq. (36) we obtain the solution

$$R_0(r)\propto\exp(-mr), \tag{43}$$

$$E=\frac{\alpha}{\beta+m}. \tag{44}$$

Thus it is easy to construct a set of solutions with the operator expression of the SSE, we must just consider a well chosen interaction. We can also find analytical solutions using Eq. (29) with a nonlocal interaction:

$$\frac{2}{\pi}\int_0^\infty\tilde{\mathscr{G}}_l(mr,mr')\,u_l(r')\,dr'=E\,u_l(r)-\int_0^\infty W(r,r')\,u_l(r')\,dr'. \tag{45}$$

Choosing, for example, $l=0$ and

$$W(r,r')=-\frac{2}{\pi}\,\mathscr{G}_0(mr,mr')\left[\frac{2\beta}{r'}+(m^2-\beta^2)\right]+\gamma r\,\exp(-\beta r-\alpha r'), \tag{46}$$

then we get the solution

$$u_0(r)\propto r\,\exp(-\beta r), \tag{47}$$

$$E=\frac{\gamma}{(\alpha+\beta)^2}. \tag{48}$$

Again we can construct with this representation of the SSE a set of analytical solutions corresponding to a set of well chosen interactions.

IV. SUMMARY

In this work our purpose was to rewrite the spinless Salpeter equation into forms more easy to handle, since its differential operator is so particular. To perform it we have extracted the hidden nonlocality of the kinetic energy operator to get it really explicit. Thus we have obtained in Sec. II the following equations:

• For $m_1=m_2$,

$$u_l(r)=\frac{1}{2\pi}\int_0^\infty \mathscr{G}_l(mr,mr')\ (E-V(r'))\ u_l(r')\ dr',$$

and

$$\left[\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}-m^2+\frac{E(E-V(r))}{4}\right]\ u_l(r)=\frac{1}{2\pi}\int_0^\infty \tilde{\mathscr{G}}_l(mr,mr')\ V(r')\ u_l(r')\ dr',$$

with

$$\tilde{\mathscr{G}}_l(mr,mr')=\mathscr{G}_l(mr,mr')\left[-\frac{d^2}{dr'^2}+\frac{l(l+1)}{r'^2}+m^2\right].$$

• And for $m_1\neq m_2$,

$$u_l(r)=\frac{1}{\pi(m_1^2-m_2^2)}\int_0^\infty \tilde{\mathscr{G}}_l(m_1r,m_1r')\ (E-V(r'))\ u_l(r')\ dr'+(m_1\rightarrow m_2).$$

The kernel $\mathscr{G}_l(mr,mr')$ was first derived in Ref. 11. It is analytic for each value of angular momentum and is given in Sec II. There are two integral equations and one integro-differential equation. These kind of equations are well known. The problem of a definition, connected to the particular kinetic energy operator, are removed when the spinless Salpeter equation is rewritten into these forms or into the form (29) (derived in Ref. 11). Moreover with this last one we can calculate the action of the square-root operator on any functions of the form $h(r)\ Y_{lm}(\hat{r})$ and find the radial part of the resulting functions. If $g(r)$ is this radial part, we have

$$g(r)=\frac{1}{\pi r}\int_0^\infty \tilde{\mathscr{G}}_l(mr,mr')\ \tilde{h}(r')\ dr',$$

with $\tilde{h}(x)=x\ h(x)$. When the integral in the right-hand side is relevant, this expression gives meaning to the action of the square-root operator on central problem wave functions. Indeed, this last relation allows us to explicitly calculate the resulting functions.

In Sec. III we have found some analytical results concerning the action of the square-root operator on some particular functions. With these results and the integral formalism we have shown how to construct a set of analytical solutions of the SSE with well chosen interactions.

ACKNOWLEDGMENTS

We thank Dr. C. Semay and Professor R. Ceuleneer for their availability and useful discussions.

APPENDIX A: FREE SOLUTIONS OF THE INTEGRAL FORM OF THE SPINLESS SALPETER EQUATION

In this section we show, with the integral equation formalism, that the free radial solutions are the regular spherical Bessel functions. It is easy to prove it with the operator expression of the SSE because one can rewrite it as

$$(\Delta+k^2)\ \Psi(\mathbf{r})=0, \tag{A1}$$

where k is given by Eq. (32).

To prove it with the integral formulation, we use the integral expression of the kernel $\mathcal{G}_l(mr,mr')$ [see Eq. (25), Ref. 11],

$$\mathcal{G}_l(mr,mr')=\frac{m^2}{2}\,r\,r'\int_0^\infty \exp\left[-\frac{1}{u}-\frac{m^2}{4}(r^2+r'^2)u\right] i_l\left(\frac{1}{2}m^2rr'u\right)\,du. \tag{A2}$$

Then knowing that

$$j_l(x)=\sqrt{\frac{\pi}{2x}}\,J_{l+1/2}(x),\quad i_l(x)=\sqrt{\frac{\pi}{2x}}\,I_{l+1/2}(x), \tag{A3}$$

and (Ref. 14, p. 718)

$$\int_0^\infty x\,\exp(-\alpha x^2)\,I_\nu(\beta x)\,J_\nu(\gamma x)\,dx=\frac{1}{2\alpha}\exp\left(\frac{\beta^2-\gamma^2}{4\alpha}\right)J_\nu\left(\frac{\beta\gamma}{2\alpha}\right) \tag{A4}$$

(with Re $\alpha>0$, Re $\nu>-1$), we have

$$\int_0^\infty \mathcal{G}_l(mr,mr')\,\gamma r'\,j_l(\gamma r')\,dr'=\frac{\sqrt{\pi}}{m}\,\gamma r\,j_l(\gamma r)\int_0^\infty\frac{\exp[(-1/u)(1+\gamma^2/m^2))]}{u^{3/2}}\,du$$

$$=\frac{\pi}{\sqrt{\gamma^2+m^2}}\,\gamma r\,j_l(\gamma r). \tag{A5}$$

APPENDIX B: SINGULARITY AND THE RELATIVISTIC COULOMB PROBLEM

We know that the relation

$$\sqrt{-\Delta+m^2}\,G(r)=\sqrt{\pm\beta^2+m^2}\,G(r), \tag{B1}$$

leads to $G(r)\propto j_l(\beta r)$ (plus sign) or $G(r)\propto i_l(\beta r)$ (minus sign). The situation is rather different for the irregular spherical Bessel functions $n_l(x)$ and $k_l(x)$ ($n_l(x)=\sqrt{(\pi/2x)}N_{l+1/2}(x)$, where $N_\nu(x)$ is the Bessel function of the second kind). To illustrate, we just consider the function $k_0(x)=\exp(-x)/x$. We see that the right-hand side of Eq. (35) is different from zero as we could expect from Eq. (B1). Then we can write

$$\sqrt{\mathbf{p}^2+m^2}\,\frac{\exp(-\gamma r)}{r}=\sqrt{m^2-\gamma^2}\,\frac{\exp(-\gamma r)}{r}+f(r), \tag{B2}$$

where

$$(\sqrt{\mathbf{p}^2+m^2}+\sqrt{m^2-\gamma^2})f(r)=4\pi\delta(\mathbf{r}), \tag{B3}$$

since

$$(\Delta-\gamma^2)\frac{\exp(-\gamma r)}{r}=-4\pi\delta(\mathbf{r}). \tag{B4}$$

From Eq. (B3) we deduce

$$f(r)=\frac{2}{\pi r}\int_0^\infty\frac{p\,\sin(pr)}{\sqrt{p^2+m^2}+\sqrt{m^2-\gamma^2}}\,dp. \tag{B5}$$

Thus we see that the singularity at the origin of the functions $n_l(x)$ and $k_l(x)$ is very important and very annoying. Indeed, if $f(r)$ was null it would be easy to solve the equal masses relativistic Coulomb problem for $l=0$. In this case one can show that

$$\sqrt{\mathbf{p}^2+m^2}\exp(-\gamma r)=\sqrt{m^2-\gamma^2}\exp(-\gamma r)+\frac{\gamma}{\sqrt{m^2-\gamma^2}}\frac{\exp(-\gamma r)}{r} \tag{B6}$$

and

$$\sqrt{\mathbf{p}^2+m^2}r\ \exp(-\gamma r)=\sqrt{m^2-\gamma^2}r\ \exp(-\gamma r)+\frac{2\gamma}{\sqrt{m^2-\gamma^2}}\exp(-\gamma r)-\frac{m^2}{(m^2-\gamma^2)^{3/2}}\frac{\exp(-\gamma r)}{r}. \tag{B7}$$

Acting as the square-root operator on the left of these relations one can verify them. The equation to solve is

$$2\sqrt{\mathbf{p}^2+m^2}R_l(r)=\left(E+\frac{\kappa}{r}\right)R_l(r). \tag{B8}$$

Using Eq. (B6), we find that the ground state is

$$R_0(r)\propto\exp\left(-\frac{\kappa m}{\sqrt{4+\kappa^2}}r\right), \tag{B9}$$

with

$$E=\frac{2m}{\sqrt{1+(\kappa/2)^2}}. \tag{B10}$$

Using Eqs. (B6) and (B7), the first excited state is given by

$$R_0(r)\propto\left(1-\frac{16\kappa m}{(16+\kappa^2)^{3/2}}r\right)\exp\left(-\frac{\kappa m}{\sqrt{16+\kappa^2}}r\right), \tag{B11}$$

with

$$E=\frac{2m}{\sqrt{1+(\kappa/4)^2}}. \tag{B12}$$

This spectrum is a rather good approximation. For example, the error is $\approx 0.25\%$, for the ground state, if $\kappa=0.456$ and $m=1$ GeV (see Ref. 13). These energies were first obtained in Ref. 17. Let us remark that the wave functions are Schrödinger-like wave functions.

[1] E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951).
[2] W. Greiner and J. Reinhardt, *Quantum Electrodynamics*, 2nd ed. (Springer-Verlag, Berlin, 1994).
[3] E. E. Salpeter, Phys. Rev. **87**, 328 (1952).
[4] D. P. Stanley and D. Robson, Phys. Rev. D **21**, 3180 (1980).
[5] P. Cea, G. Nardulli, and G. Paiano, Phys. Rev. D **28**, 2291 (1983).
[6] S. N. Gupta, S. F. Radford, and W. W. Repko, Phys. Rev. D **31**, 160 (1985).
[7] S. Godfrey and N. Isgur, Phys. Rev. D **32**, 189 (1985), and references therein.
[8] D. LaCourse and M. G. Olson, Phys. Rev. D **39**, 2751 (1989), and references therein.
[9] L. P. Fulcher, Phys. Rev. D **50**, 447 (1994), and references therein.
[10] C. Semay and B. Silvestre-Brac, Nucl. Phys. A **618**, 455 (1997), and references therein.
[11] L. J. Nickisch, L. Durand, and B. Durand, Phys. Rev. **30**, 660 (1984).
[12] L. Durand and A. Gara, J. Math. Phys. **31**, 2237 (1990).
[13] F. Brau and C. Semay, J. Comput. Phys. **139**, 127 (1998).
[14] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, corrected and enlarged edition (Academic, New York, 1980).
[15] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).
[16] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer-Verlag, Berlin, 1966).
[17] B. Durand and L. Durand, Phys. Rev. D **28**, 396 (1983); **50**, 6662 (1994) (erratum).

Representations of the Poincaré Lie algebra from unitary representations of SO(1,4)

S. A. Buchwalter
Program in Mathematics, City University of New York Graduate School, City University of New York, New York, New York 10036

P. Moylan
Physics Department, The Pennsylvania State University, Abington, Pennsylvania 19001

(Received 13 June 1997; accepted for publication 3 November 1997)

An algebraic isomorphism of an extension of the Lie field of so(1,4) and an extension of the Poincaré Lie field was given in Havlicek and Moylan [J. Math Phys. **34**, 5320 (1993)]. Here we use the isomorphism described in that paper in order to obtain matrix elements of the Poincaré Lie algebra generators in a normalizable basis for an arbitrary unitary spherical principal series representation of $SO_0(1,4)$.
© *1998 American Institute of Physics.* [S0022-2488(98)00604-5]

I. INTRODUCTION AND NOTATION

In Ref. 1 we have constructed an isomorphism between certain algebraic extensions of the Lie fields of the de Sitter algebra so(1,4) and of the Poincaré Lie algebra. The construction of this isomorphism is based on solving equations of the form $f_\nu(P_\mu)=B_\nu$ $(\mu,\nu=0,1,2,3)$ for unknown, but mutually commuting P_μ where f_ν are algebraic functions of the P_μ and of $(P_\mu P^\mu)^{1/2}$ with coefficients in the enveloping algebra of so(1,4). The B_ν are certain $SO_0(1,4)$ generators. The consequences of our algebraic results for the representation theory of these groups and algebras was studied in Ref. 2. We introduced, on the Hilbert space of a unitary spherical principal series representation of $SO_0(1,4)$,[3] a domain on which the P_μ are skew symmetric operators.

Here we apply these investigations to obtain some precise results about actions of the Poincaré Lie algebra generators in a normalizable basis for an arbitrary spherical principal series representation of $SO_0(1,4)$. Our mathematically rigorous approach is to be contrasted with the very formal results obtained in Ref. 4 for the action of the Poincaré Lie algebra in the ''Lorentz basis.''[4,5] In particular, the matrix elements of the (skew symmetric) Poincaré translation generators in our basis satisfy the properties, which one would expect of the matrix elements of a skew symmetric operator in Hilbert space.

Our notation and conventions are the same as in Ref. 1. In particular we denote by bold faced characters objects such as abstract Lie algebra elements and enveloping algebra elements.

II. MAIN THEOREM

Let $G=SO_0(1,4)$ and let H be a Lorentz subgroup of G. Denote by $\mathscr{G}$ the Lie algebra of G. Let $\mathbf{L}_{ab}$ $(a,b=0,1,2,3,4)$, be the basis for $\mathscr{G}$ as defined in Ref. 1. (They are just the usual generators of pseudo-rotations in the various planes of the five dimensional pseudo-Euclidean space in which the de Sitter space is embedded.) Denote by $\mathscr{E}(\mathscr{G})$ the universal enveloping algebra of $\mathscr{G}$. Let P be the Poincaré group, and denote its Lie algebra by $\mathscr{P}$. A basis for the Poincaré Lie algebra is given by the $\mathbf{L}_{0i}$ and $\mathbf{L}_{ij}$ $(i,j=1,2,3)$ of H together with a commuting Lorentz vector operator $\mathbf{P}_\mu$, which generates the translations.

Now we define the following operators:

$$\mathbf{L}_{12},\mathbf{L}^2,\mathbf{Q}_2=\mathbf{L}_{01}^2+\mathbf{L}_{02}^2+\mathbf{L}_{03}^2-\mathbf{L}^2$$

$$\mathbf{Q}_4=(\mathbf{L}_{12}\mathbf{L}_{30}+\mathbf{L}_{23}\mathbf{L}_{10}+\mathbf{L}_{31}\mathbf{L}_{20})^2,$$

$$\mathbf{D}_2=\mathbf{L}_{04}^2+\mathbf{L}_{01}^2+\mathbf{L}_{02}^2+\mathbf{L}_{03}^2-\mathbf{\Delta}$$

 © 1998 American Institute of Physics

$$\mathbf{D}_4=[\tfrac{1}{4}(\lambda-\rho)]^2-(\mathbf{L}_{12}\mathbf{L}_{30}+\mathbf{L}_{23}\mathbf{L}_{10}+\mathbf{L}_{31}\mathbf{L}_{20})^2 - \left(\sum_{\substack{ijk\\lm=1}}^{3}(\epsilon_{ijk}\{\tfrac{1}{2}\mathbf{L}_{04}\mathbf{L}_{jk}-\mathbf{L}_{0j}\mathbf{L}_{4k}\})(\epsilon_{ilm}\{\tfrac{1}{2}\mathbf{L}_{04}\mathbf{L}_{lm}-\mathbf{L}_{0l}\mathbf{L}_{4m}\})\right), \tag{2.1}$$

where

$$\mathbf{\Delta}=\mathbf{L}_{12}^2+\mathbf{L}_{23}^2+\mathbf{L}_{31}^2+\mathbf{L}_{14}^2+\mathbf{L}_{24}^2+\mathbf{L}_{34}^2, \tag{2.2}$$

and

$$\tfrac{1}{4}(\lambda-\rho)=\mathbf{L}_{12}\mathbf{L}_{34}+\mathbf{L}_{23}\mathbf{L}_{14}+\mathbf{L}_{31}\mathbf{L}_{24}. \tag{2.3}$$

$-\mathbf{L}^2$ and $-i\mathbf{L}_{12}$ are operators for the total angular momentum squared and the third component of the total angular momentum, respectively. $\mathbf{Q}_2$ and $\mathbf{Q}_4$ are the second order and fourth order Casimir operators of H, respectively. The center $Z(G)$ of the universal enveloping algebra $\mathscr{E}(\mathscr{G})$ of $SO_0(1,4)$ is generated by $\mathbf{D}_2$ and $\mathbf{D}_4$, which are second and fourth order Casimir operators of G, respectively.

The Poincaré Lie algebra is generated by the $\mathbf{L}_{0i}$ and $\mathbf{L}_{ij}$ $(i,j=1,2,3)$ of H together with a commuting Lorentz vector operator $\mathbf{P}_\mu$, which generates the translations. Thus, we have the following commutation relations between the $\mathbf{P}_\mu$ and the $\mathbf{L}_{\mu\nu}$:

$$\begin{aligned}[\mathbf{L}_{ij},\mathbf{P}_k]&=-\delta_{ik}\mathbf{P}_j+\delta_{jk}\mathbf{P}_i,\\ [\mathbf{L}_{0i},\mathbf{P}_j]&=-\delta_{ij}\mathbf{P}_0,\\ [\mathbf{P}_\mu,\mathbf{P}_\nu]&=0.\end{aligned} \tag{2.4}$$

Next consider the following objects constructed out of the $\mathbf{L}_{\mu\nu}$ and $\mathbf{P}_\mu$ of P:

$$\tilde{\mathbf{L}}_{\mu 4}=\frac{\lambda}{\sqrt{\mathbf{P}_\rho\mathbf{P}^\rho}}[\mathbf{Q}_2,\mathbf{P}_\mu]-\mathbf{P}_\mu \quad (\lambda\in\mathbb{R})(\mu=0,1,2,3). \tag{2.5}$$

[,] means commutator, and $\mathbf{Q}_2$ is the second order Casimir operator of H as defined in Ref. 1. The $\lambda^{-1}\tilde{\mathbf{L}}_{\nu 4}$ and the $\tilde{\mathbf{L}}_{\mu\nu}:=\mathbf{L}_{\mu\nu}$ satisfy the commutation relations of the generators of G. The $\tilde{\mathbf{L}}_{\mu 4}$ and the $\tilde{\mathbf{L}}_{\mu\nu}$ are a basis for an isomorphic copy of $\mathscr{G}$, which differs from $\mathscr{G}$ by a scaling factor λ in the $\mathbf{L}_{\mu 4}$ directions. We shall refer to this Lie algebra and the corresponding Lie group by $\mathscr{G}_\lambda$ and G_λ, respectively.

Now let $\tilde{\mathbf{C}}_2^\lambda=-\tilde{\mathbf{D}}_2^\lambda+\frac{5}{2}\lambda^2\mathbf{I}$ and $\tilde{\mathbf{C}}_4^\lambda=\tilde{\mathbf{D}}_4^\lambda+\frac{1}{4}\tilde{\mathbf{D}}_2^\lambda-\frac{9}{16}\mathbf{I}\lambda^2$. [$\mathbf{I}$ is the identity in $\mathscr{E}(\mathscr{G})$.] A proof of the following result is given in Refs. 1 and 2. It involves solving Eqs. (2.5) for the $\mathbf{P}_\mu$.

Theorem I:

Let π_ρ $(\rho\in\mathbb{R})$ be a unitary spherical principal series representation of G_λ on an Hilbert space $\mathscr{H}$, and let $d\pi_\rho$ denote the representation of $\mathscr{G}_\lambda$ on $\mathscr{H}$ corresponding to π_ρ. The representation $d\pi_\rho$ of $\mathscr{G}_\lambda$ lifts to a representation $d\pi_\rho^*$ of $\mathscr{E}(\mathscr{G}_\lambda)$ on $\mathscr{H}$. Let $\mathbf{Y}$ be given by

$$\mathbf{Y}=-\rho\mathbf{I}, \tag{2.6}$$

and let

$$\begin{aligned}\tilde{\mathbf{A}}_\mu^\nu=&-\lambda^{-2}\tilde{\mathbf{C}}_4^\lambda\delta_\mu^\nu+\frac{1}{2}\left[\left(\mathbf{Q}_2+\frac{1}{4}\right)\delta_\mu^\nu-\frac{3}{2}\tilde{\mathbf{L}}_\mu^\nu-\tilde{\mathbf{L}}_{\mu\rho}\tilde{\mathbf{L}}^{\rho\nu}-\mathbf{Q}_4\epsilon_{\mu\rho\tau}^\nu\tilde{\mathbf{L}}^{\rho\tau}\right]\frac{\mathbf{Y}}{\lambda}\\ &+\left[\left(\mathbf{Q}_2+\frac{1}{4}-\frac{1}{\lambda^2}\tilde{\mathbf{C}}_2^\lambda\right)\delta_\mu^\nu-\tilde{\mathbf{L}}_\mu^\nu-\tilde{\mathbf{L}}_{\mu\rho}\tilde{\mathbf{L}}^{\rho\nu}\right]\frac{\mathbf{Y}^2}{\lambda^2}-\left(\frac{1}{2}\delta_\mu^\nu-\tilde{\mathbf{L}}_\mu^\nu\right)\frac{\mathbf{Y}^3}{\lambda^3}\end{aligned} \tag{2.7}$$

and

$$\tilde{\mathbf{D}}=\left(\mathbf{Q}_4^2+\frac{1}{4}\,\mathbf{Q}_2-\frac{1}{\lambda^2}\,\tilde{\mathbf{C}}_4^{\lambda}+\frac{3}{16}\right)+\left(\mathbf{Q}_2+\frac{1}{2}\right)\frac{\mathbf{Y}}{\lambda}+\left(\mathbf{Q}_2-\frac{1}{\lambda^2}\,\tilde{\mathbf{C}}_4^{\lambda}-\frac{1}{2}\right)\frac{\mathbf{Y}^2}{\lambda^2}-\frac{2\mathbf{Y}^3}{\lambda^3}. \tag{2.8}$$

Then there is a skew symmetric representation $d\tilde{\pi}_\rho(\mathscr{P})$ of the Lie algebra $\mathscr{P}$ of the Poincaré group on $\mathscr{H}$ with the action of the generators $\mathbf{P}_\mu$ and $\mathbf{L}_{\mu\nu}$ given by

$$d\tilde{\pi}_\rho(\mathbf{P}_\mu)=d\pi_\rho^*(\tilde{\mathbf{D}})^{-1}d\pi_\rho^*(\tilde{\mathbf{A}}_\mu^\nu\tilde{\mathbf{L}}_{\nu 4}),\quad d\tilde{\pi}_\rho(\mathbf{L}_{\mu\nu})=d\pi_\rho(\tilde{\mathbf{L}}_{\mu\nu}). \tag{2.9}$$

Furthermore, the representation $d\tilde{\pi}_\rho(\mathscr{P})$ has a common invariant dense domain in the Hilbert space of the $SO_0(1,4)$ representation.

III. MATRIX ELEMENTS OF L_{ab} IN AN SO(1,4) BASIS

We now introduce a basis for the representation space of a given π_ρ, which is called the *K*-finite basis and which diagonalizes the following chain of subgroups:

$$SO(2)\subseteq SO(3)\subseteq SO(4)\subseteq SO_0(1,4). \tag{3.1}$$

This orthonormal basis is the set of all vectors

$$|klm\rangle \tag{3.2}$$

with k and l being positive integers and m is any integer such that $l\leqslant m\leqslant l$. We let $b(k+l)$ = linear span of $\{|klm>|k+l$ fixed$\}$. Then $\mathscr{H}$ is the completion of

$$\sum_{k+l=0}^{\infty}\oplus b(k+l) \tag{3.3}$$

with respect to the $SO_0(1,4)$ invariant norm on $\mathscr{H}$.[3]

The following result is well known:[6]

Theorem II:

The action of the $\mathbf{L}_{ij}$ $(i,j=1,2,3,4)$ and $\mathbf{L}_{04}$ on the basis fields of Eq. (2.4) in the representation π_ρ is given by

(i) The action of the $\mathbf{L}_{ij}$ $(i,j=0,1,2,3,4)$ is

$$d\pi_\rho(-i\mathbf{L}_{12})|klm\rangle=m|klm\rangle, \tag{3.4}$$

$$d\pi_\rho(\mathbf{L}_{31}-i\mathbf{L}_{23})|klm\rangle=(l-m)^{1/2}(l+m+1)^{1/2}|klm+1\rangle, \tag{3.5}$$

$$d\pi_\rho(-\mathbf{L}_{31}-i\mathbf{L}_{23})|klm\rangle=(l+m)^{1/2}(l-m+1)^{1/2}|klm-1\rangle, \tag{3.6}$$

$$\begin{aligned}d\pi_\rho(-i\mathbf{L}_{34})|klm\rangle=&\left(\frac{(l+m+1)(l-m+1)(k+2l+2)(k)}{(2l+1)(2l+3)}\right)^{1/2}|k-1l+1m\rangle\\&+\left(\frac{(l+m)(l-m)(k+2l+1)(k+1)}{(2l-1)(2l+1)}\right)^{1/2}|k+1l-1m\rangle,\end{aligned} \tag{3.7}$$

$$\begin{aligned}d\pi_\rho(\mathbf{L}_{24}-i\mathbf{L}_{14})|klm\rangle=&-\left(\frac{(l+m+1)(l+m+2)(k+2l+2)(k)}{(2l+1)(2l+3)}\right)^{1/2}|k-1l+1m+1\rangle\\&+\left(\frac{(l-m)(l-m-1)(k+2l+1)(k+1)}{(2l+1)(2l-1)}\right)^{1/2}|k+1l-1m+1\rangle,\end{aligned} \tag{3.8}$$

and

$$d\pi_\rho(\mathbf{L}_{24}+i\mathbf{L}_{14})|klm\rangle=-\left(\frac{(l-m+1)(l-m+2)(k+2l+2)(k)}{(2l+1)(2l+3)}\right)^{1/2}|k-1\,l+1\,m-1\rangle$$
$$+\left(\frac{(l+m)(l+m-1)(k+2l+1)(k+1)}{(2l-1)(2l+1)}\right)^{1/2}|k+1\,l-1\,m-1\rangle. \tag{3.9}$$

(ii) The action of $\mathbf{L}_{04}$ is:

$$d\pi_\rho(\mathbf{L}_{04})|klm\rangle=\frac{1}{2}\{k+l+3/2+i\rho\}\left[\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}\right]^{1/2}|k+1\,lm\rangle$$
$$-\frac{1}{2}\{k+l+1/2-i\rho\}\left[\frac{(k)(k+2l+1)}{(k+l+1)(k+l)}\right]^{1/2}|k-1\,lm\rangle. \tag{3.10}$$

IV. MATRIX ELEMENTS OF THE POINCARÉ TRANSLATION GENERATORS IN AN SO(1,4) BASIS

We now consider the application of these results to the problem of determining the matrix elements of a basis for the Poincaré Lie algebra in the above K-finite basis. In order to determine the action of $\mathbf{P}_0$ in this basis we use Theorem I. From Eqs. (1.6) and (1.7) of Theorem I together with Theorem II we find

$$\begin{aligned}
d\pi_\rho^*(\tilde{\mathbf{A}}_0^\nu\tilde{\mathbf{L}}_{\nu 4})|klm\rangle=&-i\rho\Bigg\{\Bigg[\frac{1}{8}\left(\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)\\
&+\frac{1}{16}\left(k+l+\frac{3}{2}+i\rho\right)\left(-\rho^2-\frac{9}{4}-(k+l+1)(k+l+3)\right)\\
&\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\Bigg)-\frac{1}{16}\left(\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)\right.\\
&\left.\times\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)\Bigg]|k+1\,lm\rangle\\
&+\Bigg[-\frac{1}{8}\left(k+l+\frac{1}{2}-i\rho\right)\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\\
&-\frac{1}{16}\left(\left(k+l+\frac{1}{2}-i\rho\right)\left(-\rho^2-\frac{9}{4}-(k+l-1)(k+l+1)\right)\right.\\
&\left.\times\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\right)\\
&+\frac{1}{16}\left(\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)\left(k+l+\frac{1}{2}-i\rho\right)\right.\\
&\left.\times\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\right)\Bigg]|k-1\,lm\rangle\Bigg\}\\
&-\rho^2\Bigg\{\Bigg[\frac{5}{8}\left(\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)\\
&+\frac{1}{4}\left(\left(k+l+\frac{3}{2}+i\rho\right)\left(-\rho^2-\frac{9}{4}-(k+l+1)(k+l+3)\right)\right.
\end{aligned}$$

$$\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}-\frac{1}{4}\left(\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)\right.$$
$$\left.\times\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)|k+1lm\rangle$$
$$+\left[-\frac{5}{8}\left(k+l+\frac{1}{2}-i\rho\right)\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}-\frac{1}{4}\left(\left(k+l+\frac{1}{2}-i\rho\right)\right.\right.$$
$$\left.\times\left(-\rho^2-\frac{9}{4}-(k+l-1)(k+l+1)\right)\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\right)$$
$$+\frac{1}{4}\left(\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)\left(k+l+\frac{1}{2}-i\rho\right)\right.$$
$$\left.\left.\times\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\right)\right]|k-1lm\rangle\Bigg\}+i\rho^3\Bigg\{\Bigg[\left(\left(k+l+\frac{3}{2}+i\rho\right)\right.$$
$$\left.\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)$$
$$+\frac{1}{4}\left(\left(k+l+\frac{3}{2}+i\rho\right)\left(-\rho^2-\frac{9}{4}-(k+l+1)(k+l+3)\right)\right.$$
$$\left.\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)-\frac{1}{4}\left(k+l+\frac{3}{2}+i\rho\right)$$
$$\times\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)$$
$$\left.\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\right)\Bigg]|k+1lm\rangle+\Bigg[-\left(k+l+\frac{1}{2}-i\rho\right)$$
$$\times\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}-\frac{1}{4}\left(\left(k+l+\frac{1}{2}-i\rho\right)\right.$$
$$\times\left(-\rho^2-\frac{9}{4}-(k+l-1)(k+l+1)\right)$$
$$\times\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}+\frac{1}{4}\left(\left(k+l+\frac{1}{2}-i\rho\right)\right.$$
$$\left.\left.\times\left(-\rho^2-\frac{9}{4}-(k+l)(k+l+2)\right)\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\right)\right]|k-1lm\rangle\Bigg\}$$
$$+\rho^4\left\{\frac{1}{2}\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\,|k+1lm\rangle\right.$$
$$\left.-\frac{1}{2}\left(k+l+\frac{1}{2}-i\rho\right)\sqrt{\frac{k(k+2l+1)}{(k+l+1)(k+l)}}\,|k-1lm\rangle\right\}. \tag{4.1}$$

Now from Eq. (2.8) we easily obtain

$$\tilde{\mathbf{D}}=\left(-i\rho+\frac{1}{2}\right)^2\left[\mathbf{Q}_2+\frac{3}{2}+\rho^2-i\rho\right] \tag{4.2}$$

and a straightforward application of Theorem II yields

$$
\begin{aligned}
d\widetilde{\pi}_\rho(\widetilde{\mathbf{D}})|klm\rangle = \Bigg(& +i\rho - \frac{3}{2} + \frac{1}{4}\frac{\left(k+l+\frac{3}{2}+i\rho\right)(k+1)(k+2l+2)\left(k+\frac{3}{2}+l-i\rho\right)}{(k+l+2)(k+l+1)} \\
& + \frac{1}{4}\frac{\left(k+\frac{1}{2}+l-s\right)k(k+1+2l)\left(k+l+\frac{1}{2}-i\rho\right)}{(k+l+1)(k+l)} \\
& - (k+l)(k+l+2)+l(l+1)\Bigg)|klm\rangle \\
& - \frac{1}{4}\left\{\left(k+l+\frac{3}{2}+i\rho\right)\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}\left(k+\frac{5}{2}+l+i\rho\right)\right. \\
& \left.\times\sqrt{\frac{(k+2)(k+3+2l)}{(k+3+l)(k+l+2)}}\right\}|k+2lm\rangle - \frac{1}{4}\left\{\left(k+l+\frac{1}{2}-i\rho\right)\right. \\
& \left.\times\sqrt{\frac{k(k+1+2l)}{(k+l+1)(k+l)}}\left(k-\frac{1}{2}+l-i\rho\right)\sqrt{\frac{(k-1)(k+2l)}{(k+l)(k-1+l)}}\right\}|k-2lm\rangle.
\end{aligned}
\tag{4.3}
$$

We now determine $d\widetilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle$ by assuming an expansion in a series of the form

$$
d\widetilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle = \sum_{k'=k+1,k+3,\ldots}^{\infty} c^+_{lm}(k,k')|k'lm\rangle \tag{4.4}
$$

for even k, and an expansion of the form

$$
d\widetilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle = \sum_{k'=0,2,\ldots}^{k-1} c^-_{lm}(k,k')|k'lm\rangle \tag{4.5}
$$

for odd k. [It is clear from Eq. (2.4) that $d\widetilde{\pi}_\rho(\mathbf{P}_0)$ cannot change l and m, and it is clear from (2.5) together with (3.10) and (4.3) that that states $|k'lm\rangle$ with $k'=k+n$ where n is an even integer cannot occur in these expansions.] The following equation is an immediate consequence of Eq. (2.9):

$$
d\widetilde{\pi}_\rho(\widetilde{\mathbf{D}})d\widetilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle = d\pi^*_\rho(\widetilde{\mathbf{A}}^\nu_\mu\widetilde{\mathbf{L}}_{\nu 4})|klm\rangle. \tag{4.6}
$$

By substituting the expansions (4.4) and (4.5) into this equation we obtain a three term recurrence relation for the expansion coefficients $c^\pm_{lm}(k,k')$. We now treat in detail the case of even k. For $k'>k+3$ we obtain:

$$
\begin{aligned}
& \frac{1}{4}\left(k'+\frac{5}{2}+l-i\pi\right)\left(k'+\frac{3}{2}+l-i\pi\right) \\
& \quad\times\sqrt{\frac{(k'+2)(k'+2l+3)(k'+1)(k'+2l+2)}{(k'+l+3)(k'+l+2)(k'+l+2)(k'+l+1)}}\,c^+_{lm}(k,k'+2) \\
& = \Bigg\{\left[-\frac{9}{4}+\rho^2\right]-(k'+l)(k'+l+2)+l(l+1) \\
& \quad + \frac{1}{4}\frac{\left(k'+\frac{3}{2}+l+i\pi\right)\left(k'+\frac{3}{2}+l-i\pi\right)(k'+1)(k'+2l+2)}{(k'+l+2)(k'+l+1)}
\end{aligned}
$$

$$+\frac{1}{4}\frac{\left(k'+\frac{1}{2}+l+i\pi\right)\left(k'+\frac{1}{2}+l-i\pi\right)(k')(k'+2l+1)}{(k'+l+1)(k'+l)}\Bigg\}\, c^{+}_{lm}(k,k')$$
$$-\frac{1}{4}\left(k'-\frac{1}{2}+l+i\pi\right)\left(k'+\frac{1}{2}+l+i\pi\right)$$
$$\times\sqrt{\frac{(k'-1)(k'+2l)(k')(k'+2l+1)}{(k'+l)(k'+l-1)(k'+l+1)(k'+l)}}\, c^{+}_{lm}(k,k'-2). \tag{4.7}$$

Now $c^{+}_{lm}(k,k'=1)$ and $c^{+}_{lm}(k,k'=3)$ are easily determined from Eq. (4.6) as special cases: evaluating the right hand side of (4.6) with the help of (4.1) gives

$$(|k-1lm\rangle, d\tilde{\pi}_{\rho}(\tilde{\mathbf{D}})d\tilde{\pi}_{\rho}(\mathbf{P}_0)|klm\rangle) = \frac{i\rho}{2}\left(\frac{1}{2}-i\rho\right)^2\left(k+l+\frac{1}{2}-i\rho\right)\left(k+l+\frac{3}{2}-i\rho\right)$$
$$\times\sqrt{\frac{(k)(k+2l+1)}{(k+l+1)(k+l)}}, \tag{4.8}$$

and, using Eqs. (4.2) and (4.3), we obtain

$$(|k-1lm\rangle, d\tilde{\pi}_{\rho}(\tilde{\mathbf{D}})d\tilde{\pi}_{\rho}(\mathbf{P}_0)|klm\rangle) = -\frac{1}{4}\left(\frac{1}{2}-i\rho\right)^2\left(k+l+\frac{1}{2}-i\rho\right)\left(k+l+\frac{3}{2}-i\rho\right)$$
$$\times\sqrt{\frac{(k)(k+1)(k+2l+1)(k+2l+2)}{(k+l+1)^2(k+l)(k+l+2)}}\, c^{+}_{lm}(k,k+1). \tag{4.9}$$

Thus by equating the right hand sides of Eqs. (4.8) and (4.9) and then solving the resulting equation for $c^{+}_{lm}(k,k+1)$, we obtain:

$$c^{+}_{lm}(k,k+1) = -2i\rho\sqrt{\frac{(k+l+1)(k+l+2)}{(k+2l+2)(k+1)}}. \tag{4.10}$$

Similarly, we have

$$(|k+1lm\rangle, d\tilde{\pi}_{\rho}(\tilde{\mathbf{D}})d\tilde{\pi}_{\rho}(\mathbf{P}_0)|klm\rangle) = \frac{i\rho}{2}\left(\frac{1}{2}-i\rho\right)^2\left(k+l+\frac{1}{2}+i\rho\right)\left(k+l+\frac{3}{2}+i\rho\right)$$
$$\times\sqrt{\frac{(k+1)(k+2l+2)}{(k+l+2)(k+l+1)}}, \tag{4.11}$$

and

$$(|k+1lm\rangle, d\tilde{\pi}_{\rho}(\tilde{\mathbf{D}})d\tilde{\pi}_{\rho}(\mathbf{P}_0)|klm\rangle) = -\frac{1}{4}\left(\frac{1}{2}-i\rho\right)^2\left(k+l+\frac{9}{2}-i\rho\right)\left(k+l+\frac{7}{2}-i\rho\right)$$
$$\times\sqrt{\frac{(k+3)(k+4)(k+2l+4)(k+2l+5)}{(k+l+4)^2(k+l+3)(k+l+5)}}$$
$$\times c^{+}_{lm}(k,k+3)+\left(\frac{1}{2}-i\rho\right)^2\Bigg\{-\left[\frac{9}{4}+\rho^2\right]$$
$$-(k+l+1)(k+l+3)+l(l+1)$$
$$+\frac{1}{4}\Bigg[\left\{\left(k+l+\frac{5}{2}\right)^2+i\rho^2\right\}\frac{(k+2)(k+2l+3)}{(k+l+3)(k+l+2)}$$

$$+\left\{\left(k+l+\frac{3}{2}\right)^2+i\rho^2\right\}\frac{(k+1)(k+2l+2)}{(k+l+1)(k+l+2)}\Bigg]\Bigg\}$$
$$\times c^+_{lm}(k,k+1). \tag{4.12}$$

By substituting $c^+_{lm}(k,k+1)$ from Eq. (4.10) into this equation and substituting the right hand side of Eq. (4.11) into the left hand side of this equation we obtain a formula for $c^+_{lm}(k,k+3)$, which simplifies to

$$c^+_{lm}(k,k+3)=+2i\rho\sqrt{\frac{(k+2)(k+2l+3)(k+l+1)(k+l+4)}{(k+2l+4)(k+2l+2)(k+3)(k+1)}}. \tag{4.13}$$

With the initial values of $c^+_{lm}(k,k+1)$ and $c^+_{lm}(k,k+3)$ given in Eqs. (4.10) and (4.13) we solve by successive iteration for higher $c^+_{lm}(k,k+n)$ and then readily establish the following result by mathematical induction:

$$\begin{aligned}c^+_{lm}(k,k+n)&=(-1)^{(n+1)/2}2i\rho\\&\quad\times\left[\frac{k+2)(k+4)\ldots(k+n-1)(k+2l+3)(k+2l+5)\ldots(k+2l+n)}{(k+1)(k+3)\ldots(k+n)(k+2l+2)(k+2l+4)\ldots(k+2l+n+1)}\right]^{1/2}\\&\quad\times\{(k+l+1)(k+l+n+1)\}^{1/2}\\&=2(-1)^{(1/2n+1/2)}i\rho\Bigg(\Gamma\left(\frac{1}{2}n+\frac{1}{2}+\frac{1}{2}k\right)\Gamma\left(\frac{1}{2}n+1+\frac{1}{2}k+l\right)\Gamma\left(\frac{1}{2}k+\frac{1}{2}\right)\\&\quad\times\Gamma\left(\frac{1}{2}k+l+1\right)(k+l+1)(k+l+n+1)\Bigg)\Bigg/\Bigg(\Gamma\left(\frac{1}{2}k\right)\Gamma\left(\frac{1}{2}k+l+\frac{1}{2}\right)\\&\quad\times\Gamma\left(\frac{1}{2}n+1+\frac{1}{2}k\right)\Gamma\left(\frac{1}{2}n+\frac{3}{2}+\frac{1}{2}k+l\right)k(k+2l+1)\Bigg)\Bigg)^{1/2}.\end{aligned} \tag{4.14}$$

Determination of the coefficients $c^-_{lm}(k,k')$ proceeds in a completely similar manner. The details of this are left to the reader. The final result is:

$$\begin{aligned}c^-_{lm}(k,k-n)&=-(-1)^{(n+1)/2}2i\rho\\&\quad\times\left[\frac{k-n+2)(k-n+4)\ldots(k-1)(k-n+2l+3)(k-n+2l+5)\ldots(k+2l)}{(k-n+1)(k-n+3)\ldots(k)(k-n+2l+2)(k-n+2l+4)\ldots(k+2l+1)}\right]^{1/2}\\&\quad\times\{(k-n+l+1)(k+l+1)\}^{1/2}.\end{aligned} \tag{4.15}$$

Evaluation of the action of $d\tilde{\pi}_\rho(\mathbf{P}_i)$ $(i=1,2,3)$ on the $|klm\rangle$ follows from the commutation relations $\mathbf{P}_i=[\mathbf{L}_{0i},\mathbf{P}_0]$. Such expressions are now determined straightforwardly, but since the answers are rather lengthy we do not bother to write them down here.

V. CONCLUSIONS

In the above we have determined the matrix elements of a set of generators for the Poincaré Lie algebra. ($\mathbf{L}_{ij},\mathbf{L}_{0i}$, and $\mathbf{P}_0$ generate $\mathscr{P}$.) It is easily seen from Eqs. (4.14) and (4.15) that

$$(|klm\rangle,d\tilde{\pi}_\rho(\mathbf{P}_0)|k'l'm'\rangle)=-(d\tilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle,|k'l'm'\rangle). \tag{5.1}$$

This equation is to be expected of a skew-symmetric operator on a Hilbert space, but, according to the authors of Ref. 4, such an equation does not hold for matrix elements of $d\tilde{\pi}_\rho(\mathbf{P}_0)$ on ''Lorentz basis'' vectors. Although our analysis does not shed any light on a possible explanation of the strange behavior of operators on ''distributional bases'' which is described in that paper, our results do show that one obtains a much cleaner analysis, when working with ''bona fida'' elements of the Hilbert space. For example we easily convince ourselves that $d\tilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle$ for k odd is in the Hilbert space, and $d\tilde{\pi}_\rho(\mathbf{P}_0)|klm\rangle$ for k even lies outside of the Hilbert space.

ACKNOWLEDGMENTS

P.M. thanks the J. W. Fulbright Foundation and the Czech Technical University for financial support. We acknowledge the help of SWI-Prolog and Maple-V for some of the calculations.

[1] M. Havlicek and P. Moylan, J. Math. Phys. **34**, 5320 (1993).
[2] M. Havlicek and P. Moylan, Proceedings of the 21st International Conference on Group Theoretical Methods in Physics, Goslar, Germany, 13–20 July 1996.
[3] Robert S. Strichartz, J. Funct. Anal. **114**, 493 (1993).
[4] A. Chakrabarti, M. Levy-Nahas, and R. Seneor, J. Math. Phys. **9**, 1274 (1968).
[5] H. Joos, Fortschr. Phys. **10**, 65 (1963).
[6] P. Moylan, J. Math. Phys. **24**, 2706 (1983).

Free boson representation of $DY_\hbar(gl_N)_k$ and $DY_\hbar(sl_N)_k$

X. M. Ding[a)]
CCAST, P.O. Box 8730, Beijing, 100080, People's Republic of China and Institute of Theoretical Physics of Acedimia Sinica, P.O. Box 2735, Beijing, 100080, People's Republic of China

B. Y. Hou[b)]
Institute of Modern Physics, Northwest University, Xian, 710069, People's Republic of China

B. Yuan Hou
Physics Department of Graduate School of Acedimia Sinica, Beijing, 100039, People's Republic of China

L. Zhao[c)]
Institute of Modern Physics, Northwest University, Xian, 710069, People's Republic of China

(Received 9 September 1997; accepted for publication 15 December 1997)

We construct a realization of the Yangian double $DY_\hbar(gl_N)_k$ and $DY_\hbar(sl_N)_k$ for an arbitrary level k in terms of free boson fields with a continuous parameter. In the $\hbar \to 0$ limit this realization becomes the Wakimoto realization of affine Lie algebras gl_N and sl_N, respectively. The vertex operators and the screening currents are also constructed with the same spirits; the screening currents commute with $DY_\hbar(sl_N)$ modulo total difference. © *1998 American Institute of Physics.* [S0022-2488(98)00104-2]

I. INTRODUCTION

Quantum algebra and Yangian were proposed by Drinfeld as generalizations of universal enveloping algebras of classical Lie algebra and loop algebra with nontrivial Hopf algebra structures.[1–3] Following the Faddeev–Reshetikhin–Takhtajan (FRT) formalism,[4] both kinds of algebras can be considered as associative algebras defined through the Yang–Baxter relation (i.e., RLL relations) with the structure constants determined by the solutions of the quantum Yang–Baxter equation (QYBE). Quantum affine algebra[3] and Yangian double with center (the central extension of the quantum double of Yangian)[5–8] are affine extensions of universal enveloping algebras of classical Lie algebra and Yangian. The latter algebras can also be viewed from the RLL viewpoint, with the R matrix being the trigonometric solution of QYBE and rational one, respectively, if one regards the Reshetikhin–Semenov–Tian–Shansky realization (RS)[9] as the affine analog of Faddeev–Reshetikhin–Takhtajan formalism. They both were proved to have important applications in certain physical problems, especially in describing the dynamical symmetries of the integrable massive quantum field theories, calculating the correlation functions and form factors of some two-dimensional exactly solvable lattice statistical model and $(1+1)$-dimensional completely integrable quantum field theories.[10–12] In these cases, they play crucial role, which is similar to the role of the Virasoro algebras and affine Lie algebras in the conformal field theories. In such physical applications, the infinite-dimensional representations of quantum affine algebra and Yangian double are required. In another words, Yangian double with central is needed, especially the representations with higher $(k>1)$ level.

In practice, free field realization is proved to be quite effective and useful approach to study complicated algebras structure. In this aspect, the Heisenberg algebra (or free boson) representation has become a common method for obtaining representations of quantum affine algebras. For examples, the free boson representations of $U_q(\widehat{sl}_2)$ with an arbitrary level have been obtained in

a)Electronic mail: xmding@itp.ac.cn
b)Electronic mail: byhou@nwu.edu.cn
c)Electronic mail: lzhao@nwu.edu.cn

 © 1998 American Institute of Physics

Refs. 13–17. Free boson representation of $U_q(\widehat{sl}_N)$ with level 1 was constructed in Ref. 18. Free boson representations of $U_q(\widehat{sl}_3)$ and $U_q(\widehat{sl}_N)$ with arbitrary level were constructed in Refs. 19 and 20, respectively. For the Yangian double with center, the free field representation of $DY_\hbar(sl_2)$ with level k was constructed in Ref. 21. The level 1 free boson representation of $DY_\hbar(gl_N)$ and $DY_\hbar(sl_N)$ was given in Ref. 22. In Ref. 23 the free field representation of the Yangian double $DY_\hbar(sl_N)$ with arbitrary level k are obtained by a simple correspondence between the quantum affine algebra $U_q(\widehat{sl_N})$ and the Yangian double $DY_\hbar(sl_N)$. This correspondence makes our derivation of free field representation for $DY_\hbar(sl_N)$ greatly simplified. However, we have been unable to obtain the vertex operators and screening currents for $DY_\hbar(sl_N)$ following the same spirit, so the relations of quantum affine algebras and the Yangian double with center is nontrivial. Construct a free field realization of the screening current operator is one of the motivations of this paper.

Another motivation is that, in Ref. 24, Frenkel and Reshetikhin used the free field realization of the quantum affine algebra $U_q(\widehat{gl_2})_k$ at critical level $k=-2$, to investigate the structure of the infinite-dimensional center of formal completion of the quantum affine algebra $U_q(\widehat{gl_2})_k$. Using the Wakimoto realization of the quantum affine algebra they define a new Poisson bracket algebra, which is nothing but the q-deformation of the classical Virasoro algebra ${}_qVir$ (in the general case are the q-deformations of the classical $\mathscr{W}_q(sl_N)$ algebras). Similarly to the quantum affine cases, infinite-dimensional centers exist for the Yangian double with center at critical level, which are just the classical $\mathscr{W}_\hbar(sl_N)$ algebras.[25] In the simplest case, ${}_\hbar$Virasoro is considered in Ref. 26 by using the free field realization of the Yangian double with center $DY_\hbar(gl_2)_k$ at level $k=-2$. Similarly, one way to study the structure of the $\mathscr{W}_\hbar(g)$ algebras is through the free field realization of the Yangian double with center $DY_\hbar(g)_k$ at critical level, where the Lie algebra g has a higher rank. However, it is difficult to obtain the free field realization of the currents for Yangian double $DY_\hbar(gl_N)$ in the higher rank cases (rank greater than one) with the boson fields using in Ref. 23, either no consistent realization of quantum determinant, which decomposes the $DY_\hbar(gl_N)_c$ to its subalgebra $DY_\hbar(sl_N)_c$ could be given.

In this paper, we introduce free fields involved with a continuous parameter, and we obtain the $DY_\hbar(gl_N)$ and $DY_\hbar(sl_N)$ currents, and the screening currents and vertex operators.

The manuscript is arranged as follows. In Sec. II we first briefly review the Drinfeld new realization of the Yangian $Y(g)$; then we give the defining relations of Drinfeld generators of the Yangian double with center $DY_\hbar(gl_N)_k$. The RS realization is also given, and the isomorphism between these two realization can be established through Ding–Frenkel correspondence. In Sec. III the three set of free boson fields with a continuous parameter are introduced; some relations, which will be used in the forthcoming sections, are listed. With the help of the results obtained in Sec. III, we give a free fields realization of $DY_\hbar(gl_N)_k$ and $DY_\hbar(sl_N)_k$ in Sec. IV. In Sec. V we give the free field realization of the screening currents and the vertex operators, and brief discussions are given at the final section of the paper.

II. DRINFELD GENERATORS OF YANGIAN DOUBLE $DY_\hbar(gl_N)$

First, we introduce some notations, which will be used in the sequel. Let $\boldsymbol{\epsilon}_\mu$ $(1\leqslant\mu\leqslant N)$ be the orthonormal basis in R^N. We have the inner product $\langle\boldsymbol{\epsilon}_\mu,\boldsymbol{\epsilon}_\nu\rangle=\delta_{\mu,\nu}$. Define

$$\bar{\boldsymbol{\epsilon}}_\mu=\boldsymbol{\epsilon}_\mu-\boldsymbol{\epsilon},\quad \boldsymbol{\epsilon}=\frac{1}{N}\sum_{\mu=1}^{N}\boldsymbol{\epsilon}_\mu. \tag{2.1}$$

Obviously, $\Sigma_{\mu=1}^{N}\bar{\boldsymbol{\epsilon}}_\mu=0$. The simple roots, fundamental weights are,

$$\alpha_\mu=\boldsymbol{\epsilon}_\mu-\boldsymbol{\epsilon}_{\mu+1},\quad \Lambda_\mu=\sum_{\nu=1}^{\mu}\bar{\boldsymbol{\epsilon}}_\nu, \tag{2.2}$$

respectively. For simple, denote half of the Cartan matrix $a_{\mu,\nu}$ as $B_{\mu,\nu}$.

There are three realizations of Yangian $Y(g)$: Drinfeld realization,[1,2] Drinfeld new realization,[3] and R-matrix realization or FRT realization.[27,4] The isomorphism between the Drinfeld new realization and the R-matrix (or FRT) realization can be established by using the Ding

and Frenkel correspondence.[28] Here, we only give the definition of the Drinfeld new realization of Yangian, for the Yangian double is defined in terms of quantum double of the Drinfeld new realization, while the Yangian double with center is the central extension of the Yangian double. There is the following theorem:

Theorem 1:[3] *The Yangian $Y(g)$ is isomorphic to the associative algebra with generators, $x^{\pm}_{\mu,r}$, $h_{\mu,r}$, $\mu=1,...,$ rank g, $r\in Z_{\geqslant 0}$, and the following defining relations:*

$$[h_{\mu,r}h_{\nu,s}]=0,$$

$$[h_{\mu,0},x^{\pm}_{\nu,s}]=\pm 2d_{\mu}B_{\mu,\nu}x^{\pm}_{\mu,s},$$

$$[h_{\mu,r+1},x^{\pm}_{\nu,s}]-[h_{\mu,r},x^{\pm}_{\nu,s+1}]=\pm\hbar d_{\mu}B_{\mu,\nu}(h_{\mu,r}x^{\pm}_{\nu,s}+x^{\pm}_{\nu,s}h_{\mu,r}),$$

$$[x^{+}_{\mu,r},x^{-}_{\nu,s}]=\delta_{\mu,\nu}h_{\mu,r+s},$$

$$[x^{\pm}_{\mu,r+1},x^{\pm}_{\nu,s}]-[x^{\pm}_{\mu,r},x^{\pm}_{\nu,s+1}]=\pm\hbar d_{\mu}B_{\mu,\nu}(x^{\pm}_{\mu,r}x^{\pm}_{\nu,s}+x^{\pm}_{\nu,s}x^{\pm}_{\mu,r}),$$

$$\sum_{\pi}[x^{\pm}_{\mu,r_{\pi(1)}},[x^{\pm}_{\mu,r_{\pi(2)}},\ldots,[x^{\pm}_{\mu,r_{\pi(m)}},x^{\pm}_{\nu,s}]\ldots]]=0,$$

for all sequences of non-negative integers $r_1,\ldots,r_{\nu}$, where $m=1-2B_{\mu,\nu}$, $2d_{\mu}B_{\mu,\nu}$ is the symmetric Cartan matrix. And the sum is over all permutations π of $1,\ldots,m$.

As an associative algebra, the Yangian double with center $DY_{\hbar}(gl_N)_c$ is generated by the Drinfeld generators $\{k^{+}_{\mu,r},k^{-}_{\mu,s}x^{+}_{\nu,m},\ x^{-}_{\nu,m}|\mu=1,2,...,N;\ \nu=1,2,...,N-1;\ r\in Z_{\geqslant 0},\ s\in Z_{<0},\ m\in Z\}$, derivation operator d and the center c.

The generating function of the Drinfeld generators of Yangian double with center $DY_{\hbar}(gl_N)_c$ are $k^{\pm}_{\mu}(u)$ and $X^{\pm}_{\nu}(u)$,

$$k^{+}_{\mu}(u)=\sum_{r\geqslant 0}k_{\mu,r}u^{-r-1},\quad k^{-}_{\mu}(u)=\sum_{s<0}h_{\nu,s}u^{-s-1},$$

$$X^{+}_{\mu}(u)=\sum_{m\in Z}x^{+}_{\mu,m}u^{-m-1},\quad X^{-}_{\mu}(u)=\sum_{m\in Z}x^{-}_{\mu,m}u^{-m-1},$$

and they satisfy the following commutation relations:

$$[d,\chi(u)]=\frac{dX}{du},\quad [c,\chi(u)]=0,\quad \chi(u)=k^{\pm}_{\mu}(u),\quad X^{+}_{\mu}(u),\quad X^{-}_{\mu}(u),$$

$$k^{\pm}_{\mu}(u)k^{\pm}_{\nu}(v)=k^{\pm}_{\nu}(v)k^{\pm}_{\mu}(u),$$

$$\rho(u-v-c\hbar)k^{+}_{\mu}(u)k^{-}_{\mu}(v)=k^{-}_{\mu}(v)k^{+}_{\mu}(u)\rho(u-v),$$

$$\rho(u-v-c\hbar)\frac{u-v-c\hbar}{u-v+\hbar-c\hbar}k^{+}_{\mu}(u)k^{-}_{\nu}(v)=k^{-}_{\nu}(v)k^{+}_{\mu}(u)\frac{u-v}{u-v+\hbar}\rho(u-v),\quad \mu<\nu,$$

$$\rho(u-v-c\hbar)\frac{u-v-\hbar-c\hbar}{u-v-c\hbar}k^{+}_{\mu}(u)k^{-}_{\nu}(v)=k^{-}_{\nu}(v)k^{+}_{\mu}(u)\frac{u-v-\hbar}{u-v}\rho(u-v),\quad \mu>\nu,$$

$$k^{\pm}_{\mu}(u)X^{+}_{\mu}(v)k^{\pm}_{\mu}(u)^{-1}=\frac{u-v}{u-v+\hbar}X^{+}_{\mu}(v),$$

$$k^{\pm}_{\mu+1}(u)X^{+}_{\mu}(v)k^{\pm}_{\mu+1}(u)^{-1}=\frac{u-v}{u-v-\hbar}X^{+}_{\mu}(v),$$

$$k_\mu^-(u)X_\mu^-(v)k_\mu^-(u)^{-1}=\frac{u-v+\hbar}{u-v}\,X_\mu^-(v),$$

$$k_{\mu+1}^-(u)X_\mu^-(v)_{\mu+1}^-(u)^{-1}=\frac{u-v-\hbar}{u-v}\,X_\mu^-(v),$$

$$k_\mu^+(u)X_\mu^-(v)k_\mu^+(u)^{-1}=\frac{u-v+\hbar-c\hbar}{u-v-c\hbar}\,X_\mu^-(v),$$

$$k_{\mu+1}^+(u)X_\mu^-(v)k_{\mu+1}^+(u)^{-1}=\frac{u-v-\hbar-c\hbar}{u-v-c\hbar}\,X_\mu^-(v), \tag{2.3}$$

$$k_\mu^\pm(u)X_\nu^+(v)k_\mu^\pm(u)^{-1}=X_\nu^+(v),\quad \mu-\nu\geqslant 2,\ \text{or}\ \ \mu-\nu\leqslant -1,$$

$$k_\mu^\pm(u)X_\nu^-(v)k_\mu^\pm(u)^{-1}=X_\nu^-(v),\quad \mu-\nu\geqslant 2,\ \text{or}\ \ \mu-\nu\leqslant -1,$$

$$(u-v\mp\hbar)X_\mu^\pm(u)X_\mu^\pm(v)=(u-v\pm\hbar)X_\mu^\pm(v)X_\mu^\pm(u),$$

$$(u-v+\hbar)X_\mu^+(u)X_{\mu+1}^+(v)=(u-v)X_{\mu+1}^+(v)X_\mu^+(u),$$

$$(u-v)X_\mu^-(u)X_{\mu+1}^-(v)=(u-v+\hbar)X_{\mu+1}^-(v)X_\mu^-(u),$$

$$X_\mu^\pm(u_1)X_\mu^\pm(u_2)X_\nu^\pm(v)-2X_\mu^\pm(u_1)X_\nu^\pm(v)X_\mu^\pm(u_2)+X_\nu^\pm(v)X_\mu^\pm(u_1)X_\mu^\pm(u_2)+\{u_1\leftrightarrow u_2\}=0,$$

$$|\mu-\nu|=1,$$

$$X_\mu^\pm(u)X_\nu^\pm(v)=X_\nu^\pm(v)X_\mu^\pm(u),\quad |\mu-\nu|\geqslant 2,$$

$$[X_\mu^+(u),X_\nu^-(v)]=\delta_{\mu,\nu}(\delta(u-(v+c\hbar))k_{\mu+1}^+(u)k_\mu^+(u)^{-1}-\delta(u-v)k_{\mu+1}^-(v)k_\mu^-(v)^{-1}),$$

in which the scalar factor

$$\rho^+(u)=\frac{\Gamma\left(\frac{1}{N}-\frac{u}{N\hbar}\right)\Gamma\left(1-\frac{1}{N}-\frac{u}{N\hbar}\right)}{\Gamma\left(-\frac{u}{N\hbar}\right)\Gamma\left(1-\frac{u}{N\hbar}\right)}, \tag{2.4}$$

and $\delta(u-v)=\Sigma_{m+n=-1}u^m v^n$. These relations are understood as relations between formal series. The above relations can be derived through the RS realization; in the following we will turn to the RS realization.

The RS realization[9] originated from the quantum inverse scattering method. One of the advantages of the RS realization is that the central elements can be written explicitly by the quantum trace.[9] For this purpose, in the following we will drop the d operator in the Yangian double with center, and denote as $DY_\hbar'(gl_N)$. Introduce the following operators:

$$L^\pm(u)=(l_{\mu\nu}^\pm(u))_{1\leqslant\mu,\nu\leqslant N}; \tag{2.5}$$

where the generating function,

$$l_{\mu\nu}^+(u)=\delta_{\mu,\nu}+\hbar\sum_{m=0}^{\infty}l_{\mu,\nu}^+[m]u^{-m-1};\quad l_{\mu\nu}^-(u)=\delta_{\mu,\nu}-\hbar\sum_{m=-1}^{-\infty}l_{\mu,\nu}^-[m]u^{-m-1}; \tag{2.6}$$

and the R-matrix $R^\pm(u)$,

$$R^\pm(u)=\rho^\pm(u)\overline{R}(u), \tag{2.7}$$

$$\rho^{-}(u)=\left(\frac{\Gamma(1/N+u/N\hbar)\Gamma(1-1/N+u/N\hbar)}{\Gamma(u/N\hbar)\Gamma(1+u/N\hbar)}\right)^{-1}=\frac{1}{\rho^{+}(-u)}, \tag{2.8}$$

and $\overline{R}(u)=(uI+\hbar P)/(u+\hbar)$ is the Yang's solution of QYBE. The P is the flip operator of $u\otimes v$, $Pu\otimes v=v\otimes u$, and the choice of the scalar function $\rho^{\pm}(u)$ can be determined by the unitarity condition and crossing symmetry.[29] But here we adopt another approach, which will be explained later in this paper.

From the universal R-matrix for the central extended Yangian double obtained in Refs. 5–7, $L^{\pm}(u)$ have the following unique decompositions:

$$L^{+}(u)=\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \hbar f^{+}_{2,1}(u-c\hbar) & 1 & & \\ & \vdots & \ddots & \ddots \\ \hbar f^{+}_{N,1}(u-c\hbar) & \cdots & \hbar f^{+}_{N,N-1}(u-c\hbar) & 1 \end{pmatrix}\begin{pmatrix} k^{+}_{1}(u) & 0 & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & \cdots & k^{+}_{N}(u) \end{pmatrix}$$
$$\times\begin{pmatrix} 1 & \hbar e^{+}_{1,2}(u) & \cdots & \hbar e^{+}_{1,N}(u) \\ & \ddots & \ddots & \\ 0 & & \ddots & \hbar e^{+}_{N-1,N}(u) \\ 0 & \cdots & \cdots & 1 \end{pmatrix},$$

$$L^{-}(u)=\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \hbar f^{-}_{2,1}(u) & 1 & & \\ & \vdots & \ddots & \ddots \\ \hbar f^{-}_{N,1}(u) & \cdots & \hbar f^{-}_{N,N-1}(u) & 1 \end{pmatrix}\begin{pmatrix} k^{-}_{1}(u) & 0 & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & \cdots & k^{-}_{N}(u) \end{pmatrix}$$
$$\times\begin{pmatrix} 1 & \hbar e^{-}_{1,2}(u) & \cdots & \hbar e^{-}_{1,N}(u) \\ & \ddots & \ddots & \\ & & \ddots & \hbar e_{N-1,N}(u) \\ 0 & \cdots & \cdots & 1 \end{pmatrix},$$

the defining relations in $DY'_{\hbar}(gl_N)$ are given as follows:

$$\begin{aligned} R^{\pm}(u-v)L^{\pm}_{1}(u)L^{\pm}_{2}(v)&=L^{\pm}_{2}(v)L^{\pm}_{1}(u)R^{\pm}(u-v), \\ R^{+}(u-v-c\hbar)L^{+}_{1}(u)L^{-}_{2}(v)&=L^{-}_{2}(v)L^{+}_{1}(u)R^{+}(u-v), \end{aligned} \tag{2.9}$$

and the notation

$$L_1(u)=L(u)\otimes 1, \quad L_2(u)=1\otimes L(u),$$

is used. From the above expression, if we set

$$X^{-}_{\mu}(u)=f^{+}_{\mu+1,\mu}(u)-f^{-}_{\mu+1,\mu}(u),$$

$$X^{+}_{\mu}(u)=e^{+}_{\mu,\mu+1}(u)-e^{-}_{\mu,\mu+1}(u),$$

which defines an isomorphism between the Drinfeld realization and the RS realization. The statement can be proved following the Ding–Frenkel equivalence.[28]

Similar to $DY_{\hbar}(gl_N)'_c$, an associative algebra, the Yangian double with center $DY_{\hbar}(sl_N)'_c$ is a subalgebra of $DY_{\hbar}(gl_N)'_c$, which is obtained with decomposing the $DY_{\hbar}(gl_N)'_c$ by the quantum determinant. If we define the following currents:

$$H_\mu^\pm(u)=k_{\mu+1}^\pm\left(u+\frac{\mu}{2}\hbar\right)k_\mu^\pm\left(u+\frac{\mu}{2}\hbar\right)^{-1},$$

$$K^\pm(u)=\prod_{\mu=1}^{N}k_\mu^\pm(u+(\mu-1)\hbar),$$

$$E_\mu(u)=\frac{1}{\hbar}X_\mu^+\left(u+\frac{\mu}{2}\hbar\right),\quad F_\mu(u)=\frac{1}{\hbar}X_\mu^-\left(u+\frac{\mu}{2}\hbar\right),\tag{2.10}$$

with the mode expansions,

$$H_\mu^+(u)=1+\hbar\sum_{m\geqslant 0}h_{\mu,m}u^{-m-1},\quad H_\mu^-(u)=1-\hbar\sum_{m<0}h_{\mu,m}u^{-m-1},$$

$$E_\mu(u)=\sum_{m\in Z}e_{\mu,m}u^{-m-1},\quad F_\mu(u)=\sum_{m\in Z}f_{\mu,m}u^{-m-1}.$$

With these notation, the Drinfeld generators of $DY'_\hbar(sl_N)_c$ are $\{h_{\nu,m},\ e_{\nu,m},\ f_{\nu,m},\ c\,|\,\nu=1,2,...,N-1;\ m\in Z.\}$

Now we can write the generating relations for $DY'_\hbar(sl_N)$ as follows:[22]

$$[H_\mu^\pm(u),\quad H_\nu^\pm(v)]=0,\quad [c,\text{everything}]=0,$$

$$(u-v+B_{\mu,\nu}\hbar\mp c\hbar)(u-v-B_{\mu,\nu}\hbar)H_\mu^\pm(u)H_\nu^\mp(v)$$
$$=(u-v-B_{\mu,\nu}\hbar\mp c\hbar)(u-v+B_{\mu,\nu}\hbar)H_\nu^\mp(v)H_\mu^\pm(u),$$

$$(u-v-B_{\mu,\nu}\hbar)H_\mu^\pm(u)E_\nu(v)=(u-v+B_{\mu,\nu}\hbar)E_\nu(v)H_\mu^\pm(u),$$

$$(u-v+B_{\mu,\nu}\hbar-c\hbar)H_\mu^+(u)F_\nu(v)=(u-v-B_{\mu,\nu}\hbar-c\hbar)F_\nu(v)H_\mu^+(u),$$

$$(u-v+B_{\mu,\nu}\hbar)H_\mu^-(u)F_\nu(v)=(u-v-B_{\mu,\nu}\hbar)F_\nu(v)H_\mu^-(u),$$

$$(u-v-B_{\mu,\nu}\hbar)E_\mu(u)E_\nu(v)=(u-v+B_{\mu,\nu}\hbar)E_\nu(v)E_\mu(u),$$

$$(u-v+B_{\mu,\nu}\hbar)F_\mu(u)F_\nu(v)=(u-v-B_{\mu,\nu}\hbar)F_\nu(v)F_\mu(u),$$

$$[E_\mu(u),F_\nu(v)]=\frac{\delta_{\mu,\nu}}{\hbar}(\delta(u-(v+c\hbar))H_\mu^+(u)-\delta(u-v)H_\mu^-(v)),$$

$$E_\mu(u_1)E_\mu(u_2)E_\nu(v)-2E_\mu(u_1)E_\nu(v)E_\mu(u_2)+E_\nu(v)E_\mu(u_1)E_\mu(u_2)+(\text{replacement:}\quad u_1\leftrightarrow u_2)$$
$$=0\ \text{ for }\ |\mu-\nu|=1,$$

$$F_\mu(u_1)F_\mu(u_2)F_\nu(v)-2F_\mu(u_1)F_\nu(v)F_\mu(u_2)+F_\nu(v)F_\mu(u_1)F_\mu(u_2)+(\text{replacement:}\quad u_1\leftrightarrow u_2)$$
$$=0\ \text{ for }\ |\mu-\nu|=1,$$

$$E_\mu(u)E_\nu(v)=E_\nu(v)E_\mu(u)\ \text{ for }\ |\mu-\nu|=1,$$

$$F_\mu(u)F_\nu(v)=F_\nu(v)F_\mu(u)\ \text{ for }\ |\mu-\nu|>1,$$

$$[K^\pm(u),K^\pm(v)]=0,$$

$$K^+(u)K^-(v)=K^-(v)K^+(u),$$

$$[K^\epsilon(u),H_\mu^\pm(v)]=[K^\epsilon(u),E_\mu(v)]=[K^\epsilon(u),F_\mu(v)]=0,\quad \forall\epsilon=\pm,\forall 1\leqslant\mu\leqslant N,$$

which are obtained by direct calculation. From the above commutation relations, $K^{\pm}(u)$ commute with all of the elements of $DY_{\hbar}(gl_N)'_c$, so the Fourier coefficients of the power series $K(u)-1$ are the central elements of the $DY_{\hbar}(gl_N)'_c$. It is obvious that the Heisenberg subalgebra of $DY_{\hbar}(gl_N)'_c$ is generated by the quotient of $DY_{\hbar}(gl_N)'_c$ by the ideal generated by this central elements. In fact we have formula $K^{\pm}(u)=\det_q L^{\pm}(u)$, and the scalar function is chosen so that the quantum determinant commutes with all of the currents and with itself.

III. FREE BOSON WITH CONTINUOUS PARAMETER

In the next section we would construct a free boson representation of $DY_{\hbar}(sl_N)_k$ with arbitrary level k ($c=k$ henceforth). For $N=2$ this problem has already been solved in Ref. 21. For generic N, the Yangian double currents can be obtained using the quantum affine-Yangian double correspondence.[23] However, screening currents and vertex operators cannot be obtained in the same way; at least, they cannot be expressed in a nice form. So this correspondence are somewhat nontrivial. In this section, we introduce another kinds Heisenberg algebras to overcome these problems.

N^2 free boson fields are needed to construct the $DY_{\hbar}(gl_N)$ generators with arbitrary level. So we introduce three set boson fields $\hat{\lambda}_{\mu}(1\leqslant\mu\leqslant N)$, $\hat{b}_{\mu,\nu}$, and $\hat{c}_{\mu,\nu}(1\leqslant\mu<\nu\leqslant N)$, and while $\hat{b}_{\mu,\nu}=\hat{c}_{\mu,\nu}\equiv 0$, for $\mu\geqslant\nu$. The quantum Heisenberg algebra $\mathscr{A}_{\hbar,k}(gl_N)$ generated by $\hat{\lambda}_{\mu}(t)$, $\hat{b}_{\mu,\nu}(t)$, $\hat{c}_{\mu,\nu}(t)$ is involved with continuous parameters $t(t\in R-\{0\})$, which enjoy the the following commutation relations

$$[\hat{\lambda}_{\mu}(t),\hat{\lambda}_{\mu}(t')]=\frac{\sinh\frac{k+g}{2}\hbar t\,\sinh\frac{(N-1)}{2}\hbar t\,\sinh\frac{\hbar}{2}t}{t\hbar^2\,\sinh\frac{N}{2}\hbar t}\,\delta(t+t'), \tag{3.1}$$

$$[\hat{\lambda}_{\mu}(t),\hat{\lambda}_{\nu}(t')]=-\frac{\sinh\frac{k+g}{2}\hbar t\,\sinh^2\frac{\hbar}{2}t}{t\hbar^2\,\sinh\frac{N}{2}\hbar t}\,e^{\mathrm{sign}(\mu-\nu)(N/2)\hbar t}\delta(t+t'), \tag{3.2}$$

$$[\hat{b}_{\mu,\nu}(t),\hat{b}_{\mu',\nu'}(t')]=-\frac{1}{\hbar^2 t}\sinh^2\frac{\hbar t}{2}\,\delta_{\mu,\mu'}\delta_{\nu,\nu'}\delta(t+t'), \tag{3.3}$$

$$[\hat{c}_{\mu,\nu}(t),\hat{c}_{\mu',\nu'}(t')]=\frac{1}{\hbar^2 t}\sinh^2\frac{\hbar t}{2}\,\delta_{\mu,\mu'}\delta_{\nu,\nu'}\delta(t+t'), \tag{3.4}$$

and the other commutators vanish identically, where $g=N$ is the dual Coxeter number for the sl_N algebra. Here $\hat{\lambda}_{\mu}(t)$ are the ''fundamental weight type'' fields, while $\hat{b}_{\mu,\nu}(t)$, $\hat{c}_{\mu,\nu}(t)$ are ''ghost type'' fields in conformal field theory alike. For the subalgebra $\mathscr{A}_{\hbar,k}(sl_N)$ of the quantum Heisenberg algebra $\mathscr{A}_{\hbar,k}(gl_N)$, one more constraint

$$\sum_{\mu=1}^{N} e^{(1-\mu)\hbar t}\hat{\lambda}_{\mu}(t)=0,$$

is needed. In fact, the free boson fields defined above are chosen in an analogous way as these yielding representation for the algebra $\mathscr{A}_{\hbar,\eta}(\widehat{sl_N})$.[30]

The Fock space corresponding to the above Heisenberg algebras can be specified as follows. Let $|0\rangle$ be the vacuum state defined by

$$\lambda_{\mu}(t)|0\rangle=b_{\mu,\nu}(t)|0\rangle=c_{\mu,\nu}(t)|0\rangle=0,\quad (t>0). \tag{3.5}$$

The Fock space $\mathscr{F}(l_{\lambda},l_b,l_c)$ is then generated by the actions of the negative modes of λ_{μ}, $b_{\mu,\nu}$, $c_{\mu,\nu}$. The normal ordering $::$ is defined as: $\hat{\chi}_{\mu,\nu}(t)\hat{\chi}_{\mu,\nu}(t'):=:\hat{\chi}_{\mu,\nu}(t')\hat{\chi}_{\mu,\nu}(t):=\hat{\chi}_{\mu,\nu}(t')\hat{\chi}_{\mu,\nu}(t)$,

for $t>0$, where $\hat{\chi}_{\mu,\nu}(t)$ stands for $\hat{\lambda}_{\mu}(t)$, $\hat{b}_{\mu,\nu}(t)$, $\hat{c}_{\mu,\nu}(t)$. We shall see later that this Fock space actually forms a (Wakimoto-like[31,32]) module for the Yangian double $DY_{\hbar}(sl_N)$ with level k.

We define the free boson fields $\chi_{\mu,\nu}(\beta|u;\alpha)$ and $\chi^{\pm}_{lm}(u)$ through

$$\chi_{\mu,\nu}(\beta|u;\alpha)=\exp\left\{-\hbar\int_{-\infty}^{0}\frac{dt}{2\pi i}\,\frac{1}{\sinh\dfrac{\beta}{2}\hbar t}\,\hat{\chi}_{\mu,\nu}(t)e^{\alpha\hbar t}e^{-iut}\right\}$$

$$\times\exp\left\{-\hbar\int_{0}^{+\infty}\frac{dt}{2\pi i}\,\frac{1}{\sinh\dfrac{\beta}{2}\hbar t}\,\hat{\chi}_{\mu,\nu}(t)e^{-\alpha\hbar t}e^{-iut}\right\}, \tag{3.6}$$

$$\chi^{+}_{\mu,\nu}(u)=\exp\left\{2\hbar\int_{0}^{+\infty}\frac{dt}{2\pi i}\,\hat{\chi}_{\mu,\nu}(t)e^{-iut}\right\}, \tag{3.7}$$

$$\chi^{-}_{\mu,\nu}(u)=\exp\left\{-2\hbar\int_{-\infty}^{0}\frac{dt}{2\pi i}\,\hat{\chi}_{\mu,\nu}(t)e^{-iut}\right\}, \tag{3.8}$$

where $\chi_{\mu,\nu}(u)$ [$\chi^{\pm}_{\mu,\nu}(u)$) stands for $\lambda_{\nu}(u)$, $b_{\mu,\nu}(u)$ or $c_{\mu,\nu}(u)(\lambda^{\pm}_{\nu}(u)$, $b^{\pm}_{\mu,\nu}(u)$ or $c^{\pm}_{\mu,\nu}(u)]$. Then following the standard procedure (normal ordering), we have (here and below, all OPE relations should be understood to hold in the analytic continuation sense):

$$b_{\mu,\nu}(u)b^{-}_{\mu',\nu'}(v)=\exp\left\{-2\int_{\widetilde{c}}\frac{dt\,\ln(-t)}{2\pi i}\,\frac{\sinh\dfrac{\hbar}{2}t}{t}\,\delta_{\mu,\mu'}\delta_{\nu,\nu'}e^{-i(u-v)t}\right\}:b_{\mu,\nu}(u)b^{-}_{\mu',\nu'}(v):$$

$$=\left(\frac{\Gamma^2\left(\dfrac{i(u-v)}{\hbar}+\dfrac{1}{2}\right)}{\Gamma\left(\dfrac{i(u-v)}{\hbar}-\dfrac{1}{2}\right)\Gamma\left(\dfrac{i(u-v)}{\hbar}+\dfrac{3}{2}\right)}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}b^{-}_{\mu',\nu'}(v)b_{\mu,\nu}(u)$$

$$=\left(\frac{u-v+i\dfrac{\hbar}{2}}{u-v-i\dfrac{\hbar}{2}}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}b^{-}_{\mu',\nu'}(v)b_{\mu,\nu}(u), \tag{3.9}$$

where the $\widetilde{c}$ indicate the circle from $+\infty$ to 0 in the upper half plain and 0 to ∞ in the lower half plain.[33,30] To derive Eq. (3.9), the identity

$$\int_{\widetilde{c}}\frac{dt\,\ln(-t)}{2i\pi t}\,\frac{e^{-ut}}{1-e^{-t/\eta}}=\ln\Gamma(\eta u)+(\eta u-1/2)(\gamma-\ln\eta)-\frac{1}{2}\ln 2\pi \tag{3.10}$$

is used, in which the γ is the Euler number. In the same way, we can get the following identities;

$$b^{+}_{\mu,\nu}(u)b_{\mu',\nu'}(v)=\left(\frac{u-v+i\dfrac{\hbar}{2}}{u-v-i\dfrac{\hbar}{2}}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}b_{\mu',\nu'}(v)b^{+}_{\mu,\nu}(u), \tag{3.11}$$

$$c_{\mu,\nu}(u)c^{-}_{\mu',\nu'}(v)=\left(\frac{u-v-i\dfrac{\hbar}{2}}{u-v+i\dfrac{\hbar}{2}}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}c^{-}_{\mu',\nu'}(v)c_{\mu,\nu}(u), \tag{3.12}$$

$$c^{+}_{\mu,\nu}(u)c_{\mu',\nu'}(v)=\left(\frac{u-v-i\,\frac{\hbar}{2}}{u-v+i\,\frac{\hbar}{2}}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}} c_{\mu',\nu'}(v)c^{+}_{\mu,\nu}(u), \tag{3.13}$$

$$\lambda^{+}_{\mu}(u)\lambda_{\mu}(k+N|v;\alpha)$$

$$=\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}+1\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}'+1\right)}\lambda_{\mu}(k+N|v;\alpha)\lambda^{+}_{\mu}(u),$$

$$\lambda_{\mu}(k+N|u;\alpha)\lambda^{-}_{\mu}(v)$$

$$=\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}+1\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}+1\right)}\lambda^{-}_{\mu}(v)\lambda_{\mu}(k+N|u;\alpha), \tag{3.14}$$

$$\lambda^{+}_{\mu}(u)\lambda_{\nu}(k+N|v;\alpha)$$

$$=\frac{\Gamma^2\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}\right)}\lambda_{\nu}(k+N|v;\alpha)\lambda^{+}_{\nu}(u),$$

$$\lambda_{\mu}(k+N|u;\alpha)\lambda^{-}_{\nu}(v)$$

$$=\frac{\Gamma^2\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}\right)}\lambda^{-}_{\nu}(v)\lambda_{\mu}(k+N|u;\alpha), \quad \mu>\nu, \tag{3.15}$$

$$\lambda^{+}_{\mu}(u)\lambda_{\nu}(k+N|v;\alpha)$$

$$=\frac{\Gamma^2\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}+1\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}+1\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}+1\right)}\lambda_{\nu}(k+N|v;\alpha)\lambda^{+}_{\mu}(u),$$

$$\lambda_{\mu}(k+N|u;\alpha)\lambda^{-}_{\nu}(v)$$

$$=\frac{\Gamma^2\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha}{N}+1\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha+1}{N}+1\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{\alpha-1}{N}+1\right)}\lambda^{-}_{\nu}(v)\lambda_{\mu}(k+N|u;\alpha), \quad \mu<\nu, \tag{3.16}$$

and

$$\lambda_{\mu}^{+}(u)\lambda_{\mu}^{-}(v)=\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}+1-\frac{1}{N}-\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{k+N}{2N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+1-\frac{1}{N}+\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}-\frac{k+N}{2N}\right)}$$

$$\times\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{1}{N}-\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+1+\frac{k+N}{2N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{1}{N}+\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+1-\frac{k+N}{2N}\right)}\lambda_{\mu}^{-}(v)\lambda_{\mu}^{+}(u),$$

$$=\frac{\rho^{+}\left(i(u-v)-\frac{k+N}{2}\hbar\right)}{\rho^{+}\left(i(u-v)+\frac{k+N}{2}\hbar\right)}\lambda_{\mu}^{-}(v)\lambda_{\mu}^{+}(u),$$

$$\lambda_{\mu}^{+}(u)\lambda_{\nu}^{-}(v)=\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}-\frac{1}{N}-\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{1}{N}-\frac{k+N}{2N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}-\frac{1}{N}+\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+\frac{1}{N}+\frac{k+N}{2N}\right)}$$

$$\times\frac{\Gamma^{2}\left(\frac{i(u-v)}{N\hbar}+\frac{k+N}{2N}\right)}{\Gamma^{2}\left(\frac{i(u-v)}{N\hbar}-\frac{k+N}{2N}\right)}\lambda_{\nu}^{-}(v)\lambda_{\mu}^{+}(u),\quad \mu>\nu, \tag{3.17}$$

$$\lambda_{\mu}^{+}(u)\lambda_{\nu}^{-}(v)=\frac{\Gamma\left(\frac{i(u-v)}{N\hbar}+1-\frac{1}{N}-\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+1+\frac{1}{N}-\frac{k+N}{2N}\right)}{\Gamma\left(\frac{i(u-v)}{N\hbar}+1-\frac{1}{N}+\frac{k+N}{2N}\right)\Gamma\left(\frac{i(u-v)}{N\hbar}+1+\frac{1}{N}+\frac{k+N}{2N}\right)}$$

$$\times\frac{\Gamma^{2}\left(\frac{i(u-v)}{N\hbar}+1+\frac{k+N}{2N}\right)}{\Gamma^{2}\left(\frac{i(u-v)}{N\hbar}+1-\frac{k+N}{2N}\right)}\lambda_{\nu}^{-}(v)\lambda_{\mu}^{+}(u),\quad \mu<\nu, \tag{3.18}$$

$$b_{\mu,\nu}^{+}(u)b_{\mu',\nu'}^{-}(v)=\left(\frac{\Gamma^{3}\left(\frac{i(u-v)}{\hbar}\right)\Gamma\left(\frac{i(u-v)}{\hbar}+2\right)}{\Gamma^{3}\left(\frac{i(u-v)}{\hbar}+1\right)\Gamma\left(\frac{i(u-v)}{\hbar}-1\right)}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}b_{\mu',\nu'}^{-}(v)b_{\mu,\nu}^{+}(u)$$

$$=\left(\frac{(u-v)^{2}}{(u-v-i\hbar)(u-v+i\hbar)}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}b_{\mu',\nu'}^{-}(v)b_{\mu,\nu}^{+}(u), \tag{3.19}$$

$$c_{\mu,\nu}^{+}(u)c_{\mu',\nu'}^{-}(v)=\left(\frac{\Gamma^{3}\left(\frac{i(u-v)}{\hbar}+1\right)\Gamma\left(\frac{i(u-v)}{\hbar}-1\right)}{\Gamma^{3}\left(\frac{i(u-v)}{\hbar}\right)\Gamma\left(\frac{i(u-v)}{\hbar}+2\right)}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}c_{\mu',\nu'}^{-}(v)c_{\mu,\nu}^{+}(u)$$

$$=\left(\frac{(u-v-i\hbar)(u-v+i\hbar)}{(u-v)^{2}}\right)^{\delta_{\mu,\mu'}\delta_{\nu,\nu'}}c_{\mu',\nu'}^{-}(v)c_{\mu,\nu}^{+}(u). \tag{3.20}$$

In the process of constructing the boson realization of $DY_{\hbar}(sl_N)_k$, the following four identities are useful. In fact, the relations of the arguments in different terms are decided by these identities, there are

$$\frac{b^+_{\mu,\nu}(u):b^-_{\mu',\nu'}(v)b_{\mu',\nu'}\left(v-i\frac{\hbar}{2}\right):}{b^+_{\mu,\nu}(u)b_{\mu',\nu'}\left(v+i\frac{\hbar}{2}\right)}=\frac{:b^-_{\mu',\nu'}(v)b_{\mu',\nu'}\left(v-i\frac{\hbar}{2}\right):b^+_{\mu,\nu}(u)}{b_{\mu',\nu'}\left(v+i\frac{\hbar}{2}\right)b^+_{\mu,\nu}(u)}, \tag{3.21}$$

$$\frac{:b^+_{\mu,\nu}(u)b^-_{\mu,\nu}\left(u+i\frac{\hbar}{2}\right):b^-_{\mu',\nu'}(v)}{b^+_{\mu,\nu}\left(u-i\frac{\hbar}{2}\right)b^-_{\mu',\nu'}(v)}=\frac{b^-_{\mu',\nu'}(v):b^+_{\mu,\nu}(u)b^-_{\mu,\nu}\left(u+i\frac{\hbar}{2}\right):}{b^-_{\mu',\nu'}(v)b^+_{\mu,\nu}\left(u-i\frac{\hbar}{2}\right)}, \tag{3.22}$$

$$\frac{:b^+_{\mu,\nu}\left(u+i\frac{\hbar}{2}\right)b^-_{\mu,\nu}(u)^{-1}:b^-_{\mu',\nu'}(v)}{b^+_{\mu,\nu}(u+i\hbar)^{-1}b^-_{\mu',\nu'}(v)}=\frac{b^-_{\mu',\nu'}(v):b^+_{\mu,\nu}\left(u+i\frac{\hbar}{2}\right)b^-_{\mu,\nu}(u)^{-1}:}{b^-_{\mu',\nu'}(v)b^+_{\mu,\nu}(u+i\hbar)^{-1}}, \tag{3.23}$$

$$\frac{b^+_{\mu,\nu}(u):b^-_{\mu',\nu'}\left(v+i\frac{\hbar}{2}\right)b_{\mu',\nu'}(v+i\hbar)^{-1}:}{b_{\mu',\nu'}(v)^{-1}b^+_{\mu,\nu}(u)}=\frac{:b^-_{\mu',\nu'}\left(v+i\frac{\hbar}{2}\right)b_{\mu',\nu'}(v+i\hbar)^{-1}:b^+_{\mu,\nu}(u)}{b^+_{\mu,\nu}(u)b_{\mu',\nu'}(v)^{-1}}. \tag{3.24}$$

The above four identities are only indicated that each of the normal ordering factor of one numerator and its denominator is the same.

IV. FREE BOSON REPRESENTATION OF $DY_\hbar(gl_N)_k$ AND $DY_\hbar(sl_N)_k$

In this section, by using the results obtained in the above sections, we can define a $DY_\hbar(gl_N)_k$ module on the $\mathscr{F}_{\lambda,b,c}$. Then with the help of Eq. (2.11), we get a module of $DY_\hbar(sl_N)_k$ on the $\mathscr{F}_{\lambda,b,c}$. The results can be expressed in terms as:

Proposition 1: There is a homomorphism $\psi_{\hbar,k}$ from $DY_\hbar(gl_N)_k$ to $\mathscr{A}_{\hbar,k}(gl_N)$, which is defined on the generators as follows:

$$k^+_\mu(u)\mapsto:\lambda^+_\mu\left(u+\frac{k-N}{4}i\hbar\right)^{-1}\prod_{\alpha=1}^{\mu-1}\prod_{\beta=\mu+1}^{N}\frac{b^+_{\alpha,\mu}\left(u+\frac{\mu-\alpha}{2}i\hbar\right)}{b^+_{\mu,\beta}\left(u+\frac{\mu-\beta}{2}i\hbar\right)}:, \tag{4.1}$$

$$k^-_\mu(u)\mapsto:\lambda^-_\mu\left(u-\frac{k-N}{4}i\hbar\right)^{-1}\prod_{\rho=1}^{\mu-1}\prod_{\alpha=1}^{\mu}\prod_{\gamma=\mu}^{N}\prod_{\beta=\mu+1}^{N}\frac{b^-_{\rho,\gamma}\left(u+\frac{\rho+\gamma}{2}i\hbar\right)}{b^-_{\alpha,\beta}\left(u+\frac{\alpha+\beta-2}{2}i\hbar\right)}:,$$

$$1\leqslant\mu\leqslant N; \tag{4.2}$$

$$X^+_\nu(u)\mapsto\sum_{m=1}^{\nu}:\left\{\frac{b^+_{m,\nu+1}\left(u-\frac{m-\nu-1}{2}i\hbar\right)}{(b+c)_{m,\nu+1}\left(u-\frac{m-\nu}{2}i\hbar\right)}-\frac{b^-_{m,\nu+1}\left(u-\frac{m-\nu-1}{2}i\hbar\right)}{(b+c)_{m,\nu+1}\left(u-\frac{m-\nu-2}{2}i\hbar\right)}\right\}(b+c)_{m,\nu}$$

$$\times\left(u-\frac{m-\nu-1}{2}i\hbar\right)\prod_{n=1}^{m-1}\frac{b^+_{n,\nu+1}\left(u-\frac{n-\nu-1}{2}i\hbar\right)}{b^+_{n,\nu}\left(u-\frac{n-\nu}{2}i\hbar\right)}:, \tag{4.3}$$

$$
\begin{aligned}
X_\nu^-(u)\mapsto:\Bigg\{ & \sum_{m=1}^{\nu} \lambda_\nu^-\left(u-\frac{k-N}{4}\,i\hbar\right)\lambda_{\nu+1}^-\left(u-\frac{k-N}{4}\,i\hbar\right)^{-1}(b+c)_{m,\nu+1}\left(u+\frac{\nu+m}{2}\,i\hbar\right)\\
&\times\Bigg[(1+\delta_{m,\nu})b_{m,\nu}^-\left(u+\frac{\nu+m}{2}\,i\hbar\right)^{-1}(b+c)_{m,\nu}\left(u+\frac{\nu+m-1}{2}\,i\hbar\right)^{-1}\\
&-b_{m,\nu}^+\left(u+\frac{\nu+m}{2}\,i\hbar\right)^{-1}(b+c)_{m,\nu}\left(u+\frac{\nu+m+1}{2}\,i\hbar\right)^{-1}\Bigg]\\
&\times\prod_{n=m+1}^{\nu}\prod_{s=\nu+1}^{N}\frac{b_{n,\nu+1}^-\left(u+\frac{\nu+n-1}{2}\,i\hbar\right)}{b_{n,\nu}^-\left(u+\frac{\nu+n}{2}\,i\hbar\right)}\frac{b_{\nu,s+1}^-\left(u+\frac{\nu+s}{2}\,i\hbar\right)}{b_{\nu+1,s}^-\left(u+\frac{\nu+s-1}{2}\,i\hbar\right)}\\
&-\sum_{m=\nu+1}^{N}\lambda_\nu^+\left(u-\frac{3k+g}{4}\,i\hbar\right)\lambda_{\nu+1}^+\left(u-\frac{3k+g}{4}\,i\hbar\right)^{-1}(b+c)_{\nu,m}\left(u-\frac{2k+m-\nu-1}{2}\,i\hbar\right)\\
&\times\Bigg[(1+\delta_{m,\nu+1})\frac{b_{\nu+1,m}^+\left(u-\frac{2k+m-\nu-1}{2}\,i\hbar\right)}{(b+c)_{\nu+1,m}\left(u-\frac{2k+m-\nu}{2}\,i\hbar\right)}\\
&-\frac{b_{\nu+1,m}^-\left(u-\frac{2K+m-\nu-1}{2}\,i\hbar\right)}{(b+c)_{\nu+1,m}\left(u-\frac{2k+m-\nu-2}{2}\,i\hbar\right)}\Bigg]\prod_{n=m}^{N}\frac{b_{\nu,n}^+\left(u-\frac{2k+n-\nu}{2}\,i\hbar\right)}{b_{\nu+1,n}^+\left(u-\frac{2k+n-\nu-1}{2}\,i\hbar\right)}:,\\
&1\leqslant\nu\leqslant N-1;
\end{aligned}
\tag{4.4}
$$

is well defined $\mathscr{F}_{\lambda,b,c}$, *and satisfies the commutation relations:*

$$k_\mu^\pm(u)k_\nu^\pm(v)=k_\nu^\pm(v)k_\mu^\pm(u),$$

$$\rho(u-v+ik\hbar)k_\mu^+(u)k_\mu^-(v)=k_\mu^-(v)k_\mu^+(u)\rho(u-v),$$

$$\rho(u-v+ik\hbar)\,\frac{u-v+ik\hbar}{u-v-i\hbar+ik\hbar}\,k_\mu^+(u)k_\nu^-(v)=k_\nu^-(v)k_\mu^+(u)\,\frac{u-v}{u-v-i\hbar}\,\rho(u-v),\quad \mu<\nu,$$

$$\rho(u-v+ik\hbar)\,\frac{u-v+i\hbar+ik\hbar}{u-v+ik\hbar}\,k_\mu^+(u)k_\nu^-(v)=k_\nu^-(v)k_\mu^+(u)\,\frac{u-v+i\hbar}{u-v}\,\rho(u-v),\quad \mu>\nu,$$

$$k_\mu^\pm(u)X_\mu^+(v)k_\mu^\pm(u)^{-1}=\frac{u-v}{u-v-i\hbar}\,X_\mu^+(v),$$

$$k_{\mu+1}^\pm(u)X_\mu^+(v)k_{\mu+1}^\pm(u)^{-1}=\frac{u-v}{u-v+i\hbar}\,X_\mu^+(v),$$

$$k_\mu^-(u)X_\mu^-(v)k_\mu^-(u)^{-1}=\frac{u-v-i\hbar}{u-v}\,X_\mu^-(v),$$

$$k_{\mu+1}^-(u)X_\mu^-(v)k_{\mu+1}^-(u)^{-1}=\frac{u-v+i\hbar}{u-v}\,X_\mu^-(v),$$

$$k_\mu^+(u)X_\mu^-(v)k_\mu^+(u)^{-1}=\frac{u-v-i\hbar+ik\hbar}{u-v+ik\hbar}\,X_\mu^-(v),$$

$$k_{\mu+1}^+(u)X_\mu^-(v)k_{\mu+1}^+(u)^{-1}=\frac{u-v+i\hbar+ik\hbar}{u-v+ik\hbar}\,X_\mu^-(v), \tag{4.5}$$

$$k_\mu^\pm(u)X_\nu^+(v)k_\mu^\pm(u)^{-1}=X_\nu^+(v),\quad \mu-\nu\geqslant 2,\ \text{or}\ \ \mu-\nu\leqslant -1$$

$$k_\mu^\pm(u)X_\nu^-(v)k_\mu^\pm(u)^{-1}=X_\nu^-(v),\quad \mu-\nu\geqslant 2,\ \text{or}\ \ \mu-\nu\leqslant -1,$$

$$(u-v\pm i\hbar)X_\mu^\pm(u)X_\mu^\pm(v)\simeq(u-v\mp i\hbar)X_\mu^\pm(v)X_\mu^\pm(u)\sim reg.,$$

$$(u-v-i\hbar)X_\mu^+(u)X_{\mu+1}^+(v)\simeq(u-v)X_{\mu+1}^+(v)X_\mu^+(u)\sim reg.,$$

$$(u-v)X_\mu^-(u)X_{\mu+1}^-(v)\simeq(u-v-i\hbar)X_{\mu+1}^-(v)X_\mu^-(u)\sim reg.,$$

$$X_\mu^\pm(u_1)X_\mu^\pm(u_2)X_\nu^\pm(v)-2X_\mu^\pm(u_1)X_\nu^\pm(v)X_\mu^\pm(u_2)+X_\nu^\pm(v)X_\mu^\pm(u_1)X_\mu^\pm(u_2)$$
$$+\{u_1\leftrightarrow u_2\}\sim 0,\quad |\mu-\nu|=1,$$

$$X_\mu^\pm(u)X_\nu^\pm(v)\simeq X_\nu^\pm(v)X_\mu^\pm(u)\sim reg.,\quad |\mu-\nu|\geqslant 2,$$

$$[X_\mu^+(u),X_\nu^{-1}(v)]\simeq reg.+\delta_{\mu,\nu}(\delta(u-(v-ik\hbar))k_{\mu+1}^+(u)k_\mu^+(u)^{-1}-\delta(u-v)k_{\mu+1}^-(v)k_\mu^-(v)^{-1}),$$

where reg. means some regular terms and $\simeq$ *and* $\sim$ *imply ''equals up to'' with difference of such terms.*

A somewhat length but straightforward calculation verifies this proposition.

Using the results Eq. (2.11) and Eq. (4.1)–Eq. (4.4), we have the following expressions:

$$H_\mu^+(u)\mapsto:\lambda_\mu^+\left(u+\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)\lambda_{\mu+1}^+\left(u+\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)^{-1}$$
$$\times\prod_{m=1}^{\mu}\prod_{n=\mu+1}^{N}\frac{b_{m,\mu+1}^+\left(u-\frac{m-1}{2}\,i\hbar\right)}{b_{m,\mu}^+\left(u-\frac{m}{2}\,i\hbar\right)}\,\frac{b_{\mu,n}^+\left(u-\frac{n}{2}\,i\hbar\right)}{b_{\mu+1,n}^+\left(u-\frac{n-1}{2}\,i\hbar\right)}:, \tag{4.6}$$

$$H_\mu^-(u)\mapsto:\lambda_\mu^-\left(u-\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)\lambda_{\mu+1}^-\left(u+\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)^{-1}$$
$$\times\prod_{m=1}^{\mu}\prod_{n=\mu+1}^{N}\frac{b_{m,\mu+1}^-\left(u+\frac{m-1}{2}\,i\hbar\right)}{b_{m,\mu}^-\left(u+\frac{m}{2}\,i\hbar\right)}\,\frac{b_{\mu,n}^-\left(u+\frac{n}{2}\,i\hbar\right)}{b_{\mu+1,n}^-\left(u+\frac{n-1}{2}\,i\hbar\right)}:, \tag{4.7}$$

$$E_\mu(u)\mapsto\frac{1}{\hbar}\sum_{m=1}^{\mu}:(b+c)_{m,\mu}\left(u-\frac{m-1}{2}\,i\hbar\right)\left\{\frac{b^+_{m,\mu+1}\left(u-\frac{m-1}{2}\,i\hbar\right)}{(c+b)_{m,\mu+1}\left(u-\frac{m}{2}\,i\hbar\right)}\right.$$

$$\left.-\frac{b^-_{m,\mu+1}\left(u-\frac{m-1}{2}\,i\hbar\right)}{(c+b)_{m,\mu+1}\left(u-\frac{m-2}{2}\,i\hbar\right)}\prod_{n=1}^{m-1}\frac{b^+_{n,\mu+1}\left(u-\frac{n-1}{2}\,i\hbar\right)}{b^+_{n,\mu}\left(u-\frac{n}{2}\,i\hbar\right)}:,\right. \tag{4.8}$$

$$F_\mu(u)\mapsto\frac{1}{\hbar}:\left\{\sum_{m=1}^{\mu}\lambda^-_\mu\left(u-\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)\lambda^-_{\mu+1}\left(u-\frac{k-N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)^{-1}(b+c)_{m,\mu+1}\right.$$

$$\times\left(u+\frac{m}{2}\,i\hbar\right)\left[(1+\delta_{m,\mu})b^-_{m,\mu}\left(u+\frac{m}{2}\,i\hbar\right)^{-1}(b+c)_{m,\mu}\left(u+\frac{m-1}{2}\,i\hbar\right)^{-1}\right.$$

$$\left.-b^+_{m,\mu}\left(u+\frac{m}{2}\,i\hbar\right)^{-1}(b+c)_{m,\mu}\left(u+\frac{m+1}{2}\,i\hbar\right)^{-1}\right]$$

$$\times\prod_{n=m+1}^{\mu}\prod_{s=\mu+1}^{N}\frac{b^-_{n,\mu+1}\left(u+\frac{n-1}{2}\,i\hbar\right)}{b^-_{n,\mu}\left(u+\frac{n}{2}\,i\hbar\right)}\frac{b^-_{\mu,s+1}\left(u+\frac{s}{2}\,i\hbar\right)}{b^-_{\mu+1,s}\left(u+\frac{s-1}{2}\,i\hbar\right)}-\sum_{m=\mu+1}^{N}\lambda^+_\mu$$

$$\times\left(u-\frac{3k+N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)\lambda^+_{\mu+1}\left(u-\frac{3k+N}{4}\,i\hbar-i\mu\,\frac{\hbar}{2}\right)^{-1}(b+c)_{\mu,m}$$

$$\times\left(u-\frac{2k+m-1}{2}\,i\hbar\right)\left[(1+\delta_{m,\mu+1})\frac{b^+_{\mu+1,m}\left(u-\frac{2k+m-1}{2}\,i\hbar\right)}{(b+c)_{\mu+1,m}\left(u-\frac{2k+m}{2}\,i\hbar\right)}\right.$$

$$\left.\left.-\frac{b^-_{\mu+1,m}\left(u-\frac{2K+m-1}{2}\,i\hbar\right)}{(b+c)_{\mu+1,m}\left(u-\frac{2k+m-2}{2}\,i\hbar\right)}\right]\prod_{n=m}^{N}\frac{b^+_{\mu,n}\left(u-\frac{2k+n}{2}\,i\hbar\right)}{b^+_{\mu+1,n}\left(u-\frac{2k+n-1}{2}\,i\hbar\right)}:,\right. \tag{4.9}$$

$(1\leqslant\mu\leqslant N-1)$ without confusion. The following corollary are the direct result of Proposition 1.

Corollary 1: The fields $H^\pm_\mu(u)$, $E^\pm_\mu(u)$ defined in equations (4.6)–Eq. (4.9) are well-defined on the Fock space $\mathscr{F}(\lambda,l_b,l_c)$ and satisfy all of the equations of Eqs. (2.11) with $\hbar$ replaced by $-i\hbar$.

The corollary is also followed by a straightforward calculation. For the sake of illustration, here we give only an example. First, consider the case

$$H^+_\mu(u)H^-_\mu(v)=\frac{u-v+ik\hbar+i\hbar}{u-v+ik\hbar-i\hbar}\,\frac{u-v-iN\hbar-i\hbar}{u-v-iN\hbar+i\hbar}$$

$$\times\prod_{n=1}^{\mu}\prod_{m=\mu+1}^{N}\left(\frac{u-v-i(n-1)\hbar}{u-v-i(n-2)}\,\frac{u-v-in\hbar}{u-v-i(n+1)\hbar}\right)$$

$$\times\left(\frac{u-v-i(m-1)\hbar}{u-v-i(m-2)\hbar}\,\frac{u-v-im\hbar}{u-v-i(m+1)\hbar}\right)H^-_\mu(v)H^+_\mu(\mu)$$

$$=\frac{u-v+ik\hbar+i\hbar}{u-v+ik\hbar-i\hbar}\frac{u-v-iN\hbar-i\hbar}{u-v+iN\hbar+i\hbar}$$
$$\times\prod_{n=1}^{N}\left(\frac{u-v-i(n-1)\hbar}{u-v-i(n-2)\hbar}\frac{u-v-in\hbar}{u-v-i(n+1)\hbar}\right)H_{\mu}^{-}(v)H_{\mu}^{+}(u)$$
$$=\frac{u-v+ik\hbar+i\hbar}{u-v+ik\hbar-i\hbar}\frac{u-v-i\hbar}{u-v+i\hbar}H_{\mu}^{-}(v)H_{\mu}^{+}(u), \tag{4.10}$$

and

$$H_{\mu}^{+}(u)H_{\mu\pm1}^{-}(v)=\frac{u-v+ik\hbar-i\frac{\hbar}{2}}{u-v+ik\hbar+i\frac{\hbar}{2}}\frac{u-v-iN\hbar+i\frac{\hbar}{2}}{u-v-iN\hbar-i\frac{\hbar}{2}}$$
$$\times\prod_{n=1}^{N}\left(\frac{(u-v-i(n-3/2)\hbar)(u-v-i(n+1/2)\hbar)}{(u-v-i(n-1/2)\hbar)^2}\right)H_{\mu\pm1}^{-}(v)H_{\mu}^{+}(u)$$
$$=\frac{u-v+ik\hbar-i\frac{\hbar}{2}}{u-v+ik\hbar+i\frac{\hbar}{2}}\frac{u-v+i\frac{\hbar}{2}}{u-v-i\frac{\hbar}{2}}H_{\mu\pm1}^{-}(v)H_{\mu}^{+}(u), \tag{4.11}$$

$$H_{\mu}^{+}(u)H_{\nu}^{-}(v)=H_{\nu}^{-}(v)H_{\mu}^{+}(u),\quad \text{for}|\mu-\nu|\geqslant 2. \tag{4.12}$$

The above three equation can be expressed as the second equation of Eqs. (2.11) with $\hbar$ replaced by $-i\hbar$. In the above calculation the identities in the above sections are used without mention. The remaining calculation are the similar.

Remark 1: Obviously, the defining relations of the Yangian double $DY_{\hbar}(gl_N)_k$ is identical to the OPEs of the currents $k_{\mu}^{\pm}$, $X_{\mu}^{\pm}$ up to the regular terms with replacing the $\hbar$ in Eqs. (2.3) by $-i\hbar$ (Wick rotation) [the relations for the Yangian double $DY_{\hbar}(sl_N)_k$ and OPEs of the currents Eq. (4.6)–Eq. (4.9) are similar]. But, the realization of Eq. (4.1) – Eq. (4.4) should not be considered to be isomorphic with the Yangian double $DY_{\hbar}(gl_N)_k$, because the free field realization carries a continuous parameter, while the Yangian double $DY_{\hbar}(sl_N)_k$ has discrete mode. For $g=sl_2$, see Ref. 30 for more information on this point.

So, we give the following proposition:

Proposition 2: Under a homomorphic map, the free field realization Eq. (4.1)–Eq. (4.4) and Eq. (4.6)–Eq. (4.9) give a representation of the Yangian double $DY_{\hbar}(gl_N)_k$ and $DY_{\hbar}(sl_N)_k$, respectively.

V. SCREENING CURRENTS

In Ref. 23 the author give a boson field realization for the currents $DY_{\hbar}(sl_N)$ by the so-called q-affine-Yangian double correspondence. There we use ordinary oscillators with discrete mode, and the screening currents and the vertex operators are difficult to obtain by such boson fields. Because of the importance of the vertex operators and screening currents in the field theories, so construction of screening currents in terms of free fields are one of the main objects in the theories. Especially if we know these quantities we could have been able to calculate the cohomology of the action of our bosonization formulas on the Fock spaces $\mathscr{F}(\lambda,l_b,l_c)$. This problem can be overcome by using the boson fields defined in Eq. (3.1), Eq. (3.3), and Eq. (3.4). The screening currents $S^{l}(u)$, $l=1,...,N-1$ can be expressed as the below nice form (under the homomorphism of course),

$$S_{\mu}(u)\mapsto:\lambda_{\mu+1}(k+N|u;(k+N)/4)\lambda_{\mu}(k+N|u;(k+N)/4)^{-1}\widetilde{S}_{\mu}(u): \tag{5.1}$$

where

$$\widetilde{S}_\mu(u)\mapsto -\frac{1}{\hbar}\sum_{m=\mu+1}^{N}(b+c)_{\mu+1,m}\left(u-\frac{N-m}{2}\,i\hbar\right)\left\{b^-_{\mu,m}\left(u-\frac{N-m}{2}-i\hbar\right)\right\}^{-1}$$

$$\times(c+b)_{\mu,m}\left(u-\frac{N-m+1}{2}\,i\hbar\right)^{-1}-b^+_{\mu,m}\left(u-\frac{N-m}{2}\,i\hbar\right)^{-1}$$

$$\times(c+b)_{\mu,m}\left(u-\frac{N-m-1}{2}\,i\hbar\right)^{-1}\prod_{n=m+1}^{N}\frac{b^-_{\mu+1,m}\left(u-\frac{N-n+1}{2}\,i\hbar\right)}{b^-_{\mu,n}\left(u-\frac{N-n}{2}\,i\hbar\right)}. \quad (5.2)$$

Then by a direct calculation, one can prove that the currents have the following properties:

$$H^\pm_\mu(u)S_\nu(v)\simeq S_\nu(v)H^\pm_\mu(u)\sim reg.,$$

$$E_\mu(u)S_\nu(v)\simeq S_\nu(v)E_\mu(u)\sim reg.,$$

$$(u-v-iB_{\mu,\nu})\widetilde{S}_\mu(u)\widetilde{S}_\nu(v)\simeq(u-v-iB_{\mu,\nu})\widetilde{S}_\nu(v)\widetilde{S}_\mu(u)\sim reg., \quad (5.3)$$

$$F_\mu(u)S_\nu(v)\simeq S_\nu(v)F(u)\sim reg.+\frac{\delta_{\mu,\nu}}{i\hbar}\left(\frac{1}{u-v+\frac{N}{2}\,i\hbar}-\frac{1}{u-v-ik\hbar-\frac{N}{2}\,i\hbar}\right)\lambda_{\mu+1}(k+N|\mu;$$

$$-(k+N)/4)\lambda_\mu(k+N|u;-(k+N)/4)^{-1}$$

$$=reg.+\delta_{\mu,\nu(k+N)}\partial_v\left(\frac{1}{u-v}\,\lambda_{\mu+1}(k+N|u;-(k+N)/4)\right.$$

$$\left.\times\lambda_\mu(k+N|u;-(k+N)/4)^{-1}\right), \quad (5.4)$$

where

$${}_{k+\alpha}\partial_u f(u)=\frac{1}{i\hbar}\left(f\left(u-i\alpha\,\frac{\hbar}{2}\right)-f\left(u+ik\hbar+i\alpha\,\frac{\hbar}{2}\right)\right). \quad (5.5)$$

For the purpose to define the vertex operators, one first introduce the state, $|p_\lambda,0,0\rangle$. It can be easy to show that this state is just the highest weight state with weight $\Lambda=l^1\Lambda_1+\ldots+l^{N-1}\Lambda_{N-1}=(l^1,\ldots,l^{N-1})$. If the vertex operators with weight is given as

$$\Phi_\Lambda(u;\alpha)\mapsto:\sum_{\mu=1}^{N-1}\exp\left\{-\hbar\int_{-\infty}^{0}\frac{dt}{2\pi i}\,\frac{\sinh\frac{l^\mu}{2}\hbar t}{\sinh\frac{k+N}{2}\hbar t\,\sinh\frac{\hbar}{2}t}\,\hat{\lambda}_\mu(t)e^{\alpha\hbar t}e^{-iut}\right\}$$

$$\times\exp\left\{-\hbar\int_{0}^{+\infty}\frac{dt}{2\pi i}\,\frac{\sinh\frac{l^\mu}{2}\hbar t}{\sinh\frac{k+N}{2}\hbar t\,\sinh\frac{\hbar}{2}t}\,\hat{\lambda}_\mu(t)e^{-\alpha\hbar t}e^{-iut}\right\}:, \quad (5.6)$$

and the highest weight state of $DY_\hbar(sl_N)$, can be obtained from the vacuum $|0\rangle$ state by this operators, so the vertex operator is similar to the primary fields in the conformal field theory. The vertex related to the intertwining operators of type one and type two will be discussed elsewhere.

VI. DISCUSSIONS

In this paper we establish the free boson representation of the Yangian double $DY_\hbar(sl_N)$ with arbitrary level k. Such representations of the Yangian double $DY_\hbar(sl_N)$ are expected to be useful in calculating the correlation functions of various quantum integrable systems in $(1+1)$-spacetime dimensions, e.g., the spin Calogero–Sutherland model,[34] the quantum nonlinear Schrödinger equation,[35] and some field theoretic models such as the Thirring model, the Gross–Neveu model with $U(N)$ gauge symmetries, etc. Our representation of $DY_\hbar(sl_N)$ may also be used to analyze the behavior of Yangian double at the critical level $k=-g$, a very fascinating area of great interest of study[24] which corresponds to the $W_\hbar$ algebra. For the length of the manuscript, we will discuss this problem in a separate paper.

ACKNOWLEDGMENTS

One of the authors (Ding) would like to thank Professor K. Wu and Professor Z. Y. Zhu for fruitful discussion. He was supported in part by the ''China postdoctoral Science Foundation.''

[1] V. G. Drinfeld, Sov. Math. Dokl. **283**, 1060 (1985).
[2] V. G. Drinfeld, in *Proceedings of the International Congress of Mathematicians*, Berkeley, 798 (1987).
[3] V. G. Drinfeld, Sov. Math. Dokl. **36**, 212 (1988).
[4] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, *Advanced Series in Mathematical Physics, Vol. 10* (World Scientific, Singapore, 1989), p. 299.
[5] S. Khoroshkin and V. Tolstoy, Lett. Math. Phys. **36**, 373 (1996).
[6] S. Khoroshkin, ''Central Extension of the Yangian Double,'' in Collection SMF, Colloque ''Septièmes Rencontres du Contact Franco-Belge en Algèbre,'' June 1995, Reins; Preprint q-alg/9602031.
[7] S. Khoroshkin and D. Lebedev, ''Intertwining operators for the central extension of the Yangian double,'' Preprint q-alg/9602030 (1996).
[8] K. Iohara and M. Konno, Lett. Math. Phys. **37**, 319 (1996).
[9] N. Yu. Reshetikhin and M. A. Semenov-Tian-Shansky, Lett. Math. Phys. **19**, 133 (1990).
[10] D. Bernard and A. LeClair, Nucl. Phys. B **399**, 709 (1993).
[11] A. LeClair and F. Smirnov, Int. J. Mod. Phys. A **7**, 2997 (1992).
[12] F. A. Smirnov, Int. J. Mod. Phys. A **7 suppl. 1B**, 813; **7 suppl. 1B**, 839 (1992).
[13] J. Shiraishi, Phys. Lett. A **171**, 243 (1992).
[14] A. Matsuo, Phys. Lett. B **308**, 260 (1993).
[15] A. Matsuo, Commun. Math. Phys. **160**, 33 (1994).
[16] K. Kimura, ''On free field representation of the quantum affine algebra $U_q(\widehat{sl_2})$,'' Preprint RIMS-910 (1992).
[17] A. Abada, A. H. Bougourzi, and M. A. El Gradechi, ''A deformation of the Wakimoto construction,'' Preprint CRM-1829 (1992).
[18] I. B. Frenkel and N. Jing, Proc. Natl. Acad. Sci. USA **85**, 9373 (1988).
[19] H. Awata, S. Odake, and J. Shiraishi, Lett. Math. Phys. **30**, 207 (1994).
[20] H. Awata, S. Odake, and J. Shiraishi, Commun. Math. Phys. **162**, 61 (1994).
[21] H. Konno, Lett. Math. Phys. **40**, 321 (1997); Preprint YITP-96-10 (1996).
[22] K. Iohara, J. Phys. A **29**, 4593 (1996) [Preprint q-alg/9603033 (1996)].
[23] B. Y. Hou, L. Zhao, and X. M. Ding, q-affine-Yangian double correspondence and free boson representation of Yangian double with arbitrary level, Preprint q-alg/9701025.
[24] E. Frenkel and N. Reshetikhin, Commun. Math. Phys. **178**, 237 (1996).
[25] B. Y. Hou and W. L. Yang, ''A $\hbar$-deformation of the W_N algebra and its Vertex operators,'' Preprint q-alg/9701023.
[26] X. M. Ding, B. Y. Hou, and L. Zhao, ''$\hbar$ (Yangian) deformation of Miura map and the Virasoro algebra,'' Preprint q-alg/9701014.
[27] A. N. Kirillov and N. Yu. Reshetikhin, Lett. Math. Phys. **12**, 199 (1986).
[28] J. Ding and I. B. Frenkel, Commun. Math. Phys. **156**, 277 (1993).
[29] I. B. Frenkel and N. Reshetikhin, Commun. Math. Phys. **146**, 1 (1992).
[30] S. Khoroshkin, D. Lebedev, and S. Pakuliak, ''Elliptic algebra $\mathcal{A}_{p,q}$ in the scaling limit,'' Preprint q-alg/9702002.
[31] M. Wakimoto, Commun. Math. Phys. **104**, 605 (1986).
[32] B. L. Feigin and E. Frenkel, Russ. Math. Surv. **43**, 221 (1989).
[33] M. Jimbo and T. Miwa, J. Phys. A **29**, 2923 (1996).
[34] K. Takimura and D. Uglov, ''The orthogonal eigenbasis and norms of eigenvectors in the spin Calogero–Sutherland model,'' Preprint RIMS-1114 (1996).
[35] T. Kojima and V. E. Korepin, ''Determinant representations for dynamical correlation functions of the quantum nonlinear Schrödinger equation,'' Preprint RIMS-1115 (1996).

Γ-conformal algebras

Maria I. Golenishcheva-Kutuzova[a)]
Institut Girard Desarques, URA 746, Université Lyon-1, France

Victor G. Kac[b)]
Department of Mathematics, MIT, Cambridge, Massachusetts 02139

(Received 29 August 1997; accepted for publication 28 October 1997)

Γ-conformal algebra is an axiomatic description of the operator product expansion of chiral fields with simple poles at finitely many points. We classify these algebras and their representations in terms of Lie algebras and their representations with an action of the group Γ. To every Γ-conformal algebra and a character of Γ we associate a Lie algebra generated by fields with the OPE with simple poles. Examples include twisted affine Kac–Moody algebras, the sin algebra (which is a ''Γ-conformal'' analogue of the general linear algebra) and its analogues, the algebra of pseudodifferential operators on the circle, etc. © *1998 American Institute of Physics.* [S0022-2488(98)00904-9]

I. INTRODUCTION

In the past there have been a number of papers in which the effect of splitting of a pole of order N in the operator product expansion (OPE) occurs. Examples include representations of twisted affine Kac–Moody algebras,[1] of the Lie algebras of quantum torus,[2] of quantum affine algebras,[3] and of the central extension of the double of Yangian.[4] These constructions found many applications in the theory of integrable systems and exactly solvable models in quantum field theory.

The simplest motivation to consider the OPE with simple poles, but in the shifted points is the following. Consider the OPE of two local fields $a(z)$ and $b(z)$:

$$a(z)b(w)\sim\sum_{j=0}^{N-1}\frac{c^j(w)}{(z-w)^{j+1}} \tag{1.1}$$

and replace $(z-w)^n$ by its q-analogue $(z-w)_q^n=(z-w)(z-qw)\cdots(z-q^{n-1}w)$. The result of this q-deformation is an OPE with only simple poles. A natural generalization is to consider Γ-deformation, where Γ is a subgroup of the multiplicative group of non-zero complex numbers $\mathbb{C}^\times$. Namely, for an n-element subset S of Γ we let

$$(z-w)_S^n=\prod_{\alpha\in S}(z-\alpha w)$$

and consider a collection of fields satisfying the condition of Γ-locality, i.e.,

$$(z-w)_S^n[a(z),b(w)]=0. \tag{1.2}$$

The study of algebraic properties of Γ-locality lead us to a formal definition of a Γ-conformal algebra given below. Property (1.2) is equivalent to the following form of the OPE (Proposition II.1):

$$a(z)b(w)\sim\sum_{\alpha\in S}\frac{c^\alpha(w)}{z-\alpha w}. \tag{1.3}$$

[a)]Electronic mail: kutuzova@geometrie.univ-lyon1.fr
[b)]Electronic mail: kac@math.mit.edu

 © 1998 American Institute of Physics

We view the field $c^{\alpha}(w)$ as an ''α-product'' of the fields $a(z)$ and $b(z)$:

$$c^{\alpha}(w)=(a_{(\alpha)}b)(w). \tag{1.4}$$

Introduce operators $T_{\alpha}(\alpha\in\mathbb{C}^{\times})$ on the space of fields by

$$T_{\alpha}a(z)=\alpha a(\alpha z). \tag{1.5}$$

Let R be a space over $\mathbb{C}$ of Γ-local fields which is closed under products (1.4) and invariant under operators (1.5) for all $\alpha\in\Gamma$. Then one can show (Proposition III.1) that the following properties hold, where $a,b,c\in R$, $\alpha,\beta\in\Gamma$:

(C0) $a_{(\alpha)}b=0$ for all but finitely many $\alpha\in\Gamma$,
(C1) $(T_{\alpha}a)_{(\beta)}b=a_{(\beta\alpha)}b$,
(C2) $a_{(\alpha)}b=-T_{\alpha}(b_{(\alpha^{-1})}a)$,
(C3) $[a_{(\alpha)},b_{(\beta)}]c=(a_{(\beta^{-1}\alpha)}b)_{(\beta)}c$.

Now we observe that one can forget that Γ is a subgroup of $\mathbb{C}^{\times}$ and that R is a space of fields. What remains is a (not necessarily abelian) group Γ, its representation space R and $\mathbb{C}$-bilinear products $a_{(\alpha)}b$ on R for each $\alpha\in\Gamma$, such that the axioms (C0)–(C3) hold. This is the basic object of our study, called a Γ-*conformal algebra*.

Recall that the OPE (1.1) of local fields similarly leads to the notion of a conformal algebra.[5,6] However, conformal algebras and Γ-conformal algebras have dramatically different properties.

Our first result on Γ-conformal algebras is Theorem IV.1 which gives their classification in terms of admissible pairs $(\mathfrak{g},\varphi)$, where $\mathfrak{g}$ is a Lie algebra and φ is a homomorphism of Γ to the group Aut$\mathfrak{g}$ of automorphisms of $\mathfrak{g}$ such that for any $a,b\in\mathfrak{g}$ one has

$$[T_{\alpha}a,b]=0 \text{ for all but finitely many } \alpha\in\Gamma. \tag{1.6}$$

Of course, if Γ is a finite subgroup of Aut$\mathfrak{g}$, then (1.6) holds. The associated Γ-conformal algebra is called a twisted current conformal algebra. A more interesting example corresponds to the admissible pair $(g\mathscr{l}_{\infty},\mathbb{Z})$, where $g\mathscr{l}_{\infty}$ is the Lie algebra of all $\mathbb{Z}\times\mathbb{Z}$-matrices (a_{ij}) over $\mathbb{C}$ with a finite number of non-zero entries and the action of $\mathbb{Z}$ is by translations along the diagonal: $n\cdot(a_{i,j})=(a_{i+n,j+n})$, $n\in\mathbb{Z}$. This $\mathbb{Z}$-conformal algebra is denoted by $gc_1(\mathbb{Z})$ for reasons explained below.

To a Γ-conformal algebra and a homomorphism χ of Γ to $\mathbb{C}^{\times}$ (or, more generally, to the group of meromorphic transformations of $\mathbb{C}$, see Sec. VI) one canonically associates an infinite-dimensional Lie algebra. This construction produces many known (and a lot of new) examples of infinite-dimensional Lie algebras which appear in deformation of conformal field theories. For example, the Lie algebra of q-differential operators on the circle (the sin-algebra) is obtained from the $\mathbb{Z}$-conformal algebra $gc_1(\mathbb{Z})$ by using $\chi:\mathbb{Z}\to\mathbb{C}^{\times}$ defined by $\chi(n)=q^n$. Taking the same $\mathbb{Z}$-conformal algebra, but the homomorphism $\chi:\mathbb{Z}\to GL_2(\mathbb{C})$ defined by $\chi(n)z=z+n$, $z\in\mathbb{C}$, instead, produces the Lie algebra of pseudodifferential operators on the circle (see Sec. VI).

Representations of Γ-conformal algebras are classified by equivariant $(\mathfrak{g},\Gamma)$-modules satisfying a finiteness condition (Theorem V.1). On the other hand, to any representation π of the group Γ we canonically associate the general Γ-conformal algebra $gc(\pi,\Gamma)$, which plays the role of the general linear group in the sense that representations of a Γ-conformal algebra R correspond to homomorphisms $R\to gc(\pi,\Gamma)$ (Theorem V.2). The most important case is when π is a free $\mathbb{C}[\Gamma]$-module of rank N. The corresponding general Γ-conformal algebra, denoted by $gc_N(\Gamma)$, is a generalization of $gc_1(\mathbb{Z})$ mentioned above. It is an analogue of gc_N in the theory of conformal algebras.[6,7]

In the present paper we consider the simplest case of simple poles and of one indeterminate z. It is straightforward to generalize this to the case of multiple poles and the case of several indeterminates. This will be discussed elsewhere.

II. *S*-LOCAL FORMAL DISTRIBUTIONS

Recall that a formal distribution with coefficients in a vector space U is an expression of the form

$$a(z)=\sum_{n\in\mathbb{Z}} a_n z^{-n-1}, \quad \text{where} \quad a_n\in\mathbb{Z}.$$

They form a vector space denoted by $U[[z,z^{-1}]]$. We shall use the standard notation:

$$\operatorname*{Res}_{z=0} a(z)=a_0.$$

Similarly, a formal distribution in z and w is an expression of the form

$$a(z,w)=\sum_{m,n\in\mathbb{Z}} a_{m,n} z^{-m-1} w^{-n-1}, \text{ where } a_{m,n}\in U.$$

It may be viewed as a formal distribution in z with coefficients in $U[[w,w^{-1}]]$.

Recall that the formal δ-function is the following formal distribution in z and w with coefficients in $\mathbb{C}$:

$$\delta(z-w)=\sum_{n\in\mathbb{Z}} z^{n-1} w^{-n}.$$

It has the following two basic properties:

$$(z-w)\,\delta(z-w)=0 \tag{2.1}$$

$$\operatorname*{Res}_{z=0} a(z)\,\delta(z-w)=a(w) \quad \text{for } a(z)\in U[[z,z^{-1}]]. \tag{2.2}$$

We shall use the formal distribution

$$\delta(\alpha z-\beta w)=\sum_{n\in\mathbb{Z}} (\alpha z)^{n-1}(\beta w)^{-n}, \quad \alpha,\beta\in\mathbb{C}^\times,$$

and the following obvious property

$$\delta(\alpha z-\beta w)=\alpha^{-1}\delta(z-\alpha^{-1}\beta w)=\beta^{-1}\delta(\alpha\beta^{-1}z-w). \tag{2.3}$$

We shall use also the following equality of distributions in z_1, z_2 and z_3:

$$\delta(z_1-z_2)\,\delta(z_2-z_3)=\delta(z_1-z_2)\,\delta(z_1-z_3) \tag{2.4}$$

Let S be an N-element set of (distinct) non-zero complex numbers. We shall use the following notation:

$$(z-w)_S^N=\prod_{\alpha\in S}(z-\alpha w). \tag{2.5}$$

Definition II.1: Two formal distributions $a(z)$ and $b(z)$ with coefficients is Lie algebra $\mathfrak{g}$ are called S-local if in $\mathfrak{g}[[z,z^{-1},w,w^{-1}]]$ *one has:*

$$(z-w)_S^N[a(z),b(w)]=0. \tag{2.6}$$

Note that S-local formal distributions are S'-local for any finite subset S' of $\mathbb{C}^\times$ containing S.

Proposition II.1: If $a(z)$ and $b(z)$ are S-local formal distributions, then there exists a unique decomposition of the form

$$[a(z),b(w)]=\sum_{\alpha\in S} c_\alpha(w)\,\delta(z-\alpha w). \tag{2.7}$$

The formal distributions $c_\alpha(w)$ are given by the formula

$$c_\alpha(w)=\operatorname*{Res}_{z=0} P_{S,\alpha}\left(\frac{z}{w}\right)[a(z),b(w)], \tag{2.8}$$

where

$$P_{S,\alpha}(u)=\prod_{\substack{\beta\in S\\ \beta\neq\alpha}}\frac{u-\beta}{\alpha-\beta}. \tag{2.9}$$

Proof of this proposition is immediate by the following lemma.

Lemma II.1: (a) Each formal distribution $a(z,w)$ *can be uniquely written in the form:*

$$a(z,w)=\sum_{\alpha\in S} c_\alpha(w)\,\delta(z-\alpha w)+b(z,w), \tag{2.10}$$

where

$$c_\alpha(w)=\operatorname*{Res}_{z=0} P_{S,\alpha}\left(\frac{z}{w}\right)a(z,w) \tag{2.11}$$

and

$$\operatorname*{Res}_{z=0} P_{S,\alpha}\left(\frac{z}{w}\right)b(z,w)=0 \quad \text{for all } \alpha\in S. \tag{2.12}$$

(b) If $b(z,w)$ *satisfies (2.12) and* $(z-w)_S^N b(z,w)=0$ then b $(z,w)=0$.

Proof: Consider the following operator on the space of formal distributions in z and w:

$$\pi_S a(z,w)=\sum_{\alpha\in S} c_\alpha(w)\,\delta(z-\alpha w),$$

where $c_\alpha(w)$ are given by (2.11). It is clear using (2.2) that $\pi_S^2=\pi_S$, which implies (a). Furthermore, suppose that $(z-w)_S^N b(z,w)=0$. It follows from Ref. 5 [Corollary 2.2] that for each $\alpha\in S$:

$$P_{S,\alpha}b(z,w)=d_\alpha(w)\,\delta(z-\alpha w)$$

for some formal distribution $d_\alpha(w)$. If also (2.12) holds, then $\operatorname{Res}_{z=0} d_\alpha(w)\ \delta(z-\alpha w)=0$, hence $d_\alpha(w)=0$ by (2.2), and

$$P_{S,\alpha}\left(\frac{z}{w}\right)b(z,w)=0 \quad \text{for all } \alpha\in S. \tag{2.13}$$

This implies that $b(z,w)=0$ since the polynomials, $P_{S,\alpha}$ are relatively prime. (Indeed $\Sigma_{\alpha\in S}P_{S,\alpha}u_\alpha=1$ for some polynomials u_α. Multiplying both sides of (2.13) by $u_\alpha(z/w)$ and summing over α, we get $b(z,w)=0$.) This proves (b). □

Corollary II.1: The formal distributions $a(z)$ *and* $b(z)$ *are S-local if there exists a decomposition of the form* (2.7)

Proof: Follows from Proposition II.1 and (2.1). □

Remark II.1: The coefficients $c_\alpha(w)$ are independent of the choice of S for which (2.6) holds. This follows from the uniqueness of the decomposition (2.10).

Remark II.2: Formula (2.7), written out in modes, looks as follows:

$$[a_m,b_n]=\sum_{\alpha\in S}\alpha^m c_{\alpha,m+n}.$$

III. α-PRODUCTS OF Γ-LOCAL FORMAL DISTRIBUTIONS

Let Γ be a subgroup of $\mathbb{C}^\times$. Two formal distributions $a(z)$ and $b(z)$ are called Γ-local if they are S-local for some finite subset S of Γ. For $\alpha\in\Gamma$ define the α-product $(a_{(\alpha)}b)(w)$ as the coefficient $c_\alpha(z)$ in the expansion (2.7). In other words,

$$(a_{(\alpha)}b)(w)=\operatorname*{Res}_{z=0}P_{S,\alpha}\left(\frac{z}{w}\right)[a(z),b(w)]. \tag{3.1}$$

Due to Remark II.1, the α-product is independent of the choice of S.

For $\beta\in\Gamma$ introduce the following operator:

$$T_\beta(a(z))=\beta a(\beta z).$$

It is clear that T_β preserves Γ-locality.

We collect below the main properties of α-products.

Proposition III.1: For Γ-local formal distributions one has the following properties $(\alpha,\beta\in\Gamma)$:

(a) (Translation invariance) $(T_\alpha a)_{(\beta)}b=a_{(\alpha\beta)}b$, and $a_{(\beta)}T_\alpha b=T_\alpha(a_{(\beta\alpha^{-1})}b)$.

(b) (Skewsymmetry) $a_{(\alpha)}b=-T_\alpha(b_{(\alpha^{-1})}a)$.

(c) (Jacobi identity) $a_{(\alpha)}(b_{(\beta)}c)=(a_{(\alpha\beta^{-1})}b)_{(\beta)}c+b_{(\beta)}(a_{(\alpha)}c)$ *provided that all the suitable localities hold.*

Proof:

(a) We have:

$$[T_\alpha a(z),b(w)]=\sum_{\beta\in\Gamma}((T_\alpha a)_{(\beta)}b)(w)\,\delta(z-\beta w),$$

$$[T_\alpha a(z),b(w)]=[\alpha a(\alpha z),b(w)]=\sum_{\gamma\in\Gamma}\alpha(a_{(\gamma)}b)(w)\,\delta(\alpha z-\gamma w)$$

$$=\sum_{\gamma\in\Gamma}(a_{(\gamma)}b)(w)\,\delta(z-\alpha^{-1}\gamma w).$$

Comparing coefficients, we get the first equation of (a).

(b) We have using (2.7) and (2.3):

$$[a(z),b(w)]=-[b(w),a(z)]=-\sum_{\beta\in\Gamma}(b_{(\beta)}a)(z)\,\delta(w-\beta z)$$

$$=-\sum_{\beta\in\Gamma}\beta(b_{(\beta^{-1})}a)(\beta w)\,\delta(z-\beta w).$$

Comparing with (2.7), we obtain (b). The second equation of (a) follows from the first and (b).

(c) We have:

$$[a(z),[b(w),c(t)]]=\sum_{\alpha,\beta\in\Gamma}(a_{(\alpha)}(b_{(\beta)}c))(t)\,\delta(z-\alpha t)\,\delta(w-\beta t),$$

$$[a(z),b(w)],c(t)]=\sum_{\alpha,\beta\in\Gamma}((a_{(\alpha\beta^{-1})}b)_{(\beta)}c)(t)\,\delta(z-\alpha\beta^{-1}w)\,\delta(w-\beta t)$$

$$=\sum_{\alpha\in\Gamma}((a_{(\alpha\beta^{-1})}b)_{(\beta)}c)(t)\,\delta(z-\alpha t)\,\delta(w-\beta t).$$

We have used here (2.3) and (2.4). We also have:

$$[b(w),[a(z),c(t)]]=\sum_{\alpha,\beta\in\Gamma} b_{(\beta)}(a_{(\alpha)}c)(t)\,\delta(w-\beta t)\,\delta(z-\alpha t).$$

Equating the first equality with the sum of the remaining two and using Jacobi identity in $\mathfrak{g}$, we obtain equality (c) due to linear independence of $\delta(w-\beta t)\,\delta(z-\alpha t)$ for $\alpha,\beta\in\Gamma$. □

IV. Γ-CONFORMAL ALGEBRAS AND LIE ALGEBRAS OF Γ-LOCAL FORMAL DISTRIBUTIONS

The above considerations motivate the following definitions. Let Γ be a group and let $\mathbb{C}[\Gamma]$ be the group algebra of Γ with basis T_α ($\alpha\in\Gamma$), so that $T_\alpha T_\beta=T_{\alpha\beta}$, $T_1=1$. Since Γ is not necessarily abelian we must distinguish the case of left and right $\mathbb{C}[\Gamma]$-modules.

Definition IV.1:A left $\mathbb{C}[\Gamma]$*-module R is called a left* Γ-conformal algebra *if it is equipped with a* $\mathbb{C}$*-bilinear product* $a_{(\alpha)}b$ *for each* $\alpha\in\Gamma$ *such that the following axioms hold* ($a,b,c\in R$, $\alpha,\beta\in\Gamma$):

(C0) $a_{(\alpha)}b=0$ for all but finitely many $a\in\Gamma$,
(C1) $(T_\alpha a)_{(\beta)}b=a_{(\beta\alpha)}b$,
(C2) $a_{(\alpha)}b=-T_\alpha(b_{(\alpha^{-1})}a)$,
(C3) $a_{(\alpha)}(b_{(\beta)}c)=(a_{(\beta^{-1}\alpha)}b)c+b_{(\beta)}(a_{(\alpha)}c)$.

A right $\mathbb{C}[\Gamma]$-module R is called a *right* Γ*-conformal algebra* if axioms (C1), (C2) and (C3) are replaced by:

$(\mathrm{C1})_\mathrm{R}$ $(aT_\alpha)_{(\beta)}b=a_{(\alpha\beta)}b$,
$(\mathrm{C2})_\mathrm{R}$ $a_{(\alpha)}b=-(b_{(\alpha^{-1})}a)T_\alpha$,
$(\mathrm{C3})_\mathrm{R}$ $a_{(\alpha)}(b_{(\beta)}c)=(a_{(\alpha\beta^{-1})}b)_{(\beta)}c+b_{(\beta)}(a_{(\alpha)}c)$.

Unless otherwise specified, we will consider left Γ-conformal algebras and will drop the adjective left.

Remark IV.1: (a) Properties (C1) and (C2) (resp. $(\mathrm{C1})_\mathrm{R}$ and $(\mathrm{C2})_\mathrm{R}$) imply
(C1′) $a_{(\beta)}T_\alpha b=T_\alpha(a_{(\alpha^{-1}\beta)}b)$ (resp. $(\mathrm{C1}')_\mathrm{R}$ $a_{(\beta)}(bT_\alpha)=(a_{(\beta\alpha^{-1})}b)T_\alpha$)
(b) (C1) and (C1′) (resp. $(\mathrm{C1})_\mathrm{R}$ and $(\mathrm{C1}')_\mathrm{R}$) imply

$$(T_\alpha a)_{(\beta)}(T_\alpha b)=T_\alpha(a_{(\alpha^{-1}\beta\alpha)}b)\quad(\text{resp.}(aT_\alpha)_{(\beta)}(bT_\alpha)=(a_{(\alpha\beta\alpha^{-1})}b)T_\alpha).$$

In particular, each T_α is an automorphism of the product $a_{(1)}b$ for each $\alpha\in\Gamma$.

(c) With respect to the product $a_{(1)}b$, R is a Lie algebra over $\mathbb{C}$. Due to (C1) all other products are expressed via the 1-product:

$$a_{(\alpha)}b=(T_\alpha a)_{(1)}b.$$

Due to Remark IV.1, we can associate to a Γ-conformal algebra R a pair $(\mathfrak{g},\varphi)$, where $\mathfrak{g}$ is a Lie algebra over $\mathbb{C}$ whose underlying space is R and bracket is $[a,b]=a_{(1)}b$, and φ is a homomorphism of Γ to the group $\mathrm{Aut}(\mathfrak{g})$ such that

$$[T_\alpha a,b]=0\ \text{ for all but finitely many }\ \widetilde{\Phi}\alpha\in\Gamma. \tag{4.1}$$

We call such $(\mathfrak{g},\varphi)$ an *admissible pair*. Conversely, given an admissible pair $(\mathfrak{g},\varphi)$ we denote by $R(\mathfrak{g},\varphi)$ the $\mathbb{C}[\Gamma]$-module $\mathfrak{g}$ with $\mathbb{C}$-bilinear products

$$a_{(\alpha)}b=[T_\alpha a,b]. \tag{4.2}$$

It is straightforward to check that $R(\mathfrak{g},\varphi)$ is a Γ-conformal algebra.

We have proved the following result:

Theorem IV.1: Γ-conformal algebras *are classified by admissible pairs* $(\mathfrak{g},\varphi)$.

Fix now a homomorphism $\chi:\Gamma\to\mathbb{C}^\times$. Let $\mathfrak{s}$ be a Lie algebra and let R be a family of pairwise $\chi(\Gamma)$-local formal distributions with coefficients in $\mathfrak{s}$, which span $\mathfrak{s}$ over $\mathbb{C}$. Assume that R is stable under all T_α and all α-products for $\alpha\in\chi(\Gamma)$. Then $\mathfrak{s}$ is called a *Lie algebra* of $\chi(\Gamma)$-*local formal distributions*.

It follows from Proposition III.1 that R is a $\chi(\Gamma)$-conformal algebra. Conversely, given a Γ-conformal algebra R, we associate to it a Lie algebra $\mathfrak{s}(R)$ of $\chi(\Gamma)$-local formal distributions as follows. Consider a vector space over $\mathbb{C}$ with the basis a_n, where $a\in R$ and $n\in\mathbb{Z}$, and denote by $\mathfrak{s}(R,\chi)$ the quotient of this space by the $\mathbb{C}$-span of elements of the form ($n\in\mathbb{Z}$):

$$(\lambda a+\mu b)_n-\lambda a_n-\mu b_n, \quad \text{where } \lambda,\mu\in\mathbb{C},\ a,b\in R, \tag{4.3a}$$

$$(T_\alpha a)_n-\chi(\alpha)^{-n}a_n, \quad \text{where } \alpha\in\Gamma,\ a\in R. \tag{4.3b}$$

Then it is straightforward to check that the formula (cf. Remark II.2)

$$[a_m,b_n]=\sum_{\gamma\in\Gamma}\chi(\gamma)^m(a_{(\gamma)}b)_{m+n} \tag{4.4}$$

gives a well-defined structure of a Lie algebra on $\mathfrak{s}(R,\chi)$ of $\chi(\Gamma)$-local formal distributions $a(z)=\Sigma_{n\in\mathbb{Z}}a_nz^{-n-1}$, $a\in R$ The associated $\chi(\Gamma)$-conformal algebra is denoted by $R(\mathfrak{s})$.

The relations between the constructed objects can be summarized by the following diagram:

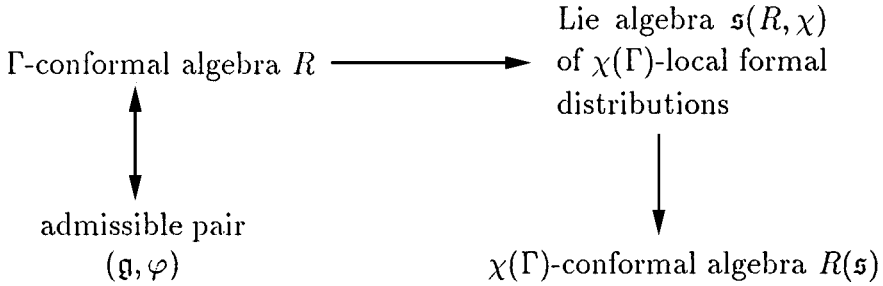

Now we turn to the most important examples of Γ-conformal algebras and related constructions.

Example IV.1: Let $(\mathfrak{g},\mathbb{Z}/N\mathbb{Z})$ be a finite-dimensional (simple) Lie algebra $\mathfrak{g}$ with the action of the group $\Gamma=\mathbb{Z}/N\mathbb{Z}$ by automorphisms of $\mathfrak{g}$. Due to this action T_α, $\alpha\in\Gamma$, we have the eigenspace decomposition ($\hat{\Gamma}$ is the group of characters of Γ):

$$\mathfrak{g}=\bigoplus_{j\in\hat{\Gamma}}\mathfrak{g}_j.$$

The Γ-conformal algebra $R(\mathfrak{g},\varphi)$ is defined as $\mathbb{C}[\Gamma]$-module with underlying space $\mathfrak{g}$ and $\mathbb{C}$-bilinear products $a_{(\alpha)}b=[T_\alpha a,b]$, $a,b\in\mathfrak{g}$, $\alpha\in\mathbb{Z}/N\mathbb{Z}$. Fix a homomorphism $\chi:\Gamma\to\mathbb{C}^\times$ such that $\chi(n)=\epsilon^n$, where $\epsilon\in\mathbb{C}$ is an N-th root of 1. The corresponding Lie algebra of $\chi(\Gamma)$-local formal distributions $\mathfrak{s}(R,\chi)$ is nothing but the twisted affine algebra:[8]

$$[a_m,b_n]=\sum_{k\in\mathbb{Z}/N\mathbb{Z}}\epsilon^{km}[T_ka,b]_{m+n}.$$

Equivalently:

$$[a(z),b(w)]=\sum_{\alpha\in\mathbb{Z}/N\mathbb{Z}}[T_\alpha a,b](w)\,\delta(z-\alpha w).$$

Suppose that $a\in\mathfrak{g}_j$, then $T_ka=\epsilon^{kj}a$ and

$$[a_m,b_n]=\begin{cases}0 & \text{if } m\neq -j \mod N\\ N[a,b]_{m+n} & \text{if } m=-j \mod N.\end{cases}$$

Thus we put $a_m=0$ for $m\neq -j \bmod N$, which is the same as to take a quotient by relations (4.3b).

Example IV.2: Consider $\Gamma=\mathbb{Z}$ and the admissible pair $(\mathfrak{g},\varphi)$, where $\mathfrak{g}=g\mathcal{A}_\infty$ is the Lie algebra of all $\mathbb{Z}\times\mathbb{Z}$ matrices over $\mathbb{C}$ with finitely many non-zero entries. Let E_{ij} $(i,j\in\mathbb{Z})$ be its standard basis, with the usual relations:

$$[E_{ij},E_{k\ell}]=\delta_{j,k}E_{i\ell}-\delta_{\ell,i}E_{kj}.$$

Let T be the image of $1\in\mathbb{Z}$ under φ. Consider two different actions of T on $g\mathcal{A}_\infty$.

(a) $T:E_{ij}\to E_{i+1,j+1}$. In this case $g\mathcal{A}_\infty$ is a free $\mathbb{C}[T,T^{-1}]$-module with the generators $A^m=E_{0,m}$, $m\in\mathbb{Z}$. As follows from Definition IV.1 of a Γ-conformal algebra, it is sufficient to define the Γ-products only on generators of the $\mathbb{C}[\Gamma]$-module. So, for the case under consideration we have:

$$A^m_{(r)}A^n=[T^rA^m,A^n]=[E_{r,m+r},E_{0,n}]=\delta_{r,-m}T^{-m}A^{m+n}-\delta_{r,n}A^{m+n},\quad r\in\mathbb{Z}. \tag{4.5}$$

Take a homomorphism $\chi=\chi_q:\mathbb{Z}\to\mathbb{C}^\times$ defined by $\chi_q(1)=q\in\mathbb{C}^\times$. The corresponding Lie algebra $\mathfrak{s}(R,\chi)$ is the Lie algebra with the basis A^m_k, $m,k\in\mathbb{Z}$, and the commutation relations:

$$[A^m_k,A^n_\ell]=\sum_{r\in\mathbb{Z}}q^{rk}(A^m_{(r)}A^n)_{k+\ell}=(q^{m\ell}-q^{kn})A^{m+n}_{k+\ell}. \tag{4.6}$$

This is the Lie algebra of q-pseudodifferential operators on the circle (sin-algebra).

(b) Now define the action of $\varphi(1)=\widetilde{T}$ by $\widetilde{T}(E_{i,j})=-E_{j+1,i+1}$. We have $\widetilde{T}=\epsilon T$, where ϵ is an order 2 automorphism of $g\mathcal{A}_\infty$ defined by $\epsilon(E_{ij})=-E_{ji}$. In this case $g\mathcal{A}_\infty$ is a free $\mathbb{C}[\widetilde{T},\widetilde{T}^{-1}]$-module with generators $A^m=E_{0,m}$ and $B^m=\epsilon(A^m)=-E_{m,0}$, $m\in\mathbb{Z}_+$. Let us compute the r-product, $r\in\mathbb{Z}$, for the elements of the basis of $\mathbb{C}[\Gamma]-module$. Suppose $n-m\geqslant 0$ (for $n-m<0$ we can use axiom (C2)):

$$A^m_{(r)}A^n=[\widetilde{T}^rA^m,A^n]=\begin{cases}\delta_{r,-m}T^{-m}A^{m+n}-\delta_{r,m}A^{m+n}, & m,n \text{ even};\\ \delta_{r,-m}T^{-m}A^{m+n}+\delta_{r,n-m}A^{n-m}, & m \text{ even}, n \text{ odd};\\ -\delta_{r,n}A^{m+n}+\delta_{r,n-m}A^{n-m}, & m \text{ odd}, n \text{ even}\\ 0, & m,n \text{ odd}.\end{cases}$$

$$B^m_{(r)}B^n=\epsilon[\widetilde{T}^rA^m,A^n]$$

$$B^m_{(r)}A^n=\begin{cases}-\delta_{r,0}\widetilde{T}^mA^{n-m}+\delta_{r,n-m}A^{n-m}, & m,n \text{ even};\\ -\delta_{r,0}\widetilde{T}^mA^{n-m}-\delta_{r,n}A^{m+n}, & m \text{ even}, n \text{ odd};\\ -\delta_{r,0}\widetilde{T}^mB^{n-m}+\delta_{r,-m}\widetilde{T}^{-m}B^{n+m}, & m \text{ odd}, n \text{ even};\\ -\delta_{r,0}\widetilde{T}^nB^{n-m}+\delta_{r,n-m}A^{n+m}\\ \quad+\delta_{r,-m}\widetilde{T}^{-m}B^{n+m}-\delta_{r,n}A^{m+n}, & m,n \text{ odd}.\end{cases}$$

The corresponding Lie algebra $\mathfrak{s}(R,\chi_q)$ is a Lie algebra with basis $\{A^m_k\}$ and $\{B^n_\ell\}$, $m,n\in\mathbb{Z}_+$, $k,\ell\in\mathbb{Z}$. From (4.4) and (4.5) we have for $n-m\geqslant 0$:

$$[A^m_k,A^n_\ell]=\begin{cases}(q^{m\ell}-q^{nk})A^{m+n}_{k+\ell}, & m,n \text{ even},\\ q^{m\ell}A^{m+n}_{k+\ell}+q^{k(n-m)}A^{n-m}_{k+\ell}, & m \text{ even}, n \text{ odd};\\ -q^{nk}A^{m+n}_{k+\ell}+q^{k(n-m)}A^{n-m}_{k+\ell}, & m \text{ odd}, n \text{ even};\\ 0, & m,n \text{ odd},\end{cases}$$

$$[B_k^m, B_{\ell}^n] = \text{ the same as for } [A_n^m, A_{\ell}^n] \text{ with change } A \to B.$$

$$[B_k^m, A_{\ell}^n] = \begin{cases} q^{-mk}(q^{kn} - q^{-m\ell})A_{k+\ell}^{n-m}, & m,n \text{ even}, \\ -q^{-m(k+\ell)}A_{k+\ell}^{n-m} - q^{kn}A_{k+\ell}^{m+n}, & m \text{ even}, n \text{ odd}; \\ -q^{-m(k+\ell)}B_{k+\ell}^{n-m} + q^{m\ell}B_{k+\ell}^{n+m}, & m \text{ odd}, n \text{ even}; \\ -q^{-m(k+\ell)}B_{k+\ell}^{n-m} + q^{k(n-m)}A_{k+\ell}^{n-m} \\ \quad + q^{m\ell}B_{k+\ell}^{n+m} - q^{nk}A_{k+\ell}^{n+m}, & m,n \text{ odd}. \end{cases}$$

This Lie algebra is apparently new. It contains as a subalgebra the sin-algebra $\{\widetilde{A}_n^k\}$, where $\widetilde{A}_k^n = A_k^n$, $n>0$, n even; $A_k^n = q^{nk}B_k^{-n}$, $n<0$, n even; and the subalgebra isomorphic to B-series of Lie algebras of quantum torus (see Ref. 2) for n odd.

Example IV.3: Consider the group $\Gamma = \mathbb{Z}_2 \times \mathbb{Z}$ and admissible pair $(g\ell_\infty, \varphi)$, where homomorphism $\varphi: \Gamma \to \operatorname{Aut} g\ell_\infty$ is determined by two elements ϵ and T as in Example IV.2 (b). With respect to the Γ-action $g\ell_\infty$ is a free module with the generators $B^m = E_{0,m}$, $m \geqslant 0$. The corresponding conformal algebra $R(g\ell_\infty, \varphi)$ is defined by the Γ-products:

$$\begin{aligned} B_{(r)}^m B^n &= [T^r B^m, B^n] = \delta_{r,-m} T^{-m} B^{n+m} - \delta_{n,r} B^{m+r}, \\ B_{(\epsilon r)}^m B^n &= [\epsilon T^r B^m, B^n] = \begin{cases} -\delta_{r,0} T^m B^{n-m} + \delta_{r,n-m} B^{n-m}, & n-m \geqslant 0 \\ \delta_{r,0} \epsilon T^n B^{n-m} - \delta_{r,n-m} B^{m-n}, & n-m < 0. \end{cases} \end{aligned} \tag{4.7}$$

Fix the homomorphism $\chi: \Gamma \to \mathbb{C}^\times$ such that $\chi(\epsilon) = -1$, $\chi(1) = q \in \mathbb{C}^\times$. The Lie algebra $\mathfrak{s}(R(g\ell_\infty, \varphi), \chi)$ is a Lie algebra with the basis B_k^m, $m \in \mathbb{Z}_+$, $k \in \mathbb{Z}$, and commutation relations:

$$[B_k^m, B_{\ell}^n] = (q^{m\ell} - q^{nk}) B_{k+\ell}^{m+n} - \begin{cases} (-1)^k (q^{k(n-m)} - q^{-m(k+\ell)}) B^{n-m}, & n-m>0, \\ (-1)^k ((-q)^{-n(k+\ell)} \\ -(-1)^{-n(k+\ell)} (-q)^{m(k+\ell)} B_{k+\ell}^{m-n}), & n-m<0. \end{cases}$$

After the renormalization of the basis $\widetilde{B}_k^m = q^{mk/2} B_k^M$ we get exactly the commutation relations for B-series of sin Lie algebras, introduced in Ref. 2:

$$[\widetilde{B}_k^m, \widetilde{B}_{\ell}^n] = (q^{(m\ell - nk)/2} - q^{[-(m\ell - nk]/2)} \widetilde{B}_{k+\ell}^{n+m} - (-1)^k (q^{(kn+m\ell)/2} - q^{-(m\ell + kn)/2}) \widetilde{B}_{k+\ell}^{n-m}.$$

V. REPRESENTATIONS OF Γ-CONFORMAL ALGEBRAS AND THE GENERAL Γ-CONFORMAL ALGEBRA $gc(\pi, \Gamma)$

Definition V.1: *A (left) module over a Γ-conformal algebra R is a $\mathbb{C}[\Gamma]$-module M with a $\mathbb{C}$-linear map $a \mapsto a_\alpha^M$ of R to $End_{\mathbb{C}} M$ for each $\alpha \in \Gamma$ such that the following properties hold (here $a, b \in R$, $\alpha, \beta \in \Gamma$ and the action of Γ on M is denoted by $\alpha \mapsto T_\alpha^M$):*

(M0) $a_{(\alpha)}^M v = 0$ *for* $v \in M$ *and all but finitely many* α.
(M1) $(T_\alpha a)_{(\beta)}^M = a_{\beta\alpha}^M$, $a_{(\beta)}^M T_\alpha^M = T_\alpha^M a_{(\alpha^{-1}\beta)}^M$,
(M2) $[a_{(\alpha)}^M, b_{(\beta)}^M] = (a_{(\beta^{-1}\alpha)} b)_{(\beta)}^M$.

For example, $a \mapsto a_{(\alpha)}^R$ is the *adjoint representation* of R on itself.
It follows from (M1) that

$$(T_\alpha a)_{(\beta)}^M (T_\alpha^M v) = T_\alpha^M (a_{(\alpha^{-1}\beta\alpha)}^M v), \quad v \in M. \tag{5.1}$$

This leads us to:

Theorem V.1: *Let R be a Γ-conformal algebra and let $(\mathfrak{g}, \varphi)$ be the associated admissible pair. Then R-modules M are classified by equivariant $(\mathfrak{g}, \Gamma)$-modules such that for any $a \in \mathfrak{g}$ and $v \in M$ one has:*

$$(T^{\mathfrak{g}}_{\alpha}a)v=0 \qquad \text{for all but finitely many } \alpha\in\Gamma. \tag{5.2}$$

Proof: It follows from (M2) that $a\mapsto a_{(1)}$ is a representation of $\mathfrak{g}$. It satisfies (5.1) due to (M2), and it is equivariant, i.e. $(T^{\mathfrak{g}}_{\alpha}a)(T^{M}_{\alpha}v)=T^{M}_{\alpha}(av)$ due to (5.1). Conversely, given an equivariant $(\mathfrak{g},\Gamma)$-module M, it becomes an R-module by letting

$$a^{M}_{\alpha}=(T^{\mathfrak{g}}_{\alpha}a)^{M}_{(1)} \text{ on } M.$$

□

Consider a representation $\alpha\mapsto T_{\alpha}$ of the group Γ in a vector space V over $\mathbb{C}$. We define a Γ-*conformal endomorphism* of V as a collection $a=\{a_{(\beta)}\}_{\beta\in\Gamma}$ of $\mathbb{C}$-endomorphisms of V such that

$$a_{(\beta)}T_{\alpha}=T_{\alpha}a_{(\alpha^{-1}\beta)}, \quad \alpha,\beta\in\Gamma, \tag{5.3}$$

and for each $v\in V$

$$a_{(\alpha)}v=0 \quad \text{for all but finitely many } \alpha\in\Gamma. \tag{5.4}$$

We denote the set of all Γ-conformal endomorphisms of V by $gc(V,\Gamma)$.

Define a $\mathbb{C}[\Gamma]$-module structure on the space $gc(V,\Gamma)$ by

$$(T_{\alpha}a)_{(\beta)}=a_{(\beta\alpha)}, \alpha,\beta\in\Gamma, \tag{5.5}$$

and α-product for each $\alpha\in\Gamma$ by

$$(a_{(\alpha)}b)_{(\beta)}=[a_{(\beta\alpha)},b_{(\beta)}]. \tag{5.6}$$

It is immediate to check that axioms (C1), (C2) and (C3) of a Γ-conformal algebra hold (though axiom (C0) probably does not hold in general).

It is clear that we have by definitions:

Proposition V.1:To give a Γ-module V a structure of a module over a Γ-conformal algebra R is the same as to give a homomorphism $R\to gc(V,\Gamma)$ (i.e. a $\mathbb{C}[\Gamma]$-module homomorphism preserving all α-products).

We will show now that axiom (C0) holds, hence $gc(V,\Gamma)$ is a Γ-conformal algebra, provided that the $\mathbb{C}[\Gamma]$-module V is finitely generated. Toward this end we shall give a different construction of $gc(V,\Gamma)$ in the case when V is a free $\mathbb{C}[\Gamma]$-module of rank 1.

Let $gc_1(\Gamma)=\oplus_{r\in\Gamma}\mathbb{C}[\Gamma]a^r$ be a free $\mathbb{C}[\Gamma]$-module with free generators a^r labeled by elements of Γ. For each $\alpha\in\Gamma$ define the α-product by

$$a^{r}{}_{(\alpha)}a^{s}=\delta_{\alpha,r^{-1}}T_{r^{-1}}a^{rs}-\delta_{\alpha,s}a^{sr}, \tag{5.7}$$

extending to the whole R by axioms (C1) and (C1').

Theorem V.2: *(a) The $\mathbb{C}[\Gamma]$-module $gc_1(\Gamma)$ with products* (5.7) *is a Γ-conformal algebra.*

(b) Let $V=\mathbb{C}[\Gamma]v$ be a free $\mathbb{C}[\Gamma]$-module of rank 1. Define the action of $gc_1(\Gamma)$ on V by letting

$$(a^{s})^{V}_{(\alpha)}v=\delta_{\alpha,s^{-1}}T_{\alpha}v \tag{5.8}$$

and extending to V by (M1). This gives V a structure of a $gc_1(\Gamma)$-module.

(c) The $gc_1(\Gamma)$-module $V=\mathbb{C}[\Gamma]v$ contains all Γ-conformal endomorphisms of V. In particular, the Γ-conformal algebras $gc(V,\Gamma)$ and $gc_1(\Gamma)$ are isomorphic.

Proof: To prove (a) we need to check that the axioms (C0)–(C3) are satisfied. (C0) is obvious. (C1) and (C2) come from the definition (5.7) of $gc_1(\Gamma)$. Check the skew symmetry (C2):

$$(a^{r}_{(\alpha)}a^{s})_{(\beta)}=\delta_{\alpha,r^{-1}}(T_{r^{-1}}a^{rs})_{(\beta)}-\delta_{\alpha,s}a^{sr}_{(\beta)},$$

$$T_{\alpha}(a^{s}_{(\alpha^{-1})}a^{r})_{\beta}=\delta_{\alpha,s}T_{\alpha s^{-1}}a^{sr}-\delta_{\alpha,r^{-1}}T_{r^{-1}}a^{rs}=-(\delta_{\alpha,r^{-1}}T_{r^{-1}}a^{rs}-\delta_{\alpha,s}a^{sr}).$$

For (C3) we need to check that both sides of

$$a^m_{(\beta\alpha)}(a^r_{(\beta)}a^s)=(a^m_{(\alpha)}a^r)_{(\beta)}a^s+a^r_{(\beta)}(a^m_{(\beta\alpha)}a^s)$$

are equal.

$$\begin{aligned}\text{LHS:}\quad a^m_{(\beta\alpha)}(a^r_{(\beta)}a^s)&=a^m_{(\beta\alpha)}(\delta_{\beta,r^{-1}}T_{r^{-1}}a^{rs}-\delta_{\beta,s}a^{sr})\\&=\delta_{\beta,r^{-1}}T_{r^{-1}}a^m_{(\alpha)}a^{rs}-\delta_{\beta,s}a^m_{(\beta\alpha)}a^{sr}\\&=\delta_{\beta,r^{-1}}\delta_{\alpha,m^{-1}}T_{(mr)^{-1}}a^{mrs}-\delta_{\beta,r^{-1}}\delta_{\alpha,rs}T_{r^{-1}}a^{rsm}\\&\quad-\delta_{\beta,s}\delta_{\beta\alpha,m^{-1}}T_{m^{-1}}a^{msr}+\delta_{\beta,s}\delta_{\beta\alpha,sr}a^{srm},\end{aligned}$$

$$\begin{aligned}\text{RHS:}\quad (a^m_{(\alpha)}a^r)_{(\beta)}a^s+a^r_{(\beta)}(a^m_{(\beta\alpha)}a^s)&=\delta_{\alpha,m^{-1}}\delta_{\beta\alpha,(mr)^{-1}}T_{(mr)^{-1}}a^{mrs}-\delta_{\alpha_1m^{-1}}\delta_{\beta\alpha,s}a^{smr}\\&\quad-\delta_{\alpha,r}\delta_{\beta,(rm)^{-1}}a^{rms}+\delta_{\alpha,r}\delta_{\beta,s}a^{srm}\\&\quad+\delta_{\beta\alpha,m^{-1}}\delta_{\beta,r}T_{(rm)^{-1}}a^{rms}-\delta_{\beta\alpha,m^{-1}}\delta_{\alpha^{-1},ms}T_{m^{-1}}a^{msr}\\&\quad-\delta_{\beta\alpha,s}\delta_{\beta,r^{-1}}a^{rsm}+\delta_{\beta\alpha,s}\delta_{\beta,sm}a^{smr},\end{aligned}$$

and we have LHS=RHS.

To prove (b) we need to check only axiom (M2) on V:

$$(a^r_{(\alpha)}a^s)_{(\beta)}v=[a^r_{(\beta\alpha)},a^s_{(\beta)}]v.$$

$$\begin{aligned}\text{LHS:}\quad (a^r_{(\alpha)}a^s)_{(\beta)}v&=(\delta_{\alpha,r^{-1}}a^{rs}_{\beta\alpha}-\delta_{\alpha,s}a^{sr}_{(\beta)})v\\&=(\delta_{\alpha,r^{-1}}\delta_{\beta,s^{-1}}T_{(rs)^{-1}}-\delta_{\alpha,s}\delta_{\beta(s2)^{-1}}T_\beta)v,\end{aligned}$$

$$\begin{aligned}\text{RHS:}\quad [a^r_{(\beta\alpha)},a^s_{(\beta)}]v&=(\delta_{\beta,s^{-1}}T_{s^{-1}}a^r_{(\alpha)}-\delta_{\beta\alpha,r^{-1}}T_{r^{-1}}a^s_{(\alpha^{-1})})v\\&=(\delta_{\beta,s^{-1}}\delta_{\alpha,r^{-1}}T_{(rs)^{-1}}-\delta_{\beta\alpha,r^{-1}}\delta_{\alpha,s}T_{(sr)^{-1}})v\end{aligned}$$

so LHS=RHS. Finally, (c) is clear. □

Corollary V.1: If V is a finitely generated $\mathbb{C}[\Gamma]$-module, then $gc(V,\Gamma)$ is a conformal algebra.

Proof: We may assume that $V=\mathbb{C}[\Gamma]^{\oplus N}$ is a free $\mathbb{C}[\Gamma]$-module of rank N. Then $gc(V,\Gamma)$ may be viewed as a $\mathbb{C}[\Gamma]$-submodule of $gc_1(\Gamma^N)$ using the diagonal homomorphism $\mathbb{C}[\Gamma]\to\mathbb{C}[\Gamma]^{\oplus N}$. Corollary now follows from Theorem V.2a. □

If V is a free $\mathbb{C}[\Gamma]$-module of rank N, we use notation $gc_N(\Gamma)=gc(V,\Gamma)$.

Example V.1: Consider $\Gamma=\mathbb{Z}^N$. For the $gc_1(\mathbb{Z}^N)$ defined by (5.7) the corresponding Lie algebra $(\mathfrak{g},\varphi)$ is a Lie algebra with generators $a^\beta_\alpha=T_\alpha a^\beta$, $\alpha,\beta\in\mathbb{Z}^N$ and commutation relations:

$$[a^\delta_\alpha,a^\gamma_\beta]=\delta_{\alpha,\beta\delta^{-1}}a^{\delta\gamma}_{\beta\delta^{-1}}-\delta_{\alpha,\beta\gamma}a^{\gamma\delta}_\beta.$$

Fix $\bar{q}=(q_1,\dots,q_N)\in(\mathbb{C}^\times)^N$ and define the homomorphism $\chi_{\bar{q}}:\mathbb{Z}^N\to\mathbb{C}^\times$, such that $\chi_{\bar{q}}(\alpha)=\bar{q}^\alpha$, where $\alpha\in\mathbb{Z}^N$ and $\bar{q}^\alpha=q_1^{\alpha_1}\cdots q_N^{\alpha_N}$. The Lie algebra $\mathfrak{s}(R,\chi_{\bar{q}})$ of formal distributions is a Lie algebra with generators a^α_m, $\alpha\in\mathbb{Z}^N$, $m\in\mathbb{Z}$. Define the generating functions $a^\alpha(z)=\sum_{m\in\mathbb{Z}}a^\alpha_m z^{-m-1}$. From (4.4) we have:

$$[a^\alpha(z),a^\beta(w)]=\sum_{\gamma\in\mathbb{Z}^N}(a^\alpha_{(\gamma)}a^\beta)(w)\,\delta(z-\chi(\gamma)w),$$

and for modes:

$$[a_m^\alpha, a_n^\beta] = \sum_{\gamma \in \mathbb{Z}^N} \chi(\gamma)^m (a_{(\gamma)}^\alpha a^\beta)_{m+n}.$$

$$= \overline{q}^{-\alpha m} (T_{-\alpha} a^{\alpha+\beta})_{m+n} - \overline{q}^{\beta m} (a^{\alpha+\beta})_{m+n}$$

$$= (\overline{q}^{\alpha n} - \overline{q}^{\beta m}) a_{m+n}^{\alpha+\beta}.$$

This is a vector generalization of sin-algebra, considered in Ref. 2.

VI. NON-COMMUTATIVE GENERALIZATION OF Γ-LOCAL FORMAL DISTRIBUTIONS

In Sec. IV we associated to a Γ-conformal algebra R a Lie algebra $\mathfrak{s}(R,\chi)$ of $\chi(\Gamma)$-local formal distributions, by fixing a homomorphism $\chi:\Gamma \to \mathbb{C}^\times$. Conversely, each Lie algebra of $\chi(\Gamma)$-conformal formal distribution defines a right $\chi(\Gamma)$-conformal algebra. In this paragraph we will consider more general case, where $\chi(\Gamma)$ is a subgroup of the group of meromorphic transformations of $\mathbb{C}$. In all examples we will consider $\chi:\Gamma \to GL_2(\mathbb{C})$. For $\alpha \in \Gamma$ we will denote by $\alpha(z)$ the corresponding under the homomorphism χ transformation from $GL_2(\mathbb{C})$. Consider the generalization of the property (2.2) of the δ-function:

Proposition VI.1: Let $\alpha(z)$ and $\beta(z)$ be from $GL_2(\mathbb{C})$, then:

$$\begin{aligned} \alpha'(z)\,\delta(\alpha(z)-w) &= \delta(z-\alpha^{-1}(w)), \\ \beta'(w)\,\delta(z-\beta(w)) &= \delta(\beta^{-1}(z)-w). \end{aligned} \tag{6.1}$$

Proof: The group $GL_2(\mathbb{C})$ is generated by the transformations a) $z \mapsto dz$, $\alpha \in \mathbb{C}$; b) $z \to z-\alpha$, $\alpha \in \mathbb{C}$; c) $z \mapsto 1/z$. It is sufficient to prove (6.1) for (a), (b) and (c) transformations. For (a) this is (2.3). To prove (b), let

$$\delta(z-(w-\alpha)) = \sum_{k \in \mathbb{Z}} z^{-k-1}(w-\alpha)^k \tag{6.2a}$$

$$\delta((z+\alpha)-w) = \sum_{k \in \mathbb{Z}} (z+\alpha)^{-k-1} w^k. \tag{6.2b}$$

We define

$$(z+\alpha)^{-\ell-1} = \sum_{k=0}^{\infty} \frac{(k+\ell)!}{k!\ell!} z^{-\ell-k-1}(-\alpha)^k \tag{6.3}$$

for $\ell \geqslant 0$. Using (6.3) and comparing coefficients of $z^{-k-1}w^\ell$ in right-hand sides of (6.2a) and (6.2b) we get (b). So, the function

$$\delta(z-w+\alpha) = \sum_k z^{-k-1}(w-\alpha)^k = \sum_\ell (z+\alpha)^{-\ell-1} w^\ell \tag{6.4}$$

is well defined. The proof of (c) is similar. □

Further we will suppose for simplicity that $\Gamma \subset GL_2(\mathbb{C})$. For an N-element subset $S \subset \Gamma$ we shall use the notation

$$(z-w)_S^N = \prod_{\alpha \in S} (z-\alpha(w)).$$

Definition VI.1: Two formal distributions $a(z)$ and $b(z)$ *with coefficients in a Lie algebra $\mathfrak{g}$ are called S-local if in $\mathfrak{g}[[z,z^{-1},w,w^{-1}]]$ one has:*

$$(z-w)_S^N [a(z), b(w)] = 0.$$

Proposition VI.2: If $a(z)$ and $b(z)$ are S-local formal distributions, then there exists a unique decomposition

$$[a(z),b(w)]=\sum_{\alpha\in S}(a_{(\alpha^{-1})}b)(w)\,\delta(z-\alpha(w)). \tag{6.5}$$

The formal distributions $(a_{(\alpha^{-1})}b)(w)$ are given by the formula:

$$(a_{(\alpha^{-1})}b)(w)=\operatorname*{Res}_x\prod_{\substack{\beta\in S\\ \beta\neq\alpha}}\frac{z-\beta(w)}{\alpha(w)-\beta(w)}[a(z),b(w)]. \tag{6.5a}$$

Proof is the same as the proof of Proposition II.1.

For $\alpha\in\Gamma$ introduce the following operator:

$$T_{\alpha^{-1}}a(z)=\alpha'(z)a(\alpha(z)). \tag{6.6}$$

It is clear, that T_α, $\alpha\in\Gamma$, preserve Γ-locality.

For α-products $a(\alpha b)(w)$ defined as (6.5a) we have all the properties of left Γ-conformal algebras (see Definition IV.1). Conversely, if we have a Γ-conformal algebra R and a homomorphism $\chi:\Gamma\to GL_2(\mathbb{C})$, we can construct a Lie algebra $\mathfrak{s}(R,\chi)$ of $\chi(\Gamma)$ local formal distributions as follows. Consider a vector space over $\mathbb{C}$ with the basis a_n, $a\in R$, $n\in\mathbb{Z}$, and denote by $\mathfrak{s}(R,\chi)$ the quotient of this space by the $\mathbb{C}$-space of elements of the form:

$$\begin{aligned}&(\lambda a+\beta b)_n-\lambda a_n-\beta b_n,\quad \lambda,\beta\in\mathbb{C},\quad a,b\in R\\ &(T_\alpha^{-1}a)_n-(\alpha'(z)a(\alpha(z)))_n,\quad \text{where}\ \ \alpha\in\Gamma,\end{aligned} \tag{6.7}$$

and $(\alpha'(z)a(\alpha(z)))_n$ is the coefficient of z^{-n-1} in the Fourier series decomposition. Then the formula

$$[a(z),b(w)]=\sum_{\alpha\in\Gamma}(a_{(\alpha^{-1})}b)(w)\,\delta(z-\alpha(w)) \tag{6.8}$$

gives a Lie algebra structure on $\mathfrak{s}(R,\chi)$ of Γ-local formal distributions $a(z)=\Sigma_{n\in\mathbb{Z}}a_nz^{-n-1}$.

Remark VI.1: If we let $a(z)T_\alpha=\alpha'(z)a(\alpha(z))$, then we have all properties of right Γ-conformal algebras. We prefer (the more customary) left Γ-modules.

Consider some important examples.

Example VI.1: Lie algebra of pseudodifferential operators on the circle.

Let R be the general $\mathbb{Z}$-conformal algebra $gc_1(\mathbb{Z})$ defined by (5.7). Fix a homomorphism $\chi:\mathbb{Z}\to GL_2(\mathbb{C})$ defined as $\chi(n)=\left(\begin{smallmatrix}1 & n\\ 0 & 1\end{smallmatrix}\right)$, $n\in\mathbb{Z}$. The corresponding Lie algebra $\mathfrak{s}(\mathbb{Z},\chi)$ is a Lie algebra with basis a_k^m, $m,k\in\mathbb{Z}$. Let $a^m(z)=\Sigma_k a_k^mz^{-k-1}$ be the generating function. Then from (6.8) we have:

$$\begin{aligned}[a^m(z),a^n(w)]&=\sum_{s\in\mathbb{Z}}(a^m_{(s)}a^n)(w)\,\delta(z-w+s)\\ &=(T_{-m}a^{m+n})(w)\,\delta(z-w-m)-a^{m+n}(w)\,\delta(z-w+n).\end{aligned} \tag{6.9}$$

To write down explicitly the commutation relations for coefficients of formal distributions $a^m(z)$, consider the Lie algebra $\mathrm{pdiff}(S^1)$ of pseudodifferential operators on the circle (see Refs. 9–11 for more details). This is a Lie algebra of the associative algebra with the basis $x^m\partial^n$, $m,n\in\mathbb{Z}$. The commutation relations are defined from the commutation relations on x and ∂:

$$\partial\circ f(x)=f(x)\circ\partial+f'(x),$$

$$\partial^{-1}\circ f(x)=\sum_{n=0}^{\infty}(-1)^nf^{(n)}(x)\partial^{-n-1},$$

where $f(x)\in C^\infty(S^1)$. Introduce a new basis of pdiff(S^1) with generators x^mD^k, $m,k\in\mathbb{Z}$, where $D=x\partial$. These generators have simpler commutation relations:

$$[x^mD^k,x^nD^\prime]=x^{m+k}((D+n)^kD^\prime-(D+m)^\prime D^k). \tag{6.10}$$

For $k<0$ we will understand $(D+n)^k$ as an expansion by the negative powers of D (see (6.3)). Introduce the generating function:

$$a^m(z)=-\sum_{k\in\mathbb{Z}}x^{-m}D^kz^{-k-1}. \tag{6.11}$$

Using the property (6.4) of δ-function, we have

$$\begin{aligned}[a^m(z),a^n(w)]&=\sum_{k,l\in\mathbb{Z}}x^{-(m+n)}((D-n)^kD^\prime z^{-k-1}w^{-\prime-1}\\&\quad-(D-m)^\prime D^kz^{-k-1}w^{-\prime-1})\\&=x^{-(m+n)}\Bigg(\sum_{k,\prime\in\mathbb{Z}}D^kD^\prime(z+n)^{-k-1}w^{-\prime-k-1}w^k\\&\quad-D^\prime D^k(w+m)^{-\prime-k-1}(w+m)^kz^{-k-1}\Bigg)\\&=a^{m+n}(w+m)\,\delta(z-w-m)-a^{m+n}(w)\,\delta(z-w-n).\end{aligned} \tag{6.12}$$

Comparing (6.9) and (6.12), we see that the Lie algebra $\mathfrak{s}(\Gamma,\chi)$ is isomorphic to the Lie algebra pdiff(S^1).

Thus, we have shown, that the Lie algebra of $q-$pdiff(S^1) [see Example IV.2 (a)] and the Lie algebra of pdiff(S^1) correspond to the same general $\mathbb{Z}$-conformal algebra $gc_1(\mathbb{Z})$, but to the different homomorphisms $\mathbb{Z}\to GL_2(\mathbb{C})$.

More generally, we can consider the general conformal algebra $gc_1(GL_2(\mathbb{C}))$ with the usual action of $GL_2(\mathbb{C})$ on $\mathbb{C}$. Let $A,B,C\in GL_2(\mathbb{C})$. By (5.7) we have:

$$a^A_{(C)}a^B=\delta_{C,A^{-1}}T_{A^{-1}}a^{AB}-\delta_{C,B}a^{BA}.$$

It is clear, that the general conformal algebra admit reduction to the subalgebra. This reduction is well defined at the level of the Lie algebra $\mathfrak{s}(g_1(GL(\mathbb{C})))$:

$$[a^A(z),a^B(w)]=A(w)^\prime_wa^{AB}(A(w))\,\delta(z-A(w))-a^{BA}(w)\,\delta(z-B^{-1}(w)). \tag{6.13}$$

As was shown, for the subgroup $H_1=\{\left(\begin{smallmatrix}q^{-m/2}&0\\0&q^{m/2}\end{smallmatrix}\right),m\in\mathbb{Z}\}$ we get the Lie algebra of $q-$pdiff(S^1); for the subgroup $H_2=\{\left(\begin{smallmatrix}1&m\\0&1\end{smallmatrix}\right),\ m\in\mathbb{Z}\}$ we get the Lie algebra pdiff(S^1). For the subgroup $H_3\subset GL_2(\mathbb{C})$, generated by two elements $a=\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ and $b=\left(\begin{smallmatrix}q^{-1/2}&\\&q^{-1/2}\end{smallmatrix}\right)$ we get the Lie algebra, that contains as subalgebra the Lie algebra of pdiff(S^1) and $q-$pdiff(S^1). Taking different subgroups H in $GL(2,\mathbb{C})$ (not necessarily discrete) we get a family of infinite-dimensional Lie algebras $\mathfrak{s}(gc_1(H))$.

Example VI.2: Consider the group Γ generated by two elements ϵ and T with relations: $\epsilon^2=1$ and $\epsilon T=T^{-1}\epsilon$. Fix two homomorphisms $\varphi_i:\Gamma\to\mathrm{Aut}(g_\infty)$, $i=1,2$.

$$\varphi_1:\begin{cases}\epsilon(E_{ij})=-E_{-j,-i}\\T(E_{ij})=E_{i+1,j+1}\end{cases}\quad\text{and}\quad\varphi_2:\begin{cases}\epsilon(E_{ij})=E_{-i,-j}\\T(E_{ij})=E_{i+1,j+1}\end{cases}.$$

In both cases g_∞ is a free $\mathbb{C}[\Gamma]$ module with the basis $A^m=E_{0,m}$, $m\in\mathbb{Z}_+$. The Γ-products on generators A^m are as follows:

(a) For φ_1 we have:

$$A^m_{(\gamma)}A^n=[T^rA^m,A^n]=\delta_{r,-m}T^{-m}A^{m+n}-\delta_{r,n}A^{m+n} \quad \text{for } \gamma=T^r,$$

$$A^m_{(\gamma)}A^n=[\epsilon T^rA^m,A^n]=-\delta_{r,0}T^{-m}A^{n+m}+\delta_{m+n,-r}A^{m+n} \quad \text{for } \gamma=\epsilon T^r.$$

(b) For φ_2 we have:

$$A^m_{(\gamma)}A^n=\delta_{r,-m}T^{-m}A^{m+n}-\delta_{r,n}A^{m+n} \quad \text{for } \gamma=T^r$$

$$A^m_{(\gamma)}A^n=-\delta_{r,-m}T^mB^{n-m}-\delta_{r,-n}B^{n-m} \quad \text{for } \gamma=\epsilon T^r \text{ and } n\geqslant m.$$

These give us two structures of a Γ-conformal algebra, which we will denote R_1 and R_2 respectively. Consider two homomorphisms $\chi_i:\Gamma\to GL_2(\mathbb{C})$, $i=1,2$:

$$\text{(a)}\chi_1(\epsilon)=\begin{pmatrix}0&1\\-1&0\end{pmatrix}, \quad \epsilon_1(z)=-\frac{1}{z};$$

$$\chi_1(T)=\begin{pmatrix}q^{1/2}&0\\0&q^{-1/2}\end{pmatrix}, \quad T_1(z)=qz,$$

$$\text{(b)}\chi_2(\epsilon)=\begin{pmatrix}1&0\\0&-1\end{pmatrix}, \quad \epsilon_2(z)=-z;$$

$$\chi_2(T)=\begin{pmatrix}1&1\\0&1\end{pmatrix}, \quad T(z)=z+1.$$

We can define four Lie algebras $\mathfrak{s}(R_i,\chi_j)$, $i,j=1,2$, of Γ-local formal distributions with generators A^m_n, $m\in\mathbb{Z}_+$, $n\in\mathbb{Z}$, and the generating function $A^m(z)=\Sigma_nA^m_nz^{-n-1}$.

(a) $S(R,\chi_1)$:

$$=q^mA^{m+n}(q^mw)\,\delta(z-q^mw)-A^{m+n}(w)\,\delta(z-q^{-n}w)$$

$$-q^mA^{m+n}(q^mw)\,\delta\left(z+\frac{1}{w}\right)+A^{m+n(w)}\,\delta\left(z+\frac{q^{m+n}}{w}\right).$$

For the coefficients A^m_k we have the commutation relations

$$[A^m_k,A^n_\ell]=(q^{-m\ell}-q^{-nk})A^{m+n}_{k+\ell}-(-1)^kq^{mk}(q^{-n\ell}-q^{nk})A^{m+n}_{\ell-k}.$$

This is the C-series of Lie algebras of quantum torus (Ref. 2).

(b) $\mathfrak{s}(R_2,\chi_1)$:

$$[A^m(z),A^n(z)]=A^{m+n}(q^mw)\,\delta(z-q^nw)-A^{m+n}(z-q^{-n}w)$$

$$-q^{-m}A^{n-m}(q^{-m}w)\,\delta\left(z+\frac{q^m}{w}\right)+A^{n-m}(w)\,\delta\left(z+\frac{q^n}{w}\right),$$

$$\text{for } m,n\in\mathbb{Z},n\geqslant m$$

and

$$[A^m_k,A^n_\ell]=(q^{-m\ell}-q^{-k\ell})A^{m+n}_{k+\ell}-(-1)^k(q^{m\ell}-q^{nk})A^{n-m}_{\ell-k}.$$

(c) $S(R_1,\chi_2)$:

$$[A^m(z),A^n(w)]=A^{m+n}(w+m)\,\delta(z-w-m)-A^{m+n}(w)\,\delta(z-w+n)$$

$$-A^{m+n}(w+m)\,\delta(z+w)+A^{m+n}(w)\,\delta(z+w-m-n). \tag{6.14}$$

Consider in pdiff(S^1) with the basis (6.11) a subalgebra, stable under the automorphism ω defined by: $\omega(x^m D^k)=x^m(m-D)^k$. This is a Lie algebra with the basis $x^m D^k-\omega(x^m D^k)$. Introduce the generating function of the form:

$$C^m(z)=-\sum_{k\in\mathbb{Z}}(x^{-m}D^k-x^{-m}(-m-D)^k)z^{-k-1}$$
$$=a^m(z)+a^m(m-z),$$

where $a^m(z)$ are given by (6.12). It is easy to check, that the fields $C^m(z)$ satisfy equation (6.15). We call this subalgebra a C-series of pdiff(S^1).

(d) $S(R_2,\chi_2)$: For generating functions we have:

$$[A^m(z),A^n(w)]=A^{m+n}(w+m)\,\delta(z-w-m)-A^{m+n}\delta(z-w+n)$$
$$+A^{n-m}(w-m)\,\delta(z+w-m)-A^{n-m}\delta(z+w+n), \tag{6.15}$$

for $m,n\in\mathbb{Z}_+$, and $n\geqslant m$. This subalgebra is called the B-series of pdiff(S^1). It can be defined as a subalgebra stable under the second-order automorphism σ defined by $\sigma(x^n D^n)=x^{-n}(-D)^n$. Relations (6.15) are exactly the relations on generating series $B^m(z)=\Sigma_{k\in\mathbb{Z}}(x^{-m}D^k+\sigma(x^{-m}D^k))z^{-k-1}=a^m(z)-a^{-m}(-z)$.

ACKNOWLEDGMENTS

The research of M. Golenishcheva-Kutuzova was supported in part by INTAS Grant No. 942317 and Dutch NWO organization and by RFFI Grant No. 96-0218046. M.G.-K. is also very grateful to the Institut Girard Desarques, URA 746, Université Lyon-1 for the kind hospitality and support.

[1] V. G. Kac, D. A. Kazhdan, J. Lepowsky, and R. L. Wilson, ''Realization of the basic representations of the Euclidean Lie algebras,'' Adv. Math. **42**, 83–112 (1981).
[2] M. I. Golenishcheva-Kutuzova and D. Lebedev, ''Vertex operator representation of some quantum tori Lie algebras,'' Commun. Math. Phys. **148**, 403–416 (1992).
[3] I. Frenkel and N. Jing, ''Vertex representations of quantum affine algebras,'' Proc. Natl. Acad. Sci. USA **85**, 9373–9377 (1988).
[4] S. Khoroshkin, D. Lebedev, and S. Pakuliak, ''Intertwining operators for the central extension of the Yangian double,'' Phys. Lett. A **222**, 381–392 (1996).
[5] V. G. Kac, *Vertex algebras for beginners*, University lecture series, ISSN, Vol. 10 (American Mathematical Society, Providence, RI, 1996).
[6] V. G. Kac, ''The idea of locality,'' in *Physical application and mathematical aspects of geometry, groups, and algebra*, edited by H.-D. Doebner *et al.* (World Scientific, Singapore, 1997), pp. 16–32.
[7] A. D'Andrea and V. G. Kac, *Structure theory of finite conformal algebra*, preprint 1997.
[8] V. G. Kac, *Infinite dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, MA, 1990).
[9] B. Khesin, V. Lyubashenko, and C. Roger, ''Extensions and contractions of the Lie Algebra of q-pseudodifferential symbols on the circle,'' J. Funct. Anal. **143**, 55–97 (1997).
[10] B. Khesin and I. Zakharevich, ''Poisson-Lie group of pseudodifferential operators,'' Commun. Math. Phys. **171**, 475–530 (1995).
[11] A. Orlov and P. Winternitz, ''P_∞ algebra of KP, free fermions and 2-cycle in the Lie algebra of pseudodifferential operators,'' solv-int/9701008 V2.

A natural basis for spinor and vector fields on the noncommutative sphere

J. Gratus[a)]
Laboratoire de Gravitation et Cosmologie Relativistes,[b)] Tour 22/12 4eme etage, Boite Courrier142, 4pl Jussieu, F75252 Paris, France

(Received 9 September 1997; accepted for publication 2 December 1997)

The product of two Heisenberg–Weil algebras contains the Jordan–Schwinger representation of $su(2)$. This algebra is quotiented by the square-root of the Casimir to produce a nonassociative algebra denoted by Ψ. This algebra may be viewed as the right-module over one of its associative subalgebras which corresponds to the algebra of scalar fields on the noncommutative sphere. It is now possible to interpret other subspaces as the space of spinor or vector fields on the noncommutative sphere. A natural basis of Ψ is given which may be interpreted as the deformed entries in the rotation matrices of $SU(2)$. © *1998 American Institute of Physics.* [S0022-2488(98)02803-5]

I. INTRODUCTION

The noncommutative or ''fuzzy'' sphere has been considered by several authors in different contexts, including the general quantization procedure, coherent states, noncommutative geometry, the theory of membranes, and the quantum Hall effect (see references in 1–4).

Normally the approximation for the algebra of functions on a sphere (scalar fields) is in terms of matrices,[2,5,6] where the functions on a sphere appear only in the limit as the size of the matrix tends to infinity. An alternative approach was presented in Ref. 1 as a two parameter algebra of polynomials $\mathscr{P}(\varepsilon,R)$ with $\varepsilon,R\in\mathbb{R}$.

The standard way of defining a vector field is to consider it as a derivation on the algebra of functions. If the latter is replaced by a matrix algebra then all such derivations are the adjoint action of elements in that algebra.[2,7] As such they do not form a module over the algebra of functions like standard vector fields do. If $\{x^1,x^2,x^3\}$ are the coordinate functions for the noncommutative sphere then the corresponding derivations $X_i=\mathrm{ad}_{x^i}=[x_i,\cdot]$ obey the equation $\Sigma_{i=1}^3 x^i X_i=0$ only in the commutative limit. Here we give an alternative definition for the analogue of vector fields which both forms a module over the algebra of functions and solves this equation. However, these vector fields are derivations only in the limit $\varepsilon=0$.

The idea of using the Jordan–Schwinger representation of $su(2)$ to describe spinor fields have been pursued by many authors.[3,4,8–10] They discovered that one could view spinor fields as rectangular matrices.

This article may be consider as an extension of Ref. 1 to spinor and vector fields by the use of the Jordan–Schwinger representation of $su(2)$.

A. Structure of the article

We present $\mathscr{W}$, an algebra which is the product of two Heisenberg–Weil algebra, generated by $[a_-,a_+]=\varepsilon$ and $[b_-,b_+]=\varepsilon$. This naturally contains the Jordan-Schwinger representation of both $su(2)$ given by $\{J_0,J_+,J_-\}$ and $su(1,1)$ given by $\{K_0,K_+,K_-\}$ defined in (3).

The idea is to consider the algebra generated by $\{J_0,J_+,J_-\}$ to be equivalent to the analogue of the algebra of functions (or scalar fields) on the sphere as presented in Ref. 1. In this paper the algebra generated by $\{J_0,J_+,J_-\}$, which is the universal covering algebra of $su(2)$, is quotiented by the ideal generated by the Casimir $J_0^2+\frac{1}{2}J_+J_-+\frac{1}{2}J_-J_+\sim R^2$ where $R^2\in\mathbb{R}$. Since the Casimir

---

[a)]Present address: Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom.
[b)]Laboratoire associé au CNRS URA 769.

 © 1998 American Institute of Physics

operator commutes with the set $\{J_0, J_+, J_-\}$ this ideal is two sided and the corresponding quotient is an algebra.

In this paper, by contrast, the Casimir operator is given by $J_0^2+\frac{1}{2}J_+J_-+\frac{1}{2}J_-J_+=K_0^2-\frac{1}{4}\varepsilon^2$. We would like to take the square root of this equation and consider quotienting $\mathscr{W}$ by the ideal generated by $K_0\sim\hat{R}=(R^2+\frac{1}{4}\varepsilon^2)^{1/2}$. However, although K_0 commutes with any polynomial in $\{J_0, J_+, J_-\}$, it does not commute with all the elements in $\mathscr{W}$. Therefore the left ideal generated by $K_0\sim\hat{R}$ is not two sided and the corresponding quotient product is not associative. However our persistence is rewarded.

We define the subspace $\Psi\subset\mathscr{W}$ as all the totally symmetric polynomials in (a_+, a_-, b_+, b_-) which contain no factor of K_0. This is achieved by defining the *formal trace*. We can define two products on Ψ, ρ, and $\rho^\star$ by quotienting by either the left or right ideal generated by $K_0\sim\hat{R}$.

There is a natural basis of Ψ given by the set $\{\psi_n^{(r,m)}|2n, n+r, n+m \in\mathbb{Z}, n\geqslant 0, |r|\leqslant n|m|\leqslant n\}$. The basis element $\psi_n^{(r,m)}$ is the unique (up to scale factor) homogeneous polynomial of degree $2n$, which is an eigenstate of both ad_{K_0} and ad_{J_0} with corresponding eigenvalues εr and εm. These basis elements are also orthogonal with respect to a sesquilinear form defined in a similar way to the sesquilinear form in Ref. 1.

These basis elements may be thought of as entries in a deformed rotation matrix, since, in the limit $\varepsilon=0$, $\psi_n^{(r,m)}$ is proportional to $D^n_{mr}(\alpha,\beta,\gamma)$ where α,β,γ are the Euler angles. As a result we may interpret the algebra (Ψ,ρ) as the noncommutative and nonassociative analogue of functions on the group $SU(2)$.

We label the eigenspace $\Psi_\bullet^{(r,\bullet)}=\mathrm{span}\{\psi_n^{(r,m)}\forall n,m\}$. Since the eigenspace $\Psi_\bullet^{(0,\bullet)}$ corresponds to all elements of Ψ which commute with K_0, these elements can be written as polynomials in $\{J_0, J_+, J_-\}$ and thus correspond to the scalar fields on the noncommutative sphere. Now, as stated, the product ρ on Ψ is not associative, but we show that $\Psi_\bullet^{(r,\bullet)}$, for each r, can be viewed as a right module over $\Psi_\bullet^{(0,\bullet)}$. This is analogous to the fact that standard spinor and vector fields are modules over the space of scalar fields.

It is natural to call the set $\Psi_\bullet^{(1,\bullet)}$ the analogue of vectors fields, and to call $\Psi_\bullet^{(-1,\bullet)}$ the analogue of covector fields or 1-forms. Using this interpretation we may define an exterior ''derivative'' $d:\Psi_\bullet^{(0,\bullet)}\mapsto\Psi_\bullet^{(-1,\bullet)}$. This map obeys the Leibniz rule only in the limit $\varepsilon=0$. Likewise the interpretation of vectors as derivations on the algebra of function is also true only in this limit. By contrast the equation $x^iX_i=0$ is true for all ε not just in the limit. Thus its swings and roundabouts: one must use one's personal conviction, or a physical interpretation, to decide which properties of vector fields are fundamental.

The spaces $\Psi_\bullet^{(1/2,\bullet)}$ and $\Psi_\bullet^{(-1/2,\bullet)}$ are the analogues of spinor fields on the noncommutative sphere. They may be viewed as contra-variant and co-variant spinor fields, or positive and negative chiral spinor fields depending on one's interpretation. We justify the definition of a spinor field on two grounds. First the product of two spinor fields is a vector, and secondly, the rotation of a spinor though 2π inverses the sign of the spinor. Alternatively, writing the spinors as 2-dimensional vectors with entries in $\Psi_\bullet^{(-1/2,\bullet)}\oplus\Psi_\bullet^{(1/2,\bullet)}$, we obtain the standard definition of spinors in the limit $\varepsilon=0$.

B. Contents of the article

In most of this article we are concerned with the mathematical structure necessary to define Ψ and its products. In section II, we examine the Heisenberg–Weil algebra $\mathscr{W}$. We define the formal trace and give some of its properties. In section III, we present the vector space Ψ and some of its subspaces. We define the nonassociative products ρ and $\rho^\star$ and show that when they are restricted to the space of scalar fields they are associative. We also define the sesquilinear Hermitian product that fails to be positive definite, (although it is positive definite in the case $\varepsilon=0$).

In section IV, we define the orthogonal basis for Ψ in a similar vain to that in Ref. 1. In section V, we give useful formulae for calculating the product ρ on Ψ. We show how to write the basis elements either as rectangular matrices or in terms of Hahn polynomials.

In section VI, we demonstrate how $\psi_n^{(r,m)}$ may be regarded as the deformed rotation matrix entry, and also how $\Psi_\bullet^{(1,\bullet)}$ can be viewed as the space of vector fields and $\Psi_\bullet^{(-1,\bullet)}$ the space of covector fields, and $\Psi_\bullet^{(1/2,\bullet)}$ as the space of spinor fields on the noncommutative sphere.

Finally, in section VII we discuss some of the problems with this interpretation. We also suggest how this methodology can be extended to scalar, spinor, and vector fields on more exotic symmetric spaces such as the Einstein–deSitter universe.

C. Notations

The summation convention is not used in this article.

The notational difference with Ref. 1 is that (1) the basis elements are not normalized; (2) we use ε instead of κ; and (3) since the product is not associative, it is written explicitly whenever used.

$\hat{R}$ is always given in relation to R by (23). In most cases we do not write the dependence on R and ε explicitly unless there is room for confusion. The implicit dependencies on $\hat{R}$ and ε are given by

	$\hat{R}$	ε
$\mathscr{W}$ as a vector space	No	No
$\mathscr{W}$ as an algebra	No	Yes
$S(s,t,u,v)$	No	No
Ψ, $\Psi_n^{(\bullet,\bullet)}$, $\Psi_\bullet^{(r,\bullet)}$, $\Psi_\bullet^{(\bullet,m)}$ as vector spaces	No	No
$\rho:\mathscr{W}\mapsto\Psi$ as a projection	Yes	No
$\rho:\Psi\times\Psi$ as a product	Yes	Yes
$\psi_n^{(r,m)}$	No	No

We shall use the following notation for elements of each set:

Space	Description	General elements	Basis
$\mathscr{W}$	Polynomials in (a_+,a_-,b_+,b_-)	$w,w_1,w_2,\ldots$	$S(s,t,u,v)$
Ψ	Formally traceless symmetric polynomials	$\xi,\zeta,\ldots$	$\psi_n^{(r,m)}$
$\Psi_\bullet^{(0,\bullet)}$	Analogue of functions	$f,g,\ldots$	$\psi_n^{(0,m)}$
$\Psi_\bullet^{(1,\bullet)}$	Analogue of vectors fields	$X,Y,\ldots$	$\psi_n^{(1,m)}$
$\Psi_\bullet^{(-1,\bullet)}$	Analogue of 1-forms	$\xi,\zeta,\ldots$	$\psi_n^{(-1,m)}$
$\Psi_\bullet^{(1/2,\bullet)}$	Analogue of spinors	$\xi,\zeta,\ldots$	$\psi_n^{(1/2,m)}$

II. $\mathscr{W}$-EXPRESSIONS: THE FREE ALGEBRA OF THE PRODUCT OF TWO HEISENBERG–WEIL ALGEBRAS

As stated in the Introduction in this section we define the algebra $\mathscr{W}=\mathscr{W}(\varepsilon)$ which is the product of two Heisenberg–Weil algebras. This naturally contains the Jordan Schwinger representation of $su(2)$ and $su(1,1)$. We define a basis for this algebra in terms totally symmetric polynomials. Since we wish to quotient out by $K_0=\hat{R}$ we need to find representative of the $\mathscr{W}$ which contain no factor of K_0. This is achieved by defining the *formal trace*. We give some basic properties of the formal trace including that it commutes with the adjoint representation of $su(2)$. Finally we show a convenient way of writing the elements of $\mathscr{W}$ as nonsymmetric polynomials.

For $\varepsilon\in\mathbb{R}$ let

$$\mathscr{W}(\varepsilon)=\Big\{\text{free algebra of finite polynomials in}(a_+,a_-,b_+,b_-)\Big\}/\sim, \tag{1}$$

where

$$[a_-,a_+]\sim\varepsilon,\quad [b_-,b_+]\sim\varepsilon,\quad [a_\pm,b_\pm]\sim 0. \tag{2}$$

This may be viewed as the tensor product of two copies of the Heisenberg–Weil algebra $\mathscr{W}(\varepsilon)=\mathscr{W}_A(\varepsilon)\otimes\mathscr{W}_B(\varepsilon)$ where $\mathscr{W}_A(\varepsilon)$ is the algebra of polynomials of (a_+,a_-), and likewise for $\mathscr{W}_B(\varepsilon)$.

From now on we write $\mathscr{W}=\mathscr{W}(\varepsilon)$ when there is no doubt about ε. A natural basis for $\mathscr{W}$ is given by $S(s,t,u,v)$ which is the totally symmetric homogeneous polynomial with each term a permutation of $a_-^s a_+^t b_-^u b_+^v$ and having coefficient 1.

Let $\mathscr{W}^n$ for $2n\in\mathbb{Z}$, $n\geqslant 0$ be the space of all symmetric homogeneous polynomials of degree $2n$. The basis for $\mathscr{W}^n$ is now given by $\{S(s,t,u,v)|s+t+u+v=2n\}$. It is easy to show the dimension of $\mathscr{W}^n$ is given by $\dim(\mathscr{W}^n)=\frac{1}{6}(2n+1)(2n+2)(2n+3)$.

The subspace $\mathscr{W}^1$ has dimension 10. Out of these 6 elements turn out to be very important and are given special names (written in a nonsymmetric but more convenient form):

$$\begin{gathered} J_0=\tfrac{1}{2}(a_+a_--b_+b_-),\quad J_+=a_+b_-,\quad J_-=a_-b_+,\\ K_0=\tfrac{1}{2}(a_+a_-+b_+b_-+\varepsilon),\quad K_+=a_+b_+,\quad K_-=a_-b_-. \end{gathered} \tag{3}$$

It is easy to show that $\{J_0,J_+,J_-\}$ forms a representation of $su(2)$ with $[J_0,J_+]=\varepsilon J_+$, etc., and that $\{K_0,K_+,K_-\}$ form a representation of $su(1,1)$ with $[K_0,K_+]=\varepsilon K_+$. The Casimirs of these are given by

$$J_0^2+\tfrac{1}{2}J_+J_-+\tfrac{1}{2}J_-J_+=K_0^2-\tfrac{1}{4}\varepsilon^2, \tag{4}$$

$$K_0^2-\tfrac{1}{2}K_+K_--\tfrac{1}{2}K_-K_+=J_0^2-\tfrac{1}{4}\varepsilon^2. \tag{5}$$

Furthermore K_0 commutes with $\{J_0,J_+,J_-\}$ and J_0 commutes with $\{K_0,K_+,K_-\}$. The entire set $\mathscr{W}^1$ forms a representation of $so(3,2)$.[11]

There is a *conjugation* of elements in $\mathscr{W}$ given by

$$\dagger:\mathscr{W}\mapsto\mathscr{W},$$

$$(w_1w_2)^\dagger=w_2^\dagger w_1^\dagger,(a_\pm)^\dagger=a_\mp,(b_\pm)^\dagger=b_\mp,\lambda^\dagger=\overline{\lambda},\forall w_1,w_2\in\mathscr{W},\lambda\in\mathbb{C}. \tag{6}$$

On the basis elements it is easy to show that

$$S(s,t,u,v)^\dagger=S(t,s,v,u). \tag{7}$$

We define the *formal trace* of an element $w\in\mathscr{W}$ as follows:

$$\mathrm{tr}:\mathscr{W}\mapsto\mathscr{W},:\mathscr{W}^n\mapsto\mathscr{W}^{n-1} \tag{8}$$

with w as a totally symmetric polynomial. There are sixteen possible combinations for the last two elements of each term. Collecting these terms together, we can now write

$$w=w_1a_+a_-+w_2a_-a_++w_3b_+b_-+w_4b_-b_++w_5a_+^2+\cdots+w_{16}b_-^2.$$

Then

$$\mathrm{tr}(w)=w_1+w_2+w_3+w_4. \tag{9}$$

The *adjoint* is given by $\mathrm{ad}_T(w)=[T,w]$ for $T,w\in\mathscr{W}$. This vanishes if $\varepsilon=0$. However the limit $\lim_{\varepsilon\to0}(1/\varepsilon)\mathrm{ad}_T(w)$ is defined. If T is in the set $\{J_0,J_+,J_-,K_0\}$ then we can write $T=\Sigma_{\mu,\nu=0}^2\chi_+^\mu T^{\mu\nu}\chi_-^\nu$. Where $\chi_\pm^1=a_\pm$ and $\chi_\pm^2=b_\pm$. In this case, which covers most situations, we have

$$\lim_{\varepsilon\to0}(1/\varepsilon)\mathrm{ad}_T(w)=\sum_{\mu,\nu=0}^2\left(\chi_+^\mu T^{\mu\nu}\frac{\partial w}{\chi_-^\nu}-\frac{\partial w}{\chi_+^\mu}T^{\mu\nu}\chi_-^\nu\right). \tag{10}$$

Lemma 1: The formal trace of a basis element is given by

$$\mathrm{tr}(S(s,t,u,v))=2S(s-1,t-1,u,v)+2S(s,t,u-1,v-1). \tag{11}$$

This commutes with the adjoint action of J_0,J_+,J_- and K_0,

$$\mathrm{tr}\circ\mathrm{ad}_{J_0}=\mathrm{ad}_{J_0}\circ\mathrm{tr},\quad \mathrm{tr}\circ\mathrm{ad}_{J_+}=\mathrm{ad}_{J_+}\circ\mathrm{tr},\quad \mathrm{tr}\circ\mathrm{ad}_{J_-}=\mathrm{ad}_{J_-}\circ\mathrm{tr},\quad \mathrm{tr}\circ\mathrm{ad}_{K_0}=\mathrm{ad}_{K_0}\circ\mathrm{tr}, \tag{12}$$

but not with K_+, K_-. *It also commutes with taking the conjugate* $\mathrm{tr}(w^\dagger)=(\mathrm{tr}(w))^\dagger$.

Proof: By reasoning similar to the appendix in Ref. 1, we can split the basis element for any $d\in\mathbb{Z}^+$,

$$S(u_1,u_2,u_3,u_4)=\sum_{v_1+v_2+v_3+v_4=d} S(u_1-v_1,u_2-v_2,u_3-v_3,u_4-v_4)S(v_1,v_2,v_3,v_4).$$

Setting $d=2$ then

$$\mathrm{tr}(S(s,t,u,v))=\widetilde{\mathrm{tr}}(S(s-1,t-1,u,v)S(1,1,0,0))+\widetilde{\mathrm{tr}}(S(s,t,u-1,v-1)S(0,0,1,1)),$$

since all other terms vanish, hence (11).

For (12) we need to set up a basis for each $\mathscr{W}_A$ and $\mathscr{W}_B$ given by $S_A(s,t)$ the totally symmetric homogeneous polynomial with each term a permutation of $a_-^s a_+^t$ and having coefficient 1. Likewise for $S_B(u,v)$. These are related to $S(s,t,u,v)$ by

$$(s+t)!(u+v)!S(s,t,u,v)=(s+t+u+v)!S_A(s,t)S_B(u,v).$$

The commutator and anti-commutator of $S_A(s,t)$ with $a_\pm$ is given by

$$[a_+,S_A(s,t)]=-\varepsilon(s+t)S_A(s,t-1),\ \ [a_-,S_A(s,t)]=\varepsilon(s+t)S_A(s-1,t),$$

$$[a_+,S_A(s,t)]_+=2s(s+t+1)^{-1}S_A(s,t-1),\ \ [a_-,S_A(s,t)]_+=2t(s+t+1)^{-1}S_A(s-1,t),$$

where $[w_1,w_2]_+=w_1w_2+w_2w_1$ is the anti-commutator. The expressions for the commutators may be proved by induction on the degree $(s+t)$. The expressions for the anti-commutators by use of the formula

$$[a_+a_-,S_A(s,t)]=\varepsilon(s+t)S_A(s,t).$$

Now Eqs. (12) follow from direct calculation.

Lemma 2: If $w\in\mathscr{W}$ *then it may be written as a sum of elements,*

$$w=\sum_{r,m} w_{rm},\quad \text{where,}\quad \mathrm{ad}_{K_0}w=\varepsilon rw,\quad \mathrm{ad}_{J_0}w=\varepsilon mw. \tag{13}$$

We can write w_{rm} *as a nonsymmetric polynomial,*

$$w_{rm}=a_\pm^{r+m}b_\pm^{r-m}p_{rm}(J_0,K_0), \tag{14}$$

where

$$a_\pm^r=\begin{cases} a_+^r, & \text{if } r>0,\\ 1, & \text{if } r=0,\\ a_-^{-r}, & \text{if } r<0,\end{cases} \tag{15}$$

and likewise for $b_\pm^r$ *and where* $p_{rm}(J_0,K_0)$ *is a polynomial in* J_0 *and* K_0.

Proof: Since ad_{K_0} and ad_{J_0} commute we can diagonalize w with respect to these operators.

Since the $a_\pm$ commute with the $b_\pm$ collect all the $a_\pm$'s together. Use $a_-a_+=J_0+K_0+\varepsilon/2$ to remove a_-a_+ and $a_+a_-=J_0+K_0-\varepsilon/2$ to remove a_+a_-. By using $[J_0,a_+]=(\varepsilon/2)a_+$, etc., we can move all J_0 and K_0 to the right. From (13) we know that for each term in w the number of a_+ minus the number of a_- is $m+r$. Thus what remains on the left is $a_\pm^{m+r}$. Similarly with $b_\pm$, hence result. □

III. (Ψ,ρ): A NONASSOCIATIVE AND NONCOMMUTATIVE ALGEBRA OF FORMALLY TRACELESS SYMMETRIC POLYNOMIALS IN $\mathscr{W}$

In this section we introduce the subspace Ψ of $\mathscr{W}$ of all formally traceless symmetric polynomials. We introduce two products ρ and $\rho^\star$ on Ψ neither of which are associative, and show they are well defined. In lemma 4 we show that Ψ contains a subspace $\Psi_\bullet^{(0,\bullet)}$ which is an associative algebra, analogous to the algebra of functions on a sphere, and that the spaces $\Psi_\bullet^{(r,\bullet)}$ may be viewed as modules over $\Psi_\bullet^{(0,\bullet)}$. We also show the order to which Ψ is associative.

Let $\Psi\subset\mathscr{W}$ be the subspace of all traceless symmetric polynomials in (a_+,a_-,b_+,b_-):

$$\Psi=\ker(\mathrm{tr})\subset\mathscr{W}. \tag{16}$$

We define the subspace

$$\Psi_n^{(\bullet,\bullet)}=\ker(\mathrm{tr})\cap\mathscr{W}^n,\ \text{for } 2n\in\mathbb{Z} \text{ and } n\geqslant 0, \tag{17}$$

i.e., the space of all formally traceless symmetric homogeneous polynomials of order n. Since $\Psi_n^{(\bullet,\bullet)}$ is the kernel of the restriction $\mathrm{tr}:\mathscr{W}^n\mapsto\mathscr{W}^{n-1}$, which is surjective, the dimension of $\Psi_n^{(\bullet,\bullet)}$ is $\dim(\Psi_n^{(\bullet,\bullet)})=(2n+1)^2$. We define the projection π_n,

$$\pi_n:\Psi\mapsto\bigoplus_{m=0}^{2n}\Psi_{m/2}^{(\bullet,\bullet)}. \tag{18}$$

We may divide the space Ψ in three different ways, since for each $\xi\in\Psi$ both $\mathrm{ad}_{J_0}(\xi),\mathrm{ad}_{K_0}(\xi)\in\Psi$:

$$\Psi_\bullet^{(\bullet,m)}=\{\xi\in\Psi\,|\,\mathrm{ad}_{J_0}\xi=\varepsilon m\xi\}, \tag{19}$$

$$\Psi_\bullet^{(r,\bullet)}=\{\xi\in\Psi\,|\,\mathrm{ad}_{K_0}\xi=\varepsilon r\xi\}. \tag{20}$$

We will show in the next section that if $2n,n+m,n+r\in\mathbb{Z}$ and $n\geqslant 0,|m|\leqslant n,|r|\leqslant n$ then $\dim(\Psi_n^{(\bullet,\bullet)}\cap\Psi_\bullet^{(r,\bullet)}\cap\Psi_\bullet^{(\bullet,m)})=1$, otherwise it has dimension 0.

So far we have considered Ψ simply as a vector space. On this vector space we define two products. These products are given by the projections ρ_R and $\rho_R^\star$.

The intention is that $R\geqslant 0$ represents the radius of the sphere or the Casimir. From (4) we want $R^2=K_0^2-\frac{1}{4}\varepsilon^2$. We take the positive square root of this equation by $K_0=\hat{R}=(R^2+\frac{1}{4}\varepsilon^2)^{1/2}$. Since K_0 is not in the center of the algebra we must be careful about quotienting by this equation. We therefore consider two projections:

$$\begin{aligned}&\rho_R:\mathscr{W}\mapsto\Psi,(\rho_R)^2=\rho_R,\\ &\rho_R^\star:\mathscr{W}\mapsto\Psi,(\rho_R^\star)^2=\rho_R^\star,\end{aligned} \tag{21}$$

defined by their kernels,

$$\begin{aligned}&\ker(\rho_R)=\{w(K_0-\hat{R})\,|\,w\in\mathscr{W}\},\\ &\ker(\rho_R^\star)=\{(K_0-\hat{R})w\,|\,w\in\mathscr{W}\},\end{aligned} \tag{22}$$

where

$$\hat{R}=\left(R^2+\tfrac{1}{4}\varepsilon^2\right)^{1/2},\quad R,\hat{R}\in\mathbb{R},\quad R\geqslant 0,\quad \hat{R}>0. \tag{23}$$

From now on we shall write ρ and $\rho^\star$ when there is only one possible R. $\hat{R}$ will always be given by (23). Note that $\hat{R}\to R$ as $\varepsilon\to 0$. These projections are surjective (onto Ψ) due to Lemma 3.

Lemma 3:

$$\mathscr{W}^{\cdot}=\Psi\oplus\ker(\rho)=\Psi\oplus\ker(\rho^{\star}). \tag{24}$$

Proof: Let $w\in\mathscr{W}^n$, $n\neq 0$; then wK_0 has a component in $\mathscr{W}^{n+1}$. Furthermore this component has nonzero trace, so $\mathrm{tr}(wK_0)$ has a component in $\mathscr{W}^n$. Now $\mathrm{tr}(w\hat{R})$ cannot have a component in $\mathscr{W}^n$ so $\mathrm{tr}(w(K_0-\hat{R}))\neq 0$ and

$$\Psi\cap\ker(\rho)=\{0\}. \tag{25}$$

We show $\mathscr{W}^{\cdot}=\Psi\oplus\ker(\rho_R)$ by dimensional argument. For this proof only let

$$\hat{\mathscr{W}}^n=\bigoplus_{m=0}^{2n}\mathscr{W}^{m/2},\quad \hat{\Psi}^n=\bigoplus_{m=0}^{2n}\Psi^{(\bullet,\bullet)}_{m/2},$$

while we also let

$$\hat{V}^n=\ker(\rho_R)\cap\hat{\mathscr{W}}^n=\{w(K_0-\hat{R})\,|\,w\in\hat{\mathscr{W}}^{n-1}\}.$$

So

$$\dim(\hat{V}^n)=\dim(\hat{\mathscr{W}}^{n-1})=\sum_{m=0}^{2n-2}\dim(\mathscr{W}^m).$$

Also,

$$\dim(\hat{\Psi}^n)=\sum_{m=0}^{2n}\dim\Psi^{(\bullet,\bullet)}_{m/2}=\sum_{m=0}^{2n}(\dim\mathscr{W}^{m/2}-\dim\mathscr{W}^{m/2-1}).$$

Thus

$$\dim(\hat{\Psi}^n)+\dim(\hat{V}^n)=\dim(\hat{\mathscr{W}}^n).$$

By intersecting (25) by $\hat{\mathscr{W}}^n$ we have $\hat{\Psi}^n\cap\hat{V}^n=\{0\}$. Combining these last two equations gives $\hat{\Psi}^n\oplus\hat{V}^n=\hat{\mathscr{W}}^n$. □

For a given ε,R we define the products, also called ρ and $\rho^{\star}$, by

$$\Psi\times\Psi\mapsto\Psi,\quad \xi,\zeta\mapsto\rho(\xi\zeta),\quad \xi,\zeta\mapsto\rho^{\star}(\xi\zeta). \tag{26}$$

Since $\ker(\rho)$ is not a two sided ideal, the product in Ψ is not associative and thus (Ψ,ρ) is not an algebra. For example $\rho(\rho(a_+a_-)a_-)=\rho(J_0a_-+(\hat{R}-\varepsilon/2)a_-)$ while $\rho(a_+\rho(a_-a_-))=\rho(a_+a_-a_-)=\rho(J_0a_-+(\hat{R}-\varepsilon)a_-)$.

However the restriction $(\Psi^{(0,\bullet)}_{\cdot},\rho)$ is an associative algebra. This algebra is identified with the analogue of the algebra of functions given in Ref. 1. Both Ψ and $\Psi^{(r,\bullet)}_{\cdot}$, for all r, are modules overs $\Psi^{(0,\bullet)}_{\cdot}$ in the same way the space of spinors and vector fields are modules over the space of functions. For reasons that will become apparent we will call the elements of $\Psi^{(1,\bullet)}_{\cdot}$ vectors, the elements of $\Psi^{(-1,\bullet)}_{\cdot}$ forms, and the elements of $\Psi^{(1/2,\bullet)}_{\cdot}$ and $\Psi^{(-1/2,\bullet)}_{\cdot}$ spinors. We also define a sesquilinear product and quadratic map:

$$\langle\bullet,\bullet\rangle:\Psi\times\Psi\mapsto\mathbb{C},\quad \langle\xi,\zeta\rangle=\pi_0(\rho_R(\xi^{\dagger}\zeta)), \tag{27}$$

$$\|\bullet\|^2:\Psi\mapsto\mathbb{R},\quad \|\zeta\|^2=\langle\zeta,\zeta\rangle. \tag{28}$$

As we will see this sesquilinear product is Hermitian but not positive definite, and that $\|\zeta\|^2$ may be positive negative or zero.

Lemma 4: For a given ε and R, the product ρ when restricted to the space $\Psi^{(0,\bullet)}_{\cdot}$ is closed and associative making $(\Psi^{(0,\bullet)}_{\cdot},\rho)$ an algebra. It is also equal to the product $\rho^{\star}$ on $\Psi^{(0,\bullet)}_{\cdot}$.

Both these algebras are equivalent to algebra $\mathscr{P}(\varepsilon,R)$ given by

$$\mathscr{P}(\varepsilon,R)=\{\text{ Free noncommuting algebra of polynomials in } (J_+,J_-,J_0)\}/\sim, \tag{29}$$

where $\sim$ *are the relations*

$$[J_0,J_+]\sim\varepsilon J_+,\ \ [J_0,J_-]\sim-\varepsilon J_-,\ \ [J_+,J_-]\sim 2\varepsilon J_0,\ \ J_0^2+\tfrac{1}{2}J_+J_-+\tfrac{1}{2}J_-J_+=R^2. \tag{30}$$

This algebra is described in detail in Ref. 1.

For general $w_1,w_2\in\Psi$ *we have* $\rho(w_1\rho(w_2))=\rho(w_1w_2)$ *whilst* $\rho(\rho(w_1)w_2)-\rho(w_1w_2)=O(\varepsilon)$. *In terms of the algebra* (Ψ,ρ) *this means that for* $\xi_1,\xi_2,\xi_3\in\Psi$,

$$\rho(\xi_1\rho(\xi_2\xi_3))=\rho(\xi_1\xi_2\xi_3),$$

$$\rho(\rho(\xi_1\xi_2)\xi_3)-\rho(\xi_1\xi_2\xi_3)=O(\varepsilon). \tag{31}$$

If $\xi\in\Psi_\bullet^{(r_1,\bullet)}$ *and* $\zeta\in\Psi_\bullet^{(r_2,\bullet)}$ *then* $\xi^\dagger\in\Psi_\bullet^{(-r_1,\bullet)}$ *and* $\rho(\xi\zeta)\in\Psi_\bullet^{(r_2+r_1,\bullet)}$.

Both Ψ *and* $\Psi_\bullet^{(r,\bullet)}$, *for all* r, *are right modules overs* $\Psi_\bullet^{(0,\bullet)}$, *since given* $\xi\in\Psi$ *and* $f,g\in\Psi_\bullet^{(0,\bullet)}$; *then*

$$\rho(\rho(\xi f)g)=\rho(\xi\rho(fg))=\rho(\xi fg). \tag{32}$$

The sesquilinear product $\langle\bullet,\bullet\rangle$ *is Hermitian:*

$$\langle\xi,\zeta\rangle=\overline{\langle\zeta,\xi\rangle},\ \ \forall\xi,\zeta\in\Psi. \tag{33}$$

Proof: If $f,g\in\Psi_\bullet^{(0,\bullet)}$ then it is clear that $\mathrm{ad}_{K_0}(fg)=0$, so $\rho(fg)=\rho^\star(fg)$. From lemma 2 we can write $fg=a_\pm^m b_\pm^{-m}p(J_0,K_0)$, thus $\rho(fg)=\Sigma_m a_\pm^m b_\pm^{-m}p(J_0,\hat{R})$. But $a_\pm^m b_\pm^{-m}=J_+^m$ for $m>0$ and $a_\pm^m b_\pm^{-m}=J_-^{-m}$ for $m<0$ so $\rho(fg)$ is a polynomial in $\{J_0,J_+,J_-\}$ and all such polynomials are elements in $\mathscr{P}(\varepsilon,R)$. From (4) these two algebras are equivalent.

Given $w_1,w_2\in\mathscr{W}$ then from (24) we may write uniquely $w_1=\rho(w_1)+w_1'(K_0-\hat{R})$. Thus $\rho(w_2w_1)=\rho(w_2\rho(w_1)+w_2w_1'(K_0-\hat{R}))=\rho(w_2\rho(w_1))$. While $\rho(\rho(w_1)w_2)-\rho(w_1w_2)=\rho(w_1'(K_0-\hat{R})w_2)=\rho(w_1'\mathrm{ad}_{K_0}(w-2))=O(\varepsilon)$.

If $f,g\in\Psi_\bullet^{(0,\bullet)}$ and $\xi\in\Psi$ then we write $\xi f=\rho(\xi f)+\xi_1(K_0-\hat{R})$ so, $\rho(\xi fg)-\rho(\rho(\xi f)g)=\rho(\xi_1(K_0-\hat{R})g)=0$ since $[K_0,g]=0$.

Given $\xi\in\Psi_\bullet^{(r_1,\bullet)}$ and $\zeta\in\Psi_\bullet^{(r_2,\bullet)}$ then $\langle\xi,\zeta\rangle\neq0$ only if $r_1=r_2$. In this case, $\rho^\star(\xi^\dagger\zeta)=\rho(\xi^\dagger\zeta)$, thus $\overline{\pi_0(\rho(\zeta^\dagger\xi))}=(\pi_0(\rho(\zeta^\dagger\xi)))^\dagger=\pi_0(\rho^\star(\xi^\dagger\zeta))=\pi_0(\rho(\xi^\dagger\zeta))$. □

IV. $\psi_n^{(r,m)}$: ORTHOGONAL BASIS FOR Ψ

Theorem 5: *There is a natural basis of* Ψ *given by*

$$\{\psi_n^{(r,m)}|2n,r+n,m+n\in\mathbb{Z},|r|\leqslant n,|m|\leqslant n\}, \tag{34}$$

where

$$\psi_n^{(r,m)}=\varepsilon^{m-n}\left(\frac{(n+m)!}{(2n)!(n-m)!}\right)^{1/2}(\mathrm{ad}_{J_-})^{n-m}(a_+^{n+r}b_-^{n-r}). \tag{35}$$

This definition is extended to the case $\varepsilon=0$ *by the use of (10). When written as formally traceless symmetric homogeneous polynomials,* $\psi_n^{(r,m)}$ *is independent of* R *and* ε. *The basis elements are orthogonal with respect to the sesquilinear form* $\langle\bullet,\bullet\rangle$:

$$\langle\psi_{n_1}^{(r_1,m_1)},\psi_{n_2}^{(r_2,m_2)}\rangle=\delta_{n_1,n_2}\delta_{r_1,r_2}\delta_{m_1,m_2}\|\psi_{n_1}^{(r_1,m_1)}\|^2, \tag{36}$$

where $\|\psi_n^{(r,m)}\|^2$ *may be positive negative or zero and is independent of* m.

For each n, *the set* $\{\psi_n^{(r,m)}|\forall r,m\}$ *forms an orthogonal basis for the set* Ψ^n. *These elements are eigenstates of the operators* ad_{J_0}, ad_{K_0}, *and* Δ:

$$\mathrm{ad}_{J_0}\psi_n^{(r,m)} = \varepsilon m \psi_n^{(r,m)}, \tag{37}$$

$$\Delta \psi_n^{(r,m)} = \varepsilon^2 n(n+1)\psi_n^{(r,m)}, \tag{38}$$

$$\mathrm{ad}_{K_0}\psi_n^{(r,m)} = \varepsilon r \psi_n^{(r,m)}, \tag{39}$$

where $\Delta = \mathrm{ad}_{J_0}\mathrm{ad}_{J_0} + \frac{1}{2}\mathrm{ad}_{J_+}\mathrm{ad}_{J_-} + \frac{1}{2}\mathrm{ad}_{J_-}\mathrm{ad}_{J_+}$, *so*

$$\Psi_n^{(\bullet,\bullet)} \cap \Psi_\bullet^{(r,\bullet)} \cap \Psi_\bullet^{(\bullet,m)} = \mathrm{span}\{\psi_n^{(r,m)}\}. \tag{40}$$

The operators ad_{J_+} *and* ad_{J_-} *act as Ladder operators within* $\Psi_n^{(\bullet,\bullet)} \cap \Psi_\bullet^{(r,\bullet)}$:

$$\mathrm{ad}_{J_+}\psi_n^{(r,m)} = \varepsilon (n-m)^{1/2}(n+m+1)^{1/2}\psi_n^{(r,m+1)}, \tag{41}$$

$$\mathrm{ad}_{J_-}\psi_n^{(r,m)} = \varepsilon (n+m)^{1/2}(n-m+1)^{1/2}\psi_n^{(r,m-1)}. \tag{42}$$

Proof: It is clear that $\psi_n^{(r,n)}$ is a formally traceless symmetric homogeneous polynomial of degree $2n$, from lemma 1 so are $\psi_n^{(r,m)}$ for all m. Equations (37)–(42) may be proved in the same way as the equivalent theorem in Ref. 1 theorem 2. Likewise for (36), where it is necessary to first show that $\pi_0(\rho(\mathrm{ad}_{J_+}x)) = \pi_0(\mathrm{ad}_{J_+}(\rho(w))) = 0$ for all $w \in \mathscr{W}$. This also comes from lemma 1. □

A formula for the conjugate of the basis element $(\psi_n^{(r,m)})^\dagger$ is given in corollary 10. As stated in lemma 4, the subspace $\Psi_\bullet^{(0,\bullet)}$ is equivalent to the algebra $\mathscr{P}(\varepsilon, R)$ given in Ref. 1. By comparing the two definitions for the basis vectors, one sees that $\psi_n^{(0,m)} = P_n^m/\alpha_n$. As a point of notation, the basis vectors here are left unnormalized, by contrast to the notation used in the above article. This avoids having to continually divide by a normalization constant, which here would depend on both r and n.

V. USEFUL FORMULAE FOR THE PRODUCT ρ ON Ψ

In this section we give a number of formulae which are useful for various calculations and the physical interpretations. All these formulae are extensions of the corresponding formulae for the basis of scalars on the noncommutative sphere,[1] but with added care taken due to the nonassociative nature of Ψ.

We start with a Hilbert space representation of $\mathscr{W}$ (subsection V A). As well as being used to calculate the value of the norm of the basis elements (subsection V B), it is also useful since it can be programmed into a symbolic mathematics program such as Maple, and hence used to calculate explicit formulae. This is followed by expressions to write $\psi_n^{(r,m)}$ as rectangular matrices (subsection V C) and in terms of Hahn polynomials (subsection V D). Finally in subsection V E we extend Ref. 1, Lemma 9 to give a formula which is useful for the definition of the exterior derivative.

A. A Hilbert space representation of $\mathscr{W}$

Let $\mathscr{H}$ be a Hilbert space with a basis given by $|k,j\rangle$ where $2k, j+k \in \mathbb{Z}$ and $k \geqslant 0$, $|j| \leqslant k$. The dual basis is represented by $\langle k,j|$ where $\langle k_1,j_1||k_2,j_2\rangle = \delta_{k_1k_2}\delta_{j_1j_2}$.

The action of $\mathscr{W}$ on $\mathscr{H}$ is given by

$$a_+|k,j\rangle = \varepsilon^{1/2}(k+j+1)^{1/2}|k+\tfrac{1}{2},j+\tfrac{1}{2}\rangle, \quad b_+|k,j\rangle = \varepsilon^{1/2}(k-j+1)^{1/2}|k+\tfrac{1}{2},j-\tfrac{1}{2}\rangle,$$

$$a_-|k,j\rangle = \varepsilon^{1/2}(k+j)^{1/2}|k-\tfrac{1}{2},j-\tfrac{1}{2}\rangle, \quad b_-|k,j\rangle = \varepsilon^{1/2}(k-j)^{1/2}|k-\tfrac{1}{2},j+\tfrac{1}{2}\rangle. \tag{43}$$

The dual action is given by $(\langle k_1,j_1|w)|k_2,j_2\rangle = \langle k_1,j_1|(w|k_2,j_2\rangle)$. The effect of the six elements J's and K's are given by

$$J_0|k,j\rangle=\varepsilon j|k,j\rangle,\quad K_0|k,j\rangle=\varepsilon(k+\tfrac{1}{2})|k,j\rangle,$$

$$J_+|k,j\rangle=\varepsilon(k-j)^{1/2}(k+j+1)^{1/2}|k,j+1\rangle,\quad K_+|k,j\rangle=\varepsilon(k-j+1)^{1/2}(k+j+1)^{1/2}|k+1,j\rangle,$$

$$J_-|k,j\rangle=\varepsilon(k+j)^{1/2}(k-j+1)^{1/2}|k,j-1\rangle,\quad K_-|k,j\rangle=\varepsilon(k-j)^{1/2}(k+j)^{1/2}|k-1,j\rangle. \tag{44}$$

One sees that the the effect of J_+,J_-,J_0 are exactly the same as the elements in the algebra in Ref. 1.

Lemma 6: Given $w\in\mathscr{W}$ which is independent of R and ε,

$$w|k,j\rangle=0\ \ \forall|k,j\rangle\in\mathscr{H}\ \ \Leftrightarrow\ \ \rho_R(w)=0\ \ \forall R\ \ \Leftrightarrow\ \ w=0. \tag{45}$$

If $\mathrm{ad}_{K_0}w=\varepsilon rw$ *and* $\mathrm{ad}_{J_0}w=\varepsilon mw$ *then*

$$w|k,j\rangle\in\mathrm{span}\{|k+r,j+m\rangle\}. \tag{46}$$

If $R^2=\varepsilon^2k(k+1)$ *then*

$$\rho_R(w)|k,j\rangle=w|k,j\rangle, \tag{47}$$

$$\pi_0(\rho_R(w))=\frac{1}{2k+1}\sum_{j=-k}^{k}\langle k,j|w|k,j\rangle. \tag{48}$$

Proof: Given $w\in\mathscr{W}$ such that $\mathrm{ad}_{K_0}w=\varepsilon rx$ and $\mathrm{ad}_{J_0}w=\varepsilon mx$ we use lemma 2 to write w in form (13). From (43), (46) is obvious.

If $\rho_R(w)=0,\forall R$ then it is clear that $w|k,j\rangle=0,\forall j,k$. To prove the reverse, $w\in\mathscr{W}$ can be written as a sum of elements $w=\Sigma_{rm}w_{rm}$ with w_{rm} in the eigenspace of ad_{K_0} and ad_{J_0}. From (46) each w_{rm} satisfies $w_{rm}|k,j\rangle=0$.

Using lemma 2 write

$$w_{rm}=a_\pm^{r+m}b_\pm^{r-m}p(K_0,J_0).$$

Then

$$0=w_{rm}|k,j\rangle=p\left(\varepsilon\left(k+\frac{1}{2}\right),\varepsilon j\right)a_\pm^{r+m}b_\pm^{r-m}|j,k\rangle,\quad\forall j,k.$$

Writing $\hat{p}(k,j)=p(\varepsilon(k+\frac{1}{2}),\varepsilon j)$ then $\hat{p}(k,j)=0$ for all j,k such that $2j,2k\in\mathbb{Z}$, $k\geqslant\max(0,-r)$ and $\max(-k,-k-m)\leqslant j\leqslant\min(k,k-m)$. Now writing $\hat{p}(k,j)=\Sigma_s p_s(j)k^s$ then $p_s(j)=0$ for above range of j. Chosing k greater than the degree of $\hat{p}$ implies $p_s\equiv0$. Thus $\hat{p}\equiv0$; thus $w=0$.

Finally, if $w\in\Psi$ write $w=\Sigma_{rm}w_{rm}$ as before. From (46) it is clear that the right hand side of (48) vanishes for all but the w_{00} term. Now w_{00} may be written as a polynomial in J_0. We can use the result of Ref. 1, Theorem 3. □

We note that the Hilbert space representation of $\psi_n^{(r,m)}$ is related to the Wigner operator,

$$\left\langle\begin{array}{ccc} & n+r & \\ 2n & & 0 \\ & n+m & \end{array}\right\rangle.$$

One can use this to find an expression for the product of two basis elements in terms of the Wigner $6j$ symbols (Ref. 12, Eqs. (3.341), (3.350), and Refs. 13,14.)

B. The value of $\|\psi_n^{(r,m)}\|^2$

Theorem 7: *The value of* $\|\psi_n^{(r,m)}\|^2$ *is given by*

$$\|\psi_n^{(r,m)}\|^2=\pi_0(a_-^{n+r}a_+^{n+r}b_+^{n-r}b_-^{n-r})=\frac{(n+r)!(n-r)!}{(2n+1)!}\prod_{s=1}^{n-r}(2\hat{R}-\varepsilon s)\prod_{s=1}^{n+r}(2\hat{R}+\varepsilon s). \quad (49)$$

The sign of this now depends on ε *and* R *and is given by*

$$\operatorname{sign}(\|\psi_n^{(r,m)}\|^2)=\begin{cases}1, & \text{if } 2\hat{R}/\varepsilon>n-r,\\ 0, & \text{if } 2\hat{R}/\varepsilon\in\mathbb{Z},1\leqslant 2\hat{R}/\varepsilon\leqslant n-r,\\ (-1)^{(n-r-\lfloor 2\hat{R}/\varepsilon\rfloor)}, & \text{if } 2\hat{R}/\varepsilon\notin\mathbb{Z},0<2\hat{R}/\varepsilon<n-r,\end{cases} \quad (50)$$

where $\lfloor 2R/\varepsilon\rfloor$ *is the largest integer less than or equal to* $2R/\varepsilon$.

Proof: This proof is similar to the proof of Ref. 1, Theorem 6. Consider $\hat{R}=\varepsilon(k+\frac{1}{2})$, we have

$$\pi_0(a_-^{n+r}a_+^{n+r}b_+^{n-r}b_-^{n-r})=\frac{1}{2k+1}\sum_{j=-k}^{k}\langle k,j|a_-^{n+r}a_+^{n+r}b_+^{n-r}b_-^{n-r}|k,j\rangle$$

$$=\frac{\varepsilon^{2k}}{2k+1}\sum_{j=-k}^{k}\frac{(k+j+r+n)!(k-j)!}{(k+j)!(k-j-n+r)!}.$$

Substituting $j=s-k$ and removing vanishing terms,

$$=\frac{\varepsilon^{2k}}{2k+1}\sum_{j=-k}^{2k-n+r}\frac{(s+r+n)!(2k-s)!}{s!(2k-n+r-s)!}$$

$$=\varepsilon^{2k}\frac{(r+n)!(2k)!}{(2k+1)(2k-n+r)!}\sum_{j=-k}^{2k-n+r}\frac{(r+n+1)_s(-2k+n-r)_s}{s!(-2k)_s}$$

$$=\varepsilon^{2k}\frac{(r+n)!(2k)!}{(2k+1)(2k-n+r)!}F(r+n+1,n-r-2k;-2k;1)$$

$$=\varepsilon^{2k}\frac{(r+n)!(2k)!}{(2k+1)(2k-n+r)!}\lim_{\kappa\to 0}F(r+n+1,n-r-2k;\kappa-2k;1)$$

$$=\varepsilon^{2k}\frac{(r+n)!(2k)!}{(2k+1)(2k-n+r)!}\lim_{\kappa\to 0}\frac{\Gamma(\kappa-2k)\Gamma(\kappa-2n-1)}{\Gamma(\kappa-2k-r-n-1)\Gamma(\kappa-n+r)}$$

$$=\varepsilon^{2k}\frac{(r+n)!(n-r)!(2k+1+r+n)!}{(2k+1)(2n+1)!(2k-n+r)!}.$$

Substituting $\hat{R}=\varepsilon(k+\frac{1}{2})$ and using lemma 6 gives us this result. □

C. Matrix representations for $\psi_n^{(r,m)}$

If $\hat{R}=\varepsilon(k+\frac{1}{2})$ then (as already noted) we can quotient the space of functions $\Psi_{\bullet}^{(0,\bullet)}$ to give the algebra of matrices $M_{2k+1}(\mathbb{C})$. This interpretation may be extended for each $\Psi_{\bullet}^{(r,\bullet)}$ but with caution to reflect its nonassociative nature. The elements of $\Psi_{\bullet}^{(r,\bullet)}$ now become rectangular matrices with $2k+2r+1$ rows and $2k+1$ columns; i.e., elements of $M_{(2k+2r+1)\times(2k+1)}(\mathbb{C})$.

Lemma 8: *Given* $\hat{R}=\varepsilon(k+\frac{1}{2})$ *then the mapping*

$$\varphi_k^r : \Psi_\bullet^{(r,\bullet)} \mapsto M_{(2k+2r+1)\times(2k+1)}(\mathbb{C}),$$

$$(\varphi_k^r(\xi))_{\mu\nu} = \langle k+r, \mu-k-r-1 | \xi | k, \nu-k-1 \rangle, \tag{51}$$

where $1 \leqslant \mu \leqslant 2k+2r+1$ *and* $1 \leqslant \nu \leqslant 2k+1$, *satisfies the following properties:*

$$\varphi_k^r(\rho(\xi f)) = \varphi_k^r(\xi)\varphi_k^0(f), \quad \forall \xi \in \Psi_\bullet^{(r,\bullet)}, \quad f \in \Psi_\bullet^{(0,\bullet)}, \tag{52}$$

$$(\varphi_k^r(\xi))^\dagger \varphi_k^r(\zeta) = \varphi_k^0(\rho(\xi^\dagger \zeta)), \tag{53}$$

$$\varphi_k^r(\psi_n^{(r,m)}) = 0, \quad \text{if } n \geqslant 2k+r+1. \tag{54}$$

Proof: Equation (52) follows from the use of (47) to remove the ρ, simple matrix multiplication, and observing that $\Sigma_{\mu=1}^{2k+1} |k, \mu-k-1\rangle\langle k, \mu-k-1|$ is the unit matrix in $M_{2k+1}(\mathbb{C})$. Likewise (53) follows from the same plus the recognition that, $\langle k, \mu-k-1 | \xi^\dagger | k+r, \nu-k-r-1 \rangle = ((\varphi_k^r(\xi))^\dagger)_{\mu\nu}$.

For $n \geqslant 2k+r+1$ then from theorem 7,

$$0 = \langle \psi^{(r,m)}{}_n, \psi_n^{(r,m)} \rangle = \pi_0(\rho(\psi_n^{(r,m)\dagger} \psi_n^{(r,m)}))$$

$$= \frac{1}{2k+1} \mathrm{TR}(\varphi_k^0(\rho(\xi^\dagger \xi))) = \frac{1}{2k+1} \sum_{\mu=0}^{2k+2r+1} \sum_{\nu=0}^{2k+1} |\varphi_k^r(\psi_n^{(r,m)})_{\mu\nu}|^2,$$

where TR represents the matrix trace. Hence (54). □

By counting dimensions we see that the $\ker(\varphi_k^r) = \mathrm{span}\{\psi_n^{(r,m)} | \forall m,n \text{ s.t. } n \geqslant 2k+r+1\}$. This is not an ideal in general but it does have the property: if $\xi \in \ker(\varphi_k^r)$ and $f \in \Psi_\bullet^{(0,\bullet)}$ then $\rho(\xi f) \in \ker(\varphi_k^r)$.

D. Writing $\psi_n^{(r,m)}$ in terms of Hahn polynomials

Lemma 9: *The highest degree term in the expansions of* $\psi_n^{(r,m)}$ *is given by*

$$\psi_n^{(r,m)} = (-1)^{n-\max(r,m)} \binom{2n}{n+m}^{1/2} a_\pm^{r+m} b_\pm^{r-m} J_0^{n-\max(|r|,|m|)} + \text{ Lower Degree Terms in } J_0, \tag{55}$$

where $a_\pm$ *and* $b_\pm$ *are given by (15).*

Proof: By repeated application of the formula

$$\mathrm{ad}_{J_-}(a_+^x b_+^{n+r-x} a_-^{n-r-v} b_-^v) = \varepsilon x a_+^{x-1} b_+^{n-r-x+1} a_-^{n-r-v} b_-^v - \varepsilon v a_+^x b_+^{n+r-x} a_-^{n-r-v+1} b_-^{v-1}, \tag{56}$$

we see that

$$(\mathrm{ad}_{J_-})^{n-m}(a_+^{n+r} b_+^{n-r})$$

$$= \varepsilon^{n-m} \sum_{s=\max(0,r-m)}^{\min(n-m,n+r)} \frac{(n-m)!(n+r)!(n-r)!(-1)^{n-m-s}}{s!(n-m-s)!(n+r-s)!(m-r+s)!} a_+^{n+r-s} b_+^s a_-^{n-m-s} b_-^{m-r+s}. \tag{57}$$

Using $a_+ a_- = J_0 + \text{ LDT}$ and $b_+ b_- = -J_0 + \text{ LDT}$, where LDT means ''Lower Degree Terms in J_0,'' we have

$$a_+^{n+r-s} b_+^s a_-^{n-m-s} b_-^{m-r+s} = a_\pm^{r+m} b_\pm^{r-m} J_0^{n+\min(r,-m)+\min(0,m-r)} (-1)^{s+\min(0,m-r)} + \text{ LDT}$$

$$= a_\pm^{r+m} b_\pm^{r-m} J_0^{n-\max(|r|,|m|)} (-1)^{s+\min(0,m-r)} + \text{ LDT},$$

where the last expression comes from considering each of the four cases below separately.

We have to consider separately the cases $m\geqslant r$ and $r\geqslant m$. We use the equations $(N+s)!=N!(N+1)_s$ and $N!=(N-s)!(-1)^s(-N)_s$ to convert our sum into a hypergeometric function. Since the coefficient is 1 we can use the summation formula to get an explicit form in terms of factorials.

For $m\geqslant r$ we have

$$
\begin{aligned}
&(\mathrm{ad}_{J_-})^{n-m}(a_+^{n+r}b_+^{n-r})\\
&\quad=\varepsilon^{n-m}a_\pm^{r+m}b_\pm^{r-m}J_0^{n-\max(|r|,|m|)}\frac{(-1)^{n-m}(n-r)!}{(m-r)!}\sum_{s=0}^{n+\min(-m,r)}\frac{(m-n)_s(-n-r)_s}{s!(m-r+1)_s}+\text{LDT}\\
&\quad=\varepsilon^{n-m}a_\pm^{r+m}b_\pm^{r-m}J_0^{n-\max(|r|,|m|)}\frac{(-1)^{n-m}(n-r)!}{(m-r)!}F(m-n,-n-r;m-r+1;1)+\text{LDT}\\
&\quad=\varepsilon^{n-m}a_\pm^{r+m}b_\pm^{r-m}J_0^{n-\max(|r|,|m|)}\frac{(-1)^{n-m}(2n)!}{(n+m)!}+\text{LDT},
\end{aligned}
$$

whilst for $m\leqslant r$,

$$
(\mathrm{ad}_{J_-})^{n-m}(a_+^{n+r}b_+^{n-r})=\varepsilon^{n-m}a_\pm^{r+m}b_\pm^{r-m}J_0^{n-\max(|r|,|m|)}\frac{(-1)^{n-r}(2n)!}{(n+m)!}+\text{LDT}.
$$

Using (35) gives this result. □

Corollary 10: The Hermitian conjugate is given by

$$
(\psi_n^{(r,m)})^\dagger=(-1)^{(r+m)}\psi_n^{(-r,-m)}. \tag{58}
$$

Proof: By putting $m=-n$ in above we have the exact formula

$$
\psi_n^{(r,-n)}=(-1)^{r+n}a_-^{n-r}b_+^{n+r}.
$$

Taking the conjugate of (35), using $(\mathrm{ad}_{J_-}(f))^\dagger=-\mathrm{ad}_{J_+}(f^\dagger)$, and substituting (59) gives (58). □

Theorem 11: *We can write the polynomials $\psi_n^{(r,m)}$ as*

$$
\psi_n^{(r,m)}=\begin{cases}
(-1)^{n-r}\varepsilon^{n-r}\dbinom{2n}{n+m}^{1/2}\dbinom{2n}{n+r}^{-1}a_+^{r+m}b_+^{r-m}h_{n-r}^{(r-m,r+m)}\left(\dfrac{(J_0+\hat R)}{\varepsilon}-\dfrac12,\dfrac{2\hat R}{\varepsilon}\right),\\
\qquad for\ -r\leqslant m\leqslant r,\\
(-1)^{n-m}\varepsilon^{n-m}\dbinom{2n}{n+m}^{1/2}a_+^{r+m}b_-^{m-r}h_{n-m}^{(m-r,r+m)}\left(\dfrac{(J_0+\hat R)}{\varepsilon}-\dfrac12,\dfrac{2\hat R}{\varepsilon}+r-m\right),\\
\qquad for\ -m\leqslant r\leqslant m,\\
(-1)^{n-r}\varepsilon^{n+m}\dbinom{2n}{n+m}^{1/2}a_-^{-r-m}b_+^{r-m}\\
\qquad\times h_{n+m}^{(r-m,-r-m)}\left(\dfrac{(J_0+\hat R)}{\varepsilon}-\dfrac12+r+m,\dfrac{2\hat R}{\varepsilon}+r+m\right),\quad for\ m\leqslant r\leqslant -m,\\
(-1)^{n-m}\varepsilon^{n+r}\dbinom{2n}{n+m}^{1/2}\dbinom{2n}{n+r}^{-1}a_-^{-r-m}b_-^{m-r}\\
\qquad\times h_{n+r}^{(m-r,-r-m)}\left(\dfrac{(J_0+\hat R)}{\varepsilon}-\dfrac12+r+m,\dfrac{2\hat R}{\varepsilon}+2r\right),\quad for\ r\leqslant m\leqslant -r
\end{cases}. \tag{59}
$$

Proof: It is clear we shall have to consider the four cases separately. For each case the proof is not too different from Ref. 1, Theorem 7. Let us here prove the theorem for the first case where $-r\leqslant m\leqslant r$, the other cases are similarly proved.

Given two basis elements $\psi_{n_1}^{(r_1,m_1)}$ and $\psi_{n_2}^{(r_2,m_2)}$ it is clear that $\langle \psi^{(r_1,m_1\rangle}{}_{n_1}, \psi_{n_2}^{(r_2,m_2)}\rangle = 0$ unless $m_1=m_2$ and $r_1=r_2$. Use lemma 2 so the $\psi_{n_1}^{(r,m)} = a_+^{r+m} b_+^{r-m} p(J_0,K_0)$. Thus $\psi_{n_1}^{(r,m)} = \rho(\psi_{n_1}^{(r,m)}) = a_+^{r+m} b_+^{r-m} p_{n_1}(J_0)$ for the polynomial $p_{n_1}(J_0)$ which we want to determine.

Let $R^2=\varepsilon^2 k(k^1)$ so $R'=\varepsilon(k+\frac{1}{2})$. Using (48) we have for $n_1\neq n_2$,

$$\begin{aligned} 0 &= \langle \psi^{(r,m\rangle}{}_{n_1}, \psi_{n_2}^{(r,m)}\rangle \\ &= \sum_{j=-k}^{k} \langle k,j| p_{n_1}(J_0) a_+^{r+m} b_+^{r-m} a_+^{r+m} b_+^{r-m} p_{n_2}(J_0) |k,j\rangle \\ &= \sum_{j=-k}^{k} p_{n_1}(\varepsilon j) p_{n_2}(\varepsilon j) \frac{(k+j+r+m)!(k-j+r-m)!}{(k+j)!(k-j)!}. \end{aligned}$$

Thus $p_n(\varepsilon j)$ are polynomials orthogonal with respect to a weight function. This weight function is precisely the same as that for the Hahn polynomials $h_{n'}^{(\alpha,\beta)}(x,N)$ where $\alpha=r-m$, $\beta=r+m$, $x=j+k$ and $N=2k+1$. We see that $n'=n-r$ since $h_{n'}^{(\alpha,\beta)}(x,N)$ is a polynomial of order n' in x, and from (55), $p_n(J_0)$ is a polynomial of order $n-r$. Comparing the coefficient for the J_0^{n-r} term we have from Ref. 14,

$$h_{n-r}^{(r-m,r+m)}\left(\frac{(J_0+\hat{R})}{\varepsilon}-\frac{1}{2},\frac{2\hat{R}}{\varepsilon}\right) = \frac{(2n)!}{(n-r)!(r+n)!}\frac{J_0^{n-r}}{\varepsilon^{n-r}} + \text{LDT}.$$

Hence the result for the range $-r\leqslant m\leqslant r$. Since this is true for all k then using lemma 6 gives result for all ε,R. The other ranges are proved similarly. □

In the limit $\varepsilon=0$ the Hahn polynomials become Jacobi Polynomials ($P_n^{(\alpha,\beta)}(z)$). Thus we have

$$\begin{aligned} \psi_n^{(r,m)}|_{\varepsilon=0} &= (-1)^{n-\max(r,m)}(2R)^{n-\max(|r|,|m|)} \\ &\quad\times\binom{2n}{n+m}^{1/2}\binom{2n}{n-\max(|r|,|m|)}^{-1} a_\pm^{r+m} b_\pm^{r-m} P_{n-\max(|r|,|m|)}^{(|r-m|,|r+m|)}\left(\frac{J_0}{R}\right). \end{aligned} \tag{60}$$

E. Another useful theorem

Here we extend Ref. 1, Lemma 9 to give a formula which is useful for the definition of the exterior derivative.

Theorem 12: *For r_1,r_2,n let*

$$\omega_n^{r_1r_2}:\Psi\mapsto\Psi,$$

$$\omega_n^{r_1r_2}(\xi) = \sum_{m=-n}^{n} \rho(\psi_n^{(r_2,m)\dagger}\xi\psi_n^{(r_1,m)}). \tag{61}$$

Then $\omega_n^{r_1r_2}$ commutes with ad_{J_0}, ad_{J_+}, ad_{J_-}, Δ *and satisfies*

$$\mathrm{ad}_{K_0}\omega_n^{r_1r_2}(\xi) = \varepsilon(r_1-r_2)\omega_n^{r_1r_2}(\xi) + \omega_n^{r_1r_2}(\mathrm{ad}_{K_0}\xi). \tag{62}$$

Its effect on the basis elements is given by

$$\omega_{n_1}^{r_1r_2}(\psi_n^{(r,m)}) = \omega_{n_1n}^{r_1r_2r}\psi_n^{(r+r_1-r_2,m)}, \tag{63}$$

where $\omega_{n_1 n}^{r_1 r_2 r} \in \mathbb{C}$. *For* $\xi = 1$,

$$\omega_n^{r_1 r_2}(1) = \delta_{r_1 r_2}(2n+1)\|\psi_n^{(r_1,m)}\|^2. \tag{64}$$

Proof: The effect of ad_{J_0}, ad_{J_+}, ad_{J_-}, and ad_{K_0} comes from direct calculation using (37), (41), (42), and (39), respectively. The effect of Δ comes from ad_{J_+} and ad_{J_-}. From this, it is clear that $\omega_{n_1}^{r_1 r_2}(\psi_n^{(r,m)})$ must be proportional to $\psi_n^{(r+r_1-r_2,m)}$ and that the proportionality constant must be independent of m since $\omega_{n_1}^{r_1 r_2}$ commutes with ad_{J_+}.

Since $\omega_n^{r_1 r_2}(1)$ is proportional to $\psi_0^{(r_1-r_2,0)}$ it is clear that it vanishes unless $r_1 = r_2$. In this case taking π_0 in the expression gives this result. □

VI. PHYSICAL INTERPRETATION

There are two ways of interpreting the algebra (Ψ,ρ) and the basis elements $\psi_n^{(r,m)}$. In subsection VI A we show that $\psi_n^{(r,m)}$ may be thought of as the deformed rotation matrices for $SU(2)$, whilst in subsections VI B–VI D we use them to construct the analogue of scalar, vector, and spinor fields.

A. Deformed rotation matrices

In the limit $\varepsilon = 0$ the elements of Ψ may be regarded as complex functions over S^3 or $SU(2)$. The product ρ becomes pointwise multiplication. The basis elements $\psi_n^{(r,m)}$ are proportional to the rotation matrix D_{mr}^n. Thus we may interpret $\psi_n^{(r,m)}$ as the deformed (or noncommutative) rotation matrices.

Just as in standard quantum mechanics the noncommutativity in the quantum domain becomes the Poisson Bracket in the commutative case, we can define a bracket (68) as the limit of the commutator. However, since for $\varepsilon \neq 0$ ρ is not associative, this bracket is neither a derivation nor satisfies the Jacobi identity. Therefore, it is not a Poisson Bracket. This is reflected in the existence of the number operator δ_N in (69). It would be nice if one could find a physical interpretation for this bracket.

Lemma 13: In the limit $\varepsilon = 0$ *the noncommutative, nonassociative algebra* (Ψ,ρ) *becomes the commutative associative algebra* $C_0(SU(2))$. *This is the algebra of complex valued polynomial functions on SU(2) with pointwise multiplication.*

If we use the Euler angles (α,β,γ) *to parameterise SU(2) then the generators of* Ψ *may be written as*

$$\begin{aligned} a_+ &= \sqrt{2R}\,\cos(\beta/2)e^{-i/2(\alpha+\gamma)}, \quad b_+ = -i\sqrt{2R}\,\sin(\beta/2)e^{i/2(\alpha-\gamma)}, \\ a_- &= \sqrt{2R}\,\cos(\beta/2)e^{i/2(\alpha+\gamma)}, \quad b_- = i\sqrt{2R}\,\sin(\beta/2)e^{-i/2(\alpha-\gamma)}. \end{aligned} \tag{65}$$

The basis elements of Ψ *are given by the rotation matrices*

$$\psi_n^{(r,m)}\big|_{\varepsilon=0} = (i)^{m-r}(-1)^{n-r}(2R)^n \binom{2n}{r}^{-1/2} D_{mr}^n(\alpha,\beta,\gamma), \tag{66}$$

where $D_{mr}^n(\alpha,\beta,\gamma)$ *uses the notation in Ref. 13. The bilinear form on* Ψ *becomes a true inner product and is given by*

$$\langle \xi,\zeta\rangle = \frac{1}{8\pi^2}\int_0^{2\pi} d\alpha \int_0^{2\pi} d\gamma \int_0^{\pi} d\beta \sin(\beta)\,\overline{\xi}\zeta. \tag{67}$$

There is a bracket defined by

$$\{\bullet,\bullet\}: C_0(SU(2))\times C_0(SU(2)) \mapsto C_0(SU(2)), \quad \{\xi,\zeta\} = \lim_{\varepsilon\to 0} \frac{1}{i\varepsilon}\rho([\xi,\zeta]), \tag{68}$$

which is given by

$$\{\xi,\zeta\}=\frac{1}{R\sin\beta}\left(\frac{\partial\xi}{\partial\alpha}\frac{\partial\zeta}{\partial\beta}-\frac{\partial\xi}{\partial\beta}\frac{\partial\zeta}{\partial\alpha}\right)+\left(\delta_N(\xi)\frac{\partial\zeta}{\partial\gamma}-\frac{\partial\xi}{\partial\gamma}\delta_N(\zeta)\right)+\frac{\cot\beta}{R}\left(\frac{\partial\xi}{\partial\beta}\frac{\partial\zeta}{\partial\gamma}-\frac{\partial\xi}{\partial\gamma}\frac{\partial\zeta}{\partial\beta}\right), \tag{69}$$

where δ_N is the number operator,

$$\delta_N:C_0(SU(2))\mapsto C_0(SU(2)),\quad \delta_N(\psi_n^{(r,m)})=n\psi_n^{(r,m)}. \tag{70}$$

Proof: By setting $\varepsilon=0$ it is clear that the algebra is commutative and associative. The condition that $K_0=R$ implies that the elements of Ψ are functions on S^3, or alternatively $SU(2)$. Assuming (65) then (66) follows from (60). The inner product (67) follows from the orthogonality of the rotation matrix elements.

In the limit $\varepsilon=0$ the algebra $\mathscr{W}(0)$ is the algebra $C_0(\mathbb{R}^4)$ of complex valued polynomial functions over $\mathbb{R}^4$. The Poisson Bracket corresponding to the limit of the commutator is given by

$$\{f,g\}_{\mathscr{W}}=\frac{\partial f}{\partial a_+}\frac{\partial g}{\partial a_-}-\frac{\partial f}{\partial a_-}\frac{\partial g}{\partial a_+}+\frac{\partial f}{\partial b_+}\frac{\partial g}{\partial b_-}-\frac{\partial f}{\partial b_-}\frac{\partial g}{\partial b_+}.$$

Writing this with respect to the coordinates (R,α,β,γ) gives (69) but with δ_N replaced by $\partial/\partial R$. However, for every polynomial $w(a_+,a_-,b_+,b_-)$ not explicitly dependent on R, $\partial w/\partial R=\delta_N(w)$. Thus restriction of $\{\bullet,\bullet\}_{\mathscr{W}}$ to $C_0(SU(2))$ is valid. □

B. The exterior derivative and 1-forms

In standard geometry there are many equivalent definitions of a vector field. However, not all of them can be extended to the noncommutative case at the same time. If one takes the definition of a vector field, that it must satisfy the Leibniz formula, and extends this definition to noncommutative geometry, then, at least for the matrix case, the vector fields are given by $X=\mathrm{ad}_f$ for some $f\in\Psi_\bullet^{(0,\bullet)}$. This can then be used to give a definition of the exterior derivative and the exterior algebra.[2]

Here we give another definition of covectors by use of the analogue of the following definition of the exterior derivative of scalar fields:

$$df=\sum_i\frac{\partial f}{\partial x^i}dx^i. \tag{71}$$

If we consider the noncommutative sphere as a three dimensional manifold with normalized basis coordinates $x^m=(2\hat{R}+\varepsilon)^{-1/2}\psi_1^{(0,m)}$ for $m\in\{-1,0,1\}$ we can define the basis 1-forms as

$$dx^m=(2\hat{R}+\varepsilon)^{-1/2}\psi_1^{(-1,m)},\ \text{for } m\in\{-1,0,1\}. \tag{72}$$

This enables us to define the *exterior derivative* on the space of functions as

$$d:\Psi_\bullet^{(0,\bullet)}\mapsto\Psi_\bullet^{(-1,\bullet)},\quad df=\varepsilon^{-1}(2\hat{R}+\varepsilon)^{-1}\omega_1^{01}(f),\quad \forall f\in\Psi_\bullet^{(0,\bullet)}. \tag{73}$$

Since $\omega_1^{01}(1)=0$, we see that

$$df=\frac{1}{\varepsilon}\sum_{m=-1}^{1}(-1)^{m+1}dx^{-m}\mathrm{ad}_{x^m}f, \tag{74}$$

which is analogous to (71). This obeys the following.

Lemma 14: The two definition of the 1-forms dx^m given by (72) and (73) agree. For the basis elements which are functions,

$$d\psi_n^{(0,m)}=2n(\hat{R}+\varepsilon n/2)(\hat{R}+\varepsilon)^{-1}\psi_n^{(-1,m)}, \tag{75}$$

and in the limit as $\varepsilon\to 0$,

$$d(fg)=d(f)g+f(dg)+O(\varepsilon). \tag{76}$$

Proof: From the definition we see that

$$J_+=\psi_1^{(0,1)}, \qquad -\sqrt{2}J_0=\psi_1^{(0,0)}, \qquad -J_-=\psi_1^{(0,-1)},$$
$$b_-^2=\psi_1^{(-1,1)}, \quad \sqrt{2}a_-b_-=\psi_1^{(-1,0)}, \quad a_-^2=\psi_1^{(-1,-1)}.$$

By manipulations of J_+, J_- we can show that

$$[J_-, J_+^n]=-2\varepsilon J_+^{n-1}(nz+\varepsilon n(n-1)/2).$$

From theorem 12 we know that $\omega_1^{01}(\psi_n^{(-1,m)})$ is proportional to $\psi_n^{(0,m)}$ and its coefficient is independent of m. Therefore taking $m=n\geqslant 1$ we have

$$\omega_1^{01}(\psi_n^{(0,n)})=a_-^2\,\mathrm{ad}_{J_+}(J_+^n)+2a_-b_-\,\mathrm{ad}_{J_0}(J_+^n)+b_-^2\,\mathrm{ad}_{J_-}(J_+^n)$$
$$=2n\varepsilon a_-b_-J_+^n-2\varepsilon b_-^2J_+^{n-1}(nz+\varepsilon n(n-1)/2).$$

Now $b_-^2J_+^{n-1}=a_+^{n-1}b_-^{n+1}=\psi_n^{(-1,n)}$ while

$$a_-b_-a_+^nb_-^n=(J_0+K_0+\varepsilon/2)a_+^{n-1}b_-^{n+1}$$
$$=a_+^{n-1}b_-^{n+1}(J_0+K_0+n\varepsilon-\varepsilon/2)).$$

Hence (75). We show that the two definitions of dx^m are equivalent by substituting $n=1$ into this formula. Finally we note that

$$\sum_{m=-1}^{1}(\psi_1^{(1,m)\dagger}fg\psi_1^{(0,m)}-\psi_1^{(1,m)\dagger}f\psi_1^{(0,m)}g-f\psi_1^{(1,m)\dagger}g\psi_1^{(0,m)})=[\psi_1^{(1,m)\dagger},f][g,\psi_1^{(0,m)}]=O(\varepsilon^2).$$

□

C. Vector fields

We can now define the vectors fields as the set $\Psi_\bullet^{(1,\bullet)}$. The action of a vector on a scalar is given by

$$X(f)=\rho((df)X). \tag{77}$$

By the same reasoning as (76), this is also only a derivation in the limit

$$X(fg)=X(f)g+fX(g)+O(\varepsilon). \tag{78}$$

However if we define X_m to be the conjugate of dx^m, $X_m=(dx^m)^\dagger=(-1)^{m+1}(2\hat{R}+\varepsilon)^{-1/2}\psi_1^{(1,-m)}$ then from $\omega_1^{1,0}(1)=\omega_1^{0,-1}(1)=0$ we have

$$\sum_{m=-1}^{1}x^mX_m=\sum_{m=-1}^{1}X_mx^m=0, \tag{79}$$

identically, for all ε, and not just in the limit $\varepsilon=0$.

We can now define a *metric*,

$$g:\Psi_\bullet^{(1,\bullet)}\times\Psi_\bullet^{(1,\bullet)}\mapsto\Psi_\bullet^{(0,\bullet)}, \quad g(X,Y)=\rho(X^\dagger Y). \tag{80}$$

This can be extended for all spaces $\Psi_\bullet^{(r,\bullet)}$.

D. Spinor fields

It is natural now to define the set $\Psi_{\bullet}^{(1/2,\bullet)}$ as the space of spinor fields and its dual $\Psi_{\bullet}^{(-1/2,\bullet)}$. This means that a vector is the product of two spinors. We may also consider the space $\Psi_n^{(r,\bullet)}=\Psi_n^{(\bullet,\bullet)}\cap\Psi_{\bullet}^{(r,\bullet)}$ as a $2n+1$ dimensional representation of $su(2)$ given by the adjoint representation, $\mathrm{ad}_u:\Psi_n^{(r,\bullet)}\mapsto\Psi_n^{(r,\bullet)}$ for any nonzero element $u\in su(2)$. The $2n+1$ eigenvalues of this mapping are given by the set $\{m\|u|/2|m+n\in\mathbb{Z},|m|\leqslant n\}$, where $\|u\|$ corresponds to the Killing form. Thus

$$e^{\pi iu/\|u\|}\psi_n^{(r,m)}e^{-\pi iu/\|u\|}=(-1)^{2n}\psi_n^{(r,m)}.$$

Hence rotation by 2π does not change the sign of $\psi_n^{(r,m)}$ if n is an integer, i.e., for scalars, vectors, and other ''Bosons.'' While $\psi_n^{(r,m)}$ changes sign under rotation of 2π for spinors tensor fields and other ''Fermions.''

An alternative way to define spinor fields is as the set

$$\mathscr{S}=\left\{\begin{pmatrix}a_+\\b_+\end{pmatrix}f_1+\begin{pmatrix}a_-\\b_-\end{pmatrix}f_2\,\middle|\,f_1,f_2\in\Psi_{\bullet}^{(0,\bullet)}\right\}. \tag{81}$$

This is usually regarded as the space of spinors. It is decomposed into $\mathscr{S}=\mathscr{S}_+\oplus\mathscr{S}_-$ corresponding to the eigenspaces of ad_{K_0} which is now regarded as the Chirality operator. This interpretation is now equivalent to the one proposed by Grosse *et al.*,[4,9,10] who go on to define and solve the Dirac equation. We note that $\mathscr{S}$ is not equivalent to simply two copies of $\Psi_{\bullet}^{(-1/2,\bullet)}\oplus\Psi_{\bullet}^{(1/2,\bullet)}$, but a proper subset. This is because the element $\binom{a_+}{0}\notin\mathscr{S}$. In the commutative limit $\varepsilon=0$ we can replace $a_+=a_-=R^{1/2}\cos(\theta/2)$ and $b_+=\overline{b_-}=R^{1/2}\sin(\theta/2)e^{-i\phi}$ to obtain the standard spinors on a sphere of radius R and coordinates $(0\leqslant\theta\leqslant\pi,0\leqslant\phi\leqslant2\pi)$.

VII. PROBLEMS AND OUTLOOK

There are many problems with this interpretation of our system. Here is a list of cases where this noncommutative geometry is different from the commutative geometry even in the limit $\varepsilon=0$.

(1) The space $\Psi_{\bullet}^{(-1,\bullet)}$ is the image of $\Psi_{\bullet}^{(0,\bullet)}$ under d. This means that all 1-forms are closed.

(2) The definition (73) can be extended as a map $d:\Psi_{\bullet}^{(r,\bullet)}\mapsto\Psi_{\bullet}^{(r-1,\bullet)}$ for all r. However this map does not satisfy $d^2=0$ even in the limit. As a result no exterior calculus is defined here.

(3) The set of vectors X_i are not dual, in the usual sense, to the set of 1-forms dx^j, or alternatively the vectors X_i are not orthogonal with respect to the metric g. This is because for $i\neq j$,

$$g(X_i,X_j)=\rho(dx^iX_j)=X_j(x^i)\neq0,$$

for all ε even when $\varepsilon=0$. However we do have $\pi_0(g(X_i,X_j))=\delta_{ij}$.

Some of these problems, together with the nonderivative nature of d may be solved by redefining the space of covectors. For example it may be similar to (81).

As stated at the end of subsection VI A the bracket (68) is in need of an interpretation.

Since ρ is not associative it would be nice to know what rules (if any) it obeys. Note, for example, even $(\rho(\xi\xi)\xi)\neq\rho(\xi\xi\xi)$ so ξ^n is not well defined.

In the limit $\varepsilon=0$ there is a formula for the combination of two matrix rotation entries D^n_{rm}. In the general case we can use the Wigner–Eckart theorem to give

$$\psi_{n_1}^{(r_1,m_1)}\psi_{n_2}^{(r_2,m_2)}=\sum_{n=|n_1-n_2|}^{n_1+n_2}C^{n_1,n_2,n}_{m_1,m_2,m_1+m_2}R^{n_1,n_2,n}_{r_1,r_2,r_1+r_2}\psi_n^{(r_1+r_2,m_1+m_2)},$$

where $C^{n_1,n_2,n}_{m_1,m_2,m_1+m_2}$ is the Wigner or Clebsch–Gordan coefficient and $R^{n_1,n_2,n}_{r_1,r_2,r_1+r_2}$ the corresponding reduced matrix element. It has since been shown[12] that $R^{n_1,n_2,n}_{r_1,r_2,r_1+r_2}$ may be written in terms of Wigner $6j$ symbols.

Since the space of spinor and vector fields naturally form modules over the space of scalar fields one can use this as a starting point for algebraic connections.[15]

As stated $\mathscr{W}$ contains a representation of $su(1,1)$. If one quotiented $\mathscr{W}$ by the ideals created by $J_0 \sim \hat{R}$, one would gain a very similar structure which could be interpreted as the basis of scalar, spinor, and vector fields on the noncommutative analogue of either the two dimensional deSitter space or the Hyperbolic plane.

The Jordan–Schwinger representation can be used for any Lie algebra. Hence one should be able to construct spinor fields on any symmetric manifold. The combination of these two steps would give one a description of scalar, spinor, and vector fields on more exotic symmetric spaces, such as the Einstein–deSitter Universe.

ACKNOWLEDGMENTS

The author would like to thank the Royal Society of London for a European Junior Fellowship which enabled him to study in Paris. The author would also like to thank Peter Prešnajder, John Madore, Luiz Saeger, and Jihad Mourad for useful discussions which aided this work and, especially, Richard Kerner and the Laboratoire de Gravitation et Cosmologie Relativistes, Paris VI for their hospitality.

[1] J. Gratus, ''A natural basis of states for the noncommutative sphere and its Moyal bracket,'' J. Math. Phys. **38**, 4283–4300 (1997).
[2] J. Madore, *An Introduction to Noncommutative Differential Geometry and its Physical Applications* (Cambridge University Press, Cambridge, 1995).
[3] U. Carow-Watamura and S. Watamura, ''Chirality and Dirac operators on the noncommutative sphere,'' Commun. Math. Phys. **183**, 365–382 (1997).
[4] H. Grosse, C. Klimčik, and P. Prešnajder, ''Topologically nontrivial field configurations in noncommutative geometry,'' Commun. Math. Phys. **178**, 507–526 (1996).
[5] J. Madore, ''*The commutative limit of a matrix geometry,*'' J. Math. Phys. **32**, 332–335 (1991).
[6] H. Grosse, C. Klimčik, and P. Prešnajder, ''*Finite quantum field theory in noncommutative geometry,*'' Int. J. Theor. Phys. **35**, 231–244 (1996).
[7] J. Gratus, ''Noncommutative differential geometry, and the matrix representations of generalized algebras,'' J. Geometry Phys., to be published, q-alq/9703034.
[8] H. Grosse and J. Madore, ''*A noncommutative version of the Schwinger model,*'' Phys. Lett. B **283**, 218–222 (1992).
[9] H. Grosse and P. Prešnajder, ''*The Dirac operator on the fuzzy sphere,*'' Lett. Math. Phys. **33**, 171–181 (1995).
[10] H. Grosse, C. Klimčik, and P. Prešnajder, ''*Field theory on a supersymmetric lattice,*'' Commun. Math. Phys. **185**, 155–175 (1997).
[11] M. Kibler, ''*On the SU_2 unit tensor,*'' J. Math. Phys. **21**, 422–439 (1980).
[12] L. C. Biedenharn and J. D. Louck, ''Angular momentum in Quantum Physics: Theory and Application,'' in the *Encyclopedia of Mathematics and its Applications* (Addison-Wesley, New York, 1981), Vol. 8.
[13] J. Gratus, ''*Wick rotations; The noncommutative hyperboloids and other surfaces of rotations,*'' submitted to Lett. Math. Phys.
[14] A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of a Discrete Variable,* (Springer-Verlag, Berlin, 1991).
[15] R. Coquereaux, ''Algebraic connections on parallel universes,'' Int. J. Mod. Phys. A **10**, 89–98 (1995).

Extremal eigenvalue gaps for the Schrödinger operator with Dirichlet boundary conditions

Samir Karaa
Laboratoire des Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, France

(Received 12 May 1997; accepted for publication 9 July 1997)

We consider the problems of minimizing and maximizing the gap between the two lowest Dirichlet eigenvalues of the Schrödinger operator $-\Delta+V(x)$ on a bounded domain in R^n, when the potential V is subjected to a p-norm constraint. We give characterization theorems for extremizing potentials. We prove in particular that a second eigenvalue corresponding to a minimizing potential is always single. © *1998 American Institute of Physics.* [S0022-2488(98)03903-6]

I. INTRODUCTION

Let Ω be an open bounded domain in R^n with smooth boundary. We consider the Schrödinger operator $L=-\Delta+V(x)$ acting on $L^2(\Omega)$, with zero Dirichlet boundary conditions. The real-valued function $V(x)$, called potential, is measurable and non-negative. As is well known, the eigenvalues of the operator L can be interpreted as the energy levels, in atomic units, of a particle traveling under the external force field of a potential $V(x)$ in R^n equal to $V(x)$ in the domain Ω and imagined as $+\infty$ outside Ω, and the corresponding eigenfunctions are wave functions of the Schrödinger equation

$$-\Delta y+V(x)y=\lambda y. \tag{1.1}$$

Moreover, for reasonable potentials V, the spectrum of L is discrete and consists of non-negative eigenvalues $\{\lambda_k\}$ which can be arranged in a nondecreasing order as follows:

$$0<\lambda_1<\lambda_2\leqslant\lambda_3\leqslant\cdots\leqslant\lambda_m\leqslant\cdots .$$

There are both physical and mathematical interests in studying the problem of minimizing and maximizing the functional $\Gamma(V)$ defined as

$$\Gamma(V)=\lambda_2(V)-\lambda_1(V),$$

when the non-negative potential $V(x)$ is subjected to some kind of constraints. The motivation of our work comes especially from quantum mechanical considerations. See Refs. 1 and 2 for concrete examples. The functional Γ will be called the *fundamental gap* of the operator L.

Bounds for the fundamental gap in terms of the geometry of Ω (V is assumed identically zero) have been the subject of many recent works (see, for example, Refs. 3–6 and the references quoted there). The most important result established recently is due to Benguria and Ashbaugh who proved that the fundamental gap achieves its maximum value at a ball having the same measure that Ω (see Refs. 3 and 4). In Ref. 6, lower and upper bounds for $\Gamma(V)$ have been obtained in terms of the geometrical invariants of Ω and the given potential V, with some assumptions imposed on both V and Ω. If d is the diameter of the domain Ω and D is the diameter of the largest inscribed ball in Ω, then with the assumption that V and Ω are convex, we have

$$C_{1.}d^{-2}\leqslant\Gamma(V)\leqslant C_{2.}D^{-2}+C_3,$$

where C_1, C_2, and C_3 are positive constants independent of Ω, whereas Kirsch and Simon[5–7] found that the fundamental gap can be estimated from below by a polynomial in d times exp

 © 1998 American Institute of Physics

$(-\text{const. } d)$. Another problem has been studied in Ref. 2. The authors considered the problem of minimizing and maximizing the fundamental gap $\Gamma(V)$ when the potential $V(x)$ is constrained to the set

$$S_{\alpha,M}=\{V:\text{supp } V\subset\Omega,\|V\|_{\alpha}\leqslant M\},$$

M is a fixed positive number, and α is assumed greater than $n/2$ when $n\geqslant 2$. In our work below we extend the analysis of Ref. 2 to the case of all $\alpha\geqslant 1$ at the price of imposing a second constraint, viz., a pointwise bound. The potential V is then allowed to vary over the set

$$K_{\alpha}=\left\{V\in L^{\infty}(\Omega),0\leqslant V(x)\leqslant H,\int_{\Omega}V^{\alpha}(x)=M\right\},$$

where H and M are positive numbers satisfying $M<H^{\alpha}|\Omega|$ and $\alpha\geqslant 1$.

It is known that for every potential in K_{α}, the operator L is strictly elliptic and self-adjoint with the quadratic-form domain $W_0^{1,2}(\Omega)$, its spectrum is discrete, and the first eigenvalue λ_1 is nondegenerate with a positive eigenfunction.[8,9] We begin by a discussion of the existence of optimal potentials and we give some regularity results. The next section deals with some characterizations of the optimal potentials.

Remark: Many authors have studied the problem of maximizing the first eigenvalue of some classes of second-order elliptic operators. See, for example, Refs. 10–13. In view of the simplicity of the first eigenvalue of the elliptic operators considered, the preoccupations of these papers are different from those presented in our work.

II. EXISTENCE

We first observe that Γ is uniformly bounded in K_{α} for any $\alpha\geqslant 1$. This follows from the fact that λ_1 and λ_2 are both bounded uniformly there because of the min–max principle and the operator inequalities

$$-\Delta\leqslant L\leqslant -\Delta+H.$$

Theorem 2.1: *Let Ω be a fixed smooth open domain in R^n and let the operator L be as above. Then the functional $\Gamma(V)$ attains its minimum and its maximum on K_{α} for any $\alpha\geqslant 1$.*

The proof is close to that given in Ref. 2. It is based on straightforward compactness arguments. We only note that the proof needs the following result:

$$\Delta y\in L^2(\Omega)\quad\text{and}\quad y\in W_0^{1,2}(\Omega)\Rightarrow y\in W^{2,2}(\Omega),$$

which is true if $\partial\Omega$ is C^2 (see Ref. 8, Chap. 8).

Proposition 2.1: Let V be in K_{α}, $\alpha\geqslant 1$. If a function $y\in W_0^{1,2}(\Omega)$ satisfies $Lu=\lambda u$ for some $\lambda\in\mathbb{R}$, then $y\in W_{\text{loc}}^{2,2}(\Omega)\cap C^1(\Omega)$. If, moreover, $\partial\Omega$ is of class C^2, then

$$y\in W^{2,2}(\Omega)\cap C_B^1(\Omega)\cap C^0(\overline{\Omega}),$$

where

$$C_B^1(\Omega)=\{u\in C^1(\Omega),D_iu\in L^{\infty}(\Omega),\ \ i=1,...,n\}.$$

Proof: Since $y\in L^2(\Omega)$, Eq. (1.1) shows that Δy is also in $L^2(\Omega)$ and hence $y\in W_{\text{loc}}^{2,2}(\Omega)$. Consequently, $y\in C^1(\Omega)$ if $n\leqslant 4$ and $y\in L_{\text{loc}}^{s}(\Omega)$ with $s=2n/(n-4)$ if $n>4$. For $n>4$, we can continue this reasoning so that $y\in W_{\text{loc}}^{2,r}(\Omega)$ for a high enough index r satisfying $2r>n$, which shows that $y\in C^1(\Omega)$. If $\partial\Omega$ is C^2, then $\Delta y\in L^2(\Omega)$ and therefore $y\in W^{2,2}(\Omega)$ (see Ref. 8). Repeating this argument we find that $y\in W^{2,r}(\Omega)$ for any $r\geqslant 1$ and hence $y\in C_B^1(\Omega)$. Furthermore, $y\in C^0(\overline{\Omega})$ because $y\in W_0^{1,r}(\Omega)$ for any $r\geqslant 1$. □

Remark: Let $\Gamma_0=\inf_{V\in K_{\alpha}}\Gamma(V)$, $\alpha\geqslant 1$ is fixed. Since Γ_0 is achieved by Γ at some $V_0\in K_{\alpha}$, then Γ_0 must be positive. Define the set

$$S(\Gamma_0)=\{V\in L^\infty(\Omega),\quad 0\leqslant V\leqslant H,\ \ \Gamma(V)=\Gamma_0\},$$

and the functional $\psi: S(\Gamma_0)\to]0,+\infty[$ by $\psi(V)=\|V\|_\alpha^\alpha$. Then it can be shown that the functional ψ attains its minimum on $S(\Gamma_0)$ at some V^* which is also a minimizer for $\Gamma(V)$ on K_α, i.e., $\|V^*\|_\alpha^\alpha=M$.

III. CHARACTERIZATION OF THE OPTIMAL POTENTIALS

We shall deal with the question of characterizing the external potentials. It is well known that in such a problem, at the optimum, the second eigenvalue λ_2 can be degenerate (i.e., multiple) and so it does not vary smoothly with V. Thus, some complications appear when we want to establish necessary (and sufficient) optimality conditions for our problem. Similar difficulties arise in the problem of Lagrange which consists in determining a non-negative function $A(x)$ defined on the interval $[0,l]$, $l>0$, which maximizes the first eigenvalue of the following problem,

$$(A^2w'')''+\lambda w''=0,$$

$$w(0)=w'(0)=0,\quad w(l)=w'(l)=0,$$

under the condition that

$$\int_0^l A(x)\ dx=\xi_0,\quad A(x)\geqslant 0,$$

ξ_0 being a given positive constant. The eigenvalue problem above does not exclude multiple first eigenvalue. For a long history upon the problem see Refs. 14–16.

Definitions: (1) A potential V in K_α is said to be bang-bang if

$$\text{meas}\ \{x\in\Omega,\ \ 0<V(x)<H\}=0.$$

(2) A real-valued bounded measurable function $P(x)$ defined on Ω is an admissible perturbation of a potential $V\in K_\alpha$ if $V+tP\in K_\alpha$ for all sufficiently small $|t|$. The function $P(x)$ is a right-admissible (resp. left-admissible) perturbation if $V+tP$ remains in K_α for positive (resp. negative) t. Any perturbation satisfies

$$\int_\Omega P(x)V(x)^{\alpha-1}\ dx=0\quad \text{if}\ \ \alpha>1,$$

and

$$\int_\Omega P(x)\ dx=0\quad \text{if}\ \ \alpha=1.$$

Let V be in K_α, α is fixed. We recall that if $\lambda_2(v)$ is nondegenerate, then it is Fréchet differentiable in a neighboorhood of V and

$$\frac{d\lambda_2(V+tP)}{dt}=\int_\Omega P(x)y_2^2(x)\ dx\quad \text{at}\ \ t=0$$

for any $P\in L^\infty(\Omega)$ (see Refs. 17, 18, and 9). Hence in the nondegenerate case we have

$$\frac{d\Gamma(V+tP)}{dt}=\int_\Omega P(x)(y_2^2-y_1^2)\ dx\quad \text{at}\ \ t=0,$$

where we assume the eigenfunctions y_1 and y_2 associated to $\lambda_2(V)$ to be normalized. If λ_2 is degenerate, then it is directionally differentiable at V in any direction P and it can split into a cluster of m eigenvalues $\lambda_{2,j}$, which can be considered as a set of analytic functions near $t=0$. If, for example, λ_2 is m-fold degenerate, then we have

$$\lambda_{2,j}(V+tP)=\lambda_2(V)+t\mu_j+o(t), \quad j=1,...,m, \tag{3.1}$$

where μ_j $(j=1,...,m)$ are the eigenvalues of the $m\times m$ self-adjoint matrix A with entries

$$a_{ij}=\int_\Omega P(x)y_{2,i}(x)y_{2,j}(x)\ dx,$$

and $\{y_{2,i}\}_{i=1}^m$ is an orthonormal eigenbasis for $\lambda_2(V)$. The eigenvalues $\lambda_{2,j}$ $(j=1,...,m)$ cannot be ordered in general in the increasing order because the lowest one for $t<0$ will typically be the highest for $t>0$. Moreover, the directional derivative of λ_2 at V in the direction P is the smallest eigenvalue of the group of eigenvalues μ_j, see Ref. 17. Finally, if Γ_j denotes the gap $\lambda_{2,j}-\lambda_1$ $(j=1,...,m)$, then the m derivatives

$$\frac{d\Gamma_j(V+tP)}{dt} \quad (t=0)$$

coincide with the eigenvalues of the $m\times m$ self-adjoint matrix B with entries

$$b_{ij}=\int_\Omega P(x)[y_{2,i}(x)y_{2,j}(x)-y_1^2(x)\,\delta_{ij}]\ dx=\langle Py_{2,i},y_{2,j}\rangle-\langle Py_1,y_1\rangle\delta_{ij}, \tag{3.2}$$

where $\langle,\rangle$ is the scalar product in $L^2(\Omega)$.

The main result of this work is the following.

Theorem 3.1: *Let Ω be a fixed smooth bounded domain in R^n and let the operator L be as above. If a function V_0 minimizes Γ over K_1, then*
(a) *$\lambda_2(V_0)$ is nondegenerate.*
(b) *In any open domain $\mathcal{O}$ in which V_0 does not achieve 0 and H, V_0 is infinitely differentiable, and is related to its first and second normalized eigenfunction y_1 and y_2 by*

$$V_0(x)=C[(\lambda_2 y_2^2(x)-\lambda_1 y_1^2(x))-(|\nabla y_2(x)|^2-|\nabla y_1(x)|^2)]$$

in $\mathcal{O}$, for some nonzero constant C.

Proof: (a) Suppose that $\lambda_2(V_0)$ is an m-fold eigenvalue. To prove that this assumption cannot hold, we shall distinguish two cases

First case. Assume that V_0 is not bang-bang and let $\mathcal{O}$ be an open domain where V_0 is between 0 and H. Since V_0 minimizes Γ over K_1, we must have in view of (3.1)

$$b_{ij}=\int_\Omega P(x)[y_{2,i}(x)y_{2,j}(x)-y_1^2(x)\,\delta_{ij}]\ dx=0, \quad 1\leqslant i,\ j\leqslant m,$$

for any admissible perturbation P satisfying supp $P\subset\mathcal{O}$. Since $\int_{\mathcal{O}}P(x)\ dx=0$, one deduces that

$$y_{2,i}^2(x)-y_1^2(x)=C_{ii} \quad \text{in } \mathcal{O}, \tag{3.3}$$

$$y_{2,i}(x)y_{2,j}(x)=C_{ij} \quad \text{in } \mathcal{O}, \tag{3.4}$$

for $1\leqslant i,\ j\leqslant m$. From (3.3) and (3.4) we obtain in particular

$$y_{2,2}^2-y_{2,1}^2=C_{22}-C_{11} \quad \text{in } \mathcal{O}, \tag{3.5}$$

$$y_{2,1}=C_{21}/y_{2,2} \quad \text{in } \mathcal{O}. \tag{3.6}$$

Thus, combining (3.5) and (3.6) we obtain the fourth-order equation

$$y_{2,2}^4(x)+(C_{22}-C_{11})y_{2,2}^2(x)-C_{21}=0,$$

holding for every x in $\mathcal{O}$. This implies that

$$y_{2,2}=\text{const.} \quad \text{in } \mathcal{O}.$$

From (3.3) we deduce that y_1 is also equal to a constant in $\mathcal{O}$ and therefore $\lambda_1=\lambda_2$, which is impossible since λ_1 is nondegenerate.

Second case. Suppose now that V_0 is bang-bang (i.e., we cannot find an admissible perturbation for V_0). Then for any right-admissible perturbation P, the eigenvalues of the matrix B must be non-negative, i.e., B must be positive semidefinite. Hence, we have in particular

$$\int_\Omega P(x)[y_{2,i}^2(x)-y_1^2(x)]\,dx\geqslant 0, \quad i=1,...,m,$$

for any right-admissible perturbation P. Put

$$B_+=\{x\in\Omega,|y_{2,1}(x)|<y_1(x)\}, \quad B_-=\{x\in\Omega,|y_{2,1}(x)|>y_1(x)\},$$

$$S_+=\{x\in\Omega,V_0(x)=H\}, \tag{3.7}$$

$$S_-=\{x\in\Omega,V_0(x)=0\}. \tag{3.8}$$

If $B_+\neq S_+$ and $B_-\neq S_-$, it is possible to find a right-admissible perturbation P (P takes negative values in S_+ and positive values in S_-) such that

$$\int_\Omega P(x)[y_{2,1}^2-y_1^2]\,dx<0.$$

This implies that either $B_+=S_+$ or $B_-=S_-$ is true. But since V_0 is bang-bang we have $B_+=S_+$ and $B_-=S_-$. The reasoning above is valid for $i=1,...,m$.

Let y be any normalized eigenfunction in the eigenspace of $\lambda_2(V_0)$ so that

$$y=\sum_{i=1}^m \alpha_i y_{2,i}, \quad \sum_{i=1}^m |\alpha_i|^2=1,$$

$$\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_m)\in R^n.$$

A straightforward calculation yields

$$\int_\Omega P(x)[y(x)^2-y_1(x)^2]\,dx=\sum_{1\leqslant i,j\leqslant m}\alpha_i\alpha_j\int_\Omega P(x)[y_{2,i}(x)y_{2,j}(x)-y_1^2(x)\,\delta_{ij}]\,dx=(B\alpha,\alpha)\geqslant 0$$

for any right-admissible perturbation P. Reasoning as above we find that $\{x\in\Omega,|y(x)|<y_1(x)\}=S_+$ and $\{x\in\Omega,|y(x)|>y_1(x)\}=S_-$. Since the interior of S_- (S_-^0) is not empty, we can find a point x_0 in S_-^0 such that $y_{2,1}(x_0)\neq 0$ and $y_{2,2}(x_0)\neq 0$. Let α_1 and α_2 be two nonzero reals such that $\alpha_1^2+\alpha_2^2=1$ and $\alpha_1 y_{2,1}(x_0)+\alpha_2 y_{2,2}(x_0)=0$, and consider the normalized eigenfunction $y=\alpha_1 y_{2,1}+\alpha_2 y_{2,2}$. Then we have $y(x_0)=0$ and $x_0\in S_-^0$, what is not compatible with the set equality: $S_-=\{x\in\Omega,|y(x)|>y_1(x)\}$. Consequently, $\lambda_2(V_0)$ is nondegenerate.

Let now y_2 be a normalized second eigenfunction for $\lambda_2(V_0)$ and suppose that V_0 is not bang-bang. Reasoning as above we arrive at

$$y_2^2(x)-y_1^2(x)=C \quad \text{in } \mathcal{O},$$

for some constant $C\neq 0$. This implies that $\Delta(y_2^2-y_1^2)=0$ in $\mathcal{O}$. However,

$$\Delta y_2^2=2|\nabla y_2|^2+2y_2\Delta y_2=2|\nabla y_2|^2+2(V_0-\lambda_2)y_2^2,$$

and similarly

$$\Delta y_1^2=2|\nabla y_1|^2+2(V_0-\lambda_1)y_1^2$$

in $\mathcal{O}$. Therefore,

$$V_0(x)=C^{-1}[(\lambda_2 y_2^2(x)-\lambda_1 y_1^2(x))-(|\nabla y_2(x)|^2-|\nabla y_1(x)|^2)]$$

in $\mathcal{O}$. Since y_1 and y_2 are $C^1(\Omega)$, we deduce that V_0 is continuous on $\mathcal{O}$. Repeating this argument leads to the conclusion that $V_0 \in C^k(\Omega)$ for any $k \geqslant 1$, which proves (b).

Remark: In the foregoing theorem we have proven that the degenerate case is excluded. Moreover, the argument used in the second part of the proof shows that one of the following cases,

$$\{|y_{2,1}|<y_1\}=S_+ \quad \text{or} \quad \{|y_{2,1}|>y_1\}=S_- \quad \text{a.e.},$$

must hold. S_+ and S_- are defined by (3.7) and (3.8), respectively. The remaining question is to know whether the potential V_0 is bang-bang. The answer is not clear even in the one-dimensional case.

Theorem 3.2: *Under the assumptions of the last theorem, if a function V_0 minimizes Γ over K_α, $\alpha>1$, then (a) $\lambda_2(V_0)$ is nondegenerate (b) In any open domain $\mathcal{O}$ in which V_0 does not equal 0 and H, V_0 is infinitely differentiable. Furthermore, y_1, y_2, and V_0 are related in $\mathcal{O}$ by*

$$y_2^2(x)-y_1^2(x)=CV_0(x)^{\alpha-1}$$

for some nonzero constant C, where y_1 (resp. y_2) is a normalized first (resp. second) eigenfunction for $\lambda_1(V_0)$ (resp. $\lambda_2(V_0)$).

The proof is entirely analogous to the one given above. To conclude that V_0 is infinitely differentiable we must take into account the fact that $y_2^2(x)$ cannot be equal to $y_1^2(x)$ for some x in $\mathcal{O}$ since $C \neq 0$.

Theorem 3.3: *Under the assumptions of Theorem 3.1, if a function V^0 maximizes Γ over K_1, then in any open domain $\mathcal{O}$ in which V_0 does not achieve 0 and H, it is infinitely differentiable. If $\lambda_2(V^0)$ is m-fold degenerate, then there exist m reals $s_1, s_2, \ldots, s_m$, such that*

$$\sum_{1 \leqslant i \leqslant m} s_i[y_{2,i}^2(x)-y_1^2(x)]=C \quad \text{in } \mathcal{O}$$

for some constant $C \neq 0$. $\{y_{2,j}\}_{i=1}^m$ is an eigenbasis for $\lambda_2(V^0)$.

Proof: Suppose that V^0 is m-fold degenerate. We have already seen that $\lambda_2(V^0)$ can split into a cluster of m eigenvalues $\lambda_{2,j}$ near $t=0$ and that if Γ_j denotes the gap $\lambda_{2,j}-\lambda_1$,

$$\Gamma_j(V^0+tP)=\Gamma(V^0)+t\nu_j+o(t),$$

where ν_j $(j=1,\ldots,m)$ are the eigenvalues of the self-adjoint matrix B with entires given by (3.2). Since V^0 maximizes Γ over K_1, for every admissible perturbation P there exist at least two eigenvalues ν_i and ν_j such that

$$\nu_i \nu_j \leqslant 0.$$

Let $\mathcal{O}$ be an open domain in which V_0 is between 0 and H. Let us consider the system of functions g_{ij} $(1 \leqslant i \leqslant j \leqslant m)$ defined by

$$g_{ij}=[y_{2,i}(x)y_{2,j}(x)-y_1^2(x)\delta_{ij}]\chi_{\mathcal{O}}, \quad 1 \leqslant i \leqslant j \leqslant m,$$

and $g_0=\chi_{\mathcal{O}}$. We prove that if g_0 and g_{ij} $(1 \leqslant i \leqslant j \leqslant m)$ are linearly independent, then there exists a bounded measurable function P defined on Ω such that supp $P \subset \overline{\mathcal{O}}$ and

$$\int_{\mathcal{O}} P(x)dx=0, \quad \int_{\mathcal{O}} P(x)g_{ij}(x)dx=\beta\delta_{ij}, \quad 1 \leqslant i \leqslant j \leqslant m,$$

where β is a given positive constant. The idea, based on Hahn–Banach theorem, is close to that used in Ref. 14 for solving a one-dimensional optimization problem.

The functions g_0 and g_{ij} are obviously in $L^1(\mathcal{O})$. Let $\mathcal{M}$ be the linear space formed by those functions and define the linear functional $F:\mathcal{M}\to R$ by $F(g_{ij})=\beta\delta_{ij}$ for $1\leqslant i\leqslant j\leqslant m$ and $F(g_0)=0$. For $g\in\mathcal{M}$, $g=\Sigma_{1\leqslant i\leqslant j\leqslant m}c_{ij}g_{ij}+c_0g_0$, we have

$$F(g)=\sum_{1\leqslant i\leqslant j\leqslant m} c_{ij}F(g_{ij})+c_0F(g_0)=\beta\sum_{i=1}^{m} c_{ii},$$

so the functional is well defined on $\mathcal{M}$. To prove the continuity of F, it is sufficient to find a number d for which

$$|F(g)|=\beta\left|\sum_{i=1}^{m} c_{ii}\right|\leqslant d\int_{\mathcal{O}}\sum_{1\leqslant i\leqslant j\leqslant m}|c_{ij}g_{ij}+c_0g_0|\,dx$$

for all $g\in\mathcal{M}$. This number can be chosen as

$$d=\sup_{\Sigma_{1\leqslant i\leqslant j\leqslant m}|c_{ij}|+|c_0|=1}\beta\left|\sum_{i=1}^{m} c_{ii}\right|\left(\int_{\mathcal{O}}\left|\sum_{1\leqslant i\leqslant j\leqslant m} c_{ij}g_{ij}+c_0g_0\right|\,dx\right)^{-1}.$$

The supremum of the last expression is finite by the assumption of the linear independence of the system of functions g_0 and g_{ij} $(1\leqslant i\leqslant j\leqslant m)$, which proves the continuity of F. The functional F constructed on the set $\mathcal{M}$ can, by the Hahn–Banach theorem, be extended by a continuous linear functional over the whole space $L^1(\mathcal{O})$. Therefore, there exists a function $P^*\in L^\infty(\mathcal{O})$ such that

$$\int_{\mathcal{O}}P^*(x)\,dx=0,\quad F(g)=\int_{\mathcal{O}}P^*(x)g(x)\,dx$$

for all $g\in\mathcal{M}$. If we extend the function P^* to the whole domain Ω by zero, we obtain a new function $P\in L^\infty(\Omega)$ satisfying

$$\text{Supp } P\subset\overline{\mathcal{O}},\quad \int_{\Omega}P(x)\,dx=0, \tag{3.9}$$

$$\int_{\Omega}P(x)[y_{2,i}y_{2,j}-y_1^2\delta_{ij}]\,dx=\beta\delta_{ij}. \tag{3.10}$$

Conditions (3.9) imply that P is an admissible perturbation and (3.10) show that the eigenvalues ν_j of the matrix B are all equal to β, i.e., the matrix B is positive definite, which is impossible. Consequently, g_0 and g_{ij} $(1\leqslant i\leqslant j\leqslant m)$ must be linearly dependent, i.e., there exist r_0 and r_{ij} $(1\leqslant i\leqslant j\leqslant m)$ nonzero reals such that

$$\sum_{1\leqslant i\leqslant j\leqslant m} r_{ij}g_{ij}(x)=r_0\quad\text{in }\mathcal{O}.$$

This expression can be written in a convenable orthonormal eigenbasis $\{z_{2,i}\}_{i=1}^m$ of $\lambda_2(V^0)$ as

$$\sum_{i=1}^{m} s_i(z_{2,i}^2-z_1^2)=s_0\quad\text{in }\mathcal{O}, \tag{3.11}$$

for some nonzero constants s_0 and s_i $(i=1,...,m)$. In (3.11), $z_1(x)=y_1(x)$. Reasoning as in the proof of Theorem 3.1, we find that

$$V^0(x)=s_0^{-1}\left(\lambda_2\sum_{i=1}^{m} z_{2,i}^2(x)-\lambda_1z_1^2(x)\right)-s_0^{-1}\left(\sum_{i=1}^{m}|\nabla z_{2,i}(x)|^2-|\nabla z_1(x)|^2\right)$$

in $\mathcal{O}$, which proves that V^0 is infinitely differentiable in $\mathcal{O}$.

[1] R. Svirsky, ''Maximally resonant potentials subject to p-Norm constraints,'' Pac. J. Math. **129**, 357–374 (1987).
[2] M. S. Ashbaugh, E. M. Harrell, and R. Svirsky, ''On minimal and maximal eigenvalues gaps and their causes,'' Pac. J. Math. **147**, 1–24 (1991).
[3] M. S. Ashbaugh and R. D. Benguria, ''A sharp bound for the ratio of the first two eigenvalues of Dirichlet Laplacians and extensions,'' Ann. Math. **135**, 601–628 (1992).
[4] M. S. Ashbaugh and R. D. Benguria, ''More bounds on eigenvalue ratios for Dirichlet Laplacians in N dimensions,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **24**, 1622–1651 (1993).
[5] W. Kirsch and B. Simon, ''Universal lower bounds on eigenvalue splitting for one dimentional Schrödinger operators,'' Commun. Math. Phys. **97**, 453–460 (1985).
[6] I. M. Singer, B. Wong, S.-T. Yau, and S. S.-T. Yau, ''An estimate of the gap of the first two eigenvalues in the Schrödinger operator,'' Ann. Scuola Norm. Sup. Pisa **4**, 319–333 (1985).
[7] W. Kirsch and B. Simon, ''Comparison theorems for the gap of Schrödinger operators,'' J. Funct. Anal. **75**, 596–410 (1987).
[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order* (Springer-Verlag, New York, 1983), 2nd ed.
[9] M. Reed and B. Simon, *Analysis of Operators, Methods of Modern Mathematical Physics IV* (Academic, New York 1978).
[10] Yu. V. Egorov and S. Karaa, ''Optimisation de la première valeur propre de l'opérateur de Sturm-Liouville,'' C. R. Acad. Sci. Paris **319**, 793–798 (1994).
[11] C. Jouron, ''Sur un problème d'optimisation où la contrainte porte sur la fréquence fondamentale,'' R.A.I.R.O Anal. Num. **12**, 349–374 (1978).
[12] S. Karaa, ''Sharp estimates for the eigenvalues of some differential equations,'' SIAM J. Math. Anal. (to appear).
[13] R. Tahraoui, ''Quelques remarques sur le controle des valeurs propres,'' in *Nonlinear Partial Differential Equations and Their Applications*, Collège de France Seminar, edited by J. L. Lions and H. Brézis (Litman, London, 1986), Vol. VIII, pp. 176–213.
[14] A. S. Bratus and A. P. Seiranian, ''Bimodal solutions in eigenvalue optimization problems,'' Prikl. Mat. Mekh. **47**, 546–554 (1983); Appl. Math. Mech. **47**, 451–457 (1983).
[15] A. P. Seiranian, ''On a problem of Lagrange,'' Inzh. Zh. Mech. Tverdogo Tela **26**, 1–7 (1984); Mech. Mater. 100–111 (1986).
[16] S. J. Cox and M. L. Overton, ''On the optimal design of columns against buckling,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **23**, 287–325 (1992).
[17] E. J. Haug, K. K. Choi, and V. Komkov, *Design Sensitivity Aanalysis and Structural Systems*, Mathematics in Science of Engineering (Academic, New York, 1986).
[18] T. Kato, *Perturbation Theory and Linear Operator* (Springer-Verlag, New York, 1976).

Perturbation theory for the Dirac equation in two-dimensional space

O. M. Kiselev[a)]
Ufa Institute of Mathematics, Russian Academy of Sciences,
112 Chernyshevsky str., Ufa, 450000, Russia

(Received 13 January 1997; accepted for publication 12 June 1997)

Basis functions associated with the two-component hyperbolic Dirac equation were obtained. The expansion via the basis functions separates time and spatial variables in a linearized Davey–Stewartson I equation. It is possible to solve this equation by the Fourier method. © *1998 American Institute of Physics.*
[S0022-2488(98)01104-9]

I. INTRODUCTION

In this work we develop the perturbation theory for the two-component hyperbolic Dirac equation:

$$\begin{pmatrix} \partial_\xi & 0 \\ 0 & \partial_\eta \end{pmatrix} \psi = -\frac{1}{2} \begin{pmatrix} 0 & q_1 \\ q_2 & 0 \end{pmatrix} \psi. \tag{1}$$

We have constructed the basis functions using formulas of the potential variations and of the scattering data variations. If we take into account $q_1 = Q$ and $q_2 = \sigma\overline{Q}$ (here the overbar denotes the complex conjugation), where the time evolution of Q is given by the Davey–Stewartson I equation (DS-I),[1,2]

$$\begin{gathered} i\partial_t Q + (\partial_\xi^2 + \partial_\eta^2)Q + (g_1 + g_2)Q = 0, \\ \partial_\xi g_1 = -\frac{\sigma}{2}\,\partial_\eta |Q|^2, \quad \partial_\eta g_2 = -\frac{\sigma}{2}\,\partial_\xi |Q|^2, \quad \sigma = \pm 1, \end{gathered} \tag{2}$$

then these basis functions make it possible to solve the DS-I equation linearized at Q and g_{12} as a background (LDS-I),

$$\begin{gathered} i\partial_t U + (\partial_\xi^2 + \partial_\eta^2)U + (g_1 + g_2)U + (V_1 + V_2)Q = F(\xi, \eta, t), \\ \partial_\xi V_1 = -\frac{\sigma}{2}\,\partial_\eta(Q\overline{U} + \overline{Q}U), \quad \partial_\eta V_2 = -\frac{\sigma}{2}\,\partial_\xi(Q\overline{U} + \overline{Q}U), \end{gathered} \tag{3}$$

by the Fourier method. These are two main results of the work, which are accurately formulated in Theorems 1 and 2.

The Davey–Stewartson equation (2) and the auxiliary linear system (1) were studied in Refs. 3–7. A system more general than (1) was studied in Ref. 8 as an auxiliary linear system for the three-wave interaction equation. A Hamiltonian structure and some properties of solutions for the Davey–Stewartson equation were studied in Refs. 9–12. A useful auxiliary linear system for a Davey–Stewartson hierarchy was studied in Ref. 13. Exact solutions of the equation were constructed in Refs. 14–17.

The results of the completeness for the squared solutions for the one-dimensional Dirac and Schrödinger equations were obtained in Refs. 18–23. They make it possible to solve the $1+1$-dimensional linearizations of the integrable equations and to investigate perturbations of the

[a)]Electronic mail: ok@imat.rb.ru

 © 1998 American Institute of Physics

$1+1$-dimensional integrable equations.[22,24–29] The same perturbation theory for the $2+1$-dimensional equation was obtained for the Kadomtsev–Petviashili equation and an auxiliary linear problem—a nonstationary Schrödinger equation in the periodic case.[30] We note also that the perturbation of the lump solution of the Davey–Stewartson II equation (which is found in Ref. 31) and the elliptic version of the two-component Dirac equation were investigated in Ref. 32.

In this work the results on the basis functions associated with the one-dimensional Dirac equation[18–20,22,23] are generalized to the two-dimensional Dirac equation. This makes it possible to solve the LDS-I equation by the Fourier method for arbitrary initial data and enables one to study a nonintegrable perturbation of some exact solutions for the DS-I in the same fashion as the perturbations of the $1+1$-dimensional integrable equations.[22–27]

There are two notes to be made now. First, the expansion on the basis functions is founded on the variation formulas for the potential and for the scattering data, hence the above expansion is the effect of a reversibility for the inverse scattering transform for (1), which was proved in Refs. 3 and 6. From this point of view we have constructed a reversible transform for the potential variations. Second, the result may be used for definition transform between the potentials and the Hamilton–Jacobi ''action-angle'' variables for the DS-I equation.[9,12]

In this work we use formulas of the inverse scattering transform for (2) from Refs. 3–7. The formulas are necessary for the accurate statement of the results for the paper. The main results of the work are formulated in Theorems 1 and 2 in the end of Sec. II. In Sec. III we deduce the formulas for the variations of the potential and of the scattering data. Section IV contains the theorem on the basic functions for finite potentials. In Sec. V we prove Theorem 1 through the approximation of an arbitrary integrable potential by a sequence of the finite potentials. Section VI contains the proof of Theorem 2.

II. THE MAIN RESULTS

After introducing some notations and results from the inverse scattering transform for the two-dimensional Dirac equation,[3–7] we state the main results of this paper in Theorems 1 and 2.

Let ψ^+ and ψ^- be the matrix solutions to the Goursat problems for the Dirac equation with the boundary conditions

$$\psi^+_{11}|_{\xi\to-\infty}=\exp\,(ik\eta),\quad \psi^+_{12}|_{\xi\to-\infty}=0,$$

$$\psi^+_{21}|_{\eta\to\infty}=0,\quad \psi^+_{22}|_{\eta\to-\infty}=\exp(-ik\xi);$$

$$\psi^-_{11}|_{\xi\to-\infty}=\exp(ik\eta),\quad \psi^-_{12}|_{\xi\to\infty}=0,$$

$$\psi^-_{21}|_{\eta\to-\infty}=0,\quad \psi^-_{22}|_{\eta\to-\infty}=\exp(-ik\xi).$$

These notations correspond to the notations in Ref. 5.

Denote by $\psi^+_{(j)}$, $j=1,2$, the column of the matrix ψ^+. Define the bilinear form:

$$(\chi,\mu)_q=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\xi\, d\eta(\chi_1\mu_1q_2+\chi_2\mu_2q_1). \tag{4}$$

Here χ_i and μ_i are elements of the columns χ and μ.

We denote by $\phi_{(1)}$ and $\phi_{(2)}$ the solutions conjugated to $\psi^+_{(1)}$ and $\psi^-_{(2)}$ with respect to bilinear form (4).

Using the formulas for the scattering data,[5] we can write the scattering data in the form

$$s_1(k,l)=\frac{1}{4\pi}\,(\psi^+_{(1)}(\xi,\eta,k),E_{(1)}(il\xi))_q, \tag{5}$$

$$s_2(k,l)=\frac{1}{4\pi}\,(\psi^-_{(2)}(\xi,\eta,k),E_{(2)}(il\eta))_q. \tag{6}$$

Here $E(z)=\mathrm{diag}\,(\exp(z),\exp(-z))$.

In Ref. 5 it was shown that the elements of matrix $\psi^{\pm}$ are analytic functions with respect to k at $\pm \mathrm{Im}(k)>0$. Using the scattering data, we can write the Riemann–Hilbert problem for ψ_{11}^{-} and ψ_{12}^{+} (see Ref. 5):

$$\psi_{11}^{-}(\xi,\eta,k)=\exp(ik\eta)+\exp(ik\eta)\left(\exp(-ik\eta)\int_{-\infty}^{\infty} dl\ s_1(k,l)\psi_{12}^{+}(\xi,\eta,l)\right)^{-},$$

$$\psi_{12}^{+}(\xi,\eta,k)=\exp(-ik\xi)\left(\exp(ik\xi)\int_{-\infty}^{\infty} dl\ s_2(k,l)\psi_{11}^{-}(\xi,\eta,l)\right)^{+}.$$

Here

$$(f(k))^{\pm}=\frac{1}{2i\pi}\int_{-\infty}^{\infty}\frac{dk'\ f(k')}{k'-(k\pm i0)}.$$

The Riemann–Hilbert problem for ψ_{21}^{-} and ψ_{22}^{+} is the following:

$$\psi_{21}^{-}(\xi,\eta,k)=\exp(ik\eta)\left(\exp(-ik\eta)\int_{-\infty}^{\infty} dl\ s_1(k,l)\psi_{22}^{+}(\xi,\eta,l)\right)^{-},$$

$$\psi_{22}^{+}(\xi,\eta,k)=\exp(-ik\xi)+\exp(-ik\xi)\left(\exp(ik\xi)\int_{-\infty}^{\infty} dl\, s_2(k,l)\psi_{21}^{-}(\xi,\eta,l)\right)^{+}.$$

Also, below we introduce the second bilinear form:

$$\langle\chi,\mu\rangle_s=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dk\ dl\,(\chi^1(l)\mu^1(k)s_2(k,l)+\chi^2(l)\mu^2(k)s_1(k,l)), \tag{7}$$

where χ^j is the element of the row χ.

Denote by $\varphi^{(j)}$, $j=1,2$, the row conjugated to $\psi^{(j)}=[\psi_{j1}^{-},\psi_{j2}^{+}]$ with respect to the bilinear form (7). Now we formulate the main result.

Theorem 1: *Let q_1 and q_2 be such that $\partial^{\alpha}q_{1,2}\in L_1\cap C^1$ for $|\alpha|\leqslant 3$ if $h_1(\xi,\eta)$ and $h_2(\xi,\eta)$ satisfy the following conditions: $\partial^{\alpha}h_{1,2}\in L_1\cap C^1$ for $|\alpha|\leqslant 4$ then h_1 and h_2 may be represented in the form*

$$h_1=\frac{-1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\hat{h}},\quad h_2=\frac{1}{\pi}\langle\psi^{(2)}(\xi,\eta,l),\varphi^{(2)}(\xi,\eta,k)\rangle_{\hat{h}},$$

where

$$\hat{h}_1=\frac{1}{4\pi}(\psi_{(1)}^{+}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_h,\quad \hat{h}_2=\frac{1}{4\pi}(\psi_{(2)}^{-}(\xi,\eta,k),\phi_{(2)}(\xi,\eta,l))_h.$$

The next statement makes it possible to solve the LDS-I equation by the Fourier method.

Theorem 2: *Let $q_1=Q$ and $q_2=\sigma\bar{Q}$ where Q is the solution of the DS-I equation with the boundary conditions $g_1|_{\xi\to-\infty}=0$ and $g_2|_{\eta\to-\infty}=0$, and Q satisfies the conditions of Theorem 1 at $t\in[0,T_0]$, $T_0>0$. If the solution of the LDS-I equation is the smooth and integrable function U with respect to ξ and η, where $\partial^{\alpha}U\in L_1\cap C^1$ and $\partial^{\alpha}F\in L_1\cap C^1$, for $|\alpha|\leqslant 4$ and $t\in[0,T_0]$, then*

$$\partial_t\hat{U}_1=i(k^2+l^2)\hat{U}_1+\hat{F}_1,$$

$$\partial_t\hat{U}_2=-i(k^2+l^2)\hat{U}_2+\hat{F}_2.$$

III. THE CONJUGATED FUNCTIONS AND VARIATIONS OF THE SCATTERING DATA

In this section we get the Volterra equations for the functions conjugated to $\psi^+_{(1)}$ and $\psi^-_{(2)}$ with respect to the first bilinear form (4). The next step is deduction of formulas for variations of the potentials and variations of the scattering data.

The Volterra equation for $\psi^+_{(1)}$ has the form

$$\psi^+_{(1)}(\xi,\eta,k)=E_{(1)}(ik\eta)+G_1[q]\psi^+_{(1)}.$$

Here $E_{(j)}(z)$ is the jth column of the matrix $E(z)$ and the operator $G_1[q]$ is

$$G_1[q]\psi^+_{(1)}=-\frac{1}{2}\begin{bmatrix}\int_{-\infty}^{\xi} d\xi'\ q_1(\xi',\eta)\psi^+_{21}(\xi',\eta,k)\\ \int_{-\infty}^{\eta} d\eta'\ q_2(\xi,\eta')\psi^+_{11}(\xi,\eta',k)\end{bmatrix}.$$

The operator conjugated to G_1 with respect to (4) is defined by the formula

$$(G_1[q]\mu,\chi)_q=(\mu,H_1[q]\chi)_q.$$

The column $\phi_{(1)}$ conjugated to $\psi^+_{(1)}$ is the solution of the equation

$$\phi_{(1)}(\xi,\eta,k)=E_{(1)}(ik\xi)+H_1[q]\phi_{(1)}.$$

Now we consider the variation formulas. Let δq_i be the variation of the potential. The equation for the variation of $\psi^+_{(1)}$ has the form

$$\delta\psi^+_{(1)}=G_1[q]\delta\psi^+_{(1)}+G_1[\delta q]\psi^+_{(1)}. \tag{8}$$

Compute the bilinear form:

$$(\delta\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_q=(G_1[q]\delta\psi^+_{(1)},\phi_{(1)})_q+(G_1[\delta q]\psi^+_{(1)},\phi_{(1)})_q.$$

After evaluations we obtain

$$(\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_{\delta q}=(\delta\psi^+_{(1)}(\xi,\eta,k),E_{(1)}(il\xi))_q+(\psi^+_{(1)}(\xi,\eta,k),E_{(1)}(il\xi))_{\delta q}.$$

On the other hand, we can derive the variation of $s_1(k,l)$:

$$\delta s_1(k,l)=\frac{1}{4\pi}(\delta\psi^+_{(1)}(\xi,\eta,k),E_{(1)}(il\xi))_q+\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),E_{(1)}(il\xi))_{\delta q}.$$

It yields

$$\delta s_1(k,l)=\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_{\delta q}. \tag{9}$$

So, we have the formula connecting the variations of the potentials with the variation of $s_1(k,l)$.

The second formula for the variation of the scattering data is obtained in a similar manner from the Volterra equation for $\psi^-_{(2)}$:

$$\psi^-_{(2)}(\xi,\eta,k)=E_{(2)}(ik\eta)+G_2[q]\psi^-_{(2)}.$$

Here

$$G_2[q]\psi^-_{(2)}=-\frac{1}{2}\begin{bmatrix} -\int_\xi^\infty d\xi' q_1(\xi',\eta)\psi^-_{22}(\xi',\eta,k) \\ \int_{-\infty}^\eta d\eta' q_2(\xi,\eta')\psi^-_{12}(\xi,\eta',k) \end{bmatrix}.$$

The column $\phi_{(2)}$ is conjugated to $\psi^-_{(2)}$. The variation formula for $s_2(k,l)$ has the form

$$\delta s_2(k,l)=\frac{1}{4\pi}(\psi^-_{(2)}(\xi,\eta,k),\phi_{(2)}(\xi,\eta,l))_{\delta q}. \tag{10}$$

Now we construct the functions conjugated with respect to the second bilinear form (7). In this way we get the formulas connecting the scattering data variations and the potential variations. These formulas are the mapping of the scattering data variations into the potential variations.

The formulas for the potentials are obtained in Ref. 5. Using the second bilinear form (7), we write them in the form

$$q_1(\xi,\eta)=\frac{-1}{\pi}\langle\Psi^{(1)}(\xi,\eta,l),E^{(1)}(ik\xi)\rangle_s, \tag{11}$$

$$q_2(\xi,\eta)=\frac{1}{\pi}\langle\Psi^{(2)}(\xi,\eta,l),E^{(2)}(ik\eta)\rangle_s. \tag{12}$$

Where $\psi^{(j)}$ is the row $[\psi^-_{j1},\psi^+_{j2}]$, analogously $E^{(j)}(z)$ is the row of the matrix $E(z)$.

From (11) the variation of the potential is

$$\delta q_1=\frac{-1}{\pi}\langle\delta\psi^{(1)}(\xi,\eta,l),E^{(1)}(ik\xi)\rangle_s-\frac{1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),E^{(1)}(ik\xi)\rangle_{\delta s}.$$

Consider the Riemann–Hilbert problem for the row $\psi^{(1)}=[\psi^-_{11},\psi^+_{12}]$:

$$\psi^{(1)}=E^{(1)}(ik\eta)+S[s]\psi^{(1)}.$$

Here $E^{(1)}(z)$ is the first row of the matrix $E(z)$ and operator $S[s]$ is defined by the formula

$$S[s]\psi^{(1)}=\Bigg[\exp(ik\eta)\Bigg(\exp(-ik\eta)\int_{-\infty}^\infty dls_1(k,l)\psi^+_{12}(\xi,\eta,l)\Bigg)^-,$$

$$\exp(-ik\xi)\Bigg(\exp(ik\xi)\int_{-\infty}^\infty dls_2(k,l)\psi^-_{11}(\xi,\eta,l)\Bigg)^+\Bigg].$$

The row conjugated to $\psi^{(1)}$ with respect to the second bilinear form (7) satisfies the Riemann–Hilbert problem:

$$\varphi^{(1)}(\xi,\eta,l)=E^{(1)}(il\xi)+T[s]\varphi^{(1)}.$$

Here

$$T[s]\varphi^{(1)}=\Bigg[-\exp(il\xi)\Bigg(\exp(-il\xi)\int_{-\infty}^\infty dks_1(k,l)\varphi^{(1)}_2(\xi,\sigma,k)\Bigg)^-,$$

$$-\exp(-il\eta)\Bigg(\exp(il\eta)\int_{-\infty}^\infty dks_2(k,l)\varphi^{(1)}_1(\xi,\eta,k)\Bigg)^+\Bigg],$$

where $\varphi^{(1)}_j$, $j=1,2$, is a component of the row $\varphi^{(1)}$.

Let δs_1 and δs_2 be the variations of the scattering data. The equation for the variation of the row $\psi^{(1)}$ has the form

$$\delta\psi^{(1)}(\xi,\eta,l)=S[s]\delta\psi^{(1)}+S[\delta s]\psi^{(1)}.$$

Compute the bilinear form:

$$\langle\delta\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_s=\langle S[s]\delta\psi^{(1)},E^{(1)}(ik\xi)\rangle_s+\langle\psi^{(1)}(\xi,\eta,l),E^{(1)}(ik\xi)\rangle_{\delta s}.$$

From (11) after evaluations we obtain

$$\delta q_1=\frac{-1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\delta s}. \tag{13}$$

In a similar manner we define the row $\phi^{(2)}$ conjugated to $\psi^{(2)}=[\psi^-_{21},\psi^-_{22}]$. The formula for the variation of q_2 has the form

$$\delta q_2=\frac{1}{\pi}\langle\psi^{(2)}(\xi,\eta,l),\varphi^{(2)}(\xi,\eta,k)\rangle_{\delta s}. \tag{14}$$

The formulas (9) and (10) define the mapping of the variations of the potentials into the variations of the scattering data, and the inverse map is defined by the formulas (13) and (14).

IV. THE FINITE POTENTIALS

In this section we prove that the formulas (9), (10), (13), and (14) define an expansion through the basis functions.

Denote $\delta q_i=f_i$, $i=1,2$. The composition of the formulas (13), (14), (9), and (10) are

$$\begin{aligned}
\hat f_1&=\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_f,\\
\hat f_2&=\frac{1}{4\pi}(\psi^-_{(2)}(\xi,\eta,k),\phi_{(2)}(\xi,\eta,l))_f;\\
h_1&=\frac{-1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\hat f},\\
h_2&=\frac{1}{\pi}\langle\psi^{(2)}(\xi,\eta,l),\varphi^{(2)}(\xi,\eta,k)\rangle_{\hat f}.
\end{aligned} \tag{15}$$

Our purpose is to show the equivalence $h_i\equiv f_i$. The proof is based on two properties of the elements of the matrix ϕ and ψ. The first property is the analyticity of the functions with respect to $k,l\in\mathbb{C}$, if q_1 and q_2 have a finite support. The analyticity is evident from the Volterra equations for these functions. The second property is the asymptotic behavior of ϕ and ψ at $|k|$ or $|l|\to\infty$.

The asymptotics of $\psi^{\pm}_{(1)}$ and $\psi^{\pm}_{(2)}$ are obtained in Ref. 5:

$$\psi^{\pm}_{(1)}(\xi,\eta,k)=\left(\begin{bmatrix}1\\0\end{bmatrix}+\frac{1}{k}\begin{bmatrix}\dfrac{-i}{4}\displaystyle\int_{-\infty}^{\xi}d\xi'\ q_1(\xi',\eta)q_2(\xi',\eta)\\ \dfrac{i}{2}q_2\end{bmatrix}+O\left(\frac{1}{k^2}\right)\right)\exp(ik\eta),$$

$$\psi^{\mp}_{(2)}(\xi,\eta,k)=\left(\begin{bmatrix}0\\1\end{bmatrix}+\frac{1}{k}\begin{bmatrix}\dfrac{-i}{2}q_1\\ \dfrac{i}{4}\displaystyle\int_{-\infty}^{\eta}d\eta'\ q_1(\xi,\eta')q_2(\xi,\eta')\end{bmatrix}+O\left(\frac{1}{k^2}\right)\right)\exp(-ik\xi).$$

The columns $\phi_{(1)}$ and $\phi_{(2)}$ have the same asymptotic behavior:

$$\phi_{(1)}(\xi,\eta,k)=\left(\begin{bmatrix}1\\0\end{bmatrix}+\frac{1}{k}\begin{bmatrix}\dfrac{-i}{4}\displaystyle\int_{-\infty}^{\eta}d\eta'\ q_1(\xi,\eta')q_2(\xi,\eta')\\ \dfrac{-i}{2}\,q_2\end{bmatrix}+O\left(\frac{1}{k^2}\right)\right)\exp(ik\xi),$$

$$\phi_{(2)}(\xi,\eta,k)=\left(\begin{bmatrix}0\\1\end{bmatrix}+\frac{1}{k}\begin{bmatrix}\dfrac{i}{2}\,q_1\\ \dfrac{i}{4}\displaystyle\int_{-\infty}^{\xi}d\xi\ q_1(\xi',\eta)q_2(\xi',\eta)\end{bmatrix}+O\left(\frac{1}{k^2}\right)\right)\exp(-ik\eta).$$

Construct the asymptotics of $\varphi^{(1)}$ at $|k|\to\infty$. The leading term and the correction of the asymptotic behavior are the following:

$$\varphi^{(1)}(\xi,\eta,k)=\Bigg([\exp(ik\xi),0]+\frac{1}{k}\Bigg[\frac{1}{2i\pi}\exp(ik\xi)\int_{-\infty}^{\infty}dk'\int_{-\infty}^{\infty}dl\ s_1(l,k')\varphi_2^{(1)}\times(\xi,\eta,l)\exp(ik'\xi),\ \frac{1}{2i\pi}\exp(-ik\eta)\int_{-\infty}^{\infty}dk'\int_{-\infty}^{\infty}dl\ s_2(l,k')\varphi_1^{(1)}(\xi,\eta,l)\exp(ik'\eta)\Bigg]+O\left(\frac{1}{k^2}\right).$$

On the other hand,

$$q_1=\frac{-1}{\pi}\langle E^{(1)}(il\eta),\varphi^{(1)}(\xi,\eta,k)\rangle_s,$$

$$\int_{-\infty}^{\eta}d\eta'\ q_1(\xi,\eta')q_2(\xi,\eta')=\frac{2}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dl\ dk\ s_2(k,l)\exp(ik\xi)\psi_{21}^{-}(\xi,\eta,l).$$

The second of these formulas was obtained in Ref. 5. It yields

$$\int_{-\infty}^{\eta}d\eta'\ q_1(\xi,\eta')q_2(\xi,\eta')=\frac{2}{\pi}\langle\psi^{(2)}(\xi,\eta,l),E^{(1)}(ik\xi)\rangle_s=\frac{2}{\pi}\langle E^{(2)}(il\eta),\varphi^{(1)}(\xi,\eta,k)\rangle_s.$$

The asymptotic of the $\varphi^{(1)}$ at $|k|\to\infty$ has the form

$$\varphi^{(1)}(\xi,\eta,k)=[\exp(ik\xi),0]+\frac{1}{k}\left[\frac{\exp(ik\xi)}{4i}\int_{-\infty}^{\eta}d\eta'\ q_1(\xi,\eta')q_2(\xi,\eta),\ \frac{\exp(-ik\eta)}{-2i}\,q_1\right]+O\left(\frac{1}{k^2}\right).$$

In a similar manner we obtain the asymptotic behavior of $\varphi^{(1)}$ at $|k|\to\infty$:

$$\varphi^{(2)}(\xi,\eta,k)=[0,\exp(-ik\eta)]+\frac{1}{k}\left[\frac{\exp(ik\xi)}{2i}\,q_2,\quad\frac{\exp(-ik\eta)}{-4i}\int_{-\infty}^{\xi}d\xi'\ q_1(\xi',\eta)q_2(\xi',\eta)\right]+O\left(\frac{1}{k^2}\right).$$

Prove the equivalence $h_i=f_i$. The composition of the formulas (13), (9), and (10) gives

$$h_1=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ dl\Big(\psi_1^{(1)}(\xi,\eta,l)\varphi_1^{(1)}(\xi,\eta,k)$$

$$\times\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi'\ d\eta'[\psi^1_{(2)}(\xi',\eta',k)\phi^1_{(2)}(\xi',\eta',l)f_2(\xi',\eta')$$

$$+\psi^2_{(2)}(\xi',\eta',k)\phi^2_{(2)}(\xi',\eta',l)f_1(\xi',\eta')]+\psi^{(1)}_2(\xi,\eta,l)\varphi^{(1)}_2$$

$$\times(\xi,\eta,k)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi'\ d\eta'[\psi^1_{(1)}(\xi',\eta',k)\phi^1_{(1)}(\xi',\eta',l)f_2(\xi',\eta')$$

$$+\psi^2_{(1)}(\xi',\eta',k)\phi^2_{(1)}(\xi',\eta',l)f_1(\xi',\eta')]\Bigg).$$

Represent the integrals with respect to k and l as a divergent improper integrals:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ dl(\cdots)=\lim_{R\to\infty}\int_{-R}^{R}\int_{-R}^{R}dk\ dl(\cdots).$$

Replace the integrals with respect to ξ' and η' with the integrals with respect to k and l. The integrand is an analytic function with respect to k and l as k and $l\in\mathbb{C}$. Hence we can transform the paths of the integration with respect to k and l to large semicircles in the upper or lower half-planes of the complex parameters k and l. We denote the paths of the integration by C_R^+ or C_R^-. We represent the integrals with respect to ξ' and η' as the sum of the integrals over $\xi'\in(-\infty,\xi]$, $\eta'\in(-\infty,\eta]$ and $\xi'\in(\xi,\infty)$, $\eta'\in(\eta,\infty)$. Transform the path of the integration with respect to k and l to C_R^+ or C_R^-, respectively. Substitute the asymptotics of the integrand at $|k|\to\infty$ and $|l|\to\infty$. As a result we obtain

$$h=\frac{1}{4\pi}\lim_{R\to\infty}\Bigg[\Bigg(\int_{-\infty}^{\xi}\int_{-\infty}^{\eta}d\xi'\ d\eta'\ f_2(f',\eta')\int_{C_R^+}dk\int_{C_R^+}dl$$

$$+\int_{\xi}^{\infty}\int_{\eta}^{\infty}d\xi'\ d\eta'\ f_2(\xi',\eta')\int_{C_R^-}dk\int_{C_R^-}dl\Bigg)$$

$$\times\left[\frac{\exp(il(\eta-\eta'))}{l}\,\frac{i}{2}\,q_1(\xi',\eta')\,\frac{\exp(ik(\xi-\xi'))}{k}\,\frac{-i}{2}\,q_1(\xi',\eta')\right]$$

$$+\Bigg(\int_{-\infty}^{\xi}\int_{-\infty}^{\eta}d\xi'\ d\eta'\ f_2(\xi',\eta')\int_{C_R^+}dk\int_{C_R^+}dl+\int_{\xi}^{\infty}\int_{\eta}^{\infty}d\xi'\ d\eta'\ f_2(\xi',\eta')\int_{C_R^-}dk\int_{C_R^-}dl\Bigg)$$

$$\times\left[\frac{\exp(il(\xi'-\xi))}{l}\,\frac{-i}{2}\,q_1(\xi,\eta)\,\frac{\exp(ik(\eta'-\eta))}{k}\,\frac{i}{2}\,q_1(\xi,\eta)\right]$$

$$+\Bigg(\int_{-\infty}^{\xi}\int_{-\infty}^{\eta}d\xi'\ d\eta'\ f_1(\xi',\eta')\int_{C_R^+}dk\int_{C_R^+}dl+\int_{\xi}^{\infty}\int_{\eta}^{\infty}d\xi'\ d\eta'\ f_1(\xi',\eta')\int_{C_R^-}dk\int_{C_R^-}dl\Bigg)$$

$$\times f_1(\xi',\eta')\int_{C_R}dk\int_{C_R}dl\Bigg[\left(1+\frac{1}{l}\,\frac{-i}{4}\int_{-\infty}^{\xi}d\xi''\ q_1(\xi'',\eta)q_2(\xi'',\eta)\right)$$

$$\times\exp(il\eta)\left(1+\frac{1}{k}\,\frac{1}{4i}\int_{-\infty}^{\eta}d\eta''\ q_1(\xi,\eta'')q_2(\xi,\eta'')\right)$$

$$\times\exp(ik\xi)\left(1+\frac{1}{k}\,\frac{i}{4}\int_{-\infty}^{\eta'}d\eta''\ q_1(\xi',\eta'')q_2(\xi',\eta'')\right)$$

$$\times\exp(-ik\xi')\left(1+\frac{1}{l}\,\frac{i}{4}\int_{-\infty}^{\xi'}d\xi''\ q_1(\xi'',\eta')q_2(\xi'',\eta')\right)\exp(-il\eta')\Bigg]\Bigg].$$

Then transform the paths of integration to the axis $\mathrm{Im}(l)=0$ and $\mathrm{Im}(k)=0$. It yields

$$h=\frac{1}{4\pi}\lim_{R\to\infty}\int_{-\infty}^{\infty}d\xi'\int_{-\infty}^{\infty}d\eta'\int_{-R}^{R}dk\int_{-R}^{R}dl\ f_1(\xi',\eta')\exp(il(\eta-\eta'))\exp(ik(\xi-\xi'))=f_1.$$

Thus we have proved the statement.

Theorem 3: *Let q_1 and q_2 be smooth functions with finite support. If f_1 and f_2 are smooth integrable functions with respect to $\xi,\eta\in\mathbb{R}$, then we may represent f_1 and f_2 in the form*

$$f_1=\frac{-1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\hat{f}},\quad f_2=\frac{1}{\pi}\langle\psi^{(2)}(\xi,\eta,l),\varphi^{(2)}(\xi,\sigma,k)\rangle_{\hat{f}},$$

where

$$\hat{f}_1=\frac{1}{4\pi}(\psi^{+}_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_f,\quad \hat{f}_2=\frac{1}{4\pi}(\psi^{-}_{(2)}(\xi,\eta,k),\phi_{(2)}(\xi,\eta,l))_f.$$

V. THE COMPLETENESS THEOREM

In the above section we proved the theorem of the representation of functions from $L_1\cap C^1$ through basis functions connecting with the Dirac system, when $q_{1,2}$ has a finite support. Here it is shown that this expansion is appropriate for the more general-type potentials. In this case there are additional restrictions on the functions to be expanded.

To prove Theorem 1 we approximate the potentials q_1 and q_2 by the potentials $\overset{N}{q}_1$ and $\overset{N}{q}_2$ with finite support. For $\varepsilon>0$ we take $N(\varepsilon)>0$ such that $\|\partial^\alpha\overset{N}{q}_{1,2}-\partial^\alpha q_{1,2}\|_{L_1}=O(\varepsilon)$, where $|\alpha|\leqslant 3$ and $\overset{N}{q}_{1,2}=q_{1,2}$ for $\xi^2+\eta^2\leqslant N(\varepsilon)$; and $\overset{N}{q}_{1,2}=0$ for $\xi^2+\eta^2>2N(\varepsilon)$. We denote an error of the approximation by $\overset{\varepsilon}{q}_{1,2}=q_{1,2}-\overset{N}{q}_{1,2}$. Throughout in this section we use a notation with an upper symbol N for solutions of the Goursat problems with potentials $\overset{N}{q}_{1,2}$. The upper symbol ε denotes functions associated with the approximation error $\overset{\varepsilon}{q}_{1,2}$.

Lemma 1: $\|\overset{N}{\psi}{}^{\pm}_{ij}-\psi^{\pm}_{ij}\|_{C^3}=O(\varepsilon)$, *for* $i,j=1,2$. *Here* $\overset{N}{\psi}{}^{\pm}_{ij}$ *are elements of the matrices being the solutions of the Goursat problems with the potentials* $\overset{N}{q}_1$ *and* $\overset{N}{q}_2$.

The proof: Denote by $\chi^{\pm}$ an expression

$$\chi^{\pm}(\xi,\eta,k)=\begin{pmatrix}\exp(-ik\eta) & 0\\ 0 & \exp(ik\xi)\end{pmatrix}\psi^{\pm}(\xi,,\eta,k). \tag{16}$$

Represent the matrix $\chi^{\pm}$ in a form $\chi^{\pm}=\overset{N}{\chi}{}^{\pm}+\overset{\varepsilon}{\chi}{}^{\pm}$. Here $\overset{N}{\chi}{}^{\pm}$ has the same representation through $\overset{N}{\psi}$ as (16). The Volterra equations for the column $\overset{\varepsilon}{\chi}{}^{+}_{(1)}$ have the form

$$\overset{\varepsilon}{\chi}{}^{+}_{(1)}=\overset{\varepsilon}{\mathscr{G}}_1[q]\overset{N}{\chi}{}^{+}_{(1)}+\overset{\varepsilon}{\mathscr{G}}_1[q]\overset{\varepsilon}{\chi}{}^{+}_{(1)}, \tag{17}$$

where

$$\overset{\varepsilon}{\mathscr{G}}_1[q]\overset{\varepsilon}{\chi}{}^{+}_{(1)}=\frac{1}{2}\begin{pmatrix}-\int_{-\infty}^{\xi}d\xi'\ q_1(\xi',\eta)\overset{\varepsilon}{\chi}{}^{+}_{21}(\xi',\eta,k)\\ \int_{\eta}^{\infty}d\eta'\ q_2(\xi,\eta')\overset{\varepsilon}{\chi}{}^{+}_{11}(\xi,\eta',k)\exp(ik(\eta'-\eta))\end{pmatrix}.$$

The Neyman series for this system converge and $|\overset{\varepsilon}{\chi}{}^{+}_{11}|+|\overset{\varepsilon}{\chi}{}^{+}_{21}|=O(\varepsilon)$ for $\forall \xi,\eta\in\mathbb{R}$. The same estimates are also true for derivatives: $|\partial^{\alpha}\overset{\varepsilon}{\chi}{}^{+}_{11}|+|\partial^{\alpha}\overset{\varepsilon}{\chi}{}^{+}_{21}|=O(\varepsilon)$ for $\forall \xi,\eta\in\mathbb{R}$ as $|\alpha|\leqslant 3$.

Similar results are obtained for $\overset{\varepsilon}{\chi}{}^{\pm}_{(2)}$ and $\overset{\varepsilon}{\chi}{}^{-}_{(1)}$. Hence the lemma is proved.

Below we shall use the asymptotic behavior of the $\overset{\varepsilon}{\chi}{}^{\pm}$ when $|k|\to\infty$.

Lemma 2: When $|k|\to\infty$ the asymptotic behavior of $\overset{\varepsilon}{\chi}{}^{\pm}$ has the form

$$\overset{\varepsilon}{\chi}{}^{\pm}_{(1)}=\frac{1}{k}\begin{pmatrix} -\frac{i}{4}\int_{-\infty}^{\xi} d\xi'(\overset{\varepsilon}{q}_1\overset{N}{q}_2+\overset{\varepsilon}{q}_2q_1) \\ \frac{i}{2}\overset{\varepsilon}{q}_2 \end{pmatrix}+O(k^{-2}). \tag{18}$$

The formula is true uniformly with respect to $\xi,\eta\in\mathbb{R}$. The remainder is differentiable with respect to ξ:

$$\overset{\varepsilon}{\chi}{}^{\pm}_{(2)}=\frac{1}{k}\begin{pmatrix} \frac{-i}{2}\overset{\varepsilon}{q}_1 \\ \frac{i}{4}\int_{-\infty}^{\xi} d\eta'(\overset{\varepsilon}{q}_2\overset{N}{q}_1+\overset{\varepsilon}{q}_1q_2) \end{pmatrix}+O(k^{-2}). \tag{19}$$

The formula is true uniformly with respect to $\xi,\eta\in\mathbb{R}$. The remainder is differentiable with respect to η.

The asymptotics can be constructed in the same way as the asymptotics of the $\psi^{\pm}$.

Prove Theorem 1. Denote

$$\overset{N}{f}_i=\frac{1}{4\pi}(\overset{N}{\psi}_{(i)}(\xi,\eta,k),\overset{N}{\phi}_{(i)}(\xi,\eta,l))_f.$$

Represent f_1 as

$$f_1=\frac{-1}{\pi}\langle\overset{N}{\psi}{}^{(1)}(\xi,\eta,l),\overset{N}{\varphi}{}^{(1)}(\xi,\eta,k)\rangle^{N}_{f}.$$

Estimate the difference f_1-h_1, where

$$h_1=\frac{-1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\hat{f}}, \tag{20}$$

$$\begin{aligned} f_1-h_1&=\frac{-1}{\pi}\langle\overset{N}{\psi}{}^{(1)}(\xi,\eta,l),\overset{N}{\varphi}{}^{(1)}(\xi,\eta,k)\rangle^{N}_{f}+\frac{1}{\pi}\langle\psi^{(1)}(\xi,\eta,l),\varphi^{(1)}(\xi,\eta,k)\rangle_{\hat{f}} \\ &=\frac{-1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dk\,dl[\overset{N}{f}_2(k,l)(\overset{N}{\psi}{}^{-}_{11}(\xi,\eta,l)\overset{N}{\varphi}_{11}(\xi,\eta,k)-\psi^{-}_{11}(\xi,\eta,l)\varphi_{11}(\xi,\eta,k)) \\ &\quad+(\overset{N}{f}_2(k,l)-\hat{f}_2(k,l))\psi^{-}_{11}(\xi,\eta,l)\varphi_{11}(\xi,\eta,k)+\overset{N}{f}_1(k,l)(\overset{N}{\psi}{}^{+}_{12}(\xi,\eta,l)\overset{N}{\varphi}_{12}(\xi,\eta,k) \\ &\quad-\psi^{+}_{12}(\xi,\eta,l)\varphi_{12}(\xi,\eta,k))+(\overset{N}{f}_1(k,l)-\hat{f}_1(k,l))\psi^{+}_{12}(\xi,\eta,l)\varphi_{12}(\xi,\eta,k)]. \end{aligned}$$

Consider the diverse terms and estimate them individually. Denote

$$I_1=\frac{-1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ dl\ \overset{N}{f}_2(k,l)(\overset{N}{\psi}{}^{-}_{11}(\xi,\eta,l)\overset{N}{\varphi}_{11}(\xi,\eta,k)-\psi^{-}_{11}(\xi,\eta,l)\varphi_{11}(\xi,\eta,k)).$$

From Lemma 1 it follows that

$$|\overset{N}{\psi}{}^{-}_{11}(\xi,\eta,l)\overset{N}{\varphi}_{11}(\xi,\eta,k)-\psi^{-}_{11}(\xi,\eta,l)\varphi_{11}(\xi,\eta,k)|=O(\varepsilon).$$

To show that $I_1=O(\varepsilon)$ it is sufficient that $\overset{N}{f}_2(k,l)\in L_1$.

Lemma 3: $\overset{N}{f}_2(k,l)\in L_1$.

The proof: Construct the asymptotics of $\overset{N}{f}_2(k,l)$ as $|k|\to\infty$ uniformly when $l\in\mathbb{R}$ and asymptotics of $\overset{N}{f}_2(k,l)$ as $|l|\to\infty$ uniformly when $k\in\mathbb{R}$:

$$\overset{N}{f}_2=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi\ d\eta\left[\left(\frac{-1}{2k}\overset{N}{q}_1(\xi,\eta)+O(k^{-2})\right)\exp(-ik\xi)\overset{N}{\phi}_{12}(\xi,\eta,l)\right.$$
$$\left.+\left(1+\frac{i}{4k}\int_{-\infty}^{\eta}d\eta'\overset{N}{q}_1(\xi,\eta')\overset{N}{q}_2(\xi,\eta')+O(k^{-2})\right)\exp(-ik\xi)\overset{N}{\phi}_{22}(\xi,\eta,l)\right].$$

Represent the integral as a sum of integrals. Integrate the terms with asymptotics $O(k^{-1})$ as $|k|\to\infty$ by parts with respect to ξ. Twice integrate by parts with respect to ξ the term

$$I_1'=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi\ d\eta\ f_2(\xi,\eta)\exp(-ik\xi)\overset{N}{\phi}_{22}(\xi,\eta,l).$$

As a result we obtain

$$\overset{N}{f}_2=O(k^{-2}),\quad \text{as}\quad |k|\to\infty\quad \text{uniformly when}\quad l\in\mathbb{R}.$$

In the same way we can show

$$\overset{N}{f}_2=O(l^{-2}),\quad \text{as}\quad |l|\to\infty\quad \text{uniformly when}\quad k\in\mathbb{R}.$$

Hence $\overset{N}{f}_2\in L_1$. The lemma is proved.

We get $I_1=O(\varepsilon)$.

Consider one more term of the formulas (20):

$$I_2=\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ dl\ \psi^{-}_{11}(\xi,\eta,l)\varphi_{11}(\xi,\eta,k)(\overset{N}{f}_2(k,l)-\hat{f}_2(k,l)).$$

Lemma 4: $\|\overset{N}{f}_2-\hat{f}_2(k,l)\|_{L_1}=O(\varepsilon)$.

The proof:

$$\overset{N}{f}_2(k,l)-\hat{f}_2(k,l)=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi\ d\eta[f_2(\xi,\eta)[(\overset{N}{\psi}_{12}(\xi,\eta,k)-\psi_{12}(\xi,\eta,k))\overset{N}{\phi}_{12}(\xi,\eta,l)$$
$$+(\overset{N}{\phi}_{12}(\xi,\eta,l)-\phi_{12}(\xi,\eta,l))\overset{N}{\psi}_{12}(\xi,\eta,k)]+f_1(\xi,\eta)[(\overset{N}{\psi}_{22}(\xi,\eta,k)$$

$$-\psi_{22}(\xi,\eta,k))\overset{N}{\phi}_{22}(\xi,\eta,l)+(\overset{N}{\phi}_{22}(\xi,\eta,l)-\phi_{22}(\xi,\eta,l))\overset{N}{\psi}_{22}(\xi,\eta,k)]]. \tag{21}$$

Represent the integral as a sum of the integrals with respect to ξ and η. Consider the asymptotic behavior of one of the terms when $|k|\to\infty$:

$$J_1=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi\,d\eta\,f_1(\xi,\eta)(\overset{N}{\psi}_{22}(\xi,\eta,k)-\psi_{22}(\xi,\eta,k))\overset{N}{\phi}_{22}(\xi,\eta,l)$$

$$=\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\xi\,d\eta\,f_1(\xi,\eta)\left[\frac{i}{4k}\int_{-\infty}^{\eta}d\eta'(\overset{\varepsilon}{q}_2(\xi,\eta')\overset{N}{q}_1(\xi,\eta')+\overset{\varepsilon}{q}_1(\xi,\eta')\overset{N}{q}_2(\xi,\eta'))+O(k^{-2})\right]\exp(ik\xi)\overset{N}{\phi}_{22}(\xi,\eta,l).$$

After integration by parts with respect to ξ we obtain

$$J_1=O(\varepsilon)O(k^{-2}) \quad \text{as} \quad |k|\to\infty, l\in\mathbb{R}.$$

By the same way we obtain

$$J_1=O(\varepsilon)O(l^{-2}) \quad \text{as} \quad |l|\to\infty, k\in\mathbb{R}.$$

Moreover, $J_1=O(\varepsilon)$ as $k,l\in\mathbb{R}$. It follows from Lemma 1. As a result we obtain

$$\|J_1\|_{L_1}=O(\varepsilon).$$

By the same way we can estimate the remaining terms in (21). As a result we obtain the statement of Lemma 4.

From Lemma 4 it follows $I_2=O(\varepsilon)$. The other terms of the formula (20) are estimated in the same way. As a result we obtain

$$|f_1-h_1|=O(\varepsilon) \quad \text{as} \quad \xi,\eta\in\mathbb{R}.$$

The similar formula is true for f_2 and h_2:

$$|f_2-h_2|=O(\varepsilon) \quad \text{as} \quad \xi,\eta\in\mathbb{R}.$$

Let $\varepsilon\to 0$. Then we obtain the statement of Theorem 1.

VI. THE TIME EVOLUTION OF THE SCATTERING DATA VARIATIONS

In this section we study the time evolution of the scattering data variations in the special case, where $q_1=Q$, $q_2=\sigma\overline{Q}$, and Q is the solution of the DS-I equation with the boundary conditions $g_1|_{\xi\to-\infty}=0$ and $g_2|_{\eta\to-\infty}=0$.

The variations of the scattering data are

$$\hat{U}_2=\frac{1}{4\pi}(\psi^-_{(2)}(\xi,\eta,k),\phi_{(2)}(\xi,\eta,l))_U, \quad \hat{U}_1=\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_U.$$

Here $U(\xi,\eta,t)$ is the solution of the LDS-I (3) and $U_2=\sigma\overline{U_1}$. In this case $\psi^+_{(1)}=\psi^+_{(1)}(\xi,\eta,k,t)$ and it is the solution of the system of the equations[4]

$$\partial_t\psi^+_{(1)}=ik^2\psi^+_{(1)}+i\begin{pmatrix}1&0\\0&-1\end{pmatrix}(\partial_\xi-\partial_\eta)^2\psi^+_{(1)}+i\begin{pmatrix}0&q_1\\q_2&0\end{pmatrix}(\partial_\xi-\partial_\eta)\psi^+_{(1)}+\begin{pmatrix}ig_1&-i\partial_\eta q_1\\i\partial_\xi q_2&-ig_2\end{pmatrix}\psi^+_{(1)}. \tag{22}$$

We can show that $\phi_{(1)}$ satisfies the system:

$$\partial_t\phi_{(1)}=ik^2\phi_{(1)}+i\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}(\partial_\eta-\partial_\xi)^2\phi_{(1)}-i\begin{pmatrix}0 & q_1\\ q_2 & 0\end{pmatrix}(\partial_\eta-\partial_\xi)\phi_{(1)}+\begin{pmatrix}ig_2 & i\partial_\xi q_1\\ -i\partial_\eta q_2 & -ig_1\end{pmatrix}\phi_{(1)}. \tag{23}$$

Take the derivative of the function $\hat{U}_1$ with respect to t:

$$\partial_t\hat{U}_1=\frac{1}{4\pi}(\partial_t\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_U+\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),\partial_t\phi_{(1)}(\xi,\eta,l))_U$$

$$+\frac{1}{4\pi}(\psi^+_{(1)}(\xi,\eta,k),\phi_{(1)}(\xi,\eta,l))_{\partial_t U}.$$

The derivative of $\psi^+_{(1)}$ is replaced by (22), the derivative of $\phi_{(1)}$ is replaced by (23), and the derivatives of U_1 and U_2 are replaced by (3). After some speculations we obtain

$$\partial_t\hat{U}_1=i(k^2+l^2)\hat{U}_1+\hat{F}_1.$$

In a similar way we obtain

$$\partial_t\hat{U}_2=-i(k^2+l^2)\hat{U}_2+\hat{F}_2.$$

The theorem 2 is proved.

ACKNOWLEDGMENT

The work was financially supported by the Russian Foundation for Basic Research under Grant No. 94-01-193-a.

[1] A. Davey and K. Stewartson, Proc. R. Soc. London, Ser. A **338**, 101 (1974).
[2] V. D. Djordjevic and L. G. Redekopp, J. Fluid Mech. **79**, 703 (1977).
[3] L. P. Niznik, *Inverse nonstationary problem of scattering theory* (Naukova Dumka, Kiev, 1973).
[4] A. S. Fokas and M. J. Ablowitz, J. Math. Phys. **25**, 2494 (1984).
[5] A. S. Fokas and P. M. Santini, Physica D **44**, 99 (1990).
[6] L. P. Niznik, *Inverse scattering problems for hyperbolic equations* (Naukova Dumka, Kiev, 1991).
[7] A. S. Fokas and L. J. Sung, Inverse Probl. **8**, 673 (1992).
[8] D. J. Kaup, Physica D **1**, 45 (1980).
[9] P. Kulish and V. Lipovskii, Zap. Nauchn. Semin. LOMI **161**, 54 (1987).
[10] B. Champaque and P. Winternitz, J. Math. Phys. **29**, 1 (1988).
[11] M. J. Ablowitz, S. V. Manakov, and C. L. Shultz, Phys. Lett. A **1148**, 50 (1990).
[12] J. Villaroel and M. J. Ablovitz, ''On the Hamiltonian formalism for the Davey–Stewartson system,'' preprint, University of Colorado, PAM #55 (1990).
[13] T. Nakao and M. Wadati, Phys. Lett. A **191**, 77 (1994).
[14] M. Boiti, J. J.-P. Leon, L. Martina, and F. Pempinelli, Phys. Lett. A **132**, 432 (1988).
[15] M. A. Salle, Zap. Nauchn. Semin. LOMI **180**, 161 (1990).
[16] J. Hietarinta and R. Hirota, Phys. Lett. A **145**, 237 (1990).
[17] P. A. Clarkson and S. Hood, J. Math. Phys. **35**, 255 (1994).
[18] D. J. Kaup, J. Math. Anal. Appl. **54**, 849 (1976).
[19] J. R. Keener and D. J. McLaughlin, J. Math. Phys. **18**, 2008 (1977).
[20] V. S. Gergjikov and E. H. Hristov, Mat. Zametki **28**, 501 (1980).
[21] R. L. Sachs, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **14**, 674 (1983).
[22] A. C. Newell, *Solitons in mathematics and physics* (SIAM, Philadelphia, 1985).
[23] L. A. Kalyakin, in *Asymptotic methods to solve differential equations* (Institute of Mathematics of Russian Acad. of Sci., Ufa, 1992), p. 46.
[24] D. J. Kaup, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **31**, 121 (1976).
[25] V. I. Karpman and E. M. Maslov, Dokl. Akad. Nauk SSSR **242**, 581 (1978).
[26] V. P. Maslov and G. A. Omel'anov, Usp. Mat. Nauk **36**, 63 (1981).
[27] Yu. S. Kivshar and B. A. Malomed, Rev. Mod. Phys. **61**, 768 (1989).
[28] L. A. Kalyakin, Teor. Mat. Fiz. **92**, 62 (1992).
[29] O. M. Kiselev, Theor. Math. Phys. **93**, 1106 (1992).
[30] I. M. Krichever, Usp. Mat. Nauk **44**, 121 (1989).
[31] V. A. Arkadied, A. K. Pogrebkov, and M. S. Polivanov, Physica D **36**, 189 (1989).
[32] R. R. Gadyl'shin and O. M. Kiselev, Teor. Mat. Fiz. **106**, 200 (1996).

Wavelets and quantum algebras

A. Ludu
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001

M. Greiner
Institut für Theoretische Physik, Technische Universität, D-01062 Dresden, Germany

J. P. Draayer
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001

(Received 23 October 1995; accepted for publication 12 November 1997)

Wavelets, known to be useful in nonlinear multiscale processes and in multiresolution analysis, are shown to have a q-deformed algebraic structure. The translation and dilation operators associate to any scaling equation a nonlinear, two parameter algebra. This structure can be mapped onto the quantum group $\mathrm{sl}_q(2)$ in one limit, and approaches a Fourier series generating algebra, in another limit. A duality between any scaling function and its corresponding nonlinear algebra is obtained. Examples for the Haar and B-wavelets are worked out in detail. © *1998 American Institute of Physics.* [S0022-2488(98)00703-8]

I. INTRODUCTION

Interest in wavelet theory and its applications in the multiresolution analysis[1] has grown over the last decade, involving new and linking disparate fields of research from pure mathematics and physics to down-to-earth signal engineering. It is being widely applied in signal processing and data compression,[2–4] pattern recognition,[5] statistical physics and turbulence,[6] jet dynamics,[7] field theory,[8,9] solid state physics,[10–12] quantum mechanical applications,[13–16] nonlinear dynamics,[6,7,17] soliton theory,[18,19] etc.

The rise in interest in wavelet theory is motivated by the fact that Fourier analysis is quite ineffective when dealing with nonlinear models and localize sharp features.[6–18] A central property of wavelet theory is the ability to expand and analyze functions with respect to a self-similar set of localized basis functions (scaling functions and wavelets) and to processes them locally, without affecting the scale. The wavelet method is recursive and therefore ideal for computational applications. Moreover, the scaling functions and the corresponding wavelets are very well localized both in the time and frequency domains. Wavelets are the natural tool in the analysis of phenomena where different space/time scales occur. They provide mathematical representations that can handle both analytical and numerical difficulties due to singular phenomena.[6,7] Like Fourier analysis, wavelet theory uses expansion functions with different characteristic scales. However, Fourier analysis has the advantage of being based on a simple and solid algebraic foundation.[20] A challenge is to see if an algebraic framework for the foundation of the wavelet analysis can be constructed.

This paper is a first step in this direction. It is mainly concerned in obtaining a closed algebraic method for the construction of the scaling and wavelet functions. We present results for the Haar and B-wavelets. The purpose of this paper is to show that wavelets are not only multiresolution bases of self-similar functions but they have a definite nonlinear symmetry associated with a q-deformation of the Fourier series generating algebra. In a following second paper we will show that wavelets also have variational properties: all multiscale equations follow from Hamilton's equation of an infinite-dimensional Hamiltonian system.

We also show that the central object of wavelet theory, the two-scale equation, has an algebraic counterpart, arising from a quantum algebra.[21] Applications of this result can be realized using supercomputers and efficient numerical schemes which have led to discrete modeling of complex continuous systems or of genuine discrete physical systems defined on lattices. The

 © 1998 American Institute of Physics

discrete symmetries related to complete solvability[22] or to discrete-continuous transitions,[23] and their basic tool, the finite difference equations, are new and central items in any analysis of discrete systems. On uniform lattices the symmetry algebras of partial differential equations is left unchanged by the discretization.[24] In the case of nonuniform lattices one needs to introduce generalized symmetries, like quantum algebras.[25] This is an example when the q-deformation of some initial symmetries can play an important role. An outcome of the present paper is an extension of the traditional Fourier analysis towards wavelet expansions that can be accomplished by q-deforming the Fourier algebra.

The finite-difference and scaling operators involved in the dilation equation are algebraically closed with respect to certain nonlinear commutation relations. This symmetry is better emphasized when the operators are realized in terms of q-deformed derivatives.[26] By expressing the translation and dilation operators as q-deformed derivatives, the scaling and difference operators can be mapped onto the generators of nonlinear algebra and the action of the q-deformed derivatives is extended to nondifferentiable functions. Moreover, such nonlinear algebras can be undeformed to the Fourier series generating algebra or can be mapped onto the $\mathrm{sl}_q(2)$ algebra.

After devoting Sec. II to definitions and notations, we present in Sec. III the deformation of a Fourier algebra into a scaling function generating algebra, and identify a spectral problem with the Haar dilation equation. In Sec. IV we introduce the definition of the most general scaling function and wavelet algebra. The uniqueness of the algebraic formulation of the dilation equation is proved. A duality relation is obtained between any scaling function and certain nonlinear algebra, which associates to any given scaling function (wavelet) an algebra and, for a given algebra of a certain type, one finds the appropriate wavelet structure. In Sec. V we give examples of scaling function algebras for the Haar and the B-wavetes. We use some new limiting procedures in order to solve the q-difference and finite-difference equations. In Sec. VI we have used a variant and extension of the deforming functional technique,[28,29,31] wherein we can obtain a mapping of the generators of nonlinear algebras to those of $\mathrm{sl}_q(2)$. Finally, Sec. VII comprises concluding comments, further extensions, and remarks.

II. BASIC ELEMENTS

An algebraic structure can be provided for any scaling function system and wavelet basis. In order to realize this construction we need three building blocks: the algebraic structure of the Fourier system (starting point), the wavelet theory (final aim), and q-deformation (intermediate tool).

A. Fourier algebras

In order to demonstrate our approach, we give an example for Fourier series, i.e., algebras having Fourier series as basis for their representation spaces.[20] In the following we shall denote $\partial^k f/\partial x^k=\partial_k f=f^k$. The trigonometric (Fourier) system $\{|k>=e^{ikx}\}_{k\in Z}$ diagonalizes all translation invariant operators acting on $L^2([0,2\pi])$. We introduce three generators within a differential realization $J_0=-i\partial$, $J_\pm=e^{\pm ix}(-i\partial)^p$, satisfying the commutation relations for $p=0$, 1 only

$$[J_0,\ J_\pm]=\pm J_\pm,\quad [J_+,\ J_-]=1-(1-2i\partial)^p. \tag{1}$$

For $p=0$ Eqs. (1) describe an algebra, denoted $\mathscr{T}_0$, that is isomorphic to an analytical prolongation, $e(2,C)$, of the Euclidean algebra $e(2,R)$ of rigid motions in the plane. The algebra $e(2,R)$, generated by $P_x=P_x^\dagger$, $P_y=P_y^\dagger$, and $R=-R^\dagger$, with the commutators $[R,\ P_{x,y}]=\pm P_{y,x}$, $[P_x,\ P_y]=0$, is mapped through the generator mapping

$$J_0=iR,\quad J_\pm=P_x\pm iP_y,$$

onto the $\mathscr{T}_0$ algebra. The unitary irreducible representations (unirreps) of $\mathscr{T}_0$ are based on the eigenvectors of the self-adjoint operator J_0, $J_0|n\rangle=n|n\rangle$. For any two distinct eigenstates, $|n\rangle$ and $|n'\rangle$, by using the first commutator in Eq. (1), we obtain $n'=n+1$ and hence the spectrum of J_0 is unbounded, discrete and consists in equidistant eigenvalues. Hence the space of representations is generated by the Fourier system and the other generators $J_\pm|n\rangle=|n\pm1\rangle$ act like ladder operators on the $|n\rangle$ states, increasing/decreasing the scale by unity. For $p=1$ Eqs. (1) describe another Lie algebra, $\mathscr{T}_1$, isomorphic with the symplectic algebra $\mathrm{sp}(2,R)\simeq\mathrm{su}(1,1)\simeq\mathrm{so}(2,1)$. This differ-

ential realization has the same representation space as $\mathscr{T}_0$ but it is not irreducible. The action of the generators of $\mathscr{T}_1$ is similar to that of the generators of $\mathscr{T}_0$ for $n\neq 0$. Unlike the $\mathscr{T}_0$ case, however, we have $J_{\pm}|\mp 1\rangle=|0\rangle$ and the subspaces $\{e^{inx}\}_{n\in N}$ and $\{e^{-inx}\}_{n\in N}$ are invariant subspaces of $\mathscr{T}_1$. The algebras $\mathscr{T}_{0,1}$ suggest the possibility of other constructions in terms of the operator ∂ and the complex exponential functions of different scales.

B. Scaling functions and wavelets

Wavelets are able to reconstruct a signal through regular sampling, i.e., by analyzing the signal at different scales (which increase/decrease exponentially) with the step size between each scale being the same. Different from the ordinary Fourier transform, which reproduces a function as a superposition of complex exponentials, or from the windowed Fourier transform, which introduces a scale into the analysis of signals, multiresolution analysis processes the signal locally, using the appropriate local scale. Basically, wavelets are constructed with a pair of operators: the dilation (scaling) and finite difference (combinations of translations) operators, acting on $L^2(R)$ and defined by $T^{\alpha}f(x)=f(x+\alpha)$, $\mathscr{D}^{\beta}f(x)=2^{\beta}f(2^{\beta}x)$, with α,β arbitrary real numbers. In the following we will use the operator $D^{\beta}=2^{-\beta}\mathscr{D}^{\beta}$ for simplicity in equations ($\mathscr{D}^{\beta}$ is the dilation operator consistent with wavelet theory). They are invertible, unitary, and fulfill $T^{\alpha}D^{\beta}=D^{\beta}T^{2^{\beta}\alpha}$, $T^{\alpha}T^{\beta}=T^{\alpha+\beta}$, $D^{\alpha}D^{\beta}=D^{\alpha+\beta}$. The formal Taylor series of these operators are $T^{\alpha}=e^{\alpha\partial}$ and $D^{\beta}=2^{\beta}e^{\beta\ln 2x\partial}$. On the space of compact supported or rapidly decreasing functions, any holomorphic function $f(T)$ is locally polynomial. Indeed, if f is holomorphic, and its action is taken on the compact supported function subspace of $\Phi\in L^2(R)$, then the action of $f(T)=\Sigma_{k\in Z}\,C_kT^k$ reduces to that of a Laurent polynomial by keeping only a finite number of terms in the sum.

The wavelet system is given by a set of scaled and translated copies of a pair of functions: the scaling function Φ and the mother wavelet Ψ. The basic fact about wavelets is that both these fundamental functions are finite linear combinations of Φ, reflecting the self-similar character of the wavelet system. The defining equation for the scaling function (the dilation equation) is a linear finite-difference equation, including a scale change (q-difference[27])

$$\Phi(x)=D\sum_{k=0}^{n}C_kT^{-k}\Phi(x)=Dg(T)\Phi(x), \tag{2}$$

where the RHS sum is a polynomial in $T,g(T)$. Equation (2) is a fixed-point equation and consequently it has only one unique solution.[1–7,11,13] The scaling function $\Phi(x)$, as a solution of Eq. (2), is required to have two properties[1–5]

(1) $\int_R\Phi(x)dx=1$, the average value property;
(2) $\langle\Phi_n,\Phi\rangle\equiv\int_R\Phi(x+n)\Phi(x)dx=\delta_{n,0}$, for any $n\in Z$, the orthogonality condition.

These conditions introduce restrictions on the coefficients C_k in Eq. (2),[1–4]

$$\sum_{k=0}^{n}C_k=2,\quad \sum_{k=0}^{n}C_kC_{k+2l}=\delta_{0l},\quad l\in Z. \tag{3}$$

The condition Eq. (3) yield a pattern of $L^2(R)$ as a chain of subspaces $V_j\subset V_{j+1}$, each one being generated by all the translations of $D^j\Phi$, $j\in Z$. By repeated application of D and T on Φ, $D^jT^n\Phi(x)=\Phi(2^jx+n)\equiv\Phi_{j,n}$, one obtains a nonorthogonal basis in $L^2(R)$. The action of the operator $-T^{-\lambda}Dg(-T^{-1})$ on Φ, λ being a unique odd integer, provides the mother wavelet function,

$$\Psi(x)=-T^{-\lambda}Dg(-T^{-1})\Phi(x)=-T^{-\lambda}\sum_k\,(-1)^kC_{-k+1}\Phi(2x-k). \tag{4}$$

The wavelet $\Psi(x)$ has the property that $\{\Psi_{j,n}\}_{j,n\in Z}$ is an orthonormal basis of $L^2(R)$ Refs. 1–3 and $L^2(R)$ is a direct sum of the orthogonal subspaces W_j (the orthogonal complements of V_j), each of them generated by all possible translations of $D^j\Psi=\Psi_{j,0}$ with integral m:$L^2(R)$

$=\oplus_{j\in Z}W_j$.[1–5] Hence translated and dilated copies of the mother wavelet, $\Psi(2^jx+k)$, generate a true orthonormal basis. Recursively applications of the Eqs. (2) and (4) relate all the scaling functions and wavelets.

The simplest example is provided by the Haar wavelet, defined by the scaling function $\Phi_{\text{Haar}}(x)=1$ if $|x-1/2|\leqslant 1/2$ and 0 otherwise. In this case we have $D(1+T^{-1})\Phi_{\text{Haar}}=\Phi_{\text{Haar}}$ and $g_{\text{Haar}}(T)=1+T^{-1}$ ($C_0=C_{-1}=1$ and the rest 0 in Eq. (2)). The corresponding Haar wavelet is $\Psi_{\text{Haar}}=D(1-T^{-1})\Phi_{\text{Haar}}$, i.e., $\lambda=1$ in Eq. (4).

There are similarities and differences between the Fourier and the wavelet approaches. From an algebraic point of view, in both approaches, there are eigenfunction equations, which keep the scale constant, and the ladder operators, which change the scale. The dilation equation, Eq. (2), has its analog in the Fourier formalism, though it is a fixed-point equation and has only one solution. The mother wavelet, Eq. (4), has no analog in Fourier analysis. Each Fourier eigenfunction carries one scale; a wavelet function $\Psi(x)$ involves two scales, e.g., $\Phi_{0,k}$ and $\Phi_{2,k}$. Because of their localization, wavelets possess a degree of freedom beyond that of a Fourier system; namely, the width of the support of Φ. In the present approach this degree of freedom is associated with the deformation parameter, q.

C. Quantum deformations

A possibility for constructing scaling functions/wavelets within an algebraic approach is to deform the Fourier algebraic structure into a nonlinear system. In general, the scaling functions and wavelets are not differentiable. This suggests the use of finite-difference operators instead of derivatives. Finite-difference operators are closed with respect to commutation but only within nonlinear algebraic constructions. Consequently, it is natural to use q-deformed derivatives to find a foundation for wavelet theory in q deformed algebras.

Quantum algebras refer to some specific deformations of Lie algebras, to which they reduce when the deformation parameter q is set equal to unity (for a recent monograph see Ref. 21, and references therein). The simplest example of a q-algebra is $\text{sl}_q(2)$ whose Jordan–Schwinger realization is given in terms of q-bosonic operators.[26] The q-deformed algebras have been applied in various branches of physics like spin-chain models, noncommutative spaces, rotational spectra of deformed nuclei, Hamiltonian quantization, and dynamical symmetry breaking. The basic element is the q-deformation of a certain object x, which can be a number, an operator, or a function,

$$[x]_s=\frac{q^x-q^{-x}}{q-q^{-1}}\xrightarrow{s\to 0}x, \tag{5}$$

where $q=e^s$. The Taylor expansion of T and D are related to the q-deformation of the derivative operator, according to the definition of the coordinate description of the q-deformed oscillator introduced in Ref. 26,

$$[\partial]_s f=\frac{T^s-T^{-s}}{2\sinh(s)}f(x). \tag{6}$$

Analogously, we introduce the operator

$$[x\partial]_{s\ln 2}f(x)=\frac{f(2^sx)-f(2^{-s}x)}{2\sinh(s\ln 2)}=\frac{D^s-D^{-s}}{2\sinh(s\ln 2)}f(x). \tag{7}$$

When $s\to 0$, $[\partial]_s\to\partial$ and $[x\partial]_{s\ln 2}\to x\partial$, Eqs. (6) and (7) can be inverted and the occurrence of the q-deformed derivative and the translation/dilation operators becomes immediate

$$T^{\pm s}=\frac{1}{2}\left(\pm\frac{[\partial]_s}{\eta(s)}+\frac{[\partial]^2_{s/2}}{\eta^2(s/2)}+2\right), \tag{8}$$

$$D^{\pm s}=\frac{1}{2}\left(\pm\frac{[x\partial]_s}{\eta(s\ln 2)}+\frac{[x\partial]^2_{s/2}}{\eta^2(s\ln 2/2)}+2\right), \tag{9}$$

where $s \in N$ (or more general $\in R$) and $\eta(s)=1/(e^s-e^{-s})$. The q-deformed algebra $\mathrm{sl}_q(2)$, generated by $\{J_0, J_\pm\}$, is defined by the q-deformed version of the commutation relation

$$[J_+, J_-]=[2J_0]_s,$$

while the other two commutator relations remain undeformed, as for with $\mathrm{sl}(2)$. In addition to this traditional version of $\mathrm{sl}_q(2)$, several generalized forms have been introduced, in two different ways. First, by deforming only the commutator $[J_+, J_-]$ by using some arbitrary function $F(J_0,q)$, $[J_+, J_-]=F(J_0,q)$ independently proposed in Refs. 28, 29, and 30. The second way is by deformations involving all three commutation relations, using two functions $G(J_0,q)$ and $F(J_0,q)=[J_+, J_-]$, introduced in Ref. 31. Unlike the former, for which the spectrum of J_0 is linear, the latter is characterized by an exponential spectrum for J_0. Since in wavelet theory the domains of analysis are divided exponentially, rather than linearly with bands of equal widths, such algebras are used in the following analyses.

III. HAAR SCALING FUNCTION ALGEBRA, $\mathscr{A}_{s,\alpha}$

The aim in this section is to obtain operators depending on D and T, closed under some nonlinear algebra. This structure must provide an algebraic form for Eqs. (2), (4) and must be mapped into one Fourier algebra, $\mathscr{T}_{0,1}$. Consider an operator depending on T and two real parameters s, α,

$$W_0(s,\alpha)=\frac{T^{s\alpha}-T^{-s\alpha}\cos s\pi}{2\xi(\alpha)\sinh s}, \tag{10}$$

where $\xi(\alpha)=[1/\sinh(1)]\sin(\alpha\pi/2)-2i\cos(\alpha\pi/2)$. The (s,α) parameters allow W_0 to approach ∂ or combinations of T,

$$W_0(0,\alpha)=\frac{\alpha}{\xi(\alpha)}\,\partial, \tag{11}$$

$$W_0\left(\frac{1}{2},\alpha\right)=\frac{1}{2\sinh(1/2)\xi(\alpha)}\,T^{\alpha/2}, \tag{12}$$

$$W_0(1,\alpha)=\frac{T^{\alpha}+T^{-\alpha}}{2\sinh(1)\xi(\alpha)}, \tag{13}$$

$$W_0(2,\alpha)=\frac{T^{2\alpha}-T^{-2\alpha}}{2\sinh(2)\xi(\alpha)}. \tag{14}$$

In the limit $s\to 0$, W_0 reduces to the normal derivative with respect to x, Eq. (11). In the limit $s=\frac{1}{2}$, we obtain a power of the translation operator T, Eq. (12). Equation (13) defines a q-deformation of unity, i.e., $W_0(1,0)=[i/4\sinh(1)]$. $W_0(s,\alpha)$ can be a Hermitian or anti-Hermitian operator, $W_0(1,\alpha)=\pm W_0(1,\alpha)^\dagger$, depending on α. The last limit, Eq. (14), represents a finite difference operator and is proportional to the q-derivative with respect to the α-deformation $[2\partial]_\alpha$, similar to Eqs. (6)–(9). For $\alpha=2k$ in Eq. (12), $\alpha=k$ in Eq. (13) and $\alpha=k/2$ in Eq. (14), the corresponding $W_0(s,\alpha)$ are Laurent polynomials in T, providing a direct connection to the wavelet operators. For some values of α, Eqs. (12)–(14) give invertible operators with respect to $T, T=T(W_0)$.

In order to introduce the dilation operator D, we define

$$W_\pm(s)=\tfrac{1}{2}e^{\mp ix(s-1)/2}D^{\mp s}e^{\mp ix(s-1)/2}(1+T^{\pm s}). \tag{15}$$

In order to fulfil the Hermiticity condition $(W_+)^\dagger=W_-$ on $L^2(R)$ these generators can be redefined in the form

$$W_\pm\to\widetilde{W}_\pm=2^{-1\mp,s/2}e^{\mp ix(s-1)/2}D^{\mp s}e^{\mp ix(s-1)/2}(1+e^{-i[s(s-1)(1+2^s)]/4}T^{\pm s(2^{(s/2)(1\mp 1)}}.$$

By using the commutators relations between D and T and by performing an integration by parts, one has a direct check of the relation $(\widetilde{W}_+)^\dagger=\widetilde{W}_-$,

$$\langle f_1,\widetilde{W}_+ f_2\rangle=\int_R f_1^*(x)\widetilde{W}_+ f_2(x)dx=\langle \widetilde{W}_- f_1,f_2\rangle.$$

From Eqs. (10) and (15) one obtains the commutators,

$$[W_0(s,\alpha),\ W_+(s)]=G(s,T)W_+(s), \tag{16}$$

$$[W_0(s,\alpha),\ W_-(s)]=-W_-(s)G(s,T), \tag{17}$$

$$[W_+(s),\ W_-(s)]=F(s,T), \tag{18}$$

with

$$G(s,T)=W_0(s,\alpha)-\frac{T^{2^s s\alpha}e^{is\alpha(1+2^s)(s-1)/2}-T^{-2^s s\alpha}e^{-is\alpha(1+2^s)(s-1)/2}\cos s\pi}{2\sinh(s)\xi(\alpha)} \tag{19}$$

and

$$F(s,T)=\frac{1}{4}\left((1+e^{is(1+2^s)(s-1)/2}T^{2^s s})(1+T^{-s})-(1+e^{is(1+2^{-s})(s-1)/2}(T^{-2^{-s}s})(1+T^s)\right). \tag{20}$$

Equations (10), (15)–(20) describe a nonlinear associative algebra, generated by $T^\pm, D^\pm$, denoted $\mathscr{A}_{s,\alpha}$. The eigenvalue problem for $W_\pm$ provides the algebraic form for the scaling equation. When $W_0(s,\alpha)$ is invertible with respect to T, the algebra has $W_0, W_\pm$ as generators which depend on two parameters (s,α) and is homeomorphic with a special q-deformed algebra, namely, the two-color quasitriangular Hopf algebra $\mathscr{A}^\pm$.[31] The spectrum of W_0 is not equidistant because $F(s,T)\neq \text{const}\times W_0$. The Casimir operator of the algebra $\mathscr{A}_{s,\alpha}$ is given by[28–31]

$$C=W_-W_++H(W_0)=W_+W_-+H(W_0)-F(W_0), \tag{21}$$

where H is a real function, holomorphic in a neighborhood of 0 and which must satisfy the functional equation

$$H(\xi)-H(\xi-G(\xi))=F(\xi) \tag{22}$$

with ξ generic. The irreps of $\mathscr{A}_{s,\alpha}$ are labeled by the eigenvalues a of W_0 and those of the Casimir operator. The commutator in Eq. (16) can be written in the form

$$W_+W_0=(W_0-G)W_+. \tag{23}$$

If $|a\langle\neq|a'\rangle$ are eigenvectors of W_0, with the corresponding eigenvalues a,a', we have from Eq. (23) a nonlinear recursion relation for all the eigenvalues

$$a'=a-G(a). \tag{24}$$

The algebra $\mathscr{A}_{s,\alpha}$ can be mapped into $e(2,C)$, the Lie algebra of a Fourier series, in the limit $s\to 0$, $\alpha\to 2$: $W_\pm\to e^{\pm ix}=J_\pm$, $W_0\to -i\partial=J_0$, where we take these limits in the order $\mathscr{A}_{s_0,\alpha_0}=\lim_{\alpha\to\alpha_0}\lim_{s\to s_0}\mathscr{A}_{s,\alpha}$. The continuous mapping $\mathscr{A}_{s,\alpha}\to\mathscr{A}_{0,2}\simeq e(2,C)$ is an algebraic morphism and one can consider wavelets as q-deformed generalizations of the Fourier series, in the above sense.

As a first example, the Haar scaling function can be associated with a particular case of $\mathscr{A}_{s,\alpha}$, i.e., the algebra $\mathscr{A}_{1,1}$. In the algebra $\mathscr{A}_{1,1}$ the eigenproblem for W_- provides the dilation equation for the Haar scaling function $\Phi(x)$ defined by Eq. (2) with $g(T)=1+T^{-1}$. We have in this case $G=W_0-2W_0^2+1$, and $W_\pm=\frac{1}{2}D^{\mp 1}(1+T^{\pm 1})$ and the commutators,

$$[W_0,\ W_+]=(W_0-2W_0^2+1)W_+,$$

$$[W_0,\ W_-]=-W_-(W_0-2W_0^2+1), \tag{25}$$

$$[W_+,\ W_-]=\tfrac{1}{4}(T^2(W_0)-T^{1/2}(W_0)-T^{-1/2}(W_0)+T^{-1}(W_0)),$$

where $T(W_0)$ is the solution of the equation $2W_0=T+T^{-1}$, a pseudodifferential operator. The Casimir operator of $\mathscr{A}_{1,1}$ is a constant. Indeed, Eq. (22) for the general Casimir operator becomes

$$H(W_0)-H(2W_0^2-1)=F(W_0), \tag{26}$$

and has, in terms of the $T=T(W_0)$ operator, the form

$$\chi(T)-\chi(T^2)=\frac{1}{4}\,(T^2-T^{1/2}-T^{-1/2}+T^{-1}), \tag{27}$$

where $\chi(T)=H[(T+T^{-1})/2]$. The unique analytical solution for Eq. (27) reads

$$\chi(T)=\tfrac{1}{4}(\text{const}-T-T^{1/2}-T^{-1/2}), \tag{28}$$

which gives for the Casimir operator a constant.

The spectrum of W_0 consists of periodic functions satisfying the equation $\Phi(x+1)+\Phi(x-1)=2\Phi(x)$. This takes one back to Fourier analysis. The way to the Haar scaling function is to use the spectrum of $W_\pm$ since in $\mathscr{A}_{1,1}$ the eigenproblem for W_- gives the dilation equation for the Haar scaling function $\Phi(x)$. Once Φ is obtained from $2W_-\Phi=\Phi$ the generator W_0 carries Φ into an infinite sequence of functions $W_0^n\Phi$. These functions are mutual orthogonal and generate the space V_0. The action of $W_\pm$ on Φ yields the subsequence $W_0^{2n}\Phi$. Consequently, V_0 is an invariant space for all the generators of the algebra $\mathscr{A}_{1,1}$.

The spectrum of W_0 depends on the values of the parameters s,α. In the case of $\mathscr{A}_{1,1}$ this relation is $a'=2a^2-1$. The corresponding representations are infinite-dimensional and the eigenvalues of $W_0|_{1,1}$ are $a_n=\cosh(2^n\tilde{a})$, $n\in Z$ for any $\tilde{a}$. The corresponding eigenfunctions have the form $|a_n\rangle_\pm=e^{\pm 2^n\tilde{a}x}$ which results in an exponential spectrum for a_n similar to the sequence of scales in wavelet theory. This is a self-similar spectrum with respect to $\tilde{a}$, like the sequence of scales in wavelet theory. A part of this spectrum is shown in Fig. 1. The action of $W_\pm$ is

$$W_\pm|a_n\rangle_\pm=\frac{1+e^{\pm 2^n\tilde{a}}}{2}\,|a_{n\pm 1}\rangle_\pm. \tag{29}$$

Another basis for a representation of $\mathscr{A}_{1,1}$ is given by $\underline{|k\rangle}=e(i\pi x/k)$, $k\in Z$. We have the action

$$W_0\underline{|k\rangle}=\cos\frac{\pi}{k}\,\underline{|k\rangle},\quad W_\pm\underline{|k\rangle}=\frac{1}{2}\,(1+e^{\pm i\pi/k})\underline{|2^{\pm 1}k\rangle},$$

and on this basis $(W_+)^\dagger=W_-$. There are two invariant spaces, $\underline{|2^k\rangle}$ and $\underline{|2^{-k}\rangle}$, $k\in N$, with $W_\pm\underline{|0\rangle}=1$.

The same procedure can be followed for any set of the parameters s,α. For example, if we choose the algebra $\mathscr{A}_{2,1/2}$ with W_0 defined in Eq. (12), we obtain the spectrum of W_0 described by the recursion relation

$$a'=a\left(a+\sqrt{a^2+1}+\frac{1}{a+\sqrt{a^2+1}}\right), \tag{30}$$

which also gives a nonlinear, unbounded, discrete representation of $\mathscr{A}_{2,1/2}$.

The last thing to prove is the uniqueness of the two-scale equation in the $\mathscr{A}_{1,1}$ scaling function generating algebra. The two-scale equation, Eq. (2), in its algebraic form $2W_-\Phi=\Phi$ is not unique

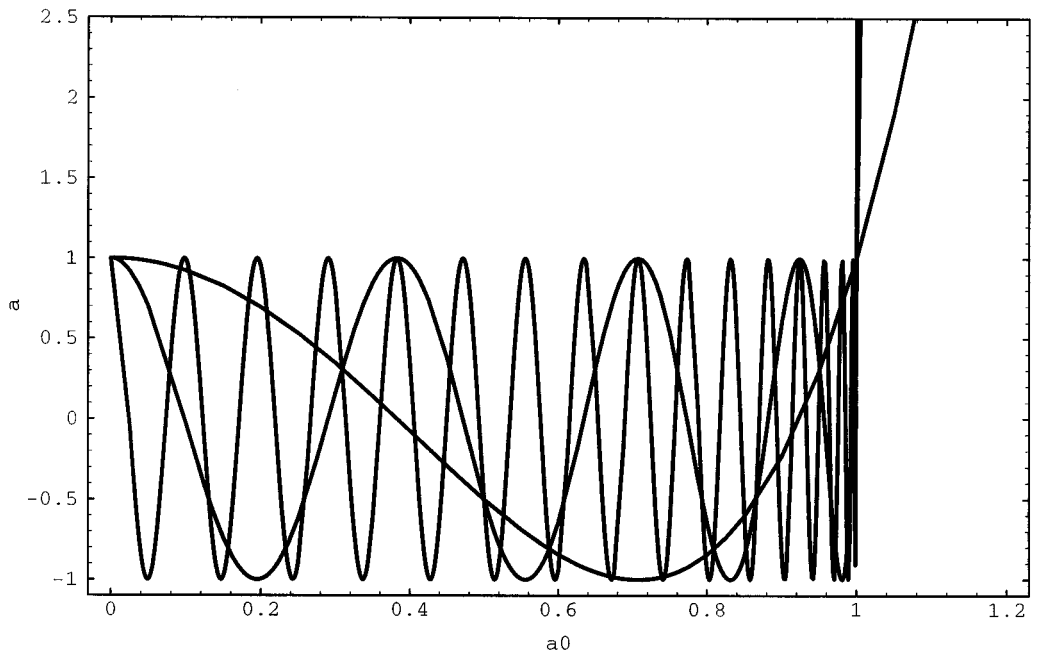

FIG. 1. The dependence on a_0 of three eigenvalues a_n of W_0 in $\mathscr{A}_{1,1}$, for $n=2$ (the largest period), 4 and 6 (the smallest period). The three spectra have self-similar structure for $a_0 \in [0, 1]$.

in $\mathscr{A}_{1,1}$ if there exists an operator, similar with that occurring in the dilation equation, which commutes with W_-. This operator should contain higher powers in T and consequently is an element of $U(\mathscr{A}_{1,1})$.

Proposition 1: In $U(\mathscr{A}_{1,1})$ *there exists a unique operator* $X=Dx(T^{-1})$ *such that* $[W_-, X]=0$. *The function* $x(\xi)$ *is an integer, it is not a polynomial and it is unbounded, for generic* ξ.

Proof: We take for X a Laurent series $x(T^{-1})=\Sigma_{k\in Z} C_k T^{-k}$ then the condition $[W_-, X]=0$ results in a recursion relation for the coefficients C_k,

$$C_{2k+1}=C_k-C_{2k-1}, \quad C_{2k}=C_k-C_{2k-2}, \tag{31}$$

for any $k\in Z$. Equations (31) have one trivial solution which reproduces W_-, $x(T^{-1})=C_{-1}(1+T^{-1})$ and only one different solution, with $C_k=\pm C_{-1}$ or 0, uniquely defined ($C_2=C_{\pm 3}=C_{-4}=C_5=C_{-6}=C_7=C_{-8}=C_{-11}\cdots=0$, $C_{-2}=C_{-5}=C_6=C_{-9}=C_{-10}=C_{10}=C_{11}=\cdots=C_{-1}$ and $C_1=C_4=C_{-7}=C_8=C_9=C_{-12}=C_{12}=\cdots-C_{-1}$, etc.). The sequence of nonzero coefficients is infinite, and $\{C_k\}$ is not a Cauchy sequence. (q.e.d.) We note that the $\mathscr{A}_{1,1}$ algebra is also a Hopf algebra,[21,30,31] defined by Eqs. (25) and by the coproduct, counit and antipode in the form,

$$\Delta T^{\pm}=T^{\pm}\otimes T^{\pm}, \quad \Delta D^{\pm}=D^{\pm}\otimes D^{\pm}, \quad \epsilon(T^{\pm})=\epsilon(D^{\pm})=1,$$

$$S(T^{\pm})=T^{\mp}, \quad S(D^{\pm})=D^{\mp}.$$

All its generators are primitive elements.

IV. GENERAL SCALING AND WAVELET ALGEBRA

This algebraic approach for the Haar scaling function can be generalized to yield a nonlinear algebra for any scaling function, and conversely, to find the dilation equation and scaling function for a certain type of algebra. The procedure starts with an algebra generated by dilation and translation and constructs, within this algebra, the two-scale equation. The generators of the $\mathscr{A}_{s,\alpha}$ algebra can be generalized as

$$W_0 \to j_0(T), \quad W_{\pm} \to j_{\pm}(D,T)=e^{\mp ix(s-1)/2} D^{\mp s} e^{\mp ix(s-1)/2} j(T^{\pm 1}), \tag{32}$$

with $j_0(T)$ and $j(T)$ being arbitrary functions of T and s, holomorphic in a neighborhood of $T=1$, with their dependence on s being such that in the limit $s\to 0$, $j_0(T)\to -i\partial$, $j(T)\to 1$. In this case the commutation relations, Eqs. (16)–(20), become

$$[j_0, j_+]=\widetilde{G}j_+, \quad [j_0, j_-]=-j_-\widetilde{G}, \quad [j_+, j_-]=\widetilde{F}, \tag{33}$$

with $\widetilde{G},\widetilde{F}$ depending on T through j_0, j, respectively,

$$\widetilde{G}=\widetilde{G}(s,T)=j_0(s,T)-j_0(s,T^{2^s}e^{i(1+2^s)(s-1)/2}), \tag{34}$$

$$\widetilde{F}=\widetilde{F}(s,T)=j(s,e^{i(s-1)/2(1+2^s)}T^{2^s})j(s,T^{-1})-j(s,T^{-2^{-s}}e^{i(s-1)/2(1+2^{-s})})j(s,T). \tag{35}$$

Equations (33)–(35) define a nonlinear algebra denoted $\mathscr{A}_{j_0,j}$ as a generalization of $\mathscr{A}_{s,\alpha}$. If the function $j(T)$ is invertible with respect to T, $\mathscr{A}_{j_0,j}$ can be expressed in terms of the generators $j_0, j_\pm$ only and then $\widetilde{G}(T)=\mathscr{G}(j_0)$, $\widetilde{F}(T)=\mathscr{F}(j_0)$. The algebraic morphism $j_0(T)\to W_0$ and $j(T)\to 1+T^s$ provides the mapping $\mathscr{A}_{j_0,j}\to\mathscr{A}_{s,\alpha}$. Since the first two commutators of $\mathscr{A}_{j_0,j}$, Eq. (33), do not depend on the function $j(T)$, and the third commutator in Eq. (33) does not depend on the function $j_0(T)$, the closure condition for $\mathscr{A}_{j_0,j}$ is independent of the functions $j_0(T)$ and $j(T)$ and hence they can be chosen in a convenient way to provide any dilation equation in the form of the eigenproblem for j_-. The algebraic closure conditions for $\mathscr{A}_{j_0,j}$, together with the definitions of j, j_0 require

$$j_0(T^2)-j_0(T)=\mathscr{G}(j_0(T^2)), \tag{36}$$

$$j(T^2)j(T^{-1})-j(T^{-1/2})j(T)=\mathscr{F}(j_0(T)). \tag{37}$$

Both Eqs. (36) and (37) are nonlinear, and in general, difficult to solve analytically. The unique solution for the dilation equation are $j_\pm$ since the Casimir operator of this algebra depends on T only. Choosing $g(T)=j(T)$ yields the dilation equation in the form $2j_-\Phi=\Phi$. In the limit of $\mathscr{A}_{s,\alpha}$ this provides again the Haar two-scale equation. The corresponding scaling function belongs to the basis of the representation of j_- with eigenvalue 1/2. The remaining arbitrary function j_0 can now be selected to obtain the wavelet generating operator $j_-j_0\Phi=\Psi$, in agreement with Eq. (4). One can formally construct a basis of the representation for j_- with the functions $|n\rangle=(\ln T)^n\Phi$. The procedure is the following: for a given algebra $\mathscr{A}_{j_0 j}$ (therefore given functions $\mathscr{G}$, $\mathscr{F}$) solve Eqs. (36) and (37) with respect to the functions $j_0, j_\pm$ and obtain the corresponding dilation equation in the form $2j_-\Phi=\Phi$. A certain combination of generators produces the corresponding wavelet equation for Ψ. Conversely, given a scaling/wavelet function, and consequently its dilation equation and $g(T)$, one can choose $j_-=Dg(T)$ and then solves the equation $Dg(-T^{-1})=j_-j_0$ with respect to j_0. With $j_0, j_\pm$ known one can solve Eqs. (36) and (37) with respect to $\mathscr{G}$, $\mathscr{F}$ to construct the algebra.

This algebraic approach is in some sense universal, what is modified is the specific realization of the nonlinear algebra in terms of D and T. Constraints are imposed by the two limiting approaches (two-scale equation and Fourier limit). Loosely speaking, the closure of the algebra provides the fixed-point dilation equation and the nonlinearity of the algebra provides the exponential scaling.

We also note that $V_0=\{\Sigma_k C_k T^k\Phi\}$, i.e., the space generated by all the integer translations of Φ, and any other $V_j=D^jV_0$, is not invariant to the action of j_-. For example, the condition $j_-D\Phi=\Sigma_k C_k T^k D\Phi$ implies the existence of an operator containing D, whose commutator with j_- is a function of T only. From the definition of the algebra this is impossible, and consequently this proves the above conjecture. The V_j spaces form, by recursion, a basis in $L^2(R)$.

In order to prove the uniqueness of the two-scale equation in the general case, we again use Proposition 1, for $W_-\to j_-=Dj(T^{-1})=D\Sigma_{k\in Z}\, j_k T^k$. We have to solve the commutation equation $[Dj(T^{-1}), X)]=0$ for $x(T)=\Sigma_{k\in Z}X_kT^k$, $X=Dx(T)$. This implies that the arbitrary functions $x(\xi)$ and $j(\xi)$ must fulfill the functional condition (ξ a generic variable)

$$x(\xi^2)j(\xi)=x(\xi)j(\xi^2). \tag{38}$$

Equation (38) does not carry any restriction with respect to the values of the functions in 0 and 1. Since $Dj(T)$ represents a scaling operator we have, according to the dilation equation, $j(1)=1$, $j(-1)=0$ which implies $x(-1)=0$. If Φ is the corresponding scaling function for j_- then $X\Phi$ is also an eigenfunction of j_-, $j_-(X\Phi)=X\Phi$. In order for $Dx(T)$ to satisfy the average value property (1) in Sec. II B, we impose the additional condition $x(1)=1$. The last condition requires $DX(T)$, the orthogonality condition (2) of Sec. II B, be equivalent to the condition[1–4]

$$|x(T)|^2+|x(-T)|^2=1. \tag{39}$$

By introducing Eq. (38) into Eq. (39) we obtain $X=j_-$ which provides the trivial indentity solution, and hence the uniqueness.

V. CONSTRUCTION OF THE SCALING AND WAVELET ALGEBRA

The link between the scaling function and wavelet, and the corresponding scaling algebra is supported by the solutions of Eqs. (36) and (37). In the following we present a method for solving these equations, from the wavelets towards the algebra. The dilation equation Eq. (2) has the form of a fixed-point equation. Therefore we must try to find its eigenvectors as limits of some functional sequences. The limits of these sequences should be compact supported or rapidly decreasing functions for two reasons; to obtain scaling functions with good localization, and to provide correct behavior of the action of the operators $f(T)$. Since the scaling function will be expressed as a limit of a sequence, one has to look for operators which commute with the limit (are absorbed in the limit). We choose a test function $\Delta_0(x)$ and a sequence of operators f_n from the universal covering $U(\mathcal{A}_{j_0,j})$ of the scaling function algebra introduced in the preceding section, such that the limit $\Phi(x)=\lim_{n\to\infty} f_n\Delta_0(x)$ exists in the weak topology and it provides the scaling function.

In the following we show how to construct the scaling algebra $\mathcal{A}_{j_0,j}$ from the dilation equation, written in the form $j_-(T,D,s)\Phi=\Phi$. We have the following proposition in the framework of the algebra $\mathcal{A}_{j_0,j}$, for $s=1$.

Proposition 2: Let $f_n(j_0)$ be a functional sequence in $U(\mathcal{A}_{j_0,j})$, $s=1$, $\Delta_0(x)$ a (test) function and $\Phi(x)$ the scaling function. If the limit $\Phi(x)=\lim_{n\to\infty} f_n(j_0)j_-^n\Delta_0(x)$ exists, and $\lim_{n\to\infty}[f_n(j_0-\mathcal{G}(s,j_0))/f_{n-1}(j_0)]=1$, then $j_-\Phi(x)=\Phi(x)$.

Proof: From the RHS of the first condition in the hypothesis and from Eq. (33) we have

$$f_n(j_0)j_-^n=j_-f_n(j_0-\mathcal{G}(s,j_0))j_-^{n-1}=j_-\frac{f_n(j_0-\mathcal{G}(s,j_0))}{f_{n-1}(j_s)}f_{n-1}(j_0)j_-^{n-1}.$$

It follows

$$j_-\Phi=\lim_{n\to\infty} j_-f_n(j_0)\Delta_0(x)=\Phi,$$

q.e.d. It follows from Proposition 2 that for a given dilation equation $j_-\Phi=\Phi$, (both j_- and Φ given) we can find out a sequence $f_n(j_0)$, the operator $\mathcal{G}(s,j_0)$ and a test function $\Delta_0(x)$, such that these objects satisfy the hypothesis of Proposition 2. Then it follows that one can construct the scaling algebra, since with j_0 and $\mathcal{G}(s,j_0)$ found, $\mathcal{F}(s,j_0)$ results from the last commutator in Eq. (33). In general, one starts with j_0 as an arbitrary function of T which can be deformed into $-i\partial$ when mapping $\mathcal{A}_{j_0,j}$ to $\mathcal{T}_{0,1}$. As j_- is provided by the two-scale equation, $j_-\Phi=\Phi$, other solutions are forbidden. If $\lim_{n\to\infty} f_n(j_0)=f_\infty(j_0)\neq$ const., this limit should commute with j_- which is forbidden by Proposition 1. This procedure closes the construction of the algebra $\mathcal{A}_{j_0,j}$.

In the following we give an algorithm for finding $f_n(j_0)$ and $j_0(T)$ and illustrate it with two examples. Since any mother scaling function is defined by the dilation equation ($Dh(T)\Phi=\Phi$ or $j_-\Phi=\Phi$) it is natural to search for solutions by using a recursion algorithm. For instance, we can find a test function $\Delta_0(x)$ and a triple of operators A,B,C such that the limit $\Phi(x)=B\lim_{n\to\infty} A^n\Delta_0(x)$ exists and in addition $CB=BA^k$, for a finite positive integer k. Then, we have the property $C\Phi(x)=\Phi(x)$. Indeed, $C\Phi=CB\lim_{n\to\infty} A^n\Delta_0=B\lim_{n\to\infty} A^{n+k}\Delta_0=B\lim_{n\to\infty} A^n\Delta_0=\Phi$. From the dilation equation we know that C should have the form C

$=Dc(T)$ with $c(T)$ a function of T. One of the simplest choices is to use $A=D$ and $B=b(T)$. Then we have $CB=Dc(T)b(T)=c(T^{1/2})b(T^{1/2})D=b(T)A$; that is, $k=1$ and a restriction for the arbitrary functions $c(T)$ and $b(T)$ arises,

$$c(T^{1/2})b(T^{1/2})=b(T). \tag{40}$$

This equation is useful in both directions (algebraic↔dilation equation), since one can start with a given scaling function (given $c(T)$) and find the operator $b(T)$ involved in the algebra and conversely.

Now consider the algorithm for the Haar scaling function, $c(T)=1+T^{-1}$. We look for solutions of Eq. (40) as Laurent series for $b(T)=\Sigma_{k\in Z}b_kT^k$. In this case Eq. (40) has an unique solution, $b(T)=\text{const.}\cdot(1-T^{-1})$. This gives again a unique solution for $\Phi=(1-T^{-1})\lim_{n\to\infty}D^n\Delta_0$. We stress that $Db(T)$ is exactly the operator which gives the Haar wavelet. Further, we can express Φ_{Haar} in the form $\Phi(x)=H(x)-H(x-1)$, where $H(x)$ is the Heaviside distribution. For any sequence of C^∞ functions $\delta_n(x)\to\delta(x)$, we have $\Delta_n(x)=\int\delta_n(x)dx\to H(x)$ and we can write

$$\Phi(x)=\lim_{n\to\infty}(\Delta_n(x)-\Delta_n(x-1)). \tag{41}$$

For example, we can use the sequences $\delta_n=(1/\pi)[2^{-n}/(x^2+2^{-2n})]$ or $\delta_n=[2^{n-1}/\cosh^2(2^nx)]$ and $\Delta_n=[\arctan(2^nx)/\pi]$ or $\Delta_n=\frac{1}{2}\tanh(2^nx)$, respectively. The latter example is a soliton-like shape, having good localization. These sequences converge to $\delta(x)$, respectively. In this way we have selected subsequences that step in powers of 2. Equation (41) can be written as

$$\Phi(x)=\lim_{n\to\infty}(1-T^{-1})D^n\Delta_0(x). \tag{42}$$

We can express this definition in terms of the algebra $\mathscr{A}_{1,1}$. By using Eq. (25) and the properties of the operators D and T, we can write Eq. (42) in the form

$$\Phi(x)=\lim_{n\to\infty}(1-T^{-2^{-n}})(2W_-)^n\Delta_0(x). \tag{43}$$

Indeed, from the commutation relation between D and T and the definition of W_- we have

$$\begin{aligned}(2W_-)^n&=(D(1+T^{-1}))^n=D^n(1+T^{-1}+\cdots+T^{-2^n+1})\\&=(1+T^{-2^{-n}}+(T^{-2^{-n}})^2+\cdots+(T^{-2^{-n}})^{2^n-1})D^n.\end{aligned} \tag{44}$$

Hence, we can write $(1-T^{-1})D^n=(1-T^{-2^{-n}})(2W_-)^n$ Moreover, we can write, by using the inverted form for $W_-(T)$, the dilation equation in a pure algebraic form,

$$\Phi(x)=\lim_{n\to\infty}(1-T^{-2^{-n}}(W_0))(2W_-)^n\Delta_0(x). \tag{45}$$

From this last equation it follows that $2W_-\Phi=\Phi$, i.e., the Haar dilation equation. If the function $\Delta_0(x)$ is chosen from a class of suitable functions for a wavelet analyzis then its scaling function, $\Phi(x)$, Eq. (45), gives a rapidly convergent wavelet expansion. In order to exemplify we show in Fig. 2 scaling functions obtained by this procedure for different values of the parameter s within $\mathscr{A}_{s,\alpha}$. For $s=0$ and 1 this functions yields the Haar scaling function.

We give another example for the Φ_2 B-scaling function and the Ψ_2 B-wavelet. We define

$$\Phi_2(x)=\begin{cases}0 & x\leqslant 0,\ x\geqslant 1\\ 4x & 0\leqslant x\leqslant 1/2\\ -4x+4 & 1/2\leqslant x\leqslant 1\end{cases}, \tag{46}$$

or

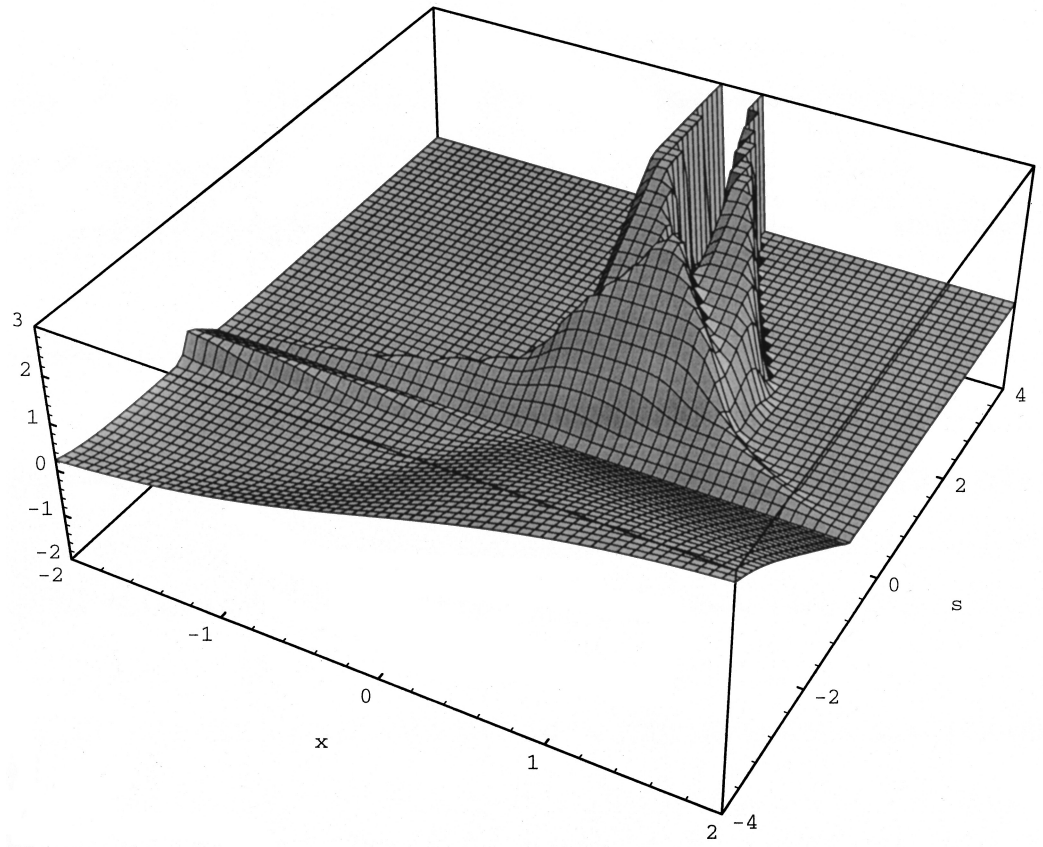

FIG. 2. The continuous deformation of the scaling function Φ in $\mathscr{A}_{s,\alpha}$ as function of the parameter s. For $s=0,1$ this is the Haar scaling function.

$$\Phi_2(x)=\lim_{n\to\infty}\frac{1}{2^{-n+1}}\,(1-2T^{-1/2}+T^{-1})D^n\Delta_{0,\mathrm{tri}}(x), \tag{47}$$

with

$$\Delta_{0,\mathrm{tri}}(x)=\frac{2x}{\pi}\arctan(2x)-\frac{\ln(1+4x^2)}{2\pi}=\int_0^{2\pi}\Delta_0(x)dx. \tag{48}$$

Following the above algorithm we obtain the corresponding dilation equation,

$$j_-|_{\mathrm{tri}}\Phi_2(x)=\frac{D}{2}\,(1+T^{-1/2})^2\Phi_2(x). \tag{49}$$

By using the commutation relation $\partial D=2D\partial$ and the dilation equation Eq. (49) we can obtain a self-contained equation for the wavelet,

$$\Psi_{\mathrm{tri}}(x)=D(1+2T^{-1/2}+T^2)^2\Psi_{\mathrm{tri}}. \tag{50}$$

The corresponding algebras for Haar and Φ_2 B-scaling functions are different. For the B-wavelet we have, in $\mathscr{A}_{1,1}$,

$$G(W_0)=W_0-W_0^2-\tfrac{1}{2}. \tag{51}$$

This new algebra, fulfills different commutation relations and consequently has a different spectrum for W_0, not the Haar scaling algebra. The algorithm introduced above can also be used to generate other scaling functions and their corresponding algebras. An interesting line is to obtain

polygon-like wavelets or, by following the same procedure which guided us to obtain the B-scaling function from the Haar scaling function, to obtain smoother scaling functions.

VI. THE $sl_q(2)$ LIMIT OF $\mathscr{A}_{j_0,j}$

The quantum group $sl_q(2)$ or its extensions[21–26,28–31] are associative algebras over C generated by three operators, $J_0=(J_0)^\dagger$, J_+, and $J_-=(J_+)^\dagger$, satisfying the commutation relations,

$$[J_0,\ J_+]=J_+,\quad [J_0,\ J_-]=-J_-,\quad [J_+,\ J_-]=[J_0]_s, \tag{52}$$

where $[J_0]$ is a deformation of J_0, that is, a real, parameter-dependent function of J_0, holomorphic in the neighborhood of zero, approaching $2J_0$ in the limit $s\to 0$. We have to find a deforming functional that transforms the $\mathscr{A}_{j_0,j}$ generators into operators satisfying the commutation relations Eq. (52). For this purpose the first equation in Eqs. (33) can be written as

$$(j_0-\mathscr{G}(s,J_0))j_+=j_+j_0, \tag{53}$$

and hence, for every entire function $p(\xi)$, we can write

$$p(j_0-\mathscr{G}(s,j_0))j_+=j_+p(j_0). \tag{54}$$

Let us consider the functional equation:

$$\Gamma(\xi-\mathscr{G}(s,\xi))=\Gamma(\xi)-1 \tag{55}$$

for a given function $\Gamma(\xi)$, with ξ generic. If this equation has a solution $\Gamma(\xi)$ that is an entire function, then Eq. (55) can be written, for $p(\xi)=\Gamma(\xi)$, in the form $[\Gamma(j_0)\ j_+]=j_+$, and correspondingly $[\Gamma(j_0),\ j_-]=-j_-$. These equations give the correspondence between the wavelet algebras and the q-deformed algebras like $sl_q(2)$ or other extensions of them. They map the structure of $\mathscr{A}_{j_0,j_\pm}$, Eq. (33), onto the structure described by the first two equations in Eq. (52). For the third commutator of each of these algebras we use the function $\Gamma(\xi)$, which allows Eqs. (33) to be reduced to the third equation in Eqs. (52) through the mapping,

$$J_0=\Gamma(j_0),\quad J_\pm=j_\pm. \tag{56}$$

If Γ is invertible, the third equation in Eqs. (33) in $\mathscr{A}_{j_0,j}$ can be reduced to the third equation of Eqs. (52) of an algebra defined by the function $[J_0]$ with the identification $\mathscr{F}\circ\Gamma^{-1}=[\,]$ where $[\,]\circ\Gamma$ means the composition of the two functions, i.e., $([\,]\circ\Gamma)(\xi)=[\Gamma(\xi)=[\Gamma(\xi)]$. In the case of $sl_q(2)$, i.e., $[J_0[=[2J_0]_q$, the function $\mathscr{F}(s,j_0,s)$ becomes

$$\mathscr{F}(s,j_0)=-\frac{(\phi(j_0))^2-(\phi(j_0))^{-2}}{q-q^{-1}}, \tag{57}$$

where $q=e^s, \phi(\xi)\equiv q^{-\Gamma(\xi)}$ has to satisfy the equation $\phi(\xi-\mathscr{G}(s,\xi))=q\,\phi(\xi)$. Consequently, Eqs. (53)–(55) provide the reduction of $sl_q(2)$ into $\mathscr{A}_{j_0,j}$, through Γ. For details of the above technique of mapping and reduction of nonlinear algebras one can see examples in Ref. 31. In the case of the algebra $\mathscr{A}_{1,1}$ Eqs. (53)–(55) give the condition

$$p(W_0-1)=W_0-G(1,W_0), \tag{58}$$

which has the solution

$$J_0=p(W_0)\big|_{\mathscr{A}_{1,1}}=\cosh(2^{-W_0}b), \tag{59}$$

for any arbitrary constant b. By inverting Eq. (59) we obtain

$$W_0=\mp\frac{1}{\ln(2)}\ln\left(\frac{\partial}{b}\right). \tag{60}$$

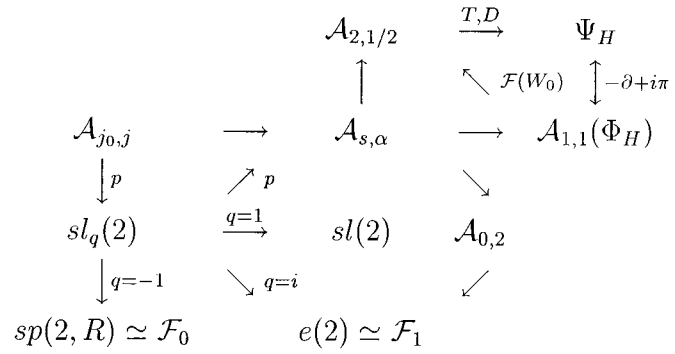

FIG. 3. A diagrammatic representation of the algebraic maps (arrows) and morphisms ($\simeq$) of $\mathscr{A}_{j_0 j}$, $\mathscr{A}_{s,\alpha}$, $\mathrm{sl}_q(2)$, sl(2), and $\mathscr{F}_{0,1}$.

Conversely, starting from $\mathrm{sl}_q(2)$ and its deformation $[\,]_q$ and mapping it into $\mathscr{A}_{1,1}$ we obtain for the operator $F(s,T)=F(s,T(W_0))$,

$$F=f(\Gamma(W_0))=f(p^{-1}(W_0))=[2p^{-1}(W_0)]. \tag{61}$$

A diagram containing these algebra morphisms is presented in Fig. 3.

VII. FURTHER EXTENSIONS, COMMENTS, AND CONCLUSIONS

In this paper we present a method for connecting classes of nonlinear (q-deformed) algebras, with scaling functions and wavelets. This connection is reciprocal. The algebra $\mathscr{A}_{j_0,j}|_{s=1}$ allows the choice of different dilation equations, i.e., for different functions $j(1,T)$ one can obtain different scaling functions. The function $j_0(1,T)$ should be chosen in order to give the correct wavelet function through the action of $j_0 j_-$. The connection with quantum groups is more transparent if we choose for $j_0=[\partial]_{s=1/2}=W_0(2,1/2)$, Eq. (14), and for Φ, the Haar scaling function, $j_-=D(1+T^{-1})$. The action of the q-derivative $[\partial]_{1/2}$ on Φ gives exactly the Haar wavelet. In this example the operator j_0 is anti-Hermitian and we can identify its spectrum (imaginary eigenvalues) from the commutation relations. We also have the relation $\Psi=(j_0-\widetilde{G}(1,j_0)\Phi$, so that one can express the eigenvectors $|a\rangle$ of J_0 in the wavelet basis of the algebra. If $|a\rangle=\Sigma_{j,n}\, C_{j,n}\Phi_{j,n}$, we obtain a recursion relation for the coefficients of $|a\rangle$ and for the eigenfunctions of $j_0\,\Sigma_k\, C_{j,p-2j_k}C_k=aC_{j,p}$ for any j and p integer, where C_k are the Taylor coefficients of the function $-\xi^{-\lambda}g(-\xi)$. We note that by using the q-derivative instead of a function of T we obtain exactly the same mother wavelet (Haar wavelet).

We want to note a natural extension of the above developments, coming from a different realization of sp(2,R) in terms of the full algebra of symmetry of the real line: $X_0=x\partial$ (dilations), $X_-=\partial$ (translations), and $X_+=x^2\partial$ (expansions). This is the maximal finitely generated real Lie algebra with generators in the form $x^n\partial$. By exponentiation, the generators X_0, X_- provide D and T and the last generator has the action $E^a=e^{aX_+}f(x)=f(x/1-ax)$. For $a>0$ the action corresponds to a contraction of the function and for $-1<a<0$ it is a dilation. But, for $a<-1$ and $|a|\geqslant 1$ we obtain a splitting of the function in two Heaviside distributions, $E^a f(x)=H(x)+H(-x+1/a)$. A possible closed nonlinear algebra containing D, T, and E could be an opportunity for wavelets with multiscale properties.[18,19]

In conclusion, we found that the operators of dilation and translation (D,T) can be combined in such a way as to generate nonlinear algebras, depending on certain parameters (s,α). We have investigated these algebras from the point of view of quantum groups, discussing their unirreps, Casimir operators, and reductions of these algebras to other q-deformed algebras, like $\mathrm{sl}_q(2)$ or to the Fourier series generating algebra. It has been shown that such algebras provide an appropriate framework for the foundation of the wavelets analysis and for the obtaining the corresponding scaling functions. We have worked out two examples, the Haar and the B-scaling functions. The

algorithm provides a general algebraic method for finding specific scaling function. Direct applications of such an approach can found in the theory of finite-difference equations and q-difference equations.[21–25,27]

This study represents only a first step towards understanding the relation between the theory of nonlinear algebras (having exponential spectra) and wavelets with their finite-difference equations. We conjecture that such an algebraic approach (nonlinear algebras) to scale invariant structures can lead to interesting mathematical constructions like tools for identifying isolated coherent structures associated with certain mother wavelets and limiting nonlinear algebraic structures. The tranzition between such self-organized structures can be carried out through the modification of the deformation parameter q, with the intermediate domain reprezenting ''noise'' (nonclosed algebraic structures).

[1] I. Daubechies, Commun. Pure Appl. Math. **49**, 909 (1988); I. Daubechies and A. Grossmann, J. Math. Phys. **21**, 2080 (1980).

[2] I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992); C. K. Chui, *An Introduction to Wavelets* (Academic, New York, 1992); G. Kaiser, *A Friendly Guide to Wavelets* (Birkhäuser, Boston, 1994); Y. Meyer, in *Wavelets*, edited by J. M. Combes, A. Grossmann, and Ph. Tchamitchian (Springer, Berlin, 1988), p. 1.

[3] Ph. Tchamitchian and B. Torresani, in *Wavelets and Their Applications*, edited by M. B. Ruskai (Jones Bartlett, Boston, 1992); G. Kaiser, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **23**, 222 (1992); C. K. Chui and A. Cohen, in *Approximation Theory VII*, edited by E. W. Cheuey, C. K. Chui, and L. L. Schumaker (Academic, Boston, 1993).

[4] J. J. Benedetto and M. W. Frazier, *Wavelets: Mathematics and Applications* (Chemical Rubber, Boca Raton, 1994); H. Ogawa and N-E. Berrached, in *Contemporary Mathematics*, edited by M. E. H. Ismail (American Mathematical Society, Providence, Rhode Island, 1995), Vol. 190, G. Strang, Physics D **60**, 239 (1992); SIAM (Soc. Ind. Appl. Math.) Rev. **31**, 614 (1989).

[5] D. Han, Y. S. Kim, and M. E. Noz, Phys. Lett. A **206**, 299 (1996); Y. S. Kim and E. P. Wigner, Phys. Rev. A **36**, 1293 (1987).

[6] C. Meneveau, Phys. Rev. Lett. **66**, 1450 (1991); M. Ima and S. Toh, Phys. Rev. E **52**, 6189 (1995); A. Grossman and J. Morlet, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **15**, 723 (1984); R. Everson, L. Sirovich, and K. R. Sreenivasan, Phys. Lett. A **145**, 314 (1990); R. Benzi and M. Vergassola, Fluid Dyn. (USSR) **8**, 117 (1991); A. Scotti and C. Meneveaux, Phys. Rev. Lett. **78**, 867 (1997).

[7] C. Meneveau and K. R. Sreenivasan, J. Fluid Mech. **224**, 429 (1991); J. F. Muzy, E. Bacry, and A. Arneodo, Phys. Rev. Lett. **67**, 3515 (1991); M. Greiner, J. Giessmann, P. Lipa, and P. Carruthers, Z. Phys. C **69**, 305 (1996).

[8] A. Krolak and P. Trzaskoma, Class. Quantum Grav. **13**, 813 (1996); J. Widjaja, Y. Tomita, and A. Wahab, Opt. Commun. **132**, 217 (1996); J. Pando and L. Z. Fang, Astrophys. J. **459** (1996).

[9] Z. Huang, I. Sarcevic, R. Thews, and X.-N. Wang, Phys. Rev. D **54**, 750 (1996).

[10] J. Kantelhardt, M. Greiner, and E. Roman, Physica A **220**, 219 (1995); H. E. Roman, J. W. Kantelhardt, and M. Greiner, Europhys. Lett. **35**, 641 (1996).

[11] K. Cho, T. A. Arios, and J. D. Joannopoulos, Phys. Rev. Lett. **71**, 1808 (1993).

[12] F. Bagarello, J. Phys. A **29**, 565 (1996).

[13] M. Farge and G. Rabreau, C. R. Acad. Sci. Paris, Ser. II **307**, 1479 (1988).

[14] W. Tobocman, Inverse Probl. **12**, 499 (1996).

[15] S. Afanasiev, M. Altaisky, and Yu. Zhestkov, Il Nuovo Cimm. **108**, 919 (1995).

[16] J. P. Modisette, P. Nordlander, J. L. Kinsey, and B. R. Johnson, Chem. Phys. Lett. **250**, 485 (1996); D. Mugnai, A. Ranfagni, R. Ruggeri, and A. Agresti, Phys. Rev. E **50**, 790 (1994); G. Berkovz, J. Elezgaray, and P. Holmes, Physica D **61**, 47 (1992); A. O. Barut, Found. Phys. **20**, 1233 (1990).

[17] I. Makoto and T. Sadayoshi, Phys. Rev. Lett. **52**, 6189 (1995); L. Gagnon and J. M. Lina, J. Phys. A **27**, 8207 (1994).

[18] A. Ludu and J. P. Draayer, in *Multiscale Phenomena in Physics and Engineering*, edited by P. Vashishta and R. Kalia, (Louisiana State University, Baton Rouge, Louisiana, 1997); *Hamiltonian System and Symmetry Algebra for Wavelets* (submitted to Mod. Phys. Lett.); A. Ludu and J. P. Draayer, *International Conference on Physics at The Turn of the Millenium*, 22–26 October 1996, Santa Fe, NM.

[19] A. Ludu and M. Greiner, ICTP/preprint IC/95/288.

[20] N. I. Vilenkin, *Special Functions and The Theory of Group Representations* (Nauka, Moscow, 1955); J. D. Talman, *Special Functions. A Group Theoretic Approach* (Benjamin, New York, 1968).

[21] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994); J. Fröchlich and T. Kerler, *Quantum Groups, Quantum Categories and Quantum Field Theory* (Springer, Berlin, 1993); E. Witten, Nucl. Phys. B **330**, 285 (1990).

[22] *Symmetries and Integrability of Difference Equations*, edited by D. Levi, L. Vinet, and P. Winter (American Mathematical Society, Providence, 1997).

[23] A. Ludu and W. Greiner, Found. Phys. **27**, 1123 (1997).

[24] R. Floreanini and L. Vinet, J. Math. Phys. **36**, 7024 (1995).

[25] R. Floreanini and L. Vinet, J. Math. Phys. **36**, 3134 (1995).

[26] A. J. Macfarlane, J. Math. Phys. A **22**, 4581 (1989).

[27] C. R. Adams, Bull. Am. Math. Soc. **37**, 361 (1931); G. E. Andrews, *q-Series: Their Development and Applications in Analysis, Number Theory, Combinatories, Physics, and Computer Algebra* (American Mathematical Society, Providence, Rhode Island, 1986).

[28] A. P. Polychronakos, Mod. Phys. Lett. A **5**, 2325 (1990).

[29] M. Rocek, Phys. Lett. B **255**, 554 (1991).
[30] A. Ludu and R. K. Gupta, J. Math. Phys. **34**, 5367 (1993).
[31] C. Delbeque and C. Quesne, J. Phys. A **26**, L127 (1993); D. Bonatsos, C. Daskaloyannis, P. Kolokotronis, A. Ludu, and C. Quesne, J. Math. Phys. **38**, 369 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 4 APRIL 1998

Existence of parallel spinors on nonsimply connected Riemannian manifolds

Brett McInnes[a)]
International Centre for Theoretical Physics, Trieste, Italy

(Received 27 May 1997; accepted for publication 30 October 1997)

It is well known, and important for applications, that Ricci-flat Riemannian manifolds of nongeneric holonomy always admit a parallel (covariant constant) spinor if they are simply connected. The nonsimply connected case is much more subtle, however. We show that a parallel spinor can still be found in this case *provided* that the (real) dimension is not a multiple of four, and *provided* that the spin structure is carefully chosen. © *1998 American Institute of Physics.*
[S0022-2488(98)02403-7]

I. INTRODUCTION

In string and brane theories, the internal part of the vacuum space–time model is taken to be a Ricci-flat Riemannian manifold. This is *necessary* if the manifold is to admit a parallel spinor, which in turn is necessary if $N=1$ supersymmetry is to be preserved.[1] One must ask, however, whether Ricci flatness is *sufficient* to ensure the existence of a parallel spinor, and this is the subject of the present work. This question is particularly topical now that internal manifolds K of dimension other than six are being considered,[2] since the spin geometry of compact Ricci-flat manifolds K varies quite strongly as the dimension n changes. (Throughout this work, ''dimension'' always means ''real dimension.'')

As the only known[3] examples of compact Ricci-flat Riemannian manifolds have nongeneric holonomy [that is, their linear holonomy groups are proper subgroups of $O(n)$, or of $\mathrm{SO}(n)$ in the orientable case] we shall assume henceforth that K is a compact, orientable, Ricci-flat manifold of nongeneric holonomy. Our question is simply this: must K admit a parallel spinor? The answer may be surprising: it is ''no.'' For example, let K be an Enriques surface,[4] so that K is a compact Kähler 4-manifold that admits a Ricci-flat metric. As we shall see, there is no parallel spinor on this manifold. Clearly, it is important to arrive at a better understanding of this situation.

Ricci flatness arises as the *local* integrability condition of the differential equations, expressing the condition that some spinor field be parallel. The question is whether the parallel spinor locally so defined can be extended globally. One finds that complications arise *only if K is not simply connected*. The complications are of two kinds. First, the existence of noncontractible loops can make it impossible to define global parallel spinors altogether; this is what happens in the case of the Enriques surfaces. The second, more subtle complication is as follows. If a spin manifold is not simply connected, then it may have *more than one* spin structure.[5] In such cases, it turns out that the success of any attempt to extend a local parallel spinor *depends on the choice of spin structure*. In fact, in a specific sense to be explained below, ''most'' spin structures on ''most'' nonsimply connected K *do not* permit the extension to be made. However, on six-dimensional Calabi–Yau manifolds, there is always (precisely) one spin structure that does admit a parallel spinor.

Recently, Joyce[6] has given examples of compact seven-manifolds with linear holonomy group isomorphic to G_2. Some of these manifolds are not simply connected, and so we must ask whether global parallel spinors exist on them, and, if so, we should identify those spin structures that permit parallel spinors and those that do not. In fact, these problems can be analyzed in a very direct and explicit way, as we shall now explain.

[a)]Permanent address: Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Republic of Singapore. Electronic mail: matmcinn@nus.edu.sg

 © 1998 American Institute of Physics

II. THE CASE $n=4m$

When n, the dimension of K, is a multiple of 4, the situation is quite clear cut. We have the following theorem.

Theorem 1: Let K be any compact, locally irreducible, Ricci-flat manifold of nongeneric holonomy and real dimension $n=4m$, where m is any integer $\geqslant 1$. Then K admits a parallel spinor if and only if it is simply connected.

Proof: If K is simply connected, its linear holonomy group must be connected, and so the holonomy group must be SU($2m$) or Sp(m) or Spin(7) if $n=8$. Recall[3] that local irreducibility means that K is not a product. All of these groups are themselves simply connected, and so K is a spin manifold.[5] Since every loop in K is contractible, parallel transport of locally constant spinors is unambiguous, and so a global parallel spinor can be constructed. Conversely, suppose that K admits a global parallel spinor with respect to some spin structure. In this case, the given information merely forces the *restricted*[3] holonomy group to be SU($2m$), Sp(m), or Spin(7). This means that a certain canonical form (the holomorphic $2m$-form in the first two cases, the Cayley form in the third) exists locally, but perhaps not globally.[7] However, via the ''squaring'' construction, these canonical forms can be expressed in terms of the parallel spinor.[5] The existence of the global parallel spinor thus forces the canonical forms to exist globally, and so the *full* linear holonomy group of K is a connected group. But this is possible, if $n=4m$, only if K is simply connected.[7,8] This completes the proof.

In connection with the second part of the proof, one should note that the Enriques surfaces mentioned earlier do not have a *connected* linear holonomy group and do not admit a global holomorphic two-form. They are not simply connected and do not admit a parallel spinor. (In fact, they are not spin manifolds.)

From the point of view of applications, Theorem 1 is rather disappointing, since one expects interesting physical consequences if K is not simply connected. For example, nontrivial *flat* gauge configurations (''Wilson loop'' or Hosotani[1] effects) are possible only if K is not simply connected. In physical language, Theorem 1 means that *if n is a multiple of 4, then $N=1$ supersymmetry is inconsistent with Wilson loop effects*.

III. THE CASE $n\neq 4m$

The situation when n is not a multiple of 4 is more satisfactory, but also more subtle. Let K be an n-dimensional spin manifold; that is, there exists a Spin(n) principal bundle Spin(K) over K such that Spin(K) is a nontrivial double cover of SO(K), a bundle of oriented orthonormal frames with respect to some fixed Riemannian structure on K. Let V be some vector space that affords a faithful representation ρ of Spin(n). A spinor on K can be interpreted as a V-valued function ψ on Spin(K) that satisfies $\psi(sg)=\rho(g^{-1})\psi(s)$ for each s in Spin(K) and each g in Spin(n). [This is precisely analogous to the interpretation[9] of vector fields as $\mathbb{R}^n$-valued function on SO(K).]

In general, there is no unique spin structure on K if it is not simply connected. In the case of compact, orientable, Ricci-flat, locally irreducible Riemannian manifolds of nongeneric holonomy, we can construct these explicitly, as follows. First, note that, by the Cheeger–Gromoll theorem,[3] K can be expressed as $K=\widetilde{K}/\Gamma$, where $\widetilde{K}$ is the universal Riemannian cover and Γ is a *finite* group of isometries acting freely on $\widetilde{K}$. If $n\neq 4m$, then both K and $\widetilde{K}$ are spin manifolds. To see this, note that in this case the *full* (not just the restricted) holonomy group of K must be connected:[7] in fact, it must be isomorphic to SU($2r+1$), $r\geqslant 1$, $n=4r+2$, or perhaps to the exceptional group G_2 if $n=7$. [Here we are using the assumption that K is orientable. Nonorientable compact, Ricci-flat manifolds of nongeneric holonomy do exist in these dimensions, and their full holonomy groups are disconnected, as is usually the case for non-simply-connected Riemannian manifolds.] Since SU($2r+1$) and G_2 are simply connected, K is spin, and similarly for $\widetilde{K}$. As $\widetilde{K}$ is simply connected, its first homology group $H_1(\widetilde{K},\mathbb{Z})$ is trivial, and hence so is its first cohomology group with $\mathbb{Z}_2$ coefficients, $H^1(\widetilde{K},\mathbb{Z}_2)$. As the latter counts[5] the spin structures on a spin manifold, we see that $\widetilde{K}$ has a unique spin structure, Spin($\widetilde{K}$).

Let f be any element of Γ, and let $\widetilde{f}$ be its natural lift,[9] so that, because f is isometric, $\widetilde{f}:\mathrm{SO}(\widetilde{K})\to\mathrm{SO}(\widetilde{K})$ is a bundle automorphism of SO($\widetilde{K}$). If we think of Spin($\widetilde{K}$) as a $\mathbb{Z}_2$ principal

bundle over $\mathrm{SO}(\widetilde{K})$, then we can construct an induced bundle[9] $\widetilde{f}^{-1}(\mathrm{Spin}(\widetilde{K}))$, which is again a $\mathbb{Z}_2$ bundle over $\mathrm{SO}(\widetilde{K})$ equipped with a bundle homomorphism $\hat{f}$ that covers $\widetilde{f}$. As usual,[9] $\hat{f}$ is unique modulo automorphisms of $\widetilde{f}^{-1}(\mathrm{Spin}(\widetilde{K}))$ that preserve $\widetilde{f}$. Now in this case $\mathrm{Spin}(\widetilde{K})$ is the only spin structure over $\widetilde{K}$, and so $\hat{f}$ is an automorphism of $\mathrm{Spin}(\widetilde{K})$ covering $\widetilde{f}$. If we denote by $-\hat{f}$ the composite of $\hat{f}$ with -1, where -1 is the action by $-1 \in \mathrm{Spin}(n)$ on $\mathrm{Spin}(\widetilde{K})$, then because -1 is central, $-\hat{f}$ is also an automorphism of $\mathrm{Spin}(\widetilde{K})$ covering $\widetilde{f}$. As f has no fixed point, nor have $\widetilde{f}$ or $\hat{f}$ or $-\hat{f}$.

Let p be the order of f. Then, by the chain rule, p is also the order of $\widetilde{f}$, and so $\hat{f}^p = \pm 1$. In the present case the minus sign can be ignored, as follows. First, if p is odd and $\hat{f}^p = -1$, then $(-\hat{f})^p = 1$ since -1, as an automorphism, commutes with every other automorphism of $\mathrm{Spin}(\widetilde{K})$. Thus if p is odd, there is exactly one automorphism of $\mathrm{Spin}(\widetilde{K})$ of order p covering $\widetilde{f}$; in this case, we pick $\hat{f}$ to be this automorphism and so there is no ambiguity here. Suppose now that p is even, and assume that $\hat{f}^p = -1$. Now K has a canonical parallel form: the holomorphic $(2r+1)$-form if the holonomy group is $\mathrm{SU}(2r+1)$, or the associative calibration[5] in the case of G_2. In each case these forms pull back to a parallel form ω on $\widetilde{K}$, and we have $f^*\omega = \omega$. [We draw the reader's attention to the fact that it is essentially this relation that distinguishes this case from the $(4m)$-dimensional case: for, in the latter, f^* cannot preserve ω. One can still take a quotient in some cases, but then the holonomy group of the quotient manifold will be disconnected.] This relation has the following interpretation in terms of holonomy theory. Let $\widetilde{P}$ be a linear holonomy bundle[9] over $\widetilde{K}$: it is a subbundle of $\mathrm{SO}(\widetilde{K})$ with a structural group isomorphic to the linear holonomy group. The relation $f^*\omega = \omega$ simply means that the natural lift, $\widetilde{f}$, maps $\widetilde{P}$ into itself, $\widetilde{f}\widetilde{P} = \widetilde{P}$. Now the Levi-Civitá connection on $\mathrm{SO}(\widetilde{K})$ induces a canonical Dirac connection on $\mathrm{Spin}(\widetilde{K})$. The holonomy bundles in $\mathrm{Spin}(\widetilde{K})$ cover the holonomy bundles in $\mathrm{SO}(\widetilde{K})$. Let $\hat{P}$ be a Dirac holonomy bundle covering $\widetilde{P}$; then we have $\hat{f}\hat{P} = \pm\hat{P}$, where again -1 acts on $\hat{P}$ through the overall action of $\mathrm{Spin}(n)$ on $\mathrm{Spin}(\widetilde{K})$. Thus $(\hat{f})^2\hat{P} = \hat{P}$, and, since p is even, $(\hat{f})^p\hat{P} = \hat{P}$. If $(\hat{f})^p = -1$, then $-\hat{P} = \hat{P}$, which is only possible if -1 is an element of the structural group of $\hat{P}$. Now $\widetilde{K}$ is simply connected, so the holonomy group of any connection on any principal bundle over $\widetilde{K}$ must be connected; therefore the structural group of $\hat{P}$ is $\mathrm{SU}(2r+1)$ or G_2. But all of these groups have centers isomorphic to an odd cyclic group, and so none contains -1. Thus $\hat{f}^p = +1$. In short, we may take it that $\hat{f}$ has the same order as f itself.

Proceeding in this way, one shows that, just as Γ acts freely on $\widetilde{K}$, so also it acts (in the form of the induced automorphisms $\hat{f}$) freely on $\mathrm{Spin}(\widetilde{K})$. Precisely because the $\hat{f}$ are automorphisms, they commute with the action of $\mathrm{Spin}(n)$ on $\mathrm{Spin}(\widetilde{K})$, and so there is a natural action of $\mathrm{Spin}(n)$ on $\mathrm{Spin}(\widetilde{K})/\Gamma$. It is not difficult to show that $\mathrm{Spin}(\widetilde{K})/\Gamma$ is in fact a principal $\mathrm{Spin}(n)$ bundle over $\widetilde{K}/\Gamma = K$. Of course, Γ also acts freely, through the natural lifts on $\mathrm{SO}(\widetilde{K})$, and we have $\mathrm{SO}(\widetilde{K})/\Gamma = \mathrm{SO}(K)$. Thus $(\mathrm{Spin}(\widetilde{K})/\Gamma)/\mathbb{Z}_2 = \mathrm{SO}(K)$, and, in fact, $\mathrm{Spin}(\widetilde{K})/\Gamma$ is a spin structure over K. However, it is clear from this explicit construction *that there is a crucial ambiguity in the way* Γ *acts on* $\mathrm{Spin}(\widetilde{K})$: namely, each f of even order can be represented *either* by $\hat{f}$ or by $-\hat{f}$. That is, there is an ambiguity not in the structure of Γ, but rather in the way it acts on $\mathrm{Spin}(\widetilde{K})$. (When Γ is not Abelian, there are additional complications because conjugation by some element of the group can mix a given element with others, so that $\hat{f}$ cannot be replaced with $-\hat{f}$ without affecting the conjugate. In this case we simply replace Γ by its Abelianization, and the above remarks apply to the Abelianization.) These various actions by Γ produce different spin structures when we take the quotient $\mathrm{Spin}(\widetilde{K})/\Gamma$, and this is how K can have more than one spin structure. One can show that the resulting spin structures are in one-to-one correspondence with the group of homomorphisms from the Abelianization of Γ to $\mathbb{Z}_2$; this group is isomorphic to $H^1(K,\mathbb{Z}_2)$, and so we recover the well-known fact that $H^1(K,\mathbb{Z}_2)$ counts the spin structures on a spin manifold K.

Some examples will clarify the basic idea. Let K be one of Joyce's[6] compact seven-manifolds with holonomy G_2 and the fundamental group isomorphic to $\mathbb{Z}_2$, so that $K = \widetilde{K}/\mathbb{Z}_2$, with $\mathbb{Z}_2$ generated by an involution f. Then $\widetilde{K}$ has a unique spin structure $\mathrm{Spin}(\widetilde{K})$, but K has two distinct spin structures, $\mathrm{Spin}(\widetilde{K})/\{1,\hat{f}\}$ and $\mathrm{Spin}(\widetilde{K})/\{1,-\hat{f}\}$. Again, suppose that K is a six-dimensional Calabi–Yau manifold of the form $\widetilde{K}/(\mathbb{Z}_2\times\mathbb{Z}_2)$, with $\mathbb{Z}_2\times\mathbb{Z}_2$ generated by a pair of involutions

(a,b). Then K has *four* distinct spin structures, given, respectively, by factoring $\mathrm{Spin}(\tilde{K})$ by $\{1,\hat{a},\hat{b},\hat{a}\hat{b}\}$, $\{1,-\hat{a},\hat{b},-\hat{a}\hat{b}\}$, $\{1,\hat{a},-\hat{b},-\hat{a}\hat{b}\}$, and $\{1,-\hat{a},-\hat{b},\hat{a}\hat{b}\}$. Finally, suppose that Γ is the non-Abelian group of order 12 generated by a,b,c satisfying $a^2=b^2=c^3=1$, $ab=ba$, $cac^{-1}=b$, $cbc^{-1}=ab$. The relation $\hat{c}\hat{a}\hat{c}^{-1}=\hat{b}$ shows that $\hat{a}$ can be replaced by $-\hat{a}$ only if $\hat{b}$ is replaced by $-\hat{b}$, but this is inconsistent with $\hat{c}\hat{b}\hat{c}^{-1}=\hat{a}\hat{b}$. Hence $K=\tilde{K}/\Gamma$ has only *one* spin structure in this case.

With the aid of this explicit construction of the spin structures over K, we can now easily prove the following result.

Theorem 2: Let K be any compact, orientable, locally irreducible, Ricci-flat manifold of nongeneric holonomy and real dimension not a multiple of four. Then there exists *exactly one* spin structure over K admitting a parallel spinor; all other spin structures admit *no* parallel spinor.

Proof: As in the proof of Theorem 1, $\tilde{K}$, the universal cover, admits a parallel spinor. If the holonomy group is G_2, this spinor is essentially unique (the space of parallel spinors is one dimensional); if the holonomy group is $\mathrm{SU}(2r+1)$, it is unique modulo complex conjugation.[10] Let ψ be the corresponding V-valued function on $\mathrm{Spin}(\tilde{K})$, and let ω be the canonical form on $\tilde{K}$. Set $K=\tilde{K}/\Gamma$, and let f be any element of Γ; then $f^*\omega=\omega$ implies, since ω may be obtained from the parallel spinor by ''squaring,'' that $\hat{f}^*\psi=\pm\psi$. But we have seen that the spin structures on K are obtained precisely by taking the quotient of $\mathrm{Spin}(\tilde{K})$ by the various possible actions of Γ on $\mathrm{Spin}(\tilde{K})$. A given spin structure over K will inherit the parallel spinor from $\mathrm{Spin}(\tilde{K})$ if and only if $\hat{f}^*\psi=+\psi$, never $-\psi$, for every f in Γ. Now since $\psi(sg)=\rho(g^{-1})\psi(s)$ for all s in $\mathrm{Spin}(\tilde{K})$ and all g in $\mathrm{Spin}(n)$, we have $\psi(-s)=-\psi(s)$, and so

$$(-\hat{f})^*\psi=-\hat{f}^*\psi.$$

Therefore, if $\hat{f}^*\psi=-\psi$, we have $(-\hat{f})^*\psi=+\psi$. By suitably distributing minus signs, we can therefore always find an action by Γ on $\mathrm{Spin}(\tilde{K})$ that preserves ψ; but only one such arrangement will work. Thus ψ projects to a parallel spinor on *one* spin structure over K, but it fails to do so for every other spin structure. This completes the proof.

For example, we saw that Joyce's manifold with a fundamental group isomorphic to $\mathbb{Z}_2$ has two spin structures, $\mathrm{Spin}(\tilde{K})/\{1,\hat{f}\}$ and $\mathrm{Spin}(\tilde{K})/\{1,-\hat{f}\}$. Since either $\hat{f}^*\psi=\psi$ or $(-\hat{f})^*\psi=\psi$, one of these has a parallel spinor, and the other does not. In the case where the fundamental group is $\mathbb{Z}_2\times\mathbb{Z}_2$, one spin structure has a parallel spinor, while the other three do not.

Readers familiar with Ref. 5 may wonder how Theorem 2 can be consistent with Proposition 10.1 in Chap. IV of that work, which claims that a global spinor always exists if the holonomy group is contained in the isotropy subgroup of the spin group at any point. The subtle point here is that the parallel transport of spinors is governed by different connections [the connections on $\mathrm{Spin}(K)^i$, where i labels the various spin structures] to the one that defines parallel transport of vectors [the connection on $\mathrm{SO}(K)$]. Thus we cannot speak of ''the'' holonomy group until we specify which of these connections is meant. In fact, one can show that while the *linear* connections of these manifolds always have a connected holonomy group, this is not necessarily true of the spinor or Dirac holonomy groups. Thus, requiring the *linear* holonomy group to be contained in the isotropy group may not suffice to ensure that the relevant *Dirac* holonomy group is so contained: this will depend on the choice of spin structure. In fact, Theorem 2 asserts that this kind of argument will usually fail.

In conclusion, then, compact Ricci-flat orientable six-manifolds of nongeneric holonomy behave very differently to their four- and eight-dimensional counterparts: even when they are not simply connected, they always admit a parallel spinor with respect to *some* spin structure. The Joyce manifolds in dimension seven behave, in this respect, exactly like six-dimensional Calabi–Yau manifolds.

IV. CONCLUSION

The results of this work are easily summarized. Let K be a compact, orientable, locally irreducible, Ricci-flat Riemannian manifold of nongeneric holonomy and dimension n. Suppose that K is not simply connected. Then if $n=4m$, K has no parallel spinor; while if $n\neq 4m$, K does

have a parallel spinor, but *only* if one makes precisely the right choice of spin structure. A bad choice of spin structure, then, will break supersymmetry. One might regard this as evidence that Calabi–Yau six-manifolds or Joyce seven-manifolds with several spin structures should be avoided. More speculatively, one might imagine that some kind of ''averaging over spin structures'' on such a manifold could shed some light on the problem of supersymmetry breaking in these theories.

ACKNOWLEDGMENTS

The author is grateful for the hospitality of the International Centre for Theoretical Physics, Trieste, and wishes to thank all those who have made his visit possible.

[1] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, 1987).
[2] B. S. Acharya, Nucl. Phys. B **475**, 579 (1996).
[3] A. L. Besse, *Einstein Manifolds* (Springer-Verlag, Berlin, 1987).
[4] A. Beauville, *Complex Algebraic Surfaces* (Cambridge University Press, Cambridge, 1983).
[5] H. B. Lawson and M. L. Michelsohn, *Spin Geometry* (Princeton University Press, Princeton, 1989).
[6] D. D. Joyce, J. Diff. Geom. **43**, 329 (1996).
[7] B. McInnes, J. Math. Phys. **34**, 4287 (1993).
[8] D. D. Joyce, Invent. Math. **123**, 507 (1996).
[9] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. I.
[10] M. Y. Wang, Ann. Global Anal. Geom. **7**, 59 (1989).

Orbit spaces of compact coregular simple Lie groups with 2, 3 and 4 basic polynomial invariants: Effective tools in the analysis of invariant potentials

G. Sartori[a)]
Dipartimento di Fisica, Università di Padova and INFN, Sezione di Padova, Italy

V. Talamini[b)]
Dipartimento di Ingegneria Civile, Università di Udine and INFN, Sezione di Padova, Italy

(Received 17 December 1996; accepted for publication 21 October 1997)

The coregular real representations of the compact simple Lie groups are classified and the equalities and inequalities defining their orbit spaces are determined in the case of 2, 3 and 4 basic polynomial invariants. The results are obtained making use of the classification of the complex coregular simple Lie groups given by G. W. Schwarz and of the determination of all the allowable metric matrices for q-dimensional ($q \leqslant 4$) orbit spaces of compact coregular linear groups, recently obtained by the present authors. The results are used to determine in a rigorous way the minima of two SO(10) and E_6 invariant Higgs potentials. © *1998 American Institute of Physics.* [S0022-2488(98)03503-8]

I. INTRODUCTION

This paper is part of a program aimed at the determination and classification of the orbit spaces of compact groups, thought of as natural domains of definition of invariant functions. In particular, in this article we shall determine all coregular real representations of simple compact Lie groups, and the corresponding orbit spaces in all the cases in which the number of basic invariants does not exceed four. The results may be used as effective tools in the analysis of invariant potentials, as we shall show working out in detail two substantial examples.

Invariant functions play an important role in all the sectors of physics where symmetries are relevant and an analysis of their properties is often required. As an example, which concerns both elementary particle and solid state physics, let us consider the determination of patterns of spontaneous symmetry breaking in theories in which the ground state of the system is determined by the minimum of an invariant potential $V(x)$ (Refs. 1–4) (V might be a Higgs potential in a gauge field theory or a thermodynamic potential in a Landau theory of structural phase transitions). The problem can be formalized in the following way. The ground state is represented by a vector x_0 belonging to the Euclidean space $\mathbb{R}^n$, on which the symmetry group G (of the formalism describing the system) acts as a group of linear transformations; x_0 is determined as the point at which $V(x)$ assumes its absolute minimum and the isotropy subgroup of G at x_0 (the little group of x_0), G_{x_0}, is the true symmetry group of the system. Generally, the potential depends also on parameters γ (for instance, scalar self-couplings in Higgs potentials, or pressure and temperature in thermodynamic potentials), that cannot be determined from invariance requirements under transformations of G. In this case x_0 and G_{x_0} can depend on the γ's, and various patterns of spontaneous symmetry breaking are allowed, corresponding to distinct structural *phases* of the system.

In supersymmetric field theories the absolute minimum of the potential controls both the spontaneous symmetry and supersymmetry breaking (see, for instance, Ref. 5 and references therein), and often the features of the two breaking schemes are related.[6]

Since the isotropy subgroups of G at nearby points may be different, in all the cases just mentioned, the determination of the point x_0 has to be analytical. This is generally a difficult

[a)]Electronic mail: gfsartori@padova.infn.it
[b)]Electronic mail: talamini@padova.infn.it

 © 1998 American Institute of Physics

computational task, also in the cases in which polynomial approximations are used for the potential, owing to the large number n of variables x_i that are generally involved. An additional difficulty is related to the degeneracy of the stationary points of the potential, which is an unavoidable consequence of the invariance properties of the potential; it prevents, in fact, a direct application[7,8] of Morse's theory.[9]

A more economical approach to the problem was proposed by Gufan,[10,11] who noted that, according to a theorem by Hilbert,[12] a G-invariant polynomial function $V(x)$ can be expressed as a function $\hat{V}(p_1(x),\ldots,p_q(x))$ of a finite set $p(x)=(p_1(x),\ldots,p_q(x))$ of basic polynomial invariant functions, i.e., of a *Minimal Integrity Basis* (abbreviated to MIB) of the ring of G-invariant polynomial functions. When the point $p\in\mathbb{R}^q$ ranges in the domain spanned by $p(x)$, $x\in\mathbb{R}^n$, the function $\hat{V}(p)$ has the same range as $V(x)$, but is not plagued by the same degeneracies. Gufan's proposal found immediate applications in crystal field theory (see, for instance, Refs. 13–17). A full and correct exploitation of his idea required, however, an exact determination of the ranges of the functions $p_i(x)$, a non-trivial problem that was solved only when it was independently remarked that any G-invariant function can be thought of as a function in the orbit space $\mathbb{R}^n/G$ of the action of G in $\mathbb{R}^n$,[18,19] where the p_i's can be used advantageously to parametrize the orbits. In $\mathbb{R}^n/G$, the images of all the points of $\mathbb{R}^n$ with the same invariance properties under G transformations form smooth sub-manifolds, which are usually called *strata*. By varying the parameters γ, the location of the minimum of $V(x;\gamma)$ may shift from a stratum to a different one, thus causing a structural phase transition of the system.

A sensible progress in the characterization of the geometry of the orbit spaces of the compact linear groups (CLG's) was achieved with the discovery that the orbit spaces of the CLG's are connected semi-algebraic varieties, whose defining equations and inequalities can be expressed in the form of positivity conditions of matrices $\hat{P}(p)$, hereafter called $\hat{P}$-matrices, determined only from the gradients of the basic polynomial invariants $p_1(x),\ldots,p_q(x)$:[18–21]

$$\hat{P}_{ab}(p(x))=\sum_1^n{}_i \frac{\partial p_a(x)}{\partial x_i}\frac{\partial p_b(x)}{\partial x_i}, \quad a,b=1,\ldots,q. \tag{1}$$

Using this result, one can obtain a concrete realization of the orbit space of all CLG's whose sets of basic invariant polynomial functions are known.

The definition of a MIB does not exclude the possibility of algebraic relations among its elements, the important point is that none of them can be expressed as a *polynomial* function of the others. If there are no algebraic relations among the elements of a MIB of a linear group G, then G is said to be *coregular*.

The *finite coregular* linear groups coincide with the finite groups generated by reflections[22] and sets of basic polynomial invariants have been given by many authors (see, e.g., Ref. 23 and references therein). However, the classification of all the coregular CLG's and the determination of their MIB's is a non-trivial problem, which is still open; but for the finite groups, partial information is available only about simple[24] and semisimple[25] Lie groups.

In the case of *non-coregular* CLG's, the non-independence of the elements of the MIB's, makes their determination much more difficult.

An alternative indirect method of computation of the $\hat{P}$-matrices, which somehow bypasses the problem of determining a complete set of basic polynomial invariants, was proposed in Refs. 26 and 27, where the $\hat{P}$-matrices were shown to be solutions of a universal *master* equation, satisfying convenient initial conditions (*allowable solutions*). The statement of the master equation does not require the explicit knowledge of a MIB and its solution involves as free parameters only the number and the degrees of the elements of a MIB.

For coregular CLG's, with less than 5 elements in the MIB's, this approach enabled us to determine the orbit spaces, thus proving that, for each fixed choice of their degrees, they can be classified in a finite (and small, for small dimensions and degrees) number of isomorphism classes.[26–28]

This result makes particularly appealing the orbit space approach to the study of invariant functions and spontaneous symmetry breaking. In fact, invariance properties are often the only bounds which are imposed on the potential (beyond regularity and stability properties and/or

bounds on the degree, when the potential is a polynomial function) and, if the symmetry groups of the potentials of different theories share isomorphic orbit spaces, the potentials have the same formal expression and the same domain when written as functions in orbit space, despite the completely different physical meaning of the variables and parameters involved in the definition of the potentials. Thus, the problems of determining the geometric features of the phase space, the location and stability properties of the minima of the potential, the number of phases and the allowed phase transitions are identical in all these theories.[19,29,26,27]

For coregular CLG's, the determination of all the allowable solutions of the master equation has been accomplished for $q \leqslant 4$ (see Refs. 27 and 28); for $q > 4$, it appears to be still possible, but extremely lengthy. This difficulty can be at least partially overcome through inductive procedures.[30]

For non-coregular CLG's, the determination of convenient additional (''initial'') conditions, allowing us to find all the significant solutions of the master equation, is still an open problem, which has been solved only in the case of only one independent relation among the elements of the MIB.

In any case, the master equation and the associated initial conditions are only a set of necessary conditions; even if quite stringent, they need not be sufficient. Therefore, once the allowable solutions of the master equation have been determined, the problem remains of selecting those which are really generated by a group.

This problem was solved in Refs. 31 and 32 with regard to the finite coregular linear groups, for which, as already remarked, complete information is available concerning classification and basic sets of invariants. The $\hat{P}$-matrices of all the finite compact linear groups with up to 4 basic invariants were explicitly calculated and checked to be present in the list of allowable solutions of the master equation determined in Refs. 26–28.

In this article we shall solve the analogous problem in the case of compact simple Lie groups, that is, we shall identify among the allowable $\hat{P}$-matrices (abbreviated to A$\hat{P}$M's) listed in Refs. 26–28, those corresponding to coregular real representations of compact simple Lie groups. As far as we know, an explicit classification of these groups cannot be found in the mathematical literature, and only partial information is available about their sets of basic polynomial invariants. We shall derive the missing pieces of information from results obtained by G. W. Schwarz,[24] who classified the *coregular complex* representations of complex simple Lie groups and determined the degrees of the corresponding basic invariant polynomials. In Schwarz's paper, hints can also be found allowing us to construct explicitly MIB's for most of the coregular complex representations of complex simple Lie groups.

Once the list of the real coregular representations of the compact simple Lie groups has been obtained, the identification of the associated $\hat{P}$-matrices turns out to be trivial or easy in the following two cases:

(1) There is only one allowable solution of the master equation in correspondence with a given set of degrees of the basic polynomial invariants.
(2) The representation is the adjoint representation. As we shall show, the $\hat{P}$-matrix of the adjoint representation is the same as the $\hat{P}$-matrix of the corresponding Weyl group, which is a finite group, so that the results of Ref. 32 can be used.

In all the other cases a sounder analysis of each particular situation has been necessary. In some cases the explicit knowledge of the full MIB could be inferred from Ref. 24, thus the determination of the $\hat{P}$-matrix has been easy. In some other cases, even if only partial information about the MIB is provided in Ref. 24, this turned out to be sufficient to single out the correct $\hat{P}$-matrix.

In order to make clear how our results can be exploited in physically interesting situations, we have worked out a rigorous and complete discussion of the minima of two Higgs potentials that have been the object of some attention in model building of grand unified gauge field theories.[33,34] This will prove that, in all the cases in which the relevant $\hat{P}$-matrix is known, like in the cases listed in Table V, the problem of determining the localization of the absolute minimum of the potential and the possible patterns of spontaneous symmetry breaking is reduced to an almost trivial exercise, at least in the case of fourth degree polynomial potentials.

The paper will be organized in the following way. We shall collect in Section II all the definitions and results concerning group actions and the $\hat{P}$-matrix approach to the determination of the orbit spaces of coregular CLG's, in order to make our work as self-contained as possible. Section III deals with adjoint representations, and Sections IV, V and VI report the theorems and the results needed to determine and classify the real coregular representations of compact simple Lie groups from the complex coregular representatations of complex simple Lie groups classified in Ref. [24]. In Section VII are identified the $\hat{P}$-matrices of all the coregular real representations of compact simple Lie groups with less than 5 basic polynomial invariants. Section VIII is devoted to a complete determination of the minima of two general fourth degree polynomial (Higgs) potentials invariant under the compact groups SO(10) and, respectively, E_6 and of the corresponding phase diagrams. The conclusions are reported in Section IX.

In the body of the paper we shall continue to use the following abbreviations:

CLG = Compact Linear Group;

MIB = Minimal Integrity Basis;

A$\hat{P}$M = Allowable $\hat{P}$ Matrix.

II. AN OVERVIEW OF THE GEOMETRY OF LINEAR GROUP ACTIONS

In this section, we shall first define most of our notations and recall, without proofs, some results concerning invariant theory and the geometry of orbit spaces of CLG's. For the unreferenced results see, for instance, Refs. 35 and 36.

A. Orbits, strata and orbit spaces

A couple (G,X) is called a *G-space* if G is topological group and X is a topological space in which G acts as a group of transformations.

We shall denote by $g\cdot x$ the image of $x\in X$ under the action of $g\in G$ and by Ω_x the *G-orbit* through x: $\Omega_x=\{g\cdot x|g\in G\}$. Distinct orbits are disjoint.

The isotropy subgroup of G at x (little group of x) will be denoted by G_x: $G_x=\{g\in G|g\cdot x=x\}$. Isotropy subgroups of G, at points lying on the same orbit Ω_x, are conjugated subgroups in G: $G_{g\cdot x}=gG_xg^{-1},\forall g\in G$.

The class $[G_x]$ of all the subgroups of G conjugated to G_x in G is called the *orbit type of* Ω_x (and of its points). In the set of all orbit types a partial ordering relation can be introduced. We shall write $[G_x]<[G_{x'}]$ if G_x is conjugated to a proper subgroup of $G_{x'}$. In this case we shall say that Ω_x (or anyone of its points) has a *lower symmetry* than $\Omega_{x'}$ (or anyone of its points).

The points $x\in\mathbb{R}^n$ (or, equivalently, the orbits of G) sharing the same orbit type are said to form an *isotropy type stratum* of the action of G in $\mathbb{R}^n$. The connected components of a stratum can be shown to be smooth iso-dimensional sub-manifolds of $\mathbb{R}^n$.

In this paper we shall be interested only in compact Lie group actions. Owing to a famous theorem of Mostow,[37] if these actions are sufficiently regular, they can be reduced to *real linear orthogonal* actions, to which we shall devote most of our attention in the following. Complex linear actions will be considered only as auxiliary tools. Therefore, in the rest of this paper, X will denote $\mathbb{R}^n$ or $\mathbb{C}^n$ and these spaces will be considered both as affine spaces and as vector spaces in which G will be assumed to act as a linear group of unitary operators. By introducing in X the standard basis and the standard Hermitian scalar product the elements of G will be represented by unitary matrices.

The *orbit space* of the action of G in $\mathbb{R}^n$ is defined as the quotient space $\mathbb{R}^n/G$ (obtained through the equivalence relation relating point belonging to the same orbit) endowed with the quotient topology and differentiable structure.

We shall denote by π the canonical projection $\mathbb{R}^n\to\mathbb{R}^n/G$. Whole orbits of G are mapped by π into single points of $\mathbb{R}^n/G$. The image through π of a stratum of $\mathbb{R}^n$ will be called an (*isotropy type*) *stratum of* $\mathbb{R}^n/G$; all its connected components turn out to be smooth iso-dimensional manifolds.

Almost all the points of $\mathbb{R}^n/G$ belong to a unique stratum Σ_p, the *principal stratum*, which is a connected open dense subset of $\mathbb{R}^n/G$. The boundary $\overline{\Sigma}_p\backslash\Sigma_p$ of Σ_p is the union of disjoint *singular* strata. All the strata lying on the boundary $\overline{\Sigma}\backslash\Sigma$ of a stratum Σ of $\mathbb{R}^n/G$ are open in $\overline{\Sigma}\backslash\Sigma$.

The orbit type $[H]$ of a stratum Σ of $\mathbb{R}^n/G$ is contained in the orbit types $[H_b]$ of all the strata Σ_b lying in its boundary; therefore, more peripheral strata are formed by points representing orbits with higher symmetry under G transformations.

The number of distinct orbit types of G is finite and there is a unique minimum orbit type, the *principal orbit type*, corresponding to the principal stratum, and also a unique maximum orbit type $[G]$, corresponding to the image through π of the set of points of $\mathbb{R}^n$, which are invariant under G; this set contains at least the origin of $\mathbb{R}^n$.

A faithful image of $\mathbb{R}^n/G$ can be obtained in the framework of the geometric invariant theory.

A function $f(x)$ is said to be G-invariant if

$$f(g\cdot x)=f(x), \quad \forall x\in\mathbb{R}^n, \quad g\in G. \tag{2}$$

The set of all real, G-invariant, polynomial functions of x forms a ring $\mathbb{R}[x]^G$. By a celebrated theorem of Hilbert, which he first proved for SL(N) and which was extended by Nagata to much more general situations,[12,38,39] including all compact linear groups, the ring of polynomial invariants of reductive algebraic groups is finitely generated, that is, there exists a finite collection of invariant polynomial functions, which we shall denote by $p(x)=(p_1(x),p_2(x),\ldots,p_q(x))$, such that any element $F\in\mathbb{R}[x]^G$ can be expressed as a polynomial function $\hat{F}$ of $p(x)$:

$$F(x)=\hat{F}(p(x)), \quad \forall x\in\mathbb{R}^n, \quad \forall F\in\mathbb{R}[x]^G. \tag{3}$$

As proved by G.W. Schwarz,[40] Eq. (3) holds true also if F is any G-invariant C^∞ function. In this case $\hat{F}$ is a C^∞ function.

The polynomial function $\hat{F}(p)$, $p\in\mathbb{R}^q$, will be said to have weight w if w is the degree of the polynomial $F(x)=\hat{F}(p(x))$, and it will be said to be w-homogeneous of *weight* w if $F(x)$ is homogeneous of degree w. We shall denote by $w(\hat{F})$ the weight of $\hat{F}$.

The elements of a basis of $\mathbb{R}[x]^G$ can be chosen to be homogeneous polynomials. The number q of elements of a MIB and their homogeneity degrees d_i's are only determined by the group G.

Without loss of generality and to avoid trivial situations, in this paper we shall only consider linear groups with no fixed points, but for the origin of $\mathbb{R}^n$. In this case, the minimum degree of the elements of a MIB is necessarily 2, and, because of the orthogonality of the G-action, the following conventions can be adopted:

$$d_1\geqslant d_2\geqslant \ldots d_q=2; \quad p_q(x)=\|x\|^2=\sum_1^n{}_i\, x_i^2. \tag{4}$$

Hereafter, by a MIB we shall mean a *minimal homogeneous* integrity basis satisfying the conventions (4).

The orbits of a compact group G are separated by the elements of its MIB's, i.e., at least one element of a MIB takes on different values on two distinct orbits. But more can be said. Each MIB $\{p_1(x),\ldots,p_q(x)\}$, defines an orbit map $p:\mathbb{R}^n\to\mathbb{R}^q$ which maps all the points of $\mathbb{R}^n$ lying on an orbit of G into a single point of $\mathbb{R}^q$:

$$p:\mathbb{R}^n\to\mathbb{R}^q$$

$$x\mapsto(p_1(x),\ldots,p_q(x)).$$

The range $p(\mathbb{R}^n)$ of the orbit map p yields a faithful image of the orbit space of G, and the elements $\{p_1,\ldots,p_q\}$ of the MIB, thought of as coordinates in the space $\mathbb{R}^q$, provide a smooth parametrization of the points of $\mathbb{R}^n/G$. In fact, every orbit map induces a diffeomorphism of $\mathbb{R}^n/G$ onto a semi-algebraic closed connected subset $\overline{\mathscr{S}}$ of $\mathbb{R}^q$

$$\overline{\mathscr{S}}=p(\mathbb{R}^n)\simeq\mathbb{R}^n/G. \tag{5}$$

The orbit space $\mathbb{R}^n/G$ will be identified to its image $\overline{\mathscr{S}}$, which is a connected semi-algebraic sub-variety of $\mathbb{R}^q$, that is a subset of $\mathbb{R}^q$ defined by polynomial equalities and inequalities. Like all

semi-algebraic varieties,[41] $\overline{\mathscr{S}}$ presents a natural stratification in connected disjoint semi-algebraic sub-varieties $\sigma_i^{(\alpha)}$, called *primary strata*, such that $\overline{\mathscr{S}}=\cup_{i,\alpha}\sigma_i^{(\alpha)}$ and the boundary of each $\sigma_i^{(\alpha)}$ is empty, or the union of lower-dimensional $\sigma_j^{(\beta)}$'s $(j>i)$. Each $\sigma_i^{(\alpha)}$ is open in its closure and it is defined recursively on i (which distinguishes semi-algebraic sets of different dimensions) as the α-th connected component of the set of regular points of the semi-algebraic varieties $\mathscr{W}_i=\overline{\mathscr{S}}\setminus\cup_{0<j<i}\sigma_j^{(\beta)}$, $i=1,2\ldots$. A simple example of a compact connected semi-algebraic variety of $\mathbb{R}^3$ is yielded by a polyhedron. Its interior points form its unique 3-dimensional primary stratum, while 2-, 1- and 0-dimensional primary strata are formed, respectively, by the interior points of each face, by the interior points of each edge and by each vertex.

A ring of polynomial functions $f(p_1,\ldots,p_q)$ is said to be *regular* if it isomorphic to a polynomial ring in q indeterminates, that is, if the elements of its integrity bases are algebraically independent.

The linear group G is said to be *coregular* if its ring of invariant polynomial functions is regular.

In this paper we shall only deal with coregular CLG's.

B. The $\hat{P}(p)$ matrix

Let $\{p(x)\}=\{p_1(x),\ldots,p_q(x)\}$ be a MIB for the group G and let $d_1,\ldots,d_q$ be the corresponding degrees. We shall associate to $\{p(x)\}$ the following square matrix whose elements, defined in terms of G-invariant Euclidean scalar products, are G-invariant polynomials in x; therefore they can be expressed in terms of polynomial functions $\hat{P}_{ab}(p)$, according to Hilbert's theorem

$$P_{ab}(x)=\sum_{1}^{n}{}_i\frac{\partial p_a(x)}{\partial x_i}\cdot\frac{\partial p_b(x)}{\partial x_i}=\hat{P}_{ab}(p(x)),\quad a,b=1,\ldots,q. \tag{6}$$

In our conventions, $\hat{P}(p)$ is a real $q\times q$ symmetric matrix, whose elements $\hat{P}_{ab}(p)$ are w-homogeneous polynomials of weight $w(\hat{P}_{ab})=d_a+d_b-2$ and $\hat{P}_{qa}(p)=\hat{P}_{aq}(p)=2d_ap_a$, $a=1,2,\ldots,q$.

A characterization of the image $\overline{\mathscr{S}}=p(\mathbb{R}^n)$ of the orbit space of G as a semi-algebraic variety can be easily obtained through the matrix $\hat{P}(p)$ associated to one of its MIB's and the following fundamental theorem:[18,19,21]

Theorem II.1: *Let G be a compact coregular matrix subgroup of $\mathrm{O}_n(\mathbb{R})$, p be the orbit map $\mathbb{R}^n\to\mathbb{R}^q$ defined by the homogeneous MIB $\{p_1(x),p_2(x),\ldots,p_q(x)\}$ and $\hat{P}(p)$ be the matrix defined in (6). Then $\overline{\mathscr{S}}=p(\mathbb{R}^n)$ is the semi-algebraic subset $\mathbb{R}^q$ where $\hat{P}(p)$ is positive semi-definite. The set $\mathscr{S}$ is connected. The k-dimensional primary strata of $\mathscr{S}$ are the connected components of the set $\hat{W}^{(k)}=\{p\in\mathbb{R}^q;|\hat{P}(p)\geqslant 0,\ \mathrm{rank}\,(\hat{P}(p))=k\}$; they are the images of the connected components of the k-dimensional isotropy type strata of $\mathbb{R}^n/G$. In particular the interior $\mathscr{S}$ of $\overline{\mathscr{S}}$, where $\hat{P}(p)$ has the rank q, is the image of the principal stratum and is connected.*

Distinct MIB's of G, $\{p(x)\}$ and $\{p'(x)\}$, are related by morphisms that we shall call MIB transformations:

$$\begin{aligned} p'_\alpha&=p'_\alpha(p),\quad \alpha=1,\ldots,q-1,\\ p'_q&=p_q\,. \end{aligned} \tag{7}$$

In (7), each polynomial function $p'_\alpha(p)$ depends only on the p_β's of weights d_β not greater than d'_α. The associated $\hat{P}$-matrices satisfy the following relation:

$$\hat{P}'(p'(p))=J(p)\cdot\hat{P}(p)\cdot J^T(p), \tag{8}$$

where $J(p)$ is the Jacobian matrix of the transformation:

$$J_{ab}(p)=\partial p'_a(p)/\partial p_b\,,\quad a,b=1,\ldots,q; \tag{9}$$

since $w(J_{ab})=d_a-d_b$, $J(p)$ is a block upper triangular matrix with constant determinant. Then $\hat{P}'(p'(p))$ and $\hat{P}(p)$ have the same rank, $\forall p \in \overline{\mathscr{S}}$.

Definition II.1: Two matrices $\hat{P}(p)$ and $\hat{P}'(p')$ will be said to be equivalent if they are connected by a relation like (8)*, where $J(p)$ is the Jacobian matrix of a MIB transformation $p'=p'(p)$.*

Thus to every real linear group an equivalence class of $\hat{P}$-matrices is uniquely associated.

If $\varphi(G)$ and $\varphi'(G)$ are unitarily equivalent real orthogonal matrix representations of G:

$$\varphi'(g)=O\varphi(g)O^{-1}, \quad g\in G, \quad OO^T=\mathbb{1}, \tag{10}$$

and if $\{p(x)\}$ is a MIB for $\varphi(G)$, then $\{p'(x)\}$, defined as

$$\{p'(x)\}=\{p(O^{-1}x)\}, \tag{11}$$

is a MIB for $\varphi'(G)$ and $\hat{P}'(p')=\hat{P}(p')$. Thus, the equivalence classes of $\hat{P}$-matrices associated to equivalent linear orthogonal groups coincide and an equivalence class of $\hat{P}$-matrices can be uniquely associated to each orthogonal representation of G.

The semi-positivity domains of equivalent $\hat{P}$-matrices are isomorphic semi-algebraic sets. Therefore, images of the orbit space of G obtained from different MIB's are isomorphic semi-algebraic sets.

The notions of MIB transformations and of equivalence of $\hat{P}$-matrices can be defined in abstract terms and extended to the case of different coregular groups G and G', provided that their MIB's have the same number of elements, with the same degrees. Thus the classification of the isomorphism classes of the orbit spaces of the coregular CLG's rests on the determination of a representative for each class of equivalent $\hat{P}(p)$ matrices.

This can be done, in principle, for all coregular CLG's, even if these groups have been only partially classified, and in many cases their MIB's are not known. The matrices $\hat{P}(p)$ have been shown,[26] in fact, to be solutions of a *master equation*, satisfying convenient initial conditions (*allowable* solutions).The master equation does involve only the number q of the elements of the MIB's as a free parameter and, in convenient bases (*A-bases*),[27] can be put in the following *canonical* form:

$$\sum_1^q {}_b\hat{P}_{ab}(p)\partial_b A(p)=2\delta_{aq}w(A)A(p), \quad a=1,\ldots,q, \tag{12}$$

where $A(p)$ is the unique generator of the ideal $\mathscr{I}(\mathscr{B})$, associated to the union $\mathscr{B}$ of all the $(q-1)$-dimensional strata of $\mathscr{S}$ (the closure of $\mathscr{B}$ forms the boundary of $\mathscr{S}$).

The following results have been proved in Ref. 27 to hold true in every A-basis:

Proposition II.1: i) The point $p^{(0)}=(0,\ldots,0,1)$ lies in the interior $\mathscr{S}$ of $\mathscr{S}$; it is the image of a particular G-orbit lying on the unit sphere of $\mathbb{R}^n$.

ii) A(p) is a factor of $\det \hat{P}(p)$*; it can be normalized to an arbitrary positive value at $p^{(0)}$:*

$$A(p^{(0)})>0, \tag{13}$$

and its weight is bounded:

$$2d_1 \leqslant w(A) \leqslant w(\det\hat{P})=2\sum_1^q {}_a\, d_a-2q. \tag{14}$$

iii) The restriction $A(p)|_{p\in\Pi}$, of $A(p)$ to the hyperplane $\Pi=\{p\in\mathbb{R}^q | p=(p_1,\ldots,p_{q-1},1)\}$ of $\mathbb{R}^q$, has a unique local non-degenerate maximum, lying at $p^{(0)}$; thus

$$\partial_\alpha A(p)|_{p=p^{(0)}}=0, \quad \alpha=1,\ldots,q-1. \tag{15}$$

iv) $\hat{P}(p^{(0)})$ *is block diagonal, each block being associated to a subset of* p_a*'s sharing the same weight, and, in a subclass of A-bases (standard A-bases), it is diagonal:*

$$\hat{P}_{ab}(p^{(0)})=\kappa_a\delta_{ab}, \quad \kappa_a>0, \quad a,b=1,\ldots,q. \tag{16}$$

Definition II.2: The solutions $(\hat{P}(p),A(p))$ *of the canonical equation satisfying the initial conditions specified in* (13) *and* (16) *will be called allowable solutions and the corresponding* $\hat{P}$ *matrices, allowable* $\hat{P}$ *matrices.*

All the $\hat{P}$-matrices associated to the different MIB's of any coregular CLG with no fixed points are necessarily equivalent to an allowable $\hat{P}$-matrix. At present we do not know if the converse holds also true, i.e., if every *allowable* $\hat{P}$-matrix is generated by a coregular CLG with no fixed points.

The allowable solutions of the canonical equations for $q\leqslant 4$ have been determined in Refs. 27 and 28. For each choice of the degrees $\{d_1,d_2,\ldots,d_q\}$, only a finite or null number of nonequivalent solutions has been found, showing the existence of *selection rules* for the degrees of the MIB's of the coregular CLG's. The solutions can be organized in *towers*; the degrees of the elements of the same tower can be written in the form $d_\alpha=sd_\alpha^{(0)}, \alpha=1,\ldots,q-1$, $d_q=2$, where s is an integer scale parameter. All $\hat{P}$-matrices of the same tower coincide on the hyperplane $p_q=1$, corresponding to the image of the unit sphere of $\mathbb{R}^n$ in $\mathbb{R}^q$; this means that the sections for $p_q=1$ of the semi-positivity regions $\overline{\mathscr{S}}$ are equal for all the $\hat{P}$-matrices in the same tower.

In this paper we shall recognize, among the allowable $\hat{P}$-matrices classified in Refs. 27 and 28, those corresponding to coregular real representations of compact simple Lie groups.

III. $\hat{P}$-MATRICES ASSOCIATED TO ADJOINT REPRESENTATIONS OF SIMPLE LIE GROUPS

The isomorphism between the ring of invariant polynomial functions of the adjoint representation of a simple compact Lie group G and the ring of invariant polynomial functions of its Weyl group has been known for a long time.[22] In this section we shall show that there is identity between the equivalence class of the $\hat{P}$-matrices associated to the adjoint representation of any compact simple Lie group and the equivalence class of the $\hat{P}$-matrices associated to the corresponding Weyl group.

Let $\mathscr{G}$ be the real Lie algebra of the compact simple Lie group G and let $\{\tau_i\}_{1\leqslant i\leqslant n}$, be an orthonormal basis of $\mathscr{G}$ with respect to the Killing form:

$$(\tau_i,\tau_j)=-\delta_{ij}. \tag{17}$$

In this basis the structure constants c_{ijk} are completely antysymmetric and the operators Adj g, $g\in G$ are represented by real orthogonal matrices $O(g)$:

$$\text{Adj } g\cdot\tau_i=g\tau_i g^{-1}=\sum_1^n {}_jO_{ji}(g)\tau_j; \quad O(g)^TO(g)=\mathbb{1}. \tag{18}$$

We shall consider the real linear group $\Gamma=\{O(g)|g\in G\}$ of the real orthogonal matrices realizing the adjoint representation in $\mathbb{R}^n$.

As is well known,[42] with reference to a fixed Cartan sub-algebra $\mathscr{H}$ of $\mathscr{G}$ and a set of primitive roots $\{\alpha_i\}_{1\leqslant i\leqslant l}$, a basis in the complexification $\mathscr{G}_c=\mathscr{G}+i\mathscr{G}$ of $\mathscr{G}$ can be chosen in the following way:

$$\{h_1,\ldots,h_l,e_\alpha,e_{-\alpha},\ldots\} \tag{19}$$

where $\{ih_j\}_{1\leqslant j\leqslant l}$ is an orthonormal basis in $\mathscr{H}$ and α ranges over all the positive roots of $\mathscr{H}$ in $\mathscr{G}$.

Correspondingly, the basis $\{\tau\}$ in $\mathscr{G}$ can be chosen so that

$$\{\tau\}=\left\{ih_1,\ldots,ih_l,\frac{e_\alpha+e_{-\alpha}}{\sqrt{2}},\frac{e_{-\alpha}-e_\alpha}{i\sqrt{2}},\ldots\right\}. \tag{20}$$

To each root α a reflection S_α can be associated in the dual space $\mathscr{H}^*$:

$$S_\alpha\cdot\beta=\beta-\frac{2(\beta,\alpha)}{(\alpha,\alpha)}\alpha. \tag{21}$$

The Weyl group is defined as the group of linear transformations generated by the reflections S_α, α ranging over all the positive roots.

The action of the Weyl group on $\mathscr{H}^*$ defines a dual action of the group on $\mathscr{H}$:

$$S_\alpha^*\cdot h=h-\frac{2(h,h_\alpha)}{(h_\alpha,h_\alpha)}h_\alpha. \tag{22}$$

With respect to the basis introduced in $\mathscr{G}$, the Weyl group is therefore represented by a linear group W of real orthogonal $l\times l$ matrices.

By defining

$$S_\alpha^*\cdot e_\beta=e_{S_\alpha\cdot\beta}, \tag{23}$$

the action of the Weyl group can be extended by linearity to the whole of $\mathscr{G}_c$. In this way, the action of S_α^* defines an automorphism of $\mathscr{G}_c$, which leaves $\mathscr{H}$ invariant and is, therefore, an inner automorphism.

Let now $p(x)$, $x\in\mathbb{R}^n$, be a Γ-invariant polynomial function. On the basis of what we have just said, $p(x)$ will also be invariant with respect to transformations induced by the dual action of the Weyl group. As a consequence, the restriction of $p(x)$ to the subspace of $\mathbb{R}^n$ determined by the vectors $(x_1,\ldots,x_l,0,\ldots,0)$ defines a W-invariant polynomial $p^{(\mathscr{H})}(\xi)$, $\xi\in\mathbb{R}^l$:

$$p^{(\mathscr{H})}(\xi)=p(\xi_1,\ldots,\xi_l,0,\ldots,0). \tag{24}$$

It has been shown by Chevalley[22] that all the W-invariant polynomials can be generated in this way. Therefore the ring $\mathbb{R}[\mathbb{R}^n]^\Gamma$ formed by the real Γ-invariant polynomials is isomorphic to the ring $\mathbb{R}[\mathbb{R}^l]^W$ formed by the real W-invariant polynomials and, if $\{p_1(x),\ldots,p_q(x)\}$ is a MIB for Γ, then $\{p_1^{(\mathscr{H})}(\xi),\ldots,p_q^{(\mathscr{H})}(\xi)\}$ is a MIB for W. Since W is an irreducible finite linear group generated by reflections, acting in an l-dimensional space, its MIB's are formed by l elements. Therefore, in the rest of this section we shall set $q=l$.

Since the group W is generated by reflections, it is a coregular group, whichever the simple Lie group G may be. As a consequence the adjoint representations of all the compact simple Lie groups define coregular real linear groups.

The MIB's for the adjoint representations of the simple Lie groups have been determined, for instance, in Refs. 43 and 44, where it has been shown that they can be built in the following way.

Let φ denote a real or complex representation of $\mathscr{G}$ and let us set

$$\mathscr{G}\ni X=\sum_1^n{}_i\, x_i\tau_i,\quad x\in\mathbb{R}^n, \tag{25}$$

$$I_\varphi^{(k)}(x)=\mathrm{Tr}[\varphi(X)]^k. \tag{26}$$

Then, for every $O(g)$, $g\in G$, making use of (18) one finds

$$\begin{aligned} I_\varphi^{(k)}(O(g)x)=\mathrm{Tr}\left[\varphi\left(\sum_1^n{}_{ij}O_{ij}(g)x_j\tau_i\right)\right]^k&=\mathrm{Tr}\left[\varphi\left(\sum_1^n{}_j\, x_j g\tau_j g^{-1}\right)\right]^k\\ &=\mathrm{Tr}[\varphi(g)\varphi(X)\varphi(g)^{-1}]^k\\ &=\mathrm{Tr}[\varphi(X)]^k=I_\varphi^{(k)}(x). \end{aligned} \tag{27}$$

Thus $I_\varphi^{(k)} \in K[\mathbb{R}^n]^\Gamma$, with $K=\mathbb{R}$ or $\mathbb{C}$, according to φ being real or complex.

The $q=l$ elements of a MIB can be obtained from the set of invariants $\{I_\varphi^{(k)}\}$ by choosing a convenient subset of φ's and k's.

Below we shall need the following Lemma:

Lemma III.1: $(\partial_j I_\varphi^{(k)})(x_1,\ldots,x_l,0,\ldots,0)=0$ *for* $j=l+1,\ldots,n$ *and* $k\geqslant 2$.

Proof: Let us set $H=\Sigma_{j=1}^{l} x_j\tau_j=\Sigma_{j=1}^{l} x_j h_j$. Then, from the definition of $I_\varphi^{(k)}(x)$ one immediately obtains

$$(\partial_j I_\varphi^{(k)})(x_1,\ldots,x_l,0,...,0)=k\ \mathrm{Tr}(\varphi(H)^{k-1}\varphi(\tau_j))=k\sum_\lambda \langle\lambda|\varphi(H)^{k-1}\varphi(\tau_j)|\lambda\rangle, \tag{28}$$

where $\{|\lambda\rangle\}$ is an orthonormal basis in W_φ, formed by weight vectors of $\varphi(\mathscr{H})$.

Owing to the form of the τ_j for $j=l+1,\ldots,n$, specified in (20), each term in (28) vanishes, since the weights of the vectors $|\lambda\rangle$ and $\varphi(H)^{k-1}\varphi(e_{\pm\alpha})|\lambda\rangle$ are different for every $j=l+1,\ldots,n$.

Theorem III.1: *Let* $\{p(x)\}=\{p_1(x),\ldots,p_q(x)\}$ *be a MIB for the linear group* Γ, *which is a real orthogonal realization of the adjoint representation. Then the* $\hat{P}$*-matrix* $\hat{P}^\Gamma(p)$ *associated to the MIB* $\{p(x)\}$ *coincides with the* $\hat{P}$*-matrix* $\hat{P}^W(p)$, *associated to the MIB* $\{p^{(\mathscr{H})}(\xi)\}=\{p_1^{(\mathscr{H})}(\xi),\ldots,p_q^{(\mathscr{H})}(\xi)\}$, *for the Weyl group W.*

Proof: By definition

$$\hat{P}^\Gamma_{ab}(p)|_{p=p(x)}=\sum_1^n{}_i\ \partial_i p_a(x)\partial_i p_b(x) \tag{29}$$

and

$$\hat{P}^W_{ab}(p)|_{p=p^{(\mathscr{H})}(\xi)}=\sum_1^l{}_i\ \partial_i p_a^{(\mathscr{H})}(\xi)\partial_i p_b^{(\mathscr{H})}(\xi). \tag{30}$$

Now, if the p_a's are chosen out of the set of invariants $I_\varphi^{(k)}$, because of Lemma III.1, we have

$$\begin{aligned}\hat{P}^\Gamma_{ab}(p)|_{p=p(\xi_1,\ldots,\xi_l,0,\ldots,0)}&=\sum_1^n{}_i\ (\partial_i p_a)(\xi_1,\ldots,\xi_l,0,...,0)(\partial_i p_b)(\xi_1,\ldots,\xi_l,0,...,0)\\&=\sum_1^l{}_i\ \partial_i p_a^{(\mathscr{H})}(\xi)\partial_i p_b^{(\mathscr{H})}(\xi)=\hat{P}^W_{ab}(p)|_{p=p(\xi_1,\ldots,\xi_l,0,\ldots,0)}.\end{aligned} \tag{31}$$

The equality just derived shows that $\hat{P}^\Gamma(p)=\hat{P}^W(p)$, for all $p\in\{p^{(\mathscr{H})}(\xi|\xi\in\mathbb{R}^l)\}$, which is a q-dimensional semialgebraic set of the space $\mathbb{R}^q$ spanned by $(p_1,\ldots,p_q)$. The equality $\hat{P}^\Gamma(p)=\hat{P}^W(p)$ holds true, therefore, for all $p\in\mathbb{R}^q$.

IV. REAL REPRESENTATIONS OF COMPACT LIE GROUPS

As claimed, one of our purposes, which is, in fact, a preliminary problem to solve, is to obtain the *real coregular* representations of *compact* simple Lie groups from Schwarz's classification of the *complex coregular* representations of *complex* simple Lie groups.

In this section, we shall characterize the complex representations of complex simple Lie groups subducing real representations for their maximal compact subgroups. We shall begin by recalling some more or less known results concerning the subject.

Let G be a compact Lie group. The connected component of the identity of G, G^0, is a real analytic group. Then G and G^0 share the same (real) Lie algebra $\mathscr{G}$.

The quotient group G/G^0 is a finite group and its order $|G/G^0|$ coincides with the number of connected components of G.

A *universal complexification* (Ref. 45, p. 204) of a real analytic Lie group G^0 is a continuous homomorphism γ of G^0 into a complex analytic group ${G^0}_c$ having the following property: for every continuous homomorphism η of G^0 into a complex analytic group H, there is one and only one complex analytic homomorphism η_c of ${G^0}_c$ into H such that $\eta_c\circ\gamma=\eta$.

The complexification $G^0{}_c$ of G^0, is so defined that its Lie algebra is the *complexification* $\mathcal{G}_c=\mathcal{G}+i\mathcal{G}$ of $\mathcal{G}$.

The *complexification* G_c of G is the Lie group defined as follows:

$$G_c=(G/G^0)\cdot G^0{}_c=G\cdot G^0{}_c.$$

The group $G^0{}_c$ is the connected component of the identity of G_c and G_c has the same number $|G/G^0|$ of connected components as G.

The following theorem states the connection between G and its complexification G_c.

Theorem IV.1:[45] *The group G_c is the smallest complex Lie group that contains G, which is a maximal compact subgroup of G_c. All the maximal compact subgroups of G_c are conjugated in G_c. Vice versa, every complex semi-simple Lie group is the complexification of a compact semi-simple Lie group.*

There is a one-to-one correspondence between the complex analytic representations of G_c and the representations of G. In fact, a complex analytic representation φ of G_c in the complex vector space V_φ defines obviously a, generally complex, representation of G. Vice versa, let φ be a representation of the compact Lie group G in a real or complex vector space W_φ. The representation φ of the group G defines uniquely a representation of the real Lie algebra $\mathcal{G}$ of G in W_φ (which we shall also denote by φ). The representation φ of $\mathcal{G}$ in W_φ defines uniquely a complex representation of $\mathcal{G}_c$ (also denoted by φ) in the complex vector space $V_\varphi=\mathbb{C}\otimes W_\varphi$:

$$\varphi(\mathcal{G}_c)=\varphi(\mathcal{G}+i\mathcal{G})=\varphi(\mathcal{G})+i\varphi(\mathcal{G}).$$

Analytic representations of $G^0{}_c$ and of G_c, which will still be denoted by φ, are then obtained from $\varphi(\mathcal{G}_c)$ by exponentiation,

$$\varphi(G^0{}_c)=\exp\ \varphi(\mathcal{G}_c),$$

and by defining

$$\varphi(G_c)=\varphi(G)\cdot\varphi(G^0{}_c).$$

With regard to their reality properties the irreducible complex matrix representations of a compact group G can be collected into three disjoint classes:[46,47]

Type 1: representations which admit a *real form*, that is representations which are equivalent to real matrix representations;

Type 2: representations φ which are equivalent to their complex conjugate representations $\overline{\varphi}$, but do not admit a real form;

Type 3: representations which are not equivalent to their complex conjugate representations.

Let us recall that the complex conjugate $\overline{\varphi}$ of a unitary matrix representation φ coincides with the contragradient representation (or star representation) φ^*:

$$\overline{\varphi}(g)=\varphi^T(g)^{-1}=\varphi^*(g),\quad \forall g\in G. \tag{32}$$

A complex representation of a group can be transformed into a real representation through the standard process of *realification*. There is, in fact, a one-to-one correspondence between points and matrices acting in $\mathbb{C}^n$ and points and matrices acting in the real space $\mathbb{R}^{2n}$, defined by

$$\begin{aligned}&\mathbb{C}^n\ni z=(z_1,\ldots,z_n)\leftrightarrow\xi=(\Re z_1,\ldots,\Re z_n,\Im z_1,\ldots,\Im z_n)\in\mathbb{R}^{2n}\\&\mathrm{GL}(n,\mathbb{C})\ni M\leftrightarrow M_R=\begin{pmatrix}M_1 & -M_2\\ M_2 & M_1\end{pmatrix}\in\mathrm{GL}(2n,\mathbb{R}),\end{aligned} \tag{33}$$

where M_1 and M_2 are the real and imaginary parts of M. In this way, starting from a complex matrix representation of $G{:}g\to\varphi(g)$, one gets a real matrix representation of $G{:}g\to\varphi_R(g)$, which is called the realification of φ.

The representation φ_R is orthogonal if and only if φ is unitary. Moreover, thinking of φ_R as acting on the complexification $\mathbb{C}^{2n}$ of $\mathbb{R}^{2n}$, φ_R is equivalent to the direct sum $\varphi \oplus \overline{\varphi}$, the equivalence being induced by the matrix

$$T = \frac{1}{\sqrt{2}} \begin{pmatrix} I_n & \imath I_n \\ I_n & -\imath I_n \end{pmatrix}.$$

Therefore, φ_R is irreducible on the reals if φ is a complex irreducible representation of type 2 or 3; it is reducible if φ is of type 1.

Since every real matrix representation of G can be decomposed into a direct sum of irreducible (on the reals) real matrix representations, the following proposition will be sufficient to characterize all the real representations of G in terms of irreducible complex representations.

Proposition IV.1: Let φ a real irreducible (on the reals) representation of a compact group G. Then either φ is irreducible on the complexes and type 1, or it is the realification of a type 2 or type 3 complex irreducible representation.

The types of the irreducible unitary representations of compact simple Lie groups were determined by Mehta and Srivastava.[48,49] Our notations will be different from the notations used in Refs. 48 and 49; we prefer in fact to order the roots in the Dynkin diagrams as in Ref. 24. To aid the reader, in our tables we shall also report the dimensions of the representations.

V. POLYNOMIAL INVARIANTS OF COMPACT LIE GROUPS AND OF THEIR COMPLEXIFICATIONS

In this section, following the line sketched in Ref. 36 and developed in Ref. 50, we shall determine the relationship between the G-invariant polynomial functions, defined in a real vector space W_φ, hosting a real representation φ of a compact Lie group G, and the G_c-invariant polynomial functions, defined in the complex vector space $V_\varphi = \mathbb{C} \otimes W_\varphi$, hosting the representation $\varphi(G_c)$ of the complexification G_c of G.

We shall need some results and standard techniques from algebraic group theory and classical invariant theory that we shall report below, trying to be as self-contained as possible, without entering too many details.These can be found, for instance, in Refs. 51–54.

Let K be the field of complex or real numbers. The set $K^n = K \times K \times \cdots \times K$ is called the *affine n-space* and denoted by A^n and the sets of common zeros in A^n of finite collections of polynomials are called *affine varieties*. The affine varieties are the closed sets in the natural topology of A^n, the *Zariski topology*.

A closed set X in the Zariski topology is a closed set in the usual Haussdorf T_2 topology in A^n. The converse does not hold true, in general.

If X is Zariski closed in A^n and T denotes a n-ple of indeterminates, a polynomial $f(T) \in K[T]$ defines a K-valued function on X: $x \mapsto f(x)$. Different polynomials need not define different functions; in fact, denoting by $\mathcal{I}(X)$ the ideal of the polynomials vanishing on X, the polynomial functions on X are in a one-to-one correspondence with the elements of the residue class ring $K[T]/\mathcal{I}(X) \equiv K[X]$, which is called the *affine algebra* on X.

Let $X \subset A^n$ and $Y \subset A^m$ be arbitrary affine varieties. A mapping $\psi: X \mapsto Y$, $\psi(x_1,\ldots,x_n) = (\psi_1(x),\ldots,\psi_m(x))$, is called a *morphism* if $\psi_i \in K[X]$, $i = 1,\ldots,m$.

Let G be an algebraic variety, endowed with a group structure such that the maps $(x,y) \mapsto x \cdot y$ and $x \mapsto x^{-1}$ are morphisms. Then G is called an (*affine*) *algebraic group*.

The group $\mathrm{GL}(n,K)$ of the $n \times n$ regular matrices over K is endowed with a structure of algebraic group if its elements M are thought of as points of coordinates $(M_{11}, M_{12}, \ldots, M_{nn}, \Delta)$ in K^{n^2+1}, lying on the affine variety defined by the equation $\Delta \cdot \det(M) - 1 = 0$.

By definition, a *linear algebraic group* is a Zarisky closed subgroup of the algebraic group $\mathrm{GL}(n,K)$.

The Zarisky closure $\mathrm{cl}(G)$ in K^{n^2+1} of any subgroup G of $\mathrm{GL}(n,K)$ is a linear algebraic group (Ref. 54, Proposition A, p. 54).

Every affine algebraic group defined over K is isomorphic to an algebraic linear group defined over K (Ref. 51, Proposition 1.10, p. 101). Therefore, in the following it will not be restrictive to identify algebraic groups with linear algebraic groups.

Well known examples of linear algebraic groups are $\mathrm{SL}(n,K)$, $\mathrm{SO}(n,K)$, and $\mathrm{Sp}(2n,K)$.

Theorem V.1 *(Ref.* 55, *Theorem 2.1.2, p. 45): Every complex (real) linear algebraic group is a complex (real) Lie group.*

The converse of Theorem V.1 does not hold true; the real Lie group of complex matrices $\mathrm{SU}(n)$, for instance, is not a linear algebraic group.

By definition, a representation φ of a group G is an *analytic* (C^∞) homomorphism, if G is a Lie group, and a *rational* (polynomial) homomorphism, if G is algebraic. As a consequence, we have the following:

Theorem V.2: *A representation φ of a (complex or real) Lie group G defines a (complex or real) Lie group of matrices $\varphi(G)$. A representation φ of a complex algebraic group G defines a linear complex algebraic group $\varphi(G)$ (Ref.* 53, *p. 122).*

The second part of the theorem is not valid for real linear algebraic groups.

Since a complex linear algebraic group is also a Lie group (Theorem V.1), it admits both rational and analytic representations.

A complex Lie group (admitting at least one complex analytic representation) is said to be *reductive* if all its complex analytic representations are completely reducible. As is well known, every finite dimensional complex analytic representation of a compact Lie group G is completely reducible. The complexification G_c of G shares with G this property, so G_c is reductive. Vice versa, every reductive complex Lie group is the complexification of a compact Lie group (Ref. 56, th. 3.1). Therefore, we have the following:

Theorem V.3: *A complex Lie group is reductive if and only if it is the complexification of a compact Lie group.*

Theorem V.4: *Every complex reductive Lie group can be identified with a complex linear algebraic group so that its complex analytic representations coincide with its rational representations.*

Proof: A compact group has at the most a finite number of connected components. By Theorem V.3, this property is shared by all complex reductive Lie groups. The thesis is then obtained from Ref. 57, th. 9.2 and th. 9.5. □

We need now Theorems V.1, V.2, V.3 and V.4 to obtain the following theorem:

Theorem V.5: *Let G be a compact Lie group and G_c be its complexification. Then*

1. G is Zariski dense in G_c.

2. For every analytic representation φ of G, or of G_c, $\varphi(G)$ is Zariski dense in $\varphi(G_c)$.

Proof:

(1) By Theorems V.3 and V.4 we can identify the complex Lie group G_c with a complex algebraic group containing G and, consequently, also $\mathrm{cl}(G)$, which is the smallest algebraic group containing G. But, by Theorem V.1, $\mathrm{cl}(G)$ is a Lie group, so $\mathrm{cl}(G) \supseteq G_c$, since G_c is the smallest complex Lie group containing G. This is sufficient to conclude that $\mathrm{cl}(G) = G_c$.

(2) As remarked in Section IV, every analytic representation φ of G can be extended to a complex analytic representation φ of G_c and, by Theorem V.4, every complex analytic representation of G_c is a rational representation. Then, by Theorem V.2, $\varphi(G_c)$ is both a complex Lie group of matrices and a complex linear algebraic group, and $\varphi(G)$ is a compact Lie group of complex matrices. Its Zariski closure $\mathrm{cl}(\varphi(G))$ is a complex linear algebraic group contained in $\varphi(G_c)$. Being φ Zarisky continuous, $\varphi^{-1}(\mathrm{cl}(\varphi(G)))$ is a Zarisky closed subgroup of G_c containing G. It must therefore coincide with G_c, because of the first part of the Theorem G is Zariski dense in G_c. This implies that $\mathrm{cl}(\varphi(G)) \equiv \varphi(G_c)$. □

Theorem V.6: *Let V be a complex vector space on which an arbitrary group $G \subseteq GL(V)$ acts and let* $\mathrm{cl}(G)$ *be the Zariski closure of G. Then, the ring of polynomial G-invariant functions defined on V and the ring of polynomial* $\mathrm{cl}(G)$*-invariant functions defined on V coincide, that is,*

$$\mathbb{C}[V]^G \equiv \mathbb{C}[V]^{\mathrm{cl}(G)}.$$

Proof (Ref. 58, p. 347): Since $\mathrm{cl}(G) \supseteq G$, the relation $\mathbb{C}[V]^{\mathrm{cl}(G)} \subseteq \mathbb{C}[V]^G$ is obviously true. Let us prove the converse relation $\mathbb{C}[V]^G \subseteq \mathbb{C}[V]^{\mathrm{cl}(G)}$. To this end, for every $p_1 \in \mathbb{C}[V]^G$ let $R_1 = \{g \cdot p_1,\ g \in \mathrm{cl}(G)\}$, with $(g \cdot p_1)(x) = p_1(g^{-1} \cdot x), \forall x \in V$. The degree of each of the polynomials $g \cdot p_1$, $\forall g \in \mathrm{cl}(G)$ does not exceed the degree of p_1, thus R_1 is a finite dimensional subspace of $\mathbb{C}[V]$. Let $p_1, \ldots, p_t$, be a basis of R_1. The action of $\mathrm{cl}(G)$ in R_1 is linear and in the

basis $\{p_i\}$ every $g \in \mathrm{cl}(G)$ is represented by a matrix $\varphi(g) \in \mathrm{GL}(t,\mathbb{C})$. By Theorem V.2, $\varphi(\mathrm{cl}(G))$ is a complex linear algebraic group. The largest subgroup H of $\varphi(\mathrm{cl}(G))$ leaving p_1 invariant is a complex linear algebraic group, since it is defined by algebraic equations ($H=\{\varphi(g), g \in \mathrm{cl}(G) | \varphi(g)p_i = \delta_{i1}p_1\}$). Obviously $H \supseteq \varphi(G)$. Since φ is Zarisky continuous, $\varphi^{-1}(H)$ is Zarisky closed, therefore $G \subseteq \varphi^{-1}(H) \subseteq \mathrm{cl}(G)$, which immediately implies $\varphi^{-1}(H) = \mathrm{cl}(G)$, that is, $\varphi(\mathrm{cl}(G)) = H$. □

Theorems V.5 and V.6 enable us to reduce the determination of the G-invariant polynomial functions, defined on complex representation spaces of G, to the determination of the G_c-invariant polynomial functions defined in the same space. This result can be summarized in the following way:

Theorem V.7: *Let G be a compact Lie group and G_c be its complexification. Then for every complex analytic representation φ of G_c in a complex vector space V_φ, there is identity between the G-invariant and the G_c-invariant polynomial functions defined on V_φ:*

$$\mathbb{C}[V_\varphi]^{\varphi(G)} \equiv \mathbb{C}[V_\varphi]^{\varphi(G_c)}.$$

An obvious consequence is the following:

Corollary V.8: The polynomials $p_1,\ldots,p_q$ form an integrity basis for $\mathbb{C}[V_\varphi]^{\varphi(G)}$ if and only if they form an integrity basis for $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$.

Let us now concentrate our attention on the real representations of compact Lie groups.

Let W be a real vector space and V its complexification. Since a polynomial function is essentially determined by its coefficients, the following identity holds true:

$$\mathbb{C}[W] = \mathbb{C} \otimes \mathbb{R}[W],$$

and there is a natural isomorphism

$$K[W] \simeq K[V], \quad K = \mathbb{R}, \mathbb{C}.$$

Theorem V.9: *Let G be a compact Lie group and G_c be its complexification. Then, for every analytic representation φ of G in the real space W_φ and of G_c in the complex space $V_\varphi = \mathbb{C} \otimes W_\varphi$, the following relation holds true:*

$$\mathbb{C}[V_\varphi]^{\varphi(G_c)} \simeq \mathbb{C} \otimes \mathbb{R}[W_\varphi]^{\varphi(G)}.$$

Proof: If W_φ is the real representation space of a compact Lie group G and V_φ is its complexification. Then from Theorem V.7,

$$\mathbb{C}[V_\varphi]^{\varphi(G)} = \mathbb{C}[V_\varphi]^{\varphi(G_c)}.$$

Moreover,

$$\mathbb{C}[V_\varphi]^{\varphi(G)} \simeq \mathbb{C}[W_\varphi]^{\varphi(G)} = (\mathbb{C} \otimes \mathbb{R}[W_\varphi])^{\varphi(G)} = \mathbb{C} \otimes \mathbb{R}[W_\varphi]^{\varphi(G)},$$

where the last equality is a consequence of the reality of $\varphi(G)$, implying that both the real and imaginary parts of a complex $\varphi(G)$-invariant function are $\varphi(G)$-invariant. □

All polynomial invariant functions in $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ can then be expressed as polynomials with complex coefficients of a polynomial basis for $\mathbb{R}[W_\varphi]^{\varphi(G)}$, with the following obvious consequences:

Corollary V.10: If the polynomial functions $p_1,\ldots,p_q$ form an integrity basis for $\mathbb{R}[W_\varphi]^{\varphi(G)}$, they form also an integrity basis for $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$.

Corollary V.11: Let φ be an analytic representation of G_c in the complex space $V_\varphi = \mathbb{C} \otimes W_\varphi$, admitting a real form as a representation of the maximal compact subgroup G of G_c. Then the ring $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ admits integrity bases formed by polynomials with real coefficients (real integrity bases).

To conclude, we have the following.

Theorem V.12: *Let G be a compact Lie group, G_c be its complexification and φ be a real analytic representation of G in W_φ. Then, $\mathbb{R}[W_\varphi]^{\varphi(G)}$ is a regular ring if and only if $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ is a regular ring.*

TABLE I. Irreducible real representations of simple compact Lie groups.

$\mathscr{G}$ and or G	Rank		Conditions on the highest weight $\mathbf{\Lambda}=(\Lambda_1,\Lambda_2,\ldots)$
A_n, SU$(n+1)$	$n\geqslant 1$	$n=0,2,3 \pmod 4$	$\Lambda_i=\Lambda_{n-i+1}$, $\forall i=1,\ldots,n$
		$n=1 \pmod 4$	$\Lambda_i=\Lambda_{n-i+1}$, $\forall i=1,\ldots,n$, and $\Lambda_{(n+1)/2}$ even
B_n, SO$(2n+1)$	$n\geqslant 2$	$n=0,3 \pmod 4$	none
		$n=1,2 \pmod 4$	Λ_n even
C_n, Sp$(2n)$	$n\geqslant 3$		$(\Lambda_1+\Lambda_3+\Lambda_5+\ldots)$ even
D_n, SO$(2n)$	$n\geqslant 4$	$n=0 \pmod 4$	none
		$n=1,3 \pmod 4$	$\Lambda_{n-1}=\Lambda_n$
		$n=2 \pmod 4$	$(\Lambda_{n-1}+\Lambda_n)$ even
E_6	6		$\Lambda_1=\Lambda_5$ and $\Lambda_2=\Lambda_4$
E_7	7		$(\Lambda_1+\Lambda_3+\Lambda_7)$ even
E_8	8		none
F_4	4		none
G_2	2		none

Proof: According to Corollary V.10, let $p_1,\ldots,p_q$ be a real integrity basis for both $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ and $\mathbb{R}[W_\varphi]^{\varphi(G)}$. If $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ is not regular, there exists a complex polynomial, $f(T)$, in the indeterminates $T_1,\ldots,T_q$, such that $f(p(z))=f(p_1(z),\ldots,p_q(z))=0,\forall z\in V_\varphi$. Then $f(p(x))=0,\forall x\in W_\varphi$ and, if f_1 and f_2 are the real and imaginary parts of f, also $f_1(p(x))=0=f_2(p(x)),\forall x\in W_\varphi$. This proves that $\mathbb{R}[W_\varphi]^{\varphi(G)}$ is not regular.

If the ring $\mathbb{R}[W_\varphi]^{\varphi(G)}$ is not regular, then there exist real polynomials, $f(T)$, in the indeterminates $T_1,\ldots,T_q$, such that $f(p(x))=0,\forall x\in W_\varphi$. As a consequence, $f(p(z))=0,\forall z\in V_\varphi$, so that the ring $\mathbb{C}[V_\varphi]^{\varphi(G_c)}$ is not regular. □

VI. REAL COREGULAR REPRESENTATIONS OF COMPACT SIMPLE LIE GROUPS

Theorem V.12 states the link enabling us to determine all the real coregular representations of compact simple Lie groups from Schwarz's classification of the complex coregular representations of complex simple Lie groups.[24] What one has to do is simply to select, for each simple complex Lie group G_c, those complex representations which are complexifications of real representations of its maximal compact subgroup G.

The conditions, which have to be satisfied by the weights of a complex irreducible matrix representation φ of a complex simple Lie group G_c, with maximal compact subgroup G, so that $\varphi(G)$ admits a real form, can be found, for example, in Refs. 59, 48, and 49. The relevant results are summarized in Tables I and II.

The irreducible representations of highest weight $\Lambda=(\Lambda_1,\Lambda_2,\ldots)$ will be denoted by the symbol ${\varphi_1}^{\Lambda_1}{\varphi_2}^{\Lambda_2}\ldots$ and ${\varphi_i}^{\Lambda_i}$ will be omitted when $\Lambda_i=0$ and shortened to φ_i when $\Lambda_i=1$.

The following theorem will make our task easier, as it assures the existence of *maximally coregular representations*.

TABLE II. Contragradient fundamental representations of simple complex Lie groups.

$\mathscr{G}$	Rank	$\varphi\leftrightarrow\varphi^*$	
A_n	$n\geqslant 1$	$\varphi_i\leftrightarrow\varphi_{n-i+1}$	$i=1,\ldots,n$
B_n	$n\geqslant 2$	$\varphi_i\leftrightarrow\varphi_i$	$i=1,\ldots,n$
C_n	$n\geqslant 3$	$\varphi_i\leftrightarrow\varphi_i$	$i=1,\ldots,n$
D_n	$n\geqslant 4$, n odd	$\varphi_i\leftrightarrow\varphi_i$, $\varphi_{n-1}\leftrightarrow\varphi_n$	$i=1,\ldots,n-2$
D_n	$n\geqslant 4$, n even	$\varphi_i\leftrightarrow\varphi_i$	$i=1,\ldots,n$
E_6	6	$\varphi_i\leftrightarrow\varphi_{6-i}$, $\varphi_6\leftrightarrow\varphi_6$	$i=1,\ldots,5$
E_7	7	$\varphi_i\leftrightarrow\varphi_i$	$i=1,\ldots,7$
E_8	8	$\varphi_i\leftrightarrow\varphi_i$	$i=1,\ldots,8$
F_4	4	$\varphi_i\leftrightarrow\varphi_i$	$i=1,2,3,4$
G_2	2	$\varphi_i\leftrightarrow\varphi_i$	$i=1,2$

Theorem VI.1: *Every proper subrepresentation φ' of a coregular representation, φ, of a reductive group G is coregular. The basic invariants for $\varphi'(G)$ can be chosen to be a proper subset of the basic invariants for $\varphi(G)$.*

Proof: Let $p_1,\ldots,p_q$ be a MIB for $\varphi(G)$ and $p'_1,\ldots,p'_{q'}$ be a MIB for $\varphi'(G)$. The proof of the Theorem stems from the observation that if $\varphi=\varphi'\oplus\varphi'_\perp$, all the elements of a MIB for $\varphi'(G)$ and for $\varphi'_\perp(G)$ can be included as elements of a MIB for $\varphi(G)$. With this choice, any algebraic relation among the elements of the MIB for $\varphi'(G)$ would be an algebraic relation among the elements of the MIB for $\varphi(G)$. □

Let us illustrate the standard procedure we followed to obtain our results, working out, step by step, a particular case.

Let us start, for instance, from Entry 3 in Table 1a of Ref. 24, where the following items are found:

(1) The symbol $\mathrm{SL}(n,\mathbb{C})$, $n\geqslant 4$ of the class of simple complex Lie groups one is dealing with;
(2) the decomposition $\varphi=3\varphi_1+\varphi_2+(n-2)\varphi_{n-1}$ of the maximally coregular complex representation φ of each group $\mathrm{SL}(n,\mathbb{C})$ in complex irreducible sub-representations;
(3) the number $\frac{1}{2}n(n+1)+1$ of elements in a MIB for the linear group $\varphi(\mathrm{SL}(n,\mathbb{C}))$;
(4) a specification of the structure of a MIB: $(3n-6)\langle 2\rangle,4\langle 3\rangle,\frac{1}{2}(n-2)(n-3)\langle 4\rangle$, meaning that there are $(3n-6)$ invariants of type $\langle 2\rangle$, 4 invariants of type $\langle 3\rangle$ and so on; the invariant of type $\langle k\rangle$ is roughly specified in Entry k of Table 1b of Ref. 24, from which one can also recover its degree.

The maximal compact subgroup of $\mathrm{SL}(n,\mathbb{C})$ is $\mathrm{SU}(n)$. Let us examine the reality properties of the sub-representations φ_1, φ_2, φ_{n-1}, contained in $\varphi(\mathrm{SU}(n))$.

(i) φ_1 and φ_{n-1}, for $n\geqslant 4$ (see Table I and Table II for A_{n-1}) are not of type 1, but are mutually contragradient, so $\varphi_{n-1}=\varphi_1^*=\overline{\varphi_1}$. The representations $k(\varphi_1+\varphi_{n-1})$ admit therefore a real form: the realification of $k\varphi_1$.
(ii) φ_2 (see Table I for A_{n-1}) is of type 1 only for $n=4$.

From Theorem VI.1, therefore, the representation $\varphi(\mathrm{SU}(n))$ contains the following coregular real representations, with the MIB's specified aside:

$\mathrm{SU}(n)$,	$n\geqslant 4$	$\varphi_1+\varphi_{n-}$	$1\langle 2\rangle$	(degree 2)	
$\mathrm{SU}(n)$,	$n\geqslant 4$	$2(\varphi_1+\varphi_{n-1})$	$4\langle 2\rangle$	(degree 2)	
$\mathrm{SU}(n)$,	$n\geqslant 5$	$3(\varphi_1+\varphi_{n-1})$	$9\langle 2\rangle$	(degree 2)	
$\mathrm{SU}(4)$		φ_2	$1\langle 3\rangle$	(degree 2)	
$\mathrm{SU}(4)$		$\varphi_1+\varphi_2+\varphi_3$	$1\langle 2\rangle$,	$1\langle 3\rangle$	(degree 2)
$\mathrm{SU}(4)$		$2\varphi_1+\varphi_2+2\varphi_3$	$4\langle 2\rangle$,	$1\langle 3\rangle$	(degree 2)
			$2\langle 3\rangle$	(degree 3)	

Among these, only the representations $\varphi_1+\varphi_{n-1}$, for $\mathrm{SU}(n)$, and φ_2, for $\mathrm{SU}(4)$, are irreducible on the reals.

The invariants listed in the last column have been obtained through a selection of the invariants reported in Schwarz's tables, made in such a way that the invariants can be constructed out of the representation spaces which are available. For instance, for $n=4$, the invariant of type $\langle 3\rangle$ in Schwarz's tables is constructed out of i vectors of the same space, V_{φ_2}, and $4-2i$ vectors of $4-2i$ different copies of spaces V_{φ_1}, with $2\leqslant 2i\leqslant 4$. Thus, in the space of the representation $2\varphi_1+\varphi_2+2\varphi_3$ of $\mathrm{SU}(4)$ it is possible to construct two invariants of type $\langle 3\rangle$ with $i=1$ (degree 3) (one using φ_1 and one using φ_3; in fact in this particular case φ_1 and φ_3 are related by an isomorphism of the Dynkin diagram of A_3), and one invariant of type $\langle 3\rangle$ with $i=2$ (degree 2). (One of the two invariants of type $\langle 3\rangle$ (degree 3) coincides also with an invariant of type $\langle 4\rangle$.)

The results we have obtained following the standard procedure described in the example are summarized in Tables III and IV, where all the real coregular representations of simple compact Lie groups are listed.

TABLE III. Real coregular representations of compact simple Lie groups (I). In each row, the following items are listed: A simple compact Lie group, G, specified by the Lie algebra of its complexification; a real coregular representations, φ, of G; the real dimensions, dim, and the reducibility properties, i/r, of φ; the indication, Source, of the Table and the Entry of Ref. 24, where φ is quoted; the number, q, and the degrees, d_i, of the basic invariants of (G,φ).

G	φ	dim	i/r	Source	q	d_i
A_1	$2\varphi_1$	4	i	1a.1	1	2
	$\varphi_1{}^2=\mathrm{Ad}$	3	i	1a.12	1	2
	$2\varphi_1{}^2$	3+3	r	1a.18	3	2,2,2
	$\varphi_1{}^4$	5	i	2a.2	2	3,2
$A_{n\geqslant 2}$	$k(\varphi_1+\varphi_n),k\leqslant n$	$k\cdot 2(n+1)$	i/r	1a.2	k^2	$k^2\times 2$
	$\varphi_1\varphi_n=\mathrm{Ad}$	$(n+1)^2-1$	i	1a.20	n	$i\ \ (i=2,\ldots,n+1)$
	$\varphi_1^2+\varphi_n^2$	$(n+1)(n+2)$	i	1a.19	$n+2$	$2\times(n+1),2i\ \ (i=1,\ldots,n)$
A_3	$k\varphi_2$	$k\cdot 6$	i/r	3a.1	$\frac{k(k+1)}{2}$	$\frac{k(k+1)}{2}\times 2$
	$\varphi_1+\varphi_2+\varphi_3$	8+6	r	1a.3	2	2,2
	$2\varphi_1+\varphi_2+2\varphi_3$	8+8+6	r	1a.3	7	$2\times 3,5\times 2$
	$\varphi_1+2\varphi_2+\varphi_3$	8+6+6	r	1a.7	5	$4,4\times 2$
	$\mathrm{Ad}+\varphi_2$	15+6	r	3a.2	6	6,4,4,3,2,2
	φ_2^2	20	i	3a.3	5	6,5,4,3,2
$A_{n\geqslant 4,\text{even}}$	$\varphi_2+\varphi_{n-1}$	$n(n+1)$	i	1a.10	$\frac{n}{2}$	$2i\ \left(i=1,\ldots,\frac{n}{2}\right)$
	$\varphi_1+\varphi_2+\varphi_{n-1}+\varphi_n$	$2(n+1)+n(n+1)$	r	1a.11	$n+2$	$2,2\times\frac{n+2}{2},2i\ \left(i=1,\ldots,\frac{n}{2}\right),$ $2i+2\ \left(i=1,\ldots,\frac{n-2}{2}\right)$
$A_{n\geqslant 5,\text{ odd}}$	$\varphi_2+\varphi_{n-1}$	$n(n+1)$	i	1a.10	$\frac{n+3}{2}$	$2\times\frac{n+1}{2},2i\ \left(i=1,\ldots,\frac{n-1}{2}\right)$
	$\varphi_1+\varphi_2+\varphi_{n-1}+\varphi_n$	$2(n+1)+n(n+1)$	r	1a.11	$n+2$	$2,2\times\frac{n+1}{2},$ $2i+2,2i\ \left(i=1,\ldots,\frac{n-1}{2}\right)$
A_5	$2\varphi_3$	40	i	2a.12	7	$6,5\times 4,2$
A_7	φ_4	70	i	2a.19	7	18,14,12,10,8,6,2
E_6	$\varphi_1+\varphi_5$	54	i	5a.6	4	4,3,3,2
	$\varphi_6=\mathrm{Ad}$	78	i	5a.7	6	12,9,8,6,5,2
E_7	$2\varphi_1$	112	i	5a.8	7	$6,5\times 4,2$
	$\varphi_6=Ad$	133	i	5a.9	7	18,14,12,10,8,6,2
E_8	$\varphi_1=\mathrm{Ad}$	248	i	5a.10	8	30,24,20,18,14,12,8,2
F_4	φ_1	26	i	5a.3	2	3,2
	$2\varphi_1$	$2\cdot 26$	r	5a.3	8	$4,4\times 3,3\times 2$
	$\varphi_4=\mathrm{Ad}$	52	i	5a.4	4	12,8,6,2
G_2	φ_1	7	i	5a.1	1	2
	$2\varphi_1$	7+7	r	5a.1	3	2,2,2
	$3\varphi_1$	$3\cdot 7$	r	5a.1	7	$3,6\times 2$
	$\varphi_2=\mathrm{Ad}$	14	i	5a.2	2	6,2

The representations listed are real, so one can always think of them as realized by real orthogonal matrices. In the column ''Source,'' we have reported the tables and the entries of Ref. 24 where the representations are obtained from. As a rule, only one source entry is indicated, also if it is not the unique one.

In all the tables, to avoid isomorphisms, the ranks are limited in the following way: $B_n, n\geqslant 2$, $C_n, n\geqslant 3$, $D_n, n\geqslant 4$. For A_n permutations of φ_i and φ_{n-i}, $i=1,\ldots,[n/2]$, for D_n permutations of φ_{n-1} and φ_n, and for D_4 permutations of φ_1, φ_3, φ_4, are avoided.

VII. THE $\hat{P}$-MATRICES OF THE REAL COREGULAR SIMPLE COMPACT LIE GROUPS WITH 2, 3, AND 4 BASIC INVARIANTS

In this section we shall identify, among the allowable $\hat{P}$-matrices (abbreviated to A$\hat{P}$M's) reported in Refs. 27 and 28, those generated by the real coregular linear compact simple Lie groups (reported in Tables III and IV) that have less than 5 basic invariants.

TABLE IV. Real coregular representations of compact simple Lie groups (II). In each row, the following items are listed: A simple compact Lie group, G, specified by the Lie algebra of its complexification; a real coregular representation, φ, of G; the real dimensions, dim, and the reducibility properties, i/r, of φ; the indication, Source, of the Table and the Entry of Ref. 24, where φ is quoted; the number, q, and the degrees, d_i, of the basic invariants of (G,φ).

G	φ	dim	i/r	Source	q	d_i
$B_{n\geqslant 2}$	$k\varphi_1, k\leqslant 2n$	$k\cdot(2n+1)$	i/r	3a.4	$\frac{k(k+1)}{2}$	$\frac{k(k+1)}{2}\times 2$
	$\varphi_2=\mathrm{Ad}$	$n(2n+1)$	i	3a.5	n	$2i\ \ (i=1,\ldots,n)$
	$\mathrm{Ad}+\varphi_1$	$n(2n+1)+2n+1$	r	3a.5	$2n+1$	$n+1,2i\ \ (i=1,\ldots,n)$, $2i\ \ (i=1,\ldots,n)$
	φ_1^2	$(n+1)(2n+1)-1$	i	3a.6	$2n$	$i\ \ (i=2,\ldots,2n+1)$
B_3	$k\varphi_3,\ k\leqslant 2n$	$k\cdot 8$	i/r	3a.11	$\frac{k(k+1)}{2}$	$\frac{k(k+1)}{2}\times 2$
	$4\varphi_3$	$4\cdot 8$	r	3a.12	11	$4,10\times 2$
	$\varphi_1+\varphi_3$	$7+8$	r	3a.9	2	2,2
	$2\varphi_1+\varphi_3$	$7+7+8$	r	3a.9	4	2,2,2,2
	$3\varphi_1+\varphi_3$	$3\cdot 7+8$	r	3a.9	8	$5,7\times 2$
	$\varphi_1+2\varphi_3$	$7+2\cdot 8$	r	3a.10	5	$3,4\times 2$
	$2\varphi_1+2\varphi_3$	$2\cdot 7+2\cdot 8$	r	3a.10	9	$4,2\times 3,6\times 2$
	$\varphi_1+3\varphi_3$	$7+3\cdot 8$	r	3a.11	10	$3\times 3,7\times 2$
B_4	φ_4	16	i	3a.17	1	2
	$\varphi_1+\varphi_4$	$9+16$	r	3a.17	3	3,2,2
	$2\varphi_1+\varphi_4$	$2\cdot 9+16$	r	3a.17	6	$2\times 3,4\times 2$
	$3\varphi_1+\varphi_4$	$3\cdot 9+16$	r	3a.17	10	$3\times 3,7\times 2$
	$4\varphi_1+\varphi_4$	$4\cdot 9+16$	r	3a.17	16	$6,4\times 3,11\times 2$
	$2\varphi_4$	$16+16$	r	3a.18	4	4,2,2,2
	$\varphi_1+2\varphi_4$	$9+2\cdot 16$	r	3a.18	9	$4,4\times 3,4\times 2$
	$2\varphi_1+2\varphi_4$	$2\cdot 9+2\cdot 16$	r	3a.18	14	$2\times 4,6\times 3,6\times 2$
	$3\varphi_4$	$3\cdot 16$	r	3a.19	12	$6\times 4,6\times 2$
B_5	$2\varphi_5$	64	i	3a.26	9	$8,6,6\times 4,2$
$D_{n\geqslant 4}$	$k\varphi_1, k<2n$	$k\cdot 2n$	i/r	3a.1	$\frac{k(k+1)}{2}$	$\frac{k(k+1)}{2}\times 2$
	$\mathrm{Ad}+\varphi_1$	$n(2n-1)+2n$	r	3a.2	$2n$	$n,2i,(i=1,\ldots,n-1)$, $2i\ \ (i=1,\ldots,n)$
	φ_1^2	$n(2n+1)-1$	i	3a.3	$2n-1$	$i\ \ (i=2,\ldots,2n)$
	$\varphi_2=\mathrm{Ad}$	$n(2n-1)$	i	3a.2	n	$n,2i\ \ (i=1,\ldots,n-1)$
D_4	$\varphi_1+\varphi_3$	$8+8$	r	3a.13	2	2,2
	$\varphi_1+\varphi_3+\varphi_4$	$8+8+8$	r	3a.15	4	3,2,2,2
	$3\varphi_1+\varphi_3$	$4\cdot 8$	r	3a.13	7	7×2
	$4\varphi_1+\varphi_3$	$5\cdot 8$	r	3a.13	12	$6,11\times 2$
	$\varphi_1+2\varphi_3$	$8+8+8$	r	3a.14	4	2,2,2,2
	$2\varphi_1+2\varphi_3$	$4\cdot 8$	r	3a.14	7	$4,6\times 2$
	$3\varphi_1+2\varphi_3$	$5\cdot 8$	r	3a.14	12	$3\times 4,9\times 2$
	$2\varphi_1+\varphi_3+\varphi_4$	$4\cdot 8$	r	3a.15	7	$2\times 3,5\times 2$
	$3\varphi_1+\varphi_3+\varphi_4$	$5\cdot 8$	r	3a.15	12	$5,3\times 3,8\times 2$
	$2\varphi_1+2\varphi_3+\varphi_4$	$5\cdot 8$	r	3a.16	12	$4,4\times 3,7\times 2$
D_5	$\varphi_4+\varphi_5$	32	i	3a.22	2	4,2
	$\varphi_1+\varphi_4+\varphi_5$	$10+32$	r	3a.22	5	$4,2\times 3,2\times 2$
	$2\varphi_1+\varphi_4+\varphi_5$	$2\cdot 10+32$	r	3a.22	10	$2\times 4,4\times 3,4\times 2$
	$3\varphi_1+\varphi_4+\varphi_5$	$3\cdot 10+32$	r	3a.22	17	$4\times 4,6\times 3,7\times 2$
D_6	$2\varphi_5$	64	i	3a.28	7	$6,5\times 4,2$
D_8	φ_7	128	i	3a.32	8	30,24,20,18,14,12,8,2
$C_{n\geqslant 3}$	$k\cdot 2\varphi_1, k\leqslant n$	$k\cdot 4n$	i/r	4a.1	$k(2k-1)$	$k(2k-1)\times 2$
	φ_2	$n(2n-1)-1$	i	4a.2	$n-1$	$i\ \ (i=2,\ldots,n)$
	$2\varphi_1+\varphi_2$	$4n+n(2n-1)-1$	r	4a.2	$2n-1$	$2,i+1,i\ \ (i=2,\ldots,n)$
	$\varphi_1^2=\mathrm{Ad}$	$n(2n+1)$	i	4a.3	n	$2i\ \ (i=1,\ldots,n)$
C_3	$2\varphi_2$	$2\cdot 14$	r	4a.5	8	$4,4\times 3,3\times 2$
	$2\varphi_3$	28	i	4a.7	7	$6,5\times 4,2$
C_4	φ_4	42	i	4a.8	6	12,9,8,6,5,2

TABLE V. Orbit spaces of real coregular representations of compact simple Lie groups with less than 5 basic invariants. In each row, the following items are listed: The row number, Entry; a simple compact Lie group G, specified by the Lie algebras of its complexification; a real coregular representation, φ, of G; the real dimensions, dim, and the reducibility properties, i/r, of φ; the degrees, d_i, of the basic invariants of (G,φ); the number, n, of nonequivalent classes of allowable $\hat{P}$-matrices with degrees d_i; a label specifying the equivalence class, Class, of the relevant allowable $\hat{P}$-matrix.

Entry	G	φ	dim	i/r	d_i	n	Class
1	A_1	$2\varphi_1$	4	i	2	1	I_1
2		$\varphi_1{}^2=\mathrm{Ad}$	3	i	2	1	I_1
3		$2\varphi_1{}^2$	3+3	r	2,2,2	2	$I(1,1)$
4		$\varphi_1{}^4$	5	i	3,2	1	I_2
5	$A_{n\geq 2}$	$\varphi_1+\varphi_n$	$2(n+1)$	i	2	1	I_1
6		$2\cdot(\varphi_1+\varphi_n)$	$2\cdot 2(n+1)$	r	2,2,2,2	5	$A1(1,1,1,1)$
7	A_2	$\varphi_1\varphi_2=\mathrm{Ad}$	8	i	3,2	1	I_1
8		$\varphi_1^2+\varphi_2^2$	12	i	4,3,3,2	1	$B2$
9	A_3	$\varphi_1\varphi_3=Ad$	15	i	4,3,2	1	$III.1$
10		φ_2	6	i	2	1	I_1
11		$\varphi_1+\varphi_2+\varphi_3$	8+6	r	2,2	1	I_2
12		$2\varphi_2$	6+6	r	2,2,2	2	$I(1,1)$
13	A_4	$\varphi_1\varphi_4=\mathrm{Ad}$	24	i	5,4,3,2	1	$E1$
14		$\varphi_2+\varphi_3$	20	i	4,2	1	I_2
15	A_5	$\varphi_2+\varphi_4$	30	i	4,3,3,2	1	$B2$
16	A_6	$\varphi_2+\varphi_5$	42	i	6,4,2	3	$III.2$
17	A_8	$\varphi_2+\varphi_7$	72	i	8,6,4,2	4	$E3$
18	$B_{n\geq 2}$	φ_1	$2n+1$	i	2	1	I_1
19		$2\varphi_1$	$2\cdot(2n+1)$	r	2,2,2	2	$I(1,1)$
20	B_2	φ_1^2	14	i	5,4,3,2	1	$E1$
21		$\varphi_2=\mathrm{Ad}$	10	i	4,2	1	I_2
22	B_3	$\varphi_2=\mathrm{Ad}$	21	i	6,4,2	3	$III.2$
23		φ_3	8	i	2	1	I_1
24		$\varphi_1+\varphi_3$	7+8	r	2,2	1	I_2
25		$2\varphi_1+\varphi_3$	7+7+8	r	2,2,2,2	5	$A3(1,1,1)$
26		$2\varphi_3$	8+8	r	2,2,2	2	$I(1,1)$
27	B_4	$\varphi_2=\mathrm{Ad}$	36	i	8,6,4,2	4	$E3$
28		φ_4	16	i	2	1	I_1
29		$\varphi_1+\varphi_4$	9+16	r	3,2,2	2	$I(2,1)$
30		$2\varphi_4$	16+16	r	4,2,2,2	10	$A1(1,1,2,2)$
31	$D_{n\geq 4}$	φ_1	$2n$	i	2	1	I_1
32		$2\varphi_1$	$2n+2n$	r	2,2,2	2	$I(1,1)$
33	D_4	$\varphi_2=\mathrm{Ad}$	28	i	6,4,4,2	7	$E2$
34		$\varphi_1+\varphi_3$	8+8	r	2,2	1	I_2
35		$\varphi_1+\varphi_3+\varphi_4$	8+8+8	r	3,2,2,2	6	$A2(1,1,1,1)$
36		$\varphi_1+2\varphi_3$	8+8+8	r	2,2,2,2	5	$A3(1,1,1)$
37	D_5	$\varphi_4+\varphi_5$	32	i	4,2	1	I_2
38	C_3	$2\varphi_1$	12	i	2	1	I_1
39		$\varphi_1^2=\mathrm{Ad}$	21	i	6,4,2	3	$III.2$
40		φ_2	14	i	3,2	1	I_2
41	C_4	$\varphi_1^2=\mathrm{Ad}$	36	i	8,6,4,2	4	$E3$
42		φ_2	27	i	4,3,2	1	$III.1$
43	C_5	φ_2	44	i	5,4,3,2	1	$E1$
44	E_6	$\varphi_1+\varphi_5$	54	i	4,3,3,2	1	$B2$
45	F_4	φ_1	26	i	3,2	1	I_2
46		$\varphi_4=\mathrm{Ad}$	52	i	12,8,6,2	3	$E4$
47	G_2	φ_1	7	i	2	1	I_1
48		$2\varphi_1$	7+7	r	2,2,2	2	$I(1,1)$
49		$\varphi_2=\mathrm{Ad}$	14	i	6,2	1	I_2

The results we have obtained are listed in the last column of Table V, where the notations are those of Refs. 27 and 28 and I_1 and I_2 denote, respectively, the unique isomorphism class of orbit spaces of dimension 1 and 2.

In the column labelled by n, we have reported the number of different A$\hat{P}$M's associated to sets of invariants with given weights d_i. When $n=1$ there is only one possible A$\hat{P}$M. When $n>1$ one has to select the right A$\hat{P}$M.

When not explicitly specified, the scale parameter s, identifying different isomorphism classes of A$\hat{P}$M's belonging to the same tower, will be assumed equal to 1.

Below, except for the cases in which the identification of the correct A$\hat{P}$M is obvious since there is only one possible choice, we shall explain the way our results were obtained, for each entry in the tables. The *standard procedure* used to calculate $\hat{P}_{ij}(p)$, which works in most of the cases, is first to evaluate $P_{ij}(x)$ from its definition (Eq. (6)), and then to guess the form of $\hat{P}(p)$, on the basis of the last equality in Eq. (6).

Then one has to perform a MIB transformation to reach the form of one of the A$\hat{P}$M's listed in Refs. 27 and 28 (*standard* A$\hat{P}$M's). In some cases the standard procedure cannot be adopted since the explicit form of some of the invariants is not known. In these cases the alternative procedure will be explained in full detail.

Let us spell out our notations. We shall denote by $x,y,\ldots$ the vectors belonging to the spaces hosting the representation we are concerned with; as a rule, different letters will be used to label vectors lying in spaces hosting unequivalent representations, while vectors $x_1,x_2,\ldots$, distinguished only by subscripts, label vectors lying in different copies of the same representation space. The first letters of the alphabet will be reserved to real constants. Scalar products will be denoted by x_1x_2 or x^2. The symbol Γ will denote a vector operator, and $z(x\Gamma y)$ will be used for the scalar product between the vectors z and $(x\Gamma y)$. To specify that x is a vector of the real space W_φ of the representation φ, we shall simply write $x\in\varphi$.

• Entries 22, 27, 33, 39, 41, 46: (B_3,φ_2), (B_4,φ_2), (D_4,φ_2), (C_3,φ_1^2), (C_4,φ_1^2), (F_4,φ_4). All of them correspond to adjoint representations. As shown in Section III, the related $\hat{P}$-matrices are equivalent to the $\hat{P}$-matrices of the associated Weyl groups, and these have been identified in Ref. 32.

• Entry 3: $(A_1,2\varphi_1^2)$. One has to choose between the following two classes of A$\hat{P}$M's: $I(1,1)$ and $II(1)$. Let $x_1,x_2\in\varphi_1^2$. The basic invariants suggested by Schwarz are the following: $q_1=x_1^2$, $q_2=x_2^2$, and $q_3=x_1x_2$. A better choice, to meet the standard form assumed for p_3, is the following: $p_1=q_1-q_2={x_1}^2-{x_2}^2$, $p_2=2q_3=2x_1x_2$, and $p_3=q_1+q_2={x_1}^2+{x_2}^2$. Through the standard procedure one finds the following non-trivial elements of the $\hat{P}$-matrix: $\hat{P}_{11}=4p_3$, $\hat{P}_{12}=0$, and $\hat{P}_{22}=4p_3$, which coincide with those of the $\hat{P}$-matrix of class $I(1,1)$.

• Entry 6: $(A_{n\geqslant 2},2\cdot(\varphi_1+\varphi_n))$. One has to choose among the following five classes of A$\hat{P}$M's: $A1(1,1,1,1)$, $A3(1,1,1)$, $A4(1,1)$, $A5(1,1,1)$, and $A7(1,1)$.

Let $x_1,x_2\in\varphi_1,y_1,y_2\in\varphi_n$. A possible choice for the elements of a MIB, leading through the standard procedure to the $\hat{P}$-matrix of class $A1(1,1,1,1)$, is the following: $p_1=2(x_1y_2+x_2y_1)$, $p_2=2(x_1x_2-y_1y_2)$, $p_3=x_1^2-x_2^2+y_1^2-y_2^2$, and $p_4=x_1^2+x_2^2+y_1^2+y_2^2$. (Note that, for $z_1=x_1+iy_1$ and $z_2=x_2+iy_2$, one obtains $p_1=2\mathrm{Im}(z_1z_2)$, $p_2=2\mathrm{Re}(z_1z_2)$, $p_3=z_1^*z_1-z_2^*z_2$, and $p_1=z_1^*z_1+z_2^*z_2$.)

• Entry 12: $(A_3,2\varphi_2)$. Let $x_1,x_2\in\varphi_2$. The elements of the MIB can be chosen as in Entry 3 and the corresponding A$\hat{P}$M turns out to be the same.

• Entry 16: $(A_6,\varphi_2+\varphi_5)$. One has to choose among the following three classes of A$\hat{P}$M's: $I(2,1)$ $(d_2=4)$, $II(2)$ $(s=2)$, and $III.2$.

Let $x\in\varphi_2$ and $y\in\varphi_5=\overline{\varphi_2}$. The complex vectors x and y can be realized as antisymmetric 7×7 complex matrices, which can be decomposed into their real and imaginary parts u and v: $x=u+iv$, $y=u-iv$. The elements of a MIB can be chosen to be the following (repeated indexes are summed):

$$q_3=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7}x_{i_6i_7}\varepsilon_{i_1i_2i_3i_4i_5j_6j_7}y^{j_6j_7},$$
$$=2\times 5!\,x_{i_6i_7}y^{i_6i_7},$$

$$q_2=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7}x_{i_4i_5}x_{i_6i_7}\varepsilon_{i_1i_2i_3j_4j_5j_6j_7}y^{j_4j_5}y^{j_6j_7}$$
$$=4\,3!\,x_{i_4i_5}x_{i_6i_7}(y^{i_4i_5}y^{i_6i_7}-2y^{i_4i_6}y^{i_5i_7}),$$

$$q_1=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7}x_{i_2i_3}x_{i_4i_5}x_{i_6i_7}\varepsilon_{i_1j_2j_3j_4j_5j_6j_7}y^{j_2j_3}y^{j_4j_5}y^{j_6j_7}$$
$$=8\,x_{i_2i_3}x_{i_4i_5}x_{i_6i_7}(y^{i_2i_3}y^{i_4i_5}y^{i_6i_7}-4y^{i_2i_3}y^{i_4i_6}y^{i_5i_7}+2y^{i_2i_4}y^{i_3i_6}y^{i_5i_7}).$$

A better choice of independent invariants is the following one:

$$p_3=-x_{i_6i_7}y^{i_6i_7}=\mathrm{Tr}(xy)=\mathrm{Tr}(uu+vv),$$

$$p_2=x_{i_4i_5}x_{i_6i_7}y^{i_4i_6}y^{i_5i_7}=\mathrm{Tr}(xyxy)=\mathrm{Tr}(u^4+v^4-2uvuv+4u^2v^2),$$

$$p_1=-x_{i_2i_3}x_{i_4i_5}x_{i_6i_7}y^{i_2i_4}y^{i_3i_6}y^{i_5i_7}=\mathrm{Tr}(xyxyxy)$$
$$=\mathrm{Tr}(u^6+6u^4v^2-6u^3vuv+3u^2vu^2v$$
$$+3uv^2uv^2-6v^3uvu+6u^2v^4+v^6).$$

The elements of the matrix $P(u,v)$ are obtained from the scalar products of the gradients of the invariants with respect to the independent variables u_{ij} and v_{ij}, $i<j$. The expressions of the non-trivial elements of $P(u,v)$ one obtains are awfully long and do not allow a determination of $\hat{P}(p)$ by simple inspection. This difficulty can be bypassed in the following way. One writes for each $\hat{P}_{ij}(p)$, $i,j=1,2$, the most general w-homogeneous polynomial in p_1,p_2 and p_3 with the correct weight d_i+d_j-2:

$$\hat{P}_{22}=a_1p_1+a_2p_2p_3+a_3p_3^3,$$

$$\hat{P}_{12}=b_1p_1p_3+b_2p_2^2+b_3p_2p_3^2+b_4p_3^4,$$

$$\hat{P}_{11}=c_1p_1p_2+c_2p_1p_3^2+c_3p_2^2p_3+c_4p_2p_3^3+c_5p_3^5,$$

where a_i, b_i, c_i are unknown constants. These unknown parameters can be determined by solving systems of linear equations obtained from the identities $\hat{P}_{ij}(p(u,v))=P_{ij}(u,v)$, evaluated for conveniently chosen numerical values of the u_{ij} and v_{ij}. By using an overdetermined system of equations, one gets, as a bonus, a check of the correctness of the results. In this way, one obtains the following expressions for the non-trivial elements of the $\hat{P}$-matrix:

$$\hat{P}_{22}=16p_1,$$

$$\hat{P}_{12}=16p_1p_3+6p_2^2-6p_2p_3^2+\frac{1}{2}p_3^4,$$

$$\hat{P}_{11}=\frac{3}{8}(40p_1p_2+20p_1p_3^2-10p_2p_3^3+p_3^5),$$

which can be reduced to the standard A$\hat{P}$M of class $III.2$ after the following MIB transformation:

$$p_1'=1296p_1-864p_2p_3+124p_3^3,$$

$$p_2'=36p_2-10p_3^2.$$

• Entry 17: $(A_8,\varphi_2+\varphi_7)$. One has to choose between the following classes of A$\hat{P}$M's: $A4(1,2)(s=2)$, $A8(3,2)(s=2)$, $B4(1)(s=2)$, and $E3$. Let $x\in\varphi_2$ and $y\in\varphi_7=\overline{\varphi_2}$. The complex vectors x and y can be realized as antisymmetric 9×9 complex matrices, which can be decomposed into their real and imaginary parts u and v: $x=u+iv$, $y=u-iv$. The elements of a MIB, as in Entry 16, can be chosen to be the following (repeated indexes are summed):

$$q_4=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7i_8i_9}x_{i_8i_9}\varepsilon_{i_1i_2i_3i_4i_5i_6i_7j_8j_9}y^{j_8j_9},$$

$$q_3=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7i_8i_9}x_{i_6i_7}x_{i_8i_9}\varepsilon_{i_1i_2i_3i_4i_5j_6j_7j_8j_9}y^{j_6j_7}y^{j_8j_9},$$

$$q_2=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7i_8i_9}x_{i_4i_5}x_{i_6i_7}x_{i_8i_9}\varepsilon_{i_1i_2i_3j_4j_5j_6j_7j_8j_9}y^{j_4j_5}y^{j_6j_7}y^{j_8j_9},$$

$$q_1=\varepsilon^{i_1i_2i_3i_4i_5i_6i_7i_8i_9}x_{i_2i_3}x_{i_4i_5}x_{i_6i_7}x_{i_8i_9}\varepsilon_{i_1j_2j_3j_4j_5j_6j_7j_8j_9}y^{j_2j_3}y^{j_4j_5}y^{j_6j_7}y^{j_8j_9}.$$

A better choice of independent invariants is the following one:

$$p_4=\mathrm{Tr}(xy)=\mathrm{Tr}(uu+vv),$$

$$p_3=\mathrm{Tr}(xyxy)=\mathrm{Tr}(u^4+v^4-2uvuv+4u^2v^2),$$

$$p_2=\mathrm{Tr}(xyxyxy)=\mathrm{Tr}(u^6+6u^4v^2-6u^3vuv+3u^2vu^2v+3uv^2uv^2-6v^3uvu+6u^2v^4+v^6),$$

$$\begin{aligned}p_1=&\mathrm{Tr}(xyxyxyxy)\\&=\mathrm{Tr}(2u^8+16u^6v^2-16u^5vuv+16u^4vu^2v+16u^4v^4-8u^3vu^3v-16u^3vuv^3+16u^3v^2uv^2\\&\quad+8u^2v^2u^2v^2-16u^2v^2uvuv+16u^2vuv^2uv+4uvuvuvuv-16u^2vuvuv^2+16u^2vu^2v^3\\&\quad-16u^3v^3uv-8uv^3uv^3+16uv^2uv^4-16uvuv^5+16u^2v^6+2v^8).\end{aligned}$$

With the same technique used to obtain the results of Entry 16, one sets

$$\hat{P}_{33}=a_1p_2+a_2p_3p_4+a_3p_4^3,$$

$$\hat{P}_{23}=b_1p_1+b_2p_2p_4+b_3p_3^2+b_4p_3p_4^2+b_5p_4^4,$$

$$\hat{P}_{22}=c_1p_1p_4+c_2p_2p_3+c_3p_2p_4^2+c_4p_3^2p_4+c_5p_3p_4^3+c_6p_4^5,$$

$$\hat{P}_{13}=d_1p_1p_4+d_2p_2p_3+d_3p_2p_4^2+d_4p_3^2p_4+d_5p_3p_4^3+d_6p_4^5,$$

$$\hat{P}_{12}=e_1p_1p_3+e_2p_1p_4^2+e_3p_2^2+e_4p_2p_3p_4+e_5p_2p_4^3+e_6p_3^3+e_7p_3^2p_4^2+e_8p_3p_4^4+e_9p_4^6,$$

$$\begin{aligned}\hat{P}_{11}=&f_1p_1p_2+f_2p_1p_3p_4+f_3p_1p_4^3+f_4p_2^2p_4+f_5p_2p_3^2+f_6p_2p_3p_4^2+f_7p_2p_4^4+f_8p_3^3p_4+f_9p_3^2p_4^3\\&+f_{10}p_3p_4^5+f_{11}p_4^7,\end{aligned}$$

where a_i, b_i, c_i, d_i, e_i, and f_i are unknown constants, and from the identities $P(u,v)=\hat{P}(p(u,v))$ one obtains the following expressions for the non-trivial elements of the $\hat{P}$-matrix:

$$\hat{P}_{33}=16p_2,$$

$$\hat{P}_{23}=12p_1,$$

$$\hat{P}_{22}=\frac{3}{32}[20(6p_1p_4+8p_2p_3-4p_2p_4^2-3p_3^2p_4+p_3p_4^3)-p_4^5],$$

$$\hat{P}_{13}=\frac{1}{6}[20(6p_1p_4+8p_2p_3-4p_2p_4^2-3p_3^2p_4+p_3p_4^3)-p_4^5],$$

$$\hat{P}_{12}=\frac{1}{8}[2(72p_1p_3+36p_1p_4^2+64p_2^2-32p_2p_4^3-12p_3^3-18p_3^2p_4^2+9p_3p_4^4)-p_4^6],$$

$$\hat{P}_{11}=\frac{1}{12}[14(32p_1p_2+24p_1p_3p_4+4p_1p_4^3+16p_2p_3^2-16p_2p_3p_4^2-4p_2p_4^4-12p_3^3p_4+2p_3^2p_4^3 + p_3p_4^5)-p_4^7],$$

which can be reduced to the standard A$\hat{P}$M of class $E3$, after the following MIB transformation:

$$p_1'=-54(8192p_1-10240p_2p_4-4096p_3^2+3648p_3p_4^2-279p_4^4),$$

$$p_2'=36(256p_2-144p_3p_4+17p_4^3),$$

$$p_3'=-3(32p_3-7p_4^2).$$

• Entry 19, 32: $(B_{n\geqslant 2},2\varphi_1)$, $(D_{n\geqslant 4},2\varphi_1)$. Let $x_1,x_2\in\varphi_1$. The elements of the MIB can be chosen as in Entry 3 and the $\hat{P}$-matrix turns out to be the same.

• Entry 25: $(B_3,2\varphi_1+\varphi_3)$. One has to choose among the following classes of A$\hat{P}$M's: $A1(1,1,1,1)$, $A3(1,1,1)$, $A4(1,1)$, $A5(1,1,1)$, and $A7(1,1)$. Let x_1, $x_2\in\varphi_1$, $y\in\varphi_3$. The basic invariants suggested by Schwarz are the following: $q_1=x_1^2$, $q_2=x_2^2$, $q_3=x_1x_2$, and $q_4=y^2$. A better choice for our purposes is the following one: $p_1=3q_1-3q_2=3{x_1}^2-3{x_2}^2$, $p_2=6q_3=6x_1x_2$, $p_3=q_1+q_2-2q^4={x_1}^2+{x_2}^2-2y^2$, and $p_4=q_1+q_2+q_4={x_1}^2+{x_2}^2+y^2$. The standard procedure leads to a $\hat{P}$-matrix with the following non-trivial elements: $\hat{P}_{11}=12(p_3+2p_4)$, $\hat{P}_{12}=0$, $\hat{P}_{13}=4p_1$, $\hat{P}_{22}=12(p_3+2p_4)$, $\hat{P}_{23}=4p_2$, and $\hat{P}_{33}=4(2p_4-p_3)$, which is equal to the $\hat{P}$-matrix of the class $A3(1,1,1)$.

• Entry 26: $(B_3,2\varphi_3)$. Let $x_1,x_2\in\varphi_3$. The elements of the MIB can be chosen as in Entry 3 and the $\hat{P}$-matrix turns out to be the same.

• Entry 29: $(B_4,\varphi_1+\varphi_4)$. One has to choose between the following two classes of A$\hat{P}$M's: $I(2,1)$ and $II(2)$. Let $x\in\varphi_1$, $y\in\varphi_4$. A possible choice for the elements of a MIB is the following: $p_1=x(y\Gamma y)$, $p_2=y^2$, and $p_3=x^2+y^2$. The following elements of the $\hat{P}$-matrix are easily obtained with the standard procedure: $\hat{P}_{12}(p)=4p_1$ and $\hat{P}_{22}(p)=4p_2$. As far as $\hat{P}_{11}(p)$ is concerned, an explicit calculation starting from $P_{11}(x,y)$ would require an explicit determination of the vector operators Γ. One can bypass this difficulty arguing as follows. In $P_{11}(x,y)$ there must be terms of the form $(y\Gamma y)(y\Gamma y)$ and x^2y^2; this suggests the following choice: $\hat{P}_{11}(p)=ap_2^2+bp_2p_3$. The determinant of $\hat{P}(p)$ has the following expression:

$$\det(\hat{P}(p))=16[p_1^2(3p_2-4p_3)-p_2^2(p_2-p_3)(ap_2+bp_3)].$$

A direct check proves that it does not satisfy the canonical equation and thus the complete active factor $A(p)$ has to factorize in $\det(\hat{P}(p))$. Owing to item *iii)* of Prop. II.1, $A(p)$ must be at least quadratic in p_1. This implies that the coefficient of p_1^2 in $\det(\hat{P}(p))$, which is irreducible, must be a factor of the whole determinant, and this requires $a=-\frac{3}{4}b$. After the following MIB transformation, $p_2'=-3p_2+2p_3$, and choosing the free parameter $b=108$, one ends up with the $\hat{P}$-matrix of the class $I(2,1)$.

• Entry 30: $(B_4,2\varphi_4)$. One has to choose among the following 10 classes of A$\hat{P}$M's: $A1(1,1,2,2)$, $A2(1,1,2,1)$, $A2(1,1,1,2)$, $A3(1,3,1)$, $A3(1,2,2)$, $A5(1,1,3)$, $A5(2,1,2)$, $A7(1,3)$, $A8(3,1)$, and $A9(3)$. Let $x_1,x_2\in\varphi_4$. A possible MIB, as suggested by Schwarz, is the following: $q_1=(x_1\Gamma x_1)(x_2\Gamma x_2)$, $q_2=x_1^2$, $q_3=x_2^2$, and $q_4=x_1x_2$. Like in Entry 3, it is convenient to chose the following form for the invariants: $p_1=q_1$, $p_2={x_1}^2-{x_2}^2$, $p_3=2x_1x_2$, and $p_4={x_1}^2+{x_2}^2$. The following elements of $\hat{P}(p)$ are easily determined through the standard procedure: $\hat{P}_{22}(p)=4p_4$, $\hat{P}_{23}(p)=0$, and $\hat{P}_{33}(p)=4p_4$. Moreover, $\hat{P}_{12}(p)=0$ because it is the difference of two equal terms, which are linear in p_1. The polynomial $P_{11}(x_1,x_2)$ is symmetric under the exchange $x_1\leftrightarrow x_2$ and contains terms of sixth degree: fourth in x_1 and second in x_2 and vice versa, thus, taking into account the symmetry $x_1\leftrightarrow x_2$, the polynomial $\hat{P}_{11}(p)$ can be written in the following general form:

$$\hat{P}_{11}(p)=ap_1p_4+bp_3^2p_4+2c^2[(p_4-p_2)^2(p_4+p_2)+(p_4+p_2)^2(p_4-p_2)].$$

The polynomial $P_{13}(x_1,x_2)$ contains terms of the form $(x_1\Gamma x_1)(x_1\Gamma x_2)+(x_1\leftrightarrow x_2)$, so its general expression in terms of the basic invariants must be $\hat{P}_{13}(p)=4dp_3p_4$. The determinant of $\hat{P}(p)$ factorizes p_4^2 and so, because of the condition in item *iii)* of Prop. II.1, the unique possibility we are left with is a complete active factor $A(p)$ of weight 8. Now the active factor $A(p)$ contains a monomial in $p_1p_4^2$ that must vanish in an A-basis, and this requires $a=0$. The canonical equation is then satisfied by $A(p)$ only if $b=4(d^2-c^2)$. After the following MIB transformation:

$$p_1'=\frac{4}{c}(2p_1-dp_3^2) \quad (c,d\neq 0),$$

$$p_2'=2p_2,$$

one ends up with the standard A$\hat{P}$M of class $A1(1,1,2,2)$. The free parameters c and d are no longer present in the $\hat{P}$-matrix, since they have been absorbed in the redefinition of the MIB.

• Entry 35: $(D_4,\varphi_1+\varphi_3+\varphi_4)$. One has to choose among the following 6 classes of A$\hat{P}$M's: $A2(1,1,1,1)$, $A3(1,2,1)$, $A5(1,1,2)$, $A5(2,1,1)$, $A7(1,2)$, and $A8(2,1)$. Let $x\in\varphi_1$, $y\in\varphi_3$, $z\in\varphi_4$. A possible MIB, as suggested by Schwarz, is the following: $p_1=x(y\Gamma z)$, $p_2=x^2$, $p_3=y^2$, and $p_4=x^2+y^2+z^2$. The following elements of $\hat{P}(p)$ are then easily determined through the standard procedure: $\hat{P}_{12}(p)=2p_1$, $\hat{P}_{13}(p)=2p_1$, $\hat{P}_{22}(p)=4p_2$, $\hat{P}_{23}(p)=0$, and $\hat{P}_{33}(p)=4p_3$. As for $\hat{P}_{11}(p)$, the polynomial $P_{11}(x,y,z)$ contains terms of fourth degree which are at most quadratic separately in x y and z; so its general expression in terms of the basic invariants is the following: $\hat{P}_{11}(p)=(p_4-p_2-p_3)(ap_2+bp_3)+cp_2p_3$.

The determinant of $\hat{P}(p)$ has the following expression:

$$\det\hat{P}(p)=64[F(p_2,p_3,p_4){p_1}^2+G(p_2,p_3,p_4)],$$

where $F(p)={p_2}^2+p_2p_3+{p_3}^2-p_2p_4-p_3p_4$ and $G(p)$ are polynomials not depending on p_1. A direct check proves that $\det\hat{P}(p)$ does not satisfy the canonical equation and, thus, the complete active factor $A(p)$ has to factorize in $\det(\hat{P}(p))$. Owing to item *iii)* of Prop. II.1, $A(p)$ must be at least quadratic in p_1, and this implies that the coefficient of p_1^2 in $\det(\hat{P}(p))$, which is irreducible, must be a factor of the whole determinant, and this requires $a=b=c$. The weight of the complete active factor is then equal to 6. After the following MIB transformation and choosing the free parameter $c=5184$,

$$p_2'=12p_2+6p_3-6p_4,$$

$$p_3'=-6p_3+2p_4,$$

one ends up with the standard A$\hat{P}$M of class $A2(1,1,1,1)$.

• Entry 36: $(D_4,\varphi_1+2\varphi_3)$. Let $x_1,x_2\in\varphi_3,z\in\varphi_1$. The elements of the MIB can be chosen as in Entry 25 and the $\hat{P}$-matrix turns out to be the same.

• Entry 48: $(G_2,2\varphi_1)$. Let $x_1,x_2\in\varphi_1$. The elements of the MIB can be chosen as in Entry 3 and the $\hat{P}$-matrix turns out to be the same.

VIII. MINIMIZATION OF HIGGS POTENTIALS: TWO EXAMPLES

In this section we shall make use of the theory and some of the results reported in this paper in order to yield a complete and rigorous discussion of the minima of two particular Higgs potentials. Our aim is just to show, in two substantial samples, that, even in relatively complex situations, the problem of minimizing a general polynomial invariant potential of degree 4 can be reduced to an almost trivial exercise in the $\hat{P}$-matrix approach, without needing to introduce any *ad hoc* assumptions or additional symmetries to eliminate intriguing terms in the potential.

In gauge theory model building the choice of the ground state is controlled by the minimum of a fourth degree invariant polinomial function $V(\phi)$ of the scalar real (Higgs) fields ϕ of the model. The fields ϕ are chosen to transform according to a real representation of the gauge group, which defines a real linear group G. The potential is generally built as the most general G-invariant fourth degree polynomial function of ϕ. The isotropy group of the point ϕ_0 at which the potential takes on its absolute minimum is the true symmetry group of the theory after spontaneous symmetry breaking. In the following, to be consistent with the notations used throughout the paper, we shall denote by $x \in \mathbb{R}^n$ the set of Higgs fields.

The two examples we shall work out are based on the grand unification gauge groups SO(10) and E_6. Models based on these groups have been extensively studied by many authors.[33,34] In the following we shall only be concerned with some technical group theoretical problems that are met in these analyses.

A. An effective approach to the determination of minima of *G*-invariant functions

In order to solve the minimization problem in the $\hat{P}$-matrix approach to the study of regular G-invariant functions mentioned in the Introduction, all one has to do is the following:

(1) Identify a MIB $p_1(x),\ldots,p_q(x)$ of G and the associated $\hat{P}$-matrix.
(2) Determine the equalities and inequalities defining the orbit space $\overline{\mathscr{S}}$ of G and its stratification, according to Theorem II.1.
(3) Write down the potential $\hat{V}(p)$ as a function of the basic invariants $(p_1,\ldots,p_q)$ and determine the conditions under which it turns out to be bounded below. It will not be restrictive to assume $\hat{V}(0)=0$.
(4) Determine the constrained minima of $\hat{V}(p)$ on the different primary strata of the orbit space $\overline{\mathscr{S}}$ of G and compare them in order to single out the absolute minimum.
(5) Identify the symmetry (isotropy type) of the stratum/a containing the orbits at which the potential can take on its absolute minimum.

The following remarks allow for considerable simplifications in the determination of the minimum of $\hat{V}(p)$, $p \in \overline{\mathscr{S}}$:

a) Owing to the w-homogeneity properties of the equalities and inequalities determining the stratification of the orbit spaces, it turns out to be convenient to adopt the following strategy:

i) introduce homogeneous variables

$$\eta_k = p_k / p_q^{d_k/2}, \quad k=1,\ldots,q-1, \tag{34}$$

so that, when p spans $\overline{\mathscr{S}}$, η spans the intersection $\overline{\mathscr{S}}_1$ of $\overline{\mathscr{S}}$ with the hyperplane of equation $p_q=1$;

ii) Define the auxiliary function $\hat{V}_{p_q}(\eta)$:

$$\hat{V}_{p_q}(\eta) = \hat{V}(\eta_1 p_q^{d_1/2},\ldots,\eta_{q-1} p_q^{d_{q-1}/2}, p_q), \quad \eta \in \overline{\mathscr{S}}_1, \ p_q>0, \tag{35}$$

which, for fixed η, is generally equal to p_q times a polynomial in $p_q^{1/2}$.

iii) After calling $\sigma^{(\alpha)}$, $\alpha=0,1,\ldots$, $(\sigma^{(0)}=\{0\})$ and $\sigma_1^{(\alpha)}$, $\alpha=1,\ldots$, the primary strata of $\overline{\mathscr{S}}$ and, respectively, of $\overline{\mathscr{S}}_1$, for each fixed p_q determine the constrained minima, $\hat{V}^{(\alpha)}(p_q)$, of the restrictions of $\hat{V}_{p_q}(\eta)$ to the strata $\sigma_1^{(\alpha)}$. We shall set $\hat{V}^{(0)}=\hat{V}(0)=0$.

iv) determine the minima of the functions $\hat{V}^{(\alpha)}(p_q)$, $p_q>0$, and compare them, in order to single out the absolute minimum of $\hat{V}(p)$, recalling that $\hat{V}(0)=0$.

b) Let us assume that $\hat{V}(p)$ depends on the invariants p_j, $j \in \mathscr{J}$, where $\mathscr{J}$ is a subset of $\{1,\ldots,q\}$ and let us denote by $\mathscr{S}'^{*}$ the orthogonal projection of $\mathscr{S}'$ in the linear subspace of $\mathbb{R}^{q-1}$ spanned by the variables η_j, $j \in \mathscr{J}\setminus\{q\}$. Then we have the following.

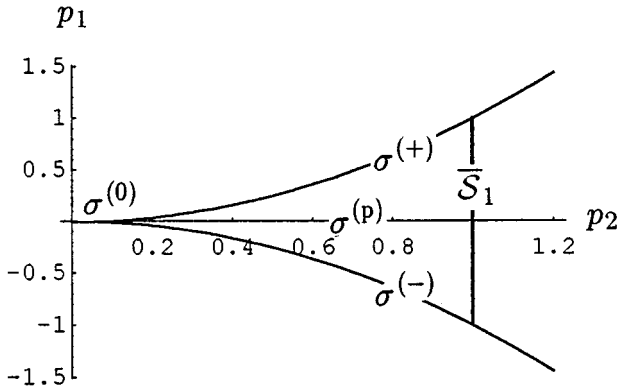

FIG. 1. Orbit space $\overline{\mathcal{S}}$ of the linear group (SO(10),**32**). $\overline{\mathcal{S}}_1$ is the intersection of $\overline{\mathcal{S}}$ with the straight line $p_2=1$.

1) If $\partial_j \hat{V}(p)$, for some $j \in \mathcal{I}\backslash\{q\}$ has no zero in $\mathcal{S}$, then the absolute minimum of $\hat{V}_{p_q}(\eta)$, $\eta \in \overline{\mathcal{S}}_1$, lies on the boundary of $\mathcal{S}_1$. This is the case[60] if there are elements of degree three and/or four in the MIB and the potential is assumed to be a fourth degree polynomial (for instance, because of renormalizability, in the case of Higgs potentials, or as a low order approximation, in the case of Landau potentials).

2) If $\hat{V}(p)$ depends only linearly on the p_j, $j \in \mathcal{I}\backslash\{q\}$ and if $\overline{\mathcal{S}}_1^*$ is contained in a convex polyhedron with vertices at points A_i which are points of $\mathcal{S}_1^*$, then $\hat{V}_{p_q}(\eta)$, $\eta \in \overline{\mathcal{S}}_1^*$ takes on its absolute minimum at one of the points A_i. The reason is that, in this case, for each fixed value of $p_q>0$, the functions $\hat{V}_{p_q}(\eta)$ are constant along linear varieties π_κ of equation $\hat{V}_{p_q}(\eta)=\kappa$. Thus, the absolute minimum of $\hat{V}_{p_q}(\eta)$, $\eta \in \overline{\mathcal{S}}_1^*$ is the lowest value of κ for which π_κ touches $\overline{\mathcal{S}}_1^*$. The set of contact points necessarily contains at least one vertex of the polyhedron.

The two examples worked out in the following subsections will clarify the procedure just described in the particular case of fourth degree polynomial potentials.

B. Minima of a general fourth degree polynomial invariant under the group (SO(10), 32)

The matrix group G is obtained by ''realizing'' a complex 16-dimensional matrix representation of the compact group SO(10).

According to Entry 37 in Table V, G is coregular and its MIB's are formed by two polynomials of degrees (4,2). If the MIB is chosen to be an A-basis, the $\hat{P}$-matrix takes on the following form:[26,27]

$$\hat{P}(p)=\begin{pmatrix} 16p_2^3 & 8p_1 \\ 8p_1 & 4p_2 \end{pmatrix}. \tag{36}$$

The active factor, $A(p)=p_2^4-p_1^2$, is proportional to the determinant of $\hat{P}(p)$ and an analysis of the semi-positivity properties of $\hat{P}(p)$ shows that the orbit space $\overline{\mathcal{S}}$ is the disjoint union of the primary strata $\sigma^{(\alpha)}$, $\alpha=0,\pm,\mathrm{p}$ (see Figure 1), whose defining equalities and inequalities are listed in Table VI.

A general G-invariant fourth degree polynomial in $x \in \mathbb{R}^{32}$ has the following form:

$$V(x)=\hat{V}(p(x)), \quad \hat{V}(p)=a_0p_2^2+a_1p_1+a_2p_2. \tag{37}$$

We shall assume $a_1 \neq 0$ in order to avoid trivial situations.

TABLE VI. Primary strata of the compact real linear group (SO(10), **32**). $\overline{x}$ denotes a conveniently chosen point in the stratum and $p(\overline{x})=(p_1(\overline{x}),p_2(\overline{x}))$. The coordinates, $(u_1,\ldots,u_{16},v_1,\ldots,v_{16})$, of $\overline{x}$ are referred to the basis specified in the Appendix.

Primary strata	Defining relations	Non-zero components of $\overline{x}$	$p(\overline{x})$	Isotropy subalgebras
$\sigma^{(0)}$	$p_1=p_2=0$	none	(0,0)	D_5
$\sigma^{(-)}$	$p_1=-p_2^2,\ p_2>0$	$u_1=1$	$(-1,1)$	A_4
$\sigma^{(+)}$	$p_1=p_2^2,\ p_2>0$	$u_1=u_{11}=\frac{1}{\sqrt{2}}$	(1,1)	B_3
$\sigma^{(p)}$	$p_1^2<p_2^4,\ p_2>0$	$2u_1=u_{11}=\frac{2}{\sqrt{5}}$	(0,1)	A_3

The function $\hat{V}(p),\ p\in\overline{\mathscr{S}}$ is bounded below only for $a_0>|a_1|$ or $a_0=|a_1|,\ a_2\geqslant 0$. Under these conditions, let us discuss the location of its absolute minimum as a function of the parameters a_i.

Using Remark b), it is easy to realize that the absolute minimum of $V(p)$ lies at $\sigma^{(0)}$ or on one of the subprincipal strata $\sigma^{(\pm)}$. The restrictions, $\hat{V}^{(\alpha)}(p_2)$, of $\hat{V}(p)$ to the strata $\sigma^{(\alpha)}$, $\alpha=0,+,-$, have the following form:

$$\hat{V}^{(0)}=0,\quad \hat{V}^{(\pm)}(p_2)=[(a_0\pm a_1)p_2+a_2]p_2, \tag{38}$$

so it is straightforward to derive the results summarized in Table VII, without exploiting the homogeneous variable η_1.

Let us stress that, up to this point, the discussion has been done with no reference to the specific linear group generating the $\hat{P}$-matrix.

In order to determine the isotropy types of the various primary strata one needs some more pieces of information about the particular MIB giving rise to the $\hat{P}$-matrix of equation (36). After setting $z=u+iv,\ u,v\in\mathbb{R}^{16}$, a general vector x of the real vector space $\mathbb{R}^{32}$ in which the compact linear group $G=(\mathrm{SO}(10),32)$ acts, may be written in the form $x=(u,v)$ and, from 3a.22 of Ref. 24 and the theorems reported in Section V, a possible choice of a MIB is the following:

$$p_1'(x)=\sum_1^{10}{}_k\ (\overline{z},\Gamma^k\overline{z})(z,\Gamma^k z)=\|(z,\Gamma^k z)\|^2,\quad p_2(x)=(\overline{z},z), \tag{39}$$

where $z\in\mathbb{C}^{16}$, (z,z') is the bilinear form $\Sigma_{j=1}^{16}z_jz_j'$ and the Γ^k are real symmetric 16×16 matrices, whose elements are proportional to Clebsch–Gordan coefficients coupling the representations 10 and 16×16 of the complex Lie algebra D_5.

The invariance properties of $p_1'(x)$ determine the relative values of the components Γ_{ij}^k, leaving indetermined only an overall numerical factor.

Let us show that it is possible to choose the normalization of the Γ's so that

TABLE VII. Primary strata, σ, hosting the absolute minimum of a general fourth degree polynomial invariant under (SO(10),**32**). The stratum is specified as a function of the parameters $a_i, i=0,1,2$, entering the definition of the potential, $\epsilon=-\mathrm{sign}(a_1)$ and a_1 is assumed $\neq 0$.

Boundedness conditions	Further conditions	σ
$a_0>\|a_1\|$	$a_2\geqslant 0$	$\sigma^{(0)}$
	$a_2<0$	$\sigma^{(\epsilon)}$
$a_0=\|a_1\|,\ a_2\geqslant 0$	none	$\sigma^{(0)}$
	$a_2=0$	$\sigma^{(\epsilon)}$

$$p_1(x)=p_1'(x)-p_2^2(x), \tag{40}$$

where $(p_1(x),p_2(x))$ form an A-basis, giving rise to the $\hat{P}$-matrix of equation (36).

With $(p_1'(x),p_2(x))$ being a MIB, only for homogeneity reasons, there will be real constants β_1,β_2,b_1,b_2 such that

$$\|\partial p_1'(x)\|^2=b_1p_1'(x)p_2(x)+b_2p_2^3(x) \tag{41}$$

and

$$p_1(x)=\beta_1p_1'(x)+\beta_2p_2^2(x), \tag{42}$$

where it is not restrictive to assume $\beta_1>0$, since a change of sign of p_1 does not change the $\hat{P}$-matrix.

By a detailed inspection of the weights of the representations 10 and 16 of D_5 (reported in the Appendix) one easily realizes that $\Gamma_{jj}^k=0$, for all $k=1,\ldots,10$, $j=1,\ldots,16$. Then, by choosing $x=x^{(j)}=(u^{(j)},0)$, where $u^{(j)}$ has components $u_i^{(j)}=\delta_{ij}$, one obtains the following relations:

$$p_1'(x^{(j)})=0,\quad \partial p_1'(x^{(j)})=0,\quad p_2(x^{(j)})=1, \tag{43}$$

which, when inserted into (41), immediately imply $b_2=0$. Moreover, since $\|\partial p_1'(x)\|^2$ and $p_1'(x)$ are positive semi-definite, by conveniently choosing the overall normalization of the Γ's it is possible to require, for instance, $b_1=32$.

By comparing the expression of $\|\partial p_1(x)\|^2=\hat{P}_{11}(p(x))=16p_2^3(x)$, obtained from (36), and the expression of $\|\partial p_1(x)\|^2$, calculated from (42) and (41), one obtains $\beta_1=-\beta_2=1$, so that (40) is satisfied.

For later use let us also note that (41) reduces to

$$\|\partial p_1'(x)\|^2=32p_1'(x)p_2(x). \tag{44}$$

The determination of the isotropy types of the different strata proceeds through the choice, for each stratum $\sigma^{(\alpha)}$, of a particular point lying in $\sigma^{(\alpha)}$ and through the identification of the corresponding isotropy group. The problem can be solved bypassing a complete determination of the Γ's. Let us prove our claim for the two subprincipal strata, using the results reported in Table VI and equation (40).

$\sigma^{(-)}$: On this stratum $p_1'(x)=0$, therefore, from (43), $p(x^{(j)})$, $j=1,\ldots16$ lies in $\sigma^{(-)}$.

$\sigma^{(+)}$: On this stratum $p_1'(x)=2p_2^2(x)$. Now, the weights of the representation 10 of D_5 are not degenerate. As a consequence, for each choice of $i,j=1,\ldots,16$, there is at most a $k=1,\ldots,10$, such that $\Gamma_{ij}^k\neq0$. Let us choose i,j,k so that $\Gamma_{ij}^k\neq0$ and define the vectors $x^{(ij)}$ so that $x_l^{(ij)}=\delta_{li}+\delta_{lj}$, $l=1,\ldots,32$. Then, from (39),

$$p_1'(x^{(ij)})=4\Gamma_{ij}^{k\,2}, \tag{45}$$

while (44) requires

$$\Gamma_{ij}^k=\pm\sqrt{2}, \tag{46}$$

so that

$$p_2(x^{(ij)})=2,\quad p_1'(x^{(ij)})=8=2p_2^2(x^{(ij)}), \tag{47}$$

assuring that $p(x^{(ij)})\in\sigma^{(+)}$.

The isotropy subalgebras of the Lie algebra of the compact group SO(10) at the points $x^{(j)}$ and $x^{(ij)}$ can be determined by standard methods and turn out to be isomorphic, respectively, to the compact forms of the Lie algebras A_4 and B_3.

Just for completeness in the Appendix we have listed the weight vectors of the two representations 10 and 16 of D_5 and the values of the non trivial components of the Γ's. These data can be used to check our arguments.

TABLE VIII. Primary strata σ, of the compact real linear group (E_6, **54**). In the columns, $\bar{x}$ denotes a conveniently chosen point in the stratum σ, $p(\bar{x})=(p_1(\bar{x}),\ldots,p_4(\bar{x}))$ and IS is the isotropy subalgebra of σ. The coordinates, $(u_1,\ldots,u_{27},v_1,\ldots,v_{27})$, of $\bar{x}$ are referred to the basis specified in the Appendix. The isotropy subalgebras of the strata $\sigma^{(3)}$ and $\sigma^{(4)}$ are distinct (so(9),**54**) subalgebras of (E_6, **54**) and $\gamma_\pm=\frac{1}{2}\sqrt{1\pm\sqrt{2/3}}$.

σ	Defining relations	Non-zero components of $\bar{x}$	p($\bar{x}$)	IS
$\sigma^{(0)}$	$p_1=p_2=p_3=p_4=0$	none	(0,0,0,0)	E_6
$\sigma^{(1)}$	$p_1=-2p_4^2$, $p_2=p_3==0$, $p_4>0$	$u_1=1$	(−2,0,0,1)	D_5
$\sigma^{(2)}$	$p_1=2p_4^2$, $p_2^2+p_3^2=4p_4^3$, $p_4>0$	$u_{1,15,27}=\frac{1}{\sqrt{3}}$	(2,2,0,1)	F_4
$\sigma^{(3)}$	$A_+(p)=0$, $p_4>0$	$u_{1,15}=\frac{1}{\sqrt{2}}$	(1,0,0,1)	B_4
$\sigma^{(4)}$	$A_-(p)=0$, $p_1>-2p_4^2$, $p_4>0$	$2u_{1,15,27}=u_{2,12}=\frac{2}{\sqrt{11}}$	$\left(\frac{-14}{121},\frac{18\sqrt{33}}{121},0,1\right)$	B_4
$\sigma^{(5)}$	$A(p)>0$, $p_1>-2p_4^2$, $p_4>0$	$u_{1,15}=\gamma_+$, $u_{2,12}=\gamma_-$	(0,0,0,1)	D_4

C. Minima of a general fourth degree polynomial invariant under the linear group (E_6,54)

The matrix group G is obtained by ''realizing'' a complex 27-dimensional matrix representation of the compact group E_6.

According to Entry 44 in Table V, G is coregular and its MIB's are formed by four polynomials of degrees (4,3,3,2). In a standard A-basis the $\hat{P}$-matrix takes on the following form (see Ref. 28, class B2):

$$\hat{P}(p)=\begin{pmatrix} 16(-p_1p_4+p_2^2+p_3^2+2p_4^3), & 24p_2p_4, & 24p_3p_4, & 8p_1 \\ 24p_2p_4, & 9(p_1+2p_4^2), & 0, & 6p_2 \\ 24p_3p_4, & 0, & 9(p_1+2p_4^2), & 6p_3 \\ 8p_1, & 6p_2, & 6p_3, & 4p_4 \end{pmatrix}. \tag{48}$$

After defining

$$A_\pm(p)=(2p_4^2-p_1)^{3/2}\pm(p_2^2+p_3^2-3p_1p_4+2p_4^3), \tag{49}$$

the active factor, $A(p)$, of the determinant can be written in the form

$$A(p)=A_+(p)A_-(p); \tag{50}$$

therefore, the singular strata are contained in the variety of equation $A(p)=0$.

Adding convenient rank and positivity conditions on the matrix $\hat{P}(p)$, one obtains the relations, reported in Table VIII, defining the primary strata $\sigma^{(\alpha)}$, $\alpha=0,\ldots,5$, of the orbit space $\bar{\mathscr{S}}$.

The intersection, $\bar{\mathscr{S}}_1$, of the orbit space $\bar{\mathscr{S}}$ with the hyperplane of equation $p_4=1$ has rotational symmetry around the p_1 axis. A section of $\bar{\mathscr{S}}_1$ with a plane through the symmetry axis is shown in Figure 2(a). Owing to this symmetry, it will not be restrictive to assume that the elements

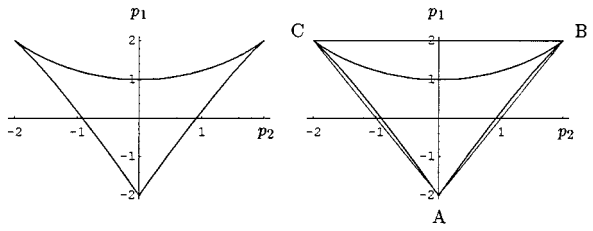

FIG. 2. Orbit space $\bar{\mathscr{S}}_1$ of the action of the linear group (E_6, **54**) in the unit sphere $p_4(x)=1$ of $\mathbb{R}^{54}$: (a) Intersection with the plane $p_3=0$; (b) orthogonal projection on the plane $p_3=0$.

TABLE IX. Primary strata, σ, hosting the absolute minimum of a general fourth degree polynomial invariant under E_6,**54**). The stratum is specified as a function of the parameters $\alpha_i=a_i/a_0$, $i=1,\ldots,4$, $a_0\neq 0$, $|\alpha_1|\leqslant 1/2$, entering the definition of the potential. $f_1(\alpha_1,\alpha_2)$ and $f_2(\alpha_1,\alpha_2)$ are functions of α_1 and α_2, defined by the relations: $f_1=\alpha_2^2/(1+2\alpha_1)$, $|\alpha_1|<1/2$; $f_2=-\alpha_2^2\sqrt{1-2\alpha_1}\,[(2+5\alpha_1)\sqrt{1-2\alpha_1}+2\sqrt{(1+\alpha_1)^3}]/[8\alpha_1^2(1+2\alpha_1)]$, for $0<\alpha_1<1/2$, and $f_2=-\infty$, for $-1/2<\alpha_1\leqslant 0$.

Boundedness conditions	Further conditions	σ
$\alpha_0>0$, $\lvert\alpha_1\rvert<\frac{1}{2}$	$(\alpha_2\neq 0,\ \alpha_4\geqslant f_1)$ or $(\alpha_2=0,\ \alpha_4\geqslant 0)$	$\sigma^{(0)}$
	$(\alpha_2\neq 0,\ f_2\leqslant\alpha_4\leqslant f_1)$ or $(\alpha_1\leqslant 0,\ \alpha_2=0,\ \alpha_4<0)$	$\sigma^{(2)}$
	$(\alpha_2\neq 0,\ \alpha_4\leqslant f_2)$ or $(\alpha_1\geqslant 0,\ \alpha_2=0,\ \alpha_4<0)$	$\sigma^{(1)}$
$a_0>0$, $\alpha_1=\frac{1}{2}$, $\alpha_4\geqslant 0$	$(\alpha_2\neq 0,\ \alpha_2^2\geqslant 2\alpha_4)$ or $\alpha_2=0$	$\sigma^{(0)}$
	$\alpha_2\neq 0,\ \alpha_2^2\geqslant 2\alpha_4$	$\sigma^{(2)}$
	$\alpha_2=0,\ \alpha_4=0$	$\sigma^{(1)}$
$\alpha_0>0$, $\alpha_1=-\frac{1}{2}$, $\alpha_2=0$, $\alpha_4\geqslant 0$	none	$\sigma^{(0)}$
	$\alpha_4=0$	$\sigma^{(2)}$

of the MIB have been chosen so that the general form of a fourth degree G-invariant real polynomial in $x\in\mathbb{R}^{54}$ is $V(x)=\hat{V}(p(x))$, where $\hat{V}(p)$ does not depend on p_3:

$$\hat{V}(p)=a_0p_4^2+a_1p_1+a_2p_2+a_4p_4,\quad a_i\in\mathbb{R}, \tag{51}$$

so that

$$\hat{V}_{p_q}(\eta)=[(a_0+a_1\eta_1)p_4+a_2\eta_2p_4^{1/2}+a_4]p_4. \tag{52}$$

Thus, the problem of determining the minima of $\hat{V}(p)$ in $\overline{\mathscr{S}}$ reduces to the analogous problem in the orthogonal projection of $\overline{\mathscr{S}}$ on an hyperplane of equation $p_3=\text{const}$.

We shall assume $a_0\neq 0$ in order to avoid trivial situations.

The conditions under which the function $\hat{V}(p)$, $p\in\overline{\mathscr{S}}$, is bounded below are not difficult to establish, if one makes use of the specification of the strata reported in Table VIII. We have reported the conclusions in the first column of Table IX.

Since the orthogonal projection, $\overline{\mathscr{S}}_1^*$, of $\overline{\mathscr{S}}_1$ on a plane of equation $p_3=\text{const}$ is contained in a triangle with vertices A, B, C (see Figure 2(b)), according to the Remark b), for fixed p_4, the function $\hat{V}_{p_4}(\eta)$, $\eta\in\overline{\mathscr{S}}_1^*$, attains its absolute minimum at A and/or B and/or C, depending on the slope of the straight lines of equations $V_{p_4}(\eta)=\kappa$ in the plane (η_1,η_2). Therefore, the absolute minimum of $\hat{V}(p)$, $p\in\overline{\mathscr{S}}$ lies on one of the primary strata $\sigma^{(\alpha)}$, $\alpha=0,1,2$.

Using the equalities and inequalities defining the primary strata reported in Table VIII, one obtains the following expressions for the restrictions of $\hat{V}(p)$ to $\sigma^{(1)}$ and $\sigma^{(2)}$:

$$\hat{V}(p)\big|_{\sigma^{(1)}}=\hat{V}_{p_4}(-2,0)=[(a_0-2a_1)p_4+a_4]p_4, \tag{53}$$

$$\hat{V}(p)\big|_{\sigma^{(2)}}=\hat{V}_{p_4}(2,\eta_2)=[(a_0+2a_1)p_4+a_2\eta_2p_4^{1/2}+a_4]p_4,\quad -2\leqslant\eta_2\leqslant 2, \tag{54}$$

whose absolute minima, for fixed p_4, are

$$\begin{aligned}\hat{V}^{(1)}(p_4)&=[(a_0-2a_1)p_4+a_4],\\ \hat{V}^{(2)}(p_4)&=[(a_0+2a_1)p_4-2|a_2|p_4^{1/2}+a_4]p_4.\end{aligned} \tag{55}$$

By comparing $\hat{V}^{(0)}$ and the minima of the funtions $\hat{V}^{(\alpha)}(p_4)$, $\alpha=1,2$, as functions of the parameters a_i, with standard manipulations, one finds the results summarized in Table IX, which give rise to the phase diagram of Figure 3, for $a_2\neq 0$.

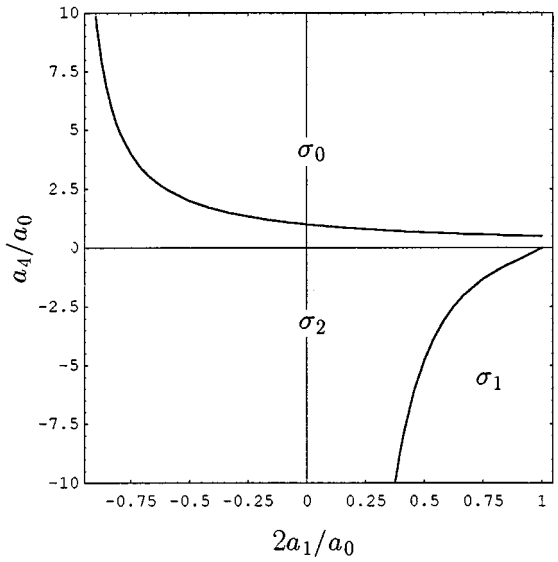

FIG. 3. Phase diagram for a general fourth degree polynomial potential invariant under the linear group (E_6, **54**). The ratio a_2/a_0 plays the role of a scale parameter and has been set equal to 1.

In order to determine the isotropy types of the various primary strata one needs some more pieces of information about the MIB giving rise to the $\hat{P}$-matrix of equation (48).

After setting $z=u+iv$, $u,v\in\mathbb{R}^{27}$, a general vector, x, of the real vector space $\mathbb{R}^{54}$ in which the compact real linear group $G=(E_6,54)$ acts, may be written in the form $x=(u,v)$ and, from 5a.6 of Ref. 24 and the theorems reported in Section V, a possible choice of a MIB can be done in the following way. Let us denote by $Q(z)$ the unique independent polynomial invariant of degree 3 of the complex linear group $(E_6,27)$:

$$Q(z)=\sum_{ijk}^{27}\Gamma_{ijk}z_iz_jz_k, \quad z\in\mathbb{C}^{27}, \tag{56}$$

where the Γ_{ijk} are real and completely symmetric in ijk. The invariance properties of $Q(z)$ only determine the relative values of the components Γ_{ijk}, leaving indetermined an overall numerical factor.

The 2 independent real trilinear invariants of G can be chosen to be the real and imaginary parts of $Q(z)$. The unique independent invariant of degree two is $p_4(x)=(\overline{z},z)$, while the unique independent fourth degree invariant can be obtained from the invariant $\|\partial p_2(x)\|^2=\|\partial p_3(x)\|^2$.

A simple inspection of the weights of the representation 27 of E_6 shows that for each choice of ij, there is at most a k such that $\Gamma_{ijk}\neq 0$ and this can occur only if i,j,k are all different. Arguing as in the preceding example, this piece of information turn out to be sufficient to conclude that, if the overall normalization factor in the definition of $Q(z)$ is conveniently chosen, an A-basis, giving rise to the $\hat{P}$-matrix of equation (48), is the following:

$$\begin{aligned} p_1(x)&=\frac{1}{3}\|\partial p_2(x)\|^2-2x^2, \quad p_2(x)=\Re\left(\sum_{ijk}^{27}\Gamma_{ijk}z_iz_jz_k\right),\\ p_3(x)&=-\Im\left(\sum_{ijk}^{27}\Gamma_{ijk}z_iz_jz_k\right), \quad p_4(x)=x^2. \end{aligned} \tag{57}$$

Like in the preceding example, the determination of the isotropy types of the various strata can be achieved bypassing the complete determination of the Γ's. We shall skip the details of the calculation and report only the final results in Table VIII.

TABLE X. Real compact coregular linear Lie groups with less than 5 basic invariants and isomorphic orbit spaces. In each row, the following items are listed: A label, Class, specifying the equivalence class of a $\hat{P}$-matrix, $\hat{P}(p)$; the degrees, d_i, of the set of basic invariants $\{p\}$; and all the linear simple compact Lie groups that generate $\hat{P}$-matrices of the given Class (they are specified by the Lie algebras, $(\mathscr{G},\varphi)$, of their complexifications).

Class	d_i	$(\mathscr{G},\varphi)$
I_1	2	All groups with 1 basic invariant
I_2	d,2	(A_1,φ_1^4), (A_2,Ad), $(A_3,\varphi_1+\varphi_2+\varphi_3)$, $(A_4,\varphi_2+\varphi_3)$, (B_2,Ad), $(B_3,\varphi_1+\varphi_3)$, $(D_4,\varphi_1+\varphi_3)(D_5,\varphi_4+\varphi_5)$, (C_3,φ_2), (F_4,φ_1),(G_2,Ad)
$I(1,1)$	2,2,2	$(A_1,2\varphi_1^2)$, $(A_3,2\varphi_2)$, $(B_{n\geqslant 2},2\varphi_1)$, $(D_{n\geqslant 4},2\varphi_1)$, $(B_3,2\varphi_3)$, $(D_4,2\varphi_3)$, $(G_2,2\varphi_1)$
$I(2,1)$	3,2,2	$(B_4,\varphi_1+\varphi_4)$
$III.1$	4,3,2	(A_3,Ad), (C_4,φ_2)
$III.2$	6,4,2	$(A_6,\varphi_2+\varphi_5)$, (B_3,Ad), (C_3,Ad)
$A1(1,1,1,1)$	2,2,2,2	$(A_{n\geqslant 2},2\varphi_1+2\varphi_n)$
$A1(1,1,2,2)$	4,2,2,2	$(B_4,2\varphi_4)$
$A2(1,1,1,1)$	3,2,2,2	$(D_4,\varphi_1+\varphi_3+\varphi_4)$
$A3(1,1,1)$	2,2,2,2	$(B_3,2\varphi_1+\varphi_3)$, $(D_4,\varphi_1+2\varphi_3)$
$B2$	4,3,3,2	$(A_2,\varphi_1^2+\varphi_2^2)$, $(A_5,\varphi_2+\varphi_4)$, $(E_6,\varphi_1+\varphi_5)$
$E1$	5,4,3,2	(A_4,Ad), (B_2,φ_1^2), (C_5,φ_2)
$E2$	6,4,4,2	(D_4,Ad)
$E3$	8,6,4,2	$(A_8,\varphi_2+\varphi_7)$, (B_4,Ad), (C_4,Ad)
$E4$	12,8,6,2	(F_4,Ad)

TABLE XI. Weight vectors for the representations 10 and 16 of SO(10) and 27 of E_6.

i	(SO(10),**10**) weight vectors	i	(SO(10),**16**) weight vectors	i	(E_6,**27**) weight vectors
1	(1,0,0,0,0)	1	(0,0,0,0,1)	1	(1,0,0,0,0,0)
2	(−1,1,0,0,0)	2	(0,0,1,0,−1)	2	(−1,1,0,0,0,0)
3	(0,−1,1,0,0)	3	(0,1,−1,1,0)	3	(0,−1,1,0,0,0)
4	(0,0,−1,1,1)	4	(1,−1,0,1,0)	4	(0,0,−1,1,0,1)
5	(0,0,0,−1,1)	5	(0,1,0,−1,0)	5	(0,0,0,−1,1,1)
6	(0,0,0,1,−1)	6	(−1,0,0,1,0)	6	(0,0,0,1,0,−1)
7	(0,0,1,−1,−1)	7	(1,−1,1,−1,0)	7	(0,0,0,0,−1,1)
8	(0,1,−1,0,0)	8	(−1,0,1,−1,0)	8	(0,0,1,−1,1,−1)
9	(1,−1,0,0,0)	9	(1,0,−1,0,1)	9	(0,0,1,0,−1,−1)
10	(−1,0,0,0,0)	10	(−1,1,−1,0,1)	10	(0,1,−1,0,1,0)
		11	(1,0,0,0,−1)	11	(0,1,−1,1,−1,0)
		12	(0,−1,0,0,1)	12	(1,−1,0,0,1,0)
		13	(−1,1,0,0,−1)	13	(1,−1,0,1,−1,0)
		14	(0,−1,1,0,−1)	14	(0,1,0,−1,0,0)
		15	(0,0,−1,1,0)	15	(−1,0,0,0,1,0)
		16	(0,0,0,−1,0)	16	(−1,0,0,1,−1,0)
				17	(1,−1,1,−1,0,0)
				18	(−1,0,1,−1,0,0)
				19	(1,0,−1,0,0,1)
				20	(−1,1,−1,0,0,1)
				21	(1,0,0,0,0,−1)
				22	(0,−1,0,0,0,1)
				23	(−1,1,0,0,0,−1)
				24	(0,−1,1,0,0,−1)
				25	(0,0,−1,1,0,0)
				26	(0,0,0,−1,1,0)
				27	(0,0,0,0,−1,0)

Just for completeness in the Appendix we have listed the weight vectors of the representation 27 of E_6 and the values of the non-trivial components of the Γ's. These data can be used to check our arguments.

IX. CONCLUSIONS

After classifying the coregular real representations of the compact simple Lie groups, we have determined the isomorphism classes of their orbit spaces in all cases in which the number q of basic polynomial invariants does not exceed 4. The identification of the isomorphism classes has been obtained by determining, for each real coregular representation, a matrix $\hat{P}(p)$, which yields a faithful image of the orbit space as the semi-algebraic set of points $\overline{\mathscr{S}}=\{p\in\mathbb{R}^q|\hat{P}(p)\geqslant 0\}$. The results we have found are a partial check of the correctness of the classification of allowable $\hat{P}$-matrices of coregular compact linear groups with up to 4 basic polynomial invariants, reported in Refs. 27 and 28 and turn out to be quite effective tools in the analysis of invariant potentials. They also confirm the following facts:

i) Not all the allowable $\hat{P}$-matrices of dimensions $\leqslant 4$ correspond to coregular compact simple Lie groups, or to coregular finite groups.[32] Part of them are surely related to coregular non-simple compact groups, some are, or might be, generated by non-coregular compact groups; and for others there might be no generating group. It has to be stressed, however, that the $\hat{P}$-matrix of whichever compact coregular linear group with less than five basic polynomial invariants should be found in the classification given in Refs. 27 and 28.

ii) Several unequivalent linear groups may share isomorphic orbit spaces; these groups may realize unequivalent representations of the same group or even representations of non homomorphic groups. To stress this phenomenon, in Table X, we have listed all the linear coregular simple Lie groups generating each allowable $\hat{P}$-matrix occurring in Table V.

iii) In all the orbit spaces realized by the compact coregular Lie groups listed in Table X, the scale parameter s,[27,28] labelling the different solutions of a tower of allowable $\hat{P}$-matrices, is always equal to 1. This supports the conjecture that only the allowable $\hat{P}$-matrices, corresponding to the lowest value of s, are significant; the other ones are spurious solutions of the canonical equation, originating from a scale symmetry that seems not to be the consequence of some property of the orbit spaces of the compact linear groups.

The method used in Refs. 27 and 28 to determine the allowable $\hat{P}$-matrices of dimensions $\leqslant 4$ may be applied also in higher dimensions, but the first attempts in this direction showed that the calculations needed to solve the canonical equation in these cases are awfully cumbersome. Some inductive methods to speed up the calculations have been discovered, but for the moment they enable us only to obtain part of the allowable solutions in dimensions $q+1$, from the solutions in lower dimensions. Thus the problems of determining all the allowable solutions of the canonical equation and of characterizing those effectively generated by some compact linear group are still open.

ACKNOWLEDGMENTS

We would like to thank Dr. G. Valente for discussions and many valuable suggestions for improving the manuscript.

This paper is partially supported by INFN and MURST 40% and 60%, and is carried out as part of the European Community Program ''Gauge Theories, Applied Supersymmetry and Quantum Gravity'' under contract SCI-CT92-D789.

APPENDIX: WEIGHT VECTORS AND CLEBSCH-GORDAN COEFFICIENTS FOR THE REPRESENTATIONS 10 AND 16 OF SO(10) AND 27 OF E_6

The bases in the spaces of the representations 10 and 16 of D_5 and of 27 of E_6 can be chosen to be the weight vectors e_k relative to the weights λ_k, ordered as indicated in Table XI.

With reference to these bases the nontrivial components of Γ^k_{ij} for $(D_5,16)$ and Γ_{ijk} for $(E_6,\mathbf{27})$ take on the values $\pm\sqrt{2}$, and $\pm\sqrt{3}$, respectively, in correspondence for the following sets $\tau^{\pm}_{D_5}$ and $\tau^{\pm}_{E_6}$ of values of the indices (k,i,j):

$$\tau^{+}_{D_5}=\{(1,1,11),(1,3,7),(2,1,13),(2,3,8),(3,1,14),(3,4,8),(4,1,15),(4,4,10),(5,1,16),$$
$$(5,7,10),(6,2,15),(6,4,13),(7,2,16),(7,7,13),(8,3,16),(8,9,13),(9,4,16),(9,9,14),$$
$$(10,6,16),(10,10,14),\quad \text{and permutations of } (i,j)\},$$

$$\tau^{-}_{D_5}=\{(1,2,9),(1,4,5),(2,2,10),(2,5,6),(3,2,12),(3,6,7),(4,3,12),(4,6,9),$$
$$(5,5,12),(5,8,9),(6,3,14),(6,6,11),(7,5,14),(7,8,11),(8,5,15),$$
$$(8,10,11),(9,7,15),(9,11,12),(10,8,15),(10,12,13),\quad \text{and permutations of } (i,j)\}.$$

$$\tau^{+}_{E_6}=\{(1,15,27),(1,18,25),(1,22,23),(2,13,26),(2,19,24),(3,10,27),(3,14,25),$$
$$(3,20,21),(4,9,26),(4,17,23),(5,6,27),(5,11,24),(5,16,21),(6,14,22),$$
$$(6,18,19),(7,8,25),(7,12,23),(8,13,20),(9,10,22),(9,15,19),$$
$$(10,16,17),(11,12,18),(13,14,15),\text{and permutations of } (i,j,k)\},$$

$$\tau^{-}_{E_6}=\{(1,16,26),(1,20,24),(2,12,27),(2,17,25),(2,21,22),$$
$$(3,11,26),(3,19,23),(4,8,27),(4,14,24),(4,18,21),(5,9,25),(5,13,23),$$
$$(6,7,26),(6,17,20),(7,10,24),(7,15,21),(8,11,22),(8,16,19),(9,12,20),$$
$$(10,13,18),(11,15,17),(12,14,16),\quad \text{and permutations of } (i,j,k)\}.$$

[1] L. O'Raifeartaigh, Rep. Prog. Phys. **42**, 159 (1979).
[2] J. Bernstein, Rev. Mod. Phys. **46**, 7 (1974).
[3] L. Michel, Rev. Mod. Phys. **52**, 617 (1980).
[4] L. Michel, in *Symmetries and broken symmetries in condensed matter physics*, edited by N. Boccara (IDSET, Paris, 1981), p. 21.
[5] J. Wess and J. Bagger, *Supersymmetry and supergravity* (Princeton U. P., Princeton, NJ, 1983).
[6] R. Gatto and G. Sartori, Commun. Math. Phys. **109**, 327 (1987).
[7] L. Michel and J. Mozrzymas, Lect. Notes Phys. **79**, 447 (1978).
[8] P. J. Houston andSiddhartha Sen, J. Phys. A **17**, 1163 (1984).
[9] Morse and S.S. Cairns, *Analysis and Differential Topology* (Academic, New York, 1969).
[10] Yu. M. Gufan, Sov. Phys. Solid State **13**, 175 (1971).
[11] Yu. M. Gufan and V. P. Sakhnenko, Sov. Phys. JETP **36**, 1009 (1973).
[12] D. Hilbert, Math. Ann. **36**, 473 (1890); **42**, 313 (1893).
[13] J. Killingbeck, J. Phys. C **5**, 2037 (1972); **5**, 2497 (1972).
[14] V. Kopsky, J. Phys. C **8**, 3251 (1975).
[15] V. Kopsky, J. Math. Phys. **19**, 574 (1978).
[16] V. Kopsky, J. Phys. A **12**, 429 (1979); **12**, 943 (1979).
[17] A. G. McLellan, J. Phys. C **7**, 3326 (1974).
[18] M. Abud and G. Sartori, Phys. Lett. B **104**, 147 (1981).
[19] M. Abud and G. Sartori, Ann. Phys. (N.Y.) **150**, 307 (1983).
[20] G. Sartori, J. Math. Phys. **24**, 765 (1983).
[21] C. Procesi and G. W. Schwarz, Invent. Math. **81**, 539 (1985).
[22] C. Chevalley, Am. J. Math. **77**, 778 (1955).
[23] J. E. Humphreys, *Reflection Groups and Coxeter Groups* (Cambridge U.P., Cambridge, England, 1990).
[24] G. W. Schwarz, Invent. Math. **49**, 167 (1978).
[25] H. Kraft, Aspekte der Matematik D1 (Vieweg, Braunschweig, 1984).
[26] G. Sartori, Mod. Phys. Lett. A **4**, 91 (1989).
[27] G. Sartori and V. Talamini, Commun. Math. Phys. **139**, 559 (1991).
[28] G. Sartori and V. Talamini, J. Group Theory Physics **2**, 13 (1994), avalaible also at http://xxx.lanl.gov/abs/hep-th/9512067.
[29] M. V. Jarić, L. Michel, and R. T. Sharp, Lect. Notes Phys. **180**, 317 (1983).
[30] V. Talamini, in *Symmetry and Structural Properties of Condensed Matter*, edited by T. Lulek, W. Florek, and S. Walcerz (World Scientific, Singapore, 1995), p. 441.
[31] G. Sartori, Riv. Nuovo Cimento **14**, 1 (1991).
[32] G. Sartori and G. Valente, J. Phys. A **29**, 193 (1996).
[33] See, for instance, R. Slansky, Phys. Rep. **79**, 1 (1981); P. Langacker, Phys. Rep. **72**, 185 (1981), and references therein.

[34] G. Anastaze, J.-P. Derendinger, and F. Buccella, Z. Phys. C **20**, 269 (1983); F. Güresey and M. Serdaroğlu, Nuovo Cimento A **65**, 337 (1982).
[35] G. E. Bredon, *Introduction to Compact Transformation Groups* (Academic, New York, 1972).
[36] G. W. Schwarz, Inst. Hautes Etudes Sci. Publ. Math. **51**, 37 (1980).
[37] G. D. Mostow, Ann. Math. **65**, 513 (1957); **65**, 432 (1957).
[38] E. Noether, Math. Ann. **77**, 89 (1916).
[39] M. Nagata, J. Math. Kyoto Univ. **3**, 369 (1963/64).
[40] G. W. Schwarz, Topology **14**, 63 (1975).
[41] H. Whitney, Ann. Math. **66**, 545 (1957).
[42] N. Jacobson, *Lie Algebras* (Interscience, New York, 1962).
[43] G. Racah, Rend. Sci. Fis. Mat. Nat. Lincei **VIII**, 108 (1950).
[44] B. Gruber and L. O'Raifertaigh, J. Math. Phys. **5**, 12 (1964).
[45] G. Hochschild, *The Structure of Lie Groups* (Holden-Day, San Francisco, 1965).
[46] G. Frobenius and I. Schur, Proc. Berlin Acad. **186** (1906).
[47] M. Hamermesh, *Group Theory and its Application to Physical Problems* (Addison–Wesley, Reading, MA, 1962).
[48] M. L. Mehta, J. Math. Phys. **7**, 1824 (1966).
[49] M. L. Mehta and P. K. Srivastava, J. Math. Phys. **7**, 1833 (1966).
[50] V. Talamini, degree thesis, Università di Padova, 1985.
[51] A. Borel, *Linear Algebraic Groups* (Benjamin, New York, 1969).
[52] A. Borel, Proc. Symp. Pure Math. **9**, 3 (1966).
[53] C. Chevalley, *Theorie des Groupes de Lie, Tome II. Groupes Algebriques* (Hermann, Paris, 1951).
[54] J. E. Humphreys, *Linear Algebraic Groups* (Springer-Verlag, New York, 1975).
[55] V. S. Varadarajan, *Lie Groups, Lie Algebras and their Representations* (Springer-Verlag, New York, 1984).
[56] G. Hochschild and G. D. Mostow, Ann. Math. **70**, 85 (1959).
[57] G. Hochschild and G. D. Mostow, Ann. Math. **66**, 495 (1957).
[58] J. E. Humphreys, Am. Math. Monthly **85**, 341 (1978).
[59] J. Tits, *Tabellen zu den Einfachen Lie Gruppen und ihren Darstellungen*, Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1967), Vol. 40.
[60] L. Michel, in *Regards sur la physique contemporaine* (CNRS, Paris, 1980) and CERN preprint TH 2716 1979).

Entropy and the Segal-Bargmann transform

Stephen Bruce Sontz[a)]
Universidad Autonoma Metropolitana, Iztapalapa, Mexico DF 09340, Mexico

(Received 21 August 1997; accepted for publication 31 October 1997)

The Segal-Bargmann transform A is a unitary isomorphism of the quantum mechanical configuration space $L^2(\mathbf{R}^n,\nu)$ onto the Segal-Bargmann, or phase, space $HL^2(\mathbf{C}^n,\mu)$ of entire functions square integrable with respect to μ. Here μ and ν are Gaussian measures, and for $z\in\mathbf{C}^n$ and measurable $f:\mathbf{R}^n\to\mathbf{C}$ $Af(z) := \int_{\mathbf{R}^n} d\nu(t) f(t)\exp(-z^2/2+\sqrt{2}z\cdot t)$ if the integral converges. Af is defined for all $f\in L^p(\mathbf{R}^n,\nu)$ if $1<p\leqslant\infty$, but A does not preserve L^p structure if $p\neq 2$. Specifically, it is proven here that A is a compact map from $L^p(\mathbf{R}^n,\nu)$ to $L^q(\mathbf{C}^n,\mu)$ if $1\leqslant q<2$ and $p>1+q/2$ (which implies $p>q$) and in this case $\|A\|_{p\to q}\leqslant((1+q/2-p'q/2)(1-q/2))^{-n/2q}$ where p' is the dual of p. Moreover, a calculation with test functions shows that $\|A\|_{p\to q}=\infty$ if $q>2$ or $p<1+q/2$ (which is implied by $p<q$). Using an improved operator norm estimate (analogous to the Hausdorff-Young inequality) we prove Hirschman type entropy inequalities for $f\in L^2(\mathbf{R}^n,\nu)$, $(p^{-1}-2^{-1})S(f)\leqslant(q^{-1}-2^{-1})S(Af)+nC_{p,q}\|f\|_2^2$, where $1\leqslant q<2$, $p>1+q/2$, $C_{p,q}$ is given explicitly as a function of p and q and $S(g)$ is the entropy of g. Examples given here show A is *not* entropy preserving. So these inequalities provide constraints on the change in entropy. Similarly, we get energy-entropy, or log-Sobolev, inequalities $a_1S(Af)+a_2S(f)\leqslant a_3\|f\|_2^2+a_4\langle f,Nf\rangle$, for f in a dense subspace of $L^2(\mathbf{R}^n,\nu)$. Here $\langle f,Nf\rangle$ is the Dirichlet form of the energy operator N in the configuration space and the a_i can have various signs and are also given explicitly. The principal results of this article are generalized Heisenberg uncertainty relations, namely they are constraints on Af, the phase space representation, given information about f, the configuration space representation. © *1998 American Institute of Physics.* [S0022-2488(98)03403-3]

I. INTRODUCTION

The Segal-Bargmann transform (see Refs. 1–3) is used in mathematical physics to ''diagonalize'' creation and annihilation operators. The Segal-Bargmann space itself dates back to the work of Fischer[4] in mathematics and Fock[5] in physics. The Segal-Bargmann space as well as the standard quantum mechanical configuration space $L^2(\mathbf{R}^n,d^nx)$ carry irreducible representations of the exponentiated (Weyl) canonical commutation relations, and these representations are unitarily equivalent by a theorem of Stone[6] and von Neumann.[7] This unitary equivalence is realized explicitly by the Bargmann transform. In this article, the point of view is taken that the usual configuration space $L^2(\mathbf{R}^n,d^nx)$ should be replaced by a unitarily equivalent configuration space $L^2(\mathbf{R}^n,\nu)$, known as the ground state representation, where a Gaussian probability measure ν has replaced the Lebesgue measure d^nx. Consequently, the Segal-Bargmann transform is defined as a unitary map from the ground state representation space to the Segal-Bargmann space. This unitarity, as well as the preservation of the creation and annihilation operators, and consequently the energy operator and its associated Dirichlet form, says that the quantum mechanical formulations are equivalent. However, these two representations are different in some sense as well. Intuitively, the difference is that the ground state representation is in a configuration space, while the Segal-Bargmann space representation is in a phase space (see Ref. 8).

Much of the analysis of this article is motivated, both physically and mathematically, by analogy to the Fourier transform. By the Plancherel theorem, the Fourier transform is also a unitary map of Hilbert spaces (actually, L^2 spaces), but the nontrivial relation between the source

a)Electronic mail: sontz@xanum.uam.mx

 © 1998 American Institute of Physics

and target Hilbert spaces of the Fourier transform can be seen by considering the action of the Fourier transform on individual functions. For example, the Heisenberg uncertainty principle, with all of its physical intuition, can be understood mathematically as an inequality involving a function and its Fourier transform. Another way of analyzing the Fourier transform is by investigating its action on spaces other than L^2. In this regard, the Hausdorff-Young inequality gives information about the L^p mapping properties of the Fourier transform, namely that $L^p(\mathbf{R}^n, d^n x)$ is mapped boundedly to $L^{p'}(\mathbf{R}^n, d^n x)$ where p' is the dual index of p, provided that $1 \leqslant p < 2$.

I analyze the Segal-Bargmann transform A, already described in the abstract, in a similar way in this article, whose contents I now outline. First, I ask what are the pairs of indices p and q for which the operator norm $\|A\|_{p\to q}$ is finite. I show in Theorem 3.1 that a sufficient condition is $1 \leqslant q < 2$ and $p > 1 + q/2$ and that the operator is actually compact in this case. Notice that these conditions on p and q imply $p > q$. Moreover by using a Hille-Tamarkin norm $|||A^T|||_{q',p'}$ (whose definition is given later), I get the estimate $\|A\|_{p\to q} \leqslant |||A^T|||_{q',p'} = (1 + q/2 - p'q/2)^{-n/2p'} (1 - q/2)^{-n/2q} (1 + q/2)^{-n(1/q - 1/p')/2}$ of the operator norm. Using test functions, I show at the end of the article that $\|A\|_{p\to q} = \infty$ if $q > 2$ or $p < 1 + q/2$. Zhou has already shown this in the case $q > 2$ in Ref. 9.

Also, the result $\|A\|_{2\to 2} = 1$ is a cornerstone to this theory; it is due to Segal in Refs. 1 and 2 and to Bargmann in Ref. 3. Specifically, they show that A is unitary onto the Segal-Bargmann space if $p = q = 2$. This last statement implies via two applications of Hölder's inequality that $\|A\|_{p\to q} = 1$ whenever $2 \leqslant p \leqslant \infty$ and $1 \leqslant q \leqslant 2$. This means that the only case for which the finiteness or infiniteness of $\|A\|_{p\to q}$ has not yet been established is when $q < 2$ and $p = 1 + q/2$. Even though the kernel of the Segal-Bargmann transform is a Gaussian, the results of Lieb[10] do not apply here to prove $\|A\|_{p\to q} < \infty$, because this occurs only when $p > q$ (except if $p = q = 2$) and the kernel of A is the exponential of a complex quadratic form. See Ref. 10 for the specific hypotheses that must be fulfilled.

The estimate for $\|A\|_{p\to q}$ goes to infinity as the pair (p,q) approaches the pair (2,2). To get an estimate on the operator norm that behaves better, I use the Riesz-Thorin interpolation theorem. Specifically, in Theorem 3.2 I get an estimate of the operator norm which is a log-convex function of the interpolating parameter. This estimate is an analogue for the Segal-Bargmann transform of the Hausdorff-Young inequality for the Fourier transform. With this information about the boundedness of A from L^p to L^q, plus the specific estimate of the operator norm, I can apply a technique of Hirschman. This technique consists in differentiating a parameterized family of inequalities with respect to the parameter to get a new inequality. (Another condition must be satisfied as discussed in the article.) In Theorem 4.1 this yields the inequality

$$(p^{-1} - 2^{-1}) S(f) \leqslant (q^{-1} - 2^{-1}) S(Af) + (\log |||A^T|||_{q',p'}) \|f\|_2^2$$

for $1 \leqslant q < 2$ and $p > 1 + q/2$ and all $f \in L^2(\mathbf{R}^n, \nu)$. This inequality is an analogue to an inequality for the Fourier transform conjectured by Hirschman in Ref. 11. Although Hirschman did not prove this inequality, which is known nonetheless as Hirschman's inequality, he did prove a version of it with a nonoptimal constant. Beckner in Ref. 12 later gave a proof of Hirschman's inequality with the best constant. My analogue to Hirschman's inequality expresses the nontriviality of the Segal-Bargmann transform as it acts on individual functions, and it brings to the forefront the entropy of the elements of the relevant Hilbert spaces as an appropriate quantity for analysis. Specifically, the Segal-Bargmann transform does not preserve entropy as calculations presented in Section VII. As we shall see the entropy may increase, decrease or remain unchanged. So my analogue to Hirschman's inequality provides a constraint on the change in entropy. As Carlen points out in Ref. 13, any result delimiting the range of the Segal-Bargmann transform can be considered as a generalized Heisenberg principle, since such a delimitation is a way of describing how a physical state can be situated in phase space. (However, beware that Carlen does not use the ground state representation in his paper.[13])

I also present a similar analysis where a weight factor is introduced into the measure of the target space. I then establish in Theorem 5.1 that there exists some Gaussian weight factor making the Segal-Bargmann transform a compact operator from L^p to ''weighted'' L^q if $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$. I also prove an estimate on the operator norm. Again, I do this by computing the Hille-Tamarkin norm for this situation. To improve the estimates, I again use an interpolation

argument in Theorem 5.3, although this time based on a theorem of Stein in Ref. 14. These interpolation estimates can be thought of as ''weighted'' analogues to the Hausdorff-Young theorem. Then I apply the differentiation technique of Hirschman, and I thereby prove in Theorem 6.1 ''weighted'' analogues to Hirschman's inequality of the form

$$a_1 S(Af) + a_2 S(f) \leqslant a_3 \|f\|_2^2 + a_4 \langle f, Nf \rangle \tag{1.1}$$

for f in a dense subspace of $L^2(\mathbf{R}^n, \nu)$ where N is the energy, or number, operator and $\langle f, Nf \rangle$ is its associated Dirichlet form. The constants a_i are given explicitly in Theorem 6.1, and they can have various algebraic signs. This energy-entropy inequality can be interpreted as a log-Sobolev inequality (see Refs. 15 and 16) for the Segal-Bargmann transform itself. This leads to a generalization of the concept of a log-Sobolev inequality, which is defined in Definition 6.2 Again, following the remark of Carlen in Ref. 13, these results are generalizations of the Heisenberg principle. The article ends by quoting a few specific calculations for some Gaussian functions.

It should be noted that the techniques of this article are mainly those of real analysis. Specifically, as is usually the case in L^p analysis, Hölder's inequality is heavily relied on. Nonetheless, interpolation theory (via the three lines lemma of Hadamard) does depend on complex analysis. However, the fact that the functions in the Segal-Bargmann space are analytic enters only in an interesting identity (6.2) of Bargmann, while the existence of a reproducing kernel function for this space is not used at all. (See Ref. 17 for a similar analysis of the reproducing kernel function.) This is a little strange, since the fact that the range of the Segal-Bargmann transform is contained in the set of entire functions is already a generalized Heisenberg principle, according to the observation of Carlen.

II. PRELIMINARIES

We begin by introducing some notation and stating some basic results. For each integer $n \geqslant 1$ we define the Segal-Bargmann space $HL^2(\mathbf{C}^n, \mu)$ (Refs. 1–3) to be the Hilbert space of all entire analytic functions $f: \mathbf{C}^n \to \mathbf{C}$ which are square integrable with respect to the Gaussian measure μ given for $z \in \mathbf{C}^n$ by

$$d\mu(z) := \pi^{-n} \exp(-z^* \cdot z) d^n x d^n y. \tag{2.1}$$

Here for each $z = (z_1, \ldots, z_n) \in \mathbf{C}^n$ and $w = (w_1, \ldots, w_n) \in \mathbf{C}^n$ we define $z^* := (z_1^*, \ldots, z_n^*)$ and $z \cdot w := z_1 w_1 + \cdots + z_n w_n$. Also λ^* denotes the complex conjugate of the complex number λ and $d^n x d^n y$ is Lebesgue measure in $\mathbf{C}^n$. We write $z^2 := z \cdot z$ for $z \in \mathbf{C}^n$. In particular, $t^2 := t \cdot t$ for $t \in \mathbf{R}^n$. The configuration space, sometimes called the ground state representation, is $L^2(\mathbf{R}^n, \nu)$ where ν is the Gaussian measure defined by

$$d\nu(t) := \pi^{-n/2} \exp(-t^2) d^n t, \tag{2.2}$$

where $t \in \mathbf{R}^n$ and $d^n t$ is Lebesgue measure in $\mathbf{R}^n$. For $z \in \mathbf{C}^n$ and measurable $f: \mathbf{R}^n \to \mathbf{C}$ the Segal-Bargmann transform A is defined by

$$Af(z) := \int_{\mathbf{R}^n} d\nu(t) A(z,t) f(t) \tag{2.3}$$

provided that the integral converges. Here the kernel function is defined by

$$A(z,t) := \exp(-z^2/2 + \sqrt{2} z \cdot t) \tag{2.4}$$

for $z \in \mathbf{C}^n$ and $t \in \mathbf{R}^n$. It follows from Hölder's inequality that Af is well defined for all $f \in L^p(\mathbf{R}^n, \nu)$ with $1 < p \leqslant \infty$, and Af is also entire in this case. A basic theorem (Refs. 1–3) says that $A: L^2(\mathbf{R}^n, \nu) \to HL^2(\mathbf{C}^n, \mu)$ is a unitary, onto map. In this formulation, the Segal-Bargmann transform A is a map between two L^2 spaces, each being with respect to a probability measure. While this is not the formulation of Bargmann in Ref. 3, where Lebesgue measure in $\mathbf{R}^n$ is used,

it is more natural in that it relates spaces for which hypercontractivity results are known (see Refs. 9, 13, 18, 19). It is however the formulation of Segal in an infinite dimensional context where Lebesgue measure does not exist (see Refs. 1 and 2).

If (X,μ) is a measure space and $1\leqslant p\leqslant\infty$, then the L^p norm is denoted by $\|f\|_p$ or by $\|f\|_{L^p(\mu)}$ depending on how explicit we need to be. The corresponding complex Banach space is denoted by $L^p(X,\mu)$, or simply by L^p. If (X,μ) and (Y,ν) are measure spaces and B is a linear transformation defined on a dense subspace D of $L^p(X,\mu)$ for $1\leqslant p\leqslant\infty$ and if $Bf: Y\to\mathbf{C}$ is measurable for all $f\in D$, then we define the operator norm by

$$\|B\|_{p\to q} := \sup\{\|Bf\|_{L^q(\nu)} : f\in D \text{ and } \|f\|_{L^p(\mu)}=1\} \tag{2.5}$$

for any $1\leqslant q\leqslant\infty$. Notice that this norm is a number in $[0,\infty]$. If this operator norm is finite, then as is well known B can be extended uniquely to a continuous linear map from $L^p(X,\mu)$ to $L^q(Y,\nu)$. Though this notation $\|\cdot\|_{p\to q}$ is standard, it is ambiguous, since the measure spaces and the measures are suppressed from it. However, context should make the meaning clear.

The spaces $HL^2(\mathbf{C}^n,\mu)$ and $L^2(\mathbf{R}^n,\nu)$ have standard unbounded self-adjoint energy, or number, operators acting in them. The operator $N := -\frac{1}{2}\Delta + t\cdot\nabla$ acts in $L^2(\mathbf{R}^n,\nu)$, where ∇ is the gradient and Δ is the Laplacian while the operator $\widetilde{N} := z_1(\partial/\partial z_1)+\cdots+z_n(\partial/\partial z_n)$ acts in $HL^2(\mathbf{C}^n,\mu)$ where we use the usual definition

$$\frac{\partial}{\partial z_k} := \frac{1}{2}\left(\frac{\partial}{\partial x_k} - i\,\frac{\partial}{\partial y_k}\right).$$

Here $z_k\in\mathbf{C}$ is written as $z_k = x_k + iy_k$ with x_k, $y_k\in\mathbf{R}$. The Segal-Bargmann transform A intertwines N and $\widetilde{N}$, namely, $AN=\widetilde{N}A$ (modulo domain considerations). We refer to this fact by saying that A preserves the energy. Another way to express this is in terms of the associated Dirichlet forms defined by

$$\langle f, Nf\rangle := \sum_{k=1}^{n}\left\|\frac{\partial f}{\partial x_k}\right\|^2_{L^2(\nu)} \tag{2.6}$$

for all $f\in L^2(\mathbf{R}^n,\nu)$ and by

$$\langle g, \widetilde{N}g\rangle := \sum_{k=1}^{n}\left\|\frac{\partial g}{\partial z_k}\right\|^2_{L^2(\mu)} \tag{2.7}$$

for all $g\in HL^2(\mathbf{C}^n,\mu)$. These Dirichlet forms are then preserved by the Segal-Bargmann transform A, namely $\langle f,Nf\rangle = \langle Af,\widetilde{N}Af\rangle$ for all $f\in L^2(\mathbf{R}^n,\nu)$.

Given a measure space (Ω,μ) and $f\in L^2(\Omega,\mu)$, define the entropy $S(f)$ by

$$S(f) := \int_\Omega d\mu |f|^2 \log|f|^2 - \|f\|_2^2 \log\|f\|_2^2. \tag{2.8}$$

Here $\log x$ is the natural logarithm (base e) of the real number $x>0$. We always use the standard convention $0\log 0 = 0$. If $\mu(\Omega)=\infty$, the integral in (2.8) may not make sense, in which case $S(f)$ is undefined. And if it does make sense, it may be $+\infty$ or $-\infty$. However, if (Ω,μ) is a probability measure space, $S(f)$ is always defined and by Jensen's inequality one gets that $S(f)\geqslant 0$ though $S(f)=\infty$ is possible. Shannon[20] originally introduced this concept of entropy in information theory.

Since the Segal-Bargmann transform A is an integral kernel operator, we will use some standard results from the theory of such operators. In the following (X,μ) and (Y,ν) are σ-finite measure spaces and $p,q\in[1,\infty]$. We say that any measurable function $K: X\times Y\to\mathbf{C}$ is a kernel function. The transpose kernel $K^T: Y\times X\to\mathbf{C}$ is defined by $K^T(y,x) := K(x,y)$ for all $(y,x)\in Y\times X$. Next, we define the Hille-Tamarkin norm (see Refs. 21, 22) of the kernel K by

$$|||K|||_{p,q} := \|K_p\|_{L^q(\mu)}, \tag{2.9}$$

where $K_p(x) := \|K(x,\cdot)\|_{L^{p'}(\nu)}$ for all $x \in X$. Here p' denotes the usual dual index of p. In particular, if $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$ then

$$|||K|||_{p,q} = \left\{ \int_X d\mu(x) \left(\int_Y d\nu(y) |K(x,y)|^{p'} \right)^{q/p'} \right\}^{1/q} \tag{2.10}$$

and in the special case $p = q = 2$, this becomes the Hilbert-Schmidt norm. Notice that the notation $|||\cdot|||_{p,q}$ has the same ambiguity as the operator norm notation, namely, the measure spaces and their measures are suppressed from the notation. Again, the meaning should be clear from context. Given a kernel function K, we define its associated integral operator by

$$Kf(x) := \int_Y d\nu(y) K(x,y) f(y) \tag{2.11}$$

for $x \in X$ and measurable $f: Y \to \mathbf{C}$ whenever this integral converges. This conforms with the usual practice, already used above, of using the same symbol for the kernel function and for its associated operator. If the integral operator K given in (2.11) has a kernel function with a finite Hille-Tamarkin norm for some $p, q \in [1, \infty]$, we say the operator K is a Hille-Tamarkin operator from $L^p(Y, \nu)$ to $L^q(X, \mu)$. This terminology is justified by the first part of the following result from Ref. 22. (We continue to use the notation and hypotheses as above.)

Proposition 2.1: $\|K\|_{p \to q} \leqslant |||K|||_{p,q}$ *and* $\|K\|_{p \to q} \leqslant |||K^T|||_{q',p'}$.

The proof is by Hölder's inequality. Here is another result from Ref. 22.

Proposition 2.2: If $|||K|||_{p,q} < \infty$ *and* $1 \leqslant q < \infty$ *and* $1 < p \leqslant \infty$ *then* K *is a compact operator from* $L^p(Y, \nu)$ *to* $L^q(X, \mu)$.

Of course, in the case $p = q = 2$, this is the well known result that a Hilbert-Schmidt operator is compact. Here is one more result from Ref. 22. It tells us which of the two upper bounds in Proposition 2.1 is better.

Proposition 2.3: If $p' \geqslant q$, *then* $|||K^T|||_{q',p'} \leqslant |||K|||_{p,q}$. *If* $p' \leqslant q$, *then* $|||K|||_{p,q} \leqslant |||K^T|||_{q',p'}$.

This follows from Minkowski's inequality for integrals in the case $1 < p \leqslant \infty$ and $1 \leqslant q < \infty$. The remaining cases are equally straightforward.

III. L^p MAPPING PROPERTIES

Theorem 3.1: *For the Segal-Bargmann transform* A *to be a Hille-Tamarkin operator from* $L^p(\mathbf{R}^n, \nu)$ *to* $L^q(\mathbf{C}^n, \mu)$ *a necessary and sufficient condition is that* p *and* q *satisfy the inequalities* $1 \leqslant q < 2$ *and* $p > 1 + q/2$. *In this case,* A *is a compact operator with*

$$\|A\|_{p \to q} \leqslant |||A|||_{p,q} = (1 + q/2 - p'q/2)^{-n/2q} (1 - q/2)^{-n/2q} < \infty \tag{3.1}$$

and

$$\|A\|_{p \to q} \leqslant |||A^T|||_{q',p'} = (1 + q/2 - p'q/2)^{-n/2p'} (1 - q/2)^{-n/2q} (1 + q/2)^{-n/2(1/q - 1/p')} < \infty. \tag{3.2}$$

Notice that these conditions on p *and* q *imply that* $p > q$.

Proof: We compute the Hille-Tamarkin norms of A, the Segal-Bargmann transform, by evaluating the Gaussian integrals explicitly. First, notice that for $z \in \mathbf{C}^n$ and $t \in \mathbf{R}^n$, and using (2.4), we have

$$|A(z,t)| = |\exp(-z^2/2 + \sqrt{2}z \cdot t)| = \exp(-x^2/2 + y^2/2 + \sqrt{2}x \cdot t), \tag{3.3}$$

where $z = x + iy$ with $x, y \in \mathbf{R}^n$. (In the following, this will always be the meaning of the symbols x, y and z.) If $1 < p \leqslant \infty$, we have

$$A_p(z) = \|A(z,\cdot)\|_{p'} = \exp((-x^2 + y^2 + p'x^2)/2) \tag{3.4}$$

by computing the Gaussian integral. Also, $A_1(z)=\|A(z,\cdot)\|_\infty=\infty$ if $x\neq 0$ so that $|||A|||_{1,q}=\infty$ for all $1\leqslant q\leqslant\infty$. Next for $1\leqslant q<\infty$ and $1<p\leqslant\infty$, again by computing the Gaussian integral, we have that

$$|||A|||_{p,q}=\|A_p\|_q=(1+q/2-p'q/2)^{-n/2q}(1-q/2)^{-n/2q}<\infty$$

if and only if $p>1+q/2$ and $1\leqslant q<2$. If either of these inequalities fails, then $|||A|||_{p,q}=\infty$, but if they both hold we have $\|A\|_{p\to q}\leqslant|||A|||_{p,q}<\infty$. Also, $|||A|||_{p,\infty}=\|A_p\|_\infty=\infty$, if $p>1$ by (3.4). The estimate (3.2) is shown by a similar calculation. The theorem now follows from Propositions 2.1 and 2.2 □

In the rest of this article we shall use the estimate (3.2) though everything works *mutatis mutandis* with the estimate (3.1). However, as we shall see later, (3.2) is the better estimate in the case which will interest us.

It is important to notice that $|||A|||_{2,2}=\infty$ even though $\|A\|_{2\to 2}=1$. This is so since the range of A is infinite dimensional (as is its domain $L^2(\mathbf{R}^n,\nu)$) and so the unitary operator A sends the closed unit ball in $L^2(\mathbf{R}^n,\nu)$ onto a closed, noncompact subset in $L^2(\mathbf{C}^n,\mu)$. In short, A is not a compact operator from L^2 to L^2, and so by Proposition 2.2, $|||A|||_{2,2}=\infty$. Worse yet, according to (3.1), $|||A|||_{p,q}\to\infty$ as $(p,q)\to(2,2)$. Also similar remarks apply to $|||A^T|||_{q',p'}$. Because we will be using interpolation theory to mollify this behavior, it is more convenient to use the points (p^{-1},q^{-1}) in the unit square $S:=[0,1]\times[0,1]$ which correspond to the indices (p,q) in $[1,\infty]\times[1,\infty]$. We will use the standard conventions that $\infty^{-1}=0$ and $0^{-1}=\infty$. We say that a point (p^{-1},q^{-1}) in S is *admissible* if $\|A\|_{p\to q}<\infty$.

Theorem 3.2: *If* $1\leqslant q<2$ *and* $p>1+q/2$ *and* $p(s)$ *and* $q(s)$ *are defined implicitly by*

$$p(s)^{-1}=sp^{-1}+(1-s)2^{-1}\qquad\text{and}\qquad q(s)^{-1}=sq^{-1}+(1-s)2^{-1}\tag{3.5}$$

for $0\leqslant s\leqslant 1$, *then*

$$1\leqslant\|A\|_{p(s)\to q(s)}\leqslant(\|A\|_{p\to q})^s\leqslant(|||A^T|||_{q',p'})^s<\infty.\tag{3.6}$$

Proof: By Theorem 3.1 (p^{-1},q^{-1}) is admissible. But for any admissible point (p^{-1},q^{-1}) in S we apply the Riesz-Thorin theorem (see Refs. 23 and 24) to interpolate along the line segment in S between (p^{-1},q^{-1}) and $(2^{-1},2^{-1})$, the latter point also being admissible. The result is

$$\|A\|_{p(s)\to q(s)}\leqslant(\|A\|_{2\to 2})^{1-s}(\|A\|_{p\to q})^s=(\|A\|_{p\to q})^s.$$

The lower bound in (3.6) follows from $\|A\|_{p_1\to q_1}\geqslant 1$ for all $p_1,q_1\in[1,\infty]$ which is true because $A1=1$ where 1 denotes the function that is constantly equal to 1. Now Theorem 3.1 and Propositions 2.1 and 2.2 establish the theorem. □

The inequality (3.6) is analogous to the Hausdorff-Young inequality, which is also obtained by using the Riesz-Thorin theorem. However, (3.6) is really a family of inequalities, one for each pair (p,q) satisfying the hypotheses, whereas there is only one Hausdorff-Young inequality in Fourier analysis. Notice that the estimate (3.6) on the operator norm goes to 1 as $s\to 0$. Also $(p(s)^{-1},q(s)^{-1})$ goes to $(2^{-1},2^{-1})$ along a line segment in S as $s\to 0$.

IV. HIRSCHMAN TYPE INEQUALITIES

We can now apply a technique of Hirschman from Ref. 11. We take (p^{-1},q^{-1}) satisfying $1\leqslant q<2$ and $p>1+q/2$ and $p(s)$, $q(s)$ with $0\leqslant s\leqslant 1$ as in equations (3.5) above. Then by (3.6)

$$\|Af\|_{q(s)}\leqslant(|||A^T|||_{q',p'})^s\|f\|_{p(s)},\tag{4.1}$$

where f is a fixed function in $L^2(\mathbf{R}^n,\nu)$. The point here is that this inequality reduces to an *equality* at $s=0$ since A is an L^2 isometry. Then formally the derivative of (4.1) at $s=0$ generates a new inequality. This is the idea behind the next result, which is analogous to an entropy inequality in Fourier analysis first proved by Hirschman in Ref. 11. But first we quote two well known results that will allow us to differentiate under the integral sign. The proofs use standard measure theory and convexity arguments. Here (Ω,μ) is a probability measure space and

$\psi:[1,\infty)\to[0,\infty]$ is defined by $\psi(r) := \int_\Omega d\mu |f|^r$ where $r\in[1,\infty)$ and $f:\Omega\to\mathbf{C}$ is measurable. The first result says that if $f\in L^2(\Omega,\mu)$, then the left side derivative of $\psi(r)$ at $r=2$ is given by

$$\psi'(2^-) := \lim_{h<0,h\to 0} \frac{\psi(2+h)-\psi(2)}{h} = \int_\Omega d\mu |f|^2 \log|f| \tag{4.2}$$

which is either a finite real number or $+\infty$. The second result says that if $f\in L^{2+\epsilon}(\Omega,\mu)$ for some $\epsilon>0$, then the right side derivative of $\psi(r)$ at $r=2$ is given by

$$\psi'(2^+) := \lim_{h>0,h\to 0} \frac{\psi(2+h)-\psi(2)}{h} = \int_\Omega d\mu |f|^2 \log|f| \tag{4.3}$$

which is a finite real number.

Theorem 4.1: *The Hirschman type entropy inequality*

$$(p^{-1}-2^{-1})S(f) \leqslant (q^{-1}-2^{-1})S(Af) + (\log|||A^T|||_{q',p'})\|f\|_2^2 \tag{4.4}$$

holds for all $f\in L^2(\mathbf{R}^n,\nu)$ *where* $1\leqslant q<2$ *and* $p>1+q/2$ *and* $|||A^T|||_{q',p'}$ *is given in (3.2).*

Proof: The result is trivial if $f=0$, so we only consider nonzero $f\in L^2(\mathbf{R}^n,\nu)$. As noted above, (4.1) holds under our hypotheses here. Next by taking the right side derivative of (4.1) at $s=0$, one obtains a new inequality

$$\left.\frac{d}{ds}\right|_{s=0+} \|Af\|_{q(s)} \leqslant \left.\frac{d}{ds}\right|_{s=0+} (|||A^T|||_{q',p'})^s \|f\|_{p(s)}$$

or, equivalently,

$$\left.\frac{d}{ds}\right|_{s=0+} \|Af\|_{q(s)} \leqslant (\log|||A^T|||_{q',p'})\|f\|_2 + \left.\frac{d}{ds}\right|_{s=0+} \|f\|_{p(s)}, \tag{4.5}$$

where the right side derivative is used since (4.1) holds for $0\leqslant s\leqslant 1$. In general, derivation is not order preserving, but in this case it is since at $s=0$ the inequality (4.1) becomes an *equality*. Of course, we must show the existence of the derivatives in (4.5). The inequality (4.1) is trivial for a fixed $f\in L^2(\mathbf{R}^n,\nu)$ if $\|f\|_{p(s)}$ is infinite for all $0<s\leqslant 1$. To make the following argument work we need $\|f\|_{p(s)}$ to be finite for all $s\in(0,\epsilon]$ for some $\epsilon>0$. If $1\leqslant p<2$, then $p(s)<2$ and so $\|f\|_{p(s)}\leqslant\|f\|_2<\infty$ for all $0<s\leqslant 1$. This means that we can take $\epsilon=1$ in this case. (As we shall see later, the case $p\geqslant 2$ is trivial.) Also notice that by differentiating (3.5) implicitly with respect to s we have

$$p'(0)=4(2^{-1}-p^{-1}) \quad\text{and}\quad q'(0)=4(2^{-1}-q^{-1}). \tag{4.6}$$

The reader should beware that $p'(0)$ here is a derivative and is totally distinct from p', the dual index.

Suppose that $p<2$. Then $p(s)$ increases to 2 as s decreases to zero. So by (4.2) and elementary calculus we have

$$\left.\frac{d}{ds}\right|_{s=0^+} \|f\|_{p(s)} = \frac{p'(0)S(f)}{4\|f\|_2}$$

with $0\leqslant S(f)\leqslant\infty$ where the entropy $S(f)$ is defined in (2.8). Similarly,

$$\left.\frac{d}{ds}\right|_{s=0^+} \|Af\|_{q(s)} = \frac{q'(0)S(Af)}{4\|Af\|_2}$$

with $0\leqslant S(f)\leqslant\infty$ follows from the hypothesis $1\leqslant q<2$. So inequality (4.5) says

$$\frac{q'(0)S(Af)}{4\|Af\|_2} \leqslant (\log|||A^T|||_{q',p'})\|f\|_2 + \frac{p'(0)S(f)}{4\|f\|_2}. \tag{4.7}$$

Notice that $p'(0)<0$ and $q'(0)<0$. So $S(f)=\infty$ implies $S(Af)=\infty$. Since the coefficients of the entropy terms in (4.4) are positive in this case we see that (4.4) is trivially true if $S(f)=\infty$. If $S(f)<\infty$, we can rearrange the terms in the inequality (4.7), multiply by $\|f\|_2=\|Af\|_2$ and use (4.6) to get (4.4).

It remains to show (4.4) in the case $p\geqslant 2$. But then $p^{-1}-2^{-1}\leqslant 0$, and so (4.4) is trivially true since the left hand side is nonpositive and the right hand side is nonnegative. (Recall that $\log|||A^T|||_{q',p'}\geqslant \log\|A\|_{p\to q}\geqslant \log 1=0$.) This terminates the proof of the theorem. □

As the proof shows, the significance of inequality (4.4) arises in the case $p<2$, which is not trivial because then the coefficients of the entropy terms in (4.4) are positive. But $p<2$ together with $p>1+q/2$ imply that $p'>q$, so that by Proposition 2.3 the estimate (3.2) is better than (3.1) in the nontrivial case. By (3.2), the coefficient of $\|f\|_2^2$ is $nC_{p,q}$ where

$$C_{p,q}=\frac{1}{n}\log|||A^T|||_{q',p'}=-\frac{1}{2}\left\{\frac{1}{p'}\log\left(1+\frac{q}{2}-\frac{p'q}{2}\right)+\frac{1}{q}\log\left(1-\frac{q}{2}\right)+\left(\frac{1}{q}-\frac{1}{p'}\right)\log\left(1+\frac{q}{2}\right)\right\}.$$

This makes it the only coefficient in (4.4) with a dependence on the dimension n. Also for any admissible (p^{-1},q^{-1}) we have by the same proof that

$$(p^{-1}-2^{-1})S(f)\leqslant (q^{-1}-2^{-1})S(Af)+(\log\|A\|_{p\to q})\|f\|_2^2.$$

It remains an open question whether an inequality of the form $S(f)\leqslant cS(Af)$ holds for all $f\in L^2(\mathbf{R}^n,\nu)$ where $c>0$ is some constant independent of f. This would hold, for example, if one could show that $\|A\|_{p\to q}=1$ for some p and q satisfying $1<p<2$ and $1\leqslant q<2$. If such an inequality holds, then it would be interesting to know the best constant c for which it holds. If such an inequality does not hold, we can rewrite (4.4) (in the nontrivial case $p<2$) in the form $S(f)\leqslant cS(Af)+\epsilon\|f\|_2^2$ with c and ϵ positive. So the best constant for the coefficient of $S(Af)$ is a function of ϵ, say $c(\epsilon)$. Then it would be interesting to know the behavior of $c(\epsilon)$ as a function of ϵ. Since $|||A^T|||_{q',p'}>1$, the coefficient of the $\|f\|_2$ term in the log-Sobolev inequality (4.4) proved here is strictly positive. Similarly, $|||A|||_{p,q}>1$.

By considering characteristic functions in $L^2(\mathbf{R}^n,\nu)$, we can easily see that states (that is, functions f satisfying $\|f\|_2=1$) which are highly concentrated have arbitrarily large entropy while states which are not concentrated have small, though nonnegative, entropy. Also, the measure space $(\mathbf{C}^n,\mu)$ is isomorphic to the measure space $(\mathbf{R}^{2n},\nu)$ since putting $2n$ in place of n in (2.2) yields (2.1) using the usual identification of $\mathbf{R}^{2n}$ and $\mathbf{C}^n$. So the same remarks hold for the entropy of states in $L^2(\mathbf{C}^n,\mu)$, and hence for its closed subspace $HL^2(\mathbf{C}^n,\mu)$. Because of this behavior, some authors refer to $S(f)$ as an information, and reserve the word entropy for the quantity $-S(f)$. Regardless of the terminology, the fact is that a sequence of states $\{f_k\}_{k\geqslant 0}$ in $L^2(\mathbf{R}^n,\nu)$, which become more and more concentrated, transforms under the Segal-Bargmann transform to a sequence of states $\{Af_k\}_{k\geqslant 0}$ in $HL^2(\mathbf{C}^n,\mu)$, which also become more and more concentrated. It should be emphasized that the concentration of the states is being measured with respect to a Gaussian measure; the nonconstant elements of the Segal-Bargmann space are unbounded near infinity, since they are entire functions. It should be noted again that the calculations in Section VII show that the Segal-Bargmann transform does not preserve entropy, so that (4.4) gives a constraint on the change in entropy.

See Hirschman's article[11] for the first appearance of the entropy inequality which bears his name. Beckner in Ref. 12 proves this inequality with the best constant. There is also a nice discussion of this in the paper of Białynicki-Birula and Mycielski.[25] Hirschman's inequality expresses the idea of the Heisenberg uncertainty principle in Fourier analysis, and the analogue proved in (4.4) does much the same in Segal-Bargmann analysis.

V. WEIGHTED ANALOGUES OF HAUSDORFF-YOUNG INEQUALITY

As we will see in Section VII, the operator norm $\|A\|_{p\to q}$ can be infinite for some $p,q\in[1,\infty]$. The next theorem shows how a change in measure can give a different result.

Theorem 5.1: *Let* $d\mu_{ab}(z) := c_{ab}\exp(-ax^2-by^2)d\mu(z)$ *define a probability measure where* $z\in\mathbf{C}^n$, *a and b are real numbers and* $c_{ab}=(1+a)^{n/2}(1+b)^{n/2}$ *is a normalization constant. For there to exist real numbers a and b such that A is a Hille-Tamarkin operator from* $L^p(\mathbf{R}^n,\nu)$ *to* $L^q(\mathbf{C}^n,\mu_{ab})$ *a necessary and sufficient condition is that p and q satisfy* $1<p\leqslant\infty$ *and* $1\leqslant q<\infty$. *If this holds, A is a compact operator, and a and b can be taken to be any real numbers satisfying*

$$a>-1-q/2+p'q/2 \quad \text{and} \quad b>-1+q/2, \tag{5.1}$$

and consequently

$$\|Af\|_{L^q(\mu_{ab})}\leqslant |||A|||_{p,q}\|f\|_{L^p(\nu)} \tag{5.2}$$

for all $f\in L^p(\mathbf{R}^n,\nu)$ *where*

$$|||A|||_{p,q}=c_{ab}^{1/q}((a+1+q/2-p'q/2)(b+1-q/2))^{-n/2q}<\infty. \tag{5.3}$$

Under the same restrictions on p, q, a and b one has

$$\|Af\|_{L^q(\mu_{ab})}\leqslant |||A^T|||_{q',p'}\|f\|_{L^p(\nu)} \tag{5.4}$$

for all $f\in L^p(\mathbf{R}^n,\nu)$ *where*

$$|||A^T|||_{q',p'}=c_{ab}^{1/q}\left(a+1+\frac{q}{2}\right)^{-(n/2)((1/q)-(1/p'))}\left(b+1-\frac{q}{2}\right)^{-(n/2q)}\left(a+1+\frac{q(1-p')}{2}\right)^{-(n/2p')}.$$

Remark: Notice that (5.1) implies $a>-1$ and $b>-1$ so that c_{ab}, and hence μ_{ab}, is well defined. Proposition 2.3 tells us which of the two inequalities (5.2) and (5.4) is better.

Proof: The question is if it is possible to find $a,b\in\mathbf{R}$ so that $|||A|||_{p,q}<\infty$, where the Hille-Tamarkin norm is with respect to the measures ν and μ_{ab}. So we compute this Hille-Tamarkin norm. Since the definition of $A_p(z)$ is the same as in Section III, we again have equation (3.4) for $1<p\leqslant\infty$. By doing the Gaussian integral for $1\leqslant q<\infty$, we have

$$|||A|||_{p,q}=\|A_p\|_{L^q(\mu_{ab})}=c_{ab}^{1/q}((a+1+q/2-p'q/2)(b+1-q/2))^{-n/2q}<\infty$$

provided that (5.1) holds. Clearly, for any $1<p\leqslant\infty$ and $1\leqslant q<\infty$ there are solutions a and b in $\mathbf{R}$ of the inequalities (5.1). So, the operator is Hille-Tamarkin in this case, and hence compact by Proposition 2.2. The inequality (5.2) follows from Proposition 2.1 This shows the sufficiency of the condition on p and q.

Let's next analyze the cases $p=1$ and $q=\infty$. First, if $p=1$, then by (3.3)

$$A_1(z)=\sup_{t\in\mathbf{R}^n}|A(z,t)|=\sup_{t\in\mathbf{R}^n}\exp(-x^2/2+y^2/2+\sqrt{2}x\cdot t)=\infty \tag{5.5}$$

for all $z=x+iy\in\mathbf{C}^n$ with $x\neq 0$. So it follows that $|||A|||_{1,q}=\|A_1\|_q=\infty$ for any $1\leqslant q\leqslant\infty$. Next, if $q=\infty$ and $p>1$, then we have by (3.4)

$$|||A|||_{p,\infty}=\sup_{z\in\mathbf{C}^n}|A_p(z)|=\sup_{z\in\mathbf{C}^n}\exp((-x^2+y^2+p'x^2)/2)=\infty. \tag{5.6}$$

Summarizing, $|||A|||_{p,q}=\infty$ if $p=1$ or $q=\infty$. This shows the necessity of the conditions on p and q and so concludes the proof up to (5.3). The remainder of the theorem follows from similar calculations. □

If we select a and b satisfying (5.1) and so that $a=b$ (which is always possible), we can think of the codomain space in terms of a Segal-Bargmann space, but with a value of Planck's constant different from that of the domain space. See Woodhouse[26] for a presentation of the Segal-Bargmann space with an explicit Planck's constant. We can consider the codomain space to be a generalized Segal-Bargmann space even in the case $a\neq b$, though now with *two* distinct values of

Planck's constant. From now on we will use estimate (5.2) rather than estimate (5.4) even though by Proposition 2.3 the latter is sometimes sharper. Everything works *mutatis mutandis* with the latter estimate.

Again, the behavior of the estimate $|||A|||_{p,q}$ of the operator norm as (p,q) approaches (2,2) needs to be improved. As the formula (5.3) shows, the bound $|||A|||_{p,q}$ depends on a and b (as well as on p and q), but a and b depend on p and q only through the inequalities (5.1), and not in a functional way. Also, when $p=q=2$, we find from (5.1) that $a>0$ and $b>0$, so that the estimate (5.2) in this case is not the statement that A is an isometry into the unweighted L^2 space. So we will apply interpolation theory to this situation in order to get better estimates of the operator norms. First, pick p and q with $1<p\leqslant\infty$ and $1\leqslant q<\infty$. Next, pick a and b in $\mathbf{R}$ satisfying inequalities (5.1). Then with respect to the measures ν and μ_{ab} one has that $\|A\|_{p\to q}\leqslant|||A|||_{p,q}<\infty$. Since A is an L^2 isometry, $\|A\|_{2\to 2}=1$ with respect to the unweighted L^2 norms. We interpolate between these two norm estimates using an interpolation theorem of E. M. Stein, which is proved as Theorem 2 in Ref. 14 and as Theorem 3.6 in Ref. 27. Since it is not as well known as the Riesz-Thorin theorem, we give a statement of this theorem taken from Ref. 14. Recall that a simple function is defined to be a measurable function f with finite range I such that $f^{-1}(i)$ has finite measure for all nonzero $i\in I$.

Theorem 5.2 (Stein): *Let (M,ν) and (N,μ) be σ-finite measure spaces. Let T be a linear transformation defined on simple functions of M to measurable functions on N. Suppose p_0, p_1, q_0 and q_1 are in $[1,\infty]$ and define p_s and q_s for $0\leqslant s\leqslant 1$ by $p_s^{-1}=(1-s)p_0^{-1}+sp_1^{-1}$ and $q_s^{-1}=(1-s)q_0^{-1}+sq_1^{-1}$. Suppose that for some weight functions $k_0,k_1:N\to[0,\infty)$ and $u_0,u_1:M\to[0,\infty)$ and for all simple functions $f:M\to\mathbf{C}$ one has*

$$\|(Tf)k_0\|_{q_0}\leqslant M_0\|fu_0\|_{p_0} \tag{5.7}$$

and

$$\|(Tf)k_1\|_{q_1}\leqslant M_1\|fu_1\|_{p_1} \tag{5.8}$$

for some M_0 and M_1 in $\mathbf{R}$. (The right hand side of either of these inequalities may be infinite for some f.) Define new weight functions $k_s:=k_0^{1-s}k_1^s$ on N and $u_s:=u_0^{1-s}u_1^s$ on M for $0\leqslant s\leqslant 1$. Then T can be uniquely extended to a linear transformation on all functions f satisfying $\|fu_s\|_{p_s}<\infty$ so that

$$\|(Tf)k_s\|_{q_s}\leqslant M_0^{1-s}M_1^s\|fu_s\|_{p_s}. \tag{5.9}$$

Let us note that this theorem can be interpreted as saying that the measures as well as the Lebesgue indices vary in the interpolation. It is a consequence of Theorem 1 of Ref. 14, which is usually called the Stein interpolation theorem. See also Refs. 23, 24, and 27. We now have the following result:

Theorem 5.3: *Suppose that $1<p\leqslant\infty$ and $1\leqslant q<\infty$. Then A is a bounded operator from $L^{p(s)}(\mathbf{R}^n,\nu)$ into $L^{q(s)}(\mathbf{C}^n,\mu_{ab}^s)$ for all $0\leqslant s\leqslant 1$ and for all real numbers a and b satisfing (5.1) where the interpolating indices are*

$$p(s)^{-1}=(1-s)2^{-1}+sp^{-1} \qquad \text{and} \qquad q(s)^{-1}=(1-s)2^{-1}+sq^{-1}$$

and the interpolating measure is

$$d\mu_{ab}^s(z)=\exp\left(-\frac{sq(s)}{q}(ax^2+by^2)\right)d\mu(z).$$

Moreover, $\|A\|_{p(s)\to q(s)}\leqslant(|||A|||_{p,q})^s$ where $|||A|||_{p,q}$ is given by (5.3). This means that

$$\|Af\|_{L^{q(s)}(\mu_{ab}^s)}\leqslant(|||A|||_{p,q})^s\|f\|_{L^{p(s)}(\nu)} \tag{5.10}$$

for all $0\leqslant s\leqslant 1$ and for all $f\in L^{p(s)}(\mathbf{R}^n,\nu)$, or equivalently,

$$\|(Af\,)k_1^s\|_{L^{q(s)}(\mu)} \leq (|||A|||_{p,q})^s \|f\|_{L^{p(s)}(\nu)}, \tag{5.11}$$

where we define

$$k_1(z) := \exp\left(-\frac{1}{q}(ax^2+by^2)\right) \tag{5.12}$$

for all $z \in \mathbf{C}^n$.

Proof: Let a and b satisfy (5.1). Then, in Stein's Theorem 5.2, we take the Lebesgue indices to be $p_0=q_0=2$ and $p_1=p$ and $q_1=q$. We also take the weight functions to be $k_0(z)=1$ and $k_1(z)$ as in (5.12) and $u_0(t)=u_1(t)=1$ where $z \in \mathbf{C}^n$ and $t \in \mathbf{R}^n$. We then have that

$$\|(Af\,)k_1\|_{L^{q(\mu)}} = \|Af\|_{L^{q(\mu_{ab})}} \leq |||A|||_{p,q} \|f\|_{L^{p(\nu)}} < \infty$$

for all $f \in L^p(\mathbf{R}^n, \nu)$ by the relations (5.2) and (5.3) in Theorem 5.1, which is applicable here. This verifies the hypothesis (5.8) of Stein's Theorem 5.2 with $M_1 = |||A|||_{p,q}$. The hypothesis (5.7) holds since A is an L^2 isometry, namely $\|Af\|_2 = M_0\|f\|_2$ where $M_0=1$. Then Stein's Theorem 5.2 gives $k_s(z)=k_1^s(z)$ and $u_s(t)=1$ for $0 \leq s \leq 1$, $z \in \mathbf{C}^n$ and $t \in \mathbf{R}^n$. So (5.9) becomes (5.11). Using the identity $\|(Af\,)k_1^s\|_{L^{q(s)}(\mu)} = \|Af\|_{L^{q(s)}(\mu_{ab}^s)}$ we get (5.10) from (5.11). □

These inequalities (5.11) are the weighted analogues of the Hausdorff-Young inequality for the current situation. This is because, as in the unweighted case, they are derived by interpolating between a (p,q) mapping estimate and a (2,2) estimate, thereby giving us an operator norm estimate that is well behaved in the interpolating parameter. Also the estimate depends only on one choice of constants a and b, namely that at the endpoint corresponding to $s=1$.

VI. LOGARITHMIC SOBOLEV INEQUALITIES

The differentiation technique of Hirschman again can be applied to the weighted analogues of the Hausdorff-Young inequality of the previous section. This now leads to entropy-energy inequalities which, as we shall see, are generalized logarithmic Sobolev inequalities.

Theorem 6.1: *Suppose that* $f \in L^2(\mathbf{R}^n, \nu)$ *satisfies* $Af \in L^{2+\epsilon}(\mathbf{C}^n, \mu)$ *for some* $\epsilon > 0$. *Suppose also that* $1 < p \leq \infty$ *and* $1 \leq q < \infty$. *If* $p > 2$, *suppose also that* $f \in L^{2+\xi}(\mathbf{R}^n, \nu)$ *for some* $\xi > 0$. *Then for any real number* a *satisfying* $a > \max(-1-q/2+p'q/2, -1+q/2)$ *we have the entropy-energy inequality*

$$(2^{-1}-q^{-1})S(Af\,) - (2^{-1}-p^{-1})S(f\,) \leq \frac{a}{q}\langle f, Nf\rangle + \Lambda\|f\|_2^2, \tag{6.1}$$

where $\Lambda = (n/q)[a - \frac{1}{2}\log((a+1+q/2-p'q/2)(a+1-q/2))]$ *and where* $\langle f, Nf\rangle$ *is the Dirichlet form defined in (2.6). Here* $S(f\,)$, $S(Af\,)$ *and* $\langle f, Nf\rangle$ *are all finite nonnegative real numbers.*

Proof: Since the results are trivial for $f=0$, we assume $f \in L^2(\mathbf{R}^n, \nu)$ is nonzero hereafter. First note that by (4.3) and the hypothesis on Af we have that $S(Af\,) < \infty$. The proof that $\langle f, Nf\rangle$ is finite is based on the following identity which comes from Bargmann's article,[3] formula (3.17). It says that for every $g \in HL^2(\mathbf{C}^n, \mu)$ one has

$$\int_{\mathbf{C}^n} d\mu(z)|g(z)|^2|z|^2 = n\|g\|_2^2 + \langle g, \widetilde{N}g\rangle, \tag{6.2}$$

where the Dirichlet form $\langle g, \widetilde{N}g\rangle$ is defined in (2.7). This is proved by expanding g in its Taylor series. But, by Hölder's inequality, we obtain

$$\int_{\mathbf{C}^n} d\mu(z)|g(z)|^2|z|^2 \leq \|g\|_{2r}^2 \|z\|_{2r'}^2, \tag{6.3}$$

for any $1 \leq r \leq \infty$. Here $\|z\|_{2r'}$ really means $\|\,|z|\,\|_{2r'}$. Taking $g=Af$ and $r=1+\epsilon/2$ where $\epsilon>0$ is chosen so that $Af \in L^{2+\epsilon}(\mathbf{C}^n, \mu)$ we have $\|g\|_{2r} = \|Af\|_{2+\epsilon} < \infty$. But $r>1$ and so we have $r' < \infty$. Consequently, $\|z\|_{2r'} < \infty$. Thus, by (6.2) and (6.3)

$$n\|Af\|_2^2+\langle Af,\widetilde{N}Af\rangle=\int_{\mathbf{C}^n}d\mu(z)|Af(z)|^2|z|^2\leqslant\|Af\|_{2+\epsilon}^2\|z\|_{2r'}^2<\infty, \tag{6.4}$$

which implies $\langle f,Nf\rangle=\langle Af,\widetilde{N}Af\rangle<\infty$ since A preserves energy. It should be noted that the identity (6.2) of Bargmann depends on the analyticity of the elements of the Segal-Bargmann space. Moreover, the inequality (6.4) is the *reverse* of a Sobolev inequality, since it says that if a function in the Segal-Bargmann space possesses just a bit more integrability than square integrability, then the function has finite expected energy.

To show the inequality (6.1) we start with the inequality (5.11) of Theorem 5.3 which does apply here because of our current hypotheses, provided that a and b also satisfy (5.1). So for $0\leqslant s\leqslant 1$ we have

$$\|(Af)k_1^s\|_{L^{q(s)}(\mu)}\leqslant(|||A|||_{p,q})^s\|f\|_{L^{p(s)}(\nu)}.$$

Since this becomes an equality at $s=0$ for $f\in L^2(\mathbf{R}^n,\nu)$, we have by the same argument which gave us (4.5) that

$$\left.\frac{d}{ds}\right|_{s=0^+}\|(Af)k_1^s\|_{q(s)}\leqslant(\log|||A|||_{p,q})\|f\|_2+\left.\frac{d}{ds}\right|_{s=0^+}\|f\|_{p(s)}, \tag{6.5}$$

provided that the derivatives on both sides of (6.5) exist. By the rules of calculus and a differentiation under the integral sign (which will be justified shortly), the derivative on the left hand side is

$$\left.\frac{d}{ds}\right|_{s=0^+}\|(Af)k_1^s\|_{q(s)}=\frac{q'(0)S(Af)}{4\|Af\|_2}+\frac{1}{\|Af\|_2}\int d\mu(z)\left.\frac{d}{ds}\right|_{s=0^+}k_1(z)^s|Af(z)|^2.$$

In Ref. 17 it is shown that

$$\left.\frac{d}{ds}\right|_{s=0^+}\int_{\mathbf{C}^n}d\mu(z)(k_1(z)^s|g(z)|)^{q(s)}$$
$$=\int_{\mathbf{C}^n}d\mu(z)\left(2|g(z)|^2\left.\frac{d}{ds}\right|_{s=0^+}(k_1(z)^s)+q'(0)|g(z)|^2\log|g(z)|\right),$$

where $g\in HL^2(\mathbf{C}^n,\mu)$, provided that $g\in L^{2+\epsilon}(\mathbf{C}^n,\mu)$ for some $\epsilon>0$. This justifies the above derivation under the integral sign. Using both

$$\left.\frac{d}{ds}\right|_{s=0^+}k_1(z)^s=\log k_1(z)=\frac{1}{q}\log c_{ab}-\frac{1}{q}(ax^2+by^2)$$

and $\|Af\|_2=\|f\|_2$ we then obtain

$$\left.\frac{d}{ds}\right|_{s=0^+}\|(Af)k_1^s\|_{q(s)}=\frac{q'(0)S(Af)}{4\|f\|_2}+\frac{1}{q}\|f\|_2\log c_{ab}-\frac{1}{q\|f\|_2}\int_{\mathbf{C}^n}d\mu(z)|Af(z)|^2(ax^2+by^2).$$

Notice that this derivative is a finite real number, since the integral term is bounded above and below by finite real numbers by Bargmann's identity (6.2) and since the entropy $S(Af)$ is also finite.

By taking $a=b>\max(-1-q/2+p'q/2,-1+q/2)$, one gets

$$\left.\frac{d}{ds}\right|_{s=0^+}\|(Af)k_1^s\|_{q(s)}=\frac{q'(0)S(Af)}{4\|f\|_2}+\frac{n}{q}\|f\|_2\log(1+a)-\frac{a}{q\|f\|_2}\int_{\mathbf{C}^n}d\mu(z)|Af(z)|^2|z|^2$$

because $c_{aa}=(1+a)^n$. Next we use Bargmann's identity (6.2), the unitarity of A, and the fact that A preserves energy to get

$$\left.\frac{d}{ds}\right|_{s=0^+} \|(Af)k_1^s\|_{q(s)} = \frac{q'(0)S(Af)}{4\|f\|_2} + \frac{n}{q}\|f\|_2 \log(1+a) - \frac{an}{q}\|f\|_2 - \frac{a}{q\|f\|_2}\langle f, Nf\rangle. \quad (6.6)$$

Now we establish

$$\left.\frac{d}{ds}\right|_{s=0^+} \|f\|_{p(s)} = \frac{p'(0)S(f)}{4\|f\|_2} \quad (6.7)$$

by examining three cases. If $p<2$, (6.7) follows from (4.2). If $p>2$, (6.7) follows from (4.3) and the extra hypothesis on f in the case $p>2$. If $p=2$, then $p(s)=2$ for all $0\leqslant s\leqslant 1$ and so $p'(0)=0$. Then both sides of (6.7) are zero, where we use the usual convention that $0\cdot\infty=0$. So $S(f)=\infty$ is still a possibility if $p\leqslant 2$. By substituting (6.6) and (6.7) into (6.5) and using (5.3) in the case $a=b$ we have (6.1). The finiteness of $S(f)$ now follows from (6.1) in the case $1<p<2$. This concludes the proof. □

It is easy to verify that the set of functions satisfying the hypotheses of Theorem 6.1 is a dense subspace of $L^2(\mathbf{R}^n,\nu)$. Also, it is straightforward to show by elementary calculus that $\Lambda\geqslant(.28)n/q>0$. The coefficient of $S(f)$ in the inequality (6.1) can be positive, negative or zero. The same holds for the coefficient of $S(Af)$. So we can have nine qualitatively different types of inequalities in (6.1). The question of best constants is a bit complicated for an inequality

$$a_1S(Af)+a_2S(f)\leqslant a_3\|f\|_2^2+a_4\langle f,Nf\rangle$$

of the form (6.1). While it is also possible to show such an inequality with $a_3=0$, it may be possible to prove it with $a_3<0$. In general, by dividing by $|a_1|$ or $|a_2|$ to normalize one of the coefficients on the left hand side, the question is to find the best constant for the other coefficient on the left hand side, given coefficients for the two terms on the right hand side. Notice that the case $a_1=a_2=0$ trivially has best constants $a_3=a_4=0$. This case arises in (6.1) for $p=q=2$, which is a valid case of Stein's theorem, namely the case when the Lebesgue indices do not vary but only the weight functions vary. In the case $a_1=0$, $a_2>0$ (respectively, $a_1>0$, $a_2=0$) one obtains a log-Sobolev inequality in the space $L^2(\mathbf{R}^n,\nu)$ (respectively, $HL^2(\mathbf{C}^n,\mu)$). It should be noted that this manner of demonstrating log-Sobolev inequalities for the individual spaces arises from the L^p mapping properties of the Segal-Bargmann transform, and not from hypercontractivity estimates. Next, let us consider the case $a_4<0$ which arises in (6.1) if $1\leqslant q<2$ and $p>1+q/2$. Notice that $q<2$ implies $a_1<0$ and that $a_2<0$ if $p>2$. By rearranging terms if $p>2$ and $q<2$ one has the energy bounded by a positive linear combination of the entropies, modulo the $\|f\|_2$ term. Similarly if $q<2$ and $p<2$ one has an upper bound on $S(f)$ in terms of $S(Af)$, again modulo the norm and energy terms. This seems weaker than (4.4), but it is not always since it includes the possibility that $a_4<0$. Finally, when $a_1>0$ and $a_2<0$ (that is, $p>2$ and $q<2$) we can bound $S(Af)$ above by $S(f)$, modulo norm and energy terms as usual. In this case, as an examination of the results in Proposition 7.1 shows, one must have $a_4>0$. Moreover, a recent result in Ref. 17 shows that one has an upper bound on $S(Af)$ by nothing more than the energy and norm terms, which means that we can take $a_2=0$ in this case.

Actually, the inequality (6.1) itself can be considered as a generalization of a log-Sobolev inequality, if it is thought of as a log-Sobolev inequality of the operator A, which preserves the norms and the Dirichlet forms. The next definition captures this generalization.

Definition 6.2: Let (X,μ) and $(\widetilde{X},\widetilde{\mu})$ be probability measure spaces with nonnegative self-adjoint operators N and $\widetilde{N}$ acting in $L^2(X,\mu)$ and in a closed subspace T of $L^2(\widetilde{X},\widetilde{\mu})$ respectively. Suppose that A is a unitary map from $L^2(X,\mu)$ onto T, and that A intertwines N and $\widetilde{N}$ in the sense that A preserves the Dirichlet forms. Then we say that A satisfies a logarithmic Sobolev inequality if there exist constants c_1, c_2, c_3 and c_4 (excluding the case when both c_1 and c_2 are nonpositive and both c_3 and c_4 are nonnegative) such that

$$c_1S(Af)+c_2S(f)\leqslant c_3\|f\|_2^2+c_4\langle f,Nf\rangle \quad (6.8)$$

for all f in some dense subspace of $L^2(X,\mu)$ for which the right hand side of (6.8) is finite.

Here the Dirichlet form $\langle f,Nf\rangle$ is defined to be $\|N^{1/2}f\|_2^2$ for all f in the domain of the operator $N^{1/2}$ (which is the unique nonnegative square root of the nonnegative self-adjoint operator N). Otherwise, $\langle f,Nf\rangle$ is defined to be $+\infty$.

An inequality of this sort has already appeared in an article of Carlen. (See Theorem 6 of Ref. 28.) In the case when (X,μ) and $(\widetilde{X},\widetilde{\mu})$ are equal, $N=\widetilde{N}$ and A is the identity operator, this definition reduces to the usual definition of a log-Sobolev inequality (see Ref. 16).

VII. CALCULATIONS

In this section we present the results of some specific calculations.

Proposition 7.1: Define $g_\lambda(t) := \exp(\lambda\cdot t-(\mathrm{Re}\,\lambda)^2/2)$ *in* $L^2(\mathbf{R}^n,\nu)$ *where* $t\in\mathbf{R}^n$ *and* $\lambda\in\mathbf{C}^n$. *Then we have* $\|g_\lambda\|_2=1$ *and*

$$S(g_\lambda)=(\mathrm{Re}\,\lambda)^2,$$

$$Ag_\lambda(z)=\exp\left(\frac{\sqrt{2}\lambda\cdot z}{2}+\frac{\lambda^2}{4}-\frac{(\mathrm{Re}\,\lambda)^2}{2}\right),$$

$$S(Ag_\lambda)=|\lambda|^2/2,$$

$$\|g_\lambda\|_{L^p(\nu)}=\exp\left(\left(\frac{p-2}{4}\right)(\mathrm{Re}\,\lambda)^2\right),$$

$$\|Ag_\lambda\|_{L^q(\mu)}=\exp\left(\left(\frac{q-2}{8}\right)|\lambda|^2\right),$$

$$\langle g_\lambda,Ng_\lambda\rangle=\langle Ag_\lambda\widetilde{N}Ag_\lambda\rangle=|\lambda|^2/2,$$

where $z\in\mathbf{C}^n$, $1\leqslant p<\infty$ *and* $1\leqslant q<\infty$.

The proof of this proposition is a computation of the Gaussian integrals (see Ref. 29). As the functions g_λ show, the Segal-Bargmann transform A can increase entropy, decrease entropy or leave the entropy unchanged.

Here is a corollary of Proposition 7.1 which complements Theorem 3.1

Corollary 7.2: $\|A\|_{p\to q}=\infty$ *if* $q>2$ *or if* $p<1+q/2$ *where* A *is the Segal-Bargmann transform. Notice these conditions on* p *and* q *are implied by* $p<q$.

Proof: If $1\leqslant p<\infty$ and $1\leqslant q<\infty$ we use the functions g_λ from Proposition 7.1 to get the lower bound

$$\|A\|_{p\to q}\geqslant\sup_{\lambda\in\mathbf{C}^n}\frac{\|Ag_\lambda\|_q}{\|g_\lambda\|_p}=\sup_{\lambda\in\mathbf{C}^n}\exp\left(\left(\frac{q-2p+2}{8}\right)(\mathrm{Re}\,\lambda)^2+\left(\frac{q-2}{8}\right)(\mathrm{Im}\,\lambda)^2\right).$$

So, $\|A\|_{p\to q}=\infty$ if $q>2$ or if $q-2p+2>0$. For $p=\infty$, we use g_λ with $\lambda\in i\mathbf{R}^n$. Then $\|g_\lambda\|_\infty=1$ and for $1\leqslant q<\infty$ we have

$$\|A\|_{\infty\to q}\geqslant\sup_{\lambda\in\mathbf{C}^n}\frac{\|Ag_\lambda\|_q}{\|g_\lambda\|_\infty}=\sup_{\lambda\in\mathbf{C}^n}\exp\left(\left(\frac{q-2}{8}\right)|\lambda|^2\right),$$

and so it follows that $\|A\|_{\infty\to q}=\infty$ if $2<q<\infty$. Finally, for $q=\infty$, we use the inequality $\|A\|_{p\to\infty}\geqslant\|A\|_{\infty\to q_1}$ for any $2<q_1<\infty$ and $1\leqslant p\leqslant\infty$ (which follows from Hölder's inequality) and the previous case. □

Zhou in Ref. 9 has shown the case $q>2$ of Corollary 7.2 by considering the functions g_λ for $\lambda\in i\mathbf{R}^n$. Also, by Hölder's inequality, for $q=2$ and $2\leqslant p\leqslant\infty$, we have $\|A\|_{p\to q}=1$. So, together with Theorem 3.1, we see that the only indices p and q for which the finiteness or infiniteness of $\|A\|_{p\to q}$ has yet to be established are those satisfying $q<2$ and $p=1+q/2$.

VIII. CONCLUDING REMARKS

The Hille-Tamarkin norm is a useful tool for analyzing the Segal-Bargmann transform because these norms become Gaussian integrals, which are then explicitly evaluated. This happens since the measures (2.1) and (2.2) and the kernel (2.4) are Gaussians. So this sort of analysis works in a more general Gaussian situation, for example in Ref. 17 where the reproducing kernel function of the Segal-Bargmann space is analyzed. However, the Hille-Tamarkin norm provides a weak estimate on the operator norm. Explicitly, this means that the inequalities in Proposition 2.1 need not be sharp. Nonetheless, this method does correctly describe the general behavior of the Segal-Bargmann transform with respect to the L^p scale of spaces. Specifically, for all pairs (p,q) except those with $q=2$ and $p\geqslant 2$ and those with $1\leqslant q<2$ and $p=1+q/2$, we have that either A is compact from $L^p(\mathbf{R}^n,\nu)$ to $L^q(\mathbf{C}^n,\mu)$ or $\|A\|_{p\to q}=\infty$. Similar behavior is seen with the integral transform whose kernel is the reproducing kernel function of the Segal-Bargmann space. This transform is also generically either compact or unbounded (see Ref. 17). So the unitarity of the Segal-Bargmann transform from $L^2(\mathbf{R}^n,\nu)$ to $HL^2(\mathbf{C}^n,\mu)$ is atypical of the behavior of this transform with respect to the L^p scale of spaces. Another way to see this is in terms of the integral transform L whose kernel function is $L(z,t) := |A(z,t)|$ for all $z\in\mathbf{C}^n$ and $t\in\mathbf{R}^n$ where $A(z,t)$ is defined in (2.4). Then $|||L|||_{p,q}=|||A|||_{p,q}$ for all $p,q\in[1,\infty]$, so that L is Hille-Tamarkin for a pair of indices if and only if A is. Moreover, by using the functions g_λ of Proposition 7.1, one sees that $\|L\|_{p\to q}=\infty$ if either $q\geqslant 2$ or if $p<1+q/2$. Note that these are not quite the same conditions on p and q as in Corollary 7.2. So the transform L is also generically compact or unbounded. Also the only indices p and q for which the finiteness or infiniteness of $\|L\|_{p\to q}$ has yet to be determined are those with $q<2$ and $p=1+q/2$, and these are precisely the same indices for which the same question remains open for $\|A\|_{p\to q}$. But the operator L is *not* a unitary operator on $L^2(\mathbf{R}^n,\nu)$. This follows from $\|L\|_{2\to 2}=\infty$ which is a particular case of the above analysis or simply by noting that $L1$ is not square integrable. So the phase information in the kernel function $A(z,t)$ is playing a critical role in the L^2 theory, but is generically unimportant on the L^p scale of spaces.

All of the results presented so far in this article are consistent with what might be called the ''one-infinity'' law for the operator norm $\|A\|_{p\to q}$. This would mean that for every pair of indices $p,q\in[1,\infty]$, one would have either $\|A\|_{p\to q}=1$ or $\|A\|_{p\to q}=\infty$. However, calculations with Gaussian functions show that $\|A\|_{p_0\to q_0}>(1.007)^n$ for $p_0=1.6$ and $q_0=1.1$. But these indices also satisfy $1\leqslant q_0<2$ and $p_0>1+q_0/2$ so that $\|A\|_{p_0\to q_0}<\infty$ by Theorem 3.1. Therefore the ''one-infinity'' law fails for the Segal-Bargmann transform, and the operator norm can have nontrivial dependence on the dimension n. This may be different from the case of the integral transform associated with the reproducing kernel function of the Segal-Bargmann space, where the question of whether the ''one-infinity'' law holds remains open (see Ref. 17).

There are various avenues for continued research here. Since there are many generalizations and deformations of the Segal-Bargmann transform, I hope that these techniques can be applied to some of them in order to understand better the nature of the range of these various transforms. Some references for these transforms are the book of Perelomov[30] and the articles of Thomas and Wassell[31] and of Hall.[32] Among the open problems in this field of research is to establish exactly the range of the L^p space under the Segal-Bargmann transform and to find the best constants for the various inequalities proved in this paper. Finally, it would be interesting to know as well the relationships between entropy and the Segal-Bargmann transform when n, the number of degrees of freedom, is not finite.

ACKNOWLEDGMENTS

I wish to thank my dissertation advisor, Lawrence Thomas, for his continual assistance and encouragement with my mathematical education and particularly with this research, most of which was done during my graduate career at the University of Virginia. I also thank T. Goebeler, B. Hall and R. Strichartz for reading a preliminary version of this paper and offering many useful comments. I also wish to thank Brian Hall for providing me with a preprint of Ref. 8. Part of this research was started during the U.S.-Sweden Workshop on Spectral Methods held at the Mittag-Leffler Institute, Djursholm, Sweden, 1992-1993 (NSF Grant No. 9217529) during the year on Schrödinger operators. I thank the organizers (B. Helffer, P. Hislop, A. Jensen, and J. Sjöstrand) and the participants for the stimulating mathematical environment at the Mittag-Leffler Institute.

[1] I. E. Segal, *Mathematical Problems of Relativistic Physics* (American Mathematical Society, Providence, 1963).
[2] I. E. Segal, *The Complex-Wave Representation of the Free Boson Field* in *Topics in Functional Analysis, Advances in Mathematics, Supplementary Studies* (Academic, New York, 1978), Vol. 3.
[3] V. Bargmann, ''On a Hilbert space of analytic functions and an associated integral transform, Part I,'' Commun. Pure Appl. Math. **14**, 187–214 (1961).
[4] E. Fischer, ''Über die differentiatensprozesse der algebra,'' J. Math. **148**, 1–78 (1917).
[5] V. Fock, ''Verallgemeinerung und lösung der diracshen statistischen gleichung,'' Z. Phys. **49**, 339–357 (1928).
[6] M. H. Stone, ''Linear transformation in Hilbert space III. Operational methods and group theory,'' Proc. Natl. Acad. Sci. USA **16**, 172–175 (1930).
[7] J. von Neumann, ''Die eindeutigkeit der Schrödingerschen operatoren,'' Math. Ann. **104**, 570–578 (1931).
[8] B. C. Hall, *Quantum Mechanics in Phase Space*, in *Proceedings of the Joint Summer Research Conference on Quantization*, edited by M. Rieffel (American Mathematical Society, Providence, 1997).
[9] Z. Zhou, ''The contractivity of the free Hamiltonian semigroup in the L_p space of entire functions,'' J. Funct. Anal. **96**, 407–425 (1991).
[10] E. H. Lieb, ''Gaussian kernels have only Gaussian maximizers,'' Invent. Math. **102**, 179–208 (1990).
[11] I. I. Hirschman, Jr., ''A note on entropy,'' Am. J. Math. **79**, 152–156 (1957).
[12] W. Beckner, ''Inequalities in Fourier analysis,'' Ann. Math. **102**, 159–182 (1975).
[13] E. Carlen, ''Some integral identities and inequalities for entire functions and their application to the coherent state transform,'' J. Funct. Anal. **97**, 231–249 (1991).
[14] E. Stein, ''Interpolation of linear operators,'' Trans. Am. Math. Soc. **83**, 482–492 (1956).
[15] P. Federbush, ''A partially alternate derivation of a proof of Nelson,'' J. Math. Phys. **10**, 50–52 (1969).
[16] L. Gross, ''Logarithmic Sobolev inequalities,'' Am. J. Math. **97**, 1061–1083 (1975).
[17] S. B. Sontz, ''On the reproducing kernel of the Segal-Bargmann space'' (preprint, 1997).
[18] E. Nelson, ''The free Markoff field,'' J. Funct. Anal. **12**, 211–227 (1973).
[19] S. Janson, ''On hypercontractivity for multipliers on orthogonal polynomials,'' Ark. Mat. **21**, 97–110 (1983).
[20] C. Shannon and W. Weaver, *A Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).
[21] E. Hille and J. D. Tamarkin, ''On the theory of linear integral equations. II,'' Ann. Math. **35**, 445–455 (1934).
[22] K. Jörgens, *Linear Integral Operators* (Pitman, Boston, 1982).
[23] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vol. 2.
[24] E. Stein and G. Weiss, *Introduction to Fourier Analysis in Euclidean Space* (Princeton University Press, Princeton, 1971).
[25] I. Białynicki-Birula and J. Mycielski, ''Uncertainty relations for information entropy in wave mechanics,'' Commun. Math. Phys. **44**, 129–132 (1975).
[26] N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980).
[27] C. Bennett and R. Sharpley, *Interpolation of Operators* (Academic, Orlando, 1988).
[28] E. Carlen, ''Superadditivity of Fisher's information and logarithmic Sobolev inequalities,'' J. Funct. Anal. **101**, 194–211 (1991).
[29] S. B. Sontz, ''L^p mapping properties of the Segal-Bargmann transform,'' Ph.D. dissertation, University of Virginia, 1994.
[30] A. M. Perelomov, *Generalized Coherent States and Their Applications* (Springer, Berlin, 1986).
[31] L. Thomas and S. Wassell, ''Semiclassical approximation for Schrödinger operators on a two-sphere at high energy,'' J. Math. Phys. **36**, 5480–5505 (1995).
[32] B. C. Hall, ''The Segal-Bargmann coherent state transform for compact Lie groups,'' J. Funct. Anal. **122**, 103–151 (1994).

Ordered products, W_∞-algebra, and two-variable, definite-parity, orthogonal polynomials

A. Verçin[a)]
Department of Physics, Ankara University, Faculty of Sciences, 06100, Tandoğan-Ankara, Turkey

(Received 10 September 1997; accepted for publication 10 November 1997)

It has been shown that the Cartan subalgebra of W_∞-algebra is the space of the two-variable, definite-parity polynomials. Explicit expressions of these polynomials, and their basic properties are presented. It also has been shown that they carry the infinite dimensional irreducible representation of the su(1,1) algebra having the spectrum bounded from below. A realization of this algebra in terms of difference operators is also obtained. For particular values of the ordering parameter s they are identified with the classical orthogonal polynomials of a discrete variable, such as the Meixner, Meixner–Pollaczek, and Askey–Wilson polynomials. With respect to variable s they satisfy a second order eigenvalue equation of hypergeometric type. Exact scattering states with zero energy for a family of potentials are expressed in terms of these polynomials. It has been put forward that it is the İnönü–Wigner contraction and its inverse that form a bridge between the difference and differential calculus. © *1998 American Institute of Physics.* [S0022-2488(98)00304-1]

I. INTRODUCTION

The classical orthogonal polynomials of a discrete variable (the Hahn, Meixner, Krawtchouk, Charlier, Pollaczek, and Meixner–Pollaczek polynomials) which are the difference analogues of the classical orthogonal polynomials of the mathematical physics have surfaced in various problems of the theoretical and mathematical physics, in group representation theory and in computational physics and techniques (Refs. 1–3 and references therein). Close relationships have been established between the generalized spherical harmonics for su(2) and the Krawtchouk polynomials, and between the Wigner $6j$-symbols and the Racah polynomials which are the discrete analogues of the Jacobi polynomials on a quadratic lattice.[1,4] The Clebsch–Gordan coefficients and $6j$-symbols for the su(1,1) can also be expressed in terms of the Hahn and Racah polynomials.[4] An important development in connection with these polynomials in the last decade was that the Hahn and Meixner polynomials can be analytically continued in the complex plane, both in variable and parameter, such that they become real polynomials which satisfy orthogonality relations with respect to continuous measure.[5,2] These polynomials are referred to as the continuous Hahn and Meixner, or, as the Hahn and Meixner polynomials of an imaginary argument (in fact, the second ones should be called the Meixner–Pollaczek polynomials[2]). The continuous Hahn polynomials are also closely related to the unitary irreducible representations of the Lorentz group SO(3,1).

Recent interest in these polynomials are mainly due to modifications and generalizations of them in connection with their q-analogues on non-uniform lattices (Ref. 6 and references therein). Another line of development, which has not attracted the sufficient interest it deserves, was determination of their connections with the ordered products of the Heisenberg–Weyl (HW) algebra.[7,8] Connection between the ordered products of the HW-algebra and particular classes of the continuous Hahn polynomials was, for the first time, investigated in Ref. 7 to find a solution for the unitarity problem in the finite element approximation to quantum field equations. This connection was established when ordering rules that gives Hermitian products are used.[7,8] In terms of ordering parameter $s \in \mathbf{C}$, henceforth used in this paper, this corresponds to pure imaginary values of s of which the well-known Weyl ordering is a special $s=0$ case. One of the purposes of

[a)]Electronic mail: vercin@science.ankara.edu.tr

 © 1998 American Institute of Physics

this study is to revive interest in this direction by carrying out a systematic investigation in the most general framework.

In two recent studies,[9,10] we managed to develop explicit expressions for the implicitly defined conventional s-ordered products widely used in the Weyl–Wigner–Groenevold–Moyal quantization and quantum optics.[9–11] In Section II, we briefly review the ordered products, W_∞-algebra and its Cartan subalgebra in their most general forms. Making use of these expressions, we firstly associate with each element of the Cartan subalgebra of the W_∞-algebra in a general s-basis a two-variable, definite-parity, and (continuous and discrete) orthogonal polynomials which are two-variable generalization of those appeared in literature. Our variables are s, and the c-number correspondence of the operator $\hat{x}=(\hat{q}\hat{p}+\hat{p}\hat{q})/2$ denoted by x (or, in some cases the variable $u=(x/i\hbar)-(1/2)$. These polynomials solve a hypergeometric type differential equation with respect to s, and a hypergeometric type difference equation with respect to x. For generic values of s this difference equation is identical with that satisfied by a particular class of Meixner polynomials. Other basic properties, and difference-differential relations are also presented. We also show that they carry the irreducible representation of the su(1,1)-algebra and obtain its realization in terms of difference operators (Section III). Secondly, for particular values of s, we identify these polynomials with particular classes of the Meixner, Meixner–Pollaczek, Askey–Wilson, and Hahn polynomials, and give the discrete and continuous orthogonality relations for them (Section IV). Thirdly, we show that exact zero-energy scattering states for a family of the Pösch–Teller type potentials are expressed in terms of these polynomials (Section V). Finally, we conclude by pointing out the role played by the İnönü–Wigner contraction and its inverse transformation in connection with differential-difference calculus.

II. ORDERED PRODUCTS, W_∞ – ALGEBRA AND ITS CARTAN SUBALGEBRA

Let us consider the HW-algebra: $[\hat{q},\hat{p}]=i\hbar\hat{I}$, where $\hbar,\hat{I},\hat{q}$ and $\hat{p}$ are the Planck's constant, the identity operator and the Hermitian position and momentum operators, respectively. Here and henceforth operators and functions of operators acting in a Hilbert space $\mathscr{H}$ are denoted by a $\hat{}$ over letters. We define the s-ordered products $\hat{t}^{(s)}_{nm}\equiv\{(\hat{q})^n(\hat{p})^m\}_s$, containing n factors of $\hat{q}$ and m factors of $\hat{p}$, by the following explicit equivalent formulas

$$\begin{aligned}\hat{t}^{(s)}_{nm}&=2^{-n}\sum_{j=0}^{n}\binom{n}{j}(1+s)^j(1-s)^{n-j}\hat{q}^j\hat{p}^m\hat{q}^{n-j}\\&=2^{-m}\sum_{k=0}^{m}\binom{m}{k}(1-s)^k(1+s)^{m-k}\hat{p}^k\hat{q}^n\hat{p}^{m-k}.\end{aligned}\tag{1}$$

These give for $s=\pm1,\hat{t}^{(1)}_{nm}=\hat{q}^n\hat{p}^m$; $\hat{t}^{(-1)}_{nm}=\hat{p}^m\hat{q}^n$ and for $s=0$, $\hat{t}^{(0)}_{nm}=2^{-n}\Sigma^n_{j=0}\binom{n}{j}\hat{q}^j\hat{p}^m\hat{q}^{n-j}=2^{-m}\Sigma^m_{k=0}\binom{m}{k}\hat{p}^k\hat{q}^n\hat{p}^{m-k}$. While the first two of these expressions exhibit the standard and antistandard rules of ordering, respectively, that corresponding to $s=0$ are two well-known expressions of the Weyl, or symmetrically ordered products. It is possible to write many equivalent forms of the above relations, but, for later use only two of them have been written. Although there are not any known physical applications apart from the three principle ones corresponding to $s=1,0,-1$, embedding orderings in a continuum provides a natural context for viewing their differences and interrelationships in a continuous manner and enable us to carry out the related analyses in their most general forms.

An arbitrary s-ordered product can be expressed in terms of a polynomial in s'-ordered product as follows

$$\hat{t}^{(s)}_{nm}=\sum_{k=0}^{(n,m)}2^{-k}\binom{n}{k}\binom{m}{k}k![i\hbar(s-s')]^k\hat{t}^{(s')}_{n-k,m-k},\tag{2}$$

where (n,m) denotes the smaller of the integers n and m, $\binom{n}{k}=n![(n-k)!k!]^{-1}$ is a binomial coefficient, and s' is also arbitrary complex number. Note that $i\hbar$ in Eq. (2) is the sign of the

commutator of the corresponding operators there. Thus, relations (1) and (2) can be used for any pair of the operators $\hat{A},\hat{B}$ of any algebra satisfying the commutation relation $[\hat{A},\hat{B}]=i\lambda,\lambda\in\mathbf{C}$. From (1) it easily follows that

$$[\hat{t}^{(s)}_{nm}]^{\dagger}=\hat{t}^{(\bar{-s})}_{nm}, \tag{3}$$

where † stands for the Hermitian conjugation and $\bar{a}$ is the complex conjugation of a. That is, for general n,m integers, $\hat{t}^{(s)}_{nm}$ are Hermitian if and only if $\bar{s}=-s$. In particular, the Weyl ordered products $\hat{t}^{(0)}_{nm}$ are Hermitian. For general $s,\alpha\in\mathbf{C}$ one can form combinations such as $\hat{\kappa}_{nm}(s)=\alpha\hat{t}^{(s)}_{nm}+\bar{\alpha}\hat{t}^{(\bar{-s})}_{nm}$ that are Hermitian.

The W_∞-algebra is the infinite algebra generated by the ordered products $\hat{t}^{(s)}_{nm}$.[10,12,13] Up to a trivial central element it is the universal enveloping algebra of the HW-algebra. In the most general basis the structure constants of the W_∞-algebra can be read off from

$$[\hat{t}^{(s)}_{kl},\hat{t}^{(s)}_{nm}]=-\sum_{j=0}^{j_{max}}\frac{i^j}{j!}\left[\sum_{r=0}^{j'}\binom{j}{r}f_{srj}a_{nmkl,rj}\right]\hat{t}^{(s)}_{n+k-j,m+l-j}, \tag{4}$$

where the prime over the second summation indicates that the maximum value that r may take is $r_{max}=(m,k)$ and

$$j_{max}=(n+r_{max},l+r_{max}),\quad a_{nmkl,rj}=\frac{n!m!k!l!}{(n+r-j)!(m-r)!(k-r)!(l+r-j)!}. \tag{5}$$

The restrictions imposed on summations also follows from the expression of $a_{nmkl,rj}$. In relation (4) $f_{srj}=(s^-)^r(-s^+)^{j-r}-(s^-)^{j-r}(-s^+)^r$, $s^{\pm}=\hbar(1\pm s)/2$, is the only factor depending on the chosen rule of ordering. Here we observe that anti-commutator of the ordered products is given by the same relation as (4) only provided that the ordering factor is replaced by $f^{+}_{srj}=(s^-)^r(-s^+)^{j-r}+(s^-)^{j-r}(-s^+)^r$. Relation (4), which was first reported in Ref. 10, for $s=0,\pm1$ coincides with those appeared in literature,[14] and generalize them for arbitrary values of s.

W_∞ has some finite and infinite dimensional subalgebras. But, for the purpose of this work we will be concerned only with the infinite abelian subalgebra consisting of the generators $\hat{H}^{(s)}_n\equiv\hat{t}^{(s)}_{nn}$. The commutativity of the generators $[\hat{H}^{(s)}_n,\hat{H}^{(s)}_k]=0$ follows from (4). But, for our purpose we prove this, and $[\hat{H}^{(s)}_n,\hat{H}^{(s')}_k]=0$ by showing that for any values of $n\geqslant0$ and $s\in\mathbf{C}$ all the ordered products of the form $\hat{H}^{(s)}_n\equiv\hat{t}^{(s)}_{nn}$ can be expressed in terms of single operator $\hat{x}\equiv(\hat{q}\hat{p}+\hat{p}\hat{q})/2$. To see this, let us first consider $\hat{H}^{(1)}_n=\hat{q}^n\hat{p}^n$ which can be written as

$$\begin{aligned}\hat{H}^{(1)}_n&=\hat{q}^{n-1}\hat{q}\hat{p}\hat{p}^{n-1}=\hat{q}^{n-1}\hat{p}^{n-1}[\hat{x}+\hat{c}(2n-1)]\\&=\hat{q}^{n-2}\hat{p}^{n-2}[\hat{x}+\hat{c}(2n-3)][\hat{x}+\hat{c}(2n-1)].\end{aligned}$$

Hence, by induction, we have

$$\hat{H}^{(1)}_n=\prod_{j=1}^{n}[\hat{x}+\hat{c}(2j-1)],\quad \hat{H}^{(-1)}_n=\prod_{j=1}^{n}[\hat{x}-\hat{c}(2j-1)], \tag{6}$$

where $\hat{c}\equiv i\hbar\hat{I}/2$ and the relations $\hat{q}\hat{p}=\hat{x}+\hat{c}$, $[\hat{x},\hat{p}^k]=2\hat{c}k\hat{p}^k$ are used. The second relation is written by making use of (3).

III. TWO-VARIABLE, DEFINITE-PARITY POLYNOMIALS

Now, evaluating (2) for $s'=\pm1$ by making use of (6), and then replacing $\hat{x}$ in results by the c-number variable x, we obtain two-variable polynomials $P_n(s,x)$ for each element of the Cartan subalgebra. Here the variables are the (dimensionless) ordering parameter $s\in\mathbf{C}$ and $x\in\mathbf{R}$ which

has dimension of angular momentum (since $\hat{x}$ is an hermitian operator we consider x as a real variable). Two equivalent, explicit expressions for these polynomials are as follows

$$\begin{aligned} P_n(s,x) &= \sum_{k=0}^{n} \binom{n}{k}^2 k![-c(1-s)]^k \prod_{j=1}^{n-k} [x+c(2j-1)] \\ &= \sum_{k=0}^{n} \binom{n}{k}^2 k![c(1+s)]^k \prod_{j=1}^{n-k} [x-c(2j-1)], \end{aligned} \tag{7}$$

where $c=i\hbar/2$. From these relations it is obvious that under the action of two dimensional parity transformation in $\mathbf{C}\times\mathbf{R}$, they transform as

$$P_n(-s,-x)=(-1)^n P_n(s,x), \tag{8}$$

that is, they have the same parity with n. The following recursion relation can also be verified

$$P_{n+1}(s,x)=[x+c(2n+1)s]P_n(s,x)+c^2(1-s^2)n^2 P_{n-1}(s,x). \tag{9}$$

The easiest way of obtaining this relation may be first noting the relation

$$[\hat{x},\hat{t}^{(s)}_{nm}]_+ = 2[\hat{t}^{(s)}_{n+1,m+1} - cs(m+n+1)\hat{t}^{(s)}_{nm} - c^2 nm(1-s^2)\hat{t}^{(s)}_{n-1,m-1}],$$

where $[,]_+$ stands for the anticommutator. They also obey the following derivatives and difference relations:

$$\partial_s^k P_n(s,x) = c^k \left[\frac{n!}{(n-k)!}\right]^2 P_{n-k}(s,x), \tag{10}$$

$$(x\pm c)P_n(s,x\pm 2c)=[x\pm c(2n+1)]P_n(s,x)+2c^2n^2(1\mp s)P_{n-1}(s,x), \tag{11}$$

with respect to s and x, respectively. This last property immediately results by first noting the relations

$$(x\pm c)P_n(\pm 1,x\pm 2c)=[x\pm c(2n+1)]P_n(\pm 1,x), \tag{12}$$

which easily result from (6).

From (9) and (10) we see that, with respect to s the lowest order differential equation obeyed by these polynomials is the following hypergeometric type differential equation

$$\{(1-s^2)\partial_s^2+[x/c+(2n-1)s]\partial_s - n^2\}P_n(s,x)=0. \tag{13}$$

The difference relations (12) can also be recast in a form in which the discrete differences are in a more conventional form by using the central first and second differences;

$$D_h f(x)=\frac{f(x+h)-f(x-h)}{2h}, \quad D_h^2 f(x)=\frac{f(x+h)-2f(x)+f(x-h)}{h^2}, \tag{14}$$

which approximate the usual first and the second derivatives on a lattice with the constant mesh $\Delta x=h$ up to second order in h.[1,8] For this purpose we take the sum and difference of two relations given by (12)

$$(x+c)P_n(s,x+2c)-(x-c)P_n(s,x-2c)=c(4n+2)P_n(s,x)-4sc^2n^2P_{n-1}(s,x), \tag{15}$$

$$(x+c)P_n(s,x+2c)+(x-c)P_n(s,x-2c)=2xP_n(s,x)+4c^2n^2P_{n-1}(s,x), \tag{16}$$

which, in terms of D_h and D_h^2 can be rewritten as follows

$$(c^2D_h^2+xD_h-n)P_n(s,x)=-scn^2P_{n-1}(s,x), \tag{17}$$

$$(xD_h^2+D_h)P_n(s,x)=n^2P_{n-1}(s,x), \tag{18}$$

where $h=2c$ and D_h denote the partial difference operation with respect to x. Note that all the difference operations are with respect to variable x and all the derivatives are with respect to the ordering variable s.

Now, by making use of the recursion relation (9), and (17), (18) we have

$$\begin{aligned} J_-P_n(s,x)&=n^2P_{n-1}(s,x),\\ J_+P_n(s,x)&=P_{n+1}(s,x),\\ J_0P_n(s,x)&=(n+\tfrac{1}{2})P_n(s,x), \end{aligned} \tag{19}$$

where

$$J_-=[xD_h^2+D_h]=\left[\frac{4x}{\hbar^2}\sin^2\left(\frac{\hbar}{2}\partial_x\right)+\frac{1}{\hbar}\sin(\hbar\partial_x)\right], \tag{20}$$

$$\begin{aligned} J_+&=c^2(1+s^2)J_-+2cs(c^2D_h^2+xD_h)+(x+cs)\\ &=c^2(1+s^2)J_-+2cs\left[-\sin^2\left(\frac{\hbar}{2}\partial_x\right)+\frac{x}{\hbar}\sin(\hbar\partial_x)\right]+(x+cs), \end{aligned} \tag{21}$$

$$\begin{aligned} J_0&=c(c+sx)D_h^2+(x+cs)D_h+\frac{1}{2}\\ &=c(c+sx)\frac{4}{\hbar^2}\sin^2\left(\frac{\hbar}{2}\partial_x\right)+(x+cs)\frac{1}{\hbar}\sin(\hbar\partial_x)+\frac{1}{2}. \end{aligned} \tag{22}$$

It is not hard to verify that by their actions on an arbitrary function, the generators $(J_0,J_\pm)$ obey the standard defining relations of the su(1,1) algebra

$$[J_+,J_-]=-2J_0,\quad [J_0,J_\pm]=\pm J_\pm. \tag{23}$$

The value of the Casimir operator $J^2=-J_-J_++J_0^2+J_0$ is found to be $J^2P_n(s,x)=-\frac{1}{4}P_n(s,x)$. Thus, we have obtained a realization of the su(1,1) algebra in terms of the difference operators, and the polynomials we have found carry the infinite dimensional irreducible representation of this algebra having the spectrum bounded from below which is designated by $D^+_{-\frac{1}{2}}$.[15]

Finally, in this section we write down explicit expressions of the first five polynomials;

$$\begin{aligned} P_0(s,x)&=1,\\ P_1(s,x)&=x+cs,\\ P_2(s,x)&=x^2+4csx+c^2(2s^2+1),\\ P_3(s,x)&=x^3+9csx^2+c^2(18s^2+5)x+3c^3s(2s^2+3),\\ P_4(s,x)&=x^4+16csx^3+2c^2(36s^2+7)x^2+16c^3s(6s^2+5)x+3c^4(8s^4+24s^2+3). \end{aligned} \tag{24}$$

The values of these polynomials at $(s,x=0)$ and at $(s=0,x=0)$ can be easily obtained from (7).[16,17] Because of (3), these are real valued for pure imaginary s. Note that the relation (2), which enables us to write $P_n(s,x)$ in terms of $P_n(s',x)$, in view of (10), is nothing more than a Taylor expansion.

IV. CONNECTIONS WITH THE MEIXNER, MEIXNER–POLLACZEK AND CONTINUOUS HAHN POLYNOMIALS

To compare the polynomials we have found with the classical orthogonal polynomials of a discrete variable we use

$$u=\frac{x}{2c}-\frac{1}{2}. \tag{25}$$

In that case we have the following difference relations

$$(u+1)P_n(s,u+1)=(u+n+1)P_n(s,u)+c(1-s)n^2P_{n-1}(s,u), \tag{26}$$

$$uP_n(s,u-1)=(u-n)P_n(s,u)+c(1+s)n^2P_{n-1}(s,u). \tag{27}$$

Eliminating the last terms between these two relations and writing out the result in terms of the difference operators

$$\Delta f(u)=f(u+1)-f(u),\quad \nabla f(u)=f(u)-f(u-1), \tag{28}$$

we arrive at

$$\left\{u\Delta\nabla-\left[\frac{2}{1-s}u+\frac{1+s}{1-s}\right]\Delta+\frac{2n}{1-s}\right\}P_n(s,u)=0. \tag{29}$$

This is the same type of difference equation satisfied by the Meixner polynomials $m_n^{(\gamma,\mu)}(u)$:

$$\{u\Delta\nabla+[(\mu-1)u+\mu\gamma]\Delta+\lambda_n\}m_n^{(\gamma,\mu)}(u)=0. \tag{30}$$

Comparing with Eq. (29) we have

$$\mu=\frac{s+1}{s-1},\quad \gamma=1,\quad \lambda_n=\frac{2n}{1-s}. \tag{31}$$

The Meixner polynomials are normalized as follows: $\Delta^n m_n^{(\gamma,\mu)}(u)=n![(\mu-1)/\mu]^n$. On the other hand, the polynomials $P_n(s,u)$ obey the relation $\Delta^nP_n(s,u)=(2c)^nn!$. (For a polynomial $q_m(u)$ of degree m the mth differences and mth derivatives are equal: $\Delta^m q_m(u)=\nabla^m q_m(u)=\partial_u^m q_m(u)$. Thus, we obtain

$$P_n(s,u)=[c(s+1)]^n m_n^{(1,(s+1)/(s-1))}(u). \tag{32}$$

If the conditions $\gamma>0$, and $0<\mu<1$ are satisfied the Meixner's polynomials obey the discrete-orthogonality relation

$$\sum_{u=0}^{\infty} m_n^{(\gamma,\mu)}(u)m_m^{(\gamma,\mu)}(u)\rho(u)=\delta_{nm}d_n^2. \tag{33}$$

Thus, provided that $s<-1$ this relation in terms of $P_n(s,u)$ is as follows:

$$\sum_{u=0}^{\infty} P_n(s,u)P_m(s,u)\rho'(u)=\delta_{nm}d_n'^2, \tag{34}$$

where, (for $s<-1$) the weight $\rho'(u)$, and the squared norm $d_n'^2$ are found to be

$$\rho'(u)=\left(\frac{s+1}{s-1}\right)^u\frac{\Gamma(u+1)}{u!},\quad d_n'^2=\frac{1}{2}(n!)^2c^{2n}(1-s)(s^2-1)^n. \tag{35}$$

Here $\Gamma(z)$ is the gamma function. It is also easy to verify that, the three-term recursion relation (9) for the polynomials $P_n(s,u)$ is the same as that obeyed by Meixner's polynomials

$$\mu m_{n+1}^{(\gamma,\mu)}(u)=[\gamma\mu+(1+\mu)n-(1-\mu)u]m_n^{(\gamma,\mu)}(u)-n(n+\gamma-1)m_{n-1}^{(\gamma,\mu)}(u), \tag{36}$$

provided that relations given by (31) are satisfied.

In Ref. 5 by considering the analytic continuation of the orthogonality relation (33) in the parameter $\mu=\exp(-2i\phi)$ the polynomials

$$P_n^{\lambda}(\phi,t)=\frac{e^{-in\phi}}{n!}m_n^{(2\lambda,\mu)}(-\lambda+it), \tag{37}$$

which, under the conditions $\lambda>0$, and $0<\phi<\pi$, obey the orthogonality relation

$$\int_{-\infty}^{\infty}P_n^{\lambda}(\phi,t)P_m^{\lambda}(\phi,t)\rho_P(t)dt=\delta_{nm}\frac{\Gamma(2\lambda+n)}{n!} \tag{38}$$

with respect to continuous measure are obtained. Here the weight ρ_P is as follows

$$\rho_P(t)=\frac{1}{2\pi}(2\sin\phi)^{2\lambda}|\Gamma(\lambda+it)|^2\exp[(2\phi-\pi)t]. \tag{39}$$

These polynomials should be called the Meixner–Pollaczek polynomials.[2] In view of the relation (32), if we identify $\hat{x}/\hbar$ with the real variable t we find that $\lambda=1/2$ and $s=i\cot\phi$. Thus, with the help of (37), we obtain

$$P_n\left(s,i\frac{x}{\hbar}-\frac{1}{2}\right)=n!\left(\frac{-\hbar}{2\sin\phi}\right)^n P_n^{1/2}\left(\phi,\frac{x}{\hbar}\right). \tag{40}$$

In terms of $P_n(s,u)$ the orthogonality relation (38) is as follows:

$$\int_{-\infty}^{\infty}P_n\left(s,i\frac{x}{\hbar}-\frac{1}{2}\right)P_m\left(s,i\frac{x}{\hbar}-\frac{1}{2}\right)\rho_0(x)dx=\delta_{nm}(n!)^2\left(\frac{\hbar}{2\sin\phi}\right)^{2n+1}, \tag{41}$$

with the weight

$$\rho_0(x)=\frac{\exp[(2\phi-\pi)(x/\hbar)]}{\cosh(\pi x/\hbar)}. \tag{42}$$

For $\lambda=1/2$, and $\phi=\pi/2$ the Meixner–Pollaczek polynomials are related to the Askey–Wilson polynomials $q_n^{(\alpha)}(x,\delta)$ as follows $q_n^{(0)}(x,1/2)=(1/2)_n P_n^{1/2}(\pi/2,x)$, where, $(a)_n=\Gamma(a+n)/\Gamma(a)$. Note that this particular case corresponds to $s=0$ (Weyl ordering) and the weight is $\rho_0(x)=\cosh(\pi x/\hbar)$. On the other hand, the Askey–Wilson polynomials are particular cases of the continuous Hahn polynomials $h_n^{(\alpha,\beta)}(z,N)$:

$$q_n^{(\alpha)}(x,\delta)=(-i)^n h_n^{(\alpha,\alpha)}(\tfrac{1}{2}ix-\tfrac{1}{2}(\alpha-\delta+1),-\alpha+\delta). \tag{43}$$

Making use of these relations we have

$$\begin{aligned}P_n\left(s=0,i\frac{x}{\hbar}-\frac{1}{2}\right)&=n!\left(\frac{-\hbar}{2}\right)^n\left[\left(\frac{1}{2}\right)_n\right]^{-1}q_n^{(0)}\left(\frac{x}{\hbar},\frac{1}{2}\right)\\&=n!\left(\frac{-\hbar}{2}\right)^n\left[\left(\frac{1}{2}\right)_n\right]^{-1}(-i)^n h_n^{(0,0)}\left(i\frac{x}{2\hbar}-\frac{1}{4},\frac{1}{2}\right).\end{aligned} \tag{44}$$

As a last remark in this section we note that for particular values, or ranges of the variables, these polynomials can also be identified with particular cases of the Jacobi polynomials and the generalized spherical harmonics (the Bargmann functions).[3]

V. RODRIGUES FORMULA, GENERATING FUNCTION, AND EXACT ZERO-ENERGY SCATTERING STATES FOR A FAMILY OF POTENTIALS

For a given x, the polynomials solve the eigenvalue equation (13), which is a differential equation of the hypergeometric type. In order to uncover more properties of these polynomials we restrict the investigation to

$$s=iy, \quad \frac{x}{c}=-iv; \quad y,v\in\mathbf{R}. \tag{45}$$

Note that in that case the corresponding ordered products are Hermitian. In terms of the new variables, Eq. (13) is as follows

$$\{(1+y^2)\partial_y^2+[v-(2n-1)y]\partial_y+n^2\}P_n(y,v)=0. \tag{46}$$

With the help of the function

$$\rho(y,v)=\frac{e^{v\tan^{-1}y}}{(1+y^2)^{(2n+1)/2}}, \tag{47}$$

which is determined from

$$\partial_y[(1+y^2)\rho]=[v-(2n-1)y]\rho, \tag{48}$$

Eq. (46) can be written in the self-adjoint form

$$\{\partial_y[(1+y^2)\rho(y,v)\partial_y]+\rho(y,v)n^2\}P_n(y,v)=0. \tag{49}$$

Thus, we have found the following Rodrigues formula, integral representation, and the generating function for these polynomials:

$$P_n(y,v)=\frac{(-ic)^n}{\rho(y,v)}\partial_y^n\left[\frac{e^{v\tan^{-1}y}}{(1+y^2)^{(1/2)}}\right], \tag{50}$$

$$P_n(y,v)=\frac{(-ic)^n n!}{2\pi i\rho(y,v)}\oint_C\frac{(1+z^2)^n}{(z-y)^{n+1}}\rho(z,v)dz, \tag{51}$$

$$\Phi(y,v,u)=(1-4uy-4u^2)^{-1/2}\frac{\rho(\xi,v)}{\rho(y,v)}, \tag{52}$$

$$(\xi=(2u)^{-1}[1-(1-4uy-4u^2)^{1/2}]).$$

In Eq. (51) C is a closed contour surrounding the point $z=y$, and the expansion of the generating function $\Phi(y,x,u)$ in power of u has, for sufficiently small $|u|$, the form

$$\Phi(y,v,u)=\sum_{n=0}^{\infty}\frac{P_n(y,v)}{n!}(-u/ic)^n. \tag{53}$$

Finally in this section, we would like to interpret these polynomials as the solutions of the Schrödinger type equations. By transforming the dependent variable as

$$\Psi_n(y,v)=[(1+y^2)\rho(y,v)]^{1/2}P_n(y,v), \tag{54}$$

the first derivative term in Eq. (46) vanishes and the equation

$$\partial_y^2\Psi_n(y,v)+\frac{y^2+2v(2n+1)y+(2n+1)^2-(v^2+3)}{4(1+y^2)^2}\Psi_n(y,v)=0 \tag{55}$$

results. η being a constant of dimension $(length)^{-1}$, Eq. (55) is the time independent Schrödinger equation for the potential

$$\frac{2m}{\eta^2\hbar^2}V_n(y,v)=-\frac{y^2+2v(2n+1)y+(2n+1)^2-(v^2+3)}{4(1+y^2)^2}. \tag{56}$$

In fact, here we have a family of potentials which are labelled by n and v, and are similar to the Pösch–Teller type ($\propto-\cosh^{-2}y$) potential holes. Thus, the wave functions given by (54) are the exact zero-energy scattering states for this family of potentials. Note that for $v=0$, $V(y,0)$ is an even function, so admit the solutions $\Psi_n(y,0)$ with the same parity as n.

VI. CONCLUSION

The main points of this study can be summarized as follows. (i) By using the explicit expression for the ordered products, which form a basis for the universal enveloping algebra of the HW-algebra, we have shown that the infinite Cartan subalgebra is, in fact, the Abelian algebra of two-variable polynomials. The variables are the ordering parameter s and the c-number correspondence of the squeeze operator $\hat{x}$ which is an element of the symplectic algebra in two dimensions. We expect that these definite-parity polynomials will play an important role in two dimensional physics in which tremendous developments are taking place in the recent years (see, for instance Ref. 13, and references therein). (ii) The realization of the su(1,1) obtained in this report, is, since it contains a single variable, a difference analogue of generalized Gelfand–Dyson realization.[18] Furthermore, the analyses of section V show that this algebra play the role of ''potential algebra'' (i.e., an algebra whose generators connect the states of the same energy in different potential strengths[18]) for a class of potentials. Regarding the role of the su(1,1) algebra in the exactly solvable problems of quantum mechanics, we expect that with its realization in terms of the difference operators found in this study, or, with appropriate generalizations, it is also the underlying algebra of the exactly solvable difference equations. This point, of course, requires further studies to be done. In particular, an algebraic approach to the polynomials of a discrete variable is now under investigation. (iii) Another particularly intriguing point requiring further studies is that the basic algebraic building block underlying the difference equations seems to be an expansion of the HW-algebra, which is the basic algebraic structure of both classical and quantum mechanics. More concretely, the algebra:

$$[x,D_h]=-\Delta_h,\quad [x,\Delta_h]=\hbar^2 D_h,\quad [D_h,\Delta_h]=0,$$

satisfied by the fundamental difference operations $\{x,D_h,2\Delta_h f(x)=f(x+h)+f(x-h)\}$ is the two-dimensional Euclidean algebra $e(2)=\{L,N_1,N_2:[L,N_1]=N_2,[L,N_2]=-N_1,[N_1,N_2]=0\}$ subjected to singular transformation (with respect to N_2): $\{x=\hbar L,\Delta_h=\hbar N_1,D_h=N_2\}$. The $\hbar\to 0$ (İnönü–Wigner) contraction of this algebra is the HW-algebra generated by

$$x,\quad \partial_x=\lim_{\hbar\to 0}D_h,\quad I=\lim_{\hbar\to 0}\Delta_h.$$

Thus, we expect that it is the İnönü–Wigner and its inverse transformation which form a bridge between the exactly solvable differential and difference equations of mathematical physics. Other results in this direction and more detailed properties of these polynomial will be given elsewhere.

ACKNOWLEDGMENTS

I wish to thank T. Dereli, M. Önder, and C. Harabati for helpful discussions. This work was supported in part by the Scientific and Technical Research Council of Turkey (TÜBİTAK).

[1] A. F. Nikiforov and V. B. Uvarov, *Special Functions of Mathematical Physics*, Translated from Russian by R. P. Boas (Birkhauser, Boston, 1988); A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of Discrete Variable* (Springer, Berlin, 1991).
[2] R. Askey, J. Phys. A **18**, L1017 (1985).
[3] J. T. Broad, Phys. Rev. A **31**, 1494 (1985); A. O. Barut and R. Wilson, J. Math. Phys. **17**, 900 (1976).
[4] Yu. F. Smirnov, S. K. Suslov, and A. M. Shirokov, J. Phys. A **17**, 2157 (1984).

[5] N. M. Atakishiyev and S. K. Suslov, J. Phys. A **18**, 1583 (1985).
[6] R. Alvarez-Nodarse and F. Marcellan, J. Math. Anal. Appl. **194**, 250 (1995); R. Alvarez-Nodarse and Yu. F. Smirnov, J. Phys. A **29**, 1435 (1996).
[7] C. M. Bender, L. R. Mead, and S. S. Pinsky, Phys. Rev. Lett. **56**, 2445 (1986); C. M. Bender, L. R. Mead, and S. S. Pinsky, J. Math. Phys. **28**, 509 (1987).
[8] C. M. Bender and G. Dunne, J. Math. Phys. **29**, 1727 (1988).
[9] T. Dereli and A. Verçin, J. Math. Phys. **38**, 5515 (1997).
[10] A. Verçin, ''Metaplectic covariance of the Weyl–Wigner–Groenewold–Moyal quantization and beyond'' (submitted).
[11] K. E. Cahill and R. J. Glauber, Phys. Rev. **177**, 1857,1882 (1969); N. L. Balazs and B. K. Jennings, Phys. Rep. **104**, 347 (1984); H. W. Lee, Phys. Rep. **259**, 147 (1994).
[12] C. N. Pope, X. Shen, and L. J. Romans, Nucl. Phys. B **339**, 191 (1990); P. Bouwknegt and K. Schoutens, Phys. Rep. **223**, 184 (1993).
[13] A. Cappelli, C. A. Trungenberger, and G. R. Zemba, Nucl. Phys. B **448**, 470 (1995); D. Karabali, Nucl. Phys. B **428**, 531 (1994).
[14] G. Dunne, J. Phys. A **21**, 2321 (1988); C. M. Bender and G. Dunne, Phys. Rev. D **40**, 3504 (1989); I. M. Gelfand and D. B. Fairlie, Commun. Math. Phys. **136**, 487 (1991).
[15] H. B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[16] For $s=0$ and $\hbar=1$ these polynomials coincide, up to factors of the form $n!/2^n$, with that $S_n(x)$ investigated in Refs. 7 and 8.
[17] For $\hbar=1$, i.e., $c^2=-1/4$ and $s=0$ the last line of Eqs. (19) is the lattice analog of the differential equation satisfied by the Hermite polynomials $He_n''(\xi)-2\xi He_n'(\xi)+2nHe_n(\xi)=0$ provided that $\xi=\sqrt{2}x$. This point was first observed in Ref. 8. Here we note that, for $s=0$ or $s\neq 0$ we have two additional difference recursion relations which, as we show in the main text, reveal important properties of these polynomials.
[18] Y. Alhassid, F. Gürsey, and F. Iachello, Ann. Phys. (N.Y.) **148**, 346 (1983); **167**, 181 (1986); A. Frank and K. B. Wolf, Phys. Rev. Lett. **53**, 1737 (1984); S. Kais and R. D. Levine, Phys. Rev. A **34**, 4615 (1986); J. Wu, F. Iachello, and Y. Alhassid, Ann. Phys. (N.Y.) **173**, 68 (1987); R. A. Ionescu, A. Ludu, and W. Scheid, J. Phys. A **29**, 3669 (1996).

Evaluation of nonrelativistic one-loop Feynman integrals by distributional methods

Peter Wagner
Institut für Mathematik und Geometrie, Universität Innsbruck, Technikerstr. 13, A-6020 Innsbruck, Austria

(Received 20 November 1996; accepted for publication 8 December 1997)

In this paper, we use Feynman's parameter formula and distribution theory in order to represent the integrals $\int_{\mathbb{R}^{n+1}}\Pi_{k=1}^{l}(it-(1)/(2m)|x-y_k|^2+\mu)^{-1}dt\,dx$ for $m>0$, $\mu>0$, $y_k\in\mathbb{R}^n$, by definite integrals over simplices. We obtain explicit expressions in the cases of $n=2$ and $n=l=3$, respectively. © *1998 American Institute of Physics.* [S0022-2488(98)03604-4]

I. INTRODUCTION AND NOTATIONS

As has been said above, we consider the integral

$$J:=\int_{\mathbb{R}^{n+1}}\prod_{k=1}^{l}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}dt\,dx \tag{1}$$

for $m>0$, $\mu>0$, $y_k\in\mathbb{R}^n$. In order to guarantee the absolute convergence of J, we suppose that $l>1+(n/2)$ and that the balls

$$B_k:=\{x\in\mathbb{R}^n:|x-y_k|^2<2m\mu\},\quad k=1,...,l,$$

are in ''general position.''

In Ref. 1, Feldman, Knörrer, Sinclair, and Trubowitz have evaluated (1) for $n=2$ by methods of algebraic geometry. In particular, for $l=3$, it was found that J vanishes whenever the disks B_1,B_2,B_3 have a point in common. This observation was made earlier by Sinclair on numerical grounds.[2]

The purpose of this paper is to apply Feynman's parameter integration method to (1) and to derive thereby a representation of (1) by integrals over simplices. (A similar treatment has been given earlier in the relativistic case.[3,4]) To avoid the divergencies introduced by the parameter formula, we use distribution theory and analytic continuation with respect to the powers of the factors in (1). The resulting representation shows that J vanishes if n is an even integer and $l=n+1$, whenever the balls $B_1,\ldots,B_l$ have nonempty intersection. Finally, we state explicit formulae for J in the cases of $n=2$ and of $n=l=3$, respectively.

Let us introduce some notations. We consider $\mathbb{R}^n$ as Euclidean space with the inner product $x\cdot y:=x_1y_1+\cdots+x_ny_n$ and write $|x|:=\sqrt{x\cdot x}$. In order to display the variable(s) referred to, we sometimes use notation as $\mathbb{R}^n_x$ or $\mathbb{R}^{n+1}_{t,x}$. We denote by Σ_{n-1} the $(n-1)$-dimensional standard simplex in $\mathbb{R}^n$, i.e.,

$$\Sigma_{n-1}:=\left\{(x_1,\ldots,x_n)\in\mathbb{R}^n:x_1\geqslant 0,...,x_n\geqslant 0,\sum_{k=1}^{n}x_k=1\right\}$$

and by $d\sigma(x)$ the measure $dx_1\cdots dx_{n-1}$ on Σ_{n-1}. When we make use of the theory of distributions, we adopt the notations from Refs. 5, 6. In particular, the Heaviside function is abbreviated by Y, $\mathscr{D}'_{L^1}$ denotes the space of integrable distributions, and $\langle\phi,T\rangle$ stands for the value of the distribution T on the test function ϕ. What concerns distributions depending analytically on parameters, we refer to Refs. 7, 8. The symbol arccot stands for the principal value of this function, i.e., that which has its range in the open interval $(0,\pi)$.

 © 1998 American Institute of Physics

II. ALGEBRAIC REDUCTION

One-loop integrals of the type (1) with more than $n+1$ factors (i.e., $l>n+1$) can algebraically be reduced to such with exactly $n+1$ factors. This goes back to Melrose[9] and Petersson.[10] We give a short deduction accomodated to our needs.

Theorem 1: Let $m,\mu>0$, $l,n\in\mathbb{N}$, $l>n+1$, *and* $y_1,\ldots,y_l\in\mathbb{R}^n$ *be such that, for each* $j=(j_1,\ldots,j_{n+1})\in\mathbb{N}^{n+1}$ *with* $1\leqslant j_1<\cdots<j_{n+1}\leqslant l$, *the vectors* M_j *and the numbers* ρ_j *determined by the equations*

$$|y_{j_r}-M_j|=\rho_j,\quad r=1,\ldots,n+1,$$

are well-defined and unique. (Hence M_j *is the center and* ρ_j *is the radius of the circumscribing sphere through* $y_{j_1},\ldots,y_{j_{n+1}}$.*) Furthermore, suppose that the numbers*

$$c_j:=(2m)^{l-n-1}\prod_{\substack{k=1\\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l}(\rho_j^2-|y_k-M_j|^2)^{-1}$$

are well-defined for j *as above. Then, for* $0\neq t\in\mathbb{R}$ *and all* $x\in\mathbb{R}^n$, *it holds*

$$\prod_{k=1}^{l}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}=\sum_{\substack{j=(j_1,\ldots,j_{n+1})\\ 1\leqslant j_1<\cdots<j_{n+1}\leqslant l}}c_j\prod_{r=1}^{n+1}\left(it-\frac{1}{2m}|x-y_{j_r}|^2+\mu\right)^{-1}.$$

Proof: Let us repeat first the following many-dimensional version of Lagrange's interpolation formula (for a proof see Lemma 1, p. 86, in Ref. 3):
Suppose that $1\leqslant\nu<l$ *and* $a_1,\ldots,a_l\in\mathbb{C}^\nu$ *are such that* $a_{j_1},\ldots,a_{j_\nu}$, *as well as* $a_{j_1}-a_k,\ldots,a_{j_\nu}-a_k$ *are linearly independent for pairwise different indices* $j_1,\ldots,j_\nu,k\in\{1,\ldots,l\}$. For $j=(j_1,\ldots,j_\nu)$, $1\leqslant j_1<\cdots<j_\nu\leqslant l$, *determine* $z(j)\in\mathbb{C}^\nu$ *by the system of linear equations* $a_{j_r}\cdot z(j)=1$, $r=1,\ldots,\nu$. *Then, for each complex homogeneous polynomial* $P(z_0,z)$ *of degree* $l-\nu$ *in the* $\nu+1$ *variables* z_0, $z=(z_1,\ldots,z_\nu)$, *we have*

$$P(z_0,z)=\sum_{\substack{j=(j_1,\ldots,j_\nu)\\ 1\leqslant j_1<\cdots<j_\nu\leqslant l}}P(1,z(j))\prod_{\substack{k=1\\ k\notin\{j_1,\ldots,j_\nu\}}}^{l}\frac{(z_0-a_k\cdot z)\det(a_{j_1},\ldots,a_{j_\nu})}{\det\begin{pmatrix}1&1&\cdots&1\\ a_k&a_{j_1}&\cdots&a_{j_\nu}\end{pmatrix}}.\tag{2}$$

We apply formula (2) to the case of $\nu=n+1$, $P=z_1^{l-n-1}$, and

$$a_k=\begin{pmatrix}|y_k|^2/2m-it-\mu\\ (x-2y_k)/\sqrt{2m}\end{pmatrix},\quad k=1,\ldots,l.$$

Then, from $z(j)\cdot a_{j_r}=1$, $r=1,\ldots,n+1$, we deduce, by Cramer's rule, that

$$c_j := P(1,z(j)) \prod_{\substack{k=1 \\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l} \frac{\det(a_{j_1},\ldots,a_{j_{n+1}})}{\det\begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_k & a_{j_1} & \cdots & a_{j_{n+1}} \end{pmatrix}}$$

$$= \prod_{\substack{k=1 \\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l} \frac{z(j)_1 \det(a_{j_1},\ldots,a_{j_{n+1}})}{\det\begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_k & a_{j_1} & \cdots & a_{j_{n+1}} \end{pmatrix}}$$

$$= \prod_{\substack{k=1 \\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l} \frac{\det\begin{pmatrix} 1 & \cdots & 1 \\ (x-2y_{j_1})/\sqrt{2m} & \cdots & (x-2y_{j_{n+1}})/\sqrt{2m} \end{pmatrix}}{\det\begin{pmatrix} 1 & 1 & \cdots & 1 \\ |y_k|^2/2m-it-\mu & |y_{j_1}|^2/2m-it-\mu & \cdots & |y_{j_{n+1}}|^2/2m-it-\mu \\ (x-2y_k)/\sqrt{2m} & (x-2y_{j_1})/\sqrt{2m} & \cdots & (x-2y_{j_{n+1}})/\sqrt{2m} \end{pmatrix}}$$

$$= \prod_{\substack{k=1 \\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l} \frac{2m \det(y_{j_2}-y_{j_1},\ldots,y_{j_{n+1}}-y_{j_1})}{\det\begin{pmatrix} |y_{j_1}|^2-|y_k|^2 & \cdots & |y_{j_{n+1}}|^2-|y_k|^2 \\ y_{j_1}-y_k & \cdots & y_{j_{n+1}}-y_k \end{pmatrix}}.$$

Because of

$$|y_{j_r}|^2-|y_k|^2-2M_j\cdot(y_{j_r}-y_k)=|y_{j_r}-M_j|^2-|y_k-M_j|^2=\rho_j^2-|y_k-M_j|^2,$$

we have

$$\det\begin{pmatrix} |y_{j_1}|^2-|y_k|^2 & \cdots & |y_{j_{n+1}}|^2-|y_k|^2 \\ y_{j_1}-y_k & \cdots & y_{j_{n+1}}-y_k \end{pmatrix}=(\rho_j^2-|y_k-M_j|^2)\det(y_{j_2}-y_{j_1},\ldots,y_{j_{n+1}}-y_{j_1}),$$

and hence

$$c_j=(2m)^{l-n-1}\prod_{\substack{k=1 \\ k\notin\{j_1,\ldots,j_{n+1}\}}}^{l}(\rho_j^2-|y_k-M_j|^2)^{-1}.$$

Therefore, Theorem 1 follows from formula (2) upon setting $z_0=0$, $z=\binom{1}{x/\sqrt{2m}}$ which implies

$$z_0-a_k\cdot z=it-\frac{1}{2m}|x-y_k|^2+\mu.$$

□

III. REPRESENTATION OF *J* BY SIMPLEX INTEGRALS

Let $\ln w$ be defined in the usual way on the slit plane $\{w\in\mathbb{C}:\operatorname{Im} w\neq 0 \text{ or } \operatorname{Re} w>0\}$. We fix $a>0$. For $z\in\mathbb{C}$ with $\operatorname{Re} z>-2$, the almost everywhere defined function

$$T_z := (it+|x|^2-a)^z = e^{z\,\ln(it+|x|^2-a)}$$

is locally integrable on $\mathbb{R}^{n+1}_{t,x}$ and of no more than polynomial growth. Furthermore,

$$T:\{z\in\mathbb{C}:\operatorname{Re} z>-2\}\to\mathscr{S}'(\mathbb{R}^{n+1}_{t,x}):z\mapsto T_z$$

is a holomorphic distribution-valued application, as one readily sees by differentiation with respect to z.

Theorem 2: *T can be extended as a holomorphic function to the whole complex plane. T_z is an integrable distribution iff* $\operatorname{Re} z<-1-(n/2)$. *In this case,*

$$\langle 1,T_z\rangle = \frac{2\pi^{n/2}\Gamma(z+1)\sin(z\pi)}{\Gamma(z+2+n/2)}\, a^{z+1+n/2}.$$

(*When the argument of either of the gamma functions vanishes or is a negative integer, then the right-hand side expression must be interpreted as a limit with respect to* z.)

Proof: (a) First notice that T_z is discontinuous at the ball $\{(0,x):|x|\leqslant a\}$ only. This implies, for Re $z>1$,

$$\frac{\partial T_z}{\partial t} = izT_{z-1} + 2i\ \sin(z\pi)\,\delta_t \otimes Y(a-|x|^2)(a-|x|^2)^z.$$

By analytic continuation, the last equation also holds for Re $z>-1$, and it allows to recursively extend T to the whole complex plane, since the same is possible for $\sin(z\pi)Y(a-|x|^2)(a-|x|^2)^z$ (comp. Ref. 7, Ch. I, 4.2, and Ch. III).

(b) On the open set $\{(t,x)\in\mathbb{R}^{n+1}: t^2+|x|^2>2a\}$, the distribution T_z is an absolutely integrable function iff Re $z<-1-n/2$. Hence $T_z\in\mathscr{D}'_{L^1}(\mathbb{R}^{n+1}_{t,x})$ iff Re $z<-1-n/2$. If, furthermore, Re $z<-2-n/2$, then $\langle 1,\partial T_{z+1}/\partial t\rangle = 0$, and hence

$$\langle 1,T_z\rangle = \frac{2}{z+1}\,\langle 1,\sin(z\pi)Y(a-|x|^2)(a-|x|^2)^{z+1}\rangle.$$

For Re $z>-2$, $U_z := \sin(z\pi)Y(a-|x|^2)(a-|x|^2)^{z+1}$ belongs to $L^1(\mathbb{R}^n_x)$, and its holomorphic extension lies in $\mathscr{D}'_{L^1}(\mathbb{R}^n)$ for all $z\in\mathbb{C}$, since it has compact support. For Re $z>-2$, we have

$$\begin{aligned}\langle 1,U_z\rangle = \sin(z\pi)\int_{|x|\leqslant\sqrt{a}}(a-|x|^2)^{z+1}\,dx &= \frac{2\pi^{n/2}\,\sin(z\pi)}{\Gamma(n/2)}\int_0^{\sqrt{a}}(a-r^2)^{z+1}r^{n-1}\,dr\\ &= \frac{\pi^{n/2}\Gamma(z+2)\sin(z\pi)}{\Gamma(z+2+n/2)}\,a^{z+1+n/2}\end{aligned}$$

(cf. Ref. 11, formula 431.2). Since the last expression for $\langle 1,U_z\rangle$ can analytically be continued to the complex plane, this completes the proof. □

Remark: Let us notice that for *negative* a and all complex z, the function $(it+|x|^2-a)^z$ is $\mathscr{C}^\infty$ on $\mathbb{R}^{n+1}$. For Re $z<-1-n/2$, this function is integrable and its integral vanishes, as readily follows from the equation

$$\frac{\partial}{\partial t}\,(it+|x|^2-a)^z = iz(it+|x|^2-a)^{z-1}$$

by partial integration with respect to t.

If we try to apply Feynman's formula

$$\prod_{k=1}^{l} w_k^{-1} = (l-1)!\int_{\Sigma_{l-1}}\frac{d\sigma(\lambda)}{(\lambda_1 w_1+\cdots+\lambda_l w_l)^l}, \tag{3}$$

valid for Re $w_1>0,\ldots,$Re $w_l>0$, (comp. Ref. 12, Th. 7-1, p. 56; Ref. 13, p. 72) in the integral (1) defining J, we obtain divergent integrals. We therefore use analytic continuation and consider, instead of the exponent -1, a complex number z. In this case, (3) takes the form

$$\prod_{k=1}^{l} w_k^{z} = \frac{\Gamma(-lz)}{\Gamma(-z)^l}\int_{\Sigma_{l-1}}\frac{(\lambda_1\cdots\lambda_l)^{-z-1}d\sigma(\lambda)}{(\lambda_1 w_1+\cdots+\lambda_l w_l)^{-lz}}. \tag{4}$$

Furthermore, we concentrate on the case of $l=n+1$, to which case the general case of $l\geqslant n+1$ can be reduced by Theorem 1.

Theorem 3: *Let* $m,\mu>0$, *and* $y_1,\ldots,y_{n+1}\in\mathbb{R}^n$ *be in general position (i.e.,* $y_2-y_1,\ldots,y_{n+1}-y_1$ *are linearly independent) and such that the balls*

$$B_k=\{x\in\mathbb{R}^n:|x-y_k|^2<2m\mu\},\quad k=1,...,n+1,$$

are in general position (i.e., no two of its boundaries touch, etc.). If Δ denotes the simplex spanned by $y_1,\ldots,y_{n+1}$, V denotes its volume, M and ρ are the center and the radius of the circumscribing sphere through $y_1,\ldots,y_{n+1}$, respectively, then

$$J:=\int_{\mathbb{R}^{n+1}}\prod_{k=1}^{n+1}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}dt\,dx$$

$$=\frac{(-1)^n 2\pi^{1+n/2}}{n!\Gamma(1+n/2)V}\left(\frac{\partial}{\partial\mu}\right)^n\int_\Delta(|x-M|^2-\rho^2+2m\mu)_+^{n/2}\,dx,$$

where, for

$$s\in\mathbb{R},\quad s_+:=\begin{cases}s: & s\geqslant 0,\\ 0: & s\leqslant 0.\end{cases}$$

Proof: Let $\tilde{y}_k:=y_k/\sqrt{2m\mu}$. Detailed inspection shows that the integral

$$I(z):=\int_{\mathbb{R}^{n+1}}\prod_{k=1}^{n+1}(it+|x-\tilde{y}_k|^2-1)^z\,dt\,dx$$

is absolutely convergent iff $-1-1/n<\mathrm{Re}\,z<-1+(n)/(2(n+1))$. For complex z in a neighbourhood of -1, we then obtain from the Feynman formula (4)

$$I(z)=\left\langle 1_{\mathbb{R}^{n+1}}(t,x),\prod_{k=1}^{n+1}(it+|x-\tilde{y}_k|^2-1)^z\right\rangle$$

$$=\frac{\Gamma(-(n+1)z)}{\Gamma(-z)^{n+1}}\left\langle 1_{\mathbb{R}^{n+1}}(t,x),\ \int_{\Sigma_n}\frac{(\lambda_1\cdots\lambda_{n+1})^{-z-1}d\sigma(\lambda)}{(it-1+\Sigma_{k=1}^{n+1}\lambda_k|x-\tilde{y}_k|^2)^{-(n+1)z}}\right\rangle.$$

We interchange the simplex integral with the evaluation of the integrable distribution in t and x in the integrand. (This process can be justified by considering the integrand as a distribution in t, x, and λ.) Since, for $\lambda\in\Sigma_n$,

$$it-1+\sum_{k=1}^{n+1}\lambda_k|x-\tilde{y}_k|^2=it+\left|x-\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-a$$

with

$$a=1+\left|\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|\tilde{y}_k|^2,$$

Theorem 2 yields

$$I(z)=\frac{2\pi^{n/2}\Gamma(-(n+1)z)\Gamma((n+1)z+1)\sin((n+1)z\pi)}{\Gamma(-z)^{n+1}\Gamma((n+1)z+2+n/2)}$$

$$\times\left\langle 1_{\Sigma_n}(\lambda),(\lambda_1\cdots\lambda_{n+1})^{-z-1}\left(1+\left|\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|\tilde{y}_k|^2\right)_+^{(n+1)z+1+n/2}\right\rangle.\qquad(5)$$

The distribution

$$V_z:=\frac{s_+^z}{\Gamma(z+1)}\in\mathscr{S}'(\mathbb{R}^1_s),\quad \mathrm{Re}\,z>-1,$$

is holomorphically extendable to the whole complex plane (cf. Ref. 7, Ch. I, 3.5). In Eq. (5), this distribution is composed with the $\mathscr{C}^\infty$ function

$$f(\lambda)=1+\left|\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|\tilde{y}_k|^2$$

(cf. Ref. 14, 7.2, p. 81) and then evaluated on Σ_n. Hence we can fix now $z=-1$ and, using

$$\lim_{z\to-1}\Gamma((n+1)z+1)\sin((n+1)z\pi)=-\frac{\pi}{n!},$$

we obtain

$$I(-1)=-2\pi^{1+n/2}\left\langle 1_{\Sigma_n}(\lambda),V_{-n/2}\left(1+\left|\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|\tilde{y}_k|^2\right)\right\rangle.$$

Since $V_{-n/2}$ is the n-th derivative of the ordinary function $s_+^{n/2}/\Gamma(1+n/2)$, we infer, with the substitutions $x=\sqrt{2m\mu}\tilde{x}$, $t=-\mu\tilde{t}$, the following:

$$\begin{aligned}J&=\int_{\mathbb{R}^{n+1}}\prod_{k=1}^{n+1}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}dt\,dx\\&=(-1)^{n+1}\left(\frac{2m}{\mu}\right)^{n/2}I(-1)\\&=\frac{(-1)^n2\pi^{1+n/2}}{\Gamma(1+(n/2))}\left(\frac{2m}{\mu}\right)^{n/2}\left\langle 1_{\Sigma_n}(\lambda),(s_+^{n/2})^{(n)}\left(1+\left|\sum_{k=1}^{n+1}\lambda_k\tilde{y}_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|\tilde{y}_k|^2\right)\right\rangle\\&=\frac{(-1)^n2\pi^{1+n/2}}{\Gamma(1+(n/2))}\left(\frac{\partial}{\partial\mu}\right)^n\int_{\Sigma_n}\left(2m\mu+\left|\sum_{k=1}^{n+1}\lambda_k y_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|y_k|^2\right)_+^{n/2}d\sigma(\lambda).\end{aligned}$$

Finally, we apply the affine transformation

$$\Sigma_n\to\Delta:\lambda\mapsto\sum_{k=1}^{n+1}\lambda_k y_k,$$

which has the Jacobian $n!V$. The quadratic form

$$\left|\sum_{k=1}^{n+1}\lambda_k y_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|y_k|^2$$

vanishes on the corners of Σ_n and hence transforms to $|x-M|^2-\rho^2$. This gives

$$J=\frac{(-1)^n2\pi^{1+n/2}}{n!\Gamma(1+(n/2))V}\left(\frac{\partial}{\partial\mu}\right)^n\int_\Delta(|x-M|^2-\rho^2+2m\mu)_+^{n/2}dx.$$

and completes the proof. □

Remark: For even n, the distribution

$$\left(\frac{\partial}{\partial\mu}\right)^n(|x-M|^2-\rho^2+2m\mu)_+^{n/2}$$

has its support on the set

$$S:=\{x\in\mathbb{R}^n:|x-M|^2=\rho^2-2m\mu\},$$

which is either the empty set $\emptyset$, or a sphere. If the balls

$$B_k:=\{x\in\mathbb{R}^n:|x-y_k|^2<2m\mu\},\quad k=1,\ldots,n+1,$$

have a point, say x_0, in common, then, for $\lambda \in \Sigma_n$ and $x=\Sigma_{k=1}^{n+1}\lambda_k y_k \in \Delta$,

$$|x-M|^2-\rho^2=\left|\sum_{k=1}^{n+1}\lambda_k y_k\right|^2-\sum_{k=1}^{n+1}\lambda_k|y_k|^2=\left|\sum_{k=1}^{n+1}\lambda_k(y_k-x_0)\right|^2-\sum_{k=1}^{n+1}\lambda_k|y_k-x_0|^2>-2m\mu,$$

and hence $S\cap\Delta=\emptyset$ and consequently $J=0$.

IV. SPECIAL CASES

Let us consider first the case of two space dimensions. By Theorem 1, it is sufficient to consider three factors, i.e.,

$$J=\int_{\mathbb{R}^3}\prod_{k=1}^{3}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}dt\,dx.$$

As above, let Δ be the triangle spanned by y_1,y_2,y_3 in $\mathbb{R}^2$, let V denote its area, M and ρ be the center and the radius of the circumscribing circle through y_1,y_2,y_3, respectively.

Since

$$\left(\frac{\partial}{\partial\mu}\right)^2(|x-M|^2-\rho^2+2m\mu)_+=4m^2\delta(|x-M|^2-\rho^2+2m\mu),$$

three cases arise: If $\rho<\sqrt{2m\mu}$, then $J=0$. Similarly, if $\rho>\sqrt{2m\mu}$ and the circle S with center M and radius $\sqrt{\rho^2-2m\mu}$, does not meet Δ, then $J=0$. (These two cases just correspond to the intersection of B_1,B_2,B_3 being non-empty; compare the Remark above.) Finally, if $\rho>\sqrt{2m\mu}$ and the curve $\emptyset\neq S\cap\Delta$ has length L, then, by Theorem 3,

$$J=\frac{2\pi^2m^2L}{V\sqrt{\rho^2-2m\mu}}.$$

(Note that the total measure of $\delta(|x-M|^2-\rho^2+2m\mu)$ is π.)

By expressing L through the lengths and angles in the triangle Δ, we arrive at a more explicit if less visual representation of J.

Theorem 4: *Let $m,\mu>0$, $l\in\mathbb{N}$, $l\geqslant 3$, and $y_1,\ldots,y_l\in\mathbb{R}^2$ be such that no three of them lie on a line and no four of them lie on a circle. For the disks*

$$B_k=\{x\in\mathbb{R}^2:|x-y_k|^2<2m\mu\},\quad k=1,\ldots,l,$$

suppose further that no two of the boundaries ∂B_k, $k=1,\ldots,l$, touch and no three of these boundaries have a point in common.

For $j=(j_1,j_2,j_3)\in\mathbb{N}^3$ with $1\leqslant j_1<j_2<j_3\leqslant l$, let Δ_j denote the triangle with corners y_{j_1},y_{j_2},y_{j_3}, let V_j denote its area, and let M_j and ρ_j be the center and the radius of the circumscribing circle through y_{j_1},y_{j_2},y_{j_3}, respectively. As in Theorem 1 define

$$c_j:=(2m)^{l-3}\prod_{\substack{k=1\\ k\notin\{j_1,j_2,j_3\}}}^{l}(\rho_j^2-|y_k-M_j|^2)^{-1}.$$

Let ϵ_j be 0 *if the disks B_{j_1},B_{j_2},B_{j_3} meet, and 1 else. Finally, denote by*

$$a_{j,r}:=|y_{j_s}-y_{j_t}|,\quad \alpha_{j,r}:=\angle(y_{j_s}-y_{j_r},y_{j_t}-y_{j_r}),\quad \{r,s,t\}=\{1,2,3\},$$

the lengths and angles in the triangle Δ_j. Then

$$J:=\int_{\mathbb{R}^3}\prod_{k=1}^{l}\left(it-\frac{1}{2m}\,|x-y_k|^2+\mu\right)^{-1}dt\;dx$$

$$=4m^2\pi^2\sum_{\substack{j=(j_1,j_2,j_3)\\1\leqslant j_1<j_2<j_3\leqslant l}}c_j\,\frac{\epsilon_j}{V_j}\left[\pi-\sum_{r=1}^{3}Y(a_{j,r}-2\sqrt{2m\mu})\operatorname{arccot}\left(\frac{2\rho_j\cos(\alpha_{j,r})}{\sqrt{a_{j,r}^2-8m\mu}}\right)\right].$$

Proof: As above, define

$$S_j:=\{x\in\mathbb{R}^2:|x-M_j|^2=\rho_j^2-2m\mu\}$$

and denote by L_j the length of $S_j\cap\Delta_j$. According to what has been said before Theorem 4 and using Theorem 1, it remains only to be shown that, in the case of $\epsilon_j=1$,

$$\frac{L_j}{\sqrt{\rho_j^2-2m\mu}}=2\pi-2\sum_{r=1}^{3}Y(a_{j,r}-2\sqrt{2m\mu})\operatorname{arccot}\left(\frac{2\rho_j\cos(\alpha_{j,r})}{\sqrt{a_{j,r}^2-8m\mu}}\right).$$

Note that S_j intersects the edge of Δ_j spanned by y_{j_s} and y_{j_t} iff $a_{j,r}>2\sqrt{2m\mu}$. In this case, the distance of the intersection points from M_j is $\sqrt{\rho_j^2-2m\mu}$ and these points lie $\sqrt{a_{j,r}^2-8m\mu}$ apart. The corresponding angle $\beta_{j,r}$ along which the circle S_j runs outside of Δ_j then either is $\gamma_{j,r}$ or $2\pi-\gamma_{j,r}$, according to whether the angle $\alpha_{j,r}$ is acute or obtuse, and where we have set

$$\gamma_{j,r}:=2\arcsin\left(\sqrt{\frac{a_{j,r}^2/4-2m\mu}{\rho_j^2-2m\mu}}\right).$$

An easy conversion yields

$$\beta_{j,r}=2\operatorname{arccot}\left(\frac{2\rho_j\cos(\alpha_{j,r})}{\sqrt{a_{j,r}^2-8m\mu}}\right).$$ □

Finally, let us consider the case of $n=l=3$. Though it is not contained in Theorem 3, a completely analogous calculation gives

$$J=\frac{4m^2\pi^2}{V}\int_{\Delta}(|x-M|^2-\rho^2+2m\mu)_+^{-1/2}d\sigma(x),\quad s_+:=\begin{cases}s: & s\geqslant 0,\\ 0: & s\leqslant 0,\end{cases}$$

where Δ denotes the triangle in $\mathbb{R}^3$ spanned by y_1,y_2,y_3, V its area, $d\sigma(x)$ its surface measure, and M and ρ the center and the radius of the circumscribing circle through y_1,y_2,y_3, respectively.

After shifting M to the origin, we can integrate radially. These radial integrals have the form

$$\int_{(\rho^2-2m\mu)_+^{1/2}}^{c/\cos\varphi}(r^2-\rho^2+2m\mu)^{-1/2}r\;dr$$

and consequently yield angular integrals of the form

$$\int\sqrt{\frac{c^2}{\cos^2\varphi}-\rho^2+2m\mu}\;d\varphi.$$

These latter integrals also give elementary transcendental functions. Several cases must be distinguished according to whether the triangle Δ is acute or obtuse, and whether $\rho<\sqrt{2m\mu}$. We merely state the final result.

Theorem 5: *Let $m,\mu>0$, and suppose $y_1,y_2,y_3\in\mathbb{R}^3$ are not collinear and are such that the boundaries of the balls*

$$B_k=\{x\in\mathbb{R}^3:|x-y_k|^2<2m\mu\},\quad k=1,2,3,$$

do not touch.

Let Δ be the triangle in $\mathbb{R}^3$ spanned by y_1,y_2,y_3, and denote by V its area, by $\alpha_1,\alpha_2,\alpha_3$ its angles, and by ρ the radius of its circumscribing circle, respectively. Then the integral defining J is absolutely convergent and it holds

$$\begin{aligned}J&:=\int_{\mathbb{R}^4}\prod_{k=1}^{3}\left(it-\frac{1}{2m}|x-y_k|^2+\mu\right)^{-1}dt\,dx\\&=-\frac{8m^2\pi^3}{V}(2m\mu-\rho^2)_+^{1/2}+\frac{4m^2\pi^2\rho}{V}\sum_{k=1}^{3}\cos\alpha_k\ln\left(\left|\frac{\sqrt{2m\mu}+\rho\sin\alpha_k}{\sqrt{2m\mu}-\rho\sin\alpha_k}\right|\right)\\&\quad+\frac{4m^2\pi^2\sqrt{|\rho^2-2m\mu|}}{V}\begin{cases}2\displaystyle\sum_{k=1}^{3}\operatorname{arccot}\left(\frac{\sqrt{2m\mu}\cot\alpha_k}{\sqrt{2m\mu-\rho^2}}\right): & \rho^2<2m\mu,\\ \displaystyle\sum_{k=1}^{3}\ln\left(\left|\frac{\sqrt{2m\mu}\cot\alpha_k-\sqrt{\rho^2-2m\mu}}{\sqrt{2m\mu}\cot\alpha_k+\sqrt{\rho^2-2m\mu}}\right|\right): & \rho^2\geqslant 2m\mu.\end{cases}\end{aligned}$$

ACKNOWLEDGMENTS

I should like to express my gratitude to Professor Horst Knörrer, who called my attention to the integrals treated in this paper, who invited me to Zürich to discuss matters, and who pointed out two errors in a preliminary version of this paper.

[1] J. Feldman, H. Knörrer, R. Sinclair, and E. Trubowitz, *Evaluation of Fermion Loops by Higher Residues*, Preprint, ETH Zürich 1996.
[2] R. Sinclair, ''The evaluation of Feynman graphs for a $(2+1)$-dimensional nonrelativistic electron gas,'' Comput. Phys. Commun. **81**, 173–184 (1994).
[3] N. Ortner and P. Wagner, ''On the evaluation of one-loop Feynman amplitudes in Euclidean quantum field theory,'' Ann. Inst. Henri Poincaré, Phys. Théor. **63**, 81–110 (1995).
[4] P. Wagner, ''A volume formula for asymptotic hyperbolic tetrahedra with an application to quantum field theory,'' Indag. Mathem. (N.S.) **7**, 527–547 (1996).
[5] L. Schwartz, *Théorie des Distributions*, Nouv. éd. (Hermann, Paris, 1966).
[6] J. Horváth, *Topological Vector Spaces and Distributions I* (Addison-Wesley, Reading, MA, 1966).
[7] I. M. Gel'fand and G. E. Shilov, *Generalized functions*, Vol. 1 (Academic, New York, 1968).
[8] J. Horváth, ''Distribuciones definidas por prolongación analítica,'' Rev. Colombiana Mat. **8**, 47–95 (1974).
[9] D. B. Melrose, ''Reduction of Feynman diagrams,'' Nuovo Cimento A **40**, 181–213 (1965).
[10] B. Petersson, ''Reduction of one-loop Feynman diagrams with n vertices in m-dimensional Lorentz space,'' J. Math. Phys. **6**, 1955–1959 (1965).
[11] W. Gröbner and N. Hofreiter, *Integraltafel, 2.Teil: Bestimmte Integrale*, 5. Aufl. (Springer, Wien, 1973).
[12] N. Nakanishi, *Graph Theory and Feynman Integrals* (Gordon and Breach, New York, 1971).
[13] L. Schwartz, *Méthodes Mathématiques pour les Sciences Physiques*, 2ème éd. (Hermann, Paris, 1965).
[14] F. G. Friedlander, *Introduction to the Theory of Distributions* (Cambridge Univ. Press, Cambridge, 1982).

Bloch and Wannier functions composed of Gaussian orbitals—analyticity, localization, and expectation values

P. Zeiner and R. Dirl
Institut für Theoretische Physik, and Center for Computational Materials Science, Technische Universität Wien A-1040 Wien, Wiedner Hauptstrasse 8-10, Austria

B. L. Davies
School of Mathematics, University of Wales, Bangor, Gwynedd LL57 1UT, Wales, United Kingdom

(Received 13 August 1997; accepted for publication 3 November 1997)

We discuss Bloch and Wannier functions composed of Gaussian orbitals. The fact that the Bloch functions may be expressed in terms of θ-functions is used to discuss them in detail. We study not only their analyticity properties, but also features of expectation values of some specifically chosen Hamiltonians. We also investigate the localization properties of various choices of Wannier functions and compare them with the original Gauss functions. Likewise, we discuss in a more general context the relationship between the analyticity properties of Bloch functions and the localization properties of corresponding Wannier functions. © *1998 American Institute of Physics.* [S0022-2488(98)00404-6]

I. INTRODUCTION

Localized but nonorthogonal functions, like Gaussian orbitals, are widely used in physical applications (see for instance Ref. 1). In the case of periodic crystals the most prominent and important functions for describing their electronic properties are Bloch[2,3] and Wannier[4] functions. Here we study in detail Bloch and Wannier functions which are composed of Gaussian orbitals.

It was G. H. Wannier[5] who first realized that these Bloch functions may be expressed in terms of θ-functions. By choosing the phase factors appropriately he was also able to give an explicit expression for the corresponding Wannier functions. These Wannier functions, however, are not symmetric due to his specific choice of the phase factors. We extend his work by considering also a symmetric choice of Wannier functions, which are shown to be the best localized Wannier functions composed of Gaussian orbitals and we also discuss in detail the totally different analyticity properties of the corresponding Bloch functions. In addition, we study expectation values of some specifically chosen Hamiltonians. We cannot only give explicit formulas for them but also prove several monotonicity properties of them. Last we discuss in general the relationship between the analyticity properties of the Bloch functions and a generalized exponential decrease of the Wannier functions.

The paper has the following structure: First we discuss the analyticity properties of the Bloch functions composed of Gaussian orbitals. Of course, the analyticity properties of Bloch functions are, in principle, well known for one dimensional crystals since this subject was extensively discussed in the famous paper of Kohn.[6] Also much is known about the analyticity properties in higher dimensions due to the classical papers of Cloizeaux[7] and the more recent paper of Nenciu.[8] Nevertheless it seems to be interesting to discuss the analytical properties for an explicit example, since it shows that the Riemann surface may be completely different for different choices of Bloch functions. Moreover recently the analyticity of Bloch functions played an important role in the theory of band representations.[9] We explicitly give the Riemann surfaces for two different choices of Bloch functions: For Wannier's choice[5] it is just the real plane with the exception of some points, in case of a symmetric choice of Bloch functions however, the Riemann surface consists of two sheets which are stuck together at an infinity of finite intervals. The second part deals with the localization properties of the corresponding Wannier functions. Apart from the exponential decrease of the Wannier functions implied by analyticity, we discuss the uncertainty Δx of the Wannier functions. We show that a certain symmetric choice of Wannier functions have the same

 © 1998 American Institute of Physics

Δx as the original Gauss functions and these Wannier functions are the best localized ones. In the third part we discuss expectation values for some Hamiltonians. Various monotonicity properties are shown, for instance we prove that the expectation value of the kinetic energy is a monotone function of the wave vector $k \in [0,\pi]$ and a monotone function of the localization parameter describing the original Gaussians. In the last section we consider the following generalization of exponential fall off: instead of requiring that the Wannier functions fall off exponentially $|W(x)| \leqslant e^{-\gamma|x|}B$, where γ and B are positive constants and $W(x)$ is the Wannier function, we require only that the integral

$$\int_{y-\epsilon}^{y+\epsilon} |W_0(x)|^2 dx \leqslant e^{-2\gamma|y|}B(\epsilon)$$

falls off exponentially. We then prove that the Wannier functions fall off exponentially in this generalized sense if and only if the corresponding Bloch functions are vector valued analytic functions.

II. BLOCH FUNCTIONS

For the sake of simplicity we confine the discussion to the one-dimensional case exclusively, its generalization to higher dimensions should be discussed elsewhere. Let $G^\beta: \mathbb{R} \to \mathbb{R}$

$$G^\beta(x) = \left(\frac{2\beta}{\pi}\right)^{1/4} e^{-\beta x^2} \tag{1}$$

be a one-dimensional Gaussian wave function normalized to unity with respect to the usual $L^2(\mathbb{R})$-norm. Here $\beta = 1/4\sigma^2$, where σ is the standard deviation of the Gaussian distribution $|G^\beta(x)|^2$, i.e., $\sigma^2 = \int_{-\infty}^{\infty} dx x^2 |G^\beta(x)|^2 = (G^\beta, x^2 G^\beta)_\infty$, where $(\cdot,\cdot)_\infty$ denotes the usual inner product of $L^2(\mathbb{R})$. The corresponding Bloch functions $\phi_k^\beta: \mathbb{R} \to \mathbb{C}$ read (lattice constant $a=1$)

$$\phi_k^\beta(x) = N^\beta(k) \sum_{n=-\infty}^{\infty} e^{ikn} G_n^\beta(x), \tag{2}$$

where $G_n^\beta(x) = G^\beta(x-n)$ are the shifted Gaussian orbitals. The normalization constants $N^\beta(k)$ are chosen real and positive, where the Bloch functions are normalized such that $(\phi_k^\beta, \phi_k^\beta)_1 = \int_0^1 dx |\phi_k^\beta(x)|^2 = 1/2\pi$, where the entry $(\cdot,\cdot)_1$ denotes the inner product on $L^2([0,1])$. By introducing the θ-function[10–12]

$$\theta_3(z|\tau) = \sum_{n=-\infty}^{\infty} e^{in^2\pi\tau} e^{i2nz}, \quad \text{Im } \tau > 0, \tag{3}$$

the Bloch function may be written in the following way:[5]

$$\phi_k^\beta(x) = N^\beta(k) \left(\frac{2\beta}{\pi}\right)^{1/4} e^{-\beta x^2} \theta_3\left(\frac{k}{2} - i\beta x \middle| \frac{i\beta}{\pi}\right), \tag{4}$$

where the normalization constant $N^\beta(k)$ is given by

$$N^\beta(k) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\theta_3\left(\frac{k}{2}\middle|\frac{i\beta}{2\pi}\right)}}. \tag{5}$$

It is easy to verify the following symmetry properties of the Bloch functions. The reality of the Gauss functions implies the time reversal symmetry

$$(\phi_k^\beta(x))^* = \phi_{-k}^\beta(x), \tag{6}$$

and the inversion symmetry of the Gauss function gives rise to the following transformation law with respect to inversion:

$$\phi_k^\beta(-x)=\phi_{-k}^\beta(x). \tag{7}$$

Note that these Bloch functions correspond to a band representation[13] of the only nontrivial one-dimensional space group characterized by the Wyckoff position $w=0$ and the trivial one-dimensional irreducible unitary representation of the corresponding site symmetry group, which is isomorphic to C_2.

The normalization constants $N^\beta(k)$ may be expressed as an infinite product[5,10–12]

$$N^\beta(k)=\frac{1}{\sqrt{2\pi}}\prod_{n=1}^{\infty}[(1-q^{2n})^{-1/2}(1+q^{2n-1}e^{ik})^{-1/2}(1+q^{2n-1}e^{-ik})^{-1/2}], \tag{8}$$

where $q:=e^{-\beta/2}$. This formula motivated Wannier to use alternative normalization constants[5]

$$\tilde{N}^\beta(k)=\frac{1}{\sqrt{2\pi}}\prod_{n=1}^{\infty}(1-q^{2n})^{-1/2}(1+q^{2n-1}e^{-ik})^{-1}. \tag{9}$$

Obviously the expression for the normalization constants $\tilde{N}^\beta(k)$ is much simpler than the formula for $N^\beta(k)$. In addition one gets an explicit expression for the Wannier functions if one chooses $\tilde{N}^\beta(k)$ instead of $N^\beta(k)$. The great advantage of $N^\beta(k)$, however, is that the corresponding Bloch functions have the symmetry properties Eqs. (6) and (7). The corresponding Bloch functions for $\tilde{N}^\beta(k)$ only fulfill Eq. (6), whereas Eq. (7) has to be replaced by

$$\phi_k^\beta(-x)=\frac{\tilde{N}^\beta(k)}{\tilde{N}^\beta(-k)}\,\phi_{-k}^\beta(x), \tag{10}$$

which implies that the corresponding Wannier functions are not symmetric. Note in passing that both normalization constants are periodic functions in k-space, namely they satisfy the condition $N^\beta(k)=N^\beta(k+2\pi N)$, and hence in both cases the Bloch functions have the same property: $\phi_k^\beta=\phi_{k+2\pi N}^\beta$, $N\in\mathbb{Z}$.

Since needed later on we summarize some further properties of the θ-functions. There exist four different types of θ-functions. We have already mentioned θ_3, hence we state only the three other ones here:

$$\theta_1(z|\tau)=-i\sum_{n=-\infty}^{\infty}(-1)^n e^{i(n+1/2)^2\pi\tau}e^{i(2n+1)z}, \tag{11}$$

$$\theta_2(z|\tau)=\sum_{n=-\infty}^{\infty}e^{i(n+1/2)^2\pi\tau}e^{i(2n+1)z}, \tag{12}$$

$$\theta_4(z|\tau)=\sum_{n=-\infty}^{\infty}(-1)^n e^{in^2\pi\tau}e^{i2nz}. \tag{13}$$

They are linearly independent but by an appropriate change of variable each one may be transformed into each other one. For these relations as well as periodicity and quasi-periodicity properties, product expansions and Jacobi's transformation, which will be used frequently in the following, we refer to the literature.[10–12]

III. ANALYTICITY PROPERTIES OF THE BLOCH FUNCTIONS

So far x and k have been real. But obviously both G^β and ϕ_k^β may be defined for complex x and complex $k=\kappa+i\rho$, too. Here $N(k)$ is defined by Eq. (8) for complex k, too. Note that with this definition the equation $\int_0^1 dx|\phi_k^\beta(x)|^2=1/2\pi$ holds true only for real k. To discuss the

dependence of the Bloch functions ϕ_k^β on k we introduce the function $\Phi^\beta: \mathsf{R}\times\mathbb{C}\to\mathbb{C}$, $\Phi^\beta(k,x) = \phi_k^\beta(x)$, where R is an appropriate subset of the complex plane or, more generally, an appropriate Riemann surface. We will use the term Bloch function for Φ^β, too, but to our mind this should not lead to any confusion. Similarly we define the function $\mathbf{N}^\beta: \mathsf{R}\to\mathbb{C}$, $\mathbf{N}^\beta(k) = N^\beta(k)$.

We want to discuss the analyticity properties of the Bloch functions now. In this context we have to refer to the famous articles of Kohn,[6] who considered the analyticity properties for the one-dimensional case, and to the work of Cloizeaux[7] and Nenciu,[8] who discussed the analyticity problem in higher dimensions.

We investigate the analyticity properties of the Bloch function Φ^β now. To this end note that θ_3 is an entire function of z, i.e., analytic for all $z\in\mathbb{C}$. Hence the Bloch function Φ^β is an analytic function of x in the whole complex plane. The analyticity of θ_3 implies also that the Bloch function Φ^β is an analytic function of k for all k for which the normalization constant $\mathbf{N}^\beta(k)$ is analytic. Obviously $\mathbf{N}^\beta(k)$ is analytic at a point k if and only if $\theta_3(k/2|i\beta/2\pi)$ is not zero. Since the zeros of this θ-function[10–12] are given by $k=(m+\frac{1}{2})2\pi+(n+\frac{1}{2})i\beta$, the normalization constant $\mathbf{N}^\beta$ and hence the Bloch function Φ^β are analytic functions of k in the strip

$$S^\beta = \left\{ k=\kappa+i\rho : |\rho| < \frac{\beta}{2} \right\}. \tag{14}$$

Note that the width of the strip is proportional to β, which is a measure of the localization of the Gauss functions. Thus if the Bloch functions are composed of strongly localized Gaussian orbitals the strip of analyticity is broad, whereas in case of weakly localized Gaussian orbitals the strip of analyticity is small. In the limit $\beta\to 0$ the domain of analyticity vanishes, and in fact, as we show in the next section, the Bloch function is not analytic in this limiting case, it is even discontinuous.

Note in passing that the domain of analyticity can be extended consistently, namely it can be enlarged to the whole plane without cuts ranging from $k=(m+\frac{1}{2})2\pi\pm i\beta/2$ to $(m+\frac{1}{2})2\pi\pm i\infty$, respectively, thus

$$D^\beta = \mathbb{C}\setminus\left\{ k:\mathrm{Re}\ k=(2m+1)\pi,\ \ |\mathrm{Im}\ k| \geqslant \frac{\beta}{2},\ \ m\in\mathbb{Z} \right\}. \tag{15}$$

In fact, we can show also more. If we go around a singularity on a closed path, then the sign of the normalization constant $\mathbf{N}^\beta$ has changed if we return to the starting point. On the other hand if the path encloses an even number of singularities, then we return to the starting point with the same phase. From this one infers easily that the domain of analyticity can be extended to

$$D_1^\beta = \mathbb{C}\setminus\{k:\mathrm{Re}\ k=(2m+1)\pi,\ \ (2n+\tfrac{1}{2})\beta \leqslant |\mathrm{Im}\ k| \leqslant (2n+\tfrac{3}{2})\beta,\ \ m\in\mathbb{Z},\ \ n\in\mathbb{N}_0\}. \tag{16}$$

The Riemann surface consists of the two sheets D_1^β, which are stuck together at the intervals $k:\mathrm{Re}\ k=(2m+1)\pi$, $(2n+\frac{1}{2})\beta<|\mathrm{Im}\ k|<(2n+\frac{3}{2})\beta$, $m\in\mathbb{Z}$, $n\in\mathbb{N}_0$. On one sheet the normalization constant $\mathbf{N}^\beta$ has the same values as before, and on the other sheet $\mathbf{N}^\beta$ is defined with the opposite sign.

Similarly one can derive the analyticity properties of $\widetilde{\mathbf{N}}^\beta(k)$. Here half of the zeros of the denominator are removed,[5] and hence $\widetilde{\mathbf{N}}^\beta(k)$ and the corresponding Bloch function are analytic in the half-plane

$$\widetilde{S}^\beta = \left\{ k=\kappa+i\rho : \rho < \frac{\beta}{2} \right\}. \tag{17}$$

Also in this case the domain of analyticity may be extended, namely,

$$\widetilde{D}^\beta = \mathbb{C}\setminus\left\{ k:\mathrm{Re}\ k=(2m+1)\pi,\ \ \mathrm{Im}\ k \geqslant \frac{\beta}{2},\ \ m\in\mathbb{Z} \right\}. \tag{18}$$

Since the singularities $k=(m+\frac{1}{2})2\pi+i\beta(n+\frac{1}{2})$, $m\in\mathbb{Z}$, $n\in\mathbb{N}_0$ are simple poles, we infer immediately that $\widetilde{\mathbf{N}}^\beta$ is analytic in

$$\widetilde{D}_1^\beta = \mathbb{C}\backslash\{k: \mathrm{Re}\ k = (2m+1)\pi,\ \ \mathrm{Im}\ k = \beta(n+1/2),\ \ m\in\mathbb{Z},\ \ n\in\mathbb{N}_0\}. \tag{19}$$

Note that this is also the Riemann surface.

From these equations we infer that $\widetilde{\mathbf{N}}^\beta$ is analytic in a larger part of the complex plane than $\mathbf{N}^\beta$, if we restrict our considerations to the complex plane. If we compare the Riemann surfaces, then $\widetilde{\mathbf{N}}^\beta$ has obviously the simpler Riemann surface. Hence from the viewpoint of analyticity, $\widetilde{\mathbf{N}}^\beta$ is the better choice of the normalization constant.

IV. WANNIER FUNCTIONS

We adopt the following conventions for the Wannier functions $W_n^\beta: \mathbb{R}\to\mathbb{R}$:

$$W_n^\beta(x) = \frac{1}{(2\pi)^{1/2}} \int_0^{2\pi} dk\ \phi_k^\beta(x) e^{-ikn}. \tag{20}$$

Note that the Wannier functions are real $W_n^\beta(x) = W_n^\beta(x)^*$ because of the symmetry $(\phi_k^\beta)^* = \phi_{-k}^\beta$ and the periodicity $\phi_k^\beta = \phi_{k+2\pi N}^\beta$ of the Bloch functions. Note that this is valid for both choices of the normalization constant. In general, however, the Wannier functions defined by $\widetilde{W}_n^\beta(x) = (1/(2\pi)^{1/2}) \int_{-\pi}^{\pi} dk e^{i\varphi(k)} \phi_k^\beta(x) e^{-ikn}$ with arbitrary real $\varphi(k)$ are complex functions. In fact, if and only if $\varphi(-k) = -\varphi(k)$, the Wannier functions $\widetilde{W}_n$ are real, too. More generally, we may define the Wannier functions by $\widetilde{W}_n^{\beta,a}(x) = [1/(2\pi)^{1/2}] \int_{-\pi+a}^{\pi+a} dk e^{i\varphi(k)} \phi_k^\beta(x) e^{-ikn}$. Unless $e^{i\varphi(k)} = e^{i\varphi(k+2\pi)}$ for all k, the Wannier functions $\widetilde{W}_n^{\beta,a}$ will be different for different values of $a\in\mathbb{R}$, respectively.

The analyticity of the Bloch function implies the exponential decay of the corresponding Wannier functions. From theorem 1 below, we infer

$$|W_0^\beta(x)| \leqslant e^{-\gamma|x|} B^\beta(\gamma) \tag{21}$$

which holds for all $\gamma < \beta/2$, where $B^\beta(\gamma)$ is an appropriate constant. In fact, for our special Wannier functions this result can be slightly improved, namely

$$|W_0^\beta(x)| \leqslant e^{-(\beta/2)|x|} B^\beta. \tag{22}$$

This is due to the fact that the poles of $(\mathbf{N}^\beta)^2$ are simple. An explicit formula for the Wannier functions cannot be given but there is an explicit formula for its Fourier transform:

$$W_0^\beta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} \frac{1}{\theta_3\left(\dfrac{i\pi k}{\beta}\middle|\ \dfrac{2\pi i}{\beta}\right)^{1/2}} dk = \frac{1}{2\pi} \left(\frac{2\pi}{\beta}\right)^{1/4} \int_{-\infty}^{\infty} e^{ikx} \frac{e^{-(1/4\beta)k^2}}{\theta_3\left(\dfrac{k}{2}\middle|\ \dfrac{i\beta}{2\pi}\right)^{1/2}} dk. \tag{23}$$

Now we turn to the Wannier functions $\widetilde{W}_0^\beta(x)$ corresponding to the normalization constant $\widetilde{\mathbf{N}}^\beta$. It is given explicitly[5] by

$$\widetilde{W}_0^\beta(x) = \left(\frac{2\beta}{\pi}\right)^{1/4} \prod_{n=1}^{\infty} (1-e^{-\beta n})^{-1/2} \left[e^{-\beta x^2} + \sum_{m=1}^{\infty} (-1)^m \frac{e^{-(\beta/2)m}}{\Pi_{p=1}^m (1-e^{-\beta p})} e^{-\beta(x-m)^2} \right]. \tag{24}$$

These functions are discussed exhaustively in Ref. 5, so we summarize just the main properties. For $x\to-\infty$ they fall off like Gaussians, which is due to the fact that the Bloch functions are analytic in the lower half of the complex plane. For positive x again Eq. (22) is valid.

At the first glance one would suppose that these Wannier functions $\widetilde{W}_0^\beta(x)$ are the better localized ones. But a more detailed analysis reveals that this is not true. This is due to the exponentially damped oscillating tail on the right side, whose oscillations are a few times greater then the two oscillating tails of $W_0^\beta(x)$. For large β, for instance, the oscillations of $\widetilde{W}_0^\beta(x)$ are twice as large as the oscillations of $W_0^\beta(x)$, since we have

$$W_0^\beta(x)=\left(\frac{2\beta}{\pi}\right)^{1/4}\left(e^{-\beta x^2}-\frac{e^{-\beta/2}}{2}\,(e^{-\beta(x+1)^2}+e^{-\beta(x-1)^2})\right)+O(e^{-\beta}), \tag{25}$$

$$\widetilde{W}_0^\beta(x)=\left(\frac{2\beta}{\pi}\right)^{1/4}(e^{-\beta x^2}-e^{-\beta/2}e^{-\beta(x-1)^2}). \tag{26}$$

A criterium for localization is the uncertainty Δx. Since the mean value vanishes because of symmetry, i.e., $\langle x\rangle=0$, it is sufficient to compute $\langle x^2\rangle_{W_0^\beta}$, where the index W_0^β denotes that the mean value $\langle x^2\rangle$ with respect to the Wannier function W_0^β is considered. A straightforward calculation yields

$$\langle x^2\rangle_{W_0^\beta}=1/4\beta. \tag{27}$$

Note that the same result is also valid for the original Gauss functions: $\langle x^2\rangle_{G^\beta}=1/4\beta$. The expectation value $\langle x\rangle_{\widetilde{W}_0^\beta}$ vanishes, too, although $\widetilde{W}_0^\beta$ is not symmetric. This fact can be easily checked by direct calculation, or alternatively one uses the concept of the Berry phase:[14,15] Using kq-representation[16] we get

$$\langle \widetilde{W}_0^\beta, x\widetilde{W}_0^\beta\rangle = i\int_0^1 dx\int_{-\pi}^{\pi} dk\ \tilde{\omega}_k^\beta(x)^*\ \frac{\partial}{\partial k}\ \tilde{\omega}_k^\beta(x)=\frac{1}{2\pi}\,a^\beta(2\pi), \tag{28}$$

where $\tilde{\omega}_k^\beta(x)$ is the periodic part of the Bloch functions $\tilde{\phi}_k^\beta(x)$, and $a^\beta(2\pi)$ is the Berry phase. Since the Berry phase does not depend on the particular choice of the Wannier functions, but only on the ''energy band,'' i.e., on the subspace which is spanned by the Wannier functions, we infer $\langle x\rangle_{\widetilde{W}_0^\beta}=\langle x\rangle_{W_0^\beta}=0$.

One quickly verifies that the expectation value $\langle x^2\rangle_{\widetilde{W}_0^\beta}$ is given by the following expression

$$\langle x^2\rangle_{\widetilde{W}_0^\beta}=\frac{1}{4\beta}+\frac{1}{2\pi}\int_{-\pi}^{\pi} dk\left|\frac{\partial\varphi^\beta(k)}{\partial k}\right|^2, \tag{29}$$

where $\varphi^\beta(k)$ is definded by

$$e^{i\varphi^\beta(k)}=\frac{\widetilde{N}^\beta(k)}{N^\beta(k)}=\prod_{n=1}^{\infty}\,(1+e^{-(\beta/2)(2n-1)}e^{ik})^{1/2}(1+e^{-(\beta/2)(2n-1)}e^{-ik})^{-1/2}. \tag{30}$$

An explicit expression for $\varphi^\beta(k)$ is given by

$$\varphi^\beta(k)=\sum_{n=1}^{\infty}\left[\frac{k}{2}-\arctan\left(\tanh\left(\frac{\beta}{4}\,(2n-1)\right)\tan\frac{k}{2}\right)\right]. \tag{31}$$

From Eq. (29) we infer immediately $\Delta x_{\widetilde{W}_0^\beta}>\Delta x_{W_0^\beta}$, i.e., the symmetric Wannier functions W_0^β are better localized than Wannier's Wannier functions $\widetilde{W}_0^\beta$. The integral in Eq. (29) can be evaluated and we obtain

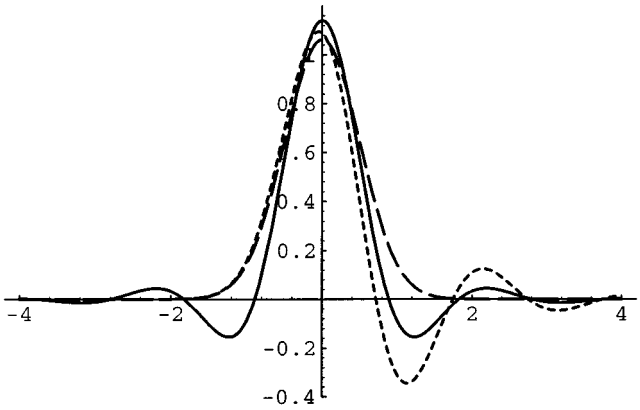

FIG. 1. Gauss function (dashed line), symmetric Wannier function and Wannier's Wannier function (dotted line) for $\beta=2$.

$$\langle x^2\rangle_{\tilde{W}_0^\beta}=\frac{1}{4\beta}+\frac{1}{2}\sum_{n=1}^{\infty}\frac{e^{-\beta n}}{(1-e^{-\beta n})^2}=\frac{1}{4\beta}+\frac{1}{48}\left(\frac{\theta_1'''\left(0\middle|\frac{i\beta}{2\pi}\right)}{\theta_1'\left(0\middle|\frac{i\beta}{2\pi}\right)}+1\right)$$

$$=\frac{1}{48}\left[1-\left(\frac{2\pi}{\beta}\right)^2\frac{\theta_1'''\left(0\middle|\frac{2\pi i}{\beta}\right)}{\theta_1'\left(0\middle|\frac{2\pi i}{\beta}\right)}\right]. \tag{32}$$

For small and large values of β one easily proves

$$(\Delta x_{\tilde{W}_0^\beta})^2=\frac{1}{4\beta}+\frac{1}{2}e^{-\beta}+O(e^{-2\beta}) \quad \text{for large } \beta \tag{33}$$

$$(\Delta x_{\tilde{W}_0^\beta})^2=\frac{\pi^2}{12\beta^2}+\frac{1}{48}+O\left(\frac{e^{-4\pi^2/\beta}}{\beta^2}\right) \quad \text{for small } \beta. \tag{34}$$

Thus for large values of β our symmetric and Wannier's Wannier functions are almost equally well localized, but for small β the symmetric Wannier functions are much better localized. See also Fig. 1 and Fig. 2.

Hence the advantage of $\tilde{W}_0^\beta(x)$ is not its localization property but the fact that there is an explicit expression for it. On the other hand, the advantage of $W_0^\beta(x)$ is its symmetry. Moreover the uncertainty Δx is smaller for the symmetric Wannier functions $W_0^\beta(x)$. In fact, this choice of Wannier functions has the smallest possible Δx. This is due to the fact that Eq. (29) is valid for any Wannier function, if $\varphi^\beta(k)$ denotes the phase by which the corresponding Bloch functions differ from our symmetric Bloch functions $\phi_k^\beta(x)$.

V. THE LIMIT $\beta\to 0$

Let us study the limit $\beta\to 0$. Note that the Gauss functions G^β converge to zero uniformly, but they do not converge to zero with respect to the $L^2(\mathbb{R})$-norm, since $\|G^\beta\|=1$ for all β. Note however that the Gauss functions converge in the weak topology to the zero function, i.e., for each square integrable function ψ the equation $\lim_{\beta\to 0}(\psi,G^\beta)_\infty=0$ holds. Since the Bloch functions and Wannier functions are normalized with respect to $L^2([0,1])$ and $L^2(\mathbb{R})$, respectively, their limits are not necessarily trivial.

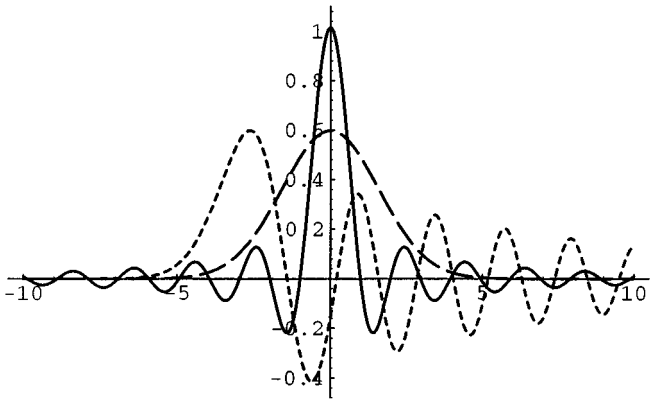

FIG. 2. Gauss function (dashed line), symmetric Wannier function and Wannier's Wannier function (dotted line) for $\beta=0.2$.

When calculating the limit of the Bloch functions $\phi_k^\beta(x)$ we use a formula which may be obtained by using the Jacobi transformation of the theta function:[5,10–12]

$$\phi_k^\beta(x)=e^{ikx}\omega_k^\beta(x), \tag{35a}$$

$$\omega_k^\beta(x)=\frac{1}{\sqrt{2\pi}}\frac{\theta_3\left(\frac{i\pi}{\beta}\frac{k}{2}+\pi x\middle|\frac{i\pi}{\beta}\right)}{\sqrt{\theta_3\left(\frac{i\pi}{\beta}k\middle|\frac{2i\pi}{\beta}\right)}}. \tag{35b}$$

Here $\omega_k^\beta(x)$ is the periodic Bloch factor. Now the limit of the Bloch functions $\phi_k^\beta(x)$ for $\beta\to 0$ is immediately obtained:

$$\phi_k^0(x)=\lim_{\beta\to 0}\phi_k^\beta(x)=\frac{1}{\sqrt{2\pi}}e^{i(k-2\pi l)x} \tag{36}$$

for $k\neq(2n+1)\pi$, where l is the unique integer for which $|k-2\pi l|<\pi$ holds, and

$$\lim_{\beta\to 0}\phi_k^\beta(x)=\frac{1}{\sqrt{\pi}}\cos(\pi x)\quad\text{for}\quad k=(2n+1)\pi. \tag{37}$$

The convergence is uniform with respect to x for any subset of $\mathbb{R}$ and uniform with respect to k for any interval $[(2n-1)\pi+\epsilon,(2n+1)\pi-\epsilon],\epsilon>0$. The convergence cannot be uniform with respect to k in a neighborhood of $(2n+1)\pi$ since the limit function $\Phi^0=\lim_{\beta\to 0}\Phi^\beta$ is not continuous at $k=(2n+1)\pi$. Note that the Bloch functions ϕ_k^β converge with respect to the $L^2([0,1])$-norm because of the uniform convergence with respect to x. Note that this limit describes the case of free electrons.

The limit of the Wannier functions is given by

$$W_0^0(x)=\lim_{\beta\to 0}W_0^\beta(x)=\frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}\phi_k^0(x)dk=\frac{\sin(\pi x)}{\pi x}. \tag{38}$$

Of course the Wannier functions converge uniformly with respect to x. Moreover, they converge also with respect to the $L^2(\mathbb{R})$-norm.

Note that the Bloch functions corresponding to $\tilde{\mathbf{N}}^\beta$ do not converge, since the quotient

$$\frac{\tilde{\mathbf{N}}^{\beta}}{\mathbf{N}^{\beta}}=\prod_{n=1}^{\infty}\frac{(1+q^{2n-1}e^{ik})^{1/2}}{(1+q^{2n-1}e^{-ik})^{1/2}} \tag{39}$$

does not converge for $\beta\to 0$. Hence there does not exist a limit [with respect to the $L^2(\mathbb{R})$-norm] for the corresponding Wannier functions either. However, numerical calculations suggest that the Wannier functions converge to zero uniformly on $\mathbb{R}$; see Figs. 1 and 2.

VI. EXPECTATION VALUES

One of the most interesting properties of a physical system is its energy. Here we consider the expectation values of some Hamiltonians $H=(1/2m)P^2+V$ with respect to the thetalike Bloch functions, where the potentials V are chosen specifically. Note that the expectation values are independent of the specific choice of the normalization constants and hence all the following results are equally valid for both cases.

A. The kinetic energy

First we compute the expectation value of the kinetic energy $T=(1/2m)P^2$. Due to the specific normalization of the Bloch functions, the expectation value $\langle P^2\rangle_k^{\beta}$ is given by

$$\langle P^2\rangle_k^{\beta}=2\pi(\phi_k^{\beta},P^2\phi_k^{\beta})_1. \tag{40}$$

A straightforward calculation gives the following result for the expectation value $\langle P^2\rangle_k^{\beta}$:

$$\langle P^2\rangle_k^{\beta}=\beta-\sum_{n=-\infty}^{\infty}e^{-\beta(n^2/2)}e^{-ikn}\beta^2n^2\Big/\sum_{n=-\infty}^{\infty}e^{-\beta(n^2/2)}e^{-ikn}. \tag{41}$$

This result may be written in terms of θ-functions, and by using Jacobi's transformation[10–12] we get an alternative formula:

$$\begin{aligned}\langle P^2\rangle_k^{\beta}&=\beta+2\beta^2\frac{\partial}{\partial\beta}\ln\,\theta_3\left(\frac{k}{2}\middle|\,\frac{i\beta}{2\pi}\right)\\&=2\beta^2\frac{\partial}{\partial\beta}\ln\left(e^{-k^2/2\beta}\theta_3\left(\frac{i\pi}{\beta}\,k\middle|\,\frac{2\pi i}{\beta}\right)\right)\\&=\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}(k-2\pi n)^2\Big/\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}.\end{aligned} \tag{42}$$

These formulas do not only hold true for real k but also for complex k, if we define

$$\langle P^2\rangle_k^{\beta}=2\pi(\phi_{k*}^{\beta},P^2\phi_k^{\beta})_1, \tag{43}$$

where k^* is the complex conjugate of k. Now we see easily that $\langle P^2\rangle_k^{\beta}$ is an analytic function of k in the strip $\{k=\kappa+i\rho:|\rho|<\beta/2\}$ for all $\beta>0$.

For the limit of large and small β one can derive simple formulas:

$$\langle P^2\rangle_k^{\beta}=\beta-2e^{-\beta/2}\beta^2\cos k+O(e^{-\beta}\beta^2)\quad\text{for}\quad\beta\to\infty, \tag{44a}$$

$$\langle P^2\rangle_k^{0}=k^2\quad\text{for}\quad k\in[-\pi,\pi]. \tag{44b}$$

There are some interesting monotonicity properties of $\langle P^2\rangle_k^{\beta}$: First, $\langle P^2\rangle_k^{\beta}$ is a strictly monotone increasing function of β. This is easily seen by considering the derivative of Eq. (42) and using Cauchy–Schwarz's inequality

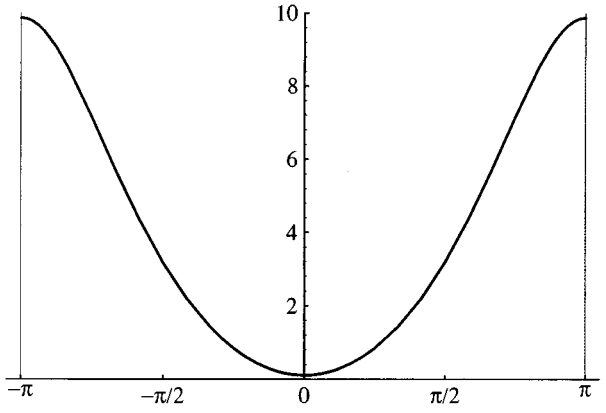

FIG. 3. $\langle P^2\rangle_k^\beta$ for $\beta=3$.

$$\frac{\partial}{\partial\beta}\langle P^2\rangle_k^\beta=\frac{1}{2\beta^2}\left(\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}(k-2\pi n)^4\Big/\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}\right)$$
$$-\frac{1}{2\beta^2}\left[\left(\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}(k-2\pi n)^2\right)^2\Big/\left(\sum_{n=-\infty}^{\infty}e^{-(k-2\pi n)^2/2\beta}\right)^2\right]>0. \tag{45}$$

$\langle P^2\rangle_k^\beta$ is also a strictly monotone increasing function of k in the interval $[0,\pi]$, because $\beta^2(\partial/\partial\beta)\ln\theta_3(k/2|i\beta/2\pi)$ is a strictly monotone increasing function of k in the interval $[0,\pi]$. A proof of this fact is given in the Appendix.

Let us have a look on the symmetry properties of $\langle P^2\rangle_k^\beta$. Obviously $\langle P^2\rangle_k^\beta$ is an even function of k. We could infer this also from the time reversal symmetry $(\phi_k^\beta)^*=\phi_{-k}^\beta$ together with the reality of P^2 or alternatively from the inversion symmetry $\phi_k^\beta(x)=\phi_{-k}^\beta(-x)$ together with the inversion symmetry of P^2. Moreover $\langle P^2\rangle_k^\beta$ is periodic in k-space.

These properties are illustrated by the following figures: Fig. 3 shows the expectation value $\langle P^2\rangle_k^\beta$ for $\beta=3$; Fig. 4 displays $\langle P^2\rangle_k^\beta$ as a function of k and β. One clearly sees the symmetry and monotonicity properties with respect to k and β, respectively.

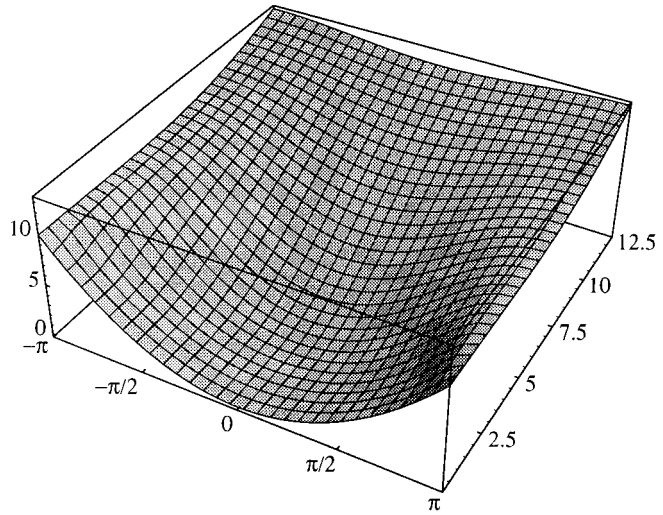

FIG. 4. $\langle P^2\rangle_k^\beta$ for $-\pi\leqslant k\leqslant\pi$ and $0.1\leqslant\beta\leqslant 12.5$.

B. The potential energy

Next we consider the expectation value of a periodic potential, whose period is equal to 1. Thus we assume $V(x)=V(x+n)$, $n\in\mathbb{Z}$. First we derive a formula for general potentials and later we discuss some special examples. To this end we need the following two θ-functions:

$$\theta_2(z|\tau)=\sum_{n=-\infty}^{\infty} e^{i(n+1/2)^2\pi\tau}e^{i(2n+1)z}, \tag{46}$$

$$\theta_4(z|\tau)=\sum_{n=-\infty}^{\infty} (-1)^n e^{in^2\pi\tau}e^{i2nz}. \tag{47}$$

We obtain the following formula:

$$\langle V\rangle_k^\beta = 2\pi(\phi_k^\beta, V\phi_k^\beta)_1 = V_+^\beta + \frac{\theta_4\left(\frac{k}{2}\middle|\frac{i\beta}{2\pi}\right)}{\theta_3\left(\frac{k}{2}\middle|\frac{i\beta}{2\pi}\right)} V_-^\beta = V_+^\beta + \frac{\theta_2\left(\frac{i\pi k}{\beta}\middle|\frac{2\pi i}{\beta}\right)}{\theta_3\left(\frac{i\pi k}{\beta}\middle|\frac{2\pi i}{\beta}\right)} V_-^\beta, \tag{48}$$

where we have used the definitions

$$V_\pm^\beta := \left(\frac{2\beta}{\pi}\right)^{1/2}\int_{-\infty}^{\infty} dx\ e^{-2\beta x^2}\frac{1}{2}\left(V(x)\pm V\left(x+\frac{1}{2}\right)\right). \tag{49}$$

Note that the expectation value of the potential depends in a simple manner on the actual shape of the potential, namely only the two parameters V_+^β and V_-^β enter in Eq. (48). Let us discuss the k-dependence of $\langle V\rangle_k^\beta$ now. First we observe that $\langle V\rangle_k^\beta$ is constant if V_-^β vanishes, i.e., if $V(x)$ has a period equal to $\frac{1}{2}$. Next assume $V_-^\beta\neq 0$. Since $\theta_4(k/2|i\beta/2\pi)$ is a strictly monotone increasing function (see the Appendix) for $k\in[0,\pi]$ and $\theta_3(k/2|i\beta/2\pi)$ is strictly monotone decreasing in the same interval, $\langle V\rangle_k^\beta$ is also strictly monotone for $k\in[0,\pi]$: it is strictly monotone increasing if V_-^β is positive and strictly monotone decreasing if V_-^β is negative.

Another interesting property of $\langle V\rangle_k^\beta$ is its inversion symmetry in k-space. This is due to the symmetry relation $|\phi_k^\beta(x)|^2=|\phi_{-k}^\beta(x)|^2$ which follows from the time reversal symmetry $(\phi_k^\beta(x))^*=\phi_{-k}^\beta(x)$. Note also that $\langle V\rangle_k^\beta$ is periodic in k-space.

Note that Eq. (48) is not only valid for real k but also for complex values of k provided that we define

$$\langle V\rangle_k^\beta = 2\pi(\phi_{k^*}^\beta, V\phi_k^\beta)_1. \tag{50}$$

With this definition $\langle V\rangle_k^\beta$ is analytic in the strip $\{k=\kappa+i\rho:|\rho|<\beta/2\}$ for all $\beta>0$. This follows immediately from Eq. (48) and the properties of the theta functions. But we can argue more generally, too. Let us assume that the potential $V(x)$ is bounded, i.e., the operator V is bounded and therefore continuous. Then the analyticity of Φ^β implies the analyticity of $V\Phi^\beta$ and together with the analyticity of Φ_*^β the analyticity of $\langle V\rangle_k^\beta$ follows. Here Φ_*^β is defined by $\Phi_*^\beta(k):=(\phi_{k^*}^\beta)^*$, where k^* is the complex conjugate of k.

Let us discuss the two coefficients $V_\pm^\beta$ in more detail. First we observe that only the symmetric part enters in the coefficients $V_\pm^\beta$, i.e., if $V(x)=V_e(x)+V_0(x)$, where $V_e(x)$ and $V_0(x)$ are the even and odd parts of the potential $V(x)$, then we have

$$V_\pm^\beta = \left(\frac{2\beta}{\pi}\right)^{1/2}\int_{-\infty}^{\infty} dx e^{-2\beta x^2}\frac{1}{2}\left(V_e(x)\pm V_e\left(x+\frac{1}{2}\right)\right). \tag{51}$$

This property is due to the inversion symmetry $|\phi_k^\beta(x)|^2=|\phi_k^\beta(-x)|^2$, which follows from Eqs. (6) and (7), and hence only the even part $V_e(x)$ contributes to the expectation value $\langle V\rangle_k^\beta$. Now let

$$V(x)=\sum_{n=-\infty}^{\infty} c_n e^{i2\pi nx} \tag{52}$$

be the Fourier expansion of the potential V. Then we have

$$V_+^\beta = c_0 + \sum_{n=1}^{\infty} e^{-2\pi^2 n^2/\beta}(c_{2n}+c_{-2n}), \tag{53a}$$

$$V_-^\beta = \sum_{n=0}^{\infty} e^{-\pi^2(2n+1)^2/2\beta}(c_{2n+1}+c_{-2n-1}), \tag{53b}$$

from which we could also infer immediately that only the even part $V_e(x)$ contributes to $\langle V\rangle_k^\beta$.

In the limit of large and small β one gets simple formulas. For small β we have the following formula:

$$\langle V\rangle_k^\beta = c_0 + \frac{2e^{-\pi^2/\beta}\cosh(k\pi/\beta)}{1+4e^{-2\pi^2/\beta}\cosh^2(k\pi/\beta)}(c_1+c_{-1}) + O(e^{-2\pi^2})$$

$$= c_0 + \frac{1}{2\cosh((\pi-k)\pi/\beta)}(c_1+c_{-1}) + O(e^{-\pi^2}), \tag{54}$$

where $k\in[-\pi,\pi]$ in the first equation and $k\in[0,\pi]$ in the second one. From this it follows immediately that the limit $\beta\to 0$ is given by

$$\langle V\rangle_k^0 = \lim_{\beta\to 0}\langle V\rangle_k^\beta = c_0 \quad \text{for} \quad k\neq(2n+1)\pi,\ \ n\in\mathbb{Z}, \tag{55a}$$

$$\langle V\rangle_k^0 = \lim_{\beta\to 0}\langle V\rangle_k^\beta = c_0+\tfrac{1}{2}(c_1+c_{-1}) \quad \text{for} \quad k=(2n+1)\pi,\ \ n\in\mathbb{Z}. \tag{55b}$$

This discontinuity is not surprising since the limit of the Bloch function for $\beta\to 0$ is not continuous either. The physical reason is evident, too: If $-\pi<k<\pi$, then $\langle V\rangle_k^0$ is simply the energy correction of the first energy band in first order perturbation theory. For $k=\pi$ perturbation theory for degenerate eigenvalues has to be used and the energy correction in first order perturbation theory is given by $c_0\pm\frac{1}{2}(c_1+c_{-1})$. Hence $\langle V\rangle_\pi^\beta$ is the energy correction for the first band if c_1+c_{-1} is negative. If $c_1+c_{-1}>0$, however, $\langle V\rangle_\pi^\beta$ is the first order energy correction for the second band. This is not surprising since $V(x)$ is attractive at $x=n$ and repulsive at $x=n+\frac{1}{2}$ for $c_1+c_{-1}<0$ (if we retain only the first terms of the Fourier expansion) and hence Bloch functions composed of Gaussian orbitals centered at $x=n$ provide us with a good approximation. For $c_1+c_{-1}>0$, however, the situation is opposite, namely, $V(x)$ is repulsive at $x=n$ and attractive at $x=n+\frac{1}{2}$, and hence it is not reasonable to use Gaussian orbitals centered at $x=n$ to construct Bloch functions for this case.

We also mention briefly the case for large β. Here we have

$$\langle V\rangle_k^\beta = V_+^\beta + (1-4e^{-\beta/2}\cos k)V_-^\beta + O(e^{-\beta}) \tag{56}$$

and in the limit $\beta\to\infty$ we obtain $\langle V\rangle_k^\infty = V(0)$.

The two cases of small and large β are illustrated in Figs. 5 and 6, respectively. In both cases we have chosen $V_-^\beta = 2\exp(-\pi^2/2\beta)$ and $V_+^\beta = 0$.

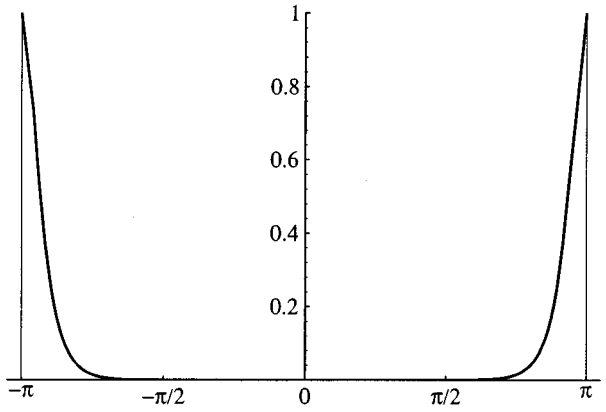

FIG. 5. $\langle V\rangle_k^\beta$ for $\beta=0.5$ with $V_-^\beta=2\exp(-\pi^2/2\beta)$ and $V_+^\beta=0$.

Example

We consider the following potential:

$$V^\alpha(x)=-\sqrt{\frac{\alpha}{\pi}}\sum_{n=-\infty}^{\infty}e^{-\alpha(x-n)^2}=-\sqrt{\frac{\alpha}{\pi}}\,e^{-\alpha x^2}\theta_3\left(i\alpha x\middle|\frac{i\alpha}{\pi}\right)$$
$$=-\theta_3\left(\pi x\middle|\frac{i\pi}{\alpha}\right)$$
$$=-1-2\sum_{n=1}^{\infty}e^{-(\pi^2/\alpha)n^2}\cos(2\pi nx). \tag{57}$$

Note that the limit of V^α for $\alpha\to\infty$ is the periodic delta function

$$\lim_{\alpha\to\infty}V^\alpha(x)=-\sum_{n=-\infty}^{\infty}\delta(x-n), \tag{58}$$

whereas for $\alpha\to 0$ we have

$$\overline{V}(x):=\lim_{\alpha\to 0}e^{\pi^2/\alpha}(V^\alpha(x)+1)=-2\cos(2\pi x). \tag{59}$$

The coefficients $V_\pm^\beta$ now depend on alpha and are denoted by $V_\pm^\beta(\alpha)$. They read as follows:

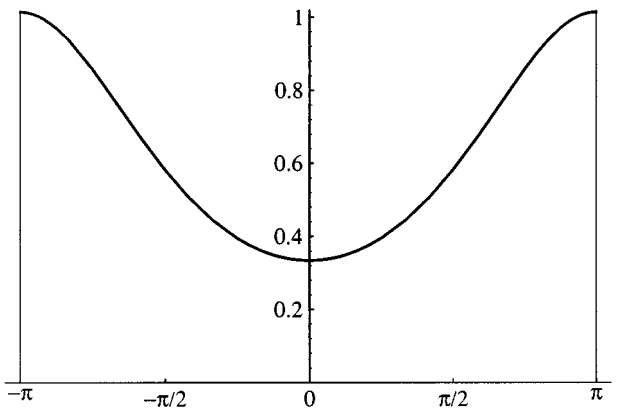

FIG. 6. $\langle V\rangle_k^\beta$ for $\beta=4$ with $V_-^\beta=2\exp(-\pi^2/2\beta)$ and $V_+^\beta=0$.

$$V_+^\beta(\alpha) = -\left(\frac{\alpha\beta}{2\pi(2\beta+\alpha)}\right)^{1/2}\theta_3\left(0\left|\frac{i\alpha\beta}{2\pi(2\beta+\alpha)}\right.\right) = -\theta_3\left(0\left|\frac{2\pi i(2\beta+\alpha)}{\alpha\beta}\right.\right), \tag{60a}$$

$$V_-^\beta(\alpha) = -\left(\frac{\alpha\beta}{2\pi(2\beta+\alpha)}\right)^{1/2}\theta_4\left(0\left|\frac{i\alpha\beta}{2\pi(2\beta+\alpha)}\right.\right) = -\theta_2\left(0\left|\frac{2\pi i(2\beta+\alpha)}{\alpha\beta}\right.\right). \tag{60b}$$

In the limit $\alpha\to\infty$ we obtain

$$V_+^\beta(\infty) = -\left(\frac{\beta}{2\pi}\right)^{1/2}\theta_3\left(0\left|\frac{i\beta}{2\pi}\right.\right) = -\theta_3\left(0\left|\frac{2\pi i}{\beta}\right.\right) = -\sum_{n=-\infty}^{\infty}\exp\left(-\frac{2\pi^2}{\beta}n^2\right), \tag{61a}$$

$$V_-^\beta(\infty) = -\left(\frac{\beta}{2\pi}\right)^{1/2}\theta_4\left(0\left|\frac{i\beta}{2\pi}\right.\right) = -\theta_2\left(0\left|\frac{2\pi i}{\beta}\right.\right) = -\sum_{n=-\infty}^{\infty}\exp\left(-\frac{\pi^2}{2\beta}(2n+1)^2\right). \tag{61b}$$

For the other limit $\alpha\to 0$ we have

$$\overline{V}_+^\beta = \lim_{\alpha\to 0} e^{\pi^2/\alpha}(V_+^\beta + 1) = 0, \tag{62a}$$

$$\overline{V}_-^\beta = -\lim_{\alpha\to 0} e^{\pi^2/\alpha}V_-^\beta = \exp\left(-\frac{\pi^2}{2\beta}\right). \tag{62b}$$

Comparing Eqs. (61) with (62) we see that they look quite similar for small β. In fact, if we introduce the rescaled potential

$$\overline{V}^\alpha(x) = e^{\pi^2/\alpha}(V^\alpha(x)+1), \tag{63}$$

we can easily prove that the corresponding coefficients $\overline{V}_\pm^\beta(\alpha)$ are nearly independent of α for small β and are approximately given by

$$\overline{V}_+^\beta(\alpha) = O\left[\exp\left(-\frac{3\pi^2}{\alpha}\right)\exp\left(-\frac{2\pi^2}{\beta}\right)\right], \tag{64a}$$

$$\overline{V}_-^\beta(\alpha) = 2\exp\left(-\frac{\pi^2}{2\beta}\right) + O\left[\exp\left(-\frac{8\pi^2}{\alpha}\right)\exp\left(-\frac{9\pi^2}{2\beta}\right)\right]. \tag{64b}$$

The reason for this is that for small β only the first terms of the Fourier expansion of the potential $V(x)$ are relevant. In Fig. 7 we present the expectation value $\langle V^\alpha\rangle_k^\beta$ of the rescaled potential $\overline{V}^\alpha(x)$ as a function of k and β. Note the symmetry and monotonicity of $V^\alpha(x)$. The expectation value of the corresponding Hamiltonian $H = P^2 + d\overline{V}^\alpha(x)$ is shown in Fig. 8, where we have chosen $\alpha = 20$ and $d = 5$.

VII. THE CONNECTION BETWEEN THE ANALYTICITY OF THE BLOCH FUNCTIONS AND THE EXPONENTIAL FALL OFF OF THE WANNIER FUNCTIONS

So far we have discussed only the special Bloch functions which are composed of Gaussian orbitals. Now we want to consider the general case of analytic Bloch functions. Here we call a Bloch function Φ analytic if the functions $\phi_x : \phi_x(k) = \Phi(k,x)$ are analytic functions of k for all $x \in \mathbb{R}$.

It is well known that the analyticity of the Bloch functions implies the exponential decrease of the Wannier functions[6,8,7] provided that the Bloch functions are continuous functions of x. But in general the Bloch functions need not to be continuous since they are only square integrable in the unit cell. Hence the Wannier functions may have infinitely many singularities and thus they do not fall off exponentially in the usual sense. Nevertheless they fall off ''almost'' exponentially in a sense explained later.

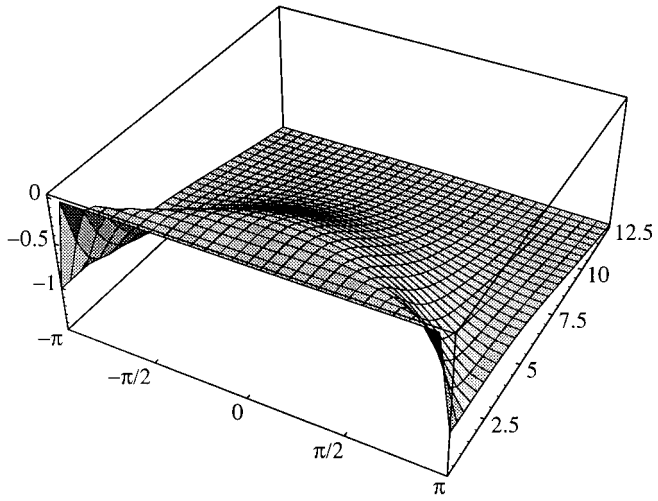

FIG. 7. $\langle V^{\alpha}\rangle_k^{\beta}$ for $\alpha=20$, $-\pi\leqslant k\leqslant\pi$, $0.1\leqslant\beta\leqslant 12.5$.

First we review the traditional arguments. Throughout this section we assume that $\Phi:\Phi(k,x)=\phi_k(x)$ be a Bloch function which is analytic in the open strip $\{k:|\mathrm{Im}\,k|<\Gamma\}$ and which is periodic in k-space: $\Phi(k+2\pi,x)=\Phi(k,x)$. Moreover we assume that Φ is normalized in the following way: $(\phi_k,\phi_k)_1=\int_0^1 dx|\phi_k(x)|^2=1/2\pi$ for real k. Again we adopt the following convention for the corresponding Wannier functions:

$$W_n(x)=\frac{1}{(2\pi)^{1/2}}\int_0^{2\pi} dk\,\phi_k(x)e^{-ikn}. \tag{65}$$

First note that we have

$$|W_0(x+n)|\leqslant e^{\mp\gamma n}\frac{1}{(2\pi)^{1/2}}\int_0^{2\pi} dk|\phi_{k\pm i\gamma}(x)| \tag{66}$$

for $\gamma<\Gamma$, which is easily proved by shifting the path of integration:

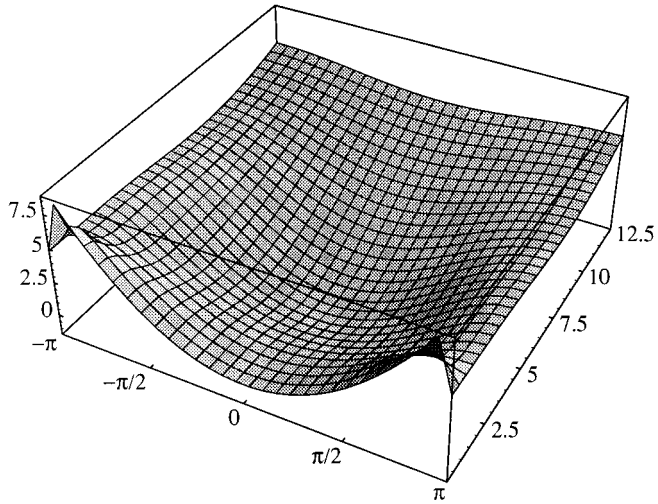

FIG. 8. $\langle H\rangle_k^{\beta}$ for $d=5$, $-\pi\leqslant k\leqslant\pi$, $0.1\leqslant\beta\leqslant 12.5$.

$$W_0(x+n)=\frac{1}{(2\pi)^{1/2}}\int_0^{2\pi}dk\,\phi_k(x)e^{ikn}=e^{\mp\gamma n}\frac{1}{(2\pi)^{1/2}}\int_0^{2\pi}dk\,\phi_{k\pm i\gamma}(x)e^{ikn}. \tag{67}$$

If we assume that the Bloch functions are bounded, e.g., if the Bloch functions are continuous with respect to x, then we derive immediately

$$|W_0(x+n)|\leqslant e^{\mp\gamma(x+n)}B(\gamma), \tag{68}$$

or equivalently

$$|W_0(x)|\leqslant e^{-\gamma|x|}B(\gamma), \tag{69}$$

where $B(\gamma)$ is defined as

$$B(\gamma)=\sup_{-1/2\leqslant x\leqslant 1/2}e^{\pm\gamma x}\frac{1}{(2\pi)^{1/2}}\int_0^{2\pi}dk|\phi_{k\pm i\gamma}(x)|. \tag{70}$$

Hence the Wannier functions fall off exponentially. The converse statement is true as well, i.e., if Eq. (69) is valid for all $\gamma<\Gamma$, then the corresponding Bloch functions are analytic in the open strip $\{k:|\mathrm{Im}\,k|<\Gamma\}$ and are periodic in k and bounded. This is easily verified by considering the expansion

$$\Phi(k,x)=\frac{1}{\sqrt{2\pi}}\sum_{n=-\infty}^{\infty}e^{ikn}W_n(x). \tag{71}$$

We summarize this in the following theorem:

Theorem 1: *Let Φ be periodic and analytic in the open strip $\{k:|\mathrm{Im}\,k|<\Gamma\}$ for fixed x, where Γ does not depend on x. If Φ is bounded in $\{k:|\mathrm{Im}\,k|\leqslant\gamma\}\times[-\frac{1}{2},\frac{1}{2}]$ with $\gamma<\Gamma$, then the Wannier functions fall off exponentially*

$$|W_0(x)|\leqslant e^{-\gamma|x|}B(\gamma), \tag{72}$$

where $B(\gamma)$ is an appropriate constant. Conversely, if Eq. (72) holds for all $\gamma<\Gamma$, then Φ is periodic and analytic in the open strip $\{k:|\mathrm{Im}\,k|<\Gamma\}$ for fixed x, and it is bounded in $\{k:|\mathrm{Im}k|\leqslant\gamma\}\times[-\frac{1}{2},\frac{1}{2}]$ for all $\gamma<\Gamma$.

In general, a square integrable[17] Bloch function ϕ_k need not be bounded, hence Φ is not bounded. So Theorem 1 is not applicable in general. Nevertheless the Wannier functions fall off ''almost'' exponentially. This means that we do not consider the probability density $|W_0(x)|^2$, instead of it we discuss the probability for finding the electron in the interval $[y-\epsilon,y+\epsilon]$, which is given by $\int_{y-\epsilon}^{y+\epsilon}|W_0(x)|^2dx$. In fact, it is this quantity that falls off exponentially.

To discuss this behavior we need the concept of a vector-valued analytic function. Let $\mathscr{H}$ be a Hilbert space and $\psi:D\to\mathscr{H}$, $D\subseteq\mathbb{C}$ a vector-valued function of the complex variable z, i.e., $\psi(z)\in\mathscr{H}$, then ψ is called a vector-valued analytic function[18,19] if for all $z\in D$ the limit $\lim_{h\to 0}[\psi(z+h)-\psi(z)]/h=:\psi'(z)$ exists as a strong limit[20] in $\mathscr{H}$. We want to apply this definition to the Bloch function Φ. Of course, for fixed k, the Bloch function ϕ_k is not square integrable on $\mathbb{R}$ and hence the Bloch function Φ cannot be interpreted as an analytic function from $\mathbb{C}$ to $L^2(\mathbb{R})$. But if we restrict the domain of definition of the Bloch functions ϕ_k to the unit cell $[-\frac{1}{2},\frac{1}{2}]$, these restricted functions are members of $L^2([-\frac{1}{2},\frac{1}{2}])$. If the domain of definition of Φ is restricted correspondingly we may view Φ as a vector valued function from $D\subseteq\mathbb{C}$ into $L^2([-\frac{1}{2},\frac{1}{2}])$. If this restriction of Φ is an analytic vector-valued function of k, we will call Φ an analytic vector-valued function.

Theorem 2: *If the Bloch function Φ is an analytic vector-valued function of k in the open strip $\{k:|\mathrm{Im}\,k|<\Gamma\}$ and if Φ is a periodic function of k with period 2π, then the following inequality holds for all $\gamma<\Gamma$:*

$$\int_{y-\epsilon}^{y+\epsilon}|W_0(x)|^2dx\leqslant e^{-2\gamma|y|}B(\gamma,\epsilon). \tag{73}$$

Proof: Squaring the inequality (66) and integrating we arrive at

$$\int_{y-\epsilon}^{y+\epsilon} |W_0(x+n)|^2 dx \leqslant e^{\mp 2\gamma n} A_\pm(\gamma,\epsilon,y), \tag{74a}$$

$$A_\pm(\gamma,\epsilon,y) = \frac{1}{(2\pi)} \int_{y-\epsilon}^{y+\epsilon} dx \left[\int_0^{2\pi} dk |\phi_{k\pm i\gamma}(x)| \right]^2. \tag{74b}$$

Since Φ is a vector-valued analytic function of k, the constant $A_\pm(\gamma,\epsilon,y)$ is finite. We rewrite the inequality in the following way:

$$\int_{y+n-\epsilon}^{y+n+\epsilon} |W_0(x)|^2 dx \leqslant e^{\mp 2\gamma(y+n)} \underbrace{e^{\pm 2\gamma y} A_\pm(\gamma,\epsilon,y)}_{=:B_\pm(\gamma,\epsilon,y)}. \tag{75}$$

Again the analyticity of Φ assures that the maximum

$$B_\pm(\gamma,\epsilon) := \max_{-\frac{1}{2} \leqslant y \leqslant \frac{1}{2}} B_\pm(\gamma,\epsilon,y)$$

is finite. Hence we have for all $y \in \mathbb{R}$ [we define $B(\gamma,\epsilon) := \max(B_\pm(\gamma,\epsilon))$]

$$\int_{y-\epsilon}^{y+\epsilon} |W_0(x)|^2 dx \leqslant e^{-2\gamma|y|} B(\gamma,\epsilon). \tag{76}$$

The converse of this theorem is true as well.

Theorem 3: *Let W_0 decrease such that the inequality*

$$\int_{y-1/2}^{y+1/2} |W_0(x)|^2 dx \leqslant e^{-2\gamma|y|} B(\gamma) \tag{77}$$

holds for all $\gamma < \Gamma$. Then the Bloch function Φ is an analytic vector-valued function of k, which has the periodicity property $\Phi(k+2\pi n,x) = \Phi(k,x)$.

Proof: Let $\|\cdot\|_1$ denote the $L^2([-\frac{1}{2}, \frac{1}{2}])$ norm. Then we have the inequality $\|W_n\|_1 \leqslant e^{-\gamma|n|} B(\gamma)$ and hence the sum

$$\Phi(k,x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} e^{ikn} W_n(x) \tag{78}$$

converges in the $L^2([-\frac{1}{2}, \frac{1}{2}])$ norm for all $k = \kappa + iq$, $|q| < \Gamma$. The convergence is uniform with respect to k. In the same way we can prove the existence of the derivative $(\partial/\partial k)\Phi(k,x)$. Thus Φ is an analytic vector-valued function of k. Periodicity is obvious.

VIII. CONCLUSIONS

We have discussed Bloch and Wannier functions composed of Gaussian orbitals. In particular, we have compared the Bloch and Wannier functions for two different choices of the normalization constants. We have stated explicitly the domain of analyticity of these Bloch functions [see Eq. (14)] as well as the exponential decrease of the corresponding Wannier functions. Also the expectation values of kinetic and potential energy have been studied in detail. We have proved several monotonicity properties of these expectation values and we have derived simple formulas for the limits of strongly and weakly localized Gaussian orbitals. Likewise, we have discussed in general the relationship of the analyticity properties of the Bloch functions and the localization properties of the corresponding Wannier functions.

ACKNOWLEDGMENTS

Two of us (P.Z. and R.D.) are very grateful to Professor Bozejko for his extremely useful information about the theta functions. The present work was financially supported by Österreichische Nationalbank under Grant No. 5917.

APPENDIX: MONOTONICITY PROPERTIES OF ϑ-FUNCTIONS

Here we prove some monotonicity properties. In the following we always assume that τ is purely imaginary with Im $\tau>0$ and $z=\xi$ is real. In addition $q=e^{i\pi\tau}$.

First we prove that $\theta_3(\xi|\tau)$ is monotone decreasing (with respect to ξ) for $\xi\in[0,\pi/2]$. This holds true if and only if $\ln\theta_3(\xi|\tau)$ is monotone decreasing for $\xi\in[0,\pi/2]$. Using the product expansion for $\theta_3(\xi|\tau)$ one has[10–12]

$$\frac{\theta_3'(\xi|\tau)}{\theta_3(\xi|\tau)}=-4\sin(2\xi)\sum_{n=1}^{\infty}\frac{q^{2n-1}}{1+2q^{2n-1}\cos(2\xi)+q^{4n-2}}, \tag{A1}$$

which is obviously negative for $\xi\in[0,\pi/2]$. Here $\theta_3'(\xi|\tau)$ means $(\partial/\partial\xi)\theta_3(\xi|\tau)$. The proof that $\theta_4(\xi|\tau)$ is monotone increasing in the same interval is completely analogous.

Also the following monotonicity property is derived by using the product expansion of $\theta_3(\xi|\tau)$. We prove that $i(\partial/\partial\tau)\ln\theta_3(\xi|\tau)$ is a monotone increasing function of ξ for $\xi\in[0,\pi/2]$. A short calculation gives

$$i\,\frac{\partial}{\partial\xi}\frac{\partial}{\partial\tau}\ln\theta_3(\xi|\tau)=2\pi\,\sin(2\xi)\sum_{n=1}^{\infty}\frac{(2n-1)\sinh(-i\pi\tau(2n-1))}{[\cosh(-i\pi\tau(2n-1))+\cos(2\xi)]^2} \tag{A2}$$

from which the assertion follows immediately.

[1] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[2] F. Bloch, Z. Phys. **52**, 555 (1928).
[3] E. E. Wittmer and L. Rosenfeld, Z. Phys. **48**, 530 (1928).
[4] G. H. Wannier, Phys. Rev. **52**, 191 (1937).
[5] G. H. Wannier, J. Math. Phys. **19**, 131 (1978).
[6] W. Kohn, Phys. Rev. **115**, 809 (1959).
[7] J. Cloizeaux, Phys. Rev. A **135**, 685 (1964); **135**, 698 (1964).
[8] G. Nenciu, Commun. Math. Phys. **91**, 81 (1983).
[9] H. Bacry, L. Michel, and J. Zak, Phys. Rev. Lett. **61**, 1005 (1988).
[10] D. F. Lawden, *Elliptic Functions and Applications* (Springer, New York, 1989).
[11] H. Rademacher, *Topics in Analytic Number Theory* (Springer, Berlin, 1973).
[12] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1962).
[13] J. Zak, Phys. Rev. B **26**, 3010 (1982); H. Bacry, L. Michel, and J. Zak, *Lecture Notes in Physics* (Springer, Berlin, 1988), Vol. 313, p. 291; L. Michel and J. Zak, Europhys. Lett. **18**, 239 (1992), and references therein.
[14] J. Zak, Phys. Rev. Lett. **62**, 2747 (1989); Europhys. Lett. **9**, 615 (1989).
[15] P. Zeiner, R. Dirl, and B. L. Davies, Phys. Rev. B **54**, 2466 (1996); **54** 16646 (1996).
[16] J. Zak, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1972), Vol. 27.
[17] Square integrable with respect to the unit cell $[-\frac{1}{2},\frac{1}{2}]$.
[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol. 1: Functional Analysis* (Academic, New York, 1970).
[19] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups* (American Mathematical Society, Providence, RI, 1957).
[20] Likewise one could define analyticity by the requirement that the limit exists in the weak sense, but one can prove that both definitions are equivalent. See Refs. 18, 19.

Erratum: "A new Hamilton operator for a massive relativistic particle with spin one in a generalized Heisenberg/Schrödinger picture" [J. Math. Phys. 38, 3457–3465 (1997)]

Rudolf A. Frick
Institute of Theoretical Physics, University of Cologne, D-50923 Cologne, Germany

(Received 17 November 1997; accepted for publication 19 November 1997)

[S0022-2488(98)02004-0]

Equation (2.10) should read

$$\mathbf{N}^{(0)}(\alpha,\mathbf{n}):=\alpha\mathbf{n}+(\mathbf{n}\times\mathbf{L}-\mathbf{L}\times\mathbf{n})/2.$$

In Eq. (3.5), the operator $\mathbf{N}:=\alpha\mathbf{n}+(\mathbf{n}\times\mathbf{J}+\mathbf{J}\times\mathbf{n})/2$ should read

$$\mathbf{N}:=\alpha\mathbf{n}+(\mathbf{n}\times\mathbf{J}-\mathbf{J}\times\mathbf{n})/2.$$

 © 1998 American Institute of Physics

Gamow vectors for degenerate scattering resonances

I. E. Antoniou
Institutes Internationaux de Physique et de Chemie Solvay, CP 231, ULB, Campus la Plaine, Boulevard du Triomphe, 1050 Brussels, Belgium and Theoretische Naturkunde, Free University of Brussels

M. Gadella
Institutes Internationaux de Physique et de Chemie Solvay, CP 231, ULB, Campus la Plaine, Boulevard du Triomphe, 1050 Brussels, Belgium and Departamento de Física Teórica, Facultad de Ciencias, Universidad de Valladolid, 47011 Valladolid, Spain

G. P. Pronko
Institutes Internationaux de Physique et de Chemie Solvay, CP 231, ULB, Campus la Plaine, Boulevard du Triomphe, 1050 Brussels, Belgium and Institute for High Energy Physics, Protvino, 142284 Moscow Region, Russia

(Received 5 May 1997; accepted for publication 16 December 1997)

In this paper we construct Gamow vectors for resonances given by poles of the analytic continuation of the S matrix of any finite order. We study their modes of decay (or growth). We obtain Jordan block structures for the extended Hamiltonians and evolution operators on the subspaces spanned by these Gamow vectors. We perform this study within the context of the rigged Hilbert space extension of quantum theory. We construct an explicit Friedrichs model with a double pole resonance to illustrate the general formulation. © *1998 American Institute of Physics.* [S0022-2488(98)03204-6]

I. INTRODUCTION

Resonances appear in quantum scattering when the S-matrix in the energy representation has singularities in the complex plane. These singularities are the scattering resonances.[1–3] In the case of simple poles,[3–5] the real part of the resonance corresponds to the resonance energy while the imaginary part corresponds to the half-width of the cross section, which is twice the mean life of the corresponding unstable state.

In order to describe the unstable quantum states, Gamow conjectured[6] the existence of eigenvectors, henceforth called *Gamow vectors*, of the Hamiltonian with complex eigenvalues. These eigenvalues were precisely the points at which the singularities of the S-matrix appear. However, this is impossible within the conventional formulation of quantum theory in Hilbert space[7,8] because the Hamiltonian (being a self-adjoint operator) can only have real eigenvalues. Despite this difficulty inherent to the Hilbert space formulation, formal constructions of Gamow vectors appeared early in the literature.[9–12]

In the case of simple scattering resonances, Bohm and Gadella constructed systematically the corresponding Gamow vectors and they provided a clear mathematical meaning to these vectors in terms of a suitable extension of quantum theory on the basis of Gel'fand triplets,[13] also called rigged Hilbert spaces.[3] These three spaces are a dual pair of locally convex topological spaces[14] plus a Hilbert space which contains one of the spaces of the dual pair and is contained in the other. The Bohm–Gadella construction is based on Hardy functions[15–17] and is systematically presented in Ref. 4 where one can find the previous references as well. The extended Schrödinger evolution is intrinsically irreversible because it splits into two distinct semigroups.[4] This result was, moreover, a first step towards a rigorous treatment of the problem of irreversibility.[18–25] In fact, Bohm conjectured[5] that extensions to Gel'fand triplets may give meaning to the Brussels–Austin formalism of irreversibility developed under the direction of Prigogine.[18–21] This conjecture was verified[22] for the Friedrichs model. The Bohm–Gadella Gamow vectors follow, in fact, from the intrinsic properties of the extended evolution operators because they actually correspond to the generalized eigenprojection constructed from subdynamics.[23]

This remark is, however, much more general. For both classical and quantum unstable sys-

 © 1998 American Institute of Physics

tems independent of scattering, there exist intrinsically irreversible extensions of dynamics to suitable Gel'fand triplets which include the resonances as generalized eigenvalues.[23]

For classical chaotic systems like the baker transformation, we find poles of all multiplicities.[23–25] These degenerate poles give rise to Jordan blocks with a size equal to the order of the pole. Therefore, one may ask whether such poles arise in unstable quantum systems as well. In the case of a spherically symmetric potential, the presence of a multiple pole resonance for a given value of the angular momentum can, in principle, be tested experimentally, since its corresponding lth partial cross section no longer has the Breit–Wigner shape. For instance, if for a certain value of l the resonance is described by a pair of double poles, the corresponding partial cross section should include, in addition to the usual Breit–Wigner component, a term of the form,[3]

$$\frac{4\pi(2l+1)}{p^2}\,\frac{\Gamma^2(E-E_0)^2}{[(E-E_0)^2-\Gamma^2/4]+\Gamma^2(E-E_0)^2}. \tag{1.1}$$

Here, the poles are located at $E_0 \pm i\Gamma/2$.

Such possibilities for decay are already explicitly mentioned in the classical paper by Goldberger and Watson.[26] Double pole resonances have already been considered in the literature. Lukierski[27,28] has studied a case of double pole resonances in field theory that appears in the interaction of two scalar fields. Furthermore, he has constructed a state operator for this resonance and shown that the rate of decay is not purely exponential, but instead a polynomial factor appears,

$$(\alpha t^2+\beta t+\gamma)e^{-\Gamma t}, \tag{1.2}$$

where α, β, γ and $\Gamma \geqslant 0$ are constants.

Katznelson[29,30] studied systematically the Jordan blocks associated to resonances of the partial resolvent using complex scaling methods.[31,32] Brändas, Chatzidimitriou-Dreismann and co-workers[33–37] combined complex scaling techniques with perturbation theory to discuss resonances and Jordan block structures in the Hamiltonian,[33,34] in density matrices,[35–37] and associated with multiple poles in the resolvent in Ref. 37. Also, Mondragon and Hernandez[38] discovered these structures in their investigation of the geometric phases for degenerate resonances. Sudarshan and Bhamati[39] also discuss Jordan blocks corresponding to the Friedrichs, the Cascade, and related models assuming the existence of multiple poles in their analytic continuation of dynamics.

The purpose of this paper is to construct Gamow vectors for scattering resonances of any finite order. After a brief discussion of simple resonance scattering in Sec. II, where we introduce the terminology and review the Bohm–Gadella work, we define, in Sec. III, the Gamow vectors when the scattering matrix has a pole of finite order m. The Jordan block structure of the extended irreversible evolution of the Gamow vectors is presented in Sec. IV. A simple Friedrichs model with doubly degenerate scattering resonance is presented in Sec. V. We finish the paper with an Appendix in which three mathematical results are proven.

II. SIMPLE RESONANCE SCATTERING

In this section we intend to present a brief summary of what is called resonance scattering for simple pole resonances. We shall also establish the basic notation.

In the conventional scattering theory,[2,8,32,40,41] the ''free'' Hamiltonian H_0 and the perturbed Hamiltonian $H=H_0+V$ satisfy the asymptotic condition. Therefore, the Møller wave operators $\mathbf{\Omega}_{\mathrm{IN}}, \mathbf{\Omega}_{\mathrm{OUT}}$ exist and provide the unitary equivalence of the absolutely continuous subspaces $\mathscr{H}_{\mathrm{ac}}(H_0)$ and $\mathscr{H}_{\mathrm{ac}}(H)$ associated with the operators H_0 and H, respectively. To simplify the notation without loss of the essence of our argument, we assume that the absolutely continuous spectrum of H_0 as well as that of H is simple and covers the positive real semiaxis $\mathbb{R}^+ \equiv [0,\infty)$.

Let U be the unitary transformation which provides the spectral representation of the absolutely continuous part of H_0 (restriction of H_0 to $\mathscr{H}_{\mathrm{ac}}(H_0)$) as the multiplication operator on $L^2(\mathbb{R}^+)$. Then, we have

$$[UH_0U^{-1}](E)=E\varphi(E), \tag{2.1}$$

$$\mathscr{H}_{\rm ac}(H) \xrightarrow{(\Omega_{\rm IN})^{-1}} \mathscr{H}_{\rm ac}(H_0) \overset{U}{\to} L^2(\mathbb{R}^+), \tag{2.2}$$

$$\mathscr{H}_{\rm ac}(H) \xrightarrow{(\Omega_{\rm OUT})^{-1}} \mathscr{H}_{\rm ac}(H_0) \overset{U}{\to} L^2(\mathbb{R}^+). \tag{2.3}$$

Now, we shall briefly present here the rigged Hilbert spaces (RHS) of Hardy functions. To construct them, consider the set of spaces $\mathscr{H}_{\pm}^2 \cap S$, where $\mathscr{H}_{\pm}^2$ are the spaces of Hardy functions on the upper and the lower half-plane[15–17] and S the Schwartz space of all functions admitting derivatives at all orders such that they and their derivatives go to zero faster than the inverse of any polynomial at the infinity.[42] The restriction of any function in $\mathscr{H}_{\pm}^2 \cap S$ to $\mathbb{R}^+$ uniquely determines its values on the whole real line[43] and, as a consequence of the Tichmarsh theorem, on the corresponding half-plane. These restrictions are dense subspaces of $L^2(\mathbb{R}^+)$ (Ref. 44) and determine two one-to-one onto maps,

$$\theta_{\pm} : \mathscr{H}_{\pm}^2 \cap S \mapsto \mathscr{H}_{\pm}^2 \cap S\big|_{\mathbb{R}^+} \equiv \Delta_{\pm}. \tag{2.4}$$

As $\mathscr{H}_{\pm}^2 \cap S$ are Fréchet nuclear spaces,[13,14] this structure is transmitted to $\Delta_{\pm}$ via $\theta_{\pm}$. Also, the identity map $I: \Delta_{\pm} \mapsto L^2(\mathbb{R}^+)$ is continuous. Therefore, if ${}^{\times}(\Delta_{\pm})$ represent the respective antidual spaces of $\Delta_{\pm}$ (the spaces of all continuous *antilinear* functionals on $\Delta_{\pm}$), the triplets,

$$\Delta_{\pm} \subset L^2(\mathbb{R}^+) \subset {}^{\times}\Delta_{\pm} \tag{2.5}$$

are Gel'fand triplets or rigged Hilbert spaces (RHS).[4]

The unitary transformations U, $\mathbf{\Omega}_{\rm IN}$, and $\mathbf{\Omega}_{\rm OUT}$ allow us to transfer the Gel'fand triplet (2.5) from the spectral representation to the original representation. Therefore, the spaces denoted by

$$\mathbf{\Phi}_+ = \mathbf{\Omega}_{\rm OUT} U^{-1} \Delta_+; \quad \mathbf{\Phi}_- = \mathbf{\Omega}_{\rm IN} U^{-1} \Delta_- \tag{2.6}$$

produce the triplets

$$\mathbf{\Phi}_{\pm} \subset \mathscr{H}_{\rm ac}(H) \subset {}^{\times}\mathbf{\Phi}_{\pm}. \tag{2.7}$$

The extensions of the transformations U, $\mathbf{\Omega}_{\rm IN}$ and $\mathbf{\Omega}_{\rm OUT}$ by duality give rise to the following formulas for the duals ${}^{\times}\mathbf{\Phi}_{\pm}$:[4]

$${}^{\times}\mathbf{\Phi}_+ = \mathbf{\Omega}_{\rm OUT} U^{-1\,\times}\Delta_+; \quad {}^{\times}\mathbf{\Phi}_- = \mathbf{\Omega}_{\rm IN} U^{-1\,\times}\Delta_-. \tag{2.8}$$

From (2.4) and (2.6), we have[4]

$$\mathbf{\Phi}_+ = \mathbf{\Omega}_{\rm OUT} U^{-1} \theta_+ [\mathscr{H}_+^2 \cap S]; \quad \mathbf{\Phi}_- = \mathbf{\Omega}_{\rm IN} U^{-1} \theta_- [\mathscr{H}_-^2 \cap S]. \tag{2.9}$$

We denote by $\phi_{\pm}$ the vectors in $\mathbf{\Phi}_{\pm}$ which correspond to the functions $\varphi_{\pm}(E)$ in $\mathscr{H}_{\pm}^2 \cap S$,

$$\phi_+ = \mathbf{\Omega}_{\rm OUT} U^{-1} \theta_+ \varphi_+; \quad \phi_- = \mathbf{\Omega}_{\rm IN} U^{-1} \theta_- \varphi_-. \tag{2.10}$$

The nuclear spectral theorem now establishes that for almost all E in the spectrum of H which is equal to $\mathbb{R}^+$ in our case, there exist functionals $|E_{\pm}\rangle \in {}^{\times}\mathbf{\Phi}_{\pm}$ such that

(i) They are (generalized) eigenvectors of H with eigenvalue E, i.e.,

$$H|E_{\pm}\rangle = E|E_{\pm}\rangle. \tag{2.11}$$

(i) For $E>0$, the functions $\varphi_{\pm}(E)$ are given by $\langle E_{\pm}|\phi_{\pm}\rangle = \langle \phi_{\pm}|E_{\pm}\rangle^*$.

Due to the regularity properties of the functions in $\mathscr{H}_{\pm}^2 \cap S$ ''almost all'' can be replaced by ''all,'' so that (2.11) is indeed true for any $E>0$. We would like to mention here in passing that this is in general one of the advantages of going beyond the Hilbert space formalism.[3]

Summarizing, the correspondence given in (2.6) can be written in the following form:

$$\phi_- \mapsto \varphi_-(E) = \langle E_- | \phi_- \rangle \in \Delta_- \Leftrightarrow \varphi_-^*(E) = \langle \phi_- | E_- \rangle \in \Delta_+ ; \quad E > 0, \tag{2.12}$$

$$\phi_+ \mapsto \varphi_+(E) = \langle E_+ | \phi_+ \rangle \in \Delta_+ \Leftrightarrow \varphi_+^*(E) = \langle \phi_+ | E_+ \rangle \in \Delta_- ; \quad E > 0. \tag{2.13}$$

Here, the symbol $*$ denotes complex conjugation.

Now, define

$$\mathbb{C}^+ = \{z \in \mathbb{C}; \quad \text{Im } z > 0\}; \quad \mathbb{C}^- = \{z \in \mathbb{C}; \quad \text{Im } z < 0\}, \tag{2.14}$$

i.e., the upper and lower open half-plane. Let $\varphi_\pm(z)$, $z \in \mathbb{C}^\pm$, be the (unique) analytic functions for which $\varphi_\pm(E) \in \Delta_\pm$, $E \in \mathbb{R}^+$, is the corresponding boundary function on the positive real semi-axis. Then, for any given $z_0 \in \mathbb{C}^-$, $z_0^* \in \mathbb{C}^+$, the following maps:

$$\phi_+ \mapsto \varphi_+(z_0^*); \quad \phi_- \mapsto \varphi_-(z_0) \tag{2.15}$$

define respective continuous linear functionals on $\mathbf{\Phi}_+$ and $\mathbf{\Phi}_-$ that we denote by $\langle z_{0+}^* |$ and $\langle z_{0-} |$. With this notation, we have

$$\langle z_{0+}^* | \phi_+ \rangle = \varphi_+(z_0^*); \quad \langle z_{0-} | \phi_- \rangle = \varphi_-(z_0). \tag{2.16}$$

Analogously, the maps

$$\phi_+ \mapsto \varphi_+^*(z_0); \quad \phi_- \mapsto \varphi_-^*(z_0^*) \tag{2.17}$$

define respective continuous *antilinear* functionals on $\mathbf{\Phi}_+$ and $\mathbf{\Phi}_-$ which we denote by $|z_{0+}\rangle$ and $|z_{0-}^*\rangle$. With this notation, which is essentially Dirac[45] notation, we have

$$\langle \phi_+ | z_{0+} \rangle = \varphi_+^*(z_0); \quad \langle \phi_- | z_{0-}^* \rangle = \varphi_-^*(z_0^*). \tag{2.18}$$

Among the functionals (2.15) [or (2.16)] and (2.17) [or (2.18)], there exist the following relations:

$$\langle z_{0-} | \phi_- \rangle^* = \langle \phi_- | z_0^* \rangle; \quad \langle z_{0+}^* | \phi_+ \rangle^* = \langle \phi_+ | z_{0+} \rangle. \tag{2.19}$$

To prove these relations, we use the Tichmarsh theorem.[15–17] For the first one, we have

$$\langle z_{0-} | \phi_- \rangle^* = [\varphi_-(z_0)]^* = \left[\frac{-1}{2\pi i} \int_{-\infty}^{\infty} \frac{\varphi_-(E) dE}{E - z_0} \right]^* = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\varphi_-^*(E) dE}{E - z_0^*} = \varphi_-^*(z_0^*)$$
$$= \langle \phi_- | z_{0-}^* \rangle. \tag{2.20}$$

The second one is proven analogously.

Next, we present a very important result.

Proposition: The antilinear functionals $|z_{0+}\rangle$ and $|z_{0-}^*\rangle$ are generalized right eigenvectors of H and the linear functionals $\langle z_{0-}^* |$ and $\langle z_{0-} |$ are generalized left eigenvectors of H, with eigenvalues z_0 and z_0^* respectively, i.e.,

$$H|z_{0+}\rangle = z_0 |z_{0+}\rangle; \quad H|z_{0-}^*\rangle = z_0^* |z_{0-}^*\rangle, \tag{2.21}$$

$$\langle z_{0+}^* | H = z_0 \langle z_{0+}^* |; \quad \langle z_{0-} | H = z_0^* \langle z_{0-} |. \tag{2.22}$$

A proof for (2.21) is given in Ref. 4 following the Tichmarsh theorem. The first identity in (2.22), follows from the first identity in (2.21),

$$z_0^* \langle z_{0+}^* | \phi_+ \rangle = (z_0 \langle \phi_+ | z_{0+} \rangle)^* = \langle \phi_+ | H | z_{0+} \rangle^* = \langle H \phi_+ | z_{0+} \rangle^* = \langle z_{0+}^* | H | \phi_+ \rangle. \tag{2.23}$$

By omitting the arbitrary ϕ_+ in (2.23), we obtain the desired result. Analogously, the second identity in (2.23) is a consequence of the first equation in (2.21). The second formula in (2.22) is obtained in a similar way.

The relevant eigenvectors for resonance scattering are those coinciding with the isolated singularities of the continuation of the analytic S-matrix, $S(E)$, $E\in\mathbb{R}^+$ in the energy representation. Simple scattering resonances are understood[1–3] as simple poles of the continued S-matrix from the semiaxis to the complex plane. In the language of Riemann surfaces, these poles are located in the second sheet.

For any simple pole

$$z_R=E_R-i\,\frac{\Gamma}{2},\ \ \Gamma>0$$

of the continued S-matrix from the upper to the lower half plane, the vectors $|z_{R+}\rangle$ and $|z_{R-}^*\rangle$ are *the decaying and the growing Gamow vectors*, respectively.[3,4]

Following the right/left eigenvector notation,[22,23] we can write the Gamow vectors as follows:

$$|f_0\rangle=|z_{R+}\rangle;\quad |\widetilde{f}_0\rangle=|z_{R-}^*\rangle, \tag{2.24}$$

$$\langle f_0|=\langle z_{R+}^*|;\quad \langle\widetilde{f}_0|=\langle z_{R-}|. \tag{2.25}$$

These forms provide the relation between the Bohm–Gadella approach and the Brussels–Austin approach.

III. DEGENERATE RESONANCE SCATTERING

A scattering resonance is degenerate when the continuation of the S-matrix, $S(E)$, $E>0$, from above to below has a pole of order m at the point $z_R=E_R-i\Gamma/2$. Correspondingly, the continuation from below to above has a pole of order m at $z_R^*=E_R+i\Gamma/2$. In this case, the construction of the Gamow vectors is a generalization of the construction for the simple pole presented in Sec. II.

Let

$$\varphi_\pm^{(n}=\frac{d^n}{dz^n}\,\varphi_\pm(z);\ \ n=0,1,2,..., \tag{3.1}$$

denote the nth derivative of the analytic function $\varphi_\pm(z)$ for which $\varphi_\pm(E)\in\Delta_\pm$ is the boundary function on the positive real semiaxis. Then, the maps

$$\phi_+\mapsto[\varphi_+^{(n}(z)]_{z=z_R^*};\ \ n=0,1,...,m-1 \tag{3.2}$$

define m linear functionals on Φ_+, which we denote by $\langle f_n|$; $n=0,1,...,m-1$. With this notation, we have,

$$\langle f_n|\phi_+\rangle=[\varphi_+^{(n}(z)]_{z=z_R^*};\ \ n=0,1,...,m-1. \tag{3.3}$$

For z_R, the maps

$$\phi_-\mapsto[\varphi_-^{(n}(z)]_{z=z_R};\ \ n=0,1,...,m-1 \tag{3.4}$$

define m linear functionals on $\mathbf{\Phi}_-$ which we denote by $\langle\widetilde{f}_n|$, $n=0,1,...,m-1$. With this notation, we have

$$\langle\widetilde{f}_n|\phi_-\rangle=[\varphi_-^{(n}(z)]_{z=z_R};\ \ n=0,1,...,m-1. \tag{3.5}$$

The corresponding antilinear functionals are defined as in the case of a simple pole,

$$\langle\phi_+|f_n\rangle=\langle f_n|\phi_+\rangle^*=[\varphi_+^{*(n}(z)]_{z=z_R}, \tag{3.6}$$

$$\langle\phi_-|\widetilde{f}_n\rangle=\langle\widetilde{f}_n|\phi_-\rangle^*=[\varphi_-^{*(n}(z)]_{z=z_R^*}. \tag{3.7}$$

We can easily see that the linear functionals $\langle f_n|$ and $\langle \widetilde{f}_n|$, $n=0,1,...,m-1$ are continuous. The proof goes as follows: Since the topology on $\mathbf{\Phi}_+$ has been constructed by transporting by $\mathbf{\Omega}_{\rm OUT}U^{-1}\theta_+$ the topology on $\mathscr{H}^2_+\cap S$ and the latter has been inhereted from the topology of the Schwartz space, the general theory of locally convex topological vector spaces,[14,42] shows that, in order to prove the continuity of $\langle f_n|$, it is sufficient to find a positive constant $C>0$ such that

$$|\langle f_n|\phi_+\rangle|\leqslant C\|\varphi_+(E)\|;\ \ \forall\phi_+\in\mathbf{\Phi}_+\,, \tag{3.8}$$

where $\varphi_+(E)$ is the function in $\mathscr{H}^2_+\cap S$ associated to ϕ_+ by the first formula in (2.10). In (3.8), $\|\varphi_+(E)\|$ denotes the $L^2(\mathbb{R})$ norm of $\varphi_+(E)$, i.e.,

$$\|\varphi_+(E)\|=\left[\int_{-\infty}^{\infty}|\varphi_+(E)|^2dE\right]^{1/2}, \tag{3.9}$$

which is one of the countable set of norms giving the topology on $\mathscr{H}^2_+\cap S$.

Analogously, the continuity of the linear functional $\langle \widetilde{f}_n|$ is assured if we can find a $C'>0$ such that

$$|\langle \widetilde{f}_n|\phi_-\rangle|\leqslant C'\|\varphi_-(E)\|;\ \ \forall\phi_-\in\mathbf{\Phi}_-\,, \tag{3.10}$$

where $\varphi_-(E)$ is the function in $\mathscr{H}^2_-\cap S$ associated to $\mathbf{\Phi}_-$ by the second formula in (2.10) Again, the norm in (3.10) is the usual $L^2(\mathbb{R})$ norm.

We shall prove here only (3.8). The proof for (3.10) is the same. A generalization of the Tichmarsh theorem (see Appendix) asserts that

$$\langle f_n|\phi_+\rangle=[\varphi_+^{(n)}(z)]_{z=z_R^*}=\frac{n!}{2\pi i}\int_{-\infty}^{\infty}\frac{\varphi_+(E)dE}{(E-z_R^*)^{n+1}}. \tag{3.11}$$

We want to recall that $\varphi_+(E)$ represents the function in $\mathscr{H}^2_+\cap S$ given by the boundary vales of $\varphi_+(z)$ on the real axis. The Cauchy–Schwarz inequality on $L^2(\mathbb{R})$ can be applied to the integral in (3.11) to give

$$|\langle f_n|\phi_+\rangle|\leqslant\frac{n!}{2\pi}\left\|\frac{1}{(E-z_R^*)^{n+1}}\right\|\|\varphi_+(E)\|, \tag{3.12}$$

where the norms in (3.12) are the standard $L^2(\mathbb{R})$ norms. Formula (3.12) shows that

$$\langle f_n|\phi_+\rangle\leqslant C\|\varphi_+(E)\|\ \ \text{with}\ \ C=\frac{n!}{2\pi}\left\|\frac{1}{(E-z_R^*)^{n+1}}\right\| \tag{3.13}$$

which proves the continuity of $\langle f_n|$.

In the previous section, following Refs. 3 and 4, we have called Gamow vectors the functionals $|f_0\rangle$, $|\widetilde{f}_0\rangle$, $\langle f_0|$, $\langle\widetilde{f}_0|$. We shall also call Gamow vectors the functionals $|f_n\rangle$, $|\widetilde{f}_n\rangle$, $\langle f_n|$, $\langle\widetilde{f}_n|$; $n=0,1,...,m-1$ corresponding to a resonance pole of order m. As in the simple pole case, the vectors $|f_0\rangle$ and $\langle\widetilde{f}_0|$ are generalized right and left eigenvectors of H with eigenvalue z_R and $|\widetilde{f}_0\rangle$ and $\langle f_0|$ are right and left eigenvectors of H with eigenvalue z_R^*, so that

$$H|f_0\rangle=z_R|f_0\rangle;\ \ \langle f_0|H=z_R^*\langle f_0|, \tag{3.14}$$

$$H|\widetilde{f}_0\rangle=z_R^*|\widetilde{f}_0\rangle;\ \ \langle\widetilde{f}_0|H=z_R\langle\widetilde{f}_0|. \tag{3.15}$$

For the vectors $|f_n\rangle$, $|\widetilde{f}_n\rangle$, $\langle f_n|$, $\langle\widetilde{f}_n|$; $n=0,1,...,m-1$, we can prove the following formulas:

$$H|f_n\rangle=z_R|f_n\rangle+n|f_{n-1}\rangle;\ \ \langle f_n|H=z_R^*\langle f_n|+n\langle f_{n-1}|, \tag{3.16}$$

$$H|\widetilde{f}_n\rangle=z_R^*|\widetilde{f}_n\rangle+n|\widetilde{f}_{n-1}\rangle;\ \ \langle\widetilde{f}_n|H=z_R\langle\widetilde{f}_n|+n\langle\widetilde{f}_{n-1}|. \tag{3.17}$$

The proof of (3.16) goes as follows: Since we have that, on their respective absolutely continuous subspaces,[8,42] $\mathbf{\Omega}_{\mathrm{OUT}}^{\dagger} H \mathbf{\Omega}_{\mathrm{OUT}} = K$ and we also have (2.1), the function in $\mathscr{H}_+^2 \cap S$ that corresponds to $H\phi_+$ by the first formula in (2.10) is $E\varphi_+(E)$. This function represents the set of boundary values of the complex analytic function $z\varphi_+(z)$ on the upper half-plane. As $\varphi_+(E)$ is in $\mathscr{H}_+^2 \cap S$, one has

$$E\varphi_+(E) \in \mathscr{H}_+^2 \cap S; \quad \frac{d^n}{dE^n}[E\varphi_+(E)] \in \mathscr{H}_+^2 \cap S. \tag{3.18}$$

Furthermore, we can show (see the Appendix) that $(d^n/dE^n)[E\varphi_+(E)]$, $E \in \mathbb{R}$, represent the set of boundary values of the analytic function $(d^n/dz^n)[z\varphi_+(z)]$ on the upper half-plane. With these ideas in mind, we use definition (3.3) and the Titchmarsh theorem to write

$$\begin{aligned}
\langle f_n | H\phi_+ \rangle &= [(z\varphi_+(z))^{(n)}]_{z=z_R^*} \\
&= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{(E\varphi_+(E))^{(n)}}{E - z_R^*}\, dE \\
&= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{E\varphi_+^{(n)}(E)}{E - z_R^*}\, dE + \frac{n}{2\pi i} \int_{-\infty}^{\infty} \frac{\varphi_+^{(n-1)}(E) dE}{E - z_R^*} \\
&= z_R^* [\varphi_+^{(n)}(z)]_{z=z_R^*} + n[\varphi_+^{(n-1)}(z)]_{z=z_R^*} \\
&= z_R^* \langle f_n | \phi^+ \rangle + n \langle f_{n-1} | \phi_+ \rangle.
\end{aligned} \tag{3.19}$$

Since ϕ_+ is an arbitrary element of $\mathbf{\Phi}_+$, (3.19) gives the second formula in (3.16). To obtain the first formula in (3.16), we use the complex conjugate of (3.19),

$$\langle H\phi_- | f_n \rangle = \langle \phi_+ | H | f_n \rangle = z_R \langle \phi_+ | f_n \rangle + n \langle \phi_+ | f_{n-1} \rangle. \tag{3.20}$$

To prove (3.17), we proceed analogously. Formulas (3.16) show that, in the subspace of ${}^{\times}\mathbf{\Phi}_+$ spanned by the vectors $|f_0\rangle$, $|f_1\rangle, \ldots, 1/(m-2)|f_{m-2}\rangle$, $1/(m-1)|f_{m-1}\rangle$, the extended Hamiltonian H is represented as an upper Jordan block of order m,

$$\begin{bmatrix} z_R & 1 & 0 & \cdots & 0 \\ 0 & z_R & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & z_R \end{bmatrix} \tag{3.21}$$

and that in the subspace of $\mathbf{\Phi}_-^{\times}$ (the space of continuous *linear* functionals on $\mathbf{\Phi}_-$) spanned by $\langle f_0|, \langle f_1|, \ldots, 1/(m-2)\langle f_{m-2}|, 1/(m-1)\langle f_{m-1}|$, H is also represented as an upper Jordan block of order m,

$$\begin{bmatrix} z_R^* & 1 & 0 & \cdots & 0 \\ 0 & z_R^* & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & z_R^* \end{bmatrix}. \tag{3.22}$$

Similar results follow from (3.17).

IV. THE IRREVERSIBLE EVOLUTION OF GAMOW VECTORS

Along the present section, we shall discuss the time evolution of Gamow vectors for degenerate resonances of any order. As the Gamow vectors live in the larger spaces ${}^{\times}\mathbf{\Phi}_+$ and ${}^{\times}\mathbf{\Phi}_-$, we have to extend the unitary evolution $U_t = e^{-itH}$ from the Hilbert space $\mathscr{H}$ to the spaces ${}^{\times}\mathbf{\Phi}_+$ and ${}^{\times}\mathbf{\Phi}_-$.

In the case of a simple scattering resonance, the extensions U_t^+ and U_t^- of U_t are given by the duality formulas,[3,4]

$$\langle \phi_+ | U_t^+ F\rangle = \langle U_t^\dagger \phi_+ | F\rangle \quad \text{for} \quad t>0, \tag{4.1}$$

$$\langle \phi_- | U_t^- \widetilde{F}\rangle = \langle U_t^\dagger \phi_- | \widetilde{F}\rangle \quad \text{for} \quad t<0, \tag{4.2}$$

where $U_t^\dagger = e^{itH}$ is the Hilbert space adjoint of U_t. Here, F and $\widetilde{F}$ are arbitrary elements of ${}^\times\mathbf{\Phi}_+$ and ${}^\times\mathbf{\Phi}_-$ and the test vectors ϕ_+ and ϕ_- are arbitrarily chosen in $\mathbf{\Phi}_+$ and $\mathbf{\Phi}_-$, respectively. The families U_t^+ for $t>0$ and U_t^- for $t<0$ define two semigroups on the topological vector spaces ${}^\times\mathbf{\Phi}_+$ and ${}^\times\mathbf{\Phi}_-$, respectively.

The action of the semigroups U_t^+ and U_t^- on the Gamow vectors is

$$U_t^+ |f_0\rangle = e^{-iz_R t}|f_0\rangle; \quad t>0, \tag{4.3}$$

$$U_t^- |\widetilde{f}_0\rangle = e^{-iz_R^* t}|\widetilde{f}_0\rangle; \quad t<0. \tag{4.4}$$

In the case of a degenerate scattering resonance associated to a pole of order m, the extensions U_t^+ and U_t^- are also defined by formulas (4.1) and (4.2). The action of the semigroups U_t^+ and U_t^- on the Gamow vectors $|f_n\rangle$, $|\widetilde{f}_n\rangle$, $n=0,1,...,m-1$, is

$$U_t^+ |f_n\rangle = e^{-iz_R t}\sum_{l=0}^{n}\binom{n}{l}(-it)^l |f_{n-l}\rangle; \quad t>0, \tag{4.5}$$

$$U_t^- |\widetilde{f}_n\rangle = e^{-iz_R^* t}\sum_{l=0}^{n}\binom{n}{l}(-it)^l |\widetilde{f}_{n-l}\rangle; \quad t<0, \tag{4.6}$$

where

$$\binom{n}{l} = \frac{n!}{l!(n-l)!}$$

are the Binomial coefficients. Note that for $n=0$, the formulas for degenerate resonances (4.5) and (4.6) coincide with (4.3) and (4.4), respectively. Let us give the proof for (4.5) and proceed by analogy to show (4.6). First of all, one can easily show that for $t>0$ the function in $\mathscr{H}_+^2 \cap S$ that corresponds to

$$U_t^\dagger \phi_+ = e^{itH}\phi_+ \tag{4.7}$$

by the first identity in (2.10) is given by

$$e^{itE}\varphi_+(E), \tag{4.8}$$

where $\varphi_+(E)$ is the function in $\mathscr{H}_+^2 \cap S$ that corresponds to ϕ_+ by the same formula. Then, we write for $t>0$,

$$\langle f_n | U_t^\dagger \phi_+\rangle = \frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{\frac{d^n}{dE^n}\{e^{itE}\varphi_+(E)\}}{E-z_R^*}\,dE = \frac{1}{2\pi i}\sum_{l=0}^{n}\binom{n}{l}(it)^l\int_{-\infty}^{\infty}\frac{e^{itE}\,\frac{d^{n-1}}{dE^{n-1}}\,\varphi(E)}{E-z_R^*}\,dE$$

$$= \sum_{l=0}^{n}\binom{n}{l}(it)^l e^{itz_R^*}\langle f_n|\phi_+\rangle. \tag{4.9}$$

In the first and third identities in (4.9), we have applied the Titchmarsh theorem,[15–17] along with the fact that, for $t>0$, the product of e^{itE} with any Hardy function on the upper half-plane, gives again a Hardy function on the upper half-plane. The complex conjugate of (4.9) gives

$$\langle U_t^\dagger \phi_+ | f_n\rangle = \langle \phi_+ | U_t^+ | f_n\rangle = e^{-itz_R}\sum_{l=0}^{n}\binom{n}{l}(-it)^l\langle \phi_+ | f_n\rangle \tag{4.10}$$

which is valid for $t>0$. Taking account of the arbitrariness of the test vector ϕ_+ in $\mathbf{\Phi}_+$, we directly obtain (4.5) from (4.10).

In the subspace of ${}^\times\mathbf{\Phi}_+$ spanned by the vectors $|f_0\rangle,|f_1\rangle,\ldots,1/(m-2)|f_{m-2}\rangle$, $1/(m-1)|f_{m-1}\rangle$, the semigroup U_t^+, $t>0$, is represented by the following upper triangular matrix,

$$e^{-itz_R}\begin{bmatrix} 1 & -it & \frac{(-it)^2}{2} & \cdots & \frac{(-it)^{m-1}}{(m-1)!} \\ 0 & 1 & -it & \cdots & \frac{(-it)^{m-2}}{(m-2)!} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad t>0. \tag{4.11}$$

This matrix can also be obtained by the exponentiation of the Jordan block (3.21) multiplied by $-it$, following the standard exponentiation formula for a Jordan block.

Analogously, in the subspace of ${}^\times\mathbf{\Phi}_-$ spanned by the vectors $|\tilde{f}_0\rangle,|\tilde{f}_1\rangle,\ldots,1/(m-2)|\tilde{f}_{m-2}\rangle$, $1/(m-1)|\tilde{f}_{m-1}\rangle$, the semigroup U_t^-, $t<0$ is also represented by an upper triangular matrix, i.e.,

$$e^{-itz_R^*}\begin{bmatrix} 1 & -it & \frac{(-it)^2}{2} & \cdots & \frac{(-it)^{m-1}}{(m-1)!} \\ 0 & 1 & -it & \cdots & \frac{(-it)^{m-2}}{(m-2)!} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad t<0. \tag{4.12}$$

Again, this matrix is obtained by exponentiation of (3.22). Time evolution for continuous functionals on $\mathbf{\Phi}_+$ and $\mathbf{\Phi}_-$ is obtained by using the complex conjugate of (4.1) and (4.2). Therefore, for the right Gamow vectors $\langle f_n|$, $\langle\tilde{f}_n|$, $n=0,1,2,\ldots,m-1$, we obtain the following time evolution:

$$\langle f_n|U_t^+ = e^{iz_R^* t}\sum_{l=0}^{n}\binom{n}{l}(it)^l\langle f_{n-l}|; \quad t>0, \tag{4.13}$$

$$\langle\tilde{f}_n|U_t^- = e^{iz_R t}\sum_{l=0}^{n}\binom{n}{l}(it)^l\langle\tilde{f}_{n-l}|; \quad t<0. \tag{4.14}$$

V. A FRIEDRICHS MODEL WITH DOUBLE RESONANCE SCATTERING

Along this last section, we shall construct an explicit potential which produces a degenerate scattering resonance in the Friedrichs model. The Friedrichs model is a typical example of resonance scattering and has been extensively discussed as a model of unstable particles as well as a toy model for quantum irreversible processes.[46,12,41,47,48,19,22,23]

In the simplest version of the Friedrichs model, we have a Hamiltonian of the following form:

$$H=H_0+\lambda V \tag{5.1}$$

in which H_0 is the free or unperturbed Hamiltonian, V is the potential and λ is a coupling parameter. The spectrum of H_0 consists of a simple eigenvalue $E_1>0$ embedded in its absolutely continuous spectrum $\mathbb{R}^+=[0,\infty)$. The interaction λV ''dissolves'' the embedded eigenvalue to the absolutely continuous spectrum and the point spectrum disappears. The operators H_0 and H satisfy the asymptotic condition for scattering mentioned in Sec. II.[12,41] The singularities of the scattering matrix $S(E)$ coincide with the singularities of the partial resolvent of H with respect to the eigenspace of E_1,

$$|1\rangle\langle 1|\,\frac{1}{E-H}\,|1\rangle\langle 1|, \tag{5.2}$$

where

$$H_0|1\rangle = E_1|1\rangle. \tag{5.3}$$

The singularities of the partial resolvent (5.2) are the zeros of the function $\eta(z)$ defined as

$$\frac{1}{\eta(z)} = \langle 1| \frac{1}{z-H} |1\rangle. \tag{5.4}$$

The function $\eta(z)$ can be calculated directly[45,12,40]

$$\eta(z) = E_1 - z + \lambda^2 \int_0^\infty dE \, \frac{|V(E)|^2}{z-E}. \tag{5.5}$$

Here, $V(E) = \langle 1|V|E\rangle$ is the form factor.

As a function on z, $\eta(z)$ is analytic on the complex plane except on the real semiaxis $\mathbb{R}^+$ on which it has a branch cut. Furthermore, it does not have zeros. This property is guaranteed if

$$\eta(0) > 0. \tag{5.6}$$

Therefore, in order to find resonances, as zeros of $\eta(z)$ we have to extend the function $\eta(z)$ through the cut. The location of the zeros of this extension depends on the value of the coupling parameter. A simple resonance corresponds to a simple zero z_R, $\mathrm{Im}\, z_R < 0$, of the analytic continuation $\eta_+(z)$ of $\eta(z)$ from the upper to the lower half-plane through the cut $\mathbb{R}^+$. The complex conjugate of z_R, z_R^*, is the zero of the analytic extension $\eta_-(z)$ of $\eta(z)$ from the lower to the upper half-plane through the cut $\mathbb{R}^+$.

The existence and the properties of the zeros of $\eta_+(z)$ and $\eta_-(z)$ depend on the form factor $V(E)$. We are interested here in the possibility of having Friedrichs models with degenerate resonances. The simplest case is a pole of second order. We shall discuss here this particular example. To this end, let us choose the following form factor:

$$|V(E)|^2 = \frac{\sqrt{E}}{P(E)}; \quad 0 \leqslant E < \infty \tag{5.7}$$

with $P(E) = (E-\alpha)(E-\alpha^*)$ for some complex number α. For the square root we choose the principal branch.[49]

Using (5.7), the function $\eta(z)$ has the form,

$$\eta(z) = E_1 - z - \pi\lambda^2 \left\{ \frac{\sqrt{-z}}{P(z)} - \frac{1}{\alpha-\alpha^*} \left(\frac{\sqrt{-\alpha}}{z-\alpha} - \frac{\sqrt{-\alpha^*}}{z-\alpha^*} \right) \right\}. \tag{5.8}$$

In order to prove it, we have to calculate the integral in (5.5) using (5.6). For any $z \notin \mathbb{R}^+$, this integral can be obtain by contour integration in the complex plane. A standard formula[49] says that

$$\int_0^\infty \frac{\sqrt{E}}{P(E)(E-z)} \, dE = \frac{-\pi \, \exp\{-\pi i(1+1/2)\}}{\sin(\pi(1+1/2)} \times \sum \left\{ \mathrm{Res} \, \frac{w^{1/2}}{P(w)(w-z)} \text{ at the poles of } \frac{1}{P(w)(w-z)} \right\}. \tag{5.9}$$

The residues in (5.9) are easily obtained and (5.8) follows straightforwardly. Since we are using the principal branch of the square root, the term $\sqrt{-z}$ shows that $\eta(z)$ has a branch cut corresponding to the positive semiaxis.

A very important remark is the following: Given an analytic function on the two sheeted Riemann surface corresponding to the square root, we obtain an analytic function on the complex plane by using the transformation $z = p^2$. These two functions have the same structure, which in particular means that if the function on the complex plane has a zero at the point p_0, the function

on the Riemann surface has a zero of the same multiplicity at the point $z_0=p_0^2$ and vice versa. We prove this fact in the Appendix. Note that the correspondence between the surface and the plane is one to one.

The above remarks will enable us to obtain the resonances by using the momentum representation, i.e., by performing the following change of variables:

$$z=p^2; \quad \alpha=b^2. \tag{5.10}$$

After this change of variables, we have the following expression for $\varphi(p)=\eta(p^2)=\eta(z)$:

$$\varphi(p)=E_1-p^2+\frac{i\pi\lambda^2}{(b-b^*)(p+b)(p-b^*)}. \tag{5.11}$$

If $\varphi(p)$ has a double zero at p_R, we must have

$$\varphi(p_R)=0; \quad \varphi'(p_R)=0; \quad \varphi''(p_R)\neq 0. \tag{5.12}$$

If this were the case, then the analytic continuation of $\eta(z)$ has a double zero at $z_R=p_R^2$ and that all the zeros of this continuation are found in the same way (see the Appendix).

As we have mentioned earlier, the analytic properties of the scattering matrix in the energy representation, $S(E)$, and $\eta^{-1}(E)$ must be identical. Correspondingly, the analytic properties of the scattering matrix in the momentum representation, $S(p)$, and $\varphi^{-1}(p)$ must be identical. We know that the singularities of $S(p)$ must lie on the lower half-plane.[1] From general considerations derived from causality conditions, we can obtain further information.[1] In fact, the possible singularities of $S(p)$ representing resonances appear as pairs of poles on the lower half-plane having symmetric positions with respect to the imaginary axis. This means that if p_R is one of these poles, the other is $-p_R^*$. These poles are zeros of $\varphi(p)$ with the same order. The condition that p_R and $-p_R^*$ must lie on the lower half-plane means that

$$\text{Im } p_R<0. \tag{5.13}$$

After all these comments, we should remark that the two first formulas in (5.12) give us four real relations for the six real parameters E_1, λ, Re p_R, Im p_R, Re b, and Im b. Therefore, we only have two free real parameters. We choose the eigenvalue E_1 of H_0 and the coupling parameter λ as free parameters. Then, we obtain for b and p_R the following expressions:

$$b=E_1^{1/2}+2i\left(\frac{\pi\lambda^2}{16E_1}\right)^{1/3}; \quad p_R=\left[E_1-\left(\frac{\pi\lambda^2}{16E_1}\right)^{2/3}\right]-\left(\frac{\pi\lambda^2}{16E_1}\right)^{1/3}. \tag{5.14}$$

The conditions required on p_R, in particular (5.13), are obviously satisfied if we impose on the independent variables E_1 and λ, the following extra condition:

$$E_1-\left(\frac{\pi\lambda^2}{16E_1}\right)^{2/3}\geqslant 0. \tag{5.15}$$

The value for p_R given in (5.14) can be carried to (5.11) to obtain

$$\varphi(p)=-\frac{(p-p_R)^2(p+p_R^*)^2}{(p+t)(p-t^*)}. \tag{5.16}$$

We observe that p_R as well as $-p_R^*$ are double zeros of $\varphi(p)$. From (5.14) and (5.15), we see that they lie on the lower half-plane. The squares $z_R=p_R^2$ and $z_R^*=(-p_R^*)^2$ are the double zeros of the analytic continuation of the extended analytic function $\eta(z)$ and, therefore, the resonances of this Friedrichs model.

A. The Gamow vectors

Now, we shall obtain the explicit form of the Gamow vectors $|f_0\rangle$ and $|f_1\rangle$ corresponding to the Friedrichs model studied above. We shall obtain the Gamow vectors directly, using the analytic properties of $\eta(z)$. To begin with, let us write the Hamiltonian (5.1) as follows:

$$H=E_1|1\rangle\langle 1|+\int_0^\infty dE E|E\rangle\langle E|+\lambda\int_0^\infty dE\{V(E)|E\rangle\langle 1|+V^*(E)|1\rangle\langle E|\}, \tag{5.17}$$

where $V(E)$ is the form factor and $|E\rangle$ is the generalized eigenvector of H_0 with positive eigenvalue E, i.e.,

$$H_0|E\rangle=E|E\rangle. \tag{5.18}$$

The next goal is to find the generalized eigenvector of H with eigenvalue $w>0$. Equivalently, we have to solve the eigenvalue equation given by

$$(H-w)\mathbf{\Psi}(w)=0. \tag{5.19}$$

Since the vectors $|1\rangle$ and $|E\rangle$, with $E>0$ form a complete set, the solution $\mathbf{\Psi}(w)$ of Eq. (5.19) must have the form

$$\mathbf{\Psi}(w)=\psi(w)|1\rangle+\int_0^\infty dE\Psi(w,E)|E\rangle, \tag{5.20}$$

where $\psi(w)$ and $\Psi(w,E)$ are functions depending on these variables, w and E. In order to obtain them, we use (5.17) and (5.20) in (5.19). We obtain the following system of equations:

$$\begin{aligned}(E_1-w)\psi(E)+\lambda\int_0^\infty dE\Psi(w,E)V^*(E)=0,\\ (E-w)\Psi(w,E)+\lambda V(E)\psi(E)=0.\end{aligned} \tag{5.21}$$

The procedure to solve the system (5.21) is standard. The second equation in (5.21) allows us to write $\psi(E)$ in terms of $\Psi(w,E)$. This relation can be carried to the first equation in (5.21) to obtain an integro-differential equation which is easily solvable. The final result is

$$\mathbf{\Psi}_+(w)=|w\rangle+\lambda V^*(w)\,\frac{1}{\eta_+(w)}\left\{|1\rangle+\lambda\int_0^\infty dE\,\frac{V(E)}{[w-E]_+}\,|E\rangle\right\}, \tag{5.22}$$

where the plus sign in the denominator of (5.22) means that the function under the integral sign in (5.22) is given by the values on $\mathbb{R}^+$ of its extension from above to below. The expression (5.26) admits an analytic extension from above to below given by $\mathbf{\Psi}_+(z)$ having poles at the zeros of $\eta_+(z)$ with the same multiplicity. Therefore, this extension has a double pole located at the point z_R. In a neighborhood of z_R, $\mathbf{\Psi}_+(z)$ has the following form, according to the Laurent[48] theorem:

$$\mathbf{\Psi}_+(z)=\frac{\mathbf{C}_1}{(z-z_R)^2}+\frac{\mathbf{C}_2}{z-z_R}+O(z), \tag{5.23}$$

where $O(z)$ in (5.27) represents an analytic function of z. Now, the vectors $\mathbf{C}_1$ and $\mathbf{C}_2$ are easily evaluated with the help of (5.22). The singular part of $\mathbf{\Psi}_+(z)$, i.e., $\mathbf{\Psi}_+(z)-O(z)$, is now given by

$$\frac{\lambda V^*(z_R)}{(z-z_R)^2}\left\{|1\rangle+\lambda\int_0^\infty dE\,\frac{V(E)}{[z-E]_+}\,|E\rangle\right\}$$
$$=\frac{\lambda V^*(z_R)}{(z-z_R)^2}\left\{|1\rangle+\lambda\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+}\,|E\rangle-\lambda(z-z_R)\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+^2}\,|E\rangle\right\}+O(z), \tag{5.24}$$

where, again, $O(z)$ represents the analytic part. From, (5.23) and (5.24), we obviously have (save for an irrelevant multiplicative constant),

$$\mathbf{C}_1=|1\rangle+\lambda\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+}\,|E\rangle, \tag{5.25}$$

$$\mathbf{C}_2=-\lambda\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+^2}\,|E\rangle. \tag{5.26}$$

We now have to show that $\mathbf{C}_1$ and $\mathbf{C}_2$ are indeed the Gamow vectors we are looking for. More precisely, we want to show that

$$|f_0\rangle=\mathbf{C}_1;\ \ |f_1\rangle=\mathbf{C}_2. \tag{5.27}$$

Due to the properties of $|f_0\rangle$ and $|f_1\rangle$ presented in Sec. III, in order to show (5.27) we have to show that

$$H|f_0\rangle=z_R|f_0\rangle;\ \ H|f_1\rangle=z_R|f_1\rangle+|f_0\rangle. \tag{5.28}$$

The first equation in (5.28) can be easily obtained by applying (5.17) directly into (5.25). We want to note that this expression for $|f_0\rangle$ has already been obtained for the case of a nondegenerate resonance.[22] In order to show that $|f_1\rangle$ fulfills the second equation in (5.28), we apply (5.17) to (5.26). One gets

$$H|f_1\rangle=-\lambda\int_0^\infty dE'E'|E'\rangle\langle E'|\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+^2}\,|E\rangle$$
$$-\lambda^2\int_0^\infty dE'V^*(E)|1\rangle\langle E'|\int_0^\infty dE\,\frac{V(E)}{[z_R-E]_+^2}\,|E\rangle$$
$$=-\lambda\int_0^\infty dE\,\frac{EV(E)}{[z_R-E]_+^2}\,|E\rangle-\left[\lambda^2\int_0^\infty dE\,\frac{|V(E)|^2}{[z_R-E]_+^2}\right]|1\rangle. \tag{5.29}$$

The coefficient of $|1\rangle$ in (5.29) is equal to one. To show it, let us go back to (5.5) and extend it analytically from above to below. Thus, we obtain the following expression:

$$\eta_+(z)=E_1-z+\lambda^2\int_0^\infty dE\,\frac{|V(E)|^2}{[z-E]_+}. \tag{5.30}$$

Its derivative with respect to z is given by

$$\eta_+'(z)=-1-\lambda^2\int_0^\infty dE\,\frac{|V(E)|^2}{[z-E]_+^2}. \tag{5.31}$$

To obtain (5.31) from (5.30) we use similar arguments than those given in the proof of the Theorem in the Appendix. Since (5.16) implies that $\eta_+'(z_R)=0$, we have that

$$\eta_+'(z_R)=0=-1-\lambda^2\int_0^\infty dE\,\frac{|V(E)|^2}{[z_R-E]_+^2}, \tag{5.32}$$

which shows our claim.

Now, (5.29) can be written as

$$H|f_1\rangle = |1\rangle + \lambda \int_0^\infty dE\, \frac{(-E+z_R)V(E)}{[z_R-E]_+^2}|E\rangle + z_R\left[-\lambda \int_0^\infty dE\, \frac{V(E)}{[z_R-E]_+^2}|E\rangle\right]$$
$$= |1\rangle + \lambda \int_0^\infty dE\, \frac{V(E)}{[z_R-E]_+}|E\rangle + z_R\left[-\lambda \int_0^\infty dE\, \frac{V(E)}{[z_R-E]_+^2}|E\rangle\right]$$
$$= |f_0\rangle + z_R|f_1\rangle. \quad \blacksquare \tag{5.33}$$

Analogously, one obtains for $|\tilde{f}_0\rangle$ and $|\tilde{f}_1\rangle$ the following expressions:

$$|\tilde{f}_0\rangle = |1\rangle + \lambda \int_0^\infty dE\, \frac{V(E)}{[z_R^*-E]_-}|E\rangle, \tag{5.34}$$

$$|\tilde{f}_1\rangle = -\lambda \int_0^\infty dE\, \frac{V(E)}{[z_R^*-E]_-^2}|E\rangle. \tag{5.35}$$

B. Remarks

The Gamow–Jordan eigenvectors $|f_0\rangle$ and $|\tilde{f}_0\rangle$ coincide with the Gamow vectors for the Friedrichs model with simple resonances.[22]

The Gamow–Jordan vectors $|f_1\rangle$ and $|\tilde{f}_1\rangle$ are not unique and can be defined up to a term proportional to $|f_0\rangle$ and $|\tilde{f}_0\rangle$, respectively.

VI. CONCLUDING REMARKS

(1) The phenomenon of resonance overlapping[50,51] might be related to the double pole resonance formalism. Resonance overlapping is a trapping effect that appears when the width of two resonances is of the same order of magnitude as their average distance. The result of this trapping effect is to replace these resonances by another pair: one with a very short mean life and a second one with a very long mean life. However, this possibility has not been explored so far.

(2) This paper actually extends the Bohm–Gadella work[3,4,5,44] from simple to degenerate scattering resonances which are also poles of the extended resolvent for the evolution operator. This is the common point between the Bohm–Gadella work and the Brussels–Austin approach[18–25] where the extension of the resolvent was obtained through subdynamics. The understanding of resonances as singularities of the extended resolvent[12,24,31,41] is the natural generalization of both Bohm–Gadella and Brussels–Austin approaches.

(3) We restricted ourselves to the extended Schrödinger evolution which is intrinsically irreversible but not intrinsically probabilistic. Intrinsically probabilistic evolutions can be obtained in the Liouville space where we study the singularities of the extended Liouville operator. The extended Liouville evolution is the natural framework for quantum statistical physics and the theory of quantum measurements. We shall discuss the rigged Liouville space and the extended dual pairs of states and observables elsewhere.[52]

ACKNOWLEDGMENTS

We want to express our gratitude to Professors I. Prigogine and A. Bohm for their constant support and encouragement. We also thank Professors E. Brändas and A. Chatzidimitriou-Dreismann who stimulated our interest in the Jordan block structures and sent us their papers. We also thank Professor L. Horwitz who demonstrated to us the natural relation of multiple resonances with Jordan blocks and introduced us to the work of Katznelson. We received financial support from the International Solvay Institutes, the Belgian Interuniversity Attraction Poles, the European Commission ESPRIT Project No. CTIAC-21042, the NATO CRG CRG940077, the Spanish DGICYT PB 95-0719, and the Junta de Castilla y León PC01/197.

APPENDIX

The aim of this appendix is to present three interesting results relative to Hardy class functions on a half-plane. These results are used in this paper and, to the knowledge of the authors, are not explicitly given in the standard literature.

Theorem: Let $\varphi_+(E)\in\mathscr{H}^2_+\cap S$ and $z\in\mathbb{C}$. Then,

$$\varphi_+^n(z)=\frac{n!}{2\pi i}\int_{-\infty}^{\infty}\frac{\varphi(E)}{(E-z)^{n+1}}\,dE. \tag{A1}$$

Proof: Let us prove it for the first derivative, then, one obtains the general result by induction. Consider

$$\frac{\varphi_+(z+w)-\varphi_+(z)}{w}=\frac{1}{2\pi i w}\int_{-\infty}^{\infty}\left\{\frac{\varphi_+(E)}{(E-z-w)}-\frac{\varphi_+(E)}{E-z}\right\}dE$$

$$=\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{\varphi_+(E)dE}{(E-z-w)(E-z)}\,. \tag{A2}$$

If we take w small enough, we always have

$$|E-(z+w)|\geqslant|E-z|-|w|\geqslant|\mathrm{Im}\ z|-|w|. \tag{A3}$$

We may always assume that $|w|\leqslant\frac{1}{2}|\mathrm{Im}\ z|=h$. Then,

$$|E-(z+w)|\geqslant h\Rightarrow\frac{1}{|E-z-w|}\leqslant\frac{1}{h}\,. \tag{A4}$$

The modulus of the last integral in (A2) is less or equal to

$$\int_{-\infty}^{\infty}\frac{|\varphi_+(E)|dE}{h|E-z|}\leqslant\frac{1}{h}\left\{\int_{-\infty}^{\infty}|\varphi_+(E)|^2dE\right\}^{1/2}\left\{\int_{-\infty}^{\infty}\frac{1}{|E-z|^2}\right\}^{1/2} \tag{A5}$$

which obviously converges. Since

$$\frac{|\varphi_+(E)|}{|E-z||E-z-w|}\leqslant\frac{1}{h}\frac{|\varphi_+(E)|}{|E-z|}\,, \tag{A6}$$

the Lebesgue theorem says that one can exchange the symbols integral and limit (as $w\to0$) in (A2) so as to have

$$\lim_{w\to0}\frac{\varphi_+(z+w)-\varphi(z)}{w}=\frac{1}{2\pi i}\int_{-\infty}^{\infty}\lim_{w\to0}\frac{\varphi_+(E)dE}{(E-z-w)(E-z)}\,dE, \tag{A7}$$

which proves our result for $n=1$. ■

Proposition: If $\varphi_+\in\mathscr{H}^2_+\cap S$, $\varphi_+^{(n}(E)$ represents the boundary values of $\varphi_+^{(n}(z)$ with $\mathrm{Im}\,z>0$.

Proof: The Payley Wiener theorem[14–16] states that if $\mathrm{Im}\,z\geqslant0$, one has

$$\varphi_+(z)=\int_{-\infty}^{0}e^{-itz}F(t)dt,\quad\text{with}\quad F(t)\in S(\mathbb{R}^+). \tag{A8}$$

Deriving n times

$$\varphi_+^{(n}(z)=\int_{-\infty}^{0}(-it)^ne^{-itz}F(t)dt. \tag{A9}$$

This procedure can be done because the function under the integral sign is absolutely integrable. Since $(z=E+iy\ y>0)$,

$$|e^{-it(E+iy)}|=e^{ty}\Rightarrow|(-it)^n e^{-itz}F(t)|\leqslant t^n e^{ty}|F(t)| \quad t<0 \tag{A10}$$

which is integrable in $(-\infty,0)$. Thus, applying the Lebesgue theorem, we obtain

$$\lim_{y\to 0}\varphi_+^{(n}(E+iy)=\int_0^\infty \lim_{y\to 0}(-it)^n e^{-it(E+iy)}F(t)dt$$

$$=\int_{-0}^{\infty}(-it)^n e^{-itE}F(t)dt=\varphi_+^{(n}(E), \tag{A11}$$

which proves the proposition. ■

The next result intends to show the following: Given an analytic function on the two sheeted Riemann surface corresponding to the square root, we obtain an analytic function on the complex plane by using the transformation $z=p^2$. These two functions have the same structure, which means in particular that if the function on the complex plane has a zero at the point p_0, the function on the Riemann surface has a zero of the same multiplicity at the point $z_0=p_0^2$ and *vice versa* as we can see from our next result. Note that the correspondence between the surface and the plane is one to one. This result remains the same if we exchange the word zero for the word pole, since a complex meromorphic function, $\varphi(z)$ has a zero of order n at a point if and only $\varphi^{-1}(z)$ has a pole of order n at the same point.

Proposition: Let $g(p)=p^2$ and $z=p^2$. The analytic function $f(z)$ has a zero of order n at p_0 where $p_0^2=z_0$.

Proof: Let us prove this result for the case of a simple zero. The proof for higher order zeros is similar. This chain rule gives

$$\frac{d}{dz}(f\circ g)(p_0)=\frac{df}{dz}(z=z_0=g(p_0))\,\frac{dg}{dp}(p_0)=2p_0\,\frac{df}{dz}(z_0). \tag{A12}$$

From (A12) it is obvious that

$$\frac{d}{dz}(f\circ g)(p_0)=0 \quad \text{if} \quad \text{and} \quad \text{only} \quad \text{if}\ \frac{df}{dz}(z_0)=0. \tag{A13}$$

This is a zero of first order if and only if the respective second derivatives are different from zero at the points p_0 and z_0. To show this, let us use the chain rule for the second derivative,

$$\frac{d^2}{dz^2}(f\circ g)(p_0)=\frac{d^2f}{dz^2}(z_0)\left(\frac{dg}{dp}(p_0)\right)^2+\frac{df}{dz}(z_0)\,\frac{d^2g}{dp^2}(z_0). \tag{A14}$$

Since $f(z)$ has a zero at z_0, (A14) yields

$$\frac{d^2}{dz^2}(f\circ g)(p_0)=4p_0^2\,\frac{d^2g}{dz^2}(z_0), \tag{A15}$$

which shows that

$$\frac{d^2}{dz^2}(f\circ g)(p_0)\neq 0 \quad \text{if} \quad \text{and only} \quad \text{if}\ \frac{d^2g}{dz^2}(z_0)\neq 0. \tag{A16}$$

This proof for simple zeros shows us how to proceed in extending it to higher order zeros. ■

[1] H. M. Nussenszveig, *Causality and Dispersion Relations* (Academic, New York, 1972).
[2] R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).
[3] A. Bohm, *Quantum Mechanics: Foundations and Applications* (Springer, New York, 1994).

[4] A. Bohm and M. Gadella, *Dirac Kets, Gamow Vectors and Gel'fand Triplets*, Springer Lecture Notes in Physics (Springer, Berlin, 1989), Vol. 348.
[5] A. Bohm, J. Math. Phys. **22**, 2813 (1981).
[6] G. Gamow, Z. Phys. **51**, 204 (1928).
[7] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955).
[8] E. Prugovecki, *Quantum Mechanics in Hilbert Space* (Academic, New York, 1981).
[9] N. Nakanishi, Prog. Theor. Phys. **19**, 607 (1958).
[10] R. E. Peierls, Proc. R. Soc. London, Ser. A **253**, 16 (1959).
[11] V. I. Kukulin, V. M. Krasnapolski, and J. Horacek, *Theory of Resonances* (Academia, Praha, 1989).
[12] L. Horwitz and P. Marchand, Rocky Mt. J. Math. **1**, 225 (1971).
[13] I Gel'fand and N. Vilenkin, *Generalized Functions* (Academic, New York, 1964), Vol. 4.
[14] H. H. Schaeffer, *Topological Vector Spaces* (Springer, New York, 1970).
[15] P. Koosis, *Introduction to H_p Spaces*, London Math. Soc. Lecture Note Series No. 40 (Cambridge University Press, Cambridge, 1980).
[16] P. Koosis, *The Logarithmic Integral* (Cambridge University Press, Cambridge, 1990).
[17] P. Duren, *Theory of H_p Spaces* (Academic, New York, 1970).
[18] I. Prigogine, *From Being to Becoming* (Freeman, New York, 1980).
[19] T. Petroski, I. Prigogine, and S. Tasaki, Physica A **173A**, 175 (1991).
[20] T. Petroski and I. Prigogine, Physica A **175**, 146 (1991).
[21] I. Prigogine, Phys. Rep. **219**, 93 (1992).
[22] I. Antoniou and I. Prigogine, Physica A **192**, 443 (1993).
[23] I. Antoniou and S. Tasaki, Int. J. Quantum Chem. **44**, 425 (1993).
[24] I. Antoniou, L. Dimitrieva, Y. Kuperin, and Y. Melnikov, Comput. Math. Appl. **34**, 399–425 (1997).
[25] I. Antoniou and S. Tasaki, Physica A **190**, 303 (1992).
[26] M. L. Goldberger and K. M. Watson, Phys. Rev. **132B**, 1472 (1964).
[27] J. Lukierski, Bull. Acad. Pol. Sci. Sér. des Sci. Math. Astron. et Phys. **XV**, 217 (1967).
[28] J. Lukierski, Bull. Acad. Pol. Sci. Sér. des Sci. Math. Astron. et Phys. **XV**, 223 (1967).
[29] K. Katznelson, *The Evolution of Nondiagonal Degenerate Hamiltonians*, Ph.D. thesis, Tel Aviv, 1980.
[30] E. Katznelson, J. Math. Phys. **21**, 1393 (1980).
[31] E. Baslev and J. M. Combes, Commun. Math. Phys. **22**, 280 (1971).
[32] M. Reed and B. Simon, *Scattering Theory* (Academic, New York, 1979).
[33] E. Brändas and P. Froelich, Int. J. Quantum Chem. **13**, 619 (1978).
[34] C. E. Reid and E. J. Brändas, Lect. Notes Phys. **325**, 475 (1989).
[35] E. J. Brändas and C. A. Chatzidimitriou-Dreissmann, Lect. Notes Phys. **325**, 485 (1989).
[36] C. A. Chatzidimitriou-Dreismann, Adv. Chem. Phys. **LXXX**, 201 (1991).
[37] E. Brändas, *Dynamics During Spectroscopic Transitions*, edited by E. Lippert and J. Macomber (Springer, New York, 1995), Chaps. 6 and 7, pp. 148–242.
[38] A. Mondragón and E. Hernández, J. Phys. A **26**, 5595 (1993).
[39] E. C. G. Sudarshan and G. Bhamati, University of Texas preprint, Physics Department CPP, University of Texas, Austin, Texas, 1995.
[40] W. O. Amrein, J. Jauch, and B. Sinha, *Scattering Theory in Quantum Mechanics* (Benjamin, Reading, 1977).
[41] P. Exner, *Open Quantum Systems and Feynman Integrals* (Reidel, Dordrecht, 1985).
[42] M. Reed and B. Simon, *Functional Analysis* (Academic, New York, 1972).
[43] C. van Winter, J. Math. Anal. **47**, 633 (1974).
[44] M. Gadella, J. Math. Phys. **24**, 1462 (1983).
[45] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Clarendon, Oxford, 1958).
[46] K. O. Friedrichs, Commun. Pure Appl. Math. **1**, 361 (1948).
[47] E. C. G. Sudarshan, C. Chiu, and V. Gorini, Phys. Rev. D **18**, 2914 (1978).
[48] G. Parravicini, V. Gorini, and E. C. G. Sudarshan, J. Math. Phys. **21**, 1980 (1980).
[49] J. E. Marsden, *Basic Complex Analysis* (Freeman, San Franciso, 1973).
[50] W. Iskra, M. Müller, and I. Rotter, Phys. Rev. C **51**, 1842 (1995).
[51] R. S. Friedman and D. G. Truhlar, *Barrier Resonances and Chemical Reactivity*, in *Multiparticle Quantum Scattering with Applications to Nuclear, Atomic and Molecular Physics*, edited by D. G. Truhlar and B. Simon (Springer, New York, 1997).
[52] I. Antoniou, M. Gadella, and Z. Suchanecki, in *Proceedings of the Goslar Semigroup Symposium*, edited by A. Bohm, H. Doebner, and P. Kielanowski (Springer, New York, 1998).

Representation-theoretic aspects of two-dimensional quantum systems in singular vector potentials: Canonical commutation relations, quantum algebras, and reduction to lattice quantum systems

Asao Arai[a)]
Department of Mathematics, Hokkaido University, Sapporo 060, Japan

(Received 23 June 1997; accepted for publication 13 January 1998)

Some representation-theoretic aspects of a two-dimensional quantum system of a charged particle in a vector potential $\mathbf{A}$, which may be singular on an infinite discrete subset $\mathbf{D}$ of $\mathbf{R}^2$ are investigated. For each vector $\mathbf{v}$ in a set $\mathbf{V}(\mathbf{D})\subset\mathbf{R}^2\backslash\{\mathbf{0}\}$, the projection $P_{\mathbf{v}}$ of the physical momentum operator $\mathbf{P}:=\mathbf{p}-\alpha\mathbf{A}$ to the direction of $\mathbf{v}$ is defined by $P_{\mathbf{v}}:=\mathbf{v}\cdot\mathbf{P}$ as an operator acting in $L^2(\mathbf{R}^2)$, where $\mathbf{p}=(-iD_x, -iD_y)[(x,y)\in\mathbf{R}^2]$ with D_x (resp., D_y) being the generalized partial differential operator in the variable x (resp., y) and $\alpha\in\mathbf{R}$ is a parameter denoting the charge of the particle. It is proven that $P_{\mathbf{v}}$ is essentially self-adjoint and an explicit formula is derived for the strongly continuous one-parameter unitary group $\{e^{itP_{\mathbf{v}}}\}_{t\in\mathbf{R}}$ generated by the self-adjoint operator $\overline{P_{\mathbf{v}}}$ (the closure of $P_{\mathbf{v}}$), i.e., the magnetic translation to the direction of the vector $\mathbf{v}$. The magnetic translations along curves in $\mathbf{R}^2\backslash\mathbf{D}$ are also considered. Conjugately to $P_{\mathbf{v}}$ and $P_{\mathbf{w}}$ $[\mathbf{w}\in\mathbf{V}(\mathbf{D})]$, a self-adjoint multiplication operator $Q_{\mathbf{v},\mathbf{w}}$ is introduced, which is a linear combination of the position operators x and y, such that, if $\mathbf{A}$ is flat on $\mathbf{R}^2\backslash\mathbf{D}$, then $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}:=\{Q_{\mathbf{v},\mathbf{w}}, Q_{\mathbf{w},\mathbf{v}}, P_{\mathbf{v}}, P_{\mathbf{w}}\}$ gives a representation of the canonical commutation relations (CCR) with two degrees of freedom. Properties of the representation $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ are analyzed. In particular, a necessary and sufficient condition for $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ to be unitarily equivalent (or inequivalent) to the Schrödinger representation of CCR is established. The case where $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is inequivalent to the Schrödinger representation corresponds to the Aharonov–Bohm effect. Quantum algebraic structures [quantum plane and the quantum group $U_q(sl_2)$] associated with the pair $\{P_{\mathbf{v}}, P_{\mathbf{w}}\}$ are also discussed. Moreover, for every $\mathbf{A}$ in a class of vector potentials having singularities on the infinite lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2):=\{m\boldsymbol{\omega}_1+n\boldsymbol{\omega}_2|m,n\in\mathbf{Z}\}$ [the case $\mathbf{D}=\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$], where $\boldsymbol{\omega}_1\in\mathbf{R}^2$ and $\boldsymbol{\omega}_2\in\mathbf{R}^2$ are linearly independent, it is shown that the magnetic translations $e^{iP\boldsymbol{\omega}_j}$, $j=1,2$, with $\mathbf{A}$ replaced by a *modified* vector potential are reduced by the Hilbert space $l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ identified with a closed subspace of $L^2(\mathbf{R}^2)$. This result, which may be regarded as one of the most important novel results of the present paper, establishes a connection of continuous quantum systems in vector potentials to lattice ones. © *1998 American Institute of Physics.*
[S0022-2488(98)02805-9]

I. INTRODUCTION

This article is a continuation of the previous articles[1–5] concerning gauge theory (quantum mechanics of a particle interacting with an external gauge potential) on a nonsimply connected space in two dimensions. In such a gauge theory, a representation of the canonical commutation relations (CCR) with two degrees of freedom is realized by the position and the physical momentum operators if the gauge potential is flat. The nonsimply connectedness of the base space is essential for this representation to be nontrivial, i.e., not necessarily be equivalent to the Schrödinger representation of CCR with two degrees of freedom. The unitary equivalence or inequivalence of the representation to the Schrödinger representation is completely characterized in terms

a)Electronic mail: arai@math.sci.hokudai.ac.jp

 © 1998 American Institute of Physics

of the local Wilson loops. An interesting feature to be noted here is that the celebrated Aharonov–Bohm effect[6] may be mathematically well understood as the representation inequivalent to the Schrödinger one. Some physical implications of the inequivalent representation in the Abelian case (the case of quantum mechanics of a charged particle in an external vector potential) are discussed in terms of the Dirac–Weyl operator (Ref. 2). Moreover, it is shown that, in the case where the vector potential is given in terms of the Weierstrass Zeta function with singularities at $z=m\omega_1+in\omega_2$, $m,n\in\mathbf{Z}$ ($\omega_j>0$, $j=1,2$), the inequivalent representation induces representations of quantum planes (rotation algebras) and the quantum group $U_q(sl_2)$ with $|q|=1$, $q\in\mathbf{C}$ (Ref. 5).

In this paper we further pursue representation-theoretic aspects of a two-dimensional Abelian gauge theory in which the vector potential may have singularities on an infinite set of points in $\mathbf{R}^2$ and show by bringing some new aspects to light that such a gauge theory may have richer and deeper structures. We first demonstrate by an explicit construction that, associated with the position and the physical momentum operators, there exists a wide class of inequivalent representations of CCR that includes those given in Refs. 1 and 5. The basic idea of the construction is to consider projections of the position and the physical momentum operators to directions of vectors in $\mathbf{R}^2$. As in the previous simpler cases discussed in Refs. 1 and 5, these new discovered representations of CCR give new representations of quantum planes and $U_q(sl_2)$ if the singularities of the vector potential form an infinite lattice.

Another new aspect of the present article is concerned with magnetic translations. These objects can be defined in both continuous and lattice quantum systems in external magnetic fields. Magnetic translations in the former systems in *uniform magnetic fields* have been discussed in some detail (e.g., Refs. 7–10), but, it seems that investigations of magnetic translations in the case of *nonuniform magnetic fields* are missing in the literature. In this paper we define, in the continuous quantum system under consideration, magnetic (parallel) translations as the strongly continuous one-parameter unitary groups generated by the projected physical momentum operators and study their properties.

On the other hand, magnetic translations on lattice quantum systems have been extensively discussed in connection with models of the Hofstadter type (e.g., Refs. 11–13 and references therein). Lattice models are usually defined by *ad hoc* procedures from continuous quantum systems (e.g., ''tight-binding approximation'' or other analogies). From a unified point of view, this situation is obviously unsatisfactory. It would be natural and interesting to investigate if there exists any *internal* (non-*ad hoc*) reduction mechanism by which a lattice quantum system ''dynamically'' emerges from a continuous quantum system. In this paper we show that such a mechanism exists. We regard this result as one of the most important new results of the present paper.

We now describe the outline of the present article in more detail. As already mentioned above, we consider a continuous quantum system of a charged particle with charge $\alpha\in\mathbf{R}\backslash\{0\}$ moving in the Euclidean plane $\mathbf{R}^2$ under the influence of a perpendicular magnetic field B. We denote by

$$\mathbf{A}(\mathbf{r}):=(A_1(\mathbf{r}),A_2(\mathbf{r})),\quad \mathbf{r}=(x,y)\in\mathbf{R}^2, \tag{1.1}$$

the vector potential of the magnetic field (up to gauge transformations), so that

$$B=D_xA_2-D_yA_1, \tag{1.2}$$

where D_x and D_y are the generalized (distributional) partial differential operators in the variables x and y, respectively.

Let

$$\mathbf{D}:=\{\mathbf{a}_n=(a_{n1},a_{n2})\}_{n\in\mathbf{N}} \tag{1.3}$$

be a set of points in $\mathbf{R}^2$ such that $\mathbf{a}_n\neq\mathbf{a}_m$ if $n\neq m$ and the set $\{a_{nj}\}_{n=1}^{\infty}$ ($j=1,2$) has no accumulation point in $\mathbf{R}$. Then

$$\mathbf{M}:=\mathbf{R}^2\backslash\mathbf{D} \tag{1.4}$$

is an open set of $\mathbf{R}^2$. We assume that $\mathbf{A}$ is continuous on $\mathbf{M}$. But $\mathbf{A}$ may be singular on $\mathbf{D}$ and B may be a distribution on $\mathbf{R}^2$ with support in $\mathbf{D}$. Except for some general aspects, it is essential for

the theory presented below to be nontrivial that **A** has singularities in **D**. The Hilbert space of state vectors of the quantum system under consideration can be taken to be $L^2(\mathbf{R}^2)$.

The operators,

$$p_1 := -iD_x, \quad p_2 := -iD_y, \tag{1.5}$$

acting in $L^2(\mathbf{R}^2)$ with domain $D(p_1) := \{\Psi \in L^2(\mathbf{R}^2) | D_x\Psi \in L^2(\mathbf{R}^2)\}$ and $D(p_2) := \{\Psi \in L^2(\mathbf{R}^2) | D_y\Psi \in L^2(\mathbf{R}^2)\}$ are self-adjoint, where, for an operator T, we denote by $D(T)$ its domain. The physical momentum operator (the velocity operator up to a constant multiple) $\mathbf{P} := (P_1, P_2)$ is given by

$$P_j := p_j - \alpha A_j, \quad j = 1,2, \tag{1.6}$$

acting in $L^2(\mathbf{R}^2)$ with $D(P_j) = D(p_j) \cap D(A_j)$.

For each vector $\mathbf{v} = (v_1, v_2) \in \mathbf{R}^2 \backslash \{0\}$, we define

$$P_{\mathbf{v}} := \mathbf{v} \cdot \mathbf{P} := \sum_{j=1}^{2} v_j P_j. \tag{1.7}$$

We call it the *projection of the physical momentum operator* (or simply *the projected physical momentum operator*) to the direction of **v**. The operator P_j is a special case of $P_{\mathbf{v}}$:

$$P_j = P_{\mathbf{e}_j}, \quad j = 1,2, \tag{1.8}$$

with

$$\mathbf{e}_1 := (1,0), \quad \mathbf{e}_2 := (0,1). \tag{1.9}$$

We introduce a subset of vectors $\mathbf{R}^2$. Let

$$\mathbf{v} \wedge \mathbf{w} := v_1 w_2 - v_2 w_1, \quad \mathbf{v}, \mathbf{w} \in \mathbf{R}^2. \tag{1.10}$$

Definition 1.1: We say that $\mathbf{v} \in \mathbf{R}^2 \backslash \{0\}$ is in the set $\mathbf{V}(\mathbf{D})$ if the sequence $\{\mathbf{v} \wedge \mathbf{a}_n\}_{n=1}^{\infty}$ has no accumulation point.

By the assumption for **D** stated above, $\mathbf{e}_j \in \mathbf{V}(\mathbf{D})$, $j = 1,2$.

In Sec. II we prove the essential self-adjointness of $P_{\mathbf{v}}$ with $\mathbf{v} \in \mathbf{V}(\mathbf{D})$ and clarify the spectral properties of the self-adjoint operator $\overline{P_{\mathbf{v}}}$ (the closure of $P_{\mathbf{v}}$) (Theorems 2.4 and 2.5).

The essential self-adjointness of $P_{\mathbf{v}}$ allows one to define the *continuous magnetic translations* to the direction of **v** as the elements of the strongly continuous one-parameter unitary group,

$$T_{\mathbf{v}}^{\mathbf{A}}(t) := e^{it\overline{P_{\mathbf{v}}}}, \quad t \in \mathbf{R}, \tag{1.11}$$

generated by $\overline{P_{\mathbf{v}}}$. We prove that there exist no nontrivial finite-dimensional subspaces of $L^2(\mathbf{R}^2)$ left invariant by $T_{\mathbf{v}}^{\mathbf{A}}(t)$ (Proposition 2.6).

Section III is devoted to a basic analysis of of the continuous magnetic translations. We derive an explicit formula for $T_{\mathbf{v}}^{\mathbf{A}}(t)$ (Theorem 3.2) and, using it, we compute commutation relations of $T_{\mathbf{v}}^{\mathbf{A}}(s)$ and $T_{\mathbf{w}}^{\mathbf{A}}(t)$ $[s,t \in \mathbf{R}, \mathbf{w} \in \mathbf{V}(\mathbf{D})]$ (Theorem 3.3).

In Sec. IV we define magnetic translations along curves in **M** and investigate their properties.

Section V is devoted to analysis of representations of CCR appearing in the quantum system under consideration. We first introduce multiplication operators given as linear combinations of the position operators,

$$q_1 := x, \quad q_2 := y. \tag{1.12}$$

Namely, for vectors $\mathbf{v}, \mathbf{w} \in \mathbf{V}(\mathbf{D})$ *linearly independent*, we define

$$Q_{\mathbf{v},\mathbf{w}} := \frac{\mathbf{q} \wedge \mathbf{w}}{\mathbf{v} \wedge \mathbf{w}}, \tag{1.13}$$

where $\mathbf{q}:=(q_1,q_2)$.

We denote by $C^k(\mathbf{M})$ the set of k times continuously differentiable functions on $\mathbf{M}$ and by $C_0^k(\mathbf{M})$ the set of functions in $C^k(\mathbf{M})$ with bounded support in $\mathbf{M}$.

We say that $\mathbf{A}$ with $A_j \in C^1(\mathbf{M})$ $(j=1,2)$ is *flat* on $\mathbf{M}$ if $B(\mathbf{r})=0$ for all $\mathbf{r}\in\mathbf{M}$. The flatness of $\mathbf{A}$ on $\mathbf{M}$ physically means the magnetic field B is concentrated on the discrete set $\mathbf{D}$ in the distribution sense.

We show that, if $A_j \in C^1(\mathbf{M})$, then

$$\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}:=\{L^2(\mathbf{R}^2), C_0^2(\mathbf{M}), Q_{\mathbf{v},\mathbf{w}}, Q_{\mathbf{w},\mathbf{v}}, \overline{P}_{\mathbf{v}}, \overline{P}_{\mathbf{w}}\} \tag{1.14}$$

is a representation of the CCR with two degrees of freedom if and only if $\mathbf{A}$ is flat on $\mathbf{M}$ (for the terminology on the representation theory of CCR, we refer to Ref. 5, Sec. 1). We analyze properties of this representation. We see that results similar to those of simpler cases in Refs. 1 and 5 hold in the present case too, generalizing them.

In Sect. VI we give some remarks on quantum algebraic structures [representations of quantum planes and $U_q(sl_2)$] associated with the representation $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$.

In the last section we consider the problem of reduction of the continuous quantum system under consideration to a lattice one. For this purpose, we fix arbitrarily two linearly independent vectors,

$$\boldsymbol{\omega}_1:=(\omega_{11},\omega_{12}), \quad \boldsymbol{\omega}_2:=(\omega_{21},\omega_{22})\in\mathbf{R}^2, \tag{1.15}$$

such that $\boldsymbol{\omega}_1\wedge\boldsymbol{\omega}_2>0$ and take as $\mathbf{D}$ an infinite lattice,

$$\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2):=\{\boldsymbol{\Omega}_{m,n}|m,n\in\mathbf{Z}\}, \tag{1.16}$$

where

$$\boldsymbol{\Omega}_{m,n}:=m\boldsymbol{\omega}_1+n\boldsymbol{\omega}_2. \tag{1.17}$$

The Hilbert space of state vectors of a quantum system on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ is taken to be

$$l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)):=\left\{\psi=\{\psi(\boldsymbol{\Omega}_{m,n})\}_{m,n\in\mathbf{Z}}\,\middle|\, \psi(\boldsymbol{\Omega}_{m,n})\in\mathbf{C}, \quad m,n\in\mathbf{Z}, \right.$$
$$\left.\sum_{m,n\in\mathbf{Z}}|\psi(\boldsymbol{\Omega}_{m,n})|^2<\infty\right\}. \tag{1.18}$$

This Hilbert space can be regarded as a closed subspace of $L^2(\mathbf{R}^2)$ in a natural way. Indeed, let $\mathbf{S}_{m,n}$ be the interior domain of the parallelogram determined by the four vectors $\boldsymbol{\Omega}_{m+1,n}-\boldsymbol{\Omega}_{m,n}$, $\boldsymbol{\Omega}_{m+1,n+1}-\boldsymbol{\Omega}_{m+1,n}$, $\boldsymbol{\Omega}_{m,n+1}-\boldsymbol{\Omega}_{m,n}$, $\boldsymbol{\Omega}_{m+1,n+1}-\boldsymbol{\Omega}_{m,n+1}$ and $\chi_{m,n}$ be the characteristic function of $\mathbf{S}_{m,n}$. Then each element $\psi\in l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ can be regarded as an element of $L^2(\mathbf{R}^2)$ by the correspondence

$$\psi\to\tilde{\psi}:=\sum_{m,n\in\mathbf{Z}}\psi(\boldsymbol{\Omega}_{m,n})\chi_{m,n}\in L^2(\mathbf{R}^2),$$

so that, under this correspondence, $l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ can be identified with the closed subspace,

$$L^2_{\omega_1,\omega_2}(\mathbf{R}^2):=\left\{\Psi\in L^2(\mathbf{R}^2)\,\middle|\,\Psi=\sum_{m,n\in\mathbf{Z}}\Psi_{m,n}\chi_{m,n},\Psi_{m,n}\in\mathbf{C},\sum_{m,n\in\mathbf{Z}}|\Psi_{m,n}|^2<\infty\right\}, \tag{1.19}$$

consisting of elements in $L^2(\mathbf{R}^2)$ being constant on each $\mathbf{S}_{m,n}$. It is easy to see that $\boldsymbol{\omega}_j \in\mathbf{V}(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$, $j=1,2$. If the unitary operators,

$$T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}:=T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}(1), \quad j=1,2, \tag{1.20}$$

leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant, then they are reduced by $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ and their restriction to $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ gives a set of magnetic translations on a quantum system on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$. In that case, models of the Hofstadter type on $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ are obtained as internal or ''dynamical'' reductions of models on the continuous space $\mathbf{R}^2\backslash\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$. Thus, the problem is to determine the class of vector potentials $\mathbf{A}$ such that $T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}$ leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant. At first sight, such vector potentials may seem not to exist, except for some physically trivial cases. This is certainly true as long as one considers *regular* vector potentials on $\mathbf{R}^2$. But, if the vector potential is allowed to be *singular* on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$, then the situation changes drastically. Indeed, in that case, we can show that there exists a method of constructing vector potentials $\mathbf{A}$ for which $T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}$ leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant. Roughly speaking, the method is as follows. Let

$$\omega_j=\omega_{j1}+i\omega_{j2},\quad j=1,2, \tag{1.21}$$

be the complex numbers corresponding to $\boldsymbol{\omega}_j\in\mathbf{R}^2$, $j=1,2(i:=\sqrt{-1})$. We take a meromorphic function $f(z)$ on $\mathbf{C}$ with poles at

$$\Omega_{m,n}:=m\omega_1+n\omega_2,\quad m,n\in\mathbf{Z}, \tag{1.22}$$

such that $df(z)/dz$ is an elliptic function with periods ω_j, $j=1,2$ and define a vector potential $\mathbf{A}$ by

$$A_1(\mathbf{r}):=\Im f(z),\quad A_2(\mathbf{r}):=\Re f(z),\quad \mathbf{r}=(x,y)\in\mathbf{R}^2,\quad z=x+iy. \tag{1.23}$$

We show that $T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}$ themselves may not leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant, but there exists a correspondence $\mathbf{A}\to\tilde{\mathbf{A}}$ of the vector potential such that $T^{\tilde{\mathbf{A}}}_{\pm\boldsymbol{\omega}_j}$ do leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant.

Unfortunately we have been unable to solve the problem of determining *all* the vector potentials $\mathbf{A}$ such that $T^{\mathbf{A}}_{\pm\boldsymbol{\omega}_j}$ $(j=1,2)$ leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant. We leave this problem for future study.

We remark that, in the same way as in Refs. 3 and 4, the results presented in Secs. II–V can be extended in a natural way to the case of non-Abelian gauge theories. If any significant aspects are discovered in the non-Abelian case, then we shall report them in a separate paper.

II. BASIC PROPERTIES OF THE PROJECTED PHYSICAL MOMENTUM OPERATORS

A. The physical momentum operator

The mathematical analysis of the physical momentum operator $\mathbf{P}$ in the present case can be made quite similarly to that in the case where $\mathbf{D}$ is a finite discrete set (Ref. 1). But, for the reader's convenience as well as for later reference, we briefly describe some basic properties of P_j, $j=1,2$. We introduce two sets:

$$\mathbf{M}_1:=\{(x,y)\in\mathbf{R}^2|x\in\mathbf{R},y\neq a_{n2},n\in\mathbf{N}\}, \tag{2.1}$$

$$\mathbf{M}_2:=\{(x,y)\in\mathbf{R}^2|x\neq a_{n1},n\in\mathbf{N},y\in\mathbf{R}\}, \tag{2.2}$$

which, by the assumed property of $\mathbf{D}$, are open sets of $\mathbf{R}^2$. Let

$$U_1(x,y):=e^{-i\alpha\int_0^x A_1(x',y)dx'},\quad U_2(x,y):=e^{-i\alpha\int_0^y A^2(x,y')dy'}. \tag{2.3}$$

Then $U_j\in C(\mathbf{M}_j)$, $j=1,2$. Since the Lebesgue measure of the set $\{(x,y)|x\in\mathbf{R},\ y=a_{n2},\ n\in\mathbf{N}\}$ (resp., $\{(x,y)|y\in\mathbf{R},\ x=a_{n1},\ n\in\mathbf{N}\}$) is zero, U_1 (resp., U_2) defines a unique multiplication unitary operator on $L^2(\mathbf{R}^2)$, which we denote by the same symbol U_1 (resp., U_2).

In what follows, we assume the following.

Hypothesis $(A)_k$: *For a non-negative integer* k, $A_j\in C^k(\mathbf{M})$, $j=1,2$.

Proposition 2.1: The operator $P_j(j=1,2)$ *is essentially self-adjoint on* $C_0^k(\mathbf{M}_j)$ *and the operator equations,*

$$\overline{P}_j = U_j^{-1} p_j U_j, \quad j=1,2, \tag{2.4}$$

hold.

Proof: The unitary operator U_j maps $C_0^k(\mathbf{M}_j)$ to itself bijectively and, for all $\Psi \in C_0^k(\mathbf{M}_j)$, $P_j\Psi = U_j^{-1} p_j U_j \Psi$, $j=1,2$. Since p_j is essentially self-adjoint on $C_0^k(\mathbf{M}_j)$, it follows that P_j is essentially self-adjoint on $C_0^k(\mathbf{M}_j)$ and (2.4) is obtained. ∎

We denote by $\sigma(P_j)$ [resp., $\sigma_{\mathrm{ac}}(P_j)$, $\sigma_p(P_j)$, $\sigma_{\mathrm{sc}}(P_j)$] the spectrum (resp., absolutely continuous, point, singular continuous spectrum) of P_j.

Proposition 2.1 (the unitary equivalence of P_j to p_j) and the well-known spectral property of p_j imply the following.

Proposition 2.2: For $j=1,2$,

$$\sigma(\overline{P}_j) = \sigma_{\mathrm{ac}}(\overline{P}_j) = \mathbf{R}, \quad \sigma_{\mathrm{p}}(\overline{P}_j) = \sigma_{\mathrm{sc}}(\overline{P}_j) = \varnothing. \tag{2.5}$$

B. The projected physical momentum operators

Let $\mathbf{V}(\mathbf{D})$ be as in Definition 1.1. For $\mathbf{v} \in \mathbf{V}(\mathbf{D})$, we define

$$L(\mathbf{a}_n;\mathbf{v}) = \{\mathbf{a}_n + s\mathbf{v} \mid s \in \mathbf{R}\}, \tag{2.6}$$

which is the straight line passing through the point $\mathbf{a}_n$ with the direction of $\mathbf{v}$. Then

$$\mathbf{M}_{\mathbf{v}}(\mathbf{D}) := \mathbf{R}^2 \backslash \cup_{n=1}^{\infty} L(\mathbf{a}_n;\mathbf{v}), \tag{2.7}$$

is an open set of $\mathbf{R}^2$.

It is well known that the angular momentum operator,

$$L := q_1 p_2 - q_2 p_1, \tag{2.8}$$

is essentially self-adjoint on $C_0^{\infty}(\mathbf{R}^2)$ (e.g., Ref. 14, Sec. 3). We denote its closure by the same symbol L. Then, for all $\theta \in \mathbf{R}$ and $\Psi \in L^2(\mathbf{R}^2)$, we have

$$(e^{i\theta L}\Psi)(\mathbf{r}) = \Psi(R(\theta)\mathbf{r}), \quad \text{a.e } \mathbf{r} \in \mathbf{R}^2, \tag{2.9}$$

where

$$R(\theta) := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{2.10}$$

is the rotation matrix in $\mathbf{R}^2$.

We write $\mathbf{v} = (v_1, v_2) \in \mathbf{V}(\mathbf{D})$ in the polar coordinate as

$$v_1 = v\cos\theta_{\mathbf{v}}, \quad v_2 = v\sin\theta_{\mathbf{v}}, \quad v = |\mathbf{v}|, \quad 0 \leqslant \theta_{\mathbf{v}} < 2\pi. \tag{2.11}$$

The function

$$A_{\mathbf{v}}(\mathbf{r}) := \frac{\mathbf{v}\cdot\mathbf{A}(R(\theta_{\mathbf{v}})\mathbf{r})}{v} \tag{2.12}$$

is in $C^k(R(-\theta_{\mathbf{v}})\mathbf{M}_{\mathbf{v}}(\mathbf{D}))$. We define

$$P_1(\mathbf{v}) := p_1 - \alpha A_{\mathbf{v}}. \tag{2.13}$$

Proposition 2.3: The operator $P_1(\mathbf{v})$ is essentially self-adjoint on $C_0^k(R(-\theta_{\mathbf{v}})\mathbf{M}_{\mathbf{v}}(\mathbf{D}))$.

Proof: We need only to apply Proposition 2.1 with A_1 and $\mathbf{a}_n$ replaced by $A_{\mathbf{v}}$ and $R(-\theta_{\mathbf{v}})\mathbf{a}_n$, respectively. ∎

The following theorem is a generalization of Proposition 2.1.

Theorem 2.4: *Let $\mathbf{v} \in \mathbf{V}(\mathbf{D})$. Then $P_{\mathbf{v}}$ is essentially self-adjoint on $C_0^k(\mathbf{M}_{\mathbf{v}}(\mathbf{D}))$ and the operator equation,*

$$e^{i\theta_{\mathbf{v}}L}\overline{P}_{\mathbf{v}}e^{-i\theta_{\mathbf{v}}L}=v\overline{P_1(\mathbf{v})}, \tag{2.14}$$

holds.

Proof: Let

$$\mathbf{p}:=(p_1,p_2). \tag{2.15}$$

Then it is easy to see that, for all $\theta\in\mathbf{R}$,

$$e^{i\theta L}\mathbf{p}e^{-i\theta L}=R(\theta)\mathbf{p}, \quad \text{on} \quad C_0^\infty(\mathbf{R}^2). \tag{2.16}$$

Hence

$$e^{i\theta_{\mathbf{v}}L}\mathbf{v}\cdot\mathbf{p}e^{-i\theta_{\mathbf{v}}L}=vp_1, \quad \text{on} \quad C_0^\infty(\mathbf{R}^2). \tag{2.17}$$

It is easy to see that

$$e^{-i\theta_{\mathbf{v}}L}C_0^k(R(-\theta_{\mathbf{v}})\mathbf{M}_{\mathbf{v}}(\mathbf{D}))=C_0^k(\mathbf{M}_{\mathbf{v}}(\mathbf{D})) \tag{2.18}$$

and

$$e^{i\theta_{\mathbf{v}}L}\mathbf{v}\cdot\mathbf{A}e^{-i\theta_{\mathbf{v}}L}=vA_{\mathbf{v}}, \quad \text{on} \quad C_0^k(R(-\theta_{\mathbf{v}})\mathbf{M}_{\mathbf{v}}(\mathbf{D})).$$

Hence

$$e^{i\theta_{\mathbf{v}}L}P_{\mathbf{v}}e^{-i\theta_{\mathbf{v}}L}=vP_1(\mathbf{v}), \quad \text{on} \quad C_0^k(R(-\theta_{\mathbf{v}})\mathbf{M}_{\mathbf{v}}(\mathbf{D})). \tag{2.19}$$

It follows from Proposition 2.3, (2.18) and (2.19), that $P_{\mathbf{v}}$ is essentially self-adjoint on $C_0^k(\mathbf{M}_{\mathbf{v}}(\mathbf{D}))$. This result implies that (2.19) can be extended to the operator equation (2.14). ■

Remark 2.1: If $\mathbf{D}$ is a finite discrete set, then $P_{\mathbf{v}}$ is essentially self-adjoint for all $\mathbf{v}\in\mathbf{R}^2\backslash\{0\}$.

By (2.14) and Proposition 2.2, we obtain the following result on the spectral property of $\overline{P}_{\mathbf{v}}$.

Theorem 2.5: *For all* $\mathbf{v}\in\mathbf{V}(\mathbf{D})$,

$$\sigma(\overline{P}_{\mathbf{v}})=\sigma_{\mathrm{ac}}(\overline{P}_{\mathbf{v}})=\mathbf{R}, \quad \sigma_{\mathrm{p}}(\overline{P}_{\mathbf{v}})=\sigma_{\mathrm{sc}}(\overline{P}_{\mathbf{v}})=\varnothing. \tag{2.20}$$

C. Nonexistence of nontrivial finite-dimensional subspaces left invariant by continuous magnetic translations

Let $T_{\mathbf{v}}^{\mathbf{A}}(t)$ be defined as in (1.11). The following fact is important in considering the algebra generated by the magnetic translations $T_{\mathbf{v}_j}^{\mathbf{A}}(t)$, $j=1,...,n$ $(n\in\mathbf{N},\ \mathbf{v}_j\in\mathbf{V}(\mathbf{D}))$.

Proposition 2.6: Let $t\in\mathbf{R}\backslash\{0\}$ *and* $\mathbf{v}\in\mathbf{V}(\mathbf{D})$. *Then there exist no nontrivial finite dimensional subspaces of* $L^2(\mathbf{R}^2)$ *that are left invariant by* $T_{\mathbf{v}}^{\mathbf{A}}(t)$.

Proof: Suppose that there exists a finite-dimensional subspace $\mathscr{K}\neq\{0\}$ of $L^2(\mathbf{R}^2)$ that is left invariant by $T_{\mathbf{v}}^{\mathbf{A}}(t)$. Then $T_{\mathbf{v}}^{\mathbf{A}}(t)$ is reduced by $\mathscr{K}$. Since $T_{\mathbf{v}}^{\mathbf{A}}(t)$ is unitary and $\mathscr{K}$ is finite dimensional, the reduced part $T_{\mathbf{v}}^{\mathbf{A}}(t)[\mathscr{K}$ has an eigenvalue λ_0 with $|\lambda_0|=1$. This is also an eigenvalue of $T_{\mathbf{v}}^{\mathbf{A}}(t)$ on $L^2(\mathbf{R}^2)$. It follows from the spectral theorem of self-adjoint operators that $P_{\mathbf{v}}$ has an eigenvalue of the form $(\arg\lambda_0+2\pi n_0)/t$ $(n_0\in\mathbf{Z})$. But this contradicts the fact that $\sigma_{\mathrm{p}}(P_{\mathbf{v}})=\varnothing$ (Theorem 2.5). Thus, we obtain the desired result. ■

III. CONTINUOUS MAGNETIC TRANSLATIONS GENERATED BY THE PROJECTED PHYSICAL MOMENTUM OPERATORS

A. Explicit representations

Let

$$p_{\mathbf{v}}=\mathbf{v}\cdot\mathbf{p}, \quad \mathbf{v}\in\mathbf{R}^2. \tag{3.1}$$

Then we have for all $t\in\mathbf{R}$ and $\Psi\in L^2(\mathbf{R}^2)$,

$$(e^{itp_{\mathbf{v}}}\Psi)(\mathbf{r})=\Psi(\mathbf{v}+t\mathbf{v}),\quad \text{a.e. } \mathbf{r}\in\mathbf{R}^2. \tag{3.2}$$

Let $T^{\mathbf{A}}_{\mathbf{v}}(t)$ be defined by (1.11). Proposition 2.1 implies that, for all $t\in\mathbf{R}$,

$$T^{\mathbf{A}}_{\mathbf{e}_j}(t)=e^{itP_j}=U_j^{-1}e^{itp_j}U_j=U_j^{-1}e^{itp_{\mathbf{e}_j}}U_j,\quad j=1,2. \tag{3.3}$$

For two vectors $\mathbf{a},\mathbf{b}\in\mathbf{R}^2$, we denote by $\int_{\mathbf{a}}^{\mathbf{b}}\mathbf{A}(\mathbf{r})\cdot d\mathbf{r}$ the line integral of $\mathbf{A}$ over the straight line from $\mathbf{a}$ to $\mathbf{b}$.

Using (3.3) and (3.2), we can obtain an explicit formula for $T^{\mathbf{A}}_{\mathbf{e}_j}(t)$.

Theorem 3.1: *For all $t\in\mathbf{R}$, and $\Psi\in L^2(\mathbf{R}^2)$,*

$$(T^{\mathbf{A}}_{\mathbf{e}_j}(t)\Psi)(\mathbf{r})=e^{-i\alpha\int_{\mathbf{r}}^{\mathbf{r}+t\mathbf{e}_j}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}\Psi(\mathbf{r}+t\mathbf{e}_j)=e^{-i\alpha\int_{\mathbf{r}}^{\mathbf{r}+t\mathbf{e}_j}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}(e^{itp_{\mathbf{e}_j}}\Psi)(\mathbf{r}),\quad \text{a.e. } \mathbf{r}. \tag{3.4}$$

Theorem 3.1 can be extended to the case of the magnetic translation $T^{\mathbf{A}}_{\mathbf{v}}(t)$ with any vector $\mathbf{v}\in\mathbf{V}(\mathbf{D})$.

Theorem 3.2: *Let $\mathbf{v}\in\mathbf{V}(\mathbf{D})$. Then, for all $t\in\mathbf{R}$ and $\Psi\in L^2(\mathbf{R}^2)$,*

$$(T^{\mathbf{A}}_{\mathbf{v}}(t)\Psi)(\mathbf{r})=e^{-i\alpha\int_{\mathbf{r}}^{\mathbf{r}+t\mathbf{v}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}\Psi(\mathbf{r}+t\mathbf{v})=e^{-i\alpha\int_{\mathbf{r}}^{\mathbf{r}+t\mathbf{v}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}(e^{itp_{\mathbf{v}}}\Psi)(\mathbf{r}),\quad \text{a.e. } \mathbf{r}. \tag{3.5}$$

Proof: By Theorem 2.4, we have, for all $t\in\mathbf{R}$,

$$T^{\mathbf{A}}_{\mathbf{v}}(t)=e^{-i\theta_{\mathbf{v}}L}e^{itv\overline{P_1(\mathbf{v})}}e^{i\theta_{\mathbf{v}}L}. \tag{3.6}$$

Let $\Psi\in L^2(\mathbf{R}^2)$ and set $R(-\theta_{\mathbf{v}})\mathbf{r}=(x(\mathbf{v}),y(\mathbf{v}))$. Then we have, by (3.4),

$$(T^{\mathbf{A}}_{\mathbf{v}}(t)\Psi)(\mathbf{r})=e^{-i\alpha\int_{x(\mathbf{v})}^{x(\mathbf{v})+tv}A_{\mathbf{v}}(x',y(\mathbf{v}))dx'}\Psi(\mathbf{r}+tvR(\theta_{\mathbf{v}})\mathbf{e}_1),\quad \text{a.e. } \mathbf{r}.$$

Noting that $vR(\theta_{\mathbf{v}})\mathbf{e}_1=\mathbf{v}$ and

$$\int_{x(\mathbf{v})}^{x(\mathbf{v})+tv}A_{\mathbf{v}}(x',y(\mathbf{v}))dx'=t\int_0^1\mathbf{v}\cdot\mathbf{A}(\mathbf{r}+\lambda t\mathbf{v})d\lambda=\int_{\mathbf{r}}^{\mathbf{r}+t\mathbf{v}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}',$$

we obtain (3.5). ■

Remark 3.1: We have, for all $t\in\mathbf{R}\backslash\{0\}$ and $\mathbf{v}\in\mathbf{V}(\mathbf{D})$,

$$t\overline{P_{\mathbf{v}}}=\overline{P_{t\mathbf{v}}}, \tag{3.7}$$

which implies that

$$T^{\mathbf{A}}_{\mathbf{v}}(t)=T^{\mathbf{A}}_{t\mathbf{v}}(1). \tag{3.8}$$

Thus, as for the magnetic translations generated by $\overline{P_{\mathbf{v}}}$, it is sufficient to consider the unitary operator,

$$T^{\mathbf{A}}_{\mathbf{v}}:=T^{\mathbf{A}}_{\mathbf{v}}(1). \tag{3.9}$$

B. Commutation relations

Let $\mathbf{v},\mathbf{w}\in\mathbf{V}(\mathbf{D})$ such that $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, and set

$$\begin{aligned}C_{\mathbf{r}}(\mathbf{v},\mathbf{w}):=&\{\mathbf{r}+\lambda\mathbf{v}|0\leqslant\lambda\leqslant1\}\circ\{\mathbf{r}+\mathbf{v}+\lambda\mathbf{w}|0\leqslant\lambda\leqslant1\}\circ\{\mathbf{r}+(1-\lambda)\mathbf{v}+\mathbf{w}|0\leqslant\lambda\leqslant1\}\\&\circ\{\mathbf{r}+(1-\lambda)\mathbf{w}|0\leqslant\lambda\leqslant1\},\end{aligned} \tag{3.10}$$

which is the closed curve starting and ending at the point $\mathbf{r}$, forming the circumference of the parallelogram with vertices $\mathbf{r}$, $\mathbf{r}+\mathbf{v}$, $\mathbf{r}+\mathbf{v}+\mathbf{w}$, $\mathbf{r}+\mathbf{w}$. We introduce

$$\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D}):=\mathbf{R}^2\backslash\cup_{n=1}^{\infty}[L(\mathbf{a}_n;\mathbf{v})\cup(L(\mathbf{a}_n;\mathbf{v})-\mathbf{w})\cup L(\mathbf{a}_n;\mathbf{w})\cup(L(\mathbf{a}_n;\mathbf{w})-\mathbf{v})]. \tag{3.11}$$

Note that, if $\mathbf{r}\in\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$, then $C_{\mathbf{r}}(\mathbf{v},\mathbf{w})$ does not intersect $\mathbf{D}$. We define

$$\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}(\mathbf{r})=\int_{C_{\mathbf{r}}(\mathbf{v},\mathbf{w})}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}', \quad \mathbf{r}\in\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D}), \tag{3.12}$$

which physically means the magnetic flux passing through the interior domain of $C_{\mathbf{r}}(\mathbf{v},\mathbf{w})$. Since the Lebesgue measure of $\mathbf{R}^2\backslash\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$ is zero, the function $\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ defines a unique multiplication self-adjoint operator on $L^2(\mathbf{R}^2)$. We denote it by the same symbol.

Theorem 3.3: *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be as above. Then*

$$T^{\mathbf{A}}_{\mathbf{v}}T^{\mathbf{A}}_{\mathbf{w}}=\exp(-i\alpha\Phi^{A}_{\mathbf{v},\mathbf{w}})T^{\mathbf{A}}_{\mathbf{w}}T^{\mathbf{A}}_{\mathbf{v}}. \tag{3.13}$$

Proof: Using Theorem 3.2, we have, for all $\Psi\in L^2(\mathbf{R}^2)$ and a.e. $\mathbf{r}$,

$$(T^{\mathbf{A}}_{\mathbf{v}}T^{\mathbf{A}}_{\mathbf{w}}\Psi)(\mathbf{r})=e^{-i\alpha\int_{\mathbf{r}}^{\mathbf{r}+\mathbf{v}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}e^{-i\alpha\int_{\mathbf{r}+\mathbf{v}}^{\mathbf{r}+\mathbf{v}+\mathbf{w}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'}\Psi(\mathbf{r}+\mathbf{v}+\mathbf{w}), \tag{3.14}$$

from which (3.13) easily follows. ■

Remark 3.2: Formula (3.13) is a generalization of a formula established in Ref. 1 [see (2.1) in Ref. 1, which corresponds to the case where $\mathbf{v}=\mathbf{e}_1$, $\mathbf{w}=\mathbf{e}_2$, and $\mathbf{A}$ may be singular on a finite discrete set of points].

IV. MAGNETIC TRANSLATIONS ALONG CURVES

In this section we consider magnetic translations along curves in $\mathbf{M}$. *Throughout this section we assume that*

$$\mathbf{V}(\mathbf{D})=\mathbf{R}^2\backslash\{0\}, \tag{4.1}$$

so that, for every vector $\mathbf{v}\in\mathbf{R}^2\backslash\{0\}$, the projected physical momentum operator $\overline{P_{\mathbf{v}}}$ can be defined as a self-adjoint operator.

An example of such $\mathbf{D}$ is given by $\mathbf{D}=\{\mathbf{\Omega}_{m,n}|m\in\mathbf{Z},n\in[-M,M']\cap\mathbf{Z}\}$, where $\mathbf{\Omega}_{m,n}$ is given by (1.17) and $M,M'\in\mathbf{N}$.

Let C be a continuous curve in $\mathbf{M}$ with parametrization $C=\{\mathbf{u}(s)|s\in[a,b]\}(-\infty<a<b<\infty)$. Let $\{s_0,s_1,\ldots,s_n\}$ be a partition of $[a,b]$ $(a=s_0<s_1<\cdots<s_{n-1}<s_n=b)$ such that $\max_{k=1,\ldots,n}(s_k-s_{k-1})\to 0(n\to\infty)$ and $\Delta\mathbf{u}_k:=\mathbf{u}(s_k)-\mathbf{u}(s_{k-1})$, $k=1,\ldots,n$. Then we define

$$U(C):=\mathrm{s}-\lim_{n\to\infty}e^{-i\overline{P}_{\Delta\mathbf{u}_n}}e^{-i\overline{P}_{\Delta\mathbf{u}_{n-1}}}\cdots e^{-i\overline{P}_{\Delta\mathbf{u}_1}}, \tag{4.2}$$

if the right-hand side (rhs) exists, where s-lim means strong limit. It follows that $U(C)$ is unitary. We call $U(C)$ the *magnetic translation along the curve* C.

We introduce

$$\Phi(C):=\int_C\mathbf{A}(\mathbf{r})\cdot d\mathbf{r}, \tag{4.3}$$

the line integral of $\mathbf{A}$ along the curve C. For $\mathbf{r}\in\mathbf{R}^2$ we define a curve $C_{\mathbf{r}}$ by

$$C_{\mathbf{r}}:=C+\mathbf{r}-\mathbf{u}(b). \tag{4.4}$$

We denote by $F(C)$ the multiplication operator by the function: $\mathbf{r}\to e^{i\alpha\Phi(C_{\mathbf{r}})}$:

$$(F(C)\Psi)(\mathbf{r})=e^{i\alpha\Phi(C_{\mathbf{r}})}\Psi(\mathbf{r}), \quad \Psi\in L^2(\mathbf{R}^2), \quad \text{a.e. } \mathbf{r}\in\mathbf{R}^2. \tag{4.5}$$

Remark 4.1: In the case where $\mathbf{A}$ has singularities on $\mathbf{D}$, the function: $\mathbf{r}\to\Phi(C_{\mathbf{r}})$ is originally defined only on $\mathbf{R}^2\backslash\cup_{n\in\mathbf{N}}\{\mathbf{a}_n+\mathbf{u}(b)-\mathbf{u}(s)|s\in[a,b]\}$. But the Lebesgue measure of the set $\cup_{n\in\mathbf{N}}\{\mathbf{a}_n+\mathbf{u}(b)-\mathbf{u}(s)|s\in[a,b]\}$ is zero, so that $e^{i\alpha\Phi(C_{\mathbf{r}})}$ defines a unique multiplication unitary operator on $L^2(\mathbf{R}^2)$.

Theorem 4.1: *The rhs of (4.2) exists and*

$$U(C)=F(C)e^{i\overline{P}_{\mathbf{u}(a)-\mathbf{u}(b)}}. \tag{4.6}$$

Proof: Let

$$U_n(C)=e^{-i\overline{P}_{\Delta\mathbf{u}_n}}e^{-i\overline{P}_{\Delta\mathbf{u}_{n-1}}}\cdots e^{-i\overline{P}_{\Delta\mathbf{u}_1}}.$$

By applying (3.5) repeatedly, we have, for all $\Psi\in L^2(\mathbf{R}^2)$,

$$(U_n(C)\Psi)(\mathbf{r})=e^{i\alpha\Phi_n(\mathbf{r})}(e^{i\overline{P}_{\mathbf{u}(a)-\mathbf{u}(b)}}\Psi)(\mathbf{r}),\quad \text{a.e.}\quad \mathbf{r},$$

with

$$\Phi_n(\mathbf{r})=\sum_{k=1}^{n-1}\int_{\mathbf{r}-\Delta\mathbf{u}_k-\Delta\mathbf{u}_{k+1}-\cdots-\Delta\mathbf{u}_n}^{\mathbf{r}-\Delta\mathbf{u}_{k+1}-\Delta\mathbf{u}_{k+2}-\cdots-\Delta\mathbf{u}_n}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'+\int_{\mathbf{r}-\Delta\mathbf{u}_n}^{\mathbf{r}}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'.$$

Hence

$$\int_{\mathbf{R}^2}|(U_n(C)\Psi)(\mathbf{r})-(F(C)e^{i\overline{P}_{\mathbf{u}(a)-\mathbf{u}(b)}}\Psi)(\mathbf{r})|^2d\mathbf{r}=\int_{\mathbf{R}^2}|e^{i\alpha\Phi_n(\mathbf{r})}-e^{i\alpha\Phi(C_{\mathbf{r}})}|^2|\Psi(\mathbf{r}-\mathbf{u}(b)+\mathbf{u}(a))|^2d\mathbf{r}.$$

It is easy to see that, for all $\mathbf{r}\in\mathbf{R}^2\backslash\cup_{n\in\mathbf{N}}\{\mathbf{a}_n+\mathbf{u}(b)-\mathbf{u}(s)|s\in[a,b]\}$, $\lim_{n\to\infty}\Phi_n(\mathbf{r})=\Phi(C_{\mathbf{r}})$, and

$$|e^{i\alpha\Phi_n(\mathbf{r})}-e^{i\alpha\Phi(C_{\mathbf{r}})}|^2|\Psi(\mathbf{r}-\mathbf{u}(b)+\mathbf{u}(a))|^2\leqslant 4|\Psi(\mathbf{r}-\mathbf{u}(b)+\mathbf{u}(a))|^2.$$

Hence, by the Lebesgue dominated convergence theorem, we have

$$\lim_{n\to\infty}\int_{\mathbf{R}^2}|e^{i\alpha\Phi_n(\mathbf{r})}-e^{i\alpha\Phi(C_{\mathbf{r}})}|^2|\Psi(\mathbf{r}-\mathbf{u}(b)+\mathbf{u}(a))|^2d\mathbf{r}=0.$$

Hence (4.6) follows. ■

Remark 4.2: (1) Suppose that C is continuously differentiable. Then we can show that $U(C)$ is the product integral of the operator-valued function: $s\to -i\overline{P}_{d\mathbf{u}(s)/ds}$ on the interval $[a,b]$ (for the product integral, see Ref. 15).

(2) By (4.6), we have, for all $\Psi\in L^2(\mathbf{R}^2)$,

$$(U(C)\Psi)(\mathbf{r})=e^{i\alpha\Phi(C_{\mathbf{r}})}\Psi(\mathbf{r}+\mathbf{u}(a)-\mathbf{u}(b)),\quad \text{a.e.}\quad \mathbf{r}. \tag{4.7}$$

Let $E=\mathbf{M}\times\mathbf{C}$ be the trivial vector bundle with base space $\mathbf{M}$, fibre $\mathbf{C}$, and structure group $U(1)$ (the one-dimensional unitary group), and $\mathbf{V}_{\mathbf{r}}=\mathbf{C}$ be the fibre at the point $\mathbf{r}\in\mathbf{M}$. Then the rhs of (4.7) geometrically means the parallel transport of the vector $\Psi(\mathbf{r}+\mathbf{u}(a)-\mathbf{u}(b))\in\mathbf{V}_{\mathbf{r}+\mathbf{u}(a)-\mathbf{u}(b)}$ to the point $\mathbf{r}$ along the curve $C_{\mathbf{r}}$ with the connection one-form $-iA:=(-i)(A_1dx+A_2dy)$ (see, e.g., Ref. 16).

As a corollary of Theorem 4.1, we obtain the following.

Theorem 4.2: (i) *We have*

$$U(C)^*=e^{-i\overline{P}_{\mathbf{u}(a)-\mathbf{u}(b)}}F(C)^* \tag{4.8}$$

$$=\operatorname*{s-lim}_{n\to\infty}e^{i\overline{P}_{\Delta\mathbf{u}_1}}e^{i\overline{P}_{\Delta\mathbf{u}_2}}\cdots e^{i\overline{P}_{\Delta\mathbf{u}_n}}. \tag{4.9}$$

(ii) *Let C_1 and C_2 be any continuous curves in $\mathbf{M}$ such that the terminal point of C_1 coincides with the initial point of C_2, so that the composition $C_1\circ C_2$ is also a continuous curve whose initial (resp., terminal) point is that of C_1 (resp., that of C_2). Then*

$$U(C_1)U(C_2)=U(C_1\circ C_2). \tag{4.10}$$

Proof: (i) Equation (4.8) follows from taking the adjoint of (4.6). In general, if unitary operators V_n, V on a Hilbert space satisfy s-$\lim_{n\to\infty} V_n = V$, then s-$\lim_{n\to\infty} V_n^* = V^*$. Applying this fact, we obtain (4.9).

(ii) This easily follows from applying formula (4.6). ■

Let $a \leqslant t \leqslant b$ and

$$C(t) = \{\mathbf{u}(s) | s \in [a,t]\}. \tag{4.11}$$

Then the correspondence $t \to U(C(t))$ gives an operator-valued function from $[a,b]$ into the set of unitary operators on $L^2(\mathbf{R}^2)$. We want to derive a differential equation for $U(C(t))$.

For a Hilbert space $\mathscr{H}$, we denote by $\mathsf{B}(\mathscr{H})$ the set of bounded linear operators on $\mathscr{H}$.

Lemma 4.3: Let $\mathscr{H}$ be a Hilbert space, and V and W be $\mathsf{B}(\mathscr{H})$-valued functions on $[a,b]$. Assume the following (i) and (ii): (i) V is strongly continuous on $\mathscr{H}$; (ii) there exist subspaces $\mathscr{D}_1$ and $\mathscr{D}_2$ of $\mathscr{H}$ such that V and W are strongly differentiable on $\mathscr{D}_1$ and $\mathscr{D}_2$, respectively, and $W(t)\mathscr{D}_2 \subset \mathscr{D}_1$ for all $t \in [a,b]$. Then the $\mathsf{B}(\mathscr{H})$-valued function: $t \to V(t)W(t)$ is strongly differentiable on $\mathscr{D}_2$ and, for all $\psi \in \mathscr{D}_2$,

$$\frac{d}{dt} V(t)W(t)\psi = \frac{dV(t)}{dt} W(t)\psi + V(t) \frac{dW(t)}{dt} \psi, \quad t \in [a,b], \tag{4.12}$$

where $dV(t)/dt$ [resp., $dW(t)/dt$] denotes the strong derivative of $V(t)$ [resp., $W(t)$] on $\mathscr{D}_1$ (resp., $\mathscr{D}_2$).

Proof: Let $X(t) = V(t)W(t)$ and $\psi \in \mathscr{D}_2$. Then we have for all $t \in (a,b)$ and $\epsilon \in \mathbf{R}$ with $|\epsilon|$ sufficiently small,

$$\begin{aligned}
&\frac{X(t+\epsilon)-X(t)}{\epsilon} \psi - \frac{dV(t)}{dt} W(t)\psi - V(t) \frac{dW(t)}{dt} \psi \\
&= V(t+\epsilon)\left\{ \frac{W(t+\epsilon)-W(t)}{\epsilon} \psi - \frac{dW(t)}{dt} \psi \right\} + [V(t+\epsilon)-V(t)] \frac{dW(t)}{dt} \psi \\
&\quad + \left\{ \frac{V(t+\epsilon)-V(t)}{\epsilon} - \frac{dV(t)}{dt} \right\} W(t)\psi.
\end{aligned}$$

By the strong continuity of V and the principle of uniform boundedness, we have $\sup_{t\in[a,b]} \|V(t)\| < \infty$. Using this fact and the assumed properties of V and W, we obtain the desired result. ■

Let

$$\mathbf{M}_C = \mathbf{R}^2 \backslash \cup_{n\in\mathbf{N}} \{\mathbf{a}_n + \mathbf{u}(t) - \mathbf{u}(s) | s,t \in [a,b]\}. \tag{4.13}$$

Theorem 4.4: *Let C be continuously differentiable. Suppose that $\mathbf{M}_C$ is an open set and Hypothesis $(A)_1$ holds. Then the operator-valued function $U(C(\cdot))^*$ on $[a,b]$ is strongly differentiable on $C_0^1(\mathbf{M}_C)$ with*

$$\frac{d}{dt} U(C(t))^* \Psi = iU(C(t))^* P_{d\mathbf{u}(t)/dt} \Psi, \quad t \in [a,b], \quad \Psi \in C_0^1(\mathbf{M}_C). \tag{4.14}$$

Proof (Outline)*:* Let $V(t) = e^{-i\overline{p}_{\mathbf{u}(a)-\mathbf{u}(t)}}$ and $W(t) = F(C(t))^*$. Then we have $U(C(t))^* = V(t)W(t)$. It is not so difficult to show that V is strongly continuous on $L^2(\mathbf{R}^2)$ and strongly differentiable on $C_0^1(\mathbf{R}^2)$ with

$$\frac{dV(t)}{dt} \Psi = iV(t) p_{d\mathbf{u}(t)/dt} \Psi, \quad \Psi \in C_0^1(\mathbf{R}^2)$$

(cf. the proof of Theorem 4.1). Let $\Psi \in C_0^1(M_C)$. Then

$$(W(t)\Psi)(\mathbf{r})=\exp\left(-i\alpha\int_a^t \mathbf{A}(\mathbf{u}(s)-\mathbf{u}(t)+\mathbf{r})\cdot\frac{d\mathbf{u}(s)}{ds}\,ds\right)\Psi(\mathbf{r}).$$

It follows from this formula and the present assumption for $\mathbf{A}$ that $W(t)C_0^1(\mathbf{M}_C)\subset C_0^1(\mathbf{M}_C)$ and $W(t)$ is strongly differentiable on $C_0^1(\mathbf{M}_C)$ with

$$\left(\frac{dW(t)}{dt}\,\Psi\right)(\mathbf{r})=\exp\left(-i\alpha\int_a^t \mathbf{A}(\mathbf{u}(s)-\mathbf{u}(t)+\mathbf{r})\cdot\frac{d\mathbf{u}(s)}{ds}\,ds\right)\left(-i\alpha\,\frac{d\mathbf{u}(t)}{dt}\cdot\mathbf{A}(\mathbf{r})\right.$$
$$\left.+i\alpha\sum_{j,k=1}^{2}\int_a^t\frac{du_j(t)}{dt}\,(\partial_j A_k)(\mathbf{u}(s)-\mathbf{u}(t)+\mathbf{r})\,\frac{du_k(s)}{ds}\,ds\right)\Psi(\mathbf{r}),$$

where $\partial_1:=\partial/\partial x$, $\partial_2:=\partial/\partial y$. On the other hand, we have

$$(ip_{d\mathbf{u}(t)/dt}W(t)\Psi)(\mathbf{r})=\exp\left(-i\alpha\int_a^t \mathbf{A}(\mathbf{u}(s)-\mathbf{u}(t)+\mathbf{r})\cdot\frac{d\mathbf{u}(s)}{ds}\,ds\right)$$
$$\times\left(-i\alpha\sum_{j,k=1}^{2}\int_a^t\frac{du_j(t)}{dt}\,(\partial_j A_k)(\mathbf{u}(s)-\mathbf{u}(t)+\mathbf{r})\,\frac{du_k(s)}{ds}\,ds\right)\Psi(\mathbf{r})$$
$$+(iW(t)p_{d\mathbf{u}(t)/dt}\Psi)(\mathbf{r}).$$

By applying Lemma 4.3, we obtain (4.14). ■

Remark 4.3: (1) It follows from Theorem 4.4 that

$$\frac{d}{dt}\,(\Psi,U(C(t))\Phi)_2=(iP_{d\mathbf{u}(t)/dt}\Psi,U(C(t))\Phi)_2,\quad \Psi\in C_0^1(\mathbf{M}_C),\quad \Phi\in L^2(\mathbf{R}^2),\quad (4.15)$$

where $(\cdot,\cdot)_2$ denotes the inner product of $L^2(\mathbf{R}^2)$. But, in the case where $\mathbf{A}$ has singularities on $\mathbf{D}$, it seems difficult to find a dense subspace $\mathscr{D}$ on which $U(C(t))$ is strongly differentiable and such that $U(C(t))\mathscr{D}\subset D(\overline{P_{d\mathbf{u}(t)/dt}})$.

(2) If $\mathbf{A}$ is differentiable on the whole space $\mathbf{R}^2$, then $U(C(t))$ is strongly differentiable on $C_0^1(\mathbf{R}^2)$, with

$$dU(C(t))\Psi/dt=iP_{d\mathbf{u}(t)/dt}U(C(t))\Psi,\quad \Psi\in C_0^1(\mathbf{R}^2).$$

V. REPRESENTATIONS OF CCR

In this section we show that, associated with the projected physical momentum operators, there exist representations of the CCR with two degrees of freedom and give a complete characterization on the unitary equivalence or inequivalence of the representations to the Schrödinger representation of the CCR with two degrees of freedom. The inequivalent representations correspond to the Aharonov–Bohm effect (Ref. 6). The results of this section include generalizations of those of the previous works (Refs. 1 and 5).

Let $Q_{\mathbf{v},\mathbf{w}}[\mathbf{v},\mathbf{w}\in\mathbf{V}(\mathbf{D})]$ be the self-adjoint multiplication operator given by (1.13). It is easy to see that, if $A_j\in C^1(\mathbf{M})$, $j=1,2$, then the set $\{Q_{\mathbf{v},\mathbf{w}},Q_{\mathbf{w},\mathbf{v}},\overline{P_{\mathbf{v}}},\overline{P_{\mathbf{w}}}\}$ of self-adjoint operators has the following commutation properties: for all $\Psi\in C_0^2(\mathbf{M})$,

$$[Q_{\mathbf{v},\mathbf{w}},\overline{P_{\mathbf{v}}}]\Psi=i\Psi,\quad [Q_{\mathbf{w},\mathbf{v}},\overline{P_{\mathbf{w}}}]\Psi=i\Psi,\quad (5.1)$$

$$[Q_{\mathbf{v},\mathbf{w}},\overline{P_{\mathbf{w}}}]\Psi=0,\quad [Q_{\mathbf{w},\mathbf{v}},\overline{P_{\mathbf{v}}}]\Psi=0,\quad (5.2)$$

$$[Q_{\mathbf{v},\mathbf{w}},Q_{\mathbf{w},\mathbf{v}}]\Psi=0,\quad (5.3)$$

$$[\overline{P_{\mathbf{v}}},\overline{P_{\mathbf{w}}}]\Psi=i\alpha(\mathbf{v}\wedge\mathbf{w})B\Psi.\quad (5.4)$$

Remark 5.1: Let G_d be the algebra generated by elements Q_j, Π_j, F_{jk}, $j=1,...,d$, with identity I, obeying the commutation relations

$$[Q_j,Q_k]=0, \quad [Q_j,\Pi_k]=i\delta_{jk}I, \quad [\Pi_j,\Pi_k]=iF_{jk}, \quad [F_{jk},Q_l]=0, \quad j,k,l=1,...,d.$$

These relations have an origin in a quantum theory of a charged particle in an external electromagnetic field on $\mathbf{R}^d$, where Q_j and Π_j are realized as the position and the physical momentum operators, and may be regarded as an extension or a deformation of the CCR with d degrees of freedom. A special class of representations of G_d is discussed in Ref. 14 in connection with exactly solvable models. Relations (5.1)–(5.4) show that, if $\mathbf{A}$ is in $C^\infty(\mathbf{M})$, then the operators $Q_{\mathbf{v},\mathbf{w}}$, $Q_{\mathbf{w},\mathbf{v}}$, $\overline{P_\mathbf{v}}$, and $\overline{P_\mathbf{w}}$ restricted to the subspace $C_0^\infty(\mathbf{M})$ give a representation of G_2 with the correspondence

$$Q_1\to Q_{\mathbf{v},\mathbf{w}}, \quad Q_2\to Q_{\mathbf{w},\mathbf{v}}, \quad \Pi_1\to\overline{P_\mathbf{v}}, \quad \Pi_2\to\overline{P_\mathbf{w}}, \quad F_{12}\to\alpha(\mathbf{v}\wedge\mathbf{w})B.$$

It would be interesting to classify Hilbert space representations of the algebra $\mathbf{G}_d$.

The following fact easily follows from (5.1)–(5.4).

Proposition 5.1: Assume Hypothesis $(A)_1$. *Then the set* $\pi^\mathbf{A}_{\mathbf{v},\mathbf{w}}$ *given by (1.14) is a representation of the CCR with two degrees of freedom if and only if* $\mathbf{A}$ *is flat on* $\mathbf{M}$.

We denote by $\mathscr{F}(\mathbf{D})$ the set of vector potentials $\mathbf{A}$ that satisfy Hypothesis $(\mathrm{A})_1$ and flat on $\mathbf{M}$.

The following example shows that the set $\mathscr{F}(\mathbf{D})$ is large to some extent.

Example 5.1: Let $a_n=a_{n1}+ia_{n2}\in\mathbf{C}$ $(n\in\mathbf{N})$ be the complex number corresponding to the vector $\mathbf{a}_n$ and $D:=\{a_n\}_{n=1}^\infty$. Without loss of generality, we can assume that $|a_1|\leqslant|a_2|\leqslant|a_3|\leqslant\cdots$. By the assumption for $\mathbf{D}$, we have $a_n\to\infty$ as $n\to\infty$. Let

$$P_n(z)=\frac{c_{n,1}}{z-a_n}+\frac{c_{n,2}}{(z-a_n)^2}+\cdots+\frac{c_{n,k_n}}{(z-a_n)^{k_n}},$$

where $k_n\in\mathbf{N}$ and $c_{n,j}\in\mathbf{C}$ $(j=1,...,k_n)$ are arbitrarily given constants. Then, by the Mittag–Leffler theorem, there exists a meromorphic function $f(z)$ on $\mathbf{C}$ having the following properties: (i) f is holomorphic on $\mathbf{C}\backslash D$; (ii) the principal part of f at $z=a_n$ is given by $P_n(z)$. Let $\mathbf{A}=(A_1,A_2)$ with A_j, $j=1,2$, given by (1.23). Then $A_j\in C^\infty(\mathbf{M})$. Moreover, the Cauchy–Riemann equation for f implies that $\mathbf{A}$ is flat on $\mathbf{M}$ and divergence-free:

$$\partial_xA_1+\partial_yA_2=0, \quad \text{on } \mathbf{M}. \tag{5.5}$$

Hence $\mathbf{A}\in\mathscr{F}(\mathbf{D})$. ■

The special case,

$$\pi^\mathbf{A}_{\mathbf{e}_1,\mathbf{e}_2}=\{L^2(\mathbf{R}^2),C_0^2(\mathbf{M}),\{q_j,\overline{P_j}\}_{j=1}^2\} \quad [\mathbf{A}\in\mathscr{F}(\mathbf{D})],$$

of the above representation of CCR with $\mathbf{D}$ being a finite discrete set has been analyzed in detail (Refs. 1, 2, and 17). Similar analyses can be made in the present general case.

We first consider irreducibility of the representation $\pi^\mathbf{A}_{\mathbf{v},\mathbf{w}}$ with $\mathbf{A}\in\mathscr{F}(\mathbf{D})$. For this purpose, we introduce two kinds of commutants.

A weak commutant of the representation $\pi^\mathbf{A}_{\mathbf{v},\mathbf{w}}$ may be defined by

$$(\pi^\mathbf{A}_{\mathbf{v},\mathbf{w}})':=\{T\in\mathsf{B}(L^2(\mathbf{R}^2))\,|\,(TS\Psi,\Phi)_2=(T\Psi,S\Phi)_2, \quad \text{for all } \Psi,\Phi\in C_0^2(\mathbf{M}) \quad \text{and}$$

$$S=Q_{\mathbf{v},\mathbf{w}},Q_{\mathbf{w},\mathbf{v}},P_\mathbf{v},P_\mathbf{w}\}. \tag{5.6}$$

We set

$$U_{\mathbf{v},\mathbf{w}}(t)=e^{itQ_{\mathbf{v},\mathbf{w}}}, \quad t\in\mathbf{R}. \tag{5.7}$$

Another commutant is associated with the operator algebra generated by

$$\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}} := \{U_{\mathbf{v},\mathbf{w}}(t), U_{\mathbf{w},\mathbf{v}}(t), T^{\mathbf{A}}_{\mathbf{v}}(t), T^{\mathbf{A}}_{\mathbf{w}}(t) | t \in \mathbf{R}\}, \tag{5.8}$$

i.e.,

$$(\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' := \{T \in \mathsf{B}(L^2(\mathbf{R}^2)) | TW = WT, \quad \text{for all} \quad W \in \mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}\}. \tag{5.9}$$

It is easy to see that

$$(\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' \subset (\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})'. \tag{5.10}$$

Theorem 5.2: *Let* $\mathbf{A} \in \mathscr{T}(\mathbf{D})$. *Then* $(\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' = (\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' = \mathbf{C}I$, *where* I *denotes the identity operator on* $L^2(\mathbf{R}^2)$.

Proof: Since $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, q_j (resp., P_j) can be written as a linear combination of $Q_{\mathbf{v},\mathbf{w}}$ and $Q_{\mathbf{w},\mathbf{v}}$ (resp., $P_{\mathbf{v}}$ and $P_{\mathbf{w}}$). Hence, we have $(\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' = (\pi^{\mathbf{A}}_{\mathbf{e}_1,\mathbf{e}_2})'$. On the other hand, in quite the same way as in the proof of Theorem 3.8 of Ref. 17, we can show that $(\pi^{\mathbf{A}}_{\mathbf{e}_1,\mathbf{e}_2})' = \mathbf{C}I$. This result and (5.10) imply $(\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})' = \mathbf{C}I$. ■

We next consider commutation properties of the operators in $\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$. Using Theorem 3.2, we can show that, for all $\mathbf{A}$ satisfying Hypothesis $(\mathrm{A})_0$,

$$U_{\mathbf{v},\mathbf{w}}(t) U_{\mathbf{w},\mathbf{v}}(s) = U_{\mathbf{w},\mathbf{v}}(s) U_{\mathbf{v},\mathbf{w}}(t), \tag{5.11}$$

$$U_{\mathbf{v},\mathbf{w}}(t) T^{\mathbf{A}}_{\mathbf{w}}(s) = T^{\mathbf{A}}_{\mathbf{w}}(s) U_{\mathbf{v},\mathbf{w}}(t), \quad U_{\mathbf{w},\mathbf{v}}(t) T^{\mathbf{A}}_{\mathbf{v}}(s) = T^{\mathbf{A}}_{\mathbf{v}}(s) U_{\mathbf{w},\mathbf{v}}(t), \tag{5.12}$$

$$U_{\mathbf{v},\mathbf{w}}(t) T^{\mathbf{A}}_{\mathbf{v}}(s) = e^{-its} T^{\mathbf{A}}_{\mathbf{v}}(s) U_{\mathbf{v},\mathbf{w}}(t), \quad U_{\mathbf{w},\mathbf{v}}(t) T^{\mathbf{A}}_{\mathbf{w}}(s) = e^{-its} T^{\mathbf{A}}_{\mathbf{w}}(s) U_{\mathbf{w},\mathbf{v}}(t). \tag{5.13}$$

A commutation relation between $T^{\mathbf{A}}_{\mathbf{v}}(s)$ and $T^{\mathbf{A}}_{\mathbf{w}}(t)$ is given by (3.13), with $\mathbf{v}$ and $\mathbf{w}$ replaced by $s\mathbf{v}$ and $t\mathbf{w}$, respectively.

Definition 5.3: We say that the magnetic flux is locally quantized (with respect to the pair $\{\mathbf{v},\mathbf{w}\}$) if the function $\Phi^{\mathbf{A}}_{s\mathbf{v},t\mathbf{w}}$ on $\mathbf{M}_{s\mathbf{v},t\mathbf{w}}(\mathbf{D})$ is $2\pi\mathbf{Z}/\alpha$ valued for all $s,t \in \mathbf{R}$.

Remark 5.2: It is easy to see that, if $A_j \in C^1(\mathbf{M})$, $j = 1,2$, and the magnetic flux is locally quantized, then $\mathbf{A}$ is flat on $\mathbf{M}$.

Theorem 5.4: *For all* $s,t \in \mathbf{R}$, *the set* $\mathscr{W}^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ *of unitary operators satisfies the Weyl relations with two degrees of freedom if and only if the magnetic flux is locally quantized.*

Proof: It follows from Theorem 3.3 and (3.8) that $T^{\mathbf{A}}_{\mathbf{v}}(s)$ and $T^{\mathbf{A}}_{\mathbf{w}}(t)$ commute for all $s,t \in \mathbf{R}$ if and only if the magnetic flux is locally quantized. This fact and (5.11)–(5.13) imply the desired assertion. ■

In the rest of this section, we consider only the case $\mathbf{A} \in \mathscr{T}(\mathbf{D})$; hence $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ *is a* representation of the CCR with two degrees of freedom.

Theorems 5.2 and 5.4 together with the von Neumann uniqueness theorem on the Weyl form of CCR[18,19] give the following result.

Theorem 5.5: *The representation* $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ *is unitarily equivalent to the Schrödinger representation* $\{q_j, p_j\}_{j=1}^2$ *if and only if the magnetic flux is locally quantized.*

Remark 5.3: As in the case of the special representation $\pi^{\mathbf{A}}_{\mathbf{e}_1,\mathbf{e}_2}$ discussed in the previous papers (Refs. 1, 2, and 5), the case where the representation $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is inequivalent to the Schrödinger representation, which, by Theorem 5.4, occurs if and only if the magnetic flux is not locally quantized, physically corresponds to the Aharonov–Bohm effect [see (3.13) and Remark 4.2(2)].

The function $\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ can be explicitly represented, as is shown below. By the assumption for $\mathbf{D}$, we have

$$\delta := \inf_{m \neq n} |\mathbf{a}_n - \mathbf{a}_m| > 0. \tag{5.14}$$

It follows from the flatness of $\mathbf{A}$ and the Green's theorem that, for all $\epsilon \in (0,\delta)$, the line integral,

$$\gamma_{\mathbf{A}}(\mathbf{a}_n) := \int_{|\mathbf{r}-\mathbf{a}_n|=\epsilon} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r}, \tag{5.15}$$

along the circle $\{\epsilon(\cos\theta,\sin\theta)+\mathbf{a}_n|0\leqslant\theta\leqslant 2\pi\}$ is independent of ϵ. It is easy to see that

$$\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}(\mathbf{r})=\sum_{\mathbf{a}_n\in D_{\mathbf{r}}(\mathbf{v},\mathbf{w})}\operatorname{sgn}(\mathbf{v}\wedge\mathbf{w})\,\gamma_{\mathbf{A}}(\mathbf{a}_n),\quad \mathbf{r}\in\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D}),\tag{5.16}$$

where $D_{\mathbf{r}}(\mathbf{v},\mathbf{w})$ is the interior domain of the curve $C_{\mathbf{r}}(\mathbf{v},\mathbf{w})$ and $\operatorname{sgn}(\lambda):=1$ (resp., -1) for $\lambda>0$ (resp., $\lambda<0$). Thus we obtain the following result.

Proposition 5.6: The magnetic flux is locally quantized if and only if $\gamma_{\mathbf{A}}(\mathbf{a}_n)\in 2\pi\mathbf{Z}/\alpha$ for all $n\in\mathbf{N}$.

This proposition and Theorem 5.4 imply the following theorem.

Theorem 5.7: *The representation $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is unitarily equivalent to the Schrödinger representation $\{q_j,p_j\}_{j=1}^2$ if and only if $\gamma_{\mathbf{A}}(\mathbf{a}_n)\in 2\pi\mathbf{Z}/\alpha$ for all $n\in\mathbf{N}$.*

Example 5.2: Let $\mathbf{A}$ be the vector potential given in Example 5.1. Then

$$\gamma_{\mathbf{A}}(\mathbf{a}_n)=2\pi\Re c_{n,1}.$$

Hence, if $\Re c_{n_0,1}\notin\mathbf{Z}/\alpha$ for some $n_0\in\mathbf{N}$, then $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is unitarily *inequivalent* to the Schrödinger representation $\{q_j,p_j\}_{j=1}^2$. Thus, there exist lots of inequivalent representations of CCR. ■

In concluding this section, we make a remark on the unitary inequivalence between the two representations $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ and $\pi^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'}$ [$\mathbf{A}'\in\mathscr{F}(\mathbf{D})$ and $\mathbf{v}',\mathbf{w}'\in\mathbf{V}(\mathbf{D})$ are linearly independent]. We introduce a 2×2 matrix:

$$K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')=\frac{1}{\mathbf{v}'\wedge\mathbf{w}'}\begin{pmatrix} v_1w_2'-w_1v_2' & w_1v_1'-v_1w_1' \\ v_2w_2'-w_2v_2' & w_2v_1'-v_2w_1'\end{pmatrix}.\tag{5.17}$$

It is easy to see that $K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')$ is bijective, with

$$\det K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')=\frac{\mathbf{v}\wedge\mathbf{w}}{\mathbf{v}'\wedge\mathbf{w}'},\tag{5.18}$$

$$K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')K(\mathbf{v}',\mathbf{w}',\mathbf{v},\mathbf{w})=I.\tag{5.19}$$

Proposition 5.8: Suppose that $\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is unitarily equivalent to $\pi^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'}$. Then, for all $s,t\in\mathbf{R}$,

$$\exp(-i\alpha\Phi^{\mathbf{A}}_{s\mathbf{v},t\mathbf{w}}(K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')\mathbf{r}))=\exp(-i\alpha\Phi^{\mathbf{A}'}_{s\mathbf{v}',t\mathbf{w}'}(\mathbf{r})),\quad\text{a.e. }\mathbf{r}\in\mathbf{R}^2.\tag{5.20}$$

Proof: To make explicit the dependence of $P_{\mathbf{v}}$ on $\mathbf{A}$, we write $\overline{P}_{\mathbf{v}}=\overline{P}_{\mathbf{v}}(\mathbf{A})$. By the present assumption, there exists a unitary operator U on $L^2(\mathbf{R}^2)$, such that

$$UQ_{\mathbf{v},\mathbf{w}}U^{-1}=Q_{\mathbf{v}',\mathbf{w}'},\quad UQ_{\mathbf{w},\mathbf{v}}U^{-1}=Q_{\mathbf{w}',\mathbf{v}'},\tag{5.21}$$

$$U\overline{P}_{\mathbf{v}}(\mathbf{A})U^{-1}=\overline{P}_{\mathbf{v}'}(\mathbf{A}'),\quad U\overline{P}_{\mathbf{w}}(\mathbf{A})U^{-1}=\overline{P}_{\mathbf{w}'}(\mathbf{A}').\tag{5.22}$$

By Theorem 3.3 and (5.22), together with functional calculus of self-adjoint operators, we have for all $s,t\in\mathbf{R}$,

$$\exp(-i\alpha U\Phi^{\mathbf{A}}_{t\mathbf{v},s\mathbf{w}}U^{-1})=\exp(-i\alpha\Phi^{\mathbf{A}'}_{t\mathbf{v}',s\mathbf{w}'}).\tag{5.23}$$

It is easy to show that

$$q_1=v_1Q_{\mathbf{v},\mathbf{w}}+w_1Q_{\mathbf{w},\mathbf{v}},\tag{5.24}$$

$$q_2=v_2Q_{\mathbf{v},\mathbf{w}}+w_2Q_{\mathbf{w},\mathbf{v}}.\tag{5.25}$$

These relations and (5.21) imply that

$$U\mathbf{q}U^{-1}=K(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')\mathbf{q}. \tag{5.26}$$

Hence, by functional calculus, the self-adjoint operator $U\Phi^{\mathbf{A}}_{\mathbf{v},s\mathbf{w}}U^{-1}$ is equal to the multiplication operator by the function $\Phi^{\mathbf{A}}_{t\mathbf{v},s\mathbf{w}}(\mathbf{K}(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}')\mathbf{r})$. This fact and (5.23) imply (5.20). ■

Remark 5.4: (1) There exist many triples $\{\mathbf{v},\mathbf{w},\mathbf{A}\}$ and $\{\mathbf{v}',\mathbf{w}',\mathbf{A}'\}$, where $\mathbf{A},\mathbf{A}'\in\mathscr{T}(\mathbf{D})$, such that (5.20) does not hold. Hence, there exist many pairs $\{\pi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}},\pi^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'}\}$ of representations that are unitarily inequivalent to each other.

(2) In the case $\mathbf{v}=\mathbf{v}'$, $\mathbf{w}=\mathbf{w}'$, we have

$$K(\mathbf{v},\mathbf{w},\mathbf{v},\mathbf{w})=I, \tag{5.27}$$

so that (5.20) becomes

$$\exp(-i\alpha\Phi^{\mathbf{A}}_{s\mathbf{v},t\mathbf{w}}(\mathbf{r}))=\exp(-i\alpha\Phi^{\mathbf{A}'}_{s\mathbf{v},t\mathbf{w}}(\mathbf{r})),\quad \text{a.e.}\quad \mathbf{r}\in\mathbf{R}^2. \tag{5.28}$$

VI. QUANTUM ALGEBRAIC STRUCTURES

In this section we discuss quantum algebraic structures associated with continuous magnetic translations $\{T^{\mathbf{A}}_{\mathbf{v}}\}_{\mathbf{v}\in\mathbf{V(D)}}$. We first introduce a special class of vector potentials.

Definition 6.1: We say that a vector potential $\mathbf{A}$ is in $\mathscr{A}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$ if the function $\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is equal to a constant on $\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$.

Example 6.1: Constant magnetic fields. Let $B_0\in\mathbf{R}$ be a constant and

$$A_1(\mathbf{r})=-\frac{B_0y}{2},\quad A_2(\mathbf{r})=\frac{B_0x}{2}.$$

Then $B(\mathbf{r})=B_0$, i.e., the magnetic field is uniformly constant. Hence, by the Green's theorem, we have

$$\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}(\mathbf{r})=(\mathbf{v}\wedge\mathbf{w})B_0,\quad \mathbf{r}\in\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D}).$$

Thus $\mathbf{A}\in\mathscr{A}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$. ■

Example 6.2: *Vector potentials singular on an infinite lattice.* Let $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ be the infinite lattice given by (1.16) and consider the case $\mathbf{D}=\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ and $\mathbf{v}=\boldsymbol{\omega}_1$, $\mathbf{w}=\boldsymbol{\omega}_2$. Let $\Omega_{m,n}$ be as in (1.22) and $f(z)$ be a meromorphic function on $\mathbf{C}$ with the following properties: (i) f is holomorphic on $\mathbf{C}\backslash\{\Omega_{m,n}\}_{m,n\in\mathbf{Z}}$; (ii) the principal part of f at $z=\Omega_{m,n}$ is of the form

$$P_{m,n}(z)=\frac{c}{z-\Omega_{m,n}}+\frac{c^{(2)}_{m,n}}{(z-\Omega_{m,n})}+\cdots+\frac{c^{(k_{m,n})}_{m,n}}{(z-\Omega_{m,n})^{k_{m,n}}},$$

where $k_{m,n}\in\mathbf{N},c,c^{(j)}_{m,n}$ are constants (c is independent of m, n). The existence of such a function f is ensured by the Mittag–Leffler theorem (see Example 5.1). As in Example 5.1, we define a vector potential $\mathbf{A}=(A_1,A_2)$ by (1.23). Then $\mathbf{A}\in\mathscr{T}(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ with A_j being infinitely many times differentiable on the open set,

$$\mathbf{M}_{\mathbf{L}}:=\mathbf{R}^2\backslash\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2). \tag{6.1}$$

As in Example 5.2, we have

$$\gamma_{\mathbf{A}}(\boldsymbol{\Omega}_{m,n})=2\pi\Re c,\quad m,n\in\mathbf{Z}. \tag{6.2}$$

For all $\mathbf{r}\in\mathbf{M}_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$, $D_{\mathbf{r}}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ contains only one point in $\{\boldsymbol{\Omega}_{m,n}\}_{m,n\in\mathbf{Z}}$. Hence, noting that $\mathrm{sgn}(\boldsymbol{\omega}_1\wedge\boldsymbol{\omega}_2)=1$, we have, from (5.16) and (6.2),

$$\Phi^{\mathbf{A}}_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{r})=2\pi\Re c. \tag{6.3}$$

Thus, $\mathbf{A}\in\mathscr{A}_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$. ■

We denote by $\mathscr{R}_{\mathbf{v},\mathbf{w},\mathbf{A}}$ the algebra generated by $T^{\mathbf{A}}_{\mathbf{v}}$ and $T^{\mathbf{A}}_{\mathbf{w}}$.

Let $\mathbf{A}\in\mathscr{A}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$. Then, by definition, $\Phi^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ is a constant on $\mathbf{M}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$. We denote the constant by $C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$. By Theorem 3.3, we have

$$T^{\mathbf{A}}_{\mathbf{v}}T^{\mathbf{A}}_{\mathbf{w}}=e^{-i\alpha C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}}T^{\mathbf{A}}_{\mathbf{w}}T^{\mathbf{A}}_{\mathbf{v}}. \tag{6.4}$$

Hence $\mathscr{R}_{\mathbf{v},\mathbf{w},\mathbf{A}}$ is a representation on $L^2(\mathbf{R}^2)$ of a rotation algebra (e.g., Ref. 20, Chap. VI) or the quantum plane with the deformation parameter $q=e^{-i\alpha C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}}$ (Ref. 21).

Proposition 6.2: There exist no nontrivial finite-dimensional subspaces of $L^2(\mathbf{R}^2)$ *left invariant by* $\mathscr{R}_{\mathbf{v},\mathbf{w},\mathbf{A}}$.

Proof: This follows from Proposition 2.6. ■

Proposition 6.3: Let $\mathbf{A}\in\mathscr{A}_{\mathbf{v},\mathbf{w}}(\mathbf{D})$ and $\mathbf{A}'\in\mathscr{A}_{\mathbf{v}',\mathbf{w}'}(\mathbf{D})$. *Suppose that* $C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}-C^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'}\notin 2\pi\mathbf{Z}/\alpha$. *Then* $\mathscr{R}_{\mathbf{v},\mathbf{w},\mathbf{A}}$ *and* $\mathscr{R}_{\mathbf{v}',\mathbf{w}',\mathbf{A}'}$ *are unitarily inequivalent.*

Proof: Suppose that there exists a unitary operator U on $L^2(\mathbf{R}^2)$ such that $UT^{\mathbf{A}}_{\mathbf{v}}U^{-1}=T^{\mathbf{A}'}_{\mathbf{v}'}$, $UT^{\mathbf{A}}_{\mathbf{w}}U^{-1}=T^{\mathbf{A}'}_{\mathbf{w}'}$. By (6.4) and the fact that $C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}$ and $C^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'}$ are constants, we have $\exp(-i\alpha C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}})=\exp(-i\alpha C^{\mathbf{A}'}_{\mathbf{v}',\mathbf{w}'})$, which is equivalent to that $C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}-C^{\mathbf{A}'}_{\mathbf{v},\mathbf{w}'}\in 2\pi\mathbf{Z}/\alpha$. Thus, the desired assertion follows. ■

Example 6.3: Let $\mathbf{A}$ be the vector potential given in Example 6.2. Then, by (6.3), we have

$$T^{\mathbf{A}}_{\boldsymbol{\omega}_1}T^{\mathbf{A}}_{\boldsymbol{\omega}_2}=e^{-2\pi i\alpha\Re c}T^{\mathbf{A}}_{\boldsymbol{\omega}_2}T^{\mathbf{A}}_{\boldsymbol{\omega}_1}. \tag{6.5}$$

Hence we have a one-parameter family $\mathscr{R}_c:=\mathscr{R}_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\mathbf{A}}$ of representations [on $L^2(\mathbf{R}^2)$] of rotation algebras (quantum planes). It follows from Proposition 6.3 that, if $\Re(c-c')\notin\mathbf{Z}/\alpha$, then the two representations $\mathscr{R}_c$ and $\mathscr{R}_{c'}$ are unitarily inequivalent to each other. Thus, associated with singular magnetic fields concentrated on $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$, there exist infinitely many representations on $L^2(\mathbf{R}^2)$ of rotation algebras inequivalent to each other. ■

Remark 6.1: The same consideration as above applies to the algebra $\mathscr{R}^*_{\mathbf{v},\mathbf{w},\mathbf{A}}$ generated by $(T^{\mathbf{A}}_{\mathbf{v}})^*$ and $(T^{\mathbf{A}}_{\mathbf{w}})^*$, which is a representation of a rotation algebra. We can also consider the algebra generated by $T^{\mathbf{A}}_{\mathbf{v}}$, $T^{\mathbf{A}}_{\mathbf{w}}$, $(T^{\mathbf{A}}_{\mathbf{v}})^*$ and $(T^{\mathbf{A}}_{\mathbf{w}})^*$, which is a $*$-subalgebra of $\mathbf{B}(L^2(\mathbf{R}^2))$.

As is shown in Ref. 5, for any representation of a quantum plane where the generators are represented as bijections, one can construct a representation of the quantum group $U_q(sl_2)$. We can apply this method to the present case to construct, from $\mathscr{R}_{\mathbf{v},\mathbf{w},\mathbf{A}}$, representations of $U_q(sl_2)$ on $L^2(\mathbf{R}^2)$ with $q=e^{-i\alpha C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}}$ or $q=e^{-i\alpha C^{\mathbf{A}}_{\mathbf{v},\mathbf{w}}/2}$ and analyze them in quite the same way as in Ref. 5. Here we only mention a basic feature of those representations: they have no weight vectors and hence no nontrivial finite-dimensional reductions (cf. Proposition 6.2). Note that this makes a big difference from the case of representations of $U_q(sl_2)$ constructed in terms of discrete magnetic translations on a lattice (Ref. 12), where finite-dimensional representations of $U_q(sl_2)$ appear (cf. also Ref. 13).

VII. REDUCTION TO LATTICE QUANTUM SYSTEMS

In this section we focus our attention on the case $\mathbf{D}=\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ and consider the problem of reduction of the continuous magnetic translations,

$$T_j:=T^{\mathbf{A}}_{\boldsymbol{\omega}_j},\quad j=1,2, \tag{7.1}$$

to the closed subspace $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$, which is given by (1.19) and naturally identified with the Hilbert space $l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$.

A. General aspects

We denote by $\mathbf{S}^{(1)}_n$ (resp., $\mathbf{S}^{(2)}_m$) the open set between the straight lines $\{n\boldsymbol{\omega}_2+s\boldsymbol{\omega}_1|s\in\mathbf{R}\}$ (resp., $\{m\boldsymbol{\omega}_1+s\boldsymbol{\omega}_2|s\in\mathbf{R}\}$) and $\{(n+1)\boldsymbol{\omega}_2+s\boldsymbol{\omega}_1|s\in\mathbf{R}\}$ (resp., $\{(m+1)\boldsymbol{\omega}_1+s\boldsymbol{\omega}_2|s\in\mathbf{R}\}$).

Let

$$\phi^{\mathbf{A}}_j(\mathbf{r}):=\int_{\mathbf{r}}^{\mathbf{r}+\boldsymbol{\omega}_j}\mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}',\quad j=1,2. \tag{7.2}$$

Definition 7.1: We say that $\mathbf{A}$ is in the class $\mathscr{C}_j$ $(j=1,2)$ if, for each $n\in\mathbf{Z}$, $\phi_j^{\mathbf{A}}$ is constant on $\mathbf{S}_n^{(j)}$, i.e., there exists a constant $c_j(n)\in\mathbf{R}$, such that

$$\phi_j^{\mathbf{A}}(\mathbf{r})=c_j(n),\quad \mathbf{r}\in\mathbf{S}_n^{(j)},\quad n\in\mathbf{Z}. \tag{7.3}$$

Remark 7.1: It is easy to see that, if $\mathbf{A}\in\mathscr{C}_j$, then the function: $\mathbf{r}\to\mathbf{A}(\mathbf{r})\cdot\boldsymbol{\omega}_j$ is periodic on $\cup_{n\in\mathbf{Z}}\mathbf{S}_n^{(j)}$ with period $\boldsymbol{\omega}_j$. But the converse is not true.

Proposition 7.2: For each $j=1,2$, T_j and T_j^{-1} leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant if and only if $\mathbf{A}\in\mathscr{C}_j$.

Proof: By Theorem 3.2, we have, for all $\Psi\in L^2(\mathbf{R}^2)$,

$$(T_j\Psi)(\mathbf{r})=e^{-i\alpha\phi_j^{\mathbf{A}}(\mathbf{r})}(e^{i\overline{p\boldsymbol{\omega}_j}}\Psi)(\mathbf{r}),\quad \text{a.e.}\quad \mathbf{r}\in\mathbf{S}_n^{(j)}.$$

Note that $e^{\pm i\overline{p\boldsymbol{\omega}_j}}$ leave $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant. Hence, T_j leaves $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$ invariant if and only if $\phi_j^{\mathbf{A}}(r)$ is constant on $\mathbf{S}_{m,n}=\mathbf{S}_n^{(1)}\cap\mathbf{S}_m^{(2)}$ for all $m,n\in\mathbf{Z}$. Since $\mathbf{S}_n^{(j)}$ is connected and $\phi_j^{\mathbf{A}}$ is continuous on $\mathbf{S}_n^{(j)}$, it follows that $\phi_j^{\mathbf{A}}(\mathbf{r})$ is constant on $\mathbf{S}_{m,n}$ for all $m,n\in\mathbf{Z}$ if and only if it is constant on $\mathbf{S}_n^{(j)}$ for all $n\in\mathbf{Z}$. Thus, we obtain the desired result. ■

We set

$$\mathscr{C}=\mathscr{C}_1\cap\mathscr{C}_2. \tag{7.4}$$

As a corollary of Proposition 7.2, we have the following.

Theorem 7.3: *The four unitary operators T_1, T_1^{-1}, T_2 and T_2^{-1} leave $L^2_{\omega_1,\omega_2}(\mathbf{R}^2)$ invariant if and only if $\mathbf{A}\in\mathscr{C}$.*

By this theorem, if $\mathbf{A}\in\mathscr{C}$, then T_1, T_1^{-1}, T_2 and T_2^{-1} are reduced by $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)$, and the restrictions

$$\hat{T}_j:=T_j\lceil L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2),\quad \widehat{T_j^{-1}}:=T_j^{-1}\lceil L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2),\quad j=1,2, \tag{7.5}$$

induce magnetic translations on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ with $(\hat{T}_j)^{-1}=\widehat{T_j^{-1}}$. But, if these reduced magnetic translations have the trivial holonomy, i.e., $\hat{T}_1\hat{T}_2\hat{T}_1^{-1}\hat{T}_2^{-1}=I$, then they are uninteresting. Thus, we need to find conditions for the magnetic translations to have a nontrivial holonomy. The following proposition gives a necessary condition for that.

Proposition 7.4: Let $\mathbf{A}\in\mathscr{C}$. Suppose that A_1 and A_2 are continuous on $\mathbf{R}^2$. Then $T_1T_2=T_2T_1$.

Proof: By the condition $\mathbf{A}\in\mathscr{C}$, we have, for all $\mathbf{r}\in\mathbf{S}_{m,n}$,

$$\Phi^{\mathbf{A}}_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{r})=\phi_1^{\mathbf{A}}(\mathbf{r})+\phi_2^{\mathbf{A}}(\mathbf{r}+\boldsymbol{\omega}_1)-\phi_1^{\mathbf{A}}(\mathbf{r}+\boldsymbol{\omega}_2)-\phi_2^{\mathbf{A}}(\mathbf{r})=c_1(n)+c_2(m+1)-c_1(n+1)-c_2(m).$$

Since A_j is continuous on $\mathbf{R}^2$, so is $\phi_j^{\mathbf{A}}$. This implies that $c_j(n)=c_j(n+1)$ for all $n\in\mathbf{Z}$ and $j=1,2$. Hence $\Phi^A_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}=0$. By this fact and Theorem 3.3 we obtain the desired result. ■

Proposition 7.4 shows that, in the case $\mathbf{A}\in\mathscr{C}$, for the magnetic translations $\hat{T}_j,\hat{T}_j^{-1}$, $j=1,2$, to have a nontrivial holonomy, it is necessary for $\mathbf{A}$ to have singularities in $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$.

We next show that, for a given vector potential $\mathbf{A}$ satisfying certain properties, we can construct an element in $\mathscr{C}$.

Definition 7.5: We say that $\mathbf{A}$ is in $\mathcal{D}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ if there exist real sequences $\{\nu_j(n)\}_{n\in\mathbf{Z}}$, $j=1,2$, and real-valued continuous functions F_j on $\cup_{n\in\mathbf{Z}}\mathbf{S}_n^{(j)}$, such that

$$\phi_j^{\mathbf{A}}(\mathbf{r})+\int_0^1 F_j(\mathbf{r}+s\boldsymbol{\omega}_j)ds=\nu_j(n),\quad \mathbf{r}\in\mathbf{S}_n^{(j)},\quad j=1,2. \tag{7.6}$$

It is obvious that $\mathscr{C}\subset\mathcal{D}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ [note that every element $\mathbf{A}\in\mathscr{C}$ satisfies (7.6) with $F_j=0$, $j=1.2$].

The matrix

$$W:=\begin{pmatrix}\omega_{11} & \omega_{12}\\ \omega_{21} & \omega_{22}\end{pmatrix} \tag{7.7}$$

is regular, since $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are linearly independent.

Proposition 7.6: Suppose that $\mathbf{A}\in\mathcal{D}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ *and let* $\nu_j(n)$, F_j *be as in Definition 7.5. Let*

$$\mathbf{F}:=(F_1,F_2), \tag{7.8}$$

$$\widetilde{\mathbf{A}}:=\mathbf{A}+W^{-1}\mathbf{F}. \tag{7.9}$$

Then $\widetilde{\mathbf{A}}\in\mathscr{C}$ *with*

$$\phi_j^{\widetilde{\mathbf{A}}}(\mathbf{r})=\nu_j(n),\quad \mathbf{r}\in\mathbf{S}_n^{(j)},\quad j=1,2. \tag{7.10}$$

Proof: We have, for all $\mathbf{r}\in\mathbf{S}_n^{(j)}$,

$$\phi_j^{\widetilde{\mathbf{A}}}(\mathbf{r})=\phi_j^{\mathbf{A}}(\mathbf{r})+\int_0^1 W^{-1}\mathbf{F}(\mathbf{r}+s\boldsymbol{\omega}_j)\cdot\boldsymbol{\omega}_j ds.$$

Since ${}^t(W^{-1})\boldsymbol{\omega}_j=\mathbf{e}_j$, we have $W^{-1}\mathbf{F}(\mathbf{r}+s\boldsymbol{\omega}_j)\cdot\boldsymbol{\omega}_j=F_j(\mathbf{r}+s\boldsymbol{\omega}_j)$. Hence, (7.10) follows. ■

Remark 7.2: The modified vector potential $\widetilde{\mathbf{A}}$ is not necessarily flat on $\mathbf{M}_\mathbf{L}$, even if $\mathbf{A}$ is flat on $\mathbf{M}_\mathbf{L}$. See the next section.

B. Construction of a subset of $\mathscr{C}$ from quasiperiodic functions

Definition 7.7: Let $f(z)$ be a holomorphic function on $\mathbf{C}\backslash\{\Omega_{m,n}\}_{m,n\in\mathbf{Z}}$ with possible poles at $z=\Omega_{m,n}$, $m,n\in\mathbf{Z}$. We say that f is in the class $\mathscr{C}$ if there exist constants $\xi_j:=\xi_{j1}+i\xi_{j2}\in\mathbf{C}$, $j=1,2$, such that

$$f(z+\omega_j)=f(z)+\xi_j,\quad j=1,2,\quad z\in\mathbf{C}\backslash\{\Omega_{m,n}\}_{m,n\in\mathbf{Z}}. \tag{7.11}$$

Remark 7.3: Every function $f\in\mathscr{C}$ is a primitive function of an elliptic function with periods ω_j, $j=1,2$. The constant ξ_j on the rhs of (7.11) is given by, e.g.,

$$\xi_j=f\left(\frac{\omega_j}{2}\right)-f\left(-\frac{\omega_j}{2}\right).$$

Let $f\in\mathscr{C}$ and A_j, $j=1,2$, be defined from f by (1.23). Set $\mathbf{A}=(A_1,A_2)$. Then the Cauchy–Riemann equation implies that $\mathbf{A}$ is flat on $\mathbf{M}_\mathbf{L}$ and

$$\partial_x A_1+\partial_y A_2=0,\quad \text{on}\quad \mathbf{M}_\mathbf{L}. \tag{7.12}$$

Let ξ_j be as in (7.11) and

$$\boldsymbol{\xi}_j:=(\xi_{j1},\xi_{j2}),\quad \boldsymbol{\xi}_j':=(\xi_{j2},\xi_{j1}),\quad j=1,2. \tag{7.13}$$

Proposition 7.8: For all $n\in\mathbf{Z}$, *there exist constants* $c_j(n)$, $j=1,2$, *such that*

$$\phi_j^{\mathbf{A}}(\mathbf{r})=\mathbf{r}\cdot\boldsymbol{\xi}_j'+c_j(n),\quad \mathbf{r}\in\mathbf{S}_n^{(j)}. \tag{7.14}$$

Proof: Let ω_j be as in (1.21) and $\theta_j:=\arg\omega_j$. Let $R(\theta)$ be defined by (2.10) and let

$$g_j(\mathbf{r}):=A(R(\theta_j)\mathbf{r})\cdot\boldsymbol{\omega}_j,\quad \phi_j'(\mathbf{r}):=\phi_j^{\mathbf{A}}(R(\theta_j)\mathbf{r}).$$

Then we have

$$\phi_j'(\mathbf{r})=\int_0^1 g_j(\mathbf{r}+s\boldsymbol{\omega}_j')ds,$$

where $\boldsymbol{\omega}_j' = |\omega_j|\mathbf{e}_1$. On the connected domain $R(-\theta_j)\mathbf{S}_n^{(j)}$, $g_j(\cdot + s\boldsymbol{\omega}_j')$, and ϕ_j' are infinitely many times differentiable with

$$\partial_x \phi_j'(\mathbf{r}) = \int_0^1 \partial_x g_j(\mathbf{r} + s\boldsymbol{\omega}_j') ds, \quad \mathbf{r} \in R(-\theta_j)\mathbf{S}_n^{(j)}.$$

It is easy to see that

$$\partial_x g_j(\mathbf{r} + s\boldsymbol{\omega}_j') = |\omega_j|^{-1} \frac{d}{ds} g_j(\mathbf{r} + s\boldsymbol{\omega}_j').$$

Hence

$$\partial_x \phi_j'(\mathbf{r}) = \frac{g_j(\mathbf{r} + \boldsymbol{\omega}_j') - g_j(\mathbf{r})}{|\omega_j|}.$$

On the other hand, we have, by (7.11) and (1.23),

$$\mathbf{A}(\mathbf{r} + \boldsymbol{\omega}_j) - \mathbf{A}(\mathbf{r}) = \boldsymbol{\xi}_j', \quad \mathbf{r} \in \mathbf{M}_\mathbf{L}. \tag{7.15}$$

Thus, we obtain

$$\partial_x \phi_j'(\mathbf{r}) = \frac{\boldsymbol{\omega}_j \cdot \boldsymbol{\xi}_j'}{|\omega_j|}, \quad \mathbf{r} \in U_{n \in \mathbf{Z}} R(-\theta_j)\mathbf{S}_n^{(j)}. \tag{7.16}$$

By the flatness of $\mathbf{A}$ on $\mathbf{M}_\mathbf{L}$ and (7.12), we can show that

$$\partial_y g_j(\mathbf{r}) = \partial_x \boldsymbol{\omega}_j \wedge \mathbf{A}(\mathbf{R}(\theta_j)\mathbf{r}).$$

Hence, in the same way as above, we obtain

$$\partial_y \phi_j'(\mathbf{r}) = \frac{\boldsymbol{\omega}_j \wedge \boldsymbol{\xi}_j'}{|\omega_j|}, \quad \mathbf{r} \in U_{n \in \mathbf{Z}} R(-\theta_j)\mathbf{S}_n^{(j)}. \tag{7.17}$$

It follows from (7.16) and (7.17) that there exist constants $c_j(n)$, $j = 1,2$, $n \in \mathbf{Z}$, such that

$$\phi_j^{\mathbf{A}}(\mathbf{r}) = \left[R(\theta_j) \left(\frac{\boldsymbol{\omega}_j \cdot \boldsymbol{\xi}_j'}{|\omega_j|}, \quad \frac{\boldsymbol{\omega}_j \wedge \boldsymbol{\xi}_j'}{|\omega_j|} \right) \right] \cdot \mathbf{r} + c_j(n), \quad \mathbf{r} \in \mathbf{S}_n^{(j)}.$$

The rhs of this equation is equal to the rhs of (7.14). ■

Theorem 7.9: *Let*

$$h_j(\mathbf{r}) := \left(\frac{\boldsymbol{\omega}_j}{2} - \mathbf{r} \right) \cdot \boldsymbol{\xi}_j', \quad \mathbf{r} \in \mathbf{R}^2, \quad j = 1,2, \quad \mathbf{h} = (h_1, h_2), \tag{7.18}$$

and

$$\tilde{\mathbf{A}} := \mathbf{A} + W^{-1}\mathbf{h}. \tag{7.19}$$

Then $\tilde{\mathbf{A}}$ *is in* $\mathscr{C}$ *with*

$$\phi_j^{\tilde{\mathbf{A}}}(\mathbf{r}) = c_j(n), \quad \mathbf{r} \in \mathbf{S}_n^{(j)}, \quad j = 1,2, \quad n \in \mathbf{Z}. \tag{7.20}$$

Moreover,

$$(T_{\boldsymbol{\omega}_j}^{\tilde{\mathbf{A}}})(\mathbf{r}) = e^{-i\alpha c_j(n)} \Psi(\mathbf{r} + \boldsymbol{\omega}_j), \quad \mathbf{r} \in \mathbf{S}_n^{(j)}, \quad j = 1,2, \quad n \in \mathbf{Z}. \tag{7.21}$$

Proof: By Proposition 7.8, (7.6) holds with $F_j = h_j$ and $\nu_j(n) = c_j(n)$. Hence, by Proposition 7.6, the desired assertion follows. Formula (7.21) follows from (3.5) and (7.20). ■

Theorem 7.9 establishes the existence of a mapping from $\mathscr{E}$ to a subset of $\mathscr{C}$ by the correspondence $f \to \widetilde{\mathbf{A}}$ defined by (7.19).

It is interesting to see when $\widetilde{\mathbf{A}}$ is flat on $\mathbf{M_L}$.

Lemma 7.10: The vector potential $\widetilde{\mathbf{A}}$ given by (7.19) is flat on $\mathbf{M_L}$ if and only if

$$\boldsymbol{\xi}_1' \cdot \boldsymbol{\omega}_2 = \boldsymbol{\xi}_2' \cdot \boldsymbol{\omega}_1 . \tag{7.22}$$

Proof: The magnetic field,

$$\widetilde{B}(\mathbf{r}) := D_x \widetilde{A}_2 - D_y \widetilde{A}_1 , \tag{7.23}$$

of the vector potential $\widetilde{\mathbf{A}}$ is constant, with

$$\widetilde{B}(\mathbf{r}) = \frac{\boldsymbol{\xi}_1' \cdot \boldsymbol{\omega}_2 - \boldsymbol{\xi}_2' \cdot \boldsymbol{\omega}_1}{\det W}, \quad \mathbf{r} \in \mathbf{M_L}. \tag{7.24}$$

Thus, the desired assertion follows. ■

We note the following fact.

Lemma 7.11: Let r_f be the residue of f at $z = \Omega_{m,n}$. Then r_f is independent of $m, n \in \mathbf{Z}$ and obeys the relation

$$\xi_1 \omega_2 - \xi_2 \omega_1 = 2\pi i r_f . \tag{7.25}$$

Proof: Let $a = (m - \frac{1}{2})\omega_1 + (n - \frac{1}{2})\omega_2$ and $C_{m,n}$ be the contour formed by the edges of the cell whose corners are $a, a + \omega_1, a + \omega_1 + \omega_2, a + \omega_2$, where the orientation of $C_{m,n}$ is taken to be anticlockwise. Then we have

$$\begin{aligned} 2\pi i r_f = \int_{C_{m,n}} f(z)dz &= \int_a^{a+\omega_1} \{f(z) - f(z+\omega_2)\}dz - \int_a^{a+\omega_2} \{f(z) - f(z+\omega_1)\}dz \\ &= -\xi_2 \omega_1 + \xi_1 \omega_2 , \end{aligned}$$

where we have used the quasiperiodicity (7.11) of f. Thus (7.25) is obtained. ■

The following proposition gives a complete characterization on the flatness of $\widetilde{\mathbf{A}}$ in terms of only f.

Proposition 7.12: The vector potential $\widetilde{\mathbf{A}}$ given by (7.19) is flat on $\mathbf{M_L}$ if and only if $\Re r_f = 0$.

Proof: We note that

$$\boldsymbol{\xi}_1' \cdot \boldsymbol{\omega}_2 - \boldsymbol{\xi}_2' \cdot \boldsymbol{\omega}_1 = \Im(\xi_1 \omega_2 - \xi_2 \omega_1).$$

By Lemma 7.11, the rhs is equal to $2\pi \Re r_f$. From this fact and Lemma 7.10, the desired result follows. ■

Remark 7.4: Let $f \in \mathscr{E}$ and $\Re r_f = 0$. Then, by Proposition 7.12, $\widetilde{\mathbf{A}}$ is flat on $\mathbf{M_L}$. We have $\Phi^{\mathbf{A}}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2} = 2\pi \Re r_f = 0$ and

$$\Phi^{\widetilde{\mathbf{A}}}_{\omega_1, \omega_2} = \Phi^{\mathbf{A}}_{\omega_1, \omega_2} + \Phi^{W^{-1}\mathbf{h}}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2} = 0.$$

Hence, in this case, the representations $\pi^{\mathbf{A}}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2}$ and $\pi^{\widetilde{\mathbf{A}}}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2}$ of CCR are unitarily equivalent to the Schrödinger representation and the algebras $\mathscr{R}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \mathbf{A}}$ and $\mathscr{R}_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \widetilde{\mathbf{A}}}$ are commutative. Thus, for $f \in \mathscr{E}$, the case $\Re r_f = 0$ is uninteresting and only the case where $\widetilde{\mathbf{A}}$ is not flat on $\mathbf{M_L}$ may be interesting.

Example 7.1: Consider the case where $f(z)$ is the Weierstrass Zeta function:

$$f(z)=\zeta(z):=\frac{1}{z}+\sum_{(m,n)\in\mathbf{Z}^2\setminus\{0\}}\left\{\frac{1}{z-\Omega_{m,n}}+\frac{1}{\Omega_{m,n}}+\frac{z}{\Omega_{m,n}^2}\right\}. \tag{7.26}$$

It is well known (e.g., Ref. 22, Chap. XX, pp. 20–41) that

$$\zeta(z+\omega_j)=\zeta(z)+2\eta_j, \quad j=1,2, \tag{7.27}$$

where

$$\eta_j=\zeta(\omega_j/2). \tag{7.28}$$

Hence, $\zeta\in\mathscr{E}$. Since $r_\zeta=1$, Lemma 7.11 implies that, in the present case, $\widetilde{\mathbf{A}}$ is not flat on $\mathbf{M_L}$. By Lemma 7.11 we have

$$\eta_1\omega_2-\eta_2\omega_1=i\pi.$$

(This is a well-known formula: e.g., Ref. 22, Chap. XX, 20·411, pp. 446–447.) Hence

$$\boldsymbol{\eta}_1'\cdot\boldsymbol{\omega}_2-\boldsymbol{\eta}_2'\cdot\boldsymbol{\omega}_1=\pi.$$

Thus, in the present example,

$$\widetilde{B}(\mathbf{r})=\frac{2\pi}{\det W}\neq 0, \quad \mathbf{r}\in\mathbf{M_L}.$$

■

C. Derivation of Hamiltonians of the Hofstadter type from continuous systems

It is well known that some transport phenomena in two-dimensional solids can be modeled in part by Hamiltonians of the Hofstadter type, which are usually defined on two-dimensional infinite lattices (e.g., Ref. 13). It may be interesting to investigate if Hamiltonians of the Hofstadter type can be obtained as reductions of self-adjoint Hamiltonians of continuous quantum systems whose Hilbert spaces of state vectors are equal to $L^2(\mathbf{R}^2)$. For that purpose, we introduce a family $\{H^{\mathbf{A}}(t)\}_{t\in\mathbf{R}}$ of bounded self-adjoint operators,

$$\begin{aligned}H^{\mathbf{A}}(t):=&T^{\mathbf{A}}_{\boldsymbol{\omega}_1}(t)+\mu T^{\mathbf{A}}_{\boldsymbol{\omega}_2}(t)+\epsilon[T^{\mathbf{A}}_{\boldsymbol{\omega}_1}(t)^2+T^{\mathbf{A}}_{\boldsymbol{\omega}_2}(t)^2]+\lambda T^{\mathbf{A}}_{\boldsymbol{\omega}_1}(t)T^{\mathbf{A}}_{\boldsymbol{\omega}_2}(t)\\&+\nu T^{\mathbf{A}}_{\boldsymbol{\omega}_1}(t)T^{\mathbf{A}}_{\boldsymbol{\omega}_2}(t)^*+\gamma T^{\mathbf{A}}_{\boldsymbol{\omega}_1}(t)^*T^{\mathbf{A}}_{\boldsymbol{\omega}_2}(t)+\text{h.c.},\end{aligned} \tag{7.29}$$

where $\mu,\epsilon,\lambda,\nu,\gamma$ are complex parameters and h.c. means the Hermitian conjugate. The operator $H^{\mathbf{A}}(t)$ is a continuous version of Hamiltonians of the Hofstadter type. By Theorem 7.3, if $\mathbf{A}\in\mathscr{E}$, then

$$H^{\mathbf{A}}:=H^{\mathbf{A}}(1)=T_1+\mu T_2+\epsilon(T_1^2+T_2^2)+\lambda T_1T_2+\nu T_1T_2^*+\gamma T_1^*T_2+\text{h.c.}, \tag{7.30}$$

is reduced by $L^2_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}(\mathbf{R}^2)\cong l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ and its reduced part yields a Hamiltonian of the Hofstadter type on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$. This shows that, in the case $\mathbf{A}\in\mathscr{E}$, it is possible to obtain Hamiltonians of the Hofstatder type on $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ as reductions of Hamiltonians of the type $H^{\mathbf{A}}$. By Theorem 7.9, there exists a wide class of vector potentials $\mathbf{A}$ that give this kind of reduction for the Hamiltonian $H^{\mathbf{A}}$. We also remark that the reduction of $H^{\mathbf{A}}$ with $\mathbf{A}\in\mathscr{E}$ to the subspace $l^2(\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2))$ makes it possible to identify the spectrum of $H^{\mathbf{A}}$ in part by analyzing the spectra of Hamiltonians of the Hofstadter type on the lattice $\mathbf{L}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ that may have interesting structures such as fractal ones (Ref. 13).

ACKNOWLEDGMENT

This work was supported by Grant-In-Aid No. 08454021 for science research from the Ministry of Education, Japan.

[1] A. Arai, ''Momentum operators with gauge potentials, local quantization of magnetic flux, and representation of canonical commutation relations,'' J. Math. Phys. **33**, 3374–3378 (1992).
[2] A. Arai, ''Properties of the Dirac–Weyl operator with a strongly singular gauge potential,'' J. Math. Phys. **34**, 915–935 (1993).
[3] A. Arai, ''Gauge theory on a nonsimply connected domain and representations of canonical commutation relations,'' J. Math. Phys. **36**, 2569–2580 (1995).
[4] A. Arai, ''Representation of canonical commutation relations in a gauge theory, the Aharonov-Bohm effect, and the Dirac–Weyl operator,'' J. Nonlinear Math. Phys. **2**, 247–262 (1995).
[5] A. Arai, ''Canonical commutation relations, the Weierstrass Zeta function, and infinite dimensional Hilbert space representations of the quantum group $U_q(sl_2)$,'' J. Math. Phys. **37**, 4203–4218 (1996).
[6] Y. Aharonov and D. Bohm, ''Significance of electromagnetic potentials in the quantum theory,'' Phys. Rev. **115**, 485–491 (1959).
[7] E. Brown, ''Bloch electrons in a uniform electromagnetic field,'' Phys. Rev. **133**, A1038–A1044 (1964).
[8] J. Zak, ''Magnetic translation group,'' Phys. Rev. **134**, A1602–A1606 (1964).
[9] W. Florek, ''Magnetic translation groups as group extensions,'' Rep. Math. Phys. **34**, 81–95 (1994).
[10] W. Florek, ''Magnetic translation groups in *n* dimensions,'' Rep. Math. Phys. **38**, 235–250 (1996).
[11] D. R. Hofstadter, ''Energy levels and wave function of Bloch electrons in rational and irrational magnetic fields,'' Phys. Rev. B **14**, 2239–2249 (1976).
[12] P. B. Wiegmann and A. V. Zabrodin, ''Quantum groups and magnetic translations. Bethe ansatz for Asbel–Hofstadter problem,'' Nucl. Phys. B **422**, 495–514 (1994).
[13] Ch. Kreft and R. Seiler, ''Models of the Hofstadter-type,'' J. Math. Phys. **37**, 5207–5243 (1996).
[14] A. Arai and N. Tominaga, ''Analysis of a family of strongly commuting self-adjoint operators with applications to perturbed d'Alembertians and the external field problem in quantum field theory,'' Hokkaido Math. J. **25**, 259–313 (1996).
[15] J. D. Dollard and C. N. Friedman, *Product Integration*, Encyclopedia of Mathematics and Its Applications, Vol. 10, Analysis (Addison-Wesley, Reading, MA, 1979).
[16] M. Göckeler and T. Schücker, *Differential Geometry, Gauge Theories, and Gravity* (Cambridge University Press, Cambridge, 1987).
[17] H. Kurose and H. Nakazato, ''Geometric construction of *-representation of the Weyl algebra with degree 2,'' Publ. RIMS Kyoto Univ. **32**, 555–579 (1996).
[18] J. von Neumann, ''Die Eindeutigkeit der Schrödingerschen Operatoren,'' Math. Ann. **104**, 570–578 (1931).
[19] C. R. Putnam, *Commutation Properties of Hilbert Space Operators* (Springer, Berlin, 1967).
[20] K. R. Davidson, *C*-Algebra by Example* (American Mathematical Society, Providence, 1996).
[21] C. Kassel, *Quantum Groups* (Springer, New York, 1995).
[22] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed. (Cambridge University Press, Cambridge, 1969).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 5 MAY 1998

Generalized Knizhnik–Zamolodchikov equations and isomonodromy quantization of the equations integrable via the Inverse Scattering Transform: Maxwell–Bloch system with pumping

H. M. Babujian[a)]
Institut für Theoretische Physik, Freie Universität Berlin, Arnimallee 14, D-14195 Berlin, Germany

A. V. Kitaev[b)]
Steklov Mathematical Institute, Fontanka 27, St.-Petersburg, 191011, Russia

(Received 21 July 1997; accepted for publication 8 December 1997)

Canonical quantization of the isomonodromy solutions of equations integrable via the Inverse Scattering Transform leads to generalized Knizhnik–Zamolodchikov equations. One can solve these equations by the off-shell Bethe ansatz method provided the Knizhnik–Zamolodchikov equations are related with the highest weight representations of the corresponding Lie algebras: These solutions can be written in terms of multi-variable generalizations of special functions of the hypergeometric type. In this work, we consider a realization of the above scheme for the Maxwell–Bloch system with pumping: quantum states for this system are found in terms of the multi-variable confluent hypergeometric function. © *1998 American Institute of Physics.* [S0022-2488(98)03203-4]

I. INTRODUCTION

Some time ago Reshetikhin[1] introduced generalized Knizhnik–Zamolodchikov equations and noticed that they can be viewed as a quantization of the non-linear system of ordinary differential equations (ODEs) defining isomonodromy deformations of linear ODEs. In particular, the original system of the Knizhnik–Zamolodchikov[2] equations can be viewed as a quantization of the classical Schlesinger equations.[3] Later, Harnad[4] gave a more ''physical'' interpretation of this quantization procedure: By considering the Schlesinger equations as a multi-time Hamiltonian system, he has shown (using the standard quantum mechanical language) that the system of Knizhnik–Zamolodchikov equations can be viewed as a multi-time non-autonomous system of Schrödinger equations.

An attractive feature of this quantization is that the Knizhnik–Zamolodchikov equations on the Lie algebras with the finite weight representations can be solved in terms of the multi-variable hypergeometric functions,[5] so that one can get explicit formulas for the wave-functions, correlators, and so on. To make this quantization procedure closer to applications one has to find physical models which can be described in terms of solutions to the system of Schlesinger equations and/or their generalizations. One regular procedure is well-known: The system of the Schlesinger equations and their generalizations govern the so-called isomonodromy solutions of the equations integrable via the Inverse Scattering Transform (IST).[6] In particular, there are special classes of the isomonodromy solutions to the self-dual Yang–Mills equations, Einstein–Maxwell equations, and the Ernst equation, which can be written in terms of the solutions to the system of the Schlesinger equations.[7,8] In general, to fix a class of the isomonodromy solutions of a partial differential equation integrable via IST means to pick out a finite dimensional non-autonomous Hamiltonian system of its solutions which defines an isomonodromy deformation of some auxiliary linear ODE (see Sec. III for details). Most of the important solutions which are studied in the IST framework; solitons, finite-gap, or similarity, belong to the isomonodromy classes of solu-

[a)]Permanent address: Yerevan Physics Institute, Alichanian Brothers 2, Yerevan 375036, Armenia. Electronic mail: babujian@lx2.yerphi.am
[b)]Electronic mail: kitaev@pdmi.ras.ru

 © 1998 American Institute of Physics

tions. The quantization of the system of Schlesinger equations (or their generalizations) which governs one such isomonodromy subsystems means to associate with the classical system a spin-like non-autonomous finite dimensional quantum system. These systems usually contain some parameters, in particular the number of ''particle (waves),'' N, so that further one can study different scaling limits $N\to\infty$.

Korotkin and Nicolai[9] used the above ideas for the quantization of a special class of the isomonodromy solutions to the Ernst equation—the class which is governed by the classical Schlesinger equations. They found that for the further interpretation of the quantum model obtained as the model of quantum gravity one has to demand that the wave function belongs to the coset space $SL(2,\mathbb{R})/SO(2)$ and satisfies an additional constraint: like the vanishing of a total spin of the system. There are no (non-trivial) solutions of the corresponding Knizhnik–Zamolodchikov equations on the representations of $sl_2(\mathbb{R})$ with the highest weights which satisfy this constraint. There are two possibilities of continuation of these studies for the Ernst equation: (1) Instead of the principal discrete series of representations of $sl_2(\mathbb{R})$ one can try to use the principal continuous series. But the corresponding system of Knizhnik–Zamolodchikov equations becomes a rather complicated system of partial differential equations which is not investigated at the present time; (2) to study an application of this quantization scheme to other classes of isomonodromy solutions to the Ernst equation. In the last case, of course, we change the objects under quantization. The corresponding systems of Knizhnik–Zamolodchikov equations also change; probably they belong to the class of the generalized Knizhnik–Zamolodchikov equations which are introduced in Ref. 1.

In this paper we would like to draw attention to the fact that the quantization procedure briefly described above is applicable to isomonodromy solutions of any IST integrable (KdV-like) differential equation. The corresponding system of the generalized Knizhnik–Zamolodchikov equations can be solved via the off-shell Bethe ansatz method.[10,11] Since there is a rather big variety of the IST integrable equations, each of these equations possesses (infinitely) many classes of the isomonodromy solutions, and the same equation very often serves for a description of different physical models, so that the different classes of the isomonodromy solutions are really involved, we hope that this quantization procedure can appear to be quite reasonable for some physical models and therefore believe that it requires more wide attention and further studies.

In this work we consider the Maxwell–Bloch system with pumping (Sec. II). From the IST point of view this is an intermediate case between the systems integrable via IST with the variable spectral parameter and IST with the spectral parameter independent of the dynamical variables. The spectral parameter in this case depends on one (space) variable and is independent of time. It is worth mentioning that the wave function, the solution of a generalized system of Knizhnik–Zamolodchikov equations, is obtained in Sec. VI in terms of the multi-variable confluent hypergeometric function.

Probably, the Maxwell–Bloch system without pumping (Sec. II) would be physically the most reasonable example for further studies. The associated auxiliary ODE with respect to the spectral parameter, which plays a key role in the definition of the isomonodromy solutions (Sec. III), has an additional pole of the second order compared to the corresponding ODE studied in this work and is therefore the next complicated case to the one considered here.

II. MAXWELL–BLOCH SYSTEM WITH PUMPING

The Maxwell–Bloch system (without pumping)[12,13] describes the propagation of a sequence of electromagnetic pulses along the positive direction of the x-axis in the two level atomic medium. It is known to be integrable via the Inverse Scattering Transform.[14–16] The Maxwell–Bloch system with pumping was introduced by Burtsev, Zakharov, and Mikhailov[17] in the framework of the Inverse Scattering Transform with a variable spectral parameter as an integrable deformation (with the deformation parameter s) of the reduced Maxwell–Bloch system (RMB). More exactly, they considered the following system of linear differential equations:

$$\partial_\xi\psi=(z\sigma_3+U_0)\psi,\quad \left(\partial_\eta+\frac{s}{z}\partial_z\right)\psi=\frac{\hat\rho}{4z}\psi, \tag{2.1}$$

where $z \in \mathbb{C}$ is the auxiliary (spectral) parameter and the 2×2-matrices σ_3, U_0, and $i\hat{\rho} \in su_2$ have the following form:

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad U_0 = \frac{1}{2}\begin{pmatrix} 0 & E \\ -E^* & 0 \end{pmatrix}, \quad \hat{\rho} = \begin{pmatrix} N & -\rho \\ -\rho^* & -N \end{pmatrix}, \tag{2.2}$$

with the functions: $E = E(\xi, \eta), N = N(\xi, \eta)$, and $\rho = \rho(\xi, \eta)$ independent of z and (*) denoting the complex conjugation. The compatibility condition for the system (2.1) reads as

$$\partial_\eta E = \rho, \quad \partial_\xi \rho = NE, \quad \partial_\xi N + \frac{1}{2}(E\rho^* + E^*\rho) = 4s. \tag{2.3}$$

This non-linear system of partial differential equations is called the Maxwell–Bloch system with pumping (PMB). The physical meaning of the notations used is as follows: dimensionless variables η and ξ are related to the space, x, and time, t, variables:

$$\eta = \Omega \frac{x}{c}, \quad \xi = \Omega\left(t - \frac{x}{c}\right); \tag{2.4}$$

E and ρ are slowly varying complex amplitudes of the electric field and polarization, respectively; N is the population inversion, therefore $-1 \leqslant N \leqslant 1$; $\Omega > 0$ is a parameter characterizing the medium; and $s > 0$ represents the ''pumping,'' i.e., an additional excitation of the active medium as it was interpreted by the authors.[17] Note that the standard RMB equations are given by (2.3) and (2.4) with $s = 0$. From the mathematical viewpoint all the techniques related with IST, including the quantization which we present in this work, can be developed for the complex function N, and the parameters ω and s. However, here we follow the restrictions mentioned above. Further details about the system (2.3), (2.4) can be found in the works.[18–22] In fact the Maxwell–Bloch system is a semiclassical system: The medium is considered quantum-mechanically while the electromagnetic field is described via the classical Maxwell equations. The quantization scheme which is presented in this work can be considered as an attempt of quantization of the electromagnetic field interacting with the active medium.

III. ISOMONODROMY SOLUTIONS

The general notion of the isomonodromy solutions of the equations integrable via IST was formulated in the work.[6] It is applicable also to the equations integrable via IST with a variable spectral parameter. The ψ-function which corresponds to such solutions, possibly after some gauge-like transformations, solves together with the system (2.1) one more linear ODE, the ODE with respect to the spectral parameter,

$$\partial_z \psi = A(z)\psi, \tag{3.1}$$

where $A(z) = A(z, \xi, \eta) \in su_2$ is a rational function of the spectral parameter z. The compatibility condition of Eq. (3.1) with the system (2.1) does not uniquely define the matrix $A(z, \xi, \eta)$. This freedom is such that all the most important solutions to the equations integrable via IST (multisoliton, finite-gap, similarity) can be specified as the solutions of the isomonodromy class: the ψ-functions corresponding to them solve Eq. (3.1) with different matrices $A(z, \xi, \eta)$ for the solutions of the different types. Some solutions, which do not belong to the isomonodromy class, e.g., rapidly decaying solutions, also show an isomonodromy behavior in many interesting asymptotic regimes.

In our case the simple possible choice of $A(z, \xi, \eta)$ which is consistent with the system (2.1) reads as

$$A(z, \xi, \eta) = (\xi - \xi_0)\sigma_3 + \sum_{j=1}^{N} \frac{A_j}{z - z_j}, \quad z_j = \sqrt{2s\eta - k_j^2}, \tag{3.2}$$

where k_j^2, $\xi_0 \in \mathbb{R}$ are the parameters and the functions $A_j = A_j(\xi, \eta)$ are independent of z. In the following, to avoid technical details, we suppose that our system has a finite size, i.e.,

$$0 \leqslant \eta \leqslant L, \quad k_j^2 > 2sL \Rightarrow z_j^* = -z_j. \tag{3.3}$$

Proposition 1: If for $j=1,\ldots,N$ the matrix-functions $A_j \in su_2$ are such that the set $\{A_j\}_{j=1,\ldots,N}$ is a solution of the system

$$\partial_\xi A_j = [z_j\sigma_3 + U_0, A_j], \quad \partial_\eta A_j = \left[\frac{\hat{\rho}}{4z_j}, A_j\right], \tag{3.4}$$

$$U_0 = \frac{1}{\xi - \xi_0}\left(\sum_{l=1}^{N} A_l - i\theta_\infty\sigma_3\right), \quad i\theta_\infty\sigma_3 \equiv \operatorname{diag}\sum_{l=1}^{N} A_l, \quad \hat{\rho} = 4s\left((\xi-\xi_0)\sigma_3 - \sum_{l=1}^{N}\frac{A_l}{z_l}\right), \tag{3.5}$$

where the brackets, $[\cdot,\cdot]$, denote the commutator of two matrices; then $\theta_\infty \in \mathbb{R}, \partial_\xi\theta_\infty = \partial_\eta\theta_\infty = 0$, the matrix U_0 is off-diagonal, and the functions E, ρ, and N, which are defined by Eqs. (2.2) and (3.5), are the solutions of PMB (2.3).

The system (3.4), (3.5) defines the deformation (with the deformation parameters ξ and η) of the coefficients, A_j, of Eq. (3.1) with the matrix $A(z)$ given by (3.2), which preserves the monodromy group or, more generally, the monodromy data[23] of this equation. This comment ''explains'' the name isomonodromy solutions. Note that for $N=2$ the system (3.4), (3.5) reduces to the classical fifth Painlevé equation.[20]

To the N-soliton solution of PMB (2.3) corresponds a particular solution of the system (3.4), (3.5): The eigenvalues of all the matrices A_j are equal to $\pm 1/2$. To each soliton corresponds one (and only one) eigenvalue z_l. So, intuitively, one can think of the isomonodromy solutions defined by the system (3.4), (3.5) as an N-wave-like object, although the behavior of the general solution is much more complicated than the N-soliton train.

IV. HAMILTONIAN STRUCTURE

It is known that the equations which define isomonodromy deformations of the linear ODEs (3.1) where the matrix $A(z)$ is a rational function of z admit a Hamiltonian structure. The Poisson brackets between the matrix elements can be written in the following compact form:[24,25]

$$\{A(z) \overset{\otimes}{,} A(w)\} = [r(z-w), A(z)\otimes I + I\otimes A(w)], \tag{4.1}$$

where $r(z-w) = P/(z-w)$ with P the permutation operator on $\mathbb{C}^2 \otimes \mathbb{C}^2$. The corresponding Hamiltonians are generated by the so-called τ-function.[23] In our case Eq. (4.1) for the matrix elements, $(A_k)_{ab}$, reads as:[26]

$$\{(A_k)_{ab}, (A_j)_{cd}\} = \delta_{kj}((A_k)_{ad}\delta_{bc} - \delta_{ad}(A_k)_{cb}). \tag{4.2}$$

Proposition 2: The system (3.4), (3.5) can be written in the Hamiltonian form:

$$\partial_\xi A_j = \{A_j, H_\xi\}, \quad \partial_\eta A_j = \{A_j, H_\eta\}, \tag{4.3}$$

where the dynamics with respect to the variables ξ and η is defined by the following non-autonomous Hamiltonians:

$$H_\xi = \sum_{k=1}^{N} z_k \operatorname{tr}(A_k\sigma_3) + \frac{S^-S^+ + S^+S^-}{\xi - \xi_0}, \quad S^+ = \left(\sum_{k=1}^{N} A_k\right)_{12}, \quad S^- = \left(\sum_{k=1}^{N} A_k\right)_{21}, \tag{4.4}$$

$$H_\eta = \sum_{k=1}^{N} H_k \frac{\partial z_k}{\partial \eta} = \sum_{k=1}^{N} \frac{H_k}{z_k}, \quad H_k = (\xi-\xi_0)\operatorname{tr}(A_k\sigma_3) + \sum_{i\neq k}^{N} \frac{\operatorname{tr}(A_kA_i)}{z_k - z_i}, \quad k = 1,\ldots,N. \tag{4.5}$$

The compatibility of the system (4.3) follows from Proposition 3.

Proposition 3: $\{H_\xi, H_\eta\}=0$, $\partial H_\xi/\partial\eta=\partial H_\eta/\partial\xi$.

V. QUANTIZATION

We consider the N-particle sector of the Fock space,

$$\mathscr{H}_N=\otimes_{k=1}^N V_k, \tag{5.1}$$

where V_k are left su_2-modules corresponding to the spin m_k representations of su_2, i.e., V_k are isomorphic to $\mathbb{C}^{2m_k+1}$, $m_k\in\mathbb{Z}/2$. Denote by $s_k^a, a=1, 2, 3$, the linear operators, which generate in V_k these representations ($s_k^a\in su_2$):

$$[s_k^a, s_j^b]=i\delta_{kj}\epsilon^{abc}s_k^c, \quad a, b, c\in\{1, 2, 3\}, \quad s_k^3\omega_k=m_k\omega_k, \tag{5.2}$$

where ω_k is the highest weight vector. In particular, for the spin 1/2 representation $s_k^a=\frac{1}{2}\sigma_k^a$, where σ_k^1, σ_k^2, σ_k^3 are multiplications on the Pauli matrices. The natural extension of s_k^a to $\mathscr{H}_N$ is denoted by $\hat{s}_k^a$,

$$\hat{s}_k^a(g)=I\otimes\cdots\underbrace{s_k^a(g)}_{k-th}\cdots\otimes I\in\mathrm{End}(\mathscr{H}), \quad g\in su_2.$$

It is convenient to also introduce the following notations:

$$\hat{s}_k^\pm=\hat{s}_k^1\pm i\hat{s}_k^2, \quad \hat{S}^3=\sum_{k=1}^N \hat{s}_k^3, \quad \hat{S}^\pm=\sum_{k=1}^N \hat{s}_k^\pm.$$

In the Schrödinger picture, quantization of the classical observables A_k reads as

$$A_k\Rightarrow\hat{A}_k=i\hbar\begin{pmatrix}\hat{s}_k^3 & \hat{s}_k^+\\ \hat{s}_k^- & -\hat{s}_k^3\end{pmatrix}, \quad (\xi-\xi_0)\sigma_3\Rightarrow i\hbar(\xi-\xi_0)\sigma_3\otimes I.$$

The generator of the quantum observables [cf. (3.2)],

$$\hat{A}(z,\xi,\eta)=i\hbar(\xi-\xi_0)\begin{pmatrix}I & 0\\ 0 & -I\end{pmatrix}+\sum_{j=1}^N\frac{\hat{A}_j}{z-z_j}, \quad z_j=\sqrt{2s\eta-k_j^2}, \tag{5.3}$$

satisfies the following commutator relation [cf. (4.1)]:

$$[\hat{A}(z)\overset{\otimes}{,}\hat{A}(w)]=i\hbar[r(z-w), \quad \hat{A}(z)\otimes\hat{I}+\hat{I}\otimes\hat{A}(w)], \tag{5.4}$$

where $\hat{I}$ is the matrix with the identity operators on the main diagonal and the classical r-matrix above is the same as in the Poisson structure (4.1). The quantum Hamiltonians in the above notations can be written as follows:

$$\hat{H}_\xi=2(i\hbar)^2\sum_{j=1}^N z_j\hat{s}_j^3+\frac{(i\hbar)^2}{\xi-\xi_0}(\hat{S}^-\hat{S}^++\hat{S}^3),$$

$$\hat{H}_k=2(i\hbar)^2(\xi-\xi_0)\hat{s}_k^3+2(i\hbar)^2\sum_{j\neq k}^N\frac{\hat{s}_k^a\hat{s}_j^a}{z_k-z_j},$$

where in the last equation the summation on the superscript a is supposed.

The time evolution of the state-vector Ψ is governed by the Schrödinger equation,

$$i\hbar\kappa\frac{\partial\Psi}{\partial\xi}=\hat{H}_\xi\Psi, \tag{5.5}$$

where $\kappa \in \mathbb{R}$ plays a role of the dimension constant; note that our ''time-variable'' ξ is dimensionless! To be consistent with the classical picture we have to impose on the vector Ψ the system of N more ''Schrödinger'' equations,

$$i\hbar\kappa\frac{\partial\Psi}{\partial z_k}=\hat{H}_k\Psi, \quad k=1,\ldots,N, \tag{5.6}$$

which govern the dependence of the state-vector on the parameters z_k and so define the space-like dependence of the state vector. The necessity of the appearance of additional equations, system (5.6), is clear since we are quantizing the (soliton-) wave-like objects (see the last paragraph in Sec. III), whilst the Schrödinger equation (5.5) defines only the time dependence of the state-vector.

The system (5.5), (5.6) is compatible. It follows from Proposition 4.

Proposition 4: $[\hat{H}_\xi, \hat{H}_\eta]=0$, $\partial_\eta\hat{H}_\xi=\partial_\xi\hat{H}_\eta$.

The system (5.5), (5.6) is a generalized system of the Knizhnik–Zamolodchikov equations. In particular, the subsystem (5.6) for $\xi-\xi_0=0$ coincides with the classical Knizhnik–Zamolodchikov equations.[2]

VI. GENERALIZED KNIZHNIK–ZAMOLODCHIKOV EQUATIONS

In the previous section we have obtained the following system of generalized Knizhnik–Zamolodchikov equations:

$$\kappa\frac{\partial\Psi}{\partial\xi}=i\hbar\left(2\sum_{j=1}^{N}z_j\hat{s}_j^3+\frac{\hat{S}^-\hat{S}^++\hat{S}^3}{\xi-\xi_0}\right)\Psi, \tag{6.1}$$

$$\kappa\frac{\partial\Psi}{\partial z_k}=i\hbar\left(2(\xi-\xi_0)\hat{s}_k^3+2\sum_{j\neq k}^{N}\frac{\hat{s}_k^a\hat{s}_j^a}{z_k-z_j}\right)\Psi. \tag{6.2}$$

We have applied the off-shell Bethe ansatz method[10,11] to construct the general solution of this system. This method has a pure algebraic nature and that is why it works not only for the real values of the parameters: ξ, ξ_0, $\hbar$, iz_k, as is assumed in the setting related with the PMB, but also in the case when one allows arbitrary complex values ($z_k\neq z_j$ for $k\neq j$) for these parameters. Thus, the following formulae are valid for complex values of the parameters too.

Since the off-shell Bethe ansatz method has already been presented in the literature[10,11] and its application to system (6.1), (6.2) requires no essential modification, here we only write down the final answer.

The vacuum solution reads as

$$\Psi_0=\chi_0|0\rangle, \quad |0\rangle=\otimes_{k=1}^{N}\omega_k,$$

$$\chi_0=(\xi-\xi_0)^{i(\hbar/\kappa)\Sigma_{k=1}^{N}m_k}\prod_{1\leqslant k<j\leqslant N}(z_k-z_j)^{2i(\hbar/\kappa)m_km_j}\exp\left(2i\frac{\hbar}{\kappa}(\xi-\xi_0)\sum_{k=1}^{N}m_kz_k\right),$$

where m_k and ω_k are the highest weights and the highest weight vectors correspondingly (5.2). For each $m\in\mathbb{Z}$ satisfying the inequality, $1\leqslant m\leqslant 2\Sigma_{k=1}^{N}m_k$, the function

$$\Psi_m(\xi,\vec{z})=\oint_{\Delta_m}\chi_m(\vec{\mu}_m|\xi,\vec{z})\hat{\varphi}_m(\vec{\mu}_m|\xi,\vec{z})d\mu_1,\ldots,d\mu_m,$$

where $\vec{z}\equiv(z_1,\ldots,z_N)$, $\vec{\mu}_m\equiv(\mu_1,\ldots,\mu_m)$,

$$\chi_m(\vec{\mu}_m|\xi,\vec{z})=(\xi-\xi_0)^{(i\hbar/\kappa)\Sigma_{k=1}^N m_k-m)}\exp\left(\frac{2i\hbar}{\kappa}(\xi-\xi_0)\sum_{k=1}^N z_k m_k\right)\prod_{1\leqslant k<j\leqslant N}(z_k-z_j)^{(2i\hbar/\kappa)\, m_k m_j}$$

$$\times\prod_{\substack{1\leqslant\alpha\leqslant m\\ 1\leqslant k\leqslant N}}(z_k-\mu_\alpha)^{-(i\hbar/2\kappa)\, m_k}\prod_{1\leqslant\alpha<\beta\leqslant m}(\mu_\alpha-\mu_\beta)^{(i\hbar/2\kappa)}$$

$$\times\exp\left(-\frac{i\hbar}{\kappa}(\xi-\xi_0)\sum_{\alpha=1}^m\mu_\alpha\right),$$

$$\hat{\varphi}_m(\vec{\mu}_m|\xi,\vec{z})=\prod_{k=1}^m\hat{S}^-(\mu_k|\xi,\vec{z})|0>,\quad \hat{S}^-(\mu|\xi,\vec{z})=\sum_{k=1}^N\frac{\hat{s}_k^-}{\mu-z_k},$$

and the cycles Δ_m are defined by the following equation:

$$\sum_{k=1}^m\oint_{\Delta_m}\frac{d}{d\mu_k}\frac{\chi_m(\vec{\mu}_m|\xi,\vec{z})\ \hat{s}_j^-\,\hat{\varphi}_m^k(\vec{\mu}_m|\xi,\vec{z})}{z_j-\mu_k}\,d\mu_1,\ldots,d\mu_m=0,$$

$$\hat{\varphi}_m^k(\vec{\mu}_m|\xi,\vec{z})=\prod_{j\neq k}^N\hat{S}^-(\mu_j|\xi,\vec{z}),$$

solves the system (6.1), (6.2). The solution, Ψ_m, also should satisfy the quantum version of the second equation in (3.4),

$$(\hbar\hat{S}^3-\theta_\infty)\Psi_m=0\Rightarrow\hbar\left(m-\sum_{k=1}^N m_k\right)=\theta_\infty\,. \tag{6.3}$$

ACKNOWLEDGMENTS

This work was done during the authors' stay at Sonderforshungsbereich 288 in May 1995. Since then different versions were presented at seminars in FU Berlin, POMI (St. Petersburg), and INFN (Padova). It is our pleasure to thank A. I. Bobenko, L. D. Faddeev, I. R. Gabitov, M. Karowski, D. A. Korotkin, I. M. Krichever, P. A. Marchetti, M. Matone, H. Nicolai, R. Seiler, R. Schrader, and F. Toppan for discussions and useful comments. One of us (A. V. K.) is grateful to the Max-Planck-Insitut für Gravitationsphysik (Potsdam) for hospitality during the preparation of this work for publication. H. M. Babujian was partially supported by Grant No. 211-5291 YPI of the German Bundesministerium für Forshung und Technologie and Sonderforschungsbereich 288. A. V. Kitaev was supported by Sonderforschungsbereich 288 and the Alexander von Humboldt Foundation.

[1] N. Reshetikhin, Lett. Math. Phys. **26**, 167 (1992).
[2] V. G. Knizhnik and A. B. Zamolodchikov, Nucl. Phys. B **247**, 83 (1984).
[3] L. Schlesinger, J. Reine Angew. Math. **141**, 96 (1912).
[4] J. Harnad, *Symmetries and Integrability of Difference Equations*, edited by Levi and Decio *et al.*, CRM Proc. Lect. Notes **9**, 155 (1996), hep-th/9406078.
[5] V. V. Schechtman and A. N. Varchenko, Lett. Math. Phys. **20**, 279 (1990); Invent. Math. **106**, 139 (1991).
[6] A. R. Its, Math. USSR Izv. **26**, 497 (1986).
[7] A. V. Kitaev, Zap. Nauchn. Semin. LOMI **181**, 65 (1990); English translation in J. Sov. Math. Plenum **62**(2), 2646 (1992).
[8] D. Korotkin and H. Nicolai, Phys. Rev. Lett. **74**, 1272 (1995).
[9] D. Korotkin and H. Nicolai, Nucl. Phys. B **475**, 397 (1996).
[10] H. M. Babujian, ''Correlation function in WZNW model as a Bethe wave function for the Gaudin magnetics,'' in *Proceedings of the XXIV International Symposium on Ahrenshoop*, Zeuthen, 1990; J. Phys. A **26**, 6981 (1993); **27**, 7753 (1982).
[11] H. M. Babujian and R. Flume, Mod. Phys. Lett. A **9**, 2029 (1994).
[12] S. L. McCall and E. L. Hahn, Phys. Rev. Lett. **18**, 908 (1967).
[13] S. L. McCall and E. L. Hahn, Phys. Rev. **183**, 457 (1969).
[14] G. L. Lamb, Jr., Rev. Mod. Phys. **43**, 99 (1971).
[15] M. J. Ablowitz, D. J. Kaup, and A. C. Newell, J. Math. Phys. **15**, 1852 (1974).

[16] I. R. Gabitov, V. E. Zakharov, and A. V. Mikhailov, Theor. Math. Phys. **63**, 328 (1985).
[17] S. P. Burtsev, V. E. Zakharov, and A. V. Mikhailov, Theor. Math. Phys. **70**, 227 (1987).
[18] D. A. Korotkin and V. B. Matveev, Leningrad. Math. J. **1**, 379 (1990).
[19] A. V. Rybin, J. Phys. A **24**, 5235 (1991).
[20] A. V. Kitaev, A. V. Rybin, and J. Timonen, J. Phys. A **26**, 3583 (1993).
[21] S. P. Burtsev, Inverse Probl. **10**, 837 (1994).
[22] S. P. Burtsev and I. R. Gabitov, Phys. Rev. A **49**, 2065 (1994).
[23] M. Jimbo, T. Miwa, and K. Ueno, Physica D **2**, 306 (1981).
[24] M. A. Semenov-Tian-Shansky, Funct. Anal. Appl. **14**, 146 (1980).
[25] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer-Verlag, Berlin, 1986).
[26] M. Jimbo, T. Miwa, Y. Môri, and M. Sato, Physica D **1**, 80 (1980).

Geometry of the Batalin–Fradkin–Vilkovisky theorem

K. Bering[a)]
Institute of Theoretical Physics, Uppsala University,
P.O. Box 803, S-751 08 Uppsala, Sweden

(Received 15 October 1997; accepted for publication 24 November 1997)

We describe gauge fixing at the level of virtual paths in the path integral as a nonsymplectic BRST-type of flow on the path phase space. As a consequence, a gauge-fixed, nonlocal symplectic structure arises. Restoring of locality is discussed. A pertinent anti-Lie-bracket and an infinite dimensional group of gauge fermions are introduced. Generalizations to Sp(2)-symmetric BLT theories are made. © *1998 American Institute of Physics.* [S0022-2488(98)00804-4]

I. INTRODUCTION

Similar to the Lagrangian BRST symmetry[1] for Yang–Mills theory, Batalin, Fradkin, and Vilkovisky[2] developed a BRST formulation of an arbitrary Hamiltonian gauge theory with reducible first class constraints and possible open gauge algebra, now known as BFV-BRST quantization.[3] The crucial ingredients turn out to be two Grassmann odd objects, the nilpotent BRST charge Ω, and the gauge fermion ψ. The BFV theorem, which is the subject of this article, simply states that the partition function $\mathcal{Z}$ is independent of the choice of ψ.

The present work was originally motivated by a concern about the stability of time locality. It is a well-known fact that time has a special status in Hamiltonian theories. One assumes that all virtual paths in the path integral are parametrized by the one and same global time parameter. As a consequence, it makes sense to speak of equal-time Poisson brackets. In Hamiltonian theories, there is no interaction between two different time slices. In other words, the symplectic two-form $\omega(t)=\omega(z(t),t)$ is a *function* of the coordinates taken in the very same time. This is often referred to as saying that the phase space Poisson structure is ultra-local in time. So the symplectic structure does not depend on the past nor the future. Nevertheless, in the standard proof[2] for the BFV theorem one performs an infinitesimal flow generated by a bosonic BRST-type vector field

$$X(t)=\mu X_{\Omega(t)}, \tag{I.1}$$

where $\mu\sim\int_0^T dt\ \delta_{\mathrm{ext}\psi}(t)$ is a *functional*. The flow therefore carries information about the past and the future of the path. In a consistent geometric formulation such flows will violate the ultra-local ansatz for the two-form $\omega(t)=\omega(z(t),t)$. The BRST variation (I.1) is transporting well-defined ultra-local theories into nonlocal theories. Naively, one would expect that the pertinent BRST transformation would map ultra-local theories in ultra-local theories. However, this is not the case. Our analysis below shows that the nonlocality arising in the canonical measure factor, the Pfaffian, remarkably can be recast into an ultra-local form, which turns out to be the gauge-fixing term in the action. This solves our original posed question, and it shows that the formalism is consistent.

Besides the time-local problem mentioned above, we introduce and give a geometric description of the gauge fixing flow directly at the level of virtual paths in the path phase space.

The article is organized as follows: A pertinent anti-Lie-bracket, that turns the set of gauge fermions into an infinite-dimensional algebra is introduced in Sec. III. This algebra is interesting in its own right from a pure mathematical point of view. There we also discuss the realizations of the algebra of gauge fermions and its corresponding group. This paves the way for introduction of a gauge-fixing flow discussed in Sec. IV. With this in hand we can prove the BFV theorem in a path space formalism; see Sec. V. We have for completeness included in Sec. V C the infinitesimal gauge-fixing flow. Finally, in Sec. VI we generalize the construction to Sp(2)-symmetric

[a)]Current address: Center for Theoretical Physics, MIT, Cambridge MA 02139, USA.

 © 1998 American Institute of Physics

Hamiltonian theories.[4] Some of the results in this article have been reported on the NATO Advanced Research Workshop, 1997, Zakopane, ''New Developments in Quantum Field Theory.''

II. LIFT TO PATH SPACE

The basic philosophy is to consider the collection $\mathscr{P}\Gamma$ of all paths as an infinite-dimensional manifold, i.e., a geometric object that does not depend on the choice of coordinates. For instance, given a volume form, the partition function is an integral over this manifold weighted with the Boltzmann factor. Let us emphasize that both the Boltzmann factor and the volume form are scalar objects, which does not depend on the choice of coordinates. In this picture the BRST transformation (I.1) is a non-Hamiltonian (and nonlocal) vector field on the manifold $\mathscr{P}\Gamma$, i.e., it is not a symplectic or equivalently a canonical transformation. So the symplectic structure is affected by the BRST vector field (I.1) and it in general becomes nonlocal.

We see that the BRST-type transformation (I.1) forces us to consider a wider class of theories than the ultra-local ones. To include nonlocal theories, we have to lift the construction from phase space Γ to a *path* phase space $\mathscr{P}\Gamma$, thereby giving room for nontrivial Poisson bracket between different times. In general, nonlocal theories are senseless, but we shall see that we can give meaning to a restricted class of nonlocal theories.

We shall here discard global obstructions and assume that objects like the gauge fermion ψ and the symplectic potential ϑ exist globally. So, in particular, our considerations do not include Gribov problems. Also we will, for simplicity, consider the case where Γ does not have boundaries. Moreover, we assume that the path integral exists for the studied gauges.

Let us introduce a gauge-fixed symplectic two-form directly in the path space $\mathscr{P}\Gamma = \Pi_{t\epsilon[0,T]}\,\Gamma$:

$$\omega_\psi = \omega_\psi(\gamma) = \frac{1}{2}\int_\gamma\int_\gamma dt\ dt' dz^A(t)\omega_{\psi AB}(t,t')\wedge dz^B(t') = \frac{1}{2}\int_\gamma\int_\gamma dt\ dt'\ \omega_{\psi AB}(t,t')dz^B(t')\wedge dz^A(t)(-1)^{\epsilon_A+1}. \tag{II.1}$$

The basic picture is that acting with a BRST-type transformation (I.1) change the symplectic two-form depending on the gauge-fixing ψ. In other words, ω_ψ is a modified symplectic two-form, whose change is an accumulated effect of acting with a BRST-type of transformations (I.1). We shall derive the explicit formula later. Here we are only interested in the principle idea. The inverse $\omega_\psi^{AB}(t,t')$ gives rise to a gauge-fixed path space Poisson bracket

$$\{F,G\}_\psi = \{F,G\}_\psi(\gamma) = \int_\gamma\int_\gamma dt\ dt'\ F\ \frac{\overleftarrow{\delta}^r}{\delta z^A(t)}\ \omega_\psi^{AB}(t,t')\ \frac{\overrightarrow{\delta}^l}{\delta z^B(t')}\ G, \tag{II.2}$$

where $F,G:\mathscr{P}\Gamma\to\mathbb{C}$. For generic ψ the elements $\omega_{\psi AB}(t,t')$ are functionals of the path γ. More precisely, we allow for the following dependence: (A) It could be a *functional* of the full path γ, (B) it could be a function of the two path values $\gamma(t)$ and $\gamma(t')$, and finally (C) it could depend explicitly on the times t and t'. In symbols,

$$\omega_{\psi AB}(t,t') = \omega_{\psi AB}(\gamma,\gamma(t),\gamma(t'),t,t'). \tag{II.3}$$

As an obvious boundary condition, when the gauge fermion is turned off, $\psi=0$, the gauge-fixed symplectic structure should coincide with the original ultra-local symplectic structure. That is,

$$\omega = \omega(\gamma) = \frac{1}{2}\int_\gamma\int_\gamma dt\ dt'\ dz^A(t)\omega_{AB}(z(t),t)\,\delta(t-t')\wedge dz^B(t'). \tag{II.4}$$

In the same manner, instead of looking at classical observables $f:\Gamma\to\mathbb{C}$, let us consider classical observables $F:\mathscr{P}\Gamma\to\mathbb{C}$ that are functions on the *path* space $\mathscr{P}\Gamma$. The usual ultra-local classical observables F are of the form

$$F(\gamma)=\int_\gamma dt\ f(\gamma(t),t) \quad \text{or} \quad F(\gamma)=f(\gamma(t_1),t_1) \tag{II.5}$$

for some function $f:\Gamma\times[0,T]\to\mathbb{C}$ and $t_1\in[0,T]$. Examples are the BRST-improved Hamiltonian $H_{\mathrm{BRST}}(z(t),t)$ and the BRST charge $\Omega(z(t),t)$. The path space versions read

$$H_{\mathrm{BRST}}=H_{\mathrm{BRST}}(\gamma)=\int_\gamma dt H_{\mathrm{BRST}}(\gamma(t),t),$$

$$\Omega=\Omega(\gamma)=\int_\gamma dt\Omega(\gamma(t),t),$$

$$\{\Omega,\Omega\}=0, \quad \{\Omega,H_{\mathrm{BRST}}\}=0. \tag{II.6}$$

The classical observables of the form (II.5) are formally the fundamental building blocks for all classical observables, i.e., a classical observable F in path space is a formal powerseries of the type (II.5). As a particular example we allow that the gauge fermion $\psi:\mathscr{P}\Gamma\to\mathbb{C}$ is a power series of these building blocks.

III. THE SPACE OF GAUGE CHOICES

Let us analyze the set of gauge fermions in more detail. First of all, it turns out that we can define a Grassmann odd bracket structure among the gauge fermions. We have chosen here to start with this construction to emphasize its fundamental nature. We shall later see how this anti-algebra can be realized in the algebra of vector fields as BRST-like vector fields. When exponentiating to the corresponding groups we realize the corresponding BRST flow or—what turn out to be the same—the gauge-fixing flow; see Sec. IV.

A. An anti-Lie-bracket

An anti-Lie-bracket in the space $C^\infty(\mathscr{P}\Gamma)$ of ''classical'' observables is defined via the Poisson bracket and the BRST charge:

$$(F,G)=\frac{i}{\hbar}(-1)^{\epsilon_F}\{\Omega,FG\}=\frac{i}{\hbar}\{FG,\Omega\}(-1)^{\epsilon_G+1}=-(-1)^{(\epsilon_F+1)(\epsilon_G+1)}(G,F). \tag{III.1}$$

The anti-Lie-bracket is of odd Grassmann parity and it has ghost number $+1$. It satisfies the correct symmetry property, the Jacobi identity, but not the Poisson property.

B. Group of gauge fermions

The space of gauge fermions $\psi:\mathscr{P}\Gamma\to\mathbb{C}$ with ghost number -1 equipped with this anti-Lie-bracket is an infinite-dimensional Lie-algebra, which we denote $\mathscr{G}$. The corresponding group G is identified with precisely the same space of ghost number -1 functions $\Psi:\mathscr{P}\Gamma\to\mathbb{C}$, so $G=\mathscr{G}$. The group G is endowed with an associative product $\diamond:G\times G\to G$,

$$\Psi_1\diamond\Psi_2=\Psi_1+\Psi_2+\frac{i}{\hbar}\Psi_1\{\Omega,\Psi_2\}, \tag{III.2}$$

and an inversion map $\tau:G\to G$,

$$\tau(\Psi)=-\frac{\Psi}{1+(i/\hbar)\{\Omega,\Psi\}}=-\frac{\Psi}{1+B}=-\Psi e^{-b}. \tag{III.3}$$

We have for convenience introduced

$$b\equiv\frac{i}{\hbar}\{\Omega,\psi\}=\ln(1+B), \quad B\equiv\frac{i}{\hbar}\{\Omega,\Psi\}=e^b-1. \tag{III.4}$$

The algebra elements are in general denoted by a lowercase ψ and the group element are denoted by uppercase Ψ. They are connected via the *bijective* exponential map $\mathrm{Exp}:\mathscr{G}\to G$:

$$\Psi=\mathrm{Exp}(\psi)=\psi e(b),\quad \psi=\mathrm{Ln}(\Psi)=\Psi l(B), \tag{III.5}$$

where e and l are

$$e(b)=\int_0^1 d\alpha\ e^{\alpha b}=\frac{e^b-1}{b}=\frac{B}{b}=\sum_{n=1}^{\infty}\frac{b^{n-1}}{n!}, \tag{III.6}$$

$$l(B)=\frac{\ln(1+B)}{B}=\frac{b}{B}=\sum_{n=1}^{\infty}\frac{(-B)^{n-1}}{n}. \tag{III.7}$$

The neutral element of the Lie group is $\Psi=0$:

$$\Psi\diamond 0=\Psi=0\diamond\Psi,\quad \mathrm{Exp}(0)=0. \tag{III.8}$$

The exponential map obeys the Baker–Campbell–Hausdorff formula:

$$\mathrm{BCH}(u\psi_1,u\psi_2)=\mathrm{Ln}(\mathrm{Exp}(u\psi_1)\diamond\mathrm{Exp}(u\psi_2))=u\psi_1+u\psi_1+\tfrac{1}{2}(u\psi_1,u\psi_2)+\mathscr{O}(u^3),$$

$$\mathrm{Exp}(-\psi)=\tau(\mathrm{Exp}(\psi)). \tag{III.9}$$

This is easiest to prove in an algebra/group representation—a so-called realization—that we describe in the next section.

C. Group of diffeomorphisms

Consider the infinite dimensional Lie group $\mathrm{Diff}(\mathscr{P}\Gamma)$ of diffeomorphisms $\sigma:\mathscr{P}\Gamma\to\mathscr{P}\Gamma$. The corresponding infinite dimensional Lie algebra $\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))$ is identified with the set of bosonic vector fields $X:C^\infty(\mathscr{P}\Gamma)\to C^\infty(\mathscr{P}\Gamma)$ with the usual Lie bracket

$$[X,Y][F]=X[Y[F]]-Y[X[F]]. \tag{III.10}$$

Here we have adapted the common definition[5] of a vector field X that it is a linear derivation on the space of functions $C^\infty(\mathscr{P}\Gamma)$, i.e., that it satisfies a graded Leibnitz' rule. Although this definition has the correct transformation properties under change of coordinate patches and it works in the general case, it has a complication that a more elementary definition of a vector field often doesn't have: It relates naturally to the *reversed* group $(\mathrm{Diff}(\mathscr{P}\Gamma),\overset{\mathrm{op}}{\circ})$ of opposite ordering, where $\sigma_1\overset{\mathrm{op}}{\circ}\sigma_2=\sigma_2\circ\sigma_1$ (or *vice-versa*, the Lie algebra of vector fields acting from the *right* corresponds to the group of usual ordering). To see this, let us write down the pertinent exponential map Exp from the algebra $(\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma)),[\cdot,\cdot])$ to the group $(\mathrm{Diff}(\mathscr{P}\Gamma),\overset{\mathrm{op}}{\circ})$. It is defined as follows. For each vector field $X\in\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))$, there exists a unique one-parameter family solution $u\mapsto\sigma_X(u)$, to the first-order differential equation

$$\begin{cases}\dfrac{d}{du}\,\sigma_X(u)=X_{\sigma_X(u)}\\ \varphi_X(u=0)=\mathrm{id}_{\mathscr{P}\Gamma}\end{cases}. \tag{III.11}$$

In greater detail, this means that for all paths γ and functions $f\in C^\infty(\mathscr{P}\Gamma)$

$$\frac{d}{du}\,f\circ\sigma_X(u,\gamma)=X_{\sigma_{X(u,\gamma)}}[f]. \tag{III.12}$$

We shall not try to justify in this infinite dimensional problem that for a fixed path γ such solutions $\sigma_X(u,\gamma)$ exist in some interval $u\in[-\epsilon(\gamma),\epsilon(\gamma)]$. And even worse, these solutions should exist uniformly in γ over path space $\mathscr{P}\Gamma$ for $u\in[0,1]$. To do this properly, we should give a rigorous

definition of an infinite dimensional manifold $\mathscr{P}\Gamma$ of paths. But this would take us too far away from the main scope of this paper. We shall simply assume this can be done. Then the exponential map Exp is defined as $\mathrm{Exp}(X)\equiv\sigma_X(u=1)$. One may show that

$$\sigma_X(u)\circ\sigma_X(v)=\sigma_X(u+v),\quad \sigma_X(u)=\mathrm{Exp}(uX),$$

$$\mathrm{Exp}(uX)\circ\mathrm{Exp}(vX)=\mathrm{Exp}((u+v)X),\quad \mathrm{Exp}(X)^{-1}=\mathrm{Exp}(-X). \tag{III.13}$$

Note that (III.12) implies that

$$\frac{d}{du}\,(\mathrm{Exp}(uX)^*f)\big|_{u=0}=X[f]\equiv\mathscr{L}_X f. \tag{III.14}$$

Together with the boundary condition this yields formally

$$f\circ\mathrm{Exp}(uX)\equiv\mathrm{Exp}(uX)^*f=\sum_{n=0}^{\infty}\frac{1}{n!}(u\mathscr{L}(X))^n f\equiv e^{u\mathscr{L}(X)}f. \tag{III.15}$$

Hence, the formulas for the pull-back and the push-forward read

$$(\mathrm{Exp}(X))^*=e^{\mathscr{L}(X)},$$

$$(\mathrm{Exp}(X))_*Y=Y\circ\sum_{n=0}^{\infty}\frac{1}{n!}X^n\equiv Y\circ e^{\circ X}. \tag{III.16}$$

Moreover, from the reversing property of the pull-back $\sigma_1^*\sigma_2^*=(\sigma_2\circ\sigma_1)^*$, it follows that

$$\mathrm{Exp}(Y)\circ\mathrm{Exp}(X)=\mathrm{Exp}(X)\overset{\mathrm{op}}{\circ}\mathrm{Exp}(Y)=\mathrm{Exp}(\mathrm{BCH}(X,Y)). \tag{III.17}$$

This shows that the Lie algebra of vector fields which acts to the left is naturally associated with the Lie group of *opposite* ordering.

The exponential map $\mathrm{Exp}:\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))\to\mathrm{Diff}(\mathscr{P}\Gamma)$ gives rise to a right action conventionally denoted by a dot

$$.:\mathscr{P}\Gamma\times:\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))\to\mathscr{P}\Gamma. \tag{III.18}$$

Its definition and main properties are

$$\gamma.X=(\mathrm{Exp}(X))\,\gamma,\quad \gamma.0=\gamma,\quad (\gamma.X).Y=\gamma.(\mathrm{BCH}(X,Y)). \tag{III.19}$$

In particular, if the path space $\mathscr{P}\Gamma$ is a right $\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))$ module, i.e., a $C^\infty(\mathscr{P}\Gamma)$ vector space endowed with an action

$$[\,\cdot\,,\cdot\,]:\mathscr{P}\Gamma\times\mathrm{Lie}(\mathrm{Diff}(\mathscr{P}\Gamma))\to\mathscr{P}\Gamma \tag{III.20}$$

satisfying

$$[\gamma,F\ X+G\ Y]=F(\gamma)[\gamma,X]+G(\gamma)[\gamma,Y],$$

$$[a\gamma_1+b\gamma_2,X]=a[\gamma_1,X]+b[\gamma_2,X],$$

$$[[\gamma,X],Y]-[[\gamma,Y]X]-[\gamma,[X,Y]]=0, \tag{III.21}$$

then the exponential map can be understood by simple Taylor expansion:

$$(\mathrm{Exp}(X))\,\gamma = \gamma e^{[\cdot,X]} = \sum_{n=0}^{\infty} \frac{1}{n!} \underbrace{[[\ldots,[\gamma,X],\ldots,X],X]}_{n \text{ commutators}}. \tag{III.22}$$

However, it is unnecessary to assume this.

IV. THE GAUGE-FIXING FLOW

We can now construct a Lie algebra homomorphism $X: \mathcal{G} \to \mathrm{Lie}(\mathrm{Diff}(\mathcal{P}\Gamma))$ that takes a gauge fermion ψ to a bosonic BRST-type vector field X^{ψ} by

$$X: \psi \mapsto X^{\psi} \equiv \frac{i}{\hbar} \psi X_{\Omega}, \quad [X^{\psi_1}, X^{\psi_2}] = X^{(\psi_1,\psi_2)}. \tag{IV.1}$$

The composition of two such vector fields is, remarkably, again a vector field of the same type

$$X^{\psi_1} \circ X^{\psi_2} = X^{\psi_1\{\psi_2,\Omega\}}, \tag{IV.2}$$

because the derivatives of second order cancel. The corresponding Lie group homomorphism

$$\sigma \equiv \mathrm{Exp} \circ X \circ \mathrm{Ln}: G \to \mathrm{Diff}(\mathcal{P}\Gamma) \tag{IV.3}$$

reads

$$\sigma: \Psi \mapsto \sigma_{\Psi} = \mathrm{id}_{\mathcal{P}\Gamma} + \frac{i}{\hbar}\, \Psi X_{\Omega},$$

$$\sigma_{\psi_1 \diamond \Psi_2} = \sigma_{\Psi_1} \overset{\mathrm{op}}{\circ} \sigma_{\Psi_2} = \sigma_{\Psi_2} \circ \sigma_{\psi_1},$$

$$\sigma_{\tau(\Psi)} = (\sigma_{\Psi})^{-1}. \tag{IV.4}$$

This is called a realization of G. As a very important consequence we have a gauge-fixing right group action

$$\diamond : \mathcal{P}\Gamma \times G \to \mathcal{P}\Gamma \tag{IV.5}$$

directly on the path space:

$$\gamma_{\Psi} = \gamma \diamond \Psi = \sigma_{\Psi}(\gamma) = \left(\mathrm{id}_{\mathcal{P}\Gamma} + \frac{i}{\hbar}\, \Psi X_{\Omega} \right) \gamma,$$

$$\gamma \diamond 0 = \gamma,$$

$$(\gamma \diamond \Psi_1) \diamond \Psi_2 = \gamma \diamond (\Psi_1 \diamond \Psi_2). \tag{IV.6}$$

This is the sought after geometric gauge-fixing. The gauge fixing can be understood at the level of paths as a modification of the individual paths!

Again, the expression $\sigma_{\Psi} = \mathrm{id}_{\mathcal{P}\Gamma} + (i/\hbar)\Psi X_{\Omega}$ is quite symbolic, but it is justified by the formula for the pull-back:

$$(\sigma_{\Psi})^{*} = \left(\mathrm{id}_{\mathcal{P}\Gamma} + \frac{i}{\hbar} \Psi X_{\Omega} \right)^{*} = 1 + \frac{i}{\hbar} \mathcal{S}(\Psi X_{\Omega}). \tag{IV.7}$$

The expression can be evaluated directly term by term on the manifold if the path space $\mathcal{P}\Gamma$ is a module wrt. the Lie algebra $(C^{\infty}(\mathcal{P}\Gamma), \{\cdot,\cdot\})$.

Before we carry on, let us for the sake of completeness also define the map

$$\sigma: \psi \mapsto \sigma_\psi = \mathrm{Exp}(X^\psi) = \mathrm{Exp}\left(\frac{1}{\hbar}\,\psi X_\Omega\right): \mathscr{G} \to \mathrm{Diff}(\mathscr{P}\Gamma),$$

$$(\sigma_\psi)^{-1} = \sigma_{-\psi}, \tag{IV.8}$$

and the corresponding right ''gauge-fixing action'' $\mathscr{P}\Gamma \times \mathscr{G} \to \mathscr{P}\Gamma$ on the path space $\mathscr{P}\Gamma$ from the Lie algebra $\mathscr{G}$:

$$\gamma.\psi = \sigma_\psi(\gamma) = \left(\mathrm{Exp}\left(\frac{i}{\hbar}\,\Psi X_\Omega\right)\right)\gamma,$$

$$\gamma.0 = \gamma,$$

$$(\gamma.\psi_1).\psi_2 = \gamma.(\mathrm{BCH}(\psi_1, \psi_2)). \tag{IV.9}$$

A. Gauge-fixed quantities

The gauge-fixed classical observables F_ψ are defined via the pull-back of the gauge-fixing map:

$$\begin{aligned} F_\psi(\gamma) &= (\sigma_\psi^* F)(\gamma) = F(\sigma_\psi(\gamma)) \\ &= F(\gamma) + \frac{i}{\hbar}\,\Psi(\gamma)\{\Omega(\gamma), F(\gamma)\} \\ &= e^{(i/\hbar)\psi(\gamma)\{\Omega(\gamma),\cdot\}} F(\gamma). \end{aligned} \tag{IV.10}$$

The BRST-improved Hamiltonian H_{BRST} and the BRST charge Ω are not changed by the gauge fixing flow, because of (II.6). The gauge-fixed symplectic two-form and the symplectic potential are remarkably simple, but clearly nonlocal:

$$\omega_\Psi = \sigma_\Psi^* \omega = \omega + \frac{i}{\hbar}\, d\Psi \wedge d\Omega, \tag{IV.11}$$

$$\vartheta_\Psi = \sigma_\Psi^* \vartheta = \vartheta + \frac{i}{\hbar}\,\Psi d\Omega + \frac{i}{\hbar}\, d[\Psi i_{X_\Omega}\vartheta]. \tag{IV.12}$$

The gauge-fixed Poisson bracket satisfies

$$\{F_\Psi, G_\Psi\}_\Psi = (\{F, G\})_\Psi, \tag{IV.13}$$

or written more explicitly

$$\begin{aligned} \{F,G\}_\Psi &= \sigma_\Psi^*(\{\sigma_{\tau(\Psi)}^* F, \sigma_{\tau(\Psi)}^* G\}) \\ &= \{F,G\} - \frac{i}{\hbar}\,\{F,\Psi\}\,\frac{1}{1+B}\,\{\Omega,G\} - \frac{i}{\hbar}\,\{F,\Omega\}\,\frac{1}{1+B}\,\{\Psi,G\} \\ &\quad + \left(\frac{i}{\hbar}\right)^2 \{F,\Omega\}\,\frac{\{\Psi,\Psi\}}{(1+B)^2}\,\{\Omega,G\} \\ &= \{F,G\} - \frac{i}{\hbar}\,\{F,\psi e(b)\} e^{-b}\{\Omega,G\} - \frac{i}{\hbar}\,\{F,\Omega\} e^{-b}\{e(b)\psi,G\} \\ &\quad + \left(\frac{i}{\hbar}\right)^2 \{F,\Omega\} e^{-b}\{e(b)\psi,\psi e(b)\} e^{-b}\{\Omega,G\}. \end{aligned} \tag{IV.14}$$

These formulas follow by straightforward application of (IV.10). The derivation becomes somewhat easier if one first derives (IV.14) for $G=F$ being Grassmann odd. Then one may use the polarization formula $\{F,G\}=\frac{1}{2}(\{F+G,F+G\}-\{F,F\}-\{G,G\})$ and finally linearity to extend to functions of both Grassmann-parities. An even simpler way is the following matrix calculation. Define, cf. (IV.11):

$$\omega_{\Psi AB}(t,t')=\omega_{AB}(t,t')+\Delta\omega_{AB}(t,t'),$$

$$\Delta\omega_{AB}(t,t')=\frac{i}{\hbar}\left(\frac{\vec{\delta}^l}{\delta z^A(t)}\,\Psi\right)\left(\Omega\,\frac{\overleftarrow{\delta}^r}{\delta z^B(t')}\right)+\frac{i}{\hbar}\left(\frac{\vec{\delta}^l}{\delta z^A(t)}\,\Omega\right)\left(\Psi\,\frac{\overleftarrow{\delta}^r}{\delta z^B(t')}\right). \tag{IV.15}$$

Then the inverse matrix is

$$[\omega_\Psi]^{-1}=([1]+[\omega]^{-1}[\Delta\omega])^{-1}[\omega]^{-1}$$

$$=\sum_{n=0}^{\infty}(-[\omega]^{-1}[\Delta\omega])^{-1})^n[\omega]^{-1}. \tag{IV.16}$$

It is easy to see that this sums up to (IV.14) with $F=z^A(t)$ and $G=z^B(t')$ substituted.

The BRST closed (and therefore BRST exact) quantities are constant under the BRST gauge-fixing flow. A BRST exact quantity

$$\{\Omega,F\}=\{\Omega,F_\psi\}_\psi,\quad \Omega_\psi=\Omega, \tag{IV.17}$$

can be written BRST exactly wrt. the gauge-fixed structure. As an important consequence the BRST cohomology is not affected by the gauge-fixing.

V. THE BFV PATH INTEGRAL

In this section we shall formally prove the BFV theorem in the path space formulation, by inspecting the various parts of the path integral. We have already argued that the BRST improved Hamiltonian $H_{\rm BRST}$ is not modified by the gauge fixing flow. Below we shall analyze the measure part and the kinetic part.

A. The volume form

Let us denote the volume form by $\Omega_{\rm vol}=\mathscr{D}z\,{\rm Pf}(\omega..)$. The gauge-fixed volume form can remarkably be written as

$$\sigma_\psi^*\Omega_{\rm vol}=\Omega_{\rm vol}\,\frac{1}{1+B}=\Omega_{\rm vol}e^{-(i/\hbar)\{\Omega,\psi\}}. \tag{V.1}$$

It is perhaps surprising that the gauge-fixing term, which can be view as a way to remove zero directions in the Hessian of the action, from a geometric point of view has little to do with the action. In fact, the above shows that it emerges from the measure part, completely dictated by the gauge-fixing flow. Because of the importance of (V.1), let us indicate a proof in the case of a finite number $2n$ of purely bosonic variables $z^A(t)=z^I$, where $(A,t)=I=1,...,2n$, so that

$$\Omega_{\rm vol}=\frac{1}{n!}\,\omega^{\wedge n}=\frac{1}{n!}\,\omega\wedge\cdots\wedge\omega. \tag{V.2}$$

If there are fermionic variables present, the volume form is no longer the nth exterior power of the symplectic two form, and the analysis becomes more involved. However, let's stick to the bosonic case. The gauge-fixed top-form is, cf. (IV.11),

$$\sigma_\Psi^*\Omega_{\rm vol}=\frac{1}{n!}\,\sigma_\Psi^*\Omega\wedge\cdots\wedge\sigma_\Psi^*\omega$$

$$
\begin{aligned}
&= \sum_{r=0}^{n} \frac{1}{(n-r)!r!}\, \omega^{\wedge(n-r)} \wedge \left(\frac{i}{\hbar}\, d\Psi \wedge d\Omega \right)^{\wedge r} \\
&= \sum_{r=0} \frac{1}{(n-r)!r!}\, \omega^{\wedge(n-r)} \wedge \sum_{I_1,\ldots,I_r} \left(\frac{i}{\hbar}\, dz^{I_1} \frac{\delta\Psi}{\delta z^{I_1}} \wedge d\Omega \right) \wedge \cdots \wedge \left(\frac{i}{\hbar}\, dz^{I_r} \frac{\delta\Psi}{\delta z^{I_r}} \wedge d\Omega \right).
\end{aligned}
\tag{V.3}
$$

Because of covariance, one may resort to Darboux coordinates $z^I=(q^i,p_i)$, $i=1,\ldots,n$. Then

$$
\{q^i,p_j\}=\delta^i_j, \quad \omega = dp_i \wedge dq^i. \tag{V.4}
$$

A simple combinatorial argument yields

$$
\begin{aligned}
\sigma^*_\Psi \Omega_{\mathrm{vol}} &= (dp_1 \wedge dq^1) \wedge \cdots \wedge (dp_n \wedge dq^n) \sum_{r=0}^{n} \sum_{i_1,\ldots,i_r} \left(-\frac{i}{\hbar} \right) \left(\frac{\delta\Psi}{\delta q^{i_1}} \frac{\delta\Omega}{\delta p_{i_1}} - \frac{\delta\Psi}{\delta p_{i_1}} \frac{\delta\Omega}{\delta q^{i_1}} \right) \\
&\quad \times \cdots \times \left(-\frac{i}{\hbar} \right) \left(\frac{\delta\Psi}{\delta q^{i_r}} \frac{\delta\Omega}{\delta p_{i_r}} - \frac{\delta\Psi}{\delta p_{i_r}} \frac{\delta\Omega}{\delta q^{i_r}} \right) \\
&= \Omega_{\mathrm{vol}} \sum_{r=0}^{n} (-B)^r = \Omega_{\mathrm{vol}}\, \frac{1-(-B)^{n+1}}{1+B}.
\end{aligned}
\tag{V.5}
$$

Now we formally take the $n=\infty$ limit to derive (V.1).

B. The kinetic part

We have already analyzed the variation of the Hamiltonian part and the measure under the gauge-fixing flow. Let us now draw our attention to the last piece of the path integral, namely the kinetic term. It is handy to introduce a vector field T that calculates the difference between the total time derivative and the explicit time derivative:

$$
T \equiv \frac{d}{dt} - \frac{\partial}{\partial t} = \int_\gamma dt\, \dot{\gamma}^A(t)\, \frac{\vec{\delta}^l}{\delta z^A(t)} : \; C^\infty(\mathscr{P}\Gamma) \to C^\infty(\mathscr{P}\Gamma). \tag{V.6}
$$

The kinetic term can be written as the contraction of the symplectic potential with this vector field:

$$
i_T \upsilon(\gamma) = \int_\gamma dt\; \vartheta_A(t)\, \dot{\gamma}^A(t), \quad \vartheta(\gamma) = \int_\gamma dt\; \vartheta_A(t) dz^A(t). \tag{V.7}
$$

We can now write the gauge-fixed BFV path integral as

$$
Z_\psi = \int \Omega_{\mathrm{vol}} e^{(i/\hbar)(i_T\vartheta - H_{\mathrm{BRST}} - \{\Omega,\psi\})}. \tag{V.8}
$$

We would like to only sum over paths in the path integral so that the kinetic term for these paths is left invariant under the gauge-fixing flow. In fact, if this is fulfilled, we have realized gauge-fixing as an internal diffeomorphism of the path integral. [Perhaps it is helpful at this point to recall that every integral $I=\int f(x)dx=\int f(\sigma(x))d\sigma(x)$ by the very definition is diffeomorphism invariant; x is just a dummy variable.] Thus the path integral cannot depend on the gauge fermion and we can conclude the BFV theorem. To achieve this conclusion, one should restrict the path space by only allowing for paths satisfying the following BRST boundary condition

$$
(i_{X_\Omega}\vartheta)(\gamma)\big|_{t=0} + \Omega(\gamma)\big|_{t=0} = (i_{X_\Omega}\vartheta)(\gamma)\big|_{t=T} + \Omega(\gamma)\big|_{t=T}. \tag{V.9}
$$

Because of the formula

$$i_{X_\Omega}\vartheta(\gamma)=\int_\gamma dt\ \vartheta_A(t)X^A_\Omega(t) \tag{V.10}$$

and because Ω is BRST invariant, we have

$$\begin{aligned} X_\Omega[i_{X_\Omega}\vartheta](\gamma)&=\int_\gamma\int_\gamma dt\ dt' X^B_\Omega(t')\ \frac{\overrightarrow{\delta}^l}{\delta z^B(t')}\,[\vartheta_A(t)X^A_\Omega(t)]\\ &=\frac{1}{2}\int_\gamma dt[X_\Omega,X_\Omega]^A(t)\vartheta_A(t)(-1)^{\epsilon_A}+\int_\gamma\int_\gamma dt\ dt' X^A_\Omega(t)X^B_\Omega(t')\ \frac{\overrightarrow{\delta}^l}{\delta z^B(t')}\ \vartheta_A(t)\\ &=\frac{1}{2}\int_\gamma dt\ X^A_{\{\Omega,\Omega\}}(t)\vartheta_A(t)(-1)^{\epsilon_A}+\frac{1}{2}\int_\gamma\int_\gamma dt\ dt' X^A_\Omega(t)X^B_\Omega(t')\omega_{BA}(t',t)=0. \end{aligned} \tag{V.11}$$

Together with $X_\Omega[\Omega]=0$ this yields that the above boundary condition on the paths is stabile under the gauge-fixing action of $\mathscr{G}$. We will assume that a vector field

$$X:\gamma\mapsto X_\gamma=\int_\gamma dt\ X^A(t)\ \frac{\overrightarrow{\delta}^l}{\delta z^A(t)} \tag{V.12}$$

has the following dependence on the path γ:

$$X^A(t)=X^A(\gamma,\gamma(t),t). \tag{V.13}$$

Under the assumption that the original symplectic potential $\vartheta_A(t)=\vartheta_A(\gamma(t))$ is ultra-local and that it has no explicit time dependence (and therefore the symplectic two form is also of that kind), one derives that a generic vector field X changes the kinetic term with

$$X[i_T\vartheta](\gamma)=i_X\vartheta|_{\gamma,\gamma(t=T),t=T}-i_X\vartheta|_{\gamma,\gamma(t=0),t=0}+\int_\gamma dt\ X^A(t)\omega_{AB}(t)\dot{z}^B(t). \tag{V.14}$$

Assuming that Ω is ultra-local and that it has no explicit time dependence and using the boundary condition (V.9), we therefore conclude that the kinetic term is invariant under the gauge-fixing flow $X=\psi X_\Omega$.

C. Infinitesimal gauge-fixing flow

Above we have argued from a formal point of view that turning on or turning off gauge fixing does not change the value of the path integral. In real applications, like quantum field theory, turning off gauge fixing makes the path integral ill-defined, as a result of integrating over an infinite gauge volume. That is why gauge fixing is introduced in the first place. It is therefore desirable to have an infinitesimal analysis.

Consider the map

$$L_{\mathrm{Exp}}:\mathscr{G}\to\mathrm{Aut}(G):\psi\mapsto L_{\mathrm{Exp}(\psi)}, \tag{V.15}$$

where L_Ψ is the left multiplication map with a fixed Ψ

$$L_\Psi(\Psi')=\Psi\diamond\Psi'. \tag{V.16}$$

Then by definition

$$L_{\mathrm{Exp}(0)}=\mathrm{id}_G. \tag{V.17}$$

Let us consider an infinitesimal change $\psi\to\psi+\delta\psi$ in the gauge fermion. The relevant object is the differential of L_{Exp} in $\psi=0$,

$$(L_{\mathrm{Exp}})_{*0}: T_0\mathscr{G} \to T_{\mathrm{id}_G}\mathrm{Aut}(G). \tag{V.18}$$

It reads

$$\Psi + \delta\Psi = ((L_{\mathrm{Exp}})_{*0}\delta\psi)(\Psi) = h_\Psi(\delta\psi) \diamond \Psi. \tag{V.19}$$

Here the function $h: \mathscr{G} \times T_0\mathscr{G} \to \mathscr{G}$ is

$$\begin{aligned} h_\psi(\delta\psi) &= (\Psi + \delta\Psi) \diamond \tau(\Psi) = \frac{\delta\Psi}{1+B} = \frac{\mathrm{Exp}_{*\psi}(\delta\psi)}{1+B} \\ &= e^{-b} e(b)\,\delta\psi + e^{-b} e'(b)\,\psi\{\Omega, \delta\psi\} \\ &= \frac{1-e^{-b}}{b}\,\delta\psi + \frac{b-1+e^{-b}}{b^2}\,\psi\{\Omega, \delta\psi\}. \end{aligned} \tag{V.20}$$

We have identified the tangent space $T_{\mathrm{id}_G}\mathrm{Aut}(G)$ with maps from the group G that to each element Ψ of the group G give an element in $\mathscr{G} \diamond \Psi$.

So the infinitesimal change in the gauge-fixing flow $\sigma_{\psi+\delta\psi} \circ \sigma_{-\psi}$ is governed by a BRST-type vector field, cf. (IV.6),

$$X^{h_\psi(\delta\psi)} = h_\psi(\delta\psi) X_\Omega. \tag{V.21}$$

Note that

$$\{\Omega, h_\psi(\mu)\} = \{\Omega, \mu\}, \quad h_{\alpha\psi}(\psi) = \alpha\psi. \tag{V.22}$$

We can now derive the infinitesimal change of various objects under an infinitesimal change $\psi \to \psi + \delta\psi$ of the gauge fermion caused by the vector field $X^{h_\psi(\delta\psi)}$. For instance, the measure changes according to

$$\mathrm{div}_\rho\, X^{h_\psi(\delta\psi)} = -\frac{i}{\hbar}\{\Omega, \delta\psi\}. \tag{V.23}$$

This is precisely the infinitesimal change in the gauge-fixed action.

VI. THE Sp(2) CASE

The whole construction can easily be extended to the Sp(2) case.[4] We shall thus be brief and just state the construction.

In the Sp(2) case we have two odd BRST charges Ω^a, $a = 1,2$, that are mutual orthogonal in the Poisson bracket sense:

$$\{\Omega^a, \Omega^b\} = 0. \tag{VI.1}$$

We define a Grassmann even bracket

$$[F, G] = Fk(G) - k(F)G = -(-1)^{\epsilon_F \epsilon_G}[G, F], \tag{VI.2}$$

where k is a Grassmann even nilpotent second order operator

$$k = \frac{i}{2\hbar}\,\epsilon_{ab} X_{\Omega^a} \circ X_{\Omega^b}, \quad k^2 = 0,$$

$$k(F) = \frac{i}{2\hbar}\,\epsilon_{ab}\{\Omega^a, \{\Omega^b, F\}\} = \frac{i}{2\hbar}\,\{\{F, \Omega^a\}, \Omega^b\}\epsilon_{ab}. \tag{VI.3}$$

The Grassmann even bracket satisfies the usual symmetry property, the Jacobi identity, but not the Poisson property. We define a degenerate metric structure (a covariant 2-tensor)

$$g(F,G)=k(F)G+Fk(G)-k(FG)$$
$$=\frac{i}{\hbar}\{F,\Omega^a\}\epsilon_{ab}\{\Omega^b,G\}$$
$$=+(-1)^{\epsilon_F\epsilon_G}g(G,F). \tag{VI.4}$$

k and g satisfy the following relations

$$g(k(F),G)=0,\quad k(g(F,G))=2k(F)k(G), \tag{VI.5}$$

$$g(F,g(G,H))=\tfrac{1}{2}\,g(F,G)k(H)+\tfrac{1}{2}\,g(F,H)k(G)(-1)^{\epsilon_G\epsilon_H}. \tag{VI.6}$$

We have a Jacobi-like identity:

$$g(F,g(G,H))-(-1)^{\epsilon_F\epsilon_G}g(G,g(F,H))=g([F,G],H). \tag{VI.7}$$

The gauge is now chosen by specifying a gauge *boson* ϕ. (Our gauge boson is *minus a half* of the BLT gauge boson.) The even bracket $[\,\cdot\,,\cdot\,]$ is a Lie-bracket in the algebra $\mathscr{G}$ of gauge bosons. The space $G=\mathscr{G}$ of gauge bosons becomes a Lie group by introducing a product and an inversion

$$\Phi_1\diamond\Phi_2=\Phi_1+\Phi_2+\Phi_1k(\Phi_2),$$
$$\tau(\Phi)=-\frac{\Phi}{1+B}=-\Phi e^{-\bar{b}}, \tag{VI.8}$$

where

$$\bar{b}=\frac{i}{2\hbar}\,\epsilon_{ab}\{\Omega^a,\{\Omega^b,\phi\}\}=k(\phi)=\mathrm{Ln}(1+\bar{B}),$$
$$\bar{B}=\frac{i}{2\hbar}\,\epsilon_{ab}\{\Omega^a,\{\Omega^b,\Phi\}\}=k(\Phi)=e^{\bar{b}}-1, \tag{VI.9}$$

and ϕ and Φ are connected via the bijective exponential map Exp: $\bar{\mathscr{G}}\to\bar{G}$

$$\Phi=\mathrm{Exp}(\phi)=\phi e(\bar{b}),\quad \phi=\mathrm{Ln}(\Phi)=\Phi l(\bar{B}). \tag{VI.10}$$

We can define a homomorphism $\Psi_a:\bar{\mathscr{G}}\to\mathscr{G}_a$ that transforms a gauge boson ϕ into a gauge fermion. [To be precise, $\Psi_a(\phi)=1/2\psi$ is *half* the corresponding gauge fermion of the standard BFV-BRST construction.] $\Psi_a(\phi)$ of the kind $a=1,2$ (corresponding to the generator Ω^a)

$$\Psi_a(\,\cdot\,)=\epsilon_{ab}X_{\Omega^b}(\,\cdot\,)$$
$$\Psi_a(F)=\epsilon_{ab}\{\Omega^b,F\}=(-1)^{\epsilon_F}\{F,\Omega^b\}\epsilon_{ba},$$
$$\Psi_a([\phi_1,\phi_2])=(\Psi_a(\phi_1),\Psi_a(\phi_2))^a,$$
$$\Psi_a(\Phi_1\diamond\Phi_2)=\Psi_a(\Phi_1)\diamond\Psi_a(\Phi_2) \tag{VI.11}$$

(no sum over a). Note that

$$\frac{i}{\hbar}\{\Omega^a,\Psi_b(F)\}=\delta^a_bk(F). \tag{VI.12}$$

The algebra homomorphism $\bar{X}:\bar{\mathscr{G}}\to\mathrm{Lie(Diff)}(\mathscr{P}\Gamma))$ is the sum of the two BRST homomorphisms

$$\overline{X}:\phi\mapsto\overline{X}^{\phi}\equiv g(\phi,\cdot)=\frac{i}{\hbar}\,\Psi_a(\phi)X_{\Omega^a},\quad [\overline{X}^{\phi_1},\overline{X}^{\phi_2}]=\overline{X}^{[\phi_1,\phi_2]}. \tag{VI.13}$$

The corresponding Lie group homomorphism $\overline{\sigma}\equiv\mathrm{Exp}\circ\overline{X}\circ\mathrm{Ln}:\overline{G}\to\mathrm{Diff}(\mathscr{P}\Gamma)$ reads

$$\begin{aligned}&\overline{\sigma}:\Phi\mapsto\overline{\sigma}_{\Phi}=\mathrm{id}_{\mathscr{P}\rho\Gamma}+g(\Phi,\cdot)+\tfrac{1}{2}\,g(\Phi,\Phi)k(\cdot),\\ &\overline{\sigma}_{\Phi_1\diamond\Phi_2}=\overline{\sigma}_{\Phi_1}\overset{\mathrm{op}}{\circ}\overline{\sigma}_{\Phi_2},\\ &\overline{\sigma}_{\tau(\Phi)}=(\overline{\sigma}_{\Phi})^{-1}.\end{aligned} \tag{VI.14}$$

The pull-back map $(\overline{\sigma}_{\Phi})^*$ is

$$(\overline{\sigma}_{\Phi})^*=1+\frac{i}{\hbar}\,\Psi_a(\Phi)\mathscr{L}(X_{\Omega^a})+\frac{i}{2\hbar}\,\epsilon_{ab}g(\Phi,\Phi)\mathscr{L}(X_{\Omega^a})\circ\mathscr{L}(X_{\Omega^b}). \tag{VI.15}$$

The gauge-fixed symplectic two form reads

$$\omega_{\Phi}=\overline{\sigma}_{\Phi}^{*}\omega=\omega+\frac{i}{\hbar}\,d\Psi_a(\Phi)\wedge d\Omega^a. \tag{VI.16}$$

The gauge-fixed volume form reads

$$\overline{\sigma}_{\phi}^{*}\Omega_{\mathrm{vol}}=\Omega_{\mathrm{vol}}\,\frac{1}{(1+\overline{B})^2}=\Omega_{\mathrm{vol}}e^{-(i/\hbar)\epsilon_{ab}\{\Omega^a,\{\Omega^b,\phi\}\}}. \tag{VI.17}$$

ACKNOWLEDGMENTS

The author would like to thank I. A. Batalin, P. H. Damgaard, A. M. Semikhatov, A. P. Nersessian, and O. M. Khudaverdian for fruitful discussions. The work is partially supported by Nordita.

[1] C. Becchi, A. Rouet, and R. Stora, Commun. Math. Phys. **42**, 127 (1975); Ann. Phys. (N.Y.) **98**, 287 (1976); I. V. Tyutin, *Gauge Invariance in Field Theory and in Statistical Physics in the Operator Formalism*, Lebedev preprint FIAN No. 39 (1975) (unpublished).
[2] E. S. Fradkin and G. A. Vilkovisky, Phys. Lett. B **55**, 224 (1975); I. A. Batalin and G. A. Vilkovisky, Phys. Lett. B **69**, 309 (1977); E. S. Fradkin and T. E. Fradkina, Phys. Lett. B **72**, 343 (1978); I. A. Batalin and E. S. Fradkin, Phys. Lett. B **122**, 157 (1983).
[3] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, NJ, 1992); M. Henneaux, Phys. Rep. **126**, 1 (1985).
[4] I. A. Batalin, P. M. Lavrov, and I. V. Tyutin, J. Math. Phys. **31**, 6 (1990); **31**, 2708 (1990); Int. J. Mod. Phys. A **6**, 3599 (1991).
[5] M. Nakahara, *Geometry, Topology and Physics* (IOP, Bristol, 1990).

Imaginary parts of Stark–Wannier resonances

Vladimir Buslaev
Saint-Petersburg State University, Department of Mathematical Physics, Saint-Petersburg 198904 Russia, and Universté de Paris-Nord, Département de Mathématiques, CNRS—URA 742, 93430 Villetaneuse, France

Alain Grigis
Université de Paris-Nord, Département de Mathématiques, CNRS—URA 742, 93430 Villetaneuse, France

(Received 19 May 1997; accepted for publication 8 December 1997)

We consider a one-dimensional Stark–Wannier Hamiltonian, $H=-d^2/dx^2+p(x)-\epsilon x$, $x\in\mathbb{R}$, where p is a smooth periodic, finite-gap potential, and $\epsilon>0$ is small enough. We compute rigorously the imaginary parts of the spectral resonances. For this purpose we develop some related elements of the adiabatic approach to the equations of the form $-\psi''+p(x)\psi+q(\epsilon x)\psi=E\psi$, $\epsilon\to 0$. © *1998 American Institute of Physics.* [S0022-2488(98)03104-1]

I. INTRODUCTION

Stark–Wannier resonances are related with the operator H, which is defined by the expression

$$H=-\frac{d^2}{dx^2}+p(x)-\epsilon x, \quad x\in\mathbb{R}. \tag{1}$$

Here $p(x)$ is a smooth real-valued periodic function, $p(x+a)=p(x)$, and ϵ is a positive small parameter. The operator H is considered as self-adjoint in $L^2(\mathbb{R})$ on the natural domain. The operator has a simple absolutely continuous spectrum coinciding with all the real axis $\mathbb{R}$.

Let us describe the resolvent $R(E)=(H-E)^{-1}$. In Ref. 1 some solutions of the equation

$$-\psi''+p(x)-\epsilon x\psi=E\psi, \tag{2}$$

with characteristic behavior as $x\to\pm\infty$, have been constructed. Assume as in Ref. 1 that

$$\int_0^a p(x)dx=0. \tag{3}$$

Then there exists the solution $g_-=g_-(x,E)$, which possesses the following asymptotic behavior as $x\to-\infty$:

$$g_-\sim\frac{1}{2|\xi+E|^{1/4}}\,e^{-(2/3\epsilon)|\xi+E|^{3/2}}, \quad \xi=\epsilon x. \tag{4}$$

It is defined for $E\in\mathbb{C}$ and is an entire function of E. There exists also the solution $f_+=f_+(x,E)$ with the following behavior for $x\to+\infty$:

$$f_+\sim\frac{1}{\sqrt{2}|\xi+E|^{1/4}}\,e^{i(2/3\epsilon)|\xi+E|^{3/2}}, \quad \xi=\epsilon x. \tag{5}$$

It is defined as a continuous function of E for $\mathrm{Im}\,E\geqslant 0$ and is a holomorphic function for $\mathrm{Im}\,E>0$.

It is clear that

$$g_-=\overline{M}f_+ + M\overline{f_+}, \tag{6}$$

 © 1998 American Institute of Physics

where $M=M(E)$ is continuous for $\operatorname{Im} E\geqslant 0$ and is holomorphic for $\operatorname{Im} E>0$. The function M is periodic: $M(E+\epsilon a)=M(E)$.

The resolvent is the integral operator with the kernel

$$R(x,y,E)=iN(x,y,E)/M(E), \quad \operatorname{Im} E>0,$$
$$N(x,y,E)=g_-(x,E)f_+(y,E), \quad x\leqslant y. \tag{7}$$

We shall assume that $p(x)$ is holomorphic in the strip $|\operatorname{Im} x|<\alpha$. Then the solution f_+ and the function M admit an analytical continuation to the extended half-plane $\operatorname{Im} E>-\epsilon\alpha$. The function M does not have roots when $\operatorname{Im} E\geqslant 0$, but can have them in the strip $-\epsilon\alpha<\operatorname{Im} E<0$. Namely these roots are Stark–Wannier resonances. They create periodic lattices with the period ϵa.

The question of the existence of the resonances and of the description on their parameters has a long history. Comments on this account can be found in Ref. 2. In the last decade new works appeared, where the question was treated from several different points of view. Let us mention some of these works: Refs. 3–8. Studies begun in these works are being continued in the last years; see, for example, Refs. 9 and 10. We will not enter in the history in more detail. Our main aim is the development of the results of the paper in Ref. 1.

In Ref. 1 the potential p was assumed to be a finite-gap potential and the parameter ϵ to be small. As a confirmation of old results of theorists (which should be considered as conjectures from the math–physics point of view), it was proved that there is a correspondence between the spectral bands of the operator H_0 with periodic potential

$$H_0=-\frac{d^2}{dx^2}+p(x), \tag{8}$$

and the periodic ladders of Stark–Wannier resonances. Their real parts were described asymptotically by some quantization conditions. The problem of the description of the imaginary parts of the resonances is more subtle. In Ref. 1, it was shown that for the periodic ladder of resonances E_n^j, l is the number of the ladder, n is the number of the resonance:

$$\operatorname{Im} E_n^l\sim -a_l\epsilon e^{-(1/\epsilon)A_l}, \quad a_l,A_l>0. \tag{9}$$

However, this result was not strictly proved in Ref. 1, since the order of error obtained in Ref. 1 was worse than the order of the main asymptotical term itself.

The goal of the present work is the justification of formula (9). We will show that this result can be obtained by the methods of Ref. 1, after their natural development. We will use the present paper also in order to improve these methods in different directions, in particular, for the description of the solutions of Eq. (2) in vicinities of turning points. Another approach to the treatment of imaginary parts of the resonances was suggested recently in Ref. 10.

As a technical basis for the study of the resonances we will use the adiabatical approach to the equations of the form

$$-\psi''+v(x,\epsilon x)\psi=E\psi, \tag{10}$$

where $v(x,\xi)$ is a periodic function of the first argument, $v(x+a,\xi)=v(x,\xi)$. General initial facts about the adiabatical method and its spectral applications can be found in Refs. 2 and 11. The adiabatical method can be considered as a generalization of the standard Wentzel–Kramers–Brillouin (WKB) method for the equation

$$-\psi''+v(\epsilon x)\psi=E\psi. \tag{11}$$

In this approach, as in the standard WKB approach, the notion of the turning point is very essential. In such a point the character of the asymptotic behavior of solutions changes crucially. For a description of the asymptotics of the solutions in vicinities of the turning points in the standard WKB method there exist different approaches. We discuss here analogs of these approaches in more detail than in the previous publications.

The structure of the text is the following. In Sec. II we are reminded of some general properties of the solutions of the Schrödinger equation with purely periodic potentials. We represent these properties in a form that is the most suitable for the further applications to the adiabatic perturbations. In particular, we pay special attention to problems with some additional external parameter. Sections III A and III B contain a list of main properties of solutions of the adiabatically perturbed problem with the linear perturbing potential. We do not prove these properties in these sections and simply formulate them. In the rest of the section we use the listed properties in order to compute the imaginary parts of the resonances. The two last sections are devoted to the justifications of the used properties of the solutions of the adiabatically perturbed equations. Section IV contains a description of the formal asymptotic solutions in different circumstances: out of vicinities of the turning points and in the vicinities of such points. In the last section, Sec. V, we construct true solutions of the perturbed equation, which can be described asymptotically by the previously introduced formal solutions. Namely, these solutions and their asymptotics are used in Sec. III in order to compute the imaginary parts of the resonances.

II. THE SCHRÖDINGER EQUATION WITH PERIODIC POTENTIALS

This part contains some necessary information on the solutions of the equations with purely periodic potentials:

$$-u''+p(x)u=Eu. \tag{12}$$

We will also describe the notion of the isoenergy curve that appears for the equation

$$-u''+v(x,\xi)u=Eu, \tag{13}$$

with the periodic potential $v(x+a,\xi)=v(x,\xi)$ depending on the external parameter ξ.

A. The Schrödinger equation with periodic potentials

Consider the solutions of (12) of the form

$$u(x)=e^{ikx}\varphi(x), \tag{14}$$

where k is a given parameter (quasimomentum), $\varphi(x+a)=\varphi(x)$. For the fixed real k there exists a countable set of such solutions:

$$u_l(x,k)=e^{ikx}\varphi_l(x,k), \quad E=\mathscr{E}_l(k)\in\mathbb{R}, \quad l=1,2,\ldots . \tag{15}$$

The functions $\mathscr{E}_l(k)$ are periodic with the period $a^*=2\pi/a$, the solutions $u_l(x,k)$ can also be considered as periodic with respect to k. Properties of the plots of $\mathscr{E}_l$ are reflected by the Fig. 1.

Let $\min \mathscr{E}_l=E_{2(l-1)}$, $\max \mathscr{E}_l=E_{2l-1}$. For special potentials the equality $E_{2l-1}=E_{2l}$ is possible. Out of such levels:

$$\mathscr{E}_1(k)=E_0+\frac{1}{2M_0}k^2+\ldots, \quad \mathscr{E}_1(k)=E_1-\frac{1}{2M_1}(k-k^*)^2+\cdots,$$

$$\mathscr{E}_2(k)=E_2+\frac{1}{2M_2}(k-k^*)^2+\cdots, \quad \mathscr{E}_2(k)=E_3-\frac{1}{2M_3}k^2+\cdots,$$

$$k^*=\tfrac{1}{2}a^*, \quad M_0,M_1,\ldots>0. \tag{16}$$

Other critical points except 0 (mod k^*), are absent. The intervals $\Delta_l=[E_{2(l-1)},E_{2l-1}]$ are spectral intervals of Eq. (12) considered on the entire axis. The intervals $\Delta_0=(-\infty,E_0)$, $\Delta_l=(E_{2l-1},E_{2l})$ are the gaps in the spectrum of the equation. The number of the nonempty gaps can be finite. The corresponding potentials are called the finite-gap potentials. If the number of the finite gaps is Q, the potential is called the Q-gap potential. We will assume that the potential p is a finite-gap potential but shall not use the known explicit formulas for such potentials. For simplicity we will assume that the open gaps are the first gaps.

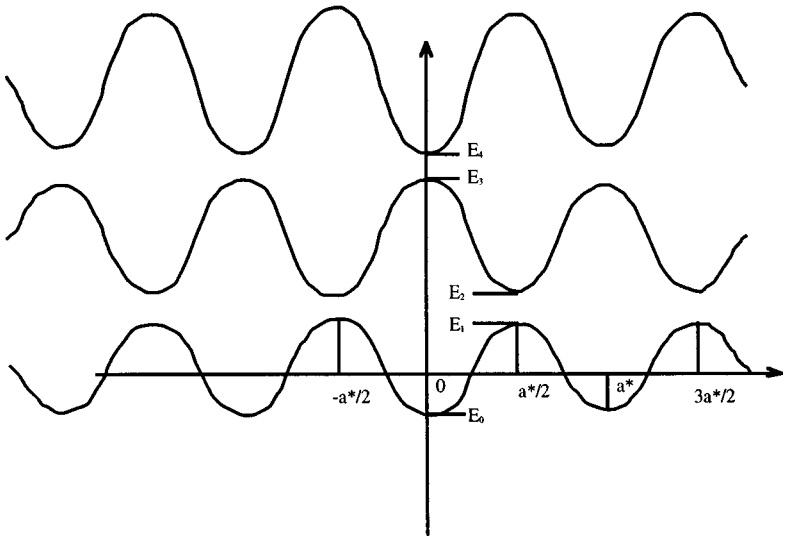

FIG. 1.

The function $\mathscr{E}_l$ can be continued analytically on the complex plane with some cuts; see Fig. 2. In fact, it is a function on the complex cylinder $\mathbb{C}/\{a^*\mathbb{Z}\}$ with the cuts. The parameters h_l, h_{l-1} depend on the potential p.

Consider the function $\mathscr{E}_l(k)$ on the intervals $((l-1)k^*, lk^*), (-lk^*, -(l-1)k^*)$. On the real axis some discontinuous function $\mathscr{E}(k)$ arises that can be analytically continued on the cut plane $\mathbb{K}$, with the cuts $it+k_l^*$, $k_l^*=lk^*$, $t\in(-h_l, h_l)$, $h_l\geqslant 0$; see Fig. 3.

The surface $\mathbb{K}$ is symmetrical with respect to the reflection $k\to -\overline{k}$. The shores of the cuts symmetrical with respect to this map should be identified. The function $\mathscr{E}(k)$ is real on the real axis and on the shores of the cuts (and only on this set). The function $\mathscr{E}(k)$ maps the broken line on Fig. 3 monotonously to the real line, and the image of the positive imaginary axis coincides with $(-\infty, E_0)$, the images of the real intervals are the spectral intervals, and the images of the loops around the cuts coincide with the gaps.

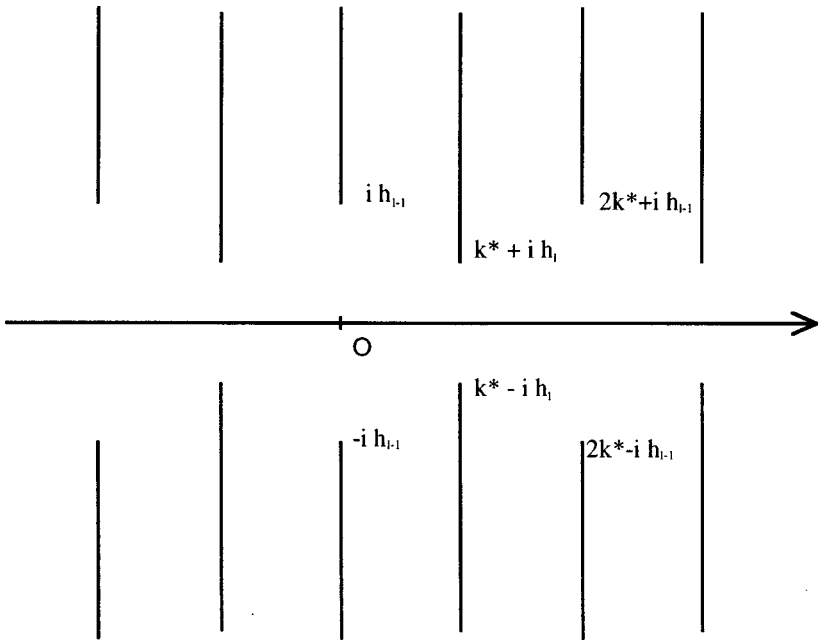

FIG. 2.

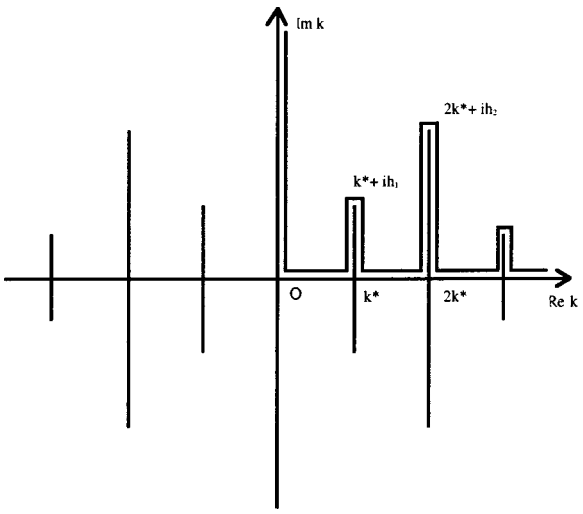

FIG. 3.

Asymptotically, when $k\to\infty$, $\mathscr{E}(k)\sim k^2$.

The solution $u_l(x,k)$ is defined by l and k up to the normalization. We can say that on the cut cylinder $\mathbb{C}/\{a^*\mathbb{Z}\}$ as a base, some bundle $\mathbb{P}_l$ is naturally defined. Also one can consider the bundle $\mathbb{P}$ over $\mathbb{K}$.

We can introduce special normalizations of $u_l(x,k)$ on $\mathbb{K}_l$ and $u(x,k)$ on $\mathbb{K}$ such that these functions will be defined as analytic functions of $k\in\mathbb{K}_l$ and $k\in\mathbb{K}$. These normalizations can be fixed appropriately such that $u(x,k)$ becomes real on the imaginary axis and $u(x,k)=e^{ikx}\varphi(x,k)\sim e^{ikx}$ as $k\to\infty$. The solutions can also be chosen real on the shores of the cuts. For all E except the boundaries of the intervals Δ_l, there are two linearly independent solutions $u(x,k)$ and $u(x,-k)$.

In order to avoid future misunderstandings it is worth mentioning that one of the real parts on the branches of the shores solutions that corresponds to a given real part nk^* of k, must have, at least, one zero in some point k of a shore; if it is chosen such that it does not have roots everywhere on the shores, it is not continuous in one of the points nk^*-0 or nk^*+0. However, this moment does not play any role in the following asymptotic constructions, since the continuity of the solutions in both the points nk^*+0 and nk^*-0 of the cut complex plane is not essential for them.

On the bundles $\mathbb{P}_l$ and $\mathbb{P}$ we can consider natural linear connections that are defined by the connection forms:

$$\omega_l=-\frac{\int_0^a\varphi_l(x,-k)d_k\varphi_l(x,k)dx}{\int_0^a\varphi(x,-k)\varphi(x,k)dx}, \tag{17}$$

$$\omega=-\frac{\int_0^a\varphi(x,-k)d_k\varphi(x,k)dx}{\int_0^a\varphi(x,-k)\varphi(x,k)dx}. \tag{18}$$

By this we can define the continuation of u_l/u along any oriented curve γ. If γ connects k_0 to k_1:

$$(u_l/u)(x,k_0)\to(u_l/u)(x,k_1)e^{\int_\gamma\omega}. \tag{19}$$

The forms are purely imaginary if $k\in\mathbb{R}$ and purely real on the imaginary axis and on the shores of the cuts.

Analogously, two similar forms corresponding to the expression

$$\widetilde{\omega}=-\frac{1}{a}\frac{\int_0^a\{u_2,d_ku_1\}dx}{\{u_2,u_1\}}, \tag{20}$$

will be useful. Here $\{f,g\}=fg'-f'g$ is the standard Wronskian and $u_1=u(x,k)$, $u_2=u(x,-k)$.

The following elementary formulas will be also essential:

$$\{u_1,u_2\}=-i\mathscr{E}'(k)\,\frac{1}{a}\int_0^a u_1u_2dx, \tag{21}$$

$$\widetilde{\omega}+i\,\frac{a}{2}\,dk=\omega-\frac{1}{2}\,d(\ln\,\mathscr{E}'). \tag{22}$$

B. Isoenergy curve

Here we will make some remarks on the structure of the curves on the ''phase plane'' defined by the equations

$$\mathscr{E}(k)+q(\xi)=E,\quad k\in\mathbb{K},\quad \xi\in\mathbb{R}, \tag{23}$$

$$\mathscr{E}_l(k)+q(\xi)=E,\quad k\in\mathbb{K}_l,\quad \xi\in\mathbb{R}. \tag{24}$$

Here $q(\xi)\in\mathbb{R}$ is a given smooth function. We will assume that on each finite interval the function q has only a finite number of nondegenerated critical points. Let us call these curves the isoenergy curves. For the Stark–Wannier case (2), $q(\xi)=-\xi$.

Consider the real plane $(\xi,\mathscr{E})$ and the curves $\mathscr{E}=E_m+q(\xi)$ on the plane. They generate on the plane some strips. The strip $\sigma(\underline{\Delta}_l)$, between $\mathscr{E}=E_{2l-2}+q(\xi)$ and $\mathscr{E}=E_{2l-1}+q(\xi)$, will be called admissible, and the strip $\sigma(\Delta_l)$, between $\mathscr{E}=E_{2l-1}+q(\xi)$ and $\mathscr{E}=E_{2l}+q(\xi)$, will be called forbidden.

Let us consider the straight line $\mathscr{E}=E$ with a fixed E, $E\in\mathbb{R}$. The values of ξ corresponding to the intersection points of this straight line and all the curves $\mathscr{E}=E_m+q(\xi)$ are called the turning points. They are solutions of the equations $E=E_m+q(\xi)$. A turning point ξ_1 is called a simple turning point if $q'(\xi_1)\neq 0$. The turning points play an essential role in the asymptotic theory of Eq. (10). In this text we will deal only with the simple turning points. For fixed E on each finite interval of the axis ξ, we can find only a finite number of the turning points.

Consider an interval between two neighboring simple turning points ξ_1 and ξ_2. As such, the intervals will be called the minimal intervals. Two cases are possible: either the points ξ_1,ξ_2 are the coordinates of the intersection of the straight line $\mathscr{E}=E$ with one curve $\mathscr{E}=E_m+q(\xi)$ [cases (a) and (b)], or they are the coordinates of the intersection with different curves $\mathscr{E}=E_m+q(\xi)$ and $\mathscr{E}=E_{m'}+q(\xi)$ [cases (c) and (d)], $m\neq m'$. Both the cases in their turn can be divided in two subcases: the line $\mathscr{E}=E$ between the intersections lies in an allowed strip [cases (a) and (c)] or it lies in a forbidden strip [cases (b) and (d)]; see Fig. 4, where the forbidden strips are striped.

Let us comment briefly on the figure. In Fig. 4(a), the straight line $\mathscr{E}=E$ intersects twice the same curve $\mathscr{E}=E_m+q(\xi)$, and its part between ξ_1,ξ_2 belongs to a forbidden strip. In Fig. 4(b), the line $\mathscr{E}=E$ again intersects the same curve, but now its part between ξ_1,ξ_2 belongs to an allowed strip. In Fig. 4(c), the line $\mathscr{E}=E$ intersects different curves and its part between ξ_1,ξ_2 belongs to a forbidden strip. In Fig. 4(d), the line $\mathscr{E}=E$ intersects different curves, but its part between ξ_1,ξ_2 belongs this time to an allowed strip.

Over the minimal interval on the plane $\mathbb{R}\times\mathbb{K}$ we can draw some branch of the isoenergy curve, and this curve has a very certain configuration. In general, properties of the configuration are in correspondence with the four indicated cases; see Fig. 5. The geometrical structure of the isoenergy curve corresponding to a forbidden strip does not depend on the case: for both cases (b) and (d) it is the same.

All the curves are symmetrical with respect to the axis ξ. The curves in Figs. 5(a) and 5(c) are real; they will be called the real branches of the isoenergy curve. The curves in Figs. 5(b) and 5(d) lie in the plane $\mathrm{Re}\,k=k_m$. They will be called the complex branches.

Let us consider the branches of the isoenergy curve over a minimal interval (ξ_1,ξ_2) up to the shift $k\to k+na^*$, $n\in\mathbb{Z}$. Then we obtain a smooth periodic curve that is, in fact, a curve on the cylinder $\mathbb{R}\times\mathbb{K}_l$. Let us denote it by γ_l if it is real and $\overline{\gamma_l}$ if it is complex.

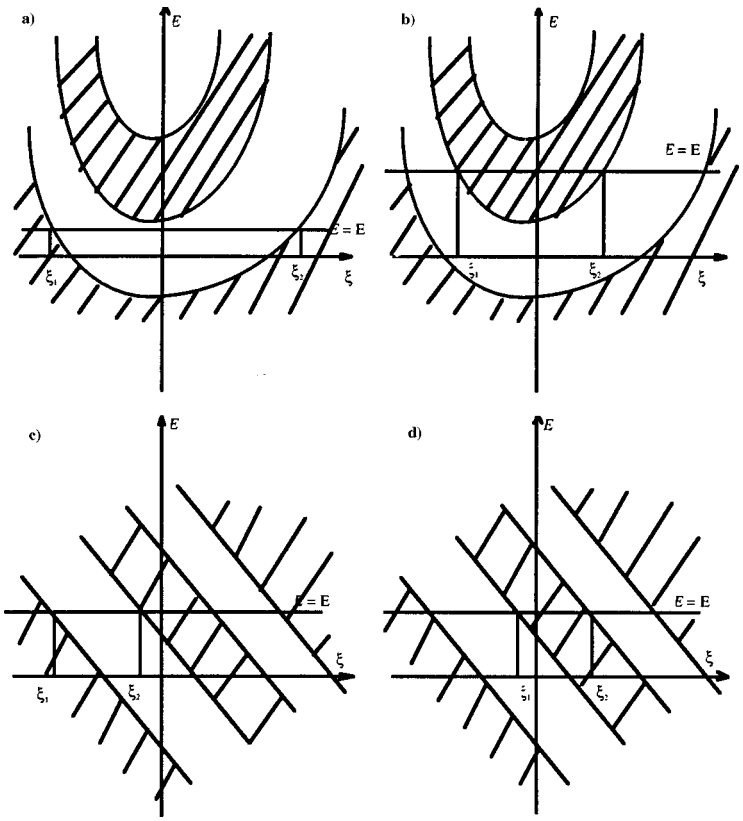

FIG. 4.

The closed curve γ_l corresponding to Fig. 5(a) can be continuously deformed to a point. The curve corresponding to Fig. 5(c) is a nontrivial cycle on the cylinder.

We make the following assumptions on the orientation: $\int_{\gamma_l}\xi d k>0$, γ_l real, $\int_{\overline{\gamma_l}}\xi(i d k)<0$, $\overline{\gamma_l}$ complex.

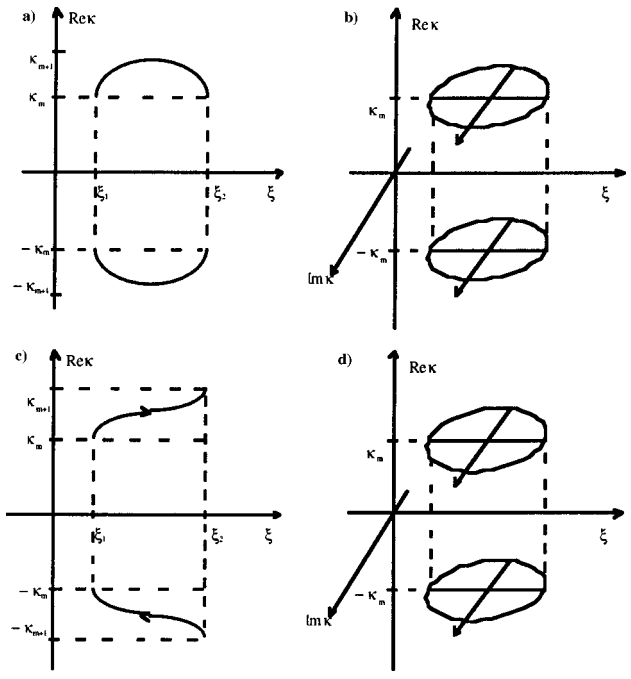

FIG. 5.

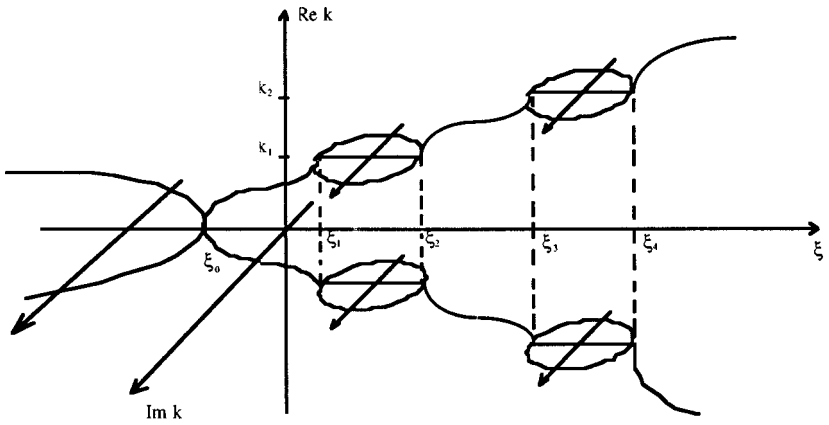

FIG. 6.

The entire axis ξ, in the general case, are divided by the turning points in a finite or an infinite number of the minimal intervals. Each of them is covered by the branch of the described type.

If p is a Q-gap potential and $q(\xi)=-\xi$, there are $2Q+1$ turning points $\xi_0<\xi_1<\cdots<\xi_{2Q}$, and $\xi_m=E_m-E$. The structure of the isoenergy curve, in the case $Q=2$, is given by Fig. 6.

III. STANDARD SOLUTIONS, TRANSITION MATRICES, IMAGINARY PARTS OF RESONANCES

A. Standard solutions and transition matrices

We begin with a short description of standard solutions of the equation

$$-\psi''+p(x)-\epsilon x\psi=E\psi, \tag{25}$$

and connections between them. The structure of such solutions and the justification of the connections between them is an object of the adiabatic approximation. A more detailed discussion of these solutions will be given in the next parts of the text.

Consider the turning points for the Q-gap potential p and for $q(\xi)=-\xi$. They are

$$\xi_m=E_m-E, \quad m=0,...,2Q. \tag{26}$$

Consider these points on the axis $\xi=\epsilon x$.

Put $\Omega_m=\Omega_m(\delta)=(\xi_{m-1}+\delta,\xi_{m+1}-\delta)$, where $\delta>0$ is fixed as a sufficiently small number such that Ω_m contains the turning point ξ_m. As ξ_{-1} and ξ_{2Q+1} we will also consider $-\infty$ and $+\infty$.

On the interval Ω_m we can construct three solutions of (25), which will be denoted by f_m, h_m, and g_m. They can be characterized by definite, relatively simple, asymptotic behavior for $\epsilon\to 0$ and $\xi\in\Omega_m$. These solutions cannot be defined uniquely, but their uncertainty is not an obstacle for our aims.

The solutions h_m, g_m are real (for $E\in\mathbb{R}$) and can be fixed such that

$$\{h_m,g_m\}=h_m g_m'-h_m' g_m=-1. \tag{27}$$

The solutions f_m and $\overline{f_m}$ are linearly independent:

$$\{f_m,\overline{f_m}\}=-i. \tag{28}$$

Two pairs of linearly independent solutions can be connected by the simple transition matrix:

$$\begin{pmatrix} h_m \\ g_m \end{pmatrix}=T\begin{pmatrix} f_m \\ \overline{f_m} \end{pmatrix}, \quad T=\frac{1}{\sqrt{2}}\begin{pmatrix} e^{-i(\pi/4)} & e^{i(\pi/4)} \\ e^{i(\pi/4)} & e^{-i(\pi/4)} \end{pmatrix}; \tag{29}$$

T is unitary, $\det T=-i$.

Our next question is the connection of the bases related with the different neighboring turning points ξ_m and ξ_{m+1}. The answer is different, depending on the nature of the interval (ξ_m,ξ_{m+1}): it can be admissible (covered by a real branch) or forbidden (covered by a complex branch).

Let the interval (ξ_m,ξ_{m+1}) be admissible and m be even. Let us put

$$\begin{pmatrix} h_{m+1} \\ g_{m+1} \end{pmatrix} = U_m \begin{pmatrix} h_m \\ g_m \end{pmatrix}. \tag{30}$$

Introduce the notation for the entries:

$$U_m = \begin{pmatrix} a_m & b_m \\ c_m & d_m \end{pmatrix}. \tag{31}$$

It is clear that U_m is real and $\det U_m=1$. Asymptotically, as $\epsilon\to 0$, U_m allows the representation

$$U_m \sim \begin{pmatrix} \cos\ \phi_m & -\sin\ \phi_m \\ \sin\ \phi_m & \cos\ \phi_m \end{pmatrix}, \tag{32}$$

where

$$\phi_m = \frac{\pi}{\epsilon a} E + \frac{1}{\epsilon} \sum_{n\geqslant 0} \epsilon^n \phi_{m,n}. \tag{33}$$

Here $\phi_{m,n}$ does not depend on E and principally can be computed, in particular,

$$\phi_{m,0}+\epsilon\phi_{m,1} = -\frac{1}{2}\int_0^{a^*} \mathscr{E}_l(k)\,dk + \frac{1}{2}\,\epsilon \int_{\gamma_l} \omega_l, \quad m=2l-2. \tag{34}$$

The formula (32) implies that for any N,

$$U_m = \begin{pmatrix} \cos\ \phi_m^N & -\sin\ \phi_m^N \\ \sin\ \phi_m^N & \cos\ \phi_m^N \end{pmatrix} + O(\epsilon^{N+1}), \tag{35}$$

where

$$\phi_m^N = \frac{\pi}{\epsilon a} E + \frac{1}{\epsilon} \sum_{n=0}^{N} \epsilon^n \phi_{m,n}. \tag{36}$$

Let now the interval (ξ_m,ξ_{m+1}) be forbidden and m be odd. Put

$$\begin{pmatrix} h_{m+1} \\ g_{m+1} \end{pmatrix} = V_m \begin{pmatrix} h_m \\ g_m \end{pmatrix}. \tag{37}$$

Again V_m is real and $\det V_m=1$. In fact, we can construct the solutions such that the diagonal elements of V_m are equal to 0:

$$V_m = \begin{pmatrix} 0 & -r_m^{-1} \\ r_m & 0 \end{pmatrix}. \tag{38}$$

The element r_m has the following asymptotic representation:

$$r_m \sim Z_m, \tag{39}$$

where

$$Z_m = 2e^{z_m}, \quad z_m = \frac{1}{\epsilon} \sum_{n \geqslant 0} z_{m,n} \epsilon^n. \tag{40}$$

$z_{m,n}$ do not depend on E, $z_{m,0} > 0$. For the main terms the explicit formula can be obtained:

$$z_{m,0} + \epsilon z_{m,1} = -\frac{1}{2} \int_{\bar{\gamma}_l} \xi(idk) + \frac{1}{2} \epsilon \int_{\bar{\gamma}_l} \omega_l, \quad m = 2l - 1. \tag{41}$$

The asymptotical sense of formula (39) is the following: for any N,

$$r_m = Z_m^N (1 + O(\epsilon^{N+1})), \quad Z_m^N = 2e^{z_m^N}, \quad z_m^N = \frac{1}{\epsilon} \sum_{n=0}^{N} \epsilon^n z_{m,n}. \tag{42}$$

B. Properties of g_-, f_+

The solutions f_0, h_0, g_0 with certain asymptotic behavior for $\xi \in (-\infty, \xi_1 - \delta)$, can be chosen, such that

$$g_- = Z_- h_0. \tag{43}$$

The coefficient $Z_- = e^{z_-}$ has the asymptotic expansion

$$z_- \sim \frac{1}{\epsilon} \sum_{n \geqslant 0} \epsilon^n z_{-,n}, \tag{44}$$

where

$$z_{-,0} + \epsilon z_{-,1} = \int_{-\infty}^{\xi_0} [ik - |E + \xi|^{1/2} \theta(-E - \xi)] d\xi - i\epsilon \int_{\xi_0}^{\infty} \omega. \tag{45}$$

The coefficients $z_{-,n}$ do not depend on E.

Similarly, the solutions f_{2Q}, h_{2Q}, g_{2Q} with certain asymptotic behavior on the interval $(\xi_{2Q-1} + \delta, \infty)$ can be chosen, such that

$$f_{2Q} = \alpha_+ f_+ + \beta_+ \bar{f}_+, \tag{46}$$

where $\alpha_+ = e^{i\varphi_+}$, φ_+ is real, allows the asymptotic expansion

$$\phi_+ \sim \frac{1}{\epsilon} \sum_{n \geqslant 0} \epsilon^n \phi_{+,n}, \tag{47}$$

and $\beta_+ = O(\epsilon^{N+1})$. For the main term one gets

$$\phi_{+,0} + \epsilon \phi_{+,1} = k_Q^* \xi_{2Q} + \int_{\xi_{2Q}}^{\infty} [k - \sqrt{(\xi + E)}\, \theta(\xi + E)] d\xi + \epsilon \int_{\xi_{2Q}}^{\infty} \omega. \tag{48}$$

There is only one term depending on E in ϕ_+; it enters in $k_Q^* \xi_{2Q} = k_Q^*(E_{2Q} - E)$.

We have described all the transition matrices and all their asymptotics for real E. The same asymptotic formulas are right for $|\mathrm{Im}\, E| < \epsilon\alpha' < \epsilon\alpha$.

C. Stark–Wannier resonances: One-gap potential

Now we have everything necessary for the description of the Stark–Wannier resonances, and, in particular, their imaginary parts. Let us begin with the one-gap potential.

Simultaneously with

$$g_- = \bar{M} f_+ + M \bar{f}_+, \tag{49}$$

consider

$$g_- = Kh_1 + Lg_1. \tag{50}$$

We will see that asymptotically the real parts of the resonances are defined by the roots of L and their imaginary parts are defined by the equation

$$L + \kappa K = 0, \tag{51}$$

with a small imaginary parameter κ.

Let us connect M with L and K. Express (h_1, g_1) in terms of $f_+, \overline{f}_+$:

$$\begin{aligned}\begin{pmatrix} h_1 \\ g_1 \end{pmatrix} &= V_1^{-1}\begin{pmatrix} h_2 \\ g_2 \end{pmatrix} = V_1^{-1}T\begin{pmatrix} f_2 \\ \overline{f}_2 \end{pmatrix} = V_1^{-1}T\begin{pmatrix} \alpha_+ & \beta_+ \\ \overline{\beta}_+ & \overline{\alpha}_+ \end{pmatrix}\begin{pmatrix} f_+ \\ \overline{f}_+ \end{pmatrix} \\ &= \frac{1}{\sqrt{2}}\begin{pmatrix} r_1^{-1}(\alpha_+ e^{i(\pi/4)} + \overline{\beta}_+ e^{-i(\pi/4)}) & r_1^{-1}(\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)}) \\ -r_1(\alpha_+ e^{-i(\pi/4)} + \overline{\beta}_+ e^{i(\pi/4)}) & -r_1(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}) \end{pmatrix}\begin{pmatrix} f_+ \\ \overline{f}_+ \end{pmatrix}.\end{aligned} \tag{52}$$

As a result,

$$M = -\frac{1}{\sqrt{2}}\, r_1(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)})[L + \kappa K], \tag{53}$$

where

$$\kappa = -r_1^{-2}\,\frac{\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)}}{\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}}. \tag{54}$$

The factor $(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)})$ does not have roots on the real axis; moreover,

$$\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)} \sim \overline{\alpha}_+ e^{i(\pi/4)}. \tag{55}$$

It is clear that it does not have roots in the strip $|\mathrm{Im}\, E| < \epsilon\alpha' < \epsilon\alpha$.

As for κ, it is clear that asymptotically,

$$\kappa = i r_1^{-2}(1 + O(\epsilon^\infty)). \tag{56}$$

In order to complete the description of the resonances, i.e., roots of M, let us find the asymptotics of L and K. From the definition

$$g_- = Z_- h_0 = Z_-(d_0 h_1 - b_0 g_1); \tag{57}$$

see (30) (31); therefore

$$K = Z_- d_0, \quad L = -Z_- b_0, \tag{58}$$

and

$$M = \frac{e^{-i\phi_+}}{\sqrt{2}}\, Z_- r_1 e^{i(\pi/4)}(b_0 - \kappa d_0). \tag{59}$$

From (32),

$$b_0 \sim -\sin\,\phi_0, \quad d_0 \sim \cos\,\phi_0. \tag{60}$$

Now we can make the final conclusions on the locations of the roots E_n of M. Up to exponentially small corrections they are defined by the roots of b_0 and, therefore, admit the asymptotic expansion as $\epsilon \to 0$,

$$\phi_0 = n\pi, \quad n \in \mathbb{Z}. \tag{61}$$

In other words, asymptotically,

$$E_n + \frac{a}{\pi} \sum_{p \geqslant 0} \epsilon^p \phi_{0,p} = \epsilon a n. \tag{62}$$

These roots of b_0 are simple and, since b_0 is real on the real axis, are real. In the strip $|\mathrm{Im}\, E| < \epsilon\alpha' < \epsilon\alpha$, other roots are absent.

The formula (62) is the so-called quantization condition. It was discussed by many authors; see, for example, Ref. 2.

The shift ΔE of the root of b_0 under the influence of the small correction to b_0 is defined by the equation

$$\left.\frac{db_0}{dE}\right|_{E_n} \Delta E \simeq \kappa d_0|_{E_n}. \tag{63}$$

From this formula, asymptotically,

$$\Delta E \simeq -\frac{\epsilon a}{\pi}\, \kappa \simeq -i\, \frac{\epsilon a}{\pi}\, (Z_1^1)^{-2}, \tag{64}$$

Z_1^1 is described by (40)–(42). So we can conclude that the real parts of the roots are defined by the equation (62), and their imaginary parts are equal to

$$\mathrm{Im}\, E \sim -\frac{\epsilon a}{4\pi}\, e^{(1/\epsilon)(\int_{\overline{\gamma}_1} \xi(idk) - \epsilon \int_{\overline{\gamma}_1} \omega_1)}. \tag{65}$$

Let us remind the reader that the integral $\int_{\overline{\gamma}_1} \xi(idk)$ is negative; it is our assumption on the orientation of the $\overline{\gamma_l}$. In fact, this integral is nothing else than the area S_l surrounded by the closed complex curve $\overline{\gamma_l}$:

$$S_l = -\int_{\overline{\gamma}_l} \xi d(ik). \tag{66}$$

It is clear, from the analytical properties of all the objects, that the small corrections do not create additional roots in the strip $|\mathrm{Im}\, E| < \epsilon\alpha' < \epsilon\alpha$.

Let us remark that the real parts of the resonances can be localized only up to the error $O(\epsilon^\infty)$. In contrast, the imaginary part allows very fine localization; it is exponentially small. The result completely coincides with (9).

D. Many-gaps potentials

Consider, at last, the Q-gap potential. Principally, some definite ϵa-periodic ladder of the resonances corresponds to every spectral interval. The real parts of such ladders are described by the same quantization conditions as in the case of a one-gap potential; see Ref. 1,

$$\sin \phi_l = 0, \tag{67}$$

or

$$\phi_l = n\pi, \quad n \in \mathbb{Z}, \tag{68}$$

or

$$E + \frac{a}{\pi} \sum_{p \geqslant 0} \epsilon^p \phi_{l,p} = \epsilon a n. \tag{69}$$

This statement is not, however, unconditional. If the distances between the real parts of the resonances corresponding to two different spectral intervals, tend to zero, the independent description of these real parts becomes impossible. We are not interested here in the description of this more complicated situation and, therefore, as in Ref. 1, we assume that some condition is fulfilled that allows us to exclude such a situation. We have to consider under this condition only a special sequence of ϵ, tending to 0. Practically, it means that we assume that for all the spectral intervals except one, let its number be l, the corresponding $\sin \phi_m^1$, $m \neq 2l-2$, are uniformly bounded from zero:

$$|\sin \phi_m^1| \geqslant \nu > 0. \tag{70}$$

Under this condition we can describe the ladder corresponding to the lth spectral interval and compute its imaginary part.

As in the case $Q=1$ introduce the new real (if $E \in \mathbb{R}$) coefficients K, L:

$$g_- = Kh_{2l-1} + Lg_{2l-1}. \tag{71}$$

Let us express M in terms of K and L. It is clear that

$$\begin{pmatrix} h_{2l-1} \\ g_{2l-1} \end{pmatrix} = A_l T \begin{pmatrix} \alpha_+ & \beta_+ \\ \overline{\beta}_+ & \overline{\alpha}_+ \end{pmatrix} \begin{pmatrix} f_+ \\ \overline{f}_+ \end{pmatrix}, \tag{72}$$

where

$$A_l = V_{2l-1}^{-1} U_{2l-2}^{-1} \cdots U_{2Q-2}^{-1} V_{2Q-1}^{-1}. \tag{73}$$

Introduce the notations for A_l entries:

$$A_l = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{74}$$

It is easy to see that

$$M = \frac{1}{\sqrt{2}} [a_{21}(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}) + a_{22}(\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)})][L + \kappa K], \tag{75}$$

$$\kappa = \frac{a_{11}(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}) + a_{12}(\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-(\pi/4)})}{a_{21}(\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}) + a_{22}(\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)})}. \tag{76}$$

Using the induction, we can prove that a single element of A, proportional to $Z_{2l-1} \cdots Z_{2Q-1}$, is a_{21}:

$$a_{21} \sim \sin \phi_{2l} \cdots \sin \phi_{2Q-2} Z_{2l-1} \cdots Z_{2Q-1} + \cdots. \tag{77}$$

Due to the condition (70) (for the E and ϵ considered) this term, indeed, defines the asymptotics of a_{21}. So we can declare that

$$\operatorname{Re} \kappa = O((Z_{2l-1}^0)^{-1} + \cdots + (Z_{2Q-1}^0)^{-1}). \tag{78}$$

As for the $\operatorname{Im} \kappa$, let us remark that because of $\det A_l = 1$,

$$\operatorname{Im} \kappa = \operatorname{Im}\left(\frac{\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}}{\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)}} \right) \left| a_{21}\left(\frac{\beta_+ e^{-i(\pi/4)} + \overline{\alpha}_+ e^{i(\pi/4)}}{\beta_+ e^{i(\pi/4)} + \overline{\alpha}_+ e^{-i(\pi/4)}} \right) + a_{22} \right|^{-2}. \tag{79}$$

Asymptotically,

$$\frac{\beta_+ e^{-i(\pi/4)}+\overline{\alpha}_+ e^{i(\pi/4)}}{\beta_+ e^{i(\pi/4)}+\overline{\alpha}_+ e^{-i(\pi/4)}}\sim i+O(\epsilon^\infty); \tag{80}$$

therefore

$$\text{Im } \kappa=\frac{1+O(\epsilon^\infty)}{(\sin \phi_{2l}^1\cdots\sin \phi_{2Q-2}^1 Z_{2l-1}^1\cdots Z_{2Q-1}^1)^2}. \tag{81}$$

Simultaneously, it is clear that $[a_{21}(\beta_+ e^{-i(\pi/4)}+\overline{\alpha}_+ e^{i(\pi/4)})+a_{22}(\beta_+ e^{i(\pi/4)}+\overline{\alpha}_+ e^{-i(\pi/4)})]$ $\sim a_{21}\overline{\alpha}_+ e^{i(\pi/4)}$ does not have roots.

Preserving the control over the roots, we can write

$$M\sim\frac{1}{\sqrt{2}}\, a_{21}\overline{\alpha}_+ e^{i(\pi/4)}(L+\kappa K), \tag{82}$$

where κ asymptotically is described by (78) and (81).

The next step is to give the asymptotic formulas for L and K. We have

$$\begin{pmatrix} h_0 \\ g_0 \end{pmatrix}=B_l\begin{pmatrix} h_{2l-1} \\ g_{2l-1} \end{pmatrix}, \tag{83}$$

where

$$B_l=U_0^{-1}V_1^{-1}\cdots U_{2l-2}^{-1}. \tag{84}$$

Let

$$B_l=\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}. \tag{85}$$

It is clear that

$$K=Z_- b_{11}, \quad L=Z_- b_{12}, \tag{86}$$

It gives

$$M\sim\frac{\alpha_+ Z_- a_{21} e^{i(\pi/4)}}{\sqrt{2}}\,(b_{12}+\kappa b_{11}). \tag{87}$$

The elements b_{21} and b_{11} contain terms proportional to $Z_1^0\cdots Z_{2l-3}^0$, asymptotically,

$$\begin{aligned} b_{11}&\sim(-1)^{l-2}Z_1\cdots Z_{2l-3}[\sin \phi_0\cdots\sin \phi_{2l-4}\cos \phi_{2l-2}+O((Z_1)^{-1}+\cdots+(Z_{2l-3})^{-1})], \\ b_{12}&\sim(-1)^{l-2}Z_1\cdots Z_{2l-3}[\sin \phi_0\cdots\sin \phi_{2l-2}+O((Z_1)^{-1}+\cdots+(Z_{2l-3})^{-1})]. \end{aligned} \tag{88}$$

As a result,

$$\begin{aligned} M\sim&\frac{1}{\sqrt{2}}\, a_{21}\overline{\alpha}_+ e^{i(\pi/4)}(-1)^{l-2}Z_1\cdots Z_{2l-3}Z_{2l-1}\cdots Z_{2Q-1} \\ &\times\sin \phi_0\cdots\sin \phi_{2l-4}\sin \phi_{2l}\cdots\sin \phi_{2Q-2}(1+O(Z_1^{-1}+\cdots+Z_{2Q-1}^{-1})) \\ &\times\Bigg[\sin \phi_{2l-2}+i\,\frac{1}{(\sin \phi_{2l}\cdots\sin \phi_{2Q-2}Z_{2l-1}\cdots Z_{2Q-1})^2}\cos \phi_{2l-2}(1+O(\epsilon^\infty)) \\ &+O(Z_1^{-1}+\cdots+Z_{2Q-1}^{-1})\Bigg]. \end{aligned} \tag{89}$$

It is essential that the second O term in [ ] is real, for $E\in\mathbb{R}$.

Repeating now the computations described in Sec. III C, we see that the real parts of the ladders are described by the quantization condition (68).

The imaginary part of this ladder is equal to

$$\operatorname{Im} E \sim -\frac{\epsilon a}{\pi} \frac{1}{(\sin \phi^1_{2l} \cdots \sin \phi^1_{2Q-2} Z^1_{2l-1} \cdots Z^1_{2Q-1})^2}, \tag{90}$$

where ϕ^1_{2m} is given in (34) and Z^1_{2m-1} in (42).

We can say that the shift of the ladder to the complex plane is induced by the tunneling of the lth spectral interval to the continuous spectrum (corresponding to the last infinite branch of the isoenergy curve) through all the intermediate gaps.

IV. FORMAL ASYMPTOTIC SOLUTIONS

Here we describe formal asymptotic solutions of the equation

$$-\psi'' + (p(x) + q(\epsilon x))\psi = E\psi. \tag{91}$$

We consider the formal solutions on the minimal intervals and more complicated solutions on the unification of two neighboring minimal intervals, including the turning point that separates them. In principal, these results do not differ from already known results; see Refs. 2 and 12; however, we describe them in a different form and it allows us to simplify the proof of properties of the transition matrices.

A. Formal solutions on the minimal interval

Fix a smooth branch γ of the isoenergy curve Γ covering the minimal interval bounded by simple turning points ξ_1, ξ_2. There exists a formal solution of the equation (91) corresponding to the branch γ, and it has the form

$$F(x, \xi, \epsilon) = e^{(i/\epsilon)\theta(\xi)} f(x, \xi, \epsilon), \quad \xi = \epsilon x, \tag{92}$$

where $\theta(\xi)$ is an integral of the differential form $k\, d\xi$ on the branch γ, $f(x, \xi, \epsilon)$ is periodic with respect to x with the period a, and the dependence on ϵ is characterized by the expansion

$$f(x, \xi, \epsilon) = \sum_{n \geqslant 0} \epsilon^n f_n(x, \xi), \quad f_0 \neq 0. \tag{93}$$

The equation (91) is equivalent to the following system of equations for f_n:

$$\left[\left(\frac{1}{i}\frac{\partial}{\partial x} + \theta'\right)^2 + p(x) + q(\xi) - E\right] f_n - i\left[2\left(\frac{1}{i}\frac{\partial}{\partial x} + \theta'\right)\frac{\partial}{\partial \xi} + \theta''\right] f_{n-1} - \frac{\partial^2}{\partial \xi^2} f_{n-2} = 0. \tag{94}$$

The equation for $n = 0$ gives

$$f_0 = N(\xi)\varphi(x, k), \quad k = \theta'(\xi). \tag{95}$$

The coefficient N can be characterized by the ''transport equation'' that arises as the condition of the existence of an a-periodic solution of the equation for f_1:

$$N(\xi) = |\mathscr{E}'(k)|^{-1/2} e^{\int \omega}, \tag{96}$$

where $\int \omega$ is an integral of the form ω on γ. As a result,

$$e^{(i/\epsilon)\theta} f_0 = c_0 |\mathscr{E}'(k)|^{-1/2} e^{(i/\epsilon)\theta(\xi)} \varphi(x, \xi) e^{\int \omega}. \tag{97}$$

This process can be continued and generates a solution F determined up to a constant factor of the form

$$c = e^{(i/\epsilon)\theta_0} \sum_{n \geqslant 0} \epsilon^n c_n . \tag{98}$$

In order to avoid misunderstandings, we should indicate that this form of the representation of f_0 is reasonable only for the case when γ is a real branch. In the internal points of this branch, $\mathscr{E}'(k) \neq 0$ and ω is purely imaginary. In the case of a complex branch the boundaries of the cuts $k = \pm 1/2na^* \pm ih_n$ are points where $(\mathscr{E}')^{-1/2}$ is finite (even zero), but not smooth. Simultaneously, in these points the solutions u_1, u_2 become orthogonal in $L_2(0,a)$, such that the form ω acquires singularities also. In order to compensate for both these singularities, it is necessary to combine the contributions of $\mathscr{E}'$ and ω in one joint expression, which is evidently smooth:

$$(\mathscr{E}')^{-1/2} e^{\int \omega} = N \exp\left[i \frac{a}{2} k + \int \tilde{\omega} \right], \tag{99}$$

where N is some constant; see (22). It is worth mentioning that ω is smooth in the ends of the cuts. Oppositely, $\tilde{\omega}$ is not smooth in these points: its singularities correspond to the singularities of $(\mathscr{E}')^{1/2.}$ Both the solutions u_1, u_2 are smooth functions of k and E in the ends of the cuts. The connection between E and k in these points is clear from Fig. 3: in particular, k is a smooth function of E. Since $\mathscr{E}(k) = E - q(\xi)$ is a smooth function of ξ, expression (99) is also smooth.

The solution F corresponding to a given branch γ of the isoenergy curve Γ can be replaced by the solution that is defined by the branch of Γ_l shifted on the period a^* with respect to k. Let us comment this declaration considering only the leading term (97). Let us assume that the solution $u(x,k)$ is a^* periodic with respect to k. Then $|\mathscr{E}'(k)|^{-1/2} e^{\int \omega}$ does not depend on the choice of a branch. The product $e^{(i/\epsilon)\theta(\xi)} \varphi(x,k)$ must be replaced by $e^{(i/\epsilon)\theta(\xi) + (i/\epsilon)a^* \xi} \varphi(x, k + a^*)$, which coincides with the initial product up to a constant factor. If $u(x,k)$ is not periodic, the transformations of $e^{\int \omega}$ and $e^{(i/\epsilon)\theta(\xi)} \varphi(x,k)$ compensate each other.

The minimal interval is covered by two branches γ_1, γ_2 of the isoenergy curve that cannot be connected by the shift $k \to k + la^*$, $l \in \mathbb{Z}$. For example, if $(k,\xi) \in \gamma_1$, then $(-k,\xi) \in \gamma_2$. Therefore, two solutions linearly independent over the field $\mathscr{A}$ of the series (98), exist on the minimal interval.

For two series F_1, F_2 of the form (92) one can consider their Wronskian:

$$\begin{aligned} \{F_1, F_2\} &= e^{(i/\epsilon)\theta_1 + (i/\epsilon)\theta_2} \left[f_1 \cdot \left(\frac{\partial}{\partial x} + \epsilon \frac{\partial}{\partial \xi} + i\theta_2' \right) f_2 - \left(\frac{\partial}{\partial x} + \epsilon \frac{\partial}{\partial \xi} + i\theta_1' \right) f_1 \cdot f_2 \right] \\ &= e^{(i/\epsilon)\theta_1 + (i/\epsilon)\theta_2} \sum_{n \geqslant 0} \epsilon^n \sum_{n_1 + n_2 = n} \left\{ f_{1,n_1} \cdot \left[\left(\frac{\partial}{\partial x} + i\theta_2' \right) f_{2,n_2} + \frac{\partial}{\partial \xi} f_{2,n_2 - 1} \right] \right. \\ &\quad \left. - \left[\left(\frac{\partial}{\partial x} + i\theta_1' \right) f_{1,n_1} + \frac{\partial}{\partial \xi} f_{1,n_1 - 1} \right] \cdot f_{2,n_2} \right\}. \end{aligned} \tag{100}$$

For any solutions this Wronskian is constant, i.e., an element of the field $\mathscr{A}$. If the solutions F_1, F_2 are linearly dependent over $\mathscr{A}$, the Wronskian is zero. If the solutions are linearly independent, the Wronskian does not equal zero.

It is easy to compute the main term of the Wronskian: if two branches γ_1, γ_2 are connected by the transformation $(k,\xi) \to (-k,\xi)$, then

$$\{f_1, f_2\} = e^{(i/\epsilon)\theta_1 + (i/\epsilon)\theta_2} |\mathscr{E}'(k)|^{-1} c_{01} c_{02} e^{\int_{\gamma_1} \omega + \int_{\gamma_2} \omega} \{u_1, u_2\}, \tag{101}$$

where $u_1 = e^{ikx} \varphi(x,k)$, $u_2 = e^{-ikx} \varphi(x,-k)$. The sum $\theta_1 + \theta_2 = \theta_0$ is constant,

$$e^{\int_{\gamma_1} \omega + \int_{\gamma_2} \omega} = N \left(\int_0^a u_1 u_2 dx \right)^{-1}, \tag{102}$$

where N is constant, and, therefore,

$$\{f_1,f_2\}\sim\frac{-2i}{a}\,e^{(i/\epsilon)\theta_0}\,\frac{\mathscr{E}'}{|\mathscr{E}'|}\,c_{01}c_{02}N; \tag{103}$$

here c_{01}, c_{02} are the corresponding constants in the formulas for the main terms of f_1,f_2.

If the branch γ (i.e., γ_1,γ_2) is real, the solutions F_1,F_2 can be normalized, such that

$$\overline{F_2}=F_1, \tag{104}$$

and $\{F_1,F_2\}$ becomes purely imaginary.

If the branch γ is complex, the solutions can be normalized such that they become real. Their Wronskian also become real.

Our last remark concerns the asymptotic character of the solutions (92).

If ξ_1,ξ_2 are simple turning points, the solution F possesses the asymptotic character on each subinterval $\delta\in(\xi_1,\xi_2)$ satisfying the condition $\mathrm{dist}(\xi_j,\delta)\geqslant\epsilon^{(2/3)(1-\gamma)}$, $\gamma>0$. It implies, more precisely, that

$$\epsilon^n|f_n(x,\xi)|\leqslant\epsilon^{\gamma n-1/6}. \tag{105}$$

On δ there exists a true solution $\mathscr{F}$ of the equation (91) for which F is its asymptotic expansion as $\epsilon\to0$:

$$\mathscr{F}=e^{(i/\epsilon)\theta}\left(\sum_{n=0}^{N}\epsilon^n f_n+o(\epsilon^{\gamma N-1/6})\right). \tag{106}$$

B. Formal solutions on the interval containing a turning point

In a vicinity of a simple turning point ξ_1, different formal solutions suitable for the asymptotic representation of the true solutions can be constructed.

The simplest one has the form

$$g=(\epsilon^{-1/6}wa+\epsilon^{1/6}w'b)e^{ik_mx}, \tag{107}$$

where $w=w(\epsilon^{-2/3}\zeta_0)$ is a solution of the Airy equation,

$$w''(z)=zw(z), \tag{108}$$

$$\zeta_0=(2M_mq'(\xi_1))^{1/3}(\xi-\xi_1); \tag{109}$$

M_m corresponds to E_m [see (16)] in the definition of the turning point, $E_m+q(\xi_1)=E$, a,b are characterized by the expansions

$$a=\sum_{n\geqslant0}\epsilon^{n/3}a_n(x,\zeta_0),\quad b=\sum_{n\geqslant0}\epsilon^{n/3}b_n(x,\zeta_0), \tag{110}$$

a_n,b_n are periodic or antiperiodic with respect to x (simultaneously with e^{ik_mx}) and are polynomials with respect to the second argument. The coefficients a,b do not depend on the choice of the solution w. In particular,

$$a_0=c_0\varphi(x,k_m). \tag{111}$$

If w is fixed, the solution g is defined up to the constant factor of the form

$$c=\sum_{n\geqslant0}\epsilon^{n/3}c_n. \tag{112}$$

The solution g has the asymptotic character on the interval $|\zeta_0|\leqslant\epsilon^{(1/2)(1+\delta)}$, $\delta>0$, where

$$\epsilon^{n/3}|a_n,b_n|\leqslant C_n\epsilon^{\delta n}. \tag{113}$$

On the same interval a true solution $\tilde{g}$ can be constructed for which g is its asymptotic expansion.

The solutions g were considered in Ref. 12; they were used there in order to describe the transformation of the solutions F in the point ξ_1. We will describe here also a different solution G that allows us to simplify essentially the description of this transformation. It is analogous to Olver's solution well known in the case of the standard WKB approach.[13]

But first let us describe shortly an analog of one more formal solution known in the standard WKB method. In Ref. 2 it was remarked that the general Maslov construction[14] can be applied also to Eq. (91), and to essentially more general partial differential equations. In particular, in vicinities of the simple turning point, the asymptotic behavior of the solutions of (91) can be described by the integrals of the form

$$\frac{e^{ik_m x}}{(2\pi\epsilon)^{1/2}} \int dp e^{(i/\epsilon)(\xi p - S(p))} \sum_{n\geqslant 0} \epsilon^n \tilde{f}_n(x,p), \quad \xi = \epsilon x, \tag{114}$$

where S is an integral of the form ξdp along the continuous branch of the isoenergy curve $\mathscr{E}(k_m+p)+q(\xi)=E$ containing the point (k_m,ξ_1); $\tilde{f}_n$ are periodic (antiperiodic) functions of x.

The asymptotic representations (107) and (114) have their own advantages and disadvantages. In particular, the representation (114) can be used on a finite (in ξ) vicinity of ξ_1. From the other side, representation (107) is more suitable for the description of the growing and decreasing solutions as ξ goes away from ξ_1 in the direction of the interval covered by the complex branch of the isoenergy curve. The Olver's-type solution possesses both these advantages.

C. Uniform asymptotics in a vicinity of the turning point

This solution has the form

$$G = (\epsilon^{-1/6} wA + \epsilon^{1/6} w' B) e^{ik_m x}, \tag{115}$$

where

$$w = w(\epsilon^{-2/3}\zeta),$$

$$A = A(x,\xi,\epsilon) = \sum_{n\geqslant 0} \epsilon^n A_n(x,\xi), \tag{116}$$

$$B = B(x,\xi,\epsilon) = \sum_{n\geqslant 0} \epsilon^n B_n(x,\xi). \tag{117}$$

In this formula w is again a solution of the Airy equation, A_n, B_n are smooth functions, periodic (antiperiodic) with respect to x, and the variable ζ is considered as a function of $\xi=\epsilon x$ defined by the equation

$$\mathscr{E}(k_m - i\sqrt{\zeta}\zeta') + q(\xi) = E. \tag{118}$$

As $\xi \to \xi_1, \zeta \sim \zeta_0$.

In order to find A_n, B_n, substitute the representation (115) to the equation (91):

$$\left[-\frac{d^2}{dx^2} + p + q - E \right] (\epsilon^{-1/6} wA + \epsilon^{1/6} w' B) e^{ik_m x} = e^{ik_m x} (\epsilon^{-1/6} wR_1 + \epsilon^{1/6} w' R_2), \tag{119}$$

$$R_1 = \sum_{n\geqslant 0} \epsilon^n R_{1,n}, \quad R_2 = \sum_{n\geqslant 0} \epsilon^n R_{2,n}; \tag{120}$$

here

$$R_{1,n}=\left[\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)^2+p+q-E-\zeta\zeta'^2\right]A_n-2i\zeta\zeta'\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)B_n-2i\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)\frac{\partial}{\partial\xi}A_{n-1}$$
$$-2\zeta\zeta'\frac{\partial}{\partial\xi}B_{n-1}-(\zeta\zeta')'B_{n-1}-\frac{\partial^2}{\partial\xi^2}A_{n-2}, \tag{121}$$

$$R_{2,n}=\left[\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)^2+p+q-E-\zeta\zeta'^2\right]B_n-2i\zeta'\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)A_n-2i\left(\frac{1}{i}\frac{\partial}{\partial x}+k_m\right)\frac{\partial}{\partial\xi}B_{n-1}$$
$$-2\zeta'\frac{\partial}{\partial\xi}A_{n-1}-\zeta''A_{n-1}-\frac{\partial^2}{\partial\xi^2}B_{n-2}. \tag{122}$$

This leads to the system for the definition of A_n, B_n:

$$R_{1,n}=0, \quad R_{2,n}=0. \tag{123}$$

In principal, the system can be studied quite analogously to system (94). The solution does not depend on the choice of w.

However, it turns out that the consideration of system (123) can be reduced almost directly to system (94).

In order to describe this reduction, consider the asymptotics of the solutions of the Airy equation as $z\to\infty$. Introduce some standard solutions of the Airy equation. There exists a unique solution w_1 with the following characteristic behavior as $z\to-\infty$:

$$w_1(z)\sim(-z)^{-1/4}e^{i(2/3)(-z)^{3/2}}\sum_{n\geqslant 0} i^n\alpha_n(-z)^{-3n/2}, \quad \alpha_0=1; \tag{124}$$

the coefficients α_n are real and can be computed explicitly. As a result,

$$w_1'(z)\sim -i(-z)^{1/4}e^{i(2/3)(-z)^{3/2}}\sum_{n\geqslant 0} i^n\beta_n(-z)^{-3n/2}, \quad \beta_0=1; \tag{125}$$

β_n can be computed in terms of α_n, α_{n-1}. The function $\overline{w_1}$ is also a solution.

The real solution,

$$v=\frac{1}{\sqrt{2}}\,(e^{-i(\pi/4)}w_1+e^{i(\pi/4)}\overline{w_1}), \tag{126}$$

is up to a constant factor a unique solution decreasing as $z\to+\infty$. It has the following asymptotic behavior:

$$v\sim\frac{1}{\sqrt{2}}\,z^{-1/4}e^{-2/3z^{3/2}}\sum_{n\geqslant 0}\alpha_n z^{-3n/2}, \tag{127}$$

$$v'\sim-\frac{1}{\sqrt{2}}\,z^{1/4}e^{-2/3z^{3/2}}\sum_{n\geqslant 0}\beta_n z^{-3n/2}. \tag{128}$$

Another real solution,

$$u=\frac{1}{\sqrt{2}}\,(e^{i(\pi/4)}w_1+e^{-i(\pi/4)}\overline{w_1}), \tag{129}$$

grows at infinity:

$$u\sim\sqrt{2}z^{-1/4}e^{2/3z^{3/2}}\sum_{n\geqslant 0}(-1)^n\alpha_n z^{-3n/2}, \tag{130}$$

$$u' \sim \sqrt{2} z^{1/4} e^{2/3 z^{3/2}} \sum_{n \geqslant 0} (-1)^n \beta_n z^{-3n/2}. \tag{131}$$

It is easy to see that

$$\{v, u\} = 2, \quad \{w_1, \overline{w_1}\} = 2i, \tag{132}$$

and

$$\begin{pmatrix} v \\ u \end{pmatrix} = T \begin{pmatrix} w_1 \\ \overline{w_1} \end{pmatrix}, \quad T = \frac{1}{\sqrt{2}} \begin{pmatrix} e^{-i(\pi/4)} & e^{i(\pi/4)} \\ e^{i(\pi/4)} & e^{-i(\pi/4)} \end{pmatrix}; \tag{133}$$

T is a unitary matrix, $\det T = -i$.

Therefore the following declarations are natural. The functions F^1, F^2,

$$\begin{aligned} F^1 &= e^{ik_m x + (i/\epsilon)(2/3)(-\zeta)^{3/2}} f^1, \quad f^1 = (-\zeta)^{-1/4} \alpha A - i(-\zeta)^{1/4} \beta B, \\ F^2 &= e^{ik_m x - (i/\epsilon)(2/3)(-\zeta)^{3/2}} f^2, \quad f^2 = (-\zeta)^{-1/4} \overline{\alpha} A + i(-\zeta)^{1/4} \overline{\beta} B, \end{aligned} \tag{134}$$

$$\alpha = \sum_{n \geqslant 0} (i\epsilon)^n \alpha_n (-\zeta)^{-3n/2}, \quad \beta = \sum_{n \geqslant 0} (i\epsilon)^n \beta_n (-\zeta)^{-3n/2},$$

satisfy system (94). The system for F^1 corresponds to the branch γ_1 of the isoenergy curve, where

$$k = k_m - (-\zeta)^{1/2} \zeta'; \tag{135}$$

F^2 corresponds to the branch γ_2, where

$$k = k_m + (-\zeta)^{1/2} \zeta'. \tag{136}$$

Since $\zeta(\xi_1) = 0$, these branches are fixed uniquely. They contain the point (ξ_1, k_m).

These branches are real. Consider now the complex branches containing the point (ξ_1, k_m).

Introduce two new series F^1, F^2,

$$F^1 = e^{ik_m x - (1/\epsilon)(2/3)\zeta^{3/2}} \frac{1}{\sqrt{2}} f^1, \quad f^1 = \zeta^{-1/4} \alpha_1 A - \zeta^{1/4} \beta_\perp B, \tag{137}$$

$$\alpha_1 = \sum_{n \geqslant 0} \epsilon^n \alpha_n \zeta^{-3n/2}, \quad \beta_1 = \sum_{n \geqslant 0} \epsilon^n \beta_n \zeta^{-3n/2},$$

$$\begin{aligned} F^2 &= e^{ik_m x + (1/\epsilon)(2/3)\zeta^{3/2}} \sqrt{2} f^2, \\ f^2 &= \zeta^{-1/4} \alpha_2 A + \zeta^{1/4} \beta_2 B, \end{aligned} \tag{138}$$

$$\alpha_2 = \sum_{n \geqslant 0} (-1)^n \epsilon^n \alpha_n \zeta^{-3n/2}, \quad \beta_2 = \sum_{n \geqslant 0} (-1)^n \epsilon^n \beta_n \zeta^{-3n/2}.$$

These solutions satisfy system (94), where correspondingly,

$$k = k_m + i\zeta^{1/2} \zeta', \quad k = k_m - i\zeta^{1/2} \zeta'. \tag{139}$$

Let us denote these branches again by γ_1, γ_2.

Let us discuss further the connection between the solutions F and G more concretely. Compute the coefficients of the series (134),

$$f^1=\sum_{n\geqslant 0}\epsilon^n f^1_n,\quad f^2=\sum_{n\geqslant 0}\epsilon^n f^2_n, \tag{140}$$

in terms of A_n, B_n:

$$f^1_n=\sum_{n_1+n_2=n} i^{n_1}(-\zeta)^{-3n_1/2}[(-\zeta)^{-1/4}\alpha_{n_1}A_{n_2}-i(-\zeta)^{1/4}\beta_{n_1}B_{n_2}], \tag{141}$$

$$f^2_n=\sum_{n_1+n_2=n} (-i)^{n_1}(-\zeta)^{-3n_1/2}[(-\zeta)^{-1/4}\alpha_{n_1}A_{n_2}+i(-\zeta)^{1/4}\beta_{n_1}B_{n_2}]. \tag{142}$$

Let us remark that we assume in that computation that solutions correspond to the real branches.

It can be checked that system (123) is equivalent to the systems (94) for f^1_n, f^2_n, respecting to the suitable branches of the isoenergy curve.

However, comparing systems (94) and (123) one has to keep in mind an essential detail principally affecting the set of the solutions of (123). In order to explain the detail, consider f^1_n, f^2_n for $n=0$:

$$f^1_0=[(-\zeta)^{-1/4}A_0-i(-\zeta)^{1/4}B_0], \tag{143}$$

$$f^2_0=[(-\zeta)^{-1/4}A_0+i(-\zeta)^{1/4}B_0]. \tag{144}$$

As a result,

$$A_0=\tfrac{1}{2}(-\zeta)^{1/4}(f^1_0+f^2_0), \tag{145}$$

$$B_0=\frac{1}{2i}(-\zeta)^{-1/4}(f^2_0-f^1_0). \tag{146}$$

With arbitrary solutions f^1_0, f^2_0, the functions A_0, B_0 satisfy the system (123) for $n=0$ everywhere except the point $\xi=\xi_1$, where f^1_0 and f^2_0 have singularities.

The solutions $f^{1,2}_0$ are fixed up to (independent) constant factors. It is essential that under the suitable choice of these constant factors, A_0 and B_0 become smooth for $\xi=\xi_1$ and are defined up to a common constant factor. The concrete structure of the singularities was briefly considered in Ref. 12, and reflected by formula (97). In order to estimate character of these singularities, remark that $\varphi(x,k)$ is a smooth function of k on the plane $\mathbb{K}$ with the cuts. Let us fix the integrals $\int_{\gamma_1}\omega$, $\int_{\gamma_2}\omega$ putting them to be zero in the point $(\xi_\perp, k_m)$. Then

$$e^{\int_{\gamma_1}\omega+\int_{\gamma_2}\omega}=\left(\int_0^a u_1(x,k_m)u_2(x,k_m)dx\right)\left(\int_0^a u_1(x,k)u_2(x,k)dx\right)^{-1}, \tag{147}$$

and

$$A_0=\tfrac{1}{2}(-\zeta)^{1/4}|\mathscr{E}'(k)|^{-1/2}[c_{01}\varphi(x,k)e^{\int_{\gamma_1}\omega}+c_{02}\varphi(x,2k_m-k)e^{\int_{\gamma_2}\omega}], \tag{148}$$

$$B_0=\frac{1}{2i}(-\zeta)^{-1/4}|\mathscr{E}'(k)|^{-1/2}[c_{02}\varphi(x,2k_m-k)e^{\int_{\gamma_2}\omega}-c_{01}\varphi(x,k)e^{\int_{\gamma_1}\omega}], \tag{149}$$

where $k=\theta_1'$. The first [ ] becomes an even function of $k-k_m$ and the second one an odd function if we put $c_{01}=c_{02}=c_0$. Both the functions A_0, B_0 become smooth even functions of $k-k_m$ and smooth functions of ξ. The coefficients are defined up to the common constant factor c_0.

We will not discuss further elementary details concerning A_n, B_n, $n\geqslant 1$. Finally, we can declare that formal solution (115) exists and is defined up to a factor of form (98). It exists on the unification of two minimal spectral intervals $\Delta=(\xi_0,\xi_1]\cup[\xi_1,\xi_2)$ containing the point ξ_1.

With any solution of system (123) we get the two-dimensional space of the solutions of (91) over the field $\mathscr{A}$ since the solution w of the Airy equation is arbitrary. The coefficients $\hat{A}=e^{ik_m x}A$ and $\hat{B}=e^{ik_m x}B$ can be found as real functions and the reality of G will be determined by the reality of w. The linear independence of G and $\overline{G}$ corresponding to w and $\overline{w}$ is also determined by the linear independence of $w,\overline{w}$.

In the conclusion of this section, consider the Wronskian of two solutions G_1,G_2:

$$\{G_1,G_2\}=G_1\frac{d}{dx}G_2-G_2\frac{d}{dx}G_1. \tag{150}$$

It is easy to show that

$$\{G_1,G_2\}=\{w_1,w_2\}[\hat{A},\hat{B}], \tag{151}$$

$$[\hat{A},\hat{B}]=\{\hat{A},\hat{B}\}+\zeta'(\hat{A}^2-\zeta\hat{B}^2), \tag{152}$$

where $\{w_1,w_2\}$ is the Wronskian of the corresponding w_1,w_2, with respect to their own argument, and $\{\hat{A},\hat{B}\}$ is the full standard Wronskian of $\hat{A},\hat{B}$ with respect to x. With an arbitrary choice of G_1,G_2 the combination $[\hat{A},\hat{B}]$ is an element of the field $\mathscr{A}$. $\{G_1,G_2\}$ equals zero if G_1 and G_2 are linearly dependent; it does not equal zero if G_1 and G_2 are independent. The expression $[\hat{A},\hat{B}]$ does not depend on the choice of w_1,w_2. Let us fix it such that

$$[\hat{A},\hat{B}]=-\tfrac{1}{2}. \tag{153}$$

This condition fixes $\hat{A},\hat{B}$ up to a common sign.

Compute the main term of the left side of (153):

$$[\hat{A},\hat{B}]\sim\frac{1}{2i}(\mathscr{E}')^{-1}c_0^2e^{\int\gamma_1\omega+\int\gamma_2\omega}\{u(x,k),u(x,2k_m-k)\}=-c_0^2\frac{1}{a}\int_0^a|u(x,k)|^2dx. \tag{154}$$

If we put

$$\int_0^a|u(x,k)|^2dx=a, \tag{155}$$

then $c_0^2=1/2$.

V. THE EXACT SOLUTIONS AND THE TRANSITION MATRICES

In this last section we construct the solutions that were used in Sec. III. We describe the asymptotic behavior of the solutions, consider the set of natural bases, and prove the asymptotic properties of the transition matrices.

A. Preliminary remarks

We remind the reader briefly of a standard idea that is used in order to prove the existence of the solution of an ordinary differential equation with a prescribed asymptotic expansion with respect to an external parameter.

Let u_1,u_2 be two formal asymptotic solutions of the equation

$$-\psi''+v(x,\epsilon)\psi=0, \tag{156}$$

where v depends on the external parameter $\epsilon,\epsilon\to 0$. Consider the finite sums of the asymptotic expansions and denote them u_1^N,u_2^N. Compute their Wronskian $\{u_1^N,u_2^N\}$. Since the formal Wronskian $\{u_1,u_2\}$ is a (formal) constant, the derivative $(d/dx)\{u_1^N,u_2^N\}$ should be small and, consequently,

$$\{u_1^N,u_2^N\}(x)=\{u_1^N,u_2^N\}(\tilde{x})+\int_{\tilde{x}}^{x}\frac{d}{dx}\{u_1^N,u_2^N\}dx \tag{157}$$

is close to a nonzero constant.

Introduce two almost new solutions:

$$y_1=\{u_1^N,u_2^N\}^{-1/2}u_1^N,\quad y_2=\{u_1^N,u_2^N\}^{-1/2}u_2^N. \tag{158}$$

Since $\{y_1,y_2\}=1$, both the functions y_1,y_2 have to satisfy the same equation,

$$-y''+\tilde{v}(x,\epsilon)y=0, \tag{159}$$

with a potential $\tilde{v}(x,\epsilon)$ close to $v(x,\epsilon)$.

Consider the integral equation

$$\psi(x)=y(x)+\int_{\tilde{x}}^{x}K(x,t)\psi(t)dt, \tag{160}$$

where $\tilde{x}$ is some fixed point, $y=\alpha y_1+\beta y_2$, α,β are some fixed constants, and

$$K(x,t)=(y_2(x)y_1(t)-y_1(x)y_2(t))(v(t,\epsilon)-\tilde{v}(t,\varepsilon)). \tag{161}$$

The solution of the integral equation satisfies the differential equation (156); its initial data in $\tilde{x}$ coincide with the initial data of y.

The kernel K is small in a suitable sense; therefore the integral equation should be effective for the asymptotic treatment of u_1,u_2.

If the asymptotic solutions possess different orders for different x, it is necessary to choose $\tilde{x}$ and α,β appropriately.

The last general remark: although the justification of the asymptotic expansion is limited by the choice of N (arbitrarily large), for the solution ψ of the integral equation the full asymptotic expansion, as $\epsilon\to 0$, can be obtained.

B. Solutions ψ_1 and ψ_2

Choose a turning point and construct real $\hat{A},\hat{B}$. Fix them such that

$$[\hat{A},\hat{B}]=-\tfrac{1}{2}. \tag{162}$$

Assume also that $\hat{A},\hat{B}$ do not contain exponential factors with respect to ϵ^{-1}. Consider four formal solutions $\mathbf{g},\mathbf{h},\mathbf{f},\overline{\mathbf{f}}$ taking as w, correspondingly, $u,v,w_1,\overline{w_1}$. It is evident that

$$\begin{pmatrix}\mathbf{h}\\ \mathbf{g}\end{pmatrix}=T\begin{pmatrix}\mathbf{f}\\ \overline{\mathbf{f}}\end{pmatrix}. \tag{163}$$

As it was shown in Sec. IV, as $\epsilon^{-2/3}\zeta\to+\infty$, the solutions $\mathbf{g},\overline{\mathbf{h}}$ can be reduced to solutions of the form F; analogously, as $\epsilon^{-2/3}\zeta\to-\infty$, the solutions $\mathbf{f},\overline{\mathbf{f}}$ can also be reduced to solutions of the form F.

Consider two neighboring turning points ξ_1,ξ_2, $\xi_1<\xi_2$. Let (ξ_1,ξ_2) be covered by a complex branch of the isoenergy curve. The minimal intervals (ξ_0,ξ_1) and (ξ_2,ξ_3) are covered by real branches. Consider in a vicinity of ξ_1 the solution $\mathbf{g}$ and denote it by $\mathbf{g}_1$. Up to a factor belonging to the field $\mathscr{A}$, it coincides (after the asymptotic reduction of w) with the solution $\mathbf{h}_2$ defined in a vicinity of ξ_2:

$$\mathbf{g}_1\sim -Z\mathbf{h}_2. \tag{164}$$

Since $\mathbf{g}_1,\mathbf{h}_2$ are fixed up to the sign, we can choose them such that $Z>0$ in the main order. Let us consider a smooth unit partition,

$$1=\theta_1+\theta_2, \tag{165}$$

where $\theta_1(\xi),\theta_2(\xi)$ differ from 0 or 1 only on a proper subinterval of (ξ_1,ξ_2). Introduce

$$\mathbf{r}=\mathbf{g}_1\theta_1-Z\mathbf{h}_2\theta_2. \tag{166}$$

Analogously, consider the solutions $\mathbf{g}_2$ and $\mathbf{h}_1$:

$$\mathbf{h}_1\sim Z_1\mathbf{g}_2, \tag{167}$$

and

$$\mathbf{s}=\mathbf{g}_2\theta_2+\frac{1}{Z_1}\,\mathbf{h}_1\theta_1. \tag{168}$$

It is easy to see that

$$\{\mathbf{r},\mathbf{s}\}=\{\mathbf{g}_1,\mathbf{h}_1\}\,\frac{1}{Z_1}=-Z\{\mathbf{h}_2,\mathbf{g}_2\}; \tag{169}$$

therefore

$$Z_1=\frac{1}{Z}, \tag{170}$$

and

$$\{\mathbf{r},\mathbf{s}\}=Z. \tag{171}$$

The factor Z will be discussed later, however, it is worth indicating here the exponential character of the dependence of Z on ϵ,

$$\ln Z=\frac{1}{\epsilon}\int_{\xi_1}^{\xi_2}|\mathrm{Im}\ k|d\xi+O(1). \tag{172}$$

Denote

$$Z_0=e^{(1/\epsilon)\int_{\xi_1}^{\xi_2}|\mathrm{Im}\ k|d\xi}. \tag{173}$$

Introduce new notations:

$$\rho_j^{\pm}(\xi)=\begin{cases} e^{\pm(2/3)\epsilon^{-1}|\zeta_j|^{3/2}}, & \zeta_j\geqslant 0,\\ 1, & \zeta_j\leqslant 0,\end{cases} \tag{174}$$

where ζ_j are the functions of the type ζ related to the points ξ_j.

Let us introduce the weights:

$$\rho_r(\xi)=\begin{cases}\rho_1^+(\xi), & \xi\leqslant\xi_2\\ Z_0\rho_2^-(\xi), & \xi\geqslant\xi_1,\end{cases}\quad \rho_s(\xi)=\begin{cases}\rho_2^+(\xi), & \xi\geqslant\xi_1,\\ Z_0\rho_1^-(\xi), & \xi\leqslant\xi_2.\end{cases} \tag{175}$$

They both are monotonous functions naturally defined for $\xi\in(\xi_0,\xi_3)$. It is clear that

$$\rho_r(\xi)\rho_s(\xi)=Z_0. \tag{176}$$

Return to the constructions of the previous section and take $\mathbf{r}$ and $\mathbf{s}$ as u_1,u_2. Then

$$|u_1^N|=|r^N|\leqslant C_N\rho_r,\quad |u_2^N|=|s^N|\leqslant C_N\rho_s, \tag{177}$$

$$\{u_1^N, u_2^N\} = \{r^N, s^N\} = Z_N + Z_0 O(\epsilon^{N+1}), \tag{178}$$

where Z_N is the cut series for Z.

It is easy to see that

$$|y_1| \leqslant C_N Z_0^{-1/2} \rho_r, \quad |y_2| \leqslant C_N Z_0^{-1/2} \rho_s,$$

$$|K| \leqslant C_N \epsilon^{N+1} Z_0^{-1} (\rho_r(\xi)\rho_s(\eta) + \rho_r(\eta)\rho_s(\xi)),$$

$$\xi = \epsilon x, \quad \eta = \epsilon t. \tag{179}$$

Consider now two exact solutions ψ_1 and ψ_2 taking as y, respectively, y_1 and y_2 and as $\widetilde{\xi}$ the points $\xi_0 + \delta, \xi_3 - \delta, \delta > 0, \delta$ is fixed.

It is almost immediately clear that on the interval $\xi \in [\xi_0 + \delta, \xi_3 - \delta]$,

$$\psi_1 = Z_N^{-1/2} r^N + Z_0^{-1/2} \rho_r(\xi) O(\epsilon^{N+1}), \tag{180}$$

$$\psi_2 = Z_N^{-1/2} s^N + Z_0^{-1/2} \rho_s(\xi) O(\epsilon^{N+1}), \tag{181}$$

$$\{\psi_1, \psi_2\} = 1 + O(\epsilon^{N+1}). \tag{182}$$

C. Solutions f, g, h

Consider the interval $(\xi_0 + \delta, \xi_3 - \delta)$. We have constructed on the interval two real solutions with a known asymptotic behavior and with a known Wronskian. As a result on finite vicinities of the turning points we have two linearly independent real solutions with a known asymptotic behavior. Introduce now on these vicinities sets of canonical solutions.

Consider on a vicinity of ξ_1 two solutions:

$$\widetilde{h}_1 = Z_N^{-1/2} \psi_2, \quad \widetilde{g}_1 = Z_N^{1/2} \psi_1. \tag{183}$$

On this vicinity, see the definition of **r** and **s**:

$$\begin{aligned} \widetilde{h}_1 &= (Z\mathbf{h}_1)^N Z_N^{-1} + \rho_1^- O(\epsilon^{N+1}) = h_1^N + \rho_1^- O(\epsilon^{N+1}), \\ \widetilde{g}_1 &= g_1^N + \rho_1^+ O(\epsilon^{N+1}), \end{aligned} \tag{184}$$

and

$$\{\widetilde{h}_1, \widetilde{g}_1\} = -\{\psi_1, \psi_2\} = -1 + O(\epsilon^{N+1}). \tag{185}$$

Let us renormalize $\widetilde{h}_1, \widetilde{g}_1$:

$$h_1 = \{\widetilde{g}_1, \widetilde{h}_1\}^{-1/2} \widetilde{h}_1, \quad g_1 = \{\widetilde{g}_1, \widetilde{h}_1\}^{-1/2} \widetilde{g}_1; \tag{186}$$

then

$$h_1 = h_1^N + \rho_1^- O(\epsilon^{N+1}), \quad g_1 = g_1^N + \rho_1^+ O(\epsilon^{N+1}), \quad \{h_1, g_1\} = -1. \tag{187}$$

Introduce f_1:

$$\begin{pmatrix} f_1 \\ \overline{f_1} \end{pmatrix} = T^{-1} \begin{pmatrix} h_1 \\ g_1 \end{pmatrix}. \tag{188}$$

It is clear that

$$f_1 = f_1^N + \rho_1^+ O(\epsilon^{N+1}), \quad \{f_1, \overline{f_1}\} = -i. \tag{189}$$

Analogously, on a vicinity of ξ_2 introduce

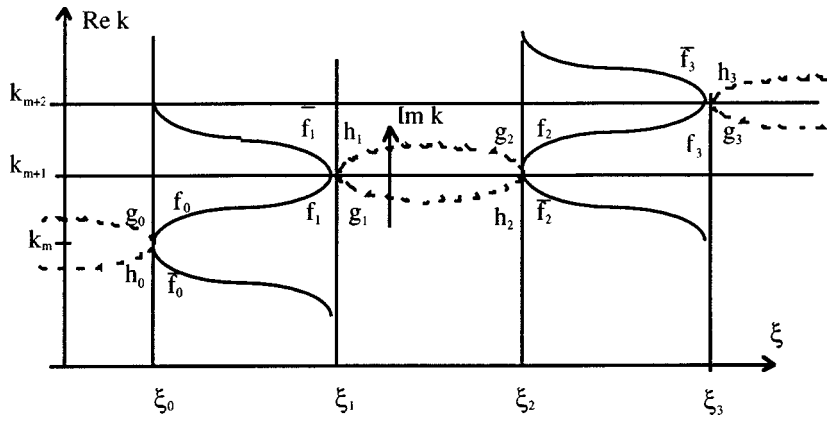

FIG. 7.

$$\widetilde{h}_2 = -Z_N^{-1/2}\psi_1, \quad \widetilde{g}_2 = Z_N^{1/2}\psi_2; \tag{190}$$

then $\{\widetilde{h}_2, \widetilde{g}_2\} = \{\psi_2, \psi_1\} = -1 + O(\epsilon^\infty)$. Consider

$$h_2 = \{\widetilde{g}_2, \widetilde{h}_2\}^{-1/2}\widetilde{h}_2, \quad g_2 = \{\widetilde{g}_2, \widetilde{h}_2\}^{-1/2}\widetilde{g}_2. \tag{191}$$

Then

$$h_2 = h_2^N + \rho_2^- O(\epsilon^{N+1}), \quad g_2 = g_2^N + \rho_2^+ O(\epsilon^{N+1}), \quad \{h_2, g_2\} = -1. \tag{192}$$

For f_2, defined by the formula

$$\begin{pmatrix} f_2 \\ \overline{f}_2 \end{pmatrix} = T^{-1}\begin{pmatrix} h_2 \\ g_2 \end{pmatrix}, \tag{193}$$

one has

$$f_2 = f_2^N + \rho_2^+ O(\epsilon^{N+1}), \quad \{f_2, \overline{f}_2\} = -i. \tag{194}$$

So one has to conclude that on a vicinity of each turning point ξ_m there exist three solutions f_m, h_m, g_m with all the set of the properties declared in Sec. III:

$$\begin{pmatrix} h_m \\ g_m \end{pmatrix} = T\begin{pmatrix} f_m \\ \overline{f}_m \end{pmatrix}, \quad f_m = f_m^N + \rho_m^+ O(\epsilon^{N+1}), \tag{195}$$

$$h_m = h_m^N + \rho_m^- O(\epsilon^{N+1}), \quad g_m = g_m^N + \rho_m^+ O(\epsilon^{N+1}),$$

$$\{f_m, \overline{f}_m\} = -i, \quad \{h_m, g_m\} = -1. \tag{196}$$

We have constructed here solutions with the finite asymptotic expansions, although N can be taken as an arbitrary natural number. These finite expansions are sufficient for the spectral applications considered in Sec. III. From the other side, it is not too difficult to complete the previous proof and to obtain solutions with the full asymptotic expansions. (See Fig. 7.)

D. Transition matrices

Now it is easy to connect two bases (h_1, g_1) and (h_2, g_2) if the minimal interval (ξ_1, ξ_2) is covered by a complex branch of the isoenergy curve. Indeed,

$$\begin{pmatrix} h_2 \\ g_2 \end{pmatrix} = \frac{1}{\{\psi_1, \psi_2\}^{1/2}}\begin{pmatrix} -Z_N^{-1/2}\psi_1 \\ Z_N^{1/2}\psi_2 \end{pmatrix} = \begin{pmatrix} -Z_N^{-1}g_1 \\ Z_N h_1 \end{pmatrix} = \begin{pmatrix} 0 & -Z_N^{-1} \\ Z_N & 0 \end{pmatrix}\begin{pmatrix} h_1 \\ g_1 \end{pmatrix}. \tag{197}$$

Let now (ξ_1,ξ_2) be separated by a minimal interval covered by a real branch of the isoenergy curve. Consider the relation

$$\begin{pmatrix} h_2 \\ g_2 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} h_1 \\ g_1 \end{pmatrix}. \tag{198}$$

It is equivalent to the formula

$$\begin{pmatrix} f_2 \\ \overline{f_2} \end{pmatrix} = T^{-1} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} T \begin{pmatrix} f_1 \\ \overline{f_1} \end{pmatrix}. \tag{199}$$

The entries of

$$\begin{pmatrix} \alpha_1 & \beta_1 \\ \overline{\beta_1} & \overline{\alpha_1} \end{pmatrix} = T^{-1} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} T \tag{200}$$

should satisfy the relation

$$|\alpha_1|^2 - |\beta_1|^2 = 1. \tag{201}$$

They can be expressed in terms of Wronskians,

$$\alpha_1 = +i\{f_2, \overline{f_1}\}, \quad \beta_1 = -i\{f_2, f_1\}. \tag{202}$$

It is clear from the constructions for f_1, f_2 that

$$\mathbf{f}_2 = e^{-i\varphi}\mathbf{f}_1, \tag{203}$$

where φ is real,

$$\varphi \sim \frac{1}{\epsilon}\,\varphi^0 + \cdots. \tag{204}$$

We will discuss the structure of φ in the next section. From (202),

$$\alpha_1 = e^{-i\varphi^N}(1 + O(\epsilon^{N+1})), \quad \beta_1 = O(\epsilon^{N+1}). \tag{205}$$

Finally, one has

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} \cos\,\phi^N & -\sin\,\phi^N \\ \sin\,\phi^N & \cos\,\phi^N \end{pmatrix} + O(\epsilon^{N+1}). \tag{206}$$

E. Main asymptotic terms of φ and Z

Let us compute the main terms of the asymptotic expansions of φ and Z. Let us begin with φ. Remember the definition,

$$\mathbf{f}_2 = e^{-i\varphi}\mathbf{f}_1. \tag{207}$$

In a vicinity of the point ξ_1,

$$\mathbf{f}_1 \simeq \epsilon^{-1/6} w_1 \hat{A} + \epsilon^{1/6} w_1' \hat{B}, \tag{208}$$

where $\hat{A}$ and $\hat{B}$ are real. After the asymptotical reconstruction,

$$\mathbf{f}_1 \sim [(-\zeta_1)^{-1/4}\hat{A} - i(-\zeta_1)^{1/4}\hat{B}]e^{(i/\epsilon)(2/3)(-\zeta)^{3/2}} = c_{01}e^{ik_m x}e^{(i/\epsilon)\int_{\xi_1}^{\xi}(k-k_m)d\xi}(\mathscr{E}')^{-1/2}\varphi(x,k)e^{\int_{\xi_1}^{\xi}\omega}, \tag{209}$$

where c_{01} is real and fixed by the normalizing condition $c_{01}^2=1/2$.

Analogously,

$$\mathbf{f}_2\sim c_{02}e^{ik_{m'}x}e^{(i/\epsilon)\int_{\xi_2}^{\xi}(k-k_{m'})d\xi}(\mathscr{E}')^{-1/2}\varphi(x,k)e^{\int_{\xi_2}^{\xi}\omega}, \tag{210}$$

and again c_{02} is real and normalized by the condition $c_{02}^2=1/2$. We have to take $k_{m'}=k_m+1/2a^*$.

The connection between the signs of c_{01},c_{02} is not essential for the spectral applications; it is enough to compute the matrix V_m up to the sign. Nevertheless, we can put $c_{01}=c_{02}$. Then

$$e^{-i\varphi}\sim e^{(i/\epsilon)k_{m'}\xi_2-(i/\epsilon)k_m\xi_1-(i/\epsilon)\int_{\xi_1}^{\xi_2}kd\xi-\int_{\xi_1}^{\xi_2}\omega}=e^{-(i/\epsilon)(\pi/a)E+(i/\epsilon)\int_{k_m}^{k_{m'}}\mathscr{E}(k)dk-\int_{\xi_1}^{\xi_2}\omega}. \tag{211}$$

It coincides with formula (33) declared in Sec. III A.

Let us consider now the case when (ξ_1,ξ_2) is covered by a complex branch of the isoenergy curve. Then, after the reconstruction,

$$\mathbf{g}_1\sim\sqrt{2}[\zeta_1^{-1/4}\hat{A}_0+\zeta_1^{1/4}\hat{B}_0]e^{(2/3)(1/\epsilon)\zeta_1^{3/2}}=\pm c_{01}\sqrt{2}e^{ik_mx}e^{(1/\epsilon)\int_{\xi_1}^{\xi}|\mathrm{Im}\,k|d\xi}(-i\mathscr{E}')^{-1/2}\varphi(x,k)e^{\int_{\xi_1}^{\xi}\omega}. \tag{212}$$

On the branch of the integration $\mathrm{Im}\,k<0$, $(\pm)$ appeared since we do not control the sign of $(-i\mathscr{E}')^{1/2}$.

Analogously,

$$\mathbf{h}_2\sim\pm c_{02}\frac{1}{\sqrt{2}}e^{ik_mx}e^{(1/\epsilon)\int_{\xi_2}^{\xi}|\mathrm{Im}\,k|d\xi}(i\mathscr{E}')^{-1/2}\varphi(x,k)e^{\int_{\xi_2}^{\xi}\omega}. \tag{213}$$

Formulas (212) and (213) are valid for different ξ: in (212) $\xi<\hat{\xi}$, in (213) $\xi>\hat{\xi}$, and $\hat{\xi}$ corresponds to the end of the cut where $\mathscr{E}'=\infty$. Both the expressions cannot be continued directly through the point $\hat{\xi}$ since the integrals $\int\omega$ are singular at this point. However, the full expressions have finite limits, as $\xi\to\hat{\xi}$. It is easy to check that

$$\frac{\mathscr{E}'(\hat{\xi}-\epsilon)}{\mathscr{E}'(\hat{\xi}+\epsilon)}\to-1,\quad \epsilon\to0. \tag{214}$$

It implies that the full expression can be defined also for intervals containing $\hat{\xi}$. For this, the integral $\int\omega$ over the interval containing the point $\hat{\xi}$ must be defined in the sense of the principal value.

Let us come to Z. The constants here are the same as in the previous computations of φ. Therefore $c_{01}^2=c_{02}^2=\frac{1}{2}$. Since the solutions $\mathbf{g}_1,\mathbf{h}_2$ should be real and

$$g_1\sim-Zh_2, Z>0, \tag{215}$$

we have to choose c_{01} and c_{02} such that

$$Z\sim2e^{(1/\epsilon)\int_{\xi_1}^{\xi_2}|\mathrm{Im}\,k|d\xi+\int_{\xi_1}^{\xi_2}\omega}=2e^{-(1/2\epsilon)\int_{\overline{\gamma}_l}\xi(idk)+1/2\int_{\overline{\gamma}_l}\omega_l}\gg1. \tag{216}$$

F. Solutions g_-,f_+

The solutions g_- and f_+ were introduced in Sec. I. Their asymptotic behavior and their connections with the solutions f_i,g_i,h_i defined in the vicinity of the turning points ξ_0 and ξ_{2N} were described in Sec. III separately. Let us justify the results listed there.

Consider first g_-. Introduce the standard formal solutions $\mathbf{f}_0,\mathbf{g}_0,\mathbf{h}_0$ defined on the interval $(-\infty,\xi_1)$ containing the turning point ξ_0.

It is well known that

$$\mathscr{E}(k)=k^2+\cdots,\quad \psi(x,k)=e^{ikx}+\cdots,\quad k\to\infty. \tag{217}$$

Fix the complex branch $\overline{\gamma}$ of the isoenergy curve for $\xi<\xi_0$, where $k=-i|k|$. Consider the corresponding formal solution $\mathbf{g}$; its main term,

$$\mathbf{g}\sim c\sqrt{2}(i\mathscr{E}'(k))^{-1/2}\varphi(x,k)e^{\int_{\xi_0}^{\xi}\omega}e^{(i/\epsilon)\int_{\xi_0}^{\xi}kd\xi},\quad \text{Im } k<0, \tag{218}$$

can be simplified as $\xi\to-\infty$:

$$\mathbf{g}\sim\pm c|\xi+E|^{-1/4}e^{-(1/\epsilon)\int_{\xi}^{\xi_0}|\xi+E|^{1/2}d\xi}. \tag{219}$$

The sign can be fixed arbitrary. So the asymptotic behavior of the main term $\mathbf{g}$ coincides up to a constant factor with the asymptotic behavior of g_-.

Fix now this formal solution such that it would possess the asymptotic character simultaneously for $\epsilon\to0$ and $\xi\to-\infty$. More precisely, we can fix it such that its coefficients g_n will decrease at infinity as $|\xi+E|^{-1/4-n}$. To construct this solution one has to choose the arbitrary limits of integrations in all the solutions of the transport equations to be equal to $-\infty$. This solution is defined up to a simple constant c. Take this constant such that the main term of the solution, as $\xi\to-\infty$, coincides with the main term of the asymptotics of g_-.

It is clear that there exists an element Z_- of the field $\mathscr{A}$, such that (after the asymptotic reduction of $\mathbf{h}_0$)

$$\mathbf{g}_-\sim-Z_-\mathbf{h}_0; \tag{220}$$

compare with (164). The main term of Z_- is exponentially large for small ϵ.

Now we can construct $\mathbf{r}_-$, the analog of the solution $\mathbf{r}$; see (166). As the analog of $\mathbf{s}$ we can simply take $\mathbf{g}_0$.

Almost literally repeating the ideas of Sec. V A, we can find on the interval $(-\infty,\xi_1-\delta)$, containing one turning point ξ_0, the solutions f_0,h_0,g_0 with the asymptotics $f_0\sim\mathbf{f}_0$, $h_0\sim\mathbf{h}_0$, $g_0\sim\mathbf{g}_0$. As almost solutions we can use properly normalized r_-^N and g_0^N.

In the integral equation for the solution ψ_-, corresponding to $\mathbf{r}_-$, see Sec. V B, one has to take $\widetilde{\xi}=-\infty$. The coincidence of the main terms of the asymptotic behavior of r_-^N and g_-, as $\xi\to-\infty$, shows that ψ_-, and g_- should coincide up to a constant factor and, therefore, h_0 and g_- also coincide up to a constant factor. Asymptotically, as $\epsilon\to0$, this constant factor can be described by a series from the field $\mathscr{A}$. The explicit formula for the main term of this constant can be found by constructions similar to the constructions of Sec. V E. The result is given by (44). For the spectral applications one does not need any formula for the constant.

The solutions f_{2Q},g_{2Q},h_{2Q} with the standard properties can be constructed simultaneously with the solutions $f_{2Q-1},g_{2Q-1},h_{2Q-1}$, as usual. In particular, $f_{2Q}\sim\mathbf{f}_{2Q}$. They are defined on the interval $(\xi_{2Q-2}+\delta,R)$; R is fixed such that $R>\xi_{2Q}$.

The solution f_+ introduced in Sec. I, is characterized by its behavior, as $\xi\to+\infty$; see (5). Since $\{f_+,\overline{f}_+\}=-i$ there exist constants α_+,β_+ such that

$$f_{2Q}=\alpha_+f_++\beta_+\overline{f}_+. \tag{221}$$

We have to describe the asymptotic behavior of α_+,β_+, as $\epsilon\to0$.

In order to obtain this asymptotic behavior, introduce the formal solution $\mathbf{f}_+$, which has an asymptotic character simultaneously, as $\epsilon\to0$ and $\xi\to\infty$. The idea is absolutely similar to the idea used for the definition of $\mathbf{g}_-$, although now the point $\xi=-\infty$ should be replaced in all the integral's limits by the point $\xi=+\infty$. This formal solution is defined up to a simple constant factor, which should be fixed, such that $\mathbf{f}_+$ coincides asymptotically, as $\xi\to\infty$, with f_+.

We have two formal solutions $\mathbf{f}_+$ and $\mathbf{f}_+$. Now we can use f_+^N and $\overline{f}_+^N$ as almost solutions on the interval $\xi\geqslant\xi_{2Q}+\delta$. Notice that the Wronskian of these almost solutions equals $-i$ as well as for f_+ and f_+. It is clear from their asymptotic behavior at infinity. In the corresponding integral equations for the true solutions we have to set $\widetilde{\xi}=\infty$. If we take as y in the integral equation the almost solution f_+^N, we obtain a solution ψ_+ which has two properties: (1) by the construction $\psi_+\sim\mathbf{f}_+$; and (2) $\psi_+=f_+$, since they have the same asymptotics as $\xi\to\infty$. Therefore, as $\epsilon\to0$,

$$f_+\sim\mathbf{f}_+. \tag{222}$$

Two formal solutions $\mathbf{f}_{2Q}$ and $\mathbf{f}_+$ correspond to the same branch of the isoenergy curve and their Wronskians with the corresponding complex conjugated solutions coincide. Therefore

$$\mathbf{f}_{2Q}=e^{-i\varphi_+}\mathbf{f}_+, \tag{223}$$

where $e^{-i\varphi_+}\in\mathscr{A}$ (in a natural sense after reconstruction). Since

$$\alpha_+=i\{f_{2Q},\overline{f_+}\},\quad \beta_+=-i\{f_{2Q},f_+\}, \tag{224}$$

we obtain

$$\alpha_+\sim e^{-i\varphi_+}, \tag{225}$$

or

$$\alpha_+=e^{-i\varphi_+^N}+O(\epsilon^{N+1}); \tag{226}$$

N is arbitrary. For β_+ we have

$$\beta_+=O(\epsilon^N). \tag{227}$$

G. An important remark

All the asymptotical considerations of the two last sections were limited by the case of real E. In fact, all the obtained asymptotic formulas remain to be right if $|\mathrm{Im}\,E|<\epsilon\alpha'<\epsilon\alpha$. If we consider the equation for such complex E, we discover that all the constructions can be repeated without any changes. It can be explained by the fact that for such small $\mathrm{Im}\,E$, all the exponential factors in the asymptotic formulas have exactly the same character of increasing and decreasing as for the real E. The most convenient variant of the treatment of the complex E is to introduce the complex turning points defined again by the equation $E_m-\xi=E$ and to consider the equation on the straight line connecting them, in other words, on the line $\mathrm{Im}\,x=(1/\epsilon)\mathrm{Im}\,E$. Some special, but very natural, details concerning the interval $(\xi_{2Q},+\infty)$, can be found in Ref. 1.

ACKNOWLEDGMENTS

This work was performed during the stay of the first author in University Paris 13. This author would like to thank the university for its hospitality. The work was supported in part by INTAS-RFBR 96-0414.

[1] V. S. Buslaev and L. A. Dmitrieva, ''A Bloch electron in an external field,'' Alg. Anal. **23**, 1–29 (1989); translated in Leningrad Math. J. **1**, 287–320 (1990).
[2] V. S. Buslaev, ''Quasiclassical approximation for equations with periodic coefficients,'' Usp. Mat. Nauk **42**, 77–98 (1987).
[3] J. Avron, ''The lifetime of Wannier ladder state,'' Ann. Phys. **143**, 33–53 (1982).
[4] J. Agler and R. Froese, ''Existence of Stark ladders resonances,'' Commun. Math. Phys. **100**, 161–171 (1985).
[5] F. Bentosela and V. Grecchi, ''Stark–Wannier ladders,'' Commun. Math. Phys. **142**, 169–192 (1991).
[6] J. M. Combes and P. D. Hislop, ''Stark ladders resonances for small electric fields,'' Commun. Math. Phys. **140**, 291–320 (1991).
[7] V. Grecchi, M. Majoli, and A. Sacchetti, ''Stark resonances in disordered systems,'' Commun. Math. Phys. **146**, 231–241 (1992).
[8] G. Nenciu, ''Adiabatic theorem and spectral concentration. I. Arbitrary order spectral concentration for the Stark effect in atomic physics,'' Commun. Math. Phys. **82**, 121–135 (1981).
[9] J. Asch and P. Briet, ''Lower bounds on the width of Stark–Wannier type resonances,'' Commun. Math. Phys. **179**, 725–736 (1996).
[10] V. Grecchi and A. Sacchetti, ''Lifetime of the Wannier–Stark resonances and perturbation theory,'' to appear in Commun. Math. Phys.
[11] V. S. Buslaev, ''On spectral properties of adiabatically perturbed Schrödinger operators with periodic potentials,'' Séminaire EDP de l'Ecole Polytechnique, exposé No. XXIII, 1990–1991.
[12] V. S. Buslaev, ''Adiabatic perturbation of a periodic potential,'' Teor. Mat. Fiz. **58**, 223–243 (1984).
[13] M. V. Fedoryuk, *Asymptotic Analysis* (Springer-Verlag, Berlin, 1993).

[14] M. V. Fedoryuk and V. P. Maslov, *Quasi-Classical Approximation for the Equations of Quantum Mechanics* (Reidel, Dordrecht, 1981).
[15] N. E. Firsova, ''The Riemann surface of quasimomentum and scattering theory for the perturbed Hill operator,'' Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **51**, 183–196 (1975).
[16] V. A. Marchenko and I. V. Ostrovskii, ''Characteristics of the spectrum of the Hill operator,'' Mat. Sb. **97**, 540–606 (1975).

An exactly soluble Schrödinger equation with smooth position-dependent mass

Liès Dekar[a)]
Centre Universitaire de Médéa, DZ-26000 Médéa, Algeria

Lyazid Chetouani
Département de Physique Théorique, Institut de Physique, Université de Constantine, Constantine, Algeria

Théophile F. Hammann
Laboratoire de Mathématique et Physique Mathématique, Faculté des Sciences et Techniques, Université de Haute Alsace, 4, rue des Frères Lumière, F-68093 Mulhouse Cédex, France

(Received 2 June 1997; accepted for publication 11 February 1998)

The one-dimensional generalized Schrödinger equation for a system with smooth potential and *mass* step is resolved exactly. The wave function depends on the *Heun's* function, which is a solution of a second-order *Fuchsian* equation with four singularities. The behavior of the transmission coefficient as a function of energy is compared to that of the case of an abrupt potential and mass step. Two limiting cases are also studied: when the width of the mass step is vanishing, and when the smooth potential and mass step tend to an abrupt potential and mass step. © *1998 American Institute of Physics.* [S0022-2488(98)02505-5]

I. INTRODUCTION

Particular interest has been given to physical systems with position-dependent mass[1–5] in recent years. One example is the effective-mass model[6] (related to the envelope-function approximation),[7,8] used to study the dynamics of electrons in semiconductor heterostructures.[9–18] Since the mass and the momentum operator no longer commute, the generalization of the standard Hamiltonian[9–16,19–21] is not trivial. The basic problems are the following: the choice of a correct ordering of the operators for the kinetic energy so that the Hamiltonian is hermitic,[9,13,21] the Galilean invariance[16,22] related to the Bargmann's theorem, and the connection rules[7,8,10,11,17,18,23] for the wave functions across an abrupt heterojunction. It appears that a consensus was made in the majority of these works about the following form of the Hamiltonian:

$$H=\frac{1}{2}\left[p\,\frac{1}{m}\,p\right]+V. \tag{1.1}$$

With the help of this Hamiltonian we have calculated[24] the Green's function via the path integral formalism for systems with potential and mass both, step, rectangular barrier, and multi-barrier. We have also deduced the reflection and transmission coefficients for the first two cases, reproducing thus the results obtained by direct resolution of the Schrödinger equation.[25] But, these applications were made for systems with piecewise flat potentials and masses. However, it is also interesting to study models with continuous and smoothly variable potentials and masses. Indeed, they are obviously much more difficult to solve, but they allow us to study the crossing of the interfaces in a smooth and not in an abrupt way. Moreover, the cases with abrupt interfaces will be deduced as being limiting cases of these latter.

In this paper, we solve the one-dimensional generalized Schrödinger equation associated with the above-mentioned form of Hamiltonian (1.1) for a system with continuous potential and mass, both of the smooth step shape. In Sec. II we will show that, thanks to a change of the independent

[a)]Present address: Laboratoire de Mathématique et Physique Mathématique, Faculté des Sciences et Techniques, Université de Haute Alsace, 4, rue des Frères Lumière, F-68093 Mulhouse Cédex, France.

 © 1998 American Institute of Physics

variable (position), followed by a change of the dependent variable (wave function), the generalized Schrödinger equation is transformed into a second-order Fuchsian equation with four singularities, the so-called Heun's equation.[26–31] With the help of the solutions of this equation—the so-called Heun's functions—we will determine the general expression of the wave function. We will deduce the reflection and transmission coefficients by studying the asymptotic behavior of the wave function when $x \to \pm\infty$. We will show that the behavior of the transmission coefficient as a function of energy is similar to that of the case of an abrupt potential and mass step.[25] We will also recover the reflection and transmission coefficients of two limiting cases, namely, when the width of the mass step is vanishing,[32,33] and when the smooth potential and mass step tend to an abrupt potential and mass step.[25] Section III will feature a brief summing up and a conclusion. In Appendix A, we will build the functional relation between the z and $1\text{-}z$ argument Heun's functions, the relation required for the study of the asymptotic behavior of the wave function when $x \to \pm\infty$. Finally, in Appendix B, we will construct a relation between two *contiguous* Heun's functions, the relation necessary for the study of the transmission coefficient as a function of the energy, and for the study of the limiting case when the smooth potential and mass tend to an abrupt potential and mass step.

II. SCHRÖDINGER EQUATION

Let us adopt the following parametrization for the smooth potential[32,33] and mass step,

$$V(x) = \frac{V_0}{2}\left(1 + th\,\frac{x}{2r}\right), \tag{2.1}$$

$$m(x) = \overline{m} + \frac{\Delta m}{2}\,th\,\frac{x}{2r}, \tag{2.2}$$

where $\overline{m} = (m_1 + m_2)/2$, $\Delta m = m_2 - m_1$. Here V_0 and r are positive constants.

The potential (respectively, the mass) increases from the value $V = 0 (m = m_1)$ for $x = -\infty$ to the value $V = V_0 (m = m_2)$ for $x = +\infty$. The main rise occurring in the $-2r < x < 2r$ range,

$$V(-2r) = 0.119 V_0, \quad m(-2r) = 0.119 \Delta m;$$

$$V(2r) = 0.881 V_0, \quad m(2r) = 0.881 \Delta m.$$

The time-independent Schrödinger equation that follows from the generalized Hamiltonian (1.1) has the form

$$-\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m(x)}\frac{d\Psi(x)}{dx}\right) + V(x)\Psi(x) = E\Psi(x), \tag{2.3}$$

or

$$\frac{d^2\Psi(x)}{dx^2} - \frac{1}{m(x)}\left(\frac{dm(x)}{dx}\right)\frac{d\Psi(x)}{dx} + \frac{2m(x)}{\hbar^2}(E - V(x))\Psi(x) = 0. \tag{2.4}$$

The change of the independent variable,

$$z = \frac{1}{2}\left(1 - th\,\frac{x}{2r}\right), \tag{2.5}$$

maps the interval $x \in]-\infty, +\infty[$ to $z \in]0,1[$, and transforms the coefficients of the differential equation (2.4) into rational functions of the new variable z. Potential (2.1) and mass (2.2) can now be written as $V(z) = -V_0(z-1)$, $m(z) = -\Delta m(z-a)$, where

$$a = \frac{m_2}{m_2 - m_1}. \tag{2.6}$$

We come to the differential equation

$$\frac{d^2\Psi(z)}{dz^2}+\left[\frac{1}{z}+\frac{1}{z-1}-\frac{1}{z-a}\right]\frac{d\Psi(z)}{dz} + \left[-\omega^2 z-\frac{\nu^2 a}{z}+\frac{\mu^2(a-1)}{z-1}+a\omega^2+\nu^2-\mu^2\right]\frac{\Psi(z)}{z(z-1)(z-a)}=0, \tag{2.7}$$

where the following abbreviations have been used:

$$\nu^2=\frac{2m_2 r^2}{\hbar^2}(V_0-E), \quad \mu^2=-\frac{2m_1 r^2 E}{\hbar^2}, \quad \omega^2=\frac{2\Delta m r^2}{\hbar^2}V_0. \tag{2.8}$$

The resulting equation (2.7) is a Heun's-type equation, i.e., a second-order linear homogeneous differential equation with four singularities $z=0$, 1, a, ∞, all regular. Since all the singularities are regular, Eq. (2.7) belongs to the class of Fuchsian equations. In order to transform the equation (2.7) into its normal form, we introduce the following change of the dependent variable:

$$\Psi(z)=z^{\nu}(z-1)^{\mu}f(z). \tag{2.9}$$

A simple calculation leads to the normal form of the Heun's-type equation,

$$\frac{d^2 f(z)}{dz^2}+\left[\frac{2\nu+1}{z}+\frac{2\mu+1}{z-1}-\frac{1}{z-a}\right]\frac{df(z)}{dz}+[\{(\nu+\mu)^2-\omega^2\}z + a\{\omega^2-(\nu+\mu)(\nu+\mu+1)\}+\nu]\frac{f(z)}{z(z-1)(z-a)}=0. \tag{2.10}$$

The solution of Eq. (2.10), which is regular in the neighborhood of $z=0$ and belongs to the exponent *zero*, is the Heun's function defined by the series[34]

$$f_1(z)=CF(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu+\omega,\nu+\mu-\omega,2\nu+1,-1;z) \equiv C\left\{1-\frac{a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu}{a(2\nu+1)}z+\sum_{s=2}^{\infty}c_s z^s\right\}. \tag{2.11}$$

Constant C will be determined by the boundary conditions. The c_s coefficients of the series (2.11) are determined by the difference equation

$$(s+2)(s+2+2\nu)ac_{s+2}=\{(s+1)^2(a+1)+(s+1)[2\nu-1+a(2\nu+2\mu+1)] +a[(\nu+\mu)(\nu+\mu+1)-\omega^2]-\nu\}c_{s+1} -(s+\nu+\mu+\omega)(s+\nu+\mu-\omega)c_s, \tag{2.12}$$

with the initial conditions

$$c_0=1, \quad c_1=\frac{a[(\nu+\mu)(\nu+\mu+1)-\omega^2]-\nu}{a(2\nu+1)}, \quad c_s=0, \quad \text{if} \quad s<0. \tag{2.13}$$

Series (2.11) converges inside a circle centered at the origin and whose radius is the distance from the origin to the nearer of the two singular points, $z=1$ and $z=a$. Thus, series (2.11) converges for $|z|<1$ if $|a|>1$, i.e., if $m_1<2m_2$ and $m_1\neq m_2$. In the case where $|a|<1$, i.e., $m_1>2m_2$, series (2.11) converges for $|z|<|a|$. In the latter case we can make an analytic continuation to ensure the convergence for $|z|<1$. The analytic continuation may be obtained by introducing into the differential equation (2.10) the change of the independent variable,

$$z'=1-z, \tag{2.14}$$

which leads to the following solution:

$$f_3(z)=C'F(1-a,(1-a)[\omega^2-(\nu+\mu)(\nu+\mu+1)] + \mu;\nu+\mu+\omega,\nu+\mu-\omega,2\mu+1,-1;1-z). \tag{2.15}$$

We first consider the limit $x\to\infty$ or $z\approx e^{-x/r}\to 0$; solution (2.11) becomes $f_1(0)=C$, or

$$\Psi(z)\to Cz^{\nu}\approx Ce^{-\nu x/r}. \tag{2.16}$$

Now two cases have to be distinguished.

(a) $E<V_0$, $\nu>0$ is real. Expression (2.16) vanishes exponentially, as would be expected for the $E<V_0$ case. We obtain $\Psi\to Ce^{-Kx}$, where

$$K^2=\frac{\nu^2}{r^2}=\frac{2m_2}{\hbar^2}(V_0-E). \tag{2.17}$$

(b) $E>V_0$, $\nu=-ik_2r$ is imaginary. Then $\Psi\to Ce^{ik_2x}$, where

$$k_2^2=\frac{2m_2}{\hbar^2}(E-V_0). \tag{2.18}$$

On the other hand, for $x\to-\infty$ or $z\to 1$, $1-z\approx e^{+x/r}\to 0$, we apply the formula (A15) of Appendix A, which links the z and $1-z$ argument Heun's functions. Thus, the Heun's function of relation (2.11) becomes

$$\begin{aligned}&F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu+\omega,\nu+\mu-\omega,2\nu+1,-1;z)\\ &\quad=C_1F(1-a,(1-a)[\omega^2-(\nu+\mu)(\nu+\mu+1)]-\mu;\nu+\mu-\omega,\nu+\mu+\omega,2\mu+1;1-z)\\ &\qquad+C_2(1-z)^{-2\mu}F(1-a,(1-a)[\omega^2-(\nu-\mu)(\nu-\mu+1)]\\ &\qquad-\mu;\nu-\mu+\omega,\nu-\mu-\omega,-2\mu+1;1-z),\end{aligned} \tag{2.19}$$

where constants C_1 and C_2 are

$$C_1=F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu-\omega,\nu+\mu+\omega,2\nu+1,-1;1), \tag{2.20}$$

$$C_2=F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\nu;\nu-\mu+\omega,\nu-\mu-\omega,2\nu+1,-1;1). \tag{2.21}$$

Since $1-z\approx e^{+x/r}$ when $x\to-\infty(1-z\to 0)$, the wave function behaves like

$$\Psi\to Ce^{i\pi\mu}\{C_1e^{\mu(x/r)}+C_2e^{-\mu(x/r)}\}. \tag{2.22}$$

Setting $\mu=ik_1r$, where

$$k_1^2=\frac{2m_1}{\hbar^2}E, \tag{2.23}$$

and

$$C=\frac{e^{i\pi\mu}}{C_1},\quad R=\frac{C_2}{C_1}, \tag{2.24}$$

wave function (2.11) can be written as

$$\Psi(x)=\left[\frac{1}{2}\left(1-th\,\frac{x}{2r}\right)\right]^{\nu}\left[\frac{1}{2}\left(1+th\,\frac{x}{2r}\right)\right]^{\mu} \times\frac{F\left(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu+\omega,\nu+\mu-\omega,2\nu+1,-1;\frac{1}{2}\left(1-th\,\frac{x}{2r}\right)\right)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu+\omega,\nu+\mu-\omega,2\nu+1,-1;1)}. \tag{2.25}$$

We can summarize the limits of the wave function when $x\to\pm\infty$ by

$$\Psi(x)=\begin{cases} e^{ik_1x}+\mathrm{Re}^{-ik_1x}, & \text{for} \quad x\to+\infty, \\ \begin{cases} Ce^{-Kx}, & \text{if } E>V_0, \quad \text{for} \quad x\to+\infty, \\ Ce^{ik_2x}, & \text{if } E<V_0, \quad \text{for} \quad x\to+\infty, \end{cases} \end{cases} \tag{2.26}$$

Thus, we recover the asymptotic behavior of a plane wave coming from the left-hand side. The quantity $|R|^2$ is the reflection coefficient,

$$|R|^2=\left|\frac{F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\nu;\nu-\mu+\omega,\nu-\mu-\omega,2\nu+1,-1;1)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu-\omega,\nu+\mu+\omega,2\nu+1,-1;1)}\right|^2. \tag{2.27}$$

Again, we distinguish the two cases $E<V_0$ and $E>V_0$:

(a) $E<V_0$, $\mu=ik_1r$ is imaginary, $\nu>0$ is real, and ω is imaginary if $m_2<m_1$ or real if $m_2>m_1$. The numerator and denominator of the reflection coefficient (2.27) are complex conjugates, we have then $|R|^2=1$. We also recover this case of total reflection when $E=V_0$, $\nu=0$, ω imaginary or real.

(b) $E>V_0$, $\mu=ik_1r$ and $\nu=-ik_2r$ are imaginary.

Thanks to relation (B22) of Appendix B between two contiguous Heun's function, we can investigate the behavior of the transmission (or reflection) coefficient as a function of the energy E. For this purpose, let us rewrite expression (2.27) of the reflection coefficient in an equivalent form:

$$|R|^2=\left|\frac{F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\nu;\nu-\mu+\omega,\nu-\mu-\omega,2\nu+1,-1;1)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu-\omega,\nu+\mu+\omega,2\nu+1,-1;1)}\right| \times\left|\frac{F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\nu;\nu-\mu-\omega,\nu-\mu+\omega,2\nu+1,-1;1)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]+\nu;\nu+\mu-\omega,\nu+\mu+\omega,2\nu+1,-1;1)}\right|, \tag{2.28}$$

where we have used the fact that Heun's function is unaltered by the interchange of α and β arguments. Knowing that

$$\omega^2=\frac{\nu^2}{a}-\frac{\mu^2}{(a-1)}, \tag{2.29}$$

we apply now relation (B22) of Appendix B for each of the four Heun's functions of (2.28),

$$|R|^2=\frac{[(a-1)\nu+a\mu]^2}{[(a-1)\nu-a\mu]^2} \times\left|\frac{F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\mu-\omega,\nu-\mu+\omega,\nu-\mu-\omega+1,2\nu+2,0;1)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]-\mu-\omega,\nu+\mu+\omega,\nu+\mu-\omega+1,2\nu+2,0;1)}\right| \times\left|\frac{F(a,a[\omega^2-(\nu-\mu)(\nu-\mu+1)]+\mu+\omega,\nu-\mu-\omega,\nu-\mu+\omega+1,2\nu+2,0;1)}{F(a,a[\omega^2-(\nu+\mu)(\nu+\mu+1)]-\mu+\omega,\nu+\mu-\omega,\nu+\mu+\omega+1,2\nu+2,0;1)}\right|. \tag{2.30}$$

(b1) The asymptotic value of the transmission coefficient when the energy goes to infinity. When $E\to\infty$ the reflection coefficient (2.30) is equivalent to

$$|R|^2 \underset{E\to\infty}{\approx} \left(\frac{\sqrt{m_1}-\sqrt{m_2}}{\sqrt{m_1}+\sqrt{m_2}}\right)^2 \left|\frac{F(a,-a(\nu-\mu)^2,\nu-\mu,\nu-\mu,2\nu,0;1)}{F(a,-a(\nu+\mu)^2,\nu+\mu,\nu+\mu,2\nu,0;1)}\right|^2.$$

Using Eq. (A16) of the Appendix A,

$$F(a,-a\alpha\beta;\alpha,\beta,\gamma,0;z)\equiv {}_2F_1(\alpha,\beta;\gamma;z),$$

and the following equivalence:

$${}_2F_1(\alpha,\alpha;\gamma;1) \underset{\alpha,\gamma\to\infty}{\approx} \exp\left(\frac{\alpha^2}{\gamma}\right),$$

we arrive at

$$|R|^2_{E\to\infty}=\left(\frac{\sqrt{m_1}-\sqrt{m_2}}{\sqrt{m_1}+\sqrt{m_2}}\right)^2. \tag{2.31}$$

Thus, the asymptotic value of the transmission coefficient is

$$|T|^2_{E\to\infty}=\frac{4\sqrt{m_1 m_2}}{(\sqrt{m_1}+\sqrt{m_2})^2}. \tag{2.32}$$

(b2) The value of the energy E for which the transmission coefficient reaches unity.

Relation (2.30) shows that $|R|^2=0$ (or $|T|^2=1$) when $(a-1)\nu+a\mu=0$, i.e., when the energy takes the value

$$E_t=\frac{m_1 V_0}{(m_1-m_2)}. \tag{2.33}$$

Hence, we see that the transmission coefficient reaches unity for the specific value E_t of the energy, only in the case $m_1>m_2$.

We come to the conclusion that the behavior of the transmission coefficient as a function of the energy E is similar to that of an abrupt potential and mass step,[25] setting

$$\tilde{k}=\sqrt{\frac{2|m_2-m_1|}{\hbar^2}\,V_0}. \tag{2.34}$$

(i) When $m_1<m_2$, $\omega=\tilde{k}r>0$ is real, we obtain

$$|R|^2=\left|\frac{F(a,ar^2[\tilde{k}^2+(k_1+k_2)^2]+ir[ak_1+(a-1)k_2];r[\tilde{k}-i(k_1+k_2)],-r[\tilde{k}+i(k_1+k_2)],1-2ik_2r,-1;1)}{F(a,ar^2[\tilde{k}^2+(k_1-k_2)^2]-ir[ak_1+(1-a)k_2];r[-\tilde{k}+i(k_1-k_2)],r[\tilde{k}+i(k_1-k_2)],1-2ik_2r,-1;1)}\right|^2. \tag{2.35}$$

The transmission coefficient increases from *zero* for $E=V_0$ to the asymptotic value (2.32) when $E\to\infty$, and never reaches unity.

(ii) When $m_1>m_2$, $\omega=i\tilde{k}r$ is imaginary, we have

$$|R|^2=\left|\frac{F(a,ar^2[-\tilde{k}^2+(k_1+k_2)^2]+ir[ak_1+(a-1)k_2];ir[\tilde{k}-k_1-k_2],-ir[\tilde{k}+k_1+k_2],1-2ik_2r,-1;1)}{F(a,ar^2[-\tilde{k}^2+(k_1-k_2)^2]-ir[ak_1+(1-a)k_2];ir[-\tilde{k}+k_1-k_2],ir[\tilde{k}+k_1-k_2],1-2ik_2r,-1;1)}\right|^2. \tag{2.36}$$

The transmission coefficient increases from *zero* for $E=V_0$, reaches unity for a specific value (2.33) of the energy, and then decreases to reach the asymptotic value (2.32).

Finally we consider the two following limiting cases.

(i) When the smooth potential and mass step tend to an abrupt potential and mass step, i.e., when the parameter r of definitions (2.1) and (2.2) tends to 0. Using the following special value of the Heun's function $F(a,0;0,1,2,0;1)=1$, we can see that each of the four Heun's functions of the relation (2.30) tend to 1 when $r\to 0$; then

$$|R|^2_{r\to 0}=\left(\frac{k_1/m_1-k_2/m_2}{k_1m_1+k_2/m_2}\right)^2. \tag{2.37}$$

Lévy-Leblond[25] found this result (2.37) by imposing, in the discontinuity of the abrupt potential and mass step, the continuity of the wave function and his derivative divided by the mass.

(ii) When the width of the mass step is vanishing, i.e., when the parameter a of the definition (2.6) tends to infinity. Using the following limit,[35] which gives one of the cases where the Heun's function degenerates into a hypergeometric function:

$$F(a,ac;\alpha,\beta,\gamma,\delta;z)_{a\to\infty}={}_2F_1(\eta+\sqrt{\eta^2+c},\eta-\sqrt{\eta^2+c};\gamma;z),$$

where

$$\eta=\frac{\alpha+\beta-\delta}{2} \quad \text{and} \quad \gamma\neq -n(n=0,1,2,...), \tag{2.38}$$

we recover all the results (wave function, reflection, and transmission coefficients) for a system with a constant mass and potential smooth step.[32] Particularly, and knowing that

$${}_2F_1(\alpha,\beta;\gamma;1)=\frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)},$$

reflection coefficient (2.27) tends to

$$|R|^2_{a\to\infty}=\left|\frac{{}_2F_1(\bar{\nu}-\bar{\mu},\bar{\nu}-\bar{\mu}+1;2\bar{\nu}+1;1)}{{}_2F_1(\bar{\nu}+\bar{\mu},\bar{\nu}+\bar{\mu}+1;2\bar{\nu}+1;1)}\right|^2=\left|\frac{\Gamma(+2\bar{\mu})\Gamma(\bar{\nu}-\bar{\mu})\Gamma(\bar{\nu}-\bar{\mu}+1)}{\Gamma(-2\bar{\mu})\Gamma(\bar{\nu}+\bar{\mu})\Gamma(\bar{\nu}+\bar{\mu}+1)}\right|^2=|\bar{R}|^2, \tag{2.39}$$

where

$$\bar{\nu}^2=\frac{2mr^2}{\hbar^2}(V_0-E), \quad \bar{\mu}^2=-\frac{2mr^2E}{\hbar^2}, \tag{2.40}$$

and $|\bar{R}|^2$ is the reflection coefficient for a system with a constant mass and potential smooth step obtained in Ref. 32.

III. CONCLUSION

We have succeeded in solving the one-dimensional generalized Schrödinger equation for a system with potential and mass, both of smooth step shape. We have found that the wave function depends on the Heun's function, which is a generalization of the hypergeometric function. We have shown that the behavior of the transmission coefficient as a function of the energy is similar to that of the case of an abrupt potential and mass step. We have also recovered the reflection and

transmission coefficients of two limiting cases, namely, when the width of the mass step is vanishing and when the smooth potential and mass step tend to an abrupt potential and mass step.

We think that our work may be used, in the future, for concrete computations in the area of semiconductor heterostructures, and more generally in that of the inhomogeneous crystals that is quickly developing.

APPENDIX A: RELATION BETWEEN THE z AND $1-z$ ARGUMENT HEUN'S FUNCTIONS

Let

$$y_1(z)=F(a,b;\alpha,\beta,\gamma,\delta;z), \tag{A1}$$

the Heun's function solution of the normal form of the Heun's equation,[36]

$$\frac{d^2y}{dz^2}+\left(\frac{\gamma}{z}+\frac{1+\alpha+\beta-\gamma-\delta}{z-1}+\frac{\delta}{z-a}\right)\frac{dy}{dz}+\frac{\alpha\beta z+b}{z(z-1)(z-a)}\,y=0, \tag{A2}$$

which is regular in the neighborhood of $z=0$ and belongs to the exponent *zero*. The two solutions linearly independent in the neighborhood of $z=1$ are[37]

$$y_3(z)=F(1-a,-b-\alpha\beta;\alpha,\beta,1+\alpha+\beta-\gamma-\delta,\delta;1-z), \tag{A3}$$

and

$$\begin{aligned} y_4(z)=(1-z)^{\gamma+\delta-\alpha-\beta}F(1-a,-b-\alpha\beta-[\gamma+\delta-\alpha-\beta][\gamma+\delta-a\gamma];\gamma \\ +\delta-\alpha,\gamma+\delta-\beta,1+\gamma+\delta-\alpha-\beta,\delta;1-z). \end{aligned} \tag{A4}$$

The solution $y_1(z)$ is a linear combination of the two solutions $y_3(z)$ and $y_4(z)$,

$$y_1(z)=C_1y_3(z)+C_2y_4(z), \tag{A5}$$

where C_1 and C_2 are independent of z, but depend on the other parameters. With the help of (A3) and (A4) the equation (A5) becomes

$$\begin{aligned} F(a,b;\alpha,\beta,\gamma,\delta;z)=C_1(a,b;\alpha,\beta,\gamma,\delta)F(1-a,-b-\alpha\beta;\alpha,\beta,1+\alpha+\beta-\gamma-\delta,\delta;1-z) \\ +C_2(a,b;\alpha,\beta,\gamma,\delta)(1-z)^{\gamma+\delta-\alpha-\beta}F(1-a,-b-\alpha\beta-[\gamma+\delta-\alpha-\beta] \\ \times[\gamma+\delta-a\gamma];\gamma+\delta-\alpha,\gamma+\delta-\beta,1+\gamma+\delta-\alpha-\beta,\delta;1-z). \end{aligned} \tag{A6}$$

Putting $z=1$ into the functional relation (A6) and knowing that $F(a,b;\alpha,\beta,\gamma,\delta;0)=1$, we obtain

$$C_1(a,b;\alpha,\beta,\gamma,\delta)=F(a,b;\alpha,\beta,\gamma,\delta;1). \tag{A7}$$

On the other hand, we have the relation[38]

$$F(a,b;\alpha,\beta,\gamma,\delta;z)=(1-z)^{\gamma+\delta-\alpha-\beta}F(a,b-a\gamma[\gamma+\delta-\alpha-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;z). \tag{A8}$$

The substitutions $\alpha\to\gamma+\delta-\alpha$, and $\beta\to\gamma+\delta-\beta$ into (A6) and (A8) give

$$\begin{aligned} F(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;z)=C_1(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta)F(1-a,-b-[\gamma+\delta-\alpha] \\ \times[\gamma+\delta-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,1+\gamma+\delta-\alpha-\beta,\delta;1-z) \\ +C_2(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta)(1-z)^{\alpha+\beta-\gamma-\delta} \\ \times F(1-a,-b-\alpha\beta+a\gamma[\alpha+\beta-\gamma-\delta];\alpha,\beta,1 \\ +\alpha+\beta-\gamma-\delta,\delta;1-z), \end{aligned} \tag{A9}$$

and

$$F(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;z)$$
$$=(1-z)^{\alpha+\beta-\gamma-\delta}F(a,b+a\gamma[\gamma+\delta-\alpha-\beta];\alpha,\beta,\gamma,\delta;z). \tag{A10}$$

Applying the relation (A9) to the Heun's function of the right-hand side of (A10), Eq. (A10) takes the form

$$\begin{aligned}F(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;z)&\\ =(1-z)^{\alpha+\beta-\gamma-\delta}C_1(a,b+a\gamma[\gamma+\delta-\alpha-\beta];a,\beta,\gamma,\delta)F(1-a,-b&\\ -a\gamma[\gamma+\delta-\alpha-\beta]-\alpha\beta,\alpha,\beta,1+\alpha+\beta-\gamma-\delta,\delta;1-z)&\\ +C_2(a,b+a\gamma[\gamma+\delta-\alpha-\beta];\alpha,\beta,\gamma,\delta)F(1-a,-b-[\gamma+\delta-\alpha]&\\ \times[\gamma+\delta-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,1+\gamma+\delta-\alpha-\beta,\delta;1-z).&\end{aligned} \tag{A11}$$

The comparison between (A9) and (A11) gives

$$C_1(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta)=C_2(a,b+a\gamma[\gamma+\delta-\alpha-\beta];\alpha,\beta,\gamma,\delta), \tag{A12}$$

and

$$C_2(a,b;\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta)=C_1(a,b+a\gamma[\gamma+\delta-\alpha-\beta];\alpha,\beta,\gamma,\delta). \tag{A13}$$

With the help of (A7), relation (A13) can be written as

$$\begin{aligned}C_2(a,b;\alpha,\beta,\gamma,\delta)&=C_1(a,b-a\gamma[\gamma+\delta-\alpha-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta)\\ &=F(a,b-a\gamma[\gamma+\delta-\alpha-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;1).\end{aligned} \tag{A14}$$

Finally, we obtain the functional relation between the z and $1-z$ argument Heun's functions,

$$\begin{aligned}F(a,b;\alpha,\beta,\gamma,\delta;z)&=F(a,b;\alpha,\beta,\gamma,\delta;1)F(1-a,-b-\alpha\beta;\alpha,\beta,1+\alpha+\beta-\gamma-\delta,\delta;1-z)\\ &+(1-z)^{\gamma+\delta-\alpha-\beta}F(a,b-a\gamma[\gamma+\delta-\alpha-\beta];\gamma+\delta-\alpha,\gamma+\delta-\beta,\gamma,\delta;1)\\ &\times F(1-a,-b-\alpha\beta-[\gamma+\delta-\alpha-\beta][\gamma+\delta-a\gamma];\gamma+\delta-\alpha,\gamma+\delta\\ &-\beta,1+\gamma+\delta-\alpha-\beta,\delta;1-z).\end{aligned} \tag{A15}$$

With[34]

$$F(a,-a\alpha\beta;\alpha,\beta,\gamma,0;z)={}_2F_1(\alpha,\beta;\gamma;z), \tag{A16}$$

we see that if we choose $b=-a\alpha\beta$ and $\delta=0$, Eq. (A15) is reduced to the well-known relation, which links the z and $1-z$ argument hypergeometric functions,

$$\begin{aligned}{}_2F_1(\alpha,\beta;\gamma;z)=&\frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}\,{}_2F_1(\alpha,\beta;\gamma;1-z)\\ &+\frac{\Gamma(\gamma)\Gamma(\alpha+\beta-\gamma)}{\Gamma(\alpha)\Gamma(\beta)}(1-z)^{\gamma-\alpha-\beta}\,{}_2F_1(\gamma-\alpha,\gamma-\beta;1+\gamma-\alpha-\beta;1-z).\end{aligned} \tag{A17}$$

APPENDIX B: RELATION BETWEEN THE *CONTIGUOUS* HEUN'S FUNCTIONS $F(a,b;\alpha,\beta,\gamma,\delta;z)$ AND $F(a,b';\alpha,\beta+1,\gamma+1,\delta+1;z)$

In order to establish this relation, we follow the Heun's method.[39] To avoid notation problems of the Heun's functions, we adopt in this appendix the Heun's notations[26] (indexed by the letter H). In the end of this appendix, we restore the Snow's[28] notations (without indexes) of the Heun's functions that we have adopted in the bulk of this paper.

Let

$$y_0 = P\left\{ \begin{array}{cccccc} 0 & 1 & a & \infty & \\ 0 & 0 & 0 & \alpha & z \\ 1-\gamma & 1-\eta & 1-\epsilon_0 & \beta & \end{array} \right\}, \tag{B1}$$

where

$$\epsilon_0 = \alpha+\beta-\gamma-\eta+1, \tag{B2}$$

the Riemann's P symbol of the Heun's function,

$$y_0 = F_{\mathrm{H}}(a,q;\alpha,\beta,\gamma,\eta;z), \tag{B3}$$

where q is the —so-called— characteristic value. Heun's function y_0 is a solution of the following differential equation;

$$\psi \frac{d^2y_0}{dz^2} + \chi \frac{dy_0}{dz} + \widetilde{\omega} y_0 = 0, \tag{B4}$$

where

$$\psi = z(z-1)(z-a),$$

$$\chi = a\gamma - \{\alpha+\beta+a\gamma+(a-1)\eta+1\}z + (\alpha+\beta+1)z^2,$$

$$\widetilde{\omega} = \alpha\beta(z-q). \tag{B5}$$

Let

$$y = P\left\{ \begin{array}{cccccc} 0 & 1 & a & \infty & \\ 0 & 0 & 0 & \alpha & z \\ -\gamma & 2-\eta & 1-\epsilon & \beta+1 & \end{array} \right\}, \tag{B6}$$

where

$$\epsilon = \alpha+\beta-\gamma-\eta+2, \tag{B7}$$

the Riemann's P symbol of the Heun's function,

$$y = F_{\mathrm{H}}(a,q';\alpha,\beta+1,\gamma+1,\eta-1;z). \tag{B8}$$

The difference between the elements (exponents) of the P's symbols (B6) and (B1) gives the following matrix:

$$\Delta P \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 \end{pmatrix}.$$

Thus, the b_i's[39] parameters are, in our case,

$$b_1 = b_3 = -1, \quad b_2 = b_4 = 0 \quad \text{and} \quad D = b_1+b_2+b_3+b_4 = -2. \tag{B9}$$

We introduce then the change of the dependent variable,

$$y = z^{b_1}(z-1)^{b_2}(z-a)^{b_3}u = z^{-1}(z-a)^{-1}u. \tag{B10}$$

This change of variable yields to the following P symbol for the u function:

$$u = P\left\{\begin{array}{cccc} 0 & 1 & a & \infty \\ 1 & 0 & 1 & \alpha-2 \quad z \\ 1-\gamma & 2-\eta & 2-\epsilon & \beta-1 \end{array}\right\}. \tag{B11}$$

Therefore function u verifies the differential equation

$$z(z-a)\psi \frac{d^2u}{dz^2} + z(z-a)\widetilde{\theta} \frac{du}{dz} + \widetilde{\sigma} u = 0, \tag{B12}$$

where

$$\widetilde{\theta} = (\gamma-1)(z-1)(z-a) + (\eta-1)z(z-a) + (\alpha+\beta-\gamma-\eta)z(z-1),$$

$$\widetilde{\sigma} = -2\psi - (2z-a)\widetilde{\theta} + \alpha(\beta+1)z(z-a)(z-q'). \tag{B13}$$

Functions y_0 and u are related by

$$u = \varphi y_0 + \psi \frac{dy_0}{dz}, \tag{B14}$$

where φ is a polynomial of second degree. Combining the relations (B14), (B12) and (B4), we can eliminate the derivative d^2y_0/dz^2, the u function, and his derivatives d^2u/dz^2, du/dz. Thus, we arrive at the relation

$$\{[\widetilde{\sigma} - z(z-a)\widetilde{\omega}]\varphi + z(z-a)(\widetilde{\theta}\varphi' + \psi\varphi'') - z(z-a)[(\widetilde{\theta}-\chi)\widetilde{\omega} + (\psi\widetilde{\omega})']\}y_0$$
$$+ \{(\widetilde{\theta}-\chi)\varphi + 2\psi\varphi' + (\widetilde{\theta}-\chi)(\psi'-\chi) + \psi(\psi''-\chi'-\widetilde{\omega}) + (z-1)\widetilde{\sigma}\} \frac{dy_0}{dz} = 0. \tag{B15}$$

Relation (B15) implicates that the coefficient of y_0 and that of dy_0/dz must identically vanish, i.e.,

$$\left.\begin{array}{c} [\widetilde{\sigma} - z(z-a)\widetilde{\omega}]\varphi + z(z-a)(\widetilde{\theta}\varphi' + \psi\varphi'') - z(z-a)[(\widetilde{\theta}-\chi)\widetilde{\omega} + (\psi\widetilde{\omega})'] = 0, \\ (\widetilde{\theta}-\chi)\varphi + 2\psi\varphi' + (\widetilde{\theta}-\chi)(\psi'-\chi) + \psi(\psi''-\chi'-\widetilde{\omega}) + (z-1)\widetilde{\sigma} = 0. \end{array}\right\} \tag{B16}$$

All the terms with the same degree in z in the two equations of (B16) must be equal to *zero*. Collecting the terms with the same degree in z in the two equations of (B16), we arrive at a system of 11 equations. Resolution of this system gives, for the function φ of (B14),

$$\varphi = \beta z(z-a). \tag{B17}$$

Relation (B14) then takes the form

$$mF_{\mathrm{H}}(a,q';\alpha,\beta+1,\gamma+1,\eta-1;z) = \beta F_{\mathrm{H}}(a,q;\alpha,\beta,\gamma,\eta;z) + (z-1)\frac{d}{dz}F_{\mathrm{H}}(a,q,\alpha,\beta,\gamma,\eta;z). \tag{B18}$$

Constant m of this last relation is determined by setting $z=0$ in (B18),

$$m = \beta - \frac{\alpha\beta q}{\gamma a}. \tag{B19}$$

Relation (B18) then takes the final form

$$(\gamma a\beta-\alpha\beta q)F_{\mathrm{H}}(a,q';\alpha,\beta+1,\gamma+1,\eta-1;z)$$
$$=\gamma a\beta F(a,q;\alpha,\beta,\gamma,\eta;z)+\gamma a(z-1)\frac{d}{dz}F_{\mathrm{H}}(a,q;\alpha,\beta,\gamma,\eta;z). \tag{B20}$$

Putting $q=1$, $a=1$ in (B20), and knowing that

$$F_{\mathrm{H}}(a,q;\alpha,\beta,\gamma,\eta;z)_{a\to 1,q\to 1}={}_2F_1(\alpha,\beta;\gamma;z),$$

relation (B20) tends to the one of the Gauss' recursion functions,[40]

$$(\gamma-\alpha){}_2F_1(\alpha,\beta+1;\gamma+1;z)=\gamma\,{}_2F_1(\alpha,\beta;\gamma;z)+\alpha(z-1){}_2F_1(\alpha+1,\beta+1;\gamma+1;z). \tag{B21}$$

Finally, we can now rewrite the relation (B20) with Snow's notations, which we have adopted in our paper,

$$(\gamma a\beta+b)F(a,b-\alpha;\alpha,\beta+1,\gamma+1\,\delta+1;z)$$
$$=\gamma a\beta F(a,b;\alpha,\beta,\gamma,\delta;z)+\gamma a(z-1)\frac{d}{dz}F(a,b;\alpha,\beta,\gamma,\delta;z), \tag{B22}$$

where we have used the following correspondence rule between the two notations:

$$b\equiv-\alpha\beta q,\quad b'\equiv-\alpha(\beta+1)q'=-\alpha\beta q-\alpha q'-\alpha\beta(q'-q)=b-\alpha. \tag{B23}$$

These latter relations are deduced from the resolution of the system of 11 equations.

[1] J.-P. Peng, Y.-M. Mu, and X.-C. Shen, ''Resonant level lifetime in the double barrier structure for both quasi-bound and extended states,'' J. Appl. Phys. **73**, 989–991 (1993).
[2] R. Kucharczyk and M. Stęslicka, ''Density-of-states formalism for multi-quantum-barrier structures,'' Solid State Commun. **84**, 727–730 (1992).
[3] F. Berz, ''Transmission coefficient and Stark shifts in double-barrier quantum well structures,'' Semicond. Sci. Technol. **8**, 243–253 (1993).
[4] J. D. Bruno and T. B. Bahder, ''Local density of states in double-barrier resonant-tunneling structures. II. Finite-width barriers,'' Phys. Rev. B **39**, 3659–3663 (1989).
[5] M. Stęslicka, R. Kucharczyk, and M. L. Glasser, ''Surface states in superlattices,'' Phys. Rev. B **42**, 1458–1461 (1990).
[6] Qi-G. Zhu and H. Kroemer, ''Interface connection rules for effective-mass wave functions at an abrupt heterojunction between two different semi-conductors,'' Phys. Rev. B **27**, 3519–3527 (1983).
[7] D. J. Ben Daniel and C. B. Duke, ''Space-charge effects on electron tunneling,'' Phys. Rev. **152**, 683–692 (1966).
[8] C. Weisbuck and B. Vinter, *Quantum Semiconductor Structures* (Academic, Boston, 1991).
[9] J. Thomsen, G. T. Einevoll, and P. C. Hemmer, ''Operator ordering in effective-mass theory,'' Phys. Rev. B **39**, 12 783–12 788 (1989).
[10] R. A. Morrow and K. R. Brownstein, ''Model effective-mass Hamiltonians for abrupt heterojunctions and the associated wave-function-matching conditions,'' Phys. Rev. B **30**, 678–680 (1984).
[11] R. A. Morrow, ''Establishment of an effective-mass Hamiltonian for abrupt heterojunctions,'' Phys. Rev. B **35**, 8074–8079 (1987).
[12] R. A. Morrow, ''Effective-mass Hamiltonians for abrupt heterojunctions in three dimensions,'' Phys. Rev. B **36**, 4836–4840 (1987).
[13] G. T. Einvoll, ''Operator ordering in effective-mass theory for heterostructures. II. Strained systems,'' Phys. Rev. B **42**, 3497–3502 (1990).
[14] K. Young, ''Position-dependent effective mass for inhomogeneous semiconductors,'' Phys. Rev. B **39**, 13 434–13 441 (1989).
[15] G. T. Einevoll and P. C. Hemmer, ''The effective-mass Hamiltonian for abrupt heterostructures,'' J. Phys. C **21**, L1193–L1198 (1988).
[16] O. Von Roos, ''Position-dependent effective masses in semiconductor theory,'' Phys. Rev. B **27**, 7547–7552 (1983).
[17] W. Trzeciakowski, ''Boundary conditions and interface states in heterostructures,'' Phys. Rev. B **38**, 4322–4325 (1988).
[18] I. Galbraith and G. Duggan, ''Envelope-function matching conditions for GaAs/(Al, Ga)As heterojunctions,'' Phys. Rev. B **38**, 10 057–10 059 (1988).
[19] T. Gora and F. Williams, ''Theory of electronic states and transport in graded mixed semiconductors,'' Phys. Rev. **177**, 1179–1182 (1969).
[20] O. Von Roos and H. Mavromatis, ''Position-dependent effective masses in semiconductor theory. II,'' Phys. Rev. B **31**, 2294–2298 (1985).
[21] G. T. Einevoll, P. C. Hemmer, and J. Thomsen, ''Operator ordering in effective-mass theory for heterostructures. I. Comparison with exact results for superlattices, quantum wells, and localized potentials,'' Phys. Rev. B **42**, 3485–3496 (1990).

[22] J.-M. Lévy-Leblond, ''Position-dependent effective mass and Galilean invariance,'' Phys. Rev. A **52**, 1845–1849 (1995).
[23] G. Bastard, ''Superlattice band structure in the envelope-function approximation,'' Phys. Rev. B **24**, 5693–5697 (1981).
[24] L. Chetouani, L. Dekar, and T. F. Hammann, ''Green's functions via path integrals for systems with position-dependent masses,'' Phys. Rev. A **52**, 82–91 (1995).
[25] J.-M. Lévy-Leblond, ''Elementary quantum models with position-dependent mass,'' Eur. J. Phys. **13**, 215–218 (1992).
[26] K. Heun, ''Zur Theorie der Riemann'schen Functionen zweiter Ordnung mit vier Verweigungspunkten,'' Math. Ann. **33**, 161–179 (1889).
[27] K. Heun, ''Beiträge zur Theorie der Lamé'schen Functionen,'' Math. Ann. **33**, 180–196 (1889).
[28] C. Snow, *Hypergeometric and Legendre Functions With Applications to Integral Equations of Potential Theory*, National Bureau of Standards Applied Mathematics Series, 1952, Vol. 19, Chap. VII, p. 87–101.
[29] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1955), Vol. III, Sec. 15.3.
[30] A. Erdélyi, ''The Fuchsian equation of second order with four singularities,'' Duke Math. J. **9**, 48–58 (1942).
[31] A. Erdélyi, ''Certain expansions of solutions of the Heun equation,'' Q. J. Math. Oxford Ser. **15**, 62–69 (1944).
[32] S. Flügge, *Practical Quantum Mechanics I* (Springer-Verlag, Berlin, 1971), p. 86–89.
[33] L. Landau and E. Lifchitz, *Mécanique Quantique* (Mir, Moscou, 1967), pp. 100–101.
[34] Reference 28, p. 89, Eqs. (6a), (7), and (7′).
[35] Reference 28, p. 91, Eq. (9b).
[36] Reference 28, p. 88, Eq. (3).
[37] Reference 28, p. 98, Eqs. (20a) and (20b).
[38] Reference 28, p. 95, Eq. (16a).
[39] Reference 26, Sec. 4, pp. 173–179.
[40] I. S. Gradshtein and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, San Diego, 1994), Eq. (9.137.7).

Intermediate level statistics with one-parameter random matrix ensembles

Hiroshi Hasegawa
Research Center of Quantum Communications, Tamagawa University, Machida, Tokyo 194, Japan

Jian-Zhong Ma
Max-Planck Institut für Kern Physik, D-69117 Heidelberg, Germany

(Received 3 November 1997; accepted for publication 2 December 1997)

We discuss a formulation of the level statistics for quantum systems that lie in an intermediate regime between the integrability subject to Poisson statistics and the full chaos subject to Gaussian statistics of the typical random matrix theory (RMT). It is based on the idea initiated by Yukawa and also by Nakamura and Lakshmanan, namely, to transcribe the eigenvalue statistics into statistical mechanics of the completely integrable Calogero system with internal degrees of rotation. By analyzing the previous works of Gaudin and Forrester from this viewpoint, we answer the question raised by Mehta and Dyson in early RMT concerning the expectation of compressible level gas, which is realized now in such an intermediate regime. It indicates the general absence of a fractional power dependence of the two-level correlation function under the ordinary thermodynamic-limit procedure as the manifestation that a transition between localized and delocalized states does not occur. The predictability of the transition by modifying the theory is remarked. Two applications of the present scheme are shown: the information loss per level and the correlation hole of survival probability for intermediate circular unitary ensembles. © *1998 American Institute of Physics.* [S0022-2488(98)03105-3]

I. INTRODUCTION

The approach of energy level statistics has been applied in many branches of physics, such as complex nuclei,[1] quantum soft chaos,[2] and disordered electronic systems.[3] In some extreme cases (the strong-chaos regime or metallic region) quantum spectra have been shown[4] to follow the prediction of random matrix theories (RMT),[5] while in the opposite cases (classical integrability; insulator region) the Poisson statistics is applicable. The problem of interpolation between the Poisson and the RMT has long been discussed for quantifying the description of level statistics of intermediate situations between these extremes.

Theoretical efforts made thus far to approach the goal may be classified into two categories: In the first category, one wishes to extend the successful periodic-orbit theory of level statistics[6] and analyzes the structure of classical phase spaces depending on the degree of the softness of chaos.[7] In the second category, one wishes to find out a systematic means to generalize the underlying ensembles of random matrices from the typical RMT extremes; the approach that started in the paper by Berry and Robnik.[8]

In the present paper, we adopt the so-called *level dynamics* approach (its name taken from Haake's book[9]) in the second category: its first idea appeared in Pechukas,[10] as motivated from the early work of Dyson,[11] and then put into explicit formulations by Yukawa[12] and by Nakamura and Lakshmanan,[13] further followed by Haake and his co-workers,[9] Nakamura and Mikeska,[14] and Gaspard, Rice, Mikeska, and Nakamura.[15]

The common idea of the level dynamics approach quoted above is the observation that a random sequence of quantum levels whose correlations between each other are of a repulsive nature could be treated as a system of interacting (classical) particles arranged on a line; specifically, the Calogero–Moser system with the pair potential of the inverse square type (for which the line is extended to infinity) or the Calogero–Sutherland system with the pair potential of the inverse square *sine* function (for which the line becomes a circle, and hence the system is peri-

 © 1998 American Institute of Physics

odic). Strictly speaking, the Calogero–Moser/Sutherland system employed here is more complex than the original model,[16,17] in that the coefficient of each pair potential is not constant but subject to a dynamical law prescribed in accordance with a set of Poisson bracket relations that has analogy to angular momentum components in the N-dimensional space. Physically, it means that each constituent particle of the system has internal degrees of freedom associated with such a rotation induced by the angular momentum; hence, the name of *generalized Calogero–Moser* (g-CM)/*Sutherland* (g-CS) system.

Here, we present its Hamiltonian function in term of three sets of dynamical variables; the first two $\{x_n\}$, $\{p_n\}$; $(n=1,2,...,N)$, for the translation of N-particles, and the second $\{f_{n,m}\}$ for the N-dimensional rotation, respectively. Also, the relevant Poisson bracket relations among them are added in accordance with a description by Gaspard *et al.*:[15]

$$\mathscr{H}_{\text{g-CM}}=\frac{1}{2}\sum_n p_n^2+\frac{1}{2}\sum_{m\neq n}\frac{\|f_{mn}\|^2}{(x_m-x_n)^2}, \tag{1.1}$$

$$\mathscr{H}_{\text{g-CS}}=\frac{1}{2}\sum_n p_n^2+\frac{1}{2}\sum_{m\neq n}\frac{\|f_{mn}\|^2}{4\sin^2(\pi/L)(x_m-x_n)}, \tag{1.2}$$

$$\{x_m,p_n\}=c\,\delta_{mn}\quad (c=1 \text{ for g-CM, and } L/2\pi \text{ for g-CS}), \tag{1.3a}$$

$$\{x_m,x_n\}=\{p_m,p_n\}=\{x_m,f_{rs}\}=\{p_m,f_{rs}\}=0, \tag{1.3b}$$

$$\{f_{mn}^{(a)},f_{rs}^{(b)}\}=-\sum_{pqc} c_{mna,rsb}^{pqc} f_{pq}^{(c)} \tag{1.3c}$$

(Superscript a,b, etc. denotes the component of a complex number, i.e. *real* and *imaginary* part for a unitary ensemble (UE) and of a quaternion for a symplectic ensemble (SE), respectively (see below), and $\|f_{mn}\|^2=\sum_{a=1}^{\nu}|f_{mn}^{(a)}|^2$.)

It has been argued[13,14] that the set of equations of motion deduced from the above Hamiltonian function with Poisson bracket relations (1.3) is completely integrable, so that any solution of it may be obtained by quadrature in terms of the $2f$ initial values of the variables, where f is the degree of freedom of the system, i.e.

$$f=N+\frac{\nu}{2}N(N-1), \tag{1.4}$$

with $\nu=1$ for an orthogonal ensemble (OE), and $\nu=2,4$ unitary and symplectic ensembles (UE), (SE), respectively.

The following fact can easily be deduced:[9,12,18] *Let H_0 and V denote two arbitrary Hermitian* (*real symmetric for $\nu=1$, complex Hermitian for $\nu=2$, quaternion real Hermitian for $\nu=4$*) *matrices*. Then, all the matrix elements of

$$H_\lambda\equiv H_0+\lambda V,\quad \text{diag } V \quad\text{and}\quad [H_\lambda,V],\quad \text{on the } H_\lambda \text{ diagonal basis}, \tag{1.5}$$

which are denoted by x_n, p_n and f_{mn}, respectively, satisfy a set of equations of motion (with time variable λ) for the g-CM *system (1.1). Similarly, all the matrix elements of*

$$U_\lambda\equiv e^{-i\lambda V}e^{-iH_0},\quad \text{diag } V \quad\text{and}\quad i(e^{iH_0}Ve^{-iH_0}-V),\quad \text{on the } U_\lambda \text{ diagonal basis}, \tag{1.6}$$

which are denoted by $e^{i(2\pi/L)x_n}$, p_n and f_{mn}, respectively, satisfy a set of equations of motion for the g-CS *system (1.2).*

The question about the complete integrability of the g-CM/S system should be the converse of the above statement, namely, an affirmative answer to the question: *Is any solution of the set of equations of motion for the g-CM/S system so deduced from the prescription (1.1)–(1.3) repre-*

sentable in a matrix form H_λ in Eq. (1.5) or U_λ in Eq. (1.6), where all the matrix elements of H_0 and V give $2f$ initial values? It is by no means elementary to answer this question affirmatively, and here we want to state clearly that ''our answer is yes.'' [19]

Surely, therefore, these two dynamical systems are completely integrable. It implies the existence of precisely f independent, involutive global constants of motion associated with g-CM (1.1) and g-CS (1.2), each Hamiltonian function being only one of them. In his 1985 paper,[12] Yukawa proposed the construction of a statistical mechanics of the g-CM system by adding another constant of motion besides the Hamiltonian $\mathscr{H}=E$ (energy of the system), namely,

$$\mathscr{Q}=\frac{1}{2}\sum_{a=1}^{\nu}\sum_{m\neq n}|f_{mn}^{(a)}|^2 \tag{1.7}$$

(square of the angular momentum),

in the canonical ensemble of statistical mechanics; the form $e^{-\beta\mathscr{H}-\gamma\mathscr{Q}}$. Our first concern in this paper is to clarify the legitimacy of this ansatz for intermediate level statistics, and then its applications to actual problems.

We can observe that considerable interests have been focused recently on the universal correlation property of the particles in a one-dimensional system of the Calogero–Moser/Sutherland type (classical, quantum, and second-quantized sense). [This topic has been surveyed recently by Guhr, Müller-Groeling and Weidenmüller.[20]] Our aim in this paper is, by confining it to the classical g-CS system (which is more primitive than g-CM), to get a definite answer to the density correlation of the particles representing the levels in a random quantum spectrum within the above framework of level dynamics. A specific feature of the framework is that the complete integrability of the system necessarily modifies the ordinary basis of *ergodicity* for equilibrium statistical mechanics.

We therefore begin with an information theoretical formulation about the most probable canonical distribution, in Sec. II, which justifies Yukawa's one-parameter exponential family; a means to enlarge the classical RMT Gaussian ensembles that conforms to Balian.[21] In Sec. III, we discuss some general properties of the two-level correlation functions of this family. An exact one-parameter model of unitary ensembles studied by Gaudin in 1966[22] and reconsidered later by Forrester[23] turns out to be an integral reduction (equivalently, a coarse graining) of this family, and is used in Sec. IV to clarify the explicit form of this correlation function. It is significant to show a violation of the compatibility relations among the series of higher-order correlation functions, when the ordinary thermodynamic-limit procedure is adopted: this is discussed in Sec. V to provide a clear answer to the question of Mehta and Dyson in the beginning stage of RMT. This feature of the thermodynamic limiting at the cost of violating the compatibility in the second-order correlation function (the so-called ''sum rule'' [24]) affects the long-range behavior of the function as regards the metal–insulator transition. We further demonstrate some more applications of the present scheme of intermediate statistics in Sec. VI.

II. INFORMATION BASIS FOR THE G-CM/S CANONICAL DISTRIBUTION

The well-established N-level joint probability density function $P_N(\{x_n\})$ [or abbreviation $P_N(x)$] in RMT can be characterized by the standard *maximum entropy* principle as follows:[21] Let H denote an element of Hermitian matrix ensembles $\mathscr{E}_\nu$ ($N\times N$ and $\nu=1,2,4$ for OE, UE, SE, respectively, as before) and $P_\nu(H)d^\nu H$ a probability of the element to lie in an interval $(H,H+dH)$ in the matrix space $\mathscr{E}_\nu$, where $d^\nu H=\Pi_n dH_{nn}\times\Pi_{n<m}dH_{mn}^{(1)}\cdots dH_{mn}^{(\nu)}$, i.e., the product of all possible independent matrix-element increments. The Gaussian probability density $P_{\mathrm{G}\nu}(H)$ is determined uniquely from

$$\max_P\left[-\int_{\mathscr{E}_\nu}d^\nu H P(H)\log P(H)\right]=-\int_{\mathscr{E}_\nu}d^\nu H P_{\mathrm{G}}(H)\log P_{\mathrm{G}}(H), \tag{2.1}$$

under the constraint

$$\langle H_{mn}\rangle=0, \quad \text{and} \quad \langle H_{nn}^2\rangle=2\langle H_{mn}^{(a)2}\rangle(m\neq n)=\sigma^2, \quad \langle H_{mn}^{(a)}H_{rs}^{(b)}\rangle((m,n)\neq(r,s) \quad or \quad a\neq b)=0. \tag{2.2}$$

Briefly, among all possible distributions of H defined in $\mathscr{E}_\nu$ for which the first and second moment of the matrix random variable H exist, the Gaussian distribution,

$$P_{G\nu}(H)=\frac{1}{Z_\nu}e^{-(1/2\sigma^2)\mathrm{Tr}\,H^2}, \tag{2.3}$$

yields the most probable distribution of all those for which the first moments and the variances are prescribed uniformly as Eq. (2.2). The standard RMT N-level distribution,

$$P_N(x)dx_1dx_2\cdots dx_N=C_{N\nu}e^{-(1/2\sigma^2)\Sigma x_n^2}\prod_{m<n}|x_n-x_m|^\nu dx_1dx_2\cdots dx_N, \tag{2.4}$$

is known[5] to be deduced from the form $P_{G\nu}(H)d^\nu H$ by a change of the variables, such that

$$\{H_{mn}\}\to\{x_n\}(\text{all eigenvalues of } H), \tag{2.5}$$

and all other cyclic variables that do not enter.

A restricted nature of the standard RMT Gaussian distribution characterized above can be seen in the uniformity of constraint on the variance, i.e., the single constant factor in Eq. (2.3); a key word of enlarging the underlying matrix ensembles. A straightforward removal of such a uniformity may be seen in the idea of *band random matrix ensembles*.[25] Here, we discuss how to remove the ansatz of uniform variances for enlargement by analyzing possible canonical distribution of the g-CM/S system (1.1)/(1.2): Our assertion is as follows.

Among all the possible canonical distributions P of the form $Z(\{\gamma_j\})^{-1}\exp[-\Sigma_j\gamma_j\mathscr{Q}_j]$ for the g-CM/S system with f constants of motion $\{\mathscr{Q}_j\}$, the form

$$P_{\beta,\gamma}=\frac{1}{Z_{\beta,\gamma}}e^{-\beta\mathscr{H}-\gamma\mathscr{Q}} \tag{2.6}$$

yields the most probable canonical one by virtue of a certain entropy principle that we set up below.

We consider a conditional expectation value (i.e., a partial integration) of the entropy function $P(y)\log P(y)$ of two sets of variables y, $\{y_i\}=\mathscr{S}\times\mathscr{R}$;

$$\mathscr{R}=\{p_n\}\times\{f_{mn}\} \tag{2.7}$$

and

$$\mathscr{S}=\{x_n\}, \tag{2.8}$$

and the integration is carried out over the $\mathscr{R}$ variables that are irrelevant for information about the space $\mathscr{S}$; thus

$$h_P(x)\equiv-\int_{\mathscr{R}}dpdfP(x,p,f\,)\log\,P(x,p,f\,). \tag{2.9}$$

This is generally a function of the $\mathscr{S}$ variables. Then, our maximum entropy principle reads as

$$\max_P h_P(x)=h_m(x), \tag{2.10}$$

under the constraint, besides $\langle p\rangle_R=\langle f\rangle_R=0$,

$$\langle p_n^2\rangle_R,\langle|f_{mn}|^2\rangle_R=\text{prescribed as a function of } \{x_n\}.$$

We now show that the maximum entropy condition (2.10) under the following variance constraint determines the distribution (2.6) uniquely:

$$\langle p_n^2\rangle_R = 1/\beta, \quad \langle |V_{mn}|^2\rangle_R = \frac{1}{(x_n - x_m)^2}\langle |f_{mn}|^2\rangle_R = \frac{1}{2\beta}\,\frac{1}{1+(\gamma/\beta)(x_n-x_m)^2}, \tag{2.11}$$

for the g-CM system, and

$$\langle p_n^2\rangle_R = 1/\beta, \quad \langle |V_{mn}|^2\rangle_R = \frac{1}{4\sin^2(\pi/L)(x_n-x_m)}\langle |f_{mn}|^2\rangle_R = \frac{1}{2\beta}\,\frac{1}{1+(\gamma/\beta)4\sin^2(\pi/L)(x_n-x_m)}, \tag{2.12}$$

for the g-CS system. The reason for this result of the constraint in the present maximization problem is as follows.

(i) The special form of canonical distribution (Yukawa) (2.4) is Gaussian with respect to the subsystem $\mathscr{R}$, i.e., p and f variables, satisfying $\langle p\rangle_R = \langle f\rangle_R = 0$ and the variances given by Eq. (2.11)/(2.12).

(ii) Any canonical distribution of the form $Z^{-1}e^{-\Sigma_j \gamma_j \mathscr{Q}_j}$, if it is Gaussian with respect to the $\mathscr{R}$ variables with $\langle p\rangle_R = \langle f\rangle_R = 0$, must be identical to the form (2.6): in other words, expression (2.6) is the only Gaussian canonical distribution (with zero mean of the p and f variables) for the g-CM/S dynamics.

The strength of this latter statement can be verified immediately, if one proves the fact that $\mathscr{H}$ and $\mathscr{Q}$ are the only two constants of motion of the g-CM/S dynamics among $\{\mathscr{Q}_j\}$ that are quadratic with respect to all the matrix elements of V defining the solution H_λ in (1.5), or U_λ in (1.6). (See Ref. 18.)

From the second part of expression (2.11)/(2.12), which is identical to the averaged pair potential, we can observe the nature of the correlation between a pair of levels (particles in the system) and also the role of the added constant $\mathscr{Q}$ with coefficient γ: the correlation retains the inverse square type of the distance between the pair (on the entire line, or within the period L) inherent to the system, but it is depending on the ratio γ/β: when $\gamma/\beta = 0$, the averaged pair potential recovers the uniformity of the RMT ansatz (i.e., $\langle |V_{mn}|^2\rangle_R = 1/2\beta$), while in the opposite limit $\gamma/\beta = \infty$, it becomes a δ-function type ($\langle |f_{mn}|^2\rangle_R = 1/2\gamma$) showing a Poisson characteristic for the pair correlation.

The N-level joint distribution that has been determined above can be related to the maximized conditional entropy $h_m(x)$, such that

$$P_N(x) \equiv \int dp\,df\,P_{\beta,\gamma}(x,p,f) = C_N e^{h_m(x)}, \tag{2.13}$$

$h_m(x)$ being parametrized by the single ratio γ/β, and hence

$$C_N^{-1} = \int dx_1 dx_2 \cdots dx_N e^{h_m(x)} = C_N^{-1}(\gamma/\beta). \tag{2.14}$$

It is a one-parameter family of distributions, explicitly given by

$$P_N(x,\gamma/\beta) = C_{N\nu}(\gamma/\beta)\prod_{m<n}|x_m - x_n|^\nu [1+(\gamma/\beta)(x_m-x_n)^2]^{-\nu/2}, \quad \text{for g-CM}, \tag{2.15}$$

$$= C_{N\nu}(\gamma/\beta)\prod_{m<n}(2\sin(\pi/L)|x_m-x_n|)^\nu \times [1+(\gamma/\beta)4\sin^2(\pi/L)(x_m-x_n)]^{-\nu/2}, \quad \text{for g-CS}. \tag{2.16}$$

Our task, then, is to determine the correlation property of the particles subject to these distributions.

III. TWO-LEVEL CORRELATION FUNCTION

Following Dyson,[26] we define the n-level correlation function for the g-CS system in length L,

$$R_n(x_1,x_2,\ldots,x_n)=\frac{N!}{(N-n)!}\int_{(-L/2,L/2)}\cdots\int\left(\frac{2\pi}{L}\right)^{N-n}dx_{n+1}dx_{n+2}\cdots dx_N P(x_1,\ldots,x_N). \tag{3.1}$$

It is related to a set of cluster functions $T_n(x)$ such that every R_n can be decomposed into a sum of products of all lower-order T_n functions, as in the ordinary cluster-expansion theory.[27] Actually, Dyson used the angular variables θ in the range $[-\pi,\pi]$ instead of x in the range $[-L/2,L/2]$ by the assumed relation

$$\theta=\frac{2\pi}{L}x=\frac{2\pi}{N}\xi;\quad \xi=\rho x,\quad \rho\equiv\frac{N}{L} \tag{3.2}$$

(ξ denotes Dyson's scaled variable $\in[0,N)$: x becomes identical to ξ, if $\rho=1$),

so that $R_n(x)=R_n(\theta(x))$ holds. Let us write the relations between $\{T_n\}$ and $\{R_n\}$ for lower-order n's, as

$$T_1(x_1)=R_1(x_1)=\frac{N}{2\pi}, \tag{3.3a}$$

$$T_2(x_1,x_2)=-R_2(x_1,x_2)+R_1(x_1)R_1(x_2), \tag{3.3b}$$

$$T_3(x_1,x_2,x_3)=R_3(x_1,x_2,x_3)-[R_1(x_1)R_2(x_1,x_2)+c.p.(i=1,2,3)]+2R_1(x_1)R_1(x_2)R_1(x_3),\ \text{etc.} \tag{3.3c}$$

By definition (3.1), an integral identity holds involving R_n and R_{n-1}, such that

$$\int_{-L/2}^{L/2}R_n(x_1,\ldots,x_{n-1},x_n)\left(\frac{2\pi}{L}\right)dx_n=(N-n+1)R_{n-1}(x_1,\ldots,x_{n-1}), \tag{3.4}$$

which may be called the ''compatibility relation for R_n's.'' The corresponding compatibility relation for T_n's may read as

$$\int_{-L/2}^{L/2}T_n(x_1,\ldots,x_{n-1},x_n)\left(\frac{2\pi}{L}\right)dx_n=(n-1)T_{n-1}(x_1,\ldots,x_{n-1})\quad(n>1). \tag{3.5}$$

An issue about this compatibility, which was raised in the early RMT[28] (see Ref. 24), is whether or not this relation remains to hold in the thermodynamic limit of the N-particle system under consideration. If this is the case, one must have, in accordance with Dyson,[26]

$$\int_{-\infty}^{\infty}Y_n(x_1,\ldots,x_{n-1},x_n)dx_n=(n-1)Y_{n-1}(x_1,\ldots,x_{n-1})\quad(n>1), \tag{3.6a}$$

where

$$\lim_{N,L(\rho\ \text{fixed})\to\infty}\left(\frac{2\pi}{L}\right)^n T_n(x)\equiv Y_n(x),\quad \text{the limit called } \textit{thermodynamic limit}. \tag{3.6b}$$

In particular, the identity

$$\int_{-\infty}^{\infty}Y_2(x_1,x_2)dx_2=1\quad[Y_1(x_1)=1,\ \text{if } \rho=1,\ \text{called } \textit{unfolding scale}), \tag{3.7}$$

must hold, which has received considerable attention recently by the name of ''sum rule.'' [24] Toward this topic, we summarize some basic properties of the two-level correlation/cluster functions next.

A. Short-range property of R_2 function

One of the established fluctuation properties of quantum levels of classically chaotic systems is the level repulsion, i.e., the resistance of levels against crossings. This phenomenon is described by the behavior in small values of the nearest-neighbor spacing (NNS) distribution $P(s)$ for a given spacing value s [conditional probability density function $p(0;s)$ for the probability that there is no level in a given interval of length s, when its both end points are occupied]. It is computed from the N-level joint probability density function $P_N(x)$ via the ''hole probability'' $E(s)$ (=the unconditional probability that there is no level in a given interval of length s)[5]: Hereafter, we assume that 2π is absorbed in the system length L, but still retain the finiteness of L and ρ. Under the circumstance, the relation between $p(0;s)$ and $E(s)$ (Ref. 5) is modified as follows: [In the unfolded scale, i.e., in terms of Dyson's ξ, the NNS function $p(0;\xi)$ is given by $p(0;\xi)=(d^2/d\xi^2)E(\xi)$: in terms of the present variable, therefore, the second derivative must be multiplied by ρ^{-2}.]

$$
\begin{aligned}
p(0;s) &= \frac{1}{\rho^2}\frac{d^2}{ds^2}E(s) \\
&= \left(\frac{L}{N}\right)^2 \frac{d^2}{ds^2}\int_{(-L/2,-s/2)\cup(s/2,L/2)} P_N(x_1,x_2,\ldots,x_N)L^{-N}dx_1dx_2\cdots dx_N \\
&= \frac{N(N-1)}{N^2}\int_{(-L/2,-s/2)\cup(s/2,L/2)} P_N(-s/2,s/2,x_3,\ldots,x_N)L^{-(N-2)}dx_3\cdots dx_N, \qquad (3.8)
\end{aligned}
$$

where the factor $N(N-1)$ arises from $\frac{1}{2}N(N-1)$ pairs of the equal integrals and twice of the first pair ($x_1=-s/2$, $x_2=s/2$), and its interchange ($1\rightleftharpoons 2$). This suggests that we redefine the low-level correlation functions (3.1), omitting each fraction p/N; $p=1,\ldots,n-1$, by

$$
\begin{aligned}
r_n(x_1,\ldots,x_n) &= (1/N)^n R_n(x_1,\ldots,x_n) \\
&= \int_{(-L/2,L/2)}\cdots\int L^{-(N-n)}dx_{n+1}dx_{n+2}\cdots dx_N P(x_1,\ldots,x_N), \qquad (3.9a)
\end{aligned}
$$

and

$$
Y_n(x_1,\ldots,x_n)=\rho^{-n}\times Y_n(x), \quad \text{defined by Eq.(3.6b).} \qquad (3.9b)
$$

Our aim in this section is to show that the small-value behavior of the NNS function $p(0;s)$ can be represented very efficiently by that of the two-level correlation function $r_2(s)$. Namely, $p(0;s)=r_2(s)+[\text{terms of higher powers of } s]$, for $0<s\rightarrow 0$. More precisely, if the lowest-power term in s, $s\ll\bar{s}\equiv\int_0^\infty p(0;s)s\,ds$ [mean level spacing over $p(0;s)$] is $A_\nu s^\nu$ ($\nu>0$), then the lowest-power term of $r_2(s)$ is also $A_\nu s^\nu$, so that

$$
\lim_{s\rightarrow 0}\frac{p(0;s)-r_2(s)}{p(0;s)}=0. \qquad (3.10)
$$

The idea has been mentioned by Bohigas, Haq, and Pandey,[29] and we formulate it quantitatively, being motivated by a formula of Wigner:[30]

$$
p(0;s)=r_2(s)+\sum_{n=3}^{N}(-1)^n\frac{\rho^{n-2}}{(n-2)!}\int_{-s/2}^{s/2}\cdots\int_{-s/2}^{s/2}dx_3\cdots dx_n r_n\left(-\frac{s}{2},\frac{s}{2},x_3,\ldots,x_n\right). \qquad (3.11)
$$

This formula can be derived from Eq. (3.8) by employing the characteristic function for the integration region $(-L/2,-s/2)\cup(s/2,L/2)$, namely, $\Pi_{n=3}^{N}(1-a(x_n))$ with $a(x)=1$ for $|x|\leqslant s/2$, and $=0$ for $|x|>s/2$: we have an expansion,

$$\prod_{n=3}^{N}(1-a(x_n))=1-\sum_{n=3}^{N}a(x_n)+\cdots+(-1)^{N-2}a(x_3)\cdots a(x_N).$$

Property (3.10) certainly stems from the level repulsion, and can be verified for a class of N-level distributions, including (2.15) and (2.16), of the form

$$P_N(x)=\prod_{m<n}|x_n-x_m|^{\nu}\times F(x),\quad \nu>0,$$

$$F(x)=\text{a positive continuous function of } x(x_1,x_2,\ldots,x_N),\quad |x|<\infty.$$

By integrating the $P_N(x)$ in the above with respect to $N-n$ variables in the finite range $[-L/2,L/2]$ for each, we can get an estimate of the r_n function in (3.1) as

$$r_n(x_1,x_2,\ldots,x_n)\leqslant \text{const.}\prod_{i<j\leqslant n}|x_i-x_j|^{\nu},$$

and for

$$|x_i-x_j|\leqslant s,\quad r_n(x_1,x_2,\ldots,x_n)\leqslant C_n s^{(\nu/2)n(n-1)}. \tag{3.12}$$

Another continuity property of the r_n function (the mean-value theorem) enables us to rewrite the right-hand side of Eq. (3.11) as

$$p(0;s)=r_2(s)-\rho s r_3(-s/2,s/2,\epsilon s)+O(s^2 r_4(-s/2,s/2,\epsilon_1 s,\epsilon_2 s)),\quad (|\epsilon|,|\epsilon_n|<1/2),$$

the problem thus reducing to $O(sr_3)$, which can be combined to the estimate (3.12).

We therefore get

$$p(0;s)\propto s^{\nu},\quad \text{and}\quad \left|\frac{p(0;s)-r_2(s)}{p(0;s)}\right|<Cs^{2\nu+1}. \tag{3.13}$$

This is our precise efficiency estimation of the correlation expansion of the NNS function (3.11): when applied to the standard RMT, the lowest-power term in the difference $p(0;s)-r_2(s)$ is proportional to $s^{3\nu+1}$, that is, very efficiently s^4 for GOE, s^7 for GUE, and s^{13} for GSE. It is expected to hold in the thermodynamic-limit situation. A general proof of this result is difficult, even for the restricted present approach. In Sec. IV, we show the proof adapted to $\nu=2$ explicitly. However, the general validity of the estimate in (3.13) is certain for the systems (2.15) and (2.16) by virtue of (3.12) and the cancellation of the normalization factor, $C_{N,\nu}$, in the relative estimation (3.13).

B. Long-range properties of R_2 function—The number variance asymptote

Level fluctuations can be characterized also in terms of the long-range correlation between two levels. Let the mean level spacing $\overline{s}$ be chosen as unity (this is the ''unfolding'' procedure: for those systems that admit the hole probability $E(s)$, $\overline{s}=1/\rho$ holds so that the two definitions are identical). Then, the short-range and long-range properties of the correlation function are referred to for $s\ll 1$ and $s\gg 1$, respectively, and we have seen that the level repulsion is a typical short-range property represented in the $r_2(s)$ function. In order to see the role of the same function in the long-range property, we first note the probabilistic meaning of the $r_2(s)$ in terms of the spacing distributions:

$$r_2(s)=\sum_{n=0}^{\infty} p(n;s) \quad \text{(the sum of all } n\text{-level spacing distributions)}, \tag{3.14}$$

which can be regarded as a converse relation to Eq. (3.11) [Eq. (5.1.39) in Ref. 5 must be a misprint].

Two commonly used concepts for the long-range statistics are the number variance Σ^2 and the averaged spectral rigidity Δ_3, which are defined, respectively, by

$$\begin{aligned}\Sigma^2(S)&\equiv\overline{(N-\bar{N})^2} \quad \text{(in the interval of length } S\text{)}\\ &=\int_{-S/2}^{S/2}\int_{-S/2}^{S/2} dxdy[r_1(x)\delta(x-y)+r_2(x,y)]-\left[\int_{-S/2}^{S/2} dx r_1(x)\right]^2\end{aligned} \tag{3.15}$$

and

$$\overline{\Delta(L)}=\min_{A,B}\frac{1}{2L}\int_{-L/2}^{L/2}\overline{(N(x)-Ax-B)^2}dx. \tag{3.16}$$

By using the unfolded scale, the two-level cluster function $1-r_2(r)$ is given by $Y_2(r)$: then, in terms of this function,

$$\Sigma^2(S)=S-S^2+2\int_0^S (S-r)r_2(r)dr=S-2\int_0^S (S-r)Y_2(r)dr. \tag{3.17}$$

An integral expression of the $\overline{\Delta}(L)$ in terms of the number variance function $\Sigma^2(S)$ was established by Pandey,[31] which reads as

$$\overline{\Delta}(L)=\frac{2}{L^4}\int_0^L (L^3-2L^2S+S^3)\Sigma^2(S)dS. \tag{3.18}$$

Bohigas *et al.*[29] pointed out that for a large class of quantum level sequences the asymptotic behavior of the number variance function $\Sigma^2(S)$ can be represented by the formula

$$\Sigma^2(S)\simeq AS+2B\ \log\ S+C, \tag{3.19a}$$

where

$$A=1-\int_{-\infty}^{\infty} Y_2(r)dr, \tag{3.19b}$$

$$B=\text{coefficient of the inverse power expansion},$$

$$Y_2(r)=\frac{B}{r^2}+O\left(\frac{1}{r^3}\right), \quad r\gg 1, \tag{3.19c}$$

and

$$C=2\int_0^1 dS\int_S^{\infty} drY_2(r)-2B. \tag{3.19d}$$

Thus, the representation (3.19) can be characterized by the total area enclosed by the Y_2 function [which must be unity, provided the sum rule (3.7) holds] and its falloff of the inverse-square type (to which the g-CM, also the g-CS statistical model in a limited range, belongs): it may be regarded as a superposition of the Poisson linear line and the Gaussian-ensemble logarithmic line in RMT, namely,

$$\Sigma^2(S)_{\text{Poisson}}=S, \quad \Sigma^2(S)_{\text{GE}}\simeq \frac{2}{\nu\pi^2}[\log(2\pi S)+\gamma_\nu]+O\left(\frac{1}{S}\right), \tag{3.20}$$

where γ_ν is a constant, depending on $\nu=1$ (GOE), 2 (GUE), and 4 (GSE). This interpretation indicates that the present model (2.15) or (2.16) provides a general long-range statistics, in agreement with the prediction by the independent superposition theory (cf. Ref. 20), for example, the formula for superposed level sequences of Seligman and Verbaarshot,[32]

$$\Sigma^2(S)=\mu_1 S+\sum_{i=2}^{k}\Sigma^2_{\text{RMT}}(\mu_i S), \tag{3.21}$$

where μ_1 is the fraction of the phase space volume of the regular region and μ_i $(i\geqslant 2)$ that of the ith chaotic region. It supports the asymptotic correctness of the Berry–Robnik picture[8] on the product hole probability $E_2(s)=\Pi E_2(\mu_i s)$. However, there exists an important reservation before drawing the conclusion: it concerns a violation of the compatibility relation (3.5) in the thermodynamic limit $N\to\infty$, or that of the sum rule (3.7), and will be discussed after presenting a concrete, exact model, namely Gaudin's model in 1966.

IV. EXACT ANALYSIS ON POISSON-CUE INTERMEDIATE STATISTICS

The one-parameter family of distributions for the g-CS system given in Eq. (2.16) can be put into a more detailed analysis, if the symmetry is restricted to $\nu=2$, i.e., one-parameter unitary ensembles. It was proposed and studied by Gaudin in 1966:[22] he considered the following N-level distribution:

$$P(\phi_1,\phi_2,\ldots,\phi_N)=C'_N\prod_{j<k}\left|\frac{\epsilon_j-\epsilon_k}{\epsilon_j-Z\epsilon_k}\right|^2, \quad \epsilon_j=e^{i2\pi\phi_j}, \quad 0\leqslant Z\leqslant 1. \tag{4.1}$$

From the present viewpoint, this model is identical to the (integral reduction of) most probable g-CS canonical distribution, expression (2.16) with $\nu=2$, since the above $P(\phi)$ can be rewritten as

$$P(\phi_1,\phi_2,\ldots,\phi_N)=C'_N\left(\frac{e^\alpha}{\sinh\,\alpha}\right)^{N(N-1)}\prod_{m<n}\left[\frac{4\,\sin^2\,\pi(\phi_m-\phi_n)}{1+\sinh^{-2}\,\alpha\,\sin^2\,\pi(\phi_m-\phi_n)}\right], \tag{4.2a}$$

where the free parameter α is related to that in Eq. (4.1) through

$$Z=e^{-2\alpha}, \quad (0<\alpha\leqslant\infty). \tag{4.2b}$$

Thus, putting

$$4\gamma/\beta=\sinh^{-2}\,\alpha, \quad \phi_n=\frac{x_n}{L}, \tag{4.2c}$$

together with $C_N=C'_N(e^\alpha/\sinh\,\alpha)^{N(N-1)}$, we complete the identification between both families. They give an interpolation between the two limiting situations, such that

$$\alpha\to 0\,(\gamma/\beta\to\infty), \quad \text{Poisson limit;} \tag{4.3a}$$

$$\alpha\to\infty\,(\gamma/\beta\to 0), \quad \text{CUE limit.} \tag{4.3b}$$

The normalization factor of the distribution (4.1) or (4.2), i.e., the partition function, can be computed exactly by the formula[22]

$$\int_0^1\cdots\int_0^1 d\phi_1\ldots d\phi_N\prod_{m<n}\frac{\sin^2\,\pi(\phi_m-\phi_n)}{\sinh^2\,\alpha+\sin^2\,\pi(\phi_m-\phi_n)}=N!\,e^{-N(N-1)\alpha}\prod_{k=1}^{N}\frac{1-e^{-2\alpha}}{1-e^{-2k\alpha}}. \tag{4.4a}$$

After a more complicated analysis, Gaudin deduced an exact formula for the two-level correlation function $R_2(\phi_1,\phi_2)=R_2(\phi_1-\phi_2)$:

$$R_2(\phi_1-\phi_2)=N(N-1)-2\,\mathrm{Re}\sum_{\lambda<\mu}\prod_{k=\lambda}^{\mu-1}\left[1+\frac{e^{2\pi(s/L)}-1}{1-e^{-2\alpha k}}\right]^{-1},\quad \phi_1-\phi_2\equiv\frac{s}{L},$$

$$R_2(s)/N^2=r_2(s)\to 1-2\,\mathrm{Re}\int_{0<x<y<1}dxdy\,\exp\left[-2\pi\rho s\int_x^y d\xi\frac{1}{1-e^{-2\pi\rho a\xi}}\right],$$

where the parameter α is now replaced by

$$a=\alpha L/\pi,\quad \text{or}\quad \rho a=\alpha N/\pi. \tag{4.4b}$$

Therefore, the correlation function with an arbitrary density parameter, ρ, is given by

$$r_2(s)=1-\left|\int_0^1 dx\,\exp\left[i\frac{s}{a}\log(e^{2\pi a\rho x}-1)\right]\right|^2=1-Y_2(r), \tag{4.5}$$

where

$$Y_2(r)=\left|\int_0^1 dx\,\exp\left[i\frac{r}{a}\log(e^{2\pi\rho a x}-1)\right]\right|^2=\left(\frac{1}{2\pi\rho a}\right)^2\left|\int_\nu^\infty\frac{e^{-i\omega r/a}}{e^\omega+1}d\omega\right|^2,\quad \nu=-\log(e^{2\pi\rho a}-1), \tag{4.6}$$

on change of the integration variable, $x\to\omega=-\log(e^{2\pi\rho a x}-1)$.

On the other hand, Forrester[23] reconsidered Gaudin's model as the grand canonical ensemble of the g-CM system, which may be viewed as the proper thermodynamic limit of the g-CS. Thus, the results of Gaudin and Forrester are summarized (notations by Gaudin) as follows:

(a) *equation of state for grand canonical ensemble with fugacity* ζ,

$$2p=\int_0^\infty\log(1+\zeta e^{-2\pi a x})dx\quad\text{and}\quad\rho=\frac{1}{2\pi a}\log(1+\zeta); \tag{4.7}$$

(b) *hole probability* $E(s)$,

$$E(s)=\det|\mathbf{I}-\rho\mathbf{G}|=\sum_{p=0}^\infty\frac{(-1)^p}{p!}\int_{(-s/2,s/2)}\cdots\int\rho^p ds_1\cdots ds_p\,\det|g(s_j-s_k)|_p, \tag{4.8a}$$

where

$$g(s)=\frac{1}{2\pi\rho a}\int_\nu^\infty\frac{d\omega}{e^\omega+1}e^{-i\omega s/a}, \tag{4.8b}$$

$g(x-y)$ being the kernel of the integral operator $\mathbf{G}$ on a function space in $(-s/2,s/2)$.

Hereafter in this section (before a discussion of thermodynamics), we assume the unfolded scale:

$$\rho=1\quad(\overline{s}=1). \tag{4.9}$$

As demonstrated by Gaudin, one gets two simplified situations:

$$(i)\quad a\to 0,\quad Y_2(r)\cong\frac{a^2}{a^2+r^2}, \tag{4.10}$$

$$(ii)\quad a\to\infty,\quad Y_2(r)\cong\left(\frac{\sin\pi r}{\pi r}\right)^2. \tag{4.11}$$

Expressions (4.10) and (4.11) are deduced by replacing the integrand in the expression (4.6); $(e^{\omega}+1)^{-1}\to e^{-\omega}$, and $\to 1$, the integration range being $[\nu,\infty]$ and $[\nu(<0),0]$, respectively.[22] Here, we demonstrate the interpolation numerically and graphically by defining the following function:[33]

$$P(x,y;a)=\frac{1}{2a}\log\frac{e^{2\pi ax}-1}{e^{2\pi ay}-1}. \tag{4.12}$$

We can show, after a few manipulations, that both the short-range and the long-range behaviors of the two-level correlation function may be exhibited in terms of this function as follows:

$$r_2(s;a)|_{s\to 0}=A(a)s^2,\quad A(a)=2\int_0^1\int_0^1 dxdy\{P(x,y)\}^2 \tag{4.13}$$

and

$$\Sigma^2(s;a)=s-\int_0^1\int_0^1 dxdy\left\{\frac{\sin sP(x,y;a)}{P(x,y;a)}\right\}^2. \tag{4.14}$$

The parameter dependence of the resulting number variance and spectral rigidity (the Δ_3 function) is exhibited in Fig. 1. Result (4.13) shows that the repulsion is always (except $a=0$) of the GUE type (square dependence): For $a\to 0$ (Poisson limit), $A(a)$ diverges, and for $a\to\infty$ (CUE limit), $A(a)\to\pi^2/3$ [in agreement with $P_{\mathrm{GUE}}(s)$].

The above short-range property agrees with the prediction by Forrester,[23] who obtained the same square dependence of the nearest-neighbor spacing distribution $p(0;s)$.

Here, we add the validity of inequalities (3.12) and (3.13) in the power series expansion (3.11) explicitly for the present model by means of the third-order correlation function,[22,23]

$$r_p(s_1,\ldots,s_p)=\det|g(s_i-s_j)|_p,\quad p=3. \tag{4.15}$$

By the power series expansion of $r_3(s_1,s_2,s_3)$ with respect to two variables, $s\equiv s_1-s_2$ and $t\equiv s_1-s_3$ (hence $t-s=s_1-s_3$), one finds that the series starts with the lowest-power terms $s^2t^4+s^4t^2$ and s^3t^3, affirming (3.12) ($\nu=2$, $n=3$); also (3.13) by using $p(0;s)\propto s^2$.

In order to analyze Fig. 1, specifically, the asymptotic behavior of the number variance curve, we first note that the formulas by Bohigas *et al.* (3.19b)–(3.19d) hold here, because, first the cluster function Y_2 given by Eq. (4.6) satisfies

$$\int_{-\infty}^{\infty}Y_2(r)dr=\frac{1}{2\pi a}\int_{\nu}^{\infty}\frac{d\omega}{(e^{\omega}+1)^2}=1-\frac{1-e^{-2\pi a}}{2\pi a},\quad \nu=-\log(e^{2\pi\rho a}-1), \tag{4.16}$$

and second for the integral in (4.6), by means of an integration by parts,

$$\frac{1}{2\pi a}\int_{\nu}^{\infty}\frac{e^{-i\omega r/a}}{e^{\omega}+1}d\omega=\frac{e^{-i\nu r/a}}{2\pi ir}\left[\frac{1}{e^{\nu}+1}+e^{i\nu r/a}\int_{\nu}^{\infty}\left(\frac{1}{(e^{\omega}+1)(1+e^{-\omega})}\right)e^{-i\omega r/a}d\omega\right].$$

Hence

$$\left|\frac{1}{2\pi a}\int_{\nu}^{\infty}\frac{e^{-i\omega r/a}}{e^{\omega}+1}d\omega\right|^2=\frac{1}{(2\pi r)^2}\left|\frac{1}{e^{\nu}+1}+e^{i\nu r/a}C\right|^2, \tag{4.17a}$$

where C denotes the integral in the square bracket above, which can be estimated as

$$|C|=\left|\int_{\nu}^{\infty}\frac{e^{-i\omega r/a}d\omega}{(e^{\omega}+1)(1+e^{-\omega})}\right|\leqslant\int_{\nu}^{\infty}\frac{d\omega}{(e^{\omega}+1)(1+e^{-\omega})}=\frac{1}{e^{\nu}+1}=1-e^{-2\pi a}. \tag{4.17b}$$

We therefore have, for the number-variance asymptote,

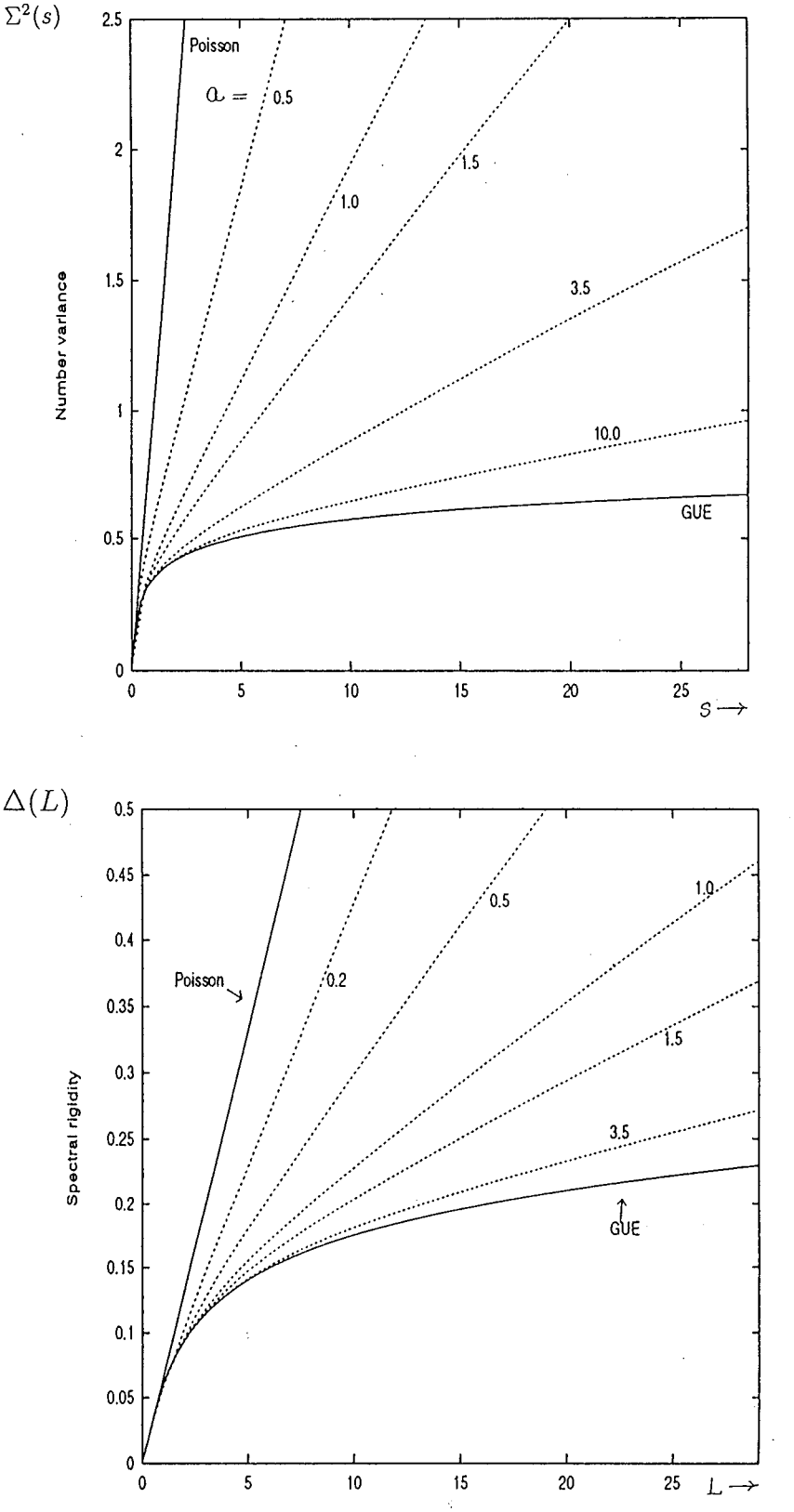

FIG. 1. Curves of number-variance $\Sigma^2(s)$ and Δ statistics $\Delta(L)$ for the exact Gaudin model (Refs. 22 and 33) of Poisson-CU ensembles.

$$\Sigma^2(S) \xrightarrow{S\to\infty} \frac{1-e^{-2\pi a}}{2\pi a}S+\frac{(1-e^{-2\pi a})^2+|C(a)|^2}{2\pi^2}(\log S+\text{const.}), \quad |C(a)|\leqslant 1-e^{-2\pi a}. \tag{4.18}$$

Result (4.16) indicates that the compatibility relation (3.7) (the sum rule) does not hold in the thermodynamic limit. The equation of state (4.7) provides its understanding from a viewpoint of compressible level gas: the presence of the linear term in the $\Sigma^2(S)$ asymptote above can be related to the isothermal compressibility of the one-dimensional interacting model gas as follows:

$$\frac{\partial}{\partial\rho}(\beta p)=\frac{d\zeta}{d\rho}\frac{\partial}{\partial\zeta}(\beta p)=\frac{2\pi a\rho}{1-e^{-2\pi a\rho}}, \tag{4.19}$$

$$\Sigma^2(S)_{\text{linear in } S}=\left(\frac{1}{\beta}\frac{d\rho}{dp}\right)S, \quad \text{with } \beta=2. \tag{4.20}$$

It corresponds to the Ornstein–Zernike formula[27] for the number variance of a three-dimensional gas in terms of the compressibility, as we discuss in more detail in Sec. V.

V. VIOLATION OF THE SUM RULE AND ABSENCE OF THE FRACTIONAL POWER LAW

The Ornstein–Zernike formula for the number variance of gas molecules in a big vessel of volume V reads[27] as

$$\langle(N-\langle N\rangle)^2\rangle=\langle N\rangle^2\frac{kT}{v}\kappa_T, \quad v\ll V, \tag{5.1}$$

where κ_T is the isothermal compressibility of the gas, given by

$$\kappa_T=-\frac{1}{v}\left(\frac{\partial\rho}{\partial p}\right)_{T,\langle N\rangle}=\frac{v}{\langle N\rangle}\left(\frac{\partial\rho}{\partial p}\right)_{T,\langle N\rangle} \quad \left(\rho=\frac{\langle N\rangle}{v}\right). \tag{5.2}$$

Hence, $\langle(N-\langle N\rangle)^2\rangle=kT(\partial\rho/\partial p)_{T,\langle N\rangle}\langle N\rangle$ holds.

For a one-dimensional level gas, it gives precisely the Poisson linear part of the $\Sigma^2(S)$ function, since $\langle N\rangle=S$, thus

$$\Sigma^2(S)_{\text{linear in } S}=\kappa S, \quad 0\leqslant\kappa\leqslant 1, \tag{5.3a}$$

where the compressibility κ now plays a role of the interpolation parameter:

$$\kappa=1: \quad \text{pure Poisson line} \quad \text{(the gas is fully compressible)}, \tag{5.3b}$$

$$\kappa=0: \quad \text{the Gaussian chaos} \quad \text{(the gas is fully incompressible)}. \tag{5.3c}$$

A contribution to the $\Sigma^2(S)$ function from the latter incompressible component is represented by a weaker logarithmic line (3.20) that originates from the inverse-square falloff of the $Y_2(r)$ function.

The feature summarized above is a necessary consequence of the violation of the sum rule (3.7), because as we have seen in the preceding sections the compressibility κ yields a measure of such a violation: $\kappa=1-\int_{-\infty}^{\infty}Y_2(r)dr$. It should be remarked that the violation $\kappa>0$ has resulted from the natural thermodynamic limit,

$$N\to\infty, \quad \text{and} \quad L\to\infty, \quad \text{with } \frac{N}{L}=\rho=\text{fixed}, \tag{5.4}$$

because for a finite g-CS level gas the integral identity must hold strictly in the form (3.5). This implies that for general compressible gases the integration and the thermodynamic limit $L\to\infty$ cannot be interchanged. This was first discussed by Mehta and Dyson[28] in early RMT without any quantitative conclusion: under the circumstance, one should adopt

$$\int_{-\infty}^{\infty} \lim_{N\to\infty}(N/L=\text{fixed}), \quad \text{rather than} \quad \lim_{N\to\infty}\int_{-L/2}^{L/2}, \tag{5.5}$$

which causes the nonvanishing linear component of the $\Sigma^2(S)$, (5.3).

One should not hasten to draw a conclusion that the above presentation be the only possible intermediate long-range statistics. There has been a prediction of a new one-parameter dependence in the Σ^2 function of the universal power type; $\Sigma^2(S)\propto S^{\gamma}$; $1>\gamma>0$, which we may call a ''fractional power law.'' The physical idea is to characterize a transition such as between insulating and metallic phases at the mobility edge,[24] and has been deduced based on the strict satisfaction of the sum rule (3.7). This behavior is certainly weaker than the Poisson straight line but weaker than the Gaussian logarithm, yielding another smooth interpolation between the two limits, as we demonstrated in Ref. 34) (cf. Ref. 20). Whether or not the fractional power law can be understood by the level dynamics is open.

VI. EXACT FORM FACTOR AND ITS APPLICATIONS

The two-level form factor $b(k)$ is defined as the Fourier transform of the (thermodynamic limit of) two-level cluster function $Y_2(x)$ [$n=2$ in (3.6b)], thus (without the unfolding, for generality)

$$b(k)=\int_{-\infty}^{\infty} Y_2(x)e^{ikx}\rho dx. \tag{6.1}$$

Gaudin's exact expression of this quantity for his linear interacting gas model (4.6) reads as

$$\begin{aligned} b(k)&=\frac{1}{2\pi a\rho}\int_{\nu}^{\infty}dx\int_{\nu}^{\infty}dx'\,\frac{\delta(x-x'-ka)}{(e^{x}+1)(e^{x'}+1)} \\ &=\frac{1}{2\sinh(|k|a/2)}\left(\frac{e^{|k|a/2}}{2\pi a\rho}\log[1+e^{-|k|a}(e^{2\pi a\rho}-1)]-e^{-|k|a/2}\right), \end{aligned} \tag{6.2}$$

where

$$b(0)=\lim_{k\to 0} b(k)=1-\frac{1}{2\pi a\rho}(1-e^{-2\pi a\rho}), \tag{6.3}$$

is in agreement with the preceding result (4.16) (if $\rho=1$).

Given a real symmetric, continuous function $f(x)$, $x\in\mathbf{R}$, which satisfies $\int_{-\infty}^{\infty}|f(x)|dx<\infty$. Then, the average of a quantity $(1/N)\Sigma_{m\neq n}f(x_m-x_n)$ over the reduced g-CS distribution (2.16), $\nu=2$, can be computed in terms of the Fourier transform of the f function $\mathscr{F}_f(k)$ and the two-level form factor $b(k)$ as follows:

$$\lim_{N\to\infty,\rho fixed}\left\langle\frac{1}{N}\sum_{m\neq n}f(x_m-x_n)\right\rangle=\rho\mathscr{F}_f(0)-\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathscr{F}_f(k)b(k)dk,$$

$$\mathscr{F}_f(k)=\int_{-\infty}^{\infty}f(k)e^{ikx}dx. \tag{6.4}$$

As an example, we show the average of the potential function, $V(x-y)$, which is identified by rewriting the distribution (2.16) in the form $\exp[-\beta\Sigma_{m<n}V(x_m-x_n)]$: in the thermodynamic limit one finds that $V(s)=\frac{1}{2}\log(1+a^2/s^2)$, which is consistent with (4.10). Then,

$$\mathscr{F}_V(k)=\frac{1}{2}\int_{-\infty}^{\infty}\log(1+a^2/x^2)e^{ikx}\,dx=\frac{\pi(1-e^{-a|k|})}{|k|},$$

and the average potential energy per particle (level) is now given by

$$\left\langle\frac{1}{N}U\right\rangle=\frac{1}{2}\left(\pi a\rho-\int_0^{\infty}\frac{dk}{k}\left\{\frac{1}{2\pi a\rho}\log[1+(e^{2\pi a\rho}-1)e^{-k}]-e^{-k}\right\}\right). \tag{6.5}$$

A. Information loss per level of a circular unitary ensemble

Dyson[35] obtained an exact result on the statistical thermodynamics of eigenvalue sequences of a randomly chosen unitary matrix; what he named *circular ensembles*—orthogonal ($\nu=1$), unitary ($\nu=2$), and symplectic ($\nu=4$). The ordinary context of statistical mechanics, i.e., free energy and mean energy, yielded the entropy per single eigenvalue lying on a unit circle; thus,

$$\begin{aligned}\frac{S}{N}&=\frac{1}{2}(1-\gamma-\log\,\pi)=-0.361,\quad \nu=1,\\ &=-\gamma=-0.577,\quad \nu=2,\\ &=1-2\gamma-\log 2=-0.848,\quad \nu=4\\ &\quad(\gamma=\text{Euler's constant}).\end{aligned} \tag{6.6}$$

Dyson interpreted these numbers as a decrease of the degree of irregularity in the three kinds of eigenvalue sequences, $\nu=1$, 2, and 4, respectively: the negativity of the numbers indicates that each sequence carries an information content in comparison with the entirely random sequence (a negative information loss or, rather, an information gain) depending on the structure of the matrix symmetry: the larger the number ν, the less irregular the sequence.

The present status having the exact formula of the internal energy for Gaudin's model enables us to compute this information quantity as a continuous function of the parameter a from 0 to infinity. The free energy can be computed from the exact partition function (4.4a) with a change of the parameter (4.4b):

$$-\frac{\beta}{N}F=\frac{1}{N}\log N!-(N-1)\alpha+\log(1-e^{-2\alpha})-\frac{1}{N}\sum_{k=1}^{N}\log(1-e^{-2k\alpha}),\quad \beta=2.$$

For $N\gg 1$,

$$\begin{aligned}&=-\pi a\rho-1+\log(2\pi a\rho)-\frac{1}{\pi a\rho}\int_0^{\pi a\rho}\log(1-e^{-2u})du,\\ &=-\pi a\rho-1+\log\frac{2\pi a\rho}{1-e^{-2\pi a\rho}}+\frac{1}{\pi a\rho}\int_0^{\pi a\rho}\frac{2u\,du}{e^{2u}-1}.\end{aligned}$$

Hence, we get

$$\begin{aligned}\frac{1}{N}S=\frac{\beta}{N}(U-F)&=\log\left(\frac{2\pi a\rho}{1-e^{-2\pi a\rho}}\right)-1+\frac{1}{\pi a\rho}\int_0^{\pi a\rho}\frac{2u\,du}{e^{2u}-1}\\ &\quad-\int_0^{\infty}\left(\frac{1}{2\pi a\rho}\log[1+(e^{2\pi a\rho}-1)e^{-u}]-e^{-u}\right)\frac{du}{u}.\end{aligned} \tag{6.7}$$

It is not difficult to check the validity of the expected limits: $S\to 0$ for $a\to 0$, and $S\to -N\gamma$ for $a\to\infty$. The computed curve of this quantity versus a is shown in Fig. 2.

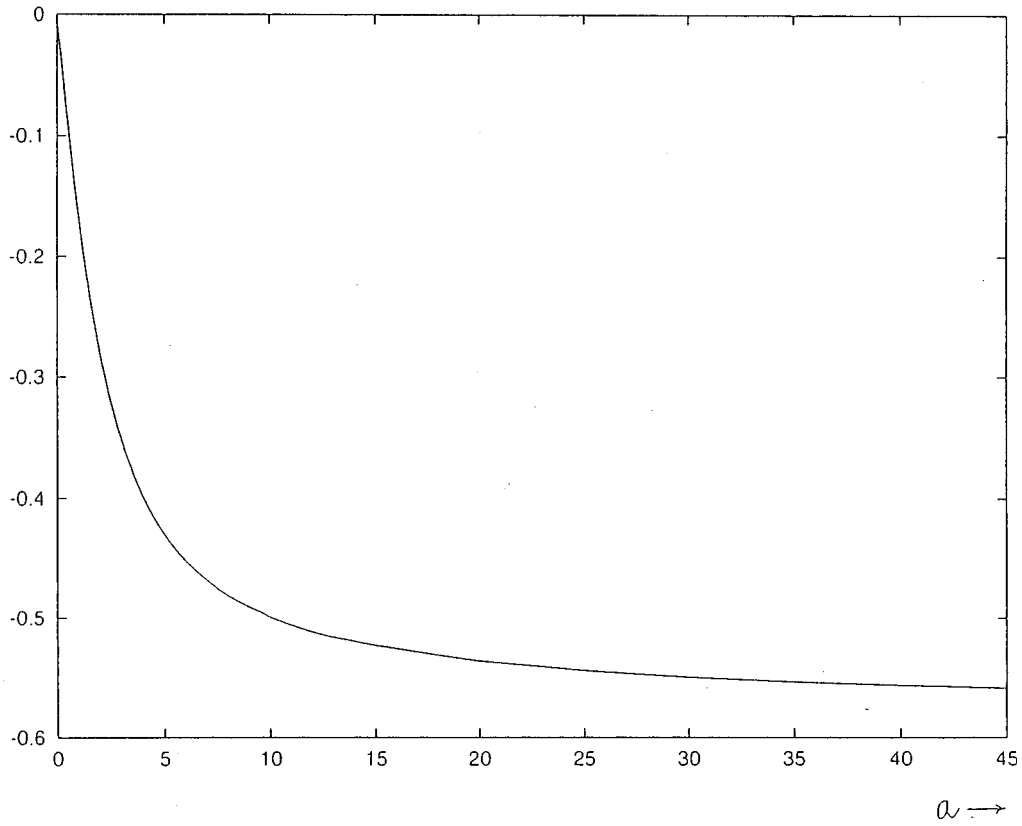

FIG. 2. Information-loss curve for the Gaudin model of Poisson-CU ensembles in accordance with Dyson (Ref. 35) computed by the formula of $b(k)$, Eq. (6.2).

B. Correlation hole of survival probability

We study a manifestation of level statistics on the so-called survival probability function (SPF), defined by

$$P(t)=|\langle\psi(0)|\psi(t)\rangle|^2, \tag{6.8}$$

i.e., the quantum probability that an initial state will return to itself after time t. It has been shown that the SPF, if averaged both over random initial states and over the Hamiltonian ensemble, exhibits a distinct difference between chaotic and regular systems. For integrable systems, the SPF decays monotonically to its long-time limit, while for chaotic systems the SPF curve exhibits a ''correlation hole,'' where the probability falls below its asymptotic value at finite times. These features can be explained in the framework of level statistics, as discussed by a number of authors.[36,37] Therefore, we investigate the intermediate situation of this phenomenon, where the present intermediate statistics is applicable, in particular, Gaudin's exact formula for intermediate circular unitary ensembles.

In Ref. 37, it was shown that the double-averaged SPF can be expressed, in terms of the two-level form factor $b(k)$ (6.2), as

$$\langle\langle P(t)\rangle\rangle=\frac{2}{N+1}+\frac{1}{N+1}\left[\frac{\sin^2(\pi\tau N)}{\pi^2\tau^2 N}-\int_{-\infty}^{\infty}dk\,b(k)\frac{\sin^2[\pi(\tau-k)N]}{\pi^2(\tau-k)^2 N}\right]. \tag{6.9}$$

Substituting the expression given by Eq. (6.2) into (6.9), we can plot the SPF, as shown in Fig. 3 for a fixed value of N $(=14)$ and four different values of the parameter a. One can see a smooth transition from the Poisson case, i.e., no correlation hole, to the GUE case, where the depth of the hole reaches its maximum. In the course of this transition, the area of the correlation hole is kept unchanged, while its depth increases as a increases. These are the manifestation of the level

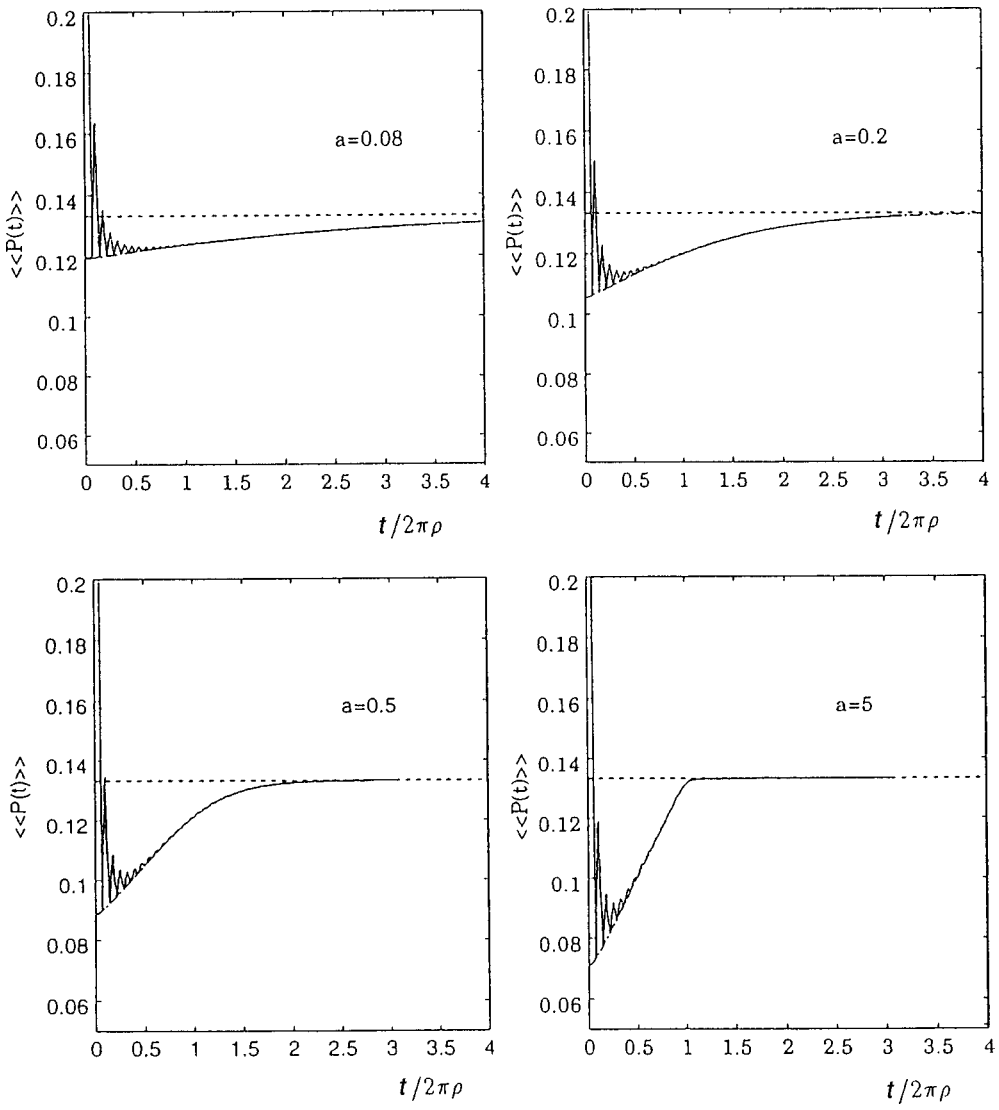

FIG. 3. Curves of survival probability for a simple quantum system without time reversal belonging to the Gaudin model (Ref. 37) of Poisson-CU ensembles.

repulsion and, in fact, due to an increase of the incompressibility of the level gas from the Poisson limit. Here, let us study another manifestation of the problem of the ''sum rule'' in the preceeding sections in the present Fourier transformed space. The depth H of the hole of SPF is given from (6.9) by

$$H(N)=\frac{1}{N+1}\left[\int_{-\infty}^{\infty}\mathrm{d}k b(k)\,\frac{\sin^2[\pi(\tau-k)N]}{\pi^2(\tau-k)^2N}\right]_{\tau=0}. \tag{6.10}$$

In deriving this expression, we have assumed that the bottom of the hole occurs close to $\tau=0$. Observe that a similar problem to that encountered in the asymptotic expansion of the number variance arises, namely, the noninterchangeability of the thermodynamic limit and other limits: If we take $\tau=0$, i.e., $t=0$ first, then we would obtain an identity for the SPF, given by

$$\langle\langle P(t=0)\rangle\rangle=1, \tag{6.11}$$

which corresponds to the exact sum rule. On the other hand, if we take the limit $N\to\infty$ first in the square bracket of Eq. (6.10), we would obtain

$$H(N)=\frac{1}{N+1}b(0) \tag{6.12}$$

which depends on the extent of the incompressibility of the level gas. Note that the conservation of the total area S of the hole is due to the identity

$$S=\frac{1}{N+1}\int_0^\infty b(\tau)d\tau=\frac{1}{2}Y_2(0). \tag{6.13}$$

Finally, we give a brief description about our ongoing investigation of the relation between the survival probability and the saturation of the long-range statistics, such as the spectral rigidity and the number variance. It was analytically shown by Berry[6] and later numerically and experimentally tested by various authors[38] that the spectral rigidity reaches a saturation value at a nonuniversal point $L_{\max}$, in contrast to the logarismic or the linear law, as predicted by the random matrix theory of treating infinitely many levels, for regular or chaotic systems, respectively. The value of $L_{\max}$ is determined by the shortest period of the corresponding classical periodic orbits. We find that in order to take into account the saturation effect the two-level form factor should be modified as

$$b(k)=1, \quad \text{if } k<k_{\text{sat}}, \tag{6.14}$$

and otherwise the same as given by the corresponding RMT model. Here, $k_{\text{sat}}\sim 1/L_{\max}$ is a nonuniversal quantity given by the classical feature of the system under consideration. Therefore, the manifestation of the saturation effect on the survival probability should be that the bottom of the correlation hole is of *finite* size, k_{sat}. The remaining question is whether or not the function $b(k)$ is discontinuous near the point k_{sat}, which must be treated separately.

VII. CONCLUDING REMARK

An important comment on the present work, which we consider as worthwhile to address in concluding it, is a possible unification of our framework of level dynamics and that of band random matrix theories (BRMT).[25,39] The common motivation in both frameworks is, being very close to each other, to take systematic account of the *structuredness* of level correlations such that a more distant pair of levels be less correlated. This nature is more pronounced in BRMT, since it starts with a straightforward destruction of the full permutation symmetry of levels (a kind of symmetry breaking of the symmetry group that is retained in our framework). It is not strange, therefore, that BRMT may be able to predict the metal–insulator transition with a fractional power law of the number variance asymptote, which should be studied.

[1] Historical surveys: for before 1980, T. A. Brody, J. Flores, J. B. French, P. A. Mello, A. Pandey, and S. S. Wong, Rev. Mod. Phys. **53**, 385 (1981); for after 1980, O. Bohigas and M.-J. Giannoni, Lect. Notes Phys. **209**, 1 (1984).
[2] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer, New York, 1990).
[3] A. D. Stone, P. A. Mello, K. A. Muttalib, and J.-L. Pichard, in *Mesoscopic Phenomena in Solids*, edited by B. L. Altshler, P. A. Lee, and R. A. Webb (North-Holland, Amsterdam, 1991).
[4] O. Bohigas, M.-J. Giannoni, and C. Schmit, Phys. Rev. Lett. **25**, 1 (1984).
[5] M. L. Mehta, *Random Matrices* (Academic, New York, 1991).
[6] M. V. Berry, Proc. R. Soc. London, Ser. A **400**, 229 (1985).
[7] O. Bohigas, S. Tomsovic, and D. Ullmo, Phys. Rep. **223**, 43 (1993).
[8] M. V. Berry and M. Robnik, J. Phys. A **17**, 2413 (1984); M. Robnik, in *Chaos, Solitons and Fractals*, edited by K. Nakamura and M. S. Elnaschie (Pergamon, New York, 1995).
[9] F. Haake, *Quantum Signature of Chaos* (Springer, Berlin, 1991).
[10] P. Pechukas, Phys. Rev. Lett. **51**, 943 (1983).
[11] F. J. Dyson, J. Math. Phys. **3**, 1191 (1962).
[12] T. Yukawa, Phys. Rev. Lett. **54**, 1883 (1985); Phys. Lett. A **116**, 227 (1986).
[13] K. Nakamura and M. Lakshmanan, Phys. Rev. Lett. **57**, 1661 (1986).
[14] K. Nakamura and H. J. Mikeska, Phys. Rev. A **35**, 5294 (1987).
[15] P. Gaspard, S. A. Rice, H. J. Mikeska, and K. Nakamura, Phys. Rev. A **42**, 4015 (1990).
[16] F. Calogero, J. Math. Phys. **10**, 2197 (1969).
[17] B. Sutherland, J. Math. Phys. **12**, 246 (1971); **12**, 251 (1971); Phys. Rev. A **4**, 2019 (1971).
[18] J.-Z. Ma and H. Hasegawa, Z. Phys. B **93**, 529 (1994).
[19] H. Hasegawa, ''*Dynamical formulation of quantum level statistics*,'' *Open Systems and Information Dynamics*, Proceed-

ings of the 3rd International Summer School, ''Let's face chaos through nonlinear dynamics,'' edited by M. Robnik (Kluwer, Dordrecht, 1997), Vol. 4, p. 359; H. Hasegawa, *Complete-Integrable Calogero Dynamics for Random Matrix Theories*, Proceedings of the 8th International Colloquim on Differential Equations, edited by D. Bainov (in press); cf. H. Hasegawa and J.-Z. Ma, J. Phys. Soc. Jpn. **63** (Suppl. A), 91 (1994).
[20] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, preprint, July, 1997.
[21] R. Balian, Nuovo Cimento B **57**, 183 (1968).
[22] M. Gaudin, Nucl. Phys. **85**, 545 (1966).
[23] P. J. Forrester, Phys. Lett. A **173**, 355 (1993).
[24] V. E. Kravtsov, I. V. Lerner, B. L. Altshuler, and A. G. Aronov, Phys. Rev. Lett. **72**, 888 (1994).
[25] F. M. Izrailev, *Structured Random Matrices and Their Applications*, Proceedings of the International Conference on Dynamical Systems and Chaos, edited by Y. Aizawa, S. Saito, and K. Shiraiwa (World Scientific, Singapore, 1995), Vol. 2, p. 497.
[26] F. J. Dyson, J. Math. Phys. **3**, 166 (1962).
[27] M. Toda, R. Kubo, and N. Saito, *Statistical Physics* (Springer, Berlin, 1992), Vol. I.
[28] M. L. Mehta and F. J. Dyson, J. Math. Phys. **4**, 713 (1963).
[29] O. Bohigas, R. U. Haq, and A. Pandy, Phys. Rev. Lett. **54**, 1645 (1985).
[30] E. P. Wigner, ''Distribution laws for the roots of a random Hermitian matrix,'' in *Statistical Theories of Spectra*, edited by C. E. Porter (Academic, New York, 1965).
[31] A. Pandey, Ann. Phys. (N.Y.) **119**, 170 (1979).
[32] T. H. Seligman and J. J. M. Verbaarshot, J. Phys. A **18**, 2227 (1985).
[33] J.-Z. Ma and H. Hasegawa, J. Phys. Soc. Jpn. **64**, 2261 (1995).
[34] H. Hasegawa, J.-Z. Ma, and T. Takami, Surf. Rev. Lett. **3**, 13 (1996).
[35] F. J. Dyson, J. Math. Phys. **3**, 140 (1963).
[36] L. Leviander, M. Lombardi, R. Jost, and J. L. Kinsey, Phys. Rev. Lett. **56**, 475 (1987); T. Guhr and H. A. Weidenmüller, Chem. Phys. **146**, 21 (1990); J. Wilkie and P. Brumer, Phys. Rev. Lett. **67**, 1185 (1991); Y. Alhassid and R. D. Levine, Phys. Rev. A **46**, 4650 (1992); Y. Alhassid and N. Whelan, Phys. Rev. Lett. **70**, 572 (1993).
[37] J.-Z. Ma, J. Phys. Soc. Jpn. **64**, 4059 (1995).
[38] See, e.g., H. Alt *et al.*, chao-dyn/9706025, and references therein.
[39] G. Casati, in *Proceedings of the 5th Wigner Symposium*, Vienna, August, 1997, edited by P. Kasperkovitz *et al.* (World Scientific, Singapore), to be published.

Path integral approach to noncommutative space–times

Gianpiero Mangano[a)]
INFN, Sezione di Napoli, and Dipartimento di Scienze Fisiche,
Università di Napoli Federico II, Italy

(Received 21 November 1997; accepted for publication 29 December 1997)

We propose a path integral formulation of noncommutative generalizations of space–time manifold in even dimensions, characterized by a length scale λ_P. The commutative case is obtaind in the limit $\lambda_P=0$. © *1998 American Institute of Physics.* [S0022-2488(98)02904-1]

I. INTRODUCTION

It is commonly believed that the picture of space–time as a smooth pseudo-Riemannian manifold should break down at most at the Planck scale $\lambda_P \sim 10^{-33}$ cm, due to the effects of quantum gravity. Many arguments on operational limits on position and time measurements have been considered in literature,[1–4] suggesting that our description of space–time as a collection of points equipped with suitable topological and metric structures, should be modified in a way which is reminiscent of the quantization of phase space in ordinary quantum mechanics. In this case too, in fact, the classical phase space, seen as the joint spectrum of commuting position and momentum operators, turns into its fuzzy quantum version, where the notion of points is replaced, in a somehow intuitive way, by fundamental cells of area $\hbar$. This quantization procedure can be formalized by adopting a *dual* point of view, namely by considering the algebra of smooth functions over the phase space, generated by positions and momenta. This algebra contains, using Gelf'and–Naimark reconstruction theorem, all informations on the underlying space. Switching on the Planck constant then amounts to considering the noncommutative algebra still generated by positions and momenta, which are now considered noncommuting elements with usual commutation relations.

Quantum phase space is perhaps the most famous example of what is now usually called noncommutative geometry. While at the classical level the phase space can be described either *geometrically*, i.e., defining its structure as a topological manifold, or algebraically, its noncommutative quantum version can only be defined via the algebra of observables generated by positions and momenta, since the notion of ordinary manifold is lost. The algebraic description of noncommutative geometries has recently received new insights, in particular due to Connes.[5]

In this paper we will assume, as first considered long ago,[6] and as suggested by the previous considerations, that the effect of gravity at very short distances is such to render positions and time X^μ noncommutative,

$$[X^\mu, X^\nu] = i\lambda_P^2 Q^{\mu\nu}, \tag{1}$$

where the antisymmetric tensor $Q^{\mu\nu}$, depending in general on X^μ, should eventually be given by a satisfactory theory of quantum gravity. The space–time is described by the $*$-algebra $\mathscr{A}$ generated by regular representations of Eq. (1), whereas the notion of points is now lost. Here and in the following we will use upper case for the elements of $\mathscr{A}$ and lower case for their commutative limit, so for example, $F(X)$ is an element of $\mathscr{A}$ and $f(x)$ is instead an ordinary smooth function over the commutative space–time manifold. Many ansatzs for Eq. (1) have been considered, in different contexts.[3,4,7,8] They predict uncertainty relations among position and time operators which in turn imply that quantum field theories built upon these generalized space–times have a more regular behavior at small distances, and, in particular, may be free of ultraviolet divergences. See for example Refs. 3, 8–10.

[a)]Electronic mail: mangano@na.infn.it

 © 1998 American Institute of Physics

The key observation of the present paper is that the analogy with ordinary quantum mechanics and quantization of phase space can also be pursued to establish a path integral formulation of these noncommutative geometries, taking the point of view that a class of linear functionals over $\mathcal{A}$, which turn into evaluation maps in the commutative limit, can be expressed as integrals, with a suitable measure, of ordinary functions over the classical, commutative space–time manifold. It is well known in fact that quantization of phase space can be introduced either in the canonical approach, namely, by defining a representation for positions and momenta as self-adjoint operators on a separable Hilbert space, or via Feynman sum over paths in classical phase space. In this case, to evaluate matrix elements of time ordered polynomials in position and momenta one should average over all trajectories with fixed boundary conditions, $q(t_0)=q_0$ and $q(t_f)=q_f$, with measure $\exp(iS/\hbar)$. For one degree of freedom

$$\langle q_f,t_f|T[Q(t_1)P(t_2)\ldots]|q_0,t_0\rangle=\int D[q]D[p]e^{iS/\hbar}(q(t_1)p(t_2)\ldots), \tag{2}$$

where T denotes time ordering and the action S is the integral over the considered path of the one-form $pdq-H(p,q)dt$. Equation (2) defines a set of linear functionals over the algebra generated by Q and P in the Heisenberg representation. In particular for $q_0=q_f$, of the two terms in S, the first one measures for each closed trajectory γ in the phase space the area of the surface Γ with boundary γ normalized to $\hbar$, i.e., the surface integral of the symplectic form $dp\wedge dq$ over Γ. Actually, the picture of quantum phase space as composed by fundamental cells of area $\hbar$ is already emerging from this term, independently of which particular dynamical system we are considering, defined by a particular choice for the Hamilton function $H(p,q)$.

These considerations suggest that a noncommutative structure of space–time could be introduced starting with the classical commutative manifold equipped with a symplectic form Ω. A path integral formalism similar to Eq. (2) may then provide a way of evaluating a class of linear functionals over polynomials of the deformed noncommutative position and time operators X^μ. Of course this will restrict our analysis to the case of even dimensional spacetime manifolds M^{2n}, in order that a nondegenerate, closed two-form Ω exists. This construction requires the introduction of a parameter with dimension of length, λ_P, which plays a role analogous of $\hbar$ in ordinary quantum mechanics. In particular in the limit $\lambda_P\to 0$ all functionals we will consider reduce to evaluation maps of the commutative algebra of smooth functions over M^{2n}. For finite λ_P instead, position and time operators X^μ satisfy nontrivial commutation relations of the form of Eq. (1) with $Q^{\mu\nu}$ related in a simple way to Ω.

II. PATH INTEGRAL OVER SPACE–TIMES

We start considering the pair (R^{2n},Ω), with Ω a symplectic form and the following generating functional:

$$Z(x_0,J)=N\int D[\gamma]\exp[i\lambda_P^{-2}(\Sigma[x_0,\gamma]+(x,J))], \tag{3}$$

where N is a normalization constant, J an arbitrary source, the integration is carried over all closed curves γ in R^{2n} with base point $x_0\in R^{2n}$ and we define

$$\Sigma[x_0,\gamma]=\int_\Gamma \Omega_{\mu\nu}(x)dx^\mu\wedge dx^\nu, \tag{4}$$

$$(x,J)=\int_\gamma x^\mu(\tau)J_\mu(\tau)d\tau, \tag{5}$$

where we have introduced a parameterization, γ: $\tau\mapsto x^\mu(\tau)$ and Γ is any two-dimensional surface with boundary γ. Finally, the antisymmetric matrix $\Omega_{\mu\nu}(x)$ is the representation of the symplectic form in the coordinate basis x^μ and, in general, explicitly depends on x^μ. Here we will assume for simplicity that $\Omega_{\mu\nu}(x)$ is a slowly varying function of x on the scale λ_P, and we consider only zero order terms in Taylor expansion around the base point x_0 in integrals (4), $\Omega_{\mu\nu}(x)$

$\sim\Omega_{\mu\nu}(x_0)$. This assumption simplifies the explicit evaluation of $Z(x_0,J)$, which is our main aim in this paper, but is not necessary. The general case will be considered elsewhere. In this case $\Sigma[x_0,\gamma]$ can therefore be written as

$$\Sigma[x_0,\gamma]=\int_\gamma x^\mu(\tau)\Omega_{\mu\nu}(x_0)\,\frac{d}{d\tau}\,x^\nu(\tau)d\tau. \tag{6}$$

$Z(x_0,J)$ is invariant under curve reparameterization, as is easy to verify, provided we redefine the arbitrary external source J. In the following we will consider $\tau\in[0,1]$.

Equation (3) is clearly inspired by the expression of path integral in phase space with external sources. However it is worth pointing out that there is one main difference with ordinary path integral approach to quantum mechanics of Eq. (2), even for $q_0=q_f$. In this case, in fact, the variables q and p are treated on different grounds, since integration is carried over all curves with boundary conditions on positions only, while momenta run over all possible values. In case of Eq. (3), instead, all curves γ should have equal end points x_0^μ for all x^μ's, $\mu=1,...,2n$.

The generating functional $Z(x_0,J)$ defines a one parameter family of (noncommutative) algebras $\mathscr{A}(\tau)$ generated by $2n$ elements $X^\mu(\tau)$, implicitly defined via the introduction of a set of linear functionals ρ_{x_0} as follows:

$$\begin{aligned}\rho_{x_0}(P[X^\mu(\tau_1)\ldots X^\nu(\tau_k)])&\equiv NZ(x_0,0)^{-1}\int D[\gamma]e^{i\Sigma[x_0,\gamma]/\lambda_P^2}x^\mu(\tau_1)\ldots x^\nu(\tau_k)\\&=(-i\lambda_P^2)^kZ(x_0,0)^{-1}\frac{\delta}{\delta J_\mu(\tau_1)}\cdots\frac{\delta}{\delta J_\nu(\tau_k)}\,Z(x_0,J)\Big|_{J=0},\end{aligned} \tag{7}$$

where P stands for path ordering, defined analogously to time ordering in quantum mechanics. Once defined on all polynomials in the X^μ, Eq. (7) can be applied to the entire algebra of continuous functions $F(X)$ in the weak topology.

To better understand Eq. (7) we consider the *classical* limit $\lambda_P\to 0$. The integral over curves γ is then dominated by the contribution at stationary points satisfying $\delta\Sigma[x_0,\gamma]=0$, i.e., $\Omega_{\mu\nu}(x_0)dx^\nu/d\tau=0$. Since Ω is not degenerate, the leading contribution satisfying $x^\mu(0)=x^\mu(1)=x_0^\mu$, is therefore given by the curve $x^\mu(\tau)=x_0^\mu$, $\forall\tau$, so that we get

$$\rho_{x_0}(P[X^\mu(\tau_1)\cdots X^\nu(\tau_k)])\to x^\mu\cdots x^\nu|_{x_0}+O(\lambda_P^2). \tag{8}$$

Thus, the maps ρ_{x_0} reduce to the evaluation maps of the commutative algebra of smooth functions over R^{2n}, $\rho_{x_0}:f(x)\mapsto f(x_0)$, independently of τ, and thus completely reconstruct R^{2n} as a topological manifold. For finite λ_P, instead, all closed curves whose projection on the $x^\mu-x^\nu$ plane enclose surfaces with area of the order $\Omega_{\mu\nu}(x_0)^{-1}\lambda_P^2$ give a relevant contribution to the right-hand side of Eq. (7), and evaluation maps are smeared over cells in R^{2n} with projection on the $x^\mu-x^\nu$ plane of this order of magnitude. It is worth mentioning that stationary phase method requires Hessian to be nondegenerate at stationary points. As for the case of wave packets of zero mass particles, this aspect deserves a careful analysis in the case under consideration, since the self-adjoint extension of the operator $\Omega_{\mu\nu}(x_0)d/d\tau$ admits zero modes. We will deal with this problem by using in the following a customary $i\epsilon$ prescription which renders the above operator invertible.

We start evaluating $Z(x_0,J)$ for the case $n=1$. Use of Darboux theorem allows for a straightforward generalization to arbitrary n. In two dimensions we write $\Omega_{\mu\nu}(x_0)$ as $\omega(x_0)\epsilon_{\mu\nu}$, with $\omega(x_0)$ a nonvanishing function on R^2. Furthermore all closed curves $x^\mu(\tau)$ with end points at x_0^μ can be written as $x^\mu(\tau)=x_0^\mu+y^\mu(\tau)$, where $y^\mu(0)=y^\mu(1)=0$. Substituting in (3) we get

$$Z(x_0,J)=N\int D[y]\exp\left[i\lambda_P^{-2}\left(\omega(x_0)\int_0^1 y^\mu(\tau)\epsilon_{\mu\nu}\,\frac{d}{d\tau}\,y^\nu(\tau)d\tau+(x_0,J)+(y,J)\right)\right]. \tag{9}$$

The operator $\epsilon_{\mu\nu}d/d\tau$ is self-adjoint with respect to the scalar product (5) on the domain of periodic two component functions $y^\mu(\tau)$. Its spectrum is given by $\mu_m=2m\pi$, $m\in Z$. To each eigenvalue correspond the two eigenfunctions, normalized to λ_P,

$$y_m^{(+)}(\tau)=\lambda_P\begin{pmatrix}\cos(\mu_m\tau)\\ \sin(\mu_m\tau)\end{pmatrix},\quad y_m^{(-)}(\tau)=\lambda_P\begin{pmatrix}\sin(\mu_m\tau)\\ -\cos(\mu_m\tau)\end{pmatrix}. \tag{10}$$

Expanding the functions $y^\mu(\tau)$ in the basis (10), $Z(x_0,J)$ becomes a usual pseudo-Gaussian integral over the Fourier components $c_m^{(\pm)}=\lambda_P^{-1}(y_m^{\pm},y)$. As mentioned before, we will introduce a $i\epsilon$ prescription as a regularization of the singularity due to the presence of a zero mode, namely shifting in the complex plane $\mu_m\to\mu_m+i\epsilon$. In this way we get, integrating over $c_m^{(\pm)}$

$$Z(x_0,J)=N'\ \exp\Bigg[i\lambda_P^{-2}x_0^\mu\int_0^1 J_\mu(\tau)d\tau-(4\epsilon\lambda_P^2)^{-1}\delta^{\mu\nu}\int_0^1 J_\mu(\tau)d\tau\int_0^1 J_\nu(\tau')d\tau'$$
$$-i\sum_{m\neq 0}(4\mu_m\omega(x_0)+i\epsilon)^{-1}((J_m^{(+)})^2+(J_m^{(-)})^2)\Bigg], \tag{11}$$

where $J_m^{(\pm)}=\lambda_P^{-1}(x_m^{(\pm)},J)$ are the Fourier components of $J_\mu(\tau)$ and notice that the zero mode does not contribute to the sum, so we can safely neglect the $i\epsilon$ term in this case. Using (10) and performing the sum over m we finally obtain

$$Z(x_0,J)=N'\ \exp\Bigg[i\lambda_P^{-2}x_0^\mu\int_0^1 J_\mu(\tau)d\tau-(4\epsilon\lambda_P^2)^{-1}\delta^{\mu\nu}\int_0^1 J_\mu(\tau)d\tau\int_0^1 J_\nu(\tau')d\tau'$$
$$+i\lambda_P^{-2}\int_0^1 d\tau\int_0^1 d\tau'\,\Omega^{-1\mu\nu}(x_0)J_\mu(\tau)J_\nu(\tau')\Delta(\tau-\tau')\Bigg], \tag{12}$$

where the kernel $\Delta(t)$ is given by

$$8\Delta(t)=\begin{cases}2t-1 & t\in\,]0,1[\\ 2t+1 & t\in\,]-1,0[\end{cases} \tag{13}$$

which is discontinuous at $t=0$. We can now proceed to explicitly evaluate the action of the maps ρ_{x_0} on path ordered polynomials of X^μ. Up to second order polynomials we get

$$\rho_{x_0}(X^\mu(\tau))=x_0^\mu,\quad \rho_{x_0}(P[X^\mu(\tau)X^\nu(\tau')])=x_0^\mu x_0^\nu+\frac{\lambda_P^2}{2\epsilon}\,\delta^{\mu\nu}-2i\lambda_P^2\Omega^{-1\mu\nu}(x_0)\Delta(\tau-\tau'). \tag{14}$$

Actually, since $Z(x_0,J)$ only contains linear and quadratic terms in J, all other higher order polynomials can be evaluated, up to combinatorial factors, in terms of these expressions.

We see from (14) that ρ_{x_0} simply associates to $X^\mu(\tau)$ the value x_0^μ, independently of τ. In fact the only dependence on the τ variable is contained in the function $\Delta(\tau-\tau')$, appearing for path ordered products of $X^1(\tau)X^2(\tau')$. As long as arbitrary polynomials $X^\mu(\tau_1)\cdots X^\mu(\tau_k)$ in only one of the X^μ operators are considered, the dependence on the parameters τ_i drops out. We also notice that we get for the *uncertainty* functional

$$\rho_{x_0}(X^\mu(\tau)X^\mu(\tau'))-\rho_{x_0}(X^\mu(\tau))\rho_{x_0}(X^\mu(\tau'))=\frac{\lambda_P^2}{2\epsilon}\ \forall\,\tau,\tau'. \tag{15}$$

These results have a customary interpretation by looking at ρ_{x_0} as corresponding to states $|x_0\rangle$ in a Hilbert space with respect to which $X^\mu(\tau)$ have mean value x_0^μ and squared uncertainty as in (15). For the equal τ commutator we find

$$\rho_{x_0}([X^\mu,X^\nu](\tau))=\lim_{\delta\to 0^+}(\rho_{x_0}(X^\mu(\tau+\delta)X^\nu(\tau))-\rho_{x_0}(X^\nu(\tau+\delta)X^\mu(\tau)))$$
$$=\frac{i}{2}\,\lambda_P^2\Omega^{-1\mu\nu}(x_0),\quad \forall\,\tau. \tag{16}$$

As expected, the commutator is proportional to the inverse symplectic form Ω^{-1} and using notation of Eq. (1), we see that $X^\mu(\tau)$ generates a noncommutative algebra with, for any τ, $\rho_{x_0}(Q^{\mu\nu}(\tau))=\Omega^{-1\mu\nu}(x_0)/2$.

The uncertainties (15) diverge if one lets the regularizing parameter ϵ go to zero. This means that in this limit the states $|x_0\rangle$ are no longer in the domain of X^μ. However, the $1/\epsilon$ singularity can be removed if one defines a *renormalized* set of linear functionals $\tilde{\rho}_{x_0}$ which satisfy the first of Eqs. (14), Eq. (16), and have minimal squared uncertainties compatible with Eq. (16), and equals for both $X^1(\tau)$ and $X^2(\tau)$, namely $\lambda_P^2(4\omega(x_0))^{-1}$. Equation (14) then becomes

$$\tilde{\rho}_{x_0}(X^\mu(\tau))=x_0^\mu, \quad \tilde{\rho}_{x_0}(P[X^\mu(\tau)X^\nu(\tau')])=x_0^\mu x_0^\nu+\frac{\lambda_P^2}{4\omega(x_0)}\,\delta^{\mu\nu}-2i\lambda_P^2\Omega^{-1\mu\nu}(x_0)\Delta(\tau-\tau'). \tag{17}$$

Accordingly, the corresponding generating functional $\tilde{Z}(x_0,J)$ is obtained from Eq. (12) by substituting ϵ with $2\omega(x_0)$. The states corresponding to this new set of linear functionals would be analogous to coherent states for ordinary quantum mechanics. They would also correspond to the states of maximal localization introduced in Ref. 8. Notice also that $\tilde{\rho}_{x_0}$ gives finite results when applied to arbitrary higher order polynomials in $X^\mu(\tau)$, which, as already stressed, can all be expressed in terms of (17).

The above results can be easily generalized to the $2n$ dimensional case. We start again with Eqs. (3)–(5), written now on R^{2n}. As before, we will consider in Eq. (4) only zero terms in the Taylor expansion of $\Omega(x)$. By the Darboux theorem it is always possible to reduce $\Omega(x_0)$, for any x_0, to the canonical form

$$\omega=\begin{pmatrix} 0 & 1_n \\ -1_n & 0 \end{pmatrix} \tag{18}$$

with a suitable coordinate transformation $z^\mu=(A^{-1})^\mu_\nu(x_0)y^\nu$. In these new coordinates and within the approximation $\Omega(x)\sim\Omega(x_0)$, $\Sigma[x_0,\gamma]$ is simply the sum of n bidimensional integrals like the first term in Eq. (9) and $Z(x_0,J)$ factorizes into the product of n bidimensional generating functionals depending on the sources $J'_\mu=(A^T)^\nu_\mu(x_0)J_\nu$, with A^T the transpose of A. Carrying out the pseudo-Gaussian integrals and transforming back the result in terms of J_μ we therefore obtain

$$Z(x_0,J)=N'\,\exp\Bigg[i\lambda_P^{-2}x_0^\mu\int_0^1 J_\mu(\tau)d\tau-(4\epsilon\lambda_P^2)^{-1}(AA^T)^{\mu\nu}(x_0)\int_0^1 J_\mu(\tau)d\tau\cdot\int_0^1 J_\nu(\tau')d\tau'$$
$$+i\lambda_P^{-2}\int_0^1 d\tau\int_0^1 d\tau'\,\Omega^{-1\mu\nu}(x_0)J_\mu(\tau)J_\nu(\tau')\Delta(\tau-\tau')\Bigg] \tag{19}$$

which gives

$$\rho_{x_0}(X^\mu(\tau))=x_0^\mu, \tag{20}$$

$$\rho_{x_0}(P[X^\mu(\tau)X^\nu(\tau')])=x_0^\mu x_0^\nu+\frac{\lambda_P^2}{2\epsilon}\,(AA^T)^{\mu\nu}(x_0)-2i\lambda_P^2\Omega^{-1\mu\nu}(x_0)\Delta(\tau-\tau'), \tag{21}$$

and for the equal τ commutator again Eq. (16).

It is worth noticing that a constant Ω in four dimensions corresponds to the case considered in Ref. 3, where the following algebra has been studied:

$$[X^\mu,X^\nu]=i\lambda_P^2Q^{\mu\nu}, \quad [X^\mu,Q^{\mu\nu}]=0,$$

$$Q_{\mu\nu}Q^{\mu\nu}=0, \quad (\tfrac{1}{8}\,\epsilon_{\mu\nu\rho\sigma}Q^{\mu\nu}Q^{\rho\sigma})^2=I. \tag{22}$$

Actually the condition $\rho_{x_0}([X^\mu,Q^{\mu\rho}](\tau))=0$, $\forall\tau$ can be easily verified using Eq. (19), while the constraint $\rho_{x_0}(\epsilon_{\mu\nu\rho\sigma}Q^{\mu\nu}Q^{\rho\sigma}(\tau))=\pm 8$ is equivalent, up to a suitable normalization, to require Ω

to be nondegenerate. The procedure of describing a noncommutative geometry with a path integral approach outlined in this paper cover, however, more general cases, since the commutator may depend on x_0, as is clear from Eqs. (16) and (19). Moreover, if one does not assume that $\Omega(x)$ is a slowly varying function of x and that for each x_0 one can approximate $\Omega(x)\sim\Omega(x_0)$ in the phase $\Sigma[x_0,\gamma]$, one may consider, at least in principle, a completely general case, although the corresponding generating functional could be difficult to handle. Nevertheless, even if $Z(x_0,J)$ is not exactly solvable for more involved choices of Ω, applications of perturbation techniques, as in ordinary path integral in quantum mechanics, may provide some information on the underlying noncommutative space–time.

III. PATH INTEGRAL OVER SPACE–TIMES WITH NONTRIVIAL TOPOLOGIES

In the previous section we considered the path integral construction of noncommutative generalizations of space–times for the case of R^{2n} as the underlying classical manifold. When the method outlined there is applied to manifolds M^{2n} with nontrivial topologies, in particular when the first and/or second homology groups are nontrivial, there are (however) two aspects which one should take into account. First of all, if $H^1(M^{2n})\neq\{0\}$, there are closed curves contributing to the sum over paths which are not boundaries of any surface, so the surface integral of the assigned symplectic form is not defined for these paths. In these cases the introduction of a one-form A, such that locally $\Omega=dA$ is a better starting point. Moreover, if $H^2(M^{2n})\neq\{0\}$, the integral $\Sigma[x_0,\gamma]$ for any γ is in general a multivalued function, since the integral of Ω over two surfaces Γ and Γ', both with boundary γ, may be different if $\Gamma-\Gamma'$ is a nontrivial 2-cycle (of course this problem is absent if Ω is exact). The ambiguity in the choice of the surface, however, can be removed by requiring *quantization* conditions on the integral of Ω over a set of generators of $H^2(M^{2n})$ such that $\exp(i\Sigma[x_0,\gamma]\lambda_P^{-2})$ is single valued, analogous to the one introduced to quantize the motion of a charged particle in the field of a magnetic charge. We will illustrate these points in two bidimensional simple examples, the sphere S^2 and the torus $S^1\times S^1$, and discuss the generalization to general cases.

Let us consider the pair (S^2,Ω). Any closed curve γ will divide S^2 into two surfaces, and therefore the value of $\Sigma[x_0,\gamma]$ of Eq. (4) will be defined only up to the integral of Ω over S^2. The measure in the path integral will be, nevertheless, single valued if the latter is a multiple of 2π, in unit λ_P^2,

$$\int_{S^2}\Omega=2n\pi\lambda_P^2. \tag{23}$$

A path integral over S^2 was already considered in Ref. 11, as a way to quantize a classical particle with spin.

Incidentally, we observe that condition (23) could have an intriguing relationship with the idea that the black hole horizon area is quantized and its spectrum is uniformly spaced.[12] In our approach, this quantization in unit of λ_P^2, would be intimately related to the fact that positions and time are noncommuting operators and their commutator is of the form reported in Eq. (1). Actually black hole physics is one of the best scenarios where the noncommutative nature of space–time at small scales, otherwise unobservable, could be felt well above the Planck length.

The ambiguity in the definition of the measure in path integral and the necessity of a *quantization* condition (23) is due to the fact that the second homology group of S^2 is nontrivial. A similar ambiguity in defining the *action* $\Sigma[x_0,\gamma]$ corresponding to a path γ on a manifold M^{2n}, and therefore the generating functional (3), will arise whenever $H^2(M^{2n})\neq\{0\}$. In particular, if its 2-Betti number is equal to $b_2(M^{2n})=k$, and denoting with Γ_i, $i=1,...,k$, a set of generators of $H^2(M^{2n})$, the measure $\exp(i\Sigma[x_0,\gamma]\lambda_P^{-2})$ will be single valued if Ω satisfies the following conditions:

$$\int_{\Gamma_i}\Omega=2n_i\pi\lambda_P^2,\quad i=1,...,k, \tag{24}$$

which of course do not depend on the choice for the elements of the basis Γ_i.

As a second example of a nontrivial topological manifold, we consider the bidimensional torus $T^2=S^1\times S^1$. In this case there are closed curves which are not boundaries of any surface and therefore the construction outlined in the previous section cannot be applied. Nevertheless it is still possible to construct a path integral over T^2 and the generating functional $Z(x_0,J)$ starting with a one-form A such that (locally) $\Omega=dA$. Let us denote with $\gamma_i(x_0)$, $i=1,2$ two generators of the T^2 1-homology group passing through the point x_0 and define

$$I_i(x_0)=\lambda_P^{-2}\int_{\gamma_i(x_0)}A. \tag{25}$$

The path integral measure we will introduce soon, will be single valued only if the values of $I_i(x_0)$ only differ for a factor $2n\pi$ when the curve γ_i is continuously moved around the manifold and go back to the original path after a complete cycle. For example, in the angular coordinates θ and ϕ on T^2, taking $A=r_1r_2(\theta d\phi-\phi d\theta)$, with r_1,r_2 the two radii, and choosing $\gamma_1(\theta_0,\phi_0)$: $\{\theta=\theta_0\}$ and $\gamma_2(\theta_0,\phi_0)$: $\{\phi=\phi_0\}$, the values of $I_1(x_0)$ for the identified points θ_0 and $\theta_0+2\pi$, and analogously for $I_2(x_0)$, will be the same up to a factor $2n\pi$ provided that

$$\int_{T^2}dA=8\pi^2r_1r_2=2n\pi\lambda_P^2. \tag{26}$$

This condition also ensures, as we have seen for the S^2 case, that for all curves which are instead boundaries of surfaces, the corresponding contribution to the generating functional is single valued. Defining

$$\Sigma[x_0,\gamma]=\int_\gamma A \tag{27}$$

the contribution to $Z(x_0,J)$ of a generic path γ on T^2 can be written as

$$\exp[i\lambda_P^{-2}\Sigma[x_0,\gamma]]=\exp\left[in_1I_1(x_0)+in_2I_2(x_0)+i\lambda_P^{-2}\int_{\gamma'}A\right] \tag{28}$$

for some integers n_1 and n_2 and with γ' a boundary of some bidimensional surface, so the generating functional takes the form

$$Z(x_0,J)=N\sum_{n_1,n_2}\int D[\gamma']\exp(in_1I_2(x_0)+in_2I_2(x_0))\exp\left(i\lambda_P^{-2}\int_\Gamma dA\right)\exp(i\lambda_P^{-2}(x,J)), \tag{29}$$

where the integration is over all 1-boundaries γ' with base point x_0 and $\partial\Gamma=\gamma'$.

This expression can now be easily generalized to an arbitrary manifold M^{2n} with Betti numbers $b_1=p$, $b_2=q$. Consider the pair (M^{2n},A), a set of basis $\gamma_i(x_0)$ and a base Γ_j, for $H^1(M^{2n})$ and $H^2(M^{2n})$, respectively. The fact that the path integral measure should be single valued enforces q integral conditions on $\Omega=dA$,

$$\int_{\Gamma_j}\Omega=2m_j\pi\lambda_P^2,\quad j=1,...,q, \tag{30}$$

and defining, as in (25)

$$I_i(x_0)=\lambda_P^{-2}\int_{\gamma_i(x_0)}A,\quad i=1,...,p, \tag{31}$$

we may introduce the generating functional as follows:

$$Z(x_0,J)=N\prod_{i=1}^{p}\sum_{n_i}\int D[\gamma']\exp(in_iI_i(x_0))\exp\left(i\lambda_P^{-2}\int_\Gamma\Omega\right)\exp(i\lambda_P^{-2}(x,J)), \tag{32}$$

where $\partial\Gamma=\gamma'$. Equation (32) represents the generalization of Eq. (3) to a generic topological manifold M^{2n}.

IV. CONCLUSIONS AND OUTLOOK

In this paper we have described a path integral approach to noncommutative geometries. In particular, using the analogy with ordinary quantum mechanics, we have described a way of expressing a set of linear functionals over noncommutative generalizations, in even dimensions, of the algebra generated by position and time operators in terms of ordinary functions over the classical manifold, which represents its commutative classical geometry. In particular all these linear functionals reduce in the limit $\lambda_P=0$ to evaluation maps over the commutative algebra of smooth functions over the manifold.

There are several points related to this approach which deserve further studies.

We have considered in more detail the simple case of a classical starting manifold with trivial first and second homology groups. It would be interesting to perform a deeper study of the properties of the generating functional over manifolds with more involved topological properties, which has been discussed in Sec. III, and of the corresponding deformed noncommutative geometries.

Another important issue is that this approach may give the possibility of introducing a differential calculus over noncommutative generalizations of the algebra of functions over classical space–times using all customary results for a commutative geometry. In particular, quite relevant is to look for a consistent definition of a metric structure, connections and curvature. Of course, when λ_P is switched on, one expects that dramatic changes affect all these structures. Nevertheless, to have a simple way of relating the algebra generated by the X^μ's to its commutative counterpart, could be a powerful way to look for their generalization.

The role of the parameter τ should be better understood. While it has a clear meaning as the parameter of closed curves over M^{2n}, its role for the noncommutative algebras $\mathcal{A}(\tau)$ is not transparent. Actually all dependence on τ is contained in the kernel $\Delta(\tau-\tau')$. One possibility is, of course, that each of these algebras, for fixed τ, is already sufficient to describe a noncommutative space–time, so that the role of the parameter τ in this case would be only auxiliary. On the other hand, the analogy with quantum phase space we started with, suggests that it could also have an intrinsic nontrivial meaning.

ACKNOWLEDGMENTS

I am pleased to thank F. Lizzi, G. Miele, and G. Sparano for discussions and comments.

This paper is dedicated to my mother.

[1] D. Amati, M. Ciafaloni, and G. Veneziano, Nucl. Phys. B **347**, 550 (1990).
[2] A. Kempf, J. Math. Phys. **35**, 4483 (1994).
[3] S. Doplicher, K. Fredenhagen, and J. E. Roberts, Commun. Math. Phys. **172**, 187 (1995).
[4] M. Maggiore, Phys. Lett. B **319**, 83 (1993).
[5] A. Connes, *Noncommutative Geometry* (Academic, New York, 1994).
[6] H. S. Snyder, Phys. Rev. **71**, 38 (1947).
[7] J. Madore, *An Introduction to Noncommutative Differential Geometry and its Physical Applications* (Cambridge University Press, Cambridge, 1995), and reference therein.
[8] A. Kempf, G. Mangano, and R. B. Mann, Phys. Rev. D **52**, 1108 (1995); A. Kempf and G. Mangano, *ibid.* **55**, 7907 (1997).
[9] A. A. Kehagias, P. A. A. Meessen, and G. Zoupanos, Phys. Lett. B **346**, 262 (1995).
[10] H. Grosse, C. Klimcik, and P. Presnajder, Commun. Math. Phys. **180**, 429 (1996).
[11] H. B. Nielsen and D. Rohrlich, Nucl. Phys. B **299**, 471 (1988).
[12] For a review, see J. D. Bekenstein, gr-qc/9710076.

From Turbiner's quasi-exactly soluble potentials in *N* dimensions to analytic solutions of the combined Coulomb plus oscillator system

Harry A. Mavromatis
Physics Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

(Received 25 June 1997; accepted for publication 6 November 1997)

Turbiner's one-dimensional, quasi-exactly soluble potentials can be interpreted as central potentials in *N* dimensions. Using a particular transformation, one shows how these potentials convert to seemingly unrelated isolated solutions of other Schrödinger equations, including solutions of the combined Coulomb plus harmonic oscillator system. © *1998 American Institute of Physics.*
[S0022-2488(98)03803-1]

I. INTRODUCTION

Quasi-analytic solutions of the one-dimensional Schrödinger equation were first pointed out by Turbiner.[1] These involve one-dimensional systems subject to the potentials

$$V(x)=2x^2[(x^2+\beta)^2-(2n+r+\tfrac{3}{2})], \quad n=0,1,2..., \quad r=0, \text{ or } 1. \tag{1}$$

For a given value of n, i.e., a particular quasi-exactly soluble potential, one can directly obtain the first $n+1$ eigenvalues and the corresponding eigenfunctions. The remaining solutions cannot be obtained analytically, hence the nomenclature ''quasi-exactly soluble potentials.''

In this paper one uses the generalization of Turbiner's work to N dimensions. Applying a recently discussed transformation,[2] to the resulting N-dimensional Schrödinger equation, it is shown to transform into other, seemingly unrelated Schrödinger equations and potentials. This yields new isolated solutions of various Schrödinger equations, including solutions of the interesting combined Coulomb plus harmonic oscillator system.

II. GENERALIZATION OF TURBINER'S RESULTS TO *N*-DIMENSIONAL SPACES

The following N-dimensional central potential $V(r)$:

$$V(r)=2r^2\left[(r^2+\beta)^2-\left(2n+\frac{k_1}{2}+1\right)\right], \quad n=0,1,2,..., \tag{2}$$

when inserted into the N-dimensional Schrödinger equation:[3]

$$\left[-\frac{1}{2}\left(\frac{d^2}{dr^2}-\frac{(k_1-1)(k_1-3)}{4r^2}\right)+V(r)-E\right]u_1(r)=0, \tag{3}$$

where $m=\hbar=c=1$, $k_1=N+2l$, and $\int_0^\infty u_1^2(r)dr=1$, $(u_1(0)=0)$ can be seen to be the N-dimensional generalization of the one-dimensional quasi-exactly soluble potentials Eq. (1) introduced by Turbiner.[1] Turbiner's one-dimensional cases are obtained from the above by setting $k_1=1(\Rightarrow N=1,\ l=0)$ or $3(\Rightarrow N=3,\ l=0)$. Equation (3) then reduces to the one-dimensional Schrödinger equation and Eq. (2) becomes equivalent to Eq. (1) for the two cases $r=0,1$. Equation (2) is a sextic potential. The factors in Eq. (1) are slightly different from those in Turbiner's papers because he works without the factor 1/2 in the first term of Eq. (3), etc.

If one assumes solutions of the form:

$$u_1(r)=Cr^{(k_1-1)/2}e^{-(r^2+\beta)^2/2}f(r), \tag{4}$$

 © 1998 American Institute of Physics

and substitutes back in Eq. (3), this yields for $f(r)$:

$$\left\{ f''(r) - \left[4r(r^2+\beta) - \frac{k_1-1}{r} \right] f'(r) - 2(\beta k_1 - 4nr^2 - E) f(r) \right\} = 0. \tag{5}$$

The $n+1$ $f(r)$'s and corresponding E's can be obtained by substituting in Eq. (5) a polynomial in r^2 of order n, i.e.,

$$f(r) = \sum_{q=0}^{n} c_q r^{2q}. \tag{6}$$

As in the one-dimensional case, the remaining solutions are unobtainable.

Special cases:

The case $n=0$ involves one exact solution:

$$E = \beta k_1, \quad f(r) = 1. \tag{7}$$

The case $n=1$ involves two exact solutions:

$$E = \beta(k_1=2) \mp 2\sqrt{\beta^2+k_1}, \quad f(r) = 1 - 2(\beta \mp \sqrt{\beta^2+k_1})\, \frac{r^2}{k_1}. \tag{8}$$

The case $n=2$ involves the three exact solutions, whose energies and coefficients c_q are obtained by solving the cubic equation:

$$E^3 - 3\beta(k_1+4)E^2 + (3\beta^2 k_1^2 + 24\beta^2 k_1 + 32\beta^2 - 16k_1 - 16)E$$

$$- \beta k_1(\beta^2 k_1^2 + 12\beta^2 k_1 + 32\beta^2 - 16k_1 - 80) = 0,$$

where

$$c_0 = 1, \quad c_1 = \beta - E/k_1, \quad c_2 = \{-8 + (\beta - E/k_1)(4\beta + \beta k_1 - E)\}/(2k_1+4). \tag{9}$$

Thus for three-dimensional systems, given the potential

$$V(r) = 2r^2[(r^2+\beta)^2 - (l+9/2)], \tag{10}$$

one has two solutions:

$$E = \beta(5+2l) \mp 2\sqrt{\beta^2+3+2l},$$

$$u_1(r) = Cr^{l+1} e^{-(r^2+\beta)^2/2} \left\{ 1 - 2(\beta \mp \sqrt{\beta^2+3+2l})\, \frac{r^2}{3+2l} \right\}, \tag{11}$$

and so on for higher n.

III. TRANSFORMATION OF TURBINER'S RESULTS

If one applies the transformation:[2] $r = \rho^m$, $u_1(r) = \rho^{(m-1)/2} u_m(\rho)$, to Eq. (3), it transforms into

$$\left[-\frac{1}{2} \left(\frac{d^2}{d\rho^2} - \frac{(k_m-1)(k_m-3)}{4\rho^2} \right) + m^2 \rho^{2(m-1)} \{V(\rho^m) - E\} \right] u_m(\rho) = 0, \tag{12}$$

where $k_m = m(k_1-2)+2$. In other words Eq. (3) transforms into a Schrödinger equation in a different dimension, with a potential and energy $v(\rho) - \epsilon = m^2 \rho^{2(m-1)} \{V(\rho^m) - E\}$.

An interesting special case to consider for the potential of Eq. (2) is $m = 1/2$. One then obtains the Schrödinger equation:

$$\left[-\frac{1}{2}\left(\frac{d^2}{d\rho^2}-\frac{(k_{1/2}-1)(k_{1/2}-3)}{4\rho^2}\right)+\frac{1}{2}(\rho+\beta)^2-\frac{E}{4\rho}-\frac{1}{2}(2n+k_{1/2})\right]u_{1/2}(\rho)=0, \tag{13}$$

with $k_{1/2}=k_1/2+1$, where $\int_0^\infty u_{1/2}^2(\rho)d\rho=1$ and $u_{1/2}(0)=0$.

One notes that for this value of m, part of Eq. (2), the generalized Turbiner potential $[-2r^2[2n+k_{1/2}]]$ has been transformed into an energy term in Eq. (13) [namely, $\frac{1}{2}(2n+k_{1/2})$]. The remainder of Turbiner's generalized potential [namely, $2r^2(r^2+\beta^2)^2$] transforms into the oscillator potential $\frac{1}{2}(\rho+\beta)^2$. The energy term (E) in Eq. (3) has meanwhile been absorbed into the new effective potential (as $-E/4\rho$). By suitable identifications one thus has a Schrödinger equation in the space specified by $k_{1/2}=N'+2l'$, with

$$v(\rho)=\frac{1}{2}(\rho+\beta)^2-\frac{E}{4\rho}, \tag{14}$$

$$\epsilon=\frac{1}{2}(2n+k_{1/2}). \tag{15}$$

For a particular n in Eq. (15), i.e., a particular ϵ, one thus has in this case from Eq. (3), $n+1$ solutions $u_1(r)$ [using Eqs. (4) and (6)] and hence directly determines the corresponding E's. Thus if $n=0$, 1, or 2, the $u_1(r)$'s and E's are given by Eqs. (7), (8), or (9), etc. Each of these particular E's then completely specifies the related potential $v(\rho)$ in Eq. (14). From the corresponding $u_1(r)$ one can also obtain the appropriate eigenfunction $u_{1/2}(\rho)=\rho^{1/4}u_1(\rho^{1/2})$ for that potential and energy.

This particular choice $m=1/2$ is of interest since it results in the potential given by Eq. (14) which includes both a (displaced if $\beta\neq 0$) harmonic oscillator and a Coulomb potential (of a specific sign and strength), treated on equal footing. If one makes the right identification (albeit isolated) solutions for this system can then be obtained. Consider, for example, the case $n=1$. For this choice of n,

$$\epsilon=\frac{1}{2}(2+k_{1/2}),\rightarrow E=2\beta k_{1/2}\mp 2\sqrt{\beta^2+2k_{1/2}-2}, \tag{16}$$

and

$$v(\rho)=\frac{1}{2}(\rho+\beta)^2-\frac{[\beta k_{1/2}\mp\sqrt{\beta^2+2k_{1/2}-2}]}{2\rho}, \tag{17}$$

with isolated solutions

$$u_{1/2}(\rho)=C\rho^{(k_{1/2}-1)/2}e^{-(\rho+\beta)^2/2}\left[1-(\beta\mp\sqrt{\beta^2+2k_{1/2}-2})\,\frac{\rho}{k_{1/2}-1}\right], \tag{18}$$

where $k_{1/2}>1$ so that $u_{1/2}(\rho)\xrightarrow[\rho\rightarrow 0]{}0$ and Eq. (18) has no singularities. If β is zero, Eq. (18) can be normalized analytically as will be shown in the two special cases discussed below. Otherwise the normalization constant C can be obtained numerically.

One notes that the two solutions of Eq. (18) have the same energy ϵ given by Eq. (16). However they are obviously not degenerate states since they correspond to the two different potentials of Eq. (17). For a particularly simple illustration consider the case $k_{1/2}=3$, i.e., $k_1=4$, and $\beta=0$, i.e., the oscillator is centered at the origin $\rho=0$. In this case, if $n=1$:

$$\epsilon=\frac{5}{2},\ \rightarrow E=\mp 4 \tag{19}$$

while

$$v(\rho)=\frac{1}{2}\,\rho^2\pm\frac{1}{\rho}, \tag{20}$$

with isolated solutions:

$$u^{\pm}_{1/2}(\rho)=C^{\pm}_1\rho e^{-\rho^2/2}[1\pm\rho]. \tag{21}$$

As required, the wave functions of Eq. (21) satisfy the condition $u_{1/2}(0)=0$. Normalizing with the condition $\int_0^\infty u^2_{1/2}(\rho)d\rho=1$, one obtains

$$C^{\pm}_1=\frac{1}{\sqrt{(5\sqrt{\pi}/8)\pm 1}}, \quad \text{i.e.,} \quad C^{+}_1\sim 0.69,\ C^{-}_1\sim 3.05.$$

One notes from Eq. (20) that one has here the two cases of a repulsive and of an attractive Coulomb potential, respectively, in addition to the harmonic oscillator.

If the quadratic potential in Eq. (13) is not displaced (i.e., $\beta=0$) and $E=0$, this equation reduces to the isotropic oscillator problem. For the n defined in Eq. (1) this yields acceptable solutions $\epsilon=n+(k_{1/2}/2)$, for $n=$an even integer or zero.

For $n=2$ (i.e., $\epsilon=7/2$), $k_{1/2}=3$ (i.e., $k_1=4$), and $\beta=0$, one has from Eq. (9) that

$$E=0, \quad E=\mp 4\sqrt{5}, \tag{22}$$

i.e.,

$$v(\rho)=\frac{1}{2}\,\rho^2, \quad v(\rho)=\frac{1}{2}\,\rho^2\pm\frac{\sqrt{5}}{\rho}, \tag{23}$$

the former is the pure harmonic oscillator potential, with solution:

$$u_{1/2}(\rho)=C_2\rho e^{-\rho^2/2}[1-\tfrac{2}{3}\rho^2], \tag{24}$$

and the latter, the harmonic oscillator plus a (repulsive or attractive) Coulomb potential, with exact solutions:

$$u^{\pm}_{1/2}(\rho)=C^{\pm}_3\rho e^{-\rho^2/2}[1\pm\sqrt{5}\rho+\rho^2]. \tag{25}$$

Normalizing one obtains

$$C_2=\left(\frac{6}{\sqrt{\pi}}\right)^{1/2}, \quad C^{\pm}_3=\left[\left(\frac{61}{16}\,\sqrt{\pi}\pm 3\sqrt{5}\right)\right]^{-1/2}.$$

Other choices of m yield different potentials and related isolated solutions. For instance, if $m=2$,

$$\left[-\frac{\hbar^2}{2m}\left(\frac{d^2}{d\rho^2}-\frac{(k_2-1)(k_2-3)}{4\rho^2}\right)+8\rho^6\left\{(\rho^4+\beta)^2-\left(2n+\frac{k_2}{4}+\frac{3}{2}\right)\right\}-4\rho^2E\right]u_2(\rho)=0, \tag{26}$$

where $k_2=2k_1-2$. Thus by the method outlined above one obtains isolated solutions of a potential of the form

$$v(\rho)=A\rho^{14}+B\rho^{10}+C\rho^6+D\rho^2, \tag{27}$$

for particular values of A, B, C, and D. If $m=3/2$, one obtains isolated solutions for a potential of the form

$$v(\rho)=F\rho^{10}+G\rho^7+H\rho^4+I\rho, \tag{28}$$

for particular values of F, G, H, and I, etc.

IV. CONCLUSIONS

The generalization to N dimensions of Turbiner's one-dimensional quasi-exactly soluble potentials is discussed in this paper, and a few specific N-dimensional solutions are listed. Applying a suitable transformation, these potentials are shown to be related to the (possibly displaced) isotropic oscillator plus Coulomb potential system. Some normalized isolated solutions for this system are obtained.

ACKNOWLEDGMENTS

KFUPM support is acknowledged.

[1] A. V. Turbiner and A. G. Ushveridze, Phys. Lett. A **126**, 181 (1987); A. Turbiner, Sov. Phys. JETP **67**, 230 (1988); A. Turbiner, Commun. Math. Phys. **118**, 467 (1988); A. V. Turbiner, Phys. Lett. B **276**, 95 (1992); M. Taler and A. V. Turbiner, J. Phys. A **26**, 697 (1993).
[2] H. A. Mavromatis, J. Phys. A **30**, 1685 (1997).
[3] A. Chatterjee, Phys. Rep. **186**, 249 (1990); H. A. Mavromatis, *Exercises in Quantum Mechanics* (Kluwer, Dordrecht, 1992), p. 120.

Quantum averaging. III. Time-dependent systems

Wolfgang Scherer[a)]
Institut für Theoretische Physik A, TU Clausthal,
Leibnizstrasse 10, 38678 Clausthal-Zellerfeld, Germany

(Received 16 July 1997; accepted for publication 23 December 1997)

We present new perturbation algorithms for time-dependent quantum systems. These new methods make use of a time-dependent averaging method and are based on analogies with the Poincaré–von Zeipel (PvZ) and Kolmogorov–Arnold–Moser (KAM) expansions. Application to a simple example shows that the PvZ as well as the KAM expansion are superior to the standard quantum mechanical time-dependent perturbation method (Dyson expansion). © *1998 American Institute of Physics.* [S0022-2488(98)01404-2]

I. INTRODUCTION

The use of HAMILTONian or *symplectic methods* has been very fruitful in dynamical systems in general and in classical mechanics in particular. It has produced deep structural insights (e.g., KAM-Theorem) as well as techniques (e.g., reduction of systems with symmetries) and algorithms (e.g., canonical perturbation theories).

Quantum mechanics, on the other hand, has mostly been treated using the tools of *functional analysis*, although it is fairly well known that the SCHRÖDINGER equation may be written as an infinite-dimensional HAMILTONian system.[1] What seems to prevent the transfer of HAMILTONian methods known in classical mechanics to quantum mechanics is the fact that classical mechanics lives on a finite-dimensional symplectic manifold, whereas quantum mechanics on an infinite-dimensional HILBERT space. This, however, is only an obstacle if the classical information is formulated in a coordinate-dependent manner. At least formally most geometric concepts of classical HAMILTONian systems can be transferred to quantum mechanics. Loosely speaking, the *basic philosophy* behind this program is the following:

P_1: formulate the classical theory in as *purely geometric* a fashion as possible
P_2: formulate the SCHRÖDINGER equation as an *infinite-dimensional* HAMILTONian system
P_3: repeat the geometric constructions of P_1 for the system in P_2.

It turns out that this program applied to the realm of *perturbation theories using the averaging method* is extremely fruitful. The method of averaging has been well known in classical mechanics and is used there to construct perturbation theories free of so called *secular terms*.[2] It is only quite recently, however, that this method has been extended to self-adjoint operators by the author, and with the help of this method analogues of the classical POINCARÉ–VON ZEIPEL[3] as well as of KOLMOGOROV's perturbation theory,[4,5] which is used in the proof of the KOLMOGOROV–ARNOLD–MOSER (KAM) theorem, have been obtained for time-independent (i.e., autonomous) systems.

In this paper we present P_1-P_3 of the above program for POINCARÉ–VON ZEIPEL as well as KAM perturbation theories of time-dependent (i.e., non-autonomous) systems. That is, two perturbation algorithms are constructed in complete analogy to the theories in classical mechanics for time-dependent quantum systems. For such systems both POINCARÉ–VON ZEIPEL, as well as the KAM algorithm, yield *new* perturbation theories which are free of secular terms and in all examples tested perform far better than the DYSON expansion which is the standard time-dependent perturbation method for quantum systems. A short outline of these algorithms has been given for the POINCARÉ–VON ZEIPEL method in Ref. 6 and for the KAM algorithm in Ref. 7. The main purpose of this paper is to present on the formal level the following:

a)Electronic mail: ptws@pt.tu-clausthal.de

 © 1998 American Institute of Physics

- the *ideas* originating from classical HAMILTONian systems as they are translated to quantum systems using *geometrization* as a crucial ingredient of the theory since this enables one to avoid the use of local coordinates which prevents copying classical methods to quantum mechanics.
- the *algorithmic procedures* in a fashion suitable to immediate applications.

This means in particular, that *functional analytic issues* such as domain questions, operator topologies, convergence, etc. will be mostly ignored in this exposition. They are currently under investigation and will be dealt with in the future.

POINCARÉ-type normal form algorithms for time-dependent quantum systems have been constructed before by KUMMER and GOMPA[8,9] and KAM-type algorithms by BELLISSARD,[10,11] COMBESCURE,[12,13] and, building on their work, recently by DUCLOT and STOVICEK.[14] While there is a certain kinship of ideas the approach presented here differs from these in that it emphasises the geometric similartity of the quantum situation with the classical one. This way the construction is more general but still only formally so and the ongoing investigation of convergence (see the remark above) will reveal how much of that generality can be maintained.

This paper is organized as follows.

In Section II the geometric formulation of time-dependent classical and quantum mechanical HAMILTONian systems is presented rather extensively. In this Section all tools necessary to construct the two perturbation algorithms are provided. This includes the notion of extended phase space (Section II B) and time-dependent averaging in classical mechanics (Section II C) as well as quantum mechanics (Sections II D and II E). Transformations on constant time slices are introduced in Section II F. These transformations which use the perturbation parameter as a flow parameter are used in the two perturbation methods (once in the POINCARÉ–VON ZEIPEL method and iteratively in the KAM method) to transform the perturbed HAMILTONian to a HAMILTONian which is such that the computation of its flow is reduced to solving a *time-independent* initial value problem.

Equipped with the averaging construction we then turn to the perturbation algorithms beginning with the POINCARÉ–VON ZEIPEL method in Section III where we first present the classical theory in Section III B and then the quantum theory in Section III C. Since the quantum theory is new it is summarized in an algorithmic fashion suitable to immediate application in Section III D.

The second perturbation method copied from classical mechanics is the time-dependent version of the KAM algorithm which is presented in Section IV beginning again with the classical case in Section IV B and continuing with the quantum case in Section IV C which is summarized in algorithmic fashion in Section IV D.

The two methods—POINCARÉ–VON ZEIPEL, and KAM—along with the standard quantum mechanical method and a numerical calculation using the Mathematica package are then applied to a simple system consisting of a spin-$\frac{1}{2}$ particle in an oscillating magnetic field in Section V and the results for a transition probability are compared in Section V E.

We conclude this paper with a discussion of the methods and associated problems in Section VI.

II. TIME-DEPENDENT HAMILTONIAN SYSTEMS

A. Introduction

The main subject of interest here are HAMILTONian systems whose HAMILTONian depends ''explicitly'' on t. Using the concept of *extended phase-space*[15,16] we shall first recall the geometric description of HAMILTONian systems of this type in Section II B, their time evolution, and the averaging method for them in Section II C. The quantum analogues of these are given in Section II D and in Section II E. The time-dependent version of the non-autonomous ε-transformations is then introduced in Section II F for classical and quantum systems.

A somewhat related approach to time-dependent quantum systems has been introduced by SAMBE,[17] HOWLAND,[18,19] and YAJIMA.[20,21] They, however, define an extended HILBERT space as the tensor product of certain functions of time with the original HILBERT space, whereas here the extended HILBERT space is built in close analogy to the classical situation by adding time and its conjugate coordinate energy to the original HILBERT space viewed as a symplectic manifold.

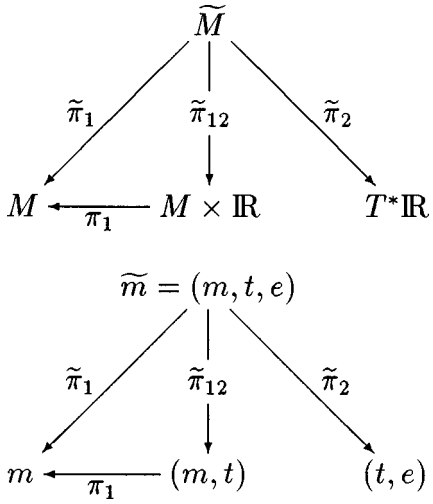

FIG. 1. Projections from extended phase space.

B. Extended phase space

Let M be a symplectic manifold whose points are denoted by m and whose symplectic form is ω and let

$$f: M \times \mathbb{R} \to \mathbb{R}, \quad (m,t) \mapsto f(m,t), \tag{2.1}$$

be a time-dependent HAMILTONian which for each $t \in \mathbb{R}$ defines a function

$$f_t: M \to \mathbb{R}, \quad m \mapsto f(m,t). \tag{2.2}$$

The time-dependent vector field generated by f is then defined as

$$X_f(\cdot,t): M \to TM, \quad m \mapsto X_f(m,t) := X_{f_t}(m), \tag{2.3}$$

where X_{f_t} is defined in the usual way such that for each $t \in \mathbb{R}$,

$$\imath(X_{f_t})\omega = -df_t. \tag{2.4}$$

Here $\imath(Z)\omega$ denotes the one-form $Y \mapsto \omega(Z,Y)$. The flow φ_f of this vector field is the solution of the following non-autonomous initial value problem,

$$\frac{d}{dt}\varphi_f(t,t_0)(m) = X_f(\varphi_f(t,t_0)(m),t), \quad \varphi_f(t_0,t_0)(m) = m, \tag{2.5}$$

which we assume to be unique. This flow still satisfies

$$\varphi_f(t_2,t_1) \circ \varphi_f(t_1,t_0) = \varphi_f(t_2,t_0), \tag{2.6}$$

but it is no longer time-translationally invariant, i.e., in general

$$\varphi_f(t+t_1,t_1) \neq \varphi_f(t+t_0,t_0), \quad \text{if } t_1 \neq t_0. \tag{2.7}$$

With the help of the extended phase space one is able to construct a flow on that space which does have this property and ''includes'' φ_f.[16] We define the extended phase space as

$$\widetilde{M} := M \times T^*\mathbb{R} \ni \widetilde{m} := (m,t,e), \tag{2.8}$$

where t is the base coordinate and e the fiber coordinate in $T^*\mathbb{R}$. This provides us with the projections depicted in Fig. 1. We will mostly consider the following classes of functions on $\widetilde{M}$.

- Lifted functions,
$$C_L(\widetilde{M}) \coloneqq \widetilde{\pi}_1^* C^\infty(M) \cup \widetilde{\pi}_{12}^* C^\infty(M \times \mathbb{R}), \tag{2.9}$$
i.e., functions $f,g,h,w,\ldots$, which are initially defined only on M (resp., $M\times\mathbb{R}$) and will be lifted with $\widetilde{\pi}_1^*$ (resp., $\widetilde{\pi}_{12}^*$) to functions on $\widetilde{M}$. But for the sake of notational sanity we shall not always write out these pull-backs and denote the functions on $\widetilde{M}$ still by the same letter. So, whenever necessary, the application of the pull-backs should be understood.
- Extended functions,
$$C_E(\widetilde{M}) \coloneqq \left\{ \begin{array}{ll} \widetilde{h}\colon & \widetilde{M} \to \mathbb{R} \\ & \widetilde{m} = (m,t,e) \mapsto h(m,t) - e \end{array} \;\middle|\; h \in C_L(\widetilde{M}) \right\}. \tag{2.10}$$

On $\widetilde{M}$ we define the symplectic form

$$\widetilde{\omega} \coloneqq \widetilde{\pi}_1^* \omega - \widetilde{\pi}_2^*(de \wedge dt). \tag{2.11}$$

For lifted functions $f \in C_L(\widetilde{M})$ we have thus the HAMILTONian vector fields,

$$\widetilde{X}_f(\widetilde{m}) = X_{f_t}(m) + \frac{\partial f}{\partial t}(m,t)\, \frac{\partial}{\partial e}, \tag{2.12}$$

where X_{f_t} is given as in (2.4) and the right hand side is to be understood as follows. The tangent maps of the projections $\widetilde{\pi}_1$ and $\widetilde{\pi}_2$ provide us with an isomorphism between $T\widetilde{M}$ and the product bundle $TM \times T^*\mathbb{R}$ over $\widetilde{M}$.[22] Using this isomorphism we may then consider $(X_{f_t}, (\partial f/\partial t) \times (\partial/\partial e))$ as a vector field on $\widetilde{M}$. For extended functions $\widetilde{f} \in C_E(\widetilde{M})$ one finds the HAMILTONian vector fields,

$$\widetilde{X}_{\widetilde{f}}(\widetilde{m}) = \widetilde{X}_f(\widetilde{m}) + \frac{\partial}{\partial t}. \tag{2.13}$$

For $f,g \in C_L(\widetilde{M})$ and $\widetilde{f},\widetilde{g} \in C_E(\widetilde{M})$ we obtain

$$\widetilde{ad}_{\widetilde{f}}(\widetilde{g}) \coloneqq \{\widetilde{\widetilde{f},\widetilde{g}}\} = \{f_t, g_t\} + \frac{\partial g}{\partial t} - \frac{\partial f}{\partial t}, \tag{2.14}$$

$$\widetilde{ad}_{\widetilde{f}}(g) \coloneqq \{\widetilde{\widetilde{f},g}\} = \{f_t, g_t\} + \frac{\partial g}{\partial t}, \tag{2.15}$$

$$\widetilde{ad}_f(g) \coloneqq \{\widetilde{f,g}\} = \{f_t, g_t\}, \tag{2.16}$$

where $\{\widetilde{\cdot,\cdot}\}$ is the POISSON bracket associated with $\widetilde{\omega}$, $\{\cdot,\cdot\}$ the one associated with ω, and, as usual, the lifts in the expressions on the right hand side are to be understood. This means that for $f,g \in C_L(\widetilde{M})$ one has originally $\{f_t,g_t\} \in C^\infty(M)$ but with its dependence on t $\{f_t,g_t\}$ may be viewed as an element of $C^\infty(\widetilde{M})$ and as such it can be lifted to an element of $C_L(\widetilde{M})$. To avoid an overburdened notation we shall thus henceforth most often write just $\{f,g\}$ for $\{f_t,g_t\}$ where $\{f,g\}$ is to be understood in the sense outlined above.

C. Time evolution and averaging on extended phase space

We will now show how to recover the flow φ_h of a time-dependent HAMILTONian h on M from the flow on $\widetilde{M}$ generated by the extended function $\widetilde{h} = h - e$. Let

$$\widetilde{\varphi}_{\widetilde{h}}\colon \mathbb{R} \times \widetilde{M} \to \widetilde{M}, \quad (\tau,\widetilde{m}) \mapsto \widetilde{\varphi}_{\widetilde{h}}(\tau)(\widetilde{m}), \tag{2.17}$$

be the flow generated by the extended function $\widetilde{h}$. It satisfies

$$\frac{d}{d\tau}\,\widetilde{\varphi}_{\widetilde{h}}(\tau)(\widetilde{m})=\widetilde{X}_{\widetilde{h}}(\widetilde{\varphi}_{\widetilde{h}}(\tau)(\widetilde{m})),\quad \widetilde{\varphi}_{\widetilde{h}}(0)(\widetilde{m})=\widetilde{m},\tag{2.18}$$

and if we project it with the $\widetilde{\pi}$'s and use the similar decomposition for the vector field given by (2.13) and (2.12), we can write the differential equation in (2.18) as

$$\frac{d}{d\tau}\begin{pmatrix}\widetilde{\pi}_1\circ\widetilde{\varphi}_{\widetilde{h}}(\tau)(\widetilde{m})\\ t(\tau)(\widetilde{m})\\ e(\tau)(\widetilde{m})\end{pmatrix}=\begin{pmatrix}X_{h_t}(\widetilde{\pi}_1\circ\widetilde{\varphi}_{\widetilde{h}}(\tau)(\widetilde{m}))\\ 1\\ \dfrac{\partial h}{\partial t}(\widetilde{\pi}_{12}(\widetilde{m}))\end{pmatrix},\tag{2.19}$$

where $t(\tau)(\widetilde{m})$ and $e(\tau)(\widetilde{m})$ are the projections down to $\mathbb{R}^2\simeq T^*\mathbb{R}$. Consequently, the extended flow is given by

$$\widetilde{\varphi}_{\widetilde{h}}(\tau)\left(\begin{pmatrix}m\\ t\\ e\end{pmatrix}\right)=\begin{pmatrix}\varphi_h(\tau+t,t)(m)\\ \tau+t\\ e+h(\varphi_h(\tau+t,t)(m),\tau+t)-h(m,t)\end{pmatrix},\tag{2.20}$$

and its projection to M is the original dynamical evolution generated by the time-dependent function h. Note that

$$\widetilde{\varphi}_{\widetilde{h}}(\tau_1+\tau_2)=\widetilde{\varphi}_{\widetilde{h}}(\tau_1)\circ\widetilde{\varphi}_{\widetilde{h}}(\tau_2)\quad\text{and}\quad\widetilde{\varphi}_{\widetilde{h}}(\tau)^*\widetilde{h}=\widetilde{h}.\tag{2.21}$$

If we choose the initial point $\widetilde{m}$ such that $e=h(m,t)$ then the motion is confined to the level set,

$$\widetilde{\Sigma}_0^{\widetilde{h}}:=\{\widetilde{m}\in\widetilde{M}\,|\,\widetilde{h}(\widetilde{m})=0\},\tag{2.22}$$

and the evolution of the coordinate e then gives the value of the energy of the time-dependent system, which is now no longer constant.

We can now present the averaging constructions for time-dependent systems in purely geometric terms.

For any map $\widetilde{\varphi}:\widetilde{M}\to\widetilde{M}$ we define its pull-back action on functions on $\widetilde{M}$,

$$\widetilde{\varphi}^*:C^\infty(\widetilde{M})\to C^\infty(\widetilde{M}),$$

$$f\mapsto\widetilde{\varphi}^*f:=f\circ\widetilde{\varphi},\tag{2.23}$$

and for any function $g\in C^\infty(\widetilde{M})$ we have

$$\frac{d}{d\tau}\,\widetilde{\varphi}_{\widetilde{f}}(\tau)^*g=\widetilde{\varphi}_{\widetilde{f}}(\tau)^*\circ\widetilde{ad}_{\widetilde{f}}(g),\quad\widetilde{\varphi}_{\widetilde{f}}(0)^*g=g,\tag{2.24}$$

as well as for the reverse flow,

$$\frac{d}{d\tau}\,\overset{-1}{\widetilde{\varphi}}_{\widetilde{f}}(\tau)^*g=-\widetilde{ad}_{\widetilde{f}}\circ\overset{-1}{\widetilde{\varphi}}_{\widetilde{f}}(\tau)^*(g),\quad\overset{-1}{\widetilde{\varphi}}_{\widetilde{f}}(0)^*g=g.\tag{2.25}$$

Let $\widetilde{f}\in C_E(\widetilde{M})$, $g\in C_L(\widetilde{M})$ and let $\widetilde{\varphi}_{\widetilde{f}}(\tau):\widetilde{M}\to\widetilde{M}$ be the flow generated by $\widetilde{f}$. Suppose $\widetilde{f}$ and g are such that

$$\lim_{T\to\infty}\frac{\overset{-1}{\widetilde{\varphi}}_{\widetilde{f}}(T)^*g-g}{T}=0\tag{2.26}$$

and

$$\overline{g^{(\tilde f)}} := \lim_{T\to\infty} \frac{1}{T} \int_0^T d\sigma\ \overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* g \in C_L(\widetilde M), \tag{2.27}$$

$$s^{(\tilde f)}(g) := \lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau \int_0^\tau d\sigma\ \overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* (g - \overline{g^{(\tilde f)}}) \in C_L(\widetilde M), \tag{2.28}$$

exist. Then it follows that

$$g - \widetilde{ad}_{\tilde f}(s^{(\tilde f)}(g)) = \overline{g^{(\tilde f)}}, \tag{2.29}$$

$$\widetilde{ad}_{\tilde f}(\overline{g^{(\tilde f)}}) = 0, \tag{2.30}$$

which means

$$g - \{f, s^{(\tilde f)}(g)\} - \frac{\partial s^{(\tilde f)}(g)}{\partial t} = \overline{g^{(\tilde f)}}, \tag{2.31}$$

$$\{f, \overline{g^{(\tilde f)}}\} + \frac{\partial \overline{g^{(\tilde f)}}}{\partial t} = 0. \tag{2.32}$$

We first prove (2.30), i.e. (2.32). Using (2.15) one has

$$\{f, \overline{g^{(\tilde f)}}\} + \frac{\partial \overline{g^{(\tilde f)}}}{\partial t} = \widetilde{ad}_{\tilde f}(\overline{g^{(\tilde f)}}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T d\sigma\ \widetilde{ad}_{\tilde f}(\overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* g),$$

where we have assumed sufficiently nice convergence properties of the limit and the integral in (2.27) such that pulling $\widetilde{ad}_{\tilde f}$ inside is permitted. From (2.25) it follows that then

$$\{f, \overline{g^{(\tilde f)}}\} + \frac{\partial \overline{g^{(\tilde f)}}}{\partial t} = -\lim_{T\to\infty} \frac{1}{T} \int_0^T d\sigma\ \frac{d}{d\sigma}\ \overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* g = -\lim_{T\to\infty} \frac{\overset{-1}{\tilde\varphi}_{\tilde f}(T)^* g - g}{T} = 0,$$

where the last line is a consequence of the assumed property (2.26) and (2.30) is thus established. Now we prove (2.31). Similar to the steps above we have

$$\{f, s^{(f)}(g)\} + \frac{\partial s^{(f)}(g)}{\partial t} = \widetilde{ad}_{\tilde f}(s^{(f)}(g)) = \lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau \int_0^\tau d\sigma\ \widetilde{ad}_{\tilde f}(\overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* g),$$

where we have used already (2.30) in the last line. Using (2.25) once more gives

$$\begin{aligned}
\{f, s^{(f)}(g)\} + \frac{\partial s^{(f)}(g)}{\partial t} &= -\lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau \int_0^\tau d\sigma \frac{d}{d\sigma}\ \overset{-1}{\tilde\varphi}_{\tilde f}(\sigma)^* g \\
&= \lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau (g - \overset{-1}{\tilde\varphi}_{\tilde f}(\tau)^* g) \\
&= g - \overline{g^{(\tilde f)}},
\end{aligned}$$

which proves (2.31), i.e., (2.29).

These proofs are identical (in geometric terms) to the ones for the time-independent case.[3–5] It should be noted, however, that while the functions $\overline{g^{(\tilde{f})}}$ and $\widetilde{s}^{(\tilde{f})}(g)$ do not depend on the extended flow parameter τ, they *do depend on* t. Equations (2.26)–(2.30) constitute the *time-dependent averaging equations* and as in the time-independent theory[3–5] they will be at the heart of the time-dependent perturbation algorithms to be constructed in the following sections.

The averaging equations above may be formulated for more general functions $f,g\in C^\infty(\widetilde{M})$ but we shall not need this generalization. Finally, we define the constant-time slices,

$$\widetilde{\Sigma}_t := \{(m,t,e)\in\widetilde{M}\,|\,m\in M, e\in\mathbb{R}\}\subset\widetilde{M}, \tag{2.33}$$

which are moved by the extended flow,

$$\widetilde{\varphi}_{\tilde{h}}(\tau)(\widetilde{\Sigma}_t)\subset\widetilde{\Sigma}_{\tau+t}. \tag{2.34}$$

D. Time-dependent Hamiltonian quantum mechanics

In this Section we will show how to transfer the geometrical concepts of Sections II B and II C to the quantum mechanical setting. This will be done on a purely formal level. Nevertheless, we will restrict the exposition to bounded operators and smooth time-dependence to avoid the otherwise omnipresent domain questions for operators. For a rigorous treatment of the geometry on HILBERT spaces in general we refer the reader to Ref. 23, Chaps. 1.1–1.4 and to Ref. 1 for the symplectic geometry on HILBERT spaces in particular. Let $\mathbb{H}$ be a HILBERT space whose vectors are denoted by ψ,ϕ. The BANACH space of linear, *bounded, and self-adjoint* operators on $\mathbb{H}$ is denoted by $\mathscr{S}_{sa}(\mathbb{H})$. Since $\mathbb{H}$ is a linear space, its tangent space $T_\phi\mathbb{H}$ at $\phi\in\mathbb{H}$ is $\mathbb{H}$ itself and we have for its tangent bundle $T\mathbb{H}\simeq\mathbb{H}\times\mathbb{H}$. Viewing a vector field $X\in\mathscr{X}(\mathbb{H})$ as a section in $T\mathbb{H}$, we write it as $X=\{(\phi,X(\phi))\,|\,\phi\in\mathbb{H}\}\subset\mathbb{H}\times\mathbb{H}$ and call $X(\phi)$ the principal part of X at ϕ. If $\mathscr{F}:\mathbb{H}\to\mathbb{R}$ is a function on $\mathbb{H}$ the action of the vector field $X=(\phi,X(\phi))$ on $\mathscr{F}$ is given by

$$X\cdot\mathscr{F}(\phi) := \frac{d}{ds}\,\mathscr{F}(\phi+sX(\phi))\Big|_{s=0}, \tag{2.35}$$

and the differntial $d\mathscr{F}_\phi$ of $\mathscr{F}$ at $\phi\in\mathbb{H}$ is the map

$$d\mathscr{F}_\phi : T_\phi\mathbb{H}\simeq\mathbb{H}\to\mathbb{R},$$
$$\psi\mapsto\frac{d}{ds}\,\mathscr{F}(\phi+s\psi)\Big|_{s=0}. \tag{2.36}$$

The imaginary part of the scalar product $\langle\cdot\,|\,\cdot\rangle$ in $\mathbb{H}$ (which is assumed anti-linear in the first and linear in the second argument) furnishes us with a constant symplectic structure Ω defined as

$$\Omega_\phi(\psi_1,\psi_2) := -2\hbar\ \mathrm{Im}\langle\psi_1|\psi_2\rangle, \tag{2.37}$$

for $\psi_1,\psi_2\in T_\phi\mathbb{H}\simeq\mathbb{H}$. The associated POISSON bracket is defined as usual.[15]

This symplectic structure leads to a nice identification between its canonical linear transformations and unitary transformations:[1]

$$\Phi:\mathbb{H}\to\mathbb{H},\ \textit{linear},\ \Phi^*\Omega=\Omega\ \Leftrightarrow\ \Phi\ \textit{unitary}. \tag{2.38}$$

We can now combine this with the results of the previous Sections and describe quantum systems whose observables depend explicitly on time as a HAMILTONian system. For simplicity we consider only *smooth* time-dependent operators $\mathbb{R}\ni t\mapsto F(t)\in\mathscr{S}_{sa}(\mathbb{H})$ which are viewed as elements of the linear space $\mathscr{C}_{sa}(\mathbb{H})$ of smooth curves in the BANACH space $\mathscr{S}_{sa}(\mathbb{H})$ of bounded self-adjoint operators on $\mathbb{H}$,

$$\mathscr{C}_{sa}(\mathbb{H}) := C^\infty(\mathbb{R},\mathscr{S}_{sa}(\mathbb{H})), \tag{2.39}$$

such that their time derivative $(\partial F/\partial t)(t)$ exists and is again an element of the space of these curves.

In $\mathscr{C}_{sa}(\mathbb{H})$ we introduce the (natural) LIE-product,

$$[F,G]_{\mathscr{C}_{sa}(\mathbb{H})}(t) := \frac{i}{\hbar}[F(t),G(t)] =: \frac{i}{\hbar}[F,G], \tag{2.40}$$

where the last equation introduces a convenient shorthand notation for the left-hand sides. Equation (2.40) makes $\mathscr{C}_{sa}(\mathbb{H})$ into a LIE-algebra.

Given such a time-dependent operator F it defines the following function on $\mathbb{H}\times\mathbb{R}$:

$$\langle F\rangle : \mathbb{H}\times\mathbb{R}\to\mathbb{R}, \quad (\psi,t)\mapsto\langle\psi|F(t)\psi\rangle, \tag{2.41}$$

which for each t defines the function

$$\langle F\rangle_t : \mathbb{H}\to\mathbb{R}, \quad \psi\mapsto\langle F\rangle(\psi,t). \tag{2.42}$$

Mathematically the definition given for $\langle F\rangle$ in (2.42) makes sense only for bounded F. For unbounded F we have to restrict the domain of $\langle F\rangle$ to the form domain of the quadratic form associated to F (Ref. 24, Chap. VIII.6). Similar to the classical case we may then define the time-dependent vector field generated by $\langle F\rangle$ as

$$X_{\langle F\rangle}(\cdot,t) : \mathbb{H}\to T\mathbb{H}, \quad \psi\mapsto X_{\langle F\rangle}(\psi,t) := X_{\langle F\rangle_t}(\psi), \tag{2.43}$$

where $X_{\langle F\rangle_t}(\psi)$ is given for each t by

$$\imath(X_{\langle F\rangle_t})\Omega = -d\langle F\rangle_t, \tag{2.44}$$

and due to the bilinearity of the generating function this vector field at $\phi\in\mathbb{H}$ and time t turns out to be essentially the application of $F(t)$ to ϕ,

$$X_{\langle F\rangle_t}(\phi) = -\frac{i}{\hbar}F(t)\phi. \tag{2.45}$$

Consequently the HAMILTONian equation satisfied by the flow $\Phi_{\langle F\rangle}(t,t_0)$ of $X_{\langle F\rangle}(\cdot,t)$ is nothing but the *time-dependent* SCHRÖDINGER-equation:

$$\frac{d}{dt}\Phi_{\langle F\rangle}(t,t_0)\phi = -\frac{i}{\hbar}F(t)\Phi_{\langle F\rangle}(t,t_0)\phi, \quad \Phi_{\langle F\rangle}(t_0,t_0)\phi = \phi. \tag{2.46}$$

We will henceforth denote this flow simply by $\Phi_F(t,t_0)$.

The extended quantum mechanical phase space is then constructed in complete analogy to the classical case outlined in Section II C. Essentially this amounts to replacing M by $\mathbb{H}$ in Section II B which yields the extended HILBERT space,

$$\widetilde{\mathbb{H}} := \mathbb{H}\times T^*\mathbb{R} \ni \widetilde{\psi} := (\psi,t,e), \tag{2.47}$$

and the symplectic form

$$\widetilde{\Omega} := \widetilde{\pi}_1^*\Omega - \widetilde{\pi}_2^*(de\wedge dt). \tag{2.48}$$

Analogous to (2.9) and (2.10) we define the lifted $C_L(\widetilde{\mathbb{H}})$ and extended functions,

$$C_E(\widetilde{\mathbb{H}}) \ni \widetilde{\langle F\rangle}(\widetilde{\psi}) := \langle F\rangle(\psi,t) - e = \langle\psi|F(t)\psi\rangle - e, \tag{2.49}$$

such that the analogues of (2.12) and (2.13) become

$$\widetilde{X}_{\langle F\rangle}(\widetilde{\psi}) = -\frac{i}{\hbar}F(t)\psi + \left\langle\psi\left|\frac{\partial F(t)}{\partial t}\psi\right\rangle\right. \frac{\partial}{\partial e}, \tag{2.50}$$

$$\widetilde{X}_{\widetilde{\langle F\rangle}}(\widetilde{\psi})=\widetilde{X}_{\langle F\rangle}(\widetilde{\psi})+\frac{\partial}{\partial t}. \tag{2.51}$$

We also have the following POISSON bracket relations:

$$\widetilde{\mathscr{AD}}_{\widetilde{\langle F\rangle}}(\widetilde{\langle G\rangle}):=\{\widetilde{\langle F\rangle},\widetilde{\langle G\rangle}\}=\left\langle\frac{i}{\hbar}[F,G]+\frac{\partial G}{\partial t}-\frac{\partial F}{\partial t}\right\rangle, \tag{2.52}$$

$$\widetilde{\mathscr{AD}}_{\widetilde{\langle F\rangle}}(\langle G\rangle):=\{\widetilde{\langle F\rangle},\widetilde{\langle G\rangle}\}=\left\langle\frac{i}{\hbar}[F,G]+\frac{\partial G}{\partial t}\right\rangle, \tag{2.53}$$

$$\widetilde{\mathscr{AD}}_{\langle F\rangle}(\langle G\rangle):=\{\widetilde{\langle f\rangle},\widetilde{\langle g\rangle}\}=\left\langle\frac{i}{\hbar}[F,G]\right\rangle. \tag{2.54}$$

As in the classical case (2.20) the flow,

$$\widetilde{\Phi}_{\widetilde{H}}(\tau):\mathbb{R}\times\widetilde{\mathbb{H}}\to\widetilde{\mathbb{H}},$$

$$(\tau,\widetilde{\psi})\mapsto\widetilde{\Phi}(\tau)(\widetilde{\psi}), \tag{2.55}$$

of the vector field $\widetilde{X}_{\widetilde{\langle F\rangle}}$ generated by an extended function $\widetilde{\langle H\rangle}$ ''contains'' the time-dependent flow (2.46):

$$\widetilde{\Phi}_{\widetilde{H}}(\tau)\begin{pmatrix}\psi\\ t\\ e\end{pmatrix}=\begin{pmatrix}\Phi_H(\tau+t,t)\psi\\ \tau+t\\ e+\langle H\rangle(\Phi_H(\tau+t,t)\psi,\tau+t)-\langle H\rangle(\psi,t)\end{pmatrix}. \tag{2.56}$$

E. Time-dependent quantum averaging

In the previous Sections we have developed the general averaging-scheme for time-dependent HAMILTONian systems (Section II C) and have given a HAMILTONian formulation of time-dependent quantum mechanics using bilinear functions $\widetilde{\langle F\rangle}$ on $\widetilde{\mathbb{H}}$ (Section II D). In this Section we develop the formalism necessary to lift these concepts to time-dependent self-adjoint operators alone (rather than their associated function).

For each $F\in\mathscr{C}_{sa}(\mathbb{H})$ we define the following derivations on $\mathscr{C}_{sa}(\mathbb{H})$:

$$\widetilde{\mathscr{AD}}_F:\mathscr{C}_{sa}(\mathbb{H})\to\mathscr{C}_{sa}(\mathbb{H}),$$

$$G\mapsto\frac{i}{\hbar}[F,G], \tag{2.57}$$

$$\widetilde{\mathscr{AD}}_{\widetilde{F}}:\mathscr{C}_{sa}(\mathbb{H})\to\mathscr{C}_{sa}(\mathbb{H}),$$

$$G\mapsto\frac{i}{\hbar}[F,G]+\frac{\partial G}{\partial t}. \tag{2.58}$$

From (2.53) and (2.54) we infer that choice of notation in (2.57) and (2.58) is justified since $\langle\cdot\rangle$ intertwines the two $\widetilde{\mathscr{AD}}$-actions:

$$\widetilde{\mathscr{AD}}_{\widetilde{\langle F\rangle}}(\langle G\rangle)=\langle\widetilde{\mathscr{AD}}_{\widetilde{F}}(G)\rangle, \tag{2.59}$$

$$\widetilde{\mathscr{AD}}_{\langle F\rangle}(\langle G\rangle)=\langle\widetilde{\mathscr{AD}}_F(G)\rangle. \tag{2.60}$$

From (2.56) we see that

$$(\widetilde{\Phi}_{\widetilde{F}}(\tau)^*\langle G\rangle)(\psi,t)=\langle\psi|\Phi_F(\tau+t,t)^\dagger G(\tau+t)\Phi_F(\tau+t,t)\psi\rangle, \tag{2.61}$$

which prompts us to define the following action on $\mathscr{C}_{sa}(\mathbb{H})$ of flows generated by self-adjoint operators:

$$\widetilde{\Phi}_{\widetilde{F}}(\tau)^{*}G(t) := \Phi_F(\tau+t,t)^{\dagger}G(\tau+t)\Phi_F(\tau+t,t). \tag{2.62}$$

An immediate consequence of the definitions (2.58) and (2.62) is

$$\frac{d}{d\tau}\overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(\tau)^{*}G := -\widetilde{\mathscr{AD}}_{\widetilde{F}} \circ \overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(\tau)^{*}G, \quad \overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(0)^{*}G = G, \tag{2.63}$$

which allows us to do averaging for time-dependent operators as in the time-dependent classical case, provided the operator analogues of the equations (2.26)–(2.30) are well defined. Let $F,G \in \mathscr{C}_{sa}(\mathbb{H})$ such that

$$\lim_{T\to\infty} \frac{\overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(T)^{*}G - G}{T} = 0 \tag{2.64}$$

and

$$\overline{G}^{(\widetilde{F})} := \lim_{T\to\infty} \frac{1}{T}\int_0^T d\sigma\ \overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(\sigma)^{*}G \in \mathscr{C}_{sa}(\mathbb{H}), \tag{2.65}$$

$$S^{(\widetilde{F})}(G) := \lim_{T\to\infty} \frac{1}{T}\int_0^T d\tau \int_0^\tau d\sigma\ \overset{-1}{\widetilde{\Phi}}_{\widetilde{F}}(\sigma)^{*}(G-\overline{G}^{(\widetilde{F})}) \in \mathscr{C}_{sa}(\mathbb{H}), \tag{2.66}$$

exist. Then it follows that

$$G - \widetilde{\mathscr{AD}}_{\widetilde{F}}(S^{(\widetilde{F})}(G)) = \overline{G}^{(\widetilde{F})}, \tag{2.67}$$

$$\widetilde{\mathscr{AD}}_{\widetilde{F}}(\overline{G}^{(\widetilde{F})}) = 0. \tag{2.68}$$

Again the proof of these statements makes repeated use of (2.63) and proceeds otherwise as in the classical case in Section II C, but they may also be verified by direct computations. Note that $\overline{G}^{(\widetilde{F})}$ as well as $S^{(\widetilde{F})}(G)$ are elements of $\mathscr{C}_{sa}(\mathbb{H})$ and as such in general time-dependent. In fact, using (2.62) we can display this time-dependence explicitly as

$$\overline{G}^{(\widetilde{F})}(t) = \lim_{T\to\infty} \frac{1}{T}\int_0^T d\sigma\ \Phi_F(t-\sigma,t)^{\dagger}G(t-\sigma)\Phi_F(t-\sigma,t),$$

$$S^{(\widetilde{F})}(G)(t) = \lim_{T\to\infty} \frac{1}{T}\int_0^T d\tau \int_0^\tau d\sigma (\Phi_F(t-\sigma,t)^{\dagger}G(t-\sigma)\Phi_F(t-\sigma,t) - \overline{G}^{(\widetilde{F})}(t)).$$

Moreover, (2.67) and (2.68) may be stated in terms commonly used in quantum mechanics, as follows:

$$G - \frac{i}{\hbar}[F, S^{(\widetilde{F})}(G)] - \frac{\partial S^{(\widetilde{F})}(G)}{\partial t} = \overline{G}^{(\widetilde{F})}, \tag{2.69}$$

$$\frac{i}{\hbar}[F, \overline{G}^{(\widetilde{F})}] + \frac{\partial \overline{G}^{(\widetilde{F})}}{\partial t} = 0. \tag{2.70}$$

It should be noted that even for the case of bounded operators treated here the existence of $S^{(\widetilde{F})}(G)$ as defined in (2.66) is not always assured. General conditions on F and G which imply the validity of the averaging constructions (2.64)–(2.66) are currently being investigated.

F. Transformations on constant time slices

We now also introduce non-autonomous flows depending on some additional parameter ε which will later play the role of the perturbation parameter. The generating functions of these flows, however, are assumed to be only lifted functions, i.e., they do not depend on e, which implies that the canonical transformations they generate do not leave the constant time slices. We begin with the classical situation. Let

$$v:\widetilde{M}\times I\to\mathbb{R},\quad (\widetilde{m},\varepsilon)\mapsto v(m,t,\varepsilon), \tag{2.71}$$

where $I:=]\varepsilon_>,\varepsilon_<[\subset\mathbb{R}$ with some $\varepsilon_>\leqslant 0\leqslant\varepsilon_<$. For each $\varepsilon\in I$ this defines a function $v_\varepsilon\in C_L(\widetilde{M})$,

$$v_\varepsilon:\widetilde{M}\to\mathbb{R},\quad m\mapsto v(m,t,\varepsilon). \tag{2.72}$$

The non-autonomous vector field generated by v is then defined as

$$\widetilde{X}_v(\cdot,\varepsilon):\widetilde{M}\to T\widetilde{M},\quad \widetilde{m}\mapsto\widetilde{X}_v(\widetilde{m},\varepsilon):=\widetilde{X}_{v_\varepsilon}(\widetilde{m}), \tag{2.73}$$

where $\widetilde{X}_{v_\varepsilon}$ is defined in the usual way such that for each $\varepsilon\in I$,

$$\imath(\widetilde{X}_{v_\varepsilon})\widetilde{\omega}=-dv_\varepsilon. \tag{2.74}$$

The non-autonomous flow $\widetilde{\varphi}_v$ of this vector field is the solution of the following initial value problem:

$$\frac{d}{d\varepsilon}\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)(\widetilde{m})=\widetilde{X}_v(\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)(\widetilde{m}),\varepsilon),\quad \widetilde{\varphi}_v(\varepsilon_0,\varepsilon_0)(\widetilde{m})=\widetilde{m}, \tag{2.75}$$

which we assume to be unique. For each t and ε, v also gives a function on M,

$$v_{t,\varepsilon}:M\to\mathbb{R},\quad m\mapsto v(m,t,\varepsilon), \tag{2.76}$$

which generates the vector field $X_{v_{t,\varepsilon}}\in\mathscr{X}(M)$ which in turn generates the non-autonomous ε-flow φ_{v_t} such that

$$\frac{d}{d\varepsilon}\varphi_{v_t}(\varepsilon,\varepsilon_0)(m)=X_{v_{t,\varepsilon}}(\varphi_{v_t}(\varepsilon,\varepsilon_0)(m)),\quad \varphi_{v_t}(\varepsilon_0,\varepsilon_0)(m)=m. \tag{2.77}$$

From (2.12) we see that

$$\widetilde{X}_v(\widetilde{m},\varepsilon)=X_{v_{t,\varepsilon}}(m)+\frac{\partial v_\varepsilon}{\partial t}(m,t)\frac{\partial}{\partial e}, \tag{2.78}$$

and we can make a similar decomposition used to arrive at (2.19) to obtain

$$\frac{d}{d\tau}\begin{pmatrix}\widetilde{\pi}_1\circ\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)((\widetilde{m}))\\ t(\varepsilon,\varepsilon_0)(\widetilde{m})\\ h(\varepsilon,\varepsilon_0)(\widetilde{m})\end{pmatrix}=\begin{pmatrix}X_{v_{t,\varepsilon}}(\widetilde{\pi}_1\circ\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)((\widetilde{m})))\\ 0\\ \dfrac{\partial v_\varepsilon}{\partial t}(m,t)\end{pmatrix}. \tag{2.79}$$

Hence, the action of the extended ε-flow is given by

$$\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)\left(\begin{pmatrix} m \\ t \\ e \end{pmatrix}\right)=\begin{pmatrix} \varphi_{v_t}(\varepsilon,\varepsilon_0)(m) \\ t \\ e+\int_{\varepsilon_0}^{\varepsilon} d\lambda\, \dfrac{\partial v}{\partial t}(\varphi_{v_t}(\lambda,\varepsilon_0)(m),t,\lambda) \end{pmatrix}. \tag{2.80}$$

Evidently, functions of the type (2.71) generate flows which do not leave the constant time slices:

$$\widetilde{\varphi}_v(\varepsilon,\varepsilon_0)(\widetilde{\Sigma}_t)\subset\widetilde{\Sigma}_t, \quad \forall t\in\mathbb{R}. \tag{2.81}$$

In our applications to time-dependent perturbation problems it will be useful to have the above formulae for the following special cases. If for some function v as in (2.71) we define

$$\widetilde{\varphi}(\varepsilon):=\overset{-1}{\widetilde{\varphi}}_{-v}(\varepsilon,0) \quad \text{and} \quad \varphi_t(\varepsilon):=\overset{-1}{\varphi}_{-v_t}(\varepsilon,0), \tag{2.82}$$

then (2.80) becomes

$$\widetilde{\varphi}^{-1}(\varepsilon)\begin{pmatrix} m \\ t \\ e \end{pmatrix}=\begin{pmatrix} \overset{-1}{\varphi}_t(\varepsilon)(m) \\ t \\ e-\int_0^{\varepsilon} d\lambda\, \overset{-1}{\varphi}_t(\lambda)^*\dfrac{\partial v_{t,\lambda}}{\partial t}(m) \end{pmatrix}, \tag{2.83}$$

and for the inverse we have

$$\widetilde{\varphi}(\varepsilon)\begin{pmatrix} m \\ t \\ e \end{pmatrix}=\begin{pmatrix} \varphi_t(\varepsilon)(m) \\ t \\ e+\varphi_t(\varepsilon)^*\int_0^{\varepsilon} d\lambda\, \overset{-1}{\varphi}_t(\lambda)^*\dfrac{\partial v_{t,\lambda}}{\partial t}(m) \end{pmatrix}. \tag{2.84}$$

Note that $\widetilde{\varphi}(\varepsilon)^*$ satisfies

$$\frac{d}{d\varepsilon}\,\widetilde{\varphi}(\varepsilon)^*g=\widetilde{ad}_{v_\varepsilon}\circ\widetilde{\varphi}(\varepsilon)^*g, \quad \widetilde{\varphi}(0)^*g=g, \quad \forall g\in C^\infty(\widetilde{M}). \tag{2.85}$$

Turning to the quantum mechanical case, suppose that we have an operator V on $\mathbb{H}$ which in addition to time also depends on some additional parameter $\varepsilon\in I:=]\varepsilon_>,\varepsilon_<[\subset\mathbb{R}$ with some $\varepsilon_>\leqslant 0\leqslant\varepsilon_<$,

$$V:\mathbb{R}\times I\to\mathscr{L}_{sa}(\mathbb{H}), \quad (t,\varepsilon)\mapsto V(t,\varepsilon). \tag{2.86}$$

Similar to the classical definitions (2.72) and (2.76) we thus define

$$\begin{matrix} V_\varepsilon:\mathbb{R}\to\mathscr{L}_{sa}(\mathbb{H}) \\ t\mapsto V(t,\varepsilon) \end{matrix} \quad \in\mathscr{C}_{sa}(\mathbb{H}), \tag{2.87}$$

and

$$V_{t,\varepsilon}\in\mathscr{L}_{sa}(\mathbb{H}). \tag{2.88}$$

Repeating the constructions of Section II F for the extended quantum mechanical phase space, we obtain for

$$\widetilde{\Phi}(\varepsilon) := \overset{-1}{\widetilde{\Phi}}_{-V}(\varepsilon,0) \quad \text{and} \quad \Phi_t(\varepsilon) := \overset{-1}{\Phi}_{-V_t}(\varepsilon,0) \tag{2.89}$$

the following action on $\widetilde{\mathbb{H}}$:

$$\widetilde{\Phi}(\varepsilon)\begin{pmatrix}\psi\\ t\\ e\end{pmatrix} = \begin{pmatrix}\Phi_t(\varepsilon)\psi\\ t\\ e + \left\langle \Phi_t(\varepsilon)^* \int_0^\varepsilon d\lambda \overset{-1}{\Phi}_t(\lambda)^* \frac{\partial V}{\partial t}\right\rangle\end{pmatrix}. \tag{2.90}$$

Note that $\Phi_t(\varepsilon)$ is the solution of

$$\frac{d}{d\varepsilon}\Phi_t(\varepsilon) = -\frac{i}{\hbar}\Phi_t(\varepsilon)V(t,\varepsilon), \quad \text{and} \quad \Phi_t(0) = \mathbb{1}, \quad \forall t \in \mathbb{R}. \tag{2.91}$$

III. POINCARÉ–VON ZEIPEL PERTURBATION THEORY

A. Introduction

Having established the geometric description of time-dependent systems both classical and quantum mechanical, we now turn to the construction of time-dependent perturbation theories. Since the time evolution of such non-autonomous systems has been formulated as HAMILTONian on extended phase space we might as well ''copy'' the geometric constructions of the time-independent constructions given in Refs. 3–5 and this is essentially what we shall do.

The aim of POINCARÉ–VON ZEIPEL perturbation theory is to provide approximations (or, if possible, exact solutions) of the flow generated by a perturbed HAMILTONian $\widetilde{h(\varepsilon)} = \widetilde{h_0} + \varepsilon h_1 + \ldots$ where $\widetilde{h_0}$ is the unperturbed part whose flow $\varphi_{h_0}(t,t_0)$ is assumed known and ε is the perturbation parameter. This is accomplished with a suitable canonical transformation $\widetilde{\varphi}(\varepsilon)$ to a new HAMILTONian $\widetilde{k(\varepsilon)} = \widetilde{h_0} + \varepsilon k_1 + \cdots$ such that $\widetilde{ad}_{\widetilde{h_0}}(k_p) = 0$. That this equation can be satisfied is assured by the averaging constructions. Moreover, this latter equation allows us to reduce the computation of the flow of $\widetilde{k(\varepsilon)}$ to a *time-independent* problem which in most cases is more amenable to exact solutions. The knowledge of the flow of $\widetilde{k(\varepsilon)}$ and of the canonical transformation $\widetilde{\varphi}(\varepsilon)$ then gives the desired flow of $\widetilde{h(\varepsilon)}$.

For HAMILTONian systems in general, time-dependent POINCARÉ–VON ZEIPEL is constructed in Section III B, whereas the quantum mechanical case is treated in Section III C, and its algorithm is summarized in Section III D.

B. General theory

Let

$$h(\varepsilon) := h_0 + \sum_{p=1}^{\infty} \frac{\varepsilon^p}{p!} h_p : M \times \mathbb{R} \to \mathbb{R},$$

$$(m,t) \mapsto h(\varepsilon)(m,t), \tag{3.1}$$

be a time-dependent HAMILTONian as in (2.1), where the flow $\varphi_{h_0}(t,t_0)$ of the unperturbed (but possibly time-dependent) part h_0 is known. With $h(\varepsilon)$ we construct the extended function

$$\widetilde{h(\varepsilon)} := h(\varepsilon) - e, \tag{3.2}$$

on $\widetilde{M}$. Let v, v_ε, and $v_{t,\varepsilon}$ be as in (2.71), (2.72), and (2.76), respectively and v_ε with an expansion,

$$v_\varepsilon := \sum_{p=0}^{\infty} \frac{\varepsilon^p}{p!} w_{p+1}, \tag{3.3}$$

in terms of ε-independent lifted functions $w_p \in C_L(\widetilde{M})$ and let $\widetilde{\varphi}(\varepsilon)$ and $\varphi_t(\varepsilon)$ be defined as in (2.82). The pull-back action of $\widetilde{\varphi}(\varepsilon)$ satisfies (2.85) and the pull-back of $\varphi_t(\varepsilon)$ which acts on functions on M satisfies

$$\frac{d}{d\varepsilon}\varphi_t(\varepsilon)^* g = ad_{v_{t,\varepsilon}} \circ \varphi_t(\varepsilon)^* g, \quad \varphi_t(0)^* g = g, \quad \forall g \in C^\infty(M), \tag{3.4}$$

and is assumed to permit the expansion

$$\varphi_t(\varepsilon)^* = \sum_{p=0}^{\infty} \frac{\varepsilon^p}{p!} t_p, \tag{3.5}$$

in terms of time-dependent differential operators acting on functions on M which may be computed recursively from

$$t_0 = id_{C^\infty(M)}, \quad t_{p+1} = \sum_{l=0}^{p} \binom{p}{l} ad_{w_{l+1}} \circ t_{p-l}. \tag{3.6}$$

Applying $\widetilde{\varphi}(\varepsilon)^*$ to $\widetilde{h(\varepsilon)}$ we find from (2.84),

$$\begin{aligned}
\widetilde{\varphi}(\varepsilon)^* \widetilde{h(\varepsilon)}(\widetilde{m}) &= h(\varepsilon)(\varphi_t(\varepsilon)(m), t) - \varphi_t(\varepsilon)^* \int_0^\varepsilon d\lambda\, \overset{-1}{\varphi_t}(\lambda)^* \frac{\partial v_{t,\lambda}}{\partial t}(m) - e \\
&=: k(\varepsilon)(m,t) - e \\
&=: \widetilde{k(\varepsilon)}(\widetilde{m}).
\end{aligned} \tag{3.7}$$

This defines the new HAMILTONian,

$$k(\varepsilon) = \varphi_t(\varepsilon)^* \left(h(\varepsilon)(m,t) - \int_0^\varepsilon d\lambda\, \overset{-1}{\varphi_t}(\lambda)^* \frac{\partial v_{t,\lambda}}{\partial t}(m) \right), \tag{3.8}$$

which we assume to have the ε-expansion

$$k(\varepsilon) = h_0 + \sum_{p=1}^{\infty} \frac{\varepsilon^p}{p!} k_p, \tag{3.9}$$

in terms of lifted functions k_p. We thus have

$$\begin{aligned}
\frac{d}{d\varepsilon}k(\varepsilon) &= \left(\frac{d}{d\varepsilon}\varphi_t(\varepsilon)^*\right)\left(h(\varepsilon)(m,t) - \int_0^\varepsilon d\lambda\, \overset{-1}{\varphi_t}(\lambda)^* \frac{\partial v_{t,\lambda}}{\partial t}(m) \right) - \frac{\partial v_{t,\varepsilon}}{\partial t} + \varphi_t(\varepsilon)^* \frac{d}{d\varepsilon} h(\varepsilon) \\
&= ad_{v_{t,\varepsilon}} \circ \varphi_t(\varepsilon)^* h(\varepsilon) - \frac{\partial v_{t,\varepsilon}}{\partial t} + \varphi_t(\varepsilon)^* \frac{d}{d\varepsilon} h(\varepsilon) \\
&= \varphi_t(\varepsilon)^* \frac{d}{d\varepsilon} h(\varepsilon) - \widetilde{ad}_{\widetilde{k(\varepsilon)}}(v_\varepsilon),
\end{aligned} \tag{3.10}$$

and arrive at the following recursive formulae for the k_p:

$$k_0 := h_0, \tag{3.11}$$

$$k_p = f_p - ad_{h_0}(w_p) - \frac{\partial w_p}{\partial t} = f_p - \widetilde{ad}_{\tilde{h}_0}(w_p), \quad \forall p \geqslant 1, \tag{3.12}$$

$$f_1 := h_1, \tag{3.13}$$

$$f_p := h_p + \sum_{l=0}^{p-2} \binom{p-1}{l} (ad_{w_{l+1}}(k_{p-l-1}) + t_{p-l-1} h_{l+1}), \quad \forall p \geqslant 2. \tag{3.14}$$

Here the h_p, k_p, and w_p are lifted functions and the expressions on the right-hand side like $ad_{w_{p+1}}(k_{p-l-1}) = \{(w_p)_t, (k_{p-l-1})_t\}$ are defined as in (2.16). How are we now to choose the w_p? We will now show that the requirement

$$\widetilde{ad}_{\tilde{h}_0}(k_p) = ad_{h_0}(k_p) + \frac{\partial k_p}{\partial t} = 0, \tag{3.15}$$

not only helps to determine the w_p but also implies a useful property for the k_p. Using the results about time-dependent averaging in Section II C we can satisfy (3.15) by the choice

$$w_p = s^{(\tilde{h}_0)}(f_p), \tag{3.16}$$

which also gives us

$$k_p = \overline{f_p}^{(\tilde{h}_0)}, \tag{3.17}$$

where the right hand sides are defined as in (2.27) and (2.28). Note that w_p as well as k_p are in general time-dependent. However, due to their construction, the time dependence of the k_p is essentially trivial since by using (2.15) and (2.13) equation (3.15) becomes

$$\{h_0, k_p\} + \frac{\partial k_p}{\partial t} = \widetilde{X}_{\tilde{h}_0} k_p = 0, \tag{3.18}$$

such that

$$(k_p)_t(m) := k_p(m,t) \tag{3.19}$$

$$= \overset{-1}{\varphi}_{h_0}(t,t_0)^*(k_p)_{t_0}(m), \tag{3.20}$$

for each p and some (arbitrarily chosen) $t_0 \in \mathbb{R}$ and hence

$$(k(\varepsilon))_t(m) := k(\varepsilon)(m,t) \tag{3.21}$$

$$= \overset{-1}{\varphi}_{h_0}(t,t_0)^*(k(\varepsilon))_{t_0}(m), \tag{3.22}$$

for $k(\varepsilon)$ as a whole. Using this we find

$$\frac{d}{dt}(\overset{-1}{\varphi}_{h_0}(t,t_0) \circ \varphi_{k(\varepsilon)}(t,t_0)(m)) = X_{\Sigma_{p=1}^{\infty}(\varepsilon^p/p!)(k_p)_{t_0}}(\overset{-1}{\varphi}_{h_0}(t,t_0) \circ \varphi_{k(\varepsilon)}(t,t_0)(m)), \tag{3.23}$$

where $\varphi_{k(\varepsilon)}(t,t_0)$ is the flow of the non-autonomous vector field generated by $k(\varepsilon)$. On the other hand, t_0 in (3.23) is arbitrary but fixed and $X_{\Sigma_{p=1}^{\infty}(\varepsilon^p/p!)(k_p)_{t_0}}$ *is an autonomous vector field*. Let us suppose then that $\varphi_{(k(\varepsilon)-h_0)_{t_0}}$ is a solution of the autonomous initial value problem,

$$\frac{d}{dt}\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)(m)=X_{\Sigma_{p=1}^{\infty}(\varepsilon^p/p!)\,(k_p)_{t_0}}(\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)(m)),\quad \varphi_{(k(\varepsilon)-h_0)_{t_0}}(t_0)(m)=m. \tag{3.24}$$

From (3.23) and (3.24) we infer that the flow generated by $k(\varepsilon)$ is given by

$$\varphi_{(k(\varepsilon))_t}(t,t_0)=\varphi_{h_0}(t,t_0)\circ\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t). \tag{3.25}$$

Knowing this flow means that we can reconstruct the extended flow,

$$\widetilde{\varphi_{\widetilde{k(\varepsilon)}}}(\tau)\begin{pmatrix}\begin{pmatrix}m\\t\end{pmatrix}\\e\end{pmatrix}=\begin{pmatrix}\varphi_{k(\varepsilon)}(\tau+t,t)(m)\\ \tau+t\\ e+k(\varepsilon)(\varphi_{k(\varepsilon)}(\tau+t,t)(m),\tau+t)-k(\varepsilon)(m,t)\end{pmatrix}, \tag{3.26}$$

generated by $\widetilde{k(\varepsilon)}$. Knowing $\widetilde{\varphi}_{\widetilde{k(\varepsilon)}}(\tau)$ means that we also know $\widetilde{\varphi}_{\widetilde{h(\varepsilon)}}(\tau)$ since $\widetilde{k(\varepsilon)}=\widetilde{\varphi}(\varepsilon)^*\widetilde{h(\varepsilon)}$ implies

$$\widetilde{\varphi}_{\widetilde{h(\varepsilon)}}(\tau)=\widetilde{\varphi}(\varepsilon)\circ\widetilde{\varphi}_{\widetilde{k(\varepsilon)}}(\tau)\circ\widetilde{\varphi}^{-1}(\varepsilon). \tag{3.27}$$

Finally the knowledge of $\widetilde{\varphi}_{\widetilde{h(\varepsilon)}}(\tau)$ implies that we have the desired non-autonomous flow $\varphi_{h(\varepsilon)}(t,t_0)$ simply by taking the projection of the motion in M. Using (2.20), (2.83), (2.84), and (3.25) in (3.27) and taking the projection on M we find then that the desired flow is given by

$$\varphi_{h(\varepsilon)}(t,t_0)=\varphi_t(\varepsilon)\circ\varphi_{h_0}(t,t_0)\circ\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)\circ\varphi_{t_0}^{-1}(\varepsilon). \tag{3.28}$$

Here $\varphi_t(\varepsilon)$ is defined as in (2.82). Equation (3.28) shows that in order to get the desired flow one first transforms on the initial time slice $\widetilde{\Sigma}_{t_0}$, evolves with the autonomous flow $\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)$, then with the *a priori* known (possibly non-autonomous) flow of the unperturbed HAMILTONian, and finally within the final time slice $\widetilde{\Sigma}_t$.

The strategy to find the flow of the time-dependent perturbed HAMILTONian $h(\varepsilon)$ is then as follows:

- construct $\widetilde{\varphi}(\varepsilon)$, i.e., calculate successively the f_p of (3.14), the w_p according to (3.16), and the k_p according to (3.17)
- find $\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)$, i.e., solve the autonomous initial value problem (3.24)
- compute $\varphi_{h(\varepsilon)}(t,t_0)$ as in (3.28).

Note that (3.24) is a time-independent (autonomous) problem and thus may be treated by means of time-independent perturbation techniques given in Refs. 3–5. In general, however, it may not be possible to compute w_p, f_p, k_p for all $p\in\mathbb{N}$ but only up to some finite N, or it may not be possible to solve (3.24) but only up to order $\mathcal{O}(\varepsilon^N)$, i.e., we may have only the following approximations available:

$$\varphi_t^N(\varepsilon):=\varphi_t(\varepsilon)+\mathcal{O}(\varepsilon^N), \tag{3.29}$$

$$\varphi^N_{(k(\varepsilon)-h_0)_{t_0}}(t):=\varphi_{(k(\varepsilon)-h_0)_{t_0}}(t)+\mathcal{O}(\varepsilon^N), \tag{3.30}$$

such that (3.28) has to be modified to

$$\varphi_{h(\varepsilon)}(t,t_0)=\varphi_t^N(\varepsilon)\circ\varphi_{h_0}(t,t_0)\circ\varphi^N_{(k(\varepsilon)-h_0)_{t_0}}(t)\circ\varphi_{t_0}^{N\,-1}(\varepsilon)+\mathcal{O}(\varepsilon^N). \tag{3.31}$$

C. Quantum mechanical case

Suppose now that we are given a time-dependent operator which also depends on the perturbation parameter ε in the usual way,

$$H(\varepsilon)=H_0+\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}H_p\in\mathscr{C}_{sa}(\mathbb{H}), \tag{3.32}$$

where the flow $\Phi_{H_0}(t,t_0)$ generated by the unperturbed (but possibly time-dependent) part H_0 is assumed known. The aim of the quantum mechanical version of time-dependent POINCARÉ–VON ZEIPEL perturbation theory is to find or approximate the time evolution generated by $H(\varepsilon)$.

Let V be an ε and time-dependent operator as in (2.86) and V_ε as in (2.87) with an ε-expansion,

$$V_\varepsilon=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}W_{p+1}\in\mathscr{C}_{sa}(\mathbb{H}), \tag{3.33}$$

in terms of ε-independent time-dependent operators $W_p\in\mathscr{C}_{sa}(\mathbb{H})$. Let $\widetilde{\Phi}(\varepsilon)$ and $\Phi_t(\varepsilon)$ be generated by V (resp., V_t) and defined as in (2.89). For each t and ε, $\Phi_t(\varepsilon)$ defines the transformation

$$\Phi_t(\varepsilon)^*:\mathscr{C}_{sa}(\mathbb{H})\to\mathscr{C}_{sa}(\mathbb{H}),$$

$$F\mapsto\Phi_t(\varepsilon)^\dagger F\Phi_t(\varepsilon). \tag{3.34}$$

Similarly to what was done in Section III B, this transformation is expanded in terms of time-dependent linear operators T_p acting now on $\mathscr{C}_{sa}(\mathbb{H})$,

$$\Phi_t(\varepsilon)^*=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}T_p, \tag{3.35}$$

where the T_p are then recursively defined through $T_0=\mathbb{1}_{\mathscr{C}_{sa}(\mathbb{H})}$ and

$$T_{p+1}=\sum_{l=0}^{p}\binom{p}{l}.\mathscr{AD}_{W_{l+1}}\circ T_{p-l}. \tag{3.36}$$

Moreover, we define

$$\widetilde{\Phi}(\varepsilon)^*:\mathscr{C}_{sa}(\mathbb{H})\to\mathscr{C}_{sa}(\mathbb{H}),$$

$$F\mapsto\widetilde{\Phi}(\varepsilon)^*F, \tag{3.37}$$

where

$$\widetilde{\Phi}(\varepsilon)^*F(t)\coloneqq\Phi_t(\varepsilon)^\dagger\left(F(t)-\int_0^\varepsilon d\lambda\,\Phi_t(\lambda)\,\frac{\partial V}{\partial t}(t,\lambda)\Phi_t(\lambda)^\dagger\right)\Phi_t(\varepsilon) \tag{3.38}$$

$$=\Phi_t(\varepsilon)^*\left(F(t)-\int_0^\varepsilon d\lambda\,\Phi_t(\lambda)\,\frac{\partial V}{\partial t}(t,\lambda)\Phi_t(\lambda)^\dagger\right). \tag{3.39}$$

With the help of this transformation we define the new HAMILTONian,

$$K(\varepsilon)\coloneqq\widetilde{\Phi}(\varepsilon)^*H(\varepsilon)\in\mathscr{C}_{sa}(\mathbb{H}), \tag{3.40}$$

which is assumed to have the usual ε-expansion,

$$K(\varepsilon)=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}K_p, \tag{3.41}$$

in terms of time-dependent operators K_p. Calculations similar to the ones performed in Section III B lead to

$$\frac{d}{d\varepsilon}K(\varepsilon)=\Phi_t(\varepsilon)^*\frac{d}{d\varepsilon}H(\varepsilon)-\mathscr{AD}_{K(\varepsilon)}(V_{t,\varepsilon})-\frac{\partial V_{t,\varepsilon}}{\partial t} \tag{3.42}$$

$$=\Phi_t(\varepsilon)^*\frac{d}{d\varepsilon}H(\varepsilon)-\widetilde{\mathscr{AD}}_{K(\varepsilon)}(V_{t,\varepsilon}), \tag{3.43}$$

and from these we derive the following recursive formulae for the operators K_p:

$$K_0=H_0, \tag{3.44}$$

$$K_p=F_p-\mathscr{AD}_{H_0}(W_p)-\frac{\partial W_p}{\partial t}=F_p-\widetilde{\mathscr{AD}}_{\tilde{H}_0}(W_p),\quad \forall p\geqslant 1, \tag{3.45}$$

$$F_1:=H_1, \tag{3.46}$$

$$F_p:=H_p+\sum_{l=0}^{p-2}\binom{p-1}{l}(\mathscr{AD}_{W_{l+1}}(K_{p-l-1})+T_{p-l-1}H_{l+1}),\quad \forall p\geqslant 2. \tag{3.47}$$

Once again we determine the W_p and thus the K_p by the requirement

$$\widetilde{\mathscr{AD}}_{\tilde{H}_0}(K_p)=\mathscr{AD}_{H_0}(K_p)+\frac{\partial K_p}{\partial t}=\frac{i}{\hbar}[H_0,K_p]+\frac{\partial K_p}{\partial t}=0. \tag{3.48}$$

Using the results about time-dependent averaging in Section II C we can satisfy (3.48) by the choice

$$W_p=S^{(\tilde{H}_0)}(F_p), \tag{3.49}$$

which also gives us

$$K_p=\overline{F_p}^{(\tilde{H}_0)}, \tag{3.50}$$

where the right hand sides are defined as in (2.65) and (2.66). As in the classical case this implies that each time-dependent operator $K_p\in\mathscr{C}_{sa}(\mathbb{H})$ is invariant under the transformation $\tilde{\Phi}_{\tilde{H}_0}(\tau)^*$, $\forall\tau\in\mathbb{R}$ and thus $K(\varepsilon)$ as a whole is also invariant. In particular we may choose a fixed $t_0\in\mathbb{R}$ and set $\tau=t-t_0$ which leads to

$$(K(\varepsilon))_t=\overset{-1}{\Phi}_{H_0}(t,t_0)^\dagger(K(\varepsilon))_{t_0}\overset{-1}{\Phi}_{H_0}(t,t_0) \tag{3.51}$$

$$=(\overset{-1}{\Phi}_{H_0}(t,t_0))^*(K(\varepsilon))_{t_0}, \tag{3.52}$$

the last equation being the equivalent of (3.22). This invariance of $K(\varepsilon)$ implies that

$$\frac{d}{dt}(\overset{-1}{\Phi}_{H_0}(t,t_0)\circ\Phi_{K(\varepsilon)}(t,t_0))=-\frac{i}{\hbar}\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}(K_p)_{t_0}(\overset{-1}{\Phi}_{H_0}(t,t_0)\circ\Phi_{K(\varepsilon)}(t,t_0)), \tag{3.53}$$

where $\Phi_{H_0}(t,t_0)$ and $\Phi_{K(\varepsilon)}(t,t_0)$ are the quantum mechanical time evolutions generated by the time-dependent HAMILTONian operators H_0 and $K(\varepsilon)$. On the other hand $\Sigma_{p=1}^{\infty}(\varepsilon^p/p!)(K_p)_{t_0}$

does not depend on t and, as in the classical case, (3.53) constitutes an autonomous problem. Let us suppose then that $\Phi_{(K(\varepsilon)-H_0)_{t_0}}$ is a solution of the autonomous initial value problem,

$$\frac{d}{dt}\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t)=-\frac{i}{\hbar}\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}(K_p)_{t_0}\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t),\quad \Phi_{(K(\varepsilon)-H_0)_{t_0}}(t_0)=\mathbb{1}_{\mathrm{H}}. \tag{3.54}$$

The time evolution generated by $K(\varepsilon)$ is thus given by

$$\Phi_{K(\varepsilon)}(t,t_0)=\Phi_{H_0}(t,t_0)\circ\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t). \tag{3.55}$$

Recalling (2.56), i.e., that

$$\widetilde{\Phi}_{\widetilde{K(\varepsilon)}}(\tau)\begin{pmatrix}\psi\\ t\\ e\end{pmatrix}=\widetilde{\Phi}_{\widetilde{\langle K(\varepsilon)\rangle}}(\tau)\begin{pmatrix}\psi\\ t\\ e\end{pmatrix}=\begin{pmatrix}\Phi_{K(\varepsilon)}(\tau+t,t)\psi\\ \tau+t\\ e+\langle K(\varepsilon)\rangle(\Phi_{K(\varepsilon)}(\tau+t,t)\psi,\tau+t)-\langle K(\varepsilon)\rangle(\psi,t)\end{pmatrix},$$

we see that a solution of (3.54) allows us to construct the extended flow $\widetilde{\Phi}_{\widetilde{\langle K(\varepsilon)\rangle}}(\tau)$. Using moreover that

$$\widetilde{\langle H(\varepsilon)\rangle}=\widetilde{\langle\widetilde{\Phi}^{-1}(\varepsilon)*K(\varepsilon)\rangle}=\widetilde{\Phi}^{-1}(\varepsilon)*\widetilde{\langle K(\varepsilon)\rangle}, \tag{3.56}$$

we may thus deduce via the following chain of equalities:

$$\widetilde{\Phi}_{\widetilde{H(\varepsilon)}}(\tau)=\widetilde{\Phi}_{\widetilde{\langle H(\varepsilon)\rangle}}(\tau) \tag{3.57}$$

$$=\widetilde{\Phi}^{-1}_{\widetilde{\Phi}(\varepsilon)*\widetilde{\langle K(\varepsilon)\rangle}}(\tau) \tag{3.58}$$

$$=\widetilde{\Phi}(\varepsilon)\circ\widetilde{\Phi}_{\widetilde{\langle K(\varepsilon)\rangle}}(\tau)\circ\widetilde{\Phi}^{-1}(\varepsilon) \tag{3.59}$$

$$=\widetilde{\Phi}(\varepsilon)\circ\widetilde{\Phi}_{\widetilde{K(\varepsilon)}}(\tau)\circ\widetilde{\Phi}^{-1}(\varepsilon), \tag{3.60}$$

that the desired non-autonomous flow $\Phi_{H(\varepsilon)}(t,t_0)$ is given by

$$\Phi_{H(\varepsilon)}(t,t_0)=\Phi_t(\varepsilon)\circ\Phi_{K(\varepsilon)}(t,t_0)\circ\overset{-1}{\Phi}_{t_0}(\varepsilon) \tag{3.61}$$

$$=\Phi_t(\varepsilon)\circ\Phi_{H_0}(t,t_0)\circ\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t,t_0)\circ\overset{-1}{\Phi}_{t_0}(\varepsilon). \tag{3.62}$$

Equation (3.62) describes how $\Phi_{H(\varepsilon)}(t,t_0)$ can be obtained from the knowledge of the W_p and the solution of the autonomous initial value problem (3.54). However, as in the classical case, it may not be possible to compute W_p, F_p, K_p for all $p\in\mathbb{N}$ but only up to some finite $N\in\mathbb{N}$, or it may be impossible to solve (3.54) exactly, but only approximately up to $\mathcal{O}(\varepsilon^N)$, i.e., we may have only the left hand sides of

$$\Phi^N_t(\varepsilon)=\Phi_t(\varepsilon)+\mathcal{O}(\varepsilon^N), \tag{3.63}$$

$$\Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t)=\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t)+\mathcal{O}(\varepsilon^N), \tag{3.64}$$

available, such that we have to modify (3.62) to the approximation

$$\Phi_{H(\varepsilon)}(t,t_0)=\Phi^N_t(\varepsilon)\circ\Phi_{H_0}(t,t_0)\circ\Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t,t_0)\circ\overset{-1}{\Phi}{}^N_{t_0}(\varepsilon)+\mathcal{O}(\varepsilon^N). \tag{3.65}$$

In fact, since in quantum mechanics we are dealing with a linear flow, we can write down an approximation $\Phi_t^N(\varepsilon)$ with the help of the W_p. Let

$$\Phi_t^N(\varepsilon) := \sum_{p=0}^{N} \frac{\varepsilon^p}{p!} \Phi_p, \tag{3.66}$$

where each $\Phi_p \in \mathscr{C}_{sa}(\mathbb{H})$ is, according to (2.91), given by the recursive relation

$$\Phi_{p+1} = -\frac{i}{\hbar} \sum_{l=0}^{p} \binom{p}{l} \Phi_{p-l} W_{l+1}, \quad \Phi_0 = \mathbb{1}_{\mathbb{H}}, \tag{3.67}$$

where only now Φ_p, $W_p \in \mathscr{C}_{sa}(\mathbb{H})$.

In general it will be only the approximation (3.65) which can be computed. As a fairly straightforward example shows (see Section V) this approximation is different from standard time-dependent perturbation theory.

D. Summary of time-dependent quantum mechanical Poincaré–von Zeipel perturbation theory

Below is a summary of the quantum mechanical time-dependent POINCARÉ–VON ZEIPEL perturbation theory developed in Section III C in an algorithmic fashion suitable to immediate application by practitioners.

Let

$$H(\varepsilon) = \sum_{p=0}^{\infty} \frac{\varepsilon^p}{p!} H_p \tag{3.68}$$

be given and suppose $\Phi_{H_0}(t,t_0)$ is known.

Step 1: Starting with $p=1$ do the following loop:

(1) Compute

$$F_p := \begin{cases} H_1, & \text{if } p=1, \\ H_p + \sum_{l=0}^{p-2} \binom{p-1}{l} (\mathscr{AD}_{W_{l+1}}(K_{p-l-1}) + T_{p-l-1} H_{l+1}), & \forall p \geqslant 2 \end{cases} \tag{3.69}$$

where $\mathscr{AD}_{W_{l+1}}(K_{p-l-1}) = i/\hbar\,[W_{l+1}, K_{p-l-1}]$.

(2) Compute

$$K_p = \overline{F_p}^{(H_0)} = \lim_{T\to\infty} \frac{1}{T} \int_0^T d\sigma\; \Phi_{H_0}(t-\sigma,t)^\dagger F_p(t-\sigma) \Phi_{H_0}(t-\sigma,t). \tag{3.70}$$

(3) Compute

$$W_p := S^{(H_0)}(F_p) = \lim_{T\to\infty} \frac{1}{T} \int_0^T d\tau \int_0^\tau d\sigma\; (\Phi_{H_0}(t-\sigma,t)^\dagger F_p(t-\sigma) \Phi_{H_0}(t-\sigma,t) - \overline{F_p}^{(\tilde{H}_0)}(t)). \tag{3.71}$$

(4) Using $T_0 = \mathbb{1}_{\mathscr{C}_{sa}(\mathbb{H})}$ compute

$$T_{p+1} = \sum_{l=0}^{p} \binom{p}{l} \mathscr{AD}_{W_{l+1}} \circ T_{p-l}. \tag{3.72}$$

(5) Using $\Phi_0 = \mathbb{1}_{\mathbb{H}}$ compute

$$\Phi_{p+1} = -\frac{i}{\hbar} \sum_{l=0}^{p-1} \binom{p-1}{l} \Phi_{p-l-1} W_{l+1}, \tag{3.73}$$

until $p = N =$ desired order of approximation.

Step 2: Compute

$$\Phi_t^N(\varepsilon)=\sum_{p=0}^{N}\frac{\varepsilon^p}{p!}\Phi_p(t). \tag{3.74}$$

Step 3: Find the solution $\Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t)$ of the autonomous initial value problem,

$$\frac{d}{dt}\Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t)=\left(-\frac{i}{\hbar}\sum_{p=1}^{N}\frac{\varepsilon^p}{p!}(K_p)_{t_0}\right)\Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t),\quad \Phi^N_{(K(\varepsilon)-H_0)_{t_0}}(t_0)=\mathbb{1}_{\mathbb{H}}. \tag{3.75}$$

Step 4: With the results of Steps 1–3 compute

$$\Phi_t^N(\varepsilon)\circ\Phi_{H_0}(t,t_0)\circ\Phi^N_{(K(\varepsilon)-H_0)(t_0)}(t,t_0)\circ\overset{-1}{\Phi}{}^N_{t_0}(\varepsilon)=\Phi_{H(\varepsilon)}(t,t_0)+\mathcal{O}(\varepsilon^{N+1}). \tag{3.76}$$

This is the desired Nth-order approximation for time evolution generated by the time-dependent HAMILTONian (3.68).

IV. TIME-DEPENDENT KAM ALGORITHM

A. Introduction

Once the time-dependent version of POINCARÉ–VON ZEIPEL is available, it is rather straightforward to construct the KAM method. This is done in analogy to the time-independent case[3–5] by a successive application of canonical transformations $\widetilde{\varphi}^n(\varepsilon)$, $n=1,2,\ldots$, with a suitable choice of generators w_p^n.

The time-dependent KAM method for HAMILTONian systems in general is presented in Section IV B, whereas the quantum mechanical case is given in Section IV C, and its algorithm summarized in Section IV D.

B. General theory

Let $h(\varepsilon)$ be given as in (3.1). Since we are going to transform successively to new HAMILTONians, we shall denote this original perturbed HAMILTONian, whose flow we would like to find (or approximate), by $h^0(\varepsilon):=h(\varepsilon)$ and accordingly $h_p^0:=h_p$ such that $h_0^0=h_0$ is our unperturbed HAMILTONian whose flow is assumed known. Moreover, for $n=1,2,\ldots$, we let v^n, v^n_ε, and $v^n_{t,\varepsilon}$ be as in (2.71), (2.72), and (2.76), respectively, and v^n_ε with an expansion,

$$v^n_\varepsilon:=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}w^n_{p+1}, \tag{4.1}$$

in terms of ε-independent lifted functions $w_p^n\in C_L(\widetilde{M})$ and let $\widetilde{\varphi^n}(\varepsilon)$ and $\varphi_t^n(\varepsilon)$ be defined as in (2.82). The pull-back action of $\widetilde{\varphi^n}(\varepsilon)$ satisfies (2.85) and the pull-back of $\varphi_t^n(\varepsilon)$ which acts on functions on M satisfies

$$\frac{d}{d\varepsilon}\varphi_t^n(\varepsilon)^*g=ad_{v^n_{t,\varepsilon}}\circ\varphi_t^n(\varepsilon)^*g,\quad \varphi_t^n(0)^*g=g,\quad \forall g\in C^\infty(M), \tag{4.2}$$

and is assumed to permit the expansion

$$\varphi_t^n(\varepsilon)^*=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}t_p^n, \tag{4.3}$$

in terms of time-dependent differential operators t_p^n acting on functions on M which may be computed recursively from

$$t_0^n=id_{C^\infty(M)},\quad t_{p+1}^n=\sum_{l=0}^{p}\binom{p}{l}ad_{w^n_{l+1}}\circ t^n_{p-l}. \tag{4.4}$$

Let $\widetilde{k^1(\varepsilon)}$ be the extended HAMILTONian obtained by pulling back the extended original perturbed HAMILTONian $\widetilde{h^0(\varepsilon)}$ with $\widetilde{\varphi^1}(\varepsilon)$,

$$\widetilde{k^1(\varepsilon)} := \widetilde{\varphi^1}(\varepsilon)^* \widetilde{h^0(\varepsilon)}. \tag{4.5}$$

This defines

$$k^1(\varepsilon) = \varphi^1_t(\varepsilon)^* \left(h^0(\varepsilon)(m,t) - \int_0^\varepsilon d\lambda \overset{-1}{\varphi^1_t}(\lambda)^* \frac{\partial v^1_{t,\lambda}}{\partial t}(m) \right), \tag{4.6}$$

which we assume to have the ε-expansion,

$$k^1(\varepsilon) = h^0_0 + \sum_{p=1}^{\infty} \frac{\varepsilon^p}{p!} k^1_p, \tag{4.7}$$

in terms of lifted functions k^1_p. Calculations similar to (3.10) lead to

$$k^1_0 := h^0_0, \tag{4.8}$$

$$k^1_p = f^1_p - ad_{h^0_0}(w^1_p) - \frac{\partial w^1_p}{\partial t} = f^1_p - \widetilde{ad}_{h^0_0}(w_p), \quad \forall p \geqslant 1, \tag{4.9}$$

$$f^1_1 := h^0_1, \tag{4.10}$$

$$f^1_p := h^0_p + \sum_{l=0}^{p-2} \binom{p-1}{l} (ad_{w^1_{l+1}}(k^1_{p-l-1}) + t^1_{p-l-1} h^0_{l+1}), \quad \forall p \geqslant 2. \tag{4.11}$$

So far we have just done POINCARÉ–VON ZEIPEL with $\widetilde{\varphi^1}(\varepsilon)$. Now, however, we shall choose the w^1_p differently:

$$w^1_1 = s^{(\widetilde{h^0_0})}(f^1_1), \tag{4.12}$$

$$w^1_p = 0, \quad \forall p \geqslant 2, \tag{4.13}$$

which implies

$$k^1_1 = \overline{h^0_0}^{(\widetilde{h^0_0})}, \tag{4.14}$$

$$k^1_{p+1} = ad_{w^1_1}(k^1_p) + \sum_{l=0}^{p} \binom{p}{l} (ad_{w^1_1})^{p-l}(h^0_{l+1}), \quad \forall p \geqslant 1. \tag{4.15}$$

After this first transformation we introduce the HAMILTONian $h^1(\varepsilon)$ which is considered as the perturbed HAMILTONian of the unperturbed part $h^0_0 + \varepsilon \overline{h^0_1}^{(\widetilde{h^0_0})}$:

$$h^1(\varepsilon) := k^1(\varepsilon) = \underbrace{h^0_0 + \varepsilon \overline{h^0_1}^{(\widetilde{h^0_0})}}_{} + \sum_{p=2}^{\infty} \frac{\varepsilon^p}{p!} k^1_p \tag{4.16}$$

$$= \quad h^1_0 \quad + \sum_{p=2}^{\infty} \frac{\varepsilon^p}{p!} h^1_p, \tag{4.17}$$

such that

$$h^1_0(\varepsilon) = h^0_0 + \varepsilon \overline{h^0_1}^{(\widetilde{h^0_0})}, \quad h^1_1 = 0, \quad h^1_p = k^1_p, \quad \forall p \geqslant 2, \tag{4.18}$$

and $h^1(\varepsilon)$ has no perturbation of first order in ε. The extended version of (4.16)–(4.17) reads as

$$\widetilde{h^1(\varepsilon)} := \widetilde{k^1(\varepsilon)} = \underbrace{\widetilde{h_0^0} + \varepsilon \overline{h_1^0}^{(\widetilde{h_0^0})}} + \sum_{p=2}^{\infty} \frac{\varepsilon^p}{p!} k_p^1 \tag{4.19}$$

$$= \widetilde{h_0^1} + \sum_{p=2}^{\infty} \frac{\varepsilon^p}{p!} h_p^1. \tag{4.20}$$

By construction $\widetilde{h_0^1}$ is invariant under $\widetilde{\varphi}_{\widetilde{h_0^0}}(\tau)$ which is equivalent to [see the discussion following (3.15)]

$$(h_0^1)_t = \overset{-1}{\varphi}_{h_0^0}(t,t_0)^*(h_0^1)_{t_0}, \tag{4.21}$$

for some $t_0 \in \mathbb{R}$ and thus

$$\frac{d}{dt}(\overset{-1}{\varphi}_{h_0^0}(t,t_0)\circ\varphi_{h_0^1}(t,t_0)(m)) = X_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(\overset{-1}{\varphi}_{h_0^0}(t,t_0)\circ\varphi_{h_0^1}(t,t_0)(m)), \tag{4.22}$$

which helps us to find the flow of h_0^1. Let $\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t)$ be the solution of the *autonomous* initial value problem,

$$\frac{d}{dt}\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t)(m) = X_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t)(m)), \quad \varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t_0)(m) = m; \tag{4.23}$$

then (4.22) implies

$$\varphi_{h_0^1}(t,t_0) = \varphi_{h_0^0}(t,t_0)\circ\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t). \tag{4.24}$$

Now we do a second transformation $\widetilde{\varphi^2}(\varepsilon)$ on $\widetilde{h^1(\varepsilon)}$ such that

$$\widetilde{k^2(\varepsilon)} := \widetilde{\varphi^2}(\varepsilon)^*\widetilde{h^1(\varepsilon)}, \tag{4.25}$$

$$= \widetilde{k_0^2} + \sum_{p=1}^{\infty} \frac{\varepsilon^p}{p!} k_p^2, \tag{4.26}$$

and in the same manner as before one finds that

$$k_0^2 = h_0^1, \tag{4.27}$$

$$k_{p+1}^2 = \sum_{l=0}^{p} \binom{p}{l} (ad_{w_{l+1}^2}(k_{p-l}^2) + t_{p-l}^2(h_{l+1}^1)), \quad \forall p \geqslant 1. \tag{4.28}$$

Here we choose the generating function $v^2(\varepsilon)$ such that

$$w_1^2 = 0, \tag{4.29}$$

$$w_p^2 = s^{(\widetilde{h_0^1})}(h_p^1), \quad p = 2,3, \tag{4.30}$$

$$w_p^2 = 0, \quad \forall p \geqslant 4, \tag{4.31}$$

which implies

$$k_1^2 = 0, \tag{4.32}$$

$$k_2^2=\widetilde{\overline{h_2^1}^{(h_0^1)}}, \tag{4.33}$$

$$k_3^2=\widetilde{\overline{h_3^1}^{(h_0^1)}}, \tag{4.34}$$

$$k_{p+1}^2=\binom{p}{2}ad_{w_2^2}(k_{p-1}^2)+\binom{p}{3}ad_{w_3^2}(k_{p-2}^2)+\sum_{l=1}^{p}\binom{p}{l}t_{p-l}^2(h_{l+1}^1). \tag{4.35}$$

Note that $k_0^2=h_0^1=h_0^0+\varepsilon\widetilde{\overline{h_1^1}^{(h_0^0)}}$ is the unperturbed HAMILTONian after the first step whose flow is found with the help of (4.24) using the solution of (4.23). Moreover, different from the POINCARÉ–VON ZEIPEL algorithm it is now $h_0^1=h_0^0+\varepsilon\widetilde{\overline{h_1^1}^{(h_0^0)}}$ (and *not* h_0^0 as in the POINCARÉ–VON ZEIPEL method) along whose flow we now compute the averages and which is used to construct w_p^2 in (4.29).

After the second transformation we introduce the HAMILTONian $\widetilde{h^2}(\varepsilon)$ which is considered as the perturbed HAMILTONian of the unperturbed part $h_0^1+(\varepsilon^2/2!)\widetilde{\overline{h_2^1}^{(h_0^1)}}+(\varepsilon^3/3!)\widetilde{\overline{h_3^1}^{(h_0^1)}}$:

$$h^2(\varepsilon):=k^2(\varepsilon)=\underbrace{h_0^0+\varepsilon\widetilde{\overline{h_1^0}^{(h_0^0)}}+\frac{\varepsilon^2}{2!}\widetilde{\overline{h_2^1}^{(h_0^1)}}+\frac{\varepsilon^3}{3!}\widetilde{\overline{h_3^1}^{(h_0^1)}}}_{=\,h_0^2}+\sum_{p=2^2}^{\infty}\frac{\varepsilon^p}{p!}k_p^2 \tag{4.36}$$

$$=h_0^2+\sum_{p=2^2}^{\infty}\frac{\varepsilon^p}{p!}h_p^2, \tag{4.37}$$

such that

$$h_0^2(\varepsilon)=h_0^0+\varepsilon\widetilde{\overline{h_1^0}^{(h_0^0)}}+\frac{\varepsilon^2}{2!}\widetilde{\overline{h_2^1}^{(h_0^1)}}+\frac{\varepsilon^3}{3!}\widetilde{\overline{h_3^1}^{(h_0^1)}}, \tag{4.38}$$

$$h_p^2=0, \quad \text{if } 1\leqslant p<2^2, \tag{4.39}$$

$$h_p^1=k_p^1, \quad \text{if } p\geqslant 2^2, \tag{4.40}$$

and $h^2(\varepsilon)$ has no perturbation up to order ε^3. Now one proceeds as before to compute the flow of h_0^2, etc.

The procedure for general n runs as follows.

Starting with $n=1$ we have after $n-1$ transformations the following HAMILTONian:

$$h^{n-1}(\varepsilon)=h_0^{n-1}+\sum_{p=2^{n-1}}^{\infty}\frac{\varepsilon^p}{p!}h_p^{n-1}, \tag{4.41}$$

such that

$$h_p^{n-1}=0, \quad \forall 1\leqslant p<2^{n-1}, \tag{4.42}$$

and the flow $\varphi_{h_0^{n-1}}(t,t_0)$ generated by h_0^{n-1} is known. To the extension of h_0^{n-1} we apply the transformation $\widetilde{\varphi^n}(\varepsilon)$ to obtain

$$\widetilde{k^n(\varepsilon)}=\widetilde{\varphi^n}(\varepsilon)*\widetilde{h^{n-1}(\varepsilon)}, \tag{4.43}$$

$$=\widetilde{k_0^n}+\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}k_p^n. \tag{4.44}$$

The w_p^n generating $\widetilde{\varphi^n}(\varepsilon)$ are chosen as

$$w_p^n=\begin{cases}0, & \text{if } p<2^{n-1},\\ s^{(\widetilde{h_0^{n-1}})}(h_p^{n-1}), & \text{if } 2^{n-1}\leqslant p<2^n,\\ 0, & \text{if } 2^n\leqslant p,\end{cases} \tag{4.45}$$

which implies

$$k_p^n=\begin{cases}h_0^{n-1}, & \text{if } p=0,\\ 0, & \text{if } 1\leqslant p<2^{n-1},\\ \overline{h_p^{n-1}}^{(\widetilde{h_0^{n-1}})}, & \text{if } 2^{n-1}\leqslant p<2^n,\\ \displaystyle\sum_{l=2^{n-1}-1}^{p-1}\binom{p-1}{l}(ad_{w_{l+1}^n}(k_{p-l-1}^n)+t_{p-l-1}^n(h_{l+1}^{n-1})), & \text{if } 2^n\leqslant p.\end{cases} \tag{4.46}$$

Thus we have

$$h^n(\varepsilon)=h_0^n+\sum_{p=2^n}^{\infty}\frac{\varepsilon^p}{p!}h_p^n, \tag{4.47}$$

where

$$h_0^n=h_0^{n-1}+\sum_{p=2^{n-1}}^{2^n-1}\frac{\varepsilon^p}{p!}\overline{h_p^{n-1}}^{(\widetilde{h_0^{n-1}})}, \tag{4.48}$$

$$h_p^n=k_p^n,\quad \forall p\geqslant 2^n. \tag{4.49}$$

Next, solve the autonomous initial value problem,

$$\frac{d}{dt}\varphi_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(t)(m)=X_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(\varphi_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(t)(m)), \tag{4.50}$$

$$\varphi_{(h_0^n(\varepsilon)-h_0^n(\varepsilon))_{t_0}}(t_0)(m)=m. \tag{4.51}$$

This solution permits us to write down $\varphi_{h_0^n}$ explicitly:

$$\begin{aligned}\varphi_{h_0^n}(t,t_0)&=\varphi_{h_0^{n-1}}(t,t_0)\circ\varphi_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(t)\\ &\ \ \vdots\\ &=\varphi_{h_0^0}(t,t_0)\circ\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t)\circ\cdots\circ\varphi_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(t).\end{aligned} \tag{4.52}$$

This way we can find the flow of $h_0^n(\varepsilon)$, but what about the original object of desire, namely the flow of $h^0(\varepsilon)$? Since

$$\widetilde{h^n(\varepsilon)}=\widetilde{\varphi^n}(\varepsilon)*\widetilde{h^{n-1}(\varepsilon)}, \tag{4.53}$$

$$=(\widetilde{\varphi^1}(\varepsilon)\circ\cdots\circ\widetilde{\varphi^n}(\varepsilon))*\widetilde{h^0(\varepsilon)}, \tag{4.54}$$

it follows from arguments similar to those used after (3.27) that

$$\varphi_{h^0(\varepsilon)}(t,t_0)=(\varphi_t^1(\varepsilon)\circ\cdots\circ\varphi_t^n(\varepsilon))\circ\varphi_{h^n(\varepsilon)}(t,t_0)\circ(\varphi_{t_0}^1(\varepsilon)\circ\cdots\circ\varphi_{t_0}^n(\varepsilon))^{-1}, \tag{4.55}$$

and since

$$\varphi_{h^n(\varepsilon)}(t,t_0)=\varphi_{h_0^n(\varepsilon)}(t,t_0)+\mathcal{O}(\varepsilon^{2^n}), \tag{4.56}$$

we deduce

$$\varphi_{h^0(\varepsilon)}(t,t_0)=(\varphi_t^1(\varepsilon)\circ\cdots\circ\varphi_t^n(\varepsilon))\circ\varphi_{h_0^0}(t,t_0)\circ\varphi_{(h_0^1(\varepsilon)-h_0^0)_{t_0}}(t)\circ\cdots\circ\varphi_{(h_0^n(\varepsilon)-h_0^{n-1}(\varepsilon))_{t_0}}(t)$$
$$\circ(\varphi_{t_0}^1(\varepsilon)\circ\cdots\circ\varphi_{t_0}^n(\varepsilon))^{-1}+\mathscr{O}(\varepsilon^{2^n}), \tag{4.57}$$

which constitutes the desired approximation to the flow of $h^0(\varepsilon)$.

C. Quantum mechanical case

Suppose now that, as in Section III C, we are given a time-dependent operator which also depends on the perturbation parameter ε in the usual way,

$$H^0(\varepsilon)=H_0^0+\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}H_p^0\in\mathscr{C}_{sa}(\mathbb{H}), \tag{4.58}$$

where the flow $\Phi_{H_0^0}(t,t_0)$ generated by the unperturbed (but possibly time-dependent) part H_0^0 is assumed known.

For $n=1,2,\ldots$, let V^n be an ε and time-dependent operator as in (2.86) and V_ε^n as in (2.87) with an ε-expansion,

$$V_\varepsilon^n=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}W_{p+1}^n\in\mathscr{C}_{sa}(\mathbb{H}), \tag{4.59}$$

in terms of ε-independent time-dependent operators $W_p^n\in\mathscr{C}_{sa}(\mathbb{H})$. Let $\widetilde{\Phi^n}(\varepsilon)$ and $\Phi_t^n(\varepsilon)$ be generated by V^n (resp., V_t^n) and defined as in (2.89). Each t and ε $\Phi_t^n(\varepsilon)$ defines a transformation $\Phi_t^n(\varepsilon)^*$ on $\mathscr{C}_{sa}(\mathbb{H})$ as given in (3.34). As in (3.35) this transformation is expanded in terms of time-dependent linear operators T_p^n acting on $\mathscr{C}_{sa}(\mathbb{H})$,

$$\Phi_t^n(\varepsilon)^*=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}T_p^n, \tag{4.60}$$

where the T_p^n are then recursively defined through $T_0^n=\mathbb{1}_{\mathscr{C}_{sa}(\mathbb{H})}$ and

$$T_{p+1}^n=\sum_{l=0}^{p}\binom{p}{l}.\mathscr{AD}_{W_{l+1}^n}\circ T_{p-l}^n. \tag{4.61}$$

For each n we define

$$\widetilde{\Phi^n}(\varepsilon)^*:\mathscr{C}_{sa}(\mathbb{H})\to\mathscr{C}_{sa}(\mathbb{H}),$$
$$F\mapsto\widetilde{\Phi^n}(\varepsilon)^*F, \tag{4.62}$$

where

$$\widetilde{\Phi^n}(\varepsilon)^*F(t):=\Phi_t^n(\varepsilon)^\dagger\left(F(t)-\int_0^\varepsilon d\lambda\,\Phi_t^n(\lambda)\,\frac{\partial V^n}{\partial t}(t,\lambda)\Phi_t^n(\lambda)^\dagger\right)\Phi_t^n(\varepsilon), \tag{4.63}$$

$$=\Phi_t^n(\varepsilon)^*\left(F(t)-\int_0^\varepsilon d\lambda\,\Phi_t^n(\lambda)\,\frac{\partial V^n}{\partial t}(t,\lambda)\Phi_t^n(\lambda)^\dagger\right). \tag{4.64}$$

With the help of $\widetilde{\Phi^1}(\varepsilon)$ we define the new HAMILTONian,

$$K^1(\varepsilon):=\widetilde{\Phi^1}(\varepsilon)^*H^0(\varepsilon)\in\mathscr{C}_{sa}(\mathbb{H}), \tag{4.65}$$

which is assumed to have the usual ε-expansion,

$$K^1(\varepsilon)=\sum_{p=0}^{\infty}\frac{\varepsilon^p}{p!}K^1_p, \tag{4.66}$$

in terms of time-dependent operators K^n_p. As shown in Section III C this leads to

$$K^1_0=H^0_0, \tag{4.67}$$

$$K^1_p=F^1_p-\mathscr{AD}_{H_0}(W^1_p)-\frac{\partial W^1_p}{\partial t}=F^1_p-\widetilde{\mathscr{AD}}_{H^0_0}(W^1_p),\quad \forall p\geqslant 1, \tag{4.68}$$

$$F^1_1:=H^0_1, \tag{4.69}$$

$$F^1_p:=H^0_p+\sum_{l=0}^{p-2}\binom{p-1}{l}(\mathscr{AD}_{W^1_{l+1}}(K^1_{p-l-1})+T^1_{p-l-1}H^0_{l+1}),\quad \forall p\geqslant 2. \tag{4.70}$$

As in the general theory the generator of $\widetilde{\Phi^1}(\varepsilon)$ is chosen as

$$W^1_1=S^{(\widetilde{H^0_0})}(F^1_1), \tag{4.71}$$

$$W^1_p=0,\quad \forall p\geqslant 2, \tag{4.72}$$

where $S^{(\widetilde{H^0_0})}(F^1_1)$ is defined as in (2.66). This implies

$$K^1_1=\overline{H^{0(\widetilde{H^0_0})}_0}, \tag{4.73}$$

$$K^1_{p+1}=\mathscr{AD}_{W^1_1}(K^1_p)+\sum_{l=0}^{p}\binom{p}{l}(\mathscr{AD}_{W^1_1})^{p-l}(H^0_{l+1}),\quad \forall p\geqslant 1. \tag{4.74}$$

After this first transformation we introduce the HAMILTONian $H^1(\varepsilon)$ which is considered as the perturbed HAMILTONian of the unperturbed part $H^0_0+\varepsilon\overline{H^{0(\widetilde{H^0_0})}_1}$:

$$H^1(\varepsilon):=K^1(\varepsilon)=\underbrace{H^0_0+\varepsilon\overline{H^{0(\widetilde{H^0_0})}_1}}+\sum_{p=2}^{\infty}\frac{\varepsilon^p}{p!}K^1_p \tag{4.75}$$

$$=\quad H^1_0\quad +\sum_{p=2}^{\infty}\frac{\varepsilon^p}{p!}H^1_p, \tag{4.76}$$

such that

$$H^1_0(\varepsilon)=H^0_0+\varepsilon\overline{H^{0(\widetilde{H^0_0})}_1},\quad H^1_1=0,\quad H^1_p=K^1_p,\quad \forall p\geqslant 2, \tag{4.77}$$

and $H^1(\varepsilon)$ has no perturbation of first order in ε.

By construction $\widetilde{H^1_0}$ is invariant under $\widetilde{\Phi}_{\widetilde{H^0_0}}(\tau)$ which is equivalent to [see the discussion following (3.15)]

$$(H^1_0)_t=\overset{-1}{\Phi}_{H^0_0}(t,t_0)^*(H^1_0)_{t_0}=\Phi_{H^0_0}(t_0,t)^\dagger(H^1_0)_{t_0}\Phi_{H^0_0}(t_0,t), \tag{4.78}$$

for some $t_0\in\mathbb{R}$ and thus

$$\frac{d}{dt}(\overset{-1}{\Phi}_{H^0_0}(t,t_0)\circ\Phi_{H^1_0}(t,t_0))=-\frac{i}{\hbar}(H^1_0(\varepsilon)-H^0_0)_{t_0}(\overset{-1}{\Phi}_{H^0_0}(t,t_0)\circ\Phi_{H^1_0}(t,t_0)). \tag{4.79}$$

Of course, $H_0^1(\varepsilon)-H_0^0=\varepsilon\widetilde{H_1^{0(H_0^0)}}$ but we retain the form on the left-hand side in the expressions here such that their generalization to $n\geqslant 2$ is not obscured. The property (4.79) helps us to find the flow of H_0^1. Let $\Phi_{(H_0^1(\varepsilon)-H_0^0)_{t_0}}(t)$ be the solution of the *autonomous* initial value problem,

$$\frac{d}{dt}\Phi_{(H_0^1(\varepsilon)-H_0^0)_{t_0}}(t)=-\frac{i}{\hbar}(H_0^1(\varepsilon)-H_0^0)_{t_0}(\Phi_{(H_0^1(\varepsilon)-H_0^0)_{t_0}}(t)),\quad \Phi_{(H_0^1(\varepsilon)-H_0^0)_{t_0}}(t_0)=\mathbb{1}_{\mathbb{H}};\tag{4.80}$$

then (4.79) implies

$$\Phi_{H_0^1}(t,t_0)=\Phi_{H_0^0}(t,t_0)\circ\Phi_{(H_0^1(\varepsilon)-H_0^0)_{t_0}}(t).\tag{4.81}$$

Now we do a second transformation $\widetilde{\Phi^2}(\varepsilon)$ on $H^1(\varepsilon)$ such that

$$K^2(\varepsilon):=\widetilde{\Phi^2}(\varepsilon)*H^1(\varepsilon),\tag{4.82}$$

$$=K_0^2+\sum_{p=1}^{\infty}\frac{\varepsilon^p}{p!}K_p^2.\tag{4.83}$$

We choose the generating operator $V^2(\varepsilon)$ of $\widetilde{\Phi^2}(\varepsilon)$ as follows:

$$W_1^2=0,\tag{4.84}$$

$$W_p^2=S^{(\widetilde{H_0^1})}(H_p^1),\quad p=2,3,\tag{4.85}$$

$$W_p^2=0,\quad \forall p\geqslant 4,\tag{4.86}$$

which implies

$$K_0^2=H_0^1,\tag{4.87}$$

$$K_1^2=0,\tag{4.88}$$

$$K_2^2=\widetilde{H_2^{1(H_0^1)}},\tag{4.89}$$

$$K_3^2=\widetilde{H_3^{1(H_0^1)}},\tag{4.90}$$

$$K_{p+1}^2=\binom{p}{2}ad_{W_2^2}(K_{p-1}^2)+\binom{p}{3}ad_{W_3^2}(K_{p-2}^2)+\sum_{l=1}^{p}\binom{p}{l}T_{p-l}^2(H_{l+1}^1).\tag{4.91}$$

Here $K_0^2=H_0^1=H_0^0+\varepsilon\widetilde{H_1^{1(H_0^0)}}$ is the unperturbed Hamiltonian after the first step whose flow is found with the help of (4.81) using the solution of (4.80).

After the second transformation we introduce the HAMILTONian $H^2(\varepsilon)$, which is considered as the perturbed HAMILTONian of the unperturbed part $H_0^1+(\varepsilon^2/2!)\widetilde{H_2^{1(H_0^1)}}+(\varepsilon^3/3!)\widetilde{H_3^{1(H_0^1)}}$:

$$H^2(\varepsilon)=K^2(\varepsilon)=H_0^1+\frac{\varepsilon^2}{2!}\widetilde{H_2^{1(H_0^1)}}+\frac{\varepsilon^3}{3!}\widetilde{H_3^{1(H_0^1)}}+\sum_{p=2^2}^{\infty}\frac{\varepsilon^p}{p!}K_p^2 \tag{4.92}$$

$$= \underbrace{\qquad\qquad H_0^2 \qquad\qquad}+\sum_{p=2^2}^{\infty}\frac{\varepsilon^p}{p!}H_p^2, \tag{4.93}$$

such that $H_p^2=K_p^2$, $\forall p\geqslant 2^2$ and $H^2(\varepsilon)$ has no perturbation of second and third order in ε. Again by construction H_0^2 is invariant under $\widetilde{\Phi}_{H_0^1}(\tau)$ and thus [see the discussion following (3.15)]

$$\frac{d}{dt}(\overset{-1}{\Phi}_{H_0^1}(t,t_0)\circ\Phi_{H_0^2}(t,t_0))=-\frac{i}{\hbar}(H_0^2(\varepsilon)-H_0^1(\varepsilon))_{t_0}(\overset{-1}{\Phi}_{H_0^1}(t,0)\circ\Phi_{H_0^2}(t,t_0)), \tag{4.94}$$

the solution of which permits us to compute

$$\Phi_{H_0^2}(t,t_0)=\Phi_{H_0^1}(t,t_0)\circ\Phi_{(H_0^2(\varepsilon)-H_0^1(\varepsilon))_{t_0}}(t). \tag{4.95}$$

Now transform $H^2(\varepsilon)$ with a third transformation $\widetilde{\Phi^3}(\varepsilon)\ldots$, and so on and so forth.

D. Summary of time-dependent quantum mechanical KAM algorithm

Let us summarize the algorithmic procedure of the time-dependent KAM method. Let N be the desired order of approximation. After $n-1$ transformations we have a HAMILTONian,

$$H^{n-1}(\varepsilon)=H_0^{n-1}(\varepsilon)+\sum_{p=2^{n-1}}^{\infty}\frac{\varepsilon^p}{p!}H_p^{n-1}, \tag{4.96}$$

where we know the time evolution $\Phi_{H_0^{n-1}(\varepsilon)}(t,t_0)$ generated by the unperturbed part after the nth step. Starting with $n=1$ do the following loop.

Step 1: For $2^{n-1}\leqslant p<\min(N,2^n)$ compute

$$S^{\widetilde{(H_0^{n-1})}}(H_p^{n-1})=\lim_{T\to\infty}\frac{1}{T}\int_0^T d\tau\int_0^\tau d\sigma\ \Phi_{H_0^{n-1}(\varepsilon)}(t-\sigma,t)^\dagger H_p^{n-1}(t-\sigma)\Phi_{H_0^{n-1}(\varepsilon)}(t-\sigma,t), \tag{4.97}$$

and set

$$W_p^n=\begin{cases}S^{\widetilde{(H_0^{n-1})}}(H_p^{n-1}), & \text{if } 2^{n-1}\leqslant p<2^n,\\ 0 & \text{otherwise}\end{cases}. \tag{4.98}$$

Step 2: For $p=1,\ldots,N$ compute the T_p^n from $T_0^n=\mathbb{1}_{\mathscr{C}_{sa}(\mathbb{H})}$ and

$$T_{p+1}^n=\sum_{l=0}^{p}\binom{p}{l}.\mathscr{AD}_{W_{l+1}^n}\circ T_{p-l}^n. \tag{4.99}$$

Step 3: For $p=0,\ldots,N$ set

$$K_0^n=H_0^{n-1}\quad\text{and}\quad K_p^n=0,\quad\text{if } 1\leqslant p<2^{n-1}, \tag{4.100}$$

and compute successively the right-hand side of

$$K_p^n=\begin{cases} \overline{H_p^{n-1}(\widetilde{H_0^{n-1}})}, & \text{if } 2^{n-1}\leqslant p<2^n,\\ \displaystyle\sum_{l=2^{n-1}-1}^{p-1}\binom{p-1}{l}(\mathscr{AD}_{W_{l+1}^n}(K_{p-l-1}^n)+T_{p-l-1}^n(H_{l+1}^{n-1})), & \text{if } 2^n\leqslant p\leqslant N,\end{cases} \tag{4.101}$$

where

$$\overline{H_p^{n-1}}(\widetilde{H_0^{n-1}})=\lim_{T\to\infty}\frac{1}{T}\int_0^T d\tau\,\Phi_{H_0^{n-1}(\varepsilon)}(t-\tau,t)^\dagger H_p^{n-1}(t-\tau)\Phi_{H_0^{n-1}(\varepsilon)}(t-\tau,t). \tag{4.102}$$

Step 4: Set

$$H_0^n=H_0^{n-1}+\sum_{p=2^{n-1}}^{2^n-1}\frac{\varepsilon^p}{p!}\overline{H_p^{n-1}}(\widetilde{H_0^{n-1}}), \tag{4.103}$$

$$H_p^n=K_p^n,\quad \text{for } 2^n\leqslant p\leqslant N, \tag{4.104}$$

and find the solution $\Phi_{(H_0^n(\varepsilon)-H^{n-1}(\varepsilon))_{t_0}}(t)$ of the autonomous initial value problem,

$$\frac{d}{dt}\Phi_{(H_0^n(\varepsilon)-H_0^{n-1}(\varepsilon))_{t_0}}(t)=\left(-\frac{i}{\hbar}(H_0^n(\varepsilon)-H_0^{n-1}(\varepsilon))_{t_0}\right)\Phi_{(H_0^n(\varepsilon)-H_0^{n-1}(\varepsilon))_{t_0}}(t),$$

$$\Phi_{(H_0^n(\varepsilon)-H^{n-1}(\varepsilon))_{t_0}}(t_0)=\mathbb{1}_{\mathrm{H}}, \tag{4.105}$$

and compute the flow of $H_0^n(\varepsilon)$ as

$$\Phi_{H_0^n(\varepsilon)}(t,t_0)=\Phi_{H_0^{n-1}(\varepsilon)}(t,t_0)\Phi_{(H_0^n(\varepsilon)-H_0^{n-1}(\varepsilon))_{t_0}}(t). \tag{4.106}$$

Step 5: Find the solution $\Phi_t^n(\varepsilon)$ of

$$\frac{d}{d\varepsilon}\Phi_t^n(\varepsilon)=-\frac{i}{\hbar}V^n(t,\varepsilon)\Phi_t^n(\varepsilon),\quad \Phi_t^n(0)=\mathbb{1}_{\mathrm{H}}, \tag{4.107}$$

and its inverse $\overset{-1}{\Phi}_t(\varepsilon)$ or, if no such solution can be found approximate this solution by computing successively for $p=1,\ldots,N-1$,

$$(\Phi_{p+1}^n)_t=-\frac{i}{\hbar}\sum_{l=0}^{p}\binom{p}{l}(\Phi_{p-l}^n)_t(W_{l+1}^n)_t, \tag{4.108}$$

starting with $(\Phi_0^n)_t=\mathbb{1}_{\mathrm{H}}$. $\Phi_t^n(\varepsilon)$ is then approximated by

$$\Phi_t^n(\varepsilon)=\underbrace{\sum_{p=0}^{N}\frac{\varepsilon^p}{p!}(\Phi_p^n)_t}_{=:\Phi_t^n(\varepsilon)^N}+\mathscr{O}(\varepsilon^{N+1}). \tag{4.109}$$

Step 6: If $N=$desired order$>2^n$ replace n by $n+1$ and start over with *Step 1* otherwise compute the desired approximation to the time-evolution generated by $H(\varepsilon)$ as

$$\Phi_t^1(\varepsilon)\cdots\Phi_t^n(\varepsilon)\Phi_{H_0^0}(t,t_0)\Phi_{(H_0^1-H_0^0)_{t_0}}(t)\cdots\Phi_{(H_0^n-H_0^{n-1})_{t_0}}(t)\overset{-1}{\Phi}{}^n_{t_0}(\varepsilon)\cdots\overset{-1}{\Phi}{}^1_{t_0}(\varepsilon)$$
$$=\Phi_{H(\varepsilon)}(t,t_0)+\mathcal{O}(\varepsilon^{N+1}). \tag{4.110}$$

V. EXAMPLE: SPIN RESONANCE

A. The system

As an illustration of the methods we apply the POINCARÉ–VON ZEIPEL and KAM methods as well as the standard DYSON expansion and a numerical integration to a simple system consisting of a spin $\frac{1}{2}$-particle in an oscillating magnetic field. More precisely, we have $\mathbb{H}=\mathbb{C}^2\ni\psi$ and

$$i\hbar\frac{d}{dt'}\psi(t',t_0')=-\mu\vec{S}\cdot\vec{B}(t')\psi(t',t_0'), \tag{5.1}$$

where μ is some constant with units $s^{-1}\cdot\text{Gauss}^{-1}$, $\vec{S}=(\hbar/2)\vec{\sigma}$ is the spin-operator expressed with the help of the PAULI-matrices σ_j, $j=1,2,3$, and

$$\vec{B}(t')\coloneqq\begin{pmatrix}B_{0x}\cos(\omega t')\\0\\B_{0z}\end{pmatrix},\quad B_{0x},B_{0z}=\text{const} \tag{5.2}$$

is the magnetic field.

Such a system has also been investigated with more standard techniques and in lower orders by SHIRLEY[25] and SAMBE.[17]

We shall only consider the ''initial time'' $t_0'=0$ and use the abbreviation $\psi(t')=\psi(t',0)$. In this system $-\mu B_{0z}S_3$ is the unperturbed part and $-\mu B_{0x}\cos(\omega t')S_1$ is the perturbation whose strength is proportional to B_{0x} and which depends on the frequency ω as an additional parameter.

Using the dimensionless variable $t=\omega t'$, the dimensionless parameters $a\coloneqq-\mu B_{0z}/\omega$ and $\varepsilon\coloneqq-\mu B_{0x}/\omega$, and $\chi(t)\coloneqq\psi(t/\omega)$ the system (5.1) may be written as

$$i\hbar\frac{d}{dt}\chi(t)=(\underbrace{aS_3}_{=:H_0}+\varepsilon\underbrace{S_1\cos(t)}_{=:H_1(t)})\chi(t) \tag{5.3}$$

and it is this system which we will investigate with the approximations presented in the previous Sections. Note that a is inversely proportional to the perturbation frequency ω and that $a=1$ is a resonant case which leads to strongly divergent behavior in all perturbation methods described below.

The unperturbed system is time-independent and its flow is

$$\Phi_{H_0}(t,0)=\exp\left(-\frac{iat}{\hbar}S_3\right)=\begin{pmatrix}e^{(-iat/2)}&0\\0&e^{(iat/2)}\end{pmatrix}. \tag{5.4}$$

Let us now turn to the different perturbation methods applied to the HAMILTONian $H(\varepsilon)=H_0+\varepsilon H_1(t)$.

B. Standard time-dependent perturbation theory

The standard perturbative approach to time-dependent problems of this kind in quantum mechanics can be found in most textbooks on quantum mechanics[26,27] and is commonly referred to as the DYSON expansion.

Suppose we know the flow $\Phi_{H_0}(t,t_0)$ of $H_0\in\mathscr{C}_{sa}(\mathbb{H})$ and would like to find the flow of the perturbed HAMILTONian $H(\varepsilon)\coloneqq H_0+\varepsilon H_1$ where $H_1\in\mathscr{C}_{sa}(\mathbb{H})$ is the perturbation. Time-dependent quantum mechanical perturbation theory consists of an iteration of the integral version of the time-dependent SCHRÖDINGER equation (2.46) in the interaction picture and leads to the following approximation to the time evolution:

$$\Phi_{H(\varepsilon)}(t,t_0)=\Phi_{H_0}(t,t_0)+\sum_{p=1}^{\infty}\varepsilon^p\Phi(t,t_0)^{(p)}_{STAN}, \tag{5.5}$$

where the operator $\Phi(t,t_0)^{(p)}_{STAN}$ is given by

$$\begin{aligned}\Phi(t,t_0)^{(p)}_{STAN}=&\frac{1}{(i\hbar)^p}\int_{t_0}^{t}dt_p\int_{t_0}^{t_p}dt_{p-1}\cdots\int_{t_0}^{t_2}dt_1\Phi_{H_0}(t,t_p)H_1(t_p)\\&\times\Phi_{H_0}(t_p,t_{p-1})\cdots\Phi_{H_0}(t_2,t_1)H_1(t_1)\Phi_{H_0}(t_1,t_0).\end{aligned} \tag{5.6}$$

To illustrate the so called secular terms it is sufficient to consider the following trivial example. Let H_0 be time-independent such that $\Phi_{H_0}(t,t_0)=\exp\{-[i(t-t_0)/\hbar\,]H_0\}$ and $H_1=H_0$. Obviously then

$$\Phi(t,t_0)^{(1)}_{STAN}=\exp\left(-\frac{i(t-t_0)}{\hbar}H_0\right)\left(-\frac{i(t-t_0)}{\hbar}H_0\right), \tag{5.7}$$

which is, of course, the first order term in the expansion,

$$\exp\left(-\frac{i(t-t_0)}{\hbar}(1+\varepsilon)H_0\right)=\exp\left(-\frac{i(t-t_0)}{\hbar}H_0\right)\left(\mathbb{1}-\varepsilon\,\frac{i(t-t_0)}{\hbar}H_0+\cdots\right), \tag{5.8}$$

of the exact solution but which is an unbounded (operator valued) function of t. Hence this method of approximating the true time evolution can only be reliable for fairly short time-differences $t-t_0$ and it will in general *violate unitarity* for sufficiently large times. It is the absence of this feature which makes the perturbation methods based on time-dependent averaging which we have presented in this paper so rather attractive. It is straightforward to compute the approximations $\Phi(t,0)^{(p)}_{STAN}$ in the expansion (5.5) for the example at hand and one finds

$$\begin{aligned}&\Phi(t,0)^{(1)}_{STAN}\\&=\frac{1}{2(a^2-1)}\begin{pmatrix}0 & ae^{(-iat/2)}-(a\,\cos(t)-i\,\sin(t))e^{(iat/2)}\\(a\,\cos(t)+i\,\sin(t))e^{(-iat/2)}-ae^{(iat/2)} & 0\end{pmatrix},\end{aligned} \tag{5.9}$$

for the first order term, still free of secular terms. The second order term has the form

$$\Phi(t,0)^{(2)}_{STAN}=\frac{-1}{16(a^2-1)^2}\begin{pmatrix}\Phi(t,0)^{(2)}_{11} & 0\\0 & \Phi(t,0)^{(2)}_{22}\end{pmatrix}, \tag{5.10}$$

where the matrix entries are

$$\begin{aligned}\Phi(t,0)^{(2)}_{11}=&e^{(-iat/2)}(1+3a^2+4ae^{(iat)}(i\,\sin(t)-a\,\cos(t))\\&+(a^2-1)(\cos(2t)+ia\,\sin(2t)+\underline{2iat})),\end{aligned} \tag{5.11}$$

$$\begin{aligned}\Phi(t,0)^{(2)}_{22}=&e^{(iat/2)}(1+3a^2-4ae^{(-iat)}(i\,\sin(t)+a\,\cos(t))\\&+(a^2-1)(\cos(2t)-ia\,\sin(2t)-\underline{2iat}))\end{aligned} \tag{5.12}$$

and we have underlined the *secular* terms which are proportional to t in this order which will lead to terms proportional to $t^2,t^3,\ldots,$ in higher orders.

C. Time-dependent Poincaré–von Zeipel

Let us now apply the POINCARÉ–VON ZEIPEL algorithm as outlined in Section III D to the system (5.3). Calculating up to $N=2$ one finds easily for the terms in *Step 1*: (in Section III D)

$$K_1 = 0_{2\times 2}, \tag{5.13}$$

$$W_1 = \frac{1}{a^2-1}(a\ \cos(t)S_2 - \sin(t)S_1), \tag{5.14}$$

$$F_2 = \frac{a\ \cos^2(t)}{a^2-1}S_3 = \frac{\cos^2(t)}{a^2-1}H_0, \tag{5.15}$$

$$K_2 = \frac{a}{2(a^2-1)}S_3 = \frac{1}{2(a^2-1)}H_0, \tag{5.16}$$

$$W_2 = \frac{a\ \sin(2t)}{4(a^2-1)}S_3 = \frac{\sin(2t)}{4(a^2-1)}H_0. \tag{5.17}$$

For $\Phi_t^2(\varepsilon)$ in *Step 2* we find

$$\Phi_t^2(\varepsilon) = \left(1 - \frac{\varepsilon^2\ \cos^2(t)}{4} - \frac{1}{4(a^2-1)}\right)\mathbb{1} - \frac{i\varepsilon a\ \cos(t)}{2(a^2-1)}\sigma_2 + \frac{i\varepsilon\ \sin(t)}{2(a^2-1)}\sigma_1 - \frac{i\varepsilon^2 a\ \sin(2t)}{8(a^2-1)}\sigma_3. \tag{5.18}$$

For $N=2$ the solution requested in *Step 3* is

$$\Phi^2_{(K(\varepsilon)-H_0)_0}(t,0) = \begin{pmatrix} e^{[-(iat)/2]-\varepsilon^2\{(iat)/[8(a^2-1)]\}} & 0 \\ 0 & e^{[(iat)/2]+\varepsilon^2\{(iat)/[8(a^2-1)]\}} \end{pmatrix}, \tag{5.19}$$

and it is now straightforward to calculate (though tedious to write down which is why we shall not do so) the second-order approximation to the time evolution as given in *Step 4*:

$$\Phi_{H(\varepsilon)}(t,0) = \Phi_t^2(\varepsilon)\circ\Phi_{H_0}(t,0)\circ\Phi^2_{(K(\varepsilon)-H_0)_0}(t,0)\circ\overset{-1}{\Phi}{}_0^2(\varepsilon) + \mathcal{O}(\varepsilon^3), \tag{5.20}$$

by insertion of the relevant terms from (5.18) and (5.19). From this it also follows that, in contrast to the second order approximation (5.11) calculated in standard perturbation theory, the second order approximation computed from the time-dependent POINCARÉ–VON ZEIPEL *is bounded as a function of* t.

D. KAM algorithm

Let us briefly summarize the results of *Step 1–Step 4* for the KAM algorithm up to $N=2$:

$$\widetilde{\overline{H_1^{0(H_0^0)}}} = 0_{2\times 2}, \tag{5.21}$$

$$W_1^1 = \frac{1}{a^2-1}(a\ \cos(t)S_2 - \sin(t)S_1), \tag{5.22}$$

$$H_2^1 = \frac{a\ \cos^2(t)}{a^2-1}S_3, \tag{5.23}$$

$$\widetilde{\overline{H_2^{1(H_0^1)}}} = \frac{a}{2(a^2-1)}S_3, \tag{5.24}$$

$$H_0^2 = \left(a + \varepsilon^2\frac{a}{4(a^2-1)}\right)S_3, \tag{5.25}$$

$$W_2^2=\frac{a\ \sin(2t)}{4(a^2-1)}S_3, \tag{5.26}$$

$$\Phi_{H_0^1}(t,0)=\Phi_{H_0^0}(t,0), \tag{5.27}$$

$$\Phi_{H_0^2}(t,0)=\begin{pmatrix} e^{[-(iat)/2]-\varepsilon^2\{(iat)/[8(a^2-1)]\}} & 0 \\ 0 & e^{[(iat)/2]+\varepsilon^2\{(iat)/[8(a^2-1)]\}} \end{pmatrix}, \tag{5.28}$$

$$=\Phi_{H_0^0}(t,0)\Phi_{(H_0^1(\varepsilon)-H_0^0)_0}(t)\Phi_{(H_0^2(\varepsilon)-H_0^1(\varepsilon))_0}(t), \tag{5.29}$$

which so far is equivalent to the POINCARÉ–VON ZEIPEL results. However, because of the choice (4.98) for the W_p^1 we can find the exact solution $\Phi_t^1(\varepsilon)$ of (4.107) as

$$\Phi_t^1(\varepsilon)=\cos\left(\frac{\varepsilon\sqrt{a^2\ \cos^2(t)+\sin^2(t)}}{2(a^2-1)}\right)\mathbb{1}-\frac{2i\ \sin\left(\frac{\varepsilon\sqrt{a^2\ \cos^2(t)+\sin^2(t)}}{2(a^2-1)}\right)}{\sqrt{a^2\ \cos^2(t)+\sin^2(t)}}(a\ \cos(t)S_3-\sin(t)S_1). \tag{5.30}$$

Hence, the KAM approximation up to second order,

$$\Phi_{H(\varepsilon)}(t,t_0)=\Phi_t^1(\varepsilon)\Phi_t^2(\varepsilon)\Phi_{H_0^0}(t,0)\Phi_{(H_0^1-H_0^0)_0}(t)\Phi_{(H_0^2-H_0^1)_0}(t)\overset{-1}{\Phi}{}_0^2(\varepsilon)\overset{-1}{\Phi}{}_0^1(\varepsilon)+\mathcal{O}(\varepsilon^3), \tag{5.31}$$

differs from the POINCARÉ–VON ZEIPEL approximation and, as can be seen in Section V E, this difference seems to further improve the quality of the approximation.

E. Comparing the transition-probability

Suppose that at the initial time $t_0=0$ the system is in a spin-up eigenstate,

$$\chi(0)\coloneqq\begin{pmatrix}1\\0\end{pmatrix}. \tag{5.32}$$

We can then compute the state at a later time t,

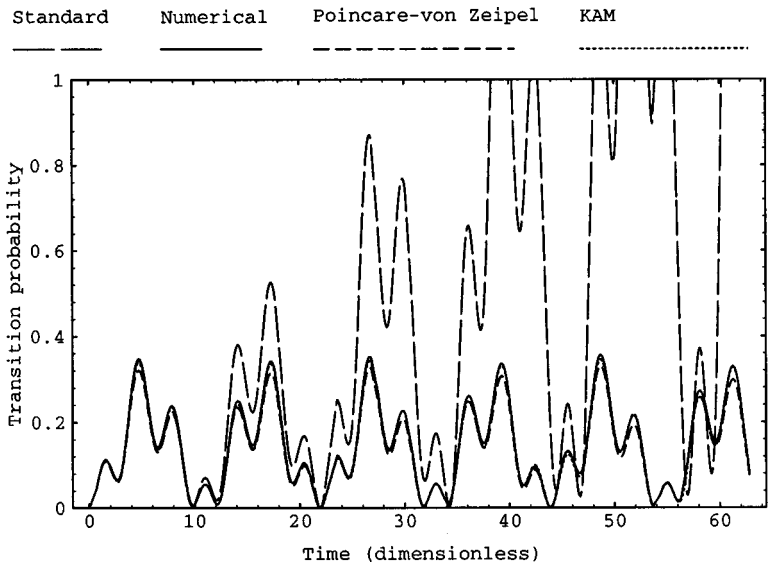

FIG. 2. Transition probability $|\chi_-(t)|^2$ for $a=0.5$ and $\varepsilon=0.7$ computed by different methods.

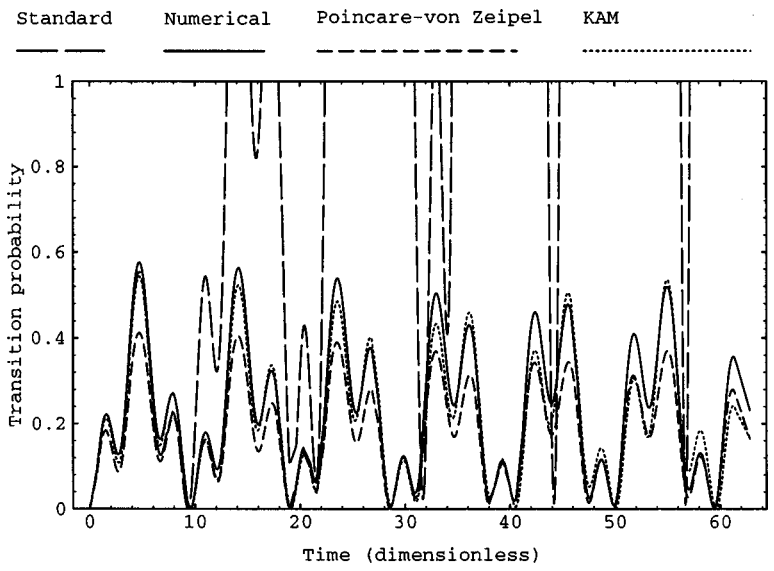

FIG. 3. Transition probability $|\chi_-(t)|^2$ for $a=0.5$ and $\varepsilon=1$ computed by different methods.

$$\chi(t) \coloneqq \begin{pmatrix} \chi_+(t) \\ \chi_-(t) \end{pmatrix} = \Phi_{H(\varepsilon)}(t,0)\chi(0), \tag{5.33}$$

to a desired order in ε with the various approximations to the time evolution given in the previous Sections. In what follows we shall compare the results for the *transition-probability*,

$$P_{+\to-}(t) = |\chi_-(t)|^2, \tag{5.34}$$

from the initial spin-up state to the spin-down state computed up to *fourth order* in ε with the standard, POINCARÉ–VON ZEIPEL and the KAM algorithm and compare it furthermore with a numerical solution of (5.3) using a numerical algorithm contained in the Mathematica package.

Figures 2–4 show $|\chi_-(t)|^2$ for $a=0.5$ and increasing perturbation $\varepsilon=0.7,1,1.3$. Figure 2 shows that the standard perturbation result is quite good for short times $t\leqslant 11$ but then increasingly deviates from the numerical result and violates unitarity due to the presence of secular terms, whereas POINCARÉ–VON ZEIPEL as well as the KAM method reproduce the numerical result with

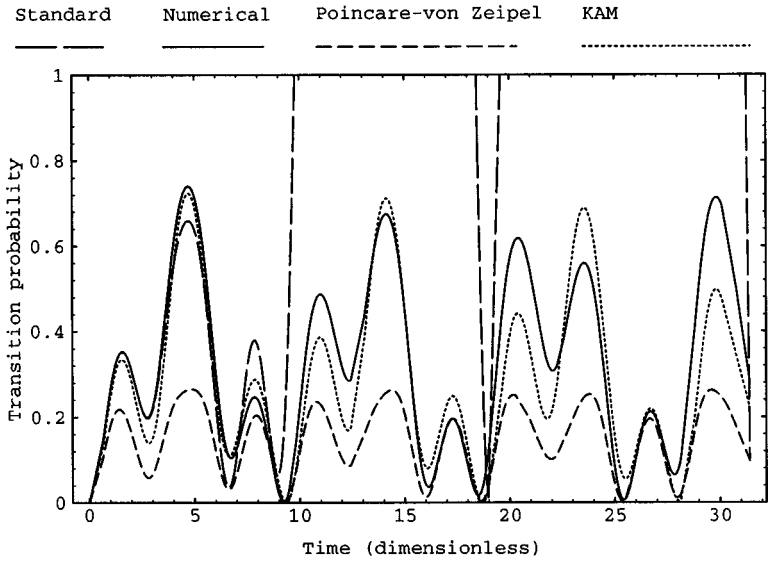

FIG. 4. Transition probability $|\chi_-(t)|^2$ for $a=0.5$ and $\varepsilon=1.3$ computed by different methods.

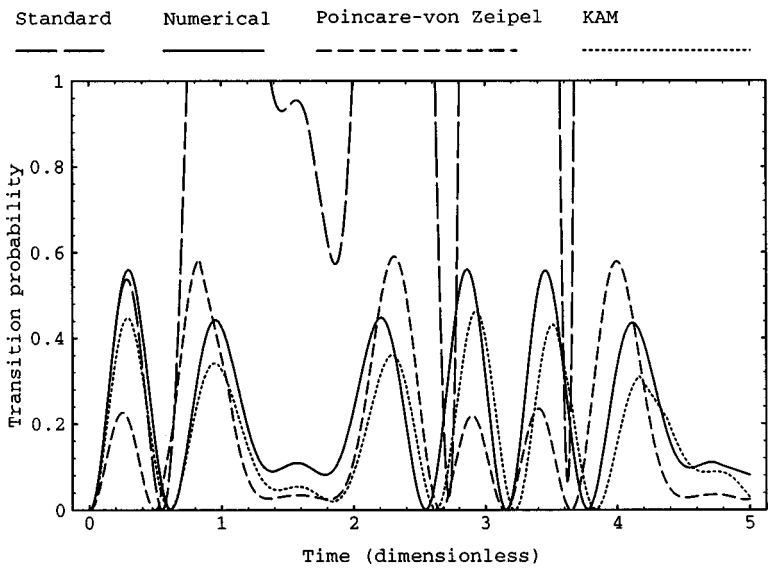

FIG. 5. Transition probability $|\chi_-(t)|^2$ for $a=7$ and $\varepsilon=8$ computed by different methods.

great accuracy and for arbitrarily long times in the case of small perturbation $\varepsilon=0.7$. As the perturbation increases in strength ($\varepsilon=1$ in Fig. 3 and $\varepsilon=1.3$ in Fig. 4) POINCARÉ–VON ZEIPEL deviates more strongly from the numerical result while the KAM method remains still close to it. This behavior is also reproduced for smaller perturbing frequencies (i.e., larger a) as can be seen from Fig. 5. As is evident from all figures, the time-span of reliability for the standard result decreases with increasing perturbation.

VI. CONCLUSION

Using time-dependent averaging we have constructed POINCARÉ–VON ZEIPEL and KAM perturbation methods originating from classical systems for self-adjoint operators. Unlike the time-independent case[3–5] here *both* algorithms give rise to *new* perturbation theories.

The POINCARÉ–VON ZEIPEL algorithm consists of constructing one ε-dependent transformation $\widetilde{\Phi}(\varepsilon)$ on the extended phase space. This transformation is built up successively by an appropriate choice of the generating function V and if projected down to the non-extended phase space it yields different transformations $\Phi_{t_0}(\varepsilon)$ at t_0 and $\Phi_t(\varepsilon)$ at t. The new HAMILTONian $K(\varepsilon)$ is then invariant under the unperturbed flow and its flow can be found by solving a time-independent initial value problem. If we can determine $\Phi_t(\varepsilon)$ exactly and find the flow of $(K(\varepsilon)-H_0)_{t_0}$, then the problem is exactly solved by

$$\Phi_t(\varepsilon)\Phi_{H_0}(t,t_0)\Phi^{-1}_{(K(\varepsilon)-H_0)_{t_0}}(t)\Phi_{t_0}(\varepsilon). \tag{6.1}$$

This, for example, is the case in an application of the theory to a driven harmonic oscillator[28] where one can show inductively that all W_p and K_p vanish for big enough p. In such a case the method tells us immediately that we must have the exact solution.

In general, however, $\Phi_t(\varepsilon)$ cannot be determined exactly, since it may be that not all W_p can be computed or we might not be able to solve (2.91). One may resort to an approximation of the type (3.74) which, however, does not respect unitarity. A unitary approximation to $\Phi_t(\varepsilon)$ can be obtained by attempting to solve

$$\frac{d}{d\varepsilon}\Phi^N_t(\varepsilon)=-\frac{i}{\hbar}\Phi_t(\varepsilon)\sum_{p=0}^{N}\frac{\varepsilon^p}{p!}W_p \quad \text{and} \quad \Phi^N_t(0)=\mathbb{1}, \quad \forall t\in\mathbb{R}. \tag{6.2}$$

Note that in general the approximation (6.1) contains all orders of ε through $\Phi_{(K(\varepsilon)-H_0)_{t_0}}(t)$ [they may also enter by solving (6.2)].

An approximation similar to the POINCARÉ–VON ZEIPEL method has also been constructed by KUMMER and GOMPA using the normal form approach.[8,9] They use this method to transform to a new HAMILTONian by algebraic means and derive rigorous bounds on the deviation from the true time evolution. Their algebraic construction requires solutions of operator equations which are provided by our time-dependent averaging.

The time-dependent KAM method consists of an iterative application of the POINCARÉ–VON ZEIPEL algorithm with a judicious choice of the generators W_p^n. The approximation to the time evolution is then

$$\Phi_t^1(\varepsilon)\cdots\Phi_t^n(\varepsilon)\Phi_{H_0^0}(t,t_0)\Phi_{(H_0^1-H_0^0)_{t_0}}(t)\cdots\Phi_{(H_0^n-H_0^{n-1})_{t_0}}(t)\overset{-1}{\Phi}{}_{t_0}^n(\varepsilon)\cdots\overset{-1}{\Phi}{}_{t_0}^1(\varepsilon). \tag{6.3}$$

Here the particular choice of the W_p^n, as given in (4.98), makes it much easier to solve for the $\Phi_t^n(\varepsilon)$ in

$$\frac{d}{d\varepsilon}\Phi_t^n(\varepsilon)=-\frac{i}{\hbar}\Phi_t(\varepsilon)V^n(t,\varepsilon) \quad \text{and} \quad \Phi_t^n(0)=\mathbb{1}, \quad \forall t\in\mathbb{R}. \tag{6.4}$$

In particular $\Phi_t^1(\varepsilon)$ can always be found by integration. Thus, if all Φ_t^n up to some N can be found and the flows of the autonomous HAMILTONians $(H_0^n-H_0^{n-1})_{t_0}$ [which are simpler than $K(\varepsilon)-H_0$ above as well] can be found, then we have an approximation to the time evolution respecting unitarity.

The applicability of the resulting theories is not confined to just quantum mechanics—or quantum field theory for that matter—but *includes any self-adjoint operator* on a complex HILBERT space for which the solution of a SCHRÖDINGER-type flow equation (2.46) is sought.

These theories are thus *generic* in the sense that they are completely independent of the particular form of the operator and, as far as the general theory is concerned, the phrases ''in quantum mechanics, in the quantum case,'' etc. are synonymous with ''for self-adjoint operators.'' From the physicist's point of view, however, it is indeed quantum theory which comprises the most prominent application of these theories.

The lack of rigorous convergence results is certainly something to be remedied in the future. In particular it will have to be established for which operators the averaging construction (2.64)–(2.66) can be defined rigorously and for which perturbations the resulting series converges to a well defined operator-valued function of ε. It would be futile, however, to insist on such results at this stage. The purpose of this work is to develop the complete formalisms suitable for immediate applications. Only now that these theories are available on the formal level, it becomes fully clear which questions have to be answered by mathematics and which tools might be most suited for that purpose.

[1] P. R. Chernoff and J. E. Marsden, *Properties of Infinite Dimensional Hamiltonian Systems*, Lecture Notes in Mathematics (Springer, Berlin, 1974), Vol. 425.
[2] J. R. Cary, Phys. Rep. **79**, 129 (1981).
[3] W. Scherer, J. Phys. A **27**, 8231 (1994).
[4] W. Scherer, Phys. Rev. Lett. **74**, 1495 (1995).
[5] W. Scherer, J. Phys. A **30**, 2825 (1997).
[6] W. Scherer, in *GROUP 21—Physical Applications and Mathematical Aspeacts of Geometry, Groups, and Algebras*, edited by H. D. Doebner, P. Naltermann, and W. Scherer (World Scientific, Singapore, 1997), Vol. 1.
[7] W. Scherer, Phys. Lett. A **233**, 1 (1997).
[8] M. Kummer and R. Gompa, J. Math. Phys. **29**, 1405 (1988).
[9] R. Gompa, ''Approximations to the quantum mechanical time evolution,'' Ph.D. thesis, University of Toledo, 1987.
[10] J. Bellisard, in *Trends and Developments in the Eighties*, edited by S. Albeverio and P. Blanchard (World Scientific, Singapore, 1985).
[11] J. Bellissard, in *Stochastic Processes in Classical and Quantum Systems*, Lecture Notes in Physics, edited by S. Albeverio, G. Casati, and D. Merlini (Springer, Berlin, 1986), Vol. 262.
[12] M. Combescure, Ann. Inst. Henri Poincaré **47**, 63 (1987).
[13] M. Combescure, Ann. Phys. (N.Y.) **185**, 86 (1988).
[14] P. Duclot and P. Stovicek, Commun. Math. Phys. **177**, 327 (1996).

[15] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, Reading, MA, 1978).
[16] G. Marmo, G. Morandi, and N. Mukunda, Riv. Nuovo Cimento **13**, 1 (1990).
[17] H. Sambe, Phys. Rev. A **7**, 2203 (1973).
[18] J. S. Howland, Math. Ann. **207**, 315 (1974).
[19] J. S. Howland, in *Mathematical Methods and Applications of Scattering Theory*, Lecture Notes in Physics, edited by J. A. Santos (Springer, Berlin, 1980), Vol. 130.
[20] K. Yajima, Commun. Math. Phys. **87**, 331 (1982).
[21] K. Yajima and H. Kitada, Ann. Inst. Henri Poincaré **39**, 145 (1983).
[22] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications* (Addison-Wesley, Reading, MA, 1983).
[23] W. Klingenberg, *Riemannian Geometry* (Walter de Gruyter, Berlin, 1982).
[24] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis* (Academic, New York, 1978).
[25] J. H. Shirley, Phys. Rev. **138**, B979 (1965).
[26] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics II* (Hermann, Paris, 1977).
[27] A. Galindo and P. Pascal, *Quantum Mechanics II* (Springer, Berlin, Heidelberg, New York, 1991).
[28] W. Scherer, *Quantum Averaging* (Habilitationsschrift, TU Clausthal, 1996).

A uniqueness theorem for distributions and its application to nonlocal quantum field theory

M. A. Soloviev[a)]
Department of Theoretical Physics, P. N. Lebedev Physical Institute, Leninsky prosp. 53, Moscow 117924, Russia

(Received 15 July 1997; accepted for publication 5 November 1997)

It is shown that if a distribution $u \in \mathscr{D}'(\mathbb{R}^n)$ has support in a convex cone not containing a straight line and its Fourier transform is carried by a closed cone K different from the whole space, i.e., $\tilde{u} \in \tilde{\mathscr{D}}'(K)$, $K \neq \mathbb{R}_n$, then u is zero. This result leads to a new and most general proof of the PCT theorem which covers nonlocal quantum fields with arbitrary high-energy behavior. © *1998 American Institute of Physics.* [S0022-2488(98)03504-X]

I. INTRODUCTION

In the classical derivation[1,2] of the PCT theorem, quantum fields are assumed to have a finite order of singularity in the space–time variables, or equivalently, to grow in momentum space no faster than a polynomial. Such behavior is inherent in the terms of formal perturbation expansions rather than in exact solutions of relativistic quantum field theories,[3] but this assumption does not give rise to serious omissions since the theorem is directly extendable to Jaffe's strictly localizable theory[4] as well as to the most general formulation of local QFT in terms of hyperfunctions,[5] which allow for the possibility that the off-mass-shell amplitudes behave at high energies worse (the polynomial bound is changed for an exponential one), but leave the analyticity properties of the vacuum expectation values in coordinate space unaltered, and these properties play a leading role in the proof set forth in Refs. 1 and 2. Nevertheless, it is of principal interest to examine the case of arbitrary singularity, when the domain of analyticity is empty, first of all in order to more deeply understand to what extent locality is essential for the existence of the fundamental PCT symmetry. This problem was raised in important papers[6,7] by Lücke who proposed using the momentum-space analyticity properties of the expectation values in overcoming it and analyzed the relevant envelopes of holomorphy. Here we present an alternative proof of the generalized PCT theorem which covers nonlocal fields with arbitrary ultraviolet behavior and which is based on the employment of the analytic wave front set notion[8] and a new type of uniqueness theorem for distributions.

As in Ref. 7, we suppose the fields under study to be averaged with the test functions whose Fourier transform is just the Shwartz space $\mathscr{D}$ of infinitely differentiable functions of compact support, and we use the notation S^0 in place of $\tilde{\mathscr{D}}$, which is intended to stress that this is the smallest space among the Gelfand–Shilov[9] spaces S^β, $0 \leqslant \beta < 1$, traditionally used in nonlocal QFT. (So S^0 gives the widest framework for the treatment of field models with singular behavior in the ultraviolet domain.) The space $S^0(\mathbb{R}^n)$ consists of (the restrictions to $\mathbb{R}^n$ of) entire analytic functions on $\mathbb{C}^n$ with the property that

$$|f(z)| \leqslant C_f (1+|x|)^{-N} e^{b|y|}, \quad (z = x + iy) \tag{1}$$

for some f-dependent $b > 0$ and each $N = 1,2,\ldots$. For definiteness, we take the norm $|\cdot|$ in $\mathbb{R}^n$ to be Euclidean and, as usual, we shall use a prime to label the continuous dual space of a topological vector space. The main problem in dealing with the generalized functions belonging to S'^0 is to correctly generalize the notion of support because its conventional definition breaks down in this case, and so does the local commutativity (microcausality) axiom for the fields with configuration-space test functions in S^0. At this point our definitions and designations differ somewhat from

[a)]Electronic mail: soloviev@td.lpi.ac.ru

 © 1998 American Institute of Physics

those used in Refs. 6 and 7, with the correlation shown in Sec. VI. Namely, first we assign a related space $S^0(U)$ to every open cone $U\subset\mathbb{R}^n$. It consists of all entire functions satisfying the inequality

$$|f(z)|\leqslant C_f(1+|x|)^{-N}e^{b|y|+bd(x,U)}, \tag{2}$$

where $d(\cdot,U)$ is the distance of the point from U, and is regarded as the inductive limit[10] of the family of countably normed spaces $S^{0,b}(U)$ whose norms are defined by

$$\|f\|_{U,b,N}=\sup_z|f(z)|(1+|x|)^N e^{-b|y|-bd(x,U)}. \tag{3}$$

For each closed cone K, we also define a space $S^0(K)$ by taking another inductive limit through those U which contain the set $K\backslash 0$ and shrink to it. A closed cone $K\subset\mathbb{R}^n$ is said to be a *carrier* of a generalized function $w\in S'^0(\mathbb{R}^n)$ if w has a continuous extension to the space $S^0(K)$. We refer to previous papers[11,12] for a motivation of these definitions and for their connection with the theory of hyperfunctions. It is essential in particular that they are in accordance with the usual definition of support of tempered distribution, in the sense that if this support is contained in K, then the restriction of the distribution to S^0 is carried by K. When applied to QFT, the notion of carrier cone leads to the corresponding generalizations of local commutativity and weak local commutativity which we call asymptotic commutativity and weak asymptotic commutativity (WAC), respectively. The main purpose of the present work is to prove the equivalence of the latter condition to the PCT invariance.

In Sec. II, the precise formulations of the WAC condition and of the generalized PCT theorem are given. In Sec. III, we show that the space S^0 is dense in every $S^0(K)$, deriving this fact (important in what follows) as a consequence of an analogous property previously established[13] for the second Gelfand–Shilov scale of test function spaces S^0_α which provides the widest framework for construction of gauge models with singular infrared behavior.[14] In Sec. IV, a relation between the analytic wave front set of Schwartz distributions and the carrier cones of their Fourier transforms is established, which improves Lemma 8.4.17 of Ref. 8. In the next section, it is used in proving a uniqueness theorem which is central in our derivation of the PCT invariance and replaces the uniqueness properties of analytic functions exploited in the usual derivation. The last section is devoted to concluding remarks.

II. GENERALIZED PCT THEOREM

We follow the notation of Ref. 2 and assume that the fields $\varphi_1,\varphi_2,\ldots$ under consideration transform according to general irreducible representations $\mathscr{D}^{(j/2,k/2)}$ of the group $SL(2,\mathbb{C})$ covering the proper Lorentz group. Recall that a point $(\xi_1,\ldots,\xi_n)\in\mathbb{R}^{4n}$, is called a Jost point if the vector $\Sigma\lambda_i\xi_i\in\mathbb{R}^4$ is spacelike for all non-negative numbers λ_i such that $\Sigma\lambda_i>0$. These points form an open cone denoted by J_n. In the conventional formalism dealing with tempered distribution fields, J_n is contained in the analyticity domain of the Wightman functions

$$W_{\mu\ldots\nu}(\xi_1,\ldots,\xi_n)=\langle\Psi_0,\varphi_\mu(x_1)\ldots\varphi_\nu(x_{n+1})\Psi_0\rangle,\quad \xi_i=x_i-x_{i+1}. \tag{4}$$

In these points, all vectors x_i-x_j, $i\neq j$, are spacelike and Jost's version of the PCT theorem says that a field theory (of tempered growth) is PCT invariant if and only if

$$\langle\Psi_0,\varphi_\mu(x_1)\ldots\varphi_\nu(x_{n+1})\Psi_0\rangle=i^F\langle\Psi_0,\varphi_\nu(x_{n+1})\ldots\varphi_\mu(x_1)\Psi_0\rangle, \tag{5}$$

where F is the number of fields with half-integer spins involved in the vacuum expectation value. This relation is referred to as the weak local commutativity (WLC) condition and it takes the form

$$W_{\mu\ldots\nu}(\xi_1,\ldots,\xi_n)-i^F W^{\mathrm{rev}}_{\nu\ldots\mu}(-\xi_n,\ldots,-\xi_1)=0,\quad \xi\in J_n, \tag{6}$$

in terms of the Wightman functions depending on the relative coordinates ξ_i.

Definition: Let $\varphi_1,\varphi_2,\ldots$ be quantum fields with configuration-space test functions in $S^0(\mathbb{R}^4)$. Then we say that they satisfy the weak asymptotic commutativity (WAC) condition if, for any n, the left-hand side of Eq. (6) is carried by the cone $\mathsf{C}J_n$ complementary to the Jost cone.

Theorem 1: A quantum field theory defined over the test function space $S^0(\mathbb{R}^4)$ and satisfying all the Wightman axioms except for locality is PCT invariant if and only if the WAC condition is fulfilled.

Proof: Recall that in terms of the Wightman functions the PCT invariance is expressed by the relation

$$W_{\mu\ldots\nu}(\xi_1,\ldots,\xi_n)=i^F(-1)^J W^{\mathrm{rev}}_{\nu\ldots\mu}(\xi_n,\ldots,\xi_1), \tag{7}$$

where $J=j_1+\cdots+j_n$ is the total number of unpointed indices of the fields involved in the vacuum expectation value. Under the usual temperedness assumption, the conversion WLC$\rightarrow$(7) is carried out in two steps. First, since the proper complex Lorentz group $L_+(\mathbb{C})$ contains the total space–time reflection, the equality $W(\zeta)=(-1)^J W(-\zeta)$ holds in the analyticity domain extended by the Bargman–Hall–Wightman theorem, which enables one to convert the WLC condition to the form (7) in the Jost points. Second, since the distributions occurring in (7) are boundary values of analytic functions from the same tube, this implies their coincidence everywhere by the uniqueness properties of analytic functions. In our case, the analyticity domain is empty. Nevertheless, the difference

$$T=W(\xi)-(-1)^J W(-\xi) \tag{8}$$

is carried by the cone $\mathsf{C}J_n$ due to the Lorentz covariance and by virtue of the spectral condition

$$\mathrm{supp}\ \widetilde{W}\subset \overline{V_+}\times\cdots\times\overline{V_+}\,. \tag{9}$$

One can make sure of this by using the regularization of the high-energy behavior by multiplication with a Lorentz invariant function $h((\Sigma p_i)^2/\Lambda^2)$, where $h\in\mathscr{D}(\mathbb{R}^1)$ and $h(t)=1$ for $|t|\leqslant 1$. Then the generalized function (8) turns into a tempered distribution which has support in $\mathsf{C}J_n$ by the BHW theorem and hence belongs to $S'^0(\mathsf{C}J_n)$. On the other hand, $\widetilde{T}_\Lambda$ coincides with $\widetilde{T}$ in a neighborhood of the origin which enlarges indefinitely as $\Lambda\rightarrow\infty$, and this fact enables T to be extended to the space $S^0(\mathsf{C}J_n)$. However, here the denseness of S^0 in $S^0(U)$ is of prime importance and we put off the details of this extension up to the end of the next section. Writing the difference of the left-hand and right-hand sides of Eq. (7) as

$$\begin{aligned}W_{\mu\ldots\nu}(\xi_1,\ldots,\xi_n)-(-1)^J W_{\mu\ldots\nu}(-\xi_1,\ldots,-\xi_n)+(-1)^J[W_{\mu\ldots\nu}(-\xi_1,\ldots,-\xi_n)\\ -i^F W^{\mathrm{rev}}_{\nu\ldots\mu}(\xi_n,\ldots,\xi_1)],\end{aligned}$$

and taking into account the symmetry property $-J_n=J_n$, we conclude that this difference is carried by $\mathsf{C}J_n$ if the WAC condition is fulfilled. Theorem 4 established below shows that such a carrier property is compatible with the spectral condition (9) only if the difference is zero, which means the PCT invariance. Conversely, the relation (7) implies the fulfillment of the WAC condition by the same argument.

III. RELEVANT APPROXIMATION THEOREM

Theorem 2: For any open cone $U\subset\mathbb{R}^n$ and for each pair of positive numbers b', b such that $b'>2neb$, the space $S^{0,b'}$ is dense in $S^{0,b}(U)$ under the topology of $S^{0,b'}(U)$. As a consequence, S^0 is dense in $S^0(U)$.

An analogous theorem has been proved in Ref. 13 for the Banach spaces $S^{0,b}_{\alpha,a}(U)$, $\alpha>1$, whose norms are defined by a formula obtainable from (3) by the replacement $(1+|x|)^N\rightarrow\exp\{|x/a|^{1/\alpha}\}$. The assertion we now are interested in results from it and from the simple fact that the space $S^0_\alpha(U)=\mathrm{inj\ lim}\, S^{0,b}_{\alpha,a}(U)$ is dense in $S^0(U)$. The latter can be demonstrated in the following way. Let $f\in S^{0,b}(U)$ and $1<\alpha'<\alpha$. We choose a function $g\in S^{0,1}_{\alpha',1}$ such that $g(0)=1$ and consider the sequence $f_\nu(z)=f(z)g(z/\nu)$. If $\epsilon>0$ and $\nu>1/\epsilon$, then

$$|f_\nu(z)|\leqslant C_\nu\ \exp\{-|x|^{1/\alpha}+(b+\epsilon)|y|+bd(x,U)\}$$

and so $f_\nu \in S^{0,b+\epsilon}_{\alpha,1}(U)$. Let us show that under the condition $b' > b+\epsilon$ the sequence f_ν tends to f in every norm of $S^{0,b'}(U)$. We have

$$\|f-f_\nu\|_{U,b',N} \leqslant \|f\|_{U,b,N+1} \sup_z |1-g(z/\nu)|(1+|x|)^{-1}e^{-\epsilon|y|}.$$

Suppose that $\nu > 2/\epsilon$. Then $|1-g(z/\nu)| \leqslant C \exp\{\epsilon|y|/2\}$. Choose R so that the supremum over $|z| \geqslant R$ is less than $\delta\|f\|^{-1}_{U,b,N+1}$, and let ν be so large that the supremum over $|z| < R$ does not exceed the same number. Then $\|f-f_\nu\|_{U,b',N} < \delta$ as was to be shown. By Theorem 5 of Ref. 13, for any $b' > 2ne(b+\epsilon)$ and $a > 1$, the functions f_ν can be approximated in the topology of $S^{0,b'}_{\alpha,a}(U)$ by functions belonging to $S^{0,b'}_{\alpha,a}$, which completes the proof of Theorem 2 since this topology is finer than that of $S^{0,b'}(U)$ and ϵ can be taken arbitrarily small.

Corollary: The generalized function $T \in S'^0(\mathbb{R}^{4n})$ defined by (8) allows a unique continuous extension to the space $S^0(CJ_n)$.

Proof: Let us denote by T_Λ the result of the regularization specified in Sec. II. As already remarked, $T_\Lambda \in S'^0(U)$, with U any open cone such that $(\complement J_n)\backslash 0 \subset U$. By construction, the regularization does not change the value of $\tilde{T}$ on the functions belonging to $\mathscr{D}$ $(|p| < \Lambda/\sqrt{n})$. Set $\Lambda(b) = 10n^{3/2}eb$, $b' = 9neb$ and denote by $\hat{T}_b$ the restriction $T_\Lambda | S^{0,b}(U)$. Then T_Λ coincides with T on the elements of $S^{0,b'}$ since the Fourier transformation converts these into functions having support in the ball of radius b'. Let $f \in S^{0,b}(U)$ and let $B > b$. Applying Theorem 2 and choosing $f_\nu \in S^{0,b'}$ so that $f_\nu \to f$ in $S^{0,b'}(U)$, we get $(\hat{T}_B, f) = \lim(T, f_\nu) = (\hat{T}_b, f)$. By the definition of inductive limit topology, a family of linear forms coordinated in this manner determines an element $\hat{T} \in S'^0(U)$, and so we have obtained the desired result.

IV. ANALYTIC WAVE FRONT SET AND CARRIER CONES

We refer to Hörmander for definition of the analytic wave front set $\mathrm{WF}_A(u)$ of a distribution $u \in \mathscr{D}'(\mathscr{O})$, $\mathscr{O} \in \mathbb{R}^n$. This is a closed subset of $\mathscr{O} \times (\mathbb{R}^n \backslash 0)$ conic relative to the second variable, that is, invariant under multiplication of it by the positive numbers. The projection of $\mathrm{WF}_A(u)$ onto $\mathscr{O}$ consists of those points that have no neighborhood wherein u is an analytic function, and the cone attached to such a point is just the set of ''high-frequency'' directions responsible for the appearance of a singularity at this point, that is, of those relevant directions in which the Fourier transform $\tilde{u}$ behaves badly at infinity.

Recall that if V is a cone in $\mathbb{R}^n$, then its dual cone is defined by $V^* = \{\eta : \eta x \geqslant 0, \text{ for all } x \in V\}$. The cone V^* is closed and convex and has a nonempty interior int V^* if and only if V is proper, that is, the closure of its convex hull does not contain a straight line. A cone V is said to be compact in another cone V' if $\overline{V \backslash 0} \subset V'$, in which case the notation $V \Subset V'$ is used.

Theorem 3: If $u \in \mathscr{D}'$ and $\tilde{u}$ is carried by a closed cone K, then

$$WF_A(u) \subset \mathbb{R}^n \times K. \tag{10}$$

Proof: We denote $\tilde{u}$ by w and at first we suppose that K is properly convex. Under the condition $\mathrm{Im}\ \zeta \in \mathrm{int}\ K^*$, the exponential $e^{iz\zeta}$ belongs to $S^0(K)$ and we can consider the (inverse) Laplace transformation

$$w \to u(\zeta) = (2\pi)^{-n}(w, e^{iz\zeta}).$$

If $\mathrm{Im}\ \zeta$ lies in a cone $V \Subset \mathrm{int}\ K^*$ (we will assume V to be open and convex) and if $|\zeta|$ is bounded, then $u(\zeta)$ satisfies

$$|u(\zeta)| \leqslant C|\mathrm{Im}\ \zeta|^{-N}. \tag{11}$$

This estimate can easily be derived in the regular way analogous to that used, for instance, in proving a Paley–Wiener–Schwartz-type theorem of Ref. 13 for generalized functions of the class S'^0_α. Namely, we choose a pair of open cones U, U' so that $K \Subset U \Subset U' \Subset \mathrm{int}\ V^*$. This is possible because $V \Subset \mathrm{int}\ K^*$ implies $K \Subset \mathrm{int}\ V^*$. For any b and some (b-dependent) N, we have

$$|u(\zeta)| \leqslant (2\pi)^{-n} \|w\|_{U,b,N} \|e^{iz\zeta}\|_{U,b,N} \tag{12}$$

and, with $\zeta = \xi + i\eta$,

$$\|e^{iz\zeta}\|_{U,b,N} = \sup_{x,y} \exp\{-x\eta - y\xi + N \ln(1+|x|) - b|y| - bd(x,U)\}.$$

Let us suppose that $|\xi| < b$ and drop the terms involving y. If $x \notin U'$, then $d(x,U) > \delta|x|$ with some positive δ and, assuming that $|\eta| < b\delta$, the exponent can be majorized by a constant. At last, by virtue of the inclusion $U' \Subset \operatorname{int} V^*$, we have the inequality

$$x\eta \geqslant \delta'|x||\eta|, \quad \text{for all} \quad x \in U', \quad \eta \in V,$$

which immediately yields (11). A simple estimation shows that, under the same restrictions on ζ, the difference quotients corresponding to the partial derivatives $\partial e^{iz\zeta}/\partial\zeta_j$ converge in the topology of $S^{0,b}(U)$ and hence the function $u(\zeta)$ is analytic in $\mathbb{R}^n + \operatorname{int} K^*$. According to Theorem 3.1.15 of Ref. 8 it has a boundary value belonging to $\mathscr{D}'(|\xi| < b)$ which is just (the restriction of) the initial distribution u because the correspondence $S'^0(K) \to \mathscr{D}'(|\xi| < b)$ is continuous due to the factor $\|w\|_{U,b,N}$ in (12) and coincides with the continuous operator $\mathscr{F}^{-1}$ on the dense subspace of $S'^0(K)$ consisting of the tempered distributions supported by K. (Indeed, even the subspace consisting of all finite linear combinations of δ-function and its derivatives is dense in $S'^0(K)$.) Taking into account that b is arbitrarily large and applying Theorem 8.4.8 of the same treatise, we obtain the inclusion $WF_A(u) \subset \mathbb{R}^n \times V^*$ which changes into (10) for this particular case as V approaches K^* since $K^{**} = K$ under the above assumptions.

In the general case, we cover the cone $K\backslash 0$ by a finite number of open circular cones U_j so that their union is close to $K\backslash 0$. Let us denote by $S^{0,b+}(U)$ the countably normed space obtained by the replacement $b \to b + 1/N$ on the right-hand side of the definition (3), whereupon the natural injections $S^{0,b,N+1}(U) \to S^{0,b,N}(U)$ turn out to be compact for any pair of adjacent spaces with the modified norms as may be proved by standard arguments based on the Montel theorem (and similar to those used in deriving Theorem 5.5 of Ref. 12.) As a consequence, the projective limit $S^{0,b+}(U)$ is a Fréchet–Schwartz (FS) space. We claim that the linear functional w regarded as an element of $S'^{0,b+}(\cup U_j)$ allows a decomposition $w = \Sigma w_j$, where $w_j \in S'^{0,b+}(U_j)$. It suffices to make sure of this for the case of two cones. Let us consider the sequence of natural mappings

$$0 \to S^{0,b+}(U_1 \cup U_2) \to S^{0,b+}(U_1) \oplus S^{0,b+}(U_2) \to S^{0,b+}(U_1 \cap U_2), \tag{13}$$

where the last arrow maps a pair (f_1, f_2) into the difference $f_1 - f_2$. The sequence (13) is exact since $d(x, U_1 \cup U_2) = \min\{d(x,U_1), d(x,U_2)\}$. Therefore, the image of $S^{0,b+}(U_1 \cup U_2)$ in the direct sum, being the kernel of a continuous mapping, is closed. So it is a FS space under the topology induced by that of the sum (see, e.g., Komatsu[15]) and, by the open mapping theorem, this topology coincides with the original topology of $S^{0,b+}(U_1 \cup U_2)$. By the Hahn–Banach theorem w has a continuous extension $\hat{w}$ to the whole sum and we obtain the desired decomposition by setting $w(f) = \hat{w}(f,0) + \hat{w}(0,f)$. Now let us consider the Laplace transform of w_j in the domain $|\xi| < \underline{b}, \eta \in V_j \Subset \operatorname{int} U_j^*, |\eta| < b\delta_j$. By the above argument we have the inclusion $\mathrm{WF}_A(u_j) \subset \mathbb{R}^n \times U_j$ for its boundary value $u_j \in \mathscr{D}'(|\xi| \leqslant \underline{b})$. Therefore, the wave front set of the restriction $u|\mathscr{D}(|\xi| < b) = \Sigma u_j$ is contained in $\mathbb{R}^n \times \cup U_j$. Since b can be taken arbitrarily large and $\cup U_j$ is arbitrarily close to K, the proof is finished.

V. UNIQUENESS THEOREM

Theorem 4: Let $u \in \mathscr{D}'(\mathbb{R}^n)$ be a nontrivial distribution whose support is contained in some proper cone V. Then only the whole of $\mathbb{R}^n$ can be a carrier cone of $\tilde{u}$.

Proof: At first we assume that $0 \in \operatorname{supp} u$. Then $(0,x)$ is an external normal to supp u for any $x \in -V^*\backslash 0$ and hence $(0, x_1 - x_2) \in \mathrm{WF}_A(u)$ for all unequal $x_1, x_2 \in V^*$ by a theorem due to Kashiwara, see Ref. 8, Corollary 9.6.8. However $V^* - V^* = \mathbb{R}^n$ since int V^* is nonempty. Therefore, $\tilde{u}$ cannot be carried by a cone different from $\mathbb{R}^n$ by virtue of Theorem 3.

The case $0 \notin \operatorname{supp} u$ may be reduced to the foregoing one by a series of suitable contractions. As before, denote $\tilde{u}$ by w. Let us suppose that a nontrivial carrier exists and so there is an open

cone U such that $\overline{U}\neq\mathbb{R}^n$ and $w\in S'^0(U)$. We may assume without loss of generality that w is bounded in the norm $\|\cdot\|_{U,b,0}$ for each b. In fact, after regularizing by a multiplier $\widetilde{h}_\Lambda(x)=\widetilde{h}(x/\Lambda)$, where $0\leqslant h\in\mathscr{D}(|\xi|<1)$, $\int h\,d\xi=1$, this property is ensured because

$$|(w,\widetilde{h}_\Lambda f)|\leqslant\|w\|_{U,B,N_B}\|\widetilde{h}_\Lambda f\|_{U,B,N_B}\leqslant C\|f\|_{U,b,0}\sup_{x,y}\frac{(1+|x|)^{N_B}}{(1+|x/\Lambda|)^N}\times\exp\{(b-B)(|y|+d(x,U))+|y/\Lambda|\},$$

and the supremum on the right-hand side is finite if $B>b+1/\Lambda$ and $N>N_B$. Furthermore, for Λ large enough, the function $u^{\mathrm{reg}}=u*h_\Lambda$ has support in a proper cone and is nontrivial because it tends to u. Denote $u_\nu(\xi)=u^{\mathrm{reg}}(\nu\xi)$, $w_\nu=\widetilde{u}_\nu$. The homothetic transformation leaves the support of distribution in the same cone and, being an isomorphism of the spaces over cones, does not alter the carrier cones of its Fourier transform. Let us consider the series

$$\sum_{\nu=1}^{\infty}a_\nu w_\nu \tag{14}$$

with positive coefficients subject to the condition

$$a_\nu\leqslant 1/(\nu^2\|w_\nu\|'_{U,\nu,0}), \tag{15}$$

where $\|w\|'_{U,\nu,0}=\sup\{|(w,f)|:f\in S^{0,\nu}(U),\|f\|_{U,\nu,0}\leqslant 1\}$. The dual norm is stronger, the greater is ν and, the strong dual space of a normed space being complete, the condition (15) ensures the absolute convergence of the series (14) in every norm $\|\cdot\|'_{U,b,0}$ and all the more its weak convergence in $S'^0(U)$ to an element which we denote by $\widetilde{\mathscr{U}}$. Now let us show that under some additional restrictions on a_ν the support of $\mathscr{U}$ contains the origin and so we return to the situation examined at first. Assume that the distance of supp u_1 from 0 is equal to 1, and let ξ_0 be its point nearest 0. For each $\mu=1,2,...$, there exists a function $g_\mu\in\mathscr{D}$ with support in the ball $|\xi-\xi_0/\mu|<(1/2)(1/\mu-1/(\mu+1))$ and such that $(u_\mu,g_\mu)=1$. Note that supp $u_\nu\cap$ supp $g_\mu=0$ for $\nu<\mu$. We specify the coefficients a_ν sequentially imposing the conditions

$$a_\nu|(u_\nu,g_\mu)|\leqslant a_\mu/2^\nu\quad\text{for}\quad\mu<\nu \tag{16}$$

in addition to (15). Then for each μ,

$$(\mathscr{U},g_\mu)=a_\mu+\sum_{\nu>\mu}a_\nu(u_\nu,g_\mu)\neq 0,$$

since the sum of the last series does not exceed $a_\mu/2$ in absolute value. Thus $0\in$ supp $\mathscr{U}$ and the proof is completed.

VI. WEAK ESSENTIAL LOCALITY VS WEAK ASYMPTOTIC COMMUTATIVITY

In Ref. 7, Lücke examined the PCT invariance of the theory of a neutral scalar field defined on the test function space $S^0(\mathbb{R}^4)$ and subject to an essential locality condition originally formulated[16] for quantum fields of any class S'^β, $0\leqslant\beta<1$. In conformity to the case $\beta=0$ of particular interest this formulation looks as follows.

(i) Let $Q\subset S^0(\mathbb{R}^n)$ and let M be a closed subset of $\mathbb{R}^n$. Then Q is called locally bounded on M if there is a positive number b such that $\sup\{\|f\|_{M,b,N}:f\in Q\}<\infty$ for every $N=1,2,...$.

(ii) A generalized function $w\in S'^0$ is said to be locally continuous on M (with respect to S^0) if $\sup\{|(w,f)|:f\in Q\}<\infty$ for each locally bounded Q.

(iii) Let φ be a scalar Hermitian field with configuration space test functions in $S^0(\mathbb{R}^4)$. Then φ is called essentially local if the generalized function $\langle\Phi,[\varphi(x),\varphi(x')]\Psi\rangle$ is locally continuous on $\{(x,x')\in\mathbb{R}^8:(x-x')^2\geqslant 0\}$ for all vectors Φ, Ψ belonging to its dense invariant domain in the Hilbert space of states.

Accordingly, the weak essential locality means that the generalized function

$$W(\xi_1,\ldots,\xi_n)-W(-\xi_n,\ldots,-\xi_1)$$

is locally continuous on the cone $\mathsf{C}J_n\subset\mathbb{R}^{4n}$, and this property certainly follows from the essential locality.

It should be noted that the definition of spaces $S^{\beta}(\mathbb{R}^n)$ used in Refs. 7 and 16 is expressed in terms of real variables and outwardly differs from the definition given above in terms of complex variables, but such a reformulation is natural for $\beta<1$ as long ago pointed out by Gelfand and Shilov. In Refs. 11 and 12, we have argued that for $\beta<1$ it is also natural to consider cones instead of arbitrary sets M and then the definition of norms $\|\cdot\|_{M,b,N}$ exploited by Bümmerstede and Lücke admits a similar restatement. Specifically, if M is a closed cone with nonempty interior, then the local boundedness (i) means that Q is contained in some space $S^{0,b}(\mathrm{int}\,M)$ and is bounded in each of its norm determined by (3).

Theorem 5: Let M be a closed cone with nonempty interior U. Then $w\in S'^0$ is locally continuous on M if and only if it is continuous under the topology induced on S^0 by that of $S^0(U)$.

Proof: The boundedness of a linear form on a countably normed space is equivalent to its continuity. Therefore, in standard terminology of the topological vector space theory, the condition (ii) means that every restriction $w|L^b$, where $L^b=S^0\cap S^{0,b}(U)$, is continuous; i.e., it is equivalent to the continuity of w under the topology of inductive limit of the subspaces L^b endowed by the norms of $S^{0,b}(U)$. This topology is not coarser than that induced on S^0 as a subspace of $S^0(U)$ but need not *a priori* coincide with the latter. However, the denseness property established in Theorem 2 ensures their coincidence by the following simple arguments. Let $\mathscr{T}$ and τ denote the topologies in question ($\mathscr{T}\geqslant\tau$) and let $\hat{S}^0(U)$ be the completion of the Hausdorff space $(S^0,\mathscr{T})$. By the definition of inductive limit topology, every mapping $L^b\to\hat{S}^0(U)$ is continuous. Since $L^{b'}$ is dense in $S^{0,b}(U)$, we are able to extend $L^{b'}\to\hat{S}^0(U)$ to a mapping $S^{0,b}(U)\to\hat{S}^0(U)$ continuous under the topology of $S^{0,b'}(U)$ and under that of $S^{0,b}(U)$ all the more. The image of $f\in S^{0,b}(U)$ in $\hat{S}^0(U)$ obtained by the extension exploiting $L^{b''}\supset L^{b'}$ is the same because if $f_\nu\in L^{b'}$ and $f_\nu\to f$ in $S^{0,b'}(U)$, then $f_\nu\to f$ in $S^{0,b''}(U)$. Thus the mappings $S^{0,b}(U)\to\hat{S}^0(U)$ are compatible and determine a homomorphism $S^0(U)\to\hat{S}^0(U)$ which is again continuous by the very definition of inductive topology. Considering its restriction to S^0, we conclude that changing from $\mathscr{T}$ to τ preserves continuity of the canonical mapping $S^0\to\hat{S}^0(U)$. This completes the proof.

Thus, the weak essential locality is a little more restrictive condition than the weak asymptotic commutativity since the space $S^0(\mathrm{int}\,\mathsf{C}J_n)$ is larger than the space $S^0(\mathsf{C}J_n)$ defined with the help of an additional inductive limit operation (whose usefulness is also explained in Refs. 11 and 12), and their dual spaces are inversely related. Therefore, the derivation of the PCT invariance set forth in the present paper covers the essentially local fields.

Theorem 5 was actually used (without a proof) by Lücke[17] in deriving a Roumieu-type representation for the generalized functions locally continuous on M through Borel measures supported by M. When being exploited in Ref. 7, this representation provides a means for the decomposition $w=\Sigma w_j$, $w_j\in S'^0(U_j)$ subordinated to a partition of the quasisupport of w. We have seen in Sec. IV that a weaker version of decomposition ensured by the exactness of the sequence (13) is sufficient to prove the generalized PCT theorem. I do not know any way to prove Theorem 5 without exploiting the denseness property. The point is that subspaces of bornologic spaces (see Ref. 10 for definition) are not necessarily bornologic. This implies that the continuity of a linear form under the topology τ induced on S^0 by that of $S^0(U)$ cannot *a priori* be expressed in simple sequential terms originally used by Gelfand and Shilov for S^0 itself. Proceeding along the same lines as in Lemma 5.2 of Ref. 12 (in particular, using the acyclicity and, as a consequence, the regularity of the injective sequence of spaces $S^{0,b}(U)$, $b=1,2,\ldots$), one can easily show that $\mathscr{T}$ is just the bornologic topology associated with τ but this fact is yet insufficient to be sure that the dual space $(S^0,\mathscr{T})'$ is not larger than $(S^0,\tau)'$. (This is the case if and only if $\mathscr{T}$ is the Mackey topology associated with τ, see Sec. IV. 3. 2 of Ref. 10.) It is also worth noting that, in deriving the Roumieu-type representation, the difficulty connected with two possible topologiza-

tions of subspaces of inductive limits occurs once again, this time in embedding $S^0(U)$ into a space of multi-index sequences of continuous functions by the correspondence $f \to f_q = \partial^q f$, as is clear from some Komatsu's considerations[18] applied at this point. The simple denseness arguments are unsuitable now since this mapping is of closed range and, in my opinion, this matter calls for a more close examination.

VII. CONCLUDING REMARKS

Analogs of Theorems 3 and 4 are valid for the ultradistributions defined on the spaces S_0^α, $\alpha > 1$. In some respects their proofs are even easier since one can apply a decomposition theorem[19] and a Paley–Wiener–Schwartz-type theorem previously established in Ref. 13. Moreover, in deriving the relevant Roumieu-type representation one can easily obviate the difficulty mentioned above because the topological aspect is simpler in this case. However, some new specific problems present themselves and we shall discuss them in an accompanying work for these results taken together with the previous ones give a consistent analytical formalism for construction and treatment of gauge models in a generic covariant gauge.

In the present paper, we have restricted ourselves to the generalized functions of class S'^0 primarily because these are traditionally used in nonlocal QFT and theoretical physicists are familiar with the Schwartz distributions into which S'^0 is converted by the Fourier transformation. From the mathematical point of view, it is also of interest that the employment of the presheaf of spaces $S^0(U)$ over cones enables one to improve considerably the well-known relation between the wave front set of a tempered distribution and the support property of its Fourier transform. A more precise version of decomposition as well as a PWS-type theorem can actually be derived in this case too but with the help of more sophisticated tools which have luckily turned out unnecessary for the purpose of this paper. Finally, it should be noted that the approach proposed here is applicable as well to the derivation of the connection between spin and statistics for the quantum fields smeared with test functions in S^0 (or S_α^0) and satisfying the asymptotic commutativity condition. Further applications, e.g., to generalization of the Borchers equivalence concept, are also possible and such studies are in progress.

ACKNOWLEDGMENTS

I would like to thank Professor V. Ya. Fainberg for helpful discussions. This work was supported in part by RFBR under Contract No. 96-01-00105 and in part by INTAS Grant No. 96-0308.

[1] R. Jost, *The General Theory of Quantum Fields* (American Mathematical Society, Providence, Rhode Island, 1965).
[2] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, Reading, 1964).
[3] A. S. Wightman, Adv. Math. Suppl. Stud. **7B**, 769 (1981).
[4] A. Jaffe, Phys. Rev. **158**, 1454 (1967).
[5] E. Bruning and S. Nagamachi, J. Math. Phys. **30**, 2340 (1989).
[6] W. Lücke, Commun. Math. Phys. **65**, 77 (1979).
[7] W. Lücke, J. Math. Phys. **27**, 1901 (1986).
[8] L. Hörmander, *The Analysis of Linear Partial Differential Operators* (Springer, Berlin, 1983), Vol. 1.
[9] I. M. Gelfand and G. E. Shilov, *Generalized Functions* (Academic, New York, 1964), Vol. 2.
[10] H. H. Schaefer, *Topological Vector Spaces* (Springer, New York, 1970).
[11] V. Ya. Fainberg and M. A. Soloviev, Theor. Math. Phys. **93**, 1438 (1992).
[12] M. A. Soloviev, in *Developments in Mathematics: The Moscow School*, edited by V. Arnold and M. Monastyrsky (Chapman and Hall, London, 1993), pp. 131–193.
[13] M. A. Soloviev, Commun. Math. Phys. **184**, 579 (1997).
[14] U. Moschella and F. Strocchi, Lett. Math. Phys. **24**, 103 (1992).
[15] H. Komatsu, J. Math. Soc. Jpn. **19**, 366 (1967).
[16] J. Bümmerstede and W. Lücke, J. Math. Phys. **16**, 1203 (1975).
[17] W. Lücke, Acta Phys. Austriaca **55**, 213 (1984).
[18] H. Komatsu, J. Fac. Sci., Univ. Tokyo, Sect. 1A **20**, 25 (1973).
[19] M. A. Soloviev, Lett. Math. Phys. **33**, 49 (1995).

Standard forms of local nets in quantum field theory

Launey J. Thomas, III and Eyvind H. Wichmann
Department of Physics, University of California, Berkeley, California 94720

(Received 22 August 1997; accepted for publication 15 January 1998)

Nets of von Neumann algebras, with local algebras associated with all compact, convex, causally complete subsets of spacetime, are constructed, assuming a particular kind of net considered by Bisognano and Wichmann. The properties of such nets are discussed, within the general framework of algebraic quantum field theory, with an emphasis on the relation between the geometry of spacetime and the structure of the nets. An ''intersection property'' for the local algebras is stated and proved. The structure of subnets of such nets is discussed. A number of particular possible features of the nets are shown to be hereditary in the sense that the subnets have the same features. Possible further extensions of the domain of definition of these nets are discussed briefly. © *1998 American Institute of Physics.*
[S0022-2488(98)00105-4]

I. INTRODUCTION

This study is within the framework of algebraic quantum field theory. A paper published in 1964 by Haag and Kastler[1] can be regarded as the seminal paper of this subject. We refer to the monographs by Haag,[2] and by Horuzhy,[3] for detailed descriptions of this approach, and for extensive references to the literature. The characteristic feature of such a theory is the existence of a *local net*, by which is understood an association of C^* - or von Neumann algebras with subsets of Minkowski spacetime M in a ''suitable'' family $\mathscr{R}$ of subsets. In this paper we will assume that the algebras are von Neumann algebras on a Hilbert space $\mathscr{H}$. We assume that $\mathscr{H}$ carries a strongly continuous unitary representation of the covering group of the Poincaré group, and that $\mathscr{H}$ contains a unique vacuum vector. We are thus only concerned with the vacuum representation of the algebras in the net. For the choice of the family of subsets one can imagine a great variety of possibilities. The family $\mathscr{R}$ should always include a sufficiently large set of *bounded* regions, and the operators in the von Neumann algebras associated with these bounded regions are the *local* operators of the net. In Section II we review some features of the theory of local nets in general. (See, in particular, Definition 2.1, which describes the notion of a local net which we adhere to in this paper.)

In Section II we also review a *specific* structure discussed by Bisognano and Wichmann,[4] and by Wichmann.[5] The subsets of M for which the von Neumann algebras are defined are the causal wedges and the closed double cones (with nonempty interiors) in M, as well as the causal complements of these double cones. Our aim in this paper is to define local nets for *larger* classes of subsets of Minkowski spacetime, and to discuss the properties of such extended nets. Our starting point is a net of the kind discussed by Bisognano and Wichmann, which we then embed in a net with a suitable larger domain of subsets of M. This involves no new assumptions, and the properties of the extended nets are thus already implicit in the structure of the ''original'' net.

A main theme of this paper is the relationship between geometric features of Minkowski spacetime and the structures of local nets. In an earlier paper,[6] hereafter referred to as TW-I, we have discussed the causal structure of Minkowski spacetime, and in a subsequent paper,[7] hereafter referred to as TW-II, we have discussed a class of tempered distributions of interest in quantum field theory, with reference to the geometric theory in TW-I. In the present paper we discuss the implications of this causal structure for the structure of local algebras, and our theory thus depends in an essential manner on the results in TW-I and TW-II. We follow the notation and terminology in these papers, to which the reader is referred for details which are not spelled out here. For the reader's convenience we briefly review here some points of particular importance.

For any subset R of Minkowski spacetime M, we denote the closure of R by $\overline{R}$, the interior of

 © 1998 American Institute of Physics

R by R^o, the complement of R in M by $\widetilde{R}$, and the boundary of R by ∂R. The *causal complement* R^c of R is defined as the set of all point of M which are spacelike relative to all points of R. The set $R^{cc}=(R^c)^c$ is called the *causal completion of* R, and we say that a subset R of M is *causally complete* if and only if $R^{cc}=R$. For a detailed discussion of these notions we refer to TW-I.[6]

We describe here the kinds of subsets of Minkowski spacetime with which we will be particularly concerned.

Definition 1.1: (a) For any $x\in\mathsf{M}$ we denote by $V_+(x)$ the *open forward lightcone*, and by $V_-(x)$ the *open backward lightcone*, with x as the apex.

(b) For any two points $x,y\in\mathsf{M}$ such that $y\in V_+(x)$ we define the *closed double cone* $D(x,y)$ with x and y as apices by $D(x,y)=\overline{V}_+(x)\cap\overline{V}_-(y)$, and we denote by $\mathscr{D}$ the set of all such double cones. Note that with this definition every double cone $D\in\mathscr{D}$ has a nonempty interior.

(c) We denote by $\mathscr{W}$ the set of all open wedge-regions W which are Poincaré-equivalent with the particular wedge-region $W_R=\{x|x_3>|x_4|\}$, i.e., an element $W\in\mathscr{W}$ is any open subset of M bounded below by a characteristic plane π' and bounded above by another characteristic plane π'' not parallel with π'.

(d) We denote by $\mathscr{K}$ the set of all *closed, convex, causally complete* subsets K of M such that K^c has a *nonempty* interior.

(e) We define $\mathscr{K}_E$ as the set of all *closed, convex, causally complete* subsets K of M which satisfy the condition that there exist four characteristic planes π_1, π_2, π_3, and π_4 in a *general relative position*, i.e., the normals of these planes are linearly independent, such that K lies *below* π_1 and π_2 and *above* π_3 and π_4.

(f) We define the subset $\mathscr{K}_C$ of $\mathscr{K}_E$ as the set of all *compact* sets $K\in\mathscr{K}_E$, i.e., $\mathscr{K}_C$ is the set of all *compact, convex, causally complete* subsets of M. We regard the empty set $\varnothing$ as an element of $\mathscr{K}_C\subset\mathscr{K}_E$. We have $\mathscr{D}\subset\mathscr{K}_C$.

(g) We denote by $\mathscr{K}_B$ the set of all *compact, causally complete* subsets K of M.

(h) We say that two wedges W_1 and W_2 are in a *general relative position* if the four boundary planes are in a general relative position. The set $\mathscr{K}_E$ can also be described as follows: A closed, convex, causally complete set K is in $\mathscr{K}_E$ if and only if there exist two wedges $W_1,W_2\in\mathscr{W}$ in a general relative position such that $K\subset W_1\cap W_2$.

In Section II we give a precise definition of what we here call a *basic local net* (see Definition 2.3). Such a net is of the kind discussed by Bisognano and Wichmann,[4] with von Neumann algebras $\mathscr{A}_o(W)$, $\mathscr{B}_o(D)$, and $\mathscr{A}_o(D^c)$ defined for all $W\in\mathscr{W}$ and all $D\in\mathscr{D}$, but we add some specific assumptions. In particular we assume that the vacuum vector Ω is cyclic for the algebra $\mathscr{B}_o(D)$ associated with any double cone $D\in\mathscr{D}$. We do assume that the algebras $\mathscr{A}_o(W)$ associated with the causal wedges W in $\mathscr{W}$ satisfy the *condition of duality*, but not necessarily the so-called *special condition of duality* discussed by Bisognano and Wichmann (see Definition 2.4). The notion of a local net is a very general notion, and it hardly makes sense to try to pursue a large number of variants. Our aim was to select a reasonable starting point for the discussion in this paper.

Sections III and IV incorporate, and depend substantially on results from the doctoral thesis of one of us (L.J.T.).[8] In Section III we discuss the construction of a specific extended local net, for which we employ the term *regular net* as a technical term in this paper, starting from a *basic net*. This extended net consists of von Neumann algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$, defined for all $W\in\mathscr{W}$, for all $K\in\mathscr{K}_C$. For any $K\in\mathscr{K}_C$ the algebra $\mathscr{B}(K)$ is defined as the intersection of *all* algebras $A_o(W)$ for which W contains K. We show that the extended net has reasonable properties. It satisfies all the usual conditions of a local net, as well as the condition of duality. Furthermore the set of all *local* operators of the extended net is equal to the set of all local operators of the original basic net. One can thus say that the extension merely amounts to a more refined association of the local operators with bounded subsets of M. The extended net is also *maximal* in the sense that the algebras $\mathscr{B}(K)$ cannot be larger without violating the condition of locality. This corresponds to a general principle which we have followed in this paper: to arrange it so that the *local* algebras are as large as they can be, consistent with the condition of locality. The extension which we describe depends in an essential way on the feature that the sets $K\in\mathscr{K}_C$ are convex, compact, and causally complete. We also discuss a rather natural association of algebras with

other closed, convex, causally complete subsets of M, but this further extended system of algebras is not necessarily a local net.

In Section IV we discuss various properties of *regular* nets. We first state and prove some continuity properties for the mapping of subsets of M into the set of von Neumann algebras. We next show that an algebra $\mathscr{B}(K)$ can be nontrivial for a $K\in\mathscr{K}_C$ with an empty interior, but that the algebra *is* trivial if K is contained in a *two*-dimensional linear manifold. Furthermore we discuss various *intersection properties*. Among other things we show that if $K_1,K_2\in\mathscr{K}_C$, then $\mathscr{B}(K_1)\cap\mathscr{B}(K_2)=\mathscr{B}(K_1\cap K_2)$. To prove this statement in full generality we have to assume that the basic net satisfies the *condition of additivity for wedges* (see Definition 2.5). However, this condition is not needed for the conclusion that $\mathscr{B}(K_1)$ and $\mathscr{B}(K_2)$ have only multiples of the identity in common if K_1 and K_2 are spacelike separated. Our discussion of intersection properties is closely related to a similar discussion in Section V in our earlier paper TW-II.[7] There the object of interest was the *weak commutant* $\mathscr{C}(K)$ of the quantum field operators in the causal complement of K, and it was shown that $\mathscr{C}(K_1)\cap\mathscr{C}(K_2)=\mathscr{C}(K_1\cap K_2)$ if $K_1,K_2\in\mathscr{K}_C$. It is not known whether the weak commutants are von Neumann algebras for an arbitrary field theory, but there is a particularly pleasing scenario in which this is the case. For *Bose fields* we then have $\mathscr{B}(K)=\mathscr{C}(K)$, and in particular this is the case for a *free* Bose field. In the present paper we do not make any assumption about associated quantum fields, and hence we cannot depend on the results in TW-II[7] concerning intersection properties.

In Section V we discuss local subnets of a regular net. This generalizes an earlier study by Wichmann[5] of subnets of a basic net. In this discussion the *special condition of duality* is assumed. There is a natural standard form of such a subnet for which the vacuum representation of the subnet is a regular local net which also satisfies the special condition of duality. We discuss a number of possible special properties of a regular net which are also inherited by the vacuum representation of the subnet. Finally we show that for any local operator B there is an *intrinsic* smallest $K\in\mathscr{K}_C$ such that $B\in\mathscr{B}(K)$.

In Section VI we briefly discuss possible further extensions of a *regular* net. The goal is to define local algebras $\mathscr{B}_e(K)$ for every K in the set $\mathscr{K}_B$ of all compact causally complete subsets of M. The methods in Section III are not applicable, because the sets in $\mathscr{K}_B$ are not all convex, and hence other procedures for the extension must be employed. We show that it is always possible to construct a particular ''standard'' extension to a local net with von Neumann algebras $\mathscr{A}_S(W)$, $\mathscr{B}_S(K)$, and $\mathscr{A}_S(K^c)$, defined for all $W\in\mathscr{W}$, for all $K\in\mathscr{K}_B$. This extension satisfies a number of reasonable conditions, but it is definitely not as canonical as the extension to a regular net in Section III. We cannot show that it is maximal or that a unique maximal extension exists, but we show that the local algebras $\mathscr{B}_e(K)$ of any maximal extension satisfy the condition $\mathscr{B}_e(K)\supset\mathscr{B}_S(K)$. We discuss some features of these extensions of regular nets. Among other things we show that the intersection properties discussed in Section IV *cannot* hold generally for the extended nets.

Finally it is worth stating that our discussion refers *specifically* to four-dimensional Minkowski spacetime. Our reasoning depends on results in our earlier papers,[6,7] some of which are specific for a four-dimensional spacetime, and not necessarily valid for other (model) spacetimes.

II. GENERALITIES ABOUT LOCAL NETS OF VON NEUMANN ALGEBRAS

Minkowski spacetime M is parameterized by the usual Cartesian coordinates $x=(x^1,x^2,x^3,x^4)=(\mathbf{x},t)$, with $t=x^4$, $\mathbf{x}=(x^1,x^2,x^3)$, with reference to some particular inertial frame. The Minkowski scalar product is denoted $x\cdot y=x^4y^4-\mathbf{x}\cdot\mathbf{y}$, where $\mathbf{x}\cdot\mathbf{y}$ is the usual scalar product on the Euclidean subspace R^3.

We denote the universal covering group of the Poincaré group by $\hat{\mathsf{P}}$, and its elements by $\lambda=\lambda(x,g)$, where $g\in SL(2,C)$ and $x\in\mathsf{M}$. The element $\lambda(x,g)=\lambda(x,I)\lambda(0,g)$ thus describes a homogeneous transformation g followed by a translation by x. In the interpretation of g we follow the usual conventions. The states of the physical system are described in terms of a separable Hilbert space $\mathscr{H}$, which carries a strongly continuous unitary representation $\lambda\to U(\lambda)$ of $\hat{\mathsf{P}}$. We will also denote the translations by $T(x)\equiv U(\lambda(x,I))=U(x,I)$, and we assume the usual spectrum condition for $T(x)$, i.e., the support of the spectral measure μ_T of $T(x)$ is contained in the closed

forward lightcone $\overline{V}_+$. We assume the existence of a normalized vacuum state vector Ω, uniquely determined (up to a factor of modulus unity) by the condition $U(\lambda)\Omega=\Omega$, for all $\lambda\in\hat{\mathsf{P}}$. We also assume the existence of a *PCT*-operator θ, which satisfies the usual conditions[9–11] relative to the $U(\lambda)$, and relative to the local algebras under discussion.

To deal efficiently with the commutation and anticommutation relations which appear in the locality conditions, we employ a device of Bisognano and Wichmann.[4] We denote the unitary operator which describes a rotation by angle 2π about any axis by $U_0=U(0,-I)$, and we define the unitary operator Z by

$$Z=\frac{I+iU_0}{1+i},$$

and we then have

$$U_0^2=I,\quad Z^2=U_0,\quad U_0\Omega=\Omega=Z\Omega,\quad \theta Z\theta^{-1}=Z^{-1},\quad U(\lambda)Z=ZU(\lambda),$$

for all $\lambda\in\hat{\mathsf{P}}$.

For any operator Q, and any set $\mathscr{S}$ of operators we write

$$Q^z=ZQZ^{-1},\quad \mathscr{S}^z=\{Q^z|Q\in\mathscr{S}\}.$$

We refer to operators which commute with the involution U_0 as *boson operators*, and we refer to operators which anticommute with U_0 as *fermion operators*. If X is a boson operator and Y is a fermion operator we thus have $X^z=ZXZ^{-1}=X$ and $Y^z=ZYZ^{-1}=iU_0Y$. Every operator Q has a unique representation as a sum of a boson operator Q_b and a fermion operator Q_f,

$$Q=Q_b+Q_f,\quad Q_b=\tfrac{1}{2}(Q+U_0QU_0),\quad Q_f=\tfrac{1}{2}(Q-U_0QU_0).$$

It is important that this resolution ''respects the locality'' of an operator. If Q is a local operator associated with some region R, then Q_b and Q_f are local operators associated with the *same* region, because the *algebra* $\mathscr{M}(R)$ of the operators associated with the region is invariant under conjugation by U_0. The *condition of locality* can now be stated as follows. If X and Y are two local operators associated with two *spacelike* separated regions, then $[X,Y^z]=0=[X^z,Y]$.

For any $R\subset\mathsf{M}$ and any $\lambda\in\hat{\mathsf{P}}$ we write $R_{(\lambda)}$ for the image of R under the usual action of λ on Minkowski spacetime. Corresponding to the *PCT* operator θ we have the mapping $R\to R_{(\theta)}=\{-x|x\in R\}$.

Throughout this paper we will assume the above structure, and follow the above notation. We next give a general description of a *local net*,[1–4] within the framework of our paper. We note that our notion of a local net is, in some respects, less general than in the papers by Haag and Kastler,[1] and by Haag and Schroer.[12]

Definition 2.1: Let $\mathscr{R}$ be a set of subsets of M, such that if $R\in\mathscr{R}$, then $R_{(\theta)}\in\mathscr{R}$, and $R_{(\lambda)}\in\mathscr{R}$ for all $\lambda\in\hat{\mathsf{P}}$. Furthermore we assume that $\mathscr{R}$ contains a subset $\mathscr{R}_L$ of *bounded* subsets of M such that if $R_1,R_2\in\mathscr{R}_L$, then there exists an $R\in\mathscr{R}_L$ such that $R\supset R_1\cup R_2$.

(a) By a *local net* $\mathscr{N}$ *with domain* $\mathscr{R}$ we understand the following structure.

I. To every $R\in\mathscr{R}$ corresponds a von Neumann algebra $\mathscr{M}(R)$ of operators on $\mathscr{H}$.

II. The net $\mathscr{N}$ satisfies the *condition of isotony*, in the sense that $\mathscr{M}(R_i)\subset\mathscr{M}(R)$ if $R_i,R\in\mathscr{R}$ and $R_i\subset R$.

III. The net $\mathscr{N}$ is *Poincaré covariant*, in the sense that $\mathscr{M}(R_{(\lambda)})=U(\lambda)\mathscr{M}(R)U(\lambda)^{-1}$ for all $R\in\mathscr{R}$ and all $\lambda\in\hat{\mathsf{P}}$.

IV. The net $\mathscr{N}$ is *local*, in the sense that $\mathscr{M}(R_1)^z\subset\mathscr{M}(R_2)'$ whenever $R_1,R_2\in\mathscr{R}$ and $R_1\subset R_2^c$. Here the prime indicates the commutant.

V. The net $\mathscr{N}$ *is determined by its local operators* in the sense that $\mathscr{M}(R)\subset\mathscr{U}''$ for all $R\in\mathscr{R}$, where $\mathscr{U}=\cup\{\mathscr{M}(R)|R\in\mathscr{R}_L\}$ is the set of all *local operators* of the net $\mathscr{N}$.

(b) In *addition* a net may, or may not, satisfy any one of the following conditions.

VI. The net $\mathscr{N}$ is *PCT-covariant*, in the sense that $\mathscr{M}(R_{(\theta)})=\theta\mathscr{M}(R)\theta^{-1}$ for all $R\in\mathscr{R}$.

VII. The net $\mathcal{N}$ is *irreducible*, in the sense that the set of operators $\cup\{\mathcal{M}(R)|R\in\mathcal{R}\}$ is1 irreducible, which is the case if and only if $\mathcal{U}''=\mathsf{B}(\mathcal{H})$. Here $\mathsf{B}(\mathcal{H})$ denotes the set of all bounded operators on $\mathcal{H}$.

VIII. The net $\mathcal{N}$ satisfies the *condition of duality*, in the sense that $R^c\in\mathcal{R}$ whenever $R\in\mathcal{R}$, *and* $\mathcal{M}(R)^z=\mathcal{M}(R^c)'$ for all $R\in\mathcal{R}$.

IX. The net has a *strictly local structure*, in the sense that $R=\cup\{R_b|R_b\in\mathcal{R}_L,\ R_b\subset R\}$ and $\mathcal{M}(R)=\{\mathcal{M}(R_b)|R_b\in\mathcal{R}_L,\ R_b\subset R\}''$, for all $R\in\mathcal{R}$.

(c) Let $\mathcal{N}$ be a local net as above, with domain $\mathcal{R}$, and let $\mathcal{N}_s$ be another local net with algebras $\mathcal{M}_s(R)$ defined on the *same* domain $\mathcal{R}$. We say that $\mathcal{N}_s$ is a *subnet of* $\mathcal{N}$ if and only if $\mathcal{M}_s(R)\subset\mathcal{M}(R)$ for all $R\in\mathcal{R}$, and we can express this relationship by the symbolism $\mathcal{N}_s<\mathcal{N}$ or $\mathcal{N}>\mathcal{N}_s$.

The purpose of the above definition is to summarize some *general* properties of local nets for later reference. The ''essence'' of a local net is, of course, the function $\mathcal{M}$. We employ the symbol $\mathcal{N}$, rather than $\mathcal{M}$, for a net as a symbol for the *entire* structure described above.

The question of how the domain $\mathcal{R}$ should be chosen is here left open temporarily. In this paper we will discuss some specific choices as explained later, but other choices can be imagined. Conditions I–IV are always imposed on local nets, and the physical motivation for these conditions is obvious. We include Condition V as a condition that there are ''enough'' *local* operators in the sense that the von Neumann algebra generated by these contains all the operators in the net. Note that the set $\mathcal{U}$ is a *unital* **-algebra* under the stated conditions. Condition IX is a much stronger condition than Condition V. The nets discussed in Sections III–V will satisfy this condition, as well as the condition of duality. One could argue that the *proper* statement of the condition of duality is that *both* Conditions VIII and IX should be satisfied. In the original formulation by Haag and Schroer[12] the condition of duality is a condition on the *local* operators of a net. We illustrate this for the case of a local net with the domain $\mathcal{R}=\mathcal{D}$. The condition that

$$\mathcal{M}(D)^z=\{\mathcal{M}(D_o)|D_o\in\mathcal{D},\ D_o\subset D^c\}', \tag{2.1}$$

for all $D\in\mathcal{D}$ involves only local operators, and a condition of this nature is commonly referred to as *Haag duality*. If we now extend the domain of the net to include all causal complements D^c of the double cones D by defining $\mathcal{M}(D^c)=\{\mathcal{M}(D_o)|D_o\in\mathcal{D},\ D_o\subset D^c\}''$ it follows that both Conditions VIII and IX are satisfied if the condition in (2.1) holds. The duality condition in (2.1) implies that the local algebras are *maximal*, i.e., if $\{\mathcal{M}_1(D)|D\in\mathcal{D}\}$ are the algebras of another local net with the same domain $\mathcal{D}$, and such that $\mathcal{M}_1(D)\supset\mathcal{M}(D)$, then $\mathcal{M}_1(D)=\mathcal{M}(D)$.

We note here a simple relationship between the condition of duality (in *our* formulation) and the conditions of isotony and locality, which we express as follows.

Lemma 2.2: Let $\mathcal{R}$ be a set of subsets of M such that $R^c\in\mathcal{R}$ whenever $R\in\mathcal{R}$. Let there be given a von Neumann algebra $\mathcal{M}(R)$ for each $R\in\mathcal{R}$, and suppose that the mapping $R\to\mathcal{M}(R)$ satisfies the *condition of duality* $\mathcal{M}(R)^z=\mathcal{M}(R^c)'$ for all $R\in\mathcal{R}$. Then the mapping satisfies the *condition of isotony*, i.e., $\mathcal{M}(R_i)\subset\mathcal{M}(R)$ whenever $R_i\subset R$, if and only if the mapping satisfies the *condition of locality*, i.e., $\mathcal{M}(R_1)^z\subset\mathcal{M}(R_2)'$ whenever $R_1\subset R_2^c$.

This trivial observation is of interest when we want to verify that a mapping $R\to\mathcal{M}(R)$ is a local net.

In a ''local theory'' we are naturally interested in local nets which are *irreducible*. Subnets of such a net are also of interest, and they are in general *reducible*. In Section V we will discuss subnets and their (irreducible) *vacuum representations*.

As we said in the Introduction, we will *assume* the existence of a local net of the kind discussed in Bisognano and Wichmann,[4] and in Wichmann.[5] The set $\mathcal{R}$ of subsets of M for which the von Neumann algebras are defined are the wedges in $\mathcal{W}$ and the closed double cones D (with *nonempty* interiors) in $\mathcal{D}$, as well as the causal complements D^c of these double cones. In the quoted papers such a local net was called an *AB-system* (to distinguish it from other kinds of local nets). In this paper we will add some specific conditions, and refer to such a net as a *basic (local) net*. This is to distinguish it from the nets which we will construct by extending the domain $\mathcal{R}$ of the mapping $\mathcal{M}$.

Definition 2.3: By an *irreducible basic local net* $\mathcal{N}_o$ is understood a system of von Neumann algebras of operators on $\mathcal{H}$ with the following properties.

(a) To every $W \in \mathscr{W}$ corresponds a von Neumann algebra $\mathscr{A}_o(W) = \mathscr{M}(W)$, and to every $D \in \mathscr{D}$ correspond two von Neumann algebras $\mathscr{B}_o(D) = \mathscr{M}(D)$ and $\mathscr{A}_o(D^c) = \mathscr{M}(D^c)$. These algebras constitute a local net $\mathscr{N}_o$ which satisfies *all* the Conditions I–IX in Definition 2.1, specifically the following:

(b) The von Neumann algebras of the net $\mathscr{N}_o$ are related by

$$\mathscr{B}_o(D) = \cap\{\mathscr{A}_o(W) | W \in \mathscr{W},\ \overline{W} \supset D\}, \quad \mathscr{A}_o(D^c) = \{\mathscr{A}_o(W) | W \in \mathscr{W},\ W \subset D^c\}'', \tag{2.2}$$

$$\mathscr{A}_o(W) = \{\mathscr{B}_o(D) | D \in \mathscr{D},\ D \subset W\}'', \quad \mathscr{A}_o(D^c) = \{\mathscr{B}_o(D_i) | D_i \in \mathscr{D},\ D_i \subset D^c\}'', \tag{2.3}$$

for all $D \in \mathscr{D}$, and all $W \in \mathscr{W}$.

(c) The net $\mathscr{N}_o$ satisfies the *condition of duality* in the sense that

$$\mathscr{A}_o(W)^z = \mathscr{A}_o(\overline{W^c})', \quad \mathscr{B}_o(D)^z = \mathscr{A}_o(D^c)', \tag{2.4}$$

for all $D \in \mathscr{D}$, and all $W \in \mathscr{W}$.

(d) For any $D \in \mathscr{D}$ we have

$$\overline{\mathscr{B}_o(D)\Omega} = \mathscr{H}. \tag{2.5}$$

(e) We denote by $\mathscr{U}_o$ the *algebra of all local operators of the net*, defined as $\mathscr{U}_o = \cup\{\mathscr{B}_o(D) | D \in \mathscr{D}\}$. The net $\mathscr{N}_o$ is *irreducible*, and hence

$$\mathscr{U}_o' = \{\mathscr{B}_o(D) | D \in \mathscr{D}\}'' = \{\mathscr{A}_o(W) | W \in \mathscr{W}\}'' = \mathsf{B}(\mathscr{H}),$$

where $\mathsf{B}(\mathscr{H})$ denotes the set of all bounded operators on $\mathscr{H}$.

The reader will note that the above is a description of a scenario, rather than a statement of a minimal set of assumptions. The various stated conditions on the net $\mathscr{N}_o$ are obviously by no means independent. It is worthwhile to discuss here some aspects of the logical structure of this scenario.

In the papers by Bisognano and Wichmann,[4] and by Wichmann,[5] the algebras $\mathscr{B}_o(D)$ and $\mathscr{A}_o(D^c)$ were in *defined* in terms of the ''wedge-algebras'' $\mathscr{A}_o(W)$, essentially as in (2.2). With this construction it follows at once that *if* the algebras $\mathscr{A}_o(W)$ satisfy the applicable conditions of Poincaré covariance, *PCT*-covariance, isotony, and locality, as was assumed in these papers, *then* the algebras $\mathscr{B}_o(D)$ and $\mathscr{A}_o(D^c)$ satisfy the corresponding conditions relevant for the regions D and D^c of these algebras.

We note here that it follows trivially from the strong continuity of the representation of the translation group that

$$\mathscr{A}_o(W) = \cap\{\mathscr{A}_o(W_o) | W_o \in \mathscr{W},\ W_o \supset \overline{W}\} = \{\mathscr{A}_o(W_i) | W_i \in \mathscr{W},\ \overline{W}_i \subset W\}''. \tag{2.6}$$

The expression in the second member could be regarded as the algebra corresponding to the *closed* wedge $\overline{W}$, properly denoted by $\mathscr{B}_o(\overline{W})$, but since it is equal to $\mathscr{A}_o(W)$ we do not employ the notation $\mathscr{B}_o(\overline{W})$. As discussed in the quoted papers, it follows that the algebras $\mathscr{B}_o(D)$ are ''continuous from the outside,'' and the algebras $\mathscr{A}_o(D^c)$ are ''continuous from the inside,'' in the sense that

$$\mathscr{B}_o(D) = \cap\{\mathscr{B}_o(D_o) | D_o \in \mathscr{D},\ D_o^o \supset D\}, \quad \mathscr{A}_o(D^c) = \{\mathscr{A}_o(D_i^c) | D_i \in \mathscr{D},\ D^c \supset \overline{D_i^c}\}''. \tag{2.7}$$

The wedge-algebras $\mathscr{A}_o(W) = \mathscr{B}_o(\overline{W})$ are continuous from both sides, in view of (2.6). Furthermore we have

$$\mathscr{B}_o(D) = \cap\{\mathscr{A}_o(W) | W \in \mathscr{W},\ W \supset D\}, \quad \mathscr{A}_o(D^c) = \{\mathscr{A}_o(W) | W \in \mathscr{W},\ \overline{W} \subset D^c\}''. \tag{2.8}$$

It is thus a matter of taste whether we define the algebras $\mathscr{B}_o(D)$ and $\mathscr{A}_o(D^c)$ through (2.2) or through (2.8). The above relations have a certain obvious *geometric* basis, which we do not

think we have to describe explicitly here. Note here that the isotony conditions include the condition that $\mathscr{B}(D)\subset\mathscr{A}(W)$ whenever $D\subset\overline{W}$, as is also implied by (2.2).

The construction of the basic net in Bisognano and Wichmann[4] was based on the premise that the vacuum vector Ω is cyclic for any wedge-algebra $\mathscr{A}_o(W)$. By a well-known argument within quantum field theory[9–11] it then follows that the set of operators $\cup\{\mathscr{A}_o(W)|W\in\mathscr{W}\}$ is irreducible, but this *does not* imply the condition in (2.5), or the relations in (2.3). Within the framework of the Bisognano–Wichmann theory it might conceivably happen that $\mathscr{B}_o(D)=\mathbf{C}^1\cdot I$ for all $D\in\mathscr{D}$, but such a theory would not be a physically sensible theory of *local* operators. Another possibility is that the algebra $\mathscr{U}_o$ is *irreducible*, but that the relation $\overline{\mathscr{B}_o(D)\Omega}=\mathscr{H}$ does not hold for *all* $D\in\mathscr{D}$. In particular one could imagine the situation in which $\mathscr{B}_o(D)=\mathbf{C}^1\cdot I$ for all ''sufficiently small'' double cones D. Such a theory would thus involve a notion of a *fundamental length*. We know of no explicit examples of such a net, and we will therefore not consider this possibility. In this paper we thus *assume* the condition in (2.5), from which it follows that the algebra $\mathscr{U}_o$ of all local operators is irreducible.

We *assume* that the net satisfies the condition of duality, as stated in Definition 2.3. The condition of duality at left in (2.4) is, of course, stronger than the condition of locality for the algebras $\mathscr{A}_o(W)$. We note that this condition of duality for the algebras $\mathscr{A}_o(W)$ implies the duality condition at right in (2.4). As far as we know the *wedge duality* $\mathscr{A}_o(W)^z=\mathscr{A}_o(\overline{W^c})'$ does not follow from the *Haag duality* in (2.1), with $\mathscr{A}_o(W)$ *defined* by (2.3) in terms of the local algebras $\mathscr{B}_o(D)=\mathscr{M}(D)$.

In the papers of Bisognano and Wichmann,[4] and Wichmann,[5] the nets were subject to a very specific further condition, called the *special condition of duality*, which we here state as follows, within the framework of the Tomita–Takesaki theory.[13]

Definition 2.4: Let the premises and notation be as in Definition 2.3. We say that the *basic* net $\mathscr{N}_o$ satisfies the *special condition of duality* if and only if the antiunitary involution $J=Z\theta U(0,\exp(-i\pi\sigma_3/2)$ is the modular conjugation, and the one-parameter group $\{V(\tau)|\tau\in\mathbf{R}^1\}$, with $V(\tau)=U(0,\exp(\tau\sigma_3/2))$, of all velocity transformations in the 3-direction (within the Poincaré group) is the modular automorphism group, for the algebra $\mathscr{A}_o(W_R)$, relative to the cyclic and separating vector Ω.

We refer to the quoted papers, as well as to the papers by Borchers,[14,15] and by Brunetti, Guido, and Longo,[16] for discussions of the implications of this condition, and of its plausibility and standing. The ''origin'' of this condition is a property of quantum fields, within the framework of local quantum field theory.[9–11] We refer to such fields as *Wightman fields*. There are good reasons to believe[4,17] that the special condition of duality holds for local nets which are ''locally related'' to an irreducible system of (finite-component) local Wightman fields, but the condition may have a more general validity. For discussions of the connection between quantum fields and local nets, we refer to the papers by Driessler, Summers, and Wichmann,[17] and by Borchers and Yngvason.[18,19] The general idea is that the unbounded local field operators have closed extensions affiliated with the local von Neumann algebras in the net. It was shown by Summers and Wichmann[20] that a local net which is related to Wightman fields in such a manner also satisfies a particular condition of *additivity*, which we state as follows.

Definition 2.5: Let $\mathscr{N}$ be a local net, with algebras $\mathscr{A}(W)$, $\mathscr{A}(D^c)$, and $\mathscr{B}(D)$ defined for all $W\in\mathscr{W}$ and $D\in\mathscr{D}$. We say that the net $\mathscr{N}$ satisfies the *condition of additivity for wedges* if and only if

$$\mathscr{A}(W)\subset\{\mathscr{B}(D)|D\in\mathscr{F}\}''$$

for any $\mathscr{F}\subset\mathscr{D}$ such that

$$W\subset\cup\{D^o|D\in\mathscr{F}\},$$

where D^o denotes the interior of D.

In the quoted paper it was explicitly proved that this condition holds for an irreducible local net locally associated with a single Hermitian scalar field, but the reasoning applies equally well, with minor modifications, to the case of any irreducible set of (finite-component) Wightman fields. For some of the conclusions in our paper we will require this additional property of the local net,

and when this is so, we will refer to this assumption explicitly. We note here that if the condition of additivity for wedges holds for a net, then Ω is a cyclic vector for $\mathscr{B}_o(D)$, for *all* $D \in \mathscr{D}$, if and only if the net is irreducible.

If a *basic* net $\mathscr{N}_o$ satisfies the *special condition of duality* in Definition 2.4, then the net has the somewhat surprising property that

$$\mathscr{A}_o(W_R) = \{V(\tau).\mathscr{B}_o(D)V(\tau)^{-1} | \tau \in \mathsf{R}^1\}'',$$

for *any* double cone $D \subset W_R$. This is suggestive, but as far as we know, the condition of additivity for wedges does *not* follow from the conditions in Definition 2.3, even if amended by the special condition of duality.

We want to contrast the above condition of additivity for wedges with the following apparently much stronger *condition of additivity for double cones*: For every $D \in \mathscr{D}$ we have $\mathscr{B}_o(D) \subset \{\mathscr{B}_o(D_c) | D_c \in \mathscr{F}\}''$ for any family $\mathscr{F}$ of doubles cones such that $D \subset \cup\{D_c^o | D_c \in \mathscr{F}\}$. It obviously implies the condition in Definition 2.5 if we also have $\mathscr{A}_o(W) = \{.\mathscr{B}_o(D) | D \in \mathscr{D}, D \subset W\}''$. This stronger condition does hold for the local nets generated by a system of free fields, and it describes a physically reasonable feature, but it is not known whether it holds for local nets $\mathscr{N}_o$ in general, under the conditions in Definition 2.3. One may thus have doubts about the validity of the condition of additivity for double cones, whereas we feel that the condition of additivity for wedges is much more plausible. In this paper we will not depend in any way on the stronger condition.

The condition of additivity for wedges trivially implies the relations in (2.3). If the condition of duality also holds, then it follows from these relations that

$$\mathscr{A}_o(W) = \cap\{.\mathscr{A}_o(D^c) | D \in \mathscr{D},\ D^c \supset \overline{W}\}, \quad \mathscr{B}_o(D) = \cap\{.\mathscr{A}_o(D_o^c) | D_o \in \mathscr{D},\ D_o^c \supset D\}, \tag{2.9}$$

for all $D \in \mathscr{D}$, and all $W \in \mathscr{W}$. Conversely the relations in (2.9) imply the relations in (2.3), if the condition of duality holds. With the situation as described in Definition 2.3 the net $\mathscr{N}_o$ is thus completely determined by the *local* algebras $\mathscr{B}_o(D)$. The relation $\mathscr{A}_o(W) = \{.\mathscr{B}_o(D) | D \in \mathscr{D}, D \subset W\}''$ in (2.3) does *not* follow from the other conditions in Definition 2.3. However, it does follow from the special condition of duality in Definition 2.4, *and* the condition that $\overline{\mathscr{B}_o(D)\Omega} = \mathscr{H}$ for at least one $D \in \mathscr{D}$.

We conclude this section with two lemmas which will be of use later. Both lemmas depend on the assumed spectrum condition on the translation group.

Lemma 2.6: (a) Let u be a *timelike* or *isotropic* four-vector. Let A and B be two bounded operators which satisfy the condition

$$\langle \Omega | [B, T(su)AT(su)^{-1}]\Omega \rangle = 0, \tag{2.10}$$

for all real s. Then,

$$\langle \Omega | BT(su)A\Omega \rangle = \langle \Omega | BA\Omega \rangle = \langle \Omega | B\Omega \rangle \langle \Omega | A\Omega \rangle,$$

for all real s.

(b) Let $\mathscr{B}_o(D)$ be the algebra of a *basic net* $\mathscr{N}_o$ corresponding to the double cone $D \in \mathscr{D}$. If A is a bounded operator which satisfies the condition in (2.10) for all $B \in \mathscr{B}_o(D)$, then $A\Omega = a\Omega$, $A^*\Omega = a^*\Omega$, with $a = \langle \Omega | A\Omega \rangle$. If furthermore $A \in \mathscr{B}_o(D_0)'$ for some $D_0 \in \mathscr{D}$, then $A = aI$.

Proof: The assertion in part (a) is a well-known result within quantum field theory, which follows from the spectrum condition on the translation group, and from the fact that the weak limit of $T(x)$ equals the vacuum projection $|\Omega\rangle\langle\Omega|$ as x tends to infinity along any fixed direction.

Under the premises in part (b) the relation $\langle B^*\Omega | A\Omega \rangle = a\langle B^*\Omega | \Omega \rangle$ holds for all $B \in .\mathscr{B}_o(D)$, and since $\mathscr{B}_o(D)\Omega$ is dense in $\mathscr{H}$, it follows that $A\Omega = a\Omega$, and similarly that $A^*\Omega = a^*\Omega$. If also $A \in \mathscr{B}_o(D_0)'$, then it follows that $A = aI$ since Ω is separating for $\mathscr{B}_o(D_0)'$.

Lemma 2.7: Let $W_1, W_2 \in \mathscr{W}$ be such that $W_1 \cap W_2 = \varnothing = W_1^c \cap W_2^c$. Let A and B be two bounded operators which satisfy the condition

$$\langle\Omega|[B,T(x)AT(x)^{-1}]\Omega\rangle=0, \tag{2.11}$$

for all $x\in W_1\cup W_2$. Then the relation in (2.11) holds for *all* x, and

$$\langle\Omega|BT(x)A\Omega\rangle=\langle\Omega|BA\Omega\rangle=\langle\Omega|B\Omega\rangle\langle\Omega|A\Omega\rangle, \tag{2.12}$$

for all $x\in\mathsf{M}$.

Proof: Under the stated premises the edges of the two wedges must be parallel, and the upper plane of one of the wedges must lie *above* the lower plane of the other. Without any loss of generality we assume that the upper plane of W_1 lies above the lower plane of W_2. The lower plane of W_1 then either coincides with the upper plane of W_2, or else lies above the upper plane of W_2. With a suitable choice of a standard coordinate system the two wedges can then be described by

$$W_1=\{x|x_3+\rho>|x_4-\tau|\},\quad W_2=\{x|x_3-\rho<-|x_4+\tau|\},$$

for some $\tau>0$, with $\rho=\tau$ if the lower plane of W_1 coincides with the upper plane of W_2, and $\rho=0$ if the lower plane of W_1 lies above the upper plane of W_2. We will depend on a result in Lemma 5.17 in TW-II,[7] and in order to conform to the notation in the quoted paper we introduce the two wedges $W_3=\{x|x_3>|x_4-\tau|+\tau/2\}$ and $W_4=\{x|x_3<-|x_4+\tau|-\tau/2\}$. We easily see that we then have $W_3\cup W_4\subset W_1\cup W_2$, for $\rho=\tau$ as well as for $\rho=0$.

We consider the function $\eta(A,B;x)=\langle B^*\Omega|T(x)A\Omega\rangle-\langle A^*\Omega|T(-x)B\Omega\rangle$. By the condition in (2.11) we have $\eta(A,B;x)=0$ for all $x\in W_3\cup W_4$. The function $\eta(A,B;x)$ is analogous to the function denoted $\eta(X,B;x)$ in the quoted paper, and the reasoning which led to the conclusion that $\eta(X,B;x)=0$ for all x also applies in the present case. We thus have $\eta(B,A;x)=0$ for all x, and from this the relations in (2.12) follow by Lemma 2.6.

III. CONSTRUCTION OF EXTENDED LOCAL NETS

In this section we discuss various extensions of the basic net $\mathscr{N}_o$ described in Definition 2.3. Our primary interest is in extensions in which we associate von Neumann algebras with the sets K and K^c, for all $K\in\mathscr{K}_C$. We recall from Definition 1.1 that $\mathscr{K}_C$ is the set of all compact, convex, causally complete subsets of M. We define a particularly canonical net of this nature as follows.

Definition 3.1: By a *regular net* $\mathscr{N}$ is understood a system of von Neumann algebras of operators on $\mathscr{H}$ with the following properties.

(a) To every $W\in\mathscr{W}$ corresponds a von Neumann algebra $\mathscr{A}(W)=\mathscr{M}(W)$, and to every $K\in\mathscr{K}_C$ correspond two von Neumann algebras $\mathscr{B}(K)=\mathscr{M}(K)$ and $\mathscr{A}(K^c)=\mathscr{M}(K^c)$. These algebras constitute an irreducible local net which satisfies *all* the Conditions I–IX in Definition 2.1.

(b) The von Neumann algebras of the net $\mathscr{N}$ are related by

$$\mathscr{B}(K)=\cap\{\mathscr{A}(W)|W\in\mathscr{W},\ \overline{W}\supset K\},\quad \mathscr{A}(K^c)=\{\mathscr{A}(W)|W\in\mathscr{W},\ W\subset K^c\}'', \tag{3.1}$$

$$\mathscr{A}(W)=\{\mathscr{B}(K)|K\in\mathscr{K}_C,\ K\subset W\}'', \tag{3.2}$$

for all $K\in\mathscr{K}_C$, and all $W\in\mathscr{W}$.

(c) The net $\mathscr{N}$ satisfies the *condition of duality* in the sense that

$$\mathscr{A}(W)'=\mathscr{A}(\overline{W^c})^z,\quad \mathscr{B}(K)'=\mathscr{A}(K^c)^z, \tag{3.3}$$

for all $W\in\mathscr{W}$ and all $K\in\mathscr{K}_C$.

(d) For any $K\in\mathscr{K}_C$ the vector Ω is *separating* for $\mathscr{B}(K)$ and *cyclic* for $\mathscr{A}(K^c)$. If K has a nonempty interior, then Ω is also *cyclic* for $\mathscr{B}(K)$ and *separating* for $\mathscr{A}(K^c)$. The vector Ω is *cyclic* and *separating* for every wedge-algebra $\mathscr{A}(W)$.

(e) We denote by $\mathscr{U}$ the *algebra of all local operators of the net*, defined as $\mathscr{U}=\cup\{\mathscr{B}(K)|K\in\mathscr{K}_C\}$. The net $\mathscr{N}$ is *irreducible* in the sense that the algebra $\mathscr{U}$ is irreducible, and hence

$$\mathscr{U}''=\{\mathscr{B}(K)|K\in\mathscr{K}_C\}''=\{\mathscr{A}(W)|W\in\mathscr{W}\}''=\mathsf{B}(\mathscr{H}),\quad \mathscr{B}(\varnothing)=\mathsf{C}^1\cdot I.$$

Like Definition 2.3 the above is also a description of a scenario. Some additional features are described in Proposition 3.6. In what follows we will mostly deal with nets so described. The term ''regular net'' has no official standing, but for the purposes of this paper we will employ it as a convenient technical term.

We now want to prove that every *basic* net has a unique extension to a *regular* net, as described in the above definition. This extension depends crucially on certain geometric relationships discussed in TW-I,[6] which we here state in the form of two propositions and two lemmas. For the proofs we refer to the quoted paper. The second proposition describes relevant characteristics of the sets $\mathscr{K}_E$ and $\mathscr{K}_C$, and the first proposition is intended to provide a more general perspective on the issues. The two lemmas which follow are of a more technical nature.

Proposition 3.2: (a) Let K be a closed, convex, causally complete subset of M such that K^c has a nonempty interior. Then

$$K=\cap\{\overline{W}|W\in\mathscr{W},\ \overline{W}\supset K\}=\cap\{W^c|W\in\mathscr{W},\ W\subset K^c\},$$

$$(K^c)^o=\cup\{W|W\in\mathscr{W},\ W^c\supset K\}=\cup\{W|W\in\mathscr{W},\ W\subset K^c\},$$

and

$$((K^c)^o)^c=K,\quad \overline{(K^c)^o}=\overline{K^c},\quad (\overline{K^c})^o=(K^c)^o.$$

The set K has an empty interior if and only if it is contained in some characteristic plane.

(b) Let $\{K_\alpha|\alpha\in A\}$ be a family of subsets of M which satisfy the premises in part (a). Then $K=\cap\{K_\alpha|\alpha\in A\}$ also satisfies the premises in part (a).

(c) Let $\mathscr{F}$ be any family of wedges in $\mathscr{W}$, and let $S=\cup\{W|W\in\mathscr{F}\}$. Then $K=\cap\{W^c|W\in\mathscr{F}\}=S^c$ satisfies the premises in part (a).

(d) Let K be a closed, convex, causally complete *proper* subset of M. Then K^c has an empty interior if and only if K contains an isotropic (straight) line, and if this is the case, then both K and K^c are contained in some characteristic plane.

The above proposition is a slightly abbreviated version of Theorem 3.2 in TW-I.[6]

Proposition 3.3: (a) Every set $K\in\mathscr{K}_E$ satisfies the premises in part (a) of Proposition 3.2, and specifically K^c is nonempty and *open.*

(b) If K_0 is any closed, convex, causally complete subset of M, then $K_0\cap K\in\mathscr{K}_E$ if $K\in\mathscr{K}_E$, and $K_0\cap K\in\mathscr{K}_C$ if $K\in\mathscr{K}_C$.

(c) If $\mathscr{F}_c\subset\mathscr{K}_C$, then $\cap\{K|K\in\mathscr{F}_c\}\in\mathscr{K}_C$. If $\mathscr{F}_e\subset\mathscr{K}_E$, then $\cap\{K|K\in\mathscr{F}_e\}\in\mathscr{K}_E$, and if at least one of the elements in $\mathscr{F}_e$ is in $\mathscr{K}_C$, then $\cap\{K|K\in\mathscr{F}_e\}\in\mathscr{K}_C$.

(d) If $K\in\mathscr{K}_E$, then there exist two *spacelike* 3-planes σ' and σ'' such that K lies *below* σ' and *above* σ''.

(e) If $K\in\mathscr{K}_E$, then K contains *no* (infinite) line, and *no* timelike or isotropic half-line.

This proposition is a verbatim quote of Proposition 3.4 in TW-I.[6] For a detailed discussion of the sets $\mathscr{K}_E$ and $\mathscr{K}_C$ we refer to Section III in the quoted paper. We note here the rather trivial fact, stated in the proposition, that the sets $\mathscr{K}_E$ and $\mathscr{K}_C$ are closed under intersections, in contrast to what the situation is for the set $\mathscr{D}$ of double cones. We also note that the sets $\mathscr{K}_E$ and $\mathscr{K}_C$ include subsets K of M with empty interiors, which are then necessarily contained in some characteristic plane.

Lemma 3.4: Let $K\in\mathscr{K}_C$, i.e., K is a compact, convex, causally complete subset of M. Then:

(a)

$$K=\cap\{D^o|D\in\mathscr{D},\ D^o\supset K\}=\cap\{D|D\in\mathscr{D},\ D\supset K\}.$$

(b) If $W\in\mathscr{W}$ and $W\supset K$, then there exists a double cone $D\in\mathscr{D}$, such that $W\supset D\supset D^o\supset K$.

The above lemma is a verbatim quote of Proposition 3.8 in TW-I.[6]

Lemma 3.5: (a) Let R be a nonempty convex subset of M, and let S be a nonempty *open* convex subset of M, such that $S\subset R^c$, and hence $R\subset S^c$. There then exists a wedge $W\in\mathscr{W}$ such that $S\subset W$ and $R\subset R^{cc}\subset W^c$.

(b) Let $K_1, K_2 \in \mathscr{K}_E$ be such that $K_1 \subset K_2^c$. Suppose furthermore that K_1 and K_2 are not both *unbounded* subsets of the *same* characteristic plane. There then exists a wedge $W \in \mathscr{W}$ such that $K_1 \subset \overline{W}$ and $K_2 \subset W^c$.

In particular a wedge W with the stated property always exists if $K_1, K_2 \in \mathscr{K}_C$.

Part (a) of this lemma is a verbatim quote of Proposition 3.1, and part (b) is a verbatim quote of Proposition 3.7 in TW-I.[6] We note here that the conclusion in part (b) can actually fail if K_1 and K_2 lie in the same characteristic plane, and neither K_1 nor K_2 is in $\mathscr{K}_C$. An example of this is given in the quoted paper.

Before we describe the construction of a regular net from a basic net we note some trivial features of a regular net.

Proposition 3.6: Let $\mathscr{N}$ be a *regular net*, with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$ for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$, as in Definition 3.1. Then:

(a) The set of algebras $\{\mathscr{A}(W) | W \in \mathscr{W}\}$ and $\{\mathscr{B}(D), \mathscr{A}(D^c) | D \in \mathscr{D}\}$ are the algebras of a *basic net*, as described in Definition 2.3, and we have

$$\mathscr{A}(W) = \{\mathscr{B}(K) | K \in \mathscr{K}_C,\ K \subset W\}'' = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset W\}''. \tag{3.4}$$

(b) Furthermore,

$$\mathscr{A}(K^c) = \{\mathscr{B}(K_i) | K_i \in \mathscr{K}_C,\ K_i \subset K^c\}'' = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset K^c\}'', \tag{3.5}$$

$$\mathscr{A}(W) = \cap\{\mathscr{A}(K^c) | K \in \mathscr{K}_C,\ K^c \supset W\} = \cap\{\mathscr{A}(D^c) | D \in \mathscr{D},\ D^c \supset W\}, \tag{3.6}$$

$$\mathscr{B}(K) = \cap\{\mathscr{A}(K_i^c) | K_i \in \mathscr{K}_C,\ K_i^c \supset K\} = \cap\{\mathscr{A}(D^c) | D \in \mathscr{D},\ D^c \supset K\}, \tag{3.7}$$

$$\mathscr{B}(K) = \cap\{\mathscr{A}(W) | W \in \mathscr{W},\ W \supset K\}, \quad \mathscr{A}(K^c) = \{\mathscr{A}(W) | W \in \mathscr{W},\ \overline{W} \subset K^c\}'', \tag{3.8}$$

for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$.

(c) For any $K \in \mathscr{K}_C$ we have

$$\mathscr{B}(K) = \cap\{\mathscr{B}(D) | D \in \mathscr{D},\ D^o \supset K\}. \tag{3.9}$$

(d) Two regular nets are equal if their algebras are equal on the subdomain $\mathscr{W}$, or on the subdomain $\mathscr{D}$.

Proof: (1) We consider part (a). Comparing the Definitions 2.3 and 3.1 we see that the only issue to settle is whether the relation in (3.2) implies the relation at left in (2.3), i.e., whether the equality of the first two members in (3.4) implies that they are equal to the third member. We first note that since $\mathscr{D} \subset \mathscr{K}_C$ it follows that $\mathscr{A}(W) = \{\mathscr{B}(K) | K \in \mathscr{K}_C,\ K \subset W\}'' \supset \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset W\}''$. Let $K \in \mathscr{K}_C$ and $K \subset W$. By Lemma 3.4 there then exists a double cone D such that $K \subset D \subset W$. By isotony this implies that $\{\mathscr{B}(K) | K \in \mathscr{K}_C,\ K \subset W\}'' \subset \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset W\}''$, and hence we have $\mathscr{A}(W) = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset W\}''$.

(2) We consider part (b). The identities in (3.5) follow from the identities in (3.4) and the relation at right in (3.1). The identities in (3.6) are equivalent with the identities in (3.4) under duality, and the identities in (3.7) are equivalent with the identities in (3.5) under duality. In view of the continuity relations in (2.6) for wedge-algebras, the relations in (3.1) trivially imply the relations in (3.8), and *vice versa.*

(3) We consider part (c). Let $K \in \mathscr{K}_C$, and let $W \in \mathscr{W}$ and $W \supset K$. By Lemma 3.4 there then exists a double cone $D_W \in \mathscr{D}$ such that $W \supset D_W \supset D_W^o \supset K$. By isotony we have $\mathscr{A}(W) \supset \mathscr{B}(D_W) \supset \cap\{\mathscr{B}(D) | D \in \mathscr{D},\ D^o \supset K\} \supset \mathscr{B}(K)$, and it follows that $\mathscr{B}(K) = \cap\{\mathscr{A}(W) | W \in \mathscr{W},\ W \supset K\} \supset \cap\{\mathscr{B}(D) | D \in \mathscr{D},\ D^o \supset K\} \supset \mathscr{B}(K)$. This proves the relation in (3.9).

The assertions in part (d) are trivial in view of the definitions in (3.1), and in view of what has been shown. This completes the proof.

We are now prepared for the extension of an ''original'' *basic* net $\mathscr{N}_o$.

Theorem 3.7: Let $\mathscr{N}_o$ be a *basic net*, with algebras $\mathscr{A}_o(W)$, $\mathscr{B}_o(D)$, and $\mathscr{A}_o(D^c)$, as described in Definition 2.3. There then exists a *unique regular net* $\mathscr{N}$, with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$, as described in Definition 3.1, such that $\mathscr{B}(D) = \mathscr{B}_o(D)$ for all $D \in \mathscr{D}$. The

net $\mathscr{N}$ is an *extension* of the net $\mathscr{N}_o$ in the sense that we also have $\mathscr{A}(D^c)=\mathscr{A}_o(D^c)$ for all $D\in\mathscr{D}$, and $\mathscr{A}(W)=\mathscr{A}_o(W)$ for all $W\in\mathscr{W}$. Furthermore, with $\mathscr{U}_o$ as in Definition 2.3, we have $\mathscr{U}=\mathscr{U}_o$, i.e.,

$$\cup\{\mathscr{B}(K)\,|\,K\in\mathscr{K}_C\}=\cup\{\mathscr{B}_o(D)\,|\,D\in\mathscr{D}\}, \tag{3.10}$$

and the *local* operators of $\mathscr{N}$ and $\mathscr{N}_o$ are thus precisely the same.

The extension $\mathscr{N}$ is also uniquely determined by the condition that $\mathscr{N}$ is a regular net which satisfies the condition $\mathscr{A}(W)=\mathscr{A}_o(W)$ for some $W\in\mathscr{W}$.

Proof: (1) For each $W\in\mathscr{W}$ we write $\mathscr{A}(W)=\mathscr{A}_o(W)$, and for each $K\in\mathscr{K}_C$ we define the von Neumann algebras $\mathscr{B}(K)$ and $\mathscr{A}(K^c)$ by

$$\mathscr{B}(K)=\cap\{\mathscr{A}_o(W)\,|\,W\in\mathscr{W},\ \overline{W}\supset K\},\quad \mathscr{A}(K^c)=\{\mathscr{A}_o(W)\,|\,W\in\mathscr{W},\ W\subset K^c\}''. \tag{3.11}$$

We want to show that the system $\mathscr{N}$ of algebras so defined is a regular net, related to the net $\mathscr{N}_o$ as described above. We first note that we have $\mathscr{B}(D)=\mathscr{B}_o(D)$ and $\mathscr{A}(D^c)=\mathscr{A}_o(D^c)$ for any $D\in\mathscr{D}\subset\mathscr{K}_C$, in view of (2.2) and the definitions in (3.11). It is trivial that the system $\mathscr{N}$ is irreducible, since the net $\mathscr{N}_o$ is irreducible.

The Poincaré- and *PCT*-covariance of the system $\mathscr{N}$ follows trivially from the Poincaré- and *PCT*-covariance of the mapping $W\to\mathscr{A}_o(W)=\mathscr{A}(W)$. We next have to show that the system $\mathscr{N}$ satisfies the conditions of isotony, locality, and duality.

(2) We consider the condition of isotony. The relation $\mathscr{A}(W_1)\subset\mathscr{A}(W_2)$ if $W_1\subset W_2$ is included in the premises. It follows at once from the definitions in (3.11) that if $K_1,K_2\in\mathscr{K}_C$ and $K_1\supset K_2$ (and hence $K_1^c\subset K_2^c$), then $\mathscr{B}(K_1)\supset\mathscr{B}(K_2)$ and $\mathscr{A}(K_1^c)\subset\mathscr{A}(K_2^c)$. It is trivial that $\mathscr{B}(K)\subset\mathscr{A}(W)$ if $K\subset\overline{W}$, and that $\mathscr{A}(W)\subset\mathscr{A}(K^c)$ if $W\subset K^c$. Suppose now that $K_1,K_2\in\mathscr{K}_C$ and $K_1\subset K_2^c$. By Lemma 3.5 there then exists a wedge $W\in\mathscr{W}$ such that $K_1\subset\overline{W}$ and $K_2\subset W^c$. The latter relation is equivalent with the relation $W\subset K_2^c$. By the definitions in (3.11) we then have $\mathscr{B}(K_1)\subset\mathscr{A}(W)\subset\mathscr{A}(K_2^c)$. Since we cannot have $W\subset K$, or $K^c\subset W$, or $K^c\subset K_1$ for any $K,K_1\in\mathscr{K}_C$ and any $W\in\mathscr{W}$, we have thus proved all the isotony relations.

(3) To prove the conditions of locality, we have to consider a number of different cases, as in the case of isotony. The locality condition for the case of two spacelike separated wedges in $\mathscr{W}$ is included in the premises. Suppose now that $K\in\mathscr{K}_C$ and $W\in\mathscr{W}$, and that $K\subset W^c$. We then have $\mathscr{B}(K)^z\subset\mathscr{A}((W^c)^o)^z\subset\mathscr{A}(W)'$, taking into account the fact that $(W^c)^o=\overline{W^c}$, and the locality condition thus holds for this case. We next consider the case when $K_1,K_2\in\mathscr{K}_C$ and $K_1\subset(K_2^c)^c=K_2$. Let $W\subset K_2^c$, and hence $K_1\subset K_2\subset W^c$. By the above we thus have $\mathscr{A}(W)\subset(\mathscr{B}(K_1)^z)'$, and since this holds for every $W\subset K_2^c$ it follows from the definition at right in (3.11) that $\mathscr{A}(K_2^z)\subset(\mathscr{B}(K_1)^z)'$, which proves the locality condition in this case.

Let $K_1,K_2\in\mathscr{K}_C$, and $K_1\subset K_2^c$. By Lemma 3.5 there then exists a wedge $W\in\mathscr{W}$ such that $K_1\subset\overline{W}$ and $K_2\subset W^c$. By the above we have $\mathscr{B}(K_2)^z\subset\mathscr{A}(W)'$, and since $\mathscr{B}(K_1)\subset\mathscr{A}(W)$, and hence $\mathscr{A}(W)'\subset\mathscr{B}(K_1)'$, it follows that $\mathscr{B}(K_2)^z\subset\mathscr{B}(K_1)'$. This proves the locality condition for two spacelike separated subsets $K_1,K_2\in\mathscr{K}_C$.

It remains to consider the cases of two regions K^c or a wedge W and a region K^c. The salient point is now that the question of locality *does not apply* to such configurations. If $K\in\mathscr{K}_C$ and $W\in\mathscr{W}$, then W cannot be spacelike separated from K^c. The relation $K^c\subset W^c$ would imply that $W\subset K$, which is impossible, and the relation $K_1^c\subset(K_2^c)^c=K_2$ is likewise impossible since K_1^c is a union of wedges in $\mathscr{W}$.

(4) We next prove that the extended net satisfies the condition of duality in part (c) in Definition 3.1. The relation at left in (3.3) is included in the premises. To prove the relation at right in (3.3) we form the commutant of both members in the relation at right in (3.11), and we obtain $\mathscr{A}(K^c)'=\cap\{\mathscr{A}(W)'\,|\,W\in\mathscr{W},\ W\subset K^c\}=\cap\{\mathscr{A}(\overline{W^c})^z\,|\,W\in\mathscr{W},\ W\subset K^c\}$. With $W_1=\overline{W^c}$ the relation $W\subset K^c$ is equivalent with the relation $\overline{W_1}\supset K$, and hence we obtain $\mathscr{A}(K^c)'=(\cap\{\mathscr{A}(W_1)\,|\,W_1\in\mathscr{W},\ \overline{W_1}\supset K\})^z=\mathscr{B}(K)^z$.

(5) We have now shown that the system $\mathscr{N}$ is a local net. By (2.3) we have $\mathscr{A}(W)=\{\mathscr{B}(D)\,|\,D\in\mathscr{D},\ D\subset W\}''$, and it then follows from the relation at right in (3.11) that we also

have $\mathscr{A}(K^c)=\{\mathscr{B}(D)|D\in\mathscr{D},\ D\subset K^c\}''$ for any $K\in\mathscr{K}_C$. Hence the net $\mathscr{N}$ satisfies Condition IX in Definition 2.1, and we conclude from the above that $\mathscr{N}$ satisfies *all* the Conditions I–IX in Definition 2.1.

We have $\mathscr{A}(W)\supset\{\mathscr{B}(K)|K\in\mathscr{K}_C,\ K\subset W\}''\supset\{\mathscr{B}(D)|D\in\mathscr{D},\ D\subset W\}''=\mathscr{A}(W)$, which proves the relations in (3.4). Since every $K\in\mathscr{K}_C$ is contained in some double cone $D\in\mathscr{D}$, and since we obviously have $\cup\{\mathscr{B}(K)|K\in\mathscr{K}_C\}\supset\cup\{\mathscr{B}(D)|D\in\mathscr{D}\}$, the identity in (3.10) follows.

It is trivial that the net of $\mathscr{N}$ satisfies the conditions in part (d) of Definition 3.1. Hence $\mathscr{N}$ is a *regular* net. The asserted uniqueness of a *regular* extension of $\mathscr{N}_o$ follows trivially from Proposition 3.6. This completes the proof.

As we see from the above, the extension of the basic net to a regular net, as described in Theorem 3.7, is not at all an ''enlargement'' of the set $\mathscr{U}_o$ of local operators. What is involved is rather a more refined assignment of the operators in $\mathscr{U}_o$ to subsets of Minkowski spacetime M. The set $\mathscr{U}_o=\mathscr{U}$ is, in fact, *maximal* in the sense that we cannot add any new local operator Q to the set. Such an operator would have to be in the commutant of one of the algebras $\mathscr{A}_o(D^c)^z$, and hence in $\mathscr{B}_o(D)\subset\mathscr{U}_o$. The existence of a unique regular net, as described in Definition 3.1, is thus *implicit* in the notion of a basic net, as described in Definition 2.3, and involves no new assumptions.

We could have shortened the above somewhat pedantic proof by relying on Lemma 2.2, according to which the conclusions in step (3) follow from the results in steps (2) and (4), but we felt that it is of interest to see in detail how the properties of isotony and locality follow from the corresponding properties of the set of wedge-algebras, independently of duality. We note that the locality condition involving two spacelike separated sets K_1 and K_2 is not entirely trivial, since the conclusion depends on the facts stated in Lemma 3.5, but the locality conditions for all the other possible configurations are indeed trivial. Likewise the proof of the corresponding isotony relation $\mathscr{B}(K_1)\subset\mathscr{A}(K_2^c)$ involves Lemma 3.5, whereas all the other isotony relations are trivial.

A *regular* net is completely determined by its ''wedge-algebras'' $\mathscr{A}(W)$, in view of the relations (3.1) or (3.8). It is also completely determined by the particular local algebras $\mathscr{B}(D)$, with $D\in\mathscr{D}$, in view of (3.4). The idea behind our notation in this paper is that a $\mathscr{B}$-algebra of a (closed) set $K\in\mathscr{K}_C$ is the intersection of all the $\mathscr{A}$-algebras of the (open) sets which contain K, and that an $\mathscr{A}$-algebra of an open region R is the von Neumann algebra generated by all $\mathscr{B}$-algebras of the closed sets in $\mathscr{K}_C$ contained in R. [We noted before the slight inconsistency in that we write $\mathscr{A}(W)$ for $\mathscr{B}(\overline{W})$.] The local $\mathscr{B}$-algebras are, in a sense, the objects of primary interest, and the $\mathscr{A}$-algebras can be regarded as ''auxiliary'' objects. The important point is that the $\mathscr{A}$-algebras are generated by the local operators which they contain, as shown in (3.4) and (3.5). The duality condition relates the $\mathscr{B}$-algebras to the $\mathscr{A}$-algebras, but since the net $\mathscr{N}$ has a *strictly local structure* in the sense of Condition IX in Definition 2.1, the duality condition is really a condition on the *local* algebras, i.e., *Haag duality* obtains for the net.

Since the algebra $\mathscr{U}$ of all local operators for the extended net $\mathscr{N}$ is the same as the algebra $\mathscr{U}_o$ for the original basic net $\mathscr{N}_o$, we expect that the algebra $\mathscr{B}(K)$ for any $K\in\mathscr{K}_C$ can be obtained ''directly'' from the algebras $\mathscr{B}(D)=\mathscr{B}_o(D)$ associated with double cones. That this is so is shown by the relation (3.9) in Proposition 3.6.

The extension principle in Theorem 3.7 has a more general applicability, which we illustrate as follows.

Proposition 3.8: Let $\mathscr{N}_a$ be a local net (not necessarily irreducible) with domain $\mathscr{D}$, with von Neumann algebras $\mathscr{B}_a(D)$. The net $\mathscr{N}_a$ thus satisfies Conditions I–V in Definition 2.1.

We define the von Neumann algebras $\mathscr{B}_e(K)$, $\mathscr{A}_e(W)$, and $\mathscr{A}_e(K^c)$ by

$$\mathscr{A}_e(W)=\{\mathscr{B}_a(D)|D\in\mathscr{D},\ D\subset W\}'' \tag{3.12}$$

and

$$\mathscr{B}_e(K)=\cap\{\mathscr{A}_e(W)|W\in\mathscr{W},\ \overline{W}\supset K\},\quad \mathscr{A}_e(K^c)=\{\mathscr{A}_e(W)|W\in\mathscr{W},\ W\subset K^c\}'', \tag{3.13}$$

for all $K\in\mathscr{K}_C$ and all $W\in\mathscr{W}$. Then:

(a) The algebras $\mathcal{B}_e(K)$, $\mathcal{A}_e(W)$, and $\mathcal{A}_e(K^c)$ are the von Neumann algebras of a local net $\mathcal{N}_e$ which satisfies Conditions I–V and IX in Definition 2.1. Furthermore the net satisfies the condition

$$\mathcal{A}_e(W)=\{\mathcal{B}_e(K)|K\in\mathcal{K}_C,\ K\subset W\}''=\{\mathcal{B}_e(D)|D\in\mathcal{D},\ D\subset W\}'', \tag{3.14}$$

for all $W\in\mathcal{W}$. The net $\mathcal{N}_e$ satisfies the condition of *PCT*-covariance if the net $\mathcal{N}_a$ satisfies this condition.

(b) The nets $\mathcal{N}_e$ and $\mathcal{N}_a$ are related by

$$\mathcal{B}_e(D)\subset\mathcal{B}_a(D), \tag{3.15}$$

for all $D\in\mathcal{D}$.

We write $\mathcal{U}_a=\cup\{\mathcal{B}_a(D)|D\in\mathcal{D}\}$ for the algebra of all local operators of the net $\mathcal{N}_a$, and we write $\mathcal{U}_e=\cup\{\mathcal{B}_e(K)|K\in\mathcal{K}_C\}$ for the algebra of all local operators of the net $\mathcal{N}_e$. We then have

$$\mathcal{U}_e\supset\mathcal{U}_a,\quad \mathcal{U}_a''=\mathcal{U}_e=\mathcal{A}_e(\mathsf{M}). \tag{3.16}$$

(c) The sub-Hilbert space $\mathcal{H}_a=\overline{\mathcal{U}_a\Omega}$ is Poincaré-invariant, and we have

$$\overline{\mathcal{A}_e(W)\Omega}=\overline{\mathcal{A}_e(K^c)\Omega}=\overline{\mathcal{U}_e\Omega}=\overline{\mathcal{U}_a\Omega}=\mathcal{H}_a, \tag{3.17}$$

for all $W\in\mathcal{W}$ and all $K\in\mathcal{K}_C$.

(d) Suppose that a local net $\mathcal{N}_b$ with domain $\mathcal{K}_C$, with von Neumann algebras $\mathcal{B}_b(K)$ defined for all $K\in\mathcal{K}_C$, is an *extension* of the net $\mathcal{N}_a$ such that $\mathcal{B}_b(D)=\mathcal{B}_a(D)$ for all $D\in\mathcal{D}$. We then have $\mathcal{U}_b=\mathcal{U}_a$, with $\mathcal{U}_b=\cup\{\mathcal{B}_b(K)|K\in\mathcal{K}_C\}$, and

$$\mathcal{A}_e(W)=\{\mathcal{B}_b(K)|K\in\mathcal{K}_C,\ K\subset W\}'', \tag{3.18}$$

$$\mathcal{B}_e(K)\supset\mathcal{B}_b(K), \tag{3.19}$$

for all $W\in\mathcal{W}$ and all $K\in\mathcal{K}_C$.

Proof: (1) The Poincaré-covariance of the mapping $W\to\mathcal{A}_e(W)$ follows from the Poincaré-covariance of the net $\mathcal{N}_a$, in view of the definition in (3.12). The Poincaré-covariance of the system N_e then follows, in view of the definitions in (3.13). Similarly we conclude that the system $\mathcal{N}_e$ satisfies the condition of *PCT*-covariance if the net $\mathcal{N}_a$ satisfies this condition.

From the definition in (3.12) it also follows trivially that the mapping $W\to\mathcal{A}_e(W)$ satisfies the conditions of isotony and locality. By the same reasoning as in steps (2) and (3) in the proof of Theorem 3.7 we then conclude that the system $\mathcal{N}_e$ satisfies all the relevant conditions of isotony and of locality.

(2) By (3.12) and (3.13) we have $\mathcal{B}_a(D)\subset\cap\{\mathcal{A}_e(W)|W\in\mathcal{W},\ \overline{W}\supset D\}=\mathcal{B}_e(D)$, for every $D\in\mathcal{D}$. Hence we have $\mathcal{U}_e\supset\mathcal{U}_a$, and we also readily conclude that $\mathcal{U}_a''=\mathcal{A}_e(\mathsf{M})=\mathcal{U}_e''$. With the results in step (1) this implies that $\mathcal{N}_e$ is a local net.

Since $\mathcal{B}_e(D)\subset\mathcal{A}_e(W)$ if $D\subset W$ it follows that $\mathcal{A}_e(W)\supset\{\mathcal{B}_e(D)|D\in\mathcal{D},\ D\subset W\}''\supset\{\mathcal{B}_a(D)|D\in\mathcal{D},\ D\subset W\}''=\mathcal{A}_e(W)$. The relation (3.14) follows, and in view of the definition of the algebras $\mathcal{A}_e(K^c)$ we conclude that the net $\mathcal{N}$ satisfies Condition IX in Definition 2.1. We have thus proved the assertions in parts (a) and (b).

(3) We consider part (c). The subspace $\mathcal{H}_a$ is trivially Poincaré-invariant. The relation $\overline{\mathcal{U}_e\Omega}=\overline{\mathcal{U}_a\Omega}$ follows from the relation $\mathcal{U}_a''=\mathcal{U}_e''$. The remaining relations in (3.17) follow by a well-known argument involving the *Reeh–Schlieder Principle*.

(4) We consider part (d). Let $K\in\mathcal{K}_C$, $W\in\mathcal{W}$, and $K\subset W$. By Lemma 3.4 there then exists a double cone D such that $K\subset D\subset W$. Hence $\mathcal{B}_b(K)\subset\mathcal{B}_b(D)=\mathcal{B}_a(D)\subset\mathcal{A}_e(W)$. From this we conclude that $\mathcal{U}_b\subset\mathcal{U}_a$, and since the opposite inclusion is trivial, it follows that $\mathcal{U}_b=\mathcal{U}_a$. Furthermore we conclude that the left member in (3.18) includes the right member, and since the latter obviously includes the right member in (3.12), we obtain the identity in (3.18). The inclusion relation in (3.19) is proved by the same reasoning as in the proof of the relation in (3.15). This completes the proof.

The net $\mathcal{N}_e$ is a rather natural extension of an original local net $\mathcal{N}_a$ with domain $\mathcal{D}$, or of an original local net $\mathcal{N}_b$ with domain $\mathcal{K}_C$. It is easy to construct examples for which $\mathcal{B}_a(D)$ is smaller than $\mathcal{B}_e(D)$. It can also happen that $\mathcal{U}_a$ is smaller than $\mathcal{U}_e$. Examples of this possibility can be constructed within the theory of generalized free fields. The local operators in $\mathcal{U}_e$ are, however, implicit in the net $\mathcal{N}_a$, ''hidden'' in the wedge-algebras $\mathcal{A}_e(W)$ constructed as in (3.12), and the relations (3.16) and (3.17) are always valid. If we apply the proposition to the case when the net $\mathcal{N}_a$ is the restriction of a *basic net* $\mathcal{N}_o$ to the domain $\mathcal{D}$, the extension $\mathcal{N}_e$ will be equal to the *regular net* $\mathcal{N}$ in Theorem 3.7. This circumstance, and the fact that the net $\mathcal{N}_e$ satisfies the relations (3.13) and (3.14), provide justification for regarding such a net as a *standard form* of a local net with a domain consisting of all wedges in $\mathcal{W}$ and all $K \in \mathcal{K}_C$ and their causal complements K^c.

In Theorem 3.7 we noted the trivial fact that the basic net $\mathcal{N}_o$ and the extended net $\mathcal{N}$ have the same set of local operators. The proposition which follows illustrates the ''rigid structure'' of the set of local operators of a regular net under certain circumstances.

Proposition 3.9: Let $\mathcal{N}$ be a *regular* net which satisfies the *special condition of duality*, as described in Definition 2.4. Let $\mathcal{N}_e$, with von Neumann algebras $\mathcal{A}_e(W)$, $\mathcal{B}_e(K)$, and $\mathcal{A}_e(K^c)$ for all $W \in \mathcal{W}$ and all $K \in \mathcal{K}_C$, be another local net (relative to the *same* representation of the group $\hat{\mathsf{P}}$, on the same Hilbert space). We assume that $\mathcal{N}_e$ is of the *standard form* described in Proposition 3.8, i.e., the net satisfies the conditions in (3.13) and in (3.14). Let $\mathcal{U}$ be the set of all local operators of the net $\mathcal{N}$, and similarly, let $\mathcal{U}_e = \cup\{\mathcal{B}_e(K) | K \in \mathcal{K}_C\}$. Suppose that Ω is *cyclic* for the set $\mathcal{S} = \mathcal{U} \cap \mathcal{U}_e$. Then $\mathcal{N} = \mathcal{N}_e$.

Proof: Let $B \in \mathcal{S}$. We then have $B \in \mathcal{B}(D_1) \cap \mathcal{B}_e(D_2)$ for some $D_1, D_2 \in \mathcal{D}$. There then exists a nonempty open subset X_B of W_R such that the translates of D_1 and D_2 by any $x \in X_B$ are contained in W_R, and hence $T(x)BT(x)^{-1} \in \mathcal{A}(W_R) \cap \mathcal{A}_e(W_R)$ for all $x \in X_B$. From this we conclude, by the *Reeh–Schlieder Principle*, that Ω is cyclic for $\mathcal{A}(W_R) \cap \mathcal{A}_e(W_R) = \mathcal{A}_R$, since Ω is cyclic for the set $\mathcal{S} = \mathcal{U} \cap \mathcal{U}_e$. The algebra $\mathcal{A}_R$ is obviously invariant under conjugation by any velocity transformation $V(\tau)$ in the 3-direction. Since the net $\mathcal{N}$ satisfies the *special condition of duality*, and since $\mathcal{A}_R \subset \mathcal{A}(W_R)$, it follows that $\mathcal{A}_R = \mathcal{A}(W_R)$, and hence $\mathcal{A}_e(W_R) \supset \mathcal{A}(W_R)$. Similarly we have $\mathcal{A}_e(W_L) \supset \mathcal{A}(W_L)$, and hence $\mathcal{A}_e(W_R)^z \subset \mathcal{A}_e(W_L)' \subset \mathcal{A}(W_L)' = \mathcal{A}(W_R)^z$. We conclude that $\mathcal{A}_e(W_R) = \mathcal{A}(W_R)$, from which it follows readily that $\mathcal{N}_e = \mathcal{N}$, i.e., we also have $\mathcal{B}_e(K) = \mathcal{B}(K)$ and $\mathcal{A}_e(K^c) = \mathcal{A}(K^c)$. This completes the proof.

That two irreducible local nets of the standard form should be identical if they have ''sufficiently many'' local operators in common seems rather plausible, and this conclusion does hold if at least one of the nets satisfies the *special condition of duality*, but the conclusion might possibly fail if this is not the case. The above proposition rules out any ''reassignment'' of regions for the local operators. We have $B \in \mathcal{B}_e(K)$ if and only if $B \in \mathcal{B}(K)$. Note that this was *not* assumed in the premises, for the operators in $\mathcal{S}$.

We next consider a proposition which provides a framework for possible further extensions of a basic net.

Proposition 3.10: Let $\mathcal{N}$ be a *regular* net as described in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$. Let $\mathcal{K}$ be the set of all *closed, convex, causally complete* subsets K of M, such that K^c has a *nonempty* interior. For each $K \in \mathcal{K}$ we define the von Neumann algebras $\mathcal{B}^\sharp(K)$ and $\mathcal{A}^\sharp(K^c)$ by

$$\mathcal{B}^\sharp(K) = \cap\{\mathcal{A}(W) | W \in \mathcal{W},\ \overline{W} \supset K\}, \quad \mathcal{A}^\sharp(K^c) = \{\mathcal{A}(W) | W \in \mathcal{W},\ W \subset K^c\}'', \tag{3.20}$$

and hence we have $\mathcal{B}^\sharp(K) = \mathcal{B}(K)$ and $\mathcal{A}^\sharp(K^c) = \mathcal{A}(K^c)$ for all $K \in \mathcal{K}_C$, and $\mathcal{A}^\sharp(W) = \mathcal{A}(W) = \mathcal{B}^\sharp(\overline{W})$ for all $W \in \mathcal{W}$. We refer to the algebras $\mathcal{B}^\sharp(K)$ and $\mathcal{A}^\sharp(K^c)$ as the *associated algebras of the net* $\mathcal{N}$, and we denote this system of algebras by $\mathcal{N}^\sharp$. Then:

(a) The mappings $K \to \mathcal{B}^\sharp(K)$ and $K^c \to \mathcal{A}^\sharp(K^c)$ are Poincaré-covariant and *PCT*-covariant.

(b) The associated algebras satisfy the duality condition

$$\mathcal{B}^\sharp(K)^z = \mathcal{A}^\sharp(K^c)', \tag{3.21}$$

for all $K \in \mathscr{K}$.

(c) If $K_s, K_l \in \mathscr{K}$, and $K_s \subset K_l$, and hence $K_s^c \supset K_l^c$, then

$$\mathscr{B}^\sharp(K_s) \subset \mathscr{B}^\sharp(K_l) = (\mathscr{A}^\sharp(K_l^c)^z)', \quad (\mathscr{B}^\sharp(K_l)^z)' = \mathscr{A}^\sharp(K_l^c) \subset \mathscr{A}^\sharp(K_s^c). \tag{3.22}$$

(d) If $K_1, K_2 \in \mathscr{K}$, and $K_1 \subset K_2^c$, and if either K_1 or K_2 has a *nonempty interior*, then

$$\mathscr{B}^\sharp(K_1)^z \subset \mathscr{B}^\sharp(K_2)', \quad \mathscr{B}^\sharp(K_1) \subset \mathscr{A}^\sharp(K_2^c). \tag{3.23}$$

(e) The algebras $\mathscr{A}^\sharp(K^c)$ are generated by the *local* operators which they contain, and specifically,

$$\mathscr{A}^\sharp(K^c) = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset K^c\}'' = (\mathscr{A}^\sharp(K^c) \cap \mathscr{U})'', \tag{3.24}$$

for any $K \in \mathscr{K}$. Here $\mathscr{U} = \cup\{\mathscr{B}(K) | K \in \mathscr{K}_C\}$ is the algebra of all local operators of the net $\mathscr{N}$.

(f) If the locality relation at left in (3.23) holds for *all* $K_1, K_2 \in \mathscr{K}$ such that $K_1 \subset K_2^c$, then the system $\mathscr{N}^\sharp$ of algebras is an irreducible local net which satisfies all the Conditions I–VIII in Definition 2.1.

Proof: (1) The duality condition in (3.21) is proved by the same kind of reasoning as in step (4) in the proof of Theorem 3.7. The conclusions in parts (a) and (c) then follow trivially from the defining relations in (3.20).

(2) We consider part (d). The proof of the *particular* locality condition at left in (3.23) is based on part (a) of Lemma 3.5. Suppose that K_1 has a nonempty interior. The interior K_1^o is convex since K_1 is convex, and we have $\overline{K_1^o} = K_1$ and $(K_1^o)^c \supset K_2$. By Lemma 3.5 there then exists a wedge $W \in \mathscr{W}$ such that $K_1^o \subset W$ and $K_2 \subset W^c$. Hence $K_1 \subset \overline{W}$ and $\mathscr{B}^\sharp(K_1) \subset \mathscr{A}(W)$, $\mathscr{B}^\sharp(K_2) \subset \mathscr{A}(\overline{W^c})$, from which the relation at left in (3.23) follows by the duality relation $\mathscr{A}(W)^z = \mathscr{A}(\overline{W^c})'$ for the wedge-algebras. The relation at right in (3.23) is equivalent with the relation at left, in view of the duality relation (3.21).

(3) The relations in (3.24) follow readily from the defining relation at right in (3.20), and from the relation $\mathscr{A}(W) = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset W\}''$ which holds for the regular net $\mathscr{N}$.

(4) We consider part (f). If $K_1, K_2 \in \mathscr{K}$, then K_1^c and K_2^c *cannot* be spacelike separated unless *both* K_1^c and K_2^c are wedges in $\mathscr{W}$. This means that the locality relations in (3.22) and (3.23) include all applicable locality conditions for the system $\mathscr{N}^\sharp$, and with the premise in part (f) the system $\mathscr{N}^\sharp$ thus satisfies the condition of locality. Since it also satisfies the condition of duality it follows, by Lemma 2.2, that it satisfies the condition of isotony, and hence $\mathscr{N}^\sharp$ is a local net of the asserted kind. This completes the proof.

We are *not* suggesting that the system $\mathscr{N}^\sharp$ is a local net in general. We have no proof, and no counterexample. The geometric basis for the above construction is described in Proposition 3.2. It is worth noting that the causal complement K^c of a set $K \in \mathscr{K}$ is not necessarily open, but it is ''close'' to its interior in the sense that $((K^c)^o)^c = K$ and $(\overline{K^c})^o = (K^c)^o$. The algebras of the system $\mathscr{N}^\sharp$ are interesting objects to consider, and the proposition which follows supports this view.

Proposition 3.11: Let $\mathscr{N}$ be a *regular net*, with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$ for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$. Let $\mathscr{N}^\sharp$ be the system of *associated algebras*, as described in Proposition 3.10. Let $\mathscr{K}_X$ be a Poincaré- and *PCT*-invariant subset of $\mathscr{K}$ which contains $\mathscr{K}_C$.

(a) Suppose that there exists a local net $\mathscr{N}_X$, with algebras $\mathscr{A}_X(W)$, $\mathscr{B}_X(K)$, and $\mathscr{A}_X(K^c)$ for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_X$, which satisfies *all* the Conditions I–IX in Definition 2.1. Suppose furthermore that the net $\mathscr{N}_X$ is an extension of the net $\mathscr{N}$ in the sense that $\mathscr{A}_X(W) = \mathscr{A}(W)$, $\mathscr{B}_X(K) = \mathscr{B}(K)$, and $\mathscr{A}_X(K^c) = \mathscr{A}(K^c)$ for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$. We then have $\mathscr{B}_X(K) = \mathscr{B}^\sharp(K)$ and $\mathscr{A}_X(K^c) = \mathscr{A}^\sharp(K^c)$ for all $K \in \mathscr{K}_X$, and an extension with the domain of $\mathscr{N}_X$ is thus *unique*.

(b) An extension of the kind described in (a) exists if and only if

$$\mathscr{B}^\sharp(K) = \{\mathscr{B}(K_c) | K_c \in \mathscr{K}_C,\ K_c \subset K\}'', \tag{3.25}$$

for all $K \in \mathscr{K}_X$.

Proof: (1) We consider part (a). Let $K \in \mathscr{K}_X$. By (3.24) we have $\mathscr{A}^\sharp(K^c) = \{\mathscr{B}(K_i) | K_i \in \mathscr{K}_C,\ K_i \subset K^c\}''$. Since the net satisfies Condition IX in Definition 2.1, we have $\mathscr{A}_X(K^c) = \{\mathscr{B}(K_i) | K_i \in \mathscr{K}_C,\ K_i \subset K^c\}''$, and hence $\mathscr{A}_X(K^c) = \mathscr{A}^\sharp(K^c)$ for all $K \in \mathscr{K}_X$. Since both the net $\mathscr{N}_X$ and the system $\mathscr{N}^\sharp$ satisfies the condition of duality, it follows that also $\mathscr{B}_X(K) = \mathscr{B}^\sharp(K)$.

(2) We consider part (b). That the condition in (3.25) is necessary follows from the assumption that $\mathscr{N}_X$ satisfies Condition IX in Definition 2.1. To prove the sufficiency we assume that this condition holds. Let $K_1, K_2 \in \mathscr{K}_X$, and $K_1 \subset K_2^c$. Let $K_{1c}, K_{2c} \in \mathscr{K}_C$, and such that $K_{1c} \subset K_1$ and $K_{2c} \subset K_2$. We then have $K_{1c} \subset K_{2c}^c$, and hence $\mathscr{B}(K_{1c})^z \subset \mathscr{B}(K_{2c})'$. With reference to (3.25) we then conclude that $\mathscr{B}^\sharp(K_1)^z \subset \mathscr{B}^\sharp(K_2)'$, which means that the system $\mathscr{N}^\sharp$ restricted to the domain of $\mathscr{N}_X$ satisfies *all* the applicable conditions of locality. Since the restricted system satisfies the condition of duality, it follows by Lemma 2.2 that the restricted system also satisfies all applicable conditions of isotony, and hence the restricted system is a local net $\mathscr{N}_X$, which satisfies the premises in part (a). This completes the proof.

The existence of an extension of the kind described above is thus implicit in the structure of the net $\mathscr{N}$, or of the basic net $\mathscr{N}_o$ of which $\mathscr{N}$ is the extension, and if it exists, it is unique. It is interesting to consider the possibility that such an extension exists for $\mathscr{K}_X$ equal to the set $\mathscr{K}_E$ in Definition 1.1. We recall that $\mathscr{K}_E$ is the set of all elements K of $\mathscr{K}$ which are contained in the intersection of some pair of wedges in $\overline{\mathscr{W}}$ which are in a general relative position. In particular $\mathscr{K}_E$ contains all sets of the form $\overline{W_1 \cap W_2}$ where W_1 and W_2 are in a general relative position. The intersection of two closed causal wedges which are *not* in a general relative position is in $\mathscr{K}$, but *not* in $\mathscr{K}_E$. The issue is then whether the locality condition at left in (3.23) holds for *all* $K_1, K_2 \in \mathscr{K}_E$. We easily see, on the basis of part (b) in Lemma 3.5, that this condition holds, with the possible exception of the case when *both* K_1 and K_2 are *unbounded* subsets of the *same* characteristic plane. This is a rather singular exceptional case, but we do not know whether this obstacle is real or illusory. For a net of algebras generated by a massive free Hermitian scalar field the set of all $\mathscr{B}$-algebras (with domain $\mathscr{K}_E$) is indeed local, and we have no example of a possible failure of this locality condition.

If an element $K \in \mathscr{K}$ has an empty interior it is always a subset of a characteristic plane. As we show in the next section, the von Neumann algebra $\mathscr{B}^\sharp(K)$ associated with such a set is not necessarily trivial, and these sets K can therefore not be dismissed outright as ''unphysical,'' although they are, in a sense, mildly pathological. We can avoid the pathology by excluding the algebras $\mathscr{B}^\sharp(K)$ and $\mathscr{A}^\sharp(K^c)$ for every $K \in \mathscr{K}$ which has an empty interior, and we then obtain a unique local net $\mathscr{N}_I$ from a regular net $\mathscr{N}$. We describe this in the theorem below. The proof follows trivially from Proposition 3.10, and we feel that we can omit the details.

Theorem 3.12: Let the premises and notation be as in Proposition 3.10. Let $\mathscr{K}_I$ be the set of all $K \in \mathscr{K}$ which have a *nonempty interior*, i.e., $\mathscr{K}_I$ is the set of all *closed convex* subsets K of M such that both K and K^c have *nonempty interiors*. Then the system $\mathscr{N}_I$ of von Neumann algebras $\{\mathscr{B}_I(K), \mathscr{A}_I(K^c) | K \in \mathscr{K}_I\}$, defined by $\mathscr{B}_I(K) = \mathscr{B}^\sharp(K)$ and $\mathscr{A}_I(K) = \mathscr{A}^\sharp(K^c)$ for all $K \in \mathscr{K}_I$, is a local net which satisfies *all* the Conditions I–VIII in Definition 2.1. Furthermore the net $\mathscr{N}_I$ satisfies the condition

$$\mathscr{A}_I(K^c) = \{\mathscr{B}_I(K_i) | K_i \in \mathscr{K}_I,\ K_i \subset K^c\}'' = \{\mathscr{B}_I(D) | D \in \mathscr{D},\ D \subset K^c\}'', \tag{3.26}$$

for all $K \in \mathscr{K}_I$. The net $\mathscr{N}_I$ is the *unique* net with the above domain which satisfies the condition in (3.26) and the condition that $\mathscr{B}_I(D) = \mathscr{B}(D)$ for all $D \in \mathscr{D}$.

We can again regard the $\mathscr{B}_I$-algebras as the primary objects of interest, and the $\mathscr{A}_I$-algebras as ''auxiliary'' constructs. It is then interesting to note that the local net with domain $\mathscr{K}_I$ consisting of the algebras $\mathscr{B}_I(K)$ satisfies a condition analogous to *Haag duality*, in view of (3.26) and the duality condition $\mathscr{A}_I(K^c) = (\mathscr{B}_I(K)^z)'$. Note also that we are *not* asserting that the net $\mathscr{N}_I$ satisfies Condition IX in Definition 2.1. The relation in (3.26) is *part* of this condition, but it might happen that the algebra $\mathscr{B}_I(K)$ for some unbounded $K \in \mathscr{K}_I$ is *not* generated by the local operators it contains.

IV. SOME PROPERTIES OF REGULAR NETS

In this section we consider some properties of a *regular* net $\mathscr{N}$, as described in Definition 3.1. Some of the properties of the algebras of such a net are shared by the associated algebras $\mathscr{B}^\sharp(K)$

and $\mathcal{A}^{\sharp}(K^c)$ of the net defined in Proposition 3.10, when K is an element of the set $\mathcal{K}_E$ in Definition 1.1, and we felt it worthwhile to state such results when applicable. The reader is warned, however, that these result might *not* be valid for an arbitrary K in the larger set $\mathcal{K}$. In many respects the elements of $\mathcal{K}_E$ are ''better behaved'' than arbitrary elements in $\mathcal{K}$.

As we noted in Section II, the algebras $\mathcal{B}_o(D)$ of a *basic* net are continuous from the outside, whereas the algebras $\mathcal{A}_o(D^c)$ are continuous from the inside, i.e., we have the relations in (2.7). A *regular* net has similar features, and one aspect of this can be stated as follows.

Proposition 4.1: Let $\mathcal{N}$ be a *regular* net, i.e., an irreducible local net as described in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{A}(K^c)$, and $\mathcal{B}(K)$ defined for all $W \in \mathcal{W}$ and all $K \in \mathcal{K}_C$. Let the associated algebras $\mathcal{A}^{\sharp}(K^c)$ and $\mathcal{B}^{\sharp}(K)$ be defined by

$$\mathcal{B}^{\sharp}(K) = \cap\{\mathcal{A}(W) \mid W \in \mathcal{W},\ \overline{W} \supset K\}, \quad \mathcal{A}^{\sharp}(K^c) = \{\mathcal{A}(W) \mid W \in \mathcal{W},\ W \subset K^c\}'', \tag{4.1}$$

for all $K \in \mathcal{K}$, as in Proposition 3.10. Hence $\mathcal{B}^{\sharp}(K) = \mathcal{B}(K)$ and $\mathcal{A}^{\sharp}(K^c) = \mathcal{A}(K^c)$ for all $K \in \mathcal{K}_C$.

Let $K \in \mathcal{K}_E$, and let K_n, $n \in \mathbb{Z}_+$, be a nested sequence of sets $K_n \in \mathcal{K}_E$ such that $K_n \supset K_{n+1}$ for all positive n. Then

$$K = \cap\{K_n \mid n \in \mathbb{Z}_+\}, \tag{4.2}$$

if and only if

$$K^c = \cup\{K_n^c \mid n \in \mathbb{Z}_+\}, \tag{4.3}$$

and if this is the case, then

$$\mathcal{B}^{\sharp}(K) = \cap\{\mathcal{B}^{\sharp}(K_n) \mid n \in \mathbb{Z}_+\}, \quad \mathcal{A}^{\sharp}(K^c) = \{\mathcal{A}^{\sharp}(K_n^c) \mid n \in \mathbb{Z}_+\}''. \tag{4.4}$$

Proof: (1) It is trivial that the relation in (4.3) implies the relation in (4.2). To prove the converse we assume that the equality in (4.2) obtains. We first note that since $K_1 \in \mathcal{K}_E$, it follows from part (d) of Proposition 3.3 that there exist two *spacelike* 3-planes σ' and σ'' such that K_1 lies *below* σ' and *above* σ''. It follows readily from this that the set $\overline{V}(x) \cap K_1$ is *compact* for any $x \in \mathsf{M}$. Suppose now that $x \in K^c$, in which case we have $\overline{V}(x) \cap K = \varnothing$. By the assumed nesting, and by (4.2), we then have $\overline{V}(x) \cap K_1 \supset \overline{V}(x) \cap K_n \supset \overline{V}(x) \cap K_{n+1}$ and $\cap\{\overline{V}(x) \cap K_n \mid n \in \mathbb{Z}_+\} = \overline{V}(x) \cap K = \varnothing$. We cannot have $\overline{V}(x) \cap K_n \neq \varnothing$ for all n, because we would then have $\overline{V}(x) \cap K \neq \varnothing$, since $\overline{V}(x) \cap K_1$ is compact. Hence $\overline{V}(x) \cap K_m = \varnothing$ for some m, which means that $x \in K_m^c \subset \cup\{K_n^c \mid n \in \mathbb{Z}_+\}$. We have thus shown that $K^c \subset \cup\{K_n^c \mid n \in \mathbb{Z}_+\}$. Since it follows trivially from (4.2) that $K^c \supset \cup\{K_n^c \mid n \in \mathbb{Z}_+\}$ we then have the equality in (4.3).

(2) We now assume that the relations in (4.2) and (4.3) hold. Let $D \in \mathcal{D}$ be such that $D \subset K^c = \cup\{K_n^c \mid n \in \mathbb{Z}_+\}$. Since D is compact it follows that $D \subset K_m^c$ for some m, and hence $\mathcal{B}(D) \subset \mathcal{A}^{\sharp}(K_m^c) \subset \{\mathcal{A}^{\sharp}(K_n^c) \mid n \in \mathbb{Z}_+\}''$. By Proposition 3.10 we then have $\mathcal{A}^{\sharp}(K^c) = \{\mathcal{B}(D) \mid D \in \mathcal{D},\ D \subset K^c\}'' \subset \{\mathcal{A}^{\sharp}(K_n^c) \mid n \in \mathbb{Z}_+\}''$. We trivially have $\mathcal{A}^{\sharp}(K^c) \supset \{\mathcal{A}^{\sharp}(K_n^c) \mid n \in \mathbb{Z}_+\}''$ since $K^c \supset K_n^c$ for all n, and hence we obtain the equality at right in (4.4). The equality at left then follows by duality. This completes the proof.

Within free-field theories one can give examples (as we will do below) of local nets for which the algebras $\mathcal{B}(K)$ are *nontrivial* for certain sets $K \in \mathcal{K}_C$ with an empty interior, and for such a net the $\mathcal{B}$-algebras can obviously not be continuous from the *inside*. Within free-field theories the algebras $\mathcal{B}(K)$ *are* continuous from the inside when K has a nonempty interior, but it is not known whether this might be true for local nets in general.

We will next show, among other things, that if $K \in \mathcal{K}_C$, with K contained in a *two*-dimensional linear manifold, then the algebra $\mathcal{B}(K)$ is trivial.

Proposition 4.2: Let $\mathcal{N}$ be a *regular* net, with algebras $\mathcal{A}(W)$, $\mathcal{A}(K^c)$, and $\mathcal{B}(K)$ defined for all $W \in \mathcal{W}$ and all $K \in \mathcal{K}_C$. Then:

(a) For every $W \in \mathcal{W}$ we have

$$\mathcal{A}(W) \cap \mathcal{A}(W)' = \mathcal{A}(W) \cap \mathcal{A}(\overline{W^c})^z = \mathbb{C}^1 \cdot I, \quad \mathcal{A}(W) \cap \mathcal{A}(\overline{W^c}) = \mathbb{C}^1 \cdot I, \tag{4.5}$$

and $\mathscr{A}(W)$ is thus a factor.

(b) If $K \in \mathscr{K}_C$ is a subset of a *two*-dimensional linear submanifold of M, then $\mathscr{B}(K) = \mathsf{C}^1 \cdot I$, and hence $\mathscr{A}(K^c) = \mathsf{B}(\mathscr{H})$.

In particular we have $\mathscr{B}(\{x\}) = \mathsf{C}^1 \cdot I$ for any singleton $\{x\}$, and $\mathscr{B}([x', x'']) = \mathsf{C}^1 \cdot I$ for any isotropic or spacelike (finite) line segment.

(c) If $K \in \mathscr{K}_C$ is a subset of a characteristic plane in M, then it can happen that $\mathscr{B}(K)$ contains elements which are *not* multiples of the identity (provided that K is not a subset of a two-plane). The vector Ω is *not* cyclic for $\mathscr{B}(K)$, and hence *not* separating for $\mathscr{A}(K^c)$.

Proof: (1) We consider part (a). Let $W \in \mathscr{W}$. With a suitable choice of a standard coordinate system we can write $W = W_R = \{x | x \in \mathsf{M},\ x_3 > |x_4|\}$. Let u be the isotropic vector $u = (0,0,1,1)$. Let $B \in \mathscr{A}(W) \cap \mathscr{A}(W)'$. Since $W(su) = \{su + x | x \in W\} \subset W$ for all $s \geqslant 0$ we then have $T(su)BT(su)^{-1} \in \mathscr{A}(W)$, and hence $[B^*, T(su)BT(su)^{-1}] = 0$ for all $s \geqslant 0$. Similarly we have $[B, T(su)B^*T(su)^{-1}] = 0$ for all $s \geqslant 0$, and it follows that $[B^*, T(su)BT(su)^{-1}] = 0$ for *all* real s. By Lemma 2.6 we then have $\langle \Omega | B^*B\Omega \rangle = |\langle \Omega | B\Omega \rangle|^2$, which implies that $(B - cI)\Omega = 0$, for some complex c, and since Ω is separating for $\mathscr{A}(W)$ it follows that $B = cI$. This proves the identities at left in (4.5).

Let B be a *boson* operator in $\mathscr{A}(W) \cap \mathscr{A}(\overline{W^c})$. Since $\mathscr{A}(\overline{W^c})$ and $\mathscr{A}(W)' = \mathscr{A}(\overline{W^c})^z$ have precisely the same boson operators, it follows from the above that B is a multiple of the identity.

Suppose that A is a *fermion* operator in $\mathscr{A}(W) \cap \mathscr{A}(\overline{W^c})$. We then have $A^*A + AA^* = 0$ by locality, and hence $A = 0$. We have thus proved the relations in (4.5).

(2) We consider part (b), and assume that $K \in \mathscr{K}_C$ is a subset of a *two*-dimensional linear manifold ω, which is then necessarily a submanifold of a characteristic plane π, since K cannot contain two timelike separated points. If ω is *spacelike*, then ω is the common edge of two wedges, $\omega = \overline{W} \cap W^c$. Since $K \subset \overline{W} \cap W^c$ we then have $\mathscr{B}(K) \subset \mathscr{A}(W) \cap \mathscr{A}(\overline{W^c}) = \mathsf{C}^1 \cdot I$ by the result in part (a).

We now consider the case when $\omega \subset \pi$ is *not* spacelike. With a suitable choice of a standard coordinate system we can write $\pi = \{x | x \in \mathsf{M},\ x_4 = x_3\}$ and $\omega = \{x | x \in \mathsf{M};\ x_4 = x_3,\ x_1 = 0\}$. Let Q be a self-adjoint operator in $\mathscr{B}(K)$. Let $K(x)$ be obtained from K by a translation by x, i.e., $K(x) = \{x + y | y \in K\}$. Let $u = (0,0,1,1)$ and $v = (1,0,0,0)$. Then $K \subset K(rv + su)^c$, and hence $[Q^z, T(rv + su)QT(rv + su)^{-1}] = 0$, for any s and any $r \neq 0$. By continuity we then have $[Q^z, T(su)QT(su)^{-1}] = 0$ for all s. By Lemma 2.6 it follows that $\langle \Omega | Q^z Q\Omega \rangle = \langle \Omega | Q^z \Omega \rangle \times \langle \Omega | Q\Omega \rangle$. If Q is a fermion operator we conclude from this that $Q = 0$, and if Q is a boson operator we conclude that Q is a multiple of the identity, and hence $\mathscr{B}(K) = \mathsf{C}^1 \cdot I$, as asserted.

(3) We consider part (c), and assume that $K \in \mathscr{K}_C$ is a subset of a characteristic plane π. To show that $\mathscr{B}(K)$ is not necessarily trivial, we depend on a result in TW-II:[7] see Example 5.20 in Section V. The example in question concerns the theory of a single Hermitian *free* scalar field $\varphi(x)$. We consider the subset $K_e = \{x | \tau \geqslant x_3 = x_4 \geqslant -\tau,\ x_1^2 + x_2^2 \leqslant \rho^2\}$ of the characteristic plane $\pi = \{x | x_3 = x_4\}$, with $\tau > 0$, $\rho > 0$. We then have $K_e \in \mathscr{K}_C$ and K_e has an empty interior. It was shown in the quoted paper that there exists a unitary operator B, not equal to a multiple of the identity, which commutes with $\varphi(x)$ for all $x \in K_e^c$. For a *free*-field theory this implies that $B \in \mathscr{B}(K_e)$.

(4) We continue the discussion of part (c), and assume that $K \in \mathscr{K}_C$ is a subset of the characteristic plane $\pi = \{x | x_3 = x_4\}$. Let $u = (0,0,1,1)$ and $v = (1,0,0,0)$, and let $K(x) = \{x + y | y \in K\}$ for any $x \in \mathsf{M}$. For any $x, y \in \pi$ we have $-(x - y) \cdot (x - y) = (x_1 - y_1)^2 + (x_2 - y_2)^2$. Since $K \subset \pi$ is compact it follows that for a sufficiently large λ the displaced set $K(\lambda v + su)$ is spacelike separated from K for *all* s. We chose such a λ, fixed in what follows. Suppose that $Q \in \mathscr{B}(K)$ is not a multiple of the identity. Let $B = T(\lambda v)QT(\lambda v)^{-1} - \langle \Omega | Q\Omega \rangle \cdot I$. For any $A \in \mathscr{B}(K)$ we then have $[A^*, T(su)B^zT(su)^{-1}] = 0$ for all s, and hence $\langle A\Omega | B^z\Omega \rangle = 0$, by Lemma 2.6. Since Ω is separating for $\mathscr{B}(K(\lambda v))$ we must have $B^z\Omega \neq 0$, and we conclude that Ω cannot be cyclic for $\mathscr{B}(K)$. This completes the proof.

That $\mathscr{A}(W)$ is a factor is a well-known old result. Part (b) of the proposition extends a result by Driessler and Summers.[21] The result in part (c) is intuitively reasonable. If $K \in \mathscr{K}_C$ has an empty interior, and hence is a subset of a characteristic plane, then $\mathscr{B}(K)$ is not necessarily trivial, but such an algebra is indeed ''smaller'' than an algebra associated with any set with a nonempty interior, in the sense that $\mathscr{B}(K)\Omega$ is not dense in $\mathscr{H}$.

We will now consider ''intersection properties'' of a *regular* net, and its associated algebras. The issue is the nature of the intersection of two or several algebras in a (regular) net $\mathcal{N}$, and how it might correspond to geometric relationships. We first consider the very simple case of two parallel wedges. We believe the result is widely known, but we present it here anyway for the sake of completeness.

Proposition 4.3: Let $W_1, W_2 \in \mathscr{W}$ be such that W_2 is obtained from W_1 by a *parallel displacement*. Let $\mathcal{A}(W)$ denote the wedge-algebras of a *regular* net. Then $(W_1 \cup W_2)^{cc} = W_u \in \mathscr{W}$, $W_1^c \cap W_2^c = W_u^c$, and

$$\mathcal{A}(\overline{W}_1^c) \cap \mathcal{A}(\overline{W}_2^c) = \mathcal{A}(\overline{W}_1^c \cap \overline{W}_2^c) = \mathcal{A}(\overline{W}_u^c), \tag{4.6}$$

$$(\mathcal{A}(W_1) \cup \mathcal{A}(W_2))'' = \mathcal{A}((W_1 \cup W_2)^{cc}) = \mathcal{A}(W_u). \tag{4.7}$$

Proof: It is trivial that $(W_1 \cup W_2)^{cc} = W_u \in \mathscr{W}$. Specifically the lower plane of W_u is the lower one of the lower planes of W_1 and W_2, and the upper plane of W_u is the upper one of the upper planes of W_1 and W_2. The relations (4.6) and (4.7) are equivalent under duality, and they are trivial if one of the wedges is contained in the other. We thus consider the case when W_2 is obtained from W_1 by a *timelike* displacement. With a suitable choice of a standard coordinate system we then have the configuration

$$W_1 = \{x | x_3 - \tau > |x_4 + \tau|\}, \quad W_2 = \{x | x_3 - \tau > |x_4 - \tau|\}, \quad W_u = \{x | x_3 > |x_4|\},$$

for some $\tau > 0$.

Let $B \in (\mathcal{A}(W_1) \cup \mathcal{A}(W_2))'$, and let $A \in \mathcal{A}(W_u)$. We write $A(x) = T(x) A T(x)^{-1}$ for the translate of A by any $x \in \mathsf{M}$. We then have $A(x) \in \mathcal{A}(W_1) \cup \mathcal{A}(W_2)$, and hence $[B, A(x)] = 0$, for any $x \in W_1 \cup W_2$, and in particular when x is in the region $\{x | x \in \mathsf{M}, x_3 > |x_4|, x_3 > 2\tau\}$. It then follows readily from the results in a paper by Araki[22] that $[B, A(x)] = 0$ for all $x \in W_u$. By the continuity of $A(x)$ we have $[A, B] = 0$. This implies that $B \in \mathcal{A}(W_u)'$, and we thus have $\mathcal{A}(W_u)' \supset (\mathcal{A}(W_1) \cup \mathcal{A}(W_2))'$, and $\mathcal{A}(W_u) \subset (\mathcal{A}(W_1) \cup \mathcal{A}(W_2))''$. By isotony we have $\mathcal{A}(W_u) \supset (\mathcal{A}(W_1) \cup \mathcal{A}(W_2))''$, and hence the relation (4.7) holds. This completes the proof.

We next discuss what amounts to an elaboration of a result by Borchers.[23]

Proposition 4.4: Let $\mathcal{N}$ be a *regular* net, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$. Let $K_1, K_2 \in \mathscr{K}_C$ be such that $K_1 \subset K_2^o$. Then:

(a) $\mathcal{B}(K_1) \cap \mathcal{B}(K_2)' = \mathsf{C}^1 \cdot I$, and hence $(\mathcal{B}(K_2) \cup \mathcal{B}(K_1)')'' = \mathsf{B}(\mathscr{H})$;

(b) $\mathcal{B}(K_1) \cap \mathcal{A}(K_2^c) = \mathsf{C}^1 \cdot I$, and hence $(\mathcal{B}(K_2) \cup \mathcal{A}(K_1^c))'' = \mathsf{B}(\mathscr{H})$.

Proof: (1) We consider part (a). The salient point is that K_1 is ''movable'' within K_2, since K_1 is compact and contained in the interior of K_2. Let u be a forward timelike vector. In view of the above there exists a $\tau_0 > 0$ such that $T(tu) \mathcal{B}(K_1) T(tu)^{-1} \subset \mathcal{B}(K_2)$ whenever $|t| < \tau_0$.

Let E be a self-adjoint projection in $\mathcal{B}(K_1) \cap \mathcal{B}(K_2)'$, and let $E^\perp = I - E$. We then have $E E^\perp = 0$, and $[T(tu) E T(tu)^{-1}, E^\perp] = 0$ whenever $|t| < \tau_0$. By Borchers' theorem this implies that $T(tu) E T(tu)^{-1} E^\perp = 0$ for *all* t. This result is a consequence of the spectrum condition on the one-parameter translation group $\{T(tu) | t \in \mathsf{R}^1\}$. From this we readily conclude that E is a multiple of the identity, which implies the conclusion in part (a).

(2) We consider part (b). The algebras $\mathcal{B}(K_1)$ and $\mathcal{A}(K_2^c)$ cannot have any nonzero fermion element Q in common, since Q would then anticommute with its adjoint. The boson elements of $\mathcal{A}(K_2^c)$ and $\mathcal{B}(K_2)' = \mathcal{A}(K_2^c)^z$ are the same, and by the result in (a), the only common boson elements of $\mathcal{B}(K_1)$ and $\mathcal{A}(K_2^c)$ are the multiples of the identity. This completes the proof.

In the further study of the intersection properties we depend on a number of results from TW-I[6] and TW-II.[7] We first consider the important notion of the *envelope of uniqueness* $\mathscr{E}(S)$ of any open subset S of M.

Lemma 4.5: Let S be an *open* subset of Minkowski spacetime M. There then exists a unique *smallest* subset X of M which satisfies the following three conditions: (I) $X \supset S$. (II) If X contains the base B_0 of an open double cone $D^o \subset \mathsf{M}$, then X contains D^o. (III) If X contains the open timelike line segment (x', x''), then X contains the open double cone $D^o = D(x', x'')^o$ with x' and x'' as apices. Here, and throughout this paper, we denote this smallest set by $\mathscr{E}(S)$, for any open $S \subset \mathsf{M}$.

See Section IV in TW-I[6] for the (trivial) proof, and for a detailed discussion of the mapping $S \to \mathscr{E}(S)$. We will not repeat this discussion here, but we note a few simple facts. The set $\mathscr{E}(S)$ is (obviously) *open*, and it satisfies the conditions $\mathscr{E}(S) \subset (S^{cc})^o \subset S^{cc} = \mathscr{E}(S)^{cc}$. If S is *causally complete*, then $\mathscr{E}(S) = S$. In general $\mathscr{E}(S)$ is smaller that S^{cc}, but for certain kinds of (physically interesting) regions S we do have $\mathscr{E}(S) = S^{cc}$, as shown in Lemma 4.6 below. The above definition of $\mathscr{E}(S)$ refers to *open* double cones, but it is obvious that if S contains either the base B of a closed double cone $D = D(x', x'') \in \mathscr{D}$, or else the closed timelike line segment $[x', x'']$, then $\mathscr{E}(S)$ contains D. The envelope $\mathscr{E}(S)$ is of interest in the theory of the Klein–Gordon equation, in view of the following well-known principle. If $\phi(x)$ is a solution of this equation which vanishes throughout S, then $\phi(x)$ vanishes throughout $\mathscr{E}(S)$. In this sense $\mathscr{E}(S)$ is an ''envelope of uniqueness.''

We state here some properties of the mapping $\mathscr{E}$ for later reference.

Lemma 4.6: (a) Let $W_1, W_2 \in \mathscr{W}$ be such that $W_1 \cap W_2 \neq \emptyset$. Then $\mathscr{E}(W_1 \cup W_2) = (W_1 \cup W_2)^{cc}$.

(b) Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$, and let $S = \{W | W \in \mathscr{F}\}$. Suppose that the normals of the boundary planes of these wedges spans M, and that $\mathscr{F}$ contains at least two wedges for which no two of the four boundary planes are parallel. Then $\mathscr{E}(S) = S^{cc}$. In particular this holds if $\mathscr{F}$ contains at least two wedges in a general relative position, i.e., if the normals of the four boundary planes are linearly independent.

(c) Let $S \subset \mathsf{M}$ be *open*, and let S_n, $n \in \mathsf{Z}_+$, be a sequence of *open* subsets of S such that $S_{n+1} \supset S_n$ for all n, and $S = \cup\{S_n | \ n \in \mathsf{Z}_+\}$. Then $\mathscr{E}(S) = \cup\{\mathscr{E}(S_n) | \ n \in \mathsf{Z}_+\}$.

For the proof of part (a), see Proposition 5.6; for the proof of part (b), see Theorem 5.7, and for the proof of part (c), see Proposition 4.12 in TW-I.[6]

We next state an important lemma which is based on some results in TW-II,[7] and ultimately on the theory of Araki referred to earlier.[22]

Lemma 4.7: Let $\{S_\alpha | \alpha \in \mathscr{I}\}$, where $\mathscr{I}$ is some index set, be a family of open regions $S_\alpha \subset \mathsf{M}$ such that S_α is bounded above by the characteristic plane π''_α and below by the characteristic plane π'_α. Let $S = \cup\{S_\alpha | \alpha \in \mathscr{I}\}$. We denote by $\mathscr{E}(S)$ the envelope of uniqueness of S, as in Lemma 4.5.

Let $\mathscr{N}$ be a *regular* net as in Definition 3.1. Let D be a double cone in $\mathscr{D}$. We denote by $D(x)$ the image of D under a translation by x. Then

$$\{\mathscr{B}(D(x)) | x \in S\}'' = \{\mathscr{B}(D(x)) | x \in \mathscr{E}(S)\}''. \tag{4.8}$$

In particular the identity in (4.8) obtains if S is the union of any family of wedges in $\mathscr{W}$.

Proof: Since $S \subset \mathscr{E}(S)$, it is trivial that the left member in (4.8) is included in the right member. We thus have to prove the opposite inclusion.

Let $A \in \{\mathscr{B}(D(x)) | x \in S\}'$ and let $B \in \mathscr{B}(D)$. We then have $[A, T(x)BT(x)^{-1}] = 0$ for all $x \in S$, and we have to prove that the commutator also vanishes for all $x \in \mathscr{E}(S)$. Let Ψ' and Ψ'' be two vectors in $\mathscr{H}$ with compact support in four-momentum space, and let $\xi(x)$ be the function on M defined by $\xi(x) = \langle \Psi' | [A, T(x)BT(x)^{-1}]\Psi'' \rangle$. This function is continuous and bounded, and it can be regarded as a tempered distribution in $\mathscr{S}'(\mathsf{M})$. We easily see that the support of the Fourier transforms $\hat{\xi}(p)$ of this distribution is contained in a region of the form $\overline{V}_+(p') \cup \overline{V}_-(p'')$, for some points p' and p'' in four-momentum space. The distribution $\xi(x)$ is in a special class of distributions studied by Araki,[22] denoted by $\mathscr{S}'_A(\mathsf{M})$ in TW-II.[7] We have $\xi(x) = 0$ for all $x \in S$, and it then follows by Theorem 2.11 in TW-II[7] that $\xi(x) = 0$ for all $x \in \mathscr{E}(S)$. Since this holds for all vectors Ψ' and Ψ'' of compact support in four-momentum space, it follows that $[A, T(x)BT(x)^{-1}] = 0$ for all $x \in \mathscr{E}(S)$. This completes the proof.

We want to state here explicitly that it is *not* true in general that if S is an open set, and if $\xi(x)$ is a distribution in the class $\mathscr{S}'_A(\mathsf{M})$ such that $\xi(x) = 0$ on S, then $\xi(x) = 0$ also on $\mathscr{E}(S)$. It is easy to find counterexamples, and some are given in TW-II.[7] On the basis of the general assumptions in *this* paper we *cannot* conclude that the relation (4.8) holds for *arbitrary* open sets S, although the relation does hold for a free-field theory. With our approach the special nature of the set S stated in the premises is crucial.

The lemma which follows is an important tool in our investigation. It depends on Lemma 4.7 and on some simple facts concerning the geometry of Minkowski spacetime.

Lemma 4.8: Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$. We write $S=\cup\{W|W\in\mathscr{F}\}$ and $K=\cap\{W^c|W\in\mathscr{F}\}=S^c$.

Let $q\in V_+$. For any wedge $W=W(\pi',\pi'')\in\mathscr{W}$ with lower plane π' and upper plane π'' we define $W[q]=W(\pi'(q),\pi''(-q))$ as the wedge in $\mathscr{W}$ with lower plane $\pi'(q)=\{x+q|x\in\pi'\}$ and upper plane $\pi''(-q)=\{x-q|x\in\pi''\}$. We write $S[q]=\cup\{W[q]|W\in\mathscr{F}\}$ and $K[q]=S[q]^c=\cap\{W[q]^c|W\in\mathscr{F}\}$.

Let $D_q=D(-q,q)\in\mathscr{D}$ be the double cone with apices $-q$ and q, and for any $x\in\mathsf{M}$, let $D_q(x)=\{x+y|y\in D_q\}$ be the double cone obtained by a displacement of D_q by x. Then:

(a) For any $W\in\mathscr{W}$ we have $D_q(x)\subset W$ if and only if $x\in W[q]$.

(b) $D_q(x)\subset S$ if and only if $x\in S[q]$.

(c) If $1>\theta>0$, then $K[q]\supset K[\theta q]\supset K$ and $S[q]\subset S[\theta q]\subset S$. Furthermore we have

$$K=\cap\{K[q/n]|n\in\mathsf{Z}_+\},\quad S=\cup\{S[q/n]|n\in\mathsf{Z}_+\},\tag{4.9}$$

$$\mathscr{E}(S)=\cup\{\mathscr{E}(S[q/n]|n\in\mathsf{Z}_+\}.\tag{4.10}$$

(d) Let $\mathscr{N}$ be a *regular net* with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$, as in Definition 3.1. Then

$$\{\mathscr{B}(D_q(x))|x\in\mathscr{E}(S[q])\}''\subset\{\mathscr{A}(W)|W\in\mathscr{F}\}''.\tag{4.11}$$

Proof: (1) Let $W=W(\pi',\pi'')\in\mathscr{W}$. Since W is causally complete, W contains $D_q(x)$ if and only if W contains the apices $x-q$ and $x+q$ of $D_q(x)$, and hence if and only if $x+q$ lies *below* the plane π'' and $x-q$ lies *above* the plane π'. These two conditions are precisely the conditions for x to be in $W[q]$, and hence we arrive at the conclusion in part (a). From this the conclusion in part (b) follows trivially, in view of the definition of the sets S and $S[q]$.

(2) We consider part (c). The inclusion relations $K[q]\supset K[\theta q]\supset K$ and $S[q]\subset S[\theta q]\subset S$, as well as the relations in (4.9), are trivial if $\mathscr{F}$ consists of a single wedge $W\in\mathscr{W}$, in which case we have $S=W$ and $K=W^c$. That these relations hold *generally*, for an arbitrary $\mathscr{F}$, then follows readily from the definitions of K, $K[q]$, S, and $S[q]$. By part (c) of Lemma 4.6 the relation in (4.10) then follows from the relation at right in (4.9).

(3) We consider part (d). Suppose that $x\in S[q]=\cup\{W[q]|W\in\mathscr{F}\}$. There thus exists a $W\in\mathscr{F}$ such that $x\in W[q]$, and hence $D_q(x)\subset W$ and $\mathscr{B}(D_q(x))\subset\mathscr{A}(W)$. We thus have $\{\mathscr{B}(D_q(x))|x\in S[q]\}''\subset\{\mathscr{A}(W)|W\in\mathscr{F}\}''$, and the relation in (4.11) follows from the identity in (4.8). This completes the proof.

It is worth noting here that the double cones $D_q(x)$ contained in a $W\in\mathscr{W}$ do not in general ''sweep out'' W, i.e., the set $\cup\{D_q(x)|x\in W[q]\}$ is *smaller* than W, unless q lies in the span of the normals u' and u''. This is in contrast to the analogous (completely trivial) scenario in a *two*-dimensional spacetime.

We also note that the open set S in the lemma is not, in general, causally complete. In particular it can happen for some $S\neq\mathsf{M}$, that $K=S^c$ is empty, in which case $S^{cc}=K^c=\mathsf{M}$. We also note here the following somewhat subtle point. Let $\mathscr{F}^*=\{W|W\in\mathscr{W},\ W^c\supset K\}$. We then have $\mathscr{F}^*\supset\mathscr{F}$ and $K=\cap\{W^c|W\in\mathscr{F}^*\}$. Let $K^*[q]=\cap\{W[q]^c|W\in\mathscr{F}^*\}$. Now one may think that $K[q]=K^*[q]$, but this is not necessarily so. The set $K^*[q]$ is in general *smaller* than the set $K[q]$, as one can see by considering examples in which $\mathscr{F}$ consists of just two intersecting wedges.

It was shown by Landau,[24] and it also follows from Proposition 4.4, that the algebras associated with two *spacelike* separated double cones have only multiples of the identity in common. The algebras associated with two regions which are merely disjoint can very well have a nontrivial intersection, but for certain special pairs of regions the intersection *is* trivial, as shown in the theorem which follows.

Theorem 4.9: Let $\mathscr{N}$ be a *regular* net, with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$, defined for all $W\in\mathscr{W}$ and all $K\in\mathscr{K}_C$.

(a) If $K_1,K_2\in\mathscr{K}_C$ are such that $K_1\cap K_2=\varnothing$, then $\mathscr{B}(K_1)\cap\mathscr{B}(K_2)=\mathsf{C}^1\cdot I$, and, equivalently, $(\mathscr{A}(K_1^c)\cup\mathscr{A}(K_2^c))''=\mathsf{B}(\mathscr{H})$.

(b) If $K_1 \in \mathscr{K}_C$ and $W_2 \in \mathscr{W}$ are such that $K_1 \cap \overline{W}_2 = \varnothing$, then $\mathscr{B}(K_1) \cap \mathscr{A}(W_2) = \mathsf{C}^1 \cdot I$, and, equivalently, $(\mathscr{A}(K_1^c) \cup \mathscr{A}(\overline{W}_2^c))'' = \mathsf{B}(\mathscr{H})$.

(c) If $W_1, W_2 \in \mathscr{W}$ are such that $\overline{W}_1 \cap \overline{W}_2 = \varnothing$, then $\mathscr{A}(W_1) \cap \mathscr{A}(W_2) = \mathsf{C}^1 \cdot I$, and, equivalently, $(\mathscr{A}(\overline{W}_1^c) \cup \mathscr{A}(\overline{W}_2^c))'' = \mathsf{B}(\mathscr{H})$.

Proof: (1) In this proof we follow the notation in Lemma 4.8. We consider part (a). Let $\mathscr{F}_1 = \{W | W \in \mathscr{W},\ W^c \supset K_1\}$, $\mathscr{F}_2 = \{W | W \in \mathscr{W},\ W^c \supset K_2\}$, and $\mathscr{F} = \mathscr{F}_1 \cup \mathscr{F}_2$. We write $K_1[q] = \cap\{W[q]^c | W \in \mathscr{F}_1\}$ and $K_2[q] = \cap\{W[q]^c | W \in \mathscr{F}_2\}$ for any $q \in V_+$. We easily see that since K_1 is compact there exists a $q \in V_+$ such that $K_1[q] \cap K_2[q] = \varnothing$. Let q be so chosen. We write $S[q] = \cup\{W[q] | W \in \mathscr{F}\}$, and we then have $S[q]^c = K_1[q] \cap K_2[q] = \varnothing$ and $S[q]^{cc} = \mathsf{M}$. Since $\mathscr{F}$ obviously contains two wedges in a general relative position it follows from part (b) of Lemma 4.6 that $\mathscr{E}(S[q]) = S[q]^{cc} = \mathsf{M}$.

By (4.11) we then have $\{\mathscr{B}(D_q(x)) | x \in \mathsf{M}\}'' \subset \{\mathscr{A}(W) | W \in \mathscr{F}\}'' = (\mathscr{A}(K_1^c) \cup \mathscr{A}(K_2^c))''$. By Lemma 2.6 the left member of this inclusion relation equals $\mathsf{B}(\mathscr{H})$, and hence the relation $(\mathscr{A}(K_1^c) \cup \mathscr{A}(K_2^c))'' = \mathsf{B}(\mathscr{H})$ follows. The relation $\mathscr{B}(K_1) \cap \mathscr{B}(K_2) = \mathsf{C}^1 \cdot I$ then follows by duality.

(2) We consider part (b), and we define $\mathscr{F}_1$ and $\mathscr{F}$ as above, with $\mathscr{F}_2 = \{\overline{W}_2^c\}$. Since K_1 is compact we can obviously find a $q \in V_+$ such that $K_1[q] \cap \overline{W}_2[q] = \varnothing$. Let q be so chosen. The family $\mathscr{F}$ again contains two wedges in a general relative position, and by the same kind of reasoning as in step (1) we conclude that $\{\mathscr{A}(W) | W \in \mathscr{F}\}'' = (\mathscr{A}(K_1^c) \cup \mathscr{A}(\overline{W}_2^c))'' = \mathsf{B}(\mathscr{H})$. The first assertion in part (b) then follows by duality.

(3) We consider part (c). It will be convenient to introduce the wedges $W_a, W_b \in \mathscr{W}$ by $W_a = \overline{W}_1^c$ and $W_b = \overline{W}_2^c$, and the premise is thus that $W_a^c \cap W_b^c = \varnothing$. The two cases $W_a \cap W_b \neq \varnothing$ and $W_a \cap W_b = \varnothing$ have to be considered separately, and we first consider the case when $W_a \cap W_b \neq \varnothing$. Let $x_i \in W_a \cap W_b$. Since the set $W_a \cap W_b$ is open we can select a $q \in V_+$ such that the closed timelike line segment $[x_i - q, x_i + q]$ is included in $W_a \cap W_b$. Let q be so chosen. As in Lemma 4.8 we consider the wedges $W_a[q] \subset W_a$ and $W_b[q] \subset W_b$, and we obviously have $x_i \in W_a[q] \cap W_b[q]$, and hence $W_a[q] \cap W_b[q] \neq \varnothing$. We assert that $W_a[q]^c \cap W_b[q]^c = \varnothing$, and to prove this we consider the possibility that there exists a point $y \in W_a[q]^c \cap W_b[q]^c$. Let l be the line passing through the points x_i and y. This line is *spacelike*, since $x_i \in W_a[q]$ and $y \in W_a[q]^c$, and its intersections with $W_a[q]^c$ and W_a^c are half-infinite intervals which do not contain x_i. Similarly the intersections of l with $W_b[q]^c$ and W_b^c are half-infinite intervals which do not contain x_i. It follows that $W_a^c \cap W_b^c$ contains a half-infinite interval on l, which is a contradiction since $W_a^c \cap W_b^c = \varnothing$. We conclude that $W_a[q]^c \cap W_b[q]^c = \varnothing$.

By the above, and by part (a) of Lemma 4.6, we then have $\mathscr{E}(W_a[q] \cup W_b[q]) = (W_a[q] \cup W_b[q])^{cc} = (W_a[q]^c \cap W_b[q]^c)^c = \mathsf{M}$. We write $S = W_a \cup W_b$ and $S[q] = W_a[q] \cup W_b[q]$. We then have $D_q(x) \subset S$, and hence $\mathscr{B}(D_q(x)) \subset (\mathscr{A}(W_a) \cap \mathscr{A}(W_b))''$, whenever $x \in S[q]$. By (4.11) we then conclude that $(\mathscr{A}(\overline{W}_1^c) \cup \mathscr{A}(\overline{W}_2^c))'' = (\mathscr{A}(W_a) \cap \mathscr{A}(W_b))'' = \mathsf{B}(\mathscr{H})$, and, by duality, that $\mathscr{A}(W_1) \cap \mathscr{A}(W_2) = \mathsf{C}^1 \cdot I$.

(4) It remains to consider the case when $W_a \cap W_b = \varnothing$ and $W_a^c \cap W_b^c = \varnothing$. The upper plane of W_a is then parallel with the lower plane of W_b, and the lower plane of W_a is parallel with the upper plane of W_b. There obviously exists a $q \in V_+$ such that $W_a[q]^c \cap W_b[q]^c = \varnothing$. Let q be so chosen. We trivially have $W_a[q] \cap W_b[q] = \varnothing$. We have $D_q(x) \subset W_a \cup W_b$, and hence $\mathscr{B}(D_q(x)) \subset \mathscr{A}(W_a) \cup \mathscr{A}(W_b)$, whenever $x \in W_a[q] \cup W_b[q]$. Let $B \in \mathscr{A}(W_1) \cap \mathscr{A}(W_2) = (\mathscr{A}(W_a)^z \cup \mathscr{A}(W_b)^z)'$ and let $A \in \mathscr{B}(D_q)^z$. We then have $[B, T(x) A T(x)^{-1}] = 0$ whenever $x \in W_a[q] \cup W_b[q]$, and by Lemma 2.7 we conclude that $\langle \Omega | BA\Omega \rangle = \langle \Omega | B\Omega \rangle \langle \Omega | A\Omega \rangle$. Since Ω is cyclic for $\mathscr{B}(D_q)^z$ it follows that $B^* \Omega = b\Omega$, for some constant b, and since Ω is separating for $\mathscr{A}(W_1) \cap \mathscr{A}(W_2)$ it follows that $B = bI$. This completes the proof.

In the discussion up to this point we have made no assumptions about the *condition of additivity for wedges*. For the discussion which follows we will need to assume this condition.

Theorem 4.10: Let $\mathscr{N}$ be a *regular* net, as described in Definition 3.1, with algebras $\mathscr{A}(W)$, $\mathscr{A}(K^c)$, and $\mathscr{B}(K)$ defined for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$. Let the associated algebras $\mathscr{A}^\sharp(K^c)$ and $\mathscr{B}^\sharp(K)$ be defined as in (4.1) for all $K \in \mathscr{K}$. In addition we assume that the net $\mathscr{N}$ satisfies the *condition of additivity for wedges*, in accordance with Definition 2.5.

Let $\mathscr{F}$ be a family of wedges in $\mathscr{W}$. We write $S=\cup\{W|W\in\mathscr{F}\}$ and $K=\cap\{W^c|W\in\mathscr{F}\}$. Let $\mathscr{E}(S)$ be defined as in Lemma 4.5.

(a) Suppose that $\mathscr{F}$ contains two wedges in a *general relative position*. Then K is an element of the set $\mathscr{K}_E$ as described in Definition 1.1, and we have $\mathscr{E}(S)=S^{cc}=K^c$, and

$$\cap\{\mathscr{A}(\overline{W}^c)|W\in\mathscr{F}\}=\mathscr{B}^\sharp(K),\quad \{\mathscr{A}(W)|W\in\mathscr{F}\}''=\mathscr{A}^\sharp(S^{cc}). \tag{4.12}$$

(b) Suppose instead that the normals of the boundary planes of the wedges in $\mathscr{F}$ span M, and suppose furthermore that $\mathscr{F}$ contains two wedges such that no two of the boundary planes of these wedges are parallel. Then the conclusions are the same as in part (a).

Proof: (1) We note that the premises in part (a) imply the premises in part (b), and it thus suffices to consider part (b). By Definition 1.1 we have $K\in\mathscr{K}_E$, and by part (b) of Lemma 4.6 we have $\mathscr{E}(S)=S^{cc}$.

Let $q\in V_+$. We define $W[q]$, as in Lemma 4.8, for any $W\in\mathscr{W}$, and we write $S[q]=\cup\{W[q]|W\in\mathscr{F}\}$. By Lemma 4.8 we then have

$$S^{cc}=\mathscr{E}(S)=\cup\{\mathscr{E}(S[q/n])|n\in\mathsf{Z}_+\}. \tag{4.13}$$

(2) Let $D_q\in\mathscr{D}$ be the double cone with apices $(-q,q)$, and let $D_q(x)=\{x+y|y\in D_q\}$, as in Lemma 4.8. We write $\mathscr{A}_S=\{\mathscr{A}(W)|W\in\mathscr{F}\}''$, and by Lemma 4.8 we then have

$$\{\mathscr{B}(D_q(x))|x\in\mathscr{E}(S[q])\}''\subset\mathscr{A}_S. \tag{4.14}$$

Let $q\in V_+$ be fixed. Let $W\in\mathscr{W}$, $\overline{W}\subset S^{cc}$, and let $x\in\overline{W}$. In view of (4.13) there then exists a positive integer n (dependent on x) such that $x\in\mathscr{E}(S[q/n])$, and by replacing q in (4.14) by q/n we conclude that $\mathscr{B}(D_{q/n}(x))\subset\mathscr{A}_S$. By the assumed condition of additivity for wedges it follows that $\mathscr{A}(W)\subset\mathscr{A}_S$. Since the wedge-algebras are continuous from the inside, as expressed in (2.6), the relation $\mathscr{A}(W)\subset\mathscr{A}_S$ also holds for all $W\in\mathscr{W}$ such that $W\subset S^{cc}$, and since $\mathscr{A}^\sharp(S^{cc})=\{\mathscr{A}(W)|W\in\mathscr{W},\ W\subset S^{cc}\}''$ by (4.1) it follows that $\mathscr{A}^\sharp(S^{cc})\subset\mathscr{A}_S$.

Since $\mathscr{F}\subset\{W|W\in\mathscr{W},\ W\subset S^{cc}\}$ it follows that $\mathscr{A}_S\subset\{\mathscr{A}(W)|W\in\mathscr{W},\ W\subset S^{cc}\}''=\mathscr{A}^\sharp(K^c)$, and hence we obtain the identity at right in (4.12). The identity at left then follows by duality. This completes the proof.

Let $\mathscr{F}^*\subset\mathscr{W}$ be defined by $\mathscr{F}^*=\{W|W\in\mathscr{W},W\supset K\}$, in which case we have $\mathscr{F}\subset\mathscr{F}^*$, and $\mathscr{B}^\sharp(K)=\cap\{\mathscr{A}(W)|W\in\mathscr{F}^*\}$ by the definition of $\mathscr{B}^\sharp(K)$. The point of the theorem is that $\mathscr{F}$ can be ''very much smaller'' than $\mathscr{F}^*$, and in general the set $S^{cc}=\cup\{W|W\in\mathscr{F}^*\}$ is strictly larger than the set $S=\cup\{W|W\in\mathscr{F}\}$. The conclusion is therefore not trivial.

Theorem 4.11: Let $\mathscr{N}$ be a *regular* net with algebras $\mathscr{A}(W)$, $\mathscr{A}(K^c)$, and $\mathscr{B}(K)$, and associated algebras $\mathscr{A}^\sharp(K^c)$ and $\mathscr{B}^\sharp(K)$, as in Theorem 4.10. In addition we assume that the net $\mathscr{N}$ satisfies the *condition of additivity for wedges*, as in Definition 2.5.

(a) Let $K_0\in\mathscr{K}_C$, and let $\mathscr{F}$ be any family of wedges in $\mathscr{W}$. We write $K=(\cap\{W^c|W\in\mathscr{F}\})\cap K_0$. Then $K\in\mathscr{K}_C$ and

$$\mathscr{B}(K)=(\cap\{\mathscr{A}(\overline{W}^c)|W\in\mathscr{F}\})\cap\mathscr{B}(K_0). \tag{4.15}$$

More generally, if $K_0\in\mathscr{K}_E$, and K is defined in terms of K_0 and $\mathscr{F}$ as above, then $K\in\mathscr{K}_E$ and

$$\mathscr{B}^\sharp(K)=(\cap\{\mathscr{A}(\overline{W}^c)|W\in\mathscr{F}\})\cap\mathscr{B}^\sharp(K_0). \tag{4.16}$$

(b) Let $\mathscr{T}_K$ be any subset of $\mathscr{K}_C$, and let $K_i=\cap\{K|K\in\mathscr{T}_K\}$. Then $K_i\in\mathscr{K}_C$ and

$$\mathscr{B}(K_i)=\cap\{\mathscr{B}(K)|K\in\mathscr{T}_K\}. \tag{4.17}$$

More generally, if $\mathscr{T}_K$ is a subset of $\mathscr{K}_E$, then $K_i=\cap\{K|K\in\mathscr{T}_K\}\in\mathscr{K}_C$ and

$$\mathscr{B}^\sharp(K_i)=\cap\{\mathscr{B}^\sharp(K)|K\in\mathscr{T}_K\}. \tag{4.18}$$

Proof: (1) Suppose that $K_0 \in \mathscr{K}_E$. Let $\mathscr{T}_0 = \{W | W \in \mathscr{W},\ W \subset K_0^c\}$ and let $\mathscr{T}_u = \mathscr{T} \cup \mathscr{T}_0$. We have $K_0 = \cap\{W^c | W \in \mathscr{T}_0\}$, $\mathscr{B}^\sharp(K_0) = \cap\{.\mathscr{A}(\overline{W^c}) | W \in \mathscr{T}_0\}$, and $K = \cap\{W^c | W \in \mathscr{T}_u\}$. Since $K_0 \in \mathscr{K}_E$, the family $\mathscr{T}_0$ contains two wedges in a general relative position, and the same then holds for the family $\mathscr{T}_u$. By Theorem 4.10 we have $\mathscr{B}^\sharp(K) = \cap\{.\mathscr{A}(\overline{W^c}) | W \in \mathscr{T}_u\}$, and the relation (4.16) follows. If K_0 is in the subset $\mathscr{K}_C$ of $\mathscr{K}_E$ we have $K \in \mathscr{K}_C$, and the relation (4.16) can then be written in the form in (4.15), since $\mathscr{B}^\sharp(K) = \mathscr{B}(K)$ for all $K \in \mathscr{K}_C$.

(2) Suppose that $\mathscr{T}_K \subset \mathscr{K}_E$. For each $K \in \mathscr{T}_K$ we define the family $\mathscr{T}[K]$ of wedges by $\mathscr{T}[K] = \{W | W \in \mathscr{W},\ W^c \supset K\}$. With $\mathscr{T} = \cup\{\mathscr{T}[K] | K \in \mathscr{T}_K\}$ we have $K_i = \cap\{W^c | W \in \mathscr{T}\}$, and the relation (4.18) follows readily on the basis of Theorem 4.10. The relation (4.17) refers to the special case when $\mathscr{T}_K \subset \mathscr{K}_C$, in which case we also have $K_i \in \mathscr{K}_C$. This completes the proof.

The special case of part (b) when $\mathscr{T}_K$ contains just two elements $K_1, K_2 \in \mathscr{K}_C$ is of particular interest. We then have the simple conclusion that $\mathscr{B}(K_1) \cap .\mathscr{B}(K_2) = \mathscr{B}(K_1 \cap K_2)$. If $K_1 \cap K_2 = \varnothing$ we have $\mathscr{B}(K_1) \cap .\mathscr{B}(K_2) = \mathsf{C}^1 \cdot I$, which is also the conclusion in Theorem 4.9, *without* the assumption that the net $\mathscr{N}$ satisfies the condition of additivity for wedges. The condition for a *regular* net that Ω is cyclic for $\mathscr{B}(D)$ for *any* $D \in \mathscr{D}$, was, however, crucial for the proof of Theorem 4.9. The condition of additivity for wedges is stronger than this condition, and the question arises as to whether it is a necessary premise of Theorem 4.11 or of Theorem 4.10. It is not hard to see that the intersection property $\mathscr{B}(K_1) \cap .\mathscr{B}(K_2) = .\mathscr{B}(K_1 \cap K_2)$ can be proved without the additivity condition for certain *special* configurations. Suppose that $D_1 = D(x', x'')$ and $D_2 = D(y', y'')$ are two double cones so located that $x'' \in D_2^o$ and $y' \in D_1^o$. We then have $D_1 \cap D_2 = D(y', x'')$, and it is easily seen that the relation $\mathscr{B}(D_1) \cap .\mathscr{B}(D_2) = .\mathscr{B}(D_1 \cap D_2)$ can be proved just on the basis of Proposition 4.3. There are, however, other configurations for which, as far as we know, the intersection properties *cannot* be proved without the condition of additivity for wedges (or some similar condition).

We next consider the intersection of two wedge-algebras.

Theorem 4.12: Let the premises and notation be as in Theorem 4.11. Let $W_1, W_2 \in \mathscr{W}$. We write $S = W_1 \cup W_2$ and $K = W_1^c \cap W_2^c = S^c$, and hence $K \in \mathscr{K}$ and we have $\mathscr{B}^\sharp(K) = \cap\{.\mathscr{A}(W) | W \in \mathscr{W},\ \overline{W} \supset K\}$ and $\mathscr{A}^\sharp(S^{cc}) = \{.\mathscr{A}(W) | W \in \mathscr{W},\ W \subset S^{cc}\}''$ in accordance with (4.1).

(a) If $W_1 \cap W_2 \neq \varnothing$, then

$$\mathscr{A}(\overline{W_1^c}) \cap .\mathscr{A}(\overline{W_2^c}) = \mathscr{B}^\sharp(W_1^c \cap W_2^c), \quad (.\mathscr{A}(W_1) \cup .\mathscr{A}(W_2))'' = \mathscr{A}^\sharp((W_1 \cup W_2)^{cc}). \tag{4.19}$$

In particular this applies if W_1 and W_2 are in a general relative position.

(b) If $K = W_1^c \cap W_2^c = \varnothing$, then the relations in (4.19) also hold, and we have $\mathscr{B}(K) = \mathsf{C}^1 \cdot I$ and $\mathscr{A}(S^{cc}) = \mathsf{B}(\mathscr{H})$.

(c) If the span of the normals of the boundary planes of the wedges W_1 and W_2 is *two-dimensional*, then the relations in (4.19) hold.

(d) If $W_1 \cap W_2 = \varnothing$, and if the span of the normals of the boundary planes of the wedges is *three-dimensional*, then $K = W_1^c \cap W_2^c \neq \varnothing$, and the relations in (4.19) *do not* hold generally.

Proof: (1) By Proposition 3.10 the two relations in (4.19) are equivalent by duality. We write $\mathscr{A}_i = (.\mathscr{A}(W_1) \cup .\mathscr{A}(W_2))''$, and we note that we trivially have $\mathscr{A}_i \subset \mathscr{A}^\sharp((W_1 \cup W_2)^{cc})$, and we thus have to prove the opposite inclusion. We first note that the conclusion in part (b) follows at once from part (c) of Theorem 4.9.

(2) We consider part (a), assuming that $W_1 \cap W_2 \neq \varnothing$. Theorem 4.10 is not directly applicable if W_1 and W_2 are *not* in a general relative position, but by part (a) of Lemma 4.6 we have $S^{cc} = \mathscr{E}(S)$, and the relation (4.13) holds as stated. The reasoning in step (2) of the proof of Theorem 4.10 is therefore also applicable to the case at hand, and we can conclude that the relations in (4.19) hold.

(3) We consider part (c). There are two possibilities. Either W_2 is obtained by a parallel displacement of W_1, or else W_2 is obtained by a parallel displacement of $\overline{W_1^c}$. In the first case the conclusion follows readily from Proposition 4.3. We consider the second case, and we can now assume that $K = W_1^c \cap W_2^c \neq \varnothing$, in view of the result in part (b). It is then easily seen that $W \supset K$ if and only if either $W \supset W_1^c$ or else $W \supset W_2^c$, and the relation at left in (4.19) then follows trivially.

(4) We consider part (d). For a counterexample which shows that the relations in (4.19) cannot be generally valid, we refer to Proposition 5.18 in TW-II,[7] and specifically to part (d) of that

proposition. The counterexample concerns the theory of a single free Hermitian massive scalar field. The von Neumann algebra $A(W)$ is then the *weak commutant* denoted $\mathscr{C}(W)$ in the quoted paper, for any $W\in\mathscr{W}$, and with the notation in the paper we have $\mathscr{C}(W_1^c\cap W_2^c)=\cap\{\mathscr{A}(W)|W\in\mathscr{W},\ W\supset W_1^c\cap W_2^c\}$. An example is presented for which $\mathscr{C}(W_1^c\cap W_2^c)\neq\mathscr{C}(W_1^c)\cap\mathscr{C}(W_2^c)$, and hence the relations in (4.19) fail to hold in this case. This completes the proof.

The relation at left in (4.19) describes an ''intersection property,'' and as we have seen, it can fail in a free-field theory for certain configurations of two wedges. When it does fail to hold, we necessarily also have

$$\mathscr{A}(\overline{W}_1^c)\cap\mathscr{A}(\overline{W}_2^c)\neq\{\mathscr{B}(K_0)|K_0\in\mathscr{K}_C,\ K_0\subset W_1^c\cap W_2^c\}'', \tag{4.20}$$

because it is easily seen that if equality obtains in (4.20), then the relation at left in (4.19) follows. The algebra $\mathscr{A}(\overline{W}_1^c)\cap\mathscr{A}(\overline{W}_2^c)$ is, of course, in the weak closure of the set $\mathscr{U}$ of *all* local operators, but it is *not* in the weak closure of the set of all local operators ''localized'' within the region $W_1^c\cap W_2^c$, contrary to what one might expect at first.

We note here that the relations in (4.19) follow at once from Theorem 4.10 in the case when W_1 and W_2 are in a general relative position, i.e., when the normals of the boundary planes of the wedges span M, in which case we have $K=W_1^c\cap W_2^c\in\mathscr{K}_E$. Thus the ''pathology'' can only occur for configurations which are essentially ''three-dimensional.'' In these cases K is an element of $\mathscr{K}$ but *not* an element of $\mathscr{K}_E$.

We conclude this section by noting a simple implication of the condition of additivity for wedges.

Proposition 4.13: Let $\mathscr{N}$ be a *regular* net, as described in Definition 3.1, with algebras $\mathscr{A}(W)$, $\mathscr{A}(K^c)$, and $\mathscr{B}(K)$. In addition we assume that the net $\mathscr{N}$ satisfies the *condition of additivity for wedges*, in accordance with Definition 2.5. Let $D\in\mathscr{D}$ be *fixed*. We denote by $D(x)$ the double cone obtained from D by a translation by any $x\in\mathsf{M}$. Then

$$\mathscr{A}(W)=\{\mathscr{B}(D(x))|x\in\mathsf{M},\ D(x)\subset W\}'',\quad \mathscr{A}(K^c)=\{\mathscr{B}(D(x))|x\in\mathsf{M},\ D(x)\subset K^c\}'',$$

for any $W\in\mathscr{W}$, and any $K\in\mathscr{K}_C$.

Proof: The relation at right follows immediately from the relation at left, since $\mathscr{A}(K^c)=\{\mathscr{A}(W)|W\in\mathscr{W},\ W\subset K^c\}''$. It thus suffices to prove the relation at left. We write $\mathscr{A}=\{\mathscr{B}(D(x))|x\in\mathsf{M},\ D(x)\subset W\}''$. We trivially have $\mathscr{A}(W)\supset\mathscr{A}$, and we thus have to prove the opposite inclusion.

With a suitable choice of standard coordinates we have $W=\{x|x\in\mathsf{M},\ x_3>|x_4|\}$, and without any essential loss of generality we can consider the double cone $D\in\mathscr{D}$ with apices $\pm(\mathbf{y},\tau)$ for some $\mathbf{y}\in\mathsf{R}^3$ and some $\tau>|\mathbf{y}|$. Let $\tau>\delta>0$, and let $W_{u,\delta}=\{x|x\in\mathsf{M},\ x_3-\tau>|x_4-\tau+\delta|\}$ and $W_{l,\delta}=\{x|x\in\mathsf{M},\ x_3-\tau>|x_4+\tau-\delta|\}$. We easily see that we then have

$$W_{u,\delta}\cup W_{l,\delta}\subset\cup\{D^o(x)|x\in\mathsf{M},\ D(x)\subset W\},$$

and hence $\mathscr{A}(W_{u,\delta})\cup\mathscr{A}(W_{l,\delta})\subset\mathscr{A}$ by the condition of additivity. We trivially have $(W_{u,\delta}\cup W_{l,\delta})^{cc}=\{x|x\in\mathsf{M},\ x_3-\delta>|x_4|\}=W_\delta$, and it then follows by Proposition 4.3 that $\mathscr{A}(W_\delta)\subset\mathscr{A}$. Since δ can be selected arbitrarily small it follows from the continuity relations in (2.6) that $\mathscr{A}(W)\subset\mathscr{A}$. This completes the proof.

The point is that the set $\cup\{D^o(x)|x\in\mathsf{M},\ D(x)\subset W\}$ is in general *smaller* than W, i.e., the double cones $D(x)$ contained in W do not in general ''sweep out'' the wedge, and hence the conclusion is not immediately obvious.

V. ABOUT SUBNETS OF REGULAR NETS

In this section we will discuss local subnets of *regular nets* which satisfy the *special condition of duality*. Before we do that we make some general observations. We refer to part (c) in Definition 2.1 for the general (obvious) notion of a subnet $\mathscr{N}_s$ of a local net $\mathscr{N}$. Let $\mathscr{N}$ be an *irreducible* local net, with algebras $\mathscr{A}(W)$, $\mathscr{B}(K)$, and $\mathscr{A}(K^c)$ for all $W\in\mathscr{W}$ and $K\in\mathscr{K}_C$. Let $\mathscr{N}_s$ be a *subnet* of $\mathscr{N}$ with von Neumann algebras $\mathscr{A}_s(W)\subset\mathscr{A}(W)$, $\mathscr{B}_s(K)\subset\mathscr{B}(K)$, and $\mathscr{A}_s(K^c)\subset\mathscr{A}(K^c)$ for all $W\in\mathscr{W}$ and $K\in\mathscr{K}_C$. Both nets $\mathscr{N}$ and $\mathscr{N}_s$ thus satisfy the basic

Conditions I–V in Definition 2.1. These include the condition of Poincaré-covariance, but not *PCT*-covariance, as it could happen that a subnet $\mathcal{N}_s$ of a *PCT*-covariant net is not *PCT*-covariant.

We consider the algebra $\mathcal{U}=\cup\{\mathcal{B}(K)|K\in\mathcal{K}_C\}=\cup\{\mathcal{B}(D)|D\in\mathcal{D}\}$ of all local operators of $\mathcal{N}$, and the algebra $\mathcal{U}_s=\cup\{\mathcal{B}_s(K)|K\in\mathcal{K}_C\}=\cup\{\mathcal{B}_s(D)|D\in\mathcal{D}\}$ of all local operators of $\mathcal{N}_s$. Our particular notion of a local net requires that $\mathcal{A}(\mathsf{M})=\mathcal{U}''$ and $\mathcal{A}_s(\mathsf{M})=\mathcal{U}_s''$. Since we assumed that the net $\mathcal{N}$ is irreducible we have $\mathcal{A}(\mathsf{M})=\mathsf{B}(\mathcal{H})$, but in general $\mathcal{A}_s(\mathsf{M})$ is smaller than $\mathsf{B}(\mathcal{H})$, and Ω is not cyclic for $\mathcal{A}_s(\mathsf{M})$. We remark here that it could happen that there exists a system of sub-algebras of the algebras of a local net (in our sense) which satisfies all the other conditions on a local net, except for the condition $\mathcal{A}_s(\mathsf{M})=\mathcal{U}_s''$. This is an interesting possibility about which we have nothing to say here.

We define the sub-Hilbert space $\mathcal{H}_s$ by $\mathcal{H}_s=F_s\mathcal{H}=\overline{\mathcal{A}_s(\mathsf{M})\Omega}$, where F_s is the self-adjoint projection onto $\mathcal{H}_s$. Hence $\mathcal{A}_s(\mathsf{M})\subset\{F_s\}'$. The subspace $\mathcal{H}_s$ is Poincaré-invariant, i.e., $U(\lambda)F_s=F_sU(\lambda)$ for all $\lambda\in\hat{\mathsf{P}}$. Corresponding to the decomposition of $\mathcal{H}$ as the direct sum $\mathcal{H}=\mathcal{H}_s\oplus\mathcal{H}_s^{\perp}$ of the two orthogonal subspaces $\mathcal{H}_s$ and $\mathcal{H}_s^{\perp}$, we can write any bounded operator X on $\mathcal{H}$ in the block-matrix form

$$X=\begin{bmatrix}\rho(X), & \alpha(X)\\ \beta(X), & \gamma(X)\end{bmatrix},\quad F_s=\begin{bmatrix}I, & 0\\ 0, & 0\end{bmatrix},\tag{5.1}$$

and $\rho(X)$ is thus the restriction of the operator F_sXF_s to $\mathcal{H}_s$. Hence $\rho(X)^*=\rho(X^*)$. For any set $\mathcal{S}$ of operators on $\mathcal{H}$ we will write $\rho(\mathcal{S})=\{\rho(X)|X\in\mathcal{S}\}$. The mapping $\lambda\to U_s(\lambda)=\rho(U(\lambda))$ is a (strongly continuous) unitary representation of the group $\hat{\mathsf{P}}$ on $\mathcal{H}_s$. We have $\rho(X^z)=Z_s\rho(X)Z_s^{-1}$, with $Z_s=\rho(Z)$, and we will write $\rho(X)^z=\rho(X^z)$. With reference to this representation of $\hat{\mathsf{P}}$, the sets $\hat{\mathcal{A}}_s(W)=\rho(\mathcal{A}_s(W))$, $\hat{\mathcal{B}}_s(K)=\rho(\mathcal{B}_s(K))$, and $\hat{\mathcal{A}}_s(K^c)=\rho(\mathcal{A}_s(K^c))$, for $W\in\mathcal{W}$ and $K\in\mathcal{K}_C$, are the von Neumann algebras of an irreducible local net $\hat{\mathcal{N}}_s$, called the *vacuum representation* of the subnet $\mathcal{N}_s$.

The general Conditions I–V in Definition 2.1 leave a considerable latitude for the structure of a net. As we showed in Proposition 3.8 it is always possible to construct a ''standard net'' which satisfies the conditions in (3.13) and (3.14) from a given net, and this new net then satisfies Condition IX in Definition 2.1, i.e., it has a *strictly local structure*. If the irreducible net $\mathcal{N}$ has this structure, it is natural to consider subnets $\mathcal{N}_s$ which also have this structure. For a particular subnet $\mathcal{N}_s$ the operators in the set $\mathcal{A}_s(\mathsf{M})\cap\mathcal{U}$ are *local* operators (of the original net $\mathcal{N}$), and we want to arrange it so that they are identified as local operators of the subnet, i.e., we should have $\mathcal{U}_s=\mathcal{A}_s(\mathsf{M})\cap\mathcal{U}$. An arbitrary subnet does not necessarily have this property.

Let $\mathcal{G}$ be a group of *local internal symmetries* for a *regular* net $\mathcal{N}$, i.e., $\mathcal{G}$ is a group of unitary operators such that $G\Omega=\Omega$, and $G\mathcal{B}(K)G^{-1}=\mathcal{B}(K)$ for all $G\in\mathcal{G}$ and all $K\in\mathcal{K}_C$. The set $\mathcal{N}_s$ of algebras defined by $\mathcal{B}_s(K)=\mathcal{B}(K)\cap\mathcal{G}'$, $\mathcal{A}_s(K^c)=\mathcal{A}(K^c)\cap\mathcal{G}'$, and $\mathcal{A}_s(W)=\mathcal{A}(W)\cap\mathcal{G}'$, for all $K\in\mathcal{K}_C$ and all $W\in\mathcal{W}$, is then, what we will call below, a *regular* subnet of $\mathcal{N}$. This is an important ''mechanism'' for the production of subnets, which has been discussed extensively by Doplicher, Haag, and Roberts.[25] For a subnet of this kind one can derive many of the results discussed in what follows *without* assuming the special condition of duality. One can show by examples, however, that a subnet is not always obtained in this fashion. See also the discussion by Araki[26] of subnets in general.

An interesting general question concerns those properties of the net $\mathcal{N}$ which are *hereditary* in the sense that they are also properties of the *reduced* subnet $\hat{\mathcal{N}}_s$. Among such properties we can consider the condition of *PCT*-covariance, the condition of duality, the property that Ω is cyclic for $\mathcal{B}(K)$ for every $K\in\mathcal{K}_C$ with a nonempty interior, or the properties discussed in Section IV. It is obvious that an *arbitrary* reduced subnet $\hat{\mathcal{N}}_s$ does not necessarily inherit such special properties of $\mathcal{N}$.

Throughout the remainder of this section we will assume that $\mathcal{N}$ is a *regular* net which satisfies the *special condition of duality*. This assumption leads to a particular satisfactory theory of subnets, and the basic reason for this is that the modular operators for the wedge-algebras (relative to the vacuum vector Ω) are of a ''geometric nature,'' i.e., related to the Poincaré group and the *PCT*-operator, as described in Definition 2.4. The three theorems which follow depend

substantially on results in a paper by Wichmann[5] on subnets of local nets. The first describes important features of a ''regular'' subnet, the second describes the generation of such a subnet by any set of local operators, and the third describes properties of the reduced subnet. It will be convenient to combine the proof of these three theorems.

Theorem 5.1: Let $\mathcal{N}$ be a *regular* net, as in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$ for all $W\in\mathcal{W}$ and $K\in\mathcal{K}_C$. Furthermore we assume that $\mathcal{N}$ satisfies the *special condition of duality*, in accordance with Definition 2.4. Let $\mathcal{N}_s$ be a subnet of $\mathcal{N}$, with von Neumann algebras $\mathcal{A}_s(W)\subset\mathcal{A}(W)$, $\mathcal{B}_s(K)\subset\mathcal{B}(K)$, and $\mathcal{A}_s(K^c)\subset\mathcal{A}(K^c)$, which satisfy the conditions

$$\mathcal{B}_s(K)=\cap\{\mathcal{A}_s(W)\,|\,W\in\mathcal{W},\ \overline{W}\supset K\},\quad \mathcal{A}_s(K^c)=\{\mathcal{A}_s(W)\,|\,W\in\mathcal{W},\ W\subset K^c\}'', \tag{5.2}$$

and

$$\mathcal{A}_s(W)=\{\mathcal{B}_s(D)\,|\,D\in\mathcal{D},\ D\subset W\}'' \tag{5.3}$$

for all $K\in\mathcal{K}_C$, and all $W\in\mathcal{W}$. With these premises we say that $\mathcal{N}_s$ is a *regular subnet* of $\mathcal{N}$.

Let $\mathcal{U}=\cup\{\mathcal{B}(K)\,|\,K\in\mathcal{K}_C\}$ and $\mathcal{U}_s=\cup\{\mathcal{B}_s(K)\,|\,K\in\mathcal{K}_C\}$. Hence $\mathcal{A}_s(\mathsf{M})=\mathcal{U}_s''$. Let $\mathcal{H}_s=\overline{\mathcal{A}_s(\mathsf{M})\Omega}=F_s\mathcal{H}$, where F_s is the self-adjoint projection onto $\mathcal{H}_s$. Then:

(a) The subnet $\mathcal{N}_s$ is *PCT*-covariant, and the subspace $\mathcal{H}_s$ is Poincaré- and *PCT*-invariant.

(b)

$$\mathcal{A}_s(W)=\mathcal{A}(W)\cap\mathcal{A}_s(\mathsf{M})=\mathcal{A}(W)\cap\{F_s\}',\quad \mathcal{B}_s(K)=\mathcal{B}(K)\cap\mathcal{A}_s(\mathsf{M})=\mathcal{B}(K)\cap\{F_s\}', \tag{5.4}$$

for all $W\in\mathcal{W}$ and all $K\in\mathcal{K}_C$, and hence $\mathcal{U}_s=\mathcal{U}\cap\mathcal{A}_s(\mathsf{M})=\mathcal{U}\cap\{F_s\}'$.

Furthermore,

$$\mathcal{A}_s(K^c)=\mathcal{A}(K^c)\cap\mathcal{A}_s(\mathsf{M})=\mathcal{A}(K^c)\cap\{F_s\}', \tag{5.5}$$

for all $K\in\mathcal{K}_C$ with a *nonempty* interior.

(c) We have $\mathcal{A}_s(\mathsf{M})\Omega=\mathcal{H}_s$, and

$$\overline{\mathcal{B}_s(K)\Omega}=\mathcal{H}_s,$$

for every $K\in\mathcal{K}_C$ with a *nonempty* interior.

(d) If $\mathcal{H}_s=\mathcal{H}$, then $\mathcal{N}_s=\mathcal{N}$, i.e., $\mathcal{A}_s(W)=\mathcal{A}(W)$, $\mathcal{B}_s(K)=\mathcal{B}(K)$, and $\mathcal{A}_s(K^c)=\mathcal{A}(K^c)$.

Theorem 5.2: Let $\mathcal{N}$ be a *regular* net, as in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$ for all $W\in\mathcal{W}$ and $K\in\mathcal{K}_C$, which satisfies the *special condition of duality*. Let $\mathcal{U}=\cup\{\mathcal{B}(K)\,|\,K\in\mathcal{K}_C\}$ be the algebra of all local operators in $\mathcal{N}$.

(a) Let $\mathcal{S}$ be a subset of $\mathcal{U}$. Let $\mathcal{P}_s$ be the *smallest Poincaré-invariant* *-algebra which contains $\mathcal{S}$. We define the von Neumann algebras $\mathcal{A}_s(W)$, $\mathcal{B}_s(K)$, and $\mathcal{A}_s(K^c)$ by

$$\mathcal{A}_s(W)=\{\mathcal{P}_s\cap\mathcal{B}(D)\,|\,D\in\mathcal{D},\ D\subset W\}'', \tag{5.6}$$

for all $W\in\mathcal{W}$, and by (5.2) for all $K\in\mathcal{K}_C$. Then this system of algebras are the von Neumann algebras of a *regular* subnet $\mathcal{N}_s$ of $\mathcal{N}$, which satisfies the premises in Theorem 5.1, and hence all the conclusions in that theorem apply. Furthermore,

$$\mathcal{P}_s\subset\mathcal{U}_s,\quad \mathcal{P}_s''=\mathcal{U}_s''=\mathcal{A}_s(\mathsf{M}),\quad \overline{\mathcal{P}_s\Omega}=\overline{\mathcal{A}_s(\mathsf{M})\Omega}=\mathcal{H}_s, \tag{5.7}$$

where $\mathcal{U}_s=\cup\{\mathcal{B}_s(K)\,|\,K\in\mathcal{K}_C\}$, and where the sub-Hilbert space $\mathcal{H}_s$ is defined as in Theorem 5.1.

We say that the subnet $\mathcal{N}_s$ is *generated by* the set of local operators $\mathcal{S}$, and we express this by the notation $\mathcal{N}_s=\mathcal{N}(\mathcal{S})$ when appropriate.

(b) Let $\mathcal{S}_1$ be another subset of $\mathcal{U}$, and let $\mathcal{H}_{s1}$ be the corresponding subspace of $\mathcal{H}$, constructed in terms of $\mathcal{S}_1$ as above. Then $\mathcal{N}(S_1)=\mathcal{N}(S)$ if $\mathcal{H}_{s1}=\mathcal{H}_s$.

Theorem 5.3: Let the premises and notation be as in Theorem 5.1. As before, we denote by $\rho(X)$ the restriction of $F_s X F_s$ to $\mathscr{H}_s$ for any bounded linear or antilinear operator X on $\mathscr{H}$, and we write $\rho(\mathscr{Q})=\{\rho(X)|X\in\mathscr{Q}\}$ for any set $\mathscr{Q}$ of such operators. Then:

(a) The sets $\hat{\mathscr{A}}_s(W)=\rho(\mathscr{A}_s(W))$, $\hat{\mathscr{B}}_s(K)=\rho(\mathscr{B}_s(K))$, and $\hat{\mathscr{A}}_s(K^c)=\rho(\mathscr{A}_s(K^c))$, for all $W\in\mathscr{W}$ and all $K\in\mathscr{K}_C$, are the von Neumann algebras of a *regular* local net $\hat{\mathscr{N}}_s$ which satisfies the *special condition of duality* (with respect to the representation $\lambda\to U_s(\lambda)=\rho(U(\lambda))$ of the group $\hat{\mathsf{P}}$).

In particular, the net $\hat{\mathscr{N}}_s$ is *PCT*-covariant, with the *PCT*-operator $\theta_s=\rho(\theta)$, and the vector Ω, regarded as a vector in $\mathscr{H}_s$, is *cyclic and separating* for $\hat{\mathscr{B}}_s(K)$ for every $K\in\mathscr{K}_C$ with a nonempty interior.

(b) Suppose that $K\in\mathscr{K}_C$ has a *nonempty* interior. Hence the vector $\Omega\in\mathscr{H}$ is cyclic and separating for $\mathscr{B}(K)$. Let $\{V_K(\tau)|\tau\in\mathsf{R}^1\}$ be the modular automorphism group, and let J_K be the modular conjugation, for the algebra $\mathscr{B}(K)$ relative to Ω. Then the self-adjoint projection F_s commutes with $V_K(\tau)$ and with J_K, and we have

$$V_K(\tau).\mathscr{B}_s(K)V_K(\tau)^{-1}=\mathscr{B}_s(K),\quad V_K(\tau).\mathscr{A}_s(K^c)V_K(\tau)^{-1}=\mathscr{A}_s(K^c),$$

for all real τ. Furthermore,

$$J_K.\mathscr{B}_s(K)J_K=\mathscr{A}_s(K^c)^z.$$

We write $J_{sK}=\rho(J_K)$ and $V_{sK}(\tau)=\rho(V_K(\tau))$. Then $\{V_{sK}(\tau)|\tau\in\mathsf{R}^1\}$ is the modular automorphism group, and J_{sK} is the modular conjugation, for the algebra $\hat{\mathscr{B}}_s(K)$, relative to Ω.

(c) As equalities between sets, we have

$$\rho(\mathscr{A}(W))=\rho(\mathscr{A}_s(W)),\quad \rho(\mathscr{B}(K))=\rho(\mathscr{B}_s(K)),\quad \rho(\mathscr{A}(K^c))=\rho(\mathscr{A}_s(K^c)),\tag{5.8}$$

for all $W\in\mathscr{W}$ and all $K\in\mathscr{K}_C$.

(d) Let $\mathscr{S}$ be any subset of $\mathscr{U}$ such that $\mathscr{N}(\mathscr{S})=\mathscr{N}_s$, with the notation in Theorem 5.2, and let $\hat{\mathscr{S}}=\rho(\mathscr{S})$. Then $\mathscr{N}(\hat{\mathscr{S}})=\hat{\mathscr{N}}_s$.

Proof of Theorems 5.1, 5.2, and 5.3: (1) Assertions very similar to the assertions in the above theorems were proved in a paper by Wichmann[5] for the case of subnets of a local net with domain $\mathscr{R}=\{D,D^c|D\in\mathscr{D}\}\cup\mathscr{W}$. We are now concerned with the rather trivial extension of these results to the kind of extended net described in Definition 3.1. We will not repeat the detailed reasoning in the earlier paper, but instead we discuss the modifications required in the present scenario. We largely follow the notation in the quoted paper, but note that the double cones are there denoted by K, rather than by D as in the present paper, and most of the formulas thus (deceptively) have the same appearance.

(2) We first consider part (a) of Theorem 5.2. The von Neumann algebras defined by $\mathscr{B}_a(D)=(\mathscr{P}_s\cap\mathscr{B}(D))''$ for all $D\in\mathscr{D}$ are the algebras of a local net $\mathscr{N}_a$ with domain $\mathscr{D}$. Let $\mathscr{N}_e$ be the net constructed from the net $\mathscr{N}_a$ as in Proposition 3.8. By (5.6) we trivially have $\mathscr{A}_s(W)=\{\mathscr{B}_a(D)|D\in\mathscr{D},\ D\subset W\}''=\mathscr{A}_e(W)$, and we thus have $\mathscr{B}_s(K)=\mathscr{B}_e(K)$ and $\mathscr{A}_s(K^c)=\mathscr{A}_e(K^c)$ for the algebras defined by (5.2). Hence $\mathscr{N}_s=\mathscr{N}_e$ is a local net which satisfies the relation (5.3) in view of the relation (3.14). Since $\mathscr{B}_a(D)\subset\mathscr{B}(D)$, and hence $\mathscr{A}_s(W)\subset\mathscr{A}(W)$, it follows that $\mathscr{N}_s$ is a subnet of $\mathscr{N}$ which thus satisfies the premises of Theorem 5.1. The relations in (5.7) follow from Proposition 3.8.

(3) Suppose that a subnet $\mathscr{N}_s$ of a *regular* net $\mathscr{N}$ satisfies all the premise in Theorem 5.1. We want to apply Theorem 4 in Wichmann,[5] with $\mathscr{S}=\mathscr{U}_s=\cup\{\mathscr{B}_s(D)|D\in\mathscr{D}\}=\mathscr{P}_s$, and with $\mathscr{S}(K)$ in the theorem given by $\mathscr{S}(K)=\mathscr{B}_s(K)$ for all $K\in\mathscr{D}$. In view of (5.3) and (5.2) the conclusions in parts (a) and (d) of Theorem 5.1, and in part (b) of Theorem 5.2, follow immediately from the quoted theorem. The relations at left in (5.4) also follow from this theorem. We now note that the reasoning in the quoted paper depends on certain geometric properties of double cones, which also obtain for all $K\in\mathscr{K}_C$, namely that every such K is compact and is the intersection of the wedges $W\in\mathscr{W}$ which contain K, and that K^c is a union of the wedges contained in K^c. (See, in particular, Proposition 6 in the paper.) The relations at right in (5.4) are thus valid for *all* $K\in\mathscr{K}_C$, and the relations in (5.5) hold as stated. It was not assumed in the quoted paper that $\mathscr{B}(D)\Omega$ is dense in

$\mathscr{H}$ for *all* $D \in \mathscr{D}$, but our Definition 3.1 of a *regular* net incorporates this assumption, and hence $\mathscr{B}(K)\Omega$ is dense in $\mathscr{H}$, and $\mathscr{B}_s(K)\Omega$ is dense in $\mathscr{H}_s$, for every $K \in \mathscr{K}_C$ with a nonempty interior.

(4) For all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$ we define the von Neumann algebras $\hat{\mathscr{A}}_w(W)$, $\hat{\mathscr{B}}_w(K)$, and $\hat{\mathscr{A}}_w(K^c)$ of operators on $\mathscr{H}_s$ by $\hat{A}_w(W) = \{\rho(\mathscr{S}(D)) | D \in \mathscr{D},\ D \subset W\}''$ and by $\hat{\mathscr{B}}_w(K) = \cap\{\hat{A}_w(W) | W \in \mathscr{W},\ W \supset K\}$ and $\hat{\mathscr{A}}_w(K^c) = \{\hat{A}_w(W) | W \in \mathscr{W},\ W \subset K^c\}''$, with $\mathscr{S}(D) = \mathscr{B}_s(D)$ as above. Note that the algebras $\hat{\mathscr{A}}_w(W)$ are denoted by $\hat{\mathscr{A}}_s(W)$ in the quoted paper, and that the algebras $\hat{\mathscr{B}}_w(K)$ and $\hat{\mathscr{A}}_w(K^c)$ are denoted by $\hat{\mathscr{B}}_s(K)$, respectively $\hat{\mathscr{A}}_s(K^c)$, for $K \in \mathscr{D}$. This (temporary) deviation from the notation in the paper is necessary in view of the different definition of the sets $\hat{\mathscr{A}}_s(W)$, $\hat{\mathscr{B}}_s(K)$, and $\hat{\mathscr{A}}_s(K^c)$ in part (a) of Theorem 5.3. By Theorem 4 in the paper we then have $\hat{\mathscr{A}}_w(W) = \rho(\mathscr{A}_s(W))$ for $W \in \mathscr{W}$, and $\hat{\mathscr{B}}_w(K) = \rho(\mathscr{B}_s(K))$ and $\hat{\mathscr{A}}_w(K^c) = \rho(\mathscr{A}_s(K^c))$ for $K \in \mathscr{D}$. The reasoning in the paper by which the last two identities are proved is, however, valid for *all* $K \in \mathscr{K}_C$, and we thus have $\hat{\mathscr{A}}_w(W) = \hat{\mathscr{A}}_s(W)$, $\hat{\mathscr{B}}_w(K) = \hat{\mathscr{B}}_s(K)$, and $\hat{\mathscr{A}}_w(K^c) = \hat{\mathscr{A}}_s(K^c)$ for all $W \in \mathscr{W}$ and all $K \in \mathscr{K}_C$. Likewise the identities in (5.8) hold as stated. By Theorem 4 in the paper, the net $\hat{\mathscr{N}}_s$ is *PCT*-covariant and satisfies the special condition of duality, and the set $\hat{\mathscr{B}}_s(K)\Omega$ is dense in $\mathscr{H}_s$ for every $K \in \mathscr{K}_C$ with a nonempty interior. We conclude that $\hat{\mathscr{N}}_s$ is a *regular* net. By the above, and by Theorem 4 in the quoted paper, all the other conclusions in Theorem 5.3 follow readily. This completes the proof.

The subnet $\mathscr{N}_s$ in Theorem 5.1 is of the standard form (of the net $\mathscr{N}_e$) described in Proposition 3.8. The term ''regular subnet'' for such a subnet is justified by the fact that the reduced subnet $\hat{\mathscr{N}}_s$ is a *regular* net in the sense of Definition 3.1.

An examination of the reasoning in the proof of Theorems 5.1 and 5.3 shows that a key point is that a wedge-algebra $\mathscr{A}_s(W)$ of the subnet $\mathscr{N}_s$ is invariant under the modular automorphism group of the algebra $\mathscr{A}(W)$ of the original net $\mathscr{N}$. This is the case when the net $\mathscr{N}$ satisfies the *special condition of duality*, and then it follows, among other things, that the subnet is *PCT*-covariant. If we relax the condition that the net $\mathscr{N}$ satisfies the *special condition of duality*, and merely assume duality, then a particular subnet might not have the above features (as far as we know). However, if we *assume* that the subnet is such that any wedge-algebra $\mathscr{A}_s(W)$ of the subnet $\mathscr{N}_s$ is invariant under the modular automorphism group of the algebra $\mathscr{A}(W)$ of the original net $\mathscr{N}$, then we arrive at results similar to those of Theorems 5.1 and 5.3, except that nothing can be concluded about the *PCT*-covariance of the subnet.

We continue the study of regular subnets of regular nets.

Proposition 5.4: Let the premises and notation be as in Theorem 5.1. Furthermore we assume that $\mathscr{H}_s$ is a *proper* subspace of $\mathscr{H}$.

Let $\mathscr{H}_s^\perp = F^\perp \mathscr{H}$ be the orthogonal complement of $\mathscr{H}_s$, with the projection $F^\perp = I - F_s$. As in (5.1) we denote by $\gamma(X)$ the restriction of the operator $F^\perp X F^\perp$ to $\mathscr{H}_s^\perp$, for any bounded operator X on $\mathscr{H}$.

(a) Let $K \in \mathscr{K}_C$, with a *nonempty interior*. Then

$$\|X\| = \|\rho(X)\| = \|\gamma(X)\|, \tag{5.9}$$

for every $X \in \mathscr{A}_s(K^c)$. In particular these relations hold for every $X \in \mathscr{U}_s$.

(b) For every $X \in \mathscr{U}$ there exists a *unique* element $\xi(X) \in \mathscr{U}_s$ such that

$$F_s X \Omega = \xi(X)\Omega,$$

or, equivalently stated, such that $\rho(X) = \rho(\xi(X))$.

If $X \in \mathscr{B}(K)$ for some $K \in \mathscr{K}_C$, then $\xi(X) \in \mathscr{B}_s(K) \subset \mathscr{B}(K)$.

(c) As a mapping of $\mathscr{U}$ onto $\mathscr{U}_s$ the mapping ξ is a *conditional expectation*, i.e., the mapping is *linear* on $\mathscr{U}$, and it satisfies the conditions

$$\xi(I) = I, \quad \xi(X^*) = \xi(X)^*, \tag{5.10}$$

$$\xi(\xi(X)) = \xi(X), \quad \xi(X)\xi(Y) = \xi(X\xi(Y)) = \xi(\xi(X)Y), \tag{5.11}$$

$$\xi(X)=\xi(X)^*\geqslant 0, \quad \text{if } X=X^*\geqslant 0, \tag{5.12}$$

$$\|\xi(X)\|\leqslant\|X\|, \tag{5.13}$$

for all $X,Y\in\mathscr{U}$.

Furthermore, we actually have $\xi(X)>0$ if $X=X^*>0$.

(d) The mapping ξ can be extended by norm-continuity to a mapping of the C^*-algebra $\mathscr{C}=\overline{\mathscr{U}}$ *onto* the C^*-algebra $\mathscr{C}_s=\overline{\mathscr{U}}_s$, which is also a conditional expectation, i.e., the extended (linear) mapping (denoted by the same symbol ξ) also satisfies the conditions in (5.10)–(5.13). The mapping ξ commutes with the adjoint actions on $\mathscr{C}$ of θ and of $U(\lambda)$ for all $\lambda\in\hat{\mathsf{P}}$.

Proof: (1) We consider part (a), and assume that $X\in\mathscr{A}_s(K^c)$, and that $K\in\mathscr{K}_C$ has a nonempty interior. We have $\alpha(X)=\beta(X)=0$, and hence $\|X\|=\max\{\|\rho(X)\|,\|\gamma(X)\|\}$. It clearly suffices to prove the equalities in (5.9) for the special case when $X=X^*>0$, and we thus assume that X is self-adjoint and positive.

We first consider the possibility that $\|\rho(X)\|<\|\gamma(X)\|$. There then exists a self-adjoint spectral projection $E\neq 0$ of X such that $\rho(E)=0$, and hence $E\Omega=0$. Since $E\in\mathscr{A}_s(K^c)$, and since Ω is separating for $\mathscr{A}_s(K^c)$ whenever K has a nonempty interior, we have $E=0$, which is a contradiction.

(2) We next consider the possibility that $\|\rho(X)\|>\|\gamma(X)\|$. We can then find a spectral projection E_0 of X which satisfies the conditions $\rho(E_0)>0$, $\gamma(E_0)=0$. Since $E_0\in\mathscr{A}_s(K^c)$ we also have $\alpha(E_0)=0=\beta(E_0)$. Let $D\in\mathscr{D}$ be such that $D\subset K^c$, and let $Y\in\mathscr{B}(D)$ but $Y\notin\mathscr{B}_s(D)$. Such a Y certainly exists, since we assumed that $\mathscr{N}\neq\mathscr{N}_s$. By (5.8) there then exists a $Y_s\in\mathscr{B}_s(D)\subset\mathscr{B}(D)$ such that $\rho(Y)=\rho(Y_s)$. We write $Q=Y-Y_s$, and hence we have $Q\in\mathscr{B}(D)$ and $\rho(Q)=0$. With the notation in (5.1) we then have

$$E_0Q=\begin{bmatrix}0, & \rho(E_0)\alpha(Q)\\ 0, & 0\end{bmatrix}, \quad E_0Q\Omega=0.$$

Since $E_0Q\in\mathscr{A}(K^c)$, and since Ω is separating for $\mathscr{A}(K^c)$, it follows that $E_0Q=0$. Let $\mathsf{N}\subset\mathsf{M}$ be a neighborhood of $0\in\mathsf{M}$ such that $D_{(x)}=\{x+y|y\in D\}\subset K^c$ for all $x\in\mathsf{N}$. Hence $T(x)QT(x)^{-1}\in\mathscr{A}(K^c)$, and $E_0T(x)Q\Omega=0$, which implies that $E_0T(x)Q=0$, for all $x\in\mathsf{N}$. We thus have $\langle\Omega|E_0T(x)QQ^*\Omega\rangle=0$ for all $x\in\mathsf{N}$, which implies that this relation holds for *all* $x\in\mathsf{M}$. Letting x tend to infinity we obtain the relation $\|E_0\Omega\|\cdot\|Q^*\Omega\|=0$. Since $E_0\Omega\neq 0$, and since Ω is separating for $\mathscr{A}(K^c)$, we obtain $Q=0$, which entails a contradiction. The assumption that $\|\rho(X)\|>\|\gamma(X)\|$ is thus untenable, and in view of the result in step (1) we conclude that $\|\rho(X)\|=\|\gamma(X)\|$, and we have thus proved the relations in (5.9) for every $X\in\mathscr{A}_s(K^c)$.

(3) Suppose that $X\in\mathscr{U}_s$. We then have $X\in\mathscr{B}_s(D)$ for some $D\in\mathscr{D}$, and since we can always find a $K\in\mathscr{K}_C$ with a nonempty interior such that $D\subset K^c$, and hence $X\in\mathscr{A}_s(K^c)$, it follows that the relations in (5.9) hold for all $X\in\mathscr{U}_s$.

(4) We consider part (b). Let $X\in\mathscr{U}$, and hence $X\in\mathscr{B}(K)$ for some $K\in\mathscr{K}_C$. By (5.9) there then exists an $X_s\in\mathscr{B}_s(K)\subset\mathscr{U}_s$ such that $\rho(X)=\rho(X_s)$. Let $X_1\in\mathscr{U}_s$ be such that $\rho(X)=\rho(X_1)$. We then have $\rho(X_s-X_1)=0$, $X_s-X_1\in\mathscr{U}_s$, and hence $(X_s-X_1)\Omega=0$. Since Ω is separating for $\mathscr{U}_s\subset\mathscr{U}$, it follows that $X_1=X_s$, and hence $X_s=\xi(X)$ is unique.

(5) We consider part (c). The assertions in (5.10) are trivial. For any $X\in\mathscr{U}$ we have $\xi(\xi(X))\Omega=F_s\xi(X)\Omega=F_sF_sX\Omega=\xi(X)\Omega$, and since $\xi(\xi(X))$, $\xi(X)\in\mathscr{U}$, and since Ω is separating for $\mathscr{U}$, it follows that $\xi(\xi(X))=\xi(X)$.

Let $X,Y\in\mathscr{U}$. We then have $\xi(X)\xi(Y)F_s=\xi(X)F_sYF_s=F_s\xi(X)YF_s=\xi(\xi(X)Y)F_s$. Hence $\xi(X)\xi(Y)\Omega=\xi(\xi(X)Y)\Omega$, and it follows that $\xi(X)\xi(Y)=\xi(\xi(X)Y)$. Forming the adjoints of both members we arrive at the relation $\xi(X)\xi(Y)=\xi(X\xi(Y))$, for all $X,Y\in\mathscr{U}$. This proves the relations in (5.11).

Suppose now that $X\in\mathscr{U}$, and $X=X^*>0$, and that $\xi(X)=0$, which implies that $\rho(X)=0$. With the notation as in (5.1) we must then have $\alpha(X)=0=\beta(X)$, since X is non-negative and self-adjoint. This implies that $X\Omega=0$, which is impossible since Ω is separating for $\mathscr{U}$. Hence we conclude that $\xi(X)>0$ if $X=X^*>0$.

To prove the inequality in (5.13) we merely note that $\|X\| \geqslant \|\rho(X)\| = \|\rho(\xi(X))\| = \|\xi(X)\|$ for any $X \in \mathscr{U}$, since the equalities in (5.9) then hold for $\xi(X) \in \mathscr{U}_s$. The assertions in part (d) are trivial. This completes the proof.

If the algebras of the subnet $\mathscr{N}_s$ are the invariance-algebras of a *compact* group $\mathscr{G}$ of local internal symmetries, i.e., $\mathscr{B}_s(K) = \mathscr{B}(K) \cap \mathscr{G}'$, $\mathscr{A}_s(K^c) = \mathscr{A}(K^c) \cap \mathscr{G}'$, and $\mathscr{A}_s(W) = \mathscr{A}(W) \cap \mathscr{G}'$, then the conditional expectation ξ is given by averaging over the group.[25]

As we have seen, certain features of a net are *hereditary* under the premises of Theorems 5.1 and 5.2, namely *PCT*-covariance; that the net satisfies the special condition of duality, and the property that Ω is cyclic for $\mathscr{B}(K)$ for every $K \in \mathscr{K}_C$ with a nonempty interior. We next consider some other properties of a net which are also hereditary.

Proposition 5.5: Let the premises and notation be as in Theorems 5.1 and 5.3. Hence $\hat{\mathscr{N}}_s$ is a *reduced regular subnet* of a *regular* net $\mathscr{N}$ which satisfies the *special condition of duality*.

Suppose that the net $\mathscr{N}$ has the property that

$$\mathscr{B}(K) = \cap\{\mathscr{B}(K_f) | K_f \in \mathscr{F}\}, \tag{5.14}$$

for any $K \in \mathscr{K}_C$ and any subset $\mathscr{F}$ of $\mathscr{K}_C$ such that $K = \cap\{K_f | K_f \in \mathscr{F}\}$. Then the nets $\mathscr{N}_s$ and $\hat{\mathscr{N}}_s$ have the same property, i.e.,

$$\mathscr{B}_s(K) = \cap\{\mathscr{B}_s(K_f) | K_f \in \mathscr{F}\}, \quad \hat{\mathscr{B}}_s(K) = \cap\{\hat{\mathscr{B}}_s(K_f) | K_f \in \mathscr{F}\}. \tag{5.15}$$

Proof: (1) The equality at left in (5.15) follows immediately if we form the intersections of both members in (5.14) with $\mathscr{A}_s(\mathsf{M})$.

(2) Let $\hat{B} \in \cap\{\hat{\mathscr{B}}_s(K_f) | K_f \in \mathscr{F}\}$, and let $K_1, K_2 \in \mathscr{F}$. There then exist two unique operators, $B_1 \in \mathscr{B}_s(K_1)$ and $B_2 \in \mathscr{B}_s(K_2)$, such that $\rho(B_1) = \hat{B} = \rho(B_2)$. We have $(B_1 - B_2)\Omega = 0$, and since Ω is trivially separating for the set $\mathscr{B}_s(K_1) \cup \mathscr{B}_s(K_2) \subset \mathscr{U}$, it follows that $B_1 = B_2$. From this we conclude that there exists a unique operator $B \in \mathscr{B}_s(K)$ such that $\rho(B) = \hat{B}$. This implies that $\rho(\mathscr{B}_s(K)) = \hat{\mathscr{B}}_s(K) \supset \cap\{\hat{\mathscr{B}}_s(K_f) | K_f \in \mathscr{F}\}$, and since the opposite inclusion follows trivially from isotony, we arrive at the identity at right in (5.15). This completes the proof.

The intersection property presented in Theorem 4.11 is thus hereditary. In that theorem the intersection property was based on the condition of additivity for wedges. In the above it is immaterial whether this condition holds or does not hold, but it is of interest to note that *this* condition is also hereditary, as we will show in Proposition 5.7 below. In preparation for this we first consider a lemma which will also permit some other interesting conclusions.

Lemma 5.6: Let the premises and notation be as in Theorems 5.1 and 5.3. Let $\mathscr{F}$ be a subset of $\mathscr{K}_C$ such that every element $K \in \mathscr{F}$ has a *nonempty* interior K^o, and such that the set $S = \cup\{K^o | K \in \mathscr{F}\}$ is *connected*. Let $\mathscr{A}_s = \cap\{\mathscr{A}_s(K^c) | K \in \mathscr{F}\}$. Then

$$\rho(\mathscr{A}_s) = \cap\{\hat{\mathscr{A}}_s(K^c) | K \in \mathscr{F}\}. \tag{5.16}$$

Proof: (1) Let $K_a, K_b \in \mathscr{F}$. Since S is open and connected it is also path-connected, and we readily conclude that there exists a finite chain, $\{K_j | j = 1,2,\ldots,n\} \subset \mathscr{F}$ such that $K_a = K_1$, $K_b = K_n$, and $K_k^o \cap K_{k+1}^o \neq \varnothing$ for $k = 1,2,\ldots,n-1$. The vector Ω is then cyclic for the algebras $\mathscr{B}(K_k)^z \cap \mathscr{B}(K_{k+1})^z$, and hence separating for the sets $\mathscr{A}_s(K_k^c) \cup \mathscr{A}_s(K_{k+1}^c) \subset \mathscr{A}(K_k^c) \cup \mathscr{A}(K_{k+1}^c)$, for $k = 1,2,\ldots,n-1$.

(2) Let $\hat{B} \in \cap\{\hat{\mathscr{A}}_s(K^c) | K \in \mathscr{F}\}$, and hence $\hat{B} \in \hat{\mathscr{A}}_s(K_j^c)$ for $j = 1,2,\ldots,n$. For each j there then exists a unique operator $B_j \in \mathscr{A}_s(K_j^c)$ such that $\rho(B_j) = \hat{B}$. We have $(B_k - B_{k+1})\Omega = 0$, and since Ω is separating for the sets $\mathscr{A}_s(K_k^c) \cup \mathscr{A}_s(K_{k+1}^c)$ it follows that $B_j = B_1 = B_n$ for all j. From this we conclude that there exists a *unique* operator $B \in \mathscr{A}_s$ such that $\rho(B) = \hat{B}$, and hence $\hat{B} \in \rho(\mathscr{A}_s)$. This proves that $\rho(\mathscr{A}_s) \supset \cap\{\hat{\mathscr{A}}_s(K^c) | K \in \mathscr{F}\}$, and since the opposite inclusion is trivial, the equality in (5.16) follows.

Proposition 5.7: Let the premises and notation be as in Theorems 5.1 and 5.3. Hence $\hat{\mathscr{N}}_s$ is a *reduced regular subnet* of a *regular* net $\mathscr{N}$ which satisfies the *special condition of duality*. In addition we assume that the net $\mathscr{N}$ satisfies the *condition of additivity for wedges*, in accordance with Definition 2.5. Then the reduced subnet $\hat{\mathscr{N}}_s$ satisfies the same condition.

Proof: Let $W \in \mathscr{W}$, and let $\mathscr{F} \subset \mathscr{D}$ be such that $W \subset \cup\{D^o | D \in \mathscr{F}\}$. Without any essential loss of generality we can assume that $D \cap W \neq \varnothing$ for all $D \in \mathscr{F}$, and hence the set $S = \cup\{D^o | D \in \mathscr{F}\}$ is *connected*. By the condition of additivity we have $\mathscr{A}(W) \subset \{\mathscr{B}(D) | D \in \mathscr{F}\}''$. This relation is, by duality, equivalent with the relation

$$\mathscr{A}(\overline{W}^c) \supset \cap\{\mathscr{A}(D^c) | D \in \mathscr{F}\}.$$

Forming the intersections of both members with $\mathscr{A}_s(\mathsf{M})$ we obtain the relation $\mathscr{A}_s(\overline{W}^c) \supset \cap\{\mathscr{A}_s(D^c) | D \in \mathscr{F}\}$. By Lemma 5.6 we have $\rho(\mathscr{A}_s) = \cap\{\hat{\mathscr{A}}_s(D^c) | D \in \mathscr{F}\}$, with $\mathscr{A}_s = \cap\{\mathscr{A}_s(D^c) | D \in \mathscr{F}\}$, and since $\hat{\mathscr{A}}_s(\overline{W}^c) = \rho(\mathscr{A}_s(\overline{W}^c)$ it follows that

$$\hat{\mathscr{A}}_s(\overline{W}^c) \supset \cap\{\hat{\mathscr{A}}_s(D^c) | D \in \mathscr{F}\}. \tag{5.17}$$

The reduced subnet $\hat{\mathscr{N}}_s$ satisfies the condition of duality, and hence the relation (5.17) is equivalent with the relation

$$\hat{\mathscr{A}}_s(W) \subset \{\hat{\mathscr{B}}_s(D) | D \in \mathscr{F}\}''.$$

The reduced subnet $\hat{\mathscr{N}}_s$ thus satisfies the condition of additivity for wedges, as asserted.

By similar reasoning we can also show that if the net $\mathscr{N}$ satisfies the *condition of additivity for double cones*, mentioned in Section II, then the reduced subnet $\hat{\mathscr{N}}_s$ also satisfies this condition.

We next show that a certain causality condition is hereditary.

Proposition 5.8: Let the premises and notation be as in Theorems 5.1 and 5.3. Hence $\hat{\mathscr{N}}_s$ is a *reduced regular subnet* of a *regular* net $\mathscr{N}$ which satisfies the *special condition of duality*. In addition we assume that the net $\mathscr{N}$ satisfies the following particular *condition of causality:*

$$\{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset S\}'' = \{\mathscr{B}(D) | D \in \mathscr{D},\ D \subset \mathscr{E}(S)\}'', \tag{5.18}$$

for every open *connected* subset S of M. Here $\mathscr{E}(S)$ is the envelope of uniqueness of S, defined as in Lemma 4.5.

Then the net $\hat{\mathscr{N}}_s$ satisfies the same causality condition.

Proof: The relation in (5.18) is, by duality, equivalent with the relation

$$\cap\{\mathscr{A}(D^c) | D \in \mathscr{D},\ D \subset S\} = \cap\{\mathscr{A}(D^c) | D \in \mathscr{D},\ D \subset \mathscr{E}(S)\}.$$

Forming the intersections of both members with $\mathscr{A}_s(\mathsf{M})$ we obtain the relation

$$\cap\{\mathscr{A}_s(D^c) | D \in \mathscr{D},\ D \subset S\} = \cap\{\mathscr{A}_s(D^c) | D \in \mathscr{D},\ D \subset \mathscr{E}(S)\}. \tag{5.19}$$

Since the set S is connected the set $\mathscr{E}(S)$ is also connected. We have $S = \cup\{D^o | D \subset S\}$ and $\mathscr{E}(S) = \cup\{D^o | D \subset \mathscr{E}(S)\}$. By an obvious application of Lemma 5.6 to both members in (5.19) we obtain the relation

$$\cap\{\hat{\mathscr{A}}_s(D^c) | D \in \mathscr{D},\ D \subset S\} = \cap\{\hat{\mathscr{A}}_s(D^c) | D \in \mathscr{D},\ D \subset \mathscr{E}(S)\},$$

which, by duality, is equivalent with the relation

$$\{\hat{\mathscr{B}}_s(D) | D \in \mathscr{D},\ D \subset S\}'' = \{\hat{\mathscr{B}}_s(D) | D \in \mathscr{D},\ D \subset \mathscr{E}(S)\}''.$$

Hence the reduced subnet $\hat{\mathscr{N}}$ satisfies the same causality condition. This completes the proof.

The above rather strong causality condition holds for a local net $\mathscr{N}$ generated by a free Hermitian scalar field associated with a single massive particle, but it is not known whether it holds more generally. The above proposition shows that counterexamples, if there are such, cannot be manufactured within the framework of finite-component free-field theories. A variety of other weaker causality conditions can be considered, and among these we note the so-called *time-slice condition*.[9,12] (This condition says that the algebra generated by all local operators associated with

a region bounded by two parallel spacelike planes is irreducible.) By the same kind of reasoning as above we see at once that the time-slice condition is also hereditary (under the general premises in the above proposition).

The algebras $\mathscr{B}(K)$ of a regular net $\mathscr{N}$ are continuous from the *outside*, but not necessarily from the inside. However, if the algebra $\mathscr{B}(K_0)$ is continuous from the inside, for some particular $K_0 \in \mathscr{K}_C$, then this property if also hereditary.

Proposition 5.9: Let the premises and notation be as in Theorems 5.1 and 5.3. Hence $\hat{\mathscr{N}}_s$ is a *reduced regular subnet* of a *regular* net $\mathscr{N}$ which satisfies the *special condition of duality*. Suppose that the algebra $\mathscr{B}(K_0)$, for some $K_0 \in \mathscr{K}_C$ with a nonempty interior, is *continuous from the inside* in the sense that

$$\mathscr{B}(K_0) = \{\mathscr{B}(K) | K \in \mathscr{K}_C,\ K \subset K_0^o\}''. \tag{5.20}$$

Then the algebra $\hat{\mathscr{B}}_s(K_0)$ of the reduced subnet $\hat{\mathscr{N}}_s$ has the same property.

Proof: Let $\mathscr{T}_0$ be the set of all $K \in \mathscr{K}_C$ with a *nonempty interior*, and such that $K \subset K_0^o$. We obviously have $\mathscr{B}(K_0) = \{\mathscr{B}(K) | K \in \mathscr{T}_0\}''$ if the relation in (5.20) holds. By duality this implies that $\mathscr{A}(K_0^c) = \cap\{\mathscr{A}(K^c) | K \in \mathscr{T}_0\}$, and from this it follows that $\mathscr{A}_s(K_0^c) = \cap\{\mathscr{A}_s(K^c) | K \in \mathscr{T}_0\}$. By Lemma 5.6 we then conclude that $\hat{\mathscr{A}}_s(K_0^c) = \cap\{\hat{\mathscr{A}}_s(K^c) | K \in \mathscr{T}_0\}$. By duality this relation is equivalent with the relation $\hat{\mathscr{B}}_s(K_0) = \{\hat{\mathscr{B}}_s(K) | K \in \mathscr{T}_0\}''$. Since we clearly have $\hat{\mathscr{B}}_s(K_0) \supset \{\hat{\mathscr{B}}_s(K) | K \in \mathscr{K}_C,\ K \subset K_0^o\}'' \supset \{\hat{\mathscr{B}}_s(K) | K \in \mathscr{T}_0\}''$, it follows that

$$\hat{\mathscr{B}}_s(K_0) = \{\hat{\mathscr{B}}_s(K) | K \in \mathscr{K}_C,\ K \subset K_0^o\}'',$$

as asserted. This completes the proof.

The so-called *condition of nuclearity* was discussed in a paper by Buchholz and Wichmann[27] as a physically motivated additional condition on the local algebras. We will not discuss this condition here, but we remark that it is trivial that it is also hereditary.

We conclude this section by showing that there is a unique $K \in \mathscr{K}_C$ associated with every local operator, under certain premises.

Proposition 5.10: Let the premises and notation be as in Theorems 5.1 and 5.3. Hence $\hat{\mathscr{N}}_s$ is a *reduced regular subnet* of a *regular* net $\mathscr{N}$ which satisfies the *special condition of duality*. In addition we assume that the net $\mathscr{N}$ has the *intersection property* described in Proposition 5.5. In particular the net $\mathscr{N}$ has this property if it satisfies the *condition of additivity for wedges* in Definition 2.5.

Let $B \in \mathscr{U} = \cup\{\mathscr{B}(K) | K \in \mathscr{K}_C\}$. There then exists a unique *smallest* $K_B \in \mathscr{K}_C$ such that $B \in \mathscr{B}(K_B)$, i.e., $K_B \subset K$ if $B \in \mathscr{B}(K)$. Furthermore, we have $\rho(B) \in \hat{\mathscr{B}}_s(K_B)$, and if $B \in \mathscr{U}_s$, then K_B is the *smallest* $K \in \mathscr{K}_C$ such that $\rho(B) \subset \hat{\mathscr{B}}_s(K)$. In particular this applies when $\mathscr{N}_s = \mathscr{N}(\{B\})$, i.e., the subnet $\mathscr{N}_s$ is *generated* by B in the sense described in Theorem 5.2.

Proof: For any $B \in \mathscr{U}$ we define K_B by $K_B = \cap\{K | K \in \mathscr{K}_C,\ B \in \mathscr{B}(K)\}$, and K_B is then the unique smallest K such that $B \in \mathscr{B}(K)$. By part (b) of Proposition 5.4 we then have $\rho(B) = \rho(\xi(B)) \in \hat{\mathscr{B}}_s(K_B)$. Suppose now that $B \in \mathscr{U}_s$, and that $\rho(B) \in \hat{\mathscr{B}}_s(K)$. We then have $B \in \mathscr{B}_s(K) \subset \mathscr{B}(K)$, and hence $K \supset K_B$, and it follows that K_B is the smallest $K \in \mathscr{K}_C$ such that $\rho(B) \subset \hat{\mathscr{B}}_s(K)$. This completes the proof.

The case when $\mathscr{N}_s = \mathscr{N}(\{B\})$ is interesting. By the above proposition K_B is the unique smallest compact, convex, causally complete subset K of M such that $[B^z, U(\lambda) B U(\lambda)^{-1}] = 0$ and $[B^z, U(\lambda) B^* U(\lambda)^{-1}] = 0$ for all $\lambda \in \hat{\mathsf{P}}$ such that $K_{(\lambda)} \subset K^c$. The region K_B is thus an *intrinsic* feature of the local operator B in the sense that it can in principle be determined without reference to the other local operators in the net.

VI. ABOUT POSSIBLE FURTHER EXTENSIONS OF SYSTEMS OF LOCAL ALGEBRAS

Given a *regular* local net $\mathscr{N}$, a rather natural project is to try to extend the domain of the $\mathscr{B}$-algebras to the set $\mathscr{K}_B$ of *all* compact causally complete subsets of M. We want to discuss this issue briefly here. For any $K \in \mathscr{K}_B$ the set K^c is open and causally complete, and we thus also like to extend the domain of the $\mathscr{A}$-algebras to include all such K^c. The *local* algebras satisfy the

condition of *Haag duality* if and only if the extended net satisfies *both* Conditions VIII and IX in Definition 2.1, and if this is the case, then the *local* algebras are maximal. We are *not* asserting here that such an extension always exists for a given regular net $\mathcal{N}$, or that it is unique if it exists. We are merely describing a desirable goal, and we have no proof one way or the other.

As in the construction of a regular net from a basic net in Section III, the set $\mathcal{U}$ of all *local* operators of the extended net remains unchanged, for the simple reason that every $K \in \mathcal{K}_B$ is contained in some double cone. To the extended net thus corresponds a particular association of the local operators with the regions in $\mathcal{K}_B$. The set $\mathcal{K}_B$ contains compact sets of considerable geometric complexity. A set $K \in \mathcal{K}_B$ may thus be a union of any number of disjoint connected components, and any component may be multiply connected. Similarly for the regions K^c. The *convex* sets in $\mathcal{K}_C$ are very much simpler. We were able to show (assuming the condition of additivity for wedges) that to each local operator B corresponds a unique smallest $K \in \mathcal{K}_C$ such that $B \in \mathcal{B}(K)$. Without a suitable ''intersection property'' for algebras corresponding to the regions in $\mathcal{K}_B$ we cannot show that there corresponds a unique smallest $K_b \in \mathcal{K}_B$ to each local operator, and it is not obvious to us that such a correspondence exists.

We will here describe the construction of a particular extended net. The main issue is the extension of the system of *local* algebras, and we will first show that given any such an extension, the extension of the domain of the auxiliary $\mathcal{A}$-algebras is then rather trivial.

Lemma 6.1: Let $\mathcal{N}$ be a *regular* local net, as in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$ for all $W \in \mathcal{W}$ and $K \in \mathcal{K}_C$. Let $\mathcal{K}_B$ be the set of all *compact and causally complete* subsets K of M. Let $\mathcal{N}_{loc}$ be a *PCT-covariant local net*, with von Neumann algebras $\mathcal{B}_e(K)$ defined for all $K \in \mathcal{K}_B$. We thus assume that the net $\mathcal{N}_{loc}$ satisfies Conditions I–VI in Definition 2.1. Furthermore we assume that $\mathcal{B}_e(K_c) = \mathcal{B}(K_c)$ for all $K_c \in \mathcal{K}_C$.

We write $\mathcal{A}_e(K^c) = (\mathcal{B}_e(K)^z)'$ for every $K \in \mathcal{K}_B$, and we write $\mathcal{A}_e(W) = \mathcal{A}(W)$ for every $W \in \mathcal{W}$. Then:

(a) $\mathcal{B}_e(K) \subset \mathcal{A}(W)$ for any $K \in \mathcal{K}_B$, $W \in \mathcal{W}$, such that $K \subset W$, and we have

$$\mathcal{A}_e(W) = \{\mathcal{B}_e(K) | K \in \mathcal{K}_B,\ K \subset W\}'' = \cap(\mathcal{A}_e(K^c) | K \in \mathcal{K}_B,\ K^c \supset W\}, \tag{6.1}$$

for every $W \in \mathcal{W}$.

(b) The system $\mathcal{N}_e$ of the von Neumann algebras $\mathcal{A}_e(W)$, $\mathcal{B}_e(K)$, and $\mathcal{A}_e(K^c)$, defined above for all $W \in \mathcal{W}$ and all $K \in \mathcal{K}_B$, is an irreducible local net, which satisfies *all* the Conditions I–VIII in Definition 2.1 (but not necessarily Condition IX). We have

$$\mathcal{B}_e(K) = \mathcal{B}(K), \quad \mathcal{A}_e(K^c) = \mathcal{A}(K^c), \quad \text{for all } K \in \mathcal{K}_C, \tag{6.2}$$

and the net $\mathcal{N}_e$ is thus an *extension of the net* $\mathcal{N}$ since we also have $\mathcal{A}_e(W) = \mathcal{A}(W)$.

(c) The nets $\mathcal{N}_e$ and $\mathcal{N}$ have precisely the same *local* operators, i.e., we have $\cup\{\mathcal{B}_e(K) | K \in \mathcal{K}_B\} = \mathcal{U} = \cup\{\mathcal{B}(K_c) | K_c \in \mathcal{K}_C\}$.

Proof: (1) It is trivial that the system $\mathcal{N}_e$ is Poincaré- and *PCT*-covariant. Furthermore the system satisfies the conditions of duality, namely $\mathcal{B}_e(K)^z = \mathcal{A}_e(K^c)'$ for all $K \in \mathcal{K}_B$ and $\mathcal{A}_e(W)^z = \mathcal{A}_e(\overline{W^c})'$ for all $W \in \mathcal{W}$, by the definitions of the algebras $\mathcal{A}_e(K^c)$ and $\mathcal{A}_e(W)$. The relation at right in (6.2) then follows by duality from the assumed relation at left.

Let $K \subset W$, with $K \in \mathcal{K}_B$ and $W \in \mathcal{W}$. Since K is compact there certainly exists a $K_c \in \mathcal{K}_C$ such that $K \subset K_c \subset W$. In particular we can select K_c as a double cone in $\mathcal{D}$. By the condition at left in (6.2), and by isotony, we then have $\mathcal{B}_e(K) \subset \mathcal{B}_e(K_c) = \mathcal{B}(K_c) \subset \mathcal{A}(W)$. From this it follows that $\mathcal{A}(W) \supset \{\mathcal{B}_e(K) | K \in \mathcal{K}_B,\ K \subset W\}'' \supset \{\mathcal{B}(K) | K \in \mathcal{K}_C,\ K \subset W\}'' = \mathcal{A}(W)$. This proves the equality of the first two members in (6.1). The equality of the first and third members then follows by duality. The remaining assertions are easily proved, and we omit the details. Note that two regions K_1^c and K_2^c can never be spacelike separated if $K_1, K_2 \in \mathcal{K}_B$, and hence the condition of locality does not apply to the pair of algebras $\mathcal{A}_e(K_1^c)$ and $\mathcal{A}_e(K_2^c)$. This completes the proof.

Condition IX in Definition 2.1 is at least *partly* satisfied, in view of the relations in (6.1). It is also easily seen that we have $\mathcal{A}_e(K^c) = \{\mathcal{B}_e(K_i) | K_i \in \mathcal{K}_B,\ K_i \subset K^c\}''$ for all $K \in \mathcal{K}_C$, but it might well happen that this relation does not hold for *all* $K \in \mathcal{K}_B$. As we said before, the goal is to find an extension such that this relation does hold for all $K \in \mathcal{K}_B$, which means that the net $\mathcal{N}_e$ satisfies *all* the Conditions I–IX.

We have arranged it so that the net $\mathcal{N}_e$ satisfies the duality condition. We could also have defined the $\mathcal{A}_e$-algebras by $\mathcal{A}_e(K^c)=\{\mathcal{B}_e(K_i)|K_i\in\mathcal{K}_B,\ K_i\subset K^c\}''$, in which case Condition IX would hold, but then our duality condition might fail. If the net $\mathcal{N}$ has no extension which satisfies all the Conditions I–IX, it becomes a matter of choice of how the auxiliary $\mathcal{A}_e$-algebras are defined. One might feel that an *essential* feature of duality is that Condition IX should also hold, and that the duality of the net $\mathcal{N}_e$ is of a ''cheap'' kind if the extension does not satisfy this condition. However, the net $\mathcal{N}_e$ does satisfy this condition for wedges, and its restriction to the domain of $\mathcal{N}$ is a net which does satisfy *all* the Conditions I–IX. These features of the net $\mathcal{N}_e$ are *not* ''cheap.''

We will now discuss a simple extension of a regular net. The strategy is to first construct algebras $\mathcal{B}_S(K)$ for all $K\in\mathcal{K}_B$, which are as large as we can make them, after which we apply Lemma 6.1 to complete the extension. A natural first step is to define a von Neumann algebra $\mathcal{B}_0(K)$ by $\mathcal{B}_0(K)=\{\mathcal{B}(K_i)|K_i\in\mathcal{K}_C,\ K_i\subset K\}''$ for every $K\in\mathcal{K}_B$. These algebras form a local net with domain $\mathcal{K}_B$, but it is possible that this net is not maximal. There is a well-known standard construction which may lead to an enlargement of these algebras, which we now describe in general terms.

Lemma 6.2: Let $\mathcal{M}$ be any mapping of $\mathcal{K}_B$ into the set of von Neumann algebras in $\mathsf{B}(\mathcal{H})$. We define the *dual mapping* $\mathcal{M}^\star$ of $\mathcal{K}_B$ into the set of von Neumann algebras by

$$\mathcal{M}^\star(K)=\{\mathcal{M}(K_i)^z|K_i\in\mathcal{K}_B,\ K_i\subset K^c\}', \tag{6.3}$$

for all $K\in\mathcal{K}_B$, and we define $\mathcal{M}^{\star\star}$ as $(\mathcal{M}^\star)^\star$ and $\mathcal{M}^{\star\star\star}$ as $(\mathcal{M}^{\star\star})^\star=(\mathcal{M}^\star)^{\star\star}$. Then:

(a) For all $K\in\mathcal{K}_B$,

$$\mathcal{M}^{\star\star}(K)\supset\mathcal{M}(K),\quad \mathcal{M}^\star(K)=\mathcal{M}^{\star\star\star}(K). \tag{6.4}$$

(b) If $\mathcal{M}$ satisfies the conditions of isotony, Poincaré-covariance, and *PCT*-covariance, then $\mathcal{M}^\star$ and $\mathcal{M}^{\star\star}$ satisfy the same conditions.

(c) If $\mathcal{M}$ is a local net with domain $\mathcal{K}_B$, then $\mathcal{M}^{\star\star}$ is also a local net with domain $\mathcal{K}_B$. Furthermore we have $\mathcal{M}^\star(K)\supset\mathcal{M}^{\star\star}(K)$ for all $K\in\mathcal{K}_B$.

Proof: (1) The inclusion relation $\mathcal{M}^{\star\star}(K)\supset\mathcal{M}(K)$ follows trivially from the definition in (6.3). Replacing $\mathcal{M}$ by $\mathcal{M}^\star$ we have $\mathcal{M}^{\star\star\star}(K)\supset\mathcal{M}^\star(K)$. Let $\mathcal{M}_l$ and $\mathcal{M}_s$ be two mappings of the stated kind, such that $\mathcal{M}_l(K)\supset\mathcal{M}_s(K)$ for all $K\in\mathcal{K}_B$. By (6.3) we then have $\mathcal{M}_l^\star(K)\subset\mathcal{M}_s^\star(K)$ for all $K\in\mathcal{K}_B$. Applying this to the relation $\mathcal{M}^{\star\star}(K)\supset\mathcal{M}(K)$ we obtain $\mathcal{M}^{\star\star\star}(K)\subset\mathcal{M}^\star(K)$, and hence the relation at right in (6.4) follows.

(2) The assertions in part (b) follow trivially from the fact that causal complementation commutes with all Poincaré transformations, and with inversion, and from the fact that $R_l^c\subset R_s^c$ whenever $R_l\supset R_s$.

(3) We consider part (c). The locality condition for $\mathcal{M}$ trivially implies that $\mathcal{M}^\star(K)\supset\mathcal{M}(K)$, in view of the definition in (6.3). By (6.4) it then follows that $\mathcal{M}^\star(K)\supset\mathcal{M}^{\star\star}(K)$. In view of (6.3) this implies that $\mathcal{M}^{\star\star}$ is a local net. This completes the proof.

In the present context *Haag duality* for a local net $\mathcal{M}$ is the condition that $\mathcal{M}^\star=\mathcal{M}$. The weaker condition that $\mathcal{M}^\star=\mathcal{M}^{\star\star}$, which is the condition that $\mathcal{M}^\star$ is local, is referred to as *essential duality* in the literature.

We are now prepared for the specific extension which we want to discuss.

Theorem 6.3: Let $\mathcal{N}$ be a *regular* local net, as in Definition 3.1, with algebras $\mathcal{A}(W)$, $\mathcal{B}(K)$, and $\mathcal{A}(K^c)$ for all $W\in\mathcal{W}$ and $K\in\mathcal{K}_C$. For each $K\in\mathcal{K}_B$ we define the von Neumann algebra $\mathcal{B}_0(K)$ by $\mathcal{B}_0(K)=\{\mathcal{B}(K_i)|K_i\in\mathcal{K}_C,\ K_i\subset K\}''$. We define the mappings $\{\mathcal{B}_0(K)\}\to\{\mathcal{B}_0^\star(K)\}$ and $\{\mathcal{B}_0(K)\}\to\{\mathcal{B}_0^{\star\star}(K)\}$ as in Lemma 6.2.

We write $\mathcal{A}_S(W)=\mathcal{A}(W)$ for $W\in\mathcal{W}$, and for each $K\in\mathcal{K}_B$ we define the von Neumann algebras $\mathcal{B}_S(K)$ and $\mathcal{A}_S(K^c)$ by $\mathcal{B}_S(K)=\mathcal{B}_0^{\star\star}(K)$ and $\mathcal{A}_S(K^c)=(\mathcal{B}_S(K)^z)'$. Then:

(a) The set of algebras $\mathcal{A}_S(W)$, for $W\in\mathcal{W}$, and $\mathcal{B}_S(K)$ and $\mathcal{A}_S(K^c)$, for $K\in\mathcal{K}_B$, forms a local net $\mathcal{N}_S$ which satisfies all the Conditions I–VIII in Definition 2.1, but possibly *not* Condition IX. It is an extension of the net $\mathcal{N}$, and we have

$$\mathscr{B}_S(K)=\mathscr{B}(K), \quad \mathscr{A}_S(K^c)=\mathscr{A}(K^c), \quad \text{for all } K\in\mathscr{K}_C,$$

and

$$\mathscr{A}_S(W)=\{\mathscr{B}_S(K)|K\in\mathscr{K}_B,\ K\subset W\}''=\cap(\mathscr{A}_S(K^c)|K\in\mathscr{K}_B,\ K^c\supset W\}.$$

The nets $\mathscr{N}$ and $\mathscr{N}_S$ have the same set of local operators, i.e.,

$$\mathscr{U}_S=\cup\{\mathscr{B}_S(K)|K\in\mathscr{K}_B\}=\mathscr{U}=\cup\{\mathscr{B}(K)|K\in\mathscr{K}_C\}.$$

(b) Let $\mathscr{N}_m$ be a local net, with von Neumann algebras $\mathscr{A}_m(W)=\mathscr{A}(W)$, $\mathscr{B}_m(K)$, and $A_m(K^c)$ for all $W\in\mathscr{W}$ and all $K\in\mathscr{K}_B$. Suppose that $\mathscr{N}_m$ is an extension of $\mathscr{N}$ such that $\mathscr{B}_m(K)=\mathscr{B}(K)$ and $\mathscr{A}_m(K^c)=\mathscr{A}(K^c)$ for all $K\in\mathscr{K}_C$, and suppose furthermore that $\mathscr{B}_m^{\star\star}(K)=\mathscr{B}_m(K)$ for all $K\in\mathscr{K}_B$. We then have

$$\mathscr{B}_0^{\star}(K)=\mathscr{B}_S^{\star}(K)\supset\mathscr{B}_m(K)\supset\mathscr{B}_S(K), \tag{6.5}$$

for all $K\in\mathscr{K}_B$.

(c) The net $\mathscr{N}_S$ satisfies Condition IX in Definition 2.1, i.e.,

$$\mathscr{A}_S(K^c)=\{\mathscr{B}_S(K_i)|K_i\in\mathscr{K}_B,\ K_i\subset K^c\}'', \tag{6.6}$$

for all $K\in\mathscr{K}_B$, if and only if $\mathscr{B}_0^{\star}(K)=\mathscr{B}_S(K)$ for all $K\in\mathscr{K}_B$, which is the case if and only if the set of algebras $\{\mathscr{B}_0^{\star}(K)|K\in\mathscr{K}_B\}$ is a *local net* with domain $\mathscr{K}_B$. If these conditions hold, then $\mathscr{N}_S$ is the *unique* extension of $\mathscr{N}$ which satisfies *all* the Conditions I–IX in Definition 2.1, and the *local* algebras $\mathscr{B}_S(K)$ are *maximal*.

(d) The mapping $K\to\mathscr{B}_0^{\star}(K)$ satisfies the conditions of isotony, Poincaré-covariance and *PCT*-covariance. We have $\mathscr{B}_0^{\star}(K)\subset\mathscr{U}$ for all $K\in\mathscr{K}_B$, and $\mathscr{B}_0^{\star}(K_c)=\mathscr{B}(K_c)$ for all $K_c\in\mathscr{K}_C$. Furthermore we have $\mathscr{B}_0^{\star}(K)\subset\mathscr{A}(W)$ for any $K\in\mathscr{K}_B$ and $W\in\mathscr{W}$ such that $K\subset W$.

Proof: (1) It follows trivially from the definition of the algebras $\mathscr{B}_0(K)$ that the mapping $K\to\mathscr{B}_0(K)$ is a local net with domain $\mathscr{K}_B$, which satisfies the condition that $\mathscr{B}_0(K_c)=\mathscr{B}(K_c)$ for all $K_c\in\mathscr{K}_C$. The conclusions in part (a) then follow readily on the basis of Lemmas 6.1 and 6.2.

(2) By part (c) of Lemma 6.2 we have $\mathscr{B}_0^{\star}(K)\supset\mathscr{B}_0^{\star\star}(K)\supset\mathscr{B}_0(K)$ for all $K\in\mathscr{K}_B$, and in particular for all $K\in\mathscr{K}_C$. Let $K_c\in\mathscr{K}_C$. Since $\mathscr{K}_C\subset\mathscr{K}_B$ it follows from (6.3) that we have $\mathscr{B}_0^{\star}(K^c)\subset\{\mathscr{B}(K_i)^z|K_i\in\mathscr{K}_C,\ K_i\subset K_c^c\}'=(\mathscr{A}(K_c^c)^z)'=\mathscr{B}(K_c)$, and hence we have $\mathscr{B}_0^{\star}(K_c)=\mathscr{B}(K_c)=\mathscr{B}_0(K_c)$ for all $K_c\in\mathscr{K}_C$. From this we conclude that $\mathscr{B}_0^{\star}(K_c)=\mathscr{B}_0^{\star\star}(K_c)=\mathscr{B}(K_c)$ for all $K_c\in\mathscr{K}_C$. Let $K\in\mathscr{K}_B$ and $W\in\mathscr{W}$ be such that $K\subset W$. There then obviously exists a $K_c\in\mathscr{K}_C$ such that $K\subset K_c\subset W$, and by the isotony of the mapping $\mathscr{B}_0^{\star}$ it follows that $\mathscr{B}_0^{\star}(K)\subset\mathscr{B}(K_c)\subset\mathscr{A}(W)$. Since every $K\in\mathscr{K}_B$ is contained in some $K_c\in\mathscr{K}_C$, we also conclude that $\mathscr{B}_0^{\star}(K)\subset\mathscr{U}$ for all $K\in\mathscr{K}_B$. We have thus proved the assertions in part (d).

(3) We consider part (b). By the isotony of the extended net $\mathscr{N}_m$ we have $\mathscr{B}_m(K)\supset\mathscr{B}_0(K)$, and hence $\mathscr{B}_m^{\star\star}\supset\mathscr{B}_0^{\star\star}=\mathscr{B}_S(K)$ and $\mathscr{B}_0^{\star}(K)\supset\mathscr{B}_m^{\star}(K)\supset\mathscr{B}_m^{\star\star}(K)$ for all $K\in\mathscr{K}_B$. Since it was also assumed that $\mathscr{B}_m(K)=\mathscr{B}_m^{\star\star}(K)$ we obtain the inclusion relations in (6.5).

(4) We consider part (c). We have $\mathscr{B}_S^{\star}(K)=\mathscr{B}_0^{\star\star\star}(K)=\mathscr{B}_0^{\star}(K)$, in view of Lemma 6.2. By the duality condition for the net $\mathscr{N}_S$ the relation (6.6) is equivalent with the relation $\mathscr{B}_S(K)=\{\mathscr{B}_S(K_i)^z|K_i\in\mathscr{K}_B,\ K_i\subset K^c\}'$, and with reference to (6.3) we conclude that the relation (6.6) holds for all $K\in\mathscr{K}_B$ if and only if $\mathscr{B}_0^{*}(K)=\mathscr{B}_S(K)=\mathscr{B}_0^{\star\star}(K)$ for all $K\in\mathscr{K}_B$. The latter relation is the condition that the mapping $K\to\mathscr{B}_0^{\star}(K)$ is a local net. If these relations hold, the stated uniqueness of the extension $\mathscr{N}_S$ then follows from the inclusion relations in (6.5), and it is trivial that the local algebras are maximal. This completes the proof.

As we see, the algebras $\mathscr{B}_0^{\star}(K)$ do consist of *local* operators in the ''original'' set $\mathscr{U}$ of local operators, but the conditions of locality involving two $\mathscr{B}_0^{\star}$-algebras might not hold for *all* relevant configurations of causally disjoint sets in $\mathscr{K}_B$, although they obviously hold for some configurations since we have $\mathscr{B}_0^{\star}(K_c)=\mathscr{B}(K_c)$ for all $K_c\in\mathscr{K}_C$. We note here that the mapping $K\to\mathscr{B}_0^{\star}(K)$ constructed from a regular net $\mathscr{N}$ associated with a *free* massive scalar field is indeed a local net with domain $\mathscr{K}_B$, and in this case $\mathscr{N}_S$ is the unique extension of $\mathscr{N}$, which satisfies *all* the Conditions I–XI in Definition 2.1.

We note that an extended net $\mathscr{N}_m^{\cdot}$ for which the local algebras $\mathscr{B}_m(K)$ are maximal trivially satisfies the premises in part (b), and this is in particular the case if the net $\mathscr{N}_m^{\cdot}$ satisfies all the Conditions I–IX. The algebras $\mathscr{B}_m(K)$ are then always pinned between the algebras $\mathscr{B}_0^{\star}(K)$ and $\mathscr{B}_S(K)$. As we said before, we have no proof that a unique extension exists which satisfies *all* the Conditions I–IX in Definition 2.1, and we tend to believe that the question cannot be decided without additional conditions on the net $\mathscr{N}^{\cdot}$. The extension $\mathscr{N}_S^{\cdot}$ described above can, in any case, be regarded as a rather natural *standard* extension, already implicit for a *basic* net as described in Definition 2.3.

It is worth noting that the construction of the extended local net $\mathscr{N}_S^{\cdot}$ depends essentially on the circumstance that the $\mathscr{B}$-algebras are defined for *compact* subsets of M. If we try to apply the above scheme to define the algebras $\mathscr{B}_S(K)$ for *all* causally complete closed sets K, it could happen that the algebras $\mathscr{A}_S(K^c)=(\mathscr{B}_S(K)^z)'$ do not satisfy *all* the conditions of locality among themselves. This issue does not arise in our case, since K_1^c and K_2^c cannot be spacelike separated if K_1 and K_2 are compact.

In Section IV we discussed various ''intersection properties'' for a *regular* net. The question arises whether similar properties can be established for an extended local net $\mathscr{N}_e^{\cdot}$ of the kind considered in this section. The results in Proposition 4.4 can indeed be generalized. Suppose that $K_1, K_0 \in \mathscr{K}_B$ are such that $K_1 \subset K_0^o$. By the same reasoning as in the proof of Proposition 4.4, based on Borchers' theorem,[23] we easily conclude that $\mathscr{B}_e(K_1) \cap \mathscr{A}_e(K_0^c) = \mathsf{C}^1 \cdot I$. This implies in particular that $\mathscr{B}_e(K_1) \cap \mathscr{B}_e(K_2) = \mathsf{C}^1 \cdot I$ if $K_1, K_2 \in \mathscr{K}_B$ are *spacelike* separated.

We also proved other ''intersection properties,'' as in Theorem 4.11, for which the convexity of the regions in $\mathscr{K}_E$ played an important role. We are inclined to believe that these results cannot be generalized, without additional conditions on the net $\mathscr{N}_e^{\cdot}$, beyond the basic assumptions in this paper. Let $K_1, K_2 \in \mathscr{K}_B$. The relation

$$\mathscr{B}_e(K_1) \cap \mathscr{B}_e(K_2) = \mathscr{B}_e(K_1 \cap K_2) \tag{6.7}$$

is, by duality, equivalent with the relation

$$(\mathscr{A}_e(K_1^c) \cup \mathscr{A}_e(K_2^c))'' = \mathscr{A}_e((K_1 \cap K_2)^c). \tag{6.8}$$

Suppose now that the extended net $\mathscr{N}_e^{\cdot}$ satisfies *all* the Conditions I–IX in Definition 2.1, the *condition of additivity*, as well as the particular *causality condition* that

$$\{\mathscr{B}_e(K) | K \in \mathscr{K}_B,\ K \subset S\}'' = \{\mathscr{B}_e(K) | K \in \mathscr{K}_B,\ K \subset \mathscr{E}(S)\}'', \tag{6.9}$$

for any open set S. For certain sets $K_1, K_2 \in \mathscr{K}_B$ it is the case that $(K_1 \cap K_2)^c = (K_1^c \cup K_2^c)^{cc} = \mathscr{E}(K_1^c \cup K_2^c)$, and if this is the case, then the relations in (6.8) and (6.7) follow from (6.9), i.e., the intersection property holds for the pair K_1, K_2. However, the intersection property in (6.7) *cannot* hold generally, for two *arbitrary* $K_1, K_2 \in \mathscr{K}_B$. For a counterexample within the theory of a free Hermitian scalar field $\varphi(x)$ associated with a massive particle, we refer to Example 5.13 in TW-II.[7] In the quoted paper the objects of interest are the *weak commutants* $\mathscr{C}(K)$ of the field operators in K^c, rather than the local von Neumann algebras, but for a free scalar field we do have $\mathscr{B}_e(K) = \mathscr{C}(K)$, as we noted earlier. The counterexample involves two particular causally complete open sets S_1 and S_2 such that $K_1 = S_1^c \in \mathscr{K}_B$, and $K_2 = S_2^c \in \mathscr{K}_B$, and such that $\mathscr{E}(S_1 \cup S_2) \neq (S_1 \cup S_2)^{cc}$. The condition in (6.9) holds for a free-field theory, but the relation in (6.7) fails to hold. This involves no contradiction to what was said above, since we have $(K_1^c \cup K_2^c)^{cc} \neq \mathscr{E}(K_1^c \cup K_2^c)$ in the counterexample.

Within the theory of a free scalar field associated with a *massless* particle one can also easily construct examples of two *disjoint* sets $K_1, K_2 \in \mathscr{K}_B$, such that $\mathscr{B}_e(K_1)$ and $\mathscr{B}_e(K_2)$ have elements in common which are not multiples of the identity. The sets K_1 and K_2 can then not be spacelike separated.

We will not pursue these issues further, since the condition in (6.9) is outside the scope of our paper.

[1] R. Haag and D. Kastler, ''An algebraic approach to quantum field theory,'' J. Math. Phys. **5**, 848–861 (1964).
[2] R. Haag, *Local Quantum Physics* (Springer-Verlag, Berlin, 1992).

[3] S. S. Horuzhy, *Introduction to Algebraic Quantum Field Theory* (Kluwer Academic, Dordrecht, 1990).
[4] J. J. Bisognano and E. H. Wichmann, ''On the Duality Condition for Quantum Fields,'' J. Math. Phys. **17**, 303–321 (1976).
[5] E. H. Wichmann, ''On systems of local operators and the duality condition,'' J. Math. Phys. **24**, 1633–1644 (1983).
[6] L. J. Thomas and E. H. Wichmann, ''On the Causal Structure of Minkowski Spacetime,'' J. Math. Phys. **38**, 5044–5086 (1997).
[7] L. J. Thomas and E. H. Wichmann, ''On a Class of Distributions of Interest in Quantum Field Theory,'' J. Math. Phys. **39**, 1680–1719 (1998).
[8] L. J. Thomas, ''About the Geometry of Minkowski Spacetime and Systems of Local von Neumann Algebras,'' Ph.D. thesis, University of California, Berkeley, 1989.
[9] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Addison-Wesley, Reading, MA, 1989).
[10] R. Jost, *The General Theory of Quantized Fields* (American Mathematical Society, Providence, RI, 1965).
[11] N. N. Bogolubov, A. A. Logunov, and I. T. Todorov, *Introduction to Axiomatic Quantum Field Theory* (Benjamin Cummings, Reading, MA, 1975).
[12] R. Haag and B. Schroer, ''Postulates of Quantum Field Theory,'' J. Math. Phys. **3**, 248–256 (1962).
[13] M. Takesaki, *Tomita's Theory of Modular Hilbert Algebras and Its Applications* (Springer-Verlag, Berlin, 1970).
[14] H.-J. Borchers, ''When does Lorentz Invariance imply Wedge-Duality?'' Lett. Math. Phys. **35**, 39–60 (1995).
[15] H. J. Borchers, ''Half-sided Modular Inclusion and the Construction of the Poincaré Group,'' Commun. Math. Phys. **179**, 703–723 (1996).
[16] R. Brunetti, D. Guido, and R. Longo, ''Group cohomology, modular theory and space–time symmetries,'' Rev. Mod. Phys. **7**, 57–71 (1994).
[17] W. Driessler, S. J. Summers, and E. H. Wichmann, ''On the Connection Between Quantum Fields and von Neumann Algebras of Local Operators,'' Commun. Math. Phys. **105**, 49–84 (1986).
[18] H. J. Borchers and J. Yngvason, ''From Quantum Fields to Local von Neumann Algebras,'' Rev. Math. Phys. Special Issue, 15–47 (1992).
[19] H. J. Borchers and J. Yngvason, ''Transitivity of locality and duality in quantum field theory. Some modular aspects,'' Rev. Math. Phys. **6**, 597–619 (1994).
[20] S. J. Summers and E. H. Wichmann, ''Concerning the Condition of Additivity in Quantum Field Theory,'' Ann. Inst. Henri Poincaré **47**, 113–124 (1987).
[21] W. Driessler and S. J. Summers, ''Nonexistence of Quantum Fields Associated with Two-Dimensional Spacelike Planes,'' Commun. Math. Phys. **89**, 221–226 (1983).
[22] H. Araki, ''A Generalization of Borchers' Theorem,'' Helv. Phys. Acta **36**, 132–139 (1963).
[23] H. J. Borchers, ''A remark on a Theorem of B. Misra,'' Commun. Math. Phys. **4**, 315–323 (1967).
[24] L. J. Landau, ''A note on extended locality,'' Commun. Math. Phys. **135**, 246–253 (1969).
[25] S. Doplicher, R. Haag, and J. E. Roberts, ''Fields, observables and gauge transformations. I,'' Commun. Math. Phys. **13**, 1–23 (1969); ''Fields, observables and gauge transformations. II,'' *ibid.* **15**, 173–200 (1969).
[26] H. Araki, ''Symmetries in theory of local observables and the choice of the net of local algebras,'' Rev. Math. Phys. Special Issue, 1–14 (1992).
[27] D. Buchholz and E. H. Wichmann, ''Causal Independence and the Energy-Level Density of States in Local Quantum Field Theory,'' Commun. Math. Phys. **106**, 321–344 (1986).

Hyperfinite-dimensional representations of canonical commutation relation

Hideyasu Yamashita
Graduate School of Human Informatics, Nagoya University, Nagoya 464-01, Japan

(Received 11 June 1997; accepted for publication 28 October 1997)

This paper presents some methods of representing canonical commutation relations in terms of hyperfinite-dimensional matrices, which are constructed by nonstandard analysis. The first method uses representations of a nonstandard extension of the finite Heisenberg group, called hyperfinite Heisenberg group. The second is based on hyperfinite-dimensional representations of $so(3)$. Then, the cases of infinite degree of freedom are argued in terms of the algebra of hyperfinite para-Fermi oscillators, which is mathematically equivalent to a hyperfinite-dimensional representation of $so(n)$. © *1998 American Institute of Physics.*
[S0022-2488(98)02205-1]

I. INTRODUCTION

In the field of mathematical foundations of quantum physics, especially of infinite degrees of freedom, it is often argued that we need any extensions of separable Hilbert space as the fundamental mathematical tools, such as the linear space of generalized functions.

In recent years, another alternative, the *nonstandard-analytical extension* method, is developed by some authors.[1–4] The present paper adopts an approach in use of hyperfinite-dimensional linear spaces, constructed by nonstandard analysis[5,6] (see the Appendix to find the basic definitions).

This approach also is closely related with that of *finite-dimensional quantum physics*;[7–10] nonstandard extensions of finite-dimensional models in quantum physics turn out to be the powerful tools for the approach.

The purpose of the present paper is to represent the canonical commutation relation (CCR) $[Q,P]=i$ (let us take the natural unit $\hbar=1$), or equivalently, $[a,a^{\dagger}]=1$, where $a=(Q+iP)/\sqrt{2}$, in a hyperfinite-dimensional linear space. Let ν be an infinite hypernatural number, and Q and P internal $\nu\times\nu$ Hermitian matrices. Q and P act on the ν-dimensional internal linear space ${}^{\star}\mathbf{C}^{\nu}$ as internal linear operators. No standard finite-dimensional matrices satisfy the CCR. By the transfer principle in nonstandard analysis, the fact also holds for any internal hyperfinite-dimensional matrices, so we have $[Q,P]\neq i$. Thus we must arrange the standard CCR for hyperfinite-dimensional matrices. First we consider the following weaker version of CCR: $[Q,P]\approx i$, that is, $\|[Q,P]-i\|_{\infty}\approx 0$, where $\|A\|_{\infty}$ is given by

$$\|A\|_{\infty}=\sup\{\|A\xi\|:\|\xi\|\leqslant 1,\xi\in{}^{\star}\mathbf{C}^{\nu}\}.$$

This is equivalent to the condition that for any $\xi\in{}^{\star}\mathbf{C}^{\nu}$ with finite norm, $[Q,P]\xi\approx i\xi$ (i.e., $\|([Q,P]-i)\xi\|\approx 0$, where $\|\cdot\|$ is the norm on ${}^{\star}\mathbf{C}^{\nu}$ defined by $\|(\xi_1,\ldots,\xi_{\nu})\|=\sqrt{|\xi_1|^2+\cdots+|\xi_{\nu}|^2}$. We see that the condition never holds as follows. Suppose the condition holds. Then any eigenvalue λ of $[Q,P]$ satisfies $\lambda\approx i$. This contradicts $\mathrm{Tr}([Q,P])=0$. Now, let us define the still weaker notion of representation of the CCR by limiting the domain of the operators Q and P.

Definition 1.1: Let Q_k, P_k be $\nu\times\nu$ internal Hermitian matrices ($k=1,2,\ldots<\infty$), and S be an external subspace of ${}^{\star}\mathbf{C}^{\nu}$ which consists only of the vectors with finite norm. The set of the triples (Q_k,P_k,S) (or of pairs (a_k,S) where $a_k=(Q_k+iP_k)/\sqrt{2}$) is called a *quasirepresentation of the CCR* if

$$[Q_k,P_l]\xi\approx i\delta_{kl}\xi,\quad [Q_k,Q_l]\xi\approx[P_k,P_l]\xi\approx 0$$

 © 1998 American Institute of Physics

for any $\xi \in S$ and $k,l=1,2,...<\infty$. A quasirepresentation of the CCR (Q_k,P_k,S) is called a *representation of CCR* if S is invariant respect to Q_k, P_k (i.e., $Q_kS,P_kS\subseteq S$).

Hereafter, we will use the following notations:

(1) $\star$ is the mapping of nonstandard extension.
(2) $\mathbf{N}(^\star\mathbf{N})$ is the set of natural (hypernatural) numbers.
(3) $\mathbf{Z}(^\star\mathbf{Z})$ is the set of integers (hyperintegers).
(4) $\mathbf{R}(^\star\mathbf{R})$ is the set of real (hyperreal) numbers.
(5) $\mathbf{C}(^\star\mathbf{C})$ is the set of complex (hypercomplex) numbers.

II. $so(3)$-REPRESENTATION OF THE CCR

Let J_1, J_2, and J_3 be an irreducible set of $(p+1)\times(p+1)$ Hermitian matrices which satisfy

$$[J_k,J_l]=i\epsilon_{klm}J_m,$$

where ϵ_{klm} is the Levi-Civita symbol. iJ_1,iJ_2, and iJ_3 generate a $p+1$ dimensional irreducible representation of Lie algebra $so(3)$. It is known that J_k has eigenvalues

$$\frac{p}{2},\quad \frac{p}{2}-1,...,\quad 1-\frac{p}{2},\quad -\frac{p}{2}$$

and corresponding eigenvectors are given by

$$\left|J_3;\frac{p}{2}\right\rangle,\quad J_-\left|J_3;\frac{p}{2}\right\rangle,...,\quad J_-^p\left|J_3;\frac{p}{2}\right\rangle,$$

where $|J_3;(p/2)\rangle$ is a normalized eigenvector with eigenvalue $p/2$ and $J_-=J_1-iJ_2$. Let $j=p/2$ and $|J_3;m\rangle=J_-^{j-m}|J_3;j\rangle/\|J_-^{j-m}|J_3;j\rangle\|$. It is shown that

$$\langle J_3;m'|J_-|J_3;m\rangle=\delta_{m',m-1}\sqrt{(j+m)(j-m+1)}.$$

We consider the nonstandard version of this representation; let p be an infinite hypernatural number, and J_k be the internal matrices satisfying the above commutation relations. Define Q and P by $Q=J_1/\sqrt{j}$, $P=J_2/\sqrt{j}$.

Theorem 2.1: Let $S\subset{}^\star\mathbf{C}^\nu$ be the subspace finitely spanned by $|J_3;j-k\rangle$, $(k=0,1,...<\infty)$, i.e.,

$$S=\left\{\sum_{k=0}^{n}c_k|J_3;j-k\rangle:\quad c_k\in{}^\star\mathbf{C},|c_k|<\infty,n\in\mathbf{N}\right\}.$$

S is an invariant subspace of P and Q, and

$$[Q,P]|\xi\rangle\approx i|\xi\rangle$$

for each $|\xi\rangle\in S$, and hence (Q,P,S) is a representation of the CCR.

Proof: The invariance of S is seen from

$$\begin{aligned}Q|J_3;j-k\rangle&=\frac{1}{\sqrt{j}}\frac{1}{2}(J_-+J_-^\dagger)|J_3;j-k\rangle\\&=\frac{1}{2}\sqrt{(2-k/j)(k+1)}|J_3;j-k-1\rangle+\frac{1}{2}\sqrt{k(2-k/j+1/j)}|J_3;j-k+1\rangle\in S,\end{aligned}$$

$$\begin{aligned}P|J_3;j-k\rangle&=\frac{1}{\sqrt{j}}\frac{i}{2}(J_-^\dagger-J_-)|J_3;j-k\rangle\\&=\frac{i}{2}\sqrt{(2-k/j)(k+1)}|J_3;j-k-1\rangle-\frac{i}{2}\sqrt{k(2-k/j+1/j)}|J_3;j-k+1\rangle\in S.\end{aligned}$$

And we have $[Q,P]|J_3;j-k\rangle=(i/j)J_3|J_3;j-k\rangle=i(1-k/j)|J_3;j-k\rangle\approx i|J_3;j-k\rangle$. Q.E.D.

We find the rotation covariance of these operators,

$$Q\ \cos\ \theta+P\ \sin\ \theta=e^{i\theta J_3}Qe^{-i\theta J_3}.$$

Moreover, it turns out that this representation is convenient in dealing with coherent states. Let $R(\theta,\phi)$ denote the rotation matrix

$$R(\theta,\phi)=e^{i\theta(J_1\ \sin\ \phi-J_2\ \cos\ \phi)}.$$

$R(\theta,\phi)$ has also the expression[11]

$$R(\theta,\phi)=e^{\mu J_-}e^{-\log(1+|\mu|^2)J_3}e^{-\mu^* J_-^\dagger},$$

where $\mu=e^{i\phi}\tan(\theta/2)$. Define $|\theta,\phi\rangle$ by $|\theta,\phi\rangle=R(\theta,\phi)|J_3;j\rangle$. It is shown that

$$|\theta,\phi\rangle=\frac{1}{(1+|\mu|^2)^j}\sum_{k=0}^{2j}\binom{2j}{k}^{1/2}\mu^k|J_3;j-k\rangle.$$

If $|\mu|\sqrt{j}<\infty$ and $k<\infty$,

$$\frac{1}{(1+|\mu|^2)^j}\approx e^{-j|\mu|^2},\quad \binom{2j}{k}^{1/2}\mu^k\approx\frac{(\sqrt{2j}\mu)^k}{\sqrt{k!}}.$$

Thus we find

$$\langle J_3;j-k|\theta,\phi\rangle\approx e^{-|z|^2/2}\frac{z^k}{k!},$$

where $z=\mu\sqrt{2j}$. Corresponding standard coherent state $|z\rangle$ satisfies

$$\langle k|z\rangle=e^{-|z|^2/2}\frac{z^k}{k!},$$

where $|k\rangle=a^{\dagger k}|0\rangle/\|a^{\dagger k}|0\rangle\|$, $a^\dagger$ being the standard bose creation operator and $|0\rangle$ the vacuum state.

III. REPRESENTATIONS OF $so(n)$

The hyperfinite-dimensional representation of the CCR which has more general rotation covariance is constructed by hyperfinite-dimensional representation of $so(n)$. This section reviews the method of *spin representation* in use of Clifford algebra. Let $C(\mathbf{R}_-^n)$ be the $\mathbf{R}$-vector space the base of which is the 2^n elements

$$1,\ e_1,\ldots,e_n,\quad e_1e_2,\ldots,e_ie_j(i<j),\ldots,e_{n-1}e_n,\ldots,$$

$$e_{i_1}\ldots,e_{i_k}(i_1<\cdots<i_k),\ldots,e_1\ldots e_n.$$

Define the product of these elements by

$$(e_{i_1}\cdots e_{i_k})(e_{j_1}\cdots e_{j_l})=e_{i_1}\cdots e_{i_k}e_{j_1}\cdots e_{j_l}$$

and require the relation

$$e_i^2=-1,\ (i=1,\ldots,n),\ e_ie_j=-e_je_i(i\neq j),\ (\lambda 1)e_i=e_i(\lambda 1)\quad(\lambda\in\mathbf{R}).$$

Assume that 1 is the unit of the product, and that the product is associative. Now, $C(\mathbf{R}_-^n)$ is the 2^n-dimensional $\mathbf{R}$-algebra. $C(\mathbf{R}_-^n)$ is called *Clifford algebra.*

Let us identify $(\lambda_1,\ldots,\lambda_n)\in\mathbf{R}^n$ with $\Sigma_{k=1}^n\lambda_k e_k$. Then we have $S^{n-1}\subset\mathbf{R}^n\subset C(\mathbf{R}^n_-)$ (S^{n-1} is the unit sphere in $\mathbf{R}^n$). The following fact is shown.

Theorem 3.1: $Spin(n)$ defined by

$$Spin(n)=\{\alpha\in C(\mathbf{R}^n_-)\,|\,\alpha=a_1\ldots a_m, a_i\in S^{n-1}, m=2,4,\ldots\}$$

is a topological group, and isomorphic to the covering group of $so(n)$. $e_{ij}=e_ie_j$ $(i,j=1,\ldots,n)$ is a base (of a representation) of $so(n)$.

If $n=2\nu+1$, $e_1,\ldots,e_n$ is represented by $2^\nu\times 2^\nu$ matrices as follows. Let $V_k\simeq\mathbf{C}^2$ $(k=1,\ldots,\nu)$. Pauli matrices $\sigma_{1,k}$, $\sigma_{2,k}$, $\sigma_{3,k}$ that act on V_k are represented as

$$\sigma_{1,k}=\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad \sigma_{2,k}=\begin{pmatrix}0&-i\\i&0\end{pmatrix},\quad \sigma_{3,k}=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

Define V by

$$V=V_1\otimes\cdots\otimes V_\nu,$$

and define $\hat\sigma_{c,k}$ and γ_i that acts on V by

$$\hat\sigma_{c,k}=\overbrace{1\otimes\cdots\otimes 1}^{k-1}\otimes\sigma_{c,k}\otimes 1\otimes\cdots\otimes 1,\quad c=1,2,3,$$

$$\gamma_{2k-1}=\hat\sigma_{2,k}\hat\sigma_{3,k+1}\cdots\hat\sigma_{3,\nu},$$

$$\gamma_{2k}=-\hat\sigma_{1,k}\hat\sigma_{3,k+1}\cdots\hat\sigma_{3,\nu},$$

$$\gamma_{2\nu+1}=\hat\sigma_{3,1}\hat\sigma_{3,2}\cdots\hat\sigma_{3,\nu}.$$

Direct calculations show the following relations:

$$\{\gamma_i,\gamma_j\}=2\delta_{ij},\quad i,j=1,\ldots,2\nu.$$

Thus, we can take $e_j=i\gamma_j$ (this is called the *spin representation*). Moreover we have the tensor product representation of $so(n)$ from one constructed above,

$$\tilde e_{ij}=\sum_{l=0}^{p-1}\overbrace{1\otimes\cdots\otimes 1}^{l}\otimes e_{ij}\otimes\overbrace{1\otimes\cdots\otimes 1}^{p-l-1}$$

also generate a representation of $so(n)$.

In the next section, we introduce a hyperfinite representation of $so(n)$, using these mathematical tools.

IV. HYPERFINITE PARA-FERMI REPRESENTATION

In this section, we construct a hyperfinite-dimensional representation of the algebra of standard infinite bose oscillators, that is, the algebra generated by bose annihilation operators $a_1,a_2,\ldots$ satisfying $[a_j,a_k^\dagger]=\delta_{jk}$ and $[a_j,a_k]=0$, with the condition that the nonzero vector $|0\rangle$ that satisfies $a_k|0\rangle=0$ is unique except the scalar multiples (*the uniqueness of vacuum*). This representation is constructed by using hyperfinite internal representations of the algebra of para-Fermi oscillators, which is mathematically equivalent to the hyperfinite-dimensional spin representation of $so(n)$.

Definition 4.1:[12,13] Let $\nu\in\mathbf{N}$, and suppose that for some $d\in\mathbf{N}$, $b_1,\ldots,b_\nu\in M(d,\mathbf{C})$ (i.e., $b_1,\ldots,b_\nu$ are finite-dimensional matrices). $b_1,\ldots,b_\nu$ is called *annihilation operators* of para-Fermi oscillators of order $p\in\mathbf{N}$ if they satisfy

$$[b_k,[b_l^\dagger,b_m]]=2\delta_{kl}b_m,$$

$$[b_k,[b_l^\dagger,b_m^\dagger]]=2\delta_{kl}b_m^\dagger-2\delta_{km}b_l^\dagger,$$

$$[b_k,[b_l,b_m]]=0,$$

and the uniqueness of vacuum $|0\rangle$, and,

$$b_kb_l^\dagger|0\rangle=\delta_{kl}p|0\rangle.$$

$b_1^\dagger,\ldots,b_\nu^\dagger$ are called *creation operators* of para-Fermi oscillators of order p. *Hyperfinite annihilation operators* of para-Fermi oscillators are the internal matrices defined in the above definition by substituting ${}^\star\mathbf{N}$ and ${}^\star\mathbf{C}$ for $\mathbf{N}$ and $\mathbf{C}$, respectively.

Green[12] has given a class of representations of the above commutation relations of para-Fermi creation and annihilation operators. In the so-called *Green representation* for the cases of order p, the para-Fermi operators b_k are expressed by the form

$$b_k=\sum_{\alpha=1}^{p} b_k^{(\alpha)},$$

where the *Green-component* operators $b_k^{(\alpha)}$ satisfy the commutation relations

$$\{b_k^{(\alpha)},b_l^{(\alpha)^\dagger}\}=\delta_{kl},\quad \{b_k^{(\alpha)},b_l^{(\alpha)}\}=0,$$

$$[b_k^{(\alpha)},b_l^{(\beta)^\dagger}]=[b_k^{(\alpha)},b_l^{(\beta)}]=0,\quad (\alpha\neq\beta),$$

where $\{A,B\}=AB+BA$, and the uniqueness of vacuum $|0\rangle$ such that

$$b_k^{(\alpha)}|0\rangle=0 \text{ for all } k,\alpha.$$

Green representation is essentially equivalent to the tensor product representation of Clifford algebra representation of $so(2\nu)$. In fact, we easily verify that $e_1,\ldots,e_{2\nu}$ defined by $e_{2k-1}=i(b_k^\dagger+b_k)$, $e_{2k}=b_k^\dagger-b_k$ form the generator of a Clifford algebra (i.e., $e_i^2=-1$ and $e_ie_j=-e_je_i$ $(i\neq j)$ hold). Thus, we can construct a $2^{p\nu}$-dimensional representation of Green components by using a spin representation of the Clifford algebra as follows. Let $V_k^{(\alpha)}\simeq\mathbf{C}^2$ $(k=1,\ldots,\nu,$ $\alpha=1,\ldots,p)$. Pauli matrices $\sigma_{1,k}^{(\alpha)}$, $\sigma_{2,k}^{(\alpha)}$, and $\sigma_{3,k}^{(\alpha)}$ that act on $V_k^{(\alpha)}$ are represented as

$$\sigma_{1,k}^{(\alpha)}=\begin{pmatrix}0&1\\1&0\end{pmatrix},\ \sigma_{2,k}^{(\alpha)}=\begin{pmatrix}0&-i\\i&0\end{pmatrix},\ \sigma_{3,k}^{(\alpha)}=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

Define $V^{(\alpha)}$ by

$$V^{(\alpha)}=V_1^{(\alpha)}\otimes\cdots\otimes V_\nu^{(\alpha)},$$

and define $\hat\sigma_{c,k}^{(\alpha)}$ and $\gamma_i^{(\alpha)}$ $(k=1,\ldots,\nu,\ i=1,\ldots,2\nu)$ that acts on $V^{(\alpha)}$ by

$$\hat\sigma_{c,k}^{(\alpha)}=\overbrace{1\otimes\cdots\otimes 1}^{k-1}\otimes\sigma_{c,k}^{(\alpha)}\otimes 1\otimes\cdots\otimes 1,\quad c=1,2,3,$$

$$\gamma_{2k-1}^{(\alpha)}=\hat\sigma_{2,k}^{(\alpha)}\hat\sigma_{3,k+1}^{(\alpha)}\cdots\hat\sigma_{3,\nu}^{(\alpha)},$$

$$\gamma_{2k}^{(\alpha)}=-\hat\sigma_{1,k}^{(\alpha)}\hat\sigma_{3,k+1}^{(\alpha)}\cdots\hat\sigma_{3,\nu}^{(\alpha)}.$$

Operators $b_k^{(\alpha)}$ $(k=1,\ldots,\nu)$ defined by

$$b_k^{(\alpha)}=\tfrac{1}{2}(\gamma_{2k-1}^{(\alpha)}-i\gamma_{2k}^{(\alpha)}),$$

satisfy the relations

$$\{b_k^{(\alpha)}, b_l^{(\alpha)\dagger}\} = \delta_{kl}, \quad \{b_k^{(\alpha)}, b_l^{(\alpha)}\} = 0,$$

for all $k,\ l = 1,...,\nu$.

Define V and $\tilde{b}_k^{(\alpha)}$ $(k = 1,...,\nu)$ acting on V by

$$V = V^{(1)} \otimes \cdots \otimes V^{(p)},$$

$$\tilde{b}_k^{(\alpha)} = \overbrace{1 \otimes \cdots \otimes 1}^{\alpha-1} \otimes b_k^{(\alpha)} \otimes 1 \otimes \cdots \otimes 1.$$

We see that for all $k,\ l = 1,...,\nu$,

$$\{\tilde{b}_k^{(\alpha)}, \tilde{b}_l^{(\alpha)\dagger}\} = \delta_{kl}, \quad \{\tilde{b}_k^{(\alpha)}, \tilde{b}_l^{(\alpha)}\} = 0,$$

$$[\tilde{b}_k^{(\alpha)}, \tilde{b}_l^{(\beta)\dagger}] = [\tilde{b}_k^{(\alpha)}, \tilde{b}_l^{(\beta)}] = 0, \quad (\alpha \neq \beta).$$

Let $|0\rangle_k^{(\alpha)} \in V_k^{(\alpha)}$ denote the normalized vector such that $b_k^{(\alpha)}|0\rangle_k^{(\alpha)} = 0$. Define $|0\rangle^{(\alpha)}$ and $|0\rangle$ by

$$|0\rangle^{(\alpha)} = |0\rangle_1^{(\alpha)} \otimes \cdots \otimes |0\rangle_\nu^{(\alpha)},$$

$$|0\rangle = |0\rangle^{(\alpha)} \otimes \cdots \otimes |0\rangle^{(p)}.$$

Now, we find that $\tilde{b}_1^{(\alpha)}, \ldots, \tilde{b}_\nu^{(\alpha)}$ are $2^{p\nu}$-dimensional representations of Green components and $|0\rangle$ is the vacuum. Thus, $b_k = \Sigma_{\alpha=1}^p \tilde{b}_k^{(\alpha)}$, $(k = 1,...,\nu)$ are $2^{p\nu}$-dimensional representations of annihilation operators of ν para-Fermi oscillators of order p. Let us call the above representation of the algebra of para-Fermi oscillators *spin representation*.

Define $\sigma_{\pm,k}^{(\alpha)}$ by

$$\sigma_{\pm,k}^{(\alpha)} = (\sigma_{1,k}^{(\alpha)} \pm i\sigma_{2,k}^{(\sigma)})/2$$

and $|1\rangle_k^{(\alpha)} \in V_k^{(\alpha)}$ by

$$|1\rangle_k^{(\alpha)} = \sigma_{+,k}^{(\alpha)}|0\rangle_k^{(\alpha)},$$

$$\{(|e_1^{(1)}\rangle_1^{(1)} \cdots |e_\nu^{(1)}\rangle_\nu^{(1)}) \cdots (|e_1^{(p)}\rangle_1^{(p)} \cdots |e_\nu^{(p)}\rangle_\nu^{(p)}) : e_k^{(\alpha)} = 0,1\}$$

($\otimes$'s are omitted) is a complete orthonormal system of V. We write the vectors simply as $|\{e_k^{(\alpha)}\}\rangle$.

Number operator N on V and the related operators N_k, $N^{(\alpha)}$ are defined as follows:

$$N_k^{(\alpha)} = \frac{1}{2}(1 + \sigma_{3,k}^{(\alpha)}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\hat{N}_k^{(\alpha)} = \overbrace{1 \otimes \cdots \otimes 1}^{k-1} \otimes N_k^{(\alpha)} \otimes \overbrace{1 \otimes \cdots \otimes 1}^{\nu-k},$$

$$\tilde{N}_k^{(\alpha)} = \overbrace{1 \otimes \cdots \otimes 1}^{\alpha-1} \otimes \hat{N}_k^{(\alpha)} \otimes \overbrace{1 \otimes \cdots \otimes 1}^{p-\alpha},$$

$$N_k = \sum_{\alpha=1}^{p} \tilde{N}_k^{(\alpha)}, \quad N^{(\alpha)} = \sum_{k=1}^{\nu} \tilde{N}_k^{(\alpha)}, \quad N = \sum_{\alpha=1}^{p} N^{(\alpha)}.$$

We see that

$$N|\{e_k^{(\alpha)}\}\rangle = n|\{e_k^{(\alpha)}\}\rangle,$$

where n is the number of $e_k^{(\alpha)}$'s that is equal to 1. It is easily shown that

$$\tilde{b}_k^{(\alpha)\dagger}\tilde{b}_k^{(\alpha)} = \tilde{N}_k^{(\alpha)}, \quad \tilde{b}_k^{(\alpha)}\tilde{b}_k^{(\alpha)\dagger} = 1 - \tilde{N}_k^{(\alpha)}, \quad N_k = \tfrac{1}{2}([b_k^\dagger, b_k] + p),$$

$$[N_k, N_l] = 0, \quad N_k b_k = b_k(N_k - 1), \quad N_k b_k^\dagger = b_k^\dagger(N_k + 1), \quad \text{etc.}$$

Lemma 4.1: Suppose that hyperfinite para-Fermi annihilation operators $b_1, \ldots, b_\nu$ are represented by spin representation, and that their order p is an infinite hypernatural number ($b_1, \ldots, b_\nu$ are $2^{p\nu} \times 2^{p\nu}$ internal matrices acting on ${}^\star\mathbf{C}^{2^{p\nu}}$). If $|\xi\rangle \in {}^\star\mathbf{C}^{2^{p\nu}}$ satisfies $\langle\xi|\xi\rangle$, $\langle\xi|N^2|\xi\rangle < \infty$, and $k \neq l$ $(k, l = 1, 2, \ldots, \nu)$, then

(i) $[\beta_k, \beta_l]|\xi\rangle \approx [\beta_k, \beta_l^\dagger]|\xi\rangle \approx 0.$
(ii) $[\beta_k, \beta_k^\dagger]|\xi\rangle \approx |\xi\rangle.$
(iii) $\beta_k \beta_k^{\dagger n}|\xi\rangle \approx (\beta_k^{\dagger n}\beta_k + n\beta_k^{\dagger n-1})|\xi\rangle,$

where $\beta_k = p^{-1/2} b_k$ (*normalization of* b_k) and $n < \infty$.

Proof:

(i) Notice that

$$[\beta_k, \beta_l] = \frac{2}{p}\sum_{\alpha=1}^{p} \tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)}, \quad [\beta_k, \beta_l^\dagger] = \frac{2}{p}\sum_{\alpha=1}^{p} \tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)\dagger}.$$

Direct calculations show that

$$b_k^{(\alpha)} b_l^{(\alpha)} = \begin{cases} -\hat{\sigma}_{3,k+1}^{(\alpha)} \cdots \hat{\sigma}_{3,l}^{(\alpha)} \hat{\sigma}_{-,k}^{(\alpha)} \hat{\sigma}_{-,l}^{(\alpha)} & (l > k) \\ -\hat{\sigma}_{3,l+1}^{(\alpha)} \cdots \hat{\sigma}_{3,k}^{(\alpha)} \hat{\sigma}_{-,k}^{(\alpha)} \hat{\sigma}_{-,l}^{(\alpha)} & (k > l) \end{cases},$$

$$b_k^{(\alpha)} b_l^{(\alpha)^\dagger} = \begin{cases} \hat{\sigma}_{3,k+1}^{(\alpha)} \cdots \hat{\sigma}_{3,l}^{(\alpha)} \hat{\sigma}_{+,k}^{(\alpha)} \hat{\sigma}_{-,l}^{(\alpha)} & (l > k) \\ \hat{\sigma}_{3,l+1}^{(\alpha)} \cdots \hat{\sigma}_{3,k}^{(\alpha)} \hat{\sigma}_{+,k}^{(\alpha)} \hat{\sigma}_{-,l}^{(\alpha)} & (k > l) \end{cases}.$$

Hence, using $N_k^{(\alpha)} + N_l^{(\alpha)} - 2N_k^{(\alpha)} N_l^{(\alpha)} = (N_k^{(\alpha)} - N_l^{(\alpha)})^2 \geqslant 0$,

$$\|\tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)}|\xi\rangle\|^2 = \langle\xi|\tilde{\sigma}_{+,k}^{(\alpha)}\hat{\sigma}_{+,l}^{(\alpha)}\tilde{\sigma}_{-,l}^{(\alpha)}\tilde{\sigma}_{-,k}^{(\alpha)}|\xi\rangle = \langle\xi|\tilde{N}_k^{(\alpha)}\tilde{N}_l^{(\alpha)}|\xi\rangle \leqslant \frac{1}{2}\langle\xi|\tilde{N}_k^{(\alpha)} + \tilde{N}_l^{(\alpha)}|\xi\rangle \leqslant \frac{1}{2}\langle\xi|\tilde{N}^{(\alpha)}|\xi\rangle.$$

Thus,

$$\|[\beta_k, \beta_l]|\xi\rangle\| = \left\| \frac{2}{p}\sum_{\alpha=1}^{p} \tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)}|\xi\rangle \right\| \leqslant \frac{2}{p}\sum_{\alpha=1}^{p} \|\tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)}|\xi\rangle\| \leqslant \frac{2}{p}\sum_{\alpha=1}^{p} \sqrt{\frac{1}{2}\langle\xi|\tilde{N}^{(\alpha)}|\xi\rangle} \approx 0.$$

The last equation is seen from $\Sigma_{\alpha=1}^{p}\langle\xi|N^{(\alpha)}|\xi\rangle = \langle\xi|N|\xi\rangle < \infty$; under this condition we find that the number of α's such that $\langle\xi|N^{(\alpha)}|\xi\rangle \not\approx 0$ is finite.

In a similar way, it is shown that

$$\|[\beta_k, \beta_l^\dagger]|\xi\rangle\| \leqslant \frac{2}{p}\sum_{\alpha=1}^{p} \|\tilde{b}_k^{(\alpha)}\tilde{b}_l^{(\alpha)\dagger}|\xi\rangle\| \leqslant \frac{2}{p}\sum_{\alpha=1}^{p} \sqrt{\langle\xi|\tilde{N}_l^{(\alpha)}(1 - \tilde{N}_k^{(\alpha)})|\xi\rangle} \leqslant \frac{2}{p}\sum_{\alpha=1}^{p} \sqrt{\langle\xi|\tilde{N}_l^{(\alpha)}|\xi\rangle} \approx 0.$$

(ii) A direct calculation shows that

$$[\beta_k, \beta_k^\dagger] = -\frac{1}{p}\sum_{\alpha=1}^{p} \tilde{\sigma}_{3,k}^{(\alpha)},$$

$$([\beta_k, \beta_k^\dagger] - 1)^2 = \frac{4}{p^2}\left(\sum_{\alpha=1}^{p} N_k^{(\alpha)}\right)^2.$$

Hence,

$$\|[\beta_k, \beta_k^\dagger]|\xi\rangle - |\xi\rangle\|^2 = \frac{4}{p^2}\langle\xi|\left(\sum_{\alpha=1}^{p} N_k^{(\alpha)}\right)|\xi\rangle.$$

$$\leqslant \frac{4}{p^2}\langle\xi|N^2|\xi\rangle \approx 0.$$

(iii) is shown by the induction for n, using (ii). Q.E.D.

Suppose the number of para-Fermi oscillators ν and their order p are infinite hypernatural numbers. When n_i is a non-negative integer for any $i=1,2,...<\infty$, and the number of n_i's such that $n_i \neq 0$ is finite, we will define $|n_1,n_2,\ldots\rangle$ by

$$|n_1,n_2,\ldots\rangle = \frac{b_1^{\dagger n_1} b_2^{\dagger n_2} \ldots |0\rangle}{\| b_1^{\dagger n_1} b_2^{\dagger n_2} \ldots |0\rangle \|}.$$

$b_1^{\dagger n_1} b_2^{\dagger n_2} \ldots$ is the product of a finite number of operators, so it is well-defined. $N_k|n_1,n_2,\ldots\rangle = n_k|n_1,n_2,\ldots\rangle$ is easily shown, and hence, since N_k is Hermitian, the set of the vectors of the form $|n_1,n_2,\ldots\rangle$ is an orthonormal system.

Lemma 4.2:

(i) $\beta_k^\dagger \beta_k |n_1,n_2,\ldots\rangle \approx n_k |n_1,n_2,\ldots\rangle$,
(ii) $\beta_k \beta_k^\dagger |n_1,n_2,\ldots\rangle \approx (n_k+1)|n_1,n_2\ldots\rangle$,
(iii) $\| \beta_1^{\dagger n_1} \beta_2^{\dagger n_2} \ldots |0\rangle \| \approx \sqrt{n_1! n_2! \cdots}$,
(iv) $\beta_k^\dagger |n_1,n_2,\ldots\rangle \approx \sqrt{n_k+1}\,|n_1,n_2,\ldots n_k+1,\ldots\rangle$,
(v) $\beta_k |n_1,n_2,\ldots\rangle \approx \sqrt{n_k}\,|n_1,n_2,\ldots,n_k-1,\ldots\rangle$.

Proof: (i) and (ii) are shown by Lemma 4.1 (i)–(iii). (iii) is shown by the induction for $n_1,n_2,\ldots$, using (ii) and Lemma 4.1 (i). (iv) and (v) are shown by (iii) and Lemma 4.1 (i) (iii). Q.E.D.

Define a set $D \subset {}^\star\mathbf{C}^{2^{p\nu}}$ by

$$D = \left\{ \left. \frac{\beta_{k_1}^\dagger \cdots \beta_{k_n}^\dagger |0\rangle}{\| \beta_{k_1}^\dagger \cdots \beta_{k_n}^\dagger |0\rangle \|} \right| n,k_1,\ldots,k_n \in \mathbf{N} \right\} \cup \{|0\rangle\}.$$

Clearly, every vector in D is a normalized eigenvector of the number operator N with a finite eigenvalue. Let S denote the external subspace of ${}^\star\mathbf{C}^{2^{p\nu}}$ spanned by D, i.e.,

$$S = \left\{ \sum_{i=1}^{n} c_i |\xi\rangle : c_i \in {}^\star\mathbf{C}, |c_i| < \infty, n \in \mathbf{N}, |\xi\rangle \in D \right\}.$$

Theorem 4.3: (β_k, S) $(k \in \mathbf{N})$ is a representation of the CCR of a countably-infinite degree of freedom, i.e., S is invariant with respect to β_k and $\beta_k^\dagger$ for every $k \in \mathbf{N}$, and

$$[\beta_k, \beta_l]|\xi\rangle \approx 0,$$

$$[\beta_k, \beta_l^\dagger]|\xi\rangle \approx \delta_{kl}|\xi\rangle,$$

for any $|\xi\rangle \in S$. Moreover, the uniqueness of the vacuum is satisfied in the following sense: if $|\xi\rangle \in S$, $\langle\xi|\xi\rangle = 1$, and $\beta_k|\xi\rangle \approx 0$ for all $k \in \mathbf{N}$, then $|\langle\xi|0\rangle| \approx 1$.

APPENDIX: NONSTANDARD ANALYSIS

This section briefly outlines the basic notions of nonstandard analysis along Hurd and Loeb,[6] and reviews some development in the field of hyperfinite-, dimensional linear space theory. Let X be a set and $\mathscr{P}(X)$ the power set of X, that is, the set of all subsets of X. The *superstructure over* X, denoted by $V(X)$, is defined by the following recursion:

$$V_0(X) = X, V_{n+1}(X) = V_n(X) \cup \mathscr{P}(V_n(X)),$$

$$V(X) = \bigcup_{n \in \mathbf{N}} V_n(X).$$

Let us regard any element of X as a nonset here; hence $x \in V(X)$ is a set iff $x \in V(X)\backslash X$. Let $\mathbf{C}$ be the set of complex numbers. $V(\mathbf{C})$ contains all the structures that we usually use in quantum physics, for instance, separable Hilbert space $\mathscr{H}$.

$V(X)$ is called a *nonstandard extension of* $V(\mathbf{C})$ if there exists a map $\star$: $V(\mathbf{C}) \to V(X)$ satisfying the following conditions:

(1) $\star$ is an injective mapping from $V(\mathbf{C})$ to $V(X)$;
(2) ${}^\star\mathbf{C} = X$;
(3) (Transfer Principle) Let ϕ be a sentence in terms of $V(\mathbf{C})$, and ${}^\star\phi$ the sentence ''transferred'' from ϕ by mapping $\star$. ϕ is true iff ${}^\star\phi$ is true.

The Transfer Principle needs more explanation. Although the exact description of it needs more definitions of some notions in mathematical logic, one can find the intuitive meaning of it without them, in an example. A sentence in terms of $V(\mathbf{C})$ is constructed from the symbols for logical connectives $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, quantifiers $\forall$, $\exists$, individual variables $x,y,z,\ldots$, two predicates $=$, $\in$, parentheses $(,)$, and elements of $V(\mathbf{C})$. We will consider an example. Let $\mathbf{R}$ denote the set of real numbers. Define $G_< \in V(\mathbf{C})$ by $G_< = \{(x,y) | x,y \in \mathbf{R}, x<y\}$, where (x,y) is identified as $\{\{x\},y\}$,

$$(\forall x)(\forall y)(x \in \mathbf{R} \wedge y \in \mathbf{R} \wedge (x,y) \in G_< \Rightarrow (\exists z)(z \in \mathbf{R} \wedge (x,z) \in G_< \wedge (z,y) \in G_<))$$

is a sentence in terms of $V(\mathbf{C})$ because $\mathbf{R}$, $G_< \in V(\mathbf{C})$. Let ϕ denote this sentence. ϕ means that $\mathbf{R}$ is dense, and hence ϕ is true. The ''transferred'' sentence ${}^\star\phi$ is as follows:

$$(\forall x)(\forall y)(x \in {}^\star\mathbf{R} \wedge y \in {}^\star\mathbf{R} \wedge (x,y) \in {}^\star G_< \Rightarrow (\exists z)(z \in {}^\star\mathbf{R} \wedge (x,z) \in {}^\star G_< \wedge (z,y) \in {}^\star G_<)).$$

By transfer principle, ${}^\star\phi$ is true (${}^\star\mathbf{R}$ is called $\star$-dense).

$u \in V(X)$ is called *standard* if there is $x \in V(\mathbf{C})$ such that $u = {}^\star x$, and called *internal* if there is $x \in V(\mathbf{C})$ such that $u \in {}^\star x$. ${}^\star V(\mathbf{C})$ is the set of all internal sets. Let A and B be internal sets. Function $f: A \to B$ is called *internal* if the graph of f (i.e., $\{(x,f(x)) | x \in A\}$) is internal. $V(X)$ is called a *countably saturated extension* of $V(\mathbf{C})$ if it satisfies the following condition:

(Saturation Principle) If countable sequence of internal sets $A_j \in V(X)\backslash X$ satisfies

$$\bigcap_{j=1}^{k} A_j \neq \phi (k = 1,2,\ldots),$$

then

$$\bigcap_{j=1}^{\infty} A_j \neq \phi.$$

Let $A \in V(\mathbf{C})\backslash\mathbf{C}$. From the Saturation Principle, we can show that if A is an infinite set, ${}^s A$ defined by

$${}^s A = \{{}^\star a | a \in A\}$$

is a proper subset of ${}^\star A$. Hence ${}^\star A$ is an extended structure of A. ${}^\star A$ is called the *nonstandard extension of* A. By renaming the elements of X, we can assume without loss of generality that $\mathbf{C}$ is a subset of ${}^\star\mathbf{C}$ and ${}^\star x = x$ for each $x \in \mathbf{C}$.

Given any subset $U \subseteq V(\mathbf{C})$, define ${}^\star U$ by

$${}^\star U = \bigcup_{n \in \mathbf{N}} {}^\star(U \cap V_n(\mathbf{C})).$$

Let $F(\mathbf{C}) \subseteq V(\mathbf{C})$ be the set of finite sets, i.e.,

$$F(\mathbf{C}) = \{A \in V(\mathbf{C})\backslash\mathbf{C} | A \text{ is a finite set}\}.$$

An element of ${}^\star F(\mathbf{C})$ is called a *hyperfinite set*. It is shown that if A is a hyperfinite set, then there is an initial segment $J=\{n\in{}^\star\mathbf{N}|n\leqslant j\}$ for some $j\in{}^\star\mathbf{N}$ and a one-to-one, onto internal mapping $f:J\to A$. Thus, we will often write a hyperfinite set A as $A=\{a_1,a_2,\ldots,a_j\}$, where $a_k=f(k)$, $k\in J$.

Any element of ${}^\star\mathbf{C}({}^\star\mathbf{R})$ is called a *hypercomplex (hyper-real)* number. We assume that ${}^\star\mathbf{R}\subseteq{}^\star\mathbf{C}$. ${}^\star\mathbf{C}$ is a proper extension field of $\mathbf{C}$, and ${}^\star\mathbf{R}$ is an ordered extension of $\mathbf{R}$. An element of ${}^\star\mathbf{N}({}^\star\mathbf{Z})$ is called a *hypernatural number (hyperinteger)*. A hypercomplex number x is called *infinite* if $|x|>n$ for any $n\in\mathbf{N}$, *finite*, $|x|<\infty$, if there is some $n\in\mathbf{N}$ such that $|x|<n$, and *infinitesimal* if $|x|<(1/n)$ for any $n\in\mathbf{N}$.

For any $x,y\in{}^\star\mathbf{C}$, we will write $x\approx y$ if $|x-y|$ is infinitesimal. For any finite hypercomplex number x, there is a unique complex number z such that ${}^\star z\approx x$; this z is called the *standard part* of x and denoted by ${}^\circ x$.

Let $\mathscr{A}$ be an internal normed linear space with norm $\|\cdot\|$. The *principal galaxy* $\mathrm{fin}(\mathscr{A})$ and the *principal monad* $\mu(0)$ are defined by

$$\mathrm{fin}(\mathscr{A})=\{x\in\mathscr{A}|\|x\|<\infty\},$$

$$\mu(0)=\{x\in\mathscr{A}|\|x\|\approx 0\}.$$

Both of them are linear spaces over $\mathbf{C}$. The *nonstandard hull of* $\mathscr{A}$ is the quotient linear space $\hat{\mathscr{A}}=\mathrm{fin}(\mathscr{A})/\mu(0)$ equipped with the norm given by

$$\|{}^\circ x\|={}^\circ\|x\|$$

for all $x\in\mathrm{fin}(\mathscr{A})$, where ${}^\circ x=x+\mu(0)$. It is shown by the Saturation Principle that $\hat{\mathscr{A}}$ is a Banach space.[6]

Let ν be an infinite hypernatural number, i.e., $\nu\in{}^\star\mathbf{N}\backslash\mathbf{N}$. Define ${}^\star\mathbf{C}^\nu$ by

$${}^\star\mathbf{C}^\nu=\{f|f:\{1,\ldots,\nu\}\to{}^\star\mathbf{C};\text{internal}\}.$$

${}^\star\mathbf{C}^\nu$ is a ν-dimensional internal inner product space with the natural inner product and the internal norm $\|\cdot\|$ derived by the inner product. Define $\mathbf{M}=M({}^\star\mathbf{C},\nu)$, the set of $\nu\times\nu$ internal matrices over ${}^\star\mathbf{C}$, by

$$\mathbf{M}=\{f|f:\{1,\ldots,\nu\}^2\to{}^\star\mathbf{C};\text{internal}\}.$$

Naturally $\mathbf{M}$ acts on ${}^\star\mathbf{C}^\nu$ as the internal linear operators. Let p_∞ be the operator norm on $\mathbf{M}$, i.e., $p_\infty(A)=\sup\{\|A\xi\||\|\xi\|\leqslant 1,\xi\in{}^\star\mathbf{C}^\nu\}$. Denote by $A^\star$ the adjoint of $A\in\mathbf{M}$. Let τ be the internal normalized trace on $\mathbf{M}$, i.e.,

$$\tau(A)=\frac{1}{\nu}\sum_{i=1}^{\nu}A_{ii}$$

for $A=(A_{ij})\in\mathbf{M}$. Then τ defines an internal inner product $(\cdot|\cdot)$ on $\mathbf{M}$ by $(A|B)=\tau(A^\star B)$, for A, $B\in\mathbf{M}$. Its derived norm called the normalized Hilbert–Schmidt norm is denoted by p_2, i.e., $p_2(A)=\tau(A^\star A)^{1/2}$ for $A\in\mathbf{M}$. Denoted by $(\mathbf{M},p_\infty)$ and $(\mathbf{M},p_2)$ the normed linear spaces are equipped with these respective norms. The principal galaxies $\mathrm{fin}_\infty(\mathbf{M})$ of $(\mathbf{M},p_\infty)$ and $\mathrm{fin}_2(\mathbf{M})$ of $(\mathbf{M},p_2)$ are defined as follows:

$$\mathrm{fin}_\infty(\mathbf{M})=\{A\in\mathbf{M}|p_\infty(A)<\infty\},$$

$$\mathrm{fin}_2(\mathbf{M})=\{A\in\mathbf{M}|p_2(A)<\infty\}.$$

The principal monads $\mu_\infty(0)$ of $(\mathbf{M},p_\infty)$ and $\mu_2(0)$ of $(\mathbf{M},p_2)$ are defined as follows:

$$\mu_\infty(0)=\{A\in\mathbf{M}|p_\infty(A)\approx 0\},$$

$$\mu_2(0)=\{A\in\mathbf{M}|p_2(A)\approx 0\}.$$

The nonstandard hull $\hat{\mathbf{M}}_2 = \mathrm{fin}_2(\mathbf{M})/\mu_2(0)$ turns out to be a Hilbert space with the inner product $\langle \cdot | \cdot \rangle$ and the norm $\| \cdot \|_2$ defined by

$$\langle A + \mu_2(0) | B + \mu_2(0) \rangle = {}^\circ(A|B)$$

and

$$\| A + \mu_2(0) \|_2 = p_2(A)$$

for $A, B \in \mathrm{fin}_2(\mathbf{M})$.

The nonstandard hull $\hat{\mathbf{M}}_\infty = \mathrm{fin}_\infty(\mathbf{M})/\mu_\infty(0)$ of $(\mathbf{M}, p_\infty)$ turns out to be a $C^\star$-algebra equipped with norm $\hat{p}_\infty$ defined by

$$\hat{p}_\infty(A + \mu_\infty(0)) = {}^\circ p_\infty(A)$$

for $A \in \mathrm{fin}_\infty(\mathbf{M})$.

Hinokuma and Ozawa[14] showed that another quotient space $\hat{\mathbf{M}}$ defined by

$$\hat{\mathbf{M}} = \mathrm{fin}_\infty(\mathbf{M})/(\mu_2(0) \cap \mathrm{fin}_\infty(\mathbf{M})).$$

is a von Neumann algebra of type II_1 factor.

ACKNOWLEDGMENTS

The author expresses his appreciation to Professor M. Ozawa for his general support.

[1] I. Ojima and M. Ozawa, ''Unitary representations of the hyperfinite Heisenberg group and the logical extension methods in physics,'' Open Syst. and Information Dynam. **2**, 107–128 (1993).
[2] S. Gudder, ''Toward a rigorous quantum field theory,'' Found. Phys. **24**, 1205–1225 (1994).
[3] M. Ozawa, ''Phase operator problem and macroscopic extension of quantum mechanics,'' Ann. Phys. (N.Y.) **257**, 65–83 (1997).
[4] T. Nakamura, ''A nonstandard representation of Feynman's path integrals,'' J. Math. Phys. **32**, 457–463 (1991).
[5] A. Robinson, *Non-Standard Analysis* (North–Holland, Amsterdam, 1966).
[6] A. E. Hurd and P. A. Loeb, *An Introduction to Nonstandard Real Analysis* (Academic, Orlando, 1985).
[7] S. Gudder and V. Naroditsky, ''Finite-dimensional quantum mechanics,'' Int. J. Theor. Phys. **20**, 619–643 (1981).
[8] A. O. Barut and A. J. Bracken, ''Compact quantum systems: Internal geometry of relativistic systems,'' J. Math. Phys. **26**, 2515–2519 (1985).
[9] R. J. B. Fawcett and A. J. Bracken, ''Simple orthogonal and unitary compact quantum systems and the Inönü-Wigner contraction,'' J. Math. Phys. **29**, 1521–1528 (1988).
[10] D. T. Pegg and S. M. Barnett, ''Phase properties of the quantized single-mode electromagnetic field,'' Phys. Rev. A **39**, 1665–1675 (1989).
[11] F. T. Arecchi *et al.*, ''Atomic coherent states in quantum optics,'' Phys. Rev. A **6**, 2211–2237 (1972).
[12] H. S. Green, ''A generalized method of field quantization,'' Phys. Rev. **90**, 270–273 (1953).
[13] Y. Ohnuki and S. Kamefuchi, *Quantum Field Theory and Parastatistics* (University of Tokyo Press, Tokyo, 1982).
[14] T. Hinokuma and M. Ozawa, ''Conversion from nonstandard matrix algebras to standard factors of type II_1,'' Ill. J. Math. **37**, 1–13 (1993).

Solutions of the modified chiral model in (2+1) dimensions

T. Ioannidou and W. J. Zakrzewski
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, United Kingdom

(Received 25 June 1997; accepted for publication 21 January 1998)

In this paper we deal with classical solutions of the modified chiral model on $\mathbf{R}^{2+1}$. Such solutions are shown to correspond to products of various factors which we call time-dependent unitons. Then the problem of solving the system of second-order partial differential equations for the chiral field is reduced to solving a sequence of systems of first-order partial differential equations for the unitons. © *1998 American Institute of Physics.* [S0022-2488(98)00505-2]

I. INTRODUCTION

As is well known,[1,2] some physical phenomena (relativity, particle physics, etc.) lead, in their mathematical description, to nonlinear partial differential equations. Some of these equations are elliptic (when we deal with time-independent problems), others are hyperbolic or parabolic. However, with few exceptions, nonlinear partial differential equations are hard to solve (especially when they involve more than one spatial variable). Although in some cases static solution can be found the generalization to the time dependent problems often presents a formidable task.

A simple model which exemplifies many of the points mentioned above is the so-called relativistic chiral model in (2+1) dimensions. This model involves a chiral field $J(x,y,t)$ which takes its values in $SU(N)$. The equation of motion for this field is given by

$$\partial_\mu(J^{-1}\,\partial^\mu J)=0, \quad J\in SU(N), \tag{1}$$

and is, in fact, the Euler–Lagrange equation corresponding to the Lagrangian density,

$$\mathscr{L}=\tfrac{1}{2}\,\mathrm{tr}[(J^{-1}\partial_\mu J)(J^{-1}\partial^\mu J)]. \tag{2}$$

Here the indices μ,ν,α range over the values 0,1,2, with $x^\mu=(t,x,y)$ and ∂_μ denotes partial differentiation with respect to x^μ. Thus J describes maps of $\mathbf{R}^{2+1}$ into $SU(N)$ and, when we restrict our attention to static fields, we have maps from $\mathbf{R}^2$ into $SU(N)$. If we restrict our attention further and consider only static fields whose energy is finite (see below) then the problem reduces to finding harmonic maps of $\mathbf{S}^2\to SU(N)$.

Such harmonic maps were studied by Uhlenbeck[3] who, in a seminal paper, showed how to construct these maps in an explicit form. Her construction involved the introduction of the concept of a uniton. She then showed how starting from a given solution one can find a new one by ''the addition of a uniton.'' This observation allowed her to demonstrate that all solutions can be obtained from the trivial ones (i.e., $J=$const) by the addition of a certain number [in fact $\leqslant(N-1)$] of unitons.

This beautiful result relied heavily on the integrability of the chiral model in (2+0) dimensions (for more details see Ref. 4). When we consider the same model in (2+1) dimensions, i.e., when we want to solve (1) we immediately run into a problem as (1) is not integrable and so Uhlenbeck's method does not apply.

Some time ago Ward[5] presented a model in (2+1) dimensions which is an integrable generalization of the static chiral model in (2+0) dimensions. The model involves an $SU(N)$ valued chiral field whose equation of motion now takes the form

$$(\eta^{\mu\nu}+\varepsilon^{\mu\nu\alpha}V_\alpha)\partial_\mu(J^{-1}\partial_\nu J)=0. \tag{3}$$

 © 1998 American Institute of Physics

Here the tensor $\eta^{\mu\nu}$ is the inverse Minkowski metric, given by $\eta^{\mu\nu}=\mathrm{diag}(-1,1,1)$; $\varepsilon^{\mu\nu\alpha}$ is the fully antisymmetric tensor with $\varepsilon^{012}=1$; and V_α is the constant unit vector pointing in the x-direction, $V_\alpha=(0,1,0)$.

Notice that although (3) is not relativistically covariant (as V_α is nonzero) the solutions of Uhlenbeck are static solutions of (3). As the model is integrable many solutions of (3) can be easily found. The first solutions of this model[5] exhibited many properties seen in other integrable models (like wave-like behavior with a phase shift, soliton-like structures passing through each other, etc.). Recently, Ward,[6] Ioannidou[7] and even more recently Anand[8] have found further interesting solutions of (3). These solutions described the evolution of soliton-like extended structures whose properties were quite similar to properties of solitonic objects in the $O(3)$ sigma model (i.e., they exhibit the 90° scattering).

The approach of Ward and Ioannidou involved treating fields of the model as the limiting cases of the Riemann problem with zeros. In this paper we return to Uhlenbeck's construction of static solutions and show that her construction can be generalized to give, in general nonstatic, solutions of the modified chiral model of Ward [i.e., of (3)]. Our approach reproduces the solutions reported in Refs. 6 and 7 and presents an unusual but interesting application of Uhlenbeck's approach.

II. THE MODIFIED MODEL

As shown by Ward[5] the modified chiral model (3) has the same conserved energy density as the original chiral model (1) and this energy is given by

$$\mathscr{E}=-\tfrac{1}{2}\,\mathrm{tr}[(J^{-1}J_t)^2+(J^{-1}J_x)^2+(J^{-1}J_y)^2]. \tag{4}$$

To specify the problem of finding solutions of (3) completely we have to state the boundary conditions satisfied by J. Thus in order to ensure the finiteness of energy of the solutions, we require

$$J=J_c+J_\theta(\theta)\ r^{-1}+O(r^{-2}),\quad r\to\infty, \tag{5}$$

where $z=r\ e^{i\theta}$ (with $z=x+iy$), J_c is a constant matrix and J_θ is independent of t.

Equation (3) is completely integrable, in the sense that it can be written as the compatibility condition for a pair of linear equations involving a spectral parameter λ. Define

$$A_z=\tfrac{1}{2}J^{-1}(\partial_z-\tfrac{1}{2}i\partial_t)J,\quad A_{\bar z}=\tfrac{1}{2}J^{-1}(\partial_{\bar z}+\tfrac{1}{2}i\partial_t)J, \tag{6}$$

where A_z, $A_{\bar z}$ are $N\times N$ matrices depending on x,y,t but not on λ.

Then the pair of linear operators,

$$\vartheta_{\bar z}+(1-\lambda)A_{\bar z},\quad \vartheta_z+(1-\lambda^{-1})A_z, \tag{7}$$

with $\vartheta_z=\partial_z+\frac{1}{2}\,i\lambda^{-1}\partial_t$, $\vartheta_{\bar z}=\partial_{\bar z}-\frac{1}{2}\,i\lambda\partial_t$, commute for all $\lambda\in\mathbf{C}^*=\mathbf{C}-\{0\}$, as a consequence of (3). Therefore one can find a solution, E_λ, of the equations

$$\vartheta_z E_\lambda=(1-\lambda^{-1})E_\lambda A_z,\quad \vartheta_{\bar z}E_\lambda=(1-\lambda)E_\lambda A_{\bar z}. \tag{8}$$

Theorem 1: If $J:\mathbf{R}^{2+1}\to SU(N)$ *is harmonic, then there exists a* $GL(N,\mathbf{C})$*-valued function* E_λ *of (8) which is analytic, holomorphic in* $\lambda\in\mathbf{C}^*$ *and unitary for* $|\lambda|=1$ *or* $\bar\lambda=\lambda^{-1}$ *such that* $E_{-1}=J$ *and*

$$E_\lambda=(E_{\bar\lambda}^{-1})^\dagger, \tag{9}$$

where † *denotes complex conjugate transpose. Conversely, if* E_λ *is holomorphic and analytic; and* A_z, $A_{\bar z}$ *defined by* (6) *are independent of* λ, *then* $J=E_{-1}$ *is harmonic.*

This follows from Ref. 3 and E_λ is called an extended solution corresponding to J. System (8) is overdetermined and in order for a solution E_λ to exist, A_z and $A_{\bar z}$ have to satisfy integrability conditions,

$$\begin{aligned}(\partial_z+\tfrac{1}{2}i\partial_t)A_{\bar z}+(\partial_{\bar z}-\tfrac{1}{2}i\partial_t)A_z=0,\\ (\partial_z-\tfrac{1}{2}i\partial_t)A_{\bar z}-(\partial_{\bar z}+\tfrac{1}{2}i\partial_t)A_z-2[A_{\bar z},A_z]=0.\end{aligned}\tag{10}$$

Notice that the first equation of (10) is equivalent to (3).

III. UHLENBECK'S CONSTRUCTION

It has been known for some time that all static solutions of Grassmannian models are also solutions of the static chiral model;[9] although not much is known about other solutions. However, Uhlenbeck[3] proved that any harmonic map from the Riemann sphere $\mathbf{S}^2$ to the unitarity group $U(N)$ can be factorized into a product of a finite number of factors (so-called unitons) involving maps into Grassmannians $G_k(\mathbf{C}^N)$ each of which satisfies a system of first-order partial differential equations. Thus the problem of solving a system of second-order partial differential equations for a static harmonic map has been reduced to having to solve a sequence of systems of first-order partial differential equations for the unitons. This observation can be generalized to the nonstatic case (3).

We will seek solutions of (8) of the form

$$E_\lambda=\prod_{k=0}^{n}(1+(\lambda-1)R_k),\tag{11}$$

for some n, where R_k are Hermitian projectors of the form $R_k=(q_k^\dagger\otimes q_k)/|q_k|^2$ (R_0 is a constant matrix, i.e., 0-uniton), which satisfy some first-order differential equations; and q_k are N-dimensional vectors which, in general, depend on $z,\bar z,t$ and $|q_k|^2=q_k\cdot q_k^\dagger$. Note that a constant E_λ is a 0-uniton. Then, following Uhlenbeck, we consider

$$\tilde E_\lambda=E_\lambda(1+(\lambda-1)R_n),\tag{12}$$

and find the following:

Theorem 2: *Let* $E_\lambda:\mathbf{C}^*\times\mathbf{R}^{2+1}\to GL(N,\mathbf{C})$ *be an extended harmonic map* $E_\lambda:\mathbf{R}^{2+1}\to SU(N)$ *for* $|\lambda|=1$. *Then* $\tilde E_\lambda=E_\lambda(1+(\lambda-1)R_n)$ *is an extended map for a Hermitian projection* (R_n) *if and only if*

$$\begin{aligned}(1-R_n)(\partial_t R_n-2iA_{\bar z}R_n)=0,\\ R_n(\partial_{\bar z}R_n-A_{\bar z}(1-R_n))=0,\end{aligned}\tag{13}$$

where $A_{\bar z}=-(A_z)^\dagger$ *given by* (6).

From the definition of the extended solution, $\tilde A_{\bar z}=(1-\lambda)^{-1}\tilde E_\lambda^{-1}\,\partial_{\bar z}\tilde E_\lambda$ and so

$$\tilde A_{\bar z}=(1+(\lambda^{-1}-1)R_n)(A_{\bar z}(1+(\lambda-1)R_n)-(\partial_{\bar z}-\tfrac{1}{2}i\lambda\partial_t)R_n),\tag{14}$$

using the fact that $\tilde E_\lambda^{-1}=(1+(\lambda^{-1}-1)R_n)E_\lambda^{-1}$ and $A_{\bar z}=(1-\lambda)^{-1}E_\lambda\,\partial_{\bar z}E_\lambda^{-1}$. But, from Theorem 1 we know that $\tilde E_\lambda$ is an extended solution if the λ and λ^{-1} terms vanish, which leads to the system (13).

Next we observe the following theorem:

Theorem 3: *Suppose that* E_λ *is an extended solution and so is* $\tilde E_\lambda=E_\lambda(1+(\lambda-1)R_n)$. *Then if* $A_{\bar z}=\frac{1}{2}E_\lambda^{-1}\,\partial_{\bar z}E_\lambda$ *and* $\tilde A_{\bar z}=\frac{1}{2}\tilde E_\lambda^{-1}\,\partial_{\bar z}\tilde E_\lambda$, *they are related by*

$$\tilde A_{\bar z}=A_{\bar z}-\partial_{\bar z}R_n+\tfrac{1}{2}i\partial_t R_n.\tag{15}$$

This follows automatically from the terms in (14) which do not contain λ. In general, by induction, one can show that

$$A_{zn}^{-}=\left(-\partial_{z}^{-}+\frac{1}{2}i\partial_{t}\right)\sum_{k=0}^{n}R_{k}, \tag{16}$$

and so, that the corresponding R_k satisfy the following equations:

$$\begin{aligned}&(1-R_n)\left(\partial_t R_n-2i\sum_{k=0}^{n-1}\left(-\partial_z^{-}+\frac{1}{2}i\partial_t\right)R_kR_n\right)=0,\\ &R_n\left(\partial_z^{-}R_n-\sum_{k=0}^{n-1}\left(-\partial_z^{-}+\frac{1}{2}i\partial_t\right)R_k\,(1-R_n)\right)=0.\end{aligned} \tag{17}$$

Observe that for $n=1$, system (17) gives $\partial_t R_1\ R_1=0$ and $R_1\partial_z^{-}R_1=0$; i.e., the equations for the holonomic static solutions (instantons) of the Grassmannian models. This is the case as

$$\partial_t R_1\ R_1\equiv q_1\ q_1^{\dagger}\ \frac{\partial_t q_1^{\dagger}\otimes q_1}{|q_1|^4}-\partial_t q_1^{\dagger}\ q_1\ \frac{q_1^{\dagger}\otimes q_1}{|q_1|^4}=0, \tag{18}$$

which is satisfied, if and only if, q_1 is independent of t. Using the same argument for $R_1\partial_z^{-}R_1=0$ we deduce that $q_1=q_1(z)$. For $n\neq 1$, we have more general and, in general t-dependent, solutions.

Now, let us concentrate on (3) and discuss the construction of its solutions. If we set $\widetilde{E}_{-1}=J$ and $E_{-1}=J_0$ in (12), then the chiral field J (due to Theorem 2) is of the form

$$J=J_0(1-2R_n), \tag{19}$$

and satisfies (3) if J_0 does and if R_n satisfies (13). This describes a time-dependent multi-soliton solution of (3) since it satisfies (10), i.e.,

$$\begin{aligned}\left(\partial_z+\frac{i}{2}\partial_t\right)\widetilde{A}_z^{-}+\left(\partial_z^{-}-\frac{i}{2}\partial_t\right)\widetilde{A}_z&\equiv\left(\partial_z+\frac{i}{2}\partial_t\right)\left(-(1-2R_n)\left(\partial_z^{-}+\frac{i}{2}\partial_t\right)R_n+(1-2R_n)A_z^{-}(1-2R_n)\right)\\&\quad+\left(\partial_z^{-}-\frac{i}{2}\partial_t\right)\left((1-2R_n)\left(\partial_z-\frac{i}{2}\partial_t\right)R_n+(1-2R_n)A_z(1-2R_n)\right)\\&=\left(\partial_z+\frac{i}{2}\partial_t\right)\left(A_z^{-}-\left(\partial_z^{-}-\frac{i}{2}\partial_t\right)R_n\right)\\&\quad+\left(\partial_z^{-}-\frac{i}{2}\partial_t\right)\left(A_z+\left(\partial_z+\frac{i}{2}\partial_t\right)R_n\right)=0,\end{aligned} \tag{20}$$

using equations (6), (13) and (19).

So, to construct time-dependent solutions of the $SU(N)$ model (3), one can start from a constant solution (zero-uniton) and add to it a one-uniton. This solution will be nonconstant — but it is static. Then, to this one uniton, one can add a second uniton, then a third one and so on. As we are going to see, the number of soliton-like structures in an n-uniton configuration is not related to n (i.e., the procedure of the ''addition of a uniton'' can both increase or decrease the number of solitons). In the static chiral model (1) Uhlenbeck's theorem shows that for $U(N)$ the largest uniton number is less than N. In our case, the t-dependence of the system (17) invalidates this argument; the number of unitons can be arbitrary. In the static $SU(2)$ case considered by Uhlenbeck the only solutions are those described by constant matrices (0-unitons) and factors constructed from holomorphic functions (1-unitons). In the modified chiral model the 0- and 1-uniton solutions are the same but then, in addition, we have further solutions corresponding to

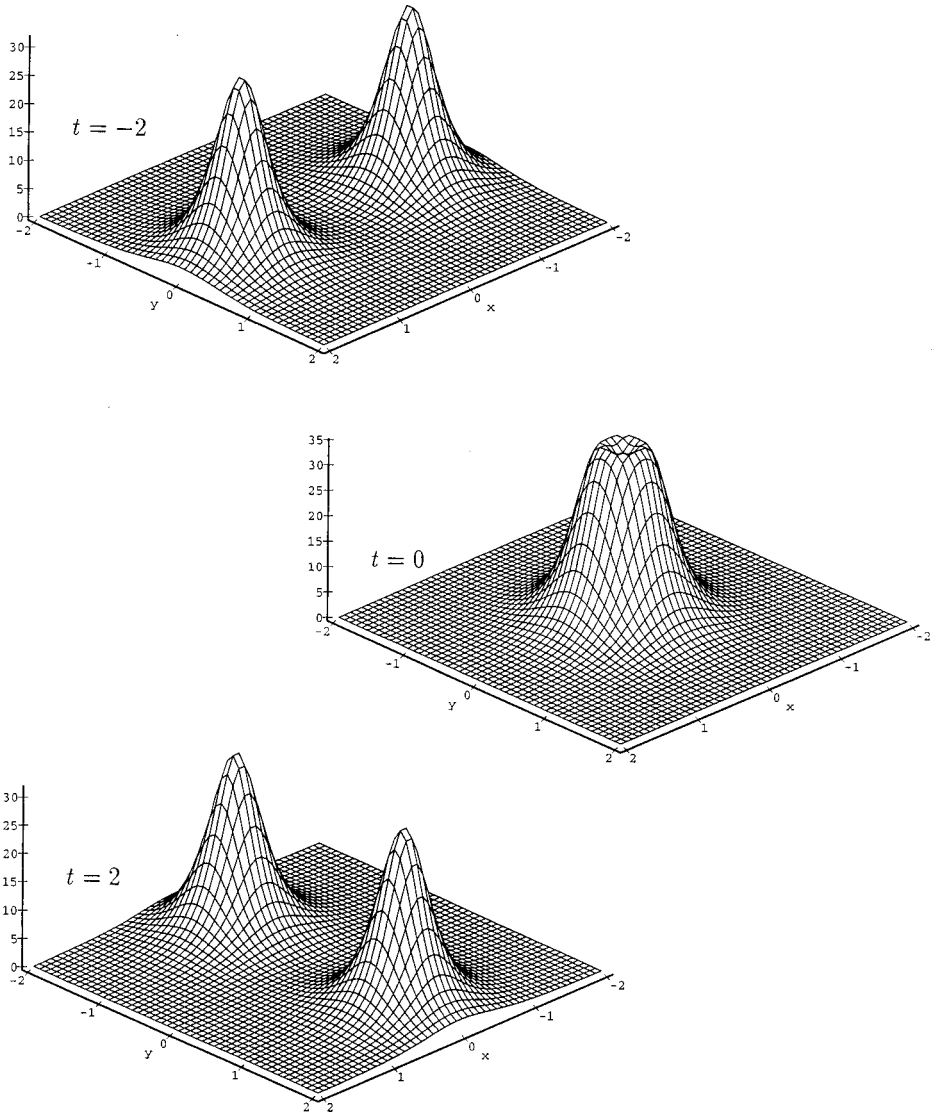

FIG. 1. Energy density of the embedding of $SU(2)$ solitons, at increasing times.

two and more unitons. These additional solutions are nonstatic. We do not know, at this stage, whether there is any bound on the uniton number so that all solutions correspond to field configurations of up to this number.

IV. THE $SU(2)$ AND $SU(3)$ CASES

In this section we shall construct and discuss, as an example, a 2-uniton solution of the $SU(3)$ model (3). First of all, observe that all fields of the $SU(2)$ model can be embedded into the $SU(3)$ model so that the $SU(2)$ solutions are automatically also solutions of the $SU(3)$ model. In what follows we will recover some of them as special cases of the derived expressions.

Let us observe that the following expression describes a nonstatic 2-uniton solution:

$$J=K\left(1-2\frac{q_1^{\dagger}\otimes q_1}{|q_1|^2}\right)\left(1-2\frac{q_2^{\dagger}\otimes q_2}{|q_2|^2}\right), \tag{21}$$

where q_i for $i=1,2$ are the three-dimensional vectors given by

$$q_1=(1,f,g), \tag{22}$$

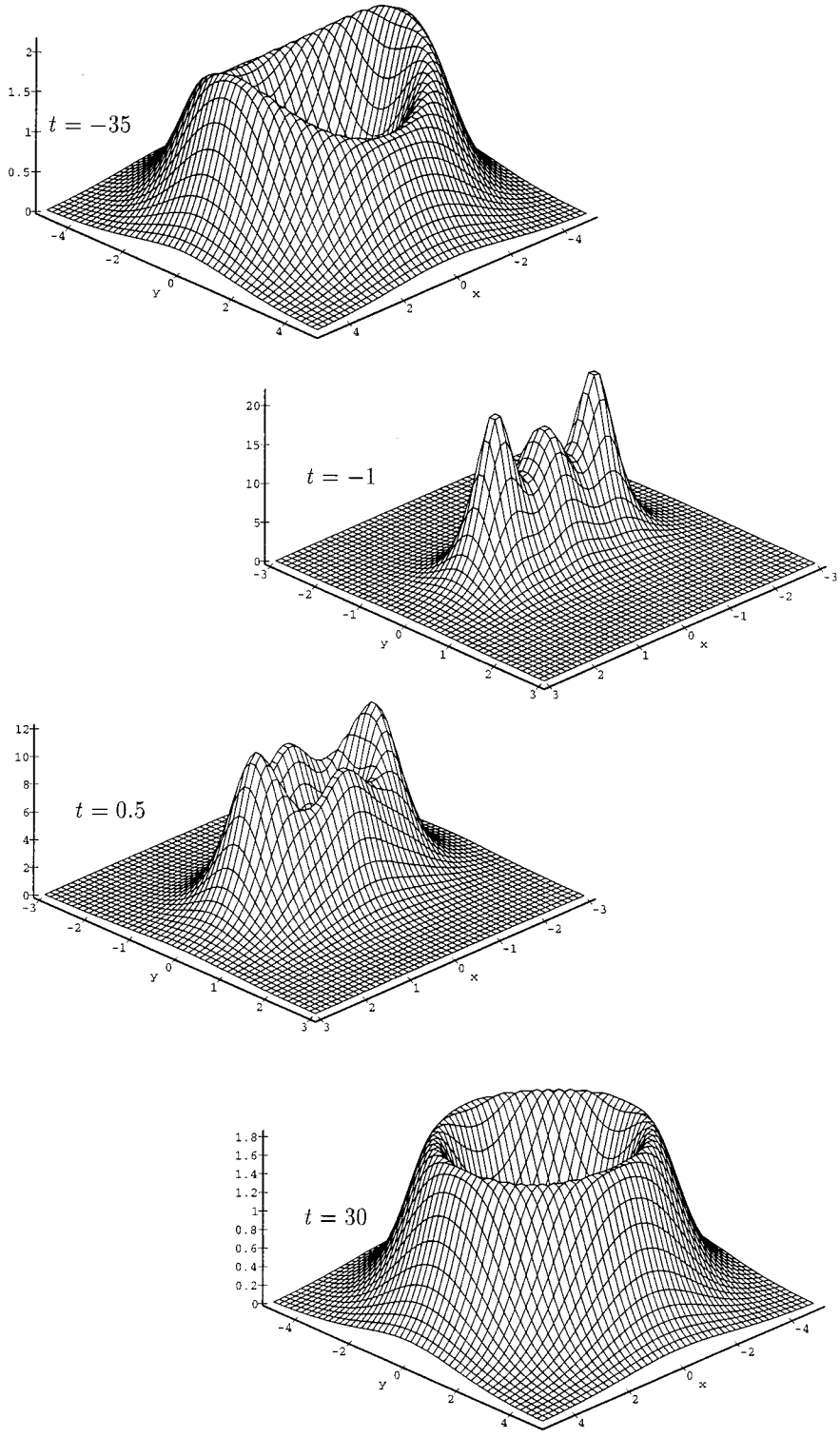

FIG. 2. Energy density at increasing times, for three ring-shaped solitons.

$$q_2=(1+|f|^2+|g|^2)(1,f,g)-2i(tf'+h)(\overline{f},-(1+|g|^2),g\overline{f})-2i(tg'+h_1)(\overline{g},\overline{g}\overline{f},-(1+|f|^2)),$$

and K is a constant matrix. Here f and g are rational meromorphic functions of z. The field J takes values in $SU(3)$, is smooth everywhere in $\mathbf{R}^{2+1}$ (as the two vectors q_i are nowhere zero). It satisfies the boundary condition (9) and the equation of motion (3).

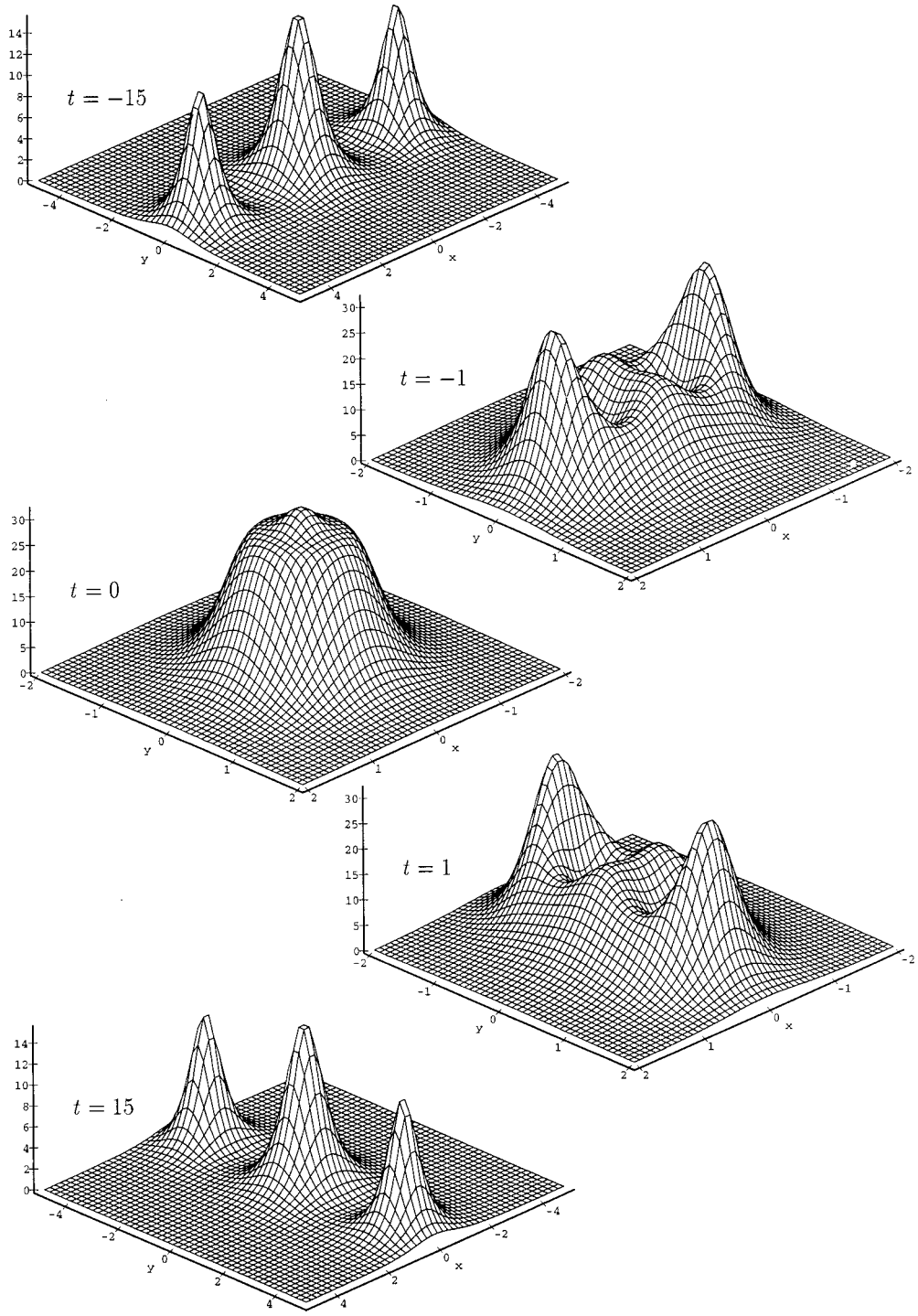

FIG. 3. Energy density at increasing times showing a 90° scattering between pure $SU(3)$ solitons.

Now let indicate how the solution (21) was constructed and how further solutions can be obtained. One way of proceeding is to take the q_i vectors of the form $q_i=(1,u_i,v_i)$. Then, for the 1-uniton solution corresponding to (17), the q_1 vector $q_1=(1,u_1,v_1)$ (as we have shown) is given in terms of the complex variable z. Thus we write $q_1=(1,f,g)$ where f and g are functions of z only. Then the system (17) for $n=2$ reduces to the following first-order differential equations for the function u_2 and v_2:

$$\partial_{\bar{z}}u_2=\frac{G}{(1+|f|^2+|g|^2)}(-u_2+f), \quad \partial_{\bar{z}}v_2=\frac{G}{(1+|f|^2+|g|^2)}(-v_2+f), \tag{23}$$

where $G=\overline{f}'(-u_2+f)+\overline{g}'(-v_2+g)+(u_2g-v_2f)(\overline{f}\overline{g}'-\overline{f}'\overline{g})$ and

$$\partial_t u_2=2i\frac{(1+u_2\overline{f}+v_2\overline{g})}{(1+|f|^2+|g|^2)}(u_2(\overline{f}f'+\overline{g}g')+\overline{g}(f'g-g'f)+f'),$$
$$\partial_t v_2=2i\frac{(1+u_2\overline{f}+v_2\overline{g})}{(1+|f|^2+|g|^2)}(v_2(\overline{f}f'+\overline{g}g')-\overline{f}(f'g-g'f)+g'), \tag{24}$$

and $'$ denotes the derivative with respect to z. A few lines of algebra allow us to show that q_2 of (22) is the general solution of the equations of the above system. Notice that when $v_i=0$ (i.e., $g=0$), (21) corresponds to the field J of the $SU(2)$ model (3).

Calculating the energy density for (21) we find

$$\mathscr{E}=\mathscr{E}_0+4\ \mathrm{tr}\left(\left(\partial_{\bar z}-\frac{i}{2}\partial_t\right)R_2\left(\partial_z+\frac{i}{2}\partial_t\right)R_2\right)-4\ \mathrm{tr}\left(\left(\partial_{\bar z}-\frac{i}{2}\partial_t\right)R_2\partial_z R_2 R_2\right)$$
$$-4\ \mathrm{tr}\left(\left(\partial_z+\frac{i}{2}\partial_t\right)R_2R_2\partial_{\bar z}R_2\right), \tag{25}$$

where $\mathscr{E}_0$ is the energy density of the 1-uniton (i.e., static) field. As is well known[3] $\mathscr{E}_0$ is a total derivative and so, probably, also is $\mathscr{E}$, but so far, we have not managed to prove this. However, we have looked at some special cases and have found that the total energy E, obtained by integrating $\mathscr{E}$, is quantized in units of 8π, i.e., is given by $E=8N\pi$. N appears to be related to the number of topological structures.

Let us discuss some special cases.

(*a*) The embedding of $SU(2)$ solitons, where we set

$$g=f, \quad h=h_1. \tag{26}$$

Such soliton solutions are localized along the direction of motion; they are not, however, of constant size. Their height (maximum of $\mathscr{E}$), is time dependent. In fact, for ($\deg f\geqslant\deg h$) no scattering occurs, i.e., the solitons are located at the center-of-mass forming a totally symmetric ring configuration. Otherwise, there are ($\deg f-1$) static solitons at the center-of-mass of the system accompanied by $N=\deg h-\deg f+1$ solitons (located at $t+z^N=0$; since J departs from its asymptotic value J_c when $tf'+h=0$) accelerating towards the solitons in the center, scattering at an angle of π/N, and then decelerating as they separate. This phenomenon has been observed, firstly, in the $SU(2)$ model (3) (cf. Refs. 6 7) and is a feature of systems (with nontrivial topology) admitting topological solitons.

As an example, let us take $f(z)=z$ and $h(z)=z^2$. The corresponding field represents a configuration of two solitons that undergo a 90° scattering. The energy density of this solution is

$$\mathscr{E}=32\frac{1+12r^2+12r^4+32r^2t^2+8t^2-16(x^2-y^2)t}{[1+12r^4+4r^2+16(x^2-y^2)t+1+8t^2]^2}, \tag{27}$$

and is symmetric under the interchange $t\to -t$, $x\to y$, $y\to x$. Figure 1 illustrates what happens close to $t=0$.

(*b*) In a more general case, i.e., for arbitrary values of our parameters, the configuration represents a multi-soliton solution that is localized close to the center of mass and deforms as the time passes. As an example, let us present a typical case by taking $f(z)=z$, $g(z)=2z$, $h(z)=z^2$ and $h_1(z)=3$. In this time-dependent solution, which corresponds to a configuration of three solitons, the energy density of the field, for large (negative) t, is in the shape of a ring which, as time passes, deforms to three peaks and then expands again to a ring. Figure 2 presents pictures of the corresponding energy densities at some representative values of time.

Returning to the scattering of our soliton-like structures we have seen that an embedding of the $SU(2)$ model provides an example of such a scattering. A further example is provided by a genuine $SU(3)$ case when we take

$$f=\rho^2(z), \quad g=\sqrt{2}\ \rho(z), \tag{28}$$

where $\rho(z)$ is an arbitrary function.

One example is presented in Figure 3, where we have chosen $\rho=z$, $h=z^3$ and $h_1=0$. This configuration can be seen to consist of two static solitons at the center-of-mass accompanied by two solitons accelerating towards the center, scattering at right angles and then decelerating as they separate.

V. CONCLUSIONS AND FURTHER COMMENTS

In this paper we have shown how to adapt Uhlenbeck's construction of uniton solutions of the chiral model in $(2+0)$ dimensions to the, time-dependent, uniton-like solutions of the modified chiral model in $(2+1)$ dimensions. It seems likely that there are many more interesting solutions still to be found. One could, for example, investigate the existence of time-dependent solutions based on other static ones constructed by Uhlenbeck.

As for static fields both models are the same, all Uhlenbeck's solutions are also solutions of the modified chiral model. However, at the level of two unitons and beyond, our construction has given us genuine time-dependent solutions of the modified chiral model. Looking at some examples we have seen that, in some cases, their field configurations can be thought of as representing scatterings of some soliton-like objects, sometimes accompanied also by static structures. The total energy appears to be quantized — thus showing that during their scattering the soliton-like structures must change their size.

We have not succeeded yet in gaining the full understanding of various properties of our solutions. These problems are currently being investigated.

ACKNOWLEDGMENTS

We thank R. S. Ward and B. Piette for helpful discussions. TI acknowledges support from EU ERBFMBICT950035.

[1] P. D. B. Collins, A. D. Martin, and E. J. Squires, *Particle Physics and Cosmology* (Wiley, New York, 1989).
[2] M. J. Ablowitz and P. A. Clarkson, *Soliton, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[3] K. Uhlenbeck, J. Diff. Geom. **30**, 1 (1989).
[4] W. J. Zakrzewski, *Low dimensional sigma models* (Institute of Physics, 1989).
[5] R. S. Ward, J. Math. Phys. **29**, 386 (1988).
[6] R. S. Ward, Phys. Lett. A **208**, 203 (1995).
[7] T. Ioannidou, J. Math. Phys. **37**, 3422 (1996).
[8] C. Anand, ''Ward's solitons II: Exact solutions,'' preprint.
[9] H. Eichenherr and M. Forger, Nucl. Phys. B **164**, 528 (1980).

Alternative potentials for the electromagnetic field

Shaun N. Mosley
Sunnybank, Albert Road, Nottingham NG3 4JD, England

(Received 24 November 1997; accepted for publication 5 February 1998)

The source-free electromagnetic field can be expressed in terms of two complex potentials α,β, which are related to the Debye potentials. The potentials $\{\alpha,\beta\}$ corresponding to various fields are discussed. A conserved radiation density ρ can be constructed in terms of these potentials, this density is positive (negative) for positive (negative) helicity radiation. © *1998 American Institute of Physics.* [S0022-2488(98)01605-3]

I. INTRODUCTION

The source-free electromagnetic field has only two degrees of freedom per space-time point. Some economy is achieved by reducing the six components of the electric and magnetic fields $(\mathbf{E},\mathbf{B})$ to the usual four potentials $(\Phi,\mathbf{A})$ which must satisfy the covariant Lorentz condition $\partial_t\Phi + \boldsymbol{\nabla}\cdot\mathbf{A}=0$, or the Φ component may be dispensed with provided that the three potentials $\mathbf{A}$ satisfy the non-covariant Coulomb gauge $\boldsymbol{\nabla}\cdot\mathbf{A}=0$. The most economical way to express the two degrees of freedom of the free electromagnetic field is in terms of the two Debye potentials[1,2] (ψ_1,ψ_2). The electric and magnetic fields are expressed in terms of these potentials by

$$\begin{bmatrix}\mathbf{E}\\ \mathbf{B}\end{bmatrix}=\begin{bmatrix}-(\mathbf{x}\times\boldsymbol{\nabla})\partial_t\\ \boldsymbol{\nabla}\times(\mathbf{x}\times\boldsymbol{\nabla})\end{bmatrix}\psi_1+\begin{bmatrix}-\boldsymbol{\nabla}\times(\mathbf{x}\times\boldsymbol{\nabla})\\ -(\mathbf{x}\times\boldsymbol{\nabla})\partial_t\end{bmatrix}\psi_2, \tag{1}$$

where ψ are solutions of the wave equation

$$(\partial_t^2-\nabla^2)\psi=0. \tag{2}$$

The standard textbooks may not refer to the Debye potentials by that name,[3] but use either ψ or $L^2\psi$ in their analysis of multipole radiation fields, where ψ are spherical solutions of the wave equation—see, for example pp. 432–433 of Reitz,[4] or pp. 745–746 of Jackson.[5] The paper by Boukamp and Casimir[6] discusses the various essentially equivalent approaches to multipole radiation, while Refs. 2, 7 show that any field $(\mathbf{E},\mathbf{B})$ outside the source region can be expressed in terms of the potentials ψ_1,ψ_2.

We express (1) more concisely by defining the complex field $\mathbf{F}\equiv\mathbf{E}+i\mathbf{B}$, then

$$\begin{aligned}\mathbf{F}&=[-(\mathbf{x}\times\boldsymbol{\nabla})\partial_t+i\boldsymbol{\nabla}\times(\mathbf{x}\times\boldsymbol{\nabla})](\psi_1+i\psi_2),\\ \mathbf{F}&=[i\partial_t+\boldsymbol{\nabla}\times]\mathbf{L}\psi,\end{aligned} \tag{3}$$

where $\mathbf{L}\equiv -i\mathbf{r}\times\boldsymbol{\nabla}$ is the angular momentum operator and $\psi\equiv-(\psi_1+i\psi_2)$. Conversely

$$\mathbf{r}\cdot\mathbf{F}=\mathbf{r}\cdot\boldsymbol{\nabla}\times\mathbf{L}\psi=iL^2\psi \tag{4}$$

and the operator L^2 is, in principle, invertible.[8] The source-free Maxwell's equations are

$$[i\partial_t-\boldsymbol{\nabla}\times]\mathbf{F}=0, \tag{5a}$$

$$\boldsymbol{\nabla}\cdot\mathbf{F}=0. \tag{5b}$$

It is readily verified that substituting (3) into the above yields an identity, given (2).

Despite the fact that the two Debye potentials (or the single complex Debye potential ψ) precisely contain the two degrees of freedom of the electromagnetic field, they are seldom used outside magnetostatics and multipole field analysis. Some disadvantages spring to mind:

 © 1998 American Institute of Physics

(a) the field $\mathbf{F}\equiv\mathbf{E}+i\mathbf{B}$ is expressed by second order differential operators acting on the potentials, so then for example the usual energy density of the field $(\mathbf{F}^*\mathbf{F})$ is a cumbersome expression in terms of ψ;
(b) there is no regular closed form potential ψ representing a plane wave field. The standard texts[9] express the plane wave potential as an infinite sum over l of the spherical harmonics $Y_{l,\pm 1}$, although this actually represents plane waves outgoing in each direction from the $(\theta=\pi/2)$ plane;
(c) the Debye potential formalism is ill suited to coping with point singularities representing charged particles in the field, i.e., there is no regular potential ψ representing the monopole field at the origin, as can be seen immediately from (4). See Ref. 1 for irregular potentials ψ representing the monopole field.

We will now show that, given (3), the field $\mathbf{F}$ can be expressed in terms of two complex potentials α,β acted on by first order differential operators. These potentials $\{\alpha,\beta\}$ admit charged particle solutions, for static and moving charges, and the potentials representing the plane wave fields are fairly simple. From (3) we calculate the field $(\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F})$ in terms of ψ, where $\hat{\mathbf{x}}\equiv\mathbf{x}/r$. We use the operator identities[2]

$$\boldsymbol{\nabla}\times\mathbf{L}=(-i\mathbf{x}\nabla^2+i\boldsymbol{\nabla}(\partial_r r)),\quad \hat{\mathbf{x}}\times\mathbf{L}=i(\boldsymbol{\nabla}r-\partial_r\mathbf{x}),\quad \hat{\mathbf{x}}\times(\boldsymbol{\nabla}\times\mathbf{L})=-\mathbf{L}(\partial_r r).$$

Then

$$\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F}=[i\partial_t\mathbf{L}-i\mathbf{x}\nabla^2+i\boldsymbol{\nabla}(\partial_r r)]\psi+i[-(\boldsymbol{\nabla}r-\partial_r\mathbf{x})\partial_t-\mathbf{L}(\partial_r r)]\psi,$$
$$\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F}=\left[\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t-\frac{1}{r}\mathbf{L}\right]\alpha, \tag{6a}$$

where

$$\alpha=i(\partial_r r\psi-\partial_t r\psi). \tag{7a}$$

To derive (6a) we have used (2). Similarly

$$\mathbf{F}-i\hat{\mathbf{x}}\times\mathbf{F}=\left[\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t+\frac{1}{r}\mathbf{L}\right]\beta \tag{6b}$$

with

$$\beta=i(\partial_r r\psi-\partial_t r\psi). \tag{7b}$$

Note that given $(\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F})$ there is no straightforward algebraic manipulation to derive $\mathbf{F}$ due to the fact that the matrix $(I+i\hat{\mathbf{x}}\times)$ is singular. We add (6a), (6b) obtaining

$$2\mathbf{F}=\left[\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t-\frac{1}{r}\mathbf{L}\right]\alpha+\left[\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t+\frac{1}{r}\mathbf{L}\right]\beta, \tag{8}$$

which is the basic formula expressing $\mathbf{F}$ in terms of $\{\alpha,\beta\}$. Dotting (6a), (6b) with $\hat{\mathbf{x}}$ yields the constraint

$$(\partial_r+\partial_t)\alpha=F_r=(\partial_r-\partial_t)\beta, \tag{9}$$

where $F_r\equiv\hat{\mathbf{x}}\cdot\mathbf{F}$. The operators $(\partial_r+\partial_t),(\partial_r-\partial_t)$ are the radial differential operators on the future, past lightcones, respectively. Given the radial field $F_r(t,\mathbf{x})$, then (9) enables us to calculate α,β as follows. To carry out the integration on the future (past) lightcone we introduce the advanced (retarded) times S and T defined by

$$S=t-r,\quad T=t+r. \tag{10}$$

The hypersurfaces $S=S_0,S_1\ldots$ are future lightcones centered at the origin, while the hypersurfaces $T=T_0,T_1\ldots$ are past lightcones. Then $F_r(S)$ or $F_r(T)$ are obtained by substituting $t=S+r$ or $t=T-r$ into the expression for $F_r(t)$. A solution of (9) is then

$$\alpha(S)=\int_0^r dr\ F_r(S),\quad \beta(T)=\int_0^r dr\ F_r(T). \tag{11}$$

Maxwell's equations can alternatively be formulated from the start using advanced or retarded time,[10] then α or β is found from F_r by a straightforward radial integration.

We now find the evolution equations for α,β, noting that $\alpha=i(\partial_r r-\partial_t r)\psi$ does not satisfy the wave equation like ψ, because the operator $(\partial_r r-\partial_t r)$ does not commute with the d'Alembertian operator $\Box\equiv(\partial_t^2-\nabla^2)$. The d'Alembertian may be expressed as follows:

$$(\partial_t^2-\nabla^2)\psi=\left[-\frac{1}{r}(\partial_r+\partial_t)(\partial_r-\partial_t)r+\frac{1}{r^2}L^2\right]\psi=0. \tag{12}$$

Now multiply (11) by $(i/r^2)(\partial_r-\partial_t)r^3$ obtaining

$$\left[-\frac{1}{r^2}(\partial_r-\partial_t)r^2(\partial_r+\partial_t)+\frac{1}{r^2}L^2\right]i(\partial_r-\partial_t)r\psi=\left[\partial_t^2-\nabla^2-\frac{2}{r}\partial_t\right]\alpha=0 \tag{13a}$$

so that the evolution operator for α is a modified d'Alembertian, having the extra term $-(2/r)\partial_t$. Similarly

$$\left[\partial_t^2-\nabla^2+\frac{2}{r}\partial_t\right]\beta=0. \tag{13b}$$

We see from (8) that adding any function $f(t-r)$ to α, or adding any function $g(t+r)$ to β leaves $\mathbf{F}$ unchanged. Also the evolution operators (13) for α,β, annihilate $f(t-r),g(t+r)$, respectively. So the gauge transformations for α,β are

$$\begin{array}{ll}\alpha\to\alpha', & \alpha'=\alpha+f(t-r),\\ \beta\to\beta' & \beta'=\alpha+g(t+r).\end{array} \tag{14}$$

II. THE POTENTIALS $\{\alpha,\beta\}$ CORRESPONDING TO RADIATION FIELDS

We first investigate some symmetries between α,β. Consider the parity transformation $\mathscr{P}:\mathbf{x}\to-\mathbf{x}$ and the time reversal transformation $\mathscr{T}:t\to-t$. Under $\mathscr{P},\mathscr{T}$ the fields transform as follows:[11]

$$\begin{array}{lll}\mathscr{P}\mathbf{E}(t,\mathbf{x})=-\mathbf{E}(t,-\mathbf{x}), & \mathscr{P}\mathbf{B}(t,\mathbf{x})=+\mathbf{B}(t,-\mathbf{x}), & \mathscr{P}\mathbf{F}(t,\mathbf{x})=-\mathbf{F}^*(t,-\mathbf{x}),\\ \mathscr{T}\mathbf{E}(t,\mathbf{x})=+\mathbf{E}(-t,\mathbf{x}), & \mathscr{T}\mathbf{B}(t,\mathbf{x})=-\mathbf{B}(-t,\mathbf{x}), & \mathscr{T}\mathbf{F}(t,\mathbf{x})=+\mathbf{F}^*(-t,\mathbf{x}),\end{array}$$

and

$$\mathscr{P}\mathscr{T}\mathbf{F}(t,\mathbf{x})=-\mathbf{F}(-t,-\mathbf{x}).$$

For brevity we will write a bar over any variable signifying its $\mathscr{P}\mathscr{T}$ transform, e.g., $\overline{\mathbf{F}}=\mathscr{P}\mathscr{T}\mathbf{F}$, $\overline{\alpha}=\mathscr{P}\mathscr{T}\alpha$, etc. Both α,β are scalars under $\mathscr{P}\mathscr{T}$, so that $\overline{\alpha}(t,\mathbf{x})=\alpha(-t,-\mathbf{x})$. Maxwell's equations are invariant under the $\mathscr{P}\mathscr{T}$ transformation, so that if $\mathbf{F}$ satisfies Maxwell's equations then so does $\mathbf{F}'=\overline{\mathbf{F}}$. In some cases of interest the field is even or odd under $\mathscr{P}\mathscr{T}$, by which we mean $\overline{\mathbf{F}}_{ev}=\mathbf{F}_{ev}$ and $\overline{\mathbf{F}}_{odd}=-\mathbf{F}_{odd}$ (from now on we will use the terms even and odd meaning even or odd under $\mathscr{P}\mathscr{T}$). Recall the fields $(\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F}),(\mathbf{F}-i\hat{\mathbf{x}}\times\mathbf{F})$ which are derived from α,β, respectively. Then if the field is even/odd,

$$(\underset{odd}{\mathbf{F}^{ev}}-i\hat{\mathbf{x}}\times\underset{odd}{\mathbf{F}^{ev}})=\pm\mathscr{P}\mathscr{T}(\underset{odd}{\mathbf{F}^{ev}}+i\hat{\mathbf{x}}\times\underset{odd}{\mathbf{F}^{ev}}). \tag{15}$$

In either of these cases the field is effectively given by one potential, α say, because after calculating $(\mathbf{F}+i\hat{\mathbf{x}}\times\mathbf{F})$ from α, one can write down $(\mathbf{F}-i\hat{\mathbf{x}}\times\mathbf{F})$ using (15). Applying the $\mathscr{PT}$ transformation to equations (6):

$$\overline{\mathbf{F}}-i\hat{\mathbf{x}}\times\overline{\mathbf{F}}=-\left[\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t+\frac{1}{r}\mathbf{L}\right]\overline{\alpha},\quad \overline{\mathbf{F}}+i\hat{\mathbf{x}}\times\overline{\mathbf{F}}=-\left[\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t-\frac{1}{r}\mathbf{L}\right]\overline{\beta},$$

so that

$$\mathbf{F}\to\overline{\mathbf{F}}\quad\text{is equivalent to}\quad\{\alpha,\beta\}\to\{-\overline{\beta},-\overline{\alpha}\}. \tag{16}$$

From (16) we see that

$$\{\mathbf{F}_{ev}:\quad\beta=-\overline{\alpha}\},\quad\{\mathbf{F}_{odd}:\quad\beta=+\overline{\alpha}\}. \tag{17}$$

We now find the potentials $\{\alpha,\beta\}$ corresponding to the radiation fields. Equations (11) enable us to calculate the potentials for a known field—which approach we follow in this section. Alternatively we can look for solutions of the evolution equations for the potentials (13), satisfying the constraint (9)—the approach we take in Appendix A. It turns out that these evolution equations (13) are separable in either spherical or parabolic coordinates—the spherical solutions yield the potentials corresponding to the multipole fields (which are singular at the origin), while the parabolic solutions are potentials for non-singular fields, including the plane wave fields.

Consider the plane wave field $\mathbf{F}$ propagating in the x_3 direction, the left circularly polarized (positive helicity) field $\mathbf{F}_L$ is

$$\mathbf{F}_L=\{1,i,0\}e^{ikx_3-ikt},\quad F_{rL}=(\hat{x}_1+i\hat{x}_2)e^{ikx_3-ikt}, \tag{18}$$

and its $\mathscr{PT}$ transform the right circularly polarized (negative helicity) field $\mathbf{F}_R$ is

$$\mathbf{F}_R=\overline{\mathbf{F}}_L=-\{1,i,0\}e^{-(ikx_3-ikt)},\quad F_{rR}=-(\hat{x}_1+i\hat{x}_2)e^{-(ikx_3-ikt)}.$$

Then recalling (10) we integrate $F_{rL}(S)=(\hat{x}_1+i\hat{x}_2)e^{-ikS}e^{-ikr(1-\hat{x}_3)}$ to find α, and follow the equivalent procedure for β, obtaining

$$\begin{aligned}\alpha_L&=\left(\frac{\hat{x}_1+i\hat{x}_2}{1-\hat{x}_3}\right)[e^{ikx_3}-e^{ikr}]\,\frac{e^{-ikt}}{ik},\\ \beta_L&=-\left(\frac{\hat{x}_1+i\hat{x}_2}{1+\hat{x}_3}\right)[e^{ikx_3}-e^{-ikr}]\,\frac{e^{-ikt}}{ik},\end{aligned} \tag{19}$$

which potentials are non-singular. The second term in the square brackets does not contribute to the field but its absence would make the potentials singular. Direct calculation verifies that

$$\left[\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t-\frac{1}{r}\mathbf{L}\right]\alpha_L=\{1+\hat{x}_3,i+i\hat{x}_3,-\hat{x}_1-i\hat{x}_2\}e^{ikx_3-ikt}$$

which is $\mathbf{F}_L+i\hat{\mathbf{x}}\times\mathbf{F}_L$, noting that the operator $[\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t-(1/r)\mathbf{L}]$ commutes with the $((x_1+ix_2)/(r-x_3))$ factor in α_L, and annihilates the term $e^{ik(r-t)}$. Similarly it can be verified that $[\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t+(1/r)\mathbf{L}]\beta_L=\mathbf{F}_L-i\hat{\mathbf{x}}\times\mathbf{F}_L$.

As $\mathbf{F}_R=\overline{\mathbf{F}_L}$, then recalling (16)

$$\alpha_R=-\overline{\beta_L}=-\left(\frac{\hat{x}_1+i\hat{x}_2}{1-\hat{x}_3}\right)[e^{-ikx_3}-e^{-ikr}]\,\frac{e^{ikt}}{ik}, \tag{20}$$

$$\beta_R = -\overline{\alpha_L} = \left(\frac{\hat{x}_1 + i\hat{x}_2}{1+\hat{x}_3}\right)[e^{-ikx_3} - e^{ikr}]\,\frac{e^{ikt}}{ik}.$$

The potentials for any plane wave propagating in the x_3 direction are a linear superposition of $\{\alpha_L,\beta_L\},\{\alpha_R,\beta_R\}$. Both $\{\alpha_L,\beta_L\},\{\alpha_R,\beta_R\}$ are eigenstates of L_3, the angular momentum operator in the direction of propagation, with eigenvalue $+1$.

We obtain the potential $\alpha_L(\mathbf{k})$ for a left circularly polarized wave propagating in the $\hat{\mathbf{k}} \equiv \mathbf{k}/|\mathbf{k}|$ direction with wavelength $2\pi/|\mathbf{k}|$, by first defining the null complex vector $\mathbf{h}$ satisfying

$$\mathbf{h}\times\hat{\mathbf{k}} = i\mathbf{h}, \quad \mathbf{h}^*\cdot\mathbf{h} = 2, \tag{21}$$

which so defined is unique up to a phase factor. Then

$$\{\alpha_L(\mathbf{k}),\beta_L(\mathbf{k})\} = \left\{ -\left(\frac{\mathbf{h}\cdot\mathbf{x}}{r-\hat{\mathbf{k}}\cdot\mathbf{x}}\right)[e^{i\mathbf{k}\cdot\mathbf{x}} - e^{i|\mathbf{k}|r}], \left(\frac{\mathbf{h}\cdot\mathbf{x}}{r+\hat{\mathbf{k}}\cdot\mathbf{x}}\right)[e^{i\mathbf{k}\cdot\mathbf{x}} - e^{-i|\mathbf{k}|r}] \right\} \frac{e^{-i|\mathbf{k}|t}}{ik}, \tag{22}$$

and $\{\alpha_R(\mathbf{k}),\beta_R(\mathbf{k})\} = \{-\beta_L(\mathbf{k}), -\alpha_L(\mathbf{k})\}$. We check that $\alpha_L(\mathbf{k}),\beta_L(\mathbf{k})$ are eigenstates of $(\hat{\mathbf{k}}\cdot\mathbf{L})$ the angular momentum operator in the direction of propagation, with eigenvalue $+1$

$$\hat{\mathbf{k}}\cdot\mathbf{L}(\mathbf{h}\cdot\mathbf{x}) = \hat{\mathbf{k}}\cdot(-i\mathbf{x}\times\mathbf{h}) = -i\mathbf{x}\cdot(\mathbf{h}\times\hat{\mathbf{k}}) = -i\mathbf{x}\cdot(i\mathbf{h}) = \mathbf{h}\cdot\mathbf{x}. \tag{23}$$

In the next section we investigate the orthonormality of the basis potentials $\{\alpha(\mathbf{k}),\beta(\mathbf{k})\}$ and $\{\alpha(\mathbf{k}'),\beta(\mathbf{k}')\}$.

We next briefly consider the multipole radiation fields and their corresponding potentials, noting some symmetry relations under the $\mathcal{PT}$ transformation. The Debye potential ψ in this case is a spherical solution of the wave equation, then the potentials $\{\alpha,\beta\}$ follow from (7). Consider

$$\psi_{lm} = j_l(kr)Y_{lm}(\theta,\phi)\cos(kt), \tag{24}$$

where j_l is the spherical Bessel function of order l, and Y_{lm} are the spherical harmonics. This ψ_{lm} when substituted into (2) yields $\mathbf{F}_{lm}$, the magnetic multipole field of order (l,m). (Multiplying the potential by i yields the corresponding electric multipole field $i\mathbf{F}_{lm}$.) Then from (7) the potentials $\{\alpha_{lm},\beta_{lm}\}$ for the magnetic multipole field of order (l,m) are

$$\alpha_{lm} = i(\partial_r - \partial_t)[rj_l(kr)Y_{lm}(\theta,\phi)\cos(kt)],$$

$$\beta_{lm} = i(\partial_r + \partial_t)[rj_l(kr)Y_{lm}(\theta,\phi)\cos(kt)].$$

We see that

$$\beta_{lm} = (-1)^l\overline{\alpha_{lm}}$$

following the parity of Y_{lm}. This means, recalling (17), that $\mathbf{F}_{lm}$ is odd or even, depending on whether l is an even or odd number. [If we had considered the potential ψ of (21) with a $(\sin(kt))$ instead of a $(\cos(kt))$ factor, then $\beta'_{lm} = (-1)^{l+1}\overline{\alpha'_{lm}}$.] Then, as discussed previously, $\mathbf{F}_{lm}$ can be calculated from either one of the potentials α,β.

III. THE CONSERVED DENSITY ρ AND THE ENERGY E

A. The radiation density ρ

The electromagnetic field is the classical ''first-quantized'' version of the massless spin-1 photon field, and as such should have a conserved density corresponding to the well-known densities for the Klein-Gordon (spin-zero) or Dirac (spin-$\frac{1}{2}$) fields. The only text on classical (non-quantized) electromagnetism which discusses a radiation density, or an inner product, for the electromagnetic field that I have come across is the book by Good and Nelson,[12] otherwise one has

to turn to texts on QED such as that by Schweber.[13] For an excellent recent review on this subject, see Ref. 14. This inner product for the electromagnetic field involves the non-local (integral) operator $(\nabla^2)^{-1/2}$.

The evolution equations (7) for α,β can be expressed in the form

$$(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\cdot(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\alpha=0, \quad (\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\cdot(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\beta=0. \tag{25}$$

It follows that

$$(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\cdot[\alpha(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\beta^*-\alpha^*(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\beta]-(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\cdot[\beta^*(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\alpha-\beta(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\alpha^*]=0. \tag{26}$$

Collecting terms and multiplying by $(-i)$ we can write (26) in the form of a continuity equation $\partial_t\rho+\boldsymbol{\nabla}\cdot\mathbf{J}=0$, where

$$\rho=-\frac{i}{4}[\alpha(\partial_r-\partial_t)\beta^*-\alpha^*(\partial_r-\partial_t)\beta+\beta^*(\partial_r+\partial_t)\alpha-\beta(\partial_r+\partial_t)\alpha^*], \tag{27}$$

$$\rho=-\frac{i}{4}(\alpha-\beta)F_r^*+\text{c.c.}=-\frac{i}{4}[(\alpha-\beta)(\partial_r+\partial_t)\alpha^*]+\text{c.c.}$$

$$\mathbf{J}=-\frac{i}{4}[\alpha(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\beta^*-\alpha^*(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_t)\beta-\beta^*(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\alpha+\beta(\boldsymbol{\nabla}+\hat{\mathbf{x}}\partial_t)\alpha^*], \tag{28}$$

where we have used (9), and c.c. stands for the complex conjugate terms. Let us substitute the $\{\alpha_L,\beta_L\}$ of (19) into (27):

$$\rho_L=-\frac{1}{4k}\left(-\left(\frac{\hat{x}_1+i\hat{x}_2}{1-\hat{x}_3}\right)[e^{ikx_3}-e^{ikr}]-\left(\frac{\hat{x}_1+i\hat{x}_2}{1+\hat{x}_3}\right)[e^{ikx_3}-e^{-ikr}]\right)((\hat{x}_1-i\hat{x}_2)e^{-ikx_3})+\text{c.c.}$$

$$=\frac{1}{2k}((1+\hat{x}_3)[1-\cos(kr-kx_3)]+(1-\hat{x}_3)[1-\cos(kr+kx_3)])\geqslant 0.$$

Inserting $\{\alpha_R,\beta_R\}=-\{\overline{\beta_R},\overline{\alpha_R}\}$ into (27), we obtain $\rho_R=-\rho_L$. Hence ρ_L is non-negative and ρ_R is non-positive.

The density ρ has close similarity with the Klein–Gordon density for a spin-zero particle: $\rho_{KG}=-(i/2)(\phi^*\partial_t\phi-\phi\partial_t\phi^*)$. In this case solutions ϕ with time dependence $e^{-i\omega t}$ with ω positive are regarded as particle solutions, those with time dependence $e^{+i\omega t}$ are regarded as anti-particle solutions. Analagously we can regard the right circularly polarized (negative helicity) field as the ''anti-photons'' of the left circularly polarized (positive helicity) photon field. There appears to be no comparable radiation density in the literature, although of course one can project out the different polarizations when the field is expressed as a Fourier integral.[15,16]

We now investigate the orthogonality of the potentials for two different fields $\mathbf{F}_1$ and $\mathbf{F}_2$. It will be convenient to represent the potentials for the field $\mathbf{F}_1$ by the 2×1 matrix $U_1\equiv[{}^{\alpha_1}_{\beta_1}]$, then we construct the following indefinite scalar product space

$$\langle U_1|U_2\rangle\equiv\int d^3x\ \rho_{12}, \tag{29}$$

where ρ_{12} is the conserved density

$$\rho_{12}=-\frac{i}{4}[U_1^T(I\partial_t+\sigma^3\partial_r)U_2^*-U_2^{T*}(I\partial_t+\sigma^3\partial_r)U_1], \tag{30}$$

$$\rho_{12}=-\frac{i}{4}[(\alpha_1-\beta_1)F_{r2}^*-(\alpha_2^*-\beta_2^*)F_{r1}].$$

Here $U^T=[\alpha\beta]$ is the transpose of U, and σ^3 is the Pauli matrix $\left[\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right]$. Note that when $U_1=U_2=U$, then $\rho_{12}=\rho$. In Appendix B we show that

$$\begin{aligned}\langle U_L(\mathbf{k})|U_L(\mathbf{k}')\rangle&=(2\pi)^3k\,\delta(\mathbf{k}-\mathbf{k}'),\\ \langle U_R(\mathbf{k})|U_R(\mathbf{k}')\rangle&=-(2\pi)^3k\,\delta(\mathbf{k}-\mathbf{k}'),\\ \langle U_L(\mathbf{k})|U_R(\mathbf{k}')\rangle&=0.\end{aligned}\tag{31}$$

The relations (31) enable one to project out the positive/negative helicity states of momentum $\mathbf{k}$, and show that ρ integrated over all space is the amount of left minus right circularly polarized radiation.

B. The energy density *E*

As a consequence of Maxwell's equations (5)

$$\partial_t\left(\frac{1}{2}\mathbf{F}\cdot\mathbf{F}^*\right)+\boldsymbol{\nabla}\cdot\left(\frac{i}{2}\,\mathbf{F}\times\mathbf{F}^*\right)=0,$$

where $E=\frac{1}{2}\mathbf{F}\cdot\mathbf{F}^*\equiv\frac{1}{2}(\mathbf{E}^2+\mathbf{B}^2)$ is the usual energy density, and $(i/2)\mathbf{F}\times\mathbf{F}^*\equiv\mathbf{E}\times\mathbf{B}$ is the momentum density. To simplify $\mathbf{F}\cdot\mathbf{F}^*$ in terms of α,β, we note the following: that from (8) we can express $\mathbf{F}$ as[17]

$$\mathbf{F}=\boldsymbol{\nabla}_\uparrow\alpha+\boldsymbol{\nabla}_\downarrow\beta+\hat{\mathbf{x}}F_r\tag{32}$$

with F_r expressed in terms of α,β by (9), and the operators $\boldsymbol{\nabla}_\uparrow,\boldsymbol{\nabla}_\downarrow$ are defined as

$$\boldsymbol{\nabla}_\uparrow\equiv\tfrac{1}{2}(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_r+i\hat{\mathbf{x}}\times\boldsymbol{\nabla}),\quad \boldsymbol{\nabla}_\downarrow=(\boldsymbol{\nabla}_\uparrow)^*\equiv\tfrac{1}{2}(\boldsymbol{\nabla}-\hat{\mathbf{x}}\partial_r-i\hat{\mathbf{x}}\times\boldsymbol{\nabla}).\tag{33}$$

These operators have the property that for any scalars χ,ζ

$$\begin{aligned}\boldsymbol{\nabla}_\uparrow\chi\cdot\boldsymbol{\nabla}_\uparrow\zeta=\boldsymbol{\nabla}_\downarrow\chi\cdot\boldsymbol{\nabla}_\downarrow\zeta=\hat{\mathbf{x}}\cdot\boldsymbol{\nabla}_\uparrow\chi=\hat{\mathbf{x}}\cdot\boldsymbol{\nabla}_\downarrow\chi=0,\\ \boldsymbol{\nabla}_\uparrow\chi\cdot\boldsymbol{\nabla}_\downarrow\zeta=\tfrac{1}{2}[(\hat{\mathbf{x}}\times\boldsymbol{\nabla}\chi)\cdot(\hat{\mathbf{x}}\times\boldsymbol{\nabla}\zeta)+i\hat{\mathbf{x}}\cdot(\boldsymbol{\nabla}\chi\times\boldsymbol{\nabla}\zeta)].\end{aligned}\tag{34}$$

Then substituting (32) into $(\mathbf{F}\cdot\mathbf{F}^*)$, and with the aid of (34), we find

$$\begin{aligned}2E=\mathbf{F}\cdot\mathbf{F}^*=[\boldsymbol{\nabla}_\uparrow\alpha+\boldsymbol{\nabla}_\downarrow\beta+\hat{\mathbf{x}}F_r]\cdot[\boldsymbol{\nabla}_\downarrow\alpha^*+\boldsymbol{\nabla}_\uparrow\beta^*+\hat{\mathbf{x}}F_r^*],\\ 2E=|\boldsymbol{\nabla}_\uparrow\alpha|^2+|\boldsymbol{\nabla}_\downarrow\beta|^2+|F_r|^2\end{aligned}\tag{35}$$

with F_r given by (9). We will further expand (35) as we shall see that E contains a divergence term, which can be removed such that the resulting energy E' is still conserved. Again using (34),

$$4E=|\hat{\mathbf{x}}\times\boldsymbol{\nabla}\alpha|^2+|\hat{\mathbf{x}}\times\boldsymbol{\nabla}\beta|^2+2|F_r|^2+i\hat{\mathbf{x}}\cdot[(\boldsymbol{\nabla}\alpha\times\boldsymbol{\nabla}\alpha^*)-(\boldsymbol{\nabla}\beta\times\boldsymbol{\nabla}\beta^*)],\tag{36}$$

and the last terms of (36) can be expressed as $i\boldsymbol{\nabla}\cdot[-\alpha(\hat{\mathbf{x}}\times\boldsymbol{\nabla}\alpha^*)+\beta(\hat{\mathbf{x}}\times\boldsymbol{\nabla}\beta^*)]$, so that the modified energy density E'

$$4E'=|\hat{\mathbf{x}}\times\boldsymbol{\nabla}\alpha|^2+|\hat{\mathbf{x}}\times\boldsymbol{\nabla}\beta|^2+2|F_r|^2\tag{37}$$

is also a conserved non-negative density, so is also a plausible candidate for the energy density.

IV. THE POTENTIALS $\{\alpha,\beta\}$ CORRESPONDING TO THE FIELD OF A CHARGED PARTICLE

First consider the static case, then the constraint $(\partial_r+\partial_t)\alpha=(\partial_r-\partial_t)\beta$ implies $\beta=\alpha$, then from (8) we obtain

$$\mathbf{F}_{\rm stc}=\boldsymbol{\nabla}\alpha_{\rm stc}$$

so that

$$\mathrm{Re}[\alpha]=-\Phi,\quad \mathrm{Im}[\alpha]=\Phi_m,$$

where Φ is the usual scalar potential, and Φ_m is the magnetic scalar potential. The potentials Φ,Φ_m for the electrostatic, magnetostatic fields are well known. The energy E' for the static field is $E'=\frac{1}{2}|\boldsymbol{\nabla}\alpha|^2$.

Magnetic monopole fields are also accomodated: the potential for a magnetic monopole at position $\mathbf{a}$ is just $\alpha=\beta=iq_m/|\mathbf{x}-\mathbf{a}|$. Note that if we use (11) to derive α from the radial field

$$F_r=q(r-\hat{\mathbf{x}}\cdot\mathbf{a})/|\mathbf{x}-\mathbf{a}|^3$$

of a charge at position $\mathbf{a}$, we obtain $\alpha=\int_0^r dr F_r=-(q/|\mathbf{x}-\mathbf{a}|)+(q/|\mathbf{a}|)$. For non-radiation fields, we can lose the constant of integration by instead obtaining α from F_r, so that (9) is satisfied, as follows:

$$\alpha(S)=\int_\infty^r dr\ F_r(S). \tag{38}$$

We will use (38) to determine the potentials $\{\alpha,\beta\}$ for a uniformly moving charge. The field of a charge passing through the origin in the x_1 direction is[18]

$$\mathbf{F}=\{x_1-vt,x_2-ivx_3,x_3+ivx_2\}\ \frac{q\gamma}{r'^{3/2}},\quad F_r=\frac{q\gamma(r-v\hat{x}_1t)}{r'^{3/2}}, \tag{39}$$

where $r'=[\gamma^2(x_1-vt)^2+x_2^2+x_3^2]^{1/2}$. Substituting $t=S+r$ into the expression for F_r,

$$F_r(S)=\frac{q\gamma(r-v\hat{x}_1S-v\hat{x}_1r)}{[\gamma^2(\hat{x}_1r-vS-vr)^2+(\hat{x}_2^2+\hat{x}_3^2)r^2]^{3/2}} \tag{40}$$

and integrating over r we obtain

$$\alpha=\left|{}^{r}_{\infty}-\frac{q\gamma[r(1-\hat{x}_1v)+S]}{S[\gamma^2(\hat{x}_1r-vS-vr)^2+(\hat{x}_2^2+\hat{x}_3^2)r^2]^{1/2}}\right|,$$
$$\alpha=-q\left(\frac{t'-r'}{t-r}\right)\frac{1}{r'}, \tag{41}$$

where $t'=\gamma(t-vx_1)$. This potential is non-singular except at the charge position, because when $t=r$, $r'=\gamma(r-vx_1)$, and $((t'-r')/(t-r))=\gamma$. The field $\mathbf{F}$ of (39) is an even field (which is only the case because the particle is passing through the origin), and so from (17)

$$\beta=-\bar{\alpha}=-q\left(\frac{t'+r'}{t+r}\right)\frac{1}{r'}.$$

V. OUTLOOK

We have calculated the potentials $\{\alpha,\beta\}$ corresponding to various fields: The simple relation (9) between the potentials and the radial field F_r makes the calculation of the potential from a given field quite straightforward, or one can solve the equations for the potentials from which one then calculates the field, as in Appendix A. Although $\{\alpha,\beta\}$ are scalars under rotations, they have complicated transformation properties under the Lorentz transformations—we will discuss these transformation properties elsewhere. (For the Lorentz transformation of the Debye potential ψ, also complicated, see the paper by Monroe.[19]) The stationary solutions for the potentials with time

dependence $e^{-ikt}(e^{+ikt})$ correspond to the left (right) circularly polarized waves: the orthogonality properties (31) suggest an alternative approach to second quantizing the electromagnetic field, which we hope to address in the future.

Under the duality transformation $(\mathbf{E}\rightarrow\mathbf{B},\mathbf{B}\rightarrow-\mathbf{E})$, or $\mathbf{F}\rightarrow i\mathbf{F}$, the transformation of the potentials is just $\{\alpha,\beta\}\rightarrow i\{\alpha,\beta\}$. The absence of any duality transformation for the usual potentials $A^{\mu}\equiv(\Phi,\mathbf{A})$ has been commented on recently by Witten.[20] On the other hand the interaction of a charged particle with the field is naturally described via the potentials A^{μ}, by replacing the free momentum p^{μ} with the gauge invariant $(p^{\mu}-eA^{\mu})$, whereas the role of the $\{\alpha,\beta\}$ potentials in gauge theory is not clear.

APPENDIX A: SOLUTIONS OF (13) IN PARABOLIC COORDINATES

We solve the equations for the potentials (13). The spherical solutions yield the multipole fields and here we discuss a few of the parabolic solutions. Defining λ,μ,ϕ such that λ,μ have the dimension of length[21,22]

$$x_1+ix_2=2\sqrt{\lambda\mu}e^{i\phi},\quad x_3=\lambda-\mu,\quad r=\lambda+\mu,$$

then[21,22]

$$\boldsymbol{\nabla}^2\equiv\frac{1}{\lambda+\mu}(\partial_\lambda\lambda\partial_\lambda+\partial_\mu\mu\partial_\mu)+\frac{1}{4\lambda\mu}\partial_\phi^2,$$

$$\boldsymbol{\nabla}^2-\partial_t^2+\frac{2}{r}\partial_t\equiv\frac{1}{\lambda+\mu}(\partial_\lambda\lambda\partial_\lambda+\partial_\mu\mu\partial_\mu+2\partial_t)+\frac{1}{4\lambda\mu}\partial_\phi^2-\partial_t^2$$

and with $\alpha=f(\lambda)g(\mu)e^{im\phi}e^{-ikt}$ then

$$\left[(\partial_\lambda\lambda\partial_\lambda+\partial_\mu\mu\partial_\mu-2ik)-\frac{m^2}{4\lambda}-\frac{m^2}{4\mu}+k^2\lambda+k^2\mu\right]f(\lambda)g(\mu)=0.$$

Inserting a separation constant $2ikc$, then

$$\left[\partial_\lambda\lambda\partial_\lambda-ik(1+2c)-\frac{m^2}{4\lambda}+k^2\lambda\right]f(\lambda)=0, \tag{A1}$$

$$\left[\partial_\mu\mu\partial_\mu-ik(1-2c)-\frac{m^2}{4\mu}+k^2\mu\right]g(\mu)=0. \tag{A2}$$

The solution of (A1) is $f(\lambda)=\lambda^{m/2}e^{-ik\lambda}{}_1F_1(1+m/2+c,m+1,2ik\lambda)$, where ${}_1F_1$ is the confluent hypergeometric function. So the solution for α is

$$\alpha=(\lambda\mu)^{m/2}e^{im\phi}e^{-ik(\lambda+\mu)}\ {}_1F_1(1+m/2+c,m+1,2ik\lambda)\,{}_1F_1(1+m/2-c,m+1,2ik\mu)e^{-ikt}. \tag{A3}$$

The confluent hypergeometric function reduces to a simpler expression for integer or half-integer values of c.

1. The case $m=0$, $c=0$

As ${}_1F_1(1,1,i\zeta)\equiv e^{i\zeta}$ we have

$$\alpha=e^{-ikt}e^{-ik(\lambda+\mu)}e^{2ik\lambda}e^{2ik\mu}=e^{-ikt}e^{ik(\lambda+\mu)}\equiv e^{ikr-ikt}$$

which potential substituted into (6a) yields a zero field.

2. The case $m=1$, $c=1/2$

Inserting these values of m,c into (A1), and using the identities ${}_1F_1(2,2,2i\zeta)\equiv e^{2i\zeta}$, ${}_1F_1(1,2,2i\zeta)\equiv(\sin\zeta/\zeta)e^{i\zeta}$, then

$$\alpha=\sqrt{\lambda\mu}e^{i\phi}e^{-ik(\lambda+\mu)}\ {}_1F_1(2,2,2ik\lambda)_1F_1(1,2,2ik\mu)e^{-ikt},$$

$$\alpha=\sqrt{\lambda\mu}e^{i\phi}e^{-ik(\lambda+\mu)}e^{2ik\lambda}\ \frac{\sin(k\mu)}{k\mu}e^{ik\mu}e^{-ikt},$$

$$\alpha=\sqrt{\lambda\mu}e^{i\phi}e^{ik\lambda}\ \frac{\sin(k\mu)}{k\mu}\ e^{-ikt},$$

which is the α_L of (19).

3. The case $m=1$, $c=0$

Substituting these values for m,c, into (A3) we have

$$\alpha_Z=\sqrt{\lambda\mu}e^{im\phi}e^{-ik(\lambda+\mu)}\ {}_1F_1(3/2,2,2ik\lambda)_1F_1(3/2,2,2ik\mu)e^{-ikt}.$$

As ${}_1F_1(3/2,2,2i\zeta)\equiv[J_0(\zeta)+iJ_1(\zeta)]e^{i\zeta}$, where J_0,J_1 are the Bessel functions of order zero and one, we have

$$\alpha_Z=\sqrt{\lambda\mu}e^{i\phi}[(J_0+iJ_1)(k\lambda)][(J_0+iJ_1)(k\mu)]e^{-ikt}, \tag{A4}$$

where we have written $(J_0+iJ_1)(k\lambda)$ as shorthand for $J_0(k\lambda)+iJ_1(k\lambda)$. For large λ, $\sqrt{\lambda}(J_0+iJ_1)(k\lambda)\simeq\sqrt{2/\pi k}e^{ik\lambda-\pi/4}$, so that α_Z is everywhere bounded. The potential β_Z such that the constraint (9) is satisfied is

$$\beta_Z=\sqrt{\lambda\mu}e^{i\phi}[(J_0-iJ_1)(k\lambda)][(J_0-iJ_1)(k\mu)]e^{-ikt}. \tag{A5}$$

We can determine the corresponding Debye potential ψ_Z, such that $i(\partial_r r\psi_Z-\partial_t r\psi_Z)=\alpha_Z$

$$\psi_Z=\frac{1}{ik}\left(\frac{\sqrt{\lambda\mu}}{\lambda+\mu}\right)e^{i\phi}[J_0(k\lambda)J_1(k\mu)+J_1(k\lambda)J_0(k\mu)]. \tag{A6}$$

The field $\mathbf{F}_Z$ may be calculated by substituting (A4), (A5) into (8), or (A6) into (3). After some labor one obtains the components of $\mathbf{F}_Z$

$$2\begin{bmatrix}F_1\\F_2\\F_3\end{bmatrix}_Z=\frac{e^{-ikt}}{(\lambda+\mu)}\left[\begin{pmatrix}\lambda[(J_0+iJ_1)(k\lambda)J_0(k\mu)]+\mu[J_0(k\lambda)(J_0-iJ_1)(k\mu)]\\ i\lambda[(J_0+iJ_1)(k\lambda)J_0(k\mu)]+i\mu[J_0(k\lambda)(J_0-iJ_1)(k\mu)]\\ -2i\sqrt{\lambda\mu}e^{i\phi}[J_0(k\lambda)J_1(k\mu)+J_1(k\lambda)J_0(k\mu)]\end{pmatrix}\right.$$
$$\left.+e^{2i\phi}\begin{pmatrix}-i\mu[J_1(k\lambda)(J_0+iJ_1)(k\mu)]+i\lambda[(J_0-iJ_1)(k\lambda)J_1(k\mu)]\\ -\mu[J_1(k\lambda)(J_0+iJ_1)(k\mu)]+\lambda[(J_0-iJ_1)(k\lambda)]J_1(k\mu)\\ 0\end{pmatrix}\right]. \tag{A7}$$

It is interesting to see the behavior of this field $\mathbf{F}_Z$ along the x_3 axis. For the positive x_3 axis $\mu=0$ and $\lambda=x_3$, and the field is just

$$2(\mathbf{F}_Z)_{\mu=0}=e^{-ikt}\begin{pmatrix}(J_0+iJ_1)(kx_3)\\ i(J_0+iJ_1)(kx_3)\\ 0\end{pmatrix}\xrightarrow{kx_3\gg1}e^{-i\pi/4}\sqrt{\frac{2}{\pi|x_3|}}\begin{pmatrix}1\\ i\\ 0\end{pmatrix}e^{ikx_3-ikt}. \tag{A8}$$

For the negative x_3 axis $\lambda=0$ and $\mu=-x_3$, and we obtain the same expression (A8) for the field along the negative x_3 axis. Along the x_3 axis the field $\mathbf{F}_Z$ appears as a right-circularly polarized plane wave propagating in the $+x_3$ direction, with amplitude decaying by $|x_3|^{-1/2}$ from the origin. Away from the x_3 axis the field decays more rapidly: the x_1,x_2 plane through the origin is parametrized by $\lambda=\mu=\sqrt{x_1^2+x_2^2}/2$, then substituting $\lambda=\mu=\rho$ into (A7) one finds that the field decays by ρ^{-1} from the origin. Thus the field $\mathbf{F}_Z$ is highly directional along the axis of propagation.

When $m=1$, $c>1/2$, the resulting field becomes infinite at spatial infinity.

APPENDIX B: ORTHONORMALITY OF THE BASIS FUNCTIONS $\boldsymbol{U}(\mathbf{k})$

First we will consider the orthogonality of two left-circularly polarized waves, $U_L(\mathbf{k})$, $U_L(\mathbf{k}')$. Due to the fact that the $U_L(\mathbf{k}), U_L(\mathbf{k}')$ are eigenstates of the angular momentum in the direction of propagation $\mathbf{k}$ or $\mathbf{k}'$, and the $\mathbf{L}$ operator is Hermitian, then

$$\langle U_L(\mathbf{k})|U_L(\mathbf{k}')\rangle = 0 \quad \text{when } \hat{\mathbf{k}} \neq \hat{\mathbf{k}}'.$$

So we need only consider the case when the waves are propagating in the same direction but with different frequencies, i.e., when $\mathbf{k}=\{0,0,k\}$ and $\mathbf{k}'=\{0,0,k'\}$. Then recalling the U_L of (19) we have the inner product

$$\begin{aligned}
&\langle U_L(\mathbf{k})|U_L(\mathbf{k}')\rangle \\
&= -\frac{1}{4}\int d^3x\left[\left(-\left(\frac{\hat{x}_1+i\hat{x}_2}{1-\hat{x}_3}\right)[e^{ikx_3}-e^{ikr}]-\left(\frac{\hat{x}_1+i\hat{x}_2}{1+\hat{x}_3}\right)[e^{ikx_3}-e^{-ikr}]\right)(k'(\hat{x}_1-i\hat{x}_2)e^{-ikx_3})\right. \\
&\quad \left.+\left(-\left(\frac{\hat{x}_1-i\hat{x}_2}{1-\hat{x}_3}\right)[e^{-ik'x_3}-e^{-ik'r}]-\left(\frac{\hat{x}_1-i\hat{x}_2}{1-\hat{x}_3}\right)[e^{-ik'x_3}-e^{ik'r}]\right)(k(\hat{x}_1+i\hat{x}_2)e^{ikx_3})\right. \\
&= \frac{1}{4}\int d^3x[k'((1+\hat{x}_3)[e^{i(k-k')x_3}-e^{i(kr-k'x_3)}]+(1-\hat{x}_3)[e^{i(k-k')x_3}-e^{-ikr-ik'x_3}]) \\
&\quad +k((1+\hat{x}_3)[e^{i(k-k')x_3}-e^{-i(k'r-kx_3)}]+(1-\hat{x}_3)[e^{i(k-k')x_3}-e^{i(k'r+ikx_3)}])] \\
&= \frac{1}{4}\int d^3x[2(k+k')e^{i(k-k')x_3}-(1+\hat{x}_3)[k'e^{i(kr-k'x_3)}+ke^{-i(k'r-kx_3)}] \\
&\quad -(1-\hat{x}_3)[k'e^{-ikr-ik'x_3}+ke^{i(k'r+ikx_3)}]].
\end{aligned}$$

The first term yields the delta function $(2\pi)^3 k\,\delta(\mathbf{k}-\mathbf{k}')$. We go over to spherical coordinates to evaluate the remaining terms, putting $\hat{x}_3=\cos\theta=\nu$, the rest of the integral is

$$\begin{aligned}
&-\frac{2\pi}{4}\int_0^\infty r^2dr\int_{-1}^1 d\nu((1+\nu)[k'e^{ir(k-k'\nu)}+ke^{-ir(k'-k\nu)}]+(1-\nu)[k'e^{-ir(k+k'\nu)}+ke^{ir(k'+k\nu)}]) \\
&= -\frac{4\pi}{4}\int_0^\infty r^2dr\int_{-1}^1 d\nu((1+\nu)[k'\cos(r(k-k'\nu))+k\cos(r(k'-k\nu))]) \\
&= -\frac{4\pi}{4}\int_0^\infty r^2dr(\partial_k-\partial_{k'})\int_{-1}^1 d\nu\left(\frac{1}{r}[k'\sin(r(k-k'\nu))-k\sin(r(k'-k\nu))]\right) \\
&= -\frac{4\pi}{4}\int_0^\infty r^2dr(\partial_k-\partial_{k'})\left|\int_{-1}^1 \frac{1}{r^2}[\cos(r(k-k'\nu))-\cos(r(k'-k\nu))]\right| = 0,
\end{aligned}$$

so that finally

$$\langle U_L(\mathbf{k})|U_L(\mathbf{k}')\rangle = (2\pi)^3 k\,\delta(\mathbf{k}-\mathbf{k}')$$

with $k\equiv|\mathbf{k}|$. The other identities of (31) follow similarly.

[1] A. C. T. Wu, ''Debye scalar potentials for the electromagnetic fields,'' Phys. Rev. D **34**, 3109–3110 (1986).
[2] C. G. Gray, ''Multipole expansions of electromagnetic fields using Debye potentials,'' Am. J. Phys. **46**, 169–179 (1978).
[3] See the notes of Ref. 2 for the original work on the Debye potentials.
[4] J. R. Reitz, F. J. Milford, and R. W. Christy, *Foundations of Electromagnetic Theory* (Addison-Wesley, Reading, MA, 1993).
[5] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed.
[6] C. J. Bouwkamp and H. G. B. Casimir, ''On multipole expansions in the theory of electromagnetic radiation,'' Physica (Amsterdam) **20**, 539–554 (1954).
[7] C. H. Wilcox, ''Debye potentials,'' J. Math. Mech. **20**, 167–201 (1957).

[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Vol. 1, pp. 378.
[9] See, for example, pp. 767–769 of Ref. 5.
[10] S. N. Mosley, ''Electromagnetics in retarded time and photon localization,'' Am. J. Phys. **65**, 1094–1097 (1997).
[11] See pp. 245–251 of Ref. 5.
[12] R. H. Good and T. J. Nelson, *Classical Theory of Electric and Magnetic Fields* (Academic, New York, 1971), pp. 609.
[13] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961), pp. 117.
[14] I. Bialynicki-Birula, ''The photon wavefunction,'' in *Progress in Optics*, edited by E. Wolf (Elsevier, Amsterdam, 1996), Vol. 36, pp. 246–294.
[15] I. Bialynicki-Birula and Z. Bialynicka-Birula, *Quantum Electrodynamics* (Pergamon, Oxford, 1975), pp. 142.
[16] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1971), 3rd revised ed., pp. 119–123.
[17] Any complex vector $\mathbf{C}$ can be split into three components with the following set of projection operators: $\mathbf{P}_{\uparrow}\equiv\frac{1}{2}(1-\hat{\mathbf{x}}\hat{\mathbf{x}}\cdot+i\hat{\mathbf{x}}\times)$, $\mathbf{P}_0\equiv\hat{\mathbf{x}}(\hat{\mathbf{x}}\cdot)$, $\mathbf{P}_{\downarrow}\equiv\frac{1}{2}(1-\hat{\mathbf{x}}\hat{\mathbf{x}}\cdot-i\hat{\mathbf{x}}\times)$, satisfying $\mathbf{P}_{\epsilon}\mathbf{P}_{\epsilon'}=\delta_{\epsilon\epsilon'}\mathbf{P}_{\epsilon}$, with $\epsilon=\{\uparrow,0,\downarrow\}$. The transverse components $(\mathbf{P}_{\uparrow}\mathbf{C})$,$(\mathbf{P}_{\downarrow}\mathbf{C})$ can then each be expressed in terms of a complex potential by $(\mathbf{P}_{\uparrow}\mathbf{C})=\boldsymbol{\nabla}_{\uparrow}\chi\equiv\mathbf{P}_{\uparrow}\boldsymbol{\nabla}\chi$ and $(\mathbf{P}_{\downarrow}\mathbf{C})=\boldsymbol{\nabla}_{\downarrow}\zeta\equiv\mathbf{P}_{\downarrow}\boldsymbol{\nabla}\zeta$. See the paper by Wilcox (Ref. 7) for equivalent methods of expressing a transverse vector $(\mathbf{P}_{\uparrow}\mathbf{C}+\mathbf{P}_{\downarrow}\mathbf{C})$ in terms of two potentials.
[18] R. Resnick, *Introduction to Special Relativity* (Wiley, New York, 1968), pp. 169.
[19] D. K. Monroe, ''Lorentz transformation of Debye potentials,'' J. Math. Phys. **25**, 1787–1790 (1984).
[20] E. Witten, ''Duality, spacetime and quantum mechanics,'' Phys. Today **50**, 28–33 (1997).
[21] H. Buchholz, *The Confluent Hypergeometric Function* (Springer-Verlag, Berlin, 1969), pp. 50.
[22] H. Hochstadt, *The Functions of Mathematical Physics* (Wiley Interscience, New York, 1971), pp. 192.

Hamiltonian time-dependent mechanics

Gennadi A. Sardanashvily
Department of Theoretical Physics, Physics Faculty, Moscow State University, 117234 Moscow, Russia

(Received 20 October 1997; accepted for publication 28 January 1998)

The usual formulation of time-dependent mechanics implies a given splitting $Y=\mathbf{R}\times M$ of an event space Y. This splitting, however, is broken by any time-dependent transformation, including transformations between inertial frames. The goal is the frame-covariant formulation of time-dependent mechanics on a bundle $Y\to\mathbf{R}$, whose fibration $Y\to M$ is not fixed. Its phase space is the vertical cotangent bundle V^*Y, provided with the canonical 3-form and the corresponding canonical Poisson structure. An event space of relativistic mechanics is a manifold Z whose fibration $Z\to\mathbf{R}$ is not fixed. © *1998 American Institute of Physics.*
[S0022-2488(98)02705-4]

I. INTRODUCTION

The symplectic technique is well known to provide an adequate mathematical formulation of autonomous mechanics, when Hamiltonians are independent of time.[1–4] Its canonical example is a mechanical system whose phase space is the cotangent bundle T^*M of an event manifold M. This phase space is provided with the canonical symplectic form

$$\Omega=dp_i\wedge dy^i,$$

written with respect to the holonomic coordinates $(y^i,\ p_i=\dot q_i)$ on T^*M. A Hamiltonian $\mathscr{H}$ is defined as a real function on T^*M. The motion trajectories are integral curves of the Hamiltonian vector field $\vartheta=\vartheta_i\partial^i+\vartheta^i\partial_i$ on T^*M, which obeys the Hamilton equations

$$\begin{gathered}\vartheta\rfloor\Omega=d\mathscr{H},\\ \vartheta^i=\partial^i\mathscr{H},\quad \vartheta_i=-\partial_i\mathscr{H}.\end{gathered}\tag{1}$$

The usual formulation of time-dependent mechanics implies a splitting $Y=\mathbf{R}\times M$ of the event manifold Y and the corresponding splitting $\mathbf{R}\times T^*M$ of the phase space.[5–10] The phase space is provided with the pull-back $\mathrm{pr}_2^*\Omega$ of the symplectic form on T^*M. By a time-dependent Hamiltonian is meant a real function on $\mathbf{R}\times T^*M$, while motion trajectories are integral curves of the time-dependent Hamiltonian vector field,

$$\vartheta:\mathbf{R}\times T^*M\to TT^*M,$$

which obeys the Hamilton equations (1). The problem is that the above-mentioned splittings are broken by any time-dependent transformation, including the inertial frame transformations. Therefore, the form $\mathrm{pr}_2^*\Omega$ on the phase space of time-dependent mechanics fails to be canonical.[11]

We will formulate first-order time-dependent mechanics as a particular field theory, when an event space is a fibered manifold $Y\to\mathbf{R}$, coordinated by (t,y^i).[5,12–14] Unless otherwise stated, the base $\mathbf{R}$ is parametrized by the coordinates t with transition functions $t'=t+\text{const}$. Relative to these coordinates, $\mathbf{R}$ is equipped with the standard vector field ∂_t and the standard 1-form dt, which is also the volume element on $\mathbf{R}$. This is not the case for relativistic mechanics.

The configuration space of time-dependent mechanics is the first-order jet manifold J^1Y of sections of $Y\to\mathbf{R}$, which is provided with the adapted coordinates (t,y^i,y^i_t). There is the canonical monomorphism,

$$\lambda:J^1Y\hookrightarrow TY,\quad \lambda=\partial_t+y^i_t\partial_i,$$

 © 1998 American Institute of Physics

over Y. It is easy to see that $\pi^1_0: J^1Y \to Y$ is an affine bundle modeled over the vertical tangent bundle $VY \to Y$. For the sake of simplicity, we will identify J^1Y with the corresponding subbundle of TY. In particular, every connection on a bundle $Y \to \mathbf{R}$ can be identified with the horizontal lift,

$$\Gamma: Y \to J^1Y \subset TY, \quad \Gamma = \partial_t + \Gamma^i \partial_i, \tag{2}$$

of the standard vector field ∂_t by this connection. We will call (2) a connection in order to refer to the standard properties of connections without an additional explanation.

The one-dimensional reduction of the polysymplectic Hamiltonian formalism[13–16] provides the adequate mathematical formulation of time-dependent Hamiltonian mechanics on the Legendre bundle $\pi_\Pi: \Pi = V^*Y \to Y$. The phase space V^*Y is endowed with the canonical 3-form,

$$\mathbf{\Omega} = dp_i \wedge dy^i \wedge dt, \tag{3}$$

written with respect to the holonomic coordinates $(t, y^i, p_i = \dot{y}_i)$ on V^*Y.

The following peculiarities of time-dependent Hamiltonian mechanics should be emphasized.

(i) The form $\mathbf{\Omega}$ (3) defines the canonical degenerate Poisson structure on the phase space V^*Y.

(ii) A Hamiltonian is not a function on a phase space. As a consequence, the evolution equation is not reduced to a Poisson bracket, and integrals of motion cannot be defined as functions in involution with a Hamiltonian.

(iii) Hamiltonian and Lagrangian formulations of time-dependent mechanics are equivalent in the case of hyperregular Lagrangians. A degenerate Lagrangian requires a set of associated Hamiltonians and Hamilton equations in order to exhaust all solutions of the Lagrange ones.

(iv) A complete connection Γ on the event space $Y \to \mathbf{R}$ defines a reference frame so that one can think of the difference $y^i_t - \Gamma^i(y)$ as being the velocity relative to the reference frame Γ. There is one-to-one correspondence between the complete connections Γ and the trivializations of $Y \to \mathbf{R}$.

For the sake of simplicity, $Y \to \mathbf{R}$ is assumed to be a bundle with a typical fiber M. It is trivial. Different trivializations,

$$Y \cong \mathbf{R} \times M, \tag{4}$$

differ from each other in fibrations $Y \to M$, while the fibration $\pi: Y \to \mathbf{R}$ is once for all. Given a trivialization (4), there are the corresponding splittings of the configuration and phase spaces,

$$J^1Y \cong \mathbf{R} \times TM, \quad V^*Y \cong \mathbf{R} \times T^*M.$$

If a fibration $Z \to \mathbf{R}$ of an event space Z is not fixed, we obtain the general formulation of relativistic mechanics, including Special Relativity on the Minkowski space $Z = \mathbf{R}^4$. Its configuration space is the first-order jet manifold $J^1_1 Z$ of one-dimensional submanifolds of Z. The bundle $J^1_1 Z \to Z$ is projective, and one can think on its fibers as being spaces of nonrelativistic velocities. Relativistic velocities are represented by elements of the tangent bundle TZ, while the cotangent bundle T^*Z plays the role of the phase space of relativistic mechanics.

All manifolds throughout are assumed to be paracompact and connected.

II. CANONICAL POISSON STRUCTURE

The Legendre bundle V^*Y of time-dependent mechanics is provided with the canonical Poisson structure as follows. Let us consider the cotangent bundle T^*Y with the holonomic coordinates (t, y^i, p, p_i), which is the homogeneous Legendre bundle of time-dependent mechanics. It admits the canonical Liouville form

$$\Xi = p\,dt + p_i dy^i, \tag{5}$$

and the canonical symplectic form

$$d\Xi = dp \wedge dt + dp_i \wedge dy^i.$$

The corresponding Poisson bracket on the space $C^\infty(T^*Y)$ of functions on T^*Y reads as

$$\{f,g\}=\partial^p f\partial_t g-\partial^p g\partial_t f+\partial^i f\partial_i g-\partial^i g\partial_i f. \tag{6}$$

Let us consider the subspace of $C^\infty(T^*Y)$ that comprises the pull-backs of functions on V^*Y by the projection $T^*Y\to V^*Y$. This subspace is closed under the Poisson bracket (6). Then there exists the canonical Poisson structure,

$$\{f,g\}_V=\partial^i f\partial_i g-\partial^i g\partial_i f, \tag{7}$$

on V^*Y induced by (6).[3] The corresponding Poisson bivector,

$$w(df,dg)=\{f,g\}_V,$$

on V^*Y is vertical with respect to the fibration $V^*Y\to\mathbf{R}$, and reads as

$$w^{ij}=0,\quad w_{ij}=0,\quad w^i_j=1.$$

A glance at this expression shows that the holonomic coordinates on V^*Y are canonical for the Poisson structure (7), which is regular and degenerate.

Given the Poisson bracket (7), the Hamiltonian vector field ϑ_f of a function f on V^*Y, defined by the relation $\{f,g\}_V=\vartheta_f\rfloor dg$, $\forall g\in C^\infty(V^*Y)$, is the vertical vector field

$$\vartheta_f=\partial^i f\partial_i-\partial_i f\partial^i, \tag{8}$$

on $V^*Y\to\mathbf{R}$. Hence, the characteristic distribution of the Poisson structure (7), generated by Hamiltonian vector fields, is precisely the vertical tangent bundle VV^*Y of $V^*Y\to\mathbf{R}$.

By virtue of the well-known theorem,[4] the Poisson structure (7) defines the symplectic foliation on V^*Y that coincides with the fibration $V^*Y\to\mathbf{R}$. The symplectic forms on the fibers of $V^*Y\to\mathbf{R}$ are the pull-backs,

$$\Omega_t=dp_i\wedge dy^i,$$

of the canonical symplectic form on the typical fiber T^*M of $V^*Y\to\mathbf{R}$ with respect to trivialization morphisms.[11]

The Poisson structure (7) can be introduced in a different way. The Legendre bundle V^*Y admits the canonical closed 3-form (3), which is the particular case of the polysymplectic form.[14–16] Then every function f on V^*Y defines the corresponding Hamiltonian vector field ϑ_f (8) by the relation

$$\vartheta_f\rfloor\mathbf{\Omega}=-df\wedge dt,$$

while the Poisson bracket (7) is recovered by the condition

$$\{f,g\}_V dt=\vartheta_g\rfloor\vartheta_f\rfloor\mathbf{\Omega}.$$

III. HAMILTONIAN FORMS

In comparison with autonomous mechanics, Hamiltonian time-dependent mechanics is formulated in the terms of Hamiltonian connections and Hamiltonian forms.

We will say that a connection,

$$\gamma=\partial_t+\gamma^i\partial_i+\gamma_i\partial^i, \tag{9}$$

on the phase space $V^*Y\to\mathbf{R}$ is locally Hamiltonian if the exterior form $\gamma\rfloor\mathbf{\Omega}$ is closed, i.e.,

$$\mathbf{L}_\gamma\mathbf{\Omega}=d(\gamma\rfloor\mathbf{\Omega})=0, \tag{10}$$

where $\mathbf{L}$ denotes the Lie derivative.

For instance, every connection Γ on the event space $Y\to\mathbf{R}$ gives rise to the locally Hamiltonian connection,

$$\widetilde{\Gamma}=\partial_t+\Gamma^i\partial_i-p_i\partial_j\Gamma^i\partial^j,$$

on the phase space, such that

$$\widetilde{\Gamma}\rfloor\mathbf{\Omega}=dH_\Gamma,\quad H_\Gamma=p_i dy^i-p_i\Gamma^i dt. \tag{11}$$

Locally Hamiltonian connections constitute an affine space modeled over the linear space of vertical vector fields ϑ on $V^*Y\to\mathbf{R}$ that obey the same condition (10), and are locally Hamiltonian vector fields as follows.

Lemma 1: Every closed form $\gamma\rfloor\mathbf{\Omega}$ on $V^*Y\to\mathbf{R}$ is exact.

Proof: Let us consider the decomposition

$$\gamma=\widetilde{\Gamma}+\vartheta, \tag{12}$$

where Γ is a connection on $Y\to\mathbf{R}$, while ϑ satisfies the relation $d(\partial\rfloor\mathbf{\Omega})=0$. It is easily seen that $\vartheta\rfloor\mathbf{\Omega}=\sigma\wedge dt$, where σ is a 1-form. Using properties of the De Rham cohomology groups of a manifold product, one can show that every closed 2-form $\sigma\wedge dt$ on V^*Y is exact, and so is $\gamma\rfloor\mathbf{\Omega}$. Moreover, in accordance with the relative Poincaré lemma, we can write locally $\vartheta\rfloor\mathbf{\Omega}=df\wedge dt$.

Definition 2: A 1-form H on the Legendre bundle V^*Y is called locally Hamiltonian if

$$\gamma\rfloor\mathbf{\Omega}=dH,$$

for a connection γ on $V^*Y\to\mathbf{R}$.

By virtue of Proposition 1, there is one-to-one correspondence between the locally Hamiltonian connections and the locally Hamiltonian forms considered throughout modulo closed forms.

Definition 3: By a Hamiltonian form H on the phase space V^*Y is meant the pull-back,

$$H=h^*\Xi=p_i dy^i-\mathscr{H}dt, \tag{13}$$

of the Liouville form Ξ (5) on T^*Y by a section h of the bundle

$$\zeta: T^*Y\to V^*Y. \tag{14}$$

Given a trivialization (4) of $Y\to\mathbf{R}$ the Hamiltonian form (13) looks like the well-known Poincaré–Cartan integral.[2] However, the Hamiltonian $\mathscr{H}$ in the expression (13) is not a function as follows. Any connection Γ on $Y\to\mathbf{R}$ defines the Hamiltonian form H_Γ (11) on V^*Y, and every Hamiltonian form on V^*Y admits the splitting

$$H=p_i dy^i-(p_i\Gamma^i+\widetilde{\mathscr{H}}_\Gamma)dt, \tag{15}$$

where Γ is a connection on $Y\to\mathbf{R}$ and $\widetilde{\mathscr{H}}_\Gamma$ is a real function on V^*Y, called the Hamiltonian function.

Proposition 4: Locally Hamiltonian forms are Hamiltonian forms locally.

Proof: Given locally Hamiltonian forms H_γ and $H_{\gamma'}$, their difference,

$$\sigma=H_\gamma-H_{\gamma'},\quad d\sigma=(\gamma-\gamma')\rfloor\mathbf{\Omega},$$

is a 1-form on V^*Y such that the 2-form $\sigma\wedge dt$ is closed and, consequently, exact. In accordance with the relative Poincaré lemma, this condition implies that $\sigma=f dt+dg$, where f and g are local functions on V^*Y. Then it follows from the splitting (12) that, in a neighborhood of every point $p\in V^*Y$, a locally Hamiltonian form H_γ coincides with the pull-back of the Liouville form Ξ on T^*Y by the local section

$$(t,y^i,p_i)\mapsto(t,y^i,p_i,p=-p_i\Gamma^i+f\ )$$

of $T^*Y\to V^*Y$, where f is a local function on V^*Y.

Proposition 5: Conversely, let H be a Hamiltonian form H on the Legendre bundle V^*Y. There exists a unique connection γ_H on $V^*Y\to\mathbf{R}$, called the Hamiltonian connection, such that $\gamma_H\rfloor\mathbf{\Omega}=dH$.

Proof: Let us introduce the Hamilton operator,

$$\mathscr{E}_H=dH-\lambda\rfloor\mathbf{\Omega}=[(y^i_t-\partial^i\cdot\mathscr{H})dp_i-(p_{ti}+\partial_i\cdot\mathscr{H})dy^i]\wedge dt,$$

on the phase space V^*Y. It is readily observed that this operator is an image of the global section,

$$\gamma_H=\partial_t+\partial^i\cdot\mathscr{H}\partial_i-\partial_i\cdot\mathscr{H}\partial^i, \tag{16}$$

of the jet bundle $J^1V^*Y\to V^*Y$. This is the unique solution of the first-order differential Hamilton equations,

$$y^i_t=\partial^i\cdot\mathscr{H},\quad p_{ti}=-\partial_i\cdot\mathscr{H}, \tag{17}$$

on V^*Y, and is a Hamiltonian connection for the Hamiltonian form H.

The integral curves of the Hamiltonian connection (16) are classical solutions of the Hamilton equations (17). Conversely, since $r(\mathbf{R})\subset V^*Y$ is closed, every classical solution $r:\mathbf{R}\to V^*Y$ of the Hamilton equations (17) can be extended to a unique Hamiltonian connection for H.

Hamiltonian connections γ_H (16) form an affine space modeled over the linear space of Hamiltonian vector fields (8).

Remark: The Hamilton equations (17) can be introduced without appealing to the Hamilton operator. They are equivalent to the relation

$$r^*(u\rfloor dH)=0,$$

which is assumed to hold for any vertical vector field u on $V^*Y\to\mathbf{R}$.

With a Hamiltonian form H (15) and the corresponding Hamiltonian connection γ_H (16), we have the Hamilton evolution equation,

$$d_{Ht}f=\gamma_H\rfloor df=(\partial_t+\partial^i\cdot\mathscr{H}\partial_i-\partial_i\cdot\mathscr{H}\partial^i)f, \tag{18}$$

on functions on the Legendre bundle V^*Y. Substituting a classical solution of the Hamilton equations (17) in (18), we obtain the time evolution of the function f. Given the splitting (15) of a Hamiltonian form H, the Hamilton evolution equation (18) is brought into the form

$$d_{Ht}f=\partial_t f+(\Gamma^i\partial_i-\partial_i\Gamma^j p_j\partial^i)f+\{\widetilde{\mathscr{H}}_\Gamma,f\}_V. \tag{19}$$

A glance at this expression shows that the Hamilton evolution equation in time-dependent mechanics does not reduce to the Poisson bracket. This fact may be relevant to the quantization problem. Under quantization, the Poisson bracket on the right-hand side of the equation (19) becomes the operator one, while the second term remains classical and depends on the choice of a reference frame.

IV. CANONICAL TRANSFORMATIONS

Canonical transformations in time-dependent mechanics are not compatible with the fibration $V^*Y\to Y$, in general.

Definition 6: By a canonical automorphism is meant an automorphism ρ over $\mathbf{R}$ of the bundle $V^*Y\to\mathbf{R}$ that preserves the canonical Poisson structure (7) on V^*Y, i.e.,

$$\{f\circ\rho,g\circ\rho\}_V=\{f,g\}_V\circ\rho,$$

and, equivalently, the canonical form $\mathbf{\Omega}$ (3) on V^*Y, i.e., $\mathbf{\Omega}=\rho^*\mathbf{\Omega}$.

The bundle coordinates on $V^*Y\to\mathbf{R}$ are called canonical if they are canonical for the Poisson structure (7). Canonical coordinate transformations satisfy the relations

$$\frac{\partial p'_i}{\partial p_j}\frac{\partial y'^i}{\partial p_k}-\frac{\partial p'_i}{\partial p_k}\frac{\partial y'^i}{\partial p_j}=0, \quad \frac{\partial p'_i}{\partial y^j}\frac{\partial y'^i}{\partial y^k}-\frac{\partial p'_i}{\partial y^k}\frac{\partial y'^i}{\partial y^j}=0, \quad \frac{\partial p'_i}{\partial p_k}\frac{\partial y'^i}{\partial y^j}-\frac{\partial p'_i}{\partial y^j}\frac{\partial y'^i}{\partial p_k}=\delta^k_j.$$

By definition, the holonomic coordinates on V^*Y are the canonical ones. Accordingly, holonomic automorphisms,

$$(y^i,p_i)\mapsto\left(y'^i,p'_i=\frac{\partial y^j}{\partial y'^i}\,p_j\right)$$

of the phase space $V^*Y\to Y$ induced by the vertical automorphisms of $Y\to\mathbf{R}$ are also canonical.

Proposition 7: Canonical automorphisms send locally Hamiltonian connections onto the locally Hamiltonian ones (and, consequently, locally Hamiltonian forms onto each other).

Proof: If γ is a locally Hamiltonian connection for H, we have

$$T\rho(\gamma)\rfloor\mathbf{\Omega}=(\rho^{-1})^*(\gamma\rfloor\mathbf{\Omega})=d((\rho^{-1})^*H).$$

It should be emphasized that, in general, canonical automorphisms do not send Hamiltonian forms onto Hamiltonian forms, but only locally.

Proposition 8: Let γ be a complete locally Hamiltonian connection on $V^*Y\to\mathbf{R}$, i.e., the vector field (9) is complete. There exist canonical coordinate transformations that bring all components of γ to zero, i.e., $\gamma=\partial_t$.

Proof: A glance at the relation (10) shows that each locally Hamiltonian connection γ is the generator of a local one-parameter group G_γ of canonical automorphisms of the phase space $V^*Y\to\mathbf{R}$. Let V^*_0Y be the fiber of $V^*Y\to\mathbf{R}$ at $0\in\mathbf{R}$. Then canonical coordinates of V^*_0Y dragged along integral curves of the complete vector field γ satisfy the statement of the proposition.

In particular, let H be a Hamiltonian form (15) such that the corresponding Hamiltonian connection γ_H (16) is complete. By virtue of Proposition 8, there exist canonical coordinate transformations that bring the Hamiltonian $\mathscr{H}$ into zero. Then the corresponding Hamilton equations reduce to the equilibrium ones. Accordingly, any Hamiltonian for H can be locally brought into the form where $\mathscr{H}=0$ by local canonical coordinate transformations.

Every canonical transformation admits a local generating function as follows. Let H be a Hamiltonian form (13) on V^*Y. Given a canonical automorphism ρ, we have

$$d(\rho^*H-H)=0,$$

$$\rho^*H-H=dS,$$

where S is a local function on V^*Y. We can write locally

$$\rho^*H=\rho_i d\rho^i-\mathscr{H}\circ\rho dt.$$

Then the corresponding coordinate relations read as

$$\partial_i S=\rho_j\partial_i\rho^j-p_i, \quad \partial^i S=\rho_j\partial^i\rho^j, \quad \mathscr{H}'-\mathscr{H}=\rho_i\partial_t\rho^i-\partial_t S.$$

Taken on the graph,

$$\Delta_\rho=\{(p,\rho(p))\in V^*Y\times V^*Y\},$$

of the canonical automorphism, the function S plays the role of a local generating function. For instance, if the graph Δ_ρ is coordinated by (t,y^i,y'^i), we obtain the familiar expression $\mathscr{H}'-\mathscr{H}=\partial_t S(t,y^i,y'^i)$.

V. REFERENCE FRAMES

Every connection Γ on the bundle $\pi:Y\to\mathbf{R}$ defines a horizontal foliation on $Y\to\mathbf{R}$ whose leaves are the integral curves of the nowhere vanishing vector field (2). Conversely, let Y admit a horizontal foliation such that, for each point $y\in Y$, the leaf of this foliation through y is locally determined by a section s_y of $V^*Y\to\mathbf{R}$ through y. Then, the map

$$\Gamma: Y \to J^1Y, \quad \Gamma(y) = j^1_t s_y, \quad \pi(y) = t,$$

is well defined. This is a connection on $Y \to \mathbf{R}$.

Given a horizontal foliation on Y, there exists the associated atlas of constant local trivializations of Y such that every leaf of this foliation is locally generated by the equations $y^i = \text{const}$, and the transition functions $y^i \to y'^i(y^j)$ are independent of the coordinate t.[16,17] Two such atlases are said to be equivalent if their union is also an atlas of constant local trivializations. They are associated with the same horizontal foliation. Thus, we have proved the following assertion.

Proposition 9: There is one-to-one correspondence between the connections Γ on $Y \to \mathbf{R}$ and the equivalence classes of atlases of constant local trivializations of Y such that $\Gamma^i = 0$ relative to the associated coordinates, called adapted to Γ.

Proposition 10: Every trivialization of $Y \to \mathbf{R}$ yields a complete connection on this bundle. Conversely, every complete connection on $Y \to \mathbf{R}$ defines a trivialization $Y \cong \mathbf{R} \times M$ such that the associated coordinates are adapted to Γ.

Proof: Every trivialization of $Y \to \mathbf{R}$ defines the horizontal lift $\Gamma = \partial_t$ onto Y of the standard field ∂_t on $\mathbf{R}$, which is obviously a complete connection on $Y \to \mathbf{R}$. Conversely, let Γ be a complete connection on $Y \to \mathbf{R}$. This is the generator of the one-parameter group G_Γ that acts freely on Y. The orbits of this action are, of course, the integral curves of Γ. Hence, we obtain a projection,

$$\pi_\Gamma : Y \to Y/G_\Gamma = M.$$

This projection together with $\pi: Y \to \mathbf{R}$ defines a trivialization of Y.

One can say that a complete connection Γ on an event space $Y \to \mathbf{R}$ describes a reference frame in time-dependent mechanics. Given a reference frame Γ, we have the corresponding covariant differential,

$$D_\Gamma: \ J^1Y \to VY, \quad (t, y^i, y^i_t) \mapsto (t, y^i, \dot{y}^i = y^i_t - \Gamma^i).$$

Let s be a (local) section of $Y \to \mathbf{R}$. One can think of $D_\Gamma \circ J^1 s$ as being the relative velocity of the motion s with respect to the reference frame Γ. Indeed, $D_\Gamma \circ J^1 s$ vanishes identically iff s is an integral curve of Γ.

Let us consider the Hamilton evolution equation (19). For any connection Γ in the splitting (19), there exist holonomic canonical transformations of V^*Y to the coordinates adapted to Γ that bring (19) into the familiar Poisson bracket form,

$$d_{Ht} f = \partial_t f + \{\widetilde{\mathscr{H}}, f\}_V.$$

VI. LAGRANGIAN POISSON STRUCTURE

In contrast with the Legendre bundle V^*Y, the configuration space J^1Y of time-dependent mechanics does not possess any canonical Poisson structure, in general. A Poisson structure on J^1Y depends on the choice of a Lagrangian,

$$L = \mathscr{L} dt, \quad \mathscr{L}: \ J^1Y \to \mathbf{R}. \tag{20}$$

We will use the notation $\pi_i = \partial^t_i \mathscr{L}$, $\pi_{ij} = \partial^t_i \partial^t_j \mathscr{L}$.

Every Lagrangian L (20) defines the Legendre map,

$$\hat{L}: J^1Y \to V^*Y, \quad p_i \circ \hat{L} = \pi_i. \tag{21}$$

The pull-back on J^1Y of the canonical 3-form $\mathbf{\Omega}$ (3) by the Legendre map $\hat{L}$ (21) reads as

$$\mathbf{\Omega}_L = \hat{L}^* \mathbf{\Omega} = d\pi_i \wedge dy^i \wedge dt.$$

By means of $\mathbf{\Omega}_L$, every vertical vector field $\vartheta = \vartheta^i \partial_i + \dot{\vartheta}^i \partial^t_i$ on $J^1Y \to \mathbf{R}$ yields the 2-form

$$\vartheta \rfloor \mathbf{\Omega}_L = \{[\dot{\vartheta}^j \pi_{ji} + \vartheta^j (\partial_j \pi_i - \partial_i \pi_j)] dy^i - \vartheta^i \pi_{ji} dy^j_t\} \wedge dt.$$

This is one-to-one correspondence, if the Lagrangian L is regular. Indeed, given any 2-form $\phi=(\phi_i dy^i+\dot\phi_i dy^i_t)\wedge dt$ on J^1Y, the algebraic equations,

$$\dot\vartheta^j\pi_{ji}+\vartheta^j(\partial_j\pi_i-\partial_i\pi_j)=\phi_i, \quad -\vartheta^i\pi_{ji}=\dot\phi_j,$$

have a unique solution,

$$\vartheta^i=-(\pi^{-1})^{ij}\dot\phi_j, \quad \dot\vartheta^j=(\pi^{-1})^{ji}[\phi_i+(\pi^{-1})^{kn}\dot\phi_n(\partial_k\pi_i-\partial_i\pi_k)].$$

In particular, every function f on J^1Y determines a vertical vector field,

$$\vartheta_f=(\pi^{-1})^{ij}\partial^t_j f\partial_i-(\pi^{-1})^{ji}[\partial_i f+(\pi^{-1})^{kn}\partial^t_n f(\partial_k\pi_i-\partial_i\pi_k)]\partial^t_j, \tag{22}$$

on $J^1Y\to\mathbf{R}$ in accordance with the relation

$$\vartheta_f\rfloor\mathbf{\Omega}_L=-df\wedge dt.$$

Then the Poisson bracket,

$$\{f,g\}_L dt=\vartheta_g\rfloor\vartheta_f\rfloor\mathbf{\Omega}_L, \quad f,g\in C^\infty(J^1Y), \tag{23}$$

can be defined on functions on J^1Y, and reads as

$$\{f,g\}_L=(\pi^{-1})^{ij}(\partial^t_i f\partial_j g-\partial^t_i g\partial_j f)+(\partial_n\pi_k-\partial_k\pi_n)(\pi^{-1})^{ki}(\pi^{-1})^{nj}\partial^t_i f\partial^t_j g.$$

The vertical vector field ϑ_f (22) is the Hamiltonian vector field of the function f with respect to the Poisson structure (23).

In particular, if the Lagrangian L is hyperregular, that is, the Legendre map $\hat L$ is a diffeomorphism, the Poisson structure (23) is obviously isomorphic to the Poisson structure (7) on the phase space V^*Y.

The Poisson structure (23) defines the corresponding symplectic foliation on J^1Y that coincides with the fibration $J^1Y\to\mathbf{R}$. The symplectic form on the leaf J^1_tY of this foliation is $\Omega_t=d\pi_i\wedge dy^i$.[18]

We will see below that the Lagrangian counterpart of Hamiltonian forms is the Poincaré–Cartan form,

$$H_L=\pi_i dy^i-(\pi_i y^i_t-\mathscr{L})dt.$$

This is the unique Lepagian equivalent of a Lagrangian L that participates in the first variational formula. The Poincaré–Cartan form defines the morphism

$$\hat H_L: J^1Y\to T^*Y, \quad p\circ\hat H_L=\mathscr{L}-\pi_i y^i_t, \quad p_i\circ\hat H_L=\pi_i. \tag{24}$$

There is the relation $\hat L=\zeta\circ\hat H_L$, where ζ is the canonical projection (14).

Let

$$u=u^t\partial_t+u^i\partial_i, \quad u^t=0,1, \tag{25}$$

be a vector field on $Y\to\mathbf{R}$. The first variational formula provides the canonical decomposition of the Lie derivative,

$$\mathbf{L}_{J^1u}L=(J^1u\rfloor d\mathscr{L})dt=(u^t\partial_t+u^i\partial_i+d_t u^i\partial^t_i)\mathscr{L}dt, \qquad d_t=\partial_t+y^i_t\partial_i+\cdots, \tag{26}$$

in accordance with the variational task.[16,19] We have

$$J^1u\rfloor d\mathscr{L}=(u^i-u^t y^i_t)\mathscr{E}_i+d_t(u\rfloor H_L), \tag{27}$$

where

$$\mathscr{E}_L=(\partial_i-d_t\partial_i^t)\mathscr{L}dy^i\wedge dt \tag{28}$$

is the Euler–Lagrange operator for L. The kernel $\mathrm{Ker}\,\mathscr{E}_L\subset J^2Y$ of this operator defines the second–order Lagrange equations on Y,

$$(\partial_i-d_t\partial_i^t)\mathscr{L}=0. \tag{29}$$

Definition 11: A connection,

$$\xi=\partial_t+\xi^i\frac{\partial}{\partial^i}+\xi_t^i\frac{\partial}{\partial\hat{y}_t^i},$$

on the configuration space $J^1Y\to\mathbf{R}$ is said to be a Lagrangian connection for the Lagrangian L if it obeys the condition

$$\xi\rfloor\Omega_L=dH_L, \tag{30}$$

which takes the coordinate form

$$(\xi^i-y_t^i)\pi_{ij}=0,$$

$$\partial_i\mathscr{L}-\partial_t\pi_i-\xi^j\partial_j\pi_i-\xi_t^j\pi_{ij}+(\xi^j-y_t^j)\partial_i\pi_j=0,$$

relative to the adapted coordinates $(t,y^i,y_t^i,\hat{y}_t^i,y_{tt})$ on J^1J^1Y.

In order to clarify the meaning of (30), let us consider the Lagrangian,

$$\overline{L}=L+(\hat{y}_t^i-y_t^i)\pi_i dt,$$

on the repeated jet manifold J^1J^1Y. The corresponding Euler–Lagrange operator, called the Euler–Lagrange–Cartan one, reads as

$$\begin{aligned}\mathscr{E}_{\overline{L}}&=[(\partial_iL-\hat{d}_t\pi_i+\partial_i\pi_j(\hat{y}_t^j-y_t^j))dy^i+\partial_i^t\pi_j(\hat{y}_t^j-y_t^j)dy_t^i]\wedge dt,\\ \hat{d}_t&=\partial_t+\hat{y}_t^i\partial_i+y_{tt}^i\partial_i^t.\end{aligned} \tag{31}$$

Then the condition (30) is equivalent to the one $\mathrm{Im}\,\xi\subset\mathrm{Ker}\,\mathscr{E}_{\overline{L}}$, and leads to the first-order differential equations on the jet manifold J^1Y, called the Cartan equations,

$$\partial_i^t\pi_j(\hat{y}_t^j-y_t^j)=0,\quad \partial_iL-\hat{d}_t\pi_i+(\hat{y}_t^j-y_t^j)\partial_i\pi_j=0. \tag{32}$$

Integral curves of Lagrangian connections ξ for L provide classical solutions $\overline{s}:\mathbf{R}\to J^1Y$ of these equations.

The restriction of $\mathscr{E}_{\overline{L}}$ to the holonomic jet manifold J^2Y defines the first-order Euler–Lagrange operator whose kernel is the system of the first-order Lagrange equations,

$$\hat{y}_\lambda^i-y_\lambda^i=0,\quad (\partial_i-d_\lambda\partial_i^\lambda)\mathscr{L}=0. \tag{33}$$

These are equivalent to the second-order Lagrange equations (29), and represent their familiar first-order reduction.

It is easily seen that the first-order Lagrange equations (33) [and consequently the second-order ones (29)] are equivalent to the Cartan equations (32) on the integrable sections $\overline{s}=J^1s$ of $J^1Y\to\mathbf{R}$. They are completely equivalent to the Cartan equations in the case of regular Lagrangians.

VII. DEGENERATE SYSTEMS

Let us consider the relations between the solutions of Hamilton and Lagrange equations. Every Hamiltonian form H on the phase space V^*Y defines the Hamiltonian map

$$\hat{H}: V^*Y \to J^1Y, \quad y^i_t \circ \hat{H} = \partial^i \mathscr{H}.$$

Its jet prolongation reads as

$$J^1\hat{H}: J^1V^*Y \to J^1J^1Y, \quad (y^i_t, \hat{y}^i_t, y^i_{tt}) \circ J^1\hat{H} = (\partial^i \mathscr{H}, y^i_t, d_t\partial^i \mathscr{H}).$$

Following the general polysymplectic scheme, we say that a Hamiltonian form H on V^*Y is associated with a Lagrangian L if H obeys the conditions[14–16]

$$\hat{L} \circ \hat{H} \circ \hat{L} = \hat{L}, \tag{34a}$$

$$p_i \partial^i \mathscr{H} - \mathscr{H} = \mathscr{L} \circ \hat{H}. \tag{34b}$$

It follows from the condition (34a) that $\hat{L} \circ \hat{H}$ is the projection operator to $Q = \hat{L}(J^1Y) \subset V^*Y$, called the Lagrangian constraint space, and $\hat{H} \circ \hat{L}$ is the projection operator to $\hat{H}(Q) \subset J^1Y$. Given a Hamiltonian form H associated with L, the Lagrangian constraint space is defined by the coordinate relation,

$$p_i = \partial^t_i \mathscr{L}(t, y^j, \partial^j \mathscr{H}).$$

If a Lagrangian L is hyperregular, there exists a unique Hamiltonian form associated with L.

Let a Lagrangian L be semiregular, i.e., the preimage $\hat{L}^{-1}(p)$ of any point $p \in Q$ is a connected submanifold of J^1Y. The following assertions issue from the corresponding theorems of a polysymplectic formalism.[14,16,20]

Proposition 12: All Hamiltonian forms H associated with a semiregular Lagrangian L coincide on the Lagrangian constraint space Q, and the Poincaré–Cartan form H_L is the pull-back of any such a Hamiltonian form H by the Legendre map $\hat{L}$, i.e.,

$$\pi_i y^i_t - \mathscr{L} = \mathscr{H}(t, y^i, \pi_i). \tag{35}$$

Proposition 13: Let a section r of the bundle $V^*Y \to \mathbf{R}$ be a solution of the Hamilton equations (17) for a Hamiltonian form H associated with a semiregular Lagrangian density L. If r lives in the Lagrangian constraint space Q, the section $s = \pi_\Pi \circ r$ of the bundle $Y \to \mathbf{R}$ satisfies the Lagrange equations (29), while its jet prolongation $\overline{s} = \hat{H} \circ r = J^1 s$ obeys the Cartan equations (32). Conversely, let a section $\overline{s}$ of $J^1Y \to \mathbf{R}$ be a solution of the Cartan equations (32) for a semiregular Lagrangian L. Let H be a Hamiltonian form associated with L so that the corresponding Hamiltonian map satisfies the condition

$$\hat{H} \circ \hat{L} \circ \overline{s} = J^1(\pi^1_0 \circ \overline{s}). \tag{36}$$

Then the section $r = \hat{L} \circ \overline{s}$ of $V^*Y \to \mathbf{R}$ is a solution of the Hamilton equations for H.

Proof: One can show that, in the case of a semiregular Lagrangian L, the Euler–Lagrange–Cartan operator (31) is the pull-back,

$$\mathscr{E}_{\overline{L}} = (J^1\hat{L})^* \mathscr{E}_H, \tag{37}$$

of the Hamilton operator for a Hamiltonian form H associated with L. In accordance with the relation (37), if γ_H is a Hamiltonian connection for H, the composition,

$$J^1\hat{H} \circ \gamma_H : V^*Y \to J^2Y, \quad (y^i_t, \hat{y}^i_t, y^i_{tt}) \circ J^1\hat{H} \circ \gamma_H = (\partial^i \mathscr{H}, \partial^i \mathscr{H}, d_{Ht}\partial^i \mathscr{H}),$$

takes its values into the kernel of the Euler–Lagrange operator $\mathscr{E}_L$. Then the morphism $J^1\hat{H}\circ\gamma_H\circ\hat{L}$ restricted to $\hat{H}(Q)$ is a local Lagrangian connection on $\hat{H}(Q)\subset J^1Y$. Conversely, the first of the Hamilton equations (17) is satisfied due to the condition (36), while the second one reduces to the second of the Cartan equations (32) because of the relation (35).

Since, by Proposition 13, solutions of the Lagrange equations for a degenerate Lagrangian may correspond to solutions of different Hamilton equations, we can conclude that, roughly speaking, the Hamilton equations involve some additional conditions in comparison with the Lagrange ones. Therefore, let us separate a part of the Hamilton equations which are defined on the Lagrangian constraint space Q in the case of almost regular Lagrangians.

Definition 14: A semiregular Lagrangian L is said to be almost regular if (i) the Lagrangian constraint space $Q\to Y$ is a closed imbedded subbundle $i_Q:Q\hookrightarrow V^*Y$ of the Legendre bundle $V^*Y\to Y$; and (ii) the Legendre map $\hat{L}:J^1Y\to Q$ is a bundle.

Let $H_Q=i_Q^*H$ be the restriction of a Hamiltonian form H associated with L to the constraint space Q. By virtue of Proposition 12, this restriction, called the constrained Hamiltonian form, is uniquely defined, and $H_L=\hat{L}^*H_Q$. For sections r of the bundle $Q\to\mathbf{R}$, we can write the constrained Hamilton equations,

$$r^*(u_Q\rfloor dH_Q)=0, \tag{38}$$

where u_Q is an arbitrary vertical vector field on $Q\to\mathbf{R}$.[14,16,21] In brief, we can identify a vertical vector field u_Q on $Q\to\mathbf{R}$ with its image $Ti_Q(u_Q)$ and can bring the constrained Hamilton equations (38) into the form

$$r^*(u_Q\rfloor dH)=0, \tag{39}$$

where r is a section of $Q\to\mathbf{R}$ and u_Q is an arbitrary vertical vector field on $Q\to\mathbf{R}$. These equations fail to be equivalent to the Hamilton equations restricted to the constraint space Q.

The following two assertions together with Proposition 13 give the relations between Cartan, Hamilton, and constrained Hamilton equations when a Lagrangian is almost regular.[16]

Proposition 15: For any Hamiltonian form H associated with an almost regular Lagrangian L, every solution r of the Hamilton equations that lives in the Lagrangian constraint space is a solution of the constrained Hamilton equations (39).

Proposition 16: A section $\overline{s}$ of $J^1Y\to\mathbf{R}$ is a solution of the Cartan equations (32) iff $\hat{L}\circ\overline{s}$ is a solution of the constrained Hamilton equations (39).

In the spirit of a well-known Dirac–Bergmann algorithm for analyzing constrained systems in symplectic mechanics, the Lagrangian constraint space Q plays the role of the primary constraint one. However, we cannot extend this algorithm to time-dependent mechanics because the Poisson bracket of a Hamiltonian and constraint functions fails to be defined.

At the same time, one can hope to find a complete family of Hamiltonian forms associated with a degenerate Lagrangian L so that, for each solution of the Lagrange equations, there exists the corresponding solution of the Hamilton ones for some Hamiltonian form from this family. By Proposition 13, if L is semiregular, such a complete family of associated Hamiltonian forms exists iff, for every solution s of the Lagrange equations, there is a Hamiltonian form H from this family such that

$$\hat{H}\circ\hat{L}\circ J^1s=J^1s.$$

In particular, a complete family of Hamiltonian forms associated with an almost regular quadratic Lagrangian always exists.[14–16]

Note that, since Hamiltonians in time-dependent mechanics are not functions on a phase space, we cannot apply to them the well-known analysis of the normal forms[22] (e.g., quadratic Hamiltonians[2]) in symplectic mechanics.

VIII. CONSERVATION LAWS

In autonomous mechanics, an integral of motion, by definition, is a function on the phase space whose Poisson bracket with a Hamiltonian is equal to zero. This notion cannot be extended to time-dependent mechanics because the Hamiltonian evolution equation (19) is not reduced to the Poisson bracket.

Let us start from conservation laws in Lagrangian mechanics. To obtain differential conservation laws, we use the first variational formula (27).[16,19] On shell, this leads to the weak identity

$$J^1u \rfloor d\mathscr{L} \approx -d_t\mathscr{T}, \tag{40}$$

where

$$\mathscr{T} = \pi_i(u^t y^i_t - u^i) - u^t\mathscr{L} \tag{41}$$

is the current along the vector field u (25). If the Lie derivative $\mathbf{L}_{J^1u}L$ (26) vanishes, we have the conservation law

$$0 \approx -d_t[\pi_i(u^t y^i_t - u^i) - u^t\mathscr{L}].$$

This is brought into the differential conservation law,

$$0 \approx -\frac{d}{dt}(\pi_i \circ s(u^t \partial_t s^i - u^i \circ s) - u^t\mathscr{L} \circ s),$$

on solutions s of the Lagrange equations. A glance at this expression shows that, in time-dependent mechanics, the conserved current (41) plays the role of an integral of motion.

Every symmetry current (41) along a vector field u (25) on Y can be represented as a superposition of the Nöther current along a vertical vector field ϑ and of the energy current along some connection Γ (2) on $Y \to \mathbf{R}$, where $u = \vartheta + \Gamma$.[19,23]

If ϑ is a vertical vector field, the weak identity (40) reads as

$$(\vartheta^i \partial_i + d_t\vartheta^i \partial^t_i)\mathscr{L} \approx d_t(\pi_i \vartheta^i).$$

If the Lie derivative of L along ϑ equals zero, we have the integral of motion $\mathscr{T} = \pi_i \vartheta^i$.

In the case of a connection Γ (2), the weak identity (40) takes the form

$$(\partial_t + \Gamma^i \partial_i + d_t\Gamma^i \partial^t_i)\mathscr{L} \approx -d_t(\pi_i(y^i_t - \Gamma^i) - \mathscr{L}), \tag{42}$$

where one can think of

$$\mathscr{T}_\Gamma = \pi_i(y^i_t - \Gamma^i) - \mathscr{L} \tag{43}$$

as being the energy function with respect to the reference frame Γ. In particular, the energy conservation law (42) written relative to the coordinates adapted to Γ takes the familiar form

$$\partial_t\mathscr{L} = -d_t(\pi_i y^i_t - \mathscr{L}). \tag{44}$$

To discover conservation laws within the framework of Hamiltonian formalism, we will do the following trick. Given a Hamiltonian form H (13) on V^*Y, let us consider the Lagrangian,

$$L_H = (p_i y^i_t - \mathscr{H})dt, \tag{45}$$

on the jet manifold J^1V^*Y. It is readily observed that the Poincaré–Cartan form of the Lagrangian L_H coincides with the Hamiltonian form H, and the Euler–Lagrange operator for L_H is precisely the Hamilton operator for H. As a consequence, the Lagrange equations for L_H are equivalent to the Hamilton equations for H. Therefore, let us apply the first variational formula (27) to the Lagrangian (45).[19]

Given a vector field (25) on the event bundle Y, its canonical lift onto V^*Y reads as

$$\widetilde{u}=u^t\partial_t+u^i\partial_i-\partial_i u^j p_j\partial^i, \quad u^t=0,1.$$

Substituting this vector field into the weak identity (40), we obtain

$$-u^i\partial_i\mathscr{H}-u^t\partial_t\mathscr{H}+p_i d_t u^i\approx -d_t(-p_i u^i+u^t\mathscr{H}), \tag{46}$$

for the current

$$\widetilde{\mathscr{T}}=-p_i u^i+u^t\mathscr{H}. \tag{47}$$

In the case of $u=\Gamma$, the weak identity (46) takes the form

$$-\partial_t\mathscr{H}-\Gamma^i\partial_i\mathscr{H}+p_i d_t\Gamma^i\approx -d_t\widetilde{\mathscr{H}}_\Gamma,$$

where $\widetilde{\mathscr{H}}_\Gamma=\mathscr{H}-p_i\Gamma^i$ is the Hamiltonian function in the splitting (15).

The following assertion shows that the Hamiltonian function $\widetilde{\mathscr{H}}_\Gamma$ is the Hamiltonian counterpart of the Lagrangian energy function $\mathscr{T}_\Gamma$ (43) in the case of semiregular Lagrangians.[16]

Proposition 17: Let a Hamiltonian form H on the Legendre bundle V^*Y be associated with a semiregular Lagrangian $\mathscr{L}$ on J^1Y. Let r be a solution of the Hamilton equations (17) for H that lives in the Lagrangian constraint space Q and s the corresponding solution of the Lagrange equations for L. Then, we have

$$\widetilde{\mathscr{T}}(r)=\mathscr{T}(\hat{H}\circ r), \quad \widetilde{\mathscr{T}}(\hat{L}\circ J^1 s)=\mathscr{T}(s),$$

where $\mathscr{T}$ is the current (41) on J^1Y and $\widetilde{\mathscr{T}}$ is the current (47) on V^*Y.

Therefore, we can treat $\widetilde{\mathscr{H}}_\Gamma$ as the energy function with respect to the reference frame Γ. In particular, if $\Gamma^i=0$, we obtain the well-known energy conservation law

$$\partial_t\mathscr{H}\approx d_t\mathscr{H},$$

which is the Hamiltonian variant of the Lagrangian conservation law (44).

IX. RELATIVISTIC MECHANICS

Let us consider a mechanic system whose event space Z has no fibration $Z\to\mathbf{R}$ or admits different such fibrations. We come to relativistic mechanics where a configuration space is the jet manifold of one-dimensional submanifolds of Y that generalizes the notion of jets of sections of a bundle.[16,24,25]

Let Z be a manifold of dimension $m+n$. The first-order jet manifold $J^1_n Z$ of n-dimensional submanifolds of Z comprises the equivalence classes $[S]^1_z$ of n-dimensional imbedded submanifolds of Z that pass through $z\in Z$ and that are tangent to each other at z. It is provided with a manifold structure as follows.

Let $Y\to X$ be an $(m+n)$-dimensional bundle over an n-dimensional base X and Φ an imbedding of Y into Z. Then there is the natural injection,

$$J^1\Phi:\ J^1Y\to J^1_n Z,$$

$$j^1_x s\mapsto[S]^1_{\Phi(s(x))}, \quad S=\mathrm{Im}(\Phi\circ s),$$

where s are sections of $Y\to X$. This injection defines a chart on $J^1_n Z$. Such charts cover the set $J^1_n Z$, and transition functions between these charts are differentiable. They provide $J^1_n Z$ with the structure of a finite-dimensional manifold.

Hereafter, we will use the following coordinate atlases on the jet manifold $J^1_n Z$ of submanifolds of Z. Let Z be endowed with a manifold atlas with coordinate charts,

$$(z^A), \quad A=1,\ldots,n+m. \tag{48}$$

Though $J^0_n Z$, by definition, is diffeomorphic to Z, let us provide $J^0_n Z$ with the atlas obtained by replacing every chart (z^A) on a domain $U \subset Z$ with the charts on the same domain U that correspond to the different partitions of the collection (z^A) in collections of n and m coordinates. We denote these coordinates by

$$(x^\lambda, y^i), \quad \lambda = 1,...,n, \quad i = 1,...,m. \tag{49}$$

The transition functions between the coordinate charts (49) of $J^0_n Z$ associated with the coordinate chart (48) of Z are reduced simply to exchange between coordinates x^λ and y_i. Transition functions between arbitrary coordinate charts of the manifold $J^0_n Z$ take the form

$$\widetilde{x}^\lambda = \widetilde{g}^\lambda(x^\mu, y^j), \quad \widetilde{y}^i = \widetilde{f}^i(x^\mu, y^j). \tag{50}$$

Given the coordinate atlas (49) of the manifold $J^0_n Z$, the jet manifold $J^1_n Z$ of Z is endowed with the adapted coordinates $(x^\lambda, y^i, y^i_\lambda)$. Using the formal total derivatives $d_\lambda = \partial_\lambda + y^i_\lambda \partial^i$, one can write the transformation rules for these coordinates in the following form. Given the coordinate transformations (50), it is easy to find that

$$d_{\widetilde{x}^\lambda} = [d_{\widetilde{x}^\lambda} g^\alpha(\widetilde{x}^\lambda, \widetilde{y}^i)] d_{x^\alpha}. \tag{51}$$

Then we have

$$\widetilde{y}^i_\lambda = \left[\left(\frac{\partial}{\partial \widetilde{x}^\lambda} + \widetilde{y}^p_\lambda \frac{\partial}{\partial \widetilde{y}^p}\right) g^\alpha(\widetilde{x}^\lambda, \widetilde{y}^i)\right]\left(\frac{\partial}{\partial x^\alpha} + y^j_\alpha \frac{\partial}{\partial y^j}\right)\widetilde{f}^i(x^\mu, y^j). \tag{52}$$

Remark: Given a manifold Z, there is one-to-one correspondence between the jets $[S]^1_z$ at a point $z \in Z$ and the n-dimensional vector subspaces of the tangent space $T_z Z$:

$$[S]^1_z \mapsto \dot{x}^\lambda(\partial_\lambda + y^i_\lambda([S]^1_z)\partial_i).$$

The bundle $J^1_n Z \to Z$ possesses the structure group $\mathrm{GL}(n,m;\mathbf{R})$ of linear transformations of the vector space $\mathbf{R}^{m+n}$ that preserve the subspace $\mathbf{R}^n$. Its typical fiber is the Grassmann manifold $\mathrm{GL}(n+m;\mathbf{R})/\mathrm{GL}(n,m;\mathbf{R})$ of n-dimensional vector subspaces of the vector space $\mathbf{R}^{m+n}$. In particular, if $n=1$, the fiber coordinates y^i_0 of $J^1_1 Z \to Z$ with the transition functions (52) are exactly the standard coordinates of the projective space $\mathbf{RP}^m$.

When $n=1$, the formalism of jets of submanifolds provides the adequate mathematical description of relativistic mechanics as follows.

Let Z be a $(m+1)$-dimensional manifold equipped with an atlas of coordinates (z^0, z^i), $i = 1,...,m$, (49) with the transition functions (50), which take the form

$$z^0 \to \widetilde{z}^0(z^0, z^j), \quad z^i \to \widetilde{z}^i(z^0, z^j). \tag{53}$$

The coordinates z^0 in different charts of Z play the role of the temporal ones.

Let $J^1_1 Z$ be the jet manifold of one-dimensional submanifolds of Z. This is provided with the adapted coordinates (z^0, z^i, z^i_0). Then one can think of z^i_0 as being the coordinates of nonrelativistic velocities. Their transition functions are obtained as follows.

Given the coordinate transformations (53), the total derivative (51) reads as

$$d_{\widetilde{z}^0} = d_{\widetilde{z}^0}(z^0) d_{z^0} = \left(\frac{\partial z^0}{\partial \widetilde{z}^0} + \widetilde{z}^k_0 \frac{\partial z^0}{\partial \widetilde{z}^k}\right) d_{z^0}.$$

In accordance with the relation (52), we have

$$\widetilde{z}^i_0 = d_{\widetilde{z}^0}(z^0) d_{z^0}(\widetilde{z}^i) = \left(\frac{\partial z^0}{\partial \widetilde{z}^0} + \widetilde{z}^k_0 \frac{\partial z^0}{\partial \widetilde{z}^k}\right)\left(\frac{\partial \widetilde{z}^i}{\partial z^0} + z^j_0 \frac{\partial \widetilde{z}^i}{\partial z^j}\right).$$

The solution of this equation is

$$\widetilde{z}^i_0=\left(\frac{\partial\widetilde{z}^i}{\partial z^0}+z^j_0\,\frac{2\partial\widetilde{z}^i}{\partial z^j}\right)\Big/\left(\frac{\partial\widetilde{z}^0}{\partial z^0}+z^k_0\,\frac{\partial\widetilde{z}^0}{\partial z^k}\right).$$

This is the transformation law of nonrelativistic velocities, which illustrates that the jet bundle $J^1_1Z\to Z$ is not affine, but projective.

To obtain the relation between nonrelativistic and relativistic velocities, let us consider the tangent bundle TZ equipped with the induced coordinates $(z^0,z^i,\dot z^0,\dot z^i)$. There is the morphism

$$\rho:\ TZ\to J^1_1Z,\quad z^i_0\circ\rho=\dot z^i/\dot z^0. \tag{54}$$

It is readily observed that the coordinate transformation laws of z^i_0 and $\dot z^i/\dot z^0$ are the same. Thus, one can think of the coordinates $(z^0,\dot z^i)$ as being relativistic velocities.

The morphism (54) is a surjection. Let us assume that the tangent bundle is equipped with a pseudo-Riemannian metric g and $Q_z\subset T_zZ$ is the hyperboloid given by the relation

$$g_{\mu\nu}(z)\dot z^\mu\dot z^\nu=1,\quad \mu,\nu=0,1,...,m. \tag{55}$$

The union of these hyperboloids over Z,

$$Q=\bigcup_{z\in Z}Q_z=Q^+\cup Q^-$$

is the union of two connected imbedded subbundles of TZ. Then the restriction of the morphism (54) to each of these subbundles is an injection of Q into J^1_1Z.

Let us consider the image of this injection in the fiber of J^1_1Z over a point $z\in Z$. There are coordinates (z^0,z^i) in a neighborhood around z such that the pseudo-Riemannian metric $g(z)$ at z comes to the Minkowski one,

$$g(z)=\eta=\mathrm{diag}(1,-1,...,-1).$$

In this coordinates the hyperboloid $Q_z\subset T_zZ$ is given by the relation

$$(\dot z^0)^2-\sum_i\,(\dot z^i)^2=1.$$

This is the union of the subsets Q^+_z where $z^0>0$ and Q^-_z where $z^0<0$. The image $\rho(Q^+_z)$ is given by the coordinate relation

$$\sum_i\,(z^i_0)<1.$$

From the physical viewpoint, this relation means that nonrelativistic velocities are bounded in accordance with Special Relativity.

In general, Lagrangian formalism fails to be appropriate to relativistic mechanics because a Lagrangian $L=\mathscr{L}dz^0$ can be defined only locally. Its Hamiltonian formulation is more promising.

The phase space of relativistic mechanics is the cotangent bundle T^*Z provided with the holonomic coordinates $(z^0,\ z^i,\ p_0=\dot z_0,\ p_i=\dot z_i)$. This admits the canonical symplectic form

$$\Omega=dp_\mu\wedge dz^\mu,\quad \mu=0,1,2,3.$$

Hamiltonian relativistic mechanics is formulated as an autonomous Hamiltonian mechanics on the symplectic manifold T^*Z. Every Hamiltonian $\mathscr{H}$ on the phase space T^*Z defines the Hamiltonian map,

$$\hat{\mathscr{H}}:\ T^*Z\to TZ,\quad \dot z^\mu\circ\hat{\mathscr{H}}=\partial^\mu\mathscr{H}.$$

Since relativistic velocities live in the hyperboloid (55), we have the similar constraint,

$$g_{\mu\nu}\frac{\partial\mathscr{H}}{\partial p_\mu}\frac{\partial\mathscr{H}}{\partial p_\nu}=1, \tag{56}$$

on the phase space T^*Z. Thus, relativistic mechanics is described as a Dirac constraint system on the primary constraint space Q (56). Its solutions are integral curves of the Hamiltonian vector field ϑ on $Q\subset T^*Z$ that obeys the Hamilton equation

$$\vartheta\rfloor i_Q^*\Omega=-i_Q^*d\mathscr{H}. \tag{57}$$

It should be emphasized that, as follows from the relation (24), the coordinate $-p$, but not a Hamiltonian $\mathscr{H}$ plays the role of a relativistic energy function. For example, let us write (i) the Hamiltonian,

$$\mathscr{H}=-\frac{1}{2m}\,\eta^{\mu\nu}p_\mu p_\nu,$$

of a free relativistic mass m in Special Relativity on the Minkowski space $\mathbf{R}^4$; (ii) the Hamiltonian,

$$\mathscr{H}=-\frac{1}{2m}\,\eta^{\mu\nu}(p_\mu-eA_\mu)(p_\nu-eA_\nu),$$

of a relativistic electric charge e in the presence of an electromagnetic field A_μ; and (iii) the Hamiltonian,

$$\mathscr{H}=-\frac{1}{2m}\,\gamma^{\mu\nu}p_\mu p_\nu,$$

of a point mass m in the presence of a metric gravitational field g. Substituting these Hamiltonians into the equations (56) and (57), we obtain the well-known solutions of relativistic mechanics.

[1] R. Abraham and J. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, London, 1978).
[2] V. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1978).
[3] P. Libermann and C-M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dordrecht, 1987).
[4] I. Vaisman, *Lectures on the Geometry of Poisson Manifolds* (Birkhäuser Verlag, Basel, 1994).
[5] E. Binz, H. Fischer, and J. Śniatycki, *Geometry of Classical Fields* (North-Holland, Amsterdam, 1988).
[6] M. de León and P. Rodrigues, *Methods of Differential Geometry in Analytical Mechanics* (North-Holland, Amsterdam, 1989).
[7] G. Morandi, C. Ferrario, and G. Lo Vecchio, Phys. Rep. **188**, 147 (1990).
[8] A. Echeverría Enríquez, M. Muñoz Lecanda, and N. Román Roy, Rev. Math. Phys. **3**, 301 (1991).
[9] J. Cariñena and J. Fernández-Núñez, Fortschr. Phys. **41**, 517 (1993).
[10] E. Massa and E. Pagani, Ann. Inst. Henri Poincaré **61**, 17 (1994).
[11] J. Cariñena and M. Rañada, J. Math. Phys. **30**, 2258 (1989).
[12] J. Cariñena, J. Fernández-Núñez, and E. Martínez, Lett. Math. Phys. **23**, 51 (1991).
[13] G. Sardanashvily and O. Zakharov, Diff. Geom. Appl. **3**, 245 (1993).
[14] G. Sardanashvily, *Generalized Hamiltonian Formalism for Field Theory. Constraint Systems* (World Scientific, Singapore, 1995).
[15] G. Sardanashvily, J. Math. Phys. **35**, 6584 (1994).
[16] G. Giachetta, L. Mangiarotti, and G. Sardanashvily, *New Lagrangian and Hamiltonian Methods in Field Theory* (World Scientific, Singapore, 1997).
[17] D. Canarutto, Bull. U.M.I. Alg. Geom. Ser. VI **V-D**, 18 (1986).
[18] I. Vaisman, Diff. Geom. Appl. **5**, 153 (1995).
[19] G. Sardanashvily, J. Math. Phys. **38**, 847 (1997).
[20] O. Zakharov, J. Math. Phys. **33**, 607 (1992).
[21] G. Giachetta and L. Mangiarotti, Int. J. Theor. Phys. **34**, 2353 (1995).
[22] A. D. Bruno, Usp. Mat. Nauk **43**, 23 (1988) (in Russian).
[23] A. Echeverría Enríquez, M. Muñoz Lecanda, and N. Román Roy, J. Phys. A **19**, 5553 (1995).
[24] I. Krasil'shchik, V. Lychagin, and A. Vinogradov, *Geometry of Jet Spaces and Nonlinear Partial Differential Equations* (Gordon and Breach, Glasgow, 1985).
[25] M. Modugno and A. Vinogradov, Ann. Mat. Pura Appl. **CLXVII**, 33 (1994).

New structures in the theory of the laser model. II. Microscopic dynamics and a nonequilibrium entropy principle

Fabio Bagarello[a] and Geoffrey L. Sewell[b]
Department of Physics, Queen Mary and Westfield College, Mile End Road, London E1 4NS, England

(Received 22 December 1997; accepted for publication 7 January 1998)

In a recent article, Alli and Sewell [J. Math. Phys. **36**, 5598 (1995)] formulated a new version of the Dicke–Hepp–Lieb laser model in terms of quantum dynamical semigroups, and thereby extended the macroscopic picture of the model. In the present article, we complement that picture with a corresponding microscopic one, which carries the following new results. (a) The local microscopic dynamics of the model is piloted by the classical, macroscopic field, generated by the collective action of its components; (b) the global state of the system carries no correlations between its constituent atoms after transient effects have died out; and (c) in the latter situation, the state of the system at any time t maximizes its entropy density, subject to the constraints imposed by the instantaneous values of its macroscopic variables. © *1998 American Institute of Physics.* [S0022-2488(98)01805-2]

I. INTRODUCTION

The classic work by Hepp and Lieb[1] (HL) on the theory of the Dicke laser model[2] represents a landmark in the constructive quantum theory of phase transitions, far from equilibrium, in open quantum systems. Most notably, it demonstrated how the model undergoes a transition from normal to coherent radiation when the optical pumping strength reaches a critical value. This substantiated ideas that had been proposed by Graham and Haken[3] at a more heuristic level.

In a recent article, which we shall refer to as (I), Alli and Sewell[4] constructed a new version[5] of the Dicke model in which the dynamics of the matter and radiation is governed by a quantum dynamical semigroup, which incorporates the dissipative action of an array of pumps and sinks: by contrast, the HL model represented a conservative system, comprising the matter, radiation, and these reservoirs. The principle new results of (I) were (a) a generalization of the HL form of the dynamics of the matter–radiation composite, (b) a demonstration that the model exhibits optically chaotic phases, in addition to the normal and coherent ones, and (c) a generalization of the theory of quantum dynamical semigroups[6] to ones with unbounded generators.

The theories of both HL and (I) were centered on the macroscopic equations of motion, derived from the underlying quantum dynamics of their respective models in a limit where their sizes become infinite. In fact, they did not provide any treatment of the microdynamics in this limit. The object of the present article is to complement the macroscopic picture of (I) with a corresponding microscopic one.

In seeking a microscopic-cum-macroscopic picture of the dynamics of the model, we note that this has already been achieved for certain *conservative* systems with long range forces. For such systems, Morchio and Strocchi[7] have shown that the evolutes, b_t, of the local microscopic observables, b, depend not only on b and the time t, but also on certain macroscopic observables, i.e., ''observables at infinity'': this represents a picture completely different from that of the Haag–Kastler scheme,[8] in which the dynamics of the micro-observables is autonomous. Furthermore, the explicit structure of the equations of motion for the microscopic and macroscopic

[a]Permanent address: Dipartimento di Matematica ed Applicazioni, Fac. Ingegneria, Universita' di Palermo, I-90128 Palermo, Italy. Electronic mail: Bagarello@ipamat.math.unipa.it
[b]Electronic mail: G.L.Sewell@qmw.ac.uk

 © 1998 American Institute of Physics

observables, b_t and B_t, has been obtained by Bagarello and Morchio[9] (BM) for a class of conservative, mean field theoretic models in the following form:

$$\frac{db_t}{dt}=f(b_t,B_t); \quad \frac{dB_t}{dt}=F(B_t). \tag{1.1}$$

Thus, the microscopic dynamics is piloted by the macroscopic observables, which, for their part, evolve according to a self-contained law.

The question now arises as to whether a similar state of affairs prevails in the case of open dissipative systems, such as the laser model, and the first thing to say about that is that the methods of BM, which led to Eq. (1.1), are not applicable there. The reason is simply that those methods depended heavily on the fact that, in a conservative system, the dynamics is automorphic and so preserves the algebraic structure of the observables. In an open dissipative system, on the other hand, the dynamics is governed by a semigroup[4] of transformations of the observables that does not preserve that structure, i.e., in general, $(ab)_t \neq a_t b_t$.

Thus, any attempt to derive equations of the form (1.1) for the laser model must be based on methods that are applicable to dissipative systems. Here, the essential problem is to pass to a limit where the size of the system becomes infinite, since the dynamics of the finite version of the model has already been formulated in (I). In fact, we surmount this problem by means of a compactness argument that lends itself to the application of the Arzela–Ascoli theorem. As a result, we do indeed obtain a microscopic-cum-macroscopic dynamics of the form (1.1), though now, of course, the functions f and F incorporate the dissipative action of the pumps and sinks on the system.

The principal new results that we obtain in this way are of a very simple character, and may be summarized as follows.

(1) The microscopic observables of *this particular model* are those of the matter only, since the radiation field reduces to a classical macroscopic one.
(2) The dynamics of the local microscopic observables is piloted by that classical radiation field.
(3) The microdynamics contains certain transients, whose lifetimes are of the order of that of an atomic excited state. This dynamics simplifies considerably after these transients have died out, and leads to the decorrelation of the observables of different atoms. Furthermore, the resultant global microstate of the system is stationary and space-translationally invariant in the normal phase, and periodic with respect to both space and time in the coherent one. Evidently, it is very complicated in the chaotic phase.
(4) Following the decay of the transients, the instantaneous state of the model is determined by a nonequilibrium variational principle. Specifically, this state maximizes the global entropy density, subject to the constraints imposed by the prevailing values of the macroscopic observables. As far as we know, this is the first example for which a principle of this kind has been proved for the nonequilibrium regime.

We shall present our arguments as follows. In Secs. II and III, we shall provide a brief summary of (I), re-expressed in a form suitable for our present purposes. Specifically, Sec. II will be devoted to a formulation of the model and Sec. III to a resume of its macroscopic properties. In Sec. IV, we shall derive the above results (1) and (2) by passing to the infinite size limit of the quantum dynamics of the finite version of the model. We shall then derive the results (3) and (4) in Secs. V and VI, respectively. We shall conclude, in Sec. VII, with a brief discussion of our main results and their possible ramifications.

II. THE MODEL

Our model, as formulated in (I), is a dissipative quantum system, $\Sigma^{(N)}$, consisting of a chain of $2N+1$ identical two-level atoms interacting with an n-mode radiation field. We label it by N only, since we shall be concerned with its properties in a limit where N tends to infinity, while n remains fixed and finite. We formulate $\Sigma^{(N)}$ as a quantum dynamical system $(\mathscr{B}^{(N)},T^{(N)},\mathscr{S}^{(N)})$, where $\mathscr{B}^{(N)}$ is a $\star$-algebra of observables, $T^{(N)}$ is a one-parameter semigroup of completely positive (CP) transformations of $\mathscr{B}^{(N)}$, and $\mathscr{S}^{(N)}$ is a family of normal states on $\mathscr{B}^{(N)}$ that is

stable under the dual of $T^{(N)}$. Most importantly, $T^{(N)}$ incorporates the dissipative action of pumps and sinks, which do not appear explicitly in the model, on the matter–radiation system. We build the model from its constituent parts as follows.

A. The single atom

This is assumed to be a two-state atom or spin, $\Sigma_{\rm at}$. Its algebra of observables, $\mathscr{A}_{\rm at}$, is that of the two-by-two matrices, and is thus the linear span of Pauli matrices $(\sigma_x,\sigma_y,\sigma_z)$ and the identity, I. Its algebraic structure is provided by the relations

$$\sigma_x^2=\sigma_y^2=\sigma_z^2=I; \quad \sigma_x\sigma_y=i\sigma_z, \quad \text{etc.} \tag{2.1}$$

We define the spin raising and lowering operators

$$\sigma_\pm=\tfrac{1}{2}(\sigma_x\pm i\sigma_y). \tag{2.2}$$

We assume that the atom is coupled to a pump and a sink, and that accordingly [cf. (I)] its dynamics is given by a one-parameter semigroup $\{T_{\rm at}(t)|t\in\mathbf{R}_+\}$ of completely positive, identity-preserving contractions of $\mathscr{A}_{\rm at}$, whose generator, $L_{\rm at}$, is of the following form:

$$L_{\rm at}\sigma_\pm=-(\gamma_1\mp i\epsilon)\sigma_\pm; \quad L_{\rm at}\sigma_z=-\gamma_2(\sigma_z-\eta I), \tag{2.3}$$

where $\epsilon\,(>0)$ is the energy difference between the ground and excited states of an atom, and the γ's and η are constants whose values are determined by the atomic coupling to the energy source and sink, and are subject to the restrictions that

$$0<\gamma_2\leqslant 2\gamma_1; \quad -1\leqslant\eta\leqslant 1. \tag{2.4}$$

In particular, η is positive or negative according to whether the coupling of the atom to the pump or the sink is the stronger. In the former case, these couplings drive the atom to a terminal mixed state with *inverted population*, i.e., with higher occupation probability for the excited state than for the ground state.

B. The matter

This consists of $2N+1$ noninteracting copies of $\Sigma_{\rm at}$, located at the sites $r=-N,\ldots,N$ of the one-dimensional lattice $\mathbf{Z}$. Thus, at each site, r, there is a copy, Σ_r, of $\Sigma_{\rm at}$, whose algebra of observables, $\mathscr{A}_r$, and dynamical semigroup, T_r, are isomorphic with $\mathscr{A}_{\rm at}$ and $T_{\rm at}$, respectively. We denote by $\sigma_{r,u}$ the copy of σ_u at r, for $u=x,y,z,\pm$.

We define the algebra of observables, $\mathscr{A}^{(N)}$, and the dynamical semigroup, $T^{(N)}_{\rm mat}$, of the matter to be $\otimes_{r=-N}^{N}\mathscr{A}_r$ and $\otimes_{r=-N}^{N}T_r$, respectively. Thus, $\mathscr{A}^{(N)}$ is the algebra of linear transformations of $\mathbf{C}^{4N+2}$. We identify elements A_r of $\mathscr{A}_r$ with those of $\mathscr{A}^{(N)}$ given by their tensor products with the identity operators attached to the remaining sites. Under this identification, the commutant, $\mathscr{A}_r'$, of $\mathscr{A}_r$ is the tensor product $\otimes_{s\neq r}\mathscr{A}_s$.

It follows from these specifications that the generator, $L^{(N)}_{\rm mat}$, of $T^{(N)}_{\rm mat}$ is given by the formula

$$L^{(N)}_{\rm mat}=\sum_{r=-N}^{N}L_r, \tag{2.5}$$

where

$$\begin{aligned} &L_r\sigma_{r,\pm}=-(\gamma_1\mp i\epsilon)\sigma_{r,\pm}; \quad L_r\sigma_{r,z}=-\gamma_2(\sigma_{r,z}-\eta I),\\ &L_r(A_rA_r')=(L_rA_r)A_r' \quad \forall A_r\in\mathscr{A}_r, \quad A_r'\in\mathscr{A}_r'. \end{aligned} \tag{2.6}$$

C. The radiation

We assume that the radiation field consists of n ($<\infty$) modes, represented by creation and destruction operators $\{a_l^\star, a_l | l=0,...,n-1\}$ in a Fock–Hilbert space $\mathscr{H}_{\rm rad}$, as defined by the standard specifications that (a) these operators satisfy the CCR,

$$[a_l,\ a_m^\star]=\delta_{lm}I; \quad [a_l,\ a_m]=0, \tag{2.7}$$

and (b) $\mathscr{H}_{\rm rad}$ contains a (vacuum) vector Φ, that is annihilated by each of the a's and is cyclic wrt the algebra of polynomials in the $a^\star$'s.

We define the Weyl map $z=(z_0,\ldots,z_{n-1})\to W(z)$, of C^n into $\mathscr{L}(\mathscr{H}_{\rm rad})$ by the formula

$$W(z)\equiv W(z_0,\ldots,z_{n-1})=\exp[i(z.a+(z.a)^*)], \quad \text{with } z.a=\sum_0^{n-1} z_l a_l. \tag{2.8}$$

We then define $\mathscr{R}$ to be the $\star$-algebra of polynomials in the a's, $a^\star$'s, and the Weyl operators $W(z)$, with z running through $\mathbf{C}^n$. Thus, in view of the CCR (2.7), this algebra is the linear span of terms of the form $W(z)a_{l_1}^\sharp \ldots a_{l_m}^\sharp$, where $a^\sharp$ denotes a or $a^\star$. Equivalently, it is the linear span of the derivatives, of all orders, of the operators $W(t_0z_0,\ldots,t_{n-1}z_{n-1})$ wrt the real variables $(t_0,\ldots,t_{n-1})$.

We assume that the radiation dynamics is given by the canonical extension to $\mathscr{R}$ of Vanheuverszwijn semigroup,[10] $T_{\rm rad}$, of quasi-free CP transformations of the Weyl algebra of linear combinations of $\{W(z)|z\in\mathbf{C}^n\}$. The formal generator of this semigroup is

$$L_{\rm rad}=\sum_{l=0}^{n-1} (i\omega_l[a_l^\star a_l,.]+2\kappa_l a_l^\star(.)a_l-\kappa_l\{a_l^\star a_l,.\}), \tag{2.9}$$

where $\{.,.\}$ denotes anticommutator, and the frequencies, ω_l, and the damping constants, κ_l, are positive.

D. The composite system

This is the coupled system, $\Sigma^{(N)}$, comprising the matter and radiation. We assume that its algebra of observables, $\mathscr{B}^{(N)}$, is the tensor product $\mathscr{A}^{(N)}\otimes\mathscr{R}$. Thus, $\mathscr{B}^{(N)}$, like $\mathscr{R}$, is an algebra of both bounded and unbounded operators in the Hilbert space $\mathscr{H}^{(N)}:=\mathbf{C}^{4N+2}\otimes\mathscr{H}_{\rm rad}$. We shall identify elements A and R, of $\mathscr{A}^{(N)}$ and $\mathscr{R}$, with $A\otimes I_{\rm rad}$ and $I_{\rm mat}\otimes R$, respectively.

We assume that the matter–radiation coupling is dipolar and is given by the interaction Hamiltonian

$$H_{\rm int}^{(N)}=i(2N+1)^{-1/2}\sum_{r=-N}^{N}\sum_{l=0}^{n-1}\lambda_l(\sigma_{r,-}a_l^\star\exp(-2\pi ilr/n)-\text{h.c.}), \tag{2.10}$$

where the λ's are real-valued, N-independent coupling constants. Correspondingly, we define the radiation field, $\phi^{(N)}$, of the model by the stipulation that its value at the site r is the coefficient of $\sigma_{r,+}$ in this formula. Thus,

$$\phi_r^{(N)}=-i(2N+1)^{-1/2}\sum_{l=0}^{n-1}\lambda_l a_l\exp(2\pi ilr/n). \tag{2.11}$$

We now need to define the state space, $\mathscr{S}^{(N)}$, and the dynamical semigroup, $T^{(N)}$, in a way that takes account of the unboundedness both of $H_{\rm int}^{(N)}$ and of some of the elements of $\mathscr{B}^{(N)}$. To this end, we start by defining $\mathscr{D}_0^{(N)}$ to be the space of density matrices in $\mathscr{H}^{(N)}$, with the trace norm topology, and $\mathscr{D}_1^{(N)}$ to be the subset of its elements, $\rho^{(N)}$, for which $\mathrm{Tr}(B^\star\rho^{(N)}B)$ is finite for all B in $\mathscr{B}^{(N)}$. We define $\mathscr{S}_1^{(N)}$ to be the set of positive, normalized, linear functionals, $\psi^{(N)}$, on $\mathscr{B}^{(N)}$, that are in the one-to-one correspondence with the $\mathscr{D}_1^{(N)}$-class density matrices, given by the standard formula $\psi^{(N)}(B)=\mathrm{Tr}(\rho^{(N)}B)$. We then adopt the formulation of $T^{(N)}$ and $\mathscr{S}^{(N)}$

provided by (I), in which $T^{(N)}$ is constructed as the modification of $T^{(N)}_{\rm mat}\otimes T_{\rm rad}$ due to the interaction $H^{(N)}_{\rm int}$, and $\mathscr{S}^{(N)}$ is a ''large'' subset of $\mathscr{S}^{(N)}_1$ that is stable wrt this dynamical semigroup. Specifically, the construction yields the following results.[4]

(1) The density matrices corresponding to $\mathscr{S}^{(N)}$ form a dense subset of $\mathscr{D}^{(N)}_0$ in the topology corresponding to the trace norm.

(2) $\mathscr{S}^{(N)}$ is stable under the transformations $\psi^{(N)}\to\psi^{(N)}\circ T^{(N)}(t):=\psi^{(N)}_t$.

(3)

$$\frac{d}{dt}\,\psi^{(N)}_t(B)=\psi^{(N)}_t(L^{(N)}B)\quad \forall B\in\mathscr{B}^{(N)},\quad t\in\mathbf{R}_+, \tag{2.12}$$

where

$$L^{(N)}=L^{(N)}_{\rm mat}+L_{\rm rad}+i[H^{(N)}_{\rm int},\ .]. \tag{2.13}$$

This completes our specification of $\Sigma^{(N)}=(\mathscr{B}^{(N)},T^{(N)},\mathscr{S}^{(N)})$.

E. The sequence of systems $\{\Sigma^{(N)}\}$

Since we shall be concerned with the properties of the above model in a limit where N tends to infinity and n remains fixed, we need to consider the sequence of systems $\{\Sigma^{(N)}|N\in\mathbf{N}\}$. Specifically, we shall investigate the evolution of these systems from initial states, $\psi^{(N)}$, for which the mean photon number is not superextensive, i.e., its ratio to N is bounded. Thus, we assume that, for some finite constant C,

$$\psi^{(N)}(a^\star_l a_l)<CN,\quad \forall N\in\mathbf{N},\quad l\in[0,n-1]. \tag{2.14}$$

This exclusion of super-extensivity was shown in (I) to persist, in a uniform way, at later times, in that

$$\psi^{(N)}_t(a^\star_l a_l)<DN,\quad \forall N\in\mathbf{N},\quad t\in\mathbf{R}_+,\quad l\in[0,n-1], \tag{2.15}$$

where D is a finite constant.

III. MACROSCOPIC DYNAMICS AND PHASE STRUCTURE

We formulate the macroscopic description of the model, as in (I), in terms of the global intensive observables

$$s^{(N)}_l=(2N+1)^{-1}\sum_{r=-N}^{N}\sigma_{r,-}\,\exp(-2\pi ilr/n);\quad l=0,...,n-1, \tag{3.1}$$

and

$$p^{(N)}_l=(2N+1)^{-1}\sum_{r=-N}^{N}\sigma_{r,z}\,\exp(-2\pi ilr/n);\quad l=0,...,n-1, \tag{3.2}$$

together with the operators

$$\alpha^{(N)}_l=(2N+1)^{-1/2}a_l;\quad l=0,...,n-1, \tag{3.3}$$

corresponding to a scaling of the number operators $a^\star_l a_l$ in units of $2N+1$. We note that the set of $p^{(N)}$'s is the same as that of their adjoints, since

$$p^{(N)\star}_0=p^{(N)}_0\quad\text{and}\quad p^{(N)\star}_l=p^{(N)}_{n-l}\quad\text{for } l=1,...,n-1. \tag{3.4}$$

The set $M^{(N)} := \{s_l^{(N)}, s_l^{(N)\star}, p_l^{(N)}, \alpha_l^{(N)}, \alpha_l^{(N)\star}\}$ of macroscopic variables is a Lie algebra wrt commutation, and, by Eqs. (2.10), (2.11), and (3.3), the radiation field and the interaction Hamiltonian are the algebraic functions of them given by

$$\phi_r^{(N)} = -i \sum_{l=0}^{n-1} \lambda_l \alpha_l^{(N)} \exp(2\pi i r l/n) \tag{3.5}$$

and

$$H_{\text{int}}^{(N)} = i(2N+1) \sum_{l=0}^{n-1} \lambda_l (s_l^{(N)} \alpha_l^{(N)\star} - \text{h.c.}). \tag{3.6}$$

In (I), the dynamics of $M^{(N)}$ was extracted from the microscopic equation of motion (2.12), in a limit where $N \to \infty$ and n remains fixed and finite, subject to the assumptions of Sec. II. As one might anticipate from the fact that the commutators of the elements of $M^{(N)}$ tend normwise to zero in this limit, this macrodynamics is classical.

To describe this dynamics, we simplify our notation by defining $x^{(N)} \equiv (x_1^{(N)}, \ldots, x_{5n}^{(N)})$ to be the $5n$ self-adjoint operators, which, by Eqs. (3.1)–(3.4), comprise the Hermitian and anti-Hermitian parts of the elements of $M^{(N)}$. Correspondingly, in anticipation of a classical limit for the dynamics of these observables, as $N \to \infty$, we introduce a phase space, $X = \mathbf{R}^{5n}$, and, for each point $x = (x_1, \ldots, x_{5n}) \in X$, we define $\{\alpha_l, s_l, p_l | l = 0, \ldots, n-1\}$ to be the complex numbers related to x in precisely the same way that the elements of $M^{(N)}$ are to $x^{(N)}$. We define $\mathscr{C}_0(X)$ to be the space of bounded continuous functions on X that tend to zero at infinity.

We formulate the macroscopic dynamics of $\Sigma^{(N)}$ in terms of the time-dependent quantum characteristic function $\mu_t^{(N)}$ on X, defined by the formula

$$\mu_t^{(N)}(y) = \psi_t^{(N)}(\exp(iy.x^{(N)})) \quad \forall y \in X, \quad t \in \mathbf{R}_+ . \tag{3.7}$$

Further [cf. (I)], we relate the limiting form of this function, as $N \to \infty$, to the following classical dynamical system.

Definition 3.1: We define $\mathscr{K}$ to be the flow in X, whose equation of motion,

$$\frac{dx_t}{dt} = F(x_t), \tag{3.8}$$

is of the explicit form

$$\frac{d\alpha_{l,t}}{dt} = -(i\omega_l + \kappa_l)\alpha_{l,t} + \lambda_l s_{l,t} \tag{3.9a}$$

$$\frac{ds_{l,t}}{dt} = -(i\epsilon + \gamma_1)s_{l,t} + \sum_{m=0}^{n-1} \lambda_m p_{[l-m],t} \alpha_{m,t} \tag{3.9b}$$

and

$$\frac{dp_{l,t}}{dt} = -\gamma_2(p_{l,t} - \eta\delta_{l,0}) - 2\sum_{m=0}^{n-1} \lambda_m (\overline{\alpha}_{m,t} s_{[l+m],t} + \alpha_{m,t} \overline{s}_{[m-l],t}), \tag{3.9c}$$

where $[l \pm m] = l \pm m (\text{mod } n)$.

The following Propositions were proved in (I).

Proposition 3.2: The equation of motion (3.8) [i.e., (3.9)] has a unique global solution, corresponding to a one-parameter semigroup of transformations $x \to \tau_t x (\equiv x_t)$ of X that maps the compacts into compacts.

Proposition 3.3: (i) Under the above assumptions, $\mu_t^{(N)}$ tends pointwise, as $N \to \infty$, to the characteristic function, μ_t, of a probability measure, m_t, on X, i.e.,

$$\lim_{N\to\infty} \mu_t^{(N)}(y) = \mu_t(y) = \int_X \exp\, i(x.y)\, dm_t(x), \tag{3.10}$$

convergence being uniform on the X-compacts.

(ii) The evolution of m_t is induced by the dynamical semigroup, τ, of $\mathscr{K}$, according to the formula

$$\int_X dm_t f = \int_X dm_0 f\circ\tau(t) \quad \forall f\in\mathscr{C}_0(X), \quad t\in\mathbf{R}_+ . \tag{3.11}$$

(iii) In the particular case of a pure phase, where m_0 is a Dirac measure, δ_x, with support at some point x, then $m_t=\delta_{x_t}$.

Comment: This last proposition signifies that, in the limit $N\to\infty$, the macroscopic variables $(\alpha^{(N)},s^{(N)},p^{(N)})$ reduce to classical ones (α,s,p). Correspondingly, the radiative field, $\phi^{(N)}$, as given by Eq. (3.5), reduces to a classical one, ϕ, whose value at the site τ and time t is

$$\phi_{r,t} = -i\sum_{l=0}^{n-1} \lambda_l \alpha_{t,l} \exp(2\pi i r l/n). \tag{3.12}$$

A. Phase structure

By Definition 3.1 and Propositions 3.2 and 3.3(iii), the properties of the pure phase are governed by the classical dynamical system $\mathscr{K}$. A study of this system yields the following results [cf. (I)].

(1) For η less than a certain specified critical value, $\eta_1(>0)$, the system has a unique stable stationary state corresponding to the fixed point, $x_0(\in X)$, given by

$$\alpha_l = 0; \quad s_l = 0; \quad p_l = \eta\delta_{l0} \quad \forall l\in[0, n-1]. \tag{3.13}$$

(2) As η increases through a certain positive value, η_1, the fixed point x_0 becomes unstable, and, by Hopf bifurcation, gives way to a periodic orbit of the form

$$\alpha_{l,t} = \alpha_l^{(0)} \exp(-i(\nu t+\theta^{(0)}))\delta_{lk}; \quad s_{l,t} = s_l^{(0)} \exp((-i\nu t+\theta^{(0)}))\delta_{lk}; \quad p_{l,t} = \eta_1\delta_{l0}, \tag{3.14}$$

where the selection of the kth mode and the values of the constants $\alpha^{(0)}$, $s^{(0)}$, and ν are determined by the parameters of the model, while the phase angle $\theta^{(0)}$ is indeterminate. Thus, in optical terms, there is a transition from normal to coherent radiation as η passes through η_1. This entails a breakdown of the gauge symmetry, represented by the transformations in which the $\sigma_{r,-}$, a_l, and correspondingly s_l, α_l, are all rephased by the same factor $\exp(i\theta)$.

(3) The model generically undergoes a further transition from coherent to chaotic radiation, as η passes through a second critical value, $\eta_2(>\eta_1)$. Specifically, the following two scenarios, both of which involve gauge symmetry breakdown, are feasible.

(a) There can be a bifurcation, at $\eta=\eta_2$, from a periodic orbit to a strange attractor, as specified in the Ruelle–Takens scheme.[11] This scenario is actually realized in the single mode case.[4]

(b) For large n, there could be a succession of bifurcations, corresponding to the activation of different modes, leading to a chaotic phase along the lines of Landau's theory of hydrodynamical turbulence.[12]

IV. THE MICROSCOPIC DYNAMICS IN PURE PHASES

Our objective now is to obtain the microscopic dynamics of the model, in the limit $N\to\infty$. We take the local microscopic observables to be those of the matter only, since, as noted in the Comment following Proposition 3.3, the radiation field reduces to a classical macroscopic one in this limit, and therefore this field carries no local microscopic observables within the terms of this model.[13]

We formulate the $C^\star$-algebra of the observables of the matter, in the limit $N\to\infty$, as the standard quasi-local one for the infinite chain of atoms of the given species, occupying the sites of the lattice $\mathbf{Z}$. Thus, in order to construct this algebra, we define L to be the family of all finite point subsets of $\mathbf{Z}$, and we assign to each $\Lambda\in L$ the $C^\star$-algebra $\mathscr{A}_\Lambda := \otimes_{r\in\Lambda}\mathscr{A}_r$, with $\mathscr{A}_r$ as defined in Sec. II. For $\Lambda\subset\Lambda'$, we identify elements A of $\mathscr{A}_\Lambda$ with $A\otimes I_{\Lambda'\setminus\Lambda}(\in\mathscr{A}_{\Lambda'})$, thereby rendering $\mathscr{A}_\Lambda$ isotonic wrt Λ. We then define $\mathscr{A}$ to be the completion of the local algebra $\mathscr{A}_L := \cup_{\Lambda\in L}\mathscr{A}_\Lambda$ wrt the norm it inherits from the $\mathscr{A}_\Lambda$'s. Thus, the algebras $\mathscr{A}^{(N)}$ of the material observables of the systems $\Sigma^{(N)}$ are just the local subalgebras $\mathscr{A}_{[-N,N]}$ of $\mathscr{A}$.

We denote by $\mathscr{S}(\mathscr{A})$ the set of all states on $\mathscr{A}$, and we represent the space translation group $\mathbf{Z}$ by the automorphisms V of this algebra, defined by the formula

$$V(m)\sigma_{r,u}=\sigma_{r+m,u}, \quad \forall r,m\in\mathbf{Z}, \quad u=\pm,z. \tag{4.1}$$

We note that it follows from our specifications of the local structure of $\mathscr{A}$ that the semigroup $T^{(N)}_{\mathrm{mat}}$, defined in Sec. II, is just the restriction to $\mathscr{A}^{(N)}$ of the global one-parameter semigroup $T_{\mathrm{mat}} := \otimes T_r$ of CP transformations of $\mathscr{A}$. The generator of T_{mat} is therefore

$$L_{\mathrm{mat}}=\sum_{r\in\mathbf{Z}} L_r, \tag{4.2}$$

where L_r is given by Eq. (2.6). The effects of the matter–field coupling on the properties of $\Sigma^{(N)}$, in the limit $N\to\infty$, will be treated presently.

Our aim now will be to obtain the microdynamics of the matter induced by that of the systems $\Sigma^{(N)}$ in the limit $N\to\infty$, subject to the following conditions, characteristic of pure phases, on their initial states.

(P1) The restrictions of the initial states, $\psi^{(N)}$, of the systems $\Sigma^{(N)}$, to the algebras $\mathscr{A}^{(N)}\equiv\mathscr{A}_{[-N,N]}$ are the same as those of a certain primary element, ω, of $\mathscr{S}(\mathscr{A})$, i.e.,

$$\omega_{|\mathscr{A}^{(N)}}=\psi^{(N)}_{|\mathscr{A}^{(N)}}, \quad \forall N\in\mathbf{N}.$$

(P2) $\omega(s_l^{(N)})$ and $\omega(p_l^{(N)})$ converge to values s_l and p_l, respectively, for each $l\in[0,n-1]$, as $N\to\infty$: Since ω is primary, the dispersions of $s_l^{(N)}$ and $p_l^{(N)}$ vanish in this limit.

(P3) The mean and dispersion of $\alpha_l^{(N)}$, for the state $\psi^{(N)}$, converge to values α_l and zero, respectively, for each $l\in[0,n-1]$, as $N\to\infty$.

The following lemma serves to confirm that the assumptions (P1)–(P3), together with Eq. (3.7), imply that the conditions of Proposition 3.3(iii) prevail.

Lemma 4.1: Under the above assumptions, the macroscopic probability measure, m_t, is simply δ_{x_t}, where $x_t=\tau(t)x$ and x is the initial phase point corresponding to (α,s,p) of conditions (P2) and (P3).

Proof: By Eq. (3.7) and the Schwartz inequality,

$$\begin{aligned}|\mu_0^{(N)}(y)-\exp(iy.x)| &\equiv |\psi^{(N)}(\exp(iy.(x^{(N)}-x))-I)|\\ &\leqslant 2\left[\psi^{(N)}\left(\sin^2\left(\frac{y.(x^{(N)}-x)}{2}\right)\right)\right]^{1/2}\\ &\leqslant[\psi^{(N)}(y.(x^{(N)}-x))^2]^{1/2},\end{aligned}$$

and, in view of (P2) and (P3), this tends to zero as $N\to\infty$. Hence, by Proposition 3.3(i), $\mu_0(y)=\exp(iy.x)$, i.e., $m_0=\delta_z$; and, consequently, by Proposition 3.3(iii), $m_t=\delta_{x_t}$, as required.

The following proposition signifies that, in the limit $N\to\infty$, the microscopic dynamics is precisely that of the matter, under the action of the classical macroscopic radiation field, ϕ.

Proposition 4.2: (i) Assuming the conditions (P1)–(P3), the microscopic dynamics of the model corresponds, in the limit $N\to\infty$, to a two-parameter family,

$$\{\mathscr{T}(s,t|\phi)|0\leqslant s\leqslant t;\mathscr{T}(s,u|\phi).\mathscr{T}_\mu(u,t|\phi)=\mathscr{T}(s,t|\phi)\},$$

of CP contractions of $\mathscr{A}$, in that

$$\lim_{N\to\infty} \psi^{(N)}(T^{(N)}(t)A) = \omega(\mathscr{T}(0,t|\phi)A) \quad \forall A \in \mathscr{A}_L, \quad t \in \mathbf{R}_+ . \tag{4.3}$$

Furthermore, the generator of $\mathscr{T}$ is

$$\mathscr{L}(t|\phi) := \frac{\partial}{\partial t}\mathscr{T}(s,t|\phi)_{|s=t} = L_{\mathrm{mat}} + \sum_{r\in\mathbf{Z}} [\phi_{r,t}\sigma_{r,+} - \mathrm{h.c.},.]. \tag{4.4}$$

Thus,

$$\mathscr{T}(s,t|\phi) = \mathbf{T}_a \exp\left(\int_s^t du\ \mathscr{L}(u|\phi)\right), \tag{4.5}$$

where $\mathbf{T}_a$ is the antichronological operator. We define

$$\omega_t := \omega \circ T(0,t). \tag{4.6}$$

(ii) $\mathscr{T}$ factorizes into a product of single site contributions according to the formula

$$\mathscr{T}(s,t|\phi) = \bigotimes_{r\in\mathbf{Z}} \mathscr{T}_r(s,t|\phi_r), \tag{4.7}$$

where

$$\mathscr{T}_r(s,t|\phi_r) = \mathbf{T}_a \exp\left(\int_s^t du\ \mathscr{L}_r(u|\phi)\right) \tag{4.8}$$

and

$$\mathscr{L}_r(t|\phi) = L_r + [\phi_{r,t}\sigma_{r,+} - \mathrm{h.c.},.]. \tag{4.9}$$

Comment: This result is similar to that obtained by Bagarello and Morchio[9] for a class of conservative systems, in that the microscopic evolution is governed by a classical field, formed by the time-dependent macroscopic observables.

The proof of the proposition requires the following lemmas.

Lemma 4.3: Defining

$$\tilde{\alpha}^{(N)}_{l,t} := \alpha^{(N)}_l - \alpha_{l,t}, \tag{4.10}$$

$$\lim_{N\to\infty} \psi^{(N)}_t(\tilde{\alpha}^{(N)\star}_{l,t}\tilde{\alpha}^{(N)}_{l,t}) = \lim_{N\to\infty} \psi^{(N)}_t(\tilde{\alpha}^{(N)}_{l,t}\tilde{\alpha}^{(N)\star}_{l,t}) = 0, \tag{4.11}$$

the convergence being uniform on the t-compacts.

Lemma 4.4: Defining

$$\tilde{s}^{(N)}_{l,t} := s^{(N)}_l - s_{l,t}, \tag{4.12}$$

$$\lim_{N\to\infty} \psi^{(N)}_t(\tilde{s}^{(N)}_{l,t}\tilde{s}^{(N)\star}_{l,t}) = 0, \tag{4.13}$$

the convergence being uniform on the t-compacts.

Proof of Proposition 4.2, assuming Lemma 4.3: (i) By Eqs. (2.5), (2.6), (2.9), (2.10), (2.12), and (2.13),

$$\psi_t^{(N)}(A)-\psi_0^{(N)}(A)=\int_0^t ds\ \psi_s^{(N)}(L_{\mathrm{mat}}A)+\int_0^t ds\sum_{l=0}^{n-1}\psi_s^{(N)}(\alpha_l^{(N)}\delta_l^+A+\alpha_l^{(N)\star}\delta_l^-A)\quad \forall A\in\mathscr{A}_L,$$

$$t\in\mathbf{R}_+, \tag{4.14}$$

where

$$\delta_l^{\pm}=\pm\lambda_l\sum_{r\in\mathbf{Z}}\exp(\pm 2\pi irl/n)[\sigma_{r,\pm},.]. \tag{4.15}$$

We note now that $\mathscr{A}_L$ is stable under $\delta_l^{\pm}$, and that, by Eqs. (2.14), (2.15), and (3.3), $\psi_t^{(N)}(\alpha_l^{(N)\star}\alpha_l^{(N)})$ is bounded, uniformly wrt N and t. It therefore follows from the Schwartz inequality that, for each $A\in\mathscr{A}_L$, the norm of the rhs of Eq. (4.14) is likewise uniformly bounded. Hence, by the Arzela–Ascoli theorem, $\psi_t^{(N)}(A)$ converges to a limit, uniformly over the t-compacts, as $N\to\infty$ over some subsequence of N. Therefore, as $\mathscr{A}$ is norm separable, it follows from the standard diagonalization procedure that there is a state ω_t' on $\mathscr{A}$, such that, over some subsequence of the integers,

$$\lim_{N\to\infty}\psi_t^{(N)}(A)=\omega_t'(A)\quad \forall\in\mathscr{A}_L, t\in\mathbf{R}_+. \tag{4.16}$$

Further, by Eqs. (2.7), (4.2), (4.4), (4.8), and (4.10), we may re-express Eq. (4.14) in the form

$$\psi_t^{(N)}(A)-\psi_0^{(N)}(A)=\int_0^t ds\ \psi_s^{(N)}(\mathscr{L}(s|\phi)A)+\int_0^t ds\sum_{l=0}^{n-1}\psi_s^{(N)}(\tilde{\alpha}_{l,s}^{(N)}\delta_l^+A+\tilde{\alpha}_{l,s}^{(N)\star}\delta_l^-A). \tag{4.17}$$

By Schwartz's inequality and Lemma 4.3, the last integral in this equation tends to zero, uniformly on the compacts, as $N\to\infty$. Hence, by Eq. (4.16), we see that Eq. (4.17) reduces to the following form in this limit:

$$\omega_t'(A)-\omega_0'(A)=\int_0^t ds\ \omega_s'(\mathscr{L}(s|\phi)A),$$

and hence

$$\frac{d}{dt}\omega_t'(A)=\omega_t'(\mathscr{L}(t|\phi)A). \tag{4.18}$$

Consequently, since Eq. (4.5) implies that

$$\frac{\partial}{\partial s}\mathscr{T}(s,t|\phi)=-\mathscr{L}(s|\phi).\mathscr{T}(s,t|\phi),$$

it follows from Eq. (4.18) that

$$\frac{\partial}{\partial s}\omega_s'(\mathscr{T}(s,t|\phi)A)=0\quad \forall 0\leqslant s\leqslant t,\quad A\in\mathscr{A}_L.$$

Hence, $\omega_s'(\mathscr{T}(s,t|\phi)A)$ is independent of s, and consequently, by assumption (P1) and Eq. (4.16),

$$\omega_t'(A)=\omega_0'(\mathscr{T}(0,t|\phi)A)\equiv\omega(\mathscr{T}(0,t|\phi)A),$$

i.e., by Eq. (4.6),

$$\omega_t'=\omega_t=\omega\circ\mathscr{T}(0,t|\phi). \tag{4.19}$$

Equations (4.3) and (4.4) now follow, at least for convergence over a subsequence of integers,

from Eqs. (4.16), (4.18), and (4.19). The removal of the restriction to a subsequence is achieved by noting that, in view of the uniqueness of the evolution $\omega \to \omega_t$, specified by Eq. (4.6), the compactness argument underlying the derivation of Eq. (4.19) may be reemployed to show that all subsequential limits of $\{\psi^{(N)}(T^{(N)}(t)A) | N \in \mathbf{N}\}$ are equal to $\omega_t(A)$.

(ii) The factorization property is an immediate consequence of the fact that, by Eqs. (4.4) and (4.9),

$$\mathscr{L}(t|\phi) = \sum_{r \in \mathbf{Z}} \mathscr{L}_r(t|\phi_r).$$

Proof of Lemma 4.3, assuming Lemma 4.4: Since, by Eqs. (2.7), (3.3), and (4.10), $[\overline{\alpha}_{l,t}^{(N)}, \overline{\alpha}_{l,t}^{(N)\star}] = N^{-1}$, it suffices to establish that the lhs of Eq. (4.11) tends to zero, uniformly on the t-compacts, as $N \to \infty$.

To this end, we define

$$J_{l,t}^{(N)} := \psi_t^{(N)}(\widetilde{\alpha}_{l,t}^{(N)\star} \widetilde{\alpha}_{l,t}^{(N)}), \tag{4.20}$$

and then infer from this equation and Eqs. (2.12) and (4.10) that

$$\frac{dJ_{l,t}^{(N)}}{dt} = \psi_t^{(N)}\left(\left(L^{(N)} + \frac{\partial}{\partial t}\right)\widetilde{a}_{l,t}^{(N)\star} \widetilde{\alpha}_{l,t}^{(N)}\right).$$

It follows from this equation, together with Eqs. (2.13), (3.1)–(3.3), (4.10), (4.12), and (4.20) that

$$\frac{dJ_{l,t}^{(N)}}{dt} = -2\kappa_l J_{l,t}^{(N)} + 2\lambda_l \,\mathrm{Re}\,(\psi_t^{(N)}(\overline{\alpha}_{l,t}^{(N)\star} \widetilde{s}_{l,t}^{(N)})), \tag{4.21}$$

and, consequently,

$$J_{l,t}^{(N)} = J_{l,0}^{(N)} \exp(-2\kappa_l t) + 2\lambda_l \,\mathrm{Re} \int_0^t dt' \exp(-2\kappa_l(t-t'))\psi_{t'}^{(N)}(\widetilde{\alpha}_{l,t'}^{(N)\star} \widetilde{s}_{l,t'}^{(N)}). \tag{4.22}$$

Therefore, defining

$$K_{l,t}^{(N)} := \psi_t^{(N)}(\widetilde{s}_{l,t}^{(N)} \widetilde{s}_{l,t}^{(N)\star}), \tag{4.23}$$

it follows from the Schwartz inequality and Eq. (4.22) that

$$J_{l,t}^{(N)} \leqslant J_{l,0}^{(N)} \exp(-2\kappa_l t) + \int_0^t dt' |\lambda_l| \exp(-2\kappa_l(t-t'))[K_{l,t'}^{(N)} J_{l,t'}^{(N)}]^{1/2}. \tag{4.24}$$

For any fixed $T(>0)$, we now define

$$C_{l,T}^{(N)} := \sup_{t \in [0,T]} J_{l,t}^{(N)} \tag{4.25}$$

and

$$D_{l,T}^{(N)} := \sup_{t \in [0,T]} K_{l,t}^{(N)}, \tag{4.26}$$

noting that these are both finite, in view of Eq. (2.15) and the uniform boundedness of $s^{(N)}$. Hence, by Eqs. (4.24)–(4.26),

$$J_{l,t}^{(N)} \leqslant J_{l,0}^{(N)} \exp(-2\kappa_l t) + 2|\lambda_l| \int_0^t dt' \exp(-2\kappa_l(t-t'))[C_{l,T}^{(N)} D_{l,T}^{(N)}]^{1/2} \quad \forall t \in [0, T]. \tag{4.27}$$

From this formula it follows easily that

$$J_{l,t}^{(N)} \leqslant J_{l,0}^{(N)} + \frac{|\lambda_l|}{\kappa_l} [C_{l,T}^{(N)} D_{l,T}^{(N)}]^{1/2},$$

and hence that

$$C_{l,T}^{(N)} \leqslant J_{l,0}^{(N)} + \frac{|\lambda_l|}{\kappa_l} [C_{l,T}^{(N)} D_{l,T}^{(N)}]^{1/2}, \tag{4.28}$$

which in turn implies that

$$C_{l,T}^{(N)} \leqslant \frac{1}{4} \left[\frac{|\lambda_1|}{\kappa_l} (D_{l,T}^{(N)})^{1/2} + \left(\frac{\lambda_l^2}{\kappa_l^2} D_{l,T}^{(N)} + 4 J_{l,0}^{(N)} \right)^{1/2} \right]^2. \tag{4.29}$$

Further, by definition (4.20) and assumption (P3), $J_{l,0}^{(N)}$ converges to zero as $N \to \infty$; and, by Lemma 4.4, the same is true of $D_{l,T}^{(N)}$. Hence, by Eq. (4.29), $C_{l,T}^{(N)}$ tends to zero as $N \to \infty$. By Eqs. (4.20) and (4.25), this is just the required result.

Proof of Lemma 4.4: Let $q_l^{(N)}$ denote the Hermitian or anti-Hermitian part of $s_l^{(N)}$; and, correspondingly, let $q_{l,t}$ denote the real or imaginary part of $s_{l,t}$. Then, defining

$$c_{l,t}^{(N)}(v) := \psi_t^{(N)}(\exp(iv q_l^{(N)})) \quad \forall v \in \mathbf{R}, \tag{4.30}$$

we see from Eq. (3.7) that $c_{l,t}^{(N)}$ is just the restriction of the characteristic function $\mu_t^{(N)}$ to one matter mode. Thus, by Proposition 3.3(iii) and Lemma 4.1,

$$\lim_{N \to \infty} c_{l,t}^{(N)}(v) = \exp(iv q_{t,l}), \quad \forall v \in \mathbf{R}, t \in \mathbf{R}_+, \tag{4.31}$$

convergence being uniform on the compacts. On integrating this equation against the Fourier transform, $\hat{f}(v)$, of any element, f, of the Schwartz space $\mathscr{D}(\mathbf{R})$, we see that

$$\lim_{N \to \infty} \psi_t^{(N)}(f(q_l^{(N)})) = f(q_{l,t}) \quad \forall f \in \mathscr{D}(\mathbf{R}), t \in \mathbf{R}_+, \tag{4.32}$$

convergence being uniform on the t-compacts for given f. We shall now use this formula to prove the required result, in the form of the equation

$$\lim_{N \to \infty} \psi_t^{(N)}((q_l^{(N)})^m) = (q_{l,t})^m \quad \text{for } m = 1,2, \tag{4.33}$$

convergence being uniform on the compacts.

To establish this formula, we first note that, by Eqs. (2.2) and (3.1), together with our definition of $q_l^{(N)}$, the norm of this operator is $\frac{1}{2}$. Thus, if f is any $\mathscr{D}(\mathbf{R})$-class function with support disjoint from $[-\frac{1}{2}, \frac{1}{2}]$, then $f(q_l^{(N)}) = 0$, and hence, by Eq. (4.32), $f(q_{l,t}) = 0$. This implies that $q_{l,t}$ must lie in the interval $[-\frac{1}{2}, \frac{1}{2}]$. We now reemploy Eq. (4.32), choosing f to be of the form $f(q) = q^m h(q)$, where $m = 1$ or 2, and h is an element of $\mathscr{D}(\mathbf{R})$, whose value is unity over the interval $[-\frac{1}{2}, \frac{1}{2}]$. This choice yields precisely the required result (4.33).

V. ASYMPTOTIC FORM OF THE DYNAMICS AND STATES OF PURE PHASES

We shall now show that the microscopic dynamics of the model simplifies greatly when transient effects, which decay in times of the order of the lifetime of the excited states of its atoms, are discarded.

We start by noting that, by Proposition 4.2, the microscopic dynamics of the model is completely determined by the action of the CP contractions, $\mathscr{T}_r$, on the single site algebras $\mathscr{A}_r$. Thus, defining

$$\sigma_{r,u}(t) := \mathscr{T}_r(0,t)\sigma_{r,u} \quad \text{for } u = +, -, x, y, z, \tag{5.1}$$

we see from Eqs. (2.1), (4.5), and (4.7)–(4.9) that

$$\frac{d\sigma_{r,-}(t)}{dt}+(i\epsilon+\gamma_1)\sigma_{r,-}(t)-i\phi_{r,t}\sigma_{r,z}(t)=0 \tag{5.2}$$

and

$$\frac{d\sigma_{r,z}(t)}{dt}+\gamma_2\sigma_{r,z}(t)-2i(\overline{\phi}_{r,t}\sigma_{r,-}(t)-\text{h.c.})=\gamma_2\eta I. \tag{5.3}$$

Thus, the spin vector $\sigma_r(t)=(\sigma_{r,x}(t),\sigma_{r,y}(t),\sigma_{r,z}(t))$ evolves according to a linear inhomogeneous equation

$$\frac{d\sigma_r(t)}{dt}=b_r(t)\sigma_r(t)+cI, \tag{5.4}$$

where $b_r(t)$ is a linear transformation of $\mathbf{C}^3$, whose explicit form is obtained from the equivalence of Eq. (5.4) to the pair of equations (5.2) and (5.3), and

$$c=(0,0,\gamma_2\eta). \tag{5.5}$$

The solution of Eq. (5.4) may be expressed in the form

$$\sigma_r(t)=g_r(t,0)\sigma_r(0)+\int_0^t dt'\ g_r(t,t')cI, \tag{5.6}$$

where g_r is the Green function for b_r, determined by the formula

$$\frac{\partial}{\partial t}\,g_r(t,t')=b_r(t)g_r(t,t') \quad \forall t\geqslant t'\geqslant 0;\ \ g_r(t,t)=I_3, \tag{5.7}$$

and I_3 is the identity operator in $\mathbf{C}^3$.

Lemma 5.1: For fixed t', $g(t,t')$ decays to zero, with increasing t, at least as fast as $\exp(-\gamma t)$, *where* $\gamma=\min(\gamma_1,\gamma_2)$.

Proof: Let ξ_r be an arbitrary element of $\mathbf{R}^3$, and, for fixed $t'(\in\mathbf{R}_+)$ and any $t>t'$, let $\xi_r(t)=(\xi_{r,x}(t),\xi_{r,y}(t),\xi_{r,z}(t)):=g_r(t,t')\xi_r$. Then, by Eq. (5.7), $\xi_r(t)$ satisfies the homogeneous linear equation obtained by replacing σ_r by ξ_r in Eq. (5.4) and discarding the term cI on the rhs. Hence, defining $\xi_{r,\pm}=1/2(\xi_{r,x}\pm i\xi_{r,y})$, it follows from the equivalence between Eq. (5.4) and the pair of equations (5.2) and (5.3) that the equation of motion for $\xi_r(t)$ is also given by replacing σ by ξ in the latter equations and discarding the rhs of (5.3). Thus,

$$\frac{d}{dt}\,\xi_{r,-}(t)+(\gamma_1-i\epsilon)\xi_{r,-}(t)-i\phi_{r,t}\xi_{r,z}(t)=0$$

and

$$\frac{d}{dt}\,\xi_{r,z}(t)+\gamma_2\xi_{r,z}(t)-2i(\overline{\phi}_{r,t}\xi_{r,-}(t)-\text{c.c.})=0.$$

Hence,

$$\begin{aligned}\frac{d}{dt}\,|\xi_r(t)|^2&\equiv\frac{d}{dt}\,(4|\xi_{r,-}(t)|^2+\xi_{r,z}(t)^2)\\&=-8\gamma_1|\xi_{r,-}(t)|^2-2\gamma_2\xi_{r,z}(t)^2\\&=-2[\gamma_1\xi_{r,x}(t)^2+\gamma_1\xi_{r,y}(t)^2+\gamma_2\xi_{r,z}(t)^2]\leqslant-2\min(\gamma_1,\gamma_2)|\xi_r(t)|^2.\end{aligned}$$

Consequently, defining $\gamma:=\min(\gamma_1,\gamma_2)$,

$$\frac{d}{dt}\left(|\xi_r(t)|^2 \exp(2\gamma t)\right) \leqslant 0,$$

which implies that $|\xi_r(t)|$, and hence $g_r(t,t')$, decay to zero at least as fast as $\exp(-\gamma t)$, with increasing t.

Comment: This lemma signifies that the first term on the rhs of Eq. (5.6) is merely a transient, which decays in a time, γ^{-1}, of the order of the lifetime of an atomic excited state. Consequently, by (5.5), the asymptotic form for $\sigma_r(t)$, obtained by discarding the transients, is

$$\sigma_r(t) = \theta_r(t) I, \tag{5.8}$$

where

$$\theta_r(t) = \int_0^t dt'\ g_r(t,t')(0,0,\gamma_2 \eta). \tag{5.9}$$

Hence, since the state ω_t is given by its action on the finite monomials in the components of the σ_r's, the following Proposition is a simple consequence of these last two equations, together with Proposition 4.1.

Proposition 5.2: The asymptotic form of the time-dependent state ω_t carries no correlations between the observables at the different sites and is therefore of the form

$$\omega_t = \bigotimes_{r \in \mathbf{Z}} \omega_{r,t}, \tag{5.10}$$

where the atomic state $\omega_{r,t}$ is given by

$$\omega_{r,t}(\sigma_r) = \theta_r(t). \tag{5.11}$$

Comment: It follows immediately from this Proposition that the asymptotic state of the model is completely determined by the form of θ_r. We shall denote the components of θ_r analogously with those of σ_r: thus $\theta_r = (\theta_{r,x}, \theta_{r,y}, \theta_{r,z})$ and $\theta_{r,\pm} = \frac{1}{2}(\theta_{r,x} \pm i\theta_{r,y})$. The following Proposition provides an explicit formula for θ_r in terms of the asymptotic forms of the macroscopic variables s and p.

Proposition 5.3: The function θ_r is given by the equations

$$\theta_{r,-}(t) = \sum_{l=0}^{n-1} s_{l,t} \exp(2\pi i r l/n) \tag{5.12}$$

and

$$\theta_{r,z}(t) = \sum_{l=0}^{n-1} p_{l,t} \exp(2\pi i r l/n). \tag{5.13}$$

Proof: By Eqs. (5.2), (5.3), (5.7), and (5.9), the equations of motion for θ_r are formally the same as those for σ_r, i.e.,

$$\frac{d\theta_{r,-}(t)}{dt} + (i\epsilon + \gamma_1)\theta_{r,-}(t) - i\phi_{r,t}\theta_{r,z}(t) = 0 \tag{5.14}$$

and

$$\frac{d\theta_{r,z}(t)}{dt} + \gamma_2 \theta_{r,z}(t) - 2i(\bar{\phi}_{r,t}\theta_{r,-}(t) - \text{c.c.}) = \gamma_2 \eta. \tag{5.15}$$

Further, defining

$$\hat{s}_r(t)=\sum_{l=0}^{n-1} s_{l,t}\exp(2\pi irl/n) \tag{5.16}$$

and

$$\hat{p}_r(t)=\sum_{l=0}^{n-1} p_{l,t}\exp(2\pi irl/n), \tag{5.17}$$

we see, by elementary Fourier analysis, that Eqs. (3.9b) and (3.9c) are equivalent to

$$\frac{d\hat{s}_r(t)}{dt}+(i\epsilon+\gamma_1)\hat{s}_r(t)-i\phi_{r,t}\hat{p}_r(t)=0 \tag{5.18}$$

and

$$\frac{d\hat{p}_r(t)}{dt}+\gamma_2\hat{p}_r(t)-2i(\overline{\phi}_{r,t}\hat{s}_r(t)-\text{c.c.})=\gamma_2\eta, \tag{5.19}$$

respectively. Hence, the equations of motion (5.14) and (5.15) for $\theta_{r,-}$ and $\theta_{r,z}$ are equivalent to those given by (5.18) and (5.19) for $\hat{s}_r$ and $\hat{p}_r$. Consequently, in the asymptotic regime, where the initial conditions are irrelevant, $\theta_{r,-}=\hat{s}_r$ and $\theta_{r,z}=\hat{p}_r$. By Eqs. (5.16) and (5.17), this is the required result.

The following Corollary is an immediate consequence of Propositions 5.2 and 5.3.

Corollary 5.4: The asymptotic state ω_t is spatially periodic, with periodicity n, i.e.,

$$\omega_t=\omega_t\circ V(n), \tag{5.20}$$

where V is the group of spatial translational automorphisms of $\mathscr{A}$ defined by Eq. (4.1).

A. Explicit asymptotic form of ω_t

Equations (5.10)–(5.13) serve to express the time-dependent microstate ω_t in terms of the macroscopic variables. We shall now present its explicit form for both the normal and coherent phases. We have no means of doing the same thing for the chaotic phase, since we have no explicit solution for the macroscopic dynamics there.

1. The normal phase

Here, by Eqs. (3.13) and (5.10)–(5.13), ω_t takes the stationary asymptotic value ω_∞, defined by the equations

$$\omega_\infty=\bigotimes_{r\in\mathbf{Z}}\omega_{\infty,r} \tag{5.21a}$$

and

$$\omega_{\infty,r}(\sigma_{r,\pm})=0,\quad \omega_{\infty,r}(\sigma_{r,z})=\eta. \tag{5.21b}$$

2. The coherent phase

In this phase, it follows from Eqs. (3.14) and (5.10)–(5.13) that the asymptotic form, ω_t^{coh}, of ω_t is given by the formula

$$\omega_t^{\text{coh}}=\bigotimes_{r\in Z}\omega_{r,t}^{\text{coh}}, \tag{5.22a}$$

where

$$\omega_{r,t}^{\text{coh}}(\sigma_{r,-})=s_l^{(0)}\exp\left(i\left(\frac{2\pi ikr}{n}-\nu t-\theta_0\right)\right),\quad \omega_{r,t}^{\text{coh}}(\sigma_{r,z})=\eta. \tag{5.22b}$$

Comment: By contrast with the thermal equilibrium states of pure phases, the stable states ω_t^{coh} are time dependent. The same is true for the pure chaotic phases, since the macroscopic variables s_t and p_t are time dependent there, too.[14]

VI. A NONEQUILIBRIUM MAXIMUM ENTROPY PRINCIPLE

We shall now establish that, as a further consequence of the results of Sec. V, ω_t maximizes the global entropy density of the matter, as defined on the states with spatial periodicity n, subject to the condition that the limiting value, as $N\to\infty$, of the expectation values of the macroscopic variables $s^{(N)},p^{(N)}$ are s_t,p_t, respectively.

For this purpose, we start by defining $\mathscr{S}_n$ to be the set of spatially periodic states, ψ, on $\mathscr{A}$, with period n, and $\mathscr{S}_{n,x_t}$ to be the subset of these states for which

$$\lim_{N\to\infty}\ \psi(s_l^{(N)})=s_{l,t}\quad\text{and}\quad\lim_{N\to\infty}\ \psi(p_l^{(N)})=p_{l,t}\quad\forall\, l\in[0,n],t\in\mathbf{R}_+ .$$

Equivalently, by Eqs. (3.1) and (3.2), $\mathscr{S}_{n,x_t}$ consists of the states, ψ, of spatial periodicity n, which satisfy the conditions

$$n^{-1}\sum_{r=0}^{n-1}\ \psi(\sigma_{r,-})\exp(-2\pi ilr/n)=s_{l,t} \tag{6.1}$$

and

$$n^{-1}\sum_{r=0}^{n-1}\ \psi(\sigma_{r,z})\exp(-2\pi ilr/n)=p_{l,t}, \tag{6.2}$$

for $l=0,...,n-1$ and $t\in\mathbf{R}_+$.

We formulate the global entropy density functional, s, on $\mathscr{S}_n$ in the following standard way.[15] For $\psi\in\mathscr{S}_n$ and $\Lambda\in L$, we define ρ_Λ^ψ to be the density matrix in $\mathscr{H}(\Lambda)$ representing the restriction of ψ to $\mathscr{A}(\Lambda)$, and $S_\Lambda(\psi)$ to be the corresponding local entropy, i.e.,

$$S_\Lambda(\psi)=-k\ \mathrm{Tr}(\rho_\Lambda^\psi\ \ln\ \rho_\Lambda^\psi), \tag{6.3}$$

k being Boltzmann's constant. We define the global entropy density functional, s, on $\mathscr{S}_n$ by the formula

$$s(\psi)=\lim_{\Lambda\uparrow\mathbf{z}}\frac{S_\Lambda(\psi)}{|\Lambda|}, \tag{6.4}$$

the convergence being guaranteed by the subadditivity of entropy[16] and the periodicity of ψ.

Proposition 6.1: The asymptotic time-dependent state ω_t maximizes the functional s, as restricted to $\mathscr{S}_{n,x_t}$.

Proof: We first note that, by Corollary 5.4, $\omega_t\in\mathscr{S}_n$, and that, by Eqs. (5.10)–(5.13), it satisfies the conditions (6.1) and (6.2). Hence, $\omega_t\in\mathscr{S}_{n,x_t}$.

Next, we observe that, if ψ is an arbitrary element of S_{n,x_t}, then it follows by elementary Fourier analysis from Eqs. (5.12), (5.13), (6.1), and (6.2) that $\psi(\sigma_r)=\theta_r(t)$, for $r\in[0,n-1]$, and, consequently, by the periodicity of ψ, for all $r\in\mathbf{Z}$. Therefore, by Eq. (5.11), the restrictions of ψ and ω_t to each single site algebra $\mathscr{A}_r$ are identical. In other words, *all the elements of $\mathscr{S}_{n,x_t}$ coincide on each of the single site algebras.*

Further, by the subadditivity of entropy,

$$S_\Lambda(\psi)\leqslant\sum_{r\in\Lambda}\ S_{\{r\}}(\psi),\quad\forall\,\psi\in\mathscr{S}_{n,x_t},$$

while, by Eqs. (5.10) and (6.3),

$$S_\Lambda(\omega_t)=\sum_{r\in\Lambda} S_{\{r\}}(\omega_t).$$

Consequently, as ψ and ω_t coincide on the single site algebras,

$$S_\Lambda(\psi)\leqslant S_\Lambda(\omega_t),$$

and therefore, by Eq. (6.4),

$$s(\psi)\leqslant s(\omega_t) \quad \forall \psi\in\mathscr{S}_{n,x_t},$$

which proves the Proposition.

VII. CONCLUDING REMARKS

We have provided a microscopic picture of the multi-mode laser model that is complementary to the macroscopic one of (I). Our principal new results are the following.

(1) The microscopic dynamics is governed by the classical, macroscopic field, ϕ, which in turn is generated by the collective action of the atoms on the radiation (Proposition 4.1).
(2) This microdynamics serves to drive the system into states in which the observables of different atoms are uncorrelated (Proposition 5.2). These states have particularly simple forms in both the normal and coherent phases, being stationary and space-translationally invariant in the former and periodic wrt both space and time, in the latter.
(3) In each phase, the eventual time-dependent microstate of the system maximizes the specific entropy of the model, subject to the constraints imposed by the instantaneous values of the macroscopic variables. To the best of our knowledge, this is the first example where such a variational principle has been established in a regime far from thermal equilibrium.[17,18]

These results concern (1) an interplay between microphysics and macrophysics, (2) the decorrelation of certain micro-observables in the course of time, and (3) a maximum entropy principle, corresponding to a conditional thermodynamic stability far from thermal equilibrium. Although the model treated here is but a caricature of a laser, one might envisage that counterparts of these results might have at least some measure of validity in the physics of real dissipative systems. For example, one might expect a local version of the maximum entropy principle (3) to prevail in hydrodynamics, in such a way as to impose local equilibrium conditions on the states of fluids.

ACKNOWLEDGMENTS

This work was carried out during a visit of FB to Queen Mary and Westfield College (QMW). He would like to thank the CNR for the financial support that enabled him to make this visit, and to express his gratitude to the staff of QMW for their hospitality.

[1] K. Hepp and E. H. Lieb, Helv. Phys. Acta **46**, 573 (1973); *Dynamical Systems, Theory and Applications*, Springer Lecture Notes in Physics 38, edited by J. Moser (Springer, Heidelberg, 1975), pp. 178–208.
[2] R. H. Dicke, Phys. Rev. **93**, 99 (1954).
[3] R. Graham and H. Haken, Z. Phys. **237**, 31 (1970); H. Haken, *Handbuch der Physik*, Bd. XXV/2C (Springer, Heidelberg, 1970).
[4] G. Alli and G. L. Sewell, J. Math. Phys. **36**, 5598 (1995).
[5] Apart from the difference in their mathematical settings, this model differs from that of HL in the following physical respect. The atoms of HL are pairs of Fermi oscillators, and can thus accommodate 0, 1, or 2 electrons each; whereas those of (I) correspond to two-level atoms, each with precisely one electron.
[6] G. Lindblad, Commun. Math. Phys. **48**, 119 (1976).
[7] G. Morchio and F. Strocchi, J. Math. Phys. **28**, 1912 (1987).
[8] R. Haag and D. Kastler, J. Math. Phys. **5**, 848 (1964).
[9] F. Bagarello and G. Morchio, J. Stat. Phys. **66**, 859 (1992).
[10] P. Vanheuverzwijn, Ann. Inst. Henri Poincaré, Sect. A **29**, 123 (1978); Erratum, *ibid.* **30**, 83 (1979); B. de Moen, P. Vanheuverzwijn, and A. Verbeure, Rep. Math. Phys. **15**, 27 (1979).
[11] D. Ruelle and F. Takens, Commun. Math. Phys. **20**, 167 (1971).
[12] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics* (Pergamon, Oxford, 1984).
[13] In the present model, this is due to the fact that the radiation field is truncated to n modes. Note that here, as in the

nontruncated model of the radiation, the Fourier coefficients of $\phi^{(N)}$ are N-independent multiples of the $\alpha^{(N)}$'s, rather than the a's; consequently, the latter do not appear in the local field algebra when one passes to the limit $N\to\infty$.

[14] If s_t and p_t were constant, then it would follow from Eq. (3.9a) that α_t, and hence x_t, would tend to a constant as t tended to infinity, and consequently that the macroscopic dynamics would not be chaotic.

[15] D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969).

[16] E. H. Lieb and M. B. Ruskai, J. Math. Phys. **14**, 1938 (1973).

[17] A different variational principle, namely, one of minimum entropy production, was established in Ref. 18 for a certain class of systems, in a regime close to thermal equilibrium.

[18] D. Goderis, A. Verbeure, and P. Vets, J. Stat. Phys. **56**, 721 (1989).

Diffusion on braided spaces

U. Franz[a)]
Institut E. Cartan and CRIN-CNRS, BP 239, Université Henri Poincaré-Nancy I, F-54506 Vandœuvre-lès-Nancy, France and Arnold Sommerfeld Institut für math. Physik, Technische Universität Clausthal, Leibnitzstr. 10, D-38678 Clausthal-Zellerfeld, Germany

R. Schott[b)]
Institut E. Cartan and CRIN-CNRS, BP 239, Université Henri Poincaré-Nancy I, F-54506 Vandœuvre-lès-Nancy, France

(Received 26 January 1996; accepted for publication 22 January 1998)

The notion of q-Brownian motion introduced by Majid is extended to braided spaces corresponding to a generic R-matrix, and combined with the theory of quantum probability. This leads to a definition of diffusions on these spaces. The corresponding heat equations (difference-differential equations) are solved in terms of Appell polynomials (i.e., shifted moment systems). Some examples of interest for applications are given. © *1998 American Institute of Physics.* [S0022-2488(98)00805-6]

I. INTRODUCTION

The applications of diffusions in physics go far beyond the description of the physical phenomenon they are named after. Functional integrals can be used to solve partial differential equations, cf. the celebrated Feynman–Kac formula. Wiener integrals are very close to Feynman path integrals. Another interesting application of diffusions is the stochastic mechanics of Nelson.[1]

We will present an approach to diffusions on braided spaces here. Diffusions on manifolds are characterized by two properties, the first being their Markov property, i.e., that at every instant t they start again, and their evolution does not depend on their history, but only on their distribution at time t. This gives rise to a semi-group and this property will also play a central role for the study of diffusions on braided spaces. The other property, i.e., that they have continuous sample paths, does not have a direct counterpart on braided spaces. We replace it by the condition that diffusions can be obtained as the limit of a (simple) random walk.

The motivations for our approach come from two directions. First, there is Majid's random walk approach to Brownian motion on the braided line and on anyspaces.[2,3] We extend his definition to (pseudo-) diffusions on multi-dimensional braided spaces and, using coalgebraic limit theorems due to Schürmann,[4] give their explicit form. This allows one to consider semigroups of functionals and Markovian transition operators, as well as the associated heat equations, and to introduce Appell systems as their polynomial solutions. Heat kernels, i.e., the densities of the functionals, are also considered.

The second ingredient, Schürmann's theory of quantum Lévy processes, comes into play to assure the existence of the associated processes, e.g., as operators on a Hilbert space. We propose a definition of *-structures for braided Hopf algebras (differing from that due to Majid), and give several examples that satisfy our axioms. Then we generalize Schürmann's definitions and a result assuring the positivity of the convolution of positive functionals under certain conditions to braided groups.

To distinguish the two levels of the constructions, we call the functionals and semigroups obtained via Majid's construction pseudo-diffusions, and reserve the name diffusion for the case where the processes can be realized as operators and are Lévy processes (i.e., independent increment processes) in the sense of Schürmann.

This paper starts with a brief summary of the definition of braided spaces and their main

[a)]Electronic mail: franzu@loria.fr
[b)]Electronic mail: schott@loria.fr

 © 1998 American Institute of Physics

properties in Sec. II. We also introduce some basic notions of quantum probability such as independence and Lévy processes, including a few generalizations that are necessary to apply the existing theory to braided spaces.

The next section (Sec. III) shows how (pseudo-) diffusion processes can be constructed on generic braided spaces associated to a pair of matrices (R,R'). The approach is based on the idea that diffusions can be approximated by random walks on lattices. This replaces the classical condition of the continuity of sample paths. We obtain the explicit form of the limiting process as the chosen time interval and the lattice parameters tend to zero.

Section IV investigates the relation between the diffusion processes and associated evolution equations. These equations are of the form $(\partial_t - L)u = 0$, where L is an operator consisting of linear and quadratic terms in the braided-partials ∂^i. Appell systems (introduced as shifted moment polynomials) are shown to be solutions of the evolution equation. We investigate their properties.

In Sec. V we discuss briefly how density functions can be introduced.

We close with a few final remarks in Sec. VI.

II. PRELIMINARIES

A. Braided spaces

Let us recall the notion of a braided group. This is $(B,m,1,\underline{\Delta},\epsilon,\underline{S},\Psi)$ where $(B,m,1)$ is a unital associative algebra, $(B,\underline{\Delta},\epsilon)$ is a coalgebra, Ψ is a braiding (satisfies the braid relations), and $\underline{S}$ is an antipode (i.e., $m\circ(\mathrm{id}\otimes\underline{S})\circ\underline{\Delta} = m\circ(\underline{S}\otimes\mathrm{id})\circ\underline{\Delta} = 1\circ\epsilon$). The difference that distinguishes braided groups from quantum groups (Hopf algebras) is that now $\underline{\Delta}$ is braided-multiplicative, i.e.,

$$\underline{\Delta}\circ m = m_\Psi^{\otimes}\circ\underline{\Delta},$$

where $m_\Psi^{\otimes}$ is defined by $m_\Psi^{\otimes}((a\otimes c)(b\otimes d)) = a\Psi(c\otimes b)d$. For a more detailed and elegant introduction, see e.g., Refs. 5 and 6.

Braided spaces are braided groups with a particular form of the product and coproduct (see below). They are a generalization of superspaces; there are braided lines, braided planes, braided matrices, etc., all in analogy with superlines, superplanes, etc. Braided spaces are the elementary building blocks of the ''braided world'' or braided linear algebra.

For a pair of invertible matrices (R,R') that satisfy

(i) $R_{12}R_{13}R_{23} = R_{23}R_{13}R_{12}$ (i.e., the Quantum–Yang–Baxter equation),
(ii) $R_{12}R_{13}R'_{23} = R'_{23}R_{13}R_{12}$, $R_{23}R_{13}R'_{12} = R'_{12}R_{13}R_{23}$,
(iii) $(PR+1)(PR'-1) = 0$,
(iv) $R_{21}R'_{12} = R'_{21}R_{12}$,

where P is the permutation matrix $(\mathbf{x}_2\otimes\mathbf{x}_1 P = \mathbf{x}_1\otimes\mathbf{x}_2)$, and $R_{12} = R\otimes\mathrm{id}$, $R_{23} = \mathrm{id}\otimes R$, etc., there exist two braided Hopf algebras $V(R,R')$ and $V^{\check{}}(R,R')$ with generators $v_1,\ldots,v_n$ and $x_1,\ldots,x_n$ and relations
(i) for $V(R,R')$,

$$\underline{\Delta}v^i = v^i\otimes 1 + 1\otimes v^i, \quad \Psi(v^i\otimes v^j) = R^{i\ j}_{k\ l}v^l\otimes v^k,$$

$$v^i v^j = R'^{i\ j}_{k\ l}v^l v^k, \quad \underline{S}(v^i) = -v_i, \quad \epsilon(v^i) = 0;$$

(ii) and for $V^{\check{}}(R,R')$:

$$\underline{\Delta}x_i = x_i\otimes 1 + 1\otimes x_i,$$

$$\Psi(x_i\otimes x_j) = x_l\otimes x_k R^{k\ l}_{i\ j},$$

$$x_i x_j = x_l x_k R'^{k\ l}_{i\ j}, \quad \underline{S}(x_i) = -x_i, \quad \epsilon(x_i) = 0$$

(see Ref. 7, Theorem 1). Here the first two indices (i,k) correspond to the first factor in the tensor product, and the last two (j,l) to the second factor. We use the summation convention, i.e., indices that appear twice (once as subscript and once as superscript) are summed over. In a shorter notation this becomes

(i) for $V(R,R')$: $\underline{\Delta}\mathbf{v}=\mathbf{v}\otimes 1+1\otimes\mathbf{v}$, $\Psi(\mathbf{v}_1\otimes\mathbf{v}_2)=R_{12}\mathbf{v}_2\otimes\mathbf{v}_1$, $R'_{12}\mathbf{v}_2\mathbf{v}_1=\mathbf{v}_1\mathbf{v}_2$, $\underline{S}(\mathbf{v})=-\mathbf{v}$, $\epsilon(\mathbf{v})=0$,

(ii) for $V^{\check{}}(R,R')$: $\underline{\Delta}\mathbf{x}=\mathbf{x}\otimes 1+1\otimes\mathbf{x}$, $\Psi(\mathbf{x}_1\otimes\mathbf{x}_2)=\mathbf{x}_2\otimes\mathbf{x}_1R_{12}$, $\mathbf{x}_2\mathbf{x}_1R'_{12}=\mathbf{x}_1\mathbf{x}_2$, $\underline{S}(\mathbf{x})=-\mathbf{x}$, $\epsilon(\mathbf{x})=0$.

The braided Hopf algebras $V(R,R')$ and $V^{\check{}}(R,R')$ are called the braided vector space and the braided covector space associated to (R,R'). For a detailed development of their properties see also Refs. 5, 6, and 8. We will summarize here only some facts that are used in this presentation.

Let $\mathcal{S}\{1,...,n\}$ be the set of all finite sequences composed of the elements 1,...,n, i.e., $\mathcal{S}\{1,...,n\}=\{\emptyset;1,...,n;11,12,...,nn;111,...\}$. Set $\mathbf{v}^a=v^{a_k}\cdots v^{a_1}$ if $a=(a_1\cdots a_k)\in\mathcal{S}\{1,...,n\}\backslash\{\emptyset\}$, and $\mathbf{v}^{\emptyset}=1$, and analagously, but in reversed order, for $\mathbf{x}_a=x_{a_1}\cdots x_{a_k}$. Then $\{\mathbf{v}^a;a\in\mathcal{S}\{1,...,n\}\}$, $\{\mathbf{x}_a;a\in\mathcal{S}\{1,...,n\}\}$ span $V(R,R')$ and $V^{\check{}}(R,R')$, respectively. Notice that in general these sets do not form a basis. But a basis can easily be extracted from them, if one orders them, and eliminates all elements that are linearly dependent on those that precede them.

The algebras $V(R,R')$ and $V^{\check{}}(R,R')$ are $\mathbb{N}$-graded with

$$V(R,R')^{(r)}=\text{span}\ \{\mathbf{v}^a;a\in S^r\{1,...,n\}\},$$

$$V^{\check{}}(R,R')^{(r)}=\text{span}\ \{x_a;a\in S^r\{1,...,n\}\},$$

where $\mathcal{S}^r\{1,...,n\}\subseteq\mathcal{S}\{1,...,n\}$ contains only the sequences of length r. This allows us to define a scaling $s(\lambda):V^{\check{}}(R,R')\to V^{\check{}}(R,R')$ by $s(\lambda)a=\lambda^{\deg(a)}a$ for homogeneous elements $a\in V^{\check{}}(R,R')$ or $V(R,R')$.

Here $V(R,R')$ and $V^{\check{}}(R,R')$ are mutually dual with the definition

$$\langle\mathbf{v}^a,\mathbf{x}_a\rangle=\delta_{|a|,|b|}[|a|,R]!^a_b,$$

where $|a|,|b|$ is the length of $a,b\in\mathcal{S}\{1,...,n\}$ and

$$[m;R]=1+(PR)_{12}+(PR)_{12}(PR)_{23}+\cdots+(PR)_{12}\cdots(PR)_{m-1,m}, \tag{1}$$

$$[m;R]!=[2;R]_{m-1,m}[3;R]_{m-2,m-1,m}\cdots[m;R]_{1,...,m} \tag{2}$$

(cf. Ref. 6, Section 5). This defines also an action of $V(R,R')$ on $V^{\check{}}(R,R')$ (and *vice versa*) by

$$\rho(v^i)=\partial^i=(\langle v^i,\cdot\rangle\otimes\text{id})\circ\underline{\Delta},$$

$$\rho(x_i)=\hat{\partial}_i=(\langle\cdot,x_i\rangle\otimes\text{id})\circ\underline{\Delta}.$$

Braided exponentials $\exp(\mathbf{x}|\mathbf{v})$ are defined as solutions of

$$\partial^i\ \exp(\mathbf{x}|\mathbf{v})=\exp(\mathbf{x}|\mathbf{v})v^i,\quad(\epsilon\otimes\text{id})\exp(\mathbf{x}|\mathbf{v})=1,$$

$$\exp(\mathbf{x}|\mathbf{v})\hat{\partial}_i=x_i\ \exp(\mathbf{x}|\mathbf{v}),\quad(\text{id}\otimes\epsilon)\exp(\mathbf{x}|\mathbf{v})=1.$$

If the braided integers $[m;R]$ are all invertible, the $\exp(\mathbf{x}|\mathbf{v})$ is given by

$$\exp(\mathbf{x}|\mathbf{v})=\sum\ \mathbf{x}_a([m;R]!^{-1})^a_b\mathbf{v}^b.$$

We will assume that this is the case in what follows, or rather that there exists a braided-exponential $\exp(\mathbf{x}|\mathbf{v})=\Sigma\mathbf{x}_aF(m;R)^a_b\mathbf{v}^b$ (see Ref. 6, Section 5.4).

B. Quantum probability and quantum Lévy processes

A quantum probability space is defined as a pair $(\mathscr{A},\Phi)$ consisting of a *-algebra $\mathscr{A}$ and a state (i.e., a normed positive linear functional) Φ on $\mathscr{A}$, and a quantum random variable j over a quantum probability space $(\mathscr{A},\Phi)$ on a *-algebra $\mathscr{B}$ is a *-algebra homomorphism $j:\mathscr{B}\to\mathscr{A}$. In particular, every element a of $\mathscr{A}$ defines a random variable on the free algebra $\mathbb{C}\langle z,z^*\rangle$ with two generators z,z^* (with the obvious *-structure); simply set $j(z)=a$, $j(z^*)=a^*$, and extend as a*-algebra homomorphism. If a is self-adjoint, then we can take $\mathscr{B}=\mathbb{C}[z]$, with the *-structure defined by $z^*=z$, i.e., the algebra of complex-valued polynomials on the real line.

A central notion to this work is that of independence. Independence in quantum probability is closely related to the notion of products of (*-)algebra, as, e.g., the tensor product or the free product. The kind of independence we use here is related to tensor product of algebras in braided categories. It generalizes the tensorial independence introduced by Schürmann for braidings that arise from an action α and a coaction γ of some group algebra $\mathbb{CL}$ as $\Psi=(\alpha\otimes\mathrm{id})\circ(\mathrm{id}\otimes\tau)\circ(\gamma\otimes\mathrm{id})$ or $\Psi=(\mathrm{id}\otimes\alpha)\circ(\tau\otimes\mathrm{id})\circ(\mathrm{id}\otimes\gamma)$. For a more general definition of independence see Ref. 9; for a general analysis of the notion of independence we refer to Refs. 10 and 11.

Definition 1: The quantum random variables $j_1,j_2,\ldots,j_n:\mathscr{B}\to\mathscr{A}$ are called Ψ-independent (wrt the state Φ), if

(i)

$$\Phi(j_{\sigma(1)}(b_1)\cdots j_{\sigma(n)}(b_n))=\Phi(j_{\sigma(1)}(b_l))\cdots\Phi(j_{\sigma(n)}(b_n))$$

for all permutations $\sigma\in\mathscr{S}(n)$ and all $b_1,\ldots,b_n\in\mathscr{B}$, and

(ii)

$$m_{\mathscr{A}}\circ(j_k\otimes j_l)=m_{\mathscr{A}}\circ(j_l\otimes j_l)=m_{\mathscr{A}}\otimes(j_l\otimes j_k)\circ\Psi$$

for all $1\leqslant k<l\leqslant n$.

Remark: If $\mathscr{A}$ is commutative, then (i) is equivalent to the condition used in classical probability theory:

(i′)

$$\Phi(j_1(b_l)\cdots j_n(b_n))=\Phi(j_l(b_1))\cdots\Phi(j_n(b_n))$$

for all $b_1,\ldots,b_n\in\mathscr{B}$.

We will call $j_1,\ldots,j_n$ pseudo-(Ψ-)independent, if they satisfy only (i′) and (ii). For many of the algebraic aspects this behaves like ''true'' independence.

Let us investigate when two quantum random variables on a braided space $V^{\vee}$ are (pseudo-) independent. Condition (ii) implies that $j_1(x_i)$ and $j_2(x_j)$ satisfy the braid relations. Therefore we can assume $\mathscr{A}=V^{\vee}\underline{\otimes}V^{\vee}$, and $j_1(u)=u\otimes 1$, $j_2(u)=1\otimes u$ [take $\hat{\Phi}=\Phi\circ(j_1\otimes j_2)$]. Condition (i′), i.e., pseudo-independence, is now equivalent to the fact that $\hat{\Phi}$ can be factorized into $\hat{\Phi}=\phi^{(2)}\otimes\phi^{(1)}$, were $\phi^{(i)}=\Phi\circ j_i$, $i=1,2$. Thus the joint distribution of j_1 and j_2 is uniquely determined by the marginal distributions.

Condition (i) is stronger; for independence in the sense of the preceding definition we need a kind of invariance or commutativity condition in addition to the factorization, namely $\phi^{(1)}\otimes\phi^{(2)}(u\otimes v)=\phi^{(2)}\otimes\phi^{(1)}(\Phi(v\otimes u))$, or $\phi^{(1)}\otimes\phi^{(2)}=\Psi(\phi^{(2)}\otimes\phi^{(1)})$, i.e., $\underline{\varepsilon}\otimes\phi^{(2)}$ and $\phi^{(1)}\otimes\underline{\varepsilon}$ have to commute.

We will study here time and space homogeneous Markov processes, i.e., Lévy processes.

Definition 2: A quantum Lévy process on a (braided)-Hopf algebra $\mathscr{B}$ is a quantum stochastic process $\{j_t:\mathscr{B}\to(\mathscr{A},\Phi)\,|\,t\in\mathbb{R}_+\}$, such that*

(i) $\{j_t|t\in\mathbb{R}_+\}$ *has independent increments, i.e., $j_{s_1t_1},\ldots,j_{s_nt_n}$ are (Ψ-)independent for all $0\leqslant s_1\leqslant t_1\leqslant\cdots\leqslant s_n\leqslant t_n$, where the increments are defined by $j_{st}=(j_s\circ S)\star j_t=m_{\mathscr{A}}\circ((j_s\circ S)\otimes j_t)\circ\Delta$,*

(ii) *the distribution $\Phi\circ j_{st}$ of an increment j_{st} depends only on $t-s$ (stationarity),*

(iii) *j_{st} converges to j_{ss} in distribution for $t\searrow s$ (weak continuity).*

We shall speak of pseudo-Lévy processes, if the increments are only pseudo-independent. In this case we will not insist on the positivity of the state either. We study pseudo-Lévy processes (and pseudo-diffusions, see below) only to illustrate the algebraic part of the theory.

It follows from the general theory (cf. Ref. 4) that a Lévy process is uniquely determined by its marginal distributions $\varphi_t=\Phi\circ j_t$, that the marginal distribution form a convolution semigroup, i.e., $\varphi_0=\underline{\varepsilon}$, $\varphi_s\star\varphi_t=(\varphi_s\otimes\varphi_t)\circ\Delta=\varphi_{s+t}$, and that there exists a unique functional L, called generator, such that $\varphi_t=\exp(tL)$. *Vice versa*, the semigroup, and thus the process, is also uniquely determined by its generator. All this is also true for pseudo-Lévy processes.

Let us now look at the differences. The independence of the increments implies $\Psi(e^{tL}\otimes e^{sL})=e^{sL}\otimes e^{tL}$, and thus $\Psi(L\otimes L)=L\otimes L$, i.e., $L\otimes\underline{\varepsilon}$ and $\underline{\varepsilon}\otimes L$ commute (in the convolution algebra of functionals on $V\check{}\underline{\otimes}V\check{}$). Also, since the semigroup is positive, L is conditionally positive (i.e., positive on the kernel of $\underline{\varepsilon}$), to see this just differentiate wrt t, and Hermitian. In the sequel, when we speak of Lévy processes, we will always assume that our states ϕ (on $\mathscr{A}$) are Ψ-invariant in the sense that $\Psi(\phi\otimes u)=u\otimes\phi$ for all functionals $u:\mathscr{B}\to\mathbb{C}$, i.e., that we can pull it across braidings.

If L is a Hermitian, conditionally positive, and Ψ-invariant functional, then all the convolution powers of L are positive (see Lemma 4), and the semigroup associated to L is positive. Thus there exists a Lévy process with $\varphi_t=\exp(tL)$ as marginal distributions.

We summarize this in the following proposition.

Proposition 3: Let $\mathscr{A}$ be a (braided) (-)Hopf algebra (cf. Sec. II C).*

(i) *There is a one-to-one correspondence between pseudo-Lévy processes $\{j_t|t\in\mathbb{R}_+\}$, normed convolution semi-groups $\{\varphi_t|t\in\mathbb{R}_+\}$, $\varphi_t=\exp(tL)=\Phi\circ j_t$, and functionals $L=d/dt\ \varphi_t|_{t=0}$ with $L(1)=0$ on $\mathscr{A}$.*

(ii) *There is also a one-to-one correspondence between Lévy processes, normed positive convolution semigroups of Ψ-invariant states, and Ψ-invariant, Hermitian, conditionally positive functionals.*

C. A remark on braided*-Hopf algebras

Majid has studied *-structures on braided Hopf algebras in a series of articles.[12–14] An essential result of his work is a new set of axioms for the braided case. However, we found it necessary to use different axioms here; in particular, it is important for our purposes that the coproduct is a *-algebra homomorphism.

We fix the *-structure in the tensor product $\mathscr{A}\underline{\otimes}\mathscr{A}$ by the condition that the canonical inclusions $\mathscr{A}\xrightarrow{i_1}\mathscr{A}\underline{\otimes}\mathscr{A}\xrightarrow{i_2}\mathscr{A}$ are *-algebra homomorphism, i.e., $(1\otimes a)^*=1\otimes a^*$ and $(a\otimes 1)^*=a^*\otimes 1$ for all $a\in\mathscr{A}$. This implies $(a\otimes b)^*=\psi(b^*\otimes a^*)$ for all $a,b\in\mathscr{A}$. Then we say that $\mathscr{A}$ is a (braided) *-Hopf algebra if $(\mathscr{A},*)$ is a *-algebra, and $\underline{\Delta}$ and $\underline{\varepsilon}$ are *-algebra homomorphism (with the *-structure of $\mathscr{A}\underline{\otimes}\mathscr{A}$ defined as above!).

For examples of braided *-Hopf algebras satisfying these conditions, see Sec. III B.

Let us outline the main step and ideas of their construction. We start with a braided-covector space $V\check{}(R,R')$ with R,R' of real type I, i.e., $\overline{R^{i\ k}_{j\ l}}=R^{l\ j}_{k\ i}$. Majid has defined a map $*$ for this case that goes from $V\check{}(R,R')$ to $V(R,R')$, and *vice versa* (cf. Ref. 13, Lemma 2.1). The free product $V(R,R')*V\check{}(R,R')$ thus becomes a *-algebra with this map, and the coproducts of $V(R,R')$ and $V\check{}(R,R')$ define a *-algebra homomorphism $\underline{\Delta}_0:V(R,R')*V\check{}(R,R')\to(V(R,R')\otimes(R,R'))*(V\check{}(R,R')\underline{\otimes}V\check{}(R,R'))$. Now it turns out that $(V(R,R')*V\check{}(R,R'))\otimes(V(R,R')*V\check{}(R,R'))$ can be obtained as a quotient of $(V(R,R')\underline{\otimes}V(R,R'))*(V\check{}(R,R')\otimes V\check{}(R,R'))$ in such a way that the canonical projection is a *-algebra homomorphism, and that $V(R,R')*V\check{}(R,R')$ is a braided *-Hopf algebra in the sense of our definition with $\underline{\Delta}=\pi\circ\underline{\Delta}_0$. Set $x_{1i}=x_i\otimes 1$, $v^i_1=v^i\otimes 1$, $x_{2i}=1\otimes x_i$, and $v^i_2=1\otimes v^i$. Then the relations defining the braided *-Hopf algebra structure of $V(R,R')*V\check{}(R,R')$ are

$$x_ix_j=x_lx_kR'^{k\ \ l}_{i\ \ j},\quad v^jv^i=R'^{j\ \ i}_{l\ \ k}v^kv^l,$$

$$x_i^*=v^i,\quad (v^i)^*=x_i,$$

$$\underline{\Delta}(x_i)=x_{1i}+x_{2i}, \quad \underline{\Delta}(v^i)=v_1{}^i+v_2{}^i,$$

$$\underline{S}(x_i)=-x_i, \quad \underline{S}(v^i)=-v^i, \quad \underline{\varepsilon}(x_i)=\underline{\varepsilon}(v^i)=0$$

$$x_{\alpha,i}x_{\alpha,j}=x_{\alpha,l}x_{\alpha,k}R'^{k}{}_{i}{}^{l}{}_{j}, \quad v_\alpha{}^j v_\alpha{}^i=R'^{j}{}_{l}{}^{i}{}_{k}v_\alpha{}^k v_\alpha{}^l, \quad \alpha=1,2,$$

$$(x_{\alpha,i})^*=v_\alpha{}^i, \quad (v_\alpha{}^i)^*=x_{\alpha,i}, \quad \alpha=1,2,$$

$$x_{2i}x_{1j}=R^{k}{}_{i}{}^{l}{}_{j}x_{1l}x_{2k}, \quad v_2{}^i x_{1j}=(R^{-1})^{i}{}_{k}{}^{l}{}_{j}x_{1l}v^{k}{}_{2},$$

$$v_1{}^j x_{2i}=(R^{-1})^{j}{}_{l}{}^{k}{}_{i}x_{2k}v_1{}^l, \quad v_1{}^j v_2{}^i=R^{j}{}_{l}{}^{i}{}_{k}v_2{}^k v_i{}^l.$$

We observe that, with the definitions $x_i^*=v_i$, $x_{\alpha,i}^*=v^i{}_\alpha$, the '$v$-$v$' relations follow from the corresponding 'x-x' relations, and the second line of the braid relations from the first (where we use that R, R' are real type I).

Note also that $x_{1i}v_1{}^i, v_1{}^i x_{1i}$ (or $x_{2i}v_2{}^i, v_2{}^i x_{2i}$, resp.), $i=1,...,n$, (no summation!) commute with all $x_{2j}, x_2{}^j$ ($x_{1j}, x_1{}^j$, resp.). Thus the functionals that are nonzero only on these elements are Ψ-invariant.

We prove the following.

*Lemma 4: Let $\mathscr{A}$ be a braided *-bialgebra (in the sense of the preceding definition, i.e., the *-structure in the braided product algebra $\mathscr{A}\underline{\otimes}\mathscr{A}$ is given by $\Psi\circ(*\otimes *)\circ\tau$, and $\underline{\Delta}$ is a *-algebra homomorphism), and let ϕ be a Ψ-invariant positive functional on $\mathscr{A}$, and θ be a positive functional on $\mathscr{A}$. Then $\phi\star\theta=(\phi\otimes\theta)\circ\underline{\Delta}$ is also positive.*

Proof: Let $a\in\mathscr{A}$, and $\underline{\Delta}(a)=\Sigma_i a_i^{(1)}\otimes a_i^{(2)}$ (Sweedler's notation). Then $\underline{\Delta}(a^*)=\Sigma_j\Psi((a_j^{(2)})^*\otimes(a_j^{(1)})^*)$ and therefore $(\phi\otimes\theta)\circ\underline{\Delta}(a^*a)=(\phi\otimes\theta)\circ(m\otimes m)\circ(\mathrm{id}\otimes\Psi\otimes\mathrm{id})\circ(\Psi\otimes id\otimes id)(\Sigma_{i,j}(a_j^{(2)})^*\otimes(a_j^{(1)})^*\otimes a_i^{(1)}\otimes a_i^{(2)}) = \theta\circ(\phi\otimes m)\circ(\Psi\circ\mathrm{id})\circ(\mathrm{id}\otimes m\otimes\mathrm{id})(\Sigma_{i,j}(a_j^{(2)})^*\otimes(a_j^{(1)})^*\otimes a_i^{(1)}\otimes a_i^{(2)})=\Sigma_{i,j}\Theta((a_j^{(2)})^*a_i^{(2)})\phi((a_j^{(1)})^*a_i^{(1)})$ is positive, since it is the Schur product of two positive definite matrices. ■

III. A CONSTRUCTION OF (PSEUDO-) DIFFUSIONS ON BRAIDED SPACES

Following Majid we first define a quantum random walk. Consider a classical random walk on the lattice. We assume that the process can only jump from a lattice site x to a finite number of other sites $x+d_i$, $i=1,...,k$. If at time t the particle is located at x, then it will be at $x+d_i$ at time $t+\Delta t$ with a probability p_i, $0\leqslant p_i\leqslant_1$, and naturally the probabilities add to one,

$$\sum_{i=1}^{k} p_i=1.$$

In classical probability theory there corresponds an evolution operator to this which maps a function (i.e., an observable) at time t to the corresponding function one time step later

$$f_t(x)\mapsto f_{t+\Delta t}(x)=\sum_{i=1}^{k} p_i f_t(x+d_i).$$

Let $\phi^{(d_i)}$ be the functional corresponding to evaluation of a function at d_i, i.e., $\phi^{(d_i)}$: $f(x)\mapsto f(d_i)$. Then the above equation can be written as

$$f_t(x)\mapsto f_{t+\Delta t}(x)=(\phi\otimes\mathrm{id})\circ\Delta f_t(x),$$

where $\phi=\Sigma p_i\phi^{(d_i)}$ and Δ stands for the usual cocommutative coproduct on $\mathbb{R}[x_1,\ldots,x_n]$. This can be done in the same manner on braided spaces and quantum groups, if one chooses an appropriate functional ϕ. The problem here (in dimension >1) is that one cannot define the $\phi^{(d_i)}$ simply as replacing the variables x_i by the components of the vectors $d_j=(d_{j1},\ldots,d_{jn})$ due to the noncommutativity. One possibility is to bring the variables into some fixed order (i.e., fix a basis of monomials) and replace them then by the d_{ji}. We shall see later that in the limit only the choice on the quadratic and linear terms matters.

A functional $\phi^{(d)}$ that corresponds to evaluation at $d=(d_1,\ldots,d_n)$ should satisfy
F1: $\phi^{(d)}(1)=1$, or equivalently $\phi^{(d)}|_{V^\vee(R,R')^{(0)}}=\epsilon$;
F2: evaluation on the generators gives the components of d, i.e.,

$$\phi^{(d)}(x_i)=d_i,$$

for $i=1,\ldots,n$;
F3: it has the correct behavior with respect to a rescaling of d, more precisely,

$$\phi^{(\lambda d)}(a)=\phi^{(d)}(s(\lambda)a)=\lambda^{\deg(a)}\phi^{(d)}(a)$$

for homogeneous $a\in V^\vee(R,R')$.
However, these conditions do not determine $\phi^{(d)}$ uniquely.
A possible choice for $\phi^{(d)}$ with $d=(d_1,\ldots,d_n)$ is to use the braided-exponential

$$\phi^{(d)}=\exp(d|\mathbf{v})=\sum_{a\in\mathcal{T}\{1,\ldots,n\}} d_a([m;R]!^{-1})^a_b\mathbf{v}^b. \tag{3}$$

This definition makes sense for all cases where the braided-integers $[m;R]$ are invertible, and can also be used if this is not the case, if an appropriate braided-exponential is chosen. For the commutative case we have $\phi^{(d)}f(\mathbf{x})=f(d)$ by Taylor's theorem, and similarly for the braided line $\mathbb{R}_q$ by a q-Taylor theorem. On anyspace this agrees with definition of the Dirac δ-function (see Ref. 3).

Another possibility is to use (positive) Dirac functionals, defined as functionals φ for which there exists a (*-)algebra $\mathcal{A}$, a state Φ on $\mathcal{A}$, and a (*-)algebra homomorphism $j:V^\vee(R,R')\to\mathcal{A}$ such that $\varphi=\Phi\circ j$ and the pair (j,j) is independent. Typically these Dirac functionals replace one generator by a real or complex number, and all other generators by zero, and lead to the same class of diffusions as the functional presented above.

Assume now that we have defined the functionals $\phi^{(d_i)}$, and chosen values for the transition probabilities p_i, i.e., fixed ϕ. Then this defines an evolution for functions, i.e., observables on our space by

$$f_t\mapsto f_{t+\Delta t}=(\phi\otimes\mathrm{id})\circ\underline{\triangle}f_t$$

as before. As explained in Ref. 2 there is a one-to-one correspondence between functionals ϕ and transition operators T; to pass from one to the other take

$$\phi\mapsto T_\phi=(\phi\otimes\mathrm{id})\circ\underline{\triangle},$$

$$T\mapsto\phi_T=\epsilon\circ T.$$

The transition operators form a (discrete) semi-group with respect to composition, and the functionals with respect to convolution. The passage from one to the other is in fact a homomorphism, i.e., $T_\phi\circ T_\psi=T_{\phi\psi}$ and $\phi_T\phi_S=\phi_{TS}$.

We study the limit that arises if the lattice parameters d_i and the time step Δt go to zero. The discrete semi-groups (indexed by $\mathbb{N}$) will then converge to continuous ones (indexed by $\mathbb{R}_+$).

Theorem 5: *Let $d_1,\ldots,d_k$ and $p_1^0,\ldots,p_k^0$ be chosen such that $\Sigma_{i=1}^k p_i^0=1$ and $\phi=\Sigma_{i=1}^k p_i^0\phi^{(d_i)}$ vanishes on $V^\vee(R,R')^{(1)}$, i.e., $\Sigma_{i=1}^k p_i^0 d_{ij}=0$ for all $j=1,\ldots,n$.*

Choose furthermore $p_1^1,\ldots,p_k^1$ such that $p_1(\mu)=p_1^0+\mu p_1^1,\ldots,p_k(\mu)=p_k^0+\mu p_k^1$ satisfy $0\leqslant p_i\leqslant 1$ and $\Sigma_{i=1}^k p_i=1$ for sufficiently small μ (this implies in particular $\Sigma_{i=1}^k p_i^1=0$).

Then for $\phi_N=\Sigma_{i=1}^k p_i(1/\sqrt{N})\phi^{(d_i/\sqrt{N})}$ we have

$$\lim_{N\to\infty}(\phi_N)^N=\exp\left(\sum_{k,l=1}^n a_{kl}v^kv^l+\sum_{j=1}^n b_iv^i\right),$$

where $a_{kl}=\Sigma_{i=1}^k p_i^0\Sigma_{r,s=1}^n d_{ir}d_{is}([2;R]^{-1})^{rs}_{kl}$ and $b_j=\Sigma_{i=1}^k p_i^1 d_{ij}$.

Proof: We apply Ref. 4, Theorem 6.1.1. Let $k_N=N$ and $\varphi_{Nk}=\phi_N$ for $k=1,\ldots,N$. We have to check the conditions of the theorem.

(i) $\varphi_{N1},\ldots,\varphi_{NN}$ commute since they are identical.
(ii) Let $c=\Sigma_{a\in\mathscr{S}\{1,\ldots,n\}}c^a\mathbf{x}_a$ with some coefficients $c^a\in\mathbb{C}$. Then

$$\max_{l\leqslant k\leqslant N}|(\varphi_{Nk}-\epsilon)(c)|=\phi_N(c)-c^{\varnothing}=O(1/\sqrt{N})$$

goes to zero as $N\to\infty$.
(iii)

$$\sum_{k=1}^{N}(\varphi_{Nk}-\epsilon)(c)=N(\phi_N(c)-c^{\varnothing})=\sum_{i=1}^{k}p_i^1\sum_{j=1}^{n}d_{ij}c^{\{j\}}+\sum_{i=1}^{k}p_i^0\sum_{r,s=1}^{n}d_{ir}d_{is}c^{rs}+O(1/\sqrt{N})$$

is bounded and tends to $(\Sigma_{k,l=1}^{n}a_{kl}v^kv^l+\Sigma_{j=1}^{n}b_iv^i)(c)$.
Thus $\lim_{N\to\infty}(\Pi_{k=1}^{N}\varphi_{Nk})(c)=\exp(\Sigma_{k,l=1}^{n}a_{kl}v^kv^l+\Sigma_{j=1}^{n}b_iv^i)(c)$.

We will consider functionals that can be obtained in this way as diffusions on braided spaces and call $L=\Sigma_{k,l=1}^{n}a_{kl}v_kv_l+\Sigma_{j=1}^{n}b_iv^i$ their generator.

Definition 6: We will call a semi-group of functionals $\exp(tL)$ *on* $V^{\check{}}(R,R')$ *a (homogeneous) pseudo-diffusion, if there exist parameters* $d_i=(d_{i1},\ldots,d_{in})$, p_i^0,p_i^1 $(i=1,\ldots,k)$ *such that the conditions of Theorem 5 are satisfied, and such that* $L=\Sigma_{k,l=1}^{n}a_{kl}v^kv^l+\Sigma_{j=1}^{n}b_iv^i$ *with* $a_{kl}=\Sigma_{i=1}^{k}p_i^0\Sigma_{r,s=1}^{n}d_{ir}d_{is}([2;R]^{-1})_{kl}^{rs}$ *and* $b_j=\Sigma_{i=1}^{k}p_i^1d_{ij}$.

If the process associated to $\exp(tL)$ *is also a Lévy process, then we will call it a diffusion.*
Remarks:

(1) We have formulated the limit theorem in terms of the functionals; analogous theorems for the associated transition operators hold also.
(2) In the statements of Theorem 5 we assumed that the braided-integers are invertible. The theorem remains valid if this is not the case; as long as a braided-exponential $\exp(\mathbf{x}|\mathbf{v})=\Sigma\mathbf{x}_aF(m;R)_b^a\mathbf{v}^b$ exists, the inverted braided-exponentials just have to be replaced by the $F(m;R)$.

A. Examples of (pseudo-) diffusions

1. Dimension n=1: The braided line $\mathbb{R}_q$

Here $R=(q)$ and $R'=(1)$. This leads to the braiding $\Psi(x^m\otimes x^p)=q^{mp}x^p\otimes x^m$ (i.e., $x'x=qxx'$, where $x=x\otimes 1$, $x'=1\otimes x$). The diffusions on the braided line have the form

$$\phi_t=\exp\left[t\left(\frac{a^2}{1+q}v^2+bv\right)\right].$$

Using the expansion $e^{-tu^2/2+xu}=\Sigma_{p=0}^{\infty}(u^p/p!)H_p(x,t)$ with the Hermite polynomials $H_p(x,t)=\Sigma_{k=0}^{[p/2]}\binom{p}{2k}[(2k)!/2^kk!]x^{p-2k}(-t)^k$ we find for the moments

$$\phi_t\left(\frac{x^p}{[p;q]!}\right)=\frac{H_p(bt,-2a^2t/(1+q))}{p!}.$$

2. Dimension n=2: The quantum plane $\mathbb{C}_q^{2|0}$

Here

$$R=\begin{pmatrix}q^2&0&0&0\\0&q&q^2-1&0\\0&0&q&0\\0&0&0&q^2\end{pmatrix},\quad R'=q^{-2}R.$$

This is the standard two-dimensional quantum plane. It has two generators x,y with relations $yx=qxy$ and braid statistics $x'x=q^2xx'$, $x'y=qyx'$, $y'y=q^2yy'$, and $y'x=qxy'+(q^2-1)yx'$. For the second braided-integer we obtain

$$[2;R]=\begin{pmatrix} 1+q^2 & 0 & 0 & 0 \\ 0 & 1 & q & 0 \\ 0 & q & q^2 & 0 \\ 0 & 0 & 0 & 1+q^2 \end{pmatrix},$$

i.e. it is not invertible. There exists nonetheless a braided-exponential (Ref. 6, Example 5.4)

$$\exp(\mathbf{x}|\mathbf{v})=\sum_{a\in\mathscr{S}\{1,\ldots,n\}}\frac{\mathbf{x}_a\mathbf{v}^a}{[|a|;q^2]!}=\sum_{p=0}^{\infty}\frac{(\mathbf{x}\cdot\mathbf{v})^p}{[p;q^{-2}]!},$$

where $[m;q^2]=(1-q^{2m})/(1-q^2)$ and $[m;q^2]!=\Pi_{\mu=1}^{m}[\mu;q^2]$. For the generator we obtain

$$L=\frac{1}{1+q^2}\sum a_{ij}\partial_i\partial_j+\sum b_i\partial_i.$$

3. The free braided-space

Here $R'=P$ and there are no relations in the algebras $V(R,R')$ and $V^{\check{}}(R,R')$ since the ideal generated by $R'_{12}\mathbf{v}_2\mathbf{v}_1=\mathbf{v}_1\mathbf{v}_2$ (respectively $\mathbf{x}_1\mathbf{x}_2R'_{12}=\mathbf{x}_2\mathbf{x}_1$) is equal to $\{0\}$, i.e., we have the free algebra with n generators. $\{\mathbf{v}^a;a\in\mathscr{S}\{1,\ldots,n\}\}$ (respectively $\{\mathbf{x}_a;a\in\mathscr{S}\{1,\ldots,n\}\}$) is thus a basis of $V(R,R')$ [respectively $V^{\check{}}(R,R')$].

We will assume that R is also equal to P, i.e., that the braiding is given by $\Psi(v^i\otimes v^j)=(v^i\otimes v^j)$ [respectively $\Psi(x_i\otimes x_j)=(x_i\otimes x_j)$]. In this case the braided-integers $[m;P]_b^a=m\,\delta_{ab}$, $m=|a|=|b|$, are invertible and thus

$$\exp(\mathbf{x}|\mathbf{v})=\sum_{k=0}^{\infty}\frac{1}{k!}\sum_{a\in\mathscr{S}^k\{1,\ldots,n\}}\mathbf{x}_a\mathbf{v}^a.$$

The dual action is

$$\rho(v^i)\mathbf{x}_a=\begin{cases}|a|\mathbf{x}_{a_2\ldots a_{|a|}} & \text{if } a_1=i,\\ 0 \quad \text{else,}\end{cases}$$

on the basis elements. Generators of diffusions have again the form

$$L=\sum_{k,l=1}^{n}a_{kl}\partial_k\partial_l+\sum_{j=1}^{n}b_j\partial_j \tag{4}$$

with symmetric a_{kl}.

B. Examples of true diffusions

1. A free two-dimensional braided plane: $V^{\check{}}=\mathbb{R}_q*\check{\mathbb{R}}_q$

The algebra underlying this example can be considered as the free product of the braided line with its dual. As an algebra it is the free unital algebra $\mathbb{C}\langle x,p\rangle$ with two generators. Let x be the generator of the braided-covector space $\mathbb{R}_q$, and p that of the braided-vector space dual to it (i.e., another copy of $\mathbb{R}_q$). Let $q\in\mathbb{R}\backslash\{0\}$, so that $R=(q)$ is of real type I in the sense of Ref. 13. Following Majid's definitions we have the braid relations

$$x'x=qxx,\quad x'p=q^{-1}px',\quad p'x=q^{-1}xp',\quad p'p=qpp',$$

and

$$*_{\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q}:\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\to\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q,\quad x^*=p,$$

$$*_{\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q}:\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q\to\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q,\quad x^*=p',\ (x')^*=p,$$

turn $\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q$ and $\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q$ into *-algebras such that $\underline{\triangle}\ \mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\to\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q$ satisfies $\underline{\triangle}\circ *_{\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q}=*_{\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q\underline{\otimes}\mathbb{R}_q\underline{\otimes}\check{\mathbb{R}}_q}\circ\tau\circ\underline{\triangle}$. Even though this *-structure does not suit our purposes, at least not in this form, it does tell how we can guess a ''better'' one.

The coproduct on $\mathbb{R}_q$ and its dual extends to a map $\underline{\triangle}_0$: $\mathbb{R}_q*\check{\mathbb{R}}_q\to(\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)=\mathbb{C}\langle p,p',x,x'\rangle/\mathscr{T}_0$, where $\mathscr{T}_0$ is the ideal generated by $x'x-qxx'$ and $p'p-qpp'$, and satisfies $\underline{\triangle}\circ *_{\mathbb{R}_q*\check{\mathbb{R}}_q}=*_{(\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)}\circ\underline{\triangle}$, if we define the *-structures in those algebras by

$$*_{\mathbb{R}_q*\check{\mathbb{R}}_q}:\quad \mathbb{R}_q*\check{\mathbb{R}}_q\to\mathbb{R}_q*\check{\mathbb{R}}_q,\quad x^*=p,$$

$$*_{(\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)}:\quad (\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)\to(\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q),\quad x^*=p',\ (x')^*=p.$$

We now note that we obtain $(\mathbb{R}_q*\check{\mathbb{R}}_q)\underline{\otimes}(\mathbb{R}_q*\check{\mathbb{R}}_q)$ if we quotient by the relations $qxp=px$ and $qp'x'=x'p'$, i.e., $(\mathbb{R}_q*\check{\mathbb{R}}_q)\underline{\otimes}(\mathbb{R}_q*\check{\mathbb{R}}_q)=(\mathbb{R}_q\underline{\otimes}\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)/\{qxp=pxp,qp'x'=x'p'\}=\mathbb{C}\langle p',x,p,x'\rangle/\mathscr{I}$, where $\mathscr{I}$ is the ideal generated by $x'x-qxx'$, $qxp-px$, $qp'x'-x'p'$, and $p'p-qpp'$. The canonical projection π: $(\mathbb{R}_q)\underline{\otimes}(\mathbb{R}_q)*(\check{\mathbb{R}}_q\underline{\otimes}\check{\mathbb{R}}_q)\to(\mathbb{R}_q*\check{\mathbb{R}}_q)\underline{\otimes}(\mathbb{R}_q*\check{\mathbb{R}}_q)$ is a *-algebra homomorphism, and $\underline{\triangle}=\pi\circ\underline{\triangle}_0$ turns $V^{\check{}}=\mathbb{R}_q*\check{\mathbb{R}}_q$ into a braided *-Hopf algebra in the sense we need. In particular, the coproduct is a *-algebra homomorphism. Note that we now have $p_1=p'$ and $x_1=x$ in the first factor of the tensor product, and $p_2=p$ and $x_2=x'$ in the second, i.e., the braid relations take the form

$$x_2x_1=qx_1x_2,\quad p_2x_1=q^{-1}x_1p_2,$$

$$x_2p_1=qp_1x_2,\quad p_2p_1=q^{-1}p_1p_2,$$

and the *-structure becomes $x_i'=p_i$, $i=1,2$.

The braiding of the real line, and thus also that of $V^{\check{}}$, are induced by an action and coaction of the group algebra $\mathbb{CZ}$. $V^{\check{}}$ is the (free) braided plane associated to the matrices

$$R'=\begin{pmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{pmatrix},\quad R=\begin{pmatrix}q&0&0&0\\0&q^{-1}&0&0\\0&0&q&0\\0&0&0&q^{-1}\end{pmatrix}.$$

The possible generators of pseudo-diffusions are of the same form as in Eq. (4), but now we are interested in generators of ''true'' diffusions, i.e., Ψ-invariant, Hermitian, conditionally positive generators. Choose the set of all words $\{1=\varnothing,x,p,xx,xp,px,pp,\ldots\}$ as basis of $V^{\check{}}$. Then ϕ_{xp} and ϕ_{px} with

$$\phi_{xp}(u)=\begin{cases}1 & \text{if } u=xp,\\ 0 & \text{else}\end{cases}\qquad \phi_{px}(u)=\begin{cases}1 & \text{if } u=px,\\ 0 & \text{else,}\end{cases}$$

on basis elements u are such generators. The processes associated to these generators have been studied by Schürmann.[4] We find the processes that have the Azéma martingales as a classical version.

Let us now consider an example whose braiding does not come from a group, but from a noncommutative and non-cocommutative bialgebra.

2. Diffusions on the four-dimensional braided-covector space $V^{\check{}}=\mathbb{C}_q^{2|0}*\check{\mathbb{C}}_q^{2|0}$

Again let q be in $\mathbb{R}\{0\}$, and set $V^{\check{}}=\mathbb{C}_q^{2|0}*\check{\mathbb{C}}_q^{2|0}=\mathbb{C}\langle v,w,x,y\rangle/\{yx-qxy,vw-qwv\}$. A similar construction as in the preceding example leads to a four-dimensional braided-covector space, where the braid relations between x_1,y_1 and x_2,y_2 are described by the R-matrix of the braided plane,

$$R=\begin{pmatrix} q^2 & 0 & 0 & 0 \\ 0 & q & q^2-1 & 0 \\ 0 & 0 & q & 0 \\ 0 & 0 & 0 & q^2 \end{pmatrix},$$

those between x_1, y_1 and v_2, w_2 by its inverse,

$$R^{-1}=\begin{pmatrix} q^{-2} & 0 & 0 & 0 \\ 0 & q^{-1} & q^{-2}-1 & 0 \\ 0 & 0 & q_{-1} & 0 \\ 0 & 0 & 0 & q_{-2} \end{pmatrix},$$

those between v_1, w_2 and x_2, y_2 again by R, and, finally, those between v_2, w_2 again by R^{-1}. The $*$-structure is defined by $x^*=v$ and $y^*=w$ (compare also Ref. 13).

Here $\{x^n y^m | n,m \in \mathbb{N}\}$ and $\{w^m v^n | n,m \in \mathbb{N}\}$ are bases of the two components of the free product defining $V^{\vee}$, and therefore we can take the finite, alternating sequences of elements of these bases as a basis of $V^{\vee}$:

$$\{1=\varnothing, x^{n_1}y^{m_1}, w^{m_1}v^{n_1}, x^{n_1}y^{m_1}w^{m_2}v^{n_2}, w^{m_1}v^{n_1}x^{n_2}y^{m_2}, \ldots | n_1, n_2, \ldots, m_1, m_2, \ldots \in \mathbb{N}\}.$$

The functionals defined on this basis by

$$\phi_{xv}(u)=\begin{cases} 1 & \text{if } u=xv, \\ 0 & \text{else,} \end{cases} \qquad \phi_{vx}(u)=\begin{cases} 1 & \text{if } u=vx, \\ 0 & \text{else,} \end{cases}$$

$$\phi_{yw}(u)=\begin{cases} 1 & \text{if } u=yw, \\ 0 & \text{else,} \end{cases} \qquad \phi_{wy}(u)=\begin{cases} 1 & \text{if } u=wy, \\ 0 & \text{else,} \end{cases}$$

as well as linear combinations thereof with positive coefficients, are generators of diffusions.

IV. APPELL SYSTEMS

Appell systems $\{h_k; k \in \mathbb{N}\}$ can be defined by the conditions
A1: h_k is a polynomial of degree k,
A2: $Dh_k = kh_{k-1}$,

where $D=d/dx$. Examples are furnished by shifted moment sequences of the form

$$h_k(x)=\int_{\mathbb{R}} (x+y)^k p(dx),$$

where p is any probability measure on $\mathbb{R}$ with all moments finite. For example, in the Gaussian case this gives the Hermite polynomials $H_n(x)=(1/\sqrt{2\pi})\int_{\mathbb{R}}(x+iy)^n e^{-y^2/2}\,dy$. Recently analogous classes of ''moment polynomials'' on Lie groups[15] and quantum groups[16] have been introduced. They cease to be polynomials in general.

The definition on quantum groups is easily adapted to braided groups.

Definition 7: We define the Appell polynomials on the braided covector space $V^{\vee}(R,R')$ with respect to the functional ϕ by

$$h_a=(\phi\otimes \mathrm{id})\circ \underline{\Delta}\mathbf{x}_a,$$

i.e., $h_a=T_\phi(\mathbf{x}_a)$, *for* $a \in \mathscr{S}\{1,\ldots,n\}$.

We are particularly interested in the Appell systems with respect to diffusions.

Proposition 8: Let ϕ_t be a diffusion on $V^{\vee}(R,R')$ with generator L and $\{h_a(t); a \in \mathscr{S}\{1,\ldots,n\}\}$ the associated Appell system. Then we have the following:

(1) $h_a(t)=\mathbf{x}_a+p_a(t)$, *where* $p_a(t) \in \otimes_{\alpha=0}^{|a|-1} V^{\vee}(R,R')^{(\alpha)}$,

(2) $D^i h_a = \Sigma_{b \in \mathcal{S}^{|a|-1}\{1,\ldots,n\}} [|a|;R]_a^{i,b} h_b$, *where* $D^i = \exp(tL)\partial^i \exp(-tL)$ *and* $i.b$ *is the concatenation of* i *with* b, *i.e.*, $i.b = ib_1 \cdots b_{|a|-1}$,
(3) *the* $h_a(t)$ *satisfy the evolution equation* $\partial_t u = Lu$.

Proof: Note that (1) and (2) are the analogues of A1 and A2.
(1) The braided-binomial theorem [Ref. 6, Equation (91)] tells us that

$$\underline{\Delta}\mathbf{x}_a = \sum_{i=0}^{|a|} \sum_{b \in S^i\{1,\ldots,n\}, c \in S^{|a|-i}\{1,\ldots,n\}} \begin{bmatrix} |a| \\ i \end{bmatrix}_a^{b.c} \mathbf{x}_b \otimes \mathbf{x}_c .$$

where the braided-binomial coefficients are defined by

$$\begin{bmatrix} m \\ r \end{bmatrix} = (PR)_{r,r+1} \cdots (PR)_{m-1,m} \begin{bmatrix} m-1 \\ r-1 \end{bmatrix} + \begin{bmatrix} m-1 \\ r \end{bmatrix},$$

$$\begin{bmatrix} m \\ 0 \end{bmatrix} = 1,$$

$$\begin{bmatrix} m \\ r \end{bmatrix} = 0 \quad \text{for } r > m.$$

The statement follows now since $\phi(t)|_{V^\vee(R,R')^{(0)}} = \epsilon$, i.e., $h_a(t) = \mathbf{x}_a + p_a(t)$ with

$$p_a(t) = \sum_{i=0}^{|a|-1} \sum_{b \in S^{|a|-i}\{1,\ldots,n\}, c \in S^{|i}\{1,\ldots,n\}} \begin{bmatrix} |a| \\ i \end{bmatrix}_a^{b.c} \phi_t(\mathbf{X}_b)\mathbf{x}_c .$$

Note that this formula can also be used to calculate the Appell systems, if the moments of the process are known.
(2)

$$\begin{aligned} D^i h_a &= \exp(tL)\partial^i \exp(-tL)\exp(tL)x_a \\ &= \exp(tL)\partial^i x_a \\ &= \exp(tL) \sum_{b \in \mathcal{S}^{|a|-1}\{1,\ldots,n\}} [|a|;R]_a^{i,b}\mathbf{x}_b = \sum_{\tilde{b} \in \mathcal{S}^{|a|-1}\{1,\ldots,n\}} [|a|;R]_a^{i,b} h_b \end{aligned}$$

with Ref. 6, Equation (77).
(3)

$$\partial_t h_a(t) = \partial_t \exp(tL)\mathbf{x}_a = L \exp(tL)\mathbf{x}_a = L h_a(t).$$

■

Property 2 shows that the D^i act as lowering operators. We also define raising operators.
Proposition 9: Let $X_i = \exp(tL)x_i \exp(-tL)$. *Then*

$$X_i h_a(t) = h_{i,a}(t).$$

A. Example 1: The braided line $\mathbb{R}_q$

The Appell polynomials associated to the functional $\phi_t = \exp(tL)$ where $L = [a^2/(1+q)]v^2 + bv$ are

$$h_k(x;t) = (\phi_t \otimes \mathrm{id}) \circ \underline{\Delta} x^k = \sum_{\nu=0}^{k} \begin{bmatrix} k \\ \nu ; q \end{bmatrix} \phi_t(x^\nu) x^{k-\nu} = \sum_{\nu=0}^{K} \frac{[k;q]! H_\nu(bt, -(2a^{2t})/(1+q))}{[k-\nu;q]!\,\nu!} x^{k-\nu}$$

These polynomials are solutions of

$$\partial_t u = \frac{a^2}{1+q} \delta^2 u + b\,\delta u$$

with $u(x,t=0)=x^k$ {where δ is the q-difference operator $(\delta f)(x)=[f(qx)-f(x)]/x(q-1)$}.

For $b=0$, $a=\sqrt{(1+q)/2}$, the Appell polynomials simplify to

$$h_k(x;t)=\sum_{\nu=0}^{[k/2]}\frac{[k;q]!t^\nu x^{k-2\nu}}{[k-2\nu;q]!2^\nu\nu!}.$$

These polynomials are a q-analogue of the Hermite polynomials (see also Ref. 16).

B. Example 2: The quantum plane $\mathbb{C}_q^{2|0}$

We choose $\phi_t=\exp tL$ with $L=[1/(1+q^2)](\partial_1^2+\partial_2^2)$. We obtain

$$\phi=\sum_{r=0}^{\infty}\frac{t^r}{(1+q^2)^r r!}\sum_{\nu=0}^{r}\begin{bmatrix}r\\ \nu;q^{-4}\end{bmatrix}\partial_2^{2\nu}\partial_1^{2(r-\nu)}$$

since $\partial_2\partial_1=q^{-1}\partial_1\partial_2$. This leads to the following formula for the Appell polynomials:

$$h_{rm}(x,y;t)=\exp(tL)x^r y^m$$
$$=\sum_{\nu=0}^{[r/2]}\sum_{\mu=0}^{[m/2]}\frac{\begin{bmatrix}\mu+\nu\\ \nu;q^{-4}\end{bmatrix}[m;q^2]![r;q^2]!q^{2\mu(r-2\nu)}t^{\mu+\nu}}{[m-2\mu-1;q^2]![r-2\nu-1;q^2]!(1+q^2)^{\mu+\nu}(\mu+\nu)!}x^{r-2\nu}y^{m-2\mu},$$

where $\begin{bmatrix}r\\ \nu;q^{-4}\end{bmatrix}$ are the braided-binomial coefficients with respect to $R=(q^{-4})\in M(1)\otimes M(1)$. These polynomials solve the evolution equation

$$\partial_t u=\frac{1}{1+q^2}(\partial_1^2+\partial_2^2)u,$$

where ∂_1, ∂_2 can be defined by

$$\partial_1 f(x,y)=\frac{f(q^2x,y)-f(x,y)}{x(q^2-1)},\quad \partial_2 f(x,y)=\frac{f(qx,q^2y)-f(qx,y)}{y(q^2-1)}.$$

C. Example 3: The free braided-space

We consider again the free braided-space associated to $R=R'=P$. It is straightforward to calculate the action of the generator of a diffusion on a basis element:

$$L\mathbf{x}_k=r(r-1)a_{k_1k_2}\mathbf{x}_{k_3\cdots k_r}+rb_{k_1}\mathbf{x}_{k_2\cdots k_r},$$

for $L=\Sigma_{k,l=1}^n a_{kl}\partial_k\partial_l+\Sigma_{j=1}^n b_j\partial_j$ and $k=(k_1,\ldots,k_r)\in\mathcal{S}\{1,...,n\}$. Thus we obtain for the Appell polynomials

$$\begin{aligned}h_k=\exp(tL)\mathbf{x}_k=\mathbf{x}_k+t\big((r-1)a_{k_1k_2}\mathbf{x}_{k_3\cdots k_r}+rb_{k_1}\mathbf{x}_{k_2\cdots k_r}\big)\\ +\frac{t^2}{2}\big(r(r-1)(r-2)(r-3)a_{k_1k_2}a_{k_3k_4}\mathbf{x}_{k_5\cdots k_r}\\ +r(r-1)(r-2)(b_{k_1}a_{k_2k_3}+a_{k_1k_2}b_{k_3})\mathbf{x}_{k_4\cdots k_r}\\ +r(r-1)b_{k_1}b_{k_2}\mathbf{x}_{k_3\cdots k_r}\big)+\cdots,\end{aligned}$$

If the drift term vanishes, i.e., $b_j=0$ for $j=1,...,n$, then

$$h_k=\sum_{\rho=0}^{[r/2]}\frac{t^\rho}{\rho!}r(r-1)\cdots(r-2\rho+1)a_{k_1k_2}\cdots a_{k_{2\rho-1}k_{2\rho}}\mathbf{x}_{k_{2\rho+1}\cdots k_r}.$$

V. DENSITIES

There are two ways in which densities can be associated to the diffusions constructed here.

This first uses the (right or left) invariant integrals that are defined on Hopf algebras. Fix such a functional $\int: V^{\check{}}(R,R') \to \mathbb{C}$. One can look for an element $\rho_t \in V^{\check{}}(R,R')$ such that

$$\phi_t(a) = \int (\rho_t a) \quad \left(\text{or} \quad \phi_t(a) = \int (a \rho_t) \right)$$

is satisfied for all $a \in V^{\check{}}(R,R')$ [where $\phi_t(a) = \langle \phi_t, a \rangle$ denotes the evaluation map between $V^{\check{}}(R,R')$ and $V(R,R')$]. This defines a (noncommutative) left (or right) density. For anyspace, this leads to the results exhibited in Ref. 3. To calculate it the Fourier transform (wrt to the integral considered) can be used. For a braided-Hopf algebra B with dual B^*, exponential exp and invariant integral $\int: B \to \mathbb{C}$, the Fourier transform $\mathscr{F}: B \to B^*$ is defined by $\mathscr{F} = (\int \otimes \mathrm{id}) \circ (m \otimes \mathrm{id}) \circ (\mathrm{id} \otimes \exp)$ (see Ref. 17). Now let $\int$ be a left-invariant integral on $V^{\check{}}(R,R')$ and $\int^*$ a right-invariant integral on $V(R,R')$. Then, since $(\mathrm{id} \otimes \langle \cdot, \cdot \rangle) \circ (\exp \otimes \mathrm{id}) = \mathrm{id}$,

$$\int (\rho a) = \langle \mathscr{F}(\rho), a \rangle$$

for all $a \in V^{\check{}}(R,R')$ and $\rho \in V(R,R')$. Thus we need to invert the Fourier transform $\mathscr{F}$ to find the density corresponding to the functional ϕ_t. By Ref. 17, Proposition 6.4, $\mathscr{F}^* \circ \mathscr{F} = \mathrm{vol}\, \underline{S}$, where vol is a constant and $\mathscr{F}^*$ the Fourier transform wrt $\int^*$, i.e., $\mathscr{F}^* = (\mathrm{id} \otimes \int^* \circ m) \circ (\Psi \otimes \mathrm{id}) \circ (\mathrm{id} \circ \exp)$. Thus

$$\rho_t = \frac{1}{\mathrm{vol}}\, \underline{S}^{-1} \circ \mathscr{F}^*(\phi_t).$$

Take, e.g., the braided-line and $\phi_t = \exp(tv^2/2)$. We shall use the Jackson integral $\int_{-\gamma \cdot \infty}^{+\gamma \cdot \infty} f(x)\, d_q x$ where γ is a parameter. For $f(x) = \Sigma f_n x^n$ we set $f(\gamma) = \Sigma f_n \gamma^n$ and

$$\int_{-\gamma \cdot \infty}^{+\gamma \cdot \infty} f(x) d_q x = (1-q) \sum_{k=-\infty}^{\infty} (f(q^k \gamma) + f(-q^k \gamma)) q^k \gamma$$

if the right-hand side converges. We will express the density in terms of the discrete q-Hermite II polynomials

$$\widetilde{h}_n(x;q) = (1-q)^n [n;q]! \sum_{k=0}^{[n/2]} \frac{(-1)^k q^{-2nk} q^{k(2k+1)} x^{n-2k}}{(1-q^2)^k [k;q^2]! (1-q)^{n-2k} [n-2k;q]!}.$$

With the orthogonality relations for the discrete q-Hermite II polynomials [Ref. 18, Equation (8.14)] and

$$\phi_t(\widetilde{h}_n(x;q)) = \begin{cases} 0 & \text{if } n \text{ is odd}, \\ (1-q)^n [n;q]! \displaystyle\sum_{k=0}^{[n/2]} \frac{(-1)^k q^{-2nk} q^{k(2k+1)} t^{n/2-k}}{(1-q^2)^k [k;q^2]! (1-q)^{n-2k} 2^{n/2-k}} & \text{if } n \text{ is even}, \end{cases}$$

we obtain

$$\rho_t = \frac{1}{c_q(\gamma)} \sum_{m=0}^{\infty} q^{4m^2} \sum_{k=0}^{m} \frac{(-1)^k q^{-4mk} q^{k(2k+1)} t^{m-k}}{(1-q^2)^k [k;q^2]!\, 2^{m-k}}\, \widetilde{h}_{2m}(x;q)\, e_{q^2}(-(1-q^2)x^2), \tag{5}$$

where $e_{q^2}(z) = \Sigma_{n=0}^{\infty} z^n / [n;q^2]!$, and $c_q(\gamma)$ is a constant depending only on q and γ [for its value see Ref. 18, Equation (8.15)].

The other way, commonly used in quantum probability (see, e.g., Ref. 19), leads to a density on the usual real line. Consider, for example, a diffusion $\phi_t = \exp(tv^2/2)$ on the braided line, and the generator x. Then $\phi_t(e^{itx})$ can be interpreted as the Fourier transform of the density of x in the

state ϕ_t. Let $0<q<1$ and $Y=(1-q)\Sigma_{i=0}^{\infty}X_i$, where the X_i are independent, and exponentially distributed with respective means q^i (see Ref. 20). Then $E(e^{iuY})=\Pi_{i=0}^{\infty}(1-iu(1-q)q^i)^{-1}$ $=\Sigma_{n=0}^{\infty}(iu)^n/[n;q]!=e_q^{iu}$. Thus $\phi_t(E(e^{iuYx}))=\phi_t(e_q^i ux)=e^{-tu^2/2}$, and it follows that Yx has a Gauss distribution. This determines the distribution of x uniquely (see Ref. 20, Theorem 3).

VI. CONCLUSION

The notion of q-Brownian motion introduced by Majid is extended in this paper to braided spaces. The resulting processes differ from the ones obtained in Refs. 21 and 22 by more algebraic considerations. To obtain these (for fixed parameter $t=1$, say), we can consider the semidirect cross product of the braided line with its dual. This algebra has basis $\{v^n\otimes x^m;n,m\in\mathbb{N}\}$ and the defining relations are $vx-qxv=1$ (where $v=v\otimes 1$, $x=1\otimes x$). Take the functional ϕ defined by $\phi(\Sigma a_{nm}v^n\otimes x^m)=a_{00}$. Then the distribution of $x+v$ in the state ϕ is equal to that of Ref. 21 for $t=1$ (for a study of its density see Ref. 23). This density is concentrated on the bounded interval $[-1/\sqrt{1-q},2/\sqrt{1-q}]$ (for $-1<q<1$). The q-deformed Brownian motion of Ref. 22 can be obtained similarly. Obviously, there is still some more nontrivial construction required to pass from the random variable presented here to the full process.

[1] E. Nelson, ''Derivation of the Schrödinger equation from Newtonian mechanics,'' Phys. Rev. **150**, 1079–1085 (1966).
[2] S. Majid, ''Quantum random walks and Time-reversal,'' Int. J. Mod. Phys. **8**, 4521–4545 (1993).
[3] S. Majid and M. J. Rodriquez-Plaza, ''Random walk and the heat equation on superspace and anyspace,'' J. Math. Phys. **35**, 3753–3760 (1994).
[4] M. Schürmann, *White Noise on Bialgebras*, Lecutre Notes in Math. 1544 (Springer-Verlag, Berlin, 1993).
[5] S. Majid, ''Beyond supersymmetry and quantum symmetry (an introduction to braided groups and braided matrices),'' in *Quantum Groups Integrable Statistical Models and Knot Theory*, edited by M.-L. Ge and H. J. de Vega (World Scientific, Singapore, 1993).
[6] S. Majid, ''Introduction to braided geometry and q-Minkowski space,'' in Proc. School 'EnricoFermi' CXXVII, (IOS, Amsterdam, 1996), pp. 267–348.
[7] S. Majid, ''Braided momentum in the q-Poincaré group,'' J. Math. Phys. **34**, 2045–2058 (1993).
[8] S. Majid, ''Free braided differential calculus, braided binomial theorem and the braided exponential map,'' J. Math. Phys. **34**, 4843–4856 (1993).
[9] L. Accardi, Y. G. Lu, and I. Volovich, ''Non-commutative (quantum) probability, master fields and stochastic bosonization,'' Preprint Centro V. Volterra 198-94, 1994.
[10] M. Schürmann, ''Direct sums of tensor products and non-commutative independence,'' J. Funct. Anal. **133**, 1–9 (1995).
[11] R. Speicher, Universal products, preprint.
[12] S. Majid, ''Quantum double as quantum mechanics,'' J. Geom. Phys. **13**, 169–202 (1994).
[13] S. Majid, ''*-Structure on braided spaces,'' J. Math. Phys. **36**, 4436–4449 (1995).
[14] S. Majid, ''Quasi-* Structure on q-Poincaré Algebras,'' to appear in J. Geom. Phys. preprint q-alg/9503014.
[15] Ph. Feinsilver and R. Schott, ''Appell systems on Lie groups,'' J. Theor. Prob. **5**(2), 251–281 (1992).
[16] Ph. Feinsilver, U. Franz, and R. Schott, ''Duality and multiplicative processes on quantum groups,'' in *Proceedings of the International Conference on ''Nonlinear, Deformed and Irreversible Quantum Systems,''* edited by H. D. Doebner *et al.* (World Scientific, Singapore, 1995), pp. 29–36.
[17] A. Kempf and S. Majid, ''Algebraic q-Integration and Fourier Theory on Quantum and Braided Spaces,'' J. Math. Phys. **35**, 6802–6837 (1994).
[18] T. H. Koornwinder, Minicourse ''Special functions and q-commuting variables'' at Workshop on ''Special functions, q-series and related topics,'' Fields Institute, Toronto, Canada, 12–23 June 1995.
[19] P.-A. Meyer, *Quantum probability for probabilities*, Lecture Notes in Math. 1538 (Springer-Verlag, Berlin, 1993).
[20] Ph. Feinsilver, ''Generalized convolutions and probability theory based on the q-Heisenberg algebra,'' Proc. First World Congr. Bernoulli Soc. (VNU Science, Utrecht, 1987), pp. 729–741.
[21] M. Bozejko and R. Speicher, ''An example of generalized Brownian motion,'' Commun. Math. Phys. **137**, 519–531 (1991).
[22] V. I. Manko and R. Vilela Mendes, ''q-deformed Brownian motion,'' Phys. Lett. A **180**, 39–42 (1993).
[23] H. van Leeuwen and H. Maassen, ''A q-deformation of the Gauss distribution,'' J. Math. Phys. **35**, 4743–4756 (1995).

Brownian motion of quantum harmonic oscillators: Existence of a subdynamics

Göran Lindblad[a)]
Theoretical Physics, Royal Institute of Technology, SE-100 44 Stockholm, Sweden

(Received 13 May 1997; accepted for publication 22 January 1998)

The effects of system-environment correlations on the dynamics of an open quantum system are investigated for the standard model of a set of quantum harmonic oscillators interacting with a heat bath of oscillators. By definition, a subdynamics is described by transformations of the open system observables. It is shown that such a construction can reproduce the observable properties of the exact dynamics only when the states of system and environment are uncorrelated, while for classical systems there is always a subdynamics. A quantum subdynamics cannot have the properties we associate with thermal fluctuations; the KMS relation at a finite temperature for the open system implies that the system must be closed. The conditions for having a subdynamics as a good approximation to the exact closed dynamics are investigated, and so are the similar but stronger conditions for a Markovian dynamics. It is also shown that a subdynamics defines the response of the open system to some types of time dependent external forces. © *1998 American Institute of Physics.* [S0022-2488(98)00705-1]

I. INTRODUCTION

In the description of irreversible quantum processes a fundamental concept is that of an *open quantum system*. We can directly observe a quantum system which evolves in contact with reservoirs and responds to driving forces. The observed system is small (sometimes of atomic or molecular size) or at least finite, and the reservoirs are often assumed infinite. The mathematical problem encountered is how the evolution of the observed subsystem can be given a reduced description, here called a *subdynamics*, where the degrees of freedom of the environment have been traced out. It is assumed that the closed composite system including the environment has a reversible, conservative dynamics, while the evolution of the small system will generally have irreversible and random properties.

Many different answers have been given to the question of what mathematical structure a subdynamics should have; here we will use a general one allowing for memory effects.[1–3] It means in essence that all probability distributions accessible by observations on the small system can be expressed as expectations on the initial partial state of this subsystem. The mathematical structure which results when a subdynamics exists was called a *quantum stochastic process* in Refs. 1 and 2, but there no emphasis was given to the environment and the system-environment correlations. In Ref. 3 the focus was precisely on the latter point, and it was shown that such a subdynamics will exist only when the state of the observed system and its environment is uncorrelated (it factorizes into a product), while it always exists for open systems based on commutative algebras of observables. This result is closely related to general theorems restricting the existence of conditional expectations in non-commutative operator algebras.[4–6] An uncorrelated initial state is also an almost universal ingredient in the derivation of master equations, but here we must demand this to hold for all times.

The goal of this paper is to display the conditions for the existence of a subdynamics and its properties for a class of models where they have a particularly simple form. The CCR algebra and the states and maps on this algebra which preserve the quadratic structure associated with multidimensional harmonic oscillators form a simple laboratory for studying the dynamics of open systems. This class of models is often given the name *quantum Brownian motion*. A very large

a)Electronic mail: gli@theophys.kth.se

 © 1998 American Institute of Physics

number of papers have been written on this subject, using many different methods, but it is not intended to give an overview here (reviews can be found in Refs. 7–10). A rather abstract, algebraic approach will be followed here, similar to the methods used in Refs. 1 and 11. However, we have dispensed with mathematical refinements which are not essential in this context. The properties we are interested in turn out to have quite simple and explicit representations as matrix relations.

The layout of the paper is as follows. We first review some well known but essential details of the formalism, introducing the Weyl form of the CCR and quasifree states (Sec. II), and quasifree maps and dynamics (Sec. III). The observable multi-time correlation functions for an open subsystem are defined and the concept of subdynamics formulated (Sec. IV). For these models the full set of correlation functions is always completely specified by a matrix-valued covariance function, in complete analogy with stationary Gaussian stochastic processes.

There is a subdynamics when this covariance function results from a dynamical transformation of the canonical operators for the subsystem. It is shown that the condition for having a subdynamics as an exact property is a factorization of the state of open system and reservoir. A classical (commutative) limit can be introduced and it is straightforward to see why there are no restrictions on the existence of a subdynamics in this case (Sec. V). It is proved that the KMS property characteristic of the quantum correlation functions in thermal equilibrium at a finite temperature, when imposed on the covariance function, is incompatible with having a non-trivial subdynamics (Sec. VI). In Sec. VII we consider the case where the initial state is not stationary. If the state of the system factorizes for all times there is still a non-stationary subdynamics which suffices to recreate the covariance function, but if it factorizes only for the initial time this no longer holds true. The structure of Markov processes and the conditions for the Markov property to hold are given in Sec. VIII.

The observable multi-time correlations not only give the statistics of observations of the system, they also define the response of the system to certain kinds of time dependent external forces acting on the small subsystem. The forces are of a harmonic type which preserves the quasifree property. In some cases the response of the system can be given an explicit form in terms of the maps which define the subdynamics (Sec. IX).

In Sec. X we find conditions which will allow a subdynamics as a good approximation to the exact evolution also for correlated system-reservoir states. These include that the back reaction of the system on the reservoir spreads effectively over a large number of degrees of freedom. A subdynamics as defined here contains a large amount of information, resulting in the strong restrictions already mentioned. If we restrict ourselves to describing the evolution of a non-stationary initial state of the open system, there are solutions also in cases where a subdynamics cannot be defined (Sec. XI). In general the dynamical maps will then fail to be positive due to the system-reservoir correlations. There is a considerable literature on the consequences of such quantum correlations. In the discussion (Sec. XII) a number of references are given to earlier work on this and similar problems.

II. QUASIFREE STATES ON THE CCR ALGEBRA

Let $\mathscr{H}$ be a separable real Hilbert space. For most of this paper we will do the algebraic manipulations as if it has a finite dimension $2N$, and thus avoid a discussion of a number of mathematical subtleties.

It is convenient to consider also operators in the complexified Hilbert space. The star (*) is used for the complex conjugation in the given real basis and real operators satisfy $A^*=A$. Let $\sigma(x,y)$ be a bounded real symplectic (antisymmetric, bilinear) form on this space. We can write it as a matrix element of a real antisymmetric operator S: $\sigma(x,y)=(x|S|y)$. The CCR algebra $\mathscr{A}(\mathscr{H})=\mathscr{A}$ is the closure in a suitable C^* norm of the $*$-algebra generated by unitary elements $\{W(x)=W(-x)^\dagger; x\in\mathscr{H}\}$ which satisfy the Weyl relation[12]

$$W(x)W(y)=\exp[-i\sigma(x,y)]W(x+y). \tag{2.1}$$

In fact, $\mathscr{A}$ is generated as a complex linear space by the operators $W(x)$. If the projector P

$\in B(\mathscr{H})$ is real, then so is $P^{\perp}:=\mathbb{1}-P$, furthermore the real Hilbert subspaces $P.\mathscr{H}$ and $P^{\perp}.\mathscr{H}$ generate CCR subalgebras. If in addition $[P,S]=0$, the symplectic form decomposes, the subalgebras commute, and there is a tensor product decomposition

$$\mathscr{A}(\mathscr{H})=\mathscr{A}(P.\mathscr{H})\otimes\mathscr{A}(P^{\perp}.\mathscr{H}).$$

We call the state ρ on $\mathscr{A}$ *quasifree* or *Gaussian* if it is defined by a bounded quadratic form g with $g(0)=0$ through

$$\rho[W(x)]=\exp[g(x)]. \tag{2.2}$$

There is no essential restriction in assuming g to be homogenous (without linear terms) and hence real. It is then defined by a symmetric, real operator G

$$g(x)=-\tfrac{1}{2}(x|G|x).$$

The positivity of the linear functional ρ is equivalent to the *positive definiteness* of the form

$$R(x,y)=\exp[g(y-x)+i\sigma(x,y)],$$

which means that for all sequences $\{x_j\in\mathscr{H},\lambda_j\in\mathbb{C}\}$ it holds that

$$\sum_{j,k}\lambda_j^{*}\lambda_k R(x_j,x_k)\geqslant 0. \tag{2.3}$$

A general argument[13] shows that this again is equivalent to the *conditional positive definiteness* of the form

$$L(x,y):=g(y-x)+i\sigma(x,y),$$

which means that for all sequences $\{x_j\in\mathscr{H},\lambda_j\in\mathbb{C};\Sigma_j\lambda_j=0\}$ it holds that

$$\sum_{j,k}\lambda_j^{*}\lambda_k L(x_j,x_k)\geqslant 0.$$

But this is again equivalent to the positive definiteness of the Hermitian form

$$(x|K|y):=L(x,y)+L(0,0)-L(x,0)-L(0,y)=g(y-x)-g(x)-g(y)+i\sigma(x,y).$$

Because $g(y-x)-g(x)-g(y)=(x|G|y)$ we can write this as

$$K=G+iS\geqslant 0. \tag{2.4}$$

The $2N$ canonical self-adjoint (unbounded) operators X_k are formally related to the Weyl ones through $W(x)=\exp(i\Sigma_k x_k X_k)$. The Weyl relation (2.1) corresponds to the CCR

$$[X_k,X_l]=2i\mathbb{1}\cdot S_{kl}.$$

The covariance matrix for the canonical operators is K: in the general case

$$\rho(X_kX_l)=K_{kl}+\rho(X_k)\rho(X_l),$$

but here we have assumed that the states satisfy $\rho(X_k)=0$ for all k. The higher order correlations are then given by the second order one in the well known way; for example,

$$\rho(X_kX_lX_mX_n)=K_{kl}K_{mn}+K_{km}K_{ln}+K_{kn}K_{lm}. \tag{2.5}$$

III. QUASIFREE MAPS ON THE CCR ALGEBRA

A set of *completely positive* (CP) maps is now introduced. They correspond to convolutions with Gaussian measures in the commutative case, and we again call them quasifree or Gaussian.[2,14–16] For general information on CP maps see also Ref. 17. The most general quasifree CP map is defined through

$$T[W(x)]=W(Ax)\exp[f(x)] \tag{3.1}$$

by a real operator $A\in B(\mathscr{H})$ and a quadratic form f such that $f(y-x)+i\sigma(x,y)-i\sigma(Ax,Ay)$ is conditionally positive definite. If $T[\mathbb{1}]=\mathbb{1}$, then $f(0)=0$ (this corresponds to conservation of probability). Again we can restrict our attention to homogenous, real f and write

$$f(y-x)-f(x)-f(y)=(x|F|y),\quad f(x)=-\tfrac{1}{2}(x|F|x),$$

where the symmetric, real operator F must satisfy

$$D:=F+iS-iA^{\dagger}SA\geqslant 0. \tag{3.2}$$

Thus the map (3.1) is uniquely defined by the pair $\{A,F\}$. It leaves the state ρ invariant if for all $y\in\mathscr{H}$

$$\rho\{T[W(y)]\}=\rho[W(y)], \tag{3.3}$$

which holds precisely when $F=G-A^{\dagger}GA$. We can rewrite this condition and (3.2) as

$$D=K-A^{\dagger}KA\geqslant 0. \tag{3.4}$$

If the state is not stationary, then the change in the state is given by $G\mapsto A^{\dagger}GA+F$, or, equivalently, by $K\mapsto A^{\dagger}KA+D$.

The transformation properties of the first and second powers of the canonical operators corresponding to (3.1) are

$$\begin{aligned} T[X_k]&=(XA)_k:=\sum_j X_jA_{jk},\\ T[X_kX_l]&=(XA)_k(XA)_l+\mathbb{1}\cdot D_{kl}. \end{aligned} \tag{3.5}$$

The transformation of the higher powers are defined by similar formulas involving the pairs of matrices $\{A,F\}$ or $\{A,D\}$. This is an obvious analog to well known properties of the commutative multidimensional Gaussian transformations. The lowest order correlation functions associated with the map is, for the Weyl operators

$$\rho\{W(-x)T[W(y)]\}=\exp[i\sigma(x,Ay)+f(x)+g(x-Ay)],$$

and for the canonical ones we obtain the covariance matrix

$$\rho[X_kT(X_l)]=(KA)_{kl}. \tag{3.6}$$

A particularly important class of maps are of the form (3.1) with $A=P$ a real projector satisfying $[P,S]=0$. If $PFP=0$, then (3.1) is the identity map on $\mathscr{A}(P.\mathscr{H})$ and maps $\mathscr{A}(\mathscr{H})$ onto this subalgebra. From the CP condition (3.2) we then find that $PFP^{\perp}=0$, and consequently find the characteristic property of a *conditional expectation*:[3–5]

$$E[W(x)W(y)W(z)]=W(x)E[W(y)]W(z),\quad \forall x,z\in P.\mathscr{H},y\in\mathscr{H}.$$

The invariance (3.3) of ρ under E means that $F=G-PGP$, and $PFP^{\perp}=0$ is then equivalent to $[P,G]=0$. This relation means that the invariant state for the composite system is a product of the partial states on the two algebras $\mathscr{A}(P.\mathscr{H}),.\mathscr{A}(P^{\perp}.\mathscr{H})$. Recalling that $[P,S]=0$ we find that there is a conditional expectation onto $\mathscr{A}(P.\mathscr{H})$ leaving ρ invariant only if

$$[P,K]=0, \tag{3.7}$$

and, conversely, if this equation holds, then

$$E[W(x)]=W(Px)\exp[g(P^{\perp}x)] \tag{3.8}$$

defines the desired conditional expectation (note that it does not depend on PGP).

The set of quasifree CP maps also includes a subset representing reversible (conservative) dynamics. In this case $f=0$ in (3.1) and A is now an invertible map V which is symplectic in the sense that the form S is left invariant

$$V^{\dagger}SV=S \tag{3.9}$$

(i.e., the CCR are invariant). Furthermore, the initial state defined by g is stationary if $V^{\dagger}GV=G$. The stationarity condition and (3.9) can be summed up in

$$V^{\dagger}KV=K. \tag{3.10}$$

We will use either a discrete or a continuous time parameter, with a group $V(t)$ of dynamical maps in the reversible case: $V(s)V(t)=V(s+t)$. For a continuous time parameter t we have to demand some continuity of $V(t)$.

In Sec. XI it will be necessary to consider maps which can fail to be CP or even positive. This means that (3.2) and (3.4) are no longer satisfied, but we can retain the rest.

IV. CORRELATION FUNCTIONS AND SUBDYNAMICS

Consider the algebra $\mathcal{A}(\mathcal{H})$ defined above to represent a closed system (observed system $\mathcal{S}$ and reservoir $\mathcal{R}$) with a given reversible quasifree dynamics defined by $V(t)$. The open subsystem $\mathcal{S}$ is defined by a finite-dimensional (say of dimension $2M \ll 2N$) real projection P commuting with S. The corresponding algebra of observables is $\mathcal{A}_{\mathcal{S}}=\mathcal{A}(P\mathcal{H})$. Assume that the operators S,G which define the CCR and the reference state ρ through (2.2) on $\mathcal{S}+\mathcal{R}$ are given.

We now want to define a notion of subdynamics for $\mathcal{S}$. In the general scheme outlined in Ref. 3 the construction of a subdynamics is based on finding a conditional expectation from the algebra of $\mathcal{S}+\mathcal{R}$ onto that of the subsystem $\mathcal{S}$, such that the state ρ is left invariant. Here we can start from the more explicit form for the correlation functions available in the CCR case and formulate the condition that the correlation functions can be written in terms of operators in the subspace $P\mathcal{H}$ only.

If we are restricted to observing $\mathcal{S}$, then the physics is given by the correlation functions of the following general structure:

$$R_n(\mathbf{x}|\mathbf{y})=\rho[W(-V_1x_1)W(-V_2x_2)\cdots W(-V_nx_n)W(V_ny_n)\cdots W(V_1y_1)], \tag{4.1}$$

where ρ is the state of $\mathcal{S}+\mathcal{R}$ at time t_0, $x_k,y_k \in P\mathcal{H}, V_k=V(t_k-t_0), t_0 \leqslant t_1 \leqslant \ldots \leqslant t_n$.[2,18] If ρ is stationary, the expression is invariant under time translation $t_k \mapsto t_k+s$. The correlation function is Gaussian, i.e., an exponent of a quadratic form in the variables (x_k,y_k):

$$R_n(\mathbf{x}|\mathbf{y})=\exp[g(\eta_n-\xi_n)+ih(\xi_n,\eta_n)],$$

$$h(\xi_n,\eta_n)=\sigma(\xi_n,\eta_n)+\sum_{k=2}^{n}[\sigma(\xi_k,\xi_{k-1})-\sigma(\eta_k,\eta_{k-1})], \tag{4.2}$$

$$\xi_m=\sum_{k=1}^{m}V_kx_k, \quad \eta_m=\sum_{k=1}^{m}V_ky_k.$$

An examination shows that the expression (4.2) is completely defined by the matrix-valued covariance kernel

$$K_P(s,t)=PV(s)^{\dagger}KV(t)P=K_P(t,s)^{\dagger} \tag{4.3}$$

(we will leave out the index P when this can be done without confusion). Recall that $V(t)$ is a real operator. This means that (4.3) determines the real and imaginary parts obtained by replacing K by G and S, respectively. The most important property of the correlation functions R_n is the condition (2.3), where we now let $x_j \in \oplus^n P\mathscr{H}$. By the line of argument leading to (2.4) this condition is equivalent to the positive definiteness of the covariance kernel: for all $\lambda_j \in \mathbb{C}, t_j \in \mathbb{R}, x_j \in P\mathscr{H}$,

$$\sum_{j,k} \lambda_j^* \lambda_k (x_j | K(t_j, t_k) | x_k) \geqslant 0. \tag{4.4}$$

If G defines a stationary state, which means that (3.10) is fulfilled by $V(t)$ for all t, then the covariance kernel simplifies to a covariance function of a single variable: for $s \leqslant t$

$$K(s,t) = K(t-s) = PKV(t-s)P. \tag{4.5}$$

We will mainly consider this stationary case.

We can now introduce the concept of a *stationary subdynamics* on the subsystem $\mathscr{S}$ as follows. Given is an operator $K_P = G_P + iS_P$ which defines the CCR (through S_P) and the stationary state (through G_P) on the Hilbert space $P\mathscr{H}$. We also need a real operator-valued function $A(t)$ in $P\mathscr{H}$ which satisfies $A(0) = P$. By the assumption of stationarity we must have (3.3), so if we put

$$F(t) = G_P - A(t)^\dagger G_P A(t), \tag{4.6}$$

then $\{A(t), F(t)\}$ define the evolution of the Weyl operators for $x \in P\mathscr{H}$ through (3.1) and the evolution of powers of the canonical operators X_k belonging to the same subspace through (3.5). From (3.6) the covariance function can be identified as

$$K_P(t) = K_P(-t)^\dagger = K_P A(t), \quad t \geqslant 0, \quad = A(-t)^\dagger K_P, \quad t \leqslant 0. \tag{4.7}$$

Provided that (4.4) is fulfilled we obtain a full set of positive definite correlation functions by replacing (4.3) by (4.7) in (4.2).

We now want to impose that the reduced dynamics is an exact consequence of the full dynamics of the system $\mathscr{S} + \mathscr{R}$. First the CCR and the state ρ on $\mathscr{S}$ are those obtained from $\mathscr{S} + \mathscr{R}$ by restriction to $\mathscr{S}$, hence $K_P = PKP$. Then we demand the identity of the correlation functions (4.1), which just means the equality of the covariance matrix functions (4.5) and (4.7) for arguments in $P\mathscr{H}$:

$$PK[A(t) - V(t)]P = 0, \quad \forall t \geqslant 0$$

(for $t \leqslant 0$ we get the conjugate relation). The imaginary part of this relation, using $[P,S] = 0$, says that

$$S[A(t) - PV(t)P] = 0. \tag{4.8}$$

If we take S to be nondegenerate (in the standard form of the CCR we have $S^\dagger S \propto \mathbb{1}$), we obtain, as the only solution

$$A(t) = PV(t)P, \quad \forall t \geqslant 0. \tag{4.9}$$

The relation (4.8) is then equivalent to

$$PGP^\perp V(t)P = 0, \quad \forall t \geqslant 0. \tag{4.10}$$

It is clear that if $[P,G] = 0$, then (4.10) is satisfied. Are there other solutions? From the form (4.1) of the correlation functions it is rather obvious that we can restrict our attention to the space $\mathscr{H}_+ \subseteq \mathscr{H}$ which is defined as the smallest Hilbert subspace satisfying the conditions

$$P\mathscr{H} \subseteq \mathscr{H}_+, \quad V(t)\mathscr{H}_+ \subseteq \mathscr{H}_+, \quad \forall t \geqslant 0.$$

It is the minimal subspace of $\mathscr{H}$ which contains $P\mathscr{H}$ and is invariant under forward time translations, and thus represents the relevant part of the total closed system. There is no real restriction in taking $\mathscr{H}_+$ to be the whole of $\mathscr{H}$, which means that in a well-defined sense the system is chosen to be minimal or irreducible. We note that if there is no shift in the dynamics $V(t)$, then $\mathscr{H}_+$ is invariant also under backward time translations, while if $V(t)$ contains a shift this need not be the case (and $\dim\mathscr{H}=\infty$, of course).[18,19] For Markov processes there must be a shift.[18]

With the minimal choice of Hilbert space it follows from (4.10) that there is an exact subdynamics only if it holds that

$$PKP^{\perp}=0, \tag{4.11}$$

which is equivalent to (3.7). But (3.7) says that there is a conditional expectation E defined by (3.8). The correlation functions for the subdynamics are then completely defined by the partial state $\rho_{\mathscr{S}}$ ($=\rho$ restricted to $\mathscr{A}(P\mathscr{H})$):

$$R_{\mathscr{S},n}(\mathbf{x}|\mathbf{y})=\rho_{\mathscr{S}}\{E[W(-V_1x_1)W(-V_2x_2)\cdots W(V_2y_2)W(V_1y_1)]\}. \tag{4.12}$$

An expansion of the exponent in the resulting expression using (3.7) shows that it is indeed identical to (4.2). In particular, the covariance function can be expressed as an expectation in the state of $\mathscr{S}$:

$$K_P(t)_{jk}=\rho_{\mathscr{S}}(X_jE[XV(t)]_k).$$

This property provides an alternative, and equivalent, definition of a subdynamics: all correlation functions are expectations in the partial initial state of $\mathscr{S}$. The matrix-valued function on $P\mathscr{H}$ defined in (3.4) is positive,

$$D(t)=PV(t)^{\dagger}P^{\perp}KP^{\perp}V(t)P=PKP-A(t)^{\dagger}KA(t)\geqslant 0,$$

which means that the pair of functions $\{A(t),F(t)=\Re D(t)\}$ define a family of CP maps $\{T(t);t\geqslant 0\}$ through (3.1). We find that

$$K_P(t)_{jk}=\rho_{\mathscr{S}}(X_jT(t)[X_k]),$$

and the stationarity of the dynamics means that $\rho_{\mathscr{S}}(T(t)[X])=\rho_{\mathscr{S}}(X)$. Note that if we want to construct $K_P(t)$ starting from $\rho_{\mathscr{S}}$ and $\{T(t)\}$, then the CP property for each t is not sufficient for (4.4) to hold, except when $\{T(t)\}$ forms a semigroup (Sec. VIII).

To sum up: the subdynamics is defined by $\{K,A(t)\}$, and from this we get a positive $D(t)$, the real part of which is $F(t)$. The pair $\{A(t),F(t)\}$ defines $\{T(t)\}$ while $\{A(t),D(t)\}$ gives a simple expression for the subdynamics as transformations of the canonical operators through (3.5).

We have already noted that $K(t)$ defines the full set of correlation functions (4.1). It is also true that this set allows us to reconstruct the state and dynamics of the system $\mathscr{S}+\mathscr{R}$, though not always uniquely[18]. Note that there is a good deal of formal similarity between the formalism used here and standard projection methods,[20] but here the projections all act in a Hilbert space.

V. THE CLASSICAL LIMIT

It is a rather elementary fact that a subdynamics always exists for dynamical systems based on commutative algebras of observables.[3] The necessary conditional expectations are directly related to the marginal distribution for the random variables describing the open system $\mathscr{S}$. For the models based on the CCR we can consider a *classical limit* simply and formally by setting $S=0$. How does the condition for having a subdynamics change?

Without a non-degenerate S there is no longer the earlier conclusion to be drawn from (4.8). For $S=0$ we obtain only

$$PG[V(t)-A(t)]P=0.$$

There is a nontrivial solution, assuming PGP to be non-degenerate for simplicity. We introduce an idempotent operator Q which satisfies $QP=P$, $PQ=Q$, but which is not self-adjoint:

$$Q=(PGP)^{-1}PG.$$

We choose

$$A(t)=QV(t)P.$$

If we replace P in (3.8) by Q and $P^{\perp}GP^{\perp}$ by $F=G-Q^{\dagger}GQ$, then (3.8) still defines a conditional expectation associated with the marginal distribution for the variables in $P\mathscr{H}$. The only thing we have to check is that $PFP=0$ and that

$$F=G-GP(PGP)^{-1}PG\geqslant 0,$$

which is a known Schwarz type inequality.[21] We note that the same method works also in the non-stationary case (see Sec. VII below).

VI. KMS CONDITION AND SUBDYNAMICS

It is well known that the KMS condition (describing the regression properties of fluctuations in thermal equilibrium) for a finite temperature is incompatible with the existence of a Markovian, exponential relaxation of the open subsystem.[22–27] Here it will be shown that in the present setup it does not help to use non-Markovian subdynamics: it is not possible to choose a dissipative quasifree subdynamics in such a way that the covariance function for the open system satisfies the KMS condition for a finite temperature. If the KMS condition holds, the dynamics is either of the conservative type of closed systems or the temperature is infinite. This is a CCR version of a more general theorem proved in Ref. 3. It is included for completeness and because of its simplicity. In this section the time parameter is continuous.

Let ρ be a quasifree state (for $\mathscr{S}$) which is stationary under the family of quasifree maps $T(t)$ defined by $A(t)$ and (4.6). We have to assume the continuity of $A(t)$, such that $\lim_{t\to 0}A(t)=\mathbb{1}$. Consider the following correlation function for $t>0$

$$R(x,y;t)=\rho\{W(-x)T(t)[W(y)]\}=\rho[W(-x)W(A(t)y)]\exp[-\tfrac{1}{2}(y|F(t)|y)]$$

$$=\exp[-\tfrac{1}{2}(x|G|x)-\tfrac{1}{2}(y|G|y)+(x|K(t)|y)],$$

where $x,y\in P\mathscr{H}$ and we have dropped the P in (4.7). It is continued to negative values of t through $R(x,y;-t)=R(y,x;t)^*$, which is equivalent to $K(t)^{\dagger}=K(-t)$, cf. (4.3). The KMS condition with parameter $\beta=\hbar/k\Theta$ reads:[28] the correlation function has an analytic continuation in the time parameter into the strip $0\leqslant\Im z\leqslant\beta$ such that for all real t the identity $R(x,y;t)=R(y,x;-t+i\beta)$ holds. Translated to the covariance function it reads

$$K(t)^*=K(t+i\beta). \tag{6.1}$$

There is also the condition that $A(t)$ must be real: for $t\geqslant 0$

$$K(t)^*=K^*K^{-1}K(t). \tag{6.2}$$

The condition (4.4) applied to $K(t)$ and Bochner's theorem implies that there is a positive-operator-valued measure $E(\cdot)$ such that

$$K(t)=\sqrt{K}\left[\int_{-\infty}^{\infty}dE(\omega)\exp(-i\omega t)\right]\sqrt{K},$$

$$\int_{-\infty}^{\infty}dE(\omega)=\mathbb{1}.$$

If (6.1) and (6.2) are translated to this measure, we find that

$$dE(-\omega)^*=\exp(\beta\omega)M^{\dagger}dE(\omega)M=M^{-1}dE(\omega)M,$$

where we have used the notation $M=K^{1/2}(K^*)^{-1/2}$ and assumed K to be nondegenerate. Simplification gives

$$MM^\dagger dE(\omega)=\exp(-\beta\omega)dE(\omega).$$

But $MM^\dagger$ is a self-adjoint operator, and its eigenspaces must be orthogonal. Using the normalization of $E(\cdot)$ we can then show that it must actually be a projection-valued measure, and

$$W(t)=\int_{-\infty}^{\infty} dE(\omega)\exp(-i\omega t)=\exp(-iHt)=K^{1/2}A(t)K^{-1/2}$$

is a group of unitary operators. Thus $\{A(t)\}$ is also a group of invertible maps satisfying (3.10), and they give a closed, conservative, evolution of $\mathscr{S}$. Note that in the classical $\hbar=0$ or ∞-temperature limit it holds that $\beta=0$, which means that $K(t)=K(t)^*$ and (6.1) and (6.2) coincide.

We note that the KMS condition is intimately related to the passivity of the thermal reservoir and hence to the validity of the second law of thermodynamics.[28] Thus losing this property is serious even though the effect may be small.

VII. THE NON-STATIONARY CASE

We now go back to (4.3) where we no longer assume stationarity; instead the state is time dependent: $G(t):=V(t)^\dagger GV(t)\neq G(0)$. In order to have a subdynamics defined for all t we need that

$$[P,G(t)]=0, \quad \forall t\geqslant 0. \tag{7.1}$$

There is then a time dependent conditional expectation which is defined by (3.8) if we replace $P^\perp GP^\perp$ by $P^\perp G(t)P^\perp$. In order to reconstruct all the correlation functions we need the function $G_P(t)=PG(t)P$ in addition to $A(t)$, and instead of (4.5) and (4.7) there is, for $s<t$,

$$K_P(s,t)=PV(s)^\dagger KV(s)PV(t-s)P=[G(s)_P+iS_P]A(t-s)=K_P(s)A(t-s).$$

There is a two-parameter family of quasifree maps $T(s,t)$ defined by $A(t-s)$ and

$$F(s,t)=G_P(t)-A(t-s)^\dagger G_P(s)A(t-s) \quad (s<t),$$

and they are all CP as we find that

$$D(s,t):=K_P(t)-A(t-s)^\dagger K_P(s)A(t-s)=PV(t-s)^\dagger P^\perp K(s)P^\perp V(t-s)P\geqslant 0.$$

The relation (3.3) is still satisfied if ρ in the RHS is replaced by the time dependent state $\rho(t)$ defined by $G(t)$ for the pertinent value of t. The relation (3.5) still holds if we replace $\{T,D\}$ by the corresponding time dependent quantities defined above.

Now consider the case where the initial state of $\mathscr{S}+\mathscr{R}$ is a product state: (7.1) holds for $t=0$ but not for $t>0$. Now we must allow the initial instant to be singled out, and there will in general only be maps from $t=0$ to arbitrary $t>0$. However, we will see that the covariance function is no longer defined by a family of dynamical maps (CP or otherwise).

At the initial time there is still a conditional expectation of the type (3.8). If $V(t)$ again represents the reversible dynamics of the closed system, we can calculate the evolution of canonical operators in $\mathscr{S}$:

$$E[X\cdot V(t)]=X\cdot A(t), \quad A(t)=PV(t)P.$$

The covariance function for $\mathscr{S}+\mathscr{R}$ is

$$K(s,t)=V(s)^\dagger KV(t)=K(t,s)^\dagger,$$

given that $V(t)$ is real. The conservation of the CCR means that the imaginary part is (for $s<t$)

$$S(s,t)=V(s)^\dagger SV(t)=-S(t,s)^\dagger=SV(t-s).$$

From this follows that $S(t,t)=S$ and that the projected part is

$$PS(s,t)P=S_P(s,t)=S_PA(t-s)\quad(s\leqslant t),\quad =A(s-t)^\dagger S_P\quad(t\leqslant s),$$

where $A(t)=PV(t)P$ as before. We reduce the covariance to the subspace by projection by P, as before, and obtain

$$\begin{aligned}PK(s,t)P=K_P(s,t)=A(s)^\dagger K_PA(t)+D(s,t),\\ D(s,t)=PV(s)^\dagger P^\perp KP^\perp V(t)P,\end{aligned}\tag{7.2}$$

where $\{K_P,D\}$ are positive definite kernels. In particular, for the imaginary part we find for $s\leqslant t$

$$PSA(t-s)=A(s)^\dagger SA(t)+\Im[D(s,t)].$$

If we assume $A(t)$ given, the CCR imply the following condition on the kernel D:

$$\Im[D(s,t)]=PSA(t-s)-A(s)^\dagger SA(t),\quad s\leqslant t.\tag{7.3}$$

From the continuity of the dynamics we get the condition

$$D(0,t)=0,\quad \forall t\geqslant 0.\tag{7.4}$$

Given a real function $A(t),t\geqslant 0$ with $A(0)=\mathbb{1}$ and a positive definite kernel $D(s,t)$ which satisfy the conditions (7.3) and (7.4) above, then (7.2) defines a non-stationary covariance kernel $K_P(s,t)$ for $s,t>0$ with a preferred initial instant $t=0$ where the state of the system factorizes. Furthermore, $\{A(t),D(t,t)\}$ defines a one-parameter family $T(t)$ of CP maps through (3.1), but this information is no longer sufficient to recreate the two-parameter function $K_P(s,t)$.

VIII. MARKOV PROCESSES

For quantum dynamics it is necessary to define what we mean by a Markov process. This subject has been treated extensively and we will use a standard definition.[1,2,23] It starts from a quantum dynamical semigroup of CP maps.[11,29–31] In general the semigroup property alone is not sufficient. The multitime correlation functions are not necessarily of the form (8.5) below. For the quasifree case we are treating here, however, we have seen already that the covariance function defines the whole process, including the higher order correlation functions. It will be evident from the following that the semigroup defines the covariance function.

Assume that on the Hilbert space $P\mathscr{H}$ we are given an operator $K_P\geqslant 0$ which defines G_P,S_P and a family of operators $A(t)$ which form a semigroup

$$A(s+t)=A(s)A(t),\quad s,t=0,1,2,\ldots,\tag{8.1}$$

where we use a discrete time parameter for the moment. The pair $\{A(t),F(t)\}$ and (4.6) define a semigroup of maps $T(s)T(t)=T(s+t)$ provided that

$$F(s+t)=F(t)+A(t)^\dagger F(s)A(t),\quad s,t>0.$$

This relation is fulfilled with the choice (3.1), and the maps are all CP if (3.4) holds:

$$K_P-A(t)^\dagger K_PA(t)\geqslant 0,\quad \forall t>0.\tag{8.2}$$

If this relation holds for $t=1$, it holds for all $t=n>0$. Furthermore the state defined by G_P is invariant.

If we are given the dynamics $V(t)$ for the system $\mathscr{S}+\mathscr{R}$, an operator $K\geqslant 0$ satisfying (4.11), and the expectation E given by (3.8), then we can introduce the characteristic Markov property for the subdynamics on $\mathscr{S}$ in the following way:[3]

$$T(t)\{E[W(x)]\}=E\{W[V(t)x]\},\quad \forall x\in\mathscr{H}. \tag{8.3}$$

If this holds for $t=1$, then it holds for $t=2,3,\ldots$, and the family $T(t)$ of maps acting on $\mathscr{A}(P\mathscr{H})$ then form a semigroup. Hence (8.1) holds, and from (8.3) it also follows that $A(t)P=PV(t)$, $\forall t\geqslant 0$. This relation implies not only (4.9) but in addition

$$PV(t)P^{\perp}=0,\quad \forall t\geqslant 0. \tag{8.4}$$

Using the invariance condition (3.10), (3.7) and (8.4) we obtain

$$V(t)^{\dagger}P^{\perp}KP^{\perp}V(t)=K-A(t)^{\dagger}KA(t)\geqslant 0.$$

Then the pair $K_P=PKP$ and $A(t)$ satisfy (8.2) and the CP property is fulfilled. The correlation function (4.12) now takes a simple form obtained just by iterating the dynamical map and the Weyl operators. Setting $t_0=t_1$ in (4.1) or (4.12) we find that

$$R_n(\mathbf{x}|\mathbf{y})=\rho_{\mathscr{A}}\{W(-x_1)T[W(-x_2)T[\ldots T[W(-x_n)W(y_n)]\ldots]]W(y_2)]W(y_1)\}. \tag{8.5}$$

We already remarked that for a dissipative Markov process the underlying closed system dynamics must contain a shift.[18] This applies to the family $V(t)$ in the present situation. If we restrict all considerations to the minimal subspace $\mathscr{H}_+\subseteq\mathscr{H}$ defined in Sec. IV we can retain the condition (4.11) for the existence of a subdynamics and the conditional expectation E.

After some manipulations the semigroup property of the maps $A(t)$ is obtained in the following form:

$$A(t)=PV(t)P=A(1)^t,\quad t=1,2,\ldots. \tag{8.6}$$

If this relation holds for all t, as indicated, then (8.4) is fulfilled. If (8.4) does not hold, simple calculations show that (8.6) holds for a finite number of values $t=1,2,\ldots,n$ if and only if the same finite order of correlation functions has the Markov form (8.5).

For completeness the properties of the CP semigroups in the continuous time case will be spelled out.[32] In the stationary case the semigroup property for $A(t), t\geqslant 0$, means that $A(t)=\exp(tL)$, where the matrix L lives in the subspace $P\mathscr{H}$, of course. From (3.4) the CP property reads

$$L^{\dagger}K_P+K_PL\leqslant 0,$$

which implies (8.2), and we get the CP property for the maps $T(t)$ defined by by $A(t)$ and (4.6), and the state on $\mathscr{S}$ defined by G_P is stationary. This is generalized to the case where L depends on time:

$$\frac{dA(0,t)}{dt}=A(0,t)L(t), \tag{8.7}$$

leading to a time-ordered exponential, of course, corresponding to the semigroup rule $T(0,s)T(s,t)=T(0,t)$, $0<s<t$. This implies for $F(0,t)$

$$F(0,t+dt)=F(t,t+dt)+A(t,t+dt)^{\dagger}F(0,t)A(t,t+dt),\quad 0<t<t+dt.$$

With the notation $M(t)=\lim_{\delta t\to 0}F(t,t+\delta t)/\delta t$ the differential equation becomes

$$\frac{dF(0,t)}{dt}=M(t)+L(t)^{\dagger}F(0,t)+F(0,t)L(t).$$

The CP property is assured by choosing $\{M(t),L(t)\}$ satisfying the differential form of (3.2):

$$M(t)-i[L(t)^{\dagger}S+SL(t)]\geqslant 0,\quad \forall t>0.$$

For dealing with Markov processes in continuous time with greater mathematical rigor one can use the formalism of quantum stochastic calculus.[33,34]

IX. RESPONSE TO TIME DEPENDENT FORCES

In this section we let the time parameter t be continuous, and consider the response of the system $\mathscr{S}+\mathscr{R}$ to external forces acting on the subsystem $\mathscr{S}$. It is well known that the equations of motion for the Hamiltonian (reversible) dynamics of a QHO can be solved explicitly for certain types of time dependent parameters and this applies to $\mathscr{S}+\mathscr{R}$.[35] It is also clear that the expression (4.1) for the full set of correlation functions of all orders [and hence the covariance function (4.3)] does in principle determine the response of the system to a large class of time dependent external perturbations acting on $\mathscr{S}$. The necessary condition is that this action can be expressed as a limit of combinations of the operations

$$X \mapsto W(-x)XW(y), \quad x,y \in P\mathscr{H}. \tag{9.1}$$

As the Weyl operators are dense in the CCR algebra this allows a large set of operations. There is an interesting subclass of harmonic forces which generate quasifree maps and conserve the quasifree form of the correlation functions.

We start with the simplest case of a force corresponding to a perturbation Hamiltonian linear in the canonical operators. In this particular case we can give an answer which is not dependent on the existence of a subdynamics. The external (generalized) force $f(t) \in P\mathscr{H}$, $t \in \mathbb{R}_+$, is understood to mean, in the absence of the intrinsic dynamics, a scalar translation of the canonical operators (a gauge transformation)

$$\frac{d}{dt}X(t) = \mathbb{1} \cdot \xi(t),$$

which is generated by a linear Hamiltonian $\Sigma_k f(t)_k X_k = f(t)X$. From the CCR we find $\xi(t) = i[f(t)X, X(t)] = -2f(t)S$. We can implement $f(t)$ as a driving force on the system defined by the correlation function (4.1) by introducing the following choices of arguments in (4.2):

$$y_k = x_k = -f(t_k)(t_{k+1} - t_k) \in P\mathscr{H}, \quad k = 1, \ldots, n-1,$$

$$y_n = x, \quad x_n = 0.$$

The exponent in (4.2) is found to be $g(x) - i(a|x)$ where the translation vector a is given by

$$a = \sum_{k=0}^{n-1} (t_{k+1} - t_k)\ \xi(t_k) A(t_n - t_k).$$

The obvious transition to a continuous time parameter gives us

$$a(t) = \int_0^t ds\, \xi(s) A(t-s) = -2 \int_0^t ds\ f(s) S A(t-s).$$

Here $A(t)$ is defined by (4.9) independently of the validity of (4.11). Using (9.1) (with $x = y$) to translate the initial state it is shown, even more directly, that $A(t)$ also is the after-effect function which describes the relaxation of an initial deviation from the equilibrium value for the expectation of the canonical operators of $\mathscr{S}$.

Now turn to the case of a time dependent quadratic Hamiltonian perturbation in the subspace $P\mathscr{H}$. This is an interesting situation which includes squeezing of oscillators with a dissipative evolution. The response of $\mathscr{S}+\mathscr{R}$ should still be defined by (4.3), but in order to obtain a more explicit formula involving the function $A(t)$ we have to assume a subdynamics.

The Hamiltonian driving means that there is a real matrix function $L(t) \in B(P\mathscr{H})$ satisfying

$$L(t)^\dagger S_P + S_P L(t) = 0, \tag{9.2}$$

which generates a family of symplectic maps through (8.7), i.e., satisfying (3.9), but which does not leave G invariant, in general. If the unperturbed evolution of the whole system is given by $V(t)$, then the response to this external driving force acting on the small subsystem is then given by standard time dependent methods as the solution of

$$U(0,t)=V(t)+\int_0^t ds\ U(0,s)L(s)V(t-s). \tag{9.3}$$

Inserting this dynamic in (4.1), instead of $V(t)$ and with appropriate values for the time parameters, we obtain all multitime correlations for the perturbed dynamics. However, stationarity will no longer hold, in general. We get a corresponding equation for the evolution of any operator in the Hilbert space

$$X_d(t)=X(t)+\int_0^t ds\ V(t-s)^\dagger[L(s)X_d(s)+X_d(s)L(s)]V(t-s), \tag{9.4}$$

using the notation $X_d(u)=U(0,u)^\dagger XU(0,u), X(u)=V(u)^\dagger XV(u)$.

We are interested in the favorable cases where the effect of the driving can be derived directly from the functions $\{A(t),F(t)\}$ describing the unperturbed stationary subdynamics. Using $P^\perp L(s)=L(s)P^\perp=0$ we find from (9.3) the corresponding equation for the family of operators $A_d(0,t)=PU(0,t)P$:

$$A_d(0,t)=A(t)+\int_0^t ds\ A_d(0,s)L(s)A(t-s),$$

with initial condition $A_d(0,0)=P$, so the solution is completely defined by $\{A(t),L(t)\}$.

We need also the evolution of the state of the system. If we put $X=K$ in (9.4) and assume (4.11) and (3.10) for all $V=V(t)$, then it follows that

$$PK_d(t)P=K_P+\int_0^t ds\ PV(t-s)^\dagger[L(s)K_d(s)+K_d(s)L(s)]V(t-s)P.$$

Now, if we assume that $[P,K_d(t)]=0,\forall t>0$, i.e., that the state of the system still factorizes for all times, then we find that

$$PK_d(t)P=K_P+\int_0^t ds\ A(t-s)^\dagger[L(s)PK_d(s)P+PK_d(s)PL(s)]A(t-s)P,$$

and the solution is again defined by $\{K_P,A(t),L(t)\}$. However, the factorization will not hold in general. In fact,

$$PK_d(t)P^\perp=\int_0^t ds\ PV(t-s)^\dagger[L(s)K_d(s)+K_d(s)L(s)]V(t-s)P^\perp,$$

and there seems to be no reason why this should have only the trivial solution in the general case. It is not a homogenous system as matrix elements in $PK_d(t)P$ enter.

In the Markov case (8.4) holds, which means that the equation above reduces to

$$PK_d(t)P^\perp=\int_0^t ds\ A(t-s)^\dagger L(s)PK_d(s)P^\perp V(t-s)P^\perp,$$

which is a homogenous equation, initial value $PK_d(0)P^\perp=0$, which will have only the zero solution under some regularity assumptions (allowing a series expansion in the time parameter, for instance). We also find from (9.3) that

$$PU(s,t)P^\perp=PV(t-s)P^\perp+\int_s^t du\ PU(s,u)PL(u)PV(t-u)P^\perp,$$

which in general has a nonzero solution. However, if (8.4) holds for $V(t)$, then it clearly holds for $U(0,t)$ as well, so we have the Markov property for the driven process. Writing $A(t)=\exp(tL_0)$ we find from the simple form of the Markovian correlation functions (8.5) that

$$\frac{d}{dt}A_d(0,t)=A_d(0,t)[L_0+L(t)],$$

$$\frac{d}{dt}F(t)_d=M(t)+[L_0+L(t)]^\dagger F(t)_d+F(t)_d[L_0+L(t)].$$

Together these equations define the driven Markovian subdynamics from which we can recover the full set of correlation functions.

The fact that the full set of quantum correlation functions define the response of the system to external forces acting on $\mathscr{S}$ means that they contain more information on the dynamics of $\mathscr{S}+\mathscr{R}$ than the corresponding data do in commutative probability theory. This is just another of the crucial non-commutative properties. The standard way of finding relations between fluctuations and the response to external forces from the Hamiltonian dynamics of the closed system $\mathscr{S}+\mathscr{R}$ is the linear response theory set down in textbooks on statistical mechanics.[36] In the present formalism, and with forces acting only on $\mathscr{S}$, the same information should be given by the covariance function (4.3). However, if we impose the condition (4.11), then there will result a classical version of the fluctuation-dissipation relation with the proportionality (4.7) between $K(t)$ and $A(t)$. This is also essentially the classical Onsager relation between the regression of fluctuations and the decay of deviations from equilibrium. The quantum version of these relations demand that we retain the exact covariance function $K(t)$ which is not of the form (4.7). This point has been discussed recently in the framework of the Langevin equation approach.[37]

X. APPROXIMATE VALIDITY OF THE SUBDYNAMICS PICTURE

Now consider the case where (3.7) does not hold. Then there will be a difference between the covariance matrix coming from the evolution of $\mathscr{S}+\mathscr{M}$ and the expression (4.7), where A is defined by (4.9), which corresponds to a subdynamics. The difference is

$$PKV(t)P-PKA(t)=PKP^\perp V(t)P. \tag{10.1}$$

As the covariance matrix defines the dynamics, if we can show that the matrix elements of (10.1) are negligible in suitable circumstances, we will also obtain conditions for $\{K_P,A(t)\}$ approximating a subdynamics.

Note that such conditions can be fulfilled for some ranges of t and not for others, so we can discuss this problem for a given t and leave out this parameter. Clearly we have to use the properties of V as well as those of K. The physical idea is that the effects of the back action of $\mathscr{S}$ on $\mathscr{R}$ spreads over a large number of degrees of freedom of $\mathscr{R}$ on a time scale which is shorter than the relaxation time for $\mathscr{S}$.

We want to see how this idea could work in the present scheme, but lacking an explicit model for the heat bath we can do this only in very general terms. Consider arbitrary unit vectors $x,y\in P\mathscr{H}$ and the matrix element $(x|KP^\perp V|y)$. Also introduce the vectors in $\mathscr{H}$: $\phi=K|x)$, $\chi=P^\perp V|y)$. We can then write

$$|(x|KP^\perp V|y)|^2=|(\phi|\chi)|^2=(\phi|\phi)(\chi|\chi)|(\hat{\phi}|\hat{\chi})|^2,$$

where $\{\hat{\phi},\hat{\chi}\}$ are normalized vectors. The simplest case is where the matrix elements of $P^\perp VP$ behave effectively as independent random variables, and all of a similar magnitude. Then $\hat{\chi}$ can be considered randomly chosen in the Hilbert space $P^\perp\mathscr{H}$. This corresponds to an equidistribution over the degrees of freedom of $\mathscr{R}$. We can then make the order of magnitude estimate $|(\hat{\phi}|\hat{\chi})|^2\simeq(\dim P^\perp\mathscr{H})^{-1}$, and for a reservoir with a large number of degrees of freedom we see why the matrix elements of (10.1) can be expected to be negligible. In more general and realistic models of heat baths we may have to replace $P^\perp\mathscr{H}$ by a Hilbert space of degrees of freedom of $\mathscr{R}$ which has an effective dimension increasing with the strength of the $\mathscr{S}-\mathscr{R}$ interaction.

Introduce $B=P^{\perp}VP$, then (10.1) equals PKB. Using the invariance of K we can write

$$K_P=A^{\dagger}KA+B^{\dagger}KB+A^{\dagger}KB+B^{\dagger}KA, \tag{10.2}$$

where the last two terms in the RHS are going to negligible if the preceding argument holds. If the dynamics mixes the degrees of freedom of $\mathscr{S}+\mathscr{R}$ thoroughly, then, by the same kind of reasoning, also the first term will be negligible and only the second term remains (under the assumption $\dim P.\mathscr{H} \ll \dim P^{\perp}.\mathscr{H}$). This situation corresponds to a final equilibrium for the open system, which means that the lowest order correlation approximately factorizes

$$R_2(\mathbf{x}|\mathbf{y}) \approx R_1(x_1|y_1)R_1(x_2|y_2).$$

In order for $\{K_P, A\}$ to reproduce $PKVP$ in a non-trivial way we then find that the first two terms in (10.2) must be of comparable magnitude, while the last two are negligible. If this holds for a useful range for the time parameter we will obtain a subdynamics for $\mathscr{S}$ which approximates the exact covariance function in this range.

We can use similar arguments to discuss the degree of approximation for the Markov property. Again using a discrete time parameter, in order to show that (8.6) holds to some approximation up to a given finite order n, we must show that the matrix elements of the following operator are small:

$$PKV(n)P-PKA(1)^n=PK\left\{P^{\perp}V(n)+\sum_{k=2}^{n}A(1)^{n-k}PV(1)P^{\perp}V(k-1)\right\}P.$$

Each term in the RHS gives matrix elements of a similar structure to those coming from the RHS of (10.1). Thus we can make the previous assumptions for each of the iterates $V(k)$ of the dynamical map to get sufficient conditions.

XI. EVOLUTION OF NON-STATIONARY INITIAL STATES

We have seen the rigorous conditions which have to be fulfilled for a subdynamics to exist and also how much information about the system is contained in this structure. If we renounce the full set of correlation functions and let ourselves be satisfied by a family of dynamical maps giving the evolution of an arbitrary initial state of $\mathscr{S}$, then the restrictions on the existence of solutions will largely disappear.

In the lowest order correlation function

$$\rho[W(-x)W(V(t)y)]=R(x,y;t), \quad x,y\in P.\mathscr{H},$$

we can set $x=0$ to find a restricted function of one argument $y\in P.\mathscr{H}$ which gives the evolution of the state of $\mathscr{S}$:

$$R(0,y;t)=\rho_{\mathscr{S}}(t)[W(y)], \quad t\geqslant 0.$$

If the initial ($t=0$) state ρ of $\mathscr{S}+\mathscr{R}$ factorizes, the family of maps $T(t)$ defined in Sec. VII by $\{A(t),D(t,t)\}$ also gives the evolution of the state

$$\rho_{\mathscr{S}}(t)[W(y)]=\rho_{\mathscr{S}}(0)\{T(t)[W(y)]\}.$$

We can now ask: if the initial state does not factorize, can we still describe the evolution of the initial state exactly by such evolution maps $T(t)$? The answer is yes, there is now no restriction, except that the maps will fail to be CP, or even positive, in general. Let $T(t)$ be defined by $\{A(t),F(t)\}$, which do not necessarily satisfy (3.2), while $\rho_R(t)$ is defined by the quadratic form $G_P(t)$. In order to have the desired evolution of the partial state of $\mathscr{S}$ we should choose F as follows:

$$F(t)=G_P(t)-A(t)^{\dagger}G_P(0)A(t)=PV(t)^{\dagger}P^{\perp}GP^{\perp}V(t)P-PV(t)^{\dagger}P^{\perp}GPA(t) \\ -A(t)^{\dagger}PGP^{\perp}V(t)P. \tag{11.1}$$

This expression is actually independent of $G_P(0)=PGP$, hence $\{A(t),F(t)\}$ defines a family of maps on $G_P(0)$. If we choose $G_P(0)$ arbitrary while $\{PGP^\perp,P^\perp GP,P^\perp GP^\perp\}$ are kept constant, then the initial ''state'' of $\mathscr{S}+\mathscr{R}$ will not necessarily be positive. Consequently the maps will not be positive in general, and they will not reproduce the covariance function unless (4.10) holds. (In the latter case the two last terms in (11.1) vanish, (3.2) holds, and we are back to a factorizing initial state!) Still, there will be a subset of initial values $G_P(0)$ which are mapped into a genuine, positive state of $\mathscr{S}$ at time t.

If the family $A(t)$ forms a semigroup, then there is a semigroup of maps $T(s,t)$ defined by $\{A(t),F(s,t)\}$ where

$$F(s,t)=G_P(t)-A(t-s)^\dagger G_P(s)A(t-s), \quad s<t.$$

Using some limiting procedures (which involves a manipulation of the time scales) one can obtain such a semigroup of maps which are generally not positive, but which still serve as an effective reduced dynamics for $\mathscr{S}$, perhaps relevant only for a restricted set of initial conditions.[25] For this allowed set there may still be convergence to a genuine positive final state of $\mathscr{S}$.

In order to see how this may work, consider a discrete semigroup. We have a map T defined by $\{A,F\}$ which need not satisfy (3.2). Assume that the semigroup has a unique limit state, which means that for any quadratic G

$$\lim_{n\to\infty} A^{n\dagger}GA^n=0.$$

Then it is easy to see that the limit obtained by iterating the transformation $G\mapsto A^\dagger GA+F$ is the following sum (if it converges):

$$G_P(\infty)=\sum_{n=0}^{\infty} A^{n\dagger}FA^n.$$

The limit is a genuine state for $\mathscr{S}$ if and only if $G_P(\infty)+iS_P=K_P(\infty)\geqslant 0$. Given any matrix $G_P(\infty)$ satisfying this condition, we can choose $F=G_P(\infty)-A^\dagger G_P(\infty)A$ to obtain a semigroup of maps such that every initial ''state'' (positive or not) converges to the final state defined by $G_P(\infty)$. The map T is CP if (8.2) holds for $K_P(\infty)$, but this is not true in general.

XII. DISCUSSION

There is an overwhelming number of papers on the dynamics of open quantum systems, even if we restrict ourselves to systems of oscillators. Here we can only mention a few which deal with the problems involved in the existence of system-reservoir correlations and of the intrinsic time scale $\beta=\hbar/k\Theta$.

The first problem we meet in the literature is that of deriving a quantum Markovian master equation for an open system $\mathscr{S}$. The fact that $\beta\neq 0$ when $\Theta<\infty$ is incompatible with a strictly exponential relaxation has been appreciated for a long time.[22–27,38,39] A derivation of master equations from the microscopic dynamics inevitably involves a separation of time scales and some form of averaging method.[40,41] The quantum case differs from the classical one mainly through the fact that the time scale β is independent of the details of the reservoir. The master equation alone is equivalent with having a semigroup of dynamical maps. We have already remarked that the full set of multitime correlation functions (8.5) follows from the semigroup property for the model treated here, but this is not so for general models of open quantum systems.

The second problem is that coming from the system-reservoir correlations. Most papers on the derivation of master equations from the microscopic dynamics accept the factorization of the initial state, at least. However, we have seen in Sec. IV how the correlations forbid an exact subdynamics, and in particular having a family of CP maps defining the covariance function. In Sec. XI we saw that there are non-positive dynamical maps which can describe the evolution of some initial conditions. The role of the correlations in destroying the positivity of the dynamical maps has been the subject of a considerable number of papers, of which we can list just a few.[25,42–45]

One possible way out of the factorization condition is thus to restrict oneself to finding a family of dynamical maps which preserve the positivity of a subset of initial states, though they cannot be extended to positive maps on the whole state space. We then need a prescription of how such states are prepared from a stationary initial state of the composite system, e.g., by the action on the subsystem $\mathscr{S}$ by a well-defined set of operations. This is a solution which has been suggested in several places.[25,43,45] In particular, certain versions of the path integral formalism allow correlated initial states.[8,41] The evolution maps can exist for some correlated states of the system, but they will not define the very strong and rigid structure we have called a subdynamics, and which contain much more information than just these maps. For this reason there is no contradiction in the body of work cited above with the results of earlier papers where the structure of a subdynamics (called a quantum stochastic process there) was imposed more or less as an axiom.[1,2] Neither is there a contradiction of the results on the form of the generators of quantum dynamical semigroups, which were assumed to act on the space of all states of $\mathscr{S}$.[29–31]

The main message of this paper, and of Ref. 3, is that the problems remain even if we drop the Markov property. The correlations in the initial state can not be accounted for just by introducing some memory in the process. In the present scheme based on the CCR, the dynamics of the open system is described completely by the time dependence of the covariance matrix, and this description does not necessarily involve any approximation. However, the non-commutativity of the observables means that this time dependence cannot be exactly reproduced by dissipative dynamical maps acting only on the observables of the open system, except when the state factorizes for all times. (This happens in particular when the reservoir has an infinite temperature, i.e., $\beta=0$.) We can still have a subdynamics as an approximation involving assumptions on the time scale and on the relaxation properties of the reservoir. These assumptions are weaker than those needed to make the dynamics Markovian, but they are of a similar type.

In this paper we have not tried to compare the formalism with the Langevin equation approach. However, it seems fairly clear that using colored noise in a quantum Langevin equation (as one should in order to stay consistent with thermodynamics) will give a covariance function which is not of the type which is reducible to a subdynamics. It was noted that the solutions will not give Markov processes even if the Langevin equation itself has no memory term.[46]

Finally we note that the present model is an exceptional one in that there are methods of finding exact solutions.[25,47,48] This allows a solution for the covariance function, which is sufficient to define the dynamics of the whole system $\mathscr{S}+\mathscr{R}$ even when no subdynamics exists.

ACKNOWLEDGMENT

This research was partially supported by the Swedish Natural Science Research Council.

[1] G. Lindblad, Commun. Math. Phys. **65**, 281 (1979).
[2] G. Lindblad, J. Math. Phys. **20**, 2081 (1979).
[3] G. Lindblad, J. Phys. A **29**, 4197 (1996).
[4] M. Takesaki, J. Funct. Anal. **9**, 306 (1972).
[5] L. Accardi and C. Cecchini, J. Funct. Anal. **45**, 245 (1982).
[6] L. Accardi, A. Frigerio, and J.T. Lewis, Pub. Res. Inst. Math. Sci. **18**, 97 (1982).
[7] H. Dekker, Phys. Rep. **80**, 1 (1981).
[8] H. Grabert, P. Schramm, and G-L. Ingold, Phys. Rep. **168**, 115 (1988).
[9] U. Weiss, *Quantum Dissipative Systems* (World Scientific, Singapore, 1992).
[10] A. Isar, A. Sandulescu, H. Scutaru, E. Stefanescu, and W. Scheid, Int. J. Mod. Phys. E **3**, 635 (1994).
[11] G. Lindblad, Rep. Math. Phys. **10**, 393 (1976).
[12] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, Commun. Math. Phys. **32**, 231 (1973).
[13] K. R. Parthasarathy and K. Schmidt, *Positive Definite Kernels, Continuous Tensor Products and Central Limit Theorems of Probability Theory*, Lecture Notes in Mathematics (Springer, Berlin, 1972) Vol. 272.
[14] B. Demoen, P. Vanheuverzwijn, and A. Verbeure, Lett. Math. Phys. **2**, 161 (1977).
[15] B. Demoen, P. Vanheuverzwijn, and A. Verbeure, Rep. Math. Phys. **15**, 27 (1979).
[16] D. E. Evans, J. T. Lewis, Comm. Dubl. Inst. Adv. Studies Ser. A, **24** (1977).
[17] V. I. Paulsen, *Completely Bounded Maps and Dilations* (Longman Scientific & Technical, London, 1986).
[18] G. Lindblad, J. Phys. A **26**, 7193 (1993).
[19] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space* (North Holland, Amsterdam, 1970).
[20] J. Rau and B. Müller, Phys. Rep. **272**, 2 (1996).
[21] E. H. Lieb and M. B. Ruskai, Adv. Math. **12**, 269 (1974).
[22] G. Lindblad, *Non-Equilibrium Entropy and Irreversibility* (Reidel, Dordrecht, 1983).
[23] P. Talkner, Ann. Phys. (N.Y.) **167**, 390 (1986).
[24] K. Lindenberg and B. J. West, Phys. Rev. A **30**, 568 (1984).

[25] F. Haake and R. Reibold, Phys. Rev. A **32**, 2462 (1985).
[26] P. Schramm and H. Grabert, J. Stat. Phys. **49**, 767 (1987).
[27] V. Gorini, M. Verri, and A. Frigerio, Physica A **161**, 357 (1989).
[28] O. Bratteli and D. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer, New York, 1979, 1981), Vols. I and II.
[29] G. Lindblad, Commun. Math. Phys. **48**, 119 (1976).
[30] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, J. Math. Phys. **17**, 821 (1976).
[31] R. Alicki and K. Lendi, *Quantum Dynamical Semigroups and Applications* (Springer, Berlin, 1987).
[32] P. Vanheuverzwijn, Ann. Inst. Henri Poincaré, Sect. A **29**, 123 (1978).
[33] K. R. Parthasarathy, *An Introduction to Quantum Stochastic Calculus* (Birkhäuser, Basel, 1992).
[34] P-A. Meyer, *Quantum Probability for Probabilists* (Springer, Berlin, 1993).
[35] K. Husimi, Prog. Theor. Phys. **9**, 381 (1953).
[36] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics, 2nd ed.* (Springer, Berlin, 1991).
[37] G. W. Ford and R. F. O'Connnell, Phys. Rev. Lett. **77**, 798 (1996).
[38] R. L. Stratonovich and O. A. Chichigina, JETP **78**, 56 (1994) [Zh. Eksp. Teor. Fiz. **105**, 106 (1994)].
[39] R. L. Stratonovich, Physica A **236**, 335 (1997).
[40] G. W. Ford, J. T. Lewis, and R. F. O'Connell, Ann. Phys. (N.Y.) **252**, 362 (1996).
[41] R. Karrlein and H. Grabert, Phys. Rev. E **55**, 153 (1997).
[42] A. Suárez, R. Silbey, and I. Oppenheim, J. Chem. Phys. **97**, 5101 (1992).
[43] P. Pechukas, Phys. Rev. Lett. **73**, 1060 (1994).
[44] A. Royer, Phys. Rev. Lett. **77**, 3272 (1996).
[45] S. Gnutzmann and F. Haake, Z. Phys. B **101**, 263 (1996).
[46] G. W. Ford, J. T. Lewis, and R. F. O'Connnell, Phys. Rev. A **37**, 4419 (1988).
[47] P. S. Riseborough, P. Hänggi, and U. Weiss, Phys. Rev. A **31**, 471 (1985).
[48] S. K. Kehrein and A. Mielke, Ann. Phys. (Leipzig) **6**, 90 (1997).

Study of hydrodynamical limits in a multicollision scale Boltzmann equation for semiconductors

Philippe Montarnal[a)]
INRIA, M3N project, 78153 Le Chesnay, Cedex, France

(Received 3 November 1997; accepted for publication 1 December 1997)

We derive in this paper a generalized hydrodynamical model for semiconductors from a multicollision scale Boltzmann equation. We apply a Chapman–Enskog method for two time-scale collisional operators. These two time scales are linked respectively with the elastic part of the electron/lattice scattering term and the electron/electron collision term. Our model has a global entropy function and a symmetric formulation. It is a generalization of the classical hydrodynamical model since we add two terms: a ''thermal friction force'' and a ''friction heat flux.'' Moreover, with some changes of parameters we obtain both the energy-transport model and the Euler–Poisson hydrodynamical model. © *1998 American Institute of Physics.* [S0022-2488(98)03404-5]

I. INTRODUCTION

In order to study submicrometer semiconductors, it is necessary to use richer macroscopic models than the drift-diffusion model. In particular, one must take into account the energy of the carriers. The two main fluid models of this type are currently the hydrodynamical model (obtained with the three first moments of the Boltzmann transport equation and some closure conditions[1,2]) and the energy-transport model (obtained with the moments of order zero and two of the Boltzmann transport equation at a diffusion scale[3–5]). Here, we develop a unified vision of these two classes of models.

We derive in this paper a generalized hydrodynamical model for semiconductors from a multicollision scale Boltzmann equation. We apply a Chapman–Enskog method[6–8] for two time-scale collisional operators. These two time scales are linked respectively with the elastic part of the electron/lattice scattering term and the electron/electron collision term. Our model has a global entropy function and a symmetric formulation. It is a generalization of the classical hydrodynamical model since we add two terms: a ''thermal friction force'' and a ''friction heat flux.'' Moreover, with some changes of parameters we obtain both the energy-transport model and the Euler–Poisson hydrodynamical model.

We consider a nondegenerate electron gas with a parabolic band structure. Using the same scaling as in Refs. 5, 9, and 10, we obtain the following perturbed Boltzmann transport equation:

$$\partial_t f + \frac{1}{\varepsilon_1}\, \underline{v}.\nabla_x f + \frac{1}{\varepsilon_1}\, \nabla_x \Phi.\nabla_v f = \frac{1}{\varepsilon_1^2}\left(\frac{1}{\varepsilon_2}\, Q_{e,e}(f,f) + Q_{\text{elas}}(f)\right) + Q_{\text{ine}}(f), \quad \text{(BTE)}$$

where $f = f(\underline{x},\underline{v},t)$ is the electron distribution function depending on the position $\underline{x}$, the velocity $\underline{v}$ and the time t, and Φ is the electric potential which is assumed to be given and smooth.

The specifications of the collision operators are the following: $Q_{e,e}$ describes binary collisions between electrons, Q_{elas} models impurity scattering and the elastic part of phonon scattering, and Q_{ine} accounts for the inelastic part of phonon scattering.

The dimensionless parameter ε_1^2 measures the typical energy gain or loss during a collision with a phonon (inelastic contribution), while ε_2 quantifies the ratio between the magnitude orders of the phonon scattering and the electron/electron collision (see Ref. 9).

We assume in this paper that

a)Electronic mail: philippe.montarnal@inria.fr

 © 1998 American Institute of Physics

$$\varepsilon_1 \ll 1, \quad \varepsilon_2 \ll 1. \tag{1}$$

In this case the elastic part of phonon scattering is considered to be dominant as compared with the inelastic part (ε_1^2, the relative energy gain or loss of electron energy during a phonon collision, is very small) and the electron/electron collision operator is of the dominant order of magnitude. This last hypothesis on ε_2 is not studied in Ref. 9 but is justified in the case of high doping.[11,12]

We introduce the density n, the mean velocity $\underline{u}$, and the energy e of the electrons by

$$\begin{pmatrix} n \\ \varepsilon_1 n\underline{u} \\ ne \end{pmatrix} = \int_{\mathbb{R}^3} \underline{K}(\underline{v}) f \, d\underline{v}, \tag{2}$$

with

$$\underline{K}(\underline{v}) = \begin{pmatrix} 1 \\ \underline{v} \\ \dfrac{|\underline{v}|^2}{2} \end{pmatrix}.$$

Thanks to the scale in the definition of the velocity $\underline{u}$, we later deal with order one quantities in the mass conservation equation [see Eq. (13)].

We also define the electronic temperature T by

$$e = \tfrac{1}{2}\varepsilon_1^2 |\underline{u}|^2 + \tfrac{3}{2} T. \tag{3}$$

In this paper we will derive from the Boltzmann transport equation (BTE) the following macroscopic system of $2+3$ equations [considering assumption (1)]:

$$\partial_t n + \operatorname{div}(n\underline{u}) = 0,$$

$$\partial_t(\varepsilon_1^2 n\underline{u}) + \operatorname{div}(\varepsilon_1^2 n\underline{u} \otimes \underline{u} + nT\underline{\underline{I_d}} + \varepsilon_2 \nu_4(T)\underline{\underline{I_d}}) - n\nabla_x \Phi$$
$$+ \left(\nu_1(T) + \frac{\varepsilon_1}{\tau_{\text{ine}}} + \varepsilon_1^2 \nu_2(T, \underline{u}) + \frac{\varepsilon_2}{n} \nu_3(T) \right) n\underline{u} = 0, \quad \text{(HD)}$$

$$\partial_t\left(\tfrac{1}{2} n(\varepsilon_1^2 |\underline{u}|^2 + 3T) \right) + \operatorname{div}\left(\tfrac{1}{2} n\underline{u}(\varepsilon_1^2 |\underline{u}|^2 + 5T) + \varepsilon_2 \nu_5(T)\underline{u} - \varepsilon_2 \kappa(T) \nabla_x T \right)$$
$$- n\underline{u}\nabla_x \Phi + n \frac{e - \frac{3}{2}T_0}{\tau_e} = 0.$$

This system contains at once the classical hydrodynamical model[1,13] (obtain by cancelling the ''thermal friction force'' and the ''friction heat flux'' defined in Sec. V), the Euler–Poisson model (obtained when $\varepsilon_2 \to 0$), and the energy-transport model[3–5] (obtained when $\varepsilon_1 \to 0$).

Moreover, our model has an entropy function

$$\eta = n \ln\left(\frac{n}{(2\pi T)^{3/2}} \right) - \frac{5}{2} n, \quad \text{(EF)}$$

which is linked with the following entropy conservation equation (which can be obtained from the H theorem)

$$\partial_t\left(n \ln\left(\frac{n}{(2\pi T)^{3/2}} \right) - \frac{5}{2} n \right) + \operatorname{div}\left(n\underline{u} \ln\left(\frac{n}{(2\pi T)^{3/2}} \right) - \frac{5}{2} n\underline{u} - \frac{\varepsilon_2}{T} \nu_5(T)\underline{u} + \varepsilon_2 \frac{\kappa(T)}{T} \nabla_x T \right)$$
$$= -\frac{n}{T}\left(\nu_1(T) + \frac{\varepsilon_1}{\tau_{\text{ine}}} + \varepsilon_1^2 \nu_2(T, \underline{u}) + \frac{\varepsilon_2}{n} \nu_3(T) \right) |\underline{u}|^2 - \frac{n}{T} \frac{e - \frac{3}{2}T_0}{\tau_e} - \varepsilon_2 \frac{\kappa(T)}{T^2} |\nabla_x T|^2 \quad \text{(EL)}.$$

This entropy function is compatible with both the hyperbolic entropy of the classical hydrodynamical model and the parabolic entropy of the energy-transport model. Moreover, the entropy function η is also linked with the existence of a symmetric formulation of the system (HD).

An outline of this paper is as follows. In Sec. II we give the definition and the principal properties of the collision operators. Then, in Sec. III, we derive the macroscopic hydrodynamical system with a Chapman–Enskog method. Section IV is devoted to the study of the entropy function and its links with the symmetric variables. Finally, in Sec. V, we describe the relations of our model with other semiconductor models.

We derive the macroscopic entropy relation from the kinetic entropy in Appendix A. In Appendix B we give the proof of Lemma 10.

II. DEFINITION AND PROPERTIES OF THE COLLISION OPERATORS

In this section, we first describe $Q_{e,e}$, the nonlinear Boltzmann collision operator for binary collisions between electrons. We present the H theorem and we introduce the associated linearized operator. Then, we present the operator Q_{elas} and its properties (H theorem), and the collision term Q_{ine} as an energy relaxation term.

The electron–electron collision operator is given by (see Ref. 14 and the references therein)

$$Q_{e,e}(f,f)=\int_{\mathbb{R}^3}\int_0^{2\pi}\int_0^{\pi/2}B_{e,e}(|\underline{v}-\underline{v}_1|,\theta)(f'f_1'-ff_1)d\theta\, d\varphi\, d\underline{v}_1,$$

with $f=f(\underline{v})$, $f=f_1(\underline{v}_1)$, $f'=f(\underline{v}')$, and $f_1'=f(\underline{v}_1')$. Here $\underline{v}$ and $\underline{v}_1$ are the velocities before the collision, $\underline{v}'$ and $\underline{v}_1'$ the velocities after the collision:

$$\underline{v}'=\underline{v}+(\underline{v}_1-\underline{v}).\underline{\omega}\,\underline{\omega},\qquad \underline{v}_1'=\underline{v}_1-(\underline{v}_1-\underline{v}).\underline{\omega}\,\underline{\omega},$$

with $\underline{\omega}\in\mathscr{S}^2_+=\{\underline{\omega}\in\mathscr{S}^2\text{ s.t.}(\underline{v}_1-\underline{v}).\underline{\omega}\geqslant 0\}$, $\mathscr{S}^2$ is the unit sphere of $\mathbb{R}^3$, and (θ,φ) are the spherical coordinates on $\mathscr{S}^2$.

The collision kernel $B_{e,e}$ satisfies the following conditions:

$$B_{e,e}\geqslant 0,\quad B_{e,e}(g,\theta)=\mathscr{B}_{e,e}(|g|,|g.\theta|). \tag{4}$$

The following H theorem holds (see Refs. 7 and 14).

Theorem 1: *Let $B_{e,e}$ satisfy (4). Then*

(i)

$$\int_{\mathbb{R}^3}Q_{e,e}(f)\ln f\,d\underline{v}\leqslant 0; \tag{5}$$

(ii) $Q_{e,e}(f)=0$ *if and only if for some* $n,T\geqslant 0$, $\underline{u}\in\mathbb{R}^3$,

$$f(\underline{v})=\frac{n}{(2\pi T)^{3/2}}\exp\left(-\frac{|\underline{v}-\underline{u}|^2}{2T}\right); \tag{6}$$

(ii)

$$\int_{\mathbb{R}^3}\underline{K}(\underline{v})Q_{e,e}(f,f)d\underline{v}=\begin{pmatrix}0\\0\\0\end{pmatrix}. \tag{7}$$

The linearized operator of $Q_{e,e}$ around a Maxwellian M is defined by (see Ref. 8)

$$\Gamma_{e,e}(g)=2M^{-1}Q_{e,e}(M,Mg).$$

Next, we introduce the notations

$$\mathscr{D}(\Gamma_{e,e})=\{g\in L^2(M\,d\underline{v})\,\text{s.t.}\,\vartheta_{e,e}g\in L^2(M\,d\underline{v})\},$$

$$\vartheta_{e,e}(\underline{v})=2\int_{\mathbb{R}^3}\int_0^{2\pi}\int_0^{\pi/2}B_{e,e}(|\underline{v}-\underline{v}_1|,\theta)M(\underline{v}_1)\,d\theta\;d\varphi\;d\underline{v}_1.$$

Also, we have the following (see Ref. 15).

Proposition 2:

(i) $\Gamma_{e,e}:\mathscr{D}(\Gamma_{e,e})\subset L^2(M\,d\underline{v})\to L^2(M\,d\underline{v})$ *is negative self-adjoint,*
(ii) $\mathrm{Ker}(\Gamma_{e,e})=\{a+\underline{b}.\underline{v}+c|\underline{v}|^2, a,c\in\mathbb{R},\underline{b}\in\mathbb{R}^3\}$,
(iii) $\forall g\in L^2(M\,d\underline{v}),\Gamma_{e,e}(f)=g$ *has a solution if and only if* $g\in(\mathrm{Ker}(\Gamma_{e,e}))^{\perp}$, i.e.,

$$\int_{\mathbb{R}^3}g\underline{K}(\underline{v})M\;d\underline{v}=0. \tag{8}$$

Moreover, we have unicity of the solution in $(\mathrm{Ker}(\Gamma_{e,e}))^{\perp}$.

We recall that Q_{elas} models impurity scattering and the elastic part of phonon scattering. Its definition is the following (see Ref. 10):

$$Q_{\mathrm{elas}}(f)=\int_{\mathscr{S}^2}B_{\mathrm{elas}}(|\underline{v}|,\alpha)(f(|\underline{v}|\underline{\omega})-f(\underline{v}))\;d\underline{\omega},$$

where α is the angle between $\underline{v}$ and $\underline{\psi}$. B_{elas} is related to the differential cross section and satisfies

$$B_{\mathrm{elas}}\geqslant 0,\;\; B_{\mathrm{elas}}(|\underline{v}|,-\alpha)=B_{\mathrm{elas}}(|\underline{v}|,\alpha). \tag{9}$$

Let us also introduce B^T_{elas} by

$$B^T_{\mathrm{elas}}=\int_{\mathscr{S}^2}B_{\mathrm{elas}}(|\underline{v}|,\alpha)\,d\underline{\omega}.$$

For Q_{elas}, we have the following H theorem:

Theorem 3: *Let* B_{elas} *satisfy (9). Then*

(i)

$$\int_{\mathbb{R}^3}Q_{\mathrm{elas}}(f)\ln f\;d\underline{v}\leqslant 0. \tag{10}$$

(ii) $Q_{\mathrm{elas}}(f)=0$ *if and only if for some* $n,T\geqslant 0$,

$$f(\underline{v})=\frac{n}{(2\pi T)^{3/2}}\exp\left(-\frac{|\underline{v}|^2}{2T}\right). \tag{11}$$

(iii) *If* $g=g(|\underline{v}|^2)$, *then*

$$\int_{\mathbb{R}^3}gQ_{\mathrm{elas}}(f)\,d\underline{v}=\begin{pmatrix}0\\0\end{pmatrix}.$$

Finally, Q_{ine} accounts for the inelastic part of phonon scattering. It expresses the exchange of energy with the high electron phonons. We will only write and use the following property:

$$\int_{\mathbb{R}^3}\underline{K}(\underline{v})Q_{\mathrm{ine}}(f)\,d\underline{v}=\begin{pmatrix}0\\-\dfrac{n\underline{u}}{\tau_{\mathrm{ine}}}\\-n\,\dfrac{e-\frac{3}{2}T_0}{\tau_e}\end{pmatrix}, \tag{12}$$

where τ_e and τ_{ine} are respectively the energy and momentum relaxation times from the inelastic contribution of phonon scattering.

III. DERIVATION OF THE HYDRODYNAMICAL MACROSCOPIC MODEL

The three first moments of the Boltzmann transport equation (BTE) give the following equations:

$$\partial_t n + \operatorname{div}(n\underline{u}) = 0, \tag{13}$$

$$\partial_t(\varepsilon_1^2 n\underline{u}) + \operatorname{div}(\varepsilon_1^2 n\underline{u}\otimes\underline{u} + \underline{\underline{p}}) - n\nabla_x\Phi = \underline{r}, \tag{14}$$

$$\partial_t(ne) + \operatorname{div}(\underline{u}(ne+\underline{\underline{p}}) + \underline{q}) - n\underline{u}\nabla_x\Phi = -n\,\frac{e-\frac{3}{2}T_0}{\tau_e}, \tag{15}$$

with

$$\underline{\underline{p}} = \int_{\mathbb{R}^3}(\underline{v}-\varepsilon_1\underline{u})\otimes(\underline{v}-\varepsilon_1\underline{u}) f\, d\underline{v},$$

$$\underline{r} = \frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}\underline{v}\,Q_{\text{elas}}(f)\,d\underline{v} - \varepsilon_1\,\frac{n\underline{u}}{\tau_{\text{ine}}}, \qquad \underline{q} = \frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}(\underline{v}-\varepsilon_1\underline{u})\,\frac{|\underline{v}-\varepsilon_1\underline{u}|^2}{2} f\, d\underline{v}.$$

In order to close this system, we will compute the asymptotic expansion of f. We will thus be able to express relations between $\underline{\underline{p}}$ (pressure tensor), r (friction force), q (heat flux), and the macroscopic quantities defined by the equation (2).

We denote by M_{ε_1} the Maxwellian with the same moments as f,

$$M_{\varepsilon_1} = \frac{n}{(2\pi T)^{3/2}}\exp\left(-\frac{|\underline{v}-\varepsilon_1\underline{u}|^2}{2T}\right).$$

We can write

$$Q_{e,e}(f,f) = M_{\varepsilon_1}\Gamma_{e,e}(M_{\varepsilon_1}^{-1}(f-M_{\varepsilon_1})) + Q_{e,e}(f-M_{\varepsilon_1}, f-M_{\varepsilon_1}),$$

where $\Gamma_{e,e}$ is the linearized operator of the collision term $Q_{e,e}$ around the Maxwellian M_{ε_1}.

We can rewrite the equation (BTE) in the following form:

$$\begin{aligned} &M_{\varepsilon_1}\Gamma_{e,e}(M_{\varepsilon_1}^{-1}(f-M_{\varepsilon_1})) + Q_{e,e}(f-M_{\varepsilon_1}, f-M_{\varepsilon_1}) \\ &\quad = -\varepsilon_2 Q_{\text{elas}}(f) + \varepsilon_1\varepsilon_2\underline{v}.\nabla_x f + \varepsilon_1\varepsilon_2\nabla_x\Phi.\nabla_v f - \varepsilon_1^2\varepsilon_2 Q_{\text{ine}}(f) + \varepsilon_1^2\varepsilon_2\partial_t f. \end{aligned} \tag{16}$$

We can also expand M_{ε_1} in ε_1:

$$\begin{aligned} M_{\varepsilon_1} = n.\mathcal{M}_0 + \varepsilon_1\,\frac{\underline{v}.\underline{u}}{T}\,n.\mathcal{M}_0 + \varepsilon_1^2\,\frac{1}{2}\left(\frac{(\underline{u}.\underline{v})^2}{T^2} - \frac{|\underline{u}|^2}{T}\right)n.\mathcal{M}_0 \\ + \varepsilon_1^3\,\frac{1}{6}\,\underline{u}.\underline{v}\left(\frac{(\underline{u}.\underline{v})^2}{T^3} - 3\,\frac{|\underline{u}|^2}{T^2}\right)n.\mathcal{M}_0 + O(\varepsilon_1^4), \end{aligned} \tag{17}$$

with

$$.\mathcal{M}_0 = \frac{1}{(2\pi T)^{3/2}}\exp\left(-\frac{|\underline{v}|^2}{2T}\right).$$

It follows

$$f = n\mathscr{M}_0 + \varepsilon_1 \frac{\underline{v}.\underline{u}}{T} n\mathscr{M}_0 + O(\varepsilon_1^2) + O(\varepsilon_1\varepsilon_2), \tag{18}$$

which gives

$$\varepsilon_2 Q_{\text{elas}}(f\ ) = \frac{\varepsilon_1\varepsilon_2}{T} n\underline{u}.Q_{\text{elas}}(\underline{v}.\mathscr{M}_0) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2), \tag{19}$$

$$\varepsilon_1\varepsilon_2\underline{v}.\nabla_x f = \varepsilon_1\varepsilon_2\underline{v}.\nabla_x(n.\mathscr{M}_0) + O(\varepsilon_1^2\varepsilon_2), \tag{20}$$

$$\varepsilon_1\varepsilon_2\nabla_x\Phi.\nabla_v f = -\frac{\varepsilon_1\varepsilon_2}{T} n.\mathscr{M}_0\underline{v}.\nabla_x\Phi + O(\varepsilon_1^2\varepsilon_2). \tag{21}$$

We further manipulate the equation (20) in order to obtain an expression depending only on the temperature gradient.

On the one hand, we have

$$\nabla_x(n.\mathscr{M}_0) = \frac{\mathscr{M}_0}{T}\nabla_x(nT) + \frac{n.\mathscr{M}_0}{T}\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\nabla_x T.$$

On the other hand, thanks to the asymptotical expansion (18), we obtain

$$\underline{\underline{p}} = nT\underline{\underline{\mathrm{I}_d}},$$

$$\underline{r} = \frac{n}{T}\left(\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\text{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v}\right)\underline{u} + O(\varepsilon_1) + O(\varepsilon_2),$$

which gives, thanks to Eq. (14),

$$\nabla_x(nT) = \frac{n}{T}\left(\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\text{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v}\right)\underline{u} + n\nabla_x\Phi + O(\varepsilon_1) + O(\varepsilon_2).$$

The operator B_{elas} being symmetric, matrix

$$\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\text{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v},$$

is diagonal and its diagonal term is defined by

$$-\int_{\mathbb{R}^3}\int_{\mathscr{S}^2} B_{\text{elas}}(|\underline{v}|,\alpha)\frac{|\underline{v}|^2}{3}\mathscr{M}_0\, d\underline{\omega}\, d\underline{v} = -\int_{\mathbb{R}^3} B^T_{\text{elas}}(|\underline{v}|)\frac{|\underline{v}|^2}{3}\mathscr{M}_0\, d\underline{v}.$$

We now define $\nu_1(T)$ by

$$\nu_1(T)\underline{\underline{\mathrm{I}_d}} = -\frac{1}{T}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\text{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v} = \frac{1}{T}\int_{\mathbb{R}^3} B^T_{\text{elas}}(|\underline{v}|)\frac{|\underline{v}|^2}{3}\mathscr{M}_0\, d\underline{v}\ \underline{\underline{\mathrm{I}_d}},$$

hence

$$\nabla_x(nT) = -n\nu_1(T)\underline{u} + n\nabla_x\Phi + O(\varepsilon_1) + O(\varepsilon_2).$$

All this together yields

$$\varepsilon_1\varepsilon_2 \underline{v}.\nabla_x f = \frac{\varepsilon_1\varepsilon_2}{T}\mathscr{M}_0 \underline{v}._x(nT) + \frac{\varepsilon_1\varepsilon_2}{T} n\mathscr{M}_0\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}.\nabla_x T + O(\varepsilon_1^2\varepsilon_2)$$

$$= -\frac{\varepsilon_1\varepsilon_2}{T} n\mathscr{M}_0 \nu_1(T)\underline{v}.\underline{u} + \frac{\varepsilon_1\varepsilon_2}{T} n\mathscr{M}_0 \underline{v}.\nabla_x \Phi + \frac{\varepsilon_1\varepsilon_2}{T} n\mathscr{M}_0\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}.\nabla_x T$$

$$+ O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2).$$

Therefore we have the following equality:

$$-\varepsilon_2 Q_{\text{elas}}(f) + \varepsilon_1\varepsilon_2 \underline{v}.\nabla_x f + \varepsilon_1\varepsilon_2 \nabla_x \Phi.\nabla_v f - \varepsilon_1^2\varepsilon_2 Q_{\text{ine}}(f) + \varepsilon_1^2\varepsilon_2 \partial_t f$$

$$= -\frac{\varepsilon_1\varepsilon_2}{T} n\underline{u}.(Q_{\text{elas}}(\underline{v}\mathscr{M}_0) + \nu_1(T)\underline{v}\mathscr{M}_0) + \frac{\varepsilon_1\varepsilon_2}{T} n\nabla_x T.\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}\mathscr{M}_0$$

$$+ O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2),$$

which gives

$$\Gamma_{e,e}(M_{\varepsilon_1}^{-1}(f - M_{\varepsilon_1})) = -M_{\varepsilon_1}^{-1}\frac{\varepsilon_1\varepsilon_2}{T} n\underline{u}.(Q_{\text{elas}}(\underline{v}\mathscr{M}_0) + \nu_1(T)\underline{v}\mathscr{M}_0)$$

$$+ M_{\varepsilon_1}^{-1}\frac{\varepsilon_1\varepsilon_2}{T} n\nabla_x T.\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}\mathscr{M}_0 + O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2).$$

We now study more precisely the linearized operator $\Gamma_{e,e}$. Moreover, we introduce $\widetilde{\Gamma}_{e,e}$, the linearized operator around $\mathscr{M}_0$:

$$\widetilde{\Gamma}_{e,e}(f) = 2\mathscr{M}_0 Q_{e,e}(\mathscr{M}_0, \mathscr{M}_0 f).$$

The operator $\widetilde{\Gamma}_{e,e}$ possesses the properties listed in Proposition 2. In the following, we use the notation $\langle f, g\rangle_0$ for the scalar product in the space $L^2(\mathscr{M}_0 d\underline{v})$.

We have the following two lemmas.

Lemma 4: There exist two functions $\underline{\Upsilon}_1$ *and* $\underline{\Upsilon}_2$ *such that* $M_{\varepsilon_1}^{-1}\underline{\Upsilon}_i$ *belongs to* $(\text{Ker}\ (\Gamma_{e,e}))^\perp$ *and*

$$\Gamma_{e,e}(M_{\varepsilon_1}^{-1}\underline{\Upsilon}_1) = -nM_{\varepsilon_1}^{-1}(Q_{\text{elas}}(\underline{v}\mathscr{M}_0) + \nu_1(T)\underline{v}\mathscr{M}_0), \tag{22}$$

$$\Gamma_{e,e}(M_{\varepsilon_1}^{-1}\underline{\Upsilon}_2) = -nM_{\varepsilon_1}^{-1}\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}\mathscr{M}_0. \tag{23}$$

Lemma 5: For the order one in ε_1, *the functions* $\underline{\Upsilon}_i$ *also satisfy*

$$\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{\Upsilon}_1) = (B_{\text{elas}}^T(|\underline{v}|) - \nu_1(T))\underline{v}, \tag{24}$$

$$\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{\Upsilon}_2) = -\left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right)\underline{v}. \tag{25}$$

Moreover, they have the following form:

$$\underline{\Upsilon}_1 = a(|\underline{v}|, T)\underline{v} + O(\varepsilon_1), \qquad \underline{\Upsilon}_2 = b(|\underline{v}|, T)\underline{v} + O(\varepsilon_1). \tag{26}$$

Before proving these lemmas, let us recall the four equalities

$$\int_{\mathbb{R}^3}\mathscr{M}_0\, d\underline{v} = 1, \quad \int_{\mathbb{R}^3}\underline{v}\otimes\underline{v}\mathscr{M}_0\, d\underline{v} = T\underline{\underline{\mathrm{I}}}_{\mathrm{d}},$$

$$\int_{\mathbb{R}^3}|\underline{v}|^2\underline{v}\otimes\underline{v}\mathscr{M}_0\, d\underline{v} = 5T^2\underline{\underline{\mathrm{I}}}_{\mathrm{d}}, \quad \int_{\mathbb{R}^3}|\underline{v}|^4\underline{v}\otimes\underline{v}\mathscr{M}_0\, d\underline{v} = 35T^3\underline{\underline{\mathrm{I}}}_{\mathrm{d}}. \tag{27}$$

Proof of Lemma 4: The proof is a direct application of Proposition 2. We have to prove that the right-hand sides of (22) and (23) belong to

$$(\mathrm{Ker}(\Gamma_{e,e}))^{\perp} \cap L^2(M_{\varepsilon_1} d\underline{v}).$$

The rhs of (22) and (23) belong obviously to $L^2(M_{\varepsilon_1} d\underline{v})$. Thus we only need the proof of the following points:

$$\int_{\mathbb{R}^3} (Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0) + \nu_1(T)\underline{v}.\mathscr{M}_0)\, d\underline{v} = 0,$$

(i)

$$\int_{\mathbb{R}^3} \left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right) \underline{v}.\mathscr{M}_0\, d\underline{v} = 0,$$

$$\int_{\mathbb{R}^3} \underline{v} \otimes (Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0) + \nu_1(T)\underline{v}.\mathscr{M}_0)\, d\underline{v} = 0,$$

(ii)

$$\int_{\mathbb{R}^3} \underline{v} \otimes \left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right) \underline{v}.\mathscr{M}_0\, d\underline{v} = 0,$$

$$\int_{\mathbb{R}^3} |\underline{v}|^2 (Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0 + \nu_1(T)\underline{v}.\mathscr{M}_0)\, d\underline{v} = 0,$$

(iii)

$$\int_{\mathbb{R}^3} |\underline{v}|^2 \left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right) \underline{v}.\mathscr{M}_0\, d\underline{v} = 0.$$

The proofs of the points (i) and (iii) are straightforward because the arguments are odd. For the point (ii), thanks to equalities (27) and the definition of ν_1, we have

$$\int_{\mathbb{R}^3} \underline{v} \otimes Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0) d\underline{v} + \nu_1(T) \int_{\mathbb{R}^3} \underline{v} \otimes \underline{v}.\mathscr{M}_0\, d\underline{v} = -T\nu_1(T)\underline{\underline{\mathrm{I}_d}} + \nu_1(T) T \underline{\underline{\mathrm{I}_d}} = 0,$$

and

$$\int_{\mathbb{R}^3} \left(\frac{|\underline{v}|^2}{2T} - \frac{5}{2}\right) \underline{v} \otimes \underline{v}.\mathscr{M}_0\, d\underline{v} = \frac{1}{2T} 5T^2 \underline{\underline{\mathrm{I}_d}} - \frac{5}{2} T \underline{\underline{\mathrm{I}_d}} = 0. \qquad \square$$

Proof of Lemma 5. First we have

$$\Gamma_{e,e}(f) = 2M_0^{-1} Q_{e,e}(M_0, M_0 f) = 2n.\mathscr{M}_0^{-1} Q_{e,e}(\mathscr{M}_0, \mathscr{M}_0 f) + O(\varepsilon_1) = n\widetilde{\Gamma}_{e,e}(f) + O(\varepsilon_1).$$

Moreover,

$$Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0) = -B^T_{\mathrm{elas}}(|\underline{v}|)\underline{v}.\mathscr{M}_0.$$

By a simple computation we obtain Eqs. (24) and (25). They are well defined because their right-hand sides belong to $(\mathrm{Ker}\ (\widetilde{\Gamma}_{e,e}))^{\perp} \cap L^2(\mathscr{M}_0 d\underline{v})$.

The second part of the lemma is due to symmetry and rotational invariance properties of the linearized operator $\widetilde{\Gamma}_{e,e}$ (see refs. 8 and 16). $\square$

Finally, we have the following asymptotic expansion for f:

$$f = M_{\varepsilon_1} + \frac{\varepsilon_1\varepsilon_2}{T}\,\underline{u}.\underline{\Upsilon}_1 - \frac{\varepsilon_1\varepsilon_2}{T}\,\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2 + O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2)$$
$$= n.\mathscr{M}_0 + \varepsilon_1\,\frac{\underline{v}.\underline{u}}{T}\,n.\mathscr{M}_0 + \varepsilon_1^2\,\frac{1}{2}\left(\frac{(\underline{u}.\underline{v})^2}{T^2} - \frac{|\underline{u}|^2}{T}\right)n.\mathscr{M}_0 + \varepsilon_1^3\,\frac{1}{6}\,\underline{u}.\underline{v}\left(\frac{(\underline{u}.\underline{v})^2}{T^3} - 3\,\frac{|\underline{u}|^2}{T^2}\right)n.\mathscr{M}_0$$
$$+ \frac{\varepsilon_1\varepsilon_2}{T}\,\underline{u}.\underline{\Upsilon}_1 - \frac{\varepsilon_1\varepsilon_2}{T}\,\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2 + O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2). \tag{28}$$

Thus, we can now compute the approximation of the macroscopic quantities $\underline{\underline{p}}$, $\underline{r}$, and $\underline{q}$:

$$\underline{\underline{p}} = nT\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_1^4), \tag{29}$$

$$\underline{r} = -\left(\nu_1(T) + \frac{\varepsilon_1}{\tau_{\mathrm{ine}}} + \varepsilon_1^2\nu_2(T,\underline{u}) + \frac{\varepsilon_2}{n}\,\nu_3(T)\right)n\underline{u} - \varepsilon_2\,\mathrm{div}(\nu_4(T)\underline{\underline{\mathrm{I}_d}}) + O(\varepsilon_1^3) + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_2^2), \tag{30}$$

$$\underline{q} = \varepsilon_2\nu_5(T)\underline{u} - \varepsilon_2\nu_6(T)\boldsymbol{\nabla}_x T + O(\varepsilon_1^3) + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_2^2), \tag{31}$$

with

$$\nu_2(T,\underline{u})\underline{\underline{\mathrm{I}_d}} = -\frac{1}{6T^3}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\mathrm{elas}}((\underline{u}.\underline{v})^2\underline{v}.\mathscr{M}_0)d\underline{v} + \frac{|\underline{u}|^2}{2T^2}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\mathrm{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v}$$
$$= \frac{1}{6T^3}\int_{\mathbb{R}^3}B^T_{\mathrm{elas}}(|\underline{v}|)(\underline{u}.\underline{v})^2\,\frac{|\underline{v}|^2}{3}\,\mathscr{M}_0\,d\underline{v}\;\underline{\underline{\mathrm{I}_d}} - \frac{|\underline{u}|^2}{2T}\,\nu_1(T)\underline{\underline{\mathrm{I}_d}}, \tag{32}$$

$$\nu_3(T)\underline{\underline{\mathrm{I}_d}} = -\frac{1}{T}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\mathrm{elas}}(\underline{\Upsilon}_1)d\underline{v} = \frac{1}{T}\int_{\mathbb{R}^3}B^T_{\mathrm{elas}}(|\underline{v}|)\,\frac{|\underline{v}|^2}{3}\,a(|\underline{v}|,T)d\underline{v}\;\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1), \tag{33}$$

$$\partial_T\nu_4(T)\underline{\underline{\mathrm{I}_d}} = \frac{1}{T}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\mathrm{elas}}(\underline{\Upsilon}_2)d\underline{v} = -\frac{1}{T}\int_{\mathbb{R}^3}B^T_{\mathrm{elas}}(|\underline{v}|)\,\frac{|\underline{v}|^2}{3}\,b(|\underline{v}|,T)d\underline{v}\;\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1), \tag{34}$$

$$\nu_5(T)\underline{\underline{\mathrm{I}_d}} = \frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^2}{2}\,\underline{v}\otimes\underline{\Upsilon}_1\,d\underline{v} = \frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^4}{6}\,a(|\underline{v}|,T)d\underline{v}\;\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1), \tag{35}$$

$$\nu_6(T)\underline{\underline{\mathrm{I}_d}} = \frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^2}{2}\,\underline{v}\otimes\underline{\Upsilon}_2\,d\underline{v} = \frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^4}{6}\,b(|\underline{v}|,T)d\underline{v}\;\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1) = \kappa(T)\underline{\underline{\mathrm{I}_d}} + O(\varepsilon_1). \tag{36}$$

Therefore we rewrite the moments system in the following conservative form:

$$\partial_t n + \mathrm{div}(n\underline{u}) = 0, \tag{37}$$

$$\partial_t(\varepsilon_1^2 n\underline{u}) + \mathrm{div}(\varepsilon_1^2 n\underline{u}\otimes\underline{u} + nT\underline{\underline{\mathrm{I}_d}} + \varepsilon_2\nu_4(T)\underline{\underline{\mathrm{I}_d}}) - n\boldsymbol{\nabla}_x\Phi$$
$$+\left(\nu_1(T) + \frac{\varepsilon_1}{\tau_{\mathrm{ine}}} + \varepsilon_1^2\nu_2(T,\underline{u}) + \frac{\varepsilon_2}{n}\,\nu_3(T)\right)n\underline{u}$$
$$= O(\varepsilon_1^3) + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_2^2), \tag{38}$$

$$\partial_t\Big(\tfrac{1}{2}\,n(\varepsilon_1^2|\underline{u}|^2+3T)\Big)+\mathrm{div}\Big(\tfrac{1}{2}\,n\underline{u}(\varepsilon_1^2|\underline{u}|^2+5T)+\varepsilon_2\nu_5(T)\underline{u}-\varepsilon_2\kappa(T)\nabla_x T\Big)$$

$$-n\underline{u}\nabla_x\Phi+n\,\frac{e-\frac{3}{2}T_0}{\tau_e}=O(\varepsilon_1^3)+O(\varepsilon_1\varepsilon_2)+O(\varepsilon_2^2). \tag{39}$$

IV. ENTROPY FUNCTION AND THE SYMMETRIZABLE SYSTEM

In this section we prove that the macroscopic entropy balance law (EL) gives a generalized entropy function of the hyperbolic–parabolic system (HD) in the sense of S. Kawashima and Y. Shizuta.[17] Moreover, using this entropy function, we derive a conservative symmetric form of the system.

Remark: At the kinetic level, the entropy is $H(f)=f\ln f$ (Theorems 1 and 3). We obtain equation (EL) by computing the asymptotic expansion of

$$\partial_t\int_{\mathbb{R}^3}H(f)\,d\underline{v}+\frac{1}{\varepsilon_1}\,\mathrm{div}_x\Big(\int_{\mathbb{R}^3}\underline{v}H(f)\,d\underline{v}\Big)+\frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}\nabla_x\Phi.\nabla_v H(f)\,d\underline{v}$$
$$=\int_{\mathbb{R}^3}(1+\ln f)\Big(\frac{1}{\varepsilon_1^2}\Big(\frac{1}{\varepsilon_2}\,Q_{e,e}(f)+Q_{\mathrm{elas}}(f)\Big)-Q_{\mathrm{ine}}(f)\Big)d\underline{v}. \tag{40}$$

This computation is presented in Appendix A.

The relation between the conservative symmetric formulation and the entropy function was first studied for hyperbolic systems of the form

$$\partial_t\underline{w}+\sum_{j=1}^{n}\partial_j\underline{g}^j(\underline{w})=0, \tag{41}$$

where $\underline{g}^j$, $j=1,...,n$, and $\underline{w}$ denote m-dimensional vectors, with $\underline{g}^j$ smooth functions of $\underline{w}$. It was shown that the existence of an entropy function for (41) is characterized by the property that (41) can be symmetrized by introducing a new dependent variable (see Refs. 18–21).

These results were generalized by S. Kawashima and Y. Shizuta[17] to the case of second-order quasilinear systems of the form

$$\partial_t\underline{w}+\sum_{j=1}^{n}\partial_j\underline{g}^j(\underline{w})=\sum_{i,j=1}^{n}\partial_i(\underline{\underline{G}}^{i,j}(\underline{w})\partial_j\underline{w})+\underline{S}(\underline{w}), \tag{42}$$

where $\underline{\underline{G}}^{ij}(\underline{w}),i,j=1,...,n$, denote $m\times m$ matrices depending smoothly on $\underline{w}$. Both $\underline{g}^j(\underline{w})$ and $\underline{\underline{G}}^{ij}(\underline{w})$ are defined on an open set $\mathcal{O}\subset\mathbb{R}$.

The following definition of an entropy function was given by Kawashima.[17,22]

Definition 6: Let $\eta=\eta(\underline{w})$ be a real-valued smooth function defined on a convex open set $\mathcal{O}_w\subset\mathcal{O}$. Then η is called an entropy function for (42), if the following properties hold:

(i) *The function η is a strictly convex function on $\mathcal{O}_{\underline{w}}$ in the sense that the Hessian matrix is positive definite on $\mathcal{O}_{\underline{w}}$.*

(ii) *There exist real-valued smooth functions $\sigma^j=\sigma^j(\underline{w})$, $j=1,...,n$, such that*

$$\partial_{\underline{w}}\eta\partial_{\underline{w}}g^j=\partial_{\underline{w}}\sigma^j,\quad j=1,...,n,$$

for $\underline{w}\in\mathcal{O}_w$. These functions are called the entropy fluxes.

(iii) *We have the property*

$$(\partial_{\underline{w}}^2\eta)^{-1}\,{}^t\underline{\underline{G}}^{i,j}=\underline{\underline{G}}^{j,i}(\partial_{\underline{w}}^2\eta)^{-1},$$

for $\underline{w}\in\mathcal{O}_{\underline{w}}$.

(iv) *The matrix $\Sigma_{i,j=1}^{n}\underline{\underline{G}}^{i,j}(\underline{w})(\partial_{\underline{w}}^2\eta)^{-1}\boldsymbol{\omega}_i\boldsymbol{\omega}_j$ is symmetric and positive semi-definite for $\underline{w}\in\mathcal{O}_w$ and $\boldsymbol{\omega}\in\mathcal{S}^{n-1}$.*

In order to define the symmetric form of system (42), we introduce a new dependent variable $\underline{\psi}$ instead of $\underline{w}$. It is assumed that there is a diffeomorphism from $\mathcal{O}_{\underline{\psi}}$ onto $\mathcal{O}_{\underline{w}}$. We denote $\underline{\psi} = \underline{\psi}(\underline{w})$. Then (42) assumes the following form:

$$\underline{\underline{A}}^{(0)}(\underline{\psi})\partial_t\underline{\psi} + \sum_{j=1}^{n} \underline{\underline{A}}^{ij}(\underline{\psi})\partial_j\underline{\psi} = \sum_{i,j=1}^{n} \partial_i(\underline{\underline{B}}^{i,j}(\underline{\psi})\partial_j\underline{\psi}) + \underline{\tilde{S}}(\underline{\psi}), \tag{43}$$

with

$$\underline{\underline{A}}^{0}(\underline{\psi}) = \partial_{\underline{\psi}}\underline{w}(\underline{\psi}), \qquad \underline{\underline{A}}^{j}(\underline{\psi}) = \partial_{\underline{\psi}} g^j(\underline{w}(\underline{\psi})) = (\partial_{\underline{w}} f^j)(\underline{w}(\underline{\psi}))\underline{\underline{A}}^{0}(\underline{\psi}),$$

$$\underline{\underline{B}}^{i,j}(\underline{\psi}) = \underline{\underline{G}}^{i,j}(\underline{w}(\underline{\psi}))\underline{\underline{A}}^{0}(\underline{\psi}).$$

The following definition of a symmetric form is due to Ref. 17.

Definition 7: The system (43) is said to be of the symmetric form if the coefficient matrices satisfy the following properties:

(i) *The matrix $\underline{\underline{A}}^{0}(\underline{\psi})$ is symmetric and positive definite for $\underline{\psi} \in \mathcal{O}_{\underline{\psi}}$.*
(ii) *The matrices $\underline{\underline{A}}^{j}(\underline{\psi}), j = 1,...,n$, are symmetric for $\underline{\psi} \in \mathcal{O}_{\underline{\psi}}$.*
(iii) *We have ${}^t(\underline{\underline{B}}^{i,j}(\underline{\psi})) = \underline{\underline{B}}^{j,i}(\underline{\psi}), i,j = 1,...,n$, for $\underline{\psi} \in \mathcal{O}_{\underline{\psi}}$.*
(iv) *The matrix $\Sigma_{i,j=1}^{n} \underline{\underline{B}}^{i,j}(\underline{\psi})\boldsymbol{\omega}_i\boldsymbol{\omega}_j$ is symmetric and positive semi-definite, for $\underline{w} \in \mathcal{O}_{\underline{w}}$ and $\underline{\omega} \in \mathcal{S}^{n-1}$.*

S. Kawashima and Y. Shizuta established the following theorem.

Theorem 8: *For the second-order quasi-linear system (42) the two following properties are equivalent:*

(i) *The system (42) has an entropy function η defined on a convex open set $\mathcal{O}_{\underline{w}} \subset \mathcal{O}$.*
(ii) *The system (42) can be symmetrized, using the symmetrizing variable $\underline{\psi}$, on a convex open set $\mathcal{O}_{\underline{w}} \subset \mathcal{O}$.*

Moreover, the symmetrizing variable $\underline{\psi}$ can be expressed in terms of the gradient of the entropy function η as

$$\underline{\psi} = {}^t(\partial_{\underline{w}}\eta). \tag{44}$$

We now consider the system (HD) and apply the previous results. It can be written in the following form:

$$\partial_t\underline{w} + \sum_{j=1}^{n} \partial_j\underline{g}^j(\underline{w}) = \sum_{j=1}^{n} \partial_j(\underline{\underline{G}}^{i,j}(\underline{w})\partial_j\underline{w}) + \underline{S}(\underline{w}), \tag{45}$$

with

$$\underline{w} = \begin{pmatrix} n \\ \varepsilon_1^2 n u_1 \\ \varepsilon_1^2 n u_2 \\ \varepsilon_1^2 n u_3 \\ \frac{1}{2} n(\varepsilon_1^2 |\underline{u}|^2 + 3T) \end{pmatrix},$$

$$\underline{g}^j = \begin{pmatrix} n u_j \\ \varepsilon_1^2 n u_1 u_j + nT\delta_{1,j} + \varepsilon_2 \nu_4(T)\delta_{1,j} \\ \varepsilon_1^2 n u_2 u_j + nT\delta_{2,j} + \varepsilon_2 \nu_4(T)\delta_{2,j} \\ \varepsilon_1^2 n u_3 u_j + nT\delta_{3,j} + \varepsilon_2 \nu_4(T)\delta_{3,j} \\ \frac{1}{2} n u_j(\varepsilon_1^2 |\underline{u}|^2 + 5T) + \varepsilon_2 \nu_5(T) u_j \end{pmatrix},$$

$$\underline{\underline{G}}^{i,j}=\underline{\underline{G}}\,\delta_{i,j},$$

$$\underline{\underline{G}}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\dfrac{\kappa(T)T}{n}+\varepsilon_1^2\dfrac{1}{3}\dfrac{\kappa(T)}{n}|\underline{u}|^2 & -\dfrac{2}{3}\dfrac{\kappa(T)}{n}u_1 & -\dfrac{2}{3}\dfrac{\kappa(T)}{n}u_2 & -\dfrac{2}{3}\dfrac{\kappa(T)}{n}u_3 & \dfrac{2}{3}\dfrac{\kappa(T)}{n} \end{pmatrix},$$

$$\underline{S}=\begin{pmatrix} 0 \\ n\partial_{x_1}\Phi-\left(\nu_1(T)+\dfrac{\varepsilon_1}{\tau_{\text{ine}}}+\varepsilon_1^2\nu_2(T,\underline{4})+\dfrac{\varepsilon_2}{n}\nu_3(T)\right)nu_1 \\ n\partial_{x_2}\Phi-\left(\nu_1(T)+\dfrac{\varepsilon_1}{\tau_{\text{ine}}}+\varepsilon_1^2\nu_2(T,\underline{u})+\dfrac{\varepsilon_2}{n}\nu_3(T)\right)nu_2 \\ n\partial_{x_3}\Phi-\left(\nu_1(T)+\dfrac{\varepsilon_1}{\tau_{\text{ine}}}+\varepsilon_1^2\nu_2(T,\underline{u})+\dfrac{\varepsilon_2}{n}\nu_3(T)\right)nu_3 \end{pmatrix}.$$

Hence we have the following.

Theorem 9: (i) *The hydrodynamical system (45) has an entropy function*

$$\eta=n\,\ln\left(\frac{n}{(2\pi T)^{3/2}}\right)-\frac{5}{2}\,n. \tag{46}$$

(ii) *The associated entropy fluxes are*

$$\sigma^j=nu_j\,\ln\left(\frac{n}{(2\pi T)^{3/2}}\right)-\frac{5}{2}\,nu_j-\frac{\varepsilon_2}{T}\,\nu_5(T)u_j,\quad j=1,\ldots,n. \tag{47}$$

(iii) *The system (45) can be symmetrized using the following symmetrizing variables:*

$$\underline{\psi}=\begin{pmatrix} \ln\left(\dfrac{n}{(2\pi T)^{3/2}}\right)-\varepsilon_1^2\dfrac{|\underline{u}|^2}{2T} \\ \dfrac{u_1}{T} \\ \dfrac{u_2}{T} \\ \dfrac{u_3}{T} \\ -\dfrac{1}{T} \end{pmatrix}. \tag{48}$$

Remark: One can obtain the entropy balance law (EL) by multiplying each term of (HD) by $\underline{\psi}$ and then adding the results.

Before giving the proof of the theorem, we need the following lemma.

Lemma 10: (i) $\nu_5(T)=T\partial_T\nu_4(T)+O(\varepsilon_1)$ and (ii) $\kappa(T)>0$ *for all* $T>0$.

The proof of this lemma is given in Appendix B.

Proof of Theorem 9: We will prove points (i) and (ii). Point (iii) is a simple consequence of Theorem 8.

The strict convexity of η is a classical result of the Euler or Navier–Stokes equations theory.[17,23] Indeed we have exactly the same entropy function.

Let us write $\underline{g}^j$ in two parts:

$$\underline{g}^j=\underline{g}_a^j+\varepsilon_2\underline{g}_b^j=\begin{pmatrix} nu_j \\ \varepsilon_1^2 nu_1u_j+nT\delta_{1,j} \\ \varepsilon_1^2 nu_2u_j+nT\delta_{2,j} \\ \varepsilon_1^2 nu_3u_j+nT\delta_{3,j} \\ \frac{1}{2}nu_j(\varepsilon_1^2|\underline{u}|^2+5T) \end{pmatrix}+\varepsilon_2\begin{pmatrix} 0 \\ \nu_4(T)\delta_{1,j} \\ \nu_4(T)\delta_{2,j} \\ \nu_4(T)\delta_{3,j} \\ \nu_5(T)u_j \end{pmatrix}.$$

For one part, we know that (Euler entropy function)

$$\partial_{\underline{w}}\eta\partial_{\underline{w}}\underline{g}_a^j=\partial_{\underline{w}}\left(nu_j\ \ln\left(\frac{n}{(2\pi T)^{3/2}}\right)-\frac{5}{2}\,nu_j\right).$$

For the other part, we have for all $l=1,\ldots,5$,

$$\begin{aligned}\partial_{w_l}\left(-\frac{1}{T}\,\nu_5(T)u_j\right)&=\nu_5(T)u_j\partial_{w_l}\left(-\frac{1}{T}\right)-\frac{1}{T}\,\partial_{w_l}(\nu_5(T)u_j)\\&=T\partial_T\nu_4(T)u_j\,\frac{1}{T^2}\,\partial_{w_l}T-\frac{1}{T}\,\partial_{w_l}(\nu_5(T)u_j)\\&=\frac{u_j}{T}\,\partial_{w_l}(\nu_4(T))-\frac{1}{T}\,\partial_{w_l}(\nu_5(T)u_j)\\&=\sum_{i=1}^{3}\frac{u_i}{T}\,\partial_{w_l}(\nu_4(T)\delta_{i,j})-\frac{1}{T}\,\partial_{w_l}(\nu_5(T)u_j)\\&=\partial_{\underline{w}}\eta\partial_{w_l}\underline{g}_b^j,\end{aligned}$$

which gives

$$\partial_{\underline{w}}\left(-\frac{1}{T}\,\nu_5(T)u_j\right)=\partial_{\underline{w}}\eta\partial_{\underline{w}}\underline{g}_b^j.$$

This ends the proof of points (i) and (ii) of the entropy definition 6. In our case, in order to obtain points (iii) and (iv), we only need to prove that the matrix

$$\partial_{\underline{w}}^2\eta\underline{\underline{G}}$$

is symmetric and positive semi-definite. With straightforward computations we obtain the symmetry of the matrix and that the positivity condition is only linked to the sign of κ. Finally, thanks to Lemma 10, we know that κ is positive. □

V. LINKS WITH OTHER SEMICONDUCTOR MODELS

In this part, we describe the relations between our model and others semiconductor models.

A. A generalized hydrodynamical model

First of all, the complete model appears like a generalization of the classical hydrodynamical model.[1,2] Indeed, we have two additional terms: a ''thermal friction force'' and a ''friction heat flux.''

More precisely, the friction force $\underline{r}$ is the sum of two contributions [see Eq. (30)]. The first one expresses the momentum relaxation with the relaxation time τ_m given by

$$\tau_m=\frac{n}{n\nu_1(T)+\varepsilon_1 n/\tau_{\text{ine}}+\varepsilon_1^2 n\nu_2(T,\underline{u})+\varepsilon_2\nu_3(T)}.$$

The second contribution is proportional to the gradient of the electron temperature:

$$\mathrm{div}(\varepsilon_2 \nu_4(T) \underline{\underline{\mathrm{I_d}}}) = \varepsilon_2 \frac{\nu_5(T)}{T} \nabla_x T.$$

This last expression is absent from the classical hydrodynamical semiconductor models. Let us call it the ''thermal friction force.''

In the same way, the heat flux $\underline{q}$ [see Eq. (31)] is the sum of a thermal heat flux associated with the heat conductivity coefficient κ and a term proportional to the velocity $\underline{u}$:

$$\varepsilon_2 \nu_5(T) \underline{u}.$$

This term, called the ''friction heat flux,'' is also new in the field of hydrodynamical models for semiconductors.

On the other hand, these two terms appear in the energy transport models.[5,9] They are linked with the Peltier and Soret effects.

Finally, let us remark that if we neglect the ε_2 terms, we obtain a hydrodynamical model without heat flux term. In fact we get a Euler–Poisson system (see Refs. 24 and 25 and the references therein).

B. Reduction to an energy-transport model

If we neglect the ε_1 terms, we obtain an energy-transport model:[3–5]

$$\partial_t n - \mathrm{div}\ \underline{J}_n = 0, \qquad \partial_t \frac{3}{2} nT + \mathrm{div}\ \underline{S} + \underline{J}_n . \nabla_x \Phi = \frac{3}{2} n \frac{T - T_0}{\tau_e}, \tag{49}$$

where the electron current density $\underline{J}_n$ and the thermal energy flow density are defined by

$$\begin{aligned} \underline{J}_n &= \mu(n,T)\left(\nabla_x nT + \frac{\varepsilon_2}{T} \nu_5(T) \nabla_x T - n \nabla_x \Phi \right), \\ \underline{S} &= -\left(\frac{5}{2} T + \frac{\varepsilon_2}{n} \nu_5(T) \right) \underline{J}_n - \varepsilon_2 \kappa(T) \nabla_x T, \end{aligned} \tag{50}$$

and $\mu(n,T)$, the electronic mobility, is defined by

$$\mu(n,T) = \frac{n}{n \nu_1(T) + \varepsilon_2 \nu_3(T)}.$$

Remark: We assume that, for ε_2 small enough, we have

$$\nu_1(T) + \frac{\varepsilon_2}{n} \nu_3(T) > 0 \quad \text{for all} \quad T > 0.$$

We will now study the entropy function and the symmetric form of this model.

First of all, we write the system in a conservative form by introducing a new flux variable $\underline{J}_T$ (we suppose that the electrostatic potential does not depend on time). Then we have

$$\partial_t n - \mathrm{div}\ \underline{J}_n = 0, \qquad \partial_t \left(\frac{3}{2} nT - n\Phi \right) - \mathrm{div}\ \underline{J}_T = \frac{3}{2} n \frac{T - T_0}{\tau_e}, \tag{51}$$

with

$$\underline{J}_T = -\underline{S} - \Phi \underline{J}_n .$$

In order to obtain a symmetric formulation, let us introduce the electron quasi-Fermi level φ_n defined, in the nondegenerate case, by a Boltzmann statistic

$$n=(2\pi T)^{3/2}\exp\left(\frac{\varphi_n+\Phi}{T}\right). \tag{52}$$

Then, the fluxes $\underline{J}_n$ and $\underline{J}_T$ have the following form:

$$\underline{J}_n=d_{1,1}\nabla_x\left(\frac{\varphi_n}{T}\right)+d_{1,2}\nabla_x\left(\frac{-1}{T}\right),\qquad \underline{J}_T=d_{2,1}\nabla_x\left(\frac{\varphi_n}{T}\right)+d_{2,2}\nabla_x\left(\frac{-1}{T}\right), \tag{53}$$

with

$$d_{1,1}=\mu nT,$$

$$d_{1,2}=d_{2,1}=\mu nT\left(\frac{5}{2}\,T+\frac{\varepsilon_2}{n}\,\nu_5(T)-\Phi\right), \tag{54}$$

$$d_{2,2}=\mu nT\left(\frac{5}{2}\,T+\frac{\varepsilon_2}{n}\,\nu_5(T)-\Phi\right)^2+\varepsilon_2 T^2\kappa(T).$$

The matrix $d_{i,j}$ is obviously symmetric positive definite. Hence we have proved the existence of a symmetric form and therefore of an entropy function (by using Kawashima and Shizuta's theorem). Let us remark that it is exactly the same entropy function as in the hydrodynamical case.

To sum up let us say that we obtain, by neglecting the ε_1 terms, an energy-transport model which can be expressed in a conservative form with the following variables:

$$\underline{w}_{\mathrm{ET}}=\begin{pmatrix} n \\ \frac{3}{2}nT-n\Phi \end{pmatrix}.$$

This model has the same entropy function as the hydrodynamical one:

$$\eta=n\,\ln\left(\frac{n}{(2\pi T)^{3/2}}\right)-\frac{5}{2}\,n.$$

With this entropy function, we obtain the following symmetric variables:

$$\underline{\psi}_{\mathrm{ET}}=\partial_{\underline{w}_{\mathrm{ET}}}\eta=\begin{pmatrix} \ln\left(\frac{n}{(2\pi T)^{3/2}}\right)-\frac{\Phi}{T} \\ -\frac{1}{T} \end{pmatrix}.$$

Moreover,we can also obtain a drift-diffusion model[26–28] by neglecting both the ε_1 and the ε_2 terms and putting an infinitely small energy relaxation time. This limit is well established in a paper by Marcati and Natalini.[29,30]

VI. CONCLUSION

In this work, we derive a generalized hydrodynamical model which gives a unified vision of the classical hydrodynamical model and the energy-transport model. We also introduce a unique entropy function for the different levels of modelization. Let us notice that this entropy function is very useful for existence and unicity results. Moreover, we also derive from this entropy function a set of symmetric variables thanks to a theorem by Kawashima and Shizuta.[17] These variables, called the entropy variables, are attractive from a numerical point of view.

The main extension of this work could be to obtain an explicit formulation of the model by computing the integral terms. In this way we would be able to make numerical comparisons with the other existing semiconductor models.

ACKNOWLEDGMENTS

The author wishes to thank B. Perthame of the LAN, UMPC, Paris for fruitful discussions and J. F. Palmier of France-Telecom/CNET for his advice and financial support.

APPENDIX A: DERIVATION OF THE ENTROPY BALANCE LAW FROM THE KINETIC ENTROPY

In this Appendix, we compute the asymptotic expansion of the equation (40) in order to obtain the entropy balance law (EL). This gives another motivation for the entropy function.

From the equation (28), we can write

$$\ln f = \ln M_{\varepsilon_1} + \frac{\varepsilon_1\varepsilon_2}{T}\,\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1 - \frac{\varepsilon_1\varepsilon_2}{T}\,\mathscr{M}_0^{-1}\nabla_x T.\underline{\Upsilon}_2 + O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2)$$

$$= \underline{\psi}..\underline{\mathscr{K}}(\underline{v}) + \frac{\varepsilon_1\varepsilon_2}{T}\,\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1 - \frac{\varepsilon_1\varepsilon_2}{T}\,\mathscr{M}_0^{-1}\nabla_x T.\underline{\Upsilon}_2 + O(\varepsilon_1^4) + O(\varepsilon_1^2\varepsilon_2) + O(\varepsilon_1\varepsilon_2^2),$$

with

$$\underline{\psi} = \begin{pmatrix} \ln\left(\dfrac{n}{(2\pi T)^{3/2}}\right) - \varepsilon_1^2\,\dfrac{|\underline{u}|^2}{2T} \\ \dfrac{u_1}{T} \\ \dfrac{u_2}{T} \\ \dfrac{u_3}{T} \\ -\dfrac{1}{T} \end{pmatrix},$$

$$\underline{\mathscr{K}}(\underline{v}) = \begin{pmatrix} 1 \\ \varepsilon_1\underline{v} \\ \dfrac{|\underline{v}|^2}{2} \end{pmatrix}.$$

Now, we will compute successively each term of (40). First, for the $\partial_t f$ term we obtain

$$\partial_t \int_{\mathbb{R}^3} H(f)\,d\underline{v} = \partial_t \int_{\mathbb{R}^3} \underline{\psi}..\underline{\mathscr{K}}(\underline{v}) M_{\varepsilon_1}\,d\underline{v} + O(\varepsilon_1^4) + O(\varepsilon_1\varepsilon_2)$$

$$= \partial_t\left(n\,\ln\left(\frac{n}{(2\pi T)^{3/2}}\right) - \frac{3}{2}\,n\right) + O(\varepsilon_1^4) + O(\varepsilon_1\varepsilon_2).$$

For the $\underline{v}.\nabla_x f$ term, we have

$$\operatorname{div}\left(\frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}\underline{v}H(f)\,d\underline{v}\right) = \operatorname{div}\left(\frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}\underline{v}\,\underline{\psi}..\underline{\mathscr{K}}(\underline{v}) f\,d\underline{v}\right) + \operatorname{div}\left(\frac{\varepsilon_2}{T}\int_{\mathbb{R}^3}\mathscr{M}_0^{-1}\underline{v}(\underline{u}.\underline{\Upsilon}_1 - \nabla_x T.\underline{\Upsilon}_2) f\,d\underline{v}\right) + O(\varepsilon_1^3) + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_2^2)$$

$$= \operatorname{div}\left(n\underline{u}\,\ln\left(\frac{n}{(2\pi T)^{3/2}}\right) - \frac{3}{2}\,n\underline{u} - \frac{\varepsilon_2}{T}\,\nu_5(T)\underline{u} + \varepsilon_2\,\frac{\kappa(T)}{T}\,\nabla_x T\right) + O(\varepsilon_1^3) + O(\varepsilon_1\varepsilon_2) + O(\varepsilon_2^2).$$

For the $\nabla_x\Phi.\boldsymbol{\nabla}_v f$ term, we have simply

$$\frac{1}{\varepsilon_1}\int_{\mathbb{R}^3}\boldsymbol{\nabla}_x\Phi.\boldsymbol{\nabla}_v H(f\,)d\underline{v}=0.$$

For the $Q_{e,e}(f,f\,)$ term, we compute

$$\frac{1}{\varepsilon_1^2\varepsilon_2}\int_{\mathbb{R}^3}(1+\ln f\,)Q_{e,e}(f,f\,)d\underline{v}$$

$$=\frac{\varepsilon_2}{T^2}\int_{\mathbb{R}^3}\mathscr{M}_0{}^{-1}(\underline{u}.\underline{\Upsilon}_1-\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2).\mathscr{M}_0\widetilde{\Gamma}_{e,e}\mathscr{M}_0^{-1}(\underline{u}.\underline{\Upsilon}_1-\boldsymbol{\nabla}_x.\underline{\Upsilon}_2)d\underline{v}$$
$$+O(\varepsilon_1^3)+O(\varepsilon_1\varepsilon_2)+O(\varepsilon_2^2)$$

$$=\frac{\varepsilon_2}{T^2}\left(\langle\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1,\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1)\rangle_0+\langle\mathscr{M}_0^{-1}\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2,\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2)\rangle_0\right.$$
$$\left.-2\langle\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1,\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2)\rangle_0\right)+O(\varepsilon_1^3)+O(\varepsilon_1\varepsilon_2)+O(\varepsilon_2^2).$$

For the $Q_{\text{elas}}(f\,)$ term we use Lemma 10 and we obtain

$$\frac{1}{\varepsilon_1^2}\int_{\mathbb{R}^3}(1+\ln f\,)Q_{\text{elas}}(f\,)d\underline{v}$$
$$=\frac{1}{\varepsilon_1^2}\int_{\mathbb{R}^3}\left(\underline{\psi}.\underline{\mathscr{K}}(\underline{v})+\frac{\varepsilon_1\varepsilon_2}{T}\mathscr{M}_0^{-1}(\underline{u}.\underline{\Upsilon}_1-\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2)+O(\varepsilon_1^4)+O(\varepsilon_1^2\varepsilon_2)+O(\varepsilon_1\varepsilon_2^2)\right)$$

$Q_{\text{elas}}(f\,)d\underline{v}$

$$=-\frac{n}{T}\left(\nu_1(T)+\varepsilon_1^2\nu_2(T,\underline{u})+\frac{\varepsilon_2}{n}\,\nu_3(T)\right)|\underline{u}|^2$$
$$-\frac{\varepsilon_2}{T}\,\underline{u}.\,\operatorname{div}(\nu_4(T)\underline{\underline{\mathrm{I}_d}})+\frac{\varepsilon_2}{T^2}\int_{\mathbb{R}^3}\mathscr{M}_0^{-1}(\underline{u}.\underline{\Upsilon}_1-\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2)n\underline{u}.Q_{\text{elas}}(\underline{v}.\mathscr{M}_0)d\underline{v}+O(\varepsilon_1^3)$$
$$+O(\varepsilon_1\varepsilon_2)+O(\varepsilon_2^2)$$
$$=-\frac{n}{T}\left(\nu_1(T)+\varepsilon_1^2\nu_2(T,\underline{u})+\frac{\varepsilon_2}{n}\,\nu_3(T)\right)|\underline{u}|^2-\frac{\varepsilon_2}{T^2}\,\nu_5(T)\underline{u}.\boldsymbol{\nabla}_x T$$
$$-\frac{\varepsilon_2}{T^2}\left(\langle\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1,\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1)\rangle_0-\langle\mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1,\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\boldsymbol{\nabla}_x T.\underline{\Upsilon}_2)\rangle_0\right)+O(\varepsilon_1^3)$$
$$+O(\varepsilon_1\varepsilon_2)+O(\varepsilon_2^2).$$

Finally, for the $Q_{\text{ine}}(f\,)$ term, we have

$$\int_{\mathbb{R}^3}(1+\ln f\,)Q_{\text{ine}}(f\,)d\underline{v}=\int_{\mathbb{R}^3}\underline{\psi}.\underline{\mathscr{K}}(\underline{v})Q_{\text{ine}}(f\,)d\underline{v}+O(\varepsilon_1^4)+O(\varepsilon_1\varepsilon_2)$$
$$=-\frac{n}{T}\,\frac{e-\frac{3}{2}T_0}{\tau_e}+O(\varepsilon_1^4)+O(\varepsilon_1\varepsilon_2).$$

Moreover, we have the following equalities:

$$\langle \mathscr{M}_0^{-1}\nabla_x T.\underline{\Upsilon}_2, \widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\nabla_x T.\underline{\Upsilon}_2)\rangle_0 = -\int_{\mathbb{R}^3} \nabla_x T.\underline{\Upsilon}_2 \nabla_x T\cdot\left(\frac{|\underline{v}|^2}{2T}-\frac{5}{2}\right)\underline{v}\ d\underline{v}$$

$$= -{}^t\nabla_x T\left(\frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^2}{2}\Upsilon_2\otimes\underline{v}\,d\underline{v}\right)\nabla_x T = -\kappa(T)|\nabla_x T|^2,$$

$$\langle \mathscr{M}_0^{-1}\underline{u}.\underline{\Upsilon}_1, \widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\nabla_x T.\underline{\Upsilon}_2)\rangle = -\int_{\mathbb{R}^3}\underline{u}.\underline{\Upsilon}_1\nabla_x T.\left(\frac{|\underline{v}|^2}{2T}-\frac{5}{2}\right)\underline{v}\ d\underline{v}$$

$$= -{}^t\underline{u}\left(\frac{1}{T}\int_{\mathbb{R}^3}\frac{|\underline{v}|^2}{2}\Upsilon_1\otimes\underline{v}\,d\underline{v}\right)\nabla_x T$$

$$= -\nu_5(T)\underline{u}.\nabla_x T.$$

With these two last equalities and Eq. (13) we obtain the balance entropy law (EL).

APPENDIX B: PROOF OF LEMMA 10

By using the fact that $M_{\varepsilon_1}^{-1}\overline{\Upsilon}_i \in (\mathrm{Ker}(\Gamma_{e,e}))^{\perp}$, let us first recall that

$$\int_{\mathbb{R}^3}\underline{v}\otimes\underline{\Upsilon}_i\ d\underline{v}=0.$$

Hence, we have on one hand

$$\partial_T\nu_4(T)\underline{\underline{\mathrm{I}_d}} = \frac{1}{T}\int_{\mathbb{R}^3}\underline{v}\otimes Q_{\mathrm{elas}}(\underline{\Upsilon}_2)d\underline{v}$$

$$= -\frac{1}{T}\int_{\mathbb{R}^3}\mathscr{M}_0(B^T_{\mathrm{elas}}(|\underline{v}|)-\nu_1(T))\underline{v}\otimes\mathscr{M}_0^{-1}\underline{\Upsilon}_2\ d\underline{v}$$

$$= -\frac{1}{T}\int_{\mathbb{R}^3}\mathscr{M}_0\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{\Upsilon}_1)\otimes\mathscr{M}_0^{-1}\underline{\Upsilon}_2 t d\underline{v}+O(\varepsilon_1)$$

$$= -\frac{1}{T}\langle\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}a(|\underline{v}|,T)\underline{v}),\mathscr{M}_0^{-1}b(|\underline{v}|,T)\underline{v}\rangle_0\underline{\underline{\mathrm{I}_d}}+O(\varepsilon_1)$$

$$= -\frac{1}{T}\langle\mathscr{M}_0^{-1}a(|\underline{v}|,T)\underline{v},\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}b(|\underline{v}|,T)\underline{v})\rangle_0\underline{\underline{\mathrm{I}_d}}+O(\varepsilon_1)$$

$$= -\frac{1}{T}\int_{\mathbb{R}^3}\underline{\Upsilon}_1\otimes\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{\Upsilon}_2)d\underline{v}+O(\varepsilon_1)$$

$$= +\frac{1}{T}\int_{\mathbb{R}^3}\underline{\Upsilon}_1\otimes\underline{v}\left(\frac{|\underline{v}|^2}{2T}-\frac{5}{2}\right)d\underline{v}+O(\varepsilon_1)=\frac{1}{T}\nu_5(T)\underline{\underline{\mathrm{I}_d}}+O(\varepsilon_1),$$

and on the other hand

$$\kappa(T)\underline{\underline{\mathrm{I}_d}} = \int_{\mathbb{R}^3}\frac{|v|^2}{2T}\underline{v}\otimes\underline{\Upsilon}_2\ d\underline{v}=\int_{\mathbb{R}^3}\left(\frac{|v|^2}{2T}-\frac{5}{2}\right)\underline{v}\otimes\underline{\Upsilon}_2\ d\underline{v}=-\int_{\mathbb{R}^3}\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}\underline{\Upsilon}_2)\otimes\underline{\Upsilon}_2\ d\underline{v}+0(\varepsilon_1)=$$

$$-\langle\widetilde{\Gamma}_{e,e}(\mathscr{M}_0^{-1}b(|\underline{v}|,T)\underline{v}),\mathscr{M}_0^{0-1}b(|vt|,T)\underline{v}\rangle_0\underline{\underline{\mathrm{I}_d}}+0(\varepsilon_1)>0+0(\varepsilon_1)$$

because the operator $\widetilde{\Gamma}_{e,e}$ is negative.

[1] K. Blotekjaer, ''Transport equations for electrons in two-valley semiconductors,'' IEEE Trans. Electron Devices **ED-17**, 38–47 (1970).
[2] M. Rudan and F. Odeh, ''Multi-dimensional discretization scheme for the hydrodynamic model of semiconductor devices,'' Compel **5**, 149–183 (1986).

[3] R. Stratton, ''Diffusion of hot and cold electrons in semiconductor barriers,'' Phys. Rev. **126**, 2002–2014 (1962).
[4] D. Chen, E. C. Kan, U. Ravaioli, C. W. Shu, and R. W. Dutton, ''An improved energy transport model including nonparabolicity and non-Maxwellian distribution effects,'' IEEE Electron Device Lett. **13**, 26–28 (1992).
[5] N. Ben Abdallah, P. Degond, and S. Genieys, ''An energy-transport model for semiconductors derived from the Boltzmann equation,'' J. Stat. Phys. **84**, 205–231 (1996).
[6] C. Bardos, ''Une interprétation des relations existant entre les équations de Boltzmann de Navier-Stokes et d'Euler à l'aide de l'entropie,'' Mat. Applic. Comp. **6**, 97–117 (1987).
[7] C. Cercignani, R. Illner, and M. Pulvirenti, *The mathematical theory of dilute gases* (Springer-Verlag, New York, 1994).
[8] S. Chapman and T. G. Cowling, *The mathematical theory of non-uniform gases* (Cambridge U. P., New York, 1958).
[9] N. Ben Abdallah and P. Degond, ''On a hierarchy of macroscopic models for semiconductors,'' J. Math. Phys. **37**, 3306–3333 (1996).
[10] S. Genieys, ''Energy-transport model for a non degenerate semiconductor. Convergence of the Hilbert expansion in the linearized case,'' Rapport interne de l'université P. Sabatier, Toulouse, France (1996).
[11] R. Stratton, ''The influence of interelectronic collisions on conduction and breakdown in covalent semi-conductors,'' Proc. R. Soc. London, Ser. A **242**, 355–373 (1957).
[12] R. Stratton, ''The influence of interelectronic collisions on conduction and breakdown in polar crystal,'' Proc. R. Soc. London, Ser. A **246**, 406–422 (1958).
[13] A. Gnudi, F. Odeh, and M. Rudan, ''Investigation of non-local transport phenomena in small semiconductor devices'' Eur. Trans. Telecommun. Related Technol. **1**, 77–82 (1990).
[14] C. Cercignani, *The Boltzman Equation and Its Applications* (Springer-Verlag, New York, 1988).
[15] H. Grad, ''Asymptotic theory of the Boltzmann equation II,'' in *Third symposium on rarefied gas dynamics I* (Academic, New York, 1963), pp. 26–59.
[16] L. Desvillettes and F. Golse, ''A remark concerning the Chapman-Enskog asymptotics,'' in *Advances in kinetic theory and computing, selected papers* (World Scientific, Singapore, 1994), pp. 191–203.
[17] S. Kawashima and Y. Shizuta, ''On the normal form of the symmetric hyperbolic-parabolic systems associated with the conservation laws,'' Tôhoku Math. J. **40**, 449–464 (1988).
[18] S. K. Godunov, ''The problem of a generalized solution in the theory of quasilinear equations and in gas dynamics,'' Russ. Math. Surv. **17**, 145–156 (1962).
[19] K. O. Friedrichs and P. D. Lax, ''Systems of conservation equations with a convex extension,'' Proc. Natl. Acad. Sci. USA **68**, 1686–1688 (1971).
[20] M. Mock, ''Systems of conservation laws of mixed type,'' J. Diff. Eqns. **37**, 70–88 (1980).
[21] Denis Serre, *Systèmes de lois de conservation. I.* (Diderot, Paris, 1996).
[22] S. Kawashima, ''Systems of hyperbolic-parabolic composite type, with applications to the equations of magnetohydrodynamics,'' Ph.D. thesis, Kyoto University, 1984.
[23] A. Harten, ''On the symmetric form of systems of conservation laws with entropy,'' J. Comp. Physiol. **49**, 151–164 (1983).
[24] T. Makino and B. Perthame, ''Sur les solutions à symétrie sphérique de l'équation d'Euler-Poisson pour l'évolution d'étoiles gazeuses,'' Jpn. J. Appl. Math. **7**, 165–170 (1990).
[25] B. Perthame, ''Non-existence of global solutions to Euler–Poisson equations for repulsive forces,'' Jpn. J. Appl. Math. **7**, 363–367 (1990).
[26] W. Van Roosbroeck, ''Theory of flow of electrons and holes in germanium and other semiconductors,'' Bell Syst. Tech. J. **29**, 560–562 (1950).
[27] F. Golse and F. Poupaud, ''Limite fluide des équations de Boltzman des semi-conducteurs pour une statistique de Fermi-Dirac,'' Asympt. Anal. **6**, 135–160 (1992).
[28] F. Poupaud, ''Runaway phenomena and fluid approximation under high fields in semiconductor kinetic theory,'' Z. Angew. Math. Mech. **72**, 359–372 (1992).
[29] P. Marcati and R. Natalini, ''Weak solution to a hydrodynamic model for semiconductors and relaxation to the drift-diffusion equation,'' Arch. Ration. Mech. Anal. **129**, 129–145 (1995).
[30] R. Natalini, ''The bipolar hydrodynamic model for semiconductors and the drift-diffusion equations,'' J. Math. Anal. Appl. **198**, 262–281 (1996).

Miura transformation between two non-linear equations in 2+1 dimensions

J. M. Cerveró and P. G. Estévez[a)]
Area de Física Teórica, Facultad de Física, Universidad de Salamanca, 37008 Salamanca, Spain

(Received 21 November 1997; accepted for publication 13 February 1998)

A Dispersive Wave equation in $2+1$ dimensions (2LDW) widely discussed by different authors is shown to be nothing but the modified version of the Generalized Dispersive Wave equation (GLDW). Using Singularity Analysis and techniques based upon the Painlevé Property leading to the Double Singular Manifold Expansion we shall find the Miura Transformation which converts the 2LDW equation into the GLDW equation. Through this Miura transformation we shall also present the Lax pair of the 2LDW equation as well as some interesting reductions to several already known integrable systems in $1+1$ dimensions. As the 2LDW equation arises from a Miura transformation we propose that it should be treated conventionally as a Modified Equation. In this case, we propose its designation as **The MGLDW Equation**. © *1998 American Institute of Physics.*
[S0022-2488(98)02405-0]

I. INTRODUCTION

Recently the following system of non-linear partial differential equations in $2+1$ dimensions has prompted a great deal of interest and several papers have appeared dealing with this system.[1,2] It could also be considered as a Long Dispersive Wave Equation (2LDW) in the form

$$\begin{aligned} 0&=\lambda q_t+q_{xx}-2qV,\\ 0&=\lambda r_t-r_{xx}+2rV,\\ 0&=(qr)_x-V_y\Rightarrow V=\int_{-\infty}^{y}(qr)_x dy'. \end{aligned} \tag{1.1}$$

(i) For $\lambda=i$ and $r=q^*$, it is identical to the equation proposed by Fokas[3] as the simplest integrable equation in $2+1$ dimensions. In this reference, the author presents the Lax pair of this particular case of (1.1). It is also worth pointing out that when $x=y$ the equation proposed by Fokas reduces to the Non-Linear Schrödinger Equation.[4]

(ii) The real version of (1.1) has been obtained by Chakravarty, Kent, and Newman[5] as a particular reduction of the General Self-Dual Yang Mills equations while the complex version of (1.1) has recently been discussed by Maccari[6] using the so-called Asymptotically Exact Reduction Method with the Kadomtsev–Petviashily equation as a starting point.

(iii) The Painlevé Property[7,8] as well as the Bilinear Hirota equations for the equation under consideration have been studied by Porsezian[2] for the complex version and by Radha and Lakhsmanan[1] for the real case. In this latter reference, a systematic use of the Bilinear Formalism is made in order to obtain soliton-like solutions of dromion type.

The aim of this paper is to show that equation (1.1) is nothing but the modified version of the Generalized Dispersive Wave Equation (GLDW) which has been thoroughly studied by Boiti, León and Pempinelli.[9] In order to show this, we shall be using the singularity analysis of the GLDW that has already been successfully applied to this equation by one of the present authors.[10] Also using techniques based upon Painlevé Analysis, we shall find the Miura transformation

[a)]Electronic mail: pilar@sonia.usal.es

 © 1998 American Institute of Physics

which converts (1.1) into the GLDW. In view of these results, it appears reasonable to refer to equation (1.1) as the Modified Generalized Dispersive Wave Equation (MGDLW). Since the Lax pair of the GLDW equation is known,[11,9] the Miura transformation that we shall present in this paper can easily be used to find the Lax pair for equation (1.1). As a final point, we shall offer several interesting reductions of our equation to some integrable equations in $1+1$ dimensions.

II. THE MIURA TRANSFORMATION BETWEEN THE GDLW AND THE MGDLW

In 1987 Boiti *et al.*[9] proposed a dispersive wave equation in $2+1$ dimensions,

$$0=u_{ty}+(\eta_{xy}+2uu_y)_x, \quad 0=\eta_{ty}+(u_{xy}+2u\eta_y)_x, \tag{2.1}$$

which we shall hereafter call the Generalized Dispersive Long Wave Equation (GDLW) since its reduction when $x=y$ is the dispersive wave equation also called the Kaup System[12] or Classical Boussinesq System.[13] Furthermore, the stationary version of (2.1) corresponds to the Sinh-Gordon equation.[9,14]

(i) **The Painlevé Property:** One of the present authors[10] has studied and discussed the system given by (2.1) from the point of view of a Painlevé Analysis. As is well known[7] the Painlevé Property (PP), for practical purposes, can be summarized in the statement that all solutions of equation (2.1) could be written in the form

$$u=\sum_{j=0}^{\infty} u_j\chi^{j-a}, \quad \eta=\sum_{j=0}^{\infty} \eta_j\chi^{j-b}, \tag{2.2}$$

where $\chi=0$ is an arbitrary function called the Movable Singularities Manifold while u_j and η_j are analytic functions in an arbitrary neighborhood of this manifold. Inserting (2.2) in (2.1) (the dominant terms balance easily) yields

$$a=b=1, \tag{2.3}$$

$$u_o=\pm\chi_x \quad \eta_o=\chi_x, \tag{2.4}$$

where $\pm$ means that two Painlevé expansions are in principle possible. In Ref. 10 a detailed discussion is presented on how to deal with both expansions simultaneously.

(ii) **The Double Singular Manifold Method:** In 1983 Weiss[8] proposed the Singular Manifold Method (SMM) which has been proved to be extremely successful in dealing with a large class of non-linear equations. The method is based on the restriction of the solutions of (2.1) for which expansion (2.2) becomes truncated at the constant level. In this case, one can prove that χ is no longer an arbitrary function since it must verify the truncation condition. Accordingly, henceforth it is called the *Singular Manifold.* As system (2.1) presents two different Painlevé branches, the suggestion made in Ref. 10 is to use **two singular manifolds** ϕ and $\hat{\phi}$ (one for each expansion branch) in such a way that the truncation of (2.2) adopts the form

$$u'=u+\frac{\phi_x}{\phi}-\frac{\hat{\phi}_x}{\hat{\phi}}, \quad \eta'=\eta+\frac{\phi_x}{\phi}+\frac{\hat{\phi}_x}{\hat{\phi}}. \tag{2.5}$$

(iii) **Superposition of solutions:** Expansion (2.5) suggests the following change of dependent variables:

$$u=m-\hat{m}, \quad \eta=m+\hat{m}, \tag{2.6}$$

such that m and $\hat{m}$ satisfy individual equations, each with just one branch of the Painlevé expansion. In order to find which equations m and $\hat{m}$ should satisfy, we add and subtract (2.1), and using (2.6) we finally obtain

$$0=m_{ty}+[m_{xy}+2(m-\hat{m})m_y]_x,$$
$$0=\hat{m}_{ty}+[-\hat{m}_{xy}+2(m-\hat{m})\hat{m}_y]_x. \tag{2.7}$$

Let us now define

$$m_t=n_x, \quad \hat{m}_t=\hat{n}_x. \tag{2.8}$$

After an integration over the variable x in (2.7) we obtain

$$0=n_y+m_{xy}+2(m-\hat{m})m_y, \tag{2.9}$$

$$0=\hat{n}_y-\hat{m}_{xy}+2(m-\hat{m})\hat{m}_y, \tag{2.10}$$

These equations allow us to obtain m and $\hat{m}$ from (2.9) and (2.10). The result is

$$m=\hat{m}+\frac{\hat{m}_{xy}-\hat{n}_y}{2\hat{m}_y}, \tag{2.11}$$

$$\hat{m}=m+\frac{m_{xy}+n_y}{2m_y}. \tag{2.12}$$

By substituting (2.11) in (2.9) and (2.12) in (2.10), one can easily decouple these equations, thus obtaining one that is satisfied by m and another one that is satisfied just by $\hat{m}$. Is trivial to check that both equations are one and the same but written with different variables. In fact (m,n) satisfies

$$0=m_t-n_x,$$
$$0=m_y^2(n_{yt}-m_{xxxy})+m_{xy}(n_y^2-m_{xy}^2)+2m_y(m_{xy}m_{xxy}-n_y n_{xy})-4m_y^3 m_{xx}, \tag{2.13}$$

while $(\hat{m},\hat{n})$ satisfies

$$0=\hat{m}_t-\hat{n}_x,$$
$$0=\hat{m}_y^2(\hat{n}_{yt}-\hat{m}_{xxxy})+\hat{m}_{xy}(\hat{n}_y^2-\hat{m}_{xy}^2)+2\hat{m}_y(\hat{m}_{xy}\hat{m}_{xxy}-\hat{n}_y\hat{n}_{xy})-4\hat{m}_y^3\hat{m}_{xx}, \tag{2.14}$$

which is obviously the same equation. Therefore, m and $\hat{m}$ are both solutions of equation (2.13) but constrained by the transformation (2.11)–(2.12). It is not difficult to recognize the constraint as an **Auto-Bäcklund Transformation** between two solutions of (2.13).

(iv) **The Miura Transformation:** With regard to u and η, these can be expressed through (2.6) as a linear superposition of m and $\hat{m}$. Taking into account (2.6) and (2.11), one can write

$$u=m-\hat{m}=\frac{\hat{m}_{xy}-\hat{n}_y}{2\hat{m}_y}=-\frac{m_{xy}+n_y}{2m_y},$$
$$\eta=m+\hat{m}=2\hat{m}+\frac{\hat{m}_{xy}-\hat{n}_y}{2\hat{m}_y}=2m+\frac{m_{xy}+n_y}{2m_y}. \tag{2.15}$$

In order to express m and $\hat{m}$ as a function of u and η with the help of (2.15), we must previously realize that the first of the equations (2.1) can be integrated in the form

$$\eta_x=-\partial_x^{-1}u_t-u_x^2, \tag{2.16}$$

in such a way that m and $\hat{m}$ can be finally expressed as

$$\begin{aligned} 2m_x &= u_x + \eta_x = u_x - u^2 - \partial_x^{-1} u_t, \\ 2\hat{m}_x &= \eta_x - u_x = -u_x - u^2 - \partial_x^{-1} u_t. \end{aligned} \tag{2.17}$$

Expressions (2.17) are Miura transformations between solutions of the non-linear systems (2.1) and (2.14). In the next Section we shall show that system (2.14) is nothing but the initial system given by (1.1). Obviously, (2.13) must be considered as the **Modified GLDW equation** since it has been transformed through the **Miura transformation** (2.17) from the GLDW equation. Hereafter, we shall refer to (2.13) as the MGLDW equation.

III. THE MODIFIED GDLW EQUATION

We now attempt to show that (1.1) must truly be considered the modified GDLW equation (2.13).

A. Equation (1.1) as the Modified GDLW equation

We first try to write down (1.1) as a equation for just one field. In order to do so, we make the change,

$$V = -m_x, \tag{3.1}$$

in such a way that (1.1) becomes

$$0 = q_t + q_{xx} + 2qm_x, \tag{3.2}$$

$$0 = r_t - r_{xx} - 2rm_x, \tag{3.3}$$

$$0 = qr + m_y, \tag{3.4}$$

where time has been rescaled in the form

$$t \to \lambda t.$$

Taking q out of (3.4) and substituting it in (3.2), we find

$$0 = \frac{m_{yt} + m_{xxy}}{r} - 2m_{xy}\frac{r_x}{r^2} + m_y\left(\frac{2m_x}{r} - \frac{r_t}{r^2} - \frac{r_{xx}}{r^2} - 2\frac{r_x^2}{r^3}\right). \tag{3.5}$$

Using (3.3) in (3.5), we also obtain

$$0 = m_{xxy} + m_{yt} - \left(2m_y\frac{r_x}{r}\right)_x, \tag{3.6}$$

which can easily be integrated by setting $m_t = n_x$, which yields

$$\frac{r_x}{r} = \frac{m_{xy} + n_y}{2m_y}. \tag{3.7}$$

Substituting (3.7) in (3.3), we obtain

$$\frac{r_t}{r} = 2m_x + \frac{m_{xxy} + n_{xy}}{2m_y} - \frac{m_{xy}^2 - n_y^2}{4m_y^2}. \tag{3.8}$$

Next, we calculate the obvious identity $(r_t/r)_x = (r_x/r)_t$ using (3.7) and (3.8), and finally we obtain for m the equation

$$0 = m_t - n_x,$$
$$0 = m_y^2(n_{yt} - m_{xxxy}) + m_{xy}(n_y^2 - m_{xy}^2) + 2m_y(m_{xy}m_{xxy} - n_y n_{xy}) - 4m_y^3 m_{xx}, \tag{3.9}$$

which is precisely equation (2.13). Taking into account (3.1) and (2.17), we also find

$$V = \tfrac{1}{2}(u_x + u^2 + \partial_x^{-1} u_t), \tag{3.10}$$

that establishes the relationship between the V-field of (1.1) and the u-field of (2.1).

Also, one can compare (3.7)–(3.8) and (2.15) yielding

$$\frac{r_x}{r} = -u, \quad \frac{r_t}{r} = \eta_x + u^2. \tag{3.11}$$

B. The Lax pair for MGDLW

The Lax pair of the GLDW equation (2.1) is well known and can easily be written as[10]

$$0 = \varphi_t - \varphi_{xx} + 2u\varphi_x, \quad 0 = \varphi_{xy} - \partial_x^{-1} u_y \varphi_x + \frac{\eta_y - u_y}{2}\varphi. \tag{3.12}$$

Alternatively, one can also write the Lax pair by using the gauge transformation:

$$\varphi = e^{\partial_x^{-1} u}\psi,$$

which leads equation (3.12) to the expression

$$0 = \psi_t - \psi_{xx} + [\partial_x^{-1} u_t - u_x + u^2]\psi, \quad 0 = \psi_{xy} + u\psi_y + \frac{\eta_y + u_y}{2}\psi. \tag{3.13}$$

This is the form of the Lax pair obtained by other authors.[11,9] Using (2.15)–(2.17), expression (3.13) becomes

$$0 = \psi_t - \psi_{xx} - 2m_x\psi,$$
$$0 = \psi_{xy} - \frac{m_{xy} + n_y}{2m_y}\psi_y + m_y\psi, \tag{3.14}$$

which can be considered as the Lax pair for the MGLDW equation (2.13). Taking into account (3.1), (3.4), and (3.8), this latter expression can also be written as

$$0 = \psi_t - \psi_{xx} + 2V\psi, \quad 0 = \psi_{xy} - \frac{r_x}{r}\psi_y - qr\psi, \tag{3.15}$$

which would be the same Lax pair but for form (3.2)–(3.4) of the MGLDW equation. Since (3.1)–(3.4) is invariant under the transformation

$$t \to -t, \quad r \to q,$$

the following pair of equations:

$$0 = \hat{\psi}_t + \hat{\psi}_{xx} - 2V\hat{\psi}, \quad 0 = \hat{\psi}_{xy} - \frac{q_x}{q}\hat{\psi}_y - qr\hat{\psi}, \tag{3.16}$$

would also be a Lax pair for the MGLDW equation.

IV. PARTICULAR CASES

Let us now examine the Miura transformations induced in some interesting reductions of the MGLDW equation.

A. The stationary case: Sinh-Gordon versus AKNS

(i) Let us assume that the fields of the **MGDLW equation** (3.9) do not depend on t. The equation takes the form

$$0=-m_y^2 m_{xxxy}-m_{xy}^3+2m_y m_{xy} m_{xxy}-4m_y^3 m_{xx}, \tag{4.1}$$

which can be integrated with respect to x to give

$$2m_{xxy}m_y+8m_x m_y^2-m_{xy}^2=0. \tag{4.2}$$

Eliminating m_{xxy} between (4.1) and (4.2), we finally obtain

$$m_{xxxy}+8m_x m_{xy}+4m_y m_{xx}=0, \tag{4.3}$$

which is the **AKNS equation in** $1+1$ **dimensions**.[15]

(ii) With the same reduction of time-independent fields, **The GLDW equation** (2.1) can also be written as

$$0=\eta_{xy}+2uu_y, \tag{4.4}$$

$$0=u_{xy}+2u\eta_y. \tag{4.5}$$

Let us now multiply (4.4) by η_y, (4.5) by u_y, and subtract. The result is

$$\frac{d(\eta_y^2-u_y^2)}{dx}=0.$$

Next, we introduce the change of variables,

$$\eta_y=a_0\cosh 2q, \quad u_y=a_0\sinh 2q, \tag{4.6}$$

and so equations (4.4)–(4.5) take the form

$$u=-q_x. \tag{4.7}$$

Alternatively, using (4.6) we finally obtain

$$q_{xy}=a_0\sinh 2q, \tag{4.8}$$

which is **the Sinh-Gordon equation**.[11,14]

(iii) Therefore, the particularization of the Miura transformation (2.17) to the stationary case,

$$2m_x=u_x-u^2=-q_{xx}-q_x^2, \quad 2\hat{m}_x=-u_x-u^2=q_{xx}-q_x^2, \tag{4.9}$$

yields the corresponding Miura transformation between solutions of the AKNS equation (4.3) and the Sinh-Gordon equation (4.8).

B. Non-local Boussinesq equation versus the Kaup system

In the next subsection we shall deal with the reduction which arises from setting $x=y$.

(i) In the present case, the MGLDW equation (3.10) is reduced to

$$0=m_t-n_x,$$
$$0=m_x^2(n_{xt}-m_{xxxx})+m_{xx}(n_x^2-m_{xx}^2)=2m_x(m_{xx}m_{xxx}-n_xn_{xx})-4m_x^3m_{xx}. \tag{4.10}$$

The set of equations (4.10) can be integrated with respect to x. One easily obtains

$$0=m_t-n_x, \quad 0=m_x(n_t-m_{xxx})-(n_x^2-m_{xx}^2)-2m_x^3. \tag{4.11}$$

Alternatively, we can eliminate the field n. Thus, for m we find the following equation:

$$m_{tt}=\left[m_{xxx}+2m_x^2+\frac{n_x^2-m_{xx}^2}{m_x}\right]_x. \tag{4.12}$$

Lambert *et al.*[16,17] have called (4.12) **the non-local Boussinesq equation (NLBq)**.

(ii) We now apply the reduction $x=y$ to the GLDW equation (2.1):

$$0=u_t+\eta_{xx}+2uu_x, \quad 0=\eta_t+u_{xx}+2u\eta_x. \tag{4.13}$$

This system has been known as the classical Boussinesq system[13] or **the Kaup system**.[18,12,19,16]

(iii) The Miura transformation (2.17) establishes a relationship between the solutions of the non-local Boussinesq equation (4.12) and the Kaup system (4.13).

(iv) An important subcase of this reduction is the **non-linear Schrödinger equation**. To see this, we must notice that the reduction $x=y$ applied to the MGLDW in its form given by (3.2)–(3.4) appears as

$$0=q_t+q_{xx}-2q^2r, \quad 0=r_t-r_{xx}+2r^2q, \tag{4.14}$$

which for imaginary time, $t\to it$ and $q=r^*$, is clearly the non-linear Schrödinger equation.[3,20]

V. CONCLUSIONS

(i) First, starting from Painlevé analysis of the Dispersive Long Wave Equation in $2+1$ dimensions (GDLW), we have been able to decompose the solutions of this equation as a **linear superposition** of two solutions of another equation in $2+1$ that has been called a Modified Dispersive Long Wave Equation (MGDLW).

(ii) The expansion of the solutions of the GDLW as a linear superposition of its modified version (MGDLW) is shown to be invertible and gives rise to the Miura transformation between both equations. This Miura transformation has been obtained explicitly.

(iii) In Section III we have shown that the MGLDW equation is nothing but the equation proposed by several authors[3,1,6,2] as one of the simplest possible integrable equation in $2+1$ dimensions.

(iv) Since the Lax pair of the GDLW equation is already known,[9] the Miura transformation allows one to pass from the GDLW equation to the MGDLW equation. In particular, we can easily find the Lax pair of the latter equation.

(v) In Section IV some interesting reductions of the MGDLW equation are presented using the Miura transformation between GDLW and MGDLW. In this way, one can obtain the Miura transformations between AKNS and Sinh-Gordon in $1+1$ and between the non-local Boussinesq equation and the Kaup system, also in $1+1$ dimensions.

(vi) In view of the above results one could conjecture that when two or more expansion branches in the Painlevé expansion appear in a given non-linear partial differential equation, a Miura transformation exists that converts the initial equation into one with just one branch in this expansion.

(vii) Systematic application of the Singular Manifold Method[8] and use of it to find the Darboux transformation for the MGDLW equation pave the way for finding extremely simple algorithms for the construction of solutions that generalize the ones already obtained by other

authors[1] using the bilinear formalism. We shall shortly be presenting these results in a forthcoming publication.

ACKNOWLEDGMENTS

This research has been supported in part by the DGICYT under Contract No. PB95-0947.

[1] R. Radha and X. Lakshmanan, J. Math. Phys. **24**, 292 (1983).
[2] K. Porsezian, J. Math. Phys. **38**, 4675 (1997).
[3] A. Fokas, Inverse Probl. **10**, L19 (1994).
[4] V. Zakharov and A. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[5] S. Chakravarty, S. L. Kent, and E. T. Newman, J. Math. Phys. **36**, 763 (1995).
[6] A. Maccari, J. Math. Phys. **37**, 6207 (1996).
[7] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[8] J. Weiss, J. Math. Phys. **24**, 1405 (1983).
[9] M. Boiti, J. Leon, and F. Pempinelli, Inverse Probl. **3**, 37 (1987).
[10] P. G. Estévez and P. R. Gordoa, Inverse Probl. **13**, 939 (1997).
[11] M. J. Ablowitz and P. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, London Mathematical Society, Lecture Note Series **149** (Cambridge University Press, Cambridge, 1991).
[12] D. Kaup, Prog. Theor. Phys. **54**, 396 (1975).
[13] P. G. Estévez, P. R. Gordoa, L. Martinez-Alonso, and E. Medina-Reus, J. Phys. A **26**, 1915 (1993).
[14] P. G. Estévez, E. Conde-Calvo, and P. R. Gordoa, J. Nonlinear Math. Phys. (in press).
[15] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[16] F. Lambert, I. Loris, J. Springael, and R. Willox, J. Phys. A **27**, 5325 (1994).
[17] R. Willox, I. Loris, and J. Springael, J. Phys. A **28**, 5963 (1995).
[18] R. Conte, M. Musette, and A. Pickering, J. Phys. A **27**, 2831 (1994).
[19] S. Kawamoto, J. Phys. Soc. Jpn. **53**, 2922 (1984).
[20] M. Tabor, *Chaos and Integrability in Nonlinear Dynamics* (Wiley, New York, 1989).

Couple coefficients for the two-block AKNS system

Gulmaro Corona Corona[a)]
Area de Análisis Matemático y sus Aplicaciones, Edif. H-122, Depto. C. B., Universidad Autónoma Metropolitana, Av. Sn. Pablo Xalpa 180, Col. Reynosa Tamaulipas, Atzcapotzalco, D. F., CP. 02200, México

(Received 15 January 1997; accepted for publication 27 January 1998)

A model for the 2×2 AKNS system is extended for the two-block AKNS system. Zeroes of the determinant of the J-diagonal scattering matrix blocks give the bound states of such a system. The couple coefficients are now matrices. © *1998 American Institute of Physics.* [S0022-2488(98)00905-0]

I. INTRODUCTION

In Ref. 1 Sasa and Satsuma found that a certain type of higher-order nonlinear Schrödinger equation can be solved by the inverse scattering transform method. This kind of equation is associated with the AKNS operator D_x given by

$$D_x = D_x(z,q) = \frac{\partial}{\partial x} - zJ - q$$

for

$$J = i\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

and z a complex number called spectral value. Here we consider the $n\times n$ constant matrix

$$J = \begin{pmatrix} p_- \mathbf{1}_{p_+} & 0_{p_+\times p_-} \\ 0_{p_-\times p_+} & -p_+ \mathbf{1}_{p_-} \end{pmatrix},$$

$p_\pm$ being positive integers whose sum is n; $\mathbf{1}_{p_\pm}$ are the $p_\pm\times p_\pm$ identity matrices, respectively, and $0_{p_\pm\times p_\mp}$ are the $p_\pm\times p_\mp$ null matrices, respectively. The $n\times n$ matrix-valued function is defined on the real line $q=\left(\begin{smallmatrix} 0_{p_+} & r_+ \\ r_- & 0_{p_-} \end{smallmatrix}\right)$ whose entries are in the Schwartz class where $0_{p_\pm}$ are the $p_\pm\times p_\pm$ null matrices, respectively, and $r_\pm$ are $(p_\pm\times p_\mp)$ matrix-valued functions. In the sequel, we refer to this system as the two-block AKNS system. For $p_+=1$ and $p_-=2$ the coupled nonlinear Schrödinger equation (cf. Ref. 2) is associated with this operator.

Three years before Ref. 1 appeared, Sattinger and Zurkowsky in Ref. 3 found a method to characterize the zeroes of the diagonal coefficients of the scattering matrix 2×2 case: $p_-=p_+=1$.

The above method cannot be extended to the generic case where the matrix J has different diagonal entries, but in this paper we extend it to study the two-block AKNS system.

II. CONSTRUCTION OF THE WAVE FUNCTIONS

We generalize the construction given in Ref. 3 of the wave functions ϕ for the present case, i.e., those functions ϕ such that $D_x(z,q)\phi=0$. These functions are also called the eigenfunctions [roughly speaking of the operator $J^{-1}(d/dx-q)$ corresponding to the eigenvalue or spectral value z see Ref. 4]. In the sequel, we refer to them as wave functions of D_x or just the wave functions.

a)Electronic mail: ccg@hp9000a1.uam.mx

 © 1998 American Institute of Physics

The $n\times p_+$ matrix e_+ and the $n\times p_-$ matrix e_- are defined by

$$e_+=\begin{pmatrix}\mathbf{1}_{p_+}\\ 0_{p_-}\times p_+\end{pmatrix},\quad e-=\begin{pmatrix}0_{p_+\times p_-}\\ \mathbf{1}_{p_-}\end{pmatrix}.$$

respectively.

The wave functions are obtained by rewriting the differential equation as a Volterra integral equation. For example, write $\phi=\ell e^{-p_+xz}$. Then we obtain differential equation

$$\frac{d\ell}{dx}=zE\ell+q\ell,\quad E=\begin{pmatrix}ni\mathbf{1}_{p_+} & 0_{p_+\times p_-}\\ 0_{p_-}\times p_+ & 0_{p_-}\end{pmatrix},$$

where 0_{p_-} is the null $p_-\times p_-$ square matrix. This may be rewritten as the Volterra integral equation; for instance,

$$\ell=e_-+\int_\infty^x e^{zE(x-y)}(q\ell)(y)\,dy.$$

The convergence of the successive approximations depends upon the potential q. For instance, if $q\in L_1(\mathbb{R})$, these converge uniformly in $\operatorname{Im} z\geqslant 0$; the solution ℓ is analytic in the upper half-plane $\operatorname{Im} z\geqslant 0$, is bounded for real x, and tends to e_- as x tends to $-\infty$.

The smoothness of the potential q gives an asymptotic expansion for ℓ. If $q\in C^r$, then

$$\ell\sim\sum_{j=0}^{r}\ell_j z^{-j},$$

where $\ell_0=e_-$, $E\ell_1=-qe_-$, and $E\ell_{j+1}=d\ell_j/dx-q\ell_j$.

So there exist wave matrix-valued functions $\phi_\pm$ and $\psi_\pm$, analytic in $\pm\operatorname{Im} z>0$ with the asymptotic behavior

$$\phi_\pm\sim e^{\mp ip_\mp xz}e_\mp,\quad x\to-\infty,\quad \psi_\pm\sim e^{\pm ip_\pm xz}e_\pm,\quad x\to\infty, \tag{1}$$

is proven in this manner.

Let Φ and Ψ denote matrices

$$\Phi(x,k)=\|\phi_-(x,k),\phi_+(x,k)\|,\quad \Psi(x,k)=\|\psi_+(x,k),\psi_-(x,k)\|$$

for real spectral value k. In general, the wave functions (with the spectral value k) Φ and Ψ are defined only for real k and satisfy the integral Volterra equations

$$\begin{aligned}\Phi(x,k)&=e^{xkJ}+\int_{-\infty}^{x}e^{(x-y)kJ}q\Phi\,dy,\\ \Psi(x,k)&=e^{xkJ}-\int_{x}^{\infty}e^{(x-y)kJ}q\Psi\,dy.\end{aligned} \tag{2}$$

Remark 1: However, if the potential q has compact support, Φ and Ψ are entire functions of the spectral value k.

III. THE ZEROES OF THE DETERMINANT OF *J*-DIAGONAL BLOCKS OF THE SCATTERING MATRIX

Each of these matrix-valued functions is a fundamental solution set on the real axis, so for each real spectral k there exists a matrix $S(k)$ such that they are related by

$$\Phi(x,k)=\Psi(x,k)s(k),\quad s(k)=\begin{pmatrix}a(k) & b(k)\\ c(k) & d(k)\end{pmatrix}. \tag{3}$$

Here $a(k)$ and $d(k)$ are $p_\pm \times p_\pm$-square matrices, respectively, and $b(k)$ and $c(k)$ are $p_\pm \times p_\pm$ matrices, respectively. From now on we call the former the J-diagnonal blocks.

By a standard argument, it is found that $\det \Phi(x,k)$ and $\det \Psi(x,k)$ are constant as functions of x for each real spectral value k. Evaluating them as $x \to \pm\infty$, respectively, we find that they are identically one. Therefore, $s(k) \in \mathrm{SL}(n,\mathbb{C})$ for each real k.

Remark 2: Solving (2) for $s(k)$ and taking into account Remark 1 we find $s(k)$ is defined for all complex numbers k and is an entire function of k whenever we let q have compact support; in particular, the matrices a and d are entire functions of k.

Substituting the left relation in (3) into the first equation in (2) and assuming that q has compact support, we obtain

$$s(k) = \mathbf{1}_{n\times n} + \int_{-\infty}^{x} e^{-ykJ} q(y,k) \Phi(y,k)\, dy, \quad x > n,$$

for some positive integer n and any fixed x.

If q' is another potential with compact support, we have for the upper J-diagonal block of $s(k) - s'(k)$ that

$$\begin{aligned} d(k) - d'(k) = & \int_{-\infty}^{x} e^{iykp_-} \| r_-(y,k) - r'_-(y,k) 0 \| \phi_+(y,k)\, dy \\ & + \int_{-\infty}^{x} e^{iykp_-} \| r'_-(y,k) 0 \| (\phi_+(y,k) - \phi'_+(y,k))\, dy. \end{aligned} \tag{4}$$

Remark 3: Since the successive approximations for ϕ_- converge uniformly in the lower half-plane $-\operatorname{Im} k \geqslant 0$, it is bounded uniformly in the same half-plane. Hence the first term in (4) is bounded by $C_- |q_1 - q'_1|_{L_1}$, where $C_- = C_-(q,x)$ is a bound for ϕ_- in the lower half-plane. It is not hard to see that $\phi_- - \phi'_-$ is bounded by $C_- = C_-(q,q',x)|q - q'|_{L_1}$ in the lower half-plane. So the second term is bounded by $C_- |q'|_{L_1} |q - q'|_{L_1}$.

As a consequence of this last remark, we have the next result.

Theorem 1: *For any potential q, the matrices a and d can be analytically extended to the lower and upper half-planes, respectively.*

Proof: Because of Remark 2, if the potential has compact support, then d is an entire function. If q does not have compact support, then we approximate it by potentials of compact support. By Remark 3, d is the uniform limit of analytic functions in $-\operatorname{Im} z \geqslant 0$. Therefore it can be extended analytically by Morera's theorem. Similarly, we proceed with a. □

Decoupling (3) by blocks, we get a matrix equation system, namely,

$$\begin{aligned} \phi_+ &= \psi_+ b + \psi_- d, \\ \phi_- &= \psi_+ a + \psi_- c. \end{aligned} \tag{5}$$

For instance, rewriting the first as a system for the unknowns $b_{1j}, \ldots b_{p_+ j}$, and $d_{1j}, \ldots d_{p_- j}$ for fixed $j = 1,\ldots,p_-$, we can see that

$$d_{lj} = |\, \psi_+ \quad \psi_-^1 \cdots \psi_-^{j-1} \quad \phi_+^l \quad \psi_-^{j+1} \cdots \psi_-^{p_-} |, \quad l,j = 1,\ldots,p_- . \tag{6a}$$

Similarly,

$$a_{lj} = |\, \psi_+^1 \cdots \psi_+^{j-1} \quad \phi_-^l \quad \psi_+^{j+l} \cdots \psi_+^{p_+} \quad \psi_- |, \quad l,j = 1,\ldots,p_+ . \tag{6b}$$

Here the superscript denotes the column of the matrix.

Definition 1: Let ψ and ϕ be two matrices with the same number of rows and with p_+, p_- columns, respectively. Its right wedge product by columns is a matrix given by

$$\psi \wedge^c \phi = \| \wedge_{\ell=1}^{p_+} \psi^\ell \wedge \wedge_{\ell=2}^{p_-} \phi^\ell \quad \cdots \quad \wedge_{\ell=1}^{p_+} \psi^\ell \wedge \wedge_{\substack{\ell=1 \\ \ell \neq j}}^{p_-} \phi^\ell \quad \cdots \quad \wedge_{\ell=2}^{p_+} \psi^\ell \wedge \wedge_{\ell=1}^{p_- - 1} \phi^\ell \|.$$

Analogously, the left wedge product $\psi^c\wedge\phi$ is defined.

If ρ is a third matrix with p_- columns, then the wedge product by columns with $\psi\wedge^c\phi$ is the matrix given by

$$(\psi\wedge^c\phi)\wedge^c\rho=(\wedge_{\ell=1}^{p_+}\psi^{\ell}\wedge\phi^1\wedge\quad\cdots\quad\wedge\phi^{j-1}\wedge\rho^l\wedge\phi^{j+1}\wedge\quad\cdots\quad\wedge\phi^{p_+})_{lj}.$$

Similarly, we define $\psi^c\wedge\phi$.

With this notation at hand, we obtain from (5), (6a), and (6b) that

$$d=(\psi_+\wedge^c\psi_-)\wedge^c\phi_+,\quad a=\phi_-{}^c\wedge(\psi_+{}^c\wedge\psi_-). \tag{7}$$

Also $\wedge^c$-multiplying (5) by $\psi_+\wedge^c\phi_+$, we obtain the relation

$$(\psi_+\wedge^c\phi_+)\wedge^c\psi_-d=(\psi_+\wedge^c\phi_+)\wedge^c\psi_+. \tag{8}$$

By a direct calculation, we may check that

$$(\psi_+\wedge^c\phi_+)\wedge^c\psi_+=\det\|\psi_+\quad\phi_+\|\mathbf{1}_{n\times n}. \tag{9}$$

On the other hand, the limits as x tends to infinity of $(\psi_+\wedge^c\psi_-)\wedge^c\phi_+$ and $(\psi_+\wedge^c\phi_+)\wedge^c\psi_-$ have the form $(\psi_+^l\wedge^c e_-)\wedge^c e_+$ and $(\psi_+^l\wedge^c e_+)\wedge^c e_-$, respectively, because of (1). It is a straightforward calculation to check that

$$(\psi_+\wedge^c\phi_+)\wedge^c\psi_-d=((\psi_+^l\wedge^c e_-)\wedge^c e_+)(\psi_+^l\wedge^c e_+)\wedge^c e_-=\det d\,\mathbf{1}_{n\times n}. \tag{10}$$

The determinants $\det\|\psi_+\quad\phi_+\|$ and $\det\|\psi_-\quad\phi_-\|$ do not depend of x, since $\|\psi_+\quad\phi_+\|$ and $\|\psi_-\quad\phi_-\|$ are solutions of $D_x\psi=0$ by construction of $\psi_\pm$ and $\phi_\pm$. Thus we have from (7)–(10), and doing a similar procedure for a, that

$$\det d=\det\|\psi_+\quad\phi_+\|,\quad \det a=\det\|\phi_-\quad\psi_-\|. \tag{11}$$

Therefore, by means of Theorem 1, we can establish the next result.

Theorem 2: *The determinants* $\det d$ *and* $\det a$ *can be analytically extended to the upper and lower half-planes, respectively. In addition, their zeroes are the same as those of* $\det\|\psi_+\quad\phi_+\|$ *and* $\det\|\phi_-\quad\psi_-\|$, *respectively*.

Remark 4: Theorem 2 can be directly proved from relations in (7) and (8). In the case that $p_+=1$ and $n=3$, the identity in (10) follows from a straight calculation, without taking the time as x goes to infinity. In addition, we have

$$a=\det\|\phi_-\quad\psi_-\|,\quad \det d=\|\psi_+\quad\phi_+\|.$$

Remark 5: From the identities in (6) we can estimate the limits of $e^{xzp_-}\phi_-$, $e^{xzp_-}\phi_+$ as $x\to\infty$. For instance, $e^{xzp_\mp}\phi_+\to\|{}^0_d\|$ as $x\to\infty$. Using this limit, one may check that the matrix functions defined on the real line by

$$\|\psi_+\quad\phi_+d^{-1}\|=\hat{m}_+e^{xzJ},\quad \operatorname{Im} z>0,\quad \|\phi_-a^{-1}\quad\psi_-\|=\hat{m}_-e^{xzJ},\quad \operatorname{Im} z<0,$$

satisfy that $\hat{m}_\pm\to\mathbf{1}$ as $x\to\infty$, and the others properties of a reduced wave function (see e.g., Ref. 5). However, we cannot obtain wave functions with the property $m_\pm\to\mathbf{1}$ as $x\to-\infty$ in this way, in contrast to the case $p_-=p_+=1$ (cf. Ref. 2). To construct those wave functions we must use another method, for instance that of X. Zhou (cf. Ref. 5).

IV. THE COUPLE MATRIX COEFFICIENTS

Definition 2: Let ϕ and ψ be two wave functions of D_x. If there exists a constant matrix c such that $\phi=c\psi$ or $c\phi=\psi$, we say that those matrices are matrix multiples. The matrix c is called the corresponding couple-matrix coefficient.

According to Theorem 2, $\det d(k^+)=0$ and $\det a(k^-)=0$ if and only if the eigenfunctions $\phi_\pm(k^\pm)$ and $\psi_\pm(k^\pm)$, corresponding to the eigenvalues $k^\pm$, respectively, are matrix multiples.

Extending the definition used in 2×2 AKNS systems to the present case, we call these wave functions the bound states associated to the potential q. With this notation at hand, as an immediate consequence of Theorem 2, we can state the following.

Corollary 1: The zeroes of det d *and* det a *give the bound states*.

Remark 4: In the case of $p_+=p_-=1$, in each zero $k^{\pm}$ of det d and det a, respectively, we have at the same time the relations

$$\phi_+=\psi_+c_+ \quad \text{and} \quad \phi_-=\psi_-c_-. \tag{12}$$

In addition, $c_\pm$ are scalars. However, this is no longer true in the other cases. For instance, we encountered that in the case $p_+\leqslant p_-$ then

$$\phi_-=\psi_-c_- \quad \text{but} \quad \phi_+c_+=\psi_+, \tag{13}$$

where $c_\pm$ are now matrices.

If we still assume that $p_+\leqslant p_-$, from (12) we obtain that

$$c_-=\frac{1}{|\wedge_{\ell=1}^{p_+}\psi_-^{\ell}|^2}(\langle\wedge_{\ell=1}^{i-1}\psi_-^{\ell}\wedge\phi_-^{j}\wedge_{\ell=i+i}^{p_+}\psi_-^{\ell},\wedge_{\ell=1}^{p_+}\psi^{\ell}\rangle)_{ij},$$

$$c_+=\frac{1}{|\wedge_{\ell=1}^{p_-}\phi_+^{\ell}|^2}(\langle\wedge_{\ell=1}^{i-1}{}_+\phi_+^{\ell}\wedge\psi_+^{j}\wedge_{\ell=i+i}^{p_+}\phi_-^{\ell},\wedge_{\ell=1}^{p_-}\phi_+^{\ell}\rangle)_{ij},$$

where $\langle\cdot,\cdot\rangle$ denotes any inner product on the complex Eculidian space $\mathbb{C}^n$, and $|\cdot|$ denotes the corresponding norm. Since we can obtain formulas similar to that in (6) for the coefficients b and c of the scattering matrix s, we recover in a different way the standard result (see Ref. 5), namely, s and $(a,b,c,b,c_\pm(k_j^\pm))$ smoothly depend of q. Here we suppose the number of zeroes $k_j^\pm$ of a and d, respectively, to be finite.

CONCLUSIONS

We have given a different method to construct wave functions from those given in the literature based on the method used for the 2×2 case by Sattinger and Zarkowsky in Ref. 3. However, other types of wave functions required in scattering theory (see Remark 5) should be constructed by using a method given in standard literature (e.g., as in Ref. 5). This method let us study the bound states in a precise sense for the two-by-two-blocks AKNS systems. The couple coefficients are now matrices. In the case that they are scalar, i.e., the 2×2 case, they need to be different from zero. For the present case, they must be left or right invertible to have, for example, the relations in (12) at the same time. A similar assumption should also be made for the model given by Satsuma and Sasa in the case $p_+=1$ and $p_-=2$ (see Ref. 1).

This is not a surprising fact, since we know that the inverse theory via a Riemman–Hilbert problem is only developed for a kind of potential (see Refs. 2 and 5).

[1] N. Sasa and J. Satsuma, ''New-type of soliton solutions for a higher-order nonlinear Schrödinger equation,'' J. Phys. Soc. Jpn. **60**, 409–417 (1987).
[2] S. V. Manakov, ''On the theory of two-dimensional stationary self-focusing of electromagnetic waves,'' Sov. Phys. JETP **38**, 248–253 (1985).
[3] D. H. Sattinger and V. D. Zurkowski, ''Gauge theory of Backlund transformation,'' in *Dynamics of Infinite Dimensional Systems*, edited by S.-N. Chow and J. K. Hale (Springer-Verlag, Berlin, 1987), Vol. F37, pp. 273–300.
[4] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution and Inverse Scattering* (Cambridge U.P., Cambridge, 1989).
[5] X. Zhou, ''Direct and inverse scattering transforms with arbitrary spectral singularities,'' Commun. Pure Appl. Math. **42**, 895–938 (1989).

The algebraic structure of the gl($n|m$) color Calogero–Sutherland models

Guo-xing Ju
Institute of Theoretical Physics, Academia Sinica, Beijing 100080, People's Republic of China, and Physics Department, Henan Normal University, Xinxiang, Henan Province 453002, People's Republic of China

Shi-kun Wang
CCAST (World Laboratory), Beijing 100080, Pepole's Republic of China, and Institute of Applied Mathematics, Academia Sinica, Beijing 100080, Pepole's Republic of China

Ke Wu
Institute of Theoretical Physics, Academia Sinica, Beijing 100080, People's Republic of China

(Received 21 July 1997; accepted for publication 5 January 1998)

We extend the study on the algebraic structure of the su(n) color Calogero–Sutherland (CS) models to the case of gl($n|m$) color CS model and show that the generators of the super-Yangian $Y(\mathrm{gl}(n|m))$ can be obtained from two gl($n|m$) loop algebras. Also, a super W_∞ algebra for the SUSY CS model is constructed. © *1998 American Institute of Physics.* [S0022-2488(98)01804-0]

I. INTRODUCTION

The Calogero–Sutherland (CS) models[1,2] are one-dimensional particle systems in which all particles interact through inverse square pairwise interactions. The systems are integrable in both classical and quantum cases and have many interesting properties (e.g., the particles have fractional statistics[3]).

There are various generalizations to the CS models. One of the examples is the su(n) color generalization in which particles have internal degrees of freedom with n possible values.[4] The su(n) extension of the CS models is also integrable and has a new symmetry—Yangian $Y(\mathrm{su}(n))$.[5,6] For the su(n) CS system, Hikami and Wadati[7–9] have made a distinction between the Calogero-type model and the Sutherland-type model and investigated in detail their integrability and algebraic structure. Hikami and Wadati found that the W_∞ algebra, which is the underlying symmetry of the quantum integrable system with long-range interactions, unifies the Yangian in the Sutherland-type model and the loop algebras in the Calogero-type model and revealed that Yangian is a subalgebra of the deformed W_∞ algebra.

Recently, Yangians associated with the simple Lie superalgebras, which are called super-Yangians, have been studied.[10–13] Based on the super-Yangian $Y(\mathrm{gl}(n|m))$, a supersymmetric extension of the su(n) color CS model is constructed and its integrability has been proved.[12] Now the question to be asked is whether the algebraic structure and some properties for the su(n) CS model discussed by Hikami, Wadati, and others still prevail in this supersymmetric extension. In this paper, we will make a study to the algebraic structure of the SUSY CS model.

The paper is organized as follows. In Sec. II, the definition of the super-Yangian $Y(\mathrm{gl}(n|m))$, the SUSY CS models, the Lax pairs for these models, and some notations are given. We then study in Sec. III the relations between the super-Yangian $Y(\mathrm{gl}(n|m))$ for the SUSY Sutherland-type model and two noncommuting loop algebras of which one is for the SUSY Calogero-type model. Moreover, we also derive a super W_∞ algebra from the Lax operators for the SUSY Calogero-type model. Finally, some discussions and remarks are made in Sec. IV.

II. SUPERSYMMETRIC CALOGERO–SUTHERLAND MODEL AND SUPER-YANGIAN $Y(\mathrm{gl}(n|m))$

We begin with the definition of the su(n) color Calogero-type and Sutherland-type models. Assume that there are N particles in these systems. Their positions are described by x_i $(i=1,\ldots,N)$

 © 1998 American Institute of Physics

and each particle carries a color degree of freedom labeled by index a, $a=1,\ldots,n$. The Hamiltonians for the su(n) color Calogero-type and Sutherland-type models are respectively defined as[4,7]

$$H_c=\frac{1}{2}\sum_i (\partial_i)^2+\frac{\lambda}{2}\sum_{i\neq j}(P_{ij}\partial_i\omega_{ij}+\lambda\omega_{ij}\omega_{ji}), \tag{II.1}$$

$$H_s=\frac{1}{2}\sum_i (x_i\partial_i)^2+\frac{\lambda}{2}\sum_{i\neq j}(P_{ij}+\lambda)\frac{x_ix_j}{(x_i-x_j)(x_j-x_i)}, \tag{II.2}$$

where

$$\omega_{ij}=\frac{1}{x_i-x_j}, \tag{II.3}$$

and P_{ij} is a permutation operator that interchanges the color degrees of freedom of the ith and jth particles. We denote by E_i^{ab}, $a,b=1,\ldots,n$, the matrices which act as $|a\rangle\langle b|$ on the color degrees of freedom of the ith particle and leave the other particles untouched. It is then straightforward to check that

$$P_{ij}=\sum_{a,b=1}^{n} E_i^{ab}E_j^{ba}. \tag{II.4}$$

The matrix unit E_i^{ab} can furnish a vector representation for the Lie algebra su(n).

We extend the su(n) color CS models to the gl($n|m$) case, that is, the particles in the systems now carry the graded color indices. Let V be an $n+m$ dimensional Z_2 graded vector space and $\{v^a, a=1,\ldots,n+m\}$ be a homogeneous basis with grading being defined as follows:

$$p(a)=\begin{cases}0, & a=1,\ldots,n\\ 1, & a=n+1,\ldots,n+m.\end{cases} \tag{II.5}$$

The graded matrix unit E^{ab} is defined as

$$E^{ab}v^c=\delta_{bc}v^a. \tag{II.6}$$

It is clear that we have a vector representation for the Lie superalgebra gl($n|m$) given by

$$[E^{ab},E^{cd}\}=\delta_{bc}E^{ad}-(-1)^{\eta}\delta_{ad}E^{cb}, \tag{II.7}$$

where $\eta=(p(a)+p(b))(p(c)+p(d))$ and the graded bracket is defined by

$$[E^{ab},E^{cd}\}\equiv E^{ab}E^{cd}-(-1)^{\eta}E^{cd}E^{ab}. \tag{II.8}$$

If we consider N copies of the matrix units E_i^{ab} ($i=1,\ldots,N$) that act on the ith space of the tensor product of graded vector spaces $V_1\otimes\cdots\otimes V_N$ with $V_i\cong V$, then we can show that the permutation operator P_{ij} defined as

$$P_{ij}=\sum_{a,b=1}^{n+m}(-1)^{p(b)}E_i^{ab}E_j^{ba} \tag{II.9}$$

exchanges the basis vectors v_i, v_j of i,j spaces and has the following properties:

$$P_{ij}=P_{ji},\quad P_{ij}P_{ij}=1,\quad P_{ij}P_{jk}=P_{ik}P_{ij}. \tag{II.10}$$

As before, we consider the one-dimensional system with N identical particles. x_i ($i=1,\ldots,N$) and index a denote, respectively, the positions and the color degrees of freedom of the particles, but now a is Z_2 graded as defined in Eq. (II.5) and takes values $a=1,\ldots,n+m$. If we replace P_{ij} in Eqs. (II.1) and (II.2) by that defined in Eq. (II.9), then we get two Hamiltonians that govern the

dynamics of the systems with graded internal degrees of freedom. In this sense, we call the models described by H_s and H_c SUSY Sutherland-type and SUSY Calogero-type models, respectively, similar to that discussed in Ref. 7.

The SUSY CS models are integrable. To prove the integrability of the systems, we may either directly construct the integrals of motion or find a Lax pair. In the former, the system is integrable if the integrals of motion commute among themselves and also commute with the Hamiltonian of the system; while in the latter, the model is integrable if two matrices L_{ij} and M_{ij} with operator entries obey

$$[H,L_{ij}]=\sum_k (L_{ik}M_{kj}-M_{ik}L_{kj}). \tag{II.11}$$

In Ref. 12, the integrability of the SUSY Sutherland-type model is guaranteed through the construction of the conserved quantities. Here we prove the integrability of SUSY CS models by giving the Lax pairs. For the SUSY Sutherland-type model, the Lax pair is given by

$$L_{ij}=\delta_{ij}\left(x_i\frac{\partial}{\partial x_i}+\frac{1}{2}\right)+\lambda(1-\delta_{ij})\theta_{ij}P_{ij}, \tag{II.12}$$

$$M_{ij}=-\delta_{ij}\lambda\sum_{k\neq i}\theta_{ik}\theta_{ki}P_{ik}+\lambda(1-\delta_{ij})\theta_{ij}\theta_{ji}P_{ij}; \tag{II.13}$$

where $\theta_{ij}=x_i/(x_i-x_j)$. While for the SUSY Calogero-type model, a possible choice is as follows:

$$L_{ij}=\delta_{ij}\frac{\partial}{\partial x_i}+\lambda(1-\delta_{ij})\omega_{ij}P_{ij}, \tag{II.14}$$

$$M_{ij}=-\delta_{ij}\lambda\sum_{k\neq i}\omega_{ik}\omega_{ki}P_{ik}+\lambda(1-\delta_{ij})\omega_{ij}\omega_{ji}P_{ij}. \tag{II.15}$$

M_{ij} in Eqs. (II.13) and (II.15) satisfies the conditions:

$$\sum_i M_{ij}=\sum_j M_{ij}=0, \tag{II.16}$$

which are the same as that in the su(n) color CS models[14] and are essential for the proof of the integrability of the models.

The SUSY Sutherland-type model possesses the super-Yangian $Y(\mathrm{gl}(n|m))$.[12] Let $\{T_p^{ab}, p\geqslant 0, a,b=1,\ldots,n+m\}$ be the generators of the super-Yangian $Y(\mathrm{gl}(n|m))$, then T_p^{ab} satisfies the following relations:[10–13]

$$[T_s^{ab},T_{p+1}^{cd}\}-[T_{s+1}^{ab},T_p^{cd}\}=\lambda(-1)^{(p(c)p(a)+p(c)p(b)+p(b)p(a))}(T_p^{cb}T_s^{ad}-T_s^{cb}T_p^{ad}) \tag{II.17}$$

for $s,p\geqslant -1$, where $T_{-1}^{ab}\equiv\lambda^{-1}\mathbf{1}\delta_{ab}$. It is easy to check that only T_0^{ab},T_1^{ab} are the basic operators, while all other T_p^{ab} $(p>1)$ are defined recursively from T_0^{ab},T_1^{ab}.[13] Therefore, we need to know the action of operators T_0^{ab},T_1^{ab} on the corresponding configuration space of the system in order to prove that the SUSY Sutherland-type model has symmetry $Y(\mathrm{gl}(n|m))$. The action of the operators T_0^{ab},T_1^{ab} on the configuration space of the system is of the following form:

$$T_0^{ab}=\sum_{i=1}^N E_i^{ab}, \tag{II.18}$$

$$T_1^{ab}=\sum_{i,j=1}^N E_i^{ab}L_{ij}, \tag{II.19}$$

where L_{ij} is given by Eq. (II.12). In addition, we also find that

$$T_p^{ab}=\sum_{i,j} E_i^{ab}(L^p)_{ij}, \quad p\geqslant 0 \tag{II.20}$$

satisfies the relation (II.17). It is straightforward to prove that generators T_0^{ab}, T_1^{ab} are conserved operators for the SUSY Sutherland-type model, that is,

$$[T_0^{ab}, H_s]=[T_1^{ab}, H_s]=0. \tag{II.21}$$

The SUSY Calogero-type model is a rational limit of the SUSY Sutherland-type model.[12] It only has gl$(n|m)$ loop algebra symmetry, which we will discuss in Sec. III.

III. THE ALGEBRAIC STRUCTURE OF THE SUSY CALOGERO–SUTHERLAND MODEL

In this section, we first discuss the defining relations of $Y(\mathrm{gl}(n|m))$ in terms of the action of T_0^{ab}, T_1^{ab} on the configuration space of the SUSY Sutherland-type model, then we study the loop algebra structure of the SUSY Calogero-type model and its relation with $Y(\mathrm{gl}(n|m))$. In the second part of this section, a super W_∞ algebra constructed from the Lax operator for the SUSY Calogero-type model is derived.

From Sec. II, we know that the SUSY Sutherland-type model possesses the super-Yangian $Y(\mathrm{gl}(n|m))$ generated by operators T_0^{ab}, T_1^{ab} in Eqs. (II.18) and (II.19). Using Eq. (II.7), it can be shown that operators T_0^{ab}, T_1^{ab} satisfy the following relations:

$$[T_0^{ab}, T_0^{cd}\}=\delta_{bc}T_0^{ad}-(-1)^{\eta}\delta_{da}T_0^{cb}, \tag{III.1}$$

$$[T_0^{ab}, T_1^{cd}\}=\delta_{bc}T_1^{ad}-(-1)^{\eta}\delta_{da}T_1^{cb}, \tag{III.2}$$

$$[T_1^{ab}, T_1^{cd}\}=\delta_{bc}T_2^{ad}-(-1)^{\eta}\delta_{da}T_2^{cb}-\lambda(-1)^{\beta}(T_0^{ad}T_1^{cb}-T_1^{ad}T_0^{cb}), \tag{III.3}$$

where $\beta=p(b)p(c)+p(c)p(d)+p(b)p(d)$ and T_2^{ab} can be written explicitly as

$$T_2^{ab}=\sum_i E_i^{ab}\left(x_i\partial_i+\frac{1}{2}\right)^2+\lambda\sum_{i\neq j}(E_iE_j)^{ab}(x_i\partial_i\theta_{ij}+\theta_{ij}(x_i\partial_i+x_j\partial_j+1))$$

$$+\lambda^2\sum_{i\neq j,j\neq k}(E_iE_jE_k)^{ab}\theta_{ij}\theta_{jk}. \tag{III.4}$$

Here we apply the conventional notations, $(E_iE_j)^{ab}=\sum_{c=1}^{n+m}(-1)^{p(c)}E_i^{ac}E_j^{cb}$, $(E_iE_jE_k)^{ab}=\sum_{c,d=1}^{n+m}(-1)^{p(c)+p(d)}E_i^{ac}E_j^{cd}E_k^{db}$. We can also obtain the following Serre-like relation for operators T_0^{ab}, T_1^{ab}:

$$[T_0^{ab},[T_1^{cd},T_1^{ef}\}\}-[T_1^{ab},[T_0^{cd},T_1^{ef}\}\}=\lambda(\delta_{bc}O_{ef}^{ad}-\delta_{de}O_{cf}^{ab}+(-1)^{\delta}\delta_{cf}O_{ed}^{ab}+(-1)^{\eta}\delta_{be}O_{af}^{cd}$$
$$-(-1)^{\eta}\delta_{ad}O_{ef}^{cb}-(-1)^{\gamma}\delta_{af}O_{eb}^{cd}), \tag{III.5}$$

with $\delta=(p(c)+p(d))(p(e)+p(f))$, $\gamma=(p(a)+p(b))(p(c)+p(d)+p(e)+p(f))$, and O_{cd}^{ab} being defined as

$$O_{cd}^{ab}=-(-1)^{\beta}(T_0^{ad}T_1^{cb}-T_1^{ad}T_0^{cb}). \tag{III.6}$$

Equations (III.1)–(III.3) and (III.5) are the defining relations for the super-Yangian $Y(\mathrm{gl}(n|m))$.

For the SUSY Calogero-type model (II.1), we introduce two operators:

$$J_0^{ab}=T_0^{ab}, \tag{III.7}$$

$$J_1^{ab}=\sum_{i,j} E_i^{ab}I_{ij}, \tag{III.8}$$

where

$$I_{ij}=\delta_{ij}\frac{\partial}{\partial x_i}+\lambda(1-\delta_{ij})\omega_{ij}P_{ij} \tag{III.9}$$

is a Lax operator given in Eq. (II.14). Here we use another notation I_{ij} for L_{ij}. It is easy to check that J_0^{ab},J_1^{ab} satisfy the following commutation relations:

$$[J_0^{ab},J_1^{cd}\}=\delta_{bc}J_1^{ad}-(-1)^{\eta}\delta_{da}J_1^{cb}, \tag{III.10}$$

$$[J_1^{ab},J_1^{cd}\}=\delta_{bc}J_2^{ad}-(-1)^{\eta}\delta_{da}J_2^{cb}, \tag{III.11}$$

$$[J_0^{ab},[J_1^{cd},J_1^{ef}\}\}-[J_1^{ab},[J_0^{cd},J_1^{ef}\}\}=0, \tag{III.12}$$

where

$$J_2^{ab}=\sum_{i,j}E_i^{ab}(I^2)_{ij}. \tag{III.13}$$

In general, we have

$$J_p^{ab}=\sum_{i,j}E_i^{ab}(I^p)_{ij}\quad(p\geqslant 0), \tag{III.14}$$

$$[J_s^{ab},J_p^{cd}\}=\delta_{bc}J_{s+p}^{ad}-(-1)^{\eta}\delta_{da}J_{s+p}^{cb}. \tag{III.15}$$

Equation (III.12) is the Serre relation for the gl$(n|m)$ loop algebra. We can show that J_p^{ab} are conserved operators for the SUSY Calogero-type model, that is, the SUSY Calogero-type model has the gl$(n|m)$ loop algebra symmetry. Furthermore, there is another representation of the gl$(n|m)$ loop algebra with the generators K_p^{ab} being defined by

$$K_p^{ab}=\sum_{i=1}^{N}E_i^{ab}x_i^p\quad(p\geqslant 0), \tag{III.16}$$

and having the relations

$$[K_s^{ab},K_p^{cd}\}=\delta_{bc}K_{s+p}^{ad}-(-1)^{\eta}\delta_{da}K_{s+p}^{cb}, \tag{III.17}$$

$$[K_0^{ab},[K_1^{cd},K_1^{ef}\}\}-[K_1^{ab},[K_0^{cd},K_1^{ef}\}\}=0. \tag{III.18}$$

The K_0^{ab},K_1^{ab} are the two basic operators with the following forms, respectively:

$$K_0^{ab}=J_0^{ab}, \tag{III.19}$$

$$K_1^{ab}=\sum_{i}E_i^{ab}x_i. \tag{III.20}$$

We consider now the algebras constructed from the elements $\{J_0^{ab},J_1^{ab},K_1^{ab}\}$. Just as in the nongraded case,[7,15] the super-Yangian $Y(\mathrm{gl}(n|m))$ appears by unifying two gl$(n|m)$ loop algebras $\{J_0^{ab},J_1^{ab}\}$ and $\{K_0^{ab},K_1^{ab}\}$, that is to say, T_1^{ab} can be obtained from the commutators between two sets of generators $\{J_0^{ab},J_1^{ab}\}$ and $\{K_0^{ab},K_1^{ab}\}$:

$$[J_1^{ab},K_1^{cd}\}+[K_1^{ab},J_1^{cd}\}-\lambda[J_0^{ab},J_0^{cd}\}+\lambda[J_0^{ab},(J_0J_0)^{cd}\}=2(\delta_{bc}T_1^{ad}-(-1)^{\eta}\delta_{da}T_1^{cb}). \tag{III.21}$$

In this sense, the algebra generated by $\{J_0^{ab},J_1^{ab},K_1^{ab}\}$ is a larger algebra including both super-Yangian $Y(\mathrm{gl}(n|m))$ and two loop algebras. In addition, we have two extra Serre-like relations:

$$
\begin{aligned}
&[J_0^{ab},[J_1^{cd}+T_1^{cd},J_1^{ef}+T_1^{ef}\}\}-[J_1^{ab}+T_1^{ab},[J_0^{cd},J_1^{ef}+T_1^{ef}\}\}\\
&\quad=\lambda(\delta_{bc}M_{ef}^{ad}-\delta_{de}M_{cf}^{ab}+(-1)^{\delta}\delta_{cf}M_{ed}^{ab}\\
&\qquad+(-1)^{\eta}\delta_{be}M_{af}^{cd}-(-1)^{\eta}\delta_{ad}M_{ef}^{cb}-(-1)^{\gamma}\delta_{af}M_{eb}^{cd}),
\end{aligned}
\tag{III.22}
$$

$$
\begin{aligned}
&[J_0^{ab},[K_1^{cd}+T_1^{cd},K_1^{ef}+T_1^{ef}\}\}-[K_1^{ab}+T_1^{ab},[J_0^{cd},K_1^{ef}+T_1^{ef}\}\}\\
&\quad=\lambda(\delta_{bc}N_{ef}^{ad}-\delta_{de}N_{cf}^{ab}+(-1)^{\delta}\delta_{cf}N_{ed}^{ab}\\
&\qquad+(-1)^{\eta}\delta_{be}N_{af}^{cd}-(-1)^{\eta}\delta_{ad}N_{ef}^{cb}-(-1)^{\gamma}\delta_{af}N_{eb}^{cd}),
\end{aligned}
\tag{III.23}
$$

where

$$
\begin{aligned}
M_{cd}^{ab}=(-1)^{\beta}&\left[\sum_{i\neq j}E_i^{ad}E_j^{cb}\left((x_i+1)\frac{\partial}{\partial x_i}-(x_j+1)\frac{\partial}{\partial x_j}\right)+\lambda\sum_{ijk}{}'\ \frac{x_i+1}{x_i-x_j}(E_iE_j)^{ad}E_k^{cb}\right.\\
&\left.-\lambda\sum_{ijk}{}'\ \frac{x_j+1}{x_j-x_k}E_i^{ad}(E_jE_k)^{cb}\right]+\lambda\sum_{i\neq j}\ \frac{x_i+x_j+2}{x_i-x_j}E_i^{ab}E_j^{cd},
\end{aligned}
\tag{III.24}
$$

$$
\begin{aligned}
N_{cd}^{ab}=(-1)^{\beta}&\left[\sum_{i\neq j}E_i^{ad}E_j^{cb}\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}+x_i-x_j\right)+\lambda\sum_{ijk}{}'\ \frac{1}{x_i-x_j}(E_iE_j)^{ad}E_k^{cb}\right.\\
&\left.-\lambda\sum_{ijk}{}'\ \frac{1}{x_j-x_k}E_i^{ad}(E_jE_k)^{cb}\right]+\lambda\sum_{i\neq j}\ \frac{2}{x_i-x_j}E_i^{ab}E_j^{cd}
\end{aligned}
\tag{III.25}
$$

and Σ' means that any two summation indices do not coincide. The Serre-like relations (III.5), (III.22), and (III.23) may be written in a more compact form. To do this, we define the operator $Q_1^{ab}(x,y)$ to be of the following form:

$$
Q_1^{ab}(x,y)\equiv T_1^{ab}+xJ_1^{ab}+yK_1^{ab},
\tag{III.26}
$$

where x and y are two complex numbers. Using the operator $Q_1^{ab}(x,y)$, we get the following relation:

$$
\begin{aligned}
&[J_0^{ab},[Q_1^{cd}(x,y),Q_1^{ef}(x,y)\}\}-[Q_1^{ab}(x,y),[J_0^{cd},Q_1^{ef}(x,y)\}\}\\
&\quad=\lambda(\delta_{bc}P_{ef}^{ad}(x,y)-\delta_{de}P_{cf}^{ab}(x,y)+(-1)^{\delta}\delta_{cf}P_{ed}^{ab}(x,y)\\
&\qquad+(-1)^{\eta}\delta_{be}P_{af}^{cd}(x,y)-(-1)^{\eta}\delta_{ad}P_{ef}^{cb}(x,y)-(-1)^{\gamma}\delta_{af}P_{eb}^{cd}(x,y)),
\end{aligned}
\tag{III.27}
$$

where

$$
\begin{aligned}
P_{cd}^{ab}(x,y)=O_{cd}^{ab}+x\left\{(-1)^{\beta}\left[\sum_{i\neq j}E_i^{ad}E_j^{cb}\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}\right)+\lambda\sum_{ijk}{}'\frac{1}{x_i-x_j}(E_iE_j)^{ad}E_k^{cb}\right.\right.\\
\left.\left.-\lambda\sum_{ijk}{}'\frac{1}{x_j-x_k}E_i^{ad}(E_jE_k)^{cb}\right]+\sum_{i\neq j}\ \frac{2}{x_i-x_j}E_i^{ab}E_j^{cd}\right\}+y(-1)^{\beta}\sum_{i\neq j}E_i^{ad}E_j^{cb}(x_i-x_j).
\end{aligned}
\tag{III.28}
$$

From Eq. (III.27), we see that generators $Q_1^{ab}(x,y)$ also form a representation of the super-Yangian $Y(\mathrm{gl}(n|m))$ for any x and y. Therefore, we can get a family of super-Yangian subalgebras from generators $Q_1^{ab}(x,y)$, all these subalgebras have a common deformation parameter λ.

Using the generators J_p^{ab} and K_p^{ab}, we can also construct another algebra. We introduce two set of operators as

$$
\mathscr{W}_p^{(s)}=\frac{1}{2(p+s)}\left[\sum_i x_i^2,\mathscr{W}_{p+2}^{(s-1)}\right],
\tag{III.29}
$$

$$\mathcal{Q}_p^{(s)ab}=\frac{1}{2(p+s)}\left[\sum_i x_i^2,\mathcal{Q}_{p+2}^{(s-1)ab}\right],\quad s=2,3,\ldots, \tag{III.30}$$

where $\mathcal{W}_p^{(1)}=J_p=\Sigma_{ij}(I^p)_{ij}$, $\mathcal{Q}_p^{(1)ab}=J_p^{ab}$. $\mathcal{W}_p^{(s)}$ and $\mathcal{Q}_p^{(s)ab}$ can be considered to have conformal spin s. We can write $\mathcal{W}_p^{(s)}$ and $\mathcal{Q}_p^{(s)ab}$ explicitly as

$$\mathcal{W}_p^{(s)}=\frac{1}{2^{s-1}(p+s)_{s-1}}[\underbrace{\sum_i x_i^2,[\cdots,[\sum_i x_i^2}_{s-1},J_{p+2s-2}]\cdots]]=\sum_j(-x_j)^{s-1}(\partial_{x_j})^{p+s-1}+\cdots, \tag{III.31}$$

$$\mathcal{Q}_p^{(s)ab}=\frac{1}{2^{s-1}(p+s)_{s-1}}[\underbrace{\sum_i x_i^2,[\cdots,[\sum_i x_i^2}_{s-1},J_{p+2s-2}^{ab}]\cdots]]=\sum_j E_j^{ab}(-x_j)^{s-1}(\partial_{x_j})^{p+s-1}+\cdots, \tag{III.32}$$

where $(p)_s$ denotes

$$(p)_s=p(p+1)\cdots(p+s-1), \tag{III.33}$$

and $(\cdots)$ represents the lower order terms of ∂_{x_j} and is not necessarily zero in the limit $\lambda\to 0$. Similar to the proofs by Hikami and Wadati,[7] we can show that $\mathcal{W}_p^{(s)}$ and $\mathcal{Q}_p^{(s)ab}$ satisfy the following commutation relations:

$$[\mathcal{W}_p^{(s)},\mathcal{W}_q^{(s')}]=((s-1)q-(s'-1)p)\mathcal{W}_{p+q}^{(s+s'-2)}+\cdots, \tag{III.34}$$

$$[\mathcal{W}_p^{(s)},\mathcal{Q}_q^{(s')ab}]=((s-1)q-(s'-1)p)\mathcal{Q}_{p+q}^{(s+s'-2)}+\cdots, \tag{III.35}$$

$$[\mathcal{Q}_p^{(s)ab},\mathcal{Q}_q^{(s')cd}\}=(\delta_{bc}\mathcal{Q}_{p+q}^{(s+s'-1)ad}-(-1)^{\eta}\delta_{ad}\mathcal{Q}_{p+q}^{(s+s'-1)cb})+\cdots, \tag{III.36}$$

where $(\cdots)$ includes lower spin operators. Equation (III.34) shows that the operators $\mathcal{W}_p^{(s)}$ constitute the quantum W_∞ algebra as in the CS models.[7] However, Eqs. (III.35) and (III.36) indicate that algebra generated by operators $\mathcal{W}_p^{(s)}$ and $\mathcal{Q}_p^{(s)ab}$ is a kind of super W_∞ algebra with color.[16]

In the case of $\lambda\to 0$, if we define a operator which is of the following form,

$$\mathcal{Q}_p^{(s)ab}=\sum_{i=1}^{N} E_i^{ab}x_i^{s-1}(\partial_{x_i})^{p+s-1}, \tag{III.37}$$

that is we only consider the first term of $\mathcal{Q}_p^{(s)ab}$ in Eq. (III.32) up to a constant factor, then it is easy to show that $\mathcal{Q}_p^{(s)ab}$ satisfy the commutation relations:

$$[\mathcal{Q}_p^{(s)ab},\mathcal{Q}_q^{(s')cd}\}=\delta_{bc}\cdot\sum_{k=0}^{p+s-1}\frac{(p+s-1)!(s'-1)!}{k!(p+s-k-1)!(s'-k-1)!}\mathcal{Q}_{p+q}^{(s+s'-1-k)ad}-(-1)^{\eta}\delta_{da}$$
$$\cdot\sum_{k=0}^{q+s'-1}\frac{(q+s'-1)!(s-1)!}{k!(q+s'-k-1)!(s-k-1)!}\mathcal{Q}_{p+q}^{(s+s'-1-k)cb}. \tag{III.38}$$

It can be seen that the algebra generated by $\mathcal{Q}_p^{(s)ab}$ in Eq. (III.37) is a kind of super W_∞ algebra without central terms.[16]

IV. REMARKS AND DISCUSSIONS

In this paper, we have discussed the algebraic structure of the SUSY CS models. We know that SUSY CS models have super-Yangian $Y(\mathrm{gl}(n|m))$ symmetry, and we give the realization of $Y(\mathrm{gl}(n|m))$ and its relations with two $\mathrm{gl}(n|m)$ loop algebras. Also, a super W_∞ algebra realized in terms of generators $\mathcal{W}_p^{(s)}$ and $\mathcal{Q}_p^{(s)ab}$ is shown. Compared with the CS models, there also exists a larger symmetry generated by two loop algebras for the SUSY CS models. But, there are three points to be remarked. First, the operator L_{ij} [Eq. (II.14)] is different from the one for the Sutherland-type model in Refs. 7 and 15, so that the right-hand sides of the Serre-like relations

(III.22) and (III.23) are nonzero. Second, in the right-hand side of Eq. (III.27), there is an x,y dependence which is absent in the nongraded CS models.[7,15] Third, the W_∞ algebra is graded with respect to the color indices a,b.

For the su(n) CS models, the problem of energy spectrum and the symmetry of the system confined in a harmonic potential have been studied in Refs. 7, 8, and 17. Because of the above similarity between SUSY CS models and the CS models, these problems for the SUSY CS models shall not be discussed further.

ACKNOWLEDGMENTS

One of the authors (G. J.) thanks Dr. Wei-Zhong Zhao for useful discussions, and also acknowledges Dr. K. Hikami for providing his publications. We gratefully acknowledge the referee for drawing our attention to Refs. 14 and 17.

[1] F. Calogero, J. Math. Phys. **10**, 2191 (1969); **10**, 2197 (1969).
[2] B. Sutherland, J. Math. Phys. **12**, 246 (1971); **12**, 251 (1971).
[3] F. D. M. Haldane, Phys. Rev. Lett. **67**, 937 (1991); D. Bernard (unpublished).
[4] J. A. Minahan and A. Polychronakos, Phys. Lett. B **302**, 265 (1993).
[5] D. Bernard, M. Gaudin, F. D. M. Haldane, and V. Pasquier, J. Phys. A **26**, 5219 (1993).
[6] V. G. Drinfeld, Sov. Math. Dokl. **32**, 245 (1985).
[7] K. Hikami and M. Wadati, J. Phys. Soc. Jpn. **62**, 4203 (1993).
[8] K. Hikami and M. Wadati, Phys. Rev. Lett. **73**, 1191 (1994).
[9] K. Hikami, J. Phys. A **28**, L131 (1995).
[10] M. L. Nazarov, Lett. Math. Phys. **21**, 123 (1991).
[11] R. B. Zhang, Lett. Math. Phys. **33**, 263 (1995); **37**, 419 (1996); J. Math. Phys. **36**, 3854 (1995).
[12] C. Ahn and W. M. Koo, Phys. Lett. B **365**, 105 (1996).
[13] G. X. Ju, J. F. Cai, H. Y. Guo, K. Wu, and S. K. Wang, J. Phys. A **30**, 6155 (1997).
[14] H. Ujino, K. Hikami, and M. Wadati, J. Phys. Soc. Jpn. **61**, 3425 (1992).
[15] D. Bernard, K. Hikami, and M. Wadati (unpublished).
[16] E. Bergshoeff, C. N. Pope, L. J. Romans, E. Sezgin, and X. Shen, Phys. Lett. B **245**, 447 (1990).
[17] H. Ujino and M. Wadati, J. Phys. Soc. Jpn. **64**, 39 (1995).

Nonalgebraic integrability of one reversible dynamical system of the Cremona type

K. V. Rerikh[a)]
Bogoliubov Laboratory of Theoretical Physics, JINR, 141980, Dubna, The Moscow Region, Russia

(Received 26 January 1996; accepted for publication 19 December 1997)

A reversible dynamical system (RDS) and a system of nonlinear functional equations, defined by a certain rational quadratic Cremona mapping and arising from the static model of the dispersion approach in the theory of strong interactions [the Chew–Low-type equations with crossing-symmetry matrix $A(l,1)$], are considered. This RDS is split into one- and two-dimensional ones. An explicit Cremona transformation that completely determines the exact solution of the two-dimensional system is found. This solution depends on an odd function satisfying a nonlinear autonomous three-point functional equation. Nonalgebraic integrability of RDS under consideration is proved using the method of Poincaré normal forms and the Siegel theorem on biholomorphic linearization of a mapping at a nonresonant fixed point. © *1998 American Institute of Physics.* [S0022-2488(98)01105-0]

I. INTRODUCTION

Reversible dynamical systems (RDS) are qualitatively very similar to Hamiltonian systems.[1–9] In particular, the existence of the Kolmogorov–Arnold–Moser tori in reversible non-Hamiltonian flows,[1–6] and nonsymplectic mappings,[4–7] has been proved. Reversible dynamical systems are of great importance in physics, which is a consequence of the time-reversible invariance of many physical laws (see surveys of Ref. 9).

One example of RDS is the system of nonlinear functional (difference) equations[10] of the static model of the dispersion approach.[11] This system is defined by a reversible mapping that is the composition of two involutions: the standard Cremona transformation[12] and the crossing-symmetry mapping. Dynamical systems (DS) of this type, distinguished only by the crossing-symmetry matrix A and defined by the quadratic Cremona mappings, have been investigated in Refs. 13–22. Interest in integrating these DS is related to besides the physical one, with a problem of integrating general Cremona mappings which, according to the M. Noether theorem, can be constructed from different quadratic Cremona mappings.[12] Moreover, such DS do not belong to the Quispel 18-parametric family of integrable mappings,[9,23] and were not considered in the survey on integrable maps.[24] The problem of integrating these two-dimensional DS with the second-order matrix A was solved long ago in Refs. 14 and 15, but only recently three-dimensional DS with the third-order matrices $A(1,1)$ and $A^{\text{Chew–Low}}$ have been integrated in Refs. 17 and 18. A general approach to integrating such n-dimensional DS was developed in Ref. 19.

This paper is devoted to the problem of integrating three-dimensional dynamical system with crossing-symmetry matrix $A(l,1)$, describing the scattering of two particles with spins equal l and 1 within the framework of the static model. This dynamical system is represented by the following system of functional equations:

$$S_i(-w)=\sum_{j=1}^{3} A_{ij}S_j(w), \quad i,j\in(1,2,3),$$

$$S_i(w+1)=1/\sum_{j=1}^{3} A_{ij}S_j(w),$$

[a)]Electronic mail: rerikh@thsun1.jinr.dubna.su

 © 1998 American Institute of Physics

where $S_i(w)$ are the S-matrix elements in the states $i(l-1,l,l+1)$ and $A_{i,j}$ are the elements of the crossing-symmetry matrix $A(l,1)$. Solutions to the system should be meromorphic functions of uniformizing variable $w=\pi^{-1}\arcsin\omega$, in which ω is the energy of a light particle scattered by heavy (static) one. (See, for example, Refs. 10 and 11 about derivation of equations of this type from the dispersion relations.)

Investigating integrability of the dynamical system at the value of $l=1$ in Ref. 17, we see that it splits into the one-dimensional and two-dimensional systems; moreover, the last system does not have a rational first integral and is not algebraically integrable but has a holomorphic first integral. It is obvious that at an arbitrary value of l the dynamical system is not algebraically integrable as well. Let us give the definition of the algebraic and nonalgebraic integrability.[25]

Definition 1: Let $f(x)$ be a birational (inverse map is also rational) mapping $C^2\to C^2$, $x\in C^2$, and consider an autonomous dynamical system in C^2 with discrete time

$$x_i(n+1)=f_i(x_1(n),x_2(n)),\quad i=(1,2), \tag{1}$$

and a system of autonomous functional equations for functions $x_i(w)$, $x\in C^2$, $w\in C$, $i=(1,2)$:

$$x_i(w+1)=f_i(x_1(w),x_2(w)),\quad i=(1,2). \tag{2}$$

Let us call the systems (1) and (2) algebraically integrable, if there exists a mapping $C^2\to C$ defined by the ratio $I_\mu(x)=g_\mu(x)/h_\mu(x)$ of two polynomials of degree μ, which is invariant with respect to the change $n\to n+1$ or $w\to w+1$:

$$I_\mu(x(n+1))=I_\mu(x(n)),\quad I_\mu(x(w+1))=I_\mu(x(w)).$$

Then, the equations

$$I_\mu(x(n))=c,\quad c=\text{const},\quad I_\mu(x(w))=\delta(w),\quad \delta(w+1)=\delta(w), \tag{3}$$

defining the level lines of the first integral or invariant of dynamical systems (1) and (2), give the one-parameter family or pencil of algebraic curves of degree μ due to a rationality $I_\mu(x)$. Otherwise, if there exists an invariant mapping $I(x)$, $I\circ f(x)=I(x)$,: $C^2\to C$, being holomorphic in a certain neighbotrhood of a fixed point of the mapping $f(x)$, let us call the systems (1) and (2) nonalgebraically integrable.

II. THE GENERAL FORMULATION OF THE PROBLEM. STRUCTURE OF THE GENERAL SOLUTION

Definition 2: Let X be an arbitrary set. A one-to-one mapping $T:X\to X$ is said to be reversible if there exists another mapping $G:X\to X$ for which $T^{-1}=G\circ T\circ G$ and G is an involution: $G^2=\mathrm{id}$.[6,8,9]

These conditions imply that $T\circ G$ is also an involution and $T=(T\circ G)\circ G$ is the composition of two involutions. Conversely, the composition of any two involutions is reversible with respect to each of them.

Consider the following two involutions of the complex plane $\mathbf{C}$: the linear one,

$$A_1:\mathbf{C}\to\mathbf{C},\quad A_1:w\mapsto -w,$$

and the affine one,

$$I_1:\mathbf{C}\to\mathbf{C},\quad I_1:w\mapsto 1-w.$$

These involutions do not commute. Their compositions $A_1\circ I_1$ and $I_1\circ A_1$ define two reciprocal reversible dynamical systems on $\mathbf{C}$.

On the other hand, consider the following two involutions of the space $\mathbf{C}^3$: the linear one, defined by the crossing symmetry involutive matrix $A\overset{\text{def}}{=}A(l,1)$,

$$z\mapsto Az\,,\quad A=\begin{pmatrix} 1-\dfrac{(2l-1)(l+1)}{(2l+1)l} & -\dfrac{1}{l} & \dfrac{2l+3}{2l+1} \\ -\dfrac{2l-1}{(2l+1)l} & 1-\dfrac{1}{l(l+1)} & \dfrac{2l+3}{(2l+1)(l+1)} \\ \dfrac{2l-1}{2l+1} & \dfrac{1}{l+1} & 1-\dfrac{(2l+3)l}{(2l+1)(l+1)} \end{pmatrix}, \tag{4}$$

with $\det A=-1$, and the nonlinear one

$$I:\mathbf{C}^3\to\mathbf{C}^3,\quad I(z_1,z_2,z_3)=\left(\frac{1}{z_1},\frac{1}{z_2},\frac{1}{z_3}\right)$$

(the latter being the standard Cremona transformation). These involutions do not commute either. Their compositions $A\circ I$ and $I\circ A$ define a two reciprocal reversible dynamical systems on $\mathbf{C}^3$. Let us remark that the parameter $l\in\mathbf{N}$, though one can consider $l\in\mathbf{R}$.

We look for a meromorphic mapping $\mathbf{C}\to\mathbf{C}^3$ which realizes the equivalence of these two reversible dynamical systems. In other words, we seek a meromorphic mapping $S:\mathbf{C}\to\mathbf{C}^3$ satisfying the following system of functional equations (FE):

$$S(-w)=AS(w), \tag{5}$$

$$S(w+1)=I\circ AS(w), \tag{6}$$

where $w\in\mathbf{C}$. This system describes a scattering of two particles with spins equal l and 1 within the framework of the static model.

Taking liberty with speech, we will sometimes call the system of FE (5) and (6) a dynamical system (DS).

One can say that the involutions A and I of $\mathbf{C}^3$ are representations of the involutions A_1 and I_1 of $\mathbf{C}$, respectively. Meromorphic mappings $S:\mathbf{C}\to\mathbf{C}^3$ satisfying FE (5) and (6), realize these representations.

Each meromorphic mapping $S:\mathbf{C}\to\mathbf{C}^3$ defines some curve in $\mathbf{C}^3$. The system of FE (5) and (6) completely determines this curve up to the group of automorphisms. One can easily verify that every transformation of the form

$$S(w)\to S(w+\beta(w))\exp((\alpha(w-1/2)), \tag{7}$$

where

$$\beta(w+1)=\beta(w),\quad \beta(-w)=-\beta(w),$$

$$\alpha(w+1)=-\alpha(w),\quad \alpha(-w)=-\alpha(w),$$

is an automorphism of the solution space for DS (5) and (6) (β defining an inner automorphism of the curve). Obviously, there should also exist an automorphism which depends on another arbitrary function in w with a period equal to 1.

As $A^2=\mathrm{id}$, the matrix A is of a simple structure (i.e., diagonalizable):

$$A=B\Lambda^a B^{-1},\quad \Lambda^a=\mathrm{diag}(\lambda_1^a,\lambda_2^a,\lambda_3^a),\quad \lambda_1^a=-1,\quad \lambda_{2,3}^a=1, \tag{8}$$

$$B_{ij}=\mu_i^{(j)},\quad A\mu^{(j)}=\lambda_j^a\mu^{(j)},\quad i,j=1,2,3, \tag{9}$$

where B is the fundamental matrix for A in the basis of eigenvectors $\mu^{(j)}$ of the matrix A,[26] which are the following:

$$\mu^{(1)}=(-2(l+1),-2,2l),\quad \mu^{(2)}=\frac{(2l+1)(2l-1)}{4}\left(\frac{2l+3}{2l-1},-\frac{2l+3}{2l+1},1\right),\quad \mu^{(3)}=(1,1,1). \tag{10}$$

According to the general approach,[19] introduce the functions $z:\mathbf{C}\to\mathbf{C}^3$, $j:\mathbf{C}^3\to\mathbf{C}^3$, $\pi:\mathbf{C}^3\to\mathbf{C}$,

$$z(w)=B^{-1}S(w), \tag{11}$$

$$j(z)=ABz, \quad \pi(z)=\prod_{i=1}^{3} j_i(z). \tag{12}$$

Define the mapping $\Phi:\mathbf{C}^3\to\mathbf{C}^3$:

$$\Phi: z\mapsto\widetilde{z}=\Phi(z)=\phi(z)/\pi(z), \tag{13}$$

where $\phi:\mathbf{C}^3\to\mathbf{C}^3$ is the mapping

$$\phi(z)=\pi(z)B^{-1}(Ij(z)). \tag{14}$$

The DS (5) and (6) can be rewritten as

$$z(-w)=\Lambda^a z(w), \tag{15}$$

$$z(w+1)=\phi(z(w))/\pi(z(w)). \tag{16}$$

Remark 1: Note that the mapping (13) is a birational Cremona transformation in $\mathbf{C}^3$. Transformations of the form (13) with different denominators $\pi(z)$ [but the same numerator $\phi(z)$] induce the same projective Cremona transformation in $\mathbf{C}P^2$:[12]

$$\widetilde{z}_1:\widetilde{z}_2:\widetilde{z}_3=\phi_1:\phi_2:\phi_3.$$

According to (8)–(12), and taking into account that $A^2=\mathrm{id}$, we have

$$\Lambda^a=B^{-1}AB, \quad S(w)=j(\Lambda^a z(w))\overset{\text{def}}{=}j'(z(w)). \tag{17}$$

A. Partial automorphic forms

Let $m=(m_1,m_2,m_3)\in\mathbf{Z}_+^3$ be a multiindex, $m_i\in\mathbf{Z}_+$, $|m|=m_1+m_2+m_3$, where $\mathbf{Z}_+$ is a set of non-negative integral numbers.

Definition 3: A polynomial $P:\mathbf{C}^3\to\mathbf{C}$ is said to be invariant if $P(\Phi(z))=0$ whenever $P(z)=0$.

Theorem 1:[19] For any invariant irreducible homogeneous polynomial $P(z)$ of degree k there exists a multiindex m with $|m|=k$ such that $P(z)$ satisfies the following system of FE:

$$P(\phi(z))=\varepsilon P(z)j^m(z), \quad j^m=j_1^{m_1}j_2^{m_2}j_3^{m_3}, \quad \varepsilon=\pm 1, \tag{18}$$

$$P(\Lambda^a z)=\nu P(z), \quad \nu=\pm 1. \tag{19}$$

We will denote any solution of these equations by $P_m(z)$ provided that it is an irreducible homogeneous polynomial of degree $|m|$.

As the matrix Λ (4) has the only eigenvalue equal to -1, we can assume that $\nu=+1$ because for $\nu=-1$ any solution P of FE (18) and (19) will be reducible:

$$P(z)=z_1^{2n+1}\hat{P}(z), \quad \deg\hat{P}=|m|-2n-1.$$

Following Ref. 19, we shall call polynomials $P_m(z)$ the partial automorphic forms (PAF) for the DS (5) and (6) of weight $m=(m_1,m_2,m_3)$.

Definition 4: PAF $P_m(z)$ is said to be an automorphic form (AF) of weight q if

$$P_m(\Phi(z))\equiv P_m(z)J^q(z),$$

where J denotes the Jacobian of the mapping Φ (13).

One can easily verify that the Jacobian J of the mapping Φ (13) is equal to $\pi^{-2}(z)$. Indeed,

$$J=-\det\,(B^{-1}\overline{J}B\Lambda^a)=-\det\,(A\overline{J})=\pi^{-2}(z),$$

where $\overline{J}=\mathrm{diag}\,(1/j_1^2(z),\ 1/j_2^2(z),\ 1/j_3^2(z))$.

A mapping Φ has a polynomial AF of weight q if and only if it possesses a PAF of weight $m=\{q,q,q\}$. It has a rational AF of weight q if there exists a rational function of the PAF that satisfies equations (18) and (19) with $m_1=m_2=m_3=q$.

According to Ref. 17, the DS (5) and (6) with the crossing-symmetry matrix $A(1,1)$ $(l=1)$ has an AF of weight one (recall that it has three PAF) and is equivalent to an APM but it is non-algebraically integrable, i.e., it possesses a first nonalgebraic integral. From this it follows that the DS (5) and (6) with matrix (4) is not also algebraic integrable. The one has the PAF $(z_1/z_3=x,\ z_2/z_3=y)$:

$$P_{(1,1,0)}(z)=z_3z_2-z_1^2=z_3^2(y-x^2)$$

—the solution of (18) with $\varepsilon=-1$.

B. Fundamental points and principal lines of a quadratic Cremona mapping

Let $\widetilde{z}=z(w+1)$. Then, according to (17), (16), and (12) we have

$$j'(\widetilde{z})=Ij(z).$$

Consider some special points and lines on the projective plane $\mathbf{C}P^2$ of homogeneous coordinates z known from the theory of quadratic Cremona transformations. Introduce six principal lines (P-lines) J_k, J_k' $(k=1,2,3)$ of the direct mapping Φ specified by (13) and the inverse one Φ^{-1}:

$$J_k=\{z:j_k(z)=0\},\quad J_k'=\{z:j_k'(z)=0\}\quad (k=1,2,3),$$

and six points of pairwise intersections of these lines $\{O_1,O_2,O_3\}$ and $\{O_1',O_2',O_3'\}$:

$$O_i=J_k\cap J_n,\quad O_i'=J_k'\cap J_n'.$$

They are called fundamental points (F-points)[12] of the direct mapping (16) and the inverse one. Here $\{ikn\}$ is a permutation of $\{123\}$.

Under the direct mapping (16) and its inverse

$$z(w)=\frac{\Lambda^a\phi(\Lambda^a\widetilde{z})}{\pi(\Lambda^a\widetilde{z})},$$

the images of the F-points O_i, O_i' and those of the P-lines J_i, J_i' are

$$\Phi:O_i\mapsto J_i',\quad \Phi:J_i\mapsto O_i',\quad \Phi^{-1}:O_i'\mapsto J_i,\quad \Phi^{-1}:J_i'\mapsto O_i.$$

Thus, the F-points and P-lines $\{O_i,J_i\}$ and $\{O_i',J_i'\}$ are the only elements of the mapping (16) where the one-to-one correspondence is violated.

Notice that the mappings Φ and Φ^{-1} realize blowing up and contraction (see the Codaira theorem, Ref. 27, Ch. 1, §4) and also the concept of σ-process of blowing up of singularities in the theory of ordinary differential equations (Ref. 28, §2).

The F- and P-elements have the form

$$J_i=\{z:j_i(z)=\mu_i^{(k)}\lambda_k^{(a)}z_k=0\},\quad J_i'=\{z:j_i'(z)=\mu_i^{(k)}z_k=0\},$$

$$O_1=\left(\frac{2}{2l-1},\frac{4}{(2l-1)^2},1\right),\quad O_2=\left(\frac{2}{4l^2+4l-1},-\frac{4}{4l^2+4l-1},1\right),\tag{20}$$

$$O_3=\left(-\frac{2}{2l+3},\frac{4}{(2l+3)^2},1\right),\quad O_i'=\Lambda^a O_i\,.$$

Remark 2: Similarly to Refs. 17 and 21, the multiindex $m=(m_1,m_2,m_3)$ has the meaning of multiplicities m_i of the algebraic invariant curve $P_m(z)=0$ at the F-points O_i [see (18) and (19)], and, consequently, at the points O_i', according to the symmetry of (19).

C. Structure of the general solution

Let us proceed from homogeneous projective coordinates $z(w)\in\mathbf{C}^3$ to the coordinates

$$(x_1(w),x_2(w),z_3(w)),\quad x_i=\frac{z_i}{z_3}\quad (i=1,2),\quad x(w)=(x_1(w),x_2(w))\in\mathbf{C}^2. \tag{21}$$

Then instead of (15) and (16) we obtain

$$x_i(-w)=\Lambda_{ii}^a x_i(w)\quad (i=1,2),\quad z_3(-w)=z_3(w), \tag{22}$$

$$x_i(w+1)=\frac{\phi_i(xz_3,z_3)}{\phi_3(xz_3,z_3)}\overset{\text{def}}{=}f_i(x)\quad (i=1,2),\quad z_3(w+1)=\Phi_3(xz_3,z_3), \tag{23}$$

where ϕ_i and Φ_3 are defined by (8)–(14) and are (below and in the sequel, we set $x_1\equiv x$, $x_2\equiv y$)

$$\phi_1(x,1)=x+3x^2-\frac{4l^2+4l-11}{4}xy-\frac{(2l+3)(2l-1)}{4}y^2, \tag{24}$$

$$\phi_2(x,1)=-y+2x^2+2xy-\frac{(2l-1)(2l+3)}{4}y^2, \tag{25}$$

$$\phi_3(x,1)=1+4x+\frac{4l^2+4l+5}{4}y-\frac{4l^2+4l-9}{2}x^2-\frac{4l^2+4l-7}{2}xy-\frac{(2l+3)(2l-1)}{4}y^2. \tag{26}$$

The following theorem on the structure of the general solution of the FE (22) and (23) is valid.

Theorem 2: The general solution of FE (22) and (23) has the form

$$x(w)=\frac{2(F(w+\frac{1}{2})+F(w-\frac{1}{2}))}{2(l-1)-(2l-3)(F(w+\frac{1}{2}-F(w-\frac{1}{2}))+4F(w+\frac{1}{2})F(w-\frac{1}{2})},$$
$$y(w)=(2l-1)^{-1}\frac{4(2(l-1)+F(w+\frac{1}{2})-F(w-\frac{1}{2}))}{2(l-1)-(2l-3)(F(w+\frac{1}{2})-F(w-\frac{1}{2}))+4F(w+\frac{1}{2})F(w-\frac{1}{2})}, \tag{27}$$

where the meromorphic function $F(w)$ is a solution of the functional equations $[F''=F(w+2),F'=F(w+1),F=F(w)]$

$$F''=\frac{2(l-1)^2(2l+3)+(l-1)(2l^2+l-5)F-2(4l-1)F'^2+2FF'-(2l-1)FF'^2}{(l-1)(2l^2+l-5)-2F'-4(l-1)F-(2l-1)F'^2}, \tag{28}$$

$$F(-w)=-F(w). \tag{29}$$

Proof: Introduce the quadratic Cremona mapping casting the functions $x(w),y(w)$ to $u_1(w),u_2(w)$:

$$u_1=\frac{y-4/(2l-1)^2}{y-[2/(2l-1)]x},\quad u_2=\frac{y-4/(2l-1)^2}{y+[2/(2l-1)]x}. \tag{30}$$

The inverse mapping is single valued and has the form

$$x=\frac{2}{2l-1}\,\frac{u_1-u_2}{u_1+u_2-2u_1u_2},\quad y=\frac{4}{(2l-1)^2}\,\frac{u_1+u_2}{u_1+u_2-2u_1u_2}. \tag{31}$$

Using (30), (31), and (24)–(26) one can obtain a simpler expression for the system (23):

$$u_1(w+1)=\frac{(4l(2l+1)u_1+4u_2-(2l-1)^2u_1u_2)(4(l-1)/(2l-1)-u_1)}{4l(2l-3)u_1+8(l-1)u_2-(2l-1)^2u_1^2},$$

$$u_2(w+1)=\frac{4(l-1)}{2l-1}-u_1, \tag{32}$$

where on the right-hand side we have set $u_1=u_1(w)$, $u_2=u_2(w)$. According to (22) an (30) we have

$$u_1(w)=u_2(-w). \tag{33}$$

Substituting (33) into the second equation of (32) and assuming

$$u_2(w)=\frac{2(l-1)}{2l-1}-\frac{2}{2l-1}\,F(w-1/2), \tag{34}$$

we obtain (29). Equations (31), (33), and (34) imply (27). Substituting (33) and (34) into the first equation of (32) we obtain Eq. (28) for $F(w)$.

This equation has a particular solution $F(w)=w$ corresponding to particular rational solution for $S(w)$ in Ref. 16.

III. INTEGRATION OF THE EQUATION FOR THE FUNCTION $F(w)$

Let $F(w-1)=u(w)$, $F(w)=v(w)$, and consider the mapping T corresponding to Eq. (28) for $F(w)[\widetilde{U}=(\widetilde{u},\widetilde{v})=(u(w+1),v(w+1)),U=(u,v)]$:

$$T:\widetilde{U}=T(U),\quad \widetilde{u}=v,$$

$$\widetilde{v}=\frac{2(l-1)^2(2l+3)+(l-1)(2l^2+l-5)u-2(4l-1)v^2+2uv-(2l-1)uv^2}{(l-1)(2l^2+l-5)-2v-4(l-1)u-(2l-1)v^2}. \tag{35}$$

The mapping T is reversible since it can be represented in the form $T=R\circ W$, where R and W are involutions:

$$R:U'=R(U),\quad u'=-u,$$

$$v'=\frac{2(l-1)^2(2l+3)-(l-1)(2l^2+l-5)v-2(4l-1)u^2+2uv+(2l-1)vu^2}{(l-1)(2l^2+l-5)+2v+4(l-1)v-(2l-1)u^2}; \tag{36}$$

$$W:U'=W(U),\quad u'=-v,\quad v'=-u.$$

Besides the parabolic fixed point $d_1:u=v=\infty$, the mapping (35) has two symmetrical hyperbolic fixed points

$$d_2:u=v=(l-1)\sqrt{\frac{2l+3}{2l-l}},\quad d_3:u=v=-(l-1)\sqrt{\frac{2l+3}{2l-1}}, \tag{37}$$

$$R(d_2)=d_3,\quad R(d_3)=d_2,\quad W(d_2)=d_3,\quad W(d_3)=d_2. \tag{38}$$

The eigenvalues λ_1,λ_2 of the linear part of the mapping (35) at the fixed point d_2 of (37) and λ_3,λ_4 at d_3 are the roots of fundamental polynomial $P_4(\lambda)$ of algebraic number λ:

$$\lambda_{1,2}=\frac{2\sqrt{(2l+3)(2l-1)}\pm\sqrt{3(2l+3)(2l-1)+1}}{\sqrt{(2l+3)(2l-1)}+1} \quad (+ \text{ for } \lambda_1); \quad \lambda_3=\lambda_1^{-1}, \quad \lambda_4=\lambda_2^{-1}. \tag{39}$$

Notice that the numbers λ_i are called conjugate algebraic numbers (see Ref. 29). The eigenvalues λ_1,λ_2 of the mapping (35) at d_2 for $l\in\mathbf{N}$ belong to Siegel's domain $(0<\lambda_2<1<\lambda_1)$. The same holds for the eigenvalues λ_3,λ_4.

Theorem 3: The set of eigenvalues λ_1,λ_2 at the fixed point d_2 and the set λ_3,λ_4 at the fixed point d_3 (39) are not resonant, i.e.,

$$\lambda_j-\lambda_1^{m_1}\lambda_2^{m_2}\neq 0 \quad \text{for } j=1,2 \quad \text{and } |m|=m_1+m_2\geqslant 2,\ m_i\in\mathbf{Z}_+,$$

$$\lambda_j-\lambda_3^{m_1}\lambda_4^{m_2}\neq 0 \quad \text{for } j=3,4 \quad \text{and } |m|=m_1+m_2\geqslant 2,\ m_i\in\mathbf{Z}_+.$$

Proof: We have to prove that algebraic numbers $\chi_1=1-\lambda_1^{m_1-1}\lambda_2^{m_2}$ and $\chi_2=1-\lambda_1^{m_1}\lambda_2^{m_2-1}$ are not equal to zero for any integers m_1 and m_2, $|m|\geqslant 2$. Consider the number χ_1. As $\lambda_1,\lambda_2\in\mathbf{Q}(\sqrt{N_1},\sqrt{N_2})$, $N_1=(2l+3)(2l-1)$ and $N_2=3(2l+3)(2l-1)+1$, a quadratic extension of the rational number field (Ref. 29, see §17), then χ_1 belongs to $\mathbf{Q}(\sqrt{N_1},\sqrt{N_2})$ and has relative degree $\leqslant 4$. The equality $\chi_1=0$ implies $\chi_1\in\mathbf{Q}(\sqrt{N_1})$. Since $\lambda_{1,2}=R_1\pm R_2\sqrt{N_2}$, where $R_1,R_2\in\mathbf{Q}(\sqrt{N_1})$, then, obviously, $\chi_1\in\mathbf{Q}(\sqrt{N_1})$ if and only if $m_2=m_1-1$. Let $m_2=m_1-1$. Then $\chi_1=1-((N_1+1-2\sqrt{N_1})/(N_1-1))^{m_2}$. It is clear that for $|m|=2m_2+1\geqslant 2$ $(m_2\geqslant 1)$ one has $\chi_1\notin\mathbf{Q}$ and, consequently, $\chi_1\neq 0$. Analogously, one can prove that $\chi_2\neq 0$. It is obvious that also the set λ_3,λ_4 is nonresonant at the fixed point d_3. ◁

The problem of linearization of the mapping (35) will be solved on the basis of the following Siegel theorem.

Definition 5: A set $(\lambda_1,\ldots,\lambda_n)\in\mathbf{C}^n$ is of the multiplicative type (C,ν), if the following inequality,

$$|\lambda_j-\lambda^m|\geqslant C|m|^{-\nu}, \quad (|m|=m_1+\cdots+m_n, \quad \lambda^m=\lambda_1^{m_1}\cdots\lambda_n^{m_n}), \tag{40}$$

is satisfied for all $j\in(1,2,\ldots,n), m_i\in\mathbf{Z}_+, |m|\geqslant 2$ $(C>0,\nu>0)$.

Theorem 4: [C. L. Siegel for $n=1$, E. Zehnder's generalization for $n>1$ (Ref. 30).] If the set $\{\lambda\}$ of eigenvalues of the linear part of a mapping, holomorphic near a fixed point, is of the multiplicative type (C,ν) for some $C>0$, $\nu>0$, then the mapping is biholomorphically equivalent to its linear part in some neighborhood of the fixed point (see Ref. 28, §28).

The following theorem on sets $\{\lambda\}$ of the multiplicative type (C,ν) is valid.[18]

Theorem 5: (See Ref. 18.) A set $(\lambda_1,\ldots,\lambda_n)\in\mathbf{C}^n$ is of the multiplicative type (C,ν), if this set is multiplicatively nonresonant and numbers $\lambda_j\in\mathbf{A}$, where $\mathbf{A}$ is an algebraic number field.

This theorem is a corollary of the Feldman theorem (Ref. 31 see chapter 10, §4.10) about the evaluation of linear forms of the logarithms of algebraic numbers.

Now we can make a general proposition about the solutions of Eqs. (28) and (29) for the function $F(w)$, defining the general solution (27) of the initial FE (23).

Theorem 6: The function $F_\delta(w)$, as a function of a one-parametric family of solutions of the system (28) and (29), is in certain finite neighborhoods of the origin in C^2, a holomorphic, function of variables z_1, z_2 or z_3, z_4:

$$z_i(w)=\delta(w)\exp((w+\beta(w)-\tfrac{1}{2})\ln\lambda_i), \quad 1\leqslant i\leqslant 4, \tag{41}$$

where λ_i are determined by (39) while arbitrary functions $\delta(w)$ and $\beta(w)$ [cf. (7)] have the following properties:

$$\begin{aligned}&\delta(-w)=\delta(w), \quad \delta(w+1)=\delta(w),\\ &\beta(-w)=-\beta(w), \quad \beta(w+1)=\beta(w).\end{aligned} \tag{42}$$

The family $F_\delta(w)$ is defined by the Taylor expansion in z_1, z_2,

$$F(w)=\sum_k f_k z^k, \quad k\in \mathbf{Z}_+^2, \quad f_k=f_{k_1,k_2}, \quad z^k=z_1^{k_1}z_2^{k_2}, \tag{43}$$

with the coefficients f_k determined by the recurrent relation

$$\begin{aligned} f_k=\Bigg\{ \sum_{|m|=1}^{|k|-1} f_{k-m}\Big[&(-((2l-1)f_0+1)\lambda^{2k-m}+((2l-1)f_0-1)\lambda^m-2(l-1)\lambda^{2k-2m} \\ &+(4l-1)\lambda^k)f_m-(2l-1)/2(\lambda^{2k-m}-\lambda^m)\sum_{|l|=1}^{|m|-1} f_{m-1}f_l\Big]\Bigg\} \\ &\times[((2l-1)f_0+l-1)(\lambda^k-\lambda_1)(\lambda^k-\lambda_2)]^{-1}, \end{aligned} \tag{44}$$

where $\lambda^k=\lambda_1^{k_1}\lambda_2^{k_2}$, $f_0=(l-1)\sqrt{(2l+3)/(2l-1)}$, $f_{0,1}$ and $f_{1,0}$ are arbitrary, and the sum is taken over all permissible values. The coefficients $\tilde{f}_k$ of the Taylor expansion of F in z_3, z_4 satisfy the same recurrent relation with λ_1, λ_2 replaced by λ_3, λ_4 and they are

$$\tilde{f}_0=-(l-1)\sqrt{\frac{2l+3}{2l-1}}, \quad \tilde{f}_{0,1}=-\lambda_3 f_{0,1}, \quad \tilde{f}_{1,0}=-\lambda_4 f_{1,0}, \quad \tilde{f}_k=-f_k/\lambda^k.$$

Each local solution of the family $F_\delta(w)$ (the germ), defined by (41)–(44), can be extended up to the global one by using $F_\delta(w)$, $F_\delta(w-1)$, and iterations of Eq. (28), and lies on the invariant manifold Γ_δ

$$\Gamma_\delta: z_1^{\gamma_1}(w+1)z_2^{\gamma_2}(w+1)=z_1^{\gamma_1}(w)z_2^{\gamma_2}(w)=\delta^2(w), \tag{45}$$

where

$$\gamma_1=\frac{-2\ln\lambda_2}{\ln\lambda_1-\ln\lambda_2}, \quad \gamma_2=\frac{2\ln\lambda_1}{\ln\lambda_1-\ln\lambda_2}.$$

Proof: According to Theorems 3 and 5, the set (λ_1,λ_2) from (39) of the eigenvalues of the mapping T of (35) at the hyperbolic fixed point d_2 given by (37) is of the multiplicative type (C,ν). Therefore, Theorem 4 guarantecs the existence of a biholomorphic mapping $U=G(z)$, $U=(u,v)$, $z=(z_1,z_2)$, and $G(0)=d_2$, which transforms a neighborhood of the origin in $\mathbf{C}^2$ to a neighborhood of the point d_2 and reduces the mapping T to the normal form at d_2:

$$G^{-1}TG:(z_1,z_2)\mapsto(\lambda_1 z_1,\lambda_2 z_2).$$

On the other hand, the biholomorphic mapping WG, where W is involution (36), transforms a neighborhood of the origin in $\mathbf{C}^2$ to a neighborhood of the point d_3 of (37) and reduces the mapping T to the normal form at d_3:

$$G^{-1}WTWG:(z_3,z_4)\mapsto(\lambda_3 z_3,\lambda_4 z_4).$$

Indeed, since involution W reverses mapping T, the mappings $G^{-1}TG$ and $G^{-1}WTWG$ are inverse to each other.

Set $(z_3,z_4)=\zeta$ and

$$G(z)=(g_1(z),g_2(z)), \quad WG(\zeta)=(-g_2(\zeta),-g_1(\zeta)).$$

If some functions $z_i(w)$, $1\leqslant i\leqslant 4$, satisfy FE

$$z_i(w+1)=\lambda_i z_i(w), \tag{46}$$

the functions $F(w)=g_2(z(w))$ or $F(w)=-g_1(\zeta(w))$ satisfy FE (28). Morcover, one will have $g_1(z(w))=F(w-1)$ or $-g_2(\zeta(w))=F(w-1)$, respectively.

Now we should determine the relations between $z(w)$ and $\zeta(w)$ which make $F(w)$ satisfy FE (29), i.e., make $F(w)$ odd. Let $F(w)=g_2(z(w))$ be close to $1/\sqrt{2}$. Then, we have $F(-w)=-g_2(\zeta(1-w))$. The equality $F(-w)=-F(w)$ is therefore ensured by

$$\zeta(1-w)=z(w). \tag{47}$$

If

$$z_i(w)=\delta_i(w)\exp((w+\alpha_i)\ln\lambda_i), \quad 1\leqslant i\leqslant 4, \tag{48}$$

then (46) is equivalent to $\delta_i(w+1)=\delta_i(w)$ and (47) is equivalent to $\zeta(w)=\Lambda z(-w)$, where $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2)$, i.e.,

$$\delta_{i+2}(-w)=\lambda_i\delta_i(w)\ \exp\ ((\alpha_i+\alpha_{i+2})\ln\lambda_i), \quad i=1,2.$$

For $\alpha_i=-1/2(1\leqslant i\leqslant 4)$ we have

$$\delta_3(-w)=\delta_1(w), \quad \delta_4(-w)=\delta_2(w). \tag{49}$$

Equations (46) and (48) result in the existence of the invariant manifold Γ_δ [cf. (45)],

$$\Gamma_\delta:\tilde{z}_1^{\gamma_1}\tilde{z}_2^{\gamma_2}=z_1^{\gamma_1}z_2^{\gamma_2}=\delta^2(w), \quad \tilde{z}_3^{\gamma_3}\tilde{z}_4^{\gamma_4}=z_3^{\gamma_3}z_4^{\gamma_4}=\delta^2(w), \tag{50}$$

where $\tilde{z}\equiv z(w+1)$, and exponents γ_1, γ_2 are determined by (45) and are invariant relative to the change $\lambda_1,\lambda_2\to\lambda_3,\lambda_4$. Substituting (48) and (49) into (50) we have

$$\delta_1^{\gamma_1}(w)\,\delta_2^{\gamma_2}(w)=\delta^2(w), \quad \delta_1^{\gamma_1}(-w)\,\delta_2^{\gamma_2}(-w)=\delta^2(w). \tag{51}$$

From (51) we have

$$\left(\frac{\delta_1(w)}{\delta_1(-w)}\right)^{\gamma_1}\left(\frac{\delta_2(w)}{\delta_2(-w)}\right)^{\gamma_2}=1. \tag{52}$$

Since $\delta_i(w+1)=\delta_i(w)$ and supposing

$$\frac{\delta_i(w)}{\delta_i(-w)}=\exp\ (\beta_i(w)\ln\lambda_i), \quad i=1,2, \tag{53}$$

we obtain from (52)

$$\exp\ [(-\beta_1(w)+\beta_2(w))\ \ln\lambda_1\ \ln\lambda_2]=1. \tag{54}$$

From (52)–(54) it follows that

$$\beta_1(w)=\beta_2(w)=2\beta(w), \tag{55}$$

where $\beta(w)$ is odd function with period equal to 1 [cf. (42)]. From (53) and (55) we have

$$\delta_i(w)=\delta(w)\ \exp\ (\beta(w)\ \ln\lambda_i), \quad \delta(-w)=\delta(w),$$

and we obtain (41) and (42).

Finally, let

$$F(w)=g_2(z(w))=\sum_k\ f_k(z(w))^k$$

for $F(w)$ close to $1/\sqrt{2}$ and

$$F(w)=-g_1(\zeta(w))=\sum_k \tilde{f}_k(\zeta(w))^k$$

for $F(w)$ close to $-1/\sqrt{2}$. Substituting these expansions into (28), we derive the desired recurrent relations on f_k and $\tilde{f}_k$. The arbitrariness in the choice of $f_{1,0}$ and $f_{0,1}$ corresponds to the ambiguity in the choice of the normalizing mapping G. ◁

Let us consider Eq. (28) at $l=-3/2$. It has the form

$$F''=\frac{5F+14F'^2+2FF'+4FF'^2}{5-2F'+10F+4F'2} \tag{56}$$

and admits a reduction of the order of (56):

$$\Phi(F'',F')=\Phi(F',F)+1, \quad \Phi(F',F)=\frac{2F'F+(5F'-7F-5)/2}{F'+F}. \tag{57}$$

It follows from (57) that $\Phi(F',F)=w+\beta(w)$ and we have the equation of lower order for the function $F(w)$ [we omit below $\beta(w)$]:

$$F(w+1)=\frac{(w+7/2)F(w)+5/2}{2F(w)-w+5/2}. \tag{58}$$

The general solution of (58) has the form

$$F(w)=w\,\frac{f(w)-1}{f(w)+1}, \quad f(-w)=f(w), \tag{59}$$

where the function $f(w)$ has the form

$$f(w)=\frac{(w^2-1/4)(w^2-9/4)}{w}\left(p\left(\frac{w+1/2}{2}\right)-p\left(\frac{w-1/2}{2}\right)+\gamma\left(\frac{w+1/2}{2}\right)\right),$$

$$p(w)=\tfrac{1}{2}\,(\Psi(w)+\Psi(-w))+\tfrac{1}{2}\,(\Psi(w+1/2)+\Psi(-w+1/2)),$$

where

$$\gamma(-w)=\gamma(w), \quad \gamma(w+1/2)=-\gamma(w).$$

The function $\Psi(z)\overset{\text{def}}{=}d/dz\,\ln\Gamma(z)$ is a solution of functional equation

$$\Psi(z+1)-\Psi(z)=\frac{1}{z}.$$

[1] J. K. Moser, SIAM (Soc. Ind. Appl. Math.) Rev. **8**, 145 (1966).
[2] J. K. Moser, Math. Ann. **169**, 136 (1967).
[3] J. K. Moser, *Stable and Random Motions in Dynamical Systems, with Special Emphasis on Celestial Mechanics* (Princeton U.P., Princeton, NJ, 1973).
[4] V. I. Arnold, in *Nonlinear and Turbulent Processes in Physics*, edited by R. Z. Sagdeev (Harwood, Chur, 1984), p. 1161.
[5] V. I. Arnold and M. B. Sevryuk, in *Nonlinear Phenomena in Plasma Physics and Hydrodynamics*, edited by R. Z. Sagdeev (Mir, Moscow, 1986), p. 31.
[6] M. B. Sevryuk, *Reversible Systems*, Lecture Notes in Mathematics, Vol. 1211 (Springer, Berlin, 1986).
[7] K. M. Khanin and Ya. G. Sinai, in *Nonlinear Phenomena in Plasma Physics and Hydrodynamics*, edited by R. Z. Sagdeev (Mir, Moscow, 1986), p. 93.
[8] R. L. Devaney, Trans. Am. Math. Soc. **218**, 89 (1976).
[9] J. A. G. Roberts and G. R. W. Quispel, Phys. Rep. **216**, 63 (1992).
[10] V. A. Meshcheryakov and K. V. Rerikh, Ann. Phys. **59**, 408 (1970).
[11] D. V. Shirkov, V. V. Serebryakov, and V. A. Meshcheryakov, *Dispersion Theories of Strong Interactions at Low Energies* (Nauka, Moscow, 1967) (in Russian).
[12] H. Hudson, *Cremona Transformations in Plane and Space* (Cambridge U. P., Cambridge, 1927).

[13] V. A. Meshcheryakov, Comm. JINR P–2369, Dubna (1965) (in Russian).
[14] G. Wanders, Nuovo Cimento **23**, 817 (1962).
[15] V. A. Meshcheryakov, Zh. Eksp. Teor. Fiz. **51**, 648 (1966) (in Russian).
[16] V. A. Meshcheryakov, ''The method of solving of scattering dispersion equations in the static limit,'' Doctoral dissertation, Dubna, 1967 (in Russian).
[17] K. V. Rerikh, Physica D **57**, 337 (1992).
[18] K. V. Rerikh, Physica D **82**, 60 (1995).
[19] K. V. Rerikh, Prep. JINR P5–93–106, Dubna (1993) (in Russian).
[20] V. I. Zhuravlev, V. A. Meshcheryakov, and K. V. Rerikh, Yad. Fiz. **10**, 168 (1969) (in Russian).
[21] K. V. Rerikh, Teor. Mat. Fiz. **50**, 251 (1982) (in Russian).
[22] K. V. Rerikh, in *Proceedings of the XIII International Conference on Differential Geometric Methods in Theoretical Physics*, Shumen, Bulgaria, 1984, edited by H. D. Doebner and T. D. Palev (World Scientific, Singapore, 1986), p. 170.
[23] G. R. W. Quispel, J. A. G. Roberts, and C. J. Thompson, Physica D **34**, 183 (1989).
[24] A. P. Veselov, Russ. Math. Surv. **46**, 1 (1991).
[25] K. V. Rerikh, J. Geom. Phys. **24**, 265 (1998).
[26] F. R. Gantmacher, *Matrix Theory* (Nauka, Moscow, 1988, in Russian); *Matrizentheorie* (Springer, Berlin, 1986).
[27] Ph. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[28] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer, New York, 1988).
[29] E. Hecke, Vorlesungen über die Theorie algebraischen Zahlen (Leipzig, 1923).
[30] E. Zehnder, in *Geometry and Topology*, Lecture Notes in Mathematics, Vol. 597 (Springer, Berlin, 1977), p. 555.
[31] N. I. Feldman, *Hilbert's Seventh Problem* (Moscow U. P., Moscow, 1982) (in Russian).

Schlesinger–Bäcklund transformations for the *N*-component KP

Johan van de Leur[a)]
Faculty of Applied Mathematics, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands

(Received 22 April 1997; accepted for publication 21 January 1998)

The n-component Kadomtsev–Petviashvili (KP) hierarchy was introduced by Sato as the set of deformation equations, i.e., Lax equations, of $n+1$ $n\times n$-matrix pseudodifferential operators. On such a set of operators we define a Schlesinger–Bäcklund transformation, which produces from a set of solutions of the KP hierarchy a new set of solutions. Applying such transformations recursively we obtain a whole family of solutions, from which we can then derive the KP tau function.
© *1998 American Institute of Physics.* [S0022-2488(98)00305-3]

I. INTRODUCTION AND FORMULATION OF THE PROBLEM

The n-component Kadomtsev–Petviashvili (KP) hierarchy was introduced by Sato in Ref. 1 (see also Refs. 2, 3) as the set of deformation equations, also called Lax equations ($i,j=1,2,...,n,\ k=1,2,...$),

$$\frac{\partial L}{\partial t_k^{(i)}}=[(L^kC^{(i)})_+,L],$$
$$\frac{\partial C^{(j)}}{\partial t_k^{(i)}}=[(L^kC^{(i)})_+,C^{(j)}], \tag{1.1}$$

of the $n\times n$-matrix pseudodifferential operators,

$$L\equiv L(x,t,\partial)=I_n\partial+U^{(1)}(x,t)\partial^{-1}+U^{(2)}(x,t)\partial^{-2}+\cdots,$$
$$C^{(j)}\equiv C^{(j)}(x,t,\partial)=E_{jj}+C^{(j1)}(x,t)\partial^{-1}+C^{(j2)}(x,t)\partial^{-2}+\cdots, \tag{1.2}$$

which satisfy the conditions

$$LC^{(j)}=C^{(j)}L,\quad C^{(i)}C^{(j)}=\delta_{ij}C^{(i)},\quad \sum_{i=1}^{n}C^{(i)}=I_n. \tag{1.3}$$

Here $\partial=\partial/\partial x$, $t=\{t_k^{(i)}|1\leqslant i\leqslant n,\ k=1,2,...\}$ and the coefficients $U^{(k)}$ and $C^{(jk)}$ are formal power series over $\mathbb{C}$ in the indeterminates x and t. $(A)_+$ stands for the differential part of a pseudodifferential operator A. The multiplication of pseudodifferential operators is given by

$$A(x,t)\partial^k\circ B(x,t)\partial^{-l}=\sum_{m=0}^{\infty}\binom{k}{m}A(x,t)\,\frac{\partial^m B(x,t)}{\partial x^m}\,\partial^{k-m-l},\quad k\in\mathbb{Z}.$$

The n-component KP hierarchy contains many of the well-known partial differential equations, e.g., the classical Kadomtsev–Petviashvili ($n=1$), Davey–Stewartson ($n=2$), and the n-wave interaction equations ($n\geqslant 3$).

Starting from the level 1 representation theory of the infinite matrix group GL_∞, Kac and the author[4] and, independently, Bergvelt and ten Kroode[5] showed that with every element τ of the

[a)]Electronic mail: vdleur@math.utwente.nl

 © 1998 American Institute of Physics

GL_∞-group orbit of the highest weight vector there exists a whole family of solutions $L(\alpha)$, $C^{(j)}(\alpha)$ of (1.1). These solutions are parametrized by the root lattice $\mathfrak{L}$ of sl_n, i.e., $\alpha \in \mathfrak{L}$.

Inspired by these articles, we want to consider the following problem in this paper. Suppose one has one nontrivial set of solutions L, $C^{(j)}$ $(1 \leqslant j \leqslant n)$, of the KP hierarchy. Is it then possible to find the corresponding tau function? So, does one such solution contain enough information? One can solve this problem using the geometry, i.e., Sato's infinite Grassmannian. However, this is, in general, not so simple. In this paper we want to give an algebraic and recursive approach. So, in order to solve this problem, one first has to find a whole family of such pseudodifferential operators that are all separate solutions of the KP hierarchy. To achieve such a result, we define a Schlesinger–Bäcklund transformation, which produces from a solution of (1.1) a new solution of (1.1). Applying such transformations recursively we obtain a whole family of solutions of (1.1), from which we can then destilate, with some restrictions, the τ function. For this construction, it will be convenient to formulate the KP hierarchy in terms of the Sato equation.

For more information on n-component KP and on some of the ideas used in this paper, we refer the reader to Ref. 4.

II. THE WAVE FUNCTION AND THE SATO EQUATION

The n-component KP hierarchy can be seen as the integrability or compatibility conditions for a linear problem. A formal oscillating function,

$$W(x,t,z) = \sum_{k=0}^{\infty} W^{(k)}(x,t) z^{-k} e^{\Xi(t,z)} e^{xz},$$

with (2.1)

$$\Xi(t,z) = \sum_{j=1}^{n} \left(\sum_{k=1}^{\infty} t_k^{(j)} z^k E_{jj} \right),$$

and $W^{(0)}(x,t) = I_n$, is called a wave function for the set of pseudodifferential operators L and $C^{(j)}$ $(1 \leqslant j \leqslant n)$ of the form (1.2) satisfying (1.3), if it satisfies the following set of linear equations:

$$LW = Wz, \quad C^{(j)}W = WE_{jj}, \quad \frac{\partial W}{\partial t_k^{(j)}} = (L^k C^{(j)})_+ W. \tag{2.2}$$

We rewrite (2.1) as follows:

$$W(x,t,z) = P(x,t,\partial) e^{\Xi(t,\partial)} e^{xz},$$

with

$$P \equiv P(x,t,\partial) = I_n + \sum_{k=1}^{\infty} W^{(k)}(x,t) \partial^{-k}. \tag{2.3}$$

Denote by A_+ (resp., $A_- = A - A_+$) the purely differential (resp., negative) part of a pseudodifferential operator A. Then clearly one has the following.

Proposition 2.1: One can find a wave function of the form (2.1) for the set of pseudodifferential operators L, $C^{(j)}$, $1 \leqslant j \leqslant n$ of the form (1.2), which satisfy (2.3) if and only if we can find an operator P of the form (2.3), which satisfies the following set of equations:

$$LP = P\partial, \quad C^{(j)}P = PE_{jj}, \quad \frac{\partial P}{\partial t_k^{(j)}} = -(L^k C^{(j)})_- P. \tag{2.4}$$

This proposition is intimately related to the Zakharov–Shabat dressing method, viz. from (2.4) we deduce that $L = P\partial P^{-1}$, $C^{(j)} = PE_{jj}P^{-1}$, and hence the last equation of (2.4) can be rewritten as

$$\frac{\partial P}{\partial t_k^{(j)}} = -(P\partial^k E_{jj} P^{-1})_- P. \tag{2.5}$$

This equation is called the Sato equation.

Proposition 2.2: Let L and $C^{(j)}$, $1 \leqslant j \leqslant n$, be of the form (1.2), satisfying (1.3). Then the equations (2.4) have a solution P of the form (2.3) if and only if L and $C^{(j)}$, $1 \leqslant j \leqslant n$, satisfy (1.1). Such a P is unique up to multiplication on the right by a pseudodifferential operator of the form

$$D(\partial) = I_n + \sum_{j=1}^{\infty} A_j \partial^{-j}, \tag{2.6}$$

where the A_j are matrices over $\mathbb{C}$.

Proof: Given L and $C^{(j)}$'s, the equations (2.4) reduce to infinitely many ordinary differential equations of the form

$$\frac{\partial W^{(j)}}{\partial x} E_{ii} = \text{linear combination of } \frac{\partial^l W^{(k)}}{\partial x^l}, \quad \text{with } k+l<j,$$

$$\frac{\partial W^{(j)}}{\partial t_m^{(i)}} = \text{linear combination of } \frac{\partial^l W^{(k)}}{\partial x^l}, \quad \text{with } k+l<j,$$

with coefficient scalar multiples of the coefficients of the $LC^{(i)}$. These equations can be solved recursively. Uniqueness up to multiplication on the right by elements $D(\partial)$ of the form (2.6) is clear. The converse is obvious, $L = P\partial P^{-1}$, $C^{(j)} = PE_{jj}P^{-1}$, and the Lax equation (1.1) follows then immediately from the Sato equation. □

Notice that from the first Lax equations (1.1) and the first Sato equations (2.5), i.e., $k=1$, we deduce that

$$\sum_{i=1}^{n} \frac{\partial L}{\partial t_1^{(i)}} = \frac{\partial L}{\partial x}, \quad \sum_{i=1}^{n} \frac{\partial C^{(j)}}{\partial t_1^{(i)}} = \frac{\partial C^{(j)}}{\partial x}, \quad \sum_{i=1}^{n} \frac{\partial P}{\partial t_1^{(i)}} = \frac{\partial P}{\partial x}. \tag{2.7}$$

Using the Sato equation we calculate explicitly

$$\frac{\partial W_{ij}^{(1)}}{\partial t_1^{(k)}} = W_{ik}^{(1)} W_{kj}^{(1)} - \delta_{jk} W_{ij}^{(2)} + \delta_{ik}\left(\frac{\partial W_{ij}^{(1)}}{\partial x} + W_{ij}^{(2)} - \sum_{l=1}^{n} W_{il}^{(1)} W_{lj}^{(1)} \right). \tag{2.8}$$

III. THE SCHLESINGER–BÄCKLUND TRANSFORMATION

Let P be of the form (2.3) satisfying the Sato equation (2.5) and let $L = P\partial P^{-1}$, $C^{(j)} = PE_{jj}P^{-1}$, $1 \leqslant j \leqslant n$, then the equations (1.1), (1.3), and (2.4) clearly hold. Suppose that the coefficient $W_{ij} := W_{ij}^{(1)} = C_{ij}^{(j1)} \neq 0$. Inspired by Proposition 3.3 of Ref. 4 and the articles of Refs. 5 and 6, we define the following differential operator, which we call a Schlesinger–Bäcklund operator for P or for the set $\{L, C^{(1)}, \ldots, C^{(n)}\}$:

$$Q(i,j;P) = \epsilon_{ij}\left(\left(\partial - \frac{\partial \log W_{ij}}{\partial t_1^{(i)}} \right) E_{ii} - \frac{1}{W_{ij}} E_{ji} - \sum_{k \neq i} W_{ik} E_{ik} \right) + \sum_{k \neq i,j} \epsilon_{ki}\epsilon_{kj}\left(E_{kk} - \frac{W_{kj}}{W_{ij}} E_{ki} \right). \tag{3.1}$$

Here ϵ_{ij} is defined by

$$\epsilon_{ij} = \begin{cases} 1, & \text{if } i \leqslant j. \\ -1, & \text{if } i > j. \end{cases}$$

Lemma 3.1: The Schlesinger–Bäcklund operator $Q(i,j;P)$ is invertible; its inverse is

$$Q(i,j;P)^{-1}=\epsilon_{ij}\Bigg(\Bigg(-\partial-\frac{\partial\log W_{ij}}{\partial t_1^{(j)}}\Bigg)E_{jj}-\frac{1}{W_{ij}}E_{ji}-\sum_{k\neq j}W_{kj}E_{kj}\Bigg)+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}\Bigg(E_{kk}-\frac{W_{ik}}{W_{ij}}E_{jk}\Bigg). \tag{3.2}$$

Proof: Multiply $Q(i,j;P)$ on the left (or on the right) by the right-hand side of (3.2). A straightforward calculation using the equations (2.8) then shows that this product is the identity matrix. □

Define the constant pseudodifferential operator,

$$R(ij)=\epsilon_{ij}(\partial E_{ii}-\partial^{-1}E_{jj})+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}E_{kk}. \tag{3.3}$$

We now calculate the coefficients of $\widetilde{P}:=Q(i,j;P)PR(i,j)^{-1}=\Sigma_l\widetilde{W}^{(l)}\partial^{-l}$ explicitly. One obtains $1\leqslant p,\ k\leqslant n,\ k\neq i,j$:

$$\begin{aligned}
\widetilde{W}_{ip}^{(l)}&=\epsilon_{ij}\epsilon_{pi}\epsilon_{pj}\Bigg(W_{ip}^{(l+1+\delta_{pj}-\delta_{pi})}+\frac{\partial W_{ip}^{(l+\delta_{pj}-\delta_{pi})}}{\partial x}\\
&\quad-W_{ip}^{(l+\delta_{pj}-\delta_{pi})}\frac{\partial\log W_{ij}^{(1)}}{\partial t_1^{(i)}}-\sum_{r\neq i}W_{ir}^{(1)}W_{rp}^{(l+\delta_{pj}-\delta_{pi})}\Bigg),\\
\widetilde{W}_{jp}^{(l)}&=-\epsilon_{ij}\epsilon_{pi}\epsilon_{pj}\frac{W_{ip}^{(l+\delta_{pj}-\delta_{pi})}}{W_{ij}^{(1)}},\\
\widetilde{W}_{kp}^{(l)}&=\epsilon_{ki}\epsilon_{kj}\epsilon_{pi}\epsilon_{pj}\Bigg(W_{kp}^{(l+\delta_{pj}-\delta_{pi})}-\frac{W_{kj}^{(1)}W_{ip}^{(l+\delta_{pj}-\delta_{pi})}}{W_{ij}^{(1)}}\Bigg).
\end{aligned} \tag{3.4}$$

From this it is easy to deduce that $\widetilde{P}$ is again of the form (2.3). The proof of the following Lemma is less straightforward.

Lemma 3.2: If P satisfies the Sato equation then so does $\widetilde{P}$, i.e.,

$$\frac{\partial\widetilde{P}}{\partial t_k^{(i)}}=-(\widetilde{P}\partial^kE_{ii}\widetilde{P}^{-1})_-\widetilde{P}.$$

Proof: We calculate

$$\begin{aligned}
\frac{\partial\widetilde{P}}{\partial t_k^{(i)}}\widetilde{P}^{-1}&=\frac{\partial Q(i,j;P)}{\partial t_k^{(i)}}Q(i,j;P)^{-1}+Q(i,j;P)\frac{\partial P}{\partial t_k^{(i)}}P^{-1}Q(ij;P)^{-1}\\
&=\frac{\partial Q(i,j;P)}{\partial t_k^{(i)}}Q(i,j;P)^{-1}-Q(i,j;P)(P\partial^kE_{ii}P^{-1})_-Q(i,j)^{-1}.
\end{aligned}$$

Now clearly the Sato equation holds if the following equation holds:

$$\begin{aligned}
\frac{\partial Q(i,j;P)}{\partial t_k^{(i)}}Q(i,j;P)^{-1}&=-Q(i,j;P)\frac{\partial P}{\partial t_k^{(i)}}P^{-1}Q(ij;P)^{-1}\\
&\quad-(Q(i,j;P)P\partial^kE_{ii}P^{-1}Q(i,j;P)^{-1})_-.
\end{aligned} \tag{3.5}$$

Denote for simplicity of notation,

$$-\frac{\partial P}{\partial t_k^{(i)}}P^{-1}=\sum_{k=1}^{\infty}S^{(k)}\partial^{-k},$$

then

$$S^{(1)}=-\frac{\partial W^{(1)}}{\partial t_k^{(i)}},\quad S^{(2)}=-\frac{\partial W^{(2)}}{\partial t_k^{(i)}}+\frac{\partial W^{(1)}}{\partial t_k^{(i)}}\,W^{(1)}. \tag{3.6}$$

And split up

$$Q=Q(i,j;P)=Q_1\partial^1+Q_0\quad\text{and}\quad Q(i,j;P)^{-1}=T_1\partial^1+T_0,$$

with

$$\begin{aligned}
&Q_1=\epsilon_{ij}E_{ii},\quad T_1=-\epsilon_{ij}E_{jj},\\
&Q_0=\epsilon_{ij}\left(\left(-\frac{\partial\log W_{ij}}{\partial t_1^{(i)}}\right)E_{ii}-\frac{1}{W_{ij}}E_{ji}-\sum_{k\neq i}W_{ik}E_{ik}\right)+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}\left(E_{kk}-\frac{W_{kj}}{W_{ij}}E_{ki}\right),\\
&T_0=\epsilon_{ij}\left(\left(-\frac{\partial\log W_{ij}}{\partial t_1^{(j)}}\right)E_{jj}-\frac{1}{W_{ij}}E_{ji}-\sum_{k\neq j}W_{kj}E_{kj}\right)+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}\left(E_{kk}-\frac{W_{ik}}{W_{ij}}E_{jk}\right);
\end{aligned} \tag{3.7}$$

then it is straightforward to check that (3.5) is equivalent to

$$-\frac{\partial Q^{(0)}}{\partial t_k^{(i)}}(T_1\partial^1+T_0)+Q^{(0)}S^{(1)}T^{(1)}+Q^{(1)}S^{(1)}T^{(0)}+Q^{(1)}\left(S^{(1)}\partial+\frac{\partial S^{(1)}}{\partial x}+S^{(2)}\right)T^{(1)}=0. \tag{3.8}$$

Next, substitute (3.6) and (3.7) onto the left-hand side of (3.8) and calculate the coefficients of E_{pq}. It is then straightforward to verify that indeed (3.8) holds. □

From all this we have the following Schlesinger–Bäcklund transformation.

Proposition 3.3: Let P be a pseudodifferential operator of the form (2.3) that satisfies the Sato equation (2.5) and let $\{L,C^{(1)},C^{(2)},\ldots,C^{(n)}\}$ be a set of pseudodifferential operators of the form (1.2) satisfying (1.1) and (1.3). If $W_{ij}^{(1)}=C_{ij}^{(j1)}\neq 0$ for $i\neq j$, then $Q(i,j;P)PR(i,j)^{-1}$ also satisfies (2.3) and (2.5), respectively, the set $\{Q(i,j;P)LQ(i,j;P)^{-1}$, $Q(i,j;P)C^{(1)}Q(i,j;P)^{-1},\ldots,Q(i,j;P)C^{(n)}Q(i,j;P)^{-1}\}$ satisfies (1.1)–(1.3).

IV. A FAMILY OF SOLUTIONS

Let P be a pseudodifferential operator satisfying (2.3) and (2.5). If $W_{ij}^{(1)}\neq 0$ for $i\neq j$, then we can use the Schlesinger–Bäcklund operator $Q(i,j;P)$ to obtain new solutions of (2.3) and (2.5). Doing this recursively, we obtain a whole family of solutions. To describe this family better we introduce the root lattice of sl_n.

Let $\tilde{\mathfrak{L}}$ be a lattice over $\mathbb{Z}$ with basis $\delta_1,\delta_2,\ldots,\delta_n$ and let $(\,|\,)$ be a symmetric bilinear form on $\tilde{\mathfrak{L}}$, defined by $(\delta_i|\delta_j)=\delta_{ij}$. Let $\delta=\delta_1+\delta_2+\cdots\delta_n$, denote by $\mathfrak{L}$ the sublattice of $\tilde{\mathfrak{L}}$ defined by $\mathfrak{L}=\{\gamma\in\tilde{\mathfrak{L}}|(\delta|\gamma)=0\}$. This lattice $\mathfrak{L}$ is the root lattice of sl_n. Using the Schlesinger–Bäcklund transformation, we can define new solutions of (2.5) parametrized by $\mathfrak{L}$. Define $P(0)=P$; suppose we have defined $P(\alpha)=\Sigma_l W^{(l)}\partial^{-l}$ for some $\alpha\in\mathfrak{L}$. If $W_{ij}^{(1)}(\alpha)\neq 0$, we define

$$P(\alpha+\delta_i-\delta_j)=Q(i,j;P(\alpha))P(\alpha)R(i,j)^{-1}. \tag{4.1}$$

However, there is a slight problem, it is not clear that in this way these solutions are well defined e.g., does the following equation hold?:

$$\begin{aligned}
&Q(i,j;Q(k,l;P(0))P(0)R(k,l)^{-1})Q(k,l;P(0))P(0)R(k,l)^{-1}R(i,j)^{-1}\\
&\quad=Q(k,l;Q(i,j;P(0))P(0)R(i,j)^{-1})Q(i,j;P(0))P(0)R(i,j)^{-1}R(k,l)^{-1}.
\end{aligned} \tag{4.2}$$

Clearly,

$$R(k,l)^{-1}R(i,j)^{-1}=R(i,j)^{-1}R(k,l)^{-1},$$

and for $\alpha\in\mathfrak{L}$ it makes sense to define $R(\alpha)$; let $\alpha=\Sigma_{i=1}^n k_i\delta_i\in\mathfrak{L}$; then we define

$$R(\alpha)=R(1,2)^{k_1}R(2,3)^{k_1+k_2}\cdots R(n-1,n)^{k_1+k_2+\cdots k_{n-1}}. \tag{4.3}$$

Notice that this is well defined. Following Ref. 4, we define a bimultiplicative function ϵ: $\mathfrak{L}\times\mathfrak{L}\to\{\pm 1\}$ by letting

$$\epsilon(\delta_i,\delta_j)=\epsilon_{ij}, \tag{4.4}$$

then

$$R(\alpha)=\sum_{i=1}^{n}\epsilon(\delta_i,\alpha)E_{ii}\partial^{(\delta_i|\alpha)}. \tag{4.5}$$

One could try to prove (4.2) by a straightforward calculation. This is tedious but possible. One has to use the equations (2.8), i.e., the first Sato equations to show that (4.2) holds. So one can say that the equations (2.8) are the Bianchi equations for the Schlesinger–Bäcklund transformations. However, such a computation does not exclude the fact that $P(\alpha)$ might be not uniquely defined, that it might depend on the path. However, this is not the case, as the following proposition shows.

Proposition 4.1: Suppose we have two paths:

$$\begin{aligned} &0=\alpha_0\mapsto\alpha_1\mapsto\alpha_2\mapsto\cdots\mapsto\alpha_p=\alpha, \quad \text{with } \alpha_k-\alpha_{k-1}=\delta_{i_k}-\delta_{j_k},\\ &0=\beta_0\mapsto\beta_1\mapsto\beta_2\mapsto\cdots\mapsto\beta_q=\beta, \quad \text{with } \beta_k-\beta_{k-1}=\delta_{l_k}-\delta_{m_k}, \end{aligned} \tag{4.6}$$

such that

$$\begin{aligned} P(\alpha_k)&=Q(i_k,j_k;P(\alpha_{k-1}))P(\alpha_{k-1})R(i_k,j_k)^{-1},\\ P(\beta_k)&=Q(l_k,m_k;P(\beta_{k-1}))P(\beta_{k-1})R(l_k,m_k)^{-1}. \end{aligned} \tag{4.7}$$

Then

$$(P(\alpha)R(\alpha-\beta)P(\beta)^{-1})_-=0, \tag{4.8}$$

and if $\alpha=\beta$, then $P(\alpha)=P(\beta)$.

Proof: Write $Q_k=Q(i_k,j_k;P(\alpha_{k-1}))$ and $\widetilde{Q}_k=Q(l_k,m_k;P(\beta_{k-1}))$ and notice that Q_k, $\widetilde{Q}_k$, Q_k^{-1}, and $\widetilde{Q}_k^{-1}$ are all first-order differential operators. Hence,

$$\begin{aligned} P(\alpha)R(\alpha-\beta)P(\beta)^{-1}&=Q_pQ_{p-1}\cdots Q_1P(0)R(\alpha)^{-1}R(\alpha-\beta)R(\beta)P(0)^{-1}\widetilde{Q}_1^{-1}\widetilde{Q}_2^{-1}\cdots\widetilde{Q}_q^{-1}\\ &=Q_pQ_{p-1}\cdots Q_1\widetilde{Q}_1^{-1}\widetilde{Q}_2^{-1}\cdots\widetilde{Q}_q^{-1} \end{aligned}$$

is a differential operator. Now if $\alpha=\beta$, then since $R(0)=I_n$ and both $P(\alpha)$ and $P(\beta)^{-1}$ are of the form (2.3), we deduce that $P(\alpha)P(\beta)^{-1}=I_n$. □

We conclude from this proposition that there is no ambiguity in the definition of $P(\alpha)$.

Now suppose $P(\alpha)$ and $P(\beta)$ exist and $\alpha-\beta=\delta_i-\delta_j$; then

$$P(\alpha)R(\delta_i-\delta_j)P(\beta)^{-1}=\epsilon_{ij}\partial E_{ii}+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}E_{kk}+\epsilon_{ij}(W^{(1)}(\alpha)E_{ii}-E_{ii}W^{(1)}(\beta)), \tag{4.9}$$

where $P(\gamma)=I_n+\Sigma_{k>0}W^{(k)}(\gamma)\partial^{-k}$. Now, expressing $P(\alpha)$ in $P(\beta)$ one obtains the following expression:

$$\epsilon_{ij}W^{(m+1)}(\alpha)E_{ii}+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}W^{(m)}(\alpha)E_{kk}-\epsilon_{ij}W^{(m-1)}(\alpha)E_{jj}$$

$$=\epsilon_{ij}E_{ii}\left(\frac{\partial W^{(m)}(\beta)}{\partial x}+W^{(m+1)}(\beta)-W^{(1)}(\beta)W^{(m)}(\beta)\right)$$

$$+\sum_{k\neq i,j}\epsilon_{ki}\epsilon_{kj}E_{kk}W^{(m)}(\beta)+\epsilon_{ij}W^{(1)}(\alpha)E_{ii}W^{(m)}(\beta). \tag{4.10}$$

Now take $m=1$ and take the (l,j)th coefficient of this expression, which gives

$$\epsilon_{ij}\delta_{li}\left(\frac{\partial W^{(1)}_{ij}(\beta)}{\partial x}+W^{(2)}_{ij}(\beta)-\sum_{k=1}^{n}W^{(1)}_{ik}(\beta)W^{(1)}_{kj}(\beta)\right)+(1-\delta_{il}-\delta_{jl})\epsilon_{li}\epsilon_{lj}W^{(1)}_{lj}(\beta)$$

$$+\epsilon_{ij}W^{(1)}_{li}(\alpha)W^{(1)}_{ij}(\beta)=-\epsilon_{ij}\delta_{lj}. \tag{4.11}$$

Taking $l=j$, we see that $W^{(1)}_{ij}(\beta)$ is invertible, and we can determine $W^{(1)}_{ji}(\alpha)$. If $l\neq i,j$ then one easily determines $W^{(1)}_{li}(\alpha)$. If $l=i$ one can use the equations (2.8) to express $W^{(1)}_{ii}(\alpha)$ in $W^{(1)}(\beta)$. In fact, one obtains that $P(\alpha)R(\delta_i-\delta_j)P(\beta)^{-1}=Q(i,j;P(\beta))$. Using (3.4) we see that since $W^{(1)}_{ij}(\beta)\neq 0$ then also $W^{(1)}_{ji}(\alpha)\neq 0$; hence one can define the inverse Schlesinger–Bäcklund operator $Q(j,i;P(\alpha))$.

Define the support $\mathrm{supp}(P)\subset\mathfrak{L}$ of P. We say that $\alpha\in\mathfrak{L}$ is in $\mathrm{supp}(P)$ if there exists a path of the form (4.6) from 0 to α such that (4.7) holds. Above we proved the following proposition:

Proposition 4.2: Let $\alpha,\beta\in supp(P)$ be such that $\alpha-\beta=\delta_i-\delta_j$; then

$$P(\alpha)=Q(i,j;P(\beta))P(\beta)R(i,j)^{-1}\quad\text{and}\quad P(\beta)=Q(j,i;P(\alpha))P(\alpha)R(j,i)^{-1}.$$

We will now introduce a wave function for every $\alpha\in\mathrm{supp}(P)$. Define

$$W(\alpha)\equiv W(\alpha,x,t,z)=P(\alpha)R(\alpha)e^{\Xi(t,z)}e^{xz}, \tag{4.12}$$

and the corresponding operators

$$L(\alpha)\equiv L(\alpha,x,t,\partial)=P(\alpha)\partial P(\alpha)^{-1},$$

$$C^{(j)}(\alpha)\equiv C^{(j)}(\alpha,x,t,\partial)=P(\alpha)E_{jj}P(\alpha)^{-1}, \tag{4.13}$$

$$N(\beta,\alpha)\equiv N(\beta,\alpha,x,t,\partial)=P(\beta)R(\beta-\alpha)P(\alpha)^{-1}.$$

Notice that by Proposition 4.1 $N(\beta,\alpha)$ is a differential operator. Since $P(\alpha)$ satisfies the Sato equation

$$\frac{\partial P(\alpha)}{\partial t^{(j)}_k}=-(P(\alpha)\partial^k E_{jj}P(\alpha)^{-1})_-P(\alpha), \tag{4.14}$$

the set of wave functions $W(\alpha)$ for $\alpha\in\mathrm{supp}(P)$ satisfy the following equations:

$$L(\alpha)W(\alpha)=W(\alpha)z,\quad C^{(j)}(\alpha)W(\alpha)=W(\alpha)E_{jj},$$

$$N(\beta,\alpha)W(\alpha)=W(\beta),\quad \frac{\partial W(\alpha)}{\partial t^{(i)}_k}=(L(\alpha)^kC^{(i)}(\alpha))_+W(\alpha), \tag{4.15}$$

and the operators $L(\alpha),C^{(j)}(\alpha),N(\beta,\alpha)$ satisfy (1.3) and the Lax equations:

$$\frac{\partial L(\alpha)}{\partial t_k^{(i)}} = [(L(\alpha)^k C^{(i)}(\alpha))_+, L(\alpha)],$$

$$\frac{\partial C^{(j)}(\alpha)}{\partial t_k^{(i)}} = [(L(\alpha)^k C^{(i)}(\alpha))_+, C^{(j)}(\alpha)], \tag{4.16}$$

$$\frac{\partial N(\beta,\alpha)}{\partial t_k^{(i)}} = (L(\beta)^k C^{(i)}(\beta))_+ N(\beta,\alpha) - N(\beta,\alpha)(L(\alpha)^k C^{(i)}(\alpha))_+ .$$

V. THE ADJOINT WAVE FUNCTIONS AND THE BILINEAR IDENTITY

On the algebra of formal $n\times n$-matrix pseudodifferential operators exists a natural linear anti-involution $*$, defined by

$$(A(x,t)\partial^k)^* = (-\partial)^k \circ {}^t A(x,t); \tag{5.1}$$

here ${}^tA$ stands for the transposed of the matrix A. The anti-involution preserves the decomposition $A = A_+ + A_-$, i.e., $(A^*)_+ = (A_+)^*$ and $(A^*)_- = (A_-)^*$. Let $W(\alpha)$ be defined by (4.12), we introduce the corresponding adjoint wave function by

$$W^*(\alpha) \equiv W^*(\alpha,x,t,z) = P(\alpha)^{*-1} R(\alpha)^{*-1} e^{-\Xi(t,z)} e^{-xz}. \tag{5.2}$$

Since $P(\alpha)^{*-1}$ satisfies the adjoint Sato equation:

$$\frac{\partial P(\alpha)^{*-1}}{\partial t_k^{(i)}} = (P(\alpha)^{*-1}(-\partial)^k E_{ii} P(\alpha)^*)_- P(\alpha)^{*-1}, \tag{5.3}$$

we have the following equations for the adjoint wave function:

$$\begin{gathered} L(\alpha)^* W^*(\alpha) = W^*(\alpha) z, \quad C^{(j)}(\alpha)^* W^*(\alpha) = W^*(\alpha) E_{jj}, \\ N(\beta,\alpha)^* W^*(\beta) = W^*(\alpha), \quad \frac{\partial W^*(\alpha)}{\partial t_k^{(i)}} = -(L(\alpha)^{*k} C^{(i)}(\alpha)^*)_+ W^*(\alpha), \end{gathered} \tag{5.4}$$

and the operators $L(\alpha)^*$, $C^{(j)}(\alpha)^*$, and $N(\beta,\alpha)^*$ satisfy (1.3) and the Lax equations:

$$\frac{\partial L(\alpha)^*}{\partial t_k^{(i)}} = -[(L(\alpha)^{*k} C^{(i)}(\alpha)^*)_+, L(\alpha)^*],$$

$$\frac{\partial C^{(j)}(\alpha)^*}{\partial t_k^{(i)}} = -[(L(\alpha)^{*k} C^{(i)}(\alpha)^*)_+, C^{(j)}(\alpha)^*], \tag{5.5}$$

$$\frac{\partial N(\beta,\alpha)^*}{\partial t_k^{(i)}} = -(L(\alpha)^{*k} C^{(i)}(\alpha)^*)_+ N(\beta,\alpha)^* + N(\beta,\alpha)^* (L(\beta)^{*k} C^{(i)}(\beta)^*)_+ .$$

The set of wave functions and their corresponding adjoint wave functions satisfy the following.

Theorem 5.1: *Let $\alpha,\beta \in supp(P)$; then one has the following bilinear identity:*

$$\mathrm{Res}_{z=0} W(\alpha,x,t,z)\, {}^t W^*(\beta,x',t',z) = 0.$$

Proof: Let $P(\alpha,x,t,z)R(\alpha) = \Sigma_i P_i \partial^{-i}$ and $P(\beta,x',t,z)^{-1} R(\beta)^{-1} = \Sigma_j Q_j \partial^{-j}$; then (4.8) is equivalent to

$$0=\left(\sum_i P_i\partial^{-i}\sum_j(-\partial)^{-j}{}^tQ_j\right)_-$$
$$=\left(\sum_{\substack{i,j\\ l\geqslant 0}}(-)^j\binom{-j-j}{l}P_i\frac{\partial^l{}^tQ_j}{\partial x^l}\partial^{-i-j-l}\right)_-.$$

Hence each coefficient of ∂^{-m-1} for $m\geqslant 0$ of this last expression, which is

$$\sum_{\substack{0\leqslant l\\ i+j+l=m+1}}(-)^j\binom{-i-j}{l}P_i\frac{\partial^l{}^tQ_j}{\partial x^l},$$

must be equal to 0. We rewrite this as

$$0=\sum_{\substack{0\leqslant l\leqslant m\\ i+j+l=m+1}}(-)^{j+l}\binom{m}{l}P_i\frac{\partial^l{}^tQ_j}{\partial x^l}$$
$$=\mathrm{Res}_{z=0}\sum_{i,j}\sum_{l=0}^{m}(-)^{j+m-l}\binom{m}{l}P_i\frac{\partial^l{}^tQ_j}{\partial x^l}z^{m-i-j-l}$$
$$=\mathrm{Res}_{z=0}\sum_{i,j}P_iz^{-i}e^{\Xi(t,z)}e^{xz}\frac{\partial^m}{\partial x^m}{}^t(Q_j(-z)^je^{-\Xi(t,z)}e^{-xz}).$$

The Lemma then follows by using the equations (5.4) and Taylor's formula. □

VI. THE TAU FUNCTION

In this section we want to introduce tau functions. For every $\alpha\in\mathrm{supp}(P)$ we introduce a tau function τ_α. We give an algebraic proof for the existence of these tau functions. This proof was inspired by the work of Ueno and Takasaki[7] on the Toda lattice hierarchy.

Before we give the main Theorem, we first introduce some notations. We write

$$P^\vee_{ij}(\alpha,x,t,z)=(P(\alpha,x,t,-z)^{*-1})_{ij},\quad R^\vee(\alpha,z)=R(\alpha,-z)^{*-1},$$

then

$$W^*(\alpha,x,t,z)=P^\vee(\alpha,x,t,z)R^\vee(\alpha,x,z)e^{-\Xi(t,z)}e^{-xz}.$$

Theorem 6.1: *For all $\alpha\in supp(P)$ we find $\tau_{\alpha+\delta_i-\delta_j}$, uniquely determined up to a constant multiple, such that*

$$P_{ij}(\alpha,x,t,z)=\epsilon_{ji}\frac{\tau_{\alpha+\delta_i-\delta_j}\left(t_k^{(p)}+\delta_{k,1}x-\delta_{pj}\frac{z^{-k}}{k}\right)}{\tau_\alpha(t_k^{(p)}+\delta_{k,1}x)},$$
$$P^\vee_{ij}(\alpha,x,t,z)=\epsilon_{ji}\frac{\tau_{\alpha+\delta_j-\delta_i}\left(t_k^{(p)}+\delta_{k,1}x+\delta_{pj}\frac{z^{-k}}{k}\right)}{\tau_\alpha(t_k^{(p)}+\delta_{k,1}x)}.\tag{6.1}$$

The proof of this theorem is based on several identities, which all follow from the bilinear identity for the wave function (Theorem 5.1). Notice that the x dependence is given by formula (2.7) and a similar formula for the adjoint wave function [which follows from (5.3)]. So from now on in the proof of the above Theorem we will concentrate on the t dependence.

Lemma 6.2: Let α *and* $\alpha+\delta_a-\delta_c\in supp(P)$ $(0\leqslant a\neq c\leqslant n)$; *then the following identities hold* $(0\leqslant b\leqslant n)$:

$$\nu P_{ab}(\alpha,x,t_r^{(p)},\nu)P^{\vee}_{bb}\left(\alpha,x,t_r^{(p)}-\delta_{pb}\frac{\nu^{-r}}{r},\nu\right)=\delta_{ab}\nu-(1-\delta_{ab})P^{\vee(1)}_{ba}\left(\alpha,x,t_r^{(p)}-\delta_{pb}\frac{\nu^{-r}}{r}\right), \tag{6.2}$$

$$\nu P^{\vee}_{ab}(\alpha,x,t_r^{(p)},\nu)P_{aa}\left(\alpha,x,t_r^{(p)}+\delta_{pa}\frac{\nu^{-r}}{r},\nu\right)=\delta_{ab}\nu-(1-\delta_{ab})P^{(1)}_{ba}\left(\alpha,x,t_r^{(p)}+\delta_{pa}\frac{\nu^{-r}}{r}\right), \tag{6.3}$$

$$\begin{aligned}&\nu P_{ac}(\alpha,x,t_r^{(p)},\nu)P^{\vee}_{bc}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},\nu\right)\\&\quad=\delta_{ab}\rho-(1-\delta_{ab}-\delta_{bc})P^{(1)}_{ab}(\alpha,x,t_r^{(p)})\\&\qquad-\rho P_{aa}(\alpha,x,t_r^{(p)},\rho)P^{\vee}_{ba}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},\rho\right),\end{aligned} \tag{6.4}$$

$$\begin{aligned}&\epsilon_{bc}\nu^{1-\delta_{bc}}P_{bb}(\alpha+\delta_a-\delta_c,x,t_r^{(p)},\nu)P^{\vee}_{cb}\left(\alpha,x,t_r^{(p)}-\delta_{pb}\frac{\nu^{-r}}{r},\nu\right)\\&\quad=-\epsilon_{ca}\epsilon_{ab}P^{(1-\delta_{ab})}_{ba}(\alpha+\delta_a-\delta_c,x,t_r^{(p)})P^{\vee(1)}_{ca}\left(\alpha,x,t_r^{(p)}+\delta_{pb}\frac{\nu^{-r}}{r}\right),\end{aligned} \tag{6.5}$$

$$\begin{aligned}&P_{aa}\left(\alpha,x,t_r^{(p)}+\delta_{pa}\frac{\nu^{-r}}{r},\nu\right)P^{\vee}_{aa}\left(\alpha,x,t_r^{(p)}-\delta_{pa}\frac{\rho^{-r}}{r},\nu\right)\\&\quad=P_{aa}\left(\alpha,x,t_r^{(p)}+\delta_{pa}\frac{\nu^{-r}}{r},\rho\right)P^{\vee}_{aa}\left(\alpha,x,t_r^{(p)}-\delta_{pa}\frac{\rho^{-r}}{r},\rho\right).\end{aligned} \tag{6.6}$$

Proof: We will give the proof of the identity (6.4); the proof of all other identities are analogous. We first introduce the formal delta function $\delta(y-z)=z^{-1}\Sigma_{k\in\mathbb{Z}}(z/y)^k$, for which one has $\Sigma_{k=0}^{\infty}(z/y)^k=z\,\delta(y-z)-\Sigma_{k=1}^{\infty}(y/z)^k$ and $\mathrm{Res}_{z=0}\,f(y,z)\delta(y-z)=f(y,y)$. Using Theorem 5.1, we deduce

$$\begin{aligned}0&=\mathrm{Res}_{z=0}\sum_{l=1}^{n}W_{al}(\alpha,x,t_r^{(p)},z)W^{*}_{bl}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},z\right)\\&=\mathrm{Res}_{z=0}\left\{P_{ac}(\alpha,x,t_r^{(p)},z)P^{\vee}_{bc}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},z\right)\sum_{k=0}^{\infty}\left(\frac{z}{\nu}\right)^k\right.\\&\quad\left.+P_{aa}(\alpha,x,t_r^{(p)},z)P^{\vee}_{bc}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},z\right)\sum_{k=0}^{\infty}\left(\frac{z}{\rho}\right)^k\right\}\\&\quad+(1-\delta_{ab}-\delta_{ac})P^{(1)}_{ab}(\alpha,x,t_r^{(p)})\\&=\mathrm{Res}_{z=0}\left\{P_{ac}(\alpha,x,t_r^{(p)},z)P^{\vee}_{bc}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},z\right)\right.\\&\quad\left.\times\left(z\,\delta(\nu-z)-\sum_{k=1}^{\infty}\left(\frac{\nu}{z}\right)^k\right)+P_{aa}(\alpha,x,t_r^{(p)},z)P^{\vee}_{bc}\left(\alpha,x,t_r^{(p)}-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},z\right)\right.\end{aligned}$$

$$\times\left(z\,\delta(\rho-z)-\sum_{k=1}^{\infty}\left(\frac{\rho}{z}\right)^k\right)\Bigg\}+(1-\delta_{ab}-\delta_{ac})P^{(1)}_{ab}(\alpha,x,t^{(p)}_r)$$

$$=\nu P_{ac}(\alpha,x,t^{(p)}_r,\nu)P^{\vee}_{bc}\left(\alpha,x,t^{(p)}_r-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},\nu\right)$$

$$+\rho P_{aa}(\alpha,x,t^{(p)}_r,\rho)P^{\vee}_{bc}\left(\alpha,x,t^{(p)}_r-\delta_{pc}\frac{\nu^{-r}}{r}-\delta_{pa}\frac{\rho^{-r}}{r},\rho\right)-\rho\,\delta_{ab}$$

$$+(1-\delta_{ab}-\delta_{ac})P^{(1)}_{ab}(\alpha,x,t^{(p)}_r).$$

□

We are now able to proof the following proposition.

Proposition 6.3: For $\alpha\in supp(P)$ *one has the following identities* $(1\leqslant i\neq j\leqslant n)$:

$$P_{jj}(\alpha,x,t^{(p)}_r,\lambda)=P_{jj}\left(\alpha,x,t^{(p)}_r-\delta_{pj}\frac{\lambda^{-r}}{r},\lambda\right)^{-1}, \tag{6.7}$$

$$P^{(1)}_{ij}(\alpha,x,t^{(p)}_r)=-P^{(1)}_{ji}(\alpha,x,t^{(p)}_r), \tag{6.8}$$

$$\lambda P_{ij}(\alpha,x,t^{(p)}_r,\lambda)=P^{(1)}_{ij}\left(\alpha,x,t^{(p)}_r-\delta_{pj}\frac{\lambda^{-r}}{r}\right)P_{jj}(\alpha,x,t^{(p)}_r,\lambda), \tag{6.9}$$

$$\lambda P^{\vee}_{ij}(\alpha,x,t^{(p)}_r,\lambda)=P^{\vee(1)}_{ij}\left(\alpha,x,t^{(p)}_r-\delta_{pj}\frac{\lambda^{-r}}{r}\right)P^{\vee}_{jj}(\alpha,x,t^{(p)}_r,\lambda). \tag{6.10}$$

Proof: The identity (6.7) is a consequence of (6.2), where one chooses $a=b=j$. To prove (6.8), we take the constant coefficient of (6.2) for $i=a\neq b=j$. The other identities follow immediately using (6.2), (6.3), and (6.8). □

Consider now the following equations $[\alpha\in\mathrm{supp}(P),\ 1\leqslant i\neq j\leqslant n]$:

$$\log P_{jj}(\alpha,0,t,z)=(e^{-\eta(t^j,z)}-1)\log\tau_a(t),\quad P^{(1)}_{ij}(\alpha,0,t)=\epsilon_{ji}\frac{\tau_{\alpha+\delta_i-\delta_j}(t)}{\tau_\alpha(t)}, \tag{6.11}$$

where

$$\eta(t,z):=\sum_{k=1}^{\infty}\frac{\partial}{\partial t_k}\frac{z^{-k}}{k}. \tag{6.12}$$

Define $p_j(\partial^{(i)})$ for $\partial^{(i)}=(\partial/\partial t^{(i)}_1,\partial/\partial t^{(i)}_2,\ldots)$ as follows:

$$\sum_{j=0}^{\infty}p_j(\partial^{(i)})z^{-j}=e^{-\eta(t^{(i)},z)},$$

and set

$$\log P_{jj}(\alpha,0,t,z)=\sum_{k=1}^{\infty}T^{(k)}_j(\alpha)z^{-k},\quad \log\epsilon_{ji}P^{(1)}_{ij}(\alpha,0,t)=T_{ij}(\alpha). \tag{6.13}$$

Of course (6.13) is not well defined if $P^{(1)}_{ij}(\alpha,0,t,z)=0$; however, this is only the case if $\alpha+\delta_i-\delta_j\notin\mathrm{supp}(P)$. In that case one simply defines $\tau_{\alpha+\delta_i-\delta_j}(t)=0$.

The equations (6.11) are solvable if the following compatibility conditions hold $(1\leqslant i,j,k\leqslant n,l,m=1,2,\ldots)$:

$$p_l(\partial^{(i)})(T^{(m)}_j(\alpha))=p_m(\partial^{(j)})(T^{(l)}_i(\alpha)), \tag{6.14}$$

$$p_l(\partial^{(i)})(T_{kj}(\alpha))=T_i^{(l)}(\alpha+\delta_k-\delta_j)-T_i^{(l)}(\alpha).$$

The ring of differential operators over $\mathbb{C}$ with generators $\partial/\partial t_l^{(i)}$, $1\leqslant i\leqslant n$, $l=1,2,...$, also have the $p_l(\partial^{(i)})$ as generators. In fact, from the first formula of (6.11) one can deduce the more explicit formula:[4]

$$\frac{\partial \log \tau_\alpha(t)}{\partial t_k^{(j)}}=\mathrm{Res}_{z=0}\, z^k\left(\frac{\partial}{\partial z}-\sum_{l>1} z^{-l-1}\frac{\partial}{\partial t_l^{(j)}}\right)\log P_{jj}(\alpha,0,t,z),$$

from which one can determine the tau function $\tau_\alpha(t)$ up to a constant factor. Hence, if the compatibility conditions (6.14) hold, then the $\tau_{\alpha+\delta_i-\delta_j}(t)$ are uniquely defined up to a constant multiple factor. These compatibility conditions follow from the next proposition.

Proposition 6.4: Let α and $\alpha+\delta_k-\delta_j\in supp(P)$, $1\leqslant j\neq k\leqslant n$; then one has the following identities $(1\leqslant i\leqslant n)$:

$$P_{ii}(\alpha,x,t_r^{(p)},\mu)P_{jj}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{\rho},\lambda\right)=P_{jj}(\alpha,x,t_r^{(p)},\lambda)P_{ii}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\lambda^{-r}}{r},\mu\right),\tag{6.15}$$

$$\begin{aligned}&\epsilon_{ji}P_{ii}(\alpha+\delta_k-\delta_j,x,t_r^{(p)},\mu)P_{ij}^{(1-\delta_{ij})}(\alpha,x,t_r^{(p)})\\&\quad=\epsilon_{jk}\epsilon_{ki}P_{ii}(\alpha,x,t_r^{(p)},\mu)P_{ik}^{(1-\delta_{ik})}(\alpha+\delta_k-\delta_j,x,t_r^{(p)})P_{kj}^{(1)}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{r}\right).\end{aligned}\tag{6.16}$$

Proof: The identity (6.15) for $i=j$ follows immediately from (6.6) and (6.7). If $i\neq j$ one uses (6.4) with $a=i$, $b=c=j$, $\nu=\lambda$, $\rho=\mu$ and with $a=j$, $b=c=i$, $\nu=\mu$, $\rho=\lambda$, respectively. One thus obtains

$$\begin{aligned}&\mu^2P_{ii}(\alpha,x,t_r^{(p)},\mu)P_{ii}^{\vee}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{r}-\delta_{jp}\frac{\lambda^{-r}}{r},\mu\right)\\&\qquad\times P_{ji}(\alpha,x,t_r^{(p)},\mu)P_{ji}^{\vee}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{r}-\delta_{jp}\frac{\lambda^{-r}}{r},\mu\right)\\&=\lambda^2P_{jj}(\alpha,x,t_r^{(p)},\lambda)P_{ii}^{\vee}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{r}-\delta_{jp}\frac{\lambda^{-r}}{r},\lambda\right)\\&\qquad\times P_{ij}(\alpha,x,t_r^{(p)},\lambda)P_{ji}^{\vee}\left(\alpha,x,t_r^{(p)}-\delta_{pi}\frac{\mu^{-r}}{r}-\delta_{jp}\frac{\lambda^{-r}}{r},\lambda\right).\end{aligned}$$

Again use (6.4), but now with $a=b=i$, $c=j$, $\nu=\lambda$, $\rho=\mu$ and $a=b=i$, $c=j$, $\nu=\mu$, $\rho=\lambda$, respectively. Then (6.15) follows immediately.

The identity (6.16) follows from (6.5), where one uses (6.7), (6.8), and (6.10). □

Corollary 6.5: Let α and $\alpha+\delta_i-\delta_k\in supp(P)$; then one has $(1\leqslant i\leqslant nl,\ m=1,2,...)$:

$$p_l(\partial^{(i)})(T_j^{(m)}(\alpha))=p_m(\partial^{(j)})(T_i^{(l)}(\alpha))$$

$$T_{kj}(\alpha)=(1-\delta_{ij})T_{ij}(\alpha)-(1-\delta_{ik})T_{ik}(\alpha+\delta_k-\delta_j)$$

$$p_l(\partial^{(i)})(T_{kj}(\alpha))=T_i^{(l)}(\alpha+\delta_k-\delta_j)-T_i^{(l)}(\alpha).$$

Hence, the compatibility conditions hold. Finally, combining Proposition 6.3, the formula (6.11) and the fact that the x dependence of the tau functions follow from the last formula of (2.7), this finishes the proof of Theorem 6.1.

We next extend the definition of the tau function (following Ref. 4). We define $\tau_\alpha=0$ for $\alpha\notin \mathrm{supp}(P)$. Let $\mathbb{C}[\tilde{\mathfrak{L}}]$ be the twisted group algebra of $\tilde{\mathfrak{L}}$, with basis e^γ, $\gamma\in\tilde{\mathfrak{L}}$ and algebra product

$$e^{\alpha}e^{\beta}=\epsilon(\alpha,\beta)e^{\alpha+\beta}. \tag{6.17}$$

Define the operator γ for $\gamma\in\tilde{\mathfrak{L}}$ on $\mathbb{C}[\tilde{\mathfrak{L}}]$ by

$$\gamma(e^{\beta})=(\gamma|\beta)e^{\beta}.$$

Notice that $\mathbb{C}[\mathfrak{L}]$ is a subalgebra of $\mathbb{C}[\tilde{\mathfrak{L}}]$. Using this group algebra we define a generating series of tau functions as follows:

$$\tau\equiv\tau(P,x,t)=\sum_{\gamma\in\mathfrak{L}}\tau_{\gamma}(x,t)e^{\gamma}. \tag{6.18}$$

Then using Theorem 5.1 and Theorem 6.1 we deduce the following important result.

Theorem 6.6: *The following bilinear identity holds for the tau function* $\tau(P,x,t)$:

$$\mathrm{Res}_{z=0}\sum_{j=1}^{n}\{e^{\delta_j z^{\delta_j}}e^{\eta(t^{(j)},z)}\tau(P,x,t)\}\{e^{-\delta_j z^{-\delta_j}}e^{-\eta(s^{(j)},z)}\tau(P,y,s)\}e^{\Sigma_{h=1}^{\infty}(t_k^{(j)}-s_k^{(j)})z^j}e^{(x-y)z}=0.$$

Remark 6.7: Notice that τ_0 *is given by the choice* $P(0,x,t)=P(x,t)$. *If one chooses a* $P(\alpha,x,t)=P(x,t)$ *for a specific* $\alpha\in\mathfrak{L}$, *one also gets a tau function. So we may conclude that the tau function belonging to* P *is uniquely defined up to a constant multiple factor and up to a translation over the root lattice* $\mathfrak{L}$. *Such a translation* $\alpha\mapsto\alpha+\beta$ *is called a Schlesinger transformation.*

Remark 6.8: Notice that with this construction it is only possible to obtain tau functions that have a connected support. It is unclear if there exist tau functions that do not have a connected support. If they exist, our construction would clearly only give a connected component. It was shown in Ref. 4 that in the polynomial case the support is always convex.

VII. THE DAVEY–STEWARTSON AND AKNS HIERARCHIES

In this section we consider the two-component KP hierarchy, which we call the Davey–Stewartson hierarchy, and its reduction the well-known AKNS hierarchy (see, e.g., Ref. 4 for more details).

From now on let $n=2$. Notice that $\alpha\in\mathfrak{L}$ is of the form $\alpha=k\delta_1-k\delta_2$ with $k\in\mathbb{Z}$. So from now on it will be convenient to write $W(k),P(k),Q(i,j,k),\ldots,$ instead of $W(\alpha),P(\alpha),Q(i,j;P(\alpha)),\ldots$. Define

$$\begin{gathered}\overline{W}(k)=P(k)R(k)=(I_2+W^{(1)}(k)\partial^{-1}+\cdots)(E_{11}\partial^k+E_{22}(-\partial)^{-k}),\\ \overline{W}^*(k)=P(k)^{*-1}R(k)^{*-1}.\end{gathered} \tag{7.1}$$

We will also write

$$\overline{W}_{ij}(k)=(\delta_{ij}+W_{ij}^{(1)}(k)\partial^{-1}+\cdots)(\delta_{j1}\partial^k+\delta_{j2}(-\partial)^{-k}),$$

and a similar expression for $\overline{W}^*_{ij}(k)$. Then one has

$$\overline{W}(k+1)=Q(i,j,k)\overline{W}(k),\quad \overline{W}^*(k+1)=Q(i,j,k)^{*-1}\overline{W}^*(k),$$

which gives for $a=1,2$:

$$\overline{W}_{1a}(k+1)=\frac{\partial\overline{W}_{1a}(k)}{\partial x}+\overline{W}_{1a}(k)\left(\partial-\frac{\partial\log W_{12}^{(1)}(k)}{\partial t_1^{(1)}}\right)-W_{12}^{(1)}(k)\overline{W}_{2a}(k), \tag{7.2}$$

$$\overline{W}_{2a}(k+1)=-\frac{\overline{W}_{1a}(k)}{W_{12}^{(1)}(k)}, \tag{7.3}$$

$$\overline{W}^*_{1a}(k+1)=-\frac{\overline{W}^*_{2a}(k)}{W^{(1)}_{12}(k)}, \tag{7.4}$$

$$\overline{W}^*_{2a}(k+1)=\frac{\partial\overline{W}^*_{2a}(k)}{\partial x}+\overline{W}^*_{2a}(k)\left(\partial-\frac{\partial \log W^{(1)}_{12}(k)}{\partial t^{(2)}_1}\right)-W^{(1)}_{12}(k)\overline{W}^*_{1a}(k). \tag{7.5}$$

Now substitute (7.3) and (7.4) in (7.2), respectively (7.5); one thus obtains the following recursion relations:

$$\begin{aligned}
\overline{W}_{1a}(k+1)&=\frac{\partial\overline{W}_{1a}(k)}{\partial x}+\overline{W}_{1a}(k)\left(\partial-\frac{\partial \log W^{(1)}_{12}(k)}{\partial t^{(1)}_1}\right)+\frac{W^{(1)}_{12}(k)}{W^{(1)}_{12}(k-1)}\,\overline{W}_{1a}(k-1),\\
\overline{W}^*_{2a}(k+1)&=\frac{\partial\overline{W}^*_{2a}(k)}{\partial x}+\overline{W}^*_{2a}(k)\left(\partial-\frac{\partial \log W^{(1)}_{12}(k)}{\partial t^{(2)}_1}\right)+\frac{W^{(1)}_{12}(k)}{W^{(1)}_{12}(k-1)}\,\overline{W}^*_{2a}(k-1).
\end{aligned} \tag{7.6}$$

The recursion formulas (7.6), when corresponding to very special tau functions, form recursion relations for string-orthogonal polynomials, as considered by Adler and van Moerbeke in Ref. 8.

Now consider the [1,1]-reduction of the Davey–Stewartson hierarchy (for more details see Ref. 4); this is the famous AKNS hierarchy. This means that $L(k)=\partial$ and $\partial W(k)/\partial x=zW(k)$, so

$$\frac{\partial\overline{W}_{1a}(k)}{\partial x}=\frac{\partial\overline{W}_{2a}(k)}{\partial x}=\frac{\partial\overline{W}^*_{1a}(k)}{\partial x}=\frac{\partial\overline{W}^*_{2a}(k)}{\partial x}=0.$$

Using (2.7) one has $\partial/\partial t^{(1)}_1=-\partial/\partial t^{(2)}_1$; hence the equations (6.7) turn into ($t_1=t^{(1)}_1$):

$$\begin{aligned}
\overline{W}_{1a}(k+1)&=\overline{W}_{1a}(k)\left(\partial-\frac{\partial \log W^{(1)}_{12}(k)}{\partial t_1}\right)+\frac{W^{(1)}_{12}(k)}{W^{(1)}_{12}(k-1)}\,\overline{W}_{1a}(k-1),\\
\overline{W}^*_{2a}(k+1)&=\overline{W}^*_{2a}(k)\left(\partial+\frac{\partial \log W^{(1)}_{12}(k)}{\partial t_1}\right)+\frac{W^{(1)}_{12}(k)}{W^{(1)}_{12}(k-1)}\,\overline{W}^*_{2a}(k-1).
\end{aligned} \tag{7.7}$$

These types of recursion relations also appear in the description of orthogonal polynomials related to the Toda lattice hierarchy, as described in Ref. 9 (see also Ref. 10).

We finally investigate the effect of the Schlesinger–Bäcklund transformation on the Lax formulation of the AKNS hierarchy. Consider again the Davey–Stewartson hierarchy and make the following change of variables:

$$x_i=t^{(1)}_i+t^{(2)}_i,\quad y_i=t^{(1)}_i-t^{(2)}_i.$$

Again consider the reduction, so $L(k)=\partial$ and $C^{(1)}(k)$ and $C^{(2)}(k)$ are independent of the x_i's and of x. Denote by $S(k)=C^{(1)}(k)-C^{(2)}(k)$; then since $S(k)^2=I_2$ and $\partial S(k)/\partial x=0$, the AKNS hierarchy is the following set of Lax equations:

$$\frac{\partial S(k)}{\partial y_i}=[(S(k)\partial^i)_+,S(k)]. \tag{7.8}$$

Let $y=y_1$, $t=-2iy_2$, $q_k=W^{(1)}_{12}(k)$, and $r_k=W^{(1)}_{21}(k)$; then it is straightforward to check that

$$\begin{aligned}
S(k)=&\begin{pmatrix}1&0\\0&-1\end{pmatrix}+\begin{pmatrix}0&-2q_k\\2r_k&0\end{pmatrix}\partial^{-1}+\begin{pmatrix}2q_kr_k&-2\partial q_k/\partial y\\-2\partial r_k/\partial y&-2q_kr_k\end{pmatrix}\partial^{-2}\\
&+\begin{pmatrix}*&-2\partial^2q_k/\partial y^2-4q_k^2r_k\\2\partial^2r_k/\partial y^2+4q_kr_k^2&*\end{pmatrix}\partial^{-3}+\cdots,
\end{aligned}$$

and that the q_k and r_k satisfy the nonlinear Schrödinger system,

$$i\,\frac{\partial q_k}{\partial t}=-\frac{1}{2}\,\frac{\partial^2 q_k}{\partial y^2}-q_k^2 r_k,\quad i\,\frac{\partial r_k}{\partial t}=\frac{1}{2}\,\frac{\partial^2 r_k}{\partial y^2}+q_k r_k^2.$$

Now

$$Q(1,2,k)=\begin{pmatrix}\partial-\partial\,\log\,q_k/\partial y & -q_k\\ -1/q_k & 0\end{pmatrix},$$

so $S(k+1)=Q(1,2,k)S(k)Q(1,2,k)^{-1}$, and this gives the following recursion relation for the q_k's and r_k's:

$$q_{k+1}=-\frac{\partial^2\,\log\,q_k}{\partial y^2}-q_k^2 r_k,\quad r_{k+1}=-\frac{1}{q_k}.$$

These are the one-dimensional Toda lattice equations. Similar equations were also obtained in Refs. 6 and 11.

ACKNOWLEDGMENTS

The author is very grateful to Gerard Helminck for stimulating discussions on the Toda lattice hierarchy and to the referee for correcting some errors. The author is supported by the ''Stichting Fundamental Onderzoek der Materie (F.O.M.).''

[1] M. Sato, ''Soliton equations as dynamical systems on infinite dimensional Grassmann manifolds,'' Res. Inst. Math. Sci. Kokyuroku **439**, 30–40 (1981).
[2] M. Jimbo, M. Kashiwara, and T. Miwa, ''Operator approach to the Kadomtsev–Petviashvili equation. Transformation groups for soliton equations. III,'' J. Phys. Soc. Jpn. **50**, 3806–3812 (1981).
[3] M. Jimbo and T. Miwa, ''Solitons and infinite dimensional Lie algebras,'' Publ. Res. Inst. Math. Sci. **19**, 943–1001 (1983).
[4] V. G. Kac and J. W. van de Leur, ''The n-component KP hierarchy and representation theory,'' in *Important Developments in Soliton Theory*, Springer Series in Nonlinear Dynamics, edited by A. S. Fokas and V. E. Zakharov (Springer, Berlin, 1993), pp. 302–343.
[5] M. Bergvelt and A. ten Kroode, ''Partitions, vertex operator constructions and multi-component KP equations,'' Pac. J. Math. **171**, 23–88 (1995).
[6] G. F. Helminck and G. F. Post, ''The geometry of differential difference equations,'' Indagationes Mathematicae (New Series) **5**, 411–438 (1994).
[7] K. Ueno and K. Takasaki, *Toda Lattice Hierarchy*, Advanced Studies in Pure Mathematics (North-Holland, New York, 1984), Vol. 4, pp. 1–95.
[8] M. Adler and P. van Moerbeke, ''String-orthogonal polynomials, string equations, and two-Toda symmetries,'' Commun. Pure Appl. Math. **L**, 241–290 (1997).
[9] M. Adler and P. van Moerbeke, ''Matrix integrals, toda symmetries, Virasoro constraints and orthogonal polynomials,'' Duke Math. J. **80**, 23–88 (1995).
[10] L. Maine and E. Horozov, ''Toda orbits of Laguerre polynomials and representations of the Virasoro algebra,'' Bull. Sci. Math **117**, 485–518 (1993).
[11] A. P. E. ten Kroode and M. J. Bergvelt, ''The homogeneous realization of the basic representation of $A_1^{(1)}$ and the Toda lattice,'' Lett. Math. Phys. **12**, 139–147 (1986).

Lax representation, *r*-matrix method, and separation of variables for the Neumann-type restricted flow

Ruguang Zhou[a)]
Department of Mathematics, Xuzhou Normal University, Xuzhou, 221009, People's Republic of China

(Received 25 November 1996; accepted for publication 9 June 1997)

The *r*-matrix method and the method of separation of variables for restricted flows with Neumann constraints are considered. Restricted Jaulent–Miodek flows are taken as illustrative examples. Lax representations are obtained from the adjoint representation of the corresponding soliton hierarchy; the *r*-matrix formula, based on this Lax representation, is derived; and the separation of variables for the Hamilton–Jacobi equation are performed. © *1998 American Institute of Physics.* [S0022-2488(98)01405-4]

I. INTRODUCTION

A great number of finite-dimensional integrable systems have arisen from the soliton equations by using the nonlinearization technique[1,2] or the restricted flow technique.[3,4] Following Ref. 3 we call these finite-dimensional integrable systems the restricted flows. Among them are the Neumann-type restricted flows which are the constrained integrable Hamiltonian system with the Neumann constraints.[1] They include the well-known Neumann model[1] and the geodesic flow on the ellipsoid.[5]

For the Neumann systems (or constrained dynamical systems) many techniques have been developed. For example, for the Neumann model, after Uhlenkbeck first found its conserved quantities, Moser,[6] Adler, van Moerbeke,[7] and Mumford[8] also obtained these conserved quantities from the geometrical interpretation. This construction enabled them to obtain the angle-action variables linearizing the Hamiltonian flow. In addition, Avan and Talon also considered the Poisson structure and the *r*-matrix formulas for the Lax operator[9,10] and again obtained the conserved quantities. These methods have been used in many other Neumann systems. But the most widely used method to study the Neumann-type restricted flows[1,11,12] is the method used by Moser in Ref. 13.

In this paper our aim is to develop the *r*-matrix method and separation of variables for Neumann-type restricted flows. As we know, the *r*-matrix formula is a powerful tool for studying the finite-dimensional integrable system.[14] From the *r*-matrix formula the involutivity of motion integrals generated from the Lax operator can be easily deduced. And the general approach to constructing separable variables for the finite-dimensional integrable systems, starting from the *r*-matrix formula, has been proposed by Sklaynin.[15–17] Crucial to our method is the use of the Dirac bracket. It is the Dirac bracket that transfers the trouble of the Hamiltonian for the constrained systems to the Poisson bracket (the standard Poisson bracket changes to the Dirac bracket) and thus enables us to more easily seek the involutive motion integrals and perform the separation of variables.

In the next section we first describe the relations between the Dirac bracket and the Neumann systems and give a general frame to treat the constrained systems using the *r*-matrix method in the Dirac bracket.

In Sec. III, we take the restricted Jaulent–Miodek flows with the Neumann constraints as examples to illustrate our method. First, according to the general procedure of the nonlinearization technique of the Lax pair,[1] we obtain two Neumann-type restricted Jaulent–Miodek flows. And then we show that the method to deduce the Lax representation for the restricted Bargmann systems from the corresponding soliton hierarchy proposed in Ref. 18 is still valid for the Neu-

[a)]Electronic mail: rgzhou@public.xz.js.cn

 © 1998 American Institute of Physics

mann system. Therefore the integrals of motion arise naturally. Moreover, we show that the Lax operator admits the representation of an r-matrix formula, both in the standard Poisson bracket and the Dirac bracket. The r-matrix formula under the Dirac bracket ensures the involutivity of the motion integrals in the constrained submanifold. Subsequently, the complete integrabilities are obtained. Finally, we discuss the separations of the Hamilton–Jacobi equation for the Neumann-type restricted Jaulent–Miodek flows using the method suggested in Ref. 19 and Ref. 20. The advantage of this method is that we can consider the separation of variables of the integrable equation hierarchy simultaneously. Finally, we reduce the Neumann-type restricted Jaulent–Miodek flows to linearizing flows. Some concluding remarks are given in Sec. IV.

II. THE *r*-MATRIX METHOD FOR THE CONSTRAINED SYSTEM

Let $\omega=\Sigma_{j=1}^{N} dq_j\wedge dp_j$ be the standard symplectic form on R^{2N} and $\{f,g\}$ the corresponding Poisson bracket between two smooth functions f and g. If $H\in C^{\infty}(R^{2N})$ we denote by V_{H} the corresponding Hamiltonian vector field, for which $dH(\cdot)=\omega(V_{\mathrm{H}}\cdot)$.

Define the submanifold M by two smooth functions F, G, satisfying $\{F,G\}\neq 0$ and dF, dG independent,

$$M=\{(q,p)\in R^{2N}|F(q,p)=G(q,p)=0\}. \tag{1}$$

Then we know that M is a symplectic submanifold with the closed 2-form $\omega|_M$, the ω restricted to TM.

In general, the vector field V_{H} need not be tangential to M, but we can construct such a vector field by replacing H with H_M, the restriction of H to M,

$$H_M=H+\frac{1}{\{F,G\}}\,(-\{H,G\}F+\{H,F\}G), \tag{2}$$

such that V_{H_M} is tangential to the submanifold M (see Ref. 6 and Ref. 21). And thus on M for any two functions $f,g\in C^{\infty}(R^{2N})$, we have

$$\{f_M,g_M\}=\{f,g\}+\frac{1}{\{F,G\}}\,(\{f,F\}\{G,g\}-\{f,G\}\{F,g\}). \tag{3}$$

However the rhs of (3) is just the definition of the Dirac bracket,[22] i.e.,

$$\{f,g\}_{\mathrm{D}}=\{f,g\}+\frac{1}{\{F,G\}}\,(\{f,F\}\{G,g\}-\{f,G\}\{F,g\}). \tag{4}$$

The Dirac bracket is a nonstandard Poisson bracket and satisfies many properties that the standard Poisson bracket satisfies, such as the antisymmetry, the linearity, the Jacobi identity, the Leibntz rule:

$$\{fg,h\}_{\mathrm{D}}=f\{g,h\}_{\mathrm{D}}+g\{f,h\}_{\mathrm{D}}, \tag{5}$$

and the derivative rule:

$$\{f(h),g\}_{\mathrm{D}}=f'\{h,g\}_{\mathrm{D}}. \tag{6}$$

These properties make the theory of the r matrix for the standard Poisson bracket in Ref. 14 hold true under the Dirac bracket. Besides we know that under the Dirac bracket, the Hamiltonian system reads as

$$q_{j,x}=\{q_j,H_M\}\equiv\{q_j,H\}_{\mathrm{D}},\quad p_{j,x}=\{p_j,H_M\}\equiv\{p_j,H\}_{\mathrm{D}},\quad j=1,2,...,N. \tag{7}$$

From the analysis above, we know that the involutivity of the restricted Hamiltonian functions of $f|_M$ and $g|_M$ in the submanifold M is equivalent to $\{f,g\}_{\mathrm{D}}=0$.

Assume we have a completely integrable Hamiltonian system H in R^{2N} with N functionally independent and involutive integrals $I_1,I_2\ldots,I_N$, i.e.,

$$\{H,I_j\}=\{I_i,I_j\}=0, \quad i,j=1,2,...,N. \tag{8}$$

Now we are concerned as to whether the Hamiltonian system restricted H to the submanifold M is still completely integrable. Namely, we want to find $N-1$ independent involutive integrals in M. The most natural idea is to make use of the original integrals $I_1,I_2,\ldots,I_N$. That is to say, we consider whether these restricted integrals are still in involution, i.e.,

$$\{H,I_j\}_{\mathrm{D}}=\{I_i,I_j\}_{\mathrm{D}}=0, \quad i,j=1,2,...,N. \tag{9}$$

From the definition of the Dirac bracket (3), it is easy to see the following.

Proposition 1: If H and F or G are the functions of $I_1,I_2,\ldots,I_N$. *i.e.,* $H=H(I_1,I_2,\ldots,I_N)$, $F=F(I_1,I_2,\ldots,I_N)$, *or* $G=G(I_1,I_2,\ldots,I_N)$, *then (9) holds.*

It is the essential element of the Moser constraint method used in Ref. 6 to derive the integrability of the Neumann model, and so on. However, for the most constrained systems the condition of the proposition above does not hold. We have to consider (9) explicitly.

For the restricted flow with the Neumann constraints, we may find the Lax representation with the aid of the adjoint representation of the corresponding soliton hierarchy, as in Ref. 18. Thus, we can obtain the motion integrals. The problem is whether the restrictions of these integrals to submanifold are in involution in the submanifold. From the general theory of the r matrix, we know if the Lax operator $L(\lambda)$ satisfies the r-matrix relation,[14]

$$\{L_1(\lambda)\overset{\otimes}{,}L_2(\mu)\}=[r_{12}(\lambda,\mu),L_1(\lambda)]+[r_{21}(\lambda,\mu),L_2(\mu)], \tag{10}$$

where the notation is as follows:[23] $L_1(\lambda)=L(\lambda)\otimes I$, $L_2(\mu)=I\otimes L(\mu)$, I is the 2×2 unit matrix; $\{L_1(\lambda)\overset{\otimes}{,}L_2(\mu)\}_{ij,kl}=\{L_1(\lambda)_{ik},L_2(\mu)_{jl}\}$; [ ] is the commutator of the matrix.

Then

$$\{L_1^2(\lambda)\overset{\otimes}{,}L_2^2(\mu)\}=[\overline{r}_{12}(\lambda,\mu),L_1(\lambda)]+[\overline{r}_{21}(\lambda,\mu),L_2(\mu)], \tag{11}$$

where

$$\overline{r}_{ij}(\lambda,\mu)=\sum_{k=0}^{1}\sum_{l=0}^{1}L_1^{1-k}(\lambda)L_2^{1-l}(\mu)r_{ij}(\lambda,\mu)L_1^k(\lambda)L_2^l(\mu). \tag{12}$$

Thus

$$\{\mathrm{Tr}\,L^2(\lambda),\mathrm{Tr}\,L^2(\mu)\}=\tfrac{1}{4}\,\mathrm{Tr}\{L_1^2(\lambda),L_2^2(\mu)\}=0, \tag{13}$$

which ensures the involution property of the integrals of motion obtained from expanding $L^2(\lambda)$ in powers of λ.

This result for the general Poisson bracket is still true for the Dirac bracket. These imply that to derive the involutivity we only need to show the r-matrix relation is true.

III. THE NEUMANN-TYPE RESTRICTED JAULENT–MIODEK FLOWS

A. The nonlinearizations of Lax systems of Jaulent–Miodek equation

The well-known Jaulent–Miodek eigenvalue problem[24] is

$$\psi_x=U(u,\lambda)\psi\equiv\begin{pmatrix}0 & 1\\ -\lambda^2+q\lambda+r & 0\end{pmatrix}\psi, \quad \psi=\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}, \quad u=\begin{pmatrix}q\\ r\end{pmatrix}, \tag{14}$$

where λ is a spectral parameter; q and r are the potentials.

To deduce the Jaulent–Miodek hierarchy, following Tu,[25] we first solve the adjoint representation of (14),

$$V_x=[U,V]\equiv UV-VU, \tag{15}$$

where V has a Laurent series expansion,

$$V(u,\lambda)=\sum_{j=1}^{\infty}\begin{pmatrix} a_m(u) & b_m(u) \\ c_m(u) & -a_m(u)\end{pmatrix}\lambda^{-m}, \tag{16}$$

which leads to the following recursion relation:

$$KG_m=JG_m, \quad a_m=-\tfrac{1}{2}b_{mx}, \quad c_m=a_{mx}-b_{m+2}+qb_{m+1}+rb_m, \tag{17}$$

$$K=\begin{pmatrix} \partial & 0 \\ 0 & -\frac{3}{4}\partial^3+r\partial+\frac{1}{2}r_x\end{pmatrix}, \quad J=\begin{pmatrix} 0 & \partial \\ \partial & -q\partial-\frac{1}{2}q_x\end{pmatrix}, \quad G_m=\begin{pmatrix} b_{m+1} \\ b_m\end{pmatrix}, \tag{18}$$

with the initial condition

$$a_0=a_1=a_2=b_0=b_1=0, \quad b_2=-1, \quad b_3=-\tfrac{1}{2}q, \quad c_0=1, \quad c_1=-\tfrac{1}{2}q. \tag{19}$$

Following Ref. 25, we know that the corresponding auxiliary spectral problem (14) is as follows:

$$\psi_{t_n}=V^{(n)}(u,\lambda)\psi, \tag{20}$$

with

$$V^{(n)}=\sum_{m=0}^{n}\begin{pmatrix} a_m & b_m \\ c_m & -a_m\end{pmatrix}\lambda^{n-m}+\begin{pmatrix} 0 & 0 \\ \lambda b_{n+1}+b_{n+2}-qb_{n+1} & 0\end{pmatrix}. \tag{21}$$

The compatible condition of (14) and (20) is

$$U_{t_n}-V_x^{(n)}+[U,V^{(n)}]=0, \tag{22}$$

which leads to the Jaulent–Miodek equation hierarchy,

$$u_{t_n}=2JG_{n+1}. \tag{23}$$

Especially, as $n=3$, the Jaulent–Miodek equation is obtained.

As usual,[2,18] we take N distinct numbers $\lambda_1,\lambda_2,\ldots,\lambda_N$ and consider the following system of equations:

$$\begin{pmatrix} \psi_{1j} \\ \psi_{2j}\end{pmatrix}_x=U(u,\lambda_j)\begin{pmatrix} \psi_{1j} \\ \psi_{2j}\end{pmatrix}\equiv\begin{pmatrix} 0 & 1 \\ -\lambda_j^2+q\lambda_j+r & 0\end{pmatrix}\begin{pmatrix} \psi_{1j} \\ \psi_{2j}\end{pmatrix}, \quad j=1,2,\ldots,N. \tag{24}$$

Under the zero boundary condition at infinity or a periodic boundary, through a direct calculation, we can get (up to a constant factor)

$$\frac{\delta\lambda_j}{\delta u}=-\begin{pmatrix} \lambda_j\psi_{1j}^2 \\ \psi_{1j}^2\end{pmatrix}, \quad j=1,\ldots,N. \tag{25}$$

As in Ref. 1, we can consider the constraints

$$G_k=\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}, \quad k=2,3,\ldots. \tag{26}$$

In this paper we focus on the Neumann constraint,[1]

$$G_2\equiv\begin{pmatrix} b_3 \\ b_2\end{pmatrix}=\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}\equiv-\begin{pmatrix} \langle\Lambda\Psi_1,\Psi_1\rangle \\ \langle\Psi_1,\Psi_1\rangle\end{pmatrix}, \tag{27}$$

i.e.,

$$\langle \Psi_1, \Psi_1 \rangle = 1, \quad q = 2\langle \Lambda \Psi_1, \Psi_1 \rangle. \tag{28}$$

Here and hereafter, we accept the following notation: $\Psi_j = (\psi_{j1}, \ldots, \psi_{jN})^T$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$; by $\langle .,. \rangle$ denote the standard inner product in space R^N.

Differentiating twice on the equation $\langle \Psi_1, \Psi_1 \rangle = 1$ yields

$$r = -\langle \Psi_2, \Psi_2 \rangle + \langle \Lambda^2 \Psi_1, \Psi_1 \rangle - 2\langle \Lambda \Psi_1, \Psi_1 \rangle^2. \tag{29}$$

Under the constraint (28) and (29), (14) is nonlinearized into the following system:

$$\Psi_{1,x} = \Psi_2,$$

$$\Psi_{2,x} = -\Lambda^2 \Psi_1 + 2\langle \Lambda \Psi_1, \Psi_1 \rangle \Lambda \Psi_1 + (\langle \Lambda^2 \Psi_1, \Psi_1 \rangle + \langle \Psi_2, \Psi_2 \rangle - 2\langle \Lambda \Psi_1, \Psi_1 \rangle^2) \Psi_1,$$

$$\langle \Psi_1, \Psi_1 \rangle = 1, \langle \Psi_1, \Psi_2 \rangle = 0. \tag{30}$$

Making use of the relation

$$K \frac{\delta \lambda_j}{\delta u} = \lambda_j J \frac{\delta \lambda_j}{\delta u},$$

we have

$$\tilde{b}_i = -\langle \Lambda^{i-2} \Psi_1, \Psi_1 \rangle, \quad \tilde{a}_i = \langle \Lambda^{i-2} \Psi_1, \Psi_2 \rangle, \quad \tilde{c}_i = \langle \Lambda^{i-2} \Psi_2, \Psi_2 \rangle, \quad i = 2,3,\ldots. \tag{31}$$

Here and hereafter, an overtilde indicates the corresponding nonlinearized quantity. We have

$$\lambda \tilde{V} = \begin{pmatrix} \sum_{k=1}^{N} \frac{\psi_{1k}\psi_{2k}}{\lambda - \lambda_k} & -\sum_{k=1}^{N} \frac{\psi_{1k}^2}{\lambda - \lambda_k} \\ \lambda - \langle \Lambda \Psi_1, \Psi_1 \rangle + \sum_{k=1}^{N} \frac{\psi_{2k}^2}{\lambda - \lambda_k} & -\sum_{k=1}^{N} \frac{\psi_{1k}\psi_{2k}}{\lambda - \lambda_k} \end{pmatrix}. \tag{32}$$

Under the condition (28) and (29), the following equation still holds:

$$\tilde{V}_x = [\tilde{U}, \tilde{V}]. \tag{33}$$

Conversely, a direct calculation can verify that (33) is just the Lax representation of equation (30). For simplicity of notation, we set $L(\lambda) = \lambda \tilde{V}$, and $M_i = \tilde{V}^{(i)}$. Therefore we have the following theorem.

Theorem 1: *The equation (30) has the Lax representation*

$$L_x = [M_0, L], \tag{34}$$

with

$$L(\lambda) = \begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & -A(\lambda) \end{pmatrix}, \quad M_0(\lambda) = \begin{pmatrix} 0 & 1 \\ D(\lambda) & 0 \end{pmatrix}, \tag{35}$$

where

$$A(\lambda) = \sum_{k=1}^{N} \frac{\psi_{1k}\psi_{2k}}{\lambda - \lambda_k}, \quad B(\lambda) = -\sum_{k=1}^{N} \frac{\psi_{1k}^2}{\lambda - \lambda_k}, \quad C(\lambda) = \lambda - \langle \Lambda \Psi_1, \Psi_1 \rangle + \sum_{k=1}^{N} \frac{\psi_{2k}^2}{\lambda - \lambda_k} \tag{36}$$

and

$$D(\lambda) = -\lambda^2 + 2\langle \Lambda \Psi_1, \Psi_1 \rangle \lambda + \langle \Lambda^2 \Psi_1, \Psi_1 \rangle - \langle \Psi_2, \Psi_2 \rangle - 2\langle \Lambda \Psi_1, \Psi_1 \rangle^2. \tag{37}$$

In the same way, we can discuss the nonlinearization of the auxiliary spectral problem,

$$\begin{pmatrix} \psi_{1j} \\ \psi_{2j} \end{pmatrix} = V^{(n)}(\lambda_j, u) \begin{pmatrix} \psi_{1j} \\ \psi_{2j} \end{pmatrix}, \quad j = 1,...,N, \tag{38}$$

under the constraints (28), (29), and (30). And we can obtain a hierarchy of finite-dimensional systems with the Neumann constraints $\langle \Psi_1, \Psi_1 \rangle = 1$ and $\langle \Psi_1, \Psi_2 \rangle = 0$. We call the restricted t_n flow. Just like in the case of the Bargmann constraint, the identity

$$V_{t_n} = [V^{(n)}, V] \tag{39}$$

is nonlinearization of the Lax representation of the restricted t_n-flow. In particular, for $n = 3$ we obtain the following system:

$$\begin{aligned} \Psi_{1,t_3} &= \langle \Lambda\Psi_1, \Psi_2 \rangle \Lambda\Psi_1 - \langle \Lambda\Psi_1, \Psi_1 \rangle \Psi_2, \\ \Psi_{2,t_3} &= E\Psi_1 - \langle \Lambda\Psi_1, \Psi_2 \rangle \Psi_2, \\ \langle \Psi_1, \Psi_1 \rangle &= 1, \langle \Psi_1, \Psi_2 \rangle = 0, \end{aligned} \tag{40}$$

with

$$\begin{aligned} E = \Lambda^3 &- \langle \Lambda\Psi_1, \Psi_1 \rangle \Lambda^2 + \Lambda(\langle \Psi_2, \Psi_2 \rangle - \langle \Lambda^2\Psi_1, \Psi_1 \rangle) - \langle \Lambda\Psi_2, \Psi_2 \rangle - \langle \Lambda^3\Psi_1, \Psi_1 \rangle \\ &+ 2\langle \Lambda\Psi_1, \Psi_1 \rangle \langle \Lambda^2, \Psi_1 \rangle, \end{aligned}$$

and the equation (40) has the Lax representation.

Theorem 2: *The equation (40) possess the following Lax representation:*

$$L_{t_3} = [M_1, L], \tag{41}$$

$$M_1(\lambda) = \begin{pmatrix} -\langle \Lambda\Psi_1, \Psi_2 \rangle & -\lambda - \frac{1}{2}\langle \Lambda\Psi_1, \Psi_1 \rangle \\ T(\lambda) & \langle \Lambda\Psi_1, \Psi_2 \rangle \end{pmatrix}, \tag{42}$$

where

$$\begin{aligned} T(\lambda) = \lambda^3 &- \langle \Lambda\Psi_1, \Psi_1 \rangle \lambda^2 + \lambda(\langle \Psi_2, \Psi_2 \rangle - \langle \Lambda^2\Psi_1, \Psi_1 \rangle) - \langle \Lambda\Psi_2, \Psi_2 \rangle \\ &- \langle \Lambda^3\Psi_1, \Psi_+ \rangle 2\langle \Lambda\Psi_1, \Psi_1 \rangle \langle \Lambda^2\Psi_1, \Psi_1 \rangle. \end{aligned}$$

B. *r* matrix and integrability

Now we turn to discuss the r-matrix relation for $L(\lambda)$. In the standard Poisson bracket of the symplectic space $(R^{2N}, d\Psi_2 \wedge d\Psi_1)$, it is not difficult to find that

$$\{A(\lambda), A(\mu)\} = \{B(\lambda), B(\mu)\} = 0, \tag{43}$$

$$\{C(\lambda), C(\mu)\} = 4(\lambda + \mu - 2\langle \Lambda\Psi_1, \Psi_1 \rangle)(A(\lambda) - A(\mu)), \tag{44}$$

$$\{A(\lambda), B(\mu)\} = \frac{2}{\mu - \lambda}(B(\mu) - B(\lambda)), \tag{45}$$

$$\{B(\lambda), C(\mu)\} = \frac{4}{\mu - \lambda}(A(\mu) - A(\lambda)), \tag{46}$$

$$\{A(\lambda), C(\mu)\} = \frac{2}{\mu - \lambda}(C(\lambda) - C(\mu)) + 2B(\lambda)(2\langle \Lambda\Psi_1, \Psi_1 \rangle - \lambda - \mu), \tag{47}$$

where λ,μ are two arbitrary parameters.

Thus we obtain the following.

Proposition 2: The L-operator $L(\lambda)$ satisfies

$$\{L_1(\lambda)\overset{\otimes}{,} L_2(\mu)\}=[r_{12}(\lambda,\mu),L_1(\lambda)]-[r_{21}(\lambda,\mu),L_2(\mu)], \tag{48}$$

with

$$r_{12}(\lambda,\mu)=\frac{2}{\mu-\lambda}P+2(-2\langle\Psi_1,\Psi_1\rangle+\lambda+\mu)\sigma_-\otimes\sigma_- \tag{49}$$

and

$$r_{21}(\lambda,\mu)=-\frac{2}{\mu-\lambda}P+2(-2\langle\Psi_1,\Psi_1\rangle+\lambda+\mu)\sigma_-\otimes\sigma_-, \tag{50}$$

where $P=1/2(I+\Sigma_{j=1}^{3}\sigma_j\otimes\sigma_j)$ is the permutation matrix, and σ_j is the Pauli matrix.

Then according to the theory of the r matrix,[14] we have

$$\{\mathrm{Tr}\ L^2(\lambda),\mathrm{Tr}\ L^2(\mu)\}=0, \tag{51}$$

which ensures the involution property of the integrals of motion obtained from expanding $L^2(\lambda)$ in powers of λ. Explicitly,

$$\det L(\lambda)=-\tfrac{1}{2}\mathrm{Tr}\ L^2(\lambda)=-A^2(\lambda)-B(\lambda)C(\lambda)=\langle\Psi_1,\Psi_1\rangle+\sum_{j=1}^{N}\frac{I_j}{\lambda-\lambda_j}, \tag{52}$$

where

$$I_k=\lambda_k\psi_{1k}^2-\langle\Lambda\Psi_1,\Psi_1\rangle\psi_{1k}^2++\sum_{\substack{i=1\\ j\neq k}}^{N}\frac{(\psi_{1j}\psi_{2k}-\psi_{1k}\psi_{2j})^2}{\lambda_k-\lambda_j},\quad k=1,2,...,N, \tag{53}$$

we have

$$\{I_j,I_k\}=0. \tag{54}$$

Moreover, it is easy to prove that $I_1,I_2,\ldots,I_N$ are N functionally independent in R^{2N}.

Since we are dealing with the involutivity in the submanifold, defined by $F=\langle\Psi_1,\Psi_1\rangle-1$ and $G=\langle\Psi_1,\Psi_2\rangle$, we have to use the Dirac bracket. Next we first give the Hamiltonian for Eqs. (30) and (40).

From the definition given in (4) as $F=\langle\Psi_1,\Psi_1\rangle-1$ and $G=\langle\Psi_1,\Psi_2\rangle$, we have

$$\{f,g\}_{\mathrm{D}}=\{f,g\}+\tfrac{1}{2}\{f,F\}\{G,g\}-\tfrac{1}{2}\{f,G\}\{F,g\}. \tag{55}$$

A direct calculation yields the following.

$$\{\psi_{1k},\psi_{1l}\}_{\mathrm{D}}=0,\quad \{\psi_{1k},\psi_{2l}\}_{\mathrm{D}}=\delta_{kl}-\psi_{1k}\psi_{1l},\quad \{\psi_{2k},\psi_{2l}\}_{\mathrm{D}}=-\psi_{1k}\psi_{2l}+\psi_{2k}\psi_{1l}. \tag{56}$$

Proposition 3: The constrained systems (30) and (40) can be written as the following Hamiltonian forms, respectively:

$$\Psi_{1,x}=\{\Psi_1,H_1\}_{\mathrm{D}},\quad \Psi_{2,x}=\{\Psi_2,H_1\}_{\mathrm{D}}. \tag{57}$$

and

$$\Psi_{1,t_3}=\{\Psi_1,H_2\}_{\mathrm{D}},\quad \Psi_{2,t_3}=\{\Psi_2,H_2\}_{\mathrm{D}}, \tag{58}$$

where

$$H_1=\tfrac{1}{2}(\langle\Lambda^2\Psi_1,\Psi_1\rangle-\langle\Lambda\Psi_1,\Psi_1\rangle^2+\langle\Psi_2,\Psi_2\rangle) \tag{59}$$

and

$$H_2=\langle\Lambda^3\Psi_1,\Psi_1\rangle-\langle\Lambda\Psi_1,\Psi_1\rangle\langle\Lambda^2\Psi_1,\Psi_1\rangle+\langle\Lambda\Psi_1,\Psi_1\rangle\langle\Psi_2,\Psi_2\rangle+\langle\Lambda\Psi_2,\Psi_2\rangle. \tag{60}$$

After a lengthy but direct calculation, we get

$$\{A(\lambda),A(\mu)\}_D=\{B(\lambda),B(\mu)\}_D=0,$$

$$\{C(\lambda),C(\mu)\}_D=4(C(\lambda)A(\mu)-A(\lambda)C(\mu))+4(\lambda+\mu-2\langle\Lambda\Psi_1,\Psi_1\rangle)(A(\lambda)-A(\mu)),$$

$$\{A(\lambda),B(\mu)\}_D=\frac{2}{\mu-\lambda}\,(B(\mu)-B(\lambda))-2B(\lambda)B(\mu),$$

$$\{B(\lambda),C(\mu)\}_D=\frac{4}{\mu-\lambda}\,(A(\mu)-A(\lambda))-4B(\lambda)A(\mu),$$

$$\{A(\lambda),C(\mu)\}_D=\frac{2}{\mu-\lambda}\,(C(\lambda)-C(\mu))+2B(\lambda)(2\langle\Lambda\Psi_1,\Psi_1\rangle-\lambda-\mu)+2B(\lambda)C(\mu).$$

From these relations Proposition 4 follows.

Proposition 4: The Lax operator $L(\lambda)$ admits the following r-matrix representation:

$$\{L_1(\lambda)\overset{\otimes}{,}L_2(\mu)\}_D=[a_{12}(\lambda,\mu),L_1(\lambda)]-[a_{21}(\lambda,\mu),L_2(\mu)], \tag{61}$$

with

$$a_{12}(\lambda,\mu)=-\frac{2}{\mu-\lambda}\,P+2(\lambda+\mu-2\langle\Lambda\Psi_1,\Psi_1\rangle)\sigma_-\otimes\sigma_-+Q_{12}, \tag{62}$$

$$a_{21}(\lambda,\mu)=\frac{2}{\mu-\lambda}\,P+2(\lambda-\mu-2\langle\Lambda\Psi_1,\Psi_1\rangle)\sigma_-\otimes\sigma_-+Q_{21}, \tag{63}$$

and

$$Q_{12}=\begin{pmatrix}1&0\\0&0\end{pmatrix}\otimes\begin{pmatrix}B(\mu)&0\\-2A(\mu)&-B(\mu)\end{pmatrix}+\begin{pmatrix}0&0\\1&0\end{pmatrix}\otimes\begin{pmatrix}0&B(\mu)\\-C(\mu)&0\end{pmatrix}, \tag{64}$$

$$Q_{21}=\begin{pmatrix}B(\lambda)&0\\-2A(\lambda)&-B(\lambda)\end{pmatrix}\otimes\begin{pmatrix}1&0\\0&0\end{pmatrix}+\begin{pmatrix}0&B(\lambda)\\-C(\lambda)&0\end{pmatrix}\otimes\begin{pmatrix}0&0\\1&0\end{pmatrix}. \tag{65}$$

Just like in the case of the standard Poisson bracket, from (61), we obtain

$$\{\mathrm{tr}\,L^2(\lambda),\mathrm{tr}\,L^2(\mu)\}_D=0. \tag{66}$$

That is to say, the restriction of $I_1,I_2,\ldots,I_N$ to M are still in involution. Since $\Sigma_{j=1}^N I_j=0$, we know $I_1,\ldots,I_N$ are not independent in M. It is not difficult to find that there are $N-1$ functions that are functionally independent among $I_1,I_2,\ldots,I_N$ in R^{2N}.

Moreover we notice that $H_1=1/2\Sigma_{j=1}^N\lambda_j I_j$, $H_2=1/2\Sigma_{j=1}^N\lambda_j^2 I_j$.

Hence we have

$$\{H_1,I_j\}_D=\{H_2,I_j\}_D=\{I_i,I_j\}_D=0,\quad i,j=1,2,...,N. \tag{67}$$

Proposition 5 follows.

Proposition 5: Both the Hamiltonian systems (30) and (40) are completely integrable systems in the Liouville's sense. Besides, the x flow and t_3 flow are commutative in the symplectic submanifold M.

C. Separation of variables

Let us introduce new variables in the $2N-2$-dimensional symplectic submanifold M. First let $u_1,\ldots,u_{N-1}$ be the $N-1$ zero points of $B(\lambda)$, i.e.,

$$\sum_{j=1}^{N} \frac{\psi_{1j}^2}{\lambda-\lambda_j} = \frac{\Pi_{k=1}^{N-1}(\lambda-u_k)}{\Pi_{k=1}^{N}(\lambda-\lambda_j)}. \tag{68}$$

It is easy to see that the choices of the λ_j's ensure $B(\lambda)$ exactly have $N-1$ simple roots.

Another $N-1$ new variables: $v_1,v_2,\ldots,v_{N-1}$, are defined by

$$v_k=\tfrac{1}{2}A(u_k), \quad k=1,2,...,N-1. \tag{69}$$

For these new variables, we have the following.

Proposition 6: New variables $\{u_k,v_k\}_{k=1}^{N-1}$ are canonically conjugated in the submanifold M, i.e.,

$$\{u_j,u_k\}_{\mathrm{D}}=\{v_j,v_k\}_{\mathrm{D}}=0, \quad \{u_j,v_k\}_{\mathrm{D}}=\delta_{jk}. \tag{70}$$

Proof: First, since the u_k's only are determined by $B(\lambda)$ and $B(\lambda)$ only depends on the variables ψ_{1j}'s, the relation $\{u_i,u_j\}_{\mathrm{D}}=0$ follows.

Second,

$$\{B(\lambda),v_k\}_{\mathrm{D}}=\{B(\lambda),\tfrac{1}{2}A(u_k)\}_{\mathrm{D}}=\tfrac{1}{2}\{B(\lambda),A(\mu)\}_{\mathrm{D},\mu=u_k}+\tfrac{1}{2}A'(u_k)\{B(\lambda),u_k\}_{\mathrm{D}}$$
$$=-\frac{1}{\mu-\lambda}\,(B(\mu)-B(\lambda))\big|_{\mu=u_k}+B(\lambda)B(\mu)\big|_{\mu=u_k}=\frac{B(\lambda)}{u_k-\lambda}.$$

However,

$$0=\{B(u_j),v_k\}_{\mathrm{D}}=\{B(\lambda),v_k\}_{\mathrm{D},\lambda=u_j}+B'(u_j)\{u_j,v_k\}_{\mathrm{D}}; \tag{71}$$

thus

$$\{u_j,v_k\}_{\mathrm{D}}=-\frac{1}{B'(u_j)}\,\{B(\lambda),v_k\}_{\mathrm{D},\lambda=u_j}. \tag{72}$$

Therefore we have

$$\{u_j,v_k\}_{\mathrm{D}}=-\frac{1}{B'(u_k)}\,\frac{B(\lambda)}{u_k-\lambda}\bigg|_{\lambda=u_j}=\delta_{jk}. \tag{73}$$

Finally, we have

$$4\{v_k,v_j\}_{\mathrm{D}}=\{A(u_k),A(u_j)\}_{\mathrm{D}}=\{A(\lambda),A(\mu)\}_{\mathrm{D},\lambda=u_k,\mu=u_j}+A'(u_j)\{v_k,u_j\}_{\mathrm{D}}+A'(u_k)\{u_k,v_j\}_{\mathrm{D}}$$
$$-A'(u_k)A'(u_j)\{u_k,u_j\}_{\mathrm{D}}=0. \tag{74}$$

This completes the proof of the proposition.

Write

$$\det L(\lambda)=\frac{P(\lambda)}{K(\lambda)}, \tag{75}$$

where

$$K(\lambda)=\prod_{k=1}^{N}(\lambda-\lambda_k)=\sum_{j=1}^{N}(-1)^j\alpha_j\lambda^{N-j} \tag{76}$$

and

$$P(\lambda)=\lambda^N+P_{N-1}\lambda^{N-1}+\cdots+P_0 \tag{77}$$

is a polynomial of order N in terms of λ.

Here (75) yields

$$P_{N-1}=-\alpha_1,\quad P_{N-2}=\alpha_2+2H_1,\quad P_{N-3}=-\alpha_3-2\alpha_1H_1+2H_2. \tag{78}$$

In general, P_k consists of the I_j's and the constants λ_j's; thus, P_k and P_j are in involution with each other. For our Hamiltonian function H_1 and H_2, we have

$$H_1=\tfrac{1}{2}(P_{N-2}-\alpha_2),\quad H_2=\tfrac{1}{2}\,(P_{N-3}+\alpha_3+\alpha_1P_{N-2}-\alpha_2\alpha_2). \tag{79}$$

Following Refs. 19 and 20, we consider the separation of variables of the Hamiltonian $P(\lambda)$ instead; directly consider the explicit H_1 and H_2, respectively.

From (75), we know that

$$P(u_k)=-4v_k^2K(u_k),\quad k=1,...,N-1. \tag{80}$$

Using the Lagrange interpolation, $P(\lambda)$ can be expressed as

$$P(\lambda)=Q(\lambda)\left(\lambda-\alpha_1+\sum_{k=1}^{N-1}u_k\right)-4\sum_{k=1}^{N-1}\frac{Q(\lambda)K(u_k)}{(\lambda-u_k)Q'(u_k)}\,v_k^2, \tag{81}$$

where

$$Q(\lambda)=\prod_{k=1}^{N-1}(\lambda-u_k). \tag{82}$$

By $\partial S/\partial u_k$, replace v_k and interpret the coefficient of $P(\lambda)$: $P_0,\ldots,P_1,\ldots,P_{N-2}$ as the integration constants. The Hamilton–Jacobi equation follows:

$$Q(\lambda)\left(\lambda-\alpha_1+\sum_{k=1}^{N-1}u_k\right)-4\sum_{k=1}^{N-1}\frac{Q(\lambda)K(u_k)}{(\lambda-u_k)Q'(u_k)}\left(\frac{\partial S}{\partial u_k}\right)^2=P(\lambda). \tag{83}$$

We wish to look for the action function S with the form $S(u_1,\ldots,u_{N-1})=\Sigma_{k=1}^{N-1}s_k(u_k)$.

Inserting into Eq. (83), dividing two sides by $Q(\lambda)$, and taking residue at $\lambda=u_k$, we get

$$4\left(\frac{\partial s_k}{\partial u_k}\right)^2=-\frac{P(u_k)}{K(u_k)},\quad k=0,...,N-2. \tag{84}$$

Finally, we have

$$S=\frac{1}{2}\sum_{k=1}^{N-1}\int_0^{u_k}\sqrt{-\frac{P(\lambda)}{K(\lambda)}}\,d\lambda. \tag{85}$$

Therefore the linearizing coordinates are

$$Q_j=\frac{\partial S}{\partial P_j}=\frac{1}{4}\sum_{k=1}^{N-1}\int_0^{u_k}\frac{\lambda^k}{\sqrt{-K(\lambda)P(\lambda)}}\,d\lambda,\quad j=0,...,N-2. \tag{86}$$

Thus, in the symplectic submanifold $(M,\Sigma_{j=0}^{N-2}dQ_j\wedge dP_j)$, the flows generating from the Hamiltonians H_1 and H_2 are, respectively,

$$Q_j=Q_j^0+\tfrac{1}{2}\,\delta_{j,N-2}x, \tag{87}$$

and

$$Q_j = Q_j^0 + \tfrac{1}{2}(\delta_{j,N-3} + \alpha_1 \delta_{j,N-2}) t_3, \quad j = 0,...,N-2, \tag{88}$$

where Q_0^0, $Q_1^0,\ldots,Q_{N-2}^0$ are $N-1$ constants.

These involve Abelian integrals of the first kind and essentially define the Abel map to the Jacobi variety. Using the reciprocity theorem relating Abelian integrals of the first kind, the equations (87) and (88) may be solved in terms of the Riemann theta function.

IV. CONCLUDING REMARKS

We have shown how to use the r-matrix approach to the Neumann system and how to separate the variables for the Neumann system. From the procedure we can see that since the use of the Dirac bracket the r matrix must be a dynamical r matrix. This will provide us with a large number of examples of the dynamical. In addition, we notice that the constraint between the potentials of the soliton hierarchy and the eigenfunctions may give the finite-gap solutions of the soliton hierarchy.[1,5] Since the system of the equation with respect to the eigenfunctions can be solved, we can search for the finite-gap solution for the soliton equation. This will be discussed in the future. Besides, as we know that the r-matrix formula and separation of variables have a direct link between the classical integrable problem and the quantum integrable problem. We hope our results obtained here will be helpful to the quantizations of the Neumann systems.

ACKNOWLEDGMENTS

The author is grateful to Professor Chaohao Gu and Hesheng Hu, Cewen Cao, Zixiang Zhou, and Zhijun Qiao for their valuable suggestions.

This project is supported in part by The Doctoral Programme Foundation of Institute High Education and NSFECJS.

[1] C. W. Cao and X. G. Geng, ''Classical integrable systems generated through nonlinearization of eigenvalue problems,'' *Proceedings of the Conference on Nonlinear Physics*, Shanghai, 1989, Research Reports in Physics (Springer-Verlag, Berlin, 1990), pp. 66–78.
[2] C. W. Cao, Sci. China A **33**, 528 (1990).
[3] S. Rauch-Wojciechowiski, Phys. Lett. A **160**, 241 (1991).
[4] O. Ragnisco and S. Rauch-Wojciechowski, Inverse Probl. **8**, 245 (1992).
[5] R. G. Zhou, Phys. Lett. A **220**, 320 (1996).
[6] J. Moser, ''Geometry of quadrics and spectral theory,'' in *The Chern Symposium*, 1979 (Springer-Verlag, Berlin 1980).
[7] M. Adler and P. van Moerbeke, Acta Metall. **38**, 267 (1980).
[8] D. Mumford, *Tata Lectures on Theta* (Birkhauser, Basel, 1984), Vol. II.
[9] J. Avan and M. Talon, Phys. Lett. B **268**, 209 (1991).
[10] J. Avan and M. Talon, Int. J. Mod. Phys. A **5**, 4477 (1990).
[11] C. W. Cao and X. G. Geng, J. Math. Phys. **32**, 2325 (1991).
[12] Z. J. Qiao, J. Nonlinear Math. Phys. **4**, 221 (1997).
[13] J. Moser, ''Various aspects of integrable Hamiltonian systems,'' *Proceedings of Centro Internazionale Mathematico Estivo* held in Bressone, Italy (Ligouri Editions, Naples, 1980), pp. 37–195.
[14] O. Babelon and C. M. Villet, Phys. Lett. B **237**, 411 (1990).
[15] E. Sklyanin, Lect. Notes Phys. **226**, 196 (1985).
[16] E. K. Sklyanin, Commun. Math. Phys. **150**, 18 (1992).
[17] E. K. Sklyanin, Prog. Theor. Phys. Suppl. **118**, 35 (1995).
[18] Y. B. Zeng and Y. S. Li, J. Phys. A **26**, L273 (1993).
[19] O. Babelon and M. Talon, Nucl. Phys. B **379**, 321 (1992).
[20] J. Harnad and P. Winternitz, Commun. Math. Phys. **172**, 263 (1995).
[21] P. Deift, F. Lund, and E. Trubowitz, Commun. Math. Phys. **74**, 141 (1980).
[22] K. Sundermeyer, ''Constrained dynamics with applications to Yang–Mills theory,'' *General Relativity, Classical Spin, Dual String Model*, Lectures Notes in Physics 169 (Springer-Verlag, Berlin, 1982).
[23] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Soliton* (Spring-Verlag, Berlin, 1978).
[24] M. Jaulent and K. Miodek, J. Math. Phys. **17**, 1351 (1978).
[25] G. Z. Tu, J. Phys. A **22**, 2375 (1989).

Curvature-free asymmetric metric connections and Lanczos potentials in Kerr–Schild spacetimes

F. Andersson[a] and S. Brian Edgar[b]
Department of Mathematics, Linköping University, Linköping, S581 83 Sweden

(Received 14 October 1997; accepted for publication 2 February 1998)

The set of spinor equations linking the curvature spinors of a Riemann–Cartan space with the curvature spinors of the corresponding Riemann space is derived, and these are used to establish a simple relationship between Ψ-flat connections of the Riemann–Cartan space and Lanczos potentials of the Riemann space. This not only yields, very easily, a recent result of Bergqvist for the Kerr metric, but also enables Bergqvist's result to be generalized; specifically we show that a curvature-free connection, associated with a class of Kerr–Schild metrics, can be identified as a Lanczos potential for the Weyl conformal curvature spinor of these spaces. © *1998 American Institute of Physics.* [S0022-2488(98)01205-5]

I. RIEMANN–CARTAN SPINOR EQUATIONS

When a symmetric metric connection ∇_a of a Riemannian space (with curvature tensor R_{abcd}) is generalized by the addition of a tensor $\Gamma_{abc}(=\Gamma_{[ab]c})$ to give a new asymmetric connection $\hat{\nabla}_a$ by

$$\hat{\nabla}_a\xi_b=\nabla_a\xi_b-2\Gamma_b{}^c{}_a\xi_c\,, \tag{1}$$

this new Riemann–Cartan space is still metric compatible and has curvature tensor $\hat{R}_{abcd}$ given by

$$\hat{R}_{abcd}=R_{abcd}+4\nabla_{[a}\Gamma_{|cd|b]}+8\Gamma_{ed[a}\Gamma_{|c|}{}^e{}_{b]}\,. \tag{2}$$

(The standard notation of Ref. 1 is used.) The spinor version of Γ_{abc} is given by

$$\Gamma_{abc}=\Gamma_{ABCC'}\epsilon_{A'B'}+\overline{\Gamma}_{A'B'C'C}\epsilon_{AB}\,, \tag{3}$$

where $\Gamma_{ABCC'}=\Gamma_{(AB)CC'}$, so that (1) becomes[1]

$$\hat{\nabla}_{AA'}\xi_B=\nabla_{AA'}\xi_B-2\Gamma^E{}_{BAA'}\xi_E\,. \tag{4}$$

The spinor version of $\hat{R}_{abcd}$ is given by[1]

$$\begin{aligned}\hat{R}_{abcd}=\epsilon_{A'B'}\epsilon_{C'D'}[&\hat{\Psi}_{ABCD}+2(\epsilon_{B(C}\hat{\Sigma}_{D)A}+\epsilon_{A(C}\hat{\Sigma}_{D)B})+\hat{\Lambda}(\epsilon_{AD}\epsilon_{BC}+\epsilon_{AC}\epsilon_{BD})]\\&+\hat{\Phi}_{ABC'D'}\epsilon_{A'B'}\epsilon_{CD}+\text{c.c.}\end{aligned} \tag{5}$$

After the substitution of (3) and the usual spinor decomposition for R_{abcd} into the right hand side of (2), the resulting spinor expression can be equated to the right hand side of (5). Then, this spinor version of (2) can be written out as a number of equations,[2]

$$\begin{aligned}\hat{\Psi}_{ABCD}&=\Psi_{ABCD}-2\nabla_{(A}{}^{E'}\Gamma_{BCD)E'}-4\Gamma_{E(AB}{}^{E'}\Gamma^E{}_{CD)E'},\\3\hat{\Lambda}&=3\Lambda-\nabla_A{}^{E'}\Gamma^{AB}{}_{BE'}-\Gamma_{EABF'}\Gamma^{BEAF'}+\Gamma^B{}_{EBF'}\Gamma^{EA}{}_A{}^{F'},\end{aligned}$$

[a]Electronic mail: frand@mai.liu.se
[b]Electronic mail: bredg@mai.liu.se

 © 1998 American Institute of Physics

$$4\hat{\Sigma}_{AD}=\nabla^{BE'}\Gamma_{B(AD)E'}-\nabla_{(A}{}^{E'}\Gamma^{B}{}_{D)BE'}-2\Gamma_{E(D|B|}{}^{F'}\Gamma^{BE}{}_{A)F'}-2\Gamma_{E(AD)}{}^{F'}\Gamma^{BE}{}_{BF'},$$
$$\hat{\Phi}_{ABC'D'}=\Phi_{ABC'D'}+2\nabla_{(B|E'}\overline{\Gamma}_{C'D'|}{}^{E'}{}_{A)}+4\overline{\Gamma}_{E'D'F'(A}\overline{\Gamma}_{|C'|}{}^{E'F'}{}_{B)}. \tag{6}$$

To our knowledge, the precise set of equations (6) has not appeared before in the literature, although an NP spin-coefficient version of related equations has been given.[3,4]

II. LANCZOS POTENTIALS AND FLAT CONNECTIONS

Illge[5] has shown that the Weyl curvature spinor Ψ_{ABCD} of a Riemann space can always be given locally in terms of a symmetric potential $L_{ABCA'}(=L_{(ABC)A'})$ as

$$\Psi_{ABCD}=2\nabla_{(A}{}^{A'}L_{BCD)A'}. \tag{7}$$

(It should be noted that Lanczos[6] originally proposed a potential for the Weyl tensor C_{abcd}; this 3-tensor H_{abc} had the properties $H_{abc}=H_{[ab]c}$ and $H_{[abc]}=0$, with an associated algebraic gauge $H_{ab}{}^{a}=\chi_b$. It is easy to see that the gauge choice $\chi_b=0$ reduces the 3-tensor H_{abc} to be equivalent to the symmetric spinor $L_{ABCA'}$. A formal local existence proof for this tensor version has been given by Bampi and Caviglia.[7] See Ref. 8 for a recent review on the Lanczos spinor/tensor.)

We therefore have the following result, directly from the first of equations (6).

Lemma: For an asymmetric metric connection $\hat{\nabla}_{AA'}$ defined as in (4), any two of the following properties implies the third:

(i) $\hat{\Psi}_{ABCD}=0$, *i.e., the Riemann–Cartan space is Ψ-flat.*
(ii) $\Gamma_{E(AB}{}^{A'}\Gamma^{E}{}_{CD)A'}=0$.
(iii) $L_{ABCA'}=\Gamma_{(ABC)A'}$ *is a Lanczos potential for the Weyl spinor Ψ_{ABCD} of the original Riemann space, i.e., (7) is satisfied.*

III. FLAT CONNECTIONS WHICH YIELD LANCZOS POTENTIALS

A few years ago Bergqvist and Ludvigsen[9] constructed, for the Kerr spacetime, a metric-compatible asymmetric connection with vanishing curvature, which they have proposed has significant physical relevance. Shortly afterwards, Harnett[10] showed a similar construction can be carried out for the whole Kerr–Schild class of metrics,[11]

$$g_{ab}=g_{o^{ab}}+fl_al_b, \tag{8}$$

where $g_{o^{ab}}$ is the Minkowski metric, f is a scalar function of the coordinates and l_a is a null vector with respect to both metrics. An associated metric-compatible asymmetric connection, which has vanishing curvature, is constructed from the symmetric connection via (4) by choosing

$$\Gamma_{ABCC'}=\nabla_{(A}{}^{A'}(f\,o_{B)}o_Co_{A'}o_{C'}), \tag{9}$$

where $l_a=o_Ao_{A'}$.

Recently, Bergqvist[12] has shown that the flat connection found in Ref. 9 can be used to determine a Lanczos potential for the Weyl spinor of the Kerr metric.

We will now generalize Bergqvist's result and show—for a large class of the Kerr–Schild spaces—that the flat connection found in Ref. 10 can be used to determine a Lanczos potential for the Weyl spinor of these Kerr–Schild metrics.

Theorem: *For those Kerr–Schild metrics given by (8), where l_a is tangent to a shear-free null geodesic congruence, the Weyl spinor has a Lanczos potential given by*

$$L_{ABCA'}=\nabla_{(A}{}^{C'}(f\,o_Bo_{C)}o_{A'}o_{C'}).$$

Proof: As noted above, from Ref. 10 the connection given by substituting (9) in (4) is curvature-free, and so, in particular, condition (i) holds in the Lemma above. Hence $\Gamma_{(ABC)A'}$ is a Lanczos potential if and only if condition (ii) holds.

Noting that

$$o^{A'}\Gamma_{ABCA'} = o^{A'}\nabla_{(A}{}^{A'}(f\, o_{B)}o_C o_{A'}o_{C'}) = f\, o_C o_{(B}o^{B'}o^{A'}\nabla_{A)B'}o_{A'} = 0,$$

where the fact that o^A represents a shear-free null geodesic congruence is used in the last step, we can write

$$\Gamma_{ABCA'} = M_{ABC}o_{A'}$$

for some M_{ABC}. Hence condition (ii) follows.

This proof, for the whole Kerr–Schild subclass, is much more concise than Bergqvist's proof[12] for the special case of Kerr spacetime. This result will enable the physical applications of the Lanczos spinor, which Bergqvist has been developing for Kerr space, to be applicable to this larger class of Kerr–Schild spaces.

In conclusion, we point out that there are a number of algorithms in the literature for constructing Lanczos potentials for various spacetimes;[13–17] the Lemma in Section II gives us a means of checking whether such Lanczos potentials enable us to construct a $\Psi-$flat connection, while the equations in Section I enable us to test which other parts of the curvature tensor also vanish. It is straightforward to show that the Lanczos potentials defined in Refs. 14 and 15 do not permit a $\Psi-$flat connection by the construction in this paper; on the other hand, the Lanczos potential defined in Ref. 17 is easily confirmed to permit such a Ψ-flat connection, and indeed some other parts of the connection are also flat. We shall look at these Lanczos potentials in more detail elsewhere.

[1] R. Penrose and W. Rindler, *Spinors and Spacetime, Vols. 1 and 2* (Cambridge U. P., Cambridge, 1984).
[2] F. Andersson, *Linköping Studies in Science and Technology. Theses. No. 647* (Linköping University, Sweden, 1997).
[3] R. Gambini and L. Herrera, J. Math. Phys. **21**, 1449 (1980).
[4] S. Jogia and J. B. Griffiths, Gen. Relativ. Gravit. **12**, 597 (1980).
[5] R. Illge, Gen. Relativ. Gravit. **20**, 551 (1988).
[6] C. Lanczos, Rev. Mod. Phys. **34**, 379 (1962).
[7] F. Bampi and G. Caviglia, Gen. Relativ. Gravit. **15**, 375 (1983).
[8] S. B. Edgar and A. Höglund, Proc. R. Soc. London, Ser. A **453**, 835 (1997).
[9] G. Bergqvist and M. Ludvigsen, Class. Quantum Grav. **8**, 697 (1991).
[10] G. Harnett, Class. Quantum Grav. **10**, 407 (1993).
[11] R. P. Kerr and A. Schild, in *Atti Conv. Relativa Generale* (Barbera, Firenze, 1965), p. 222.
[12] G. Bergqvist, J. Math. Phys. **38**, 3142 (1997).
[13] M. Novello and V. Velloso, Gen. Relativ. Gravit. **19**, 1251 (1987).
[14] G. Ares de Parga, J. Math. Phys. **30**, 1294 (1989).
[15] J. L. López-Bonilla *et al.*, Class. Quantum Grav. **10**, 2153 (1993).
[16] P. Dolan and C. W. Kim, Proc. R. Soc. London, Ser. A **447**, 577 (1994).
[17] G. F. Torres del Castillo, J. Math. Phys. **36**, 195 (1995).

"Clock rate synchronizable" reference frames in curved space–times

Sijie Gao
Department of Physics, Beijing Normal University, Beijing, People's Republic of China

Zhiquan Kuang[a] and Canbin Liang[b]
Center of Theoretical Physics, China Center of Advanced Science and Technology (World Lab), Beijing, People's Republic of China

(Received 12 September 1997; accepted for publication 6 January 1998)

It is well known that all reference frames that are clock synchronizable are hypersurface orthogonal. A weaker requirement for clock synchronization was first suggested by Zhao [Z. Zhao, Sci. China **34**(7), 835 (1991)] using coordinate language. Reference frames satisfying this requirement are hence referred to as class Z in the present paper. We give a geometrical definition of class Z reference frames and investigate their physical meaning as well as geometrical properties in detail. Finally, another kind of reference frame referred to as class $\tilde{\mathrm{z}}$ is introduced and its possible relation to another of Zhao's claims is pointed out. © *1998 American Institute of Physics.* [S0022-2488(98)02604-8]

I. INTRODUCTION

It is well accepted in general relativity that an observer in a space–time (M,g_{ab}) is a future pointing timelike curve (his world line) in M with $g_{ab}(\partial/\partial\tau)^a(\partial/\partial\tau)^b=-1$, where τ is the proper time of the observer, and a reference frame $\mathscr{R}$ in (M,g_{ab}) is a vector field, each of whose integral curves is an observer.[1] (In this paper we use the "abstract index notation,"[2] where a letter with Latin letter indices represents a tensor itself while the same letter with Greek letter indices represent the components of that tensor with respect to some basis.) Since each observer has his own clock, there exists the issue of synchronization among clocks of different observers in $\mathscr{R}$. In the literature there is a series of definitions of synchronizability of reference frames:[1] $\mathscr{R}$ is called (a) *proper time synchronizable* iff $\exists$ smooth function t on M such that $\nu_a=-(dt)_a$, where ν^a is the four-velocity of observers in $\mathscr{R}$; (b) *locally proper time synchronizable* iff $(d\nu)_{ab}=0$; (c) *synchronizable* iff $\exists$ smooth functions t and $h(>0)$ on M such that $\nu_a=-h(dt)_a$; (d) *locally synchronizable* iff $\nu_a\wedge(d\nu)_{bc}=0$. The last condition, being the weakest one of the four, is equivalent to requiring that $\mathscr{R}$ be hypersurface orthogonal. Physically, orthogonal hypersurfaces (short for hypersurface orthogonal to all world lines in $\mathscr{R}$) are interpreted as simultaneity surfaces, and hence all (standard or coordinate) clocks in $\mathscr{R}$ are required be adjusted (synchronized) so that they give the same reading on each such surface. To this end one needs only to choose an arbitrary observer (denoted γ_P) in $\mathscr{R}$ to take the lead and command all other observers in $\mathscr{R}$ to adjust their clock settings. In the case where $\mathscr{R}$ is not hypersurface orthogonal, one can choose any "one-dimensional family" of observers in $\mathscr{R}$ (their world lines comprise a two-dimensional submanifold of M) and require that all clocks in this family be adjusted so that they give the same reading on each "simultaneity curve" (a curve orthogonal to all world lines in this family). The problem, however, is that orthogonal lines might not be closed curves, and, as a result, after observers γ_Q and $\gamma_{Q'}$ in $\mathscr{R}$ adjust their clocks to agree with γ_P, they will find that they do not agree with each other, i.e., if η is a simultaneity curve connecting γ_Q to $\gamma_{Q'}$, the readings of γ_Q and $\gamma_{Q'}$ on $\gamma_Q\cap\eta$ and $\gamma_{Q'}\cap\eta$ will be different. An interesting point suggested by Ref. 3 is that there does exist a class of reference frames satisfying the following requirement: if γ_Q and $\gamma_{Q'}$ are two clocks in $\mathscr{R}$ and η_1 and η_2 are any simultaneity curves connecting γ_Q to $\gamma_{Q'}$ and if $q_1=\gamma_Q\cap\eta_{1'}$, $q_2=\gamma_Q\cap\eta_2$, $q_1'=\gamma_{Q'}\cap\eta_1$, and $q_2'=\gamma_{Q'}\cap\eta_2$ (Fig. 1), then $\Delta t\equiv t(q_2)-t(q_1)$ equals $\Delta t'$

[a]Mailing address: Institute of Mathematics, Academia Sinica, Beijing, People's Republic of China.
[b]Mailing address: Department of Physics, Beijing Normal University, Beijing, People's Republic of China.

 © 1998 American Institute of Physics

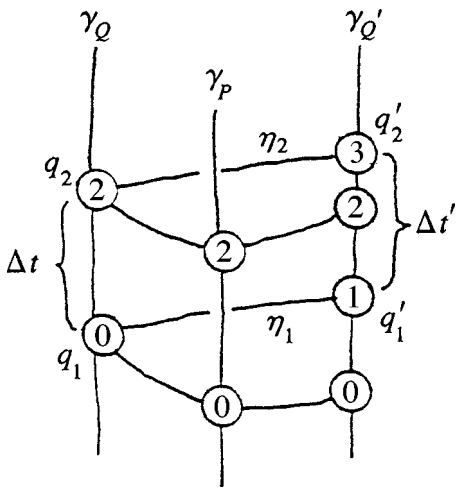

FIG. 1. Here $\Delta t' = \Delta t$ implies that all clocks in reference frame $\mathscr{R}$ have the same *rate* despite the fact that their initial settings *cannot* be adjusted to give the same readings on all simultaneity curves.

$\equiv t(q_2') - t(q_1')$, where t is the time coordinate of the comoving coordinate system. This can be interpreted physically as that all coordinate clocks in $\mathscr{R}$ have the same *rate* despite the fact that their initial settings *cannot* be adjusted to give the same readings on all simultaneity curves. This property of reference frames is referred to as ''transitivity of clock rate synchronizability'' in Ref. 3. Reference frames satisfying any one of the previous conditions (a)–(d) all admit this property since they are all hypersurface orthogonal, but the converse is not true. To avoid awkward wording, the jargon ''clock rate synchronizable reference frames'' is replaced by ''reference frames of class Z'' in this paper.

II. REFERENCE FRAMES OF CLASS Z

Throughout the paper we use (M, g_{ab}), $\mathscr{R}$, γ, and ν^a to represent respectively a space–time, a reference frame with domain $M_{\mathscr{R}} \subseteq M$, an observer in this frame, and the four velocity of the observer. Let $\hat{M}_{\mathscr{R}}$ be the set consisting of all world lines γ in $\mathscr{R}$ as elements. For the purpose of this paper it is enough to confine ourselves to reference frames in each of which $\hat{M}_{\mathscr{R}}$ is a three-dimensional manifold and there is a natural projection map $\pi: M_{\mathscr{R}} \to \hat{M}_{\mathscr{R}}$. A smooth curve $\hat{\eta} \subset \hat{M}_{\mathscr{R}}$ can be interpreted as a one-dimensional family of observers. A simultaneity curve $\eta \subset M_{\mathscr{R}}$ is called a *horizontal lift* of $\hat{\eta}$ if $\pi[\eta] = \hat{\eta}$. A reference frame $\mathscr{R}$ is said to belong to class Z if there exists a smooth function t on $M_{\mathscr{R}}$ satisfying (a) $t \circ \gamma$ is a monotone increasing function of the proper time τ for any observer γ in $\mathscr{R}$ and (b) if η and η' are simultaneity curves with $\pi[\eta] = \pi[\eta']$, γ_P and γ_Q are world lines in $\mathscr{R}$ intersecting η and η' at p, p', q, and q' (Fig. 2), then

$$\Delta t_{p'p} = \Delta t_{q'q}, \tag{1}$$

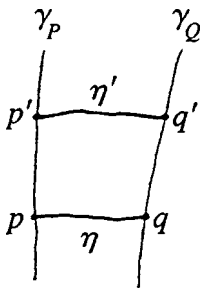

FIG. 2. Any pair of world lines γ_P and γ_Q in a class Z reference frame satisfies $\Delta t_{p'p} = \Delta t_{q'q}$ if η and η' are simultaneity curves with $\pi[\eta] = \pi[\eta']$.

where $\Delta t_{p'p} \equiv t(p') - t(p)$ and $\Delta t_{q'q} \equiv t(q') - t(q)$. Here $\mathscr{R}$ is said to belong to class Z locally iff $\forall\ p \in M_{\mathscr{R}}$ $\exists$ a neighborhood O_p of p such that $\exists$ a smooth function on O_p satisfying conditions (a) and (b). We now investigate the geometrical properties of class Z reference frames.

A smooth vector field w^a on γ is called a *neighbor* of γ in $\mathscr{R}$ iff $g_{ab} w^a \nu^b = 0$ and

$$\mathscr{L}_\nu w^a = \alpha \nu^a \tag{2}$$

with α a smooth function on γ. Note that this definition is equivalent to that given in Ref. 1.

Proposition 1: $\mathscr{R}$ belongs to class Z iff there exists a smooth function $f \neq 0$ on $M_{\mathscr{R}}$ such that

$$\mathscr{L}_\xi w^a = 0 \tag{3}$$

for any neighbor w^a, where $\xi^a \equiv f \nu^a$.

Proof: Suppose there exists f such that Eq. (3) holds. Then there exists a smooth function t on $M_{\mathscr{R}}$ such that $\xi^a = (\partial/\partial t)^a$. Let $\hat{\eta}$ be a nonclosed smooth curve in $\hat{M}_{\mathscr{R}}$. Then $\mathscr{S} \equiv \pi^{-1}[\hat{\eta}]$ is a two-dimensional submanifold in M. The commutivity of ξ^a and w^a [Eq. (3)] ensures the existence of coordinates t' and x' on $\mathscr{S}$ such that

$$(\partial/\partial t')^a = \xi^a = (\partial/\partial t)^a \quad \text{and} \quad (\partial/\partial x')^a = w^a. \tag{4}$$

The orthogonality of w^a to ν^a implies that each integral curve of w^a is a simultaneity curve, and Eq. (4) guarantees $\Delta t_{p'p} = \Delta t_{q'q}$ between any two simultaneity curves η and η' (Fig. 2), and hence $\mathscr{R}$ belongs to class Z by definition. Conversely, suppose $\mathscr{R}$ belongs to class Z. Then there exists a function t on $M_{\mathscr{R}}$ such that $\Delta t_{p'p} = \Delta t_{q'q} \equiv \Delta t$. Using t as a new parameter of each world line γ in $\mathscr{R}$, then the tangent vector field $\xi^a = (\partial/\partial t)^a$ defines, for any Δt, a diffeomorphism $\phi_{\Delta t}$ with $\phi_{\Delta t}[\eta] = \eta'$ (Fig. 2). Therefore $\phi_{\Delta t *} w^a|_p = w^a|_{p'}$ for any neighbor w^a, and hence $\mathscr{L}_\xi w^a = 0$. □

Proposition 2: $\mathscr{R}$ belongs to class Z iff there exists a smooth function f on $M_{\mathscr{R}}$ such that

$$H^a \nabla_a \ln f = H^a A_a, \quad \forall p \in M_{\mathscr{R}}, \tag{5}$$

where A^a is the four-acceleration of the observers in $\mathscr{R}$, and H^a satisfies $g_{ab} H^a \nu^b = 0$.

Proof: Suppose $\mathscr{R}$ belongs to class Z. Then Proposition 1 ensures the existence of a smooth function f such that

$$0 = \mathscr{L}_{f\nu} w^a = f(\nu^b \nabla_b w^a - w^b \nabla_b \nu^a - \nu^a w^b \nabla_b \ln f). \tag{6}$$

Contracting both sides with ν_a, making use of $\nu_a w^a = 0$ and $\nu^a \nu_a = -1$, one obtains $w^a A_a = w^b \nabla_b \ln f$. For any vector H^a orthogonal to γ there is a neighbor w^a such that $H^a = \lambda w^a$, $\lambda \in \mathbb{R}$, therefore Eq. (5) follows. Conversely, if there is a smooth function f satisfying Eq. (5), then $\mathscr{L}_{f\nu} w^a = 0$ for any neighbor w^a since it follows from Eqs. (6), (5), and (2) that

$$\nu_a \mathscr{L}_{f\nu} w^a = -w^a A_a + w^b \nabla_b \ln f = 0$$

and

$$H_a \mathscr{L}_{f\nu} w^a = f H_a \mathscr{L}_\nu w^a = f H_a \alpha \nu^a = 0 \quad \forall H^a \text{ orthogonal to } \gamma. \qquad \square$$

Proposition 3: $\mathscr{R}$ belongs to class Z locally iff $\forall p \in M_{\mathscr{R}}$ there exists a comoving coordinate system (t, x^i) $(i = 1,2,3)$ with a coordinate patch containing p such that

$$\frac{\partial}{\partial t}\left(\frac{g_{0i}}{g_{00}}\right) = 0, \tag{7}$$

where $g_{\mu\nu}$ are components of g_{ab} with respect to this coordinate system.

Note: Equation (7) is used in Ref. 3 to define ''transitivity of clock rate synchronizability.''

Proof: Suppose there exist comoving coordinates (t, x^i) satisfying Eq. (7). Let

$$e^a_{(i)}=\left(\frac{\partial}{\partial x^i}\right)^a-\frac{g_{0i}}{g_{00}}\left(\frac{\partial}{\partial t}\right)^a,\quad i=1,2,3. \tag{8}$$

Then it is easily checked that $e^a_{(i)}$ is a neighbor, and any neighbor w^a can be written as $w^a=w^{(i)}e^a_{(i)}$ with $w^{(i)}$ satisfying $\partial w^{(i)}/\partial t=0$, therefore

$$\mathscr{L}_{\partial/\partial t}w^a=\mathscr{L}_{\partial/\partial t}w^{(i)}e^a_{(i)}=-w^{(i)}\left(\frac{\partial}{\partial t}\right)^a\frac{\partial}{\partial t}\left(\frac{g_{0i}}{g_{00}}\right)=0,$$

where Eqs. (2)–(7) are used in the last step. Consequently $\mathscr{R}$ belongs to class Z locally. Conversely, if $\mathscr{R}$ belongs to class Z locally, then $\forall p\in M_{\mathscr{R}}$ there is a comoving coordinate system (t,x^i) satisfying $\mathscr{L}_{\partial/\partial t}w^a=0$ $\forall$ neighbors w^a confined in the coordinate patch, hence

$$0=\mathscr{L}_{\partial/\partial t}e^a_{(i)}=-\left(\frac{\partial}{\partial t}\right)^a\frac{\partial}{\partial t}\left(\frac{g_{0i}}{g_{00}}\right),$$

implying Eq. (7). □

Note that the time coordinate t in the comoving coordinate systems involved in Proposition 3 can be chosen to agree with the function involved in the definition of class Z.

It follows from Propositions 2 and 3 that (a) all reference frames with $(dA)_{ab}=0$ belong to class Z at least locally; (b) all hypersurface orthogonal reference frames belong to class Z at least locally; (c) all conformally stationary reference frames belong to class Z; and (d) $\mathscr{R}$ belongs to class Z iff there exists a conformal transformation that turns the world lines of observers in $\mathscr{R}$ to geodesics.

III. REFERENCE FRAMES OF CLASS $\tilde{Z}$ AND ZHAO'S CONCLUSION

A reference frame $\mathscr{R}$ is said to belong to class $\widetilde{\mathrm{Z}}$ iff the four-acceleration A^a of observers in $\mathscr{R}$ satisfies $\oint_\eta A_a=0$, where η is any closed, continuous, piecewise differentiable simultaneity curve.

We introduce class $\widetilde{\mathrm{Z}}$ here because this class is closely related to class Z. In fact, we have shown (although not presented in this paper) that they are equivalent locally in most cases. In addition, Zhao claimed in Ref. 3 the equivalence of transitivity of clock rate synchronizability and transitivity of thermal equilibrium. Although we have found some clues implying that class $\widetilde{\mathrm{Z}}$ reference frames may be interpreted as thermal equilibrium transitive reference frames, our investigation on this issue is far from mature.

ACKNOWLEDGMENTS

We would like to thank the referee for his valuable suggestion of writing this revised version of our paper. The first author would like to thank Professor Zhao Zheng for calling his attention to Ref. 3 and his encouragement during the investigation of this work. This work has been supported by the Natural Science Foundation of China.

[1] R. Sachs and H. Wu, *General Relativity for Mathematicians* (Springer-Verlag, New York, 1977).
[2] R. Wald, *General Relativity* (Univ. of Chicago, Chicago, 1984).
[3] Z. Zheng, Sci. China A **34(7)**, 835 (1991).

Multidimensional classical and quantum cosmology with intersecting *p*-branes

V. D. Ivashchuk and V. N. Melnikov[a)]
Center for Gravitation and Fundamental Metrology, VNIIMS, 3-1 M. Ulyanovoy Str., Moscow, 117313, Russia

(Received 9 September 1997; accepted for publication 6 October 1997)

A multidimensional cosmological model describing the evolution of $(n+1)$ Einstein spaces in the theory with several scalar fields and forms is considered. When a (electro-magnetic composite) p-brane ansatz is adopted the field equations are reduced to the equations for the Toda-like system. The Wheeler–De Witt equation is obtained. In the case when n ''internal'' spaces are Ricci-flat, one space M_0 has a non-zero curvature, and all p-branes do not ''live'' in M_0; the classical and quantum solutions are obtained if certain orthogonality relations on parameters are imposed. Spherically symmetric solutions with intersecting non-extremal p-branes are singled out. A non-orthogonal generalization of intersection rules corresponding to (open, closed) Toda lattices is obtained. A chain of bosonic $D \geqslant 11$ models (that may be related to hypothetical higher dimensional supergravities and F-theories) is suggested. © *1998 American Institute of Physics.* [S0022-2488(98)02504-3]

I. INTRODUCTION

In this paper we continue our investigations of p-brane solutions (see, for example, Refs. 1–3 and references therein) based on the sigma-model approach.[4–6] (For pure gravitational sector see Refs. 7 and 8.)

Here we consider a cosmological case, when all functions depend upon one variable (time). The model under consideration contains several scalar fields and antisymmetric forms and is governed by action (2.1).

The considered cosmological model contains some stringy cosmological models (see, for example, Ref. 9). It may be obtained (at the classical level) from the multidimensional cosmological model with perfect fluid[10–14] as an interesting special case.

The paper is organized as follows. In Sec. II the model with composite electro-magnetic p-branes is described. In Sec. III the σ-model representation (under certain constraints on p-branes) is considered and certain scalar invariants are calculated. In Sec. IV the Wheeler–De Witt equation is obtained. Section V is devoted to exact solutions (classical and quantum) for the orthogonal case, when one factor space is curved. In Subsec. V C spherically symmetric configurations with non-extremal p-branes are singled out. In Sec. VI the intersection rules are generalized to the non-orthogonal case and a chain of bosonic B_D-models ($D \geqslant 11$) containing a ''truncated'' $D=11$ supergravity model[15] and a $D=12$ model from Ref. 16 is suggested. (It may be supposed that these models or their analogues may be connected with higher dimensional generalization of M- and F-theories.[17–20])

II. THE MODEL

Here as in Ref. 5 we consider the model governed by the action

$$S=\frac{1}{2\kappa^2}\int_M d^D z\sqrt{|g|}\{R[g]-2\Lambda-h_{\alpha\beta}g^{MN}\partial_M\varphi^\alpha\partial_N\varphi^\beta \tag{2.1}$$

a)Electronic mail: melnikov@rgs.phys.msu.su

 © 1998 American Institute of Physics

$$-\sum_{a\in\Delta}\frac{\theta_a}{n_a!}\exp[2\lambda_a(\varphi)](F^a)^2_g\}+S_{GH},$$

where $g=g_{MN}dz^M\otimes dz^N$ is the metric $(M,N=1,\ldots,D)$, $\varphi=(\varphi^\alpha)\in\mathbb{R}^l$ is a vector from dilatonic scalar fields, $(h_{\alpha\beta})$ is a symmetric non-degenerate $l\times l$ matrix $(l\in\mathbb{N})$, $\theta_a=\pm 1$,

$$F^a=dA^a=\frac{1}{n_a!}F^a_{M_1\ldots M_{n_a}}dz^{M_1}\wedge\ldots\wedge dz^{M_{n_a}} \tag{2.2}$$

is an n_a-form $(n_a\geqslant 1)$ on a D-dimensional manifold M, Λ is a cosmological constant and λ_a is a 1-form on $\mathbb{R}^l$: $\lambda_a(\varphi)=\lambda_{a\alpha}\varphi^\alpha$, $a\in\Delta$, $\alpha=1,\ldots,l$. In (2.1) we denote $|g|=|\det(g_{MN})|$,

$$(F^a)^2_g=F^a_{M_1\ldots M_{n_a}}F^a_{N_1\ldots N_{n_a}}g^{M_1N_1}\ldots g^{M_{n_a}N_{n_a}}, \tag{2.3}$$

$a\in\Delta$, where Δ is some finite set, and $S_{\rm GH}$ is the standard Gibbons-Hawking boundary term.[21] In the models with one time all $\theta_a=1$ when the signature of the metric is $(-1,+1,\ldots,+1)$.

The equations of motion corresponding to (2.1) have the following form:

$$R_{MN}-\frac{1}{2}g_{MN}R=T_{MN}-\Lambda g_{MN}, \tag{2.4}$$

$$\Delta[g]\varphi^\alpha-\sum_{a\in\Delta}\theta_a\frac{\lambda^\alpha_a}{n_a!}e^{2\lambda_a(\varphi)}(F^a)^2_g=0, \tag{2.5}$$

$$\nabla_{M_1}[g](e^{2\lambda_a(\varphi)}F^{a,M_1\ldots M_{n_a}})=0, \tag{2.6}$$

$a\in\Delta$; $\alpha=1,\ldots,l$. In (2.5) $\lambda^\alpha_a=h^{\alpha\beta}\lambda_{a\beta}$, where $(h^{\alpha\beta})$ is matrix inverse to $(h_{\alpha\beta})$. In (2.4)

$$T_{MN}=T_{MN}[\varphi,g]+\sum_{a\in\Delta}\theta_a e^{2\lambda_a(\varphi)}T_{MN}[F^a,g], \tag{2.7}$$

where

$$T_{MN}[\varphi,g]=h_{\alpha\beta}\left(\partial_M\varphi^\alpha\partial_N\varphi^\beta-\frac{1}{2}g_{MN}\partial_P\varphi^\alpha\partial^P\varphi^\beta\right), \tag{2.8}$$

$$T_{MN}[F^a,g]=\frac{1}{n_a!}\left[-\frac{1}{2}g_{MN}(F^a)^2_g+n_aF^a_{MM_2\ldots M_{n_a}}F^{a,M_2\ldots M_{n_a}}_N\right]. \tag{2.9}$$

In (2.5) and (2.6) $\Delta[g]$ and $\nabla[g]$ are Laplace–Beltrami and covariant derivative operators, respectively, corresponding to g.

Let us consider the manifold

$$M=\mathbb{R}\times M_0\times\ldots\times M_n \tag{2.10}$$

with the metric

$$g=we^{2\gamma(u)}du\otimes du+\sum_{i=0}^{n}e^{2\phi^i(u)}g^i, \tag{2.11}$$

where $w=\pm 1$, u is a distinguished coordinate which, by convention, will be called ''time''; $g^i=g^i_{m_in_i}(y_i)dy_i^{m_i}\otimes dy_i^{n_i}$ is a metric on M_i satisfying the equation

$$R_{m_in_i}[g^i]=\xi_i g^i_{m_in_i}, \tag{2.12}$$

$m_i, n_i = 1, \ldots, d_i$; $d_i = \dim M_i$, $\xi_i = \text{const}$, $i = 0, \ldots, n$; $n \in \mathbf{N}$. Thus, (M_i, g^i) are Einstein spaces. The functions γ, ϕ^i: $(u_-, u_+) \to \mathbf{R}$ are smooth.

Each manifold M_i is assumed to be oriented and connected, $i = 0, \ldots, n$. Then the volume d_i-form

$$\tau_i = \sqrt{|g^i(y_i)|}\, dy_i^1 \wedge \ldots \wedge dy_i^{d_i}, \tag{2.13}$$

and the signature parameter

$$\varepsilon(i) = \text{ sign } \det(g^i_{m_i n_i}) = \pm 1, \tag{2.14}$$

are correctly defined for all $i = 0, \ldots, n$.

Let

$$\Omega_0 = \{\varnothing, \{0\}, \{1\}, \ldots, \{n\}, \{0,1\}, \ldots, \{0,1, \ldots, n\}\} \tag{2.15}$$

be a set of all subsets of

$$I_0 \equiv \{0, \ldots, n\}. \tag{2.16}$$

For any $I = \{i_1, \ldots, i_k\} \in \Omega_0$, $i_1 < \ldots < i_k$, we define a form

$$\tau(I) \equiv \tau_{i_1} \wedge \ldots \wedge \tau_{i_k} \tag{2.17}$$

of rank

$$d(I) \equiv \sum_{i \in I} d_i = d_{i_1} + \ldots + d_{i_k}, \tag{2.18}$$

and a corresponding p-brane submanifold

$$M_I \equiv M_{i_1} \times \ldots \times M_{i_k}, \tag{2.19}$$

where $p = d(I) - 1$ ($\dim M_I = d(I)$). We also define an ε-symbol

$$\varepsilon(I) \equiv \varepsilon(i_1) \ldots \varepsilon(i_k). \tag{2.20}$$

For $I = \varnothing$ we put $\tau(\varnothing) = \varepsilon(\varnothing) = 1$, $d(\varnothing) = 0$.

For fields of forms we adopt the following ''composite electro-magnetic'' ansatz

$$F^a = \sum_{I \in \Omega_{a,e}} F^{(a,e,I)} + \sum_{J \in \Omega_{a,m}} F^{(a,m,J)}, \tag{2.21}$$

where

$$F^{(a,e,I)} = d\Phi^{(a,e,I)} \wedge \tau(I), \tag{2.22}$$

$$F^{(a,m,J)} = \mathrm{e}^{-2\lambda_a(\varphi)} * (d\Phi^{(a,m,J)} \wedge \tau(J)), \tag{2.23}$$

$a \in \Delta$, $I \in \Omega_{a,e}$, $J \in \Omega_{a,m}$ and

$$\Omega_{a,e}, \Omega_{a,m} \subset \Omega_0. \tag{2.24}$$

(For empty $\Omega_{a,v} = \varnothing$, $v = e, m$, we put $\Sigma_\varnothing = 0$ in (2.21).) In (2.23) $* = *[g]$ is the Hodge operator on (M, g).

For the potentials in (2.22) and (2.23) we put

$$\Phi^s = \Phi^s(u), \tag{2.25}$$

$s \in S$, where

$$S = S_e \sqcup S_m, \quad S_v \equiv \sqcup_{a \in \Delta} \{a\} \times \{v\} \times \Omega_{a,v}, \tag{2.26}$$

$v = e, m$.

For dilatonic scalar fields we put

$$\varphi^\alpha = \varphi^\alpha(u), \tag{2.27}$$

$\alpha = 1, \ldots, l$.

From (2.22) and (2.23) we obtain the relations between dimensions of p-brane world-sheets and ranks of forms

$$d(I) = n_a - 1, \quad I \in \Omega_{a,e}, \tag{2.28}$$

$$d(J) = D - n_a - 1, \quad J \in \Omega_{a,m}, \tag{2.29}$$

in the electric and magnetic cases, respectively.

III. σ-MODEL REPRESENTATION

Restrictions on $\Omega_{a,v}$. Let

$$w_1 \equiv \{i | i \in \{1, \ldots, n\}, d_i = 1\}. \tag{3.1}$$

The set w_1 describes all 1-dimensional manifolds among M_i ($i \geqslant 0$). We impose the following restrictions on the sets $\Omega_{a,v}$ (2.24):

$$W_{ij}(\Omega_{a,v}) = \varnothing, \tag{3.2}$$

$a \in \Delta$; $v = e, m$; $i, j \in w_1$, $i < j$ and

$$W_j^{(1)}(\Omega_{a,m}, \Omega_{a,e}) = \varnothing, \tag{3.3}$$

$a \in \Delta$; $j \in w_1$. Here

$$W_{ij}(\Omega_*) \equiv \{(I,J) | I, J \in \Omega_*, I = \{i\} \sqcup (I \cap J), J = \{j\} \sqcup (I \cap J)\}, \tag{3.4}$$

$i, j \in w_1$, $i \neq j$, $\Omega_* \subset \Omega_0$ and

$$W_j^{(1)}(\Omega_{a,m}, \Omega_{a,e}) \equiv \{(I,J) \in \Omega_{a,m} \times \Omega_{a,e} | \overline{I} = \{j\} \sqcup J\}, \tag{3.5}$$

$j \in w_1$. In (3.5)

$$\overline{I} \equiv I_0 \backslash I \tag{3.6}$$

is a ''dual'' set. (The restrictions (3.2) and (3.3) are trivially satisfied when $n_1 \leqslant 1$ and $n_1 = 0$, respectively, where $n_1 = |w_1|$ is the number of 1-dimensional manifolds among M_i.)

It follows from Ref. 5 (see Proposition 2 in Ref. 5) that the equations of motion (2.4)–(2.6) and the Bianchi identities

$$dF^s = 0, \quad s \in S, \tag{3.7}$$

for the field configuration (2.11), (2.21)–(2.25), and (2.27) with the restrictions (3.2) and (3.3) imposed are equivalent to equations of motion for the σ-model with the action

$$S_\sigma = \frac{\mu}{2} \int du \mathscr{N} \left\{ G_{ij} \dot{\phi}^i \dot{\phi}^j + h_{\alpha\beta} \dot{\varphi}^\alpha \dot{\varphi}^\beta + \sum_{s \in S} \varepsilon_s \exp[-2U^s(\phi, \varphi)] (\dot{\Phi}^s)^2 - 2\mathscr{N}^{-2} V(\phi) \right\}, \tag{3.8}$$

where $\dot{x} \equiv dx/du$,

$$V = V(\phi) = -w\Lambda\, \mathrm{e}^{2\gamma_0(\phi)} + \frac{w}{2}\sum_{i=0}^{n} \xi_i d_i\, \mathrm{e}^{-2\phi^i + 2\gamma_0(\phi)} \tag{3.9}$$

is the potential with

$$\gamma_0(\phi) \equiv \sum_{i=0}^{n} d_i \phi^i, \tag{3.10}$$

$$\mathcal{N} = \exp(\gamma_0 - \gamma) > 0 \tag{3.11}$$

is the lapse function,

$$U^s = U^s(\phi, \varphi) = -\chi_s \lambda_{a_s}(\varphi) + \sum_{i \in I_s} d_i \phi^i, \tag{3.12}$$

$$\varepsilon_s = (-\varepsilon[g])^{(1-\chi_s)/2} \varepsilon(I_s)\, \theta_{a_s} \tag{3.13}$$

for $s = (a_s, v_s, I_s) \in S$, $\varepsilon[g] = \mathrm{sign}\ \det(g_{MN})$,

$$\chi_s = +1, \quad v_s = e, \tag{3.14}$$

$$\chi_s = -1, \quad v_s = m, \tag{3.15}$$

and

$$G_{ij} = d_i \delta_{ij} - d_i d_j \tag{3.16}$$

are components of the ''pure cosmological'' minisupermetric, $i, j = 0, \ldots, n$.[22]

In the electric case ($F^{(a,m,I)} = 0$) for finite internal space volumes V_i the action (3.8) coincides with the action (2.1) if $\mu = -w/\kappa_0^2$, $\kappa^2 = \kappa_0^2 V_0 \ldots V_n$.

Action (3.8) may be also written in the form

$$S_\sigma = \frac{\mu}{2} \int du \mathcal{N} \{ \mathcal{G}_{\hat{A}\hat{B}}(X) \dot{X}^{\hat{A}} \dot{X}^{\hat{B}} - 2\mathcal{N}^{-2} V(X) \}, \tag{3.17}$$

where $X = (X^{\hat{A}}) = (\phi^i, \varphi^\alpha, \Phi^s) \in \mathbf{R}^N$, and minisupermetric

$$\mathcal{G} = \mathcal{G}_{\hat{A}\hat{B}}(X) dX^{\hat{A}} \otimes dX^{\hat{B}} \tag{3.18}$$

on minisuperspace

$$\mathcal{M} = \mathbf{R}^N, \quad N = n + 1 + l + |S| \tag{3.19}$$

($|S|$ is the number of elements in S) is defined by the relation

$$(\mathcal{G}_{\hat{A}\hat{B}}(X)) = \begin{pmatrix} G_{ij} & 0 & 0 \\ 0 & h_{\alpha\beta} & 0 \\ 0 & 0 & \varepsilon_s\, \mathrm{e}^{-2U^s(X)} \delta_{ss'} \end{pmatrix}. \tag{3.20}$$

A. Scalar products

The minisuperspace metric (3.18) may be also written in the form

$$\mathcal{G} = \overline{G} + \sum_{s \in S} \varepsilon_s\, \mathrm{e}^{-2U^s(x)} d\Phi^s \otimes d\Phi^s, \tag{3.21}$$

where $x=(x^A)=(\phi^i,\varphi^\alpha)$,

$$\overline{G}=\overline{G}_{AB}dx^A\otimes dx^B=G_{ij}d\phi^i\otimes d\phi^j+h_{\alpha\beta}d\varphi^\alpha\otimes d\varphi^\beta, \tag{3.22}$$

$$(\overline{G}_{AB})=\begin{pmatrix} G_{ij} & 0 \\ 0 & h_{\alpha\beta}\end{pmatrix}, \tag{3.23}$$

$U^s(x)=U^s_A x^A$ is defined in (3.12) and

$$(U^s_A)=(d_i\delta_{iI_s},-\chi_s\lambda_{a_s\alpha}). \tag{3.24}$$

Here

$$\delta_{iI}\equiv\sum_{j\in I}\delta_{ij}=\begin{matrix}1, & i\in I,\\ 0, & i\notin I,\end{matrix} \tag{3.25}$$

is an indicator of i belonging to I. The potential (3.9) reads

$$V=(-w\Lambda)\,\mathrm{e}^{2U^\Lambda(x)}+\sum_{j=0}^{n}\frac{w}{2}\xi_j d_j\,\mathrm{e}^{2U^j(x)}, \tag{3.26}$$

where

$$U^j(x)=U^j_A x^A=-\phi^j+\gamma_0(\phi), \tag{3.27}$$

$$U^\Lambda(x)=U^\Lambda_A x^A=\gamma_0(\phi), \tag{3.28}$$

$$(U^j_A)=(-\delta^j_i+d_i,0), \tag{3.29}$$

$$(U^\Lambda_A)=(d_i,0). \tag{3.30}$$

The integrability of the Lagrange system (3.17) crucially depends upon the scalar products of co-vectors U^Λ, U^j, U^s corresponding to $\overline{G}$:

$$(U,U')=\overline{G}^{AB}U_AU'_B, \tag{3.31}$$

where

$$(\overline{G}^{AB})=\begin{pmatrix} G^{ij} & 0 \\ 0 & h^{\alpha\beta}\end{pmatrix} \tag{3.32}$$

is matrix inverse to (3.23). Here (as in Ref. 22)

$$G^{ij}=\frac{\delta^{ij}}{d_i}+\frac{1}{2-D}, \tag{3.33}$$

$i,j=0,\ldots,n$. These products have the following form:

$$(U^i,U^j)=\frac{\delta_{ij}}{d_j}-1, \tag{3.34}$$

$$(U^i,U^\Lambda)=-1, \tag{3.35}$$

$$(U^\Lambda,U^\Lambda)=-\frac{D-1}{D-2}, \tag{3.36}$$

$$(U^s,U^{s'})=q(I_s,I_{s'})+\chi_s\chi_{s'}\lambda_{a_s}\cdot\lambda_{a_{s'}}, \tag{3.37}$$

$$(U^s,U^i)=-\delta_{iI_s}, \tag{3.38}$$

$$(U^s,U^\Lambda)=\frac{d(I_s)}{2-D}, \tag{3.39}$$

where $s=(a_s,v_s,I_s)$, $s'=(a_{s'},v_{s'},I_{s'})\in S$,

$$q(I,J)\equiv d(I\cap J)+\frac{d(I)d(J)}{2-D}, \tag{3.40}$$

$$\lambda_a\cdot\lambda_b\equiv\lambda_{a\alpha}\lambda_{b\beta}h^{\alpha\beta}. \tag{3.41}$$

Relations (3.34)–(3.36) were found in Ref. 14 and (3.37) in Ref. 5.

IV. WHEELER–DE WITT EQUATION

Here we fix the gauge as follows:

$$\gamma_0-\gamma=f(X),\quad \mathcal{N}=e^f, \tag{4.1}$$

where $f:\mathcal{M}\to\mathbf{R}$ is a smooth function. Then we obtain the Lagrange system with the Lagrangian

$$L_f=\frac{\mu}{2}\,\mathrm{e}^f\mathcal{G}_{\hat{A}\hat{B}}(X)\dot{X}^{\hat{A}}\dot{X}^{\hat{B}}-\mu\,\mathrm{e}^{-f}V \tag{4.2}$$

and the energy constraint

$$E_f=\frac{\mu}{2}\,\mathrm{e}^f\mathcal{G}_{\hat{A}\hat{B}}(X)\dot{X}^{\hat{A}}\dot{X}^{\hat{B}}+\mu\,\mathrm{e}^{-f}V=0. \tag{4.3}$$

Using the standard prescriptions of (covariant and conformally covariant) quantization (see, for example, Refs. 22–24) we are led to the Wheeler–De Witt (WDW) equation

$$\hat{H}^f\Psi^f\equiv\left(-\frac{1}{2\mu}\Delta[e^f\mathcal{G}]+\frac{a}{\mu}R[e^f\mathcal{G}]+e^{-f}\mu V\right)\Psi^f=0, \tag{4.4}$$

where

$$a=a_N=\frac{N-2}{8(N-1)}. \tag{4.5}$$

Here $\Psi^f=\Psi^f(X)$ is the so-called ''wave function of the universe'' corresponding to the f-gauge (4.1), and satisfying the relation

$$\Psi^f=e^{bf}\Psi^{f=0},\quad b=(2-N)/2, \tag{4.6}$$

($\Delta[\mathcal{G}_1]$ and $R[\mathcal{G}_1]$ denote the Laplace–Beltrami operator and the scalar curvature corresponding to $\mathcal{G}_1$).

For the scalar curvature of minisupermetric (3.21) we get [see (2.29) in Ref. 5)]

$$R[\mathcal{G}]=-\sum_{s\in S}(U^s,U^s)-\sum_{s,s'\in S}(U^s,U^{s'}). \tag{4.7}$$

For the Laplace operator we obtain

$$\Delta[\mathcal{G}] = \mathrm{e}^{U(x)} \frac{\partial}{\partial x^A}\left(\overline{G}^{AB} \mathrm{e}^{-U(x)} \frac{\partial}{\partial x^B}\right) + \sum_{s\in S} \varepsilon_s \mathrm{e}^{2U^s(x)} \left(\frac{\partial}{\partial \Phi^s}\right)^2, \tag{4.8}$$

where $U(x)=\Sigma_{s\in S}U^s(x)$.

Harmonic-time gauge. The WDW Eq. (4.4) for $f=0$,

$$\hat{H}\Psi \equiv \left(-\frac{1}{2\mu}\Delta[\mathcal{G}] + \frac{a}{\mu} R[\mathcal{G}] + \mu V\right)\Psi = 0, \tag{4.9}$$

may be rewritten, using relations (4.7), (4.8) and

$$U^{si} = G^{ij}U^s_j = \delta_{iI_s} - \frac{d(I_s)}{D-2}, \quad U^{s\alpha} = -\chi_s \lambda^{\alpha}_{a_s}, \tag{4.10}$$

as follows:

$$\begin{aligned} 2\mu \hat{H}\Psi = \Bigg\{ & -G^{ij}\frac{\partial}{\partial \phi^i}\frac{\partial}{\partial \phi^j} - h^{\alpha\beta}\frac{\partial}{\partial \varphi^\alpha}\frac{\partial}{\partial \varphi^\beta} - \sum_{s\in S}\varepsilon_s \mathrm{e}^{2U^s(\phi,\varphi)}\left(\frac{\partial}{\partial \Phi^s}\right)^2 \\ & + \sum_{s\in S}\left[\sum_{i\in I_s}\frac{\partial}{\partial \phi^i} - \frac{d(I_s)}{D-2}\sum_{j=0}^{n}\frac{\partial}{\partial \phi^j} - \chi_s \lambda^{\alpha}_{a_s}\frac{\partial}{\partial \varphi^\alpha}\right] + 2aR[\mathcal{G}] + 2\mu^2 V\Bigg\}\Psi = 0. \end{aligned} \tag{4.11}$$

Here $\hat{H}\equiv \hat{H}^{f=0}$ and $\Psi \equiv \Psi^{f=0}$.

V. EXACT SOLUTIONS WITH ONE CURVED FACTOR SPACE AND ORTHOGONAL U^s

Here we put the following restrictions on the parameters of the model:

$$\textbf{(i)} \quad \Lambda = 0, \tag{5.1}$$

i.e., the cosmological constant is zero;

$$\textbf{(ii)} \quad \xi_0 \neq 0, \quad \xi_1 = \ldots = \xi_n = 0, \tag{5.2}$$

i.e., one space is curved and others are Ricci-flat;

$$\textbf{(iii)} \quad 0 \notin I_s, \quad \forall s = (a_s, v_s, I_s) \in S, \tag{5.3}$$

i.e., all ''brane'' submanifolds M_{I_s} [see (2.19)] do not contain M_0.

We also put the following orthogonality restriction on the vectors U^s:

$$\textbf{(iv)} \quad (U^s, U^{s'}) = d(I_s \cap I_{s'}) + \frac{d(I_s)d(I_{s'})}{2-D} + \chi_s \chi_{s'} \lambda_{a_s \alpha}\lambda_{a_{s'}\beta}h^{\alpha\beta} = 0, \tag{5.4}$$

for $s\neq s'$, $s=(a_s, v_s, I_s)$, $s'=(a_{s'}, v_{s'}, I_{s'})\in S$ and

$$\textbf{(v)} \quad (U^s, U^s) \neq 0 \tag{5.5}$$

for all $s\in S$.

From **(i)** and **(ii)** we get for the potential (3.26)

$$V = \frac{1}{2} w \xi_0 d_0 \mathrm{e}^{2U^0(x)}, \tag{5.6}$$

where

$$(U^0,U^0)=\frac{1}{d_0}-1<0 \tag{5.7}$$

[see (3.34)].

From **(iii)** and (3.38) we get

$$(U^0,U^s)=0 \tag{5.8}$$

for all $s\in S$. Thus, all co-vectors U^0, U^s ($s\in S$) belonging to dual space $(\mathbf{R}^{n+1+l})^*\simeq\mathbf{R}^{n+1+l}$ are orthogonal with respect to the scalar product (3.31).

Let

$$\eta_s\equiv \operatorname{sign}(U^s,U^s), \tag{5.9}$$

and $|S|_\pm$ is the number of U^s with $\eta_s=\pm 1$; $|S|_+ + |S|_- = |S|$. Let the matrix $(h_{\alpha\beta})$ have the signature $(-1,\ldots,-1_{l_-},+1,\ldots,+1_{l_+})$ ($l_- + l_+ = l$). The matrix (G_{ij}) has a pseudo-Euclidean signature $(-1,+1,\ldots,+1_n)$.[22] Then from (5.7)–(5.8) we obtain $|S|_-\leqslant l_-$, $|S|_+\leqslant n+l_+$ and hence

$$|S|\leqslant n+l. \tag{5.10}$$

Thus we obtain the restriction on the number of orthogonal p-brane configurations.

A. Quantum solutions

The truncated minisuperspace metric (3.23) may be diagonalized by the linear transformation

$$z^A=S^A{}_B x^B, \quad (z^A)=(z^0,z^a,z^s) \tag{5.11}$$

as follows:

$$\overline{G}=\overline{G}_{AB}dx^A\otimes dx^B=-dz^0\otimes dz^0+\sum_{s\in S}\eta_s dz^s\otimes dz^s+dz^a\otimes dz^b\,\eta_{ab}\,, \tag{5.12}$$

where $a,b=1,\ldots,n+l-|S|$; $\eta_{ab}=\eta_{aa}\delta_{ab}$; $\eta_{aa}=\pm 1$, and

$$q_0 z^0=U^0(x), \tag{5.13}$$

$$q_s z^s=U^s(x), \tag{5.14}$$

with

$$q_0\equiv\sqrt{|(U^0,U^0)|}=\sqrt{1-\frac{1}{d_0}}>0, \tag{5.15}$$

$$q_s=\nu_s^{-1}\equiv\sqrt{|(U^s,U^s)|}=\sqrt{\left|d(I_s)\left(1+\frac{d(I_s)}{2-D}\right)+\lambda_{a_s}^2\right|}>0, \tag{5.16}$$

$s=(a_s,v_s,I_s)\in S$.

From (4.8), (5.11), (5.12) and (5.14) we get

$$\Delta[\mathcal{G}]=-\left(\frac{\partial}{\partial z^0}\right)^2+\eta^{ab}\frac{\partial}{\partial z^a}\frac{\partial}{\partial z^b}+\sum_{s\in S}\eta_s\,\mathrm{e}^{q_s z^s}\frac{\partial}{\partial z^s}\left(\mathrm{e}^{-q_s z^s}\frac{\partial}{\partial z^s}\right)$$

$$+\sum_{s\in S}\varepsilon_s\,\mathrm{e}^{2q_s z^s}\left(\frac{\partial}{\partial\Phi^s}\right)^2. \tag{5.17}$$

The relation (4.7) in the orthogonal case reads as

$$R[\mathcal{G}] = -2\sum_{s\in S}\,(U^s,U^s) = -2\sum_{s\in S}\,\eta_s q_s^2\,. \tag{5.18}$$

We are seeking the solution to WDW Eq. (4.9) by the method of the separation of variables, i.e., we put

$$\Psi_*(z) = \Psi_0(z^0)\left(\prod_{s\in S}\,\Psi_s(z^s)\right)\,\mathrm{e}^{\mathrm{i}P_s\Phi^s}\,\mathrm{e}^{\mathrm{i}p_a z^a}. \tag{5.19}$$

It follows from (5.17) that $\Psi_*(z)$ satisfies WDW Eq. (4.9) if

$$2\hat{H}_0\Psi_0 \equiv \left\{\left(\frac{\partial}{\partial z^0}\right)^2 + \mu^2 w\xi_0 d_0\,\mathrm{e}^{2q_0 z^0}\right\}\Psi_0 = 2\mathcal{E}_0\Psi_0\,, \tag{5.20}$$

$$2\hat{H}_s\Psi_s \equiv \left\{-\eta_s\,\mathrm{e}^{q_s z^s}\frac{\partial}{\partial z^s}\left(\,\mathrm{e}^{-q_s z^s}\frac{\partial}{\partial z^s}\right) + \varepsilon_s P_s^2\,\mathrm{e}^{2q_s z^s}\right\}\Psi_s = 2\mathcal{E}_s\Psi_s\,, \tag{5.21}$$

$s\in S$, and

$$2\mathcal{E}_0 + \eta^{ab}p_a p_b + 2\sum_{s\in S}\,\mathcal{E}_s + 2aR[\mathcal{G}] = 0, \tag{5.22}$$

with a and $R[\mathcal{G}]$ from (4.5) and (5.18), respectively.

Using the relations from Appendix A we obtain linearly independent solutions to (5.20) and (5.21), respectively,

$$\Psi_0(z^0) = B^0_{\omega_0(\mathcal{E}_0)}\left(\sqrt{-w\mu^2\xi_0 d_0}\,\frac{\mathrm{e}^{q_0 z^0}}{q_0}\right), \tag{5.23}$$

$$\Psi_s(z^s) = \,\mathrm{e}^{q_s z^s/2}B^s_{\omega_s(\mathcal{E}_s)}\left(\sqrt{\eta_s\varepsilon_s P_s^2}\,\frac{\mathrm{e}^{q_s z^s}}{q_s}\right), \tag{5.24}$$

where

$$\omega_0(\mathcal{E}_0) = \sqrt{2\mathcal{E}_0}/q_0\,, \quad \omega_s(\mathcal{E}_s) = \sqrt{\frac{1}{4} - 2\,\eta_s\mathcal{E}_s\nu_s^2}, \tag{5.25}$$

$s\in S$ and $B^0_\omega, B^s_\omega = I_\omega, K_\omega$ are the modified Bessel function.

The general solution of the WDW Eq. (4.9) is a superposition of the ''separated'' solutions (5.19):

$$\Psi(z) = \sum_B\,\int dp\,dP\,d\mathcal{E}\,C(p,P,\mathcal{E},B)\Psi_*(z|p,P,\mathcal{E},B), \tag{5.26}$$

where $p=(p_a)$, $P=(P_s)$, $\mathcal{E}=(\mathcal{E}_s)$, $B=(B^0,B^s)$, $B^0,B^s = I,K$; and $\Psi_* = \Psi_*(z|p,P,\mathcal{E},B)$ is given by relation (5.19) and (5.23)–(5.25) with $\mathcal{E}_0$ from (5.22). Here $C(p,P,\mathcal{E},B)$ are smooth enough functions. In the non-composite electric case these solutions were considered recently in Ref. 25.

B. Classical solutions

1. Toda-like representation

Here we will integrate the Lagrange equations corresponding to the Lagrangian (4.2) with the energy constraint (4.3) and hence we will find classical exact solutions when the restrictions (5.1)–(5.5) are imposed. We put $f=0$, i.e., the harmonic time gauge is considered.

The problem of integrability may be simplified if we integrate the Maxwell equations (for $s \in S_e$) and Bianchi identities (for $s \in S_m$):

$$\frac{d}{du}[\exp(-2U^s)\dot{\Phi}^s]=0 \Leftrightarrow \dot{\Phi}^s=Q_s \exp(2U^s), \tag{5.27}$$

where Q_s are constants, $s=(a_s,v_s,I_s)\in S$.

Let

$$Q_s \neq 0, \quad s \in S_*; \tag{5.28}$$

$$Q_s=0, \quad s \in S \backslash S_*,$$

where $S_* \subset S$ is a non-empty subset of S.

For fixed $Q=(Q_s, s\in S_*)$ the Lagrange equations for the Lagrangian (4.2) with $f=0$ corresponding to $(x^A)=(\phi^i,\varphi^\alpha)$, when equations (5.27) are substituted, are equivalent to the Lagrange equations for the Lagrangian

$$L_Q=\frac{1}{2}\overline{G}_{AB}\dot{x}^A\dot{x}^B-V_Q, \tag{5.29}$$

where

$$V_Q=V+\frac{1}{2}\sum_{s\in S_*}\varepsilon_s Q_s^2 \exp[2U^s(x)], \tag{5.30}$$

and $(\overline{G}_{AB})$ and V are defined in (3.23) and (3.9), respectively. The zero-energy constraint (4.3) reads

$$E_Q=\frac{1}{2}\overline{G}_{AB}\dot{x}^A\dot{x}^B+V_Q=0. \tag{5.31}$$

2. Exact solutions for one curved space and orthogonal U^s

When the conditions **(i)**–**(v)** are satisfied exact solutions for Lagrangian (5.29) with the potential (5.30), and V from (5.6) could be readily obtained using the relations from Appendix B.

The solutions read

$$x^A(u)=-\frac{U^{0A}}{(U^0,U^0)}\ln|f_0(u-u_0)|-\sum_{s\in S_*}\frac{U^{sA}}{(U^s,U^s)}\ln|f_s(u-u_s)|+c^A u+\overline{c^A}, \tag{5.32}$$

where u_0 and u_s are constants and $s\in S_*$. Functions f_0 and f_s in (5.32) are the following:

$$f_0(\tau)=\left|\frac{\xi_0(d_0-1)}{C_0}\right|^{1/2}\mathrm{sh}(\sqrt{C_0}\tau), \quad C_0>0, \quad \xi_0 w>0; \tag{5.33}$$

$$=\left|\frac{\xi_0(d_0-1)}{C_0}\right|^{1/2}\sin(\sqrt{|C_0|}\tau), \quad C_0<0, \quad \xi_0 w>0; \tag{5.34}$$

$$=\left|\frac{\xi_0(d_0-1)}{C_0}\right|^{1/2}\mathrm{ch}(\sqrt{C_0}\tau), \quad C_0>0, \quad \xi_0 w<0; \tag{5.35}$$

$$=|\xi_0(d_0-1)|^{1/2}\tau, \quad C_0=0, \quad \xi_0 w>0, \tag{5.36}$$

and

$$f_s(\tau)=\frac{|Q_s|}{\nu_s|C_s|^{1/2}}\,\mathrm{sh}(\sqrt{C_s}\tau),\quad C_s>0,\quad \eta_s\varepsilon_s<0; \tag{5.37}$$

$$=\frac{|Q_s|}{\nu_s|C_s|^{1/2}}\,\sin(\sqrt{|C_s|}\tau),\quad C_s<0,\quad \eta_s\varepsilon_s<0; \tag{5.38}$$

$$=\frac{|Q_s|}{\nu_s|C_s|^{1/2}}\,\mathrm{ch}(\sqrt{C_s}\tau),\quad C_s>0,\quad \eta_s\varepsilon_s>0; \tag{5.39}$$

$$=\frac{|Q^s|}{\nu_s}\tau,\quad C_s=0,\quad \eta_s\varepsilon_s<0, \tag{5.40}$$

where C_0 and C_s are constants. Here we used the relations (5.7)–(5.9). The contravariant components $U^{0A}=\overline{G}^{AB}U^0_B$ are

$$U^{0i}=-\frac{\delta^i_0}{d_0},\quad U^{0\alpha}=0. \tag{5.41}$$

Corresponding relations for U^{sA}, $s\in S$, were presented in (4.10).

Using (5.32), (5.7), (5.16), (4.10) and (5.41) we obtain

$$\phi^i=\frac{\delta^i_0}{1-d_0}\,\ln|f_0|+\sum_{s\in S_*}\alpha^i_s\,\ln|f_s|+c^iu+\overline{c^i}, \tag{5.42}$$

where

$$\alpha^i_s=-\eta_s\nu_s^2\left(\delta_{iI_s}-\frac{d(I_s)}{D-2}\right), \tag{5.43}$$

$s\in S$, and

$$\varphi^\alpha=\sum_{s\in S_*}\eta_s\nu_s^2\chi_s\lambda^\alpha_{a_s}\,\ln|f_s|+c^\alpha u+\overline{c^\alpha}, \tag{5.44}$$

$\alpha=1,\ldots,l$.

Vectors $c=(c^A)$ and $\overline{c}=(\overline{c^A})$ satisfy the linear constraint relations (see Appendix B)

$$U^0(c)=U^0_Ac^A=-c^0+\sum_{j=0}^n d_jc^j=0, \tag{5.45}$$

$$U^0(\overline{c})=U^0_A\overline{c^A}=-\overline{c^0}+\sum_{j=0}^n d_j\overline{c^j}=0, \tag{5.46}$$

$$U^s(c)=U^s_Ac^A=\sum_{i\in I_s}d_ic^i-\chi_s\lambda_{a_s\alpha}c^\alpha=0, \tag{5.47}$$

$$U^s(\overline{c})=U^s_A\overline{c^A}=\sum_{i\in I_s}d_i\overline{c^i}-\chi_s\lambda_{a_s\alpha}\overline{c^\alpha}=0, \tag{5.48}$$

$s\in S$. The harmonic gauge function (3.10) reads

$$\gamma_0(\phi)=\frac{d_0}{1-d_0}\,\ln|f_0|+\sum_{s\in S_*}\frac{d(I_s)}{D-2}\,\eta_s\nu_s^2\,\ln|f_s|+c^0u+\overline{c^0}. \tag{5.49}$$

The zero-energy constraint reads (see Appendix B)

$$E = E_0 + \sum_{s \in S_*} E_s + \frac{1}{2} \overline{G}_{AB} c^A c^B = 0, \tag{5.50}$$

where $C_0 = 2E_0(U^0, U^0)$, and $C_s = 2E_s(U^s, U^s)$. Using the relations (5.15), (5.16), and (5.45) we rewrite (5.50) as

$$C_0 \frac{d_0}{d_0 - 1} = \sum_{s \in S_*} C_s \nu_s^2 \eta_s + h_{\alpha\beta} c^\alpha c^\beta + \sum_{i=1}^{n} d_i (c^i)^2 + \frac{1}{d_0 - 1} \left(\sum_{i=1}^{n} d_i c^i \right)^2. \tag{5.51}$$

From the relation

$$\exp 2U^s = f_s^{-2}, \tag{5.52}$$

following from (5.32), (5.4), (5.8), (5.47), and (5.48) we get for electric-type forms (2.22)

$$F^s = Q_s f_s^{-2} du \wedge \tau(I_s), \tag{5.53}$$

$s \in S_e$, and for magnetic-type forms (2.23)

$$F^s = \mathrm{e}^{-2\lambda_a(\varphi)} * [Q_s f_s^{-2} du \wedge \tau(I_s)] = \overline{Q}_s \tau(\overline{I}_s), \tag{5.54}$$

$s \in S_m$, where $\overline{Q}_s = Q_s \varepsilon(I_s) \mu(I_s) w$ and $\mu(I) = \pm 1$ is defined by the relation $\mu(I) du \wedge \tau(I_0) = \tau(\overline{I}) \wedge du \wedge \tau(I)$. The relation (5.54) may be readily obtained using the formula (5.28) from Ref. 5 (for $\gamma = \gamma_0$).

Relations for the metric follows from (5.42), (5.43) and (5.49):

$$g = \left(\prod_{s \in S_*} [f_s^2(u - u_s)]^{\eta_s d(I_s) \nu_s^2/(D-2)} \right) \left\{ [f_0^2(u - u_0)]^{d_0/(1-d_0)} \mathrm{e}^{2c^0 u + 2\overline{c^0}} \right.$$

$$\left. \times [w du \otimes du + f_0^2(u - u_0) g^0] + \sum_{i \neq 0} \left(\prod_{\substack{s \in S_* \\ I_s \ni i}} [f_s^2(u - u_s)]^{-\eta_s \nu_s^2} \right) \mathrm{e}^{2c^i u + 2\overline{c^i}} g^i \right\}. \tag{5.55}$$

Thus we obtained exact solutions for multidimensional cosmology, describing the evolution of $(n+1)$ spaces $(M_0, g_0), \ldots, (M_n, g_n)$, where (M_0, g_0) is an Einstein space of non-zero curvature, and (M_i, g^i) are ''internal'' Ricci-flat spaces, $i = 1, \ldots, n$; in the presence of several scalar fields and forms. The solution is presented by relations (5.44) and (5.53)–(5.55) with the functions f_0, f_s defined in (5.33)–(5.40) and the relations on the parameters of solutions c^A, $\overline{c^A}$ $(A = i, \alpha)$, C_0, C_s $(s \in S_*)$, and ν_s, imposed in (5.45)–(5.48), (5.51), and (5.16), respectively.

This solution describes a set of charged (by forms) overlapping p-branes ($p_s = d(I_s) - 1$, $s \in S_*$) ''living'' on submanifolds M_{I_s} (2.19), where the sets I_s do not contain 0, i.e., all p-branes live in ''internal'' Ricci-flat spaces.

The solution is valid if the dimensions of the p-branes and the dilatonic coupling vector satisfy the relations (5.4). In the non-composite case these solutions were considered recently in Refs. 25,26 (electric case), and 27 (electro-magnetic case). For $n = 1$, see also Refs. 9 and 28.

C. Spherically symmetric solutions

To illustrate the general solution from Subsec. V B let us consider also the spherically symmetric case

$$w = 1, \quad M_0 = S^{d_0}, \quad g^0 = d\Omega_{d_0}^2, \tag{5.56}$$

where $d\Omega^2_{d_0}$ is canonical metric on unit sphere S^{d_0}. We also assume that $M_1=\mathbf{R}$, $g^1=-dt\otimes dt$ and

$$1\in I_s, \quad \forall s\in S_*, \tag{5.57}$$

i.e., all p-branes have common time direction t. Let

$$\eta_s\varepsilon_s=-1, \tag{5.58}$$

$s\in S_*$. For integration constants we put $\overline{c^A}=0$,

$$c^A=\overline{\mu}\sum_{r\in\overline{S}}\frac{U^{rA}}{(U^r,U^r)}-\overline{\mu}\delta^A_1, \tag{5.59}$$

$$C_0=C_s=\overline{\mu^2}, \tag{5.60}$$

where $\overline{\mu}>0$ and $\overline{S}=\{0\}\cup S$. Here $A=(i_A,\alpha_A)$ and $A=1$ means $i_A=1$. It may be verified that the restrictions (5.45)–(5.48) and (5.51) are satisfied identically.

We also introduce a new radial variable $R=R(u)$ by relations

$$\exp(-2\overline{\mu}u)=1-\frac{2\mu}{R^{\overline{d}}}, \quad \mu=\overline{\mu^{\overline{d}}}>0, \quad \overline{d}=d_0-1, \tag{5.61}$$

and put $u_0=0$, $u_s<0$,

$$\frac{|Q_s|}{\overline{\mu}\nu_s}\sinh\beta_s=1, \quad \beta_s\equiv\overline{\mu}|u_s|, \quad s\in S_*. \tag{5.62}$$

Then, solutions for the metric and scalar fields (see (5.44) and (5.55)) are the following

$$g=\left(\prod_{s\in S_*}H_s^{2\eta_s d(I_s)\nu_s^2/(D-2)}\right)\left\{\frac{dR\otimes dR}{1-2\mu/R^{\overline{d}}}+R^2d\Omega^2_{d_0}\right.$$
$$\left.-\left(\prod_{s\in S_*}H_s^{-2\eta_s\nu_s^2}\right)\left(1-\frac{2\mu}{R^d}\right)dt\otimes dt+\sum_{i=2}^{n}\left(\prod_{\substack{s\in S_*\\ I_s\ni i}}H_s^{-2\eta_s\nu_s^2}\right)g^i\right\}, \tag{5.63}$$

$$\varphi^\alpha=\sum_{s\in S_*}\eta_s\nu_s^2\chi_s\lambda^\alpha_{a_s}\ln H_s, \tag{5.64}$$

where

$$H_s=1+\frac{\mathscr{P}_s}{R^d}, \quad \mathscr{P}_s\equiv\frac{|Q_s|\overline{d}}{\nu_s}e^{-\beta_s}, \tag{5.65}$$

$s\in S_*$.

The fields of forms are given by (2.22) and (2.23) with

$$\Phi^s=\frac{\nu_s}{H_s'}, \tag{5.66}$$

$$H_s'=\left(1-\frac{\mathscr{P}_s'}{H_sR^d}\right)^{-1}=1+\frac{\mathscr{P}_s'}{R^d+\mathscr{P}_s-\mathscr{P}_s'}, \tag{5.67}$$

$$\mathscr{P}'_s \equiv -\frac{Q_s \overline{d}}{\nu_s}, \tag{5.68}$$

$s \in S_*$. It follows from (5.62), (5.65) and (5.68) that

$$|\mathscr{P}'_s| = \frac{\mu}{\sinh \beta_s} = \mathscr{P}_s e^{\beta_s} = \sqrt{\mathscr{P}_s(\mathscr{P}_s + 2\mu)}, \tag{5.69}$$

$s \in S_*$.

The solutions obtained describe non-extremal charged intersecting p-branes and agree with those from Refs. 29,30,31 ($d_1 = \ldots = d_n = 1$, $\eta_s = +1$) and Ref. 27 ($\eta_s = +1$, non-composite case).

We note that Hawking temperature corresponding to the solution is (see also Refs. 27 and 31)

$$T_H = \frac{\overline{d}}{4\pi (2\mu)^{1/\overline{d}}} \prod_{s \in S_*} \left(\frac{2\mu}{2\mu + \mathscr{P}_s} \right)^{\eta_s \nu_s^2}. \tag{5.70}$$

Recall that $\eta_s \nu_s^2 = (U^s, U^s)^{-1}$.

D. WDW equation with fixed charges

We may consider also another scheme based on zero-energy constraint relation (5.31). The corresponding WDW equation in the harmonic gauge reads

$$\hat{H}_Q \Psi \equiv \left(-\frac{1}{2\mu} \overline{G}^{AB} \frac{\partial}{\partial x^A} \frac{\partial}{\partial x^B} + \mu V_Q \right) \Psi = 0, \tag{5.71}$$

where potential V_Q is defined in (5.30). This equation describes quantum cosmology with classical fields of forms and quantum scale factors and dilatonic fields. Such an approach is equivalent to the scheme of quantization of multidimensional perfect fluid considered in Ref. 10.

Equation (5.71) is readily solved in the orthogonal case (5.4). The basis of solutions is given by the following replacements in (5.19), (5.22), (5.24) and (5.25), respectively:

$$P_s \mapsto Q_s, \tag{5.72}$$

$$2aR[\mathscr{S}] \mapsto 0, \tag{5.73}$$

$$\Psi_s(z^s) \mapsto B^s_{\omega_s}\left(\sqrt{\eta_s \varepsilon_s Q_s^2} \frac{e^{q_s z^s}}{q_s} \right), \tag{5.74}$$

$$\omega_s \mapsto \sqrt{-2\eta_s \mathscr{E}_s \nu_s^2}. \tag{5.75}$$

We note that recently such a solution for the special case with one internal space ($n = 1$) and non-composite p-branes was considered in Ref. 32.

VI. INTERSECTION RULES FOR INTEGRABLE CONFIGURATIONS WITH ORTHOGONAL AND NON-ORTHOGONAL U^s

In this section we analyze in detail the orthogonality relation (5.4). According to (2.28), (2.29), (5.3) and (5.4) there exist some obstacles for the existence of binary configurations, i.e., solutions with two p-branes. For these reasons some restrictions on the model (2.1) (e.g., on dimension D, ranks of forms n_a and scalar products for dilatonic couplings $\lambda_a \cdot \lambda_b$) and the manifold (2.10) (e.g., on dimensions $d_0, d_1, \ldots d_n$) should be imposed. For example, there are no binary configurations when: (i) all $\lambda_a \cdot \lambda_b$ are irrational; (ii) d_0 is big enough, i.e., the dimension of ''internal'' space (where all p-branes live) $D - d_0 - 1$ is too small; (iii) dimensions of factor spaces d_i ($i \geqslant 1$) do not fit the ranks of forms n_a [see (2.28) and (2.29)].

In Subsec. VI A we propose a sufficiency condition for the existence of ''full'' spectrum of p-brane pairs (see Propositions 1 and 2). This condition is formulated in terms of the so-called fundamental matrix of the model.

In Subsec. VI B we suggest a chain of $D \geqslant 11$ models (''beautiful'' models). These models satisfy conditions of Proposition 1 for $d_0=2$. They may be considered as a nice polygon for investigation of solutions with intersecting p-branes and possible generalization to higher dimensional supergravitational and super-p-brane models (F_D- or M_D-theories).

In Subsec. VI C we consider the generalization of the orthogonality relation leading to some other integrable Euclidean Toda-like Lagrangians and find corresponding intersection rules. This result may be considered as a ''bridge'' between Toda lattices (open, closed, etc.) and intersecting p-branes. It opens a new area for use of Lie algebras (e.g., affine Lie algebras, etc.) in the (classical and quantum) models with p-branes.

A. Orthogonal U^s

In Sec. V we obtained exact cosmological solutions for the model (2.1) with $\Lambda=0$ defined on the manifold (2.10). In (2.10) M_0 is an Einstein space of non-zero curvature and hence $d_0 = \dim M_0 > 1$. All p-branes ''live'' in internal space

$$\overline{M}=M_1\times \ldots \times M_n \tag{6.1}$$

((M_i,g^i) are Ricci-flat). Here we treat the existence of the solutions with two p-branes. Let us consider such a solution with a set of indices $S_*=\{s_1,s_2\}$, $s_1\neq s_2$, where $s_1=(a_{s_1},v_{s_1},I_{s_1})$, $s_2=(a_{s_2},v_{s_2},I_{s_2})$; $a_{s_1},a_{s_2}\in\Delta$; $v_{s_1},v_{s_2}\in\{e,m\}$; $I_{s_1},I_{s_2}\subset\{1,\ldots,n\}$.

The orthogonality relation (5.4) defines the intersection rules

$$d(I_{s_1}\cap I_{s_2})=\Delta(s_1,s_2), \tag{6.2}$$

where

$$\Delta(s_1,s_2)\equiv \frac{d(I_{s_1})d(I_{s_2})}{D-2}-\chi_{s_1}\chi_{s_2}\lambda_{a_{s_1}}\cdot\lambda_{a_{s_2}}, \tag{6.3}$$

$s_1\neq s_2$. It follows from (6.2) that in orthogonal case

$$\Delta(s_1,s_2)\in \mathbf{Z}_+\equiv\{0,1,2,\ldots\}. \tag{6.4}$$

The intersection symbol (6.3) is symmetric: $\Delta(s_1,s_2)=\Delta(s_2,s_1)$. In the orthogonal case it satisfies the obvious restrictions

$$\Delta(s_1,s_2)\leqslant \min(d(I_{s_1}),d(I_{s_2})), \tag{6.5}$$

$$d(I_{s_1}\cup I_{s_2})=d(I_{s_1})+d(I_{s_2})-\Delta(s_1,s_2)\leqslant D-d_0-1=\dim\ \overline{M}. \tag{6.6}$$

The latter follows from the inclusion $M_{I_{s_1}\cup I_{s_2}}\subset\overline{M}$.

We recall that dimensions $d(I_s)$ obey the relations (2.28) and (2.29) that may be written as follows:

$$d(I_s)=\overline{D}\overline{\chi_s}+\overline{n}_{a_s}\chi_s, \tag{6.7}$$

where $\overline{D}=D-2$, $\overline{n_a}=n_a-1$ and $\overline{\chi_s}=\frac{1}{2}(1-\chi_s)=0,1$ $(\chi_s=1,-1)$ for $v_s=e,m$, respectively.

Let

$$N(a,b)\equiv \frac{(n_a-1)(n_b-1)}{D-2}-\lambda_a\cdot\lambda_b, \tag{6.8}$$

$a,b \in \Delta$. We call $[N(a,b)]$ fundamental matrix of the model (2.1).

From (6.3) and (6.7) we obtain

$$\Delta(s_1,s_2)=\overline{D\chi_{s_1}}\overline{\chi}_{s_2}+\overline{n}_{a_{s_1}}\chi_{s_1}\overline{\chi}_{s_2}+\overline{n}_{a_{s_2}}\chi_{s_2}\overline{\chi}_{s_1}+N(a_{s_1},a_{s_2})\chi_{s_1}\chi_{s_2}, \tag{6.9}$$

or, more explicitly,

$$\Delta(s_1,s_2)=N(a_{s_1},a_{s_2}), \quad v_{s_1}=v_{s_2}=e; \tag{6.10}$$

$$=\overline{n}_{a_{s_1}}-N(a_{s_1},a_{s_2}), \quad v_{s_1}=e, \quad v_{s_2}=m; \tag{6.11}$$

$$=\overline{D}-\overline{n}_{a_{s_1}}-\overline{n}_{a_{s_2}}+N(a_{s_1},a_{s_2}), \quad v_{s_1}=v_{s_2}=m. \tag{6.12}$$

Thus $N(a,b)$ is the dimension of intersection for two electric p-branes charged by forms F^a and F^b, respectively. The fundamental matrix defines the intersection rules in the electro-magnetic case, too.

Remark 1: The matrix (6.8) is unchanged if $\overline{n}_a \mapsto k\overline{n}_a$, $\overline{D}\mapsto k^2\overline{D}$, $k\in\mathbf{N}$. For example, we may consider the chain of models with $D=9k^2+2$, $n=\text{rank } F=3k+1$ originating from the truncated bosonic part of $D=11$ supergravity.

Proposition 1: Let matrix (6.8) satisfy the following restrictions:

$$(\mathbf{A}) \quad N(a_1,a_2)\in\mathbf{Z}_+; \quad N(a_1,a_2)\geqslant d_0-1,$$

$$(\mathbf{B}) \quad N(a_1,a_2)\leqslant \min(n_{a_1},n_{a_2})-d_0,$$

$$(\mathbf{C}) \quad N(a_1,a_2)\geqslant n_{a_1}+n_{a_2}-D+d_0-1,$$

for any $a,b\in\Delta$ ($d_0>1$). Then $\Delta(s_1,s_2)$ from (6.9) satisfies consistency relations (6.4)–(6.6) for any $s_1\neq s_2$ [with $d(I_s)$ from (6.7)].

Proof: Let us consider three cases: (ee) $v_{s_1}=v_{s_2}=e$; (em) $v_{s_1}=e$, $v_{s_2}=m$; (mm) $v_{s_1}=v_{s_2}=m$. Relation (6.4) follows from restrictions (**A**), (**B**) and (**C**) for (ee), (em) and (mm) cases, respectively. Analogously (6.5) follows from (**B**), (**C**) and (**B**) and (6.6) follows from (**C**), (**B**) and (**A**) for (ee), (em) and (mm) cases, respectively. The proposition is proved.

From relation (**B**) and $d_0>1$ we get

$$K(a)\equiv n_a-1-N(a,a)>0 \tag{6.13}$$

and hence

$$(U^s,U^s)=K(a_s)>0. \tag{6.14}$$

Thus, inequalities (5.5) are satisfied.

Definition 1: The model (2.1) with $\Lambda=0$ defined on the manifold (2.10) is called binary complete if for any $a_1,a_2\in\Delta$, $v_1,v_2\in\{e,m\}$ there exists cosmological solution from V B.2 with two intersecting p-branes defined by indices $s_1=(a_1,v_1,I_1)\neq s_2=(a_2,v_2,I_2)$.

For the binary complete model we have the maximal number $|\Delta|(2|\Delta|+1)$ of different types of binary configurations (formed by $|\Delta|$ electric and $|\Delta|$ magnetic p-branes).

Definition 2: The matrix (6.8) satisfying the restrictions (**A**), (**B**) and (**C**) of Proposition 1 is called d_0-proper.

From Proposition 1 and (6.14) we obtain the following.

Proposition 2: The model (2.1) with $\Lambda=0$ defined on the manifold (2.10) with $M_1=\ldots M_n=\mathbf{R}$ ($n=D-1-d_0$, $d_0>1$) is binary complete if fundamental matrix (6.8) of this model is d_0-proper.

In the next subsection we apply these propositions for $D=11$ supergravity and the chain of $D>11$ models.

B. Example: chain of B_D-models

Let us consider the action in dimension D

$$S_D=\int_M d^Dz\sqrt{|g|}\left\{R[g]+g^{MN}\partial_M\vec{\varphi}\partial_N\vec{\varphi}-\sum_{a=4}^{D-7}\frac{1}{a!}\exp[2\vec{\lambda}_a\vec{\varphi}](F^a)^2\right\},\tag{6.15}$$

where $\vec{\varphi}=(\varphi^1,\ldots,\varphi^l)\in\mathbf{R}^l$, $\vec{\lambda}_a=(\lambda_{a1},\ldots,\lambda_{al})\in\mathbf{R}^l$, $l=D-11$, rank $F^a=a$, and $a=4,\ldots,D-7$. Here vectors $\vec{\lambda}_a$ satisfy the relations

$$\vec{\lambda}_a\vec{\lambda}_b=N(a,b)-\frac{(a-1)(b-1)}{D-2},\tag{6.16}$$

$$N(a,b)=\min(a,b)-3,\tag{6.17}$$

$a,b=4,\ldots,D-7$.

The vectors $\vec{\lambda}_a$ are linearly dependent:

$$\vec{\lambda}_{D-7}=-2\vec{\lambda}_4.\tag{6.18}$$

For $D>11$ vectors $\vec{\lambda}_4,\ldots,\vec{\lambda}_{D-8}$ are linearly independent.

The model (6.15) contains l scalar fields with negative kinetic term (i.e., $h_{\alpha\beta}=-\delta_{\alpha\beta}$ in (2.1)) coupled with $l+1$ forms. For $D=11$ ($l=0$) the model (6.15) coincides with the truncated bosonic sector of $D=11$ supergravity (''truncated'' means without Chern–Simons term). For $D=12$ ($l=1$) (6.15) coincides with the truncated $D=12$ model from Ref. 16 (see also Ref. 5). We call the model (6.15) the B_D-model.

The matrix (6.17) is a fundamental matrix [see (6.8)] of the model (since $\lambda_a\cdot\lambda_b=-\vec{\lambda}_a\vec{\lambda}_b$, $n_a=a$).

The dimensions of p-brane worldsheets are [see (6.7)]

$$d(I)=\begin{cases}3,\ldots,D-8, & I\in\Omega_{a,e},\\ D-5,\ldots,6, & I\in\Omega a,m.\end{cases}\tag{6.19}$$

(6.20)

We have $(l+1)$ electric and $(l+1)$ magnetic p-branes $p=d(I)-1$. The intersection rules are given by (6.9). For $D=12$ see also Ref. 5.

Fundamental matrix (6.17) is d_0-proper, iff $d_0=2$ and hence the B_D-model is binary complete if the manifold decomposition (2.10) with $M_1=\ldots M_n=\mathbf{R}$, $n=D-3$, $d_0=2$ is chosen. So, for spherically symmetric solutions the binary-completeness takes place only in ''non-Tangerlini'' case $M_0=S^2$.

We note that in the B_D-models $(U^s,U^s)=2>0$, since

$$K(a)=2\tag{6.21}$$

[see (6.14)] for all $a=4,\ldots,D-7$.

Remark 2: Here two problems arise. The first one is the existence of the chain of extensions

$$B_D\leftarrow\hat{B}_D\leftarrow SG_D\leftarrow M_D=F_D,\tag{6.22}$$

where $M_D=F_D$ is M- or F-theory in dimension D, SG_D is supergravitational (field) theory coinciding with low-energy limit of F_D and $\hat{B}_D$ is the bosonic sector of SG_D. $\hat{B}_D$-action is the sum of the B_D-action and the Chern–Simons terms. The second problem is the existence of relations between B_D-models via dimensional reduction

$$B_{11}\leftarrow B_{12}\leftarrow B_{13}\leftarrow B_{14}\leftarrow\ldots.\tag{6.23}$$

For $B_{11} \leftarrow B_{12}$ see Ref. 16.

We note that B_D-models have also rather interesting classes of Toda lattice solutions. The intersection rules for these ''Toda p-branes'' may be obtained using general relations described in the next subsection. (To our knowledge p-brane solutions governed by open Toda lattices with a_n Lie algebras were studied first in Refs. 9 and 28.)

C. Generalizations to Toda lattices

Let us consider the Lagrange system (5.29) and (5.30), where V (curvature part of potential) is defined in (5.6). Here we put

$$(U^s, U^s) > 0, \quad s \in S_*. \tag{6.24}$$

The Lagrangian (5.29) may be ''block-diagonalized'' in z-variables (5.11) satisfying (5.12), (5.13), $\eta_s = +1$ and

$$B^s_{s'} z^{s'} = U^s(x), \tag{6.25}$$

$s \in S_*$, instead (5.14). Vectors $\vec{B}^s = (B^s_{s'})$ obey the relations

$$\vec{B}^{s_1} \vec{B}^{s_2} = (U^{s_1}, U^{s_2}), \tag{6.26}$$

$s_1, s_2 \in S_*$. In z-variables the Lagrangian (5.29) reads

$$L_Q = L_0 + L_{T,Q} + L_f, \tag{6.27}$$

where

$$L_0 = -\frac{1}{2}(\dot{z}^0)^2 - \frac{1}{2} w \xi_0 d_0 \, \mathrm{e}^{2q_0 z^0}, \tag{6.28}$$

$$L_{T,Q} = \frac{1}{2} \sum_{s \in S_*} (\dot{z}^s)^2 - \frac{1}{2} \sum_{s \in S_*} \varepsilon_s Q_s^2 \exp(2 \vec{B}^s \vec{z}), \tag{6.29}$$

$$L_f = \frac{1}{2} \eta_{ab} \dot{z}^a \dot{z}^b, \tag{6.30}$$

$\vec{z} = (z^s)$; $\varepsilon_s Q_s \neq 0$, $s \in S_*$.

Thus, the problem of integrability of the Lagrange system (6.27) is reduced to the problem of integrability of the Euclidean Toda-like system with the Lagrangian (6.29).

Let vectors $\vec{B}^s$ (or, equivalently, U^s), $s \in S_*$, be linearly independent. According to Adler–van Moerbeke criterion[33] the equations of motion corresponding to (6.29) are integrable in quadratures if and only if

$$\frac{2 \vec{B}^{s_1} \vec{B}^{s_2}}{\vec{B}^{s_2} \vec{B}^{s_2}} = \frac{2(U^{s_1}, U^{s_2})}{(U^{s_2}, U^{s_2})} = C_{s_1 s_2}, \tag{6.31}$$

where $C = (C_{s_1 s_2})$ is the Cartan matrix for some semisimple Lie algebra $\mathbf{g}$. When $\mathbf{g}$ is simple $L_{T,Q}$ describes a Toda lattice corresponding to $\mathbf{g}$[34–36]. From (5.4), (6.14), (6.3) and (6.31) we obtain the intersection rules

$$d(I_{s_1} \cap I_{s_2}) = \Delta(s_1, s_2) + \frac{1}{2} K(a_{s_2}) C_{s_1 s_2}, \tag{6.32}$$

$s_1 \neq s_2$, where $\Delta(s_1, s_2)$ is defined in (6.3) and $K(a)$ in (6.13).

When $\mathrm{rank}(\vec{B}^s, s \in S_*) = |S_*| - 1$ (i.e., only one vector linearly depends upon others) the Adler–van Moerbeke criterion has the form (6.31) with the Cartan matrix corresponding to some affine Lie algebra.[33] In this case a closed Toda lattice arises.

Thus, when relations (6.13) and (6.32) are satisfied, the cosmological model is integrable since it is reduced effectively to (open or closed) a Toda lattice.

Uninorm models. Let $K(a) = K$, $a \in \Delta$, and $|S| \geqslant 2$. Then from (6.32) we get that the Cartan matrix is symmetric. Hence $\mathbf{g}$ is of A-D-E type (or simply laced), i.e., it belongs to one of a_n, d_n, e_n series. In this case $C_{s_1 s_2} = 0, -1$, for $s_1 \neq s_2$ and -1 takes place for some $s_1 \neq s_2$. Let $\Delta(s_1, s_2) \in \mathbf{Z}$. Then, it follows from (6.32) that K should be even.

Example 1: In $D = 11$ supergravity $K(a) = 2$. For $S = \{s_1, s_2\}$ and $\mathbf{g} = a_2 = \mathrm{sl}(3)$ we get the intersection rule

$$d(I_{s_1} \cap I_{s_2}) = \Delta(s_1, s_2) - 1 = 0, 1, 3 \tag{6.33}$$

for $\{d(I_{s_1}), d(I_{s_2})\} = \{3,3\}, \{3,6\}, \{6,6\}$, respectively. We see that two membranes are intersecting in a point.

Example 2: Let us consider truncated $D = 11$ supergravity and put $|S| = n = 3$, $d_1 = d_2 = d_3 = 3$, $d_0 = 1$. Let $\Omega_e = \{1,2,3\}$, i.e., we consider three non-intersecting electric 2-branes (Euclidean for $w = -1$) attached to M_1, M_2, M_3, respectively. In this case $C_{s_1 s_2} = -1$ for $s_1 \neq s_2$ (the Dynkin diagram is a triangle) and the Lagrangian $L_{T,Q}$ from (6.29) is closed Toda-lattice Lagrangian corresponding to affine Lie algebra $a_2^{(1)}$. (In this case the curvature term in Lagrangian L_0 from (6.28) is absent and (5.13) should be modified.) The corresponding solutions are expressed in terms of θ-functions.[37]

Example 3: Let $K(a) = 2$ and $n_a = 1$ for all $a \in \Delta$. From (6.3) and (6.32) we get

$$\chi_{s_1} \chi_{s_2} \lambda_{a_{s_1}} \cdot \lambda_{a_{s_2}} = C_{s_1 s_2}. \tag{6.34}$$

This relation implies the appearance of E_N (open) Toda lattices in maximal supergravities in D dimensions coming from $D = 11$ supergravity.[32]

We see that the appearance of A-D-E algebras seems to be rather typical for supergravitational models (with $K(a) = 2$).

VII. CONCLUSIONS

In this paper we obtained exact solutions to Einstein and Wheeler–De Witt equations for the multidimensional cosmological model describing the evolution of n Ricci-flat spaces and one Einstein space M_0 of non-zero curvature in the presence of composite electro-magnetic p-branes. The solutions were obtained in orthogonal case (5.4), when p-branes do not ''live'' in M_0. As an illustration we singled out the spherically-symmetric solutions with intersecting non-extremal p-branes.

In the classical case we reduced the model to the Euclidean Toda-like system. Applying the Adler–van Moerbeke criterion we obtained (Toda) intersection rules (6.32) that imply the reduction of the cosmological model to (open or closed) Toda lattice. For uniform models [$K(a) = K$, $a \in \Delta$] simply laced (A-D-E) Lie algebras appear. We gave some examples of reduction to Toda lattice (e.g., closed one). Concrete exact solutions will be considered in a separate publication.

Also we developed (general) formalism of (orthogonal) intersections using the notions of fundamental matrix and binary completeness of the model and gave examples of binary complete models for $d_0 = \dim M_0 = 2$ (B_D-models). We may suppose that investigations of intersecting (e.g., Toda) p-branes may be a powerful tool in constructing new higher dimensional models (that may have supersymmetric or non-local extensions).

ACKNOWLEDGMENTS

This work was supported in part by DFG grants 436 RUS 113/7, 436 RUS 113/236/O(R) and by the Russian Ministry for Science and Technology, Russian Fund for Basic Research, project N 95-02-05785-a.

APPENDIX A: SOLUTIONS WITH BESSEL FUNCTIONS

Let us consider two differential operators

$$2\hat{H}_0 = -\frac{\partial^2}{\partial z^2} + 2A\mathrm{e}^{2qz}, \tag{A1}$$

$$2\hat{H}_1 = -\ \mathrm{e}^{qz}\frac{\partial}{\partial z}\left(\ \mathrm{e}^{-qz}\frac{\partial}{\partial z}\right) + 2A\ \mathrm{e}^{2qz}. \tag{A2}$$

Equation

$$H_k\Psi_k = \mathscr{E}\Psi_k \tag{A3}$$

has the following linearly independent solutions for $q \neq 0$:

$$\Psi_k(z) = \mathrm{e}^{kqz/2}B_{\omega_k(\mathscr{E})}\left(\sqrt{2A}\frac{\mathrm{e}^{qz}}{q}\right), \tag{A4}$$

$$\omega_k(\mathscr{E}) = \sqrt{\frac{k}{4} - \frac{2\mathscr{E}}{q^2}}, \tag{A5}$$

where $k = 0,1$ and $B_\omega = I_\omega, K_\omega$ are modified Bessel function.

APPENDIX B: CLASSICAL SOLUTIONS WITH ORTHOGONAL VECTORS

Let

$$L = \frac{1}{2}\langle \dot{x}, \dot{x}\rangle - \sum_{s=1}^{m} A_s\ \exp[2\langle b_s, x\rangle] \tag{B1}$$

be a Lagrangian, defined on $V \times V$, where V is n-dimensional vector space, over $\mathbf{R}$, $A_s \neq 0$, $s = 1,\ldots,m$; $m \leqslant n$, and $\langle \cdot,\cdot \rangle$ is non-degenerate real-valued quadratic form on V. Let

$$\langle b_s, b_s\rangle \neq 0, \tag{B2}$$

$$\langle b_s, b_l\rangle = 0, \quad s \neq l, \tag{B3}$$

$s, l = 1,\ldots,m$. Denote

$$\eta_s = \ \mathrm{sign}\langle b_s, b_s\rangle. \tag{B4}$$

Then, the Euler–Lagrange equations for the Lagrangian (B1) have the following solutions:[14]

$$x(t) = -\sum_{s=1}^{m}\frac{b_s}{\langle b_s, b_s\rangle}\ \ln|f_s(t - t_{0s})| + t\alpha + \beta, \tag{B5}$$

where $\alpha, \beta \in V$,

$$\langle \alpha, b_s\rangle = \langle \beta, b_s\rangle = 0, \tag{B6}$$

and

$$f_s(\tau)=\left|\frac{A_s}{E_s}\right|^{1/2}\mathrm{sh}(\sqrt{C_s}\tau),\quad C_s>0,\quad A_s\eta_s<0, \tag{B7}$$

$$=\left|\frac{A_s}{E_s}\right|^{1/2}\sin(\sqrt{|C_s|}\tau),\quad C_s<0,\quad A_s\eta_s<0, \tag{B8}$$

$$=\left|\frac{A_s}{E_s}\right|^{1/2}\mathrm{ch}(\sqrt{C_s}\tau),\quad C_s>0,\quad A_s\eta_s>0, \tag{B9}$$

$$=|2A_s\langle b_s,b_s\rangle|^{1/2}\tau,\quad C_s=0,\quad A_s\eta_s<0, \tag{B10}$$

$C_s=2E_s\langle b_s,b_s\rangle$, and E_s, t_{0s} are constants.

For the energy corresponding to the solution (B5) we have

$$E=\sum_{s=1}^{m}E_s+\frac{1}{2}\langle\alpha,\alpha\rangle. \tag{B11}$$

For dual vectors $u^s\in V^*$ defined as $u^s(x)=\langle b_s,x\rangle$, $\forall x\in V$, we have $\langle u^s,u^l\rangle_*=\langle b_s,b_l\rangle$, where $\langle\cdot,\cdot\rangle_*$ is dual form on V^*. The orthogonality conditions (B6) read

$$u^s(\alpha)=u^s(\beta)=0, \tag{B12}$$

$s=1,\ldots,m$.

[1] M. J. Duff, R. R. Khuri, and J. X. Lu, Phys. Rep. **259**, 213 (1995).
[2] K. S. Stelle, ''Lectures on Supergravity p-branes,'' hep-th/9701088.
[3] I. Ya. Aref'eva and O. A. Rytchkov, ''Incidence Matrix Description of Intersecting p-brane Solutions,'' hep-th/9612236; I. Ya. Aref'eva, M. G. Ivanov, and O. A. Rytchkov, ''Properties of Intersecting p-branes in Various Dimensions,'' hep-th/9702077.
[4] V. D. Ivashchuk and V. N. Melnikov, ''Intersecting p-brane Solutions in Multidimensional Gravity and M-theory,'' hep-th/9612089; Gravit. Cosmol. **2**(4), 297 (1996); V. D. Ivashchuk and V. N. Melnikov, Phys. Lett. B **403**, 23 (1997).
[5] V. D. Ivashchuk and V. N. Melnikov, ''Sigma-model for the Generalized Composite p-branes,'' hep-th/9705036; Class. Quantum Gray **14**, 3001 (1997).
[6] V. D. Ivashchuk, V. N. Melnikov, and M. Rainer, ''Multidimensional σ-models with Composite Electric p-branes,'' gr-qc/9705005; to appear in Gravit. Cosmol.
[7] M. Rainer and A. Zhuk, Phys. Rev. D **54**, 6186 (1996).
[8] V. D. Ivashchuk and V. N. Melnikov, ''Multidimensional Gravity with Einstein Internal spaces,'' hep-th/9612054; Gravit. Cosmol. **2**(3), 177 (1996).
[9] H. Lü, S. Mukherji, C. N. Pope, and K.-W. Xu, ''Cosmological Solutions in String Theories,'' hep-th/9610107.
[10] V. D. Ivashchuk and V. N. Melnikov, Int. J. Mod. Phys. D **3**(4), 795 (1994).
[11] U. Bleyer, V. D. Ivashchuk, and V. N. Melnikov, Nucl. Phys. B **429**, 177 (1994).
[12] V. D. Ivashchuk and V. N. Melnikov, Gravit. Cosmol. **1**(3), 204 (1995).
[13] V. D. Ivashchuk and V. N. Melnikov, Class. Quantum Grav. **11**, 1793 (1994).
[14] V. R. Gavrilov, V. D. Ivashchuk, and V. N. Melnikov, J. Math. Phys. **36**, 5829 (1995).
[15] E. Cremmer, B. Julia, and J. Scherk, Phys. Lett. B **76**, 409 (1978).
[16] N. Khvengia, Z. Khvengia, H. Lü, and C. N. Pope, ''Toward Field Theory of F-Theory,'' hep-th/9703012.
[17] J. M. Schwarz, ''Lectures on Superstring and M-theory Dualities,'' hep-th/9607201.
[18] M. J. Duff, ''M-theory (the Theory Formerly Known as Strings),'' hep-th/9608117.
[19] C. Hull and P. Townsend, ''Unity of Superstring Dualities,'' Nucl. Phys. B **438**, 109 (1995); P. Horava and E. Witten, Nucl. Phys. B **460**, 506 (1996).
[20] C. M. Hull, ''String dynamics at strong coupling,'' Nucl. Phys. B **468**, 113 (1996); C. Vafa, ''Evidence for F-Theory,'' Nucl. Phys. B **469**, 403 (1996).
[21] G. W. Gibbons and S. W. Hawking, Phys. Rev. D **15**, 2752 (1977).
[22] V. D. Ivashchuk, V. N. Melnikov, and A. I. Zhuk, Nuovo Cimento B **104**, 575 (1989).
[23] C. W. Misner, in *Magic without Magic: John Archibald Wheeler*, edited by J. R. Klauder (Freeman, San Francisco, 1972).
[24] J. J. Halliwell, Phys. Rev. D **38**, 2468 (1988).
[25] M. A. Grebeniuk, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable Multidimensional Quantum Cosmology for Intersecting p-branes,'' gr-qc/9708031; Gravit. Cosmol. **3**(11), 240 (1997).
[26] K. A. Bronnikov, M. A. Grebeniuk, V. D. Ivashchuk, and V. N. Melnikov, ''Integrable Multidimensional Cosmology for Intersecting p-branes,'' Gravit. Cosmol. **3**(2), 105 (1997).
[27] K. A. Bronnikov, V. D. Ivashchuk, and V. N. Melnikov, ''The Reissner-Nordström Problem for Intersecting Electric and Magnetic P-branes,'' gr-qc/9710054; Gravit. Cosmol. **3**(11), 203 (1997).

[28] H. Lü, C. N. Pope, and K. W. Xu, ''Liouville and Toda Solitons in M-theory,'' hep-th/9604058.
[29] M. Cvetic and A. A. Tseytlin, Nucl. Phys. B **478**, 181 (1996).
[30] I. Ya. Aref'eva, M. G. Ivanov, and I. V. Volovich, ''Non-extremal Intersecting p-branes in Various Dimensions,'' hep-th/9702079.
[31] N. Ohta, ''Intersection Rules for Non-extreme p-branes,'' hep-th/9702164.
[32] H. Lü, J. Maharana, S. Mukherji and C. N. Pope, ''Cosmological Solutions, p-branes and the Wheeler–De Witt Equation,'' hep-th/9707182.
[33] M. Adler and P. van Moerbeke, Commun. Math. Phys. **83**, 83 (1982).
[34] O. I. Bogoyavlensky, Commun. Math. Phys. **51**, 201 (1976).
[35] B. Kostant, Adv. Math. **34**, 195 (1979).
[36] M. A. Olshanetsky and A. M. Perelomov, Invent. Math. **54**, 261 (1979).
[37] I. M. Krichever, Usp. Mat. Nauk **33(4)**, 215 (1978) (in Russian).

Green functions of higher-order differential operators

Ivan G. Avramidi[a)]
Department of Mathematics, University of Greifswald,
F.-L.-Jahnstrasse 15a, D–17489 Greifswald, Germany
and Department of Mathematics, The University of Iowa, Iowa City, Iowa 52242-1419

(Received 10 July 1997; accepted for publication 4 December 1997)

The Green functions of the partial differential operators of even order acting on smooth sections of a vector bundle over a Riemannian manifold are investigated via the heat kernel methods. We study the resolvent of a special class of higher-order operators formed by the products of second-order operators of Laplace type defined with the help of a unique Riemannian metric but with different bundle connections and potential terms. The asymptotic expansion of the Green functions near the diagonal is studied in detail in any dimension. As a by-product a simple criterion for the validity of the Huygens principle is obtained. It is shown that all the singularities as well as the non-analytic regular parts of the Green functions of such high-order operators are expressed in terms of the usual heat kernel coefficients a_k for a special Laplace type second-order operator. © *1998 American Institute of Physics.* [S0022-2488(98)02905-3]

I. INTRODUCTION

The Green functions of partial differential operators are of great importance in mathematical physics and quantum theory.[1–4] In particular, the singularities of the Green functions on the diagonal play a crucial role in the renormalization procedure of quantum field theory.[5]

In this paper we study a special class of differential operators of higher orders that appear, in particular, in higher-derivative field theories (for example, in higher-derivative quantum gravity[6]). Namely, we consider differential operators of higher *even* orders, $2N$, of the form

$$H=(F_N+m^2)\cdots(F_2+m^2)(F_1+m^2), \tag{1}$$

where F_i, $(i=1,\ldots,N)$ are Laplace type operators, i.e. second-order operators with the scalar leading symbol determined by the metric of the manifold

$$\sigma_L(F_i)=|\xi|^2, \tag{2}$$

where $|\xi|^2\equiv g^{\mu\nu}(x)\xi_\mu\xi_\nu$, and m is some constant. In the next section we give precise definition of Laplace type operators and describe their properties. Obviously, the operator H also has a scalar leading symbol

$$\sigma_L(H)=|\xi|^{2N}, \tag{3}$$

and is a particular case of the so called minimal higher-derivative operators.

We would like to stress from the beginning that the operators F_i, even if of the same Laplace type, are *different* and, in general, *do not commute*. They only have the same leading, second-derivatives, part, but differ in the first and zero order parts.

The Green function, or the resolvent kernel, $G_{H-\lambda}(x,x')$ of the operator H is the kernel of the Green, or resolvent, operator $G_{H-\lambda}=(H-\lambda)^{-1}$ and is defined by requring it to satisfy the differential equation

$$(H-\lambda)G_{H-\lambda}(x,x')=\delta(x,x'), \tag{4}$$

[a)]On leave of absence from Research Institute for Physics, Rostov State University, Stachki 194, 344104 Rostov-on-Don, Russia. Electronic-mail: iavramid@math.uiowa.edu

 © 1998 American Institute of Physics

with H acting on the first argument of the Green function and $\delta(x,x')$ being the covariant delta-distribution.

In the present paper we will consider mainly the case of compact complete manifolds without boundary and assume the metric of the manifold to be positive definite. Then the Laplace type operators F_i as well as the operator H are *elliptic* and eq. (4) defines a unique Green function. We will also assume for simplicity the constant m to be sufficiently large so that all operators $(F_i + m^2)$ and also H are *positive*.

However, we would like to make some remarks concerning this subject:

(i) First, for manifolds with boundary (in elliptic case) one has to impose additionally some suitable boundary conditions to make the solution of eq. (4) unique. However, the singularities of the Green function near the diagonal and the asymptotic expansions of it as $m \to \infty$ do not depend on the boundary conditions.

(ii) Second, when the metric of the manifold has Minkowski type signature, sign $g=(-,+,\ldots,+)$, the Laplace type operators F_i are *hyperbolic*. In this case the Green function can be fixed by the Wick rotation, i.e. the analytic continuation from the Euclidean case, which is equivalent to adding an infinitesimal negative imaginary part to the constant m^2, $m^2 \to m^2 - i\varepsilon$. In other words, we consider the *Feynman propagators*. The formulas for the hyperbolic case can be obtained just by analytic continuation.

II. LAPLACE TYPE OPERATORS

Let (M,g) be a smooth Riemannian manifold of dimension d with a metric g. To simplify the exposition we assume the manifold M to be complete and compact, i.e., without boundary, $\partial M = \varnothing$, and the metric g to be positive definite.

Let $V(M)$ be a smooth vector bundle over the manifold M, End (V) be the bundle of all smooth endomorphisms of the vector bundle V, and $C^\infty(M,V)$ and $C^\infty(M,\mathrm{End}\,(V))$ be the spaces of all smooth sections of the vector bundles V and End (V). We assume, as usual, the vector bundle V to be Hermitian, i.e., there is a Hermitian pointwise fibre scalar product. Then the dual vector bundle V^* is naturally identified with V and a natural L^2 inner product is defined using the invariant Riemannian volume element $d\mathrm{vol}\,(x)$ on the manifold M. The completion of $C^\infty(M,V)$ in this norm defines the Hilbert space of square integrable sections of the vector bundle $L^2(M,V)$.

Let, further, ∇ be a connection, or covariant derivative, on the vector bundle V which is compatible with the Hermitian metric on the vector bundle V. Denoting by T^*M the cotangent bundle we define the tensor product connection on the tensor product bundle $T^*M \otimes V$ by means of the Levi-Civita connection. Similarly, we extend the connection ∇ with the help of the Levi-Civita connection to $C^\infty(M,V)$-valued tensors of all orders and denote it just by ∇. Usually there is no ambiguity and the precise meaning of the covariant derivative is always clear from the nature of the object it is acting on.

Let, further, tr_g denote the contraction of sections of the bundle $T^*M \otimes T^*M \otimes V$ with the metric on the cotangent bundle, and $Q \in C^\infty(M,\mathrm{End}(V))$ be a smooth Hermitian section of the endomorphism bundle. Then we define the generalized Laplacian

$$\Box = \mathrm{tr}_g \nabla\nabla \tag{5}$$

and a *Laplace type* differential operator F: $C^\infty(M,V) \to C^\infty(M,V)$ by

$$F = -\Box + Q. \tag{6}$$

Let x^μ, $(\mu = 1,2,\ldots,d)$, be a system of local coordinates and ∂_μ and dx^μ be the local coordinate frames for the tangent and the cotangent bundles. We adopt the notation that the Greek indices label the tensor components with respect to local coordinate frame and range from 1 through d. Besides, a summation is always carried out over repeated indices. Let $g_{\mu\nu} = (\partial_\mu, \partial_\nu)$ be the metric on the tangent bundle, $g^{\mu\nu} = (dx^\mu, dx^\nu)$ be the metric on the cotangent bundle, $g = \det g_{\mu\nu}$, and $\mathscr{A}_\mu$ be the connection 1-form on the vector bundle V. Then the generalized Laplacian reads

$$\Box = g^{\mu\nu}\nabla_\mu\nabla_\nu = g^{-1/2}(\partial_\mu + \mathscr{A}_\mu) g^{1/2} g^{\mu\nu} (\partial_\nu + \mathscr{A}_\nu). \tag{7}$$

It is worth noting that every second-order differential operator with a scalar leading symbol given by the metric tensor is of Laplace type and can be put in the form (6) by choosing the appropriate connection 1-form $\mathscr{A}$ and the endomorphism Q.

The operator F has a positive leading symbol and is elliptic. It is symmetric with respect to the natural L^2 inner product. Moreover, the operator F is essentially self-adjoint, i.e., there is a unique self-adjoint extension $\overline{F}$ of the operator F. However, we will not be very careful about distinguishing between the operator F and its closure $\overline{F}$, and will simply say that the operator F is elliptic and self-adjoint.

A. Heat kernel of a Laplace type operator

Hence, the operator $U_F(t)=\exp(-tF)$ for $t>0$ is well defined as a *bounded* operator on the Hilbert space $L^2(M,V)$ of square integrable sections of the vector bundle V. This operator forms a one-parameter semi-group. The kernel $U_F(t|x,x')$ of this operator satisfies the heat equation and is called the heat kernel. For $t>0$ the heat kernel is a smooth section of the external tensor product of the vector bundles V and V^* over the tensor product manifold $M\times M$.

In the following we are going to study the Green function and the heat kernel only locally (like in Ref. 7), i.e. in the neighborhood of the diagonal of $M\times M$. We will keep a point x' of the manifold fixed and consider a small geodesic ball, i.e. a small neighborhood of the point x': $B_{x'}=\{x\in M|r(x,x')<\varepsilon\}$, $r(x,x')$ being the geodesic distance between the points x and x'. We will take the radius of the ball sufficiently small, so that each point x of the ball of this neighborhood can be connected by a unique geodesic with the point x'. This can always be done if the size of the ball is smaller than the injectivity radius of the manifold at x', $\varepsilon<r_{\text{inj}}(x')$.

Let $\sigma(x,x')$ be the geodetic interval, also called world function, defined as one half the square of the length of the geodesic connecting the points x and x'

$$\sigma(x,x')=\tfrac{1}{2}r^2(x,x'). \tag{8}$$

The first derivatives of this function with respect to x and x' define the tangent vector fields to the geodesic at the end points x and x' and the determinant of the mixed second derivatives defines a so called Van Vleck-Morette determinant[2,8,9]

$$\Delta(x,x')=g^{-1/2}(x)\det\left(-\nabla_\mu\nabla'_{\nu'}\sigma(x,x')\right)g^{-1/2}(x'). \tag{9}$$

Let, finally, $\mathscr{P}(x,x')$ denote the parallel transport operator along the geodesic from the point x' to the point x. It is a section of the external tensor product of the vector bundle V and V^* over $M\times M$, or, in other words, it is an endomorphism from the fiber of V over x' to the fiber of V over x. Near the diagonal of $M\times M$ all these two-point functions are smooth single-valued functions of the coordinates of the points x and x'. To simplify the consideration, one can assume that these functions are *analytic* in the ball $B_{x'}$.

The heat kernel of a Laplace type operator is well described by factorizing out the semi-classical factor[2,6,7]

$$U_F(t|x,x')=(4\pi t)^{-d/2}\Delta^{1/2}(x,x')\exp\left(-\frac{1}{2t}\sigma(x,x')\right)\Omega_F(t|x,x'). \tag{10}$$

The function $\Omega_F(t|x,x')$ is called the *transfer function* of the operator F. Obviously,

$$U_{F+m^2}(t)=e^{-tm^2}U_F(t),\quad \Omega_{F+m^2}(t)=e^{-tm^2}\Omega_F(t). \tag{11}$$

It can be proved[7] that for a positive operator $F+m^2$ the transfer function $\Omega_{F+m^2}(t)$ can be presented in form of an inverse Mellin transform (we slightly change notation here in comparison to Ref. 7)

$$\Omega_{F+m^2}(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}dq\ t^q\Gamma(-q)b_{F+m^2}(q), \tag{12}$$

where c is a negative constant. The function $b_{F+m^2}(q)$ is an *entire* function of q, so that the integrand has simple poles at the positive integer points $q=k$ $(k=0,1,\dots)$. The function $b_{F+m^2}(q|x,x')$ satisfies a functional-differential equation[7]

$$\left(1+\frac{1}{q}D\right)b_{F+m^2}(q)=L\ b_{F+m^2}(q-1), \tag{13}$$

where

$$L=\Delta^{-1/2}F\Delta^{1/2}+m^2=-\Delta^{-1/2}\Box\Delta^{1/2}+Q+m^2, \tag{14}$$

$$D=\sigma^\mu\nabla_\mu, \quad \sigma_\mu=\nabla_\mu\sigma, \tag{15}$$

and the initial condition

$$b_{F+m^2}(0|x,x')=\mathscr{P}(x,x'). \tag{16}$$

This equation (together with the condition of analyticity and some asymptotic condition at $q\to\text{const}\pm i\infty$) enables one to compute $b_{F+m^2}(q)$ if one fixes its initial value at some arbitrary point $q=q_0$. (For more details, see Ref. 7.) The initial condition at the origin produces the values of $b_{F+m^2}(k)$ at positive integer points, the initial condition at the point $q=1/2$ would give the values of $b_{F+m^2}(q)$ at all half-integer positive points $b_{F+m^2}(k+1/2)$, etc. Differentiating eq. (13) with respect to q one obtains another recursion

$$\left(1+\frac{1}{q}D\right)b'_{F+m^2}(q)=L\ b'_{F+m^2}(q-1)+\frac{1}{q^2}Db_{F+m^2}(q), \tag{17}$$

where

$$b'_{F+m^2}(q)=\frac{\partial}{\partial q}b_{F+m^2}(q), \tag{18}$$

which enables one to compute the derivatives of the function $b_{F+m^2}(q)$ at positive integer points if one fixes its value $b'_{F+m^2}(0)$.

Moving the contour of integration to the right and taking into account that the residues of the gamma-function $\Gamma(-q)$ at the points $q=k$ equal $(-1)^k/k!$ one obtains[7]

$$\Omega_{F+m^2}(t)=\sum_{k=0}^{n-1}\frac{(-t)^k}{k!}a_k(F+m^2)+\frac{1}{2\pi i}\int_{c_n-i\infty}^{c_n+i\infty}dq\ t^q\Gamma(-q)b_{F+m^2}(q), \tag{19}$$

where

$$a_k(F+m^2)\equiv b_{F+m^2}(k)=\left(-\frac{\partial}{\partial t}\right)^k\Omega_{F+m^2}(t)|_{t=0}, \tag{20}$$

n is an arbitrary positive integer, $n\geqslant 1$, and c_n ranges in the interval $n-1<c_n<n$. The coefficients $a_k(F+m^2)$ are the famous Hadamard-Minakshisundaram-De Witt-Seeley (HMDS)-coefficients[1,10,2,11] to the operator $F+m^2$. They can be computed in form of covariant Taylor series from the recursion relations that are obtained from eq. (13) (by putting $q\to k$) with the initial condition $a_0(F+m^2)=\mathscr{P}$. The diagonal values of the HMDS-coefficients are known in general case up to a_4.[12,6,13,7,14] In some particular cases, e.g., in flat space, there are results for higher-order coefficients. For a review of the methods for calculation of HMDS-coefficients and further references see Refs. 15 and 16.

Taking into account eq. (11) we find from (20)

$$a_k(F+m^2)=\sum_{n=0}^{k}\binom{k}{n}m^{2(k-n)}a_n(F). \tag{21}$$

Moreover, one can obtain an asymptotic expansion of the function $b_{F+m^2}(q)$ as $m\to\infty$[7]

$$b_{F+m^2}(q)\sim\sum_{n\geqslant 0}\frac{\Gamma(q+1)}{n!\Gamma(q-n+1)}\,m^{2(q-n)}a_n(F). \tag{22}$$

From (19) there follows an asymptotic expansion of the transfer function as $t\to 0$ (called Schwinger-De Witt expansion in the physical literature)[2–4,6,7]

$$\Omega_{F+m^2}(t)\sim\sum_{k\geqslant 0}\frac{(-t)^k}{k!}a_k(F+m^2)=e^{-tm^2}\sum_{k\geqslant 0}\frac{(-t)^k}{k!}a_k(F). \tag{23}$$

Using this equation one obtains from (10) the well known formula for the asymptotic expansion of the functional trace of the heat kernel

$$\mathrm{Tr}_{L^2}\exp(-tF)\sim(4\pi)^{-d/2}t^{-d/2}\sum_{k\geqslant 0}\frac{(-t)^k}{k!}A_k(F), \tag{24}$$

where

$$A_k(F)=\mathrm{Tr}_{L^2}a_k(F)=\int_M d\mathrm{vol}\,(x)\ \mathrm{tr}_V a_k(F|x,x), \tag{25}$$

with tr_V being the fiber trace.

B. Green function of a Laplace type operator

The Laplace type operator F is a self-adjoint elliptic operator with a real spectrum bounded from below. Thus for sufficiently large m the operator $(F+m^2)$ is positive and its Green operator, or the resolvent operator, $G_{F+m^2}=(F+m^2)^{-1}$ is a bounded operator. It can be defined by

$$G_{F+m^2}=\int_0^\infty dt\ e^{-tm^2}\exp(-tF). \tag{26}$$

The Green function of the operator $(F+m^2)$ is obtained by the kernel form of this equation

$$G_{F+m^2}=(4\pi)^{-d/2}\Delta^{1/2}\int_0^\infty dt\ t^{-d/2}\exp\left(-tm^2-\frac{\sigma}{2t}\right)\Omega_F(t). \tag{27}$$

This integral converges at infinity for sufficiently large m. It also converges at $t=0$ outside the diagonal, i.e. for $\sigma\neq 0$. By using the ansatz (10) and (12) for the heat kernel one obtains a corresponding ansatz for the Green function

$$G_{F+m^2}=(4\pi)^{-d/2}\Delta^{1/2}\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}dq\,\left(\frac{\sigma}{2}\right)^{q+1-d/2}\Gamma(-q)\Gamma(-q-1+d/2)b_{F+m^2}(q), \tag{28}$$

where $c<-1/2$.

This ansatz is especially useful for studying the singularities of the Green function, or more general, for constructing the Green function as a power series in σ. The integrand in (28) is again a meromorphic function. However, contrary to (12), we have now a more complicated structure of the poles. There are always poles at the points $q=k$ and $q=k-1+d/2$, $(k=0,1,2,\dots)$. Here one has to distinguish between odd and even dimensions. In odd dimensions, the poles are at the points $q=k$ and $q=k+[d/2]-1/2$ and are *simple*, whereas in even dimension there are simple poles at $q=0,1,2,\dots,d/2-2$ and *double* poles at the points $q=k+d/2-1$.

Moving the contour of integration in (28) to the right one can obtain an expansion of the Green function in powers of σ (Hadamard series). Generally, we obtain

$$G_{F+m^2}=G^{\mathrm{sing}}_{F+m^2}+G^{\text{non-anal}}_{F+m^2}+G^{\mathrm{reg}}_{F+m^2}. \tag{29}$$

Here $G^{\rm sing}_{F+m^2}$ is the singular part which is polynomial in the inverse powers of $\sqrt{\sigma}$

$$G^{\rm sing}_{F+m^2}=(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{[(d+1)/2]-2}\frac{(-1)^k}{k!}\Gamma(d/2-k-1)\left(\frac{2}{\sigma}\right)^{d/2-k-1}a_k(F+m^2), \tag{30}$$

For the rest we get in *odd* dimensions

$$\begin{aligned}G^{\rm non\text{-}anal}_{F+m^2}+G^{\rm reg}_{F+m^2}=&(-1)^{(d-1)/2}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{n-(d+1)/2}\frac{\pi}{\Gamma(k+(d+1)/2)\Gamma(k+3/2)}\\&\times\left(\frac{\sigma}{2}\right)^{k+1/2}a_{k+(d-1)/2}(F+m^2)\\&+(-1)^{(d+1)/2}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{n-(d+1)/2}\frac{\pi}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^{k}b_{F+m^2}(k-1+d/2)\\&+(4\pi)^{-d/2}\Delta^{1/2}\frac{1}{2\pi i}\int_{c_n-i\infty}^{c_n+i\infty}dq\left(\frac{\sigma}{2}\right)^{q+1-d/2}\Gamma(-q)\Gamma(-q-1+d/2)b_{F+m^2}(q),\end{aligned} \tag{31}$$

where $n-1<c_n<n-1/2$ and $n>(d-1)/2$. Thus, by putting $n\to\infty$ we recover herefrom the Hadamard power series in σ for *odd* $d=1,3,5,\ldots$

$$\begin{aligned}G^{\rm non\text{-}anal}_{F+m^2}\sim&(-1)^{(d-1)/2}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{\infty}\frac{\pi}{\Gamma(k+(d+1)/2)\Gamma(k+3/2)}\left(\frac{\sigma}{2}\right)^{k+1/2}\\&\times a_{k+(d-1)/2}(F+m^2)\ ,\end{aligned} \tag{32}$$

$$G^{\rm reg}_{F+m^2}\sim(-1)^{(d+1)/2}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{\infty}\frac{\pi}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^{k}b_{F+m^2}(k-1+d/2). \tag{33}$$

In *even* dimensions, the point is more subtle due to the presence of double poles. Moving the contour in (28) to the right and calculating the contribution of the residues at the simple and double poles we obtain

$$\begin{aligned}&G^{\rm non\text{-}anal}_{F+m^2}+G^{\rm reg}_{F+m^2}\\&\quad=(-1)^{d/2-1}(4\pi)^{-d/2}\Delta^{1/2}\log\left(\frac{\mu^2\sigma}{2}\right)\sum_{k=0}^{n-1}\frac{1}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^{k}a_{k-1+d/2}(F+m^2)\\&\quad+(-1)^{d/2-1}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{n-1}\frac{1}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^{k}\Big\{b'_{F+m^2}(k-1+d/2)\\&\quad-[\log\mu^2+\Psi(k+1)+\Psi(k+d/2)]a_{k-1+d/2}(F+m^2)\Big\}\\&\quad+(4\pi)^{-d/2}\Delta^{1/2}\frac{1}{2\pi i}\int_{c_n-i\infty}^{c_n+i\infty}dq\left(\frac{\sigma}{2}\right)^{q+1-d/2}\Gamma(-q)\Gamma(-q-1+d/2)b_{F+m^2}(q),\end{aligned} \tag{34}$$

where μ is an arbitrary mass parameter introduced to preserve dimensions, $n-1<c_n<n$ and $\Psi(z)=(d/dz)\log\Gamma(z)$. If we let $n\to\infty$ we obtain the Hadamard expansion of the Green function for *even* $d=2,4,\ldots$

$$G^{\text{non-anal}}_{F+m^2}\sim(-1)^{d/2-1}(4\pi)^{-d/2}\Delta^{1/2}\log\left(\frac{\mu^2\sigma}{2}\right)\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^k a_{k-1+d/2}(F+m^2)\ , \tag{35}$$

$$G^{\text{reg}}_{F+m^2}\sim(-1)^{d/2-1}(4\pi)^{-d/2}\Delta^{1/2}\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^k \times\left\{b'_{F+m^2}(k-1+d/2)-[\log\mu^2+\Psi(k+1)+\Psi(k+d/2)]a_{k-1+(d/2)}(F+m^2)\right\}. \tag{36}$$

Note that the singular part (which is a polynomial in inverse powers of $\sqrt{\sigma}$) and the *non-analytical* parts (proportional to $\sqrt{\sigma}$ and $\log\sigma$) are expressed *uniquely only in terms of the local HMDS-coefficients* $a_k(F+m^2)$, whereas the regular analytical part contains the values of the function $b_{F+m^2}(q)$ at half-integer positive points and the derivatives of the function $b_{F+m^2}(q)$ at integer positive points, which are *not* expressible in terms of the local information. These objects are global and cannot be expressed further in terms of the local HMDS-coefficients. However, they can be computed from eqs. (13) and (17) almost by the same algorithm as the HMDS-coefficients in terms of the value of the function $b(q)$ at some fixed q_0 (see Refs. 7).

The regular part of the Green function has a well defined diagonal value and the functional trace. It reads in odd dimensions ($d=1,3,5,\dots$):

$$\begin{aligned}\mathrm{Tr}_{L^2}\,G^{\text{reg}}_{F+m^2}&=(-1)^{(d+1)/2}(4\pi)^{-d/2}\frac{\pi}{\Gamma(d/2)}B_{F+m^2}(d/2-1)\\&\overset{m\to\infty}{\sim}(-1)^{(d+1)/2}(4\pi)^{-d/2}\pi\sum_{k\geqslant 0}\frac{m^{d-2-2k}}{k!\Gamma(d/2-k)}A_k(F)\end{aligned} \tag{37}$$

and in even dimensions ($d=2,4,6,\dots$)

$$\begin{aligned}\mathrm{Tr}_{L^2}\,G^{\text{reg}}_{F+m^2}&=(-1)^{d/2-1}\frac{(4\pi)^{-d/2}}{\Gamma(d/2)}\{B'_{F+m^2}(d/2-1)-[\log\,\mu^2+\Psi(d/2)-\mathbb{C}]A_{d/2-1}(F+m^2)\}\\&\overset{m\to\infty}{\sim}(-1)^{d/2-1}(4\pi)^{-d/2}\Bigg\{\sum_{k=0}^{d/2-1}\frac{m^{d-2-2k}}{k!\Gamma(d/2-k)}[\mathbb{C}-\Psi(d/2-k)\\&\quad+\log\,(m^2/\mu^2)]A_k(F)+\sum_{k\geqslant d/2}\frac{(-1)^{k-d/2}}{k!}m^{d-2-2k}\Gamma(k+1-d/2)A_k(F)\Bigg\},\end{aligned} \tag{38}$$

where

$$B_{F+m^2}(q)=\mathrm{Tr}_{L^2}b_{F+m^2}(q)=\int_M d\mathrm{vol}\,(x)\mathrm{tr}_V\,b_{F+m^2}(x,x), \tag{39}$$

and $\mathbb{C}=-\Psi(1)=0.577\dots$ is the Euler's constant.

Thus, we see that

(i) all the singularities of the Green function and the non-analytical parts thereof (proportional to $\sqrt{\sigma}$ in odd dimensions and to $\log\sigma$ in even dimensions) are determined by the HMDS-coefficients $a_k(F)$;

(ii) there are no power singularities, i.e., $G^{\text{sing}}_{F+m^2}=0$, in lower dimensions $d=1,2$;

(iii) there is no logarithmic singularity (more generally, no logarithmic part at all) in odd dimensions;

(iv) the regular part depends on the values of the function $b_{F+m^2}(q)$ at half-integer points and its derivative at integer points and is a global object that cannot be reduced to purely local information like the HMDS-coefficients.

The logarithmic part of the Green function is very important. On the one hand it determines, as usual, the renormalization properties of the regular part of the Green function, i.e. the derivative $\mu(\partial/\partial\mu)G^{\rm reg}_{F+m^2}$. In particular,

$$\mu\frac{\partial}{\partial\mu}{\rm Tr}_{L^2}G^{\rm reg}_{F+m^2}=\begin{cases}0 & \text{for odd } d\\ \dfrac{(4\pi)^{-d/2}}{\Gamma(d/2)}A_{d/2-1}(F+m^2) & \text{for even } d.\end{cases}\tag{40}$$

On the other hand, it is of crucial importance in studying the Huygens principle. Namely, the absence of the logarithmic part of the Green function is a necessary and sufficient condition for the validity of the Huygens principle for hyperbolic operators.[17–19] The HMDS-coefficients and, therefore, the logarithmic part of the Green function are defined for the hyperbolic operators just by analytic continuation from the elliptic case. Thus, the condition of the validity of Huygens principle reads

$$\sum_{k=0}^{\infty}\frac{\Gamma(d/2)}{k!\Gamma(k+d/2)}\left(\frac{\sigma}{2}\right)^k a_{k-1+d/2}(F+m^2)=0,\tag{41}$$

or, by using (21),

$$\sum_{k=0}^{\infty}\sum_{n=0}^{k-1+d/2}\frac{\Gamma(d/2)}{k!n!\Gamma(k-n+d/2)}\left(\frac{\sigma}{2}\right)^k m^{2(k-n)}a_n(F)=0.\tag{42}$$

By expanding this equation in covariant Taylor series using the methods of Ref. 7 one can obtain an infinite set of local conditions for validity of the Huygens principle.

$$\sum_{k=0}^{[n/2]}\frac{n!\Gamma(d/2)}{4^k k!(n-2k)!\Gamma(k+d/2)}(\vee^k g)\vee\langle n-2k|a_{k-1+d/2}(F+m^2)\rangle=0,\tag{43}$$

where $\vee$ is the exterior symmetric tensor product, g is the metric tensor on the tangent bundle, and $\langle n|a_k\rangle$ denotes the diagonal value of the symmetrized covariant derivative of nth order.[7] More explicitly,

$$[a_{d/2-1}(F+m^2)]=0,\tag{44}$$

$$[\nabla_\mu a_{d/2-1}(F+m^2)]=0,\tag{45}$$

$$[\nabla_{(\mu}\nabla_{\nu)}a_{d/2-1}(F+m^2)]+\frac{1}{2d}g_{\mu\nu}[a_{d/2}(F+m^2)]=0,\tag{46}$$

$$\dots,$$

where the square brackets denote the diagonal value of two-point quantities: $[f(x,x')]=f(x,x)$.

III. HIGHER ORDER OPERATORS

Let us now describe in detail the class of higher order operators we are going to study. We consider again a Riemannian manifold (M,g) and a vector bundle $V(M)$ with a connection ∇. Let $\mathscr{A}_{(i)}$, $(i=1,\dots,N)$ be a set of different smooth sections of the vector bundle $T^*M\otimes{\rm End}(V)$. They define a set of *different* connections

$$\nabla_{(i)}=\nabla+\mathscr{A}_{(i)}\tag{47}$$

on the vector bundle $V(M)$, and therefore, a set of different generalized Laplacians

$$\begin{aligned}\Box_i&={\rm tr}_g\nabla_{(i)}\nabla_{(i)}=g^{\mu\nu}(\nabla_\mu+\mathscr{A}_{(i)\,\mu})(\nabla_\nu+\mathscr{A}_{(i)\,\nu})\\&=\Box+g^{\mu\nu}(\mathscr{A}_{(i)\mu}\nabla_\nu+\nabla_\nu\mathscr{A}_{(i)\mu})+g^{\mu\nu}\mathscr{A}_{(i)\mu}\mathscr{A}_{(i)\nu}.\end{aligned}\tag{48}$$

Let, further, Q_i be a set of different smooth sections of the vector bundle $\mathrm{End}(V)$. Then, we define a set of *different* Laplace type operators

$$F_i = -\Box_i + Q_i \tag{49}$$

and a higher order operator of a special form

$$H = (F_N + m^2)\cdots(F_2 + m^2)(F_1 + m^2). \tag{50}$$

A. Algebraical reduction to Laplace type operators

We are going to study the Green function $G_{H-\lambda}(x,x')$ of the higher order operator $H-\lambda$, where λ is an arbitrary sufficiently large negative constant. To do this, we show, first, that the Green operator $G_{H-\lambda}=(H-\lambda)^{-1}$ can be reduced to the Green operator of an auxiliary Laplace type operator. This will mean that the Green function of the higher order differential operator will be reduced to the Green function of a *second-order* differential operator of Laplace type, which is described in previous section.

Theorem 1: *Let F_i, $(i=1,\ldots,N)$, be some positive elliptic operators and H be an operator defined by*

$$H = (F_N + m^2)\cdots(F_2 + m^2)(F_1 + m^2), \tag{51}$$

where m is a sufficiently large constant. Let I_{N-1} be the unit $(N-1)\times(N-1)$ matrix and O_{N-1} be the square $(N-1)\times(N-1)$ null matrix. Let P and Π be square $N\times N$ matrices defined by the following block form

$$P = \begin{pmatrix} 0 & I_{N-1} \\ \lambda & 0 \end{pmatrix}, \quad \Pi = \frac{\partial}{\partial\lambda} P = \begin{pmatrix} 0 & O_{N-1} \\ 1 & 0 \end{pmatrix}, \tag{52}$$

where λ is a constant, and let $\widetilde{F} = F_1 \oplus \cdots \oplus F_N$ be a $N\times N$ diagonal matrix formed by the operators F_i

$$\widetilde{F} = \begin{pmatrix} F_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & F_N \end{pmatrix}. \tag{53}$$

Let $\mathscr{F}$ be an operator defined by

$$\mathscr{F} = \widetilde{F} - P. \tag{54}$$

Let $G_{H-\lambda}$ and $G_{\mathscr{F}+m^2}$ be the Green operators of the operators $(H-\lambda)$ and $(\mathscr{F}+m^2)$ satisfying the equations

$$(H-\lambda)G_{H-\lambda} = 1 \tag{55}$$

and

$$(\mathscr{F}+m^2)G_{\mathscr{F}+m^2} = 1. \tag{56}$$

Then there holds

$$G_{H-\lambda} = \mathrm{tr}\ \Pi\ G_{\mathscr{F}+m^2}, \tag{57}$$

$$\frac{\partial}{\partial m^2} G_{H-\lambda} = -\frac{\partial}{\partial\lambda}\mathrm{tr}\ G_{\mathscr{F}+m^2}, \tag{58}$$

Proof: Let us introduce some auxiliary operators

$$Z_1 = G_{H-\lambda}$$

$$Z_2 = (F_1 + m^2) Z_1$$

$$\vdots$$

$$Z_N = (F_{N-1} + m^2) Z_{N-1}. \tag{59}$$

Then eq. (55) can be rewritten as

$$(F_N + m^2) Z_N - \lambda G_{H-\lambda} = 1. \tag{60}$$

Let us collect the operators Z_i in a column-vector

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix}. \tag{61}$$

The Green operator $G_{H-\lambda}$ is the first component of the vector Z and can be obtained by multiplying with the row-vector $\Pi_1^\dagger = (1,0,\ldots,0)$

$$G_{H-\lambda} = \Pi_1^\dagger Z. \tag{62}$$

Further, let us define another column-vector

$$\Pi_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{63}$$

so that

$$\Pi = \Pi_N \otimes \Pi_1^\dagger. \tag{64}$$

It is not difficlut to show that eqs. (59) can be rewritten in a matrix form

$$(\mathscr{F} + m^2) Z = \Pi_N. \tag{65}$$

Using the Green operator $G_{\mathscr{F}+m^2}$ of the operator $(\mathscr{F} + m^2)$ we obtain

$$Z = G_{\mathscr{F}+m^2} \Pi_N. \tag{66}$$

Finally, from (62) by taking into account (64) we get the Green operator $G_{H-\lambda}$

$$G_{H-\lambda} = \Pi_1^\dagger \; G_{\mathscr{F}+m^2} \Pi_N = \mathrm{tr}\; \Pi \; G_{\mathscr{F}+m^2}, \tag{67}$$

which proves eq. (57). To prove eq. (58) we note that

$$\frac{\partial}{\partial \lambda} \mathrm{tr}\; G_{\mathscr{F}+m^2} = -\,\mathrm{tr}\; G_{\mathscr{F}+m^2} \frac{\partial \mathscr{F}}{\partial \lambda} G_{\mathscr{F}+m^2}$$

$$= -\,\mathrm{tr}\; \Pi \; G^2_{\mathscr{F}+m^2}. \tag{68}$$

Further, from the definition of the Green operator we have, obviously,

$$\frac{\partial}{\partial m^2} \mathrm{tr}\; \Pi \; G_{\mathscr{F}+m^2} = -\,\mathrm{tr}\; \Pi \; G^2_{\mathscr{F}+m^2}. \tag{69}$$

Comparing (68) and (69) and using eq. (67) we convince ourselves that eq. (58) is correct too.

Let us make some remarks. First, ''tr'' denotes here the usual matrix trace and has nothing to do with the trace over the bundle indices which are left intact. Second, the operator $\widetilde{F}$ has, obviously, the form

$$\widetilde{F}=-\widetilde{\Box}+\widetilde{Q}, \tag{70}$$

where $\widetilde{\Box}=\Box_1\oplus\cdots\oplus\Box_N$ and $\widetilde{Q}=Q_1\oplus\cdots\oplus Q_N$ have the same diagonal structure. Obviously

$$\widetilde{\Box}=\mathrm{tr}_g\widetilde{\nabla}\widetilde{\nabla} \tag{71}$$

where $\widetilde{\nabla}=\nabla_1\oplus\cdots\oplus\nabla_N$ is a new connection which also has the same diagonal structure. Therefore, the operator $\mathscr{F}$ reads

$$\mathscr{F}=-\widetilde{\Box}+\widetilde{Q}-P. \tag{72}$$

Remember that the matrix P is constant. Thus, both the operator $\widetilde{F}$ and $\mathscr{F}$ are *second-order* elliptic Laplace type differential operators. However, the operator $\mathscr{F}$ is *not self-adjoint* because the matrix P is not symmetric. In spite of this, the operator $(\mathscr{F}+m^2)$ for sufficiently large m is non-degenerate, i.e., there exists a well defined Green operator $G_{\mathscr{F}+m^2}$, which justifies made assumptions.

1. Properties of the matrix P

The matrix P plays the role of the matrix term and is of great importance in further considerations. That is why, we state below some important properties of it.

First of all, we note that the matrix Π is nilpotent

$$\Pi^2=0 \tag{73}$$

and, of course, degenerate. The matrix P is not degenerate. The matrix P as well as its powers belong to the class of two-diagonal matrices. We compute first the powers of this matrix.

$$P=\begin{pmatrix}0 & I_{N-1}\\ \lambda & 0\end{pmatrix},\quad P^2=\begin{pmatrix}0 & I_{N-2}\\ \lambda I_2 & 0\end{pmatrix},\cdots,\quad P^{N-1}=\begin{pmatrix}0 & 1\\ \lambda I_{N-1} & 0\end{pmatrix}, \tag{74}$$

$$P^N=\lambda I_N. \tag{75}$$

It is also not difficult to show that

$$\det(z-P)=z^N-\lambda. \tag{76}$$

Therefore, the eigenvalues p_j of the matrix P are just the Nth roots of λ

$$p_j=(-\lambda)^{1/N}e^{i\,\pi(2j-1)/N},\quad j=1,\ldots,N, \tag{77}$$

where $|\arg(-\lambda)|<\pi$. The roots p_j lie on the circle $|z|=|\lambda|^{1/N}$.

The higher powers of the matrix P are expressed in terms of the lower order powers (74) and (75)

$$P^{iN+j}=\lambda^i P^j, \tag{78}$$

where $j=0,1,2,\ldots,N-1$ and $i=0,1,2,\ldots$. Using these formulas we find that for any function $f(z)$ that is analytic at the points $p_j(\lambda)$ there holds

$$f(P)=\frac{1}{2\pi i}\int_{C_P} dz\ f(z)\ (z-P)^{-1}$$

$$=\sum_{j=0}^{N-1}\beta_{N-1-j}(\lambda;f)P^j, \tag{79}$$

where

$$\beta_n(\lambda;f)=\frac{1}{2\pi i}\int_{C_P} dz\ \frac{z^n}{z^N-\lambda}f(z) \tag{80}$$

and the contour C_P goes counter-clockwise around the points p_j in such a way that it does not contain the singularities of the function $f(z)$ — inside the contour C_P there must be no other singularities except for the points p_j. Evidently, the coefficients $\beta_n(\lambda;f)$ possess the following ''periodicity'' property

$$\beta_{n+kN}(\lambda;f)=\lambda^k\beta_n(\lambda;f),\quad k=0,1,2,\ldots, \tag{81}$$

so that there are essentially only N independent coefficients $\beta_0,\beta_1,\ldots,\beta_{N-1}$.

Equations (79) and (85) determine any function of the matrix P. In particular, using the traces of the powers of the matrix P

$$\text{tr}\ P^n=\begin{cases} N\lambda^i & \text{for}\quad n=iN,\ i=0,1,2,\ldots, \\ 0 & \text{otherwise}\end{cases} \tag{82}$$

we can compute the trace of any function of P

$$\text{tr}\ f(P)=N\beta_{N-1}(\lambda;f). \tag{83}$$

More generally,

$$\text{tr}\ P^n\ f(P)=N\beta_{N-1+n}(\lambda;f). \tag{84}$$

From here we obtain the coefficients $\beta_n(\lambda;f)$

$$\beta_n(\lambda;f)=\frac{1}{N\lambda}\sum_{j=1}^{N} p_j^{n+1}f(p_j)$$

$$=-\frac{1}{N}(-\lambda)^{(n-N+1)/N}\sum_{j=1}^{N} e^{i\,\pi(2j-1)(n+1)/N}f(p_j). \tag{85}$$

From (82), it follows, in particular,

$$\sum_{j=1}^{N} p_j^n=0\quad \text{for}\quad n\neq 0,\pm N,\pm 2N,\ldots. \tag{86}$$

Moreover, using the definition (52) we calculate the traces of the product of the powers of the matrix P with the matrix Π

$$\text{tr}\ \Pi\ P^n=\frac{1}{(n+1)}\frac{\partial}{\partial\lambda}\text{tr}\ P^{n+1}=\begin{cases}\lambda^i & \text{for}\quad n=iN+N-1,\ i=0,1,2,\ldots, \\ 0 & \text{otherwise}.\end{cases} \tag{87}$$

By using the equation

$$\frac{\partial}{\partial\lambda}\text{tr}\ f(P)=\text{tr}\ \frac{\partial P}{\partial\lambda}\,f'(P) \tag{88}$$

where $f'(z)=(\partial/\partial z)f(z)$, and eq. (52) or using directly eq. (87) we calculate the trace of the product of the function $f(P)$ with the matrix Π

$$\text{tr}\ \Pi\ f(P)=\beta_0(\lambda;f)=-\frac{1}{N}(-\lambda)^{-(N-1)/N}\sum_{j=1}^{N} e^{i\,\pi(2j-1)/N}f(p_j). \tag{89}$$

More generally, one can obtain also

$$\text{tr}\ \Pi(z-P)^{-1}=-\frac{\partial}{\partial\lambda}\log\,\det\,(z-P)=\frac{1}{z^N-\lambda} \tag{90}$$

as well as

$$\text{tr}\ \Pi\ e^{tP}=\frac{1}{2\pi i}\int_{C_P} dz\,\frac{e^{tz}}{z^N-\lambda}=-\frac{1}{N}(-\lambda)^{-(N-1)/N}\sum_{j=1}^{N} e^{i\,\pi(2j-1)/N}e^{tp_j}. \tag{91}$$

In the next section we will also need the following lemma, which is actually a generalization of eq. (82).

Lemma 1: Let P and Π be the $N\times N$ matrices defined by (52) and $B_1,\dots,B_{N-2}$ be arbitrary diagonal matrices. Then for any $k\leqslant N-2$ and arbitrary non-negative integers $n_1,\dots,n_k$ satisfying the condition $n_1+\cdots+n_k\leqslant N-2$ there holds

$$\text{tr}\ \Pi\ B_1P^{n_1}B_2P^{n_2}\cdots B_kP^{n_k}=0. \tag{92}$$

2. Properties of the symbol $\sigma(\mathscr{F})$

Let us consider now the symbol $\sigma(\mathscr{F})$ of the operator $\mathscr{F}$. It has the form

$$\sigma(\mathscr{F})=\sigma(\widetilde{F})-P, \tag{93}$$

where $\sigma(\widetilde{F})$ is a diagonal matrix.

That is why we consider the following matrix:

$$M\equiv\widetilde{\mu}-P, \tag{94}$$

where $\widetilde{\mu}=\mu_1\oplus\cdots\oplus\mu_N$ is a diagonal matrix with some constants μ_j on the diagonal. Note that the matrices $\widetilde{\mu}$ and P do not commute with each other $[\widetilde{\mu},P]\neq 0$.

We are interested in the eigenvalues q_j of the matrix M. It is not difficult to calculate

$$\text{tr}\,M^n=\text{tr}\,(\widetilde{\mu})^n=\sum_{j=1}^{N}\mu_j^n \qquad (n=1,2,\dots,N-1), \tag{95}$$

$$\text{tr}\,M^N=\sum_{j=1}^{N}\mu_j^N+(-1)^N N\lambda, \tag{96}$$

$$\det\,(z-M)=\prod_{j=1}^{N}(z-\mu_j)-(-1)^N\lambda. \tag{97}$$

Thus the eigenvalues q_j of the matrix M are determined either from one algebraical equation of Nth order

$$(\mu_1-z)\cdots(\mu_N-z)-\lambda=0 \tag{98}$$

or from the system of N equations

$$q_1+\cdots+q_N=\sum_{j=1}^{N}\mu_j, \tag{99}$$

$$\cdots$$

$$q_1^{N-1}+\cdots+q_N^{N-1}=\sum_{j=1}^{N}\mu_j^{N-1}, \tag{100}$$

$$q_1^{N}+\cdots+q_N^{N}=\sum_{j=1}^{N}\mu_j^{N}+(-1)^N N\lambda. \tag{101}$$

This enables one to calculate the traces

$$\mathrm{tr}\ f(M)=\sum_{j=1}^{N} f(q_j). \tag{102}$$

One can also get analogous to (80) integral representation

$$\mathrm{tr}\ f(M)=\frac{1}{2\pi i}\int_{C_M} dz\ f(z)\frac{A(z,\mu,\lambda)}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda}, \tag{103}$$

where C_M is a counter-clockwise contour around the spectrum of the matrix M and

$$A(z,\mu,\lambda)=\sum_{j=1}^{N}\prod_{k=1;\ k\neq j}^{N}(z-q_k) \tag{104}$$

is a polynomial of order $(N-1)$.

Further, by using the equation

$$\frac{\partial}{\partial\lambda}\mathrm{tr}\ f(M)=\mathrm{tr}\ \frac{\partial M}{\partial\lambda}f'(M), \tag{105}$$

where $f'(z)=(\partial/\partial z)f(z)$, and the obvious equation

$$\frac{\partial}{\partial\lambda}M=-\Pi, \tag{106}$$

which follows from the definition of the matrix M and eq. (52), we calculate the trace of the product of the function $f'(M)$ with the matrix Π

$$\mathrm{tr}\ \Pi\ f'(M)=-\frac{\partial}{\partial\lambda}\mathrm{tr}\ f(M). \tag{107}$$

Using eqs. (95) and (96) we get, in particular

$$\mathrm{tr}\ \Pi\ M^n=-\frac{1}{n+1}\frac{\partial}{\partial\lambda}\mathrm{tr}\ M^{n+1}=0 \quad \text{for} \quad n=0,1,\ldots,N-2, \tag{108}$$

$$\mathrm{tr}\ \Pi\ M^{N-1}=-\frac{1}{N}\frac{\partial}{\partial\lambda}\mathrm{tr}\ M^N=(-1)^{N-1}. \tag{109}$$

Note that (108) is a particular case of eq. (92). Further, from (99)–(101) one obtains

$$\sum_{j=1}^{N}\frac{\partial q_j}{\partial\lambda}q_j^n=0 \quad \text{for}\ n=0,1,2,\ldots,N-2, \tag{110}$$

$$\sum_{j=1}^{N} \frac{\partial q_j}{\partial \lambda} q_j^{N-1} = (-1)^N. \tag{111}$$

Further, we also get

$$\mathrm{tr}\ \Pi\ (z-M)^{-1} = \frac{\partial}{\partial \lambda} \log \det (z-M) = \frac{(-1)^{N-1}}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda}, \tag{112}$$

$$\mathrm{tr}\ \Pi\ M^{-k} = \frac{(-1)^N}{(k-1)!}\left(\frac{\partial}{\partial z}\right)^{k-1} \frac{1}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda}\bigg|_{z=0}, \tag{113}$$

$$\mathrm{tr}\ \Pi\ M^{k} = -\frac{1}{\Gamma(k-N+2)}\left(\frac{\partial}{\partial y}\right)^{k-N+1} \frac{1}{(y\mu_1-1)\cdots(y\mu_N-1)-y^N\lambda}\bigg|_{y=0}. \tag{114}$$

Note that for $k \leqslant N-2$ the last equation gives zero in accordance with (109).

This leads to a very simple and convenient representation

$$\mathrm{tr}\ \Pi\ f(M) = (-1)^{N-1} \frac{1}{2\pi i} \int_{C_M} dz\ \frac{f(z)}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda}. \tag{115}$$

We will also need, in particular, the following formula

$$\begin{aligned} \mathrm{tr}\ \Pi\ e^{-tM} &= (-1)^{N-1} \frac{1}{2\pi i} \int_{C_M} dz\ \frac{e^{-tz}}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda} \\ &= \frac{1}{t}\frac{\partial}{\partial \lambda} \mathrm{tr}\ e^{-tM} = -\sum_{j=1}^{N} \frac{\partial q_j}{\partial \lambda} e^{-tq_j}. \end{aligned} \tag{116}$$

From (108) it follows that this function is of order $O(t^{N-1})$ as $t \to 0$. Using eq. (98) we obtain

$$\frac{\partial q_j}{\partial \lambda} = -\left\{ \sum_{k=1}^{N} \prod_{i=1;\ i\neq k}^{N} (\mu_i - q_j) \right\}^{-1}. \tag{117}$$

In particular, for case $\lambda = 0$ we have

$$q_j|_{\lambda=0} = \mu_j, \tag{118}$$

hence,

$$\frac{\partial q_j}{\partial \lambda}\bigg|_{\lambda=0} = -\prod_{k=1;\ k\neq j}^{N} \frac{1}{(\mu_k - \mu_j)} \tag{119}$$

and

$$\mathrm{tr}\ \Pi\ e^{-tM}|_{\lambda=0} = \sum_{j=1}^{N} e^{-t\mu_j} \prod_{k=1;\ k\neq j}^{N} \frac{1}{(\mu_k - \mu_j)}. \tag{120}$$

These formulas will be useful when studying the heat kernel for the operator $\mathscr{F}$.

B. Green function of the operator $H-\lambda$

This shows that the problem is reduced to a Laplace type operator $\mathscr{F}$ with an additional matrix structure. Using the usual heat kernel representation for the Green function $\mathscr{G}$ of the Laplace type operator $\mathscr{F}$ we have from eq. (57)

Corollary 1:

$$G_{H-\lambda}=\int_0^\infty dt\ e^{-tm^2}\mathrm{tr}\ \Pi\ \exp(-t\mathscr{F}), \tag{121}$$

or in the kernel form

$$G_{H-\lambda}=(4\pi)^{-d/2}\Delta^{1/2}\int_0^\infty dt\ t^{-d/2}\exp\left(-tm^2-\frac{\sigma}{2t}\right)\omega_{\mathscr{F}}(t), \tag{122}$$

where

$$\omega_{\mathscr{F}}(t)=\mathrm{tr}\ \Pi\ \Omega_{\mathscr{F}}(t), \tag{123}$$

$\Omega_{\mathscr{F}}(t)=\Omega_{\mathscr{F}}(t|x,x')$ *is the transfer function of the operator* $\mathscr{F}$ *(54).*

Thus we see that there is analogous heat kernel representation for the Green function of higher-order operator $H-\lambda$ with a new ''transfer function'' $\omega_{\mathscr{F}}(t)$.

1. Power of a Laplace type operator

The simplest case is, of course, the case of equal potential terms Q_j and the connections $\nabla^{(i)}$, i.e.,

$$Q_i=Q,\quad \nabla^{(i)}=\nabla. \tag{124}$$

In this case all operators F_j are equal to each other

$$F_j=F=-\Box+Q, \tag{125}$$

and the operator H is the Nth power of the Laplace type operator $F+m^2$

$$H=(F+m^2)^N. \tag{126}$$

Further, in this simple case we have

$$\mathscr{F}=F-P \tag{127}$$

and the operator F commutes with the matrix P

$$[F,P]=0. \tag{128}$$

Therefore,

$$\exp(-t\mathscr{F})=\exp(tP)\exp(-tF). \tag{129}$$

and, using eqs. (57) and (26) we get

$$\omega_{\mathscr{F}}(t)=f(t)\Omega_F(t),\quad f(t)=\mathrm{tr}\ \Pi\exp(tP), \tag{130}$$

which is determined by (91) and (85)

$$f(t)=\frac{1}{2\pi i}\int_{C_P}dz\frac{e^{tz}}{z^N-\lambda}=-\frac{1}{N}(-\lambda)^{-(N-1)/N}\sum_{j=1}^{N}e^{i\ \pi(2j-1)/N}e^{tp_j(\lambda)}. \tag{131}$$

On the other hand, expanding the integrand in powers of λ and using the formula (163) we obtain $\omega_{\mathscr{F}}(t)$ in form of a power series

$$f(t)=t^{N-1}\sum_{k\geq 0}\frac{\lambda^k t^{Nk}}{(Nk+N-1)!}. \tag{132}$$

Thus, the function $\omega_{\mathscr{F}}(t)$ is of order $O(t^{N-1})$ as $t\to 0$.

In case $\lambda=0$ we obtain herefrom the usual heat kernel representation for the inverse power of the operator $(F+m^2)$

$$G_H=(F+m^2)^{-N}=\frac{1}{(N-1)!}\int_0^\infty dt\ t^{N-1}e^{-tm^2}\exp(-tF). \tag{133}$$

2. Constant perturbations of the potential terms

Let us consider now the case when the Laplace type operators F_j differ from each other only by a constant part of the potential terms, i.e., the connection $\nabla^{(j)}$ is the same, but the endomorphisms Q_j are different

$$Q_j=Q+\mu_j, \tag{134}$$

so that

$$F_j=F+\mu_j, \tag{135}$$

where

$$F=-\Box+Q. \tag{136}$$

The operator H has the form

$$H=(F+m^2+\mu_N)\cdots(F+m^2+\mu_2)(F+m^2+\mu_1). \tag{137}$$

The operators F_i, even if different, commute. The corresponding operator $\mathscr{F}$ has the form

$$\mathscr{F}=F+M, \tag{138}$$

where $M=\widetilde{\mu}-P$ and $\widetilde{\mu}=\mu_1\oplus\cdots\oplus\mu_N$ is a diagonal matrix with the constants μ_j on the diagonal. The matrix M commutes with the operator F

$$[F,M]=0. \tag{139}$$

Therefore,

$$\omega_{\mathscr{F}}(t)=h(t)\Omega_F(t),\quad h(t)=\mathrm{tr}\ \Pi\exp(-tM). \tag{140}$$

Using the eigenvalues of the matrix M, which are determined by eq. (98), we obtained this function in the form

$$h(t)=(-1)^{N-1}\frac{1}{2\pi i}\int_{C_M}dz\ \frac{e^{-tz}}{(z-\mu_1)\cdots(z-\mu_N)-(-1)^N\lambda}=-\sum_{j=1}^N\frac{\partial q_j}{\partial\lambda}e^{-tq_j}. \tag{141}$$

As it has been shown in section III A, this function is of order $O(t^{N-1})$ as $t\to0$. For $\lambda=0$ we obtained

$$h(t)|_{\lambda=0}=\sum_{j=1}^N\left[\prod_{k=1;\ k\neq j}^N\frac{1}{(\mu_k-\mu_j)}\right]e^{-t\mu_j}. \tag{142}$$

3. General case

As we have seen, in particular cases studied before the function $\omega_{\mathscr{F}}(t)$ defined by (123) is of order $O(t^{N-1})$ as $t\to0$. It is almost clear that this is also valid in general, since the main contribution to the asymptotics give the constant background fields. This can be formulated in the form of a lemma.

Lemma 2: Let $\mathscr{F}$ be the Laplace type operator defined by (54) and Π be the matrix defined by (52). Then for its HMDS-coefficients $a_k(\mathscr{F})$ there holds

$$\mathrm{tr}\ \Pi\ a_k(\mathscr{F})=0 \quad \text{for}\ \ k=0,1,\dots,N-2. \tag{143}$$

In other words, for the asymptotic expansion as $t\to 0$ of its transfer function $\omega_{\mathscr{F}}(t)$ there holds

$$\omega_{\mathscr{F}}(t)\big|_{t\to 0}\sim \frac{(-t)^{N-1}}{(N-1)!}\sum_{k\geqslant 0}\frac{(-t)^k}{k!}c_k(\mathscr{F}), \tag{144}$$

where

$$c_k(\mathscr{F})=\frac{(N-1)!k!}{(N+k-1)!}\mathrm{tr}\ \Pi\ a_{k+N-1}(\mathscr{F}). \tag{145}$$

Proof: The coefficients a_k can be calculated in form of a covariant Taylor series near the diagonal (for details and notation see Ref. 7)

$$a_k=\sum_{n\geqslant 0}|n\rangle\langle n|a_k\rangle, \tag{146}$$

where $|n\rangle$ is the covariant Taylor basis and $\langle n|a_k\rangle$ are the diagonal values of the symmetrized covariant derivatives. $\langle n|a_k\rangle$ have the following general form[7]

$$\langle n|a_k\rangle=\sum_{n_1,\dots,n_{k-1}\geqslant 0}c(n_1,\dots,n_{k-1})\langle n|M|n_{k-1}\rangle\langle n_{k-1}|M|n_{k-2}\rangle\cdots\langle n_1|M|0\rangle, \tag{147}$$

where $c(n_1,\dots,n_{k-1})$ are some constants and $\langle m|M|n\rangle$ are some End (V)-valued tensors that vanish for $n>m+2$ and for $n=m+1$. Important for us is that, since the matrix P is constant, it appears only in $\langle n|M|n\rangle$

$$\langle n|M|n\rangle=P+B, \tag{148}$$

where B is a diagonal matrix. *All* others tensors $\langle n|M|m\rangle$ do not depend on the matrix P and are diagonal—they are actually polynomials in the curvatures and their covariant derivatives.[7] So, all the terms in the sum (147) have the form

$$B_1P^{n_1}\dots B_jP^{n_j},\quad j=1,2,\dots,k, \tag{149}$$

where B_i are some diagonal matrices and the total power of the matrix P is restricted by $n_1+\cdots+n_j\leqslant k$. Therefore, from (92) we find that $\mathrm{tr}\,\Pi\langle n|a_k\rangle=\mathrm{tr}\ \Pi\ a_k=0$ for $k\leqslant N-2$, which proves the lemma.

Actually, using (148) and (87) one can prove more, namely,

$$c_0(\mathscr{F})=\mathrm{tr}\ \Pi\ a_{N-1}(\mathscr{F})=\mathrm{tr}\,\Pi\ P^{N-1}=1. \tag{150}$$

4. Hadamard expansion of the Green function of the operator $H-\lambda$

The behavior of the Green function of Laplace type operators near the diagonal is studied in the previous section. So, by using the equations of section II we obtain from (57) the Hadamard series for the Green function of the higher order operator $H-\lambda$

$$G_{H-\lambda}=G^{\mathrm{sing}}_{H-\lambda}+G^{\text{non-anal}}_{H-\lambda}+G^{\mathrm{reg}}_{H-\lambda}, \tag{151}$$

where the singular part reads

$$G^{\mathrm{sing}}_{H-\lambda}=(4\pi)^{-d/2}\frac{\Delta^{1/2}}{(N-1)!}\sum_{k=0}^{[(d+1)/2]-N-1}(-1)^{N+k}\frac{\Gamma\left(\frac{d}{2}-k-N\right)}{k!}\left(\frac{2}{\sigma}\right)^{d/2-k-N}c_k(\mathscr{F}+m^2). \tag{152}$$

Further, we get in odd dimensions:

$$G^{\text{non-anal}}_{H-\lambda}\sim(-1)^{(d-1)/2}(4\pi)^{-d/2}\frac{\Delta^{1/2}}{(N-1)!}\sum_{k=0}^{\infty}\frac{\pi(\sigma/2)^{k+1/2}c_{k-N+(d+1)/2}(\mathscr{F}+m^2)}{\Gamma\left(k-N+1+\frac{d+1}{2}\right)\Gamma\left(k+\frac{3}{2}\right)} \tag{153}$$

$$G^{\text{reg}}_{H-\lambda}\sim(-1)^{(d+1)/2}(4\pi)^{-d/2}\frac{\Delta^{1/2}}{(N-1)!}\sum_{k=0}^{\infty}\frac{\pi(\sigma/2)^{k}}{k!\Gamma(k-N+1+d/2)}\varphi_{\mathscr{F}+m^2}(k-N+d/2), \tag{154}$$

and in even dimensions:

$$G^{\text{non-anal}}_{H-\lambda}\sim(-1)^{d/2-1}(4\pi)^{-d/2}\frac{\Delta^{1/2}}{(N-1)!}\log\left(\frac{\mu^2\sigma}{2}\right)\sum_{k=0}^{\infty}\frac{(\sigma/2)^{k}c_{k-N+d/2}(\mathscr{F}+m^2)}{k!\Gamma(k-N+1+(d/2))}, \tag{155}$$

$$G^{\text{reg}}_{H-\lambda}\sim(-1)^{d/2-1}(4\pi)^{-d/2}\frac{\Delta^{1/2}}{(N-1)!}\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k-N+1+d/2)}\left(\frac{\sigma}{2}\right)^{k}$$
$$\times\{\varphi'_{\mathscr{F}+m^2}(k-N+d/2)-[\log\mu^2+\Psi(k+1)+\Psi(k+d/2)]c_{k-N+(d/2)}(\mathscr{F}+m^2)\}, \tag{156}$$

where

$$\varphi_{\mathscr{F}+m^2}(q)\equiv\frac{(N-1)!\Gamma(q+1)}{\Gamma(q+N)}\text{tr}\ \Pi\ b_{\mathscr{F}+m^2}(q+N-1). \tag{157}$$

Thus, we find that

(i) the structure of the diagonal singularities of the Green function $G_{H-\lambda}$ is determined by the coefficients $c_k(\mathscr{F}+m^2)$;
(ii) in odd dimensions there is no logarithmic singularity;
(iii) for $N>d/2$ there are *no* singularities at all and the Green function is regular on the diagonal, i.e. there exists finite coincidence limit $G^{\text{diag}}_{H-\lambda}\equiv G_{H-\lambda}(x,x)=G^{\text{reg}}_{H-\lambda}(x,x)$;
(iv) for $N=[(d+1)/2]$ there are no singularities for odd dimension $d=2N-1$ and there is only logarithmic singularity in even dimension $d=2N$.
(v) The condition for the validity of the Huygens principle for the operator $H-\lambda$ in even dimensions reads

$$\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k-N+1+d/2)}\left(\frac{\sigma}{2}\right)^{k}c_{k-N+d/2}(\mathscr{F}+m^2)=0. \tag{158}$$

Similarly to (43) this also produces an infinite set of local conditions.

C. Heat kernel of the operator *H*

The heat semigroup for the operator H can be obtained by using the standard formula

$$\exp(-tH)=-\frac{1}{2\pi i}\int_{C_H}d\lambda e^{-t\lambda}G_{H-\lambda}$$
$$=\frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda e^{-t\lambda}\left(\frac{\partial}{\partial\lambda}\right)^{k}G_{H-\lambda}, \tag{159}$$

where the contour of integration C_H goes counter-clockwise around the spectrum of the operator H, k is some sufficiently large integer and c is a negative constant. Since the operator H is a positive self-adjoint operator its spectrum lies on the positive real half-axis, i.e., the contour C_H should go from $i\varepsilon+\infty$ to $-i\varepsilon+\infty$ enclosing the origin. Using the Theorem 1 one can express the heat semigroup for the higher order operator H in terms of the heat semigroup of the Laplace type operator $\mathscr{F}$. We have first

$$\exp(-tH)=\frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\left(\frac{\partial}{\partial\lambda}\right)^k \mathrm{tr}\ \Pi\ G_{\mathscr{F}+m^2}, \tag{160}$$

Now, using the heat kernel representation of the Green operator $G_{\mathscr{F}+m^2}$ we obtain

Corollary 2: The heat semigroup for the operator H is determined by the heat semigroup for the operator $\mathscr{F}$

$$\exp(-tH)=\frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\int_0^\infty ds\ e^{-sm^2}\left(\frac{\partial}{\partial\lambda}\right)^k \mathrm{tr}\ \ \Pi\ \exp(-s\mathscr{F}) \tag{161}$$

and, consequently, the heat kernel for the operator H is given by

$$U_H(t)=(4\pi)^{-d/2}\Delta^{1/2}\frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\int_0^\infty ds\ s^{-d/2}\exp\left(-sm^2-\frac{\sigma}{2s}\right)\left(\frac{\partial}{\partial\lambda}\right)^k \omega_{\mathscr{F}}(s). \tag{162}$$

We studied the expansion of the Green function of a Laplace type operator in the previous section in detail. The important fact is, that the HMDS-coefficients $a_k(\mathscr{F}+m^2)$ for the Laplace type operator $(\mathscr{F}+m^2)$ are polynomial in λ, they are, of course, also polynomial in m^2. This means that the singular part of the Green function (including the logarithmic singularity), which are expressed in terms of $a_k(\mathscr{F}+m^2)$, are also polynomial in λ. On the other hand, it is well known that

$$-\frac{1}{2\pi i}\int_{C_H}d\lambda\ e^{-\lambda}(-\lambda)^z=\frac{1}{\Gamma(-z)}. \tag{163}$$

For integer $z=k=0,1,2,\ldots$ we have, therefore,

$$-\frac{1}{2\pi i}\int_{C_H}d\lambda\ e^{-\lambda}(-\lambda)^k=0. \tag{164}$$

Thus, the singular part, which is polynomial in λ, does not contribute to the integral (160). Therefore, one can substitute in this equation instead of the whole Green function only the regular part of it G^{reg}, go to the diagonal and take the trace

$$\mathrm{Tr}_{L^2}\exp(-tH)=\frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\left(\frac{\partial}{\partial\lambda}\right)^k \mathrm{Tr}_{L^2}\mathrm{tr}\ \Pi\ G^{\mathrm{reg}}_{\mathscr{F}+m^2}, \tag{165}$$

Finally, using eqs. (37) and (38) or (154) and (156) we have: in odd dimensions

$$\mathrm{Tr}_{L^2}\exp(-tH)=\frac{(-1)^{(d+1)/2}(4\pi)^{-d/2}\pi}{(N-1)!\Gamma(d/2)-N+1)}\ \ \frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\left(\frac{\partial}{\partial\lambda}\right)^k \Phi_{\mathscr{F}+m^2}(d/2-N), \tag{166}$$

and in even dimensions

$$\mathrm{Tr}_{L^2}\exp(-tH)=\frac{(-1)^{d/2-1}(4\pi)^{-d/2}}{(N-1)!\Gamma(d/2-N+1)}\ \ \frac{t^{-k}}{2\pi i}\int_{c-i\infty}^{c+i\infty}d\lambda\ e^{-t\lambda}\left(\frac{\partial}{\partial\lambda}\right)^k \Phi'_{\mathscr{F}+m^2}(d/2-N), \tag{167}$$

where $\Phi_{\mathscr{F}+m^2}(q)=\mathrm{Tr}_{L^2}\varphi_{\mathscr{F}+m^2}(q)$ and $\Phi'_{\mathscr{F}+m^2}(q)=(d/dq)\Phi_{\mathscr{F}+m^2}(q)$.

These formulas may be used, for example, to compute the asymptotic expansion of the functional trace of the heat kernel for the higher order operator H, which is known to have the following general form:[3]

$$\mathrm{Tr}_{L^2}\exp(-tH)\sim(4\pi)^{-d/2}\frac{\Gamma[d/(2N)]}{N\Gamma(d/2)}\,t^{-d/(2N)}\sum_{k\geqslant 0}t^{k/N}E_k(H). \tag{168}$$

We normalize the coefficients $E_k(H)$ so that $E_0(H)=\mathrm{Tr}_{L^2}1$. For $N=1$ the coefficients E_k are determined by the HMDS-coefficients $E_k=[(-1)^k/k!]A_k$. However, one should note that to compute the coefficients $E_k(H)$ one needs to know the asymptotics of the function $\omega_{\mathscr{F}}(t)$ [or the functions $\Phi_{\mathscr{F}+m^2}(d/2-N)$ and $\Phi'_{\mathscr{F}+m^2}(d/2-N)$] as $\lambda\to-\infty$, which is unknown, in general, in terms of the known asymptotic expansion as $t\to 0$, i.e. in terms of the HMDS-coefficients for a Laplace type operator. The known asymptotics of these functions as $s\to 0$ and $m\to\infty$ do not contribute at all in eqs. (162), (166), and (167).

ACKNOWLEDGMENTS

I am greatly indebted to Peter Gilkey and Rainer Schimming for stimulating and fruitful discussions. This work was supported by the Deutsche Forschungsgemeinschaft.

[1] J. Hadamard, ''Lectures on Cauchy's Problem,'' in: *Linear Partial Differential Equations* (Yale U. P., New Haven, 1923).
[2] B. S. De Witt, *Dynamical Theory of Groups and Fields* (New York, Gordon and Breach, 1965).
[3] P. B. Gilkey, *Invariance Theory, the Heat Equation and the Atiyah—Singer Index Theorem* (Publish or Perish, Wilmington, DE, 1984).
[4] A. O. Barvinsky and G. A. Vilkovisky, Phys. Rep. **C 119**, 1 (1985).
[5] N. N. Bogolyubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Nauka, Moscow, 1976).
[6] I. G. Avramidi, *Covariant Methods for the Calculation of the Effective Action in Quantum Field Theory and Investigation of Higher-Derivative Quantum Gravity*, Ph.D. thesis, Moscow State University, 1986, UDK 530.12:531.51, [in Russian]; Transl.: hep-th/9510140.
[7] I. G. Avramidi, Nucl. Phys. B **355**, 712 (1991).
[8] B. S. De Witt and R. W. Brehme, Ann. Phys. **9**, 220 (1960).
[9] J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1960).
[10] S. Minakshisundaram, J. Indian Math. Soc. **17**, 158 (1953).
[11] R. T. Seeley, ''Complex powers of an elliptic operator,'' in *Singular integrals*, Proceeding of Symposium in Pure Mathematics, Chicago, IL, 1966 (American Mathematical Society, Providence, RI, 1967), Vol. 10, pp. 288–307.
[12] P. B. Gilkey, J. Diff. Geom. **10**, 601 (1975).
[13] I. G. Avramidi, Phys. Lett. B **238**, 92 (1990).
[14] P. Amsterdamski, A. L. Berkin, and D. J. O'Connor, Class. Quantum Grav. **6**, 1981 (1989).
[15] I. G. Avramidi and R. Schimming, ''Algorithms for the calculation of the heat kernel coefficients,'' in *Quantum Field Theory under the Influence of External Conditions*, edited by M. Bordag, Teubner-Texte zur Physik, Vol. 30 (Teubner, Stuttgart, 1996), p. 150.
[16] R. Schimming, ''Calculation of the heat kernel coefficients,'' in *Analysis, Geometry and Groups. A Riemann Legacy Volume*,'' edited by H. M. Srivastava and Th. M. Rassias (Hadronic, Palm Harbour, 1993), p. 627.
[17] R. Schimming, Beitr. z. Analysis **11**, 45 (1978).
[18] R. Schimming, Z. f. Analysis u. ihre Anw. **2**, 71 (1982).
[19] R. Schimming, Math. Nachr. **147**, 217 (1990).

Quasifinite highest weight modules over the Lie algebra of matrix differential operators on the circle

Carina Boyallian[a)]
FAMAF Universidad Nacional de Córdoba—(5000) Córdoba, Argentina

Victor G. Kac
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

José I. Liberati[a)]
FAMAF Universidad Nacional de Córdoba—(5000) Córdoba, Argentina

Catherine H. Yan
Courant Institute of Mathematical Sciences, New York, New York 10012

(Received 8 December 1997; accepted for publication 14 January 1998)

We give a complete description of the quasifinite highest weight modules over the central extension of the Lie algebra of $M\times M$-matrix differential operators on the circle and obtain them in terms of representation theory of the Lie algebra $\hat{\mathfrak{gl}}(\infty,R_m)$ of infinite matrices with only finitely many nonzero diagonals over the algebra $R_m=\mathbb{C}[t]/(t^{m+1})$. We also classify the unitary ones, and construct them in terms of charged free fermions. This construction provides a large (and conjecturally complete) family of irreducible modules over the associated vertex algebra $W^M_{1+\infty,c}$, where c is a positive integer. © *1998 American Institute of Physics.* [S0022-2488(98)03005-9]

I. INTRODUCTION

The study of representation theory of the Lie algebra $\hat{\mathscr{D}}$ (the universal central extension of the Lie algebra of differential operators on the circle, also denoted by $W_{1+\infty}$), was initiated in Ref. 1. In that paper, Kac and Radul classified the irreducible quasifinite highest weight representations of $\hat{\mathscr{D}}$, realized them in terms of irreducible highest weight representations of the Lie algebra of infinite matrices, and described the unitary ones. This study was continued in Refs. 2 and 3 in the framework of vertex algebra theory.

It was mentioned at the end of Ref. 1 that similar results can be obtained in the matrix case, and this is the main goal of the present paper.

The paper is organized as follows. In Secs. II and III we study the structure of the central extension $\widehat{\mathscr{D}^M}$ of the Lie algebra of $M\times M$-matrix differential operators on the circle, its parabolic subalgebras, and the relation with $\hat{\mathfrak{gl}}(\infty,R_m)$. In Secs. IV and V we classify and construct irreducible quasifinite highest weight modules over $\widehat{\mathscr{D}^M}$ and classify the unitary ones.

In Sec. VI, we consider the simple vertex algebra $W^M_{1+\infty,c}$ constructed on the irreducible vacuum module of $\widehat{\mathscr{D}^M}$, and we construct a large family of representations of this vertex algebra using twisted modules over MN free charged fermions, proving thereby that all primitive $\widehat{\mathscr{D}^M}$-modules are vertex algebra modules for $c\in\mathbb{Z}_+$. We conjecture that these are all irreducible modules over the vertex algebra $W^M_{1+\infty,c}$ if c is a positive integer.

II. LIE ALGEBRAS $\widehat{\mathscr{D}^M}$ AND $\widehat{\mathscr{D}^{M\mathscr{O}}}$

Let M be a positive integer. Denote by $\mathscr{D}^M_{\mathrm{as}}$ the associative algebra of all regular matrix differential operators on $\mathbb{C}^\times$, i.e., the operators on $\mathbb{C}^M[z,z^{-1}]$ of the form

[a)]Current address: Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

 © 1998 American Institute of Physics

$$E=e_k(z)\partial_z^k+e_{k-1}(z)\partial_z^{k-1}+\cdots+e_0(z), \quad \text{where } e_i(z)\in \mathrm{Mat}_M\mathbb{C}[z,z^{-1}],$$

and denote by $\mathscr{D}^M$ the corresponding Lie algebra. Here and further we denote by $\mathrm{Mat}_M R$ the associative algebra of all $M\times M$ matrices over an algebra R. It is more convenient to write the differential operators as linear combinations of elements of the form $z^k f(D)A$, where f is a polynomial, $D=z\partial_z$, $k\in\mathbb{Z}$, and $A\in\mathrm{Mat}_M\mathbb{C}$. The product in $\mathscr{D}^M_{\mathrm{as}}$ is then given by

$$(z^r f(D)A)(z^s g(D)B)=z^{r+s}f(D+s)g(D)AB. \tag{2.1}$$

Fix a *trace functional* on $\mathrm{Mat}_M\mathbb{C}[w]$, which is a linear map $T:\mathrm{Mat}_M\mathbb{C}[w]\to\mathbb{C}$, such that $T(ab)=T(ba)$. Then we have the following 2-cocycle on $\mathscr{D}^M$, where $f(w)$, $g(w)\in\mathrm{Mat}_M\mathbb{C}[w]$ (see Ref. 1):

$$\Psi_T(z^r f(D),z^s g(D))=\begin{cases} T\left(\sum\limits_{-r\leqslant m\leqslant -1} f(w+m)g(w+m+r)\right), & \text{if } r=-s\geqslant 0,\\ 0, \quad \text{if } r+s\neq 0. \end{cases} \tag{2.2}$$

In particular, if $T:\mathrm{Mat}_M\mathbb{C}[w]\to\mathbb{C}$ is the usual trace tr composed with the evaluation map at $w=0$, we obtain the following 2-cocycle on the Lie algebra $\mathscr{D}^M$, where $r,s\in\mathbb{Z}$, $f,g\in\mathbb{C}[w]$, $A,B\in\mathrm{Mat}_M\mathbb{C}$:

$$\Psi(z^r f(D)A,z^s g(D)B)=\begin{cases} \mathrm{tr}(AB)\sum\limits_{-r\leqslant m\leqslant -1} f(m)g(m+r), & \text{if } r=-s\geqslant 0,\\ 0, \quad \text{if } r+s\neq 0. \end{cases} \tag{2.3}$$

Later we shall need another formula for this cocycle (see Ref. 4):

$$\Psi(f(z)\partial_z^m A,g(z)\partial_z^n B)=\frac{(\mathrm{tr}\,AB)m!n!}{(m+n+1)!}\,\mathrm{Res}_{z=0}\, f^{(n+1)}(z)g^{(m)}(z)dz. \tag{2.4}$$

Let

$$\widehat{\mathscr{D}^M}=\mathscr{D}^M+\mathbb{C}C$$

denote the central extension of $\mathscr{D}^M$ corresponding to the cocycle (2.3), i.e., the bracket in the subspace $\mathscr{D}^M$ is given by

$$[E_1E_2]=E_1E_2-E_2E_1+\Psi(E_1,E_2)C, \tag{2.5}$$

and C is a nonzero central element.

The elements $z^kD^me_{ij}$ ($k\in\mathbb{Z}$, $m\in\mathbb{Z}_+$, $i,j=1,...,M$) form a basis of $\mathscr{D}^M$. Here and further e_{ij} is the standard basis of $\mathrm{Mat}_M\mathbb{C}$. Define the *weight* wt on $\mathscr{D}^M$ by

$$\mathrm{wt}\, z^k f(D)e_{ij}=kM+i-j, \quad \mathrm{wt}\, C=0. \tag{2.6}$$

This gives us the *principal* $\mathbb{Z}$-gradation of $\widehat{\mathscr{D}^M}$:

$$\widehat{\mathscr{D}^M}=\bigoplus_{j\in\mathbb{Z}}\widehat{\mathscr{D}^M}_j.$$

Let $\mathscr{O}$ be the algebra of all holomorphic functions on $\mathbb{C}$ with the topology of uniform convengence on compact sets. We consider the vector space $\mathscr{D}^{M\mathscr{O}}_{\mathrm{as}}$ spanned by the matrix differential operators (of infinite order) of the form $z^k f(D)A$, where $f\in\mathscr{O}$. The product in $\mathscr{D}^M_{\mathrm{as}}$ extends to $\mathscr{D}^{M\mathscr{O}}_{\mathrm{as}}$ by formula (2.1). The principal gradation extends as well by (2.6). Denote by $\mathscr{D}^{M\mathscr{O}}$ the

corresponding Lie algebra. Then the cocycle Ψ extends to a 2-cocycle on $\mathscr{D}^{M\mathscr{O}}$ by formula (2.3). Let $\widehat{\mathscr{D}^{M\mathscr{O}}}=\mathscr{D}^{M\mathscr{O}}+\mathbb{C}C$ be the corresponding central extension with the principal gradation as above.

Observe that for elements $z^re^{\lambda D}A$ ($r\in\mathbb{Z}$, $\lambda\in\mathbb{C}$, $A\in \mathrm{Mat}_M\mathbb{C}$) the commutator in $\widehat{\mathscr{D}^{M\mathscr{O}}}$ is given by a very simple formula (cf. Ref. 1):

$$[z^re^{\lambda D}A,z^se^{\mu D}B]=z^{r+s}e^{(\lambda+\mu)D}(e^{\lambda s}AB-e^{\mu r}BA)+\delta_{r,-s}(\mathrm{tr}\,AB)\,\frac{e^{-\lambda r}-e^{\mu s}}{1-e^{\lambda+\mu}}\,C.$$

Remark 2.1: Consider the following traces on $\mathrm{Mat}_M\mathscr{O}$:

$$\mathrm{tr}_{a,b}\,f(w)=\mathrm{tr}(f(a)-f(b)),\quad \text{where}\quad f\in\mathrm{Mat}_M\mathscr{O},\quad \text{and}\quad a,b\in\mathbb{C},$$

$$\mathrm{tr}_s^{[m]}\,f(w)=\frac{d^m}{dw^m}\bigg|_{w=s}\mathrm{tr}(f(w)),\quad \text{where}\quad s\in\mathbb{C},\quad m\in\mathbb{Z}_+,$$

and denote the corresponding cocycles by $\Psi_{a,b}:=\Psi_{\mathrm{tr}_{a,b}}$ and $\Psi_s^{(m)}:=\Psi_{\mathrm{tr}_s^{[m]}}$; see (2.2). On $\mathscr{D}^{M\mathscr{O}}$ these cocycles are nontrivial. However, when we restrict to $\mathscr{D}^M$ they become trivial [see Ref. 1, 1.3(d)]. We have the following explicit formulas for these trivial cocycles on $\mathscr{D}^M$:

$$\Psi_{a,b}(z^rf(D)A,z^sg(D)B)=\delta_{r,-s}\Lambda_{a,b}([z^rf(D)A,z^sg(D)B]),\tag{2.7}$$

$$\Psi_s^{(m)}(z^rf(D)A,z^sg(D)B)=\delta_{r,-s}\Lambda_s^{[m]}([z^rf(D)A,z^sg(D)B]),\tag{2.8}$$

where $\Lambda_{a,b}$ and $\Lambda_s^{[m]}$ are the linear functions on $\mathrm{Mat}_M\mathbb{C}[w]$ defined by the following generating series in x:

$$\Lambda_{a,b}(e^{xw}A)=-\frac{e^{ax}-e^{bx}}{e^x-1}\,\mathrm{tr}\,A,\quad \Lambda_s^{[m]}(e^{xw}A)=-\frac{x^me^{sx}}{e^x-1}\,\mathrm{tr}\,A.\tag{2.9}$$

Define a *parabolic* subalgebra $\mathfrak{p}$ of $\widehat{\mathscr{D}^M}$ as a subalgebra of the following form:

$$\mathfrak{p}=\bigoplus_{j\in\mathbb{Z}}\mathfrak{p}_j,\quad \text{where}\quad \mathfrak{p}_j=\widehat{\mathscr{D}^M}_j,\quad \text{for}\quad j\geqslant 0,\quad \text{and}\quad \mathfrak{p}_j\subset\widehat{\mathrm{D}^M}_j,\quad \text{for}\quad j<0.$$

Observe that

$$\widehat{\mathscr{D}^M}_j=\bigoplus_{k,l,m:kM+l-m=j}z^k\mathbb{C}[D]e_{lm};$$

hence, for each $j\in\mathbb{N}$ we have

$$\mathfrak{p}_{-j}=\bigoplus_{k,l,m:kM+l-m=-j}z^kI_{-j}^le_{lm},\tag{2.10}$$

where I_{-j}^l is a subspace of $\mathbb{C}[w]$. Since

$$[f(D)e_{ll},p(D)e_{lm}]=f(D)p(D)e_{lm},\quad l\neq m,$$

and

$$[f(D)I_M,z^kp(D)e_{lm}]=z^k(f(D+k)-f(D))p(D)e_{lm},$$

we see that I_{-j}^l is an ideal of $\mathbb{C}[w]$.

Remark 2.2: There exist parabolic subalgebras $\mathfrak{p}$ of $\widehat{\mathscr{D}^M}$ for $M>1$, such that $\mathfrak{p}_j\neq 0$ for some $j<0$, but $\mathfrak{p}_j=0$ for $j\ll 0$ (cf. Ref. 1, Sec. 2.4). For example,

$$\mathfrak{p}=\bigoplus_{j\geqslant 0}(\widehat{\mathscr{D}^M})_j\oplus\mathbb{C}[w]e_{12}.$$

Given a parabolic subalgebra $\mathfrak{p}$, denote by $b_j^l(w)$ a generator of the ideal $I_{-j}^l\subset\mathbb{C}[w]$. We will assume that $b_j^l(w)$ is a monic polynomial if I_{-j}^l is nonzero, otherwise $b_j^l(w)=0$. Thus, we have associated to a parabolic subalgebra $\mathfrak{p}$ a collection $\{b_j^l(w)\}_{j\in\mathbb{N},1\leqslant l\leqslant M}$, consisting of monic polynomials or zero, called *characteristic* polynomials of $\mathfrak{p}$.

Lemma 2.1: Let $\{b_j^l\}$ *be the set of characteristic polynomials of a parabolic subalgebra* $\mathfrak{p}$ *of the Lie algebra* $\widehat{\mathscr{D}^M}$. *Then*

(1) *If* $b_j^{i+1}(w)\neq 0$, *then it divides* $b_{j+1}^i(w)$ *for all* $j\in\mathbb{N}$ *and* $i=1,...,M$ (*we let* $b_j^{M+1}=b_j^1$).
(2) *If* $b_j^i(w)\neq 0$, *then it divides* $b_{j+1}^i(w)$, *except in the special case*: $-j=kM+i$, *when it divides* $b_{j+1}^i(w+1)$.
(3) *If* $b_{(k+l)M+n-i}^i(w)\neq 0$, *then it divides* $b_{kM+p-i}^i(w-l)b_{lM+n-p}^p(w)$ *for all* $k,l\in\mathbb{N}$, $1\leqslant i$, $p\leqslant M$.

Proof: Using $[\mathfrak{p}_1,\mathfrak{p}_{-j-1}]\subseteq\mathfrak{p}_{-j}$ and the commutator

$$[e_{i+1,i},z^{-k}b_{j+1}^i(D)e_{il}]=z^{-k}b_{j+1}^i(D)e_{i+1,l}-z^{-k}\delta_{i+1,l}b_{j+1}^i(D)e_{ii},$$

we see that $b_{j+1}^i(D)\in I_{-j}^{i+1}$ for $1\leqslant i\leqslant M-1$. Since

$$[ze_{1,M},z^{-k}b_{j+1}^i(D)e_{im}]=z^{-k+1}b_{j+1}^i(D)\,\delta_{i,M}e_{1,m}-z^{-k+1}b_{j+1}^i(D+1)\,\delta_{1,m}e_{i,M},$$

we obtain that $b_j^1(w)$ divides $b_{j+1}^M(w)$, proving (1). The proof of (2) follows from the above commutator and

$$[e_{l,l-1},z^{-k}b_{j+1}^i(D)e_{il}]=z^{-k}b_{j+1}^i(D)\,\delta_{l-1,i}e_{l,l}-z^{-k}b_{j+1}^i(D)e_{i,l-1}.$$

Finally, the proof of (3) is similar by computing the commutator

$$[z^{-k}b_{kM+m-i}^ie_{i,m},z^{-l}b_{lM+n-p}^p(D)e_{p,n}].$$ □

A parabolic subalgebra is called *nondegenerate* if all $\mathfrak{p}_j$, $j\in\mathbb{Z}$, are nonzero. It is immediate to deduce from Lemma 2.1 the following corollary.

Corollary 2.1: Let $\mathfrak{p}$ *be a parabolic subalgebra of* $\widehat{\mathscr{D}^M}$. *Then* $\mathfrak{p}$ *is nondegenerate if and only if all polynomials* $b_1^1(w),\ldots,b_1^M(w)$ *are nonzero.*

Given M monic polynomials $\mathbf{b}=(b^1(w),\ldots,b^M(w))$, we define (cf. Lemma 2.1)

$$\begin{aligned}(b^{\min})_{kM+m-i}^i(w)=&\,b^i(w-k)b^{i+1}(w-k)\cdots b^{M-1}(w-k)b^M(w-k+1)\\&\times b^1(w-k+1)\cdots b^{M-1}(w-k+1)b^M(w-k+2)b^1(w-k+2)\cdots b^M(w-1)\\&\times b^1(w-1)\cdots b^{M-1}(w-1)b^M(w)b^1(w)\cdots b^{m-1}(w).\end{aligned}$$

It is clear that there exists (a unique nondegenerate) parabolic subalgebra that we denote by $\mathfrak{p}_{\min}(\mathbf{b})$, for which the characteristic polynomials are $\{(b^{\min})_j^i\}$. We also have

$$\dim(\widehat{\mathscr{D}^M}_k/\mathfrak{p}_{\min}(\mathbf{b})_k)<\infty,\quad k\in\mathbb{Z}.\tag{2.11}$$

Remark 2.3: Any parabolic subalgebra $\mathfrak{p}$ such that $b_1^i=b^i$, satisfies $\mathfrak{p}_{\min}(\mathbf{b})\subseteq\mathfrak{p}$. This follows from Lemma 2.1.

Given M polynomials $\mathbf{b}=(b^1(w),\ldots,b^M(w))$, consider the following subspace of $\widehat{\mathscr{D}^M}_0$:

$$\begin{aligned}(\widehat{\mathscr{D}^M}_0)^{\mathbf{b}}=&\{b^i(D)g(D)e_{i+1,i+1}-b^i(D)g(D)e_{ii}|1\leqslant i\leqslant M-1,g(w)\in\mathbb{C}[w]\}\oplus\{b^M(D)g(D)e_{1,1}\\&-b^M(D+1)g(D+1)e_{MM}+b^M(0)g(0)C|g(w)\in\mathbb{C}[w]\}.\end{aligned}$$

We shall need the following proposition to study modules over $\widehat{\mathscr{D}^M}$ induced from its parabolic subalgebras.

Proposition 2.1: Let $\mathfrak{p}$ *be a parabolic subalgebra of* $\widehat{\mathscr{D}^M}$ *and let* $\mathbf{b}=(b^1(w),\dots,b^M(w))$ *be its first* M *characteristic polynomials. Then*

$$[\mathfrak{p},\mathfrak{p}]=(\underset{k\neq 0}{\oplus}\mathfrak{p}_k)\bigoplus(\widehat{\mathscr{D}^M}_0)^{\mathbf{b}}.$$

Proof: First, clearly $[\mathfrak{p}_0,\mathfrak{p}_k]=\mathfrak{p}_k$ for $k\neq 0$. Next, since $[\mathfrak{p}_1,\mathfrak{p}_k]=\mathfrak{p}_{k+1}$ for $k\geqslant 1$, we have

$$\begin{aligned}[\mathfrak{p}_{1+k},\mathfrak{p}_{-(1+k)}]&\subseteq[[\mathfrak{p}_1,\mathfrak{p}_k],\mathfrak{p}_{-(1+k)}]\\&\subseteq[[\mathfrak{p}_k,\mathfrak{p}_{-(k+1)}],\mathfrak{p}_1]+[[\mathfrak{p}_{-(k+1)},\mathfrak{p}_1],\mathfrak{p}_k]\\&\subseteq[\mathfrak{p}_{-1},\mathfrak{p}_1]+[\mathfrak{p}_{-k},\mathfrak{p}_k].\end{aligned}$$

Hence, by induction, $[\mathfrak{p},\mathfrak{p}]_0=[\mathfrak{p}_{-1},\mathfrak{p}_1]$. Now, using that

$$[g(D)e_{i+1,i},b^i(D)e_{i,i+1}]=b^i(D)g(D)e_{i+1,i+1}-b^i(D)g(D)e_{ii}$$

and

$$[zg(D+1)e_{1,M},z^{-1}b^M(D)e_{M,1}]=b^M(D)g(D)e_{1,1}-b^M(D+1)g(D+1)e_{MM}+b^M(0)g(0)C,$$

we see that $[\mathfrak{p}_{-1},\mathfrak{p}_1]=(\widehat{\mathscr{D}^M}_0)^{\mathbf{b}}$; hence the proposition follows. □

III. CONNECTION BETWEEN $\widehat{\mathscr{D}^M}$ AND $\widehat{\mathfrak{gl}}(\infty)[m]$

We will discuss the relation between $\widehat{\mathscr{D}^M}$ and the Lie algebra of infinite matrices with finitely many nonzero diagonals over the algebra of truncated polynomials.

Let R be an associative algebra over $\mathbb{C}$ and denote by R^∞ a free R-module with a fixed basis $\{v_j\}_{j\in\mathbb{Z}}$ and define the operators E_{ij} on R^∞ by $E_{ij}v_k=\delta_{jk}v_i$.

Let $\widetilde{M}(\infty,R)$ be the associative subalgebra of $\operatorname{End}R^\infty$ consisting of all operators $\Sigma_{i,j\in\mathbb{Z}}a_{ij}E_{ij}$, where $(a_{ij})_{i,j\in\mathbb{Z}}$ has a finite number of nonzero diagonals. Recall that $\widetilde{M}(\infty,R)$ is a $\mathbb{Z}$-graded algebra with the *principal* gradation defined by $\deg E_{ij}=j-i$.

Let $\varphi:R^M[z,z^{-1}]\to R[z,z^{-1}]$ be the isomorphism defined by

$$e_iz^j\longrightarrow z^{jM+i-1}.$$

Fix $s\in\mathbb{C}$ and a nilpotent element $t\in R$. Consider $R[z,z^{-1}]z^s$ as a free R-module with basis $v_j=z^{-j+s}$, $j\in\mathbb{Z}$. Using the isomorphism φ we associate to an element $z^kf(D)e_{ij}\in\mathscr{D}^M_{\text{as}}$ the operator $z^kf(D+t)e_{ij}$ on $R[z,z^{-1}]z^s$. More precisely, $(z^kf(D)e_{ij})(v_{lM-m+1})=(z^kf(D+t)e_{ij})(e_mz^{-l+s})=\delta_{jm}f(-l+s+t)e_iz^{k-l+s}=\delta_{jm}f(-l+s+t)v_{(l-k)M-i+1}$.

Thus, we obtain an embedding $\varphi_{s,t}:\mathscr{D}^M_{\text{as}}\to\widetilde{M}(\infty,R)$ of associative algebras over $\mathbb{C}$, which is compatible with the principal gradation. Explicitly,

$$\varphi_{s,t}(z^kf(D)e_{ij})=\sum_{l\in\mathbb{Z}}f(-l+s+t)E_{lM-kM-i+1,lM-j+1},\tag{3.1}$$

and $\varphi_{s,t}$ extends to a homomorphism $\varphi_{s,t}:\mathscr{D}^{M\mathscr{O}}_{\text{as}}\to\widetilde{M}(\infty,R)$.

Consider the truncated polynomial algebra $R_m=\mathbb{C}[t]/(t^{m+1})$, where $m\in\mathbb{Z}_+$, and let $\widetilde{M}(\infty)[m]=\widetilde{M}(\infty,R_m)$. The homomorphism $\varphi_{s,t}:\mathscr{D}^{M\mathscr{O}}_{\text{as}}\to\widetilde{M}(\infty)[m]$ given by (3.1) will be denoted by $\varphi_s^{[m]}$. It follows from Taylor's formula that

$$\varphi_s^{[m]}(z^kf(D)e_{ij})=\sum_{r=0}^m\sum_{l\in\mathbb{Z}}\frac{f^{(r)}(-l+s)}{r!}t^rE_{lM-kM-i+1,lM-j+1}.\tag{3.2}$$

Let

$$I_s^{[m]}=\{f\in\mathscr{O}\,|\,f^{(i)}(j+s)=0,\quad \text{for all } j\in\mathbb{Z},\ i=0,1,...,m\}$$

and

$$J_s^{[m]}=\bigoplus_{i,j=1}^{M}\ \bigoplus_{k\in\mathbb{Z}} z^k I_s^{[m]}e_{ij}\in\mathscr{D}_{\text{as}}^{M\mathscr{O}}.$$

We obviously have

$$\text{Ker }\varphi_s^{[m]}=J_s^{[m]}. \tag{3.3}$$

Now, let us fix $\mathbf{s}=(s_1,\ldots,s_n)\in\mathbb{C}^n$, such that

$$s_i-s_j\notin\mathbb{Z},\quad \text{for } i\neq j, \tag{3.4}$$

and fix $\mathbf{m}=(m_1,\ldots,m_n)\in\mathbb{Z}_+^n$.

Let $\widetilde{M}(\infty)[\mathbf{m}]=\oplus_{i=1}^n\widetilde{M}(\infty)[m_i]$. Consider the homomorphism

$$\varphi_{\mathbf{s}}^{[\mathbf{m}]}=\bigoplus_{i=1}^{n}\varphi_{s_i}^{[m_i]}:\mathscr{D}_{\text{as}}^{M\mathscr{O}}\to\widetilde{M}(\infty)[\mathbf{m}].$$

Proposition 3.1: We have an exact sequence of $\mathbb{Z}$*-graded associative algebras*:

$$0\to J_{\mathbf{s}}^{[\mathbf{m}]}\to\mathscr{D}_{\text{as}}^{M\mathscr{O}}\xrightarrow{\varphi_{\mathbf{s}}^{[\mathbf{m}]}}\widetilde{M}(\infty)[\mathbf{m}]\to 0,$$

where $J_{\mathbf{s}}^{[\mathbf{m}]}=\cap_{i=1}^n J_{s_i}^{[m_i]}$.

Proof: The first part is clear from (3.3). The surjectivity of $\varphi_{\mathbf{s}}^{[\mathbf{m}]}$ follows from the following well-known fact: for every discrete sequence of points in $\mathbb{C}$ and a nonnegative integer m, there exists $f(w)\in\mathscr{O}$ having prescribed values of its first m derivatives at these points. We have used (3.4) here. □

Now, denote by $\widetilde{\mathrm{gl}}(\infty,R_m)$ the Lie algebra corresponding to the associative algebra $\widetilde{M}(\infty)[m]$. Let $\widehat{\mathrm{gl}}(\infty,R_m)=\widetilde{\mathrm{gl}}(\infty,R_m)+R_m$ be the central extension with this 2-cocycle,

$$\Phi(A,B)=\mathrm{tr}([J,A]B), \tag{3.5}$$

where $J=\Sigma_{i\leqslant 0}E_{ii}$. The principal $\mathbb{Z}$-gradation of this Lie algebra extends from $\widetilde{\mathrm{gl}}(\infty,R_m)$ by letting wt $R_m=0$.

The associative algebra homomorphism $\varphi_s^{[m]}$ defines a Lie algebra homomorphism, which we denote by the same letter.

The restriction of the cocycle Φ to $\varphi_s^{[m]}(\mathscr{D}^{M\mathscr{O}})$ gives us the following R_m-valued 2-cocycle on $\mathscr{D}^{M\mathscr{O}}$:

$$\Psi_s^{[m]}:=\Psi+\Psi_{s,0}+\sum_{j=1}^{m}\Psi_s^{(j)}\frac{t^j}{j!},$$

where the cocycle Ψ is given by (2.3) and the cocycles $\Psi_{s,0}$ and $\Psi_s^{(m)}$ are defined in Remark 2.1. Thus, formulas (2.7)–(2.9) imply the following result.

Proposition 3.2: The $\mathbb{C}$*-linear map* $\hat{\varphi}_s^{[m]}:\widehat{\mathscr{D}^M}\to\widehat{\mathrm{gl}}(\infty,R_m)$*, defined by*

$$\hat{\varphi}_s^{[m]}|_{\widehat{\mathscr{D}^M}_j}=\varphi_s^{[m]}|_{\mathscr{D}^M{}_j},\quad \text{if } j\neq 0,$$

$$\hat{\varphi}_s^{[m]}(e^{xD}e_{ii}) = \varphi_s^{[m]}(e^{xD}e_{ii}) - \frac{e^{sx}-1}{e^x-1} - \sum_{j=1}^{m} \frac{x^j e^{sx}}{e^x-1}\, t^j/j!, \tag{3.6}$$

$$\hat{\varphi}_s^{[m]}(C) = 1 \in R_m,$$

is a homomorphism of Lie algebras over $\mathbb{C}$. □

IV. QUASIFINITE HIGHEST WEIGHT MODULES OVER $\widehat{\mathscr{D}^M}$

Recall that, given $\lambda \in (\widehat{\mathscr{D}^M}_0)^*$, the *Verma module* over $\widehat{\mathscr{D}^M}$ is defined as follows:

$$M(\lambda) = M(\widehat{\mathscr{D}^M}, \lambda) := \mathscr{U}(\widehat{\mathscr{D}^M}) \oplus_{\mathscr{U}(\oplus_{j\in\mathbb{Z}_+} \widehat{\mathscr{D}^M}_j)} \mathbb{C}_\lambda,$$

where $\mathbb{C}_\lambda$ is the one-dimensional $\oplus_{j\in\mathbb{Z}_+} \widehat{\mathscr{D}^M}_j$-module given by $h \mapsto \lambda(h)$ if $h \in \widehat{\mathscr{D}^M}_0$, $\widehat{\mathscr{D}^M}_j \mapsto 0$ for $j>0$, and the action of $\widehat{\mathscr{D}^M}$ is induced by the left multiplication in $\mathscr{U}(\widehat{\mathscr{D}^M})$. The vector $v_\lambda := 1 \otimes 1$ is called the *highest weight vector*.

Let $L(\lambda) = L(\widehat{\mathscr{D}^M}, \lambda)$ denote the unique irreducible quotient of $M(\widehat{\mathscr{D}^M}, \lambda)$. Modules $M(\lambda)$ and their quotients [including $L(\lambda)$] are called *highest weight modules* (with highest weight λ). Note that the principal gradation on $\widehat{\mathscr{D}^M}$ induces the principal gradation on any highest weight module V:

$$V = \bigoplus_{j\in\mathbb{Z}+} V_{-j}, \quad \text{where} \quad V_0 = \mathbb{C}v_\lambda. \tag{4.1}$$

Consider a parabolic subalgebra $\mathfrak{p}$ and let $\mathbf{b}$ be its first M characteristic polynomials. Let $\lambda \in (\widehat{\mathscr{D}^M}_0)^*$ be such that $\lambda|_{(\widehat{\mathscr{D}^M}_0)^{\mathfrak{b}}} = 0$. Then the $\oplus_{j\in\mathbb{Z}_+} \widehat{\mathscr{D}^M}_j$-module $\mathbb{C}_\lambda$ extends to $\mathfrak{p}$ by letting $[\mathfrak{p},\mathfrak{p}] \mapsto 0$ (see Proposition 2.1). Denote by $M(\lambda;\mathbf{b})$ the *generalized Verma module*, i.e., the $\widehat{\mathscr{D}^M}$-module,

$$M(\widehat{\mathscr{D}^M}, \mathfrak{p}, \lambda) = \mathscr{U}(\widehat{\mathscr{D}^M}) \bigoplus_{\mathscr{U}(\mathfrak{p})} \mathbb{C}_\lambda.$$

Recall that a nonzero vector v in a highest weight module over $\widehat{\mathscr{D}^M}$ is called *singular* if $(\widehat{\mathscr{D}^M})_j v = 0$, for all $j>0$.

Definition 4.1: A Verma module $M(\lambda)$ over $\widehat{\mathscr{D}^M}$ is called *highly degenerate* if there exists a singular vector $v \in M(\lambda)_{-1}$ of the following form: $v = Av_\lambda$, where $A \in (\widehat{\mathscr{D}^M})_{-1}$ and $\det A \neq 0$.

Proposition 4.1: $M(\lambda)$ is highly degenerate if and only if $\lambda((\widehat{\mathscr{D}^M}_0)^{\mathbf{b}}) = 0$ for some $\mathbf{b} = (b^1(w),\ldots,b^M(w))$ where all $b^i(w)$ are monic polynomials.

Proof: Suppose that $v = Av_\lambda$, with $A \in (\widehat{\mathscr{D}^M})_{-1}$ and $\det A \neq 0$. Then

$$A = \begin{pmatrix} 0 & b^1(D) & 0 & \ldots & 0 \\ 0 & 0 & b^2(D) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & b^{M-1}(D) \\ z^{-1}b^M(D) & 0 & 0 & \ldots & 0 \end{pmatrix};$$

hence $b^i(D) \neq 0$ for all i. Observe that v is a singular vector if and only if $(\widehat{\mathscr{D}^M})_1 v = 0$. Given $B \in (\widehat{\mathscr{D}^M})_1$, we have $Bv = BAv_\lambda = [B,A]v_\lambda + ABv_\lambda = \lambda([B,A])v_\lambda$; hence the proposition follows. □

An irreducible highest weight $\widehat{\mathscr{D}^M}$-module V is called *quasifinite* if $\dim V_{-j}<\infty$ for all $j \in \mathbb{Z}_+$. The following key proposition follows from Propositions 2.1 and 4.1, and formula (2.11).

Proposition 4.2: The following conditions on $\lambda \in (\widehat{\mathscr{D}^M}_0)^*$ *are equivalent*:
(1) $M(\lambda)$ is *highly degenerate*;
(2) $L(\lambda)$ *is quasifinite*;
(3) $L(\lambda)$ *is the irreducible quotient of a generalized Verma module* $M(\lambda;\mathbf{b})$ *for some monic polynomials* $\mathbf{b}=(b^1(w),\ldots,b^M(w))$.

Let $L(\lambda)$ be a quasifinite highest weight module over $\widehat{\mathscr{D}^M}$. By Proposition 4.2 there exist some monic polynomials $b^1(w),\ldots,b^M(w)$, such that

$$(b^i(D)e_{i,i+1})v_\lambda=0 \quad (i=1,\ldots,M-1), \quad (z^{-1}b^M(D)e_{M,1})v_\lambda=0.$$

We shall call such monic polynomials of minimal degree the *characteristic* polynomials of $L(\lambda)$. Note that $L(\lambda)$ is the irreducible quotient of $M(\lambda;\mathbf{b})$, where $\mathbf{b}=(b^1(w),\ldots,b^M(w))$ are the characteristic polynomials of $L(\lambda)$.

A functional $\lambda \in (\widehat{\mathscr{D}^M}_0)^*$ is described by its *labels* $\Delta_{i,m}=-\lambda(D^m e_{ii})$ and the *central charge* $c=\lambda(C)$. We shall consider the generating series

$$\Delta_i(x)=\sum_{m=0}^{\infty} \frac{x^m}{m!}\Delta_{i,m}.$$

Recall that a *quasipolynomial* is a linear combination of functions of the form $p(x)e^{\alpha x}$, where $p(x)$ is a polynomial and $\alpha \in \mathbb{C}$.

Theorem 4.1: *A* $\widehat{\mathscr{D}^M}$*-module* $L(\lambda)$ *is quasifinite if and only if*

$$\Delta_M(x)=\frac{\varphi_M(x)}{e^x-1},$$

$$\Delta_i(x)=\frac{\varphi_i(x)(e^x-1)+\varphi_M(x)}{e^x-1}, \quad \text{for} \quad i=1,\ldots,M-1,$$

where $\varphi_i(x)$ *are all quasipolynomials and* $\varphi_M(0)=0$.

Proof: From Propositions 4.1, 4.2, and 2.1, we have that $L(\lambda)$ is quasifinite if and only if there exist polynomials

$$b^i(w)=w^{m_i}+f_{i,m_i-1}w^{m_i-1}+\cdots+f_{i,0}, \quad \text{for} \quad i=1,\ldots,M,$$

such that for all $s=0,1,\ldots$, we have

$$\lambda(b^i(D)D^s e_{i+1,i+1}-b^i(D)D^s e_{ii})=0, \quad i=1,\ldots,M-1,$$

$$\lambda(b^M(D)D^s e_{1,1}-b^M(D+1)(D+1)^s e_{MM}+\delta_{s,0}b^M(0)C)=0.$$

These conditions can be rewritten as follows:

$$\sum_{n=0}^{m_i} f_{i,n}F_{i,n+s}=0, \quad \text{for all} \quad s=0,1,\ldots, \tag{4.2}$$

where

$$F_{i,n}=\Delta_{i,n}-\Delta_{i+1,n}, \quad \text{for} \quad i=1,\ldots,M-1,$$

$$F_{M,n}=\delta_{n,0}c-\Delta_{1,n}+\sum_{j=0}^{n}\Delta_{M,j}.$$

Introducing the generating series,

$$F_i(x)=\sum_{n=0}^{\infty}\frac{x^n}{n!}F_{i,n},$$

we may rewrite (4.2) in the form

$$\left(\sum_{j=0}^{m_i} f_{i,j}\left(\frac{d}{dx}\right)^j\right)F_i(x)=0.$$

Thus, $L(\lambda)$ is quasifinite if and only if $F_i(x)$ are quasipolynomials. But

$$\begin{aligned} &F_i(x)=\Delta_i(x)-\Delta_{i+1}(x), \quad \text{for} \quad i=1,...,M-1,\\ &F_M(x)=c-\Delta_1(x)+e^x\Delta_M(x); \end{aligned} \tag{4.3}$$

hence the theorem follows. □

Definition 4.2: Each of the quasipolynomials $F_i(x)$ $(i=1,...,M)$ in (4.3) can be (uniquely) written in the form

$$F_i(x)=\sum_{r\in\mathbb{C}} p_{i,r}(x)e^{rx}, \quad p_{i,r}(x)\neq 0. \tag{4.4}$$

The numbers r appearing in (4.4) are called *exponents* of the $\widehat{\mathscr{D}^M}$-module $L(\lambda)$.

Corollary 4.1: Let $L(\Lambda)$ be a quasifinite irreducible highest weight module over $\widehat{\mathscr{D}^M}$, and let $\mathbf{b}=(b^1(w),\ldots,b^M(w))$ be its first M characteristic polynomials. By Theorem 4.1, $F_i=\Delta_i(x)-\Delta_{i+1}(x)$, $i=1,...,M-1$, and $F_M=c+e^x\Delta_M(x)-\Delta_1(x)$ are quasipolynomials. Let $(F_i)^{(m_i)}+f_{i,m_i-1}(F_i)^{(m_i-1)}+\cdots+f_{i,0}F_i=0$ be the minimal-order homogeneous linear differential equations with constant coefficients satisfied by $F_i(x)$. Then $b^i(w)=w^{m_i}+f_{i,m_i-1}w^{m_i-1}+\cdots+f_{i,0}$. Moreover, the exponents appearing in (4.4) are all the roots of the polynomials $b^i(w)$.

We shall need the following proposition.

Proposition 4.3: Let V be a quasifinite $\widehat{\mathscr{D}^M}$-module. Then the action of $\widehat{\mathscr{D}^M}$ on V naturally extends to the action of $\widehat{\mathscr{D}^M\mathscr{O}}_k$ on V for any $k\neq 0$.

Proof: Replace $B=\operatorname{ad} D^2-k^2$ by $B=\frac{1}{2}\operatorname{ad}(D^2e_{i,i}-(D-k)^2e_{j,j})$ in the proof of Proposition 4.3 in Ref. 1. The rest of the proof remains the same. □

We return now to the $\mathbb{Z}$-graded complex Lie algebra $\mathfrak{g}^{[m]}:=\widehat{\mathrm{gl}}(\infty,R_m)=\widetilde{\mathrm{gl}}(\infty,R_m)+R_m$ considered in Sec. III. Recall that it is a central extension of the Lie algebra $\widetilde{\mathrm{gl}}(\infty,R_m)$ over $\mathbb{C}$ by the $m+1$-dimensional space R_m.

An element $\lambda\in(\mathfrak{g}_0^{[m]})^*$ is characterized by its *labels*,

$$\lambda_k^{(j)}=\lambda(t^jE_{kk}), \quad k\in\mathbb{Z}, \quad j=0,...,m,$$

and *central charges*,

$$c_j=\lambda(t^j), \quad j=0,1,...,m.$$

Let

$$h_k^{(j)}=\lambda_k^{(j)}-\lambda_{k+1}^{(j)}+\delta_{k,0}c_j, \quad k\in\mathbb{Z}, \quad j=0,...,m. \tag{4.5}$$

As usual, we have the irreducible highest weight $\mathfrak{g}^{[m]}$-module $L(\mathfrak{g}^{[m]},\lambda)$ associated to λ. The proof of the following proposition is standard.

Proposition 4.4: The $\mathfrak{g}^{[m]}$-module $L(\mathfrak{g}^{[m]},\lambda)$ is quasifinite if and only if for each $j=0,1,...,m$ all but finitely many of the $h_k^{(j)}$ are zero.

Given $\mathbf{m}=(m_1,\dots,m_n)\in\mathbb{Z}_+^n$, we let $\mathfrak{g}^{[\mathbf{m}]}=\oplus_{i=1}^n \mathfrak{g}^{[m_i]}$. By Proposition 3.1, we have a surjective homomorphism of Lie algebras over $\mathbb{C}$ for $\mathbf{s}$ satisfying (3.4):

$$\hat{\varphi}_{\mathbf{s}}^{[\mathbf{m}]}=\bigoplus_{i=1}^{n}\hat{\varphi}_{s_i}^{[m_i]}:\widehat{\mathscr{D}^{M\mathscr{O}}}\to\mathfrak{g}^{[\mathbf{m}]}. \tag{4.6}$$

Take $\lambda(i)\in(\mathfrak{g}_0^{[m_i]})^*$ such that $L(\mathfrak{g}^{[m_i]},\lambda(i))$ is a quasifinite irreducible $\mathfrak{g}^{[m_i]}$-module. Then

$$L(\mathfrak{g}^{[\mathbf{m}]},\boldsymbol{\lambda})=\bigotimes_{i=1}^{n}L(\mathfrak{g}^{[m_i]},\lambda(i))$$

is an irreducible $\mathfrak{g}^{[\mathbf{m}]}$-module. Now, via the homomorphism (4.6), we consider $L(\mathfrak{g}^{[\mathbf{m}]},\boldsymbol{\lambda})$ as a $\widehat{\mathscr{D}^M}$-module, and we denote it by $L_{\mathbf{s}}^{[\mathbf{m}]}(\boldsymbol{\lambda})$.

We have the following important theorem.

Theorem 4.2: *Let* $\mathbf{s}=(s_1,\dots,s_n)$ *satisfy (3.4). Consider the embedding* $\hat{\varphi}_{\mathbf{s}}^{[\mathbf{m}]}:\widehat{\mathscr{D}^M}\to\mathfrak{g}^{[\mathbf{m}]}$ *and let V be a quasifinite* $\mathfrak{g}^{[\mathbf{m}]}$*-module. Then any* $\widehat{\mathscr{D}^M}$*-submodule of V is a* $\mathfrak{g}^{[\mathbf{m}]}$*-submodule as well. In particular, the* $\widehat{\mathscr{D}^M}$*-modules* $L_{\mathbf{s}}^{[\mathbf{m}]}(\boldsymbol{\lambda})$ *are irreducible.*

Proof: The same proof as of Theorem 4.5 in Ref. 1. □

By Proposition 4.4 and Theorem 4.2, we have that the $\widehat{\mathscr{D}^M}$-modules $L_{\mathbf{s}}^{[\mathbf{m}]}(\boldsymbol{\lambda})$ are irreducible quasifinite highest weight modules. Using formulas (3.2) and (3.6), it is easy to calculate the generating series $\Delta_{\mathbf{m},\mathbf{s},\boldsymbol{\lambda},i}(x)=\Sigma_{n\geqslant 0}\Delta_{i,n}(x^n/n!)$ of the highest weight and the central charge c of the $\widehat{\mathscr{D}^M}$-module $L_{\mathbf{s}}^{[\mathbf{m}]}(\boldsymbol{\lambda})$. We have

$$\Delta_{m,s,\lambda,i}(x)=-\sum_{p=0}^{m}\sum_{l\in\mathbb{Z}}\frac{\lambda_{lM-i+1}^{(p)}}{p!}x^p e^{(s-l)x}+\frac{e^{sx}-1}{e^x-1}c_0+\sum_{p=1}^{m}\frac{x^p e^{sx}}{e^x-1}\frac{c_p}{p!}, \tag{4.7}$$

$$c=c_0, \tag{4.8}$$

and

$$\Delta_{\mathbf{m},\mathbf{s},\boldsymbol{\lambda},i}(x)=\sum_j \Delta_{m_j,s_j,\lambda(j),i}(x),\quad c=\sum_j c_0(j). \tag{4.9}$$

Introduce the polynomials [see (4.5)]

$$g_k(x)=\sum_{j=0}^{m}h_k^{(j)}x^j/j!\quad(k\in\mathbb{Z}). \tag{4.10}$$

Then (4.7) can be rewritten as follows:

$$\Delta_{m,s,\lambda,i}(x)=\frac{\Sigma_{l\in\mathbb{Z}}e^{(s-l)x}(g_{lM-i+1}(x)+g_{lM-i+2}(x)+\cdots+g_{lM+M-i}(x))-c_0}{e^x-1}. \tag{4.11}$$

Furthermore, notice that

$$\Delta_{m,s,\lambda,i}(x)-\Delta_{m,s,\lambda,i+1}(x)=\sum_{l\in\mathbb{Z}}e^{(s-l)x}g_{lM-i}(x),\quad\text{for}\quad i=1,\dots,M-1. \tag{4.12}$$

Note that these formulas imply that any irreducible quasifinite highest weight module $L(\widehat{\mathscr{D}^M},\lambda)$ can be obtained as above in an essentially unique way. More precisely, we have the following result.

Theorem 4.3: *Let* $L=L(\widehat{\mathscr{D}^M},\lambda)$ *be an irreducible quasifinite highest weight module with central charge* c *and generating series* $\Delta_i(x)$*, such that*

$$\Delta_i(x)-\Delta_{i+1}(x)=F_i(x) \quad (i=1,...,M-1), \quad \Delta_M(x)=\frac{\Sigma_{i=1}^M F_i(x)-c}{e^x-1},$$

where $F_i(x)$ *are quasipolynomials and* $\Sigma_{i=1}^M F_i(0)=c$ *(see the proof of Theorem 4.1). We write* $F_i(x)=\Sigma_{r\in\mathbb{C}} p_{i,r}(x)e^{rx}$*, where* $p_{i,r}(x)$ *are nonzero polynomials. We decompose the set of exponents* $\{r\}$ *in a disjoint union of congruence classes mod* $\mathbb{Z}$*. Let* $S=\{s,s-k_1,s-k_2,\ldots\}$ *be such a congruence class, let* $m=\max_{r\in S}\deg p_{i,r}$ *and let* $h^{(j)}_{k_rM-i}=(d/dx)^j p_{i,s-k_r}(0)$*, for* $i=1,2,...,M$*. We associate to* S *the* $\mathfrak{g}^{[m]}$*-module* $L^{[m]}(\lambda_S)$ *with central charges,*

$$c_j=\sum_{k_rM-i} h^{(j)}_{k_rM-i},$$

and labels,

$$\lambda^{(j)}_l=\sum_{(k_r+\delta_{i,M})M-i\geq l}(h^{(j)}_{k_rM-i}-c_j\delta_{(k_r+\delta_{i,M})M-i,0}).$$

Then the $\widehat{\mathscr{D}^M}$*-module* L *is isomorphic to the tensor product of all the modules* $L_s^{[m]}(\lambda_S)$*.*

Proof: The tensor product L' of all the modules $L_s^{[m]}(\lambda_S)$ is irreducible due to Theorem 4.2. It remains to show that L' has the same highest weight as L does. This is done by exploiting the observation that $-\Delta_i(x)$ is the value of the highest weight of L' on $e^{xD}e_{ii}$, and using the formulas (4.7)–(4.12). □

V. UNITARY QUASIFINITE HIGHEST WEIGHT MODULES OVER $\widehat{\mathscr{D}^M}$

In this section we shall classify all *unitary* (irreducible) quasifinite highest weight modules over $\widehat{\mathscr{D}^M}$ with respect to the anti-involution ω on $\mathscr{D}^M_{\text{as}}$, defined by

$$\omega(z^k f(D)e_{ij})=z^{-k}\overline{f}(D-k)e_{ji},$$

where for $f(D)=\Sigma_i f_i D^i$ we let $\overline{f}(D)=\Sigma_i \overline{f_i}D^i (f_i\in\mathbb{C})$. The anti-involution ω extends to the whole algebra $\mathscr{D}^{M\mathscr{O}}_{\text{as}}$.

Observe that

$$\Psi(\omega(A),\omega(B))=\omega(\Psi(B,A)), \quad A,B\in\mathscr{D}^{M\mathscr{O}}_{\text{as}}.$$

Therefore, the anti-involution ω of the Lie algebras $\mathscr{D}^M$ and $\mathscr{D}^{M\mathscr{O}}$, lifts to an anti-involution of their central extensions, $\widehat{\mathscr{D}^M}$ and $\widehat{\mathscr{D}^{M\mathscr{O}}}$, such that $\omega(C)=C$, which we again denote by ω.

Remark 5.1: (a) Under the homomorphism $\varphi_s=\varphi_{s,0}:\mathscr{D}^{M\mathscr{O}}_{\text{as}}\to\widetilde{M}(\infty,\mathbb{C})$ we have

$$(\varphi_s(z^k f(D)e_{ij}))^*=\varphi_{\bar{s}}(\omega(z^k f(D)e_{ij})).$$

Here A^* stands for the complex conjugate transpose of the matrix $A\in\widetilde{M}(\infty,\mathbb{C})$.

(b) (see, e.g., Ref. 5). For the anti-involution ω of $\widehat{\mathrm{gl}}(\infty,\mathbb{C})$ defined by $\omega(A)={}^t\overline{A}$, $\omega(1)=1$, a highest weight $\widehat{\mathrm{gl}}(\infty,\mathbb{C})$-module with highest weight λ and central charge c is unitary if and only if the numbers $h_i^{(0)}$ are non-negative integers and $c=\Sigma_i h_i^{(0)}$.

We shall need the following lemma.

Lemma 5.1: Let V *be a unitary quasifinite highest weight module over* $\widehat{\mathscr{D}^M}$ *and let* $\mathbf{b}=(b^1(w),\ldots,b^M(w))$ *be its first* M *characteristic polynomials. Then each* $b^i(w)$ *has only simple real roots.*

Proof: The proof is the same as that of Lemma 5.2 in Ref. 1 with the following changes. Let $S_i=De_{ii}$ for $i=1,2,...,M-1$, $S_M=-\frac{1}{2}(D^2+\Delta_{1,2}-1)e_{MM}$ and $v_i=e_{i,i+1}v_\lambda$, $i=1,2,...,M-1$,

$v_M = z^{-1} e_{M,1} v_\lambda$. Then replace $(s, z^{-1} v_\lambda, b(w))$ in the proof of the Lemma 5.2 in Ref. 1 by $(S_i, v_i, b^i(w))$. □

Now we can formulate the main result of this section.

Theorem 5.1: *(a) A nontrivial quasifinite irreducible highest weight module* $L(\widehat{\mathscr{D}^M}, \lambda)$ *is unitary if and only if the central charge* c *is a positive integer and there exist* $r_1, \ldots, r_c \in \mathbb{R}$ *and a partition* $\{1,...,c\} = I_0 \cup \cdots \cup I_{M-1}$, *such that*

$$\Delta_i(x) = \sum_{j \in I_0 \cup \cdots \cup I_{i-1}} \frac{e^{r_j x} - 1}{e^x - 1} + \sum_{j \in I_i \cup \cdots \cup I_{M-1}} \frac{e^{(r_j+1)x} - 1}{e^x - 1}. \tag{5.1}$$

(b) Any unitary quasifinite $\widehat{\mathscr{D}^M}$*-module* V *is obtained by taking tensor product of* n *unitary irreducible quasifinite highest weight modules over* $\widehat{\mathrm{gl}}(\infty, \mathbb{C})$ *and restricting to* $\widehat{\mathscr{D}^M}$ *via an embedding* $\hat{\varphi}_{\mathbf{s}}^{[0]}$, *where* $\mathbf{s} = (s_1, \ldots, s_n)$ *is a real vector and* $s_i - s_j \notin \mathbb{Z}$ *if* $i \neq j$.

Proof: By Theorem 4.3, being a quasifinite irreducible highest weight $\widehat{\mathscr{D}^M}$-module, V is isomorphic to one of the modules $L_{\mathbf{s}}^{[\mathbf{m}]}(\boldsymbol{\lambda})$. It follows from Lemma 5.1 and Corollary 4.1 that $\mathbf{m} = 0$. Now the claim (b) follows from Remark 5.1. The claim (a) follows from (b) and (4.11)–(4.12). □

Corollary 5.1: Suppose that only finitely many labels $\Delta_{i,n}$ *of* λ *are nonzero. Then the* $\widehat{\mathscr{D}^M}$*-module* $L(\lambda)$ *is unitary if and only if*

$$\Delta_{i,0} \in \mathbb{Z}_+ \quad (i = 1,...,M), \quad 0 \leqslant \Delta_{M,0} \leqslant \Delta_{M-1,0} \leqslant \cdots \leqslant \Delta_{1,0} \leqslant c \in \mathbb{Z}_+$$

and

$$\Delta_{i,n} = 0, \quad \text{for} \quad n > 0, \quad i = 1,...,M.$$

Proof: By the hypothesis, $\Delta_i(x)$ are polynomials for $i = 1,...,M$. Using Theorem 5.1, it follows that $\Delta_i(x) - \Delta_{i+1}(x) = \Sigma_{j \in I_i} e^{r_j x}$ $(i = 1,...,M-1)$ and $\Delta_M(x) = \Sigma_{j=1}^{c} (e^{r_j x} - 1)/(e^x - 1)$ are constant. Therefore $\Delta_{i,n} = 0$ for $n > 0$, $i = 1,...,M$. Moreover, we have that $r_j = 0$ for all $j \in I_i$ $(i = 1,...,M-1)$, $\Delta_{i,0} - \Delta_{i+1,0} = \#I_i$ and $\Delta_{M,0} = \#\{r_j | r_j = 1 \text{ and } j \in I_0\} \leqslant \#I_0$. Observe that $c = \Sigma \#I_i \geqslant \Delta_{1,0} \geqslant \Delta_{2,0} \geqslant \cdots \geqslant \Delta_{M,0} \geqslant 0$. The converse is immediate from Theorem 5.1. □

Now we consider the ''analytic continuation'' of quasifinite irreducible unitary modules of $\widehat{\mathscr{D}^M}$.

Definition 5.1: A $\widehat{\mathscr{D}^M}$-module $L(\widehat{\mathscr{D}^M}, \lambda)$ is called *primitive* if $\Delta_i(x)$ are of the form (5.1) with $r_j \in \mathbb{C}$. (Recall that the numbers r_j are the exponents of this module.) Note that primitive $\widehat{\mathscr{D}^M}$-modules are parametrized by the set of exponents $E = \{r_1, \ldots, r_c\}$ (with multiplicities) and a partition p of this set in M disjoint parts. We therefore will denote this $\widehat{\mathscr{D}^M}$ module by $L(\widehat{\mathscr{D}^M}, E, p)$.

Denote $\widetilde{\mathrm{gl}} = \widetilde{\mathrm{gl}}(\infty, \mathbb{C})$ and $\widehat{\mathrm{gl}} = \widehat{\mathrm{gl}}(\infty, \mathbb{C})$ (see Sec. III). Given $\lambda \in \widetilde{\mathrm{gl}}_0^*$ and $c \in \mathbb{C}$, we have a unique irreducible $\widehat{\mathrm{gl}}$-module $L(\lambda, c) = L(\widehat{\mathrm{gl}}, \lambda, c)$ with $C = cI$, which admits a nonzero vector v_λ, such that

$$E_{ij} v_\lambda = 0, \quad \text{for } i < j,$$

$$E_{ii} v_\lambda = \lambda(E_{ii}) v_\lambda, \quad \text{for } i \in \mathbb{Z}.$$

Define $\Lambda_j \in \widetilde{\mathrm{gl}}_0^*$, $j \in \mathbb{Z}$, as follows:

$$\Lambda_j(E_{ii}) = \begin{cases} 1, & \text{for } 0 < i \leqslant j, \\ -1, & \text{for } j < i \leqslant 0, \\ 0, & \text{otherwise.} \end{cases}$$

Let $H_N = \{(\nu_1, \ldots, \nu_N) \in \mathbb{Z}^N | \nu_1 \geqslant \cdots \geqslant \nu_N\}$, and consider for each $\nu \in H_N$,

$$\Lambda_\nu=\Lambda_{\nu_1}+\cdots+\Lambda_{\nu_N}. \tag{5.2}$$

Due to Remark 5.1(b), a $\widehat{\mathrm{gl}}$-module $L(\widehat{\mathrm{gl}},\lambda,c)$ is a nontrivial unitary module if and only if c is a positive integer and $\lambda=\Lambda_\nu$ for some $\nu\in H_c$.

Let

$$A=-\left(MDI_M+\sum_{i=1}^{M}(i-1)e_{ii}\right)\in\widehat{\mathscr{D}^M}. \tag{5.3}$$

Fix a positive integer N and $\nu\in H_N$. Observe that all the eigenspaces of $\widehat{\varphi_s}(A)$ in $L(\widehat{\mathrm{gl}},\Lambda_\nu,N)$ are finite dimensional, and recall that one has the following ''specialized'' character formula for these modules (cf. Ref. 2):

$$\mathrm{tr}_{L(\widehat{\mathrm{gl}},\Lambda_\nu,N)}\,q^{\widehat{\varphi_s}(A)}=q^{a(\Lambda_\nu)}\prod_{1\leqslant i<j\leqslant N}(1-q^{\nu_i-\nu_j+j-i})/\phi(q)^N, \tag{5.4}$$

where $a(\Lambda_\nu)=\Sigma_k\nu_k((\nu_k+1)/2-Ms)$, and $\phi(q)=\Pi_{j=1}^{\infty}(1-q^j)$ is the Euler product.

The following result is a special case of Theorem 4.2 and formulas (4.11)–(4.12).

Proposition 5.1: Let $\nu\in H_N$, and consider the sets $I_i=\{j\,|\,\nu_j=l_jM-i\}$, $i=0,...,M-1$. Then, when restricted to $\hat{\varphi}_s(\widehat{\mathscr{D}^M})$ the module $L(\Lambda_\nu,N)$ becomes an irreducible quasifinite $\widehat{\mathscr{D}^M}$-module with central charge N and

$$\Delta_i(x)=\sum_{j\in I_0\cup\cdots\cup I_{i-1}}\frac{e^{(s-l_j)x}-1}{e^x-1}+\sum_{j\in I_i\cup\cdots\cup I_{M-1}}\frac{e^{(s-l_j+1)x}-1}{e^x-1}.$$

In other words, it is the primitive $\widehat{\mathscr{D}^M}$-module with exponents $\{s-l_1,\ldots,s-l_N\}$ and partition $\{I_i\}_{i=0}^{M-1}$.

A primitive $\widehat{\mathscr{D}^M}$-module with positive integral central charge N is called *integral* if all exponents $r_1,\ldots,r_N$ are integral.

Recall that primitive $\widehat{\mathscr{D}^M}$-modules are completely determined by the set E of exponents and a partition p of this set in M disjoint parts.

Remark 5.2: There is a correspondence between H_N and the set of integral primitive $\widehat{\mathscr{D}^M}$-modules. More precisely, given $\nu\in H_N$ consider the partition p_ν defined by the sets of congruence classes $\{\nu_j\,|\,\nu_j=l_jM-i\}$, $i=0,...,M-1$, and take $E_\nu=\{-l_1,\ldots,-l_N\}$.

Conversely, given a set of exponents E and a partition p of this set in M disjoint parts $E=\cup_{i=0}^{M-1}I_i$, we can construct $\nu\in H_N$: take the N integers $\{-r_jM-i\,|\,r_j\in I_i\}$ and then order them in decreasing order.

Thus we have the following proposition for primitive $\widehat{\mathscr{D}^M}$-modules.

Proposition 5.2. *Let V be a primitive $\widehat{\mathscr{D}^M}$-module with exponents $\{r_i\}$. Decompose the set $\{r_i\}$ into a disjoint union of congruence classes $\mathrm{mod}\,\mathbb{Z}$, i.e.,*

$$\{r_i\}=S_1\cup\cdots\cup S_m,$$

where $S_i=\{s_i-n_1^{(i)},\ldots,s_i-n_{N_i}^{(i)}\}$, $n_j^{(i)}\in\mathbb{Z}$ and $s_i-s_j\notin\mathbb{Z}$. Let $\mathbf{s}=(s_1,\ldots,s_m)$ and $\Lambda^{(i)}(V)=\Lambda_{n_1^{(i)}}+\ldots+\Lambda_{n_{N_i}^{(i)}}$. Then the $\widehat{\mathscr{D}^M}$-module V is isomorphic to the $\widehat{\mathrm{gl}}^m$-module,

$$\bigotimes_{i=1}^{m}L(\widehat{\mathrm{gl}},\Lambda^{(i)}(V),N_i),$$

by restricting to $\hat{\varphi}_{\mathbf{s}}(\widehat{\mathscr{D}^M})$. In particular, the specialized character $\mathrm{tr}_V\,q^A$ is equal to the product of the corresponding characters of the irreducible $\widehat{\mathrm{gl}}$-modules [given by formula (5.4)]. □

VI. VERTEX ALGEBRA STRUCTURE ON THE VACUUM MODULE OF $\widehat{\mathscr{D}^M}$ AND VERTEX ALGEBRA MODULES

In this section we will construct the structure of a vertex algebra on the vacuum module M_c and hence on the irreducible quotient module L_c over $\widehat{\mathscr{D}^M}$. The original definition of a vertex algebra was given by Ref. 6. We will, however, use a different approach, inspired by Ref. 7 and developed in Refs. 2 and 8.

The following theorem allows one to construct vertex algebras.

Theorem 6.1: (Existence Theorem, Refs. 2, and 8) *Let V be a vector space, let $|0\rangle$ be a vector of V and T an endomorphism of V. Let $\{a^\alpha(x)\}_{\alpha\in I}$ be a collection of fields: $a^\alpha(x)=\Sigma_m a^\alpha_{(m)}x^{-m-1}$, where $a^\alpha_{(m)}\in End\ V$, such that the following properties hold.*

(1) *$a^\alpha(x)$ and $a^\beta(x)$ are mutually local, $(\alpha,\beta\in I):(x-y)^N[a^\alpha(x),a^\beta(y)]=0$, for $N\gg 0$.*

(2) *$a^\alpha(x)|0\rangle|_{x=0}=a^\alpha$ $(\alpha\in I)$, where the a^α are linearly independent.*

(3) *The space V is spanned by the vectors*

$$a^{\alpha_s}_{(-j_s-1)}\cdots a^{\alpha_1}_{(-j_1-1)}|0\rangle,\quad j_1,\ldots,j_s\in\mathbb{Z}_+.$$

(4) $[T,a^\alpha(x)]=\partial_x a^\alpha(x)$ $(\alpha\in I)$, $\quad T|0\rangle=0$.

Then the formula

$$Y(a^{\alpha_s}_{(-j_s-1)}\cdots a^{\alpha_1}_{(-j_1-1)}|0\rangle,x)=(j_1!\cdots j_s!)^{-1}:\partial_x^{j_s}a^{\alpha_s}(x)\cdots\partial_x^{j_2}a^{\alpha_2}(x)\partial_x^{j_1}a^{\alpha_1}(x):$$

(where the normal ordering of more than two fields is from right to left as usual), defines a unique structure of a vertex algebra on V such that $|0\rangle$ is the vacuum vector, T is the infinitesimal translation operator, and $Y(a^\alpha,x)=a^\alpha(x)$, $\alpha\in I$. □

In this case, we shall say that the vertex algebra is *generated* by the fields $a^\alpha(x)$, $\alpha\in I$.

Let

$$J^n_{(k)}=-z^k(\partial_z)^n\quad(k\in\mathbb{Z},n\in\mathbb{Z}_+).$$

These elements form a basis of $\mathscr{D}$. The elements

$$J^n_{(k)}e_{ij}\quad(k\in\mathbb{Z},n\in\mathbb{Z}_+,1\leqslant i,j\leqslant M)$$

form a basis of $\mathscr{D}^M$. Denote by $\mathscr{P}$ the subalgebra of $\mathscr{D}^M$, which consists of ($\mathrm{Mat}_M\mathbb{C}$ valued) operators that can be extended to regular differential operators on $\mathbb{C}$. We have

$$\mathscr{P}=\text{linear span of }\{J^n_{(k)}e_{ij}|k\geqslant 0,1\leqslant i,j\leqslant M\}.$$

The cocycle of our central extension [see (2.4)] restricts to a zero cocycle on $\mathscr{P}$. It follows that $\mathscr{P}$ is a subalgebra of $\widehat{\mathscr{D}^M}$. Let $\hat{\mathscr{P}}=\mathscr{P}\oplus\mathbb{C}C$; note that $[\hat{\mathscr{P}},\hat{\mathscr{P}}]=\mathscr{P}$.

Now, fix $c\in\mathbb{C}$, and denote by M_c the $\widehat{\mathscr{D}^M}$-module $M(\widehat{\mathscr{D}^M},\hat{\mathscr{P}},c)$ defined in Sec. IV, i.e., the $\widehat{\mathscr{D}^M}$-module induced from the one-dimensional module of the subalgebra $\hat{\mathscr{P}}$ defined by $\mathscr{P}\mapsto 0$, $C\mapsto c$. The module M_c is called the *vacuum $\widehat{\mathscr{D}^M}$-module* with central charge c. Denote by L_c the (unique) irreducible quotient of M_c, called the *irreducible vacuum $\widehat{\mathscr{D}^M}$-module*.

Recall that $(J^n_{(k)}e_{ij})|0\rangle=0$ for $n\in\mathbb{Z}_+$, $k\geqslant 0$ and $1\leqslant i,j\leqslant M$, where $|0\rangle$ is the highest weight vector of M_c. Note that the vectors of the form

$$(J^{n_r}_{(-k_r-1)}e_{i_r,j_r})\cdots(J^{n_1}_{(-k_1-1)}e_{i_1,j_1})|0\rangle,$$

where $(n_l,k_l)\in\mathbb{Z}_+^2$ and $1\leqslant i_l,j_l\leqslant M$, span M_c. It follows that the fields

$$J^n_{ij}(x)=\sum_{k\in\mathbb{Z}}J^n_{(k)}e_{ij}x^{-k-1}$$

satisfy conditions (2) and (3) of Theorem 6.1. Condition (1) follows from the operator product expansion [see Ref. 2, (3.9)]:

$$J_{ij}^m(x)J_{kl}^n(y)\sim\delta_{jk}\Bigg\{\sum_{a=1}^{m+n}([n]_aJ_{il}^{m+n-a}(y)-(-1)^a[m]_aJ_{il}^{m+n-a}(x))/(x-y)^{a+1}$$

$$+(-1)^m\delta_{il}m!n!c/(x-y)^{m+n+2}\Bigg\}.$$

It is easy to check condition (4) with $T=-\partial_z I$. Hence, the vacuum $\widehat{\mathscr{D}^M}$-module M_c is a vertex algebra. It follows that any quotient of the $\widehat{\mathscr{D}^M}$-module M_c is a vertex algebra. In particular, the irreducible vacuum module L_c is a vertex algebra, which is denoted by $W_{1+\infty,c}^M$.

A vertex algebra is called *conformal* of rank $c\in\mathbb{C}$ if there exists a vector $\nu\in V$ (called the Virasoro element), such that the corresponding vertex operator $Y(\nu,x)=\Sigma_{n\in\mathbb{Z}}L_nx^{-n-2}$ satisfies the following properties:
(1) $\{L_n\}_{n\in\mathbb{Z}}$ generates a Virasoro algebra with central charge c;
(2) $L_{-1}=T$, and L_0 is diagonalizable in V.

Let $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_M)\in\mathbb{C}^M$ and $\nu(\boldsymbol{\beta})=(J_{(-1)}^1I-\Sigma_{i=1}^M\beta_iJ_{(-2)}^0e_{ii})|0\rangle$. Then it is easy to check that $Y(\nu(\boldsymbol{\beta}),x)=\Sigma_{k\in\mathbb{Z}}L_k(\boldsymbol{\beta})x^{-k-2}$, where

$$L_k(\boldsymbol{\beta})=J_{(k+1)}^1I+\sum_{i=1}^M\beta_i(k+1)J_{(k)}^0e_{ii}.$$

generate a Virasoro algebra $\mathrm{Vir}(\boldsymbol{\beta})$ with central charge

$$c_{\boldsymbol{\beta}}=-\sum_i(12\beta_i^2-12\beta_i+2)c.$$

It is easy to check that condition (2) holds for any element $\nu(\boldsymbol{\beta})$. Thus, we have established the following result.

Theorem 6.2: *For any c and $\boldsymbol{\beta}$, the quadrupole $(M_c,|0\rangle,\nu(\boldsymbol{\beta}),Y(\cdot,x))$ is a conformal vertex algebra of rank $c_{\boldsymbol{\beta}}$ generated by the fields $J_{ij}^n(x)$ $(n=0,1,2,\ldots,1\leqslant i,j\leqslant M)$. The same holds for L_c.* □

Remark 6.1: (a) Define

$$L_k^-(\boldsymbol{\beta})=J_{(k+1)}^1I+\sum_{i=1}^M(k+\beta_i(-k+1))J_{(k)}^0e_{ii}.$$

These elements generate a new family of Virasoro subalgebras $\mathrm{Vir}^-(\boldsymbol{\beta})$ with the same central charge $c_{\boldsymbol{\beta}}$, and the two families $\mathrm{Vir}^-(\boldsymbol{\beta})$ and $\mathrm{Vir}(\boldsymbol{\beta})$ intersect at $\beta_i=\underline{\frac{1}{2}}$. In this case $c_{\mathbf{1}/2}=Mc$.

(b) Observe that $\omega(L_k(\boldsymbol{\beta}))=L_{-k}^-(\boldsymbol{\beta}')$, where $\boldsymbol{\beta}'$ has coordinates β_i. Therefore, $\mathrm{Vir}(\boldsymbol{\beta})$ is ω-stable iff $\beta_i=\frac{1}{2}$ for all i.

(c) The operator A in (5.3) is equal to $ML_0(\boldsymbol{\beta})$ with $\beta_i=(i-1)/M$.

Consider the *Clifford algebra* C_{NM} over $\mathbb{C}$, on generators $\psi_{m[k]}^i$, $\psi_{m[k]}^{*i}$ $(k\in\mathbb{Z},\ i=1,\ldots,N,$ and $m=1,\ldots,M)$, with defining relations

$$[\psi_{m[k]}^{*i},\psi_{n[l]}^j]_+\equiv\psi_{m[k]}^{*i}\psi_{n[l]}^j+\psi_{n[l]}^j\psi_{m[k]}^{*i}=\delta_{ij}\delta_{mn}\delta_{k,-l},$$

and all other anticommutators equal to zero.

There exists a unique irreducible representation of the algebra C_{NM} in a vector space $\mathscr{F}=\mathscr{F}^{NM}$, called the *spin representation*, which admits a nonzero vacuum vector $|0\rangle$, satisfying

$$\psi_{m[k]}^i|0\rangle=0,\quad\text{for}\quad k\geqslant0,\quad\psi_{m[k]}^{*i}|0\rangle=0,\quad\text{for}\quad k>0.$$

In other words, we have NM pairs of free charged fermionic fields:

$$\psi_m^i(x)=\sum_{k\in\mathbb{Z}}\psi_{m[k]}^i x^{-k-1},\quad \psi_m^{*i}(x)=\sum_{k\in\mathbb{Z}}\psi_{m[k]}^{*i}x^{-k},\tag{6.1}$$

i.e., local odds fields with OPE:

$$\psi_m^{*i}(x)\psi_n^j(y)\sim\frac{\delta_{mn}\delta_{ij}}{x-y},\quad \text{and all other OPE}\sim 0.$$

Using the Existence Theorem, the space $\mathscr{F}^{NM}$ of the spin representation has a structure of a vertex algebra with vacuum vector $|0\rangle$ and generated by the fields given in (6.1).

Now, set

$$e_{mn}^{ij}(x)=:\psi_m^i(x)\psi_n^{*j}(x):\equiv\sum_{k\in\mathbb{Z}}e_{mn(k)}^{ij}x^{-k-1}\quad(1\leqslant i,j\leqslant N;1\leqslant m,n\leqslant M)$$

and

$$E^{ij}(x,y)=\sum_{p=1}^{N}:\psi_i^p(x)\psi_j^{*p}(y):\equiv\sum_{k,l\in\mathbb{Z}}E_{kl}^{ij}x^{k-1}y^{-l}\quad(1\leqslant i,j\leqslant M).$$

Then, using Wick's formula, we obtain that the operators $e_{mn(k)}^{ij}$ define a representation in $\mathscr{F}$ of the affine Lie algebra $\widehat{\mathrm{gl}}_{NM}$ of level 1:

$$[e_{mn(p)}^{ij},e_{kl(q)}^{st}]=\delta_{js}\delta_{nk}e_{ml(p+q)}^{it}-\delta_{it}\delta_{ml}e_{kn(p+q)}^{sj}+p\,\delta_{p,-q}\delta_{js}\delta_{nk}\delta_{it}\delta_{ml}.\tag{6.2}$$

Letting charge $\psi_i^p=1$, charge $\psi_i^{*p}=-1$, charge $|0\rangle=0$ defines the *charge decomposition* of $\mathscr{F}$:

$$\mathscr{F}^{NM}=\bigoplus_{m\in\mathbb{Z}}\mathscr{F}_{(m)}^{NM}.$$

Clearly each $\mathscr{F}_{(m)}^{NM}$ is a submodule of the $\widehat{\mathrm{gl}}_{NM}$-module $\mathscr{F}$.

Proposition 6.1: Each $\widehat{\mathrm{gl}}_{NM}$-module $\mathscr{F}_{(m)}^{NM}$ is irreducible.

Proof: This follows from the boson–fermion correspondence (see, e.g., Ref. 8). □

Denote by $V^1(\mathrm{gl}_N)$ the vertex algebra constructed on the irreducible vacuum module of level 1 over the affine Lie algebra $\widehat{\mathrm{gl}}_N$ (see, e.g., Ref. 8). Proposition 6.1 implies the following.

Corollary 6.1: The subspace $\mathscr{F}_{(0)}^{NM}$ is a subalgebra of the vertex algebra $\mathscr{F}^{NM}$ isomorphic to the vertex algebra $V^1(\mathrm{gl}_{NM})$.

It follows from (6.2) that the diagonal embedding Δ of gl_N into gl_{NM} defined by

$$\Delta(e_{ij}):=\sum_{m=1}^{M}e_{mm(0)}^{ij},$$

defines a representation of the general linear Lie algebra gl_N in $\mathscr{F}^{NM}$.

Moreover, the operators E_{kl}^{ij} $(k,l\in\mathbb{Z};\ 1\leqslant i,j\leqslant M)$ form a representation in $\mathscr{F}^{NM}$ of a central extension of the Lie algebra $\mathrm{gl}_M\otimes\mathrm{gl}_\infty$, denoted by $\widehat{\mathrm{gl}_M\otimes\mathrm{gl}_\infty}$, with central charge N:

$$[E_{kl}^{ij},E_{mn}^{st}]=\delta_{js}\delta_{lm}E_{kn}^{it}-\delta_{it}\delta_{kn}E_{ml}^{sj}+N\Phi_1(E_{kl}^{ij},E_{mn}^{st}),\tag{6.3}$$

where the 2-cocycle Φ_1 is defined by

$$\Phi_1(E^{ij}\otimes A,E^{st}\otimes B)=\delta_{it}\delta_{js}\Phi(A,B),$$

and Φ is given by (3.5). Observe that the map $\mathrm{gl}_M\otimes\mathrm{gl}_\infty\to\mathrm{gl}_\infty$, defined by

$$E^{ij}\otimes E_{kl}\mapsto E^{(k-1)M+i,(l-1)M+j},\tag{6.4}$$

is an isomorphism of associative algebras, and this map induces an isomorphism of Lie algebras: $\widehat{\mathrm{gl}_M \otimes \mathrm{gl}_\infty} \simeq \widehat{\mathrm{gl}}_\infty$.

The two actions defined above commute:

$$[e_{ij}, E^{st}_{kl}] = 0, \quad \text{for} \quad \text{all} \quad k,l \in \mathbb{Z}; \quad 1 \leqslant i,j \leqslant N; \quad 1 \leqslant s,t \leqslant M,$$

since

$$[e^{ij}_{mn(0)}, E^{st}_{kl}] = \sum_{q \in \mathbb{Z}} \sum_{p=1}^{N} [:\psi^i_{m[-q]} \psi^{*j}_{n[q]}:, :\psi^p_{s[-k]} \psi^{*p}_{t[l]}:] = \delta_{ns} \psi^i_{m[-k]} \psi^{*j}_{t[l]} - \delta_{mt} \psi^i_{s[-k]} \psi^{*j}_{n[l]}.$$

It is easy to see the following commutation relations:

$$[e_{ij}, \psi^s_{t[l]}] = \delta_{js} \psi^i_{t[l]}, \quad [e_{ij}, \psi^{*s}_{t[l]}] = -\delta_{is} \psi^j_{t[l]}, \tag{6.5}$$

and

$$[E^{ij}_{kl}, \psi^s_{t[-m]}] = \delta_{ml} \delta_{jt} \psi^s_{i[-k]}, \quad [E^{ij}_{kl}, \psi^{*s}_{t[m]}] = -\delta_{it} \delta_{km} \psi^{*s}_{j[l]}. \tag{6.6}$$

Let us define the following subspaces of the algebra C_{NM}:

$$U_{t,m} = \sum_{i=1}^{N} \mathbb{C} \psi^i_{t[-m]}, \quad U^*_{t,m} = \sum_{i=1}^{N} \mathbb{C} \psi^{*i}_{t[m]},$$

$$U = \sum_{m \in \mathbb{Z}} \sum_{t=1}^{M} U_{t,m}, \quad U^* = \sum_{m \in \mathbb{Z}} \sum_{t=1}^{M} U^*_{t,m},$$

$$U_+ = \sum_{m>0} \sum_{t=1}^{M} U_{t,m}, \quad U^*_- = \sum_{m \leqslant 0} \sum_{t=1}^{M} U^*_{t,m}.$$

Using (6.5), we have the following observations.

Remark 6.2: (a) The subspaces $U_{t,m}$ (resp., $U^*_{t,m}$) are gl_N-submodules of C_{NM} isomorphic to the standard gl_N-module $\mathbb{C}^N$ (resp., its dual).

(b) C_{NM} is isomorphic to $\wedge(U^* + U)$ as a gl_N-module.

(c) As a gl_N-module, $\mathscr{F}$ is isomorphic to $\wedge(U^*_- + U_+)$.

Proposition 6.2: (a) The centralizer $(C_{NM})^{\mathrm{gl}_N}$ *is the associative subalgebra generated by the elements*

$$E^{ij}_{kl} = \sum_{p=1}^{N} :\psi^{\ p}_{i[-k]} \psi^{*p}_{j[l]}:,$$

where $k,l \in \mathbb{Z}$ *and* $1 \leqslant i,j \leqslant M$.

(b) *Let* $\mathscr{F} = \Sigma_E \mathscr{F}_E$ *be the decomposition of* $\mathscr{F}$ *in a direct sum of isotypic components with respect to* gl_N. *Then each* $\mathscr{F}_E$ *is irreducible with respect to the Lie algebra* $\mathrm{gl}_N \oplus \mathrm{gl}_\infty$ *[cf. (6.4)].*

Proof: The proof of (a) is the same as that of Proposition 2.1 in Ref. 3. Part (b) follows from (a), Remark 6.2, and the modification of the theory of dual pairs of Howe given in Ref. 3, Theorem 1.1. We are also using the isomorphism (6.4). □

Consider

$$\mathrm{gl}_M \otimes \mathrm{gl}_- = \sum_{k,l \leqslant 0} \sum_{i,j=1}^{M} \mathbb{C} E^{ij}_{kl} \subset C^-_{NM},$$

$$\mathrm{gl}_M \otimes \mathrm{gl}_+ = \sum_{k,l > 0} \sum_{i,j=1}^{M} \mathbb{C} E^{ij}_{kl} \subset C^+_{NM}.$$

Under the isomorphism (6.4), these subalgebras correspond to the subalgebras gl_- and gl_+ of gl_∞ defined in Ref. 3, that is, $\mathrm{gl}_- = \Sigma_{k,l\leqslant 0}\ \mathbb{C}E_{kl}$, and $\mathrm{gl}_+ = \Sigma_{k,l>0}\ \mathbb{C}E_{kl}$. From now on we will identify these subalgebras by isomorphism (6.4).

Recall that the standard gl_+ (resp., gl_-)-module is $\mathbb{C}^{+\infty} = \oplus_{j>0}\mathbb{C}v_j$ (resp., $\mathbb{C}^{-\infty} = \oplus_{j\leqslant 0}\mathbb{C}v_j$), with the usual action $E_{ij}v_k = \delta_{jk}v_i$.

Let $\mathfrak{g} = \mathrm{gl}_N \oplus \mathrm{gl}_- \oplus \mathrm{gl}_+$ be a direct sum of Lie algebras. Using (6.5) and (6.6), we obtain the following.

Proposition 6.3: (a) Both $U_\pm^*$ and U_+ are $\mathfrak{g}$-modules. The $\mathfrak{g}$-module U_+ (resp., U_-^*) is isomorphic to the $\mathrm{gl}_N \oplus \mathrm{gl}_+$-(resp., $\mathrm{gl}_N \oplus \mathrm{gl}_-$)-module $\mathbb{C}^N \otimes \mathbb{C}^{+\infty}$ (resp., $\mathbb{C}^N \otimes \mathbb{C}^{-\infty}$) with the trivial action of gl_- (resp., gl_+).

(b) As a $\mathfrak{g}$-module, $\mathscr{F}^{NM}$ is isomorphic to $\wedge(U_-^*) \otimes \wedge(U_+)$.

Thus, we can use the results of Ref. 3 in order to obtain the decomposition of $\mathscr{F}^{NM}$ as a $\mathrm{gl}_N \oplus \widehat{\mathrm{gl}}_\infty$-module.

Theorem 6.3: (cf. Refs. 2 and 3) *The decomposition of the spin representation $\mathscr{F}^{NM}$ as $\mathrm{gl}_N \oplus \mathrm{gl}_\infty$-module is as follows*:

$$\mathscr{F}^{NM} = \bigoplus_{\nu \in H_N} L_N(\nu) \otimes L(\Lambda_\nu, N), \tag{6.7}$$

where $L_N(\nu)$ is the finite-dimensional irreducible representation of gl_N with highest weight ν [i.e., $\nu(e_{ii}) = \nu_i$], and Λ_ν is given by (5.2).

Corollary 6.2. $(\mathscr{F}^{NM})^{\Delta(\mathrm{gl}_N)} \simeq L(0,N)$ *as $\widehat{\mathrm{gl}}_\infty$-modules. Hence, the vertex algebras $(\mathscr{F}^{NM})^{\Delta(\mathrm{gl}_N)}$, $W^M_{1+\infty,N}$, and $V^1(\mathrm{gl}_{NM})^{\Delta(\mathrm{gl}_N)}$ are isomorphic.*

Recall that the labels of the highest weight $\Lambda = (\Lambda_j)_{j\in\mathbb{Z}}$ of the $\widehat{\mathrm{gl}}_\infty$-module $L(\Lambda, c)$ are the numbers $h_i = \Lambda_i - \Lambda_{i+1} + \delta_{i,0}c$, $i \in \mathbb{Z}$; cf. (4.5). It is clear that a sequence $\{h_i\}_{i\in\mathbb{Z}}$ is a sequence of labels of a highest weight Λ that occurs in the decomposition (6.7), iff

$$h_i \in \mathbb{Z}_+, \quad \text{for all} \quad i \in \mathbb{Z}, \quad \text{and} \quad \sum_i h_i = N. \tag{6.8}$$

Remark 6.3: The embedding φ_s of $\mathscr{D}^M$ into $\mathrm{gl}(\infty,\mathbb{C})$ given in (3.1) corresponds to the embedding,

$$\mathscr{D}^M = \mathrm{gl}_M \otimes \mathscr{D} \hookrightarrow \mathrm{gl}_M \otimes \mathrm{gl}_\infty \xrightarrow{\sim} \mathrm{gl}_\infty.$$

Using Theorem 6.3 and Proposition 5.1, we obtain the following.

Theorem 6.4: *With respect to the commuting pair $(\mathrm{gl}_N, \widehat{\mathscr{D}^M})$, the module $\mathscr{F}^{NM}$ decomposes as follows:*

$$\mathscr{F}^{NM} = \bigoplus_{\nu \in H_N} L_N(\nu) \otimes L(\widehat{\mathscr{D}^M}, E_\nu, p_\nu),$$

where E_ν and p_ν are given in Remark 5.2.

Let $s \in \mathbb{C}$ and denote by Γ_s the additive subgroup of $\mathbb{C}$ generated by 1 and s. Consider a Γ_s-twisting of $\mathscr{F}$ by letting the twist of

$$\psi^{i_1}_{m_1[k_1]} \cdots \psi^{i_u}_{m_u[k_u]} \psi^{*j_1}_{n_1[l_1]} \cdots \psi^{*j_t}_{n_t[l_t]} |0\rangle, \quad \text{equal} \quad s(u-t) + \mathbb{Z}.$$

Now, we define a Γ_s-twisted $\mathscr{F}$-module $\mathscr{F}_s$ as follows (cf. Ref. 3). As a vector space we consider $\mathscr{F}_s = \mathscr{F}$, and let

$$Y^{\mathscr{F}_s}(\psi^p_{i[-1]}|0\rangle, x) := x^{-s}\psi^p_i(x), \quad Y^{\mathscr{F}_s}(\psi^{*p}_{j[0]}|0\rangle, x) := x^s \psi^{*p}_j(x),$$

with $1 \leqslant i,j \leqslant M$. In the same way as in Ref. 3, this can be extended inductively to a Γ_s-twisted module structure on $\mathscr{F}_s$.

Let $W^M_{1+\infty,N}$ be the vertex subalgebra of the vertex algebra $\mathscr{F}$ generated by the fields

$$J^l_{ij}(x)=\sum_{p=1}^{N} :\psi_i^p(x)\partial_x^l\psi_j^{*p}(x):, \quad 1\leqslant i,j\leqslant M, \quad l\in\mathbb{Z}_+;$$

in other words, $W^M_{1+\infty,N}$ is the subspace of $\mathscr{F}$ given by the polynomials of the modes of $J^l_{ij}(x)$ applied to $|0\rangle$.

From Corollary 6.2, we obtain the isomorphism of vertex algebras:

$$W^M_{1+\infty,N}=(\mathscr{F}^{NM})^{\Delta(\mathrm{gl}_N)}.$$

Finally, using Theorem 6.3 and Proposition 5.1, we obtain the following result.

Theorem 6.5: *The twisted $\mathscr{F}$-module $\mathscr{F}_s$, viewed as a (nontwisted) module over the vertex algebra $W^M_{1+\infty,N}$, decomposes in a direct sum (with multiplicities) of all primitive modules for which the set of exponents lies in $s+\mathbb{Z}$. In particular, all these primitive $\widehat{\mathscr{D}^M}$-modules are modules over the vertex algebra $W^M_{1+\infty,N}$.*

Corollary 6.3: Any primitive $\widehat{\mathscr{D}^M}$-module extends to a (irreducible) module over the vertex algebra $W^M_{1+\infty,N}$.

*Conjecture 6.1: Let N be a positive integer. Then all irreducible modules over the vertex algebra $W^M_{1+\infty,N}$ are primitive $\widehat{\mathscr{D}^M}$-*modules.

For $M=1$ this has been proved in Ref. 2 using the representation theory of $W(\mathrm{gl}_N)=V^1(\mathrm{gl}_N)^{\mathrm{gl}_N}$. In order to prove Conjecture 6.1 for $M>1$, one needs to develop a representation theory of the vertex algebra $V^1(\mathrm{gl}_{NM})^{\Delta(\mathrm{gl}_N)}$.

ACKNOWLEDGMENT

The third author was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

[1] V. Kac and A. Radul, ''Quasifinite highest weight modules over the Lie algebra of differential operators on the circle,'' Commun. Math. Phys. **157**, 429–457 (1993).
[2] E. Frenkel, V. Kac, A. Radul, and W. Wang, ''$W_{1+\infty}$ and $W(\mathrm{gl}_N)$ with central charge N,'' Commun. Math. Phys. **170**, 337–357 (1995).
[3] V. Kac and A. Radul, ''Representation theory of the vertex algebra $W_{1+\infty}$,'' Transformation Groups **1**, 41–70 (1996).
[4] V. Kac and D. H. Peterson, ''Spin and wedge representations of infinite-dimensional Lie algebras and groups,'' Proc. Natl. Acad. Sci. USA **78**, 3308–3312 (1981).
[5] V. Kac, *Infinite-Dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).
[6] R. Borcherds, ''Vertex algebras, Kac–Moody algebras, and the Monster,'' Proc. Natl. Acad. Sci. USA **83**, 3068–3071 (1986).
[7] P. Goddard, ''Meromorphic conformal field theory,'' in *Infinite-Dimensional Lie Algebras and Groups*, Advanced Series in Mathematical Physics, edited by V. Kac (World Scientific, Singapore, 1989), Vol. 7, pp. 556–587.
[8] V. Kac, *Vertex Algebras for Beginners* (American Mathematical Society, Providence, 1996).

Algebraic structures and invariant manifolds of differential systems

A. Figueiredo[a] and T. M. Rocha Filho[b]
Departamento de Física, Universidade de Brasília, 70.910-900 Brasília-DF, Brazil

L. Brenig[c]
Service de Physique Statistique, Université Libre de Bruxelles, CP-231 1050 Brussels, Belgium

(Received 13 August 1997; accepted for publication 17 December 1997)

Algebraic tools are applied to find integrability properties of ODEs. Bilinear nonassociative algebras are associated to a large class of polynomial and nonpolynomial systems of differential equations, since all equations in this class are related to a canonical quadratic differential system: the Lotka–Volterra system. These algebras are classified up to dimension 3 and examples for dimension 4 and 5 are given. Their subalgebras are associated to nonlinear invariant manifolds in the phase space. These manifolds are calculated explicitly. More general algebraic invariant surfaces are also obtained by combining a theorem of Walcher and the Lotka–Volterra canonical form. Applications are given for Lorenz model, Lotka, May–Leonard, and Rikitake systems. © *1998 American Institute of Physics.* [S0022-2488(98)01905-7]

I. INTRODUCTION

Given a dynamical system it is relevant to determine if the system is integrable. This can be achieved by obtaining constants of motion. If the system is not integrable, the existence of a constant of motion gives substantial information about its solutions. If the system depends on parameters, it is useful to obtain the parameter values for which such constants of motion exist. Even though this is a difficult problem some methods offer a partial answer. We can cite the Painlevé analysis,[1–4] the Lie symmetry method,[5–8] and the compatibility analysis.[9]

In this paper we present a new approach based on an algebraic theory of differential equations initiated by L. Markus[10] and allow for a systematic study of quasi-polynomial dynamical systems. The fundamental contribution of Markus has been the discovery of a link between systems of differential equations with homogeneous quadratic polynomial vector fields and nonassociative commutative algebras.[11] To such a differential system $\dot{x}=p(x)$ defined on an *n-dimensional* vector space V one can associate an algebra with a bilinear multiplication operation as defined below.

Markus showed that two quadratic differential systems are linearly equivalent if and only if their respective associated algebras are isomorphic. He also established a direct connection between geometrical objects such as invariant hyperplanes in V and subalgebras belonging to the above described algebras. Thus, certain properties of the flows generated by differential equations can be studied by investigating the structure of the associated algebras.[12–14]

These concepts have been generalized to polynomial differential systems with degrees higher than 2.[12–14] To a differential system with homogeneous polynomial vector field of degree $m>1$ may be associated a nonassociative commutative m-ary algebra. Their subalgebras are also related to invariant hyperplanes in the phase space of the differential system.

Nonhomogeneous polynomial differential systems[13,14] may also be related to such nonassociative algebras by a straightforward embedding. However, these more general nonassociative algebras are less popular than the bilinear algebras since calculations may become very complex when $m>2$.

[a]Electronic mail: annibal@helium.fis.unb.br
[b]Electronic mail: marciano@helium.fis.unb.br
[c]Electronic mail: lbrenig@ulb.ac.be

 © 1998 American Institute of Physics

In this article we present a different generalization of the ideas of Markus. We show that bilinear algebras may be associated to any polynomial (even of degree larger than 2 and nonhomogeneous) differential systems. This property is even deeper since it applies also to a larger class of differential systems including ''monomials'' with noninteger exponents in their vector fields, the so-called quasi-polynomial systems for which the previous algebras do not exist. This association may be established thanks to the fact that any differential system in the quasi-polynomial class of equations may be related to a universal quadratic canonical form, the Lotka–Volterra system.[15–19] This canonical form is also combined with a theorem due to Walcher relating ideals of polynomials and invariant surfaces. These concepts lead to a constructive theory of very general invariant surfaces and invariants for the quasi-polynomial class of ordinary differential equations (ODEs).

The article is divided into four sections plus four appendices. In Sec. II we present the association of bilinear Lotka–Volterra algebras to general quasi-polynomial differential systems. In Sec. III, the general properties of the Lotka-Volterra nonassociative algebras are studied. A general algorithm is constructed for analyzing the subalgebras. It is illustrated by classifying the 2.3-dimensional Lotka–Volterra algebras. The method presented in Sec. III is applied on two examples: May–Leonard and Lorenz. In Sec. IV the conditions for the existence of algebraic invariant surfaces are obtained and applied on four systems: (α,β,δ) Lotka–Volterra, May–Leonard, Lorenz, and Rikitake. New integrability cases are found. The results of this method are given in the four tables of Appendix D.

II. DIFFERENTIAL EQUATIONS AND NONASSOCIATIVE ALGEBRAS

Let us define the class of quasi-polynomial differential equations as follows:

$$\dot{x}_i = x_i \sum_{j=1}^{m} A_{ij} \prod_{k=1}^{n} x_k^{B_{jk}}, \quad i = 1,...,n, \tag{1}$$

which is defined on R^n or C^n and where the dot means time derivative.

In Eq. (1), A and B are real or complex, constant and rectangular matrices. Moreover, the number m is related to the number of monomials appearing in the vector field of Eq. (1). Thus, m is generally different from n. The use of the term ''quasi-monomial'' for $\Pi_{k=1}^{n}(x_k)^{B_{jk}}$ is, of course, more accurate since the exponents B_{jk} may be real or complex. This explains the denomination of quasi-polynomial differential equation. These equations are ubiquitous in physics, chemical physics, and biomathematics.

The class (1) includes all polynomial differential equations. Indeed, the factor x_i in front of the right side in Eq. (1) may disappear by cancellation with some factors included in $\Pi_{k=1}^{n}(x_k)^{B_{jk}}$. The factor x_i is extracted in Eq. (1) in order to yield certain useful properties for the matrices A and B. Moreover, linear terms may be hidden in the expression (1) depending on the values of certain entries of the B matrix. In order to give a more systematic idea of the domain of applications of the equations of type (1), let us remark that many equations with right-hand sides more general than quasi-polynomials [like $\sin(x)$, $\exp(x)$,...] may be brought to form (1) by an appropriate embedding.[20–22]

A first step in establishing a connection between Eq. (1) and bilinear nonassociative commutative algebra (NACA) is based on the following lemma:

Lemma 1: Any differential system of class (1) is associated to a homogeneous Lotka–Volterra system.

Proof: One way to exhibit this property is to consider the time evolution of all the quasi-monomials in Eq. (1). Thus, let us define new variables $U_j = \Pi_{k=1}^{n}(x_k)^{B_{jk}}$, $j = 1,...,m$. The differential equations obeyed by these variables can easily be derived from Eq. (1) and read

$$\dot{U}_i = U_i \sum_{l=1}^{m} M_{jl} U_l, \quad j = 1,...,m, \tag{2}$$

where $M = BA$ (matrix product). The same result has been shown earlier by using quasi-monomials transformations and an appropriate embedding.[15,16]

Thus, Eq. (1) can be recast in the following system:

$$\dot{x}_i = x_i \sum_{j=1}^{n} A_{ij} U_j, \quad \dot{U}_j = U_j \sum_{l=1}^{m} M_{jl} U_l, \quad i = 1,...,n, \quad j = 1,...,m. \tag{3}$$

Since the first equation in this system is linear in x, the essential properties of the solutions of Eq. (1) are already embedded in Eq. (2) which is closed in U.

We want to stress here that Eq. (2) is a homogeneous quadratic equation but not the most general one which would be given by

$$\dot{V}_i = \sum_{j,k=1}^{m} C_{ijk} V_j V_k, \quad i = 1,...,m. \tag{4}$$

In fact, Eq. (2), the so-called homogeneous Lotka–Volterra equation,[23–25] is a particular case of (4) which amounts to choosing the tensor C as follows:

$$C_{ijk} = \tfrac{1}{2}(\delta_{ik} M_{kj} + \delta_{ij} M_{jk}). \tag{5}$$

We are now ready for the second step, i.e., the connection of (1) with a nonassociative commutative algebra (NACA).[26,27] This is made along the definition given by Markus in 1960[10] for homogeneous quadratic differential equations:

Definition 1: To a homogeneous quadratic differential system of type (4), where C_{ijk} is symmetric in j,k, one can associate an algebra based on the vector space R^n or C^n and whose multiplication is defined for a basis $e_1,\ldots,e_m$ as follows:

$$e_i \cdot e_j = \sum_{k=1}^{m} C_{kij} e_k. \tag{6}$$

We may thus apply this definition to Eq. (2) which is homogeneous and quadratic. This leads to the following theorem.

Theorem 1: *Differential equations with quasi-polynomial vector fields (1) are associated to bilinear NACAs, whose structure tensor is given by (5), where $M = BA$.*

We would like to stress at this level that in previous works differential systems with homogeneous polynomial vector fields of degree k were only associated to a k-ary NACA.[12–14] Furthermore, such algebras are not defined whenever noninteger exponents appear in the vector field.

In contrast, we have shown here the following: *any quasi-polynomial differential system is associated to a bilinear NACA.*

To end this section, let us remark that the bilinear NACAs we found have a universal structure based on an object which is a matrix: the Lotka–Volterra interaction matrix M. All differential systems of type (1) with matrices A and B such that the product BA is equal to the same matrix M have the same bilinear algebra. They belong to the same equivalence class.

III. STRUCTURE OF LOTKA–VOLTERRA ALGEBRAS

A. Definitions and theorems

The qualitative behavior of the solutions to the systems belonging to the class (1) will now be analyzed through their associated Lotka–Volterra algebras. Indeed, Markus[10] proved two basic theorems for quadratic differential systems. They are the following:

Theorem 2: *Two quadratic differential systems are equivalent under a nonsingular linear transformations iff their related NACAs are isomorphic.*

Theorem 3: *A quadratic m-dimensional differential system has an invariant r-plane* ($1 \leqslant r \leqslant m$) *through the origin iff the related NACA has an r-dimensional subalgebra.*

These results may be directly applied to the Lotka–Volterra equation associated to a given quasi-polynomial system. We consequently may find invariant r-planes for this Lotka–Volterra

equation. Now, how are these r-planes related to the geometry of the phase portrait of the original equation (1)? To clarify this, let us first write the equation of an r-plane passing through the origin in the m-dimensional phase space of the Lotka–Volterra system (2):

$$\sum_{j=1}^{m} l_{ij} U_j = 0, \quad i = 1,...,m-r.$$

It is an invariant r-plane if the coefficients l_{ij} satisfy the existence conditions of a r-dimensional subalgebra of the Lotka–Volterra NACA. Next, we know that the U_i are quasi-monomials in the variable x_i of the equation (1):

$$U_i = \prod_{k=1}^{n} x_k^{B_{ik}}, \quad i = 1,...,m.$$

We have thus shown the following result: *the invariant surface for the original system (1) associated to the existence of an r-dimensional subalgebra of the Lotka–Volterra NACA is an r-dimensional manifold:*

$$\sum_{j=1}^{m} l_{ij} \prod_{k=1}^{n} x_k^{B_{jk}} = 0, \quad i = 1,...,m-r. \tag{7}$$

The geometric properties of the phase portrait for (1) can thus be classified along the structures of the Lotka–Volterra NACAs. We propose now to carry on this algebraic classification. First of all we must introduce some useful definitions:

Definition 2: Let us define a basis of the Lotka–Volterra algebra—LV-NACA—as the set of linearly independent vectors $e_1,\ldots,e_m$ *for which the multiplication is defined by Eq. (6) and tensor (5). We shall call axial algebras the one-dimensional subalgebras generated by each of the* e_i. $\langle e_i \rangle$.

Definition 3: The algebras $\langle e_{i_1},\ldots,e_{i_p} \rangle$ *will be called faces or facial algebras. Here the indices* $i_1,\ldots,i_p$ *are all different and take the values* $1,...,m$.

Definition 4: We define $v_{(i_1,\ldots,i_p)}$ *as the generator of a one-dimensional subalgebra which belongs to the face* $\langle e_{i_1},\ldots,e_{i_p} \rangle$ *and does not belong to any subface belonging to the face* $\langle e_{i_1},\ldots,e_{i_p} \rangle$.

Since $v_{(i_1,\ldots,i_p)}$ generates a one-dimensional algebra, it must satisfy

$$v_{(i_1,\ldots,i_p)} \cdot u_{(i_1,\ldots,i_p)} = \lambda v_{(i_1,\ldots,i_p)}, \tag{8}$$

where $\lambda \in C$. When $\lambda = 0$, $v_{(i_1,\ldots,i_p)}$ is called a *nilpotent* element; if $\lambda = 1$, $v_{(i_1,\ldots,i_p)}$ is an *idempotent* element.

Definition 5: The $p \times p$ *submatrix of M obtained by erasing the lines and columns of indices not in* $\{i_1,\ldots,i_p\}$ *will be denoted* $M^{(i_1,\ldots,i_p)}$.

Definition 6: The matrix obtained by replacing the column i of a matrix N by a column where all entries are 1 is denoted $N^{(i)}$.

Another useful property concerns the multiplication rule. Let us compute the product of two elements of the algebra

$$w_1 \cdot w_2 = \left(\sum_{i=1}^{m} w_{1i} e_i \right) \cdot \left(\sum_{j=1}^{m} w_{2j} e_j \right).$$

Hence

$$w_1 \cdot w_2 = \sum_{i,j=1}^{m} w_{1i} w_{2j} e_i \cdot e_j$$

and finally

$$w_1 \cdot w_2 = \sum_{k=1}^{m} [w_{1k}(Mw_2)_k + w_{2k}(Mw_1)_k]e_k, \tag{9}$$

where Mw_i means the action of matrix M on vector w_i. We now state two fundamental theorems:

Theorem 4: *Let us consider* $N = M^{(i_1,\ldots,i_p)}$. *Two cases are possible: (a) If* $\det(N) \neq 0$ *and* $\det(N^{(i)}) \neq 0$, *then a one-dimensional subalgebra generated by* $v_{(i_1,\ldots,i_p)}$ *exists. (b) If* $\det(N) = 0$, *then the kernel of N generates a nilpotent subalgebra (called in the following nul-algebra).*

Proof: Let us expand $v_{(i_1,\ldots,i_p)}$ in the basis $e_{i_1},\ldots,e_{i_p}$:

$$v_{(i_1,\ldots,i_p)} = \sum_{k=1}^{p} \gamma_k e_{i_k}, \tag{10}$$

where by definition $\gamma_1,\ldots,\gamma_p$ are all different of zero. We now insert Eq. (10) into Eq. (8), yielding

$$\sum_{k,l=1}^{p} \gamma_l \gamma_k (e_{i_k} \cdot e_{i_l}) = \lambda \sum_{r=1}^{p} \gamma_r e_{i_r}. \tag{11}$$

We use formula (9) for the left-hand side of (11) and obtain

$$(M^{(i_1,\ldots,i_p)}\gamma)_i = \frac{\lambda}{2}, \quad i = 1,\ldots,p, \tag{12}$$

where γ is the vector $(\gamma_1,\ldots,\gamma_p)$. From (12) it is clear that if $\det(M^{(i_1,\ldots,i_p)}) \neq 0$, then necessarily $\lambda \neq 0$ and the vector γ is uniquely determined. The solution of (12) is given by

$$\gamma_i = \frac{\det(N^{(i)})}{\det(N)} \frac{\lambda}{2}. \tag{13}$$

Otherwise, if $\det(M^{(i_1,\ldots,i_p)}) = 0$, the kernel of $M^{(i_1,\ldots,i_p)}$ generates a nilpotent algebra. Some comments on these results are here necessary:

(1) The property of idempotence or nilpotence depends entirely on the rank of the corresponding submatrix $M^{(i_1,\ldots,i_p)}$.
(2) When $M^{(i_1,\ldots,i_p)}$ is not of maximal rank, then its kernel is itself a nul-algebra.
(3) In some exceptional but evident cases the condition $\det(M^{(i_1,\ldots,i_p)}) = 0$ may not necessarily imply $\lambda = 0$.

This result exhausts all the possible structures of the one-dimensional subalgebras of a LV-NACA. The second fundamental theorem is the following:

Theorem 5: *A p-dimensional subalgebra of the LV-NACA is always generated by a set of one-dimensional subalgebras.*

Proof: The proposition is clearly true for Lotka–Volterra algebras of dimension 1. Let us suppose the theorem to be true for a Lotka–Volterra with dimension $m-1$. Now, let us consider a p-dimensional subalgebra $B_{(p)}$ which belongs to an m-dimensional LV-NACA. It is easy to see that the intersection between $B_{(p)}$ and any face in this LV-NACA is a $(p-1)$-dimensional subalgebra belonging to the face. However, any face is a LV-NACA of dimension $m-1$, hence this intersection is generated by $(p-1)$ one-dimensional subalgebras. This shows that there exists a set of one-dimensional subalgebras which generates $B_{(p)}$.

Those two theorems provide an algorithm for constructing the existence conditions of any $(m-r)$-dimensional subalgebra of a LV-NACA of dimension m.

Indeed, let us consider an $(m-r)$-dimensional subalgebra $B_{(m-r)}$ and the facial algebra $\langle e_1,\ldots e_{m-1}\rangle$. Then

$$B_{(m-r)} \cap \langle e_1,\ldots,e_{m-1}\rangle = A_{(m-r-1)},$$

which is an $(m-r-1)$-dimensional subalgebra of the face.

By construction, there always exist $r+1$ axial vectors which are transversal to $A_{(m-r-1)}$. These vectors constitute a subface $\langle e_{i_1},\ldots,e_{i_r},e_m\rangle$ with $i_1,\ldots,i_r\in\{1,...,m-1\}$. Now, the intersection

$$B_{(m-r)}\cap\langle e_{i_1},\ldots,e_{i_r},e_m\rangle$$

is a one-dimensional subalgebra generated by a vector $v_{(j_1,\ldots,j_p)}$ (with $p\leqslant r+1$ and $j_1,\ldots,j_p\in\{i_1,\ldots,i_r,m\}$). Thus

$$B_{(m-r)}=\langle A_{(m-r-1)},v_{(j_1,\ldots,j_p)}\rangle.$$

If $A_{(m-r-1)}$ is known, this yields $B_{(m-r)}$.

We now illustrate this algorithm on 2-dimensional and 3-dimensional algebras.

B. Two-dimensional LV-NACAs

Let us construct the conditions for the existence of one-*dimensional* nonaxial algebras for a two-dimensional LV-NACA. By definition, the generator satisfies

$$v_{(1,2)}\cdot v_{(1,2)}=\lambda v_{(1,2)}.$$

Here $v_{(1,2)}=w_1e_1+w_2e_2$.

Using formula (9) we obtain

$$w_1(Mw)_1=\frac{\lambda}{2}w_1,\quad w_2(Mw)_2=\frac{\lambda}{2}w_2, \tag{14}$$

where M is a 2×2 matrix and w_1 and w_2 are nonvanishing (if this would not be the case, the one-dimensional subalgebra would be axial). Now, there are two possibilities:

(1) $\det(M)\neq0$: In this case the solution of (14) is the generator of a one-dimensional subalgebra:

$$v_{(1,2)}=(M_{12}-M_{22})e_1+(M_{21}-M_{11})e_2, \tag{15}$$

where $(M_{12}-M_{22})(M_{21}-M_{11})\neq0$.

(2) $\det(M)=0$: Here again, there are two possibilities:

(a) $(M_{12}-M_{22})(M_{21}-M_{11})\neq0$. Here, formula (8) is still true but $v_{(1,2)}\cdot v_{(1,2)}=0$.

(b) $M_{12}=M_{22}$ and $M_{21}=M_{11}$. Then w_1 and w_2 are arbitrary and $v_{(1,2)}=w_1e_1+w_2e_2$ is any vector. There is an infinite family of one-dimensional subalgebras.

C. Three-dimensional LV-NACAs

We first analyze the one-dimensional subalgebras. The possible generators are, in this case, e_1, e_2, e_3, $v_{(1,2)}$, $v_{(1,3)}$, $v_{(2,3)}$, and $v_{(1,2,3)}$. The algebras generated respectively by e_1,e_2,e_3 are obvious. As to the generators $v_{(1,2)},v_{(1,3)},v_{(2,3)}$, the analysis is the same as for the two-dimensional LV-NACA with the matrix M successively replaced by $M^{(1,2)}$, $M^{(2,3)}$, and $M^{(1,3)}$. Finally, concerning the one-dimensional algebras generated by $v_{(1,2,3)}=w_1e_1+w_2e_2+w_3e_3$ (with $w_1,w_2,w_3\neq0$), we have

$$Mv_{(1,2,3)}=\frac{\lambda}{2}. \tag{16}$$

Hence, there are again two cases:

(a) $\det(M)\neq0$: thus, $v_{(1,2,3)}=M^{-1}\frac{\lambda}{2}$. This yields a one-dimensional non-nul-algebra.

(b) $\det(M)=0$: in this case $\lambda=0$. If rank $(M)=2$, there exists a one-dimensional nul-algebra, and if rank $(M)=1$, there exists a two-dimensional nul-algebra. This closes the analysis of the one-dimensional subalgebras.

Next, we study the two-dimensional subalgebras of a three-dimensional LV-NACA. Clearly, we may limit the existence analysis to the nonfacial two-dimensional subalgebras. We now apply the second fundamental theorem to this case: let B_2 be a two-dimensional subalgebra. Then, the intersection $B_2 \cap \langle e_1, e_2 \rangle = A_1$ may be generated by three possible generators: e_1, e_2, and $v_{(1,2)}$. For each of these generators we may proceed as follows: we consider the two-dimensional facial subalgebra $\langle e_i, e_3 \rangle$ $(i=1,2)$ which excludes A_1. The last step in building the two-dimensional subalgebras amounts to taking $B_2 \cap \langle e_i, e_3 \rangle$, which yields as possible generators e_3 or $v_{(i,3)}$. Thus, the admitted two-dimensional subalgebras are $B_2 = \langle A_1, e_3 \rangle$ or $B_2 = \langle A_1, v_{(i,3)} \rangle$, $i=1,2$.

Let us now explicitly list these B_2 subalgebras (excluding from it the facial subalgebras) and their respective existence conditions:

$$
\begin{aligned}
&B_2 = \langle e_1, v_{(2,3)} \rangle, \quad M_{21} = M_{31}, \\
&B_2 = \langle e_2, v_{(1,3)} \rangle, \quad M_{12} = M_{32}, \\
&B_2 = \langle v_{(1,2)}, e_3 \rangle, \quad M_{13} = M_{23}, \\
&B_2 = \langle v_{(1,2)}, v_{(2,3)} \rangle, \\
&(M_{12} - M_{22})(M_{23} - M_{33})(M_{31} - M_{11}) + (M_{13} - M_{33})(M_{21} - M_{11})(M_{32} - M_{22}) = 0.
\end{aligned}
\tag{17}
$$

The above considerations yield the existence conditions for all the possible subalgebras of a three-dimensional LV-NACA. It should be stressed that these conditions are all expressed in terms of the matrix $M = BA$ of the quasi-polynomial system (1). We end this chapter with two applications.

D. Applications

We apply the above method to the May–Leonard and Lorenz systems. For both examples the Lotka–Volterra matrix $M = BA$ is easily computed. For the May–Leonard system the M matrix is 4×4, whereas for the Lorenz system M is 5×5. The algebraic structures of both systems are to be obtained as follows.

1. May–Leonard system

The May–Leonard system[28] is given by

$$
\begin{aligned}
\dot{x}_1 &= l_1 x_1 - x_1(x_1 + \alpha x_2 + \beta x_3), \\
\dot{x}_2 &= l_2 x_2 - x_2(\beta x_1 + x_2 + \alpha x_3), \\
\dot{x}_3 &= l_3 x_3 - x_3(\alpha x_1 + \beta x_2 + x_3).
\end{aligned}
\tag{18}
$$

It can be cast in the Lotka–Volterra form using the monomials $U_1 = 1$, $U_2 = x_1$, $U_3 = x_2$, and $U_4 = x_3$. The corresponding M matrix as defined in Eq. (2) is given by

$$
M = \begin{bmatrix} 0 & 0 & 0 & 0 \\ l_1 & -1 & -\alpha & -\beta \\ l_2 & -\beta & -1 & -\alpha \\ l_3 & -\alpha & -\beta & -1 \end{bmatrix}. \tag{19}
$$

The generators of one-dimensional sub-algebras are obtained solving equation (12) and using Theorem 3. The results are summarized in Table I (Appendix A) for the case (1) and (2) (a) in Sec. III B corresponding to face $U_1 = 0$. The degenerate case $(\alpha = \beta = 1)$, which corresponds to case (2)(b) in Sec. III B for the face $U_1 = 0$, has a different structure as shown in Table II (Appendix A).

Two- and three-dimensional subalgebras existence conditions are given in Tables IV and VI (Appendix B) for the general and degenerate cases, respectively. For the degenerate case the two- and three-dimensional subalgebras are given in Table V (Appendix B). We obtained one- and

two-dimensional invariant manifolds for the May–Leonard system for the parameters values given in the tables of Appendix B. This is, to our knowledge, a result not found in the literature.

For the three-dimensional subalgebras the existence conditions reported in tables of Appendix B correspond to similar result by Cairo and Feix[29] who obtained algebraic equations for the components of M. In contrast to the method of Cairo and Feix our approach yields all the solutions of these equations.

2. Lorenz system

The Lorenz system[30,31] reads

$$\begin{aligned} \dot{x}_1 &= \sigma(x_2 - x_1), \\ \dot{x}_2 &= \rho x_1 - x_2 - x_1 x_3, \\ \dot{x}_3 &= -\beta x_3 + x_1 x_2. \end{aligned} \tag{20}$$

The Lotka–Volterra form is obtained using the quasi-monomials $U_1=1$, $U_2=x_1^{-1}x_2$, $U_3=x_1x_2^{-1}$, $U_4=x_1x_2^{-1}x_3$, and $U_5=x_1x_2x_3^{-1}$. Its form is given by Eq. (2) with the following matrix:

$$M=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \sigma-1 & -\sigma & \rho & -1 & 0 \\ 1-\sigma & \sigma & -\rho & 1 & 0 \\ 1-\sigma-\beta & \sigma & -\rho & 1 & 1 \\ \beta-1-\sigma & \sigma & \rho & -1 & -1 \end{bmatrix}. \tag{21}$$

In our analysis the parameters σ, ρ, and β are considered as nonvanishing.

The generator of one-dimensional subalgebras are given in Table III (Appendix A). The higher dimensional subalgebras are given in Tables VII (Appendix C). The subalgebras existence conditions are given in Table VIII (Appendix C).

IV. ALGEBRAIC INVARIANT SURFACES

As seen above, the LV-NACA's subalgebras yield invariant manifolds whose equations may only contain quasi-monomials already appearing in the vector field of Eq. (1). In this section, we briefly present a further extension of the algebraic method in order to study a larger class of invariant manifolds including more general quasi-monomials.

We define the flow of the Lotka–Volterra system (2) by the function $U(U_0,t)$ which satisfies (2) and such that $U(U_0,0)=U_0$. In fact, $U(U_0,t)$ is the solution curve to the system (2) with initial conditions U_0. We say that a surface $S\subset R^m$ is an invariant surface of the Lotka–Volterra system (2) if, for any $U_0\in S$, $U(U_0,t)\in S$ for all t.

Now, we denote the ring of all polynomial functions by $\mathrm{Pol}\,[U]$. We consider a surface $S\subset\mathbb{R}^m$, an ideal $I\subset\mathrm{Pol}\,[U]$, and define two special sets: (a) the ideal $J(S)=\{f\in\mathrm{Pol}\,[U]/f(S)=0\}$ and (b) the surface $S(I)=\{U\in\mathbb{R}^m/\forall f\in I, f(U)=0\}$.

According to the above definitions and using results due to Walcher[14] we can state the following theorem.

Theorem 6: *For any autonomous polynomial dynamical system we have the following properties:*

(a) *If* $(d/dt)(I)\subset I$, *then* $S(I)$ *is an invariant surface.*
(b) *If* $S\subset\mathbb{R}^m$ *is an invariant surface, then* $(d/dt)(J(S))\subset J(S)$.

Here, we remember that

$$\frac{d}{dt}=\sum_{i=1}^{m}\left(U_i\sum_{j=1}^{m}M_{ij}U_j\right)\frac{\partial}{\partial U_i}$$

is a differential operator and $\dot{f}=(d/dt)(f)$ is the derivative of f along the flow of the Lotka–Volterra system (2).

We know that any ideal $I\subset \mathrm{Pol}[U]$ is finitely generated. Hence, if $I=\langle P_1,\ldots P_q\rangle$, then the surface $S(I)$ is defined by the system of q polynomials equations

$$P_i(U_1,\ldots,U_m)=0, \quad i=1,...,q. \tag{22}$$

Here $S(I)$ is a surface of dimension $m-q$. In fact, the polynomials invariant surfaces S given by Eqs. (22) generalize the invariant hyperplanes presented in Sec. IV. Indeed, the invariant hyperplanes correspond to the case where the P_i are linear.

If there exists a polynomial $P(U)$ satisfying the equation $P=\lambda P$, with $\lambda\in \mathrm{Pol}[U]$, then, by Theorem 6, we conclude that the surface, given by $P(U)=0$, is an $(m-1)$-dimensional invariant surface for the system (2). In this case $P(U)$ is called a semi-invariant or a Darboux polynomial[32,33] and we denote the set of all semi-invariants of a given Lotka–Volterra system by $SI[U]$. Moreover, it can be shown that λ (see Ref. 14) is necessarily a linear function.

Let us show now the link between semi-invariants and a particular class of invariants for the system (2). Consider a class of invariants of the form

$$K=\prod_{i=1}^{m} U_i^{\alpha_i} P(U_1,\ldots,U_m), \tag{23}$$

where the α_i are real numbers and $P(U)$ is a polynomial function. Cairo, Feix, and Hua[29,33] have studied invariants of this form. The equation for an invariant is

$$\dot{K}=\sum_{j=1}^{m} \frac{\partial K}{\partial U_j}\,\dot{U}_j=0. \tag{24}$$

Then, by inserting (2) and (23) in (24) we obtain

$$\prod_{i=1}^{m} U_i^{\alpha_i}(\dot{P}-\lambda P)=0, \tag{25}$$

where $\lambda=-\Sigma_{j=1}^{m}(\alpha M)_j U_j$ and $(\alpha M)_j=\Sigma_{k=1}^{m}\alpha_k M_{kj}$. The condition (25) is equivalent to

$$\dot{P}=\lambda P. \tag{26}$$

Thus, P is a semi-invariant.

Let us define the set $L[U]$ of all linear functions on U, and the set $SI[U]$ of all semi-invariants of the system (3). We also define the function $\Lambda: SI[U]\to L[U]$ associating to any $P\in SI[U]$ its corresponding eigenvalue given in Eq. (26), and the mapping $\hat{M}: R^m\to L[U]$ which associates the vector $\alpha\in R^m$ to the linear function $\Sigma_{i=1}^{m}(\alpha M)_i U_i$. With these definitions, we conclude that a function of the type (23) is an invariant for the system (2) if two conditions are satisfied:

(a) There is a semi-invariant for the system.
(b) Image $(\hat{M})\cap$ Image $(\Lambda)\neq\varnothing$.

To end this section, we observe that the invariant surface given by (22) and the invariant of the form (23) in the Lotka–Volterra variables correspond respectively to the following invariant surface and invariant in the original variable x of system (1):

$$P_i\left(\prod_{k=1}^{n} x^{B_{1k}},\ldots,\prod_{k=1}^{n} x^{B_{mk}}\right)=0, \quad i=1,...,q,$$

and

$$K=\prod_{k=1}^{n} x_i^{(\alpha B)_i} P\left(\prod_{k=1}^{n} x^{B_{1k}},\ldots,\prod_{k=1}^{n} x^{B_{mk}}\right).$$

Such invariants are clearly more general than those furnished by the NACA approach.

A. Applications

In this section we apply our method to the well studied (α,β,δ) Lotka–Volterra. May–Leonard, Lorenz, and Rikitake systems. To obtain the solutions for Eq. (25) we used a computer algebra in MAPLE written by the authors.[34] All computations for the semi-invariants are performed to polynomials of degree up to 5. The results are shown in Appendix D. We observe that since any quasi-monomial is a semi-invariant, they are not reported in the tables.

1. (α,β,δ) Lotka–Volterra system

The (α,β,δ) Lotka–Volterra system with vanishing eigenvalues[35] is given by

$$\begin{aligned}\dot{x}_1&=x_1(\delta x_2+x_3),\\ \dot{x}_2&=x_2(x_1+\alpha x_3),\\ \dot{x}_3&=x_3(\beta x_1+x_2).\end{aligned}\tag{27}$$

The monomials are trivially obtained: $U_1=x_1$, $U_2=x_2$, and $U_3=x_3$. The matrix M is given by

$$M=\begin{bmatrix}0 & \delta & 1\\ 1 & 0 & \alpha\\ \beta & 1 & 0\end{bmatrix}.\tag{28}$$

The semi-invariants and the corresponding quasi-polynomials invariants (QP-invariants) obtained for this system are given in Table IX (Appendix D). We obtained more cases than in a similar study performed by Labrunie and Conte,[32] who looked for rational invariants obtained from semi-invariants with the same eigenvalue.

Our analysis yields some interesting results: cases 4 and 10–20 are new integrable cases not found in the literature to the authors' knowledge (an exhaustive study of all known cases can be found in Ref. 35). Cases 3, 5, 8, and 9 are particular cases for more general classes of integrable Lotka–Volterra systems for which only one invariant was known[35] and the invariant obtained in this work is a new one not yet obtained. Case 2 is a known integrable class with a new invariant obtained here. Cases 6 and 7 are particular cases of integrable classes for which our approach does not yield any invariant. Nonetheless we obtained an invariant surface that, together with the known invariant, determines the orbit of the system. All know quasi-polynomial invariants for this system plus some new cases were obtained.

2. May–Leonard system

This system, the corresponding monomials, and the matrix M are given in Sec. III D by Eqs. (18) and (19). The semi-invariants and the corresponding QP-invariants are given in Table IX (Appendix D). The linear semi-invariants correspond to cases already obtained by the three-dimensional subalgebras in Table I (Appendix B).

3. Lorenz system

This system is described in Sec. III D by Eqs. (20) and (21). The results obtained are given in Table XI (Appendix C). Most known integrable cases[31] are obtained here while Labrunie and Conte[32] obtained only the cases ($\beta=2\sigma$) and ($\beta=1$, $\rho=0$).

4. Rikitake system

The Rikitake system[36] is given by

$$\dot{x}_1=-\mu x_1+x_2(x_3+\beta),$$

$$\dot{x}_2=-\mu x_2+x_1(x_3-\beta), \tag{29}$$

$$\dot{x}_3=\alpha-x_1x_2.$$

The monomials are $U_1=1$, $U_2=x_1^{-1}x_2x_3$, $U_3=x_1^{-1}x_2$, $U_4=x_1x_2^{-1}x_3$, $U_5=x_1x_2^{-1}$, $U_6=x_3^{-1}$, and $U_7=x_1x_2x_3^{-1}$. The matrix M is given by

$$M=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -\beta & 1 & -\beta & 1 & -1 \\ 0 & -1 & -\beta & 1 & -\beta & 0 & 0 \\ 0 & 1 & \beta & -1 & \beta & 1 & -1 \\ 0 & 1 & \beta & -1 & \beta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ -2\mu & 1 & \beta & 1 & -\beta & -1 & 1 \end{bmatrix}. \tag{30}$$

Our results are shown in Table XII (Appendix D) and coincide with those obtained by Labrunie and Conte.[32]

V. CONCLUSION

We have shown how some algebraic concepts such as nonassociative algebras and polynomial ideals associated to systems of ODEs generate a systematic theory for finding invariant surfaces and invariants for the quasi-polynomial ODEs systems. These results are both exhaustive and constructive in a given functional class. Our next step will be, using the same tools, the generalization of these results to more general functional classes of invariants and the exploration of nonintegrability conditions.

ACKNOWLEDGMENTS

We are all indebted to Mr. L. Ferier for useful discussions during the preparation of this manuscript. AF wishes to thank CAPES and CNPq for financial support and the Department of Optics, Plasma and Nonlinear Physics of the Free Brussel University for its hospitality. AF also acknowledges the support of the bilateral agreement between the Universidade de Brasilia and the Université Libre de Bruxelles. LB benefited from the support of the working group CATHODE (ESPRIT Program, CEE).

APPENDIX A: ONE-DIMENSIONAL SUBALGEBRAS

TABLE I. May–Leonard generators—general case.

Nilpotents	Belong to two-dim faces $v_{(12)}=[1,l_1,0,0]$, $v_{(13)}=[1,0,l_2,0]$, $v_{(14)}=[1,0,0,l_3]$	
	Belong to three-dim faces $v_{(123)}=[1-\alpha\beta,l_1-\alpha l_2,l_2-\beta l_1,0]$, $v_{(124)}=[1-\alpha\beta,l_1-\beta l_3,0,l_3-\alpha l_1]$ $v_{(134)}=[1-\alpha\beta,0,l_2-\alpha l_3,l_3-\beta l_2]$	
	Nonfacial $v_{(1234)}=[\alpha^3+\beta^3-3\alpha\beta+1,(\alpha\beta-1)l_1+(\alpha-\beta^2)l_2+(\beta-\alpha^2)l_3,$ $(a^2-\beta)l_1+(1-\alpha\beta)l_2+(\beta^2-\alpha)l_3,(\alpha-\beta^2)l_1+(\beta-\alpha^2)l_2+(\alpha\beta-1)l_3]$	
Idempotents	Belong to two-dim faces $v_{(23)}=[0,\alpha-1,\beta-1.0]$ $v_{(24)}=[0,\beta-1,0,\alpha-1]$ $v_{(34)}=[0,0,\alpha-1,\beta-1]$	$v_{(23)}^2=(1-\alpha\beta)v_{(23)}$ $v_{(24)}^2=(1-\alpha\beta)v_{(24)}$ $v_{(34)}^2=(1-\alpha\beta)v_{(34)}$
	Belong to three-dim faces $v_{(234)}=[0,1,1,1]$, $v_{(234)}^2=(\alpha^3+\beta^3-3\alpha\beta+1)v_{(234)}$	

TABLE II. May–Leonard generators—degenerate case (c_1, c_2, c_3, x_1, x_2, y_1, y_2, z_1, and z_2 are nonvanishing real numbers).

Nilpotents	Belong to two-dim faces $v_{(12)}=[1,l_1,0,0]$, $v_{(13)}=[1,0,l_2,0]$, $v_{(14)}=[1,0,0,l_3]$ $v_{(23)}=[0,-1,1,0]$, $v_{(24)}=[0,-1,0,1]$, $v_{(34)}=[0,0,-1,1]$	
	Belong to three-dim faces $v_{(234)}=[0,-c_1,c_1-c_2,c_2]$	
Idempotents	Belong to two-dim faces $v_{(23)}=[0,0,-x_1,x_2]$ $v_{(24)}=[0,-y_1,0,y_2]$ $v_{(34)}=[0,0,z_1,z_2]$	$v_{(23)}^2=-(x_1+x_2)v_{(23)}$ $v_{(24)}^2=-(y_1+y_2)v_{(23)}$ $v_{(34)}^2=-(z_1+z_2)v_{(34)}$

TABLE III. Lorenz generators (w_1, w_2, y_1, y_2, z_1, z_2, c_1, and c_2 are nonvanishing real numbers).

Nilpotents	Belong to two-dim faces $v_{(12)}=[\sigma,\sigma-1,0,0,0]$, $v_{(13)}=[\rho,0,1-\sigma,0,0]$ $v_{(14)}=[1,0,0,\sigma+\beta-1,0]$, $v_{(15)}=[1,0,0,0,\beta-\sigma-1]$ $v_{(23)}=[0,\rho,\sigma,0,0]$, $v_{(24)}=[0,1,0,-\sigma,0]$ $v_{(45)}=[0,0,0,-1,-1]$	
	Belong to three-dim faces $v_{(125)}=[\sigma,\sigma-1,0,0,\sigma(\beta-2)]$, $v_{(135)}=[\rho,0,1-\sigma,0,\rho(\beta-2\sigma)]$ $v_{(235)}=[0,\rho,\sigma,0,2\rho\sigma]$, $v_{(123)}=[w_1\sigma,w_1(\sigma-1)+w_2\rho,w_2\rho,0]$ $v_{(234)}=[0,y_1\rho,y_1\rho+y_2,y_2\rho,0]$	
	Belong to four-dim faces $v_{(1245)}=[1,1,0,-1,\beta]$ $v_{(1235)}=[z_1\rho,z_2\rho,z_1(1-\sigma)+z_2\sigma,0,z_1\rho(\beta-2\sigma)+2z_2\rho\sigma]$	
Idempotents	Belong to two-dim faces $v_{(25)}=[0,-1,0,0,-2\sigma]$ $v_{(35)}=[0,0,-1,0,-2\rho]$ $v_{(34)}=[0,0,c_1,c_2,0]$	$v_{(25)}^2=\sigma v_{(25)}$ $v_{(35)}^2=\rho v_{(35)}$ $v_{(34)}^2=(c_2-\rho c_1)v_{(34)}$
	Belong to three-dim faces $v_{(245)}=[0,-2,0,4\sigma,-4\sigma]$,	$v_{(245)}^2=-2\sigma v_{(245)}$

APPENDIX B: MAY–LEONARD SUBALGEBRAS

TABLE IV. Subalgebra existence conditions—May–Leonard general case. (Terms of the form $\langle e_i, v_{(jk)}\rangle$ on the third column correspond to the existence conditions for these two-dimensional subalgebras.)

Faces	Two-dim subalgebras	Existence conditions
$U_1=0$	$\langle e_2, v_{(34)}\rangle$, $\langle e_3, v_{(24)}\rangle$, $\langle e_4, v_{(23)}\rangle$ $\langle v_{(23)}, v_{(34)}\rangle$	$\alpha=\beta$ $(\alpha+\beta-2)\times(\alpha(\alpha-1)+\beta(\beta-1)-\alpha\beta+1)=0$
$U_2=0$	$\langle e_1, v_{(34)}\rangle$ $\langle e_3, v_{(14)}\rangle$ $\langle e_4, v_{(13)}\rangle$ $\langle v_{(23)}, v_{(34)}\rangle$	$l_2=l_3$ $\beta=0$ $\alpha=0$ $l_2(1-\beta)=l_3(\alpha-1)$
$U_3=0$	$\langle e_1, v_{(24)}\rangle$ $\langle e_2, v_{(14)}\rangle$ $\langle e_4, v_{(12)}\rangle$ $\langle v_{(12)}, v_{(24)}\rangle$	$l_1=l_3$ $\alpha=0$ $\beta=0$ $l_3(1-\beta)=l_1(\alpha-1)$
$U_4=0$	$\langle e_1, v_{(23)}\rangle$ $\langle e_2, v_{(13)}\rangle$ $\langle e_3, v_{(12)}\rangle$ $\langle v_{(12)}, v_{(23)}\rangle$	$l_1=l_2$ $\beta=0$ $\alpha=0$ $l_1(1-\beta)=l_2(\alpha-1)]$
Nonfacial subalgebras	$\langle v_{(12)}, v_{(34)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_3=l_2+l_1(\alpha-\beta)$
	$\langle v_{(13)}, v_{(24)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_3=l_1+l_2(\beta-\alpha)$
	$\langle v_{(23)}, v_{(14)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_2=l_1+l_3(\alpha-\beta)$
	$\langle v_{(23)}, v_{(134)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_1(1-\beta)=l_2(\alpha-1)$
	$\langle v_{(24)}, v_{(134)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_3(1-\beta)=l_1(\alpha-1)$
	$\langle v_{(34)}, v_{(124)}\rangle$	$\alpha(\alpha-1)+\beta(\beta-1)=0$ $l_2(1-\beta)=l_3(\alpha-1)$
	$\langle v_{(12)}, v_{(234)}\rangle$ $\langle v_{(13)}, v_{(234)}\rangle$ $\langle v_{(14)}, v_{(234)}\rangle$ $\langle v_{(123)}, v_{(234)}\rangle$	$l_2=-l_1(\alpha+1),\ l_3=-l_1(\beta+1)$ $l_1=-l_2(\beta+1),\ l_3=-l_2(\alpha+1)$ $l_2=-l_3(\beta+1),\ l_1=-l_3(\alpha+1)$ $l_2(1+\alpha)=l_3(1+\beta)$ $l_3(\alpha+1)(\alpha\beta-1)=(1-2\beta-3\alpha\beta)$
	Three-dim subalgebras	Existence conditions
	$\langle e_1, e_2, v_{(34)}\rangle$ $\langle e_1, e_3, v_{(24)}\rangle$ $\langle e_2, e_3, v_{(14)}\rangle$ $\langle e_1, e_4, v_{(23)}\rangle$ $\langle e_2, e_4, v_{(13)}\rangle$ $\langle e_3, e_4, v_{(12)}\rangle$ $\langle e_1, v_{(23)}, v_{(34)}\rangle$ $\langle e_2, v_{(13)}, v_{(34)}\rangle$ $\langle e_3, v_{(12)}, v_{(24)}\rangle$ $\langle e_4, v_{(12)}, v_{(23)}\rangle$ $\langle v_{(12)}, v_{(23)}, v_{(34)}\rangle$	$\langle e_1, v_{(34)}\rangle$, $\langle e_2, v_{(34)}\rangle$ $\langle e_1, v_{(24)}\rangle$, $\langle e_3, v_{(24)}\rangle$ $\langle e_2, v_{(14)}\rangle$, $\langle e_3, v_{(14)}\rangle$ $\langle e_1, v_{(23)}\rangle$ $\langle e_4, v_{(23)}\rangle$ $\langle e_2, v_{(13)}\rangle$, $\langle e_4, v_{(13)}\rangle$ $\langle e_3, v_{(12)}\rangle$, $\langle e_4, v_{(12)}\rangle$ $\langle e_1, v_{(23)}\rangle$, $\langle e_1, v_{(34)}\rangle$, $\langle v_{(23)}, v_{(34)}\rangle$ $\langle e_2, v_{(13)}\rangle$, $\langle e_2, v_{(34)}\rangle$, $\langle v_{(13)}, v_{(34)}\rangle$ $\langle e_3, v_{(12)}\rangle$, $\langle e_3, v_{(24)}\rangle$, $\langle v_{(12)}, v_{(24)}\rangle$ $\langle e_4, v_{(12)}\rangle$, $\langle e_4, v_{(23)}\rangle$, $\langle v_{(12)}, v_{(23)}\rangle$ $\langle v_{(12)}, v_{(23)}\rangle$, $\langle v_{(23)}, v_{(34)}\rangle$ $2(\alpha+\beta+2)(\alpha^2+\beta^2-\alpha\beta-\alpha)l_1-(\alpha-1)^2(l_3-l_2)=0$

TABLE V. Subalgebra structure—May–Leonard degenerate case.

Faces	Two-dim subalgebras	Multiplicative structure
$U_1=0$	$\langle e_2, v_{(34)}\rangle$ $\langle e_3, v_{(24)}\rangle$ $\langle e_4, v_{(23)}\rangle$ $\langle v_{(23)}, v_{(34)}\rangle$	$e_2\cdot v_{(34)}=-1/2(z_1+z_2)e_2-1/2v_{(34)}$ $e_3\cdot v_{(24)}=-1/2(y_1+y_2)e_3-1/2v_{(24)}$ $e_4\cdot v_{(23)}=-1/2(x_1+x_2)e_4-1/2v_{(23)}$ $v_{(23)}\cdot v_{(34)}=-1/2(z_1+z_2)v_{(23)}-1/2(x_1+x_2)v_{(34)}$
$U_2=0$	$\langle v_{(13)}, v_{(34)}\rangle$ $(z_1=l_2, z_2=-l_3)$	$v_{(13)}\cdot v_{(34)}=1/2(l_3-l_2)v_{(34)}$
$U_3=0$	$\langle v_{(12)}, v_{24}\rangle$ $(y_1=l_1, y_2=-l_3)$	$v_{(12)}\cdot v_{(24)}=1/2(l_3-l_1)v_{(24)}$
$U_4=0$	$\langle v_{(12)}, v_{(23)}\rangle$ $(x_1=l_1, x_2=-l_2)$	$v_{(12)}\cdot v_{(23)}=1/2(l_2-l_1)v_{(23)}$
Nonfacial	$\langle v_{(12)}, v_{(23)}, v_{(34)}\rangle$ $(x_1=l_1, x_2=-l_2)$ $(z_1=l_2, z_2=-l_2)$	$v_{(12)}\cdot v_{(34)}=1/2(l_3-l_2)v_{(23)}+1/2(l_3-l_1)v_{(34)}$

TABLE VI. Subalgebra existence conditions—May–Leonard degenerate case. (Terms of the form $\langle e_i, v_{(jk)}\rangle$ on the third column correspond to the existence conditions for these two-dimensional subalgebras. The * corresponds to the subalgebras in Table V.)

Faces	Two-dim subalgebras	Existence conditions
$U_2=0$	$\langle e_1, v_{(34)}\rangle$	$l_2=l_3$
$U_3=0$	$\langle e_1, v_{(24)}\rangle$	$l_1=l_3$
$U_4=0$	$\langle e_1, v_{(23)}\rangle$	$l_1=l_2$
	$\langle e_1, v_{(234)}\rangle$	$l_1=l_2=l_3$
	$\langle v_{(13)}, v_{(24)}\rangle$ $(y_2=-y_1)$	$l_1=l_3$
	$\langle v_{(12)}, v_{(34)}\rangle$ $(z_2=-z_1)$	$l_2=l_3$
	$\langle v_{(23)}, v_{(14)}\rangle$ $(x_2=-x_1)$	$l_1=l_2$
	$\langle v_{(12)}, v_{(234)}\rangle$	$l_1=l_2=l_3$
Nonfacial	$\langle v_{(14)}, v_{(234)}\rangle$	$l_1=l_2=l_3$
subalgebras	$\langle v_{(13)}, v_{(234)}\rangle$	$l_1=l_2=l_3$
	Three-dim subalgebras	Existence conditions
	$\langle e_1, e_2, v_{(34)}\rangle$	$\langle e_2, v_{(34)}\rangle^*$ $\langle e_1, v_{(34)}\rangle$
	$\langle e_1, e_3, v_{(24)}\rangle$	$\langle e_3, v_{(24)}\rangle^*$ $\langle e_1, v_{(24)}\rangle$
	$\langle e_1, e_4, v_{(23)}\rangle$	$\langle e_4, v_{(23)}\rangle^*$ $\langle e_1, v_{(23)}\rangle$
	$\langle e_1, v_{(23)}, v_{(34)}\rangle$	$\langle v_{(23)}, v_{(34)}\rangle^*$ $\langle e_1, v_{(23)}\rangle$ $\langle e_1, v_{(34)}\rangle$

APPENDIX C: LORENZ SUBALGEBRAS

TABLE VII. Subalgebra structure—Lorenz system.

	Faces	Two-dim subalgebras	
	$U_1=U_5=0$	$\langle v_{(12)}, v_{(23)}\rangle$	
Nil	$U_4=U_5=0$	$\langle v_{(23)}, v_{(34)}\rangle$ $(c_1=1, c_2=\rho)$	
subalgebras	$U_4=0$	$\langle v_{(123)}, v_{(235)}\rangle$ $(w_1=2\rho, w_2=2-\beta)$	
	Nonfacial	$\langle v_{(34)}, v_{(1245)}\rangle$ $(c_1=1, c_2=\rho)$	
	Faces	Two-dim subalgebras	Multiplicative structure
	$U_3=U_4=0$	$\langle e_5, v_{(12)}\rangle$	$e_5\cdot v_{(12)}=1/2\sigma(\beta-2)e_5$
	$U_2=U_5=0$	$\langle v_{(13)}, v_{(34)}\rangle$	$v_{(13)}\cdot v_{(34)}=1/2\rho(2\sigma-\beta)v_{(34)}$
		$[c_1=\sigma-1, c_2=\rho(\beta-\sigma-1)]$	
	$U_2=U_4=0$	$\langle e_5, v_{(13)}\rangle$	$e_5\cdot v_{(13)}=1/2\rho(\beta-2\sigma)e_5$
	$U_1=U_5=0$	$\langle e_5, v_{(34)}\rangle$	$e_2\cdot v_{(34)}=1/2(\rho c_1-c_2)e_2+1/2\sigma v_{(34)}$
Non-nil	$U_1=U_4=0$	$\langle e_2, v_{(35)}\rangle$	$e_2\cdot v_{(35)}=-1/2\rho e_2+1/2\sigma v_{(35)}$
subalgebras		$\langle e_3, v_{(25)}\rangle$	$e_3\cdot v_{(25)}=-1/2\sigma e_3+1/2\rho v_{(25)}$
		$\langle e_5, v_{(23)}\rangle$	$e_5\cdot v_{(23)}=\rho\sigma e_5$
	$U_1=U_3=0$	$\langle e_2, v_{(45)}\rangle$	$e_2\cdot v_{(45)}=-1/2e_2+1/2\sigma v_{(45)}$
		$\langle e_4, v_{(25)}\rangle$	$e_4\cdot v_{(25)}=-3/2\sigma e_4-1/2v_{(25)}$
	$U_1=U_2=0$	$\langle v_{(34)}, v_{(45)}\rangle$ $(c_1=1, c_2=2\rho)$	$v_{(34)}\cdot v_{(45)}=-1/2v_{(34)}-1/2\rho v_{(45)}$
	$U_1=0$	$\langle v_{(34)}, v_{(245)}\rangle$ $(c_1=1, c_2=2\rho)$	$v_{(34)}\cdot v_{(245)}=\sigma v_{(34)}-\rho v_{(245)}$
		$\langle v_{(35)}, v_{(234)}\rangle$ $(y_1=1, y_2=-2\sigma)$	$v_{(35)}\cdot v_{(234)}=\sigma\rho v_{(35)}-1/2\rho v_{(234)}$
	$U_4=0$	$\langle e_5, v_{(123)}\rangle$	$e_5\cdot v_{(123)}=1/2\sigma(x_1(\beta-2)+2x_2\rho)e_5$
		Three-dim subalgebras	Multiplicative structure
	$U_1=0$	$\langle e_3, e_4, v_{(25)}\rangle$	$\langle e_3, v_{(25)}\rangle$ $\langle e_4, v_{(25)}\rangle$
		$\langle e_2, v_{(34)}, v_{(45)}\rangle$ $(c_1=1, c_2=2\rho)$	$\langle e_2, v_{(34)}\rangle$ $\langle e_2, v_{(45)}\rangle$ $\langle v_{(34)}, v_{(45)}\rangle$
	$U_4=0$	$\langle e_5, v_{(12)}, v_{(23)}\rangle$	$\langle e_5, v_{(12)}\rangle$ $\langle e_5, v_{(23)}\rangle$ $\langle v_{(12)}, v_{(23)}\rangle$

TABLE VIII. Subalgebra existence conditions—Lorenz system. (The * correspond to the two-dimensional subalgebras in Table VII.)

Faces	Two-dim subalgebras	Existence conditions
$U_4=U_5=0$	$\langle e_1, v_{(23)}\rangle$	$\sigma+1=0$
$U_3=U_5=0$	$\langle e_1, v_{(24)}\rangle$	$\beta+2\sigma-2=0$
$U_3=U_4=0$	$\langle e_1, v_{(25)}\rangle$	$\beta-2\sigma=0$
	$\langle v_{(12)}, v_{(25)}\rangle$	$\sigma+\beta-3=0$
$U_2=U_4=0$	$\langle e_1, v_{(35)}\rangle$	$\beta-2=0$
	$\langle v_{(13)}, v_{(35)}\rangle$	$3\sigma-\beta-1=0$
$U_2=U_3=0$	$\langle e_1, v_{(45)}\rangle$	$\beta-1=0$
$U_2=0$	$\langle v_{(34)}, v_{(15)}\rangle$ $[c_1=1, c_2=\rho]$	$\sigma+1=0$
$U_3=0$	$\langle e_1, v_{(245)}\rangle$	$\sigma=1/2\beta=1$
	$\langle v_{(14)}, v_{(245)}\rangle$	$\sigma=-1\beta=-2$
$U_4=0$	$\langle e_1, v_{(235)}\rangle$	$\sigma=1\beta=2$
	$\langle v_{(13)}, v_{(235)}\rangle$	$\beta-2\sigma=0$
	Three-dim subalgebras	Existence conditions
$U_3=0$	$\langle e_1, e_2, v_{(45)}\rangle$	$\langle e_2, v_{(45)}\rangle^* \langle e_1, v_{(45)}\rangle$
	$\langle e_1, e_4, v_{(25)}\rangle$	$\langle e_4, v_{(25)}\rangle^* \langle e_1, v_{(25)}\rangle$
$U_4=0$	$\langle e_1, e_2, v_{(35)}\rangle$	$\langle e_2, v_{(35)}\rangle^* \langle e_1, v_{(35)}\rangle$
	$\langle e_1, e_3, v_{(25)}\rangle$	$\langle e_3, v_{(25)}\rangle^* \langle e_1, v_{(25)}\rangle$
	$\langle e_1, e_5, v_{(23)}\rangle$	$\langle e_5, v_{(23)}\rangle^* \langle e_1, v_{(23)}\rangle$
	Two-dim subalgebras	Existence conditions
Nonfacial	$\langle v_{(34)}, v_{(125)}\rangle$ $[c_1=1, c_2=\rho]$	$\sigma\beta-\beta+2=0$
	Four-dim subalgebras	Existence conditions
	$\langle e_1, e_3, e_4, v_{(25)}\rangle$	$\langle e_3, v_{(25)}\rangle^* \langle e_4, v_{(25)}\rangle^* \langle e_1, v_{(25)}\rangle$

APPENDIX D: POLYNOMIAL SEMI-INVARIANTS (SI) AND QUASI-POLYNOMIAL INVARIANTS (QP-INVARIANTS)

TABLE IX. Semi-invariants and invariants of the Lotka system.

No.	Parameter conditions	Semi-invariant	Eigenvalue	QP-invariant
1	$\alpha\beta\delta=-1$	$\alpha\beta U_1+U_2-\alpha U_3$	0	SI
2	$3=1$	αU_3-U_2	U_1	$U_1^{\alpha/1+\alpha\delta}U_2^{-1/1+\alpha\delta}U_3^{-\alpha\delta/1+\alpha\delta}$
3	$\beta=1,\ \delta=2$	$(1+\alpha)(U_1U_3-U_2^2)-U_1U_2$	U_1+2U_2	$U_1^{-\alpha/1+2\alpha}U_2^{1/1+2\alpha}U_3^{-2\ 1+\alpha/1+2\alpha}$
4	$\beta=2,\ \delta=-1+\alpha/\alpha$	$\alpha^2U_3(2U_1-U_3)+U_2(2\alpha U_3-U_2)$	$2U_1$	$U_1^{-2\alpha/1+2\alpha}U_2^{2/1+2\alpha}U_3^{-2(1+\alpha)/1+2\alpha}$
5	$\alpha=1/2,\ \beta=\delta=1$	$-U_2^2-U_3^2+U_2U_3$	$2U_1$	$U_1^{1/3}U_2^{4/3}U_3^{2/3}$
6	$\alpha=1,\ \beta=1/3,\ \delta=-3$	$U_1^2U_2+9U_2^3+6U_1U_2^2-9U_1U_2U_3+27U_2^2U_3+27U_2U_3^2-9U_3^3$	U_1	Not found
7	$\alpha=-1/3,\ \beta=3,\ \delta=1$	$9U_1^2U_3-6U_1U_2^2+27U_2^3-27U_1U_2U_3+27U_2^2U_3+9U_2U_3^2+U_3^3$	$3U_1$	Not found
8	$\beta=1,\ \alpha=-2,\ \delta=-3/2$	$4U_1^2U_2+72U_1^2U_3-144U_1U_2^2+9U_2^3+12U_1U_2^2+144U_1U_2U_3+54U_2^2U_3+108U_2U_3^2+72U_2^3$	U_1	$U_1^{-1/2}U_2^{-1/4}U_3^{-3/4}$
9	$\beta=1,\ \alpha=-2/3,\ \delta=1/2$	$108U_1^2U_2+8U_1^2U_3-16U_1U_2^2+27U_2^3+108U_1U_2^2-104U_1U_2U_3+54U_2^2U_3+36U_2U_3^2+8U_3^3$	U_1	$U_1^{-1/2}U_2^{-3/4}U_3^{-1/4}$
10	$\beta=2,\ \alpha=-3/2,\ \delta=-1/3$	$4U_1^2U_2+72U_1^2U_3-144U_1U_2^2+9U_2^3+12U_1U_2^2+144U_1U_2U_3+54U_2^2U_3+108U_2U_3^2+72U_3^3$	U_1	$U_1^{-1/2}U_2^{-1/4}U_3^{-3/4}$
11	$\beta=1/3,\ \alpha=-4,\ \delta=-3/4$	$16U_1^2U_2+9U_2^3+24U_1U_2^2+96U_1U_2U_3+108U_2^2U_3+432U_2U_3^2+576U_3^3$	U_1	$U_1^{-2}U_2^{-1/2}U_3^{-3/2}$
12	$\beta=1/2,\ \alpha=6,\ \delta=-2/3$	$9U_1^2U_2+432U_1U_3^2+4U_2^3+12U_1U_2^2+432U_2U_3^2-144U_1U_2U_3-72U_2^2U_3-864U_3^2$	U_1	$U_1^{-6}U_2U_3^{-4}$
13	$\beta=2,\ \alpha=-1/6,\ \delta=4$	$4U_1^2U_3-4U_1U_3^2+216U_2^3-54U_1U_2^2$ $U_3^3-36U_1U_2U_3+108U_2^2U_3+18U_2U_3^2$	$2U_1$	$U_1U_2^6U_3^{-4}$
14	$\beta=1/2,\ \alpha=1/4,\ \delta=-6$	$32U_1^2U_2+9U_1U_3^2+1152U_2^3+384U_1U_2^2+216U_2U_3^2-144U_1U_2U3-864U_2^2U_3-18U_3^3$	U_1	$U_1U_2^{-4}U_3^6$
15	$\beta=1/3,\ \alpha=-4/5,\ \delta=-3/4$	$2000U_1^2U_2+1125U_2^3+3000U_1U_2^2+2700U_2^2U_3-7200U_1U_2U_3+2160U_2U_3^2+576U_3^3$	U_1	$U_1^{-2/3}U_2^{-5/6}U_3^{-1/2}$
16	$\beta=1/2,\ \alpha=-3/4,\ \delta=-2/3$	$288U_1^2U_2-27U_1U_3^2+128U_2^3+384U_1U_2^2+54U_3^3-720U_1U_2U_3+288U_2^2U_3+216U_2U_3^2$	U_1	$U_1^{-3/5}U_2^{-4/5}U_3^{-2/5}$
17	$\beta=2,\ \alpha=-3/2,\ \delta=1/6$	$108U_1^2U_3-108U_1U_3^2+8U_2^2-48U_1U_2^2$ $36U_2^2U_3-144U_1U_2U_3+54U_2U_3^2+27U_3^3$	$2U_1$	$U_1^{-6}U_2^{-4}U_3$
18	$\beta=1/2,\ \alpha=-3,\ \delta=-2/3$	$9U_1^2U_2-54U_1U_3^2+4U_2^3+12U_1U_2^2+18U_1U_2U_3+36U_2^2U_3+108U_2U_3^2+108U_3^3$	U_1	$U_1^{-3/2}U_2^{-1/2}U_3^{-1}$
19	$\beta=3,\ \alpha=-4/3,\ \delta=-1/4$	$576U_1^2U_3-384U_1U_3^2+27U_2^3-288U_1U_2U_3+108U_2^2U_3+144U_2U_3^2+64U_3^3$	$3U_1$	$U_1^{-2}U_2^{3/2}U_3^{-1/2}$
20	$\beta=3,\ \alpha=-4/3,\ \delta=-5/4$	$576U_1^2U_3-384U_1U_3^2+27U_2^3+64U_3^2+864U_1U_2U_3+108U_2^2U_3+144U_2U_3^2$	$3U_1$	$U_1^{-2/3}U_2^{-1/2}U_3^{-5/6}$
21	$\alpha=\beta=\delta=-1$	$1/3(U_1^3+U_2^3+U_3^3)+U_1^2U_2+U_1^2U_3+U_2^2U_1+U_2^2U_3+U_3^2U_1+U_3^2U_2$	0	SI
		$U_1U_2U_3$	0	SI

TABLE X. Semi-invariants and invariants of the May–Leonard system.

No.	Parameter conditions	Semi-invariant	Eigenvalue	QP-invariant
1	$\alpha=2-\beta$	$U_2+U_3+U_4$	$l_1U_1-U_2-U_3-U_4$	$(U_2U_3U_4)^{-1/3}$
	$l_1=l_2=l_3$	$l_1U_1-U_2-U_3-U_4$	$-(U_2+U_3+U_4)$	Not found
2	$\alpha=\beta$			
	$l_1=l_2$	U_3-U_2	$l_2U_1-U_2-U_3-\alpha U_4$	Not found
3	$\alpha=\beta=1$	$l_1l_2l_3-l_2l_3U_2-l_1l_3U_3-l_1l_2U_4$	$-U_2-U_3-U_4$	$U_2^aU_3^{[a(l_1-l_3)-l_3]/(l_3-l_2)}\times U_4^{[a(l_1-l_2)-l_2]/(l_2-l_3)}$
		$l_3U_1-U_4$	$-U_4$	Not found
4	$\alpha=\beta=0$	$l_2U_1-U_3$	$-U_3$	Not found
		$l_1U_1-U_2$	$-U_2$	Not found
5	$\alpha=\beta=0$			
	$l_1=-l_2$	$l_2U_1+U_2-U_3$	$-(U_2+U_3)$	$(U_2U_3)^{-1}$
6	$\alpha=\beta=0$	$l_1(-4U_1U_2+U_1U_3+2l_1U_1^2)+$		
	$l_3=-2l_1$	$+2U_2^2-2U_2U_4$	$-(2U_2+U_4)$	$U_2^{-2}U_4^{-1}$
7	$\alpha=\beta=0$			
	$l_1=2l_2$	$-2U_2U_3+l_2U_1U_2+2U_3^2$	$2l_2U_1-U_2-2U_3$	Not found
8	$\alpha=\beta=0$	$l_3(l_3U_1^2-U_1U_4+U_1U_3+U_1U_2)+$		
	$l_1=-l_3l_2=-l_3$	$+U_2U_3-U_3U_4$	$-(U_2+U_3+U_4)$	Not found
9	$\alpha=\beta=1$	$1/2(l_3^2U_1+U_2^2+U_3^2+U_4^2)+$		
	$l_1=l_3l_2=-l_3$	$+l_3U_1U_3-l_3U_1U_2$	$-2(U_2+U_3+U_4)$	$(U_2U_3)^{-1}$

TABLE XI. Semi-invariants and invariants of the Lorenz system.

No.	Parameter conditions	Semi-invariant	Eigenvalue	QP-invariant
1	$\sigma=0$, $\beta=-1$	$2U_1-2\rho U_3+U_4+U_5$	$-\rho U_3+U_4-U_5$	$U_1^aU_2^{(2+c)}\times U_3^cU_4^{(b+1)}U_5^b$
2	$\sigma=0$, $\rho=0$	$(\beta-1)U_1-U_4-U_5$	U_4-U_5	Not found
3	$\beta=1$, $\rho=0$	U_4+U_5	$-\sigma U_1+\sigma U_2+U_4-U_5$	Not found
4	$\beta=2\sigma$	$-2\sigma U_2+U_5$	$(\sigma-1)U_1+\sigma U_2$	Not found
			$+\rho U_3-U_4-U_5$	
5	$\beta=1$, $\sigma=1/2$, $\rho=0$	$-U_2^2+U_4U_5-2U_4U_2$	$-(U_1+U_2)$	Not found
6	$\beta=1$, $\sigma=1$	$-\rho U_3U_5+U_2U_5+U_2U_4$	$-(U_1+U_5)$	Not found
7	$\beta=1$, $\sigma=1/3$	$4U_2^2+8U_1U_2-12\rho U_2U_3$	$-2/3(U_1+U_2)$	Not found
		$+12U_2U_4-9U_4U_5$		
8	$\sigma=1$, $\beta=4$	$4U_2(U_4U_2+4U_1^2+U_4U_5$	$-U_2+\rho U_3-U_4-U_5$	Not found
		$-2U_1U_4)-U_4^2U_5+4\rho U_2U_3(U_4-4U_2)$		

TABLE XII. Semi-invariants and invariants of the Rikitake system.

No.	Parameter conditions	Semi-invariant	Eigenvalue	QP-invariant
1	$\beta=0$	U_2-U_4	$1/49(U_2-U_4+\alpha U_6-U_7)$	Not found
2	$\beta=\mu=0$	U_2-U_4	$1/49(U_2-U_4+\alpha U_6-U_7)$	U_6
		$U_2-2\beta U_3+U_7$	$1/49(-U_2-\beta U_3+U_4-\beta U_5+U_7)$	$U_2^{1/49}U_3^{-2/49}$
3	$\alpha=\mu=0$	$-U_4+2\beta U_5-U_7$	$1/49(-U_2-\beta U_3+U_4-\beta U_5+U_7)$	$U_2^{1/49}$
		$U_2U_7-2\beta U_3U_4-U_4U_7$	0	SI

[1] M. J. Ablowitz, A. Ramani, and H. Segur, ''Nonlinear evolutions equations and ordinary differential equations of Painlevé type,'' Lett. Nuovo Cimento **23**, 333–338 (1978).
[2] M. J. Ablowitz, A. Ramani, and H. Segur, ''A connexion between nonlinear evolution equations and ordinary differential equations of *P*-type I,'' J. Math. Phys. **21**, 715–721 (1980).
[3] A. Ramani, B. Grammaticos, and T. Bountis, ''The Painlevé property and singularity analysis of integrable and non-integrable systems,'' Phys. Rep. **180**, 159–245 (1990).
[4] T. Bountis, A. Ramani, B. Grammaticos, and B. Dorizzi, ''On the complete and partial integrability of non-Hamiltonian systems,'' Physica A **128**, 268–288 (1984).
[5] S. Lie, ''Theorie der Transformationgruppen,'' unter Mitwirkung von F. Engel (Teubner, Leipzig, 1893), Vols. 1–3.
[6] L. E. Dixon, ''Differential equations form the group standpoints,'' Ann. Math. **25**(1), 287–378 (1924).

[7] P. J. Olver, *Applications of Lie groups to differential equations*, Graduate Text in Mathematics 107 (Springer Verlag, New York, 1986).
[8] F. Schwartz, ''Simmetries of differential equations: from Sophus Lie to computer algebra,'' SIAM (Soc. Ind. Appl. Math.) Rev. **30**, 450–481 (1988).
[9] J. M. Strelcyn and S. Wojciechowski, ''A method of finding integrals for 3-dimensional dynamical systems,'' Phys. Lett. A **133**, 207–212 (1988).
[10] L. Markus, ''Quadratic differential equations and non-associative algebras,'' in *Contributions to the Theory of Nonlinear Oscillations*, edited by L. Cesari, J. P. La Salle, and S. Lefshetz (Princeton U.P. Princeton, NJ, 1960), Vol. 5, pp. 185–213.
[11] J. Kaplan and J. Yorke, ''Non associative real algebras and quadratic differential equations,'' Nonlinear Anal. TMA, **3**, 49–51 (1979).
[12] H. Röhrl, ''Algebras and differential equations,'' Nagoya Math. J. **68**, 59–122 (1977).
[13] H. Röhrl, ''A theorem on non-associative algebras and its application to differential equations,'' Manuscr. Math. **21**, 181–187 (1977).
[14] S. Walcher, *Algebras and Differential Equations* (Hadronic, Palm Harbor, 1991).
[15] L. Brenig and A. Goriely, ''Universal canonical forms for the time-continuous dynamical systems,'' Phys. Rev. A **40**(7), 4119–4122 (1989).
[16] L. Brenig and A. Goriely, ''Painlevé Analysis and Normal Forms,'' in London Math. Soc. Lecture Note Series 195, edited by E. Tournier (Cambridge U.P., Cambridge, 1994), pp. 211–237.
[17] B. Hernández-Bermejo and V. Fairen, ''Nonpolynomial vector fields under Lotka-Volterra normal form,'' Phys. Lett. A **206**(1,2), 31–37 (1995).
[18] J. L. Gouzé, ''Transformation of polynomial differential systems in the positive orthant,'' Rapport *INRIA*, Sophia-Antilopis, 06561, Valbonne, France (1990).
[19] E. O. Voit and M. A. Savageau, ''Equivalence between *S*-systems and Volterra systems,'' Math. Biosci. **78**, 47–55 (1986).
[20] B. Hernández-Bermejo and V. Fairén, ''Lotka-Volterra representations of general nonlinear systems,'' Math. Biosci. (in press).
[21] H. E. Kerner, ''Universal formats for nonlinear ordinary differential systems,'' J. Math. Phys. **22**, 1366–1371 (1981).
[22] M. A. Savageau and E. O. Voit, ''Recasting nonlinear differential equations as *S*-systems: a canonical nonlinear form,'' Math. Biosci. **87**, 83–115 (1987).
[23] A. J. Lotka, ''Analytical note on certain rhythmic relations in organic systems,'' Proc. Natl. Acad. Sci. USA **6**, 410 (1920).
[24] V. Volterra, *Leçons sur la théorie mathématique de la lutte pour la vie* (Gauthier Villars, Paris, 1931).
[25] R. D. Jenks, ''Quadratical differential systems for interactive populations models,'' J. Diff. Eqns. **5**, 497–514 (1969).
[26] D. K. Harrison, ''Commutative nonassociative algebras and cubic forms,'' J. Algebra. **32**, 518–528 (1974).
[27] R. D. Schafer, *An Introduction to Nonassociative Algebras* (Academic, New York, 1966).
[28] R. H. May and W. J. Leonard, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **29**, 243–253 (1975).
[29] L. Cairo and M. R. Feix, ''Families of invariants of the motion for the Lotka-Volterra equations: the linear polynomials family,'' J. Math. Phys. **33**, 2440–2455 (1992).
[30] E. N. Lorenz, ''Deterministic nonperiodic flow,'' J. Atmos. Sci. **20**, 130–141 (1963).
[31] T. Sen and M. Tabor, ''Lie symmetries of the Lorenz model,'' Physica D **44**, 313–339 (1990).
[32] S. Labrunie and R. Conte, ''A geometrical method towards first integrals for dynamical systems,'' J. Math. Phys. **37**, 6198–6206 (1996).
[33] D. D. Hua, L. Cairo, and M. R. Feix, ''Time-independent invariants of motion for the quadratic system,'' J. Phys. A **26**, 7097–7114 (1993).
[34] T. M. Rocha Filho, A. Figueiredo, and L. Brenig, ''QPSI—A Maple package for the determination of quasi-polynomial symmetries and invariants'' (to be published).
[35] B. Grammaticos, J. Moulin-Ollagnier, A. Ramani, J. M. Strelcyn, and S. Wojciechowski, ''Integrals of quadratic ordinary differential equations in R^3: the Lotka-Volterra system,'' Physica A **163**, 683–722 (1990).
[36] T. C. Bountis, A. Ramani, B. Grammaticos, and B. Dorizzi, ''On the complete and partial integrability of non-Hamiltonian systems,'' Physica A **128**, 268–288 (1984).

Davey–Stewartson type equations of the second and third order: Derivation and classification

Alejandra Folguera
Department of Theoretical and Applied Mechanics, University of Illinois, Urbana, Illinois 61801

(Received 17 November 1997; accepted for publication 17 December 1997)

The inverse spectral method provides an algorithm for generating and solving integrable equations. Here we apply this method to study a family of nonlinear partial differential equation (PDEs) which are two spatial dimensional generalizations of the nonlinear Schrödinger (NLS), the Korteweg–de Vries (KdV) and the modified Korteweg–de Vries (m-KdV) equations. We find simple transformations which reduce this family to a few canonical systems which include the Davey–Stewartson, the Nizkhik–Veselov–Novikov (NVN), and the modified NVN (mod-NVN) equations, and the equations recently derived by Fokas. We also construct Miura maps between the modified NVN and the NVN equations which are valid without any restrictions in the space of solutions. Finally, we present the bi-Hamiltonian formulation of some of the systems presented here. © *1998 American Institute of Physics.* [S0022-2488(98)02005-2]

I. INTRODUCTION

There exists the misconception that there are fewer nonlinear integrable equations in $2+1$ (two space, one time) dimensions than in $1+1$ (one space, one time). It was noticed in Ref. 1 that the $2+1$ generalization of $1+1$ systems corresponding to the NLS equation yields a family of $2+1$ systems that can be parametrized by two real numbers. As it turns out, this family reduces to a few canonical systems. In this paper we study the general structure of the $2+1$ equations generated by the procedure proposed in Ref. 1. The main contributions of this work are:

(a) we derive the two-parameter family of $2+1$ equations of second and third order whose one-dimensional reduction is the nonlinear Schrödinger equation, the KdV, and the modified KdV equations;
(b) we give the transformations that reduce this family to a few canonical systems;
(c) we give the complete list of all canonical systems;
(d) for the third-order systems, we construct Miura maps; and
(e) we derive the bi-Hamiltonian formulation for the above systems.

The equations considered here support *dromion* solutions. That is, there is a family of traveling wave solutions that is localized in space whose path is determined by the so-called boundary functions [the terms containing U and V's in Eqs. (6) and (11)]. The problem of finding such solutions is considered in Ref. 2 and in a companion paper.[3]

Consider the prototypical example of an integrable equation in $1+1$, the nonlinear Schrödinger equation,

$$iq_t+q_{xx}\pm\tfrac{1}{2}q|q|^2=0, \tag{1}$$

and its natural integrable generalization to $2+1$ dimensions, the Davey–Stewartson (DS) equation,

$$iq_t=q_{\eta\eta}+q_{\xi\xi}+\frac{q}{2}\left[\partial_\xi^{-1}(|q|^2)_\eta+\partial_\eta^{-1}(|q|^2)_\xi\right]. \tag{2}$$

The DS equation was introduced in Ref. 4 and by Ref. 5 in the context of water waves.

 © 1998 American Institute of Physics

For our purposes, a system of nonlinear partial differential equations (PDEs) is said to be integrable if it admits a Lax pair formulation. The Lax pair is completely determined by the Lax equation and the degree of the dispersion relation. In the case of the NLS equation, the Lax equation is

$$L\Psi=\left[\begin{pmatrix}\partial_x+ik & 0\\ 0 & \partial_x-ik\end{pmatrix}+Q\right]\Psi=0,\quad \text{where } Q=\begin{pmatrix}0 & q\\ r & 0\end{pmatrix}, \tag{3}$$

and $r=\pm q$. We postulate a solution of the form $\Psi=\mu\ \exp(ikJ+\Omega(k)t)$, where $J=\mathrm{diag}\{1,-1\}$, $\mu=I+\mathcal{O}(1/k)$, and Ω is called the *dispersion function*. For a dispersion of order two, i.e., $\Omega(k)=k^2$, we denote the second equation in the Lax pair by

$$L_2\Psi=0.$$

For order three, i.e., $\Omega(k)=k^3$, we denote the second equation in the Lax pair by

$$L_3\Psi=0.$$

The compatibility of L and L_2 yields the system of two nonlinear equations for q and r:

$$iq_t+q_{xx}+\tfrac{1}{2}q(qr)=0,\quad ir_t+r_{xx}+\tfrac{1}{2}r(rq)=0.$$

These equations reduce to (1) in the case $q=\pm\overline{r}$. The compatibility of L and L_3 yields the third-order system

$$q_t+q_{xxx}+6qrq_x=0,\quad r_t+r_{xxx}+6qrr_x=0.$$

This system has two possible reductions. If $r=q$, then the system reduces to the modified KdV equation,

$$q_t+q_{xxx}+6q^2q_x=0.$$

If $r=-1$, then the system reduces to the KdV equation,

$$q_t+q_{xxx}-6qq_x=0.$$

The following chart summarizes the list of $1+1$ nonlinear equations stemming from the Lax equation (3), together with $\Omega(k)=k^2$ and $\Omega(k)=k^3$.

2nd Order Equations	3rd Order Equations
Nonlinear Schrödinger	KdV modified KdV

The above scheme generalizes to $2+1$ equations by replacing the ik appearing in (3) by ∂_y. A problem will be said to be of the DS type if it can be obtained as the compatibility condition of a pair of matrix differential operators, one of which is

$$L=\begin{pmatrix}\partial_x+\sigma\partial_y & 0\\ 0 & \partial_x-\sigma\partial_y\end{pmatrix}+\begin{pmatrix}0 & q\\ r & 0\end{pmatrix},\quad \sigma^2=\pm1. \tag{4}$$

As before, the operator which involves the variable t is completely determined once the order of the dispersion function Ω is chosen. In $2+1$, the dispersion relation $\Omega(k)=k^n$ can be generalized by introducing two real parameters α and β as follows:

$$\Omega(k)=k^n\begin{pmatrix}\alpha+\beta & 0\\ 0 & \alpha-\beta\end{pmatrix}. \tag{5}$$

We call DS-type systems of order two and three those equations corresponding to $n=2$ and 3, respectively. Thus the DS-type systems are parametrized by two real parameters α and β. It turns out that simple transformations reduce them to a few canonical systems. We summarize the comprehensive list of all such systems here and refer the reader to Sec. II for the actual equations.

The canonical second-order systems of order three: [i.e., equations corresponding to (4) and (5) with $n=2$]

	$\sigma=1$	$\sigma=i$
$\alpha=0,\ \beta=1$	DST stystem I	DST system II
$\alpha=1,\ \beta=0$	DST system III	DST system II
$\alpha=\beta=1/2$	2nd order F T system I	2nd order F T system II

where DST is DS type and FT is Fokas' type. We note that for $\sigma=i$ the equations for $(\alpha,\beta)=(1,0)$ and $(\alpha,\beta)=(0,1)$ are equivalent.

These systems reduce to a single equation when certain relations are imposed on q and r. In the case of the second-order system, the reduction $r=\epsilon\overline{q}$ ($\epsilon=\pm 1$ and $\overline{q}$ denotes the complex conjugate of q) yields

	$\sigma=1$	$\sigma=i$
$\alpha=0,\ \beta=1$	DS I	DS II
$\alpha=1,\ \beta=0$	DS III	DS II
$\alpha=\beta=1/2$	2nd order Fokas I	2nd order Fokas II

The canonical systems of order three: [i.e., equations corresponding to (4) and (5) with $n=3$]

	$\sigma=1$	$\sigma=i$
$\alpha=1,\ \beta=0$	NVN system I	NVN system II
$\alpha=\beta=1/2$	3rd order F T system I	3rd order F T system II

In the case of third-order systems, there are two different reductions that reduce the systems to a single equation, yielding generalizations to $2+1$ dimensions of the KdV and m-KdV equations. These are

	$q=-r,\ (\sigma=1/\sigma=i)$	$r=-1(\sigma=1/\sigma=i)$
$\alpha=1,\ \beta=0$	modified NVN (I/II)	NVN (I/II)
$\alpha=\beta=1/2$	3rd order Fokas (I/II)-1	3rd order Fokas (I/II)-2

Several of these systems have appeared in the literature previously. The DS I and DS II equations represent the amplitude of water waves in a shallow medium, and were derived in Ref. 5. The DS III equation can be found in Refs. 6–8. The second-order equation that we are calling Fokas' can be found in Ref. 1. The third-order systems are first found in the literature in Ref. 9, where the system with $\sigma=1$ was studied; the case $\sigma=i$ was treated in Ref. 10; Ref. 6 has a brief comment on the third-order case; and for a more organized treatment see Ref. 11. None of these references consider the more general situation arising from introducing the two real parameters α and β in the third-order system.

This paper is organized as follows. In the next section we give the complete list of DS-type systems of the second and third order. In Sec. III we derive the relevant systems and their reductions. The method for solving the equations corresponding to $\sigma=i$ is sketched in the proof of Propositions 5 and 6. Section IV contains the transformations. These include transformations used

to classify the systems in terms of canonical systems, as well as the Miura map between mod-NVN and NVN. In the last section we give a brief description of the bi-Hamiltonian operators for this hierarchy of equations.

II. LIST OF DS-TYPE EQUATIONS OF THE SECOND AND THIRD ORDER

In this section we give a comprehensive list of all the DS-type equations of the second and third order. We give the general systems (that is with the parameters α and β in them) as well as the canonical ones. For each of them we also consider their reductions to one equation.

(1) General DS type systems and their reductions.

The general second-order DST system I ($\sigma=1$)*:*

$$iq_t=(\beta-\alpha)q_{\xi\xi}+(\beta+\alpha)q_{\eta\eta}-\frac{q}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right)$$
$$+q\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right), \tag{6}$$

$$ir_t=(\beta-\alpha)r_{\xi\xi}+(\beta+\alpha)r_{\eta\eta}-\frac{r}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right)$$
$$+r\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right), \tag{7}$$

where $\eta=x-y$, $\xi=x+y$, and $U(\xi,t)$ and $V(\eta,t)$ are arbitrary functions.

General DS equation I: System (6) is reduced to a single equation by the choice $r=\epsilon\bar{q}$ where $\epsilon=\pm1$ and the bar denotes complex conjugation:

$$iq_t=(\beta-\alpha)q_{\xi\xi}+(\beta+\alpha)q_{\eta\eta}-\epsilon\frac{q}{2}\left((\beta-\alpha)\partial_\eta^{-1}(|q|^2)_\xi+(\beta+\alpha)\partial_\xi^{-1}(|q|^2)_\eta\right)$$
$$+\epsilon q\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right). \tag{8}$$

The general second-order DST system II ($\sigma=i$)*:*

$$iq_t=(\beta-\alpha)q_{zz}+(\beta+\alpha)q_{\bar{z}\bar{z}}-\frac{q}{2}\left((\beta-\alpha)\partial_{\bar{z}}^{-1}(qr)_z+(\beta+\alpha)\partial_z^{-1}(qr)_{\bar{z}}\right),$$
$$ir_t=(\beta-\alpha)r_{zz}+(\beta+\alpha)r_{\bar{z}\bar{z}}-\frac{r}{2}\left((\beta-\alpha)\partial_{\bar{z}}^{-1}(qr)_z+(\beta+\alpha)\partial_z^{-1}(qr)_{\bar{z}}\right), \tag{9}$$

where $z=x+iy$ and $z=x-iy$.

General DS equation II: As before, this system reduces to the single equation

$$iq_t=(\beta-\alpha)q_{zz}+(\beta+\alpha)q_{\bar{z}\bar{z}}-\epsilon\frac{q}{2}\left((\beta-\alpha)\partial_{\bar{z}}^{-1}(|q|^2)_z+(\beta+\alpha)\partial_z^{-1}(|q|^2)_{\bar{z}}\right). \tag{10}$$

The general third-order DST system I ($\sigma=1$)*:*

$$q_t=(\beta-\alpha)q_{\xi\xi\xi}+(\alpha+\beta)q_{\eta\eta\eta}-\tfrac{3}{4}(\beta-\alpha)q\partial_\eta^{-1}(rq_\xi)_\xi-\tfrac{3}{4}(\alpha+\beta)q\partial_\xi^{-1}(rq_\eta)_\eta$$
$$-\tfrac{3}{4}(\beta-\alpha)q_\xi\partial_\eta^{-1}(qr)_\xi-\tfrac{3}{4}(\alpha+\beta)q_\eta\partial_\xi^{-1}(qr)_\eta+(\beta-\alpha)q_\xi U_0(\xi,t)$$
$$+(\beta+\alpha)q_\eta V_0(\eta,t)+q\left((\beta-\alpha)U_1(\xi,t)+(\beta+\alpha)V_1(\eta,t)\right), \tag{11}$$

$$
\begin{aligned}
r_t=&(\beta-\alpha)r_{\xi\xi\xi}+(\alpha+\beta)r_{\eta\eta\eta}-\tfrac{3}{4}(\beta-\alpha)r\partial_\eta^{-1}(qr_\xi)_\xi-\tfrac{3}{4}(\alpha+\beta)r\partial_\xi^{-1}(qr_\eta)_\eta\\
&-\tfrac{3}{4}(\beta-\alpha)r_\xi\partial_\eta^{-1}(rq)_\xi-\tfrac{3}{4}(\alpha+\beta)r_\eta\partial_\xi^{-1}(rq)_\eta\\
&+(\beta-\alpha)r_\xi U_0(\xi,t)+(\beta+\alpha)r_\eta V_0(\eta,t)+r((\beta-\alpha)(U_0'-U_1)(\xi,t)\\
&+(\beta+\alpha)(V_0'-V_1)(\eta,t)).
\end{aligned}
$$

There are two reductions of this system.
(i) *General NVN equation I* $(r=-1)$*:*

$$
\begin{aligned}
q_t=&(\beta-\alpha)q_{\xi\xi\xi}+(\beta+\alpha)q_{\eta\eta\eta}+\tfrac{3}{4}(\alpha+\beta)(q\partial_\xi^{-1}q_\eta)_\eta+\tfrac{3}{4}(\beta-\alpha)(q\partial_\eta^{-1}q_\xi)_\xi\\
&+q((\beta-\alpha)U_0'(\xi,t)+(\beta+\alpha)V_0'(\eta,t))+(\alpha-\beta)U_0(\xi,t)q_\xi+(\beta+\alpha)V_0(\eta,t)q_\eta. \qquad (12)
\end{aligned}
$$

(ii) *General modified NVN equation I* $(r=-q)$:

$$
\begin{aligned}
q_t=&(\beta-\alpha)q_{\xi\xi\xi}+(\alpha+\beta)q_{\eta\eta\eta}+\tfrac{3}{8}(\beta-\alpha)q\partial_\eta^{-1}(q^2)_{\xi\xi}+\tfrac{3}{8}(\alpha+\beta)q\partial_\xi^{-1}(q^2)_{\eta\eta}\\
&+\tfrac{3}{4}(\beta-\alpha)q_\xi\partial_\eta^{-1}(q^2)_\xi+\tfrac{3}{4}(\alpha+\beta)q_\eta\partial_\xi^{-1}(q^2)_\eta+\frac{q}{2}((\beta-\alpha)U_0'(\xi,t)\\
&+(\beta+\alpha)V_0'(\eta,t))+(\beta-\alpha)U_0(\xi,t)q_\xi+(\beta+\alpha)V_0(\eta,t)q_\eta. \qquad (13)
\end{aligned}
$$

(iii) *General third-order DST system II* $(\sigma=i)$*:*

$$
\begin{aligned}
q_t=&(\beta-\alpha)q_{zzz}+(\alpha+\beta)q_{\bar z\bar z\bar z}-\tfrac{3}{4}(\alpha+\beta)q\partial_z^{-1}(rq_{\bar z})_{\bar z}-\tfrac{3}{4}(\beta-\alpha)q\partial_{\bar z}^{-1}(rq_z)_z\\
&-\tfrac{3}{4}(\alpha+\beta)q_{\bar z}\partial_z^{-1}(qr)_{\bar z}-\tfrac{3}{4}(\beta-\alpha)q_z\partial_{\bar z}^{-1}(qr)_z,\\
r_t=&(\beta-\alpha)r_{zzz}+(\alpha+\beta)r_{\bar z\bar z\bar z}-\tfrac{3}{4}(\alpha+\beta)r\partial_z^{-1}(qr_{\bar z})_{\bar z}-\tfrac{3}{4}(\beta-\alpha)r\partial_{\bar z}^{-1}(qr_z)_z\\
&-\tfrac{3}{4}(\alpha+\beta)r_{\bar z}\partial_z^{-1}(rq)_{\bar z}-\tfrac{3}{4}(\beta-\alpha)r_z\partial_{\bar z}^{-1}(rq)_z.
\end{aligned} \qquad (14)
$$

There exist two reductions of this system.
(i) *General NVN equation II* $(r=-1)$*:*

$$
q_t=(\beta-\alpha)q_{zzz}+(\beta+\alpha)q_{\bar z\bar z\bar z}+\tfrac{3}{4}(\alpha+\beta)(q\partial_z^{-1}q_{\bar z})_{\bar z}+\tfrac{3}{4}(\beta-\alpha)(q\partial_{\bar z}^{-1}q_z)_z. \qquad (15)
$$

(ii) *General modified NVN-II* $(r=-q)$*:*

$$
\begin{aligned}
q_t=&(\beta-\alpha)q_{zzz}+(\alpha+\beta)q_{\bar z\bar z\bar z}+\tfrac{3}{4}(\alpha+\beta)q\partial_z^{-1}(qq_{\bar z})+\tfrac{3}{4}(\beta-\alpha)q\partial_{\bar z}^{-1}(qq_z)_z\\
&+\tfrac{3}{4}(\alpha+\beta)q_{\bar z}\partial_z^{-1}(q^2)_{\bar z}+\tfrac{3}{4}(\beta-\alpha)q_z\partial_{\bar z}^{-1}(q^2)_z. \qquad (16)
\end{aligned}
$$

(2) Canonical DS type systems and their reductions.

A. Canonical second-order systems and their reductions

DST I system $(\alpha=0,\ \beta=1)$*:*

$$
\begin{aligned}
&iq_t+q_{\xi\xi}+q_{\eta\eta}-\frac{q}{2}(\partial_\eta^{-1}(qr)_\xi+\partial_\xi^{-1}(qr)_\eta)+q(U(\xi,t)+V(\eta,t)),\\
&ir_t+r_{\xi\xi}+r_{\eta\eta}-\frac{r}{2}(\partial_\eta^{-1}(qr)_\xi+\partial_\xi^{-1}(qr)_\eta)+r(U(\xi,t)+V(\eta,t)),
\end{aligned} \qquad (17)
$$

and its reduction, *DS I*,

$$iq_t = q_{\xi\xi} + q_{\eta\eta} - \epsilon \frac{q}{2} (\partial_\eta^{-1}(|q|^2)_\xi + \partial_\xi^{-1}(|q|^2)_\eta) + q(U(\xi,t) + V(\eta,t)). \tag{18}$$

DST III system ($\alpha = 1$, $\beta = 0$)*:*

$$\begin{aligned} iq_t &= q_{\xi\xi} - q_{\eta\eta} - \frac{q}{2} (\partial_\eta^{-1}(qr)_\xi - \partial_\xi^{-1}(qr)_\eta) + q(U(\xi,t) - V(\eta,t)), \\ ir_t &= r_{\xi\xi} - r_{\eta\eta} - \frac{r}{2} (\partial_\eta^{-1}(qr)_\xi - \partial_\xi^{-1}(qr)_\eta) + r(U(\xi,t) - V(\eta,t)), \end{aligned} \tag{19}$$

and its reduction, *DS III*,

$$iq_t = q_{\xi\xi} - q_{\eta\eta} - \epsilon \frac{q}{2} (\partial_\eta^{-1}(|q|^2)_\xi - \partial_\xi^{-1}(|q|^2)_\eta) + q(U(\xi,t) + V(\eta,t)). \tag{20}$$

Second-order Fokas type I system ($\alpha = \beta = \frac{1}{2}$)*:*

$$iq_t = q_{\xi\xi} - \frac{q}{2} \partial_\eta^{-1}(qr)_\xi + qU(\xi,t), \quad ir_t = r_{\xi\xi} - \frac{r}{2} \partial_\eta^{-1}(qr)_\xi + rU(\xi,t), \tag{21}$$

and its reduction *second-order Fokas I equation*,

$$iq_t = q_{\xi\xi} - \frac{q}{2} \partial_\eta^{-1}(|q|^2)_\xi + \epsilon q U(\xi,t). \tag{22}$$

DST II system:

$$\begin{aligned} iq_t &= q_{zz} + q_{\bar{z}\bar{z}} - \frac{q}{2} (\partial_z^{-1}(qr)_z + \partial_{\bar{z}}^{-1}(qr)_{\bar{z}}), \\ ir_t &= r_{zz} + r_{\bar{z}\bar{z}} - \frac{r}{2} (\partial_z^{-1}(qr)_z + \partial_{\bar{z}}^{-1}(qr)_{\bar{z}}), \end{aligned} \tag{23}$$

and its reduction, *DS II*,

$$iq_t = q_{zz} + q_{\bar{z}\bar{z}} - \epsilon \frac{q}{2} (\partial_z^{-1}(|q|^2)_z + \partial_{\bar{z}}^{-1}(|q|^2)_{\bar{z}}). \tag{24}$$

Fokas type system II:

$$iq_t = q_{zz} + \frac{q}{2} \partial_z^{-1}(qr)_z, \quad ir_t = r_{zz} + \frac{r}{2} \partial_z^{-1}(qr)_z, \tag{25}$$

and its reduction, *Fokas' simplest II*,

$$iq_t = q_{zz} + \epsilon \frac{q}{2} \partial_z^{-1}(|q|^2)_z. \tag{26}$$

B. Canonical third-order systems and their reductions

Third-order DST system I:

$$\begin{aligned} q_t = {} & q_{\xi\xi\xi} + q_{\eta\eta\eta} - \tfrac{3}{4} q \partial_\eta^{-1}(rq_\xi)_\xi - \tfrac{3}{4} q \partial_\xi^{-1}(rq_\eta)_\eta - \tfrac{3}{4} q_\xi \partial_\eta^{-1}(qr)_\xi \\ & - \tfrac{3}{4} q_\eta \partial_\xi^{-1}(qr)_\eta + q_\xi U_0(\xi,t) + q_\eta V_0(\eta,t) + q(U_1(\xi,t) + V_1(\eta,t)), \end{aligned} \tag{27}$$

$$r_t = r_{\xi\xi\xi} + r_{\eta\eta\eta} - \tfrac{3}{4} r \partial_\eta^{-1}(q r_\xi)_\xi - \tfrac{3}{4} r \partial_\xi^{-1}(q r_\eta)_\eta - \tfrac{3}{4} r_\xi \partial_\eta^{-1}(rq)_\xi - \tfrac{3}{4} r_\eta \partial_\xi^{-1}(rq)_\eta + r_\xi U_0(\xi,t) + r_\eta V_0(\eta,t) + r(U_2(\xi,t) + V_2(\eta,t)).$$

There exist two reductions for this system: *NVN-I equation* $(r=-1)$,

$$q_t = q_{\xi\xi\xi} + q_{\eta\eta\eta} + \tfrac{3}{4}(q\partial_\xi^{-1} q_\eta)_\eta + \tfrac{3}{4}(q\partial_\eta^{-1} q_\xi)_\xi + q \times (U_0'(\xi,t) + V_0'(\eta,t)) + U_0(\xi,t) q_\xi + V_0(\eta,t) q_\eta, \tag{28}$$

and *modified NVN-I equation* $(r=-q)$,

$$q_t = q_{\xi\xi\xi} + q_{\eta\eta\eta} + \frac{3}{8} q \partial_\eta^{-1}(q^2)_{\xi\xi} + \frac{3}{8} q \partial_\xi^{-1}(q^2)_{\eta\eta} + \frac{3}{4} q_\xi \partial_\eta^{-1}(q^2)_\xi + \frac{3}{4} q_\eta \partial_\xi^{-1}(q^2)_\eta + \frac{q}{2}(U_0'(\xi,t) + V_0'(\eta,t)) + U_0(\xi,t) q_\xi + V_0(\eta,t) q_\eta. \tag{29}$$

Third Order Fokas system I:

$$\begin{aligned} q_t &= q_{\xi\xi\xi} - \tfrac{3}{4} q \partial_\eta^{-1}(r q_\xi)_\xi - \tfrac{3}{4} q_\xi \partial_\eta^{-1}(qr)_\xi + q_\xi U_0(\xi,t) + + q U_1(\xi,t), \\ r_t &= r_{\xi\xi\xi} - \tfrac{3}{4} r \partial_\eta^{-1}(q r_\xi)_\xi - \tfrac{3}{4} r_\xi \partial_\eta^{-1}(rq)_\xi + r_\xi U_0(\xi,t) + r U_2(\xi,t), \end{aligned} \tag{30}$$

and its reductions, *third-order Fokas equation I-1* $(r=-1)$,

$$q_t = q_{\xi\xi\xi} + \tfrac{3}{4} q \partial_\eta^{-1} q_{\xi\xi} + \tfrac{3}{4} q_\xi \partial_\eta^{-1} q_\xi + q_\xi U_0(\xi,t) + q U_1(\xi,t), \tag{31}$$

and *third-order Fokas equation I-2* $(r=-q)$,

$$q_t = q_{\xi\xi\xi} + \tfrac{3}{4} q \partial_\eta^{-1}(q q_\xi)_\xi + \tfrac{3}{4} q_\xi \partial_\eta^{-1}(q^2)_\xi + q_\xi U_0(\xi,t) + q U_1(\xi,t). \tag{32}$$

Third-order DST system II:

$$\begin{aligned} q_t &= q_{zzz} + q_{\overline{zzz}} - \tfrac{3}{4} q \partial_z^{-1}(r q_{\bar z})_{\bar z} - \tfrac{3}{4} q \partial_{\bar z}^{-1}(r q_z)_z - \tfrac{3}{4} q_{\bar z} \partial_z^{-1}(qr)_{\bar z} - \tfrac{3}{4} q_z \partial_{\bar z}^{-1}(qr)_z, \\ r_t &= r_{zzz} + r_{\overline{zzz}} - \tfrac{3}{4} r \partial_z^{-1}(q r_{\bar z})_{\bar z} - \tfrac{3}{4} r \partial_{\bar z}^{-1}(q r_z)_z - \tfrac{3}{4} r_{\bar z} \partial_z^{-1}(rq)_{\bar z} - \tfrac{3}{4} r_z \partial_{\bar z}^{-1}(rq)_z, \end{aligned} \tag{33}$$

and its reductions, *NVN-II* $(r=-1)$,

$$q_t = q_{zzz} + q_{\overline{zzz}} + \tfrac{3}{4}(q\partial_z^{-1} q_{\bar z})_{\bar z} + \tfrac{3}{4}(q\partial_{\bar z}^{-1} q_z)_z, \tag{34}$$

and *modified NVN-II* $(r=-q)$,

$$q_t = q_{zzz} + q_{\overline{zzz}} + \tfrac{3}{4} q \partial_z^{-1}(q q_{\bar z})_{\bar z} + \tfrac{3}{4} q \partial_{\bar z}^{-1}(q q_z)_z + \tfrac{3}{4} q_{\bar z} \partial_z^{-1}(q^2)_{\bar z} + \tfrac{3}{4} q_z \partial_{\bar z}^{-1}(q^2). \tag{35}$$

Third-order Fokas' system II:

$$\begin{aligned} q_t &= q_{zzz} + \tfrac{3}{4} q \partial_{\bar z}^{-1}(r q_z)_z - \tfrac{3}{4} q_z \partial_{\bar z}^{-1}(qr)_z, \\ r_t &= r_{zzz} + \tfrac{3}{4} r \partial_{\bar z}^{-1}(q r_z)_z - \tfrac{3}{4} r_z \partial_{\bar z}^{-1}(rq)_z, \end{aligned} \tag{36}$$

and its reductions, *third-order reduced Fokas II-1* $(r=-1)$,

$$q_t = q_{zzz} - \tfrac{3}{4}(q \partial_{\bar z}^{-1} q_z)_z + \tfrac{3}{4} q_z \partial_{\bar z}^{-1}(q)_z, \tag{37}$$

and *third-order reduced Fokas II-2* $(r=-q)$,

$$q_t = q_{zzz} - \tfrac{3}{4} q \partial_{\bar z}^{-1}(q q_z)_z + \tfrac{3}{4} q_z \partial_{\bar z}^{-1}(q^2)_z. \tag{38}$$

III. DERIVATION OF SYSTEMS OF INTEGRABLE EQUATIONS OF THE DS TYPE

We present the derivation of the systems of integrable equations named *general* in Sec. II, that is, where the parameters (α,β) have not yet been eliminated. The reduction of these systems to their canonical versions is discussed in the next section. This section is organized as follows. We first derive the DS-type systems with (4) as their Lax equation where σ has been set equal to 1. We refer to these systems as *systems of type I*. They admit $\xi=x+y$ and $\eta=x-y$ as characteristic variables. We see that these systems arise in connection with a nonlocal matrix Riemann–Hilbert problem on the complex plane. We do this in two steps. In Sec. III A, we derive the systems for zero boundary functions in two propositions. In Proposition 1, we derive the second-order system and Proposition 2 contains the derivation of the third-order system. The equations resulting from this method do not contain the boundary functions, that is, the additional terms in only one of the variables η and ξ that appear in the equations in the list of the Introduction. In Sec. III B, we modify the t-part of the Lax pair to include nonzero boundary functions. This is done in two propositions. Proposition 3 contains the derivation of the second-order system, while Proposition 4 contains the derivation of the third-order system. This is the formulation which yields the equations as listed in Sec. II. Subsection III C, contains the derivation of the second and third-order systems arising from a slight modification of the Lax equation resulting in characteristic variables being complex numbers $z=x+iy$ and $\bar{z}=x-iy$. This is done in two propositions (Propositions 5 and 6). Again, these two systems have a common solution technique, based on a nonlocal DBAR problem. We shall call this systems *systems of type II*.

In what follows, α and β are real numbers, $J=\mathrm{diag}\{1,-1\}$ is a 2×2 matrix, and q and r are complex-valued functions of x, y, and t. In order to match the equations in this work with the ones appearing in the literature the parameter λ should be replaced by ik. This change accounts for the missing i multiplying the derivatives with respect to t in the second-order equations. The change $t\to it$ restores the equations to the usual form.

A. Derivations of type I systems for zero boundary functions

Proposition 1: The Lax pair formed by L and L_2, where L is as in (4),

$$L_2=\partial_t-(\alpha+\beta J)\partial_y^2-\beta P\partial_y+Q,$$

and Q is defined by (42), gives rise to the system of equations

$$q_t=(\beta-\alpha)q_{\xi\xi}+(\beta+\alpha)q_{\eta\eta}-\frac{q}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right),$$

$$r_t=(\beta-\alpha)r_{\xi\xi}+(\beta+\alpha)r_{\eta\eta}-\frac{r}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right).$$

Proposition 2: The Lax pair formed by L and L_3, where

$$L_3=\partial_t+(\alpha+\beta J)\partial_y^3+\beta P\partial_y^2+C\partial_y+Q,$$

and C and Q are defined by (44) and (45), respectively, gives rise to the system of equations

$$\begin{aligned}q_t=&(\beta-\alpha)q_{\xi\xi\xi}+(\alpha+\beta)q_{\eta\eta\eta}-\tfrac{3}{4}(\beta-\alpha)q\partial_\eta^{-1}(rq_\xi)_\xi-\tfrac{3}{4}(\alpha+\beta)q\partial_\xi^{-1}(rq_\eta)_\eta\\&-\tfrac{3}{4}(\beta-\alpha)q_\xi\partial_\eta^{-1}(qr)_\xi-\tfrac{3}{4}(\alpha+\beta)q_\eta\partial_\xi^{-1}(qr)_\eta,\\ r_t=&(\beta-\alpha)r_{\xi\xi\xi}+(\alpha+\beta)r_{\eta\eta\eta}-\tfrac{3}{4}(\beta-\alpha)r\partial_\eta^{-1}(qr_\xi)_\xi-\tfrac{3}{4}(\alpha+\beta)r\partial_\xi^{-1}(qr_\eta)_\eta\\&-\tfrac{3}{4}(\beta-\alpha)r_\xi\partial_\eta^{-1}(rq)_\xi-\tfrac{3}{4}(\alpha+\beta)r_\eta\partial_\xi^{-1}(rq)_\eta.\end{aligned}$$

Proof of Proposition 1:

We will show that the second-order system possesses a Lax pair; the system arises as the compatibility condition for a pair of overdetermined linear differential operators with matrix

coefficients. In what follows, let Ψ be a matrix function with variables λ, x, y, and t and let $\mathscr{L}_1$ and $\mathscr{L}_2$ be differential operators with matrix coefficients so that $\mathscr{L}_1\Psi=0$ and $\mathscr{L}_2\Psi=0$. We shall write

$$\Psi=\mu(x,y,t,\lambda)\exp\{\lambda(Jx-y)+\lambda^2(\alpha+\beta J)t\},$$

where μ is a matrix function. It is convenient to introduce the operators

$$D_x\mu=\mu_x+\lambda\mu J,\quad D_y\mu=\mu_y-\lambda\mu,$$

$$D_t\mu=\mu_t+\lambda^2\mu(\alpha+\beta J).$$

Thus $\mathscr{L}\Psi=(\Sigma_{i=1}^N A_i\partial_x^i+\Sigma_{i=1}^N B_i\partial_y^i+\partial_t)\Psi=0$ if and only if $L\mu=(\Sigma_{i=1}^N A_iD_x^i+\Sigma_{i=1}^N B_iD_y^i+D_t)\mu=0$.

We begin by considering a nonlocal Riemann–Hilbert problem,

$$\mu^+(k,x,y,t)=\int_R dl\ \mu^-(l,x,y,t)F(l,k,x,y,t),\quad k\in R$$
$$\mu\sim 1+\frac{\mu_1(x,y,t)}{\lambda}+\frac{\mu_2(x,y,t)}{\lambda^2}+O\left(\frac{1}{\lambda^3}\right),\quad \lambda\to\infty. \tag{39}$$

In this equation μ^+, μ^-, and F are matrix-valued functions and the super indices $+$ and $-$ denote holomorphicity in the upper and lower complex planes, respectively. We assume that (39) has unique solution μ^+,μ^- for a given function F. Denote by P_F the rhs of Eq. (39), that is,

$$P_F\mu\equiv\int_R dl\ \mu(l,x,y,t)F(l,k,x,y,t),$$

and choose F so that P_F and the operators D_x, D_y, and D_t commute. The conditions on F that arise from this are

$$F_x=(lJF-kFJ),\quad F_y=(l-k)F,$$

$$F_t=l^2F(\alpha+\beta J)-k^2(\alpha+\beta J)F.$$

Next, look for linear combinations of $D_x\mu$, $D_y\mu$, and $D_t\mu$ that vanish for $\lambda\to\infty$. This guarantees that these combinations must be identically zero since we are assuming that the homogeneous Riemann–Hilbert problem has unique zero solution. The simplest such linear combination L_1 is of the form $L_1=D_x+JD_y+P$. The $\mathscr{O}(\lambda)$ term cancels and the $\mathscr{O}(1)$ also cancels by choosing $P=[J,\mu_1]=\left(\begin{smallmatrix}0 & q\\ r & 0\end{smallmatrix}\right)$. This is the time-independent part of the Lax pair.

Using (39) and grouping powers of λ we find

$$\mathscr{O}(\lambda^{-i}):\mu_{ix}+J\mu_{iy}+[\mu_{i+1},J]+P\mu_i=0. \tag{40}$$

These are recursive equations for the coefficients $\mu_i=\left(\begin{smallmatrix}a_i & b_i\\ c_i & d_i\end{smallmatrix}\right)$ of the function μ. These coefficients are determined only up to the addition of arbitrary functions of ξ or η and t. That is,

$$\begin{aligned}0=L_1\mu&=\mu_x+\lambda\mu J+J\mu_y-\lambda J\mu+P\mu\\&=(\mu_x+J\mu_y)+\lambda[\mu_1,J]+P\mu\\&=2\sum_{k=0}^{\infty}\left\{\begin{pmatrix}a_{k\xi} & b_{k\xi}\\ c_{k\eta} & d_{k\eta}\end{pmatrix}+\begin{pmatrix}0 & b_{k+1}\\ c_{k+1} & 0\end{pmatrix}+\frac{1}{2}\begin{pmatrix}qc_k & qd_k\\ ra_k & rb_k\end{pmatrix}\right\}\lambda^{-k}.\end{aligned}$$

So for $k=0$ one obtains $b_1=q/2$ and $c_1=-r/2$ from the off diagonal entries. From the diagonal entries one has that $a_1=\frac{1}{4}\partial_\xi^{-1}(qr)+v_0(\eta,t)$ and $d_1=-\frac{1}{4}\partial_\eta^{-1}(qr)+u_0(\xi,t)$, where u_0 and v_0 are arbitrary functions of ξ or η and t. We will assume here that the functions u_0 and v_0 are zero. To find the time-dependent part of the Lax pair we set

$$\begin{aligned} L_2\mu &= D_t\mu + AD_y^2\mu + BD_y\mu + Q\mu \\ &= \mu_t + \lambda^2\mu(\alpha+\beta J) + A(\mu_{yy} - 2\lambda\mu_y + \lambda^2\mu) + B(\mu_y - \lambda\mu) + Q\mu = 0. \end{aligned} \tag{41}$$

Collecting the powers of the parameter λ we obtain the following equations for the coefficients A, B, and Q:

$$\mathcal{O}(\lambda^2):(\alpha+\beta J)+A=0,$$

$$\mathcal{O}(\lambda):\mu_1(\alpha+\beta J)-(\alpha+\beta J)\mu_1-B=\beta[\mu_1,\ J]-B=0,$$

$$\mathcal{O}(1):\mu_2(\alpha+\beta J)-2A\mu_{1y}-(\alpha+\beta J)\mu_2-B\mu_1+Q=0.$$

Solving these equations one obtains $A=-(\alpha+\beta J)$, $B=\beta[\mu_1,\ J]=-\beta P$ and $Q=\left(\begin{smallmatrix} Q^{11} & Q^{12} \\ Q^{21} & Q^{22}\end{smallmatrix}\right)$ with

$$\begin{aligned} Q^{11} &= -\tfrac{1}{2}\alpha qr + \tfrac{1}{2}(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta, \quad Q^{12} = -\alpha q_\xi + (\beta+\alpha)q_\eta, \\ Q^{22} &= -\tfrac{1}{2}\alpha qr - \tfrac{1}{2}(\beta-\alpha)\partial_\eta^{-1}(qr)_\xi, \quad Q^{21} = -\alpha r_\eta - (\beta-\alpha)r_\xi. \end{aligned} \tag{42}$$

The $\mathcal{O}(\lambda^{-1})$ order of (41) produces the equation

$$\begin{aligned} 0 &= \mu_{1t} + \mu_3(\alpha+\beta J) + A\mu_{1yy} - 2A\mu_{2y} + A\mu_3 + B\mu_{1y} - B\mu_2 + Q\mu_1 \\ &= \mu_{1t} + \beta[\mu_3, J] + 2(\alpha+\beta J)\mu_{2y} - (\alpha+\beta J)\mu_{1yy} \\ &\quad + 2(\alpha+\beta J)\mu_{2y} - \beta P\mu_{1y} + \beta P\mu_2 + Q\mu_1. \end{aligned} \tag{43}$$

From the off diagonal entries in the equation above and using the specific formulas for Q and the recursion relations (40), we obtain the desired equations for q and r.

Remark 1: The same equations can be obtained as a compatibility condition, that is, from the statement that $\partial_x\partial_t\Psi=\partial_t\partial_x\Psi$. □

Proof of Proposition 2:

Let Ψ be of the form

$$\Psi=\mu\ \exp\{\lambda(Jx-y)+\lambda^3(\alpha+\beta J)t\}.$$

The difference between this and the previous Ψ is the cube power of λ that accompanies the variable t in the exponential. Now we search for operators $\mathcal{L}_1$ and $\mathcal{L}_3$ such that $\mathcal{L}_1\Psi=0$ and $\mathcal{L}_3\Psi=0$.

Remark 2: The operator $\mathcal{L}_1$ obtained above together with the recursion relations (40) satisfies $\mathcal{L}_1\Psi=0$.

Set $L_3\mu=0$ for a general operator $L_3=D_t+AD_y^3+BD_y^2+CD_y+Q$:

$$\begin{aligned} L_3\mu &= \mu_t + \lambda^3\mu(\alpha+\beta J) + A(\mu_{yyy} - 3\lambda\mu_{yy} + 3\lambda^2\mu_y - \lambda^3\mu) \\ &\quad + B(\mu_{yy} - 2\lambda\mu_y + \lambda^2\mu) + C(\mu_y - \lambda\mu) + Q\mu. \end{aligned}$$

Collecting powers of λ one obtains the following equations:

$$\mathcal{O}(\lambda^3):(\alpha+\beta J)-A=0, \quad \mathcal{O}(\lambda^2):\beta[\mu_1,\ J]+B=0,$$

$$\mathcal{O}(\lambda^1):\beta[\mu_2,\ J]+B\mu_1+3A\mu_{1y}-C=0,$$

$$\mathcal{O}(1):\beta[\mu_3,\ J]+3(\alpha+\beta J)\mu_{2y}-3(\alpha+\beta J)\mu_{1yy}-\beta P\mu_2-2\beta P\mu_{1y}-C\mu_1+Q=0.$$

Solving these equations we obtain $A=(\alpha+\beta J)$, $B=\beta P$, $C=\begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix}$ and $Q=\begin{pmatrix} Q^{11} & Q^{12} \\ Q^{21} & Q^{22} \end{pmatrix}$ with

$$\begin{aligned} C^{11}&=(\tfrac{3}{4}\alpha+\tfrac{1}{4}\beta)qr-\tfrac{3}{4}(\alpha+\beta)\partial_\xi^{-1}(qr)_\eta,\\ C^{12}&=(\tfrac{3}{4}\alpha+\tfrac{1}{2}\beta)q_\xi-\tfrac{3}{2}(\alpha+\beta)q_\eta,\\ C^{21}&=(\tfrac{3}{2}\alpha-\tfrac{1}{2}\beta)r_\eta-\tfrac{3}{2}(\alpha-\beta)r_\xi,\\ C^{22}&=(\tfrac{3}{4}\alpha-\tfrac{1}{4}\beta)qr-\tfrac{3}{4}(\alpha-\beta)\partial_\eta^{-1}(qr)_\xi, \end{aligned} \tag{44}$$

and

$$Q=-3(\alpha+\beta J)\mu_{2y}+3(\alpha+\beta J)\mu_{1yy}+\beta P\mu_2+2\beta P\mu_{1y}+C\mu_1+\beta[J,\ \mu_3]. \tag{45}$$

Compatibility of the operators L_1 and L_3 produce the following equations for q and r:

$$\begin{aligned} q_t&=2Q^{12}_\xi+q(Q^{22}-Q^{11})-C^{11}q_y-(\alpha+\beta)q_{yyy},\\ r_t&=2Q^{21}_\eta+r(Q^{11}-Q^{22})-C^{22}r_y-(\alpha-\beta)r_{yyy}. \end{aligned} \tag{46}$$

When the expressions for Q and C are substituted in we obtain the equations in Propositions 2 and 6. □

B. Modified *t*-part of the Lax pair for type I systems

Proposition 3: The system (6),

$$\begin{aligned} iq_t=&(\beta-\alpha)q_{\xi\xi}+(\beta+\alpha)q_{\eta\eta}-\frac{q}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right)\\ &+q\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right),\\ ir_t=&(\beta-\alpha)r_{\xi\xi}+(\beta+\alpha)r_{\eta\eta}-\frac{r}{2}\left((\beta-\alpha)\partial_\eta^{-1}(qr)_\xi+(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta\right)\\ &+r\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right), \end{aligned}$$

can be obtained from the following Lax pair:

$$\begin{aligned} &L\Psi=0,\\ &\Psi_t=(\alpha+\beta J)\Psi_{yy}+\beta P\Psi_y+Q\Psi+\int_R dl\ \Psi(l)\mathbf{f}(k,l), \end{aligned} \tag{47}$$

where L is as in (4), Q is defined by (49) below, and $\mathbf{f}=\mathrm{diag}\{f_1,f_2\}$:

$$\begin{aligned} f_1(k,l)&=(\beta+\alpha)\left[k^2\delta(k-l)+\frac{1}{2\pi}\int_R d\eta\ V(\eta)\exp\left(-i(k-l)\eta\right)\right],\\ f_2(k,l)&=(\beta-\alpha)\left[k^2\delta(k-l)+\frac{1}{2\pi}\int_R d\xi\ U(\xi)\exp\left(-i(k-l)\xi\right)\right]. \end{aligned}$$

If U and V are real valued functions, this system admits the reduction $r=\epsilon\bar{q}$:

$$\begin{aligned} q_t=&(\beta-\alpha)q_{\xi\xi}+(\beta+\alpha)q_{\eta\eta}-\epsilon\frac{q}{2}\left((\beta-\alpha)\partial_\eta^{-1}(|q|^2)_\xi+(\beta+\alpha)\partial_\xi^{-1}(|q|^2)_\eta\right)\\ &+\epsilon q\left((\beta-\alpha)U(\xi,t)+(\beta+\alpha)V(\eta,t)\right). \end{aligned}$$

Proof: Recall that the t-part of the Lax pair for the system (6) is given by

$$\Psi_t=(\alpha+\beta J)\Psi_{yy}+\beta P\Psi_y+Q\Psi \tag{48}$$

with $Q=\left(\begin{smallmatrix}Q^{11} & Q^{12}\\ Q^{21} & Q^{22}\end{smallmatrix}\right)$ as in (42) and where $\Psi=(\Psi_1,\Psi_2)^T$. In the case that the boundary functions are not zero, the coefficients in Q can be expanded as follows:

$$\begin{aligned}
Q^{11}&=-\tfrac{1}{2}\alpha qr+\tfrac{1}{2}(\beta+\alpha)\partial_\xi^{-1}(qr)_\eta+(\beta+\alpha)v_0'(\eta),\\
Q^{12}&=-\alpha q_\xi+(\beta+\alpha)q_n,\\
Q^{22}&=-\tfrac{1}{2}\alpha qr-\tfrac{1}{2}(\beta-\alpha)\partial_\eta^{-1}(qr)_\xi-(\beta-\alpha)u_0'(\xi),\\
Q^{21}&=-\alpha r_\eta-(\beta-\alpha)r_\xi.
\end{aligned} \tag{49}$$

In this case, the equation (48) for Ψ must be modified. We start by defining

$$\Psi_1=\mu_{12}^- e^{-ik\xi},\quad \Psi_2=\mu_{22}^- e^{-ik\xi},$$

so from (4)

$$\Psi_1=\frac{1}{2}\int_\xi^\infty d\xi'\ q\Psi_2,\quad \Psi_2=e^{-ik\xi}-\frac{1}{2}\int_{-\infty}^\eta d\eta'\ r\Psi_1.$$

Now add an extra term of the form $\int_R dl\ f(k,l)\Psi(l)$ to the t-part of the Lax pair (48),

$$\Psi_t=(\alpha+\beta J)\Psi_{yy}+\beta P\Psi_y+Q\Psi+\int_R dl\ f(k,l)\Psi(l),$$

and compute the limits of the equation above as $\eta\to-\infty$. Taking into account that $\Psi_1\to 0$, $\Psi_2\sim e^{-ik\xi}$, and $q,r\sim 0$, we obtain

$$0=-(\beta-\alpha)(-ik)^2e^{-ik\xi}-(\beta-\alpha)U(\xi)e^{-ik\xi}+\int dl\ f(k,l)e^{-il\xi}.$$

So

$$\begin{aligned}
\int_R d\xi\int_R dl'\ f(k,l')\exp(-i(l'-l)\xi)&=(\beta-\alpha)k^2\int_R d\xi\ \exp(-i(k-l)\xi)\\
&\quad+(\beta-\alpha)\int_R d\xi\ U(\xi)\exp(-i(k-l)\xi),
\end{aligned}$$

or

$$f(k,l)=(\beta-\alpha)k^2\delta(k-l)+(\beta-\alpha)\frac{1}{2\pi}\int_R d\xi\ U(\xi)\exp(-i(k-l)\xi).$$

□

Proposition 4: The system (11) can be obtained from the following Lax pair:

$$L\Psi=0,$$

$$-\Psi_t=(\alpha+J\beta)\Psi_{yyy}+\beta P\Psi_{yy}+C\Psi_y+Q\Psi+\mathbf{v},$$

where C and Q are defined by (53) and (54) and $\mathbf{v}=\operatorname{diag}\{v,u\}$:

$$v=-(\beta+\alpha)\left((ik)^3\Psi+\int_R\gamma_1(k-l)\Psi(l)\,dl+(ik)\int\gamma_2(k-l)\Psi(l)\,dl\right),$$

$$u=(\beta-\alpha)\left((ik)^3\Psi+\int_R\delta_1(k-l)\Psi(l)\,dl+(ik)\int\delta_2(k-l)\Psi(l)\,dl\right),$$

where

$$\gamma_1(n)=\frac{1}{2\pi}\int_R V_2(\eta)e^{-in\eta}\,d\eta,\quad \delta_1(s)=\frac{1}{2\pi}\int_R U_2(\xi)e^{-is\xi}\,d\xi,$$

$$\gamma_2(n)=\frac{1}{2\pi}\int_R V_0(\eta)e^{-in\eta}\,d\eta,\quad \delta_2(s)=\frac{1}{2\pi}\int_R U_0(\xi)e^{-is\xi}\,d\xi.$$

The functions U_0,U_1,U_2 *and* V_0,V_1,V_2 *are related by the equations*

$$\begin{aligned}U_1(\xi,t)&=-U_2(\xi,t)+U_0'(\xi,t),\\ V_1(\eta,t)&=-V_2(\eta,t)+V_0'(\eta,t).\end{aligned}\tag{50}$$

This system admits the following reductions:

(1) $r=-1$. *Setting* $r=-1$ *in the above system, imposing the condition* $U_2=V_2=0$ [*so that* $U_0'(\xi,t)=U_1(\xi,t)$ *and* $V_0'(\eta,t)=V_1(\eta,t)$ *by virtue of (50)*], *and writing* $q=u$, *the system reduces to the single equation*

$$\begin{aligned}u_t=&(\beta-\alpha)u_{\xi\xi\xi}+(\beta+\alpha)u_{\eta\eta\eta}+\tfrac{3}{4}(\alpha+\beta)(u\partial_\xi^{-1}u_\eta)_\eta+\tfrac{3}{4}(\beta-\alpha)(u\partial_\eta^{-1}u_\xi)_\xi\\&+u((\beta-\alpha)U_1(\xi,t)+(\beta+\alpha)V_1(\eta,t))+(\alpha-\beta)U_0(\xi,t)u_\xi+(\beta+\alpha)V_0(\eta,t)u_\eta.\end{aligned}\tag{51}$$

(2) $r=-q$. *Setting* $-r=q=v$ *and imposing the condition* $U_1=U_2$ *and* $V_1=V_2$, *the system reduces to the single equation*

$$\begin{aligned}v_t=&(\beta-\alpha)v_{\xi\xi\xi}+(\alpha+\beta)v_{\eta\eta\eta}+\tfrac{3}{8}(\beta-\alpha)v\partial_\eta^{-1}(v^2)_{\xi\xi}+\tfrac{3}{8}(\alpha+\beta)v\partial_\xi^{-1}(v^2)_{\eta\eta}\\&+\tfrac{3}{4}(\beta-\alpha)v_\xi\partial_\eta^{-1}(v^2)_\xi+\tfrac{3}{4}(\alpha+\beta)v_\eta\partial_\xi^{-1}(v^2)_\eta+v((\beta-\alpha)U_1(\xi,t)\\&+(\beta+\alpha)V_1(\eta,t))+(\beta-\alpha)U_0(\xi,t)v_\xi+(\beta+\alpha)V_0(\eta,t)v_\eta.\end{aligned}\tag{52}$$

Proof: When the additional terms coming from the integration of the recursion relations (40) are considered, the following terms must be added to the coefficients C and Q in (44) and (45), respectively:

$$C_{11}=\text{old } C_{11}-\tfrac{3}{4}(\beta-\alpha)v_0'(\eta,t),\quad C_{22}=\text{old } C_{22}-\tfrac{3}{4}(\beta+\alpha)u_0'(\xi,t),\tag{53}$$

$$\begin{aligned}Q_{12}&=\text{old } Q_{12}+\tfrac{3}{8}q[(\beta-\alpha)u_0'(\xi,t)-(\beta+\alpha)v_0'(\eta,t)],\\ Q_{21}&=\text{old } Q_{21}+\tfrac{3}{8}r[(\beta-\alpha)u_0'(\xi,t)-(\beta+\alpha)v_0'(\eta,t)],\\ Q_{11}&=\text{old } Q_{11}+3(\beta+\alpha)[-v_1'(\eta)+\tfrac{1}{4}v_0''(\eta)-\tfrac{1}{8}(v_0^2(\eta))'],\\ Q_{22}&=\text{old } Q_{22}+3(\beta-\alpha)[-u_1'(\xi)+\tfrac{1}{4}u_0''(\xi)-\tfrac{1}{8}(u_0^2(\xi))'].\end{aligned}\tag{54}$$

In turn, these add new terms to Eqs. (46):

$$q_t = \text{old terms} - \tfrac{3}{4}(\beta+\alpha) q_\eta v_0'(\eta) + (\beta+\alpha) q(-\tfrac{3}{4} v_0''(\eta) + 3v_1'(\eta) + \tfrac{3}{8}(v_0^2(\eta))')$$
$$- \tfrac{3}{4}(\beta-\alpha) q_\xi u_0'(\xi) + (\beta-\alpha) q(-3u_1'(\xi) - \tfrac{3}{8}(u_0(\xi)^2)'),$$

$$r_t = \text{old terms} - \tfrac{3}{4}(\beta-\alpha) r_\xi u_0'(\xi) + (\beta-\alpha) r(-\tfrac{3}{4} u_0''(\xi) + 3u_1'(\tau) + \tfrac{3}{8}(u_0(\xi)^2)')$$
$$- \tfrac{3}{4}(\alpha+\beta) r_\eta v_0'(\eta) + (\beta+\alpha) r(-3v_1'(\eta) - \tfrac{3}{8}(v_0^2(\eta))').$$

We define

$$U_0 = -\tfrac{3}{4} u_0', \quad V_0 = -\tfrac{3}{4} v_0',$$

$$U_1 = (-3u_1' - \tfrac{3}{8}(u_2^2)'), \quad V_1 = (-\tfrac{3}{4} v_0'' + 3v_1' + \tfrac{3}{8}(v_0^2)'),$$

$$U_2 = (-\tfrac{3}{4} u_0'' + 3u_1' + \tfrac{3}{8}(u_0^2)'), \quad V_2 = (-3v_1' - \tfrac{3}{8}(v_0^2)').$$

The original equation for Ψ_t for the third-order system is

$$\Psi_t + (\alpha+\beta J)\Psi_{yyy} + \beta P \Psi_{yy} + C\Psi_y + Q\Psi = 0. \tag{55}$$

We define

$$\Psi_1 = \mu_{11}^+ e^{ik\eta}, \quad \Psi_2 = \mu_{21}^+ e^{ik\eta}.$$

Then, from (4),

$$\Psi_1 = e^{ik\eta} - \frac{1}{2} \int_{-\infty}^{\xi} d\xi' \; q\Psi_2, \quad \Psi_2 = \frac{1}{2} \int_{\eta}^{\infty} d\eta' \; r\Psi_1.$$

In order to find the t-part of the Lax pair in the case that the functions U_0, U_1, U_2 and V_0, V_1 and V_2 are not zero, we add an extra term of the form $\int_R dl \; f(k,l)\Psi(l)$ to the t-part of the Lax pair (55):

$$-\Psi_t = (\alpha+\beta J)\Psi_{yyy} + \beta P \Psi_{yy} + C\Psi_y + Q\Psi + \int_R dl \; f(k,l)\Psi(l). \tag{56}$$

Writing (56) in components,

$$-\Psi_{1t} = (\alpha+\beta)\Psi_{1yyy} + q\Psi_{2yy} + C^{11}\Psi_{1y} + C^{12}\Psi_{2y} + Q^{11}\Psi_1 + Q^{12}\Psi_2 + \int f(k,l)\Psi_1(l) \; dl. \tag{57}$$

$$-\Psi_{2t} = (\alpha-\beta)\Psi_{2yyy} + r\Psi_{1yy} + C^{21}\Psi_{1y} + C^{22}\Psi_{2y} + Q^{21}\Psi_1 + Q^{22}\Psi_2 + \int f(k,l)\Psi_2(l) \; dl.$$

When $\xi \to -\infty$ we have $\binom{\Psi_1}{\Psi_2} \sim \binom{e^{ik\eta}}{0}$. Taking the limit $\xi \to -\infty$ in (57) we find

$$\lim_{\xi \to -\infty} (Q^{21} e^{-ik\eta}) = 0$$

and

$$0 = (\beta+\alpha)(ik)^3 e^{ik\eta} - \frac{3}{4}(\beta+\alpha) v_0' \eta (ik) e^{ik\eta} + \lim_{\eta \to -\infty} Q^{11} e^{ik\eta} + \int_R f(k,l) e^{ik\eta} \, dl. \tag{58}$$

From the expression for Q^{11} we obtain

$$\lim_{\eta \to \infty} Q^{11} = (\beta+\alpha)(-3v_1'(\eta) + \tfrac{3}{4} v_0''(\eta) + \tfrac{3}{8}(v_0(\eta)^2)').$$

Assumption 1: In order to justify the limits above we only need to assume that the functions U_n and V_n are bounded, and not that they go to zero necessarily, since q decays as η and ξ go to either infinity.

From (58),

$$\int dl'\ f(k,l)e^{-il'\eta} = -(\beta+\alpha)((ik)^3 e^{ik\eta} + V_0(\eta)e^{ik\eta} + e^{ik\eta}V_2(\eta)),$$

where we have used the definitions of V_0 and V_2 given in the proof of Proposition 2. Multiplying both sides by $e^{il\eta}$ and integrating the variable η,

$$f(k,l) = -(\beta+\alpha)\left((ik)^3\delta(k-l) + \frac{ik}{2\pi}\int_R V_0(\eta)e^{-i(k-l)\eta}\,d\eta + \frac{1}{2\pi}\int_R V_2(\eta)e^{-i(k-l)\eta}\,d\eta \right).$$

Thus, the correct Lax pair is obtained by adding the vector

$$\mathrm{v} = -(\beta+\alpha)\left((\mathrm{ik})^3\Psi + \int_{\mathrm{R}} \gamma_1(\mathrm{k}-\mathrm{l})\Psi(\mathrm{l})\,\mathrm{dl} + (\mathrm{ik})\int \gamma_2(\mathrm{k}-\mathrm{l})\Psi(\mathrm{l})\,\mathrm{dl} \right),$$

where

$$\gamma_1(n) = \frac{1}{2\pi}\int_R V_2(\eta)e^{-in\eta}\,d\eta, \qquad \gamma_2(n) = \frac{1}{2\pi}\int_R V_0(\eta)e^{-in\eta}\,d\eta. \qquad \square$$

C. Derivation of type II systems

In what follows, we assume that $\sigma = i$ in (4), thus ∂_η becomes ∂_z and ∂_ξ becomes $\partial_{\bar z}$.

Proposition 5: The Lax pair formed by L and L_2, where L is as in (4),

$$L_2 = \partial_t + (\alpha+\beta J)\partial_y^2 + i\beta P\partial_y + Q,$$

and Q is defined by (42), gives rise to the system of equations

$$q_t = (\beta-\alpha)q_{zz} + (\beta+\alpha)q_{\bar z\bar z} - \frac{q}{2}\left((\beta-\alpha)\partial_{\bar z}^{-1}(qr)_z + (\beta+\alpha)\partial_z^{-1}(qr)_{\bar z}\right),$$

$$r_t = (\beta-\alpha)r_{zz} + (\beta+\alpha)r_{\bar z\bar z} - \frac{r}{2}\left((\beta-\alpha)\partial_{\bar z}^{-1}(qr)_z + (\beta+\alpha)\partial_z^{-1}(qr)_{\bar z}\right).$$

This system admits the reduction $r = \epsilon\bar q$ and in this case it reduces to

$$q_t = (\beta-\alpha)q_{zz} + (\beta+\alpha)q_{\bar z\bar z} - \epsilon\frac{q}{2}\left((\beta-\alpha)\partial_{\bar z}^{-1}(|q|^2)_z + (\beta+\alpha)\partial_z^{-1}(|q|^2)_{\bar z}\right).$$

Proposition 6: The Lax pair formed by L and L_3, where

$$L_3 = \partial_t + i(\alpha+\beta J)\partial_y^3 - \beta P\partial_y^2 - iC\partial_y + Q,$$

and C and Q are defined by (44) and (45), respectively, gives rise to the system of equations

$$\begin{aligned} q_t &= (\beta-\alpha)q_{zzz} + (\alpha+\beta)q_{\bar z\bar z\bar z} - \tfrac{3}{4}(\alpha+\beta)q\partial_z^{-1}(rq_{\bar z})_{\bar z} - \tfrac{3}{4}(\beta-\alpha)q\partial_{\bar z}^{-1}(rq_z)_z \\ &\quad - \tfrac{3}{4}(\alpha+\beta)q_{\bar z}\partial_z^{-1}(qr)_{\bar z} - \tfrac{3}{4}(\beta-\alpha)q_z\partial_{\bar z}^{-1}(qr)_z, \\ r_t &= (\beta-\alpha)r_{zzz} + (\alpha+\beta)r_{\bar z\bar z\bar z} - \tfrac{3}{4}(\alpha+\beta)r\partial_z^{-1}(qr_{\bar z})_{\bar z} - \tfrac{3}{4}(\beta-\alpha)r\partial_{\bar z}^{-1}(qr_z)_z \\ &\quad - \tfrac{3}{4}(\alpha+\beta)r_{\bar z}\partial_z^{-1}(rq)_{\bar z} - \tfrac{3}{4}(\beta-\alpha)r_z\partial_{\bar z}^{-1}(rq)_z. \end{aligned} \tag{59}$$

This system admits the following reductions:

(1) $r=-1$. Setting $r=-1$, the system reduces to the single equation

$$u_t=(\beta-\alpha)u_{zzz}+(\beta+\alpha)u_{\bar{z}\bar{z}\bar{z}}+\tfrac{3}{4}(\alpha+\beta)(u\partial_z^{-1}u_{\bar{z}})_{\bar{z}}+\tfrac{3}{4}(\beta-\alpha)(u\partial_{\bar{z}}^{-1}u_z)_z, \tag{60}$$

where $q=u$.

(2) $r=-q=-v$. Setting $r=-q=-v$, the system reduces to the single equation

$$\begin{aligned} v_t=&(\beta-\alpha)v_{zzz}+(\alpha+\beta)v_{\bar{z}\bar{z}\bar{z}}+\tfrac{3}{4}(\alpha+\beta)v\partial_z^{-1}(vv_{\bar{z}})_{\bar{z}}+\tfrac{3}{4}(\beta-\alpha)v\partial_{\bar{z}}^{-1}(vv_z)_z\\ &+\tfrac{3}{4}(\alpha+\beta)v_{\bar{z}}\partial_z^{-1}(v^2)_{\bar{z}}+\tfrac{3}{4}(\beta-\alpha)v_z\partial_{\bar{z}}^{-1}(v^2)_z. \end{aligned} \tag{61}$$

Proof of Propositions 5 and 6: The starting point in this case is the DBAR problem,

$$\begin{aligned} &\partial_\lambda^{-}\mu(k,x,y,t)=\int dl\ \mu(l,x,y,t)R_0(l,k,x,y,t),\\ &\mu(k,x,y,t)\sim 1+\frac{\mu_1(x,y,t)}{\lambda}+\frac{\mu_2(x,y,t)}{\lambda^2}+O\left(\frac{1}{\lambda^3}\right),\quad \lambda\to\infty. \end{aligned} \tag{62}$$

The functions μ and R_0 are matrix-valued functions. We introduce the operators

$$\begin{aligned} &D_x\mu=\mu_x+\lambda\mu J,\quad D_y\mu=\mu_y-i\lambda\mu,\\ &D_t\mu=\mu_t+\lambda^2\mu(\alpha+\beta J). \end{aligned} \tag{63}$$

Again we assume that the $\bar{\partial}$ problem (62) has a unique solution. Let P_R be given by the rhs of (62), that is

$$P_R\mu(k,x,y,t)\equiv\int dl\ \mu(l,x,y,t)R(l,k,x,y,t),$$

and choose R so that P_R commutes with the operators (63). This produces the following equations for the function R:

$$\begin{aligned} &R_x=-kJR+lRJ,\quad R_y=i(k-l)R,\\ &R_t=-l^2R(\alpha+\beta J)+k^2(\alpha+\beta J)R. \end{aligned}$$

These equations are satisfied if, for example,

$$R(k,l,x,y,t)=\exp\{-k(xJ-iy)+k^2(\alpha+\beta J)t\}R_0(k,l)\exp\{k(xJ-iy)-k^2(\alpha+\beta J)t\}.$$

Now we look for linear combinations of the operators D_x, D_y, and D_t that vanish for $k\to\infty$. By the assumption of uniqueness of solution for the DBAR problem, these linear combinations must be identically zero. □

IV. RELATIONS BETWEEN THE VARIOUS EQUATIONS

A. Simple transformations

In this section we analyze the relation between the various systems derived in the previous chapter. We show that there are three different cases of the system (6) in Proposition 1 corresponding to the values $(\alpha,\beta)=(0,1)$, $(\alpha,\beta)=(1,0)$, and $(\alpha,\beta)=(\frac{1}{2},\ \frac{1}{2})$; and only two cases of the systems considered in Propositions 5, 2, and 6. These correspond to $(\alpha,\beta)=(0,1)$ and $(\alpha,\beta)=(\frac{1}{2},\ \frac{1}{2})$.

Proposition 7: If q and r satisfy (6) with α and β such that $\alpha+\beta>0$ and $\beta-\alpha>0$, then the transformed $\tilde{q}$ and $\tilde{r}$ given by

$$\tilde{q}(\eta,\xi,t)=\frac{1}{a}\,q\left(\frac{\eta}{a},\frac{\xi}{b},t\right),\quad \tilde{r}(\eta,\xi,t)=\frac{1}{b}\,r\left(\frac{\eta}{a},\frac{\xi}{b},t\right),\tag{64}$$

where $a^2=\alpha+\beta$ *and* $b^2=\beta-\alpha$, *satisfy the system (6) with* $\alpha=0$, $\beta=1$, $\tilde{U}(\eta)=U(\eta/a)$, *and* $\tilde{V}(\xi)=V(\xi/b)$.

If q *and* r *satisfy (6) with* $\alpha+\beta>0$ *and* $\beta-\alpha<0$, *then* $\tilde{q}$ *and* $\tilde{r}$ *defined by (64), where* $a^2=\beta+\alpha$ *and* $b^2=\alpha-\beta$, *satisfy (6) with* $\alpha=1$ *and* $\beta=0$.

If q *and* r *satisfy (11) with* $\beta+\alpha\neq 0$ *and* $\beta-\alpha\neq 0$, *then* $\tilde{q}$ *and* $\tilde{r}$ *defined by (64), where* $a^3=\beta+\alpha$ *and* $b^3=\alpha-\beta$, *satisfy (11) with* $\alpha=0$ *and* $\beta=1$.

Proof: The proof is a straightforward computation that makes use of the fact that

$$\tilde{f}_{\eta}(\eta,\xi,t)=\frac{1}{a}\,f\left(\frac{\eta}{a},\frac{\xi}{b},t\right),\quad \tilde{f}_{\xi}(\eta,\xi,t)=\frac{1}{b}\,f\left(\frac{\eta}{a},\frac{\xi}{b},t\right),$$

and that

$$\partial_{\eta}^{-1}\tilde{f}(\eta,\xi)=\int_{-\infty}^{\eta}\tilde{f}(s,\xi)\,ds=\int_{-\infty}^{\eta}f\left(\frac{s}{a},\frac{\xi}{b}\right)\,ds=a\int_{-\infty}^{\eta/a}f\left(s,\frac{\xi}{b}\right)\,ds=a\,\partial_{\eta}^{-1}f\left(\frac{\eta}{a},\frac{\xi}{b}\right).$$

From this computation we see that if we assume that $\alpha+\beta>0$ and $\beta-\alpha>0$ and we take $a^2=\beta+\alpha$ and $b^2=\beta-\alpha$, then we can transform (6) to the same equation but with $\beta+\alpha=1$ and $\beta-\alpha=1$. This transformation will preserve the reality of the variables ξ and η.

Performing the transformation (64) to a solution of the system in (11) with $a^3=\beta+\alpha$ and $b^3=\beta-\alpha$ we produce a solution to (11) with $(\alpha,\beta)=(0,1)$. In this case no restriction is necessary in the signs of $\alpha+\beta$ and $\beta-\alpha$, provided that neither of those quantities is zero, since we can always find a real cube root for the equations $a^3-(\beta+\alpha)=0$ and $b^3-(\beta-\alpha)=0$. □

Proposition 8: Let q *and* r *satisfy (9). Define*

$$\tilde{q}(z,\bar{z},t)=\frac{1}{a}\,q\left(\frac{z}{a},\frac{\bar{z}}{b},t\right),\quad \tilde{r}(z,\bar{z},t)=\frac{1}{b}\,r\left(\frac{z}{a},\frac{\bar{z}}{b},t\right).$$

where $a^2=\beta+\alpha$ *and* $b^2=\beta-\alpha$. *Then* $\tilde{q}$ *and* $\tilde{r}$ *satisfy the system (9) with* $\alpha=0$ *and* $\beta=1$. *Let* q *and* r *be solutions to (59). Then* $\tilde{q}$ *and* $\tilde{r}$ *given by (64), where* $a^3=\beta+\alpha$ *and* $b^3=\beta-\alpha$, *solve (59) with* $\alpha=0$ *and* $\beta=1$.

Proof: The proof is similar to that of the previous proposition. We show in detail the action of the change of variables on the operators ∂_z^{-1} and $\partial_{\bar{z}}^{-1}$. For any function $f(x,y)=f(z)=f(z,\bar{z})$, we have

$$(\partial_z^{-1}f\,)(z)=\frac{1}{\pi}\int dx\,dy\,\frac{f(\zeta,\bar{\zeta})}{z-\zeta},$$

where $z=x+iy$ and $\bar{z}=x-iy$. Setting $f(z,\bar{z})=\hat{f}(az,b\bar{z})$, the above transforms to

$$\begin{aligned}(\partial_z^{-1}f\,)(z)&=\frac{1}{\pi}\int dx\,dy\,\frac{\tilde{f}(a\zeta,b\bar{\zeta})}{z-\zeta}\\&=\frac{1}{\pi}\int dx'\,dy'\,\frac{\partial(x,y)}{\partial(x',y')}\,\frac{\tilde{f}(x',y')}{z-\eta/a}=\frac{1}{b\pi}\int dx'\,dy'\,\frac{\tilde{f}(x',y')}{az-\eta}=\frac{1}{b}\,(\partial_z^{-1}\tilde{f}\,)(az),\end{aligned}$$

where we have used the change of variables $x'+iy'=a\zeta$ and $x'-iy'=b\bar{\zeta}$ inside the integral, so that the Jacobian $\partial(x,y)/\partial(x',y')$ has the value $(ab)^{-1}$. Similarly,

$$(\partial_{\bar z}^{-1} f\,)(\bar z) = \frac{1}{a}\,\partial_{\bar z}^{-1}\tilde f(b\bar z). \qquad \square$$

Remark 3: An interesting feature of this transformation is that in order to consider the reductions of the systems, the substitution $r=\epsilon\bar q$ in the case of the second-order systems or $r=-q$ in the case of the third-order systems must be done after the change of variables.

B. The Miura map between NVN and modified NVN

Proposition 9: If v is a solution of (61) and u is defined by

$$u = v\,\partial_{\bar z}^{-1} v_z - v_z, \tag{65}$$

then u is a solution of (60).

Remark 4: The same transformation holds in the case of the systems (11) with z substituted for ξ and $\bar z$ substituted for η.

Proof: This proof is a generalization to two dimensions of the following one-dimensional argument to derive the Miura map for the KdV equation. Let $v=\Psi_1/\Psi_2$ with Ψ_1 and Ψ_2 such that

$$\left[\begin{pmatrix}\partial_z & 0\\ 0 & \partial_{\bar z}\end{pmatrix}+\begin{pmatrix}0 & u\\ 1 & 0\end{pmatrix}\right]\begin{pmatrix}\Psi_1\\ \Psi_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}. \tag{66}$$

Define (Φ_1,Φ_2) solutions to

$$\left[\begin{pmatrix}\partial_z & 0\\ 0 & \partial_{\bar z}\end{pmatrix}+\begin{pmatrix}0 & v\\ -v & 0\end{pmatrix}\right]\begin{pmatrix}\Phi_1\\ \Phi_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}. \tag{67}$$

This can be written as

$$\Psi_2\Phi_{1z}+\Psi_1\Phi_2=0, \tag{68a}$$

$$\Psi_2\Phi_{2\bar z}-\Psi_1\Phi_1=0. \tag{68b}$$

Differentiating (68a) with respect to $\bar z$ and (68b) with respect to z we obtain

$$\Psi_{2\bar z}\Phi_{1z}+\Psi_2\Phi_{1z\bar z}+\Psi_{1\bar z}\Phi_2+\Psi_1\Phi_{2\bar z}=0,$$

$$\Psi_{2z}\Phi_{2\bar z}+\Psi_2\Phi_{2z\bar z}-\Psi_{1z}\Phi_1-\Psi_1\Phi_{1z}=0. \tag{69}$$

From (67)

$$\Phi_{1z\bar z}=(-v\Phi_2)_{\bar z}=-v_{\bar z}\Phi_2-v\Phi_{2\bar z}=-v_{\bar z}\Phi_2-v^2\Phi_1,$$

$$\Phi_{2z\bar z}=(v\Phi_1)_z=v_z\Phi_1+v\Phi_{1z}=v_z\Phi_1-v^2\Phi_2.$$

Using this and (67) to substitute all Φ_{1z} and $\Phi_{2\bar z}$, (69) can be written as

$$-v\Psi_{2\bar z}\Phi_2-v_{\bar z}\Psi_2\Phi_2-v^2\Phi_2\Phi_1+\Psi_{1\bar z}\Phi_2+v\Psi_1\Phi_1=0,$$

$$v\Psi_{2z}\Phi_1+v_z\Psi_2\Phi_1-v^2\Psi_2\Phi_2-\Psi_{1z}\Phi_1+v\Psi_1\Phi_2=0,$$

or

$$\begin{aligned}(-v^2\Psi_2+v\Psi_1)\Phi_1-(v\Psi_{2\bar z}+v_{\bar z}\Psi_2-\Psi_{1\bar z})\Phi_2=0,\\ (v\Psi_{2z}+v_z\Psi_2-\Psi_{1z})\Phi_1+(-v^2\Psi_2+v\Psi_1)\Phi_2=0.\end{aligned} \tag{70}$$

However,

$$v^2\Psi_2 - v\Psi_1 = \left(\frac{\Psi_1}{\Psi_2}\right)^2 \Psi_2 - \frac{\Psi_1}{\Psi_2}\Psi_1 = \frac{\Psi_1^2}{\Psi_2} - \frac{\Psi_1^2}{\Psi_2} = 0$$

and $\Psi_1 = v\Psi_2$, so $\Psi_{1\bar{z}} = v_{\bar{z}}\Psi_2 + v\Psi_{2\bar{z}}$. Therefore (70) is always satisfied.

The reasoning above shows that taking $v = \Psi_1/\Psi_2$ where (Ψ_1, Ψ_2) solves the system (66) with u a solution to (60) imposes no condition on Ψ_1 and Ψ_2, that is that this is a compatible procedure; here (Φ_1, Φ_2) solve (67) with $v = \Psi_1/\Psi_2$. Now we derive the relation between u and v. From (66) we find that $u = -\Psi_{1z}/\Psi_2$. Then,

$$v_z = \left(\frac{\Psi_1}{\Psi_2}\right)_z = \frac{\Psi_{1z}}{\Psi_2} - \frac{\Psi_1}{\Psi_2}\frac{\Psi_{2z}}{\Psi_2} = -u - v\frac{\Psi_{2z}}{\Psi_2} \tag{71}$$

or

$$\frac{\Psi_{2z}}{\Psi_2} = -\frac{u + v_z}{v}. \tag{72}$$

Differentiating (72) with respect to $\bar{z}$,

$$\frac{\Psi_{2z\bar{z}}}{\Psi_2} - \frac{\Psi_{2z}}{\Psi_2}\frac{\Psi_{2\bar{z}}}{\Psi_2} = -\left(\frac{u + v_z}{v}\right)_{\bar{z}}. \tag{73}$$

Substituting $\Psi_{2\bar{z}z} = -\Psi_1$ from (66) on the left hand side and using (71)

$$\text{lhs} = -\frac{\Psi_{1z}}{\Psi_2} + \frac{\Psi_{2z}}{\Psi_2}\frac{\Psi_1}{\Psi_2} = -v_z.$$

Therefore (73) can be written as

$$u = v\partial_{\bar{z}}^{-1}v_z - v_z.$$

□

V. BI-HAMILTONIAN STRUCTURE

We omit in this section the preliminary material and refer the reader to the paper by Fokas and Gel'fand.[12] We give the results on the bi-Hamiltonian nature of the DS-type equation that are not of the degenerate ($\alpha = \beta$) kind.

The Hamiltonian approach to integrability of nonlinear evolution equations in $1+1$, in one spatial and one temporal dimension, has its origin in the works of Zakharov and Fadeev[13,14] and Gardner[15] who interpreted the KdV equation

$$u_t = u_{xxx} + 6uu_x$$

as a completely integrable Hamiltonian system in an infinite-dimensional phase space. It was shown in Ref. 13 that the inverse spectral method yields action-angle variables for this equation. Hierarchies of infinitely many commuting vector fields and constants of motion in involution for the KdV equation were constructed by Lax[16,17] and by Gel'fand and Dikii[18–20] using the equation for the resolvent of the Schrödinger operator (which is related to the equation for the squared eigenfunctions). The approach of Lax is based on what became later known as the Lenard scheme (see below), while Gel'fand and Dikii obtained the constants of motion explicitly in terms of fractional powers of the Schrödinger operator, using what is known as the resolvent scheme.

In this work we use the fact that the squared eigenfunction operator can be factorized in terms of two compatible Hamiltonian operators to compute the pair of Hamiltonian operators associated to systems of the DS type.

We consider the ring $\mathscr{R}$ consisting of smooth functions of the basic dynamical variables $u(x,y,t)$, of its x and y derivatives, and of primitives (integrals) with respect to ∂_x of smooth functions of u and of $\partial^{\alpha+\beta}/\partial_x^\alpha \partial_y^\beta$. We assume that the operator ∂_x^{-1} is skew-symmetric, for example, if the dynamical variables decay fast enough at infinity, $\partial_x^{-1}=\frac{1}{2}(\int_{-\infty}^{x}-\int_{x}^{\infty})$. The inclusion of ∂^{-1} in the construction of this ring is a necessary consequence of the fact that the bi-Hamiltonian equations in $2+1$ are nonlocal equations. With the ring $\mathscr{R}$ we consider the ring $\mathscr{R}\{\partial_y\}$ of formal pseudodifferential operators in ∂_y with coefficients in $\mathscr{R}$. The multiplication law is uniquely defined by $\partial_y \cdot a = a\partial_y + (\partial_y a)$, $\partial_y^{-1}\cdot a = a\partial_y^{-1} - (\partial_y a)\partial_y^{-2} + (\partial_y^2 a)\partial_y^{-3} - \cdots$, $a \in \mathscr{R}$. Under this multiplication, $\mathscr{R}\{\partial_y\}$ becomes an associative ring.

Vector fields, denoted by v, are elements of $\mathscr{R}\{\partial_y\}$ which are equipped with the commutator

$$L_{v_1} v_2 = v_2' v_1 - v_1' v_2, \quad v_1, v_2 \in \mathscr{R}\{\partial_y\}.$$

We refer the reader to Ref. 12 for the construction of the complex Ω^k of differential forms and proceed to review the *square eigenfunction* method of deriving the Hamiltonian operator for $1+1$ and $2+1$ systems.

The equation $Lv=0$ can be written in the form $\Psi = W\Psi$, where W is an operator in ∂_y. We can define the adjoint of this equation and obtain

$$\Phi = [W, \Phi], \Phi = \Psi\hat{\Psi}. \tag{74}$$

Hamiltonian operators can be used to derive integrable Hamiltonian equations via the *Lenard scheme*.

Proposition 10: The Lenard scheme.[12] *Let Θ and Λ be two compatible Hamiltonian operators $\Omega^1 \to V$ and let v_1 be a vector field in V such that*

$$v_1 = \Theta\xi_0 = \Lambda\xi_1, \quad \xi_0, \xi_1 \in \Omega^1, \quad d\xi_0, d\xi_1 = 0.$$

Consider the hierarchy

$$v_i = \Theta\xi_{i-1} = \Lambda\xi_i, \quad i = 1,2,...,$$

and assume that if some $\xi \in \Omega^1$ is closed on the image of Λ, then $d\xi = 0$. Then,
(i) all ξ_i are closed, i.e., $d\xi_i = 0$;
(ii) all v_i commute in V; and
(iii) all ξ_i commute.

Thus, if $\xi_i = d\uparrow f_i\downarrow$, then

$$\{f_i, f_j\}_\Theta = \{f_i, f_j\}_\Lambda = 0,$$

and $u_t = v_i$, $i=1,2,...$, is a completely integrable Hamiltonian system.

A. Bi-Hamiltonian structures for equations of the DS type

Consider the wave function Ψ and the Lax equation

$$\Psi_x = W\Psi, \quad \text{with} \quad W = -J\partial_y - P - ikJ,$$

and its *adjoint* modified version

$$\hat{\Psi}_x = -\hat{\Psi}AWA^{-1} \quad \text{with} \quad A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}.$$

Then the squared eigenfunction $\Phi \equiv \Psi\hat{\Psi}$ satisfies the equation

$$\Phi = W\Phi - \Phi AWA^{-1}.$$

Therefore, the operator

$$\partial_x + \cdot AWA^{-1} - W = \partial_x + \cdot (-J\partial_y - APA^{-1} - ikJ) + (J\partial_y + P + ikJ)\cdot$$
$$= \mathscr{R} = \partial_x \cdot + [J\partial_y, \cdot] + \{P\cdot - \cdot (APA^{-1})\} + ik[J, \cdot]$$

is Hamiltonian. This suggests the following two compatible Hamiltonian operators:

$$\Theta \cdot = \partial_x \cdot + [J\partial_y, \cdot] + \{P\cdot - \cdot (APA^{-1})\}, \quad \Lambda \cdot = [iJ, \cdot]. \tag{75}$$

Applying (75) to a vector $v = \begin{pmatrix} v^{(11)} & v^{(12)} \\ v^{(21)} & v^{(22)} \end{pmatrix}$ one obtains

$$\Theta v = \begin{pmatrix} 2v_\xi^{(11)} + qv^{(21)} - \dfrac{a}{b} v^{(12)} r & 2v_\xi^{(12)} + 2v^{(12)}v^{(22)}\partial_y + qv^{(22)} - \dfrac{a}{b} v^{(11)} q \\ rv^{(11)} - \dfrac{a}{b} v^{(22)} r + 2v_\eta^{(21)} - 2v^{(21)}v^{(22)}\partial_y & 2v_\eta^{(22)} + rv^{(12)} - \dfrac{a}{b} v^{(21)} q \end{pmatrix},$$

$$\Lambda v = \begin{pmatrix} 0 & 2iv^{(12)} \\ -2iv^{(21)} & 0 \end{pmatrix}.$$

Proposition 11: The second- and third-order systems possess a bi-Hamiltonian structure with the operators (75).

Proof: We apply Proposition 10 and note that the combination $i\Theta v_2 - 2\Theta v_1' - i\Theta v_0''$ is a function (as opposed to a pseudodifferential operator). This combination yields the right-hand side of the second-order system (6) with boundary functions equal to zero and $a^2 = \beta + \alpha$ and $b^2 = \beta - \alpha$. □

Remark 5: If a or b is zero (that is if $\alpha = \pm\beta$), then the above construction breaks down. The problem of computing the bi-Hamiltonian structure for the Fokas-type systems is, to the best of our knowledge, not solved.

ACKNOWLEDGMENT

I would like to thank Thanasis Fokas for his guidance and encouragement throughout the course of this work.

[1] A. S. Fokas, ''On the simplest integrable equation in $2+1$,'' Inverse Probl. **10**, L19–L22 (1994).
[2] A. Folguera, ''Second and third order systems of integrable equations of the Davey–Stewartson type,'' Ph.D. thesis, University of Illinois, 1997.
[3] A. S. Fokas and A. Folguera, ''Davey–Stewartson type equations of the 2nd and 3rd order: Dromion solutions,'' in preparation.
[4] D. J. Benney and G. J. Roskes, ''Wave instabilities,'' Stud. Appl. Math. **48**, 377–385 (1969).
[5] A. Davey and K. Stewartson, ''On three-dimensional packets of surface waves,'' Proc. R. Soc. London, Ser. A **338**, 101–110 (1974).
[6] E. I. Shul'man, ''On the integrability of equations of Davey–Stewartson type,'' Theor. Math. Phys. **56**, 720–724 (1984).
[7] A. S. Fokas and P. M. Santini, ''Recursion operators and bi-Hamiltonian structures in multidimensions. I.,'' Commun. Math. Phys. **115**, 375–419 (1988).
[8] F. Pempinelli, M. Boiti, and P. C. Sabatier, ''First and second order nonlinear evolution equations from an inverse spectral problem,'' Inverse Probl. **9**, 1–37 (1993).
[9] M. D. Pochinaiko, *Spectral Theory of Differential-Operator Equations* (Inst. Math., Kiev, 1986).
[10] L. V. Bogdanov, ''The Veselov–Novikov equation as a natural two-dimensional generalization of the Korteweg–de Vries equation,'' Teor. Mat. Fiz. **70**(2), 309–314 (1987).
[11] B. G. Konopelchenko, *Introduction to Multidimensional Integrable Equations* (Plenum, New York, 1992).
[12] A. S. Fokas and I. M. Gel'fand, ''Bi-Hamiltonian structures and integrability,'' in *Important Developments in Soliton Theory* (Springer-Verlag, New York, 1993), pp. 259–282.
[13] V. E. Zaharov and L. D. Faddeev, ''The Korteweg–de Vries equation is a fully integrable Hamiltonian system,'' Funkcional. Anal. i Prilož**en** **5**(4), 18–27 (1971).
[14] L. D. Fadeev, ''Symplectic structure and quantization of the Einstein gravitation theory,'' Actes du Congres International des Mathematiciens, **3**, 35–39 (1970).
[15] C. S. Gardner, ''Korteweg–de Vries equation and generalization. IV. The Korteweg–de Vries equation as a Hamiltonian system,'' J. Math. Phys. **12**, 1548–1551 (1971).
[16] P. D. Lax, ''Integrals of nonlinear equations of evolution and solitary waves,'' Commun. Pure Appl. Math. **21**, 467–490 (1968).

[17] P. D. Lax, ''Almost periodic solutions to the KdV equation,'' SIAM (Soc. Ind. Appl. Math.) Rev. **18**(3), 351–375 (1976).
[18] I. M. Gel'fand and L. A. Dikii, ''Asymptotic properties of the resolvent of Sturm–Liouville equations, and the algebra of Korteweg–de Vries equations,'' Usp. Mat. Nauk **30**(5), 67–100 (1975).
[19] I. M. Gel'fand and L. A. Dikii, ''A Lie algebra structure in the formal calculus of variations,'' Funct. Anal. Appl. **10**(1), 18–25 (1976).
[20] I. M. Gel'fand and L. A. Dikii, ''Fractional powers of operators and Hamiltonian systems,'' Funct. Anal. Appl. **10**(4), 13–29 (1976).

Irreducible representations of the Heisenberg algebra in Krein spaces

M. Mnatsakanova
Institute for Nuclear Physics of Moskow State University, Moskow, Russia

G. Morchio
Dipartimento di Fisica dell'Università and INFN, Pisa, Italy

F. Strocchi
Scuola Normale Superiore and INFN, Pisa, Italy

Yu. Vernov
Institute for Nuclear Research of Russian Academy of Sciences, Moskow, Russia

(Received 9 June 1997; accepted for publication 21 January 1998)

The representations of the Heisenberg algebra in Krein spaces, more generally in weakly complete inner product spaces, are classified under general regularity and irreducibility conditions. Besides the Fock representation, two other types appear; one with negative, the other with a two-sided discrete spectrum of the number operator. © *1998 American Institute of Physics.* [S0022-2488(98)00405-8]

I. INTRODUCTION

The main motivation of this paper is the classification of the representations of the canonical commutation relations in Krein spaces; this problem arises for the mathematical foundations of the quantization in indefinite metric spaces proposed by Dirac and Heisenberg[1,2] and crucially enters in the local formulation of gauge QFTs,[3,4] where locality and covariance of the gauge fields are incompatible with positivity of the inner product. In this paper we start by looking at the case of finite degrees of freedom and classify the irreducible representations of the Heisenberg algebra in Krein spaces, actually in all weakly complete indefinite inner product spaces.

The Heisenberg algebra is the algebraic structure at the basis of Quantum Mechanics (QM), and it is of great interest, not only from a physical point of view, but also as a basic mathematical structure. It is the Lie algebra generated by the so-called canonical variables, q_i, p_i, $i=1,\dots n$, satisfying the following commutation relations (CCR in unbounded form):[5–7]

$$[p_i,q_j]=-i\delta_{i,j},\quad [q_i,q_j]=0=[p_i,p_j]. \tag{1.1}$$

The foundations of QM rely on the analysis of the representations of such an algebra and great clarification of the mathematical structures of QM has been obtained by this analysis.[8–10]

From a technical point of view, nontrivial problems arise because the commutation relations (1.1) imply that q_i and p_i cannot be represented by bounded operators (in fact, the corresponding group is not compact); as a consequence, domain questions become important, the concept of irreducibility is delicate (see below), etc. For these reasons, one usually considers the algebra $\mathscr{A}$ generated by the Weyl exponentials $\exp i\alpha q$, $\exp i\beta p$, $\alpha,\beta\in\mathbb{R}^n$ (briefly, the Weyl algebra). This amounts to consider, instead of the Lie algebra (1.1), the associated group (the Heisenberg group), and to analyze its representations. A systematic classification of such representations in a Hilbert space, under general regularity conditions, is available.[9,10] It should, however, be remarked that the choice of the Weyl algebra, generated by rather special functions of the canonical variables q_i,p_i, gives rise to problems, e.g., for the algebraic setting of the dynamics, since, in general, the time evolution does not leave the Weyl algebra stable.[11] Actually, most of the standard treatments of QM models deal with the canonical variables directly, and therefore it is of some interest to consider the problem of analyzing the representations of the Heisenberg algebra directly.

This strategy is actually the most convenient one if one wants to consider the case of the representations of the CCRs in indefinite metric spaces, in particular in Krein spaces, since in this

 © 1998 American Institute of Physics

case q_i and p_i are not self-adjoint operators, the construction of the Weyl exponentials is, in general, problematic and requires domain and regularity assumptions, and the advantage is not significant since the Weyl exponentials need not be bounded operators.

A notable case of an analysis of the Hilbert space representations of the CCRs based on the Heisenberg algebra is the theorem by Rellich and Dixmier,[12–14] which overcomes the difficulties mentioned above, by assuming that $N=1/2(p^2+q^2-1)$ is essentially self-adjoint on a dense domain invariant under the closure of p and q. The result of this paper is a generalization of the theorem by Rellich and Dixmier, which covers the case of representations of the Heisenberg algebra in Krein spaces, together with a discussion of irreducibility.

Crucial technical points of our approach are the following: (i) a proper position of the irreducibility problem in a situation in which the algebra consists of unbounded operators; and (ii) regularity assumptions in the same spirit of Rellich and Dixmier. The result is that the spectrum of $N=1/2(p^2+q^2-1)$ is discrete (but not semibounded, in general), and in terms of it one has a classification of the irreducible representations of the Heisenberg algebra in Krein spaces (more generally, in weakly complete nondegenerate inner product spaces). Up to isometries between indefinite spaces, such representations are completely described by the action of the algebra on the (weakly) dense domain generated by the eigenvector of N.

II. GENERAL SETTING

*Definition 1: The Heisenberg algebra $\mathscr{H}$ for n degrees of freedom is the complex * algebra generated by p_i, q_i, $i=1,\ldots n$, with the relations (1.1) and $p_i^*=p_i$, $q_i^*=q_i$.*

In the following, we will use the standard notation

$$a_i^*\equiv 1/\sqrt{2}(q_i-ip_i),\quad a\equiv 1/\sqrt{2}(q_i+ip_i),\quad N=\sum_i a_i^* a_i .$$

The Heisenberg relations become

$$[a_i,a_j^*]=\delta_{i,j},\quad [a_i,a_j]=[a_i^*,a_j^*]=0.$$

For simplicity, in the following we restrict our attention to the Heisenberg algebra for one degree of freedom, the extension to a finite number of degrees of freedom being straightforward.

A Krein space $\mathscr{K}$[15,16] is a vector space with indefinite inner product $\langle\cdot,\cdot\rangle$ and a linear operator J, such that

(i) $J^2=1$,
(ii) $(x,x)\equiv\langle x,Jx\rangle\geqslant 0,\quad \forall x\in\mathscr{K}$,
(iii) $\mathscr{K}$ is complete in the norm $\|x\|=(x,x)^{1/2}$.

We recall that in a Krein space the weak Krein topology, defined by the inner product $\langle\cdot,\cdot\rangle$, coincides with the weak Hilbert topology, defined by the Hilbert scalar product $(\cdot,\cdot)$. In the following, we will only consider separable Krein spaces.

For any densely defined operator A on $\mathscr{K}$, the adjoint $A^\dagger$ is defined, as in the Hilbert space case, by

$$\langle A^\dagger x,y\rangle=\langle x,Ay\rangle,\quad \forall y\in\mathscr{D}_A .$$

An operator A is closable (in the strong Hilbert topology) iff its adjoint $A^\dagger$ is densely defined, and then its closure $\bar{A}$ coincides with $(A^\dagger)^\dagger$. An operator A is said to be weakly closable if $x_n\to 0$ weakly and Ax_n weakly convergent implies $Ax_n\to 0$ weakly. Since weak and strong closures of operators coincide in Hilbert spaces, they also coincide in Krein spaces.

Given Krein spaces $\mathscr{K}_1,\mathscr{K}_2$, a linear operator U, defined on a dense domain $\mathscr{D}_1\subset\mathscr{K}_1$ with range in $\mathscr{K}_2$ is called an *isometry* if $\forall x,y\in\mathscr{D}_1$,

$$\langle Ux,Uy\rangle_{\mathscr{K}_2}=\langle x,y\rangle_{\mathscr{K}_1}.$$

From the nondegeneracy of the inner product and the density of $\mathscr{D}_1$, it follows that U has an inverse, defined on its range.

Definition 2: A representation of a $*$ algebra $\mathscr{A}$ in a Krein space $\mathscr{K}$ is a $*$ morphism ρ of $\mathscr{A}$ into the algebra of operators acting on a dense invariant common domain $\mathscr{D}\subset\mathscr{K}$, namely

$$\rho(\mathscr{A})\mathscr{D}\subset\mathscr{D},$$

and, $\forall x,y\in\mathscr{D}$, $A\in\mathscr{A}$,

$$\langle x,\rho(A)y\rangle=\langle\rho(A^*)x,y\rangle. \tag{2.1}$$

Equation (2.1) implies that all the operators $\rho(A)$, $A\in\mathscr{A}$, have a densely defined adjoint $\rho(A)^\dagger$, with respect to the indefinite inner product, and $\rho(A^*)\subset\rho(A)^\dagger$; all the operators $\rho(A)$ are therefore closable. For a discussion of representations of unbounded operator algebras, see Refs. 17 and 18.

In general, the common domain of the closures $\overline{\rho(A)}$, $A\in\mathscr{A}$, may not be invariant under the operators $\overline{\rho(A)}$, and therefore it may not define a representation of $\mathscr{A}$. However, any dense subspace $\mathscr{D}'$, invariant under all the closures $\overline{\rho(A)}$, $A\in\mathscr{A}$, provides a representation of the algebra in the sense of Definition 1. In fact, Eq. (2.1) implies $\overline{\rho(A)}\subset\rho(A^*)^\dagger$, and the invariance of $\mathscr{D}'$ implies $\forall x\in\mathscr{D}'$, $\overline{\rho(B)}x\in\mathscr{D}_{\overline{\rho(A)}}\subset\mathscr{D}_{\rho(A^*)^\dagger}$, so that $\forall y\in\mathscr{D}$, $x_n\in\mathscr{D}$, with $x_n\to x$ and $\rho(AB)x_n$ convergent,

$$\begin{aligned}\langle y,\overline{\rho(AB)}x\rangle&=\lim_n\langle y,\rho(AB)x_n\rangle=\lim_n\langle\rho(B^*)\rho(A^*)y,x_n\rangle\\&=\langle\rho(B^*)\rho(A^*)y,x\rangle\\&=\langle\rho(A^*)y,\overline{\rho(B)}x\rangle=\langle y,\overline{\rho(A)}\ \overline{\rho(B)}x\rangle.\end{aligned} \tag{2.2}$$

Thus, any representation, in the sense of Definition 2, is naturally associated with all the representations defined by dense domains invariant under the closures $\overline{\rho(A)}$, $A\in\mathscr{A}$. Since such representations are uniquely determined by ρ, it is convenient to introduce a notion of equivalence that identifies representations determined one from the other in the above sense.

Definition 3: Given two representations ρ_1, ρ_2 of a $*$ algebra $\mathscr{A}$ in Krein spaces $\mathscr{K}_1,\mathscr{K}_2$, with domains $\mathscr{D}_1,\mathscr{D}_2$, respectively, ρ_1 is said to be in the closure of ρ_2, briefly $\rho_1\subset\overline{\rho_2}$, if there exists an isometry $U:\mathscr{D}_1\to\mathscr{K}_2$, such that

$$U\rho_1(A)U^{-1}\subset\overline{\rho_2(A)},\quad\forall A\in\mathscr{A}.$$

The two representations are said to be equivalent if $\rho_1\subset\overline{\rho_2}$ and $\rho_2\subset\overline{\rho_1}$.
Clearly, $\rho_1\subset\overline{\rho_2}$ implies that $U\mathscr{D}_1$ is an invariant domain for the closures $\overline{\rho_2(A)}$.

We stress that for the analysis of the representations it is essential to consider the closures of the operators, since otherwise the restrictions of the same representation to different domains would appear as different, in general inequivalent, representations. Such phenomena occur in general when representations of unbounded algebras are considered,[17,18] but their control is more difficult in Krein spaces, due to the failure of some of the corresponding Hilbert space constructions.

According to the discussion given above, we will consider representations of the Heisenberg algebra that satisfy an additional *regularity condition*.

Definition 4: A representation ρ of the Heisenberg algebra $\mathscr{H}$ on a domain $\mathscr{D}$ will be called regular iff it satisfies the following condition:

(A) *(Regularity condition). The operator $\rho(N)=\frac{1}{2}\rho(q^2+p^2-1)$ is the weak derivative, on $\mathscr{D}$, of a one-parameter group $U(s)$, $s\in\mathbb{R}$:*

$$\frac{d}{ds}\langle y,U(s)x\rangle|_{s=0}=i\langle y,\rho(N)x\rangle,\quad\forall y\in\mathscr{K},\quad x\in\mathscr{D}, \tag{2.3}$$

with

$$U(s)\mathscr{D}\subset\mathscr{D}, \tag{2.4}$$

$$U(s)U(t)=U(s+t), \quad \text{on } \mathscr{D}. \tag{2.5}$$

Condition (A) implies

$$\langle U(s)x,U(s)y\rangle=\langle x,y\rangle, \quad \forall x,y\in D, \tag{2.6}$$

and therefore $U(-s)\subset U(s)^{\dagger}$, so that $U(s)$, $s\in\mathbb{R}$, are closable operators. Moreover, weakly on $\mathscr{D}$,

$$\frac{d}{ds}\,(U(s)\rho(a)U(-s))=iU(s)\rho([a^*a,a])U(-s)=-iU(s)\rho(a)U(-s),$$

so that

$$U(s)\rho(a)U(-s)=e^{is}\rho(a), \quad U(s)\rho(a^*)U(-s)=e^{is}\rho(a^*), \quad \text{on } \mathscr{D}. \tag{2.7}$$

Therefore, $U(s)$ implements the one-parameter group of ''gauge'' automorphisms of $\mathscr{H}$,

$$\gamma^s(a)=e^{-is}a, \quad \gamma^s(a^*)=e^{is}a^*. \tag{2.8}$$

Condition (A), Eq. (2.3), also implies that $U(s)$ is weakly continuous at the origin, and then, by the group law, for all $s\in\mathbb{R}$. Equations (2.3)–(2.5) immediately imply that

$$\frac{d}{ds}\,U(s)=i\rho(N)U(s), \quad \forall s\in\mathbb{R}, \tag{2.9}$$

as a weak derivative, on $\mathscr{D}$.

By Definition 2, a regular representation ρ defines an *extended algebra* $\mathscr{H}_{\text{ext}}$, as the concrete $*$ algebra generated by $\rho(\mathscr{H})$ and the group $U(s)$, $s\in\mathbb{R}$, satisfying $U(s)^*=U(-s)$ and the relations (2.7). In the following, when discussing a fixed representation ρ, we will denote, for simplicity, $\rho(A)$ by A. We will denote by $\mathscr{D}_c$ the common domain of the closure of the operators of $\mathscr{H}_{\text{ext}}$; we will omit the bar that denotes the closure, when no ambiguity may arise.

The regularity condition plays the rôle of the Rellich–Dixmier condition (essential self-adjointness of N on $\mathscr{D}$), and, in fact, in the ordinary Hilbert space case, the stability of $\mathscr{D}$ under $U(s)$ implies the essential self-adjointness of its generator N on $\mathscr{D}$. A stronger form of the Rellich–Dixmier condition is justified here because, without positivity of the inner product, the $U(s)$ need not to be continuous everywhere defined operators.

The next issue is a proper notion of irreducibility. As it is well known, the concept of irreducibility is delicate in the case of algebras of unbounded operators. Algebraic reducibility is a very weak property for algebras of *unbounded* operators, but also the existence of a proper closed subspace $\mathscr{K}_1\subset\mathscr{K}$ containing a common linear domain $\mathscr{D}_1$ invariant under $\mathscr{H}$ *(topological reducibility)*[19] is too weak, in particular for the Heisenberg algebra; in fact, in the ordinary Schrödinger representation in $L^2(\mathbb{R})$, which is irreducible for the Weyl algebra, the Heisenberg algebra, defined on the dense domain of C^∞ functions $f(x)$ of a fast decrease, with all derivatives vanishing at $x=0$, is topologically *reducible*, since the subdomain of the function with positive support is invariant and not dense (see the Appendix for more details).

A more relevant notion, which uses the regularity properties of the representation, is that of topological irreducibility *for the extended algebra* $\mathscr{H}_{ext}$:

Definition 5: A regular representation of $\mathscr{H}$ *is called domain irreducible if there is no nontrivial domain* $\mathscr{D}_1\subset\mathscr{D}_c$ *(the common dense domain of the weak closures of the elements of* $\mathscr{H}_{ext}$*), such that*

(i) $\mathscr{D}_1$ *is invariant under* $\mathscr{H}_{ext}$*; and*

(ii) $\mathscr{D}_1$ *is not dense in* $\mathscr{K}$.

The above definition recognizes the crucial role played by the ''integrated'' elements $U(s)$, very close to the role of the one-parameter group generated by the Nelson elliptic operator Δ in the classification of the Hilbert space representations of unbounded Lie algebras.[19,20]

It is immediate to see that domain irreducibility in the above sense is equivalent to cyclicity of all vectors x belonging to subspaces of $\mathscr{D}_c$ invariant under $\mathscr{H}_{\text{ext}}$, i.e., $\mathscr{H}_{\text{ext}}x$ is dense in $\mathscr{K}$. In fact, if x belongs to an invariant domain, by applying elements of $\mathscr{H}_{\text{ext}}$ to x, one gets an invariant subspace of $\mathscr{D}_c$, which must be dense, by Definition 5 (the converse is obvious).

An alternative definition of irreducibility is the one related to the triviality of the commutant, and it is often used in the analysis of unbounded operator algebras.[19] Given the representation of $\mathscr{H}$ on the domain $\mathscr{D}$, one could formulate commutant irreducibility by asking that any (weakly closable) operator $\mathscr{B}$ that leaves $\mathscr{D}$ invariant and commutes on $\mathscr{D}$ with $\mathscr{H}$ and $U(s)$, is a multiple of the identity.

The relation between this notion and Definition 5 does not appear as trivial as in the bounded positive case, because a spectral theory of unbounded operators in Krein spaces is not available. Moreover, reduction may take place only after the operators have been extended by closure, and, in fact a more relevant notion of commutant irreducibility must be given in terms of the extended domain $\mathscr{D}_c$.

Definition 6: A regular representation of $\mathscr{H}$ is said to be commutant irreducible if whenever a (densely defined) weakly closable operator B has the properties

(i) *B leaves invariant a nontrivial domain $\mathscr{D}_1 \subset \mathscr{D}_c$, invariant under $\mathscr{H}_{\text{ext}}$; and*

(ii) *B commutes with $\mathscr{H}_{\text{ext}}$ on $\mathscr{D}_1$,*

then B is a multiple of the identity.

It is easy to see that commutant irreducibility in the sense of Definition 6 implies domain irreducibility (Definition 5).

In Sect. III, we analyze irreducible representations in the sense of Definition 6. In Sect. IV, the same classification will be obtained (Theorem 13), assuming only domain irreducibility (Definition 5), but using additional regularity assumptions on $U(s)$.

III. REGULAR COMMUTANT IRREDUCIBLE REPRESENTATIONS OF THE HEISENBERG ALGEBRA IN KREIN SPACES

The general setting discussed in the previous section allows for the following characterization of the regular (Definition 4) irreducible (Definition 6) representations of the Heisenberg algebra. The proof essentially relies on the fact that commutant irreducibility implies the compactness of the group $U(s)$.

Theorem 7: *All regular, commutant irreducible representations of the Heisenberg algebra $\mathscr{H}$ in a Krein space $\mathscr{K}$ fall in the equivalence classes (Definition 3) of the following representations.*

(1) π_+: *The operators $U(s)$, $s \in \mathbb{R}$ have a common complete set of eigenvectors ψ_n, $n \in \mathbb{N}$, with*

$$U(s)\psi_n = e^{ins}\psi_n, \quad \langle \psi_l, \psi_m \rangle = \eta_0 \delta_{l,m}, \quad \eta_0 = \pm 1, \quad J = \eta_0 \mathbb{1}. \tag{3.1}$$

The representation is also characterized by the existence of a vector $\psi_0 \in \mathscr{D}_c$ (Fock vector), satisfying

$$a\psi_0 = 0. \tag{3.2}$$

(2) π_-: *The operators $U(s)$ have a common complete set of eigenvectors ψ_n, $n = -1, -2, \ldots$, satisfying Eq. (3.1), with*

$$\langle \psi_l, \psi_m \rangle = \eta_1 (-1)^m \delta_{l,m}, \quad \eta_1 = \pm 1, \quad J = \eta_1 (-1)^N.$$

The representation is also characterized by the existence of a vector $\psi_{-1} \in \mathscr{D}_c$ (anti-Fock vector), satisfying

$$a^* \psi_{-1} = 0. \tag{3.3}$$

(3) π_θ, $\theta \in (0,1)$: *The operators $U(s)$ have a common complete set of eigenvectors ψ_n, $n \in \mathbb{Z}$, with*

$$U(s)\psi_n = e^{i(\theta+n)s}\psi_n, \tag{3.4}$$

$$\langle \psi_l, \psi_m \rangle = \eta_\theta \delta_{l,m}, \quad m \geqslant 0, \quad \langle \psi_l, \psi_m \rangle = \eta_\theta \delta_{l,m}(-1)^{m+1}, \quad m < 0,$$

$$J\psi_m = \eta_\theta \psi_m, \quad m \geqslant 0, \quad J\psi_m = \eta_\theta(-1)^{m+1}\psi_m, \quad m < 0, \quad \eta_\theta = \pm 1 \tag{3.5}$$

(''Dirac sea states'').

In all the cases, the domain $\mathscr{D}$ can be taken as the linear span of the vectors ψ_n; the (closure of the) operators $U(s)$, $s \in \mathbb{R}$, commute with the metric operator J and are therefore unitary, with a simple point spectrum given by Eqs. (3.1), (3.4).

The proof follows from a few lemmas.

Lemma 8: The closures of $U(s)$, $s \in \mathbb{R}$, have common eigenvectors ψ_n, $n \in \mathbb{Z}$,

$$U(s)\psi_n = e^{i(n+\theta)s}\psi_n, \tag{3.6}$$

with θ fixed in [0,1); the linear span of such vectors is dense in $\mathscr{H}$.

Proof: By Eq. (2.4), the operator $U(2\pi)$ leaves $\mathscr{D}$ invariant, and by Eq. (2.8) it commutes with $\mathscr{H}$ and obviously with $U(s)$, $\forall s \in \mathbb{R}$. Therefore, by the commutant irreducibility of the representation, $U(2\pi)$ is a multiple of the identity, and actually a phase, by Eq. (2.6):

$$U(2\pi) = e^{2\pi i\theta}, \quad \theta \in [0,1). \tag{3.7}$$

Thus, the one-parameter group,

$$V(s) \equiv U(s)e^{-is\theta}, \quad s \in \mathbb{R}, \tag{3.8}$$

is a $U(1)$ group:

$$V(2\pi) = 1, \tag{3.9}$$

and $\forall \psi \in \mathscr{D}$, the vector-valued function $V(s)\psi$ is periodic and weakly continuous, by condition *A*. Hence, the Riemann sums,

$$\psi_n^{(\epsilon)} \equiv \frac{\epsilon}{2\pi} \sum_{s_k = k\epsilon, 0 \leqslant k < 2\pi/\epsilon} e^{-ins_k} V(s_k)\psi, \tag{3.10}$$

converge weakly, as $\epsilon \to 0$, to the (weak) integral,

$$\psi_n \equiv \frac{1}{2\pi} \int_0^{2\pi} e^{-ins} V(s)\psi ds, \tag{3.11}$$

which defines a vector in $\mathscr{H}$, by weak completeness. Moreover, $\forall t \in \mathbb{R}$, $V(t)\psi_n^{(\epsilon)}$ also converges weakly as $\epsilon \to 0$:

$$V(t)\psi_n^{(\epsilon)} = \frac{\epsilon}{2\pi} \sum_{s_k = k\epsilon, 0 \leqslant k < 2\pi/\epsilon} e^{-ins_k} V(t+s_k)\psi = \frac{\epsilon}{2\pi} \sum_{s_k} e^{-ins_k} e^{int} V(s_k)\psi \to e^{int}\psi_n, \tag{3.12}$$

where the periodicity of $V(s)$ has been used. Thus, all the vectors ψ_n are in the domain of the closures of $V(t)$, and

$$V(s)\psi_n = e^{ins}\psi_n. \tag{3.13}$$

Furthermore, the vector-valued weakly continuous periodic functions $V(s)\psi$ can be expressed (pointwise in s) by their (weakly convergent) Fourier series,

$$V(s)\psi=\sum_{n\in\mathbb{Z}} e^{ins}\psi_n . \tag{3.14}$$

For $s=0$, Eq. (3.14) expresses all $\psi\in\mathscr{D}$ as weak limits of linear combinations of the vectors ψ_n, so that their linear span is dense in $\mathscr{K}$.

Lemma 9: The vectors ψ_n constructed in Lemma 8 are eigenvectors of the closure of a^*a:

$$a^*a\psi_n=(n+\theta)\psi_n . \tag{3.15}$$

Proof: With the notations of the proof of Lemma 8, by condition (A), in the sense of weak derivatives:

$$\begin{aligned}(a^*a-\theta)\psi_n^{(\epsilon)}&=\frac{-i}{2\pi}\,\epsilon\sum_{s_k} e^{-ins_k}\frac{d}{dt}\,V(t)V(s_k)\psi\Big|_{t=0}\\&=\frac{-i}{2\pi}\,\epsilon\sum_{s_k} e^{-ins_k}V(s_k)\,\frac{d}{dt}\,V(t)\psi\Big|_{t=0}\\&\to\frac{-i}{2\pi}\int_0^{2\pi}ds e^{-ins}\frac{d}{ds}\,V(s)\psi=n\,\frac{1}{2}\int_0^{2\pi}ds e^{-ins}V(s)\psi,\end{aligned}$$

where an integration by parts has been performed in the last step.

Lemma 10: The linear span $\mathscr{D}_1$ of all the vectors ψ_n constructed in Lemma 8, as ψ varies on $\mathscr{D}$, is contained in $\mathscr{D}_c$ (the common domain of the weak closures of the elements of $\mathscr{K}_{\rm ext}$), and it is left invariant by $\mathscr{K}_{\rm ext}$. The Heisenberg relations hold on $\mathscr{D}_1$, and each vector in $\mathscr{D}_1$ is cyclic for $\mathscr{K}_{\rm ext}$. $\mathscr{D}_1$ provides a regular representation ρ_1 of $\mathscr{K}$, equivalent (Definition 3) to the representation ρ on $\mathscr{D}$. For any $\psi\in\mathscr{D}$ there is an integer $\bar{n}$ such that the vectors $(a^*)^k\psi_{\bar{n}}$, $a^l\psi_{\bar{n}}$ generate a dense subspace of $\mathscr{D}_1$.

Proof: By Eq. (3.10),

$$a\psi_n^{(\epsilon)}=\frac{\epsilon}{2\pi}\sum_{s_k} e^{-ins_k}aV(k\epsilon)\psi=\frac{\epsilon}{2\pi}\sum_{s_k} e^{-ins_k}e^{is_k}V(s_k)a\psi=(a\psi)_{n-1}^{(\epsilon)}, \tag{3.16}$$

and similarly,

$$a^*\psi_n^{(\epsilon)}=(a^*\psi)_{n+1}^{(\epsilon)} . \tag{3.17}$$

Since the rhs of Eqs. (3.16) and (3.17) converge by Lemma 8, the vectors ψ_n belong to the domain of the weak closure of a and a^*, and

$$\bar{a}\psi_n=(a\psi)_{n-1}, \quad \bar{a^*}\psi_n=(a^*\psi)_{n+1} . \tag{3.18}$$

The same treatment applies to all the products of a and a^*, so that $\mathscr{D}_1\subset\mathscr{D}_c$ and $\mathscr{K}_{\rm ext}\mathscr{D}_1\subset\mathscr{D}_1$. By irreducibility, $\mathscr{K}_{\rm ext}\chi$ is dense in $\mathscr{K}$, $\forall\chi\in\mathscr{D}_1$. The Heisenberg commutation relations hold on $\mathscr{D}_1$, as a consequence of their validity on $\mathscr{D}$, by Eq. (3.18):

$$(aa^*-a^*a)\psi_n=a(a^*\psi)_{n+1}-a^*(a\psi)_{n-1}=((aa^*-a^*a)\psi)_n=\psi_n . \tag{3.19}$$

Thus, $\mathscr{D}_1$ defines a representation ρ_1 contained in the closure of ρ; moreover, Eqs. (3.18) and (3.14) imply that $\mathscr{D}$ is contained in the domain of the closures of $\rho_1(A)$, for all $A\in\mathscr{A}$, so that $\rho\subset\bar{\rho_1}$, and the two representations are equivalent.

From Eqs. (3.19) and (3.15) one easily gets

$$aa^*\psi_n=(n+1+\theta)\psi_n \tag{3.20}$$

(an equation that could have been derived directly as in Lemma 9). For each ψ in $\mathscr{D}$, by Eq. (3.14), there is at least one $\psi_{\bar{n}}$ that does not vanish, and is then cyclic. Given any product of operators a,a^* applied to $\psi_{\bar{n}}$, by using Eqs. (3.18), (3.15), and (3.20), one can write it as a constant times

$(a^*)^k\psi_{\overline{n}}$, if the number of a^* operators exceeds, by k, that of the a's, or, otherwise, as $a^k\psi_{\overline{n}}$. The cyclicity of $\psi_{\overline{n}}$ implies therefore the density in $\mathscr{K}$ of the linear span of the vectors $\{(a^*)^k\psi_{\overline{n}}, a^l\psi_{\overline{n}}, k,l\in\mathbb{N}\}$.

Proof of Theorem 7: By Lemma 10, the analysis of regular irreducible representations of $\mathscr{H}$ can be done on the domain $\mathscr{D}_1$. Given a vector $\psi_{\overline{n}}\in\mathscr{D}_1$, we consider first the case in which at least one of the sequences $a^k\psi_{\overline{n}}$ terminates. If, for some $k>0$,

$$a^k\psi_{\overline{n}}=0, \quad a^{k-1}\psi_{\overline{n}}\neq 0, \tag{3.21}$$

then, by Eqs. (3.15) and (3.18) one has

$$0=a^*a\,a^{k-1}\psi_{\overline{n}}=a^*a(a^{k-1}\psi)_{\overline{n}-k+1}=(\overline{n}-k+1+\theta)a^{k-1}\psi_{\overline{n}},$$

i.e.,

$$\overline{n}-k+1+\theta=0, \tag{3.22}$$

which implies $\theta=0$, since $\theta\in[0,1)$. Similarly,

$$(a^*)^l\psi_{\overline{n}}=0, \quad (a^*)^{l-1}\psi_{\overline{n}}\neq 0 \tag{3.23}$$

implies, by Eqs. (3.18) and (3.20),

$$\overline{n}+l+\theta=0, \tag{3.24}$$

and therefore again $\theta=0$, Eqs. (3.22) and (3.24) are incompatible, and therefore only one sequence can terminate.

Case 1: Equation (3.21) holds. Then, putting

$$\psi_0\equiv a^{k-1}\psi_{\overline{n}}=(a^{k-1}\psi)_0, \tag{3.25}$$

one has

$$a\,\psi_0=0, \tag{3.26}$$

and $\mathscr{D}_1$ is generated by the vectors $(a^*)^l\psi_0, l\geqslant 0$. (*Fock representation*). By using Eqs. (3.18), (3.20), and (3.25) one gets

$$\langle(a^*)^l\psi_0,(a^*)^m\psi_0\rangle=m!\,\delta_{lm}\langle\psi_0,\psi_0\rangle. \tag{3.27}$$

Since the linear span of the vectors $(a^*)^m\psi_0$ is dense in $\mathscr{K}$ (Lemma 10), the nondegeneracy of the inner product implies $\langle\psi_0,\psi_0\rangle\neq 0$, and Eq. (3.27) fully determines the inner product in $\mathscr{K}$, which turns out to be definite. Without loss of generality we can normalize $\langle\psi_0,\psi_0\rangle\equiv\eta_0=\pm 1$, and the Krein metric operator is then

$$J=\eta_0\mathbb{1}.$$

Apart from the sign of the inner product, the representation is the usual Fock representation in a Hilbert space.

Case 2. Equation (3.23) holds. Then, putting

$$\psi_{-1}\equiv(a^*)^{l-1}\psi_{\overline{n}}=((a^*)^{l-1}\psi)_{-1}, \tag{3.28}$$

one has

$$a^*\psi_{-1}=0, \tag{3.29}$$

and $\mathscr{D}_1$ is generated by the vectors $a^l\psi_0, l>0$ (''*anti-Fock representation*''). By using Eqs. (3.15), (3.18), and (3.28), one gets

$$\langle a^l\psi_{-1},a^m\psi_{-1}\rangle=\delta_{l,m}m!(-1)^m\langle\psi_{-1},\psi_{-1}\rangle. \tag{3.30}$$

The nondegeneracy of the inner product implies $\langle \psi_{-1}, \psi_{-1} \rangle \neq 0$, and the inner product in $\mathscr{K}$ is fully determined by Eq. (3.30), and it is not definite.

The metric operator that defines the Hilbert space structure of $\mathscr{K}$ is not uniquely determined by Eq. (3.30). There is, however, a distinguished Krein structure associated to Eq. (3.30), defined by the metric operator

$$J = \eta_1 (-1)^N, \tag{3.31}$$

where $\eta_1 \equiv \langle \psi_{-1}, \psi_{-1} \rangle = \pm 1$.

In general, the representation defined (on $\mathscr{D}_1$) by Eq. (3.30) may live in a Krein space obtained by closing $\mathscr{D}$ with a (weak) topology different from that defined by the above metric operator. However, all such Krein spaces contain dense domains $\mathscr{D}_1$, on which the corresponding representations are related by indefinite space isometries. Thus, any representation of type 2 is equivalent to the representation defined by Eq. (3.30) on the Krein space given by the metric (3.31).

Case 3. None of the sequences $a^k \psi_{\overline{n}}, (a^*)^l \psi_{\overline{n}}$ terminates. Then one of the sequences contains a vector ψ_θ with

$$a^* a \psi_\theta = \theta \psi_\theta, \tag{3.32}$$

and $\mathscr{D}_1$ is generated by the vectors $a^l \psi_\theta, (a^*)^k \psi_\theta, l, k \in \mathbb{N}$ (''Dirac sea states''). Different eigenvalues of $U(s)$ define orthogonal eigenspaces (with the indefinite inner product), and, using the commutation relations of Lemma 10, one has

$$\langle (a^*)^l \psi_\theta, (a^*)^m \psi_\theta \rangle = \delta_{l,m} (\theta+1)(\theta+2) \cdots (\theta+m) \langle \psi_\theta, \psi_\theta \rangle, \tag{3.33}$$

$$\begin{aligned} \langle a^l \psi_\theta, a^m \psi_\theta \rangle &= \delta_{l,m} \theta(\theta-1) \cdots (\theta-m+1) \langle \psi_\theta, \psi_\theta \rangle \\ &= \delta_{l,m} (-1)^{m-1} \theta(1-\theta) \cdots (m-1-\theta) \langle \psi_\theta, \psi_\theta \rangle. \end{aligned} \tag{3.34}$$

Again $\langle \psi_\theta, \psi_\theta \rangle \neq 0$ follows from nondegeneracy of the inner product, which is fully determined by Eqs. (3.33) and (3.34), and is not definite. Again, the metric operator is not uniquely determined, and a distinguished Krein structure is defined by

$$J(a^*)^m \psi_\theta = \eta_\theta (a^*)^m \psi_\theta, \quad J a^l \psi_\theta = \eta_\theta (-1)^{l-1} a^l \psi_\theta, \tag{3.35}$$

with $\eta_\theta \equiv \langle \psi_\theta, \psi_\theta \rangle = \pm 1$. The same arguments as in case 2 then apply to the equivalence of representations in different Krein spaces.

In order to control all the common eigenvectors of $U(s)$, $s \in \mathbb{R}$, assume that a vector χ, in the domain of the closure of all the operators $U(s)$, satisfies

$$U(s)\chi = \lambda(s)(\chi).$$

By weak continuity, and Eq. (2.5),

$$\lambda(s) = e^{i\mu s}.$$

From Eq. (2.6) (extended to the closure of the operators), it follows that χ is orthogonal to all the eigenvectors of $U(s)$ with different eigenvalues, and therefore, by the density of the domains $\mathscr{D}_1$ constructed above, $\exp i\mu s$ is, in the three different cases, one of the eigenvalues classified above,

$$e^{i\mu s} = e^{i(n+\theta)s}.$$

For each $\psi \in \mathscr{D}$ the set of eigenvectors constructed in Lemma 10 generates a dense subspace; since, by the argument given above, χ is orthogonal to all such eigenvectors, except for the eigenvalue $\exp i(\overline{n}+\theta)s$, it follows that χ is a multiple of the corresponding eigenvector, which is therefore unique. It also follows that all the subspaces of $\mathscr{D}_1$ introduced in Lemma 10 actually coincide with $\mathscr{D}_1$.

In order to complete the classification, it is necessary to show that indeed all the representations constructed in cases 1–3 fulfill the regularity and irreducibility conditions. The regularity is obvious, and the proof of irreducibility follows from the absence of multiplicities for the common eigenvalues of $U(s)$, $s\in\mathbb{R}$.

In fact, if a closable operator B leaves a domain $\mathscr{D}'\subset\mathscr{D}_c$ invariant, with $\mathscr{H}_{\text{ext}}\mathscr{D}'\subset\mathscr{D}'$, then for any $\psi\in\mathscr{D}'$ one can construct vectors $\psi_n^{(\epsilon)}$ as in Lemma 8, Eq. (3.10), which converge weakly to vectors ψ_n, Eq. (3.12), still belonging to an invariant domain contained in $\mathscr{D}_c$; if B commutes with $\mathscr{H}_{\text{ext}}$, then

$$B\psi_n^{(\epsilon)}\to(B\psi)_n, \quad \text{weakly.} \tag{3.36}$$

We can fix n so that ψ_n is nonvanishing; since B is (weakly) closable, ψ_n belongs to the domain of its weak closure $\bar{B}$, and

$$\bar{B}\psi_n=(B\psi)_n. \tag{3.37}$$

Both ψ_n and $(B\psi)_n$ are eigenvectors of $U(s)$, $\forall s\in\mathbb{R}$, as in Lemma 8, Eq. (3.6). Since there are no multiplicities, Eq. (3.37) implies

$$B\psi_n=\lambda\psi_n. \tag{3.38}$$

Since B commutes with $\mathscr{H}$, Eq. (3.38) also holds, with the same λ, for the vectors $\{a^k\psi_n^-,(a^*)^l\psi_n^-\}$, generating a dense domain. B is closable, and therefore $B=\lambda 1$.

IV. TEMPERED DOMAIN-IRREDUCIBLE REPRESENTATIONS OF THE HEISENBERG ALGEBRA IN KREIN SPACES

The classification of the *a priori* more general set of domain irreducible representations of the Heisenberg algebra in Krein spaces is more difficult, since, in general, domain irreducibility does not seem to imply compactness of the group $U(s)$. This property, however, follows if the behavior of $U(s)$ at infinity is not too singular. We therefore introduce the following notion.

Definition 11: A representation of the Heisenberg algebra $\mathscr{H}$ on a domain $\mathscr{D}$ will be called tempered if

(i) *It is regular (Definition 4);*

(ii) *$U(s)$ is weakly polynomially bounded, i.e., if $\forall x\in\mathscr{D}$, there exists a polynomial $P_x(n)$ such that $\forall y\in\mathscr{K}$ there exists a constant C_y, such that*

$$|\langle y,U(2\pi n)x\rangle|\leqslant C_yP_x(n), \quad \forall n\in\mathbb{Z}. \tag{4.1}$$

We then have the following.

Proposition 12: For tempered domain irreducible representations of the Heisenberg algebra in Krein spaces, the operator $U(2\pi n)$ is a multiple of the identity,

$$U(2\pi n)=e^{2\pi i\theta}\mathbb{1}. \tag{4.2}$$

From the proof of Theorem 7 we have the following.

Theorem 13: *The tempered domain irreducible representations of the Heisenberg algebra in Krein spaces are classified as in Theorem 7.*

Proof of Proposition 12: By Eq. (2.7), $U(2\pi n)$ commutes with the algebra $\mathscr{H}$ on $\mathscr{D}$, but Eq. (4.2) does not immediately follow, as in the standard case (bounded operators in Hilbert spaces), where a crucial role is played by the spectral theorem. A substitute for the spectral theory of the group $U(2\pi n)$ is necessary, and can be based on temperedness. The Fourier transform of $U(2\pi n)$,

$$T(t)=\sum_{n=-\infty}^{\infty}U(2\pi n)e^{int},$$

is, in fact, well defined as an operator-valued distribution on $\mathscr{D}$: denote

$$T_N(t) \equiv \sum_{n=-N}^{N} U(2\pi n) e^{int},$$

and, for all $f(t)$ infinitely differentiable on the circle, $t \in [0,2\pi)$,

$$T_N(f) \equiv \int_0^{2\pi} T_N(t) f(t) dt = \sum_{n=-N}^{N} c_f^{(n)} U(2\pi n), \tag{4.3}$$

with $c_f^{(n)}$ the Fourier coefficients of f. Since such coefficients decrease faster than any polynomial, the temperedness condition immediately implies that $T_N(f)$ converges weakly on $\mathscr{D}$, as $N \to \infty$. By the weak completeness of $\mathscr{K}$, the limit defines an operator that maps $\mathscr{D}$ in $\mathscr{K}$. Furthermore, since all the elements $A \in \mathscr{H}$ are closable,

$$AT_N(f)x = T_N(f)Ax \to T(f)Ax, \quad \forall x \in \mathscr{D},$$

implies

$$AT_N(f)x \to AT(f)x, \tag{4.4}$$

and therefore

$$[A, T(f)] = 0, \quad \text{on } \mathscr{D}. \tag{4.5}$$

The operator-valued distribution $T(f)$ satisfies on $\mathscr{D}$,

$$T(1) = \sum_{n=-\infty}^{\infty} U(2\pi n) \int_0^{2\pi} e^{int} dt = U(0) = 1.$$

If $f_1(t)$ and $f_2(t)$ have disjoint support, $f_1(t) f_2(t) = 0$, then

$$T(f_1) T(f_2) = 0. \tag{4.6}$$

In fact, on $\mathscr{D}$,

$$T(f_1) T(f_2) = \sum_{n,m} U(2\pi n) U(2\pi m) c_{f_1}^{(n)} c_{f_2}^{(m)} = \sum_{k,m} U(2\pi k) c_{f_1}^{(k-m)} c_{f_2}^{(m)} \equiv \sum_{k} U(2\pi k) c^{(k)},$$

with

$$\begin{aligned} c^{(k)} &= \sum_m c_{f_1}^{(k-m)} c_{f_2}^{(m)} \\ &= \sum_m \int_0^{2\pi} e^{ikt} e^{-imt} f_1(t) dt \int_0^{2\pi} e^{imt'} f_2(t') dt' \\ &= \sum_m \int_0^{2\pi} e^{ikt} e^{-im(t-t')} f_1(t) f_2(t') dt dt' = \int_0^{2\pi} e^{ikt} f_1(t) f_2(t) dt = 0. \end{aligned}$$

In order to prove that $U(2\pi n)$ is a multiple of the identity, we have to consider two possibilities.

(i) The support of T consists of only one point, denoted by t_0. Let us fix $x \in \mathscr{D}$; the bound (4.1) and the support property of T imply $\forall y \in \mathscr{D}$,

$$\langle y, U(2\pi n) e^{-int_0} x \rangle = \sum_{k=0}^{N} a_k(y) n^k. \tag{4.7}$$

Then, by putting

$$x_k \equiv \sum_{m=0}^{N} A_{kn} U(2\pi n) e^{-int_0} x, \quad k=0,1,...N, \tag{4.8}$$

with A_{kn} the inverse matrix of $B_{nj} \equiv n^j$,

$$\sum_{n=0}^{N} A_{kn} n^j = \delta_{kj},$$

one has

$$a_k(y) = \langle y, x_k \rangle,$$

and therefore

$$U(2\pi n)x = \sum_{j=0}^{N} n^j e^{int_0} x_j. \tag{4.9}$$

The lowest integer N for which Eq. (4.7) holds will be called the degree of x. The subspace $\mathscr{D}_M \subset \mathscr{D}$ of vectors of degree less or equal to some M is invariant under $\mathscr{H}_{\text{ext}}$, and nonempty for sufficiently large M; we have therefore constructed operators O_j, $j=0,\ldots M$, defined on $\mathscr{D}_M$ by $O_j x = x_j$; by Eq. (4.8), the operators O_j are given by finite linear combinations of operators $U(2\pi n)$, so that they are closeable and leave $\mathscr{D}_M$ invariant. The operator O_M does not vanish, if M is chosen small enough, and, if $M>0$, it satisfies, by the group law for $U(s)$,

$$O_M^2 = 0. \tag{4.10}$$

$O_M \mathscr{D}_M$ is invariant under $\mathscr{H}_{\text{ext}}$ and is therefore dense, by domain irreducibility; this is incompatible with Eq. (4.10); all vectors in $\mathscr{D}$ are therefore of degree 0, i.e.,

$$U(2\pi n)x = e^{int_0} x, \quad \forall x \in \mathscr{D}. \tag{4.11}$$

(ii) The support contains at least two points. There are then two functions, f_1, f_2, with $T(f_i) \neq 0$, $i=1,2$, and $f_1 f_2 = 0$. Then, $T(f_1)\mathscr{D}$, by Eq. (4.4) a subspace of $\mathscr{D}_c$, must be dense, since it is invariant under $\mathscr{H}$, and the representation is domain irreducible. But then $T(f_2)T(f_1) = 0$ is incompatible with $T(f_2) \neq 0$, since all $T(f)$ are closable.

The same analysis as in Theorems 7 and 13 can be done under the weaker assumption that $\mathscr{D}$ is a dense subspace of a sequentially weakly complete nondegenerate inner product space $\mathscr{V}$, rather than a Krein space. The assumption is indeed weaker, since there exist weakly complete inner product spaces that are not Krein spaces. In fact, all the assumptions of Theorems 7 and 13 make sense in sequentially weakly complete spaces, and we have the following.

Proposition 14: The regular, commutant irreducible or tempered domain irreducible, representations of the Heisenberg algebra in sequentially weakly complete spaces are classified as in Theorem 7, up to isometries defined on dense subspaces.

Proof: The proofs of Theorems 7 and 13 only use the representation of $\mathscr{H}$ on a weakly dense domain $\mathscr{D}$, the regularity condition (A) (and temperedness) on $\mathscr{D}$, and construct the common eigenvectors of $U(s)$ by taking weak limits and extending weakly closable operators.

ACKNOWLEDGMENT

M. Mnatsakanova and Yu. Vernov were supported by the Russian Foundation for Basic Research.

APPENDIX: REGULARITY, DOMAIN AND COMMUTANT IRREDUCIBILITY, WEAK COMPLETENESS

The notion of irreducibility discussed in Sec. II crucially relies on the regularity of the representation and on making reference to the extended algebra $\mathscr{H}_{\text{ext}}$. In fact, even for the standard regular irreducible representations of the Weyl algebra in a Hibert space, neither domain

nor commutant irreducibility hold, in general, for the Heisenberg algebra, on an invariant domain $\mathscr{D}$, either because on $\mathscr{D}$ regularity fails, or because the subdomain $\mathscr{D}_1$ (see Definitions 5 and 6) is not stable under $\mathscr{H}_{\text{ext}}$.

For example, if for $\mathscr{D}$ one takes the dense domain $\mathscr{D}_0$ of C^∞ functions $f(x)$ of fast decrease, vanishing together with all their derivatives at $x=0$, $\mathscr{D}_0$ is clearly invariant under the Heisenberg algebra, but the representation of the Heisenberg algebra on $\mathscr{D}_0$ does *not* satisfy the regularity condition, since the operators $U(s)$ exist but do not leave $\mathscr{D}_0$ invariant. The subdomains $\mathscr{D}_0^+$ and $\mathscr{D}_0^-$, consisting of a function with positive, respectively negative, support are invariant, and their closure are proper orthogonal subspaces of $L^2(\mathbb{R})$. The representation is therefore *domain reducible*. It is also *commutant reducible*, since the projector on $L^2(0,\infty)$ commutes with the Heisenberg algebra on $\mathscr{D}_0$ (not only in the weak sense, since the domain is left invariant by the projector).

If the representation of $\mathscr{H}$ is defined, on a domain $\mathscr{D}\supset\mathscr{D}_0$, invariant under $\mathscr{H}$ and $U(s)$, $s\in\mathbb{R}$, then the regularity condition is satisfied. However, domain irreducibility does *not* hold for the *Heisenberg algebra* on $\mathscr{D}$, since the previously considered domains are obviously still invariant under $\mathscr{H}$. Domain irreducibility *holds for the extended algebra* $\mathscr{H}_{\text{ext}}$; in fact, invariance of a domain $\mathscr{D}'$ under $U(s)$ implies, using the construction given in the proof of Theorem 7, that one, and therefore every, eigenvector of the operators $U(s)$ belongs to the closure of $\mathscr{D}'$, and $\mathscr{D}'$ is therefore dense.

Identical conclusions hold for commutant reducibility, since one may still consider the same projectors as above, as operators defined on $\mathscr{D}_0$, commuting there with $\mathscr{H}$. On the other hand, commutativity with $\mathscr{H}_{\text{ext}}$ on an invariant domain (Definition 6) implies triviality, as in the proof of Theorem 7.

The operator N is *not* essentially self-adjoint on $\mathscr{D}_0$, and the extended algebra [with the validity of relations (2.7)] can be constructed only on the basis of *one* choice for its self-adjoint extension (namely by taking functions continuous, together with their derivative, at $x=0$). All other choices lead to different extensions of $\mathscr{H}$, with algebraic relations different from (2.7), and, in general, to different spectra of N.

A weakly complete inner product space that is not a Krein space is easily constructed as follows: we start with the space V_0 of *finite* sequences a_n, $n\in\mathbb{Z}$, with the inner product

$$\langle a_n, b_m\rangle \equiv \sum_n \bar{a}_{-n} b_n .$$

On the space V_0 it is easy to define positive scalar products that majorize the indefinite inner product (see Ref. 15), and therefore V_0 can be completed to a Krein space (inequivalent scalar products defining different Krein completions).

However, not all weakly complete inner product spaces that contain V_0 as a dense subspace are Krein spaces. Consider, in fact, the space $V\supset V_0$ consisting of *all* sequences that terminate on the right, $a_n=0$, for n large enough. The space V is weakly complete, because (i) weak convergence implies convergence of each term of the sequences, (ii) by completeness of the l^2 spaces, the limit sequences x_n belong to all the spaces l^2 with *any* weight function $p_n>0$ on $n\in\mathbb{N}$,

$$\sum_{n=0}^{\infty} |x_n|^2 p_n < \infty,$$

and therefore x_n vanishes for n large enough.

Moreover, V is not a Krein space, since in any Krein space containing V_0 there are always sequences that are infinite on the right, and have weakly convergent truncations; one such sequence is immediately constructed, e.g., as $1/(c_k k^2)$, $k>0$, with c_k the Hilbert norm of the sequence $a_n\equiv\delta_{n,k}$, in the given Krein space. The argument also shows that V is not a Banach space, i.e., it cannot be obtained from V_0 by a norm completion.

[1] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **180**, 1 (1942).
[2] W. Pauli, Rev. Mod. Phys. **15**, 175 (1943), W. Heisenberg, Nucl. Phys. **4**, 532 (1957).
[3] F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974).
[4] F. Strocchi, *Selected Topics on the General Properties of Quantum Field Theory* (World Scientific, Singapore, 1993).
[5] W. Heisenberg, *The Physical Principles of the Quantum Theory* (Dover, New York, 1930).

[6] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford University Press, Oxford, 1986).
[7] J. Von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, 1955).
[8] J. M. Chaiken, Ann. Phys. (N.Y.) **42**, 23 (1967).
[9] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer-Verlag, Berlin, 1967), Vol. II.
[10] C. R. Putnam, *Commutation Properties of Hilbert Space Operators and Related Topics* (Springer-Verlag, Berlin, 1967); P. Jorgensen and R. Moore, *Operator Commutation Relations* (Kluwer, Dordrecht, 1984).
[11] M. Fannes and A. Verbeure, Commun. Math. Phys. **35**, 257 (1974).
[12] F. Rellich, Nachr. Akad. Wiss. Goett. II, Math.-Phys. Kl. II, **11A**, 107 (1946).
[13] J. Dixmier, Compositio Math. Appl. **13**, 263 (1958).
[14] C. R. Putnam, loc. cit., Ch. IV, 4.6; N. S. Poulsen, Math. Scand. **32**, 112 (1973).
[15] J. Bognar, *Indefinite Inner-Product Spaces* (Springer-Verlag, Berlin, 1974).
[16] T. Ya. Azizov and I. S. Iokhvidov, *Linear Operators in Spaces with an Indefinite Metric* (Wiley, New York, 1989).
[17] R. T. Powers, Commun. Math. Phys. **21**, 85 (1971); Trans. Am. Math. Soc. **187**, 261 (1974).
[18] P. Jorgensen, Publ. RIMS Kyoto Univ. **25**, 923 (1989)
[19] A. O. Barut and R. Rackza, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986).
[20] E. Nelson and W. F. Stinespring, Am. J. Math. **81**, 547 (1959).

Quantum group structure on the quantum plane

E. H. Tahri[a)]
Université Mohammed V, Faculté des Sciences, Laboratoire de Physique Théorique, Av. Ibn Battota B.P. 1014, Rabat, Morocco

(Received 6 October 1997; accepted for publication 22 January 1998)

We show that the quantum plane introduced by Manin is a quantum group according to the Faddeev–Reshetikhin–Takhtajan approach. We give its matrix representation, R-matrix, and its Hopf structure. Some quantum properties of the quantum matrix whose elements generate the quantum plane are also discussed. © *1998 American Institute of Physics.* [S0022-2488(98)00205-9]

I. INTRODUCTION

In recent years, there has been much interest in the concept of quantum groups.[1–4] This is due to its close connection to the Yang–Baxter equation, which is central in various physically interesting models and theories, such as the Knizhnik–Zamolodchikov equation, two-dimensional conformal field theory, integrable models, and so on.[5–8] Quantum groups are noncommutative Hopf algebras. The noncommutativity is controlled by the so-called R-matrices that are solutions of the Yang–Baxter equation and whose entries are complex numbers called deformation parameters. Ordinary Lie groups (commutative Hopf algebra) are recovered when the latter parameters go to particular values. So, quantum groups are deformations (and in some sense generalizations) of usual groups.

Manin has discribed quantum general linear group as a natural symmetry of noncommutative algebraic variety defined by quadratic equations (see Refs. 9 and 10). This new object is called ''quantum plane'' or ''quantum space'' or also ''Manin plane.'' It is defined as the deformation of coordinate functions and is mathematically viewed as comodule over the corresponding quantum group. Since its inception, the quantum plane has been attracting a growing amount of attention; it has been envisioned by many as a paradigm for the general program of quantum deformed physics. Indeed, many efforts have been made to set adequate mathematical tools and then attempts to build some quantum deformed models (see, for example, Refs. 11–18, and references therein).

In our knowledge the quantum plane has never been considered a quantum group. In this work we give its matrix representation, which leads to the structure of the quantum group. In this new approach the quantum plane appears as a quantum subgroup of a larger quantum group, with special conditions on deformation parameters.

The paper is organized as follows. In Sec. II, we concern ourself with the two-dimensional case, because of its simplicity and importance in the literature. In the first part we recall the Manin approach and at the same time we set the basic tools for the proof of our main results, which are in the second part. We also discuss some important quantum properties of quantum matrices considered to be the group elements of the quantum plane. In Sec. III, we generalize the above results to a high dimension. Finally, we give some concluding remarks in Sec. IV.

II. TWO-DIMENSIONAL QUANTUM PLANE

A. Review of Manin approach

The simplest quantum group is a deformation $GL_q(2)$ of the classical group $GL(2)$ of 2×2 invertibles matrices. The more general deformed version depends on two parameters, p and q, denoted by $GL_{p,q}(2)$, is well studied in Refs. 19 and 20. It consists of co-commutative Hopf algebra generated by noncommuting elements of the matrix:

[a)]Current mailing address: Université Med I, Faculté des Sciences, Lab. de Physique Théorique et de Particules, B.P. 524, 60000, Oujda, Morocco.

 © 1998 American Institute of Physics

$$M=\begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{2.1}$$

The noncommutativity is expressed in terms of the R matrix as

$$RM_1M_2=M_2M_1R, \quad \text{i.e.,} \quad R^{ij}{}_{kl}M^k{}_mM^l{}_n=M^j{}_lM^i{}_kR^{kl}{}_{mn}, \tag{2.2}$$

called the RTT equation, where the matrix R is given by

$$R=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & q^{-1} & 0 & 0 \\ 0 & 1-(qp)^{-1} & p^{-1} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.3}$$

which is the solution of the Yang–Baxter equation:

$$R_{12}R_{13}R_{23}=R_{23}R_{13}R_{12}, \quad \text{i.e.,} \quad R^{ij}{}_{kl}R^{km}{}_{rs}R^{ls}{}_{uv}=R^{jm}{}_{kl}R^{il}{}_{sv}R^{sk}{}_{ru} \tag{2.4}$$

(p and q are arbitrary complex numbers). Explicitly, the commutation relations between the generators a, b, c, and d are given by

$$\begin{gathered} ab=pba, \quad cd=pdc, \\ ac=qca, \quad bd=qdb, \\ pbc=qcb, \quad ad-da=(p-q^{-1})bc. \end{gathered} \tag{2.5}$$

Further, one has to introduce the quantum determinant defined by[19,21]

$$D=ad-pbc=da-q^{-1}bc, \tag{2.6}$$

which satisfies, using (2.5), the following relations:

$$\begin{gathered} Da=aD, \quad Dd=dD, \\ qDb=pbD, \quad pDc=qcD. \end{gathered} \tag{2.7}$$

Then one extends the algebra $GL_{p,q}(2)$ by an element D^{-1}, such that

$$DD^{-1}=D^{-1}D=1, \tag{2.8}$$

from which follows:

$$\begin{gathered} D^{-1}a=aD^{-1}, \quad D^{-1}d=dD^{-1}, \\ pD^{-1}b=qbD^{-1}, \quad qD^{-1}c=pcD^{-1}. \end{gathered} \tag{2.9}$$

This permits the introduction of the left M_L^{-1} and right M_R^{-1} quantum inverse matrices of M as

$$M^{-1}=M_L^{-1}=D^{-1}\begin{pmatrix} d & -q^{-1}b \\ -qc & a \end{pmatrix}=\begin{pmatrix} d & -p^{-1}b \\ -pc & a \end{pmatrix}D^{-1}=M_R^{-1}. \tag{2.10}$$

So, the Hopf structure is given by the following: *(i) the coproduct:*

$$\Delta(M)=M\dot{\otimes}M \quad \text{i.e.,} \quad \Delta\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)=\begin{pmatrix} a\otimes a+b\otimes c & a\otimes b+b\otimes d \\ c\otimes a+d\otimes c & c\otimes b+d\otimes d \end{pmatrix}, \tag{2.11a}$$

$$\Delta(D^{-1})=D^{-1}\otimes D^{-1}; \tag{2.11b}$$

(ii) the counit:

$$\epsilon(M)=1, \quad \text{i.e.} \quad \epsilon\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.12a}$$

$$\epsilon(D^{-1})=1; \tag{2.12b}$$

and *(iii) the antipode:*

$$S(M)=M^{-1}, \quad \text{i.e.,} \quad S\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)=\begin{pmatrix} dD^{-1} & -p^{-1}bD^{-1} \\ -pcD^{-1} & aD^{-1} \end{pmatrix}, \tag{2.13a}$$

$$S(D^{-1})=D. \tag{2.13b}$$

Now let us recall some quantum properties[10] of the quantum matrix M considered as a group element (or a point[10]) of $GL_{p,q}(2)$. First, if one takes another copy M' of the same type as M, that is, M' is also a group element of $GL_{p,q}(2)$ such that its elements mutually commute with those of M, then MM' is an element of $GL_{p,q}(2)$. Second, M^{-1} is a group element of $GL_{p^{-1},q^{-1}}(2)$. Next the nth power $M^n (n \in \mathbb{Z}_+)$ is a group element of $GL_{p^n,q^n}(2)$.

Manin[9,10] has defined in correspondence with the quantum group, a quantum plane and its dual. This consists of free algebras noted $K_q^{2|0}$ and $K_p^{0|2}$ generated by the pairs of coordinates (x,y) and (ξ,η), respectively, such that for $K_q^{2|0}$ we have

$$xy=qyx, \tag{2.14}$$

which is the explicit from the following relation:

$$x^i x^j = R^{ji}{}_{kl} x^k x^l, \tag{2.15}$$

where $x^1=x$, $x^2=y$ and R is the R matrix (2.3).

For $K_p^{0|2}$, we have

$$\xi\eta=-\frac{1}{p}\,\xi\eta, \quad \xi^2=\eta^2=0, \tag{2.16}$$

which also can be expressed in terms of the R matrix (2.3) as

$$\xi^i \xi^j = -pq R^{ji}_{kl} \xi^k \xi^l, \tag{2.17}$$

where $\xi^1=\xi$, $\xi^2=\eta$. In this point of view, the quantum group $GL_{p,q}(2)$ appears as endomorphisms of the quantum planes $K_q^{2|0}$ and $K_p^{0|2}$ under which the quadratic relations (2.14) and (2.16) remain invariant. So, Manin proved the equivalence between the defining quantum group commutation relations (2.5) and the quadratic relations (2.14) and (2.16). This is known in the literature as the Manin approach to the quantum group.

The quantum planes $K_q^{2|0}$ and $K_p^{0|2}$ are viewed mathematically as comodules over $GL_{p,q}(2)$ via the (left) coactions

$$\delta: K_q^{2|0} \mapsto GL_{p,q}(2) \otimes K_q^{2|0}, \quad \delta^*: K_p^{0|2} \mapsto GL_{p,q}(2) \otimes K_p^{0|2},$$

such that

$$\delta(X)=M \otimes X, \quad \text{i.e.,} \quad \begin{cases} \delta(x)=a \otimes x + b \otimes y, \\ \delta(y)=c \otimes x + d \otimes y \end{cases}, \tag{2.18a}$$

$$\delta^*(\Theta)=M \otimes \Theta, \quad \text{i.e.,} \quad \begin{cases} \delta^*(\xi)=a \otimes \xi + b \otimes \eta, \\ \delta^*(\eta)=c \otimes \xi + d \otimes \eta \end{cases}, \tag{2.18b}$$

where X and Θ are column vectors of components x, y and ξ, η, respectively.

B. Quantum plane as quantum group

Before giving the quantum group structure on $K_q^{2|0}$, let us look at the particular case of $GL_{p,q}(2)$ when $p=q^{-1}$. In this case the R matrix (2.3) reduces to

$$R=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & q^{-1} & 0 & 0 \\ 0 & 0 & q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.19}$$

and the commutation relations (2.5) become

$$\begin{gathered} ab=q^{-1}ba, \quad cd=q^{-1}dc, \\ ac=qca, \quad bd=qdb, \\ bc=q^2cb, \quad ad=da. \end{gathered} \tag{2.20}$$

Now, let us observe that if we take $a=d=y$, $b=x$ and $c=0$, then the commutation relations (2.20) reduce to the quadratic relation (2.14) defining the quantum plane $K_q^{2|0}$. It is the fondamental observation that proves the following main result.

Theorem 1: The quantum plane $K_q^{2|0}$ generated by the coordinates x and y satisfying the quadratic relation $xy=qyx$ is a quantum group generated by the elements of the matrix of the type

$$A=\begin{pmatrix} y & x \\ 0 & y \end{pmatrix}, \tag{2.21}$$

modulo the RTT equation:

$$RA_1A_2=A_2A_1R, \quad \text{i.e.,} \quad R^{ij}{}_{kl}A^k{}_mA^l{}_n=A^j{}_lA^i{}_kR^{kl}{}_{mn}, \tag{2.22}$$

where R is solution of the Yang–Baxter equation given by (2.19).

The costructure is given by the following:

(i) the coproduct:

$$\Delta(A)=A\dot{\otimes}A, \quad \text{i.e.,} \quad \begin{cases} \Delta(x)=y\otimes x+x\otimes y, \\ \Delta(y)=y\otimes y; \end{cases} \tag{2.23}$$

(ii) the counit:

$$\epsilon(A)=1, \quad \text{i.e.,} \quad \begin{cases} \epsilon(x)=0, \\ \epsilon(y)=1. \end{cases} \tag{2.24}$$

Moreover, the bialgebra $K_q^{2|0}$ extended by y^{-1}, such that

$$yy^{-1}=y^{-1}y=1 \tag{2.25}$$

is a Hopf algebra with an antipode structure given by

(iii) the antipode:

$$S(A)=A^{-1}, \quad \text{i.e.,} \quad \begin{cases} S(x)=-y^{-1}xy^{-1}, \\ S(y)=y^{-1}. \end{cases} \tag{2.26}$$

The costructure is immediately obtained from the one on $GL_{p,q}(2)$ (2.11a) (2.12a) when $p=q^{-1}$ and M in (2.1) is replaced by A in (2.21). The quantum inverse of A is given by

$$A^{-1}=D^{-1}\begin{pmatrix} y & -q^{-1}x \\ 0 & y \end{pmatrix}=\begin{pmatrix} y & -qx \\ 0 & y \end{pmatrix}D^{-1}, \tag{2.27}$$

where the quantum determinant is now

$$D=y^2, \tag{2.28}$$

from which it follows that $D^{-1}=y^{-2}$, which is an element of the quantum plane extended by y^{-1}. Further, from the quadratic relation (2.14) and (2.25) it follows that

$$y^{-1}x=qxy^{-1}. \tag{2.29}$$

So we get the explicit form of the quantum inverse in (2.27) of A as

$$A^{-1}=\begin{pmatrix} y^{-1} & -y^{-1}xy^{-1} \\ 0 & y^{-1} \end{pmatrix}, \tag{2.30}$$

which gives the relations (2.26). Finally, from (2.25), (2.23), (2.24), and (2.26) it follows that

$$\Delta(y^{-1})=y^{-1}\otimes y^{-1}, \quad \epsilon(y^{-1})=1, \quad S(y^{-1})=y. \tag{2.31}$$

Now all quantum properties of the quantum matrix M in (2.1), considered as a group element of $GL_{p,q}(2)$ when $p=q^{-1}$, hold in the case of a quantum matrix of type A considered as a group element of $K_q^{2|0}$. So, we have the following.

Proposition 1: (a) If A and A' are two copies of group elements of $K_q^{2|0}$ such that their elements mutually commute, then their product is commutative ($AA'=A'A$) and a group element of $K_q^{2|0}$.

(b) The nth power of A is given by

$$A^n=\begin{pmatrix} y^n & [n]_q y^{n-1}x \\ 0 & y^n \end{pmatrix}, \tag{2.32}$$

where $[n]_q=(1-q^n)/(1-q)$, and it is a group element of $K_{q^n}^{2|0}$.

(c) The quantum inverse of A is a group element of $K_{q^{-1}}^{2|0}$.

Proof: (a) The commutativity of product is easily seen. Indeed, we have

$$A''=AA'=\begin{pmatrix} yy' & yx'+xy' \\ 0 & yy' \end{pmatrix}=\begin{pmatrix} y'y & y'x+x'y \\ 0 & y'y \end{pmatrix}=A'A, \tag{2.33}$$

and $\det_q(A'')=\det_q(A)\det_q(A')=y^2y'^2$.

The statements (b) and (c) can be proved by a direct calculus using the quadratic relations (2.14) and (2.29).

Remarks: (1) The power n of A in (b) can also be a negative integer and when it is equal to -1 the expression (2.32) coincides exactly with that of the quantum inverse of A given by (2.30) (since $[-1]_q=-q^{-1}$).

(2) A quantum matrix of type A in (2.21) is actually a particular copy of a general group element M in (2.1) of quantum group $GL_{q^{-1},q}(2)$, so the quantum plane $K_q^{2|0}$ may be viewed as a quantum subgroup of $GL_{q^{-1},q}(2)$.

In a more rigorous mathematical language, the quantum group $GL_{q^{-1},q}(2)$ is bicovariant bicomodule over $K_q^{2|0}$, as can be easily checked in the following statement.

Proposition 2: There exists algebra morphisms:

$$\Delta_L: GL_{q^{-1},q}(2)\mapsto K_q^{2|0}\otimes GL_{q^{-1},q}(2), \quad \Delta_R: GL_{q^{-1},q}(2)\mapsto GL_{q^{-1},q}(2)\otimes K_q^{2|0},$$

such that

$$\Delta_L(M)=A\otimes M, \quad \text{i.e.,} \quad \Delta_L(M_j^i)=A_k^i\otimes M_j^k, \tag{2.34}$$

$$\Delta_R(M)=M\otimes A, \quad \text{i.e.,} \quad \Delta_R(M_j^i)=M_k^i\otimes A_j^k, \tag{2.35}$$

satisfying the properties of left and right coactions:

$$(\Delta\otimes \mathrm{id})\Delta_L=(\mathrm{id}\otimes\Delta_L)\Delta_L,$$
$$(\mathrm{id}\otimes\epsilon)\Delta_L(M)=M, \tag{2.36}$$

and

$$(\mathrm{id}\otimes\Delta)\Delta_R=(\Delta_R\otimes\mathrm{id})\Delta_R,$$
$$(\mathrm{id}\otimes\epsilon)\Delta_R(M)=M, \tag{2.37}$$

respectively, and the condition of bicovariance:

$$(\mathrm{id}\otimes\Delta_R)\Delta_L=(\Delta_L\otimes\mathrm{id})\Delta_R. \tag{2.38}$$

Now let us return to the expression (2.32) of A^n and suppose that the deformation parameter is root of unity, we remark the following.

Remark: If there exists $m\in\mathbb{Z}_+$ such that $q^m=1$, then A^m becomes

$$A^m=\begin{pmatrix} y^m & 0 \\ 0 & y^m \end{pmatrix}. \tag{2.39}$$

Furthermore, (a) if $y^m=0$, then $A^m=0$; (b) if $y^m=1$, then $A^m=1$.

In case (a) we obtain a nilpotent quantum matrix. The quantum plane generated by such coordinates is closely related to the so-called fractional superspace that unifies the boson with a particle of fractional spin. In this case there is no antipode on $K_q^{2|0}$. In case (b) we have a cyclic quantum plane, in the sense that $K_{q^{n+1}}^{2|0}\equiv K_q^{2|0}$. Finally, let us remark that there is no additional condition on the coordinate x as that on y in contrast to the classical plane where we must introduce at the same time the inverses of x and y to get the structure of the group. We interpret this as a broken symmetry between the two directions.

III. HIGH-DIMENSIONAL CASE

The multiparametric deformation of $GL(n)$ generalizing that of $GL(2)$ of Sec. I is well studied in Refs. 22–24. It is generated by the elements $M^i{}_j(i,j=1,...,n)$ of the $n\times n$ matrix M, such that (for $i<j$ and $k<l$)

$$M^i{}_kM^i{}_l=p_{kl}M^i{}_lM^i{}_k,$$
$$M^i{}_kM^j{}_k=q_{ij}M^j{}_kM^i{}_k,$$
$$M^i{}_lM^j{}_k=\frac{q_{ij}}{p_{kl}}\,M^j{}_kM^i{}_l, \tag{3.1}$$
$$M^i{}_kM^j{}_l=\frac{p_{kl}}{p_{ij}}\,M^j{}_lM^i{}_k+(p_{kl}-q_{kl}^{-1})M^i{}_lM^j{}_k,$$

with $p_{ij}q_{ij}=r$, where r is a complex number. These commutation relations are the explicit form of the RTT equation as in (2.2) with the R matrix is now given by[22]

$$R^{ij}{}_{kl}=\delta^i{}_k\delta^j{}_l\left(\delta^{ij}+\Theta^{ji}\,\frac{1}{q_{ij}}+\Theta^{ij}\,\frac{q_{ji}}{r}\right)+\delta^i{}_l\delta^j{}_k\Theta^{ij}\left(1-\frac{1}{r}\right), \tag{3.2}$$

where $\Theta^{ij}=1$ if $i>j$ and 0 elsewhere. This quantum group is noted $GL_{q_{ij},r}(n)$. The corresponding quantum plane $K_{q_{ij}}^{n|0}$ is generated by the set of coordinates $(x^i)_{1\leqslant i\leqslant n}$ satisfying the following quadratic relations:

$$x^ix^j=q_{ij}x^jx^i, \tag{3.3}$$

such that

$$q_{ii}=1, \quad q_{ij}q_{ji}=1, \tag{3.4}$$

for $i,j=1,...,n$. The quadratic relations (3.3) may be expressed in terms of the R matrix (3.2) as in (2.15).

Now let us look at the following special case. Let us take $r=1$ and the only matrix elements different from zero are $M^i{}_i$ and $M^i{}_n$ for all $i=1,...,n$, i.e.,

$$M=\begin{pmatrix} M^1{}_1 & 0 & \dots & 0 & M^1{}_n \\ 0 & M^2{}_2 & \ddots & \vdots & M^2{}_n \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & M^{n-1}{}_{n-1} & M^{n-1}{}_n \\ 0 & \dots & 0 & 0 & M^n{}_n \end{pmatrix}. \tag{3.5}$$

Then the commutation relations (3.1) reduce to $(i<j\leqslant n)$

$$M^i{}_iM^i{}_n=q_{ni}M^i{}_nM^i{}_i, \tag{3.6a}$$

$$M^i{}_nM^j{}_n=q_{ij}M^j{}_nM^i{}_n, \tag{3.6b}$$

$$M^i{}_nM^j{}_j=q_{ij}q_{jn}M^j{}_jM^i{}_n, \tag{3.6c}$$

$$M^i{}_iM^j{}_n=q_{ni}q_{ij}M^j{}_nM^i{}_i, \tag{3.6d}$$

$$M^i{}_iM^j{}_j=M^j{}_jM^i{}_i. \tag{3.6e}$$

Furthermore, if we put $M^i{}_i=x^n$ and $M^i{}_n=x^i$ for $i=1,...,n$, with the following additional conditions on the deformation parameters:

$$q_{ij}q_{jn}q_{ni}=1, \tag{3.7}$$

then we prove the following.

Theorem 2: The n-dimensional quantum plane $K^{n|0}_{q_{ij}}$ generated by the set of coordinates $\{x^i,i=1,...,n\}$ satisfying the following quadratic relations:

$$x^ix^j=q_{ij}x^jx^i, \tag{3.8}$$

such that $q_{ii}=1$, $q_{ij}q_{ji}=1$, and $q_{ij}q_{jn}q_{ni}=1$ for $i,j=1,...,n$, is a quantum group generated by the elements of the matrix of type

$$A=\begin{pmatrix} x^n & 0 & \dots & 0 & x^1 \\ 0 & x^n & \ddots & \vdots & x^2 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & x^n & x^{n-1} \\ 0 & \dots & 0 & 0 & x^n \end{pmatrix}, \tag{3.9}$$

and the quadratic relations (3.8) are just the explicit form of the RTT equation, where R is given by

$$R^{ij}{}_{kl}=\delta^i{}_k\delta^j{}_l\left(\delta^{ij}+\Theta^{ji}\frac{1}{q_{ij}}+\Theta^{ij}q_{ji}\right). \tag{3.10}$$

The costructure is given by the following:
(i) the coproduct:

$$\Delta(A)=A\dot{\otimes}A, \quad \text{i.e.,} \quad \begin{cases} \Delta(x^i)=x^n\otimes x^i+x^i\otimes x^n, & 1\leqslant i\leqslant n-1, \\ \Delta(x^n)=x^n\otimes x^n. \end{cases} \tag{3.11}$$

(ii) the counit:

$$\epsilon(A)=1, \quad \text{i.e.,} \quad \begin{cases} \epsilon(x^i)=0, & 1\leqslant i\leqslant n-1, \\ \epsilon(x^n)=1. \end{cases} \tag{3.12}$$

Moreover, the bialgebra $K_{q_{ij}}^{n|0}$ extended by $(x^n)^{-1}$, such that

$$(x^n)^{-1}x^n=x^n(x^n)^{-1}=1, \tag{3.13}$$

is a Hopf algebra with an antipode structure given by
(iii) the antipode:

$$S(A)=A^{-1} \quad \text{i.e.,} \quad \begin{cases} S(x^i)=-(x^n)^{-1}x^i(x^n)^{-1}, & 1\leqslant i\leqslant n-1, \\ S(x^n)=(x^n)^{-1}. \end{cases} \tag{3.14}$$

Let us prove the explicit form (3.14) of the quantum inverse of A. The quantum determinant of A is given by

$$D=\det_q(A)=(x^n)^n, \tag{3.15}$$

[the second n in $(x^n)^n$ designs the nth power of coordinate x^n] from which it follows that $D^{-1}=(x^n)^{-n}$, which is an element of $K_{q_{ij}}^{n|0}$ extended by $(x^n)^{-1}$. From (3.8) and (3.15) we have

$$Dx^i=(q_{in})^{-n}x^iD, \tag{3.16}$$

from which follows

$$D^{-1}x^i=(q_{in})^nx^iD^{-1}, \tag{3.17}$$

where $i=1,...,n$. Furthermore, the quantum inverse of A is given by

$$\begin{aligned} A^{-1}=D^{-1}&\begin{pmatrix} (x^n)^{n-1} & 0 & \dots & 0 & -q_{1n}^{-1}(x^n)^{n-2}x^1 \\ 0 & (x^n)^{n-1} & \ddots & \vdots & -q_{2n}^{-1}(x^n)^{n-2}x^2 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & (x^n)^{n-1} & -q_{n-1,n}^{-1}(x^n)^{n-2}x^{n-1} \\ 0 & \dots & 0 & 0 & (x^n)^{n-1} \end{pmatrix} \\ =&\begin{pmatrix} (x^n)^{n-1} & 0 & \dots & 0 & -q_{1n}x^1(x^n)^{n-2} \\ 0 & (x^n)^{n-1} & \ddots & \vdots & -q_{2n}x^2(x^n)^{n-2} \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & (x^n)^{n-1} & -q_{n-1,n}x^{n-1}(x^n)^{n-2} \\ 0 & \dots & 0 & 0 & (x^n)^{n-1} \end{pmatrix} D^{-1}, \end{aligned} \tag{3.18}$$

which is consistent with (3.17). Then, from (3.8) and (3.13), it follows that

$$A^{-1}=\begin{pmatrix} (x^n)^{-1} & 0 & \dots & 0 & -(x^n)^{-1}x^1(x^n)^{-1} \\ 0 & (x^n)^{-1} & \ddots & \vdots & -(x^n)^{-1}x^2(x^n)^{-1} \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & (x^n)^{-1} & -(x^n)^{-1}x^{n-1}(x^n)^{-1} \\ 0 & \dots & 0 & 0 & (x^n)^{-1} \end{pmatrix}, \tag{3.19}$$

which proves the explicit forms given in (3.14).

Finally, using (3.13), (3.11), (3.12), and (3.14), we obtain the actions of Δ, ϵ, and S on the inverse of x^n as

$$\Delta((x^n)^{-1})=(x^n)^{-1}\otimes(x^n)^{-1},\quad \epsilon((x^n)^{-1})=1,\quad S((x^n)^{-1})=x^n. \tag{3.20}$$

Remark: As in the two-dimensional case, the quantum plane $K_{q_{ij}}^{n|0}$ appears as a quantum subgroup of $GL_{q_{ij},1}(n)$ when the additional conditions (3.7) on the parameters are satisfied. The coactions defined in Proposition 2 generalize to the n dimension [as in (2.34) and (2.35), $i,j=1,...,n$], which provides the structure of the bicovariant bicomodule on $GL_{q_{ij},1}(n)$ over $K_{q_{ij}}^{n|0}$.

Now let us look at the quantum properties of quantum matrix A and generalize the two-dimensional case.

Proposition 3: (a) If A and A' are two copies of group elements of $K_{q_{ij}}^{n|0}$ such that their elements pairwise commute, then their product is commutative ($AA'=A'A$) and a group element of $K_{q_{ij}}^{n|0}$.

(b) The mth power of A is a group element of $K_{q_{ij}^m}^{n|0}$ given by

$$A^m=\begin{pmatrix} (x^n)^m & 0 & \dots & 0 & [m]_{q_1}(x^n)^{m-1}x^1 \\ 0 & (x^n)^m & \ddots & \vdots & [m]_{q_2}(x^n)^{m-1}x^2 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & (x^n)^m & [m]_{q_{n-1}}(x^n)^{m-1}x^{n-1} \\ 0 & \dots & 0 & 0 & (x^n)^m \end{pmatrix}, \tag{3.21}$$

where $[m]_{q_i}=[1-(q_{in})^m]/(1-q_{in})$.

(c) The quantum inverse of A is a group element of $K_{q_{ij}^{-1}}^{n|0}$.

Proof: (a) The only nonvanishing elements of AA' are the diagonal and nth column elements, which are, at the ith row, $x^nx^{n'}$ and $x^nx^{i'}+x^ix^{n'}$, respectively, the corresponding elements of $A'A$ are $x^{n'}x^n$ and $x^{n'}x^i+x^{i'}x^n$, which proves the statement.

(b) The form of A^m can be deduced from a straightforward inductive way. To show that it is a group element of $K_{q_{ij}^m}^{n|0}$, let us set $x^{i'}=[m]_{q_i}(x^n)^{m-1}x^i$, then

$$\begin{aligned} x^{i'}x^{j'}&=[m]_{q_i}(x^n)^{m-1}x^i[m]_{q_j}(x^n)^{m-1}x^j=(q_{in})^{m-1}[m]_{q_i}[m]_{q_j}(x^n)^{m-1}(x^n)^{m-1}x^ix^j \\ &=q_{ij}(q_{in})^{m-1}(q_{jn})^{-(m-1)}[m]_{q_i}[m]_{q_j}(x^n)^{m-1}x^j(x^n)^{m-1}x^i=q_{ij}q_{jn}q_{ni}(q_{in})^m(q_{jn})^{-m}x^{j'}x^{i'} \\ &=(q_{ij})^mx^{j'}x^{i'}. \end{aligned} \tag{3.22}$$

In the last equality we used the conditions (3.7) on the deformation parameters.

Remark: (1) The power of A may be negative and when $m=-1$ the expression (3.21) coincides with that of the quantum inverse of A in (3.19).

(2) We check immediately from the expression (3.21) of A^m that if there exists integers m_{in} such that $(q_{in})^{m_{in}}=1$, then we have

$$A^N=\begin{pmatrix} (x^n)^N & 0 & \dots & 0 & 0 \\ 0 & (x^n)^N & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & (x^n)^N & 0 \\ 0 & \dots & 0 & 0 & (x^n)^N \end{pmatrix}, \tag{3.23}$$

where $N=\mathrm{lcm}(m_{in})_{1\leqslant i\leqslant n-1}$ (here the abbreviation lcm denotes the least common multiple). Furthermore, (a) if $(x^n)^N=0$ then $A^N=0$; (b) if $(x^n)^N=1$ then $A^N=1$.

Note that if q_{in} are roots of unity, then all q_{ij} are also roots of unity because of (3.7), namely, there exist integers $m_{ij}=\mathrm{lcm}(m_{in},m_{jn})$ such that $(q_{ij})^{m_{ij}}=1$. In case (a) there is no antipode on $K_{q_{ij}}^{n|0}$; it is only a bialgebra. This constitutes the so-called fractional superspace.

IV. CONCLUDING REMARKS

The concept of q deformation of Lie groups and spaces is well physically consistent when it is performed in the context of Hopf algebras. This is the motivation of this paper. So, we have shown that the well-known Manin plane, which is viewed mathematically as a comodule over its corresponding quantum group, may be incorporated into the Faddeev–Reshetikhin–Takhtajan approach to the quantum group.[3] We have given its matrix representation, the R-matrix, and its Hopf structure.

We expect that this new approach to the quantum plane will be a starting point to construct more consistent deformed physical models. Of course, this requires more mathematical tools; the bicovariant differential calculus à la Woronowicz[25] and the extension to the quantum superplane are in preparation.[26]

ACKNOWLEDGMENTS

I would like to thank M. Baghdadi for his permanent encouragement, A. El Hassouni for introducing me to this interesting topic, and E. M. Souidi for many valuable comments on a preliminary version of this manuscript. I am also grateful to the referee for the useful remarks.

[1] V. G. Drinfeld, *Quantum Groups*, Proc. Int. Congress of Mathematicians, Berkeley, CA, 1986 (AMS, Providence, RI, 1987), Vol. 1, pp. 798–820.
[2] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985); **11**, 247 (1986); Commun. Math. Phys. **102**, 537 (1986).
[3] L. D. Faddeev, N. Yu. Reshetikhin, and L. Takhtajan, *Algebraic Analysis* (Academic, 1988), Vol. 1, p. 129.
[4] S. L. Wororonowicz, Commun. Math. Phys. **111**, 613 (1987).
[5] A. Connes, Publ. Math. Inst. Hautes Etudes Sci. **62**, 257 (1986).
[6] C. N. Yang, Phys. Rev. Lett. **19**, 1312 (1967).
[7] L. Alvarez-Gaume, C. Gomez, and G. Sierra, Phys. Lett. B **220**, 142 (1989).
[8] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
[9] Yu. I. Manin, Ann. Inst. Fourier **37**, 191 (1987).
[10] Yu. I. Manin, ''Quantum group and non-commutative geometry,'' preprint, Montreal University, CRM-1561, 1988.
[11] J. Wess and B. Zumino, Nucl. Phys. B (Proc. Suppl.) **18**, 302 (1990).
[12] T. Brezezinski and J. Rembielinski, J. Phys. A **25**, 1945 (1992).
[13] C. Chryssomalakos and B. Zumino, ''Translation, integrals and Fourier transforms in the quantum plane,'' LBL-34803, UCB-PTII 93/30.
[14] J. Wess, Talk given at the Workshop de l'Ouest Mediterranéen en Physique Théorique, Rabat, 1997.
[15] M. Dubois-Violette, J. Madore, T. Masson, and J. Mourad, ''Linear connections on the quantum plane,'' LPTHE Orsay, 94/94.
[16] A. Kempf, ''Algebraic q-integration and Fourier transform on the quantum and braid spaces,'' DAMTP/94-7.
[17] U. Carow-Watamura, M. Schlieker, M. Scholl, and S. Watamura, Z. Phys. C **48**, 159 (1990); **49**, 439 (1991); Int. J. Mod. Phys. A **6**, 3081 (1991); O. Ogievetsky, W. B. Schmidke, J. Wess, and B. Zumino, Commun. Math. Phys. **150**, 495 (1992).
[18] M. Arik, Z. Phys. C **51**, 627 (1991).
[19] A. Schirmacher, J. Wess, and B. Zumino, Z. Phys. C **49**, 317 (1991); O. Ogievetsky and J. Wess, *ibid.* **50**, 123 (1991).
[20] V. K. Dobrev, J. Math. Phys. **33**, 3419 (1992).
[21] E. E. Demidov, Yu. I. Manin, E. E. Mukhin, and D. V. Zhdanovich, Prog. Theor. Phys. Suppl. **102**, 203 (1990).
[22] A. Schirmacher, Z. Phys. C **50**, 321 (1991).
[23] A. Sudbery, J. Phys. A **23**, L697 (1990).
[24] V. K. Dobrev and P. Parashar, J. Phys. A **26**, 6991 (1993).
[25] S. L. Woronowicz, Commun. Math. Phys. **122**, 125 (1989).
[26] E. H. Tahri, in preparation.

Relativistic Gamow vectors

I. Antoniou
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium
and Theoretische Natuurkunde, Free University of Brussels

M. Gadella
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium
and Departamento de Física Teórica, Facultad de Ciencias, E-47011, Valladolid, Spain

I. Prigogine
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium
and Center for Studies in Statistical Mechanics and Complex Systems, The University of Texas at Austin, Austin, Texas 78712

G. P. Pronko
International Solvay Institutes for Physics and Chemistry, C.P. 231, Campus Plaine ULB, Bd. du Triomphe, 1050 Brussels, Belgium
and Institute for High Energy Physics, Protvino, 142284 Moscow region, Russia

(Received 2 July 1997; accepted for publication 24 February 1998)

The Friedrichs model has often been used in order to obtain explicit formulas for eigenvectors associated to complex eigenvalues corresponding to lifetimes. Such eigenvectors are called Gamow vectors and they acquire meaning in extensions of the conventional Hilbert space of quantum theory to the so-called rigged Hilbert space. In this paper, Gamow vectors are constructed for a solvable model of an unstable relativistic field. As a result, we obtain a time asymmetric relativistic extension of the Fock space. This extension leads to two distinct Poincaré semigroups. The time reversal transformation maps one semigroup to the other. As a result, the usual PCT invariance should be extended. We show that irreversibility as expressed by dynamical semigroups is compatible with the requirements of relativity. © *1998 American Institute of Physics.* [S0022-2488(98)01206-7]

I. INTRODUCTION

As is well known, in order to describe the unstable quantum states, Gamow conjectured[1] the existence of eigenvectors of the Hamiltonian with complex eigenvalues. These eigenvalues were precisely the points at which the singularities of the S-matrix appear. However, this is impossible within the conventional formulation of quantum theory in Hilbert space[2,3] because the Hamiltonian, being a self-adjoint operator, can only have real eigenvalues. As a result, formal constructions of Gamow vectors appeared in the literature.[4–7] These Gamow vectors describe the scattering resonances.[8–10] In the case of simple poles,[10–12] the real part of the resonance corresponds to the resonance energy while the imaginary part corresponds to the half-width of the cross section, which is twice the mean life of the corresponding unstable state.

In the case of simple scattering resonances, Bohm and Gadella[11] constructed systematically the corresponding Gamow vectors and they provided a clear mathematical meaning to these vectors in terms of a suitable extension of quantum theory on the basis of Gel'fand triplets,[13] also called rigged Hilbert spaces, constructed with Hardy functions.[14] The extended Schrödinger evolution is ''intrinsically'' irreversible because it splits into two distinct semigroups.[11] This result is, moreover, a step towards a rigorous solution of the problem of irreversibility.[15] In fact, Bohm conjectured[16] that extensions to Gel'fand triplets may give meaning to the Brussels–Austin formalism of irreversibility developed under the direction of I. Prigogine.[15,17,18] This conjecture was verified[19] for the Friedrichs model, which has been extensively studied[17–19] as it leads to a time-symmetry breaking. The Bohm–Gadella Gamow vectors follow in fact from the intrinsic proper-

 © 1998 American Institute of Physics

ties of the extended evolution operators as they reduce in special cases[20,21] to the generalized eigenprojections constructed from subdynamics developed by the Brussels–Austin group.[20,22]

This method is, however, much more general. For both classical and quantum unstable systems independent of scattering, there exist intrinsically irreversible extensions of dynamics to suitable Gel'fand triplets which include the resonances as generalized eigenvalues.[20,21]

The purpose of this paper is to discuss relativistic resonances and the corresponding Gamow vectors. In this case, the transformation properties of the Gamow vectors with respect to Lorentz boosts pose nontrivial questions due to the appearance of complex eigenvalues. *The model we study in this paper is a relativistic Friedrich's model* which involves a quadratic coupling of a local scalar field with a bilocal scalar field. This model is presented in Sec. II and its solution given in Sec. III. The redefined vacuum is constructed in Sec. IV and the relativistic Gamow vectors are presented in Sec. V. Their meaning is given in terms of a rigged Fock space of Hardy functions following the similar construction in the nonrelativistic case.[11,19] Finally, we address the question of the action of the Poincaré semigroups on the Gamow state. Some calculations have been reported in the Appendices.

In the case of the Friedrich's model the time symmetry breaking is due to the resolution of the unstable eigenstate to the continuum.[17–19] Evolution is just a shift in the continuum. In order to describe the approach of the system to equilibrium such as the canonical distribution, we have to go to a statistical description that includes density operators in appropriate extensions of the conventional Hilbert space formulation. This formulation has been discussed in Refs. 22 and 23 for systems of nonrelativistic particles.

The main conclusion of this paper is that the construction of dynamical semigroups breaking the time symmetry is compatible with relativistic field theory. This result provides the first example, known to us, of an extension of the PCT theorem, because time reversal leads from one semigroup (for example, the one pointing towards the future) to the other (pointing towards the past).

II. THE MODEL

We shall consider an exactly solvable model for relativistic unstable quantum fields. This model involves a quadratic interaction between a local scalar field φ and a bilocal scalar field ψ. The interaction makes the local scalar field φ unstable. The model is an extension of the well-known Friedrich's model[4,7,17–19] where a single eigenstate interacts with the continuum and becomes unstable.

The local scalar field $\varphi(x)$ with mass M may be written in terms of creation and annihilation operators as follows:

$$\varphi(\mathbf{x},t)=\int d\tilde{\mathbf{k}}[a^{\dagger}(\mathbf{k})e^{ik\cdot x}+a(\mathbf{k})e^{-ik\cdot x}]. \tag{1}$$

By boldface letters we denote three-dimensional vectors. Four-dimensional vectors in Minkowski space are denoted by Roman style letters. The products of two vectors in Minkowski space are characterized by a dot. In Minkowski space we use the metric $+---$. For example, $k\cdot x=k_0x_0-\mathbf{kx}$. The Lorentz invariant measure in (1) is

$$d\tilde{\mathbf{k}}=\frac{d^3\mathbf{k}}{(2\pi)^3 2\omega(\mathbf{k})}, \quad \omega(\mathbf{k})=(\mathbf{k}+M^2)^{1/2}. \tag{2}$$

The creation and annihilation operators in (1) satisfy the usual commutation relations:

$$[a(\mathbf{k}),\ a^{\dagger}(\mathbf{k}')]=(2\pi)^3 2\omega(\mathbf{k})\,\delta(\mathbf{k}-\mathbf{k}'). \tag{3}$$

The Hamiltonian of the field $\varphi(x)$ is given by

$$\mathrm{H}_M=\int d\tilde{\mathbf{k}}\ \omega(\mathbf{k})a^{\dagger}(\mathbf{k})a(\mathbf{k}) \tag{4}$$

We shall consider, in addition, a bilocal scalar field, $\psi(x_\mu,q)$, with continuous mass spectrum. Bilocal fields were introduced first long ago by Yukawa[24] and Markov[25] in their discussion of extended elementary particles. It is used here in a different context. The field depends also on an additional real variable q representing an internal degree of freedom. For simplicity we shall assume that $\psi(x_\mu,q)$ is an even function of q. This means that

$$\psi(x_\mu,-q)=\psi(x_\mu,q). \tag{5}$$

The eigenvalues of the mass operator are defined by the internal variable. The simplest form of the mass operator $\mathbf{M}$ is given by

$$\mathbf{M}^2=4m^2-\frac{\partial^2}{\partial q^2}. \tag{6}$$

The field $\psi(x_\mu,q)$ satisfies then the following generalized Klein–Gordon equation:

$$(\Box-\mathbf{M}^2)\psi(x_\mu,q)=0, \tag{7}$$

where $\Box$ is the usual d'Alembert operator. The solution of Eq. (6) can be written in the following form:

$$\psi(x_\mu,q)=\int d^4k_\mu\int_{-\infty}^{\infty}d\kappa\,\frac{\cos\,q\kappa}{(2\pi)^4}\,e^{-ik_\mu x^\mu}\delta(k^2-4m^2-\kappa^2)B(k_\mu,\kappa), \tag{8}$$

where we took into account (5). The amplitude $B(k_\mu,\kappa)$ is also an even function of κ. Integrating the rhs of (7) over k_0 gives

$$\psi(x_\mu,q)=\int_{-\infty}^{\infty}d\kappa\,\frac{d^3\mathbf{k}\,\cos\,\kappa q}{(2\pi)^4 2E(\mathbf{k},\kappa)}\,(B^*(\mathbf{k},\kappa)e^{ik\cdot x}+B(\mathbf{k},\kappa)e^{-ik\cdot x}), \tag{9}$$

where $k_\mu=(E,\mathbf{k})$ and

$$E(\mathbf{k},\kappa)=[4m^2+\kappa^2+\mathbf{k}^2]^{1/2}. \tag{10}$$

We can change the variables in (8) so that E is the new independent variable instead of κ:

$$\kappa=(E^2-\mathbf{k}^2-4m^2)^{1/2},\quad \frac{d\kappa}{E}=\frac{dE}{\kappa}. \tag{11}$$

Then,

$$\psi(x_\mu,q)=\int_0^{\infty}\frac{dE\;d^3\mathbf{k}\,\cos\,\kappa(k_\mu)q}{(3\pi)^4\kappa(k_\mu)}\,(B^*(\mathbf{k},E)e^{ik\cdot x}+B(\mathbf{k},E)e^{-ik\cdot x}). \tag{12}$$

Following the routine of second quantization of a scalar field, we may introduce the Hamiltonian operator:

$$\mathrm{H}_0=\int d^3\mathbf{x}\;dq\;\frac{1}{2}\left[\dot{\psi}^2+(\boldsymbol{\nabla}\psi)^2+4m^2\psi^2+\left(\frac{\partial\psi}{\partial q}\right)^2\right]. \tag{13}$$

The canonical commutation relations are

$$[\psi(x,q),\;\pi(x',q')]=i\delta^4(x-x')\,\frac{\delta(q+q')+\delta(q-q)}{2}. \tag{14}$$

The canonical commutation relations follow from Dirac's quantization: The classical nonlocal field satisfies the Poisson bracket relation,

$$\{\psi(x,q),\pi(x',q')\}=-\delta^4(x-x')\,\frac{\delta(q+q')+\delta(q-q)}{2}, \tag{15}$$

which results from the canonical formalism.

The classical formula (11) gives the expression of the quantum field $\psi(x,q)$ in terms of the creation and annihilation operators $B^\dagger$ and B,

$$\psi(x_\mu,q)=\int_0^\infty \frac{dE\ d^3\mathbf{k}\ \cos\ \kappa(k_\mu)q}{(2\pi)^4\kappa(k_\mu)}\,[B^\dagger(\mathbf{k},E)e^{ik\cdot x}+B(\mathbf{k},E)e^{-ik\cdot x}], \tag{16}$$

with the following commutation relations:

$$[B(\mathbf{k},E),\ B^\dagger(\mathbf{k}',E')]=(2\pi)^4\kappa(k_\mu)\delta^4(k_\mu-k'_\mu). \tag{17}$$

We now introduce a quadratic interaction Hamiltonian as

$$\mathrm{H_{int}}=-\lambda\int d^3\mathbf{x}\int_{-\infty}^{\infty}dq\ \psi(x,q)f(q)\varphi(x), \tag{18}$$

where the even function $f(q)$ is Lorentz scalar and has the Fourier transform

$$f(q)=\int dy\ \alpha(y)\cos yq. \tag{19}$$

Clearly, $f(q)$ plays the role of a form factor and, to avoid the divergencies, we choose it to have good asymptotic behavior.

The representation of the ten generators of the Poincaré group in terms of creation and annihilation operators may be obtained as usual (see, e.g., Ref. 26). The Hamiltonian is

$$\begin{aligned}P_0=&\int\frac{d^3\mathbf{k}\ dE}{(2\pi)^4\kappa(\mathbf{k},E)}\,EB^\dagger(\mathbf{k},E)B(\mathbf{k},E)+\int\frac{d^3\mathbf{k}}{(2\pi)^3 2\omega(\mathbf{k})}\,\omega(\mathbf{k})a^\dagger(\mathbf{k})a(\mathbf{k})\\&+\int\frac{d^3\mathbf{k}\ dE}{(2\pi)^3 2\omega}\,\frac{\lambda\alpha(\kappa(\mathbf{k},E))}{\kappa(\mathbf{k},E)}\,(a(\mathbf{k})+a^\dagger(-\mathbf{k}))(B^\dagger(\mathbf{k},E)+B(-\mathbf{k},E)).\end{aligned} \tag{20}$$

The generator of three-dimensional translations is

$$\mathbf{P}=\int\frac{d^3\mathbf{k}\ dE}{(2\pi)^4\kappa(\mathbf{k},E)}\,\mathbf{k}B^\dagger(\mathbf{k},E)B(\mathbf{k},E)+\int\frac{d^3\mathbf{k}\ \mathbf{k}}{(2\pi)^3 2\omega(\mathbf{k})}\,a(\mathbf{k}), \tag{21}$$

The generators of three-dimensional rotations are

$$\begin{aligned}J_{ij}=&-i\int\frac{d^3\mathbf{k}}{(2\pi)^4\kappa(\mathbf{k},E)}\,B^\dagger(\mathbf{k},E)\left(k_j\frac{\partial}{\partial k_j}-k_j\frac{\partial}{\partial k_i}\right)B(\mathbf{k},E)\\&-i\int\frac{d^3\mathbf{k}}{(2\pi)^3 2\omega(\mathbf{k})}\,a^\dagger(\mathbf{k})\left(k_i\frac{\partial}{\partial k_j}-k_j\frac{\partial}{\partial k_i}\right)a(\mathbf{k}).\end{aligned} \tag{22}$$

The perturbation does not influence the spatial translation and rotations. However, the Lorentz-boost generators acquire an additional term due to the interaction:

$$\begin{aligned}J_{oi}=&\,i\int\frac{d^3\mathbf{k}\ dE}{(2\pi)^4\kappa(\mathbf{k},E)}\,B^\dagger(\mathbf{k},E)\left(E\frac{\partial}{\partial k_i}+k_i\frac{\partial}{\partial E}\right)B(\mathbf{k},E)+i\int\frac{d^3\mathbf{k}}{(2\pi)^3 2\omega(\mathbf{k})}\,a^\dagger\left(\omega(\mathbf{k})\frac{\partial}{\partial k_i}\right)a(\mathbf{k})\\&+\frac{i}{2}\int\frac{d^3\mathbf{k}\ dE}{(2\pi)^3}\,\frac{\lambda\alpha(\kappa(\mathbf{k},E))}{\kappa(\mathbf{k},E)}\left(E\frac{\partial}{\partial k_i}+k_i\frac{\partial}{\partial E}\right)\frac{B(\mathbf{k},E)+B^\dagger(-\mathbf{k},E)}{E}\,\frac{a(-\mathbf{k})+a^\dagger(\mathbf{k})}{\omega(\mathbf{k})}.\end{aligned} \tag{23}$$

The analogy between the Hamiltonian (20) and the interacting Hamiltonian in the usual Friedrich's model is obvious. We consider here an ''infinite collection'' of Friedrich's models, each corresponding to a value of the momentum **k**. This model is relativistic and contains both ''virtual'' and ''real'' interactions.

III. THE EIGENVALUE PROBLEM

The solution of the eigenvalue problem in our case is a straightforward generalization of the usual Friedrich's solution. The eigenvalue problem can be written as

$$[P_\mu,\ b^\dagger(E,\mathbf{k})]=k_\mu b^\dagger(E,\mathbf{k}), \tag{24}$$

where $k_\mu=(E,\mathbf{k})$ and P_μ is given by (19) and (20). Let $b^\dagger(E,\mathbf{k})$ be a linear combination of $B^\dagger(E,\mathbf{k}),B(E,\mathbf{k});a^\dagger(\mathbf{k}),a(\mathbf{k})$:

$$b^\dagger(E,\mathbf{k})=\int dE'\ (T(E,E',\mathbf{k})B^\dagger(E',\mathbf{k})+R(E,E',\mathbf{k})B(E',-\mathbf{k})) + t(E,\mathbf{k})a^\dagger(\mathbf{k})+r(E,\mathbf{k})a(-\mathbf{k}), \tag{25}$$

which due to the simple form of **P**-operator (24) satisfies

$$[\mathbf{P},\ b^\dagger(E,\mathbf{k})]=\mathbf{k}b^\dagger(E,\mathbf{k}). \tag{26}$$

Note that covariance with respect to three-dimensional rotations implies that coefficients in (24) depend only on $\mathbf{k}^2$. Equation (23) for $\mu=0$ gives

$$(E'-E)T(E,E',\mathbf{k})+\lambda\ \frac{\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\ (t(E,\mathbf{k})-r(E,\mathbf{k}))=0, \tag{27a}$$

$$-(E'+E)R(E,E',\mathbf{k})+\lambda\ \frac{\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\ (t(E,\mathbf{k})-r(E,\mathbf{k}))=0, \tag{27b}$$

$$(\omega-E)t(E,\mathbf{k})+\frac{\pi}{\omega}\int dE'\ \lambda\alpha(\kappa(E',\mathbf{k}))(T(E,E',\mathbf{k})-R(E,E',\mathbf{k}))=0, \tag{27c}$$

$$-(\omega+E)r(E,\mathbf{k})+\frac{\pi}{\omega}\int dE'\ \lambda\alpha(\kappa(E',\mathbf{k}))(T(E,E',\mathbf{k})-R(E,E',\mathbf{k}))=0. \tag{27d}$$

First, we express R and r in terms of T and t:

$$R(E,E',\mathbf{k})=-\frac{E'-E}{E'+E}\ T(E,E',\mathbf{k}),\quad r(E,\mathbf{k})=-\frac{\omega(\mathbf{k})-E}{\omega(\mathbf{k})+E}\ t(E,\mathbf{k}). \tag{28}$$

The equations for $T(E,E',\mathbf{k})$ and $t(E,\mathbf{k})$ become

$$(\omega-E)t(E,\mathbf{k})+\frac{\lambda\pi}{\omega(\mathbf{k})}\int dE'\,\alpha(\kappa(E',\mathbf{k}))\,\frac{2E'}{E'+E}\ T(E,E',\mathbf{k})=0, \tag{29a}$$

$$(E'-E)T(E,E',\mathbf{k})+\lambda\ \frac{\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\ \frac{2\omega(\mathbf{k})}{\omega(\mathbf{k})+E}\ t(E,\mathbf{k})=0. \tag{29b}$$

Equation (29b) gives

$$T(E,E',\mathbf{k})=C\,\delta(E'-E)-\frac{1}{E'-E}\ \frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\ \frac{2\omega(\mathbf{k})}{\omega(\mathbf{k})+E}\ t(E,\mathbf{k}). \tag{30}$$

Substituting (29) into (28) we obtain

$$\left[\omega^2-E^2-\int \frac{dE'}{E'^2-E^2}\,2E'\rho(E',\mathbf{k})\right]t(E,\mathbf{k})=-C\,\frac{\lambda\pi}{\omega(\mathbf{k})}\,\alpha(\kappa(E,\mathbf{k})), \tag{31}$$

where we have introduced the function $\rho(E,\mathbf{k})$:

$$\rho(E,\mathbf{k})=2\pi\,\frac{\lambda^2\alpha^2(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}. \tag{32}$$

Note that $\rho(E,\mathbf{k})$ depends only of $(E^2-\mathbf{k}^2)$ due to (10). The function in square brackets on the lhs of (30) is the relativistic generalization of the inverse one-particle resolvent. Equation (30) can be written as

$$\frac{1}{G(E,\mathbf{k})}\,t(E,\mathbf{k})=-C\,\frac{\lambda\pi}{\omega(\mathbf{k})}\,\alpha(\kappa(E,\mathbf{k})) \tag{33}$$

where

$$G(E,\mathbf{k})=\frac{1}{\omega^2-E^2-\Pi(E,\mathbf{k})}, \tag{34}$$

$$\Pi(E,\mathbf{k})=\int_{E_0}^{\infty}dE'\;2E'\;\frac{\rho(E',\mathbf{k})}{E'^2-E^2} \tag{35}$$

and $E_0=(4m^2+\mathbf{k}^2)^{1/2}$.

It is clear that the properties of $\Pi(E,\mathbf{k})$ and $G(E,\mathbf{k})$ depend upon the function $\rho(E,\mathbf{k})$ which is defined by the form factor $f(q)$. Equations (33) and (34) are also valid for complex values of E^2. We shall specify their meaning for real E^2 later. First of all, we shall suppose that the integral on the rhs of (34) exists and, moreover,

$$\Pi(E,\mathbf{k})\to 0 \quad \text{as} \quad E\to\infty. \tag{36}$$

From (34) we see that $\Pi(E,\mathbf{k})$ is an analytic function of E^2 with a cut on the physical sheet from $E^2=E_0^2$ to infinity. The jump of $\Pi(E,\mathbf{k})$ at the cut is given by

$$\Delta\Pi(E,\mathbf{k})=\Pi(E+i0,\mathbf{k})-\Pi(E-i0,\mathbf{k})=2\pi i\rho(E,\mathbf{k}). \tag{37}$$

We assume that $\rho(E,\mathbf{k})$ vanishes at the threshold

$$\rho(E_0(\mathbf{k}),\mathbf{k})=0. \tag{38}$$

The function $G(E,\mathbf{k})$ also has the same analytic properties, and its discontinuity has the form

$$\begin{aligned}\Delta G(E,\mathbf{k})&=G(E+i0,\mathbf{k})-G(E-i0,\mathbf{k})\\&=2\pi i\rho(E,\mathbf{k})G(E+i0,\mathbf{k})G(E-i0,\mathbf{k})=2\pi i\rho(E,\mathbf{k})|G(E,\mathbf{k})|^2.\end{aligned} \tag{39}$$

Therefore $G(E,\mathbf{k})$ satisfies the dispersion relation

$$G(E,\mathbf{k})=\int dE'^2\,\frac{1}{E'^2-E^2}\,\rho(E',\mathbf{k})[G(E',\mathbf{k})]^2, \tag{40}$$

This dispersion relation is actually the Källen–Lehmann representation for the propagator. Under the assumption (35), it is evident that

$$G(E,\mathbf{k})\to-\frac{1}{E^2} \tag{41}$$

as $E\to\infty$. From (40) and (39) we conclude therefore that

$$\int_{E_0^2}^{\infty} dE'^2\ \rho(E',\mathbf{k})|G(E',\mathbf{k})|^2=1. \tag{42}$$

From (35) we have

$$G(E,\mathbf{k})\to \frac{1}{\omega^2(\mathbf{k})-E^2}+o\left(\frac{1}{E^4}\right), \tag{43}$$

as $|\omega^2(\mathbf{k})-E^2|\to\infty$

Now, let us calculate the difference $G(E,\mathbf{k})-(\omega^2-E^2)^{-1}$:

$$G(E,\mathbf{k})-\frac{1}{\omega^2-E^2}=\int dE'^2\left(\frac{1}{E'^2-E^2}-\frac{1}{\omega^2-E^2}\right)\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2, \tag{44}$$

where we have used (43). Furthermore,

$$\begin{aligned} G(E,\mathbf{k})-\frac{1}{\omega^2-E^2}&=\frac{1}{\omega^2-E^2}\int dE'^2\,\frac{\omega^2-E'^2}{E'^2-E^2}\,\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2 \\ &\to\frac{1}{E^4}\int dE'^2(\omega^2-E'^2)\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2+o\left(\frac{1}{E^4}\right)\quad \text{as } E\to\infty. \end{aligned} \tag{45}$$

This result is compatible with (42) only if

$$\int_{E_0^2}^{\infty} dE'^2\ (\omega^2(\mathbf{k})-E'^2)\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2=0. \tag{46}$$

The sum rules (41) and (45) will be used in the sequel.

The unstable case which we are going to consider here corresponds to the extended Green's functions $G_+(E,\mathbf{k})$ and $G_-(E,\mathbf{k})$ from the upper to the lower half-plane and from the lower to the upper half-plane, respectively.

In the unstable case, $G_+(E,\mathbf{k})$ has a pole at $E^2=\mathbf{k}^2+\mu_c^2$ and $G_-(E,\mathbf{k})$ has a pole at the conjugate point $E^2=\mathbf{k}^2+\mu_c^{*2}$, where

$$\mu_c^2=\mu^2-i\mu\Gamma \tag{47}$$

and μ and Γ are real positive numbers defined by the form factor $f(q)$.

The equation for the complex pole μ_c^2 is

$$\omega^2(\mathbf{k})-E^2-\int dE'^2\,\frac{\rho(E',\mathbf{k})}{E'^2-E^2}=0$$

or

$$M^2-(E^2-\mathbf{k}^2)-\mathscr{P}\int dE'^2\,\frac{\rho(E',\mathbf{k})}{E'^2-E^2}-i\pi\rho(E,\mathbf{k})=0, \tag{48}$$

where $\mathscr{P}$ stands for the Cauchy principal value. For small values of the coupling constant λ we can omit the integral term

$$\mu_c^2=M^2-i\pi\rho(E,\mathbf{k})|_{E^2-\mathbf{k}^2=M^2}=M^2-2i\pi^2\lambda^2\,\frac{[\alpha(\sqrt{M^2-4m^2})]^2}{\sqrt{M^2-4m^2}}. \tag{49}$$

The pole appears only for $M\geqslant 2m$.

With Green's function, $G(E,\mathbf{k})$, we can proceed further and find the general solution of (27):

$$t(E,\mathbf{k}) = -C\lambda\pi\alpha(\kappa(E,\mathbf{k}))G(E,\mathbf{k})\,\frac{E+\omega(\mathbf{k})}{\omega(\mathbf{k})},$$

$$T(E,E',\mathbf{k}) = C\left[\delta(E-E') + 2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(E,\mathbf{k})\,\frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\,\frac{1}{E'-E}\right], \tag{50}$$

$$R(E,E',\mathbf{k}) = -C2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(E,\mathbf{k})\,\frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\,\frac{1}{E'+E},$$

$$r(E,\mathbf{k}) = -C\,\frac{E-\omega(\mathbf{k})}{\omega(\mathbf{k})}\,\lambda\pi\alpha(\kappa(E,\mathbf{k}))G(E,\mathbf{k}).$$

The formal solution of eigenvalue problem (23) is, for positive energies,

$$b^{\dagger}(E,\mathbf{k}) = C\Bigg\{B^{\dagger}(E,\mathbf{k}) + 2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(E,\mathbf{k})\Bigg[\int dE'\,\frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\,(E',\mathbf{k}) \times\left(\frac{B^{\dagger}(E,\mathbf{k})}{E'-E} - \frac{B(E,-\mathbf{k})}{E'+E}\right) - \frac{(E+\omega(\mathbf{k}))a^{\dagger}(\mathbf{k}) + (E-\omega(\mathbf{k}))a(-\mathbf{k})}{2\omega(\mathbf{k})}\Bigg]\Bigg\}. \tag{51}$$

The negative energy solution of (23) which corresponds to the annihilation operator $b(E,\mathbf{k})$ is

$$b(E,\mathbf{k}) = C\Bigg\{B(E,\mathbf{k}) + 2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(e,\mathbf{k})\Bigg[\int dE'\,\frac{\lambda\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\left(\frac{B(E,\mathbf{k})}{E'-E} - \frac{B^{+}(E,-\mathbf{k})}{E'+E}\right) - \frac{(E+\omega(\mathbf{k}))a(\mathbf{k}) + (E-\omega(\mathbf{k}))a^{+}(-\mathbf{k})}{2\omega(\mathbf{k})}\Bigg]\Bigg\}. \tag{52}$$

We call this solution formal because up to now we did not specify the singular integrals in (50) and in the definition of Green's function for real energy. We have different solutions of (23). These correspond to the incoming and outgoing boundary conditions. The incoming operators are

$$b^{\dagger}_{\mathrm{in}}(E,\mathbf{k}) = B^{\dagger}(E,\mathbf{k}) + 2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(E+i0,\mathbf{k})\Bigg[\int dE'\,\frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\left(\frac{B^{\dagger}(E,\mathbf{k})}{E'-E-i0} - \frac{B(E',-\mathbf{k})}{E'+E}\right) - \frac{(E+\omega(\mathbf{k}))a^{\dagger}(\mathbf{k}) + (E-\omega(\mathbf{k}))a(-\mathbf{k})}{2\omega(\mathbf{k})}\Bigg],$$

$$b_{\mathrm{in}}(E,\mathbf{k}) = B(E,\mathbf{k}) + 2\pi\lambda\alpha(\kappa(E,\mathbf{k}))G(E-i0,\mathbf{k})\Bigg[\int dE'\,\frac{\lambda\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\left(\frac{B(E',\mathbf{k})}{E'-E+i0} - \frac{B^{\dagger}(E',-\mathbf{k})}{E'+E}\right) - \frac{(E+\omega(\mathbf{k}))a(\mathbf{k}) + (E-\omega(\mathbf{k}))a^{\dagger}(-\mathbf{k})}{2\omega(\mathbf{k})}\Bigg], \tag{53}$$

$$E \geqslant E_0(\mathbf{k}) = (4m^2 + \mathbf{k}^2)^{1/2},$$

where we have set the normalization constant $c=1$. The outgoing set of operators corresponds to change $+i0 \to -i0$ in (52).

We end this section with a remark concerning the solution of (23) in the stable case. The latter corresponds to Green's function $G(E,\mathbf{k})$ with simple pole on real axis E which lies below threshold $E=E_0$ in the point $E=\omega(\mathbf{k})$. In this case there appears an additional solution of Eqs. (28) which corresponds to perturbed operators $\tilde{a}(\mathbf{k})$ and $\tilde{a}^{\dagger}(\mathbf{k})$. Together with the operators $b(E,\mathbf{k})$ and $b^{\dagger}(E,\mathbf{k})$, these operators form a complete set which is connected with the initial operators $a(k), a^{\dagger}(\mathbf{k}); B(E,\mathbf{k}), B^{\dagger}(E,\mathbf{k})$ through the Bogolubov transformation. In the unstable case the operators $\tilde{a}(\mathbf{k})$ and $a^{\dagger}(\mathbf{k})$ do not appear although the Bogolubov transformation exists.

IV. THE VACUUM STATE

For the commutation relations of the incoming and outgoing operators $b^{\text{in}}_{\text{out}}(E,\mathbf{k})$ and $b^{\dagger}_{\substack{\text{in}\\\text{out}}}(E,\mathbf{k})$, we have to use the explicit form of $b(E,\mathbf{k})$ and the fundamental commutation relations of the initial operators (3) and (16) together with the properties of Green's function $G(E,\mathbf{k})$ (see Appendix A):

$$[b_{\text{in}}(E,\mathbf{k}),\ b^{\dagger}_{\text{in}}(E',\mathbf{k}')]=[b_{\text{out}}(E,\mathbf{k}),\ b^{\dagger}_{\text{out}}(E',\mathbf{k})]=(2\pi)^4\kappa(E,\mathbf{k})\,\delta(E-E')\,\delta^3(\mathbf{k}-\mathbf{k}'),$$

$$[b_{\text{out}}(E,\mathbf{k}),\ b^{\dagger}_{\text{in}}(E',\mathbf{k}')]=(2\pi)^4\kappa(E,\mathbf{k})\,\delta(E-E')\,\delta^3(\mathbf{k}-\mathbf{k}')\,\frac{G(E+i0,\mathbf{k})}{G(E-i0,\mathbf{k})}, \tag{54}$$

$$[b_{\text{in}}(E,\mathbf{k}),\ b^{\dagger}_{\text{out}}(E',\mathbf{k})]=(2\pi)^4\kappa(E,\mathbf{k})\,\delta(E-E')\,\delta^3(\mathbf{k}-\mathbf{k}')\,\frac{G(E-i0,\mathbf{k})}{G(E+i0,\mathbf{k})};$$

all other commutators vanish.

In terms of creation and annihilation operators, the energy momentum vector is diagonal, as follows from (25) and from the fact that b and $b^{\dagger}$ solve the eigenvalue problem. A direct calculation gives

$$P_{\mu}=\int\ \frac{d^3\mathbf{k}\,dE}{(2\pi)^4\kappa(E,\mathbf{k})}\,k_{\mu}b^{\dagger}_{\substack{\text{in}\\\text{out}}}(E,\mathbf{k})b^{\text{in}}_{\text{out}}(E,\mathbf{k}), \tag{55}$$

where $k_{\mu}=(E,\mathbf{k})$ and the lhs of (54) is given by Eqs. (19) and (20) up to a constant in P_0, which arises due to operator ordering. A further important property of the new operators is the way they change under Lorentz transformations. Three-dimensional rotations present no difficulties and therefore we present only the commutation rules of the b-operators with the Lorentz boosts:

$$[J_{oi},\ b^{\dagger}_{\substack{\text{in}\\\text{out}}}(E,\mathbf{k})]=i\left(E\,\frac{\partial}{\partial k_i}+k_i\,\frac{\partial}{\partial E}\right)b^{\dagger}_{\substack{\text{in}\\\text{out}}}(E,\mathbf{k}). \tag{56}$$

The commutation relation (53) provides a simple way for the construction of the space of states which is the subspace of the whole Fock space formed by the operators $B^{\dagger}(E,\mathbf{k})$ and $a^{\dagger}(\mathbf{k})$ and the initial vacuum state $|0\rangle$ which was defined by

$$B(E,\mathbf{k})|0\rangle=0,\quad a(\mathbf{k})|0\rangle=0. \tag{57}$$

The explicit form of $b(E,k)$ shows us that

$$b(E,\mathbf{k})|0\rangle\neq 0, \tag{58}$$

i.e., there arises a new vacuum state which will be a superposition of states with an arbitrary number of particles of B and a types. The new vacuum Ω has the following form:

$$\Omega=e^{V}|0\rangle, \tag{59}$$

where V is a functional of creation operators $B^{\dagger}(E,\mathbf{k})$ and $a^{\dagger}(\mathbf{k})$.

It should be noted that the operator e^{V} is a part of the general unitary Bogolubov transformation which, besides dressing the vacuum state, transforms also the operators $B(E,\mathbf{k})$ and $a(\mathbf{k})$ into $b(E,\mathbf{k})$. In order to simplify the calculations, we have subdivided the problem into two parts, (1) the construction of the creation and annihilation operators and (2) the introduction of the new vacuum state. Moreover, from the standard theory of Bogolyubov transformations, it follows that the functional V, which gives the solution of the equation

$$b_{\text{in}}(E,\mathbf{k})\Omega=0, \tag{60}$$

is at most a quadratic functional of $B^{\dagger}(E,\mathbf{k})$ and $a^{\dagger}(\mathbf{k})$:

$$V(B^\dagger,a^\dagger)=\int d^3\mathbf{k}\Bigg\{\frac{1}{2}\int dE\ dE'\ f(E,E',\mathbf{k})B^\dagger(E,\mathbf{k})B^\dagger(E,-\mathbf{k}) + \int dE\ B^\dagger(E,\mathbf{k})a^\dagger(-\mathbf{k})f(E,\mathbf{k})+\frac{1}{2}f(\mathbf{k})a^\dagger(\mathbf{k})a^\dagger(-\mathbf{k})\Bigg\}. \tag{61}$$

In Appendix B, we shall see how to obtain the functions $f(E,E',\mathbf{k})$, $f(E,\mathbf{k})$, and $f(\mathbf{k})$. The most important part of the solution is the problem of factorization of Green's function $G(E,\mathbf{k})$ which we introduced in Sec. II. The factorization problem is to find a function $\gamma(E,\mathbf{k})$, analytic in the right half-plane, such that

$$\gamma(E,\mathbf{k})\gamma(-E,\mathbf{k})=G(E,\mathbf{k}) \tag{62}$$

on the whole complex E plane. The additional conditions which guarantee the uniqueness of $\gamma(E,\mathbf{k})$ are discussed in Appendix C. The most important properties of the function $\gamma(E,\mathbf{k})$ are given by the equations

$$\frac{1}{\gamma(E,\mathbf{k})}+E+\omega(\mathbf{k})+2\xi(\mathbf{k})+\int_{E_0}^{\infty}dE'\ \frac{\rho(E',\mathbf{k})\gamma(E',\mathbf{k})}{E'+E}=0, \tag{63}$$

$$\int_{E_0}^{\infty}dE'\ \rho(E',\mathbf{k})\gamma(E',\mathbf{k})=2\xi(\mathbf{k})(\omega(\mathbf{k})+\xi(\mathbf{k})), \tag{64}$$

$$\left(E\frac{\partial}{\partial k_i}+k_i\frac{\partial}{\partial E}\right)\gamma(E,\mathbf{k})+\gamma(E,\mathbf{k})\frac{\partial}{\partial k_i}(\omega(\mathbf{k})+2\xi(\mathbf{k}))=0, \tag{65}$$

where $\rho(E,\mathbf{k})$ has been defined in (22) and $\omega=(\mathbf{k}^2+M^2)^{1/2}$. The new variable $\xi(\mathbf{k})$ describes the shift of the second term of the asymptotic expansion of $\gamma(E,\mathbf{k})$ for $E\to\infty$ due to the interaction. The detailed definition of $\xi(\mathbf{k})$ as well as the derivation of the properties of $\gamma(E,\mathbf{k})$ are discussed in Appendix C.

In terms of the solution of the factorization problem, $\gamma(E,\mathbf{k})$, the functions which define the functional V have the form

$$\begin{aligned} f(E,E',\mathbf{k})&=\frac{1}{(2\pi)^3 2(\omega(\mathbf{k})+\xi(\mathbf{k}))}\,q(E,\mathbf{k})q(E',\mathbf{k})\left(1+\frac{2(\omega(\mathbf{k})+\xi(\mathbf{k}))}{E'+E}\right),\\ f(E,\mathbf{k})&=\frac{1}{(2\pi)^3 2(\omega(\mathbf{k})+\xi(\mathbf{k}))}\,q(E,\mathbf{k}),\\ f(\mathbf{k})&=\frac{1}{(2\pi)^3 2(\omega(\mathbf{k})+\xi(\mathbf{k}))}\left(-\frac{\xi(\mathbf{k})}{\omega(\mathbf{k})}\right),\\ q(E,\mathbf{k})&=\frac{\lambda\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\,\gamma(E,\mathbf{k}). \end{aligned} \tag{66}$$

Substituting (65) into the definition of V we finally arrive at

$$V(B^\dagger,a^\dagger)=\int\frac{d^3\mathbf{k}}{(2\pi)^3 2(\omega(\mathbf{k})+\xi(\mathbf{k}))}\cdot\Bigg\{\frac{1}{2}\int_{E_0}^{\infty}dE\ dE'\ q(E,\mathbf{k})q(E',\mathbf{k}) \times\left(1+\frac{2(\omega(\mathbf{k})+\xi(\mathbf{k}))}{E+E'}\right)B^\dagger(E,\mathbf{k})B^\dagger(E,-\mathbf{k}) + \int_{E_0}^{\infty}dE\ q(E,\mathbf{k})B^\dagger(E,\mathbf{k})a^\dagger(-\mathbf{k})-\frac{\xi(\mathbf{k})}{2\omega(\mathbf{k})}\,a^\dagger(\mathbf{k})a^\dagger(-\mathbf{k})\Bigg\}. \tag{67}$$

This operator V defines the new Lorentz invariant vacuum state. Note that Ω is a new vacuum for the incoming–outgoing operators $b(E,\mathbf{k})$. The creation operators $b^\dagger(E,\mathbf{k})$ act on Ω and define a new Fock space. In particular the one-particle state is

$$b^\dagger_{\rm in}(E,\mathbf{k})\Omega=\left\{B^\dagger(E,\mathbf{k})+\frac{2\pi\lambda\,\alpha(\kappa(E,\mathbf{k}))}{2(\omega(\mathbf{k})+\xi(\mathbf{k}))}\;\gamma(-E-i0,\mathbf{k})\right.$$
$$\left.\times\left[\int dE'\;q(E',\mathbf{k})\left(1+\frac{2(\omega(\mathbf{k})+\xi(\mathbf{k}))}{E'-E-i0}\right)B^\dagger(E,\mathbf{k})+a^\dagger(\mathbf{k})\right]\right\}\Omega, \tag{68}$$

where we have used explicit form of $B^\dagger(E,\mathbf{k})$ and Ω to eliminate the annihilation operators $B(E,\mathbf{k})$ and $a(\mathbf{k})$.

Comparing Eq. (67) with its nonrelativistic analog,[18] we see that the role of the partial resolvent is now played by $\gamma(-(E+i0),\mathbf{k})$. In particular, Green's function has the following behavior on the vicinity of its pole on the lower half-plane with respect to the Lorentz invariant variable $z=E^2-\mathbf{k}^2$:

$$G_+(E,\mathbf{k})=\gamma_+(E)\,\gamma_+(-E)=\frac{c^2}{-E^2+(\mu_c^2+\mathbf{k}^2)}+\text{regular terms.} \tag{69}$$

At this point $z=\mu_c^2$ the solution of the factorization problem with required asymptotic properties has the form

$$\gamma_+(E)=-\frac{c}{E+(\mu_c^2+\mathbf{k}^2)^{1/2}}+\text{regular terms},$$
$$\gamma_+(-E)=+\frac{c}{E-(\mu_c^2+\mathbf{k}^2)^{1/2}}+\text{regular terms}. \tag{70}$$

V. THE GAMOW VECTORS

Now we are ready to discuss the main point of the present paper: the eigenvectors corresponding to the unstable relativistic field. Consider, for example, the state (67). Due to the presence of $\gamma(-E-i0,\mathbf{k})$, the state

$$\Phi_{\rm in}(E,\mathbf{k})=b^\dagger_{\rm in}(E,\mathbf{k})\Omega \tag{71}$$

has a pole in the second sheet of complex E-plane where we make the analytic extension from above to below. This statement, as well as all other states in the present section, acquire mathematical meaning in the framework of rigged Hilbert spaces[11,13] and rigged Fock spaces.[27] On the energy plane, the resonance pole z_R, which corresponds to the pole μ_c^2 on the plane of the Lorentz invariant variable $E^2-\mathbf{k}^2$, is

$$z_R=(\mathbf{k}+\mu^2-i\mu\Gamma)^{1/2}=A(\mathbf{k})-iB(\mathbf{k}), \tag{72}$$

where

$$A(\mathbf{k})=\left\{\frac{[(\mathbf{k}^2+\mu^2)^2+\mu^2\Gamma^2]^{1/2}+(\mathbf{k}^2+\mu^2)}{2}\right\}^{1/2},$$
$$B(\mathbf{k})=\left\{\frac{[(\mathbf{k}^2+\mu^2)^2+\mu^2\Gamma^2]^{1/2}-(\mathbf{k}^2+\mu^2)}{2}\right\}^{1/2}.$$

Note that, because the imaginary part of z_R defines the decay law, the life time is also $\mathbf{k}$ dependent. Usually, it is supposed that

$$\Gamma(\mathbf{k}) = \Gamma \frac{m}{\omega(\mathbf{k})},$$

which follows from the Lorentz transformation of time scale. From (71) follows that this rather naive answer is valid only for $\Gamma \to 0$.

The state $\Phi_{\text{in}}(E,\mathbf{k})$ can be written in a neighborhood of z_R as

$$\Phi_{\text{in}}(E,\mathbf{k}) = \frac{1}{E - z_R} \varphi^G_{\text{in}}(\mathbf{k}) + \text{regular part.} \tag{73}$$

Here, $\varphi^G_{\text{in}}(\mathbf{k})$ is the Gamow vector associated with the resonance pole z_R.

The Gamow vectors are linear functionals over a suitable space of ''test functions.'' As was done in Ref. 27, this space can be written in the following form:

$$\mathbf{Y}_- = C \oplus \Phi \oplus (\Phi \otimes \Phi) \oplus (\Phi \otimes \Phi \otimes \Phi) \oplus \cdots . \tag{74}$$

Following our previous work[11,21] we may choose as a representation space for the one-particle test function space Φ the following:

$$\Psi = (\mathscr{H}^2_- \cap S) \otimes S(R^3), \tag{75}$$

where the following hold.

(i) $\mathscr{H}^2_-$ is the space of Hardy functions on the lower half-plane.[14]
(ii) S is the space of all real-valued complex functions that are differentiable at all orders such that they as well as their derivatives vanish at the infinity faster than the inverse of any polynomial (the one-dimensional Schwartz space).
(iii) $S(R^3)$ represents the space of all complex functions on the three-dimensional real space R^3 having the same properties as S in (ii).
(iv) The tensor product that we are using in (74) and throughout this paper is the standard π-tensor product of topological vector spaces.[28]

A typical function of $(\mathscr{H}^2_- \cap S) \otimes S(R^3)$ is of the form $\varphi(E,\mathbf{k})$. For a fixed value of the momentum $\mathbf{k}$, this is a function of the energy and belongs to $\mathscr{H}^2_- \cap S$. For a fixed value of the energy E, $\varphi(E,\mathbf{k}) \in S(R^3)$.

Then, if $\varphi(E,\mathbf{k})$ is in the space $(\mathscr{H}^2_- \cap S) \otimes S(R^3)$, we can write explicitly the vectors in Φ (written in bra form) as

$$\langle \varphi | = \int \frac{d^3\mathbf{k}\, dE}{(2\pi)^4 \kappa(E,\mathbf{k})} \langle \Omega | b_{\text{out}}(E,\mathbf{k}) \varphi(E,\mathbf{k}). \tag{76}$$

The above integral goes from E_0 to ∞, which in particular means that the energy is always positive. It is also important to remark that a Hardy function on the lower half-plane, $\varphi_-(E)$, is completely defined by its given boundary values on the semiaxis $[E_0,\infty)$ (and the same occurs for a Hardy function on the upper half-plane). The reason is very simple: take $f(E) = \varphi_-(E + E_0)$, which is also a function in $\mathscr{H}^2_-$. After the van Winter theorem, $f(E)$ is determined by its boundary values on the positive semiaxis and those coincide with the values of $\varphi_-(E)$ on $[E_0,\infty)$.

Thus, there is a bijective (one-to-one and onto) correspondence between Ψ and Φ, so that we can transport the topology from Ψ to Φ. As mentioned earlier, $\Phi \otimes \Phi$ has the π-tensor product topology.

The space $\mathbf{Y}_-$ can be endowed with a stronger topology than the Fock space topology. The action of a functional, F, on $\mathbf{y} \in \mathbf{Y}_-$ is represented here as $\langle \mathbf{y} | F \rangle$.

In order to construct the topology on $\mathbf{Y}_-$, we note that the space Φ_n, which is defined as n times the π-tensor product of Φ by itself, is algebraically isomorphic and topologically homeomorphic to a proper subspace of Φ_{n+1}. Thus, we can write $\Phi_n \subset \Phi_{n+1}$. In this sense, we can say that the original topology of Φ_n coincides with the topology it has as a subspace of Φ_{n+1}. Therefore, from this point of view, it makes sense to write

$$\mathbf{Y}_- = \bigcup_{n=0}^{\infty} \Phi_n, \quad \Phi_0 = C,$$

and to endow $\mathbf{Y}_-$ with the strict inductive limit topology induced by the Φ_n. Then, $\mathbf{Y}_-$ has the following properties:

(i) The spaces Φ_n are nuclear because they are π-products of nuclear spaces. Since the strict inductive limit of nuclear spaces is nuclear, we conclude that $\mathbf{Y}_-$ is nuclear.

(ii) Since Φ is a Fréchet (metrizable) space, all Φ_n, $n=0,1,2,...$, are Fréchet. Therefore, $\mathbf{Y}_-$ is a LF space.[28]

(iii) By the dual space, $\mathbf{Y}_-^\times$, we mean the strong dual of $\mathbf{Y}_-$. Since Φ is a Fréchet space, $(\Phi \otimes \Phi \otimes \cdots \otimes \Phi)^\times = \Phi^\times \otimes \Phi^\times \otimes \cdots \otimes \Phi^\times$. This is a proper subspace of $\mathbf{Y}_-^\times$, so that

$$C \oplus \Phi^\times \oplus (\Phi^\times \otimes \Phi^\times) \oplus \cdots \subset \mathbf{Y}^\times.$$

The Gamow vector $\varphi_{\text{in}}^G(\mathbf{k})$ is a continuous functional on $\mathbf{Y}_-$. When applied to a space of the form C or $\Phi \otimes \cdots \otimes \Phi$ it gives zero, provided that the tensor product involves two or more copies of Φ. On Φ itself $\varphi_{\text{in}}^G(\mathbf{k})$ is nonzero.

Taking into account the commutation relations (53), we can compute the bracket $\langle \varphi | \Phi_{\text{in}}(E,\mathbf{k}) \rangle$:

$$\begin{aligned}
\langle \varphi | \Phi_{\text{in}}(E,\mathbf{k}) \rangle &= \int \frac{dE'\ d^3\mathbf{k}'}{(2\pi)^4 \kappa(E',\mathbf{k}')} \langle \Omega | b_{\text{out}}(E',\mathbf{k}') \varphi(E',\mathbf{k}') b_{\text{in}}^\dagger(E,\mathbf{k}) | \Omega \rangle \\
&= \int dE' t d^3\mathbf{k}'\ \varphi(E'\mathbf{k}') \delta(E-E') \delta(\mathbf{k}-\mathbf{k}') \frac{G(E+i0,\mathbf{k})}{G(E-i0,\mathbf{k})} \langle \Omega | \Omega \rangle \\
&= \varphi(E,\mathbf{k}) \frac{G(E+i0,\mathbf{k})}{G(E-i0,\mathbf{k})} = \varphi(E,\mathbf{k}) \frac{\gamma(E)\gamma(-(E+i0))}{\gamma(E)\gamma(-(E-i0))} = \varphi(E,\mathbf{k}) \frac{\gamma_+(-E)}{\gamma_-(-E)},
\end{aligned} \tag{77}$$

which is well defined because $\Phi_{\text{in}}(E,\mathbf{k})$ is a continuous functional on $\mathbf{Y}_-$. This equation shows that $\langle \varphi | \Phi_{\text{in}}(E,\mathbf{k}) \rangle$ admits a meromorphic extension to the lower half-plane having a pole at z_R.

So far, we have made our calculations for real values of E. If we want to to find the pole of the analytic continuation of (74) from above to below, we must remember that the continuation $\gamma(-(E+i0)) = \gamma_+(-E)$ has a pole at z_R with residue c and that the continuation of $\gamma(-(E+i0)) = \gamma_+(-E)$ has neither a singularity nor a zero at that point. Therefore, on a neighborhood of z_R,

$$\frac{\gamma_+(-E)}{\gamma_-(-E)} \quad \text{behaves as} \quad \frac{c}{E-z_R} \frac{1}{\gamma_-(-E)},$$

and we obtain finally

$$\langle \varphi | \varphi_{\text{in}}^G(\mathbf{k}) \rangle = \varphi(E,\mathbf{k}) \left. \frac{c}{\gamma_-(-E)} \right|_{E=z_R}. \tag{78}$$

Since

$$P_0 b_{\text{out}}^\dagger(E,\mathbf{k})\Omega = E b_{\text{out}}^\dagger(E,\mathbf{k})\Omega, \tag{79}$$

we immediately conclude that the test function space $\mathbf{Y}_-$ is stable under P_0, i.e., $P_0 \mathbf{Y}_- \subset \mathbf{Y}_-$. In addition, one can show that P_0 is continuous on $\mathbf{Y}_-$. Therefore, P_0 can be extended continuously to the dual of $\mathbf{Y}_-$ (the space of continuous functionals on $\mathbf{Y}_-$). Then, the action of P_0 on the Gamow state $\varphi_{\text{in}}^G(\mathbf{k})$ gives the expected formula:

$$P_0 \varphi_{\text{in}}^G(\mathbf{k}) = z_R \varphi_{\text{in}}^G(\mathbf{k}). \tag{80}$$

As in nonrelativistic quantum mechanics, we can extend the evolution operator e^{-itP_0} to the dual space of $\mathbf{Y}_-$, $\mathbf{Y}_-^\times$ for certain values of time only. Formula (79) and the properties of Hardy functions establish that $\mathbf{Y}_-$ is stable under e^{itP_0} if and only if $t\geqslant 0$. The proof of this statement is identical in the nonrelativistic case and is given in Ref. 11. Therefore, e^{-itP_0} can act on $\varphi_{\text{in}}^G(\mathbf{k})$ if and only if $t\geqslant 0$ and gives

$$e^{-itP_0}\varphi_{\text{in}}^G(\mathbf{k})=e^{-itz_R}\varphi_{\text{in}}^G(\mathbf{k}); \quad t\geqslant 0, \tag{81}$$

which shows the exponential decay of the Gamow vector $\varphi_{\text{in}}^G(\mathbf{k})$.

The same procedure can be done for the outgoing Gamow vectors $\varphi_{\text{out}}^G(\mathbf{k})$. In this case, we have to replace the space of Hardy functions on the lower half-plane by the space $\mathscr{H}_+^2$ of Hardy functions on the upper half-plane and b_{in} by b_{out}. Also, the following properties hold:

$$P_0\varphi_{\text{out}}^G(\mathbf{k})=z_R^*\varphi_{\text{out}}^G(\mathbf{k}), \quad e^{-itP_0}\varphi_{\text{out}}^G(\mathbf{k})=e^{-itz_R^*}\varphi_{\text{out}}^G(\mathbf{k}) \quad \text{for } t\leqslant 0 \text{ only.} \tag{82}$$

In this case, $\mathbf{Y}_+$ is the space of test functions and $\mathbf{Y}_+^\times$ its dual.

As in the nonrelativistic case, the group e^{-itP_0}, when extended to the duals, splits into two distict semigroups, one for $\mathbf{Y}_-^\times$ and the other one for $\mathbf{Y}_+$. This demonstrates the intrinsic irreversibility of the relativistic Friedrich's model. The implications for the Poincaré group will be discussed in the next section.

Going back to the definition of creation and annihilation operators $b_{\text{in(out)}}^\dagger(E,\mathbf{k})$ and $b_{\text{in(out)}}(E,\mathbf{k})$ we can also define operators which ''create'' and ''annihilate'' Gamow states. Indeed, let us consider

$$\psi^\dagger(E,\mathbf{k})=\int dE'\,\frac{\lambda\alpha(E',\mathbf{k})}{\kappa(E',\mathbf{k})}\left(\frac{B^\dagger(E',\mathbf{k})}{E'-E}-\frac{B(E',-\mathbf{k}')}{E'+E}\right)-\frac{(E+\omega(\mathbf{k}))a^\dagger(\mathbf{k})+(E-\omega(\mathbf{k}))a(-\mathbf{k})}{2\omega(\mathbf{k})} \tag{83}$$

and

$$\psi(E,\mathbf{k})=\int dE'\,\frac{\lambda\alpha(E',\mathbf{k})}{\kappa(E',\mathbf{k})}\left(\frac{B(E',\mathbf{k})}{E'-E}-\frac{B^\dagger(E',-\mathbf{k})}{E'+E}\right)-\frac{(E+\omega(E',\mathbf{k}))a(\mathbf{k})+(E-\omega(E',\mathbf{k}))a^\dagger(\mathbf{k})}{2\omega(E',\mathbf{k})}. \tag{84}$$

The analytic continuation described above for the state $\psi^+(E,\mathbf{k})\Omega$ produces the states $\varphi_G^{\text{in}}(\mathbf{k})$ or $\varphi_G^{\text{out}}(\mathbf{k})$, and we can consider $\psi^+(\mathbf{k},z_R)$ and $\psi^+(\mathbf{k},z_R^*)$ as the operators which create Gamow ket states. In the same sense $\psi(\mathbf{k},z_R)$ and $\psi(\mathbf{k},z_R^*)$ create Gamow bra states. The nonvanishing commutation relations for these operators are

$$\begin{aligned}&[\psi(\mathbf{k},z_R), \quad \psi^+(\mathbf{k}',z_R)]=(2\pi)^3\delta(\mathbf{k}-\mathbf{k}')2z_RZ,\\ &[\psi(\mathbf{k},z_R^*),\ \psi^+(\mathbf{k}',z_R^*)]=(2\pi)^3\delta(\mathbf{k}-\mathbf{k}')2z_R^*Z^*,\end{aligned} \tag{85}$$

where

$$Z=\frac{d}{dE^2}\,G^{-1}(E,\mathbf{k})|_{E^2-\mathbf{k}^2=\mu_c^2}.$$

These commutation relation may be a starting point for the introduction of a Gamow field, which we intend to consider elsewhere.

VI. THE POINCARÉ SEMIGROUPS

Now let us see how the Gamow vectors transform under Lorentz boosts. Commutation relations (53) together with the Lorentz invariance of vacuum state give us the action of the boost generators on the state (70):

$$J_{oi}\Phi_{\text{in}}(E,\mathbf{k})=i\left(k_i\frac{\partial}{\partial E}+E\frac{\partial}{\partial k_i}\right)\Phi_{\text{in}}(E,\mathbf{k}). \tag{86}$$

The Lorentz transformation generated by J_{0i} is

$$U(\alpha)\Phi_{\text{in}}(\mathbf{k},E)=\exp\{i\alpha_i J_{oi}\}\Phi_{\text{in}}(\mathbf{k},E)=\Phi_{\text{in}}(\mathbf{k}'_\alpha,E'_\alpha), \tag{87}$$

where

$$\begin{aligned}E'_\alpha&=E\cosh\alpha-(\mathbf{nk})\sinh\alpha,\\ \mathbf{k}'_\alpha&=\mathbf{k}-\mathbf{n}(\mathbf{kn})+\mathbf{n}((\mathbf{kn})\cosh\alpha-E\sin\alpha),\\ \alpha_i&=\alpha n_i,\quad \mathbf{n}^2=1.\end{aligned} \tag{88}$$

The rhs of (86) has a pole at the point z'_R which, of course, coincides with z_R on the E plane because the complex mass does not change. Comparing the pole terms in both sides of (86) we come to conclusion that

$$U(\alpha)\varphi^G_{\text{in}}(\mathbf{k})=\varphi^G_{\text{in}}(\mathbf{k}'_\alpha) \tag{89}$$

with

$$\mathbf{k}'_\alpha=\mathbf{k}-\mathbf{n}(\mathbf{kn})+\mathbf{n}((\mathbf{kn})\cosh\alpha-z_R\sinh\alpha).$$

This equation is meaningless because the transformed space components of the momentum become complex. The difficulty can be resolved by using again the rigged Fock spaces. The action of the Lorentz boost $U(\alpha)$ on the Gamow vectors $\varphi^G_{\text{in}}(\mathbf{k})$ is given by

$$\langle f|U(\alpha)\varphi^G_{\text{in}}(\mathbf{k})\rangle\equiv\langle U^\dagger(\alpha)f|\varphi(E,\mathbf{k})\rangle|_{E=z_R} \tag{90}$$

for any test function $f\in\mathbf{Y}_-$.

The action of the general element of the translation subgroup parametrized by the four-vector a_μ on the Gamov ket $\varphi^G_{\text{in}}(\mathbf{k})$ is defined by

$$\langle f|T(a)\varphi^G_{\text{in}}(\mathbf{k})\rangle\equiv\langle T(a)^\dagger f|\varphi^G_{\text{in}}(\mathbf{k})\rangle \tag{91}$$

with

$$T(a)=\exp(-iP_\mu a^\mu).$$

Now let us combine the translation $T(a)$ with a Lorentz transformation:

$$\begin{aligned}\langle f|T(a)U(\alpha)\varphi^G_{\text{in}}(\mathbf{k})\rangle&\equiv\langle U^+(\alpha)T^+(a)f,\varphi^G_{\text{in}}(\mathbf{k})\rangle\\ &=\langle U^+(\alpha)T^+(a)U^{+-1}(\alpha)U^+(\alpha)f|\varphi^G_{\text{in}}(\mathbf{k})\rangle\\ &=\langle T^+(a'(-\alpha))U^+(\alpha)f|\varphi^G_{\text{in}}(\mathbf{k})\rangle\\ &=e^{i(a'_0(-\alpha)z_R-\overline{a}(-\alpha)\mathbf{k})}\langle U^+(\alpha)f|\varphi^G_{\text{in}}(\mathbf{k})\rangle.\end{aligned} \tag{92}$$

In Eq. (91) we have used the fact that $\varphi^G_{\text{in}}(\mathbf{k})$ is the generalized eigenvector of P_μ with eigenvalue $(z_R,\mathbf{k})$. The $a'_\mu(\alpha)$ in (91) is given by

$$\begin{aligned}a'_0(-\alpha)&=a_0\cosh\alpha+\mathbf{na}\sinh\alpha,\\ \mathbf{a}'(-\alpha)&=\mathbf{a}-\mathbf{n}(\mathbf{an})+\mathbf{n}(\mathbf{an}\cosh\alpha+a_0\sinh\alpha).\end{aligned} \tag{93}$$

The space of test functions will remain invariant only if

$$a_0'(-\alpha) \geqslant 0. \tag{94}$$

This inequality gives us the admissible translations. Indeed, if the vector a_μ is timelike with non-negative zero component, then it has this property in each reference frame. So, in the relativistic case, the admissible translations of $\varphi^G_{\text{in}}(\mathbf{k})$ belong to the future cone V_+. Alternatively permitted translations of $\varphi^G_{\text{out}}(\mathbf{k})$ belong to the past cone V_-. Note that in the nonrelativistic case all translations are separated into future and past-oriented translations.

In the relativistic case there exist translations onto spacelike vectors, which have no analog in the nonrelativistic case and which are admissible for neither $\varphi^G_{\text{in}}(\mathbf{k})$ nor $\varphi^G_{\text{out}}(\mathbf{k})$.

The semidirect products of the Lorentz group with the two translation semigroups define the two Poicaré semigroups $P_\pm$ which describe the spacetime transformations of the two Gamow vectors.

VII. CONCLUDING REMARKS

In this paper, we have presented an extension of the conventional Fock space to a rigged Fock space. The extended evolution is no longer the unitary time symmetric group, but instead splits into two distinct semigroups, one in the forward and the other in the backward time direction. The time reversal transformation maps one semigroup onto the other. As a result, the PCT transformation acquires a new meaning and has to be extended.

The form of the Lorentz transformation of the four momentum of unstable particles has been discussed by many authors. For example, Kawai and Goto,[29] Comi *et al.*,[30] and Beltrametti and Luzzatto[31] discussed nonunitary irreducible representations of the Poincaré group. Raczka[32] discussed nonunitary indecomposable representations and Ben Yaacov[33] considered the Lorentz transformation in Liouville space in the context of relativistic subdynamics.

In the nonrelativistic case, the construction of Gamow vectors and the study of their time evolution has been formulated in terms of rigged Hilbert spaces of Hardy functions.[11] In this paper, we have shown that the rigged Hilbert spaces of Hardy functions also give meaning to the relativistic Gamow vectors. We have to replace the usual rigged Hilbert space by the rigged Fock space[27] $\mathbf{Y} \subset \mathscr{F} \subset \mathbf{Y}^\times$, where $\mathscr{F}$ is the Fock space based on the Hilbert space $\mathscr{H} = L^2[0,\infty) \otimes L^2(R^3)$.

In a recent paper,[34] Pelc and Horwitz considered a different type of rigged Fock space. Although inspired by similar principles, there are some differences between their and our constructions. First of all, they introduce statistics, so that N-particle test spaces include symmetric or antisymmetric vectors only. Second, for the basis representation space, our Ψ, they use $\mathscr{D}(\Omega)$, the space of C^∞ functions with compact support on Ω, endowed with an inductive limit topology. While our Ψ is a Frechét space, their $\mathscr{D}(\Omega)$ is not. Also, the configuration space $\dot{\Omega}$ is quite different from ours. In Ref. 34, Ω is written as a product of three components: the spectra of the generators for space–time translations and rotations of the Poincaré group, the points on a mass shell, and an index to specify individual mass shells.

The purpose of the paper by Pelc and Horwitz is also different from ours, because they introduce a mathematical formalism in order to (i) extend the Dirac formulation to the relativistic case, (ii) discuss some representations of the Poincaré group on a rigged Fock space, and (iii) serve as a basis to extend a mathematical theorem in a separate article.[35] Although all this is made compatible from the very beginning with the S-matrix formalism of scattering, no resonance phenomena are addressed in their paper.

The representation theory of noncompact groups in rigged Hilbert space has been discussed by some authors, for example, Maurin,[36] Antoine,[37] and Lindblad and Nagel.[38] The importance of the representations of noncompact groups in rigged Hilbert space for the introduction of irreversibility has been discussed also in previous works.[19,20,21]

In the past it has often been questioned whether irreversibility is compatible with relativity.[39] If a thermodynamic time asymmetry evolution exists for some observer, it should be present for all inertial observers who should observe the same arrow of time. Therefore, special relativity cannot destroy the universality of the second law of thermodynamics. Our paper provides a simple example where the compatibility of irreversibility and relativity is made explicit.

ACKNOWLEDGMENTS

The need for the extension of Gamow vectors to relativistic systems was suggested to us by Professor A. Bohm. We also thank Dr. G. Ordóñez for interesting comments. We received financial support from the International Solvay Institutes, the Belgian Interuniversity Attraction Poles, the European Commission ESPRIT project CTIAC 21042/DG III and the Grant NPSS No. 0992/DG XII, the NATO CRG940077, the Spanish DGICYT PB 95-0719, and the Junta de Castilla y León CP 001/96.

APPENDIX A: COMMUTATION RELATIONS OF THE CREATION AND ANNHILATION OPERATORS b_{in}, b_{out}, $b_{\text{in}}^\dagger$, AND $b_{\text{out}}^\dagger$

In order to obtain the commutation relation (53) for different types of new creation and annihilation operators we shall make use of the commutation relations (3) and (16). Consider, for example, the first commutator in (53). Using the explicit expression (52) for $b_{\text{in}}(E,\mathbf{k})$ and $b_{\text{in}}^\dagger(E,\mathbf{k})$, we obtain

$$\begin{aligned}[b_{\text{in}}(E,\mathbf{k}),\, b_{\text{in}}^\dagger(E',\mathbf{k}')] &= (2\pi)^4\kappa(E,\mathbf{k})\delta^4(k_\mu - k'_\mu) + (2\pi\lambda)^2\alpha(\kappa(E,\mathbf{k}))\alpha(\kappa(E',\mathbf{k}')) \\ &\times G(E-i0,\mathbf{k})G(E'+i0,\mathbf{k}')(2\pi)^3\delta(\mathbf{k}-\mathbf{k}') \\ &\times\Bigg[\int dE''\ \rho(E'',\mathbf{k})\left(\frac{1}{E''-E+i0}\,\frac{1}{E''-E'-i0} - \frac{1}{E''+E}\,\frac{1}{E''+E'}\right) \\ &+ \frac{1}{E-E'-i0}\left(\frac{1}{G(E-i0,\mathbf{k})} - \frac{1}{G(E'+i0,\mathbf{k})}\right) + (E+E')\Bigg]. \end{aligned} \tag{A1}$$

The expression within the square brackets may be written as follows:

$$\begin{aligned}&\int dE''\ \rho(E'',\mathbf{k})\,\frac{2E''}{E-E'}\left(\frac{1}{E''^2-E^2+i0} - \frac{1}{E''^2-E'^2-i0}\right) + \frac{1}{E-E'}\left(\frac{1}{G(E-i0,\mathbf{k})}\right. \\ &\left. - \frac{1}{G(E'+i0,\mathbf{k})}\right) + \frac{1}{E-E'}\,(E^2-E'^2) \\ &= \frac{1}{E-E'}\left\{\left(\frac{1}{G(E-i0,\mathbf{k})} - \frac{1}{G(E'+i0,\mathbf{k})}\right)\right. \\ &\left. - \left(\int dE''^2\,\frac{\rho(E'',\mathbf{k})}{E''^2-E^2+i0} - \int dE''^2\,\frac{\rho(E'',\mathbf{k})}{E''^2-E^2-i0}\right) + E^2-E'^2\right\}. \end{aligned} \tag{A2}$$

Now, using the definition of Green's function given in (33) and (34), we can show that (A2) vanishes and, therefore, one obtains that

$$[b_{\text{in}}(E,\mathbf{k}),\, b_{\text{in}}^\dagger(E,\mathbf{k})] = (2\pi)^4\kappa(E,\mathbf{k})\delta^4(k_\mu - k'_\mu). \tag{A3}$$

The rest of the commutation relations (53) are obtained in a similar way.

APPENDIX B: THE NEW VACUUM STATE

The new vacuum state Ω is defined as the solution to the equation

$$b_{\text{in}}(E,\mathbf{k})\Omega = 0. \tag{B1}$$

The same equation holds true also for $b_{\text{out}}(E,\mathbf{k})$, and this may be checked explicitly. However, in order to fix ideas, we shall consider (B1).

We will be looking for Ω in the following form:

$$\Omega = e^V|0\rangle, \tag{B2}$$

where $|0\rangle$ is the vacuum state in the initial Fock space and V is a quadratic function of creation operators $B^\dagger(E,\mathbf{k})$ and $a^\dagger(\mathbf{k})$:

$$V=\int d^3\mathbf{k}\left\{\frac{1}{2}\int dE\ dE'\ f(E,E',\mathbf{k})B^\dagger(E,\mathbf{k})B^\dagger(E',-\mathbf{k})\right.$$
$$\left.+\int dE\ f(E,k)B^\dagger(E,\mathbf{k})a^\dagger(-\mathbf{k})+\frac{1}{2}f(\mathbf{k})a^\dagger(\mathbf{k})a^\dagger(-\mathbf{k})\right\}. \tag{B3}$$

While parametrizing V in (B3) we already took into account the invariance of Ω with respect to three-dimensional translations and rotations, generated by (17) and (18). The function $f(E,E',\mathbf{k})$ is symmetric with respect to E and E'

The equations (B1) and (B2) give us

$$b_{\text{in}}(E,\mathbf{k})e^V|0\rangle=e^V([b_{\text{in}}(E,\mathbf{k}),\ V]+b_{\text{in}}(E,\mathbf{k}))|0\rangle=0, \tag{B4}$$

because

$$[V,[b_{\text{in}}(E,\mathbf{k}),\ V]]=0. \tag{B5}$$

In order to show (B4) and (B5), let us expand e^V into formal series:

$$e^V=\sum_{n=0}^{\infty}\frac{1}{n!}V^n. \tag{B6}$$

Then, we apply the annihilation operator $b_{\text{in}}(E,\mathbf{k})$ to (B6) on the left and assume that it commutes with the infinite sum on the right-hand side of (B6). Then, each term of the formal series has a product of the following form:

$$b_{\text{in}}V^n=Vb_{\text{in}}V^{n-1}+[b_{\text{in}},\ V]V^{n-1}. \tag{B7}$$

The commutator $[b_{\text{in}},\ V]$ contains only the *creation operators* $B^\dagger(E,\mathbf{k})$ and $a^\dagger(\mathbf{k})$ because of the structure of b_{in} and V. Therefore, this commutator commutes with V, which is a functional of creation operators only:

$$[[b_{\text{in}},\ V],V]=0. \tag{B8}$$

This is the equation (B5). After (B7), we can rewrite (B6) in the following form:

$$b_{\text{in}}V^n=Vb_{\text{in}}V^{n-1}+V^{n-1}[b_{\text{in}},\ V]. \tag{B9}$$

If we use (B9) in the first term of the right-hand side of (B9), we obtain the following equation:

$$b_{\text{in}}V^n=V^2b_{\text{in}}V^{n-2}+2V^{n-1}[b_{\text{in}},\ V]. \tag{B10}$$

After n steps, we arrive at

$$b_{\text{in}}V^n=V^nb_{\text{in}}+nV^n[b_{\text{in}},\ V]. \tag{B11}$$

If we use (B11) in (B6), we obtain

$$b_{\text{in}}e^V=\sum_{n=0}^{\infty}\frac{1}{n!}b_{\text{in}}V^n=\sum_{n=0}^{\infty}\frac{1}{n!}(V^nb_{\text{in}}+nV^{n-1}[b_{\text{in}},\ V])=e^V(b_{\text{in}}+[b_{\text{in}},\ V]). \tag{B12}$$

Equation (B12) implies (B4) and (B5).

Equation (B4) together with the explicit form of $b_{\text{in}}(E,\mathbf{k})$ [see Eq. (93)] gives us the following equations for functions $f(E,E',\mathbf{k})$, $f(E,\mathbf{k})$, and $f(\mathbf{k})$:

$$f(E,E',\mathbf{k})+G(E^2-\mathbf{k}^2-i0)2\pi\lambda^2\,\frac{\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\int_{E_0}^{\infty}dE''\,\frac{\alpha(\kappa(E',\mathbf{k}))}{E''-E+i0}\,f(E'',E',\mathbf{k})$$
$$=\frac{G(E^2-\mathbf{k}^2-i0)}{(2\pi)^3}\,\lambda\,\frac{\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\left[\lambda\,\frac{\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\,\frac{1}{E'+E}+(E+\omega(\mathbf{k}))(2\pi)^3 f(E',\mathbf{k})\right], \tag{B13}$$

$$f(E,\mathbf{k})+G(E^2-\mathbf{k}^2-i0)2\pi\lambda^2\,\frac{\kappa(\alpha(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\int_{E_0}^{\infty}dE'\,\frac{\alpha(\kappa(E',\mathbf{k}))}{E'-E+i0}\,f(E',\mathbf{k})$$
$$=\frac{G(E^2-\mathbf{k}^2-i0)}{(2\pi)^3}\,\lambda\,\frac{\alpha(\kappa(E',\mathbf{k}))}{\kappa(E',\mathbf{k})}\left[\frac{E+\omega(\mathbf{k})}{2\omega(\mathbf{k})}\,(1+(2\pi)^3 2\omega(\mathbf{k})f(\mathbf{k}))-1\right]. \tag{B14}$$

In order to simplify these equations, let us introduce instead of the f's new functions ϕ as follows:

$$f(E,E',\mathbf{k})=\frac{q(E,\mathbf{k})q(E',\mathbf{k})}{2(2\pi)^3}\,\phi(E,E',\mathbf{k}), \tag{B15}$$

$$f(E,\mathbf{k})=\frac{q(E,\mathbf{k})}{2(2\pi)^3}\,\phi(E,\mathbf{k}), \tag{B16}$$

$$f(\mathbf{k})=\frac{1}{2(2\pi)^3}\,\phi(\mathbf{k}), \tag{B17}$$

where

$$q(E,\mathbf{k})=\lambda\,\frac{\alpha(\kappa(E,\mathbf{k}))}{\kappa(E,\mathbf{k})}\,\gamma(E,\mathbf{k}), \tag{B18}$$

and $\gamma(E,\mathbf{k})$ is the solution of the factorization problem (see Appendix C).

Substituting (B15)–(B17) into (B13) and (B14) and taking into account the properties of $\gamma(E,\mathbf{k})$, we obtain a system of singular integral equations for ϕ:

$$\phi(E,E',\mathbf{k})+\gamma(-E+i0,\mathbf{k})\left[\int_{E_0}^{\infty}dE''\,\frac{\rho(E'',\mathbf{k})\gamma(E'',\mathbf{k})}{E''-E+i0}\,\phi(E'',E',\mathbf{k})-\frac{2}{E+E'}\,\frac{1}{\gamma(E',\mathbf{k})}\right.$$
$$\left.-(E+\omega)\phi(E',\mathbf{k})\right]=0, \tag{B19}$$

$$\phi(E,\mathbf{k})+\gamma(-E+i0,\mathbf{k})\left[\int_{E_0}^{\infty}dE'\,\frac{\rho(E',\mathbf{k})\gamma(E',\mathbf{k})}{E'-E+i0}\,\phi(E',\mathbf{k})-\frac{E-\omega(\mathbf{k})}{\omega(\mathbf{k})}-(E+\omega(\mathbf{k}))\phi(\mathbf{k})\right]=0, \tag{B20}$$

where $\rho(E,\mathbf{k})$ was introduced in (31).

In contradiction with (B13) and (B14), the equations (B20) and (B13) may be considered not only for real values of energy but also on the complex E plane. From (B20) we see that the function $\phi(E,\mathbf{k})$ is analytic in the left half-plane. Let us denote the jump of $\phi(E,k)$ on the cut (E_0,∞) in the right semiplane via $\Delta\phi(E,k)$:

$$\Delta\phi(E,\mathbf{k})=\phi(E+i0,\mathbf{k})-\phi(E-i0,\mathbf{k}). \tag{B21}$$

Further, let us rewrite (B20) for complex E in the following form:

$$\frac{\phi(E,\mathbf{k})}{\gamma(E,\mathbf{k})}=-\int_{E_0}^{\infty}dE'\,\frac{\rho(E',\mathbf{k})\gamma(E',\mathbf{k})}{E'-E}\,\phi(E',\mathbf{k})+\frac{E-\omega(\mathbf{k})}{\omega(\mathbf{k})}+(E+\omega(\mathbf{k}))\phi(\mathbf{k}). \tag{B22}$$

The calculation of the jump on the cut in both sides of (B22) gives the equation,

$$-\Delta\phi(E,\mathbf{k})\left[\frac{1}{\gamma(-E-i0,\mathbf{k})}+2\pi i\gamma(E,\mathbf{k})\rho(E,\mathbf{k})\right]-2\pi i\rho(E,\mathbf{k})\gamma(E,\mathbf{k})\phi(E,\mathbf{k})$$
$$=-2\pi i\rho(E,\mathbf{k})\gamma(E,\mathbf{k})\phi(E,\mathbf{k}). \tag{B23}$$

From (B23) we see that

$$\Delta\phi(E,\mathbf{k})=0, \tag{B24}$$

i.e., the function $\phi(E,\mathbf{k})$ has discontinuity neither in the left, nor in the right E half-plane, so $\phi(E,\mathbf{k})$ is a meromorphic function of E. We shall now show that $\phi(E,\mathbf{k})$ may have poles at the infinity only. Let us rewrite Eq. (B20):

$$\phi(E,\mathbf{k})=-\gamma(-E,\mathbf{k})\left[\int_{E_0}^{\infty}dE'\,\frac{\rho(E',\mathbf{k})\gamma(E',\mathbf{k})}{E'-E}\,\phi(E',\mathbf{k})-\frac{E+\omega(\mathbf{k})}{\omega(\mathbf{k})}-(E+\omega(\mathbf{k}))\phi(\mathbf{k})\right]. \tag{B25}$$

The singularities of $\phi(E,\mathbf{k})$ may come either from $\gamma(E,\mathbf{k})$, or from the integral. Both factors have the branching points at $E=E_0$ and different singularities on the different sheets, but because $\phi(E,\mathbf{k})$ has no branching point, it is sufficient to consider it only on the first sheet, where both $\gamma(E,\mathbf{k})$ and the integral in (B25) may have poles in the infinity only. Therefore, the function $\phi(E,\mathbf{k})$ may have poles only at the infinity, i.e., $\phi(E,\mathbf{k})$ should be a polynomial of E. Now let us make use of the asymptotic expansion of the rhs of (B25):

$$\phi(E,\mathbf{k})=-\left(\frac{1}{E}+\frac{\omega(\mathbf{k})+2\xi(\mathbf{k})}{E^2}+O\left(\frac{1}{E^3}\right)\right)\left[-E\left(\frac{1}{\omega(\mathbf{k})}+\phi(\mathbf{k})\right)+(1-\omega(\mathbf{k})\phi(\mathbf{k}))+O\left(\frac{1}{E}\right)\right]$$
$$=\left(\frac{1}{\omega(\mathbf{k})}+\phi(\mathbf{k})\right)+\frac{1}{E}\left[(\omega(\mathbf{k})+2\xi(\mathbf{k}))\left(\frac{1}{\omega(\mathbf{k})}+\phi(\mathbf{k})\right)-(1-\omega\phi(\mathbf{k}))\right]+O\left(\frac{1}{E^2}\right), \tag{B26}$$

where we have used (C4). It is obvious that our conclusion that $\phi(E,k)$ should be a polynomial of E is compatible with (B19) iff only the first term does not vanish, i.e.,

$$\phi(E,\mathbf{k})=\frac{1}{\omega(\mathbf{k})}+\phi(\mathbf{k}),$$

$$(\omega(\mathbf{k})+2\xi(\mathbf{k}))\left(\frac{1}{\omega(\mathbf{k})}+\phi(\mathbf{k})\right)-(1-\omega(\mathbf{k})\phi(\mathbf{k}))=0, \tag{B27}$$

and all other terms also vanish. From (B27) we obtain

$$\phi(\mathbf{k})=-\frac{\xi(\mathbf{k})}{\omega(\mathbf{k})(\omega(\mathbf{k})+\xi(\mathbf{k}))},$$

$$\phi(E,\mathbf{k})=\frac{1}{\omega(\mathbf{k})+\xi(\mathbf{k})}. \tag{B28}$$

The substitution of (B28) into (B25) cancels all terms of the asymptotic expansion (B26); therefore (B28) is the solution of (B20).

Now, using (B21) we can rewrite Eq. (B19):

$$\phi(E,E',\mathbf{k})+\gamma(-E+i0,\mathbf{k})\left[\int_{E_0}^{\infty}dE''\,\frac{\rho(E'',\mathbf{k})\,\gamma(E'',\mathbf{k})}{E''-E+i0}\,\phi(E'',E',\mathbf{k})\right.$$
$$\left.-\frac{2}{E+E'}\,\frac{1}{\gamma(E',\mathbf{k})}-\frac{E+\omega(\mathbf{k})}{\xi(\mathbf{k})+\omega(\mathbf{k})}\right]=0. \tag{B29}$$

The same arguments as above may be used to prove the absence of discontinuities of $\phi(E,E',\mathbf{k})$ in the E plane. As $\phi(E,E',\mathbf{k})$ is symmetric with respect to E and E', the same it true for E'. Therefore, we conclude the only singularities of $\phi(E,E',k)$ are the poles at $E=-E_0$ and at infinity. Thus, $\phi(E,E',\mathbf{k})$ has the form

$$\phi(E,E',\mathbf{k})=\frac{C(\mathbf{k})}{E+E'}+P(E,E',\mathbf{k}), \tag{B30}$$

where $P(E,E',\mathbf{k})$ is a polynomial of E and E'. The asymptotic expansion of both sides of (B30) gives

$$C(\mathbf{k})=2,\quad P(E,E',\mathbf{k})=\frac{1}{\omega(\mathbf{k})+\xi(\mathbf{k})}, \tag{B31}$$

so

$$\phi(E,E',\mathbf{k})=\frac{1}{\omega(\mathbf{k})+\xi(\mathbf{k})}+\frac{2}{E+E'}. \tag{B32}$$

We may verify that (B32) is the solution of (B29), if we substitute (B32) into (B29) using the integral representation for $\gamma(E,\mathbf{k})$ given in Appendix C [see formula (C9)].

Formula (B32) completes the solution of (B19) and (B20) and gives the expression for functions the $f(E,E',\mathbf{k})$, $f(E,\mathbf{k})$, and $f(\mathbf{k})$ via (B15)–(B17).

APPENDIX C: FACTORIZATION OF GREEN'S FUNCTION

The problem of factorization of Green's functions or the S matrix often arises in different parts of theoretical and mathematical physics. The solutions of the problem may be very different in each case depending on the physical context. In our case the factorization problem is formulated as follows: given Green's function $G(E,\mathbf{k})$ with analytic properties discussed in Sec. III, i.e., on the first sheet the only singularities are the branching points at $E=\pm E_0$ and the asymptotic behavior is given by (40), find a function $\gamma(E,\mathbf{k})$ analytic in the right half-plane such that

$$\gamma(E,\mathbf{k})\,\gamma(-E,\mathbf{k})=G(E,\mathbf{k}) \tag{C1}$$

is valid. Obviously, this condition does not fix the $\gamma(E,\mathbf{k})$, and in order to remove all ambiguities we must require that $\gamma(E,\mathbf{k})$ has neither poles nor essential singularities on the first sheet of E plane and that the only singularity in the whole complex plane is the cut between $E=-E_0$ and infinity in the left half-plane Re $E<0$. Further, because neither $G(E,\mathbf{k})$ nor $G(E,\mathbf{k})^{-1}$ has poles in the finite domain on the first sheet, one could easily prove that $\gamma^{-1}(E,\mathbf{k})$ also has no poles, so $\gamma(E,\mathbf{k})$ has neither poles nor zeros in the finite domain of the first sheet of E plane.

In the absence of interaction ($\lambda=0$), the function $G(E,\mathbf{k})$ becomes Green's function of a Klein–Gordon field:

$$G(E,\mathbf{k})=(\omega^2(\mathbf{k})-E^2)^{-1}. \tag{C2}$$

In this case, the solution of the factorization problem is

$$\gamma(E,\mathbf{k})=-(\omega(\mathbf{k})+E)^{-1}, \tag{C3}$$

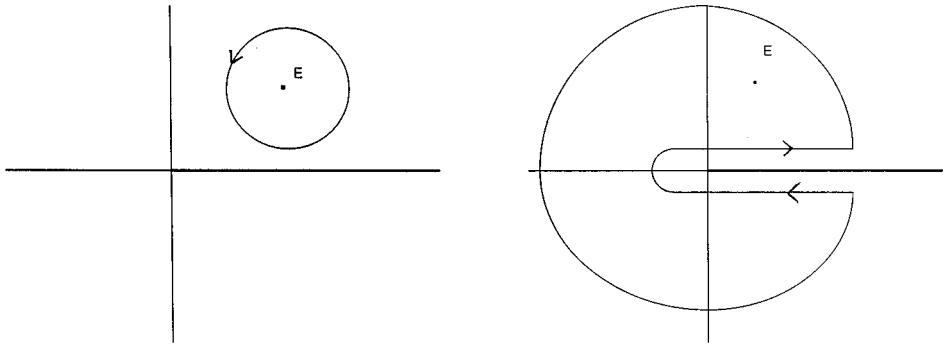

FIG. 1. The contour of integration in Eqs. (C6) and (C7).

where we fixed the sign on the r.h.s. The interaction does not change the asymptotic behavior of Green's function and therefore we can impose the asymptotic condition on $\gamma(E,\mathbf{k})$ in the following form:

$$\gamma^{-1}(E,\mathbf{k}) = -E - (\omega(\mathbf{k}) + 2\xi(\mathbf{k})) + O(1/E), \quad E \to \infty. \tag{C4}$$

The energy shift in the second term $\xi(\mathbf{k})$ appears due to interaction.

Now we can consider the Cauchy integral for function $r(E,\mathbf{k})$, which is defined by

$$r(E,\mathbf{k}) = \gamma^{-1}(-E,\mathbf{k}) - E + \omega(\mathbf{k}) + 2\xi(\mathbf{k}). \tag{C5}$$

This function vanishes at infinity, may have poles in only at infinity, and has a cut in the right half-plane on E from $E = E_0$ to infinity. Thus, in the domain of analyticity of $r(E,\mathbf{k})$, the Cauchy formula holds:

$$r(E,\mathbf{k}) = \frac{1}{2\pi i} \int_C dE' \, \frac{r(E',\mathbf{k})}{E' - E}, \tag{C6}$$

where the contour C lays on the first sheet around point E. Deforming the contour C to the infinite circle excluding the cut on the real axes (Fig. 1), we get

$$r(E,\mathbf{k}) = \frac{1}{2\pi i} \int_{E_0}^{\infty} dE' \, \frac{1}{E' - E} \, (r(E' + i0,\mathbf{k}) - r(E' - i0,\mathbf{k})), \tag{C7}$$

where the integral along infinite circle disappeared due to the asymptotic behavior of $r(E,\mathbf{k})$. The discontinuity of the integrand in (C7) is easily calculated:

$$\begin{aligned} r(E + i0,\mathbf{k}) - r(E - i0,\mathbf{k}) &= \gamma^{-1}(-E - i0,\mathbf{k}) - \gamma^{-1}(-E + i0,\mathbf{k}) \\ &= \gamma(E,\mathbf{k})(G^{-1}(E + i0,\mathbf{k}) - G^{-1}(E - i0,\mathbf{k})) = -2\pi i \gamma(E,\mathbf{k})\rho(E,\mathbf{k}), \end{aligned} \tag{C8}$$

where in the second step we used the fact that $\gamma(E,\mathbf{k})$ has no discontinuity in the right half-plane and in the last step we used Eq. (38).

Taking into account Eq. (C8) we finally obtain the equation

$$\gamma^{-1}(-E,\mathbf{k}) = E - (\omega(\mathbf{k}) + 2\xi(\mathbf{k})) - \int_{E_0}^{\infty} dE' \, \frac{\rho(E',\mathbf{k})\,\gamma(E',\mathbf{k})}{E' - E}, \tag{C9}$$

which is valid for complex E on the first sheet. This equation obviously is the same as Eq. (62).

From (B9) we can obtain the third term of asymptotic expansion of $\gamma^{-1}(E,\mathbf{k})$:

$$\gamma^{-1}(-E,\mathbf{k})=E-(\omega(\mathbf{k})+2\xi(\mathbf{k}))+\frac{1}{E}\int_{E_0}^{\infty}dE'\ \rho(E',\mathbf{k})\gamma(E',\mathbf{k})+o(1/E). \tag{C10}$$

Using (A1) and (38) we obtain the relation (63):

$$\int_{E_0}^{\infty}dE\ \rho(E,\mathbf{k})\gamma(E,\mathbf{k})=2\xi(\mathbf{k})(\omega(\mathbf{k})+\xi(\mathbf{k})). \tag{C11}$$

Now if we will consider instead of the function $r(E,\mathbf{k})$ the $\gamma(-E,\mathbf{k})$ itself, we will obtain the relation

$$\gamma(-E,\mathbf{k})=\int_{E_0}^{\infty}dE'\ \frac{\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2}{\gamma(E',\mathbf{k})}\frac{1}{E'-E}, \tag{C12}$$

which is also valid for complex E on the first sheet. Here, for the calculation of the discontinuity, we used Eq. (39). Combining both (C9) and (C12), which are nonlinear integral equations, we can obtain a linear integral equation of Fredholm type for $\gamma(E,\mathbf{k})$:

$$\gamma(E,\mathbf{k})-\int_{E_0}^{\infty}dE'\ \rho(E',\mathbf{k})\gamma(E',\mathbf{k})v(E,E')=U(E), \tag{C13}$$

where

$$v(E,E')=\frac{R(E,\mathbf{k})-R(E'\mathbf{k})}{E-E'}, \tag{C14}$$

$$U(E)=1+(-E+\omega(\mathbf{k})+2\xi(\mathbf{k}))R(E,\mathbf{k}), \tag{C15}$$

$$R(E,\mathbf{k})=\int_{E_0}^{\infty}dE'\ \frac{\rho(E',\mathbf{k})|G(E',\mathbf{k})|^2}{E+E'}. \tag{C16}$$

The solution of (C13) is the standard Fredholm solution which can be obtained from the expression of the form factor $\rho(E,\mathbf{k})$.

The factorization problem described here is similar although not the same as the factorization problem considered by Krein.[40] The latter is formulated as follows:

Given a function $F(t)$ defined on the real axis R, find two functions of the complex variable z, $f_+(z)$ and $f_-(z)$, analytic on the upper and lower half-plane, respectively, such that

$$f_+(t)f_-(t)|_{t\in R}=F(t) \tag{C17}$$

satisfying the following normalization condition:

$$f_{\pm}(\infty)=1. \tag{C18}$$

This problem is obviously related to the Riemann–Hilbert problem for $f_+(z)$ and $f_-^{-1}(z)$.[41,42]

In our case, we need to construct a solution of (C1) such that $f_+(z)$ [in our notation $\gamma(E)$] coincides with $f_-(-z)$. In addition, our function $F(t)[=G(E)]$ is known for complex E and admits the representation (C1) in the whole complex plane. A *direct* relation between the problem described in (C1) and the Riemann–Hilbert factorization problem is unknown to us. Also, we did not find in the literature any discussion concerning this point.

[1] G. Gamow, Z. Phys. **51**, 204 (1928).
[2] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1955).
[3] E. Prugovecki, *Quantum Mechanics in Hilbert Space* (Academic, New York, 1981).
[4] N. Nakanishi, Prog. Theor. Phys. **19**, 607 (1958).
[5] R. E. Peierls, Proc. R. Soc. London, Ser. A **253**, 16 (1959).
[6] V. I. Kukulin, V. M. Krasnapolski, and J. Horacek, *Theory of Resonances* (Academia, Praha, 1989).
[7] L. Horwitz and P. Marchand, Rocky Mt. J. Math. **1**, 225 (1971).

[8] H. M. Nussenzveig, *Causality and Dispersion Relations* (Academic, New York, 1972).
[9] R. G. Newton, *Scattering Theory of Waves and Particles* (Springer Verlag, New York, 1982).
[10] A. Bohm, *Quantum Mechanics: Fundations and Applications* (Springer Verlag, New York, 1994).
[11] A. Bohm and M. Gadella, *Dirac Kets, Gamow Vectors and Gel'fand Triplets*, Springer Lecture Notes in Physics, No. 348 (Springer Verlag, New York, 1989).
[12] A. Bohm, J. Math. Phys. **22**, 2813 (1981).
[13] I. Gel'fand and N. Vilenkin, *Generalized Functions* (Academic, New York, 1964), Vol. 4; H. H. Schaeffer, *Topological Vector Spaces* (Springer Verlag, New York, 1970).
[14] P. Koosis, *Introduction to H_p Spaces*, London Math. Soc. Lecture Notes Series No. 40 (Cambridge U.P., Cambridge, 1980); P. Koosis, *The Logaritmic Integral*, (Cambridge U.P., Cambridge, 1990); P. Duren, *Theory of H_p Spaces* (Academic, New York, 1970).
[15] I. Prigogine, *From Being to Becoming* (Freeman, New York, 1980).
[16] A. Bohm, J. Math. Phys. **22**, 2813 (1981).
[17] A. Grecos and I. Prigogine, Physica **59**, 77 (1972); Proc. Natl. Acad. Sci. USA **60**, 1629 (1972).
[18] T. Petroski, I. Prigogine, and S. Tasaki, Physica A **175**, 175 (1991).
[19] I. Antoniou and I. Prigogine, Physica A **192**, 443 (1993).
[20] I. Antoniou and S. Tasaki, Int. J. Quantum Chem. **44**, 425 (1993).
[21] I. Antoniou, L. Dmitrieva, Y. Kuperin, and Y. Melnikov, ''Resonances and the Extension of Dynamics in Rigged Hilbert Spaces,'' Comput. Math. Appl. **34**, 399–425 (1997).
[22] T. Petrovski and I. Prigogine, Adv. Chem. Phys. **99**, 1 (1997).
[23] I. Antoniou and Z. Suchanecki, Adv. Chem. Phys. **99**, 299 (1997).
[24] H. Yukawa, Phys. Rev. **77**, 219 (1950); **80**, 1047 (1950).
[25] M. A. Markov, J. Phys. USSR **2**, 453 (1940).
[26] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).
[27] M. Gadella, J. Math. Phys. **26**, 725 (1985).
[28] H. H. Schaeffer, *Topological Vector Spaces* (Springer Verlag, New York, 1970).
[29] T. Kawai and M. Goto, Nuovo Cimento **60**, 21 (1969).
[30] M. Comi, L. Lanz, L. Lugiato, and G. Ramella, J. Math. Phys. **16**, 910 (1975).
[31] E. Beltrametti and G. Luzatto, Nuovo Cimento **36**, 1217 (1965).
[32] R. Raczka, Ann. Inst. Henri Poincaré, Sect. A **19**, 341 (1973).
[33] U. Ben Yaacov, Physica A **222**, 307 (1995).
[34] O. Pelc and L. P. Horwitz, J. Math. Phys. **38**, 115 (1997).
[35] O. Pelc and L. P. Horwitz, J. Math. Phys. **38**, 139 (1997).
[36] K. Maurin, *General Eigenfunction Expansions and Unitary Representation of Topological Groups* (Polish Scientific Publishers, Warsaw, 1968).
[37] J. P. Antoine, J. Math. Phys. **10**, 53 (1969).
[38] G. Lindblad and B. Nagel, Ann. Inst. Henri Poincaré, Sect. A **13**, 27 (1970).
[39] P. Davies, *Einstein's Unfinished Revolution* (Simon and Schuster, New York, 1995).
[40] M. G. Krein, Usp. Mat. Nauk **13**, 1 (1958).
[41] F. D. Gakhov, *Boundary Value Problems* (Pergamon, Oxford, 1966).
[42] A. S. Fokas and M. J. Ablowitz, in *Nonlinear Phenomena*, edited by K. B. Wolf, Springer Lecture Notes in Physics, Vol. 189 (Springer Verlag, New York, 1983).

Quantum martingale measures and stochastic partial differential equations in Fock space

David Applebaum
Department of Mathematics, Statistics and Operational Research,
The Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, England

(Received 21 April 1997; accepted for publication 13 January 1998)

A concept of quantum martingale measure is introduced and examples are constructed as quantum stochastic spectral integrals in Fock space. These are then utilized as space–time noise to drive a parabolic stochastic partial differential equation (spde). We establish the existence and uniqueness of the solutions as families of densely defined closable operators in Fock space that are jointly continuous in time and space variables and satisfy a Markov property. © *1998 American Institute of Physics.* [S0022-2488(98)00206-0]

I. INTRODUCTION

A stochastic partial differential equation (spde) is effectively a partial differential equation that is perturbed by a space–time random field. These arise in a number of different contexts including branching processes, neurophysiology, and constructive quantum field theory, and there is an enormous literature on the subject. Our aim in this paper is to give the first example of an spde, which is perturbed by a quantum noise.

We begin by briefly describing the classical setup that we intend to quantize. We will in this paper be solely concerned with quantum analogs of parabolic spdes of the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(t,x,u) + g(t,x,u)\dot{W}(t,x), \tag{1.1}$$

with, e.g., reflecting boundary conditions on $\mathbb{R}^+ \times [0,a]$ and suitable Lipshitz conditions on the functions f and g. Here W is a martingale measure, i.e., a martingale with respect to time and a measure with respect to space. The solution of (1.1) when it exists will be a random field. The main technique for studying (1.1) is to first write it in a weak form and then treat an equivalent ''auxiliary equation'' wherein the spatial derivatives are absorbed into the Green's function for the homogeneous version of (1.1). For more information on the existence and uniqueness, see the papers in Refs. 1 and 2, which will be our standard references for the subject.

For more recent progress (especially in hyperbolic spdes)—see the recent review,[3] and for applications to quantum fields see Ref. 4 and references therein.

Most authors have studied (1.1), with W being a Gaussian space–time white noise (i.e., a Brownian sheet). Recently there has been some interest in using different types of noise—so W has been a Poisson white noise in Ref. 5 and a Lévy white noise in Ref. 6. A related equation (in three space dimensions) driven by α-stable Lévy noise has been considered in Ref. 7. Here we will take W to be a quantum noise in a suitable Fock space so that the solution of (1.1) when it exists will be a family of linear operators indexed by space and time. One reason for doing this is to provide an example of a new type of irreversible quantum dynamics constructed by means of the powerful tools of quantum stochastic calculus. Another motivation comes from the modeling of electric currents in neurons.

In fact, we shall only be concerned in this paper with the case where $-f(t,x,u) = g(t,x,u) = u$ in (1.1). This equation is usually called the cable equation, and it was proposed in Refs. 8 and 1 as a model for the evolution in space and time of the electric potential on a neuron when the latter is treated as a long thin cylinder of length a. It is argued that this potential evolves through a combination of diffusion and random fluctuations, the latter of which should be modeled by a Poisson noise or (in the limit of infinitely many small jumps) a Gaussian noise (for a lengthy

 © 1998 American Institute of Physics

discussion of this in the context of an infinite-dimensional stochastic model see the introduction to Ref. 9). The reasoning behind the introduction of a quantum noise in this equation is the recent proposal of Penrose that the propagation of nerve signals might be influenced by quantum coherence effects within the microtubules in the cytoskeletons of the neurons (see Chap. 7 of Ref. 10). The quantum version of (1.1) would then describe the microtubular potential as an observable that evolves in space and time through diffusion and quantum effects, as described by the noise terms. Although this proposed application is highly speculative, it is sufficiently intriguing to merit further investigation.

The organization of this paper is as follows. In Sec. II, we introduce the basic framework for quantum space–time noise—namely, the quantum martingale measure. As its name suggests, this is an obvious analog of the classical concept described above, but the martingale concept is now quantum probabilistic (see Ref. 11 and references therein) and the measure is Hilbert space valued. Our main examples of these objects are quantum stochastic spectral integrals with respect to creation, conservation, and annihilation processes in Fock space.

These involve both a quantum stochastic integral in the sense of Ref. 12 (time) and an integration with respect to a projection-valued measure (space) and were first introduced to realize stochastic flows driven by Lévy processes in Fock space (see Refs. 13 and 14). We review those aspects of this theory that we will need in Sec. III.

With quantum stochastic spectral integrals playing the role of the noise W, we are now able to formulate the spde (1.1) in Fock space.

In Secs. IV and V we follow the line of reasoning of Ref. 1 in first proving that the existence and uniqueness problem for the weak form of (1.1) is equivalent to that for the conceptually simpler auxiliary equation constructed with the aid of the Green's function and then establishing the existence and uniqueness for the solution of the auxiliary equation. We then show that the solution $[u(t,x),\ t\in\mathbb{R}^+,\ 0\leqslant x\leqslant a]$ consists of densely defined closable linear operators that are jointly continuous when evaluated on exponential vectors and give rise to adapted processes for each $0\leqslant x\leqslant a$. We also present an appropriate form of the Markov property and show that this is satisfied by the solution.

If the equation (1.1) is to be taken seriously as a model for a quantum potential, the solution should be essentially self-adjoint. It is easy to show (with appropriate restrictions on the noise) that the solution is symmetric, but we will postpone further consideration of this question to a later paper. It is worth pointing out at this stage that even in the relatively simpler case of the Hudson–Parthasarthy theory there are few results known to the author about self-adjointness of quantum stochastic integrals or of solutions to quantum stochastic differential equations, and he hopes that this paper will succeed in stimulating work on this and related problems. Some interesting recent work in this direction appears in Ref. 15.

Finally, we emphasize that we will only deal with quantum spdes involving one spatial dimension in this paper. By analogy with the classical case, we expect that we will not, in general, be able to find operator-valued solutions for higher-dimensional equations but that generalized (distribution-valued) solutions will exist.

II. QUANTUM MARTINGALE MEASURES

Let S be a measurable space equipped with a σ-algebra of subsets $\mathcal{F}$ and let $\mathcal{H}$ be a complex separable Hilbert space. By an *$\mathcal{H}$-valued measure* on $(S,\mathcal{F})$ we will mean a map $Q\colon \mathcal{F}\to\mathcal{H}$, which is such that, for any sequence $(E_n,\ n\in\mathbb{N})$ of mutually disjoint subsets in $\mathcal{F}$, we have

$$\text{(i)}\quad \sum_{n=1}^{\infty} Q(E_n)\in\mathcal{H}$$

and

$$\text{(ii)}\quad Q\Big(\bigcup_{n\in\mathbb{N}} E_n\Big)=\sum_{n=1}^{\infty} Q(E_n).$$

Any projection-valued measure acting on a fixed vector gives rise to an $\mathscr{H}$-valued measure in the above sense.

Now, let $\{\mathscr{H}_t,\ t\in\mathbb{R}^+\}$ be a family of closed linear subspaces of $\mathscr{H}$ for which

$$\text{(i)}\quad \mathscr{H}_s\subseteq\mathscr{H}_t,\quad \text{whenever } 0\leqslant s\leqslant t<\infty;$$

$$\text{(ii)}\quad \mathscr{H}=\{\bigcup_{t\geqslant 0}\mathscr{H}_t\}^-.$$

We call $\{\mathscr{H}_t,\ t\in\mathbb{R}^+\}$ a *filtration* of $\mathscr{H}$ and denote as Π_t the orthogonal projection from $\mathscr{H}$ into $\mathscr{H}_t$, for each $t\in\mathbb{R}^+$.

A family of linear operators $[X(t),\ t\in\mathbb{R}^+]$ in $\mathscr{H}$ is said to be *adapted* to the filtration $\{\mathscr{H}_t,\ t\in\mathbb{R}^+\}$, if for each $t\in\mathbb{R}^+$,

(i) Π_t leaves Dom $X(t)$ invariant;

(ii) $X(t)\Pi_t\psi=\Pi_t X(t)\psi$, for all $\psi\in\text{Dom}(X(t))$.

Let ϕ be a fixed unit vector in $\mathscr{H}$ such that $\Pi_t\phi=\phi$ for each $t\in\mathbb{R}^+$.

Now take $(S,\mathscr{F})$ to be as above, let $\mathscr{U}$ be a dense linear subspace of $\mathscr{H}$ that contains ϕ and for which $\Pi_t\mathscr{U}\subseteq\mathscr{U}$ for all $t\in\mathbb{R}^+$ and let $M=(M(t,E),t\in\mathbb{R}^+,E\in\mathscr{F})$ be a family of densely defined linear operators in $\mathscr{H}$.

We say that M is a *quantum martingale measure* relative to the filtration $\{\mathscr{H}_t,\ t\in\mathbb{R}^+\}$ and the subspace $\mathscr{U}$ if (i) $\mathscr{U}\subseteq\text{Dom}(M(t,E))$ for each $t\in\mathbb{R}^+$, $E\in\mathscr{F}$; (ii) for each $E\in\mathscr{F}$, $M(\cdot,E)$ is adapted to the filtration $\{\mathscr{H}_t,\ t\in\mathbb{R}^+\}$; (iii) for each $E\in\mathscr{F}$, $M(\cdot,E)$ is a martingale, i.e., for each $s\leqslant t$,

$$\Pi_s M(t,E)\Pi_s\psi=M(s,E)\psi,\quad \text{for all}\quad \psi\in\mathscr{U};$$

(iv) for each $t\in\mathbb{R}^+$, $\psi\in\mathscr{U}$, $M(t,\cdot)\psi$ is an $\mathscr{H}$-valued measure.

Example 1 (classical martingale measures): The concept of an L^2-martingale measure that is used in probability theory is essentially a special case of the above, wherein we take $\mathscr{H}=L^2(\Omega,\mathscr{F},P)$ for some probability space equipped with a filtration $(\mathscr{F}_t,\ t\in\mathbb{R}^+)$, each $\mathscr{H}_t$ to be $L^2(\Omega,\mathscr{F}_t,P)$, ϕ to be the constant function 1, $\mathscr{U}=\{f\phi,f\in L^\infty(\Omega,\mathscr{F},P)\}$ and Π_t to be the usual conditional expectation (see, e.g., Refs. 16 and 1).

Example 2 (W-martingale measures):* Let $\mathscr{A}$ be a von Neumann algebra acting in $\mathscr{H}$ and ϕ be a cyclic and separating vector for $\mathscr{A}$. Suppose that $(\mathscr{A}_t,\ t\in\mathbb{R}^+)$ is a filtration of $\mathscr{A}$ and that there exists a family of ω-compatible conditional expectations $(E_t,\ t\in\mathbb{R}^+)$, where each $E_t:\mathscr{A}\to\mathscr{A}_t$ and ω is the vector state determined by ϕ. In this case we may take $\mathscr{H}_t$ to be the completion of $\mathscr{A}_t\phi$, $\mathscr{U}=\mathscr{A}\phi$ and define each Π_t by $\Pi_t(x\phi)=E_t(x)\phi$ for $t\in\mathbb{R}^+$, $x\in\mathscr{A}$. Now, if $M(t,E)$ is a family of operators in $\mathscr{A}$ such that $M(\cdot,E)$ is an $\mathscr{A}$-valued martingale (see, e.g., Ref. 17) for each $E\in\mathscr{F}$, then for each $x\in\mathscr{A}$, $M(t,E)x\phi$ will be a quantum martingale measure whenever $M(t,\cdot)x\phi$ is an $\mathscr{H}$-valued measure.

Many of the concepts of classical martingale measures such as orthogonality and nuclear covariance can be extended to the present context, but we will not need to develop this here. We will require just one further definition.

A quantum martingale measure will be said to be *dominated* by a family of σ-finite measures $\{\mu_\psi,\ \psi\in\mathscr{U}\}$ on $(S,\mathscr{F})$ if there exists a measurable function h on $\mathbb{R}^+$ such that

$$\|M_t(A)\psi\|^2\leqslant h(t)\mu_\psi(A),\tag{2.1}$$

for all $t\in\mathbb{R}^+$, $A\in\mathscr{F}$, $\psi\in\mathscr{U}$.

We will now give some examples of martingale measures in quantum stochastic calculus. To this end, we will, from now on, take $\mathscr{H}$ to be the symmetric Fock space $\Gamma(L^2(\mathbb{R}^+,K))$, where K is a complex separable Hilbert space, which we will identify, when convenient with $\Gamma(L^2(\mathbb{R}^+)\otimes K)$. Fix a dense linear manifold $\mathscr{D}$ in K and let $\mathscr{M}$ denote the dense linear manifold in $L^2(\mathbb{R}^+,K)$ of locally bounded functions, taking values in $\mathscr{D}$. We denote by $\mathscr{E}$ the linear span of the exponential vectors $\{e(f),\ f\in\mathscr{M}\}$ that is dense in $\mathscr{H}$. We will find it convenient to denote the vacuum vector $e(0)$ as e_0. Given $f\in L^2(\mathbb{R}^+,K)$, the map that sends $e(f)$ to $e(f\chi_{[0,t)})$

$\otimes e(f\chi_{[t,\infty)})$ extends to a canonical isomorphism between $\mathscr{H}$ and $\Gamma(L^2([0,t),K))\otimes\Gamma(L^2([t,\infty),K))$ which we use to identify these two spaces. We denote as e_{0t} and e_0^t the vacuum vectors in $\Gamma(L^2([0,t),K))$ and $\Gamma(L^2([t,\infty),K))$, respectively.

Now for each $f\in L^2(\mathbb{R}^+,K)$ let $a(f)$ and $a^\dagger(f)$ denote the annihilation and creation operators (respectively) in $\mathscr{H}$ and for T a bounded self-adjoint operator in $L^2(\mathbb{R}^+,K)$, let $\lambda(T)$ be the conservation operator in $\mathscr{H}$. We note that $a(f)$ and $a^\dagger(f)$ are densely defined and mutually adjoint on $\mathscr{E}$ and that $\lambda(T)$ is essentially self-adjoint therein. Finally, if T is a contraction on $L^2(\mathbb{R}^+,K)$, then $\Gamma(T)$ denotes its second quantization in $\mathscr{H}$, which is also a contraction. In particular, the functorial properties of Γ ensure that it maps projections in $L^2(\mathbb{R}^+,K)$ to projections in $\mathscr{H}$. Precise definitions and further developments of the concepts described above can be found in Ref. 11, pp. 123–52.

We now give some examples of quantum martingale measures in the following context:

$$\mathscr{H}_t=\Gamma(L^2([0,t),K))\otimes\mathbb{C}e_0^t,$$

$$\Pi_t=\Gamma(\chi_{[0,t)}),\quad \mathscr{U}=\mathscr{E},\quad \phi=e_0.$$

In the sequel, we will also find it useful to denote $\mathscr{M}_t=\mathscr{M}\cap L^2([0,t),K)$.

Now, with $(S,\mathscr{F})$ arbitrary, let P be a projection-valued measure defined on $\mathscr{F}$ and taking values in K. For each $t\in\mathbb{R}^+$, $E\in\mathscr{F}$, define

$$\begin{aligned}A(t,E)&=a(\chi_{[0,t)}\otimes P(E)u),\\ \Lambda(t,E)&=\lambda(\chi_{[0,t)}\otimes P(E)),\\ A^\dagger(t,E)&=a^\dagger(\chi_{[0,t)}\otimes P(E)u),\end{aligned}\tag{2.2}$$

where u is a fixed unit vector in $\mathscr{D}$.

Let $c_j\in\mathbb{C}$ for $j=1,2,3$ and define

$$M(t,E)=c_1A(t,E)+c_2\Lambda(t,E)+c_3A^\dagger(t,E).\tag{2.3}$$

Proposition 2.1: M is a quantum martingale measure.

Proof: Adaptedness and the martingale property is a straightforward exercise in Fock space manipulation. We will obtain countable additivity from a more general result, which we will prove in the next section. □

We now write $\psi=f\otimes k$, where f is a locally bounded function on $\mathbb{R}^+$ and $k\in\mathscr{D}$. A straightforward but tedious computation with the extended CCRs shows that M is dominated by the measure μ_ψ with $h(t)=t(1+\sup_{0\leqslant s\leqslant t}|f(s)|^2)$, where

$$\mu_\psi(E)=(|c_1|^2+|c_2|^2+|c_3|^2)(1+\|\psi\|^2)(\|P(E)k\|^2+\|P(E)u\|^2).$$

Examples 1 (Gaussian white noise): Let $c_2=0$ and $c_3=\overline{c_1}$ in (2.1) with $|c_1|=1$; then it is well known that each $M(t,E)$ is Gaussian distributed in the vacuum state, i.e., the characteristic function (where $x\in\mathbb{R}$) is

$$\langle e_0,\exp(ixM(t,E))e_0\rangle=\exp\{-\tfrac{1}{2}tx^2\|P(E)u\|^2\}.$$

If we take S to be $[0,a]$ equipped with its Borel σ algebra $\mathscr{B}([0,a])$, $K=L^2([0,a])$, each $P(E)=\chi_E$ and $u=a^{-1/2}$ for $a>0$ then the process $M_{t,u}=M(t,[0,u))$ is a Brownian sheet on $\mathbb{R}^+\times[0,a]$.

Example 2 (Poisson random measures): Let $c_j=1$ $(j=1,2,3)$; then $M(t,E)$ is a realization in $\mathscr{H}$ of a compensated Poisson random measure on $\mathbb{R}^+\times\mathscr{F}$ with intensity measure ν given by $\nu(E)=\|P(E)u\|^2$.

In particular, if we define $Y(t,E)=M(t,E)+t\nu(E)$, then for all $x\in\mathbb{R}$,

$$\langle e(0),\exp(ixY(t,E))e(0)\rangle=\exp\{t\nu(E)(e^{ix}-1)\}.$$

If we take S to be of the form $[0,a]\times\mathscr{S}$, $\mathscr{T}=\mathscr{B}([0,a])\times\mathscr{G}$, $K=L^2([0,a],\mathfrak{h})$ where $\mathfrak{h}$ is a complex separable Hilbert space and $P(E\times F)=\chi_E\otimes Q(F)$, where Q is a projection-valued measure on $\mathscr{G}$ taking values in $\mathfrak{h}$; then for $0\leqslant u\leqslant a$, $F\in\mathscr{G}$ if we define

$$Y_{t,u}(F)=Y(t,[0,u)xF),$$

then we have a space–time Poisson random measure on $\mathbb{R}^+\times[0,a]\times\mathscr{T}$ for which M is the compensating martingale measure (Ref. 6).

III. QUANTUM SPECTRAL STOCHASTIC INTEGRALS (Refs. 13,14,18,19)

In this section, we will assume familiarity with the theory of quantum stochastic integration developed in Ref. 12 (see also Ref. 11), and we begin by briefly reviewing how this theory was extended in Ref. 13 to include an additional integration with respect to a projection-valued measure. The aim is to make sense of integrals of the form

$$\int_a^b\int_0^t (F_1(s,x)A^\dagger(ds,dx)+F_2(s,x)\Lambda(ds,dx)+F_3(s,x)A(ds,dx)), \tag{3.1}$$

as linear operators on the domain $\mathscr{E}$, where $A^\dagger$, Λ, and A are the processes (2.2) and F_j $(j=1,2,3)$ are *admissible families* in that they satisfy the following.

A(i) For each $x\in[a,b]$, $(F_j(t,x),t\in\mathbb{R}^+)$ is adapted and stochastically integrable in the sense of Ref. 11 (p. 189).

A(ii) For each $t\in\mathbb{R}^+$, $\psi\in\mathscr{M}$, and given $\epsilon>0$, there exists $\delta>0$, such that whenever $a<x,y<b$ with $|x-y|<\delta$, then

$$\sup_{0\leqslant s\leqslant t}\|(F_j(s,x)-F_j(s,y))e(\psi)\|<\epsilon.$$

Note: (i) Condition (ii) is considerably weakened in Ref. 19.

(ii) In the next section, we will encounter a situation where A(i) and A(ii) do not hold on the whole of $\mathbb{R}^+$.

In the sequel, we will find it useful to introduce the notation

$$\alpha(t,\psi;F_1,F_2,F_3)=\max_{1\leqslant j\leqslant 3}\sup_{a\leqslant x\leqslant b}\sup_{0\leqslant s\leqslant t}\{\|F_j(s,x)e(\psi)\|^2\},$$

and note that the admissibility conditions ensure that this expression is finite.

The procedure of Ref. 13 is to define a sequence $\mathscr{P}_n=\{a=x_0^n<x_1^n<\cdots<x_{m(n)+1}^n=b\}$ of partitions of $[a,b]$, such that $\lim_{n\to\infty}|\mathscr{P}_n|=0$ and define the operators (3.1), which we denote as $(M(t),t\in\mathbb{R}^+)$ on $\mathscr{E}$ by $M(t)e(\psi)=\lim_{n\to\infty}M_n(t)e(\psi)$ for each $t\in\mathbb{R}^+$, where

$$M_n(t)=\sum_{r=0}^{m(n)}\int_0^t (F_1(s,y_r^n)A^\dagger(ds,(x_r^n,x_{r+1}^n))+F_2(s,y_r^n)\Lambda(ds,(x_r^n,x_{r+1}^n))$$
$$+F_3(s,y_r^n)A(ds,(x_r^n,x_{r+1}^n))),$$

and each $x_r^n<y_r^n<x_{r+1}^n$.

$(M(t),t\in\mathbb{R}^+)$ is itself an adapted process and each of the maps $t\to M(t)e(\psi)$ is continuous from $\mathbb{R}^+$ to $\mathscr{H}$.

The extension of these results to the case $a\to-\infty$, $b\to\infty$ can be found in the papers quoted at the section head, but we will not require this herein.

In the sequel, we will need the following two results from Ref. 13, Theorem (3.4)—[note that here and in the sequel, $(\cdot,\cdot)$ denotes the inner product in K and $\langle\cdot,\cdot\rangle$ is that in $\mathscr{H}$].

For any ψ_1, ψ_2, $\psi\in\mathscr{M}$, $t\in\mathbb{R}^+$,

$$\langle e(\psi_1), M(t)e(\psi_2)\rangle$$
$$= \int_0^t \int_a^b \langle e(\Psi_1), [(\psi_1(s), P(dx)u)F_1(s,x) + (\psi_1(s), P(dx)\psi_2(s))F_2(s,x)$$
$$+ (P(dx)u, \psi_2(s))F_3(s,x)]e(\psi_2)\rangle ds; \tag{3.2}$$

$$\text{(ii)} \quad \|M(t)e(\psi)\|^2 \leqslant \int_0^t \int_a^b e^{t-s}\left(\sum_{j=1}^3 \|F_j(s,x)e(\psi)\|^2\right)\rho_{s,\psi}(dx)ds, \tag{3.3}$$

where, for each $0 \leqslant s \leqslant t$, $G \in \mathscr{B}([a,b])$,

$$\rho_{s,\psi}(G) = K(\psi,t)(\|P(G)u\|^2 + \|P(G)f(s)\|^2), \tag{3.4}$$

with each $K(\psi,\cdot)$ being locally bounded.

Now let $E \in \mathscr{B}([a,b])$ and let P_E denote the projection-valued measure defined on $\mathscr{B}(E)$ by

$$P_E(A) = P(E \cap A).$$

Note: Of course, $P_E(E) \neq I$, but this causes us no difficulty. Now replace P by P_E in (2.2) and (3.1). We denote the corresponding noise processes as A_E, Λ_E, and $A_E^\dagger$, and the corresponding quantum stochastic spectral integrals as

$$M(t,E) = \int_E \int_0^t (F_1(s,x)A^\dagger(ds,dx) + F_2(s,x)\Lambda(ds,dx) + F_3(s,x)A(ds,dx))$$
$$= \int_a^b \int_0^t (F_1(s,x)A_E^\dagger(ds,dx) + F_2(s,x)\Lambda_E(ds,dx) + F_3(s,x)A_E(ds,dx)). \tag{3.5}$$

From now on we will take S to be $[a,b]$ and $\mathscr{T}$ to be $\mathscr{B}([a,b])$.

Theorem 3.1: M is a quantum martingale measure.

Proof: The martingale property is a standard exercise in quantum stochastic calculus that we outline for completeness. Let $s<t$ and let $\psi_{1,s}$, $\psi_{2,s} \in \mathscr{M}_s$; then, by adaptedness,

$$\langle e(\psi_{1,s}) \otimes e_0^s, M(t,E)e(\psi_{2,s}) \otimes e_0^s\rangle$$
$$= \langle e(\psi_{1,s}), M(s,E)e(\psi_{2,s})\rangle\rangle + \langle e(\psi_{1,s}), e(\psi_{2,s})\rangle\langle e_0^s, (M(t,E) - M(s,E))e_0^s\rangle$$
$$= \langle e(\psi_{1,s}), M(s,E)e(\psi_{2,s})\rangle \text{ by (3.2).}$$

Hence, for arbitrary ψ_1, $\psi_2 \in \mathscr{M}$, we have

$$\langle e(\psi_1), \Pi_s M(t,E)\Pi_s e(\psi_2)\rangle = \langle e(\psi_1), M(s,E)e(\psi_2)\rangle,$$

and the result follows from the totality in $\mathscr{H}$ of the exponential vectors.

To show that M is a martingale measure, we again write $\psi = f \otimes k$ as in Sec. II and let $(E_n, n \in \mathbb{N})$ be a disjoint partition of $[a,b]$.

Our first step is to show that $\Sigma_{n=1}^\infty M(t,E_n)$ exists on $\mathscr{E}$.

By (3.5), (3.2), and the finite additivity of $A^\dagger$, Λ, and A, it is not difficult to verify that

$$M_n(t) = \sum_{j=1}^n M(t,E_j) = M\left(t, \bigcup_{j=1}^n E_j\right), \quad \text{on } \mathscr{E}.$$

Now, by (3.3), we have

$$\|(M_n(t)-M_m(t))e(\psi)\|^2 \leqslant e^t \alpha(t,\psi;F_1,F_2,F_3) \sup_{o\leqslant s\leqslant t}[K(\psi,s)(1+|f(s)|^2)]$$

$$\times \sum_{j=m+1}^{n} (\|P(E_j)u\|^2+\|P(E_j)k\|^2)\to 0, \quad \text{as } n,m\to\infty.$$

So the sequence $(M_n(t)e(\psi), n\in\mathbb{N})$ is uniformly Cauchy and hence uniformly convergent on each $[0,t)$.

To verify countable additivity, it follows from (3.2) that

$$\left\langle e(\psi_1)\left(\sum_{n=1}^{\infty} M(t,E_n)-M(t,\bigcup_{n=1}^{\infty} E_n)\right)e(\psi_2)\right\rangle = 0,$$

and the result follows. □

Note: (i) The countable additivity result in Proposition 2.1 is obtained by taking each $F_j(x,t)=c_j$ in the above.

(ii) From (3.3), it is easily seen that on taking $\psi=f\otimes k$, M is dominated with $h(t)=e^t t(1+\sup_{0\leqslant s\leqslant t}|f(s)|^2)K(\psi,t)\alpha(t,\psi;F_1,F_2,F_3)$ and

$$\mu_\psi(E)=3(\|P(E)u\|^2+\|P(E)k\|^2).$$

The following notation will be useful in the sequel. Let $(V(t,x),\ t\in\mathbb{R}^+,\ x\in[a,b])$ be a family of densely defined linear operators in $\mathscr{H}$, such that (i) each Dom $V(t,x)\subseteq$Ran $F_j(t,x)$ $(j=1,2,3)$; (ii) the families $(V(t,x)F_j(t,x),\ t\in\mathbb{R}^+,\ x\in[a,b])$ satisfy the admissibility conditions.

We can then make the definition

$$\int_a^b\int_0^t V(s,x)M(ds,dx)=\int_a^b\int_0^t (V(s,x)F_1(s,x)A^\dagger(ds,dx)+V(s,x)F_2(s,x)\Lambda(ds,dx)$$
$$+V(s,x)F_3(s,x)A(ds,dx)). \tag{3.6}$$

We need a final technical result. Suppose that V satisfies all the conditions quoted above for $0\leqslant s<t$ but is undefined at $s=t$. We may still want to make sense of (3.6). We proceed as follows. Define for all $t\in\mathbb{R}^+$ the sequence of linear operators $(I_n(t),n\in\mathbb{N})$ on the domain $\mathscr{E}$ by

$$I_n(t)=\int_a^b\int_0^{t-1/n} V(s,x)M(ds,dx).$$

Lemma 3.2: $(I_n(t)e(\psi),n\in\mathbb{N})$ converges in $\mathscr{H}$ for each $t\in\mathbb{R}^+$.

Proof: The fact that the sequence is Cauchy is a consequence of the estimate (3.3) and the result follows immediately. □

We may now define, for each $\psi\in\mathscr{M}$,

$$\int_a^b\int_0^t V(s,x)M(ds,dx)e(\psi)=\lim_{n\to\infty} I_n(t)e(\psi). \tag{3.7}$$

IV. QUANTUM STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS

In the sequel we will, for $a>0$, take $S=[0,a]$ and $\mathscr{T}=\mathscr{B}([0,a])$. We will also take $K=L^2([0,a])$, $\mathscr{D}$ to be the locally bounded functions on $[0,a]$, and for each $E\in\mathscr{T}$, $P(E)=\chi_E$. We will also fix M to be the martingale measure defined by (3.5), which we call a *Fock space–time noise* on $\mathbb{R}^+\times[0,a]$ controlled by F_1, F_2, and F_3. We will also take each F_j to be such that each $F_j(t,x)$ leaves the domain $\mathscr{E}$ invariant.

From now on we will always take $\psi=f\otimes k$, and from (3.3) we obtain, for each $E\in\mathscr{T}$, $t\in\mathbb{R}^+$,

$$\rho_{t,\psi}(E)\leqslant\tau_{t,\psi}(E),$$

where

$$\tau_{s,\psi}(E)=K(\psi,t)(1+\sup_{0\leqslant s\leqslant t}|f(s)|^2)\left(\int_E(|u(x)|^2+|k(x)|^2)\right)dx. \tag{4.1}$$

In the future, we will find it more useful to employ the estimate (3.3) with the measure τ replacing ρ. We will denote the Radon–Nikodym derivatives $d\tau_{t,\psi}/dx$ as $g_{t,\psi}$ and note that our assumptions imply that $g_{t,\psi}$ is locally bounded.

Our goal in this and subsequent sections is to study the *quantum cable equation*, given formally as

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}-U+U\dot{M}, \tag{4.2}$$

whose solution $U=(U(t,x),t\in\mathbb{R}^+,x\in[0,a])$ will be a family of linear operators in $\mathscr{H}$ with each $\mathrm{Dom}(U(t,x))\supseteq\mathscr{E}$. We impose the following reflecting boundary conditions where $u_0\in L^2([0,a])$:

$$\begin{gathered}\frac{\partial U}{\partial t}(t,0)=\frac{\partial U}{\partial t}(t,a)=0,\quad \text{for all } t\in\mathbb{R}^+,\\ U(0,x)=u_0(x),\quad \text{for all } x\in[0,a].\end{gathered} \tag{4.3}$$

To make rigorous sense of (4.2), we write it in weak form. To be precise, we say that U is a solution of (4.2) subject to (4.3) if and only if, for each $\phi\in C^\infty([0,a])$ with $\phi'(0)=\phi'(a)=0$, $\psi\in\mathscr{M}$, $t\in\mathbb{R}^+$, we have

$$\begin{aligned}\int_0^a(U(t,x)-u_0(x))\phi(x)e(\psi)dx&=\int_0^t\int_0^a U(s,x)(\phi''(x)-\phi(x))e(\psi)dxds\\&\quad+\int_0^a\int_0^t U(s,x)\phi(x)M(ds,dx)e(\psi),\end{aligned} \tag{4.4}$$

where the first two integrals in (4.4) are Bochner integrals and the last is a quantum spectral stochastic integral of the form (3.6).

Let G: $(0,\infty)\times[0,a]^2\to\mathbb{R}^+$ be the Green's function for the homogeneous equation associated to (4.2) and (4.3). An explicit form for G can be found in Ref. 1, p. 312. We will quote the following properties that we will need below.

(G1) G is continuous in all three variables (but not at $t=0$, where it is undefined),

(G2) $\int_0^a G_s(x,y)G_t(y,z)=G_{s+t}(x,z)$, for all $x,z\in[0,a]$,

(G3) $G_t(x,y)=G_t(y,x)$, for each $t\in\mathbb{R}^+,x,y\in[0,a]$,

(G4) for each $0<r<3$, $\int_0^a G_t^r(x,y)dy\leqslant Ce^{-\mathrm{tr}}t^{(1-r)/2}$,

which is integrable on $(0,\infty)$.

We note that if ϕ is as in (4.4), then $G_t(\phi,y)$ also solves the homogeneous equation, where

$$G_t(\phi,y)=\int_0^a G_t(x,y)\phi(x)dx, \tag{4.5}$$

for $t\in\mathbb{R}^+$, $0\leqslant y\leqslant a$.

Now consider the following integrodifferential equation in $\mathscr{H}$ for $0\leqslant x\leqslant a$, $t\in\mathbb{R}^+$,

$$V(t,x)=\int_0^a u_0(y)G_t(x,y)dx+\int_0^a\int_0^t V(s,y)G_{t-s}(x,y)M(ds,dy), \tag{4.6}$$

whose solution (if it exists) will be a family $V=(V(t,x),t\in\mathbb{R}^+,0\leqslant x\leqslant a)$ of linear operators on $\mathscr{H}$ with each $\mathrm{Dom}(V(t,x))\supseteq\mathscr{E}$. We call (4.6) the *auxiliary equation* to (4.4) and we will discuss the existence and uniqueness of solutions in the next section. We close this one by establishing the following key result.

Theorem 4.1: There is a one-to-one correspondence between solutions U to (4.4) and solutions V to (4.6), for which for all $T>0$, $\psi\in\mathscr{M}$, $\sup\{\|V(t,x)e(\psi)\|,\|U(t,x)e(\psi)\|,0\leqslant x\leqslant a,0<t\leqslant T\}<\infty$.

Proof: This is essentially along the same lines as Ref. 1.

First, suppose that U is a solution to (4.3). We extend (4.4) to appropriate functions φ of both x and t as in Ref. 1, p. 313 and then argue as on p. 314 therein. So, for each $0\leqslant y\leqslant a$, $t\in\mathbb{R}^+$, define $\varphi^n_{y,t}(s,x)=G_{t-s}(\phi^n_y,x)$ by (4.5), where $\lim_{n\to\infty}\phi^n_y(x)=\delta(x-y)$ in the sense of distributions.

Now, the extended version of (4.4) becomes, for each $\psi\in\mathscr{M}$, $n\in\mathbb{N}$,

$$\int_0^a U(t,x)\phi^n_y(x)e(\psi)dx=\int_0^a u_0(x)G_t(\phi^n_y,x)e(\psi)dx + \int_0^a\int_0^t U(s,x)G_{t-s}(\phi^n_y,x)M(dx,ds)e(\psi),$$

and if ψ_1, $\psi_2\in\mathscr{M}$ it is not difficult to establish that (in the sense of distributions), and for all $t\in\mathbb{R}^+$, $0\leqslant y\leqslant a$,

$$\lim_{n\to\infty}\left\langle e(\psi_1),\int_0^a U(t,x)\phi^n_y(x)dxe(\psi_2)\right\rangle=\langle e(\psi_1),U(t,y)e(\psi_2)\rangle,$$

and, using (3.2) and (G3),

$$\lim_{n\to\infty}\left\langle e(\psi_1),\int_0^a\int_0^t U(s,x)G_{t-s}(\phi^n_y,x)M(dx,ds)e(\psi_2)\right\rangle =\left\langle e(\psi_1),\int_0^a\int_0^t U(s,y)G_{t-s}(x,y)M(dx,ds)e(\psi_2)\right\rangle,$$

from which the required result follows.

Now, suppose that V is a solution of (4.6). We simply imitate the argument of Ref. 1, p. 317, utilizing (3.2) and Fubini's theorem, where it is required, to show that

$$\left\langle e(\psi_1),\left\{\int_0^a(U(t,x)-u_0(x))\phi(x)dx-\int_0^t\int_0^a U(s,x)(\phi''(x)-\phi(x))dxds - \int_0^a\int_0^t U(s,x)\phi(x)M(ds,dx)\right\}e(\psi_2)\right\rangle=0.$$

V. EXISTENCE, UNIQUENESS, AND PROPERTIES OF SOLUTIONS

In this section, we will make a further simplification on the noise, namely, we will require that each F_j is a jointly continuous complex-valued function on $\mathbb{R}^+\times[0,a]$.

We also introduce the notation

$$\beta(t,\psi,F_1,F_2,F_3)=3e^t\sup_{U\leqslant x\leqslant a}g_{t,\psi}(x)\alpha(t,\psi;F_1,F_2,F_3),$$

for $\psi=f\otimes k$, $t\in\mathbb{R}^+$.

Following Ref. 1, we seek a solution to (4.6) by using Picard iteration, and we have the following.

Theorem 5.1: There exists a sequence $(V_n(t,x),\ 0\leqslant x\leqslant a,\ t>0,\ n\in\mathbb{N}\cup\{0\})$ of linear operators in $\mathscr{H}$ for which

$$V_0(t,x)=\int_0^a u_0(y)G_t(x,y)dy,$$

$$V_n(t,x)=V_0(t,x)+\int_0^a\int_0^t V_{n-1}(s,y)G_{t-s}(x,y)M(ds,dy),$$

which is such that (a) each V_n satisfies the admissibility conditions A(i) and A(ii); (b) each map $s\to V_n(s,x)e(\psi)$ is continuous on $(0,t)$ for $0\leqslant x\leqslant a,\ \psi\in\mathscr{M}$.

Proof: We proceed by induction on n. Clearly V_0 is well defined and by (G1) we deduce that it is continuous on $(0,t)$ and that it satisfies the admissibility condition (ii) on $(0,t)$. Now, suppose that V_{n-1} is well defined, continuous on $(0,t)$, and satisfies the admissibility conditions.

Define for $0<s<t-1/m,\ 0\leqslant x,y\leqslant a,\ 1\leqslant j\leqslant 3$,

$$T^j_{n-1}(x,y,s)=V_{n-1}(s,y)G_{t-s}(x,y)F_j(s,y);$$

then the map $s\to T^j(x,y,s)e(\psi)$ is clearly adapted and is continuous on $(0,t-1/m)$, hence, by proposition 25.7 of Ref. 11 (p. 190), it follows that the admissibility condition (i) is satisfied.

The second admissibility condition (ii) is verified by a straightforward $\varepsilon/2$ argument. Hence, each $V_n(s,x)e(\psi)$ is well defined on each $(0,t-1/m)\times(0,a)$, is clearly continuous in s, and is easily seen via (3.3) to satisfy the admissibility conditions. Each V_n can now be extended to $(0,t]$ by the procedure of (3.7). □

Theorem 5.2: There exists a unique solution to the auxiliary equation (4.6)

Proof (existence)*:* For $n\in\mathbb{N},\ \psi=f\otimes k,\ t\in\mathbb{R}^+$ define

$$S_n(t)=\sup_{0\leqslant x\leqslant a}\|(V_n(t,x)-V_{n-1}(t,x))e(\psi)\|.$$

It is easily deduced from A(ii) that each $S_n(t)<\infty$.

Now, by (3.3) we have

$$S_n(t)^2\leqslant\sup_{0\leqslant x\leqslant a}\int_0^t\int_0^a e^{t-s}\left(\sum_{j=1}^3\|(V_{n-1}(s,y)-V_{n-2}(s,y))G_{t-s}(x,y)F_j(s,y)e(\psi)\|^2\right)\tau_{s,\psi}(dy)ds$$

$$\leqslant\beta(t,\psi,F_1,F_2,F_3)\sup_{0\leqslant x\leqslant a}\int_0^t\int_0^a G_{t-s}(x,y)^2\|(V_{n-1}(s,y)-V_{n-2}(s,y))e(\psi)\|^2dyds.$$

We can now follow the procedure of Ref. 1, pp. 315–316 and use Hölder's inequality to obtain, for a given $p\geqslant 2$, that there exists $c>-1$, $K>0$, such that

$$S_n(t)^p\leqslant K\beta(t,\psi,F_1,F_2,F_3)^{p/2}\int_0^t S_{n-1}(s)^p(t-s)^c ds;$$

thus, by Lemma 3.3 of Ref. 1, for each $T\geqslant 0$, there exists $K_1>0$ and $q\in\mathbb{N}$, such that for all $n,m\in\mathbb{N}$ and $0\leqslant t\leqslant T$,

$$S_{n+mq}(t)^p\leqslant K_1^m\int_0^t S_n(s)^p\,\frac{(t-s)}{(m-1)!}\,ds.$$

Since each S_n is continuous on $[0,t]$, we easily deduce that

$$\sum_{n=0}^{\infty} S_n(t) = \sum_{k=0}^{q-1} \sum_{m=0}^{\infty} S_{k+mq}(t) \quad \text{converges uniformly on } [0,T].$$

Hence, $(V_n(t,x)e(\psi),\ n\in\mathbb{N})$ is uniformly Cauchy and hence uniformly convergent on $[0,a]\times[0,T]$. We define

$$V(t,x)e(\psi) = \lim_{n\to\infty} V_n(t,x)e(\psi).$$

It is easily verified that each $V(t,x)$ is a densely defined linear operator in $\mathscr{H}$ with each Dom $V(t,x)\supseteq\mathscr{E}$.

To see that V satisfies the equation (4.6), we have by (3.3) that

$$\left\| \int_0^a \int_0^t (V(s,y)-V_{n-1}(s,y))G_{t-s}(x,y)M(ds,dy) \right\|^2$$

$$\leqslant \beta(t,\psi,F_1,F_2,F_3) \sup_{0\leqslant s\leqslant t} \sup_{0\leqslant y\leqslant a} \|(V(s,y)-V_{n-1}(s,y))e(\psi)\|^2 \int_0^t \int_0^a G_{t-s}(x,y)^2 dyds,$$

and the right hand tends to zero as $n\to\infty$ by (G4) and the uniformity of the convergence.

(Uniqueness). This is established by an obvious extension of the argument of Ref. 1, pp. 314–315.

We now investigate some of the properties of the solution. It is clear from the construction that V is adapted to the filtration $(\mathscr{H}_t,\ t\in\mathbb{R}^+)$. By Theorem 5.1 and the uniformity of the convergence on compacts, it is easily deduced that the maps $s\to V(s,y)e(\psi)$ and $y\to V(s,y)e(\psi)$ are continuous on each $(0,t)$ and $(0,a)$, respectively, where $\psi\in\mathscr{M}$ is fixed, from which it follows that the map $(s,y)\to V(s,y)e(\psi)$ is jointly continuous on each $[0,t)\times(0,a]$.

By Theorem 4.1, we immediately deduce that V is the unique solution to the quantum cable equation (4.4).

Now, we formulate and prove the Markov property of the solution. Suppose that $X=(X(t,y),t\in\mathbb{R}^+,0\leqslant y\leqslant a)$ is a family of linear operators in $\mathscr{H}$ with each $\mathrm{Dom}(X(t,y))\supseteq\mathscr{E}$ and such that each $X(\cdot,y)$ is an adapted process. For each $t\in\mathbb{R}^+$, define the Hilbert space $\mathscr{H}_{\{t\}}$ to be the closed linear span of $\{X(t,y),\ 0\leqslant y\leqslant a\}$. We say that X satisfies the *Markov property* if for each $\psi\in\mathscr{M},\ 0\leqslant y\leqslant a,\ s>0$,

$$\Pi_t X(t+s,y)\Pi_t e(\psi)\in\mathscr{H}_{\{t\}}.$$

Remark: This clearly gives the classical Markov property in the case where M is a Brownian sheet and $\mathscr{H}$ is realized as Wiener space. It is interesting to compare this concept of Markovianity with the more algebraic notion of Ref. 20 (see also Ref. 21).

Theorem 5.3: V satisfies the Markov property.

Proof: For each ψ_1,ψ_2 we find by (3.2), (G2), and Fubini's theorem that

$$\begin{aligned}\langle e(\psi_1),V(s+t,x)e(\psi_2)\rangle &= \langle e(\psi_1),V_0(s+t,x)e(\psi_2)\rangle \\ &\quad + \left\langle e(\psi_1), \int_0^a \int_0^{t+s} V(u,y)G_{t+s-u}(x,y)M(du,dy)e(\psi_2)\right\rangle \\ &= \int_0^a G_s(x,z)\Bigg\{\langle e(\psi_1),V_0(s,z)e(\psi_2)\rangle \\ &\quad + \left\langle e(\psi_1), \int_0^a \int_0^t V(u,y)G_{t-u}(z,y)M(du,dy)e(\psi_2)\right\rangle\Bigg\}dz \\ &\quad + \left\langle e(\psi_1), \int_0^a \int_t^{t+s} V(u,y)G_{t+s-u}(x,y)M(du,dy)e(\psi_2)\right\rangle.\end{aligned}$$

So we find that

$$V(s+t,x)=\int_0^a G_s(x,z)V(t,z)dz+\int_0^a\int_t^{t+s}V(u,y)G_{t+s-u}(x,y)M(du,dy).$$

Hence, by adaptedness and (3.2), we have, for each $\psi\in\mathscr{M}$ and writing $\psi_t=\psi\chi_{[0,t)}$,

$$\Pi_t V(s+t,x)\Pi_t e(\psi)=\Pi_t\int_0^a G_s(x,z)V(t,z)dz\Pi_t e(\psi)=\int_0^a G_s(x,z)V(t,z)e(\psi_t)dz.$$

Now, since each $G_s(x,z)V(t,z)e(\psi_t)\in\mathscr{H}_{\{t\}}$, it follows that the Bochner integral $\int_0^a G_s(x,z)V(t,z)e(\psi_t)dz\in\mathscr{H}_{\{t\}}$ and the required result follows. □

Now, let V be as above and W be the unique solution of another quantum cable equation (4.4) that is related to U as follows—the initial condition $w(y)=\overline{u(y)}$ and the admissible families G_1, G_2, and G_3 are given as $G_1(t,y)=\overline{F_3(t,y)}$, $G_3(t,y)=F_1(t,y)$ and $G_2(t,y)=F_2(t,y)$ for all $t\in\mathbb{R}^+$. $0\leqslant y\leqslant a$. A standard exercise in quantum stochastic calculus shows that for all $\psi_1,\psi_2\in\mathscr{M}$,

$$\langle e(\psi_1),V(t,y)e(\psi_2)\rangle=\langle W(t,y)e(\psi_1),e(\psi_2)\rangle,$$

from which we deduce that each $V(t,y)$ is closable. If we take each $u(y)=\overline{u(y)}$, $F_3(t,y)=\overline{F_1(t,y)}$ and $F_2(t,y)=\overline{F_2(t,y)}$ we can further deduce that each $V(t,y)$ is symmetric. In a future paper we will investigate the conditions under which the $V(t,y)$'s are essentially self-adjoint.

ACKNOWLEDGMENT

I would like to thank Jiang-Lun Wu for introducing me to the wonderful world of spdes.

[1] J. Walsh, *An Introduction to Stochastic Partial Differential Equations* (Springer, Berlin, 1986), Vol. 1180, pp. 266–439.
[2] E. Pardoux, ''Stochastic partial differential equations, a review,'' Bull. Sci. Math. 2 Sér. **117**, 29–47 (1993).
[3] S. Albeverio and R. Russo, ''Stochastic partial differential equations, infinite dimensional stochastic processes and random fields: A short introduction,'' in *Proceedings of the Euroconference on Non-linear Klein–Gordon and Schrödinger Systems: Theory and Applications*, edited by L. Vasquez, L. Streit, and V. M. Perez-Garcia (World Scientific, Singapore, 1996), pp. 68–86.
[4] S. Albeverio and R. Høegh-Krohn, ''Construction of interacting local relativistic quantum fields in four spacetime dimensions,'' Phys. Lett. B **200**, 108–114 (1988).
[5] S. Albeverio, J-L Wu, and T-S Zhang, ''Parabolic SPDE's driven by Poisson white noise,'' Stochastic Processes and their Applications (in press).
[6] D. Applebaum and J-L Wu, ''Stochastic PDE's driven by Lévy white noise,'' Nottingham Trent University preprint, 1997.
[7] C. Mueller, ''The heat equation with Lévy noise,'' preprint, 1996.
[8] J. Walsh, ''A stochastic model of neural response,'' Adv. Appl. Prob. **13**, 231–281 (1981).
[9] G. Kallianpur and R. Wolpert, ''Infinite dimensional stochastic differential equation models for spatially distributed neurons,'' Appl. Math. Opt. **12**, 125–172 (1984).
[10] R. Penrose, *Shadows of the Mind* (Oxford University Press, Oxford, 1994).
[11] K. R. Parthasarathy, *An Introduction to Quantum Stochastic Calculus* (Birkhäuser Basle, 1992).
[12] R. Hudson and K. R. Parthasarathy, ''Quantum Itô's formula and stochastic evolutions,'' Commun. Math. Phys. **93**, 301–323 (1984).
[13] D. Applebaum, ''Spectral families of quantum stochastic integrals,'' Commun. Math. Phys. **170**, 607–628 (1995).
[14] D. Applebaum and M. Brooks, ''Infinite series of quantum stochastic spectral integrals,'' J. Operator Theory **36**, 295–316 (1996).
[15] G. Vincent-Smith, ''The Itô formula for quantum semimartingales,'' Proc. London. Math. Soc. **75**, 671–720 (1997).
[16] H. Zhiyuan, ''Stochastic integrals on general topological measurable spaces,'' Z. Wahrscheinlichkeitstheor. Verwandte Geb. **66**, 25–40 (1984).
[17] C. Barnett and I. Wilde, ''Random times and time projections,'' Proc. Am. Math. Soc. **110**, 425–440 (1990).
[18] M. Brooks, ''On the construction of quantum stochastic spectral integrals,'' to appear in Quantum Probab. Commun.
[19] M. Brooks, Nottingham Trent University Ph.D. thesis, 1998.
[20] L. Accardi, A. Frigerio, and J. T. Lewis, ''Quantum stochastic processes,'' Publ. Res. Inst. Math. Sci. Kyoto Univ. **18**, 97–133 (1982).
[21] A. Frigerio, ''Covariant Markov dilations of quantum dynamical semigroups,'' Publ. Res. Inst. Math. Sci. Kyoto Univ. **21**, 657–675 (1985).

Interaction energy of generalized Abelian Higgs vortices

K. Arthur
Clockworks Multimedia, Unit 8 IDA Centre, Pierce Street, Dublin 2, Ireland

Y. Brihaye
Physique-Mathematique, Universite de Mons-Hainaut, Mons, Belgium

D. H. Tchrakian
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland, and School of Theoretical Physics — DIAS, 10 Burlington Road, Dublin 4, Ireland

(Received 2 December 1997; accepted for publication 21 January 1998)

The interaction energy between vortices of the generalized Abelian Higgs vortices is studied numerically. These are the solitons of the p-th member of the hierarchy of Maxwell–Higgs systems, the $p=1$ member of which is the usual Abelian Higgs model. It is found that for all p, these vortices have both attractive and repulsive phases, the strengths of these interactions decreasing successively for increasing p. A detailed comparison of the $p=1,2$ and $p=3$ cases is made quantitatively, and an extended model, composed of the $p=1$ and $p=2$ members, is studied in some detail and shown to exhibit attractive and repulsive phases. © *1998 American Institute of Physics.* [S0022-2488(98)02106-9]

I. INTRODUCTION

The hierarchy of generalized Abelian Higgs models[1] arise by subjecting the p-th member of the Yang–Mills hierarchy,[2] with gauge group (chiral) $SO_{\pm}(4p)$ and on the direct product space $\mathbb{R}_2 \times S^{4p-2}$, to dimensional descent. (Indeed by performing the descent on $\mathbb{R}_d \times S^{4p-d}$, one obtains the hierarchies of non-Abelian Higgs models[3] on $\mathbb{R}_d$.) These yield the hierarchy of $U(1)$ gauged Higgs models on $\mathbb{R}_2$ which differ from the usual Abelian Higgs model[4] in that they are Skyrme-like gauge field systems.[5] In the present paper we restrict ourselves to the case where the $U(1)$ gauge dynamics is given by a Maxwell-like term. A corresponding generalization of the usual Chern–Simons–Higgs model employing instead Chern–Simons dynamics was also given in Ref. 6 which is not considered here. Accordingly we will refer to these models as Maxwell–Higgs (MH) models in the following. As far as numerical computations are concerned, we restrict ourselves to the $p=2$ and $p=3$ members of the hierarchy.

In Ref. 1, the self-dual solutions of the p-th member of the MH hierarchy were found by integrating the corresponding Bogomol'nyi equations numerically. Subsequently an analytic proof for the existence of self-dual solutions to the $p=2$ MH model, as well as the corresponding generalized Chern–Simons–Higgs model,[6] was given in Ref. 7. In all aspects except their Skyrme-like features, the $p \geqslant 2$ MH models are similar to the $p=1$ MH model, namely the usual Abelian Higgs model.[4] One such example is illustrated by the study[8] of the zero modes of the rotationally symmetric vortices of the $p=2$ MH system and their scattering, in which the head on collision of n vortices exhibits the π/n symmetry present in the $p=1$ model.

In the background of the above described status of the $p \geqslant 2$ MH models, it is natural to ask yet another question pertinent to soliton systems supporting self-dual solutions, namely, whether there exist attractive and repulsive phases like in the $p=1$ case, discovered long ago by Jacobs and Rebbi.[9] The physical relevance of this question is evident in all areas of vortex dynamics, and it is the aim of the present work to study this problem quantitatively using similar numerical methods as in Ref. 9. We shall consider the cases $p=1,2$ in detail and their interaction properties will be compared. In addition, the simplest example of an extended model composed of both $p=1$ and $p=2$ subsystems will be considered.

In Section II we introduce the p-th member of the hierarchy and discuss the topological lower bound on the energy. In Section III we write down the spherically symmetric one-dimensional

 © 1998 American Institute of Physics

subsystems and give the asymptotic solutions and proceed with the numerical integration. Explicit integrations are carried out for the $p=1,2,3$ MH models, the first of which is a repetition of the work of Ref. 9. Section IV is concerned with the extended model whose Hamiltonian is the sum of the Hamiltonians of several MH models of distinct p, and in Section V we end up with a summary of our results.

II. THE MODELS

The (static) energy density functional for the arbitrary-p MH model is given by

$$\mathcal{H}^{(p)}=(\eta^2-|\varphi|^2)^{2(p-2)}(\lambda_2[(\eta^2-|\varphi|^2)F_{ij}+i(p-1)D_{[i}\varphi^*D_{j]}\varphi]^2 +4p(2p-1)\lambda_1(\eta^2-|\varphi|^2)^2|D_i\varphi|^2+2(2p-1)^2\lambda_0(\eta^2-|\varphi|^2)^4), \tag{1}$$

in 2 Euclidean dimensions. For $p=1$, (1) reduces to the usual Abelian Higgs model. $\varphi=\phi_1+i\phi_2$ is the complex Higgs field, $D_\mu\phi=\partial_i\varphi+iA_i\varphi$, $i=1,2$, is the covariant derivative, and the gauge fields F_{ij} are defined in terms of the real gauge potentials A_i as $F_{ij}=\partial_{[i}A_{j]}$. The square brackets mean antisymmetrization in the indices. Note that the constant η has the same dimensions as the Higgs field φ.

Before proceeding to the main subject of this paper, namely the study of the interaction energies of the vortices of the system (1), we comment on an interesting general feature of the $p\geqslant 2$ MH models. Setting the connection and curvature fields A_μ and $F_{\mu\nu}$ equal to zero in the usual Abelian Higgs model, or the $p=1$ MH model, the resulting mutilated system on $\mathbb{R}_2$ does not support soliton solutions because the Derrick scaling requirements are not then satisfied. This is in contrast to the corresponding situation for all the other $p\geqslant 2$ members of the MH hierarchy, where the Derrick requirement is no longer violated and solitons exist. This was discussed in detail in the Appendix of Ref. 10 and such solitons were studied in detail numerically in Ref. 11. What relation the solitons of the ungauged systems[10,11] have to the vortices of the MH models[1] will be addressed quantitatively in our discussion of the vortices of the $p=2$ MH model.

The densities (1) differ from those derived by dimensional reduction and employed in Refs. 1, 7, 8 in respect of the appearance of the dimensionless parameters λ_0, λ_1, and λ_2. When $\lambda_0=\lambda_1=\lambda_2=1$, the topological inequalities bounding the density (1) from below can be saturated and the resulting Bogomol'nyi equations can then be solved to yield the self-dual solutions presented and employed in Refs. 1, 7, 8. We will not repeat these arguments here as they are given in detail in Ref. 1, save to say that when any one of the three dimensionless coupling constants $\lambda_a\neq 1$, $a=0,1,2$, it is impossible to saturate this bound and hence the finite energy topologically stable soliton solutions we seek satisfy only the second order Euler–Lagrange equations rather than the first order Bogomol'nyi equations as in Refs. 1, 7, 8.

Concerning the appearance of the dimensionless coupling constants λ_a in (1), we note that some of these can be eliminated. If we divide the energy,

$$E^{(p)}(\lambda_0,\lambda_1,\lambda_2)=\int d^2x\mathcal{H}^{(p)}(\lambda_0,\lambda_1,\lambda_2), \tag{2}$$

by, say, λ_1, and suitably rescale the coordinate x_i, we find the relation

$$E^{(p)}(\lambda_0,\lambda_1\lambda_2)=\lambda_1E^{(p)}\left(\frac{\lambda_0\lambda_2}{\lambda_1^2},1,1\right). \tag{3}$$

Accordingly, we shall set $\lambda_1=\lambda_2=1$ in (1) and will retain only λ_0 without loss of generality. When in Section IV we consider the extended model, we shall return to this scaling argument.

The topological stability of the solutions is guaranteed by the lower bound on the energy E, given by the topological charge which can be calculated as a surface integral and is equal to the winding number coinciding with the vorticity n. With the appropriate normalizations, this reads as

$$E\geqslant n. \tag{4}$$

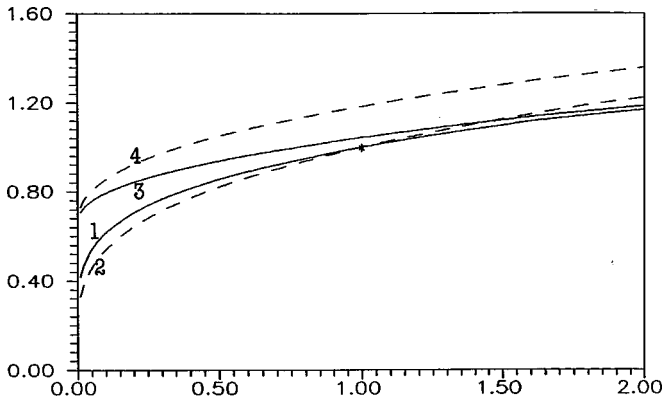

FIG. 1. The energy *per unit charge* of the solution is plotted as a function of λ_0 for the $p=2$ gauged and ungauged models. The curve 1 (resp., 3) refers to the energy of the $n=1$ vortex in the gauged (resp., ungauged) model. The curves 2 and 4 refer to the same data for the $n=2$ vortex. The star represents the self-dual point.

The detailed topological charge densities for $p=1$, 2, and 3 are given explicitly in Ref. 1 by equations (1.4), (2.7a), and (2.7b), respectively, which we do not repeat here. When $\lambda_0=1$ the inequality (4) is saturated.[4,1] It is clear that when $\lambda_0 \geqslant 1$, the topological inequality (4) remains valid. When $0 \leqslant \lambda_0 \leqslant 1$, it can be shown using the argument in Ref. 9 given for the $p=1$ case, that this inequality is altered to read as

$$E \geqslant \lambda_0 n. \tag{5}$$

In Ref. 9 the quantity $\mathcal{E}(n,\lambda_0)=E/n$ was plotted as a function of λ_0 for the solutions to the $p=1$ MH system (1). We shall refer to such graphs as Jacobs–Rebbi plots. At $\lambda_0=1$ all these curves coincide at $\mathcal{E}(n,\lambda_0)=1$ since this point corresponds to all the self-dual solutions for which $E=n$. The stress tensors of all these solutions vanish and the corresponding solitons are therefore noninteracting. It was found[9] that for $\lambda_0>1$ the $n=1$ curve lay below the $n=2$ curve implying repulsion, while for $\lambda_0<1$ the $n=2$ curve lay below the $n=1$ curve implying attraction. We have reproduced this result by repeating the work of Ref. 9 and do not present the graphs of the Jacobs–Rebbi plot here because they are qualitatively the same as those to be given below for the $p=2$ MH model, in Figure 1. We expect to find that this situation holds qualitatively for MH models of arbitrary p, and we have verified it quantitatively for the $p=2$ and $p=3$ cases.

III. RADIAL SYMMETRY

A. The equations

To convert the Euler–Lagrange equations of the system (1) to ordinary differential equations which can readily be integrated numerically, we proceed to subject the system (1) to the radially symmetric Ansatz,

$$\varphi = \eta g(r) e^{-in\theta}, \tag{6}$$

$$A_i = \frac{a(r)-n}{r^2}\varepsilon_{ij}x_j. \tag{7}$$

The resulting reduced one-dimensional energy density functional $H[a(r),g(r)]$ defined by

$$\int H^{(p)}[a(r),g(r)]dr = \frac{1}{2\pi}\int \mathcal{H}^{(p)}d^2x, \tag{8}$$

is then

$$H^{(p)}=\frac{\lambda_2}{r}([(1-g^2)^{p-1}a]')^2+2p(2p-1)\lambda_1\eta^2(1-g^2)^{2(p-1)}$$
$$\times\left[rg'^2+\frac{a^2g^2}{r}\right]+(2p-1)^2\lambda_0\eta^4r(1-g^2)^{2p}, \tag{9}$$

where we have used the notation $g'=dg/dr$; from now on we set $\lambda_1=\lambda_2=1$ in this section.

As a check it is interesting to verify that adding the inequality

$$r(1-g^2)^{2(p-2)}\left(\frac{1}{r}[(1-g^2)a'-2(p-1)agg']+(2p-1)\eta^2(1-g^2)^2\right)^2\geqslant 0, \tag{10}$$

to $2p$ times the inequality

$$\eta^2r(1-g^2)^{2(p-1)}\left(g'-\frac{ag}{r}\right)^2\geqslant 0, \tag{11}$$

yields the functional $H^{(p)}[a,g]$ in (9), plus the total derivative term

$$\sigma_p \overset{def.}{=} 2(2p-1)\eta^2\frac{d}{dr}[(1-g^2)^{2p-1}a], \tag{12}$$

which with a suitable normalization is the topological charge density. The integral of (12) then equals $2(2p-1)\eta^2$ times the vortex number $|n|$ provided that the solution has the following asymptotic values:

$$\lim_{r\to 0}g(r)=0, \quad \lim_{r\to\infty}g(r)=1, \tag{13}$$

$$\lim_{r\to 0}a(r)=n, \quad \lim_{r\to\infty}a(r)=0. \tag{14}$$

We note moreover that the conditions (13)–(14) are consistent with the finiteness of the energy, by inspection of (9).

Anticipating the subsequent numerical integrations, we introduce the function

$$b(r)=(1-g(r)^2)^{p-1}a(r), \tag{15}$$

which is more convenient to employ in the numerical integrations than the function $a(r)$. The Euler–Lagrange equations arising from the variation of $H^{(2)}$ with respect to the arbitrary variations of the function $a(r)$ and of $b(r)$ are identical, so the transformation given by (15) is legitimate. In terms of the dimensionless radial variable $\rho=\eta r$, the Euler–Lagrange equations of (9) with respect to the variations with $g(\rho)$ and $b(\rho)$ are, respectively,

$$[\rho(1-g^2)^{2(p-1)}g_\rho]_\rho+2(p-1)\rho(1-g^2)^{2p-3}g_\rho^2g+(2p-1)\lambda_0\rho(1-g^2)^3g-\frac{1}{\rho}b^2g=0, \tag{16}$$

$$\rho\left[\frac{b_\rho}{\rho}\right]_\rho-2p(2p-1)g^2b=0, \tag{17}$$

where we have used the notation $g_\rho=dg/d\rho$.

The asymptotic solutions in the $r\gg 1$ region can easily be found by linearizing (16) and (17) with the results

$$g(\rho)=1-[K_0(\sqrt{2p(2p-1)\lambda_0}\rho)]^{1/p}, \tag{18}$$

$$b(\rho)=\rho K_1(\sqrt{2p(2p-1)}\rho), \tag{19}$$

in terms of the Bessel functions K_0 and K_1, consistent with the decays of the selfdual solutions in Ref. 1, namely exponential decays. In the $r \ll 1$ region it can be verified that the asymptotic solutions are

$$g(\rho) = A\rho^n + o(\rho^{n+2}), \tag{20}$$

$$a(\rho) = n - B\rho^2 + o(\rho^4), \tag{21}$$

which are consistent with the requirements of analyticity of the fields (6) and (7) at the origin. The constants A and B are determined by the numerical inegration. Note here that except for the $\lambda_0 = \lambda_1 = \lambda_2 = 1$ case when the equations (16)–(17) are solved by the Bogomol'nyi equations arising from the saturation of the inequalities (10)–(11), A and B are independent, while in the self-dual case they are related as given in equations (3.6a)–(3.6b) of Ref. 1.

B. Numerical results

We have integrated numerically the equations of motion (16) and (17) for $p = 1,2,3$ and for vorticities $n = 1,2,3$. The most relevant quantity, for our purpose, is the energy $E(n,\lambda_0)$ per unit charge,

$$\mathcal{E}(n,\lambda_0) = \frac{E(n,\lambda_0)}{n},$$

which we have computed as a function of λ_0 for each type of solution.

The case $p = 1$ was treated in detail by Jacobs and Rebbi;[9] we therefore refrain from presenting it in detail here but in due course we will compare the results obtained for $p > 1$ with their results.

Let us first discuss our results for the $p = 2$ model. The solutions of vorticity $n = 1$ and $n = 2$ exhibit the same property as for the case $p = 1$, namely a repulsion of the two-soliton for $\lambda_0 > 1$ and an attraction for $\lambda_0 < 1$. This behavior is illustrated on Fig. 1 by the curves labelled 1 ($n = 1$) and 2 ($n = 2$). We shall refer to such plots as Jacobs–Rebbi plots in the rest of this paper.

Plots of the profiles of the functions $g(\rho)$ and $b(\rho)$ [or $a(\rho)$] are not presented since they are qualitatively the same as the corresponding curves for the self-dual solutions, exhibited, for example, in Ref. 1.

One of the remarkable features of the $p > 1$ models is that soliton solutions exist also in their ungauged version[10,11] (for $p = 1$ the ungauged model does not support a soliton because Derrick's scaling argument is violated). In Fig. 1 we have superposed the energy of the solitons of the ungauged $p = 2$ model; the curves labelled 3 (resp., 4) representing the energy of the soliton of vorticity 1 (resp., 2). It is seen on the figure that curves 1 and 2 are, respectively, lower than curves 3 and 4. Completing the analysis of Ref. 11 this result provides a nice check of our numerics because it is expected on general grounds, namely that introducing a new set of fields to a system results in a new landscape of energy contours, where the new minima arising because of the new coordinates can only be lower than the old ones.

On the same figure, we also observe that the curves 1,3 (and independently 2,4) approach each other while the parameter λ_0 increases; moreover, the profiles of the function $g(r)$ for a gauged and an ungauged soliton hardly differ in this region. This incited us to analyze the Maxwell field of the gauged solitons. In addition to this, the way in which the Maxwell term enters the $p \geqslant 2$ systems leads to a different dependence of the electromagnetic field energy,

$$E_{e.m.} = \int d^3x F_{ij}^2 \sim \int \frac{a'^2}{r} dr,$$

on the Higgs self-interaction coupling constant λ_0, from that for the $p = 1$ system. The graphs of $E_{e.m.}$ given in Figure 2 show that this quantity increases with λ_0 for the $p = 1$ system and decreases for the $p = 2$ system, which is typical for all $p \geqslant 2$ systems.

The attraction-repulsion phenomenon discovered by Jacobs and Rebbi in the $p = 1$ model is also present in the $p = 3$ model and the energy plot of the $n = 1$ and $n = 2$ solution display the same

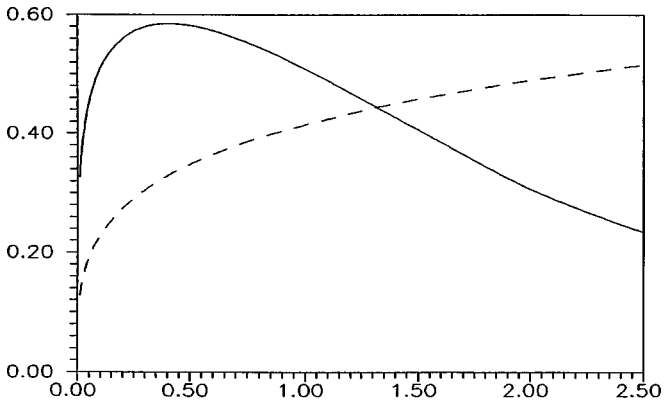

FIG. 2. The energy of the Maxwell field of the $n=1$ vortex as a function of λ_0. The solid (resp., dashed) curve refers to the vortex of the $p=2$ (resp., $p=1$) model.

pattern as in Fig. 1. Because the differences between the magnitudes of the interaction energies for the $p=1,2,3$ models, are not sufficiently large, we have calculated the quantities

$$\Delta E(p,\lambda_0)=\mathcal{E}^{(p)}(2,\lambda_0)-\mathcal{E}^{(p)}(1,\lambda_0), \tag{22}$$

as functions of λ_0. Thus for $\Delta E(p,\lambda_0)>0$ there is repulsion and for $\Delta E(p,\lambda_0)<0$ there is attraction, and the magnitudes of these quantities represent the strengths of the interactions. To illustrate the fact that these interaction energies decrease in magnitude progressively for the higher p MH models, we have listed the values of the three interaction energies $\Delta E(1,\lambda_0)$, $\Delta E(2,\lambda_0)$, and $\Delta E(3,\lambda_0)$ for a small range of λ_0 in Table I.

It is our guess that the attraction-repulsion phenomenon persists for higher p models. Table I manifestly suggests that interaction energies of the higher p MH vortices are of progressively smaller magnitudes. This question arises naturally in the background of our knowledge, from the work of Ref. 1, that the solutions of the higher p models are progressively more tightly localized, i.e., they decay faster at large distances. More generally, this may be of physical interest in quantitative questions arising in the physics of vortices.

IV. EXTENDED MODELS

With the hierarchy of models (1) we can manufacture a family of extended models by considering the linear combination

$$\mathcal{H}=\sum_{p=1}^{N}\eta^{-4(p-1)}\hat{\mathcal{H}}^{(p)}(\lambda^{(p)}), \tag{23}$$

with

$$\lambda^{(p)}\equiv(\lambda_0^{(p)},\lambda_1^{(p)},\lambda_2^{(p)}).$$

TABLE I. Comparison of the quantity (22) for some values of λ_0 and for $p=1,2,3$.

λ_0	$\Delta E(1,\lambda_0)$	$\Delta E(2,\lambda_0)$	$\Delta E(3,\lambda_0)$
0.8000	−0.0134	−0.0126	−0.0123
0.9000	−0.0065	−0.0061	−0.0059
1.0000	0.0000	0.0000	0.0000
1.1000	0.0062	0.0059	0.0057
1.2000	0.0122	0.0115	0.0112

The densitites $\hat{\mathcal{H}}^{(p)}$ are normalized versions of the members (1) of the MH hierarchy:

$$\hat{\mathcal{H}}^{(p)}(\lambda^{(p)})=\frac{1}{2(2p-1)N}\mathcal{H}^{(p)}(\lambda^{(p)}). \tag{24}$$

The normalization, which follows from (12), is made such that the energy of the winding number n vortex should be normalized to n, i.e., if the solution did saturate the topological lower bound, then the energy would equal the topological charge.

The reason for considering systems like (23) is that it may be desirable to have a propagating photon in the theory and hence it would be necessary to include the $p=1$ MH model, which is the only MH system featuring the usual Maxwell term F_{ij}^2 without field dependent coefficients. This may be a particularly relevant requirement in applications to Cosmic Strings.[12]

The mixing parameters giving the relative strengths of each of the component subsystems are absorbed in the coupling constants. Applying the scaling argument leading to the relation (3) to the energy integral pertaining to (23), we can set, as before, $\lambda_1^{(1)}=\lambda_2^{(1)}=1$. This leads to the natural normalization of the kinetic terms and leaves a number of $3N-2$ independent dimensionless coupling constants.

There is one important respect in which the systems (23) differ from the hierarchy of MH systems (1). This is, the absence of solutions which saturate the corresponding topological lower bounds. We illustrate this explicitly in the special case $N=2$ [i.e., the simplest of systems (23)] that we conveniently parametrize as follows:

$$\mathcal{H}=\tfrac{1}{4}\mathcal{H}^{(1)}(\lambda_0,1,1)+\tfrac{1}{12}\eta^{-4}\mathcal{H}^{(2)}(\mu_0,\mu_1,\mu_2). \tag{25}$$

In principle the lower bound on the energy density functional can be saturated when $\lambda_0=1$ and $\mu_0=\mu_1=\mu_2=1$. This absolute minimum will be acheived if all four topological inequalities (10)–(11), namely both for $p=1$ and for $p=2$ are simultaneously saturated. As a result of this we have the following three Bogomol'nyi equations to satisfy:

$$\frac{dg}{dr}=\frac{ag}{r}, \tag{26}$$

$$\frac{da}{dr}=-\eta^2 r(1-g^2), \tag{27}$$

$$(1-g^2)\frac{da}{dr}=-3\eta^2 r(1-g^2)^2+\frac{2}{r}a^2g^2, \tag{28}$$

which can readily be verified to yield the *unacceptable* solution $g^2=\eta^2r^2$, and hence for our purposes (26)–(28) are to be considered overdetermined.

As a consequence of the Bogomol'nyi equations (26)–(28) not being satisfied by the solution to the second-order Euler–Lagrange equations, the stress tensor for this field configuration is *nonzero*, and hence the static soliton solutions do not describe noninteracting vortices. The solutions in question are of course topologically stable by virtue of the pair of topological inequalities (10) and (11) for $p=1$ and $p=2$, respectively. [For the more general systems (23), N pairs of inequalities (10)–(11) are employed for this purpose.] The lower bound on the energy of the system (25) is 1 for all n because of the normalizations, when all the coupling constants $\lambda_0,\mu_0,\mu_1,\mu_2$ *are greater than 1*. When any of these coupling constants *are smaller than 1*, then it follows from an extension of the argument that led to the lower bound (5), that we have

$$E\geqslant \min\{\lambda_0,\mu_0,\mu_1,\mu_2\}. \tag{29}$$

Given this situation, it is interesting to find out by how much the solution differs from the field configuration that saturates the topological lower bound. We refer to this as the putative self-dual field configuration. The most important question is: does the system (25), and hence also all

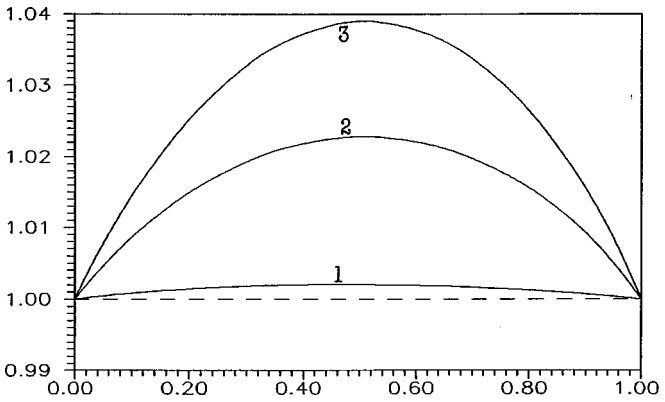

FIG. 3. The energy *per unit charge* of the vortex of the extended model (30) as a function of the mixing parameter θ. The labels 1,2,3 refer to the vorticity.

systems (23), support vortices of both attractive and repulsive phases like the individual members (1) of the MH hierarchy? Both these questions can be answered by performing the relevant numerical experiments on (25).

We have first considered the system

$$\mathcal{H}=\frac{1-\theta}{2}\mathcal{H}^{(1)}(1,1,1)+\frac{\theta}{6}\eta^{-4}\mathcal{H}^{(2)}(1,1,1), \tag{30}$$

where the $p=1$ and $p=2$ self-dual models are superposed by means of a mixing parameter θ. (By self-dual models we mean models whose lower bound on the energy can be saturated by the Bogomol'nyi equations, *and,* the latter support finite energy solutions.) As a check of our equations, we found that the energies *per unit charge* corresponding to the self-dual solutions (i.e., for $\theta=0$ and $\theta=1$) are indeed equal to 1.0. In between, the energy becomes slightly higher than $E=1$ and the numerical result indicates that the deviation culminates at $\theta=\frac{1}{2}$. At this value, the energy of solutions with vorticities $n=1$, $n=2$, and $n=3$ are, respectively, 1.0020, 1.0228, and 1.0390. These values differ very little from the putative self-dual case ($E=1.0$) and their departure from the latter grows with the vorticity n, nonetheless staying within 4 *percent* up to $n=3$.

The energy curves correponding to the $n=1,2,3$ solitons as a function of θ are depicted in Figure 3; it clearly shows that the self-duality property is maximally violated for the model (30) for $\theta=\frac{1}{2}$, or, equivalently, the model (25) for $\lambda_0=\mu_0=\mu_1=\mu_2=1$. Remarkably the three curves are almost symmetric under the reflection $\theta\rightarrow(\theta-1)$; we do not have any analytical explanation for this observation.

In order to offer some answers to the questions posed above, we solved numerically the equations associated with (25) for numerous values of the coupling constants. In fact we studied the equations for all but one of the coupling constants set equal to 1 and varying the remaining one. This way, the set of models addressed was ''relatively close'' to the model (30) with $\theta=\frac{1}{2}$ (i.e., where self-duality is maximally violated).

The evaluation of the energy of $n=1$ and $n=2$ vorticity solutions as a function of λ_0 (with $\mu_0=\mu_1=\mu_2=1$) is summarized by the two solid curves in Fig. 4. The energy *per unit charge* of the $n=2$ solution lies always above the energy of the $n=1$ solution. We conclude that two *unit* vortices at large distances apart will tend to increase their mutual distances to minimize the energy, i.e., repel.

Exactly the same scenario occurs when we vary the coupling constant μ_2 (with $\lambda_0=\mu_0=\mu_1=1$), as depicted by the two dashed lines of Fig. 4. The situation becomes, however, physically more interesting when the parameters μ_0 and/or μ_1 are varied, as discussed now.

Varying the parameter μ_0 (with $\lambda_0=\mu_1=\mu_2=1$), we observe that the $n=1$ and $n=2$ vortices energy are in an attractive phase for $\mu_0<\mu_0^{cr}\approx 0.64$ and in a repulsive phase for $\mu_0>\mu_0^{cr}$. This is clearly indicated in Fig. 5 by the two solid curves. Considering finally the variation of the parameter μ_1, we observe from the dashed curves in Fig. 5 the occurrence of a repulsion phase for

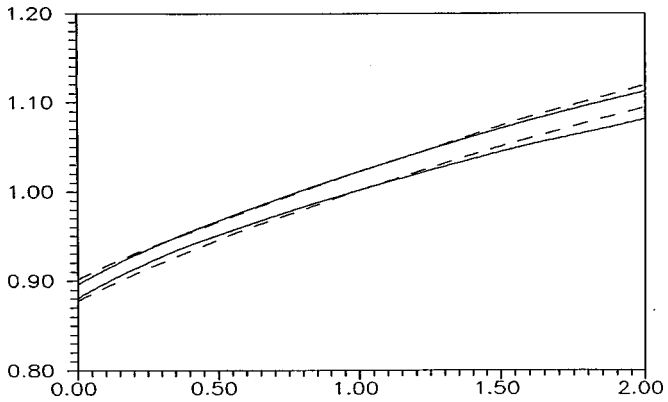

FIG. 4. The energy *per unit charge* of solutions of the extended model (25). The two solid (resp., dashed) lines represent the energy as a function of λ_0 (resp., μ_2); all other coupling constants are set to 1. In the two cases, the upper (resp., lower) line refers to the $n=2$ (resp., $n=1$) vortex.

$\mu_1<\mu_1^{cr}\approx 1.45$ and of an attraction phase for $\mu_1>\mu_1^{cr}$. We note from Fig. 5 that, contrary to the case of μ_0 varying, the two-vortex system is in an attractive phase for $\mu_1>\mu_1^{cr}$ ($\mu_1\approx 1.45$) and in a repulsive phase if $\mu_1<\mu_1^{cr}$.

The occurrence of Jacobs–Rebbi plots or, equivalently, of attraction-repulsion phase transitions in some specific regions of the space of the coupling constants deserves to be studied in more detail. To this end, we remark on two important features of self-dual models.

(i) The energies *per unit charge* of the n-vortex solutions are all equal to 1, irrespectively of the interger n.

(ii) All components of the stress energy tensor T_{ij} vanish, implying noninteracting solitons.

The extended models considered in this Section *do not* support self-dual solutions, thus neither (i) nor (ii) above are satisfied. Nevertheless, in some regions of the space of coupling parameters these criteria are minimally violated. We refer to these regions of the parameter space as ''almost-self-dual regions.''

For our quantitative considerations, it will be convenient to restrict just to the (2,2) component of T_{ij}, whose restriction to the one-dimensional spherically symmetric subsystem is given by

$$T_{22}=\frac{1}{8}\left(\left[\frac{a_\rho^2}{\rho}-\rho(1-g^2)^2\right]+2\left[\rho g_\rho^2-\frac{a^2g^2}{\rho}\right]\right)$$
$$+\frac{1}{24}\left(\left[\frac{[(1-g^2)a]_\rho^2}{\rho}-9\mu_0\rho(1-g^2)^4\right]+12\mu_1(1-g^2)^2\left[\rho g_\rho^2-\frac{a^2g^2}{\rho}\right]\right). \tag{31}$$

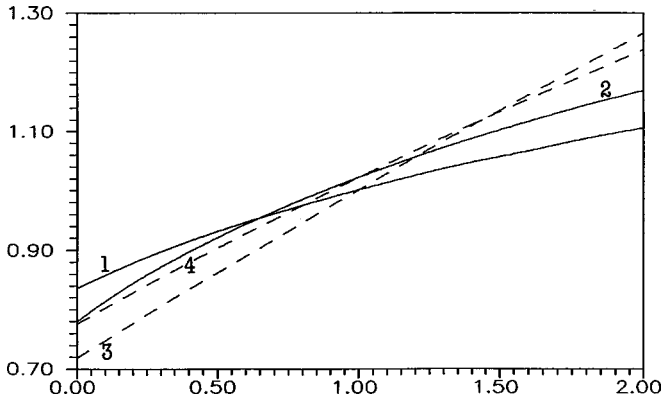

FIG. 5. The energy *per unit charge* of solutions of the extended model (25). The two solid (resp., dashed) lines represent the energy as a function of μ_0 (resp., μ_1); all other coupling constants are set to 1. Curves 1 and 3 (resp., 2 and 4) refer to energy of the $n=1$ (resp., $n=2$) vortex.

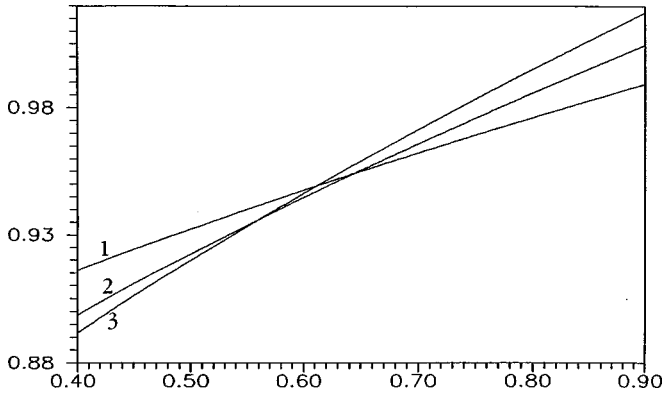

FIG. 6. The details of the energy *per unit charge* for the solutions of vorticity $n=1,2,3$ for the model (25) as a function of μ_0 (other coupling constants are set equal to 1).

In order to characterize more quantitatively the ''almost self-dual region'' of the space of coupling constants obtained above, we calculated *also* the $n=3$ vortex solutions in this region. We evaluated, for the solutions with vorticities $n=1,2,3$, the quantity

$$F=\int (T_{22})^2 dr. \tag{32}$$

We now discuss the results obtained for when we vary the coupling constant μ_0, with all others set equal to 1 as before. The curves of the energies *per unit charge* of the $n=1,2,3$ vorticities are given in Fig. 6. The three curves intersect at the following points:

$$n=1 \text{ intersects } n=2 \quad \text{at } \mu_0=\mu_0^{cr}\approx 0.64, \quad E\approx 0.954,$$

$$n=1 \text{ intersects } n=3 \quad \text{at } \mu_0=\mu_0^{cr}\approx 0.61, \quad E\approx 0.948,$$

$$n=2 \text{ intersects } n=3 \quad \text{at } \mu_0=\mu_0^{cr}\approx 0.56, \quad E\approx 0.941.$$

The values of μ_0 where the intersections occur are indeed very close to each other. Next, the three, $n=1,2,3$, curves of the positive definite quantitity (32) versus μ_0 are depicted in Fig. 7. We see that this quantity F attains very sharp minima at

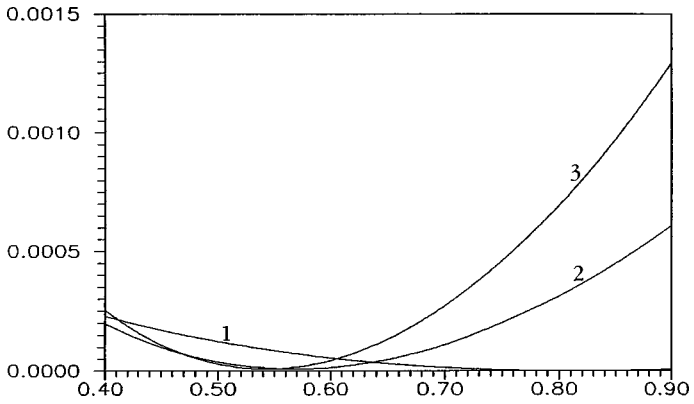

FIG. 7. The quantity (32) for the solutions of vorticity $n=1,2,3$ for the model (25) as a function of μ_0 (other coupling constants are set to 1).

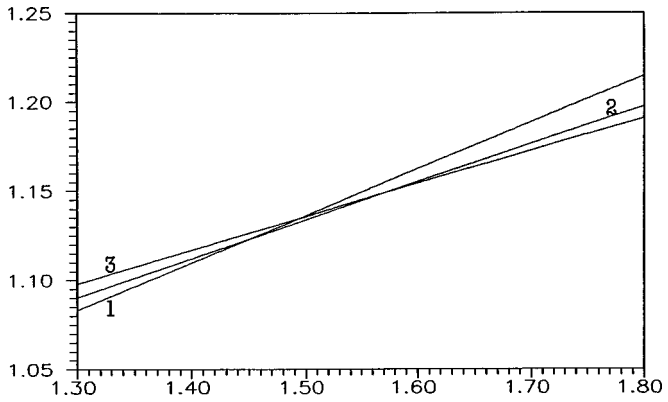

FIG. 8. The details of the energy *per unit charge* for the solutions of vorticity $n=1,2,3$ for the model (25) as a function of μ_1 (other coupling constants are set equal to 1).

$$\mu_0=\mu_0^{min}\approx 0.81, \quad F\approx 0.2\times 10^{-6}, \quad E\approx 0.977, \quad \text{for } n=1,$$

$$\mu_0=\mu_0^{min}\approx 0.57, \quad F\approx 0.8\times 10^{-4}, \quad E\approx 0.938, \quad \text{for } n=2,$$

$$\mu_0=\mu_0^{min}\approx 0.54, \quad F\approx 0.7\times 10^{-4}, \quad E\approx 0.931, \quad \text{for } n=3.$$

A similar analysis was carried out when we varied the coupling constant μ_1, leading to very similar results. The energy is depicted in Fig. 8, and the counterpart of the corresponding data above is

$$n=1 \text{ intersects } n=2 \text{ at } \mu_1=\mu_1^{cr}\approx 1.45, \quad E\approx 1.125,$$

$$n=1 \text{ intersects } n=3 \text{ at } \mu_1=\mu_1^{cr}\approx 1.50, \quad E\approx 1.136,$$

$$n=2 \text{ intersects } n=3 \text{ at } \mu_1=\mu_1^{cr}\approx 1.57, \quad E\approx 1.150.$$

Similarly we calculated the quantity (32) and found its minima. Since these curves are similar to those in Fig. 7, we do not present them but suffice to give the data as in the previous case,

$$\mu_1=\mu_1^{min}\approx 1.20, \quad F\approx 0.9\times 10^{-6}, \quad E\approx 1.056 \quad \text{for } n=1,$$

$$\mu_1=\mu_1^{min}\approx 1.55, \quad F\approx 0.4\times 10^{-5}, \quad E\approx 1.145 \quad \text{for } n=2,$$

$$\mu_1=\mu_1^{min}\approx 1.57, \quad F\approx 0.3\times 10^{-5}, \quad E\approx 1.150 \quad \text{for } n=3.$$

The above data implies the following.

(i) The energies *per unit charge* of the n-vortex solutions for the models characterized by the critical values of the parameters μ_0^{cr} and μ_1^{cr}, or the models at the crossing points of the Jacobs–Rebbi plots, are all close to the topological lower bound 1.0. This departure increases with the vorticity n.

(ii) The quantity F, (32), which is a measure of the departure of the vortices from the strictly noninteracting self-dual case, has minimum values which are very close to *zero* for some values of the coupling contants μ_0 and μ_1. In these regions of the parameters, the systems admit solutions describing ''almost noninteracting'' vortices. This departure also increases with increasing n.

The conclusion is then that for certain values of the coupling constants, the extended models while not admitting strictly noninteracting self-dual solutions, nevertheless support ''almost self-dual'' vortices. This is because for these values of the coupling constants the vortices are almost noninteracting. The energies of these ''almost self-dual'' vortices, while not exactly coincident with, are close to, the energies of the vortices of the systems characterized by the critical values of

the parameters μ_0^{cr} and μ_1^{cr}. It should be pointed out in addition that the critical values of the coupling constants μ^{cr} are fairly close to those values of the coupling constants μ^{min}, for which the quantity F pertaining to $n=2$ and $n=3$ vortices is minimal.

V. SUMMARY

We have calculated the interaction energy of the vortices of the generalized Abelian Higgs models, referred to as the hierarchy of MH models the first member of which is the usual Abelian Higgs model. We find that the vortices of all members of this hierarchy exhibit both an attractive and a repulsive phase, as was shown for the usual case by Jacobs and Rebbi.[9] We have shown, moreover, that the strengths of these interactions become progressively smaller in the case of higher members of the hierarchy. Our numerical demonstration is restricted to the comparative study of the first three members of the hierarchy, which we present as a demonstration by induction of this property for all members of the hierarchy. It is hoped that this quantitative feature of the generalized MH vortices may be of some interest in various areas of vortex physics.

We have also presented a hybrid model consisting of the sum of several members of the hierarchy of MH models, *always including the first member*. The inclusion of the usual Abelian Higgs model endows the photon in this model with the usual propagation properties. This is the physical motivation for introducing these extended models. The solitons of these extended models have nonzero stress tensor because they are not self-dual. It is found that the energies of such solitons are within *a few percent* of the topological lower bound, or the energy of the putative self-dual soliton, this departure growing with the vorticity of the soliton. Most importantly, we have demonstrated that the vortices of these extended models exhibit both attractive and repulsive phases like the individual members of the MH hierarchy, in spite of there being no self-dual solutions to these systems for any values of the parameters in the theory. This has been done for two-vortex and three-vortex systems.

An important and interesting feature of these extended models is that for certain values of the coupling constants the solutions are very close indeed to self-dual field configurations describing noninteracting systems, and that these points in parameter space are close but not coincident with the critical values of the parameters which demarcate the attractive and repulsive phases. In this sense, these field configurations are *almost* self-dual and describe *almost* noninteracting systems of vortices. This analysis been performed for systems of up to three vortices, but it can be extended to an arbitrary number.

ACKNOWLEDGMENTS

This work was carried out under Project No. IC/97/019 of FORBAIRT.

[1] J. Burzlaff, A. Chakrabarti, and D. H. Tchrakian, J. Phys. A **27**, 1617 (1994).
[2] D. H. Tchrakian, J. Math. Phys. **21**, 166 (1980); D. H. Tchrakian, Phys. Lett. B **150**, 360 (1985); D. H. Tchrakian, ''Yang–Mills Hierarchy,'' Int. J. Mod. Phys. A (Proc. Suppl.) 3 A, 584 (1993).
[3] Z.-Qi Ma and D. H. Tchrakian, Lett. Math. Phys. **26**, 111 (1992).
[4] See A. Jaffe and C. J. Taubes, *Monopoles and Vortices* (Birkhäuser, Basel, 1980).
[5] D. H. Tchrakian, ''Skyrme-like models in gauge theory,'' in *Constraint Theory and Quantisation Methods,* edited by F. Colomo, L. Lusanna, and G. Marmo (World Scientific, Singapore).
[6] J. Burzlaff, A. Chakrabarti, and D. H. Tchrakian, Phys. Lett. B **293**, 127 (1992).
[7] D. H. Tchrakian and Yisong Yang, Lett. Math. Phys. **36**, 403 (1996).
[8] J. Burzlaff and D. H. Tchrakian, J. Math. Phys. **37**, 650 (1996).
[9] L. Jacobs and C. Rebbi, Phys. Rev. B **19**, 4486 (1979).
[10] H. J. W. Müller-Kirsten and D. H. Tchrakian, Phys. Rev. D **44**, 1204 (1991).
[11] See K. Arthur and D. H. Tchrakian, J. Math. Phys. **36**, 6566 (1995), and references therein.
[12] T. W. B. Kibble, Phys. Rep. **69**, 83 (1980).

A relativistic supersymmetry and some algebraic consequences

Ion I. Cotăescu
The West University of Timişoara, V. Pârvan Ave. 4, RO-1900 Timişoara, Romania

(Received 15 August 1997; accepted for publication 23 February 1998)

We analyze the supersymmetry and the shape invariance of the potentials which appear in the Klein–Gordon equations of the $(1+1)$ relativistic Pöschl–Teller oscillators we have recently proposed. We point out how the main operators depend on the value of the unique parameter involved in the shape invariance of the supersymmetric partner potentials, and we propose a unitary transformation that changes its value. © *1998 American Institute of Physics.*
[S0022-2488(98)01306-1]

I. INTRODUCTION

The problem of relativistic generalization of the isotropic harmonic oscillator has been discussed by many authors.[1] In the context of general relativity, this has been defined as a free (test) particle on an anti-de-Sitter static background.[2,3] Recently, we have generalized the $(3+1)$ anti-de-Sitter metric of the bag model of Ref. 2 to a family of metrics depending on a real parameter.[4] These generate analytically solvable quantum models which have as nonrelativistic limit just the nonrelativistic harmonic oscillator (NRHO). All these models have interesting energy spectra[5] which could allow one to identify new specific relativistic effects. However, to this end, we have to solve many sensitive problems starting with that of the correct definition of observables up to those concerning the properties of the whole operator algebra. A guide in this direction could be the study of another family of simpler models, obtained as the restriction to $(1+1)$ dimensions of the mentioned ones.[6]

These have the advantage of the special natural frames in which the metrics are conformally flat and, consequently, the relativistic scalar product coincides with the usual one.[7] In these frames the models with countable energy spectra of our family of $(1+1)$ models appear as the relativistic correspondents of the well-known nonrelativistic Pöschl–Teller systems.[8,9] They will be referred as relativistic Pöschl–Teller models (RPTMs). In Ref. 7 we have studied the algebraic properties of their main operators including those of the supersymmetry[10] but without pointing out the mechanisms of supersymmetry and shape invariance of the RPTM. For this reason we consider this problem in the present article. Moreover, we analyze the relation between the algebras of the RPTM with supersymmetric partner (superpartner) potentials with the hope of obtaining a complete image about the algebraic properties of this class of models.

We start in Sec. II with a short review of our previous results concerning the RPTM for which we choose a special parametrization with a unique variable parameter which should determine the mass of the test particle while the background remains unchanged. In Sec. III, we show that the mechanisms of supersymmetry work like in the nonrelativistic case, involving only the mentioned parameter in the shape invariance of the superpartner relativistic potentials. This leads to the conclusion that these correspond to a pair of RPTMs with particles of different masses but the same background. The problem of the structure and the properties of the operators we need in our theory is treated in Sec. IV, where we identify the main operators which depend on the variable parameter, while in Sec. V we introduce transformations able to change its value.

Here we shall use the traditional wave function formalism of the coordinate representations since, in our program, the standard Dirac notation is reserved for the Fock space of the quantum field theory. However, in order to make precise the definitions of several operators introduced here, we give in the Appendix their spectral representation in Dirac notation.

We shall work in natural units with $\hbar = c = 1$.

 © 1998 American Institute of Physics

II. THE PARAMETRIZATION OF THE MODELS

Let us start with a short review of the main results concerning the RPTM. We have shown[7] that there are holonomic frames, (t,x), in which the metrics of our family of models[6] are given by the line elements

$$ds^2=\left(1+\frac{1}{\epsilon^2}\tan^2\hat{\omega}x\right)(dt^2-dx^2), \tag{1}$$

where we have denoted $\hat{\omega}=\epsilon\omega$ understanding that ω is the frequency of the harmonic oscillator obtained as the nonrelativistic limit of these models. Here ϵ is a new parameter that determines the domain of the quantum free motion, $D=(-\pi/2\hat{\omega},\pi/2\hat{\omega})$, for a fixed ω. This has been introduced so that the exact $(1+1)$ anti-de Sitter metric should have $\epsilon=1$. We must specify that in the frame (t,x) the scalar product of the functions defined on D is just the usual one.[7] This means that here the definition of the Hermitian conjugation is also the usual one (e.g., $\partial_x^+=-\partial_x$).

The Klein–Gordon equation of a test particle of the mass m and energy E has the form

$$\left(-\frac{d^2}{dx^2}+\frac{m^2}{\epsilon^2}\tan^2\hat{\omega}x\right)U(x)=(E^2-m^2)U(x). \tag{2}$$

It has a countable discrete energy spectrum, the corresponding square-integrable energy eigenfunctions of which are

$$U_{n_s,s}(x)=N_{n_s,s}\cos^k\hat{\omega}x\,\sin^s\hat{\omega}xF(-n_s,k+s+n_s,s+\tfrac{1}{2},\sin^2\hat{\omega}x), \tag{3}$$

where $N_{n_s,s}$ is the normalization factor[7] and

$$k=\frac{1}{2}\left[1+\sqrt{1+4\frac{m^2}{\epsilon^2\hat{\omega}^2}}\right]>1 \tag{4}$$

is the positive solution of the equation

$$k(k-1)=\frac{m^2}{\epsilon^2\hat{\omega}^2}. \tag{5}$$

The quantum numbers, $n_s=0,1,2,...$ and $s=0,1$, can be embedded into the main quantum number $n=2n_s+s$. This will take even values if $s=0$ and odd values for $s=1$. Hence, the functions $U_{n_s,s}(x)$ are completely determined by n only. They are real polynomials of the degree n in $\sin\hat{\omega}x$, with the factor $\cos^k\hat{\omega}x$, which vanishes at $x=\pm\pi/2\hat{\omega}$. The energy levels also depend on the main quantum number. These can be obtained from the quantization condition

$$E_n^2-m^2\left(1-\frac{1}{\epsilon^2}\right)=\hat{\omega}^2(n+k)^2,\quad n=0,1,2... . \tag{6}$$

We specify that, for the pure anti-de-Sitter oscillator, the resulting energy spectrum is equidistant[3] since $\epsilon=1$.

All these models depend only on three parameters, the mass m, and the parameters which determine the metric, i.e., the frequency ω and the new parameter ϵ. In the following we shall consider that the background remains unchanged (which means that the last two parameters are fixed) and we shall study the free motion of test particles with different masses on this background. In these conditions it is convenient to use the parameter k instead of m. This will be the unique arbitrary parameter giving the properties of a RPTM which will be denoted from now by (k). The mass of its test particle results from (5),

$$m(k)^2=k(k-1)\epsilon^2\hat{\omega}^2, \tag{7}$$

while the energy spectrum, defined by the quantization condition (6), has the energy levels

$$E_n(k)^2=\hat{\omega}^2[(n+k)^2+(\epsilon^2-1)k(k-1)]. \tag{8}$$

Moreover, the second term on the left-hand side of Eq. (2), which is the relativistic potential,[7] can be put in the form

$$V_{PT}(k,x)=k(k-1)\hat{\omega}^2\ \tan^2\ \hat{\omega}x. \tag{9}$$

Based on the previous observation that the eigenfunctions (3) are defined by the main quantum number only, we can denote $U_{n_s,s}\equiv U_{k,n}$. For each $k>1$ the orthonormal and complete set $U_{k,n}, n=0,1,2,...$, represents a basis of the Hilbert state space, $L^2(D)$, of the square integrable functions defined on D. This is the energy basis of the model (k). It is important to specify that all the models (with $k>1$) have the same Hilbert space since the domain D depends only on ω and ϵ, which remain fixed.

III. SUPERSYMMETRY AND SHAPE INVARIANCE

We can construct the supersymmetric quantum mechanics of our RPTM by using adjusted Klein–Gordon operators instead of the Hamiltonians of the nonrelativistic theory. These must admit as eigenvalues the differences

$$d_{k,n}=E_n(k)^2-E_0(k)^2=\hat{\omega}^2n(n+2k), \tag{10}$$

which satisfy $d_{k,0}=0$. Thus, we obtain the new operator $\mathbf{\Delta}[V]$ defined by

$$\{\mathbf{\Delta}[V]U\}(x)=\left(-\frac{d^2}{dx^2}+V(x)\right)U(x), \tag{11}$$

which allows us to rewrite the Klein–Gordon equation in the form

$$\mathbf{\Delta}[V_-(k)]U_{k,n}=d_{k,n}U_{k,n} \tag{12}$$

where we have denoted

$$V_-(k,x)=V_{PT}(k,x)+m(k)^2-E_0(k)^2=\hat{\omega}^2[k(k-1)\ \tan^2\ \hat{\omega}x-k]. \tag{13}$$

Furthermore, from the (normalized[11]) ground-state eigenfunction,

$$U_{k,0}(x)=\left(\frac{\hat{\omega}^2}{\pi}\right)^{1/4}\left[\frac{\Gamma(k+1)}{\Gamma(k+\frac{1}{2})}\right]^{1/2}\ \cos^k\ \hat{\omega}x, \tag{14}$$

we obtain the superpotential,

$$W(k,x)=-\frac{1}{U_{k,0}(x)}\frac{dU_{k,0}(x)}{dx}=\hat{\omega}k\ \tan\ \hat{\omega}x, \tag{15}$$

and the superpartner of V_-,

$$V_+(k,x)=-V_-(k,x)+2W(k,x)^2=\hat{\omega}^2[k(k+1)\ \tan^2\ \hat{\omega}x+k]. \tag{16}$$

According to the standard procedure,[10] we define the pair of adjoint operators, $\mathbf{A}_k$ and $\mathbf{A}_k^+$, having the action

$$(\mathbf{A}_kU)(x)=\left(\frac{d}{dx}+W(k,x)\right)U(x), \tag{17}$$

$$(\mathbf{A}_k^+U)(x)=\left(-\frac{d}{dx}+W(k,x)\right)U(x), \tag{18}$$

and the commutation rule

$$[\mathbf{A}_k,\ \mathbf{A}_k^+]=2k\hat{\omega}^2\mathbf{1}+\frac{1}{2k}\,(\mathbf{A}_k+\mathbf{A}_k^+)^2, \tag{19}$$

where $\mathbf{1}$ is the unit operator. Then, the operators $\mathbf{\Delta}$ of the models with superpartner potentials can be written as

$$\mathbf{\Delta}[V_-(k)]=\mathbf{A}_k^+\mathbf{A}_k,\quad \mathbf{\Delta}[V_+(k)]=\mathbf{A}_k\mathbf{A}_k^+. \tag{20}$$

Other algebraic properties of the operator $\mathbf{A}_k$ can be derived starting with the observation that it is proportional with the operator $\tilde{\mathbf{a}}_k$ we have studied in Ref. 7 (i.e., $\mathbf{A}_k=\hat{\omega}\sqrt{2k}\tilde{\mathbf{a}}_k$).

Let us observe now that the potentials $V_-(k)$ and $V_+(k)$ are shape invariant since

$$V_+(k,x)=V_-(k+1,x)+(2k+1)\hat{\omega}^2. \tag{21}$$

Consequently, we can verify that

$$\mathbf{\Delta}[V_+(k)]U_{k+1,n-1}=d_{k,n}U_{k+1,n-1},\quad n=1,2,..., \tag{22}$$

which means that the spectrum of $\mathbf{\Delta}[V_+(k)]$ coincides with that of $\mathbf{\Delta}[V_-(k)]$, apart from the lowest eigenvalue of the last one, $d_{k,0}=0$. Furthermore, from (22) combined with (12) it results that the normalized energy eigenfunctions satisfy

$$\mathbf{A}_kU_{k,n}=\sqrt{d_{k,n}}U_{k+1,n-1},\quad \mathbf{A}_k^+U_{k+1,n-1}=\sqrt{d_{k,n}}U_{k,n}, \tag{23}$$

for $n>0$, while for $n=0$ we have $\mathbf{A}_kU_{k,0}=0$, according to (15) and (17). Hence, we have obtained the desired relation between the energy basis of the models (k) and $(k+1)$ with the superpartner potentials $V_-(k)$ and $V_+(k)$, respectively. In general, any normalized energy eigenfunction of the model (k) can be written as

$$U_{k,n}=\frac{1}{\hat{\omega}^n\sqrt{n!}}\left[\frac{\Gamma(n+2k)}{\Gamma(2n+2k)}\right]^{1/2}\mathbf{A}_k^+\mathbf{A}_{k+1}^+\cdots\mathbf{A}_{k+n-1}^+U_{k+n,0}, \tag{24}$$

where $U_{k+n,0}$ is the normalized ground-state eigenfunction of the model $(k+n)$, given by (14).

A special case is that of $\epsilon=0$ when the energy eigenfunction, $U_n^{(0)}$, of the relativistic model coincides with those of the NRHO.[7] This is because, in the limit of $\epsilon\to 0$, when $\hat{\omega}\to 0$, $k\to\infty$, but $\epsilon^2k\to m/\omega$, we have

$$\lim_{\epsilon\to 0}V_{PT}(k,x)=\lim_{\epsilon\to 0}V_-(k,x)=\lim_{\epsilon\to 0}V_+(k,x)=m^2\omega^2x^2 \tag{25}$$

and

$$\lim_{\epsilon\to 0}\mathbf{A}_k=\sqrt{2m\omega}\mathbf{a}, \tag{26}$$

where $\mathbf{a}$ is the usual lowering differential operator of the NRHO. Moreover, since $\Gamma(z+a)/\Gamma(z+b)\sim z^{(a-b)}$ for very large z,[11] we can verify that for $\epsilon=0$ Eq. (24) takes the usual form, $U_n^{(0)}=(\mathbf{a}^+)^nU_0^{(0)}/\sqrt{n!}$.

IV. THE OPERATOR ALGEBRA

The operator algebra of our RPTM contains operators independent on k as well as specific operators which depend on this parameter. In order to analyze this dependence it would be useful to introduce an operator with the eigenvalues k, but, in our opinion, this cannot be done. The argument is that the wave functions $U_{k,n}$ and $U_{k',n}$ generally are not orthogonal for $k'\neq k$ and, consequently, they could not represent eigenvectors of a self-adjoint operator. In these conditions

we are forced to consider that the specific operators of these models[7] are functions of k. In the following we restrict ourselves to the precise structure of the most important of them, without entering into other details.

The differential operators of the coordinate representation generally do not depend on k if they are not explicitly defined as functions of k. For example, the coordinate and momentum operators, $\mathbf{X}\sim x$ and $\mathbf{P}\sim i\partial_x$, are the same for all the models (k). In other respects, the main differential operators of the supersymmetry, $\mathbf{A}_k$ and $\mathbf{A}_k^+$, depend on k according to (15) and (17) from which it results that

$$\mathbf{A}_k=-i(\cos\ \hat{\omega}\mathbf{X})^k\mathbf{P}(\cos\ \hat{\omega}\mathbf{X})^{-k}. \tag{27}$$

However, the nondifferential operators can be defined only by giving their action in the energy basis of a model (k) and, therefore, all of them depend on k. In general, they are analytic functions of the operators $\mathbf{a}_k$ and $\mathbf{a}_k^+$ (with $[\mathbf{a}_k,\ \mathbf{a}_k^+]=\mathbf{1}$) of a unitary representation of the standard oscillator (Heisenberg–Weyl) algebra defined directly in the energy basis by

$$\mathbf{a}_kU_{k,n}=\sqrt{n}U_{k,n-1},\quad \mathbf{a}_k^+U_{k,n}=\sqrt{n+1}U_{k,n+1}, \tag{28}$$

or by (A3), so that $\mathbf{N}_k=\mathbf{a}_k^+\mathbf{a}_k$ should be the operator of number of quanta of the model (k). As mentioned, for $\epsilon=0$ the energy basis coincides with that of the NRHO and, consequently,

$$\lim_{\epsilon\to 0}\ \mathbf{a}_k=\mathbf{a}. \tag{29}$$

The above introduced operators allow one to put the wave equations of the RPTM with superpartner potentials, (12) and (22), in the operator form

$$\mathbf{A}_k^+\mathbf{A}_k=\hat{\omega}^2\mathbf{N}_k(\mathbf{N}_k+2k\mathbf{1}), \tag{30}$$

$$\mathbf{A}_k\mathbf{A}_k^+=\hat{\omega}^2(\mathbf{N}_{k+1}+\mathbf{1})[\mathbf{N}_{k+1}+(2k+1)\mathbf{1}], \tag{31}$$

which is in accordance with (10).

Each model (k) has its own pair of shift (i.e., lowering or raising) operators of the energy basis,[9]

$$\mathbf{A}_{k,(-)}=\frac{1}{\hat{\omega}\sqrt{2k}}[(\cos\ \hat{\omega}\mathbf{X})\mathbf{A}_k+\hat{\omega}(\sin\ \hat{\omega}\mathbf{X})\mathbf{N}_k], \tag{32}$$

$$\mathbf{A}_{k,(+)}=\frac{1}{\hat{\omega}\sqrt{2k}}[(\cos\ \hat{\omega}\mathbf{X})\mathbf{A}_k^++\hat{\omega}(\sin\ \hat{\omega}\mathbf{X})\mathbf{N}_k], \tag{33}$$

which decrease or increase the main quantum number with one unit, keeping the value of k unchanged. We have shown[7] that the operators $\sqrt{2k}\mathbf{A}_{k,(-)}$, $\sqrt{2k}\mathbf{A}_{k,(+)}$, and $\mathbf{N}_k+k\mathbf{1}$ are the generators of a nonunitary representation of the so(1,2) algebra which is equivalent with the unitary discrete representation of the minimal weight k generated by

$$\begin{aligned}\mathbf{K}_{k,(+)}&=\mathbf{a}_k^+(\mathbf{N}_k+2k\mathbf{1})^{1/2},\\ \mathbf{K}_{k,(-)}&=(\mathbf{N}_k+2k\mathbf{1})^{1/2}\mathbf{a}_k,\\ \mathbf{K}_{k,3}&=\mathbf{N}_k+k\mathbf{1}.\end{aligned} \tag{34}$$

Notice that, when $\epsilon\to 0$, this algebra degenerates in the standard one of the NRHO[7] (i.e., $\mathbf{K}_{k,(+)}/\sqrt{2k}\to\mathbf{a}^+$, $\mathbf{K}_{k,(-)}/\sqrt{2k}\to\mathbf{a}$, and $\mathbf{K}_{k,3}/k\to\mathbf{1}$).

V. *k*-TRANSITIONS

In the supersymmetric quantum mechanics we have to work simultaneously with two or more different RPTMs, each one having its own specific algebra. For this reason, it is useful to introduce new operators which help us to transform the algebra of a given model into that of another one, performing the change of the value of the parameter k. These transformations will be called k-transitions.

They can be correctly defined since the energy basis of different RPTMs are different orthonormal basis of the *same* Hilbert space. This means that there are unitary operators, $\mathbf{U}(k',k)$, which transform between themselves the basis of two different models, (k) and (k'), as

$$U_{k',n}=\mathbf{U}(k',k)U_{k,n}, \quad n=0,1,2,\ldots . \tag{35}$$

Obviously, these operators may have properties similar to those of the time evolution operators, i.e.,

$$\mathbf{U}(k'',k')\mathbf{U}(k',k)=\mathbf{U}(k'',k), \quad \mathbf{U}(k,k)=\mathbf{1}, \tag{36}$$

$$\mathbf{U}(k',k)^{+}=\mathbf{U}(k',k)^{-1}=\mathbf{U}(k,k'). \tag{37}$$

In the limit of $\epsilon\to 0$ (when $k\to\infty$) we have

$$\lim_{\epsilon\to 0}\mathbf{U}(k,k')=\mathbf{1}. \tag{38}$$

Let us observe that the k-transition is, in fact, the unitary transformation between the equivalent representations (28) of the Heisenberg–Weyl algebra of the models (k) and (k'). This has the usual form

$$\mathbf{a}_{k'}=\mathbf{U}(k',k)\mathbf{a}_k\mathbf{U}(k,k'). \tag{39}$$

On the other hand, for the unitary representations (34) of the so(1,2) algebra, which are not equivalent, we have a more complicated transformation rule, namely

$$\begin{aligned}\mathbf{K}_{k',(+)}&=(\mathbf{K}_{k',(-)})^{+}=\mathbf{U}(k',k)\mathbf{K}_{k,(+)}\left(\frac{\mathbf{N}_k+2k'\mathbf{1}}{\mathbf{N}_k+2k\mathbf{1}}\right)^{1/2}\mathbf{U}(k,k')\\ \mathbf{K}_{k',3}&=\mathbf{U}(k',k)\mathbf{K}_{k,3}\mathbf{U}(k,k')+(k'-k)\mathbf{1}.\end{aligned} \tag{40}$$

However, the major advantage of introducing the unitary operators of the k-transitions is to point out that the differential operator $\mathbf{A}_k$ can be put in the form

$$\mathbf{A}_k=\hat{\omega}\mathbf{U}(k+1,k)\mathbf{a}_k(\mathbf{N}_k+2k\mathbf{1})^{1/2}, \tag{41}$$

and transforms like

$$\mathbf{A}_{k'}=\mathbf{U}(k'+1,k+1)\mathbf{A}_k\left(\frac{\mathbf{N}_k+2k'\mathbf{1}}{\mathbf{N}_k+2k\mathbf{1}}\right)^{1/2}\mathbf{U}(k,k'). \tag{42}$$

Its spectral representation as well as that of $\mathbf{U}$ are given in the Appendix.

VI. COMMENTS

As a conclusion, we can say that our example of relativistic supersymmetric quantum mechanics of a massive scalar particle follows the same general rules as in the nonrelativistic case. The major difference is that the Hamiltonian of the nonrelativistic theory is replaced here by the operator $\mathbf{\Delta}$, which is linearly dependent on the squared energy operator.[7] For this reason, the role of the energies of the nonrelativistic theory of supersymmetry is played now by the differences (10). Moreover, the homogeneous structure of the Klein–Gordon equation induces some kind of regularities which lead to a very simple parametrization of the theory in such a manner that for any

pair of models with superpartner potentials we have $\Delta k=1$. Thus k simulates the behavior of a quantum number even though it cannot be the eigenvalue of a self-adjoint operator.

On the other hand, we have seen that in this parametrization we can easily identify the operators which explicitly depend on k and introduce the unitary operators of k-transitions. Thus we have obtained a large and complicated algebra which seems to have new and interesting properties. Of course, there are many fundamental problems or simple technical details to be solved before drawing definitive conclusions. However, in our opinion, no major difficulties could arise therein since, fortunately, we know the spectral representations of all the above-defined operators in the energy basis where we have the form of the normalized energy eigenfunctions. This means that we can calculate all the (usual or generalized) matrix elements of these operators, either in one of the energy bases or in the coordinate representation, thus having the possibility to find their main algebraic properties.

APPENDIX: SPECTRAL REPRESENTATIONS

In the standard Dirac notation, where $U_{k,n}(x)=\langle x|k,n\rangle$, the energy basis of the model (k) is $\{|k,n\rangle|n=0,1,2,...\}$. It is orthonormal,

$$\langle k,n|k,n'\rangle=\delta_{n,n'}, \tag{A1}$$

and complete,

$$\mathbf{1}=\sum_{n=0}^{\infty}|k,n\rangle\langle k,n|. \tag{A2}$$

According to our previous definitions, we can write the following spectral representations,

$$\mathbf{a}_k=\sum_{n=1}^{\infty}|k,n-1\rangle\sqrt{n}\langle k,n|, \tag{A3}$$

$$\mathbf{A}_k=\sum_{n=1}^{\infty}|k+1,n-1\rangle\sqrt{d_{k,n}}\langle k,n|, \tag{A4}$$

$$\mathbf{U}(k',k)=\sum_{n=0}^{\infty}|k',n\rangle\langle k,n|, \tag{A5}$$

from which we can calculate the matrix elements in one of the energy bases or the generalized matrix elements (i.e., the kernels) in the coordinate representation. For example, herein we have

$$\langle x|\mathbf{U}(k',k)|y\rangle=\sum_{n=0}^{\infty}U^{*}_{k',n}(x)U_{k,n}(y). \tag{A6}$$

[1] H. Yukawa, Phys. Rev. **91**, 416 (1953); R. P. Feynman, M. Kislinger, and F. Ravndal, Phys. Rev. D **3**, 2706 (1971); Y. S. Kim and M. E. Noz, Am. J. Phys. **46**, 480 (1978); Y. S. Kim and E. P. Wigner, Phys. Rev. A **38**, 1159 (1988); D. Itô, K. Mori, and E. Carrieri, Nuovo Cimento A **51**, 1119 (1967); M. Moshinsky and A. Szczepaniak, J. Phys. A **22**, L817 (1989); V. Aldaya, J. Bisquert, R. Loll, and J. Navarro-Salas, J. Math. Phys. **33**, 3087 (1992).
[2] C. Dullemond and E. van Beveren, Phys. Rev. D **28**, 1028 (1983).
[3] D. J. Navarro and J. Navarro-Salas, J. Math. Phys. **37**, 6006 (1996).
[4] I. I. Cotăescu, Mod. Phys. Lett. A **12**, 685 (1997).
[5] G. Ardelean and I. I. Cotăescu, Int. J. Mod. Phys. C **8**, 1237 (1997).
[6] I. I. Cotăescu, Int. J. Mod. Phys. A **12**, 3545 (1997).
[7] I. I. Cotăescu and G. Drăgănescu, J. Math. Phys. **38**, 5505 (1997).
[8] G. Pöschl and E. Teller, Z. Phys. **83**, 149 (1933).
[9] M. M. Nieto and L. M. Simmons, Jr., Phys. Rev. Lett. **41**, 207 (1978); Phys. Rev. D **20**, 1332 (1979).
[10] R. Dutt, A. Khare, and U. P. Sukhatme, Am. J. Phys. **56**, 163 (1987); F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **251**, 267 (1995).
[11] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).

Strong asymptotics of Laguerre polynomials and information entropies of two-dimensional harmonic oscillator and one-dimensional Coulomb potentials

J. S. Dehesa[a)]
Departamento de Física Moderna and Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071—Granada, Spain

R. J. Yáñez[b)]
Departamento de Matemática Aplicada and Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071—Granada, Spain

A. I. Aptekarev[c)] and V. Buyarov
Keldysh Institute of Applied Mathematics (Russian Academy of Sciences) and Faculty of Mechanics and Mathematics, Moscow State University, Moscow, Russia

(Received 29 October 1997; accepted for publication 7 January 1998)

The information entropies of the two-dimensional harmonic oscillator, $V(x,y)=1/2\lambda(x^2+y^2)$, and the one-dimensional hydrogen atom, $V(x)=-1/|x|$, can be expressed by means of some entropy integrals of Laguerre polynomials whose values have not yet been analytically determined. Here, we first study the asymptotical behavior of these integrals in detail by extensive use of strong asymptotics of Laguerre polynomials. Then, this result (which is also important by itself in a context of both approximation theory and potential theory) is employed to analyze the information entropies of the aforementioned quantum-mechanical potentials for the very excited states in both position and momentum spaces. It is observed, in particular, that the sum of position and momentum entropies has a logarithmic growth with respect to the main quantum number which characterizes the corresponding physical state. Finally, the rate of convergence of the entropies is numerically examined. © *1998 American Institute of Physics.* [S0022-2488(98)00806-8]

I. INTRODUCTION

The two-dimensional $(2D)$ harmonic oscillator $V(\vec{r})=1/2\lambda r^2$, $\vec{r}=(x,y)$, and the one-dimensional $(1D)$ hydrogen atom $V(x)=-1/|x|$, are two quantum-mechanical potentials currently used to model and explain in a simple way various natural phenomena of optical and condensed-matter character.[1–3] Both systems received a great deal of physico-mathematical attention[4–6] because they are prototypes of numerous physical situations.

However, only recently[6] the rigorous study of the entropies of these two systems has been initiated from a technical point of view, i.e., in the sense of information theory.[7–9] The starting point is Born's probability density, which for a single-particle system is $\rho(\vec{r})=|\Psi(\vec{r})|^2$ in the position space and $\gamma(\vec{p})=|\tilde{\Psi}(\vec{p})|^2$ in the momentum space. The symbols $\Psi(\vec{r})$ and its Fourier transform $\tilde{\Psi}(\vec{p})$ denote the position and momentum eigenfunctions of the corresponding potential.

It is well known that the normalized-to-unity position and momentum densities are, respectively,

$$\rho(\vec{r})=\frac{1}{\pi}\,\lambda^{|m|+1}r^{2|m|}e^{-\lambda r^2}[\hat{L}_n^{|m|}(\lambda r^2)]^2, \tag{1a}$$

$$\gamma(\vec{p})=\frac{1}{\lambda^2}\,\rho\left(\frac{\vec{p}}{\lambda}\right); \quad n=0,1,2,..., \tag{1b}$$

a)Electronic mail: dehesa@ugr.es
b)Electronic mail: ryanez@ugr.es
c)Electronic mail: aptekaa@rfbr.ru

 © 1998 American Institute of Physics

for the $2D$-harmonic oscillator[6,10] in the spherical coordinates (r,ϕ) and

$$\rho(x)=\frac{2}{n^4}\,e^{-(2/n)|x|}x^2\left[\hat{L}^1_{n-1}\left(\frac{2}{n}\,|x|\right)\right]^2, \tag{2a}$$

$$\gamma(p)=\frac{2n}{\pi}\,\frac{1}{(1+n^2p^2)^2};\quad n=1,2,..., \tag{2b}$$

for the $1D$ hydrogen atom, where $\hat{L}^\alpha_n(x)$ denotes an orthonormal Laguerre polynomial. Note that the physical densities are characterized by the integer quantum numbers $n=1,2,...,\infty$ and, in the oscillator case, $|m|$ correspondingly to its number of degrees of freedom. The corresponding bound-state energies $\mathscr{E}_n$ of both systems are

$$\mathscr{E}_n=\begin{cases}\lambda(2n+|m|+1), & \text{for the } 2D \text{ oscillator,}\\ -\dfrac{1}{n^2}, & \text{for the } 1D \text{ hydrogen atom.}\end{cases}$$

The lowest (unphysical) state of the hydrogen system with a density $\rho(x)=\delta(x)$ will not be considered here.

The Boltzmann–Shannon information entropy, which measures the spread or extent of the probability density and is closely related to the concept of entropy and disorder in thermodynamics, is defined as[7,11]

$$S_\rho=-\int \rho(\vec{r})\ln\,\rho(\vec{r})d\vec{r} \tag{3}$$

in the position space, and, corresponding,

$$S_\gamma=-\int \gamma(\vec{p})\ln\,\gamma(\vec{p})d\vec{p} \tag{4}$$

in the momentum space. For a D-dimensional single-particle system, these quantities are mutually related by means of the Bialynicki-Birula and Mycielski[12] inequality,

$$S_\rho+S_\gamma\geqslant D\,\ln(e\pi), \tag{5}$$

which is a stronger version of the Heisenberg uncertainty relation.

It has been recently shown[6] that the position and momentum entropies of the two aforementioned quantum-mechanical systems can be expressed by means of the entropy-like functionals of the following general type;

$$S_n(\alpha,\beta):=\int_0^\infty x^\beta\omega_\alpha(x)[\hat{L}^\alpha_n(x)]^2\,\ln[\hat{L}^\alpha_n(x)]^2dx,\quad \beta\geqslant 0, \tag{6}$$

with $\omega_\alpha(x)=x^\alpha e^{-x}$. Notice that this reduces to

$$E_n(\alpha):=-S_n(\alpha,0)=-\int_0^\infty \omega_\alpha(x)[\hat{L}^\alpha_n(x)]^2\,\ln[\hat{L}^\alpha_n(x)]^2dx, \tag{7}$$

for $\beta=0$, and to

$$F_n(\alpha):=-S_n(\alpha,1)=-\int_0^\infty x\omega_\alpha(x)[\hat{L}^\alpha_n(x)]^2\,\ln[\hat{L}^\alpha_n(x)]^2dx, \tag{8}$$

for $\beta=1$. The quantity $E_n(\alpha)$ is referred to as ''entropy for the orthonormal Laguerre polynomial $\hat{L}^\alpha_n(x)$'' or simply as ''entropy of Laguerre polynomials'';[7,13] it plays a very important role in the determination of the information entropies of exactly and quasi-exactly solvable potentials.

The exact evaluation of these entropy-like functionals is a very difficult problem, not yet solved. Here we will attack this problem by means of the determination of the asymptotical behavior (i.e., at large n) of the general functional $S_n(\alpha,\beta)$ by extensive use of the strong asymptotics of Laguerre polynomials, as discussed in detail in Section II. Then, in Section III, the asymptotics of the Laguerre entropy $E_n(|m|)$ is used to evaluate the information entropies of the $2D$-harmonic oscillator in both position and momentum spaces for very excited states. Third, in Section IV, the asymptotical value of the entropic integral $F_n(\alpha)$, evaluated at $\alpha=1$, allows us to find the position entropy of the $1D$ hydrogen atom. The rate of convergence of the information entropies of the aforementioned systems is also numerically discussed. Finally, some concluding remarks and the pertinent bibliography are given.

II. ASYMPTOTICS OF ENTROPY-LIKE FUNCTIONALS OF LAGUERRE POLYNOMIALS

Here the asymptotical behavior (i.e., at large n) of the quantity $S_n(\alpha,\beta)$, given by Eq. (6) is analyzed in detail. It is found that

$$\begin{aligned} S_n(\alpha,\beta) &:= \int_0^\infty x^\beta \omega_\alpha(x)[\hat{L}_n^\alpha(x)]^2 \ln[\hat{L}_n^\alpha(x)]^2 dx \\ &= \frac{2^{2\beta+2}\Gamma(\beta+3/2)}{\sqrt{\pi}\Gamma(\beta+2)} n^{\beta+1} - \frac{2^{2\beta}(\alpha+1)\Gamma(\beta+1/2)}{\sqrt{\pi}\Gamma(\beta+1)} n^\beta \ln n \\ &\quad + \frac{2^{2\beta-1}\Gamma(\beta+1/2)}{\sqrt{\pi}\Gamma(\beta+1)} [2(\alpha+1)\psi(\beta+1) - (2\alpha+1)\psi(\beta+1/2) - 2\ln \pi \\ &\quad - 4(\alpha-1)\ln 2 + \gamma_E + 4 + 2(\alpha+2\beta) + 4\alpha\beta] n^\beta + o(n^\beta), \end{aligned} \tag{9}$$

which holds for any real $\alpha > -1$. For the particular cases $\beta=0$ and $\beta=1$, this expression leads straightforwardly to

$$E_n(\alpha) = -2n + (\alpha+1)\ln n - \alpha - 2 + \ln(2\pi) + o(1), \tag{10}$$

$$F_n(\alpha) = -6n^2 + 2(\alpha+1)n \ln n + 2(\ln 2\pi - 2\alpha - 4)n + o(n). \tag{11}$$

To find the main asymptotic formula (9) we use a general method which consists of the three following steps. First, to express $S_n(\alpha,\beta)$ in terms of the quantities of L_p-norm character,

$$N_n(\beta,p) = \int_0^\infty x^\beta |\omega_\alpha^{1/2} \hat{L}_n^\alpha(x)|^{2p} dx. \tag{12}$$

Second, to find the asymptotical value (i.e., at large n) to be denoted by $N_n(\beta,p)$, of this quantity. The latter will be done (i) by substitution of the strong asymptotics of Laguerre polynomials for some integrals of trigonometric functions in the finite interval of Plancherel–Rotach uniform convergence, and (ii) by calculation of the asymptotics of these integrals via the generalized Bernstein–Fejer lemma of Aptekarev *et al.* (Ref. 14, Lemma 2.1). Third, to evaluate the asymptotical value $S(\alpha,\beta)$ of $S_n(\alpha,\beta)$ from the quantity $N(\beta,p)$.

To start with, let us consider the first derivative of $N_n(\beta,p)$ with respect to p. Its value $N_n'(\beta,p)$ at $p=1$ is given by

$$N_n'(\beta,1) \equiv \frac{d}{dp} N_n(\beta,p)\bigg|_{p=1} = \int_0^\infty x^\beta \omega_\alpha(x) q_n^2(x) \ln[\omega_\alpha(x) q_n^2(x)] dx, \tag{13}$$

where the notation $q_n(x) \equiv \hat{L}_n^\alpha(x)$ will be used hereinafter. By keeping in mind the definition (6) of $S_n(\alpha,\beta)$ one has that

$$S_n(\alpha,\beta) = \frac{d}{dp} N_n(\beta,p)\bigg|_{p=1} + \mathscr{T}_1(\beta) - \alpha \mathscr{T}_2(\beta), \tag{14}$$

where the integrals $\mathscr{T}_1$ and $\mathscr{T}_2$ are given by

$$\mathscr{T}_1(\beta)=\int_0^\infty x^{\beta+1}\omega_\alpha(x)q_n^2(x)dx, \tag{15}$$

$$\mathscr{T}_2(\beta)=\int_0^\infty x^{\beta}\ln x\,\omega_\alpha(x)q_n^2(x)dx. \tag{16}$$

Moreover, the definition of $N_n(\beta,p)$ given by Eq. (12) allows us to write

$$\mathscr{T}_1(\beta)=N_n(\beta+1,1), \tag{17}$$

$$\mathscr{T}_2(\beta)=\frac{d}{d\beta}\mathscr{T}_1(\beta-1)=\frac{d}{d\beta}N_n(\beta,1). \tag{18}$$

Then

$$S_n(\alpha,\beta)=\frac{d}{dp}N_n(\beta,p)\bigg|_{p=1}+N_n(\beta+1,1)-\alpha\frac{d}{d\beta}N_n(\beta,1), \tag{19}$$

so that the study of the asymptotics of the entropy-like functional $S_n(\alpha,\beta)$ transforms into the analysis of the asymptotic behavior of the L_p-norm quantities $N_n(\beta,p)$. To do that we will follow the general procedure of Aptekarev *et al.*,[14] where the asymptotics of L_p-norms of orthogonal polynomials on unbounded sets is analyzed in detail. This procedure allows us to show firstly that the main contribution to $N_n(\beta,p)$ is coming from the interval where the Planchere–Rotach formula approximates polynomials uniformly so that

$$N_n(\beta,p)=\int_\epsilon^{x_n(1-\epsilon)}x^\beta|\omega_\alpha^{1/2}(x)q_n(x)|^{2p}dx[1+o(1)], \tag{20}$$

p being in some interval which includes $p=1$; and, second, we use the asymptotic behavior of the Laguerre polynomials supplied by the Plancherel–Rotach formula (Ref. 15, Theorem 8.22.9) which for

$$x_n=4n+2\alpha+2,\quad N=n+\frac{\alpha+1}{2}, \tag{21}$$

and the ϕ-definition $x=x_n\cos^2\phi$ gives that

$$\omega_\alpha^{1/2}(x)q_n(x)=\left[\frac{2}{\pi\sqrt{x(x_n-x)}}\right]^{1/2}\cos\left[N(\sin 2\phi-2\phi)+\frac{\pi}{4}\right]+o(1), \tag{22}$$

uniformly for $\epsilon\leqslant\phi\leqslant\pi/2-\epsilon/\sqrt{n}$. Then, the integral on the right-hand side of Eq. (20) gets reduced so that one obtains

$$N_n(\beta,p)=\int_\epsilon^{x_n(1-\epsilon)}x^\beta\left|\sqrt{2}\cos\left[N(\sin 2\phi-2\phi)+\frac{\pi}{4}\right]\right|^{2p}\frac{dx}{[\pi\sqrt{x(x_n-x)}]^p}[1+o(1)], \tag{23}$$

in a similar way as described in Ref. 14.

Moreover, taking into account that we may put $\epsilon=0$ in Eq. (23) and with the notation $\zeta(\phi)=\sin 2\phi-2\phi$, one has

$$N_n(\beta,p)=x_n^{\beta+1-p}2\left(\frac{2}{\pi}\right)^p\int_0^{\pi/2}(\cos\phi)^{2\beta}\left|\cos\left[N\zeta(\phi)+\frac{\pi}{4}\right]\right|^{2p}|\cos\phi\sin\phi|^{1-p}d\phi[1+o(1)], \tag{24}$$

where the change of variable $x \to \phi$: $x = x_n \cos^2 \phi$ was carried out. Therefore,

$$\lim_{n\to\infty} \frac{N_n(\beta,p)}{x_n^{\beta+1-p}} = \lim_{n\to\infty} \left(\frac{2}{\pi}\right)^p \int_0^{\pi/2} 2(\cos\ \phi)^{2\beta} |\cos\ \phi\ \sin\ \phi|^{1-p} \left|\cos\left[N\zeta(\phi) + \frac{\pi}{4}\right]\right|^{2p} d\phi,$$

where the right-hand side goes to a constant to be calculated. To go ahead, it is convenient to do another change of variable,

$$\phi \to \theta: \phi = \zeta^{-1}(\zeta(\phi)) \equiv \zeta^{-1}(\theta) = \zeta_1(\theta).$$

Then,

$$\lim_{n\to\infty} \frac{N_n(\beta,p)}{x_n^{\beta+1-p}} = \lim_{n\to\infty} \left(\frac{4}{\pi}\right)^p \int_0^{\pi} [\cos\ \zeta_1(\theta)]^{2\beta} |\sin 2\zeta_1(\theta)|^{1-p} \left|\cos\left(N\theta + \frac{\pi}{4}\right)\right|^{2p} \zeta_1'(\theta) d\theta.$$

Now, the use of Lemma 2.1 of Aptekarev *et al.*[15] leads directly to the searched constant,

$$\lim_{n\to\infty} \frac{N_n(\beta,p)}{x_n^{\beta+1-p}} = \left(\frac{4}{\pi}\right)^p \frac{1}{\pi} \int_0^{\pi} [\cos\ \zeta_1(\theta)]^{2\beta} |\sin 2\zeta_1(\theta)|^{1-p} \zeta_1'(\theta) d\theta \cdot \int_0^{\pi} |\cos\ \theta|^{2p} d\theta + o(1).$$

Recovering back the variable ϕ in the first integral on the right-hand side, one can obtain its exact value as

$$\int_0^{\pi/2} [\cos\ \phi]^{2\beta} |\sin 2\phi|^{1-p} d\phi = \frac{\Gamma(\beta+1-p/2)\Gamma(1-p/2)}{2^p \Gamma(\beta+2-p)}.$$

And the second integral is

$$\int_0^{\pi} |\cos\ \theta|^{2p} d\theta = \frac{\sqrt{\pi}\Gamma(p+1/2)}{\Gamma(p+1)}.$$

Finally we have that

$$\lim_{n\to\infty} \frac{N_n(\beta,p)}{x_n^{\beta+1-p}} = \left(\frac{2}{\pi}\right)^p \frac{\Gamma(\beta+1-p/2)\Gamma(1-p/2)}{\sqrt{\pi}\Gamma(\beta+2-p)} \frac{\Gamma(p+1/2)}{\Gamma(p+1)} \equiv \overline{N}(\beta,p), \tag{25}$$

which is valid uniformly for p in some interval which contains $p=1$, according to the aforementioned indications from Ref. 15.

This limiting expression, which gives the asymptotics of the quantities of L_p-norm type $N_n(\beta,p)$ defined by Eq. (12), allows us to find the asymptotic values of the three components $(d/dp)N_n(\beta,p)|_{p=1}$, $N_n(\beta+1,1)$ and $(d/d\beta)N_n(\beta,1)$ of the searched entropy-like functional $S_n(\alpha,\beta)$. Indeed, because of the uniform convergence[14] one has

$$\left.\frac{d}{dp}\left[\frac{N_n(\beta,p)}{x_n^{\beta+1-p}}\right]\right|_{p=1} \to \left.\frac{d}{dp}\,\overline{N}(\beta,p)\right|_{p=1}$$

$$= \frac{\Gamma(\beta+1/2)}{2\pi^{1/2}\Gamma(\beta+1)} [2 + \gamma_E - 2\ln\ \pi - \psi(\beta+1/2) + 2\psi(\beta+1)] \tag{26}$$

(where $\gamma_E = 0.577215664...$ is the Euler constant) so that

$$\left.\frac{d}{dp} N_n(\beta,p)\right|_{p=1} = x_n^{\beta} \left.\frac{d}{dp}\,\overline{N}(\beta,p)\right|_{p=1} - N_n(\beta,1)\ln\ x_n + o(x_n^{\beta}). \tag{27}$$

On the other hand, taking into account the uniformity in β of the asymptotical formula (25) for $\bar{N}(\beta,p)$ we can derive straightforwardly the first terms of the remaining components $N_n(\beta+1,1)$ and $(d/d\beta)N_n(\beta,1)$ of $S_n(\alpha,\beta)$. Namely,

$$N_n(\beta+1,1)=\frac{\Gamma(\beta+3/2)}{\sqrt{\pi}\Gamma(\beta+2)}\,x_n^{\beta+1}+o(x_n^{\beta+1}), \tag{28}$$

$$\frac{d}{d\beta}N_n(\beta,1)=\frac{\Gamma(\beta+1/2)}{\sqrt{\pi}\Gamma(\beta+1)}\{x_n^{\beta}\ln x_n+(\psi(\beta+1/2)-\psi(\beta+1))x_n^{\beta}\}+o(x_n^{\beta}\ln x_n). \tag{29}$$

Taking into account the definition of x_n given by Eq. (21), the expressions (27)–(29) become as

$$\left.\frac{d}{dp}N_n(\beta,p)\right|_{p=1}=-\frac{2^{2\beta}\Gamma(\beta+1/2)}{\pi^{1/2}\Gamma(\beta+1)}\,n^{\beta}\ln n+\frac{2^{2\beta-1}\Gamma(\beta+1/2)}{\pi^{1/2}\Gamma(\beta+1)}$$
$$\times[2+\gamma_E-2\ln\pi-4\ln 2-\psi(\beta+1/2)+2\psi(\beta+1)]n^{\beta}+o(n^{\beta}), \tag{30}$$

$$N_n(\beta+1,1)=\frac{2^{\beta+2}}{\sqrt{\pi}}\frac{\Gamma(\beta+3/2)}{\Gamma(\beta+2)}\,n^{\beta+1}+o(n^{\beta+1}), \tag{31}$$

and

$$\frac{d}{d\beta}N_n(\beta,1)=\frac{2^{\beta+2}}{\sqrt{\pi}}\frac{\Gamma(\beta+1/2)}{\Gamma(\beta+1)}\{n^{\beta}\ln n+(\psi(\beta+1/2)-\psi(\beta+1))n^{\beta}\}+o(n^{\beta}\ln n). \tag{32}$$

To sum appropriately these three quantities it is convenient to have more accurate expressions for $N_n(\beta+1,1)$ and $(d/d\beta)N_n(\beta,1)$ than Eqs. (31) and (32), respectively. To this end we will use the following formula, first given by Schrödinger[16] and later rediscovered by other authors,[17,18] for the orthonormal Laguerre polynomials $q_n(x)=\hat{L}_n^{\alpha}(x)$:

$$\int_0^{\infty}x^s e^{-x}[q_n(x)]^2dx=\frac{n!}{\Gamma(n+\alpha+1)}\sum_{r=0}^{n}\binom{s-\alpha}{n-r}^2\frac{\Gamma(s+r+1)}{r!},$$

valid for any real $s>-1$. By means of this formula one can exactly evaluate the integral $\mathscr{T}_1(\beta)$ and $\mathscr{T}_2(\beta)$ given by Eqs. (15) and (16) as

$$\mathscr{T}_1(\beta)=\frac{n!}{\Gamma(n+\alpha+1)}\sum_{r=0}^{n}\binom{\beta+1}{n-r}^2\frac{\Gamma(\beta+\alpha+r+2)}{r!}, \tag{33a}$$

$$\mathscr{T}_2(\beta)=\frac{n!}{\Gamma(n+\alpha+1)}\sum_{r=0}^{n}\binom{\beta}{n-r}^2\frac{\Gamma(\beta+\alpha+r++1)}{r!}$$
$$\cdot[2\psi(\beta+1)+\psi(\beta+\alpha+r+1)-2\psi(\beta+r+1-n)]. \tag{33b}$$

On the other hand, it follows straightforwardly that

$$N_n(\beta+1,1)=\frac{2^{2\beta+2}\Gamma(\beta+3/2)}{\sqrt{\pi}\Gamma(\beta+2)}\left[n^{\beta+1}+\frac{1}{2}(\beta+1)(\alpha+1)n^{\beta}\right]+o(n^{\beta}) \tag{34}$$

and

$$\frac{d}{d\beta}N_n(\beta,1)=\frac{2^{2\beta}\Gamma(\beta+1/2)}{\sqrt{\pi}\Gamma(\beta+1)}\{n^{\beta}\ln n+[2\ln 2+\psi(\beta+1/2)-\psi(\beta+1)]n^{\beta}\}+o(n^{\beta}). \tag{35}$$

Finally, by summing up the expressions (30), (34) and (35) in the sense indicated by Eq. (19) one obtains the searched asymptotical formula (9) for the entropy-like functional $S_n(\alpha,\beta)$.

III. ENTROPIES FOR RYDBERG STATES OF THE TWO-DIMENSIONAL HARMONIC OSCILLATOR

Here we study analytically and numerically the information entropies of the two-dimensional harmonic oscillator $V(x,y)=1/2\lambda(x^2+y^2)$ in both position and momentum spaces for any excited state with an emphasis in the Rydberg region which corresponds to a large principal quantum number n.

Recently it has been shown[6,19] that the position and momentum entropies of the system are

$$S_\rho=-\ln\lambda+\ln(e\pi), \tag{36a}$$

$$S_\gamma=\ln\lambda+\ln(e\pi), \tag{36b}$$

for the ground state $(n=0,\ m=0)$ and

$$S_\rho=-\ln\lambda-\ln\left(\frac{n!}{(n+|m|)!\,\pi}\right)+2n+|m|+1-\frac{n!}{(n+|m|)!}\,(|m|I_1+I_2), \tag{37a}$$

$$S_\gamma=\ln\lambda-\ln\left(\frac{n!}{(n+|m|)!\,\pi}\right)+2n+|m|+1-\frac{n!}{(n+|m|)!}\,(|m|I_1+I_2), \tag{37b}$$

for an arbitrary excited state characterized by the quantum numbers $(n,|m|)$. The symbols I_1 and I_2 denote the integrals

$$I_1=\int_0^\infty \ln t\,\omega_{|m|}(t)[L_n^{|m|}(t)]^2dt=\frac{|m|}{n!}\,(n+|m|)!\,\psi(n+|m|+1), \tag{38a}$$

$$I_2=\int_0^\infty \omega_{|m|}(t)[L_n^{|m|}(t)]^2\,\ln[L_n^{|m|}(t)]^2dt=h_n(|m|)^2[\ln h_n(|m|)^2-E_n(|m|)], \tag{38b}$$

where $h_n(\alpha)=[\Gamma(n+\alpha+1)/n!]^{1/2}$ is the Laguerre norm, so that the orthonormal polynomials $\hat{L}_n^\alpha(x)=[h_n(\alpha)]^{-1}L_n^\alpha(x)$. Notice that the value of I_1 immediately follows from Eqs. (16) and (33b).

The substitution of the integral values (38a) and (38b) into Eqs. (37a) and (37b) leads us to the following exact expressions for the position and momentum entropies:

$$S_\rho=2n+|m|+1+\ln\frac{\pi}{\lambda}-|m|\psi(n+|m|+1)+E_n(|m|), \tag{39a}$$

$$S_\gamma=2n+|m|+1+\ln(\pi\lambda)-|m|\psi(n+|m|+1)+E_n(|m|), \tag{39b}$$

respectively, where $E_n(|m|)$ is the entropy of the Laguerre polynomial $\hat{L}_n^{|m|}(x)$.

One observes that the states $(n,\pm m)$ have the same entropies. Moreover, these expressions allow us to obtain easily the exact values of a few of the lowest excited states $(n,|m|)$ because the associated Laguerre entropic integral (7) can be analytically solved in such cases. So, for example, the first excited state $(n=0,\ |\mathrm{m}|=1)$ has the entropies

$$S_\rho=\ln\frac{e\pi}{\lambda}+\gamma_E,$$

$$S_\gamma=\ln(e\pi\lambda)+\gamma_E,$$

where γ_E is Euler's constant.

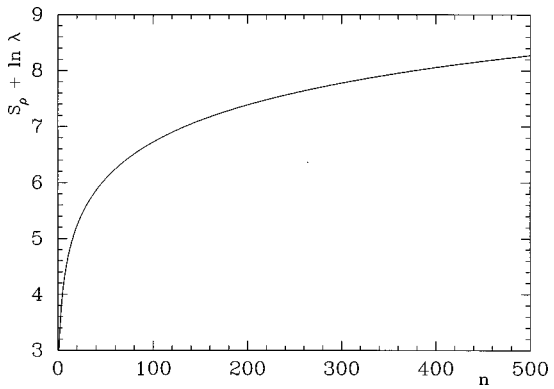

FIG. 1. Shifted position information entropy, $S_\rho + \ln \lambda$, of the two-dimensional harmonic oscillator $V(x,y) = 1/2\lambda(x^2 + y^2)$ for the excited states $(n, m=0)$ with $n = 0-500$.

However, the calculation of the entropic integral associated to an arbitrary state is a quite involved problem, still open.[6,13,14,19–21] Up to now, this integral has been analytically evaluated with high accuracy only in the asymptotic case (i.e., for large values of the degree of the involved polynomial) which physically corresponds to the Rydberg region of very excited states of the system. The asymptotical behavior of the entropy of the Laguerre polynomial was thoroughly discussed in Section II. The use of Eq. (10) together with Eqs. (39a) and (39b) leads us to the following behavior of the position and momentum entropies for very excited states of the two-dimensional harmonic oscillator:

$$S_\rho = \ln n - \ln \lambda + \ln(2\pi^2) - 1 + o(1), \tag{40a}$$

$$S_\gamma = \ln n + \ln \lambda + \ln(2\pi^2) - 1 + o(1), \tag{40b}$$

respectively. Notice that both Rydberg entropies do not depend on the quantum number m. More important it is the identification of a logarithmic dependence of both entropies on the principal quantum number n. Then, the entropy sum $S_\rho + S_\gamma$ has a logarithmic growth with n. In addition, this sum does not depend on the potential strength λ.

To know the accuracy of the position and momentum entropies for Rydberg states, as given by Eqs. (40a) and (40b) respectively, we have numerically analyzed the rate of convergence of the exact entropies [as given by Eqs. (39a) and (39b)] to these values for the excited states $(n, |m|=0)$ with $n = 0-500$. The results in position space are shown in Figures 1 and 2, where the quantity $S_\rho + \ln \lambda$ [as calculated numerically from Eq. (39a)] and the difference between this quantity and its asymptotical value (40a) are given, respectively. Similar results can be obtained for the momentum entropy case.

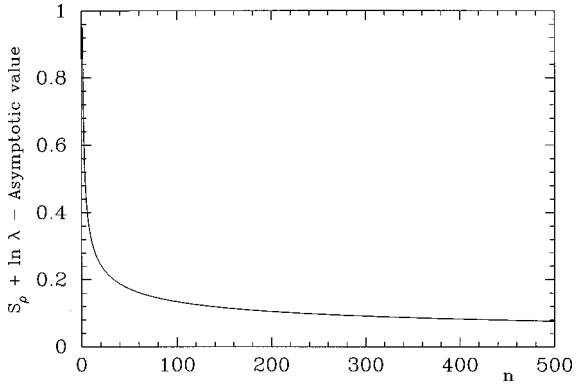

FIG. 2. The difference between the shifted position information entropy, $S_\rho + \ln \lambda$, and its asymptotical value for the excited states $(n, m=0)$, $n = 0-500$, of the two-dimensional harmonic oscillator $V(x,y) = 1/2\lambda(x^2 + y^2)$.

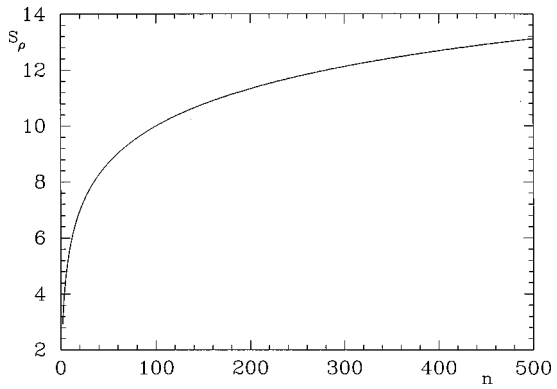

FIG. 3. Position information entropy S_ρ of the one-dimensional hydrogen atom, $V(x)=-1/|x|$, for the excited states characterized by the quantum number $n=0-500$.

IV. ENTROPIES FOR RYDBERG STATES OF THE ONE-DIMENSIONAL HYDROGEN ATOM

Here we study analytically and numerically the information entropies of the $1D$ hydrogen atom in both position and momentum spaces for an arbitrary state. They are[6]

$$S_\rho = \ln(2n^3) + 3n - \frac{1}{n^2}\left[J_2' + \frac{1}{2}J_3\right], \tag{41a}$$

$$S_\gamma = -\ln n + \ln(8\pi) - 2, \tag{41b}$$

respectively. The symbols J_2' and J_3 denote the integrals

$$J_2' = \int_0^\infty x \ln x \omega_1(x)[L_{n-1}^1(x)]^2 dx = n\int_0^\infty x \ln x \omega_1(x)[\hat{L}_{n-1}^1(x)]^2 dx = 2n^2 - n + 2n^2\psi(n+1), \tag{42a}$$

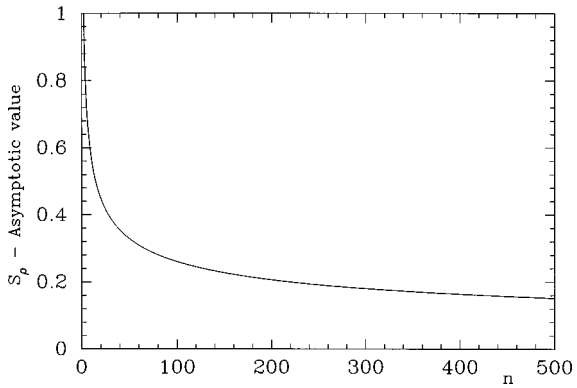

FIG. 4. The difference between the position entropy of the one-dimensional hydrogen atom and its asymptotical value given by Eq. (44) for the excited states with $n=0-500$.

$$J_3=\int_0^\infty x\omega_1(x)[L_{n-1}^1(x)]^2\ \ln[L_{n-1}^1(x)]^2dx=[h_{n-1}(1)]^2$$
$$\times\left\{\ln[h_{n-1}(1)]^2\int_0^\infty x\omega_1(x)[\hat{L}_{n-1}^1(x)]^2dx+S_{n-1}(1,1)\right\}$$
$$=n\{2n\ \ln\ n-F_{n-1}(1)\}. \tag{42b}$$

Notice that the value (42a) for the integral J_2' immediately follows from the general expressions (16) and (33b), and that to find the value (42b) of J_3 we have used Eqs. (15) and (33a) as well as Eqs. (6) and (8).

Putting the values (42a) and (42b) into Eq. (41a) one finds that

$$S_\rho=3n+\ln(2n^2)+\frac{1}{n}-2-2\psi(n+1)+\frac{1}{2n}\,F_{n-1}(1). \tag{43}$$

This expression gives the exact position entropy for a quantum-mechanical state of the $1D$ hydrogen atom characterized by the quantum number n. However, the analytical evaluation of the involved integral $F_{n-1}(1)$ is not generally possible. Only in the asymptotic case (i.e., for large n) this integral can be calculated by means of the expression (11) found in Section II. This observation allows us to find the value

$$S_\rho=2\ln\ n+2\ln 2+\ln\ \pi-2+o(1), \tag{44}$$

for the position entropy of the $1D$ hydrogen atom at a very excited (Rydberg) state. One notices again the logarithmic dependence for this entropy on the quantum number n characterizing the state under consideration. Then, the entropy sum $S_\rho+S_\gamma$ has again a logarithmic growth with n.

To know the accuracy of the Rydberg position entropy (44), we have numerically analyzed the rate of convergence of the exact entropy (43) to this value for the excited states with $n=0-500$. The results are shown in Figures 3 and 4, where the quantity S_ρ [as numerically calculated from Eq. (43)] and the difference between this quantity and its asymptotical value (44) are given, respectively.

Finally, as a concluding remark, we want to emphasize the logarithmically increasing behavior with the principal quantum number n of both the position entropy S_ρ as well as the entropy sum $S_\rho+S_\gamma$ for the two quantum-mechanical systems considered in this work. This behavior is the same in spite of the different increasing behaviors with n of the energies $\mathscr{E}_n$ of the quantum-mechanical states associated to these two systems with different power type potentials. Most probably, this logarithmic behavior of the Boltzmann–Shannon entropies is shared by a much larger set of physical systems, particularly those which are characterized by a potential of r^α type, $\alpha\in\mathbb{R}$, with bound states.

ACKNOWLEDGMENTS

We are very grateful for the partial support of the European Project No. INTAS-93-209-ext. Also, two of us (JSD, RJY) appreciate the financial help of the Junta the Andalucia under Project No. FQM0207 and the Spanish DGES Project No. PG95-1205.

[1] U. Schwengelbeck and F. H. M. Faisal, Phys. Rev. A **50**, 632 (1994).
[2] R. Loudon, Am. J. Phys. **27**, 649 (1959).
[3] G. F. Saville and J. M. Goodkind, Phys. Rev. A **50**, 2059 (1994).
[4] A. J. Bracken, D. S. McAnally, and O. A. Odundun, J. Math. Phys. **28**, 397 (1987).
[5] A. Luks, V. Perinova, and J. Krepelka, Phys. Rev. A **50**, 818 (1994).
[6] R. J. Yáñez, W. van Assche, and J. S. Dehesa, Phys. Rev. A **50**, 3065 (1994).
[7] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 and 623 (1949).
[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[9] S. Guiasu, *Information Theory with Applications* (McGraw-Hill, New York, 1977).
[10] J. Avery, *Hyperspherical Harmonics: Applications in Quantum Theory* (Kluwer, Dordrecht, 1988).
[11] L. Boltzmann, *Vorlesungen über Gastheorie* (Barth, Leipzig, 1896).
[12] I. Bialynicki-Birula and J. Mycielski, Commun. Math. Phys. **44**, 129 (1975).
[13] A. I. Aptekarev, J. S. Dehesa, and R. J. Yáñez, J. Math. Phys. **35**, 4423 (1994).

[14] A. I. Aptekarev, V. Buyarov, and J. S. Dehesa, Russ. Acad. Sci. Sb. Math. **185**, 3 (1994); English translation RAS Math. Sbo. **82**, 373 (1995).
[15] G. Szëgo, *Orthogonal Polynomials* (American Mathematical Society, Providence, RI, 1978).
[16] E. Schrödinger, Ann. Phys. (Leipzig) **79**, 361 (1926); **79**, 489 (1926); **80**, 437 (1926); **81**, 109 (1926).
[17] M. M. Nieto, Am. J. Phys. **47**, 1067 (1979).
[18] M. M. Nieto and L. M. Simmons, Phys. Rev. A **19**, 438 (1979).
[19] J. S. Dehesa, W. van Assche, and R. J. Yáñez, Methods Appl. Anal. **4**, 91 (1997).
[20] W. van Assche, R. J. Yáñez, and J. S. Dehesa, J. Math. Phys. **36**, 4106 (1995).
[21] A. I. Aptekarev, V. S. Buyarov, W. van Assche, and J. S. Dehesa, Dokl. Akad. Nauk **346**, 439 (1996) (in Russian); Dokl. Math. **53**, 47 (1996).

The modified Newton–Sabatier method for the coupled channel inverse scattering problem with charged particles at fixed energy

Matthias Eberspächer and Werner Scheid
Institut für Theoretische Physik der Justus-Liebig-Universität, Giessen, Germany

(Received 2 February 1998; accepted for publication 24 February 1998)

The modified Newton–Sabatier method was recently developed for inelastic inverse problems for the scattering of neutral particles. In this paper, the method is extended to the solution of the inelastic inverse scattering problem with charged particles. The bases are radial Schrödinger equations coupled by a local potential matrix known from a certain distance on. The inversion procedure starts with the S-matrix fixing the asymptotic wave functions of the charged scattering system. This S-matrix can be transformed to another one for asymptotic constant potentials instead of Coulomb potentials. The new S-matrix belongs to the same inner potential matrix as the original one and is solved by the modified Newton–Sabatier method similar to the case of neutral particles. An application with a given S-matrix, belonging to a coupled square-well potential matrix with an outer Coulomb potential yields good agreement between the inverted potential matrix and the original one. © *1998 American Institute of Physics.* [S0022-2488(98)01106-2]

I. INTRODUCTION

The inverse scattering problem (ISP) in quantum mechanics serves to determine potentials from measured cross sections. In the first step, the so-called phase-shift analysis, a set of phase shifts or S-matrix elements is calculated from the differential cross section.[1] Then these phase shifts are used in the inversion procedure to give a local optical potential.[2,3] A first solution for the ISP in quantum mechanics for a spherical potential modeling the elastic scattering of spinless particles at fixed energy was developed by Newton.[4] The method uses the integral representation of the radial Schrödinger equation (Gel'fand–Levitan equation) and requires the asymptotic wave functions for $r\rightarrow\infty$ as input data. Sabatier *et al.* extended the Newton(–Sabatier) method and studied the nonuniqueness of the ISP.[3] They were the first who tested the procedure in numerical calculations.[5] Münchow and Scheid[6] modified the Newton–Sabatier method under the assumption that the spherical potential $V(r)$ is known from a certain radial distance r_0 up to infinity. Thus, the potential is only unknown in the finite interval $[0,r_0]$, where it can be uniquely determined from a finite set of phase shifts. The modified method was worked out for the elastic scattering of neutral[6] and charged particles.[7] The modification made the method applicable to experimental data of heavy ion collisions.[8] Recent developments include a spin-orbit inversion scheme,[9] inversion of polarization-corrected phase shifts[10] and an extension to coupled channels for neutral particles.[11]

Other methods for the solution of the inverse scattering problem at fixed energy are, for example, the Bargmann method of Lipperheide and Fiedeldey,[12] the Wentzel–Kramers–Brillouin method of Kujawski and others,[13] the iterative-perturbative method of Mackintosh and others,[14] and the generalized Darboux transformations of Schnizer and Leeb.[15] The coupled channel inverse scattering problem at fixed angular momentum is discussed by Chadan and Sabatier[3] and Zakhariev and Suzko.[16] Numerical solutions of this problem were given by Kohlhoff *et al.*[17] and Sofianos *et al.*[18] Reviews on the ISP are given in the textbooks.[2,3,16,19,20] The most recent reviews of this field can be found in the Proceedings of the International Conference on ''Inverse and Algebraic Quantum Scattering Theory.''[21]

Recently the modified Newton–Sabatier method was extended to a coupled channel problem for the scattering of neutral particles at a fixed energy.[11] In this paper we generalize this method to the case that the potentials are of Coulomb type for radii larger than some radius r_0. In Section II we derive the coupled channel Schrödinger equations under the assumption that the coupling

 © 1998 American Institute of Physics

matrix does not depend on the angular momentum of relative motion. This special type of potential matrix can only induce monopole transitions. The Coulomb potentials for $r \geqslant r_0$ and the connection of the wave functions with the S-matrix are discussed in Section III. The general method to solve the ISP for coupled channels is described in Section IV. In order to simplify the solution of the ISP we transform the S-matrix belonging to asymptotic Coulomb potentials to an S-matrix belonging to asymptotic constant potentials and solve the ISP with this new S-matrix. The constant potentials have similar solutions to the asymptotic potentials of neutral particles. As explained in Section V both S-matrices yield the same potential matrix in the interior region. The procedure is tested in Section VI by applying it to the case of an S-matrix belonging to a potential matrix which consists of two coupled square wells with outer Coulomb potentials. A summary and comments are given in Section VII. In the Appendix we analytically calculate the S-matrix for a coupled square well model with Coulomb potentials in the outer region.

II. THE COUPLED CHANNEL SCHRÖDINGER EQUATIONS

We consider a scattering system with intrinsic dynamics, e.g., a scattering system consisting of two nuclei. The relative motion between the scattering partners and the intrinsic motion are described by the relative coordinate $\mathbf{r}$ and by a set of coordinates ξ, respectively. The Hamiltonian is assumed to be of the form

$$H = T(\mathbf{r}) + h(\xi) + W(r,\xi), \tag{1}$$

where $T = \mathbf{p}^2/2\mu$ denotes the kinetic energy operator, $h(\xi)$ the Hamiltonian of the intrinsic dynamics and $W(r,\xi)$ the coupling between relative and intrinsic motion. W is a local interaction operator depending on the intrinsic coordinates ξ and on the relative distance r of the colliding partners only. Equation (1) represents a special choice of the Hamiltonian which does not couple the orbital relative motion and the intrinsic motion. The operator of relative angular momentum commutes with H and, therefore, the wave function of the Schrödinger equation $H\Psi = E\Psi$ can be expanded as

$$\Psi_{\ell m,n} = \sum_{\alpha=1}^{N} R^{\ell}_{\alpha n}(r)\chi_\alpha(\xi)Y_{\ell m}(\theta,\phi). \tag{2}$$

$Y_{\ell m}$ describes the orbital motion, $\chi_\alpha(\xi)$ are the eigenfunctions of the intrinsic Hamiltonian

$$h(\xi)\chi_\alpha(\xi) = \varepsilon_\alpha \chi_\alpha(\xi). \tag{3}$$

with $\alpha = 1,\ldots,N$. We assume that N eigenstates exist which are characterized by the quantum number α. The eigenenergies ε_α are the internal excitation energies and are ordered with increasing energy $\varepsilon_1 = 0 \leqslant \varepsilon_2 \leqslant \ldots \leqslant \varepsilon_N$. The radial wave functions $R^{\ell}_{\alpha n}$ depend on the angular momentum quantum number ℓ, on the quatum number α of the intrinsic state, and on the number n of the degenerate solution to the same given channel energy $E_\alpha = E - \varepsilon_\alpha$. N also denotes the number of coupled channels.

With this special form of the Hamiltonian (1) and the wave function (2) only monopole transitions may be described. As an application one might consider transfer reactions, where all scattering partners are in the ground state. A possible system is the $^{16}\mathrm{O}(\alpha,{}^{8}\mathrm{Be}_{gs})^{12}\mathrm{C}$ reaction with the channels $\alpha + {}^{16}\mathrm{O}$ and ${}^{8}\mathrm{Be}_{gs} + {}^{12}\mathrm{C}$. In this case one has to regard different reduced masses in the channels.

Inserting the wave function (2) in the Schrödinger equation $(H-E)\Psi_{\ell m,n} = 0$, we get the following equation:

$$\sum_{\alpha=1}^{N} \chi_\alpha(\xi)\left(-\frac{\hbar^2}{2\mu}\frac{1}{r}\frac{d^2}{dr^2}r + \frac{\ell(\ell+1)\hbar^2}{2\mu r^2} + \varepsilon_\alpha - E + W(r,\xi)\right) R^{\ell}_{\alpha n}(r) = 0. \tag{4}$$

Under the simplifying assumption that the intrinsic wave functions form a complete set of functions, we obtain a set of N coupled equations for the radial wave functions by projecting with the intrinsic channel states $(\alpha = 1,\ldots,N)$:

$$\left[-\frac{\hbar^2}{2\mu}\frac{1}{r}\frac{d^2}{dr^2}r+\frac{\hbar^2}{2\mu}\frac{\ell(\ell+1)}{r^2}+\varepsilon_\alpha-E\right]R^{\ell}_{\alpha n}(r)+\sum_{\beta=1}^{N}V_{\alpha\beta}(r)R^{\ell}_{\beta n}(r)=0, \tag{5}$$

where $V_{\alpha\beta}(r)=<\chi_\alpha|W|\chi_\beta>$ is the effective potential matrix. Introducing the dimensionless quantities

$$\rho=kr=\sqrt{\frac{2\mu E}{\hbar^2}}r, \quad \varphi^{\ell}_{\alpha n}(\rho)=\rho R^{\ell}_{\alpha n}(r), \quad U_{\alpha\beta}(\rho)=\frac{V_{\alpha\beta}(r)}{E}, \tag{6}$$

we rewrite Eq. (5) in the form

$$\sum_{\beta=1}^{N}D^{U}_{\alpha\beta}(\rho)\varphi^{\ell}_{\beta n}(\rho)=\ell(\ell+1)\varphi^{\ell}_{\alpha n}(\rho), \tag{7}$$

with the differential operator

$$D^{U}_{\alpha\beta}(\rho)=\rho^2\left\{\left[\frac{d^2}{d\rho^2}+\frac{E_\alpha}{E}\right]\delta_{\alpha\beta}-U_{\alpha\beta}(\rho)\right\}. \tag{8}$$

III. ASYMPTOTIC COULOMB POTENTIALS

In this paper we consider the scattering of charged particles, e.g., nuclei. Therefore, we assume the charges of the projectile and target to be $Z_p e$ and $Z_t e$, respectively. Then, the potential matrix $U_{\alpha\beta}$ is diagonal and Coulombian from some radius ρ_0 up to infinity:

$$U_{\alpha\beta}(\rho>\rho_0)=\delta_{\alpha\beta}\frac{Z_p Z_t e^2 k}{\rho E}. \tag{9}$$

For this potential matrix, the asymptotic solutions can be written as

$$\varphi^{\ell}_{\alpha n}(\rho>\rho_0)=\sqrt{k_\alpha}\rho(\delta_{\alpha n}H^{-}_{\ell}(k_\alpha\rho)-S^{\ell}_{\alpha n}H^{+}_{\ell}(k_\alpha\rho)) \tag{10}$$

with

$$H^{\pm}_{\ell}(k_\alpha\rho)=\frac{e^{i\sigma^{\ell}_\alpha}}{k_\alpha\rho}(G_\ell(k_\alpha\rho)\pm iF_\ell(k_\alpha\rho)), \quad k_\alpha=\sqrt{1-\frac{\varepsilon_\alpha}{E}}. \tag{11}$$

Here, G_ℓ and F_ℓ are the irregular and regular Coulomb functions, respectively. The quantities σ^{ℓ}_α are the Coulomb phases:

$$\sigma^{\ell}_\alpha=\arg\Gamma(\ell+1+i\eta_\alpha), \quad \eta_\alpha=\frac{Z_p Z_t e^2\mu}{\hbar^2 k k_\alpha}. \tag{12}$$

The S-matrix contains the information about the potential matrix. In the following we use the extended modified Newton–Sabatier method to solve the ISP given by (7) and (10).

IV. THE ISP FOR COUPLED CHANNELS AT FIXED ENERGY

In order to solve the ISP we first introduce a symmetric reference potential matrix $U^0_{\alpha\beta}(\rho)=U^0_{\beta\alpha}(\rho)$. We assume that the regular solutions $\varphi^{0\ell}_{\alpha n}$ of the coupled channel Schrödinger equations are known for the reference potential matrix:

$$\sum_{\beta=1}^{N}D^{U_0}_{\alpha\beta}(\rho)\varphi^{0\ell}_{\beta n}(\rho)=\ell(\ell+1)\varphi^{0\ell}_{\alpha n}(\rho) \tag{13}$$

with

$$D^{U_0}_{\alpha\beta}(\rho)=\rho^2\left\{\left[\frac{d^2}{d\rho^2}+\frac{E_\alpha}{E}\right]\delta_{\alpha\beta}-U^0_{\alpha\beta}(\rho)\right\}. \tag{14}$$

Then it is possible to write the wave function of the coupled channel equations in the form

$$\varphi^{\ell}_{\alpha n}(\rho)=\varphi^{0\ell}_{\alpha n}(\rho)-\sum_{\beta=1}^{N}\int_0^\rho \frac{d\rho'}{\rho'^2}K^{UU_0}_{\alpha\beta}(\rho,\rho')\varphi^{0\ell}_{\beta n}(\rho'), \tag{15}$$

if the potential matrix is given by

$$U_{\alpha\beta}(\rho)=U^0_{\alpha\beta}(\rho)-\frac{2}{\rho}\frac{d}{d\rho}\frac{K^{UU_0}_{\alpha\beta}(\rho,\rho)}{\rho}. \tag{16}$$

The kernel matrix $K^{UU_0}_{\alpha\beta}$ has to fulfill the differential equation

$$\sum_{\beta=1}^{N}D^{U}_{\alpha\beta}(\rho)K^{UU_0}_{\beta\gamma}(\rho,\rho')=\sum_{\beta=1}^{N}D^{U_0}_{\gamma\beta}(\rho')K^{UU_0}_{\alpha\beta}(\rho,\rho'), \tag{17}$$

with the boundary conditions

$$K^{UU_0}_{\alpha\beta}(\rho=0,\rho')=K^{UU_0}_{\alpha\beta}(\rho,\rho'=0)=0. \tag{18}$$

A possible kernel matrix which fulfills Eqs. (17) and (18) can be expanded in terms of the radial functions $\varphi^{\ell}_{\alpha n}$ and $\varphi^{0\ell}_{\alpha n}$ with coefficients $c^{\ell}_{nn'}$:

$$K^{UU_0}_{\alpha\beta}(\rho,\rho')=\sum_{\ell=0}^{\infty}\sum_{n,n'=1}^{N}c^{\ell}_{nn'}\varphi^{\ell}_{\alpha n}(\rho)\varphi^{0\ell}_{\beta n'}(\rho'). \tag{19}$$

Inserting this kernel into Eq. (15) we get the fundamental coupled equations for the solution of the ISP:

$$\varphi^{\ell}_{\alpha n}(\rho)=\varphi^{0\ell}_{\alpha n}(\rho)-\sum_{\ell'=0}^{\infty}\sum_{n'',n'=1}^{N}\varphi^{\ell'}_{\alpha n''}(\rho)c^{\ell'}_{n''n'}L^{\ell'\ell}_{n'n}(\rho), \tag{20}$$

where $L^{\ell\ell'}$ is a matrix determined by the known reference solutions:

$$L^{\ell'\ell}_{n'n}(\rho)=\sum_{\beta=1}^{N}\int_0^\rho\frac{d\rho'}{\rho'^2}\varphi^{0\ell'}_{\beta n'}(\rho')\varphi^{0\ell}_{\beta n}(\rho'). \tag{21}$$

Equation (20) depends on the so called spectral coefficients $c^{\ell}_{nn'}$ which are first unknown. Since the wave functions $\varphi^{\ell}_{\alpha n}(\rho)$ are known for $\rho>\rho_0$ if the S-matrix is given [see Eq. (10)], Eq. (20) can be used to calculate the coefficients $c^{\ell}_{nn'}$ by solving it with the wave functions at values of $\rho>\rho_0$. Then, Eq. (20) can be applied for $\rho<\rho_0$ to obtain the unknown wave functions $\varphi^{\ell}_{\alpha n}(\rho)$ in the inner region, which finally yield the potential matrix $U_{\alpha\beta}(\rho)$ by using Eqs. (16) and (19). This procedure provides us a unique determination of the potential matrix for a given S-matrix inside the family of potential matrices defined by Eq. (16).

V. REDUCTION TO THE CASE OF AN ASYMPTOTIC CONSTANT POTENTIAL

For the case of an asymptotic Coulomb potential (9) one may choose a Coulomb potential as reference potential and Coulomb wave functions as reference wave functions. In order to avoid the difficulties with these functions with respect to the singularity of the Coulomb potential at $\rho=0$

and with respect to the calculation of the matrix elements (21), we first transform the S-matrix to a new S-matrix for a constant asymptotic potential and then solve the ISP with a constant reference potential $U^0_{\alpha\beta}$.

A. Transformation of the *S*-matrix

Let us consider a potential matrix $U^B_{\alpha\beta}(\rho)$ with the properties

$$U_{\alpha\beta}{}^B(\rho)=\begin{cases} U_{\alpha\beta}(\rho) & \rho<\rho_B \\ \delta_{\alpha\beta}\dfrac{Z_pZ_te^2k}{\rho_BE} & \rho\geqslant\rho_B, \end{cases} \tag{22}$$

where ρ_B is a fixed radius situated outside the nuclear interaction region. $U^B_{\alpha\beta}(\rho)$ is schematically shown together with $U_{\alpha\beta}(\rho)$ in Fig. 1. Both potential matrices, $U_{\alpha\beta}(\rho)$ and $U^B_{\alpha\beta}(\rho)$, are identical in the inner region. The S-matrices $S^{\ell}_{\alpha n}$ and $S^{\ell B}_{\alpha n}$ belong to the same potential matrix $U_{\alpha\beta}(\rho)$ for $\rho<\rho_B$, but different matrices for $\rho\geqslant\rho_B$. Solving the ISP we will get the same potential matrix $U_{\alpha\beta}(\rho)$ from $S^{\ell}_{\alpha n}$ and $S^{\ell B}_{\alpha n}$ for $\rho<\rho_B$. However, the solution of the ISP is much easier to handle by starting with $S^{\ell B}_{\alpha n}$ because the reference potential can be chosen as constant, i.e., as independent of ρ. The S-matrix $S^{\ell B}_{\alpha n}$ is obtained by a transformation from $S^{\ell}_{\alpha n}$. The Schrödinger equation with the Coulomb potential is given for $\rho\geqslant\rho_B$ by

$$\left(\frac{1}{\rho}\frac{d^2}{d\rho^2}\rho-\frac{\ell(\ell+1)}{\rho^2}+k_\alpha^2-\frac{Z_pZ_te^2k}{\rho E}\right)T^{\ell}_{\alpha n}(\rho)=0. \tag{23}$$

Its solutions are of the form (10):

$$T^{\ell}_{\alpha n}(\rho\geqslant\rho_B)=\sqrt{k_\alpha}(\delta_{\alpha n}H^-_{\ell}(k_\alpha\rho)-S^{\ell}_{\alpha n}H^+_{\ell}(k_\alpha\rho)), \tag{24}$$

with $H^{\pm}_{\ell}$ already defined by Eq. (11). The Schrödinger equation for $\rho\geqslant\rho_B$ for the potential matrix $U^B_{\alpha\beta}$ is of the form

$$\left(\frac{1}{\rho}\frac{d^2}{d\rho^2}\rho-\frac{\ell(\ell+1)}{\rho^2}+(k^B_\alpha)^2\right)T^{\ell B}_{\alpha n}(\rho)=0, \tag{25}$$

with the abbreviations

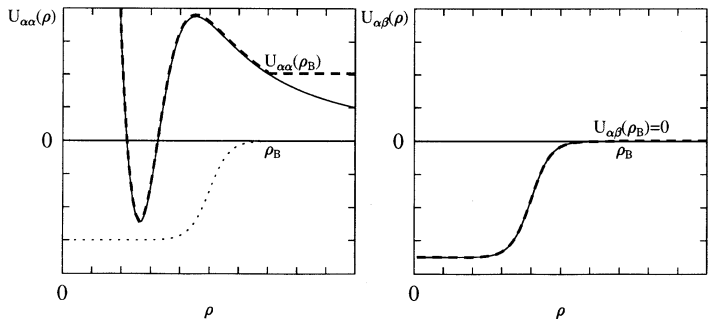

FIG. 1. Schematic picture to the transformation of the S-matrix. The left diagram shows the diagonal elements of $U_{\alpha\alpha}$ (real part: —, imaginary part: $\cdots$) and $U^B_{\alpha\alpha}$ (real part: - - -, imaginary part: $\cdots$). The right diagram gives the off-diagonal elements $U_{\alpha\beta}$ and $U^B_{\alpha\beta}$. ρ_B is the transformation radius.

$$(k_\alpha^B)^2=\frac{E_B-\varepsilon_\alpha}{E},\quad E_B=E-\frac{Z_pZ_te^2k}{\rho_B}. \tag{26}$$

The solutions of Eq. (25) are the modified spherical Hankel functions:

$$T_{\alpha n}^{\ell B}(\rho\geqslant\rho_B)=\sqrt{k_\alpha^B}(\delta_{\alpha n}h_\ell^-(k_\alpha^B\rho)-S_{\alpha n}^{\ell B}h_\ell^+(k_\alpha^B\rho)). \tag{27}$$

The modified Hankel functions are defined in terms of regular and irregular spherical Bessel functions (that is why we use the suffix ''B''):

$$h_\ell^\pm(\rho)=-n_\ell(\rho)\pm ij_\ell(\rho). \tag{28}$$

In order to get the new S-matrix elements $S_{\alpha n}^{\ell B}$, we equate the wave functions composed of the solutions (24) and (27) and their derivatives at $\rho=\rho_B$, e.g.,

$$\sum_{n'=1}^{N}\sqrt{k_\alpha}(\delta_{\alpha n'}H_\ell^-(k_\alpha\rho_B)-S_{\alpha n'}^{\ell}H_\ell^+(k_\alpha\rho_B))D_{n'n}^{\ell}=\sqrt{k_\alpha^B}(\delta_{\alpha n}h_\ell^-(k_\alpha^B\rho_B)-S_{\alpha n}^{\ell B}h_\ell^+(k_\alpha^B\rho_B)). \tag{29}$$

For fixed ℓ this procedure yields two sets of $N\times N$ linear equations for the $N\times N$ unknown coefficients $D_{nn'}^{\ell}$ and the $N\times N$ new S-matrix elements $S_{\alpha n}^{\ell B}$. Similar transformations can also be done for other asymptotic potentials.

B. Solution of ISP with a constant reference potential

With the S-matrix elements $S_{\alpha n}^{\ell B}$ we can solve the ISP with a method which is very similar to the one for neutral particles. For the scattering of neutral particles the potential matrix is zero from a certain radius ρ_0 and the reference potential can be set $U_{\alpha\beta}^0=0$. In our case we choose $U_{\alpha\beta}^0=\delta_{\alpha\beta}(Z_pZ_te^2k/\rho_BE)$ [see Eq.(22)]. The regular solutions of Eq. (13) are then

$$\varphi_{\alpha n}^{0\ell}(\rho)=\rho T_{\alpha n}^{0\ell}(\rho):=\rho j_\ell(k_\alpha^B\rho)\,\delta_{\alpha n}, \tag{30}$$

where the function j_ℓ is the regular spherical Bessel function. The asymptotic solution for the potential matrix $U_{\alpha\beta}^B(\rho)$ was already given in terms of the S-matrix in Eq. (27) and is repeated here:

$$T_{\alpha n}^{\ell B}(\rho\geqslant\rho_B)=\sqrt{k_\alpha^B}(\delta_{\alpha n}h_\ell^-(k_\alpha^B\rho)-S_{\alpha n}^{\ell B}h_\ell^+(k_\alpha^B\rho)), \tag{31}$$

where $h_\ell^\pm$ are modified spherical Hankel functions defined in Eq. (28). The wave functions $\varphi_{\alpha n}^{\ell}(\rho)$ are written as superpositions of all the degenerate solutions with the asymptotic (31) belonging to a single channel α with yet unknown coefficients $A_{nn'}^{\ell}$:

$$\varphi_{\alpha n}^{\ell}(\rho)=\sum_{n'=1}^{N}\rho T_{\alpha n'}^{\ell}(\rho)A_{n'n}^{\ell}. \tag{32}$$

By introducing the wave functions (30) and (32) into Eq. (20), we get the coupled equations

$$\sum_{n'=1}^{N}\left(T_{\alpha n'}^{\ell}(\rho)A_{n'n}^{\ell}+\sum_{\ell'=0}^{\ell_{\max}}T_{\alpha n'}^{\ell'}(\rho)b_{n'n}^{\ell'}L_n^{\ell'\ell}(\rho)\right)=T_{\alpha n}^{0\ell}(\rho) \tag{33}$$

with new matrices

$$L_n^{\ell\ell'}(\rho)=\int_0^{\rho} j_\ell(k_n^B\rho')j_{\ell'}(k_n^B\rho')d\rho', \quad b_{nn'}^{\ell}=\sum_{n''=1}^{N} A_{nn''}^{\ell}c_{n''n'}^{\ell}. \tag{34}$$

In order to get a finite set of equations, we limit the sum over the angular momenta in Eq.(33) to $\ell_{\max}$. The effect of this limitation is discussed in Ref. 22. Equation (33) is twofold used—first, for the calculation of the coefficient matrices $A_{nn'}^{\ell}$ and $b_{nn'}^{\ell}$ at $\rho>\rho_B$ and then for the calculation of the wave functions $T_{\alpha n}^{\ell}(\rho)$ at $\rho<\rho_B$. For the first step, two arbitrary radii $\rho_1,\rho_2>\rho_B$ are chosen to obtain a set of $2\times(\ell_{\max}+1)\times N\times N$ linear equations for the $(\ell_{\max}+1)\times N\times N$ unknown coefficients $A_{nn'}^{\ell}$ and the $(\ell_{\max}+1)\times N\times N$ unknown spectral coefficients $b_{nn'}^{\ell}$. After determining these coefficients the wave functions $T_{\alpha n}^{\ell}(\rho)$ are calculated at discrete points ρ_i in the inner region $0<\rho_i<\rho_B$. Then the kernel matrix (19) needed for the potential matrix is given by

$$K_{\alpha\beta}^{UU_0}(\rho,\rho')=\sum_{\ell=0}^{\ell_{\max}}\sum_{n=1}^{N} b_{\beta n}^{\ell}T_{\alpha n}^{\ell}(\rho)j_\ell(k_\beta^B\rho)\rho^2. \tag{35}$$

The uniqueness of this method of the solution of the ISP was proved and discussed in Ref. 11.

VI. APPLICATIONS

We apply our method to the case of coupled square well potentials with asymptotic Coulomb potentials. For simplicity we restrict ourselves to two channels only. In this case, analytic formulas for the S-matrix for the scattering of neutral particles were derived by Lovas.[23] In the Appendix we extend these calculations to charged particles. This provides us with analytically exact input data for the inversion procedure.

As an example, we assume a potential matrix for two channels with charged particles:

$$\mathbf{V}(r)=\begin{cases}\begin{pmatrix}-6-3i & -4\\ -4 & -5-4i\end{pmatrix}\text{MeV}, & r\leqslant 5 \text{ fm},\\ \dfrac{12e^2}{r}\ \text{MeV}, & r>5 \text{ fm}.\end{cases} \tag{36}$$

The excitation energies are $\varepsilon_1=0$ MeV and $\varepsilon_2=5$ MeV. The charge numbers $Z_p=2$ and $Z_t=6$ and the reduced mass $\mu=2794.50$ MeV/c^2 are chosen to model the scattering system ${}_2^4$He $+{}_6^{12}$C. The scattering energy is $E=200$ MeV. The S-matrix is calculated with the formulas of the Appendix.

Figure 2 shows the obtained potential matrices calculated with three different transformation radii r_B. The diagrams at the top and bottom of Fig. 2 present the diagonal elements V_{11} and V_{22}, respectively. The two diagrams in the middle correspond to the coupling potentials V_{12} and V_{21}. The left side shows the real part, the right side the imaginary part of the potential matrix elements.

The transformation radii are $r_B=5$,6, and 7 fm for the full, dotted and dashed curve, respectively. The radii needed for the calculation of the spectral coefficients in Eq. (33) are $r_1=r_B+0.05$ fm and $r_2=r_B+0.1$ fm. The step width for the discrete points r_i for the calculation of the wave functions in Eq. (33) is $\Delta r=r_{i+1}-r_i=0.05$ fm. All three results reproduce the initial potential matrix (36) with small oscillations, which are caused by a pole at $r=0$ in Eq. (16) and by the discontinuity of the square well potential at $r=5$ fm. The amplitude of the oscillations gets slightly larger, the further out the transformation radius r_B is chosen. The best reproduction is given, when the transformation radius r_B coincides with the radius $r_0=5$ fm.

Let us now consider a way of diminishing the small oscillations which have a wavelength of about $\lambda\approx 0.5$ fm in our example. In order to calculate the potential matrix (16) we take the derivative with respect to ρ by evaluating the kernel matrix $K_{\alpha\beta}^{UU_0}$ at distances r_i separated by a step width Δr and by applying a three-point formula for the derivative. Figure 3 shows the

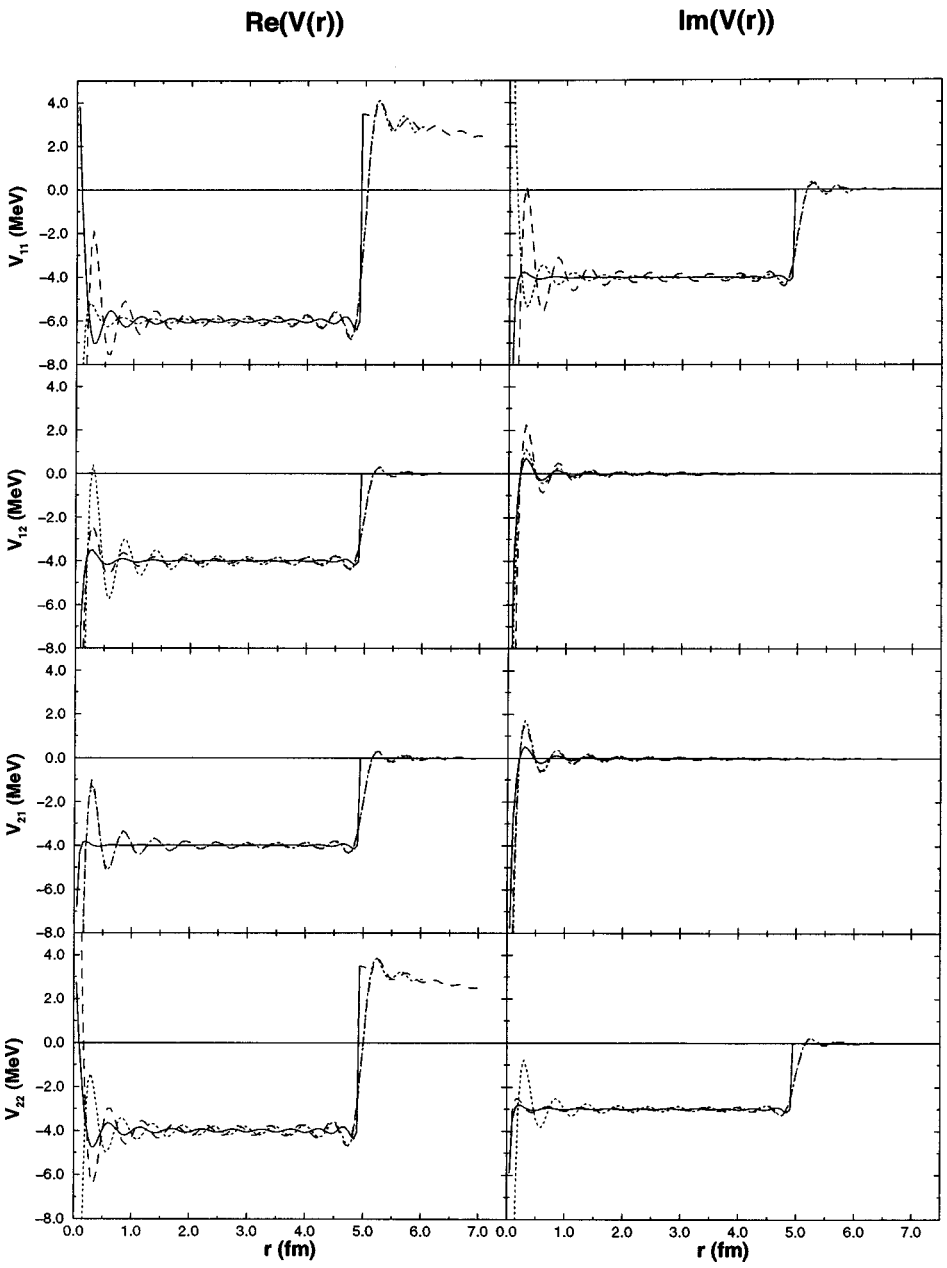

FIG. 2. Inverted potential matrix $V_{\alpha\beta}(r)$ for different transformation radii: $r_B=5$ fm (—) with $\ell_{\max}=24$, 6 fm $(\cdots)$ with $\ell_{\max}=28$, and 7 fm (- - -) with $\ell_{\max}=32$. For further details see text.

potential matrices obtained from the calculations with three different step widths for the discrete points r_i. The original potential matrix is the same as in Eq. (36). The transformation radius is $r_B=6$ fm, the radii for the calculation of the spectral coefficients are $r_1=6.05$ fm and $r_2=6.1$ fm. The step widths used are $\Delta r=0.01,\ 0.05$, and 0.25 fm for the full, dotted and dashed curves, respectively. The inverted potentials are given at 606, 120 and 22 radii for the full, dotted and dashed curve, respectively. The dotted curve is the same as in Fig. 2. As can be seen in Fig. 3, the inverted potential can not be improved by decreasing the step width from 0.05 to 0.01 fm. If we increase the step width to half of the wavelength $\Delta=\lambda/2=0.25$ fm (dashed curve) the oscillations are canceled out. This is a nontrivial process, as the numerical method for the potential matrix (16) takes the kernel (19) at three neighboring points into account. As a result of this improvement, the inverted potential is now only known at 22 radii. If the potential is needed on a finer grid (for example, in order to calculate the S-matrix back numerically), one might apply an interpolation

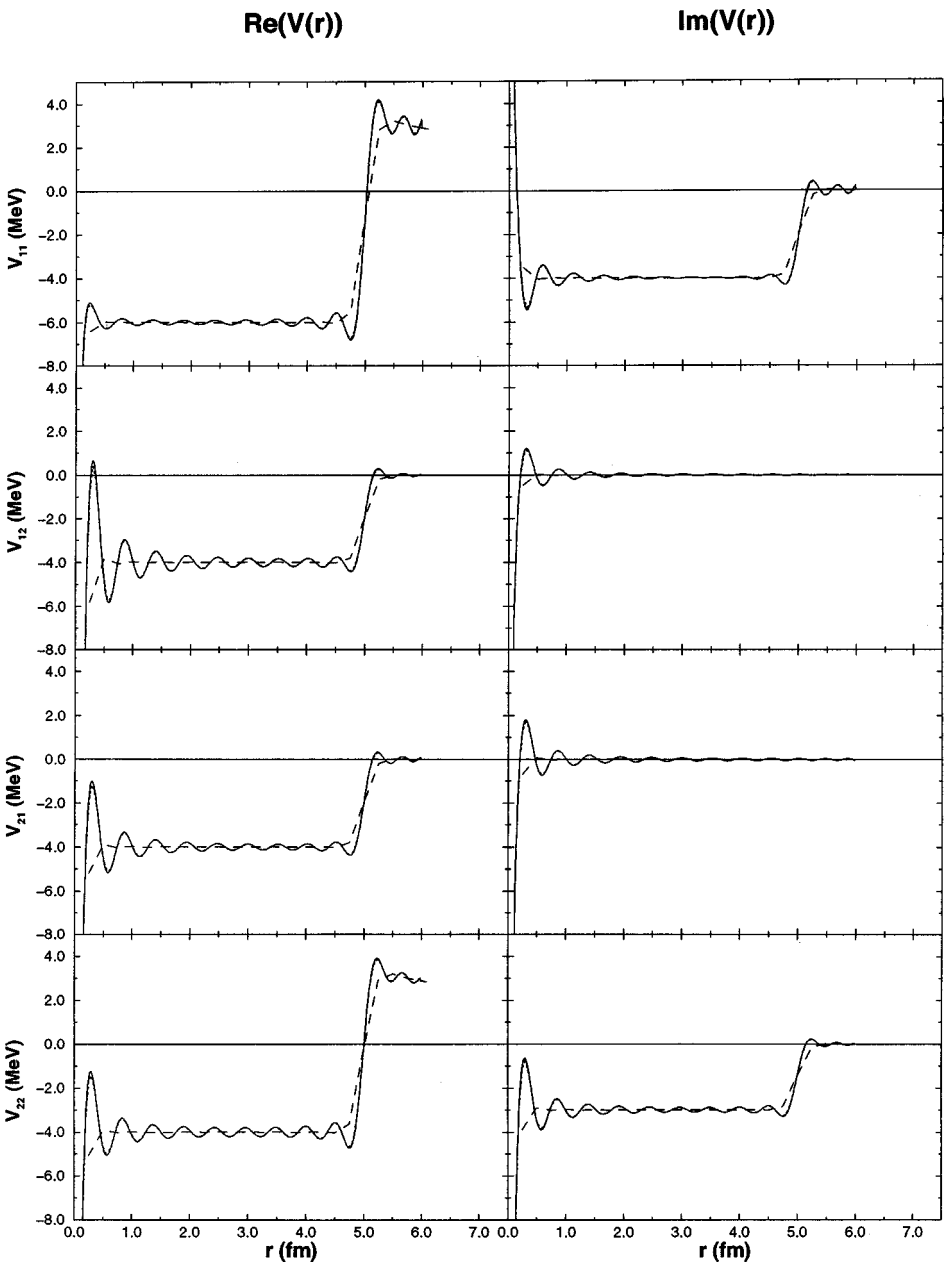

FIG. 3. Inverted potential matrix $V_{\alpha\beta}(r)$ for different step widths: $\Delta r = 0.01$ fm (—), 0.05 fm ($\cdots$), and 0.25 fm (- - -) all with $\ell_{\max} = 28$. For further details see text.

procedure. However, any structure of the potential smaller than the step width is then lost. Thus, the described method is only applicable for smoothly behaving well potentials discussed in this paper. For other potentials different regularization procedures must be considered.

VII. SUMMARY AND CONCLUSIONS

We extended the modified Newton–Sabatier method for coupled channels and neutral particles to the case of charged particles. The method is restricted to cases in which the potential matrix does not depend on the angular momentum of the relative motion. To remove this restriction may be very difficult inside the modified Newton–Sabatier method.

The inversion method for charged particles proposed here is carried out in two steps. First, the S-matrix of the charged particle problem is transformed to an S-matrix belonging to asymptotic

constant potentials. Then, the modified Newton–Sabatier method is applied to this new S-matrix in a similar way as for neutral particles. This procedure shows good agreement with the initial potential matrix for the presented example of two channels coupled by a square well potential matrix. Applications to more realistic potentials are in progress.

Applications to transfer reactions require a channel-dependent reduced mass μ_α and channel-dependent charge numbers Z_p^α and Z_t^α. Both features can be easily included in our present method.

For the inversion procedure in the case of coupled channels the complete S-matrix is needed as input data. This confronts us with the fundamental problem that a complete S-matrix at fixed energy is generally not known from experimental data. Usually in an experiment, the particles are scattered in their ground state. Thus, only one column of the S-matrix can be measured, describing the excitation from this state. Only for two channels and no absorption one might use the unitary condition for the S-matrix and calculate the missing matrix elements.[1]

ACKNOWLEDGMENTS

The authors are thankful to Professor B. Apagyi (Technical University of Budapest) for his advice in several details. This work was supported by the Graduiertenkolleg ''Theoretische und experimentelle Schwerionenphysik'' (DFG) and in the framework of the Hungarian–German scientific collaboration (MTA/DFG).

APPENDIX: SOLUTION OF A COUPLED SQUARE WELL MODEL WITH CHARGED PARTICLES

We limit ourselves to two channels. The potential matrix is

$$\mathbf{U}(\rho)=\begin{cases}\begin{pmatrix}U_{11} & U_{12}\\ U_{21} & U_{22}\end{pmatrix}, & \rho\leqslant\rho_0\\ \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\dfrac{Z_pZ_te^2k}{\rho E}, & \rho>\rho_0\end{cases} \tag{A1}$$

where the matrix elements U_{11}, U_{22} and $U_{12}=U_{21}\equiv U$ are constant. Z_p, Z_t are the charge numbers of the projectile and target, respectively. Then we can rewrite Eq. (5) for the inner region:

$$\left[\frac{d^2}{d\rho^2}+k_\alpha^2-\frac{\ell(\ell+1)}{\rho^2}\right]\varphi_{\alpha n}^{\ell}(\rho)=\sum_{\beta=1}^{2}U_{\alpha\beta}\varphi_{\beta n}^{\ell}(\rho),\rho\leqslant\rho_0 \tag{A2}$$

and for the outer region:

$$\left[\frac{d^2}{d\rho^2}+k_\alpha^2-\frac{\ell(\ell+1)}{\rho^2}-\frac{Z_pZ_te^2k}{\rho E}\right]\varphi_{\alpha n}^{\ell}(\rho)=0,\rho>\rho_0 \tag{A3}$$

with $k_\alpha^2=1-\varepsilon_\alpha/E$. For the solution with number n, let the channel $\alpha=n$ be the one that contains an incoming wave ($\rho>\rho_0$):

$$\varphi_{\alpha n}^{\ell}(\rho)=\sqrt{k_\alpha}\rho(\delta_{\alpha n}H_\ell^-(k_\alpha\rho)-S_{\alpha n}H_\ell^+(k_\alpha\rho)) \tag{A4}$$

with

$$H_\ell^\pm(k_\alpha\rho)=\frac{e^{i\sigma_\alpha^\ell}}{k_\alpha\rho}(G_\ell(k_\alpha\rho)\pm iF_\ell(k_\alpha\rho)), \tag{A5}$$

where G_ℓ and F_ℓ are the irregular and regular Coulomb wave functions and σ_α^ℓ are the Coulomb phases:

$$\sigma_\alpha^\ell=\arg\Gamma(\ell+1+i\eta_\alpha),\quad \eta_\alpha=\frac{Z_pZ_te^2\mu}{\hbar^2kk_\alpha}. \tag{A6}$$

For the solution in the internal region, which is regular at the origin we introduce the ansatz:

$$\varphi^{\swarrow}_{1n}(\rho)=\rho\sum_{n'=1}^{2} j_{\swarrow}(\tilde{k}_{n'}\rho)a_{n'n}, \quad \varphi^{\swarrow}_{2n}(\rho)=\rho\sum_{n'=1}^{2} j_{\swarrow}(\tilde{k}_{n'}\rho)b_{n'n}. \tag{A7}$$

Inserting these expansions into Eq. (A2), we obtain a set of homogeneous linear equations for the coefficients $a_{\alpha n}$, $b_{\alpha n}$:

$$(k_1'^2-\tilde{k}_\alpha^2)a_{\alpha n}-Ub_{\alpha n}=0, \tag{A8}$$

$$Ua_{\alpha n}-(k_2'^2-\tilde{k}_\alpha^2)b_{\alpha n}=0 \tag{A9}$$

with $k_\alpha'^2=k_\alpha^2-U_{\alpha\alpha}$. This system has a nontrivial solution for

$$\left.\begin{matrix}\tilde{k}_1^2\\ \tilde{k}_2^2\end{matrix}\right\}=\frac{k_1'^2+k_2'^2}{2}\pm\sqrt{\frac{(k_1'^2-k_2'^2)^2}{4}+U^2}. \tag{A10}$$

Using Eqs. (A8) and (A9) let us define the ratios

$$\gamma_\alpha\equiv\frac{a_{\alpha n}}{b_{\alpha n}}=\frac{U}{k_1'^2-\tilde{k}_\alpha^2}=\frac{k_2'^2-\tilde{k}_\alpha^2}{U}. \tag{A11}$$

Next we require the continuity of the logarithmic derivatives of the wave functions at $\rho=\rho_0$ in each channel separately. This yields the following ratios of the coefficients:

$$\begin{aligned}&\frac{b_{11}}{b_{21}}=-\frac{W_{22}^+}{W_{21}^+}, \quad \frac{a_{12}}{a_{22}}=-\frac{W_{12}^+}{W_{11}^+}, \quad -\omega_1\equiv\frac{a_{11}}{a_{21}}=-\frac{W_{12}^--S_{11}W_{12}^+}{W_{11}^--S_{11}W_{11}^+},\\ &-\omega_2\equiv\frac{b_{12}}{b_{22}}=-\frac{W_{22}^--S_{22}W_{22}^+}{W_{21}^--S_{22}W_{21}^+}.\end{aligned} \tag{A12}$$

Here, $W_{\alpha\beta}^{\pm}$ is the Wronskian:

$$W_{\alpha\beta}^{\pm}=\tilde{k}_\beta j'_{\swarrow}(\tilde{k}_\beta\rho_0)H_{\swarrow}^{\pm}(k_\alpha\rho_0)-k_\alpha H'^{\pm}_{\swarrow}(k_\alpha\rho_0)j_{\swarrow}(\tilde{k}_\beta\rho_0). \tag{A13}$$

Using Eqs. (A11) and (A12) we obtain

$$S_{11}=\frac{\gamma_2W_{21}^+W_{12}^--\gamma_1W_{22}^+W_{11}^-}{\gamma_2W_{21}^+W_{12}^+-\gamma_1W_{22}^+W_{11}^+}, \quad S_{22}=\frac{\gamma_1W_{11}^+W_{22}^--\gamma_2W_{12}^+W_{21}^-}{\gamma_1W_{11}^+W_{22}^+-\gamma_2W_{12}^+W_{21}^+}. \tag{A14}$$

The coefficients a_{11} and b_{12} can be calculated if we require the continuity of the wave functions at $\rho=\rho_0$:

$$a_{11}=\frac{\sqrt{k_1}(H_{\swarrow}^-(k_1\rho_0)-S_{11}H_{\swarrow}^+(k_1\rho_0))}{(j_{\swarrow}(\tilde{k}_1\rho_0)-j_{\swarrow}(\tilde{k}_2\rho_0)/\omega_1)}, \quad b_{12}=\frac{\sqrt{k_2}(H_{\swarrow}^-(k_2\rho_0)-S_{22}H_{\swarrow}^+(k_2\rho_0))}{(j_{\swarrow}(\tilde{k}_1\rho_0)-j_{\swarrow}(\tilde{k}_2\rho_0)/\omega_2)}. \tag{A15}$$

Repeated applications of Eqs. (A11) and (A12) give $a_{21}, a_{12}, a_{22}, b_{11}, b_{21}$ and b_{22}. Then the off-diagonal elements of the S-matrix can be calculated as:

$$S_{12}=-\frac{a_{12}j_{\swarrow}(\tilde{k}_1\rho_0)+a_{22}j_{\swarrow}(\tilde{k}_2\rho_0)}{\sqrt{k_1}H_{\swarrow}^+(k_1\rho_0)}, \quad S_{21}=-\frac{b_{11}j_{\swarrow}(\tilde{k}_1\rho_0)+b_{21}j_{\swarrow}(\tilde{k}_2\rho_0)}{\sqrt{k_2}H_{\swarrow}^+(k_2\rho_0)}. \tag{A16}$$

All derived formulas are reduced to the form presented by R.G. Lovas [23] if $\eta_\alpha=0$. In this case the Coulomb functions $H_{\swarrow}$ are the modified Hankel functions $h_{\swarrow}^{\pm}$:

$$H_\ell^\pm(\rho) \overset{\eta\to 0}{\to} h_\ell^\pm(\rho) = -n_\ell(\rho) \pm i j_\ell(\rho), \tag{A17}$$

where n_ℓ and j_ℓ are the irregular and regular spherical Bessel functions.

[1] D. R. Lun, L. J. Allen, and K. Amos, Phys. Rev. A **50**, 4000 (1994); V. Chiste, R. Lichtenthäler, A. C. C. Villari, and L. C. Gomes, Phys. Rev. C **54**, 784 (1996).
[2] R. G. Newton, *Scattering of Waves and Particles* (McGraw-Hill, New York, 1966).
[3] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1989).
[4] R. G. Newton, J. Math. Phys. **3**, 75 (1962).
[5] P. C. Sabatier and F. Queyen Van Phu, Phys. Rev. D **4**, 127 (1971); C. Coudray, Lett. Nuovo Cimento **19**, 319 (1977).
[6] M. Münchow and W. Scheid, Phys. Rev. Lett. **44**, 1299 (1980).
[7] K. E. May, M. Münchow, and W. Scheid, Phys. Lett. **141B**, 1 (1984).
[8] K. E. May and W. Scheid, Nucl. Phys. A **466**, 157 (1987); B. Apagyi, A. Ostrowski, W. Scheid, and H. Voit, J. Phys. G **18**, 195 (1992); B. Apagyi, A. Schmidt, W. Scheid, and H. Voit, Phys. Rev. C **49**, 2608 (1994).
[9] H. Leeb, H. Huber, and H. Fiedeldey, Phys. Lett. B **344**, 18 (1995); N. Alexander, K. Amos, B. Apagyi, and D. R. Lun, Phys. Rev. C **53**, 88 (1996); D. R. Lun, and S. J. Buckman, Phys. Rev. Lett. **79**, 541 (1997).
[10] B. Apagyi, P. Lévay, and W. Scheid, in *Inverse and Algebraic Quantum Scattering Theory*, edited by B. Apagyi, G. Endrédi, and P. Lévay (Springer, Berlin, 1997), pp. 158–168.
[11] M. Eberspächer, B. Apagyi, and W. Scheid, Phys. Rev. Lett. **77**, 1921 (1996).
[12] R. Lipperheide and H. Fiedeldey, Z. Phys. A **286**, 45 (1978); Z. Phys. A **301**, 81 (1981); K. Naidoo, H. Fiedeldey, S. A. Sofianos, and R. Lipperheide, Nucl. Phys. A **419**, 13 (1984); S. A. Sofianos, H. Fiedeldey, L. J. Allen, and R. Lipperheide, Phys. Rev. C **31**, 2300 (1985).
[13] E. Kujawski, Phys. Rev. C **6**, 709 (1972); H. Fiedeldey, R. Lipperheide, K. Naidoo, and S. A. Sofianos *ibid.* **30**, 434 (1984); L. J. Allen and H. Burger, Phys. Rev. A **30**, 1237 (1984); L. J. Allen, K. Amos, C. Steward, and H. Fiedeldey, Phys. Rev. C **41**, 2021 (1990); D. R. Lun, K. Amos, and L. J. Allen, Phys. Rev. A **53**, 831 (1996).
[14] A. A. Ioannides and R. S. Mackintosh, Nucl. Phys. A **467**, 482 (1987); S. G. Cooper and R. S. Mackintosh, Inverse Probl. **5**, 707 (1989); S. Shi and H. Rabitz, Comput. Phys. Rep. **10**, 2 (1989); S. G. Cooper and R. S. Mackintosh, Phys. Rev. C **43**, 1001 (1991); S. G. Cooper *ibid.* **50**, 359 (1994).
[15] W. A. Schnizer and H. Leeb, J. Phys. A **26**, 5145 (1993); **27**, 2605 (1994); F. Korinek, H. Leeb, M. Braun, and S. A. Sofianos, in Ref. 10, pp. 177–186.
[16] B. N. Zakhariev and A. A. Suzko, *Direct and Inverse Problems* (Springer Verlag, Berlin, 1990).
[17] H. Kohlhoff, M. Küker, H. Freitag, and H. V. von Geramb, Phys. Scr. **48**, 238 (1993); H. Kohlhoff, thesis, Hamburg, 1993.
[18] S. A. Sofianos, M. Braun, R. Lipperheide, and H. Leeb, in Ref. 10, pp. 54–63.
[19] R. G. Newton, *Inverse Schrödinger Scattering in Three Dimensions* (Springer, Berlin, 1989).
[20] *Quantum Inversion Theory and Applications*, edited by H. V. von Geramb (Springer, Berlin, 1994).
[21] In Ref. 10.
[22] M. Eberspächer, Diploma thesis, Gießen, 1996; M. Eberspächer, B. Apagyi, and W. Scheid, in Ref. 10, pp. 98–111.
[23] R. G. Lovas, Atomki Közlemények **14**, 161 (1972).

A non-Abelian square root of Abelian vertex operators

K. Frieler and K.-H. Rehren[a)]
II. Institut für Theoretische Physik, Universität Hamburg, Germany

(Received 1 December 1997; accepted for publication 5 February 1998)

The quantum field theoretical formulation of Kadanoff's ''correlations along a line'' in the critical two-dimensional Ising model is given in terms of p-products of Wightman fields. It provides a quadratic factorization, in the sense of operator-valued distributions, of Abelian chiral vertex operators into non-Abelian exchange fields. This basic result implies p-polynomial relations between fields from various conformal models in two dimensions; notably the two-dimensional local fields with chiral $SU(2)$ symmetry at level 2 are shown to share the nontrivial peculiarities of the Ising model order and disorder fields. These examples have interesting implications for the properties of p-products in general. © *1998 American Institute of Physics.* [S0022-2488(98)02706-6]

I. INTRODUCTION AND OUTLINE

Wightman distributions can be pointwise multiplied, unlike general distributions, and the product is again a Wightman distribution. This fact (Ref. 1, Chap. V) is due to the spectral properties, i.e., support properties in momentum space, of Wightman fields. The new Wightman field resulting from this prescription is known as the p-product of the factor fields.

The converse is in general not true: the square root of a Wightman distribution is in general not a Wightman distribution. There is, however, at least one famous nontrivial example, which can be retraced to a factorization property of the anisotropic XY model[2] and was put into quantum field theoretical terms by Schroer and Truong.[3] The exponentials of the potential for the axial current of a free Dirac field in two dimensions can be defined (with an appropriate renormalization, see Sec. II A below) and give rise to local Wightman fields. The sine and the cosine, at a specific value of the coefficient in the exponential, turn out to be the p-squares of the order and disorder fields of the Ising model. This result establishes a deep field theoretical relation between the ''doubled'' Ising model and the free Dirac theory. Analytically it means that the vacuum expectation values of the latter can be computed by taking square roots of v.e.v.'s of the former; the double-valuedness of the square root gives rise to the mutual nonlocality of the order and disorder fields.

The purpose of this paper is the discussion of another instance of nontrivial p-product factorization, related to the (nondoubled) Ising field theory, and some of its consequences pertaining to various other models. The basic statement is formula (3.4) which represents a chiral Abelian vertex operator (cf. Sec. II A) as a p-quadratic expression in terms of chiral non-Abelian exchange fields (Sec. II B); the latter are the well-known chiral constituents of the order and disorder fields of the critical Ising field theory.

The p-product of chiral abelian vertex operators is another abelian vertex operator, with a quadratic addition law for the charges. Conversely, the p-factorization of Abelian vertex operators into Abelian vertex operators of smaller charges is an almost trivial fact (cf. Sec. II A). In contrast, the new feature of our basic result is the non-Abelian nature of the constituents, reflected in their non-Abelian braiding and fusion properties.

The p-product is well-defined also for nonlocal fields, as long as they satisfy the remaining Wightman axioms. We shall refer to such fields as nonlocal Wightman fields. Indeed, our basic result (3.4) is in terms of nonlocal (chiral) fields, while some of its applications pertain to local fields in two dimensions.

[a)]Present address: Institut für Theoretische Physik, Universität Göttingen, Germany.

 © 1998 American Institute of Physics

Viewing Wightman fields as operator-valued distributions, the p-product is the pointwise multiplication as a distribution, and the tensor product as an operator. The p-product, denoted by $\Phi\odot\Psi$, of two Wightman fields Φ and Ψ on Hilbert spaces H_Φ and H_Ψ therefore lives on $H_\Phi \otimes H_\Psi$ (its cyclic subspace may be smaller). Symbolically we write

$$(\Phi\odot\Psi)(f) = \int d^d x f(x)\Phi(x)\otimes\Psi(x). \tag{1.1}$$

The precise statement is that the tensor product of Wightman distributions extends to singular smearing functions of the form $f(x)\delta^d(x-y)$ where f is a smooth test function.

The first input for the basic result (cf. Sec. III) is the old observation in the critical Ising model[4] due to Kadanoff (for the magnetization σ alone) and Kadanoff and Ceva (for mixed correlations involving also the disorder parameter μ) that upon restriction along a line in two-dimensional space, the correlations of σ and μ satisfy certain ''Γ-selection rules'' and simplify considerably. Their analytical form becomes that of the correlations of appropriately assigned electric charges in a one-dimensional Coulomb gas.

Translated to the associated quantum field theory in $1+1$ dimensions, this observation is tantamount to a simple identification between the order and disorder fields along a time-like axis and chiral Abelian vertex operators E_α of charge $\alpha=\pm 1/2$ and scaling dimension $h=1/8$. Explicitly, the Kadanoff–Ceva formulas read as

$$\sigma|_{x=0} \cong E_{+1/2}P_+ + E_{-1/2}P_-, \quad \mu|_{x=0} \cong E_{-1/2}P_+ + E_{+1/2}P_-, \tag{1.2a}$$

$$\Leftrightarrow E_{\pm 1/2} \cong \sigma|_{x=0}P_\pm + \mu|_{x=0}P_\mp, \tag{1.2b}$$

where $P_\pm$ are the projections onto the subspaces of integer and half-integer charge. (For the precise notations see Sec. II A below.)

The general theory says that the restriction of a Wightman field in d dimensions to a time-like hyperplane (such as $x^1=0$) yields another Wightman field in one dimension less.[5] For $d=1+1$, the resulting field is a chiral field; the ''positive'' direction (pertinent to the spectrum condition) is inherited from the forward time-like direction.

In order to prevent confusion, we note that the commutation relations obtained by restricting a $1+1$-dimensional field to the line $x=0$ of course test the time-like commutation relations of that field. There is no conflict between σ and μ being local fields and their representation (1.2a) in terms of nonlocal vertex operators.

The second input is the well-known fact that the (local, but not mutually local) $1+1$-dimensional order and disorder fields σ and μ possess a bilinear factorization into chiral exchange fields a and b along with their adjoints.[6] Exchange fields are also known as non-Abelian vertex operators, referring to the non-Abelian superselection structure of the underlying chiral observables. In the present instance, the chiral observables are generated by the stress-energy tensor T with central charge $c=1/2$, and a and b are the primary fields of scaling dimension $h=1/16$,[7] pertaining to various transitions between the three superselection sectors of T. We shall call a, b, a^+ and b^+ the ''elementary fields'' in the sequel. They are nonlocal Wightman fields. (Still, they are relatively local with respect to the chiral observables; in this sense, all nonlocal fields throughout this article are not as nonlocal as could be. Genuinely nonlocal fields without any organizing local observables might be far more intricate objects.)

We have to distinguish the factorization in two independent light-cone coordinates $u=t+x$, $v=t-x$, as mentioned here, from p-factorization described before. Let Φ and Ψ be two chiral (nonlocal) Wightman fields; then the $\times$-product,

$$(\Phi\times\Psi)(f) = \int dt dx f(t,x)\Phi(t+x)\otimes\Psi(t-x), \tag{1.3}$$

defines a $1+1$-dimensional (*a priori* also nonlocal) Wightman field. The $\times$-product is just the ordinary tensor product of operator-valued distributions (except for the passage to light-cone coordinates), and is therefore the tensor product both as a distribution and as an operator.

The bilinear factorization of the order and disorder fields is given in Eqs. (3.1) below. For the present outline of results, the gross structure,

$$(\sigma \quad \text{and} \quad \mu) \sim (\text{linear}) \times (\text{linear}), \tag{1.4}$$

is sufficient. (Here and in the sequel expressions like ''linear'' or ''quadratic'' mean ''linear or p-quadratic polynomial in the elementary fields.'')

The point is now that the $\times$-product, when restricted along a time-like axis (such as $x=0$), turns into the p-product; symbolically,

$$(\Phi \times \Psi)(t,0) = (\Phi \odot \Psi)(t). \tag{1.5}$$

Thus, restricting a $1+1$-dimensional field which is obtained as a $\times$-product of chiral fields produces the p-product of the chiral fields.

We apply this observation to the decomposition (1.4) and insert it into the Kadanoff-Ceva result (1.2b). This yields a representation of the chiral Abelian vertex operators of charge $\pm 1/2$ (scaling dimension $h=1/8$) as a quadratic p-polynomial in the (nonlocal and non-Abelian) elementary fields of scaling dimension $h=1/16$,

$$E_{\pm 1/2} \sim (\text{quadratic in the elementary fields}). \tag{1.6}$$

The exact formula, Eq. (3.4), is the basic result of the present paper. While it is quite surprising as it stands, it becomes even more interesting by its consequences. Namely, we apply it in several ways by exhibiting various familiar fields of various chiral and two-dimensional models as homogeneous p- and $\times$-product combinations in the elementary fields a and b; we thereby reproduce several well-known observations from a unifying point of view, as well as produce several new results.

(1) In Sec. IV A we give another proof for the massless case of the previously mentioned observation[3] in the ''doubled'' Ising model. It relates the p-squares of the order and disorder fields to the axial potential of the massless Dirac theory. The latter has the form $\varphi_R(t-x)-\varphi_L(t+x)$, so its exponentials are $\times$-product of (oppositely charged) chiral Abelian vertex operators. The identifications given in Ref. 3 are the sine and cosine formulas,

$$\sigma_D \equiv \sigma \odot \sigma \cong \frac{1}{\sqrt{2}i}\,(E_{-1/2} \times E_{1/2} - E_{1/2} \times E_{-1/2}), \tag{1.7a}$$

$$\mu_D \equiv \mu \odot \mu \cong \frac{1}{\sqrt{2}}\,(E_{-1/2} \times E_{1/2} + E_{1/2} \times E_{-1/2}). \tag{1.7b}$$

Since σ and μ are $\times$-bilinear in the elementary fields, their p-squares are $\times$-products of p-quadratic expressions in the elementary fields. The latter arrange in such a way that the basic result turns them into Abelian chiral vertex operators, thus reproducing Eqs. (1.7) (up to a unitary similarity transformation, and only on the joint cyclic Hilbert space of the fields under consideration).

(2) We apply our methods to conformal coset models.[8] In these models, the primary fields with respect to a current algebra are factorized into primary fields with respect to some current subalgebra and primary fields with respect to the corresponding coset algebra (''branching''). This separation of degrees of freedom is another example of p-factorization.

In particular, we consider the $U(2)$ currents of a doublet of complex chiral fermions (the primary fields), and its subalgebra of $SU(2)$ currents at level 1 (Sec. II C). The separation of the $SU(2)$ and $U(1)$ ($=$coset) degrees of freedom of the complex fermion doublet[9] is an instance of

the trivial Abelian p-factorization of Abelian vertex operators as mentioned before. It gives rise to the pseudo-real doublet of $SU(2)$-primary fields represented in terms of chiral Abelian vertex operators with scaling dimension $h=1/4$.

In the next step (Sec. IV B), the latter are p-factorized further into two Abelian vertex operators with scaling dimension $h=1/8$, which in turn are p-quadratic in the elementary fields by the basic result. The resulting p-quartic expression for the level 1 primary fields, when written in the form (linear) $\odot$ (cubic), is precisely the coset model factorization with respect to the coset stress-energy tensor (with central charge $c=1/2$) and the level 2 currents, when the level 2 theory is embedded into the tensor product of two level 1 theories.[8]

The primary chiral exchange fields of scaling dimension $h=3/16$ at level 2 are pseudo-real isospin doublets of non-Abelian exchange fields A_i and B_i along with their adjoints, making transitions between the three sectors of the $SU(2)$ current algebra. The previous discussion yields a p-cubic representation in terms of the elementary fields,

$$A_i \quad \text{and} \quad B_i \sim (\text{cubic in the elementary fields}). \tag{1.8}$$

The exact formulas are Eqs. (4.12) below. This representation explains the observation that the solutions to the Knizhnik–Zamolodchikov equation[10] for isospin 1/2 at level 2 are cubic expressions in the $h=1/16$ conformal blocks of the $c=1/2$ minimal model.[7]

(3) In Sec. IV C we consider the defining matrix field g_{ij} of the Wess–Zumino–Witten (WZW) model (=generalized Thirring model)[11] with $SU(2)_L \times SU(2)_R$ symmetry. It is $\times$-bilinear in the primary chiral exchange field doublet (A_i and B_i and their adjoints) of the level 2 $SU(2)$ current algebra, just as σ and μ factorize into the chiral a and b fields.

According to our previous result (1.8), the chiral exchange field doublets are p-cubic in the elementary fields, and hence g_{ij} are of the form (cubic)$\times$(cubic). It is possible to split off a p-quadratic part in the form of Abelian vertex operators $E_{\pm 1/2}$ in either chiral factor, and the remaining p-factor (linear)$\times$(linear) can be identified with σ and μ for the diagonal and off-diagonal matrix elements. The gross structure is, up to some Klein transformation,

$$\begin{aligned} g_{\text{diag}} &\sim \sigma \odot (E_{\pm 1/2} \times E_{\mp 1/2}), \\ g_{\text{offd}} &\sim \mu \odot (E_{\pm 1/2} \times E_{\pm 1/2}). \end{aligned} \tag{1.9}$$

The trace of g involves σ as well as a field which we already know how to identify with $\sigma \odot \sigma$ by Eq. (1.7a); hence

$$\mathrm{Tr}\, g \cong \sqrt{2}(\sigma \odot \sigma \odot \sigma). \tag{1.10}$$

Next, we restrict g_{ij} to a time-like axis and obtain a p-polynomial of order six in the elementary fields. By Eqs. (1.2a) and (1.9), these polynomials turn into cubic p-products of vertex operators $E_{\pm 1/2}$. The resulting representation (4.19) of $g_{ij|x=0}$ by Abelian chiral vertex operators with scaling dimension 3/8 generalizes Kadanoff's result (1.2a) to the WZW model.

(4) In Sec. IV D we consider p-quadratic expressions in the $1+1$-dimensional matrix field g_{ij}, notably the p-determinant. In terms of the elementary fields, this determinant field is of the form (order six)$\times$(order six). Again arranging p-quadratic polynomials in either chiral factor to yield Abelian vertex operators according to the basic result, and performing some astute similarity transformations, the p-determinant field is finally found to be

$$\mathrm{Det}_{\odot} g \cong \frac{1}{\sqrt{2}i} (E_{-\beta} \times E_{\beta} - E_{\beta} \times E_{-\beta}), \tag{1.11}$$

with charge $\beta = \sqrt{3}/2$ [rather than 1/2 in Eq. (1.7a)]. This generalizes the ''doubling'' result[3] (in the massless case) to the WZW model.

A technical comment might be in order. Throughout the paper, we prove identities between various fields only ''on the cyclic Hilbert space, and up to a unitary similarity transformation which preserves the vacuum vector,'' with the notation $\cong$. This means, of course, no weakening of the statement, since the vacuum expectation values which determine the fields are never af-

fected. A system of fields must of course be subjected to the same transformation. One should, however, keep in mind other fields which are possibly present, and which might or might not be transformed at the same time. Notably, the property of being primary with respect to some chiral stress-energy tensor or current algebra is defined by specific commutation relations with these chiral observables, and will be preserved only if the latter are. This is, e.g., guaranteed if the unitary operator which implements the similarity transformation is a constant on each irreducible representation space of the chiral observables. Otherwise, the latter would have to be transformed along with the primary fields. Even if the state of affairs is not explicitly mentioned in every single instance, it is contextually evident everywhere in the paper by inspection of the specific form of the similarity transformation at hand.

The present work contains the results of the diploma thesis of the first author (KF).[12]

II. PREPARATION OF THE GROUND

A. Abelian vertex operators

Unlike the massless free scalar field in $1+1$ dimensions, its exponentials exist on a positive definite Hilbert space.[1,13–15] Mandelstam[16] has introduced such exponentials depending on only one chiral coordinate when he bosonized the massless Thirring model (at a vanishing coupling constant) in terms of the sine-Gordon model (at $\beta^2=4\pi$). These fields, denoted by $E_\alpha = {:}\exp i\alpha\varphi{:}$ in the sequel, are now known as (Abelian) chiral vertex operators. The ''triple ordering'' [13] refers to the creation and annihilation parts of the auxiliary chiral scalar field and involves a renormalization parameter μ of the dimension of mass which is understood in the limit $\mu \searrow 0$. For rather streamlined treatments see, e.g., Refs. 14 and 17.

The complex free chiral fermion field ψ can be identified with a vertex operator of unit charge $\alpha=1$. Its current $j=:\psi\psi^*:=\partial\varphi$ generates a $U(1)$ symmetry with charge operator $Q=1/(2\pi)\int j(x)dx$. By virtue of

$$QE_\alpha = E_\alpha(Q+\alpha), \tag{2.1}$$

the $U(1)$ symmetry extends to the chiral vertex operators of arbitrary charge α.

Chiral vertex operators of general charge are nonlocal chiral Wightman fields, satisfying anyonic commutation relations,

$$E_\alpha(x)E_\beta(y) = e^{\pm i\pi\alpha\beta}E_\beta(y)E_\alpha(x), \quad \text{if } x\neq y, \tag{2.2}$$

with the sign $\pm=\mathrm{sign}(x-y)$ in the exponent. They are conformally covariant fields with scaling dimension $h=\alpha^2/2$ with respect to the stress-energy tensor $i/(4\pi):\psi\overleftrightarrow{\partial}\psi^*:=1/(4\pi):j^2:$ of the complex fermion (with central charge $c=1$).

The vacuum expectation values of chiral vertex operators satisfy charge conservation and are given by

$$\langle\Omega, E_{\alpha_1}(x_1)\ldots E_{\alpha_n}(x_n)\Omega\rangle = \prod_{i<j}\Delta(x_i-x_j)^{-\alpha_i\alpha_j}, \quad \text{provided } \sum \alpha_i = 0. \tag{2.3}$$

The distributions,

$$\Delta(x-y)^{2h} = \left[\frac{-i}{x-y-i\epsilon}\right]^{2h} = \frac{1}{\Gamma(2h)}\int_0^\infty dk k^{2h-1}e^{-ik(x-y)}, \tag{2.4}$$

contributing to these v.e.v.'s have spectral support on the positive momenta.

Abelian chiral vertex operators with multi-component charges can be defined as exponentials of several independent massless free fields, e.g.,

$$E_{(\alpha,\beta)} = {:}\exp i(\alpha\varphi_1+\beta\varphi_2){:} \cong E_\alpha \odot E_\beta. \tag{2.5}$$

Their vacuum expectation values are given by the same formula (2.3), replacing only products of charges by scalar products of multi-component charges. This replacement results in the factorization of the total v.e.v.'s into the v.e.v.'s for the charge components and therefore justifies the equivalence with the p-product in Eq. (2.5)

On the other hand, it implies an obvious $O(N)$ symmetry for the vertex operators with N-component charges, corresponding to rotations and reflections in charge space. Therefore on its own cyclic Hilbert space, the field (2.5) is also equivalent to $E_{\sqrt{\alpha^2+\beta^2}}$.

B. Exchange fields

Chiral exchange fields arise by a spectral decomposition with respect to global conformal transformations of $1+1$-dimensional fields with conformal symmetry[18] and subsequent chiral factorization. They arise also as point-like limits of bounded intertwining operators in the framework of algebraic quantum field theory,[19,20] or as intertwiners between modules of chiral symmetry algebras such as Virasoro or affine Kac-Moody algebras.[21] They are the fundamental charge-carrying entities which make transitions between the various superselection sectors of a theory of local chiral observables. As such, they are defined on a ''source'' Hilbert space and take values in a ''range'' Hilbert space which are both irreducible sectors of the observables, and are relatively local with respect to the latter.

In specific models where the chiral observables are generated by the stress-energy tensor or by a current algebra, it is convenient to distinguish ''primary'' exchange fields by their specific commutation relations with the observables. These imply certain partial differential equations, reflecting the symmetry of the model, for the vacuum expectation values of primary exchange fields whose solutions are known as ''conformal blocks.'' [7,10] The commutation relations of chiral exchange fields with each other are given by a non-Abelian representation of the braid group. Abelian vertex operators (restricted to a charge sector) are the exchange fields with respect to an Abelian current algebra.

The primary chiral exchange fields of the Ising model were first treated in detail in Ref. 6. In this model, the chiral observables are generated by the stress-energy tensor with central charge $c=1/2$. There are three superselection sectors, which carry representations of the conformal group with primary scaling dimensions 0, 1/16 and 1/2, respectively.

The primary field for the trivial sector 0 is just the unit operator. The primary field for the sector 1/2 is a real free fermion. There are three exchange fields corresponding to this primary field, which are given by the free fermion field in the Neveu-Schwarz representation restricted to the subspaces of even and odd Fermi number, and the free Fermi field in the Ramond representation. The former two are each other's adjoints $\psi_{\mathrm{NS}}:\mathscr{H}_0\to\mathscr{H}_{1/2}$ and $\psi_{\mathrm{NS}}^+:\mathscr{H}_{1/2}\to\mathscr{H}_0$, and the latter is a self-adjoint field $\psi_{\mathrm{R}}:\mathscr{H}_{1/16}\to\mathscr{H}_{1/16}$.

The primary exchange fields for the sector 1/16 are the chiral constituents of the order and disorder parameter of the $1+1$-dimensional Ising model. There are two ''raising'' operators a and b along with their adjoints:

$$\begin{aligned} &a:\mathscr{H}_0\to\mathscr{H}_{1/16} \quad \text{and its adjoint} \quad a^+:\mathscr{H}_{1/16}\to\mathscr{H}_0,\\ &b:\mathscr{H}_{1/16}\to\mathscr{H}_{1/2} \quad \text{and its adjoint} \quad b^+:\mathscr{H}_{1/2}\to\mathscr{H}_{1/16}. \end{aligned} \tag{2.6}$$

The commutation relations of these chiral exchange fields are given by a non-Abelian representation of the braid group of the Hecke type.[6]

We have chosen the normalization such that the two-point function is $\langle\Omega,a^+(x)a(y)\Omega\rangle = \Delta(x-y)^{1/8}$. The only possible four-point functions are given by

$$\begin{aligned} \langle\Omega,a^+(x_1)a(x_2)a^+(x_3)a(x_4)\Omega\rangle &= V^{1/8}\cdot f(x),\\ \langle\Omega,a^+(x_1)b^+(x_2)b(x_3)a(x_4)\Omega\rangle &= V^{1/8}\cdot g(x), \end{aligned} \tag{2.7}$$

where the first factor with

$$V(x_1,x_2,x_3,x_4)=\frac{\Delta(x_1-x_2)\Delta(x_3-x_4)\Delta(x_1-x_4)\Delta(x_2-x_3)}{\Delta(x_1-x_3)\Delta(x_2-x_4)} \tag{2.8}$$

is the four-point function of Abelian vertex operators with alternating charges and the proper scaling dimension 1/16, and the ''conformal block'' functions,

$$f(x)=\sqrt{\tfrac{1}{2}\,(1+\sqrt{1-x})}\quad\text{and}\quad g(x)=\sqrt{\tfrac{1}{2}\,(1-\sqrt{1-x})} \tag{2.9}$$

depend only on the conformally invariant cross ratio $x=\Delta(x_1-x_3)\Delta(x_2-x_4)/\Delta(x_1-x_2)\Delta(x_3-x_4)$.

The higher $2n$-point functions were derived in closed form in Ref. 6.

C. The $SU(2)$ primary fields at level 1 in terms of vertex operators

A doublet of complex free chiral fermions $\psi_\pm$ gives rise to a theory with $U(2)$ symmetry. Its stress-energy tensor with $c=2$ is given by $T=(i/4\pi)(:\psi_+\overleftrightarrow{\partial}\psi_+^*:+:\psi_-\overleftrightarrow{\partial}\psi_-^*:)$, and the currents are

$$j^A=:\psi_i\tau^A_{ij}\psi_j^*:\quad (A=0,1,2,3), \tag{2.10}$$

where $\tau^0=\mathbb{1}$ and τ^a $(a=1,2,3)$ are the Pauli matrices.

The component j^0 generates the Abelian $U(1)$ current algebra with stress-energy tensor $T_U=1/(8\pi):(j^0)^2:(c=1)$, and j^a generate the $SU(2)$ current algebra at level 1 with stress-energy tensor T_S given by the Sugawara formula quadratic in the currents $(c=1)$, such that $T=T_U+T_S$. The $U(1)$ and the $SU(2)$ currents commute with each other, and both transform the fermion doublet in the obvious way.

The $U(1)$ and the $SU(2)$ degrees of freedom of the fermion doublet can be separated as follows in a very explicit manner.[9] The crucial point is the fact that chiral charged fermions can be represented by Abelian vertex operators of unit charge.

The doublet of charged chiral fermions can be represented by a pair of Abelian vertex operators with orthogonal unit charge vectors $\alpha_\pm$. Since the latter constitute a pair of commuting fermions, a Klein transformation is needed to make them anti-commute. Rather than the standard choice [multiplying, e.g., E_{α_+} by the exponential $\exp(i\pi\alpha_- Q)$ of the charge operator of the other component], we prefer a more symmetric one and multiply both components by $\exp[i\pi(\alpha_+-\alpha_-)Q/2]$. Both choices are unitarily equivalent by conjugation with the exponential of some quadratic polynomial of the charge operators.

We choose $\alpha_\pm=1/\sqrt{2}\cdot(1,\pm 1)$ and apply the p-factorization of multi-component vertex operators [Eq. (2.5)]. This gives rise to the simple p-factorization,

$$\psi_\pm=\phi\odot\phi_\pm, \tag{2.11}$$

where

$$\phi=E_\alpha \tag{2.12}$$

is an ordinary Abelian vertex operator of charge $\alpha=1/\sqrt{2}$ carrying the $U(1)$ symmetry, and

$$\phi_\pm=E_{\pm\alpha}e^{i\pi\alpha Q} \tag{2.13}$$

is a pseudo-real doublet carrying the $SU(2)$ symmetry. Pseudo-reality means

$$\phi_i^*=i\epsilon_{ij}\phi_j C=-i\epsilon_{ij}C\phi_j, \tag{2.14}$$

where $\epsilon_{+-}=1$ and $C=e^{2\pi i\alpha Q}=C^{-1}$ is a Casimir operator which equals $+1$ (-1) on all states of integer (half-integer) isospin and thus commutes with the $SU(2)$ symmetry.

In this representation, the factorization of the current algebra becomes manifest: the $U(1)$ current is $j^0=\sqrt{2}\partial\varphi\otimes\mathbb{1}$, and the $SU(2)$ currents are found in the Frenkel–Kac[22] representation $j^3=\mathbb{1}\otimes\sqrt{2}\partial\varphi$ and $j^\pm=\mathbb{1}\otimes E_{\pm\sqrt{2}}$.

Both the $U(1)$ field and the $SU(2)$ doublet are primary fields w.r.t. the respective current algebra, and have scaling dimension $h=1/4$. They both satisfy anyonic commutation relations according to Eq. (2.2), modified for the doublet by the Klein factor $e^{i\pi\alpha Q}$. The two-point functions are $\langle\Omega,\phi^*\phi\Omega\rangle=\Delta^{1/2}$ and $\langle\Omega,\phi_i^*\phi_j\Omega\rangle=\delta_{ij}\Delta^{1/2}$ or equivalently $\langle\Omega,\phi_i\phi_j\Omega\rangle=-i\epsilon_{ij}\Delta^{1/2}$.

The cyclic Hilbert space of the pseudo-real doublet contains all charges which are integer multiples of $\alpha=1/\sqrt{2}$. The primary exchange fields which interpolate the two sectors of the $SU(2)$ current algebra at level 1 (integer and half-integer isospin), are obtained by restricting $\phi_\pm$ to the subspaces of even or odd charge in units of α, or equivalently of integer or half-integer isospin.

D. Goddard–Kent–Olive (GKO) coset construction

Consider two current algebras $\mathscr{A}_i$ for some simple Lie group G at level k_i generated by the currents j_i^a, and with stress-energy tensor T_i given by the affine Sugawara formula. The diagonal currents,

$$J^a=j_1^a\otimes\mathbb{1}+\mathbb{1}\otimes j_2^a, \tag{2.15}$$

generate a current algebra $\mathscr{A}$ at level k_1+k_2 within $\mathscr{A}_1\otimes\mathscr{A}_2$. This current algebra has its own stress-energy tensor T given by the Sugawara formula in terms of J^a which has the same commutation relations with J^a as the total stress-energy tensor $T_1\otimes\mathbb{1}+\mathbb{1}\otimes T_2$. The difference T_{coset} is a stress-energy tensor which commutes with the current algebra $\mathscr{A}$. It belongs (possibly along with other operators) to the coset model $\mathscr{C}$ of fields in $\mathscr{A}_1\otimes\mathscr{A}_2$ commuting with $\mathscr{A}$.

Embedding a primary exchange field at level k_1 by the prescription $\phi\otimes\mathbb{1}$ into the Hilbert space of $\mathscr{A}_1\otimes\mathscr{A}_2$, yields operators which have primary commutation relations with both the diagonal currents J^a and the coset stress-energy tensor. Therefore, according to the same scheme of separation of degrees of freedom as in Sec. II C, there is a factorization of the form

$$\phi\otimes\mathbb{1}\cong\sum\ \varphi_c\odot\varphi, \tag{2.16}$$

where φ_c are primary exchange fields w.r.t. the coset theory and φ are primary exchange fields w.r.t. the current algebra at level k_1+k_2. The quantum numbers of the exchange fields on the right hand side are determined by the branching rules of representations of $\mathscr{A}_1\otimes\mathscr{A}_2$ in restriction to $\mathscr{C}\otimes\mathscr{A}$; in particular, φ carries the G quantum numbers, and the scaling dimensions are additive.

The symbol $\odot$ on the right hand side of the GKO decomposition (2.16) refers to the tensor product of representation spaces of $\mathscr{C}\otimes\mathscr{A}$, while in contrast, on the left hand side $\otimes$ refers to the tensor product $\mathscr{A}_1\otimes\mathscr{A}_2$. The passage is made by a reorganization of the Hilbert space which, in a special case, will become explicit along the way in Sec. IV B.

III. THE BASIC RESULT

Chiral exchange fields can be combined in various ways to yield local $1+1$-dimensional fields. A systematic way to do so was described in Ref. 23 in terms of the algebraic theory of superselection sectors. (Helas, the assignment of σ and μ was interchanged in Ref. 23, and no care was taken to construct local fields which are closed under conjugation; in order to repair that defect, the self-adjoint local fields below differ from the result in Ref. 23 by a nonunitary similarity transformation. One can actually adjust the correct formulas ''by hand'' from the known commutation relations and four-point functions of the elementary a and b fields[6] by imposing the correct commutation relations[3] and four-point functions[6] for σ and μ.)

The upshot is the following. The local field σ interpolates between the ''diagonal'' charge sectors $\mathscr{H}_{ss}\equiv\mathscr{H}_s\otimes\mathscr{H}_s$ where $\mathscr{H}_s$ $(s=0,\frac{1}{16},\frac{1}{2})$ are the three superselection sectors of the chiral observables. It is given by the standard formula $\sigma=a\times a+b\times b+(\text{h.c.})$. The field μ, however carries an ''excess $\mathbb{Z}_2$ charge'' which makes transitions to the ''nondiagonal'' sectors $\mathscr{H}_{0(1/2)}$, $\mathscr{H}_{(1/2)0}$ and another copy of $\mathscr{H}_{(1/16)(1/16)}$. In order to describe both fields simultaneously, σ has of course to be defined also on the nondiagonal sectors. There is some ambiguity on how to define these fields which consists in unitary similarity transformations (which commute with the chiral

observables in order not to spoil the primary commutation relations). These unitaries are given by a complex phase on each of the four simple sectors, and a unitary 2×2 matrix which mixes the two copies of $\mathscr{H}_{(1/16)(1/16)}$.

Our representation of choice is the following:

$$\sigma = u\otimes(a\times a + b^+\times b^+) + v\otimes(a\times b^+ + b^+\times a) + (\text{h.c.}). \tag{3.1a}$$

$$\mu = v\times(a\times a - b^+\times b^+) + iu\otimes(a\times b^+ - b^+\times a) + (\text{h.c.}). \tag{3.1b}$$

We introduced two orthogonal unit vectors u and v in $\mathbb{C}^2$ in order to distinguish the two copies of $\mathscr{H}_{(1/16)(1/16)}$ within $\mathscr{H}_{(1/16)(1/16)}\oplus\mathscr{H}_{(1/16)(1/16)} = \mathbb{C}^2\otimes\mathscr{H}_{(1/16)(1/16)}$; they are multiplied like rectangular matrices. Obviously, σ preserves both $\mathscr{H}_{00}\oplus(u\otimes\mathscr{H}_{(1/16)(1/16)})\oplus\mathscr{H}_{(1/2)(1/2)}$ and $\mathscr{H}_{0(1/2)}\oplus(v\otimes\mathscr{H}_{(1/16)(1/16)})\oplus\mathscr{H}_{(1/2)0}$, while μ alternates between these spaces. Conversely, μ preserves the analogous spaces with u and v interchanged, while σ alternates between them. The following diagram summarizes the situation: σ interpolates according to the horizontal arrows, while μ interpolates according to the diagonal arrows:

$$\begin{array}{ccccc} (00) & \longleftrightarrow & u\otimes(\frac{1}{16}\frac{1}{16}) & \longleftrightarrow & (\frac{1}{2}\frac{1}{2}) \\ & \nwarrow\!\!\searrow & & \nwarrow\!\!\searrow & \\ (0\tfrac{1}{2}) & \longleftrightarrow & v\otimes(\frac{1}{16}\frac{1}{16}) & \longleftrightarrow & (\tfrac{1}{2}0) \end{array} \tag{3.2}$$

Every single term in (3.1) corresponds to an arrow in (3.2).

Our normalization is such that the two-point functions are $\langle\Omega,\sigma\sigma\Omega\rangle = \langle\Omega,\mu\mu\Omega\rangle = (\Delta_L\Delta_R)^{1/8}$ in terms of the left- and right-moving light-cone coordinates. All mixed four-point functions,[6] e.g., $\langle\Omega,\mu\mu\sigma\sigma\Omega\rangle = (V_LV_R)^{1/8}(f_Lf_R - g_Lg_R)$, can be read off Eqs. (2.7) and (3.1).

Each of the fields σ and μ satisfies local commutativity at a space-like distance with itself, but they are not mutually local. Instead, they satisfy the ''dual'' commutation relations,[4]

$$\sigma(x)\mu(y) = \pm\mu(y)\sigma(x), \quad \text{if} \quad (x-y)^2<0, \tag{3.3}$$

with the sign $\pm = \text{sign}\,(x^1-y^1)$.

Let us consider now the Kadanoff-Ceva formulas (1.2). Evaluating the fields σ and μ at $x=0$ means replacing the $\times$-product in (3.1) by the p-product $\odot$. The projections P_+ and P_- project onto the subspaces generated from the vacuum by an even, resp., odd number of applications of the fields σ and μ, i.e., P_+ projects onto $\mathscr{H}_+ = \oplus_{s,t=0,1/2}\,\mathscr{H}_{st}$ while P_- projects onto $\mathscr{H}_- = \mathbb{C}^2\otimes\mathscr{H}_{(1/16)(1/16)}$.

Thus inserting (3.1) into (1.2b) yields the basic result,

$$\begin{aligned} E_{+1/2} = {} & u\otimes(a\odot a + b^+\odot b^+) - iu^+\otimes(a^+\odot b - b\odot a^+) \\ & + v\otimes(a\odot b^+ + b^+\odot a) + v^+\otimes(a^+\odot a^+ - b\odot b), \end{aligned} \tag{3.4}$$

and $E_{-1/2} = E^*_{+1/2}$

Like all formulas below, Eq. (3.4) can easily be checked in all mixed four-point functions of $E_{\pm 1/2}$, using Eqs. (2.3) and (2.7). The important point is, however, its validity on the entire joint cyclic Hilbert space of these fields.

The charge conjugation invariance $E_{+1/2}\leftrightarrow E_{-1/2}$ is manifest in this representation. It is implemented by the involutive unitary operator $\Pi\cdot(P_{00} + iP_{0(1/2)} - iP_{(1/2)0} - P_{(1/2)(1/2)} + M\otimes P_{(1/16)(1/16)})$ involving the discrete ''flip'' operator Π: $\mathscr{H}_{st}\to\mathscr{H}_{ts}$ which interchanges the two tensor factors, and a matrix M in the multiplicity space of $\mathscr{H}_{(1/16)(1/16)}$ which interchanges u and v.

Equation (3.4) shows a $\mathbb{Z}_4$ charge structure as follows. We redisplay the diagram (3.2) with only the arrows pertaining to the combinations $\varepsilon := \sigma P_+ + \mu P_-$ or $E_{+1/2} \equiv \varepsilon|_{x=0}$:

$$\begin{array}{ccccc} (00) & \longrightarrow & u \otimes (\frac{1}{16}\frac{1}{16}) & \longleftarrow & (\frac{1}{2}\frac{1}{2}) \\ & \nwarrow\!\!\searrow & & \swarrow\!\!\nearrow & \\ (0\frac{1}{2}) & \longrightarrow & v \otimes (\frac{1}{16}\frac{1}{16}) & \longleftarrow & (\frac{1}{2}0) \end{array} \tag{3.5}$$

It becomes apparent that the subspaces H_i given by

$$\begin{aligned} H_0 &= \mathscr{H}_{00} \oplus \mathscr{H}_{(1/2)(1/2)}, \quad H_1 = u \otimes \mathscr{H}_{(1/16)(1/16)}, \\ H_2 &= \mathscr{H}_{0(1/2)} \oplus \mathscr{H}_{(1/2)0}, \quad H_3 = v \otimes \mathscr{H}_{(1/16)(1/16)}, \end{aligned} \tag{3.6}$$

are the subspaces of charge $(i \bmod 4)\frac{1}{2}$. The four terms in Eq. (3.4) correspond in turn to $P_{i+1} E_{+1/2} P_i = E_{+1/2} P_i$, $i = 0,1,2,3$.

In terms of these subspaces we have the interpolation diagrams

$$\sigma: \begin{cases} H_0 \leftrightarrow H_1 \\ H_3 \leftrightarrow H_2, \end{cases} \quad \mu: \begin{cases} H_0 & H_1 \\ \updownarrow & \updownarrow, \\ H_3 & H_2 \end{cases} \quad \varepsilon, E_{1/2}: \begin{cases} H_0 & \rightarrow & H_1 \\ \uparrow & & \downarrow. \\ H_3 & \leftarrow & H_2 \end{cases} \tag{3.7}$$

The $\mathbb{Z}_4$ symmetry of the $1+1$-dimensional field ε together with the two $\mathbb{Z}_2$'s $\sigma \to -\sigma$ and $\mu \to -\mu$ generate a total discrete symmetry $\mathbb{Z}_4 \times \mathbb{Z}_2$ (the diagonal of the two $\mathbb{Z}_2$'s coinciding with the $\mathbb{Z}_2$ subgroup of $\mathbb{Z}_4$). Upon restriction to the line, ε turns into the chiral field $E_{1/2}$ and the $\mathbb{Z}_4$ symmetry is enhanced to U(1), corresponding to the Kadanoff–Ceva Γ-charge.

In this picture, the real chiral Ising fermions ψ_L and ψ_R arising in the operator product expansions of $\mu\sigma$ and $\sigma\mu$ carry two units of $\mathbb{Z}_4$ charge.

IV. APPLICATIONS

A. The doubled Ising model

In contrast to the order and disorder fields $\varphi = \sigma, \mu$ with their dual commutation relations (3.3), the ''doubled'' fields $\varphi_D = \varphi \odot \varphi$ satisfy local commutativity also with each other. The asserted identifications (1.7) which we are going to establish are equivalent to

$$\mu_D + i\sigma_D \cong \sqrt{2} \cdot E_{-1/2} \times E_{1/2}, \tag{4.1}$$

where the fields on the right hand side are also local fields made from anyonic chiral constituents.[17]

We compute $E_{-1/2} \times E_{1/2}$ from the basic result (3.4). Its cyclic Hilbert space involves only the conjugate charge sectors $H_{-i} \otimes H_i$. For the vacuum expectation values, we may therefore omit all terms which are defined on $H_j \otimes H_i$, $j \neq -i$. The remaining terms are

$$\begin{aligned} E_{-1/2} \times E_{1/2} \cong\ & Y \otimes (aa - b^+b^+) \times (aa + b^+b^+) + iX \otimes (ab^+ - b^+a) \times (ab^+ + b^+a) \\ & + X^+ \otimes (a^+a^+ + bb) \times (a^+a^+ - bb) - iY^+ \otimes (a^+b + ba^+) \times (a^+b - ba^+), \end{aligned} \tag{4.2}$$

on the cyclic subspace. We omitted the $\odot$ symbols, and put $X = u \otimes v$, $Y = v \otimes u$.

On the other hand, we compute the doubled fields from Eqs. (3.1). Each of the fields σ and μ, according to the scheme (3.7), takes each of the subspaces H_i into a specific H_j. Hence repeated application of the doubled fields connects the vacuum sector $H_0 \otimes H_0$ only with the diagonal sectors $H_i \otimes H_i$, and the latter again exhaust the joint cyclic Hilbert space. Thus we obtain

$$\begin{aligned}\mu_D+i\sigma_D\cong&(V+iU)\otimes(aa\times aa+b^+b^+\times b^+b^++iaa\times b^+b^++ib^+b^+\times aa)\\&-(V-iU)\otimes(ab^+\times ab^++b^+a\times b^+a-iab^+\times b^+a-ib^+a\times ab^+)\\&+(V^++iU^+)\otimes(a^+a^+\times a^+a^++bb\times bb+ia^+a^+\times bb+ibb\times a^+a^+)\\&-(V^+-iU^+)\otimes(a^+b\times a^+b+ba^+\times ba^+-ia^+b\times ba^+-iba^+\times a^+b),\end{aligned}\tag{4.3}$$

on the joint cyclic subspace. We put $U=u\otimes u$, $V=v\otimes v$ and arranged the factors as $(A\times B)\odot(C\times D)\simeq(A\odot C)\times(B\odot D)$. Note that due to the different factor ordering, the individual subspaces $H_j\otimes H_i$ in Eq. (4.2) are different from $H_j\otimes H_i$ in Eq. (4.3).

Equation (4.3) is transformed into $\sqrt{2}$ times Eq. (4.2) by conjugation with a unitary operator which preserves the vacuum vector. The operator which does the job equals (in an obvious notation) $(P_{00\times00}-P_{(1/2)(1/2)\times(1/2)(1/2)}+P_{0(1/2)\times(1/2)0}-P_{(1/2)0\times0(1/2)})+i(P_{00\times(1/2)(1/2)}-P_{(1/2)(1/2)\times00}+P_{0(1/2)\times0(1/2)}-P_{(1/2)0\times(1/2)0})+M\otimes P_{(1/16)(1/16)\times(1/16)(1/16)}$, with a unitary matrix M in the multiplicity space of $\mathscr{K}_{(1/16)(1/16)}\otimes\mathscr{K}_{(1/16)(1/16)}$ taking $V+iU\mapsto\sqrt{2}Y$ and $V-iU\mapsto\sqrt{2}X$. This similarity transformation commutes with the chiral observables of the doubled Ising field theory (two stress-energy tensors with central charge $c=1/2$ on each light-cone), and therefore preserves the structure of the full model.

It might be objected that one obtains apparently different results when one restricts the doubled fields according to Eqs. (1.7) to the line $x=0$, and when one doubles the restricted fields according to Eqs. (1.2a); namely,

$$\begin{aligned}\sigma_D|_{x=0}&=\frac{1}{\sqrt{2}i}\,(E_\alpha-E_{-\alpha}) & (\sigma|_{x=0})_D&=E_\alpha P_++E_{-\alpha}P_-,\\ &\text{vs}\\ \mu_D|_{x=0}&=\frac{1}{\sqrt{2}}\,(E_\alpha+E_{-\alpha}) & (\mu|_{x=0})_D&=E_{-\alpha}P_++E_\alpha P_-,\end{aligned}\tag{4.4}$$

where $P_\pm$ now project on the spaces of even, resp., odd charge in units of $\alpha=1/\sqrt{2}$. In fact, the latter form can be transformed into the former by a chain of similarity transformations exploiting $SU(2)$ invariance. It is implemented by the unitary operator $\mathscr{T}e^{-i\pi(\alpha Q)^2}\mathscr{U}e^{i\pi(\alpha Q)^2}$ where first $e^{i\pi(\alpha Q)^2}$ transforms the Abelian vertex operators into the pseudo-real $SU(2)$ doublet $\phi_\pm$ along with the $SU(2)$-invariant Casimir operator $C=P_+-P_-$ (cf. Sec. II C); next the $SU(2)$ transformation $\mathscr{U}$ takes $\phi_\pm\mapsto\sqrt{(\mp i/2)}(\phi_+\mp\phi_-)$, then $e^{-i\pi(\alpha Q)^2}$ reproduces the vertex operators, and finally $\mathscr{T}=P_++\sqrt{-i}P_-$ (which commutes with the other three operators) removes the charge dependence. We refrain from an explicit demonstration, since a similar transformation will be discussed in detail later (Sec. IV D).

B. The $SU(2)$ primary fields at level 2

We consider the GKO coset $(SU(2)_1\times SU(2)_1)/SU(2)_2$. The coset algebra is the chiral stress-energy tensor with $c=1/2$ whose primary exchange fields were described in Sec. II B.

The level 2 current algebra has also three sectors $\mathscr{K}_I$ distinguished by the isospin $I=0,1/2,1$ of the ground states of the conformal Hamiltonian. The primary exchange fields of isospin $I=1/2$ are four $SU(2)$ doublets which make transitions in analogy to Eqs. (2.6):

$$\begin{aligned}A_i&:\mathscr{K}_0\to\mathscr{K}_{1/2}\quad\text{and}\quad A_i^+:\mathscr{K}_{1/2}\to\mathscr{K}_0,\\ B_i&:\mathscr{K}_{1/2}\to\mathscr{K}_1\quad\text{and}\quad B_i^+:\mathscr{K}_1\to\mathscr{K}_{1/2}.\end{aligned}\tag{4.5}$$

The fields A_i^+ and B_i^+ are not the adjoints of A_i and B_i, since the latter transform in the conjugate representation. We may, however, choose the normalizations such that

$$X_i^*=-\epsilon_{ij}X_j^+,\quad (X_i^+)^*=\epsilon_{ij}X_j\quad(X=A,B,\ \ i,j=\pm).\tag{4.6}$$

Like the elementary fields, the chiral exchange fields of the $SU(2)$ level 2 current algebra satisfy braid group commutation relations given by another non-Abelian Hecke type representation.

The two-point function is $\langle\Omega,A_i^+A_j\Omega\rangle=\epsilon_{ij}\Delta^{3/8}$. The only nonvanishing four-point functions of these exchange fields have the structure $\langle\Omega,A_i^+A_jA_k^+A_l\Omega\rangle=V^{3/8}\cdot F_{ijkl}(x)$ and $\langle\Omega,A_i^+B_j^+B_kA_l\Omega\rangle=V^{3/8}\cdot G_{ijkl}(x)$, with V and x as in Sec. II B. The $SU(2)$ tensors F and G are the two independent ''conformal block'' solutions to the Knizhnik-Zamolodchikov differential equation,

$$\left[(k+2)\partial_x-\frac{C_{12}}{x}+\frac{C_{23}}{1-x}\right]W(x)=0, \tag{4.7}$$

where $k=2$ is the level and $C_{mn}=\Sigma_a(I_m^a+I_n^a)^2$ are the quadratic $SU(2)$ Casimir operators on the product representations due to the m-th and n-th field entry. The solutions

$$\begin{aligned} F_{ijkl}&=\epsilon_{ij}\epsilon_{kl}\cdot f(f^2+g^2)-\epsilon_{ik}\epsilon_{jl}\cdot 2\,fg^2,\\ G_{ijkl}&=-\epsilon_{ij}\epsilon_{kl}\cdot g(f^2+g^2)+\epsilon_{ik}\epsilon_{jl}\cdot 2\,f^2g, \end{aligned} \tag{4.8}$$

involve the same functions $f(x)$ and $g(x)$ as in Eq. (2.9).

Our intention is to give an explanation for the cubic structure of these vacuum expectation values in terms of the v.e.v.'s of the elementary fields. For this purpose, we recall how the level 2 primary exchange fields can be obtained by the GKO coset construction from the level 1 primaries $\phi_\pm\otimes\mathbb{1}$.

According to the general discussion in Sec. II D, the branching rules (in particular, the $SU(2)$ quantum numbers and scaling dimensions) determine the possible terms in the coset factorization of $\phi_\pm\otimes\mathbb{1}$. The result is

$$\phi_i\otimes\mathbb{1}=a\odot A_i-ia^+\odot A_i^++b\odot B_i+ib^+\odot B_i^+\,. \tag{4.9}$$

Here the absolute coefficients 1 in front of the first and third term are chosen by a normalization freedom, and the relative coefficients $\pm i$ in front of the adjoint terms are determined by the pseudo-reality property (2.14) of ϕ_i along with Eq. (4.6). Equation (4.9) can be cross-checked by inserting the four-point functions (2.7) and (4.8) and comparing with the four-point functions of $\phi_\pm$ which are computed with the help of Eqs. (2.13) and (2.3).

Now comes the crucial step. According to Eq. (2.13), $\phi_\pm$ have a representation in terms of vertex operators of charge $\pm\alpha$ with $\alpha^2=\frac{1}{2}$. According to Eq. (2.5) with the choice $\alpha=(\frac{1}{2},\frac{1}{2})$, the latter can be written in the form $E_{\pm1/2}\odot E_{\pm1/2}$, and according to Eq. (3.4), the vertex operators of charge $\pm\frac{1}{2}$ can be written as p-quadratic polynomials in the elementary exchange fields. The operator $e^{i\pi\alpha Q}$ in (2.13) takes values i^p in the sectors of charge $p\alpha$, and the charge quantum numbers p mod 4 in the representation (3.4) have been described in Eqs. (3.5) and (3.6). Since in the cyclic Hilbert space of $E_{\pm1/2}\odot E_{\pm1/2}$ only these ''diagonal'' charges occur, the mixed terms can be omitted from the p-square of (3.4). The result is

$$\begin{aligned} \phi_+\otimes\mathbb{1}\cong&[u\otimes(a\odot a+b^+\odot b^+)]^{\odot2}+i[-iu^+\otimes(a^+\odot b-b\odot a^+)]^{\odot2}\\ &-[v\otimes(a\odot b^++b^+\odot a)]^{\odot2}-i[v^+\otimes(a^+\odot a^+-b\odot b)]^{\odot2}. \end{aligned} \tag{4.10}$$

We reorder this expression by separating the first p-factor (and at the same time make it more lucid by omitting the $\otimes$ and $\odot$ signs):

$$\begin{aligned} \phi_+\otimes\mathbb{1}\cong&\,a\odot[Ua(aa+b^+b^+)-Vb^+(ab^++b^+a)]+a^+\odot[-iU^+b(a^+b-ba^+)\\ &-iV^+a^+(a^+a^+-bb)]+b\odot[iU^+a^+(a^+b-ba^+)+iV^+b(a^+a^+-bb)]\\ &+b^+\odot[Ub^+(aa+b^+b^+)-Va(ab^++b^+a)], \end{aligned} \tag{4.11}$$

where $U=u\otimes u$, $V=v\otimes v$ are just another pair of orthogonal unit vectors pointing to two of the four copies of $\mathscr{H}^{\otimes2}_{(1/16)(1/16)}$ in the representation space of $E_{\pm1/2}\odot E_{\pm1/2}$.

Of course, also the embedded field $1\otimes\phi_\pm$ has an equivalent representation in terms of exchange fields of the coset and level 2 chiral observables. It does, however, not live on the same cyclic Hilbert space $(\mathscr{H}_0\otimes\mathscr{H}_0)\oplus(\mathscr{H}_{1/16}\otimes\mathscr{H}_{1/2})\oplus(\mathscr{H}_{1/2}\otimes\mathscr{H}_1)$ as $\phi_\pm\otimes1$. For example, the field

$\phi_+ \odot \phi_+$ belongs to a triplet of the diagonal $SU(2)$ and has scaling dimension 1/2. These quantum numbers are found in the sectors $\mathscr{H}_0 \otimes \mathscr{H}_1$ and $\mathscr{H}_{1/2} \otimes \mathscr{H}_0$. In fact, the joint cyclic Hilbert space contains apart from the mentioned five sectors also a second copy of $\mathscr{H}_{1/16} \otimes \mathscr{H}_{1/2}$, where the two embedded doublets have similar interpolation properties as σ and μ in (3.2). We refrain from computing these formulas.

By comparing (4.11) with (4.9), we read off the p-cubic representations of the level 2 primary exchange doublet:

$$\begin{aligned} A_+ &= Ua(aa + b^+b^+) - Vb^+(ab^+ + b^+a), \\ A_+^+ &= U^+b(a^+b - ba^+) + V^+a^+(a^+a^+ - bb), \\ B_+ &= iU^+a^+(a^+b - ba^+) + iV^+b(a^+a^+ - bb), \\ B_+^+ &= -iUb^+(aa + b^+b^+) + iVa(ab^+ + b^+a), \end{aligned} \tag{4.12}$$

while the $-$ components are determined by (4.6).

This interpretation of (4.11) means that we choose the embedding in such a way that the coset stress-energy tensor (with $c = 1/2$) acts on the Hilbert spaces belonging to the first p-factor. Then the Sugawara stress-energy tensor of the level 2 current algebra (with $c = 3/2$) acts on the triple tensor products of Hilbert spaces belonging to the remaining three p-factors in the form of three commuting $c = 1/2$ ''stress-energy tensors.'' The latter, of course, do not individually generate the conformal transformations of any level 2 fields, nor are the level 2 currents defined on the individual factor Hilbert spaces; the true stress-energy tensor with $c = 3/2$ is the sum of the three $c = 1/2$ fields.

A nontrivial cross-check of our result (4.12) is that it reproduces the four-point functions (4.8) [using Eq. (2.7)].

Finally, for later use, we display another representation which amounts to viewing the last two p-factors in (4.12) together with the multiplicity vectors as contributions to the Abelian vertex operators from which they originate [cf. Eq. (3.4)]. In this form,

$$\begin{aligned} A_+ &= a \odot E_+ P_0 - b^+ \odot E_+ P_2, \\ A_+^+ &= ib \odot E_+ P_1 + a^+ \odot E_+ P_3, \\ B_+ &= -a^+ \odot E_+ P_1 + ib \odot E_+ P_3, \\ B_+^+ &= -ib^+ \odot E_+ P_0 + ia \odot E_+ P_2, \end{aligned} \tag{4.13}$$

where $E_+ \equiv E_{+1/2}$ and P_i project onto the subspaces (3.6). We observe that each of the three sectors between which the elementary fields in these expressions interpolate arises in two orthogonal copies, distinguished by different Abelian charge projections in the second p-factor. Again, the $-$ components are determined by (4.6).

One might proceed and compute also the primary isospin 1 triplet of real fermions, e.g., from the operator product expansion of the fields (4.12). Inserting the o.p.e.'s for the elementary fields, it is evident that one will obtain the three permutations of $\psi \odot 1 \odot 1$ where ψ is the real Ising chiral fermion, along with some Klein transformations which make the three components anti-commute. Since nothing exciting happens, we do not present the explicit calculations.

C. The 1+1-dimensional Wess–Zumino–Witten (WZW) field at level 2

The $1+1$-dimensional fundamental field of the WZW model is a 2×2 matrix with $SU(2)_L \times SU(2)_R$ symmetry (by left and right matrix multiplication). It is a local field which carries the quantum numbers of the nonlocal primary exchange doublet with respect to both chiral current algebras.

At level 2, it can be represented in the form [23]

$$g_{ij}=(A_i\times A_k+A_i^+\times A_k^++B_i\times B_k+B_i^+\times B_k^+)\,\epsilon_{kj}\,, \tag{4.14}$$

where the ϵ_{kj} is necessary because the right action of $SU(2)$ is in the conjugate representation. This form parallels the representation (3.1a) of the order parameter on its own cyclic subspace [we do not consider the analogue of the disorder parameter, which would enlarge the joint cycle Hilbert space and require a second copy of $\mathscr{H}_{(1/2)(1/2)}$ as in Eqs. (3.1b)]. Due to Eq. (4.2), one has

$$g_{ij}^*=\epsilon_{ik}\epsilon_{jl}g_{kl}\,. \tag{4.15}$$

We insert the expressions (4.13) for the primary exchange fields into the definition (4.14). As before, we first consider the interpolation scheme within the joint cyclic subspace. Every component g_{ij} is a sum of contributions of the form $X_{ij}^{kl}\odot(E\times E)P_{kl}$ where $(E\times E)P_{kl}$ is the $\times$-product of the Abelian vertex operators from Eq. (4.13) on the charge sector $H_k\otimes H_l$ according to Eq. (3.7), the accompanying $\times$-product of the elementary fields from Eq. (4.13) being called X_{ij}^{kl}. As it turns out, on the cyclic Hilbert space every such operator X_{++}^{kl} (resp., X_{+-}^{kl}) can be identified with a contribution to σ (resp. μ) as in Eqs. (3.1), restricted to a Γ-charge sector H_i according to Eq. (3.6). For example, the component g_{+-} involves the combination $X_{00}=a\times a-b^+\times b^+$ which appears in the representation (3.1a) of μ on H_0. For this identification, no multiplicity vectors u, v are needed to keep the two copies H_1 and H_3 of $\mathscr{H}_{(1/16)(1/16)}$ apart, since the accompanying Abelian charge projection operators P_{kl} already have the same effect.

The result of these considerations is the following interpolation diagram in which the horizontal arrows represent the action of the component g_{+-} which involves μ and $E_{+1/2}\times E_{+1/2}$ while the diagonal arrows represent g_{++} which involves σ and $E_{+1/2}\times E_{-1/2}$ [the other two matrix elements interpolate with the reversed arrows according to Eq. (4.15)]:

$$\begin{array}{ccccccccc} 0\odot 00 & \longrightarrow & 3\odot 11 & \longrightarrow & 0\odot 22 & \longrightarrow & 3\odot 33 & \longrightarrow & 0\odot 00 \\ & \searrow\!\!\!\times & & \searrow\!\!\!\times & & \searrow\!\!\!\times & & \searrow\!\!\!\times & \\ 2\odot 02 & \longrightarrow & 1\odot 13 & \longrightarrow & 2\odot 20 & \longrightarrow & 1\odot 31 & \longrightarrow & 2\odot 02 \end{array} \tag{4.16}$$

The vertices $i\odot kl$ of this diagram are labelled by the Γ-charge sectors H_i between which σ and μ interpolate, and the Abelian charge sectors $H_k\otimes H_l$ between which the vertex operators $E_{\pm 1/2}$ interpolate, according to Eqs. (3.7).

The explicit computation of the coefficients yields

$$\begin{aligned} g_{++}&\cong\sigma\odot(E_{+1/2}\times E_{-1/2})(-P_{00}+iP_{11}-iP_{22}-P_{33}+iP_{02}-iP_{13}+P_{20}+P_{31}),\\ g_{+-}&\cong\mu\odot(E_{+1/2}\times E_{+1/2})(P_{00}+P_{11}-P_{22}+P_{33}+iP_{02}+P_{13}-iP_{20}-P_{31}). \end{aligned} \tag{4.17}$$

The charge dependences in Eqs. (4.17) can be removed to a large extent, but not entirely, by a unitary similarity transformation of the form $\mathbb{1}\odot\Sigma_{ij}\omega_{ij}P_{ij}$ with suitable complex phases ω_{ij} on the charge sectors $H_i\otimes H_j$. For later use, we display two convenient simplifications of Eqs. (4.17):

$$g_{++}\cong\sigma\odot(E_{+1/2}\times E_{-1/2}),\quad g_{+-}\cong\mu\odot(E_{+1/2}\times E_{+1/2}e^{-i\pi Q}), \tag{4.18a}$$

or

$$\begin{aligned} g_{++}&\cong\sigma\odot(E_{+1/2}e^{-i\pi Q}\times E_{-1/2}e^{-i\pi Q}),\\ g_{+-}&\cong\mu\odot(E_{+1/2}e^{-i\pi Q}\times E_{+1/2}). \end{aligned} \tag{4.18b}$$

The other matrix elements are given by Eq. (4.15), i.e., $g_{--}=g_{++}^*$ and $g_{-+}=-g_{+-}^*$.

It is amusing to observe how the Klein transformations $e^{-i\pi Q}$ in these expressions ''mimic'' the same dual commutation relations for the Abelian vertex operator fields as for the order and disorder fields [Eq. (3.3)], and hence serve to cancel the corresponding signs such that the

p-products are indeed mutually local. It is also possible to transfer the Klein transformation in Eq. (4.18) from the vertex operators to the order and disorder fields, changing their dual commutation relations into anyonic ones.

In the representation (4.18a), the trace field Tr g looks particularly simple, and with Eq. (1.7a) turns into

$$\mathrm{Tr}\, g \cong \sqrt{2}(\sigma \odot \mu \odot \mu) \cong \sqrt{2}\sigma \odot \sigma \odot \sigma. \tag{4.19}$$

The latter equivalence holds since the cosine $\mu \odot \mu$ differs from the sine $\sigma \odot \sigma$ by a $U(1)$ transformation, cf. Sec. II A.

On the line $x=0$, the fields g_{ij} can be represented as Abelian vertex operators, e.g., inserting (1.2a) into (4.18a),

$$\begin{aligned} g_{++}|_{x=0} &\cong E_\beta(P_0+P_2)+E_\gamma(P_1+P_3), \\ g_{+-}|_{x=0} &\cong E_\delta(P_0-P_2)+iE_\epsilon(P_1-P_3), \end{aligned} \tag{4.20}$$

where $\beta=\frac{1}{2}(1,1,-1)$, $\gamma=\frac{1}{2}(-1,1,-1)$, $\delta=\frac{1}{2}(-1,1,1)$, $\epsilon=\frac{1}{2}(1,1,1)$ (with $\beta+\delta=\gamma+\epsilon$), and P_i project on the subspaces of charge $(\gamma-\delta)Q=(i \bmod 4)\frac{1}{2}$, or equivalently on the subspaces of right-handed WZW isospin component $j^3_R=(i \bmod 4)\frac{1}{2}$. These formulas are another example of ''Abelianization'' of correlations along the line, just like Kadanoff's result for the Ising model. Note, however, the nontrivial angles between the charge vectors.

It is interesting to verify the unitarity of the matrix-valued field g_{ij} in these representations, and to compute the associated chiral $SU(2)$ currents. Of course, one has to consider operator quadratic expressions of the form $g_{ij}g^*_{kj}$ or $g_{ij}\partial g^*_{kj}$ which have to be renormalized. For conformal fields of chiral scaling dimensions (h,h) the proper way of doing so is a point-split regularization and renormalization by $\Delta_L^{-2h}\Delta_R^{-2h}$ in order to single out the leading contribution to the operator product expansion. This indeed reproduces unitarity (using ${}^\circ_\circ\sigma\sigma^\circ_\circ={}^\circ_\circ\mu\mu^\circ_\circ={}^\circ_\circ E_{\pm 1/2}E_{\mp 1/2}{}^\circ_\circ=\mathbb{1}$; actually one finds that rather $g/\sqrt{2}$ is unitary since we have normalized every matrix element such that its two-point function $\langle \Omega, g^*_{ij}g_{ij}\Omega\rangle$ is $(\Delta_L\Delta_R)^{3/8}$), and gives the level 2 currents $J^a_L=i\mathrm{Tr}\tau^a{}^\circ_\circ\partial_+ g g^*{}^\circ_\circ \cong J^a\times\mathbb{1}$ in the form

$$\begin{aligned} J^3 &\cong -\mathbb{1}\odot 2\partial\varphi \cong (\alpha_1+\alpha_2)\partial\varphi, \\ J^+ &\cong (E_{+1}+E_{-1})\odot E_{-1} \cong E_{\alpha_1}+E_{\alpha_2}, \end{aligned} \tag{4.21}$$

where $\alpha_1=(1,-1)$, $\alpha_2=(-1,-1)$, and similar for the right-moving currents $i\mathrm{Tr}\tau^a{}^\circ_\circ\partial_- g^* g^\circ_\circ \cong \mathbb{1}\times J^a$. In order to obtain this result, we have used the relevant leading contributions to the operator product expansions $E_{\pm 1/2}(x)E_{\pm 1/2}(y)\approx\Delta(x-y)^{-1/4}E_{\pm 1}+\cdots$ and $E_{\pm 1/2}E_{\mp 1/2}\approx\Delta^{1/4}(\mathbb{1}\pm(i/2)(x-y)\partial\varphi+\cdots)$, as well as $\sigma\sigma\approx\mu\mu\approx(\Delta_L\Delta_R)^{1/8}\mathbb{1}+\cdots$ and $\mu\sigma\approx i\sigma\mu\approx\sqrt{i}\Delta_L^{-3/8}\Delta_R^{1/8}\psi_L+\cdots$ where the left-moving real chiral Ising fermion ψ_L can in turn be represented in the form $\frac{1}{2}(E_{+1}+E_{-1})\times\mathbb{1}$, up to another Klein transformation which makes ψ_L and ψ_R anti-commute. The form (4.21) complies with Eq. (2.15) with two Frenkel–Kac representations at level 1, if one changes the basis to $\alpha_1=(\sqrt{2},0)$, $\alpha_2=(0,\sqrt{2})$.

D. The determinant field

Every p-quadratic expression in either the diagonal or the off-diagonal matrix elements g_{ij}, Eqs. (4.18), can be expressed in terms of Abelian vertex operators alone, since the doubled non-Abelian factors μ and σ can. This observation generalizes the Schroer–Truong formulas (1.7).

A particularly simple representation of the sine form (1.11) can be obtained for the p-determinant field [which is invariant under $SU(2)_L\times SU(2)_R$],

$$\mathrm{Det}_\odot g := \tfrac{1}{2}\,\epsilon_{ik}\epsilon_{jl}g_{ij}\odot g_{kl} = \tfrac{1}{2}\,(g_{11}\odot g_{22}+g_{22}\odot g_{11}-g_{12}\odot g_{21}-g_{21}\odot g_{12}). \tag{4.22}$$

In order to compute the p-determinant, we take the liberty to use the representations (4.18a) and (4.18b) for the first and second p-factor, respectively. As for the Abelian vertex operators in Eqs. (4.18), we observe that they always combine in (4.22) in the form $E_{\pm 1/2}\odot E_{\mp 1/2}\cong E_{\pm\alpha}$ where $\alpha=1/\sqrt{2}$. Consequently, all charges in the cyclic Hilbert space are integer multiples of α, and the charge operators involved may be replaced by $e^{i\pi Q}\odot\mathbb{1}\cong e^{i\pi\alpha Q}$ and $\mathbb{1}\odot e^{i\pi Q}\cong e^{-i\pi\alpha Q}=e^{i\pi\alpha Q}\cdot C$ where $C=e^{2\pi i\alpha Q}=C^{-1}$ is the Casimir operator introduced previously in Sec. II C.

The crucial point now is that an enhanced symmetry emerges on the vertex operators of charge $\pm\alpha$ which was not present on the vertex operators of charge $\pm 1/2$ in the representation (4.13), i.e., before taking the p-square [and which spoils the original (squared) level 2 affine chiral $SU(2)$ symmetries generated by (4.21)]. Namely, all terms contributing to (4.22) involve $E_{\pm\alpha}e^{i\pi\alpha Q}\times E_{\pm\alpha}e^{i\pi\alpha Q}$, i.e., the components of the pseudo-real $SU(2)$ level 1 doublet $\phi_\pm$ as in Eq. (2.13), possibly along with some Casimir operators. Explicitly, we have

$$\begin{aligned}\mathrm{Det}_\odot g &\cong \tfrac{1}{2}\left[\sigma_D\odot(\phi_+C\times\phi_-C+\phi_-\times\phi_+)-i\mu_D\odot(\phi_+C\times\phi_+C+\phi_-\times\phi_-)\right]\\ &\cong\frac{1}{2\sqrt{2}i}\left[(E_{-1/2}\times E_{+1/2})\odot(\phi_++\phi_-)\times(\phi_++\phi_-)\right.\\ &\quad\left.+(E_{+1/2}\times E_{-1/2})\odot(\phi_+-\phi_-)\times(\phi_+-\phi_-)\right].\end{aligned}\tag{4.23}$$

For the second line, we observed that on the cyclic Hilbert space $C\otimes C$ equals $\mathbb{1}$, and we evaluated σ_D and μ_D according to Eqs. (1.7).

The new level 1 chiral $SU(2)$ symmetries permit us to rotate the components $(\phi_+\pm\phi_-)$ into $\sqrt{2}\phi_\mp$ in the left chiral factor and into $\pm\sqrt{2}\phi_\pm$ in the right chiral factor. This unitary similarity transformation turns the determinant field into

$$\mathrm{Det}_\odot g\cong\frac{1}{\sqrt{2}i}\left[(E_{-1/2}\times E_{+1/2})\odot(\phi_-\times\phi_+)-(E_{+1/2}\times E_{-1/2})\odot(\phi_+\times\phi_-)\right].\tag{4.24}$$

After this transformation, the cyclic Hilbert space is seen to carry opposite charges in the two chiral factors. We may therefore insert $\mathbb{1}=\mathbb{1}\odot(e^{i\pi\alpha Q}\otimes e^{i\pi\alpha Q})$ and re-substitute the vertex operators $E_{\pm\alpha}$ for the doublet $\phi_\pm$. This gives finally the sine formula, to be compared with Eq. (1.7a),

$$\mathrm{Det}_\odot g\cong\frac{1}{\sqrt{2}i}\,(E_{-\beta}\times E_\beta-E_\beta\times E_{-\beta}),\tag{4.25}$$

with $\beta=\frac{1}{2}(1,\sqrt{2})$ or, after an $O(2)$ rotation in charge space, $\beta=\frac{1}{2}\sqrt{3}$. The nontriviality of this formula can be esteemed upon the intricate exercise to test it only at the four-point level.

V. CONCLUSIONS

We have established a sort of ''Quantum Field Lego'' with chiral non-Abelian exchange fields which can be recombined in a large variety of ways to produce several local and nonlocal, chiral and $1+1$-dimensional Wightman fields belonging to different models. It originates in some well-known remarkable properties of the (critical) Ising model,[4,2,3] and entails similar but new results for a WZW (=generalized Thirring) model, and gives rise to relations between various models.

A prominent feature of the basic decomposition (3.4) is that it produces non-Abelian exchange fields, satisfying commutation relations according to a non-Abelian representation of the braid group and possessing a non-Abelian fusion structure, by factorization of Abelian vertex operators with anyonic commutation relations and additive charge structure.

We have seen along the way (Secs. II C and IV C) how the generators of the chiral symmetry can be explicitly computed in our operator representations and are found, of course, in complete agreement with well-known results.

Throughout the paper, we have mentioned only en passant the issue of braid group commutation relations. Let us just add that p-products go along with the tensor products of representa-

tions of the braid group; notably the basic result singles out an Abelian subrepresentation within the square of the Hecke type representations associated with the elementary fields.

Let us consider the question whether one may expect parallel findings in other minimal or coset models. Since Abelian vertex operators are the primary fields for stress-energy tensors with integer central charge c, one may expect representations of Abelian vertex operators as p-products of non-Abelian ones whenever a stress-energy tensor with integer central charge occurs in the product of the chiral observables. In the case of the doubled Ising model, the latter is just the sum of the two stress-energy tensors with $c=\frac{1}{2}$, and in the case of the determinant field, it is the coset stress-energy tensor with $c=1$ for the coupling of two level 2 current algebras into the level 4 current algebra. Admitting sufficiently high p-powers, any model with rational central charge presumably provides another set of non-Abelian roots of Abelian vertex operators. On the other hand, the ''Kadanoff mechanism'' of quadratic Abelianization upon restriction to a time-like axis, which in the present model we consider to be the reason for the possibility of several independent recombinations of the same elementary fields, can only work when the two opposite chiral central charges add up to an integer. This property of course distinguishes the Ising model with $c=\frac{1}{2}$ and the $SU(2)$ model at level 2 with $c=\frac{3}{2}$ studied in this paper from other minimal or WZW models.

An immediate further application, not treated in this paper, is the construction of the primary field ϕ_{21} with dimension $\frac{7}{16}$ of the minimal model with $c=\frac{7}{10}$ which arises in the coset construction as the singlet contribution to the p-product of the $SU(2)$ primary doublets at level 1 [Eq. (2.13)] and at level 2 [Eq. (4.12)], commuting with the level 3 currents.

Although the basic result involves chiral and nonlocal fields, several of our applications pertain to ordinary local Wightman fields in $1+1$-dimensional Minkowski space-time.

We want to draw the attention to a point of general importance which is illustrated by Eq. (4.25). By definition, the determinant field is p-quadratic in the four mutually local fields g_{ij}. On the other hand, the right hand side of Eq. (4.25), by elementary identities using the representation (1.7), equals one half of

$$\sigma_D\odot\sigma_D\odot\sigma_D-\sigma_D\odot\mu_D\odot\mu_D-\mu_D\odot\sigma_D\odot\mu_D-\mu_D\odot\mu_D\odot\sigma_D\,. \tag{5.1}$$

[This formula would not be guessed from Eqs. (4.18) according to which g_{ij} are not p-cubic in σ and μ.] The determinant field is thus at the same time also p-cubic in local fields. Yet, a common sixth-order factorization into mutually local fields cannot be perceived (at least not in terms of our elementary fields). This indicates, with due precaution, that something like a ''prime decomposition'' with respect to the p-product does not exist in general within the class of local Wightman fields. In this context we notice that two-dimensional local and scale-invariant fields such as $E_\alpha\times E_{-\alpha}$ are infinitely factorizable with respect to the p-product. This is certainly impossible in higher dimensions since the scaling dimension is bounded from below by $(d-2)/2$. For the analogous question with respect to the s-product, cf. Ref. 24.

We consider our work as an exercise with noncanonical fields from which several lessons on the general theory of Wightman fields can be drawn. We want to refute the possible impression that the construction of p-products is just an artificial trick to construct new Wightman fields without a proper physical meaning from old ones. On the contrary, all the product fields in this article were previously known and considered in their own model context. The new aspect is the passage between different models enabled by p-products.

ACKNOWLEDGMENTS

The second author (K.-H.R.) thanks B. Schroer for many helpful and interesting comments during the final preparation of this article.

[1] A. S. Wightman, ''Introduction to some aspects of the relativistic dynamics of quantized fields,'' in Cargèse Lectures in Theoretical Physics, edited by M. Levy (Gordon and Breach, New York, 1967), pp. 171–291.
[2] E. H. Lieb, T. D. Schultz, and C. D. Mattis, ''Two soluble models of an antiferromagnetic chain,'' Ann. Phys. (N.Y.) **16**, 407–466 (1961); B. M. McCoy, ''Spin correlation functions of the X-Y model,'' Phys. Rev. **173**, 531–541 (1968).
[3] B. Schroer and T. T. Truong, ''The order/disorder quantum field operators associated with the two-dimensional Ising model in the continuum limit,'' Nucl. Phys. B **144**, 80–122 (1978).
[4] L. P. Kadanoff, ''Correlations along a line in the two-dimensional Ising model,'' Phys. Rev. **188**, 859–863 (1969); H.

Ceva and L. P. Kadanoff, ''Determination of an operator algebra for the two-dimensional Ising model,'' Phys. Rev. B **3**, 3918–3939 (1971).
[5] H.-J. Borchers, ''Field operators as C^∞ functions in spacelike directions,'' Nuovo Cimento **33**, 1600–1613 (1964).
[6] K.-H. Rehren and B. Schroer, ''Exchange algebra on the light-cone and order/disorder $2n$-point functions in the Ising field theory,'' Phys. Lett. B **198**, 84–88 (1987).
[7] A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, ''Infinite dimensional symmetries in two-dimensional quantum field theory,'' Nucl. Phys. B **241**, 333–380 (1984).
[8] P. Goddard, A. Kent, and D. Olive, ''Unitary representations of the Virasoro and super-Virasoro algebras,'' Commun. Math. Phys. **103**, 105–119 (1986).
[9] L. V. Belvedere, K. D. Rothe, B. Schroer, and J. A. Swieca, ''Generalized two-dimensional abelian gauge theories and confinement,'' Nucl. Phys. B **153**, 112–140 (1979).
[10] V. G. Knizhnik and A. B. Zamolodchikov, ''Current algebra and Wess-Zumino model in two dimensions,'' Nucl. Phys. B **247**, 83–103 (1984).
[11] E. Witten, ''Non-abelian bosonization in two dimensions,'' Commun. Math. Phys. **92**, 455–472 (1984).
[12] K. Frieler, ''Nicht-abelsche Faktorisierungen abelscher Vertex-Operatoren,'' Diploma thesis, Hamburg University, 1997 (in German).
[13] H. Lehmann and J. Stehr, ''The Bose field structure associated with a free massive Dirac field in one space dimension,'' DESY preprint 76-29, 1976.
[14] B. Schroer and J. A. Swieca, ''Conformal transformations for quantized fields,'' Phys. Rev. D **10**, 480–485 (1974).
[15] A. L. Carey, S. N. M. Ruijsenaars, and J. D. Wright, ''The massless Thirring model: positivity of Klaiber's n-point functions,'' Commun. Math. Phys. **99**, 347–364 (1985).
[16] S. Mandelstam, ''Soliton operators for the quantized sine-Gordon equation,'' Phys. Rev. D **11**, 3026–3030 (1975).
[17] K.-H. Rehren, ''Bounded Bose fields,'' Lett. Math. Phys. **40**, 299–306 (1997).
[18] B. Schroer, J. A. Swieca, and A. H. Völkel, ''Global operator expansions in conformally invariant relativistic quantum field theory,'' Phys. Rev. D **11**, 1509–1520 (1975).
[19] K. Fredenhagen, K.-H. Rehren, and B. Schroer, ''Superselection sectors with braid group statistics and exchange algebras, II,'' Rev. Math. Phys. (Special Issue), **113–157** (1992).
[20] M. Jörß, ''The construction of pointlike localized charged fields from conformal Haag-Kastler nets,'' Lett. Math. Phys. **38**, 257–274 (1996).
[21] I. B. Frenkel, J. Lepowsky, and A. Meurman, ''Vertex operator algebras and the monster,'' *Pure and Applied Mathematics* (Boston, Academic, New York, 1988); Vol. 8, especially Chap. 8.
[22] I. B. Frenkel and V. G. Kac, ''Basic representations of affine Lie algebras and dual resonance models,'' Invent. Math. **62**, 23–66 (1980).
[23] K.-H. Rehren, ''Space-time fields and exchange fields,'' Commun. Math. Phys. **132**, 461–483 (1990).
[24] G. C. Hegerfeldt, ''Prime field decomposition and infinitely divisible states on Borchers' tensor algebra,'' Commun. Math. Phys. **45**, 137–151 (1975).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 6 JUNE 1998

Constrained dynamics for quantum mechanics. I. Restricting a particle to a surface

Hendrik Grundling[a)]
Department of Pure Mathematics, University of New South Wales, Sydney, New South Wales 2052, Australia

C. A. Hurst[b)]
Department of Physics and Mathematical Physics, University of Adelaide, GPO Box 498, Adelaide, South Australia 5001, Australia

(Received 15 July 1997; accepted for publication 17 February 1998)

We analyze constrained quantum systems where the dynamics do not preserve the constraints. This is done, in particular, for the restriction of a quantum particle in $\mathbb{R}^n$ to a curved submanifold, and we propose a method of constraining and dynamics adjustment that produces the right Hamiltonian on the submanifold when tested on known examples. This method will be the germ of a ''Dirac algorithm for quantum constraints.'' We generalize it to the situation where the constraint is a general self-adjoint operator with some additional structures. © *1998 American Institute of Physics.* [S0022-2488(98)00906-2]

I. INTRODUCTION

The problem of enforcing constraints in classical mechanics has had satisfactory solutions for some time now.[1–4] The same is not true in quantum mechanics, contrary to claims of ''quantizations'' of classical constrained systems after constraining. Quantization maps are rarely unique nor well defined.[5–8] Moreover, using noncommutativity, some constraints can be defined in quantum mechanics with zero classical limit. This means that the general problem of quantum constraints needs to be solved in the quantum arena without appealing to classical methods.

There is presently a wide variety of methods for imposing quantum constraints,[1,9–13] but these methods (except Ref. 13) deal with the kinematics only at the quantum level, and their interrelations are unclear. (The method of Ref. 13 does consider the dynamics, producing for the constrained Hamiltonian the projection of the unconstrained one onto the physical subspace. This will not work for the system considered in this paper.) When the given quantum dynamics preserves the constrained subsystem that is fine, one only needs to restrict it to the constrained subsystem to obtain the constrained dynamics. However, when this is not the case, the problem arises of how to appropriately adjust the constraint set and the dynamics into a stable system. (New constraints produced by such a method are generally called secondary constraints). The obvious strategy of fixing the dynamics and extending the constraint set to its orbit under the dynamics, produces the wrong physics in examples. It is our opinion that the question of secondary quantum constraints and dynamics adjustment is a physical problem, and cannot be decided by mathematical arguments alone. In this paper we set out to solve this problem of secondary quantum constraints in the limited class of systems consisting of the restriction of a quantum particle on $\mathbb{R}^n$ to a smooth (possibly curved) submanifold, i.e., the case of holonomic constraints (cf. Ref. 14, p. 75). The motivation for this choice is as follows.

(1) There are several examples of quantum mechanics on surfaces available (e.g., $\mathbf{S}^2$) against which we can test the results of our analysis.

(2) In $\mathbb{R}^n$ there is ample geometry available to guide the intuition, e.g., a metric, hence normal vectors and Lebesgue measure.

(3) Generalizations are easy from this class, in fact, in Sec. VI we propose a generalization of it to the case of a general self-adjoint operator on a Hilbert space with additional structures.

a)Electronic mail: hendrik@maths.unsw.edu.au
b)Electronic mail: ahurst@physics.adelaide.edu.au

 © 1998 American Institute of Physics

(4) A constraint in this situation is of the form $\varphi(q)=0$, which has an immediate obvious quantization in the Schrodinger representation, whereas a constraint of the form $\varphi(q,p)=0$ need not have a unique quantization.

(5) Physically, these systems in $\mathbb{R}^3$ can be approximated by very thin films or wires and so are close to experimental verification.

(6) The exposition of constraint methods for these systems is more transparent than for more general ones.

(7) Holonomic quantum constraints have been considered in the physics literature.[15,16]

The kinematics of quantum systems of this type has already been done concretely by Landsman[11] and abstractly at the C^* level by Grundling and Hurst,[9] so any new proposed method should be compatible with these. We will not consider BRST methods,[10] as we believe these not to be equivalent at the quantum level to the other methods.[17–19] So, to summarize, we aim to solve for this class of systems the problem of secondary quantum constraints, i.e., given a constrained system where the dynamics do not preserve the constraints, to find a method of adjustment that agrees on known constrained systems with the right physics.

The architecture of the paper is as follows; in Sec. II we summarize the Dirac procedure for constraining a classical particle to a submanifold, both local and global, and we also do two examples. In Sec. III we consider the question of quantum constraints, we first summarize the usual Dirac procedure, then start to analyze the situation of restricting a quantum particle in $\mathbb{R}^n$ to a lower-dimensional submanifold Γ. We obtain the constraining map κ from the Hilbert space $L^2(\mathbb{R}^n)$ of the unconstrained particle to that of the constrained particle $L^2(\Gamma)$, and we discuss generalizations to other types of constrained systems. In Sec. IV we solve the problem of how to obtain from the Hamiltonian of the original particle a constrained Hamiltonian on $L^2(\Gamma)$, and we apply it to two examples; restricting a free particle in $\mathbb{R}^3$ to a sphere and a cylinder, and we obtain in each case the correct Hamiltonian. In Sec. V, we consider how to obtain observables on $L^2(\Gamma)$ from the original observables on $L^2(\mathbb{R}^n)$, i.e., we analyze when we obtain sensible liftings of operators through the map κ, and how to obtain a suitable field algebra for the constrained particle. In Sec. VI we suggest a generalization of the method of the preceding sections to the constraining situation where the constraint is a self-adjoint operator on an abstract Hilbert space, with zero in the continuous part of the spectrum. We find that we need to assume some additional structure to do the work that the geometry of $\mathbb{R}^n$ does in the method of the preceding sections. The reader in a hurry can start with Sec. III.

II. CLASSICAL SECONDARY CONSTRAINTS

We start by recalling the basic Dirac–Bergman method,[1,20,21] which is, of course, a local procedure. However, to keep this brief, we only present it in the context of a particle in $\mathbb{R}^n$ constrained to an $(n-1)$-dimensional submanifold Γ, which is the system at the focus of this paper.

The full phase space is $\mathbb{R}^{2n}$ with generic point $(\mathbf{q},\mathbf{p})$ and the usual Poisson algebra $\mathscr{P}=(C^\infty(\mathbb{R}^{2n}),\{\cdot,\cdot\})$ is the setting for the problem. We specify a smooth $(n-1)$-dimensional submanifold Γ in $\mathbb{R}^n$ as the zero set of a smooth function $\varphi\in C^\infty(\mathbb{R}^n)$, i.e., $\Gamma=\varphi^{-1}(0)$. We also assume $\nabla\varphi\neq 0$ on Γ so we can use the gradient to define a normal on Γ. The Hamiltonian for a particle in a potential V is

$$H_c=\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}+V(\mathbf{q}). \tag{2.1}$$

So we have a constrained system with a single primary constraint $\varphi(\mathbf{q},\mathbf{p}):=\varphi(\mathbf{q})$ and primary constraint manifold in phase space $\varphi^{-1}(0)=\Gamma\times\mathbb{R}^n=:\Gamma_p$. The Dirac procedure starts with the new Hamiltonian:

$$H_p:=\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}+V(\mathbf{q})+\mu\varphi,$$

where $\mu\in\mathscr{P}$ is a free multiplier, to be determined below. So the new time evolution for $A\in\mathscr{P}$ on Γ_p is

$$\dot{A}\restriction\Gamma_p=\{A,H_p\}\restriction\Gamma_p=(\{A,H_c\}+\mu\cdot\{A,\varphi\})\restriction\Gamma_p.$$

Since time evolution must preserve Γ_p, we require that

$$0=\dot{\varphi}\restriction\Gamma_p=\{\varphi,H_p\}\restriction\Gamma_p=\frac{1}{m}\,\mathbf{p}\cdot(\boldsymbol{\nabla}\varphi)\restriction\Gamma_p.$$

Thus $\chi:=(1/m)\mathbf{p}\cdot(\boldsymbol{\nabla}\varphi)$ is a secondary constraint determining a smaller submanifold $\Gamma_f:=\varphi^{-1}(0)\cap\chi^{-1}(0)$, which again must be preserved by time evolution. In the examples in the rest of this paper we will constrain free motion to a surface, so for this case ($V=0$),

$$\dot{\xi}\restriction\Gamma_f=0=(\{\xi,H_c\}+\mu\cdot\{\xi,\varphi\})\restriction\Gamma_f,\quad\text{where}\quad\xi\ \text{denotes}\ \chi,\varphi,$$

so

$$\begin{aligned}0&=\left(\left\{\frac{1}{m}\,\mathbf{p}\cdot(\boldsymbol{\nabla}\varphi),\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}\right\}+\mu\left\{\frac{1}{m}\,\mathbf{p}\cdot(\boldsymbol{\nabla}\varphi),\varphi\right\}\right)\restriction\Gamma_f\\&=\left(\frac{1}{2m^2}\,\mathbf{p}\cdot\{\boldsymbol{\nabla}\varphi,\mathbf{p}\cdot\mathbf{p}\}+\frac{1}{m}\,\mu(\boldsymbol{\nabla}\varphi)\cdot\{\mathbf{p},\varphi\}\right)\restriction\Gamma_f\\&=\left(\frac{1}{m^2}\,\mathbf{p}\cdot(\boldsymbol{\nabla}(\mathbf{p}\cdot\boldsymbol{\nabla}\varphi))-\frac{1}{m}\,\mu(\boldsymbol{\nabla}\varphi)\cdot(\boldsymbol{\nabla}\varphi)\right)\restriction\Gamma_f.\end{aligned}$$

Since $\boldsymbol{\nabla}\varphi\neq 0$ on Γ, we solve on Γ_f,

$$\mu=\frac{\mathbf{p}\cdot(\boldsymbol{\nabla}(\mathbf{p}\cdot\boldsymbol{\nabla}\varphi))}{m\|\boldsymbol{\nabla}\varphi\|^2}=\frac{(\mathbf{p}\cdot\boldsymbol{\nabla})^2\varphi}{m\|\boldsymbol{\nabla}\varphi\|^2},$$

thus

$$H_p=\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}+\varphi\cdot\frac{(\mathbf{p}\cdot\boldsymbol{\nabla})^2\varphi}{m\|\nabla\varphi\|^2}.$$

The secondary constraint, looked at on configuration space, restricts momenta by

$$\chi=\left(\frac{1}{m}\,\mathbf{p}\cdot\boldsymbol{\nabla}\varphi\right)\restriction\Gamma=0,\tag{2.2}$$

to guarantee that the motion is tangent to the surface. There are no more constraints. Moreover, since

$$\{\varphi,\chi\}\restriction\Gamma_f=\frac{1}{m}\,\|\boldsymbol{\nabla}\varphi\|^2\restriction\Gamma_f\neq 0,$$

the constraints are second class. The Dirac algorithm continues to construct a Dirac bracket, but we will not need this. For later use we specialize the above to a sphere and a cylinder.

Let Γ be a sphere of radius a in $\mathbb{R}^3$. The phase space is $\mathbb{R}^6$, $\Gamma_p=\mathbf{S}^2\times\mathbb{R}^3$, and we have

$$\varphi(\mathbf{q})=\mathbf{q}\cdot\mathbf{q}-a^2,\quad\text{so}\ \ \chi=\frac{1}{m}\,\mathbf{p}\cdot\boldsymbol{\nabla}\varphi=\frac{2}{m}\,\mathbf{p}\cdot\mathbf{q},$$

and since $\boldsymbol{\nabla}(\mathbf{p}\cdot\boldsymbol{\nabla}\varphi)=2\mathbf{p}$, here we get

$$H_p=\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}\left(1+\frac{\mathbf{q}\cdot\mathbf{q}-a^2}{\|\mathbf{q}\|^2}\right).\tag{2.3}$$

An explicit calculation produces

$$H_p \restriction \Gamma_f = \frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}\Big|_{\substack{|\mathbf{q}|=\alpha \\ \mathbf{q}\cdot\mathbf{p}=0}} = \frac{|\mathbf{L}|^2}{2ma^2}, \tag{2.4}$$

where $\mathbf{L}=\mathbf{q}\times\mathbf{p}|_{r=a}$ is angular momentum on the sphere.

If we take Γ to be a cylinder of radius a around the z axis in $\mathbb{R}^3$, we have

$$\varphi(\mathbf{q})=q_1^2+q_2^2-a^2, \quad \text{so} \quad \chi=\frac{2}{m}\,(p_1q_1+p_2q_2),$$

$$H_p=\frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}+(q_1^2+q_2^2-a^2)\,\frac{p_1^2+p_2^2}{2m(q_1^2+q^2)}. \tag{2.5}$$

And now,

$$H_p \restriction \Gamma_f = \frac{1}{2m}\,\mathbf{p}\cdot\mathbf{p}\Big|_{\substack{q_1^2+q_2^2=a^2 \\ p_1q_1=-p_2q_2}} = \frac{1}{2m}\left(\frac{L_3^2}{a^2}+p_3^2\right)\restriction \Gamma_f, \tag{2.6}$$

where $L_3 := q_1p_2-q_2p_1$ is the usual angular momentum around the q_3 axis.

Next we recall the global version of the Dirac–Bergman method of constraints as worked out by Gotay, Nester, and Hinds,[2] which is a considerable simplification. Start with a symplectic manifold (M,ω) (finite dimensional here) and a Hamiltonian $H\in C^\infty(M)$. Let $M_1\subset M$ be a given primary constraint submanifold. Now the evolution equation $\iota_X\omega=dH$ may not have solutions for the vector field X tangential to M_1, so let

$$M_2 := \{m\in M_1 \,|\, dH(m)=\iota_X\omega(m), \quad \text{for some} \quad X\in T_mM_1\}.$$

However, the time evolution must now preserve M_2, so iterate

$$M_{k+1} := \{m\in M_k \,|\, dH(m)=\iota_X\omega(m), \quad \text{for some} \quad X\in T_mM_k\}.$$

When this iteration converges sensibly, we obtain a final constraint manifold M_c on which $dH=\iota_X\omega$ has solutions $X\in\Gamma^\infty(M_c)\equiv$smooth vector fields on M_c, which provide the desired time evolutions. Geometrically we can think of M_c as the largest submanifold, which has a vector field X that is tangential at each $m\in M_c$ to a trajectory of the time evolution of the original unconstrained manifold M. The completeness of X is still an open question for the general analysis. (Below we will refer to the above as the GNH algorithm).

We will take this geometric picture as the guiding principle for constraining the dynamics of a quantum system below in Sec. IV.

III. CONSTRAINING THE REPRESENTATION SPACE

For reference, we recall the usual Dirac prescription for first class quantum constraints. Given a field algebra $\mathscr{A}$ of operators preserving a dense domain Δ in a Hilbert space $\mathscr{H}$, as well as a Hamiltonian $H\in\mathscr{A}_{sa}$ and a set of constraints $\{\hat{\varphi}_i\in\mathscr{A}_{sa}\,|\,i\in I\}$ (where I is an abstract index set), one selects the physical subspace by

$$\mathscr{H}_{\text{phys}} := \overline{\{\psi\in\Delta\,|\,\hat{\varphi}_i\psi=0, \quad \forall i\}},$$

and this is assumed to be nonzero (hence zero must be in the discrete spectrum of each $\hat{\varphi}_i$). Decompose $\mathscr{H}=\mathscr{H}_{\text{phys}}\oplus\mathscr{H}_{\text{phys}}^{\perp}$, so for each $A\in\mathscr{A}$ we have a decomposition

$$A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Clearly, the only $A\in\mathscr{A}$, which, together with A^* can be restricted to $\mathscr{H}_{\text{phys}}$, are those of the form

$$A=\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}.$$

Denote the * algebra of these by $\mathscr{O}$. Those that restrict trivially are of the form

$$A=\begin{pmatrix} 0 & 0 \\ 0 & d \end{pmatrix}$$

and comprise an ideal $\mathscr{D}$ of $\mathscr{O}$. Then the * algebra of physical observables are

$$\mathscr{O}\restriction\mathscr{H}_{\text{phys}}\cong\mathscr{O}/\mathscr{D}=:\mathscr{R}.$$

If $\exp itH$ preserves $\mathscr{H}_{\text{phys}}$, then Ad $\exp(itH)$ lifts to automorphisms on $\mathscr{R}$. Otherwise, we need to alter it (cf. Sec. IV). Let P_{phys} be the projection on $\mathscr{H}_{\text{phys}}$ and define the sesquilinear form

$$(\psi_1,\psi_2)_D:=(P_{\text{phys}}\psi_1,P_{\text{phys}}\psi_2)=(\psi_1,P_{\text{phys}}\psi_2); \tag{3.1}$$

then $\text{Ker}(\cdot,\cdot)_D=\mathscr{H}_{\text{phys}}^{\perp}$, and we write P_{phys} as a factorization $P_{\text{phys}}:\mathscr{H}\to\mathscr{H}/\text{Ker}(\cdot,\cdot)_D\cong\mathscr{H}_{\text{phys}}$ (an obvious identification), and we can think of $\mathscr{O}$ as all $A\in\mathscr{A}$, which together with A^* can lift through the factorization.

Next, we introduce our main object of study. Consider a quantum particle in $\mathbb{R}^n$ in the Schrödinger representation. Its basic data consists of the Hilbert space $\mathscr{H}=L^2(\mathbb{R}^n,\mu)$ (with μ the Lebesgue measure), and on the dense subspace $C_c^\infty(\mathbb{R}^n)$ of smooth functions of compact support, we have the position operator $(\hat{\mathbf{q}}\psi)(\mathbf{x}):=\mathbf{x}\psi(\mathbf{x})$, the momentum operator $(\hat{\mathbf{p}})\psi(\mathbf{x}):=(i\nabla\psi)(\mathbf{x})$, and a Hamiltonian H [which is $(1/2m)\hat{\mathbf{p}}\cdot\hat{\mathbf{p}}$ when the particle is free]. These operators are all essentially self-adjoint and preserve $C_c^\infty(\mathbb{R}^n)$. We wish to constrain this particle to a given smooth $(n-1)$-dimensional submanifold $\Gamma\subset\mathbb{R}^n$. Assume we have a bounded real-valued constraint function $\varphi\in C^\infty(\mathbb{R}^n)$ with $\nabla\varphi$ nonzero and bounded on a neighborhood of Γ, and

$$\Gamma=\{x\in\mathbb{R}^n\,|\,\varphi(\mathbf{x})=0\}=\varphi^{-1}(0).$$

(Notice that the level hypersurfaces of φ form a foliation of this neighborhood.[22]) Quantize φ by the multiplication operator $\varphi(\hat{\mathbf{q}})$, i.e.,

$$(\hat{\varphi}\psi)(\mathbf{x})=\varphi(\mathbf{x})\cdot\psi(\mathbf{x}),\quad\forall\psi\in\mathscr{H},$$

which is self-adjoint since φ is real valued, and bounded because φ is.

Assume we are given a concrete field algebra $\mathscr{F}\subset\mathscr{B}(\mathscr{H})$, which should be a C^* algebra containing all relevant operators in bounded form, e.g., $\hat{\varphi}$, $\exp i\mathbf{a}\cdot\hat{\mathbf{q}}$, $\exp i\mathbf{a}\cdot\hat{\mathbf{p}}$ ($\mathbf{a}\in\mathbb{R}^n$). The choice of field algebra will turn out to be important for producing a meaningful algebra of observables in the constrained system, but we will return to this matter in Sec. V. In summary, we are assuming the following data: (a) the operators $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, and H on $C_c^\infty(\mathbb{R}^n)\subset L^2(\mathbb{R}^n)$; (b) a bounded $\varphi\in C^\infty(\mathbb{R}^n)$ with $\Gamma=\varphi^{-1}(0)$ a smooth $(n-1)$-dimensional submanifold, and with $\nabla\varphi$ nonzero and bounded in a neighborhood of Γ; and (c) a unital field algebra $\mathscr{F}$ on $L^2(\mathbb{R}^n)$.

We would like to find a way of imposing the constraint ''$\hat{\varphi}=0$'' on this system. However, since Γ is a null set of μ, we have $\text{Ker}\,\hat{\varphi}=\{0\}$, and so Dirac's method of restriction to the kernel of the constraint fails. The fact that zero is in the continuous spectrum of $\hat{\varphi}$ is physically due to the uncertainty principle; a particle confined to Γ will violate Heisenberg's uncertainty principle in a direction normal to Γ.

At present, there are two methods that can handle the above situation: the T procedure of Ref. 9, as well as Landsman's application of Rieffel induction.[11] The T procedure ''ignores'' the original representation of $\mathscr{F}$, considers $\mathscr{F}$ as an abstract C^* algebra, for which it then finds those representations π for which zero is in the discrete part of the spectrum of $\pi(\hat{\varphi})$, and works out the algebraic structures associated with factoring $\text{Ker}\,\pi(\hat{\varphi})$ out of each of those representations. On the other hand, Landsman's method assumes a locally compact group H with a continuous proper action on $\mathbb{R}^n$ such that Γ is precisely the set of points left invariant. Assuming some initial representation π of H on $L^2(\mathbb{R}^n)$, he makes $C_c(\mathbb{R}^n)$ into an $(\mathscr{A}$–$\mathscr{B})$ bimodule, where $\mathscr{B}$ is the

algebra $C_c(H)$ (with convolution for multiplication) and $\mathscr{A}$ is the part of the commutant of $\pi(H)$ preserving $C_c(\mathbb{R}^n)$. There is then a natural rigging map, and this allows a Rieffel induction from the trivial representation of $\mathscr{B}$ to $\mathscr{A}$.

The T procedure selects the set of all representations in which the Dirac algorithm makes sense, which necessarily excludes the original representation of the current system. This means that one loses the original physical interpretation of the state vectors, and one has the problem of which representation to choose for the constrained system. From the point of view of the physics, one may sometimes be more interested in constructing the constrained operators directly out of the original ones. Landsman's method does this, but requires some additional group structure, and quantizes a fairly small algebra. Both methods do only the kinematics, and if the dynamics do not preserve Γ it is hard to see from these structures what should be done.

We will propose a method for constraining the kinematics of a quantum particle to Γ, which generalizes the usual Dirac prescription and in which the constrained operators are explicitly constructed from the original ones. It will then be fairly easy to see what to do with the dynamics.

In a sense the solution of the kinematics problem is already known. One takes the subspace $C_c(\mathbb{R}^n)\subset L^2(\mathbb{R}^n)$ and restricts it to Γ, thus obtaining a dense subset of the physical Hilbert space $L^2(\Gamma,\gamma)$ (γ denotes the measure induced on Γ by μ). The restriction is the same as factoring out the subspace,

$$N:=\{f\in C_c(\mathbb{R}^n)\,|\, f\restriction\Gamma=0\},$$

from $C_c(\mathbb{R}^n)$. The observables are those operators that ''restrict to Γ,'' i.e., operators A that preserve $C_c(\mathbb{R}^n)$ and N, in which case A lifts through the restriction to define an operator on $C_c(\Gamma)\subset L^2(\Gamma)$. This is the procedure that we intend to refine. Note, however, that the restriction map $\rho: C_c(\mathbb{R}^n)\to C_c(\Gamma)$ is unbounded with respect to the L^2 norms and it is not closable as an operator [let $\psi_k\in L^2(\mathbb{R}^n)$ be a sequence of continuous functions with compact support, all of which have the same restriction to a nonzero $\theta\in L^2(\Gamma)$, and which converge to zero w.r.t. the L^2 norm. Then $\psi_k\to 0$, $\rho(\psi_k)\to\theta\neq 0$, so $\lim \rho(\psi_k)\neq\rho(\lim \psi_k)$]. So we expect some pathologies to arise in operator questions. We would also like to build the inner product of $L^2(\Gamma)$ explicitly out of the inner product of $L^2(\mathbb{R}^n)$. To do this, we adapt a well-known argument from statistical mechanics,[23] and p. 80 of Ref. 24.

Since $\nabla\varphi(\mathbf{x})\neq 0$ on a neighborhood of Γ, locally (i.e., in open bounded sets) the critical points of φ are a finite distance away from Γ. So we have the geometrically obvious lemma:

Lemma 3.2: There is a $t>0$ such that in $\varphi^{-1}[-t,t]$ we have locally a smooth curvilinear orthogonal coordinate system $(\mathbf{y},\varphi)$ where $\mathbf{y}=$ const are curves normal to the level hypersurfaces of φ.

Proof: Consider the vector field $X(\mathbf{x})=\mathbf{n}(\mathbf{x})/|\nabla\varphi(\mathbf{x})|$ on $\varphi^{-1}(-t,t)$, where $\mathbf{n}=\nabla\varphi/|\nabla\varphi|$. Then its integral curves are normal to all level surfaces of φ in $\varphi^{-1}(-t,t)$, and, moreover, for each $s\in(-t,t)$ we have a diffeomorphism $\alpha_s:\varphi^{-1}(0)\to\varphi^{-1}(-t,t)$, such that $\alpha_0=t$ and $(d/ds)\alpha_s(\mathbf{x})=X(\alpha_s(\mathbf{x}))$.[25] Now

$$\frac{d}{ds}\,\varphi(\alpha_s(\mathbf{x}))=X(\varphi)(\alpha_s(\mathbf{x}))=1,$$

because

$$X(\varphi)(\mathbf{x})=\frac{1}{|\nabla\varphi|^2}\left(\sum_{i=1}^{3}\frac{\partial\varphi}{\partial x_i}\frac{\partial}{\partial x_i}\right)\varphi(\mathbf{x})=1.$$

So by the initial condition $\varphi(\alpha_0(\mathbf{x}))=\varphi(\mathbf{x})$ we get $\varphi(\alpha_s(\mathbf{x}))=s+\varphi(\mathbf{x})$, and if $\mathbf{x}\in\Gamma$, the last term is zero. Thus, we have diffeomorphisms $\alpha_s:\varphi^{-1}(0)\to\varphi^{-1}(s)$. Now equip an open neighborhood $U\subset\varphi^{-1}(0)$ with local coordinates $\mathbf{y}$ by a chart, then α_s will equip $\alpha_s(U)\subset\varphi^{-1}(s)$ with the same coordinates. Thus, we have the desired local coordinates $(\mathbf{y},\varphi)$ on the set $V:=\varphi^{-1}(-t,t)\cap\alpha_{(-t,t)}(U)$, providing there are no critical points in V. If there are, we just take a bounded U or a smaller t. (The incompleteness of X on $\varphi^{-1}[-t,t]$ is not an obstacle to the

argument—just attenuate X with appropriate smooth bump functions to make it integrable with speed 1 on φ, where we want it.) ■

For a function $f \in L^1(\mathbb{R}^n)$, we have

$$\int_V f(\mathbf{x})\ d\mu(\mathbf{x}) = \int_V f(\mathbf{y},\varphi) J(\mathbf{y},\varphi)\ d\mathbf{y} d\varphi,$$

where J is the Jacobian and V is as in the preceding proof. Observe that for a fixed $\varphi = s$ that $J(\mathbf{y},s)d\mathbf{y}$ is not yet the surface measure $d\gamma_s$ on $\varphi^{-1}(s)$. For that the orthogonal coordinate φ needs to be expressed in terms of the length of the curves $\mathbf{y} = \text{const}$. Since $d\varphi = |\nabla\varphi| dl$, with dl the length measure on a curve $\mathbf{y} = \text{const}$, we conclude that $d\gamma_s(\mathbf{y}) = [J \cdot |\nabla\varphi|](\mathbf{y},s) d\mathbf{y}$. Now

$$\int_V |\nabla\varphi(\mathbf{x})| f(\mathbf{x}) d\mu(\mathbf{x}) = \int_V f(\mathbf{y},\varphi)[J \cdot |\nabla\varphi|]\ d\mathbf{y} d\varphi = \int_{-t}^{t} \left(\int_U f \cdot J \cdot |\nabla\varphi| d\mathbf{y} \right) d\varphi, \tag{3.3}$$

and this is the expression we wish to exploit. Let $f \in C_c(\mathbb{R}^n)$, and let the thickness of the shell $\varphi^{-1}[-t,t]$ around Γ approach zero; then

$$\lim_{s\to 0} \frac{1}{2s} \int_{V \cap \varphi^{-1}[-s,s]} |\nabla\varphi| \cdot f d\mu = \lim_{s\to 0} \frac{1}{2s} \int_{-s}^{s} \left(\int_U f \cdot J \cdot |\nabla\varphi| d\mathbf{y} \right) d\varphi$$

$$= \int_U (f \cdot J \cdot |\nabla\varphi|)(\mathbf{y},0) d\mathbf{y},$$

where the use of the fundamental theorem of calculus is justified because the function

$$\varphi \to \int_U (f \cdot J \cdot |\nabla\varphi|)(\mathbf{y},\varphi) d\mathbf{y}$$

is continuous due to the uniform continuity of the integrand—a consequence of $f \in C_c(\mathbb{R}^n)$ and $J \cdot |\nabla\varphi| \in C^\infty(V)$. Thus

$$\lim_{s\to 0} \frac{1}{2s} \int_{V \cap \varphi^{-1}[-s,s]} |\nabla\varphi| \cdot f d\mu = \int_U (f \restriction \Gamma) d\gamma.$$

By doing this for all open patches $U \subset \Gamma$ (equipped with charts), we conclude

$$\lim_{s\to 0} \frac{1}{2s} \int_{\varphi^{-1}[-s,s]} |\nabla\varphi| \cdot f d\mu = \int_\Gamma (f \restriction \Gamma) d\gamma, \tag{3.4}$$

for all $f \in C_c(\mathbb{R}^n)$.

In particular, let $\psi_1, \psi_2 \in C_c(\mathbb{R}^n)$; then we have

$$(\psi_1 \restriction \Gamma, \psi_2 \restriction \Gamma)_{L^2(\Gamma)} = \lim_{t\to 0} \frac{1}{2t} \int_{\varphi^{-1}[-t,t]} \overline{\psi_1} \cdot \psi_2 \cdot |\nabla\varphi| d\mu$$

$$= \lim_{t\to 0} \frac{1}{2t} (\psi_1, |\nabla\varphi| \cdot \chi_{\varphi^{-1}[-t,t]} \psi_2) = \lim_{t\to 0} \frac{1}{2t} (\psi_1, \hat{h} P_t \psi_2), \tag{3.5}$$

where we used the notation $(\hat{h}\psi)(\mathbf{x}) := |\nabla\varphi(\mathbf{x})| \psi(\mathbf{x})$ and $P_t := \hat{\chi}_{\varphi^{-1}[-t,t]}$ and (3.5) is the desired relation between the inner products of the initial space $L^2(\mathbb{R}^n)$ and of the physical space $L^2(\Gamma)$. We use the notation χ_S for the characteristic function of S. (Note that P_t is just the spectral projection of $\hat{\varphi}$ on the interval $[-t,t]$, and that $\hat{h}$ is a positive bounded operator on $P_t\mathscr{H}$.) Now the right-hand side of this equation will exist for a much larger class of functions in $L^2(\mathbb{R}^n)$ than $C_c(\mathbb{R}^n)$, though for some of these the restrictions to Γ on the left-hand side may not be defined. However, the limit in (3.5) will definitely fail to exist for some elements of $L^2(\mathbb{R}^n)$. Define

$$(\psi_1,\psi_2)_\Gamma := \lim_{t\to 0}\frac{1}{2t}(\psi_1,\hat{h}P_t\psi_2), \tag{3.6}$$

for all pairs ψ_1,ψ_2 for which the rhs is defined and finite. To obtain a positive sesquilinear form from (3.6), we need to decide on a dense domain on which $(\cdot,\cdot)_\Gamma$ is defined and finite. Clearly, $C_c(\mathbb{R}^n)$ is one such domain, but it is not maximal. In fact, $(\cdot,\cdot)_\Gamma\restriction C_c(\mathbb{R}^n)$ has no closed extensions as a sesquilinear form (cf. Ref. 26, p. 278), so there is no canonical way of getting a maximal domain. The set

$$S := \{\psi\in L^2(\mathbb{R}^n)\,|\,(\psi,\psi)_\Gamma \text{ exists and is finite}\}$$

is not a linear space since $(\psi_1,\psi_2)_\Gamma$ need not exist for $\psi_1,\psi_2\in S$, as one can verify with easy examples. Nevertheless, some domains containing $C_c(\mathbb{R}^n)$ seem quite natural, e.g., the domain C_φ consisting of those $\psi\in L^2(\mathbb{R}^n)$ for which there is some $t>0$ (depending on ψ) such that $P_t\psi\in C_c(\varphi^{-1}[-t,t])$, and clearly the argument that led to Eq. (3.5) is still valid for $\psi_i\in C_\varphi$. Now C_φ is still not maximal and we can, of course, choose a maximal defining subset for $(\cdot,\cdot)_\Gamma$ in S containing C_φ, but this choice is not unique. Equip the factor space $C_\varphi/\mathrm{Ker}(\cdot,\cdot)_\Gamma$ with the inner product $(\cdot,\cdot)_\Gamma$, making it into a pre-Hilbert space that we complete to obtain the Hilbert space $\mathscr{H}_\Gamma$. We identify $\mathscr{H}_\Gamma$ with $L^2(\Gamma)$ in a natural way:

Theorem 3.7: *With the structures above, denote by* $\mathscr{C}:C_\varphi\to C_\varphi/\mathrm{Ker}(\cdot,\cdot)_\Gamma$ *the factor map; then (1)* $(\mathscr{C}\psi_1,\mathscr{C}\psi_2)_\Gamma=(\psi_1\restriction\Gamma,\psi_2\restriction\Gamma)_{L^2(\Gamma)}$ *for all* $\psi_i\in C_c(\mathbb{R}^n)$, *(2) there is a unitary* $U:\mathscr{H}_\Gamma\to L^2(\Gamma,\gamma)$ *such that* $U\mathscr{C}\psi=\psi\restriction\Gamma$ *for all* $\psi\in C_c(\mathbb{R}^n)$, *and* $(U\mathscr{C})(C_c(\mathbb{R}^n))=C_c(\Gamma)$.

Proof: (1) Since $(\psi_1,\psi_2)_\Gamma=(\mathscr{C}\psi_1,\mathscr{C}\psi_2)_\Gamma$, this is already proven above. (2) Let $\psi\in C_c(\mathbb{R}^n)$ and define $U\mathscr{C}\psi := \psi\restriction\Gamma$, which is in $C_c(\Gamma)$, since Γ is a closed subset of $\mathbb{R}^n$. To see that U is well defined on $C_c(\mathbb{R}^n)/\mathrm{Ker}(\cdot,\cdot)_\Gamma$, let $\psi_1,\psi_2\in C_c(\mathbb{R}^n)$ with $(\psi_1-\psi_2)\restriction\Gamma=0$. Then by (3.4) we find $\psi_1-\psi_2\in\mathrm{Ker}(\cdot,\cdot)_\Gamma$, i.e., $\mathscr{C}\psi_1=\mathscr{C}\psi$ so U is well defined. Moreover, from (1) we see that U is unitary on the dense subspace $\mathscr{C}[C_c(\mathbb{R}^n)]$; hence it extends to a unitary on $\mathscr{H}_\Gamma$. Clearly $C_c(\mathbb{R}^n)\restriction\Gamma=C_c(\Gamma)$, so since the image of U contains a dense subspace and U is unitary, we find the map $U:\mathscr{H}_\Gamma\to L^2(\Gamma)$ is onto. ∎

This means the restriction map can be written as $\kappa := U\circ\mathscr{C}:C_\varphi\to L^2(\Gamma)$, but now it implicitly involves a limiting process, which we will exploit below. Note that $C_c(\mathbb{R}^n)\cap\mathrm{Ker}(\cdot,\cdot)_\Gamma=\{\psi\in C_c(\mathbb{R}^n)\,|\,\psi\restriction\Gamma=0\}$ and that κ is unbounded (w.r.t. the Hilbert space norms).

Example: We want to see what the above structure looks like in the restriction of a quantum particle in $\mathbb{R}^3$ to a sphere. So take $\mathscr{H}=L^2(\mathbb{R}^3)$, $\hat{\mathbf{p}}\psi=i\boldsymbol{\nabla}\psi$, $\hat{\mathbf{q}}\psi(\mathbf{x})=\mathbf{x}\psi(\mathbf{x})$ as usual, and for the constraint $\varphi(\mathbf{x})=|\mathbf{x}|^2-a^2$, $a>0$. Now $(\nabla\varphi)(\mathbf{x})=2\mathbf{x}\neq 0$ except if $\mathbf{x}=0$, and so the critical point of φ is distance a away from $\Gamma=\varphi^{-1}(0)$. We have $\varphi^{-1}[-s,t]=\{\mathbf{x}\in\mathbb{R}^3\,|\,|\mathbf{x}|^2\in[a^2-s,a^2+t]\}$, $s,t>0$. So

$$\begin{aligned}(\psi_1,\psi_2)_\Gamma=\lim_{t\to 0}\frac{1}{2t}(\psi_1,\hat{h}P_t\psi_2)&=\frac{1}{2}\frac{d}{dt}\int_{\varphi^{-1}[-t,t]}2|\mathbf{x}|(\overline{\psi_1}\psi_2)(\mathbf{x})d^3x\Big|_{t=0}\\&=\frac{d}{dt}\int_{\sqrt{a^2-t}}^{\sqrt{a^2+t}}r^2dr\int d\Omega\, r(\overline{\psi_1}\psi_2)(\mathbf{x})\Big|_{t=0},\end{aligned}$$

where we used polar coordinates, and Ω denotes the measure on the unit sphere $\mathbf{S}^2$. For functions of the form $\psi_i(\mathbf{x})=f_i(r)\xi_i(\theta,\phi)$, $\xi_i\in L^2(\mathbf{S}^2)$, and f_i continuous on $(a-\epsilon,a+\epsilon)$ we have, by the fundamental theorem of calculus,

$$(\psi_1,\psi_2)_\Gamma=\left(\frac{d}{dt}\int_{\sqrt{a^2-t}}^{\sqrt{a^2+t}}r^3\overline{f_1}f_2(r)dr\Big|_{t=0}\right)\int\overline{\xi_1}\xi_2(\theta,\phi)d\Omega=a^2\overline{f_1}(a)f_2(a)\int\overline{\xi_1}\xi_2\,d\Omega,$$

as expected from (3.5). Moreover, we have that $\psi_i\in\mathrm{Ker}(\cdot,\cdot)_\Gamma$ whenever $f_i(a)=0$.

One may ask how the method above should be generalized to bigger classes of constraints, so toward that a few remarks.

Remarks: (1) If Γ has corners (so φ is not smooth) we can develop an ''approximation'' to it by a sequence of smooth submanifolds $\{\Gamma_k\}$, but need to choose a notion of convergence for $\Gamma_k \to \Gamma$.

(2) If Γ has edges, we can combine the procedure above with enforcing a Dirac constraint. For example, if we want to constrain a particle in $\mathbb{R}^3$ to the upper hemisphere Γ_+ of the sphere Γ of the last example,

$$\Gamma_+ := \{\mathbf{x} \in \mathbb{R}^3 |\ |\mathbf{x}| = a,\ x_1 \geqslant 0\},$$

then first apply the method above to the constraint $\varphi(\mathbf{x}) = |\mathbf{x}|^2 - a^2$ to obtain a dense subspace of $L^2(\Gamma)$, and follow this by applying a Dirac procedure via (3.1) to the constraint $\zeta(\theta, \phi) = \chi_{[0,\pi/2]}(\theta)$ in $L^2(\Gamma)$.

(3) To constrain a particle in $\mathbb{R}^n$ to a smooth $(n-k)$-dimensional submanifold Γ, we need k independent constraints $\varphi_1, \ldots, \varphi_k$ such that $\Gamma = \varphi_1^{-1}(0) \cap \cdots \cap \varphi_k^{-1}(0)$ with critical points well away from Γ. Now it is clear how to adapt the method based on (3.2)—we choose a curvilinear coordinate system $(\mathbf{y}, \varphi_1, \ldots, \varphi_k)$, where φ_i are the coordinates measured along the trajectories of the gradient fields $\boldsymbol{\nabla}\varphi_i$, and $\mathbf{y}$ are the coordinates along the manifolds $\varphi_1^{-1}(c_1) \cap \cdots \cap \varphi_k^{-1}(c_k)$, $\mathbf{c}$ in some ϵ neighborhood of zero in $\mathbb{R}^k$. Then with J the Jacobian, the measure on Γ is $d\gamma = (J \cdot |\boldsymbol{\nabla}\varphi_1| \cdots |\boldsymbol{\nabla}\varphi_k|)(\mathbf{y}, 0, \ldots, 0) d\mathbf{y}$ so (3.3) becomes, for $f \in C_c(\mathbb{R}^n)$,

$$\lim_{s \to 0} [2^k s_1 \cdots s_k]^{-1} \int_{\varphi^{-1}[-\mathbf{s}, \mathbf{s}]} |\boldsymbol{\nabla}\varphi_1| \cdots |\boldsymbol{\nabla}\varphi_k| \cdot f d\mu = \int_\Gamma (f \restriction \Gamma) d\gamma,$$

where $\varphi^{-1}[-\mathbf{s}, \mathbf{s}] := \varphi_1^{-1}[-s_1, s_1] \cap \cdots \cap \varphi_k^{-1}[-s_k, s_k]$. Further adaptations are straightforward.

Alternatively, we can impose the constraints φ_i one by one, but for this we need to generalize the procedure above for $\mathbb{R}^n$ to general Riemannian manifolds.

(4) For φ that are positive, $\Gamma = \varphi^{-1}(0)$ consists only of critical points, though if other critical points are well away from Γ, there is still a neighborhood of Γ foliated by the level hypersurfaces of φ, and it is possible to adapt the method above to this situation.

(5) In the case where $\varphi^{-1}(0) = \Gamma \cup \Delta$ with $\mu(\Gamma) = 0 \neq \mu(\Delta)$, $\Gamma \cap \Delta = \varnothing$, neither the Dirac method [which will only produce $L^2(\Delta)$] nor the method above seem appropriate. So what we want is a method that will give the method above on Γ, and the Dirac procedure on Δ. (We choose not to allow surface terms on Δ.) Assume the critical points of φ are well away from Γ, and that Δ is the closure of an open set. That is, we want a densely defined map $\kappa: L^2(\mathbb{R}^n) \to L^2(\Gamma) \oplus L^2(\Delta)$. Recalling that $L^2(\Delta) = P_{\{0\}} L^2(\mathbb{R}^n)$, this can be done by defining

$$(\psi_1, \psi_2)_R := (\psi_1, P_{\{0\}} \psi_2) + \lim_{t \to 0} \frac{1}{2t} (\psi_1^N, \hat{h} P_t \psi_2^N),$$

for all $\psi_i \in \mathbf{D}_R := \{\psi \in \mathscr{H} | \psi^N \in C_\varphi\}$, where $\psi^N := \chi_N \cdot \psi$ and N is an open neighborhood of Γ such that $N \cap \Delta = \varnothing$. Then proceed as before, taking $\mathbf{D}_R / \mathrm{Ker}(\cdot, \cdot)_R$ with κ the composition of the factor map and the unitary from $\mathbf{D}_R / \mathrm{Ker}(\cdot, \cdot)_R$ to $L^2(\Gamma) \oplus L^2(\Delta)$.

(6) At the cost of complicating the current exposition, we can enlarge the domain C_φ of κ considerably. For instance, by splitting the limit (3.6) into the average of the limits from above and below in t, and letting the domain consist of those $\psi \in L^2(\mathbb{R}^n)$ for which $\varphi \to \psi(\mathbf{y}, \varphi)$ is continuous from above and below at zero (where we used the curvilinear coordinates of Lemma 3.2), we obtain another useful domain for κ. We chose not to do this because the additional analytic details would have obscured the simplicity of our central idea.

(7) The methods proposed here for dynamics reduction can be extended to other holonomic constraints, i.e., if we are given a constraint $\varphi(\mathbf{q}, \mathbf{p})$, which involves only one member of each canonical pair (q_i, p_i), then a partial Fourier transform can convert it to the type $\varphi(\mathbf{q})$ considered here. So there is a unitary transformation of such a system to one of the present type.

IV. CONSTRAINING THE DYNAMICS

In this section we continue the analysis of the problem of constraining a quantum particle in $\mathbb{R}^n$ to a subset and address the problem of how to constrain the dynamics, i.e., how to construct out of the given time evolution on $L^2(\mathbb{R}^n)$ an acceptable time evolution on the physical Hilbert space. We consider four cases.

(1) For an ordinary Dirac constraining we restrict, for example, a quantum particle in $\mathbb{R}^n$ with Hamiltonian H to live on a set T, which is the closure of an open set S [hence $\mu(T)\neq 0$]. Kinematically, one constrains via the projection $P_{\text{phys}}: L^2(\mathbb{R}^n)\to L^2(T)$, as described at the start of Sec. III. If the dynamics preserve T, i.e., each $U_t := \exp itH$ is of the form

$$U_t = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

with respect to the decomposition $L^2(\mathbb{R}^n)=L^2(T)\oplus L^2(T)^{\perp}$, then there is no problem; one just restricts U_t to $L^2(T)$ to obtain the constrained dynamics. In practice this is a rare occurrence.

(2) For the case of a Dirac constraining as in (1), where the dynamics does not preserve T, e.g., if the particle is free, $H=\hat{\mathbf{p}}^2/2m$ and T is compact, then $H\restriction L^2(T)$ is not self-adjoint. To equip the constrained particle with a time evolution, we need to choose some self-adjoint extension of the symmetric operator $H\restriction C_c^{\infty}(S)$ [for this to make sense we need to assume that H preserves $C_c^{\infty}(S)$, which will be true if it is a differential operator]. This amounts to the choice of boundary conditions, i.e., deciding how the particle should behave at the walls (e.g., reflection), and this is a physical choice that cannot be determined from mathematical considerations alone.

(3) For the problem of the last section where we constrain a particle in $\mathbb{R}^n$ with Hamiltonian H to a lower-dimensional submanifold Γ, we will need to assume that H is ''smooth enough'' near Γ, i.e., $H(\text{Dom } H\cap C_{\varphi})\subset C_{\varphi}$ and $\kappa(\text{Dom } H\cap C_{\varphi})$ is dense in $L^2(\Gamma)$. In the case where H ''preserves'' Γ, i.e., H preserves Dom $H\cap$Ker κ, H will clearly lift through κ, and we only need to define the constrained Hamiltonian H_{Γ} on $L_2(\Gamma)$ by $H_{\Gamma}\kappa(\psi) := \kappa(H\psi)$ for all $\psi \in \text{Dom } H\cap C_{\varphi}$, and require that it is essentially self-adjoint.

(4) The case of constraining a particle as in (3), but where H does not preserve Γ, will be our object of study for the rest of this section. By analogy with case (2), we expect that some physical choices will need to be made. When H is a Schrödinger operator with smooth potential, physicists already know what to do. One lets the constrained Hamiltonian be $(1/2m)\Delta_{\Gamma}+V\restriction\Gamma$, where Δ_{Γ} is the Laplacian on Γ.

Henceforth, we assume that H on $L^2(\mathbb{R}^n)$ does not preserve Γ in the sense of (3) above. Recalling that $\kappa: C_{\varphi}\to L^2(\Gamma)$ involves a factorization, it is natural to look for linear sections of κ that one can use to construct liftings of operators on C_{φ} to $L^2(\Gamma)$. The image of such a section deserves a name of its own.

Definition: A *transverse space* is a linear space $\mathscr{T}\subset C_{\varphi}$ such that Ker $\kappa\cap\mathscr{T}=\{0\}$ and $\kappa(\mathscr{T})$ is dense in $L^2(\Gamma)$.

Transverse spaces are abundant, and below we will use physical arguments to make a choice of a transverse space but first we show how one can use a transverse space to constrain H. Given a transverse space $\mathscr{T}\subset\text{Dom } H\cap C_{\varphi}$ such that $H.\mathscr{T}\subset C_{\varphi}$, we have the following options.

(i) Define the constrained Hamiltonian $H_{(i)}$ with dense domain $\kappa(\mathscr{T})$ by

$$H_{(i)}\kappa(\psi) := \kappa(H\psi), \quad \forall\,\psi\in\mathscr{T}.$$

This is well defined since, by definition of transverse spaces, κ is injective on $\mathscr{T}$.

(ii) Given that the choice of $\mathscr{T}$ has some physical content (see below), one may object to (i) in that H is not forced to preserve $\mathscr{T}$ in any sense. If we want to include such a restriction, since it is only the behavior near Γ that should be important, we can restrict H to $\mathscr{T}$ ''in the limit,'' i.e., we consider the projection $P^t_{\mathscr{T}}H$ as t goes to zero in the definition of κ, where $P^t_{\mathscr{T}}$ denotes the projection onto $P_t.\mathscr{T}$. So we propose an alternative constrained Hamiltonian $H_{(ii)}$ on $\kappa(\mathscr{T})$ by

$$(H_{(ii)}\kappa(\psi),\ \kappa(\phi))_{L^2(\Gamma)}=\lim_{t\to 0}\frac{1}{2t}\,(P^t_{\mathscr{T}}H\psi,\ \hat{h}P_t\phi),$$

for all $\psi,\ \phi\in\mathscr{T}$.

Below we will test both methods for a suitable choice of transverse space. This choice is the issue we now want to address. Recall that in the classical global method of Gotay, Nester, and Hinds,[2] the idea was to adjust the dynamics so that the motion is always tangential to Γ, thus forcing the particle to remain on Γ. We look for some quantum mechanical version of this.

First we want to give meaning to the concept of motion ''tangential to Γ'' for a quantum particle. While the classical state of a particle is a point in phase space, quantum mechanically a state is here a vector ψ in $L^2(\mathbb{R}^n)$. Classically, a particle with constant mass at position $\mathbf{q}$ moves tangentially to Γ if its momentum $\mathbf{p}=m\dot{\mathbf{q}}$ is tangential to Γ, i.e., $\mathbf{p}\in T_{\mathbf{q}}\Gamma$. Inspired by this, in a quantum mechanical setting, we say a particle in a state $\psi\in C_c^\infty(\mathbb{R}^n)$ moves tangentially to Γ if $(\hat{\mathbf{p}}\psi)(\mathbf{q})\in T_{\mathbf{q}}\Gamma$ for all $\mathbf{q}\in\Gamma$. Now, given the normal vectors $\boldsymbol{\nabla}\varphi$ to Γ, we have $\mathbf{v}\in T_{\mathbf{q}}\Gamma$ iff $\boldsymbol{\nabla}\varphi(\mathbf{q})\cdot\mathbf{v}=0$. So a particle in a state $\psi\in C_c^\infty(\mathbb{R}^n)$ moves tangentially to Γ when $(\boldsymbol{\nabla}\varphi\cdot\hat{\mathbf{p}}\psi)(\mathbf{x})=0$ for all $\mathbf{x}\in\Gamma$, i.e., iff the component of the momentum of ψ normal to Γ is zero. We would like to generalize this notion away from $C_c^\infty(\mathbb{R}^n)$. Since Γ is of μ measure zero, we cannot, in general, specify a property for an L^2 function $\psi\in L^2(\mathbb{R}^n)$ on Γ. Recall that κ is a limiting procedure on the shells $\varphi^{-1}[-t,t]$, in which eventually $t\to 0$. So it may be enough to define tangentiality on the level sets of φ in such a shell, and depend on κ to restrict the property to Γ. Fix a shell $S_{t_0}:=\varphi^{-1}[-t_0,t_0]$ around Γ; then we say a $\psi\in\mathrm{Dom}(\hat{\mathbf{p}}\restriction L^2(S_{t_0}))$ has momentum tangential to the level sets of φ iff $\boldsymbol{\nabla}\varphi\cdot\hat{\mathbf{p}}\psi=0$. Since in terms of the local coordinate system $(\mathbf{y},\varphi)$ of Sec. III this means $\partial\psi/\partial\varphi=0$ on S_{t_0}, we see that $\psi\restriction S_{t_0}$ must be constant in the normal direction (i.e., along the trajectories of the vector field $\boldsymbol{\nabla}\varphi$). So $\psi\restriction S_{t_0}$ is uniquely determined by its restriction to Γ. This defines the notion of ''states with momentum tangential to the level sets of φ in S_{t_0}.''

Conversely, given any $\psi\in L^2(\Gamma)$, we can make out of it a state $\tilde{\psi}\in L^2(\mathbb{R}^n)$ by extending it constantly along the normals in S_{t_0} and set it equal to zero outside S_{t_0}, i.e.,

$$\tilde{\psi}(\mathbf{x})=\psi(\alpha^{-1}_{\varphi(\mathbf{x})}(\mathbf{x})),\qquad \text{for } \mathbf{x}\in S_{t_0} \text{ and zero for } \mathbf{x}\notin S_{t_0}$$

(cf. the proof of Lemma 3.2 for α). Denote the space of these by $\mathscr{H}^T_{t_0}$, and the projection onto $\mathscr{H}^T_{t_0}$ by $P^T_{t_0}$. Then $\mathscr{H}^T_{t_0}$ is thought of as a ''thickening'' of $L^2(\Gamma)$ in $L^2(\mathbb{R}^n)$, and note that $\mathscr{H}^T_{t_0}\cap C_\varphi$ is a transverse space. However, when we deal with differential operators, the discontinuity at the boundary of the shell S_{t_0} is a problem, so we prefer the following smooth version. Let $\zeta_{t_0}\in C^\infty(\mathbb{R})$ be a bump function that is unity on $[-t_0,t_0]$ and zero outside $[-t_0-\epsilon,t_0+\epsilon]$ for a given ϵ. Then define for a $\psi\in L^2(\Gamma)$ the new thickening,

$$\check{\psi}(\mathbf{x})=\zeta_{t_0}(\varphi(\mathbf{x}))\cdot\psi(\alpha^{-1}_{\varphi(\mathbf{x})}(\mathbf{x})).$$

Denote the space spanned by these by $\mathscr{H}^\zeta_{t_0}\subset L^2(\mathbb{R}^n)$, then clearly $P_{t_0}\mathscr{H}^\zeta_{t_0}=\mathscr{H}^T_{t_0}$.

We now want to constrain the Hamiltonian in such a way that it can be thought of as ''projecting the force down to its tangential component,'' where a Hamiltonian tangential to Γ will be one that keeps tangential motion to Γ tangential, i.e., preserves the tangential states on some shell S_{t_0}. Recall that in the GNH algorithm,[2] one restricted the Hamiltonian to all states with motion tangential to Γ. To do the same here, we can now use either of the two proposed methods (i) or (ii) above with the choice of transverse space as

$$\mathscr{T}_{t_0}:=\mathscr{H}^\zeta_{t_0}\cap C_\varphi\cap\mathrm{Dom}\ H.$$

As long as $C_c^\infty(\mathbb{R}^n)\subset\mathrm{Dom}\ H$, we have that $\kappa(\mathscr{T}_{t_0})$ is dense, in which case this is indeed a transverse space. So assuming the latter, and that $H\mathscr{T}_{t_0}\subset C_\varphi$, we define

$$H_{(i)}\cdot\kappa(\psi):=\kappa(H\psi),\quad \forall\,\psi\in\mathscr{T}_{t_0}. \tag{4.1}$$

In fact, when ψ is smooth and H is a differential operator, this becomes $H_{(i)}(\psi\restriction\Gamma)=(H\psi)\restriction\Gamma$, and then the method is nothing but the one used by Helgason, pp. 251–252[27] for the restriction of a differential operator to a submanifold. This is the first reasonable method to consider.

To motivate the use of method (ii), consider how a physicist might object to (4.1). Since H need not preserve $\mathscr{T}_{t_0}$, it is possible that $H\psi$ has momentum that is not tangential to Γ, in which case it seems there is a force acting on the particle, forcing it off Γ. To project this force out of the total force acting on the particle, the restriction procedure above may not be appropriate. Instead, we intend to use the Hilbert space projections to first project out of H the parts that affect the momentum component normal to the level surfaces of φ in the shell $\varphi^{-1}[-t_0,t_0]$; then in the limit of κ when the thickness of the shell goes to zero, we will be left with the appropriate Hamiltonian on $L^2(\Gamma)$ (having maintained tangentiality of momentum to Γ during the limiting process). In the light of this we define $H_{(ii)}$ by

$$(H_{(ii)}\cdot\kappa(\psi_1),\ \kappa(\psi_2))_{L^2(\Gamma)}:=\lim_{t\to 0}\frac{1}{2t}\,(P_t^T H\psi_1,\ \hat{h}P_t\psi_2), \tag{4.2}$$

for all $\psi_1\in\mathscr{T}_{t_0}$ for which this exists for all $\psi_2\in\mathscr{T}_{t_0}$ [so that (4.2) defines a vector $H_{(ii)}\kappa(\psi_1)\in L^2(\Gamma)$]. Since $P_t\overline{\mathscr{T}}_{t_0}\subset\mathscr{H}_t^T$, (4.2) coincides with method (ii). This is the second reasonable method.

One can easily conceive of other methods apart from (i) and (ii) for constraining dynamics, for instance, forcing the particle onto Γ by infinite potential walls.

(iii) Given a system $L^2(\mathbb{R}^n)$, φ, H as before, restrict the particle to a box around Γ, say the shell S_t, but for it to be well defined we need to assume $HC_c^\infty(S_t)\subseteq C_c^\infty(S_t)$ and that we have a self-adjoint extension H_t on $L^2(S_t)$ of the symmetric operator $H\restriction\{\psi\in C_c^\infty(S_t)\,|\,\psi(\varphi^{-1}(\pm t))=0\}$. In the case when H is a differential operator, this means one needs to decide how it behaves at the walls. Henceforth, we assume that H is a differential operator. If we choose ordinary reflection at the walls (quantum billiards), then a $\psi\in\mathrm{Dom}\,H_t$ must satisfy

$$\frac{\partial\psi}{\partial\varphi}=0,\quad \text{when}\quad \varphi=\pm t,$$

but without more detailed knowledge of H we do not know what additional boundary conditions ψ must satisfy to make H_t self-adjoint. So Dom H_t is a subspace of

$$\mathscr{D}_t:=\left\{\psi\in P_tC^\infty(\mathbb{R}^n)\,\middle|\,\psi\in L^2(\mathbb{R}^n),\quad \frac{\partial\psi}{\partial\varphi}=0,\quad \text{when}\quad \varphi(\mathbf{x})=\pm t\right\}\subset C_\varphi.$$

Then, in the limit when t goes to zero, the boxes force the particle (hence H_t) onto Γ. The appropriate way to take this limit, is through κ, i.e.,

$$(H_{(iii)}\cdot\kappa(\psi_1),\ \kappa(\psi_2))_{L^2(\Gamma)}:=\lim_{t\to 0}\frac{1}{2t}\,(H_tP_t\psi_1,\ \hat{h}P_t\psi_2), \tag{4.3}$$

where $\psi_i\restriction S_t\in\mathrm{Dom}\,H_t$ for all $t\in(0,t_0)$ and some fixed t_0. Note that the possible ψ_i's are restricted by the behavior at the boundary. So here for reflection, we get that $(\partial\psi_i/\partial\varphi)(\mathbf{y},t)=0$ for all $t\in[-t_0,0)\cup(0,t_0]$. We recognize that $P_t\psi$ is in our transverse space of earlier on. When the space of $\kappa(\psi)$ for such ψ is dense in $L^2(\Gamma)$, $H_{(iii)}$ is well defined. Other versions of this are possible if we change the behavior at the walls, e.g., introduce a phase with the reflection.

We will not discuss any other methods, but now the surprise is that we have the following equivalences.

Theorem 4.4: (1) *With the assumptions above,* $H_{(i)}=H_{(ii)}=:H_\Gamma$; (2) *moreover, if H is a differential operator, then* $H_{(i)}=H_{(iii)}$ *on Dom* $H_{(iii)}$.

Proof: (1) We first show $H_{(i)}=H_{(ii)}$. Recall that $C_c^\infty(\mathbb{R}^n)\subset\mathrm{Dom}\,H$, and that $\kappa(\mathscr{T}_{t_0}\cap C_c^\infty(\mathbb{R}^n))=C_c^\infty(\Gamma)$; hence it suffices to show that

$$((H_{(i)}-H_{(ii)})\kappa(\psi_1),\ \kappa(\psi_2))_{L^2(\Gamma)}=0,$$

for all $\psi_i \in \mathscr{T}_{t_0} \cap C_c^\infty(\mathbb{R}^n)$. By (4.1) and (4.2), the left-hand side is

$$\lim_{t\to 0} \frac{1}{2t} ((\mathbb{I}-P_t^T)H\psi_1,\ \hat{h}P_t\psi_2) = \lim_{t\to 0} \frac{1}{2t} \int_{S_t} \overline{(\mathbb{I}-P_t^T)H\psi_1}\psi_2 |\nabla\varphi| d\mu.$$

Partition Γ into patches U on which local coordinates exist as in Lemma 3.2; then the right-hand side becomes a finite sum of terms:

$$\lim_{t\to 0} \frac{1}{2t} \int_{-t}^{t} d\varphi \int_U d\mathbf{y} \overline{(\mathbb{I}-P_t^T)H\psi_1}\psi_2 |\nabla\varphi| \qquad (*)$$

(alternatively, take ψ_2 with support in U and let U vary). Now $(\mathbb{I}-P_t^T)H\psi_1 =: \rho_t \perp \mathscr{H}_t^T$, so for all $\phi \in \mathscr{H}_t^T \cap C_c^\infty(S_t)$ we have

$$0 = (\rho_t, \phi) = \int_U d\mathbf{y}\phi(\mathbf{y}) \int_{-t}^{t} d\varphi\ \overline{\rho_t},$$

where we used the fact that on S_t, ϕ is independent of φ, so since this holds for all ϕ [which will span a dense subspace of $L^2(U)$], we have that $\int_{-t}^{t} \rho_t(\mathbf{y},\varphi)d\varphi = 0$, a.e. in $\mathbf{y}$. Now using the continuity of $\psi_2|\nabla\varphi|$, $(*)$ becomes

$$\lim_{t\to 0} \frac{1}{2t} \int_U d\mathbf{y}(\psi_2|\nabla\varphi|)(\mathbf{y},0) \int_{-t}^{t} d\varphi\ \overline{\rho_t} = 0.$$

Since this holds for all U, we find that $H_{(i)} = H_{(ii)}$.

(2) Next we prove that $H_{(i)} = H_{(iii)}$ on $\mathrm{Dom}\, H_{(iii)}$. Clearly $\cap_{t\in[0,t_0]} \mathscr{D}_t \restriction S_{t_0} = \mathscr{H}_{t_0}^T$, so it suffices to show that

$$((H_{(i)} - H_{(ii)})\kappa(\psi_1),\ \kappa(\psi_2))_{L^2(\Gamma)} = 0, \quad \forall \psi_i \in C_c^\infty(S_{t_0}) \cap \Big(\bigcap_{t\in(0,t_0]} \mathrm{Dom}\, H_t\Big),$$

i.e., that

$$\lim_{t\to 0} \frac{1}{2t} ((H - H_tP_t)\psi_1,\ \hat{h}P_t\psi_2) = 0.$$

Now

$$((H - H_tP_t)\psi_1,\ \hat{h}P_t\psi_2) = ((P_tH - H_tP_t)\psi_1,\ \hat{h}P_t\psi_2),$$

and let

$$\mathscr{Q}_t := \{\phi \in C_c^\infty(S_t) \mid \phi(\varphi^{-1}(\pm t)) = 0\},$$

which is dense in $L^2(S_t)$, and $H = H_t$ on $\mathscr{Q}_t$. Now, given an open precompact set $X \subset S_t$, we can write any $\phi \in C_c^\infty(\mathbb{R}^n)$ as $\phi = \phi_X + \tilde{\phi}_X$, where $\phi_X \in \mathscr{Q}_t$ and $\phi \restriction X = \tilde{\phi}_X \restriction X$. Let $\psi \restriction S_t \in \mathrm{Dom}\, H_t \cap C_c^\infty(S_t)$ for all $t \in (0,t_0]$, let $\phi \in \mathscr{Q}_t$ and X be the interior of supp ϕ. Then

$$\begin{aligned}((P_tH - H_tP_t)\psi, \phi) &= ((P_tH - H_tP_t)(\psi_X + \tilde{\psi}_X), \phi) \\ &= ((H - H_t)\psi_X + (P_tH - H_tP_t)\tilde{\psi}_X, \phi) \\ &= ((P_tH - H_tP_t)\tilde{\psi}_X, \phi).\end{aligned} \qquad (\dagger)$$

Since H and H_t are differential operators, they preserve supports; hence $\mathrm{supp}((P_tH - H_tP_t)\tilde{\psi}_X) \subseteq \mathrm{supp}\ \tilde{\psi}_X$, which is disjoint from the support of ϕ, hence, since the inner product in (†) is an integral, we conclude that (†) is zero, i.e., $(P_tH - H_tP_t)\psi \perp \mathscr{Q}_t$; hence $(P_tH - H_tP_t)\psi = 0$. ∎

We remark that any reasonable method of dynamics reduction should produce a constrained Hamiltonian H_Γ on $L^2(\Gamma)$ that is essentially self-adjoint, and the time evolution it generates, $\exp itH_\Gamma$, must preserve the algebra of observables $\mathscr{R}_S$ when we have obtained this (see the next section). We now test the above methods on two examples.

Example 4.5: Recall the previous example of the sphere of radius a, $\Gamma = \mathbf{S}_a^2$ in $\mathbb{R}^3$; $\mathscr{H} = L^2(\mathbb{R}^3) = L^2(\mathbb{R}_+, r^2dr) \otimes L^2(\mathbf{S}_1^2, d\Omega)$ with Hamiltonian $H = \hat{\mathbf{p}}^2/2m = -(1/2m)\Delta$, having domain of essential self-adjointness $C_c^\infty(\mathbb{R}^3)$ and that acts on the decomposable functions $\psi(\mathbf{x}) = f(r) \cdot \xi(\theta, \phi) \in C_c^\infty(\mathbb{R}^3)$ by

$$(-\Delta\psi)(\mathbf{x}) = \left(-\frac{d^2}{dr^2} - \frac{2}{r} \cdot \frac{d}{dr} \right) f(r) \cdot \xi(\theta, \phi) + \frac{1}{r^2} f(r) \cdot B\xi(\theta, \phi),$$

where $B = |\mathbf{L}|^2$ is the Laplace–Beltrami operator on $L^2(\mathbf{S}_1^2)$. Now for $t \in (0, a)$ we have

$$\mathscr{H}_t^T := \{\psi \in \mathscr{H} \mid \psi(\mathbf{x}) = \chi_{[\sqrt{a^2-t}, \sqrt{a^2+t}]}(r) \cdot \xi(\theta, \phi), \quad \xi \in L^2(\Gamma)\}.$$

To obtain the smoothed space, choose a $\zeta_t \in C_c^\infty(\mathbb{R})$, which is one on $[-t, t]$ and zero on $[-t-\epsilon, t+\epsilon]$, and set

$$\mathscr{H}_t^\zeta := \{\psi \in \mathscr{H} \mid \psi(\mathbf{x}) = \zeta_t(r^2 - a^2) \cdot \xi(\theta, \phi), \quad \xi \in L^2(\Gamma)\},$$

and note that if $\psi \restriction \Gamma = \xi \in C^\infty(\Gamma)$, then $\psi \in C^\infty(\mathbb{R}^n) \subset \mathrm{Dom}\ H$. So we choose the transverse space $\mathscr{T}_t = \{\psi \in \mathscr{H}_t^\zeta \mid \psi \restriction \Gamma \in C^\infty(\Gamma)\}$. Now for method (i) for constraining H, let $\psi = (\zeta_t \circ \varphi) \cdot \xi \in \mathscr{T}_t$. Then from the explicit formula for H, we see

$$(H\psi)(\mathbf{x}) = \frac{1}{2mr^2} (B\xi)(\theta, \phi) \cdot \chi_{(\sqrt{a^2-t}, \sqrt{a^2+t})}(r) + \rho(\mathbf{x}),$$

where ρ is a function with support disjoint from $\varphi^{-1}(-t, t)$. So, since $H\psi$ is continuous near Γ, we have $\kappa(H\psi) = (H\psi) \restriction \Gamma$, i.e., $\kappa(H\psi)(\theta, \phi) = (B\xi)(\theta, \phi)/2ma^2$. So

$$H_{(i)}\kappa(\psi) = \kappa(H\psi) = \frac{1}{2ma^2} B\kappa(\psi),$$

for all $\psi \in \mathscr{T}_t$, so, since $\kappa(\mathscr{T}_t) = C^\infty(\Gamma)$ is dense in $L^2(\Gamma)$, $H_{(i)}\xi = (1/2ma^2)B\xi$ for all $\xi \in C^\infty(\Gamma)$ is densely defined, agrees with the classical Hamiltonian $|\mathbf{L}|^2/2ma^2$ obtained in Sec. II [Eq. (2.4)], and is essentially self-adjoint: a good result.

Next, as an exercise in method (ii), we calculate

$$(P_t^T H\psi_1, \hat{h} P_t \psi_2), \quad \psi_i \in \mathscr{T}_{t_0},$$

where $\psi_i = (\zeta_t \circ \varphi) \cdot \xi_i$ as above. Now

$$(P_t^T H\psi_1)(\mathbf{x}) = \left(P_t^T \frac{B}{2mr^2} \right) \xi(\theta, \phi),$$

if $r \in [\sqrt{a^2-t}, \sqrt{a^2+t}] =: I_t$ and zero otherwise. To work out the projection, we only need to consider the radial coordinate. The one-dimensional space N generated by $\chi_{I_t} \in L^2(\mathbb{R}_+, r^2dr)$ corresponds to $\mathscr{H}_t^T$. So we need to decompose the function $\chi_{I_t}(r)/2mr^2$ according to $N \oplus N^\perp$, i.e., $(1/2mr^2)\chi_{I_t}(r) = \lambda_t \cdot \chi_{I_t}(r) + h(r)$, where λ_t is a constant and $h \in N^\perp$, i.e., $\int_{I_t} r^2 h(r) dr = 0$. Now $(\chi_{I_t}, (1/2mr^2)\chi_{I_t}) = (1/2m)\int_{I_t} dr = \lambda_t \int_{I_t} r^2 dr$, i.e.,

$$\lambda_t = \frac{3(\sqrt{a^2+t}-\sqrt{a^2-t})}{2m[(a^2+t)^{3/2}-(a^2-t)^{3/2}]},$$

and so

$$(P_t^T H\psi_1)(\mathbf{x}) = \lambda_t \cdot \chi_{I_t}(r)\cdot (B\xi_1)(\theta,\phi).$$

Thus

$$\begin{aligned}(P_t^T H\psi_1, \hat{h}P_{[-t,t]}\psi_2) &= \int \lambda_t \cdot \chi_{I_t}(r)(B\overline{\xi_1})(\theta,\phi)\cdot 2r\psi_2(\mathbf{x})r^2\, dr\, d\Omega \\ &= 2\lambda_t \int_{\sqrt{a^2-t}}^{\sqrt{a^2+t}} r^3 dr \int (B\overline{\xi_1})\cdot \xi_2 d\Omega \\ &= 2\lambda_t \frac{1}{4a^2}((a^2+t)^2-(a^2-t)^2)\cdot (B\xi_1,\xi_2)_{L^2(\Gamma)} = 2t\lambda_t \cdot (B\xi_1,\xi_2)_{L^2(\Gamma)},\end{aligned}$$

where we used the fact that the measure on Γ is $a^2 d\Omega$ with $d\Omega$ the usual measure on the unit sphere. Now

$$\lim_{t\to 0}\lambda_t = \frac{3}{2m}\lim_{t\to 0}\frac{\sqrt{a^2+t}-\sqrt{a^2-t}}{(a^2+t)^{3/2}-(a^2-t)^{3/2}} = \frac{1}{2ma^2},$$

hence

$$\begin{aligned}\lim_{t\to 0}\frac{1}{2t}(P_t^T H\psi_1, \hat{h}P_{[-t,t]}\psi_2) &= \lim_{t\to 0}\lambda_t (B\xi_1,\xi_2)_{L^2(\Gamma)} \\ &= \left(\frac{B}{2ma^2}\,\xi_1,\xi_2\right)_{L^2(\Gamma)} \\ &= (H_{(ii)}\xi_1,\xi_2), \quad \text{for all } \xi_i \in C^\infty(\Gamma).\end{aligned}$$

Thus, $H_{(ii)} = B/2ma^2 = H_{(i)} = H_\Gamma$ on $C^\infty(\Gamma)$, in agreement with Theorem 4.4.

Without much extra effort, we can also show that Schrödinger operators $H = (1/2m)\hat{\mathbf{p}}^2 + V(\hat{\mathbf{q}})$ for V smooth near Γ, produce, for $\xi \in C^\infty(\Gamma)$,

$$H_{(i)}\xi(\theta,\phi) = H_{(ii)}\xi(\theta,\phi) = \left(\frac{B}{2ma^2}\,\xi\right)(\theta,\phi) + V(a,\theta,\phi)\cdot\xi(\theta,\phi).$$

Example 4.6: We would like to restrict a free quantum particle in $\mathbb{R}^3$ to a cylinder Γ of radius a around the z axis. We do this in exact analogy with the sphere. Let the constraint be $\varphi(\mathbf{x}) = x^2+y^2-a^2$; then $\nabla\varphi(\mathbf{x}) = (2x,2y,0)$; hence the critical points are well away from Γ. So

$$\varphi^{-1}[-s,t] = \{\mathbf{x}\in\mathbb{R}^3 | r^2 \in [a^2-s, a^2+t]\},$$

where we henceforth use cylindrical coordinates (z,r,θ). Then

$$(\psi_1,\psi_2)_\Gamma = \lim_{t\to 0}\frac{1}{2t}(\psi_1,\hat{h}P_t\psi_2) = \frac{1}{2}\frac{d}{dt}\int_{\sqrt{a^2-t}}^{\sqrt{a^2+t}} 2r^2 dr \int_{-\infty}^{\infty} dz \int_0^{2\pi} d\theta\, (\overline{\psi_1}\psi_2)(\mathbf{x})|_{t=0}.$$

Let ψ_i be decomposable, $\psi_i(\mathbf{x}) = f_i(r)\xi_i(z,\theta)$, where $\xi_i \in L^2(\Gamma)$ and f_i is continuous at $r=a$. Then

$$(\psi_1,\psi_2)_\Gamma = \int \overline{\xi_1}\xi_2\, dz d\theta \cdot \frac{d}{dt}\int_{\sqrt{a^2-t}}^{\sqrt{a^2+t}} r^2 \overline{f_1} f_2(r) dr|_{t=0} = a\overline{f_1}(a) f_2(a)\int \overline{\xi_1}\xi_2\, d\theta dz.$$

The free Hamiltonian $H=-\Delta/2m$ in cylindrical coordinates acts by

$$\Delta\psi=\frac{1}{r}\frac{\partial}{\partial r}\left(r\,\frac{\partial\psi}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2\psi}{\partial\theta^2}+\frac{\partial^2\psi}{\partial z^2}.$$

For $t\in(0,a)$, we have

$$\mathscr{H}_t^T=\{\psi\in\mathscr{H}\mid\ \psi(\mathbf{x})=\chi_{I_t}(r)\cdot\xi(z,\theta),\xi\in L^2(\Gamma)\},$$

where $I_t:=[\sqrt{a^2-t},\sqrt{a^2+t}]$. Then with a bump function $\zeta_t\in C^\infty(\mathbb{R})$, which is unity on $[-t,t]$ and zero outside $[-t-\epsilon,t+\epsilon]$, we have

$$\mathscr{H}_t^\xi:=\{\psi\in\mathscr{H}\mid\ \psi(\mathbf{x})=\zeta_t(r^2-a^2)\cdot\xi(z,\theta),\xi\in L^2(\Gamma)\}.$$

Since the smooth functions of compact support are in the domain of H, we choose

$$\mathscr{T}_t=\{\psi\in\mathscr{H}_t^\xi\mid\ \psi\restriction\Gamma\in C_c^\infty(\Gamma)\}.$$

Now to obtain $H_{(i)}$, let $\psi=(\zeta_t\circ\varphi)\cdot\xi\in\mathscr{T}$, then

$$(H\psi)(\mathbf{x})=\left(\frac{-1}{2m}\right)\chi_{J_t}(r)\left(\frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}+\frac{\partial^2}{\partial z^2}\right)\xi(z,\theta)+\rho(\mathbf{x}),$$

where $J_t:=(\sqrt{a^2-t},\sqrt{a^2+t})$ and ρ has support disjoint from J_t. Now, since $\kappa(H\psi)=(H\psi)\restriction\Gamma$, we get

$$H_{(i)}\xi(z,\theta)=\left(\frac{-1}{2m}\right)\left(\frac{1}{a^2}\frac{\partial^2}{\partial\theta^2}+\frac{\partial^2}{\partial z^2}\right)\xi(z,\theta),$$

for all $\xi\in C_c^\infty(\Gamma)$. Thus, via Theorem 4.4,

$$H_\Gamma=\left(\frac{-1}{2m}\right)\left(\frac{1}{a^2}\frac{\partial^2}{\partial\theta^2}+\frac{\partial^2}{\partial z^2}\right),$$

on $C_c^\infty(\Gamma)$. This is precisely the quantization one would expect of the constrained classical Hamiltonian (2.6).

Remarks: (1) We regard the choice of a transverse space $\mathscr{T}$ as a decision on the direction from which H should be reduced to Γ. Above, we chose the normal direction. In quantum systems with less geometry (e.g., no metric on an underlying manifold), it may be difficult to decide on an appropriate $\mathscr{T}$. This choice has physical content, and seems analogous to the choice of a self-adjoint extension for a Hamiltonian in the Dirac approach sketched early in this section. An alternative way of expressing the choice of the transverse space is to observe that on $L^2(S_{t_0})$ for t_0 small enough, the pair of operators $\hat{\varphi}$ and $P_\varphi:=i\partial/\partial\varphi$ form a canonical pair, i.e., $[P_\varphi,\hat{\varphi}]\psi=i\psi$ for all smooth ψ, with compact support in the interior of S_{t_0}. Then $\mathscr{T}\restriction S_{t_0}=\mathrm{Ker}\ P_\varphi\cap\mathrm{Dom}(H\restriction L^2(S_{t_0}))$. So, in a general quantum system one can look for a given ''local canonical conjugate'' to the constraint to obtain a transverse space as its kernel. In fact, this observation also gives a clue on how to enforce second-class constraints, in the sense that if we are given a canonical pair φ, P_φ to enforce, then we do it as above, by using P_φ to *locally* select a transverse space to $\varphi^{-1}(0)$.

(2) Recall that in Sec. II, the classical secondary constraint we obtained for φ was $(\nabla\varphi)\cdot\mathbf{p}\restriction\Gamma$. The selection of the normal transverse space above can be thought of as the enforcement of the constraint $\nabla\varphi\cdot\hat{\mathbf{p}}\psi=0$ near Γ, which looks very much like what one would expect a ''quantization'' of the classical secondary constraint to be. This leads one to ask whether we can quantize the infinitesimal Dirac procedure that produced this secondary constraint as a whole. Unfortunately, this does not seem to work; we sketch how it goes awry. Given the primary constraint $\hat{\varphi}$ and the Hamiltonian $H=(1/2m)\hat{\mathbf{p}}^2+V(\hat{\mathbf{q}})$, the secondary constraint should be $[H,\hat{\varphi}]$ ''restricted'' to Γ. Now

$$[H,\hat{\varphi}]\psi = -\frac{1}{2m}(\Delta\varphi)\psi - \frac{1}{m}(\nabla\varphi)\cdot(\nabla\psi),$$

for all $\psi \in C_c^\infty(\mathbb{R}^n)$. Restriction to Γ is done by κ. However, while $-(1/m)(\nabla\varphi)\cdot(\nabla\psi)$ is proportional to the desired secondary constraint $\nabla\varphi\cdot\hat{\mathbf{p}}\psi$, the term $-(1/2m)(\Delta\varphi)\psi$ need not vanish on Γ. Moreover, if we use $[H,\hat{\varphi}]$ as a new secondary constraint to select the transverse space instead of $\nabla\varphi\cdot\hat{\mathbf{p}}$, we obtain the wrong result on the sphere. Nevertheless, given the close parallel that the selection of the transverse space has to the classical secondary constraint, we regard the use of a transverse space here as the imposition of a secondary quantum constraint.

(3) We remark that one needs to reduce the dynamics infinitesimally, i.e., through the Hamiltonian, not directly on the time evolution unitaries, $\exp(itH)$.

(4) In general, there is no guarantee that the constraining of a self-adjoint Hamiltonian H to Γ produces an essentially self-adjoint operator on $L^2(\Gamma)$. This is a difficult question that needs further investigation, and its classical equivalent in the GNH algorithm also is unsolved, that is, we do not know whether reducing a given complete Hamiltonian vector field to a submanifold Γ produces a complete vector field on Γ. When Γ has edges, it is very easy to get examples of complete vector fields that will not be complete when constrained to Γ.

(5) The confining potential approach, developed in Refs. 15 and 16, also treats the dynamics of quantum particles on (or near) surfaces. In this, one assumes there is a confining potential depending only on the normal coordinates, and with a deep minimum on Γ. The system is then taken to be in a fixed eigenstate in the normal direction (''normal degrees of freedom are frozen''), and this was applied to some physical examples where such confining potentials actually occur.[15] By contrast, our object is to impose the constraints exactly in order to understand the nature of the mathematical problems involved with the construction of a quantum version of Dirac's theory of constraints. As can be seen from this paper, such a program encounters deep pathologies that do not arise when the potentials only provide approximate confinement. Furthermore, if one has a physical principle that requires the particle to be exactly on Γ, or if one is defining quantum mechanics on a curved manifold, it also seems more appropriate to solve the constraint exactly. In addition to the dynamics, we also do the kinematics.

V. CONSTRAINED OBSERVABLES

In this section we continue the analysis of the problem of the previous sections specifically in regard to the observables. That is, given the unbounded map $\kappa: C_\varphi \to L^2(\Gamma)$ above, we wish to examine how self-adjoint operators and unitaries on $L^2(\mathbb{R}^n)$ lift through κ to produce operators on $L^2(\Gamma)$. Due to the unboundedness and nonclosability of κ there will be some pathology, even for bounded operators on $L^2(\mathbb{R}^n)$.

The choice of field algebra $\mathcal{F} \subseteq \mathcal{B}(\mathcal{H})$ will turn out to be important for obtaining a nontrivial constrained field algebra on $L^2(\Gamma)$. In fact, the CCR algebra $\overline{\Delta(\mathbb{R}^{2n})} = C^*\{\exp(i\hat{\mathbf{q}}\cdot\mathbf{a}), \exp(i\hat{\mathbf{p}}\cdot\mathbf{a}) \,|\, \mathbf{a}\in\mathbb{R}^n\}$ will be too small if Γ is a curved manifold (see later in this section). A more suitable field algebra for curved Γ is $C_b(\mathbb{R}^n) \rtimes \mathrm{Diff}\ \mathbb{R}^n$, which is here concretely the C^* algebra generated in $\mathcal{B}(L^2(\mathbb{R}^n))$ by the multiplication operators $\{T_f \,|\, f \in C_b(\mathbb{R}^n)\}$,

$$(T_f\psi)(\mathbf{x}) := f(\mathbf{x})\psi(\mathbf{x}), \quad \psi \in L^2(\mathbb{R}^n) = \mathcal{H},$$

and the set of unitaries $\{V_\beta \,|\, \beta \in \mathrm{Diff}\ \mathbb{R}^n\}$, where

$$(V_\beta\psi)(\mathbf{x}) := J_\beta(\mathbf{x})^{1/2}\psi(\beta\mathbf{x}), \quad \psi \in L^2(\mathbb{R}^n),$$

with J_β the Jacobian of $\beta \in \mathrm{Diff}\ \mathbb{R}^n$. We will concentrate on these two classes of operators below.

Now an operator $A \in \mathcal{B}(\mathcal{H})$ will lift through κ to a densely defined operator $\Lambda(A)$ on $L^2(\Gamma)$ if the following exists:

(i) there is a space $S \subset C_\varphi$ such that $AS \subseteq C_\varphi$ and $\kappa(S)$ is dense in $L^2(\Gamma)$; (ii) $A\psi \in \mathrm{Ker}(\cdot,\cdot)_\Gamma = \mathrm{Ker}\ \kappa$, for all $\psi \in S \cap \mathrm{Ker}\ \kappa$.

In this case,

$$\Lambda(A)\kappa(\psi) := \kappa(A\psi), \quad \forall \psi \in S, \tag{5.1}$$

or equivalently,

$$(\Lambda(A)\kappa(\phi),\ \kappa(\psi))_\Gamma = \lim_{t\to 0}\frac{1}{2t}\,(A\phi,\ \hat{h}P_t\psi), \quad \forall\,\psi,\phi\in S.$$

Since we are interested in $*$ algebras of operators, we will concentrate on situations where the dense subspace S is invariant under the class of operators under consideration. Define the $*$ algebra:

$$\mathscr{T}_S := \{A\in\mathscr{T} \mid A\psi\in S\ni A^*\psi, \quad \forall\,\psi\in S\}. \tag{5.2}$$

There are three obvious dense subspaces S we can ask to be preserved, C_φ, $C_c(\mathbb{R}^n)$, and the transverse space $\mathscr{T}_t=\mathscr{H}_t^\zeta\cap C_\varphi\cap\text{dom } H$ of the last section (or better still, the space $\mathscr{T}^{(t_0)}$ spanned by all $\mathscr{T}_t$ for $t\in(0,t_0]$ and all possible smoothings ζ), and we will consider all these in due course. A useful way of characterizing $\mathscr{T}^{(t_0)}$ is as

$$\mathscr{T}^{(t_0)}=\left\{\psi\in C_c(\mathbb{R}^n)\,\middle|\,\frac{\partial\psi}{\partial\varphi}(\mathbf{x})=0, \quad \forall\,\mathbf{x}\in S_{t_\psi},\ 0<t_\psi<t_0\right\}\cap\text{Dom } H.$$

The choices $\mathscr{T}_t$ or $\mathscr{T}^{(t_0)}$ can be thought of as enforcing secondary quantum constraints on the observables; cf. Remark 2 of Sec. IV. However, we will see that in the quantum picture there is no compelling reason to do this.

Unless we specify what S is below, we will assume that some choice has been made. The elements of $\mathscr{T}_S$ that will lift via (5.1) are those that preserve $S\cap\text{Ker }\kappa$, and for these we have $\Lambda(A)\Lambda(B)\kappa(\psi)=\kappa(AB\psi)=\Lambda(AB)\kappa(\psi)$ for all $\psi\in S$, i.e., Λ is an algebra homomorphism. However, there are two pathologies associated with Λ; (i) given an $A\in\mathscr{B}(\mathscr{H})$ for which $\Lambda(A)$ exists, then $\Lambda(A)$ need not be bounded; (ii) given an $A\in\mathscr{B}(\mathscr{H})$ for which both $\Lambda(A)$ and $\Lambda(A^*)$ exist, we need not have that $\Lambda(A^*)\subseteq\Lambda(A)^*$.

5.3. Example of (i): Let $\mathscr{H}=L^2(\mathbb{R}^2)$, $\varphi(\mathbf{x})=x_2$ so $\Gamma=\varphi^{-1}(0)$ is the x_1 axis. Choose $S=C_c(\mathbb{R}^n)$. Clearly all V_β preserve $S=C_c(\mathbb{R}^2)$ for all $\beta\in\text{Diff }\mathbb{R}^2$, so since $(V_\beta^*\psi)(\mathbf{x})=J_\beta^{-1/2}(\beta^{-1}\mathbf{x})\cdot\psi(\beta^{-1}\mathbf{x})$ we have $V_\beta\in\mathscr{T}_S$. If $\beta\Gamma\subseteq\Gamma$, we get that V_β preserves $C_c(\mathbb{R}^2\backslash\Gamma)=\text{Ker}(\cdot,\cdot)_\Gamma\cap C_c(\mathbb{R}^2)$, and so $\Lambda(V_\beta)$ exists. Consider now the $\beta\in\text{Diff }\mathbb{R}^2$ given by $\beta(x_1,x_2):=(ax_1,x_2e^{x_1})$, $a>0$, so clearly $\beta\Gamma=\Gamma$ and $J_\beta(\mathbf{x})=ae^{x_1}\neq 0$. [That β is a diffeomorphism is clear since its inverse is $\beta^{-1}(x_1,x_2)=(x_1/a,x_2e^{-x_1/a})$, which is also differentiable]. Now, for $\psi\in C_c(\mathbb{R}^2)$, we have

$$\begin{aligned}(\Lambda(V_\beta)\kappa(\psi),\ \Lambda(V_\beta)\kappa(\psi))_{L^2(\Gamma)} &= (V_\beta\psi,\ V_\beta\psi)_\Gamma = (V_\beta\psi\restriction\Gamma,\ V_\beta\psi\restriction\Gamma)_{L^2(\Gamma)}\\ &= \int|(V_\beta\psi\restriction\Gamma)|^2dx_1 = \int J_\beta(x_1,0)|\psi(ax_1,0)|^2dx_1\\ &= \int e^{x/a}|\psi(x,0)|^2dx =: I_\psi,\end{aligned}$$

and now we can choose a $\psi\in C_c(\mathbb{R}^2)$ with $\|\kappa(\psi)\|=\|\psi\restriction\Gamma\|_{L^2(\Gamma)}=1$, which can make I_ψ arbitrarily large, e.g., $\psi=\chi_{[n,n+1]\times[0,1]}$ and consider I_ψ as $n\to\infty$. Thus $\Lambda(V_\beta)$ is unbounded.

We remark that if we chose a more restrictive space S, e.g., $\mathscr{T}^{(t_0)}$, then the diffeomorphism in the last example will preserve $\mathscr{T}^{(t_0)}$, so this pathology cannot be removed by enforcing ''secondary quantum constraints.''

Example of (ii): Continue the previous example, noting that $\Lambda(V_\beta^*)$ exists because $\beta^{-1}\Gamma=\Gamma$. However, if $\Lambda(V_\beta^*)\subseteq\Lambda(V_\beta)^*$, we have, for all $\psi\in C_c(\mathbb{R}^2)$, that

$$\begin{aligned}(\Lambda(V_\beta)\kappa(\psi),\ \Lambda(V_\beta)\kappa(\psi))_{L^2(\Gamma)} &= (\kappa(\psi),\ \Lambda(V_\beta)^*\kappa(V_\beta\psi))_{L^2(\Gamma)}\\ &= (\kappa(\psi),\ \Lambda(V_\beta^*)\kappa(V_\beta\psi))_{L^2(\Gamma)}\\ &= (\kappa(\psi),\ \kappa(V_\beta^*V_\beta\psi))_{L^2(\Gamma)} = (\kappa(\psi),\kappa(\psi))_{L^2(\Gamma)},\end{aligned}$$

which makes $\Lambda(V_\beta)$ unitary, hence bounded, in contradiction with example (i), so it is false that $\Lambda(V_\beta^*)\subseteq\Lambda(V_\beta)^*$.

Now, it appears reasonable to the authors that in a constraining method, boundedness and the adjoint operation should be preserved, at least on the physical variables. So we want to restrict the set of operators under consideration to those satisfying these two requirements. Observe that an $A\in\mathscr{F}_S$ will lift to a bounded operator $\Lambda(A)$ on $L^2(\Gamma)$ iff

$$(A\psi,A\psi)_\Gamma\leqslant M\cdot(\psi,\psi)_\Gamma \quad \forall\psi\in S, \quad \text{and a fixed } M<\infty.$$

[Clearly, if $\psi\in S\cap\mathrm{Ker}\ \kappa$ then the last inequality implies $(A\psi,A\psi)_\Gamma=0$, i.e., $A\psi\in\mathrm{Ker}\ \kappa$, so $\Lambda(A)$ exists.]

Definition:

$$\mathscr{O}_S:=\{A\in\mathscr{F}_S \mid (A\psi,\ A\psi)_\Gamma\leqslant M_A\cdot(\psi,\psi)_\Gamma\geqslant(A^*\psi,\ A^*\psi)_\Gamma,$$

$$\forall\psi\in S, \quad \text{and some } M_A<\infty, \quad \text{and } \Lambda(A^*)\subseteq\Lambda(A)^*\},$$

$$\mathscr{D}_S:=\{A\in\mathscr{O}_S \mid AS\subseteq\mathrm{Ker}\ \kappa\supseteq A^*S\}.$$

Note that if $A\in\mathscr{O}_S$, then $\Lambda(A)$ is bounded, so we have, in fact, that $\Lambda(A)^*=\overline{\Lambda(A^*)}$, so there are no problems with self-adjointness if A is self-adjoint. These will occur, however, if we consider Λ on unbounded operators.

Lemma 5.4: $\mathscr{O}_S$ and $\mathscr{D}_S$ are *-algebras.

Proof: Both sets are linear spaces. To see that $\mathscr{O}_S$ is closed under taking of adjoints, take the adjoint of $\Lambda(A^*)\subseteq\Lambda(A)^*$ to get $\Lambda(A^{**})=\Lambda(A)\subseteq\overline{\Lambda(A)}=\Lambda(A)^{**}\subseteq\Lambda(A^*)^*$, i.e., A^* is in $\mathscr{O}_S$ if A is. Then clearly $\mathscr{D}_S$ is also closed under taking of adjoints. To see that $\mathscr{O}_S$ is an algebra, let $A,\ B\in\mathscr{O}_S$; then

$$(AB\psi,\ AB\psi)_\Gamma\leqslant M_A\cdot(B\psi,B\psi)_\Gamma\leqslant M_AM_B(\psi,\psi)_\Gamma,$$

for all $\psi\in S$, and similarly for B^*A^*. Moreover,

$$\Lambda(AB)^*=[\Lambda(A)\Lambda(B)]^*\supseteq\Lambda(B)^*\Lambda(A)^*\supseteq\Lambda(B^*)\Lambda(A^*)=\Lambda((AB)^*),$$

and thus $AB\in\mathscr{O}_S$. That $\mathscr{D}_S$ is an algebra is obvious. ■

Now $\mathscr{D}_S=\mathrm{Ker}(\Lambda\restriction\mathscr{O}_S)$ is an * ideal of $\mathscr{O}_S$, and $\mathscr{R}_S:=\mathscr{O}_S/\mathscr{D}_S\cong\Lambda(\mathscr{O}_S)\subset\mathscr{B}(L^2(\Gamma))$. We think of $\mathscr{R}_S$ as the ''physical observables'' obtained from enforcing the constraint on $\mathscr{O}_S$, in analogy to the algebra $\mathscr{R}$ of the T procedure for Dirac constraining. $\mathscr{R}_S$ cannot be zero because the identity operator $\mathbb{I}\in\mathscr{O}_S$ and $\overline{\Lambda(\mathbb{I})}=\mathbb{I}$. We do not expect $\mathscr{R}_S$ to be a C^* algebra, but it can easily generate a C^* algebra since $\mathscr{R}_S\cong\Lambda(\mathscr{O}_S)\subset\mathscr{B}(L^2(\Gamma))$.

Given that the commutant $\hat{\varphi}'$ is the traditional observables, we find here the following.

Theorem 5.5: $\mathscr{F}_S\cap\hat{\varphi}'\cap\hat{h}'\subset\mathscr{O}_S$.

Proof: Let $A\in\mathscr{F}_S\cap\hat{\varphi}'\cap\hat{h}'$, which is an * algebra, so it also contains A^*.

Then

$$(A\psi,\ \hat{h}P_tA\psi)=\|\widehat{h^{1/2}}P_tA\psi\|^2=\|A\widehat{h^{1/2}}P_t\psi\|^2\leqslant\|A\|^2\cdot\|\widehat{h^{1/2}}P_t\psi\|^2=\|A\|^2\cdot(\psi,\ \hat{h}P_t\psi),$$

for all $\psi\in S$. Thus

$$(A\psi,\ A\psi)_\Gamma=\lim_{t\to0}\frac{1}{2t}\,(A\psi,\ \hat{h}P_tA\psi)\leqslant\|A\|^2\lim_{t\to0}\frac{1}{2t}\,(\psi,\ \hat{h}P_t\psi)=\|A\|^2\cdot(\psi,\psi)_\Gamma.$$

The same is true for A^*. Now, for all $\psi,\ \phi\in S$,

$$
\begin{aligned}
(\Lambda(A)\kappa(\psi),\ \kappa(\phi))_{L^2(\Gamma)} &= (\kappa(A\psi),\kappa(\phi))_{L^2(\Gamma)} = (A\psi,\phi)_\Gamma \\
&= \lim_{t\to 0}\frac{1}{2t}\,(A\psi,\ \hat{h}P_t\phi) = \lim_{t\to 0}\frac{1}{2t}\,(\psi,\ A^*\hat{h}P_t\phi) \\
&= \lim_{t\to 0}\frac{1}{2t}\,(\psi,\ \hat{h}P_tA^*\phi) = (\psi,A^*\phi)_\Gamma \\
&= (\kappa(\psi),\ \Lambda(A^*)\kappa(\phi))_{L^2(\Gamma)},
\end{aligned}
$$

so $\Lambda(A^*)\subseteq\Lambda(A)^*$. Thus $A\in\mathscr{O}_S$. ■

Remarks: (1) Note that due to the limit in $(\cdot,\cdot)_\Gamma$, it is only the behavior near Γ, which contributes in the constraining of an operator. In fact, the proof actually shows that an $A\in\mathscr{F}_S$ will be in $\mathscr{O}_S$ if

$$
\lim_{t\to 0}\frac{1}{t}\,\|\widehat{h^{1/2}}P_tA\psi\|^2 = \lim_{t\to 0}\frac{1}{t}\,\|A\widehat{h^{1/2}}P_t\psi\|^2
$$

and

$$
\lim_{t\to 0}\frac{1}{t}\,(\psi,[A^*,\hat{h}P_t]\phi) = 0,
$$

for all $\psi,\ \phi\in S$ and the same for A^*. So, if A commutes with $\hat{\varphi}$ and $\hat{h}$ on a small neighborhood of Γ, it will also be in $\mathscr{O}_S$. The behavior of A away from Γ is irrelevant.

(2) We note by Theorem 5.5 that all multiplication operators by bounded Borel functions that preserve S will be in $\mathscr{O}_S$. In particular, for a reasonable physical system we expect $\hat{\varphi}$ to preserve S, so the constraint $\hat{\varphi}$ is in $\mathscr{O}_S$. In this case $\hat{\varphi}$ is in $\mathscr{D}_S$ by the following argument:

$$
\begin{aligned}
\lim_{t\to 0}\|\widehat{h^{1/2}}P_t\hat{\varphi}\psi\|^2 &= \lim_{t\to 0}\frac{1}{t}\,\|\hat{\varphi}\widehat{h^{1/2}}P_t\psi\|^2 \leqslant \lim_{t\to 0}\frac{1}{t}\,\sup\{|\varphi(\mathbf{x})|^2 \mid \mathbf{x}\in\varphi^{-1}[-t,t]\}\cdot\|\widehat{h^{1/2}}P_t\psi\|^2 \\
&= (\lim_{t\to 0}\sup\{|\varphi(x)|^2 \mid \mathbf{x}\in\varphi^{-1}[-t,t]\})\cdot\lim_{t\to 0}\frac{1}{t}\,\|\widehat{h^{1/2}}P_t\psi\|^2 = 0, \quad \forall\,\psi\in S,
\end{aligned}
$$

using the fact that the limit $\lim_{t\to 0}(1/t)\|\widehat{h^{1/2}}P_t\psi\|^2 = \lim_{t\to 0}(1/t)(\psi,\ \hat{h}P_t\psi)$ exists due to $\psi\in S\subset C_\varphi$, and that the limit of the supremum is zero. Hence $(\hat{\varphi}\psi,\hat{\varphi}\psi)_\Gamma = 0$, i.e., $\hat{\varphi}\psi\in\mathrm{Ker}\ \kappa$ for all $\psi\in S$, so $\Lambda(\hat{\varphi}) = 0$, as expected.

Theorem 5.6: *An $A\in\mathscr{F}_S$ preserves* Ker $\kappa\cap S$ *iff*

$$
\lim_{t\to 0}\frac{1}{t}\,\|P_tA(\mathbb{I}-P_t)\psi\|^2 = 0,
$$

for all $\psi\in\mathrm{Ker}\ \kappa\cap S$.

Proof: We need the following lemma:

Lemma: With assumptions and notation as above, we have

$$
\psi\in\mathrm{Ker}(\cdot,\cdot)_\Gamma, \quad \textit{iff}\ \lim_{t\to 0}\frac{1}{t}\,\|P_t\psi\|^2 = 0.
$$

Proof: $\psi\in\mathrm{Ker}(\cdot,\cdot)_\Gamma$, iff $0 = (\psi,\psi)_\Gamma = \lim_{t\to 0}(1/2t)(\psi,\ \hat{h}P_t\psi)$. Now

$$(\psi,\ \hat{h}P_t\psi)=\int_{\varphi^{-1}[-t,t]}|\nabla\varphi(\mathbf{x})|\cdot|\psi(\mathbf{x})|^2d\mu(\mathbf{x})$$

$$\geqslant \inf\{|\nabla\varphi(\mathbf{x})|\mid \mathbf{x}\in\varphi^{-1}[-t,t]\}\int_{\varphi^{-1}[-t,t]}|\psi(\mathbf{x})|^2d\mu(\mathbf{x}).$$

Since $|\nabla\varphi(\mathbf{x})|$ is assumed to be bounded, continuous, and nonzero on a neighborhood of Γ, we have

$$\lim_{t\to 0}\inf\{|\nabla\varphi(\mathbf{x})|\mid \mathbf{x}\in\varphi^{-1}[-t,t]\}=:M>0.$$

Thus

$$\lim_{t\to 0}\frac{1}{t}(\psi,\ \hat{h}P_t\psi)\geqslant M\ \lim_{t\to 0}\frac{1}{t}\int_{\varphi^{-1}[-t,t]}|\psi(\mathbf{x})|^2d\mu(\mathbf{x})=M\ \lim_{t\to 0}\frac{1}{t}\|P_t\psi\|^2,$$

so $\psi\in\mathrm{Ker}(\cdot,\cdot)_\Gamma$ implies

$$0=\lim_{t\to 0}\frac{1}{t}(\psi,\ \hat{h}P_t\psi)\geqslant M\ \lim_{t\to 0}\frac{1}{t}\|P_t\psi\|^2,$$

and so

$$\lim_{t\to 0}\frac{1}{t}\|P_t\psi\|^2=0.$$

On the other hand, given a $\psi\in\mathscr{H}$ satisfying the last equation, we have

$$(\psi,\ \hat{h}P_t\psi)=\int_{\varphi^{-1}[-t,t]}|\nabla\varphi(\mathbf{x})|\cdot|\psi(\mathbf{x})|^2d\mu(\mathbf{x})$$

$$\leqslant\sup\{|\nabla\varphi(\mathbf{x})|\mid \mathbf{x}\in\varphi^{-1}[-t,t]\}\int_{\varphi^{-1}[-t,t]}|\psi(\mathbf{x})|^2d\mu(\mathbf{x}),$$

and the limit of the supremum (denoted N) exists by the boundedness assumption, and is strictly positive. So

$$\lim_{t\to 0}\frac{1}{t}(\psi,\ \hat{h}P_t\psi)\leqslant N\cdot\lim_{t\to 0}\frac{1}{t}\|P_t\psi\|^2=0,$$

i.e., $\lim_{t\to 0}(1/t)(\psi,\ \hat{h}P_t\psi)=0$, i.e., $\psi\in\mathrm{Ker}(\cdot,\cdot)_\Gamma$. ▼

So by this lemma, $A\in\mathscr{T}_S$ preserves $\mathrm{Ker}\,\kappa\cap S$ iff $\lim_{t\to 0}(1/t)\|P_tA\psi\|^2=0$ for all $\psi\in\mathrm{Ker}\,\kappa\cap S$. Now

$$\lim_{t\to 0}\frac{1}{t}\|P_tAP_t\psi\|^2\leqslant\lim_{t\to 0}\frac{1}{t}\|P_tAP_t\|^2\|P_t\psi\|^2\leqslant\|A\|^2\cdot\lim_{t\to 0}\frac{1}{t}\|P_t\psi\|^2=0,$$

whenever $\psi\in\mathrm{Ker}\,\kappa$. So, by the triangle inequality:

$$\|P_tAP_t\psi\|+\|P_tA(\mathbb{I}-P_t)\psi\|\geqslant\|P_tA\psi\|\geqslant|\|P_tAP_t\psi\|-\|P_tA(\mathbb{I}-P_t)\psi\||,$$

we obtain

$$\lim_{t\to 0}\frac{1}{\sqrt{t}}\|P_tA\psi\|=\lim_{t\to 0}\frac{1}{\sqrt{t}}\|P_tA(\mathbb{I}-P_t)\psi\|,$$

for all $\psi \in \mathrm{Ker}\ \kappa \cap S$. ∎

An immediate consequence is that $\Lambda(A)$ exists for all $A \in \hat{\varphi}' \cap \mathscr{T}_S$ [though $\Lambda(A)$ need not be bounded in general], and, as before, we only need commutativity close to Γ to get this. Also note the similarity with the condition $P_{\mathrm{phys}} A(\mathbb{I} - P_{\mathrm{phys}}) = 0$ for an observable in Dirac constraining. Moreover, the proof of Theorem 5.6 used only the existence of the limit (3.6) for ψ, so it will work for larger domains than C_φ.

In the particular cases where S is either C_φ or $C_c(\mathbb{R}^n)$, we shorten the notation to $\mathscr{A}_\varphi := \mathscr{A}_{C_\varphi}$ and $\mathscr{A}_c := \mathscr{A}_{C_c(\mathbb{R}^n)}$, where $\mathscr{A}$ can be $\mathscr{T}$, $\mathscr{O}$, $\mathscr{D}$, or $\mathscr{R}$. Note that

$$C_c(\mathbb{R}^n) \cap \mathrm{Ker}\ \kappa = \{\psi \in C_c(\mathbb{R}^n) \mid \psi \restriction \Gamma = 0\} = C_c(\mathbb{R}^n \backslash \Gamma),$$

$$\mathscr{D}_c = \{A \in \mathscr{O}_c \mid AC_c(\mathbb{R}^n) \subseteq C_c(\mathbb{R}^n \backslash \Gamma) \supseteq A^* C_c(\mathbb{R}^n)\},$$

and for an $A \in \mathscr{O}_c$ we have $\Lambda(A)\kappa(\psi) = \kappa(A\psi) = (A\psi) \restriction \Gamma$ for all $\psi \in C_c(\mathbb{R}^n)$. Analogous statements hold if we replace $C_c(\mathbb{R}^n)$ by C_φ.

On comparing the method above to that of Landsman,[11] we note that the algebra that Landsman selects to impose the constraint on is the subalgebra of the commutant $\hat{\varphi}'$ that preserves $C_c(\mathbb{R}^n)$, i.e., $\hat{\varphi}' \cap \mathscr{T}_c$. In our case the algebra that we constrain, $\mathscr{O}_c$ can be considerably larger than that, with consequently larger algebra of observables $\mathscr{R}_c$, but, on the other hand, there are also nonzero $A \in \hat{\varphi}' \cap \mathscr{T}_c \backslash \mathscr{O}_c$ (see below).

Next, we wish to examine whether particular classes of operators are in $\mathscr{O}_S$.

(i) Consider the multiplication operator $(T_f\psi)(\mathbf{x}) := f(\mathbf{x}) \cdot \psi(\mathbf{x})$, $\forall \psi \in \mathscr{H}$, where f is bounded and Borel. It is not automatic that $T_f \in \mathscr{T}_S$. In fact, if $S = C_c(\mathbb{R}^n)$, then $T_f C_c(\mathbb{R}^n) \subseteq C_c(\mathbb{R}^n)$ iff f is continuous (in which case $T_f \in \mathscr{O}_c$ by Theorem 5.5). For the choice $S = C_\varphi$, we see that if a Borel function f is discontinuous on Γ, then the restriction to Γ may not be defined, i.e., we may have $T_f \notin \mathscr{T}_\varphi$. (This also shows $\hat{\varphi}' \backslash \mathscr{T}_\varphi \neq \emptyset$ hence that $\hat{\varphi}' \not\subset \mathscr{O}_\varphi \cup \mathscr{O}_c$). On the other hand, if f is continuous and bounded on some shell S_t, we have $T_f \in \mathscr{T}_\varphi$. So $\mathscr{T}_\varphi$ contains a larger class of multiplication operators than $\mathscr{T}_c$, and these are all in $\mathscr{O}_\varphi$. In particular, if $\exp i\hat{\mathbf{q}} \cdot \mathbf{a} \in \mathscr{T}$, then $\exp i\hat{\mathbf{q}} \cdot \mathbf{a} \in \mathscr{O}_c \cap \mathscr{O}_\varphi$. However, a T_f is only in $\mathscr{T}_{\mathscr{T}^{(t)}}$ if $f \in \mathscr{T}^{(t)}$.

(ii) Now consider the unitaries V_β, $\beta \in \mathrm{Diff}\ \mathbb{R}^n$. For the choice $S = C_c(\mathbb{R}^n)$, we have that both V_β and V_β^* preserve $C_c(\mathbb{R}^n)$, hence $V_\beta \in \mathscr{T}_c$ for all $\beta \in \mathrm{Diff}\ \mathbb{R}^n$. This is not, however, true for the choice $S = C_\varphi$. In fact, let β be a fixed translation $\beta\mathbf{x} = \mathbf{x} + \mathbf{a}$ not preserving Γ, and let $\psi \in C_\varphi$ be continuous on a neighborhood of Γ but so discontinuous on $\beta\Gamma$ that the restriction to $\beta\Gamma$ is not defined. Then $V_{\beta^{-1}}\psi \notin C_\varphi$, and so $V_{\beta^{-1}} \notin \mathscr{T}_\varphi$. Thus $\mathscr{T}_\varphi$ contains a smaller set of the unitaries V_β than $\mathscr{T}_c$. Below we will examine when these are in $\mathscr{O}_\varphi$.

(iii) Let $S = C_c(\mathbb{R}^n) \ni \psi$, then

$$\Lambda(V_\beta)\kappa(\psi) = \kappa(V_\beta\psi) = (V_\beta\psi) \restriction \Gamma = (J_\beta^{1/2} \restriction \Gamma) \cdot (\psi \circ \beta) \restriction \Gamma.$$

So $\overline{\Lambda(V_\beta)} = \mathbb{I}$ iff $\beta\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in \Gamma$ and $J_\beta \restriction \Gamma = 1$, i.e., $V_\beta - \mathbb{I} \in \mathscr{D}_c$ iff $\beta\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in \Gamma$ and $J_\beta \restriction \Gamma = 1$.

Since $\kappa(\psi) = \psi \restriction \Gamma$ for $\psi \in C_c(\mathbb{R}^n)$, one might surmise that $V_\beta\Gamma \subseteq \Gamma$ iff $V_\beta \in \mathscr{O}_c$. However, the example (5.3) shows that it is not true that $\beta\Gamma \subseteq \Gamma \Rightarrow V_\beta \in \mathscr{O}_c$. For the converse, we have the following.

Lemma 5.7: If $V_\beta \in \mathscr{O}_c \cup \mathscr{O}_\varphi$, then $\beta\Gamma \subseteq \Gamma$.

Proof: Let $V_\beta \in \mathscr{O}_c$, hence V_β preserves $\mathrm{Ker}(\cdot,\cdot)_\Gamma \cap C_c(\mathbb{R}^n) = C_c(\mathbb{R}^n \backslash \Gamma)$.

This is equivalent to $\beta(\mathbb{R}^n \backslash \Gamma) \subseteq \mathbb{R}^n \backslash \Gamma$, which is equivalent to $\beta\Gamma \subseteq \Gamma$. This argument directly adapts to C_φ. ∎

Note that if Γ is curved and nonperiodic, then the translations will not preserve it, so $\exp(i\hat{\mathbf{p}} \cdot \mathbf{a}) \notin \mathscr{O}_c \cup \mathscr{O}_\varphi$, hence of the generating unitaries of $\overline{\Delta(\mathbb{R}^{2n})}$, only the commutative set $\{\exp(i\hat{\mathbf{q}} \cdot \mathbf{a}) \mid \mathbf{a} \in \mathbb{R}^n\}$ is in $\mathscr{O}_c \cup \mathscr{O}_\varphi$. This is why we consider the CCR algebra as too small a choice of field algebra for this type of constraining.

Theorem 5.8: *Given the notation above, we have*

$$V_\beta \in \mathscr{O}_c \cup \mathscr{O}_p \quad \textit{iff} \quad \beta\Gamma \subseteq \Gamma \quad \textit{and} \quad J_\beta(\mathbf{x}) = \mathbb{J}_\beta(\mathbf{x}), \quad \textit{for all} \quad \mathbf{x} \in \Gamma,$$

where J_β (resp., $\mathbb{J}_\beta$) is the Jacobian of β (resp., $\beta \restriction \Gamma$) with respect to μ (resp., γ).

Proof: Let $V_\beta \in \mathscr{O}_c \cup \mathscr{O}_\varphi$, so by Theorem 5.7, $\beta\Gamma \subseteq \Gamma$. Now, since $\Lambda(V_\beta^*) = \Lambda(V_\beta)^*$, we get

$$\Lambda(V_\beta)^*\Lambda(V_\beta)\kappa(\psi)=\Lambda(V_\beta)^*\kappa(V_\beta\psi)=\Lambda(V_\beta^*)\kappa(V_\beta\psi)=\kappa(V_\beta^*V_\beta\psi)=\kappa(\psi)$$
$$=\Lambda(V_\beta)\Lambda(V_\beta)^*\kappa(\psi),\quad \forall\,\psi\in C_c(\mathbb{R}^n)\quad\text{or}\quad C_\varphi.$$

Thus $\overline{\Lambda(V_\beta)}$ is unitary, so

$$(\psi,\psi)_\Gamma=(V_\beta\psi,\ V_\beta\psi)_\Gamma=(V_\beta\psi\restriction\Gamma,\ V_\beta\psi\restriction\Gamma)_{L^2(\Gamma)}=\int_\Gamma|(V_\beta\psi)(\mathbf{x})|^2d\gamma(\mathbf{x})$$
$$=\int_\Gamma J_\beta(\mathbf{x})\cdot|\psi(\beta\mathbf{x})|^2d\gamma(\mathbf{x})=\int_\Gamma J_\beta(\mathbf{x})\cdot\mathbb{J}_\beta^{-1}(\mathbf{x})|\psi(\beta\mathbf{x})|^2d\gamma(\beta\mathbf{x})$$
$$=\int_\Gamma J_\beta(\beta^{-1}\mathbf{x})\cdot\mathbb{J}_\beta^{-1}(\beta^{-1}\mathbf{x})\cdot|\psi(\mathbf{x})|^2d\gamma(\mathbf{x})=\int_\Gamma|\psi(\mathbf{x})|^2d\gamma(\mathbf{x}),$$

$$\forall\,\psi\in C_c(\mathbb{R}^n),\quad\text{or}\quad C_\varphi,$$

iff $J_\beta(\beta^{-1}\mathbf{x})\cdot\mathbb{J}_\beta^{-1}(\beta^{-1}\mathbf{x})=1$ for all $\mathbf{x}\in\Gamma$, i.e., $J_\beta(\mathbf{x})=\mathbb{J}_\beta(\mathbf{x})$ for all $\mathbf{x}\in\Gamma$.

Conversely, observe that $\beta\Gamma\subseteq\Gamma$ guarantees that both $\Lambda(V_\beta)$ and $\underline{\Lambda(V_\beta^*)}$ exist, so by the reversibility of the previous calculation when $J_\beta=\mathbb{J}_\beta$ on Γ, we see that $\overline{\Lambda(V_\beta)}$ is unitary, hence bounded, and similarly the same is true for $\underline{\Lambda(V_\beta^*)}$. Since Λ is a homomorphism and inverses are unique, we get that $\overline{\Lambda(V_\beta^*)}=\Lambda(V_\beta)^*$, i.e., $V_\beta\in\mathscr{O}_c\cup\mathscr{O}_\varphi$. ■

Remarks: (1) Note that if $\beta\in\mathrm{Diff}\,\mathbb{R}^n$ preserves all the level sets of φ we have $\varphi\circ\beta=\varphi$, so $[V_\beta,\hat{\varphi}]=0$. However, since we need not have $\mathbb{J}_\beta=J_\beta$ on Γ, there are certainly such β for which $V_\beta\notin\mathscr{O}_c$. Thus, $\mathscr{F}_c\cap\hat{\varphi}'\setminus\mathscr{O}_c\neq\emptyset$, i.e., Landsman[11] quantizes some operators that we have excluded from our observables. (Some of these can still be taken through Λ using Theorem 5.6, but they need not preserve the adjoint or boundedness). Nevertheless, we conclude that the field algebra $C_b(\mathbb{R}^n)\rtimes\mathrm{Diff}\,\mathbb{R}^n$ has ample elements in its physical algebras $\mathscr{R}_c$ or $\mathscr{R}_\varphi$.

(2) There is no reason, in general, to expect that the time evolutions will preserve $\mathscr{R}_S$, so we will need to extend it by taking the constrained field algebra as the C^* algebra generated in $\mathscr{B}(L^2(\Gamma))$ by $\mathscr{R}_S$ and the unitaries $\exp itH_\Gamma$, $t\in\mathbb{R}$.

(3) Recall that we consider the enforcement of secondary quantum constraints as the selection of the set $S=\mathscr{T}^{(t_0)}$ that the observables should preserve. Since this is an analogy to the classical procedure, it is natural to look for reasons to justify such a choice, and a first attempt may be to ask whether the choice of transverse states for S will make the Jacobian condition in Theorem 5.8 obsolete, i.e., whether for a $V_\beta\in\mathscr{F}_{\mathscr{T}^{(t)}}$ we have $V_\beta\in\mathscr{O}_{\mathscr{T}^{(t)}}$ iff $\beta\Gamma\subset\Gamma$. This is not true, as we can see by an easy counterexample. Continue the example (5.3) with the additional assumption that the Hamiltonian has $\mathrm{Dom}\,H=C_c^\infty(\mathbb{R}^2)$ and that $\beta\in\mathrm{Diff}\,\mathbb{R}^2$ is not the given one in (5.3), but $\beta(x_1,x_2)=(x_1,2x_2)$. Then both V_β and V_β^* preserve

$$\mathscr{T}^{(t_0)}=\left\{\psi\in C_c^\infty(\mathbb{R}^n)\;\middle|\;\frac{\partial\psi}{\partial x_2}(\mathbf{x})=0,\quad\forall\,\mathbf{x}\in S_{t_\psi},\ \ t_\psi<t_0\right\};$$

hence $V_\beta\in\mathscr{F}_{\mathscr{T}^{(t_0)}}$, and $\beta\Gamma=\Gamma$, $\beta\restriction\Gamma=\mathbb{I}$, $\mathbb{J}_\beta=1$, $J_\beta=2$; hence

$$\Lambda(V_\beta)^*=\sqrt{2}\mathbb{I}\neq\Lambda(V_\beta^*)=\frac{1}{\sqrt{2}}\,\mathbb{I},$$

so $V_\beta\notin\mathscr{O}_{\mathscr{T}^{(t_0)}}$.

Another possible reason one may want to use to justify the enforcement of secondary quantum constraints, is to produce a common dense invariant domain $\kappa(S)$ for the observables $\mathscr{R}_S$. However, for the choice $S=\mathscr{T}^{(t)}$ there is no reason why the Hamiltonian H_Γ should preserve $\kappa(\mathscr{T}^{(t)})$, so it will be unreasonable to require the observables to do so. So at this stage we fail to see why secondary quantum constraints should be imposed on the observables.

(4) It is interesting to observe that for any Hilbert space operator $K:\mathscr{H}_1\to\mathscr{H}_2$ with dense range, we can obtain the structure above. That is, given a dense space $S\subseteq\mathrm{dom}\,K$ for which $K(S)$

is dense in $\mathscr{H}_2$, we can define $\mathscr{T}_S$, $\mathscr{O}_S$, $\mathscr{D}_S$, $\mathscr{R}_S$ exactly as before by replacing $\mathscr{T}$ with $\mathscr{B}(\mathscr{H}_1)$, and κ by K. So every such operator K and space S defines a lifting problem, hence, a short exact sequence of $*$ algebras,

$$0\to\mathscr{D}_S\to\mathscr{O}_S\to\mathscr{R}_S\to 0.$$

This then produces a similar short exact sequence for the automorphism groups and their lifting through the factorization. In particular, for ordinary Dirac constraining, the operator K is the projection $P_{\text{phys}}:\mathscr{H}\to\mathscr{H}_{\text{phys}}$, the space S is $\mathscr{H}$, and $\mathscr{A}_S\cap\mathscr{T}_S$ is just $\mathscr{A}$, where $\mathscr{A}$ denotes either $\mathscr{O}$ or $\mathscr{D}$. This way of thinking nicely unifies the current construction (where $K=\kappa$) with that of the Dirac approach. The main difference is that P_{phys} is bounded while κ is unbounded and nonclosable.

VI. CONSTRAINING BY A GENERAL SELF-ADJOINT OPERATOR

For later reference, we start by summarizing the constraining algorithm developed in the preceding sections. Assume the following data: operators $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, and H on $C_c^\infty(\mathbb{R}^n)\subset L^2(\mathbb{R}^n)$, a bounded $\varphi\in C^\infty(\mathbb{R}^n)$ such that $\Gamma=\varphi^{-1}(0)$ is a smooth $(n-1)$-dimensional submanifold, and, moreover, $\nabla\varphi$ is bounded and nonzero on a neighborhood of Γ and a unital field algebra $\mathscr{T}\subset\mathscr{B}(L^2(\mathbb{R}^n))$ containing $\hat{\varphi}$, $\exp i\hat{\mathbf{p}}\cdot\mathbf{a}$, $\exp i\hat{\mathbf{q}}\cdot\mathbf{a}$.

(1) Define $(\psi_1,\psi_2)_\Gamma:=\lim_{t\to 0}(1/2t)(\psi_1,\hat{h}P_t\psi_2)$, $\forall\psi_i\in C_\varphi(\mathbb{R}^n)$, $L^2(\Gamma)=[C_\varphi(\mathbb{R}^n)/\mathrm{Ker}(\cdot,\cdot)_\Gamma]^{-(\cdot,\cdot)_\Gamma}$, $\kappa:C_\varphi\to L^2(\Gamma)$ is the factorization map, and $L^2(\Gamma)$ is the constrained space.

(2) Let $\mathscr{T}_{t_0}:=\mathscr{H}_{t_0}^{\zeta}\cap C_\varphi\cap\mathrm{Dom}\,H$, check that $H\mathscr{T}_{t_0}\subset C_\varphi$ for t_0 small enough, and if so, define the constrained Hamiltonian H_Γ by $H_\Gamma\cdot\kappa(\psi):=\kappa(H\psi)$ for all $\psi\in\mathscr{T}_{t_0}$.

(3) Choose a space S in $\mathscr{H}$ for which $\kappa(S)$ is dense (e.g., C_φ), and define

$$\mathscr{T}_S=\{A\in\mathscr{T}\mid AS\subseteq S\supseteq A^*S\},$$

$$\mathscr{O}_S=\{A\in\mathscr{T}_S\mid(A\psi,A\psi)_\Gamma\leqslant M_A(\psi,\psi)_\Gamma\geqslant(A^*\psi,A^*\psi)_\Gamma,\quad\Lambda(A^*)\subseteq\Lambda(A)^*\},$$

where $\Lambda(A)$ denotes the closure of the operator defined by

$$\Lambda(A)\kappa(\psi)=\kappa(A\psi),\quad\forall\psi\in S.$$

Define $\mathscr{D}_S:=\{A\in\mathscr{O}_S\mid AS\subseteq\mathrm{Ker}\,\kappa\supseteq A^*S\}$; then $\Lambda(\mathscr{O}_S)\cong\mathscr{O}_S/\mathscr{D}_S=:\mathscr{R}_S$, and the algebra of the constrained observables on $L^2(\Gamma)$ is the C^* algebra generated by $\mathscr{R}_S$ and $\exp(i\mathbb{R}H_\Gamma)$.

In this section we want to generalize the method above to impose a constraint ''$C\psi=0$,'' where C is a general bounded self-adjoint operator on a Hilbert space $\mathscr{H}$ with zero in its continuous spectrum. [If we start with an unbounded self-adjoint operator C, we can, without loss of generality, convert the problem to a bounded one by replacing C with $f(C)$, where f is a continuous bounded real-valued function with $f(x)=x$ on a neighborhood of zero.] As before, we still assume that there is a unital field algebra $\mathscr{T}$ acting on $\mathscr{H}$, containing C, and that there is also a (possibly unbounded) Hamiltonian H given on $\mathscr{H}$, $\exp(itH)\in\mathscr{T}$ for all t. We will be concerned with the construction of three objects: the constrained Hilbert space (and the constraining map to it from the original space), the constrained Hamiltonian, and the algebra of constrained observables.

At the abstract C^* level, the kinematics part has already been solved,[9] but one obtained the set of all representations in which the constraint can be imposed as an eigenvalue condition. Here we want to build a particular concrete constrained system out of the given unconstrained one.

Now geometry was paramount in the analysis of the previous sections, in that we needed a metric (on an underlying space) to define gradients, norms of vectors, normals to surfaces and the Lebesgue measure. This vital piece of information is missing in the problem under consideration, and we somehow need to augment the given data $\{\mathscr{T},\mathscr{H},H,C\}$ in order to adapt the method previously found to this problem. The extra information we will assume is the following.

(i) A maximally commutative C^* algebra $\mathscr{A}\subset\mathscr{B}(\mathscr{H})$ containing C.

[Call $\mathscr{A}$ a *polarization*, and we know that it always has a cyclic and separating vector (Ref. 28, Proposition 2.5.3.]

(ii) A choice of a cyclic and separating vector Ω for $\mathscr{A}$. (Call Ω the vacuum.)

(iii) A choice of scaling operator $K\in\mathscr{A}_+$, Ker $K=\{0\}$.

(iv) A self-adjoint operator $\mathscr{P}_C$ such that $[\mathscr{P}_C,C]\psi=i\psi$ for all $\psi\in\mathscr{D}\equiv$ a dense invariant subspace of $P_{t_0}\mathscr{H}$, where P_t denotes the spectral projection of C of the interval $[-t,t]$. This is thought of as a ''local'' canonical momentum for C, which will define the normal direction to the constrained system.

Now define the space $\mathscr{L}:=\{A\in\mathscr{A}\mid \omega_0(A) \text{ exist}\}$, where

$$\omega_0(A):=\lim_{t\to 0}\frac{(P_t\Omega,\ AP_t\Omega)}{\|P_t\Omega\|^2}.$$

Since obviously $\mathbb{I}$ and $C\in\mathscr{L}$, this space is not zero. Moreover, the space is self-adjoint, and we also have that $|\omega_0(A)|\leqslant\|A\|$ for all $A\in\mathscr{L}$ because $|(P_t\Omega,\ AP_t\Omega)/\|P_t\Omega\|^2|\leqslant\|A\|$, and so ω_0 extends to the closure of $\mathscr{L}$. Now since ω_0 is obviously positive on the positive elements of $\mathscr{L}$, we can apply Ref. 29, Theorem 4.3.13 to conclude that ω_0 extends to a state on $\mathscr{A}$. Henceforth we fix a choice of extension and still denote it by ω_0.

Denote the GNS representation of ω_0 by $(\pi_0,\Omega_0,\mathscr{H}_0)$. Noting that $\omega_0(K)>0$, we define a state ω_K on $\mathscr{A}$ by

$$\omega_K(A):=\omega_0(KA)/\omega_0(K),\quad A\in\mathscr{A}.$$

Then, inspired by the following lemma and subsequent example, we identify the constrained Hilbert space with $\mathscr{H}_{\omega_K}$.

Lemma 6.1: ω_K is a ''Dirac state'' in the sense that $\omega_K(AC)=0$ for all $A\in\mathscr{A}$, i.e., $\pi_{\omega_K}(C)\Omega_{\omega_K}=0$.

Proof: It suffices to show that $\omega_K(C^2)=0$, since C is self-adjoint. So

$$\frac{\omega_0(KC^2)}{\omega_0(K)}=\lim_{t\to 0}\frac{(P_t\Omega,\ KC^2P_t\Omega)}{\|P_t\Omega\|^2\omega_0(K)}.$$

Now using the spectral decomposition $C=\int\lambda\,dP(\lambda)$, we have

$$\begin{aligned}|(KP_t\Omega,\ P_tC^2\Omega)|&\leqslant\|KC\|\cdot\|P_t\Omega\|\cdot\|P_tC\Omega\|\\&=\|KC\|\cdot\|P_t\Omega\|\cdot\left\|\int_{-t}^{t}\lambda\,dP(\lambda)\Omega\right\|\\&\leqslant\|KC\|\cdot\|P_t\Omega\|\cdot t\left\|\int_{-t}^{t}dP(\lambda)\Omega\right\|=t\|KC\|\cdot\|P_t\Omega\|^2.\end{aligned}$$

Hence $|\omega_K(C^2)|\leqslant[1/\omega_0(K)]\lim_{t\to 0}\ t\|KC\|=0$. ■

Remark: The proof above is easily adapted to show that ω_0 is also a Dirac state on $\mathscr{A}$. More precisely, we have for the left kernels that $N_{\omega_0}=N_{\omega_K}$ due to Ker $K=\{0\}$ and the fact that $\mathscr{A}$ is commutative. Hence we have an identification map $\theta:\mathscr{A}/N_{\omega_0}\to\mathscr{A}/N_{\omega_K}$ by $\theta(A+N_{\omega_0})=A+N_{\omega_K}$ and θ extends to a map $\theta:\mathscr{H}_{\omega_0}\to\mathscr{H}_{\omega_K}$ because

$$\begin{aligned}\|\theta(A+N_{\omega_0})\|_{\mathscr{H}_{\omega_K}}&=\omega_K(\theta(A+N_{\omega_0})^*\theta(A+N_{\omega_0}))^{1/2}\\&=\omega_K(A^*A)^{1/2}=\omega_0(KA^*A)^{1/2}\\&\leqslant\|K\|\cdot\omega_0(A^*A)^{1/2}=\|K\|\cdot\|A+N_{\omega_0}\|_{\mathscr{H}_0}.\end{aligned}$$

To define a constraining map $\kappa:\mathscr{H}\to\mathscr{H}_{\omega_K}$ from the unconstrained to the constrained space, recall that Ω separates $\mathscr{A}$; hence the map

$$\kappa(A\Omega):=\pi_{\omega_K}(A)\Omega_{\omega_K},\quad A\in\mathscr{A},$$

is well defined on the dense subspace $\mathscr{A}\Omega$ with a dense range. This will be our choice of constraining map for this context. Since $\|\kappa(A\Omega)\|^2=\omega_K(A^*A)$, we see that Ker $\kappa=N_{\omega_K}\Omega=N_{\omega_0}\Omega$.

Example: To motivate the preceding structures, and to see what is involved in the choices of $\mathscr{A}$, Ω, K, we now reconsider the constraint situation $\hat{\varphi}$, $L^2(\mathbb{R}^n)$ of the preceding sections.

Starting with $\hat{\varphi}=\varphi(\hat{\mathbf{q}})$, the natural choice of polarization is $\mathscr{A}:=\{f(\hat{\mathbf{q}})|f\in L^\infty(\mathbb{R}^n)\}=\{\exp(i\hat{\mathbf{q}}\cdot\mathbf{a})|\mathbf{a}\in\mathbb{R}^n\}''$, and for Ω we can then choose any positive nowhere vanishing L^2 function. Choose the Gaussian $\Omega(\mathbf{x})=\exp(-a|\mathbf{x}|^2)$, $a>0$ and note that $C_c(\mathbb{R}^n)=C_c(\mathbb{R}^n)\Omega$, hence $C_c(\mathbb{R}^n)\subset\mathscr{A}\Omega$. Now for K we must have $K\in\mathscr{A}_+$, so $K=k(\hat{\mathbf{q}})$ for some $k\in L^\infty_+(\mathbb{R}^n)$, $k^{-1}(0)=\varnothing$. For the moment we will choose k to be continuous, and then below deduce the precise choice that will correspond with the previous results for this situation. In particular, what we want to show is that

$$(\kappa(A\Omega),\kappa(B\Omega))_{\mathscr{H}_{\omega_K}}=b((A\Omega\restriction\Gamma),(B\Omega\restriction\Gamma))_{L^2(\Gamma)}, \qquad (*)$$

for all A, $B\in C_c(\mathbb{R}^n)$ for the right choice of K, where κ is the map defined in this section and b is a normalizing constant. Now

$$(\kappa(A\Omega),\kappa(B\Omega))_{\mathscr{H}_{\omega_K}}=\omega_K(A^*B)=\lim_{t\to 0}\frac{(A\Omega,KP_tB\Omega)}{\|P_t\Omega\|^2\omega_0(K)},$$

whenever $A^*B\in\mathscr{L}$ [which we will see below is all $C_c(\mathbb{R}^n)$], so if we write $A=f_A(\hat{\mathbf{q}})$, $B=f_B(\hat{\mathbf{q}})$ with f_A, $f_B\in C_c(\mathbb{R}^n)$ and remember that

$$(P_t\psi)(\mathbf{x})=\chi_{\varphi^{-1}[-t,t]}(\mathbf{x})\cdot\psi(\mathbf{x}),$$

we have

$$\frac{(A\Omega,\ KP_tB\Omega)}{\|P_t\Omega\|^2}=\frac{\int_{\varphi^{-1}[-t,t]}\overline{f_A}(\mathbf{x})k(\mathbf{x})f_B(\mathbf{x})\ \exp(-2a|\mathbf{x}|^2)d\mu(\mathbf{x})}{\int_{\varphi^{-1}[-t,t]}\exp(-2a|\mathbf{x}|^2)d\mu(\mathbf{x})}.$$

Now, recall that for $f\in C_c(\mathbb{R}^n)$ and small t, we have for a patch $U\cap\varphi^{-1}[-t,t]$ with local coordinates

$$\int_{U\cap\varphi^{-1}[-t,t]}f(\mathbf{x})d\mu(\mathbf{x})=\int_{-t}^{t}\left(\int\frac{f(\mathbf{y},\varphi)}{|\boldsymbol{\nabla}\varphi(\mathbf{y},\varphi)|}\,d\gamma_\varphi(\mathbf{y})\right)d\varphi$$
$$=2t\int_{U\cap\Gamma}\frac{f(\mathbf{y},0)}{|\nabla(\mathbf{y},0)|}\,d\gamma(\mathbf{y})+O(t)=2t\int_{U\cap\Gamma}(f/|\boldsymbol{\nabla}\varphi|)\restriction\Gamma\ d\gamma+O(t).$$

Thus

$$(\kappa(A\Omega),\kappa(B\Omega))_{\mathscr{H}_{\omega_K}}=\lim_{t\to 0}\frac{(A\Omega,\ KP_tB\Omega)}{\|P_t\Omega\|^2\omega_0(K)}=\frac{\int_\Gamma(\overline{f_A}kf_Be/|\boldsymbol{\nabla}\varphi|)\restriction\Gamma\ d\gamma}{\omega_0(K)\int_\Gamma(e/|\boldsymbol{\nabla}\varphi|)\restriction\Gamma\ d\gamma},$$

where e denotes the function $\exp(-2a|\mathbf{x}|^2)$. Now, since by a similar argument

$$\omega_0(K)=\lim_{t\to 0}\frac{(P_t\Omega,\ KP_t\Omega)}{\|P_t\Omega\|^2}=\frac{\int_\Gamma(ke/|\boldsymbol{\nabla}\varphi|)\restriction\Gamma\ d\gamma}{\int_\Gamma(e/|\boldsymbol{\nabla}\varphi|)\restriction\Gamma\ d\gamma}$$

[the existence of this limit shows $\mathscr{L}\supseteq C_c(\mathbb{R}^n)$], we have

$$(\kappa(A\Omega),\kappa(B\Omega))_{\mathscr{H}_{\omega_K}}=\frac{\int_\Gamma(\overline{f_A}kf_Be/|\nabla\varphi|)\restriction\Gamma\ d\gamma}{\int_\Gamma(ke/|\nabla\varphi|)\restriction\Gamma\ d\gamma}$$

$$=\frac{\int_\Gamma(\overline{f_A}f_Be)\restriction\Gamma\ d\gamma}{\int_\Gamma e\restriction\Gamma\ d\gamma}=\frac{((A\Omega\restriction\Gamma),(B\Omega\restriction\Gamma))_{L^2(\Gamma)}}{\|\Omega\restriction\Gamma\|^2_{L^2(\Gamma)}}$$

(where we made the choice $k=|\nabla\varphi|$), which is just the normalized inner product of $L^2(\Gamma)$, and this will produce the same κ that we had before.

If there is also a group of symmetries that needs to be unitarily implemented on the constrained Hilbert space $\mathscr{H}_{\omega_K}$, we can use this unitarity to determine K.

So now that we have the constraining map $\kappa:\mathscr{H}\to\mathscr{H}_{\omega_K}$ as above, we can proceed to constrain the dynamics. This is where we will use the assumed operator $\mathscr{P}_C$ in the role of the ''normal derivative $i(\partial/\partial\varphi)$ near Γ.'' Thus, start by selecting a space of transverse states by

$$\mathscr{H}^T_t:=\overline{\{P_t\psi\mid\psi\in\mathrm{Dom}\ \mathscr{P}_C,\ P_t\mathscr{P}_C\psi=0\}}$$

and

$$\mathscr{T}_t:=\{\psi\in\mathrm{Dom}\ H\mid P_t\psi\in\mathscr{H}^T_t\},$$

and note that $\mathscr{T}_t\subset\mathscr{T}_s$ if $t>s$. For a consistent constraining we need to assume there is a $t_0>0$ such that the set $\mathscr{S}_{t_0}:=\{\psi\in\mathscr{T}_{t_0}\mid H\psi\in\mathscr{A}\Omega\}$ satisfies $\mathrm{Ker}\ \kappa\cap\mathscr{S}_{t_0}=\{0\}$ and $\kappa(\mathscr{S}_{t_0})$ is dense in $\mathscr{H}_{\omega_K}$. Given this, we define the constrained Hamiltonian H_C on domain $\kappa(\mathscr{S}_{t_0})$ to be

$$H_C\cdot\kappa(\psi):=\kappa(H\psi),\quad\forall\psi\in\mathscr{S}_{t_0}.$$

Example: Continue the previous example with $C=\hat{\varphi}$ and the choice $K=|\nabla\varphi|$. Choose

$$\mathscr{P}_{\hat{\varphi}}\psi:=if(\varphi)\frac{\partial}{\partial\varphi}\psi,\quad\forall\psi\in C^\infty_c(\mathbb{R}^n),$$

where f is a smooth bump function that is unity on $[-t_0,t_0]$ and zero outside $[-t_0-\epsilon,\ t_0+\epsilon]$ for some $\epsilon>0$. Then

$$\mathscr{H}^T_{t_0}=\{\psi\in\mathscr{H}\mid\mathrm{supp}\ \psi\subset\varphi^{-1}([-t_0,t_0]),\quad\psi(\mathbf{y},\varphi)\ \text{independent of}\ \varphi\ \text{on}\ \varphi^{-1}[-\mathrm{t}_0,\mathrm{t}_0]\}.$$

If H is a differential operator with domain $C^\infty_c(\mathbb{R}^n)$, we have

$$\mathscr{T}_{t_0}=\{\psi\in C^\infty_c(\mathbb{R}^n)\mid\psi\restriction\varphi^{-1}((-t_0,t_0))\ \text{is constant in the direction}\ \nabla\varphi\}.$$

Since $\mathscr{H}^\zeta_{t_0}\cap C^\infty_c(\mathbb{R}^n)\subset\mathscr{T}_{t_0}$, we conclude that

$$H_{\hat{\varphi}}\kappa(\psi)=\kappa(H\psi),\quad\forall\psi\in\mathscr{T}_{t_0}$$

defines the same operator as before in Sec. IV.

Finally, to obtain the constrained kinematics as in Sec. V, let $S\subseteq\mathscr{A}\Omega$ be a subspace such that $\kappa(S)$ is dense in $\mathscr{H}_{\omega_K}$. Define, in complete analogy with Sec. V,

$$\mathscr{T}_S:=\{A\in\mathscr{T}\mid A\psi\in S\ni A^*\psi,\quad\forall\psi\in S\}$$

$$\mathscr{O}_S:=\{A\in\mathscr{T}_S\mid(A\psi,A\psi)_K\leqslant M_A\cdot(\psi,\psi)_K\geqslant(A^*\psi,A^*\psi)_K,\quad\forall\psi\in S,$$

$$M_A<\infty\quad\text{and}\ \Lambda(A^*)\subseteq\Lambda(A)^*\},$$

$$\mathscr{D}_S:=\{A\in\mathscr{O}_S\mid AS\subseteq\mathrm{Ker}\ \kappa\supseteq A^*S\},$$

where $\Lambda(A)\kappa(\psi):=\kappa(A\psi),\ \forall\psi\in S,\ A\in\mathscr{O}_S$.

The previous proofs carry over verbatim, so $\mathscr{D}_S = \mathrm{Ker}(\Lambda \restriction \mathscr{O}_S)$ and the algebra of observables acting on $\mathscr{H}_{\omega_K}$ is the C^* algebra generated by $\exp i\mathbb{R}H_C$ and

$$\mathscr{R}_S := \mathscr{O}_S / \mathscr{D}_S \cong \Lambda(\mathscr{O}_S) \subset \mathscr{B}(\mathscr{H}_{\omega_K}).$$

Example: In a continuation of the last example, it is obvious that if we take $\mathscr{F} = C_b(\mathbb{R}^n) \rtimes \mathrm{Diff}\ \mathbb{R}^n$ and $S = C_c(\mathbb{R}^n)$, then we will get the same observable algebra as before.

Remark: Another way to proceed could have been to extend the Dirac state ω_K on $\mathscr{A}$ to $\mathscr{F}$. However, this would have enlarged the GNS space (producing a wrong result for the example of this paper), and, in addition, there is the problem that the extension need not be unique.

VII. CONCLUSIONS AND DISCUSSION

To summarize, we took the problem of restricting a quantum particle in $\mathbb{R}^n$ to a lower-dimensional submanifold not conserved by the dynamics. We found a method to do the restriction, and this came in three parts—a map κ from a dense subspace of $L^2(\mathbb{R}^n)$ to the constrained Hilbert space $L^2(\Gamma)$, a method for constructing the constrained Hamiltonian from the unconstrained one, and a method for the construction of the algebra of observables $\mathscr{R}_\Gamma$ on $L^2(\Gamma)$. For each of these three parts there were choices to be made. For the first part we had to choose a domain for κ [$C_c(\mathbb{R}^n)$ and C_φ seem the canonical choices], for the second part the transverse space [here tangentiality of momentum to the level sets of φ near Γ seems to be the physically justified criterion], and for the third part the space S on which to reduce the observables [likely choices here are $C_c(\mathbb{R}^n)$, C_φ, or the transverse space used]. We regard these choices to be made as physical ones, and in the two examples, the choices made were evident, and produced results in agreement with the known solutions. The necessity of these choices should be compared with the choice of boundary conditions required in the constraining of a free quantum particle to a box.

There are mathematical pathologies associated with our proposed method, of which the central one is that the map κ is neither bounded nor even closable. As a consequence, the constraining of observables need not preserve boundedness or involution, so we had to restrict the method to those observables for which these pathologies do not occur. The method is also sensitive with respect to the original choice of field algebra, and may not provide enough observables for some choices.

While this method was developed in analogy with the classical constraint method, the analogy became quite vague at certain points, for instance, one can interpret the necessity for a choice of transverse space as an imposition of a secondary constraint, but this constraint could not be interpreted as obtained from a similar method as the classical one. Moreover, secondary constraints do not seem to be necessary for the constraining of the observables.

We expect from our work in Ref. 9 that the pathologies that occurred here can be circumvented by a suitable generalization to a C^*-algebra framework, insofar as the T procedure solves already without pathology the kinematics part of the question there; but this would involve losing the original representation, which may contain some physical information. The advantage of the current method is that it constructs the constrained system concretely from the given representation of the original system. Nevertheless, we intend to further pursue this line of thought.

ACKNOWLEDGMENTS

One of us (H.G.) would like to thank Alan Weinstein for his question during a seminar, which lead to this paper. H.G. would also like to thank and acknowledge the support of the Erwin Schrödinger Institute of Mathematical Physics during the last phase of this project. We are grateful to Klaas Landsman for drawing our attention to the work of Ref. 15.

[1] P. A. M. Dirac, ''Lectures on quantum mechanics,'' Belfer Graduate School of Science, Yeshiva University, New York, 1964.
[2] M. J. Gotay, J. M. Nester, and G. Hinds, ''Presymplectic manifolds and the Dirac–Bergman theory of constraints,'' J. Math. Phys. **19**, 2388–2399 (1978).
[3] J. E. Marsden and A. Weinstein, ''Reduction of symplectic manifolds with symmetry,'' Rep. Math. Phys. **5**, 121–130 (1974).
[4] L. Faddeev and R. Jackiw, ''Hamiltonian reduction of unconstrained and constrained systems,'' Phys. Rev. Lett. **60**, 1692 (1988).

[5] L. Van Hove, ''Sur certaines représentations unitaires d'un groupe infini de transformations,'' Proc. R. Acad. Sci. Belgium **26**, 1–102 (1951).
[6] N. M. J. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1992).
[7] M. A. Rieffel, ''Deformation quantization of Heisenberg manifolds,'' Commun. Math. Phys. **122**, 531–562 (1989).
[8] M. J. Gotay, H. B. Grundling, and G. M. Tuynman, ''Obstruction results in quantization theory,'' J. Nonlinear Sci. **6**, 469–498 (1996).
[9] H. Grundling and C. A. Hurst, ''Algebraic quantization of systems with a gauge degeneracy,'' Commun. Math. Phys. **98**, 369–390 (1985); H. Grundling and C. A. Hurst, ''A note on regular states and supplementary conditions,'' Lett. Math. Phys. **15**, 205–212 (1988); and errata, *ibid.* **17**, 173–174 (1989).
[10] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992).
[11] N. P. Landsman, ''Rieffel induction as generalised quantum Marsden–Weinstein reduction,'' J. Geom. Phys. **15**, 285–319 (1995).
[12] F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974); **17**, 1930 (1976).
[13] J. Klauder, Ann. Phys. (N.Y.) **254**, 419–453 (1997).
[14] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, 1989).
[15] P. Maraner, ''Monopole gauge fields and quantum potentials induced by the geometry in simple dynamical systems,'' Ann. Phys. (N.Y.) **246**, 325–346 (1996).
[16] K. Fujii, N. Ogawa, S. Uchiyama, and N. Chepilko, ''Geometrically induced gauge structure on manifolds embedded in a higher dimensional space,'' preprint hep-th/9702191.
[17] H. Grundling, ''BRST quantum theory: A functional analytic approach,'' UNSW preprint, 1991.
[18] N. P. Landsman and N. Linden, ''Superselection rules from Dirac and BRST quantization of systems with constraints,'' Nucl. Phys. B **371**, 415–433 (1992).
[19] D. McMullan, ''Classical states and the BRST charge,'' Commun. Math. Phys. **149**, 161–174 (1992).
[20] K. Sundermeyer, ''Constrained dynamics,'' *Lecture Notes in Physics* (Springer-Verlag, New York, 1982), Vol. 169.
[21] E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974).
[22] Ph. Tondeur, *Foliations on Riemannian Manifolds* (Springer-Verlag, New York, 1988).
[23] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics* (Dover, New York, 1949).
[24] I. E. Farquhar, *Ergodic Theory in Statistical Mechanics* (Interscience, New York, 1964).
[25] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin, London, 1982).
[26] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I* (Academic, New York, 1972).
[27] S. Helgason, *Groups and Geometric Analysis, Integral Geometry, Invariant Differential Operators, and Spherical Functions* (Academic, New York, 1984).
[28] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I* (Springer-Verlag, New York, 1987).
[29] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras I* (Academic, New York, 1983).

de Rham cohomology of SO(n) and some related manifolds by supersymmetric quantum mechanics

Kazuto Oshima
Gunma College of Technology, Maebashi 371-8530, Japan

(Received 11 February 1998; accepted for publication 17 March 1998)

We study supersymmetric quantum mechanics on RP_n, $\mathrm{SO}(n)$, G_2, and $\mathrm{U}(2)$ to examine Witten's Morse theory concretely. We confirm the simple instanton picture of the de Rham cohomology that has been given in a previous paper. We use a reflection symmetry of each theory to select the true vacuums. The number of selected vacuums agrees with the de Rham cohomology for each of the above manifolds. © *1998 American Institute of Physics.* [S0022-2488(98)03206-X]

I. INTRODUCTION

In a previous paper,[1] the author has investigated Witten's Morse theory[2] for $\mathrm{SO}(n)$.[3] Instanton effects between adjacent classical vacuums have been studied. Using reflection symmetries of the theory, a selection rule for true vacuums has been given. This selection rule works well, at least for $n \leqslant 5$;[1] the number of selected vacuums agrees with the de Rham cohomology of $\mathrm{SO}(n)$. The main purpose of this paper is to show that the selection rule works well for arbitrary n. We also apply the selection rule to some related manifolds RP_n, G_2, and $\mathrm{U}(2)$; $\mathrm{RP}_1 \simeq \mathrm{SO}(2)$, $\mathrm{RP}_3 \simeq \mathrm{SO}(3)$, $\mathrm{G}_2 \subset \mathrm{O}(7)$, and $\mathrm{U}(2) \simeq \mathrm{SO}(2) \times \mathrm{SU}(2)$.

This paper is organized as follows: In Sec. II we review supersymmetric quantum mechanics on a manifold M. The selection rule for true vacuums is stated in this section. The manifold RP_n are studied in the third section. The manifold RP_n is simple, and this section will be helpful to understand the following sections. In Sec. IV the de Rham cohomology of $\mathrm{SO}(n)$ is derived for arbitrary n by the selection rule. The manifolds G_2 and $\mathrm{U}(2)$ are discussed in Secs. V and VI, respectively. The results are summarized in the last section.

II. SUPERSYMMETRIC QUANTUM MECHANICS ON A MANIFOLD

The supersymmetric Hamiltonian on a manifold with Morse function h is given by

$$\hat{H} = \tfrac{1}{2}(d_h d_h^\dagger + d_h^\dagger d_h), \tag{1}$$

where $d_h = e^{-h} d e^h$, $d_h^\dagger = e^h d^\dagger e^{-h}$, d is the exterior derivative, and $d^\dagger$ is its adjoint operator. In (any) coordinates $\{x^\mu\}$, the exterior multiplication e_{dx^μ} and the interior multiplication $i\partial/\partial x^\mu$ can be identified with the fermion creation operator $\hat{\psi}^{*\mu}$ and the annihilation operator $\hat{\psi}_\mu$, and we have

$$d = \hat{\psi}^{*\mu} \boldsymbol{\nabla}_\mu, \quad d^\dagger = g^{-1/2} \boldsymbol{\nabla}_\mu g^{1/2} g^{\mu\nu} \hat{\psi}_\nu, \tag{2}$$

where g is the determinant of the metric tensor $g_{\mu\nu}$ and $\boldsymbol{\nabla}_\mu$ is the covariant derivative,

$$\boldsymbol{\nabla}_\mu = \frac{\partial}{\partial x^\mu} - \Gamma^\lambda_{\mu\nu} \hat{\psi}^{*\nu} \hat{\psi}_\lambda . \tag{3}$$

The Hamiltonian $\hat{H}$ (1) takes the form

$$2\hat{H} = -g^{-1/2} \boldsymbol{\nabla}_\mu g^{1/2} g^{\mu\nu} \boldsymbol{\nabla}_\nu + R_{\mu\nu\sigma\tau} \hat{\psi}^\sigma \hat{\psi}^{*\tau} \hat{\psi}^{*\nu} \hat{\psi}^\mu + g^{\mu\nu} \frac{\partial h}{\partial x^\mu} \frac{\partial h}{\partial x^\nu} + H_{\mu\nu}[\hat{\psi}^{*\mu}, \hat{\psi}^\nu], \tag{4}$$

where $R_{\mu\nu\sigma\tau}$ is the Riemann tensor, and $H_{\mu\nu}$ is the Hessian matrix,

 © 1998 American Institute of Physics

$$H_{\mu\nu}=(\partial_\mu\partial_\nu-\Gamma^\lambda_{\mu\nu}\partial_\lambda)h. \tag{5}$$

The corresponding Lagrangian is

$$\mathscr{L}=\frac{1}{2}\,g_{\mu\nu}\,\frac{dx^\mu}{dt}\,\frac{dx^\nu}{dt}+\frac{1}{2}\,g^{\mu\nu}\,\frac{\partial h}{\partial x^\mu}\,\frac{\partial h}{\partial x^\nu}+\psi^{*\mu}\left(\frac{d}{dt}\,\psi_\mu-\Gamma^\lambda_{\mu\nu}\psi_\lambda\,\frac{dx^\nu}{dt}\right)$$
$$+H_{\mu\nu}\psi^{*\nu}\psi^\mu+\tfrac{1}{4}\,R_{\mu\nu\sigma\tau}\psi^\mu\psi^\nu\psi^{*\sigma}\psi^{*\tau}. \tag{6}$$

The gradient flow equation of (6) is

$$\frac{dx^\mu}{dt}=\pm g^{\mu\nu}\,\frac{\partial h}{\partial x^\nu}. \tag{7}$$

A relevant instanton solution satisfies (7) and connects adjacent critical points.

There is one classical vacuum for each critical point of h.[2] Around each critical point, h has the expansion

$$h=h_0+\sum_i\,\lambda_i\xi_i^2, \tag{8}$$

where λ_i are eigenvalues of the Hessian matrix and ξ_i are the corresponding local coordinates. The classical vacuum around a critical point corresponds to an l form:

$$|l\rangle=\prod_{\lambda_i<0}\,\hat{\psi}^*_{\xi_i}|0\rangle, \tag{9}$$

where l represents the number of the excited fermions (the negative eigenvalues) and $|0\rangle$ is the bosonic vacuum.

The true vacuums are determined by the quantum tunneling between adjacent classical vacuums (9). According to Witten,[2] one has

$$\langle l+1|d_h|l\rangle=\sum_\gamma\,n_\gamma e^{-(h(P^{(l+1)})-h(P^{(l)}))}, \tag{10}$$

where n_γ is an integer assigned for each instanton path γ from $P^{(l)}$ to $P^{(l+1)}$. If a state $|l\rangle$ does not couple with any adjacent classical vacuums, that is, if $d_h|l\rangle=0$ and $\langle l|d_h=0$, $|l\rangle$ is a true vacuum.

If there are plural instanton paths connecting a pair of adjacent critical points, it is not so easy to determine whether the matrix element (10) is zero or not, because of the notorious minus signs associated with fermions. In fact, for SO(n), there are exactly two instanton solutions between each pair of adjacent classical vacuum.[1] The selection rule given in Ref. 1 is as follows. Find a reflection transformation that leaves $h, d_h, d_h^\dagger$, and, consequently, $\hat{H}$ invariant, and that interchanges the two instanton paths. If the parities of a pair of adjacent classical vacuums are the same, the corresponding matrix element does not vanish, and the two classical vacuums are not true vacuums. If the parities of the two classical vacuums are different, the corresponding matrix element vanishes, and the two classical vacuums can be true vacuums. Repeat this for each pair of adjacent classical vacuums, and one obtains all true vacuums.

In the following sections, we apply this selection rule to SO(n) and some related manifolds. We see that the number of true vacuums agrees with the de Rham cohomology for each manifold.

III. RP_n

In this section, we discuss the real projective space RP_n, which is very simple but nontrivial. The manifold RP_n can be identified with $(n+1)\times(n+1)$ real matrices A with conditions $A^t=A$, $A^2=A$, and tr $A=1$.[4] Concretely, we have

$$A=(x_ix_j),\quad i,j=0,1,2,...,n, \tag{11}$$

where $\Sigma_{i=1}^{n} x_i^2 \leqslant 1$ and $x_0 = \sqrt{1-\Sigma_{i=1}^{n} x_i^2}$. Let us introduce the following polar coordinates:

$$x_1 = \sin\ \theta_1\ \cos\ \theta_2,$$

$$x_2 = \sin\ \theta_1\ \sin\ \theta_2\ \cos\ \theta_3,$$

$$\vdots$$

$$x_{n-2} = \sin\ \theta_1\ \sin\ \theta_2 \cdots \sin\ \theta_{n-2}\ \cos\ \theta_{n-1},$$

$$x_{n-1} = \sin\ \theta_1\ \sin\ \theta_2 \cdots \sin\ \theta_{n-2}\ \sin\ \theta_{n-1}\ \cos\ \theta_n,$$

$$x_n = \sin\ \theta_1\ \sin\ \theta_2 \cdots \sin\ \theta_{n-2}\ \sin\ \theta_{n-1}\ \sin\ \theta_n, \tag{12}$$

with $-\pi/2 \leqslant \theta_i \leqslant \pi/2$.

Fix real numbers $c_0, c_1, \ldots, c_n$ with $c_{i+1} > 2c_i > 0$. Then

$$h = \sum_{i=0}^{n} c_i x_i^2 \tag{13}$$

is a Morse function on RP_n[4] with critical points $P^{(l)}$,

$$P^{(l)} = \mathrm{diag}(0,0,\ldots,0,x_l^2 = 1,0,\ldots,0). \tag{14}$$

Around $P^{(l)}$, h has the expansion

$$h = (c_0 - c_l)(x_0 x_l)^2 + \cdots + (c_{l-1} - c_l)(x_{l-1} x_l)^2 + c_l + (c_{l+1} - c_l)(x_{l+1} x_l)^2 + \cdots + (c_n - c_l)(x_n x_l)^2. \tag{15}$$

The classical vacuum around $P^{(l)}$ is an l fermions excited state,

$$|l\rangle = \hat{\psi}_{0l}^* \hat{\psi}_{1l}^* \cdots \hat{\psi}_{l-1l}^* |0\rangle, \tag{16}$$

where $\hat{\psi}_{ij}^*$ is the fermionic mode corresponding to $x_i x_j$.

We introduce a metric on RP_n and find reflection symmetries of the theory and instanton solutions concretely. We introduce the following metric:

$$g_{ij} = \frac{1}{2}\ \mathrm{tr}\ \frac{\partial A^t}{\partial \theta_i}\ \frac{\partial A}{\partial \theta_j}. \tag{17}$$

The nonzero components of g_{ij} and Christoffel symbols are

$$\begin{gathered} g_{11} = 1, \quad g_{ii} = \sin^2\ \theta_1\ \sin^2\ \theta_2 \cdots \sin^2\ \theta_{i-1}, \quad i = 2,\ldots,n, \\ \Gamma_{ij}^{j} = \cot\ \theta_i, \quad i = 1,\ldots,n-1, \quad j = i+1,\ldots,n. \end{gathered} \tag{18}$$

The Morse function h (13), the metric g_{ij} and the corresponding supercharge,

$$d = \hat{\psi}^{*i} \nabla_i = \hat{\psi}^{*i} \left(\frac{\partial}{\partial \theta_i} - \cot\ \theta_i \sum_{j=i+1}^{n}\ \hat{\psi}^{*j} \hat{\psi}_j \right), \tag{19}$$

are invariant under the following transformations:

$$[i]: \quad \theta_i, \hat{\psi}^{*i}, \hat{\psi}_i \to -\theta_i, -\hat{\psi}^{*i}, -\hat{\psi}_i, \quad i = 1,\ldots,n. \tag{20}$$

Subsequently, the theory is invariant under these reflection transformations. Under the transformation $[i]$, x_j transform to $-x_j$ if $j \geqslant i$, and x_j are invariant if $j < i$. The fermionic modes $\hat{\psi}_{jk}^*$ have the same transformation properties as $x_j x_k$.

From the gradient flow equation (6), we have

$$\begin{aligned}\frac{d\theta_i}{dt}&=(c_0-c_{i-1}+(c_i-c_0)\cos^2\ \theta_{i+1}+(c_{i+1}-c_0)\sin^2\ \theta_{i+1}\ \cos^2\ \theta_{i+2}\\&\quad+\cdots+(c_n-c_0)\sin^2\ \theta_{i+1}\ \sin^2\ \theta_{i+2}\cdots\sin^2\ \theta_n)\sin 2\theta_i\,,\quad i=1,...,n-1,\\\frac{d\theta_n}{dt}&=(c_n-c_{n-1})\sin 2\theta_n\,.\end{aligned}\tag{21}$$

One finds that there are exactly a pair of instanton paths that connect $|l-1\rangle$ and $|l\rangle$ (16),

$$A=\begin{matrix}\\ l-1\\ l\end{matrix}\begin{matrix}\begin{matrix}\quad l-1 & \quad l\end{matrix}\\ \begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & & & \\ & & 0 & & & & & \\ & & & \cos^2\ \theta_l & \pm\sin\ \theta_l\ \cos\ \theta_l & & & \\ & & & \pm\sin\ \theta_l\ \cos\ \theta_l & \sin^2\ \theta_l & & & \\ & & & & & 0 & & \\ & & & & & & \ddots & \\ & & & & & & & 0\end{pmatrix}\end{matrix},\tag{22}$$

where $0\leqslant\theta_l\leqslant\pi/2$. Each of these two instanton solutions causes a nonzero instanton effect between $|l-1\rangle$ and $|l\rangle$.

Let us consider the transformation $[l]$. This transformation interchanges the two instanton paths. Under this transformation the two classical vacuums transform as

$$|l-1\rangle\to|l-1\rangle,\quad |l\rangle\to(-1)^l|l\rangle.\tag{23}$$

Thus, if l is even, the two classical vacuums $|l-1\rangle$ and $|l\rangle$ have the same parities, which means that the classical vacuums $|l-1\rangle$ and $|l\rangle$ are quantum mechanically coupled, and are not true vacuums.[1] If l is odd, the two classical vacuums $|l-1\rangle$ and $|l\rangle$ have the opposite parities and the two instanton effects cancel each other.

One obtains the following results. For $n=$even there is only one true vacuum $|0\rangle$, and for $n=$odd there are two true vacuums $|0\rangle$ and $|n\rangle$. These results agrees with the de Rham cohomology of RP_n; for $n=$even, there is one nonzero Betti number $b_0=1$, and for $n=$odd, there are two nonzero Betti numbers, $b_0=b_n=1$.

IV. SO(n)

In Ref. 1, all the classical vacuums, all the instanton solutions, and symmetry transformations are identified for a certain Morse function on SO(n). We review them to find true vacuums for arbitrary n.

Let $A=(a_{ij})$ be a group element of SO(n), and fix real numbers $c_1,c_2,\ldots,c_n$ with $c_i>2c_{i+1}>0$. Then

$$h=\sum_{i=1}^{n} c_i a_{ii}\,,\tag{24}$$

is a Morse function[4] on SO(n) with critical points $P^{(l)}$,

$$P^{(l)}=\mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_n),\quad \epsilon_i=\pm1,\prod_i\ \epsilon_i=1.\tag{25}$$

The Morse index of $P^{(l)}$ is $\Sigma_{i=1}^{n}(n-i)[(1+\epsilon_i)/2]$. Around each critical point, h has the expansion

$$h=\sum_i \epsilon_i c_i+\sum_{i<j}(\lambda_{ij}\xi_{ij}^2+\mu_{ij}\eta_{ij}^2), \tag{26}$$

where

$$\lambda_{ij}=-\frac{\epsilon_j-\epsilon_i}{4}(c_j-c_i), \quad \xi_{ij}=a_{ij}+a_{ji}, \tag{27}$$

$$\mu_{ij}=-\frac{\epsilon_j+\epsilon_i}{4}(c_j+c_i), \quad \eta_{ij}=a_{ij}-a_{ji}. \tag{28}$$

A vacuum state that is localized around a critical point with the Morse index l is

$$|l\rangle=\prod_{i<j}\prod_{\epsilon_i=\epsilon_j=1}\hat{\psi}^*_{\eta_{ij}}\prod_{\epsilon_i>\epsilon_j}\hat{\psi}^*_{\xi_{ij}}|0\rangle. \tag{29}$$

There are exactly a pair of instanton solutions between the critical points $\mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_{i-1},-1,1,\epsilon_{i+2},\ldots,\epsilon_n)$ and $\mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_{i-1},1,-1,\epsilon_{i+2},\ldots,\epsilon_n)$:

$$\begin{pmatrix} \epsilon_1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & \epsilon_{i-1} & & & & & & \\ & & & \cos\theta & \mp\sin\theta & & & & \\ & & & \mp\sin\theta & -\cos\theta & & & & \\ & & & & & \epsilon_{i+2} & & & \\ & & & & & & \ddots & & \\ & & & & & & & \epsilon_n & \end{pmatrix}. \tag{30}$$

Some changes of the signs are needed in (30) for the instanton solutions between the critical points $\mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_{n-2},-1,-1)$ and $\mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_{n-2},1,1)$.

The theory is invariant under each of the supersymmetric generalizations[1] of the following reflection transformations:

$$a_{ij}\to -a_{ij} \quad \text{and} \quad a_{ji}\to -a_{ji}, \quad j\neq i, i=2,3,\ldots,n, \tag{31}$$

and the transposition

$$a_{ij}\to a_{ji}. \tag{32}$$

We call the corresponding supersymmetric transformations as $[i]$ and $[t]$, respectively. Supersymmetric generalization means that the corresponding fermionic modes are also transformed. A transformation that is useful to elucidate true vacuums is the transformation that interchanges the two instanton solutions (30).

For the instanton process corresponding to (30), the transformation $[i]$ is relevant. From (29), one sees what kind of fermionic modes are newly excited or suppressed in this process. In this process, only the fermionic modes, including the indices i or $i+1$, can newly be exited or suppressed. At first, the fermionic modes $\hat{\psi}^*_{ki}$ and $\hat{\psi}^*_{ki+1}$, $k=1,\ldots,i-1$, are excited in the pair $\hat{\psi}^*_{\xi_{ki}}\hat{\psi}^*_{\eta_{ki+1}}$ if $\epsilon_k=1$. They change into $\hat{\psi}^*_{\eta_{ki}}\hat{\psi}^*_{\xi_{ki+1}}$ in this process, but do not cause any changes in their parities. In this process, the mode $\hat{\psi}_{\xi_{ii+1}}$ is newly exited. As for the fermionic modes $\hat{\psi}^*_{ik}$ and $\hat{\psi}^*_{i+1k}$, $k=i+1,\ldots,n$, the corresponding exited modes are $\hat{\psi}^*_{\eta_{i+1k}}$ and $\hat{\psi}^*_{\xi_{i+1k}}$, for $\epsilon_k=1$ and $\epsilon_k=-1$, respectively. In this process they change into $\hat{\psi}^*_{\eta_{ik}}$ and $\hat{\psi}^*_{\xi_{ik}}$, respectively. Thus, in this

process the number of the index i increases by $n-i$, and the number of the index $i+1$ diminishes by $n-i-2$ in the excited fermionic modes. The number of excited η modes does not change, so this process is not prohibited by the transformation $[t]$ [the instanton process below (30) is prohibited by $[t]$]. Taking the transformation $[i]$ or $[i+1]$ into consideration, we see that the two instanton effects cancel each other if $n-i$ is odd, and they add up if $n-i$ is even. Applying this simple rule to all the instanton solutions, one can easily select all true vacuums.

Let us first consider SO$(2n-1)$. We mark the elements of the critical points off two by two,

$$\mathrm{diag}(\epsilon_1,\epsilon_2|\epsilon_3,\epsilon_4|\cdots|\epsilon_{2n-3},\epsilon_{2n-2}|\epsilon_{2n-1}). \tag{33}$$

From the above rule, one sees that there are instanton effects between the two states around the following critical points:

$$\mathrm{diag}(\epsilon_1,\epsilon_2|\cdots|1,-1|\cdots|\epsilon_{2n-1})\leftrightarrow\mathrm{diag}(\epsilon_1,\epsilon_2|\cdots|-1,1|\cdots|\epsilon_{2n-1}). \tag{34}$$

On the contrary, if $\epsilon_{2i-1}=\epsilon_{2i}(i=1,2,...,n-1)$, the corresponding state is not affected by instanton effects, and it is decided to be a true vacuum. Thus, the critical points that correspond to true vacuums are

$$\mathrm{diag}(\epsilon_1,\epsilon_1|\epsilon_3,\epsilon_3|\cdots|\epsilon_{2n-3},\epsilon_{2n-3}|1). \tag{35}$$

Their Morse indices are $\Sigma_{i=\mathrm{odd}}^{2n-3}(4n-3-2i)\epsilon_i$. This formula agrees with the de Rham cohomology of SO$(2n-1)$:$\Lambda(x_3,x_7,\ldots,x_{4n-5})$.[5]

In the same way, for SO$(2n)$, the critical points that correspond to true vacuums are found to be

$$\mathrm{diag}(\epsilon_{2n-1}|\epsilon_1,\epsilon_1|\epsilon_3,\epsilon_3|\cdots|\epsilon_{2n-3},\epsilon_{2n-3}|\epsilon_{2n-1}). \tag{36}$$

Their Morse indices are $\Sigma_{i=\mathrm{odd}}^{2n-3}(4n-3-2i)\epsilon_i+(2n-1)[(\epsilon_{2n-1}+1)/2]$, which agrees with the de Rham cohomology of SO$(2n)$:$\Lambda(x_3,x_7,\ldots,x_{4n-5},x_{2n-1})$.[5]

V. G_2

The group G_2 is a 14-dimensional submanifold of O(7). An element A of G_2 is given by three seven-component vectors $a_i(i=1,2,4)$;

$$M(7,3;\mathbf{R})\supset \mathrm{G}_2\ni A=(a_1,a_2,a_4), \tag{37}$$

with conditions $|a_i|=1$ $(i=1,2,4)$ and $(a_1,a_2)=(a_1,a_4)=(a_2,a_4)=(a_1a_2,a_4)=0$. The inner product and the absolute value are defined by $(a_i,a_j)=\Sigma_{k=1}^{7}a_{ki}a_{kj}$, $|a_i|=\sqrt{(a_i,a_i)}$ and the products a_ia_j are defined using octonians (see the Appendix).

Fix real numbers c_1, c_2 and c_4 with $c_1>c_2>c_4>0$. Then

$$h=c_1a_{11}+c_2a_{22}+c_4a_{44} \tag{38}$$

is a Morse function on G_2.[4] This Morse function has eight critical points with the Morse indices $l=0$, 3, 5, 6, 8, 9, 11, and 14. We are interested in the following four classical vacuums $|5\rangle$, $|6\rangle$, $|8\rangle$, and $|9\rangle$ that are localized around the critical points $(a_{11},a_{22},a_{44})=(-1,1,-1)$, $(1,-1,-1)$, $(-1,1,1)$, and $(1,-1,1)$, respectively. Around each critical point $(\epsilon_1,\epsilon_2,\epsilon_4)$, h has the expansion[4]

$$\begin{aligned}h=&\sum_i \epsilon_ic_i+\frac{1}{2}\sum_{i<j}(\lambda_{ij}\xi_{ij}^2+\mu_{ij}\eta_{ij}^2)-\epsilon_1c_1(a_{31}^2+a_{51}^2+a_{71}^2)-\epsilon_2c_2(a_{32}^2+a_{62}^2+a_{72}^2)\\&-\epsilon_4c_4(a_{54}^2+a_{64}^2+a_{74}^2)-\tfrac{1}{9}(a_{34}\ \ a_{52}\ \ a_{61})B(a_{34}\ \ a_{52}\ \ a_{61})^t,\end{aligned} \tag{39}$$

where $i,j=1,2,4$, λ_{ij}, μ_{ij}, ξ_{ij}, and η_{ij} are defined by (27) and (28), and B is the following 3 $\times$3 matrix:

$$B=\begin{pmatrix} 4\epsilon_4c_4+\epsilon_2c_2+\epsilon_1c_1 & 2\epsilon_2c_4+2\epsilon_4c_2-\epsilon_1\epsilon_2\epsilon_4c_1 & 2\epsilon_1c_4-\epsilon_1\epsilon_2\epsilon_4c_2+2\epsilon_4c_1 \\ 2\epsilon_2c_4+2\epsilon_4c_2-\epsilon_1\epsilon_2\epsilon_4c_1 & \epsilon_4c_4+4\epsilon_2c_2+\epsilon_1c_1 & \epsilon_1\epsilon_2\epsilon_4c_4-2\epsilon_1c_2-2\epsilon_2c_1 \\ 2\epsilon_1c_4-\epsilon_1\epsilon_2\epsilon_4c_2+2\epsilon_4c_1 & \epsilon_1\epsilon_2\epsilon_4c_4-2\epsilon_1c_2-2\epsilon_2c_1 & \epsilon_4c_4+\epsilon_2c_2+4\epsilon_1c_1 \end{pmatrix}.$$

From (39), one finds classical vacuums as

$$\begin{aligned} |5\rangle &= |a_{32},a_{62},a_{72},a_{24}+a_{42},(a_{34},a_{52},a_{61})\rangle, \\ |6\rangle &= |a_{31},a_{51},a_{71},a_{12}+a_{21},a_{14}+a_{41},(a_{34},a_{52},a_{61})\rangle, \\ |8\rangle &= |a_{32},a_{62},a_{72},a_{24}-a_{42},(a_{34},a_{52},a_{61}),a_{54},a_{64},a_{74}\rangle, \\ |9\rangle &= |a_{31},a_{51},a_{71},a_{12}-a_{21},a_{14}-a_{41},(a_{34},a_{52},a_{61}),a_{54},a_{64},a_{74}\rangle, \end{aligned} \tag{40}$$

where the modes in the ket vectors denote excited fermionic modes and (a_{34},a_{52},a_{61}) represents a certain linear combination of a_{34}, a_{52}, and a_{61}.

Let us discuss instanton solutions of G_2. Embedding G_2 into O(7) (see the Appendix), we see that the critical points that correspond to $|5\rangle$, $|6\rangle$, $|8\rangle$, and $|9\rangle$ are elements of SO(7). The Morse function (38) will be obtained from (24) by imposing some restrictions on the coordinates for SO(7). Accordingly, instanton solutions of G_2 should be included in that of SO(7). Consider the pair $|5\rangle$ and $|6\rangle$. There are four instanton solutions of SO(7) that connect $(a_{11},a_{22},a_{44})=(-1,1,-1)$ and $(1,-1,-1)$ [these are diag$(-1,1,-1,-1,1,-1,1)$ and diag$(1,-1,-1,-1,-1,1,1)$ in SO(7), respectively]:

$$\begin{pmatrix} \cos\alpha & \pm\sin\alpha & & & & & & \\ \pm\sin\alpha & -\cos\alpha & & & & & & \\ & & -1 & & & & & \\ & & & -1 & & & & \\ & & & & -\cos\beta & \pm\sin\beta & & \\ & & & & \pm\sin\beta & \cos\beta & & \\ & & & & & & 1 & \end{pmatrix}. \tag{41}$$

Setting $\alpha=\beta$, Eq. (41) represents two paths in G_2 and gives two instanton solutions in G_2. Thus, we have found that there are exactly two instanton solutions in G_2 that connect $|5\rangle$ and $|6\rangle$. Under the transformation $a_{12}\to -a_{12}, a_{21}\to -a_{21}$, which induces the transformation $a_{56}\to -a_{56}$, $a_{65}\to -a_{65}$ in terms of O(7), the two instanton solutions are interchanged. We use this transformation to elucidate true vacuums. In the same way, one finds that there are two similar instanton solutions between $|8\rangle$ and $|9\rangle$.

We will determine whether there are instanton effects or not between the classical vacuums (40). A reflection transformation that is useful to this purpose is such transformation under which the following each set of modes (a_{12},a_{21}), (a_{34},a_{52},a_{61}), (a_{14},a_{41}), and (a_{24},a_{42}) has the same parities. For SO(n), we have seen that a reflection transformation that transform a group element to a group element makes the theory invariant.[1] As G_2 is a subgroup of O(7), some reflection symmetries for SO(n) will survive as symmetries for G_2 with suitable restrictions. For G_2, one finds eight reflection transformations that transform a group element to a group element. These eight transformations, which are certain combined transformations of $[i]$ for SO(n), will be symmetries of the theory. Among these eight transformations, the four transformations in Table I interchange the two instanton solutions (41) with $\alpha=\beta$.

The classical vacuums $|5\rangle$ and $|6\rangle$ have the same parities under each of the four transformations. Thus, the matrix element $\langle 6|d_h|5\rangle$ is nonvanishing, and $|5\rangle$ and $|6\rangle$ are not true vacuums. In the same way, one finds that $|8\rangle$ and $|9\rangle$ are not true vacuums. We have found that true vacuums are $|0\rangle$, $|3\rangle$, $|11\rangle$, and $|14\rangle$, in accordance with de Rham cohomology of G_2; $b_p=1$ for $p=0$, 3, 11, 14, and $b_p=0$ for the others.

TABLE I. Symmetry transformations that interchange the two instanton solutions (41) with $\alpha=\beta$. The minus signs mean that the corresponding elements a_{ki}, $i=1,...,4$, $k=1,...,7$, reverse the signs (a_3 is not independent).

	1	2	3	4		1	2	3	4		1	2	3	4		1	2	3	4
1	+	−	−	+	1	+	−	+	+	1	+	−	+	−	1	+	−	−	−
2	−	+	+	−	2	−	+	−	−	2	−	+	−	+	2	−	+	+	+
3	−	+	+	−	3	+	−	+	+	3	+	−	+	−	3	−	+	+	+
4	+	−	−	+	4	+	−	+	+	4	−	+	−	+	4	−	+	+	+
5	+	−	−	+	5	−	+	−	−	5	+	−	+	−	5	−	+	+	+
6	−	+	+	−	6	+	−	+	+	6	−	+	−	+	6	+	−	−	−
7	−	+	+	−	7	−	+	−	−	7	+	−	+	−	7	+	−	−	−

VI. U(2)

In this section we discuss U(2), which is topologically identical with SO(2)×SU(2). We parametrize an element A of U(2) as

$$A=\begin{pmatrix} x_{11}+iy_{11} & x_{12}+iy_{12} \\ x_{21}+iy_{21} & x_{22}+iy_{22} \end{pmatrix}=\begin{pmatrix} \cos\ \theta e^{i(\alpha+\phi+\psi)} & -\sin\ \theta e^{i(\alpha+\phi-\psi)} \\ \sin\ \theta e^{i(\alpha-\phi+\psi)} & \cos\ \theta e^{i(\alpha-\phi-\psi)} \end{pmatrix},$$

$$0\leqslant\theta\leqslant\pi,\quad 0\leqslant\alpha,\phi,\psi\leqslant 2\pi. \tag{42}$$

Fix real numbers c_1, c_2 with $c_2>2c_1>0$. Then

$$h=c_1x_{11}+c_2x_{22}=(c_1\ \cos(\alpha+\phi+\psi)+c_2\ \cos(\alpha-\phi-\psi))\cos\ \theta \tag{43}$$

is a Morse function on U(2).[4] This Morse function has four critical points:

$$P^{(0)}=\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},\quad P^{(1)}=\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},\quad P^{(3)}=\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},\quad P^{(4)}=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{44}$$

We denote the corresponding classical vacuums as $|0\rangle$, $|1\rangle$, $|3\rangle$, and $|4\rangle$, respectively.

Around each critical point (ϵ_1,ϵ_2), h has the expansion

$$8h=(\epsilon_1c_1+\epsilon_2c_2)(8-(\epsilon_1x_{12}-\epsilon_2x_{21})^2-(\epsilon_1y_{12}+e_2y_{21})^2)-4\epsilon_1c_1y_{11}^2-4\epsilon_2c_2y_{22}^2. \tag{45}$$

From this expansion, one finds that

$$\begin{gathered} |0\rangle=|0\rangle,\quad |1\rangle=|\xi\rangle, \\ |3\rangle=|\eta,x_{12}+x_{21},y_{12}-y_{21}\rangle,\quad |4\rangle=|\xi,\eta,x_{12}-x_{21},y_{12}+y_{21}\rangle, \end{gathered} \tag{46}$$

where $\xi=\alpha+\phi+\psi$ and $\eta=\alpha-\phi-\psi$. We introduce the following metric:

$$g_{ij}=\frac{1}{2}\ \mathrm{tr}\ \frac{\partial A^\dagger}{\partial\theta^i}\frac{\partial A}{\partial\theta^j}, \tag{47}$$

where the indices i denote θ, α, ψ, and ϕ. The nonzero components of g_{ij} and Christoffel symbols are

$$\begin{gathered} g_{ii}=1,\quad g_{\phi\psi}=g_{\psi\phi}=\cos 2\theta, \\ \Gamma^0_{\phi\psi}=\Gamma^0_{\psi\phi}=\sin 2\theta,\quad \Gamma^\phi_{\theta\psi}=\Gamma^\psi_{\theta\phi}=-\frac{1}{\sin 2\theta}. \end{gathered} \tag{48}$$

The covariant derivatives are

$$\nabla_\alpha=\partial_\alpha,\quad \nabla_\theta=\partial_\theta+\frac{1}{\sin 2\theta}\,(\hat{\psi}^{*\phi}\hat{\psi}_\psi+\hat{\psi}^{*\psi}\hat{\psi}_\phi),$$

$$\nabla_\phi=\partial_\phi-\sin 2\theta\hat{\psi}^{*\psi}\hat{\psi}_\theta+\frac{1}{\sin 2\theta}\,\hat{\psi}^{*\theta}\hat{\psi}_\psi, \tag{49}$$

$$\nabla_\psi=\partial_\psi-\sin 2\theta\hat{\psi}^{*\phi}\hat{\psi}_\theta+\frac{1}{\sin 2\theta}\,\hat{\psi}^{*\theta}\hat{\psi}_\phi.$$

Consider the following simultaneous transformation:

$$\theta^i,\hat{\psi}^{*\theta^i},\hat{\psi}_{\theta^i}\to-\theta^i,-\hat{\psi}^{*\theta^i},-\hat{\psi}_{\theta^i},\quad \theta^i=\alpha,\psi,\phi, \tag{50}$$

which makes the Morse function h invariant. Under this transformation the covariant derivatives ∇_α, ∇_ψ, and ∇_ϕ reverse the sings, and ∇_θ is invariant. Accordingly, the supercharges $d_h, d_h^\dagger$, and the Hamiltonian is invariant under (50).

The gradient flow equations are

$$\frac{d\theta}{dt}=-(c_1\cos\xi+c_2\cos\eta)\sin\theta,$$

$$\frac{d\alpha}{dt}=-(c_1+c_2)\sin\theta\sin\alpha\cos(\phi+\psi)+(c_2-c_1)\cos\theta\cos\alpha\sin(\phi+\psi),$$

$$\frac{d\phi}{dt}=\frac{d\psi}{dt}=((c_2-c_1)\sin\alpha\cos(\phi+\psi)-(c_1+c_2)\cos\alpha\sin(\phi+\psi))\,\frac{\cos\theta}{1+\cos 2\theta}. \tag{51}$$

The degrees of the freedom has been reduced to three. Let us consider the equations in the (θ,η) plane. There are four fixed points $(\theta,\eta)=(0,0)$, $(0,\pi)$, $(\pi,0)$, and (π,π). Around these fixed points the gradient flow equations (51) become simple:

$$\begin{aligned}
&\theta\simeq 0,\quad \frac{d\eta}{dt}=-2c_2\sin\eta,\\
&\theta\simeq\pi,\quad \frac{d\eta}{dt}=2c_2\sin\eta,\\
&\eta\simeq 0,\quad \frac{d\theta}{dt}=-(c_2+c_1\cos\xi)\sin\theta,\\
&\eta\simeq\pi,\quad \frac{d\theta}{dt}=(c_2-c_1\cos\xi)\sin\theta.
\end{aligned} \tag{52}$$

One sees that $(\theta,\eta)=(0,0),(\pi,\pi)$ are stable fixed points and that $(\theta,\eta)=(0,\pi),(\pi,0)$ are unstable fixed points irrespective of the value of ξ.

The initial and the final conditions of a solution that connects $P^{(0)}$ and $P^{(1)}$ are both $\cos\theta\cos\eta=-1$. So, for solutions of (51) that connect $P^{(0)}$ and $P^{(1)}$, one concludes that $(\theta,\eta)\equiv(0,\pi)$ or $(\pi,0)$. Under these conditions, the gradient flow equation of ξ becomes

$$\frac{d\xi}{dt}=\pm 2c_1\sin\xi. \tag{53}$$

There are exactly two instanton solutions that connect $P^{(0)}$ and $P^{(1)}$,

TABLE II. Multiplication table of octonians e_j.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7
e_1	-1	e_3	$-e_2$	e_5	$-e_4$	e_7	$-e_6$
e_2	$-e_3$	-1	e_1	$-e_6$	e_7	e_4	$-e_5$
e_3	e_2	$-e_1$	-1	e_7	e_6	$-e_5$	$-e_4$
e_4	$-e_5$	e_6	$-e_7$	-1	e_1	$-e_2$	e_3
e_5	e_4	$-e_7$	$-e_6$	$-e_1$	-1	e_3	e_2
e_6	$-e_7$	$-e_4$	e_5	e_2	$-e_3$	-1	e_1
e_7	e_6	e_5	e_4	$-e_3$	$-e_2$	$-e_1$	-1

$$A=\begin{pmatrix} e^{\pm i\xi} & 0 \\ 0 & -1 \end{pmatrix}, \quad 0\leqslant\xi\leqslant\pi. \tag{54}$$

For solutions that connect $P^{(3)}$ and $P^{(4)}$, one sees that $(\theta,\eta)\equiv(0,0)$ or (π,π) and that there are exactly two instanton solutions:

$$A=\begin{pmatrix} e^{\pm i\xi} & 0 \\ 0 & 1 \end{pmatrix}, \quad 0\leqslant\xi\leqslant\pi. \tag{55}$$

By the transformation (50), ξ reverses the sign and the two instanton solutions (54) [(55)] are interchanged with each other. Under this transformation $|0\rangle$ and $|3\rangle$ have even parities and $|1\rangle$ and $|4\rangle$ have odd parities. Therefore, there are not instanton effects in the processes $P^{(0)}\to P^{(1)}$ and $P^{(3)}\to P^{(4)}$. Our result for U(2) is that $|0\rangle$, $|1\rangle$, $|3\rangle$, and $|4\rangle$ are all true vacuums. This result agrees with the de Rham cohomology of U(2); $b_p=1$ for $p=0$, 1, 3, 4 and $b_p=0$ for the others. Our result is one of the examples of the Kunneth formula for a product manifold $M\times N$: $b_k(M\times N)=\Sigma_{p+q=k}\, b_p(M)b_q(N)$.

VII. SUMMARY

We have studied supersymmetric quantum mechanics on RP_n, $\mathrm{SO}(n)$, G_2, and U(2) as concrete examples of Witten's Morse theory. We have identified all the true vacuums for each theory. The number of the true vacuums agrees with the de Rham cohomology for each manifold. We have seen that the simple instanton picture that the author proposed for $\mathrm{SO}(n)^4$ holds for the above manifolds. For each pair of adjacent classical vacuums, there are exactly two instanton solutions that are interchanged by a reflection transformation of the theory. Owing to the reflection symmetry, it is easy to decide whether the two instanton effects cancel each other or not.

ACKNOWLEDGMENTS

The author thanks Dr. Y. Yasui for introducing him to this subject. The author thanks Dr. H. Usui for useful discussions.

APPENDIX: MULTIPLICATION RULE IN G_2 AND EMBEDDING

We note the multiplication rule in G_2. For elements a_i of G_2, the following expansions by octonians e_j are taken into account:

$$a_i=\sum_{j=1}^{7} a_{ji}e_j=\begin{pmatrix} a_{1i} \\ \vdots \\ a_{7i} \end{pmatrix}.$$

Note that $e_0=1$ does not appear in the expansion.

The multiplication rules of octonians are defined as usual (see Table II); for example, $e_1e_2=e_3$, $e_5e_3=-e_6$, and $e_i^2=-1$, $e_ie_j=-e_je_i$ $(i\neq j)$. Explicitly, $a_3=a_1a_2$ is given by

$$a_3=\begin{pmatrix} -a_{31}a_{22}+a_{21}a_{32}-a_{51}a_{42}+a_{41}a_{52}-a_{71}a_{62}+a_{61}a_{72} \\ a_{31}a_{12}-a_{11}a_{32}+a_{61}a_{42}-a_{41}a_{62}-a_{71}a_{52}+a_{51}a_{72} \\ -a_{21}a_{12}+a_{11}a_{22}-a_{71}a_{42}-a_{61}a_{52}+a_{51}a_{62}+a_{41}a_{72} \\ a_{51}a_{12}-a_{61}a_{22}+a_{71}a_{32}-a_{11}a_{52}+a_{21}a_{62}-a_{31}a_{72} \\ -a_{41}a_{12}+a_{71}a_{22}+a_{61}a_{32}+a_{11}a_{42}-a_{31}a_{62}-a_{21}a_{72} \\ a_{71}a_{12}+a_{41}a_{22}+a_{51}a_{32}-a_{21}a_{42}-a_{31}a_{52}-a_{11}a_{72} \\ -a_{61}a_{12}-a_{51}a_{22}-a_{41}a_{32}+a_{31}a_{42}+a_{21}a_{52}+a_{11}a_{62} \end{pmatrix}.$$

An element A of G_2 is embedded into O(7) as follows: $A=(a_1,a_2,a_4)\rightarrow(a_1,a_2,a_1a_2,a_4,a_1a_4,a_4a_2,a_1a_4a_2)\in \mathrm{O}(7)$.

[1] K. Oshima, J. Math. Phys. **38**, 6281 (1997).
[2] E. Witten, J. Diff. Geom. **17**, 661 (1982).
[3] T. Hirokane, M. Miyajima, and Y. Yasui, J. Math. Phys. **34**, 2789 (1993); K. Oshima, Prog. Theor. Phys. **96**, 1237 (1996).
[4] I. Yokota, *Manifold and Morse Theory* (in Japanese) (Gendai Suugakusha, Kyoto, 1991).
[5] S. Iyanaga and Y. Kawada, *Encyclopedic Dictionary of Mathematics* (MIT Press, Cambridge, 1977).

Quantization of minisuperspaces as ordinary gauge systems

Claudio Simeone[a)]
Departamento de Física, Comisión Nacional de Energía Atómica, Av. del Libertador 8250, 1429 Buenos Aires, Argentina and Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón I, 1428, Buenos Aires, Argentina

(Received 24 November 1997; accepted for publication 17 March 1998)

Simple cosmological models are used to show that gravitation can be quantized as an ordinary gauge system if the Hamilton–Jacobi equation for the model under consideration is separable. In this situation, a canonical transformation can be performed such that in terms of the new variables the model has a linear and homogeneous constraint, and therefore canonical gauges are admissible in the path integral. This has the additional practical advantage that gauge conditions that do not generate Gribov copies are then easy to choose. © *1998 American Institute of Physics.* [S0022-2488(98)02906-5]

I. INTRODUCTION

The evolution of a cosmological model is given by the dynamics of the fields $g_{\mu\nu}(X)$ and of the matter field ϕ, which yield from the extremal condition imposed on the Einstein action. The change from the variables $g_{\mu\nu}$ to the set

$$g_{ab}(\mathbf{x},\tau),\ N(\mathbf{x},\tau),\ N^a(\mathbf{x},\tau)$$

where N and N^a are the lapse and shift functions defined by Kuchař[1] as a generalization of those introduced by Arnowitt, Deser and Misner,[2] $N=(-g^{00})^{-1/2}$, $N^a=g^{ab}g_{b0}$, gives the Einstein action the form[3]

$$S[g_{ab},p^{ab},N,N^a,\phi]=\int d\tau\int d^3x\ (p^{ab}\dot{g}_{ab}-N\mathcal{H}-N^a\mathcal{H}_a), \tag{1}$$

where the Hamiltonian and momentum constraints

$$\begin{aligned}\mathcal{H}&=\tfrac{1}{2}G_{abcd}\ p^{ab}p^{cd}-({}^3g)^{1/2}({}^3R-2\Lambda)+\mathcal{H}_{\mathrm{matt}}\approx 0,\\ \mathcal{H}_a&=-2\,g_{ac}\nabla_d\ p^{cd}+\mathcal{H}_{a,\mathrm{matt}}\approx 0\end{aligned} \tag{2}$$

reflect the general covariance of the theory.

The restriction to minisuperspace models and the choice of an homogeneous lapse and zero shift lead to a single constraint of the form[4]

$$H=\tfrac{1}{2}f_{ij}\ p^ip^j+V(q)\approx 0, \tag{3}$$

where f_{ij} is the reduced version of the DeWitt supermetric. The constraint H is a generator of gauge transformations giving rise to an infinite number of physically equivalent paths in phase space. If the action

$$S[q^i,p_i,N]=\int d\tau\ (p_i\dot{q}^i-NH) \tag{4}$$

a)Electronic mail: simeone@tandar.cnea.edu.ar

 © 1998 American Institute of Physics

is used to quantize the system by means of a path integral, the propagator diverges because of the sum over equivalent paths. This is solved by imposing a gauge condition that selects one path from each class of equivalent paths. Admissible gauge conditions are those which can be reached from any path by performing a gauge transformation that does not modify the action, so that a first restriction on admissible gauge conditions appears. A second—and general—condition is that gauge fixing must not generate Gribov copies, that is, the orbits (i.e., the points of the phase space connected by gauge transformations) must have no more than one intersection with the surface defined by the gauge.

Under the gauge transformation

$$\delta_\epsilon q^i = \epsilon(\tau)[q^i, H],\ \delta_\epsilon p_i = \epsilon(\tau)[p_i, H],\ \delta_\epsilon N = \dot{\epsilon}(\tau), \tag{5}$$

the action changes by

$$\delta_\epsilon S = \left[\epsilon(\tau) \left(p_i \frac{\partial H}{\partial p_i} - H \right) \right]_{\tau_1}^{\tau_2}, \tag{6}$$

so that for a quadratic Hamiltonian like that of Eq. (4) $\delta_\epsilon S$ vanishes only if

$$\epsilon(\tau_1) = 0 = \epsilon(\tau_2). \tag{7}$$

This introduces a restriction on admissible gauges: derivative gauge conditions like

$$\chi \dot{N} - \chi_0(q^i, p_i, N) = 0 \tag{8}$$

should be used,[5,6] reflecting a difference between gravitation and ordinary gauge systems, which have linear and homogeneous constraints so that $\delta_\epsilon S$ vanishes even if the boundaries are not mapped onto themselves, allowing then for canonical gauge conditions

$$\chi(q^i, p_i, \tau) = 0. \tag{9}$$

However, a system with a quadratic Hamiltonian can be made gauge invariant at the end points, improving the action with end point terms.[7] In a previous paper[8] we have shown that these terms can be seen as the result of a canonical transformation which identifies the constraint with one of the new momenta, yielding an action like that of an ordinary gauge system. In the present work we apply this idea to some minisuperspace models; we solve the Hamilton–Jacobi equation for them, and find the gauge-invariant action which makes simple canonical gauges admissible in the path integral. We show that this gauge fixing procedure makes it easy to avoid the Gribov problem, and, from a different point of view, we find that the path integral resulting after the gauge choice clearly manifests the (arbitrary) separation between true degrees of freedom and time.

II. GAUGE-INVARIANT ACTION

Let us consider the canonical transformation $(q^i, p_i) \to (\overline{Q}^i, \overline{P}_i)$ generated by the solution $W(q^i, \alpha_\mu, E)$ of the τ-independent Hamilton–Jacobi equation

$$H\left(q^i, \frac{\partial W}{\partial q^i} \right) = E, \tag{10}$$

matching the constants (α_μ, E) to the new momenta $(\overline{P}_\mu, \overline{P}_0)$. The new coordinates and momenta verify

$$\begin{aligned} [\overline{Q}^\mu, \overline{P}_0] &= [\overline{Q}^\mu, H] = 0, \\ [\overline{P}_\mu, \overline{P}_0] &= [\overline{P}_\mu, H] = 0, \\ [\overline{Q}^0, \overline{P}_0] &= [\overline{Q}^0, H] = 1. \end{aligned} \tag{11}$$

$\overline{Q}^{\mu}$ and $\overline{P}_{\mu}$ do not change under a gauge transformation and are then called observables, while $\overline{Q}^{0}$ has nonzero bracket with the constraint, suggesting its use to fix the gauge: In terms of the new variables the constraint is linear and homogeneous in the momenta, so that canonical gauges are admissible, and the condition that guarantees that the surface $\chi=0$ is not tangent to the orbits[9] (which will be used to ensure that it does not intersect them more than once) is fulfilled by $\chi \equiv \overline{Q}^{0}$:

$$\det([\chi,H])=[\overline{Q}^{0},H]\neq 0. \tag{12}$$

A second transformation, now in the space of observables, generated by

$$F=P_0\overline{Q}^{0}+f(\overline{Q}^{\mu},P_{\mu},\tau), \tag{13}$$

yields a new nonzero Hamiltonian

$$K=NP_0+\frac{\partial f}{\partial \tau}, \tag{14}$$

and a new set of nonconserved observables (Q^{μ},P_{μ}) that are therefore appropriate to characterize the trajectories in phase space. As a functional of Q^{i} and P_i the new gauge-invariant action reads

$$\mathcal{S}=\int_{\tau_1}^{\tau_2}\left(P_i\frac{dQ^{i}}{d\tau}-NP_0-\frac{\partial f}{\partial \tau}\right)d\tau, \tag{15}$$

so that the system is now an ordinary gauge system, i.e., one with a linear and homogeneous constraint $P_0\approx 0$ and a nonzero Hamiltonian $\partial f/\partial\tau$. The path integral then has the form

$$\int DQ^{0}\,DP_0\,DQ^{\mu}\,DP_{\mu}\,DN\;\delta(\chi)\;|[\chi,P_0]|\exp\left[i\int_{\tau_1}^{\tau_2}\left(P_i\frac{dQ^{i}}{d\tau}-NP_0-\frac{\partial f}{\partial\tau}\right)d\tau\right], \tag{16}$$

where $|[\chi,P_0]|$ is the Fadeev–Popov determinant and admits any canonical gauge. In terms of the original variables the gauge invariant action is written

$$\mathcal{S}=\int_{\tau_1}^{\tau_2}\left(p_i\frac{dq^{i}}{d\tau}-NH\right)d\tau+B \tag{17}$$

with[8]

$$B=[\overline{Q}^{i}\overline{P}_i-W+Q^{\mu}P_{\mu}-f]_{\tau_1}^{\tau_2}. \tag{18}$$

The existence of a gauge such that $\tau=\tau(q^{i})$ (intrinsic time)[10] and the vanishing of the end point terms B in that gauge and on the constraint surface assure that the new action $\mathcal{S}$ and the original one S weigh the paths in the same way.[8] This requirement determines the function f in Eq. (13).

III. MINISUPERSPACES

The procedure we have shown depends on the possibility of finding a solution to the Hamilton–Jacobi equation for the system. The separability of this equation and its application to cosmological models has been widely studied.[11,12] Here we will study simple models for which the Hamilton–Jacobi equation can be easily solved. Consider an isotropic and homogeneous Friedmann–Robertson–Walker (FRW) metric

$$ds^2=N^2d\tau^2-a^2(\tau)\left(\frac{dr^2}{1-kr^2}+r^2d\theta^2+r^2\,\sin^2\,\theta d\varphi^2\right).$$

The Hamiltonian constraint

$$H=\tfrac{1}{4}e^{-3\Omega}(\pi_\phi^2-\pi_\Omega^2)+\Lambda e^{3\Omega}\approx 0, \tag{19}$$

with $\Omega\sim\ln a(\tau)$ corresponds to a FRW metric with massless scalar field, nonzero cosmological constant and null curvature. The evolution is restricted to one of the two surfaces

$$\pi_\Omega=\pm\sqrt{\pi_\phi^2+4\Lambda e^{6\Omega}}$$

separated by $\pi_\Omega=0$; from a geometrical point of view this means that the topology of the surface where the gauge choice must select only one point from each orbit is that of half a plane. The τ-independent Hamilton–Jacobi equation has the solutions

$$W_\pm=\overline{P}\phi\pm\int d\Omega\sqrt{\overline{P}^2-4\overline{P}_0e^{3\Omega}+4\Lambda e^{6\Omega}}, \tag{20}$$

which are of the form

$$W=w(q^0,\overline{P}_0,\overline{P})+C(q^i)\overline{P}, \tag{21}$$

with $q^0=\Omega$, $q=\phi$. Fixing the gauge by means of the canonical condition

$$\chi\equiv\overline{Q}^0-g(\overline{P},T(\tau))=0 \tag{22}$$

with $T(\tau)$ a monotonous function of τ, as $\overline{Q}^0=\partial W_\pm/\partial\overline{P}_0=\overline{Q}^0(q^0,\overline{P}_0,\overline{P})$, if we choose

$$g(\overline{P},T(\tau))=\overline{Q}^0(q^0=T(\tau),\overline{P}_0=0,\overline{P})$$

in terms of the original variables we have

$$q^0=T(\tau). \tag{23}$$

The surface $\chi=0$ is thus a plane $\Omega=$constant for each value of τ. This guarantees that the gauge (22) does not produce Gribov copies, because if it did so at any τ, at another one it would be $[\chi,H]=0$, which is prevented by the gauge fixing procedure.

The additional end point terms B for the action of this system vanish in gauge (22) if

$$f=\overline{Q}P-w(T(\tau),\overline{P}_0=0,\overline{P}=P).$$

Thus, in this gauge and on the constraint surface we have

$$Q=\frac{\partial f}{\partial P}=C(q^i),$$

which together with Eq. (23) means that Q and τ define a hypersurface in the original configuration space. The Hamiltonian for the reduced system (Q,P) is

$$\frac{\partial f}{\partial\tau}=\mp\sqrt{P^2+4\Lambda e^{6T}}\,\frac{dT}{d\tau},$$

and it resembles that of a relativistic particle of T-dependent mass $m=2\Lambda^{1/2}e^{3T}$. In gauge (22) and on the constraint surface the path integral for the system then has the simple form

$$\langle\phi'',\Omega''|\phi',\Omega'\rangle=\int DQDP\exp\left[i\int_{T'}^{T''}(PdQ\pm\sqrt{P^2+4\Lambda e^{6T}}dT)\right], \tag{24}$$

where the endpoints are given by $T'=\Omega'$ and $T''=\Omega''$, while the paths go from $Q'=\phi'$ to $Q''=\phi''$.

Another easily solvable constraint is

$$H=\tfrac{1}{4}e^{-3\Omega}3(\pi_\phi^2-\pi_\Omega^2)+e^{3\Omega}V(\phi)\approx 0 \tag{25}$$

with $V(\phi)=\lambda e^{6\Omega}$, which corresponds to a FRW metric with a massless scalar field and an exponential potential which can reproduce an increasing cosmological constant. The topology is again like that of two disjoint planes. Multiplying by the positive-definite function $e^{3\Omega}$ an equivalent constraint is obtained:

$$H'=\tfrac{1}{4}(\pi_\phi^2-\pi_\Omega^2)+\lambda e^{6(\Omega+\phi)}\approx 0. \tag{26}$$

This is an example in which the new momentum $\overline{P}_0$ differs from the constraint H by a nonzero factor, so that the bracket of $\overline{Q}$ with the constraint is zero only on the constraint surface, but $\overline{Q}$ is still an observable. The solution of the τ-independent Hamilton–Jacobi equation is

$$W=\overline{P}(\phi-\Omega)+\frac{1}{\overline{P}}\left(\overline{P}_0(\phi+\Omega)-\frac{\lambda}{6}e^{6(\Omega+\phi)}\right) \tag{27}$$

and has the form (21) if $q^0=\phi+\Omega$ and $q=\phi-\Omega$. The canonical gauge

$$\chi\equiv\overline{Q}^0-\frac{T(\tau)}{\overline{P}}=0 \tag{28}$$

gives

$$q^0=\phi+\Omega=T(\tau),$$

corresponding to a plane $\phi+\Omega=$constant for each value of τ. This gauge does not generate Gribov copies, because if at any τ the plane intersected an orbit, at another τ they should be tangent, yielding $[\chi,H]=0$. As in the former example, with the choice of the function f such that the end points B vanish in gauge (28) the new coordinate Q and τ define a hypersurface in the original configuration space. The Hamiltonian for the reduced system described by Q and P is

$$\frac{\partial f}{\partial\tau}=\frac{\lambda}{P}e^{6T(\tau)}\frac{dT}{d\tau},$$

and the path integral in gauge (28) and on the constraint surface has the form

$$\langle\phi'',\Omega''|\phi',\Omega'\rangle=\int DQDP\exp\left[i\int_{T'}^{T''}\left(PdQ-\frac{\lambda}{P}e^{6T}dT\right)\right], \tag{29}$$

where $T'=\phi'+\Omega'$ and $T''=\phi''+\Omega''$, and the paths go from $Q'=\phi'-\Omega'$ to $Q''=\phi''-\Omega''$. As before, the evolution of the system described by the observables (Q,P) is given by a T-dependent Hamiltonian.

In a more general case,[8] whenever the Hamiltonian constraint has the form

$$H=G(\phi,\Omega)(\pi_\phi^2-\pi_\Omega^2)+V(\phi,\Omega)\approx 0 \tag{30}$$

with G and V positive definite functions such that $V(\phi,\Omega)/G(\phi,\Omega)=L_1(\phi+\Omega)L_2(\phi-\Omega)$ the change to the null coordinates

$$u=R_1(\phi-\Omega),\ \ v=R_2(\phi+\Omega)$$

with the choice $L_2(z)=R'(z)$ yields the equivalent constraint

$$H'=\pi_u\pi_v+\tfrac{1}{4}L(v) \tag{31}$$

related to H by $H=V(\phi,\Omega)H'$, with $L(v)\equiv L_1(v)/L_2(v)$. The generator function W which identifies the set $(\overline{Q}^i,\overline{P}_i)$ is

$$W=\overline{P}u+\frac{\overline{P}_0}{\overline{P}}v-\frac{1}{4\overline{P}}\int L(v)dv, \tag{32}$$

and has the form

$$W=w(v,\overline{P}_0,\overline{P})+C(u)\overline{P}. \tag{33}$$

The canonical gauge

$$\chi\equiv\overline{Q}^0-\frac{T(\tau)}{\overline{P}}=0 \tag{34}$$

is equivalent to

$$v=R_2(\phi+\Omega)=T(\tau).$$

The Gribov problem is avoided in the same way it was before, as the constraint hypersurface splits in

$$\pi_\Omega=\pm\sqrt{\frac{V(\phi,\Omega)}{G(\phi,\Omega)}+\pi_\phi^2}.$$

The generator function f giving the change to nonconserved observables is

$$f=\overline{Q}P+\frac{1}{4P}\int L(T)dT \tag{35}$$

and the new variables are related to the null ones by

$$\begin{gathered}Q^0=\frac{v}{P},\quad Q=u+\frac{v}{P^2}(1-P_0)-\frac{1}{4P^2}\int L(T)dT,\\ \pi_u=P,\quad \pi_v=\frac{P_0-L(v)}{P}.\end{gathered} \tag{36}$$

The system (Q,P) is governed by the nonzero Hamiltonian $\partial f/\partial\tau=(1/4P)\,L(T)\,dT/d\tau$, so that the path integral on the constraint surface and in gauge (34) is

$$\langle\phi'',\Omega''|\phi',\Omega'\rangle=\int DQDP\exp\left[i\int_{T'}^{T''}\left(PdQ-\frac{L(T)}{4P}dT\right)\right], \tag{37}$$

where $T'=R_2(\phi'+\Omega')$ and $T''=R_2(\phi''+\Omega'')$, and the paths go from $Q'=R_1(\phi'-\Omega')$ to $Q''=R_1(\phi''-\Omega'')$. The evolution of the reduced system is given by a T-dependent Hamiltonian. As in the preceding examples, this is nothing more, however, than a consequence of the fact that the gauge choice is not only a procedure to remove divergences from the path integral, but also a reduction procedure to physical degrees of freedom, giving rise to time and dynamical evolution.[3] This can soon be realized by examining the result of a ''wrong'' gauge choice, like a τ-independent one. We would then obtain a zero Hamiltonian for the reduced system, yielding no dynamical evolution.

IV. CONCLUSIONS

The presence of a quadratic constraint in the action of gravitation would make impossible, in principle, the quantization of cosmological models by means of the usual path integral procedure for gauge systems.[6] However, when the Hamilton–Jacobi equation for the system can be solved, this can mean nothing more than an inappropriate choice of coordinates and momenta. The change to variables such that the constraint is linear and homogeneous in the momenta provides the action

with invariance at the end points, making canonical gauges admissible in the path integral. In terms of the original variables the gauge-invariant action has additional end point terms, but these terms do not modify the dynamical trajectories of the system. The vanishing of the end point terms in a gauge that defines a hypersurface in the original configuration space guarantees that the improved action weighs the paths in the same way as the original one. However, because the new action is gauge invariant, the path integral can be computed in any canonical gauge. Canonical gauges make it easy to avoid Gribov copies, and also to define a global time.[3]

It should be emphasized that our procedure will only work in a limited class of minisuperspace models; even in the case of homogeneous and isotropic models, for an arbitrary potential $V(\phi,\Omega)$—yielding a topologically nontrivial constraint surface—it is most unlikely to work in general. However, the Hamilton–Jacobi equation is separable for constraints more general than those considered here: For example, solutions are known for models with more than one scalar field, or with a potential $V(\phi)\sim 1/\phi^2$, which reproduces an inflationary model with a decaying cosmological constant.[12] The existence of an intrinsic time and the possibility to avoid the Gribov problem in these and other separable models should be the object of a further analysis.

ACKNOWLEDGMENT

I wish to thank Rafael Ferraro for reading the manuscript and making helpful comments.

[1] K. V. Kuchař, J. Math. Phys. **17**, 777 (1976).
[2] R. Arnowitt, S. Deser, and C. Misner, in *Gravitation, an Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[3] A. O. Barvinsky, Phys. Rep. **230**, 237 (1993).
[4] J. J. Halliwell, in *Introductory Lectures on Quantum Cosmology*, Proceedings of the Jerusalem Winter School on Quantum Cosmology and Baby Universes, edited by T. Piran (World Scientific, Singapore, 1990).
[5] C. Teitelboim, Phys. Rev. D **25**, 3159 (1982).
[6] J. J. Halliwell, Phys. Rev. D **38**, 2468 (1988).
[7] M. Henneaux, C. Teitelboim, and J. D. Vergara, Nucl. Phys. B **387**, 391 (1992).
[8] R. Ferraro and C. Simeone, J. Math. Phys. **38**, 599 (1997).
[9] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, 1992).
[10] K. V. Kuchař, in *Proceedings of the Fourth Canadian Conference on General Relativity and Relativistic Astrophysics*, edited by G. Kunstatter, D. Vincent, and J. Williams (World Scientific, Singapore, 1992).
[11] D. S. Salopek and J. R. Bond, Phys. Rev. D **42**, 3936 (1990).
[12] D. S. Salopek and J. M. Stewart, Class. Quantum Grav. **9**, 1943 (1992).

The Maxwell–Vlasov equations in Euler–Poincaré form

Hernán Cendra[a)]
Control and Dynamical Systems, California Institute of Technology 107-81, Pasadena, California 91125

Darryl D. Holm[b)]
Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Mark J. W. Hoyle[c)]
Department of Mathematics, University of California, Santa Cruz, California 95064

Jerrold E. Marsden[d)]
Control and Dynamical Systems, California Institute of Technology 107-81, Pasadena, California 91125

(Received 22 September 1997; accepted for publication 21 January 1998)

Low's well-known action principle for the Maxwell–Vlasov equations of ideal plasma dynamics was originally expressed in terms of a mixture of Eulerian and Lagrangian variables. By imposing suitable constraints on the variations and analyzing invariance properties of the Lagrangian, as one does for the Euler equations for the rigid body and ideal fluids, we first transform this action principle into purely Eulerian variables. Hamilton's principle for the Eulerian description of Low's action principle then casts the Maxwell–Vlasov equations into Euler–Poincaré form for right invariant motion on the diffeomorphism group of position-velocity phase space, $\mathbb{R}^6$. Legendre transforming the Eulerian form of Low's action principle produces the Hamiltonian formulation of these equations in the Eulerian description. Since it arises from Euler–Poincaré equations, this Hamiltonian formulation can be written in terms of a Poisson structure that contains the Lie–Poisson bracket on the dual of a semidirect product Lie algebra. Because of degeneracies in the Lagrangian, the Legendre transform is dealt with using the Dirac theory of constraints. Another Maxwell–Vlasov Poisson structure is known, whose ingredients are the Lie–Poisson bracket on the dual of the Lie algebra of symplectomorphisms of phase space and the Born–Infeld brackets for the Maxwell field. We discuss the relationship between these two Hamiltonian formulations. We also discuss the general Kelvin–Noether theorem for Euler–Poincaré equations and its meaning in the plasma context. © *1998 American Institute of Physics.* [S0022-2488(98)00506-4]

I. INTRODUCTION

A. Reduction of action principles

Due to their wide applicability, the Maxwell–Vlasov equations of ideal plasma dynamics have been studied extensively. In 1958 Low[1] wrote down an action principle for them in preparation for studying stability of plasma equilibria. Low's action principle is expressed in terms of a mixture of Lagrangian particle variables and Eulerian field variables.

Following the initiative of Arnold[2] and its later developments (see Ref. 3 for background), we start with a purely Lagrangian description of the plasma and investigate the invariance properties of the corresponding action. Using this setup and recent developments in the theory of the Euler–

[a)]Permanent address: Universidad Nacional del Sur, 8000 Bahia Blanca, Argentina.
Electronic mail: uscendra@criba.edu.ar
[b)]Electronic mail: dholm@lanl.gov
[c)]Electronic mail: mjwh1@cats.ucsc.edu
[d)]Electronic mail: marsden@cds.caltech.edu

 © 1998 American Institute of Physics

Poincaré equations[4,5] due to Holm, Marsden, and Ratiu,[6] we are able to cast Low's action principle into a purely Eulerian description.

In this paper, we start with the *standard* form of Hamilton's variational principle (in the Lagrangian representation) and *derive* the new Eulerian action principle by a systematic reduction process, much as one does in the corresponding derivation of Poisson brackets in the Hamiltonian formulation of the Maxwell–Vlasov equations starting with the *standard canonical brackets* and proceeding by symmetry reduction (as in Ref. 7). In particular, the Eulerian action principle we obtain in this way is different from the ones found in Ye and Morrison[8] by *ad hoc* procedures. We also mention that the method of reduction of variational principles we develop naturally justifies constraints on the variations of the so-called ''Lin constraint'' form, well known in fluid mechanics.

The methods of this paper are based on reduction of variational principles, that is, on Lagrangian reduction (see Refs. 9–12). These methods have also been useful for systems with nonholonomic constraints. This has been demonstrated in the work of Bloch *et al.*,[13] who derived the reduced Lagrange d'Alembert equations for nonholonomic systems, which also have a constrained variational structure. The methods of the present paper should enhance the applicability of the Lagrangian reduction techniques for even wider classes of continuum systems.

B. Passage to the Hamiltonian formulation

The Hamiltonian structure and nonlinear stability properties of the equilibrium solutions for the Maxwell–Vlasov system have been thoroughly explored. Some of the key references are Iwínski and Turski,[14] Morrison,[15] Marsden and Weinstein,[7] and Holm, Marsden, Ratiu, and Weinstein.[16] See also the introduction and bibliography of Marsden *et al.*[17] for a guide to the history and literature of this subject.

In our approach, Lagrangian reduction leads to the Euler–Poincaré form of the equations, which is still in the Lagrangian formulation. Using this setup, one may pass from the Lagrangian to the Hamiltonian formulation of the Maxwell–Vlasov equations by Legendre transforming the action principle in the Eulerian description at either the level of the group variables (the level that keeps track of the particle positions), or at the level of the Lie algebra variables. One must be cautious in this procedure because the relevant Hamiltonian and Lagrangian are degenerate. We deal with this degeneracy by using a version of the Dirac theory of constraints.

Legendre transforming at the group level leads to a canonical Hamiltonian formulation and the latter leads to a new Hamiltonian formulation of the Maxwell–Vlasov equations in terms of a Poisson structure containing the Lie–Poisson bracket on the dual of a semidirect product Lie algebra. This new formulation leads us naturally to the starting point for Hamiltonian reduction used by Marsden and Weinstein[7] (see also Refs. 15 and 18).

C. Stability and asymptotics

The new Hamiltonian formulation of the Maxwell–Vlasov system places these equations into a framework in which one can use the energy-momentum and energy-Casimir methods for studying nonlinear stability properties of their relative equilibrium solutions. This is directly in line with Low's intended program, since the study of stability was Low's original motivation for writing his action principle. Sample references in this direction are Holm, Marsden, Weinstein, and Ratiu,[16] Morrison,[19] Morrison and Pfirsch,[20] Wan,[21] Batt and Rein[22] and Batt, Morrison, and Rein.[23] Other historical references for the Lagrangian approach to the Maxwell–Vlasov equations include Sturrock,[24] Galloway and Kim,[25] and Dewar.[26]

The Eulerian formulation of Low's action principle also casts it into a form that is amenable to asymptotic expansions and creation of approximate theories (such as guiding center theories) possessing the same mathematical structure arising from the Euler–Poincaré setting. See, for example, Ref. 27 for applications of this approach of Hamilton's principle asymptotics in geophysical fluid dynamics.

D. Comments on the Maxwell–Vlasov system

The rest of this paper will be concerned with variational principles for the Maxwell–Vlasov system of equations for the dynamics of an ideal plasma. These equations have a long history

dating back at least to Jeans,[28] who used them in a simpler form known as the Poisson–Vlasov system to study structure formation on stellar and galactic scales. Even before Jeans, Poincaré[4,29] had investigated the stability of equilibrium solutions of the Poisson–Vlasov system for the purpose of determining the stability conditions for steller configurations. The history of the efforts to establish stellar stability conditions using the Poisson–Vlasov system is summarized by Chandrasekhar.[30] The Poisson–Vlasov system is also used to describe the self-consistent dynamics of an electrostatic collisionless plasma, whereas the Maxwell–Vlasov system is used to describe the dynamics of a collisionless plasma evolving self-consistently in an electromagnetic field.

E. Organization of the paper

The paper is organized as follows. In Sec. II we introduce the Maxwell–Vlasov equations. In Sec. III we state the Euler–Poincaré theorem for Lagrangians depending on parameters along with the associated Kelvin–Noether theorem. This general theorem plays a key role in our analysis. In Sec. V we reformulate these equations in a purely Eulerian form and show how they satisfy the Euler–Poincaré theorem. The following section reviews some aspects of the Legendre transformation for degenerate Lagrangians. In Sec. IV we reprise Low's action principle for the Maxwell–Vlasov equations. In Sec. VII we cast the Euler–Poincaré formulation of the Maxwell–Vlasov equations into Hamiltonian form possessing a Poisson structure that contains a Lie–Poisson bracket. In Sec. VIII we summarize our conclusions.

II. THE MAXWELL–VLASOV EQUATIONS

The Maxwell–Vlasov system of equations describes the single particle distribution for a set of charged particles of one species moving self-consistently in an electromagnetic field. In this description, the Boltzmann function $f(\mathbf{x},\mathbf{v},t)$ is viewed as the instantaneous probability density function for the particle distribution, i.e., given a region Ω of phase space, the probability of finding a particle in that region is

$$\int_\Omega d\mathbf{x}\, d\mathbf{v} f(\mathbf{x},\mathbf{v},t), \tag{2.1}$$

where $\mathbf{x}$ and $\mathbf{v}$ are the current positions and velocities of the plasma particles. Thus, if the phase-space domain Ω is the whole $(\mathbf{x},\mathbf{v})$ space, the value of this integral at a certain time t is normalized to unity.

As is customary, we assume that the particles of the plasma obey dynamical equations and that the plasma density f is advected as a scalar along the particle trajectories in phase space, i.e.,

$$\frac{\partial f}{\partial t} + \dot{\mathbf{x}}\cdot\nabla_{\mathbf{x}} f + \dot{\mathbf{v}}\cdot\nabla_{\mathbf{v}} f = 0. \tag{2.2}$$

In this equation, and in the sequel, an overdot refers to a time derivative along a phase space trajectory, and $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{v}}$ denote the gradient operators with respect to position and velocity, respectively. For pressureless motion in the electromagnetic field of the charged particle distribution, the acceleration of a particle is given by

$$\ddot{\mathbf{x}} = -\frac{q}{m}\left[\nabla_{\mathbf{x}}\Phi + \frac{\partial \mathbf{A}}{\partial t} - \mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right], \tag{2.3}$$

where (q/m) denotes the charge-to-mass ratio of an individual particle, Φ is the electric potential, and $\mathbf{A}$ is the magnetic vector potential. Substituting this expression for $\dot{\mathbf{v}}$ in Eq. (2.2) yields

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_{\mathbf{x}} f - \frac{q}{m}\left[\nabla_{\mathbf{x}}\Phi + \frac{\partial \mathbf{A}}{\partial t} - \mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right]\cdot\nabla_{\mathbf{v}} f = 0. \tag{2.4}$$

This is the ***Vlasov equation*** (also called the collisionless Boltzmann, or Jeans equation). The system is completed by the Maxwell equations with sources:

$$\nabla_{\mathbf{x}}\cdot\mathbf{E} = \rho, \quad \nabla_{\mathbf{x}}\times\mathbf{B} = \frac{\partial \mathbf{E}}{\partial t} + \mathbf{j}, \tag{2.5}$$

where **E** and **B** are the electric and magnetic field variables, respectively, ρ is the charge density, and **j** is the current density. These quantities are expressed in terms of the Boltzmann function f and the Maxwell scalar and vector potentials Φ and **A** by

$$\mathbf{E}=-\nabla_{\mathbf{x}}\Phi-\frac{\partial \mathbf{A}}{\partial t},\quad \mathbf{B}=\nabla_{\mathbf{x}}\times\mathbf{A},$$
$$\rho(\mathbf{x},t)=q\int d\mathbf{v}\ f(\mathbf{x},\mathbf{v},t),\quad \mathbf{j}(\mathbf{x},t)=q\int d\mathbf{v}\ \mathbf{v}f(\mathbf{x},\mathbf{v},t). \tag{2.6}$$

By their definitions, **E** and **B** satisfy the kinematic Maxwell equations

$$\nabla_{\mathbf{x}}\cdot\mathbf{B}=0,\quad \nabla_{\mathbf{x}}\times\mathbf{E}=-\frac{\partial \mathbf{B}}{\partial t}. \tag{2.7}$$

Equations (2.4)–(2.7) comprise the ***Maxwell–Vlasov equations***. When **A** is absent, the field is electrostatic and one obtains the Poisson–Vlasov equations. The Poisson–Vlasov system can also be used to describe a self-gravitating collisionless fluid, and so it forms a model for the evolution of galactic dynamics (see, e.g., Ref. 31).

Note that the integral in (2.1) is independent of time (as the region and the function f evolve), since the vector field defining the motion of particles [see Eq. (2.3)] is divergence free with respect to the standard volume element on velocity phase space. Thus, one may interpret f either as a density or as a scalar. For our purposes later, we will need to be careful with the distinction, since the volume-preserving nature of the flow of particles will be a consequence of our variational principle and will not be imposed at the outset.

III. THE EULER–POINCARÉ EQUATIONS, SEMIDIRECT PRODUCTS, AND KELVIN'S THEOREM

A. The general Euler–Poincaré equations

Here we recall from Ref. 6 the general form of the Euler–Poincaré equations and their associated Kelvin–Noether theorem. In the next section, we will immediately specialize these statements for a general invariance group G to the case of plasmas when G is the diffeomorphism group, Diff($T\mathbb{R}^3$). We shall state the general theorem for right actions and right invariant Lagrangians, which is appropriate for the Maxwell–Vlasov situation. The notation is as follows.

(i) There is a *right* representation of the Lie group G on the vector space V and G acts in the natural way from the *right* on $TG\times V^*:(v_g,a)h=(v_gh,ah)$.

(ii) $\rho_v:\mathfrak{g}\to V$ is the linear map given by the corresponding right action of the Lie algebra on $V:\rho_v(\xi)=v\xi$, and $\rho_v^*:V^*\to\mathfrak{g}^*$ is its dual. The $\mathfrak{g}$–action on $\mathfrak{g}^*$ and V^* is defined to be *minus* the dual map of the $\mathfrak{g}$–action on $\mathfrak{g}$ and V, respectively, and is denoted by $\mu\xi$ and $a\xi$ for $\xi\in\mathfrak{g}$, $\mu\in\mathfrak{g}^*$, and $a\in V^*$. For $v\in V$ and $a\in V^*$, it will be convenient to write
$$v\diamond a=\rho_v^*a,\quad \text{i.e.,}\quad \langle v\diamond a,\xi\rangle=\langle a,v\xi\rangle=-\langle v,a\xi\rangle,$$
for all $\xi\in\mathfrak{g}$. Note that $v\diamond a\in\mathfrak{g}^*$.

(iii) Let $\mathcal{Q}$ be a manifold on which G acts *trivially* and assume that we have a function $L:TG\times T\mathcal{Q}\times V^*\to\mathbb{R}$ which is right G-invariant.

(iv) In particular, if $a_0\in V^*$, define the Lagrangian $L_{a_0}:TG\times T\mathcal{Q}\to\mathbb{R}$ by $L_{a_0}(v_g,u_q)=L(v_g,u_q,a_0)$. Then L_{a_0} is right invariant under the lift to $TG\times T\mathcal{Q}$ of the right action of G_{a_0} on $G\times\mathcal{Q}$.

(v) Right G-invariance of L permits us to define $l:\mathfrak{g}\times T\mathcal{Q}\times V^*\to\mathbb{R}$ by
$$l(v_gg^{-1},\ u_q,\ ag^{-1})=L(v_g,u_q,a).$$
Conversely, this relation defines for any $l:\mathfrak{g}\times T\mathcal{Q}\times V^*\to\mathbb{R}$ a right G-invariant function $L:TG\times T\mathcal{Q}\times V^*\to\mathbb{R}$.

(vi) For a curve $g(t)\in G$, let $\xi(t):=\dot{g}(t)g(t)^{-1}$ and define the curve $a(t)$ as the unique solution of the linear differential equation with time-dependent coefficients $\dot{a}(t)=-a(t)\xi(t)$

with initial condition $a(0)=a_0$. The solution can be equivalently written as $a(t)=a_0 g(t)^{-1}$.

Theorem 3.1: The following are equivalent:

(i) Hamilton's variational principle holds:

$$\delta\int_{t_1}^{t_2} L_{a_0}(g(t),\dot{g}(t),q(t),\dot{q}(t))\,dt=0, \tag{3.1}$$

for variations of g and q with fixed endpoints.

(ii) $(g(t),q(t))$ satisfies the Euler–Lagrange equations for L_{a_0} on $G\times Q$.

(iii) The constrained variational principle,[32]

$$\delta\int_{t_1}^{t_2} l(\xi(t),q(t),\dot{q}(t),a(t))\,dt=0, \tag{3.2}$$

holds on $\mathfrak{g}\times Q$, upon using variations of the form

$$\delta\xi=\frac{\partial\eta}{\partial t}-ad_\xi\eta=\frac{\partial\eta}{\partial t}-[\xi,\eta],\quad \delta a=-a\eta, \tag{3.3}$$

where $\eta(t)\in\mathfrak{g}$ vanishes at the endpoints and $\delta q(t)$ is unrestricted except for vanishing at the endpoints.

(iv) The following system of Euler–Poincaré equations (with a parameter) coupled with Euler–Lagrange equations holds on $\mathfrak{g}\times TQ\times V^*$:

$$\frac{\partial}{\partial t}\frac{\delta l}{\delta\xi}=-\mathrm{ad}^*_\xi\frac{\delta l}{\delta\xi}+\frac{\delta l}{\delta a}\diamond a \tag{3.4}$$

and

$$\frac{\partial}{\partial t}\frac{\partial l}{\partial\dot{q}^i}-\frac{\partial l}{\partial q^i}=0. \tag{3.5}$$

The strategy of the proof is simple: one just determines the form of the variations on the reduced space $\mathfrak{g}\times Q\times V^*$ that are induced by variations on the unreduced space $TG\times TQ$ and includes the relation of $a(t)$ to a_0. One then carries the variational principle to the quotient. See Ref. 6 for details. Here we have included the extra factor of Q which is needed in the present application; this will be the space of potentials for the Maxwell field. This extra factor does not substantively alter the arguments.

B. The Kelvin–Noether Theorem

We start with a Lagrangian L_{a_0} depending on a parameter $a_0\in V^*$ as above and introduce a manifold $\mathscr{C}$ on which G acts. We assume this is also a right action and suppose we have an equivariant map $\mathscr{K}:\mathscr{C}\times V^*\to\mathfrak{g}^{**}$.

In the case of continuum theories, the space $\mathscr{C}$ is chosen to be a loop space and $\langle\mathscr{K}(c,a),\mu\rangle$ for $c\in\mathscr{C}$ and $\mu\in\mathfrak{g}^*$ will be a circulation. This class of examples also shows why we *do not* want to identify the double dual $\mathfrak{g}^{**}$ with $\mathfrak{g}$.

Define the ***Kelvin–Noether quantity*** $I:\mathscr{C}\times\mathfrak{g}\times TQ\times V^*\to\mathbb{R}$ by

$$I(c,\xi,q,\dot{q},a)=\left\langle\mathscr{K}(c,a),\frac{\delta l}{\delta\xi}(\xi,q,\dot{q},a)\right\rangle. \tag{3.6}$$

Theorem 3.2 (Kelvin–Noether): Fixing $c_0\in\mathscr{C}$, let $\xi(t),q(t),\dot{q}(t),a(t)$ satisfy the Euler–Poincaré equations and define $g(t)$ to be the solution of $\dot{g}(t)=\xi(t)g(t)$ and, say, $g(0)=e$. Let $c(t)=g(t)^{-1}c_0$ and $I(t)=I(c(t),\xi(t),q(t),\dot{q}(t),a(t))$.

Then

$$\frac{d}{dt}I(t)=\left\langle \mathscr{K}(c(t),a(t)),\frac{\delta l}{\delta a}\diamond a\right\rangle. \tag{3.7}$$

The proof of this theorem is relatively straightforward; we refer to Ref. 6. We shall express the relation (3.7) explicitly for Maxwell–Vlasov plasmas at the end of Sec. VII.

IV. AN ACTION FOR THE MAXWELL–VLASOV EQUATIONS

A typical element of $T\mathbb{R}^3\cong\mathbb{R}^3\times\mathbb{R}^3$ will be denoted $\mathbf{z}=(\mathbf{x},\mathbf{v})$. We let $\pi_s:T\mathbb{R}^3\to\mathbb{R}^3$ and $\pi_v:T\mathbb{R}^3\to\mathbb{R}^3$ be the projections $\pi_s(\mathbf{z})=\mathbf{x}$ and $\pi_v(\mathbf{z})=\mathbf{v}$ onto the first and second factors, respectively.

A. Spaces of fields

We let $\mathrm{Diff}(T\mathbb{R}^3)$ denote the group of C^∞-diffeomophisms from $T\mathbb{R}^3$ onto itself. An element $\psi\in\mathrm{Diff}(T\mathbb{R}^3)$ maps plasma particles having initial position and velocity $(\mathbf{x}_0,\mathbf{v}_0)$ to their current position and velocity $(\mathbf{x},\mathbf{v})=\psi(\mathbf{x}_0,\mathbf{v}_0)$. This is the particle evolution map. We shall sometimes abbreviate $(\mathbf{x}_0,\mathbf{v}_0)=\mathbf{z}_0$, $(\mathbf{x},\mathbf{v})=\mathbf{z}$, etc. The spatial components of $\psi(\mathbf{x}_0,\mathbf{v}_0)$ are written as $\mathbf{x}(\mathbf{x}_0,\mathbf{v}_0)$ and the velocity components as $\mathbf{v}(\mathbf{x}_0,\mathbf{v}_0)$. We shall also use the following notation:

(i) $\mathscr{V}=C^\infty(\mathbb{R}^3,\mathbb{R})$ is the space of electric potentials $\Phi(\mathbf{x})$;
(ii) $\mathscr{A}$ is the space of magnetic potentials $A(\mathbf{x})$;
(iii) $\mathscr{F}=C^\infty(T\mathbb{R}^3,\mathbb{R})$ is the space of plasma densities $f(\mathbf{x},\mathbf{v})$;
(iv) $\mathscr{F}_0=C_0^\infty(T\mathbb{R}^3,\mathbb{R})$ is the space of plasma densities with compact support; and
(v) $\mathscr{D}_0=C_0^\infty(\mathbb{R}^3,\mathbb{R})$ is a space of test functions, denoted $\varphi(\mathbf{x})$.

The test functions $\varphi(\mathbf{x})$ are used to localize the variational principle. Thus, once one obtains Euler–Lagrange equations depending on f_0 and φ_0, if their validity can be naturally extended for any f_0 and φ_0, which will happen in our case, then we shall consider those extended equations to be the Euler–Lagrange equations of the system. We will usually be interested in the Euler–Lagrange equations for $f_0>0$ and $\varphi_0=1$.

B. The Lagrangian and the action

For each choice of the initial plasma distribution function f_0 and the test function φ_0, we define the Lagrangian

$$\begin{aligned}L_{f_0,\varphi_0}(\psi,\dot\psi,\Phi,\dot\Phi,\mathbf{A},\dot{\mathbf{A}})=&\int d\mathbf{x}_0\,d\mathbf{v}_0\,f_0(\mathbf{x}_0,\mathbf{v}_0)\Big(\frac{1}{2}m|\dot{\mathbf{x}}(\mathbf{x}_0,\mathbf{v}_0)|^2+\frac{1}{2}m|\dot{\mathbf{x}}(\mathbf{x}_0,\mathbf{v}_0)-\mathbf{v}(\mathbf{x}_0,\mathbf{v}_0)|^2\\&+q\dot{\mathbf{x}}(\mathbf{x}_0,\mathbf{v}_0)\cdot\mathbf{A}(\mathbf{x}(\mathbf{x}_0,\mathbf{v}_0))-q\Phi(\mathbf{x}(\mathbf{x}_0,\mathbf{v}_0))\Big)\\&+\frac{1}{2}\int d\mathbf{r}\,\varphi_0(\mathbf{r})\Big(\Big|\nabla_{\mathbf{r}}\Phi+\frac{\partial\mathbf{A}}{\partial t}(\mathbf{r})\Big|^2-|\nabla_{\mathbf{r}}\times\mathbf{A}(\mathbf{r})|^2\Big).\end{aligned} \tag{4.1}$$

This Lagrangian is the natural generalization of that for an N-particle system, with terms corresponding to kinetic energy, electric and magnetic field energy, the usual magnetic coupling term with coupling constant q (the electric charge), and a constraint that ties the Eulerian fluid velocity $\mathbf{v}$ to $\dot{\mathbf{x}}$, the material derivative of the Lagrangian particle trajectory. Here $\mathbf{x}$ and $\mathbf{v}$ are Lagrangian phase space variables, while $\mathbf{A}$ and Φ are Eulerian field variables. Thus, there should be no confusion created by the slight abuse of notation in abbreviating $\partial\mathbf{A}/\partial t$ and $\partial\Phi/\partial t$ as $\dot\Phi$ and $\dot{\mathbf{A}}$, respectively, in the arguments of the Lagrangian. This Lagrangian is inspired by Low.[1] However, we have added the term

$$\tfrac{1}{2}m|\dot{\mathbf{x}}(\mathbf{x}_0,\mathbf{v}_0)-\mathbf{v}(\mathbf{x}_0,\mathbf{v}_0)|^2,$$

which allows $\mathbf{v}$ to be varied independently in the variational treatment.

Consider the action

$$\mathfrak{S}=\int dt\; L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}}),$$

defined on the family of curves $(\psi(t),\Phi(t),\mathbf{A}(t))$ satisfying the usual fixed-endpoint conditions $(\psi(t_i),\Phi(t_i),\mathbf{A}(t_i))=(\psi_i,\Phi_i,\mathbf{A}_i)$, $i=1,2$. One now applies the standard techniques of the calculus of variations. In particular, integration by parts can be performed since f_0 and φ_0 have compact support. Moreover, once the Euler–Lagrange equations have been obtained, their validity can be easily extended in a natural way for $f_0>0$ and $\varphi_0=1$.

C. Derivation of the equations

To write the equations of motion, we need some additional notation. Consider the evolution map $\psi_t(\mathbf{x}_0,\mathbf{v}_0)=(\mathbf{x},\mathbf{v})$ so that ψ_t relates the initial positions and velocities of fluid particles to their positions and velocities at time t. Let $\mathbf{u}$ be the corresponding vector field:

$$\mathbf{u}(\mathbf{x},\mathbf{v}):=\dot{\psi}_t\circ\psi_t^{-1}(\mathbf{x},\mathbf{v})=:\dot{\mathbf{x}}\,\frac{\partial}{\partial\mathbf{x}}+\dot{\mathbf{v}}\,\frac{\partial}{\partial\mathbf{v}},$$

so the components of $\mathbf{u}$ are $(\dot{\mathbf{x}},\dot{\mathbf{v}})$. Recall that the transport of f_0 as a scalar is given by $f(\mathbf{x},\mathbf{v},t)=f_0\circ\psi_t^{-1}(\mathbf{x},\mathbf{v})$, which satisfies

$$\frac{\partial f}{\partial t}+\mathbf{u}\cdot\nabla_{\mathbf{z}}\,f=0, \tag{4.2}$$

where $\nabla_{\mathbf{z}}=(\nabla_{\mathbf{x}},\nabla_{\mathbf{v}})$ is the six-dimensional gradient operator in $(\mathbf{x},\mathbf{v})$ space. Let J_ψ be the Jacobian determinant of the mapping $\psi\in\mathrm{Diff}(T\mathbb{R}^3)$, that is, the determinant of the Jacobian matrix $\partial(\mathbf{x},\mathbf{v})/\partial(\mathbf{x}_0,\mathbf{v}_0)$.

Define $F(\mathbf{x},\mathbf{v},t)$ to be f_0, transported as a *density*:

$$F(\mathbf{x}(\mathbf{x}_0,\mathbf{v}_0),\mathbf{v}(\mathbf{x}_0,\mathbf{v}_0),t)J_\psi(\mathbf{x}_0,\mathbf{v}_0)=f_0(\mathbf{x}_0,\mathbf{v}_0),$$

so that

$$\frac{\partial F}{\partial t}+\nabla_{\mathbf{z}}\cdot(F\mathbf{u})=0. \tag{4.3}$$

Taking variations in our Lagrangian (4.1) and making use of the preceding equation for F, we obtain the following equations (taking $\varphi_0=1$):

$$\begin{aligned}
&\delta\mathbf{x}:\ \ m\ddot{\mathbf{x}}+m(\ddot{\mathbf{x}}-\dot{\mathbf{v}})=-q\nabla_{\mathbf{x}}\Phi-q\frac{\partial\mathbf{A}}{\partial t}+q\dot{\mathbf{x}}\times(\nabla_{\mathbf{x}}\times\mathbf{A}),\\
&\delta\mathbf{v}:\ \ \dot{\mathbf{x}}-\mathbf{v}=0,\\
&\delta\Phi:\ \ \nabla_{\mathbf{x}}\cdot\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)=-q\int d\mathbf{v}\;F(\mathbf{x},\mathbf{v},t),\\
&\delta\mathbf{A}:\ \ \nabla_{\mathbf{x}}\times(\nabla_{\mathbf{x}}\times\mathbf{A})=-\frac{\partial}{\partial t}\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)+q\int d\mathbf{v}\;\mathbf{v}F(\mathbf{x},\mathbf{v},t).
\end{aligned} \tag{4.4}$$

The second equation in (4.4) treats the Eulerian velocity $\mathbf{v}$ as a Lagrange multiplier, and ties its value to the fluid velocity $\dot{\mathbf{x}}$, hence $\dot{\mathbf{v}}=\ddot{\mathbf{x}}$ as well. The first two variational equations in the set (4.4) provide the desired relation for particle acceleration and the last two equations are the Maxwell equations with source terms. Thus, Hamilton's principle with Low's action provides the equations for self-consistent particle motion in an electromagnetic field, as required, and the description is completed by substituting

$$\left(\mathbf{v}, -\frac{q}{m}\left[\nabla_{\mathbf{x}}\Phi + \frac{\partial \mathbf{A}}{\partial t} - \mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right]\right)$$

for the components of $\mathbf{u}$ in the transport equation (4.2) to give the Vlasov equation (2.4).

V. THE MAXWELL–VLASOV SYSTEM AS EULER–POINCARÉ EQUATIONS

We will now specialize the general Euler–Poincaré theorem to the case of plasmas. The Lagrangian $L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}})$ in Eq. (4.1) has a right Diff$(T\mathbb{R}^3)$-symmetry. Let $\eta \in \mathrm{Diff}(T\mathbb{R}^3)$, $F \in \mathscr{F}$, and define the action of η on F by $F\eta = (F\circ\eta)J_\eta$ where, as above, J_η is the Jacobian determinant of η.

The symmetry of $L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}})$ is the property

$$L_{f_0\eta,\varphi_0}(\psi\eta,\dot{\psi}\eta,\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}}) = L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}}),$$

for all $\eta \in \mathrm{Diff}(T\mathbb{R}^3)$.

A. Ingredients for Euler–Poincaré

Now we apply the general Euler–Poincaré Theorem 3.1, taking $G = \mathrm{Diff}(T\mathbb{R}^3)$ and $\mathscr{Q} = \mathscr{V}\times\mathscr{A}$ and the parameter $a_0 = f_0$. As we have explained before, φ_0 is an auxiliary quantity that will ultimately take the value unity. In the general Euler–Poincaré Theorem 3.1 we take

$$\delta\mathbf{u} = \frac{\partial \mathbf{w}}{\partial t} - \mathrm{ad}_{\mathbf{u}}\mathbf{w}, \quad \delta a = -\pounds_{\mathbf{w}} a, \tag{5.1}$$

where $\mathbf{w} \in \mathfrak{g}$ is a vector field on $T\mathbb{R}^3$, $\pounds_{\mathbf{w}}$ is the Lie derivative, and $\mathrm{ad}_{\mathbf{u}}\mathbf{w} = -[\mathbf{u},\mathbf{w}]$ defines $\mathrm{ad}_{\mathbf{u}}\mathbf{w}$ in terms of the Lie bracket of vector fields, $[\mathbf{u},\mathbf{w}]$. The Euler–Poincaré equations (3.4) are

$$\frac{\partial}{\partial t}\frac{\delta l}{\delta \mathbf{u}} = -a\,d^*_{\mathbf{u}}\frac{\delta l}{\delta \mathbf{u}} + \frac{\delta l}{\delta a}\diamond a, \tag{5.2}$$

where $\mathrm{ad}^*_{\mathbf{u}}$ is the dual of $\mathrm{ad}_{\mathbf{u}}$ and $\delta l/\delta \mathbf{u}$ is a one-form density. The one-form density $(\delta l/\delta a)\diamond a$ is defined by

$$\left\langle \frac{\delta l}{\delta a}\diamond a, \mathbf{w}\right\rangle = -\int \frac{\delta l}{\delta a}\cdot \pounds_{\mathbf{w}} a. \tag{5.3}$$

When the quantities a are tensor fields, $\delta l/\delta a$ will be elements of the dual space under the natural pairing.

We shall apply this result to obtain the Maxwell–Vlasov system (2.4)–(2.7) as Euler–Poincaré equations. We begin by recording a formula that will be needed later. Let $\mathbf{u},\mathbf{w}$ be two elements of $\mathfrak{g}$, the Lie algebra of vector fields for the diffeomorphism group on a manifold $\mathscr{M}$. Choose the one-form density $\mathbf{c} \in \mathfrak{g}^*$, and let the pairing $\langle \mathbf{c},\mathbf{u}\rangle : \mathfrak{g}^*\times\mathfrak{g}\to\mathbb{R}$ be given by

$$\langle \mathbf{c},\mathbf{u}\rangle = \int_{\mathscr{M}} d\mathbf{z}\ \mathbf{c}\cdot\mathbf{u} = \int_{\mathscr{M}} d\mathbf{z}\ c_j u^j, \tag{5.4}$$

where c_j and u^j, $j = 1,\ldots,n$, are components of $\mathbf{c}$ and $\mathbf{u}$ in $\mathbb{R}^n$ and $d\mathbf{z}$ is the volume form on $\mathscr{M}$. Then we can write the desired formula,

$$\begin{aligned}\langle \mathrm{ad}^*_{\mathbf{u}}\mathbf{c},\mathbf{w}\rangle &= \int d\mathbf{z}\ \mathrm{ad}^*_{\mathbf{u}}\mathbf{c}\cdot\mathbf{w}\\ &= \int d\mathbf{z}\ \mathbf{c}\cdot\mathrm{ad}_{\mathbf{u}}\mathbf{w}\\ &= -\int d\mathbf{z}\ c_i\left(u^j\frac{\partial w^i}{\partial z^j}-w^j\frac{\partial u^i}{\partial z^j}\right)\\ &= \int d\mathbf{z}\ w^i\left(c_j\frac{\partial u^j}{\partial z^i}+c_i(\nabla\cdot\mathbf{u})+(\mathbf{u}\cdot\nabla)c_i\right)=\langle \pounds_{\mathbf{u}}\mathbf{c},\mathbf{w}\rangle. \end{aligned} \tag{5.5}$$

Here $\pounds_{\mathbf{u}}\mathbf{c}$ is the Lie derivative of the the one-form density $\mathbf{c}$ with respect to the vector field $\mathbf{u}$, z^j is the coordinate chart, and c_j, u^j, w^j are the components of vectors in $\mathbb{R}^n$. Unless otherwise stated, we sum repeated indices over their range, $i,j=1,\ldots,n$, where n is the dimension of $\mathscr{M}$. We assume that the vector fields and one-form densities are defined so that integration by parts gives no contribution at the boundary (inclusion of nonzero boundary terms is straightforward). Formula (5.5) for $\mathrm{ad}^*_{\mathbf{u}}\mathbf{c}$ will be useful later.

By definition, $\mathbf{u}=(\dot{\mathbf{x}},\dot{\mathbf{v}})$; we will denote $\mathbf{u}_s=\dot{\mathbf{x}}$, the spatial part of the phase space velocity field.

B. The reduced action

We may transform the action (4.1) into the Eulerian description as the reduced action

$$\begin{aligned}\mathfrak{S}_{\mathrm{red}} &= \int dt\ l(\mathbf{u},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}})\\ &= \int dt\ \int d\mathbf{x}\ d\mathbf{v}\ F(\mathbf{x},\mathbf{v},t)\left(\frac{1}{2}m|\mathbf{u}_s|^2+\tfrac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2-q\Phi+q\mathbf{u}_s\cdot\mathbf{A}\right)\\ &\quad+\frac{1}{2}\int dt\int d\mathbf{x}\left|\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right|^2-|\nabla_{\mathbf{x}}\times\mathbf{A}|^2. \end{aligned} \tag{5.6}$$

We vary this action with respect to $\mathbf{u}_s$, F, Φ and $\mathbf{A}$:

$$\begin{aligned}\delta\mathfrak{S}_{\mathrm{red}} &= \int dt\int d\mathbf{x}\ d\mathbf{v}\{F\,[(m\mathbf{u}_s+m(\mathbf{u}_s-\mathbf{v})+q\mathbf{A})\cdot\delta\mathbf{u}_s-q\,\delta\Phi+\mathbf{u}_s\cdot\delta\mathbf{A}]\\ &\quad+\delta F[\tfrac{1}{2}m|\mathbf{u}_s|^2+\tfrac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2-q\Phi+q\mathbf{u}_s\cdot\mathbf{A}]\}\\ &\quad+\int dt\int d\mathbf{x}\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)\cdot\left(\nabla_{\mathbf{x}}\delta\Phi+\delta\frac{\partial\mathbf{A}}{\partial t}\right)-(\nabla_{\mathbf{x}}\times\mathbf{A})\cdot(\nabla_{\mathbf{x}}\times\delta\mathbf{A}). \end{aligned} \tag{5.7}$$

Stationary variations in Φ and $\mathbf{A}$ yield

$$\begin{aligned}\nabla_{\mathbf{x}}\cdot\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)&=-q\int d\mathbf{v}\ F(\mathbf{x},\mathbf{v},t),\\ \nabla_{\mathbf{x}}\times(\nabla_{\mathbf{x}}\times\mathbf{A})&=-\frac{\partial}{\partial t}\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)+q\int d\mathbf{v}\ F(\mathbf{x},\mathbf{v},t)\mathbf{u}_s. \end{aligned} \tag{5.8}$$

Thus, Maxwell's equations for the electromagnetic field of the plasma are recovered by requiring $\delta l=0$ for all variations of the field potentials Φ and $\mathbf{A}$. To continue toward the Euler–Poincaré form of the Maxwell–Vlasov equations, one must determine the forms of the variations $\delta\mathbf{u}_s$ and δF in (5.7).

According to the general theory, variations in the particle evolution map ψ lead to variations in the phase space velocity $\delta\mathbf{u}$ of the form

$$\delta\mathbf{u}=\frac{\partial\mathbf{w}}{\partial t}+[\mathbf{u},\mathbf{w}]\equiv\frac{\partial\mathbf{w}}{\partial t}-\mathrm{ad}_{\mathbf{u}}\mathbf{w}. \tag{5.9}$$

This Euler–Poincaré form of the variations may also be verified by a direct tensorial calculation, which is given in Ref. 6. The spatial part of this equation gives the variation of the spatial part of the field $\mathbf{u}$.

Variations of the field ψ also induce variations of the density F, in the same way as the parameter variations are induced in the general theory for the Euler–Poincaré equations [see Eq. (5.1)]. Either from that equation, or by direct calculations, these variations are computed to be

$$\delta F=-\nabla_{\mathbf{z}}\cdot(F\mathbf{w}), \tag{5.10}$$

which is equivalent to the formula

$$\delta(F\ d\mathbf{x}\ d\mathbf{v})=-\pounds_{\mathbf{w}}(F\ d\mathbf{x}\ d\mathbf{v}).$$

C. Computation of the variations

With these formulas for $\delta\mathbf{u}$ and δF in place, we compute

$$\begin{aligned}\delta\mathfrak{S}_{\mathrm{red}}=\int dt\int d\mathbf{x}\ d\mathbf{v}\ F\Big[&(m\mathbf{u}_s+m(\mathbf{u}_s-\mathbf{v})+q\mathbf{A})\cdot\left(\frac{\partial}{\partial t}\mathbf{w}+[\mathbf{u},\mathbf{w}]\right)\Big]\\&-\nabla_{\mathbf{z}}\cdot(F\mathbf{w})\left(\frac{1}{2}(m|\mathbf{u}_s|^2+m|\mathbf{u}_s-\mathbf{v}|^2)+q\mathbf{u}_s\cdot\mathbf{A}-q\Phi\right).\end{aligned} \tag{5.11}$$

Integrating by parts and dropping boundary terms gives

$$\begin{aligned}\delta\mathfrak{S}_{\mathrm{red}}=\int dt\int d\mathbf{x}\ d\mathbf{v}\ \mathbf{w}\cdot\Big[&-\frac{\partial}{\partial t}\left(Fm\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)\right)-\mathrm{ad}^*_{\mathbf{u}}\left(Fm\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)\right)\\&+F\nabla_{\mathbf{z}}\left(\frac{1}{2}m|\mathbf{u}_s|^2+\tfrac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2+q\mathbf{u}_s\cdot\mathbf{A}-q\Phi\right)\Big].\end{aligned} \tag{5.12}$$

Expanding the ad* term using formula (5.5) results in

$$\begin{aligned}\delta\mathfrak{S}_{\mathrm{red}}=\int dt\int d\mathbf{x}\ d\mathbf{v}\ \mathbf{w}\cdot\Big[&-\frac{\partial F}{\partial t}m\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)-Fm\frac{\partial}{\partial t}\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)\\&-(\mathbf{u}\cdot\nabla_{\mathbf{z}})\left(Fm\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)\right)-Fm\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)(\nabla_{\mathbf{z}}\cdot\mathbf{u})\\&-\left(Fm\left(\mathbf{u}_{sj}+(\mathrm{u}_{sj}-\mathrm{v}_j)+\frac{q}{m}\mathrm{A}_j\right)\right)\nabla_{\mathbf{z}}u^j+F\nabla_{\mathbf{z}}\left(\tfrac{1}{2}m|\mathbf{u}_s|^2+\tfrac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2+q\mathbf{u}_s\cdot\mathbf{A}-q\Phi\right)\Big].\end{aligned} \tag{5.13}$$

We expand the products to obtain

$$\begin{aligned}\delta\mathfrak{S}_{\mathrm{red}}=\int dt\int d\mathbf{x}\ d\mathbf{v}\ \mathbf{w}\cdot\Big\{&-m\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)\left(\frac{\partial F}{\partial t}+\mathbf{u}\cdot\nabla_{\mathbf{z}}F\right)\\&-Fm\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)(\nabla_{\mathbf{z}}\cdot\mathbf{u})-Fm\left[\left(\frac{\partial}{\partial t}+(\mathbf{u}\cdot\nabla_{\mathbf{z}})\right)\left(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v})+\frac{q}{m}\mathbf{A}\right)+\frac{q}{m}\nabla_{\mathbf{z}}\Phi\right]\\&-Fm\left(\mathrm{u}_{sj}+(\mathrm{u}_{sj}-\mathrm{v}_j)+\frac{q}{m}\mathrm{A}^j\right)\nabla_{\mathbf{z}}u^j+Fm\mathrm{u}_{sj}\nabla_{\mathbf{z}}\mathrm{u}_s^j+Fq\mathrm{A}^j\nabla_{\mathbf{z}}\mathrm{u}_{sj}\\&+Fm(\mathrm{u}_{sj}-\mathrm{v}_j)\nabla_{\mathbf{z}}(\mathrm{u}_s^j-\mathrm{v}_s^j)+Fq\mathrm{u}_{sj}\nabla_{\mathbf{z}}\mathrm{A}^j\Big\}.\end{aligned} \tag{5.14}$$

Consider the last two lines of Eq. (5.14). Upon writing $\mathbf{w}=(\mathbf{w}_1,\mathbf{w}_2)$, where $\mathbf{w}_1,\mathbf{w}_2\in\mathbb{R}^3$, these lines reduce to

$$\begin{aligned}
&-Fm(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v}))\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{u}+Fm(\mathbf{u}_s+(\mathbf{u}_s-\mathbf{v}))\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{u}_s\\
&\quad -Fq\mathbf{A}\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{u}+Fq\mathbf{A}\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{u}_s+Fq\mathbf{u}_s\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{A}\\
&\quad -Fm(\mathbf{u}_s-\mathbf{v})\cdot(\mathbf{w}_1\cdot\nabla_{\mathbf{x}}+\mathbf{w}_2\cdot\nabla_{\mathbf{v}})\mathbf{v} \;=\; Fq\mathbf{u}_s(\mathbf{w}_1\cdot\nabla_{\mathbf{x}})\mathbf{A}-Fm(\mathbf{u}_s-\mathbf{v})\cdot\mathbf{w}_2. \qquad (5.15)
\end{aligned}$$

The first three lines cancel to zero because they only involve spatial velocity projections, where $\mathbf{u}=\mathbf{u}_s$. The last line follows upon using $\nabla_{\mathbf{x}}\mathbf{v}=0$ and $\nabla_{\mathbf{v}}\mathbf{A}=0$, which hold, respectively, because $\mathbf{v}$ is an independent coordinate and $\mathbf{A}$ is a function of space alone. Similarly, and under the additional observation that $\nabla_{\mathbf{z}}\Phi=(\nabla_{\mathbf{x}}\Phi,0)$ because the potential Φ also does not depend on velocity, the other three lines of Eq. (5.14) are purely spatial, i.e., the projection onto the last three coordinates would give zero, and hence the contribution to the variation of the action $\delta\mathfrak{S}_{\text{red}}$ from $\mathbf{w}_2$ comes only from the calculation in Eq. (5.15). Stationarity of the action under the velocity components of the variation, $\mathbf{w}_2$, then implies

$$Fm(\mathbf{u}_s-\mathbf{v})=0, \quad \text{i.e.,} \quad \mathbf{u}_s=\mathbf{v}. \qquad (5.16)$$

Consequently, in Eq. (5.14) we can write $\mathbf{u}$ as $(\mathbf{v},\mathbf{a})$ where $\mathbf{a}$ is yet to be determined, and we can also replace $\mathbf{u}_s-\mathbf{v}$ with zero. On doing this, the contribution to the variation of the action from $\mathbf{w}_1$ becomes

$$\begin{aligned}
\delta\mathfrak{S}_{\text{red}}=\int dt\int d\mathbf{x}\; d\mathbf{v}\;\mathbf{w}_1\cdot\Bigg[&-(m\mathbf{v}+q\mathbf{A})\left(\frac{\partial F}{\partial t}+\nabla_{\mathbf{z}}\cdot(F\mathbf{u})\right)\\
&-F\left(m\frac{\partial\mathbf{v}}{\partial t}+m(\mathbf{v}\cdot\nabla_{\mathbf{x}})\mathbf{v}+m(\mathbf{a}\cdot\nabla_{\mathbf{v}})\mathbf{v}+q\frac{\partial\mathbf{A}}{\partial t}+q\nabla_{\mathbf{x}}\Phi+q\mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right)\Bigg]. \qquad (5.17)
\end{aligned}$$

Here, we have used standard vector identities in obtaining the result

$$\mathbf{w}\cdot qF(\mathrm{u}_{sj}\nabla_{\mathbf{z}}\mathrm{A}_s^j-(\mathbf{u}\cdot\nabla_{\mathbf{z}})\mathbf{A})=qF\mathbf{w}_1\cdot(\mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})). \qquad (5.18)$$

Referring to the continuity equation (4.3) for F and using the identities $\partial\mathbf{v}/\partial t=0$ and $\nabla_{\mathbf{x}}\mathbf{v}=0$ reduces Eq. (5.17) to

$$\delta\mathfrak{S}_{\text{red}}=-\int dt\int d\mathbf{x}\; d\mathbf{v}\;\mathbf{w}_1\cdot F\left(m\mathbf{a}+q\nabla_{\mathbf{x}}\Phi+q\frac{\partial\mathbf{A}}{\partial t}-q\mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right).$$

Therefore, $\delta\mathfrak{S}_{\text{red}}=0$ implies that

$$m\mathbf{a}=-q\nabla_{\mathbf{x}}\Phi-q\frac{\partial\mathbf{A}}{\partial t}+q\mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A}). \qquad (5.19)$$

Now consider what the invariance of the Boltzmann function f implies. By Eq. (4.2) and substitution for $\mathbf{u}=(\mathbf{v},\mathbf{a})$ we obtain

$$\frac{\partial f}{\partial t}+\mathbf{v}\cdot\nabla_{\mathbf{x}}f-\frac{q}{m}\left[\left(\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right)-\mathbf{v}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\right]\cdot\nabla_{\mathbf{v}}f=0, \qquad (5.20)$$

and so, along with Eqs. (5.8), we have recovered the full Maxwell–Vlasov system from stationarity of the action (5.6) entirely in the Eulerian description.

VI. THE GENERALIZED LEGENDRE TRANSFORMATION

A. Introduction

Before passing to the Hamiltonian description of the Maxwell–Vlasov equations, we pause to explain the theoretical background of how one does this when there are degeneracies. This section can be skipped if one is willing to simply take on faith that one should *do the Legendre transformation slowly and carefully* when there are degeneracies.

As explained in Ref. 3, one normally thinks of passing from Euler–Poincaré equations on a Lie algebra $\mathfrak{g}$ to Lie–Poisson equations on the dual $\mathfrak{g}^*$ by means of the Legendre transformation. In some situations involving the Euler–Poincaré equations, one starts with a Lagrangian on $\mathfrak{g}\times V^*$ and performs a *partial* Legendre transformation, in the variable ξ only, by writing

$$\mu=\frac{\delta l}{\delta \xi}, \quad h(\mu,a)=\langle \mu,\xi\rangle - l(\xi,a). \tag{6.1}$$

Since

$$\frac{\delta h}{\delta \mu}=\xi+\left\langle \mu,\frac{\delta \xi}{\delta \mu}\right\rangle-\left\langle \frac{\delta l}{\delta \xi},\frac{\delta \xi}{\delta \mu}\right\rangle=\xi, \tag{6.2}$$

and $\delta h/\delta a=-\delta l/\delta a$, we see that the Euler–Poincaré equations (3.4) for $\xi\in\mathfrak{g}$ and $\dot{a}(t)=-a(t)\xi(t)$ imply the Hamiltonian semidirect-product Lie–Poisson equations for $\mu\in\mathfrak{g}^*$. Namely,

$$\frac{\partial}{\partial t}\mu=-\mathrm{ad}^*_{(\delta h/\delta \mu)}\mu-\frac{\delta h}{\delta a}\diamond a=\{\mu,h\}_{LP}, \quad \frac{\partial}{\partial t}a=-a\,\frac{\delta h}{\delta \mu}=\{a,h\}_{LP}, \tag{6.3}$$

with $(+)$ Lie–Poisson bracket on $\mathfrak{g}^*\times V^*$ given by

$$\{g,h\}_{LP}=-\left\langle \mu,\mathrm{ad}_{(\delta h/\delta \mu)}\frac{\delta g}{\delta \mu}\right\rangle+\left\langle a,\frac{\delta g}{\delta a}\frac{\delta h}{\delta \mu}-\frac{\delta h}{\delta a}\frac{\delta g}{\delta \mu}\right\rangle. \tag{6.4}$$

If the Legendre transformation (6.1) is invertible, then one can also pass Lie–Poisson equations to the Euler–Poincaré equations together with the equations $\dot{a}(t)=-a(t)\xi(t)$.

It is important in this paper to give a detailed explanation that incorporates the degeneracy of the parameter-dependent system together with the role of symmetry. Unlike the examples considered in Ref. 6 such as compressible flow or MHD, in the case of the Maxwell–Vlasov system or even the Vlasov–Poisson system, the Lagrangian L_{a_0} corresponding to the action in Eq. (5.6) is *degenerate*, since it does not depend on the variables $\dot{\Phi}$ and $\dot{\mathbf{v}}$. In other words, *the degeneracy and corresponding constraints that appear in Vlasov plasmas are more serious than for fluids or the heavy top, etc*. To deal with this degeneracy, we shall use the generalized Legendre transformation in the context of Lagrangian submanifolds, as described in Ref. 32. This is also related to the Dirac theory of constraints (see Ref. 33). In particular, we shall take special care to ensure that the Hamiltonian formulation of the Maxwell–Vlasov system preserves the constraints associated with the degeneracy of its Lagrangian.

B. The general construction

Let Q be a manifold and $\pi:T^*Q\to Q$ be the cotangent bundle of Q. Then TT^*Q is a symplectic manifold with a symplectic form that can be written in two distinct ways as the exterior derivative of two intrinsic one-forms. These two one-forms are denoted λ and χ and are given in coordinates by

$$\lambda=\dot{p}\,dq+p\,d\dot{q} \tag{6.5}$$

and

$$\chi=\dot{p}\,dq-\dot{q}\,dp, \tag{6.6}$$

where (q,p) are coordinates for T^*Q and $(q,p,\dot{q},\dot{p})$ are the corresponding coordinates for TT^*Q. For the intrinsic definitions of these forms, see Ref. 33.

Let $L:J\to\mathbb{R}$ be a Lagrangian defined on a submanifold $J\subset TQ$ called the ***Lagrangian constraint***. The Legendre transformation is a procedure to obtain a Hamiltonian $H:K\to\mathbb{R}$ defined on a submanifold $K\subset T^*Q$, called the ***Hamiltonian constraint***. The Euler–Lagrange equations are

$$\lambda = dL \quad \text{on } J, \tag{6.7}$$

while the Hamilton equations are

$$\chi = -dH \quad \text{on } K. \tag{6.8}$$

The abbreviated expressions (6.7) and (6.8) stand for

$$\lambda = d(L\circ T\pi) \quad \text{on} \quad (T\pi)^{-1}(J) \tag{6.9}$$

and

$$\chi = -d(H\circ\tau^{-1}) \quad \text{on} \quad (\tau)^{-1}(K), \tag{6.10}$$

where τ is the canonical projection $\tau:TT^*Q\to T^*Q$, given in coordinates by $\tau(q,p,\dot{q},\dot{p}) = (q,p)$. The map $T\pi$ is given by $T\pi(q,p,\dot{q},\dot{p}) = (q,\dot{q})$.

Both the Euler–Lagrange and Hamilton equations define the same Lagrangian submanifold D of TT^*Q. The Lagrangian and Hamiltonian L and H are the generating functions with respect to the one-forms λ and χ, respectively.

The ***generalized Legendre transformation*** consists of the following steps:

Step 1: For each $(q,p)\in T^*Q$ define

$$K(q,p) = \left\{ (q,\dot{q}) \in T_qQ \,\middle|\, \frac{\partial}{\partial\dot{q}}(p\dot{q} - L(q,\dot{q})) = 0 \right\}, \tag{6.11}$$

and let

$$K = \{(q,p)\in T^*Q \,|\, K(q,p)\neq\varnothing\}. \tag{6.12}$$

Assumption: Assume that for each $(q,p)\in K$, the submanifold $K(q,p)$ is connected. This implies that the stationary value

$$\mathrm{stat}_{\dot{q}}(p\dot{q} - L(q,\dot{q})) \tag{6.13}$$

of $p\dot{q} - L(q,\dot{q})$ on $K(q,p)$ is uniquely defined; that is, it does not depend on $\dot{q}$.

Step 2: Define $H:K\to\mathbb{R}$ as follows:

$$H(q,p) = \mathrm{stat}_{\dot{q}}(p\dot{q} - L(q,\dot{q})). \tag{6.14}$$

C. The generalized Legendre transformation with parameters and symmetry

Now we adapt this methodology to the case of parameter-dependent Lagrangians with symmetry. Let $L_{a_0}:TG\times T\mathcal{Q}\to\mathbb{R}$ be a Lagrangian depending on a parameter $a_0\in V^*$. Assume that G acts on V^* on the right and denote by ag the action of $g\in G$ on $a\in V^*$. Assume also the following invariance property:

$$L_{ah}(gh,\dot{g}h,q,\dot{q}) = L_a(g,\dot{g},q,\dot{q}), \tag{6.15}$$

for all $g,h\in G$, all $(q,\dot{q})\in T\mathcal{Q}$, and all $a\in V^*$. A typical element of $T^*G\times T^*\mathcal{Q}$ will be denoted (g,α_g,q,ν_q) or simply (g,α,q,ν). For each $a_0\in V^*$ and $(g,\alpha)\in T^*G$, define

$$K_{a_0}(g,\alpha,q,\nu)=\left\{(g,\dot{g},q,\dot{q})\left|\frac{\partial}{\partial\dot{g}}(\alpha\dot{g}+\nu\dot{q}-L_{a_0}(g,\dot{g},q,\dot{q}))=0\right.\right.$$
$$\left.\left.\text{and }\frac{\partial}{\partial\dot{q}}(\alpha\dot{g}+\nu\dot{q}-L_{a_0}(g,\dot{g},q,\dot{q}))=0\right\}.\tag{6.16}$$

One can immediately check for any $a_0\in V^*$, $h\in G$, and $(g,\alpha,q,\nu)\in T^*G\times T^*\mathcal{Q}$ that $K_{a_0h}(gh,\alpha h,q,\nu)=K_{a_0}(g,\alpha,q,\nu)h$. Define

$$K_{a_0}=\{(g,\alpha,q,\nu)|K_{a_0}(g,\alpha,q,\nu)\neq\varnothing\}.\tag{6.17}$$

Then one can easily prove for any $h\in G$ that $K_{a_0h}=K_{a_0}h$. Define

$$K=\{(g,\alpha,q,\nu,a)|K_a(g,\alpha,q,\nu)\neq\varnothing\}.\tag{6.18}$$

Then $K\subset T^*G\times T^*\mathcal{Q}\times V^*$ is an invariant subset under the action of G given by $(g,\alpha,q,\nu,a)h=(gh,\alpha h,q,\nu,ah)$. Now for each $a_0\in V^*$ we define $H_{a_0}:K_{a_0}\to\mathbb{R}$ by

$$H_{a_0}(g,\alpha,q,\nu)=\alpha\dot{g}+\nu\dot{q}-L_{a_0}(g,\dot{g},q,\dot{q}),\tag{6.19}$$

for any $(g,\dot{g},q,\dot{q})\in K_{a_0}(g,\alpha,q,\nu)$. Then, according to the general theory explained above, Hamilton's equations are, for each $a_0\in V^*$, $-dH_{a_0}=\chi$ on K_{a_0}, where

$$\chi=\dot{\alpha}\,dg-\dot{g}\,d\alpha+\dot{\nu}\,dq-\dot{q}\,d\nu.\tag{6.20}$$

One can also easily prove, using the previous equalities, that $H_{a_0}(g,\alpha,q,\nu)$ has the following invariance property:

$$H_{a_0h}(gh,\alpha h,q,\nu)=H_{a_0}(g,\alpha,q,\nu).\tag{6.21}$$

Let $\mathfrak{s}^*$ be the dual of the semidirect product Lie algebra $\mathfrak{s}=\mathfrak{g}\circledS V$. Then define $\mathcal{K}\subset\mathfrak{s}^*\times T^*\mathcal{Q}$ by

$$\mathcal{K}=\{(\alpha,q,\nu,a)\in\mathfrak{s}^*\times T^*\mathcal{Q}|(e,\alpha,q,\nu,a)\in K\},$$

and the Hamiltonian $h_{\mathcal{K}}:\mathcal{K}\to\mathbb{R}$ by $h_{\mathcal{K}}(\alpha,a,q,\nu)=H_a(e,\alpha,q,\nu)$. Thus, $h_{\mathcal{K}}$ is the restriction to $\mathcal{K}\subset\mathfrak{s}^*$ of the right invariant Hamiltonian $H:K\to\mathbb{R}$ given by $H(g,\alpha,q,\nu,a)=H_a(g,\alpha,q,\nu)$. Then, by a natural generalization of semidirect product theory to include constrained Hamiltonian systems, we have that Hamilton's equations on $\mathcal{K}\subset\mathfrak{s}^*$ generated by $h_{\mathcal{K}}$ give the evolution of the system on $\mathcal{K}$ determined by the Poisson–Hamilton equations $\dot{f}=\{f,h_{\mathcal{K}}\}$ on the Poisson submanifold $\mathcal{K}\subset\mathfrak{s}^*\times T^*\mathcal{Q}$, where the Poisson structure is defined in a natural way. More precisely, we have the Dirac brackets on K (see, for instance, Ref. 34 or 3) which, by reduction, give the brackets on $\mathcal{K}$. This is the abstract procedure underlying the computations we do in the specific case of plasmas given in the next section.

VII. HAMILTONIAN FORMULATION

We now pass to the corresponding Hamiltonian formulation of the Maxwell–Vlasov system (2.4) and (2.5) in the Eulerian description by taking the Legendre transform of the reduced action (5.6).

A. The role of the general theory

From the geometrical point of view, we simply apply the generalized Legendre transformation described abstractly in Sec. VI to the degenerate Lagrangian

$$L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}}).$$

This Lagrangian is degenerate because it does not depend on the variables $\dot{\Phi}$ and $\dot{\mathbf{v}}$. The theory described in Sec. VI may be applied to this action on $T(\mathscr{F}\times\mathscr{V}\times\mathscr{A})$. The action of the group $\mathrm{Diff}(T\mathbb{R}^3)$ on the factor $\mathscr{F}$ for this Lagrangian is given as before, while the actions on the factors $\mathscr{V}$ and $\mathscr{A}$ are trivial. It is easy to see that the Hamiltonian constraint for each f_0 is $K_{f_0}\subset T^*(\mathrm{Diff}(T\mathbb{R}^3)\times\mathscr{V}\times\mathscr{A})$, defined by the conditions

$$\Psi=\frac{\delta L}{\delta\dot{\Phi}}=0 \quad \text{and} \quad \mathbf{m}_v=\frac{\delta L}{\delta\dot{\mathbf{v}}}=0.$$

These conditions impose constraints, which for consistency must be dynamically preserved.

B. Calculation of the transformed equations

We will perform the calculations in detail, working with the reduced Lagrangian rather than the Lagrangian

$$L_{f_0,\varphi_0}(\psi,\dot{\psi},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}})$$

and setting $\varphi_0=1$ as usual.

We start with the action (5.6) for the Maxwell–Vlasov system in the Eulerian description,

$$\mathfrak{S}_{\mathrm{red}}(\mathbf{u},\Phi,\dot{\Phi},\mathbf{A},\dot{\mathbf{A}})=\int dt\int d\mathbf{x}\,d\mathbf{v}F(\mathbf{x},\mathbf{v},t)\left(\frac{1}{2}m|\mathbf{u}_s|^2+\frac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2-q\Phi+q\mathbf{u}_s\cdot\mathbf{A}\right)$$
$$+\frac{1}{2}\int dt\int d\mathbf{x}\left|\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}\right|^2-|\nabla_{\mathbf{x}}\times\mathbf{A}|^2. \tag{7.1}$$

This leads immediately to

$$\frac{\delta l}{\delta\dot{\mathbf{A}}}=\nabla_{\mathbf{x}}\Phi+\frac{\partial\mathbf{A}}{\partial t}=-\mathbf{E}, \tag{7.2}$$

and so (minus) the electric field variable $\mathbf{E}$ is the field momentum density canonically conjugate to the magnetic potential. Let us define the material momentum density in six dimensions,

$$\mathbf{m}\equiv\frac{\delta l}{\delta\mathbf{u}}. \tag{7.3}$$

We write $\mathbf{m}=(\mathbf{m}_s,\mathbf{m}_v)$, where $\mathbf{m}_s$ is the projection of $\mathbf{m}$ onto the first three coordinate positions, and $\mathbf{m}_v$ is the projection onto the last three places. We think of $\mathbf{m}_s$ and $\mathbf{m}_v$ also as vectors in six dimensions. From the Lagrangian we see that

$$\mathbf{m}_s=F(m\mathbf{u}_s+m(\mathbf{u}_s-\mathbf{v})+q\mathbf{A}) \quad \text{and} \quad \mathbf{m}_v=0. \tag{7.4}$$

Proceeding with the Legendre transform of our action (7.1) results in a corresponding (reduced) Hamiltonian function written in terms of the velocities,

$$h=\int d\mathbf{x}\,d\mathbf{v}\,F\left(m|\mathbf{u}_s|^2-\frac{1}{2}m|\mathbf{v}|^2+q\Phi\right)+\mathbf{m}_v\cdot\mathbf{a}+\frac{1}{2}\int d\mathbf{x}(|\mathbf{E}|^2+|\nabla_{\mathbf{x}}\times\mathbf{A}|^2+2\mathbf{E}\cdot\nabla_{\mathbf{x}}\Phi), \tag{7.5}$$

where $\mathbf{a}$ denotes the projection of $\mathbf{u}$ onto its last three entries. Transforming this to the momentum variables gives

$$h=\int d\mathbf{x}\; d\mathbf{v}\;\frac{1}{4Fm}|\mathbf{m}_s+mF\mathbf{v}-qF\mathbf{A}|^2-\frac{1}{2}mF|\mathbf{v}|^2+qF\Phi+\mathbf{m}_v\cdot\mathbf{a}$$
$$+\frac{1}{2}\int d\mathbf{x}\Big(|\mathbf{E}|^2+|\nabla_{\mathbf{x}}\times\mathbf{A}|^2+2\mathbf{E}\cdot\nabla_{\mathbf{x}}\Phi\Big). \tag{7.6}$$

The variation of this Hamiltonian with respect to **m**, **a**, **E**, **A**, F, and Φ is given by

$$\delta h=\int d\mathbf{x}\; d\mathbf{v}\Big[\mathbf{u}\cdot\delta\mathbf{m}+\mathbf{m}_v\cdot\delta\mathbf{a}-qF\mathbf{u}_s\cdot\delta\mathbf{A}+qF\,\delta\Phi-\Big(\frac{1}{2}m|\mathbf{u}_s|^2+\frac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2+q\mathbf{u}_s\cdot\mathbf{A}$$
$$-q\Phi\Big)\delta F\Big]+\int d\mathbf{x}(\mathbf{E}+\nabla_{\mathbf{x}}\Phi)\cdot\delta\mathbf{E}-(\nabla_{\mathbf{x}}\cdot\mathbf{E})\,\delta\Phi+\nabla_{\mathbf{x}}\times(\nabla_{\mathbf{x}}\times\mathbf{A})\cdot\delta\mathbf{A}. \tag{7.7}$$

This expression allows one to read off the evolution equations for the electromagnetic field:

$$\frac{\partial\mathbf{A}}{\partial t}=-\frac{\delta h}{\delta\mathbf{E}}=-\mathbf{E}-\nabla_{\mathbf{x}}\Phi,\quad \text{i.e.,}\quad \mathbf{E}=-\nabla_{\mathbf{x}}\Phi-\frac{\partial\mathbf{A}}{\partial t},$$

$$\frac{\delta h}{\delta\Phi}=0=-\nabla_{\mathbf{x}}\cdot\mathbf{E}+q\int d\mathbf{v}\;F,\quad \text{i.e.,}\quad \nabla_{\mathbf{x}}\cdot\mathbf{E}=q\int d\mathbf{v}F:=\rho, \tag{7.8}$$

$$\frac{\partial\mathbf{E}}{\partial t}=\frac{\delta h}{\delta\mathbf{A}}=\nabla_{\mathbf{x}}\times(\nabla_{\mathbf{x}}\times\mathbf{A})-q\int d\mathbf{v}\;F\mathbf{u}_s,\quad \text{i.e.,}\quad \frac{\partial\mathbf{E}}{\partial t}=\nabla_{\mathbf{x}}\times\mathbf{B}-\mathbf{j}.$$

Note that the constraint $\delta h/\delta\Phi=0$ (Gauss' law) arises from the absence of $\dot{\Phi}$ dependence in l.

The general theory of Sec. VI shows that F is an element of the second factor of the semi-direct product and so its evolution is given by Lie dragging as a density. Likewise, f is Lie dragged as a scalar and m_i satisfies a Lie–Poisson evolution equation:

$$\frac{\partial F}{\partial t}=-\nabla_{\mathbf{z}}\cdot(F\mathbf{u}),\quad \frac{\partial f}{\partial t}=-\mathbf{u}\cdot\nabla_{\mathbf{z}}\,f, \tag{7.9}$$

$$\frac{\partial m_i}{\partial t}=-\frac{\partial}{\partial z^j}m_iu^j-m_j\frac{\partial}{\partial z^i}u^j-F\frac{\partial}{\partial z^i}\frac{\delta h}{\delta F}.$$

The first two of these equations reflect the assumptions that were made in the definitions of f and F, while the last equation encodes the dynamics for the system. We first consider the case where the momentum component i takes the values 4,5,6. In this case,

$$-\frac{\partial m_i}{\partial t}=m_{sj}\frac{\partial}{\partial z^i}u^j+m_{vj}\frac{\partial}{\partial z^i}u^j-F\frac{\partial}{\partial z^i}\,(\tfrac{1}{2}m|\mathbf{u}_s|^2+\tfrac{1}{2}m|\mathbf{u}_s-\mathbf{v}|^2+q\mathbf{u}_s\cdot\mathbf{A}-q\Phi)$$
$$=m_{vj}\frac{\partial}{\partial z^i}u^j+Fm(u_{sj}-v_j)\,\frac{\partial u^j}{\partial z^i}-Fm(u_{sj}-v_j)\,\frac{\partial}{\partial z^i}(u_s^j-v_s^j)-qFu_{sj}\frac{\partial A_s^j}{\partial z^i}+Fq\frac{\partial}{\partial z^i}\Phi, \tag{7.10}$$

where $i=4,5,6$. In the second line of Eq. (7.10), we have substituted for $\mathbf{m}_s$ from Eq. (7.4) and rearranged terms. Here $\mathbf{m}_v=0$, because l does not depend on $\dot{\mathbf{v}}$. Setting $\mathbf{m}_v=0$ initially in Eq. (7.10) ensures that $\mathbf{m}_v\equiv0$ persists throughout the ensuing motion, for potentials Φ and **A** that are independent of **v** and provided the constraint holds that $\mathbf{u}_s=\mathbf{v}$, as in Eq. (5.16). Likewise, the Gauss' law constraint imposed by $\delta h/\delta\Phi=0$ also persists during the ensuing motion, as seen from the last equation of (7.8) and the first equation of (7.9), provided the constraint $\mathbf{u}_s=\mathbf{v}$ holds and F vanishes in the limit as $|\mathbf{v}|\to\infty$.

The spatial part of the evolution equation of **m** will produce the required single-particle dynamics. From Eq. (7.9), we have

$$\frac{\partial m_i}{\partial t} = -\frac{\partial}{\partial z^j} m_i u^j - m_j \frac{\partial}{\partial z^i} u^j - F \frac{\partial}{\partial z^i} \frac{\delta h}{\delta F}. \tag{7.11}$$

Setting $i=1,2,3$, in Eq. (7.11), then substituting for $\delta h/\delta F$ and using the relations

$$\mathbf{m}_s = F(m\mathbf{u}_s + m(\mathbf{u}_s - \mathbf{v}) + q\mathbf{A}), \quad \mathbf{m}_v \equiv 0,$$

and

$$\nabla_{\mathbf{v}}\Phi = 0, \quad \nabla_{\mathbf{v}}\mathbf{A} = 0,$$

yields the spatial components of the motion equation,

$$\frac{\partial m_{si}}{\partial t} = -\frac{\partial}{\partial z^j} m_{si} u^j - m_{sj} \frac{\partial}{\partial z^i} u^j - F \frac{\partial}{\partial z^i} \left(\frac{1}{2} m |\mathbf{u}_s|^2 + \frac{1}{2} m |\mathbf{u}_s - \mathbf{v}|^2 + q\mathbf{u}_s \cdot \mathbf{A} - q\Phi \right). \tag{7.12}$$

Substituting for $\mathbf{m}_s$ and then using the continuity relation $\partial F/\partial t + \nabla_{\mathbf{z}} \cdot (F\mathbf{u}) = 0$ gives

$$m \frac{\partial \mathrm{u}_{si}}{\partial t} + q \frac{\partial \mathrm{A}_{si}}{\partial t} = -u^j \frac{\partial}{\partial z^j} m\mathrm{u}_{si} - \mathrm{u}_s^j \frac{\partial}{\partial z^j} q\mathrm{A}_{si} - q\mathrm{A}_j \frac{\partial}{\partial z^i} u^j - q \frac{\partial \Phi}{\partial z^i} + q \frac{\partial}{\partial z^i} (\mathbf{u}_s \cdot \mathbf{A})$$
$$- \frac{1}{2} m \frac{\partial}{\partial z^i} |\mathbf{u}_s - \mathbf{v}|^2. \tag{7.13}$$

Rearranging this equation results in

$$m \left(\frac{\partial}{\partial t} + u^j \frac{\partial}{\partial z^j} \right) \mathbf{u}_s = q\mathbf{E} + q\mathbf{u}_s \times (\nabla_{\mathbf{x}} \times \mathbf{A}) - \frac{1}{2} m \frac{\partial}{\partial \mathbf{z}} |\mathbf{u}_s - \mathbf{v}|^2. \tag{7.14}$$

We may now evaluate this on the constraint set $\mathbf{u}_s = \mathbf{v}$ and thereby obtain the Lorentz force,

$$m\mathbf{a} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{7.15}$$

where $\mathbf{a}$ is the acceleration of a fluid parcel [the last three components in $\mathbf{u} = (\mathbf{v}, \mathbf{a})$]. As we have seen, in this Hamiltonian formulation of the Maxwell–Vlasov equations in the Eulerian description, the acceleration $\mathbf{a}$ in $\mathbf{u}$ is a vector Lagrange multiplier which imposes $\mathbf{m}_v = 0$. Equation (7.15) provides an expression for this Lagrange multiplier in terms of known dynamical variables and, as a consequence, we regain the equation for the acceleration of a charged particle in an electromagnetic field. The momentum constraint $\mathbf{m}_v = 0$ remains invariant when the electromagnetic potentials are independent of the phase space velocity coordinate $\mathbf{v}$ and the velocity constraint $\mathbf{u}_s = \mathbf{v}$ holds. Perhaps not unexpectedly, one finds that $\nabla_{\mathbf{z}} \cdot \mathbf{u} = 0$. Also, (minus) the electric field is canonically conjugate to the vector potential, and the electrostatic potential Φ plays the role of a Lagrange multiplier which imposes Gauss's law. Thus, our Hamiltonian formulation augments the usual Maxwell–Vlasov description of plasma dynamics by self-consistently deriving the particle acceleration by the Lorentz force $m\mathbf{a} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$ instead of assuming it *a priori*.

C. The Poisson Hamiltonian structure

The general theory outlined briefly in Sec. VI also leads to the Poisson bracket structure for the Maxwell–Vlasov theory on the Hamiltonian side. However, our Hamiltonian description has a redundancy, namely the information for the particle trajectories can be recovered from the spatial plasma density. Explicitly, if we let $H(f) = (1/2)|\mathbf{v}|^2 + \Phi_f(\mathbf{x})$ be the single-particle Hamiltonian determined by the plasma density f, then the flow of this Hamiltonian function can be identified with the particle evolution map ψ. We can also think of this as a constraint on the level of equations of motion, as the Hamiltonian vector field of $H(f)$ must equal the time derivative of the map ψ, i.e., the particle velocity field in phase space. In other words, as is well known, the particle dynamics is completely determined by the plasma density dynamics. This may be regarded as a

constraint on the system that leads to the elimination of the forward map as a dynamical variable. This ''redundancy'' is, of course, one of the sources of degeneracy of the Lagrangian and Hamiltonian structures.

Thus, the constraint of explicitly enforcing this consistency condition leads to a further ''reduction'' which again may be handled by the Dirac theory of constraints to arrive at the Hamiltonian structure in terms of the variables F (or equivalently f in view of the canonical nature of the particle transformations) and the electromagnetic potentials. The resulting Poisson bracket structure is given by the Lie–Poisson structure for the f's plus the canonical structure for the electromagnetic potentials, which was the starting point for Marsden and Weinstein,[7] who carried out the reduction of this bracket with respect to the action of the electromagnetic gauge group to obtain the final Maxwell–Vlasov bracket on the space with variables f, **E**, and **B**. This procedure was motivated by and corrected a bracket found by *ad hoc* methods in Ref. 15. We need not repeat this construction.

D. The Kelvin–Noether theorem

A final result worth mentioning is Kelvin's theorem for the Maxwell–Vlasov particle dynamics. These dynamics, given in the last equation in (7.9), may be rewritten as

$$\left(\frac{\partial}{\partial t}+\pounds_{\mathbf{u}}\right)\left(\frac{1}{F}m_i dz^i\right)+d\frac{\delta h}{\delta F}=0, \tag{7.16}$$

so that

$$\frac{d}{dt}\oint_{\gamma(t)}\frac{1}{F}m_i\,dz^i=0, \tag{7.17}$$

for a loop $\gamma(t)$ which follows the particle trajectories in phase space. The Kelvin circulation integral in phase space,

$$I=\oint_{\gamma(t)}\frac{1}{F}m_i\,dz^i, \tag{7.18}$$

may be evaluated on the invariant constraint manifold $\mathbf{m}_v=0$ as

$$I=\oint_{\gamma(t)}(mu_{si}+qA_i)dx^i. \tag{7.19}$$

We recognize this integral as the *Poincaré invariant* for the single-particle motion in phase space.

The above result follows from the abstract Kelvin–Noether theorem by letting $\mathscr{C}:=\{\gamma:S^1\to T\mathbb{R}^3|\,\gamma$ continuous$\}$ be the space of continuous loops in single-particle velocity phase space and letting the group $\mathrm{Diff}(T\mathbb{R}^3)$ act on $\mathscr{C}$ on the right by $(\eta,\gamma)\in\mathrm{Diff}(T\mathbb{R}^3)\times\mathcal{C}\mapsto\gamma\circ\eta\in\mathcal{C}$. The quantity $\mathscr{K}$ is chosen to be

$$\langle\mathscr{K}(\gamma,F),a\rangle=\oint_{\gamma}\frac{1}{F}a. \tag{7.20}$$

The abstract Kelvin–Noether theorem for the Maxwell–Vlasov equations in Euler–Poincaré form then reproduces the version of Kelvin's theorem given in (7.17).

VIII. CONCLUSION

In this paper we have cast Low's mixed Eulerian–Lagrangian action principle for Maxwell–Vlasov theory into a purely Eulerian description. In this description we find that Maxwell–Vlasov dynamics are governed by the Euler–Poincaré equations for right invariant motion on the diffeomorphism group of $\mathbb{R}^n$ ($n=6$ for three-dimensional Maxwell–Vlasov motion). These equations were recently discovered by Holm *et al.*[6] who investigated the class of Hamilton's principles which are right invariant under the subgroup of the diffeomorphisms which leaves invariant a set

$\mathscr{T}$ of tensor fields in the Eulerian variables. The Maxwell–Vlasov motions invariant under this subgroup are the steady Eulerian solutions, which, thus, are identified as relative equilibria. This identification of steady Eulerian Maxwell–Vlasov solutions as right invariant equilibria places these solutions into the Hamiltonian framework required for investigating their nonlinear stability characteristics using, e.g., the energy-Casimir method (see Ref. 16). It was this stated goal that first motivated Low to write his Lagrangian for Maxwell–Vlasov dynamics.

Thus, our formulation of a purely Eulerian action principle and its associated Euler–Poincaré equations and Hamiltonian framework advances Low's original intention of using his action principle for studying stability of plasma equilibria by placing the entire Maxwell–Vlasov equations (including the particle dynamics, field dynamics and probability distribution dynamics) into one self-consistent Hamiltonian picture in the Eulerian description. (As we discussed, Low used mixed aspects of both Eulerian and Lagrangian phase space descriptions in his action principle.)

Our Eulerian Hamilton's principle for Maxwell–Vlasov dynamics is constrained, and all of the corresponding Lagrange multipliers have been resolved. This Hamilton's principle is thus available for further approximations, e.g., by Hamilton's principle asymptotics (see, e.g., Ref. 27).

In summary, we have taken an existing action, due to Low,[1] for the Maxwell–Vlasov system of equations and demonstrated how to rederive this system as Euler–Poincaré equations. The Euler–Poincaré form emerges from Hamilton's principle for a system whose configuration space is a group and whose action is right invariant under a subgroup. This situation commonly appears in the Eulerian description of continuum mechanics. In the case of continuum mechanics, the dynamics takes place on the group of diffeomorphisms and the Eulerian variables are invariant under a subgroup of the diffeomorphism group. (This subgroup corresponds to steady Eulerian flows with nonzero velocity and vorticity.) We showed that this situation also occurs for the Maxwell–Vlasov equations of plasma dynamics in the Eulerian description, by showing that the variations considered take the appropriate form, and then deriving the Maxwell–Vlasov equations from the Hamilton's principle for the right invariant action (5.6) in Eulerian variables. We then passed to the Hamiltonian formulation of this system and found its Lie–Poisson structure.

As discussed in the Introduction, the Euler–Poincaré form of the dynamics is naturally adapted for applying Lagrange–D'Alembert methods for geometrical constraints and control as in Ref. 13. In future work, our Euler–Poincaré form of the Maxwell–Vlasov system shall be implemented to describe the control features of a plasma driven by an external antenna, following the lines of inquiry begun in the oscillation center approximation for plasmas by Similon *et al.*[35]

ACKNOWLEDGMENTS

We would all like to extend our gratitude to T. Ratiu for his time and invaluable input. Some early (unpublished) work by J. Marsden, P. Morrison, and H. Gumral on aspects of the problem addressed in this paper helped us in the formulation given here and is gratefully acknowledged. In addition, M. Hoyle would like to thank Dion Burns, Ion Georgiou, Don Korycansky, Shinar Kouranbaeva, Myung Kim and David Schneider for their time and the discussions which helped him understand this material. Work by D. Holm was conducted under the auspices of the US Department of Energy, supported (in part) by funds provided by the University of California for the conduct of discretionary research by Los Alamos National Laboratory. J. Marsden was partially supported by US Department of Energy Contract No. DE-FG0395-ER25251.

[1] F. E. Low, ''A Lagrangian formulation of the Boltzmann–Vlasov equation for plasmas,'' Proc. R. Soc. London, Ser. A **248**, 282–287 (1958).
[2] V. I. Arnold, ''Sur la géometrie differentielle des groupes de Lie de dimenson infinie et ses applications à l'hydrodynamique des fluids parfaits,'' Ann. Inst. Fourier, Grenoble **16**, 319–361 (1966).
[3] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*. Texts in Applied Mathematics, Vol. 17 (Springer-Verlag, Berlin, 1994).
[4] H. Poincaré, ''Sur la stabilité de l'équilibre des figures piriformes affectées par une masse fluide en rotation,'' Philos. Trans. R. Soc. London, Ser. A **198**, 333–373 (1901).
[5] H. Poincaré, ''Sur une forme nouvelle des équations de la méchanique,'' C. R. Hebd. Seances Acad. Sci. **132**, 369–371 (1901).
[6] D. D. Holm, J. E. Marsden, and T. Ratiu, ''Euler-Poincaré equations and semidirect products with applications to continuum theories,'' to appear in Adv. Math. (1998).
[7] J. E. Marsden and A. Weinstein, ''The Hamiltonian structure of the Maxwell-Vlasov equations,'' Physica D **4**, 394–406 (1982).

[8] H. Ye and P. J. Morrison, ''Action principles for the Vlasov equation,'' Phys. Fluids B **4**, 771–777 (1992).
[9] H. Cendra and J. E. Marsden, ''Lin constraints, Clebsch potentials and variational principles,'' Physica D **27**, 63–89 (1987).
[10] H. Cendra, A. Ibort, and J. E. Marsden, ''Variational principal fiber bundles: a geometric theory of Clebsch potentials and Lin constraints,'' J. Geom. Phys. **4**, 183–206 (1987).
[11] J. E. Marsden and J. Scheurle, ''Lagrangian reduction and the double spherical pendulum,'' Z. Angew. Math. Mech. **44**, 17–43 (1993).
[12] J. E. Marsden and J. Scheurle, ''The reduced Euler–Lagrange equations,'' Fields Institute Comm. **1**, 139–164 (1993).
[13] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. Murray, ''Nonholonomic mechanical systems with symmetry,'' Arch. Ration. Mech. Anal. **136**, 21–99 (1996).
[14] Z. R. Iwiński and L. A. Turski, ''Canonical theories of systems interacting electromagnetically,'' Letters in Applied and Engineering Sciences **4**, 179–191 (1976).
[15] P. J. Morrison, ''The Maxwell–Vlasov equations as a continuous Hamiltonian system,'' Phys. Lett. **80**, 383–386 (1980).
[16] D. D. Holm, J. E. Marsden, T. S. Ratiu, and A. Weinstein, ''Nonlinear stability of fluid and plasma equilibria,'' Phys. Rep. **123**, 1–116 (1985).
[17] J. E. Marsden, A. Weinstein, T. S. Ratiu, R. Schmid, and R. G. Spencer, ''Hamiltonian systems with symmetry, coadjoint orbits and plasma physics,'' in Proc. IUTAM-IS1MM Symposium on *Modern Developments in Analytical Mechanics*, Torino 1982, Atti della Acad. della Sc. di Torino **117**, 289–340 (1983).
[18] A. N. Kaufman and R. L. Dewar, ''Canonical derivation of the Vlasov–Coulomb noncanonical Poisson structure,'' Cont. Math. AMS **28**, 51–54 (1984).
[19] P. J. Morrison, ''Variational principle and stability of nonmonotone Vlasov–Poisson equilibria,'' Z. Naturforsch. A **42a**, 1115–1123 (1987).
[20] P. J. Morrison and D. Pfirsch, ''Free-energy expressions for Vlasov equilibria,'' Phys. Rev. A **40**, 3898–3910 (1989).
[21] Y. H. Wan, ''Nonlinear stability of stationary spherically symmetric models in stellar dynamics,'' Arch. Rational Mech. Anal. **112**, 83–95 (1990) and preprint, (1997).
[22] J. Batt and G. Rein, ''A rigorous stability result for the Vlasov–Poisson system in three dimensions,'' Ann. Mat. Pura Appl. **164**, 133–154 (1993).
[23] J. Batt, P. J. Morrison, and G. Rein, ''Linear stability of stationary solutions of the Vlasov–Poisson system in three dimensions,'' Arch. Ration. Mech. Anal. **130**, 163–182 (1995).
[24] P. A. Sturrock, ''A variational principle and an energy theorem for small amplitude disturbances of electron beams and of electron plasmas,'' Ann. Phys. (N.Y.) **4**, 306–324 (1958).
[25] J. J. Galloway and H. Kim, ''Lagrangian approach to non-linear wave interaction in a warm plasma,'' J. Plasma Phys. **6**, 53 (1971).
[26] R. L. Dewar, ''A Lagrangian theory for non-linear wave packets in a collisionless plasma,'' J. Plasma Phys. **7(2)**, 267–284 (1972).
[27] D. D. Holm, ''Hamiltonian balance equations,'' Physica D **98**, 379–414 (1996).
[28] J. H. Jeans, ''The stability of a spherical nebula,'' Phil. Trans. R. Soc. London **199**, 1–53 (1902).
[29] H. Poincaré, *Théorie des tourbillons* (Éditions Jacques Gabay, Paris, 1890) (reprint).
[30] K. Chandrasekhar, *Ellipsoidal Figures of Equilibrium* (Dover, New York, 1977).
[31] J. Binney and S. Tremaine, *Galactic Dynamics*, 1st ed. (Princeton U.P., Princeton, NJ, 1987).
[32] Strictly speaking, the constrained variational principle is not a variational principle because of the constraints imposed on the variations. Rather, this principle is more like the Lagrange d'Alembert principle used in nonholonomic mechanics.
[33] W. M. Tulczyjew, ''The Legendre transformation,'' Ann. Inst. Henri Poincaré, Sect. A **27**, 101–114 (1977).
[34] P. A. M. Dirac, ''Generalized Hamiltonian mechanics,'' Can. J. Math. **2**, 129–148 (1950).
[35] P. L. Similon, A. N. Kaufman, and D. D. Holm, ''Oscillation center theory and pondermotive stabilization of the low-frequency plasma modes,'' Phys. Fluids **29**, 1908–1922 (1986).

Maxwell equations and the optical geometry

Sebastiano Sonego[a)]
Università degli Studi di Udine, DIC—Via delle Scienze 208, 33100 Udine, Italy and International School for Advanced Studies, Via Beirut 2-4, 34014 Trieste, Italy

Marek A. Abramowicz[b)]
Department of Astronomy and Astrophysics, Chalmers University of Technology, 41296 Göteborg, Sweden and International School for Advanced Studies, Via Beirut 2-4, 34014 Trieste, Italy and Università degli Studi di Udine, DIC—Via delle Scienze 208, 33100 Udine, Italy

(Received 14 April 1997; accepted for publication 10 February 1998)

We show that, in conformally static spacetimes, Maxwell equations take their simplest form when written in terms of the so-called ''optical metric.'' © *1998 American Institute of Physics.* [S0022-2488(98)01006-8]

I. INTRODUCTION

In an inertial frame[1] of Minkowski spacetime, Maxwell equations take their simplest form:

$$\nabla\cdot\mathbf{B}=0, \tag{1.1}$$

$$\nabla\times\mathbf{E}+\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t}=\mathbf{0}, \tag{1.2}$$

$$\nabla\cdot\mathbf{E}=4\pi\rho, \tag{1.3}$$

$$\nabla\times\mathbf{B}-\frac{1}{c}\frac{\partial\mathbf{E}}{\partial t}=\frac{4\pi}{c}\,\mathbf{j}. \tag{1.4}$$

In fact, one can define a notion of electromagnetically preferred frames as those in which Maxwell equations become simple, and then verify that such frames coincide with those identified by the other requirement, deriving from particle dynamics, that there are no ''apparent'' forces of inertial origin. Thus, mechanics and electromagnetism identify the same class of privileged frames in flat spacetime. Of course, these are nothing else than the inertial frames of special relativity.

An alternative way to characterize an inertial frame is to require that the family of observers associated with it be nonaccelerating, nonrotating, and nondeforming. This immediately creates a difficulty if one tries to carry over this notion to general relativity. Let n^a be a field of four-velocities, and $h_a{}^b:=\delta_a{}^b+n_a n^b$ the associated transverse projector.[2] One can define in a unique way the acceleration a_a, vorticity ω_{ab}, shear σ_{ab}, and expansion θ of the frame n^a, through the decomposition[3]

$$\nabla_b n_a=-n_b a_a+\omega_{ab}+\sigma_{ab}+\tfrac{1}{3}\theta h_{ab}. \tag{1.5}$$

For a nonaccelerating, nonrotating, and nondeforming frame, all the quantities a_a, ω_{ab}, σ_{ab}, and θ must vanish, which amounts to saying that $\nabla_a n^b=0$. However, it is not difficult to see[4] that vector fields satisfying this condition can only exist in the so-called ''ultrastatic'' spacetimes,[5] i.e., those whose metric admits the coordinate representation

$$g=-\mathrm{d}t^2+h_{ij}\mathrm{d}x^i\mathrm{d}x^j, \tag{1.6}$$

with i,j taking values from 1 to 3, and $\partial h_{ij}/\partial t=0$. These spacetimes are physically rather dull; they describe situations in which there are no gravitational forces, and differ from Minkowski spacetime only by having curved $t=$ const spatial sections. It seems therefore that no extension of the idea of inertial frames is possible in a general-relativistic context of physical interest.

a)Electronic mail: sebastiano.sonego@dic.uniud.it
b)Electronic mail: marek@tfa.fy.chalmers.se

 © 1998 American Institute of Physics

Recently, however, a proposal has been made for a definition of inertial and gravitational forces in general relativity.[6–8] Contrary to what one might think at first, this idea is not at odds with the equivalence principle.[9,10] Furthermore, it has proved itself very useful for the analysis of several nontrivial general-relativistic effects.[11] Since the concept of inertial forces is associated with the one of deviation from inertial motion, it seems therefore that, at least in those spacetimes for which such forces can be defined, one should be able to introduce a concept of mechanically inertial frames.[9] Of course, the condition $\nabla_a n^b = 0$ must be replaced by a weaker one, in a way that we now briefly describe. A much more detailed discussion can be found in Refs. 8 and 9.

Let us first consider Newtonian mechanics in a static gravitational field with potential ϕ. An inertial observer at rest at a point $\mathbf{x}$ is subject to the gravitational force $-\nabla\phi(\mathbf{x})$, and therefore needs a thrust (i.e., a force of nongravitational origin) $\nabla\phi(\mathbf{x})$ per unit mass in order to remain at $\mathbf{x}$. This is exactly what happens, within a general-relativistic context, to static observers in a static spacetime $(\mathcal{M}, g_{ab})$. For, if we denote by η^a the timelike Killing vector field, the four-velocity of static observers is given by $n^a = \eta^a/(-\eta_b \eta^b)^{1/2}$, and one can easily check that their four-acceleration a_a (which is directly related to the thrust per unit mass that they need in order to remain static) can be expressed as $a_a = \nabla_a \Phi$, where

$$\Phi = \tfrac{1}{2} \ln(-\eta_a \eta^a) . \tag{1.7}$$

The function Φ can thus be thought of as a generalization of the Newtonian potential,[6,12] and the static frame as the counterpart of the Newtonian rest frame.[12] It is then reasonable to require that any definition of inertial observers should include the static ones as a particular case, in spite of the fact that they have $a_a \neq 0$. We now look for a property of the static frame n^a that could be taken as a more general characterization of inertial frames.

The vector field n^a has $\omega_{ab} = \sigma_{ab} = \theta = 0$, so that $\nabla_b n_a = -n_b \nabla_a \Phi$. Introducing the new metric

$$\tilde{g}_{ab} := e^{-2\Phi} g_{ab} , \tag{1.8}$$

which is of the ultrastatic type, this relationship can be reexpressed as $\tilde{\nabla}_a \tilde{n}^b = 0$, where $\tilde{\nabla}$ is the Riemannian connection associated to $\tilde{g}_{ab}$, and $\tilde{n}^a := e^{\Phi} n^a = \eta^a$ has unit norm with respect to $\tilde{g}_{ab}$. This suggests that $(\mathcal{M}, \tilde{g}_{ab})$, rather than the physical spacetime $(\mathcal{M}, g_{ab})$, is the natural structure to use in identifying inertial observers. Inspired by these heuristic remarks, we shall say that a frame n^a in $(\mathcal{M}, g_{ab})$ is inertial if there exists a conformally related $(\mathcal{M}, \tilde{g}_{ab})$ in which $\tilde{n}^a = e^{\Phi} n^a$ satisfies the condition $\tilde{\nabla}_a \tilde{n}^b = 0$.[13]

Of course, this definition allows one to define inertial frames only in spacetimes that are conformally static. A further generalization is possible;[6,9] however, as far as the present paper is concerned, we prefer to remain within a limited context, rather than introducing technical complications that would make the discussion less clear. Also, it must be pointed out that conformally static spacetimes already include several examples of great physical interest.

In a conformally static spacetime, g_{ab} admits at least one timelike conformal Killing vector field[3] η^a that satisfies the Frobenius condition, so that the inertial observers have $\tilde{n}^a = \eta^a$. In the ultrastatic spacetime $(\mathcal{M}, \tilde{g}_{ab})$, with $\tilde{g}_{ab}$ defined by Eqs. (1.7) and (1.8), the frame $\tilde{n}^a$ *is* indeed nonaccelerating, nonrotating, and nondeforming. The kinematical parameters of $\tilde{n}^a$ are related to those of n^a by the transformation:

$$\tilde{a}_a = a_a - h_a{}^b \nabla_b \Phi, \tag{1.9}$$

$$\tilde{\theta} = e^{\Phi}(\theta - 3\, n^a \nabla_a \Phi), \tag{1.10}$$

$$\tilde{\omega}_{ab} = e^{-\Phi} \omega_{ab} , \tag{1.11}$$

$$\tilde{\sigma}_{ab} = e^{-\Phi} \sigma_{ab} . \tag{1.12}$$

Therefore, in the physical spacetime $(\mathcal{M}, g_{ab})$, n^a has acceleration

$$a_a = h_a{}^b \nabla_b \Phi\,, \tag{1.13}$$

expansion

$$\theta = 3\, n^a \nabla_a \Phi\,, \tag{1.14}$$

and vanishing rotation and shear:

$$\omega_{ab} = \sigma_{ab} = 0\,. \tag{1.15}$$

Of course, the set of Eqs. (1.13)–(1.15) could also be taken as definition of the frame n^a and of the function Φ, but it is more convenient, in practice, to look simply for a conformal Killing vector field η^a.

This characterization of inertial frames may look somewhat artificial, as it requires one to introduce an unphysical spacetime $(\mathcal{M},\tilde{g}_{ab})$. Nevertheless, it is fully in line with the common ideas about what an inertial frame should be. In fact, when expressed in the frame $\tilde{n}^a$ in $(\mathcal{M},\tilde{g}_{ab})$, particle dynamics takes its simplest form, which closely mirrors Newtonian dynamics in an inertial frame.[6–9] In addition, the spatial trajectories of massless particles are geodesics of $\tilde{h}_{ab} = e^{-2\Phi} h_{ab}$.[6,7,9,14] Thus, light propagates along ''straight lines'' according to $\tilde{h}_{ab}$; for this reason, $\tilde{h}_{ab}$ is called the *optical metric*.[14]

Given the fact that one can successfully introduce a notion of mechanically inertial observers in conformally static spacetimes, it is natural to ask whether this is also true for electromagnetism; to wit, whether Maxwell equations also take a simpler form when expressed in the optical metric. In this article we show that this is indeed the case, i.e., that Maxwell equations, if written in terms of electric and magnetic fields, take their simplest form not in the physical spacetime $(\mathcal{M},g_{ab})$, but rather in $(\mathcal{M},\tilde{g}_{ab})$, when the fields are defined with respect to the frame $\tilde{n}^a$.

The presentation is organized as follows. In Sec. II we derive the form of Maxwell equations with respect to an arbitrary frame in the spacetime $(\mathcal{M},g_{ab})$. In Sec. III we do the same in $(\mathcal{M},\tilde{g}_{ab})$, and show that in this case one can choose the frame in such a way that Maxwell equations take the very simple form of Eqs. (1.1)–(1.4). In Sec. IV we consider two examples of physical interest, in order to show how problems that are otherwise complicated, simplify within this formalism. Section V contains some concluding remarks.

II. DECOMPOSITION OF MAXWELL EQUATIONS

In Minkowski spacetime, Maxwell equations

$$\nabla_{[a} F_{bc]} = 0 \tag{2.1}$$

and

$$\nabla^b F_{ba} = -4\pi J_a\,, \tag{2.2}$$

when decomposed in an inertial frame, take the familiar form of Eqs. (1.1)–(1.4).

In a general spacetime $(\mathcal{M},g_{ab})$, one can still define the electric and magnetic fields with respect to an arbitrary reference frame n^a as[3]

$$E_a := F_{ab} n^b \tag{2.3}$$

and

$$B^a := -\tfrac{1}{2} \epsilon^{abcd} n_b F_{cd}\,. \tag{2.4}$$

The electromagnetic tensor can then be expressed in the form[15]

$$F_{ab} = n_a E_b - n_b E_a + \epsilon_{abcd} n^c B^d. \tag{2.5}$$

This follows immediately on rewriting Eq. (2.4) as

$$\epsilon_{abcd}B^a=-\frac{1}{2}\,\epsilon_{abcd}\epsilon^{aefg}n_eF_{fg}=\frac{3!}{2}\,\delta^{[e}{}_b\delta^f{}_c\delta^{g]}{}_d n_eF_{fg}=n_bF_{cd}+n_dF_{bc}+n_cF_{db}\,, \tag{2.6}$$

and then contracting Eq. (2.6) with n^b, and using the definition (2.3) of E_a.

We now want to write differential equations for E_a and B^a. For this purpose, we must decompose Eqs. (2.1) and (2.2) in the frame n^a.

Let us begin by rewriting Eq. (2.1) in the form $\epsilon^{abcd}\nabla_bF_{cd}=0$. Using expression (2.5) for F_{cd} and the property

$$\epsilon^{abcd}\epsilon_{cdef}=2(\delta^a{}_f\delta^b{}_e-\delta^a{}_e\delta^b{}_f)\,, \tag{2.7}$$

we arrive at

$$\epsilon^{abcd}n_b(D_cE_d+a_cE_d)+\epsilon^{abcd}\omega_{bc}E_d+\nabla_bn^aB^b-\nabla_bn^bB^a+n^a\nabla_bB^b-n^b\nabla_bB^a=0\,. \tag{2.8}$$

Here, D denotes the transverse Riemannian connection associated with h_{ab},[3]

$$D_aT^{b_1\cdots b_r}{}_{c_1\cdots c_s}:=h_a{}^dh_{e_1}{}^{b_1}\cdots h_{e_r}{}^{b_r}h_{c_1}{}^{f_1}\cdots h_{c_s}{}^{f_s}\nabla_dT^{e_1\cdots e_r}{}_{f_1\cdots f_s} \tag{2.9}$$

(see the Appendix for the geometrical meaning of D when $\omega_{ab}\neq0$). Since

$$\nabla_bB^c=D_bB^c-n_bn^dh^c{}_e\nabla_dB^e-h_b{}^dn^cn_e\nabla_dB^e+n_bn^dn^cn_e\nabla_dB^e\,, \tag{2.10}$$

Eq. (2.8) can be rewritten as

$$\epsilon^{abcd}n_b(D_cE_d+a_cE_d)+\epsilon^{abcd}\omega_{bc}E_d+n^aD_bB^b-h^a{}_bn^c\nabla_cB^b+\omega^a{}_bB^b+\sigma^a{}_bB^b-\tfrac{2}{3}\,\theta B^a=0. \tag{2.11}$$

An expression similar to Eq. (2.11), with the roles of E_a and B^a exchanged, is easily obtained by substituting Eq. (2.5) into the left-hand side of Eq. (2.2), and using the analogue of Eq. (2.10) for E_a:

$$-\epsilon_{abcd}n^b(D^cB^d+a^cB^d)-\epsilon_{abcd}\omega^{bc}B^d+n_aD^bE_b-h_a{}^bn^c\nabla_cE_b+\omega_a{}^bE_b+\sigma_a{}^bE_b-\tfrac{2}{3}\theta E_a=4\pi J_a\,. \tag{2.12}$$

Now, on projecting Eqs. (2.11) and (2.12) along directions parallel and orthogonal to n^a, and defining the density of charge ρ and density of current j_a through $J_a=\rho\,n_a+j_a$, with $j_a:=h_a{}^bJ_b$, we finally obtain:

$$D_aB^a-\epsilon^{abcd}n_a\omega_{bc}E_d=0, \tag{2.13}$$

$$\epsilon^{abcd}n_b(D_cE_d+a_cE_d)-h^a{}_bn^c\nabla_cB^b+\omega^a{}_bB^b+\sigma^a{}_bB^b-\tfrac{2}{3}\theta B^a=0\,, \tag{2.14}$$

$$D^aE_a+\epsilon_{abcd}n^a\omega^{bc}B^d=4\pi\rho, \tag{2.15}$$

$$-\epsilon_{abcd}n^b(D^cB^d+a^cB^d)-h_a{}^bn^c\nabla_cE_b+\omega_a{}^bE_b+\sigma_a{}^bE_b-\tfrac{2}{3}\theta E_a=4\pi j_a\,. \tag{2.16}$$

As is obvious by inspection of Eqs. (2.13)–(2.16), the decomposition of Maxwell equations (2.1) and (2.2) with respect to a frame n^a does not produce, in general, a set of equations of the same form of Eqs. (1.1)–(1.4), because of the occurrence of extra terms arising from the derivatives of n^a. Also strictly related to this circumstance is the fact that charge is *apparently* not conserved in the frame n^a:

$$n^a\nabla_a\rho+D_aj^a=-\rho\,\theta-j^aa_a\,. \tag{2.17}$$

The additional terms in Eqs. (2.13)–(2.17) have exactly the same status of the so-called apparent forces that one must introduce when formulating dynamics in a noninertial reference

frame. Only if $\nabla_a n^b = 0$, does one recover the simple form of Eqs. (2.1)–(2.2) and of the continuity equation. However, this is possible only in the very special case of an ultrastatic spacetime.

III. OPTICAL GEOMETRY RECOVERED

It is a well-known property of conformal transformations that, in four spacetime dimensions,[3]

$$\tilde{\nabla}_{[a}F_{bc]} = \nabla_{[a}F_{bc]} \tag{3.1}$$

and

$$\tilde{\nabla}^b F_{ba} = e^{2\Phi}\nabla^b F_{ba}\,. \tag{3.2}$$

[Hereafter, indices of geometrical objects pertaining to the spacetime $(\mathcal{M},\tilde{g}_{ab})$ are raised and lowered by using $\tilde{g}_{ab}$ and its inverse $\tilde{g}^{ab}$. Thus we have, for example, $\tilde{X}^a := \tilde{g}^{ab}\tilde{X}_a$.] If we define $\tilde{J}_a := e^{2\Phi}J_a$, Eqs. (3.1) and (3.2) imply that Maxwell equations (2.1) and (2.2) hold in $(\mathcal{M},g_{ab})$ iff their counterparts

$$\tilde{\nabla}_{[a}F_{bc]} = 0 \tag{3.3}$$

and

$$\tilde{\nabla}^b F_{ba} = -4\pi\tilde{J}_a \tag{3.4}$$

hold in $(\mathcal{M},\tilde{g}_{ab})$. Notice that if we write $\tilde{J}_a = \tilde{\rho}\,\tilde{n}_a + \tilde{j}_a$, we have that $\tilde{\rho} = e^{3\Phi}\rho$; this is perfectly reasonable, if we think of $\tilde{\rho}$ as the electric charge per unit volume as measured according to the spatial metric $\tilde{h}_{ab}$. Similarly, we have $\tilde{j}_a = e^{2\Phi}j_a$, consistent with $\tilde{j}_a = \tilde{\rho}\tilde{v}_a$ for some three-velocity $\tilde{v}_a = e^{-\Phi}v_a$.

Since $\tilde{\epsilon}_{abcd} = e^{-4\Phi}\epsilon_{abcd}$, the electric and magnetic vectors in $(\mathcal{M},\tilde{g}_{ab})$ are

$$\tilde{E}_a := F_{ab}\tilde{n}^b = e^{\Phi}E_a \tag{3.5}$$

and

$$\tilde{B}^a := -\tfrac{1}{2}\tilde{\epsilon}^{abcd}\tilde{n}_b F_{cd} = e^{3\Phi}B^a\,. \tag{3.6}$$

Repeating the same procedure that led us to Eqs. (2.13)–(2.16), we write:

$$\tilde{D}_a\tilde{B}^a - \tilde{\epsilon}^{abcd}\tilde{n}_a\tilde{\omega}_{bc}\tilde{E}_d = 0\,, \tag{3.7}$$

$$\tilde{\epsilon}^{abcd}\tilde{n}_b(\tilde{D}_c\tilde{E}_d + \tilde{a}_c\tilde{E}_d) - h^a{}_b\tilde{n}^c\tilde{\nabla}_c\tilde{B}^b + \tilde{\omega}^a{}_b\tilde{B}^b + \tilde{\sigma}^a{}_b\tilde{B}^b - \tfrac{2}{3}\tilde{\theta}\tilde{B}^a = 0\,, \tag{3.8}$$

$$\tilde{D}^a\tilde{E}_a + \tilde{\epsilon}_{abcd}\tilde{n}^a\tilde{\omega}^{bc}\tilde{B}^d = 4\pi\tilde{\rho}, \tag{3.9}$$

$$-\tilde{\epsilon}_{abcd}\tilde{n}^b(\tilde{D}^c\tilde{B}^d + \tilde{a}^c\tilde{B}^d) - h_a{}^b\tilde{n}^c\tilde{\nabla}_c\tilde{E}_b + \tilde{\omega}_a{}^b\tilde{E}_b + \tilde{\sigma}_a{}^b\tilde{E}_b - \tfrac{2}{3}\tilde{\theta}\tilde{E}_a = 4\pi\tilde{j}_a\,. \tag{3.10}$$

If $(\mathcal{M},g_{ab})$ is conformally static with conformal Killing vector field η^a, we can choose Φ as in Eq. (1.7). It is then obvious that, in the reference frame $\tilde{n}^a = \eta^a$ in $(\mathcal{M},\tilde{g}_{ab})$, Eqs. (3.7)–(3.10) become:

$$\tilde{D}_a\tilde{B}^a = 0\,, \tag{3.11}$$

$$\tilde{\epsilon}^{abcd}\tilde{n}_b\tilde{D}_c\tilde{E}_d - \tilde{n}^b\tilde{\nabla}_b\tilde{B}^a = 0\,, \tag{3.12}$$

$$\tilde{D}^a\tilde{E}_a = 4\pi\tilde{\rho}\,, \tag{3.13}$$

$$-\tilde{\epsilon}_{abcd}\tilde{n}^b\tilde{D}^c\tilde{B}^d - \tilde{n}^b\tilde{\nabla}_b\tilde{E}_a = 4\pi\tilde{j}_a\,. \tag{3.14}$$

The second term of Eqs. (3.8) and (3.10) has been simplified by noticing that, e.g.,

$$h^a{}_b\tilde{n}^c\tilde{\nabla}_c\tilde{B}^b = \tilde{h}^a{}_b\tilde{n}^c\tilde{\nabla}_c\tilde{B}^b = \tilde{n}^c\tilde{\nabla}_c(\tilde{h}^a{}_b\tilde{B}^b) = \tilde{n}^b\tilde{\nabla}_b\tilde{B}^a\,, \tag{3.15}$$

because $\tilde{\nabla}_a\tilde{h}^b{}_c = 0$. These equations have exactly the same form as Eqs. (1.1)–(1.4). Also, in the frame $\tilde{n}^a$ charge conservation is expressed by the familiar continuity equation

$$\tilde{n}^a\tilde{\nabla}_a\tilde{\rho} + \tilde{D}_a\tilde{j}^a = 0\,, \tag{3.16}$$

without the ''apparent sources'' that are present in the right-hand side of Eq. (2.17). Therefore, in the reference frame $\tilde{n}^a$ of $(\mathcal{M},\tilde{g}_{ab})$, electromagnetism is described by the same equations that hold in an inertial frame of Minkowski spacetime.

IV. EXAMPLES

The formulation of particle dynamics in the optical geometry leads not only to a better intuitive understanding, but also to formal simplifications that allow one to have better control over complicated physical problems.[11] One expects that something similar should happen for electromagnetism as well. In this section we illustrate this point with the help of two simple examples, leaving the discussion of more technical problems for a future publication.

A. Magnetic field around a Schwarzschild black hole

Equations (3.11)–(3.14) allow one to introduce potentials exactly in the same way as is usually done in flat spacetime. We now present an example of a complicated problem that can be easily solved by this procedure.

Consider a Schwarzschild black hole immersed into an otherwise homogeneous, constant magnetic field. (We assume that the intensity of the latter is sufficiently low that it does not perturb spacetime appreciably.) We want to know how the field is modified by the presence of the black hole. This problem was solved by Wald in an indirect way,[16] and by Hanni and Ruffini through a rather lengthy calculation.[17]

For a general magnetostatic problem, Eq. (3.14) implies that, in a region in which there are no currents, one can express $\tilde{B}^a$ as the gradient of a scalar function ψ,

$$\tilde{B}^a = \tilde{D}^a\psi\,. \tag{4.1}$$

Substituting into Eq. (3.11), we see that ψ must be harmonic with respect to the optical metric:

$$\tilde{D}_a\tilde{D}^a\psi = 0\,. \tag{4.2}$$

In the particular case of Schwarzschild spacetime, $e^{\Phi} = (1-2M/r)^{1/2}$, and the optical metric is

$$\tilde{h} = \left(1-\frac{2M}{r}\right)^{-2}\mathrm{d}r^2 + \left(1-\frac{2M}{r}\right)^{-1} r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\ \mathrm{d}\varphi^2)\,. \tag{4.3}$$

Orienting the coordinate system in such a way that the field at $r\to+\infty$ points in the azimuthal direction, ψ does not depend on φ, and Eq. (4.2) can be written explicitly as

$$\left(1-\frac{2M}{r}\right)\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\psi}{\partial r}\right) + \frac{1}{\sin\vartheta}\,\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\,\frac{\partial\psi}{\partial\vartheta}\right) = 0\,. \tag{4.4}$$

It is very easy to see that the solution of Eq. (4.4) which satisfies the asymptotic condition $\psi(r \to +\infty) \propto r\cos\vartheta$ is $\psi = K(r-2M)\cos\vartheta$, where K is a constant that represents the intensity of the field at infinity. The magnetic field B^a is then trivially found using Eqs. (4.1) and (3.6), and has the components:

$$B^r = K\left(1-\frac{2M}{r}\right)^{1/2}\cos\vartheta, \tag{4.5}$$

$$B^\vartheta = -K\left(1-\frac{2M}{r}\right)^{1/2}\frac{\sin\vartheta}{r}, \tag{4.6}$$

$$B^\varphi = 0. \tag{4.7}$$

B. Electromagnetism in FLRW cosmological models

Our second example is that of a Friedmann–Lemaitre–Robertson–Walker (FLRW) cosmological model, with metric

$$g = -\mathrm{d}t^2 + a(t)^2[\mathrm{d}\chi^2 + f(\chi)^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\ \mathrm{d}\varphi^2)], \tag{4.8}$$

where $a(t)$ depends on the particular matter content of the model, and $f(\chi)$ is either $\sin\chi$, $\sinh\chi$, or χ, according to whether the curvature of the spatial sections $t=$ const is positive, negative, or zero. It is obvious that this metric is conformally ultrastatic, with

$$\tilde{h} = \mathrm{d}\chi^2 + f(\chi)^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\ \mathrm{d}\varphi^2) \tag{4.9}$$

and $e^\Phi = a/a_0$, where $a_0 =$ const.

Suppose that we are given a configuration of charges and currents in a FLRW universe, and that we look for the electric and magnetic fields measured by the observers comoving with the cosmological fluid. Equations (3.11)–(3.14) tell us that we can consider a much simpler problem for the fields $\tilde{E}_a = (a/a_0)E_a$ and $\tilde{B}^a = (a/a_0)^3 B^a$, corresponding to the distribution of charges and currents described by $\tilde{\rho} = (a/a_0)^3\rho$ and $\tilde{j}_a = (a/a_0)^2 j_a$. In Eqs. (3.12) and (3.14), the terms $\tilde{n}^b\tilde{\nabla}_b\tilde{B}^a$ and $\tilde{n}^b\tilde{\nabla}_b\tilde{E}_a$ reduce to $\partial\tilde{B}^a/\partial t$ and $\partial\tilde{E}_a/\partial t$, respectively, so the only difference between Eqs. (1.1)–(1.4) and Eqs. (3.11)–(3.14) is that the latter are formulated in a curved space with the metric (4.9). This simplification might be of some help in investigations concerning the evolution of cosmic magnetic fields. Of course, in the particular case of a spatially flat model, Eqs. (3.11)–(3.14) exactly *coincide* with Eqs. (1.1)–(1.4); every problem can thus be mapped into a corresponding one formulated in Minkowski spacetime, as is already evident from Eqs. (3.1) and (3.2).

V. CONCLUSIONS

It is by now rather well established that the analysis of mechanical and optical phenomena in conformally static spacetimes can be greatly simplified working in the optical geometry.[6,7,11] In this article we have shown that the same happens for electromagnetism. *A posteriori*, this conclusion may perhaps sound not surprising, because electromagnetism (and, more generally, the physics of any massless field) must be consistent with the description of light propagation, and in particular with geometric optics.

Apart from its theoretical interest, the formulation of Maxwell equations in terms of the optical metric is potentially of practical usefulness, as we have argued in the previous section. The underlying reason for the simplification so achieved is that the influences of gravity and of the geometry of space on the electric and magnetic fields are separated out in this approach. In the ultrastatic spacetime $(\mathcal{M},\tilde{g}_{ab})$ there is no gravity, so Eqs. (3.11)–(3.14) just describe electromagnetism on a curved static space. It is the factors e^Φ in Eqs. (3.5) and (3.6) that account for the effect of gravity on E_a and B^a. In regions that do not contain charges or currents, this effect is simply a rescaling for the values of the fields, and does not alter the shape of the lines of force.

As we said at the beginning of the article, all this applies to conformally static spacetimes. The formalism can be generalized, e.g., by introducing a gravomagnetic field in addition to the potential Φ,[6,9] but this is beyond the scope of the present paper.

APPENDIX: TRANSVERSE COVARIANT DERIVATIVE

If the vector field n^a satisfies the Frobenius condition $n_{[a}\nabla_b n_{c]}=0$, i.e., if $\omega_{ab}=0$, there exist (at least locally) a family of hypersurfaces $\mathcal{S}$ orthogonal to n^a. In this case it is easy to show[3] that D_a defined by Eq. (2.9) coincides with the covariant derivative associated with the Riemannian connection induced on $\mathcal{S}$ by h_{ab}. The purpose of this appendix is to investigate the extent to which this situation is modified when $\omega_{ab}\neq 0$.

In this case the spaces of vectors orthogonal to n^a at different points of spacetime do not mesh in such a way as to coincide with the tangent spaces of any family of hypersurfaces, and there is no hope of defining a ''connection on a submanifold,'' because the submanifold simply does not exist. We can still look, however, for a transverse derivative that characterizes the rate of change of tensors as one moves along a spacelike direction. Of course, a first candidate is the projection of the Riemannian covariant derivative ∇_a associated with g_{ab}, defined by Eq. (2.9). It is straightforward to check that $D_a h_{bc}=0$, but also that for an arbitrary function f,

$$(D_aD_b-D_bD_a)f=-2\,\omega_{ab}n^c\nabla_c f\,. \tag{A1}$$

Thus, D_a does not define a connection, not even one with torsion. For if there were a torsion tensor $T^c{}_{ab}$ associated with D_a, the right-hand side of Eq. (A1) could be written as $-T^c{}_{ab}D_c f$, which is impossible unless $\omega_{ab}=T^c{}_{ab}=0$. This is unpleasant, and one might wonder whether another derivative exists that behaves in a more reasonable way. It is not difficult to see, however, that *no* derivative operator $\mathcal{D}_a$ can satisfy the following two requirements: (i) for any $T^{a_1\cdots a_r}{}_{b_1\cdots b_s}$ which is totally transverse (i.e., such that all its contractions with n^a vanish), $\mathcal{D}_c T^{a_1\cdots a_r}{}_{b_1\cdots b_s}$ is also totally transverse; (ii) there exists $T^c{}_{ab}$ such that $(\mathcal{D}_a\mathcal{D}_b-\mathcal{D}_b\mathcal{D}_a)f=-T^c{}_{ab}\mathcal{D}_c f$, $\forall f$. On the other hand, the notion of transverse derivative can be generalized on replacing (ii) by the weaker condition that there exist a *defect tensor* $\Delta^c{}_{ab}$,[18] such that

$$(\mathcal{D}_a\mathcal{D}_b-\mathcal{D}_b\mathcal{D}_a)f=-2\,\Delta^c{}_{ab}\nabla_c f\,, \tag{A2}$$

for any f. One can then define the torsion as twice the transverse part of the defect tensor, $T^c{}_{ab}:=2\,h^c{}_d\Delta^d{}_{ab}$, so that D_a turns out to be torsion-free (although its defect tensor is nonvanishing, equal to $2\,\omega_{ab}n^c$). It is then easy to prove that D_a is the only torsion-free derivative that satisfies (i), (ii), and which also preserves h_{ab}. We can thus regard it as the closest generalization of the notion of transverse Riemannian covariant derivative to the case in which the Frobenius condition fails to be satisfied.

[1] We adopt the following terminology. An *observer* is a future-directed, differentiable timelike curve. A *reference frame* is a congruence of observers, that can also be identified with the field of their four-velocities.

[2] We use latin indices a, b, ... from the beginning of the alphabet as abstract indices (see, e.g., Ref. 3), which just indicate the tensorial nature of an object without requiring the specification of a chart. We choose $+2$ as signature of the metric, $\epsilon_{0123}>0$ for the sign convention of the Levi–Civita tensor ϵ_{abcd}, and work in units in which $c=1$.

[3] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).

[4] D. Kramer, H. Stephani, E. Herlt, and M. MacCallum, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980), pp. 343–344; H. Stephani, *General Relativity*, 2nd ed. (Cambridge University Press, Cambridge, 1990), p. 174.

[5] S. A. Fulling, ''Alternative vacuum states in static space-times with horizons,'' J. Phys. A **10**, 917–951 (1977).

[6] M. A. Abramowicz, B. Carter, and J. P. Lasota, ''Optical reference geometry for stationary and static dynamics,'' Gen. Relativ. Gravit. **20**, 1173–1183 (1988).

[7] M. A. Abramowicz, ''Centrifugal force: a few surprises,'' Mon. Not. R. Astron. Soc. **245**, 733–746 (1990).

[8] M. A. Abramowicz, ''Inertial forces in general relativity,'' in *The Renaissance of General Relativity and Cosmology*, edited by G. Ellis, A. Lanza, and J. Miller (Cambridge University Press, Cambridge, 1993), pp. 40–58; M. A. Abramowicz, P. Nurowski, and N. Wex, ''Covariant definition of inertial forces,'' Class. Quantum Grav. **10**, L183–L186 (1993); S. Sonego and M. Massar, ''Covariant definition of inertial forces: Newtonian limit and time-dependent gravitational fields,'' *ibid.* **13**, 139–144 (1996).

[9] S. Sonego (unpublished).

[10] S. Sonego and M. Massar, ''On the notions of gravitational and centrifugal force in static spherically symmetric space-times,'' Mon. Not. R. Astron. Soc. **281**, 659–665 (1996).

[11] M. A. Abramowicz and A. R. Prasanna, ''Centrifugal-force reversal near a Schwarzschild black hole,'' Mon. Not. R. Astron. Soc. **245**, 720–728 (1990); M. A. Abramowicz and J. C. Miller, ''Ellipticity behavior of relativistic Maclaurin spheroids,'' *ibid.* **245**, 729–732 (1990); M. A. Abramowicz and J. Bičák, ''The interplay between relativistic gravitational, centrifugal and electric forces: A simple example,'' Gen. Relativ. Gravit. **23**, 941–946 (1991) M. A. Abramowicz, J. C. Miller, and Z. Stuchlík, ''Concept of radius of gyration in general realitivity,'' Phys. Rev. D **47**, 1440–1447 (1993); S. Sonego and A. Lanza, ''Relativistic perihelion advance as a centrifugal effect,'' Mon. Not. R. Astron. Soc. **279**, L65–L66 (1996); M. A. Abramowicz, A. Lanza, J. C. Miller, and S. Sonego, ''Curving Newtonian space,'' Gen. Relativ. Gravit. **29**, 1585–1596 (1997); M. A. Abramowicz, N. Andersson, M. Bruni, P. Ghosh, and S. Sonego, ''Gravitational waves from ultracompact stars: the optical geometry view of trapped modes,'' Class. Quantum Grav. **14,** L189–L194 (1997)

[12] R. D. Greene, E. L. Schucking, and C. V. Vishveshwara, ''The rest frame in stationary space-times with axial symmetry,'' J. Math. Phys. **16**, 153–162 (1975).

[13] In spacetimes with a sufficiently high degree of symmetry, this definition turns out to be too weak, and one must supplement it with additional constraints (Ref. 9). However, these considerations do not affect the following discussion, and we shall therefore ignore them.

[14] J. S. Dowker and G. Kennedy, ''Finite temperature and boundary effects in static space-times,'' J. Phys. A **11**, 895–920 (1978); G. W. Gibbons and M. J. Perry, ''Black holes and thermal Green functions,'' Proc. R. Soc. London, Ser. A **358**, 467–494 (1978); G. Kennedy, R. Critchley, and J. S. Dowker, ''Finite temperature field theory with boundaries: Stress tensor and surface action renormalisation,'' Ann. Phys. (N.Y.) **125**, 346–400 (1980).

[15] G. F. R. Ellis, ''Relativistic cosmology,'' in Cargèse Lectures in Physics, edited by E. Schatzmann (Gordon & Breach, New York, 1973), Vol. VI, pp. 1–60.

[16] R. M. Wald, ''Black hole in a uniform magnetic field,'' Phys. Rev. D **10**, 1680–1685 (1974).

[17] R. S. Hanni and R. Ruffini, ''Schwarzschild black hole in an asymptotically uniform magnetic field,'' Lett. Nuovo Cimento **15**, 189–194 (1976).

[18] G. F. R. Ellis, M. Bruni, and J. Hwang, ''Density-gradient-vorticity relation in perfect-fluid Robertson–Walker perturbations,'' Phys. Rev. D **42**, 1035–1046 (1990), and references therein.

Microscopic dynamics from a coarsely defined solution to the protein folding problem

Ariel Fernández[a] and Andrés Colubri
Instituto de Matemática (INMABB), Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur, Avenida Alem 1253, Bahía Blanca 8000, Argentina

(Received 16 December 1997; accepted for publication 27 February 1998)

In this work we introduce a least-action formulation of the protein folding problem casted within a coarse description of the soft mode dynamics of the peptide chain. Ultimately, we show that this coarse variational approach can be lifted to yield the microscopic long-time torsional dynamics responsible for the actual folding process. As a first step, a binary coding of local topological constraints associated to each structural motif is introduced to coarsely mimic the long-time dynamics. Folding pathways are initially resolved as transitions between patterns of locally encoded structural signals. Our variational approach is aimed at identifying the most economic pathway with respect to the stepwise cost in conformational freedom. Our treatment allows us to account for the expediency of the process in proteins effectively capable of *in vitro* renaturation. We identify the dominant pathway by introducing a coarse version of Lagrangian microscopic dynamics. The coarse folding pathways are generated by a parallel search for structural patterns in a matrix of local topological constraints (LTM) of the chain. Each local topological constraint represents a coarse description of a local torsional state and each pattern is evaluated, translated, and finally recorded as a contact matrix (CM), an operation that is subject to a renormalization feedback loop. The renormalization operation periodically introduces long-range correlations on the LTM according to the latest CM generated by translation. Local topological constraints may form consensus regions in portions of the chain that translate as secondary structure motifs or tertiary interactions. Nucleation steps and cooperative effects are accounted for by means of the renormalization operation, which warrants the persistence of seeding patterns upon successive LTM evaluations. Relevant folding time scales beyond the realm of molecular dynamics simulations become accessible through the coarsely codified representation of local torsional constraints. The validity of our approach is tested *vis à vis* experimentally probed folding pathways generating tertiary interactions in proteins that may recover their active structure under *in vitro* renaturation conditions. We focus on determining significant folding intermediates and the late kinetic bottlenecks that occur within the first $\frac{1}{100}$ s of the renaturation process. After the computational accessibility of this coarse solution of the folding problem becomes apparent, we show how to lift our variational problem to microscopic dynamics of the peptide chain. The consistency of our approach is revealed by an actual generation of Newton's laws at the microscopic level through an inverse projection of the coarse dynamics originally generated through the pattern recognition computation. © *1998 American Institute of Physics.*
[S0022-2488(98)01506-0]

I. INTRODUCTION

This work is concerned with the theoretical underpinnings of the microscopic expedient by which natural proteins reach their active conformation under *in vitro* renaturation conditions. Specifically, we shall focus on proteins which recover their activity as a result of the actual dominance of robust folding pathways.[1–8] In this regard we shall concentrate on two critical

[a]Corresponding author; electronic mail: inmabb@criba.edu.ar

 © 1998 American Institute of Physics

aspects: (i) The *parallel* nature of the folding process, which implies that different portions of the chain might search *concurrently* for structural patterns, and (ii) the expediency of the folding process, which suggests the existence of a least-action principle underlying the parallel search for folding possibilities.

In general, the time scales involved in the folding process are incommensurably shorter than ergodic or thermodynamic time scales.[1–6] The question of how natural proteins fold as fast as they do remains elusive, mainly because most predictive approaches to the protein folding problem are rooted in the paradigm of equilibrium thermodynamics, which does not deal with the stringent timing of the process.[5]

We have asserted that the expediency of the folding process calls for a variational principle that should single out the dominant folding pathway.[9] But, how do we introduce such a principle for a process which is inherently parallel? This concurrence of folding events precludes any meaningful isolation of a single reaction coordinate at any given time. In order to address this question we shall first coarsely identify foldings of the protein as patterns of locally encoded signals. Within this frame, an underlying coarse Lagrangian or least-action formulation will be introduced over trajectories coarsely defined as sequences of pattern transitions. This Lagrangian will be conceived to single out the folding pathway that at each step minimizes the conformational entropy cost while maximizing the decrease in enthalpy. These tenets warrant a protein folding collapse following the most economic pathway.[1,9,10] Furthermore, even if a coarse description of conformation space might make longer time scales accessible to computation, ultimately, the problem of lifting the coarse dynamics to the level of actual torsional microscopic motions of the flexible chain should be addressed, as it is the case in this work.

The definition of the Lagrangian demands that we first introduce the appropriate coarse graining of conformation space upon which pathways are to be drawn. This leads to the first problem: The vast gap between the time scales accessible to molecular dynamics computations, typically in the range 10^{-12}–10^{-9} s, and those inherent to transitions between contact patterns (CPs) formed by the chain, typically in the range 10^{-4}–10^{3} s, suggests the need for a *semiempirical* model to elucidate the microscopic origin of the *expediency* of the folding process.[11,12] Then, our first problem becomes how to coarsely codify the local torsional states and local correlations and provide an effective dynamical picture introducing long-range correlations to account for the long-time behavior of the chain. Accordingly, the first goal of this work is to solve this problem by regarding foldings as patterns of locally encoded coarse structural signals.[13]

In consonance with this first goal, our working strategy may be sketched as follows.

(a) First, we introduce an ensemble of locally correlated two-state oscillators or spin flippers to coarsely simulate torsional isomerizations and search for *consensus* regions of torsional isomers along the chain. By consensus we simply mean regions of the chain where the local topological constraints associated to the formation of a particular folding are satisfied. In this way, a consensus window emerges as a pattern of structural signals encoded *locally* along the sequence. The broad latitude in local torsional coordinates within the Ramachandran, or local correlation maps of the protein chain,[14] and the vast structural distorsions it leads to implies that the binary codification cannot be implemented at the geometric level. Rather, the spin flippers are meant to mimic changes in the local topological constraints to which the flexible chain is subject in order to reach specific structural patterns.

(b) In a simplified version, a residue is classified using three information bits. The first two indicate the qualitative topographical features of its Ramachandran map,[14] defining the type of local torsional correlations (C), while the other bit indicates its hydrophobicity or polarity (h/p). This classification begets a fixed $3\times N$ matrix denoted the C-h/p matrix, with N being the length of the chain. Torsional topologies of an individual residue generated by the oscillators defined in (a) are rejected in the evaluation as part of the consensus if they are incompatible with the C class of the residue.

(c) We generate structural patterns as consensus regions within a matrix, the LTM, of local topological constraints (LTCs) of the peptide chain.

(d) We evaluate and translate such patterns into a contact matrix (CM) evolving within the time scale range 10^{-4}–10^{2} s. Thus, translation becomes a pattern recognition and, therefore, a *parallel* operation taking place at regular intervals. Each pattern within a LTM emerges with a certain probability that is effectively computed as the number of evaluations of the LTM that yield

the particular structural motif associated to the pattern divided by the total number of evaluations of the same LTM.

(e) The translation operation is subject to a feedback loop, whereby a renormalization operation readjusts the oscillator periods (or spin-flipping frequencies) according to the latest CP translated, and the contour ranges of intrachain interactions and contour distances are renormalized relative to the latest CP formed. In other words, *the renormalization operation introduces long-range correlations on the LTM* by slowing down or speeding up specific oscillators, depending on whether new interactions are formed or dismantled.

(f) Nucleation steps and the cooperativity in the formation of a secondary structure are accounted for by means of the renormalization operation: Suppose the LTM is evaluated at a given time and a short consensus region is detected. Then, the oscillators that generated this initial consensus window become endowed with frequencies that are lower than those of the neighboring residues and, consequently, the consensus region initially formed has a chance to grow upon successive evaluations of the LTM.

The fact that we characterize folding steps as CP transitions does not imply that any ''adiabatic assumption'' has been introduced *a priori*, in the sense that no enslavement or subordination of the fast-evolving microscopic degrees of freedom to CP transitions has been imposed.[15] Equivalently, CPs are not treated *a priori* as quasiequilibrium states by integrating out the relatively fast torsions as conformational entropy. Thus, the CM is generated by a parallel search for consensus windows in the LTM. In turn, the LTM is built upon the C-h/p matrix by adding two rows recording the phases of $2N$ two-state oscillators evolving within the broad period range 10^{-11}–10^{-3} s. Defined in this way, a LTM represents a coarse microscopic realization of a CP such that the consensus windows reflect the fulfillment of the LTCs determined by the CP. This sketch of the operational tenets reveals that, although advantage is taken of the fact that there exists a wide separation between characteristic time scales associated to folding events (typically in the range 10^{-4}–10^{3} s) and chain torsions (typically in the range 10^{-11}–10^{-7} s), no ''adiabatic assumption'' is introduced.

In this way, a computational strategy will be devised to provide theoretical underpinnings of the folding dynamics emerging as the evolution of patterns of locally encoded signals whose coherence reflects both cooperativity and nucleation effects. The key features of our approach are (a) the two-state coarse codification of local topological constraints of the flexible chain; (b) the renormalization of time scales for torsional isomerizations and their correlation decays relative to the successive stages of folding, thus introducing long-range correlations due to large-scale motions; (c) The identification of structural patterns with consensus regions in which specific topological constraints are fulfilled; and (d) the vast range of time scales 10^{-11}–10^{-2} s covered[2] (see Secs. III and IV) to account for the microscopic origin of cooperativity. In addition, our semiempirical approach has predictive potential, as revealed in Sec. IV, where we contrast a sequence of CP transitions with experimentally probed folding pathways involving secondary and tertiary interactions formed cooperatively.

At this point we may return to the original problem posed. Our representation of foldings as patterns of structural signals encoded locally along the peptide chain allows us to identify the most economic pathway defined as a sequence of CP transitions. This pathway entails at each step the minimal cost in conformational entropy while forming as many contacts as possible.[9,10] This fact will enable us to readily define a Lagrangian underlying the generation of trajectories regarded as sequences of LTM transitions. The variational perspective on the protein folding problem is developed in Sec. V. The LTM action leads naturally to a stochastic formulation of the problem, where the LTM transition probabilities are determined by the action, as indicated in Sec. V. This probability, in turn, begets a measure in pathway space that is typically very concentrated—almost Dirac delta-like—for folding chains that are the products of natural selection, as revealed in the detailed example expounded in Sec. IV. Furthermore, stochasticity is, in principle, inherent to our representational context of parallel computing, since different—and even conflicting—patterns in the same LTM may be simultaneously recognized by different processors, leading to competing folding pathways. The probability of recognizing a specific pattern is computed as the number of processors that recognize that pattern divided by the total number of processors involved in pattern recognition at that time.

Sections VI and VII are devoted to the elucidation of the microscopic torsional dynamics of

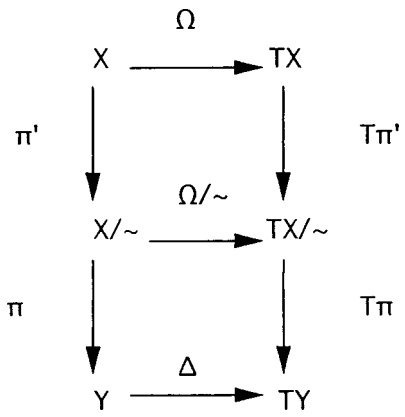

FIG. 1. Commutative scheme reflecting the compatibility of different coarse grainings of the folding dynamics.

the chain as stemming from the coarse solution to the folding problem. Thus, the coarse folding dynamics are lifted to the microscopic level, making the relevant long-time microscopic dynamics computationally accessible for the first time, at least to the best of our knowledge.

II. THE LEVELS OF COARSENING OF THE FOLDING DYNAMICS

The interplay between the different levels of coarseness with which the folding dynamics are simulated in this work require a clear representation scheme, such as the one presented in Fig. 1. Our main concern in this section is to rigorously define what is meant by a consistent treatment of the folding problem in which different coarse levels of description of the chain dynamics are compatible with each other. Standard notation has been adopted;[14] thus the conformation of each residue along the chain is specified by the two torsional or dihedral angular variables ϕ and ψ, each of which is taking values in a circle.

Let X denote the $2N$ torus (Cartesian product of $2N$ circles) corresponding to the full-detailed or (ϕ,ψ)-conformation space for a chain of length N. Then, at the richest level of the description, the chain dynamics is given by a map $\Omega: X \rightarrow TX$, where TX is the tangent space of X, that is, the space of all possible vector fields over X. Very limited time ranges for the full-detail dynamics (1–100 ps) have been explored by molecular dynamics (MD) simulations.[12] Since the time range effectively covered by MD is far shorter than the time ranges of interest to the folding context, we shall consistently introduce coarse descriptions of X, as indicated in Fig. 1. However, *our ultimate aim in this work is to define the dynamics over X. But we shall do so by first noting the existence of a variational principle at a coarser level of description, and, subsequently, lifting this variational formulation to X.*

Let $X/\sim$ be the quotient space consisting in the set of equivalence classes of (ϕ,ψ) isomers modulo a relation ''$\sim$,'' that is, conformations defined modulo their local torsional isomers, regardless of their difference at a finer level of detail. In accord with known stereochemical constraints, there are only two distinctive torsional species or isomers for each torsional variable. Two conformations of the entire chain are regarded as $\sim$ equivalent if their ϕ,ψ values for each residue lie within the same regions of the circle corresponding to the same local torsional isomers. Since there are two torsional isomers defined for each dihedral variable, $X/\sim$ may be represented as a lattice of $2\times 2N$ points drawn on the torus X. Actually, since each point represents one of the two basins of attraction in the unit circle, with each basin representing a torsional isomer, we get the isomorphism $X/\sim \approx (\mathbf{Z}_2)^{2N}$, where $\mathbf{Z}_2$ represents the cyclic field of integers modulo 2. This space is also isomorphic to the space of all possible LTMs, since local topological constraints are specified by regions within the circle for certain combinations of the canonical dihedral variables, as indicated in Sec. III. The dynamics at the LTM level are given by the projected map $\Omega/\sim: (\mathbf{Z}_2)^{2N} \rightarrow T(\mathbf{Z}_2)^{2N}$, where, as usual, ''$T$'' denotes the tangent space in this context.

Let Y denote an even coarser description of X. In this case, two conformations in X are regarded as equivalent if their respective CPs are identical, regardless of finer detail. Thus, at the CP level, the folding dynamics is given by a map $\Delta: Y \rightarrow TY$. Then, all three levels of description

of the folding dynamics are compatible with each other if and only if the diagram given in Fig. 1 is commutative. That means that the following relations between map compositions must hold:

$$\Delta\pi=(T\pi)(\Omega/\sim), \tag{1a}$$

$$(\Omega/\sim)\pi'=(T\pi')\Omega, \tag{1b}$$

where π is the canonical projection (in the computational context already described referred to as the translation operation) which associates to each LTM its equivalence class modulo CPs, π' is the canonical projection associating its LTM to each microscopic conformation belonging to X, and $T\pi$ and $T\pi'$ are their respective projections between tangent spaces, as indicated in Fig. 1.

The commutativity of the diagram would not hold valid unless the torsional variables divided in classes according to the potential basin where they belong (*cis-trans* isomers) are regarded as topological rather than geometric constraints: The vast latitude (30°–60°) in local torsional correlations within specific *cis* or *trans* basins[14] would lead otherwise to vast geometric distortions that would make specific structural motifs unrecognizable at a more detailed level. This point is developed in Sec. III. Likewise, excluded volume effects on the chain would lead to severe structural distortions for a putative element of a class modulo $\sim$, taking that microscopic realization to another $\sim$ class, and the diagram would fail to commute unless the elements of the space $X/\sim$ would again represent sets of topological rather than geometric constraints.

To summarize, the commutativity of the diagram displayed in Fig. 1 defines the consistent theoretical framework dealt with in this work: All three descriptions of conformation space, the one with the richest detail, X, the LTM description, $X/\sim$, and the CP description Y will represent mutually compatible frameworks within which the folding dynamics will be studied.

III. SIMULATING FOLDING DYNAMICS AT THE LTM LEVEL AND AT THE CP LEVEL

The essential operation in our parallel model and algorithm, the translation operation, lies within a renormalization feedback loop and consists in the transference of information cast in terms of the state of internal microscopic degrees of freedom into a coarser representation, a CP defined by a CM. Thus, patterns of locally encoded structural signals contained in the LTM are translated as foldings of the chain. The CM is, in turn, built upon the map of putative hydrophobic interactions, or h/p map. The algorithm and its underlying semiempirical model identify the CP class to which a LTM belongs at certain time intervals according to a set of renormalizable rules. The generation of the LTM is, in turn, affected by the last CP transition that has taken place, since a new set of constraints arises with each new CP formed, and the prescription for the evaluation operation itself is renormalized and thus depends on the last CP that has occurred. Prior to defining the folding process explicitly within our model, we must specify the basic representational elements and their interrelationships with regard to the basic operations:

A. The local topological two-state (or spin) variables η, ξ versus the geometric torsional variables ϕ,ψ

A basic tenet underlying our approach is that structural patterns are identified through the translation of consensus windows of a spin state matrix, the LTM. The consensus represents the fulfillment of local topological constraints required to form the structural pattern. In turn, the recognition of the structural pattern determines a renormalization of the LTM and of its evaluation code. Thus, it becomes convenient to define two-spin or two-state variables η, ξ defining local topological features. The identification of these variables requires a ribbon-like representation of the chain made up of tetragons with hinges at each α carbon (see Fig. 2), each of which contains the torsionally rigid CO–NH group in its planar trans conformation, a standard constraint of the system.[14]

The first variable η, also named the vertebral variable, is defined with respect to a specified *local* orientation, or + side, of the ribbon independently of twistings (see Fig. 2), so that $\eta_i=+1$ indicates local concavity at the ith α carbon of the chain (or hinge in the ribbon representation) with respect to the local + side of the ribbon, and $\eta_i=-1$, local convexity at the same contour location.

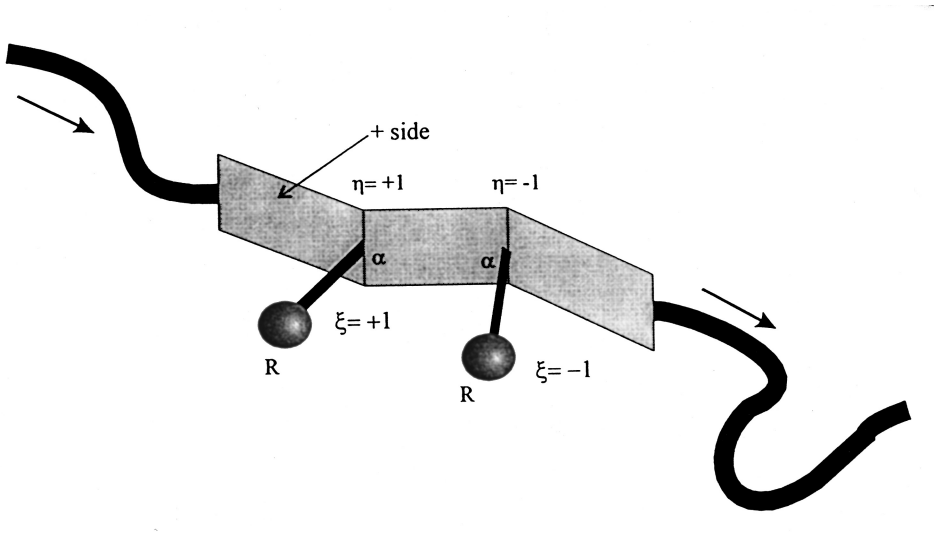

FIG. 2. Local topology of a ribbon-like model of the peptide chain, as represented by the two two-state variables η, ξ. The local + side of the ribbon faces the background and the balls labeled ''*R*'' represent the side chains attached to the hinge-like α carbons indicated by ''α.'' Each tetragon contains the torsionally rigid CO–NH group.

On the other hand, by spanning two adjacent tetragons into planes intersecting along a line containing the α carbon, two hinge regions are locally defined: an inner or concave domain and an outer or convex domain with respect to the + side. Thus, $\xi_i = +1$ indicates the fact that the side chain attached to the ith α carbon is contained in the convex region, while $\xi_i = -1$ indicates that it is contained in the concave domain, as indicated in Fig. 2.

B. The local topological matrix (LTM)

This two-state-entry matrix represented in Fig. 3 has five rows and N columns, with the ith column, $i = 1,2,...,N$, representing the residue with contour number i. The first two rows indicate, respectively, the local topological variables η, ξ. The correlation pattern C for each residue requires two bits of information and thus is indicated in the third and fourth rows, as indicated in Fig. 4. Thus, we get four kinds of correlation diagrams depending on the type of residue: one for glycine, one for proline, one for residues preceding a proline, except when that residue is glycine, and one for L-alanyl-like residues.[14] The hydrophobicity or polarity, denoted, respectively, with the $+1$ and -1 state, is indicated in the fifth row, which constitutes the h/p map. Conventionally,

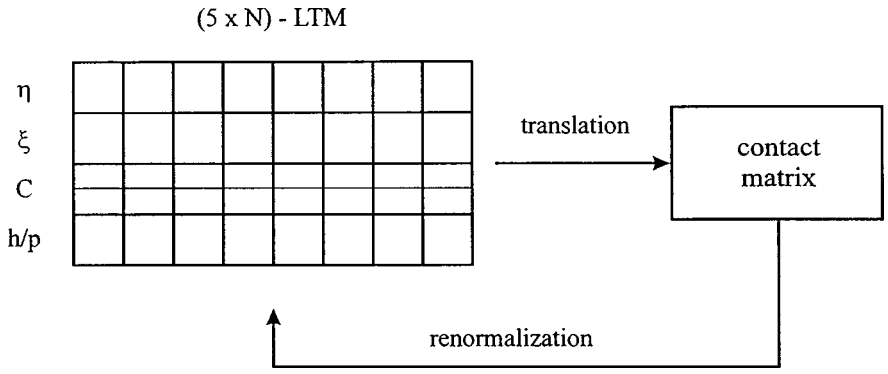

FIG. 3. The translation–renormalization feedback loop. It recognizes structural patterns as consensus windows in the local topological constraints matrix (LTM) and records them in the contact matrix (CM) as contact patterns (CPs). The LTM is built upon the primary sequence that determines its last three rows (the two C rows and the h/p row), which are fixed. The purpose of the renormalization operation is twofold: Taking a CM as input, it redistributes the frequency domain of the oscillators generating new LTMs and redefines contour distances *vis-à-vis* the latest CP generated, thus represcribing the evaluation of the LTM.

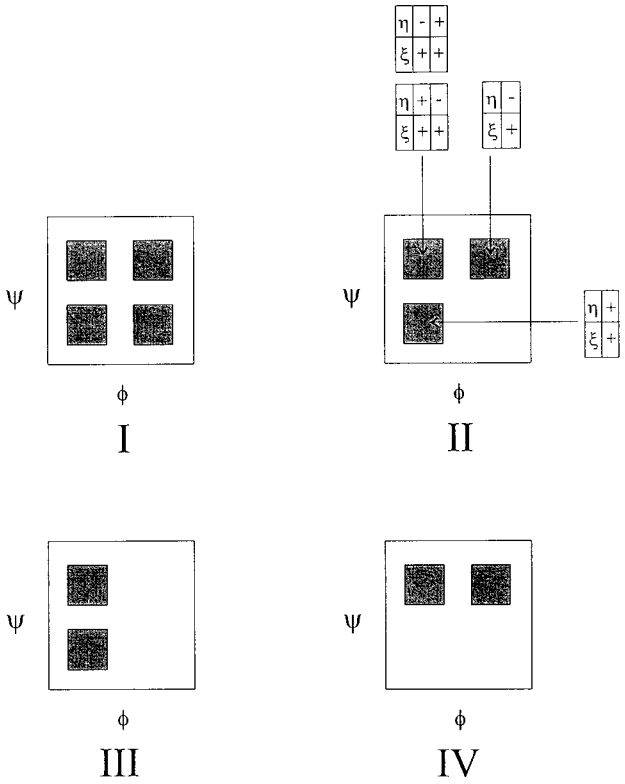

FIG. 4. Local topological constraints associated to the four classes of correlation diagrams based upon Ramachandran maps. In the caricatures, the regions of high probability have been shaded. The diagrams I–IV are labeled, respectively, $(-+)$, $(++)$, $(--)$, and $(+-)$ in the C representation. They correspond, respectively, to Gly, L-alanyl-like, Pro, and a residue preceding Pro different from Gly. Three of the plausible correlated zones correspond to structural motifs that may be characterized in terms of (η,ξ) states with $\xi=+1$. Thus, local torsional correlations are characterized by their local topological constraints.

a residue that is not hydrophobic is assigned the -1 sign. The structural patterns are progressively registered in the CM according to a renormalization feedback loop schematized in Fig. 3.

C. Folding as pattern recognition

The dominant secondary structure motifs can be identified in the LTM. Thus, the right-handed α helix requires a window of continuous concavity, that is, the vertebral variable should not change sign within the window, while the side chains are all oriented toward bulk solvent, or outward with respect to the helix interior. Without loss of generality and for the sake of notation, we shall identify this motif by a window of the LTM with $\eta=+1$, $\xi=+1$, while each local $\eta_i=+1$, $\xi_i=+1$ isomer is compatible with the ith Ramachandran correlation diagram C_i, and the h/p map shows a periodic $(i,i+3)$ or $(i,i+4)$ hydrophobicity [if $\text{sign(h/p)}_i=-1$, then the next residue for which $\text{sign(h/p)}=-1$ is $i+3$ or $i+4$]. Because of the highly helix-disruptive tendencies of glycine,[12,14] if its correlation diagram appears within a consensus window, the entire pattern is obliterated from the CM. The disrupting tendencies of the proline, on the other hand, do not require special instructions as the $(\eta_i=+1,\ \xi_i=+1)$ pattern is incompatible with a residue preceding proline, as shown in Fig. 4.

A consensus window translating into the α helix (R) is indicated in Fig. 5(a). Similarly, for a left-handed α helix, we must demand permanence of the -1 sign in the vertebral variable, while retaining all other conditions regarding side chain orientation, C compatibility, and hydrophobic periodicity.

Likewise, being pleated structures, β sheets are characterized by an alternation of the vertebral sign, while the side chain lies always within the convex region. In order to fulfill topological as well as h/p compatibilities, the η, ξ, and h/p sign pattern of a window must be preserved in other windows in a parallel or antiparallel fashion, depending on the relative orientation of the strands

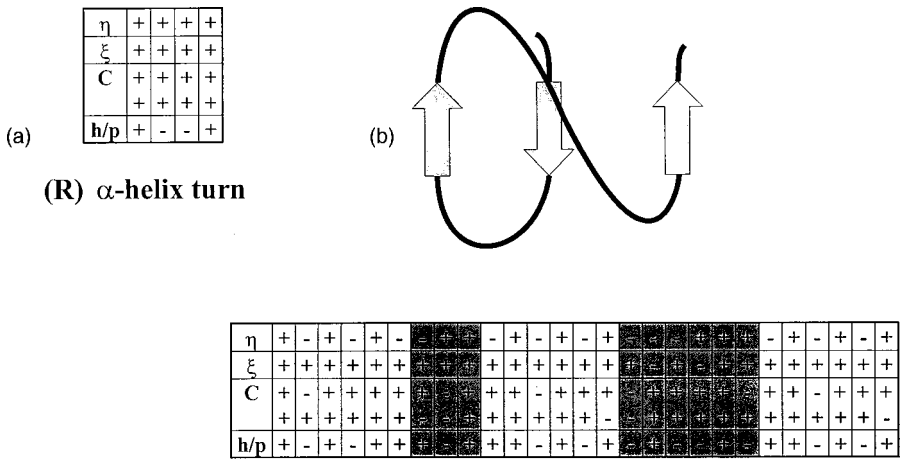

FIG. 5. (a) A consensus window for a right-handed α-helix turn. (b) A consensus window for the complex three-strand antiparallel β-sheet motif shown. The β bend and the turn consensus windows have been shaded in the LTM.

in the β sheet. For the sake of illustration, a three-strand β-sheet topology is displayed in Fig. 5(b), together with its LTM consensus pattern. A structural pattern in the same topology class[16] will be generated in our time-dependent simulation of the folding of bovine pancreatic trypsin inhibitor (BPTI),[17] as shown in Sec. IV.

Turns and bends may be a determinant of the β sheet or simply required to form hydrophobic contacts, thus, they will be treated generically, regardless of whether or not they realize β-sheet topologies. Should such motifs require closure of chain loops, they would pose demands in terms of LTM consensus, as indicated below.

Closure of a loop formed by a rod-like object determines two solvent environments: an inner cluster-like and an outer bulk-like domain. Since polar groups tend to be oriented toward the high dielectric medium, no orientational constraints are imposed upon loop closure when the inner cluster becomes like bulk solvent,[18] that is, it allows for four solvation layers on the polar groups.[19] Given the dimensions of the polypeptide chain, and the fact that the loop must accommodate seven water molecules across (the cluster diameter in a planar loop), a rough calculation gives a critical loop length of $N_c \approx 22$ for proteins. This means that a loop of size $N(\text{loop}) = N \geqslant 22$ poses no orientational demands on the polar groups.

These situations easily translate into patterns of the LTM. A bend or turn requiring loop closure is characterized by the following features: (1) a permanence of sign in the η row (this

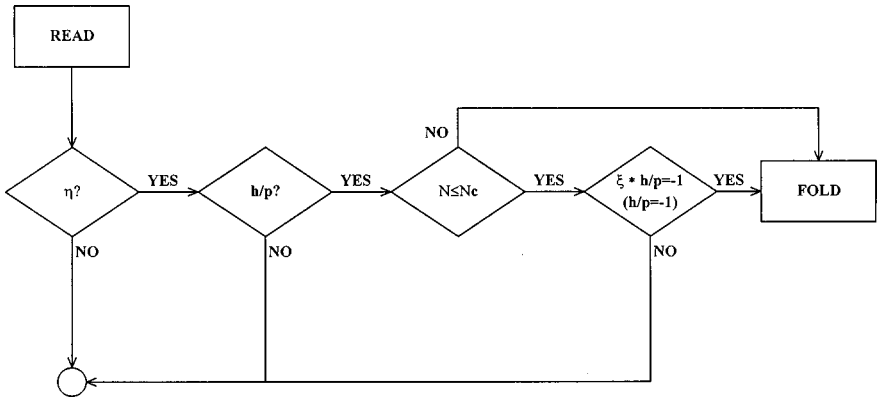

FIG. 6. The flow chart for evaluation of LTM-consensus representing a loop closure of the peptide chain. Evaluation asks first for fulfillment of vertebral constraints (η?), then for the possibility of hydrophobic contacts at the loop extremes (h/p?), then whether the size of the loop is larger or smaller than the critical size ($N \leqslant N_c$), and, finally, whether polar side chains are oriented toward bulk solvent for $N(\text{loop}) = N \leqslant N_c$ ($\xi \times (\text{h/p}) = -1$ for $(\text{h/p}) = -1$).

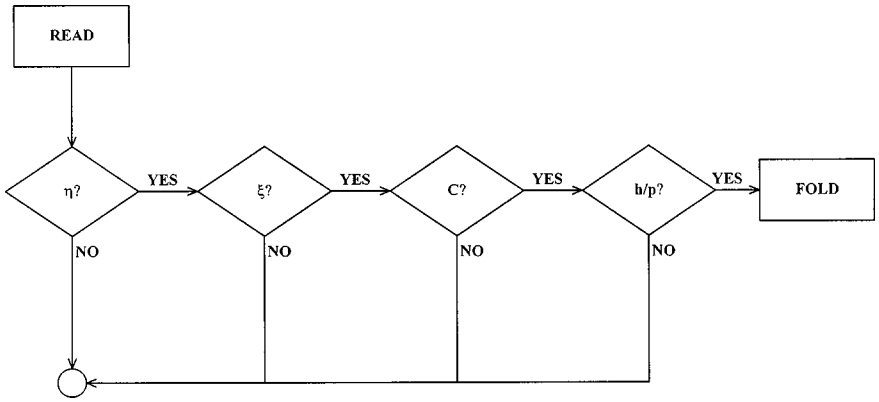

FIG. 7. The flow chart for the recognition of consensus windows as structural patterns. Evaluation asks first for fulfillment of a vertebral consensus (η?), then for the fulfillment of side-chain orientational constraints (ξ?), then for C compatibility (C?), and finally, for the occurrence of hydrophobic/polar periodicities or correspondences (h/p?).

condition may be somewhat relaxed for very large loops with $N(\text{loop}) \gg N_c$); (2) the outward orientation of polar groups in small loops becomes $[\text{sign(h/p)}]_i \times \xi_i = -1$ when $\text{sign(h/p)}_i = -1$ and $N(\text{loop}) < 22$; (3) the η,ξ patterns for individual residues must conform to their correlation diagrams when $\xi = +1$. On the other hand, crossover turns, such as the one featured in the complex β-sheet topology displayed in Fig. 5(b), pose no orientational demands on the polar groups, since they do not define local solvent environments. The only requirements of such turns are the sign inversion of the vertebral variable, which marks the existence of an inflection point [see Fig. 5(b)] and the C compatibility of the local topological patterns.

The structural patterns for bends and turns entailing loop closure are recognized according to the flow chart displayed in Fig. 6. On the other hand, once bends and turns have formed at the highest expense in conformational entropy (see below), the secondary structural patterns are identified according to the flow chart displayed in Fig. 7. This latter chart has been developed following all the operational tenets for the pattern identification described above.

D. The distribution of flipping periods in the generation of the LTM

In attempting to project torsions into the CP space, where the folding process is recorded, the time scale limitations of molecular dynamics simulations must be circumvented. This explains the need to introduce models in which activated molecular motions in the 10 ps to 1 MS range are considered.[12] Thus, faster diffusion-like unhindered torsions in a free residue are integrated out as conformational entropy of the state defined by the coarse LTM representation. Such motions, in the picosecond (ps) range, determine the rod-like shape of the protein molecule when viewed within the time scale window between two LTM states. Precisely, this interrelation between shape and time scale justifies the concept of the inner and outer solvent domain defined by an intramolecular loop, as put forth in our consensus analysis of the entropic cost of loop closure. Thus, the microscopic mean time range relevant to LTM transitions is 10^{-11}–10^{-3} s, covering the time scale for internal motions (10^{-11} s) of the order of the calculated diffusional displacements of flexible hinged domains,[12] and, at the other end of the spectrum, the limiting values (10^{-4}–10^{-3} s), typical mean time scales for the fast exchange between folded and unfolded states with respect to tertiary interactions engaging two secondary elements.[17] Within this range, we encompass the mean time frame of 10^{-7} s, typical for a localized helix-unwinding event leading to a bubble.[12]

These considerations lead us to define a temperature-dependent normalized distribution of periods, $w = w(\tau)$ for the independent $2N$ η,ξ oscillators. This distribution features three Gaussian peaks centered at characteristic periods 10^{-11}, 10^{-7}, and 10^{-3} s. The periods of the oscillators representing the topological *cis-trans* flipping or actually the characteristic times for changes in the local topological constraints, are assigned from this distribution in such a way that the effect of thermal fluctuations on the formation of consensus, and thus on structural transitions, is incor-

porated. Each Gaussian peak has a dispersion $\sigma^2 = gT$, where the constant g depends on the actual denaturation temperature $T(\text{denat.})$ and on the consensus interpretation of denaturation.

The period distribution allows us to classify residues in three classes: Class I contains all free residues, that is, residues not engaged in any structural motif, with mean period 10^{-11} s; class II contains all residues with mean period 10^{-7} s engaged in secondary structure but not in tertiary interactions; and class III contains all residues engaged in tertiary structure, whose mean period is 10^{-3} s. This classification of residues according to their inherent oscillator time scales is compatible with fluorescence depolarization probes for unhindered torsional motions,[12] with typical time scales for localized helix disruptions, and with diffusion–collision models,[20] in which secondary structure is stabilized further by forming tertiary contacts.[17]

In this way, long-wave or slower modes with mean frequencies in the range $10^7 - 10^3$ Hz result from the coupling of several local (η,ξ) modes which originally were high-frequency modes with a mean frequency of 10^{11} Hz. This is so since the formation of secondary or tertiary patterns involves the recognition of a consensus region, and thus the concurrent prescribed slowing down of the oscillators involved corresponds to the coupling associated to the consensus region.

Accordingly, the LTM evolves through a lottery from which periods of the oscillators for free residues are assigned from within a period range centered at 10 ps. A new period is assigned from the lottery to each oscillator each time a whole previously assigned period has been completed. The frequency $f = 1/\tau$ of an oscillator in the first two rows of the LTM corresponding to a residue not engaged in an intrachain interaction or loop satisfies the inequality

$$|f^{-1} - 10^{-11}\ \text{s}| \leqslant |\tau' - 10^{-11}\ \text{s}|, \quad \text{with } \tau' \text{ satisfying } w(\tau') = \text{Infimum}_{\tau}\{w(\tau) \geqslant 1/[2N]\}. \quad (2)$$

The condition yielding the extreme period τ' arises from the fact that there are, at most, $2N$ free oscillators in the peptide chain.

The other two peaks in the distribution correspond, respectively, to mean periods for residues in secondary and tertiary structural elements. Again, the same considerations apply in regards to the period assignment to oscillators for class II and class III residues. These rules imply that the residues in loops, bends, or turns concurrently formed with any secondary or tertiary structural element adopt the cadence of the structural element itself.

The mechanistic aspects of period distribution imply that this operation is subject to renormalization with each CP transition: A CP determines which columns in the LTM correspond to free or class I residues and that correspond to residues engaged in an intrachain interaction, or, equivalently, they belong to class II or III. Thus, since a CP transition reclassifies the residues, it also dictates the range from which new period assignments are drawn. The period range for a specific residue remains the same as before the CP transition if the transition does not alter the class of the residue, and changes if the CP transition transfers the residue to a different class. Thus, the renormalization operation is essentially the means by which long-range correlations in the LTM are introduced as a consequence of the latest large-scale motions.

E. The built-in internal clock

In order to satisfy the basic tenets of Shannon's information transmission while making all operations synchronized, the intrinsic beat period of the folding machine should be taken to represent $1/2\tau'$, one-half of the shortest possible period to be assigned to an oscillator. Since τ' depends on $w(\tau)$, it is itself T dependent. The time interval between two consecutive readings of the LTM, $\tau(\text{read})$, is taken as constant and, in order to lower the operational cost, it is fixed at the shortest time that a CP transition could possibly take. Thus, $\tau(\text{read}) = 2^{3+1}\tau' \approx 64$ ps at 298 K, the shortest time to close a three-loop with a single hydrophilic side chain in order to form a hydrophobic contact engaging the fastest oscillators at a typical folding temperature.

F. The temperature (T) in the semiempirical model

Secondary structure dismantling materializes and is recorded as such by deletion in the CM whenever a consensus bubble forms among class II residues engaged in the structure. By ''consensus bubble'' we mean that in any of the first two rows, a consecutive sequence of oscillators of length 30% of the total consensus window length[14] must be out of phase with the consensus

required values at the time when the reading of the LTM takes place. Because of the renormalization loop, this transition at the CM level immediately transfers a new set of constraints for the generation of new LTMs: The residues previously engaged in the structure and in its concurrent loops are reclassified, being transferred from class II to the higher-frequency class I.

Cooperative effects reflect themselves mechanistically in the formation of the consensus bubble: For instance, in the α-helix motif, the larger the helix, the more improbable to find a 30% out-of-phase subsequence of class II oscillators. Furthermore, these considerations enable us to estimate the constant g that determines the effect of thermal fluctuations on the period distribution: At the denaturation or melting temperature $T(\text{denat.})$, every helix formed must develop a consensus bubble evaluated and recorded with the next reading of the LTM. Thus, if σ is ''large enough,'' the period distribution in the helix is broad enough so that consensus cannot be preserved: The period range, of the order of σ, is such that a helix consensus cannot survive two consecutive readings. From these considerations, and taking into account our empirical estimate of the denaturation dispersion fixed at $\sigma = 10^{-8}$ s, and the typical experimental $T(\text{denat.}) = 303$ K for proteins such as the ones studied in this work[17]), we get $g \approx 3.2 \times 10^{-23}$ s^2/K.

G. Nucleation effects and protein folding cooperativity

Structural patterns such as the one presented in Fig. 5(b) do not result from all-or-none processes. Rather, a nucleating event involving the largest cost in conformational entropy takes place first and is subsequently followed by a sequence of folding events that seek to minimize additional losses in conformational entropy. Two theoretical approaches to biopolymer folding, the variational treatment rooted in the sequential minimization of entropy loss (SMEL),[9,21] and the theory of hydrophobic zippers,[10] take into account the cooperative nature of the formation of structural motifs triggered by nucleating events in their attempt to account for entropically driven pathways.

In our model, cooperativity and nucleation work through successive renormalization loops as expedients for the folding of peptide chains: Suppose a nucleating pattern has been identified in a reading of the LTM, be it a helix turn such as the one depicted in Fig. 5(a), or a turn involving a loop closure. Then, the residues engaged in the respective nucleating consensus window are automatically locked into lower frequencies of the order of 10^7 Hz ($1\ \text{Hz} = 1\ \text{s}^{-1}$) via the renormalization operation that immediately follows the translation of the nucleation pattern into a CP registered in the CM. Thus, the nucleation consensus will ''survive'' or be preserved for at least approximately 200 evaluations of the LTM, since the evaluation frequency is of the order of 10^9 Hz at 298 K [cf. Eq. (2)], the typical working temperature for *in vitro* folding of BPTI. This time span allows for a chance to enlarge the original consensus window by progressive torsional isomerizations of residues adjacent to those engaged in the nucleating step. Once adjacent residues form a consensus compatible with the original one in relation to a specific structural motif, they are also locked into the 10^7 Hz frequency domain and thus they contribute to the seeding process.

Formation of complex motifs such as the one shown in Fig. 5(b) might be triggered by more than one nucleating event taking place asynchronically, such as the closure of the two loops that ultimately generate the two pleated antiparallel pairings in this β-sheet topology. The asynchronous occurrence of more than one seeding event is compatible with the parallel nature of the renormalization operation, as illustrated in Sec. IV.

Summarizing, the key feature of renormalization that enables us to deal with cooperativity in the formation of a secondary structure is the survival of a nucleating pattern warranted by the reclassification of class I residues as class II residues once the nucleating pattern occurs. The argument holds *mutatis-mutandis* for tertiary interactions formed cooperatively from a secondary structure. In this case, the reclassification of class II residues as class III residues translates into a drastic shift in the oscillator frequency, from 10^7 to 10^3 Hz, thus warranting the survival of the respective nucleation pattern.

H. The renormalization operation and renormalized contour distances

As stated above, the renormalization operation is the expedient by which long-range correlations are introduced. It involves two concurrent roles: Redefining contour distances between different residues relative to the latest CP translated, thus determining how to read and evaluate

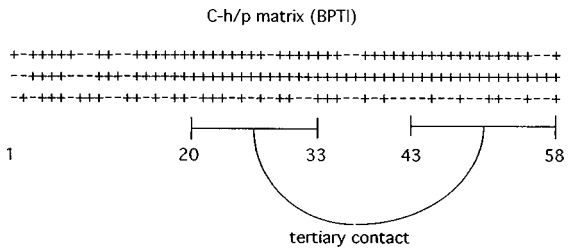

FIG. 8. The C-h/p matrix for a bovine pancreatic trypsin inhibitor (BPTI). The notation is consistent with that adopted in Figs. 3 and 5.

the LTM, and placing a new set of constraints on the generation of new LTMs after a CP transition has been recorded as a change in the CM. Both roles are specified at a given time according to the latest CP generated. The renormalization of constraints for the oscillators is articulated through the operational feedback loop (see Fig. 3), introducing the latest CP generated as input for the reclassification of residues. On the other hand, the renormalization of contour distances points to the root of cooperativity, since unfavorable long-range interactions with $N(\text{loop}) \approx 22$ or $N(\text{loop}) \gg 22$, might eventually materialize with lower entropic costs, as illustrated in Sec. III. Such situations take place when nucleating events bring the renormalized loop lengths, respectively, into the favorable ranges $N(\text{loop}) \ll 22$, or $22 < N(\text{loop}) < 30$, the upper bound being the critical size beyond which excluded-volume effects start becoming important.[19]

IV. RESULTS: THE CM PATHWAY FOR A BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI)

The aim of this section is to predict the dominant structural features of BPTI by a mechanistic analysis of its long-time dynamics, resulting from the simplified model of torsional motion at 298 K.[7,17,22] We shall focus on determining significant folding intermediates and the late kinetic bottlenecks that occur within the first 10^{-2} s of the renaturation process, a time span that requires 10^7 evaluations and renormalizations of the LTM. The C-h/p matrix for BPTI is displayed in Fig. 8. Our aim is to show how the dominant sequence of CM transitions for the BPTI describes the dominant folding kinetics, reproducing the essential cooperative features of the experimentally probed folding pathways, including the late scenario in which tertiary interactions direct and stabilize the native Cys–Cys (5,55) contact (cf. Refs. 7, 17). The actual refolding conditions involve special thiol and disulfide concentrations, so that the kinetics of formation, dismantling, and recombination of intrachain Cys–Cys disulfide bonds lies within the 10^{-5} s time scale,[7,22] and does not interfere with the folding process. This is so since the fastest Cys5–Cys30 interaction takes 4×10^{-5} s to form, as shown below.

In carrying out the simulation, special precautions have been taken to relax the conformational entropy demands associated with polar orientations in closed loops near the critical size. First of all, an operational definition of contact based on a proximity of 7 Å has been adopted in the CM development, in consonance with previous treatments.[10] Second, loops need not be, in general, planar, thus the orientational demands beyond the critical loop size must be relaxed smoothly and not abruptly: The entropic contribution, $-T\Delta S(\text{loop})$, will be linearly reduced in the interval $22 \leqslant N(\text{loop}) \leqslant 25$, so as to yield the following mean times of formation:

$$\tau(N(\text{loop}) = 22) = 2^{22 + N(\text{polar},22)} \times 10^{-11} \text{ s}, \tag{3a}$$

$$\tau(N(\text{loop}) = 25) = 2^{25} \times 10^{-11} \text{ s} \approx 3.2 \times 10^{-4} \text{ s}, \tag{3b}$$

where $N(\text{polar},22)$ is the number of polar side groups of a particular loop of size 22, and $\tau(N(\text{loop}))$ is the mean time of formation of a loop of size $N(\text{loop})$. Thus, the orientational demands are removed completely for $N(\text{loop}) > 25$, while they drop gradually and linearly within the interval $22 \leqslant N(\text{loop}) \leqslant 25$.

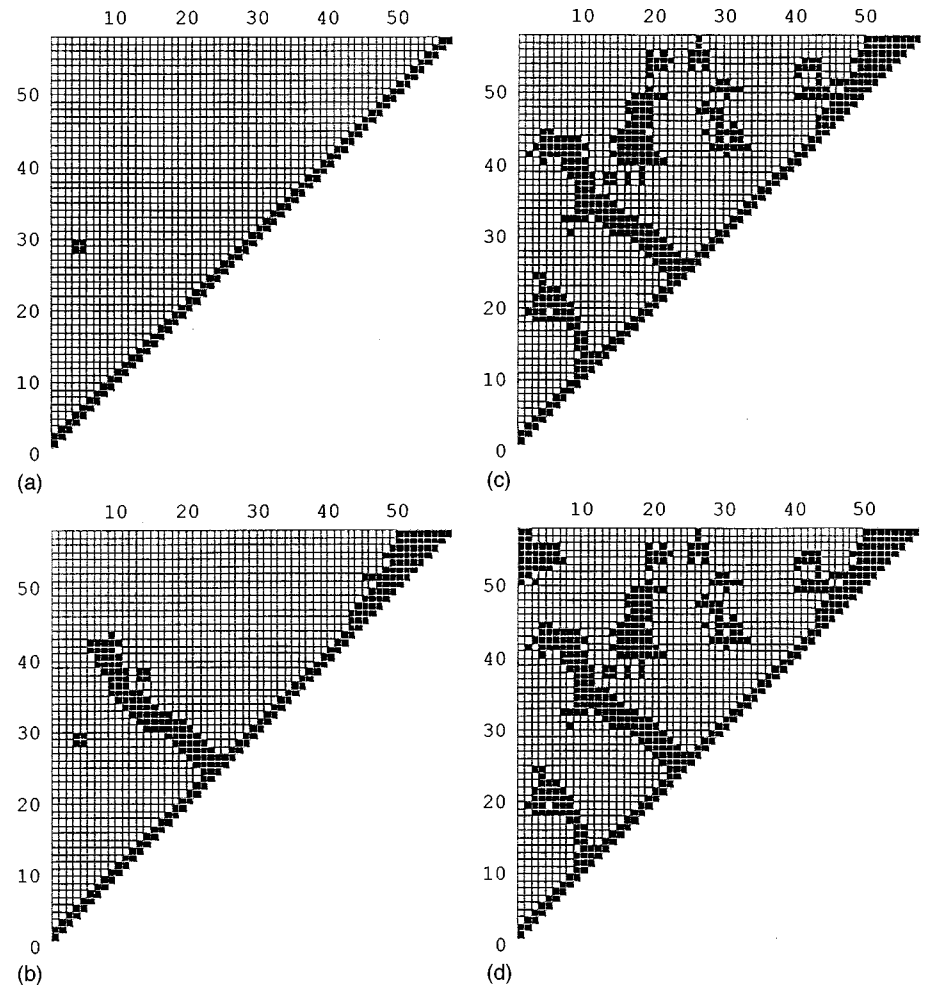

FIG. 9. (a)–(d). Four snapshots of the CM taken, respectively, at 3.2×10^{-4} s, 1.3×10^{-3} s, $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}$ s, and $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}\,\mathrm{s}+0.5\times10^{-2}$ s in the course of 10^7 translation–renormalization loops for the BPTI.

How does cooperativity operate in the folding of BPTI? To answer this question we can examine the estimated mean times to form native Cys–Cys disulfide contacts starting from a random coil conformation:

$$\tau(5,55)=2^{50}\times10^{-11}\ \mathrm{s}\approx10^{4}\ \mathrm{s};\quad \tau(30,51)=2^{21+13}\times10^{-11}\ \mathrm{s}\approx1.6\times10^{-1}\ \mathrm{s}, \tag{4}$$

where the number of polar side chains between residues 30 and 51 is $N(\mathrm{polar})=13$. Direct inspection of Eq. (4) reveals that such native contacts take a long time to form and can only be created cooperatively within the time scales under investigation. On the other hand, the third native Cys–Cys contact (14,38) may form directly within time scales commensurate with the occurrence of tertiary contacts:

$$\tau(14,38)\approx2^{24+1/3\times10}\times10^{-11}\ \mathrm{s}\approx1.3\times10^{-3}\ \mathrm{s}, \tag{5}$$

where $N(\mathrm{polar},(14,38))=10$ for the BPTI.

The actual sequence of CMs taken within the time span of 10^7 evaluations of the LTM is consistent with the previous analysis. Four significant snapshots of the time evolution of the CM have been taken and are displayed in Figs. 9(a)–9(d). The corresponding times are, respectively, 3.2×10^{-4} s, 1.3×10^{-3} s, $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}$ s (the third snapshot is taken 320 LTM readings after the second), and $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}\,\mathrm{s}+0.5\times10^{-2}$ s. A direct analysis of the CM evolution indicates that the non-native (5,30) disulfide bond forms first because a loop of size 25

poses no orientational constraints on its polar groups. Thus, we get the estimate $\tau(5,30) \approx 2^{25}\times 10^{-11}$ s$\approx 3.2\times 10^{-4}$ s. This is a good estimate, since the (5,30) contact indeed shows up on the CM snapshot taken precisely at that time, as featured in Fig. 9(a).

Nucleation windows of the form shown in Fig. 5(a) seeding the formation of an α helix appear in the (43–58) extreme of the molecule within the range of time scales 8.8×10^{-4}–9.8×10^{-4} s. The time scale of formation of (14,38) given in Eq. (5) is also a good estimate. The snapshot taken at 1.3×10^{-3} s [Fig. 9(b)] displays this native contact, as well as fully and partially developed secondary structure elements such as the α helix and a two-strand portion involving the contour region (20–33) of the β-sheet topology shown in Fig. 5(b). The nucleating events leading to the two-strand portion of the β-sheet topology take place in the (20–33) region of the chain within the same time interval as those triggering the formation of the helix.

Tertiary contacts between the α helix and the complex β sheet require the closure of the loop in the contour region (33–43) and start developing between the incomplete β sheet and the helix 320 LTM evaluations after the time when the last snapshot displayed in Fig. 9(b) was taken. This coincides precisely with the time estimation for closure of the ten loop with five polar groups within the (33–43) region of the BPTI:

$$\tau(\text{tertiary interaction}) \approx 2^{10+5}\times 10^{-11} \text{ s} \approx 3.2\times 10^{-7} \text{ s}. \tag{6}$$

We emphasize that this is a renormalized calculation that assumes the previous formation of a secondary structure.

At this point, the non-native (5,30) Cys–Cys bond is completely dismantled to give rise to the native (30–51) contact induced and stabilized by the tertiary interaction, as a direct observation of the third snapshot [Fig. 9(c)] reveals. This pathway reveals how the formation of (30,51) is expedited by cooperative folding, in agreement with recent findings.[7] Furthermore, during the time span of development of the first tertiary contact, the complex β-sheet motif continues to develop, fostering other tertiary interactions between the two dominant secondary structure elements. Since folding and unfolding of tertiary structure exchanges on the 10^{-3} s fast NMR time scale, the locking of oscillators at the mean 10^3 Hz frequency peak warrants the survival of the initial tertiary consensus while contact (30–51) forms and the β sheet is completed.

Finally, the (5,55) disulfide bond that would initially take a forbiddingly long time to form, now entails the closure of a complex 29 loop with no polar orientation demands: This loop is made up of the quasicoil (5–20) region, the β-sheet (20–30) region, and the quasicoil strand (51–55). Notice that the formation of the native (30,51) contact has short-circuited the loop closure for the (5,55) contact, so that the estimated time for this interaction is

$$\tau(5,55) \approx 2^{29}\times 10^{-11} \text{ s} \approx 0.5\times 10^{-2} \text{ s}. \tag{7}$$

This estimation of the rate-determining step in BPTI folding is corroborated by examination of the fourth snapshot [Fig. 9(d)], and confirms previous estimations,[7,17,22] in the sense that contact (30,51) occurs 10^5-fold more rapidly than (5,55).

Thus, the long-time dynamics obtained by means of our semiempirical microscopic model not only predicts with good accuracy the tertiary structural elements of the BPTI, but it also shows how cooperative effects can serve as an expedient to aid the formation of native interactions shaping the hydrophobic core.[17] A good agreement with experimental kinetic probes[7,17,22] has been found.

V. ($\pi-\rho$)-FOLDING PATHWAYS AND THE LEAST-ACTION PRINCIPLE

As indicated in the previous sections, structural patterns have been identified with consensus windows within which local topological constraints are fulfilled. It follows that the preferred folding pathway resolved as a sequence of CP transitions is the one for which each transition entails the minimal cost in conformational entropy, while maximizing the enthalpy decrease.[9] An example of such a pathway has been described in Sec. IV, where a single folding pathway happens to be dominant. On the other hand, the possibility of competing or multiple folding pathways is naturally incorporated in our computations, since different—and even conflictive—patterns may

be recognized in a single LTM. Thus, the probability of a single pattern is actually computed as the number of processors that recognize the pattern divided by the total number of processors involved in the evaluation of LTMs.

The highly dominant CM pathway computed in Sec. IV is the easiest to follow by the BPTI chain, and therefore by far the most probable as it entails the maximum economy of means for each step. This may be rationalized on account of two facts: (a) consensus, as identified in a LTM, implies the fulfillment of local topological constraints, which, in turn, determine the CP; (b) the enthalpic content is exclusively dependent on the contacts that have materialized.[9] Thus, whenever a single pathway dominates at the CM level, our semiempirical approach is compatible with the principle of sequential minimization of entropy loss (SMEL).[9,10,21]

In view of these facts, the following problem arises: Is it possible to define a Lagrangian L at the coarse microscopic level of LTM transitions that reproduces the SMEL behavior when a LTM trajectory is coarsely translated as a sequence of CP transitions? *In simple terms, is there a Lagrangian formulation to account for the expediency of the protein folding*? According to our treatment, the solution to this central problem involves determining $L=K-H$, where H is the enthalpy content, or potential in our context, and K is the ''kinetic energy'' of a LTM, $K=K(\mathrm{LTM})$, which must satisfy the following conditions.

(a) It must be uniquely defined for all LTMs associated to a specific CP, since otherwise many LTM trajectories, among which is the one singled out by L, would translate into the SMEL pathway.

(b) A CP transition is favored if it entails a minimal entropic cost; therefore its corresponding LTM transition must entail a minimization of the kinetic energy loss.

In accord with conditions (a) and (b), we define

$$K(\mathrm{LTM})=1/2\sum_{i=1,\ldots,2N} f_i^2, \tag{8}$$

where the f_i's, given in Hz, are the mean frequencies of the oscillators of the LTM and hence H must be given in Hz^2 units.

The choice given in Eq. (8) is a good one, as we shall demonstrate, making use of the fact that the oscillator frequencies of the LTM are determined by the renormalization operation: Let us suppose that a pattern of local signals has been identified and recorded in the CM at a given time by the translation operation. Then, the renormalization operation reclassifies the residues engaged in the latest CP generated, from class I into class II or, if a tertiary interaction has taken place, from class II into class III. The reader is reminded that residues belonging to loops whose closure is concurrent with secondary or tertiary interactions adopt the same cadence of those residues that form the respective contacts. Thus, *the loss in conformational entropy associated to the CP transition corresponds with a loss in kinetic energy of the LTM in such a way that a minimization of the entropy loss corresponds with a minimization of the loss in kinetic energy. This implies that the trajectory which maximizes the L action is the one that, projected onto CP space, represents the SMEL behavior*. For clarity, consider the possibility of two competitive interactions entailing closure of different size loops. Obviously, according to the SMEL principle, the interaction involving the formation of the smallest loop will be preferred. On the other hand, the closure of the smallest loop implies the slowing down of a smaller number of oscillators in the LTM, either from 10^{11} to 10^7 Hz or from 10^7 to 10^3 Hz, depending on whether a secondary or tertiary interaction has taken place. Thus, at the LTM level, the formation of the smallest loop is also selected by L, as it minimizes the kinetic energy loss with respect to the formation of the competing larger loop.

On the other hand, the renormalization operation warrants the uniqueness of K for every LTM, which translates into the same CP, since the renormalization operation establishes the new frequencies exclusively on the basis of the latest CP formed. By virtue of the renormalization operation, the occurrence of intrachain contacts is reflected as an alteration of oscillator frequencies. *This is the reason why the Lagrangian L underlies the coarse microscopic behavior which shows up at the level of CP transitions as the SMEL principle*.

In regards to the T dependence of K, we must take into account that at the denaturation T $(T(\mathrm{denat.}))$ consensus bubbles are created which dismantle every CP, as indicated in Sec. III. This is the result of a broad spreading of frequencies beyond the critical spreading value and the

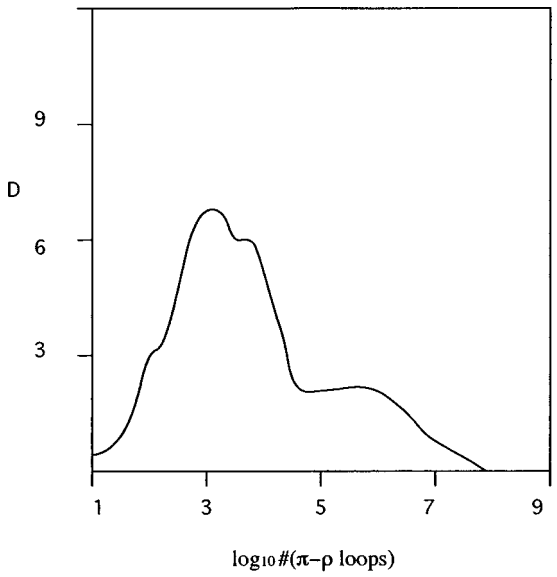

FIG. 10. Plot of the Hamming distance ($\times 100$) between consecutively evaluated LTMs along the folding pathway generated by iterating $(\pi-\rho)$ loops for BPTI.

concurrent impossibility of the preserving phase in consensus windows, as indicated in Sec. III. Thus, the renormalization operation near criticality will speed up all the oscillators for consensus windows where critical size bubbles occur, while such bubbles will become more and more ubiquitous as T approaches $T(\text{denat.})$. Then, K increases with T and the following relation holds as T approaches $T(\text{denat.})$ from below:

$$\lim_{T\to T(\text{denat.})^-} K=\text{Maximum } K=N\times 10^{22}\ \text{Hz}^2. \tag{9}$$

To conclude, we may state that the most economic folding pathway resolved at the coarse microscopic level, y^*, satisfies the following classical dynamics result:

$$L\ \text{action}(y^*)=\max_{\text{all LTM trajectories } y} L\ \text{action}(y). \tag{10}$$

Here ''L action(y)'' denotes the sum of all $L(\text{LTM}(i))$'s for all the LTM(i)'s, with i = discretized time index, along the trajectory y defined by mapping the previously allotted time interval over $(\mathbf{Z}_2)^{2N}$.

Without loss of generality we may assume that there exists a positive constant q so that $q+L=L'>0$. We may define a stochastic process Ξ at the semiempirical microscopic level of LTM transitions by defining transitional probabilities with respect to L':

$$p(V\to W)=[L'(W)-L'(V)]/[\Sigma_{W'}L'(W')-L'(V)], \tag{11}$$

where V, W, W' are LTMs, and the sum is extended over all W' states accessible from V within a predetermined Hamming distance $d=d(\text{max.})$ between LTM matrices (see Fig. 10). This distance is defined as

$$d(V,W)=(1/3N)\sum_{i,j}\Delta_{ij}(V,W),\quad \Delta_{ij}(V,W)=1\ \text{ if }\ V_{ij}\neq W_{ij},\quad \text{and}\quad =0\ \text{ otherwise.} \tag{12}$$

Prior to carrying out the stochastic computation using the Lagrangian, we determine the time dependence of $d(\text{max.})$ using the basic operational tenets defined in Sec. III. A $D-t$ plot ($D=100d(\text{max.})$) has been generated by performing 10^7 $(\pi-\rho$ loops) of the type indicated in Fig. 3, where π denotes translation and ρ renormalization. Thus, D is determined by computing after

every τ(read) interval the distance between two consecutive LTMs evaluated through two consecutive π operations. A computation for the BPTI is displayed in Fig. 10.

VI. THE LAGRANGIAN STRUCTURE OF THE $(\pi-\rho)$ LOOP COMPUTATION

This section is aimed at rigorously establishing the validity of the Lagrangian defined in Sec. V as the underlying functional selecting the optimal pathway among those generated by the $(\pi-\rho)$ loop iterative method. *We shall accomplish this by first noting that the Lagrangian-induced stochastic process* Ξ *and the stochastic process determined by an ensemble of realizations of the* $(\pi-\rho)$*-loop iterative computation are actually one and the same.*

The comparison between both processes requires a computation of the coarse informational entropy, $\sigma(t)$, and the computation of $\sigma_L(t)$, the informational entropy determined by the stochastic process Ξ. These quantities defined with respect to the CP partition of X, measure the spreading of probability in Shannon's sense due to the exploration of X that takes place during the folding process.[23,24] Thus, $\sigma=\sigma(t)$ is defined as follows:

$$\sigma(t)=-\sum_{j=1,\ldots,M} P_j(t)\ln P_j(t), \tag{13}$$

where M is the number of CP classes, and $P_j(t)$ is the probability that a $(\pi-\rho)$ loop computation reaches CP(j) at time t. Thus, this probability is evaluated as

$$P_j(t)=Q_j(t)/Q, \tag{14}$$

where $Q_j(t)$ is the number of $(\pi-\rho)$-loop pathways passing through CP(j) at time t and $Q\approx 100$ is the cardinal of the ensemble of $(\pi-\rho)$-loop pathways generated.

On the other hand, $\sigma_L(t)$ is defined as follows:

$$\sigma_L(t)=-\sum_{i=1,\ldots,M} \pi_t\mu\pi^{-1}\mathrm{CP}(i)[\ln(\pi_t\mu\pi^{-1}\mathrm{CP}(i))], \tag{15}$$

where π is the translation or projection of a LTM into its CP class, μ is the probability measure determined by Ξ over the space of LTM pathways, and $\pi_t\mu$ is the projection of μ over the time t replica of the LTM space. Thus, for a given CP, indicated by CP(i), the quantity $\pi_t\mu\pi^{-1}$ CP(i) is computed as follows:

$$\pi_t\mu\pi^{-1}\mathrm{CP}(i)=\#\{\Xi \text{ generated LTM trajectories intersecting } \pi^{-1}\ \mathrm{CP}(i) \text{ at time } t\}/\#\{\Xi \text{ generated LTM trajectories}\}. \tag{16}$$

Both coarse Shannon entropies defined, respectively, by Eqs. (13) and (15) are compared for the protein species BPTI (Fig. 11), giving a highly satisfactory agreement. Within the time scale range corresponding to the lifetime of the so-called molten globule intermediate[2] that corresponds to a plateau in the plots, the agreement is actually perfect. *We may conclude that the* $(\pi-\rho)$ *dynamics is underlied by the variational principle induced by the Lagrangian* L*, a result already implied in principle in Sec. V, and ultimately confirmed by the fact that both stochastic processes yield the same dispersion of probability at all times within relevant folding time scales.*

VII. GENERATING THE FOLDING DYNAMICS AT THE MICROSCOPIC LEVEL

The aim of this section is to determine the long-time microscopic dynamics of the chain torsional motion relevant to the folding process circumventing the computational limitations of Molecular Dynamics. This is done at the expense of approximating trajectories by coarse graining the time axis. Thus, in order to obtain the variational equations of microscopic motion, our strategy is to lift the coarsely defined Lagrangian introduced in Sec. V to the torus X. *Ultimately, we want to prove that the same coarse Lagrangian description that accounted for the SMEL, or maximum-economy-of-means behavior*[21] *at the level of contact pattern transitions, also yields Newton's equations of motion when carried over to the microscopic level.*

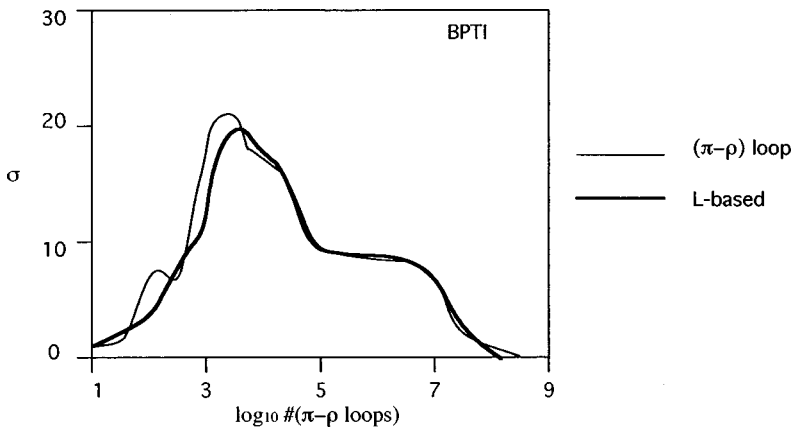

FIG. 11. Time dependence of the coarse information entropies $\sigma(t)$ and $\sigma_L(t)$ relative to the CP partition for BPTI. The abscissas indicate time in $\#(\pi-\rho)$ loop units.

To start with, the $(\pi-\rho)$ loop may be alternatively generated by a simulation of the continuous Φ, Ψ dynamics of the chain. Thus, if the dihedral torsions Φ_k, Ψ_k of the kth residue are modeled by oscillators with range $[0,2\pi]$, the LTM could simply be generated by transforming via a simple projection γ (see Fig. 12) the canonical torsional variables into the topological spin or two-state η, ξ variables used to define the LTM. This extended scheme of the $(\pi-\rho)$ loop is displayed in Fig. 12. One should notice that the computations using the extended scheme are actually identical to those using the scheme of Fig. 3, if the frequencies of the Φ, Ψ oscillators are taken to be exactly those of the η, ξ oscillators, and they are renormalized through the same rules via the map ρ.

Using the scheme of Fig. 12, the coarse Lagrangian defined in Sec. V may be now expressed in terms of the Φ, Ψ phase variables evaluated at time i on the microscopic conformation $\{\Phi_{k,i}, \Psi_{k,i}\}_k$ ($\Phi_{k,i}, \Psi_{k,i}$ denote the phases Φ_k, Ψ_k at time i):

$$L(\{\Phi_{k,i}, \Psi_{k,i}\}_k) = 1/2 \sum_{k=1,\ldots,2N} f_{k,i}^2 - H(\pi\gamma\{\Phi_{k,i}, \Psi_{k,i}\}_k), \tag{17}$$

where $f_{k,i}$ now represents the frequency of the Φ_k, Ψ_k phase oscillators at time i, which is identical to the frequency of the two-state η_k, ξ_k oscillators at time i, as indicated above. Thus, the following equations hold modulo 2π:

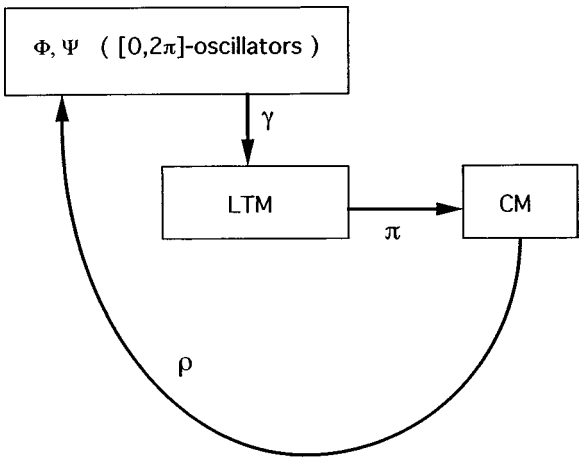

FIG. 12. Alternative scheme for the generation of $(\pi-\rho)$ loops using continuous oscillators with range $[0,2\pi]$ to mimic the Φ, Ψ torsions of the chain.

$$\Phi_{k,i+1}=\Phi_{k,i}+2\pi f_{k,i}h, \tag{18a}$$

$$\Psi_{k,i+1}=\Psi_{k,i}+2\pi f_{k,i}h, \tag{18b}$$

where $h=\tau'\approx 4$ ps at 298 °K ($\tau'=$minimum period, as defined in Sec. III) is a fixed time interval that enables us to discretize time so that we shall only consider the dynamics defined at multiples $t=ih$. The frequencies involved in Eqs. (18a) and (18b) are selected from the distribution ensemble whose width reflects the thermal fluctuations produced by the solvent environment, as indicated in Sec. III. With the time discretization, the variational calculus using L becomes an extremal problem in $2N\times J$ variables, with Jh being the time interval. The L action A reads as

$$A(\{\Phi_{k,i},\Psi_{k,i}\}_{k,i})=\sum_{i=1,\ldots,J} L(\{\Phi_{k,i},\Psi_{k,i}\}_k). \tag{19}$$

Thus, the pathway $\{\Phi_{k,i},\Psi_{k,i}\}_{k,i}$ realizing the extreme of A satisfies the following conditions:

$$\partial A/\partial\Phi_{k,i}=0; \quad \partial A/\partial\Psi_{k,i}=0; \quad k=1,\ldots,N; \quad i=1,\ldots,J. \tag{20}$$

These equations yield the following relations:

$$\rho(f_{k,i})=f_{k,i+1}=f_{k,i}-2\pi h[\partial H/\partial\Phi_k]_{i+1}, \tag{21a}$$

$$\rho(f_{k,i})=f_{k,i+1}=f_{k,i}-2\pi h[\partial H/\partial\Psi_k]_{i+1}. \tag{21b}$$

Since the reasoning is identical for both dihedral torsions, we shall henceforth focus only on Φ_k. In Eq. 21, $[\partial H/\partial\Phi_k]_{i+1}$ indicates that the derivative is evaluated at the conformation $\{\Phi_{k,i+1},\Psi_{k,i+1}\}_k$.

Equation (21) provides a thermodynamic interpretation of the $(\pi-\rho)$ loop dynamics whose characteristic period given by the time elapsed between one loop, and the next is $2^4h=64$ ps (see Sec. III). Thus, examining Eq. (21) once every 64 ps, we can identify the thermodynamic force—or torque—$[\partial H/\partial\Phi_k]_{i+1}$ as being the determinant of the frequency change. Thus, if a change of the phase $\Phi_{k,i}\rightarrow\Phi_{k,i+1}$ is responsible for an intrachain contact formation registered at the CP level, and thereby of a decrease in H, then the force $[\partial H/\partial\Phi_k]_{i+1}$ can only be positive or eventually zero, since the contact formed can only be dismantled $(\Delta H>0)$, or remain the same $(\Delta H=0)$ upon alterations of the phase Φ_k. Thus, the decrease in frequency that materializes upon the action of the renormalization at time $i+1$ is determined by the thermodynamic force $[\partial H/\partial\Phi_k]_{i+1}$ acting at time $i+1$. The stronger the contact formed as a result of the phase change $\Phi_{k,i}\rightarrow\Phi_{k,i+1}$, the bigger the decrease in frequency for the Φ_k oscillator.

One of the operational tenets of the $(\pi-\rho)$ loop is that if a transition $\Phi_{m,i}\rightarrow\Phi_{m,i+1}$ is responsible for the occurrence of an intrachain contact, with concurrent loop closure recorded as the CP$(i)\rightarrow$CP$(i+1)$ transition, then, *not only the frequency of the Φ_m oscillator is lowered upon renormalization, but also the f_k for any residue k within the loop*. This fact finds a natural interpretation once the thermodynamic forces governing the $(\pi-\rho)$ loop have been identified and Eq. (21) is used: *Only a Φ_k transition for a residue k within the loop can alter the contact formed*. Thus, all the f_k's in the loop are renormalized according to Eq. (21), while all other dihedrals retain their mean frequency because they cannot contribute to alter the contact in the CP(i) $\rightarrow$CP$(i+1)$ transition.

At this point we may obtain the microscopic equations of motion by combining Eqs. (18) and (21):

$$\Phi_{k,i+2}-2\Phi_{k,i+1}+\Phi_{k,i}=-4\pi^2h^2[\partial H/\partial\Phi_k]_{i+1}. \tag{22}$$

Since we can get the thermodynamic torques $[\partial H/\partial\Phi_k]_{i+1}$ by measuring the decrease in frequency using Eq. (18), together with the $(\pi-\rho)$ loop computation, *the folding dynamics have been essentially obtained at the microscopic level, except that we have coarse-grained time to multiples of 64h*.

Furthermore, Eq. (22) yields Newton's equations of motion at the microscopic level if we identify H with the potential and $(4\pi^2)^{-1}$ with the effective moment of inertia. Thus, in the limit $h \to 0$, Eq. (22) reads as

$$(4\pi^2)^{-1}\partial^2/\partial t^2 \Phi_k = -\partial H/\partial \Phi_k . \tag{23}$$

VIII. DISCUSSION

The semiempirical microscopic theory and its associated predictive algorithm expounded in this work provide the underpinnings to some issues that are central to the understanding of the long-time dynamics of a peptide chain in its relation to protein folding. Because our approach encompasses slower collective degrees of freedom than what detailed Molecular Dynamics computations may handle at present, we have initially dealt with a coarse description of the soft mode or manifold that nevertheless captures the dynamic hierarchy of torsional isomerization time scales. This has been done with the aim of eventually lifting the coarse dynamics to the level of soft-mode dynamics in a consistent manner.

In the first part of the work we obtain a coarse version of the soft-mode torsional dynamics of a peptide chain averaging out hard modes—stretching vibrations and planar angular vibrations,[25]—by means of a parallel computation which enabled us to penetrate into the time scale range of 1–10 ms, relevant to the protein folding process. Our treatment hinges upon a binary codification of local torsional isomerizations subject to the constraints imposed by both local and long-range correlations. In this way, patterns of locally encoded structural signals represent actual patterns of fulfilled local topological constraints that are recognized and translated as foldings of the chain. In turn, a feedback loop operates, whereby a new folding generated imposes new constraints upon the formation of patterns. Long-range intrachain contacts and secondary and tertiary interactions are induced by the drastic reduction in the available conformation space due to local correlations and by the solvent generation of local conformation-dependent dielectric environments. The actual time step of the parallel computation fixed at 64 ps is longer than that in previous treatments of the long-time torsional dynamics.[26] While such a long time step allows us to coarsely determine the microscopic origin of nucleation and cooperativity in the folding process by penetrating well into the 1–10 ms time range, it demands a more complex treatment of the solvent, which can no longer be treated as a hydrodynamic drag term: The recognition of a structural pattern becomes conditioned by the type of putative conformation-dependent dielectric domains that the solvent would determine if the corresponding folding would actually form.

Then, by introducing a variational principle underlying the generation of trajectories within our coarse microscopic description, we have reproduced and explained the expediency of the folding process at a meaningful level of understanding: The selected pathway reflects the maximum economy of each folding step. This nongeneric least-action formulation has been implemented, taking into account the inherently parallel nature of folding, demanding a concurrent evaluation of folding possibilities as consensus windows of locally encoded structural signals. The predictive potential of such an approach has been corroborated by contrasting the pathway singled out by the variational principle with the experimentally probed folding pathway.

The second aspect covered in this work has been to lift the coarse variational principle to the actual microscopic level of description of the long-time dynamics of chain torsions. The result reveal the consistency of our semiempirical variational approach, which becomes apparent by examining the $(\pi-\rho)$ loop dynamics at both ends of the spectrum of coarsenings of conformation space. Thus, at the finest level, we obtain Newton's equations of motion, while at the coarser level of contact patterns, we obtain a folding pathway entailing the maximum economy of means. The paramount advantage of our approach is given by the fact that we are able to actually compute ''relatively often'' (in a precise sense given in Sec. VII) the forces acting at the microscopic level and thus, elucidate the microscopic dynamics that project directly onto the dominant folding pathway.

ACKNOWLEDGMENTS

This work was produced during the tenure of a fellowship from the J. S. Guggenheim Memorial Foundation awarded to A. F. Part of this research was conceived during a visit to the

Courant Institute of Mathematical Sciences of New York University. Insightful conversations with Professor R. Stephen Berry, Professor Stuart A. Rice, Professor Jerome K. Percus, and Professor Karl Freed are gratefully acknowledged.

[1] R. Jaenicke, ''Is there a code for protein folding?,'' in *Protein Structure and Protein Engineering*, edited by E. L. Winnacker and R. Huber (Springer-Verlag, Berlin, 1988), pp. 16–36.
[2] O. B. Ptitsyn and G. V. Semisotnov, ''The mechanism of protein folding,'' in *Conformations and Forces in Protein Folding*, edited by B. Nall and K. Dill (American Association for the Advancement of Science, Washington, 1991), pp. 155–168.
[3] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. USA **89**, 20 (1992).
[4] R. L. Baldwin, Proc. Natl. Acad. Sci. USA **93**, 2627 (1996).
[5] K. A. Dill and H. S. Chan, Nature Struct. Biol. **4**, 10 (1997).
[6] J. Bohr, H. Bohr, and S. Brunak, Europhys. News **27**, 50 (1996).
[7] T. E. Creighton, N. J. Darby, and J. Kemmink, FASEB J. **10**, 110 (1996).
[8] M. Jamin and R. L. Baldwin, Nature Struct. Biol. **3**, 613 (1996).
[9] A. Fernández and G. Appignanesi, Phys. Rev. Lett. **78**, 2668 (1997).
[10] K. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. USA **90**, 1942 (1993).
[11] M. Guenza and K. F. Freed, J. Chem. Phys. **105**, 3823 (1996).
[12] C. Brooks III, M. Karplus, and B. Montgomery Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, Advances in Chemical Physics (Wiley, New York, 1988), Vol. LXXI.
[13] J. G. Saven and P. G. Wolynes, J. Mol. Biol. **257**, 199 (1996).
[14] C. Cantor and P. Schimmel, *Biophysical Chemistry* (Freeman, New York, 1980), Vols. I–III.
[15] A. Fernández, G. A. Appignanesi, and R. A. Montani, Phys. Rev. E **56**, 927 (1997).
[16] J. S. Richardson and D. C. Richardson, ''The origami of proteins,'' in *Protein Folding*, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 5–18.
[17] T. G. Oas and P. S. Kim, ''A protein folding intermediate analog,'' in Ref. 16, pp. 123–128.
[18] O. Sinanoglu, Chem. Phys. Lett. **81**, 188 (1981).
[19] A. Fernández, G. A. Appignanesi, and H. Cendra, Chem. Phys. Lett. **242**, 460 (1995).
[20] D. Bashford, M. Karplus, and D. Weaver, ''The diffusion–collision model of protein folding,'' in Ref. 16, pp. 283–290.
[21] A. Fernández, H. Arias, and D. Guerín, Phys. Rev. E **52**, R1299 (1995); A. Fernández, H. Arias, and D. Guerín, *ibid.* **54**, 1005 (1996).
[22] T. E. Creighton, *Understanding Protein Folding Pathways and Mechanisms in ''Protein Folding,''* edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 157–170.
[23] A. Fernández, Physica A **233**, 226 (1996).
[24] A. Fernández and A. Colubri, Physica A **248**, 336 (1998).
[25] N. Go and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[26] S. He and H. A. Scheraga, J. Chem. Phys. **108**, 271 (1998).

Free energy and correlations of the number-theoretical spin chain

Francesco Guerra[a)]
Dipartimento di Fisica, Università di Roma ''La Sapienza,''
Piazzale Aldo Moro, 2, I-00185 Roma, Italy

Andreas Knauf[b)]
Max-Planck-Institute for Mathematics in the Sciences,
Inselstr. 22–26, D-04103 Leipzig, Germany

(Received 26 September 1996; accepted for publication 2 February 1998)

We analyze the number-theoretical spin chain with partition function $Z(\beta)=\zeta(\beta-1)/\zeta(\beta)$ using the polymer model technique. The finite (grand) canonical chains give bounds for the limit free energy and internal energy. The correlation functions for inverse temperature $\beta=-1$ are products of two-point functions. A combinatorial result for general interval graphs is derived. © *1998 American Institute of Physics.* [S0022-2488(98)00606-9]

I. INTRODUCTION

The simplest kind of number theory consists of adding integers and observing the regularities that arise. For example, one may add integers at neighboring positions in a row and write the results in the next row. If one starts with the row number $k=0$ containing a single one, then the result is Pascal's triangle [Fig. 1(a)], with the binomial coefficient $\binom{k}{l}$ at position (k,l). The defining relations of the binomial coefficients are thus

$$\binom{0}{l}:=\delta_0(l) \quad \text{and} \quad \binom{k+1}{l}:=\binom{k}{l-1}+\binom{k}{l}.$$

Although Pascal's triangle was already described in the 1303 book *The Precious Mirror of the Four Elements* by Chu Shih-chieh and thus has a rather long history, there are still many unsolved questions connected with it. Anybody who is now and then summing series knows some nontrivial identities met by the binomial coefficients.

Now we change the defining rule a bit. Instead of simply forgetting the values written in the kth row, we copy them to the row number $k+1$, thus obtaining *Pascal's triangle with memory.*

We want 2^k instead of $2^{k+1}-1$ entries in each row. Thus we consider these entries as being circularly ordered. For practical purposes we write a second copy of the first entry 1 at the end of the row but encircle it in order to mark its identification with the first entry [Fig. 1(b)].

The 2^k entries can be indexed by the base 2 k-digit numbers $0\ldots0$, $0\ldots01$, $\ldots$, $1\ldots1$, writing them in ascending order. Thus we may consider these entries as the values of a function $\mathbf{h}_k^C:\mathbf{G}_k\to\mathbb{N}$ with the k-fold Cartesian product group $\mathbf{G}_k:=\mathbb{Z}_2^k$, $\mathbb{Z}_2=(\{0,1\},+)$ being the two-element group in additive notation. A moment's thought shows that these functions can be defined by $\mathbf{h}_0^C:=1$,

$$\mathbf{h}_{k+1}^C(\sigma_1,\ldots,\sigma_{k+1}):=\mathbf{h}_k^C(\sigma_1,\ldots,\sigma_k)+\sigma_{k+1}\cdot\mathbf{h}_k^C(1-\sigma_1,\ldots,1-\sigma_k). \tag{1}$$

If one defines the *energy function* $\mathbf{H}_k^C:\mathbf{G}_k\to\mathbb{R}$ of the *canonical ensemble* by $\mathbf{H}_k^C=\ln(\mathbf{h}_k^C)$, then the functions

a)Electronic mail: guerra@roma1.infn.it
b)Electronic mail: knauf@mis.mpg.de

 © 1998 American Institute of Physics

a)

								$l=0$											
$k=0$								1			$l=1$								
$k=1$							1			1			$l=2$						
$k=2$					1			2			1			$l=3$					
$k=3$			1			3			3			1			$l=4$				
$k=4$	1			4			6			4			1						

b)

$k=0$	1															①
$k=1$	1						2									①
$k=2$	1			3			2			3						①
$k=3$	1		4		3		5		2		5		3		4	①
$k=4$	1	5	4	7	3	8	5	7	2	7	5	8	3	7	4	5	①

FIG. 1. Pascal's triangle (a) without memory and (b) with memory.

$$Z_k^C(\beta) := \sum_{\sigma \in \mathbf{G}_k} \exp(-\beta \mathbf{H}_k^C(\sigma))$$

can be interpreted as partition functions of a spin chain of length k, the superscript C referring to the canonical ensemble. Here the spins take the values $\sigma_i = 0,1$ instead of the more conventional choice $(-1)^{\sigma_i} = 1,-1$.

It turns out that they converge for $\mathrm{Re}(\beta) > 2$ to

$$Z(\beta) = \lim_{k \to \infty} Z_k^C(\beta)$$

which is the quotient

$$Z(\beta) := \frac{\zeta(\beta - 1)}{\zeta(\beta)} = \sum_{n=1}^{\infty} \varphi(n) \cdot n^{-\beta}$$

of Riemann zeta functions. Here $\varphi(n) := |\{l \in \{1, \ldots, n\} | \gcd(l,n) = 1\}|$ is Euler's totient function.

In Ref. 1 one of us showed that the interaction coefficients $j_k^C : \mathbf{G}_k \to \mathbb{R}$ with

$$\mathbf{H}_k^C(\sigma) = -\sum_{t \in \mathbf{G}_k^*} j_k^C(t) \cdot (-1)^{\sigma \cdot t}$$

are asymptotically translation invariant and saturate in the length k of the chain. Moreover, the effective interaction between spins at positions r and l was shown to decay like $1/(r-l)^2$.

This decay rate $\mathrm{dist}^{-\alpha}$ with $\alpha = 2$ is known to be the borderline case for phase transitions of a spin chain.

The dual group $\mathbf{G}_k^*$ of $\mathbf{G}_k$ being isomorphic to $\mathbf{G}_k$, the Fourier transform of $f : \mathbf{G}_k \to \mathbb{R}$ is given by

$$(\mathcal{F}_k f)(t) := 2^{-k} \sum_{\sigma \in \mathbf{G}_k} f(\sigma) \cdot (-1)^{\sigma \cdot t} \tag{2}$$

and thus $j_k^C = -\mathcal{F}_k \mathbf{H}_k^C$. In Ref. 1 the interaction was shown to be weakly ferromagnetic in the sense that $j_k^C(t) \geqslant 0$ for $t \neq 0$.

It turned out to be useful to introduce the *grand canonical* energy functions $\mathbf{H}_k^G : \mathbf{G}_k \to \mathbb{R}$, $\mathbf{H}_k^G := \ln(\mathbf{h}_k^G)$, the values

$$\mathbf{h}_k^G(\sigma) := \mathbf{h}_{k+1}^C(\sigma,1) \tag{3}$$

being the new elements appearing in each row of Fig. 1(b). They have the symmetries

$$\mathbf{H}_k^G(1-\sigma_1, \ldots, 1-\sigma_k) = \mathbf{H}_k^G(\sigma_1, \ldots, \sigma_k) = \mathbf{H}_k^G(\sigma_k, \ldots, \sigma_1).$$

In Ref. 2 the existence of the thermodynamic limit $F(\beta) := \lim_{k\to\infty} F_k^C(\beta)$ of the free energy $F_k^C(\beta) := -(\beta\cdot k)^{-1}\ln(Z_k^C(\beta))$ was shown (due to the lack of strict translation invariance of the interaction standard estimates cannot be applied here). Due to the long-range character of the interaction the chain is in a frozen state above the phase transition which gives rise to the somewhat unusual constancy of the free energy [$F(\beta)=0$ for $\beta\geqslant 2$].

In Ref. 3 thermodynamic expectations

$$\langle O\rangle_k^C(\beta) := \frac{\Sigma_{\sigma\in\mathbf{G}_k} O(\sigma)\exp(-\beta\mathbf{H}_k^C(\sigma))}{Z_k^C(\beta)}$$

of quantities like the internal energy $U_k^C := \langle (1/k)\mathbf{H}_k^C\rangle_k^C$ and magnetization $M_k := \langle (1/k)\times\Sigma_{l=1}^k(-1)^{\sigma_l}\rangle_k^C$ were analyzed. Due to the absolute convergence of the Dirichlet series $Z(\beta) = \Sigma_{n=1}^\infty \varphi(n) n^{-\beta}$ for $\mathrm{Re}(\beta)>2$ the limits $U(\beta) := \lim_{k\to\infty} U_k^C(\beta)$ and $M(\beta) := \lim_{k\to\infty} M_k(\beta)$ can be easily shown to have the values $U(\beta)=0$, $M(\beta)=1$ in the frozen low-temperature phase $\beta>2$.

In the high-temperature regime $\beta<2$ the estimates

$$\frac{\ln 2-\beta\,\ln(3/2)}{2-\beta} \leqslant U(\beta) \leqslant \ln(3/2) \tag{4}$$

were derived.

The corresponding grand canonical quantities are defined similarly.

In Ref. 4 it was shown that the values $\mathbf{H}_k^C(\sigma)$ of the canonical energy function can be naturally interpreted in terms of geodesics coming from and going to the cusp in the modular domain.

These geodesics are naturally organized in families of 2^k members and labelled by $\sigma\in\mathbf{G}_k$. Then $\mathbf{H}_k^C(\sigma)$ is the length difference between the geodesic with index σ and the shortest such geodesic (which carries index 0).

This related the spin chain to the quantum scattering theory in the modular domain, since these length differences arise naturally in the Wentzel–Kramers–Brilloin (WKB) expansion of the scattering matrix.

Moreover, numerical evidence was presented that the Riemann zeta function can be well approximated *within its critical strip* by statistical-mechanical expectations. The reasons for this observation are discussed in Ref. 5 by one of us.

In Ref. 6 Contucci and Knauf showed that there exists only one phase transition, at $\beta=2$. There the magnetization jumps from 1 to 0. The phase transition is at most of second order.

In Ref. 7 the relations between number theory and polymer models are treated.

The same partition function (but without the interpretation as coming from a spin chain) was discussed by Artuso *et al.* in Ref. 8. See also the related work by Feigenbaum[9] and the overview by Cvitanović.[10]

The study of the number-theoretical spin chain is mainly motivated by the idea that probabilistic properties of statistical mechanics, and in particular ferromagnetism, could be helpful in deriving new results in number theory (see Ref. 5).

For example, an interpretation of the functional equation of Z in terms of a Wannier-type duality of polymer model interpretations would be welcome.

In Sec. II we calculate the Fourier transforms of the functions $\mathbf{h}_k^C$ and $\mathbf{h}_k^G$. This leads directly to expressions for the correlation functions at $\beta=-1$.

The polymer model technique is then used in Sec. III in order to derive a formula for the interaction coefficients. They and not the free energy are described using connected multipolymers.

Like any Dirichlet series with a Euler product the number-theoretical spin chain itself has an interpretation as a polymer model (see Ref. 7), but this interpretation is not well adapted to the peculiarities of this chain. So, we would like to see another such interpretation in the spirit of Sec. III.

In Sec. IV we deal with general interval graphs, that is, graphs whose vertices correspond to intervals and which are connected by an edge if the corresponding intervals overlap. These types of graphs are used in our context.

In Sec. V the limit free energy is estimated from above and below using the k-spin (grand) canonical ensembles.

This result and a polymer inequality finally lead in Sec. VI to inequalities for the internal energy.

II. CORRELATION FUNCTIONS AT $\beta=-1$

In this section we analyze the Fourier transforms [see (2)] $J_k^C := \mathcal{F}_k(\mathbf{h}_k^C)$, $J_k^G := \mathcal{F}_k(\mathbf{h}_k^G)$ of the exponentiated (grand) canonical energy functions $\mathbf{h}_k^C$, resp. $\mathbf{h}_k^G$, inductively defined in (1) and (3). Using the shorthand $|s| := \Sigma_{i=1}^l s_i$ for $s \in \mathbf{G}_l$, we have the following.

Lemma 1: $J_0^C = 1$, and for $t = (t_1, \ldots, t_{k+1}) \in \mathbf{G}_{k+1}^*$

$$J_{k+1}^C(t_1, \ldots, t_{k+1}) = \tfrac{1}{2}(1 + (-1)^{t_{k+1}} + (-1)^{|t|}) J_k^C(t_1, \ldots, t_k). \tag{5}$$

The grand canonical Fourier coefficients are given by

$$J_{k+1}^G(t) = (1 + (-1)^{|t|}) J_k^C(t) \quad (t \in \mathbf{G}_k^*). \tag{6}$$

Proof: If $k=0$, then $|\mathbf{G}_k| = 1$ and $J_0^C = \mathbf{h}_0^C = 1$. To show the inductive formula (5), we use the abbreviation $t' := (t_1, \ldots, t_k) \in \mathbf{G}_k^*$. Then

$$\begin{aligned}
J_{k+1}^C(t) &= 2^{-k-1} \sum_{\sigma \in \mathbf{G}_{k+1}} \mathbf{h}_{k+1}^C(\sigma) \cdot (-1)^{\sigma \cdot t} \\
&= 2^{-k-1} \sum_{\sigma' \in \mathbf{G}_k} (\mathbf{h}_{k+1}^C(\sigma', 0) + (-1)^{t_{k+1}} \mathbf{h}_{k+1}^C(\sigma', 1)) \cdot (-1)^{\sigma' \cdot t'} \\
&= \frac{1}{2} 2^{-k} \sum_{\sigma' \in \mathbf{G}_k} \mathbf{h}_k^C(\sigma') \cdot (-1)^{\sigma' \cdot t'} + \frac{1}{2} (-1)^{t_{k+1}} 2^{-k} \sum_{\sigma' \in \mathbf{G}_k} (\mathbf{h}_k^C(\sigma') \\
&\quad + \mathbf{h}_k^C(1 - \sigma')) \cdot (-1)^{\sigma' \cdot t'} \\
&= \tfrac{1}{2}(1 + (-1)^{t_{k+1}} + (-1)^{|t|}) J_k^C(t'),
\end{aligned}$$

using the iterative definition (1) of $\mathbf{h}_k^C$ and the relation

$$\sum_{\sigma' \in \mathbf{G}_k} \mathbf{h}_k^C(1 - \sigma') \cdot (-1)^{\sigma' \cdot t'} = (-1)^{|t'|} \sum_{\sigma' \in \mathbf{G}_k} \mathbf{h}_k^C(\sigma') \cdot (-1)^{\sigma' \cdot t'}.$$

Equation (6) follows from the definition $\mathbf{h}_k^G(\sigma) = \mathbf{h}_{k+1}^C(\sigma, 1)$ of $\mathbf{h}_k^G$ and (5), since this implies

$$\begin{aligned}
J_k^G(t) &= 2^{-k} \sum_{\sigma \in \mathbf{G}_k} \mathbf{h}_{k+1}^C(\sigma, 1) \cdot (-1)^{\sigma \cdot t} \\
&= 2^{-k} \sum_{\sigma \in \mathbf{G}_k} (\mathbf{h}_k^C(\sigma) + \mathbf{h}_k^C(1 - \sigma)) \cdot (-1)^{\sigma \cdot t} \\
&= (1 + (-1)^{|t|}) J_k^C(t).
\end{aligned}$$

□

The above formulas allow us to write the Fourier coefficients as products over contributions from so-called polymers. The set

$$P_k := \{p_{l,r}, p_l \in \mathbf{G}_k^* \,|\, 1 \leqslant l, r \leqslant k;\ l < r\}$$

of *polymers* in $\mathbf{G}_k^*$ is given by

$$p_{l,r}(i) := \delta_l(i) + \delta_r(i), \quad p_l(i) := \delta_l(i) \quad (i = 1, \ldots, k). \tag{7}$$

We define their *support* by $\mathrm{supp}(p_{l,r}) := \{i \,|\, l \leqslant i \leqslant r\}$ for the *even* polymers and $\mathrm{supp}(p_l) := \{i \,|\, l \leqslant i \leqslant k\}$ for the *odd* ones. Two polymers $\gamma_1, \gamma_2 \in P_k$ are called overlapping or *incompatible* if $\mathrm{supp}(\gamma_1) \cap \mathrm{supp}(\gamma_2) \neq \varnothing$. So, in particular, all the odd polymers p_l, $1 \leqslant l \leqslant k$, are mutually incompatible.

Now an arbitrary group element $t = (t_1, \ldots, t_k) \in \mathbf{G}_k^*$ can be uniquely decomposed into the sum of $n := [\frac{1}{2}(|t| + 1)]$ compatible polymers:

$$t = \begin{cases} \displaystyle\sum_{i=1}^{n} p_{l_i, r_i}, & |t| \text{ even}, \\ \displaystyle\sum_{i=1}^{n-1} p_{l_i, r_i} + p_{l_n}, & |t| \text{ odd}, \end{cases} \tag{8}$$

whose indices l_i, r_i may be assumed to be increasing in i and are determined by $t_{l_i} = t_{r_i} = 1$. As an example $(0,1,0,1,1,0) = p_{2,4} + p_5 \in \mathbf{G}_6^*$.

We write this decomposition in the short form $t = \Sigma_{i=1}^{n} \gamma_i$ and set the activity of a polymer $\gamma \in \mathbf{G}_k^*$ equal to $z(\gamma) := -3^{-|\mathrm{supp}(\gamma)|}$.

Theorem 2: *With the above decomposition (8) of* $t \in \mathbf{G}_k^*$

$$J_k^C(t) = (3/2)^k \prod_{i=1}^{n} z(\gamma_i) \tag{9}$$

and

$$J_k^G(t) = (1 + (-1)^{|t|})(3/2)^k \prod_{i=1}^{n} z(\gamma_i). \tag{10}$$

Proof: We proceed by induction in the number k of spins. For $k = 0$, Eq. (9) gives $J_0^C = 1$, as in Lemma II.

(i) If $|t|$ is even and $t_{k+1} = 0$, then the polymers of the representation (8) of $(t_1, \ldots, t_{k+1})$ coincide with the polymers of $(t_1, \ldots, t_k)$. So, by (9), $J_{k+1}^C(t_1, \ldots, t_{k+1}) = \frac{3}{2} J_k^C(t_1, \ldots, t_k)$, in accordance with (5).

(ii) If $|t|$ is even and $t_{k+1} = 1$, then $(t_1, \ldots, t_{k+1})$ contains the additional polymer p_{k+1} so that both (5) and (9) imply

$$J_{k+1}^C(t_1, \ldots, t_{k+1}) = -\tfrac{1}{2} J_k^C(t_1, \ldots, t_k).$$

(iii) If $|t|$ is odd, then the odd polymer p_{l_n} in the decomposition (8) of $(t_1, \ldots, t_k)$ has activity $z(p_{l_n}) = -3^{l_n - k - 1}$. If $t_{k+1} = 0$, then the polymer decomposition for $(t_1, \ldots, t_{k+1})$ remains the same, but now $z(p_{l_n}) = -3^{l_n - k - 2}$. If $t_{k+1} = 1$, the odd polymer p_{l_n} is changed into the even polymer $p_{l_n, k+1}$ which, too, has activity $z(p_{l_n, k+1}) = -3^{l_n - k - 2}$. So both (5) and (9) predict $J_{k+1}^C(t_1, \ldots, t_{k+1}) = \frac{1}{2} J_k^C(t_1, \ldots, t_k)$, independent of whether $t_{k+1} = 0$ or 1.

The corresponding equation (10) for the grand canonical coefficients $J_k^G(t)$ follows directly from (9) and (6). □

Theorem 2 has the following corollary concerning the (grand) canonical expectations

$$\langle f\rangle_k^{C/G}(\beta)=\frac{\Sigma_{\sigma\in\mathbf{G}_k} f(\sigma)e^{-\beta\mathbf{H}_k^{C/G}(\sigma)}}{Z_k^{C/G}(\beta)}$$

of $f:\mathbf{G}_k\to\mathbb{R}$ at inverse temperature β:

Corollary 3: For $1\leqslant i_1<\cdots<i_{2l}\leqslant k$ *the even (grand) canonical correlation functions at inverse temperature* $\beta=-1$ *equal*

$$\left\langle \prod_{m=1}^{l}\sigma_{i_{2m-1}}\sigma_{i_{2m}}\right\rangle_k^{C/G}(-1)=\prod_{m=1}^{l}\langle\sigma_{i_{2m-1}}\sigma_{i_{2m}}\rangle_k^{C/G}(-1),$$

whereas the grand canonical correlation functions between an odd number of spins vanish for any $\beta\in\mathbb{R}$.

Proof: Using (9), resp. (10), we obtain for $|t|$ even the general identity

$$\left\langle \prod_{i:t_i=1}\sigma_i\right\rangle_k^{C/G}(-1)=\frac{J_k^{C/G}(t)}{J_k^{C/G}(0)}=\prod_{i=1}^{n}z(\gamma_i)$$

for the correlation functions, proving the product formula. The odd grand canonical correlation functions vanish identically since, by (10), $J_k^G(t)=0$ for $|t|$ odd. □

III. POLYMERS AND INTERACTION COEFFICIENTS

Now we will express the (grand) canonical interaction coefficients

$$\mathbf{j}_k^{C/G}(t)=-\mathcal{F}_k(\mathbf{H}_k^{C/G})(t)\quad(t\in\mathbf{G}_k^*)$$

in terms of the Fourier transforms $J_k^{C/G}(t)=\mathcal{F}_k(\mathbf{h}_k^{C/G})(t)$ of the exponentiated energy functions $\mathbf{h}_k^{C/G}=\exp(\mathbf{H}_k^{C/G})$ [observe that for historical reasons the $\mathbf{j}_k^{C/G}(t)$ are the *negative* Fourier coefficients].

Our motivation for considering that functional relationship is the existence of the explicit formulas (9) and (10) for J_k^C, resp. J_k^G.

The technique we apply stems from the theory of polymer models whose general setting we now recall (see, e.g., Refs. 11–13).

In an abstract setting one starts with a denumerable set P whose elements are called *polymers*. Two given polymers $\gamma_1,\gamma_2\in P$ may or may not overlap (be *incompatible*). Incompatibility is assumed to be a reflexive and symmetric relation on P.

Thus one may associate to a *k-polymer* $X:=(\gamma_1,\ldots,\gamma_k)\in P^k$ an undirected graph $G(X)=(V(X),E(X))$ without loops (edges connecting a vertex with itself) with vertex set $V(X):=\{1,\ldots,k\}$, vertices $i\neq j$ being connected by the edge $\{\gamma_i,\gamma_j\}\in E(X)$ if γ_i and γ_j are incompatible. Accordingly the k-polymer X is called *connected* if $G(X)$ is path-connected and *disconnected* if it has no edges $[E(X)=\emptyset]$.

The corresponding subsets of P^k are called C^k, resp. D^k, with $D^0:=P^0:=\{\emptyset\}$ consisting of a single element. Moreover $P^\infty:=\cup_{k=0}^\infty P^k$ with the subsets $D^\infty:=\cup_{k=0}^\infty D^k$ and $C^\infty:=\cup_{k=1}^\infty C^k$. We write $|X|:=k$ if $X\in P^k$.

Statistical weights or activities $z:P\to\mathbb{C}$ of the polymers are multiplied to give the activities $z^X:=\Pi_{i=1}^k z(\gamma_i)$ of k-polymers.

Finally, in order to work with finite sums, we use finite subsets $P_\Lambda\subset P$ of polymers, which are thought of as being geometrically enclosed in a finite region $\Lambda\subset S$.

A system of statistical mechanics is called a *polymer model* if its partition functions Z_Λ have the form

$$Z_\Lambda=\sum_{X\in D_\Lambda^\infty}\frac{z^X}{|X|!}. \tag{11}$$

Then, up the normalization factor $|\Lambda|^{-1}$, the free energy is given by

$$\ln(Z_\Lambda) = \sum_{X \in C_\Lambda^\infty} \frac{n(X)}{|X|!} z^X, \tag{12}$$

with $n(X) := n_+(X) - n_-(X)$, $n_\pm(X)$ being the number of subgraphs of $G(X)$ connecting the vertices of $G(X)$ with an even, resp. odd, number of edges (see Ref. 11).

In the present context we argue as follows: Our polymers are elements of the Abelian group

$$\mathbf{G}_\infty^* := \oplus_{n=1}^\infty \mathbb{Z}_2 \subset \otimes_{n=1}^\infty \mathbb{Z}_2,$$

which is the direct sum of the $\mathbb{Z}_2$ groups and thus consists of those infinite sequences $t = (t_1, t_2, \dots) \in \otimes_{n=1}^\infty \mathbb{Z}_2$ which have only finitely many entries $t_i \neq 0$. The set P of polymers and their supports is defined similarly to the set P_k [see (7)] except that now $k = \infty$.

This infinite volume case is then approximated by considering the set P_k of polymers in the finite subset $\Lambda = \{1, \dots, k\} \subset \mathbb{N}$. The group $\mathbf{G}_k^*$ is identified with the subgroup of those $t \in \mathbf{G}_\infty^*$ with $t_n = 0$ for $n > k$.

Then the map $D_k^\infty \to \mathbf{G}_k^*$, $(\gamma_1, \dots, \gamma_n) \mapsto \Sigma_{i=1}^n \gamma_i$ is an isomorphism between the disconnected multi-polymers in $\{1, \dots, k\}$ and the Abelian group $\mathbf{G}_k^*$.

We will now use this isomorphism to calculate the interaction coefficients.

For the canonical ensemble formula (9) in Theorem 2 implies

$$\begin{aligned} j_k^C(t) &= -2^{-k} \sum_{\sigma \in \mathbf{G}_k} \mathbf{H}_k^C(\sigma) \cdot (-1)^{\sigma \cdot t} \\ &= -2^{-k} \sum_{\sigma \in \mathbf{G}_k} \ln\left[\sum_{s \in \mathbf{G}_k^*} J_k^C(s) \cdot (-1)^{s \cdot \sigma} \right] \cdot (-1)^{\sigma \cdot t} \\ &= -\delta_{t,0} \cdot k \, \ln(3/2) - 2^{-k} \sum_{\sigma \in \mathbf{G}_k} \ln\left[\sum_{X \in D_k^\infty} \frac{\tilde{z}_\sigma^X}{|X|!} \right] \cdot (-1)^{\sigma \cdot t}, \end{aligned} \tag{13}$$

where the redefined single-polymer activities $\tilde{z}_\sigma(\gamma)$, $\gamma \in P_k$, are given by $\tilde{z}_\sigma(\gamma) := z_\sigma(\gamma) \cdot (-1)^{\sigma \cdot \gamma}$, that is

$$\tilde{z}_\sigma(p_{l,r}) := z(p_{l,r}) \cdot (-1)^{\sigma_l + \sigma_r}, \quad \tilde{z}_\sigma(p_l) := z(p_l) \cdot (-1)^{\sigma_l}.$$

By (12) we get

$$\begin{aligned} j_k^C(t) + \delta_{t,0} \cdot k \, \ln(3/2) &= -2^{-k} \sum_{\sigma \in \mathbf{G}_k} \sum_{X \in C_k^\infty} \frac{n(X)}{|X|!} \tilde{z}_\sigma^X \cdot (-1)^{\sigma \cdot t} \\ &= -\sum_{X = (\gamma_1, \dots, \gamma_{|X|}) \in C_k^\infty} \frac{n(X)}{|X|!} z^X \cdot 2^{-k} \sum_{\sigma \in \mathbf{G}_k} (-1)^{\sigma \cdot \left(t + \sum_{i=1}^{|X|} \gamma_i \right)} \\ &= -\sum_{\substack{X = (\gamma_1, \dots, \gamma_{|X|}) \in C_k^\infty \\ \sum_{i=1}^{|X|} \gamma_i = t}} \frac{n(X)}{|X|!} z^X, \end{aligned} \tag{14}$$

using the identity $\Sigma_{\sigma \in \mathbf{G}_k} (-1)^{\sigma \cdot s} = 2^k \delta_{s,0}$.

For the grand canonical ensemble $j_k^G(t)=0$ if $|t|$ is odd. If $|t|$ is even, then by (10) and a similar derivation as above

$$j_k^G(t)=-\delta_{t,0}\cdot(\ln(2)+k\ \ln(3/2))-\sum_{\substack{X=(\gamma_1,\ldots,\gamma_{|X|})\in C_k^{\infty,e}\\ \Sigma_{i=1}^{|X|}\gamma_i=t}}\frac{n(X)}{|X|!}z^X, \tag{15}$$

where now $C_k^{\infty,e}$ contains precisely those multi-polymers $X=(\gamma_1,\ldots,\gamma_{|X|})$ consisting only of even polymers γ_i of the form $p_{l,r}$ and not the odd polymers p_l.

IV. COMBINATORICS OF INTERVAL GRAPHS

The graphs $G(X)$ of the connected multi-polymers $X\in C_k^\infty$, resp. $C_k^{\infty,e}$, occurring in (14), resp. (15), have a special structure which enables us to obtain a simple explicit formula for the combinatorial factor $n(X)$.

The supports of the even polymers $p_{l,r}$ were of the form $\{l,\ldots,r\}$, whereas for the odd polymers p_l one had $\mathrm{supp}(p_l)=\{l,\ldots,k\}$. Polymers in P_k with overlapping supports were called incompatible.

These definitions imply that the graph $G(X)$ of a multi-polymer is a so-called interval graph. In general, a graph $G=(V,E)$ is called *interval graph* if there exists an isomorphism $I:V\to\overline{V}$ between the set V of vertices and a set $\overline{V}$ of intervals in $\mathbb{R}$ such that $v_i,v_j\in V$ are connected by the edge $\{v_i,v_j\}\in E$ precisely if the corresponding intervals $I(v_i)$ and $I(v_j)$ have nonempty intersection.

By renaming the vertices we can assume that $V=V_r:=\{v_1,\ldots,v_r\}$, $I(v_i)=[a_i,b_i]$, and $a_i\leqslant a_j$ for $i<j$. Then $G_i:=(V_i,E_i)$ with $E_i:=E\cap(V_i\times V_i)$ is the induced subgraph of G on the first i vertices, $1\leqslant i\leqslant r$.

Theorem 4: *For interval graphs G with the above notation*

$$n(G)=\prod_{i=1}^{r-1}(|E_i|-|E_{i+1}|).$$

Proof: We prove the formula by induction in the number $|V|$ of vertices.

If $|V|=1$, the index set of the product is empty so that the product gives 1. But $n(G)=n_+(G)-n_-(G)=1$, too, since $n_+(G)=1$ and $n_-(G)=0$.

The subgraphs G_i of G are themselves interval graphs. We assume that $n(G_{l-1})=\Pi_{i=1}^{l-2}(|E_i|-|E_{i+1}|)$ and show that

$$n(G_l)=\prod_{i=1}^{l-1}(|E_i|-|E_{i+1}|)=n(G_{l-1})\cdot(|E_{l-1}|-|E_l|)=-n(G_{l-1})\cdot|\hat{E}_l|, \tag{16}$$

where $\hat{E}_l:=E_l-E_{l-1}$ is the set of edges in G_l connecting the vertex v_l with the subgraph G_{l-1}. The set $\mathcal{C}_l$ of connected subgraphs of G_l is the disjoint union

$$\mathcal{C}_l=\dot{\cup}_{\hat{E}\subset\hat{E}_l}\mathcal{C}_l(\hat{E})$$

of $\mathcal{C}_l(\hat{E}):=\{(V_l,\tilde{E})\in\mathcal{C}_l|\tilde{E}\cap\hat{E}_l=\hat{E}\}$, i.e., the subsets of those connected graphs where the edges in $\hat{E}$ emanate from the vertex v_l. So

$$n(G_l)=\sum_{\hat{E}\subset\hat{E}_l}\ \sum_{(V_l,\tilde{E})\in\mathcal{C}_l(\hat{E})}(-1)^{|\tilde{E}|}. \tag{17}$$

We will show now that in fact

$$n(G_l)=\sum_{\substack{\hat{E}\subset\hat{E}_l\\ |\hat{E}|=1}}\ \sum_{(V_l,\tilde{E})\in\mathcal{C}_l(\hat{E})}(-1)^{|\tilde{E}|}, \tag{18}$$

that is, only subgraphs with exactly one edge adjacent to v_l contribute.

It is obvious that $\mathcal{C}_l(\varnothing)=\varnothing$, since in that case the vertex v_l is disconnected.

So we are left to show that terms in (17) with $|\hat{E}|\geqslant 2$ give no contribution.

We choose two different edges $\{v_x,v_l\},\{v_y,v_l\}\in\hat{E}$. Then the graph G_l contains the edge $e:=\{v_x,v_y\}$, as we show now. Namely the intervals $[a_x,b_x]=I(v_x)$ and $[a_y,b_y]=I(v_y)$ both intersect the interval $[a_l,b_l]=I(v_l)$ so that $a_l\leqslant b_x,b_y$, and they have the additional property $a_x,a_y\leqslant a_l$ since $x,y\leqslant l$. Thus $a_l\in[a_x,b_x]\cap[a_y,b_y]$ showing that these intervals intersect.

The set $A:=\{(V_l,\tilde{E})\in\mathcal{C}_l(\hat{E})\,|\,e\in\tilde{E}\}$ consists of the connected subgraphs containing the edge $e=\{v_x,v_y\}$. The map

$$J:A\to J(A),\quad (V_l,\tilde{E})\mapsto(V_l,\tilde{E}-\{e\}),$$

deletes the edge e and is obviously one-to-one. Moreover,

$$\mathcal{C}_l(\hat{E})=A\cup J(A),$$

since the graph $(V_l,\tilde{E}-\{e\})$ is still connected (v_x and v_y are connected by the edges $\{v_x,v_l\}$, $\{v_y,v_l\}$).

So if $|\hat{E}|\geqslant 2$, then

$$\sum_{(V_l,\tilde{E})\in\mathcal{C}_l(\hat{E})}(-1)^{|\tilde{E}|}=\sum_{(V_l,\tilde{E})\in A}((-1)^{|\tilde{E}|}+(-1)^{|\tilde{E}-\{e\}|})=0,$$

showing (18).

Now we resume (18):

$$n(G_l)=-\sum_{\hat{e}\in\hat{E}_l}\ \sum_{(V_l,\tilde{E})\in\mathcal{C}_l(\{\hat{e}\})}(-1)^{|\tilde{E}-\{\hat{e}\}|}.$$

However, the map K on $\mathcal{C}_l(\{\hat{e}\})$ given by $K((V_l,\tilde{E})):=(V_{l-1},\tilde{E}-\{\hat{e}\})$ is one-to-one and $K(\mathcal{C}_l(\{\hat{e}\}))=\mathcal{C}_{l-1}$, that is, its image consists of the connected subgraphs on the first $l-1$ vertices (by removing $\hat{e}$ we disconnect the vertex v_l from the vertices $v_1,\ldots,v_{l-1}$, but we do not disconnect any of these). Thus

$$n(G_l)=-\sum_{\hat{e}\in\hat{E}_l}\ \sum_{(V_{l-1},\tilde{E})\in\mathcal{C}_{l-1}}(-1)^{|\tilde{E}|}=-n(G_{l-1})\cdot|\hat{E}_l|,$$

proving (16). □

Remark: As the example of the graph $G=(V,E)$ of a square with $V:=\{v_1,v_2,v_3,v_4\}$ and $E:=\{\{v_1,v_2\},\{v_2,v_3\},\{v_3,v_4\},\{v_4,v_1\}\}$ shows, an analogous result does not hold for general graphs. Namely here $n(G_3)=1-0=1$ but $n(G)=1-4=-3$, although there are only two edges adjacent to the vertex v_4.

It was already shown in Ref. 1 that the (grand) canonical interaction is weakly ferromagnetic, i.e., that

$$j_k^C(t)\geqslant 0\quad\text{and}\quad j_k^G(t)\geqslant 0\quad\text{for}\quad t\in\mathbf{G}_k^*\backslash\{0\}.$$

That fact can now be easily seen to be a by-product of our above theorem for interval graphs $G=(V,E)$, since Theorem 4 implies

$$\operatorname{sign}(n(G))=\begin{cases}0, & G \text{ not connected},\\ -(-1)^{|V|}, & G \text{ connected}.\end{cases}$$

On the other hand, the sign of the activity of a multi-polymer X is $\operatorname{sign}(z^X)=(-1)^{|X|}$. Thus all connected multi-polymers give a positive contribution to the interaction coefficients (14) and (15).

By estimating the contribution of the simplest multi-polymers, we may obtain more precise lower bounds for the interaction coefficients. As in Ref. 1 we denote the positions of the first, resp. last, one in $t=(t_1,\ldots,t_k)\in\mathbf{G}_k^*\backslash\{0\}$ by

$$P_k^l(t):=\min\{i\in\{1,\ldots,k\}|t_i=1\}$$

and

$$P_k^r(t):=\max\{i\in\{1,\ldots,k\}|t_i=1\}.$$

Corollary 5: Let $t=(t_1,\ldots,t_k)\in\mathbf{G}_k^*\backslash\{0\}$.
(i) *If* $|t|$ *is odd, then* $j_k^G(t)=0$ *and* $j_k^C(t)\geqslant 3^{-2s}$, *with* $s:=k+1-P_k^l(t)$.
(ii) *If* $|t|$ *is even, then* $j_k^G(t)\geqslant 3^{-2s}$ *and* $j_k^C(t)\geqslant 3^{-2s}$, *with* $s:=P_k^r(t)+1-P_k^l(t)$.

The grand canonical interaction coefficients increase under concatenation, that is,

$$j_{k+l}^G(s,t)\geqslant \delta_{t,0}\cdot j_k^G(s)+\delta_{s,0}\cdot j_l^G(t)\quad (s\in\mathbf{G}_k^*,t\in\mathbf{G}_l^*). \tag{19}$$

Proof: We already stated that $j_k^G(t)=0$ for $|t|$ odd. The other lower bounds are achieved by first decomposing t according to (8) into disconnected polymers $t=\Sigma_{i=1}^n p_{l_i,r_i}$ for $|t|$ even, resp. $t=\Sigma_{i=1}^{n-1}p_{l_i,r_i}+p_{l_n}$ for $|t|$ odd. Then one remarks that for
(i) $|t|$ even the connected multi-polymer

$$X=(\gamma_1,\ldots,\gamma_n):=(p_{l_1,r_n},p_{r_1,l_2},\ldots,p_{r_{n-1},l_n})\in C_n^{\infty,e},$$

and for
(ii) $|t|$ odd the connected multi-polymer

$$X=(\gamma_1,\ldots,\gamma_n):=(p_{l_1},p_{r_1,l_2},\ldots,p_{r_{n-1},l_n})\in C_n^{\infty},$$

and their $n!$ different permutations have the property $\Sigma_{i=1}^n\gamma_i=t$. So they give contributions to $j_k^C(t)$ and $j_k^G(t)$ in (14) and (15).

The factor $|X|!=n!$ in the denominator is cancelled by the multiplicity of the permutations. The graphs $G(X)$ are trees, namely all γ_i, $i>1$, are only connected to γ_1. So, $n(X)\equiv n(G(X))=-(-1)^n$. The activities are $z^X=(-1)^n\cdot 3^{l_1-r_n-1}\cdot 3^{r_1-l_2-1}\cdots 3^{r_{n-1}-l_n-1}$ for $|t|$ even resp. $z^X=(-1)^n\cdot 3^{l_1-k-1}\cdot 3^{r_1-l_2-1}\cdots 3^{r_{n-1}-l_n-1}$ for $|t|$ odd. Thus multiplying these factors leads to the estimates of the corollary.

One obtains inequality (19) by noticing that in the grand canonical case the activities of the occurring (even) polymers are translation invariant. If both s and t are nonzero, then (19) simply states that $j_{k+l}^G(s,t)\geqslant 0$, i.e., the ferromagnetic property. If s or t is zero, then all connected multi-polymers contributing to the right-hand side of (19) contribute to the lhs, too (after a translation by k for the t-terms). □.

Remark: Proposition 3 of Ref. 2 says that, for nonzero $t\in\mathbf{G}_n^*$,

$$j_{n+1}^C(t,0)\leqslant j_n^C(t)\leqslant j_{n+1}^C(0,t).$$

This implies in particular that the canonical interaction coefficients neither increase nor decrease under concatenation.

V. ESTIMATES FOR THE FREE ENERGY

In this section we will show that the limit free energy is majorized by the canonical and minimized by the grand canonical k-spin free energies.

In order to derive these estimates, we compare the (grand) canonical energy values of the spin chain of length $k+l$ with the ones of the chains with lengths k and l. We state these inequalities for the exponentiated energies.

Lemma 6: For $\sigma \in \mathbf{G}_k$, $\tau \in \mathbf{G}_l$ *one has*

$$\mathbf{h}^G_{k+l}(\sigma,\tau) \leqslant \mathbf{h}^G_k(\sigma)\cdot \mathbf{h}^G_l(\tau), \tag{20}$$

$$\mathbf{h}^C_{k+l}(\sigma,\tau)+\mathbf{h}^C_{k+l}(1-\sigma,\tau)=(\mathbf{h}^C_k(\sigma)+\mathbf{h}^C_k(1-\sigma))\cdot \mathbf{h}^C_l(\tau) \tag{21}$$

and

$$|\mathbf{h}^C_{k+l}(\sigma,\tau)-\mathbf{h}^C_{k+l}(1-\sigma,\tau)| \leqslant |\mathbf{h}^C_k(\sigma)-\mathbf{h}^C_k(1-\sigma)|\cdot \mathbf{h}^C_l(\tau). \tag{22}$$

Proof: As in Ref. 1 we use the auxiliary functions $\mathbf{r}_k(s_0,s_1):\mathbf{G}_k \to \mathbb{R}$ with real parameters s_0,s_1 which are defined by $\mathbf{r}_1(s_0,s_1)(\sigma):=s_\sigma$ and

$$\mathbf{r}_{k+1}(s_0,s_1)(\sigma_1,\ldots,\sigma_{k+1}):=\mathbf{r}_k(s_0,s_1)(\sigma_1,\ldots,\sigma_k) + \sigma_{k+1}\cdot \mathbf{r}_k(s_0,s_1)(1-\sigma_1,\ldots,1-\sigma_k).$$

In particular we have $\mathbf{h}^C_k=\mathbf{r}_k(1,2)$, $k\in\mathbb{N}$, and $\mathbf{h}^G_k(\sigma)=\mathbf{h}^C_{k+1}(\sigma,1)=\mathbf{r}_{k+1}(1,2)(\sigma,1)$, $\sigma\in\mathbf{G}_k$.

By Lemma 4.6 of Ref. 1

$$\mathbf{r}_{k+l}(s_0,s_1)(\sigma,\tau)=\mathbf{r}_{l+1}(\mathbf{r}_k(s_0,s_1)(\sigma),\mathbf{r}_k(s_0,s_1)(1-\sigma))(0,\tau).$$

Moreover, we have for $\rho\in\mathbf{G}_l$

$$\mathbf{r}_{l+1}(s,s)(0,\rho)=\mathbf{r}_l(s,2s)(\rho),$$

since for $l=1$ both sides equal $s\cdot(1+\rho)$, whereas for $\rho'\in\mathbf{G}_l$, $\rho''\in\mathbf{G}_1$, and $\rho=(\rho',\rho'')\in\mathbf{G}_{l+1}$,

$$\begin{aligned}\mathbf{r}_{l+2}(s,s)(0,\rho)&=\mathbf{r}_{l+1}(s,s)(0,\rho')+\rho''\mathbf{r}_{l+1}(s,s)(1,1-\rho')\\ &=\mathbf{r}_l(s,2s)(\rho')+\rho''\mathbf{r}_{l+1}(s,s)(0,1-\rho')\\ &=\mathbf{r}_l(s,2s)(\rho')+\rho''\mathbf{r}_l(s,2s)(1-\rho')\\ &=\mathbf{r}_{l+1}(s,2s)(\rho).\end{aligned}$$

Now we apply these identities:

$$\begin{aligned}\mathbf{h}^G_{k+l}(\sigma,\tau)&=\mathbf{r}_{k+l+1}(1,2)(\sigma,\tau,1)\\ &=\mathbf{r}_{l+2}(\mathbf{r}_k(1,2)(\sigma),\mathbf{r}_k(1,2)(1-\sigma))(0,\tau,1)\\ &=\mathbf{r}_{l+2}(\mathbf{h}^C_k(\sigma),\mathbf{h}^C_k(1-\sigma))(0,\tau,1)\\ &=\mathbf{h}^C_k(\sigma)\cdot\mathbf{r}_{l+2}(1,0)(0,\tau,1)+\mathbf{h}^C_k(1-\sigma)\cdot\mathbf{r}_{l+2}(0,1)(0,\tau,1)\\ &\leqslant(\mathbf{h}^C_k(\sigma)+\mathbf{h}^C_k(1-\sigma))\cdot\mathbf{r}_{l+2}(1,1)(0,\tau,1)\\ &=\mathbf{h}^G_k(\sigma)\cdot\mathbf{r}_{l+2}(1,1)(0,\tau,1)\\ &=\mathbf{h}^G_k(\sigma)\cdot\mathbf{r}_{l+1}(1,2)(\tau,1)\\ &=\mathbf{h}^G_k(\sigma)\cdot\mathbf{h}^G_l(\tau),\end{aligned}$$

showing (20).

Similarly

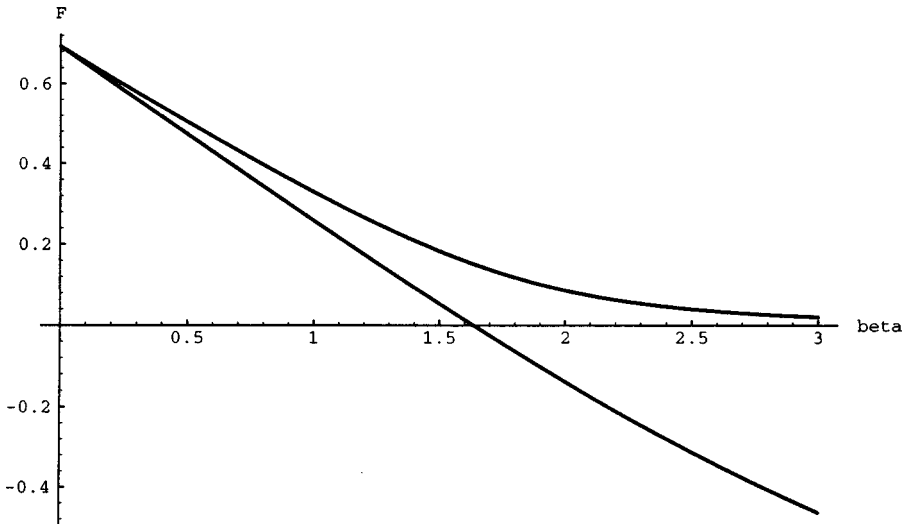

FIG. 2. Upper and lower bounds on the free energy $F(\beta)$.

$$\mathbf{h}^C_{k+l}(\sigma,\tau)=\mathbf{h}^C_k(\sigma)\cdot\mathbf{r}_{l+1}(1,0)(0,\tau)+\mathbf{h}^C_k(1-\sigma)\cdot\mathbf{r}_{l+1}(0,1)(0,\tau)$$

so that

$$\begin{aligned}\mathbf{h}^C_{k+l}(\sigma,\tau)+\mathbf{h}^C_{k+l}(1-\sigma,\tau)&=(\mathbf{h}^C_k(\sigma)+\mathbf{h}^C_k(1-\sigma))\cdot(\mathbf{r}_{l+1}(1,0)(0,\tau)+\mathbf{r}_{l+1}(0,1)(0,\tau))\\&=(\mathbf{h}^C_k(\sigma)+\mathbf{h}^C_k(1-\sigma))\cdot\mathbf{r}_{l+1}(1,1)(0,\tau)\\&=(\mathbf{h}^C_k(\sigma)+\mathbf{h}^C_k(1-\sigma))\cdot\mathbf{r}_l(1,2)(\tau),\end{aligned}$$

which implies (21).

The difference

$$\begin{aligned}|\mathbf{h}^C_{k+l}(\sigma,\tau)-\mathbf{h}^C_{k+l}(1-\sigma,\tau)|&=|\mathbf{h}^C_k(\sigma)-\mathbf{h}^C_k(1-\sigma)|\cdot|\mathbf{r}_{l+1}(1,0)(0,\tau)-\mathbf{r}_{l+1}(0,1)(0,\tau)|\\&\leqslant|\mathbf{h}^C_k(\sigma)-\mathbf{h}^C_k(1-\sigma)|\cdot\mathbf{h}^C_l(\tau),\end{aligned}$$

which is (22). □

Theorem 7: *For inverse temperature $\beta>0$ the limit free energy $F(\beta)$ is bounded by the k-spin free energies, $k\in\mathbb{N}$:*

$$F^C_k(\beta)\leqslant F(\beta)\leqslant F^G_k(\beta).$$

(*See Fig. 2, where* $k=15$).

Proof: We have for $n\in\mathbb{N}$

$$F^G_{nk}(\beta)=-\frac{1}{\beta k}\ln(Z^G_{nk}(\beta))^{1/n}\leqslant-\frac{1}{\beta k}\ln(Z^G_k(\beta))=F^G_k(\beta),$$

since by (20)

$$Z^G_{k+l}(\beta)=\sum_{\sigma\in\mathbf{G}_k}\sum_{\tau\in\mathbf{G}_l}(\mathbf{h}^G_{k+l}(\sigma,\tau))^{-\beta}$$

$$\geqslant\sum_{\sigma\in\mathbf{G}_k}\sum_{\tau\in\mathbf{G}_l}(\mathbf{h}^G_k(\sigma)\cdot\mathbf{h}^G_l(\tau))^{-\beta}$$

$$=Z^G_k(\beta)\cdot Z^G_l(\beta).$$

So $F(\beta)=\lim_{n\to\infty}F^G_{nk}(\beta)\leqslant F^G_k(\beta)$.

To show the lower estimate we need the converse inequality

$$Z^C_{k+l}(\beta)\leqslant Z^C_k(\beta)\cdot Z^C_l(\beta), \tag{23}$$

which then implies $F(\beta)=\lim_{n\to\infty}F^C_{nk}(\beta)\geqslant F^C_k(\beta)$.

To show (23), we use (21) and (22). For $a,b,c,d>0$, $a+b=c+d$, and $|a-b|\leqslant|c-d|$ the convexity of the function $x\mapsto x^{-\beta}$ implies $a^{-\beta}+b^{-\beta}\leqslant c^{-\beta}+d^{-\beta}$. Thus

$$(\mathbf{h}^C_{k+l}(\sigma,\tau))^{-\beta}+(\mathbf{h}^C_{k+l}(1-\sigma,\tau))^{-\beta}\leqslant(\mathbf{h}^C_k(\sigma)^{-\beta}+\mathbf{h}^C_k(1-\sigma)^{-\beta})\cdot\mathbf{h}^C_l(\tau)^{-\beta},$$

which after summation over σ and τ leads to (23). □

VI. THE INTERNAL ENERGY

From the estimates for the free energy we will derive upper and lower bounds for the internal energy per particle

$$U(\beta)=\lim_{k\to\infty}\frac{1}{k}\langle\mathbf{H}^C_k\rangle^C_k(\beta)=\lim_{k\to\infty}\frac{1}{k}\langle\mathbf{H}^G_k\rangle^G_k(\beta).$$

The first thing to say is that, in general, the k-spin energy expectation $\langle H_k\rangle(\beta)$ is *not* a monotone increasing function of the energy values $H_k(\sigma)$. But if

$$H_k(\sigma)=-\sum_{t\in\mathbf{G}^*_k}j(t)\cdot(-1)^{\sigma\cdot t}$$

is weakly ferromagnetic ($j(t)\geqslant 0$ for $t\neq 0_k$), then by the first and second GKS inequality for $\beta\geqslant 0$ and $s\in\mathbf{G}^*_k$

$$\frac{\partial\langle H_k\rangle}{\partial j(s)}=-\langle\chi_s\rangle-\beta\sum_{t\in\mathbf{G}^*_k}j(t)(\langle\chi_s\chi_t\rangle-\langle\chi_s\rangle\langle\chi_t\rangle)\leqslant 0 \tag{24}$$

with the character functions $\chi_s:\mathbf{G}_k\to\{-1,1\}$,

$$\chi_s(\sigma):=(-1)^{\sigma\cdot s}=\prod_{i:s_i=1}(-1)^{\sigma_i}.$$

Theorem 8: *For $\beta\geqslant 0$ and $k\in\mathbb{N}$*

$$U(\beta)\leqslant\frac{1}{k}\langle\mathbf{H}^G_k\rangle^G_k(\beta).$$

Proof: We start from the inequality (19),

$$j^G_{k+l}(s,t)\geqslant\delta_{t,0}\cdot j^G_k(s)+\delta_{s,0}\cdot j^G_l(t)\quad(s\in\mathbf{G}^*_k,t\in\mathbf{G}^*_l),$$

for the interaction coefficients. Thus by (24)

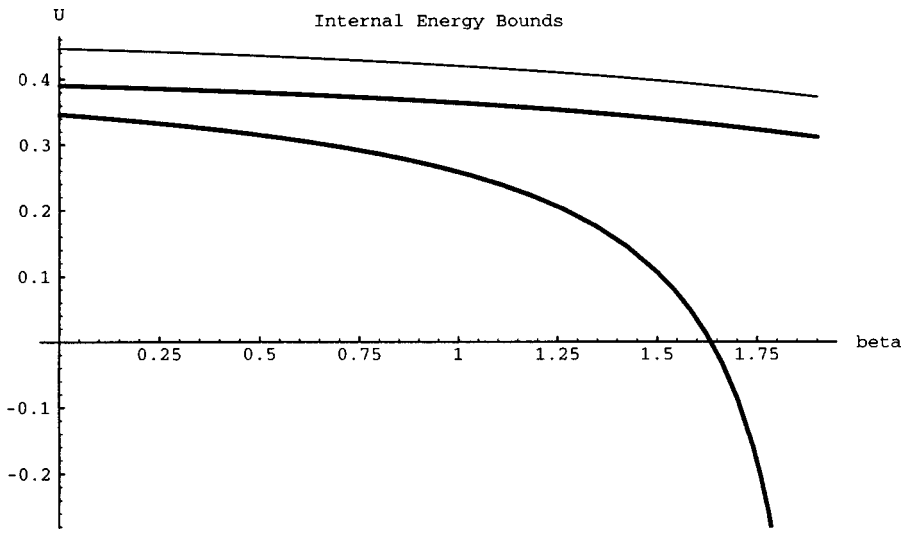

FIG. 3. Upper and lower bounds on the internal energy $U(\beta)$ (bold: Theorem 9).

$$\langle \mathbf{H}^G_{k+l}\rangle^G_{k+l} \leqslant \langle \mathbf{H}^G_k\rangle^G_k + \langle \mathbf{H}^G_l\rangle^G_l .$$

In particular we have $(1/n)\,\langle \mathbf{H}^G_{nk}\rangle^G_{nk} \leqslant \langle \mathbf{H}^G_k\rangle^G_k$ so that the theorem follows. □

We can also use the concavity of the function $\beta \mapsto \beta F(\beta)$ and our knowledge of the two values $\lim_{\beta \searrow 0} \beta F(\beta) = -\ln 2$ and $\beta_{\rm cr} F(\beta_{\rm cr}) = 0$ for the inverse critical temperature $\beta_{\rm cr} = 2$.

Theorem 9: *For* $0 < \beta < \beta_{\rm cr} = 2$ *and* $k \in \mathbb{N}$

$$\frac{-\beta F^G_k(\beta)}{2-\beta} \leqslant U(\beta) \leqslant \frac{\ln(2) + \beta F^G_k(\beta)}{\beta}.$$

Proof: In the above inverse temperature range the free energy is real-analytic, as shown in Ref. 6. Thus we can write the internal energy in the form

$$U(\beta) = \frac{d}{d\beta}(\beta F(\beta)).$$

By concavity of the function $\beta \mapsto \beta F(\beta)$ the derivative is bounded by the slope of the secants:

$$\frac{\beta_{\rm cr} F(\beta_{\rm cr}) - \beta F(\beta)}{\beta_{\rm cr} - \beta} \leqslant U(\beta) \leqslant \frac{\beta F^G_k(\beta) - (-\ln(2))}{\beta - 0}.$$

The bounds then follow from this and Theorem 7. □

In Fig. 3 we show the different bounds on the internal energy $U(\beta)$ for $k = 15$.

Although for this number k of spins Theorem 9 leads to a better upper bound than Theorem 8, only the latter estimate converges in the thermodynamic limit $k \to \infty$.

Since the methods used here are high-temperature expansions, the lower estimate for the internal energy derived in Ref. 6 is better near the critical temperature $\beta = 2$ but worse near $\beta = 0$.

ACKNOWLEDGMENTS

The authors would like to thank the people of the Erwin-Schrödinger-Institut for their hospitality. A. K. thanks Pierluigi Contucci for introducing him to polymer models and for proposing their use in Number Theory.

[1] A. Knauf, ''On a Ferromagnetic Spin Chain,'' Commun. Math. Phys. **153**, 77–115 (1993).
[2] A. Knauf, ''On a Ferromagnetic Spin Chain. Part II: Thermodynamic Limit,'' J. Math. Phys. **35**, 228–236 (1994).
[3] A. Knauf, ''Phases of the Number-Theoretical Spin Chain,'' J. Stat. Phys. **73**, 423–431 (1993).

[4] A. Knauf, ''Irregular Scattering, Number Theory, and Statistical Mechanics,'' in *Stochasticity and Quantum Chaos*, edited by Z. Haba *et al.* (Kluwer, Dordrecht, 1995).
[5] A. Knauf, ''The Number-Theoretical Spin Chain and the Riemann Zeroes,'' Max-Planck-Institute for Mathematics in the Sciences Preprint No. 15 (http://www.mis.mpg.de) (1997), to appear in Commun. Math. Phys.
[6] P. Contucci, and A. Knauf, ''The Phase Transition of the Number-Theoretical Spin Chain,'' Forum Mathematicum **9**, 547–567 (1997).
[7] P. Contucci and A. Knauf, ''The Low Activity Phase of Some Dirichlet Series,'' J. Math. Phys. **37**, 5458–5475 (1996).
[8] R. Artuso, P. Cvitanović, and B. G. Kenny, ''Phase Transitions on Strange Irrational Sets,'' Phys. Rev. A **39**, 268–281 (1989).
[9] M. Feigenbaum, ''Presentation Functions, Fixed Points, and a Theory of Scaling Function Dynamics,'' J. Stat. Phys. **52**, 527–569 (1988).
[10] P. Cvitanović ''Circle Maps: Irrationally Winding,'' in *From Number Theory to Physics*, edited by M. Waldschmidt, P. Moussa, J. M. Luck, and C. Itzykson (Springer, Berlin, 1992).
[11] G. Gallavotti, A. Martin-Löf, and S. Miracle-Solé, ''Some Problems Connected with the Description of Coexisting Phases at low Temperatures in the Ising Model,'' in *Statistical Mechanics and Mathematical Problems* (Batelle, 1971), edited by A. Lenard, Lecture Notes in Physics, Vol. 20 (Springer, Berlin, 1973).
[12] B. Simon, *The Statistical Mechanics of Lattice Gases*, Princeton Series in Physics (Princeton U.P., NJ, 1993), Vol. I.
[13] J. Glimm and A. Jaffe, *Quantum Physics*, 2nd ed. (Springer, New York, 1987).

Generalized similarity solutions for the type D fluid in five-dimensional flat space

O. P. Bhutani[a] and K. Singh[b]
Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India

(Received 29 August 1997; accepted for publication 18 February 1998)

The partial differential equation corresponding to the metric function in a five-dimensional flat space describing the perfect fluid distribution has been examined for generalized symmetries via a new approach. It is shown that the approach utilized here has, besides providing generalized symmetries, for some particular values of the parameters involved, yielded new metrics and Petrov type D solutions. Further, for the general case the problem is reduced to the solution of an Abel's equation of second kind, which obviously generalizes the scope for new metrics and new Petrov type D solutions. © *1998 American Institute of Physics.* [S0022-2488(98)00306-5]

I. INTRODUCTION

This is a sequel to the paper by Bhutani and Kalra,[1] in which we demonstrated the efficacy and importance of the generalized symmetry approach. In fact, the said approach[2,3] combined with variational symmetry[4] has not only provided us a breakthrough in obtaining the exact solutions of the coupled system of nonlinear partial differential equations corresponding to Einstein exterior equations, which, in turn, yielded new solutions of Ernst form of Einstein equations, but has also enabled us to arrive at a generalized form of Weyl[5] and Schwarzschild[6] solutions.

Here, we have utilized the same technique for examining a problem from the realm of general relativity recently investigated[7] using the Lie continuous group of point transformation method.[8] More specifically, the authors[7] have derived solutions of type D for the perfect fluid distribution expressed by a five-dimensional flat metric of embedding class I. The carryover of the generalized symmetry technique has not only provided the generalized symmetries but has also enlarged the scope for the determination of type D solutions, the importance of which is well known.

The paper is structured as follows. The basic equations are briefly recalled in Sec. II. In Sec. III, the symmetries in the generalized form are derived, which are then used to reduce the main partial differential equation (pde) to a general second-order ordinary differential equation (ode). Some special cases are treated in Sec. IV. Section V is primarily devoted to determining the transformation group for a further reduction of second-order ode to a general first-order ode. We end this paper with some discussions. The detailed computations regarding the Fréchet derivative are shown in Appendix A.

II. RECALLING THE MAIN EQUATIONS

A five-dimensional flat metric,

$$ds^2=-(dx_1)^2-(dx_2)^2-(dx_3)^2+(dx_4)^2\pm(dx_5)^2, \tag{2.1}$$

on introducing the plane symmetry by

$$x_1=(t-r)\theta\cos\phi, \quad x_2=(t-r)\theta\sin\phi,$$

[a]Electronic mail: bhutani@maths.iitd.ernet.in
[b]Electronic mail: karan@maths.iitd.ernet.in

 © 1998 American Institute of Physics

$$x_3=\frac{\theta^2}{2}(t-r)+r, \quad x_4=\frac{\theta^2}{2}(t-r)+t,$$

$$x_5=u(r,t), \tag{2.2}$$

assumes the following form:

$$ds^2=-t^2(d\theta^2+\theta^2 d\phi^2)+dt^2+2drdt-du^2. \tag{2.3}$$

The metric function $u(r,t)$ is governed by the differential equation,

$$L(u)\equiv -u_{rr}u_{tt}+u_{rt}^2+\frac{1}{t}u_r^3u_{tt}+\frac{2}{t}u_{rt}u_r(1-u_ru_t)$$

$$-\frac{1}{t}(1-u_t^2)u_ru_{rr}+\frac{1}{t^2}u_r^2(1-2u_ru_t+u_r^2)=0. \tag{2.4}$$

The associated expressions for pressure and density are given by

$$8\pi p=\frac{u_r^2}{mt^2}, \tag{2.5}$$

$$8\pi(\rho-p)=\frac{2(u_{rr}u_{tt}-u_{rt}^2)}{m^2}, \tag{2.6}$$

$$\text{where} \quad m=1+u_r^2-2u_ru_t. \tag{2.7}$$

For further details see Ref. 7.

III. DETERMINATION OF SYMMETRIES

Since for the case under consideration there is only one spatial variable, the symmetry operator has the form

$$S(u)=A(r,t,u)\frac{\partial u}{\partial t}+B(r,t,u)\frac{\partial u}{\partial r}+C(r,t,u). \tag{3.1}$$

Calculating the Fréchet derivative of L defined in Eq. (2.4), we get

$$F(L,u,\nu)=\nu_{tt}\left(\frac{u_r^3}{t}-u_{rr}\right)+\frac{2}{t}u_r\nu_t\left(u_tu_{rr}-u_ru_{rt}-\frac{1}{t}u_r^2\right)$$

$$+\frac{1}{t}\nu_r\left[3u_r^2u_{tt}+2u_{rt}-u_{rr}(1-u_t^2)+2u_r\left(\frac{1}{t}-2u_tu_{rt}\right)-\frac{2}{t}u_r^2(3u_t-2u_r)\right]$$

$$+2\nu_{rt}\left(u_{rt}+\frac{1}{t}u_r-\frac{1}{t}u_r^2u_t\right)-\nu_{rr}\left[u_{tt}+\frac{u_r}{t}(1-u_t^2)\right]. \tag{3.2}$$

On replacing ν by $S(u)$ in Eq. (3.2) and substituting the values of different derivatives of $S(u)$ in the resulting equation and collecting the coefficients of various powers of the derivatives of u in F, we get a polynomial expansion in u_r, u_t, u_ru_t,..., etc. On making use of Eq. (2.4) in the polynomial expression for F, rearranging the terms of various powers of derivatives of u, and equating them to zero, leads to the following determining equations (see Appendix A):

$$A_{rr}=0=A_u, \quad C_{rr}=0, \quad C_{rt}+\frac{1}{t}C_r=0,$$

$$C_{tt}+\frac{1}{t}C_r=0,\quad A_{uu}-\frac{1}{t}A_r=0,\quad A_{rt}+\frac{1}{t}A_r+C_{ru}=0,$$

$$A_{tt}+\frac{1}{t}A_r+2C_{tu}=0,\quad 2A_{tu}+C_{uu}-\frac{1}{t}C_r=0,\quad 2C_{ru}+B_{rr}-\frac{2}{t}A_r=0,$$

$$B_{tu}-\frac{1}{t}(B_u+A_u+C_t)=0,\quad B_{uu}-\frac{1}{t}B_r-\frac{2}{t}C_u-\frac{A}{t^2}=0,$$

$$C_{uu}+2B_{ru}-\frac{1}{t}(3C_r+2A_u)=0,\quad 2A_{tu}+\frac{2}{t}A_u-2C_{rt}-3B_{ru}=0,$$

$$B_{rt}+C_{tu}-\frac{1}{t}A_t-\frac{1}{t}A_r+\frac{A}{t^2}=0,\quad B_{tt}+\frac{1}{t}(B_r-2B_t-2A_t)+\frac{A}{t^2}=0,$$

$$A_{tt}-2B_{rt}+\frac{1}{t}(3A_r+2A_t)-\frac{2A}{t^2}=0,\quad C_{rt}-C_{uu}-B_{ru}-A_{tu}+\frac{1}{t}(A_u-3C_r)=0. \tag{3.3}$$

These equations, when solved, yield the following generalized Lie symmetries of the equation (2.4):

$$A=\alpha t,$$

$$B=b_1 r+\beta t^3+(1/2)(b_1-\alpha)t+\gamma, \tag{3.4}$$

$$C=-(1/2)(\alpha+b_1)u-\delta,$$

where α, β, γ, δ, and b_1 are arbitrary constants.

It is interesting to point out here that the case $b_1=\alpha$ corresponds to the symmetries that are reported in Ref. 7.

In our opinion it is a significant improvement in the available symmetries of the equation (2.4) as it involves an additional parameter and hence widens the scope of obtaining new possible metrics. In the next section it has been demonstrated through special cases, by way of generating a class of metrics belonging to Petrov type D. Before that, we determine the similarity variable $\xi=\xi(r,t)$ and new dependent variable $f(\xi)$, in terms of which Eq. (2.4) gets reduced to an ode, by solving, for the general case, the equations

$$\frac{dt}{\alpha t}=\frac{dr}{b_1 r+\beta t^3+(1/2)(b_1-\alpha)t+\gamma}=\frac{du}{(1/2)(b_1+\alpha)u+\delta}, \tag{3.5}$$

which yield

$$u=t^{(b_1+\alpha)/2\alpha}f(\xi)-\frac{2\delta}{b_1+\alpha}, \tag{3.6}$$

where

$$\xi(r,t)=t^{-b_1/\alpha}\left[r+\frac{\beta}{(b_1-3\alpha)}t^3+\frac{1}{2}t+\frac{\gamma}{b_1}\right]. \tag{3.7}$$

On substituting (3.6) in Eq. (2.4), we get

$$f''[(\alpha^2-b_1^2)f+(\alpha+b_1)^2 f^2 f'+4b_1(b_1-3\alpha)\xi f']$$
$$-4b_1(b_1-3\alpha)\xi f'^4+(\alpha+b_1)(3b_1-7\alpha)ff'^3+(b_1-3\alpha)^2 f'^2=0. \tag{3.8}$$

As Eq. (3.8) is highly nonlinear in character, it cannot be solved through available techniques for all values of the parameters involved in it. Consequently, some particular cases are in order.

IV. SOME SPECIAL CASES

A. Case (i) $b_1=-\alpha\neq 0$, $\delta\to 0$

For the case under consideration, ode (3.8) reduces to

$$\xi(f''-f'^3)+f'=0. \tag{4.1}$$

On solving the above equation, we get

$$f'(\xi)=\frac{\sqrt{c_0}}{\sqrt{(\xi+c_0)^2-c_0^2}}, \tag{4.2}$$

where c_0 is a positive constant.

Equation (4.2) readily yields

$$f(\xi)=\sqrt{c_0}\,\ln[\xi+c_0+\sqrt{\xi(\xi+2c_0)}]+d_0, \tag{4.3}$$

where d_0 is a constant.

Thus, we have obtained $u=f(\xi)$, with

$$\xi=\frac{t}{4b_1}(4b_1r+\beta t^3+2b_1t+4\gamma).$$

The metric (2.3) can now be evaluated as follows:

$$\begin{aligned}ds^2=&-\frac{16b_1^2c_0t}{(4b_1r+\beta t^3+2b_1+4\gamma)\{t(4b_1r+\beta t^3+2b_1t+4\gamma)+8c_0b_1\}}\,dr^2\\&+2\left[1-\frac{16b_1c_0(b_1r+\beta t^3+b_1t+\gamma)}{(4b_1r+\beta t^3+2b_1t+4\gamma)\{t(4b_1r+\beta t^3+2b_1t+4\gamma)+8c_0b_1\}}\right]dr\,dt\\&+\left[1-\frac{16c_0(b_1r+\beta t^3+b_1t+\gamma)^2}{t(4b_1r+\beta t^3+2b_1t+4\gamma)\{t(4b_1r+\beta t^3+2b_1t+4\gamma)+8c_0b_1\}}\right]dt^2\\&-t^2(d\theta^2+\theta^2d\phi^2).\end{aligned} \tag{4.4}$$

The corresponding expressions for pressure and density can be expressed as

$$8\pi p=\frac{16b_1^2c_0}{t^2[16b_1^2r^2+8b_1r(\beta t^3+2b_1t+4\gamma)+\beta^2t^6+4\beta b_1t^4+8\beta\gamma t^3+4b_1(b_1-6c_0\beta)t^2+16b_1\gamma t+16\gamma^2]} \tag{4.5}$$

and

$$8\pi(\rho-p)=\frac{32b_1^2c_0[16b_1^2r^2-16b_1r(\beta t^3-b_1t-2\gamma)-5\beta^2t^6-8\beta b_1t^4-16\beta\gamma t^3+4b_1(b_1-6c_0\beta)t^2+16b_1\gamma t+16\gamma^2]}{t^2[16b_1^2r^2+8b_1r(\beta t^3+2b_1t+4\gamma)+\beta^2t^6+4\beta b_1t^4+8\beta\gamma t^3+4b_1(b_1-6c_0\beta)t^2+16b_1\gamma t+16\gamma^2]^2}. \tag{4.6}$$

It is not difficult to see that the solution obtained is of Petrov type D. Further, it is shown in the later section that the metric (4.4) corresponds to nongeodesic flow.

B. Case (ii) $b_1=3\alpha\neq 0$, $\beta\to 0$

For this case, Eq. (3.8) assumes the following form:

$$f''-2ff'f''-f'^3=0. \tag{4.7}$$

On introducing $Z=f'$ in (4.7) and integrating the resulting equation, we get

$$Z^2 f = Z - a, \tag{4.8}$$

where ''a'' is a constant of integration.

(A) For the case when a is zero, Eq. (4.8) yields

$$f(\xi) = (2\xi + c_1)^{1/2}.$$

Thus, $u(r,t)$ assumes the following form:

$$u = t^2(2\xi + c_1)^{1/2} - \frac{\delta}{2\alpha}. \tag{4.9}$$

The corresponding metric, and the expressions for pressure and density, can be written as

$$ds^2 = -\frac{3\alpha t}{(6\alpha r + 3c_1\alpha t^3 + 3\alpha t + 2\gamma)}\,dr^2 + \frac{6(\alpha r - c_1\alpha t^3 - \gamma)}{(6\alpha r + 3c_1\alpha t^3 + 3\alpha t + 2\gamma)}\,dr dt + \left[1 - \frac{(3\alpha r + 6c_1\alpha t^3 + 3\alpha t + \gamma)^2}{3\alpha t(6\alpha r + 3c_1\alpha t^3 + 3\alpha t + 2\gamma)}\right]dt^2 - t^2(d\theta^2 + \theta^2 d\phi^2), \tag{4.10}$$

$$8\pi\rho = 3(8\pi p) = -\frac{1}{c_1 t^4}. \tag{4.11}$$

Clearly, (4.11) demands that $c_1 < 0$. The solution obviously belongs to Petrov type O.

(B) For $a \neq 0$, Eq. (4.8) can be written as

$$\frac{f\,df}{1 \pm \sqrt{1 - 4af}} = \frac{1}{2}\,d\xi, \tag{4.12}$$

which on integrating yields

$$6af \pm (1 - 4af)^{3/2} = 12a^2\xi + b. \tag{4.13}$$

On letting $f = (1 - q^2)/4a$, Eq. (4.13) can be expressed in the following convenient form:

$$q^2(\pm 2q - 3) = 24a^2\xi + 2b - 3. \tag{4.14}$$

In Eq. (4.14), ''b'' is a constant of integration.

The metric (2.3) in terms of implicit function $q = q(\xi)$ assumes the following forms:

$$ds^2 = -\frac{4a^2}{t^2(1 \mp q)^2}\,dr^2 + 2\left[1 - \frac{\{\alpha t^3(1 \mp q)^2(1 \pm q) - 4a^2(3\alpha r + \alpha t + \gamma)\}}{\alpha t^3(1 \mp q)^2}\right]dr dt + \left[1 - \frac{\{\alpha t^3(1 \mp q)^2(1 \pm q) - 4a^2(3\alpha r + \alpha t + \gamma)\}^2}{4a^2\alpha^2 t^4(1 \mp q)^2}\right]dt^2 - t^2(d\theta^2 + \theta^2 d\phi^2). \tag{4.15}$$

The pressure and density can be evaluated as

$$8\pi p = \frac{4a^2\alpha}{t[4a^2\alpha t - \alpha t^3(1 \mp q)^2(1 \pm 2q) + 8a^2(3\alpha r + \alpha t + \gamma)]}, \tag{4.16}$$

$$8\pi(\rho - p) = \frac{8a^2\alpha[4a^2(6\alpha r + 3\alpha t + 2\gamma) \mp \alpha t^3(1 \mp q)^2(1 \pm 2q)]}{tq[4a^2\alpha t - \alpha t^3(1 \mp q)^2(1 \pm 2q) + 8a^2(3\alpha r + \alpha t + \gamma)]^2}. \tag{4.17}$$

It is demonstrated in later sections that the metric (4.15) belongs to nongeodesic fluid distribution.

C. Case (iii) $\alpha \neq 0$, $b_1 = 0$, $\gamma \to 0$

The related ode for the case can be expressed as

$$ff'' + f^2 f' f'' - 7ff'^3 + 9f'^2 = 0. \tag{4.18}$$

Observing the invariance of (4.18) under the group of translation of independent variable ξ, we introduce $f' = \psi$ so that the equation (4.18) reduces to the following first-order equation:

$$\psi(9 - 7f\psi)df + f(1 + f\psi)d\psi = 0. \tag{4.19}$$

It can be easily integrated by multiplying with a suitable integrating factor to yield the following:

$$2f^9 f' = c(ff' - 1)^2, \tag{4.20}$$

where c is a positive constant. Equation (4.20) can be integrated further to give the following solution in the implicit form:

$$f^{10} + 5cf^2 \mp 10\int f^5 \sqrt{f^8 + 2c}\, df = 10c\xi + d. \tag{4.21}$$

As the algebra involved in evaluating the corresponding expressions for pressure, density, and metric turns out to be highly complex, we are leaving it for out later studies.

V. REDUCTION TO FIRST-ORDER ORDINARY DIFFERENTIAL EQUATION

As pointed out earlier, the main difficulty in solving Eq. (3.8) for all values of the parameters involved in it, is due to its intrinsic nonlinearity. To have an insight to it and for the sake of completeness, we have, here, using the Bluman–Cole technique,[8] made an attempt to determine a transformation group that leaves Eq. (3.8) invariant. This eventually helps to reduce it to a first-order differential equation.

To this effect we proceed as follows:

Consider the one-parameter transformation group,

$$\xi_1 = \xi + \epsilon\phi(\xi, f) + O(\epsilon^2), \quad f_1 = f + \epsilon\psi(\xi, f\ ) + O(\epsilon^2). \tag{5.1}$$

Invoking the invariance criteria as given in Ref. 8, Eq. (3.8), under the transformation (5.1), yields a highly overdetermined system of equations for the unknowns ϕ and ψ. Since the computations involved are too cumbersome but straightforward we, therefore, give the final results as under:

$$\phi(\xi, f) = 2k\xi, \quad \psi(\xi, f) = kf, \tag{5.2}$$

(one can also arrive at the same result via scale analysis) where k is an arbitrary constant.

In order to determine the invariants $x = x(\xi, f\ )$ and $y = y(\xi, f, f')$, we have to solve the following set of equations:

$$\frac{d\xi}{2k\xi} = \frac{df}{kf} = \frac{df'}{-kf'}, \tag{5.3}$$

which readily yields

$$x = \xi^{-1/2} f, \quad y = ff'.$$

In terms of the new independent variable x and the new dependent variable y, Eq. (3.8) assumes the following form:

$$x[(b_1^2-\alpha^2)x^4+2(\alpha^2+6\alpha b_1-3b_1^2)x^2y-(\alpha+b_1)^2x^4y$$
$$+2(\alpha+b_1)^2x^2y^2+8b_1(b_1-3\alpha)y^2]\frac{dy}{dx}=4y^2[2b_1(b_1-3\alpha)y(y+1)$$
$$+(\alpha+b_1)(4\alpha-b_1)x^2y-(4\alpha^2-3\alpha b_1+b_1^2)x^2]. \tag{5.4}$$

This is Abel's equation of the second kind, which is known, in general, to be not readily amenable to any of the standard devices of deriving the exact solutions available in the literature. The form of Eq. (5.4) suggests that it brings forth a tremendous scope. It generates an almost exhaustive class of solutions whenever a way of handling Abel's equation of the second kind, in general, is discovered.

For the sake of ready reference, a few special cases of Eq. (5.4) are listed below:

(i) $b_1=0, \quad \gamma\to 0,$

$$[x^3+(x^2-2)xy-2xy^2]\frac{dy}{dx}=16y^2(1-y), \tag{5.5}$$

(ii) $b_1=-\alpha, \quad \delta\to 0,$

$$x(x^2-2y)\frac{dy}{dx}=2y(x^2-y-y^2). \tag{5.6}$$

VI. DISCUSSIONS AND CONCLUSIONS

The study carried out above has established the superiority of the generalized symmetry approach utilized here over the Bluman and Cole technique used in Ref. 7 for the same problem. In short, the generalized symmetries arrived at in Eq. (3.4) have not merely provided, for some particular values of the parameters involved in it, new metrics [see (4.4) and (4.15)], but have also enlarged the scope for new metrics and Petrov type D solutions, the importance of which is well known.

Metric (4.4): As the reduction of it into normal form is rather difficult, the direct comparison with the most general metric due to Barnes[9] for geodesic flow is not possible.

Nevertheless, it can be shown that it possesses nonvanishing flow vector $(\nu^1,0,0,\nu^4)$:

$$\nu^1=-S\left[\frac{2\xi(\xi+2c_0)}{\{\xi(2b_1r-\beta t^3+b_1t+2\gamma)-3\beta c_0t^3\}(R+tPQ)}\right]^{1/2}$$

and

$$\nu^4=\frac{b_1}{2}(tP+T)\left[\frac{2\xi(\xi+2c_0)}{\{\xi(2b_1r-\beta t^3+b_1t+2\gamma)-3\beta c_0t^3\}}\right]^{1/2},$$

where

$$S=-b_1(2b_1r-\beta t^3+2\gamma)\xi^3-b_1c_0(4b_1r-5\beta t^3)\xi^2+c_0t\{(b_1r+\gamma)^2-t(b_1r+\gamma)(\beta t^2-b_1)$$
$$-2\beta t^4(\beta t^2+b_1)+6\beta b_1c_0t^2\}\xi-3\beta c_0^2t^4(b_1r+\beta t^3+b_1t+\gamma),$$

$$R=4b_1^3t^2\xi^6(\xi+6c_0)-4b_1^2c_0t^2(3b_1tr+3\beta t^4+2b_1t^2+3\gamma t-12b_1c_0)\xi^5$$
$$-16b_1^2c_0^2t^2(3b_1tr+3\beta t^4+2b_1t^2+3\gamma t-2b_1c_0)\xi^4+b_1c_0^2t^3\{b_1t(2b_1r-\beta t^3+b_1t+2\gamma)$$
$$-16b_1c_0(3b_1r+3\beta t^3+2b_1t+3\gamma)+8t(b_1r+\beta t^3+\gamma)^2+12b_1t^2(b_1r+\beta t^3+\gamma)$$
$$+4b_1^2t^3\}\xi^3+b_1c_0^3t^4\{b_1(4b_1r-5\beta t^3+2b_1t+4\gamma)+16(b_1r+\beta t^3+\gamma)^2$$
$$+24b_1t(b_1r+\beta t^3+\gamma)+8b_1^2t^2\}\xi^2-c_0^3t^4\{6\beta c_0b_1^2t^3$$
$$+(2b_1r-\beta t^3+b_1t+2\gamma)(b_1r+\beta t^3+b_1t+\gamma)^2\}\xi+3\beta c_0^4t^7(b_1r+\beta t^3+b_1t+\gamma)^2,$$

$$P^2=4b_1^2\xi^6+8b_1^2c_0\xi^5-4b_1c_0(2b_1tr+2\beta t^4+b_1t^2+2\gamma t-4b_1c_0)\xi^4$$
$$-8b_1c_0^2t(2b_1r+2\beta t^3+b_1t+2\gamma)\xi^3+c_0^2t^2(2b_1r-\beta t^3+b_1t+2\gamma)^2\xi^2$$
$$-6\beta c_0^3t^5(2b_1r-\beta t^3+b_1t+2\gamma)\xi+9\beta^2c_0^4t^8,$$

$$Q=2b_1^2\xi^4+8b_1^2c_0\xi^3-b_1c_0(4b_1tr+4\beta t^4+3b_1t^2+4\gamma t-8b_1c_0)\xi^2$$
$$-2b_1c_0^2t(4b_1r+4\beta t^3+3b_1t+4\gamma)\xi+c_0^2t^2(b_1r+\beta t^3+b_1t+\gamma)^2,$$

$$T=-2b_1t\xi^3-4b_1c_0t\xi^2+c_0t^3(3\beta t^2+b_1)\xi+3\beta c_0^2t^5,$$

and

$$\xi=(t/4b_1)(4b_1r+\beta t^3+2b_1t+4\gamma).$$

An attempt was made to calculate the corresponding acceleration vector $(\dot{\nu}^1,0,0,\dot{\nu}^4)$ too, but the algebra involved blew out of proportion, and hence we are omitting it. However, it is easily realizable that it would turn out to be nonvanishing. It can be concluded that the fluid distribution is with nongeodesic flow.

Metric (4.15): In this case the associated velocity and acceleration vectors are obtained as follows:

$$\nu^1=\frac{[\pm\alpha t^3q(1\mp q)^2-4a^2(3\alpha r+\alpha t+\gamma)]}{2a\alpha t[(4a^2/t)(6r+3t+2\gamma/\alpha)-t^2(1\mp q)^2(1\pm 2q)]^{1/2}},$$

$$\nu^2=\nu^3=0,$$

$$\nu^4=2a[(4a^2/t)(6r+3t+2\gamma/\alpha)-t^2(1\mp q)^2(1\pm 2q)]^{-1/2},$$

and the acceleration vectors $(\dot{\nu}^1,0,0,\dot{\nu}^4)$ are

$$\dot{\nu}^1=(1/2atq)[4a^2t(6r+3t+2\gamma/\alpha)-t^4(1\mp q)^2(1\pm 2q)]^{-3/2}[\pm t^6(1\pm 2q)q^2(1\mp q)^4$$
$$\pm 4a^2t^3(1\mp q)^2\{(1\pm q-4q^2)(3r+\gamma/\alpha)+t(1-7q^2)\}$$
$$\mp 16a^4(6r+3t+2\gamma/\alpha)\{(1\mp q)(3r+\gamma/\alpha)+t(1\mp 2q)\}],$$

and

$$\dot{\nu}^4=\frac{2at[t^3(1\pm 2q)(1\mp q)^2-4a^2(6r+3t+2\gamma/\alpha)]}{[4a^2t(6r+3t+2\gamma/\alpha)-t^4(1\mp q)^2(1\pm 2q)]^{3/2}}.$$

It is clear that velocity and acceleration vectors do not vanish and that the metric and hence the related fluid distribution is with nongeodesic flow.

ACKNOWLEDGMENTS

The second author (K.S.) is thankful to Dr. T. Gnana Bhaskar for his constant encouragement and interest in the preparation of this paper, and to the Council of Scientific and Industrial Research, India, for providing the financial support. Authors are also thankful to the referee for the constructive suggestions made on the earlier version of this paper.

APPENDIX A: FRECHET DERIVATIVE

On replacing ν by $S(u)$ in Eq. (3.5), we get

$$F(L,u,S(u))=[S(u)]_{tt}\left(\frac{u_r^3}{t}-u_{rr}\right)+\frac{2}{t}\,u_r\left(u_tu_{rr}-u_ru_{rt}-\frac{1}{t}\,u_r^2\right)[S(u)]_t$$
$$+\frac{1}{t}\left[3u_r^2u_{rt}+2u_{rt}-u_{rr}(1-u_t^2)-4u_tu_ru_{rt}+\frac{2}{t}\,u_r(1-3u_ru_t+2u_r^2)\right][S(u)]_r$$
$$+2\left(u_{rt}+\frac{1}{t}\,u_r-\frac{1}{t}\,u_r^2u_t\right)[S(u)]_{rt}-\left[u_{tt}+\frac{1}{t}\,u_r(1-u_t^2)\right][S(u)]_{rr}. \tag{A1}$$

Substituting the values of different derivatives of $S(u)$ in Eq. (A1) and collecting the coefficients of various powers of u_r, u_t, u_tu_r,..., etc., we get

$$F(L,u,S(u))=u_{rt}^2\Bigg[C_{rr}+A_{rr}u_t+\left(B_{uu}-\frac{1}{t}\,B_r-\frac{2}{t}\,C_u-\frac{A}{t^2}\right)u_r^3+\left(A_{uu}-\frac{1}{t}\,A_r\right)u_r^2u_t$$
$$+\left\{C_{uu}+2B_{ru}-\frac{1}{t}\,(3C_r+2A_u)\right\}u_r^2+\left(B_{rr}+2C_{ru}-\frac{2}{t}\,A_r\right)u_r+2A_{ru}u_ru_t\Bigg]$$
$$+2u_{rt}u_{rr}\Bigg[-\left(C_{rt}+\frac{1}{t}\,C_r\right)-\left(A_{rt}+C_{ru}+\frac{A_r}{t}\right)u_t-A_{ru}u_t^2+\left(\frac{1}{t}\,A_r-A_{uu}\right)u_t^2u_r$$
$$+\left\{\frac{1}{t}\,(A_u+B_u+C_t)-B_{tu}\right\}u_r^2+\left\{\frac{A}{t^2}+\frac{1}{t}\,(B_r+2C_u)-B_{uu}\right\}u_r^2u_t$$
$$+\left\{\frac{1}{t}\,(A_r+A_t)-\frac{A}{t^2}-B_{rt}-C_{tu}\right\}u_r+\left(\frac{1}{t}\,A_u-\frac{2}{t}\,C_r-A_{tu}-C_{uu}-B_{ru}\right)u_ru_t\Bigg]$$
$$+\frac{2}{t}\,u_{rt}\Bigg[C_{rr}u_r+A_{rr}u_ru_t+2\left\{B_{tu}-\frac{1}{t}\,(A_u+B_u+C_t)\right\}u_r^5-A_{ru}u_r^3u_t^2$$
$$+\left(B_{uu}+B_{rt}-\frac{1}{t}\,B_r+C_{tu}-\frac{2}{t}\,C_u-\frac{1}{t}\,A_t-\frac{1}{t}\,A_r\right)u_r^4$$
$$+\left(A_{tu}+\frac{1}{t}\,A_u+\frac{1}{t}\,C_r-B_{ru}\right)u_r^4u_t+\left(C_{rt}-\frac{2}{t}\,C_r-\frac{2}{t}\,A_u+C_{uu}+2B_{ru}\right)u_r^3$$
$$+\left(\frac{2}{t}\,A_r+A_{uu}-C_{ru}-B_{rr}+A_{rt}\right)u_r^3u_t+\left(B_{rr}+2C_{ru}-\frac{2}{t}\,A_r\right)u_r^2$$
$$+(2A_{ru}-C_{rr})u_r^2u_t-A_{rr}u_r^2u_t^2\Bigg]+u_{rr}^2\Bigg[C_{tt}+\frac{1}{t}\,C_r+\left(A_{tt}+2C_{tu}+\frac{1}{t}\,A_r\right)u_t$$
$$+\left(2A_{tu}+C_{uu}-\frac{1}{t}\,C_r\right)u_t^2+\left(A_{uu}-\frac{1}{t}\,A_r\right)u_t^3+\left(\frac{A}{t^2}-\frac{2}{t}\,A_t+B_{tt}+\frac{1}{t}\,B_r-\frac{2}{t}\,B_t\right)u_r$$
$$+2\left(B_{tu}-\frac{1}{t}\,B_u-\frac{1}{t}\,A_u-\frac{1}{t}\,C_t\right)u_ru_t+\left(B_{uu}-\frac{1}{t}\,B_r-\frac{2}{t}\,C_u-\frac{A}{t^2}\right)u_ru_t^2\Bigg]$$
$$+u_{rr}\Bigg[-\frac{2}{t}\left(C_{rt}+\frac{1}{t}\,C_r\right)u_r-\frac{2}{t}\left(A_{rt}+\frac{1}{t}\,A_r+C_{ru}\right)u_ru_t-\frac{2}{t}\,A_{ru}u_ru_t^2$$
$$-\frac{2}{t}\left\{C_{tu}+B_{rt}-\frac{1}{t}\,(A_r+A_t)+\frac{A}{t^2}\right\}u_r^2+\frac{2}{t}\left\{C_{rt}-C_{uu}-B_{ru}-A_{tu}\right.$$
$$\left.+\frac{1}{t}\,(A_u-3C_r)\right\}u_r^2u_t+\frac{2}{t}\left(C_{ru}+A_{rt}-A_{uu}+\frac{2}{t}\,A_r\right)u_r^2u_t^2+\frac{2}{t}\,A_{ru}u_r^2u_t^3$$

$$
\begin{aligned}
&-\frac{1}{t}\left(A_{uu}-\frac{1}{t}A_r\right)u_r^3u_t^3-\frac{2}{t}\left\{C_{tt}+B_{tu}-\frac{1}{t}(B_u-C_r+C_t+A_u)\right\}u_r^3\\
&+\frac{2}{t}\left\{B_{ru}-A_{tu}-\frac{1}{t}(A_u+C_r)\right\}u_r^3u_t^2-\frac{2}{t}\left\{A_{tt}+B_{uu}-B_{rt}+C_{tu}\right.\\
&\left.+\frac{1}{t}(A_t+2A_r-2C_u-B_r)-\frac{2}{t^2}A\right\}u_r^3u_t\Big]+\frac{1}{t^2}\left\{B_{tt}+\frac{1}{t}(B_r-2B_t-2A_t)+\frac{A}{t^2}\right\}u_r^7\\
&+\frac{1}{t^2}\left\{A_{tt}-2B_{rt}+\frac{1}{t}(3A_r+2A_t)-\frac{2A}{t^2}\right\}u_r^6u_t+\frac{1}{t^2}\left\{C_{tt}+2B_{tu}\right.\\
&\left.+\frac{1}{t}(C_r-2C_t-2B_u-2A_u)\right\}u_r^6+\frac{1}{t^2}\left(B_{rr}-2A_{rt}-\frac{4}{t}A_r\right)u_r^5u_t^2\\
&+\frac{1}{t^2}\left\{2C_{tu}+2B_{rt}+B_{uu}-\frac{1}{t}(B_r+2C_u+2A_t+2A_r)+\frac{A}{t^2}\right\}u_r^5\\
&+\frac{1}{t^2}\left\{2A_{tu}-2C_{rt}-3B_{ru}+\frac{2}{t}A_u\right\}u_r^5u_t\\
&+\frac{1}{t^2}\left(C_{uu}+2C_{rt}+2B_{ru}-\frac{1}{t}(2A_u+C_r)\right)u_r^4\\
&+\frac{1}{t^2}\left(2A_{rt}-2C_{ru}+A_{uu}-2B_{rr}+\frac{5}{t}A_r\right)u_r^4u_t+\frac{1}{t^2}(C_{rr}+2A_{ru})u_r^4u_t^2\\
&+\frac{1}{t^2}A_{rr}u_r^4u_t^3+\frac{1}{t^2}\left(B_{rr}+2C_{ru}-\frac{2}{t}A_r\right)u_r^3\\
&+\frac{2}{t^2}(A_{ru}-C_{rr})u_r^3u_t-\frac{2}{t^2}A_{rr}u_r^3u_t^2+\frac{1}{t^2}C_{rr}u_r^2+\frac{1}{t^2}A_{rr}u_r^2u_t\,.
\end{aligned}
$$

[1] O. P. Bhutani and D. K. Kalra, ''Asymptotically flat rotating solutions of Einstein's exterior equations: A Unified Approach,'' preprint.
[2] O. P. Bhutani and P. Mital, J. Meteorol. Soc. Jpn. **64**, 593 (1986).
[3] O. P. Bhutani, K. Vijay Kumar, P. Mital, and G. Chandrasekaran, Int. J. Eng. Sci. **27**, 921 (1989).
[4] O. P. Bhutani and P. Mital, Int. J. Eng. Sci. **23**, 353 (1985).
[5] H. Weyl, Ann. Phys. **54**, 117 (1917).
[6] K. Schwarzschild, *Sitzungsber* (Preuss-Akad, Wiss, Berlin, 1916), p. 424.
[7] Y. K. Gupta and J. R. Sharma, J. Math. Phys. **37**, 1962 (1996).
[8] G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer-Verlag, New York, 1974).
[9] A. Barnes, Gen. Relativ. Gravit. **5**, 147 (1974).

On separability of bi-Hamiltonian chain with degenerated Poisson structures

Maciej Błaszak[a)]
Department of Physics, A. Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland

(Received 13 June 1997; accepted for publication 15 October 1997)

Separability of bi-Hamiltonian finite-dimensional chains with two degenerated Poisson tensors, which have Pfaffian quasi-bi-Hamiltonian representation, is proved. © *1998 American Institute of Physics.* [S0022-2488(98)00106-6]

I. INTRODUCTION

In recent years considerable progress has been made in the construction of new integrable finite-dimensional dynamical systems showing bi-Hamiltonian property. The majority of them come from stationary flows, restricted flows, or nonlinearization of Lax equations of underlying soliton systems.[1–12] Although such constructed finite-dimensional systems are integrable, in principle, as follows from the existence of an appropriate number of integrals of motion, we still lack a general method of finding solutions given in the original variables. In this paper we announce significant progress in this direction. It is proved that most (perhaps all) bi-Hamiltonian finite-dimensional chains, which start with a Casimir of the first Poisson structure and terminate with a Casimir of the second Poisson structure, consist of nonlinear dynamical systems that are separable and, hence, through the solutions of the Hamilton–Jacobi equation, integrable by quadratures. Actually, first we consider in some phase space $M(\dim M=2n)$ so-called quasi-bi-Hamiltonian systems[13] that are Pfaffian, i.e., which in the Nijenhuis coordinates (λ,μ) (coordinates that diagonalize the appropriate Nijenhuis tensor) take the following form:

$$\begin{pmatrix}\lambda\\ \mu\end{pmatrix}_t=\begin{pmatrix}0 & I\\ -I & 0\end{pmatrix}\nabla H(\lambda,\mu;c)=\frac{(-1)^{n+1}}{\lambda_1\cdot\ \cdots\ \cdot\lambda_n}\begin{pmatrix}0 & \Lambda\\ -\Lambda & 0\end{pmatrix}\nabla F(\lambda,\mu;c), \tag{1}$$

where $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_n)$ and c is a constant parameter, or are inverse Pfaffian, i.e., in the Nijenhuis coordinates, take the form

$$\begin{pmatrix}\lambda\\ \mu\end{pmatrix}_t=\begin{pmatrix}0 & I\\ -I & 0\end{pmatrix}\nabla H(\lambda,\mu;c)=(-1)^{n+1}\lambda_1\cdot\ \cdots\ \cdot\lambda_n\begin{pmatrix}0 & \Lambda^{-1}\\ -\Lambda^{-1} & 0\end{pmatrix}\nabla F(\lambda,\mu;c), \tag{2}$$

where $\Lambda^{-1}=\mathrm{diag}(1/\lambda_1,\ldots,1/\lambda_n)$. Next, in both cases we prove the existence of n independent constants of motion $h_i(\lambda,\mu;c)$ $i=1,...,n$ all separable. The separability of $h_1(\lambda,\mu;c=0)\equiv H(\lambda,\mu)$ and $h_n(\lambda,\mu;c=0)\equiv F(\lambda,\mu)$ of the case (1) was recently proved in Ref. 14. Finally, we prove that most bi-Hamiltonian finite-dimensional systems with two degenerated Poisson tensors, the chain of which, in extended phase space $\bar{M}(\dim \bar{M}=2n+1)$, starts with the Casimir of the canonical (noncanonical) Poisson structure and terminates with the Casimir of the noncanonical (canonical) Poisson tensor, in the original phase space M and in Nijenhuis coordinates are equivalent to Pfaffian (inverse-Pfaffian) quasi-bi-Hamiltonian systems. Hence, they are separable! Moreover, a canonical transformation to the Nijenhuis coordinates is constructed in a very natural way. The procedure is applicable to all known bi-Hamiltonian systems being stationary flows, restricted flows, or nonlinearized spectral problems of underlying soliton systems. Few nontrivial examples illustrate the results.

a)Electronic mail: blaszakm@main.amu.edu.pl

 © 1998 American Institute of Physics

II. PRELIMINARIES

Let us reexamine some facts about bi-Hamiltonian systems, both finite and infinite dimensional. We recall some definitions. Let M be a differentiable manifold, TM and T^*M its tangent and cotangent bundle. At any point $u \in M$, the tangent and cotangent spaces are denoted by T_uM and T_u^*M, respectively. The pairing between them is given by the map $\langle \cdot, \cdot \rangle : T_u^*M \times T_uM \to R$. For each smooth function $F \in C^\infty(M)$, dF denotes the differential of F (gradient ∇F for finite systems and variation δF for field systems). M is said to be a Poisson manifold if it is endowed with a Poisson bracket $\{\cdot, \cdot\} : C^\infty(M) \times C^\infty(M) \to C^\infty(M)$, degenerated in general. The related Poisson tensor θ is defined by $\{F,G\}_\theta(u) := \langle dG, \theta \circ dF \rangle(u) = \langle dG(u), \theta(u) dF(u) \rangle$. So, at each point u, $\theta(u)$ is a linear map $\theta(u) : T_u^*M \to T_uM$, which is skew symmetric and fulfills the Jacobi identity. Any function $c \in C^\infty(M)$, such that $dc \in \ker\theta$, is called a Casimir of θ. Let $\theta_0, \theta_1 : T^*M \to TM$ be two Poisson tensors on M. A vector field K is said to be a bi-Hamiltonian with respect to θ_0 and θ_1 if there exist two smooth functions $H, F \in C^\infty(M)$, such that

$$K = \theta_0 \circ dH = \theta_1 \circ dF. \tag{3}$$

Poisson tensors θ_0 and θ_1 are said to be compatible if the associated pencil $\theta^\epsilon = \theta_1 - \epsilon\theta_0$ is itself a Poisson tensor for any $\epsilon \in R$. Moreover, if θ_0 is invertible, the tensor $\Phi = \theta_1 \circ \theta_0^{-1}$, called a recursion operator, is a Nijenhuis (hereditary) tensor of such a property, that when it acts on a given bi-Hamiltonian vector field K, produces another bi-Hamiltonian vector field being a symmetry generator of K. Hence, having the Nijenhuis tensor, one can construct a hierarchy of Hamiltonian symmetries and related hierarchy of constants of the motion for underlying system, such as is important for its integrability. Finally, arbitrary tensor field $T^{(k,l)}$ is called a tensor invariant[15] of dynamical system $u_t = K(u)$ if

$$\left(\frac{\partial}{\partial t} + L_K \right) T^{(k,l)} = 0, \tag{4}$$

where L_K denotes the Lie derivative in the direction of vector field K. Examples of various tensor invariants are both Poisson tensors θ_1 and θ_2, the recursion operator Φ, symmetry generators $K_n = \Phi^n \circ K$, or conserved functions H and F.

Unfortunately, for a majority of bi-Hamiltonian finite-dimensional systems, both Poisson structures are degenerated, so one cannot construct the recursion Nijenhuis tensor inverting one of the Poisson structures. Nevertheless, due to the nonuniqueness of Hamiltonian functions, determined up to an appropriate Casimir function, it is always possible to construct a finite bi-Hamiltonian chain starting and terminating with Casimirs of θ_0 and θ_1, respectively. Actually, in the extended phase $\overline{M} \ni (q,p,c)$, where $q = (q_1, \ldots, q_n)^T$ and $p = (p_1, \ldots, p_n)^T$ are generalized coordinates and c is a Casimir coordinate; a typical bi-Hamiltonian chain reads as

$$\begin{aligned}
&\theta_0 \circ \nabla h_0 = 0, \\
&\theta_0 \circ \nabla h_1 = K_1 = \theta_1 \circ \nabla h_0, \\
&\theta_0 \circ \nabla h_2 = K_2 = \theta_1 \circ \nabla h_1, \\
&\qquad \vdots \\
&\theta_0 \circ \nabla h_n = K_n = \theta_1 \circ \nabla h_{n-1}, \\
&\qquad 0 = \theta_1 \circ \nabla h_n,
\end{aligned} \tag{5}$$

where $K \equiv K_1$, $H \equiv h_1$, $F \equiv h_0$, and $d \equiv \nabla$. A canonical θ_0 and noncanonical θ_1 Poisson structures (both degenerated) take the general form

$$\theta_0 = \begin{pmatrix} \theta_c & 0 \\ 0 & 0 \end{pmatrix}, \quad \theta_1 = \begin{pmatrix} \theta_{\mathrm{nc}} & K \\ -K^T & 0 \end{pmatrix}, \tag{6}$$

where

$$\theta_c=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \theta_{\mathrm{nc}}=\begin{pmatrix} D & A \\ -A^T & B \end{pmatrix} \tag{7}$$

are a nondegenerated matrix from the space $M\ni(q,p)$, $K=K(q,p,c)\equiv K_1$ is an appropriate vector field from the hierarchy (5), and A,B,D are entries depending on generalized coordinates (q,p).

In our further considerations we shall concentrate on the following form of θ_{nc}:

$$\theta_{\mathrm{nc}}=\begin{pmatrix} 0 & A(q) \\ -A^T(q) & B(q,p) \end{pmatrix}. \tag{8}$$

It should be stressed that for all known examples of constrained flows of Lax equations and stationary flows in a Newton representation,[4,5,8,10,11,16,12,17] the matrix θ_{nc} takes the form (8), and for stationary flows in the Ostrogradsky representation[1,3,9] θ_{nc} can be put into the form (8) by an appropriate canonical transformation of variables. Moreover, for matrix θ_{nc} in the form (8), the Hamiltonians $h_r(q,p,c)$ take the general form

$$h_r(q,p,c)=h_r(q,p)+c\rho_\tau(q). \tag{9}$$

III. QUASI-BI-HAMILTONIAN SYSTEMS

For finite-dimensional systems the bi-Hamiltonian property is a very strong condition and, in fact, too restrictive for integrability. There are some physically interesting dynamical systems that in their natural phase space $M\ni(q,p)$ satisfy a weaker condition than the bi-Hamiltonian one. So, before we continue our considerations of bi-Hamiltonian chains we shall introduce the concept of quasi-bi-Hamiltonian systems.[13]

Definition 1: A vector field K is called a *quasi-bi-Hamiltonian* with respect to Poisson tensors θ_0 and θ_1 if there exist three smooth functions $\rho, H,F,$ such that

$$K=\theta_0\circ\nabla H=\frac{1}{\rho}\cdot\theta_1\circ\nabla F; \tag{10}$$

both Poisson matrix are compatible and nondegenerated (invertible). The function ρ is called the *integrating factor*.

To make this concept more familiar and relate it to the bi-Hamiltonian formalism let us review the well-known example of the integrable case of the Henon–Heiles system,[10,13]

$$q_{1tt}=-3q_1^2-\tfrac{1}{2}q_2^2, \qquad q_{2tt}=-q_1q_2, \tag{11}$$

which has two constants of motion of the form

$$\begin{aligned} H&=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+q_1^3+\tfrac{1}{2}q_1q_2^2, \\ F&=-2q_2p_1p_2+2q_1p_2^2-\tfrac{1}{4}q_2^4-q_1^2q_2^2, \end{aligned} \tag{12}$$

where $p_1=q_{1t}$ and $p_2=q_{2t}$. In the phase space $M\ni(q_1,q_2,p_1,p_2)$, the related vector fields admit the following quasi-bi-Hamiltonian formulations:

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix}_{t_{\mathrm{H}}} = \begin{pmatrix} p_1 \\ p_2 \\ -3q_1^2-\tfrac{1}{2}q_2^2 \\ -q_1q_2 \end{pmatrix} \equiv K_H=\theta_0\circ\nabla H=-\frac{1}{q_2^2}\cdot\theta_1\circ\nabla F, \tag{13}$$

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix}_{t_F} = \begin{pmatrix} -2q_2p_2 \\ -2q_2p_1+4q_1p_2 \\ 2q_1q_2^2-2p_2^2 \\ 2p_1p_2+q_2^3+2q_1^2q_2 \end{pmatrix} \equiv K_F = \theta_0 \circ \nabla F = -q_2^2 \cdot \theta_{-1} \circ \nabla H, \tag{14}$$

where

$$\theta_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad \theta_1 = \begin{pmatrix} 0 & 0 & q_1 & \frac{1}{2}q_2 \\ 0 & 0 & \frac{1}{2}q_2 & 0 \\ -q_1 & -\frac{1}{2}q_2 & 0 & \frac{1}{2}p_2 \\ -\frac{1}{2}q_2 & 0 & -\frac{1}{2}p_2 & 0 \end{pmatrix},$$

$$\theta_{-1} = \begin{pmatrix} 0 & 0 & 0 & 1/q_2 \\ 0 & 0 & 1/q_2 & -2q_1/q_2^2 \\ 0 & -1/q_2 & 0 & p_2/q_2^2 \\ -1/q_2 & 2q_1/q_2^2 & -p_2/q_2^2 & 0 \end{pmatrix}, \tag{15}$$

are appropriate compatible Poisson tensors. The Nijenhuis operator,

$$\Phi = \theta_1 \circ \theta_0^{-1}, \tag{16}$$

relates to both systems, i.e., applying Φ to Eq. (14) we get Eq. (13). Note that H, F, and θ_0 are tensor invariants of $u_t = K(u)$, but θ_1 and ρ are not. Hence, Φ is not a recursion operator.

Now, let us consider the first generalization of the Henon–Heiles system,

$$q_{1tt} = -3q_1^2 - \tfrac{1}{2}q_2^2 + c, \qquad q_{2tt} = -q_1q_2, \tag{17}$$

with the following constants of motion:

$$\overline{H} = \tfrac{1}{2}p_1^2 + \tfrac{1}{2}p_2^2 + q_1^3 + \tfrac{1}{2}q_1q_2^2 - cq_1,$$

$$-\tfrac{1}{4}\overline{F} = \tfrac{1}{2}q_2p_1p_2 - \tfrac{1}{2}p_1p_2^2 + \tfrac{1}{16}q_2^4 + \tfrac{1}{4}q_1^2q_2^2 - \tfrac{1}{4}q_2^2c, \tag{18}$$

$$\overline{G} = c.$$

In the extended phase space $\overline{M} \ni (q_1, q_2, p_1, p_2, c)$, the related dynamical systems admit the following bi-Hamiltonian formulations:

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \\ c \end{pmatrix}_{t_H} = \begin{pmatrix} p_1 \\ p_2 \\ -3q_1^2 - \frac{1}{2}q_2^2 + c \\ -q_1q_2 \\ 0 \end{pmatrix} \equiv \overline{K_{\overline{H}}} = \overline{\theta_0} \circ \overline{\nabla H} = \overline{\theta_1} \circ \overline{\nabla G}, \tag{19}$$

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \\ c \end{pmatrix}_{t_F} = \begin{pmatrix} \frac{1}{2} q_2 p_2 \\ \frac{1}{2} q_2 p_1 - q_1 p_2 \\ -\frac{1}{2} q_1 q_2^2 + \frac{1}{2} p_2^2 \\ -\frac{1}{2} p_1 p_2 - \frac{1}{4} q_2^3 - \frac{1}{2} q_1^2 q_2 + \frac{1}{2} q_2 c \\ 0 \end{pmatrix} \equiv \overline{K_F} = \overline{\theta_0} \circ \overline{\nabla} \left(-\frac{1}{4} \overline{F} \right) = \overline{\theta_1} \circ \overline{\nabla} \overline{H}, \tag{20}$$

where compatible Poisson tensors are of the following form:

$$\overline{\theta_0} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\overline{\theta_1} = \begin{pmatrix} 0 & 0 & q_1 & \frac{1}{2} q_2 & p_1 \\ 0 & 0 & \frac{1}{2} q_2 & 0 & p_2 \\ -q_1 & -\frac{1}{2} q_2 & 0 & \frac{1}{2} p_2 & -3q_1^2 - \frac{1}{2} q_2^2 + c \\ -\frac{1}{2} q_2 & 0 & -\frac{1}{2} p_2 & 0 & -q_1 q_2 \\ -p_1 & -p_2 & 3q_1^2 + \frac{1}{2} q_2^2 - c & q_1 q_2 & 0 \end{pmatrix}, \tag{21}$$

each degenerated. Note that $\overline{G}$ is a Casimir of $\overline{\theta_0}$ and $\overline{F}$ is a Casimir of $\overline{\theta_1}$.

Finally, let us consider the second generalized Henon–Heiles system in the form

$$q_{1tt} = -3q_1^2 - \tfrac{1}{2} q_2^2, \qquad q_{2tt} = -q_1 q_2 + \frac{d}{q_2^3}, \tag{22}$$

with the following constants of motion:

$$-2\overline{H} = -p_1^2 - p_2^2 - 2q_1^3 - q_1 q_2^2 - \frac{d}{q_2^2},$$

$$\overline{F} = -2q_2 p_1 p_2 + 2q_1 p_2^2 - \frac{1}{4} q_2^4 - q_1^2 q_2^2 + \frac{2dq_1}{q_2^2}, \tag{23}$$

$$\overline{G} = d.$$

In the extended phase space $\overline{M} \ni (q_1, q_2, p_1, p_2, d)$, the related dynamical systems admit the following bi-Hamiltonian formulations:

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \\ c \end{pmatrix}_{t_H} = \begin{pmatrix} -2p_1 \\ -2p_2 \\ 6q_1^2+q_2^2 \\ 2q_1q_2-2d/q_2^3 \\ 0 \end{pmatrix} \equiv \overline{K_H} = \overline{\theta}_0 \circ \overline{\nabla}(-2\overline{H}) = \overline{\theta}_{-1} \circ \overline{\nabla F}, \tag{24}$$

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \\ c \end{pmatrix}_{t_F} = \begin{pmatrix} -2q_2p_2 \\ -2q_2p_1+4q_1p_2 \\ 2q_1q_2^2-2p_2^2-2d/q_2^2 \\ 2p_1p_2+q_2^3+2q_1^2q_2+4dq_1/q_2^3 \\ 0 \end{pmatrix} \equiv \overline{K_F} = \overline{\theta}_0 \circ \overline{\nabla F} = \overline{\theta}_{-1} \circ \overline{\nabla G}, \tag{25}$$

where compatible Poisson tensors are $\overline{\theta}_0$ and

$$\overline{\theta}_{-1} = \begin{pmatrix} 0 & 0 & 0 & 1/q_2 & -2q_2p_2 \\ 0 & 0 & 1/q_2 & -2q_1/q_2^2 & -2q_2p_1+4q_1p_2 \\ 0 & -1/q_2 & 0 & p_2/q_2^2 & 2q_1q_2^2-2p_2^2-2d/q_2^2 \\ -1/q_2 & 2q_1/q_2^2 & -p_2/q_2^2 & 0 & 2p_1p_2+q_2^3+2q_1^2q_2+4dq_1/q_2^3 \\ * & * & * & * & 0 \end{pmatrix}, \tag{26}$$

again both degenerated. Stars in the last row mean matrix elements that make (26) skew symmetric. Notice that $\overline{G}$ and $\overline{H}$ from (23) are respective Casimirs of $\overline{\theta}_0$ and $\overline{\theta}_{-1}$.

The generalized Henon–Heiles systems (19) and (25) additionally have a quasi-bi-Hamiltonian representation. Actually, for the system (19) we find

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix}_{t_H} = \begin{pmatrix} p_1 \\ p_2 \\ -3q_1^2-\frac{1}{2}q_2^2+c \\ -q_1q_2 \end{pmatrix} \equiv \overline{K_H} = \theta_0 \circ \overline{\nabla H} = -\frac{1}{q_2^2} \cdot \theta_1 \circ \overline{\nabla F}, \tag{27}$$

where Poisson tensors θ_0, θ_1 are given by (15) and conserved functions $\overline{H}, \overline{F}$ take the form (18), while for the system (25) its quasi-bi-Hamiltonian representation reads as

$$\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix}_{t_F} = \begin{pmatrix} -2q_2p_2 \\ -2q_2p_1+4q_1p_2 \\ 2q_1q_2^2-2p_2^2-2d/q_2^2 \\ 2p_1p_2+q_2^3+2q_1^2q_2+4dq_1/q_2^3 \end{pmatrix} \equiv \overline{K_F} = \theta_0 \circ \overline{\nabla F} = -q_2^2 \cdot \theta_{-1} \circ \overline{\nabla H}, \tag{28}$$

where conserved functions $\overline{H}, \overline{F}$ take the form (23). Hence, one concludes that quasi-bi-Hamiltonian representation survives the reduction $c=d=0$ (13), (14) while bi-Hamiltonian representation does not.

The Henon–Heiles example gave us some insight into the mutual relations between bi-Hamiltonian and quasi-bi-Hamiltonian representations. Now we shall pass to a more systematic treatment of the problem. In further considerations we shall only use the canonical coordinates (q,p) with respect to which θ_0 takes the canonical form $\theta_0=\theta_c$ (7). As θ_0 and θ_1 are compatible and invertible, the tensor $\Phi=\theta_1 \circ \theta_0^{-1}$ is the Nijenhuis (hereditary) tensor. As Φ is implectic–symplectic factorizable, all its eigenvalues have at least algebraic multiplicity 2.[18] Here, we assume the existence of an open set $U \subset M$ such that in U matrix Φ is maximal, i.e., has n distinct eigenvalues $\lambda=(\lambda_1,\ldots,\lambda_n)$. So, in a neighborhood of a regular point, where the eigenvalues λ are independent, one can construct a canonical transformation $(q,p) \to (\lambda,\mu)$ [(λ,μ) referred to as *Nijenhuis coordinates*], such that θ_1 and Φ take the Darboux form

$$\theta_1=\begin{pmatrix} 0 & \Lambda \\ -\Lambda & 0 \end{pmatrix}, \quad \Phi=\begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda \end{pmatrix}, \quad \Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_n). \tag{29}$$

Definition 2:

(i) A quasi-bi-Hamiltonian vector field (10) is said to be Pfaffian if, in Nijenhuis coordinates, it takes the form

$$\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\nabla H(\lambda,\mu)=\frac{(-1)^{n+1}}{\lambda_1\cdot\ \cdots\lambda_n}\begin{pmatrix} 0 & \Lambda \\ -\Lambda & 0 \end{pmatrix}\nabla F(\lambda,\mu), \tag{30}$$

i.e., the integrating factor ρ is the product of the eigenvalues of Φ,

$$\rho=(-1)^{n+1}\prod_{i=1}^{n}\lambda_i. \tag{31}$$

(ii) A quasi-bi-Hamiltonian vector field (10) is said to be inverse Pfaffian if in Nijenhuis coordinates if it takes the form

$$\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\nabla H(\lambda,\mu)=(-1)^{n+1}\lambda_1\cdot\ \cdots\ \cdot\lambda_n\begin{pmatrix} 0 & \Lambda^{-1} \\ -\Lambda^{-1} & 0 \end{pmatrix}\nabla F(\lambda,\mu), \tag{32}$$

where $\Lambda^{-1}=\mathrm{diag}(1/\lambda_1,\ldots,1/\lambda_n)$.

Notice that forms (31) and (32) are related by the Nijenhuis operator Φ (29) and its inverse Φ^{-1}.

Lemma 1: In the Pfaffian case (30) and in the inverse-Pfaffian case (32), the general solution of Eq. (10) is given by

$$H=\sum_{i=1}^{n}\frac{1}{\Delta_i}f_i(\lambda_i,\mu_i)+g_H(\lambda),$$

$$F=\sum_{i=1}^{n}\frac{\rho}{\lambda_i}\frac{1}{\Delta_i}f_i(\lambda_i,\mu_i)+g_F(\lambda), \tag{33}$$

where

$$\frac{\partial g_H}{\partial\lambda_i}=\frac{\lambda_i}{\rho}\frac{\partial g_F}{\partial\lambda_i}, \quad \Delta_i=\prod_{i\neq j}\lambda_{ij}, \quad \lambda_{ij}:=\lambda_i-\lambda_j, \tag{34}$$

and the n functions $f_i(\lambda_i,\mu_i)$ (each one depending on one pair of coordinates) are arbitrary smooth functions.

The proof for $g_H(\lambda)=g_F(\lambda)=0$ can be found in Ref. 14. Of course, the addition of an arbitrary function $g_H(\lambda)$ to $H=\Sigma_{i=1}^{n}(1/\Delta_i)f_i$, generates in $F=\Sigma_{i=1}^{n}(\lambda_i/\rho)(1/\Delta_i)f_i$ an additional function $g_F(\lambda)$ related to $g_H(\lambda)$ through relation (34).

Theorem 1: An arbitrary bi-Hamiltonian finite-dimensional system with two degenerated Poisson tensors of the form (6), where θ_{nc} takes the form (8), whose chain starts with a Casimir of the canonical Poisson structure and terminates with a Casimir of a noncanonical Poisson structure, is equivalent to a quasi-bi-Hamiltonian system, which in the Nijenhuis coordinates takes a Pfaffian form.

Proof: Let

$$\overline{\theta_0}\circ\overline{\nabla h_0}=0,$$
$$\overline{\theta_0}\circ\overline{\nabla h_1}=\overline{K_1}=\overline{\theta_1}\circ\overline{\nabla h_0},$$
$$\overline{\theta_0}\circ\overline{\nabla h_2}=\overline{K_2}=\overline{\theta_1}\circ\overline{\nabla h_1},$$

$$\vdots$$

$$\overline{\theta_0}\circ\overline{\nabla h_n}=\overline{K_n}=\overline{\theta_1}\circ\overline{\nabla h_{n-1}},$$
$$0=\overline{\theta_1}\circ\overline{\nabla h_n}, \tag{35}$$

be a bi-Hamiltonian chain on $U\subset\overline{M}\ni(q,p,c)$ with two degenerated Poisson structures, each with one Casimir, so that dim $\overline{M}=2n+1$ and rank $\overline{\theta_0}(\overline{\theta_1})=2n$, where

$$\overline{\theta_0}=\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad \overline{\theta_1}=\begin{pmatrix}0 & A(q) & \dfrac{\partial\overline{h_1}}{\partial p}\\ -A^T(q) & B(q,p) & -\dfrac{\partial\overline{h_1}}{\partial q}\\ -\left(\dfrac{\partial\overline{h_1}}{\partial p}\right)^T & -\left(\dfrac{\partial\overline{h_1}}{\partial q}\right)^T & 0\end{pmatrix} \tag{36}$$

are compatible, $\overline{h_r}(q,p,c)=\overline{h_r}(q,p)+c\rho_r(q)$, and $\overline{\nabla}=(\partial/\partial q,\partial/\partial p,\partial/\partial c)^T$. In order to construct a bi-Hamiltonian hierarchy also in the original phase space $M\ni(q,p)$, one can try to project the chain (35) onto M. If $\pi:M\to M:(q,p,c)\to(q,p)$ is the projection map, the Hamiltonians h_r and the vector fields K_r cannot be projected onto M, because they depend on the fiber coordinate c. Instead, the Poisson tensors θ_0 and θ_1 are projected onto

$$\theta_0\equiv\theta_c=\pi'\circ\overline{\theta_0}\circ\pi'^T=\begin{pmatrix}0 & I\\ -I & 0\end{pmatrix},$$

$$\theta_1\equiv\theta_{\mathrm{nc}}=\pi'\circ\overline{\theta_1}\circ\pi'^T=\begin{pmatrix}0 & A\\ -A^T & B\end{pmatrix}, \tag{37}$$

with $A=A(q)$ and $B=B(q,p)$, where π' is the derivative of the map π. Because these Poisson tensors are compatible and invertible, one obtains the following Nijenhuis tensor:

$$\Phi=\theta_1\circ\theta_0^{-1}=\begin{pmatrix}A & 0\\ B & A^T\end{pmatrix}. \tag{38}$$

Nevertheless, as the Poisson tensor θ_1 is not invariant along the flows of vector fields K_r, where $\overline{K_r}=(K_r,0)^T$, so Φ is not a recursion operator for any K_r. However, vector fields K_r are still Hamiltonian with respect to θ_0, i.e.

$$K_r=\theta_0\circ\nabla h_r, \tag{39}$$

where $h_r=\overline{h_r}$, c is treated as a parameter, and $\nabla=(\partial/\partial q,\partial/\partial p)^T$.

Consider the last equation from the hierarchy (35): $0=\theta_1\circ\nabla h_n$. On M it reads as

$$\theta_1\circ\nabla h_n=-\rho_n K_1,\quad \rho_n=\frac{\partial h_n}{\partial c}. \tag{40}$$

As $h_0=c$, hence on M, $\nabla h_0=0$, and the hierarchy (35) turns into the form

$$\theta_0\circ\nabla h_1=K_1=-\frac{1}{\rho_n}\cdot\theta_1\circ\nabla h_n,$$

$$\theta_0\circ\nabla h_2=K_2=-\frac{\rho_1}{\rho_n}\cdot\theta_1\circ\nabla h_n+\theta_1\circ\nabla h_1,$$

$$\vdots$$

$$\theta_0 \circ \nabla h_n = K_n = -\frac{\rho_{n-1}}{\rho_n} \cdot \theta_1 \circ \nabla h_n + \theta_1 \circ \nabla h_{n-1}, \tag{41}$$

where $\rho_r = \partial h_r / \partial c$. The last equation in (41) terminates the sequence of vector fields K_r in the hierarchy, as for the next equation from the chain we have

$$\theta_0 \circ \nabla h_{n+1} = K_{n+1} = -\theta_1 \circ \nabla h_n + \theta_1 \circ \nabla h_n = 0. \tag{42}$$

Because of the form of the first equation from the hierarchy (41), vector field $K \equiv K_1$ is quasi-bi-Hamiltonian. Now, from the relation

$$\Phi \circ K_r = \theta_1 \circ \nabla h_r = K_{r+1} + \frac{\rho_r}{\rho_n} \cdot \theta_1 \circ \nabla h_n = K_{r+1} - \rho_r K_1, \tag{43}$$

one finds that

$$K_{r+1} = \Phi_r K_1, \quad \Phi_r = \sum_{i=0}^{r} \rho_{r-i} \Phi^i, \quad \rho_0 = 1, \quad r = 0,...,n-1, \tag{44}$$

and for $r = n$,

$$\Phi_n \circ K_1 = K_{n+1} = 0; \tag{45}$$

hence

$$\Phi^n + \rho_1 \Phi^{n-1} + \cdots + \rho_{n-1} \Phi + \rho_n I = 0. \tag{46}$$

In Nijenhuis coordinates, operator Φ is diagonal, so Eq. (46) is equivalent to the following system:

$$\begin{pmatrix} \lambda_1^{n-1} & \cdots & \lambda_1 & 1 \\ \lambda_2^{n-1} & \cdots & \lambda_2 & 1 \\ \vdots & & \vdots & \vdots \\ \lambda_n^{n-1} & \cdots & \lambda_n & l \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{pmatrix} = \begin{pmatrix} -\lambda_1^n \\ -\lambda_2^n \\ \vdots \\ -\lambda_n^n \end{pmatrix}. \tag{47}$$

For all λ_i different, there is a unique solution to (47) in the form

$$\rho_r(q) = (-1)^r \sum_{\substack{j_1,..,j_r \\ j_1<..<j_r}} \lambda_{j_1} \cdot \cdots \cdot \lambda_{j_r} := \rho_r(\lambda), \quad r = 1,...,n. \tag{48}$$

So, the quasi-bi-Hamiltonian representation (41) for $K_1 \equiv K$ takes the Pfaffian form (30), where now $H = h_1$, $F = h_n$ and $\rho = -\rho_n$. □

The explicit form of the canonical transformation $(q,p) \to (\lambda,\mu)$ to the Nijenhuis coordinates is given in what follows. From relation (48) we get $q_k = \varphi_k(\lambda)$. Then, we introduce the generating function $S = \Sigma_{i=1}^n p_i \varphi_i(\lambda)$, such that

$$q_i = \frac{\partial S}{\partial p_i}, \quad \mu_i = \frac{\partial S}{\partial \lambda_i}. \tag{49}$$

The first equation reconstructs $q_k = \varphi_k(\lambda)$ and solving the second one with respect to p_k, we get the missing part of the canonical transformation in the form $p_k = \psi_k(\lambda,\mu)$.

Theorem 2: An arbitrary bi-Hamiltonian finite-dimensional system with two degenerated Poisson tensors of the form (6), where θ_{nc} takes the form (8), whose chain starts with a Casimir of

a noncanonical Poisson structure and terminates with a Casimir of the canonical structure, is equivalent to a quasi-bi-Hamiltonian system, which in the Nijenhuis coordinates takes an inverse-Pfaffian form.

Proof: Let

$$\begin{aligned}
&\overline{\theta}_{-1}\circ\overline{\nabla h_1}=0,\\
&\overline{\theta}_{-1}\circ\overline{\nabla h_2}=\overline{K}_1=\overline{\theta}_0\circ\overline{\nabla h_1},\\
&\overline{\theta}_{-1}\circ\overline{\nabla h_3}=\overline{K}_2=\overline{\theta}_0\circ\overline{\nabla h_2},\\
&\qquad\vdots\\
&\overline{\theta}_{-1}\circ\overline{\nabla h_{n+1}}=\overline{K}_n=\overline{\theta}_0\circ\overline{\nabla h_n},\\
&\qquad 0=\overline{\theta}_0\circ\overline{\nabla h_{n+1}}
\end{aligned}\tag{50}$$

be a bi-Hamiltonian chain on $U\subset\overline{M}\ni(q,p,c)$ with two degenerated Poisson tensors, each with one Casimir, so that $\dim\overline{M}=2n+1$ and $\operatorname{rank}\overline{\theta}_0(\overline{\theta}_{-1})=2n$, where

$$\overline{\theta}_{-1}=\begin{pmatrix} 0 & A(q) & \dfrac{\partial\overline{h}_n}{\partial p}\\ -A^T(q) & B(q,p) & -\dfrac{\partial\overline{h}_n}{\partial q}\\ -\left(\dfrac{\partial\overline{h}_n}{\partial p}\right)^T & -\left(\dfrac{\partial\overline{h}_n}{\partial q}\right)^T & 0\end{pmatrix},\quad \overline{\theta}_0=\begin{pmatrix} 0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\tag{51}$$

are compatible and where $\overline{h}_r(q,p,c)=\overline{h}_r(q,p)+c\tilde{\rho}_r(q)$. Like in the previous case (Theorem 6.4), in the original phase space $M\ni(q,p)$, $\theta_c=\theta_0$ and $\theta_{\mathrm{nc}}=\theta_{-1}$ constitute a compatible pair of Poisson tensors, but as in the previous case, the tensor θ_{-1} is not invariant along any flow of K_r. Because θ_0 is invertible, hence $\Phi^{-1}=\theta_{-1}\circ\theta_0^{-1}$ is the inverse of the Nijenhuis operator Φ but is not a recursion operator for K_r.

Consider the first equation from the hierarchy (50): $\overline{\theta}_{-1}\circ\overline{h}_1=0$. On M it reads as

$$\theta_{-1}\circ\nabla h_1=-\tilde{\rho}_1K_n,\quad \tilde{\rho}_1=\frac{\partial h_1}{\partial_c},\tag{52}$$

and the whole hierarchy turns into

$$\begin{aligned}
&-\frac{1}{\tilde{\rho}_1}\cdot\theta_{-1}\circ\nabla h_1=K_n=\theta_0\circ\nabla h_n,\\
&\theta_{-1}\circ\nabla h_n-\frac{\tilde{\rho}_n}{\tilde{\rho}_1}\cdot\theta_{-1}\circ\nabla h_1=K_{n-1}=\theta_0\circ\nabla h_{n-1},\\
&\qquad\vdots\\
&\theta_{-1}\circ\nabla h_2-\frac{\tilde{\rho}_2}{\tilde{\rho}_1}\cdot\theta_{-1}\circ\nabla h_1=K_1=\theta_0\circ\nabla h_1,
\end{aligned}\tag{53}$$

where $\tilde{\rho}_r=\partial h_r/\partial c$, $\tilde{\rho}_{n+1}=1$, $h_r=\overline{h}_r$, and $\overline{K}_r=(K_r,0)^T$. The last equation terminates the sequence of vector fields K_r in the hierarchy; as for the next equation from the chain we have

$$\theta_{-1}\circ\nabla h_1-\theta_{-1}\circ\nabla h_1=K_0=0.\tag{54}$$

Because of the form of the first equation from the hierarchy (53), the vector field $K \equiv K_n$ is called quasi-bi-Hamiltonian. Now, from the relation

$$\Phi^{-1} \circ K_r = \theta_{-1} \circ \nabla h_r = K_{r-1} + \frac{\tilde{\rho}_r}{\tilde{\rho}_1} \cdot \theta_{-1} \circ \nabla h_1 = K_{r-1} - \tilde{\rho}_r K_n, \tag{55}$$

one finds that

$$K_{n-r} = \Phi_r^{-1} \circ K_n = \sum_{i=0}^{r} \tilde{\rho}_{r+1-i} \Phi^{-r+i} \circ K_n, \tag{56}$$

and for $r=n$,

$$K_0 = \Phi_n^{-1} K_n = \sum_{i=0}^{n} \tilde{\rho}_{n+1-i} \Phi^{-n+i} \circ K_n = 0; \tag{57}$$

hence

$$\Phi^{-n} + \tilde{\rho}_n \Phi^{-n+1} + \cdots + \tilde{\rho}_2 \Phi^{-1} + \tilde{\rho}_1 I = 0. \tag{58}$$

In Nijenhuis coordinates we have

$$\theta_0 = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \theta_{-1} = \begin{pmatrix} 0 & \Lambda^{-1} \\ -\Lambda^{-1} & 0 \end{pmatrix}, \quad \Phi^{-1} = \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Lambda^{-1} \end{pmatrix}, \tag{59}$$

where $\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_n)$, and for all λ_i different there is a unique solution to Eq. (58) in the form

$$\tilde{\rho}_{n+1-k} = (-1)^k \sum_{\substack{j_1,\ldots,j_k \\ j_1 < \ldots < j_k}} \frac{1}{\lambda_{j_1} \cdot \cdots \cdot \lambda_{j_k}}, \quad k = 1,\ldots,n. \tag{60}$$

So, the first equation (53) takes the inverse-Pfaffian form (32), where now $H = h_n$, $F = h_1$, and $\rho = -\tilde{\rho}_1$. □

IV. SEPARABILITY OF BI-HAMILTONIAN AND QUASI-BI-HAMILTONIAN SYSTEMS

In the previous section we introduced the notion of quasi-bi-Hamiltonian systems, their Pfaffian, and inverse-Pfaffian forms. Moreover, we proved that each bi-Hamiltonian system with two degenerated Poisson tensors always has a quasi-bi-Hamiltonian representation. In the following section we consider the inverse statement, which says that each Pfaffian or inverse-Pfaffian quasi-bi-Hamiltonian system with $c \neq 0$ has a bi-Hamiltonian representation with a bi-Hamiltonian chain, and, moreover, we prove the separability of such a class of dynamical systems.

Theorem 3: For a quasi-bi-Hamiltonian system, the Hamiltonian functions H and F, written in terms of the Nijenhuis coordinates (λ_i, μ_i) in the form (33), are separable for each n-ple of functions $f_i(\lambda_i, \mu_i)$ when

$$g_H(\lambda) = 0 \quad \text{and} \quad g_F(\lambda) = 0,$$

$$g_H(\lambda) = c(\lambda_1 + \cdots + \lambda_n) \quad \text{and} \quad g_F(\lambda) = c\rho,$$

$$g_H(\lambda) = \frac{c}{\rho} \quad \text{and} \quad g_F(\lambda) = c\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}\right), \tag{61}$$

where c is an arbitrary constant.

Proof: The Hamilton–Jacobi equation for H is separable iff H verifies the Levi-Civita condition. It is in the form of a set of nonlinear partial differential equations,

$$\frac{\partial H}{\partial \lambda_i}\frac{\partial H}{\partial \lambda_j}\frac{\partial^2 H}{\partial \mu_i \partial \mu_j}+\frac{\partial H}{\partial \mu_i}\frac{\partial H}{\partial \mu_j}\frac{\partial^2 H}{\partial \lambda_i \partial \lambda_j}-\frac{\partial H}{\partial \lambda_i}\frac{\partial H}{\partial \mu_j}\frac{\partial^2 H}{\partial \mu_i \partial \lambda_j}-\frac{\partial H}{\partial \mu_i}\frac{\partial H}{\partial \lambda_j}\frac{\partial^2 H}{\partial \lambda_i \partial \mu_j}=0, \tag{62}$$

for all pairs (i,j), $i\neq j$. Applying the relations

$$\frac{\partial}{\partial \lambda_i}\frac{1}{\Delta_i}=-\frac{1}{\Delta_i}\sum_{\alpha\neq i}\frac{1}{\lambda_{i\alpha}},\quad \frac{\partial}{\partial \lambda_\beta}\frac{1}{\Delta_i}=-\frac{1}{\Delta_i}\frac{1}{\lambda_{i\beta}},\quad \beta\neq i, \tag{63}$$

one directly verifies the relation (62) for $H(\lambda,\mu)$ and $F(\lambda,\mu)$. □

The case $g_H(\lambda)=g_F(\lambda)=0$ was proved for the first time in Ref. 14.

Theorem 4:

(i) An arbitrary quasi-bi-Hamiltonian system, which in Nijenhuis coordinates has the Pfaffian form (30), has n independent constants of motion of the form

$$h_r=\sum_{i=1}^{n}\frac{\rho_{r-1}^{i}}{\Delta_i}f_i(\lambda_i,\mu_i)+c\rho_r,\quad r=1,...,n, \tag{64}$$

where

$$\Delta_i=\prod_{j\neq i}\lambda_{ij},\quad \lambda_{ij}=\lambda_i-\lambda_j,$$
$$\rho_r^i:=(-1)^r\sum_{\substack{j_1,...,j_r\\ j_1<..<j_r\\ j_1\neq i,..,j_r\neq i}}\lambda_{j_1}\cdot\ \cdots\ \cdot\lambda_{j_r}, \tag{65}$$

ρ_r are defined by (48), and n functions f_i are arbitrary smooth functions.

(ii) Arbitrary function h_r is separable for each n-ple of functions $f_i(\lambda_i,\mu_i)$, that is, an appropriate Hamilton–Jacobi equation can be solved.

(iii) On the phase space $M\ni(\lambda,\mu)$, functions h_k, $k=1,...,n$, are in involution with respect to Poisson tensors,

$$\theta_0=\begin{pmatrix}0 & I\\ -I & 0\end{pmatrix},\quad \theta_1=\begin{pmatrix}0 & \Lambda\\ -\Lambda & 0\end{pmatrix}.$$

(iv) On the extended phase space $\overline{M}\ni(\lambda,\mu,c)$, functions h_r with $c\neq 0$ form a bi-Hamiltonian chain with respect to degenerated Poisson tensors,

$$\overline{\theta_0}=\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad \overline{\theta_1}=\begin{pmatrix}0 & \Lambda & \dfrac{\partial h_1}{\partial \mu}\\ -\Lambda & 0 & -\dfrac{\partial h_1}{\partial \lambda}\\ -\left(\dfrac{\partial h_1}{\partial \mu}\right)^T & \left(\dfrac{\partial h_1}{\partial \lambda}\right)^T & 0\end{pmatrix}.$$

The chain starts with the Casimir $h_0=c$ of the first Poisson tensor $\overline{\theta_0}$ and terminates with the Casimir h_n of the second Poisson tensor $\overline{\theta_1}$.

Proof:

(iii) Let $\gamma_r=\nabla h_r$; then by direct calculations one can verify that

$$\gamma_{r+1}=\Phi^{\dagger}\gamma_r+\rho_r\gamma_1, \tag{66}$$

where

$$\Phi^{\dagger}=\theta_0^{-1}\circ\theta_1=\begin{pmatrix}\Lambda & 0\\ 0 & \Lambda\end{pmatrix},$$

and ρ_r is given by Eq. (48). In the calculus the following relations are helpful:

$$\frac{\partial\rho_r}{\partial\lambda_i}=-\rho^i_{r-1},\quad \lambda_i\frac{\partial\rho_r}{\partial\lambda_i}=\rho_r-\rho^i_r,\quad \lambda_i\rho^i_{r-1}=-\rho_r+\rho^i_r,$$

$$\frac{\partial\rho^k_r}{\partial\lambda_i}=-\rho^{ki}_{r-1},\quad \lambda_i\frac{\partial\rho^k_r}{\partial\lambda_i}=\rho^k_r-\rho^{ki}_r,\quad \lambda_i\rho^{ki}_{r-1}=-\rho^k_r+\rho^{ki}_r, \tag{67}$$

$$\rho^{ki}_r:=(-1)^r\sum_{\substack{j_1,\ldots,j_r\\ j_1<\ldots<j_r\\ j_1\neq i,k,\ldots,j_r\neq i,k}}\lambda_{j_1}\cdot\ \cdots\ \cdot\lambda_{j_r},$$

as well as relations (63). From (66), it follows that

$$\gamma_{r+1}=\Phi^{\dagger}_r\circ\gamma_1=\sum_{i=0}^{r}\rho_{r-i}(\Phi^{\dagger})^i\circ\gamma_1,$$

and thus, because $\theta_0\circ\Phi^{\dagger}=\Phi\circ\theta_0$, we have

$$\begin{aligned}\{h_l,h_k\}_{\theta_0}&=\langle\nabla h_k,\theta_0\circ\nabla h_l\rangle\\ &=\sum_{i=0}^{l}\sum_{j=0}^{k}\rho_{r-i}\rho_{k-j}\langle\Phi^{\dagger j}\circ\gamma_1,\theta_0\circ\Phi^{\dagger i}\circ\gamma_1\rangle\\ &=\sum_{i=0}^{l}\sum_{j=0}^{k}\rho_{r-i}\rho_{k-j}\langle\gamma_1,\Phi^{i+j}\circ\theta_0\circ\gamma_1\rangle=0,\end{aligned}$$

where the last equality follows from skew symmetry of the tensor $\Phi^m\circ\theta_0$ for any m. As $\theta_1=\Phi\circ\theta_0$ we have

$$\begin{aligned}\{h_l,h_k\}_{\theta_1}&=\langle\nabla h_k,\theta_k,\theta_1\circ\nabla h_l\rangle\\ &=\sum_{i=0}^{l}\sum_{j=0}^{k}\rho_{r-i}\rho_{k-j}\langle\Phi^{\dagger j}\circ\gamma_1,\theta_1\circ\Phi^{\dagger i}\circ\gamma_1\rangle\\ &=\sum_{i=0}^{l}\sum_{j=0}^{k}\rho_{r-i}\rho_{k-j}\langle\gamma_1,\Phi^{i+j}\circ\theta_1\circ\gamma_1\rangle\\ &=\sum_{i=0}^{l}\sum_{j=0}^{k}\rho_{r-i}\rho_{k-j}\langle\gamma_1,\Phi^{i+j+1}\circ\theta_0\circ\gamma_1\rangle=0.\end{aligned}$$

□

(i) Follows immediately from (iii) as

$$L_K h_{r+1}=\langle\nabla\gamma_{r+1},\theta_0\circ\gamma_1\rangle=0.$$

□

(ii) We prove the statement by direct verification of the Levi-Civita condition (62) for each h_r. Notice that $h_1=H$ and $h_n=F$, where H and F are given in Lemma 1, as $\rho^i_0=1$ and $\rho^i_{n-1}=-\rho_n/\lambda_i=\rho/\lambda_i$. □

(iv) The relation $\overline{\theta_0}\circ\nabla h_{r+1}=\overline{\theta_1}\circ\nabla h_r$ is equivalent to the following equalities:

$$\frac{\partial h_{r+1}}{\partial \mu_i}=\lambda_i \frac{\partial h_r}{\partial \mu_i}+\rho_r \frac{\partial h_1}{\partial \mu_i},$$

$$\frac{\partial h_{r+1}}{\partial \lambda_i}=\lambda_i \frac{\partial h_r}{\partial \lambda_i}+\rho_r \frac{\partial h_1}{\partial \lambda_i}, \quad i=1,...,n,$$

$$\left(\frac{\partial h_1}{\partial \lambda}\right)^T \frac{\partial h_r}{\partial \mu}-\left(\frac{\partial h_1}{\partial \mu}\right)^T \frac{\partial h_r}{\partial \lambda}=0.$$

The first two are equivalent to the relation (66) while the last one is equal to $\{h_r,h_1\}_{\theta_0}=0$.□

Theorem 5:

(i) An arbitrary quasi-bi-Hamiltonian system, which in Nijenhuis coordinates has the inverse-Pfaffian form (32), has n independent constants of motion of the form

$$h_r=\sum_{i=1}^{n} \frac{\rho_{r-1}^i}{\Delta_i} f_i(\lambda_i,\mu_i)+c\tilde{\rho}_r, \quad r=1,...,n, \tag{68}$$

where Δ_i, ρ_r^i, and f_i are the same as in Theorem 4.

(ii) Arbitrary function h_r is separable for each n-ple of functions $f_i(\lambda_i,\mu_i)$, that is, an appropriate Hamilton–Jacobi equation can be solved.

(iii) On the phase space $M\ni(\lambda,\mu)$, functions h_k, $k=1,...,n$, are in involution with respect to Poisson tensors,

$$\theta_0=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \theta_{-1}=\begin{pmatrix} 0 & \Lambda^{-1} \\ -\Lambda^{-1} & 0 \end{pmatrix}. \tag{69}$$

(iv) On the extended phase space $\overline{M}\ni(\lambda,\mu,c)$, h_r functions with $c\neq 0$ form a bi-Hamiltonian chain with respect to degenerated Poisson tensors,

$$\overline{\theta}_0=\begin{pmatrix} 0 & I & 0 \\ -I & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \overline{\theta}_{-1}=\begin{pmatrix} 0 & \Lambda^{-1} & \dfrac{\partial h_n}{\partial \mu} \\ -\Lambda^{-1} & 0 & -\dfrac{\partial h_n}{\partial \lambda} \\ -\left(\dfrac{\partial h_n}{\partial \mu}\right)^T & \left(\dfrac{\partial h_n}{\partial \lambda}\right)^T & 0 \end{pmatrix}. \tag{70}$$

The chain starts with the Casimir h_1 of the noncanonical Poisson tensor $\overline{\theta}_{-1}$ and terminates with the Casimir $h_{n+1}=c$ of the canonical Poisson tensor $\overline{\theta}_0$.

Proof: The proof is analogous to the one of Theorem 4, where we additionally use the relation $\tilde{\rho}_r=-\rho_{r-1}/\rho_n$.

V. EXAMPLES

The theory developed in previous sections will be illustrated by a few representative examples. We start with the Henon–Heiles system introduced in the beginning of this paper.

Example 1: Separability of the Henon–Heiles system and its generalizations.

For the first generalization of the Henon–Heiles system (17)–(21), we find

$$\frac{\partial h_1}{\partial c}=\frac{\partial \overline{H}}{\partial c}=-q_1=-\lambda_1-\lambda_2,$$

$$\frac{\partial h_2}{\partial c}=\frac{\partial\left(-\frac{1}{4}\overline{F}\right)}{\partial c}=-\frac{1}{4}q_2^2=\lambda_1\lambda_2, \tag{71}$$

and hence

$$q_1=\lambda_1+\lambda_2, \quad q_2=2\sqrt{-\lambda_1\lambda_2}. \tag{72}$$

From the generating function $S=p_1(\lambda_1+\lambda_2)+2p_2\sqrt{-\lambda_1\lambda_2}$, we get the conjugate momenta

$$p_1=\sqrt{-\lambda_1\lambda_2}\left(\frac{\mu_1}{\Delta_1}+\frac{\mu_2}{\Delta_2}\right), \quad p_2=\frac{\lambda_1\mu_1}{\Delta_1}+\frac{\lambda_2\mu_2}{\Delta_2}, \tag{73}$$

where $\Delta_1=-\Delta_2=\lambda_1-\lambda_2$. In Nijenhuis coordinates, the conserved functions read as

$$\begin{aligned} h_1&=\frac{\lambda_1(\mu_1^2+2\lambda_1^3)}{2\Delta_1}+\frac{\lambda_2(\mu_2^2+2\lambda_2^3)}{2\Delta_2}-c(\lambda_1+\lambda_2),\\ h_2&=-\frac{\lambda_2\lambda_1(\mu_1^2+2\lambda_1^3)}{2\Delta_1}-\frac{\lambda_1\lambda_2(\mu_2^2+2\lambda_2^3)}{2\Delta_2}+c\lambda_1\lambda_2. \end{aligned} \tag{74}$$

The Hamilton–Jacobi equations $h_i(\lambda,\partial W_i/\partial\lambda)=E_i$, $i=1,2$, are separable and have the complete integrals $W_i=\Sigma_{k=1}^2 W_i^{(k)}(\lambda_k;c_0,c_1)$, where $W_i^{(k)}$ are solutions of the following equations:

$$\frac{dW_i^{(k)}}{d\lambda_k}=\left[\frac{1}{\lambda_k}\,(-2\lambda_k^4+c\lambda_k^2+c_1\lambda_k+c_0)\right]^{1/2}, \quad c_{2-i}=2E_i, \quad i,k=1,2. \tag{75}$$

For the second generalized Henon–Heiles representation (22)–(26), from the relations

$$\begin{aligned} \frac{\partial h_1}{\partial d}&=\frac{\partial(-2\overline{H})}{\partial d}=-\frac{1}{q_2^2}=\frac{1}{\lambda_1\lambda_2},\\ \frac{\partial h_2}{\partial d}&=\frac{\partial\overline{F}}{\partial d}=\frac{2q_1}{q_2^2}=-\frac{1}{\lambda_1}-\frac{1}{\lambda_2}, \end{aligned} \tag{76}$$

we find the following Nijenhuis coordinates:

$$\begin{aligned} q_1&=\frac{1}{2}\,(\lambda_1+\lambda_2), \quad q_2=\sqrt{-\lambda_1\lambda_2},\\ p_1&=2\sqrt{-\lambda_1\lambda_2}\left(\frac{\mu_1}{\Delta_1}+\frac{\mu_2}{\Delta_2}\right), \quad p_2=2\left(\frac{\lambda_1\mu_1}{\Delta_1}+\frac{\lambda_2\mu_2}{\Delta_2}\right), \end{aligned} \tag{77}$$

and the following conserved functions:

$$\begin{aligned} h_1&=-\frac{\lambda_1(16\mu_1^2+\lambda_1^3)}{4\Delta_1}-\frac{\lambda_2(16\mu_2^2+\lambda_2^3)}{4\Delta_2}+\frac{d}{\lambda_1\lambda_2},\\ h_2&=\frac{\lambda_2\lambda_1(16\mu_1^2+\lambda_1^3)}{4\Delta_1}+\frac{\lambda_1\lambda_2(16\mu_2^2+\lambda_2^3)}{4\Delta_2}-d\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right). \end{aligned} \tag{78}$$

Again, the Hamilton–Jacobi equations $h_i(\lambda,\partial W_i/\partial\lambda)=E_i$, $i=1,2$, are separable and have the complete integrals $W_i=\Sigma_{k=1}^2 W_i^{(k)}(\lambda_k;c_0,c_1)$, where $W_i^{(k)}$ are solutions of the following equations:

$$\frac{dW_i^{(k)}}{d\lambda_k}=\left[\frac{1}{\lambda_k}\left(-16\lambda_k^4-c_1\lambda_k-c_0-\frac{4d}{\lambda_k}\right)\right]^{1/2}, \quad c_{2-i}=4E_i, \quad i,k=1,2. \tag{79}$$

Example 2: Separability of the second Newton representation for the stationary flow of the Harry Dym equation.

Newton dynamics,

$$q_{1_{xx}}=8q_1^{-5}q_2, \qquad q_{2_{xx}}=-2q_1^{-4}-d, \tag{80}$$

is another example of a bi-Hamiltonian system recently derived in Ref. 12. With $p_1=q_{1_x}$ and $p_2=q_{2_x}$ the related bi-Hamiltonian chain (35) is given by Hamiltonian functions

$$\begin{aligned} h_0&=d,\\ h_1&=\tfrac{1}{2}p_1^2+\tfrac{1}{2}p_2^2+2q_1^{-4}q_2+dq_2,\\ h_2&=\tfrac{1}{2}q_2p_1^2-\tfrac{1}{2}q_1p_1p_2+\tfrac{1}{2}q_1^{-2}+2q_1^{-4}q_2^2-\tfrac{1}{4}dq_1^2, \end{aligned} \tag{81}$$

canonical Poisson tensor $\overline{\theta_0}$, and the noncanonical one,

$$\overline{\theta_1}=\begin{pmatrix} 0 & 0 & 0 & -\frac{1}{2}q_1 & p_1\\ 0 & 0 & -\frac{1}{2}q_1 & -q_2 & p_2\\ 0 & \frac{1}{2}q_1 & 0 & \frac{1}{2}q_1 & -8q_1^{-5}q_2\\ \frac{1}{2}q_1 & q_2 & -\frac{1}{2}q_1 & 0 & 2q_1^{-4}+d\\ -p_1 & -p_2 & 8q_1^{-5}q_2 & -2q_1^{-4}-d & 0 \end{pmatrix}. \tag{82}$$

Hence, we have

$$\rho_1=\frac{\partial h_1}{\partial d}=q_2=-\lambda_1-\lambda_2, \qquad \rho_2=\frac{\partial h_2}{\partial d}=-\tfrac{1}{4}q_1^2=\lambda_1\lambda_2, \tag{83}$$

and then

$$q_1=2\sqrt{-\lambda_1\lambda_2}, \quad q_2=-\lambda_1-\lambda_2. \tag{84}$$

Introducing a generating function $S=2p_2\sqrt{-\lambda_1\lambda_2}-p_2(\lambda_1+\lambda_2)$ and inverting the system $\mu=\partial S/\partial\lambda$, we get the second part of the canonical transformation,

$$p_1=\sqrt{-\lambda_1\lambda_2}\left(\frac{\mu_1}{\Delta_1}+\frac{\mu_2}{\Delta_2}\right), \quad p_2=-\lambda_1\frac{\mu_1}{\Delta_1}-\lambda_2\frac{\mu_2}{\Delta_2}, \tag{85}$$

and hence

$$\begin{aligned} h_1&=\frac{\lambda_1(4\mu_1^2+\lambda_1^{-3})}{8\Delta_1}+\frac{\lambda_2(4\mu_2^2+\lambda_2^{-3})}{8\Delta_2}-d(\lambda_1+\lambda_2),\\ h_2&=-\frac{\lambda_2\lambda_1(4\mu_1^2+\lambda_1^{-3})}{8\Delta_1}-\frac{\lambda_1\lambda_2(4\mu_2^2+\lambda_2^{-3})}{8\Delta_2}+d\lambda_1\lambda_2. \end{aligned} \tag{86}$$

The Hamilton–Jacobi equations $h_i(\lambda,\partial W_i/\partial\lambda)=E_i$, $i=1,2$, are separable and have the complete integrals $W_i=\Sigma_{k=1}^{2}W_i^{(k)}(\lambda_k;c_0,c_1)$, where $W_i^{(k)}$ are solutions of the following equations:

$$\frac{dW_i^{(k)}}{d\lambda_k}=\left[\frac{1}{4\lambda_k}\left(-\lambda_k^{-2}+4d\lambda_k^2+c_1\lambda_k+c_0\right)\right]^{1/2},\quad c_{2-i}=8E_i,\quad i,k=1,2. \tag{87}$$

Example 3: Separability of the first Newton representation of the seventh-order stationary flow of the KdV.

We consider a Newton system,

$$\begin{aligned}0&=r_{1_{xx}}+10r_1^2-4r_2,\\ 0&=r_{2_{xx}}+16r_1r_2-10r_1^3-4r_3,\\ 0&=r_{3_{xx}}+20r_1r_3+8r_2^2-30r_1^2r_2+15r_1^4-c,\end{aligned} \tag{88}$$

with a related bi-Hamiltonian chain given by the conserved functions

$$\begin{aligned}h_0&=c,\\ h_1&=s_1s_3+\tfrac{1}{2}s_2^2+10r_1^2r_3-4r_2r_3+8r_1r_2^2-10r_1^3r_2+3r_1^5-cr_1,\\ h_2&=\tfrac{1}{2}r_3s_3^2-\tfrac{1}{2}r_1s_2^2+\tfrac{1}{2}r_2s_2s_3-\tfrac{1}{2}s_1s_2-\tfrac{1}{2}r_1s_1s_3+2r_1^2r_2^2+\tfrac{5}{2}r_1^4r_2-\tfrac{5}{4}r_1^6-2r_2^3+r_3^2-6r_1r_2r_3\\ &\quad+\tfrac{1}{4}cr_1^2+\tfrac{1}{2}cr_2,\\ h_3&=\tfrac{1}{8}r_2^2s_3^2+\tfrac{1}{8}r_1^2s_2^2+\tfrac{1}{8}s_1^2+\tfrac{1}{4}r_1s_1s_2+\tfrac{1}{4}r_2s_1s_3-\tfrac{1}{4}r_1r_2s_2s_3-\tfrac{1}{2}r_3s_2s_3-3r_1^3r_2^2+r_1r_2^3+\tfrac{5}{4}r_1^5r_2\\ &\quad+2r_3^2r_1+\tfrac{5}{4}r_1^4r_3+r_2^2r_3-r_1^2r_2r_3-\tfrac{1}{4}cr_1r_2-\tfrac{1}{4}cr_3,\end{aligned} \tag{89}$$

with the 7×7 Poisson matrices $\overline{\theta_0}$ and $\overline{\theta_1}$, where

$$\overline{\theta_1}=\begin{pmatrix}0&0&0&r_1/2&-1/2&0&s_3\\ 0&0&0&r_2/2&0&-1/2&s_2\\ 0&0&0&r_3&r_2/2&r_1/2&s_1\\ -r_1/2&-r_2/2&-r_3&0&s_2/2&s_3/2&c-\partial V/\partial r_1\\ 1/2&0&-r_2/2&-s_2/2&0&0&-\partial V/\partial r_2\\ 0&1/2&-r_1/2&-s_3/2&0&0&-\partial V/\partial r_3\\ -s_3&-s_2&-s_1&\partial V/\partial r_1-c&\partial V/\partial r_2&\partial V/\partial r_3&0\end{pmatrix}, \tag{90}$$

and $V(r)=10r_1^2r_3-4r_2r_3+8r_1r_2^2-10r_1^3r_2+3r_1^5$. This bi-Hamiltonian chain contains three nontrivial vector fields that commute in pairs. Integrals of motion also commute and are functionally independent, since the Jacobian $\det[\partial(h_1,h_2,h_3)/\partial(s_1,s_2,s_3)]$ does not vanish. Hence, we have

$$\begin{aligned}\rho_1&=\frac{\partial h_1}{\partial c}=-r_1=-\lambda_1-\lambda_2-\lambda_3,\\ \rho_2&=\frac{\partial h_2}{\partial c}=\frac{1}{4}r_1^2+\frac{1}{2}r_2=\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3,\\ \rho_3&=\frac{\partial h_3}{\partial c}=-\frac{1}{4}r_1r_2-\frac{1}{4}r_3=-\lambda_1\lambda_2\lambda_3\end{aligned} \tag{91}$$

and then

$$r_1=\lambda_1+\lambda_2+\lambda_3,$$

$$r_2=(\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3)-\tfrac{1}{2}(\lambda_1^2+\lambda_2^2+\lambda_3^2), \tag{92}$$

$$r_3=4\lambda_1\lambda_2\lambda_3+\tfrac{1}{3}(\lambda_1+\lambda_2+\lambda_3)[4(\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3)-\tfrac{1}{3}(\lambda_1^2+\lambda_2^2+\lambda_3^2)].$$

From the appropriate generating function, we calculate

$$s_1=-\frac{1}{8}\left\{[(-3\lambda_1^2+\lambda_2^2+\lambda_3^2)+2(\lambda_1\lambda_2+\lambda_1\lambda_3-\lambda_2\lambda_3)]\,\frac{\mu_1}{\Delta_1}+\text{c.p.}\right\},$$

$$s_2=\frac{1}{4}\left[(-\lambda_1+\lambda_2+\lambda_3)\,\frac{\mu_1}{\Delta_1}+(\lambda_1-\lambda_2+\lambda_3)\,\frac{\mu_2}{\Delta_2}+(\lambda_1+\lambda_2-\lambda_3)\,\frac{\mu_3}{\Delta_3}\right], \tag{93}$$

$$s_3=\frac{1}{4}\left(\frac{\mu_1}{\Delta_1}+\frac{\mu_2}{\Delta_2}+\frac{\mu_3}{\Delta_3}\right),$$

where $\Delta_1=(\lambda_1-\lambda_2)(\lambda_1-\lambda_3),\ldots,$ and c.p. stands for cyclic permutations of the indices 1,2,3, in λ, μ, and Δ. The three Hamiltonians in the Nijenhuis coordinates take the form

$$h_1=\frac{f_1}{\Delta_1}+\frac{f_2}{\Delta_2}+\frac{f_3}{\Delta_3}-c(\lambda_1+\lambda_2+\lambda_3),$$

$$h_2=-\frac{(\lambda_2+\lambda_3)f_1}{\Delta_1}-\frac{(\lambda_1+\lambda_3)f_2}{\Delta_2}-\frac{(\lambda_1+\lambda_2)f_3}{\Delta_3}+c(\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3), \tag{94}$$

$$h_3=\frac{\lambda_2\lambda_3 f_1}{\Delta_1}+\frac{\lambda_1\lambda_3 f_2}{\Delta_2}+\frac{\lambda_1\lambda_2 f_3}{\Delta_3}-c\lambda_1\lambda_2\lambda_3,$$

where $f_i(\lambda_i,\mu_i)=\frac{1}{8}(\mu_i^2+128\lambda_i^7)$. The appropriate Hamilton–Jacobi equations $h_i(\lambda,\partial W_i/\partial\lambda)=E_i$ are separable and have the complete integrals

$$W_i=\sum_{k=1}^{3} W_i^{(k)}(\lambda_k;c_0,c_1,c_2),\quad i=1,2,3, \tag{95}$$

with $W_i^{(k)}$ being solutions of the following equations:

$$\frac{dW_i^{(k)}}{d\lambda_k}=(-128\lambda_k^7+8c\lambda_k^3+c_2\lambda_k^2+c_1\lambda_k+c_0)^{1/2},\quad c_{3-i}=8E_i. \tag{96}$$

Example 4: Separability of the second Newton representation of the seventh-order stationary flow of the KdV.

Let us consider a Newton system,[12]

$$0=q_{1_{xx}}+10q_1^2-4q_2,$$

$$0=q_{2_{xx}}+20q_1q_2-10q_1^3+2q_3, \tag{97}$$

$$0=q_{3_{xx}}+4q_1q_3-dq_3^{-3}.$$

The related bi-Hamiltonian chain takes the form (50); h_r and $\overline{\theta}_{-1}$ are

$$h_1=-2p_1p_2-2p_3^2-20q_1^2q_2+4q_2^2+5q_1^4-4q_1q_3^2-dq_3^{-2},$$

$$h_2=-p_1^2-2q_1p_1p_2+2q_2p_2^2-2q_1p_3^2+2q_3p_2p_3+2q_1^2q_3^2-4q_2q_3^2-16q_1q_2^2+4q_1^5-2dq_2q_3^{-2},$$

$$h_3=-q_3^2p_2^2+2q_3p_1p_3+2q_1q_3p_2p_3-q_1^2p_3^2+8q_2p_3^2+q_3^4$$

$$+8q_1q_2q_3^2-4q_1^3q_3^2-dq_1^2q_3^{-2}-2dq_2q_3^{-2},$$

$$h_4=d, \tag{98}$$

and

$$\overline{\theta}_{-1}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1/q_3 & A_1 \\ 0 & 0 & 0 & 1 & 0 & -q_1/q_3 & A_2 \\ 0 & 0 & 0 & -q_1/q_3 & -1/q_3 & B_1 & A_3 \\ 0 & -1 & q_1/q_3 & 0 & 0 & B_2 & A_4 \\ 0 & 0 & 1/q_3 & 0 & 0 & -p_3/q_3^2 & A_5 \\ 1/q_3 & q_1/q_3 & -B_1 & -B_2 & p_3/q_3^2 & 0 & A_6 \\ -A_1 & -A_2 & -A_3 & -A_4 & -A_5 & -A_6 & 0 \end{pmatrix}, \tag{99}$$

where

$$p_1=q_{2_x}, \quad p_2=q_{1_x}, \quad p_3=q_{3_x}, \quad (A_1,\ldots,A_6,0)^T=\overline{\theta}_0\circ\nabla h_3,$$

$$B_1=(q_1^2+2q_2)/q_3^2, \quad B_2=(q_3p_2-q_1p_3)/q_3^2.$$

Hence, we find

$$\begin{aligned} \tilde{\rho}_1&=-q_3^{-2}=-\frac{1}{\lambda_1\lambda_2\lambda_3}, \\ \tilde{\rho}_2&=-2q_1q_3^{-2}=\frac{1}{\lambda_1\lambda_2}+\frac{1}{\lambda_1\lambda_3}+\frac{1}{\lambda_2\lambda_3}, \\ \tilde{\rho}_3&=-q_1^2q_3^{-2}-2q_2q_3^{-2}=-\frac{1}{\lambda_1}-\frac{l}{\lambda_2}-\frac{1}{\lambda_3}, \end{aligned} \tag{100}$$

and then the Nijenhuis map

$$\begin{aligned} q_1&=-\tfrac{1}{2}(\lambda_1+\lambda_2+\lambda_3), \\ q_2&=\tfrac{1}{2}(\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3)-\tfrac{1}{8}(\lambda_1+\lambda_2+\lambda_3)^2, \\ q_3&=(\lambda_1\lambda_2\lambda_3)^{1/2}, \\ p_1&=\lambda_1(-\lambda_1+\lambda_2+\lambda_3)\frac{\mu_1}{\Delta_1}+\lambda_2(\lambda_1-\lambda_2+\lambda_3)\frac{\mu_2}{\Delta_2}+\lambda_3(\lambda_1+\lambda_2-\lambda_3)\frac{\mu_3}{\Delta_3}, \\ p_2&=-2\left(\lambda_1\frac{\mu_1}{\Delta_1}+\lambda_2\frac{\mu_2}{\Delta_2}+\lambda_3\frac{\mu_3}{\Delta_3}\right), \\ p_3&=2(\lambda_1\lambda_2\lambda_3)^{1/2}\left(\frac{\mu_1}{\Delta_1}+\frac{\mu_2}{\Delta_2}+\frac{\mu_3}{\Delta_3}\right). \end{aligned} \tag{101}$$

The three Hamiltonians (98) in the Nijenhuis coordinates take the form

$$h_1=\frac{f_1}{\Delta_1}+\frac{f_2}{\Delta_2}+\frac{f_3}{\Delta_3}-d\,\frac{1}{\lambda_1\lambda_2\lambda_3},$$

$$h_2=-\frac{(\lambda_2+\lambda_3)f_1}{\Delta_1}-\frac{(\lambda_1+\lambda_3)f_2}{\Delta_2}-\frac{(\lambda_1+\lambda_2)f_3}{\Delta_3}+d\left(\frac{1}{\lambda_1\lambda_2}+\frac{1}{\lambda_1\lambda_3}+\frac{1}{\lambda_2\lambda_3}\right), \tag{102}$$

$$h_3=\frac{\lambda_2\lambda_3 f_1}{\Delta_1}+\frac{\lambda_1\lambda_3 f_2}{\Delta_2}+\frac{\lambda_1\lambda_2 f_3}{\Delta_3}-d\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right),$$

where $f_i(\lambda_i,\mu_i)=\lambda_i(\lambda_i^5-4\mu_i^2)$. The appropriate Hamilton–Jacobi equations $h_i(\lambda,\partial W_i/\partial\lambda)=E_i$ are separable and have the complete integrals

$$\frac{dW_i^{(k)}}{d\lambda_k}=\left[\frac{1}{4\lambda_k}\left(\lambda_k^6-c_2\lambda_k^2-c_1\lambda_k-c_0-\frac{d}{\lambda_k}\right)\right]^{1/2},\quad c_{3-i}=E_i,\quad k=1,2,3. \tag{103}$$

Example 5: A bi-Hamiltonian formulation for *elliptic separable potentials*. Present results allow us to reexamine the theory of classical Hamiltonian systems with separable potentials in the frame of the bi-Hamiltonian theory. In Refs. 19 and 20, there has been formulated and proved a theorem that says that the Hamilton–Jacobi equation for a natural Hamiltonian system $H=T+V(q)$ is separable in generalized elliptic coordinates if it has N functionally independent and commuting integrals of motion of the form

$$K_i(q,p)=\frac{1}{2}\sum_{j=1,j\neq i}^{N}(\alpha_i-\alpha_j)^{-1}l_{ij}^2+p_i^2+V_i(q)=P_i(q,p)+V_i(q), \tag{104}$$

where q_i,p_i are canonically conjugated variables, $l_{ij}=q_ip_j-q_jp_i$ $(i,j=1,...,N)$ are angular momenta, α_i are different, positive constants, and $V_i(q)$ are functions of $q=(q_1,\ldots,q_N)$ such that the Poisson brackets $\{H,K_i\}=0$, $i=1,...,N$. The Hamiltonian is given as

$$H=\frac{1}{2}\sum_{i=1}^{N}K_i(q,p)=\frac{1}{2}(p,p)+\sum_{i=1}^{N}\frac{1}{2}V_i(q)=T+V(q). \tag{105}$$

Then, in Ref. 17, a theorem was proved that says that every natural Hamiltonian system $H=T+V(q)$ separable in generalized elliptic coordinates admits, in the extended phase space of variables (q,p,c), the bi-Hamiltonian formulation

$$\begin{pmatrix} q\\ p\\ c\end{pmatrix}_t=\theta_0\circ\nabla h_1=\theta_1\circ\nabla h_0,\quad \theta_0=\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix},$$
$$\theta_1=\begin{pmatrix}0 & A-\frac{1}{2}q\otimes q & q_t\\ -A+\frac{1}{2}q\otimes q & \frac{1}{2}p\otimes q-\frac{1}{2}q\otimes p & p_t\\ -q_t & -p_t & 0\end{pmatrix}, \tag{106}$$

where $A=\mathrm{diag}(\alpha_1,...,\alpha_N)$, $h_0=c$, $h_1=H+\frac{1}{2}c(q,q)$, and in the remaining column and row of θ_1 $q_t=\partial h_1/\partial p=p$, $p_t=-\partial h_1/\partial q=-\partial(V+\frac{1}{2}cq^2)/\partial q$ denote the components of the Hamiltonian vector field. The potential V satisfies the equations

$$0=(\alpha_i-\alpha_r)\frac{\partial^2 V}{\partial q_i\partial q_r}+\frac{3}{2}\left(q_r\frac{\partial V}{\partial q_i}-q_i\frac{\partial V}{\partial q_r}\right)+\frac{1}{2}\sum_{j=1}^{N}\left(q_jq_r\frac{\partial^2 V}{\partial q_i\partial q_j}-q_iq_j\frac{\partial^2 V}{\partial q_j\partial q_r}\right),\quad i,r=1,...,N, \tag{107}$$

which are the iff condition for θ_1 to fulfill the Jacobi identity.

This bi-Hamiltonian formulation generates an infinite chain of commuting bi-Hamiltonian vector fields,

$$\theta_0 \circ \nabla h_0 = 0, \ldots, \theta_0 \circ \nabla h_{r+1} = \theta_1 \circ \nabla h_r, \ldots, \quad r = 0,1,2,\ldots, \tag{108}$$

with $h_0 = c$ and

$$h_r = \frac{1}{2} \sum_{i=1}^{N} \alpha_i^{r-1} \tilde{K}_i = \frac{1}{2} \sum_{i=1}^{N} \alpha_i^{r-1} K_i + \frac{1}{2} c(q, A^{r-1} q), \quad r = 1,2,\ldots, \tag{109}$$

where $\tilde{K}_i = K_i + c q_i^2$ and K_i are given by (104). Moreover, we can define new Hamiltonians,

$$\tilde{h}_r = h_r + \sum_{k=1}^{r} \beta_{N-k} h_{r-k}, \quad r = 0,\ldots,N, \tag{110}$$

where β_r are solutions of the set of linear equations,

$$\alpha_i^N + \sum_{r=0}^{N-1} \beta_r \alpha_i^r = 0, \tag{111}$$

so that the new chain starts with $\tilde{h}_0 = h_0 = c$, a Casimir of θ_0, and terminates with

$$\tilde{h}_N = \beta_0 \left(\frac{1}{2} \sum_{i=1}^{N} \alpha_i^{-1} \tilde{K}_i - c \right), \tag{112}$$

which is a Casimir of θ_1.

Here we prove the inverse statement that each bi-Hamiltonian chain (106) is separable in generalized elliptic coordinates. Let us construct the Nijenhuis coordinates. The solution to the set of equations (111) takes the form

$$\beta_{N-r} = (-1)^r \sum_{\substack{j_1,\ldots,j_r \\ j_1 < \ldots < j_r}} \alpha_{j_1} \cdot \cdots \cdot \alpha_{j_r} := \rho_r(\alpha), \quad \beta_N = 1. \tag{113}$$

Hence, the c-dependent part of

$$\tilde{h}_r(q,p,c;\alpha) = \tilde{h}_r(q,p;\alpha) + c \rho_r(q;\alpha) \tag{114}$$

reads as

$$\rho_r(q;\alpha) = \beta_{N-r} + \frac{1}{2} \sum_{k=0}^{r-1} \beta_{N-k}(q, A^{r-k-1} q), \tag{115}$$

and defines the Nijenhuis coordinates (48) through the relation

$$\rho_r(\alpha) + \frac{1}{2} \sum_{k=0}^{r-1} \rho_k(\alpha)(q, A^{r-k-1} q) = \rho_r(\lambda). \tag{116}$$

But so defined Nijenhuis coordinates are just the generalized elliptic coordinates. Indeed, the generalized elliptic coordinates $\lambda_1, \ldots, \lambda_N$ defined by the relation[20]

$$1 + \frac{1}{2} \sum_{k=1}^{N} \frac{q_k^2}{z - \alpha_k} = \frac{\Pi_{j=1}^{N} (z - \lambda_j)}{\Pi_{k=1}^{N} (z - \alpha_k)}, \tag{117}$$

are equivalent to the Nijenhuis coordinates defined by relations (116). To see this let

$$A(z) := \prod_{k=1}^{N} (z - \alpha_k) = \sum_{r=0}^{N} \rho_r(\alpha) z^{N-r},$$

$$\Lambda(z) := \prod_{k=1}^{N} (z-\lambda_k) = \sum_{r=0}^{N} \rho_r(\lambda) z^{N-r}, \tag{118}$$

$$A_k(z) := \frac{A(z)}{z-\alpha_k} = \sum_{r=1}^{N} \rho_{r-1}^{k}(\alpha) z^{N-r},$$

where

$$\rho_r^i(\alpha) = (-1)^r \sum_{\substack{j_1,\ldots,j_r \\ j_1<\ldots<j_r \\ j_1\neq i,\ldots,j_r\neq i}} \alpha_{j_1} \cdot \cdots \cdot \alpha_{j_r}.$$

From (117), it follows that

$$\begin{aligned}\Lambda(z) = A(z) + \sum_{k=1}^{N} A_k(z) q_k^2 &= \sum_{r=0}^{N} \rho_r(\alpha) z^{N-r} + \frac{1}{2} \sum_{k=1}^{N} \sum_{r=1}^{N} \rho_{r-1}^{k}(\alpha) z^{N-r} q_k^2 \\ &= z^N + \sum_{r=1}^{N} \left[\rho_r(\alpha) + \frac{1}{2} \sum_{k=1}^{N} \rho_{r-1}^{k}(\alpha) q_k^2 \right] z^{N-r}. \end{aligned} \tag{119}$$

Comparing the coefficients standing by the same powers of z and applying the relation

$$\sum_{l=0}^{r} \rho_l(\alpha) \alpha_i^{r-l} = \rho_r^i(\alpha), \tag{120}$$

we get

$$\begin{aligned}\rho_r(\lambda) = \rho_r(\alpha) + \frac{1}{2} \sum_{k=1}^{N} \rho_{r-1}^{k}(\alpha) q_k^2 &= \rho_r(\alpha) + \frac{1}{2} \sum_{k=1}^{N} \sum_{l=0}^{r-1} \rho_l(\alpha) \alpha_k^{r-l-1} q_k^2 \\ &= \rho_r(\alpha) + \frac{1}{2} \sum_{l=0}^{r-1} \rho_l(\alpha) (q, A^{r-l-1} q), \end{aligned} \tag{121}$$

that is, relation (116).

The basic example here is the Garnier system, with the potential $V^{(2)}(q) = \frac{1}{4}(q,q)^2 + \frac{1}{2}(q,Aq)$. Its bi-Hamiltonian formulation with the bi-Hamiltonian infinite chain is the special case of the system (106) and the chain (108), (109). This potential is a member of an infinite family of permutationally symmetric potentials separable in generalized elliptic coordinates that were found and characterized recursively in Ref. 20. This family contains, for instance, $V^{(3)}(q) = \frac{1}{8}(q,q)^3 - \frac{1}{2}(q,q)(q,Aq) + \frac{1}{2}(q,A^2q)$ and $V^{(-1)}(q) = [1 - \frac{1}{2}(q,A^{-1}q)]^{-1}$. All these potentials come up as constraint spectral problems of the coupled KdV systems and their bi-Hamiltonian formulation has been initially derived in Ref. 5. Nevertheless, all potentials, not necessarily related to the coupled KdV hierarchies, which have the bi-Hamiltonian formulation given in (106), are separable in generalized elliptic coordinates (117).

[1] M. Antonowicz, A. Fordy, and S. Wojciechowski, ''Integrable stationary flows: Miura maps and bi-Hamiltonian structures,'' Phys. Lett. A **124**, 143 (1987).
[2] M. Antonowicz and S. Rauch-Wojciechowski, ''Constrained flows of integrable PDE's and bi-Hamiltonian structure of Garnier system,'' Phys. Lett. A **147**, 455 (1990).
[3] M. Antonowicz and S. Rauch-Wojciechowski, ''Restricted flows of soliton hierarchies: Coupled KdV and Harry Dym case,'' J. Phys. A **24**, 5043 (1991).
[4] M. Antonowicz and S. Rauch-Wojciechowski, ''Bi-Hamiltonian formulation of the Henon–Heiles system and its multidimensional extension,'' Phys. Lett. A **163**, 167 (1992).
[5] M. Antonowicz and S. Rauch-Wojciechowski, ''How to construct finite-dimensional bi-Hamiltonian systems from soliton equations: Jacobi integrable potentials,'' J. Math. Phys. **33**, 2115 (1992).
[6] M. Błaszak and S. Wojciechowski, ''Bi-Hamiltonian dynamical systems related to low-dimensional Lie algebras,'' Physica A **155**, 545 (1989).

[7] M. Błaszak and P. Basarab-Horwath, ''Bi-Hamiltonian formulation of a finite dimensional integrable system reduced from a Lax hierarchy of the KdV,'' Phys. Lett. A **171**, 45 (1992).
[8] M. Błaszak and S. Rauch-Wojciechowski, ''Newton representation of nonlinear ordinary differential equations,'' Physica A **197**, 191 (1993).
[9] M. Błaszak, ''Miura map and bi-Hamiltonian formulation for restricted flows of the KdV hierarchy,'' J. Phys. A **26**, 5985 (1993).
[10] M. Błaszak and S. Rauch-Wojciechowski, ''A generalized Henon–Heiles system and related integrable Newton equations,'' J. Math. Phys. **35**, 1693 (1994).
[11] M. Błaszak, ''Newton representation for stationary flows of some class of nonlinear dynamical systems,'' Physica A **215**, 201 (1995).
[12] S. Rauch-Wojciechowski, K. Marciniak, and M. Błaszak, ''Two Newton decompositions of stationary flows of KdV and Harry Dym hierarchies,'' Physica A **233**, 307 (1996).
[13] R. Brouzet, R. Caboz, J. Rabenivo, and V. Ravoson, ''Two degrees of freedom quasi-bi-Hamiltonian systems,'' J. Phys. A **29**, 2069 (1996).
[14] C. Morosi and G. Tondo, ''Quasi-bi-Hamiltonian systems and separability,'' J. Phys. A **30**, 2799 (1997).
[15] H. M. M. Ten Eikelder, ''Symmetries of the massive Thiring model,'' J. Math. Phys. **27**, 1404 (1986).
[16] S. Rauch-Wojciechowski, ''Newton representation for stationary flows of the KdV hierarchy,'' Phys. Lett. A **170**, 91 (1992).
[17] S. Rauch-Wojciechowski, ''A bi-Hamiltonian formulation for separable potentials and its application to the Kepler problem and the Euler problem of two centers of gravitation,'' Phys. Lett. A **160**, 149 (1991).
[18] H. M. M. Ten Eikelder, ''On the local structure of recursion operators,'' Proc. K. Ned. Akad. A **89**, 386 (1986).
[19] I. Marshall and S. Wojciechowski, ''When is a Hamiltonian system separable,'' J. Math. Phys. **29**, 1338 (1988).
[20] S. Wojciechowski, in *Proceedings of the SMS on Systemes Dynamiques Nonlineaires*, edited by P. Winternitz (Press Universite Montreal, Montreal, 1986), p. 294.

An application of Lax's formalism to a class of curvilinear axially symmetric coordinates in three-dimensional space

Marcin Cerkaski
Henryk Niewodniczanski Institute of Nuclear Physics, Department of Theoretical Physics, Ul. Radzikowskiego 152, 31-342, Kraków, Poland

(Received 20 May 1997; accepted for publication 2 February 1998)

Solutions of a nonlinear system of differential partial equations in two-dimensional space are studied. We prove that the considered system is fully integrable and that its solutions determine an infinite class of mappings of the space of curvilinear axially symmetric coordinates onto the space of cylindric coordinates. The corresponding family of reverse transformations is found in a closed form. An infinite number of the first integrals of the nonlinear system is obtained. Symmetries of a class of electrostatic systems associated with the considered family of curvilinear coordinates are discussed. © *1998 American Institute of Physics.*
[S0022-2488(98)01305-X]

I. INTRODUCTION

The theory of the solvable nonlinear partial differential equations have been initiated by finding analytic solutions of the Korteweg–de Vries's equation.[1] General tools for the investigation of solvable nonlinear systems have been found.[2–4] Actually, there is a rich literature discussing a large number of interesting examples of solvable nonlinear partial equations and presenting the application of these systems to studying solitons dynamics and the theory of instantons.[3–9]

The main aim of this work is the investigation of the possibility of applying Lax's formalism to the study of a space of the axially symmetric surfaces in the three-dimensional space. In particular, we want to answer the following two questions: (a) Does there exist the Poisson action for which we can find a complete system of mutually commuting functionals? (b) Is it possible to distinguish a class of the curvilinear coordinates that are a geometric generalization of the spherical ones?

The solution of these problems appears to be related with the discussion of the following equation:

$$\partial_q{}^2 \ln \rho^2 + \partial_c{}^2 \rho^2 = 0, \tag{1.1}$$

where ρ is the real function of coordinates $(q,c) \in M$:

$$M = \{(q,c), \infty > q \geqslant 0, |c| \leqslant 1\}. \tag{1.2}$$

Space S_ρ of the solution of Eq. (1.1) that we will investigate is defined as follows:

$$\rho \in S_\rho: \ \rho^2(q,c) = \frac{1-c^2}{q^2}\, w(q,c), \quad w(q,c) = 1 + \sum_{n=1}^{\infty} w_n(c) q^n, \tag{1.3}$$

where functions $w_n(c)$ are polynomials. Equation (1.1) formulates the condition of integrability for differential form dz given by the relation

$$dz(q,c) = \rho \rho_{,c}\, dq - (\rho_{,q}/\rho) dc. \tag{1.4}$$

Denoting $\partial_q f = \dot{f}$ and $\partial_c f = f'$, we immediately see that the pair of functions z, ρ results from the system of equations

 © 1998 American Institute of Physics

$$\dot{z}=\rho\rho', \quad \dot{\rho}=-\rho z'. \tag{1.5}$$

An essential point of the discussion of solution S: $\rho\in S_\rho$ is based on the Lax's theory, and it is presented in Sec. III. The result of this section proves that system (1.5) can be presented as an infinite-dimensional Hamiltonian system for which the pair of functions $(z(c),\rho(c))$ plays the role of canonical coordinates while variable q is the evolution parameter giving a mapping $q\mapsto(z(c,q),\rho(c,q))$. This Hamiltonian system is fully integrable. We will prove this by constructing an infinite set of commuting functionals $H_n:(z,\rho)=G\mapsto J_n=J_n(G)\in R$, $n=1,2,...$, $J=(J_1,J_2,\ldots)$.

One of the possible interpretations of functions z,ρ and coordinates q,c is shortly discussed below and will be studied in Sec. II. In Sec. II we will show that interpretation results from the considering classes of axially symmetric curvilinear coordinates in R^3 space. We will call them q coordinates and let $\mathbf{q}_f=(q_f,c_f,\varphi)$ be some elements of this class. The space of q coordinates is obtained by considering the following conditions:

$$\mathbf{q}_f\in N=M\times(0,2\pi), \tag{1.6a}$$

$$*dq_f=dc_f\wedge d\varphi, \tag{1.6b}$$

where $*$ is the Hodge isomorphism (see Sec. II). For all q coordinates the metric tensor g and the volume element $d\mu(q,c,\varphi)$ have a very specific simple form:

$$g(q,c,\varphi)=\mathrm{diag}(g_q,g_q/\rho^2,\rho^2), \quad g_q(q,c)=\rho_{,q}^2+(\rho\rho_{,c})^2, \tag{1.7}$$

$$d\mu(q,c,\varphi)=g_q(q,c)dqdcd\varphi, \tag{1.8}$$

where functions $\rho(q,c)\in\overline{S_\rho}\subset S_\rho$, and for simplicity, label f has been omitted.

Simultaneously, if coordinates z,ρ,φ are interpreted as standard cylindric coordinates then elements $(z(q,c),\rho(q,c))$ define the coordinates transformation G,

$$G: \quad N\ni(q,c,\varphi)\mapsto(z,\rho,\varphi)\in D, \tag{1.9}$$

where D denotes the space of cylindrical coordinates. The simplest example of q coordinates is obtained through substitution:

$$q_e=r^{-1}=\frac{1}{\sqrt{\rho^2+z^2}}, \quad c_e=z/r, \quad \varphi=\arctan(y/x). \tag{1.10}$$

Here, x,y,z are Cartesian coordinates. The latter coordinates are called nondeformed q coordinates and they are an elementary modification of spherical coordinates. For all other q coordinates the solution of equation $q_f=\mathrm{const}$ leads to families of an axially symmetric closed surface prolate—deformed along the z axis.

The problem of consistency of condition (1.6a) with condition (1.6b) is not trivial and it is studied in Sec. II. This consistency is the result of an explicit construction of an infinite set of mappings F_f, $f\in\mathscr{F}$, of space D onto space N:

$$F_f: \quad D\ni(z,\rho,\varphi)\mapsto(q,c,\varphi)\in N. \tag{1.11}$$

Specific properties of the family of q coordinates are revealed under their application to studying the Poisson equation.

Let $\Psi(z,\rho)$ be the electric potential in the cylindrical frame, (f,ϕ)—the subspace of electrostatic states where $\phi(\mathbf{q})\equiv\phi(q)$ is the potential in the q frame of coordinates, while f determines the transformation F_f of $D\mapsto N$, i.e., let us have $\Psi=\phi\circ F_f$. Considering the integral of Coulomb energy for state (f,ϕ) and performing the transformation of coordinates to q coordinates, one finds

$$\mathscr{E}(f,\phi)=\frac{1}{8\pi}\int_{R^3}d^3x\mathbf{E}^2(\mathbf{x})=\frac{1}{8\pi}\int_N d\mu(q,c,\varphi)g^{qq}[\phi'(q)]^2=\frac{1}{2}\int_0^\infty dq[\phi'(q)]^2. \tag{1.12}$$

For all (f,ϕ) states this integral is factorized into the ''radial part''×''angular part.'' The angular factor is equal to 4π. Coulomb energy depends on the choice of function ϕ only. Thus, Coulomb energy is an invariant of transformations: $(f,\phi)\mapsto(\overline{f},\phi)$.

There is one more interesting and rather unexpected symmetry of (f,ϕ) states that stems from the hidden symmetry of equations (1.5). Consider spherical moments of charge density $\varrho_f(\mathbf{q})$. From the Poisson equation and Eq. (1.7), we get

$$\varrho_f(\mathbf{q})=-\frac{1}{4\pi\sqrt{|g|}}\frac{\partial}{\partial q^\alpha}g^{\alpha\beta}\sqrt{|g|}\frac{\partial}{\partial q^\beta}\phi_f(q)=-\frac{1}{4\pi g_q}\phi_f''(q). \tag{1.13}$$

Let $\mathscr{Q}_l(f,\phi)$ denote moments associated with spherical functions: $Q_l=r^lP_l(\cos\theta)$, and $\overline{h_l}(z,\rho)$ be homogeneous polynomials of degree n obtained by the transformation of coordinates:

$$\overline{h_l}(z,\rho)=r^l(z,\rho)P_l(z/r(z,\rho)). \tag{1.14}$$

Then, in the frame of cylindrical coordinates we have

$$\mathscr{Q}_l(f,\phi)=2\pi\int_{-\infty}^\infty\int_0^\infty dzd\rho\rho\overline{h_l}(z,\rho)(\varrho_f\circ F_f)(z,\rho), \tag{1.15}$$

and changing the latter into $\mathbf{q}_f$ [see Eqs. (1.8) and (1.13)], we get

$$\mathscr{Q}_l(f,\phi)=-\frac{1}{2}\int_0^\infty dq\,\phi_f''(q)\times H_l(G_f), \tag{1.16}$$

where

$$H_l(G_f)=\int_{-1}^1 dc(\overline{h_l}\circ G_f)(q,c),\quad G_f\in\{G\}:G_f\circ F_f=\mathrm{id}. \tag{1.17}$$

In Sec. III we will prove that $H_l(G_f)$ are just invariant functionals of the Lax's system (1.5); thus they are independent of coordinate q. Hence, integrals $\mathscr{Q}_l(\varrho_f)$ are also factorized into the ''radial part''×''angular part.''

In the same section we find too that $H_0(G_f)\equiv 2$; hence the total charge Q_0 is equal to $\mathscr{Q}_0(\phi)=\phi'(0)-\phi'(\infty)$, and we can rewrite Eq. (1.16) in the form

$$\mathscr{Q}_l(f,\phi)=\tfrac{1}{2}H_l(G_f)Q_0(\phi). \tag{1.18}$$

As we can see, only constant Q_0 depends on potential functions ϕ, so if $Q_0(\phi)=Q_0(\overline{\phi})$ then $\mathscr{Q}_l$ are invariants of the transformations $(f,\phi)\mapsto(f,\overline{\phi})$.

In the Appendix we discuss solutions for polynomials $w_n(c)$ [see Eq. (1.3)], and we find the recurrence algorithm for obtaining them. Coefficients of polynomials $w_n(c)$ are functions of some constants $I_2,I_3,\ldots,I_n$. We present explicit expressions for a few lowest polynomials and we find values of functionals H_n for a few lowest values of parameter n.

II. THE CLASS OF q COORDINATES

The Hodge isomorphism * in the three-dimensional space associates the space of one-forms with the space of two-forms. For basis one-forms dq^i, we have

$$*dq^i=\frac{1}{2}\sqrt{|g|}(g)^{ij_1}\epsilon_{j_1j_2j_3}dq^{j_2}\wedge dq^{j_3}, \tag{2.1}$$

where $g=g(\mathbf{q})$ is a matrix of the metric tensor and $(g)^{ij}$ are elements of the reverse matrix $(g)^{-1}$. In the frame of cylindrical coordinates the metric tensor $g_{\mathrm{cyl}}(z,\rho,\varphi)=\mathrm{diag}(1,1,\rho^2)$; hence, for the squared length element we get

$$ds^2(dz,d\rho,d\varphi)=dz^2+d\rho^2+\rho^2 d\varphi^2, \tag{2.2}$$

Considering transformations G of q coordinates into cylindrical coordinates, we obtain

$$(G^*ds^2)(dq,dc,d\varphi)=g_{ij}(\mathbf{q})dq^i dq^j=(z_{,q}^2+\rho_{,q}^2)dq^2+2(z_{,q}z_{,c}+\rho_{,q},\rho_{,c})dqdc + (z_{,c}^2+\rho_{,c}^2)dc^2+\rho^2 d\varphi^2, \tag{2.3}$$

Applying the last formula to Eq. (2.1), we find

$$^*dq=\frac{\rho}{|z_{,q}\rho_{,c}-z_{,c}\rho_{,q}|}\,[(z_{,c}^2+\rho_{,c}^2)dq^{c\varphi}-(z_{,q}z_{,c}+\rho_{,q}\rho_{,c})dq^{\varphi q})], \tag{2.4}$$

where, for brevity, we denoted $dq^i\wedge dq^j=dq^{ij}$. Comparing this result with condition in Eq. (1.6b), we get

$$\frac{\rho(z_{,c}^2+\rho_{,c}^2)}{|z_{,q}\rho_{,c}-z_{,c}\rho_{,q}|}=1, \quad z_{,q}z_{,c}+\rho_{,q},\rho_{,c}=0, \tag{2.5}$$

and after eliminating function $z_{,q}$ from the latter equation we arrive at

$$0=\rho_{,q}\rho_{,c}^2/\rho+\rho_{,c}^2 z_{,c}+(\rho_{,q}/\rho)z_{,c}^2+z_{,c}^3=(z_{,c}+\rho_{,q}/\rho)(z_{,c}^2+\rho_{,c}^2). \tag{2.6}$$

Hence, because the second factor is positive we obtain the pair of Eqs. (1.5): $\rho_{,q}=-\rho z_{,c}$, $z_{,q}=\rho\rho_{,c}$. This proves immediately the validity of Eqs. (1.1,1.4) and (1.7). Considering reverse transformations F described by functions $q(z,\rho)$ and $c(z,\rho)$ and the rules of composition for matrices describing the tangent transformations dF and dG, we get

$$\frac{q_{,z}}{c_{,\rho}}=\frac{\rho_{,c}}{z_{,q}}=\frac{1}{\rho}, \quad \frac{q_{,\rho}}{c_{,z}}=\frac{z_{,c}}{\rho_{,q}}=-\frac{1}{\rho}\,; \tag{2.7}$$

hence,

$$\rho q_{,z}-c_{,\rho}=0, \quad \rho q_{,\rho}+c_{,z}=0, \tag{2.8}$$

which lead to the system of differential equations,

$$dq(z,\rho)=\frac{1}{\rho}\,(c_{,\rho}dz-c_{,z}d\rho), \tag{2.9}$$

$$dc(z,\rho)=-\rho(q_{,\rho}dz-q_{,z}d\rho). \tag{2.10}$$

Let us consider the conditions of integrability of differentials dz and dq. They read as

$$0=\frac{1}{\rho}\,(c_{,zz}+c_{,\rho\rho}-c_{,\rho}/\rho), \tag{2.11a}$$

$$0=\rho(q_{,zz}+q_{,\rho\rho}+q_{,\rho}/\rho). \tag{2.11b}$$

Equation (2.11b) is fulfilled identically for $\rho=0$. For the $\rho>0$ function q has to satisfy the three-dimensional cylindrical Laplace equation. Thus, function $q(z,\rho)$ is electrical potential that we obtain by considering a charged string placed along the z axis: $\varrho_e(x,y,z)=\delta(x)\delta(y)f(z)$. It enables us to write a complete solution of system (2.9),(2.10) in the following form:

$$q_f(z,\rho)=q(z,\rho)=\int_{-\infty}^{\infty} dtf(t)\ \frac{1}{\sqrt{(z-t)^2+\rho^2}}, \tag{2.12}$$

$$c_f(z,\rho)=c(z,\rho)=\int_{-\infty}^{\infty} dtf(t)\ \frac{z-t}{\sqrt{(z-t)^2+\rho^2}}. \tag{2.13}$$

Let us consider functions (q_f,ρ_f) as elements of a family of transformation F_f, $f\in\mathscr{F}$, where $\mathscr{F}$ is define as follows: (a) $f>0$, (b) $\int_R f=1$, and (c) supp $f=\cup_{i=1}^{p}[a_i,b_i]$, $a_i<b_i<a_{i+1}$, (support).

Introducing the notation $Z(m)=\{(z,\rho),z\in m,\rho=0\}$, $M(h,g)=\{(q,c),q\in h,c\in g\}$ and taking into account the formula

$$c(z,0)=\lim_{\rho\to 0} c(z,\rho)=\int_R dtf(l)\mathrm{sgn}(z-t)=-1+2\int_{-\infty}^{z} dtf(t), \tag{2.14}$$

we find that F_f are one-to-one mappings of the following sets:

$$D_1=Z(\{z,z<a_1\})\leftrightarrow M(R^+,-1)=M_1, \tag{2.15}$$

$$D_2=Z(\text{supp } f)\leftrightarrow M(\infty,[-1,1])=M_2, \tag{2.16}$$

$$D_3=Z(\{z,z>b_p\})\leftrightarrow M(R^+,-1)=M_3, \tag{2.17}$$

$$D_4=\lim_{r\to\infty}\{(z,\rho),z^2+\rho^2=r^2\}\leftrightarrow M(0,[-1,1])=M_4. \tag{2.18}$$

For Jacobian $J(z,\rho)$ of transformations F_f, we find

$$J(z,\rho)=q_{,z}c_{,\rho}-q_{,\rho}c_{,z}=\rho(q_{,z}^2+q_{,\rho}^2),\quad J>0 \quad \text{if } \rho>0; \tag{2.19}$$

hence, if $p=1$ then F_f: $\partial D\mapsto\partial M$ and F_f are one-to-one transformations of the space of cylindrical coordinates D onto M [see Eq. (1.2)].

Let us consider the case $p>1$. Then, in each interval $b_i\leqslant a_{i+1}$ there exists a single point z_i: $q_{,z}(z_i,0)=q_{,\rho}(z_i,0)=0$, $F_f(z_i,0)=(q_i,c_i)$. Using formula (2.14) we get

$$F_f(Z([b_i,z_i]))=F_f(Z([z_i,a_{i+1}])=M([q_i,\infty],c_i)=N_i, \tag{2.20}$$

which shows that functions $z(q,c)$ are discountinous at points of one-dimensional set $N=\cup_{i=1}^{p-1}N_i$. All other conclusions about transformations F_f and F_f^{-1} remain unchanged, hence points (a)–(c) and Eqs. (2.12) and (2.13) define a class of coordinates $\mathbf{q}_f$ that were postulated in (1.6a) and (1.6b).

III. LAX'S SYSTEM

The class of q coordinates studied in the previous section forms an infinite family. Below we will discuss a family of reverse transformations G: $(q,c)\to(z,\rho)$. Functions (z,ρ) describing these transformations obey a pair of nonlinear partial equations (1.5). Because set $\{G\}$ must also consist of an infinite family we can expect that these equations are integrable.

Below we develop Lax's formalism for discussing this question. The basic point of this idea is determining a Poisson algebra for functions z,ρ and then obtaining Eq. (1.5) as a pair of Hamiltonian equations, i.e., let us have

$$\dot{z}=\{z,H\},\quad \dot{\rho}=\{\rho,H\}, \tag{3.1}$$

and let Poisson brackets be defined by the following relations:

$$\{z(c),z(\overline{c})\}=\{\rho(c),\rho(\overline{c})\}=0, \tag{3.2}$$

$$\{\rho(c),z(\overline{c})\}=\rho(c)\,\delta'(\overline{c}-c), \tag{3.3a}$$

$$\{z(c),\rho(\overline{c})\}=\rho(\overline{c})\,\delta'(\overline{c}-c), \tag{3.3b}$$

where $\delta'(x)\equiv-\delta'(-x)$ denotes derivative of Dirac distribution. The introduced Poisson brackets obey Jacobi conditions: $W(A,B,C)=0$ for arbitrary elements A,B,C, where

$$W(A,B,C)=\{A,\{B,C\}\}+\{B,\{C,A\}\}+\{C,\{A,B\}\}. \tag{3.4}$$

Indeed, we have

$$W(z(a),z(b),\rho(c))=\rho(c)\,\delta'(b-c)[\,\delta'(a-c)+\delta'(c-a)\,]=0, \tag{3.5}$$

while all remaining Jacobi conditions are fulfilled trivially. The physics interpretation of Eqs. (3.1) is to be found through considering the space of curves (or axially symmetric surfaces) $\{\Gamma\}$, $\Gamma=\{[-1,1]\ni c\mapsto(z(c),\rho(c))\}$. For given $\Gamma_0=\Gamma(q_0)$ they determine the evolution of this curve: $q\mapsto\Gamma(q)$.

Theorem 1: (a) *Let* $\{\overline{\Gamma}\}$ *denote subspace of a set* $\{\Gamma\}$ *of curves* Γ: $[-1,1]\in c\to(z(c),\rho(c))$ *for which* $\rho(\pm1)=0$, $\mathscr{H}=\{H_n(\Gamma)\}$ *is a set of functionals:*

$$H_l(\Gamma)=\int_{-1}^{1}dc(h_l\circ\Gamma)(c),\quad l=0,1,2,..., \tag{3.6}$$

$$h_l(z,\rho)=\sum_{m=0}^{[l/2]}h_{lm}z^{l-2m}\rho^{2m},\quad h_{lm}=\frac{(-4)^{-m}l!}{(l-2m)!m!^2}, \tag{3.7}$$

and let $A=(\mathscr{H},\{,\})$ *denote the algebra of a group of transformations of* $\{\overline{\Gamma}\}$.

(a) $\{\overline{\Gamma}\}$ *is an A-invariant subspace of* $\{\Gamma\}$; (b) *algebra A is Abelian on* $\{\overline{\Gamma}\}$; (c) *Eqs. (1.5) and (3.1) coincide if*

$$H=\frac{1}{2}H_2=\frac{1}{4}\int_{-1}^{1}d\overline{c}[2z^2(\overline{c})-\rho^2(\overline{c})]; \tag{3.8}$$

(d) *for* $\Gamma_q=G(q,c)$ *being a solution of Eqs. (1.5) integrals* $H_l(\Gamma_q)$ *do not depend on q.*

Proof: Functions $h_n(z,\rho)$ obey the following relations:

$$h_{n,z}=nh_{n-1},\quad \rho h_{n,\rho}=n(h_n-zh_{n-1}), \tag{3.9}$$

$$(n+1)h_{n+1}-(2n+1)zh_n+n(z^2+\rho^2)h_{n-1}=0. \tag{3.10}$$

For functionals

$$A=\int_{-1}^{1}dca(z,\rho), \tag{3.11}$$

Poisson brackets $\{x^i(c),A\}$, $(x^1=z,\ x^2=\rho)$, are obtained from the textbook expression

$$\{x^i(c),A\}=\sum_j\int_{-1}^{1}d\overline{c}a_{,j}(\mathbf{x}(\overline{c}))\{x^i(c),x^j(\overline{c})\}. \tag{3.12}$$

Hence,

$$\{z(c),H_k\}=\int_{-1}^{1}d\overline{c}.\mathscr{H}_{k,\rho}(\overline{c})\rho(\overline{c})\,\delta'(\overline{c}-c)=-(d_c\rho.\mathscr{H}_{k,\rho})(c), \tag{3.13}$$

$$\{\rho(c),H_k\}=\rho(c)\int_{-1}^{1}d\overline{c}.\mathscr{H}_{k,z}(\overline{c})\,\delta'(\overline{c}-c)=-(\rho d_c.\mathscr{H}_{k,z})(c), \tag{3.14}$$

where $d_c = z' d_z + \rho' d_\rho$ and $\mathscr{H}_{l,i}(c) = (h_{l,i} \circ \Gamma)(c)$. While arriving at the last formulas, we used integration by parts and the fact that function ρ vanishes at the integration boundary. Using relations (3.9) we can present this result in the form

$$\begin{pmatrix} \{z, H_k\} \\ \{\rho, H_k\} \end{pmatrix} = -k\rho \begin{pmatrix} h_{k-1,\rho} & -h_{k-1,z} \\ h_{k-1,z} & h_{k-1,\rho} \end{pmatrix} \begin{pmatrix} z' \\ \rho' \end{pmatrix}, \tag{3.15}$$

from which we immediately see the equivalence of Eq. (1.5) and Eqs. (3.1) and (3.8). Next, we have

$$\{h_l, H_k\} = h_{l,z}\{z(c), H_k\} + h_{l,\rho}\{\rho(c), H_k\} = -(k-1)kl(A_{lk} z' + B_{lk} \rho'), \tag{3.16}$$

where

$$A_{lk} = h_{k-1} h_{l-1} - 2h_{k-2} h_{l-1} z + h_{k-2} h_{l-1}, \tag{3.17}$$

$$B_{lk} = \frac{1}{\rho} [h_{k-1} h_l - z(h_{k-1} h_{l-1} + h_{k-2} h_l) + (z^2 - \rho^2) h_{k-2} h_{l-1}], \tag{3.18}$$

and where A_{lk} as well as B_{lk} are polynomials of variables z, ρ. If we define

$$\Omega_{lk} = \rho(\partial_\rho A_{lk} - \partial_z B_{lk}), \tag{3.19}$$

then, with the help of formulas (3.9), we arrive at

$$\begin{aligned} \Omega_{lk} &= (k+1) h_{k+1} h_l + (l+1) h_k h_{l+1} - 2(k+l+1) z h_k h_l + (z^2 + \rho^2)(l h_k h_{l-1} + k h_{k-1} h_l) \\ &= h_k[(l+1) h_{l+1} - (2l+1) z h_l + l(z^2 + \rho^2) h_{l-1}] \\ &\quad + h_l[(k+1) h_{k+1} - (2k+1) z h_k + k(z^2 + \rho^2) h_{k-1}] \\ &= 0. \end{aligned}$$

In the last step we used Eq. (3.10). This result proves that the two terms on the right-hand side of formula (3.16) give derivative d_c, i.e. we have

$$\{h_l, H_k\} = \frac{d}{dc} U_{lk}(z, \rho), \tag{3.20}$$

where functions U_{lk} are homogeneous polynomials of variables z, ρ obtained from the formula

$$U_{lk}(z, \rho) = -(k-1)kl\left(\int_a^z ds A_{lk}(s, \rho) + \int_0^\rho dt B_{lk}(a, t) \right), \tag{3.21}$$

We have $h_l(z,0) = z^l$, $A_{lk}(z,0) = 0$; hence $U_{lk}(z,0) = 0$. Integrating formula (3.20), we get

$$\{H_l, H_k\} = \int_{-1}^{1} dc \{h_l, H_k\} = U_{lk}(z, \rho) \Big|_{c=-1}^{c=1} = U_{lk}(z,0)|_{c=-1}^{c=1} = 0, \tag{3.22}$$

which completes the proof point (c). Then we get immediately point (d): $2\dot{H}_m = \{H_m, H_2\} = 0$. The second equation in formulas (3.15) has the following form: $\dot{\rho} = \rho F_n$. Hence, if $\Gamma(q_0) \in \{\Gamma\}$ then all derivatives $(d_q^p \rho)(\pm 1)$ must vanish and for all q: $\Gamma(q) \in \{\Gamma\}$; thus the point (a) is fulfilled too.

Below we present the result of explicit calculation of the values of integrals H_n. They are obtained using the solution given in the Appendix. H_n are functions of a set of constants I. Let H_n' be the value of H_n for $I_1 = 0$. Then we have

$$H_0' = 2, \quad H_1' = 0, \quad H_2' = I_2, \quad H_3' = \tfrac{3}{2} I_3, \quad H_4' = I_2^2 + 2I_4, \quad H_5' = 5I_2 I_3 + \tfrac{5}{2} I_5,$$

$$H_6'=I_2^3+\frac{9}{2}I_3^2+6I_2I_4+3I_6, \quad H_7'=\frac{21}{2}I_2^2I_3+\frac{63}{5}I_3I_4+\frac{21}{2}I_2I_5+\frac{7}{2}I_7,$$

and

$$H_n(I)=\sum_{k=0}^{n}\binom{n}{k}H'_{n-k}(I_1)^k. \tag{3.23}$$

The structure of these relations is very simple: $H_l=\gamma_k I_k+\psi_k(I_1,\ldots,I_{k-1})$. Mappings $H_l\to c^l H_l$, $I_l\to c^l I_l$ are symmetry transformations of Eq. (3.23). The one nontrivial observation results from the above numerical calculus: $\gamma_l=l/2\neq 0$. The nonvanishing of coefficients γ_l leads immediately to one-to-one correspondence between the elements of set I and between the elements of set of constants $H=\{H_1,H_2,\ldots\}$. Hence, H is the complete set of the first integrals of equations (1.5) in respect to the considered class of solutions.

The proof of this fact for all n results from noticing that in Eq. (A8) only the first term in round parentheses gives contributions to function ρ, which are proportional to I_n, and that the appropriate numerical coefficients are known. However, details of this proof will not be discussed here.

Polynomials $h_n(z,\rho)$ are homogeneous functions of degree n. The recurrence relation (3.10) shows that $h_n(x,\sqrt{1-x^2})=P_n(x)$; hence h_n and polynomials h_n appearing in Eq. (1.14) coincide. This conclusion proves the validity of formulas (1.14)–(1.18).

Formula (1.18) shows that the spherical moments $\mathcal{O}(f,\phi)$ are invariants of transformations $(f,\phi)\mapsto(f,\tilde{\phi})$: $Q_0(\phi)=Q_0(\tilde{\phi})$. Taking into account the above fact and noticing that for $\tilde{\phi}(q)=q$ we have $\rho(f,\tilde{\phi})(z,\rho)=f(z)\delta(\rho)/(2\pi\rho)$, we can change $\phi\to Q_0(\phi)q$ and calculate moments $\mathcal{O}_l$ in an elementary way in the frame of cylindrical coordinates. We get

$$\mathcal{O}_l(f,\phi)=\mathcal{O}_l(f,Q_0(\phi)q)=f_l(f)Q_0(\phi), \quad f_l(f)=\int_{-\infty}^{\infty}dt\,t^l f(t). \tag{3.24}$$

Comparing this result to formula (1.18), we find

$$f_l=\tfrac{1}{2}H_l(I), \quad l=0,1,\ldots. \tag{3.25}$$

The last formula enables us to express parameters of the solution for functions z,ρ by moments f_n describing transformation F_f. This means that for all q-coordinates $\mathbf{q}_f(z,\rho,\varphi)$ we can find reverse transformations G_f: $G_f\circ F_f=\mathrm{id}$, and we can calculate elements of the metric tensor $g(\mathbf{q}_f)$ [see Eq. (1.7)].

ACKNOWLEDGMENT

The author wishes to express his thanks for financial support by KBN Grant No. 2 PO3B 14010.

APPENDIX: DERIVATION OF POLYNOMIALS *w*

We will study a class of solutions of the system (1.5) for which function ρ fulfills the condition (1.3). Substituting Eq. (1.3) to Eq. (1.1) and changing variables (q,c) into (x,y), we obtain the equation for function w,

$$x^2(ww_{,xx}-w_{,x}^2)+w^2[-2(w-1)+(1-y^2)w_{,yy}-4yw_{,y}]=0. \tag{A1}$$

Searching w in the form of a power series,

$$w(x,y)=\sum_{n=0}^{\infty}w_n(y)x^n, \tag{A2}$$

one gets the following system for function coefficients $w_n(y)$:

$$\sum_p [p(2p-n-1)+2]w_p w_{n-p} - \sum_{p,r} w_p w_r (2w_{n-p-r} + 4y w'_{n-p-r} - (1-y^2) w''_{n-p-r}) = 0, \tag{A3}$$

valid for $n=0,1,2,\ldots$. It can be rewritten in the form

$$w_0^2\{(1-y^2)w''_n - 4y w'_n\} + \gamma_n w_n + d_n(y) = 0, \tag{A4}$$

where

$$d_0(y)=0, \quad \gamma_0 = 2w_0(1-w_0), \quad \gamma_n = [n(n-1)+4]w_0 - 6w_0^2, \quad n \geqslant 1, \tag{A5}$$

$$d_n(y) = \sum_{i=1}^{n-1} \{i(2i-n-1)w_{n-i}w_i - [d_i/w_0 + i(i-1)w_i]g_{n-i}/w_0\} + \gamma_0 \overline{d}_n(y), \tag{A6}$$

$$g_k(y) = \sum_{i=0}^{k} w_{k-i} w_i .$$

If we put $w_0(y)=1$, $w_1(y)=0$ then $\gamma_0=0$, $\gamma_n=(n-2)(n+1)$ and functions $w_n(y)$, $d_n(y)$ are polynomials of parity $(-1)^n$,

$$w_n(y) = \sum_{k \in Z_n} w_{nk} y^k, \quad d_n(y) = \sum_{k \in Z_{n-2}} d_{nk} y^k, \tag{A7}$$

where $Z_n = n-2, n-4, \ldots, 0, (1)$. We get coefficients w_{nk}, $n \geqslant 2$ with the help of Eq. (A4),

$$w_{nk} = \frac{(n+k-1)!!}{k!(n-k-2)!!} \left((-1)^{(n-k)/2} I_n + \sum_{p=0,2,\ldots,}^{2[k/2]-2} (-1)^{p/2+1} \mathscr{N}^n_{kp} d_{nk-p-2} \right), \tag{A8}$$

$$\mathscr{N}^n_{kp} = \frac{(k-p-2)!(n-k+p-2)!!}{(n+k-2p-1)!!}. \tag{A9}$$

Here, $[n/m]$ is an integer quotation and $I=\{I_2, I_3, \ldots\}$ is an infinite set of integration constants. Because $\gamma_0=0$, Eqs. (A6) and (A8) determine the recurrence algorithm with respect to index n. Using it, one obtains

$$\begin{aligned} w(x,y) = 1 &- I_2 x^2 - 3I_3 y x^3 + [(-2^{-1}3\ 5 I_4 + 3I_{2^2})y^2 + 2^{-1}3 I_4]x^4 + [(-2^{-1}5\ 7 I_5 + 2\ 5 I_{32})y^3 \\ &+ 2^{-1}3\ 5 I_5 y]x^5 + [(2^{-1}3\ 5\ 7 I_{42} + 2^{-2}3^2 5 I_{3^2} - 2^{-3}3^2 5\ 7 I_6 - 3\ 5 I_{2^3})y^4 \\ &+ (-2^{-1}3^2 5 I_{42} + 2^{-2}3\ 5\ 7 I_6 + 5 I_{2^3})y^2 - 2^{-3}3\ 5\ I_6]x^6 + [(2^{-1}3^2 5^{-1}7\ 13 I_{43} \\ &+ 2^{-1}3^3 7 I_{52} - 2^{-3}3^2 7\ 11 I_7 - 3^2 7 I_{32^2})y^5 + (-2^{-1}3^2 7 I_{43} - 2^{-1}3\ 5\ 7 I_{52} + 2^{-2}3^2 5\ 7 I_7 \\ &+ 3\ 7 I_{32^2})y^3 - 2^{-3}3\ 5\ 7 I_7 y]x^7 + \cdots, \end{aligned} \tag{A10}$$

where $I_{p_1{}^{k_1},\ldots p_s{}^{k_s}}$ denote $\Pi_{i=1}^s I_{p_i}{}^{s_i}$. We find function $z(x,y)$ we by integrating of formula (1.4):

$$z(x,y) = I_1 + \frac{y}{x} + \frac{y}{2x} \sum_{n=2}^{\infty} \sum_{k \in Z_n} \frac{k+2}{n-1} (w_{nk+2} - w_{nk}) x^n y^k, \tag{A11}$$

[1] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).
[2] P. D. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[3] F. Calogero and A. Degasperis, *Spectral Transforms and Solitons: Tools to Solve and Investigate Nonlinear Evolution Equations* (North-Holland, Amsterdam, 1982).
[4] L. A. Takhtajan and L. D. Faddeew, *Hamiltonian Approach in Soliton Theory* (Science, Moscow, 1986) (in Russian).
[5] G. L. Lamb, *Elements of Soliton Theory* (Wiley, New York, 1980).
[6] R. Rajaraman, *Solitons and Instantons* (North-Holland, Amsterdam, 1982).

[7] A. N. Leznov and M. V. Saveliev, *Group Methods of Integration of Nonlinear Dynamical Systems* (Science, Moscow, 1985) (in Russian).
[8] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevski, *Solitons Theory: The Inverse Scattering Transform Method* (Science, Moscow, 1980).
[9] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).

The potential constraints of the complex (1+2)-dimensional soliton systems and related Hamiltonian equations

Deng-yuan Chen
Department of Mathematics, Shanghai University, Shanghai 201800, People's Republic of China

(Received 15 January 1997; accepted for publication 6 November 1997)

Potential complete constraints (symmetric and nonsymmetric) of the complex KP and MKP systems are derived by using the general theory for a nonlinear equation to be a Hamiltonian system. All the complex Hamiltonian equations corresponding to this kind of constraint are obtained, which contain the famous nonlinear Schrödinger equations and several new Hamiltonian equations. In addition, the classical Kaup–Newell system and the Heisenberg system are also led by a new reduction approach from the MKP system. © *1998 American Institute of Physics.* [S0022-2488(98)01406-6]

I. INTRODUCTION

It is well known that many $(1+1)$-dimensional soliton equations can arise as the nonlinearization of real linear problems of a $(1+2)$-dimensional integrable system. In the nonlinearized procedure, potential symmetric constraints play crucial roles.[1–4] For example, by restricting the potential u to the symmetry $\phi\psi$ of the Kadomtsev–Petviashvili (KP) system, the associated real linear problem for ϕ and its adjoint for ψ are reduced to the first two equations of the famous AKNS hierarchy. As a result, one can obtain solutions for the KP equation by solving the lower-dimensional constrained equations.

Recently, a new notion called the potential complete constraint (contains symmetric and nonsymmetric) of the $(1+2)$-dimensional integrable system was introduced by this author with N. S. Zhu and M. Zhu.[5] Based on the idea that the linear problem and its adjoint must be restricted to an infinite-dimensional real Hamilton equation, we derived the concrete expressions of the complete constraint of the KP system[5] and modified Kadomtesev–Petviashvili (MKP) system,[6] and obtained further some new Hamiltonian equations.

In this paper, we would like to extend the method of finding the complete constraint to a high-dimensional integrable system associated with a complex linear problem (named complex integrable system for short). To this end, we first need to establish the theory of the Hamiltonian equation on the linear space constructed by complex vector fields. Then, along the line provided in Ref. 5, we derive the potential complete constraints for the complex KP and MKP systems and the related Hamiltonian equation. Consequently, the famous nonlinear Schrödinger (NS) equations and several other new complex Hamiltonian equations are deduced as particular examples. Finally, by the new nonlinearized method of linear problem of the complex MKP system, we obtain the classical Kaup–Newell (KN) equations and Heisenberg (HB) equations and give their Hamiltonian structures.

The organization of the paper is as follows. In Sec. II we give some notions of the Gateaux derivative and the functional derivative and prove a fundamental theorem of the complex Hamiltonian equation. In Sec. III we deduce the potential constraints of the complex KP system and related Hamiltonian equations. (The detailed computation is given in the Appendix.) We also discuss the potential constraints of the complex MKP system by a particular reduction approach in Sec. IV.

II. GATEAUX DERIVATIVE, FUNCTIONAL DERIVATIVE, AND COMPLEX HAMILTONIAN EQUATION

Let $t\in R$, $x\in R^n$, $\phi=\phi(t,x)=(\phi_1,\phi_2,\ldots,\phi_s)^T$, and for any fixed $t,\phi_j=\phi_j(t,x)$ $(j=1,2,\ldots,s)$ belong to the space $S(R^n)$ of the rapidly decreasing complex-valued functions. For two

 © 1998 American Institute of Physics

series of non-negative integers $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)$ and $\beta=(\beta_1,\beta_2,\ldots,\beta_n)$, we write $D^\alpha=D_1^{\alpha_1}D_2^{\alpha_2}\cdots D_n^{\alpha_n}$, $\alpha=\Sigma_{j=1}^n\alpha_j$, and $I^\beta=I_1^{\beta_1}I_2^{\beta_2}\cdots I_n^{\beta_n}$, $\beta=\Sigma_{j=1}^n\beta_j$, where $D_j=\partial/\partial x_j$, $I_j=\frac{1}{2}(I_j^- -I_j^+)$, $I_j^-=\int_{-\infty}^{x_j}dx_j$, $I_j^+=\int_{x_j}^{+\infty}dx_j$, and $D_jI_j=I_jD_j=1$, $(j=1,2,...,n)$. Suppose $f=f(t,x,\phi)=(f_1,f_2,\ldots,f_m)^T$ is a complex smooth vector field with the components of $f_j=f_j(t,x,D\phi,\ldots,D^\alpha\phi,I\phi,\ldots,I^\beta\phi)$ $(j=1,2,...,m)$, which, together with all its partial derivatives, tends to be zero fast enough. $V^{(m)}$ denotes the complex linear space constructed by these vector fields. For two arbitrary vector fields f and g in $V^{(m)}$, we define their real inner product,[7]

$$\langle f,g\rangle=\frac{1}{2}\int_{-\infty}^{\infty}(f\cdot\overline{g}+\overline{f}\cdot g)dx, \tag{2.1}$$

where $f\cdot g=\Sigma_{j=1}^m f_jg_j$, and the overbar denotes the complex conjugate. Thus it can be seen that inner product (2.1) is commutative,

$$\langle f,g\rangle=\langle g,f\rangle. \tag{2.2}$$

Let $Q^{(m)}$ denote a set composed of all the linear operators $L=L(t,x,\phi)$ on $V^{(m)}$ into $V^{(m)}$ that are infinitely differentiable with respect to x and smooth with respect to ϕ.

Definition 1: Let θ be an operator in $Q^{(m)}$. If there exists another operator $J\in Q^{(m)}$ such that

$$\langle\theta f,g\rangle=\langle f,Jg\rangle, \tag{2.3}$$

for any $f,g\in V^{(m)}$, then J is named the conjugate operator of θ and denoted as θ^*. It is named a self-conjugate operator if $\theta^*=\theta$, while an anti-self-conjugate operator if $\theta^*=-\theta$.

Obviously, the differential operator D_j, integral operator I_j, and complex matrix

$$\sigma=\begin{pmatrix}-i & 0\\ 0 & i\end{pmatrix}\in Q^{(2)}, \tag{2.4}$$

are all anti-self-conjugate operators.

Definition 2: Let $f\in V^{(m)}$, $h\in V^{(s)}$, and ϵ is an arbitrary complex number; then

$$f'[h]=f'(\phi)[h]=\frac{d}{d\epsilon}f(\phi+\epsilon h)\big|_{\epsilon=0} \tag{2.5}$$

is named the Gateaux derivative of the complex vector field f in the direction h.

By simple calculation, we obtain easily that f' is a linear operator on $V^{(s)}$ into $V^{(m)}$ containing differential operator D_j and integrable operator I_j. In fact, if we set

$$f'_{l\phi_j}=\sum_\alpha\frac{\partial f_l}{\partial(D^\alpha\phi_j)}D^\alpha+\sum_\beta\frac{\partial f_l}{\partial(I^\beta\phi_j)}I^\beta, \tag{2.6}$$

we have

$$f'=(f'_{l\phi_j})_{m\times s}. \tag{2.7}$$

For an operator $L\in Q^{(m)}$, we can also define their Gateaux derivative operator L' on $V^{(s)}$ into $Q^{(m)}$ and the Leibniz formula holds.

Concerning the high-order Gateaux derivative, we have the following relation.

Lemma 1: Let $f,g\in V^{(s)}$ and $L\in Q^{(m)}$; then

$$(L')'[f]g=(L')'[g]f; \tag{2.8a}$$

$$(L'[f])'[g]-(L'[g])'[f]=L'[[\![f,g]\!]], \tag{2.8b}$$

where $[\![f,g]\!]=f'[g]-g'[f]$ is the commutator of f and g.

Definition 3: If $\theta=\theta(t,x,\phi)\in Q^{(s)}$ is an anti-self-conjugate operator and satisfies the identity

$$\langle f,\theta'[\theta g]h\rangle+\langle g,\theta'[\theta h]f\rangle+\langle h,\theta'[\theta f]g\rangle=0, \tag{2.9}$$

for any $f,g,h\in V^{(s)}$, then θ is named an implectic operator.

Clearly, the above D_j, I_j, and σ are all implectic operators.

Definition 4: For a given real-valued functional $H=H(t,x,\phi)$, if there exits a complex vector field $f\in V^{(s)}$ satisfying

$$H'[h]=\langle f,h\rangle, \tag{2.10}$$

for any $h\in V^{(s)}$, then f is named functional derivative or Gradient of H and is denoted as

$$\frac{\delta H}{\delta\phi}=\text{grad } H=f. \tag{2.11}$$

The above Gateaux derivative and functional derivative also have a similar close relationship to the real vector fields.

Theorem 1: Suppose that a complex vector field $f\in V^{(s)}$ is the functional derivative of a real-valued functional w, i.e., $f=\delta w/\delta\phi$, then the Gateaux derivative of f is a self-conjugate operator. Inversely, if $f'=f'^*$, then f is the functional derivative of the functional,

$$w=\int_0^1\langle f(\rho\phi),\phi\rangle d\rho. \tag{2.12}$$

Proof: Let f be the functional derivative of w; then for any complex vector field $g,h\in V^{(s)}$, we have

$$w'[g]=\langle f,g\rangle,\quad w'[h]=\langle f,h\rangle. \tag{2.13}$$

It follows that

$$(w'[g])'[h]=\langle f'[h],g\rangle+\langle f,g'[h]\rangle \tag{2.14}$$

and

$$(w'[h])'[g]=\langle f'[g],h\rangle+\langle f,h'[g]\rangle. \tag{2.15}$$

Using the formula (2.8b), from (2.14) and (2.15), we obtain

$$w'[[\![g,h]\!]]=\langle f'[h],g\rangle-\langle f'[g],h\rangle+\langle f,[[\![g,h]\!]]\rangle, \tag{2.16}$$

which leads to

$$\langle g,f'[h]\rangle=\langle f'[h],g\rangle=\langle f'[g],h\rangle, \tag{2.17}$$

i.e., f' is a self-conjugate operator.

On the other hand, calculating the Gateaux derivative of (2.12) in the direction $h\in V^{(s)}$ and noting $\phi'[h]=h$, we obtain

$$w'[h]=\int_0^1(\langle\rho f'(\rho\phi)[h],\phi\rangle+\langle f(\rho\phi),h\rangle)d\rho. \tag{2.18}$$

According to the hypothesis $f'=f'^*$, the above equality can also be read as

$$w'[h]=\int_0^1\langle\rho f'(\rho\phi)[\phi]+f(\rho\phi),h\rangle d\rho. \tag{2.19}$$

Because

$$f'(\rho\phi)[\phi]=\frac{d}{d\epsilon}f(\rho\phi+\epsilon\phi)|_{\epsilon=0}=\frac{d}{d\rho}f(\rho\phi), \tag{2.20}$$

we finally have

$$w'[h]=\int_0^1\left\langle\frac{d}{d\rho}\rho f(\rho\phi),h\right\rangle d\rho=\langle f,h\rangle, \tag{2.21}$$

which is just the desired result.

Definition 5: Suppose that there is a complex nonlinear equation,

$$\phi_t=\theta(\phi)f(t,x,\phi), \tag{2.22}$$

if $\theta(\phi)\in Q^{(s)}$ is an implectic operator and $f(t,x,\phi)\in V^{(s)}$ is a functional derivative of some real-valued functional $w=w(t,x,\phi)$; then (2.22) is called a Hamiltonian equation and w is called its Hamiltonian functional.

In light of Theorem 1, we have immediately the following corollary.

Corollary: Let $\theta(\phi)\in Q^{(s)}$ be an implectic operator; then the sufficient and necessary condition for (2.22) to be a Hamiltonian equation is that the Gateaux derivative of f is a self-conjugate operator,

$$f'=f'^*, \tag{2.23}$$

and the corresponding Hamiltonian functional is given by (2.12).

III. COMPLETE CONSTRAINTS OF THE COMPLEX KP SYSTEM

Consider the linear problem of the complex KP system,

$$i\phi_y=\phi_{xx}+2u\phi, \tag{3.1a}$$

$$\phi_t=\phi_{xxx}+3u\phi_x+\tfrac{3}{2}u_x\phi+\tfrac{3}{2}i(Iu_y)\phi; \tag{3.1b}$$

and its conjugate linear problem,

$$i\psi_y=-\psi_{xx}-2v\psi, \tag{3.2a}$$

$$\psi_t=\psi_{xxx}+3v\psi_x+\tfrac{3}{2}v_x\psi-\tfrac{3}{2}i(Iv_y)\psi, \tag{3.2b}$$

where u and $v=\overline{u}$ both satisfy the KP equation,

$$q_t=\tfrac{1}{4}q_{xxx}+3qq_x-\tfrac{3}{4}Iq_{yy}, \tag{3.3}$$

$\psi=\overline{\phi}$, $i^2=-1$, and

$$I=\frac{1}{2}(I^--I^+),\quad I^-=\int_{-\infty}^{x}dx,\quad I^+=\int_x^{\infty}dx. \tag{3.4}$$

If the potentials u and v are constrained to the symmetry $\phi\psi$ of the complex KP system, i.e., $u=v=\phi\psi$, then (3.1a) and (3.2a) become[1]

$$i\phi_y=\phi_{xx}+2\phi^2\psi, \tag{3.5a}$$

$$i\psi_y=-\psi_{xx}-2\phi\psi^2, \tag{3.5b}$$

which is the NS equation. Also, (3.1b) and (3.2b) become

$$\phi_t=\phi_{xxx}+6\phi\psi\phi_x, \tag{3.6a}$$

$$\psi_t=\psi_{xxx}+6\phi\psi\psi_x, \tag{3.6b}$$

which is the first high-order NS equation.

In order to derive other nonsymmetric constraints of the complex KP system, we choose the implectic operator σ and rewrite the linear problems (3.1) and (3.2) as the vector form

$$\omega_y=\sigma f; \tag{3.7a}$$

$$\omega_t=\sigma g, \tag{3.7b}$$

where $\omega,f,g\in V^{(2)}$ and

$$\omega=\begin{pmatrix}\phi\\ \psi\end{pmatrix},\quad f=\begin{pmatrix}f_1\\ f_2\end{pmatrix}=\begin{pmatrix}\phi_{xx}+2u\phi\\ \psi_{xx}+2v\psi\end{pmatrix}; \tag{3.8a}$$

$$g=\begin{pmatrix}g_1\\ g_2\end{pmatrix}=\begin{pmatrix} i\phi_{xxx}+3iu\phi_x+\frac{3}{2}iu_x\phi-\frac{3}{2}(Iu_y)\phi\\ -i\psi_{xxx}-3iv\psi_x-\frac{3}{2}iv_x\psi-\frac{3}{2}(Iv_y)\psi\end{pmatrix}. \tag{3.8b}$$

Now, let the general form of the potential constraint of the complex KP system be

$$(u,v)=(u(\phi,\psi,\phi_x,\psi_x),v(\phi,\psi,\phi_x,\psi_x)); \tag{3.9}$$

then the sufficient and necessary conditions for (3.7a) and (3.7b) to be Hamiltonian equations are

$$f'=f'^*; \tag{3.10a}$$

$$g'=g'^*, \tag{3.10b}$$

where

$$f'=\begin{pmatrix}f'_{1,\phi} & f'_{1,\psi}\\ f'_{2,\phi} & f'_{2,\psi}\end{pmatrix},\quad f'^*=\begin{pmatrix}f'^*_{1,\phi} & f'^*_{2,\phi}\\ f'^*_{1,\psi} & f'^*_{2,\psi}\end{pmatrix}; \tag{3.11a}$$

$$g'=\begin{pmatrix}g'_{1,\phi} & g'_{1,\psi}\\ g'_{2,\phi} & g'_{2,\psi}\end{pmatrix},\quad g'^*=\begin{pmatrix}g'^*_{1,\phi} & g'^*_{2,\phi}\\ g'^*_{1,\psi} & g'^*_{2,\psi}\end{pmatrix}. \tag{3.11b}$$

From (3.10), one can obtain by direct computation (see the Appendix)

$$u=-ia\psi\phi_x+\alpha\phi\psi+\beta, \tag{3.12a}$$

$$v=ia\phi\psi_x+\alpha\phi\psi+\beta, \tag{3.12b}$$

where a, α, and β are three arbitrary real constants. Therefore we have the following.

Theorem 2: Assume that the complex vector fields f and g of (3.8) belong to $V^{(2)}$; then the necessary and sufficient condition for (3.7) to be Hamiltonian equations is that the potential constraints take the form of (3.12).

Substituting the potential (3.12) into (3.7a) and (3.7b), respectively, gives

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_y=\sigma\begin{pmatrix}\phi_{xx}-2(ia\psi\phi_x-\alpha\phi\psi-\beta)\phi\\ \psi_{xx}+2(ia\phi\psi_x+\alpha\phi\psi+\beta)\psi\end{pmatrix}, \tag{3.13a}$$

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_t=\sigma\begin{pmatrix} i\phi_{xxx}+3a\psi(\phi\phi_x)_x-3ia^2(\phi\psi)^2\phi_x+3a\alpha(\phi\psi)^2\phi+6i\alpha\phi\psi\phi_x+3i\beta\phi_x\\ -i\psi_{xxx}+3a\phi(\psi\psi_x)_x+3ia^2(\phi\psi)^2\psi_x+3a\alpha(\phi\psi)^2\psi-6i\alpha\phi\psi\psi_x-3i\beta\psi_x\end{pmatrix}, \tag{3.13b}$$

which are Hamiltonian equations, and the corresponding Hamiltonian functionals are

$$w_1=\int_{-\infty}^{\infty}[\tfrac{1}{2}(\psi\phi_{xx}+\phi\psi_{xx})+\tfrac{1}{2}ia\phi\psi(\phi\psi_x-\psi\phi_x)+\alpha(\phi\psi)^2+2\beta\phi\psi]dx, \tag{3.14a}$$

$$w_2=\int_{-\infty}^{\infty}\{\tfrac{1}{2}i(\psi\phi_{xxx}-\phi\psi_{xxx})+\tfrac{3}{4}a[\psi^2(\phi\phi_x)_x+\phi^2(\psi\psi_x)_x]$$
$$+a\alpha(\phi\psi)^3-\tfrac{1}{2}i(a^2\phi^2\psi^2-3\alpha\phi\psi-3\beta)(\psi\phi_x-\phi\psi_x)\}dx. \tag{3.14b}$$

In particular, let $\alpha=1$ and $a=\beta=0$; the potential (3.12) becomes the well-known symmetry constraint of the linear problem of the complex KP system. Equations (3.13a) and (3.13b) become the NS equations (3.5) and (3.6), respectively, and the corresponding Hamiltonian functionals are

$$w_1=\frac{1}{2}\int_{-\infty}^{\infty}(\psi\phi_{xx}+\phi\psi_{xx}+2\phi^2\psi^2)dx, \tag{3.15a}$$

$$w_2=\frac{1}{2}i\int_{-\infty}^{\infty}[\psi\phi_{xxx}-\phi\psi_{xxx}+3\phi\psi(\psi\phi_x-\phi\psi_x)]dx. \tag{3.15b}$$

Let $a=1$ and $\alpha=\beta=0$; the potential (3.12) is

$$(u,v)=(-i\psi\phi_x,i\phi\psi_x), \tag{3.16}$$

and (3.13a) and (3.13b) give, respectively,

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_y=\sigma\begin{pmatrix}\phi_{xx}-2i\phi\psi\phi_x\\ \psi_{xx}+2i\phi\psi\psi_x\end{pmatrix}, \tag{3.17a}$$

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_t=\sigma\begin{pmatrix}i\phi_{xxx}+3\psi(\phi\phi_x)_x-3i(\phi\psi)^2\phi_x\\ -i\psi_{xxx}+3\phi(\psi\psi_x)_x+3i(\phi\psi)^2\psi_x\end{pmatrix}, \tag{3.17b}$$

which are new Hamiltonian equations. To our knowledge, their integrabilities have not been reported in any literature. The corresponding Hamiltonian functionals are

$$w_1=\frac{1}{2}\int_{-\infty}^{\infty}[\psi\phi_{xx}+\phi\psi_{xx}+i\phi\psi(\phi\psi_x-\psi\phi_x)]dx, \tag{3.18a}$$

$$w_2=\frac{1}{2}i\int_{-\infty}^{\infty}\{\psi\phi_{xxx}-\phi\psi_{xxx}-\tfrac{3}{2}i[\psi^2(\phi\phi_x)_x+\phi^2(\psi\psi_x)_x]-\phi^2\psi^2(\psi\phi_x-\phi\psi_x)\}dx. \tag{3.18b}$$

IV. THE REDUCTION TECHNIQUE OF THE COMPLEX MKP SYSTEM

For the sake of finding more $(1+1)$-dimensional integrable systems, in this section we would like to consider the potential constraints and the new reduction method of the complex MKP system.

It is known that the linear problem of the complex MKP system is

$$i\phi_y=\phi_{xx}+2iu\phi_x, \tag{4.1a}$$

$$\phi_t=\phi_{xxx}+3iu\phi_{xx}+\tfrac{3}{2}(iu_x-u^2-Iu_y)\phi_x; \tag{4.1b}$$

and the conjugate linear problem is

$$i\psi_y=-\psi_{xx}+2iv\psi_x, \tag{4.2a}$$

$$\psi_t=\psi_{xxx}-3iv\psi_{xx}-\tfrac{3}{2}(iv_x+v^2+Iv_y)\psi_x, \tag{4.2b}$$

where u and $v=\overline{u}$ both satisfy the MKP equation,

$$q_t=\tfrac{1}{4}q_{xxx}+\tfrac{3}{2}q^2q_x-\tfrac{3}{2}q_xIq_y-\tfrac{3}{4}Iq_{yy}, \tag{4.3}$$

and $\psi=\overline{\phi}$.

If the potentials u and v are constrained into the symmetry of the MKP system, i.e., $u=v=-\phi\psi$, then (4.1a) and (4.2a) become the integrable equation (3.17a). Also, (4.1b) and (4.2b) become a high-order integrable equation (3.17b).

Generally, if we write the linear problems (4.1) and (4.2) as the vector form (3.7), where

$$f=\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}=\begin{pmatrix} \phi_{xx}+2iu\phi_x \\ \psi_{xx}-2iv\psi_x \end{pmatrix}, \tag{4.4a}$$

$$g=\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}=\begin{pmatrix} i\phi_{xxx}-3u\phi_{xx}-\frac{3}{2}(u_x+iu^2+iIu_y)\phi_x \\ -i\psi_{xxx}-3v\psi_{xx}-\frac{3}{2}(v_x-iv^2-iIv_y)\psi_x \end{pmatrix}, \tag{4.4b}$$

then, using the approach ($f'=f'^*$, $g'=g'^*$) provided in the previous section, we can deduce the potential complete constraint of the MKP system.

Theorem 3: Assume the vector fields f and g in (4.4) belong to $V^{(2)}$; then the necessary and sufficient condition for corresponding (3.7) to be Hamiltonian equations is that the complete constraint takes the functional form

$$u=v=\alpha\phi\psi+\beta\phi+\gamma\psi+\delta, \tag{4.5}$$

where α and δ are the real constants and β and $\gamma=\overline{\beta}$ are the conjugate complex numbers. Substituting (4.5) into (3.7a) and (3.7b) yields

$$\begin{pmatrix} \phi \\ \psi \end{pmatrix}_y=\sigma\begin{pmatrix} \phi_{xx}+2i(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)\phi_x \\ \psi_{xx}-2i(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)\psi_x \end{pmatrix}; \tag{4.6a}$$

$$\begin{pmatrix} \phi \\ \psi \end{pmatrix}_t=\sigma\begin{pmatrix} i\phi_{xxx}-3(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)\phi_{xx}-3[\alpha\psi\phi_x+\beta\phi_x+i(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)^2]\phi_x \\ -i\psi_{xxx}-3(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)\psi_{xx}-3[\alpha\phi\psi_x+\gamma\psi_x-i(\alpha\phi\psi+\beta\phi+\gamma\psi+\delta)^2]\psi_x \end{pmatrix}, \tag{4.6b}$$

which are the Hamiltonian equations, and the corresponding Hamiltonian functionals are

$$w_1=\int_{-\infty}^{\infty}\{\tfrac{1}{2}(\psi\phi_{xx}+\phi\psi_{xx})+i[\tfrac{1}{2}\alpha\phi\psi+\tfrac{2}{3}(\beta\phi+\gamma\psi)+\delta](\psi\phi_x-\phi\psi_x)\}dx; \tag{4.7a}$$

$$\begin{aligned} w_2=\int_{-\infty}^{\infty}\{&\tfrac{1}{2}i(\psi\phi_{xxx}-\phi\psi_{xxx})-\tfrac{3}{4}\alpha[\psi^2(\phi\phi_x)_x+\phi^2(\psi\psi_x)_x]-(\beta\phi+\gamma\psi+\tfrac{3}{2}\delta) \\ &\times(\psi\phi_{xx}+\phi\psi_{xx})-\beta\psi\phi_x^2-\gamma\phi\psi_x^2+i[\tfrac{1}{2}\alpha^2\phi^2\psi^2+\tfrac{6}{5}\alpha\phi\psi(\beta\phi+\gamma\psi) \\ &+\tfrac{3}{4}(\beta\phi+\gamma\psi)^2+\tfrac{3}{2}\alpha\delta\phi\psi+2\delta(\beta\phi+\gamma\psi)+\tfrac{3}{2}\delta^2](\phi\psi_x-\psi\phi_x)\}dx. \end{aligned} \tag{4.7b}$$

Specially, if we set $\alpha=-1$, $\beta=\delta=0$, then (4.6a) and (4.6b) become (3.17a) and (3.17b), and the Hamilton functionals (4.7a) and (4.7b) become (3.18a) and (3.18b). If we set $\beta=1$, $\alpha=\delta=0$, then (4.6a) and (4.6b) become

$$\begin{pmatrix} \phi \\ \psi \end{pmatrix}_y=\sigma\begin{pmatrix} \phi_{xx}+2i(\phi+\psi)\phi_x \\ \psi_{xx}-2i(\phi+\psi)\psi_x \end{pmatrix}; \tag{4.8a}$$

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_t=\sigma\begin{pmatrix} i\phi_{xxx}-3(\phi+\psi)\phi_{xx}-3[\phi_x+i(\phi+\psi)^2]\phi_x\\ -i\psi_{xxx}-3(\phi+\psi)\psi_{xx}-3[\psi_x-i(\phi+\psi)^2]\psi_x\end{pmatrix}, \tag{4.8b}$$

which are new Hamiltonian equations. The corresponding Hamiltonian functionals are

$$w_1=\int_{-\infty}^{\infty}[\tfrac{1}{2}(\psi\phi_{xx}+\phi\psi_{xx})+\tfrac{2}{3}i(\phi+\psi)(\psi\phi_x-\phi\psi_x)]dx; \tag{4.9a}$$

$$w_2=\int_{-\infty}^{\infty}[\tfrac{1}{2}i(\psi\phi_{xxx}-\phi\psi_{xxx})-(\phi+\psi)(\psi\phi_{xx}+\phi\psi_{xx}) -(\psi\phi_x^2+\phi\psi_x^2)+\tfrac{3}{4}i(\phi+\psi)^2(\phi\psi_x-\psi\phi_x)]dx. \tag{4.9b}$$

Similarly, if setting $\beta=i$, $\alpha=\delta=0$, from (4.6), we can obtain another new Hamilton equation,

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_y=\sigma\begin{pmatrix}\phi_{xx}-2(\phi-\psi)\phi_x\\ \psi_{xx}+2(\phi-\psi)\psi_x\end{pmatrix}; \tag{4.10a}$$

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix}_t=\sigma\begin{pmatrix} i\phi_{xxx}-3i(\phi-\psi)\phi_{xx}-3i[\phi_x-(\phi-\psi)^2]\phi_x\\ -i\psi_{xxx}-3i(\phi-\psi)\psi_{xx}+3i[\psi_x-(\phi-\psi)^2]\psi_x\end{pmatrix}, \tag{4.10b}$$

the corresponding Hamilton functionals are

$$w_1=\int_{-\infty}^{\infty}[\tfrac{1}{2}(\psi\phi_{xx}+\phi\psi_{xx})-\tfrac{2}{3}(\phi-\psi)(\psi\phi_x-\phi\psi_x)]dx; \tag{4.11a}$$

$$w_2=\int_{-\infty}^{\infty}[\tfrac{1}{2}i(\psi\phi_{xxx}-\phi\psi_{xxx})-i(\phi-\psi)(\psi\phi_{xx}+\phi\psi_{xx}) -i(\psi\phi_x^2-\phi\psi_x^2)-\tfrac{3}{4}i(\phi-\psi)^2(\phi\psi_x-\psi\phi_x)]dx. \tag{4.11b}$$

In what follows, we shall propose a new reduction approach that can generate the KN equations and the HB equations from the linear problem of the complex MKP system.

Differentiating (4.1a) and (4.2a) with respect to x and replacing ϕ_x and ψ_x by ϕ and ψ, we have

$$i\phi_y=(\phi_x+2iu\phi)_x, \tag{4.12a}$$

$$i\psi_y=(-\psi_x+2iv\psi)_x. \tag{4.12b}$$

If the potentials u and v are both constrained to the symmetry $\phi\psi$, i.e., $u=v=\phi\psi$, then (4.12) becomes the KN equation,[8,9]

$$i\phi_y=(\phi_x+2i\phi^2\psi)_x, \tag{4.13a}$$

$$i\psi_y=(-\psi_x+2i\phi\psi^2)_x, \tag{4.13b}$$

which is a Hamiltonian equation,

$$\omega_y=Df=D\frac{\delta w_1}{\delta\omega}, \tag{4.14a}$$

where $f\in V^{(2)}$,

$$f=\begin{pmatrix}f_1\\ f_2\end{pmatrix}=\begin{pmatrix}-i\phi_x+2\phi^2\psi\\ i\psi_x+2\psi^2\phi\end{pmatrix}, \tag{4.14b}$$

and w_1 is a Hamiltonian functional,

$$w_1=\frac{1}{2}\int_{-\infty}^{\infty}[i(\phi\psi_x-\psi\phi_x)+2\phi^2\psi^2]dx. \tag{4.14c}$$

Still applying the same operation to (4.1b) and (4.2b) i.e., differentiating (4.1a) and (4.2b) and replacing ϕ_x, ψ_x by ϕ, ψ, yields easily

$$\phi_t=[\phi_{xx}+3iu\phi_x+\tfrac{3}{2}(iu_x-u^2-Iu_y)\phi]_x, \tag{4.15a}$$

$$\psi_t=[\psi_{xx}-3iv\psi_x-\tfrac{3}{2}(iv_x+v^2+Iv_y)\psi]_x. \tag{4.15b}$$

Setting $u=v=\phi\psi$ and noticing $Iu_y=Iv_y=i(\phi\psi_x-\psi\phi_x)+3\phi^2\psi^2$, we can derive the high-order KN equation,[8,9]

$$\phi_t=(\phi_{xx}+6i\phi\psi\phi_x+6\phi^3\psi^2)_x, \tag{4.16a}$$

$$\psi_t=(\psi_{xx}-6i\phi\psi\psi_x+6\phi^2\psi^3)_x, \tag{4.16b}$$

which is also a Hamiltonian equation,

$$\omega_t=Dg=D\frac{\delta w_2}{\delta\omega}, \tag{4.17a}$$

where $g\in V^{(2)}$,

$$g=\begin{pmatrix}g_1\\g_2\end{pmatrix}=\begin{pmatrix}\phi_{xx}+6i\phi\psi\phi_x+6\phi^3\psi^2\\ \psi_{xx}-6i\phi\psi\psi_x+6\phi^2\psi^3\end{pmatrix}, \tag{4.17b}$$

and w_2 is a Hamiltonian functional,

$$w_2=\frac{1}{2}\int_{-\infty}^{\infty}[\psi\phi_{xx}+\phi\psi_{xx}+3i\phi\psi(\psi\phi_x-\phi\psi_x)+4\phi^3\psi^3]dx. \tag{4.17c}$$

In (4.12) and (4.15), we introduce a new real function ν satisfying

$$\nu^2+\phi\psi=1, \tag{4.18}$$

and set the potential constraints to be

$$u=\frac{1}{2}i\left(\nu_x+\frac{1}{1+\nu}\psi\phi_x\right), \tag{4.19a}$$

$$v=-\frac{1}{2}i\left(\nu_x+\frac{1}{1+\nu}\phi\psi_x\right); \tag{4.19b}$$

then (4.12a) and (4.12b) become

$$\phi_y=i(\phi\nu_{xx}-\nu\phi_{xx}), \tag{4.20a}$$

$$\psi_y=i(\nu\phi_{xx}-\psi\nu_{xx}), \tag{4.20b}$$

$$\nu_y=\tfrac{1}{2}i(\psi\phi_{xx}-\phi\psi_{xx}), \tag{4.20c}$$

which is the HB equation.[10]

Inserting (4.19a) and (4.19b) into (4.15a) and (4.15b), respectively, and noticing

$$Iu_y=-\frac{1}{4}\left(3\phi_x\psi_x+3\nu_x^2+2\nu_{xx}+\frac{2}{1+\nu}\psi\phi_{xx}\right), \tag{4.21a}$$

$$Iv_y=-\frac{1}{4}\left(3\phi_x\psi_x+3\nu_x^2+2\nu_{xx}+\frac{2}{1+\nu}\,\phi\psi_{xx}\right), \tag{4.21b}$$

we obtain

$$\phi_t=\phi_{xxx}+\tfrac{3}{2}[(\phi_x\psi_x+\nu_x^2)\phi]_x, \tag{4.22a}$$

$$\psi_t=\psi_{xxx}+\tfrac{3}{2}[(\phi_x\psi_x+\nu_x^2)\psi]_x, \tag{4.22b}$$

$$\nu_t=\nu_{xxx}+\tfrac{3}{2}[(\phi_x\psi_x+\nu_x^2)\nu]_x, \tag{4.22c}$$

which is the high-order HB equation.[10] Equations (4.20b) and (4.22b) are both the Hamiltonian equations. But here the differential operator D is not their implectic operator. Consider the $V^{(2)}$'s linear subspace $V_0^{(2)}$ constructed by all vector fields (f_1,f_2) with $f_2=f_1$ and the complex function matrix

$$\theta=\sigma\nu=\begin{pmatrix}-i\nu & 0\\ 0 & i\nu\end{pmatrix}, \tag{4.23}$$

with $\nu=\sqrt{1-\phi\psi}$. It is easy to check that θ is an anti-Hermite matrix, and for any $f,g,h\in V_0^{(2)}$, θ satisfies (2.9); therefore it is the implectic operator.

Now we rewrite the equations (4.20b) and (4.22b) as the vector form

$$\omega_y=\theta f,\quad f=\begin{pmatrix}f_1\\ f_2\end{pmatrix}=\frac{1}{\nu}\begin{pmatrix}\nu\phi_{xx}-\phi\nu_{xx}\\ \nu\psi_{xx}-\psi\nu_{xx}\end{pmatrix}, \tag{4.24a}$$

$$\omega_t=\theta g,\quad g=\begin{pmatrix}g_1\\ g_2\end{pmatrix}=\frac{i}{\nu}\begin{pmatrix}\phi_{xxx}+\tfrac{3}{2}[(\phi_x\psi_x+\nu_x^2)\phi]_x\\ -\psi_{xxx}-\tfrac{3}{2}[(\phi_x\psi_x+\nu_x^2)\psi]_x\end{pmatrix}, \tag{4.24b}$$

where f and g belong to $V_0^{(2)}$. By direct computation, we can verify that $f'=f'^*$ and $g'=g'^*$. Consequently, (4.24a) and (4.24b) are both the Hamiltonian equations, and the corresponding Hamiltonian functionals are

$$w_1=\int_{-\infty}^{\infty}\left[\frac{1}{2}\,(\phi\psi)_x(\ln\,\nu)_x-\phi_x\psi_x\right]dx, \tag{4.25a}$$

$$w_2=\int_{-\infty}^{\infty}\left\{\frac{i}{1+\nu}\,(\psi\phi_{xxx}-\phi\psi_{xxx})+\frac{i}{2(1+\nu)^2}\,(\psi\phi_x-\phi\psi_x)\left[(\nu+2)\phi_x\psi_x+\frac{\nu(\nu+3)}{1+\nu}\,\nu_x^2\right]\right\}dx. \tag{4.25b}$$

This is a new result. To our knowledge, the Hamiltonian property of the HB equations have not been reported in any literature.

For a complex operator, $T=\Sigma a_{ijk}(u,v)\partial_x^i\partial_y^j\partial_t^k$; according to Definition 1, its conjugate (adjoint) operator is $T^*=\Sigma(-1)^{i+j+k}\partial_x^i\partial_y^j\partial_t^k\overline{a_{ijk}(u,v)}$; then we find easily that (4.12a), (4.15a) and (4.12b), (4.15b) are just the adjoints of the linear problems (4.1) and (4.2), respectively. Therefore, the new reduction approach is substantially to nonlinearize the adjoint linear problems of the MKP system. (These adjoint linear problems are completely different from the adjoint of the real MKP system defined by Cheng *et al.* in Ref. 3.)

ACKNOWLEDGMENT

The author's sincere thanks are extended to the referees who have offered me valuable suggestions. The author also thanks Da-jun Zhang and Hong-wei Xin for their helpful discussions and careful verifications to all the results.

This project is supported by the National Science Foundation of China.

APPENDIX: PROOF OF CONSTRAINED FORMULAS OF COMPLEX KP SYSTEM

In the Appendix, we will prove the formulas (3.12).

From (3.8a) and (3.9), by using (2.6), it is easily calculated that

$$f'_{1,\phi}=D^2+2\frac{\partial\phi u}{\partial\phi_x}D+2\frac{\partial\phi u}{\partial\phi},\quad f'_{1,\psi}=2\phi\left(\frac{\partial u}{\partial\psi_x}D+\frac{\partial u}{\partial\psi}\right),$$
$$f'_{2,\phi}=2\psi\left(\frac{\partial v}{\partial\phi_x}D+\frac{\partial v}{\partial\phi}\right),\quad f'_{2,\psi}=D^2+2\frac{\partial\psi v}{\partial\psi_x}D+2\frac{\partial\psi v}{\partial\psi}. \tag{A1}$$

Thus, we have

$$f'^{*}_{1,\phi}=D^2-2D\frac{\partial\psi v}{\partial\psi_x}+2\frac{\partial\psi v}{\partial\psi},\quad f'^{*}_{2,\phi}=-2\left(D\phi\frac{\partial u}{\partial\psi_x}-\phi\frac{\partial u}{\partial\psi}\right),$$
$$f'^{*}_{1,\psi}=-2\left(D\psi\frac{\partial v}{\partial\phi_x}-\psi\frac{\partial v}{\partial\phi}\right),\quad f'^{*}_{2,\psi}=D^2-2D\frac{\partial\phi u}{\partial\phi_x}+2\frac{\partial\phi u}{\partial\phi}. \tag{A2}$$

Substituting (A1) and (A2) into (3.10a) and comparing the coefficients of D and the functional term of the corresponding elements leads to

$$\phi\frac{\partial u}{\partial\phi_x}=-\psi\frac{\partial v}{\partial\psi_x}, \tag{A3a}$$

$$\frac{\partial u}{\partial\psi_x}=0,\quad \frac{\partial v}{\partial\phi_x}=0, \tag{A3b}$$

$$\frac{\partial\phi u}{\partial\phi}=-\left(\psi\frac{\partial v}{\partial\psi_x}\right)_x+\frac{\partial\psi v}{\partial\psi}, \tag{A3c}$$

$$\frac{\partial\psi v}{\partial\psi}=-\left(\phi\frac{\partial u}{\partial\phi_x}\right)_x+\frac{\partial\phi u}{\partial\phi}. \tag{A3d}$$

From Eqs. (A3), we have immediately

$$u=-ia(\phi,\psi)\psi\phi_x+b_1(\phi,\psi); \tag{A4a}$$

$$v=ia(\phi,\psi)\phi\psi_x+b_2(\phi,\psi), \tag{A4b}$$

where a is the real function of ϕ and ψ, $b_2=\overline{b_1}$ and

$$\frac{\partial\phi b_1}{\partial\phi}=\frac{\partial\psi b_2}{\partial\psi}. \tag{A5}$$

In order to determine the unknown function $a(\phi,\psi)$ and $b_1(\phi,\psi)$, we have to use the equality (3.10b).

It follows from (3.8b), (A4), and (2.6) that

$$g'_{1,\phi}=iD^3+3iuD+3i\phi_x\left(\frac{\partial u}{\partial\phi_x}D+\frac{\partial u}{\partial\phi}\right)+\frac{3}{2}i\phi D\left(\frac{\partial u}{\partial\phi_x}D+\frac{\partial u}{\partial\phi}\right)+\frac{3}{2}iu_x-\frac{3}{2}\phi I(u_y)'_\phi-\frac{3}{2}Iu_y. \tag{A6}$$

Noticing the following equalities:

$$\frac{\partial u_y}{\partial\phi_{xxx}}=-i\frac{\partial u}{\partial\phi_x},\quad \frac{\partial u_y}{\partial\phi_{xx}}=-i\frac{\partial u}{\partial\phi}-2i\phi\left(\frac{\partial u}{\partial\phi_x}\right)^2, \tag{A7}$$

we calculate easily

$$\phi I(u_y)'_\phi = -i\left\{ D\phi\,\frac{\partial u}{\partial \phi_x}\,D + 2\phi\,\frac{\partial u}{\partial \phi_x}\,D\phi\,\frac{\partial u}{\partial \phi_x} - \left[\phi_x\,\frac{\partial u}{\partial \phi_x} + 2\phi\left(\frac{\partial u}{\partial \phi_x}\right)_x - \phi\,\frac{\partial u}{\partial \phi}\right]D + \phi\left(\frac{\partial u}{\partial \phi_x}\right)_{xx} - 4\phi\,\frac{\partial u}{\partial \phi_x}\left(\phi\,\frac{\partial u}{\partial \phi_x}\right)_x - 2\phi^2\,\frac{\partial u}{\partial \phi_x}\left(\frac{\partial u}{\partial \phi_x}\right)_x - \phi\left(\frac{\partial u}{\partial \phi}\right)_x \right\} + \phi\,\frac{\partial u_y}{\partial \phi_x} + \phi I\,\frac{\delta u_y}{\delta \phi}, \tag{A8}$$

where

$$\frac{\delta u_y}{\delta \phi} = \frac{\partial u_y}{\partial \phi} - \left(\frac{\partial u_y}{\partial \phi_x}\right)_x + \left(\frac{\partial u_y}{\partial \phi_{xx}}\right)_{xx} - \left(\frac{\partial u_y}{\partial \phi_{xxx}}\right)_{xxx}. \tag{A9}$$

Then (A6) can be rewritten as

$$\begin{aligned} g'_{1\phi} = {} & iD^3 + 3iD\phi\,\frac{\partial u}{\partial \phi_x}\,D + 3i\phi\,\frac{\partial u}{\partial \phi_x}\,D\phi\,\frac{\partial u}{\partial \phi_x} + 3i\left[u - \phi\left(\frac{\partial u}{\partial \phi_x}\right)_x + \phi\,\frac{\partial u}{\partial \phi}\right]D \\ & + 3i\phi_x\,\frac{\partial u}{\partial \phi} + \frac{3}{2}\,i\phi\left(\frac{\partial u}{\partial \phi_x}\right)_{xx} - 6i\phi\,\frac{\partial u}{\partial \phi_x}\left(\phi\,\frac{\partial u}{\partial \phi_x}\right)_x \\ & - 3i\phi^2\,\frac{\partial u}{\partial \phi_x}\left(\frac{\partial u}{\partial \phi_x}\right)_x - \frac{3}{2}\,\phi\,\frac{\partial u_y}{\partial \phi_x} + \frac{3}{2}\,iu_x - \frac{3}{2}\,Iu_y - \frac{3}{2}\,\phi I\,\frac{\delta u_y}{\delta \phi}. \end{aligned} \tag{A10}$$

Furthermore, we have

$$\begin{aligned} g'^{*}_{1\phi} = {} & iD^3 - 3iD\psi\,\frac{\partial v}{\partial \psi_x}\,D + 3i\psi\,\frac{\partial v}{\partial \psi_x}\,D\psi\,\frac{\partial v}{\partial \psi_x} + 3i\left[v - \psi\left(\frac{\partial v}{\partial \psi_x}\right)_x + \psi\,\frac{\partial v}{\partial \psi}\right]D - 3i\psi_x\,\frac{\partial v}{\partial \psi} \\ & - \frac{3}{2}\,i\psi\left(\frac{\partial v}{\partial \psi_x}\right)_{xx} + 6i\psi\,\frac{\partial v}{\partial \psi_x}\left(\psi\,\frac{\partial v}{\partial \psi_x}\right)_x + 3i\psi^2\,\frac{\partial v}{\partial \psi_x}\left(\frac{\partial v}{\partial \psi_x}\right)_x - 3i\left[\psi\left(\frac{\partial v}{\partial \psi_x}\right)_x\right]_x \\ & + 3i\left(\psi\,\frac{\partial v}{\partial \psi}\right)_x - \frac{3}{2}\,\psi\,\frac{\partial v_y}{\partial \psi_x} + \frac{3}{2}\,iv_x - \frac{3}{2}\,Iv_y + \frac{3}{2}\,\frac{\delta v_y}{\delta \psi}\,I\psi. \end{aligned} \tag{A11}$$

In $g'_{1\phi} = g'^{*}_{1\phi}$, using the formula

$$Ir = \sum_{k=0}^{\infty} (-1)^k r^{(k)} I^{k+1}, \quad r^{(k)} = \frac{\partial^k r}{\partial x^k}, \tag{A12}$$

and comparing the coefficients of the same powers of D or I lead to

$$u - \phi\left(\frac{\partial u}{\partial \phi_x}\right)_x + \phi\,\frac{\partial u}{\partial \phi} = v - \psi\left(\frac{\partial v}{\partial \psi_x}\right)_x + \psi\,\frac{\partial v}{\partial \psi}, \tag{A13a}$$

$$\begin{aligned} & 2\phi_x\,\frac{\partial u}{\partial \phi} + \phi\left(\frac{\partial u}{\partial \phi_x}\right)_{xx} - 4\phi\,\frac{\partial u}{\partial \phi_x}\left(\phi\,\frac{\partial u}{\partial \phi_x}\right)_x - 2\phi^2\,\frac{\partial u}{\partial \phi_x}\left(\frac{\partial u}{\partial \phi_x}\right)_x + i\phi\,\frac{\partial u_y}{\partial \phi_x} + u_x + iIu_y \\ & = -2\psi_x\,\frac{\partial v}{\partial \psi} - \psi\left(\frac{\partial v}{\partial \psi_x}\right)_{xx} + 4\psi\,\frac{\partial v}{\partial \psi_x}\left(\psi\,\frac{\partial v}{\partial \psi_x}\right)_x + 2\psi^2\,\frac{\partial v}{\partial \psi_x}\left(\frac{\partial v}{\partial \psi_x}\right)_x \\ & \quad - 2\left[\psi\left(\frac{\partial v}{\partial \psi_x}\right)_x\right]_x + 2\left(\psi\,\frac{\partial v}{\partial \psi}\right)_x + i\psi\,\frac{\partial v_y}{\partial \psi_x} + v_x + iIv_y, \end{aligned} \tag{A13b}$$

$$\phi \frac{\partial^k}{\partial x^k} \frac{\delta u_y}{\delta \phi} = - \frac{\delta v_y}{\delta \psi} \frac{\partial^k \psi}{\partial x^k}, \quad k = 0,1,2,\ldots . \tag{A13c}$$

Inserting the expressions (A4) into (A13a), we obtain easily $a_x = 0$, which means a is a real constant. So

$$\begin{aligned} u_y = {} & a(\phi_x \psi_x - \psi \phi_{xx})_x + 2ia^2(\phi \psi^2 \phi_x)_x - 4ab_{1x}\phi\psi + 2a(b_2 - b_1)\psi\phi_x \\ & - i\left[(\phi_{xx} + 2b_1\phi)\frac{\partial}{\partial \phi} - (\psi_{xx} + 2b_2\psi)\frac{\partial}{\partial \psi}\right] b_1, \end{aligned} \tag{A14a}$$

and

$$\begin{aligned} v_y = {} & a(\phi_x \psi_x - \phi \psi_{xx})_x - 2ia^2(\phi^2 \psi \psi_x)_x - 4ab_{2x}\phi\psi + 2a(b_1 - b_2)\phi\psi_x \\ & - i\left[(\phi_{xx} + 2b_1\phi)\frac{\partial}{\partial \phi} - (\psi_{xx} + 2b_2\psi)\frac{\partial}{\partial \psi}\right] b_2. \end{aligned} \tag{A14b}$$

It follows that (A13b) is

$$\begin{aligned} & 2\phi_x \frac{\partial b_1}{\partial \phi} - 2\psi\left(\frac{\partial b_2}{\partial \psi}\right)_x + (b_1 - b_2)_x - 6ia(b_1 - b_2)\phi\psi \\ & + I\left\{\left[(\phi_{xx} + 2b_1\phi)\frac{\partial}{\partial \phi} - (\psi_{xx} + 2b_2\psi)\frac{\partial}{\partial \psi}\right](b_1 - b_2) - 2ia(b_1 - b_2)_x\phi\psi\right\} = 0. \end{aligned} \tag{A15}$$

Differentiating the above equality with respect to x, we have

$$\begin{aligned} & 2\phi_{xx}\frac{\partial b_1}{\partial \phi} + 2\phi_x\left(\frac{\partial b_1}{\partial \phi}\right)_x - 2\psi_x\left(\frac{\partial b_2}{\partial \psi}\right)_x - 2\psi\left(\frac{\partial^2 b_2}{\partial \phi \partial \psi}\phi_x + \frac{\partial^2 b_2}{\partial \psi^2}\psi_x\right)_x \\ & + \left(\frac{\partial (b_1 - b_2)}{\partial \phi}\phi_x + \frac{\partial (b_1 - b_2)}{\partial \psi}\psi_x\right)_x - 6ia[(b_1 - b_2)\phi\psi]_x \\ & + \left[(\phi_{xx} + 2b_1\phi)\frac{\partial}{\partial \phi} - (\psi_{xx} + 2b_2\psi)\frac{\partial}{\partial \psi}\right](b_1 - b_2) - 2ia(b_1 - b_2)_x\phi\psi = 0. \end{aligned} \tag{A16}$$

In (A16), setting the coefficients of ϕ_{xx} and ψ_{xx} to be zero, we obtain

$$\frac{\partial b_1}{\partial \phi} - \psi \frac{\partial^2 b_2}{\partial \phi \partial \psi} + \frac{\partial (b_1 - b_2)}{\partial \phi} = 0, \tag{A17a}$$

$$\frac{\partial^2 b_2}{\partial \psi^2} = 0, \tag{A17b}$$

and combining (A17a) and (A5), we also have

$$\frac{\partial^2 b_1}{\partial \phi^2} = 0. \tag{A18}$$

Furthermore, setting the coefficients of ϕ_x^2, $\phi_x\psi_x$, and ψ_x^2 to be zero yields

$$\frac{\partial^2 b_1}{\partial \psi^2} = 0, \quad \frac{\partial^2 (b_1 - b_2)}{\partial \phi \partial \psi} = 0, \quad \frac{\partial^2 b_2}{\partial \phi^2} = 0. \tag{A19}$$

Setting the coefficients of ϕ_x and ψ_x to be zero, if $a \neq 0$ we obtain

$$4\phi\,\frac{\partial(b_1-b_2)}{\partial\phi}+3(b_1-b_2)=0, \tag{A20a}$$

$$4\psi\,\frac{\partial(b_1-b_2)}{\partial\psi}+3(b_1-b_2)=0. \tag{A20b}$$

At last, setting the sum of the remaining items to be zero gives

$$\phi b_1\,\frac{\partial(b_1-b_2)}{\partial\phi}-\psi b_2\,\frac{\partial(b_1-b_2)}{\partial\psi}=0. \tag{A21}$$

Equation (A20) coupled with (A21) implies $b_1=b_2=b$, i.e., b is a real function. Thus, from (A18), (A19), and (A5), we obtain

$$b=\alpha\phi\psi+\beta, \tag{A22}$$

where α and β are real constants.

If $a=0$, one also obtains (A22) from (A17b), (A18), (A19), (A21) and (A5).

Hence, the formulas (3.12) are derived. Then, it is easy to verify that (3.12) satisfy (A13) and (3.10b). So we have proved the desired result completely.

[1] Y. Cheng and Y. S. Li, Phys. Lett. A **157**, 22 (1991).
[2] B. Konopelchenko, J. Sidorenko, and W. Strampp, Phys. Lett. A **157**, 17 (1991).
[3] Y. Cheng and Y. S. Li, J. Phys. A **25**, 419 (1992).
[4] Y. Cheng, J. Math. Phys. **33**, 3774 (1992).
[5] D. Y. Chen, N. S. Zhu, and M. Zhu, J. Math. Phys. **35**, 4725 (1994).
[6] M. Zhu, ''Potential constraint of MKP system and associated Hamiltonian equations,'' preprint, Shanghai University of Science and Technology, 1993.
[7] B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47 (1981).
[8] D. J. Kaup and A. C. Newell, J. Math. Phys. **19**, 798 (1978).
[9] Y. S. Li and D. W. Zhang, Acta Math. Sin. **25**, 464 (1982).
[10] D. Y. Chen and Y. B. Zeng, Acta Math. Sin. **27**, 624 (1984).

Thermal conduction before relaxation in slowly rotating fluids

L. Herrera[a)]
Área de Física Teórica, Facultad de Ciencias, Universidad de Salamanca, 37008, Salamanca, España

J. Martínez
Grupo de Física Estadística, Departamento de Física, Universidad Autónoma de Barcelona, 08193 Bellaterra, Barcelona, España

(Received 5 February 1998; accepted for publication 16 March 1998)

For slowly rotating fluids, we establish the existence of a critical point similar to the one found for nonrotating systems. As the fluid approaches the critical point, the effective inertial mass of any fluid element decreases, vanishing at that point and changing sign beyond it. This result implies that the first-order perturbative method is not always reliable to study dissipative processes occurring before relaxation. Physical consequences that might follow from this effect are commented on. © *1998 American Institute of Physics.* [S0022-2488(98)03406-9]

I. INTRODUCTION

An alternative path to the study of self-gravitating systems, which avoids the use of numerical procedures and/or the introduction of too restrictive simplifying assumptions, consists in perturbing the system, compelling it to withdraw from an equilibrium state. Then, evaluating it after its departure from equilibrium, it is possible to study the tendency of the evolution of the object. This is usually done following a first-order perturbative method that neglects quadratic and higher terms in the perturbed quantities. This applies whenever the relevant processes occurring in the self-gravitating object take place on time scales that are of the order of, or smaller than, the hydrostatic time scale. In this case the quasistatic approximation fails[1] (e.g., during the quick collapse phase preceding neutron star formation), and the system is evaluated immediately after its departure from equilibrium, where *immediately* means on a time scale of the order of relaxation times.

Recently, it has been shown[2–5] that, for systems out of a quasistatic approximation, a first-order perturbative theory is not always satisfactory. In fact, there exist systems for which this method seems to be inadequate, however small is the perturbation. These are those for which the parameter

$$\alpha = \frac{\kappa T}{\tau(\rho + p)}$$

is close to, or beyond the so-called critical point ($\alpha = 1$). This combination of the temperature T, the heat conduction coefficient κ, the relaxation time τ, the energy density ρ, and the pressure p, has been found to be the same in spherically symmetric systems,[2,3] and axially symmetric systems with reflection symmetry.[5] Also, the viscous spherically symmetric case has been studied with similar results.[4]

The astrophysical interest of the study of relativistic rotating fluids is past all doubt. Therefore, it seems important to establish, for such systems, the existence, or not, of a critical point as described above. With this aim, we assume that, initially, a nonviscous slowly rotating object is close to the hydrostatic equilibrium (along the r coordinate) and nearly to thermal adjustment [the so-called *complete equilibrium* (Ref. 1, p. 66)], as measured by a local Minkowskian observer.

[a)]On leave from the Departamento de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela and Centro de Astrofísica Teórica, Mérida, Venezuela.

 © 1998 American Institute of Physics

Therefore, the time derivatives of the radial velocity and heat flow can be neglected. At that time, we perturb the radial velocity and the heat flow, and we evaluate conservation equations and the heat transport equation just after the perturbation takes place, neglecting quadratic and higher terms in the perturbed quantities. Here *just after the perturbation* means on a time scale that is of the order of the relaxation time. This is necessary if the relevant processes take place on time scales that are of the order of, or smaller than, the hydrostatic time scale. This meaning of *just after the perturbation* implies that physical quantities remain unchanged, but not the time derivatives of the perturbed quantities. These are still small, but they cannot be neglected since the system is departing from the complete equilibrium.

As has been mentioned above, it is necessary to use a heat transport equation, together with the conservation equations, to find out the existence, or not, of the critical point. In order to keep clear of inconsistencies, the heat transport equation cannot be the well-known Eckart one,[6,7] because it assumes a vanishing relaxation time. Furthermore, this theory suffers from two importants drawbacks: Noncausality (the thermal signals propagate at infinite speed), and unstability (all the predicted equilibrium states are unstable). Fortunately, there exist well physically founded thermodynamical theories that avoid these problems and that can deal with prerelaxation processes.[8–11] In this work, we shall use the Israel–Stewart heat transport equation. Nevertheless, it is important to emphasize that, as in Refs. 2–5, the results found are also valid in the context of the *Extended Irreversible Thermodynamics*.[10,11]

The paper is organized as follows. The next section is devoted to introduce the interior and exterior metrics used, and to construct the stress–energy tensor. Also, the validity of the slow rotating limit is discussed. In Sec. III, the conservation equations and heat transport equation are evaluated just after perturbation, and we find the expression for the critical point. Finally, we discuss the results in the last section.

We adopt metric of signature -2 and geometrized units $c=G=1$. The quantities subscripted with r_1 denote that they are evaluated at the surface of the object, whereas a partial derivative with respect time is denoted by subscript 0.

II. ENERGY–MOMENTUM TENSOR

We consider a nonstatic and axisymmetric distribution of matter and radiation. Let us assume that the interior metric is given by[12]

$$ds^2=Y^2du^2+2\frac{Y}{X}\,dudr+2a\,\sin^2\,\theta\left(\frac{Y}{X}-Y^2\right)dud\phi-2a\,\sin^2\,\theta\left(\frac{Y}{X}\right)drd\phi-R^2d\theta^2$$
$$-\sin^2\,\theta\left[r^2+a^2\,\cos^2\,\theta+2a^2\,\sin^2\,\theta\left(\frac{Y}{X}-\frac{Y^2}{2}\right)\right]d\phi^2, \tag{1}$$

where $u=x^0$ is a time-like coordinate, $r=x^1$ is the null coordinate, and $\theta=x^2$ and $\phi=x^3$ are the usual angle coordinates. Here $R^2=r^2+a^2\,\cos^2\,\theta$, a is the angular momentum per unit mass in the weak field limit—the Kerr parameter—and X and Y are arbitrary functions of u, r, and θ. The u-coordinate is related to retarded time in a flat space–time, and therefore, u-constant surfaces are null cones open to the future. In these coordinates r-constant surfaces are oblate spheroids.

The energy–momentum tensor may be expressed in the above coordinates (1). Nevertheless, the physical quantities appearing in it will be those measured by a local Minkowskian observer comoving with the fluid. Thus, it is necessary to introduce the local Minkowski coordinates (t,x,y,z) related to these by

$$dt=Ydu+\frac{dr}{X}+a\,\sin^2\,\theta\left(\frac{1}{X}-Y\right)d\phi,$$

$$dx=\frac{dr}{X}+\frac{a\,\sin^2\,\theta}{X}\,d\phi,$$

$$dy=(r^2+a^2\cos^2\theta)^{1/2}d\theta,$$
$$dz=(r^2+a^2\cos^2\theta]^{1/2}\sin\theta d\phi. \tag{2}$$

The radial velocity of matter is given by

$$\frac{dr}{du}=\frac{XYR\omega_x-Ya\omega_z\sin\theta}{R(1-\omega_x)+Ya\omega_z\sin\theta}, \tag{3}$$

and the orbital velocity is

$$\Omega=\frac{d\phi}{du}=\frac{Y\omega_z}{R\sin\theta(1-\omega_x)+Ya\omega_z\sin^2\theta}, \tag{4}$$

where ω_x and ω_z are the corresponding components of the velocity of a fluid element as measured in the locally Minkowski frame.

In the slow rotating limit, $a\ll 1$, and consequently $\Omega\ll 1$. Thus, from (4),

$$\omega_z=\frac{\Omega R\sin\theta(1-\omega_x)}{Y(1-\Omega a\sin^2\theta)}=\frac{\Omega r\sin\theta(1-\omega_x)}{Y}+\mathcal{O}(\Omega^2),$$

and ω_z is also much less than unity. Note that in the static case (i.e., $\omega_x=0$), and for a local Minkowskian observer (i.e., $Y=1$); this means that every fluid element must move at a nonrelativistic velocity.[13] A simple calculus shows that this condition is accomplished by most of the known pulsars (Ref. 14, p. 146).

The interior metric (1) can be matched to the Kerr–Vaidya exterior metric,[15]

$$ds^2=\left(1-\frac{2mr}{\tilde{R}^2}\right)du^2+2dudr+\frac{4mr\tilde{a}\sin^2\theta}{\tilde{R}^2}dud\phi-2\tilde{a}\sin^2\theta drd\phi-\tilde{R}^2d\theta^2$$
$$-\sin^2\theta\left(r^2+\tilde{a}^2+\frac{2mr\tilde{a}^2\sin^2\theta}{\tilde{R}^2}\right)d\phi, \tag{5}$$

where $\tilde{R}=r^2+\tilde{a}^2\cos^2\theta$, $\tilde{a}$ is the exterior Kerr parameter, and m is the total mass. It is worth mentioning at this point that the metric above is not a pure radiation solution and may be interpreted as such only asymptotically.[16] A pure rotating radiation solution may be found in Ref. 17. However, although the interpretation of the Carmeli–Kaye metric is not completely clear, the appearance of the critical point is independent of the shape and the intensity of the emission pulse, and will be put in evidence for small values of luminosity (see below).

A particular solution can be found in Ref. 12. Nevertheless, in this work we shall not restrict ourselves to a particular solution, and we shall go on using the unknown functions $X(u,r,\theta)$ and $Y(u,r,\theta)$. These are constrained by the following junction conditions at the surface $(r=r_1)$,[12]

$$X_{r=r_1}=Y_{r=r_1}=\left(1-\frac{2mr}{R^2}\right)^{1/2}_{r=r_1},$$
$$\left(\frac{\partial X}{\partial\theta}\right)_{r=r_1}=\left(\frac{\partial Y}{\partial\theta}\right)_{r=r_1}=-\left(\frac{mra^2\sin(2\theta)}{X(r^2+a^2\cos^2\theta)^2}\right)_{r=r_1}, \tag{6}$$
$$\left(\frac{\partial X}{\partial r}\right)_{r=r_1}=\left(\frac{\partial Y}{\partial r}\right)_{r=r_1}=-\left(\frac{m(r^4-a^4\cos^2\theta+2r^2a^2\cos^2\theta)}{X(r^2-a^2\cos^2\theta)(r^2+a^2\cos^2\theta)^2}\right)_{r=r_1}.$$

Next, we assume that, for a local Minkowskian observer, comoving with the fluid, the space–time contains the following.

(1) An anisotropic fluid of density $\rho^{\rm mat}$, radial pressure $p^{\rm mat}$, and tangential pressure $p_\perp^{\rm mat}$.
(2) A radiation field of specific intensity $I(x,t;\mathbf{n},\nu)$, radiation energy flow q, radiation energy density $\rho^{\rm rad}$, and radiation pressure $p^{\rm rad}$.

The specific intensity of the radiation field $I(x,t;\mathbf{n},\nu)$, is measured at the position x and time t, traveling in the direction $\mathbf{n}$ with a frequency ν. The moments of $I(x,t;\mathbf{n},\nu)$ for a planar geometry can be written as[18]

$$\rho^{\rm rad}=\frac{1}{2}\int_0^\infty d\nu\int_{-1}^1 d\mu I(x,t;\mathbf{n},\nu), \tag{7}$$

$$q=\frac{1}{2}\int_0^\infty d\nu\int_{-1}^1 d\mu\mu I(x,t;\mathbf{n},\nu), \tag{8}$$

and

$$p^{\rm rad}=\frac{1}{2}\int_0^\infty d\nu\int_{-1}^1 d\mu\mu^2 I(x,t;\mathbf{n},\nu), \tag{9}$$

where $\mu=\cos\theta$. In classical radiative transfer theory, the specific intensity of the radiation field, $I(x,t;\mathbf{n},\nu)$, at the position x and time t, traveling in the direction $\mathbf{n}$ with a frequency ν, is defined so that

$$d\mathscr{E}=I(x,t;\mathbf{n},\nu)dS\cos\alpha d\vartheta d\nu dt, \tag{10}$$

is the energy crossing a surface element dS, into solid angle $d\vartheta$ around $\mathbf{n}$ (α is the angle between $\mathbf{n}$ and the normal to dS), transported by radiation of frequencies $(\nu,\nu+d\nu)$, in time dt (see Ref. 18 for details).

For a nonrotating observer, the radiation portion of the stress–energy tensor reads as[18,19]

$$\hat{T}^R_{\mu\nu}=\begin{pmatrix}\rho^{\rm rad} & -q & 0 & 0\\ -q & p^{\rm rad} & 0 & 0\\ 0 & 0 & \frac{1}{2}(\rho^{\rm rad}-p^{\rm rad}) & 0\\ 0 & 0 & 0 & \frac{1}{2}(\rho^{\rm rad}-p^{\rm rad})\end{pmatrix}. \tag{11}$$

The radiation part of the energy momentum tensor as seen by an observer comoving with the fluid can be found by means of a local rotation to (11),

$$\hat{T}^R_{\mu\nu}=\begin{pmatrix}\rho^{\rm rad}+D^2p_\perp^{\rm rad} & -q & 0 & DG\\ -q & p^{\rm rad} & 0 & -Dq\\ 0 & 0 & p_\perp^{\rm rad} & 0\\ DG & -Dq & 0 & D^2\rho^{\rm rad}+p_\perp^{\rm rad}\end{pmatrix}, \tag{12}$$

where D is an unknown function of u, r and θ associated with the local dragging of inertial frames effect, $p_\perp^{\rm rad}=\frac{1}{2}(\rho^{\rm rad}-p^{\rm rad})$, and $G=\frac{1}{2}(3\rho^{\rm rad}-p^{\rm rad})$.

The material part of the energy–momentum tensor for this observer is given by

$$\hat{T}^{\rm M}_{\mu\nu}=(\rho_{\rm M}+p_\perp)\hat{U}_\mu\hat{U}_\nu-p_\perp\eta_{\mu\nu}+(p-p_\perp)\hat{s}_\mu\hat{s}_\nu, \tag{13}$$

where the Minkowski metric is denoted by $\eta_{\mu\nu}$, $\hat{s}_\mu=\delta^x_\mu$, and $\hat{U}_\mu=\delta^t_\mu$. Thus, the energy–momentum tensor, as seen by a Minkowskian observer comoving with the fluid, can be written as

$$\hat{T}_{\mu\nu}=\hat{T}^R_{\mu\nu}+\hat{T}^{\rm M}_{\mu\nu}. \tag{14}$$

In the slow rotation limit D is taken up to first order. Thus, in virtue of (12) and (13), (14) $\hat{T}_{\mu\nu}$ can be expressed as

$$\hat{T}_{\mu\nu}=(\rho+p_\perp)\hat{U}_\mu\hat{U}_\nu-P_\perp\eta_{\mu\nu}+(p-p_\perp)\hat{s}_\mu\hat{s}_\nu+2\hat{q}_{(\mu}\hat{U}_{\nu)}+2\hat{q}_{(\mu}\hat{D}_{\nu)}+2G\hat{U}_{(\mu}\hat{D}_{\nu)}, \tag{15}$$

where $\hat{q}_\mu=-q\hat{s}_\mu$, $\hat{D}_\mu=D\delta^z_\mu$, $\rho=\rho^{\rm rad}+\rho^{\rm mat}$ is the total energy density, and $p=p^{\rm mat}+p^{\rm rad}$ and $p_\perp=p_\perp^{\rm mat}+p_\perp^{\rm rad}$ are the total radial pressure and the total tangential pressure, respectively.

It remains to express the energy–momentum tensor in curvilinear coordinates (1), as seen by an observer at rest with respect to the Minkowskian coordinates given by (2). Thus, we apply a Lorentz boost and the coordinate transformation defined in (2). The boost velocity is, in the rotating case, $\boldsymbol{\omega}=(\omega_x,0,\omega_z)$—see Ref. 20 for details. Assuming a slow rotation limit, D, a, and ω_z are taken up to first order. Thus,

$$T_{\mu\nu}=(\rho+p_\perp)U_\mu U_\nu-p_\perp g_{\mu\nu}+(p-p_\perp)s_\mu s_\nu+2q_{(\mu}U_{\nu)}+2q_{(\mu}D_{\nu)}+2GU_{(\mu}D_{\nu)}, \tag{16}$$

where $g_{\mu\nu}$ is given by (1),

$$U_\mu=\gamma Y\delta^u_\mu+\frac{\gamma(1-\omega_x)}{X}\,\delta^r_\mu+\gamma\left[a\,\sin^2\,\theta\left(\frac{1-\omega_x}{X}-Y\right)-\omega_z r\,\sin\,\theta+\mathcal{O}(\omega_z^2)\right]\delta^\phi_\mu, \tag{17}$$

$$\begin{aligned}s_\mu=&-\gamma\omega_x Y\delta^u_\mu+\left(\frac{\gamma(1-\omega_x)}{X}+\mathcal{O}(\omega_z^2)\right)\delta^r_\mu\\&+\left[r\,\sin\,\theta\,\frac{\omega_z}{\omega_x}\,(\gamma-1)+\frac{\gamma a\,\sin^2\,\theta}{X}\,[1-\omega_x(1-YX)]+\mathcal{O}(\omega_z^2)\right]\delta^\phi_\mu,\end{aligned} \tag{18}$$

$$q_\mu=-qs_\mu, \tag{19}$$

$$D_\mu=\mathcal{O}(\omega_z^2)\,\delta^u_\mu+\mathcal{O}(\omega_z^2)\,\delta^r_\mu+[Dr\,\sin\,\theta+\mathcal{O}(\omega_z^2)]\delta^\phi_\mu, \tag{20}$$

$$\omega=\sqrt{\omega_x^2+\omega_z^2}=\omega_x+\mathcal{O}(\omega_z^2),$$

and

$$\gamma=\frac{1}{\sqrt{1-\omega^2}}=\frac{1}{\sqrt{1-\omega_x^2}}+\mathcal{O}(\omega_z^2). \tag{21}$$

Here, $\mathcal{O}(\omega_z^2)$ corresponds to quadratic and higher terms in ω_z, D, and a.

III. DEPARTURE FROM COMPLETE EQUILIBRIUM

As has been mentioned in the Introduction, we assume that, before perturbation, the slowly rotating system is evolving along a sequence of states in which it is close complete equilibrium. Therefore, u derivatives of ω_x and q can be neglected because it is close to hydrostatic equilibrium (along the r coordinate) and nearly thermally adjusted. A system that is thermally adjusted changes its properties considerably only within a time scale $\tau_{\rm cha}$ that is large as compared with the Kelvin–Helmholtz time scale $\tau_{\rm KH}$ (Ref. 1, p. 66). Thus, before perturbation we can assume that the u derivatives of ρ, p, and $p_\perp$ are small, and consequently ω_x too (i.e., we can neglect quadratic and higher terms in ω_x). On the other hand, if the system is close to hydrostatic equilibrium, then the hydrostatic time scale $\tau_{\rm hyd}\sim\sqrt{r^3/m}$ is much shorter than the Kelvin–Helmholtz time scale $\tau_{\rm KH}\sim m^2/2rl$, and inertial terms in the equation of motion $T^\mu_{r;\mu}=0$ can be ignored. This condition will be accomplished for small values of luminosity l, and consequently for small values of q. Thus, before perturbation $q\sim\mathcal{O}(\omega_x)$.

We shall evaluate the system just after perturbation. Where, as stated before, *just after perturbation* means on a time scale of the order of the relaxation time. Physically, this implies that the

perturbed quantities (ω_x and q) are still much less than unity. Nevertheless, the system is departing form complete equilibrium and the u derivatives of ω_x and q must be small but different from zero [i.e., $q_{,0}\sim\omega_{x,0}\sim\mathscr{O}(\omega_x)$].

Thus, the system is characterized by the following: Before perturbation,

$$\rho_{,0}\sim p_{,0}\sim p_{\perp,0}\sim\omega_x\sim q\sim\mathscr{O}(\omega_x), \tag{22}$$

$$\omega_{x,0}\sim q_{,0}\sim\mathscr{O}(\omega_x^2); \tag{23}$$

after perturbation,

$$\rho_{,0}\sim p_{,0}\sim p_{\perp,0}\sim\omega_x\sim q\sim\omega_{x,0}\sim q_{,0}\sim\mathscr{O}(\omega_x). \tag{24}$$

In order to clarify the existence of a critical point in slowly rotating fluids, we shall use conservation equations ($T^{\mu}_{\nu;\mu}=0$).

A. Conservation equations

Before perturbation, conservation equations read as

$$R_\nu := T^{\mu}_{\nu;\mu}=0. \tag{25}$$

After perturbation, physical quantities contained in (16) remain unchanged since we are evaluating the system on a time scale of the order of the relaxation time. Therefore, the only new terms appearing in conservation equations are those containing u derivatives of ω_x and q, and conservation equations can be written as

$$\tilde{T}^{\mu}_{\nu;\mu}=\tilde{R}_\nu+\tilde{\omega}_{x,0}\mathscr{F}_\nu+\tilde{q}_{,0}\mathscr{G}_\nu=0, \tag{26}$$

where an overtilde denotes that the quantity is evaluated after perturbation, and $\mathscr{F}_\nu$ and $\mathscr{G}_\nu$ do not depend on ω_x, q, or u derivatives of physical variables, since we are applying first-order perturbation theory. The only terms that can contain $\tilde{\omega}_{x,0}$ and $\tilde{q}_{,0}$ in $\tilde{T}^{\mu}_{\nu;\mu}=0$ are of the form $\tilde{T}^{0}_{\nu,0}$. By means of (1), (17–20), and (16), we find four equations of the form (26)—see Appendix A for details—

$$\begin{aligned}
\tilde{R}_u&=(\rho+p)\tilde{\omega}_{x,0}+\tilde{q}_{,0},\\
\tilde{R}_r&=\frac{2}{XY}\left[(\rho+p)\tilde{\omega}_{x,0}+\tilde{q}_{,0}\right],\\
\tilde{R}_\theta&=0,\\
\tilde{R}_\phi&=0.
\end{aligned} \tag{27}$$

Note that $\tilde{R}_\phi$ is not the total meridional force acting on a given fluid element since it contains terms in $\tilde{\omega}_{z,0}$. Nevertheless, $\tilde{R}_r$ does not contain u-derivatives of physical variables. Thus, $\tilde{R}_r>0$ is the total outward force (pressure gradient+gravitational) along the r-coordinate acting on a given fluid element after perturbation.

The u derivative of the heat flow $\tilde{q}_{,0}$ can be connected with $\tilde{\omega}_{x,0}$ by means of an adequate heat transport equation.

B. Heat transport equation

As it is well known, the Eckart–Landau transport equation[6,7] assumes a vanishing relaxation time. This fact leads to undesirable predictions: An infinite speed for the propagation of the thermal signals and unstable equilibrium states.[21] Thus, it is necessary to adopt a relativistic thermodynamic theory leading to a hyperbolic equation for the propagation of thermal signals. On

the other hand, we are evaluating the system *just after* its departure form hydrostatic equilibrium and thermal adjustment (in the sense described above). Thus, to be consistent with this choice we must use a heat transport equation with a nonvanishing relaxation time.

We shall use the Israel–Stewart relativistic transport equation.[8,9] For viscous free fluid distributions, this one can be written as[22]

$$\tau h^{\mu\nu}U^{\alpha}q_{\nu;\alpha}+q^{\mu}=\kappa h^{\mu\nu}(T_{,\nu}-TU^{\alpha}U_{\nu;\alpha})-\frac{1}{2}\,\kappa T^{2}\left(\frac{\tau}{\kappa T^{2}}\,U^{\beta}\right)_{;\beta}q^{\mu}+\tau\omega^{\mu\nu}q_{\nu}, \tag{28}$$

where κ, τ, and T denote thermal conductivity, thermal relaxation time, and temperature, respectively, $h^{\mu\nu}=U^{\mu}U^{\nu}-g^{\mu\nu}$ is the projector onto the hypersurface orthogonal to U^{μ} and $\omega_{\mu\nu}=h^{\alpha}_{\mu}h^{\beta}_{\nu}U_{[\alpha;\beta]}$ is the vorticity.

Before perturbation, transport equations (28) can be symbolized as

$$\mathscr{H}^{\mu}=0. \tag{29}$$

Just after perturbation, (28) can be written as

$$\tilde{\mathscr{H}}^{\mu}+\tilde{\omega}_{x,0}\mathscr{T}^{\mu}+\tilde{q}_{,0}\mathscr{T}^{\mu}=0. \tag{30}$$

Nevertheless, physical quantities contained on $\mathscr{H}^{\mu}$ do not change just after perturbation, and $\tilde{\mathscr{H}}^{\mu}=\mathscr{H}^{\mu}$. Thus, from (29), expression (30) takes the form

$$\tilde{\omega}_{x,0}\mathscr{T}^{\mu}+\tilde{q}_{,0}\mathscr{T}^{\mu}=0. \tag{31}$$

Vectors $\mathscr{T}^{\mu}$ and $\mathscr{T}^{\mu}$, as $\mathscr{T}_{\nu}$ and $\mathscr{G}_{\nu}$ in the preceding section, do not depend on ω_x, q, or u derivatives of physical variables.

The components of (28) containing u-derivatives of q and ω_x up to first order are

$$\tau h^{\mu\nu}U^{\alpha}q_{\nu;\alpha} \tag{32}$$

and

$$-\kappa T h^{\mu\nu}U^{\alpha}U_{\nu;\alpha}. \tag{33}$$

Therefore, the heat transport equation (28) just after perturbation is given by—see Appendix B for details—

$$\tilde{q}_{,0}=-\frac{\kappa T}{\tau}\,\tilde{\omega}_{x,0}, \tag{34}$$

for any value of μ.

C. Equation of motion

We are now in a position to find the equation of motion just after perturbation. From (27) and (34), we can write

$$\tilde{R}_{r}=\frac{2(\rho+p)}{XY}\,(1-\alpha)\tilde{\omega}_{x,0}, \tag{35}$$

where

$$\alpha=\frac{\kappa T}{\tau(\rho+p)}. \tag{36}$$

As it has been noted in Sec. III A, $\tilde{R}_{r}>0$ is the total outward force along the r-coordinate acting on a given fluid element. Note that it vanishes for $\alpha=1$ (the critical point). This fact has an important consequence: For $\alpha=1$, $\tilde{R}_{r}$ vanishes, even though the u-derivative of the radial velocity

is different from zero. This method also predicts an anomalous behavior beyond the critical point—Eq. (35). If $\alpha>1$, then an outward force ($\tilde{R}_r>0$) implies an inward acceleration ($\tilde{\omega}_{x,0}<0$). Therefore, it seems that, for systems out of the quasistatic approximation, first-order perturbation theory cannot be applied close to the critical point or beyond it.

On the other hand, (35) may be compared with the Newtonian form

$$Force = mass \times acceleration, \tag{37}$$

where here, the term

$$\frac{2(\rho+p)}{XY}(1-\alpha), \tag{38}$$

stands for the *effective inertial mass*. Below the critical point, this *inertial mass* decreases as α grows up. This seems to be connected with the dynamical stability of the system, leading to a minimum stability for $\alpha=1$.[23]

IV. CONCLUSIONS

In this paper we have studied the departure, of a slowly rotating fluid distribution, from a state close to hydrostatic equilibrium (along the r coordinate) and nearly thermally adjusted. Our aim has been to elucidate the existence of a critical point similar to that found for nonrotating systems.[2–5] The existence of this critical point implies that the first-order perturbative method is not always satisfactory to study prerelaxation processes (i.e., processes that take place on time scales smaller than the hydrostatic time scale).

We have found that, also in this case, there exists such a critical point. This one is given by the condition

$$\alpha=\frac{\kappa T}{\tau(\rho+p)}=1, \tag{39}$$

and it coincides with this one found in a spherically symmetric case[2–4] and in an axially symmetric case.[5] Therefore, condition $\alpha=1$ establishes an upper limit for which prerelaxation processes can be studied by means of a first-order perturbative method. This result is also valid if the initial system configuration is strictly in complete equilibrium and radially static (i.e., $\omega_x=\omega_{x,0}=q=q_{,0}=0$).

Note that this method predicts, for values of α less than unity, that the effective inertial mass decreases as α grows. Intuitively, this means that the departure from equilibrium or quasiequilibrium will be steeper for larger α's, or, in other words, that the smaller the α, the larger the dynamical stability. This point has been recently illustrated, by means of an expression for the active gravitational mass in terms of α.[23] It is interesting to emphasize that, at least in nonrotating configurations, causality and stability conditions[21] do not always forbid the existence of the critical point.[3,4]

Finally, it is also worth noticing that the critical point and the inflationary equation of state for nondissipative systems ($p=-\rho$) are similar, in that they imply the vanishing of the inertial mass term. Therefore one might wonder about the plausibility of an inflationary scenario in a Universe at, or close to, the critical point.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Ministry of Education under Grant No. PB94-0718.

APPENDIX A: CONSERVATION EQUATIONS JUST AFTER PERTURBATION

In the slow rotating limit, U^μ, s^μ, and D^μ read as

$$U^\mu=\left(\frac{\gamma(1-\omega_x)}{Y}+\mathcal{O}(\omega_z^2)\right)\delta^\mu_u+[\omega_x X+\mathcal{O}(\omega_z^2)]\delta^\mu_r+\left(\frac{\omega_z}{r\ \sin\ \theta}+\mathcal{O}(\omega_z^2)\right)\delta^\mu_\phi, \tag{A1}$$

$$s^\mu=\left(\frac{\gamma(1-\omega_x)}{Y}+\mathcal{O}(\omega_z^2)\right)\delta^\mu_u+[-\gamma X+\mathcal{O}(\omega_z^2)]\delta^\mu_r+\left[-\frac{\omega_z}{r\ \sin\ \theta}\left(\frac{\gamma-1}{\omega_x}\right)+\mathcal{O}(\omega_z^2)\right]\delta^\mu_\phi, \tag{A2}$$

$$D^\mu=\mathcal{O}(\omega_z^2)\,\delta^\mu_u+\mathcal{O}(\omega_z^2)\,\delta^\mu_r+\left(-\frac{D}{r\ \sin\ \theta}+\mathcal{O}(\omega_z^2)\right)\delta^\mu_\phi, \tag{A3}$$

and $q^\mu=-qs^\mu$. From (16) the only terms that contain u derivatives of ω_x and q (up to first order in ω_x and ω_z) in conservation equations $T^\mu_{\nu;\mu}=0$ are of the form $T^0_{\nu,0}$. In particular,

$$\begin{aligned}
(\rho+p_\perp)(U^0U_\nu)_{,0}&=(\rho+p_\perp)\left(\frac{\gamma(1-\omega_x)}{Y}\,U_\nu\right)_{,0},\\
(p-p_\perp)(s^0s_\nu)_{,0}&=(p-p_\perp)\left(\frac{\gamma(1-\omega_x)}{Y}\,s_\nu\right)_{,0},\\
G(U^0D_\nu+U_\nu D^0)_{,0}&=G\left(\frac{\gamma(1-\omega_x)}{Y}\,D_\nu\right)_{,0},\\
(q^0U_\nu+q^0D_\nu+q_\nu U^0+q_\nu D^0)_{,0}&=-q_{,0}\left(\frac{\gamma(1-\omega_x)}{Y}\,(U_\nu+D_\nu+s_\nu)\right)\\
&\quad-q\left(\frac{\gamma(1-\omega_x)}{Y}\,(U_\nu+D_\nu+s_\nu)\right)_{,0}.
\end{aligned} \tag{A4}$$

Thus, from (17)–(20), (21), (A4), and following the definition of $\mathcal{F}_\nu$ and $\mathcal{G}_\nu$ given in (26), we find, up to first order in ω_x,

$$\begin{aligned}
\omega_{x,0}\mathcal{F}_u&=(\rho+p_\perp)\left(\frac{1}{1+\omega_x}\right)_{,0}-(p-p_\perp)\left(\frac{\omega_x}{1+\omega_x}\right)_{,0}-q\left(\frac{1-\omega_x}{1+\omega_x}\right)_{,0}=-\omega_{x,0}(\rho+p),\\
\omega_{x,0}\mathcal{F}_r&=\left(\frac{\rho+p-2q}{XY}\right)\left(\frac{1-\omega_x}{1+\omega_x}\right)_{,0}=-\omega_{x,0}\,\frac{2(\rho+p)}{XY},\\
\omega_{x,0}\mathcal{F}_\theta&=0,\quad \omega_{x,0}\mathcal{F}_\phi=0,
\end{aligned} \tag{A5}$$

$$\begin{aligned}
q_{,0}\mathcal{G}_u&=-q_{,0},\quad q_{,0}\mathcal{G}_r=-q_{,0}\,\frac{2}{XY},\\
q_{,0}\mathcal{G}_\theta&=0,\quad q_{,0}\mathcal{G}_\phi=0,
\end{aligned} \tag{A6}$$

where we have neglected in $\omega_{x,0}\mathcal{F}_\phi$ terms of the form $\omega_{x,0}D$, $\omega_{x,0}a$, and $\omega_{x,0}\omega_z$, and in $q_{,0}\mathcal{G}_\phi$ terms of the form $q_{,0}D$, $q_{,0}a$, and $q_{,0}\omega_z$. Thus,

$$\mathcal{F}_\mu=-(\rho+P)(1,2/XY,0,0), \tag{A7}$$

and

$$\mathcal{G}_\mu=-(1,2/XY,0,0). \tag{A8}$$

The expression of conservation equations can be found by means of (26):

$$\tilde{R}_\nu=[(\rho+p)\tilde{\omega}_{x,0}+\tilde{q}_{,0}](1,2/XY,0,0). \tag{A9}$$

APPENDIX B: HEAT TRANSPORT EQUATION JUST AFTER PERTURBATION

The right-hand terms in (28),

$$-\frac{1}{2}\kappa T^2\left(\frac{\tau}{\kappa T^2}U^\beta\right)_{;\beta}q^\alpha,$$

and

$$\tau\omega^{\mu\nu}q_\nu,$$

contain factors of the form $\omega_{x,0}q$, that are of second order. Therefore, the only terms in (28) that contain u derivatives of ω_x and q up to first order are of the form

$$\tau h^{\mu\nu}U^0q_{\nu,0} \tag{B1}$$

and

$$-\kappa T h^{\mu\nu}U^0U_{\nu,0}. \tag{B2}$$

Using (A1), (18), and (19) in (B1) and (B2),

$$\begin{aligned}\tau h^{\mu\nu}U^0q_{\nu,0}&=\frac{\gamma(1-\omega_x)}{Y}\tau h^{\mu\nu}(-q_{,0}s_\nu-qs_{\nu,0}),\\ -\kappa T h^{\mu\nu}U^0U_{\nu,0}&=-\frac{\gamma(1-\omega_x)}{Y}\kappa T h^{\mu\nu}U_{\nu,0}.\end{aligned} \tag{B3}$$

Vectors $\mathscr{T}^\mu$ and $\mathscr{T}^\mu$ (31) do not depend on ω_x, q, and their u derivatives because we are using first-order perturbation theory. Thus, from (B3),

$$\tilde{\omega}_{x,0}\mathscr{T}^\mu=\tilde{\omega}_{x,0}\frac{\kappa T}{XY}h^{\mu r}, \tag{B4}$$

and

$$\tilde{q}_{,0}\mathscr{T}^\mu=\tilde{q}_{,0}\frac{\tau}{XY}h^{\mu r}, \tag{B5}$$

where we have neglected terms of the form $\tilde{q}_{,0}a$. Therefore, from (B4) and (B5), expression (31) takes the form

$$\tilde{q}_{,0}=-\frac{\kappa T}{\tau}\tilde{\omega}_{x,0}, \tag{B6}$$

which is valid for any μ.

[1] R. Kippenhan and A. Weigert, *Stellar Structure and Evolution* (Springer-Verlag Berlin, 1994).
[2] L. Herrera, A. Di Prisco, J. L. Hernández-Pastora, J. Martín, and J. Martínez, Class. Quantum Grav. **14**, 2239 (1997).
[3] L. Herrera and J. Martínez, Class. Quantum Grav. **14**, 2697 (1997).
[4] L. Herrera and J. Martínez, Class. Quantum Grav. **15**, 407 (1998).
[5] L. Herrera, A. Di Prisco, and J. Martínez, ''Breakdown of the linear approximation in the perturbative analysis of heat conduction in relativistic systems,'' to appear in Phys. Lett. A.
[6] C. Eckart, Phys. Rev. **15**, 919 (1940).
[7] L. Landau and E. Lifshitz, *Fluid Mechanics* (Pergamon, London, 1959).
[8] W. Israel, Ann. Phys. (N.Y.) **100**, 310 (1976).

[9] W. Israel and J. M. Stewart, Ann. Phys. (N.Y.) **118**, 341 (1979).
[10] D. Pavón, D. Jou, and J. Casas-Vázquez, Ann. Inst. Henri Poincaré, Sect. A **36**, 79 (1982).
[11] D. Jou, J. Casas-Vázquez, and G. Lebon, *Extended Irreversible Thermodynamics* (Springer-Verlag, Berlin, 1993).
[12] L. Herrera and J. Jiménez, J. Math. Phys. **23**, 2339 (1982).
[13] J. B. Hartle, Astrophys. J. **150**, 1005 (1967).
[14] M. Demiański, *Relativistic Astrophysics*, International Series in Nat. Phyl. V110, edited by D. Ter Haar (Pergamon, Oxford, 1985).
[15] M. Carmeli and M. Kaye, Ann. Phys. (N.Y.) **103**, 97 (1977).
[16] C. González, L. Herrera, and J. Jiménez, J. Math. Phys. **20**, 836 (1979).
[17] D. Kramer and U. Hahner, Class. Quantum Grav. **12**, 2287 (1995).
[18] D. Mihalas and B. Mihalas, *Foundations of Radiation Hydrodynamics* (Oxford University Press, Oxford, 1984).
[19] R. W. Lindquist, Ann. Phys. (N.Y.) **37**, 487 (1966).
[20] L. Herrera, A. Melfo, L. A. Núñez, and A. Patiño, Astrophys. J. **421**, 677 (1994).
[21] W. A. Hiscock and L. Lindblom, Ann. Phys. (N.Y.) **151**, 466 (1983).
[22] R. Maartens and J. Triginer, Phys. Rev. D **56**, 4640 (1997).
[23] L. Herrera and A. Di Prisco, submitted to Gen. Rel. Grav.

On different integrable systems sharing the same nondynamical *r*-matrix

Zhijun Qiao[a)]
Institute of Mathematics and School of Mathematical Science, Peking University, Beijing 100871, Peoples Republic of China

Walter Strampp[b)]
Fachbereich 17, Mathematik, Universitaet-Gh Kassel, Hollaendische Str. 36, 34109 Kassel, Deutschland

(Received 25 November 1997; accepted for publication 6 February 1998)

In a recent paper [Zhijun Qiao and Ruguang Zhou, Phys. Lett. A **235**, 35 (1997)], the amazing fact was reported that a discrete and a continuous integrable system share the same r-matrix with the interesting property of being nondynamical. Now, we present three further pairs of different continuous integrable systems sharing the same r-matrix again being nondynamical. The first pair is the finite-dimensional constrained system (FDCS) of the famous AKNS hierarchy and the Dirac hierarchy; the second pair is the FDCS of the well-known geodesic flows on the ellipsoid and the Heisenberg spin chain hierarchy; and the third pair is the FDCS of one hierarchy studied by Xianguo Geng [Phys. Lett. A **162**, 375 (1992)] and another hierarchy proposed by Zhijun Qiao [Phys. Lett. A **192**, 316 (1994)]. All those FDCSs possess Lax representations and from the viewpoint of r-matrix can be shown to be completely integrable in Liouville's sense. © *1998 American Institute of Physics.* [S0022-2488(98)02506-7]

I. INTRODUCTION

The well-known nonlinearization method[1] is a powerful tool to generate completely integrable finite-dimensional Hamiltonian systems (CIFDHSs) from the eigenvalue problems or Lax pairs of nonlinear evolution equations (NLEEs). With the help of this method many new CIFDHSs have been successively found, and solutions to NLEEs have been given by involutive or parametric representations.[2–12] The r-matrix[13] method is quite an effective approach to the Lie–Poisson structure in $1+1$-dimensional space,[14,15] providing fundamental commutator relations in the quantum inversing scattering.[16] Recently, the study of CIFDHSs admitting an r-matrix has already received attention.[17–22] It has been shown that r-matrices play a very important part in proving the integrability of CIFDHSs. Besides, the Lax matrix associated with the r-matrix of a CIFDHS can be applied for obtaining algebraic–geometric solutions of soliton equations or NLEEs.[23]

In a recent paper,[24] we reported an interesting and surprising result: a discrete and a continuous integrable system possess the same nondynamical r-matrix being independent from the dynamical variables p,q. The question of whether there are other pairs of continuous or discrete finite-dimensional integrable systems sharing the same r-matrix which should be nondynamical arises. In the present paper, on the basic idea of Ref. 24 we have considered a wide class of 2×2 spectral problems and finally succeeded in finding three other pairs of different continuous integrable systems with the common nondynamical r-matrix. The first pair is formed by the finite-dimensional constrained system (FDCS) of the famous AKNS and Dirac hierarchies; the second pair is formed by the FDCS of the well-known geodesic flow on the ellipsoid and the Heisenberg spin chain hierarchy; while the third pair consists of the FDCS of a hierarchy studied by Xianguo Geng[7] and a hierarchy proposed by Zhijun Qiao.[11] All of these FDCSs possess Lax representations, and by using the r-matrix approach can be shown to be completely integrable in Liouville's sense.

a)Electronic mail: qiaozj@sxx0.math.pku.edu.cn
b)Electronic mail: strampp@hrz.uni-kassel.de

 © 1998 American Institute of Physics

Before displaying these results, let us first introduce some basic symbols and notations. Let $(R^{2N},dp\wedge dq)$ stand for the standard symplectic structure in the Euclidic space $R^{2N}=\{(p,q)|p=(p_1,\ldots,p_N),\ q=(q_1,\ldots,q_N)\}$, where p_i,q_i $(i=1,...,N)$ are N pairs of canonical coordinates. The standard inner product in R^N will be denoted by $\langle\cdot,\cdot\rangle$ and the Poisson bracket of two Hamiltonian functions F,G is given through[25]

$$\{F,G\}=\sum_{i=1}^{N}\left(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i}-\frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i}\right)=\left\langle\frac{\partial F}{\partial q},\frac{\partial G}{\partial p}\right\rangle-\left\langle\frac{\partial F}{\partial p},\frac{\partial G}{\partial q}\right\rangle.$$

With N arbitrary distinct constants $\lambda_1,\ldots,\lambda_N$ we form the diagonal matrix $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_N)$, while λ, μ are used for two different spectral parameters. By $C^\infty(R)$ we denote the set of all functions on real field R being infinitely many times differentiable. Finally, x stands for the spatial continuous variable.

II. THE CONSTRAINED AKNS AND DIRAC (D) SYSTEM

We consider the following two 2×2 traceless Lax matrices:

$$L^{\mathrm{AKNS}}=L^{\mathrm{AKNS}}(\lambda)=\begin{pmatrix}1&0\\0&-1\end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix}p_jq_j&-q_j^2\\p_j^2&-p_jq_j\end{pmatrix}\equiv\begin{pmatrix}A_{\mathrm{AKNS}}(\lambda)&B_{\mathrm{AKNS}}(\lambda)\\C_{\mathrm{AKNS}}(\lambda)&-A_{\mathrm{AKNS}}(\lambda)\end{pmatrix},\tag{1}$$

$$L^{D}=L^{D}(\lambda)=\begin{pmatrix}0&1\\-1&0\end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix}p_jq_j&-q_j^2\\p_j^2&-p_jq_j\end{pmatrix}\equiv\begin{pmatrix}A_D(\lambda)&B_D(\lambda)\\C_D(\lambda)&-A_D(\lambda)\end{pmatrix}.\tag{2}$$

Let us choose two 2×2 matrices M_{AKNS} and M_D:

$$M_{\mathrm{AKNS}}=\begin{pmatrix}\lambda&-\langle q,q\rangle\\\langle p,p\rangle&-\lambda\end{pmatrix},\tag{3}$$

$$M_D=\begin{pmatrix}\langle p,q\rangle&\lambda+\frac{1}{2}(\langle p,p\rangle-\langle q,q\rangle)\\-\lambda+\frac{1}{2}(\langle p,p\rangle-\langle q,q\rangle)&-\langle p,q\rangle\end{pmatrix}.\tag{4}$$

Then we have the following.

Theorem 1: The Lax equations

$$L_x^{\mathrm{AKNS}}=[M_{\mathrm{AKNS}},L^{\mathrm{AKNS}}],\tag{5}$$

$$L_x^{D}=[M_D,L^{D}],\tag{6}$$

respectively describe the following finite-dimensional Hamiltonian systems (FDHSs) (H_{AKNS}) and (H_D),

$$(H_{\mathrm{AKNS}}):\begin{cases}q_x=-\langle q,q\rangle p+\Lambda q=\dfrac{\partial H_{\mathrm{AKNS}}}{\partial p},\\[2ex]p_x=\langle p,p\rangle q-\Lambda p=-\dfrac{\partial H_{\mathrm{AKNS}}}{\partial q};\end{cases}\tag{7}$$

$$(H_D):\begin{cases}q_x=\langle p,q\rangle q+\frac{1}{2}(\langle p,p\rangle-\langle q,q\rangle)p+\Lambda p=\dfrac{\partial H_D}{\partial p},\\[2ex]p_x=-\langle p,q\rangle p-\frac{1}{2}(\langle p,p\rangle-\langle q,q\rangle)q-\Lambda q=-\dfrac{\partial H_D}{\partial q},\end{cases}\tag{8}$$

with Hamiltonian functions

$$H_{\mathrm{AKNS}}=\langle \Lambda q,p\rangle-\tfrac{1}{2}\langle p,p\rangle\langle q,q\rangle, \tag{9}$$

$$H_D=\tfrac{1}{2}(\langle \Lambda p,p\rangle+\langle \Lambda q,q\rangle)+\tfrac{1}{2}(\langle p,q\rangle^2-\langle q,q\rangle\langle p,p\rangle)+\tfrac{1}{8}(\langle p,p\rangle+\langle q,q\rangle)^2. \tag{10}$$

Proof: By direct calculation.

Through setting

$$u=-\langle q,q\rangle,\quad v=\langle p,p\rangle, \tag{11}$$

Eq. (7) becomes

$$\begin{aligned} q_{jx}&=\lambda_j q_j+up_j,\quad j=1,...,N,\\ p_{jx}&=vq_j-\lambda_j p_j,\quad j=1,...,N, \end{aligned} \tag{12}$$

which is nothing other than the well-known Zakharov–Shabat–AKNS spectral problem (ZS-AKNSSP)[26]

$$y_x=\begin{pmatrix}\lambda & u\\ v & -\lambda\end{pmatrix}y \tag{13}$$

where $\lambda=\lambda_j$, $y=(q_j,p_j)^T$. The potentials u,v defined by Eq. (11) are correspond exactly to the Bargmann–Garnier constraint[1(b)]

$$G_0=(v,u)^T=\sum_{j=1}^{N}\left(\frac{\delta\lambda_j}{\delta u},\frac{\delta\lambda_j}{\delta v}\right)^T \tag{14}$$

of ZS-AKNSSP (13), where $\delta\lambda_j/\delta u$, $\delta\lambda_j/\delta v$ are the two spectral gradients of spectral parameters λ_j with respect to the potentials u and v. Therefore Eq. (7) coincides with the constrained AKNS system (c-AKNSS).

Similarly, after setting

$$u=-\tfrac{1}{2}(\langle p,p\rangle-\langle q,q\rangle),\quad v=-\langle p,q\rangle, \tag{15}$$

it is easily seen that Eq. (8) becomes the Dirac spectral problem (DSP)[27]

$$y_x=\begin{pmatrix}-v & \lambda-u\\ -\lambda-u & v\end{pmatrix}y \tag{16}$$

with $\lambda=\lambda_j$, $y=(q_j,p_j)^T$. Thus, Eq. (8) is the constrained Dirac system (c-DS).

Let us return to the Lax matrix (1) and (2). By a simple calculation we obtain the following.

Proposition 1: For $J=\mathrm{AKNS}$ and $J=D$ the following holds:

$$\begin{aligned} &\{A_J(\lambda),A_J(\mu)\}=\{B_J(\lambda),B_J(\mu)\}=\{C_J(\lambda),C_J(\mu)\}=0,\\ &\{A_J(\lambda),B_J(\mu)\}=\frac{2}{\mu-\lambda}(B_J(\mu)-B_J(\lambda)),\\ &\{A_J(\lambda),C_J(\mu)\}=\frac{2}{\mu-\lambda}(C_J(\lambda)-C_J(\mu)),\\ &\{B_J(\lambda),C_J(\mu)\}=\frac{4}{\mu-\lambda}(A_J(\mu)-A_J(\lambda)). \end{aligned} \tag{17}$$

Set $L_1^J(\lambda)=L^J(\lambda)\otimes I$, $L_2^J(\mu)=I\otimes L^J(\mu)$, where I is the 2×2 unit matrix. Then the above proposition yields the following.

Theorem 2: The Lax matrices $L^J(\lambda)$ ($J=$AKNS,D) defined by Eqs. (1) and (2) satisfy the fundamental Poisson bracket

$$\{L_1^J(\lambda)\overset{\otimes}{,}L_2^J(\mu)\}=[r_{12}(\lambda,\mu),L_1^J(\lambda)]-[r_{21}(\mu,\lambda),L_2^J(\mu)]. \tag{18}$$

Here $\{L_1^J(\lambda)\overset{\otimes}{,}L_2^J(\mu)\}$ is a 4×4 matrix,[14] $[\cdot,\cdot]$ is the usual commutator of the matrix, and the r-matrices $r_{12}(\lambda,\mu)$, $r_{21}(\mu,\lambda)$ are exactly given by the following standard r-matrices:

$$r_{12}(\lambda,\mu)=\frac{2}{\mu-\lambda}P,\quad r_{21}(\mu,\lambda)=Pr_{12}(\mu,\lambda)P, \tag{19}$$

$$P=\begin{pmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{pmatrix}=\frac{1}{2}\left(I+\sum_{i=1}^{3}\sigma_i\otimes\sigma_i\right) \tag{20}$$

(here σ_i are the Pauli matrices).

So, the c-AKNSS (7) and c-DS (8) share the same standard r-matrix (19), which is obviously nondynamical.

Remark 1: In fact, since the r-matrix relation (18) is concerned only with the commutator, the r-matrix $r_{12}(\lambda,\mu)$ satisfying (18) in the case of the c-AKNSS (7) and c-DS (8) can be also chosen as

$$r_{12}(\lambda,\mu)=\frac{2}{\mu-\lambda}P+I\otimes\tilde{S},\quad \tilde{S}=\begin{pmatrix}a&b\\c&d\end{pmatrix}, \tag{21}$$

where the elements a,b,c,d can be arbitrary functions $a(\lambda,\mu,p,q)$, $b(\lambda,\mu,p,q)$, $c(\lambda,\mu,p,q)$, $d(\lambda,\mu,p,q)\in C^\infty(R)$ with respect to the spectral parameters λ, μ and the dynamical variables p,q. This shows that for a given Lax matrix, the associated r-matrix is not uniquely defined (there are even infinitely many r-matrices possible). Here we give the simplest case: $a=b=c=d=0$, i.e., the standard r-matrix (19).

III. THE CONSTRAINED HARRY–DYM (HD) AND HEISENBERG SPIN CHAIN (HSC) SYSTEM

The constrained Harry–Dym (HD) system describes the geodesic flow on an ellipsoid and shares the same r-matrix with the constrained Heisenberg spin chain (HSC).

To see this, we follow the process given in Sec. II, considering the following Lax matrices:

$$\begin{aligned}L^{\mathrm{HD}}=L^{\mathrm{HD}}(\lambda)&=\begin{pmatrix}-\langle p,q\rangle\lambda^{-1}&\lambda^{-2}+\langle q,q\rangle\lambda^{-1}\\-\langle p,p\rangle\lambda^{-1}&\langle p,q\rangle\lambda^{-1}\end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix}p_jq_j&-q_j^2\\p_j^2&-p_jq_j\end{pmatrix}\\&\equiv\begin{pmatrix}A_{\mathrm{HD}}(\lambda)&B_{\mathrm{HD}}(\lambda)\\C_{\mathrm{HD}}(\lambda)&-A_{\mathrm{HD}}(\lambda)\end{pmatrix},\end{aligned} \tag{22}$$

$$\begin{aligned}L^{\mathrm{HSC}}=L^{\mathrm{HSC}}(\lambda)&=\begin{pmatrix}-\langle p,q\rangle\lambda^{-1}&\langle q,q\rangle\lambda^{-1}\\-\langle p,p\rangle\lambda^{-1}&\langle p,q\rangle\lambda^{-1}\end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix}p_jq_j&-q_j^2\\p_j^2&-p_jq_j\end{pmatrix}\\&\equiv\begin{pmatrix}A_{\mathrm{HSC}}(\lambda)&B_{\mathrm{HSC}}(\lambda)\\C_{\mathrm{HSC}}(\lambda)&-A_{\mathrm{HSC}}(\lambda)\end{pmatrix}.\end{aligned} \tag{23}$$

Set

$$M_{\mathrm{HD}}=\begin{pmatrix} 0 & 1 \\ -\dfrac{\langle \Lambda p,p\rangle}{\langle \Lambda^2 q,q\rangle}\lambda & 0 \end{pmatrix}, \tag{24}$$

$$M_{\mathrm{HSC}}=\begin{pmatrix} -i\lambda\langle \Lambda p,q\rangle & i\lambda\langle \Lambda q,q\rangle \\ -i\lambda\langle \Lambda p,p\rangle & i\lambda\langle \Lambda p,q\rangle \end{pmatrix}, \quad i^2=-1. \tag{25}$$

Then a direct calculation leads to the following theorem.

Theorem 3: The Lax representations

$$L_x^{\mathrm{HD}}=[M_{\mathrm{HD}},L^{\mathrm{HD}}], \tag{26}$$

$$L_x^{\mathrm{HSC}}=[M_{\mathrm{HSC}},L^{\mathrm{HSC}}], \tag{27}$$

respectively give the following FDHS:

$$(H_{\mathrm{HD}}):\begin{cases} q_x=p=\dfrac{\partial H_{\mathrm{HD}}}{\partial p}\Big|_{TQ^{N-1}}, \\ p_x=-\dfrac{\langle \Lambda p,p\rangle}{\langle \Lambda^2 q,q\rangle}\Lambda q=-\dfrac{\partial H_{\mathrm{HD}}}{\partial q}\Big|_{TQ^{N-1}}, \\ \langle \Lambda q,q\rangle=1; \end{cases} \tag{28}$$

$$(H_{\mathrm{HSC}}):\begin{cases} q_x=i\langle \Lambda q,q\rangle\Lambda p-i\langle \Lambda p,q\rangle\Lambda q=\dfrac{\partial H_{\mathrm{HSC}}}{\partial p}, \\ p_x=i\langle \Lambda p,q\rangle\Lambda p-i\langle \Lambda p,p\rangle\Lambda q=-\dfrac{\partial H_{\mathrm{HSC}}}{\partial q}, \end{cases} \tag{29}$$

with the Hamiltonian functions

$$H_{\mathrm{HD}}=\frac{1}{2}\langle p,p\rangle-\frac{\langle \Lambda p,p\rangle}{2\langle \Lambda^2 q,q\rangle}(\langle \Lambda q,q\rangle-1), \tag{30}$$

$$H_{\mathrm{HSC}}=\tfrac{1}{2}\,i\langle \Lambda p,p\rangle\langle \Lambda q,q\rangle-\tfrac{1}{2}\,i\langle \Lambda p,q\rangle^2. \tag{31}$$

Here, in Eq. (28) TQ^{N-1} is a tangent bundle in R^{2N}:

$$TQ^{N-1}=\{(p,q)\in R^{2N}|F\equiv\langle \Lambda q,q\rangle-1=0,G\equiv\langle \Lambda p,q\rangle=0\}. \tag{32}$$

Obviously, Eq. (28) is equivalent to

$$q_{xx}+\frac{\langle \Lambda q_x,q_x\rangle}{\langle \Lambda^2 q,q\rangle}\Lambda q=0, \quad \langle \Lambda q,q\rangle=1, \tag{33}$$

which is nothing but the equation of the geodesic flow on the surface $\langle \Lambda q,q\rangle=1$ in R^N space.[3,22] On the other hand, by setting

$$u=\frac{\langle \Lambda q_x,q_x\rangle}{\langle \Lambda^2 q,q\rangle}, \tag{34}$$

Eq. (33) becomes the well-known Harry–Dym spectral problem

$$y_{xx}+\lambda uy=0 \tag{35}$$

with $\lambda=\lambda_j$, $y=q_j$, $j=1,...,N$. Simultaneously, Eq. (34) gives the constraint condition considered by Cao,[3] so that Eq. (28) coincides with the constrained HD system (c-HDS).

In addition, after setting

$$u=-\langle \Lambda q,q\rangle, \quad v=\langle \Lambda p,p\rangle, \quad w=-\langle \Lambda p,q\rangle, \tag{36}$$

we can see that Eq. (29) becomes the Heisenberg spin chain spectral problem[28]

$$y_x=\begin{pmatrix} -i\lambda w & -i\lambda u \\ -i\lambda v & i\lambda w \end{pmatrix} y, \quad i^2=-1, \tag{37}$$

with $\lambda=\lambda_j$, $y=(q_j,p_j)^T$. Thus, Eq. (29) reads as the constrained Heisenberg spin chain system (c-HSCS). The constraint defined by Eq. (36) has been studied and applied to obtaining involutive solutions for the Heisenberg spin chain equations in Ref. 29. However, the r-matrix was not given.

Now, we construct the r-matrix of c-HDS (28) and c-HSCS (29). Their Lax matrices (22) and (23) share all elements except one, namely

$$\begin{pmatrix} 0 & \lambda^{-2} \\ 0 & 0 \end{pmatrix}.$$

Since this element does not affect the calculations concerning the fundamental Poisson bracket $\{L_1^J(\lambda)\overset{\otimes}{,}L_2^J(\mu)\}$, one can readily deduce that the c-HDS and c-HSCS possess the same r-matrix.

Theorem 4: The Lax matrix $L^J(\lambda)$ ($J=$HD,HSC) defined by Eqs. (22) and (23) satisfies the fundamental Poisson bracket (18) with the common nondynamical r-matrix

$$r_{12}(\lambda,\mu)=\frac{2\lambda}{\mu(\mu-1)}P, \quad r_{21}(\mu,\lambda)=Pr_{12}(\mu,\lambda)P. \tag{38}$$

Apparently the r-matrix (38) solves the classical Yang–Baxter equation (CYBE)

$$[r_{ij},r_{ik}]+[r_{ij},r_{jk}]+[r_{kj},r_{ik}]=0, \quad i,j,k=1,2,3. \tag{39}$$

Remark 2: The r-matrix $r_{12}(\lambda,\mu)$ satisfying Eq. (18) in the case of c-HDS (28) and c-HSCS (29) can also be chosen as

$$r_{12}(\lambda,\mu)=\frac{2\lambda}{\mu(\mu-\lambda)}P+I\otimes\tilde{S}. \tag{40}$$

Evidently, Eq. (38) is the simplest case: $\tilde{S}=0$ of Eq. (40).

IV. THE CONSTRAINED *G* AND *Q* SYSTEM

In this section, we introduce the following Lax matrices:

$$\begin{aligned} L^G=L^G(\lambda)&=\begin{pmatrix} (\frac{1}{2}+\langle p,q\rangle)\lambda^{-1} & \langle q,q\rangle\lambda^{-1} \\ 0 & -(\frac{1}{2}+\langle p,q\rangle)\lambda^{-1} \end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix} p_jq_j & -q_j^2 \\ p_j^2 & -p_jq_j \end{pmatrix} \\ &\equiv\begin{pmatrix} A_G(\lambda) & B_G(\lambda) \\ C_G(\lambda) & -A_G(\lambda) \end{pmatrix}, \end{aligned} \tag{41}$$

$$L^Q=L^Q(\lambda)=\begin{pmatrix} -\lambda^{-1} & \langle q,q\rangle\lambda^{-1} \\ 0 & \lambda^{-1} \end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix} p_jq_j & -q_j^2 \\ p_j^2 & -p_jq_j \end{pmatrix}\equiv\begin{pmatrix} A_Q(\lambda) & B_Q(\lambda) \\ C_Q(\lambda) & -A_Q(\lambda) \end{pmatrix}. \tag{42}$$

If we set

$$M_G=\begin{pmatrix} -\frac{1}{\alpha}\lambda & \frac{1}{\alpha}(\langle p,p\rangle-\langle q,q\rangle)-1 \\ \frac{1}{\alpha}(\langle p,p\rangle-\langle q,q\rangle+1)\lambda & \frac{1}{\alpha}\lambda \end{pmatrix}, \tag{43}$$

$$M_Q=\begin{pmatrix} \lambda+\frac{1}{2\beta^2}\langle \Lambda q,q\rangle\langle p,p\rangle & \frac{1}{\beta}\langle \Lambda q,q\rangle \\ -\frac{1}{\beta}\langle p,p\rangle\lambda & -\lambda-\frac{1}{2\beta^2}\langle \Lambda q,q\rangle\langle p,p\rangle \end{pmatrix}, \tag{44}$$

with

$$\alpha=\sqrt{(\langle p,p\rangle-\langle \Lambda q,q\rangle)^2-4\langle \Lambda q,p\rangle},\quad \beta=1-\langle p,q\rangle,$$

then, by a lengthy and straightforward calculation we obtain the following.

Theorem 5: The following Lax representations,

$$L_x^G=[M_G,L^G] \tag{45}$$

and

$$L_x^Q=[M_Q,L^Q], \tag{46}$$

where the first one is restricted to the surface $\Gamma=\{(p,q)\in R^{2N}|\langle p,q\rangle=0,\langle \Lambda q,q\rangle\langle p,p\rangle+\langle \Lambda q,p\rangle=0\}$ in the space R^{2N}, respectively produce the finite-dimensional systems

$$\begin{aligned} q_x&=\frac{1}{\alpha}(-\Lambda q+(\langle p,p\rangle-\langle \Lambda q,q\rangle)p)-p,\\ p_x&=\frac{1}{\alpha}(\Lambda p+(\langle p,p\rangle-\langle \Lambda q,q\rangle)\Lambda q)+\Lambda q, \end{aligned} \tag{47}$$

and

$$\begin{aligned} q_x&=\Lambda q+\frac{1}{\beta}\langle \Lambda q,q\rangle p+\frac{1}{2\beta^2}\langle p,p\rangle\langle \Lambda q,q\rangle q,\\ p_x&=-\Lambda p-\frac{1}{\beta}\langle p,p\rangle\Lambda q-\frac{1}{2\beta^2}\langle p,p\rangle\langle \Lambda q,q\rangle p. \end{aligned} \tag{48}$$

In Eqs. (47) and (48), respectively insert

$$\begin{aligned} u&=\frac{1}{\alpha}=\frac{1}{\sqrt{(\langle p,p\rangle-\langle \Lambda q,q\rangle)^2-4\langle \Lambda q,p\rangle}},\\ v&=\frac{\langle p,p\rangle-\langle \Lambda q,q\rangle}{\alpha}=\frac{\langle p,p\rangle-\langle \Lambda q,q\rangle}{\sqrt{(\langle p,p\rangle-\langle \Lambda q,q\rangle)^2-4\langle \Lambda q,p\rangle}}; \end{aligned} \tag{49}$$

and

$$\begin{aligned} u&=\frac{\langle \Lambda q,q\rangle}{\beta}=\frac{\langle \Lambda q,q\rangle}{1-\langle q,p\rangle}\\ v&=\frac{-\langle p,p\rangle}{\beta}=-\frac{\langle p,p\rangle}{1-\langle q,p\rangle}. \end{aligned} \tag{50}$$

Then, Eqs. (47) and (48) turn out to become the spectral problem studied by Geng (simply called G-spectral problem),[7]

$$y_x=\begin{pmatrix} -\lambda u & v-1 \\ \lambda(v+1) & \lambda u \end{pmatrix} y, \tag{51}$$

with $\lambda=\lambda_j$, $y=(q_j,p_j)^T$, and the spectral problem proposed by Qiao (simply called Q-spectral problem),[11]

$$y_x=\begin{pmatrix} \lambda-\frac{1}{2}uv & u \\ \lambda v & -\lambda+\frac{1}{2}uv \end{pmatrix} y, \tag{52}$$

with $\lambda=\lambda_j$, $y=(q_j,p_j)^T$, respectively. So, Eqs. (47) and (48) are nothing but the constrained Geng system (c-GS) and constrained Qiao system (c-QS) under the constraint conditions (49) and (50). Since the Lax equation (45) gives the c-GS (47) on the surface Γ, the Lax matrix L^G should become

$$L_\Gamma^G=L_\Gamma^G(\lambda)=\begin{pmatrix} \frac{1}{2}\lambda^{-1} & \langle q,q\rangle\lambda^{-1} \\ 0 & -\frac{1}{2}\lambda^{-1} \end{pmatrix}+\sum_{j=1}^{N}\frac{1}{\lambda-\lambda_j}\begin{pmatrix} p_jq_j & -q_j^2 \\ p_j^2 & -p_jq_j \end{pmatrix}, \tag{53}$$

which is almost the same as L^Q. Hence, through calculating the fundamental Poisson bracket and commutator we have the following theorem.

Theorem 6: For the c-GS (47) and c-QS (48), their Lax matrices $L_\Gamma^G(\lambda)$ and $L^Q(\lambda)$ satisfy the fundamental Poisson bracket (18) with the same nondynamical r-matrix:

$$r_{12}(\lambda,\mu)=\frac{2}{\mu-\lambda}P-\frac{2}{\mu}S,\quad S=\begin{pmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&1&0&0\\ 0&0&0&0 \end{pmatrix}=\sigma_-\otimes\sigma^+. \tag{54}$$

We can easily show that Eq. (54) satisfies the CYBE (39).

Remark 3: The r-matrix $r_{12}(\lambda,\mu)$ in Theorem 6 can be also chosen as

$$r_{12}(\lambda,\mu)=\frac{2}{\mu-\lambda}P-\frac{2}{\mu}S+I\otimes\tilde{S}. \tag{55}$$

Equation (54) is the simplest case: $\tilde{S}=0$ of Eq. (55).

V. CONCLUSION

In this article, we present three pairs of different finite-dimensional constrained systems with common nondynamical r-matrices. Along the discrete Toda symplectic map and continuous constrained c-KdV system discovered in Ref. 24 these three pairs form (to the authors' knowledge), the only four examples of pairs of different finite-dimensional integrable systems possessing the above property. The question of whether or not there are any other pairs like them arises. It seems that this is not the case.

From Remarks 1–3, we see that the r-matrix $r_{12}(\lambda,\mu)$ satisfying the fundamental Poisson bracket (18) is composed of two parts, the first one being their main term, and the second one being the common term $I\otimes\tilde{S}$ of Eqs. (21), (40), and (55). Usually when proving the integrability of FDHS we choose their main term as the simplest nondynamical one in order to reduce the calculations.

Apparently, the r-matrix is not uniquely defined. In fact, there are infinitely many r-matrices since the elements a,b,c,d in the matrix $\tilde{S}$ of Eqs. (21), (40), and (55) can be arbitrarily chosen, and they may be constants as well as functions with respect to the spectral parameters λ, μ and the dynamical variables p,q. For a given Lax matrix $L^J(\lambda)$, our r-matrices formulas (21),

(40), and (55) admit an infinite set of solutions $r_{12}(\lambda,\mu)$ of Eq. (18), which includes both the dynamical case [as $a=a(\lambda,\mu,p,q)$, $b=b(\lambda,\mu,p,q)$, $c=c(\lambda,\mu,p,q)$, $d=d(\lambda,\mu,p,q)\in C^{\infty}(R)$] and the nondynamical or constant case [as $a=a(\lambda,\mu)$, $b=b(\lambda,\mu)$, $c=c(\lambda,\mu)$, $d=d(\lambda,\mu)$, or $a,b,c,d=\text{const}$].

In analogy to the first author's thesis,[30] we can further discuss the involutive sets, integrability, separation of variables, and algebraic–geometric solutions for these constrained FDHSs by using the determinant of Lax matrix, r-matrix relation, Poisson bracket, and further modern algebraic–geometric tools.

ACKNOWLEDGMENTS

The first author would like to express his sincere thanks to Prof. Gu Chaohao and Prof. Hu Hesheng for their enthusiastic instructions and help. The last two sections of this work were finished when the first author visited Gh-Kassel University, Germany in 1997. He is very grateful to Prof. Strampp, Prof. Blum, and Prof. Barner for their friendly help and warm invitations. This work has been supported by the National Natural Science Foundation of China, China Postdoctoral Science Foundation, Chinese National Basic Research Project ''Nonlinear Science'' and the Doctoral Programme Foundation of the Institution of High Education, China.

[1] (a) C. W. Cao, Chin. Quart. J. Math. **3**, 90 (1988); (b) Sci. Sin. A (Chinese Series) **32**, 701 (1989); (c) Sci. Sinica A (English Series) **33**, 528 (1990).
[2] M. Antonowicz and S. R. Wojciechowski, Phys. Lett. A **147**, 455 (1990).
[3] C. W. Cao, Acta Math. Sin. New Series **6**, 35 (1990).
[4] M. Antonowicz and S. R. Wojciechowski, J. Phys. A **24**, 5043 (1991).
[5] C. W. Cao and X. G. Geng, J. Math. Phys. **32**, 2323 (1991).
[6] C. W. Cao, Acta Math. Sin. New Series **7**, 216 (1991).
[7] X. G. Geng, Phys. Lett. A **162**, 375 (1992).
[8] Z. J. Qiao, J. Math. Phys. **35**, 3110 (1993).
[9] C. W. Cao, Phys. Lett. A **184**, 333 (1994).
[10] W. X. Ma and W. Strampp, Phys. Lett. A **185**, 277 (1994).
[11] Z. J. Qiao, Phys. Lett. A **192**, 316 (1994).
[12] Z. J. Qiao, J. Math. Phys. **36**, 3535 (1995).
[13] M. A. Semenov-Tian-Shansky, Funct. Anal. Appl. **17**, 259 (1983).
[14] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1986).
[15] A. G. Reyman and M. A. Semenov-Tian-Shansky, Phys. Lett. A **130**, 456 (1988).
[16] A. G. Reyman and M. A. Semenov-Tian-Shansky, *Dynamical Systems 7*, edited by V. I. Arnol'd and S. P. Novikov (Springer, Berlin, 1993), Vol. EMS 16.
[17] E. K. Sklyanin, J. Sov. Math. **47**, 2473 (1989).
[18] O. Babelon and C. M. Viallet, Phys. Lett. B **237**, 411 (1990).
[19] J. C. Eilbeck, V. Z. Enol'skii, V. B. Kuznetsov, and A. V. Tsiganov, J. Phys. A **27**, 567 (1994).
[20] Z. J. Qiao, Physica A **243**, 129 (1997).
[21] F. W. Nijhoff, O. Ragnisco, and V. B. Kuznetsov, Commun. Math. Phys. **176**, 681 (1996).
[22] R. G. Zhou, Phys. Lett. A **220**, 320 (1996).
[23] Z. J. Qiao, Non-dynamical r-matrix and algebraic-geometric solution for the integrable symplectic map, 1996, preprint, to appear in Chin. Sci. Bull. **43**, No. 10 (1998); R. G. Zhou, J. Math. Phys. **38**, 2535 (1997).
[24] Z. J. Qiao and R. G. Zhou, Phys. Lett. A **235**, 35 (1997).
[25] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
[26] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[27] B. M. Levitan and I. S. Sargsjan, *Introduction to Spectral Theory: Self-Adjoint Ordinary Differential Operator* (American Mathematical Society, Providence, RI, 1975).
[28] L. A. Takhtajan, Phys. Lett. A **64**, 235 (1977).
[29] Z. J. Qiao, Phys. Lett. A **186**, 97 (1994).
[30] Z. J. Qiao, ''Generalized Lax algebra, r-matrices and algebraic geometry solutions for the integrable systems,'' preprint, June 1996, Ph.D. thesis (10246/950011), Fudan University, China.

Fluid space–times and conharmonic symmetries

Abdussattar and Babita Dwivedi
Department of Mathematics, Banaras Hindu University, Varanasi 221 005, India

(Received 29 January 1997; accepted for publication 27 January 1998)

The conharmonic curvature tensor is considered as an invariant of the conharmonic transformation defined by Ishii and the necessary and sufficient conditions for the conharmonic curvature tensor in a perfect fluid space–time to be divergence free has been obtained. Conharmonic motion, conharmonic collineation, and conharmonic curvature collineation are introduced as subcases of conformal motion, conformal collineation, and Weyl conformal collineation, respectively, and relations of conharmonic symmetries with inheriting symmetries are investigated. In the case of an existing conharmonic Killing vector along the flow vector, along the anisotropy vector, and perpendicular to both in an anisotropic fluid space–time it is found that no equation of state is singled out unless the conharmonic Killing vector is also a curvature inheritance vector. Conditions are obtained for the symmetries of the anisotropic fluid space–time admitting a conharmonic Killing vector to be inherited. In the case of a conharmonic symmetric space–time and also in the case of a space–time with a divergence-free conharmonic curvature tensor it is found that if the space–time admits an infinitesimal conharmonic Killing vector then the scalar curvature of the space–time vanishes and the space–time is either conharmonically flat or has four distinct principal null directions. © *1998 American Institute of Physics.* [S0022-2488(98)02605-X]

I. INTRODUCTION

Katzin *et al.*[1] initiated the study of a fundamental symmetry property of the space–time of general relativity defined by the vanishing Lie deformation of the Riemann curvature tensor. A vector ξ is said to define a ''curvature collineation'' (CC) if $L_\xi R^h_{ijk}=0$, where R^h_{ijk} is the Riemann curvature tensor of the Riemannian V_4 and L_ξ denotes the Lie derivative with respect to an infinitesimal vector ξ. The contraction of CC, i.e., $L_\xi R_{ij}=0$, where R_{ij} is the Ricci tensor, defines ''Ricci collineation'' (RC). Further studies on CC (including RC) indicated that their relations with conformal and projective symmetries are restricted to a rare special case. In order to get rid of this situation, Duggal[2] modified the concept of CC suitable for proper conformal and other related symmetries and introduced another symmetry called ''curvature inheritance'' (CI). A V_4 is said to admit curvature inheritance along a vector ξ if

$$L_\xi R^h_{ijk}=2\alpha R^h_{ijk}, \tag{1.1}$$

where α is a scalar function of space–time coordinates. The ''Ricci inheritance'' (RI) is similarly defined by

$$L_\xi R_{ij}=2\alpha R_{ij}. \tag{1.2}$$

We also have

$$L_\xi R^j_k=2\alpha R^j_k-R^i_k h^j_i \tag{1.3}$$

and

$$L_\xi R=2\alpha R-R_{ij}h^{ij}, \tag{1.4}$$

where

 © 1998 American Institute of Physics

$$h_{ij} \equiv L_{\xi} g_{ij} = \xi_{i;j} + \xi_{j;i}, \tag{1.5}$$

The existence of symmetry vectors is of interest in the study of exact solutions of Einstein field equations and in the determination of related conservation laws.[3–13] In particular a number of recent articles[14–17] have been concerned with conformal killing vectors (CKVs) in fluid space–times. Much of the analysis that has been done in the subject of CKVs concerns special conformal killing vectors (SCKVs).[18–21]

Ishii[22] introduced the conharmonic transformation as a subgroup of the conformal transformation

$$\overline{g_{ij}} = e^{2\sigma} g_{ij}, \tag{1.6}$$

where σ is a scalar function, satisfying the condition

$$\sigma^{i}_{;i} + \sigma_{;i}\sigma^{i}_{;} = 0. \tag{1.7}$$

A fourth-rank tensor Z^{h}_{ijk} which remains invariant under conharmonic transformation is given by

$$Z^{h}_{ijk} = R^{h}_{ijk} + \tfrac{1}{2}(\delta^{h}_{j} R_{ik} - \delta^{h}_{k} R_{ij} + g_{ik} R^{h}_{j} - g_{ij} R^{h}_{k}) \tag{1.8}$$

and its contraction

$$Z_{ij} = -\tfrac{1}{2} R g_{ij} \tag{1.9}$$

is also invariant under (1.7). It may be noted that the conharmonic curvature tensor Z^{h}_{ijk} and the Riemann curvature tensor R^{h}_{ijk} become identical in an empty space–time ($R_{ij} = 0$). The conformal curvature tensor

$$C^{h}_{ijk} = R^{h}_{ijk} + \tfrac{1}{2}(\delta^{h}_{j} R_{ik} - \delta^{h}_{k} R_{ij} + g_{ik} R^{h}_{j} - g_{ij} R^{h}_{k}) + \frac{R}{6}(\delta^{h}_{k} g_{ij} - \delta^{h}_{j} g_{ik}) \tag{1.10}$$

may be expressed in terms of conharmonic curvature tensor Z^{h}_{ijk} and its contraction Z_{ij} as

$$C^{h}_{ijk} = Z^{h}_{ijk} - \tfrac{1}{3}(\delta^{h}_{k} Z_{ij} - \delta^{h}_{j} Z_{ik}). \tag{1.11}$$

The two tensors C^{h}_{ijk} and Z^{h}_{ijk} become identical if the space–time is with zero scalar curvature. Recently we have obtained a spherically symmetric conharmonically flat space–time

$$ds^2 = \frac{[f(r-t) + F(r+t)]^2}{r^2}[dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\phi^2 - dt^2], \tag{1.12}$$

which is found to describe the gravitational field due to a distribution of pure radiation in the presence of disordered radiation.[23]

The plan of the present paper is as follows. In Sec. II a perfect fluid space–time is considered with a divergence-free conharmonic curvature tensor and it is found that if the conharmonic curvature tensor is divergence-free then either $(\mu + p) = 0$ or the fluid is shear-free, vorticity-free, acceleration-free, and the energy density and isotropic pressure are constant over the spacelike hypersurface orthogonal to the fluid four-velocity vector. Conharmonic motion (Conh M), conharmonic collineation (Conh C), and conharmonic curvature collineation (Conh CC) are defined as subcases of conformal motion (Conf M), conformal collineation (Conf C), and Weyl conformal collineation (W Conf C) respectively, in Sec. III. It is shown that if a space–time admits a proper curvature inheritance vector (CIV) which is also a conharmonic killing vector (Conh KV), then the space–time is necessarily conharmonically flat and if a space–time admits a proper Ricci inheritance vector (RIV) which is also a Conh KV, then the scalar-curvature of the space–time vanishes. In Sec. IV an anisotropic fluid space–time admitting a Conh KV is considered and the equation of state is determined in three special cases of the Conh KV which is also a CIV. In Sec. V, a proper Conh KV is studied in two different space–times—one with a symmetric conharmonic curvature tensor and the other with a divergence-free conharmonic curvature tensor. In both cases

it is found that the scalar curvature of the space–time vanishes and the space–time is either conharmonically flat or of type N, i.e., has four distinct principal null directions.

II. PERFECT FLUID SPACE–TIME WITH A DIVERGENCE-FREE CONHARMONIC CURVATURE TENSOR

The energy momentum tensor for a perfect-fluid distribution is given by

$$T_{ij}=(\mu+p)u_iu_j+pg_{ij}, \tag{2.1}$$

where μ is the energy density, p is the isotropic pressure, and u^i is the fluid four-velocity vector. We consider a perfect fluid space–time which has a divergence-free conharmonic curvature tensor. If the divergence of the conharmonic curvature tensor vanishes, i.e., $Z^h_{ijk;h}=0$, we obtain from Eq. (1.8)

$$(R_{ij}-\tfrac{1}{2}Rg_{ij})_{;k}=(R_{ik}-\tfrac{1}{2}Rg_{ik})_{;j}. \tag{2.2}$$

From the field equations

$$R_{ij}-\tfrac{1}{2}Rg_{ij}=T_{ij}, \tag{2.3}$$

Eq. (2.2) may be expressed as

$$T_{ij;k}=T_{ik;j}. \tag{2.4}$$

Substituting for T_{ij} from Eq. (2.1) in Eq. (2.4) we get

$$\begin{aligned}&[\langle\mu+p\rangle_{;k}u_iu_j+(\mu+p)u_{i;k}u_j+(\mu+p)u_{j;k}u_i+p_{;k}g_{ij}]\\&\quad-[\langle\mu+p\rangle_{;j}u_iu_k+(\mu+p)u_{i;j}u_k+(\mu+p)u_{k;j}u_i+p_{;j}g_{ik}]=0.\end{aligned} \tag{2.5}$$

Contraction by u^k yields

$$(\mu+p)^{\cdot}u_iu_j+(\mu+p)\dot{u}_iu_j+(\mu+p)\dot{u}_ju_i+\dot{p}g_{ij}+(\mu+p)_{;j}u_i+(\mu+p)u_{i;j}-p_{;j}u_i=0. \tag{2.6}$$

Here and in what follows an overhead dot denotes the covariant derivative along the fluid flow vector.

From the conservation equation $T^{ij}_{;j}=0$ we get

$$(\mu+p)\dot{u}_i=-p_{;i}-\dot{p}u_i \quad \text{(force equation)}, \tag{2.7}$$

$$\text{and}\quad \dot{\mu}=-(\mu+p)\theta \quad \text{(energy equation)}, \tag{2.8}$$

where $\theta\equiv u^i_{;i}$ is the expansion scalar. The covariant derivative of the unit flow vector may be decomposed as

$$u_{i;j}=\sigma_{ij}+w_{ij}+\tfrac{1}{3}\theta(g_{ij}+u_iu_j)-\dot{u}_iu_j, \tag{2.9}$$

where σ_{ij} is the shear, w_{ij} is the rotation, and $\dot{u}_i$ is the acceleration of the congruence of the fluid flow vector. Equation (2.6) with the help of Eq. (2.7) takes the form

$$(\mu-p)^{\cdot}u_iu_j+(\mu-p)_{;j}u_i-p_{;i}u_j+\dot{p}g_{ij}+(\mu+p)u_{i;j}=0. \tag{2.10}$$

Contracting by u^i we get

$$(\mu-p)_{;j}=-(\mu-p)^{\cdot}u_j. \tag{2.11}$$

Making use of Eq. (2.7) we get Eq. (2.11)

$$(\mu+p)\dot{u}_i=-\mu_{;i}-\dot{\mu}u_i. \tag{2.12}$$

Equation (2.10) in view of Eq. (2.11) reduces to

$$-p_{;i}u_j+\dot{p}g_{ij}+(\mu+p)u_{i;j}=0. \tag{2.13}$$

Contracting by g^{ij} we get

$$\dot{p}=-\tfrac{1}{3}(\mu+p)\,\theta. \tag{2.14}$$

In view of Eq. (2.8) this equation yields

$$\dot{\mu}=3\dot{p}. \tag{2.15}$$

Equation (2.13) with the help of Eq. (2.7) may be written as

$$(\mu+p)\dot{u}_iu_j+\dot{p}(g_{ij}+u_iu_j)+(\mu+p)u_{i;j}=0. \tag{2.16}$$

Using Eq. (2.14) in Eq. (2.16) we get

$$(\mu+p)[\dot{u}_iu_j-\tfrac{1}{3}\theta(g_{ij}+u_iu_j)+u_{i;j}]=0. \tag{2.17}$$

This suggests that either

$$(\mu+p)=0 \tag{2.18}$$

or

$$u_{i;j}=\tfrac{1}{3}\theta(g_{ij}+u_iu_j)-\dot{u}_iu_j\,. \tag{2.19}$$

If $(\mu+p)=0$ we have either $\mu=0$, $p=0$ (empty space–time) or the perfect fluid may be regarded as if endowed with a vacuumlike equation of state.[24]

Comparing Eq. (2.19) with Eq. (2.9) we get

$$\sigma_{ij}+\omega_{ij}=0. \tag{2.20}$$

That is, $\sigma_{ij}=\omega_{ij}=0$. Making use of Eqs. (2.12) and (2.19), Eq. (2.5) becomes

$$\begin{aligned}&[p_{;k}-(\mu+p)\dot{u}_k]u_iu_j+\tfrac{1}{3}(\mu+p)\,\theta u_jg_{ik}-(\mu+p)\dot{u}_ju_ku_i+p_{;k}g_{ij}\\&\quad-[p_{;i}-(\mu+p)\dot{u}_j]u_iu_k-\tfrac{1}{3}(\mu+p)\,\theta g_{ij}u_k+(\mu+p)\dot{u}_ku_iu_j-p_{;j}g_{ik}=0.\end{aligned} \tag{2.21}$$

This equation with the help of Eqs. (2.7) and (2.14) may be written as

$$p_{;k}u_iu_j-p_{;j}u_iu_k-(\mu+p)\dot{u}_kg_{ij}+(\mu+p)\dot{u}_jg_{ik}-(\mu+p)\dot{u}_ku_iu_j+(\mu+p)\dot{u}_ju_iu_k=0. \tag{2.22}$$

Contracting by g^{ij} and making use of Eq. (2.7), Eq. (2.22) yields

$$(\mu+p)\dot{u}_k=0, \tag{2.23}$$

i.e.,

$$\dot{u}_k=0 \quad (\text{since } \mu+p\neq 0). \tag{2.24}$$

In view of this equation, Eqs. (2.7) and (2.12) yield

$$\mu_{;i}=-\dot{\mu}u_i \tag{2.25}$$

and

$$p_{;i}=-\dot{p}u_i\,. \tag{2.26}$$

From these equations and Eq. (2.15) we obtain

$$\mu-3p=\text{constant}. \tag{2.27}$$

Thus if $(\mu+p)\neq 0$, then the fluid is shear-free, vorticity-free, acceleration-free, and the energy density and pressure are constant over the spacelike hypersurface orthogonal to the fluid four-velocity vector. These are precisely the conditions for the space–time to be FRW.

Conversely, if the perfect fluid does not have a vacuumlike equation of state and $(\mu+p)=0$ (i.e., $\mu=0$, $p=0$), then obviously the conharmonic curvature tensor of the space–time is divergence-free. Also if the fluid is shear-free, vorticity-free, acceleration-free, and the energy density and pressure are constant over the spacelike hypersurface orthogonal to the fluid four-velocity vector, from Eq. (2.1) we see that

$$T_{ij;k}-T_{ik;j}=[(\mu+p)u_iu_j+pg_{ij}]_{;k}-[(\mu+p)u_iu_k+pg_{ik}]_{;j}. \tag{2.28}$$

Using Eqs. (2.19), (2.24), (2.25), and (2.26) in Eq. (2.28) we get Eq. (2.4), i.e., the divergence of the conharmonic curvature tensor vanishes. Thus we may state the following theorem.

Theorem (2.1): The conharmonic curvature tensor of a perfect-fluid space–time will be divergence-free if and only if either $(\mu+p)=0$ (which is not really a perfect fluid but is an Einstein space) or the space–time is a FRW model satisfying $(\mu-3p)=$constant.

III. CONHARMONIC SYMMETRIES AND THE RELATION OF CONH KV WITH INHERITING SYMMETRIES

A V_4 is said to admit a Conf M generated by a conformal Killing vector (CKV) ξ if

$$h_{ij}\equiv L_\xi g_{ij}=2\psi g_{ij}, \tag{3.1}$$

where ψ is a scalar function of the space–time coordinates given by $\psi=\frac{1}{4}\xi^k_{;k}$. In particular ξ is called a special conformal Killing vector (SCKV) if $\psi_{;ij}=0$, $\psi_{,i}\neq 0$. Other subcases of CKV are the homothetic vector (HV) if $\psi_{,i}=0$, $\psi\neq 0$ and the Killing vector (KV) if $\psi=0$.

A V_4 is said to admit a conformal collineation (Conf C) if there exists a vector ξ such that

$$L_\xi\Gamma^i_{jk}=\delta^i_j\psi_{;k}+\delta^i_k\psi_{;j}-g_{jk}g^{im}\psi_{;m} \tag{3.2}$$

and V_4 admits a Weyl conformal collineation (W Conf C) along the vector field ξ if

$$L_\xi C^h_{ijk}=0. \tag{3.3}$$

It follows that every Conf M is a Conf C and W Conf C but the converse is not necessarily true.

Now we define another subcase of conformal symmetries called conharmonic symmetries. A Conf M is defined as conharmonic motion (Conh M) if

$$\Box\psi\equiv g^{ij}\psi_{;ij}=0,\quad \psi_{;ij}\neq 0. \tag{3.4}$$

A Conf C is similarly defined as Conh C if Eq. (3.2) holds together with condition (3.4). A V_4 is said to admit a conharmonic curvature collineation (Conh CC) if there exists a vector ξ which satisfies

$$L_\xi Z^h_{ijk}=0. \tag{3.5}$$

Obviously every conharmonic motion is necessarily a Conh C and Conh CC but the converse is not necessarily true. From Eq. (1.11) it is clear that every Conh CC is a W Conf C. Every Conh M must also satisfy

$$L_\xi R^h_{ijk}=\delta^h_j\psi_{;ik}-\delta^h_k\psi_{;ij}+\psi^h_{;j}g_{ik}-\psi^h_{;k}g_{ij}, \tag{3.6}$$

$$L_\xi R_{ij}=-2\psi_{;ij}, \tag{3.7}$$

$$L_\xi R^j_k=-2\psi^j_{;k}-2\psi R^j_k, \tag{3.8}$$

and

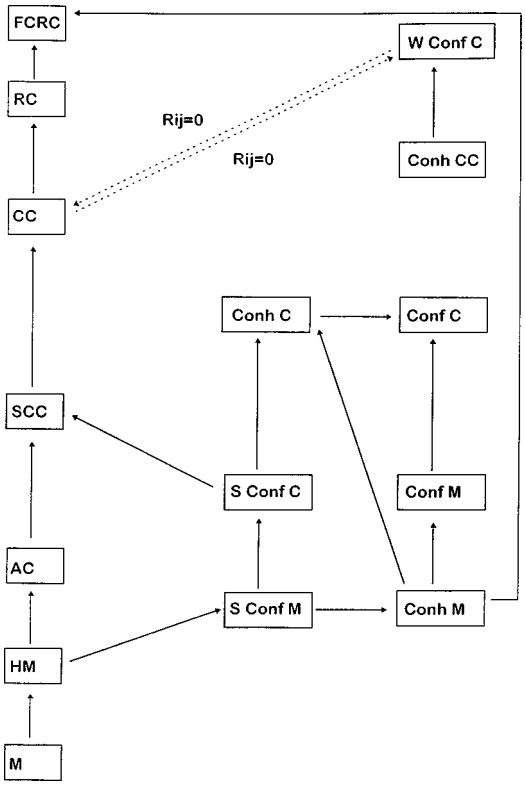

FIG. 1. Relation between conharmonic and other space–time symmetries.

$$L_\xi R = -2\psi R. \tag{3.9}$$

From Eq. (3.7)—in view of Eq. (3.4)—we observe that

$$g^{ij} L_\xi R_{ij} = 0, \tag{3.10}$$

showing that every Conh M is a contracted Ricci collineation but the converse is not necessarily true. Consequently, the existence of a Conh KV ξ also gives rise to a conservation law generator of the form[25]

$$[\sqrt{-g}(T_i^j - \tfrac{1}{2}\delta_i^j T)\xi^i]_{;j} = 0. \tag{3.11}$$

A block diagram (Fig. 1) is constructed showing the relation of conharmonic symmetries with some other well-known symmetries. Now we have the following theorems based upon the relation of Conh KV with inheriting symmetries.

Theorem 3.1: A V_4, which admits a Conh KV and also a proper CIV, is necessarily a conharmonically flat space.

Proof: Taking the Lie derivative of (1.8) and making use of Eqs. (1.1), (1.2), (1.3), and (1.8) we get

$$L_\xi Z_{ijk}^h = 2\alpha Z_{ijk}^h + E_{ijk}^h, \tag{3.12}$$

where

$$E_{ijk}^h = \tfrac{1}{2}[h_{ik}R_j^h - h_{ij}R_k^h + g_{ij}R_k^m h_m^h - g_{ik}R_j^m h_m^h]. \tag{3.13}$$

Using Eq. (3.1) in Eq. (3.13) we get

$$E_{ijk}^h = 0. \tag{3.14}$$

In view of Eqs. (3.5) and (3.14), Eq. (3.12) gives

$$Z^h_{ijk}=0 \quad [\text{since } \alpha\neq 0), \tag{3.15}$$

showing that V_4 is conharmonically flat.

In order to illustrate this theorem we consider a plane symmetric perfect fluid cosmological model obtained by Singh and Singh[26] which is of class one but not conformal to flat space–time. The geometry of the model is described by the metric

$$ds^2=(1+at)^2(-dt^2+dx^2+dy^2)+(1+bt)dz^2, \tag{3.16}$$

where a and b are either both positive or both negative arbitrary constants. The space–time (3.16) is found to admit a Conh KV $\xi^i=(A/a)\delta^i_0$, which is also a CIV only when $a=b$ with $\psi=-\alpha=A/(1+at)$, where A is an arbitrary constant. However, when $a=b$ the model becomes conharmonic to flat space–time and reduces to a special case $(k=0)$ of the FRW model representing a universe filled with disordered radiation.

Theorem 3.2: If ξ is a RIV as well as a Conh KV in a V_4 then

$$L_\xi R^h_{ijk}=2\alpha(R^h_{ijk}-Z^h_{ijk}). \tag{3.17}$$

Proof: Taking the Lie derivative of Eq. (1.8) and then using Eqs. (1.2), (1.3), (3.1), and (3.5) we get

$$L_\xi R^h_{ijk}+\alpha(\delta^h_j R_{ik}-\delta^h_k R_{ij}+g_{ik}R^h_j-g_{ij}R^h_k)=0. \tag{3.18}$$

Using Eq. (1.8) in Eq. (3.18) we get the following result.

Corollary: If ξ is a RIV as well as a Conh KV in a conharmonically flat space–time then the symmetry of the Riemann tensor of the space–time is inherited.

Proof: Proof is obvious from Eq. (3.17).

Remark: In the conharmonically flat cosmological model of Singh and Singh given by Eq. (3.16) with $a=b$, $\xi^i=(A/a)\delta^i_0$ being a Conh KV and RIV with $\psi=-\alpha=A/(1+at)$ is found to be a CIV also.

Theorem 3.3: If ξ is a proper RIV as well as a Conh KV in a V_4 then the scalar curvature of the space–time vanishes and if the Riemann curvature tensor of the space–time is divergence-free then we have

$$(\alpha+\psi)_{;j}R^j_k=0. \tag{3.19}$$

Proof: Equations (1.2) and (3.7) give

$$\alpha R_{ij}=-\psi_{;ij}. \tag{3.20}$$

Contracting by g^{ij} and using (3.4) we get

$$\alpha R=0, \tag{3.21}$$

that is, the scalar curvature R of the space–time vanishes (since $\alpha\neq 0$). Differentiating (3.20) covariantly we get

$$\alpha_{;k}R_{ij}+\alpha R_{ij;k}=-\psi_{;ijk}. \tag{3.22}$$

Antisymmetrizing this equation with respect to j and k and making use of the divergence-free condition of the Riemann tensor,

$$R^h_{ijk;h}\equiv R_{ij;k}-R_{ik;j}=0, \tag{3.23}$$

and the identity

$$\psi_{;ikj}-\psi_{;ijk}=\psi_{;h}R^h_{ijk}, \tag{3.24}$$

we get

$$[\alpha_{;k}R_{ij}-\alpha_{;j}R_{ik}]=\psi_{;h}R^{h}_{ijk}. \tag{3.25}$$

Contracting this equation by g^{ij} and making use of the fact that $R=0$ we get the theorem.

Corollary: If a space–time V_4, with perfect fluid distribution admits a RIV which is also a Conh KV, then this space–time necessarily admits a distribution of disordered radiation or photon fluid. Also if the Riemann curvature tensor of the space–time is divergence free then we have $(\alpha+\psi)=$constant.

Proof: From Eqs. (2.1), (2.3), and the fact that $R=0$ we get the first part of the corollary with

$$R_{ij}=\frac{\mu}{3}(4u_iu_j+g_{ij}). \tag{3.26}$$

In view of Eq. (3.26), Eq. (3.19) yields

$$\frac{\mu}{3}[4(\alpha+\psi)^{\cdot}u_k+(\alpha+\psi)_{;k}]=0. \tag{3.27}$$

Contracting Eq. (3.27) with u_k we get

$$(\alpha+\psi)^{\cdot}=0 \quad (\text{if } \mu\neq 0), \tag{3.28}$$

that is

$$(\alpha+\psi)=\text{constant}. \tag{3.29}$$

Remark: In the plane symmetric model of Singh and Singh given by Eq. (3.16), $\xi^i=A/a\,\delta^i_0$ is a Conh KV and a RIV leads to $R=0$. Since the model is filled with a perfect fluid distribution, in this situation $(R=0)$ it represents a distribution of disordered radiation.

We now consider a static spherically symmetric metric in curvature coordinates

$$ds^2=-e^{\nu}dt^2+e^{\lambda}dr^2+r^2(d\theta^2+\sin^2\theta\,d\phi^2), \tag{3.30}$$

where λ and ν are functions of r alone. If this space–time is admitting a one parameter group of Conf M defined by Eq. (3.1) with ψ as a function of r alone, then we have[18,27,28]

$$e^{\nu}=a^2r^2, \tag{3.31}$$

$$e^{\lambda}=b^2/\psi^2, \tag{3.32}$$

and

$$\xi^i=\psi r\,\delta^i_1, \tag{3.33}$$

where a and b are arbitrary constants.

In order that the metric (3.30) may admit a one parameter group of Conh M, the function ψ should satisfy Eq. (3.4), which gives

$$\psi_{,rr}+\psi^2_{,r}/\psi+3\psi_{,r}/r=0. \tag{3.34}$$

Integrating Eq. (3.34) we get

$$\psi^2=A/r^2+B, \tag{3.35}$$

where A and B are arbitrary constants. The line-element (3.30) admitting a one parameter group of Conh M is obtained as

$$ds^2=-a^2r^2dt^2+b^2(A/r^2+B)^{-1}dr^2+r^2(d\theta^2+\sin^2\theta\,d\phi^2). \tag{3.36}$$

This space–time is obviously found to admit contracted Ricci collineation defined by Eq. (3.10) along the Conh KV $\xi^i=(A/r^2+B)^{1/2}r\,\delta^i_1$. The scalar curvature R of the space–time (3.36) vanishes if we take $B=b^2/3$. The space–time with scalar curvature $R=0$ is admitting a Conh KV $\xi^i=(A/r^2+b^2/3)^{1/2}r\,\delta^i_1$ but it is verified that this ξ^i is not a RIV, showing that the converse of the first part of theorem (3.3) is not necessarily true.

IV. ANISOTROPIC FLUID SPACE–TIMES ADMITTING CONH KV

Harrera *et al.*[18] initiated the study of Conf M for anisotropic matter in the context of general relativity. Maartens *et al.*[20] examined in detail the kinematic and dynamic properties of anisotropic fluids that admit a Conf M. The relations of energy density with the anisotropic pressure components have been explicitly obtained by considering the conformal killing vector to be special and parallel to u^i, parallel to n^i, and perpendicular to u^i and n^i. They have further extended their results to a fluid with arbitrary stress tensor and with vanishing energy flux vector. In this section we consider the anisotropic fluid space–time to admit a Conh KV which is also a ClV and study its consequences.

If ξ is a CKV and X^i is any non-null unit vector, then Maartens *et al.*[20] have shown that

$$L_\xi X^i=-\psi X^i+Y^i \tag{4.1}$$

and

$$L_\xi X_i=\psi X_i+Y_i, \tag{4.2}$$

where Y^i is a vector perpendicular to X^i. Making use of these results for a timelike unit vector u^i and spacelike unit vector n^i we get

$$L_\xi u^i=-\psi u^i+v^i, \tag{4.3}$$

$$L_\xi u_i=\psi u_i+v_i, \tag{4.4}$$

where $u_iv^i=0$ and

$$L_\xi n^i=-\psi n^i+m^i, \tag{4.5}$$

$$L_\xi n_i=\psi n_i+m_i, \tag{4.6}$$

where $n_im^i=0$. These vectors u^i and n^i are identified as the unit fluid four-velocity vector and the unit anisotropy vector in the energy momentum tensor for a fluid with anisotropic pressure. It is clear from Eqs. (4.3) and (4.6) that if $n_iu^i=0$ then we have

$$n_iv^i+u^im_i=0. \tag{4.7}$$

Maartens *et al.*[20] have shown that (i) if a CKV ξ is parallel to u^i, i.e., $\xi^i=\lambda u^i$, then

$$\psi=\dot\lambda+\dot u_i\xi^i \tag{4.8}$$

and

$$v_i=2\omega_{ij}\xi^j+\lambda[\dot u_i+(\log\lambda^{-1})_{;j}h^j_i], \tag{4.9}$$

where $\lambda=-\xi_iu^i$ is a scalar function, ω_{ij} is the vorticity tensor, $\dot u_i$ is the four-acceleration vector, and $h_{ij}=g_{ij}+u_iu_j$ is the projection tensor. (ii) If ξ is perpendicular to u^i, i.e., $\xi_iu^i=0$, then

$$\psi=\dot u_i\xi^i, \tag{4.10}$$

$$v_i=2\omega_{ij}\xi^j. \tag{4.11}$$

The expression for the energy–momentum tensor of an anisotropic fluid is given by

$$T_{ij}=\mu u_i u_j+P_{11}n_i n_j+p_1 p_{ij}, \tag{4.12}$$

where μ is the total energy density as measured by an observer with four-velocity u^i, n^i is a unit spacelike vector of anisotropy orthogonal to u^i, p_{11} and p_1 are the components of the anisotropic pressure parallel to and perpendicular to n^i, and $p_{ij}=g_{ij}+u_i u_j-n_i n_j$ is the projection tensor onto the two plane orthogonal to u^i and n^i. We substitute for T_{ij} from Eq. (4.12) in Eq. (2.3) and then take the Lie derivative. This gives, on making use of Eqs. (3.1), (3.7), (3.9), (4.4), and (4.6),

$$\begin{aligned}-2\psi_{;ij}=&(L_\xi\mu+2\psi\mu)u_i u_j+(L_\xi p_{11}+2\psi p_{11})n_i n_j+(L_\xi p_1+2\psi p_1)p_{ij}\\&+2(\mu+p_1)u_{(i}v_{j)}+2(p_{11}-p_1)n_{(i}m_{j)}.\end{aligned} \tag{4.13}$$

Contractions by $u^i u^j$, $n^i n^j$, p^{ij}, $u^i n^j$, $u^i p^{jk}$, $n^i p^{jk}$, and $p^{ik}p^{jl}-\frac{1}{2}p^{ij}p^{kl}$, respectively, yield

$$L_\xi\mu+2\psi\mu=-2\psi_{;ij}u^i u^j, \tag{4.14}$$

$$L_\xi p_{11}+2\psi p_{11}=-2\psi_{;ij}n^i n^j, \tag{4.15}$$

$$L_\xi p_1+2\psi p_1=-\psi_{;ij}p^{ij}, \tag{4.16}$$

$$(\mu+p_{11})n^j v_j=2\psi_{;ij}u^i n^j, \tag{4.17}$$

$$(\mu+p_1)p^{jk}v_j=2\psi_{;ij}u^i p^{jk}, \tag{4.18}$$

$$(p_1-p_{11})p^{jk}m_j=2\psi_{;ij}n^i p^{jk}, \tag{4.19}$$

$$\psi_{;ij}p^{ik}p^{jl}-\tfrac{1}{2}(\psi_{;ij}p^{ij})p^{kl}=0. \tag{4.20}$$

With the help of Eqs. (4.14)–(4.20) we see that if ξ is a Conh KV of an anisotropic fluid space–time which is parallel to u^i, parallel to n^i, and perpendicular to u^i and n^i, then no equation of state is singled out. However, when we take ξ to be a CIV also then it may be possible to get an equation of state. If ξ is a CKV of a space–time of general relativity with anisotropic fluid given by (4.12), which is also a CIV, and if $(\mu+p_{11})\neq 0$ and $(\mu+p_1)\neq 0$ then Duggal[2] has obtained the following inheritance equations:

$$L_\xi\mu=2(\alpha-\psi)\mu, \tag{4.21a}$$

$$L_\xi p_{11}=2(\alpha-\psi)p_{11}, \tag{4.21b}$$

$$L_\xi p_1=2(\alpha-\psi)p_1, \tag{4.21c}$$

$$L_\xi u_i=\psi u_i(\text{i.e., } v_i=0). \tag{4.21d}$$

A CKV ξ satisfies[20]

$$(R^{ij}\xi_j)_{;i}=-3\Box\psi. \tag{4.22}$$

For a Conh KV ξ this reduces to

$$(R^{ij}\xi_j)_{;i}=0. \tag{4.23}$$

If ξ is a Conh KV which is also a CIV of the space–time then from theorem (3.3) we have in the case of anisotropic fluid

$$\mu=2p_1+p_{11}. \tag{4.24}$$

Since the scalar curvature of the space–time is zero; we may write Eq. (4.23) with the help of Eqs. (2.3) and (4.12) as

$$[(\mu u^i u^j+p_{11}n^i n^j+p_1 p^{ij})\xi_j]_{;i}=0. \tag{4.25}$$

From (3.1) we get

$$\xi^i_{;i}=4\psi. \tag{4.26}$$

Now we consider three special cases of the vector ξ: (i) ξ parallel to n^i, (ii) ξ parallel to u^i, and (iii) ξ perpendicular to u^i and n^i.

Case (i): Using $\xi^i=\lambda n^i$ in (4.25) we get

$$[p_{11}\xi^i]_{;i}=0. \tag{4.27}$$

With the help of Eqs. (4.21b) and (4.26), Eq. (4.27) obtains

$$(\alpha+\psi)p_{11}=0. \tag{4.28}$$

If $(\alpha+\psi)\neq 0$, then $p_{11}=0$. Thus Eq. (4.24) gives

$$\mu=2p_1. \tag{4.29}$$

Case (ii): Using $\xi^i=\lambda u^i$ in (4.25) we get

$$[\mu\xi^i]_{;i}=0. \tag{4.30}$$

With the help of Eqs. (4.21a) and (4.26), Eq. (4.30) obtains

$$(\alpha+\psi)\mu=0. \tag{4.31}$$

Since $(\alpha+\psi)\neq 0$ we get $\mu=0$ and Eq. (4.24) reduces to

$$2p_1+p_{11}=0. \tag{4.32}$$

Case (iii): Using $\xi^i u_i=\xi^i n_i=0$ in (4.25) we get

$$(p_1\xi^i)_{;i}=0. \tag{4.33}$$

With the help of Eqs. (4.21c) and (4.26), Eq. (4.33) gives

$$(\alpha+\psi)p_1=0. \tag{4.34}$$

Since $(\alpha+\psi)\neq 0$ we get $p_1=0$ and Eq. (4.24) gives

$$\mu=p_{11}. \tag{4.35}$$

Remark: In the case of a perfect fluid distribution (i.e., when $p_1=p_{11}$) cases (i) and (iii) give the stiff equation of state and case (ii) leads to an empty space–time.

Now we consider the case of an anisotropic fluid with arbitrary stress tensor π^{ij} ($\pi^{ij}u_j=0$, $\pi^i_i=0$) given by

$$T^{ij}=\mu u^i u^j+ph^{ij}+\pi^{ij}, \tag{4.36}$$

where p is the isotropic pressure. Taking the Lie derivative of Eq. (2.3) with T_{ij} given by (4.36) and making use of Eqs. (3.1), (3.7), (3.9), and (4.4) we get

$$-2\psi_{;ij}=(L_\xi\mu+2\psi\mu)u_iu_j+(L_\xi p+2\psi p)h_{ij}+L_\xi\pi_{ij}. \tag{4.37}$$

Contracting Eq. (4.37) by u^iu^j, h^{ij}, $u^ih^j_k$, and $h^i_kh^j_l-\frac{1}{3}h^{ij}h_{kl}$, respectively, we get

$$L_\xi\mu+2\psi\mu=-2\psi_{;ij}u^iu^j, \tag{4.38}$$

$$L_\xi p+2\psi p=-\tfrac{2}{3}\psi_{;ij}u^iu^j, \tag{4.39}$$

$$u^ih^j_kL_\xi\pi_{ij}=-2\psi_{;ij}u^ih^j_k, \tag{4.40}$$

$$\pi_{ik}v^i = -2\psi_{;ij}(h^i_k h^j_l - \tfrac{1}{3}h^{ij}h_{kl}). \tag{4.41}$$

Using Eqs. (2.3) and (4.36) in Eq. (4.23) we get

$$\left[(\mu+p)u^i(u^j\xi_j)+\left(\frac{\mu}{2}-\frac{p}{2}\right)\xi^i+\pi^{ij}\xi_j\right]_{;i}=0. \tag{4.42}$$

Considering $\xi_i\, u^i=0$ and making use of Eq. (4.26), Eq. (4.42) yields

$$\left(\frac{\mu}{2}-\frac{p}{2}\right)_{;i}\xi^i+2\psi(\mu-p)+\pi^{ij}_{;i}\xi_j=0. \tag{4.43}$$

With the help of Eqs. (4.38) and (4.39), Eq. (4.43) takes the form

$$\psi(\mu-p)+\pi^{ij}_{;i}\xi_j=\tfrac{2}{3}\psi_{;ij}u^iu^j. \tag{4.44}$$

For the T^{ij} given by (4.36) the momentum conservation equation is obtained as

$$(\mu+p)\dot{u}^i=-h^{ik}(p_{;k}+\pi^j_{k;j}). \tag{4.45}$$

Multiplying Eq. (4.45) by ξ_i and making use of Eqs. (4.39) and (4.44) we get

$$(\mu+p)[\dot{u}^i\xi_i-\psi]=0. \tag{4.46}$$

In view of Eq. (4.10), Eq. (4.46) is identically satisfied for any μ and p.

When we choose ξ parallel to u^i then Eq. (4.42) is the same as in the case of perfect fluid distribution because $u_i\pi^{ij}=0$. Thus in this case the equation of state may be singled out only when ξ is also a CIV and this leads to an empty space–time.

The space–time symmetries of a CKV ξ are said to be inherited[29] if fluid flow lines are mapped conformally by ξ, i.e., if Eq. (4.21d) holds. We find that the space–time symmetries are not inherited in the case of Conh KV ξ in general but when we take u_i as an eigenvector of $\psi_{;ij}$, i.e.,

$$\psi_{;ij}u^j=\nu u_i, \tag{4.47}$$

where ν is a scalar function then Eqs. (4.17) and (4.18) are reduced to

$$(\mu+p_{11})n^jv_j=0 \tag{4.48}$$

and

$$(\mu+p_1)p^{jk}v_j=0. \tag{4.49}$$

Now if $(\mu+p_{11})\neq 0$ and $(\mu+p_1)\neq 0$, then Eqs. (4.48) and (4.49) give

$$n^jv_j=0 \tag{4.50}$$

and

$$p^{jk}v_j=0. \tag{4.51}$$

Using Eq. (4.50) in Eq. (4.51) and remembering the fact that $u^iv_i=0$, we get

$$v_i=0. \tag{4.52}$$

Thus the Conh KV symmetries of the fluid space–time are inherited when u_i is an eigenvector of $\psi_{;ij}$.

Recently, Coley and Tupper[30] have investigated all the spherically symmetric space–times representing an anisotropic fluid and admitting a proper inheriting conformal Killing vector (ICKV). We consider the case of class B2S space–times,

$$ds^2=S^{-2}(r)(-dt^2+dr^2+d\theta^2+\sin^2\ \theta d\phi^2) \tag{4.53}$$

with $\xi^i=\delta^i_1$ and $\psi=-S^{-1}S,r$. In order that the space–time (4.53) may admit a proper inheriting conharmonic Killing vector (I Conh KV), ψ should satisfy Eq. (3.4), which gives

$$\psi_{,rr}+2\psi\psi_{,r}=0. \tag{4.54}$$

Integrating Eq. (4.54) we get

$$\psi_{,r}+\psi^2=K, \tag{4.55}$$

where K is an arbitrary constant. The solution of Eq. (4.55) is obtained as

$$\psi=\frac{\sqrt{K}(1-Ce^{-2\sqrt{K}r})}{(1+Ce^{-2\sqrt{K}r})}, \tag{4.56}$$

where C is an arbitrary constant. Consequently a spherically symmetric static metric representing an anisotropic fluid and admitting a proper inheriting Conh KV $\xi^i=\delta^i_1$ is obtained as

$$ds^2=(Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r})^2[-dt^2+dr^2+d\theta^2+\sin^2\ \theta\ d\phi^2], \tag{4.57}$$

where A and B are arbitrary constants satisfying $B=AC$.

The expressions for μ, p_{11}, and p_1 in this model are obtained as

$$\mu=(Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r})^{-2}\left[1-2K+K\left\{\frac{Ae^{\sqrt{K}r}-Be^{-\sqrt{K}r}}{Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r}}\right\}^2\right], \tag{4.58}$$

$$p_{11}=(Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r})^{-2}\left[-1+3K\left\{\frac{Ae^{\sqrt{K}r}-Be^{-\sqrt{K}r}}{Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r}}\right\}^2\right], \tag{4.59}$$

and

$$p_1=(Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r})^{-2}\left[2K-K\left\{\frac{Ae^{\sqrt{K}r}-Be^{-\sqrt{K}r}}{Ae^{\sqrt{K}r}+Be^{-\sqrt{K}r}}\right\}^2\right]. \tag{4.60}$$

The equation of state satisfied by the anisotropic fluid in this model is obtained as

$$\mu=\frac{p_{11}+4p_1-6Kp_1}{(6K-1)}. \tag{4.61}$$

It may be noted that when we take $K=1/3$ then

$$\mu=p_{11}+2p_1. \tag{4.62}$$

V. SOME SPACE–TIMES WITH PROPER CONHARMONIC SYMMETRY

A. Conharmonic symmetric space–times

Sharma[31] has studied the proper conformal symmetries of conformal symmetric space–times defined by[32]

$$C_{hijk;l}=0. \tag{5.1}$$

It is shown that conformal symmetric space–times, admitting an infinitesimal conformal symmetry, are either of type O or N. Those of type N represent plane gravitational waves with parallel rays, provided the Einstein tensor is invariant of the infinitesimal conformal symmetry. In this section we consider a conharmonic symmetric space–time defined by

$$Z_{hijk;l}=0 \tag{5.2}$$

and study the consequences of an existing Conh KV. The commutation formula[33] for the conharmonic curvature tensor gives

$$L_\xi(Z^h_{ijk;l})-(L_\xi Z^h_{ijk})_{;l}=(L_\xi\Gamma^h_{lm})Z^m_{ijk}-(L_\xi\Gamma^m_{li})Z^h_{mjk}-(L_\xi\Gamma^m_{lj})Z^h_{imk}-(L_\xi\Gamma^m_{lk})Z^h_{ijm}. \tag{5.3}$$

Using Eqs. (3.2), (3.5), and (5.2) in Eq. (5.3) we get

$$\begin{aligned}&(\delta^h_l\psi_{;m}+\delta^h_m\psi_{;l}-g_{lm}\psi^h_{;})Z^m_{ijk}-(\delta^m_l\psi_{;i}+\delta^m_i\psi_{;l}-g_{li}\psi^m_{;})Z^h_{mjk}\\&\quad-(\delta^m_l\psi_{;j}+\delta^m_j\psi_{;l}-g_{lj}\psi^m_{;})Z^h_{imk}-(\delta^m_l\psi_{;k}+\delta^m_k\psi_{;l}-g_{lk}\psi^m_{;})Z^h_{ijm}=0.\end{aligned} \tag{5.4}$$

Contracting the above equation at h and l we get

$$\psi_{;m}Z^m_{ijk}-\psi_{;m}(Z^m_{ijk}+Z^m_{kij}+Z^m_{jki})+\psi_{;j}Z_{ik}-\psi_{;k}Z_{ij}=0. \tag{5.5}$$

From Eq. (1.8) it may be seen that

$$Z^m_{\{ijk\}}=0. \tag{5.6}$$

Thus Eq. (5.5) reduces to

$$\psi_{;m}Z^m_{ijk}+\psi_{;j}Z_{ik}-\psi_{;k}Z_{ij}=0. \tag{5.7}$$

Contracting by g^{ij} and then making use of Eq. (1.9) we get

$$R\psi_{;k}=0. \tag{5.8}$$

Since ξ is a proper conharmonic vector field we have $\psi_{;i}\neq 0$ and Eq. (5.8) leads to $R=0$. Consequently, Eq. (1.9) gives

$$Z_{ij}=0 \tag{5.9}$$

and from Eq. (1.11) the conharmonic curvature tensor and the conformal curvature tensor become identical. Equation (5.7) now takes the form

$$\psi_{;m}Z^m_{ijk}=0. \tag{5.10}$$

Making use of Eq. (5.10) in Eq. (5.4) and then multiplying the obtained equation by $\psi^l_{;}$ we get

$$\psi_{;l}\psi^l_{;}Z^h_{ijk}=0. \tag{5.11}$$

In view of the above equation, either $Z^h_{ijk}=0$, i.e., the space–time is conharmonically flat or $\psi^l_{;}\,\psi_{;l}=0$. Since space–time admits proper conharmonic vector field ξ, $\psi^l_{;}\neq 0$, therefore, if $\psi^l_{;}\psi_{;l}=0$ then $\psi^l_{;}$ must be null. Thus from Eq. (5.10) the space–time is of type N and the four repeated principal null directions of the conharmonic curvature tensor are given by $\psi^l_{;}$. Thus we may state the following theorem.

Theorem 5.1: If a conharmonic symmetric space–time admits a proper conharmonic killing vector field ξ, then the scalar curvature of the space–time vanishes and the space–time is either conharmonically flat or of type N.

B. Space–times admitting a divergence-free conharmonic curvature tensor

Sharma[34] has shown that if ξ is a CKV of a space–time admitting a divergence-free conformal curvature tensor then this space–time will be either conformally flat or of type N. We have studied the above case when ξ is a Conh KV of a space–time with a divergence-free conharmonic curvature tensor. From Eq. (1.8) we obtain

$$Z^h_{ijk;l}+Z^h_{ikl;j}+Z^h_{ilj;k}=\tfrac{1}{2}[\,\delta^h_j(R_{ik;l}-R_{il;k})+\delta^h_k(R_{il;j}-R_{ij;l})+\delta^h_l(R_{ij;k}-R_{ik;j}) +g_{ik}(R^h_{j;l}-R^h_{l;j})+g_{il}(R^h_{k;j}-R^h_{j;k})+g_{ij}(R^h_{l;k}-R^h_{k;l})]. \tag{5.12}$$

The divergence-free condition of the conharmonic curvature tensor given by Eq. (2.2) reduces Eq. (5.12) to

$$Z^h_{ijk;l}+Z^h_{ikl;j}+Z^h_{ilj;k}=\tfrac{1}{2}[\,\delta^h_j(R_{;l}g_{ik}-R_{;k}g_{il})+\delta^h_k(R_{;j}g_{il}-R_{;l}g_{ij})+\delta^h_l(R_{;k}g_{ij}-R_{;j}g_{ik})]. \tag{5.13}$$

Now taking the Lie derivative of Eq. (5.13) we get

$$\begin{aligned}L_\xi[Z^h_{ijk;l}+Z^h_{ikl;j}+Z^h_{ilj;k}]&=\psi[\,\delta^h_j(R_{;l}g_{ik}-R_{;k}g_{il})+\delta^h_k(R_{;j}g_{il}-R_{;l}g_{ij})+\delta^h_l(R_{;k}g_{ij}-R_{;j}g_{ik})]\\&\quad+\tfrac{1}{2}[\,\delta^h_j\{(L_\xi R)_{;l}g_{ik}-(L_\xi R)_{;k}g_{il}\}+\delta^h_k\{(L_\xi R)_{;j}g_{il}-(L_\xi R)_{;l}g_{ij}\}\\&\quad+\delta^h_l\{(L_\xi R)_{;k}g_{ij}-(L_\xi R)_{;j}g_{ik}\}].\end{aligned} \tag{5.14}$$

With the help of Eq. (3.9), Eq. (5.14) takes the form

$$L_\xi(Z^h_{ijk;l}+Z^h_{ikl;j}+Z^h_{ilj;k})=-R[\,\delta^h_j(\psi_{;l}g_{ik}-\psi_{;k}g_{il})+\delta^h_k(\psi_{;j}g_{il}-\psi_{;l}g_{ij})+\delta^h_l(\psi_{;k}g_{ij}-\psi_{;j}g_{ik})]. \tag{5.15}$$

In view of Eq. (5.3) we may write two more equations,

$$L_\xi(Z^h_{ikl;j})-(L_\xi Z^h_{ikl})_{;j}=(L_\xi\Gamma^h_{jm})Z^m_{ikl}-(L_\xi\Gamma^m_{ji})Z^h_{mkl}-(L_\xi\Gamma^m_{jk})Z^h_{iml}-(L_\xi\Gamma^m_{jl})Z^h_{ikm} \tag{5.16}$$

and

$$L_\xi(Z^h_{ilj;k})-(L_\xi Z^h_{ilj})_{;k}=(L_\xi\Gamma^h_{km})Z^m_{ilj}-(L_\xi\Gamma^m_{ki})Z^h_{mlj}-(L_\xi\Gamma^m_{kl})Z^h_{imj}-(L_\xi\Gamma^m_{kj})Z^h_{ilm}. \tag{5.17}$$

Adding Eqs. (5.3), (5.16), and (5.17) and making use of Eqs. (3.2), (3.5), and (5.15) we get

$$\begin{aligned}&-R[\,\delta^h_j(\psi_{;l}g_{ik}-\psi_{;k}g_{il})+\delta^h_k(\psi_{;j}g_{il}-\psi_{;l}g_{ij})+\delta^h_l(\psi_{;k}g_{ij}-\psi_{;j}g_{ik})]\\&=(\delta^h_l\psi_{;m}+\delta^h_m\psi_{;l}-g_{lm}\psi^h_{;})Z^m_{ijk}-(\delta^m_l\psi_{;i}+\delta^m_i\psi_{;l}-g_{li}\psi^m_{;})Z^h_{mjk}\\&\quad+(\delta^h_j\psi_{;m}+\delta^h_m\psi_{;j}-g_{jm}\psi^h_{;})Z^m_{ikl}-(\delta^m_j\psi_{;i}+\delta^m_i\psi_{;j}-g_{ji}\psi^m_{;})Z^h_{mkl}\\&\quad+(\delta^h_k\psi_{;m}+\delta^h_m\psi_{;k}-g_{km}\psi^h_{;})Z^m_{ilj}-(\delta^m_k\psi_{;i}+\delta^m_i\psi_{;k}-g_{ki}\psi^m_{;})Z^h_{mlj}.\end{aligned} \tag{5.18}$$

A straightforward contraction of the above equation at h and l with the help of Eq. (5.6) yields

$$\psi_{;m}Z^m_{ijk}+g_{ij}\psi^m_{;}Z_{mk}-g_{ki}\psi^m_{;}Z_{mj}=2R(\psi_{;i}g_{ik}-\psi_{;k}g_{ij}). \tag{5.19}$$

Again contracting the above equation by g^{ij} and making use of Eq. (1.9) we get $R=0$ (since $\psi_{;i}\neq 0$) and hence Eq. (1.9) gives $Z_{ij}=0$. Making use of these facts in Eq. (5.19) we get Eq. (5.10) and consequently Eq. (5.18) leads to

$$\psi_{,h}Z_{lijk}+\psi_{;i}Z^h_{ljk}=0. \tag{5.20}$$

Multiplying this equation by $\psi^l_{;}$ we get Eq. (5.11). Thus we may state the following theorem.

Theorem 5.2: If ξ is a proper Conh KV of a space–time with divergence-free conharmonic curvature tensor then the scalar curvature of the space–time vanishes and the space–time becomes either conharmonically flat or of type N.

Corollary 1: If ξ is a proper Conh KV in a space–time with a divergence-free conharmonic curvature tensor then we have

$$Z^h_{ijk;l}+Z^h_{ikl;j}+Z^h_{ilj;k}=0. \tag{5.21}$$

Proof: Proof is obvious from Eq. (5.13) and the fact that $R=0$.

Corollary 2: If a space–time V_4 has a divergence-free conharmonic curvature tensor and ξ is a proper Conh KV of the space–time, then the space–time must have a divergence-free Riemann curvature tensor.

Proof: Making use of the fact that $R=0$ in Eq. (2.2) we get Eq. (3.23) and hence the corollary.

[1] G. H. Katzin, J. Levine, and W. R. Davis, J. Math. Phys. **10**, 617 (1969).
[2] K. L. Duggal, J. Math. Phys. **33**, 2989 (1992).
[3] C. D. Collinson, Gen. Relativ. Gravit. **1**, 137 (1970).
[4] M. E. Cahill and A. H. Taub, Commun. Math. Phys. **21**, 1 (1971).
[5] H. Michalshi and J. Wainwright, Gen. Relativ. Gravit. **6**, 289 (1975).
[6] K. P. Singh and D. N. Sharma, Gen. Relativ. Gravit. **6**, 397 (1975).
[7] J. Wainwright and P. E. A. Yaremovicz, Gen. Relativ. Gravit. **7**, 345 (1976).
[8] J. Wainwright and P. A. E. Yaremovicz, Gen. Relativ. Gravit. **7**, 595 (1976).
[9] D. R. Oliver, Jr. and W. R. Davis, J. Math. Phys. **17**, 1790 (1976).
[10] L. K. Norris, L. H. Green, and W. R. Davis, J. Math. Phys. **18**, 1305 (1977).
[11] L. H. Green, L. K. Norris, D. R. Oliver, Jr., and W. R. Davis, Gen. Relativ. Gravit. **8**, 731 (1977).
[12] W. R. Davis and D. R. Olliver, Jr., Ann. Inst. Henri Poincaré **28**, 197 (1978).
[13] D. Eardley, J. Isenberg, J. Morsden, and V. Moncrief, Commun. Math. Phys. **106**, 137 (1986).
[14] G. Prasad and B. B. Sinha, Ann. Inst. Henri Poincaré **32**, 15 (1980).
[15] K. L. Duggal, J. Math. Phys. **30**, 1316 (1989).
[16] A. A. Coley and B. O. J. Tupper, Class. Quantum Grav. **7**, 2195 (1990).
[17] A. A. Coley and B. O. J. Tupper, Gen. Relativ. Gravit. **22**, 241 (1990).
[18] L. Herrera, J. Jimenez, L. Leal, and J. Ponce de Leon, J. Math. Phys. **25**, 3274 (1984).
[19] D. P. Mason and M. Tsamparlis, J. Math. Phys. **26**, 2881 (1985).
[20] R. Maartens, D. P. Mason, and M. Tsamparlis, J. Math. Phys. **27**, 2987 (1986).
[21] A. A. Coley and B. O. J. Tupper, J. Math. Phys. **30**, 2616 (1989).
[22] Y. Ishii, Tensor (N.S.) **7**, 73 (1957).
[23] Abdussattar and Babita Dwivedi, Bull. Calcutta Math. Soc. **88**, 465 (1996).
[24] D. Kalligas, P. Wesson, and C. W. F. Everitt, Gen. Relativ. Gravit. **24**, 351 (1992).
[25] W. R. Davis, L. H. Green, and L. K. Norris, Nuovo Cimento B **34B**, 256 (1976).
[26] K. P. Singh and D. N. Singh, Mon. Not. R. Astron. Soc. **140**, 453 (1968).
[27] L. Herrera and J. Ponce de Leon, J. Math. Phys. **26**, 2302 (1985).
[28] R. Maartens and M. S. Maharaj, J. Math. Phys. **31**, 151 (1990).
[29] A. A. Coley and B. O. J. Tupper, Class. Quantum Grav. **7**, 1961 (1990).
[30] A. A. Coley and B. O. J. Tupper, Class. Quantum Grav. **11**, 2553 (1994).
[31] R. Sharma, J. Math. Phys. **29**, 2421 (1988).
[32] M. C. Chaki and B. Gupta, Indian J. Math. **5**, 113 (1963).
[33] K. Yano, *The Theory of Lie Derivatives and its Applications* (Interscience, New York, 1957).
[34] R. Sharma, J. Math. Phys. **34**, 3582 (1993).

Relativistic spin networks and quantum gravity

John W. Barrett[a)]
Department of Mathematics, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom

Louis Crane[b)]
Department of Mathematics, Kansas State University, Manhattan, Kansas 66502

(Received 3 November 1997; accepted for publication 27 January 1998)

Relativistic spin networks are defined by considering the spin covering of the group SO(4), SU(2)×SU(2). Relativistic quantum spins are related to the geometry of the two-dimensional faces of a 4-simplex. This extends the idea of Ponzano and Regge that SU(2) spins are related to the geometry of the edges of a 3-simplex. This leads us to suggest that there may be a four-dimensional state sum model for quantum gravity based on relativistic spin networks that parallels the construction of three-dimensional quantum gravity from ordinary spin networks. © *1998 American Institute of Physics.* [S0022-2488(98)01005-6]

I. INTRODUCTION

In Refs. 1 and 2, it was proposed that the quantum theory of gravity could take the form of a very special type of discrete model on a triangulated 4-manifold, which has the property that the propagation of physical states does not depend on the choice of triangulation. Such a model is called a topological state sum.

Topological state sums are closely related to topological quantum field theories. Arguments that the quantum theory of gravity should be closely related to a topological quantum field theory are given in Ref. 3.

In Ref. 4, a specific family of four-dimensional topological state sums were constructed. It was suggested in Refs. 1 and 2 that the topological state sum of Ref. 4 might be a representation of a topological state for the quantum theory of gravity, and that the physical hypothesis that the universe is in that state could be studied.

Ooguri,[5] and, more recently, Baez[6] have argued that the topological state sum of Ref. 4 (reformulated in Ref. 7), which is a precise realization of Ref. 5, is a construction of the quantum theory of the BF Lagrangian.

In fact, the Plebanski Lagrangian[8,9] for general relativity has the form of BF theory[10] with gauge group $\mathrm{SL}(2,C)$, with an added term that is a constraint. Smolin[11] has argued that for general relativity with a cosmological constant the constraint vanishes for the self-dual sector. This suggests[1,2] that the self-dual sector can be identified with the topological state sum of Ref. 4.

The purpose of this paper is to outline a new state sum model, which is a sort of constrained doubling of the topological state sum of Ref. 4. This constraint is a quantum version of the Plebanski constraint.

In three dimensions, quantum gravity is constructed by the Ponzano–Regge state sum model in which representations of SU(2) label edges.[12] This corresponds to a path integral in which the metric is the variable. However, this model can also be understood as a path integral in the first-order formalism using a frame field and connection form.[13]

It is worth pursuing the link with the classical frame field to get a heuristic understanding of the discrete models. Integrating the frame field e, an $\mathbb{R}^3$-valued 1-form, along an edge, gives a vector in $\mathbb{R}^3$. This can also be regarded as an element of the dual of the Lie algebra so(3).

The idea is that to pass from three dimensions to four dimensions one has to replace SO(3) with SO(4), or its spin covering, SU(2)×SU(2).

a)Electronic mail: jwb@maths.nott.ac.uk
b)Electronic mail: crane@math.ksu.edu

 © 1998 American Institute of Physics

The topological home for this spin is now on triangles. This is in accord with the assignment of spins to triangles in Ref. 4, which can be regarded as a special case of $SU(2)\times SU(2)$, where one of the copies of $SU(2)$ acts trivially, thus rendering all variables self-dual. We are also making contact with the loop variables picture of Rovelli and Smolin,[14] in which the fundamental operators live on surfaces. The relation of the loop variables to state sum models, and the difference between BF theory and gravity in this context, was investigated in Refs. 15 and 16.

In a BF theory, the B field is an $so(4)^*$-valued 2-form, which can be integrated on a triangle to give an element of $so(4)^*$, the dual of the Lie algebra $so(4)$. However, in Einstein gravity it is important that $so(4)^*$ is identified with $\mathbb{R}^4\wedge\mathbb{R}^4$, as B is constrained to be of the form $B=*e\wedge e$, where e is a vector-valued one-form. In this formula, $e\wedge e$ is an $\mathbb{R}^4\wedge\mathbb{R}^4$-valued 2-form, and $*$ is the Hodge $*:\mathbb{R}^4\wedge\mathbb{R}^4\to\mathbb{R}^4\wedge\mathbb{R}^4$.

Assuming the frame field gives a linear embedding of the triangle in $\mathbb{R}^4$, the integral over this triangle,

$$\int e\wedge e,$$

is a *simple* bivector, i.e., of the form $f\wedge g\in\mathbb{R}^4\wedge\mathbb{R}^4$. Thus, the essential constraint that marks Einstein gravity from BF theory is the simplicity of bivectors.

This remark is the key to our construction of the discrete model, even though we do not have a classical frame field in the construction. The discrete models and the path integral formalism with classical frame fields do have close parallels, however.

A discretization of the classical Einstein gravity using bivectors on triangles that are constrained to be simple was previously studied in Refs. 17 and 18. Some related discrete classical models with self-dual bivectors, rather than simple bivectors, were studied in Refs. 18 and 19.

The new state sum model gives a quantum mechanical version of these constraints. The translation is to replace bivectors with bivector operators. The simplicity condition is then translated into a condition in the representation theory for these operators.

At present, we have an axiomatic description of the space of vertices, but not an explicit construction. If the only allowable vertex is the one we explicitly construct, then the state sum does not seem to give a topological invariant of manifolds. Thus, the relationship of our new state sum to topology is still unclear.

However, it is already very close to a topological theory, in that the deviation from the topological theory[4] is only in the imposed constraints. This should make an investigation of its ''ultraviolet'' behavior much simpler than for a continuum theory. Our hope is that the ''doubling'' of the topological state sum of Ref. 4 should allow us to pass from the self-dual sector to the full quantum theory of gravity.

The new state sum has a very natural relationship with Regge calculus in a new formulation, in that the classical version of the theory has the metric as its variable in the same way as Regge calculus. We believe, though do not have a complete argument, that it will contain ''local degrees of freedom'' as does general relativity, and that propagating plane waves will occur in a linearization of the theory, as in Refs. 20 and 21. However, we do not yet have an analog of the Ponzano–Regge asymptotic formula that would give a direct link with the Einstein action.

The heart of our construction is a reformulation of the geometry of simplexes in Euclidean (or Minkowski) 4-space in terms of variables on their 2-subsimplexes. This paper, accordingly, begins with a theorem in classical four-dimensional Euclidean geometry. The theorem allows us to describe 4-simplexes in terms of oriented area variables on 2-simplexes, giving a more precise version of the idea in Ref. 17. The constraints on the variables have a very elegant form, which can be carried over to the quantum theory.

Having discovered this form for the quantum constraints, it then becomes relatively straightforward to produce a new state sum.

II. BIVECTOR GEOMETRY OF A 4-SIMPLEX

An abstract 4-simplex is determined by an ordered set of five points, which we denote (01234). The faces of the abstract 4-simplex are just the subsets of this set.

A geometric 4-simplex is the convex hull of 5 points in $\mathbb{R}^4$. This embedding is required to be nondegenerate, i.e., the five points do not lie in any hyperplane. To each face now corresponds an affine subspace in $\mathbb{R}^4$, and also a linear subspace through the origin in $\mathbb{R}^4$ that is parallel to it. In particular, each triangle (2-simplex) determines a plane through the origin.

An embedding of an abstract 4-simplex in $\mathbb{R}^4$ is simply the set of position coordinates $x_0,x_1,x_2,x_3,x_4\in\mathbb{R}^4$ that determine the vertices of a geometric 4-simplex. The ordering of the vertices also determines an orientation for the simplex.

An oriented triangle in $\mathbb{R}^4$ determines a bivector. The position coordinates $x_0,x_1,x_2\in\mathbb{R}^4$ determine the displacement vectors for the edges $e_{01}=x_0-x_1$, $e_{02}=x_0-x_2$ and hence the bivector $b=\frac{1}{2}e_{01}\wedge e_{02}\in\mathbb{R}^4\wedge\mathbb{R}^4$. This bivector is invariant under an even permutation of the vertices but changes sign for an odd permutation.

In this way, a geometric 4-simplex determines a set of ten bivectors, one for each triangle. We show that the bivectors characterize the geometric 4-simplex uniquely, and give the conditions that an assignment of ten bivectors to the triangles of an abstract 4-simplex determines an embedding.

The properties of the bivectors for a 4-simplex are the following.

(1) The bivector changes sign if the orientation of the triangle is changed.

(2) Each bivector is simple, i.e., of the form $b=f\wedge g$.

(3) If two triangles share a common edge, then the sum of the two bivectors is also simple.

(4) The sum of the four bivectors corresponding to the faces of a tetrahedron is zero. This sum is calculated using the orientation of the triangles given by the boundary of the tetrahedron.

(5) The assignment of bivectors is nondegenerate. This means that for six triangles sharing a common vertex, the six bivectors are linearly independent.

(6) Using the Euclidean metric, the bivectors can be considered as linear operators. Then for three triangles meeting at a vertex of a tetrahedron, $\mathrm{Tr}\, b_1[b_2,b_3]>0$. The sign in this formula is determined by a convention relating a choice for the orientation for the boundary of the tetrahedron to the order in which the three bivectors appear in this formula.

Conditions (2), (3), and (4) were introduced in Ref. 17. We are indebted to John Baez,[22] who pointed out the necessity of (6) and the inversion ambiguity in the theorem that were missing in our first version of this paper.

Theorem: *Each geometric 4-simplex determines a set of bivectors satisfying these conditions, and each set of bivectors satisfying these conditions determines a geometric 4-simplex unique up to parallel translation and inversion through the origin.*

Proof: A simple bivector determines a plane (of dimension 2) through the origin in $\mathbb{R}^4$. Two simple bivectors b_1,b_2 determine two planes lying in a hyperplane (of dimension 3) if and only if b_1+b_2 is simple. If the planes lie in a hyperplane, this is clear because all bivectors are simple in three dimensions. Conversely, the simplicity condition can be understood as $b\wedge b=0$. Then

$$0=(b_1+b_2)\wedge(b_1+b_2)=2b_1\wedge b_2.$$

If $b_1=x\wedge y$, $b_2=z\wedge t$, then $0=x\wedge y\wedge z\wedge t$ implies a linear relation between x,y,z,t. This calculation also shows that (3) is insensitive to the orientation of the triangles.

For the bivectors for a geometric 4-simplex (4) is satisfied by Stokes' theorem. Condition (5) is true because the edges meeting at a vertex can be taken as axes for $\mathbb{R}^4$, and then the bivectors are clearly linearly independent. Condition (6) follows from the calculation $\mathrm{Tr}\, b_1[b_2,b_3]=\frac{9}{8}V^2$, where V is the volume of the tetrahedron.[23] This completes the proof that the conditions are satisfied by a geometric 4-simplex.

Now suppose a set of bivectors b_i satisfies (1)–(6). The simplicity conditions and the nondegeneracy imply that the planes corresponding to the four faces of a tetrahedron either all lie in a common hyperplane or share a common direction. However, if the bivectors share a common direction, then $\mathrm{Tr}\, b_1[b_2,b_3]=0$, contradicting (6). Therefore the four planes for a tetrahedron lie in a hyperplane. Furthermore, each plane is determined as the intersection of two of these hyperplanes, as each triangle is contained in two tetrahedra. The nondegeneracy condition implies that the five hyperplanes intersect generically, that is, any four of them intersect at a point.

To construct the geometric 4-simplex, shift one of the five hyperplanes away from the origin by parallel translation. The hyperplanes now bound a geometric 4-simplex, with two-dimensional faces parallel to the planes corresponding to the original bivectors b_i, $i=1,...,10$. This simplex has

ten bivectors, $b_i' = \lambda_i b_i$, for some ten scalars, $\lambda_i \neq 0 \in \mathbb{C}$. Condition (4) is the only linear relation on the bivectors for the faces of a tetrahedron. This implies that the λ_i are all equal for the faces of a tetrahedron, and hence are all equal for all ten triangles.

The geometric 4-simplex can be scaled by moving the hyperplane. If its distance from the origin is scaled by a parameter μ, then the bivectors all scale as μ^2. Therefore the λ_i can be scaled so that they are all equal to 1 or all equal to -1. Then the area bivectors b_i' of the reconstructed 4-simplex are either exactly equal to the original b_i, or all equal to $-b_i$. However, the b_i' also satisfy (6) as they are the bivectors of a geometric 4-simplex. Hence, the inequality in (6) forces the b_i' to be exactly equal to the b_i and not to their negatives.

The scaling parameter μ is only determined up to a sign, and so the geometric 4-simplex is determined up to inversion about the origin. This operation does not affect the area bivectors.

Metric geometry. Now introduce the standard Euclidean metric on $\mathbb{R}^4$ (one could also use the standard Minkowski metric). If the bivectors are replaced by their duals, $*b$, then conditions (1)–(5) are still satisfied, but $\mathrm{Tr}\, b_1[b_2, b_3] = 0$. This duality interchanges a set of planes lying in a hyperplane with a set of planes sharing a common direction.

The bivectors can be identified with elements of the Lie algebra so(4). (Also, it may be profitable to think of them as the elements of the dual of this Lie algebra.) The splitting $\mathrm{so}(4) \simeq \mathrm{su}(2) \oplus \mathrm{su}(2)$ is then the same as the splitting of the space of bivectors into self-dual and anti-self-dual parts, $\mathbb{R}^4 \wedge \mathbb{R}^4 = \Lambda^2_+ \oplus \Lambda^2_-$. The condition that a bivector b is simple is

$$0 = \langle b, *b \rangle = \langle b^+, b^+ \rangle - \langle b^-, b^- \rangle,$$

so that the norm of the self-dual and anti-self-dual parts is equal. The area A of the triangle is determined by

$$A^2 = \langle b, b \rangle = \langle b^+, b^+ \rangle + \langle b^-, b^- \rangle.$$

A geometric simplex has a metric geometry, which is determined by the lengths of its ten edges. Thus, two geometric 4-simplexes have the same edge lengths if and only if they are related by an isometry of $\mathbb{R}^4$. Therefore the bivector data determines a set of edge lengths. Two sets of bivector data determine the same edge lengths if and only if the bivectors are all transformed simultaneously by an orthogonal rotation.

By contrast, specifying the ten areas does not specify the metric geometry uniquely,[24] although it does for almost all configurations. However, giving the areas of the four faces of a tetrahedron does not specify the geometry of the tetrahedron uniquely, so that two 4-simplexes cannot be glued together by specifying the areas of the 16 triangles in their union.[25] For this reason, we decided that the theory should be based on bivectors rather than areas.

III. RELATIVISTIC SPIN NETWORKS

In the theory of spin networks, using the Lie group SU(2), each representation determines a map from vectors in $\mathbb{R}^3$ to operators in the representation space, namely the representation of the Lie algebra. In quantum physics, these are called vector operators. The intuition is that in the classical limit, when the representation is ''large,'' these behave as ordinary vectors in three-dimensional space, with probability near 1.

Given three irreducible representations on spaces V_a, V_b, V_c, the spin network trivalent vertex is a uniquely determined element of

$$V_a \otimes V_b \otimes V_c.$$

In a three-dimensional theory, this is the vertex for a triangle. This element is invariant under the group action. The corresponding vector operators satisfy

$$J_a + J_b + J_c = 0,$$

when applied to this state, as a consequence. This is the relation for the edge vectors of a triangle, or, equivalently, the bivectors in $\mathbb{R}^3$ that are dual to the edge vectors. Note that for other groups,

the space of vertices may form a vector space of dimension greater than unity. In these cases, it is necessary to pick a basis for these spaces and sum over the basis in a state sum model.

To define a theory of relativistic spin networks, we label the edges of a graph with irreducible representations of $SU(2)\times SU(2)$. These are pairs (j,k) of representations of $SU(2)$. Then the representation of the Lie algebra gives bivector operators, by analogy with the preceding case. The condition that these are simple bivectors can be transcribed into the quantum theory by requiring that $j=k$. This is the correct condition because classically it is requiring the self-dual and anti-self-dual parts to have the same norm.

We also have in mind using the q deformation of spin networks.[26] This is algebraically very similar to the undeformed case, with the advantage that the computations can be made finite by the choice of a root of unity. A disadvantage is the loss of the simple geometrical picture of bivectors. Also, the representations are braided. We take the braiding for the two copies of $SU(2)$ to be opposite, i.e., related by an orientation reversing map.

Now in the relativistic case, the theory is for four-dimensional manifolds and the relativistic spin network vertices reside in a tetrahedron. There will be one bivector operator for each face of a tetrahedron, and so this will be a four-valent vertex. The idea for relating spins on faces to the geometry of a tetrahedron is from Barbieri,[23] who considered a single $SU(2)$ spin.

The labeling of a 4-simplex is as follows. The triangles are labeled with representations of $SU(2)\times SU(2)$, and each tetrahedron is labeled with a tensor in the product of the four spaces on its faces.

(1) Changing the orientation of a triangle changes the representation to its dual.

(2) The representations on the triangles are of the form (j,j). These are called simple representations.

(3) For any pair of faces of a tetrahedron, the pair of representations can be decomposed into its Clebsch–Gordan series for $SU(2)\times SU(2)$. Under this isomorphism, the tensor for the tetrahedron decomposes into summands that are nonzero only for the simple representations of $SU(2)\times SU(2)$.

(4) The tensor for the tetrahedron is an invariant tensor.

These conditions are the quantum analogs of the properties (1)–(4) of bivectors on a 4-simplex given earlier. For the first condition, note that in the case of a classical Lie algebra, the action of the Lie algebra in the dual representation is minus the adjoint of the original action. We have not analyzed the inequalities (5) and (6), but note that a simple operator analog of (6) seems feasible. Operator versions of the expression in (6) are considered in the context of canonical quantization in Ref. 27.

We do not have a description of a basis for the space of vertices. However we can show that the space is nonempty in many cases, by exhibiting a canonical element. This vertex is defined by

$$\sum_j \dim j \;\rangle\!\!-\!\!\langle$$

where the middle edge is labeled by (j,j), and $\rangle\!-$ is the usual $SU(2)$ vertex doubled.

This looks unsymmetric because the faces of the tetrahedron have been split into two pairs. Each term in the sum obviously satisfies (2) and (4), and satisfies (3) for the two preferred pairs. The sum of the bivector operators for a pair is equal to the bivector operator for the middle edge in the vertex. This is a simple bivector operator because only simple representations of the form (j,j) appear.

However, the pairs can be switched by fusion, with the coefficients determined by the $6j$ symbol. The orthogonality relation of the $6j$ symbol shows that the vertex is given by the same formula if the top two and bottom two pairs are chosen. For the third splitting, the braidings of self-dual and anti-self-dual pieces cancel, so the same remark applies.

The classical limit. The simple representations of $SU(2)\times SU(2)$ occur as the summands of the Hilbert space of functions on S^3, which has the obvious action of $SO(4)$. The corresponding classical phase space is therefore the cotangent space T^*S^3. A Lie algebra element $l\in so(4)$ acts on functions on S^3 by the formula

$$l = l^{\mu\nu} x_\mu \frac{\partial}{\partial x^\nu}.$$

The corresponding momentum in the classical case is $l^{\mu\nu} x_\mu p_\nu$, for a point $(x,p) \in T^*S^3$. This is the dual pairing of so(4) with the simple bivector $x \wedge p$ in $\mathbb{R}^4$, which explains how simple bivectors are associated to triangles.

This phenomenon is familiar from the standard relativistic angular momentum theory, where the failure of simplicity of the angular momentum is associated with intrinsic spin. In our picture, intrinsic spin occurs for nonsimple representations, which would arise from sections of spinor bundles over S^3.

IV. THE NEW MODEL

It is easy to see how to transform the state sum of Ref. 4 to create a quantum version of summation over all assignments of bivector data to the 2-simplexes of a triangulated 4-manifold. We pick labelings of the 2-simplex of the triangulation by irreducible representations of $U_q(SU_2)$, just as in Ref. 4. For each such labeling of the 2-simplexes, we label the tetrahedra in each 4-simplex by elements of a basis for the vector space of vertices defined in the previous section (so that each tetrahedron in the interior of the manifold is labeled twice). In Ref. 4, a basis for the space of vertices was determined by specifying a labeling of the tetrahedron by a spin j. In the new model, each basis vertex will be a certain linear combination of the $\rangle\!\!-\!\!\langle$ labeled on the middle edge with (j,j).

There are two slightly different possibilities here. The first is to label tetrahedra in each 4-simplex with vertices as just described. The second possibility is to label the tetrahedron in the manifold with $\rangle\!\!-\!\!\langle$ labeled on the middle edge with (j,j) and use this in each of the two 4-simplexes it shares. The extra label j is then summed with the same weights as given by the vertex. The second sum is thus the ''diagonal'' subset of the first. Thus, we actually have two slightly different models, both of which deserve careful study. If vertices other than the canonical vertex exist, there will be still other variant state sums to consider.

In either case, a weight (in $\mathbb{C}$) is defined for the labeling, and the state sum model determines a number by summing the weights over all labelings.

For each 4-simplex, a weight is determined by taking the product of all the vertices for its tetrahedra to form a closed relativistic spin network. The weight for the manifold is determined by taking the product of all these weights with correction factors for the simplexes of dimension less than four. For example, on a tetrahedron, this is the inner product between the basis elements in the vector spaces of vertices for the tetrahedron in each of the two 4-simplexes it shares. This ensures that the state sum does not depend on the choice of basis for the space of vertices on tetrahedra in the interior. The lower-dimensional corrections would be analogous to the ones in Ref. 4, but modified to take into account the fact that we are quantum tracing over two copies of each representation at each step.

In Ref. 4 we combined the 15 spins on each 4-simplex into a diagram called a $15j$-q symbol that depended on the orientation of the 4-simplex. Since reversing orientation switches self-dual with anti-self-dual bivectors, in the new state sum the weights are linear combinations of the product of two $15j$-q symbols with identical representations but opposite orientation for each 4-simplex of the triangulated 4-manifold. Thus, the model is a ''constrained'' doubling of the Crane–Yetter model.

V. CONCLUSIONS

The one issue that still needs to be addressed in order for the state sum we are proposing to be a candidate for a quantum theory of gravity is how probability amplitudes computed with it behave as we refine the triangulation. This is analogous to the ultraviolet problems long familiar to quantum field theory and quantum gravity. It is important to note, however, that the nature of the problem has changed, since we no longer have a continuum metric. Instead of short distance behavior, we are concerned with fluctuations within fluctuations, with each fluctuation having a characteristic Planck length ''size.'' This at any rate is the most natural interpretation of the labelings in our state sum.

There are several possible avenues for constructing a finite theory from our state sum. The most elegant solution would be if there is topological invariance. It will take a good deal of careful calculation of relativistic spin networks to determine if this is the case.

Another possibility is to consider the possibility that below a certain subdivision the universe is in a self-dual, i.e., topological, state. However, this would require relaxing the constraints to allow self-dual states in the model. We could think of this hypothesis as related to the idea that the universe began in a topological state.[28]

Another possibility is to try to imitate our construction using either the category of representations of a larger quantum group or possibly a quantum supergroup. It is plausible that the sort of cancellations familiar from supergravity and supersymmetry will reappear in our context, and give us a finite theory. The advantage of this line of development over superstring theory is that if the cancellations do give a finite theory, it will be an exact theory rather than a perturbative expansion to something more mysterious.

It is interesting that the sort of tensor categories that go into the state sum we are proposing are so similar to the ones invented in constructing string theories.[29] Terms in our state sum can be interpreted as diagrams in string perturbation theory, by connecting together the diagrams we are associating to the 4-simplexes, and interpreting elements in the representations as string states.

At this point, perhaps the strongest thing that we can say of our model is that it transforms a number of important and familiar questions into statements about finite calculations that are well defined and can be investigated. We believe that this is a good sign in itself.

[1] L. Crane, ''Clock and category, is quantum gravity algebraic?,'' J. Math. Phys. **36**, 6180–6193 (1995).
[2] L. Crane, ''Topological field theory as the key to quantum gravity,'' *Knots and Quantum Gravity*, edited by J. C. Baez (Clarendon, Oxford, 1994).
[3] J. W. Barrett, ''Quantum gravity as topological quantum field theory,'' J. Math. Phys. **36**, 6161–6179 (1995).
[4] L. Crane, L. Kauffman, and D. Yetter, ''State sum invariants of 4-manifolds,'' J. Knot Theory Ram. **6**, 177–234 (1997).
[5] H. Ooguri, ''Topological lattice models in four dimensions Hirosi Ooguri,'' Mod. Phys. Lett. A **7**, 2799–2810 (1992).
[6] J. C. Baez, ''Four-dimensional BF theory as a topological quantum field theory,'' Lett. Math. Phys. **38**, 129–143 (1996).
[7] J. Roberts, ''Skein theory and Turaev–Viro invariants,'' Topology **34**, 771–788 (1995).
[8] J. F. Plebański, ''On the separation of Einsteinian substructures,'' J. Math. Phys. **18**, 2511–2520 (1977).
[9] R. Capovilla, J. Dell, T. Jacobson, and L. Mason, ''Self-dual 2-forms and gravity,'' Class. Quantum Grav. **8**, 41–58 (1991).
[10] G. T. Horowitz, ''Exactly soluble diffeomorphism invariant theories,'' Commun. Math. Phys. **125**, 417–437 (1989).
[11] L. Smolin, ''Linking topological quantum field theory and nonperturbative quantum gravity,'' J. Math. Phys. **36**, 6417–6455 (1995).
[12] G. Ponzano and T. Regge, ''Semiclassical limit of Racah coefficients,'' *Spectroscopic and Group Theoretical Methods in Physics*, edited by F. Bloch (North-Holland, Amsterdam, 1968), pp. 1–58.
[13] E. Witten, ''2+1 gravity as an exactly soluble system,'' Nucl. Phys. B **311**, 46–78 (1988).
[14] C. Rovelli and L. Smolin, ''Discreteness of area and volume in quantum gravity,'' Nucl. Phys. B **442**, 593–622 (1995).
[15] C. Rovelli, ''The basis of the Ponzano–Regge–Turaev–Viro–Ooguri model is the loop representation basis,'' Phys. Rev. D **48**, 2702–2707 (1993).
[16] M. P. Reisenberger and C. Rovelli, ''Sum over surfaces form of loop quantum gravity,'' Phys. Rev. D **56**, 3490–3508 (1997).
[17] H. Waelbroeck and J. A. Zapata, ''A Hamiltonian lattice formulation of topological gravity,'' Class. Quantum Grav. **11**, 989–998 (1994).
[18] J. A. Zapata, ''Topological lattice gravity using self-dual variables,'' Class. Quantum Grav. **13**, 2617–2634 (1996).
[19] M. P. Reisenberger, ''A left-handed simplicial action for Euclidean general relativity,'' Class. Quantum Grav. **14**, 1753–1770 (1997).
[20] J. W. Barrett, ''A convergence result for linearised Regge calculus,'' Class. Quantum Grav. **5**, 1187–1192 (1988).
[21] J. W. Barrett and R. M. Williams, ''The convergence of lattice solutions of linearised Regge calculus,'' Class. Quantum Grav. **5**, 1543–1556 (1988).
[22] J. C. Baez, ''Spin foam models,'' gr-qc/9709052.
[23] A. Barbieri, ''Quantum tetrahedra and simplicial spin networks,'' gr-qc/9707010.
[24] J. W. Barrett, ''First order Regge calculus,'' Class. Quantum Grav. **11**, 2723–2730 (1994).
[25] J. W. Barrett, M. Rocek, and R. M. Williams, ''A note on area variables in Regge calculus,'' gr-qc/9710056.
[26] L. H. Kauffman, ''Spin networks and knot polynomials,'' Int. J. Mod. Phys. A **5**, 93–115 (1990).
[27] J. Lewandowski, ''Volume and quantizations,'' Class. Quantum Grav. **14**, 71–76 (1997).
[28] E. Witten, ''Topological quantum field theory,'' Commun. Math. Phys. **117**, 353–386 (1988).
[29] G. Moore and N. Seiberg, ''Classical and quantum conformal field theory,'' Commun. Math. Phys. **123**, 177–254 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 6 JUNE 1998

Effects of spin-torsion in gauge theory gravity

Chris Doran,[a] Anthony Lasenby, Anthony Challinor, and Stephen Gull
MRAO, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 5 February 1997; accepted for publication 17 February 1998)

The spin-torsion sector of a new gauge-theoretic formulation of gravity is analyzed and the relationship to the Einstein–Cartan–Kibble–Sciama theory of gravity is discussed. The symmetries of the Riemann tensor and the conservation laws of the theory are derived. This formalism is applied to the problem of a Dirac field coupled self-consistently to gravity. The equations derived from a minimally coupled gauge-invariant Lagrangian naturally give the gauge-theoretic analogs of the Einstein–Cartan–Dirac equations. Finally, a semiclassical model for a spinning point-particle moving in a gravitational background with torsion is considered. © *1998 American Institute of Physics.* [S0022-2488(98)00706-3]

I. INTRODUCTION

The problem of formulating gravity as a gauge theory has been considered by many authors (see Refs. 1–8 for a representative sample). The gauge-theoretic approach leads naturally to an extension of general relativity known as spin-torsion theory.[9] Such theories build upon Cartan's suggestion[10] that torsion (the antisymmetric part of the connection) should be identified as a possible physical field. The connection between torsion and quantum spin emerged later[2,11,12] when it was realized that the stress-energy tensor for a massive fermion field was asymmetric.[13,14]

A new approach to gauge theory gravity (GTG) was developed in Refs. 15 and 16. In this approach, gravitational effects are described by a pair of gauge fields defined over a flat Minkowski background space–time. The gauge fields ensure the invariance of the theory under arbitrary local displacements and rotations in the background spacetime. All physical predictions are extracted in a gauge-invariant manner, which ensures that the background space–time plays no dynamic role in the physics. Most of the discussion in Ref. 15 was simplified by setting the torsion to zero. Even if the torsion does vanish, differences between GTG and general relativity still arise over global issues such as the role of topology and horizons. These differences are highlighted by the discussion of the Kerr singularity in Doran *et al.*[17–19] Further applications of GTG to situations with vanishing torsion can be found in Ref. 20.

In this paper we consider the spin-torsion sector of GTG in more detail. This is important both for theoretical considerations, such as spinor-driven inflationary models, and for comparisons with possible experiments or observations. Within GTG, torsion is viewed as a physical field derived from the gravitational gauge fields. This viewpoint has some conceptual advantages over that used in differential geometry, where torsion is regarded as a property of a non-Riemannian manifold. A further feature of the GTG approach is that the type of torsion which can arise physically is constrained.[15] This constraint follows from the requirement that minimally coupled equations for the matter fields should be derivable from a minimally coupled, gauge-invariant action.

The paper is arranged as follows. We begin with an introduction to the ''spacetime algebra.''[21,22] This algebraic structure is well suited to many aspects of relativistic physics, enabling manipulations to be carried out in a coordinate-free manner. The theory developed in Ref. 15 takes on its most compelling form when written in the space–time algebra, so we employ it here. One aim of this paper is to advertise the advantages of the algebraic techniques provided by space–time algebra, which extends the language of differential forms to incorporate spinors and quaternions, and provides very clean and clear methods for encoding rotations. These advantages are perhaps not as widely known as they deserve to be. The relations derived in this paper may be

[a]Electronic mail: C.Doran@mrao.cam.ac.uk

 © 1998 American Institute of Physics

translated into more conventional languages using the schemes described in Appendix D and in Ref. 15.

We next introduce the equations of GTG and discuss the relationship between GTG and the Einstein–Cartan–Kibble–Sciama (ECKS) theory of gravity. We then proceed to a covariant expression for the Riemann tensor in the presence of torsion. From this we are able to relate the Riemann tensor, and its contractions, to the equivalent tensors in the absence of torsion. We then consider the symmetries of the Riemann tensor and the gauge theory version of the Bianchi identity. As is well known, much of the symmetry of the Riemann tensor is lost when torsion is included.

The Dirac field coupled self-consistently to gravity provides an example of a system with torsion, which may be relevant to discussions of the early universe. The equations for this setup are derived from a minimally coupled, gauge-invariant action. The resulting equations are the GTG analogs of the Einstein–Cartan–Dirac equations. Appendix A contains a number of results for the gravitational fields particular to the case of spin generated by a single Dirac fermion. The analogs of the Einstein–Dirac equations are also given. These equations describe the Dirac field coupled to gravity via the (symmetric) stress-energy tensor only. In this arrangement the torsion vanishes. The Einstein–Dirac equations can also be derived from an action, but only if one is prepared to sacrifice the gauge structure of the theory. In Appendix B we give the GTG proof of the useful result, first given by Seitz,[23] relating solutions of the Einstein–Cartan–Dirac equations and solutions of the Einstein–Dirac equations.

Finally, we consider a simple, semiclassical model for a spin-1/2 point-particle moving in a gravitational background with torsion, which builds upon the model discussed in Ref. 15. This displays a coupling of the particle's motion to the torsion, unlike some previous models.[24] This model is used in a separate paper[25] to discuss the anisotropy of a new cosmological solution of the Einstein–Cartan–Dirac equations. Throughout the main part of this paper we assume that the spin tensor has a vanishing contraction, which is necessary for consistency with the minimal coupling procedure. Results for the case where the contraction of the spin tensor does not vanish are contained in Appendix C.

Natural units ($G=c=\hbar=1$) are used throughout this paper.

II. GAUGE THEORY GRAVITY

We begin with a brief introduction to the ''spacetime algebra''—the geometric (or Clifford) algebra of space–time. This is familiar to physicists in the guise of the algebra generated from the Dirac γ-matrices. Further details may be found in Refs. 21 and 22. The spacetime algebra (STA) is generated by four vectors $\{\gamma_\mu\}$, $\mu=0,\ldots,3$, equipped with an associative (Clifford) product, denoted by juxtaposition. The symmetrized and antisymmetrized products define the inner and outer products between vectors, denoted by a dot and a wedge, respectively:

$$\gamma_\mu\cdot\gamma_\nu\equiv\tfrac{1}{2}(\gamma_\mu\gamma_\nu+\gamma_\nu\gamma_\mu)=\eta_{\mu\nu}=\text{diag}(+---), \tag{1}$$

$$\gamma_\mu\wedge\gamma_\nu\equiv\tfrac{1}{2}(\gamma_\mu\gamma_\nu-\gamma_\nu\gamma_\mu). \tag{2}$$

The outer product of two vectors defines a bivector — a directed plane segment, representing the plane including the two vectors.

A full basis for the STA is provided by the set

1	$\{\gamma_\mu\}$	$\{\sigma_k, i\sigma_k\}$	$\{i\gamma_\mu\}$	i
1 scalar	4 vectors	6 bivectors	4 trivectors	1 pseudoscalar
grade 0	grade 1	grade 2	grade 3	grade 4

(3)

where $\sigma_k\equiv\gamma_k\gamma_0, k=1\ldots3$, and $i\equiv\gamma_0\gamma_1\gamma_2\gamma_3=\sigma_1\sigma_2\sigma_3$. The pseudoscalar i squares to -1 and anticommutes with all odd-grade elements. The $\{\sigma_k\}$ generate the geometric algebra of Euclidean 3-space which will be familiar as the algebra of the Pauli spin matrices. An arbitrary real super-

position of the basis elements (3) is called a ''multivector,'' and these inherit the associative Clifford product of the $\{\gamma_\mu\}$ generators. For a grade-r multivector A_r and a grade-s multivector B_s we define the inner and outer products via

$$A_r \cdot B_s \equiv \langle A_r B_s \rangle_{|r-s|}, \quad A_r \wedge B_s \equiv \langle A_r B_s \rangle_{r+s}, \tag{4}$$

where $\langle M \rangle_r$ denotes the grade-r part of M. The subscript 0 is dropped when denoting the scalar part of a multivector, $\langle M \rangle = \langle M \rangle_0$. We shall also make use of the commutator product,

$$A \times B \equiv \tfrac{1}{2}(AB - BA). \tag{5}$$

The operation of reversion, denoted by a tilde, is defined by

$$(AB)^{\sim} \equiv \tilde{B}\tilde{A}, \tag{6}$$

and the rule that vectors are unchanged under reversion. We adopt the convention that, in the absence of brackets, inner, outer, and commutator products take precedence over Clifford products.

Vectors are usually denoted in lower case Latin or Greek for a frame of orthonormal vectors. Let $\{e_\mu\}$ be an arbitrary basis for the grade-1 vector space. The reciprocal basis, denoted by $\{e^\mu\}$, satisfies $e_\mu \cdot e^\nu = \delta_\mu^\nu$. The vector derivative $\partial_a F$ of a multivector-valued function $F(a)$, with respect to the vector argument a, is defined by

$$\partial_a F(a) \equiv e^\mu e_\mu \cdot \partial_a F(a), \tag{7}$$

with the directional derivative operator $e_\mu \cdot \partial_a$ defined by the limit

$$e_\mu \cdot \partial_a F(a) \equiv \lim_{\epsilon \to 0} \left(\frac{F(a + \epsilon e_\mu) - F(a)}{\epsilon} \right). \tag{8}$$

We shall use the symbol ∇ for the vector derivative with respect to space–time position, so that $\nabla \equiv \partial_x$.

Linear functions mapping vectors to vectors are usually denoted with an underbar, $\underline{f}(a)$ (where a is the vector argument). The adjoint function $\overline{f}(a)$ satisfies

$$\underline{f}(a) \cdot b = a \cdot \overline{f}(b). \tag{9}$$

Linear functions extend to act on multivectors via the rule

$$\underline{f}(a \wedge b \wedge \cdots \wedge c) \equiv \underline{f}(a) \wedge \underline{f}(b) \wedge \cdots \wedge \underline{f}(c), \tag{10}$$

which defines a grade-preserving linear operation. The determinant of a linear function is defined by

$$\underline{f}(i) \equiv \det(\underline{f}) i. \tag{11}$$

Contractions of a linear function may be written as $\partial_a \cdot \underline{f}(a)$, and the protraction of a linear function is defined to be $\partial_a \wedge \underline{f}(a)$. Throughout, the argument of a linear function is assumed to independent of position, unless stated otherwise.

Gravitational effects are introduced via two gauge fields, $\overline{h}(a) = \overline{h}(a,x)$ and $\Omega(a) = \Omega(a,x)$, where x is the STA position vector $x = x^\mu \gamma_\mu$ and the dependence on x is often left implicit. The first of these, $\overline{h}(a)$, is a position-dependent linear function mapping the vector argument a to vectors. The gauge-theoretic purpose of $\overline{h}(a)$ is to ensure covariance of the equations under arbitrary local displacements of the matter fields in the background space–time.[15,16] The second gauge field, $\Omega(a)$, is a position-dependent linear function which maps the vector a onto the space of bivectors. Its introduction ensures covariance of the equations under local (Lorentz) rotations of fields, at a point, in the background space–time. It is useful to define another gauge field $\omega(a)$ by

$$\omega(a,x)\equiv\Omega(\underline{h}(a),x), \tag{12}$$

which also maps vectors to bivectors. This field has the advantage of transforming covariantly under position-gauge transformations $x\mapsto x'$:

$$\omega(a,x)\mapsto\omega'(a,x)\equiv\omega(a,x'). \tag{13}$$

From these gauge fields we assemble the covariant derivative $\mathcal{D}$, whose action on a multivector M is given by

$$\begin{aligned}\mathcal{D}M&\equiv\partial_a a\cdot\mathcal{D}M\\ &=\partial_a(a\cdot\overline{h}(\nabla)M+\omega(a)\times M),\end{aligned} \tag{14}$$

which also defines the operator $a\cdot\mathcal{D}$. We shall also make use of the operator $\mathcal{D}_a$ defined by

$$\mathcal{D}_a M\equiv a\cdot\nabla M+\Omega(a)\times M. \tag{15}$$

The covariant derivative contains a grade-raising and -lowering component, so that we may write

$$\mathcal{D}M=\mathcal{D}\cdot M+\mathcal{D}\wedge M, \tag{16}$$

where

$$\mathcal{D}\cdot M\equiv\partial_a\cdot(a\cdot\mathcal{D}M),\quad \mathcal{D}\wedge M\equiv\partial_a\wedge(a\cdot\mathcal{D}M). \tag{17}$$

The field strength corresponding to the $\Omega(a)$ gauge field is defined by

$$R(a\wedge b)\equiv a\cdot\nabla\Omega(b)-b\cdot\nabla\Omega(a)+\Omega(a)\times\Omega(b). \tag{18}$$

From this we define the covariant Riemann tensor

$$\mathcal{R}(a\wedge b)\equiv R[\underline{h}(a\wedge b)]. \tag{19}$$

The Ricci tensor, Ricci scalar and Einstein tensor are formed from contractions of the Riemann tensor:

$$\text{Ricci tensor:}\quad \mathcal{R}(a)=\partial_b\cdot\mathcal{R}(b\wedge a), \tag{20}$$

$$\text{Ricci scalar:}\quad \mathcal{R}=\partial_a\cdot\mathcal{R}(a), \tag{21}$$

$$\text{Einstein tensor:}\quad \mathcal{G}(a)=\mathcal{R}(a)-\tfrac{1}{2}a\mathcal{R}. \tag{22}$$

In any expression containing the symbol $\mathcal{R}$ the argument determines whether the Riemann or Ricci tensors, or Ricci scalar is implied.

Differential operators act on the multivector to their immediate right unless brackets are present. When acting on multivectors to which they are not immediately adjacent, the ''overdot'' notation is employed to denote the scope of a differential operator, so that

$$\dot{\nabla}A\dot{B}=\partial_a A a\cdot\nabla B. \tag{23}$$

As an example, Leibniz' rule can be written

$$\nabla(AB)=\nabla AB+\dot{\nabla}A\dot{B}. \tag{24}$$

The overdot notation extends to covariant derivatives. For example, for a general multivector-valued tensor $T(b,\ldots c)$ we define

$$a\cdot\dot{\mathcal{D}}\dot{T}(b,\ldots c)\equiv a\cdot\mathcal{D}T(b,\ldots c)-T(a\cdot\mathcal{D}b,\ldots c)-\cdots-T(b,\ldots a\cdot\mathcal{D}c). \tag{25}$$

The resultant object $a\cdot\dot{\mathcal{D}}\dot{T}(b,\ldots c)$ is also a covariant tensor. The utility of the $a\cdot\dot{\mathcal{D}}$ operator for tensors is that it commutes with contractions and protractions. The proof is straightforward, involving only the results that

$$\partial_b a\cdot\mathcal{D}\dot{T}(b,c,\ldots d)=a\cdot\dot{\mathcal{D}}\dot{T}'(c,\ldots d)-\omega(a)\cdot\partial_b T(b,c,\ldots d)-\partial_b T(\omega(a)\cdot b,c,\ldots d), \tag{26}$$

where $T'(c,\ldots d)\equiv\partial_b T(b,c,\ldots d)$, and

$$\partial_b\cdot\omega(a)T(b,\ldots)-\partial_b T(\omega(a)\cdot b,\ldots)=\partial_b T((\omega(a)\cdot b)\cdot\partial_e e,\ldots)-\partial_b T(\omega(a)\cdot b,\ldots)=0. \tag{27}$$

The field equations are obtained from variation of a suitable gauge-invariant action with respect to the gravitational gauge fields and the matter fields.[15] Varying with respect to the $\overline{h}$-field leads to the GTG analog of the Einstein equation:

$$\mathcal{G}(a)=\kappa\mathcal{T}(a), \tag{28}$$

where $\mathcal{T}(a)$ is the covariant matter stress-energy tensor (which need not be symmetric if spin is present) and $\kappa\equiv 8\pi$. On varying the $\Omega(a)$-field we find that

$$\mathcal{D}\wedge\overline{h}(a)=\kappa\mathcal{S}[\overline{h}(a)]+\tfrac{1}{2}\kappa[\partial_b\cdot\mathcal{S}(b)]\wedge\overline{h}(a), \tag{29}$$

where $\mathcal{S}(a)$ is the covariant spin tensor and a is independent of position. The definition of $\mathcal{S}(a)$ in terms of the matter Lagrangian is given in Ref. 15 and for the Dirac field it is discussed in Sec. V below.

The variation principle leading to (29) is different from that usually considered in the context of Poincaré gauge theory.[5,26] There one varies the action with respect to the metric and the torsion tensors, which are considered as separate dynamical variables. In GTG we vary with respect to the $\overline{h}(a)$ and $\Omega(a)$ gauge fields, the latter of which contains information about both the Christoffel connection and the torsion. In this respect the GTG approach is closer to the Palatini principle. The fact that $\Omega(a)$ is bivector valued and coordinate free removes the problems with the Palatini approach in ECKS theory first discussed by Sandberg.[27] This is because the ''projective transformation'' discussed in Ref. 27 is not a permissible transformation in GTG. Treating $\Omega(a)$ as a separate dynamical variable also recovers the link with a genuine gauge theory since it is the full $\Omega(a)$ that appears in the expression for minimal coupling to a Dirac wave function.

An important consequence of Eq. (29) is that minimally coupled equations for the matter fields are obtained from a minimally coupled Lagrangian only if the contraction of the spin tensor vanishes, that is, if

$$\partial_a\cdot\mathcal{S}(a)=0. \tag{30}$$

This is because Eq. (29) implies

$$\mathcal{D}_a[\overline{h}(\partial_a)\det(\underline{h})^{-1}]=-\tfrac{1}{2}\kappa\partial_a\cdot\mathcal{S}(a), \tag{31}$$

and the left-hand side must vanish for the minimally coupled equations to arise from the minimally coupled action. Minimal coupling at the level of the field equations is crucial in ensuring that GTG incorporates the equivalence principle (both classical and quantum). Minimal coupling at the level of the action, on the other hand, ensures that one has a genuine gauge theory and is important for constructing a quantum field theory. We therefore believe that Eq. (30) is a necessary requirement of all physical fields and this condition is assumed throughout the main part of this paper. Results for the case when $\partial_a\cdot\mathcal{S}(a)$ is not zero are given in Appendix C. For scalar fields and Yang–Mills gauge fields the spin tensor vanishes. For spin-1/2 fields the spin tensor is of the form

$$\mathcal{S}(a)=T\cdot a, \tag{32}$$

where T is the spin trivector. This form necessarily has vanishing contraction. The picture is slightly more complicated for spin-1 and spin-3/2 matter fields, where there is some freedom in the equations that one can work with. In both these cases, however, it is possible to construct theories in which Eq. (30) does hold. This work will be presented elsewhere.

Given that the spin tensor satisfies Eq. (30), Eq. (29) reduces to the ''wedge'' equation

$$\mathcal{D}\wedge\bar{h}(a)=\partial_b\wedge[b\cdot\mathcal{D}\bar{h}(a)]=\kappa\mathcal{S}[\bar{h}(a)], \tag{33}$$

or, in the case where a is position dependent,

$$\mathcal{D}\wedge\bar{h}(a)=\bar{h}(\nabla\wedge a)+\kappa\mathcal{S}[\bar{h}(a)]. \tag{34}$$

We refer to the left-hand side of Eq. (33) as the torsion. In the spin-torsion extension of general relativity, the torsion is measured by the antisymmetric part of the connection (the symmetric part is fixed by the requirement of metric compatibility) on the space–time manifold. The relationship between these two expressions is described in Appendix D.

The criterion that led us to restrict the torsion type also constrains the Lagrangian for the free gravitational field. In particular, it rules out the quadratic curvature terms in the Lagrangian often considered in the context of Poincaré gauge theory[5,28–30] This is because the higher-order terms in the field equations prevent one from deriving $\mathcal{D}_a\bar{h}(\partial_a\det(\underline{h})^{-1})=0$. The only additional terms that can be added to the Lagrangian are quadratic in the torsion.[5,31] Such terms contain derivatives of the $\bar{h}$-field, however, and so considerably complicate the theory. Indeed, the theory looks so simple and compelling with no derivatives of the $\bar{h}$-function present that it is tempting to view their absence as suggestive of a deeper principle. We discuss some ways that derivatives of the $\bar{h}$-field can be incorporated into the action in Sec. V, however, and it is a simple matter to adapt the present work for the presence of quadratic torsion terms.

Equations (28) and (29) are locally the same as those of ECKS theory. This can be made clear by introducing a coordinate frame and following the scheme described in Appendix D. Globally, however, the theories are not equivalent. This is because GTG is explicitly constructed as a gauge theory in a flat space–time and there is no possibility of the fields altering the topology of this background space–time. This leads to a number of differences between the two theories,[15,17,18] though these differences are not the subject of the present paper. Our aim here is to show that GTG, together with the techniques of space–time algebra, simplify many otherwise difficult derivations required in all of GTG, ECKS theory and general relativity.

This completes the definitions of the quantities required in this paper. Further details may be found in Ref. 15.

III. THE RIEMANN TENSOR

The definition (19) for the Riemann tensor can be manipulated into a manifestly covariant form. We begin by introducing the derivatives

$$L_a\equiv a\cdot\bar{h}(\nabla) \tag{35}$$

and writing

$$\mathcal{R}(a\wedge b)=L_a\omega(b)-L_b\omega(a)-\omega[\underline{h}^{-1}(L_a\underline{h}(b)-L_b\underline{h}(a))]+\omega(a)\times\omega(b), \tag{36}$$

where in this section the vectors a,b, etc. are assumed to be arbitrary functions of position. To eliminate the derivatives of the $\underline{h}$-function, we first write the ''wedge'' Eq. (33) in the form

$$\bar{h}(\dot{\nabla})\wedge\dot{\bar{h}}(c)=-\partial_d\wedge[\omega(d)\cdot\bar{h}(c)]+\kappa\mathcal{S}[\bar{h}(c)]$$

$$\Rightarrow\quad\langle b\wedge a\ \bar{h}(\dot{\nabla})\wedge\dot{\bar{h}}(c)\rangle=-\langle b\wedge a\ \partial_d\wedge[\omega(d)\cdot\bar{h}(c)]\rangle+\kappa\langle b\wedge a\ \mathcal{S}[\bar{h}(c)]\rangle$$

$$\Rightarrow\quad[\dot{L}_a\dot{\underline{h}}(b)-\dot{L}_b\dot{\underline{h}}(a)]\cdot c=[a\cdot\omega(b)-b\cdot\omega(a)]\cdot\bar{h}(c)-\kappa\bar{\mathcal{S}}(a\wedge b)\cdot\bar{h}(c), \tag{37}$$

where overdots denote the scope of the differential operator and $\overline{\mathcal{S}}(a)$ is the adjoint to the spin tensor, defined by

$$\overline{\mathcal{S}}(B)\cdot a = B\cdot\mathcal{S}(a), \tag{38}$$

for an arbitrary vector a and bivector B. The adjoint function $\overline{\mathcal{S}}(B)$ is a vector-valued linear function of its bivector argument B. It follows that

$$\begin{aligned}\underline{h}^{-1}(L_a\underline{h}(b)-L_b\underline{h}(a)) &= L_a b - L_b a + a\cdot\omega(b) - b\cdot\omega(a) - \kappa\overline{\mathcal{S}}(a\wedge b)\\ &= a\cdot\mathcal{D}b - b\cdot\mathcal{D}a - \kappa\overline{\mathcal{S}}(a\wedge b).\end{aligned} \tag{39}$$

If we denote the right-hand side of (39) by c, then we can write the Riemann tensor as

$$\mathcal{R}(a\wedge b) = L_a\omega(b) - L_b\omega(a) + \omega(a)\times\omega(b) - \omega(c), \tag{40}$$

which only differs from the torsion-free expression given in Ref. 15 by the definition of the vector c.

Evaluating the commutator of L_a and L_b gives

$$[L_a, L_b] = [L_a\underline{h}(b) - L_b\underline{h}(a)]\cdot\nabla = c\cdot\overline{h}(\nabla) = L_c, \tag{41}$$

where c is given by the right-hand side of (39). This bracket structure summarizes the content of the ''wedge'' equation (33) in a form that is independent of the torsion. The structure is essentially that of the Lie bracket, though its use in GTG is somewhat different. A new method of solving the field equations, which exploits this bracket structure, was given in Ref. 15. This ''intrinsic'' method has the advantage of yielding a set of first-order equations which, although still nonlinear, are generally more tractable than their second-order counterparts in general relativity.

When solving the field equations in the presence of torsion it proves useful to introduce the new field

$$\omega'(a) = \omega(a) + \kappa\mathcal{S}(a) - \tfrac{3}{2}\kappa a\cdot T, \tag{42}$$

where

$$T \equiv \tfrac{1}{3}\partial_a\wedge\mathcal{S}(a) \tag{43}$$

is the spin trivector. The $\omega'(a)$-field satisfies the equation

$$\mathcal{D}'\wedge\overline{h}(a) \equiv \partial_b\wedge(\dot{L}_b\dot{\overline{h}}(a) + \omega'(b)\cdot\overline{h}(a)) = 0 \tag{44}$$

and is the ω-function in the absence of torsion. We can derive an expression for $\omega'(a)$ in terms of the $\overline{h}$-function alone by inverting (44) to yield[15]

$$\omega'(a) = -H(a) + \tfrac{1}{2}a\cdot(\partial_b\wedge H(b)), \tag{45}$$

where

$$H(a) \equiv \overline{h}(\nabla\wedge\overline{h}^{-1}(a)) = -\overline{h}(\dot{\nabla})\wedge\dot{\overline{h}}(\overline{h}^{-1}(a)). \tag{46}$$

In GTG, $\omega'(a)$ plays the same role as the Christoffel symbol (the symmetric part of the connection) in the spin-torsion extension of general relativity.

On defining

$$\varpi(a) \equiv \omega(a) - \omega'(a) = -\kappa\mathcal{S}(a) + \tfrac{3}{2}\kappa a\cdot T, \tag{47}$$

we see that the Riemann tensor can be written as

$$\mathcal{R}(a\wedge b)=\mathcal{R}'(a\wedge b)+a\cdot\dot{\mathcal{D}}'\dot{\varpi}(b)-b\cdot\dot{\mathcal{D}}'\dot{\varpi}(a)+\varpi(a)\times\varpi(b), \tag{48}$$

where the primes denote the equivalent quantities in the absence of torsion. (In the literature these are often denoted with a pair of curly braces to denote the use of the Christoffel connection.)

On contracting Eq. (48) with ∂_a we find the following expression for the Ricci tensor:

$$\begin{aligned}\mathcal{R}(b)&=\mathcal{R}'(b)+\dot{\mathcal{D}}'\cdot\dot{\varpi}(b)-b\cdot\mathcal{D}'[\partial_a\cdot\varpi(a)]+\partial_a\cdot[\varpi(a)\times\varpi(b)]\\&=\mathcal{R}'(b)-\kappa\dot{\mathcal{D}}'\cdot\dot{\mathcal{S}}(b)-\tfrac{3}{2}\kappa b\cdot(\mathcal{D}'\cdot T)+\langle\partial_a\varpi(a)\varpi(b)\rangle_1-\partial_a\,\varpi(a)\cdot\varpi(b)\\&=\mathcal{R}'(b)-\kappa\dot{\mathcal{D}}'\cdot\dot{\mathcal{S}}(b)-\tfrac{3}{2}\kappa b\cdot(\mathcal{D}'\cdot T)-\kappa^2\overline{\mathcal{S}}[\mathcal{S}(b)]+\tfrac{3}{2}\kappa^2\overline{\mathcal{S}}(b\cdot T),\end{aligned} \tag{49}$$

where again $\mathcal{R}'(a)$ is the Ricci tensor in the absence of torsion and we have assumed a vanishing contraction of the torsion tensor. Finally, we make a further contraction to obtain the following expression for the Ricci scalar:

$$\mathcal{R}=\mathcal{R}'+\tfrac{9}{2}\kappa^2T^2-\kappa^2\mathcal{S}(\partial_a)\cdot\mathcal{S}(a). \tag{50}$$

No derivatives of $\mathcal{S}(a)$ appear in this expression, which is a feature of assuming that the spin tensor has a zero contraction.[5,31]

IV. SYMMETRIES AND CONSERVATION LAWS

In this section we discuss some of the symmetries of the Riemann tensor and conservation laws of GTG when torsion is present. These have already been given in Ref. 15, for the case of vanishing torsion. The zero torsion results may be combined with the expressions of the previous section to extend them to the torsion sector. However, for reasons of completeness, we shall not follow this route but shall give derivations from ''first principles.'' Many of these results have immediate counterparts in differential geometry, though the conciseness of their derivations here compares favorably with tensor-calculus techniques.

Our starting point is the equation

$$\mathcal{D}\wedge\overline{h}(c)=\kappa\mathcal{S}[\overline{h}(c)], \tag{51}$$

where c is independent of position. Taking the covariant exterior derivative of this equation gives

$$\mathcal{D}\wedge(\mathcal{D}\wedge\overline{h}(c))=\kappa\mathcal{D}\wedge\mathcal{S}[\overline{h}(c)]. \tag{52}$$

However,

$$\begin{aligned}\mathcal{D}\wedge(\mathcal{D}\wedge\overline{h}(c))&=\overline{h}(\partial_a)\wedge\mathcal{D}_a[\overline{h}(\partial_b)\wedge(\mathcal{D}_b\overline{h}(c))]\\&=[\mathcal{D}\wedge\overline{h}(\partial_b)]\wedge[\mathcal{D}_b\overline{h}(c)]+\overline{h}(\partial_a)\wedge\overline{h}(\partial_b)\wedge[\mathcal{D}_a\mathcal{D}_b\overline{h}(c)]\\&=\kappa\mathcal{S}[\overline{h}(\partial_b)]\wedge[\mathcal{D}_b\overline{h}(c)]+\tfrac{1}{2}\overline{h}(\partial_a)\wedge\overline{h}(\partial_b)\wedge[R(a\wedge b)\cdot\overline{h}(c)],\end{aligned} \tag{53}$$

so that

$$\begin{aligned}(2\kappa)^{-1}\partial_a\wedge\partial_b\wedge[\mathcal{R}(a\wedge b)\cdot\overline{h}(c)]&=\mathcal{D}\wedge\mathcal{S}[\overline{h}(c)]-\mathcal{S}[\overline{h}(\partial_b)]\wedge[\mathcal{D}_b\overline{h}(c)]\\&=\dot{\mathcal{D}}\wedge\dot{\mathcal{S}}[\overline{h}(c)]+\partial_a\wedge\mathcal{S}[a\cdot\mathcal{D}\overline{h}(c)]-\mathcal{S}(\partial_b)\wedge[b\cdot\mathcal{D}\overline{h}(c)].\end{aligned} \tag{54}$$

The final two terms on the right-hand side of Eq. (54) are

$$\begin{aligned}\partial_a\wedge\mathcal{S}\,[a\cdot\mathcal{D}\bar{h}(c)]-\mathcal{S}(\partial_b)\wedge[b\cdot\mathcal{D}\bar{h}(c)]&=\mathcal{S}(\partial_a)\wedge[\partial_b\ a\cdot(b\cdot\mathcal{D}\bar{h}(c))-a\cdot\mathcal{D}\bar{h}(c)]\\&=-\mathcal{S}(\partial_a)\wedge[a\cdot(\mathcal{D}\wedge\bar{h}(c))]\\&=-\kappa\mathcal{S}(\partial_a)\wedge[a\cdot\mathcal{S}(\bar{h}(c))],\end{aligned}\tag{55}$$

so Eq. (54) reduces to the covariant equation

$$\partial_a\wedge\partial_b\wedge[\mathcal{R}(a\wedge b)\cdot c]=2\kappa\dot{\mathcal{D}}\wedge\dot{\mathcal{S}}(c)-2\kappa^2\mathcal{S}(\partial_a)\wedge[a\cdot\mathcal{S}(c)].\tag{56}$$

Protracting Eq. (56) with ∂_c yields

$$\partial_a\wedge\partial_b\wedge\mathcal{R}(a\wedge b)=3\kappa\mathcal{D}\wedge T+\kappa^2\mathcal{S}(\partial_a)\wedge\mathcal{S}(a),\tag{57}$$

and dotting back with c gives

$$2\partial_b\wedge\mathcal{R}(c\wedge b)-\partial_a\wedge\partial_b\wedge[\mathcal{R}(a\wedge b)\cdot c]=3\kappa c\ \mathcal{D}\wedge T+\kappa^2 c\ \mathcal{S}(\partial_a)\wedge\mathcal{S}(a).\tag{58}$$

Hence, applying Eq. (56) again, we find that

$$\partial_a\wedge\mathcal{R}(a\wedge b)=-\kappa[\dot{\mathcal{D}}\wedge\dot{\mathcal{S}}(b)+\tfrac{3}{2}b\ \mathcal{D}\wedge T]+\kappa^2\mathcal{S}(\partial_a)\wedge[a\cdot\mathcal{S}(b)]-\tfrac{1}{2}\kappa^2 b\ \mathcal{S}(\partial_a)\wedge\mathcal{S}(a).\tag{59}$$

In the absence of spin the right-hand side of (59) vanishes, leaving the simple equation $\partial_a\wedge\mathcal{R}(a\wedge b)=0$ which, as noted in Ref. 15, encodes all the symmetries of the Riemann tensor in the absence of torsion.

The adjoint function to $\mathcal{R}(B)$ is written $\bar{\mathcal{R}}(B)$ and is defined by

$$A\cdot\mathcal{R}(B)=B\cdot\bar{\mathcal{R}}(A)\tag{60}$$

for any two bivectors A and B. The Riemann tensor and its adjoint are related by

$$\mathcal{R}(B)-\bar{\mathcal{R}}(B)=-\partial_b\wedge[B\cdot(\partial_a\wedge\mathcal{R}(a\wedge b))]+\tfrac{1}{2}B\ \partial_b\wedge\partial_a\wedge\mathcal{R}(a\wedge b)\tag{61}$$

and Eq. (59) can be used to express the right-hand side in terms of the spin. Clearly a vanishing spin tensor implies that $\mathcal{R}(B)=\bar{\mathcal{R}}(B)$, though the relation $\partial_a\wedge\mathcal{R}(a\wedge b)=0$ contains more information and is more useful in practice.

Returning to Eq. (59) and contracting with ∂_b we obtain

$$\partial_a\wedge\mathcal{R}(a)=\partial_a\wedge\mathcal{G}(a)=-\kappa\dot{\mathcal{S}}(\dot{\mathcal{D}})+\kappa^2[\partial_b\cdot\mathcal{S}(\partial_a)]\wedge[a\cdot\mathcal{S}(b)],\tag{62}$$

and the same expression must hold for the matter stress-energy tensor. The final term can be simplified by noting that

$$i\bar{\mathcal{S}}(iB)=i\partial_a\langle iB\mathcal{S}(a)\rangle=\mathcal{S}(a)\wedge(\partial_a\cdot B)\tag{63}$$

[since $\partial_a\cdot\mathcal{S}(a)=0$]. We can therefore write

$$[\partial_b\cdot\mathcal{S}(\partial_a)]\wedge[a\cdot\mathcal{S}(b)]=\partial_b\cdot[i\bar{\mathcal{S}}(i\mathcal{S}(b))]=-i\partial_b\wedge\bar{\mathcal{S}}(i\mathcal{S}(b)).\tag{64}$$

The gauge theory version of the Bianchi identity follows from the Jacobi identity applied to covariant derivatives of an arbitrary multivector M:

$$[\mathcal{D}_a,[\mathcal{D}_b,\mathcal{D}_c]]M+\text{ cyclic permutations}=0.\tag{65}$$

Evaluating the commutators we find that

$$\mathcal{D}_a R(b\wedge c)+\text{ cyclic permutations}=0,\tag{66}$$

which we want to replace by a covariant expression. We first extend to the case where a,b and c are position dependent:

$$\mathcal{D}_a R(b\wedge c)-R[(a\cdot\nabla b-b\cdot\nabla a)\wedge c]+\text{ cyclic permutations}=0. \tag{67}$$

Next, we replace a with $\underline{h}(a)$, etc. to find

$$a\cdot\mathcal{D}\mathcal{R}(b\wedge c)-\mathcal{R}[(\underline{h}^{-1}L_a\underline{h}(b)-\underline{h}^{-1}L_b\underline{h}(a))\wedge c]+\text{ cyclic permutations}=0. \tag{68}$$

Recalling Eq. (39), Eq. (68) becomes

$$a\cdot\mathcal{D}\mathcal{R}(b\wedge c)-\mathcal{R}[(a\cdot\mathcal{D}b-b\cdot\mathcal{D}a)\wedge c]+\kappa\mathcal{R}[\overline{\mathcal{S}}(a\wedge b)\wedge c]+\text{ cyclic permutations}=0, \tag{69}$$

which simplifies to give

$$a\cdot\dot{\mathcal{D}}\dot{\mathcal{R}}(b\wedge c)+\kappa\mathcal{R}[\overline{\mathcal{S}}(a\wedge b)\wedge c]+\text{ cyclic permutations}=0. \tag{70}$$

The adjoint form of this equation is

$$\partial_c\wedge\partial_b\wedge\partial_a\langle a\cdot\dot{\mathcal{D}}\dot{\mathcal{R}}(b\wedge c)B\rangle=-\kappa\partial_c\wedge\partial_b\wedge\partial_a\langle\mathcal{R}[\overline{\mathcal{S}}(a\wedge b)\wedge c]B\rangle, \tag{71}$$

where B is an arbitrary bivector. Equation (71) evaluates to

$$\dot{\mathcal{D}}\wedge\dot{\overline{\mathcal{R}}}(B)=\kappa\partial_a\ \mathcal{S}(a)\wedge\overline{\mathcal{R}}(B), \tag{72}$$

or, alternatively,

$$\partial_a\wedge[a\cdot\mathcal{D}\overline{\mathcal{R}}(B)-\overline{\mathcal{R}}(a\cdot\mathcal{D}B)]=\kappa\partial_a\ \mathcal{S}(a)\wedge\overline{\mathcal{R}}(B), \tag{73}$$

which is our gauge theory version of the Bianchi identity.

On contracting Eq. (72) we obtain

$$(\partial_a\wedge\partial_b)\cdot[\dot{\mathcal{D}}\wedge\dot{\overline{\mathcal{R}}}(a\wedge b)]=\kappa(\partial_a\wedge\partial_b\wedge\partial_c)\ \mathcal{S}(c)\wedge\overline{\mathcal{R}}(a\wedge b),$$

$$\Rightarrow\quad 2b\cdot\dot{\mathcal{D}}\ \dot{\overline{\mathcal{R}}}(\partial_b)-\mathcal{D}\mathcal{R}=\kappa(\partial_a\wedge\partial_b\wedge\partial_c)\ \mathcal{S}(c)\wedge\overline{\mathcal{R}}(a\wedge b), \tag{74}$$

which may be written as

$$\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}})\equiv a\cdot\dot{\mathcal{D}}\dot{\overline{\mathcal{G}}}(\partial_a)=-\kappa\partial_a\cdot\overline{\mathcal{R}}[\mathcal{S}(a)]+\kappa\partial_a\cdot\mathcal{S}[\overline{\mathcal{R}}(a)]. \tag{75}$$

Taking the inner product with an arbitrary vector b, we obtain the adjoint form

$$\dot{\mathcal{D}}\cdot\dot{\mathcal{G}}(a)=\kappa\mathcal{S}(\partial_b)\cdot\mathcal{R}(b\wedge a)-\kappa[a\cdot\mathcal{S}(\partial_b)]\cdot\mathcal{R}(b). \tag{76}$$

In the absence of torsion, we obtain the simple result

$$\dot{\mathcal{G}}(\dot{\mathcal{D}})=0, \tag{77}$$

which is the GTG version of covariant conservation of the Einstein tensor.

V. THE SELF-CONSISTENT DIRAC FIELD

A simple way to produce gauge fields with torsion is to consider matter described by a Dirac spinor field $|\psi(x)\rangle$. In the STA, it is convenient to represent such a spinor by an even multivector ψ (see Refs. 32 and 33 for an explicit map for the Dirac–Pauli representation). The $\{\hat{\gamma}_\mu\}$ operators, and the conventional unit scalar imaginary j, have actions which are represented by

$$\begin{aligned}\hat{\gamma}_\mu|\psi\rangle&\leftrightarrow\gamma_\mu\psi\gamma_0\quad(\mu=0\ldots3),\\ j|\psi\rangle&\leftrightarrow\psi i\sigma_3.\end{aligned} \tag{78}$$

In this manner, all matrix manipulations are eliminated, and the geometric content of the Dirac theory is made manifest.[34]

Spinors transform single sidedly under Lorentz rotations, which leads to the introduction of the spinor covariant derivative $D\psi \equiv \partial_a a \cdot D\psi$ where

$$a \cdot D\psi \equiv L_a \psi + \tfrac{1}{2}\omega(a)\psi. \tag{79}$$

The equations of motion for the Dirac field and the gauge fields are derived from the minimally coupled action,[15]

$$S_{\mathrm{ECD}} = \int |d^4x| \det(\underline{h})^{-1} \left[\tfrac{1}{2}\mathcal{R} - \kappa\langle D\psi i\gamma_3\tilde{\psi} - m\psi\tilde{\psi}\rangle\right], \tag{80}$$

where m is the mass of the particle. Varying with respect to ψ and the gravitational gauge fields leads to the system of equations

$$\begin{aligned} &\text{``wedge'':} && \mathcal{D}\wedge\overline{h}(a) = \kappa T \cdot \overline{h}(a) \\ &\text{Einstein:} && \mathcal{G}(a) = \kappa\mathcal{T}(a), \\ &\text{Dirac:} && D\psi i\sigma_3 = m\psi\gamma_0, \end{aligned} \tag{81}$$

where

$$T \equiv \tfrac{1}{2}\psi i\gamma_3\tilde{\psi} \tag{82}$$

is the spin trivector, and the matter stress-energy tensor $\mathcal{T}(a)$ is given by[15]

$$\mathcal{T}(a) = \langle a \cdot D\psi i\gamma_3\tilde{\psi}\rangle_1. \tag{83}$$

These equations are the gauge theory equivalents of the Einstein–Cartan–Dirac equations. The ''wedge'' equation summarizes Cartan's suggestion[10] that the torsion be identified with the spin of the matter present. Note that the minimally coupled Lagrangian has given rise to the minimally coupled Dirac equation, on variation of ψ. As noted in the Introduction, this is because ''trivector'' spin of the type $\mathcal{S}(a) = T \cdot a$ automatically has a vanishing contraction. A number of results for the gravitational field particular to this case of trivector spin are contained in Appendix A.

Protracting the stress-energy tensor (83) with ∂_a, we find that

$$\begin{aligned} \partial_a \wedge \mathcal{T}(a) &= \partial_a \wedge \langle a \cdot D\psi i\gamma_3\tilde{\psi}\rangle_1 \\ &= \tfrac{1}{2}\langle D\psi i\gamma_3\tilde{\psi} - \partial_a\psi i\gamma_3(a \cdot D\psi)^{\sim}\rangle_2. \end{aligned} \tag{84}$$

However, the Dirac equation implies that $\langle D\psi i\gamma_3\tilde{\psi}\rangle_2 = 0$, so we can write

$$\partial_a \wedge \mathcal{T}(a) = -\tfrac{1}{2}\langle \partial_a[a \cdot D\psi i\gamma_3\tilde{\psi} + \psi i\gamma_3(a \cdot D\psi)^{\sim}]\rangle_2 = -\langle \partial_a(a \cdot \mathcal{D}T)\rangle_2 = -\mathcal{D} \cdot T. \tag{85}$$

This demonstrates that the Dirac equation and the definition (83) for the stress-energy tensor are consistent with Eq. (62), since the final term on the right-hand side of Eq. (62) vanishes for the case of trivector spin. It can also be shown that the Dirac equation is consistent with the contracted Bianchi identity (75). The proof of this is simplified by employing the results of Appendix A. A new, exact solution to the Einstein–Cartan–Dirac equations in GTG, describing an almost homogeneous universe with a spin-induced anisotropy, is given in Ref. 25.

The gauge-invariant action (80) naturally gives rise to a spin-torsion theory, with the torsion (measured by the left-hand side of the ''wedge'' equation) given by the spin of the matter. Such an interaction is inevitable if the gravitational fields are introduced via a gauge theoretic route. This is good reason to prefer the Einstein–Cartan–Dirac equations over the Einstein–Dirac equa-

tions, which result from coupling the Dirac equation with the torsion-free space–time of general relativity. However, as most of the literature focuses on solutions of the Einstein–Dirac equations (see, for example, Refs. 35–38), it is interesting to see how the present techniques can be adapted to such a setup. We start by replacing $\omega(a)$ in the action by the equivalent expression in terms of $\overline{h}(a)$ and its derivatives. The action then becomes

$$S_{\mathrm{ED}}=\int |d^4x|\det(\underline{h})^{-1}\,[\tfrac{1}{2}\mathcal{R}'-\kappa\langle\overline{h}(\nabla)\psi i\gamma_3\tilde{\psi}+\tfrac{1}{4}\partial_a\wedge H(a)\psi i\gamma_3\tilde{\psi}-m\psi\tilde{\psi}\rangle], \tag{86}$$

where $H(a)$ is as defined in Eq. (46) and $\mathcal{R}'$ is now a function of $\overline{h}(a)$ and its first and second derivatives. The dynamical variables are now $\overline{h}(a)$ and ψ, with the ''wedge'' equation replaced by the identity

$$\mathcal{D}'\wedge\overline{h}(a)=0. \tag{87}$$

Variation with respect to ψ produces the equation

$$D'\psi i\sigma_3=m\psi\gamma_0, \tag{88}$$

where the prime again denotes that the covariant derivative does not contain a contribution from the spin. Variation with respect to $\overline{h}(a)$ yields

$$\mathcal{G}'(a)=\kappa\mathcal{T}'(a), \tag{89}$$

where $\mathcal{T}'(a)$ is obtained by variation of the matter action with respect to $\overline{h}(a)$. The derivation of $\mathcal{T}'(a)$ is more complicated now due to the presence of derivatives of the $\overline{h}(a)$-function in the matter Lagrangian. The calculations are not too difficult, however, and yield

$$\mathcal{T}'(a)=\langle a\cdot D'\psi i\gamma_3\tilde{\psi}\rangle_1+\tfrac{1}{2}a\cdot(\mathcal{D}'\cdot T). \tag{90}$$

Satisfyingly $\mathcal{T}'(a)$ is a symmetric tensor, which it has to be since $\mathcal{G}'(a)$ is automatically symmetric. The pair of equations (88) and (89) constitute the GTG version of the Einstein–Dirac equations. These are usually obtained from the system (81) by the Belifante–Rosenfield procedure.[39,40] It is perhaps surprising that to derive the equations rigorously from an action involves replacing the symmetric metric tensor $g_{\mu\nu}$ with $\overline{h}(a)$ as the dynamical variable, since it is the latter that unambiguously yields a symmetric stress-energy tensor.

Despite the neat features of the GTG derivation of the Einstein–Dirac set of equations, there are still good reasons to prefer the Einstein–Cartan–Dirac set, though of course no experiment has yet been devised which could distinguish between them. The key reason, as stated earlier, is that the Einstein–Cartan–Dirac set maintains the link with gauge theories, which is lost if one pushes the Riemannian geometry viewpoint of the Einstein–Dirac set. Such considerations are clearly important in attempting to construct a fully quantum theory. Furthermore, the clean separation into the $\overline{h}(a)$ and $\Omega(a)$ gauge fields of the GTG approach is particularly appealing in looking for extensions to a multiparticle quantum theory of gravity.

It is worth pointing out that if $\overline{h}(a),\omega(a)$ and ψ satisfy Eq. (81), then the fields $\overline{h}(a),\omega'(a)$ and ψ satisfy Eqs. (88) and (89) if the condition

$$T\psi=\tfrac{1}{2}\psi i\gamma_3\tilde{\psi}\psi=0 \tag{91}$$

is satisfied. It is a simple matter to show that this is equivalent to the condition[21]

$$\tilde{\psi}\psi=\psi\tilde{\psi}=0. \tag{92}$$

The proof of the analogous result between Einstein–Dirac and Einstein–Cartan–Dirac theory was given by Seitz in Ref. 23. The gauge theory proof is given in Appendix B. Condition (92) is always satisfied for neutrino solutions.

VI. SPINNING POINT-PARTICLE MODELS

In recent years there has been considerable interest in obtaining semiclassical equations describing the dynamics of a fermion, while maintaining the concept of a definite particle trajectory $x(\lambda)$ (see, for example, Refs. 41 and 42). One of the aims of this work is to obtain a quantum theory via a path-integral quantization of a suitable classical model.[43,44] In Ref. 15, an action was proposed to describe the dynamics of a spin-1/2 particle in a gravitational background, but the resulting equations were only given in a full classical approximation, which neglected the spin. In this section we derive the semiclassical equations of motion resulting from this action, including the effects of torsion in the gravitational background.

The action is given by

$$S=\int d\lambda \left\langle \dot{\Psi} i\sigma_3\tilde{\Psi}+\frac{1}{2}\Omega(\dot{x})\Psi i\sigma_3\tilde{\Psi}+p(v-m_p e\Psi\gamma_0\tilde{\Psi})+m_p^2 e\right\rangle, \tag{93}$$

where $v\equiv \underline{h}^{-1}(\dot{x})$ is the covariant tangent vector to the trajectory and, for this section, overdots denote differentiation with respect to λ. The Lagrange multiplier p is identified with the momentum of the particle (which has mass m_p) and an einbein e is included to ensure invariance of the equations under reparametrization of the path. The spinor $\Psi(\lambda)$ contains information about the spin and velocity of the particle. In general, p and v will not be collinear. The dynamical variables are $x(\lambda),\Psi(\lambda),p(\lambda)$ and $e(\lambda)$. Varying Ψ,p and e gives the equations

$$\dot{\Psi} i\sigma_3+\tfrac{1}{2}\omega(v)\Psi i\sigma_3=m_p ep\Psi\gamma_0, \tag{94}$$

$$v=m_p e\Psi\gamma_0\tilde{\Psi}, \tag{95}$$

$$p\cdot v=em_p^2. \tag{96}$$

Equation (94) may be written in manifestly covariant form

$$v\cdot D\Psi i\sigma_3=m_p ep\Psi\gamma_0. \tag{97}$$

Equations of motion for v and the spin bivector $S\equiv\Psi i\sigma_3\tilde{\Psi}$ follow from (94). We find that

$$ev\cdot\mathcal{D}(v/e)=-2m_p^2e^2p\cdot S, \tag{98}$$

$$v\cdot\mathcal{D}S=2p\wedge v. \tag{99}$$

In the presence of trivector type torsion the geodesic equation reduces to $v\cdot\mathcal{D}v=0$, where v is the unit tangent vector to the path (this is not true for more general types of torsion, in which case it is necessary to distinguish between geodesics and auto-parallels). It follows from (98) that, in general, the motion is not geodesic. Furthermore, if we construct the unit spin vector along $\Psi\gamma_3\tilde{\Psi}$, we find that this is not Fermi transported in general. These effects, which are present even in the absence of gravitation,[33] were present in the classical dipole model of Papapetrou.[45]

The equation of motion for p is obtained by variation of x. We find that

$$\partial_\lambda\overline{h}^{-1}(p)+\tfrac{1}{2}\partial_\lambda\partial_a(\Omega(a)\cdot S)=\overset{*}{\nabla}(\dot{x}\overset{*}{\overline{h}}^{-1}(p))+\tfrac{1}{2}\partial_a\langle a\cdot\overset{*}{\nabla}\overset{*}{\Omega}(\dot{x})S\rangle, \tag{100}$$

where, for this section, we employ overstars in place of overdots for the scope of a differential operator. This equation yields

$$\begin{aligned}\overline{h}^{-1}(\dot{p})+\underline{h}(v)\cdot(\overset{*}{\nabla}\wedge\overset{*}{\overline{h}}^{-1}(p))&=\tfrac{1}{2}\partial_a\langle(a\cdot\overset{*}{\nabla}\overset{*}{\Omega}(\dot{x})-\dot{x}\cdot\overset{*}{\nabla}\overset{*}{\Omega}(a))S-\Omega(a)\dot{S}\rangle\\&=\tfrac{1}{2}\overline{h}^{-1}[v\cdot\overline{\mathcal{R}}(S)]-\tfrac{1}{2}\partial_a\,\Omega(a)\cdot[\dot{S}+\omega(v)\times S].\end{aligned} \tag{101}$$

We now employ Eq. (99) on the right-hand side of (101) to find (after application of $\overline{h}$ to both sides)

$$\dot{p}-v\cdot[\overline{h}(\overset{*}{\nabla})\wedge\overset{*}{\overline{h}}(\overline{h}^{-1}(p))]=\tfrac{1}{2}v\cdot\overline{\mathcal{R}}(S)-\partial_a\,\omega(a)\cdot(p\wedge v). \tag{102}$$

Hence, recalling Eq. (33), we find the manifestly covariant equation of motion for p

$$v\cdot\mathcal{D}p=\tfrac{1}{2}v\cdot\overline{\mathcal{R}}(S)+\kappa v\cdot\mathcal{S}(p). \tag{103}$$

The presence of $\overline{\mathcal{R}}(S)$ and $\mathcal{S}(p)$ in this equation implies a coupling of the particle to the torsion, unlike some previous models.[24] We exploit this coupling in another paper[25] to discuss the torsion-induced anisotropy of a cosmological solution for a self-consistent Dirac field. We can use Eq. (99) to substitute for p in (103). This gives the equation

$$v\cdot\mathcal{D}\left(\frac{em_p^2v}{v^2}-\frac{v\cdot(v\cdot\mathcal{D}S)}{2v^2}\right)=\frac{1}{2}v\cdot\overline{\mathcal{R}}(S)+\kappa v\cdot\mathcal{S}(p), \tag{104}$$

which is the analog of the result found by Papapetrou[45] in the absence of torsion [his equation (5.7)].

The constraint equation (96) is consistent with the equations of motion for v and p since

$$\partial_\lambda(p\cdot v)=(v\cdot\mathcal{D}p)\cdot v+p\cdot(v\cdot\mathcal{D}v)=(\dot{e}/e)p\cdot v$$

$$\Rightarrow(p\cdot v)/e=\text{ constant}. \tag{105}$$

We thus have a complete, consistent set of semiclassical equations describing the dynamics of a massive spin-1/2 point-particle in a gravitational background. The equations are under-determined until a choice is made for the einbein. A convenient choice for comparison to classical models is to take $v^2=1$, so that λ is the proper time for the particle.

VII. CONCLUSIONS

In this paper we have discussed the symmetry of the Riemann tensor and the conservation laws of GTG in the presence of torsion. This extends many of the results discussed in Ref. 15, where the effects of torsion were not considered. Many of the results presented here have counterparts in the standard tensor-calculus formulation of ECKS theory. A full translation between the two approaches is contained in Appendix D.

As an example of a model with torsion, we considered the Dirac field coupled self-consistently to gravity. A minimally coupled gauge-invariant Lagrangian leads to the GTG version of the Einstein–Cartan–Dirac equations. The Einstein–Dirac equations may also be formulated consistently but they cannot be derived from minimal coupling and gauge-theoretic arguments alone. This is a consequence of the natural association of torsion with the spin of the matter field which arises in GTG. Finally, we considered a semiclassical model for a massive spinning point-particle, moving in a gravitational background with torsion. This showed that the motion is not generally geodesic, the spin vector is not Fermi-transported and the particle couples to the torsion.

In a separate paper[25] we apply much of this formalism to obtain a new self-consistent cosmological solution describing a universe arising from a massive Dirac field. This solution provides a counterexample to the claim of Ref. 26 that Einstein–Cartan theories of gravity and general relativity are observationally equivalent. The model presented in Ref. 25 contains an isotropic line element, but spinning point particles see a preferred direction in space due to the spin of the matter field. It is certainly true that for every solution of the Einstein–Cartan (or GTG) equations there is an equivalent solution in general relativity. There must be because every metric gives a solution to the Einstein field equations; the matter distribution is whatever the Einstein tensor dictates. This misses the point, however, that in theories with spin there are extra physical fields present which have observable consequences. These cannot just be hidden in an effective stress-energy tensor.

APPENDIX A: RESULTS FOR TRIVECTOR SPIN

The specialization of the results in the main text to the case of trivector spin is useful for studying the Einstein–Cartan–Dirac set of equations. In this setup the spin tensor is given entirely by

$$\mathcal{S}(a)=T\cdot a, \tag{A1}$$

where T is the spin trivector. First we have the following expressions for the Riemann tensor:

$$\mathcal{R}(a\wedge b)=\mathcal{R}'(a\wedge b)+\tfrac{1}{4}\kappa^2(a\wedge b)\cdot T\ T-\tfrac{1}{2}\kappa[(a\wedge b)\cdot\partial_c]\cdot(c\cdot\mathcal{D}T), \tag{A2}$$

with the prime denoting the equivalent tensor in the absence of spin. Contracting Eq. (A2) we obtain expressions for the Ricci tensor and scalar:

$$\mathcal{R}(a)=\mathcal{R}'(a)+\tfrac{1}{2}\kappa^2(a\cdot T)T-\tfrac{1}{2}\kappa a\cdot(\mathcal{D}\cdot T), \tag{A3}$$

$$\mathcal{R}=\mathcal{R}'+\tfrac{3}{2}\kappa^2T^2. \tag{A4}$$

These expressions are equivalent to those used in Ref. 23, but express the intrinsic content in a coordinate-free fashion.

The protraction of the Riemann tensor is

$$\partial_a\wedge\mathcal{R}(a\wedge b)=-\kappa(b\cdot\mathcal{D}T+\tfrac{1}{2}b\ \mathcal{D}\wedge T), \tag{A5}$$

which produces an antisymmetric term of

$$\overline{\mathcal{R}}(B)-\mathcal{R}(B)=\kappa(B\cdot\mathcal{D})\cdot T-\tfrac{1}{2}\kappa B\ \mathcal{D}\wedge T. \tag{A6}$$

Equation (A5) contracts to give

$$\partial_a\wedge\mathcal{R}(a)=\partial_a\wedge\mathcal{G}(a)=-\kappa\mathcal{D}\cdot T, \tag{A7}$$

so the antisymmetric part of the Einstein tensor is determined by the covariant divergence of the spin trivector.

The Bianchi identity simplifies to

$$\dot{\mathcal{D}}\wedge\dot{\overline{\mathcal{R}}}(B)=-\kappa\overline{\mathcal{R}}(B)\times T. \tag{A8}$$

The contracted result is

$$\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}})=\tfrac{1}{2}\kappa(\partial_b\wedge\partial_a)\cdot[\overline{\mathcal{R}}(a\wedge b)\times T]. \tag{A9}$$

Taking the inner product with an arbitrary vector b, we obtain the adjoint form

$$\dot{\mathcal{D}}\cdot\dot{\mathcal{G}}(b)=\tfrac{1}{2}\kappa(\partial_c\wedge\partial_a)\cdot\mathcal{R}[(a\wedge b\wedge c)\times T]. \tag{A10}$$

Equation (A9) may be written as

$$\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}})=\kappa i\partial_a\wedge\mathcal{R}(a\wedge(iT)), \tag{A11}$$

which, when combined with (A5) gives

$$\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}})=-\tfrac{1}{2}\kappa^2(s\mathcal{D}\cdot s+2s\cdot\mathcal{D}s), \tag{A12}$$

where the ''Pauli–Lubansky'' vector $s\equiv -iT$ is the dual of the spin trivector. Equation (A12) shows that $\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}})$ is second order in the spin.

APPENDIX B: SOLUTIONS OF THE EINSTEIN–CARTAN–DIRAC EQUATIONS AND THE EINSTEIN–DIRAC EQUATIONS

In this appendix we give the GTG proof that solutions of the equations (81) satisfying condition (92) are also solutions of the pair of equations (88) and (89). The proof is similar to that of Seitz.[23]

Assume that $\overline{h}(a), \omega(a)$, and ψ satisfy (81) and define $\omega'(a)$ via Eq. (45). The Dirac stress-energy tensor $\mathcal{T}(a)$, defined in Eq. (83), decomposes as follows:

$$\mathcal{T}(a)=\langle a\cdot D\psi i\gamma_3\tilde{\psi}\rangle_1=\langle a\cdot D'\psi i\gamma_3\tilde{\psi}\rangle_1+\tfrac{1}{2}\langle\varpi(a)\psi i\gamma_3\tilde{\psi}\rangle_1$$
$$=\mathcal{T}'(a)+\tfrac{1}{2}\kappa^2(a\cdot T)\cdot T-\tfrac{1}{2}\kappa a\cdot(\mathcal{D}\cdot T). \tag{B1}$$

We next use Eqs. (A3) and (A4) to write

$$\mathcal{G}(a)=\mathcal{G}'(a)-\tfrac{1}{2}\kappa a\cdot(\mathcal{D}\cdot T)+\tfrac{1}{2}\kappa^2(a\cdot T)\cdot T-\tfrac{3}{4}\kappa^2 aT^2. \tag{B2}$$

So, given that $\mathcal{G}(a)=\kappa\mathcal{T}(a)$, the equivalent equation for the primed tensors will also be satisfied provided that $T^2=0$. This holds if and only if $T\psi=0$.

It remains to consider the Dirac equation. Noting that

$$D'\psi=D\psi-\tfrac{1}{4}\kappa\partial_a a\cdot T\psi=D\psi-\tfrac{3}{4}\kappa T\psi, \tag{B3}$$

we see that Eq. (88) will also be satisfied provided that the condition $T\psi=0$ is satisfied. This completes the proof. Clearly, the converse is also true.

APPENDIX C: RESULTS FOR ARBITRARY SPIN TENSORS

In this Appendix we generalize the main results of this paper to a general spin tensor $\mathcal{S}(a)$ which is not required to have a vanishing contraction. The results are quoted without proof since the derivations follow those given in the main text.

For consistency with the previous results, we continue to denote the protraction of $\mathcal{S}(a)$ by $3T$. The contraction of the spin tensor is denoted t, so we have

$$\partial_a\mathcal{S}(a)=\partial_a\cdot\mathcal{S}(a)+\partial_a\wedge\mathcal{S}(a)=t+3T. \tag{C1}$$

Equation (39) generalizes to

$$\underline{h}^{-1}(L_a\underline{h}(b)-L_b\underline{h}(a))=a\cdot\mathcal{D}b-b\cdot\mathcal{D}a-\kappa\overline{\mathcal{S}}(a\wedge b)-\tfrac{1}{2}\kappa(a\wedge b)\cdot t, \tag{C2}$$

where $\overline{\mathcal{S}}(a)$ is the adjoint to the spin tensor, as defined at Eq. (38). The right-hand side of (C2) now defines the vector c which appears in the covariant expression (40) for the Riemann tensor. The $\omega'(a)$ field (the equivalent field in the absence of torsion) is given by

$$\omega'(a)=\omega(a)+\kappa\mathcal{S}(a)-\tfrac{1}{2}\kappa(3a\cdot T+a\wedge t). \tag{C3}$$

The Riemann tensor is still given by Eq. (48). On contracting we arrive at the following expression for the Ricci tensor:

$$\mathcal{R}(a)=\mathcal{R}'(a)-\kappa\dot{\mathcal{D}}'\cdot\dot{\mathcal{S}}(a)-\tfrac{3}{2}\kappa a\cdot(\mathcal{D}'\cdot T)-\tfrac{1}{2}\kappa a\mathcal{D}'\cdot t$$
$$-\kappa^2\overline{\mathcal{S}}(\mathcal{S}(a))+\tfrac{3}{2}\kappa^2\overline{\mathcal{S}}(a\cdot T)+\tfrac{1}{2}\kappa^2\overline{\mathcal{S}}(a\wedge t), \tag{C4}$$

which is the generalization of (49). Contracting this expression with ∂_a, we find the generalization of (50):

$$\mathcal{R}=\mathcal{R}'-\kappa\mathcal{D}'\cdot t+\tfrac{9}{2}\kappa^2T^2-\tfrac{1}{2}\kappa^2t^2-\kappa^2\mathcal{S}(\partial_a)\cdot\mathcal{S}(a). \tag{C5}$$

Of the full spin tensor $\mathcal{S}(a)$ only t is differentiated in this expression, though the corresponding term in the action is a total divergence, and so does not contribute to the field equations.

The protraction of the Riemann tensor [Eq. (59)] is given by

$$\partial_a \wedge \mathcal{R}(a \wedge b) = -\tfrac{1}{2}\kappa(b \wedge \mathcal{D} \wedge t + 2\dot{\mathcal{D}} \wedge \dot{\mathcal{S}}(b) + 3b\ \mathcal{D} \wedge T) + \tfrac{1}{4}\kappa^2(4i\overline{\mathcal{S}}[i\mathcal{S}(b)] - 2b\ \mathcal{S}(\partial_a) \wedge \mathcal{S}(a) + 2b \wedge \mathcal{S}(t) - 2\mathcal{S}(b) \wedge t + 3b\ T \wedge t). \tag{C6}$$

The protraction of the Ricci tensor [Eq. (62)] generalizes to

$$\partial_a \wedge \mathcal{R}(a) = -\kappa\dot{\mathcal{S}}(\dot{\mathcal{D}}) + \kappa^2 \partial_a \cdot (i\overline{\mathcal{S}}[i\mathcal{S}(a)]) - \tfrac{1}{2}\kappa^2 \mathcal{S}(t). \tag{C7}$$

For general torsion, the Bianchi identity is still given by the compact expression (72). The contracted Bianchi identity now becomes

$$\dot{\overline{\mathcal{G}}}(\dot{\mathcal{D}}) = \kappa[\overline{\mathcal{G}}(t) - \partial_a \cdot \overline{\mathcal{R}}(\mathcal{S}(a)) + \partial_a \cdot \mathcal{S}(\overline{\mathcal{R}}(a))]. \tag{C8}$$

Taking the inner product with a vector a gives the adjoint relation

$$\dot{\mathcal{D}} \cdot \dot{\mathcal{G}}(a) = \kappa[t \cdot \mathcal{G}(a) + \mathcal{S}(\partial_b) \cdot \mathcal{R}(b \wedge a) - (a \cdot \mathcal{S}(\partial_b)) \cdot \mathcal{R}(b)]. \tag{C9}$$

APPENDIX D: THE RELATION WITH TENSOR CALCULUS

In this appendix we relate the frame-free approach of GTG to the standard approaches of tensor calculus. These results extend those given in Appendix C of Ref. 15. Throughout this Appendix we follow the conventions of Ref. 9.

We start with a set of coordinates $\{x^\mu\}$ and introduce the vectors

$$e_\mu \equiv \frac{\partial x}{\partial x^\mu}, \quad e^\mu \equiv \nabla x^\mu. \tag{D1}$$

From these we construct the vectors

$$g_\mu \equiv \underline{h}^{-1}(e_\mu), \quad g^\mu \equiv \overline{h}(e^\mu). \tag{D2}$$

These vectors satisfy the relation

$$g_\mu \cdot g^\nu = e_\mu \cdot e^\nu = \delta_\mu^\nu. \tag{D3}$$

The metric is given by the 4×4 matrix

$$g_{\mu\nu} \equiv g_\mu \cdot g_\nu. \tag{D4}$$

If the x dependence in $g_{\mu\nu}$ is replaced by dependence on the coordinates $\{x^\mu\}$ alone, then Riemann–Cartan geometry is recovered.

The connection is defined by

$$\mathcal{D}_\mu g_\nu = \Gamma^\alpha_{\mu\nu} g_\alpha, \tag{D5}$$

so that

$$\Gamma^\lambda_{\mu\nu} = g^\lambda \cdot (\mathcal{D}_\mu g_\nu). \tag{D6}$$

Unlike the coordinate-free object $\Omega(a)$, the connection contains artifacts from the chosen coordinate frame. Since

$$\partial_\mu g_{\nu\lambda} = (\mathcal{D}_\mu g_\nu) \cdot g_\lambda + g_\nu \cdot (\mathcal{D}_\mu g_\lambda), \tag{D7}$$

we find that

$$\partial_\mu g_{\nu\lambda} = \Gamma^\alpha_{\mu\nu} g_{\alpha\lambda} + \Gamma^\alpha_{\mu\lambda} g_{\alpha\nu}, \tag{D8}$$

which recovers ''metric compatibility'' of the connection. The covariant derivative of a covariant vector $\mathcal{A} = A^\alpha g_\alpha = A_\alpha g^\alpha$ is

$$\mathcal{D}_\mu \mathcal{A} = \mathcal{D}_\mu (A^\alpha g_\alpha) = (\partial_\mu A^\alpha) g_\alpha + A^\alpha \Gamma^\beta_{\mu\alpha} g_\beta = (\partial_\mu A^\alpha + \Gamma^\alpha_{\mu\beta} A^\beta) g_\alpha, \tag{D9}$$

as expected.

Equation (D8) inverts to show that the connection contains a component given by the standard Christoffel symbol and a ''contorsion'' term

$$\Gamma^\nu_{\lambda\mu} = \{^\nu_{\lambda\mu}\} - K_{\lambda\mu}{}^\nu. \tag{D10}$$

The components of the contorsion tensor are given by

$$\begin{aligned} K_{\lambda\mu}{}^\nu = \{^\nu_{\lambda\mu}\} - g^\nu \cdot (\mathcal{D}_\lambda g_\mu) &= g^\nu \cdot [(\omega'(g_\lambda) - \omega(g_\lambda)) \cdot g_\mu] \\ &= \kappa (g^\nu \wedge g_\mu) \cdot [\mathcal{S}(g_\lambda) - \tfrac{3}{2} g_\lambda \cdot T - \tfrac{1}{2} g_\lambda \wedge t]. \end{aligned} \tag{D11}$$

The torsion tensor is defined by the antisymmetric part of the connection:

$$\begin{aligned} S_{\lambda\mu}{}^\nu = \tfrac{1}{2}(\Gamma^\nu_{\lambda\mu} - \Gamma^\nu_{\mu\lambda}) &= \tfrac{1}{2} g^\nu \cdot (\mathcal{D}_\lambda g_\mu - \mathcal{D}_\mu g_\lambda) \\ &= -\tfrac{1}{2}[g_\mu \cdot (\mathcal{D}_\lambda g^\nu) - g_\lambda \cdot (\mathcal{D}_\mu g^\nu)] \\ &= -\tfrac{1}{2}(g_\mu \wedge g_\lambda) \cdot (\mathcal{D} \wedge g^\nu) = \tfrac{1}{2}(g_\lambda \wedge g_\mu) \cdot [\mathcal{S}(g^\nu) + \tfrac{1}{2} t \wedge g^\nu], \end{aligned} \tag{D12}$$

and the modified torsion tensor is defined by

$$\begin{aligned} T_{\lambda\mu}{}^\nu &= S_{\lambda\mu}{}^\nu + \delta^\nu_\lambda S_{\mu\alpha}{}^\alpha - \delta^\nu_\mu S_{\lambda\alpha}{}^\alpha \\ &= \tfrac{1}{2}\kappa (g_\lambda \wedge g_\mu) \cdot \mathcal{S}(g^\nu). \end{aligned} \tag{D13}$$

For covariant quantities such as the Riemann tensor the translation to tensor calculus is straightforward:

$$R^\mu{}_{\nu\rho\sigma} = (g^\mu \wedge g_\nu) \cdot \mathcal{R}(g_\sigma \wedge g_\rho). \tag{D14}$$

A vierbein $e_\mu{}^i$ (essentially an orthonormal tetrad) is given by

$$e_\mu{}^i = g_\mu \cdot \gamma^i, \quad e^\mu{}_i = g^\mu \cdot \gamma_i, \tag{D15}$$

where $\{\gamma^i\}$ is a fixed orthonormal frame. Any position dependence in the $\{\gamma^i\}$ is eliminated with a suitable rotor transformation.

[1] R. Utiyama, ''Invariant theoretical interpretation of interaction,'' Phys. Rev. **101**, 1597 (1956).
[2] T. W. B. Kibble, ''Lorentz invariance and the gravitational field,'' J. Math. Phys. **2**, 212 (1961).
[3] F. Mansouri and L. N. Chang, ''Gravitation as a gauge theory,'' Phys. Rev. D **13**(12), 3192 (1976).
[4] Y. M. Cho, ''Gauge theory of Poincaré symmetry,'' Phys. Rev. D **14**(12), 3335 (1976).
[5] K. Hayashi and T. Shirafuji, ''Gravity from Poincaré gauge theory of the fundamental particles. I,'' Prog. Theor. Phys. **64**(3), 866 (1980).
[6] T. Eguchi, P. B. Gilkey, and A. J. Hanson, ''Gravitation, gauge theories and differential geometry,'' Phys. Rep. **66**(6), 213 (1980).
[7] E. A. Ivanov and J. Niederle, ''Gauge formulation of gravitation theories I. The Poincaré, de Sitter, and conformal cases,'' Phys. Rev. D **25**(4), 976 (1982).
[8] D. Ivanenko and G. Sardanashvily, ''The gauge treatment of gravity,'' Phys. Rep. **94**(1), 1 (1983).
[9] F. W. Hehl, P. von der Heyde, G. D. Kerlick, and J. M. Nester ''General relativity with spin and torsion: Foundations and prospects,'' Rev. Mod. Phys. **48**, 393 (1976).
[10] E. Cartan, ''Sur un généralisation de la notion de courbure de Riemann et les espaces à torsion,'' C. R. Hebd. Seances Acad. Sci. **174**, 593 (1922).
[11] H. Weyl, ''A remark on the coupling of gravitation and electron,'' Phys. Rev. **77**(5), 699 (1950).
[12] D. Sciama, ''The physical structure of general relativity,'' Rev. Mod. Phys. **36**, 463, 1103 (1964).

[13] J. Weyssenhoff and A. Raabe, ''Relativistic dynamics of spin-fluids and spin-particles,'' Acta Phys. Pol. **9**, 7 (1947).
[14] O. Costa de Beauregard, ''Translational inertial spin effect,'' Phys. Rev. **129**(1), 466 (1963).
[15] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''Gravity, gauge theories and geometric algebra,'' Philos. Trans. R. Soc. Lond., Ser. A **356**, 487 (1998).
[16] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''Astrophysical and cosmological consequences of a gauge theory of gravity,'' in *Advances in Astrofundamental Physics, Erice 1994*, edited by N. Sánchez and A. Zichichi (World Scientific, Singapore, 1995), p. 359.
[17] C. J. L. Doran, ''Integral equations and Kerr–Schild fields I. Spherically-symmetric fields,'' submitted to Class. Quantum Grav. (1997).
[18] C. J. L Doran, A. N. Lasenby, and S. F Gull, ''Integral equations and Kerr–Schild fields II. The Kerr solution,'' submitted to Class. Quantum Grav. (1997).
[19] A. N. Lasenby, C. J. L. Doran, Y. Dabrowski, and A. D. Challinor, ''Rotating astrophysical systems and a gauge theory approach to gravity,'' in *Current Topics in Astrofundamental Physics, Erice 1996*, edited by N. Sánchez and A. Zichichi (World Scientific, Singapore, 1997), p. 380.
[20] C. J. L. Doran, A. N. Lasenby, and S. F. Gull, ''Physics of rotating cylindrical strings,'' Phys. Rev. D **54**(10), 6021 (1996).
[21] D. Hestenes, *Space-Time Algebra* (Gordon and Breach, New York, 1966).
[22] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).
[23] M. Seitz, ''On solutions of the Einstein–Cartan–Dirac theory,'' Class. Quantum Grav. **2**, 919 (1985).
[24] A. Barducci, R. Casalbuoni, and L. Lusanna ''Classical spinning particles interacting with external gravitational fields,'' Nucl. Phys. B **124**, 521 (1977).
[25] A. D. Challinor, A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''Massive, non-ghost solutions for the Dirac field coupled self-consistently to gravity,'' Gen. Relativ. Gravit. **29**(12), 1527 (1997).
[26] J. M. Nester, ''Effective equivalence of the Einstein–Cartan and Einstein theories of gravity,'' Phys. Rev. D **16**, 2395 (1977).
[27] V. D. Sanberg, ''Are torsion theories of gravitation equivalent to metric theories?'' Phys. Rev. D **12**(10), 3013 (1975).
[28] R. T. Rauch, ''Equivalence of an $R+R^2$ theory of gravity to Einstein-Cartan-Sciama-Kibble theory in the presence of matter,'' Phys. Rev. D **26**(4), 931 (1982).
[29] R. D. Hecht, J. Lemke, and R. P. Wallner, ''Can Poincaré gauge theory be saved?'' Phys. Rev. D **44**(8), 2442 (1991).
[30] A. V. Khodunov and V. V. Zhytnikov, ''Gravitational equations in space–time with torsion,'' J. Math. Phys. **33**, 3509 (1992).
[31] K. Sundermeyer, ''Extra 'gauge' symmetries in $R+T^2$ gravitational theories,'' Phys. Lett. **134**(6), 415 (1984).
[32] C. J. L. Doran, A. N. Lasenby, and S. F. Gull, ''States and operators in the spacetime algebra,'' Found. Phys. **23**, 1239 (1993).
[33] C. J. L. Doran, A. N. Lasenby, S. F. Gull, S. S. Somaroo, and A. D. Challinor, ''Spacetime algebra and electron physics,'' Adv. Imaging Electron Phys. **95**, 271 (1996).
[34] D. Hestenes, ''Observables, operators, and complex numbers in the Dirac theory,'' J. Math. Phys. **16**, 556 (1975).
[35] M. G. Alé and L. P. Chimento, ''Self-consistent solutions of the semiclassical Einstein–Dirac equations with cosmological constant,'' Class. Quantum Grav. **12**, 101 (1995).
[36] U. Ochs and M. Sorg, ''Non-singular, cosmological solutions for the coupled Dirac–Einstein equations,'' J. Phys. A **28**, 7263 (1995).
[37] C. J. Radford and A. H. Klotz, ''An exact solution of the Einstein–Dirac equations,'' J. Phys. A **16**, 317 (1983).
[38] C. J. Isham and J. E. Nelson, ''Quantization of a coupled Fermi field and Robertson–Walker metric,'' Phys. Rev. D **10**, 3226 (1974).
[39] F. J. Belifante, ''On the spin angular momentum of mesons,'' Physica **6**, 887 (1939).
[40] F. J. Belifante, ''On the current and density of the electric charge,'' Physica **7**, 449 (1940).
[41] A. O. Barut and N. Zanghi, ''Classical models of the Dirac electron,'' Phys. Rev. Lett. **52**(23), 2009 (1984).
[42] S. M. Kuzenko, S. L. Lyakhovich, A. Yu. Segal, and A. A. Sharapov, ''Massive spinning particle on anti-deSitter space,'' Int. J. Mod. Phys. A **11**, 3307 (1996).
[43] A. O. Barut and I. H. Duru, ''Path integral formulation of quantum electrodynamics from classical particle trajectories,'' Phys. Rep. **172**(1), 1 (1989).
[44] P. L. Nash, ''Path-integral quantization of the dynamics of a classical point particle with intrinsic spin,'' Nuovo Cimento B **110**, 913 (1995).
[45] A. Papapetrou, ''Spinning test-particles in general relativity I.,'' Proc. R. Soc. London, Ser. A **209**, 248 (1951).

A dynamical symmetry breaking model in Weyl space

A. Feoli
Dipartimento di Scienze Fisiche, ''E. R. Caianiello,'' Università di Salerno, I-84081 Baronissi, Salerno, Italy

W. R. Wood
Faculty of Natural and Applied Sciences, Trinity Western University, 7600 Glover Road, Langley, British Columbia, V2Y 1Y1, Canada

G. Papini
Department of Physics, University of Regina, Regina, Saskatchewan, S4S 0A2, Canada

(Received 24 March 1997; accepted for publication 12 March 1998)

The dynamical process following the breaking of Weyl geometry to Riemannian geometry is considered by studying the motion of de Sitter bubbles in a Weyl vacuum. The bubbles are given in terms of an exact, spherically symmetric thin shell solution to the Einstein equations in a Weyl–Dirac theory with a time-dependent scalar field of the form $\beta=f(t)/r$. The dynamical solutions obtained lead to a number of possible applications. An important feature of the thin shell model is the manner in which β provides a connection between the interior and exterior geometries since information about the exterior geometry is contained in the boundary conditions for β. © *1998 American Institute of Physics.* [S0022-2488(98)02806-0]

I. INTRODUCTION

The success of models like the Glashow–Weinberg–Salam electroweak model have provided strong motivation for believing that symmetries of nature that were once manifest are no longer directly observable; one must extrapolate backwards from our broken symmetry state to the symmetric state that is described by the Lagrangian. The ultimate success of the standard model is contingent upon the existence of the Higgs particle—unless an additional degree of symmetry is broken.[1] This latter possibility, which involves the breaking of Weyl's scale invariance, has recently received further consideration.[2] While the topic of breaking Weyl geometry to Riemannian geometry has a relatively long history (see, e.g., Ref. 3), the analysis of the dynamics of thin shells that form in this process has been neglected. This problem is considered here by studying the dynamical properties of spherically symmetric thin-shell solutions to Einstein's equations with Weyl's scale invariance broken in the interior of the bubble.[4]

Shortly after Einstein presented his general theory of relativity, Weyl[5] attempted to unify the electromagnetic and gravitational fields within a geometrical framework. Einstein had shown that the rotation of a vector under parallel transport was related to the gravitational field; Weyl argued that the electromagnetic field would acquire a similar geometric status if the length of the vector was also allowed to vary according to $\delta\ell=\ell\kappa_\mu\delta x^\mu$. The covector field κ_μ, together with the metric tensor $g_{\mu\nu}$ that is defined modulo an equivalence class, comprised the fundamental fields of the new geometry. Weyl's geometry was rejected at the time because it did not permit an atomic standard of length. In more recent years, Weyl's theory has been reconsidered from several points of view (see, e.g., Refs. 4 and 6–11). Of particular interest here is the suggestion that atomic standards of length are introduced into a pure Weyl geometry through a symmetry breaking process.

The possibility that regions of space–time with broken scale invariance can coexist within regions of unbroken scale invariance has been demonstrated[4] by applying the Gauss–Mainardi–Codazzi (GMC) formalism in Weyl geometry to a Weyl–Dirac theory with a real scalar field β. The surface tension of the resulting static, spherically symmetric bubble arose from the boundary conditions imposed on the normal component of the derivative of the field β. In this way, a stable ''particle'' was constructed entirely from the fields $g_{\mu\nu}$, κ_μ, and β. By setting $\kappa_\mu=0$ and

 © 1998 American Institute of Physics

$\beta=$ constant in the interior space, the conformal invariance is broken and an ''atomic'' scale can be introduced into the theory. Generalizations of this model have been considered,[9–11] including the development of a geometric formulation of the causal interpretation of quantum mechanics[10] where β is a complex scalar field. In the present work, β is taken to be real and of the form $\beta(r,t)=f(t)/r$ in the exterior space V^E. This choice is useful in focusing on the dynamical properties of bubbles after they emerge in the ambient Weyl geometry.

The absence of a generalized Birkhoff theorem in Weyl space allows dynamical solutions to exist in the spherically symmetric case in the model under consideration. The analysis of exact, time-dependent solutions to the radial equation of motion shows that spherically symmetric bubbles that form in the ambient Weyl vacuum may, depending on the circumstances, collapse, remain at a fixed radius, oscillate indefinitely, or diverge to infinity. The spherically symmetric solution to the Einstein equations for the ansatz $\beta(r,t)=f(t)/r$ is presented in Sec. II. The analysis of the radial motion of the thin shell is provided in Sec. III and some concluding remarks are given in Sec. IV.

II. THE SPHERICALLY SYMMETRIC SOLUTION

The GMC formalism is here applied to the Weyl–Dirac theory given by

$$I_D=\int\left\{-\frac{1}{4}f_{\mu\nu}f^{\mu\nu}+\beta^2R+6\beta_{,\mu}\beta^{,\mu}-\lambda\beta^4\right\}\sqrt{-g}\;d^4x, \tag{2.1}$$

where $f_{\mu\nu}=\kappa_{\nu,\mu}-\kappa_{\mu,\nu}$. This action was introduced by Dirac[6] in order to study his Large Numbers hypothesis within the context of a theory of gravitation. Dirac's eloquent reasoning led him to propose a two-metric view that served to revive interest in Weyl's geometry. Generalizations of Dirac's action were subsequently proposed (see, e.g., Ref. 7) which accommodated Weyl's geometric interpretation of the electromagnetic field within a single-metric theory. This latter view is also maintained in the present paper.

The field equations that follow from (2.1) are the Maxwell and Einstein equations

$$\Box_\nu f^{\mu\nu}=0 \tag{2.2}$$

and

$$G_{\mu\nu}=\frac{1}{2\beta^2}E_{\mu\nu}+I_{\mu\nu}+\frac{1}{2}\lambda g_{\mu\nu}\beta^2\equiv T_{\mu\nu}, \tag{2.3}$$

where $E_{\mu\nu}$ is the usual Maxwell tensor and $\Box$ is the Riemannian covariant derivative,

$$I_{\mu\nu}=\frac{2}{\beta}(\Box_\nu\Box_\mu\beta-g_{\mu\nu}\Box_\alpha\Box^\alpha\beta)-\frac{1}{\beta^2}(4\beta_{,\mu}\beta_{,\nu}-g_{\mu\nu}\beta_{,\alpha}\beta^{,\alpha}), \tag{2.4}$$

and the field equation for β is identically the trace of (2.3).

For simplicity, the GMC formalism is applied to a spherically symmetric shell. The most general line element is then given by[12]

$$ds^2=-e^{\nu(r,t)}\,dt^2+e^{\mu(r,t)}\,dr^2+r^2(d\theta^2+\sin^2\theta\,d\phi^2). \tag{2.5}$$

The exterior and interior geometries are distinguished by writing $t_{E,I}$, $\nu_{E,I}$ and $\mu_{E,I}$ in the exterior and interior space–times $V^{E,I}$, respectively. It will be assumed here that the bubbles which form in the symmetry breaking process have interior geometries with $\kappa_\mu=0$ (which establishes length integrability in the region of space–time occupied by the bubble) and $\beta=\beta_0$ (which breaks the conformal invariance and fixes the scale in V^I according to the value of the constant β_0). Due to their continuity properties, $\kappa_\mu=0$ and $\beta=\beta_0$ on the thin shell Σ as well.[4] With the Maxwell tensor and $I_{\mu\nu}$ vanishing in V^I, the interior stress-energy tensor reduces to

$$T^I_{\mu\nu}=\tfrac{1}{2}\lambda g_{\mu\nu}\beta_0^2. \tag{2.6}$$

In what follows, λ will be taken to be negative. For this choice, and assuming no additional matter sources in V^I, the tensor (2.6) represents a de Sitter space and the interior metric is given by

$$e^{-\mu_I}=1+\tfrac{1}{6}\lambda\beta_0^2r^2=e^{\nu_I}, \tag{2.7}$$

with a horizon at $r=\beta_0^{-1}\sqrt{6/|\lambda|}\equiv r_h^I$.

In the exterior geometry V^E, β is expressed in the form

$$\beta(r,t)=f(t)/r, \tag{2.8}$$

where $f(t)$ is an arbitrary dimensionless function of t, and where $t\equiv t_E$ here and henceforth. Normally, one would proceed in applying the GMC formalism by solving the Einstein equations in V^E for the ansatz (2.8) and then using the resulting solution to obtain the equation of motion for the thin shell. In the present model a number of possible exterior metrics are known to exist. However, the condition that $\beta=\beta_0$ in V^I together with the boundary condition that β be continuous across Σ for all t requires that β also be a constant with respect to the intrinsic time τ of the thin shell defined at $r=R(\tau)$. That is,

$$\left.\frac{d\beta}{d\tau}\right|_{r=R}=\beta_{,t}X_E+\beta'\dot{R}=0, \tag{2.9}$$

where $X_E\equiv dt/d\tau$ and the prime and dot denote differentiation with respect to r and τ, respectively. The resulting condition,

$$X_E=-\frac{\beta'}{\beta_{,t}}\dot{R}, \tag{2.10}$$

leads to a restriction on the set of possible solutions to (2.3) in V^E. In effect, the condition (2.10) establishes a relationship between the radial and time derivatives of β at Σ that constrains the exterior metric since it is a functional of β. This constraint is satisfied when the exterior metric takes the form (see the Appendix)

$$e^{\nu_E}=-\frac{\beta_{,t}^2}{\beta'^2(1+\tfrac{1}{6}\lambda\beta^2r^2)} \tag{2.11}$$

and

$$e^{\mu_E}=-\frac{1+\tfrac{1}{6}\lambda\beta^2r^2}{\gamma^2\beta^2r^4}, \tag{2.12}$$

where γ is an arbitrary constant with dimension (length)$^{-1}$. If one assumes that the sign of the constant λ does not change during the formation of the bubble, then it must be taken to be negative to ensure the correct signature of the metric (see Ref. 4 for the conventions used here). For this choice, the signature $(-,+,+,+)$ is preserved for $r^2<(r_h^I)^2$ in V^I and for $r^2>(r_h^I)^2\beta_0^2/\beta^2$ in V^E. For the ansatz (2.8) this latter condition becomes $f^2(t)>(r_h^I)^2\beta_0^2$, so that the possibility of a time-dependent change of signature exists.

It is interesting to note that, in the spherically symmetric case with β of the form given in (2.8), the Einstein equations (2.3) also support the time-dependent solution

$$e^{\nu_E}=-\frac{\beta_{,t}^2}{\beta'^2(1+\tfrac{1}{6}\lambda\beta^2r^2+q^2/4\beta^2r^2)} \tag{2.13}$$

and

$$e^{\mu_E} = -\frac{1+\frac{1}{6}\lambda\beta^2 r^2 + q^2/4\beta^2 r^2}{\gamma^2\beta^2 r^4}. \tag{2.14}$$

However, when the exterior metric is matched onto the thin shell via the field β, the boundary conditions lead to the requirement that $q=0$. This suggests that, in the time-dependent, spherically symmetric case, a charged particle solution is not allowed. In contrast, the surface charge was an integral component of the static solution presented in Ref. 4.

III. THE MOTION OF THE THIN SHELL

With the interior and exterior metrics in hand, it is now possible to study the motion of the thin shell. In the GMC formalism, the equation of motion follows from the jump in the $\theta\theta$-component of the extrinsic curvature $K_{\mu\nu}$ across Σ. In terms of the interior metric (2.7) together with the ansatz (2.8), this equation can be expressed in the form [see (A12)]

$$1+\dot{R}^2+\frac{1}{6}\lambda\beta_0^2 R^2 = e^{(\mu-\nu)_E}\frac{\dot{R}^4}{X_E^2}. \tag{3.1}$$

Substituting (A8) for X_E^2 into (3.1) yields

$$1+\dot{R}^2+\frac{1}{6}\lambda\beta_0^2 R^2 = \frac{\dot{R}^4}{\dot{R}^2+e^{-\mu_E}}. \tag{3.2}$$

Using (2.12) at $r=R$ (with $\beta=\beta_0$) and solving for $\dot{R}^2$, one finds

$$\dot{R}^2 = \frac{a^2(1-R^2/R_{eq}^2)}{(R_{eq}^2/R^2-1)^2-a^2}, \tag{3.3}$$

where $a^2=\gamma^2\beta_0^2 R_{eq}^4$ and

$$R_{eq} = \frac{1}{\beta_0}\sqrt{\frac{6}{|\lambda|}} = r_h^I \tag{3.4}$$

results when $\dot{R}=0$. It is possible to obtain an analytical description of the radial motion from (3.3). The analysis of the motion of the thin shell is facilitated by breaking the problem into the following two cases.

A. Case 1: $R<R_{eq}$

The situation considered here applies to the case where a bubble of Riemannian geometry with $R<R_{eq}$ forms in the ambient Weyl space. In this case, the substitution of variables given by

$$x \equiv \frac{R_{eq}^2}{R^2} - 1 \tag{3.5}$$

allows (3.3) to be transformed into the integral form

$$I_1 \equiv \int \frac{1}{1+x}\sqrt{\frac{x^2-a^2}{x}}dx = \pm 2\gamma\beta_0 R_{eq}\tau. \tag{3.6}$$

Setting $x=\cosh p \equiv \mathrm{ch}p$, the integral I_1 becomes

$$I_1 = \frac{1}{\sqrt{a}}\left[\int \frac{(a\,\mathrm{ch}p-1)}{\sqrt{\mathrm{ch}p}}dp + (1-a^2)\left(\int \frac{dp}{\sqrt{\mathrm{ch}p}} - a\int \frac{\sqrt{\mathrm{ch}p}}{a\,\mathrm{ch}p+1}dp\right)\right]. \tag{3.7}$$

Expressing the three integrals in (3.7) in terms of elliptic integrals [see Eqs. (2.464.9), (2.464.10) and (2.464.15) in Ref. 13], one has

$$I_1=\sqrt{2a}\left\{\sqrt{2}\frac{\mathrm{sh}p}{\sqrt{\mathrm{ch}p}}-2E\left(\alpha,\frac{1}{\sqrt{2}}\right)+(1-a)\left[F\left(\alpha,\frac{1}{\sqrt{2}}\right)-\Pi\left(\alpha,\frac{1}{a+1},\frac{1}{\sqrt{2}}\right)\right]\right\}, \tag{3.8}$$

where $p\neq 0$, $\mathrm{sh}p\equiv\sinh p$, and

$$\alpha\equiv\arcsin\sqrt{\frac{\mathrm{ch}p-1}{\mathrm{ch}p}}=\arcsin\sqrt{\frac{R_{eq}^2-2R^2}{R_{eq}^2-R^2}}. \tag{3.9}$$

The case $a=1$ can be investigated easily because it involves only the elliptic integral of the second kind $E(\alpha,1/\sqrt{2})$. This choice does not represent a serious limitation due to the arbitrariness of the parameters γ and β_0 that appear in a. With this choice, Eq. (3.6) becomes

$$-\sqrt{2}E\left(\alpha,\frac{1}{\sqrt{2}}\right)+\frac{\mathrm{sh}p}{\sqrt{\mathrm{ch}p}}=\pm\frac{\tau}{R_{eq}}, \tag{3.10}$$

which is valid for $0<R<R_{eq}/\sqrt{2}$. For $R=0$ ($\alpha=\pi/2$, $\mathrm{ch}p=\infty$) the second term in (3.10) diverges, while $R=R_{eq}/\sqrt{2}$ leads to $\alpha=0$ and $\mathrm{ch}p=1$ that invalidate the integrations leading to (3.8). While R cannot be set equal to the endpoints of the interval over which it is defined (for $a=1$), it is of interest to consider the properties of the thin shell in the neighborhood of these points by means of expansions. To this end, it is useful to first invert (3.9), so that

$$\frac{R}{R_{eq}}=\sqrt{\frac{1-\sin^2\alpha}{2-\sin^2\alpha}}, \tag{3.11}$$

and to choose units for which $R_{eq}=1$.

1. Upper limit: $R\approx R_{eq}/2^{1/2}$

This limit corresponds to $\alpha\approx 0$. The expansion of (3.10) to third order in α yields

$$\pm\frac{\tau}{2}-\frac{\alpha^3}{3}\left(\frac{1}{4}+\frac{1}{2^{3/2}}\right)=0, \tag{3.12}$$

which has the solutions $\alpha=\pm 1.52042\tau^{1/3}$. Substituting this result into (3.11), one obtains an oscillatory motion with a period that increases with τ and that has a maximum amplitude $R\sim R_{eq}/\sqrt{2}$.

2. Lower limit: $R\approx 0$

This limit corresponds to $\alpha\approx\pi/2$. Expanding (3.10) to second order in $\alpha-\pi/2$ for the $+$ sign on the right-hand side, one obtains

$$-\frac{1}{\alpha-\pi/2}+\tau E\left(\frac{1}{\sqrt{2}}\right)+\left(\sqrt{1-\frac{1}{\sqrt{2}}}-\frac{1}{3(2)^{3/2}}\right)\left(\alpha-\frac{\pi}{2}\right)=0, \tag{3.13}$$

where $E(1/\sqrt{2})$ is the complete elliptic integral of the second type. The two solutions of (3.13), which are given by

$$\alpha=0.109314-0.835143\tau\pm 0.835143\sqrt{5.45723+3.4996\tau+\tau^2},$$

also lead to oscillatory motion when they are substituted into (3.11).

The expansion for the $-$ sign on the right-hand side of (3.10) leads to the two solutions

$$\alpha = 0.109314 + 0.835143\tau \pm 1.67249\sqrt{1.36071 - 0.872685\tau + 0.249342\tau^2}.$$

Of these, the one with the + sign in front of the square root yields oscillatory motion. The other solution, on the contrary, expands from $R \sim 0.4R_{eq}$ at $\tau = 0$ to the maximum $R \sim R_{eq}$ at $\tau \sim 1.5$, and then decreases asymptotically to $R = 0$.

In summary, it is has been shown that bubbles which are created in the Weyl vacuum with initial radii near either of the endpoints of the interval $0 < R < R_{eq}/\sqrt{2}$ exist indefinitely with a finite radius or else collapse to $R = 0$.

B. Case 2: $R > R_{eq}$

The situation considered here applies to the case where a bubble of Riemannian geometry with $R > R_{eq}$ is formed in the ambient Weyl vacuum. The substitution of variables given by

$$y \equiv 1 - \frac{R_{eq}^2}{R^2} \tag{3.14}$$

leads to the equation

$$I_2 \equiv \int \frac{1}{1-y}\sqrt{\frac{a^2-y^2}{y}}\, dy = \mp \frac{2a}{R_{eq}}\tau. \tag{3.15}$$

1. Upper limit: $R^2 \gg R_{eq}^2$

This limit corresponds to $y \sim 1$. Expanding the integrand of I_2 about this upper limit to order $(y-1)^{-1}$ one obtains

$$I_2 \sim \int \left[\frac{a^2+1}{2\sqrt{a^2-1}} - \frac{\sqrt{a^2-1}}{y-1}\right] dy = \mp \frac{2a}{R_{eq}}\tau. \tag{3.16}$$

Integration of (3.16) yields the result

$$y = 1 + \exp\left[\frac{a^2+1}{2(a^2-1)}y \pm \frac{2a\tau}{R_{eq}\sqrt{a^2-1}}\right]. \tag{3.17}$$

For sufficiently large values of τ, the result with the + sign leads to values of $y \gg 1$ (for $a^2 \neq 1$) which invalidates the approximation, whereas $y \to 1$ ($R \to \infty$) exponentially as $\tau \to \infty$ for the − sign in (3.17).

2. Lower limit: $R \approx R_{eq}$

By expanding the integrand of I_2 in the neighborhood of $y = 0$ to lowest order and integrating, one obtains

$$y^{1/2}\left(1 + \frac{y}{3}\right) \simeq \mp \frac{\tau}{R_{eq}}. \tag{3.18}$$

This equation has the solutions

$$y = 0 \quad \text{and} \quad y = -3 \pm 3\sqrt{1 + \left(\frac{\tau}{R_{eq}}\right)^2}. \tag{3.19}$$

The case $y = 0$ is trivial: if the bubble is created at $\tau = 0$ with $R = R_{eq}$, it will remain at the stable radius forever. The solution with the lower negative sign requires $y < 0$ and is unphysical in the present case ($0 < y < 1$). The remaining solution leads to

$$\frac{R_{eq}^2}{R^2} = 4 - 3\sqrt{1 + \left(\frac{\tau}{R_{eq}}\right)^2}. \tag{3.20}$$

In this case, a bubble that has a radius $R \gtrsim R_{eq}$ at $\tau = 0$ will expand to $R = \infty$ in a time $\tau = \sqrt{7/9} R_{eq}$.

3. Special case: a=1

In this case I_2 can be integrated exactly and yields [see Eq. (2.618.2) in Ref. 13]

$$I_2 = -2\sqrt{2} E\left(\alpha, \frac{1}{\sqrt{2}}\right) = \mp \frac{2\tau}{R_{eq}}, \tag{3.21}$$

where $\sqrt{2}\sin\varphi/2 = R_{eq}/R$, $0 < \varphi \leqslant \pi/2$ $(0 < R_{eq}/R \leqslant 1)$, and $\alpha = \arcsin(\sqrt{2}\sin\varphi/2)$. To second order in α, Eq. (3.21) becomes

$$\alpha = \mp \frac{\tau}{\sqrt{2} R_{eq}} \text{ or } \frac{R}{R_{eq}} = \mp \frac{1}{\sin(\tau/\sqrt{2} R_{eq})}.$$

In this case, R diverges whenever $\tau = n\pi\sqrt{2} R_{eq}$, and has minima of $R = R_{eq}$ for $\tau = (2n + 1)\pi^2/\sqrt{2}$.

In summary, bubbles that are created in the Weyl vacuum with an initial radius greater than R_{eq} appear to be unstable in all cases, except the trivial case of $R = R_{eq}$.

IV. DISCUSSION

It has been suggested, in the spirit of the electroweak model, that Weyl geometry should be considered a more fundamental geometry than Riemannian geometry since the latter represents a state of broken scale invariance. The feasibility of this proposal has been explored here by considering a Weyl–Dirac theory that supports a dynamical thin-shell solution to the Einstein equations. The fields that occur in the conformally invariant action are the source of the spherically symmetric bubbles that are posited; the geometric fields are the fabric from which the ''particles'' are constructed. The conformal symmetry is broken in the interior de Sitter geometry, and the surface tension is provided by the scalar field β. In principle, the bubbles could be associated with microscopic particles (see Ref. 10), perhaps born in a transition phase in the early universe, or they could be viewed on the cosmological scale. In the latter case, one could consider a single bubble of Riemannian geometry as representing our observable universe. In either case, the extrapolation backwards from our present state of broken scale invariance to the pre-particle Weyl vacuum lies within the domain of cosmology.

A common feature of the most successful cosmological theories is the presence of an inflationary epoch. An outstanding problem in such theories concerns the period of transition from the initial inflationary stage to the final stage of decelerating Friedmann–Robertson–Walker expansion. This transition is often described in terms of a first-order phase transition of the vacuum, wherein true vacuum bubbles are created and expand, leaving isolated regions of false vacuum. However, the precise physical process involved in the transition remains poorly understood. Kodama *et al.*[14] provided one of the earliest models of the transition phase in which an infinite number of true vacuum bubbles are nucleated simultaneously on a sphere, forming a shell-like true vacuum region around a de Sitter bubble. They showed that the trapped false vacuum domain either disappears leaving a black hole, or else remains as an ever-expanding domain connected with the outer region through a wormhole. However, it has been shown that the formation of bubbles in this manner is highly improbable.[15] Furthermore, Linde has pointed out[16] that the constant vacuum energy of de Sitter space, which has traditionally been viewed as necessarily leading to inflation (see, e.g., Ref. 17), can support other interpretations depending upon one's choice of coordinates. In the context of the present model, it is of interest to note that a variety of radial motions are possible, including an exponentially expanding solution. However, it remains an open question as to what interpretation should be given with regard to the identification of ''true'' and ''false'' vacua in the present model. Also, the precise conditions that precipitate the symmetry breaking process require further study.

In recent years, it has become popular to consider inflationary scenarios within the context of conformally invariant scalar field theories. In this paper, a theory of this type has been used to

study the dynamical properties of a bubble once it has formed in the Weyl vacuum. The development of the bubble model in the time-dependent case places a significant constraint on the set of possible spherically symmetric solutions to Einstein's equations. This constraint strongly links the properties of the thin shell to those of the exterior geometry. In the microscopic interpretation, a similar link has been shown[10] to lead to the guiding force postulated in the causal interpretation of quantum mechanics. In the cosmological interpretation, such a link is of interest because it allows for the possibility that information about the external metric may be available to an internal observer. This could prove to be of significance in the problem of boundary conditions of the universe.

The analysis of the radial equation of motion of the thin shell revealed the following results. Bubbles that emerge in the Weyl vacuum with $R>R_{eq}$ are unstable (except the trivial static case of $R=R_{eq}$) as they variously diverge to $R=\infty$. For bubbles with $R<R_{eq}$ initially, a number of possibilities exist. Of particular interest are those oscillatory solutions for which the radius remains finite. The freedom allowed by the parameters in the theory permit the scale associated with these bubbles to take on any value from the microscopic to the cosmological. In the first case, the results obtained here may prove to be of interest in the geometric causal interpretation of Ref. 10. In the latter case, one may find the present model useful in addressing some of the unresolved issues in inflationary cosmology. Indeed, the intimate relationship between particle physics and cosmology suggests that we should be considering theories of the very early universe that find application at any scale.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. One of the authors (G.P.) wishes to thank Dr. K. Denford, Dean of Science, University of Regina, for continued research support.

APPENDIX A: THE TIME-DEPENDENT MODEL

A time-dependent, spherically symmetric solution to the Einstein field equations (2.3) is, in general, obtained by solving the three equations

$$\frac{d}{dr}\,(re^{-\mu})=1+r^2\left[-\frac{q^2}{4\beta^2 r^4}-e^{-\mu}\left(2\frac{\beta''}{\beta}-\frac{\beta'^2}{\beta^2}-\mu'\frac{\beta'}{\beta}+\frac{4}{r}\frac{\beta'}{\beta}\right)\right.$$
$$\left.+e^{-\nu}\left(\mu_{,t}\frac{\beta_{,t}}{\beta}+3\frac{\beta_{,t}^2}{\beta^2}\right)+\frac{1}{2}\lambda\beta^2\right], \tag{A1}$$

$$\frac{d}{dr}\,(\mu+\nu)=r\left[2\frac{\beta''}{\beta}-4\frac{\beta'^2}{\beta^2}-(\mu'+\nu')\frac{\beta'}{\beta}\right]+re^{\mu-\nu}\left[2\frac{\beta_{,tt}}{\beta}-4\frac{\beta_{,t}^2}{\beta^2}-(\mu+\nu)_{,t}\frac{\beta_{,t}}{\beta}\right], \tag{A2}$$

and

$$\frac{d}{dt}\,(e^{-\mu})=-re^{-\mu}\left(2\frac{\beta'_{,t}}{\beta}-\nu'\frac{\beta_{,t}}{\beta}-\mu_{,t}\frac{\beta'}{\beta}-4\frac{\beta'}{\beta}\frac{\beta_{,t}}{\beta}\right). \tag{A3}$$

The solutions to these equations in the interior and exterior spaces are then generally used in determining the equation of motion $r=R(\tau)$ for the thin shell in the frame comoving with the thin shell with proper time τ. In the GMC formalism, one introduces the spherically symmetric intrinsic metric

$$ds_\Sigma^2=-d\tau^2+R^2(\tau)(d\theta^2+\sin^2\theta\,d\phi^2) \tag{A4}$$

and the extrinsic curvature[4]

$$K_{\mu\nu}=-\tfrac{1}{2}n^\alpha h_{\mu\nu,\alpha}, \tag{A5}$$

where n^μ is the unit spacelike vector field normal to Σ and $h_{\mu\nu}$ are the coefficients of the intrinsic metric on Σ. The equation of motion of the thin shell can be obtained from the jump in the $\theta\theta$-component of the extrinsic curvature:[4]

$$(K^\theta_{\ \theta})^E-(K^\theta_{\ \theta})^I=\sigma, \tag{A6}$$

where 2σ is the surface energy density of the thin shell,

$$(K^\theta_{\ \theta})^{E,I}=-\frac{1}{R}e^{-\mu_{E,I}}X_{E,I}, \tag{A7}$$

and

$$X^2_{E,I}=e^{(\mu-\nu)_{E,I}}\dot{R}^2+e^{-\nu_{E,I}}. \tag{A8}$$

The surface energy σ is generally taken to be an arbitrary parameter in models involving domain walls. However, by identifying this surface energy with the discontinuity in the normal derivative of the field β across Σ, the ''particle'' proposed in Ref. 4 was shown to be an entity that is derived entirely from the fields in the theory.

In the time-dependent, spherically symmetric theory under consideration, the scalar field β establishes an intimate link between V^I, the thin shell, and V^E. The dynamical implications of this link can be seen as follows. In the time-dependent case,

$$n^\mu=e^{(\mu+\nu)/2}(e^{-\nu}\dot{R},e^{-\mu}X,0,0) \tag{A9}$$

and

$$\sigma=e^{-(\mu+\nu)_E/2}\frac{\omega}{X_E}, \tag{A10}$$

where $\omega\equiv(\beta'/\beta)|_{r=R}$. Equation (A10) provides a correction to (4.3) in Ref. 11 where n^μ was not normalized to unity. Using (A9) and (A10), the equation of motion (A6) for the thin shell becomes

$$e^{(\nu-\mu)_I/2}X_I=e^{(\nu-\mu)_E/2}X_E+e^{-(\mu+\nu)_E/2}\frac{\omega R}{X_E}. \tag{A11}$$

For $\beta(t,r)=f(t)/r$, it follows that $\omega R=-1$. Using this, together with the interior metric (2.7), the square of (A11) becomes

$$1+\dot{R}^2+\frac{1}{6}\lambda\beta_0^2R^2=e^{(\mu-\nu)_E}\frac{\dot{R}^4}{X_E^2}. \tag{A12}$$

Now, there are two distinct expressions for X_E^2 in the present model. First, there is (A8) that follows from the form of $dt/d\tau$ in the spherically symmetric case. Then there is (2.10) that follows from the choice of boundary conditions for β. In general, there is no guarantee that the two equations of motion that result when these two expressions for X_E^2 are substituted into (A12) will be consistent with each other. Of course, the form of the equation of motion also depends on the form of the solutions for e^{μ_E} and e^{ν_E}. Clearly then, this model is rather tightly constrained. Indeed, the equation of motion of the thin shell could be viewed as being somewhat prescriptive of the nature of the exterior geometry. In other words, the intimate link between the thin shell and the exterior geometry suggests that the properties of the thin shell may provide information regarding properties of the exterior geometry. This in fact is the case. Equating the two expressions for X_E^2 given in (A8) and (2.10) one finds

$$e^{(\mu-\nu)_E}\dot{R}^2+e^{-\nu_E}=\frac{\beta'^2}{\beta_{,t}^2}\dot{R}^2. \tag{A13}$$

Substitution of $e^{(\mu-\nu)_E}\dot{R}^2$ from (A13) into the equation of motion (A12) leads to the result

$$e^{\nu_E} = -\frac{\beta_{,t}^2}{\beta'^2(1+\frac{1}{6}\lambda\beta_0^2R^2)}. \tag{A14}$$

Hence, the equation of motion, which applies only at $r=R$, leads to an algebraic expression for $e^{\nu_E}|_{r=R}$. One may view (A14) as a necessary condition for the projection of $e^{\nu_E}(r)$ onto Σ in order for the present model to be internally consistent. Adopting this viewpoint, the most obvious form for $e^{\nu_E}(r)$ to satisfy this condition is

$$e^{\nu_E} = -\frac{\beta_{,t}^2}{\beta'^2(1+\frac{1}{6}\lambda\beta^2r^2)}. \tag{A15}$$

For this choice, the Einstein equations (A1)–(A3) are satisfied when

$$e^{\mu_E} = -\frac{1+\frac{1}{6}\lambda\beta^2r^2}{\gamma^2\beta^2r^4}, \tag{A16}$$

where γ is an arbitrary constant with dimensions $(\text{length})^{-1}$. In the present model of a spherically symmetric region of Riemannian space in an exterior Weyl geometry, the equations of the dynamical model will be self-consistent provided the exterior metric is of the form given in (A15) and (A16) (the time-dependent solution presented in Ref. 11 satisfies this condition for $c_1 = -c_2^2$ and $q=0$. However, the oscillatory nature of the thin shell solution given in that paper is superseded by the results contained herein).

[1] H. Cheng, Phys. Rev. Lett. **61**, 2182 (1988).
[2] M. Pawłowski and M. Rączka, ''A Higgs-Free Model for Fundamental Interactions, Part I: Formulation of the Model,'' in *Modern Group Theoretical Methods in Physics*, edited by J. Bertrand *et al.* (Kluwer Academic, Dordrecht, 1995), p. 221; E. Tsantilis, R. A. Puntigam, and F. W. Hehl, ''A Quadratic Curvature Lagrangian of Pawłowski and Rączka: A Finger Exercise with MathTensor,'' in *Relativity and Scientific Computing—Computer Algebra, Numerics, Visualizations*, edited by F. W. Hehl, R. A. Puntigam, and H. Ruder (Springer, Berlin, 1996).
[3] G. Domokos, ''Broken Weyl Symmetry,'' DESY preprint 76/24 (1976).
[4] W. R. Wood and G. Papini, Phys. Rev. D **45**, 3617 (1992).
[5] H. Weyl, Sitzungsber. Preuss. Akad. Wiss., 465 (1918).
[6] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **333**, 403 (1973).
[7] D. Gregorash and G. Papini, Nuovo Cimento B **63**, 487 (1981); G. Papini, in *High-Energy Physics*, edited by S. L. Mintz and A. Perlmutter (Plenum, New York, 1985), p. 179.
[8] E. Santamato, Phys. Rev. D **29**, 216 (1984); J. T. Wheeler, Phys. Rev. D **41**, 431 (1990); C. Castro, Found. Phys. **22**, 569 (1992); M. Israelit and N. Rosen, Found. Phys. **26**, 585 (1996).
[9] W. R. Wood, N. Mobed, and G. Papini, Phys. Rev. D **48**, 5009 (1993).
[10] W. R. Wood and G. Papini, Found. Phys. Lett. **6**, 207 (1993); G. Papini and W. R. Wood, Phys. Lett. A **170**, 409 (1992); G. Papini and W. R. Wood, Phys. Lett. A **202**, 46 (1995).
[11] W. R. Wood, A. Feoli, and G. Papini, J. Math. Phys. **34**, 5916 (1993).
[12] J. L. Synge, *Relativity: The General Theory* (North–Holland, Amsterdam, 1960).
[13] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1967).
[14] H. Kodama, M. Sasaki, K. Sato, and K. Maeda, Prog. Theor. Phys. **65**, 1443 (1981); **66**, 2052 (1981); Phys. Lett. B **108**, 103 (1982).
[15] A. H. Guth and E. J. Weinberg, Nucl. Phys. B **212**, 321 (1983).
[16] A. Linde, astro-ph/9610077.
[17] S. K. Blau, E. I. Guendelman, and A. H. Guth, Phys. Rev. D **35**, 1747 (1987).

Stable singularities of wave-fronts in general relativity

Robert Low[a)]
Department of Mathematics, Coventry University, Coventry CV1 5FB, United Kingdom

(Received 23 December 1997; accepted for publication 23 February 1998)

A wave-front in a space-time $\mathcal{M}$ is a family of null geodesics orthogonal to a smooth spacelike two-surface in $\mathcal{M}$; it is of some interest to know how a wave-front can fail to be a smoothly immersed surface in $\mathcal{M}$. In this paper we see that the space of null geodesics $\mathcal{N}$ of $\mathcal{M}$, considered as a contact manifold, provides a natural setting for an efficient study of the stable singularities arising in the time evolution of wave-fronts. © *1998 American Institute of Physics.*
[S0022-2488(98)03506-3]

Consider a smooth spacelike two-surface in a space-time $\mathcal{M}$. Associated with this two-surfaces are two families of null geodesics, the ingoing and the outgoing null congruence. Congruences of this form are called wave-fronts, and it is of some interest to analyze the behavior of the evolution of these surfaces. In particular, we wish to know in what ways the surface can fail to be smoothly immersed, at least generically. For example, information gained from such an analysis is relevant to the consideration of images formed by gravitational lenses.

Analyses of the singularities of evolving wave-fronts in GR have already been presented in the literature.[1,2] This analysis differs from earlier work on the problem in that it automatically considers variations through wave-fronts, and requires no noncanonical choices, thereby facilitating a geometrical understanding of the situation. It also establishes a framework in which the application of Arnol'd's classification of stable singularities of Legendre mappings can be applied without the need for a great deal of intervening analysis.

We will begin by reviewing the differentiable and contact structures of the space of null geodesics, $\mathcal{N}$, for a strongly causal space-time, $\mathcal{M}$, and in particular for the case when $\mathcal{M}$ is globally hyperbolic. We then go on to see how to use the contact geometry to characterize wave-fronts and the stable singularities arising in their evolution, making heavy use of the analysis Arnol'd carried out for wave-fronts in flat space.[3]

So let $\mathcal{M}$ be a strongly causal space-time, and let $T^{*\prime}\mathcal{M}$ be the cotangent bundle of $\mathcal{M}$, minus the zero section. Take local coordinates (x^a,p_a) on $T^{*\prime}\mathcal{M}$. Then $T^{*\prime}\mathcal{M}$ has two natural forms defined on it, the canonical one-form θ given locally by $p_a dx^a$ and the symplectic form $\omega = d\theta = dp_a \wedge dx^a$.

In addition to these two differential forms, there are a couple of vector fields on $T^{*\prime}\mathcal{M}$ that we will need.

The first of these is the Euler field Δ, given by $\Delta = p_a(\partial/\partial p_a)$, which generates dilatations on each fiber of the cotangent bundle.

The second is the geodesic flow on the (reduced) cotangent bundle, defined by the Lorentz metric on $\mathcal{M}$ as follows: the Hamiltonian is defined by $H(x^a,p_a) = \frac{1}{2}g^{ab}p_a p_b = \frac{1}{2}p^a p_a$, and the geodesic spray is the vector field X_G on $T^{*\prime}\mathcal{M}$ defined by $X_G \lrcorner \omega + dH = 0$, or, in coordinates,

$$X_G = p^a \frac{\partial}{\partial x^a} - \Gamma_{abc} p^a p^b \frac{\partial}{\partial p_c},$$

where $\Gamma_{abc} = g_{ad}\Gamma^d{}_{bc}$ and $\Gamma^d{}_{bc}$ is the usual Christoffel symbol. We note also that $\mathcal{L}_\Delta X_G = X_G$.

Now, to construct $\mathcal{N}$, the space of null geodesics, we restrict our attention to $N'\mathcal{M}$, the bundle of (future pointing) null vectors, i.e., that part of $T^{*\prime}\mathcal{M}$ given by $p^a p_b = 0$ with p^a future pointing. Furthermore, since Δ and X_G are linearly independent and $[\Delta, X_G] = X_G$, the two-surfaces ele-

a)Electronic mail: r.low@coventry.ac.uk

 © 1998 American Institute of Physics

ments defined at each point by these two vector fields form an integral distribution. We could consider the space of null geodesics of $\mathcal{M}$ as the quotient manifold of leaves of this distribution.

It is sometimes more useful to consider taking the quotient by each vector field in turn. We could take the quotient by the Euler field, to obtain the space of unscaled null geodesics of $\mathcal{M}$ lifted to $N'\mathcal{M}$; since the Lie derivative of X_G along Δ is just X_G again, although X_G does not project to a vector field on this space, it still gives a one dimensional distribution on the resulting space. The integral curves of this distribution form a regular distribution[4] if $\mathcal{M}$ is strongly causal, and so the resulting space, $\mathcal{N}$, is a manifold. We could equally carry out these two identifications in the other order, with the same result.

It is also worth noting that under a conformal transformation of the metric, $N'\mathcal{M}$ is unchanged, and the restriction of the geodesic spray, X_G, to $N'\mathcal{M}$ is simply rescaled, so that all the geometry we are considering is conformally invariant.

As a notational convention, we will use lower-case Greek letters to represent points of $\mathcal{N}$ and the corresponding upper-case Greek letters to represent the corresponding subset of $\mathcal{M}$.

Lemma 1: If $\mathcal{M}$ is a globally hyperbolic space-time, and $\mathcal{S}$ is any smooth Cauchy surface for $\mathcal{M}$, then $\mathcal{N}$, the space of null geodesics of $\mathcal{M}$, is diffeomorphic to $T_1^*\mathcal{S}$, the cotangent sphere bundle of $\mathcal{S}$.

Proof: Give $\mathcal{S}$ coordinates, x_α, where α runs from 1 to 3, and let h be the Riemannian metric on $\mathcal{S}$ induced by g, the Lorentz metric on $\mathcal{M}$. Then a unit covector at x in $\mathcal{S}$ can be identified with a future pointing codirection at x in $\mathcal{M}$, by adding the unit future pointing timelike conormal to $\mathcal{S}$ at x, and then identifying this covector with all its positive multiples. Thus $T_1^*\mathcal{S}$ may be identified with the restriction to $\mathcal{S}$ of $N'\mathcal{M}$.

This gives us a bijection from $\mathcal{N}$ to $T_1^*\mathcal{S}$, by mapping any null geodesic γ in $\mathcal{N}$ to (x,v) where x is the point at which γ meets $\mathcal{S}$ and v is the unit covector in $T\mathcal{S}$ associated with the future pointing cotangent to Γ at x. It remains only to show that this bijection is a diffeomorphism. However, this now follows immediately from the fact that $T_1^*\mathcal{S}$ gives a slice of the foliation of $N'\mathcal{M}$ provided by the null geodesics, and thus defines the differentiable structure of $\mathcal{N}$.[4] □

Given two Cauchy surfaces, $\mathcal{S}_1$ and $\mathcal{S}_2$, the geodesic flow on $T^{*\prime}\mathcal{M}$ defines a function $f_g: T_1^*\mathcal{S}_1 \to T_1^*\mathcal{S}_2$ by mapping the point (x_1,v_1) to the point (x_2,v_2) where the null geodesic through $x_1 \in \mathcal{S}_1$ with a cotangent specified by v_1 meets $\mathcal{S}_2$ at x_2 with cotangent specified by v_2.

Corollary 1: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are any Cauchy surfaces of $\mathcal{M}$, then $f_g: T_1^*\mathcal{S}_1 \to T_1^*\mathcal{S}_2$ is a diffeomorphism. □

We can now see how the canonical form on $T^{*\prime}\mathcal{M}$ gives a contact structure on $\mathcal{N}$. To this end it is convenient to list some relationships between the Euler and geodesic vector fields and the canonical and symplectic forms,

$$\Delta \lrcorner \theta = 0, \quad \Delta \lrcorner \omega = \theta, \quad X_G \lrcorner \theta = 2H,$$

$$X_G \lrcorner \omega = -dH, \quad \mathcal{L}_\Delta \theta = \theta, \quad \mathcal{L}_\Delta \omega = \omega,$$

$$\mathcal{L}_{X_G} \theta = dH, \quad \mathcal{L}_{X_G} \omega = 0.$$

First, we restrict everything to $N'\mathcal{M}$. This is the surface $H=0$ in $T^{*\prime}\mathcal{M}$, and on this surface we observe that $X_G \lrcorner \theta$, $X_G \lrcorner \omega$, $\mathcal{L}_{X_Q}\theta$ and $\mathcal{L}_{X_G}\omega$ are all zero. It follows that if we simply take the quotient space of integral curves of the geodesic spray restricted to $N'\mathcal{M}$, θ will project to the resulting space of scaled null geodesics.

But we also want to quotient out by the action of the Euler field. θ cannot be projected to this quotient: however, it is homogeneous, and the distribution of planes that it defines does project. θ thus defines a 2-form on $\mathcal{N}$ only up to scale. Denote this object by $[\theta]$. The distribution of planes defined by the vanishing of $[\theta]$ gives a geometric structure on $\mathcal{N}$ which is well defined, even though θ itself need not be globally defined by this procedure. We now have to show that we do indeed have a contact structure.

So let S be a two-surface in $\mathcal{N}$. Then S defines Σ, a two-parameter family of null geodesics in $\mathcal{M}$, i.e., a three-surface $\tilde{S}$ in $\mathcal{M}$ which is ruled by null geodesics. This surface will, in general, have self-intersections and singularities, but will not, in general, be a wave-front.

Lemma 2: The tangent plane to S lies in the kernel of $[\theta]$ at each point of S iff the corresponding family of null rays Σ in $\mathcal{M}$ forms a wave-front, i.e., it comprises the outgoing (or ingoing) null congruence to some spacelike two-surface Σ.

Proof: The tangent place to S lies in the kernel of $[\theta]$ at each point iff θ vanishes on Σ, which is well known to be equivalent to Σ being orthogonal to any spacelike slice.[5] □

Lemma 3: The planes so defined on $\mathcal{N}$ are precisely the contact planes of $T_1^*\mathcal{S}$.

Proof: The contact form on $T_1^*\mathcal{S}$ is simply the restriction to $T_1^*\mathcal{S}$ of the canonical form on $T\mathcal{S}$. A Legendre manifold of $T_1^*\mathcal{S}$ is therefore a manifold L such that for any point $(q,v)\in L$, the projection map carries v to a vector orthogonal to the projection of L. But on identifying $T_1^*\mathcal{S}$ with N, we see that these are precisely the wave-fronts, and so $[\theta]$ defines a contact structure on $\mathcal{N}$. □

Corollary 2: The mapping $f_g: T_1^*\mathcal{S}_1 \to T_1^*\mathcal{S}_2$ is a contact mapping. □

Note that in the case where $\mathcal{M}$ is static, and $\mathcal{S}_1$ and $\mathcal{S}_2$ are both hypersurfaces orthogonal to the timelike Killing vector, this is essentially a statement of Huygens' principle; the above corollary is then a generalization of Huygens' principle to a nonstatic space-time.

Corollary 3: The Legendre sub-manifolds of $\mathcal{N}$ are precisely those families of null geodesics which form wave-fronts. □

A particularly important example of this is when $\tilde{S}$ consists of all those null geodesics which pass through some point, say p, of $\mathcal{M}$. In this case, one can show[6,7] that S is a smooth S^2 in $\mathcal{N}$, called the sky of p and denoted P. Thus a special case of what we will consider is the evolution of the light cone of a point.

We can now use this framework to provide the classification on the stable singularities of wave-fronts,[1,2] with a considerable saving of labor.

Theorem 1: Generically, wave-fronts will contain only singular points of types A_2 and A_3. At isolated instants, they will also contain transitions of type A_4 and D_4.

Proof: This now follows immediately from the classifications of Legendre mappings given in Ref. 3. □

Since the Legendre manifolds of $\mathcal{N}$ are precisely the lifts of wave-fronts, we immediately see that this resolution of singularities is obtained by perturbation through Legendre sub-manifolds of wave-fronts, so that singularities may be resolved by a small perturbation of the surface in which a wave-front intersects any Cauchy surface. This provides some saving of labor over the space-time approach, such as is provided in Ref. 2; the analysis presented therein may be related to this one by the observation that the conical Lagrange manifolds of $T^{*\prime}\mathcal{M}$ are precisely those which project to Legendre manifolds of $\mathcal{N}$.

Finally, we should note that although the analysis carried out above has been presented for the case where $\mathcal{M}$ is globally hyperbolic, this classification of stable singularities holds in rather greater generality. If, now, Γ is some generator of a wave-front, we can consider the intersection of some neighborhood of Γ in the wave-front with $\mathcal{S}$, a small portion of the spacelike surface through which Γ passes. We continue to require that $\mathcal{M}$ be strongly causal; first, so that $\mathcal{N}$ will retain its structure as a contact manifold, and second, so that we can guarantee that by taking $\mathcal{S}$ sufficiently small, no null geodesic in our neighborhood of Γ will intersect $\mathcal{S}$ more than once. Then by the same argument as before, wave-fronts are precisely the Legendre submanifolds of $\mathcal{N}$. Furthermore, if we consider the projection of a neighborhood of Γ to $\mathcal{S}$, then the classification of stable singularities will be exactly as before.

ACKNOWLEDGMENTS

It is a pleasure to thank the Erwin Schrödinger International Institute for Mathematical Physics, Vienna, for hospitality and facilities, and in particular for providing such a congenial environment, while some of this work was carried out.

[1] H. Friedrich and J. M. Stewart, ''Characteristic initial data and wavefront singularities in general relativity,'' Proc. R. Soc. London, Ser. A **385**, 345–371 (1983).
[2] W. Hasse, M. Kriele, and V. Perlick, ''Caustics of wavefronts in general relativity,'' Class. Quantum Grav. **13**, 1161–1182 (1996).
[3] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, Berlin, 1978).

[4]F. Brickell and R. S. Clarke, *Differentiable Manifolds* (Van Nostrand Reinhold, London, 1970).
[5]R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge University Press, Cambridge, 1986), Vol. 2.
[6]R. J. Low, ''The geometry of the space of null geodesics,'' J. Math. Phys. **30**, 809–811 (1989).
[7]R. J. Low, ''Twistor linking and causal relations,'' Class. Quantum Grav. **7**, 177–187 (1990).

Static perfect fluid cylinders

Ulf S. Nilsson
Department of Physics, Stockholm University, S-113 85 Stockholm, Sweden

Claes Uggla
Department of Physics, Luleå University of Technology, S-951 87 Luleå, Sweden

Mattias Marklund
Department of Plasma Physics, Umeå University, S-901 87 Umeå, Sweden

(Received 3 November 1997; accepted for publication 29 January 1998)

The field equations for cylindrically symmetric perfect fluid models with a non-scale invariant equation of state where the energy density is linearly related to the pressure are rewritten as a set of first order coupled ordinary differential equations. Variables are chosen such that the resulting phase space is compact and everywhere regular. The dynamical systems approach, used in the past mainly for scale invariant matter sources, is subsequently used to gain qualitative information about the space of solutions. It leads to a simple way of demonstrating the existence and properties of interior sources for cylindrically symmetric vacuum solutions. The models are shown to exhibit asymptotic scale invariance. © *1998 American Institute of Physics.* [S0022-2488(98)00605-7]

I. INTRODUCTION

There has been a continuing interest in cylindrically symmetric models since the first vacuum solution was found by Weyl and Levi-Civita in 1917–1918. This is despite the fact that it is hard to find any natural phenomena which could give rise to this particular geometry. Instead, cylindrical symmetry can be used as a probe to examine what features survive when generalizing spherical symmetry to axial symmetry as the infinite cylinder is the simplest axially symmetric system possible. Thin shells can be used as sources, which was shown by Wang *et al.*[1] Philbin[2] recently found an exact perfect fluid source but the equation of state is somewhat complicated. Stela and Kramer[3] have studied the case of static models with an incompressible fluid, i.e., constant energy density, numerically using the formulation of Kramer.[4]

In this paper we will use the dynamical systems approach which has been used with great success for over twenty years now in spatially homogeneous cosmology where considerable qualitative information has been gained, see, e.g., Ref. 5 and references therein. Usually one has considered scale invariant matter sources. For a perfect fluid this implies an equation of state where the pressure is directly proportional to the energy-density. In this paper we will give an example of a source which is not scale invariant and show that the dynamical system can be used to study such sources as well.

The matter source is assumed to be a perfect fluid with a stress-energy tensor of the form

$$T_{ab}=\mu u_a u_b+p(g_{ab}+u_a u_b), \tag{1}$$

where μ, p and u^a are the energy density, the pressure and the 4-velocity of the fluid, respectively. The equation of state is assumed to be of the form $\mu=\mu_0+(\eta-1)p$, where η and μ_0 are constants. The incompressible fluid case is given by $\eta=1$ and $\mu_0>0$. The case $\eta>1$ and $\mu_0 \geqslant 0$ yields a ''generalized barotropic'' equation of state where the genuine barotropic case is found by setting $\mu_0=0$. The equation of state for this case is often written in the form $p=(\gamma-1)\mu$, resulting in the relationship $\eta=\gamma/(\gamma-1)$. The physical range of γ is $1<\gamma\leqslant 2$, which corresponds to $2\leqslant\eta<\infty$ ($\gamma=1\Leftrightarrow\eta=\infty$). The choice of equation of state includes the explicitly solvable case of Evans,[6] $\eta=6$ ($\gamma=\frac{6}{5}$), which was also studied by Stela and Kramer.[3]

The static cylindrically symmetric models are a subclass of the more general static Bianchi type I models. The general line element of these models is

 © 1998 American Institute of Physics

$$ds^2 = -C_1(\rho)^2 dt^2 + d\rho^2 + C_2(\rho)^2 d\varphi + C_3(\rho)^2 dz^2, \tag{2}$$

with the 4-velocity of the fluid given by $u^a = C_1^{-1}(1,0,0,0)$.

For a solution to be cylindrically symmetric it has to have flat geometry in the vicinity of the central axis (pages 192 and 217 in Ref. 7). In the present case this implies that the above metric coefficients satisfy the following conditions:

$$\lim_{\rho\to 0} C_{1,3} = \lim_{\rho\to 0} \rho^{-1} C_2 = 1. \tag{3}$$

Thus one might expect that in order to distinguish a particular solution as cylindrically symmetric or not, the metric functions must be known explicitly in order to check the above limits. However, in Ref. 8 it was shown that the dynamical systems approach to the field equations yields a very simple way of implementing the above constraint. In the present paper we will study a relatively simple class of models. We thus hope that the presentation will not be clouded by too many technical details. We will return to possible generalizations in the discussion at the end of the paper.

The outline of the paper is as follows. In section II we write the field equations so that the phase space of the equations becomes compact and everywhere regular. In section III we perform a local analysis of the dynamical system. In section IV the different invariant boundary submanifolds of the phase space are discussed and the corresponding phase portraits are given. In section V we show how an interior solution is matched with the vacuum solution while in section VI we examine the cylindrically symmetric submanifold explicitly. In section VII we focus on the incompressible fluid while the more general models are investigated in section VIII. We end with a discussion in section IX. In Appendix A we discuss the Levi-Civita vacuum models while Appendix B gives the relation to the variables used by Kramer[4] and Stela and Kramer.[3]

II. THE DYNAMICAL SYSTEM

We will now rewrite the field equations in an autonomous form with variables adapted to the normal congruence associated with the symmetry surfaces $\rho = \text{const}$. Note that this congruence is spacelike. Misner has introduced a convenient parametrization of the metric coefficients in the context of spatially homogeneous cosmology.[9] Here we use an analogous parametrization. The metric functions C_1, C_2 and C_3 can be written as

$$C_1 = e^{\beta^0 - 2\beta^+}, \quad C_2 = e^{\beta^0 + \beta^+ + \sqrt{3}\beta^-}, \quad C_3 = e^{\beta^0 + \beta^+ - \sqrt{3}\beta^-}. \tag{4}$$

The variables β^0, $\beta^\pm$ are related to the kinematics of the normal congruence of the symmetry surfaces,

$$\theta = 3\dot{\beta}^0, \quad \sigma_+ = 3\dot{\beta}^+, \quad \sigma_- = 3\dot{\beta}^-, \tag{5}$$

where θ is the expansion and $\sigma_\pm$ describes the shear. A dot indicates $d/d\rho$. The field equations $G_{ab} = T_{ab}$ can now be written as the following:

Evolution equations:

$$\dot{\theta} = -\theta^2 - \frac{1}{2}(\mu - 5p), \tag{6}$$

$$\dot{\sigma}_+ = -\theta\sigma_+ - (\mu + p), \tag{7}$$

$$\dot{\sigma}_- = -\theta\sigma_-. \tag{8}$$

Defining equation for p:

$$p = \tfrac{1}{3}\theta^2 - \tfrac{1}{3}\sigma_+^2 - \tfrac{1}{3}\sigma_-^2. \tag{9}$$

The spacetime is fully described by the variables θ and $\sigma_{\pm}$, which can be understood by either studying the automorphism group[10] or using the equivalence problem approach.[11,12] However, the variables β^0 and $\beta^{\pm}$, and thus the metric variables C_1, C_2 and C_3, can be found in terms of quadratures using the decoupled equations in Eq. (5) when θ and $\sigma_{\pm}$ are known.

The equations of motion for the fluid, $\nabla_a T^{ab}=0$, yield an evolution equation for the pressure p, which also can be found by differentiating Eq. (9),

$$\dot{p}=-\tfrac{1}{3}(\theta-2\sigma_+)(\mu+p). \tag{10}$$

For the purpose of this paper it is more useful to examine instead the evolution equation for the variable $w=\mu+p$. Without specifying an equation of state, we obtain

$$\dot{w}=(v_s^{-2}+1)\dot{p}=-\tfrac{1}{3}(\theta-2\sigma_+)(v_s^{-2}+1)w, \tag{11}$$

where $v_s^2=dp/d\mu$ is the square of the sound velocity in the fluid. This velocity is assumed to be a real quantity. We now make the assumption that $w=\mu+p\geqslant 0$. This is just the weak energy condition[13] and w can be interpreted as a ''weak energy variable.'' Thus, as $w=0$ is an invariant submanifold of the system of differential equations, it is obvious that w is an important dynamical quantity. For the models considered in this paper, the evolution equation for w becomes

$$\dot{w}=-\tfrac{1}{3}\eta(\theta-2\sigma_+)w. \tag{12}$$

Note that the above equations also can be found by specializing the equations of Ref. 14 to the static Bianchi type I case. We can now write Eq. (9) as

$$\frac{3w}{\eta}=\theta^2+\frac{3\mu_0}{\eta}-\sigma_+^2-\sigma_-^2. \tag{13}$$

As w is assumed positive, we see that the $D=\sqrt{\theta^2+3\mu_0/\eta}$ is a ''dominant quantity.'' By introducing new bounded variables according to

$$Q=\frac{\theta}{D},\quad \Sigma_+=\frac{\sigma_+}{D},\quad \Sigma_-=\frac{\sigma_-}{D},\quad W=\frac{3w}{\eta D^2}, \tag{14}$$

along with a new independent variable ξ,

$$3d\xi=Dd\rho, \tag{15}$$

the system of differential equations can be written as the following.

Evolution equations:

$$Q'=-\tfrac{1}{2}(1-Q^2)(6+(\eta-6)W), \tag{16}$$

$$\Sigma_+'=-\tfrac{1}{2}(2\eta-(\eta-6)Q\Sigma_+)W, \tag{17}$$

$$\Sigma_-'=\tfrac{1}{2}(\eta-6)Q\Sigma_-W. \tag{18}$$

Defining equation for W:

$$W=1-\Sigma_+^2-\Sigma_-^2, \tag{19}$$

where a prime denotes $d/d\xi$. The μ_0-terms have been eliminated using

$$\frac{3\mu_0}{\eta D^2}=1-Q^2. \tag{20}$$

The above system of differential equations is invariant under the discrete symmetry

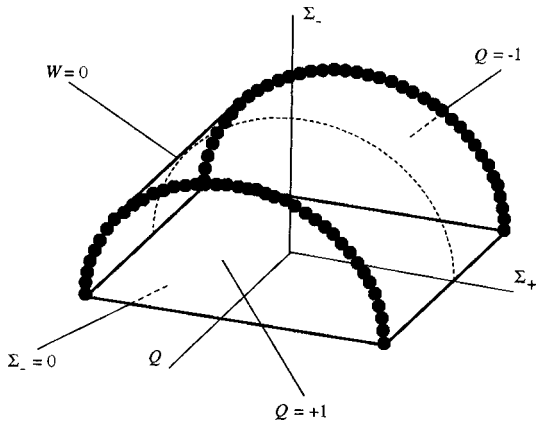

FIG. 1. The structure of the phase space with the different boundaries indicated.

$$(Q,\Sigma_+,\Sigma_-,\eta)\rightarrow -(Q,\Sigma_+,\Sigma_-,\eta). \tag{21}$$

This transformation can be achieved by choosing $D=-\sqrt{\theta^2+3\mu_0/\eta}$ instead of $D=\sqrt{\theta^2+3\mu_0/\eta}$. The line element is invariant under this discrete symmetry and solutions are thus represented by two orbits in the phase space related by the above discrete symmetry. The system is also invariant under the discrete symmetry

$$(Q,\Sigma_+,\Sigma_-)\rightarrow(Q,\Sigma_+,-\Sigma_-), \tag{22}$$

which corresponds to the discrete isometry $z\leftrightarrow\varphi$ in the line element [see Eq. (2)]. Knowing the structure of the phase space for $\Sigma_-\geqslant 0$ means that by using the above discrete symmetry, the structure of the phase space where $\Sigma_-\leqslant 0$ is easily found. For convenience and without loss of generality we will choose $\Sigma_-\geqslant 0$. The boundary of the reduced phase space is given by a number of invariant submanifolds. They are (i) $W=1-\Sigma_+^2-\Sigma_-^2=0$, (ii) $\Sigma_-=0$ and (iii) $Q=\pm 1$.

From Eq. (16), it follows that Q is a decreasing monotone function in phase space except on the boundaries $Q=\pm 1$. This prevents the existence of periodic orbits, recurrent orbits and homoclinic orbits as well as equilibrium points anywhere in phase space except on the $Q=\pm 1$ boundaries.[5] Therefore the only equilibrium points of the system of differential equations must lie on the boundary submanifolds $Q=\pm 1$. The structure of the phase space is given in Fig. 1 where we also indicate the different boundary submanifolds.

III. A LOCAL ANALYSIS OF THE DYNAMICAL SYSTEM

Here we give the equilibrium points of the system together with their eigenvalues. The value of W is also given as it shows if the point is on the boundary $W=0$ or not. As stated earlier, all equilibrium points are located on the $Q=\pm 1$ boundaries and thus have two eigenvectors which are associated with directions within the $Q=\pm 1$ submanifold while the third direction is directed away from this submanifold. We refrain from giving the explicit eigenvectors but the eigenvalues are ordered so that the first two are associated with directions within the $Q=\pm 1$ submanifolds while the third eigenvalue corresponds to a direction pointing out of these submanifolds.

The equilibrium points $K^{\pm}$:

$$Q=\pm 1,\quad \Sigma_+^2+\Sigma_-^2=1,\quad W=0; \tag{23}$$

$$0,2\eta\Sigma_+\pm(6-\eta),\pm 6. \tag{24}$$

The zero eigenvalue indicates that this is a line of equilibrium points. The eigenvector direction which points out of the $Q=\pm 1$ boundaries lies in the $W=0$ submanifold. There is one special point, denoted T, for which $Q=1$, $\Sigma_+=\frac{1}{2}$ and $\Sigma_-=\sqrt{3}/2$. It corresponds to flat space written on the so called Taub form, $ds^2=-dt^2+d\rho^2+\rho^2 d\varphi^2+dz^2$, which is explicitly cylindrically symmetric. It corresponds to the choice $m=0$ in the Levi-Civita line element, see Appendix A. Orbits

which start from this point will therefore correspond to solutions which are cylindrically symmetric. By only studying orbits starting from this point we have thus ''automatically'' satisfied the constraint equation (3), (see Ref. 8 for a more detailed discussion). There is of course a corresponding point on the $Q=-1$ boundary, but using the discrete symmetry [Eq. (21)] we only need to study the $Q=+1$ case. Being of such importance, we will give the eigenvalues and eigenvectors for T explicitly. They are

$$0, \quad (0,-\sqrt{3},1); \quad 6, \quad (0,-\sqrt{3}(2+\eta),\eta-6); \quad 6, \quad (1,0,0). \tag{25}$$

The equilibrium points $B^{\pm}$:

$$Q=\pm 1, \quad \Sigma_+=\pm\frac{2\eta}{\eta-6}, \quad \Sigma_-=0, \quad W=\frac{3(2-\eta)(6+\eta)}{(\eta-6)^2}; \tag{26}$$

$$\pm\frac{3(2-\eta)(\eta+6)}{2(\eta-6)}, \quad \pm\frac{3(2-\eta)(\eta+6)}{2(\eta-6)}, \quad \mp\frac{3(2+\eta)}{\eta-6}. \tag{27}$$

The point $B^+(B^-)$ exists only for $1\leqslant\eta<2$ ($\gamma>2$). At $\eta=2$ ($\gamma=2$) the point $B^+(B^-)$ coincides with a point on the line $K^+(K^-)$. The eigenvector directions which point out of the $Q=\pm 1$ boundary submanifolds lie in the $\Sigma_-=0$ submanifold.

IV. BOUNDARY SUBMANIFOLDS

When studying the orbits in the interior of phase space it is always advantageous to know the dynamics of the boundary. The different invariant boundary submanifolds were given in section II but are discussed here in more detail.

A. The scale invariant submanifolds

The invariant submanifolds $Q=\pm 1$ correspond to $\mu_0/D^2=0$, implying that μ_0 is dynamically unimportant. For $\eta\neq 1$ the ''effective'' equation of state is $\mu=(\eta-1)p$, i.e., $p=(\gamma-1)\mu$ which is *scale invariant*. Such models, with $1<\gamma<2$, were studied in Ref. 8. For the incompressible fluid, $\eta=1$, one has a perfect fluid with vanishing energy-density but non-zero pressure. Although being physically questionable, this type of source is also scale invariant. There is a bifurcation at $\eta=\gamma=2$ when the points $B^{\pm}$ leave the phase space. The general solution for the $Q=\pm 1$ submanifolds is

$$\Sigma_+=\frac{2\eta Q}{\eta-6}(1-A_0\Sigma_-), \quad \eta\neq 6, \tag{28}$$

where A_0 is a constant. When $\eta=6$, the general solution is instead $\Sigma_-=\text{const}$. The phase portraits for the $Q=1$ submanifold is given in Fig. 2(a)–2(d). The corresponding phase portraits for the $Q=-1$ submanifold can be found by using the discrete symmetries Eq. (21) and Eq. (22).

B. The ''weak-energy condition'' submanifold

The dynamics on the $W=0$ boundary is independent of the parameter η and the solution is

$$\Sigma_+=s_+, \quad \Sigma_-=s_-, \tag{29}$$

where s_+ and s_- are constants satisfying $s_+^2+s_-^2=1$. The phase portrait is given in Fig. 2(e).

C. The LRS submanifold

The submanifold $\Sigma_-=0$ corresponds to locally rotationally symmetric (LRS) solutions of Petrov type D (see e.g., Ref. 7). For these models the metric variables C_2 and C_3 in Eq. (2) can be chosen to be equal. There is a bifurcation at $\eta=\gamma=2$ resulting from $B_{\pm}$ leaving the phase space. The phase portraits are given in Fig. 2(f) and 2(g).

When $\mu_0\neq 0$ the source of the field equations is no longer scale invariant. Therefore the equations only allow self-similar solutions when $\mu_0=0$. The situation is similar to that of cos-

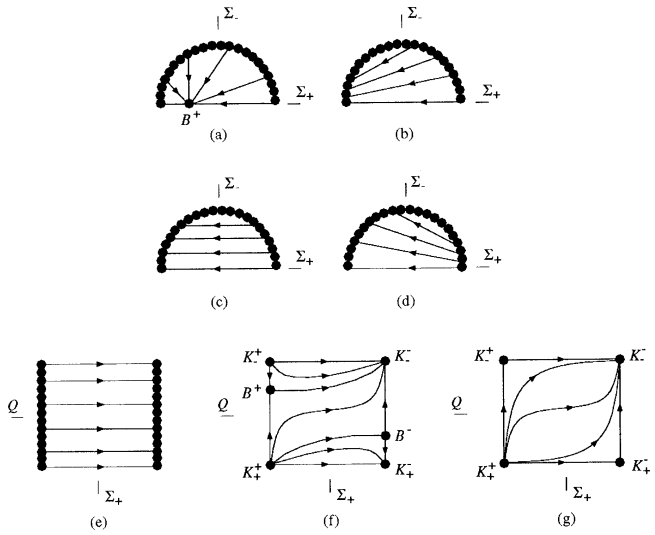

FIG. 2. The structure of the various boundary submanifolds for different values of the parameter η when (a) $Q=1$ and $1\leqslant\eta<2$, (b) $Q=1$ and $2\leqslant\eta<6$, (c) $Q=1$ and $\eta=6$, (d) $Q=1$ and $\eta>6$, (e) $W=0$, (f) $\Sigma_-=0$ and $1\leqslant\eta<2$ and (g) $\Sigma_-=0$ and $\eta\geqslant 2$. The dynamics on the boundary $Q=-1$ can be found by applying the discrete symmetries Eq. (21) and Eq. (22).

mology where a non-zero cosmological constant forbids the existence of self-similar solutions. As there exists a monotone function everywhere in phase space except on the $Q=\pm 1$ submanifolds there can be no complicated attractors inside phase space. Therefore all orbits end at equilibrium points which are characterized by a scale invariant matter source yielding an asymptotic scale invariance for the models. Also, as long as orbits in phase space are ''shadowing'' the $Q=\pm 1$ submanifolds, they will behave effectively as having a scale invariant perfect fluid source.

V. MATCHING WITH THE VACUUM SOLUTION

A source to the vacuum solution is characterized by a positive pressure p which vanishes on a surface at a finite distance from the center. On this surface the interior solution is matched with the vacuum solution. This is analogous to a spherically symmetric star where the interior solution is matched with the Schwarzschild solution. The continuity of the metric and its first derivatives is equivalent to the continuity of the variables θ, σ_+ and σ_- on the boundary. From Eq. (9) it follows that $p=0$ implies

$$Q^2-\Sigma_+^2-\Sigma_-^2=0. \tag{30}$$

This is a surface in the reduced phase space which is not an invariant submanifold, as is easily seen by differentiating the above expression. It divides the phase space into disjoint parts characterized by $p>0$ and $p<0$, respectively (see Fig. 3). Physical solutions are represented by orbits starting from an equilibrium point in a direction which corresponds to positive pressure. Cylindrically symmetric models start at the equilibrium point T. Instead of continuing the orbits through the surface $p=0$, they are stopped at the $p=0$ surface and the explicit values of the phase space variables there serve as ''initial'' data for the vacuum equations.

The matching conditions on the surface $p=0$ implies that we have to match the interior solution with a vacuum solution described by

$$\sigma_+=\frac{\Sigma_+^{(\mathrm{B})}}{Q^{(\mathrm{B})}}\,\theta, \quad \sigma_-=\theta\sqrt{1-\left(\frac{\Sigma_+^{(\mathrm{B})}}{Q^{(\mathrm{B})}}\right)^2}, \tag{31}$$

where $\Sigma_+^{(\mathrm{B})}$ and $Q^{(\mathrm{B})}$ are the values of the solutions in the limit $p\to 0$. The ''mass-parameter'' m of the Levi-Civita line element (see Appendix A) is related explicitly to the boundary values $\Sigma_\pm^{(\mathrm{B})}$ and $Q^{(\mathrm{B})}$ by

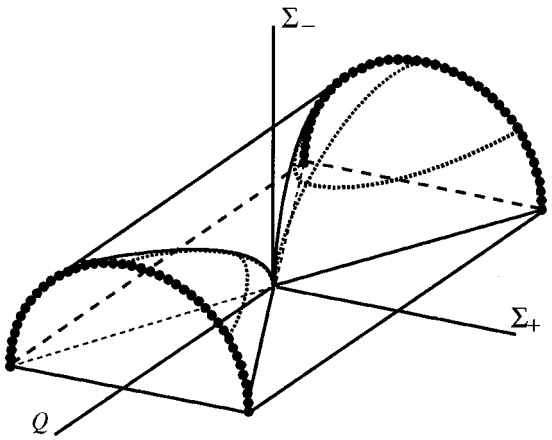

FIG. 3. The surfaces $p=0$. The scale of Q has been exaggerated for visual reasons. The region beneath the paraboloid is characterized by $p>0$.

$$m=\frac{2Q^{(\mathrm{B})}-\Sigma_{+}^{(\mathrm{B})}-\sqrt{3}\Sigma_{-}^{(\mathrm{B})}}{Q^{(\mathrm{B})}-2\Sigma_{+}^{(\mathrm{B})}}, \tag{32}$$

showing that m is manifestly constant. As orbits corresponding to physical sources are stopped at some finite ξ_0 and not at an equilibrium point of the system, it follows from the above expression and the differential equation for Q that all cylinders have a finite size. The proper radius R of the cylinder can be found by integrating Eq. (15),

$$R=3\int \frac{d\xi}{D}=\sqrt{\frac{3\eta}{\mu_0}}\int_{-\infty}^{\xi_0}\sqrt{1-Q^2}\,d\xi. \tag{33}$$

VI. THE CYLINDRICALLY SYMMETRIC SUBMANIFOLD

Within the general class of static Bianchi type I models there exists a particularly interesting subclass of solutions, namely the cylindrically symmetric solutions. As discussed previously, all the cylindrically symmetric solutions correspond to orbits starting from the point T. As this point is a source; there will always be a one-parameter set of orbits entering the interior phase space. Using the eigenvectors given in Eq. (25), we can write the linearized solution near the point T as

$$Q=1-\epsilon_Q e^{6\xi}, \tag{34}$$

$$\Sigma_{+}=\tfrac{1}{2}-\sqrt{3}\,\epsilon_\Sigma(2+\eta)e^{6\xi}, \tag{35}$$

$$\Sigma_{-}=\frac{\sqrt{3}}{2}+\epsilon_\Sigma(\eta-6)e^{6\xi}, \tag{36}$$

where $\epsilon_Q,\epsilon_\Sigma$ are small positive parameters. A translation of the parameter ξ just leads to a rescaling of the parameters $\epsilon_Q,\epsilon_\Sigma$. The quotient p/μ at the center $\rho=0$ of the cylinder is given by

$$\left.\frac{p}{\mu}\right|_{\xi\to-\infty}=\frac{4\sqrt{3}(\epsilon_\Sigma/\epsilon_Q)-1}{\eta+(\eta-1)[4\sqrt{3}(\epsilon_\Sigma/\epsilon_Q)-1]}, \tag{37}$$

which implies that orbits corresponding to solutions with positive pressure have $\epsilon_\Sigma/\epsilon_Q>1/4\sqrt{3}$ (assuming $\eta\geqslant 1$). Orbits with $\epsilon_\Sigma/\epsilon_Q<1/4\sqrt{3}$ have negative pressure while if $\epsilon_\Sigma/\epsilon_Q=1/4\sqrt{3}$ the orbits start out exactly in the $p=0$ surface. Enforcing the energy condition $p/\mu\leqslant 1$ at the center of the cylinder implies that for $1\leqslant\eta<2$ we must also have $\epsilon_\Sigma/\epsilon_Q\leqslant 1/(2\sqrt{3}[2-\eta])$. For $\eta\geqslant 2$

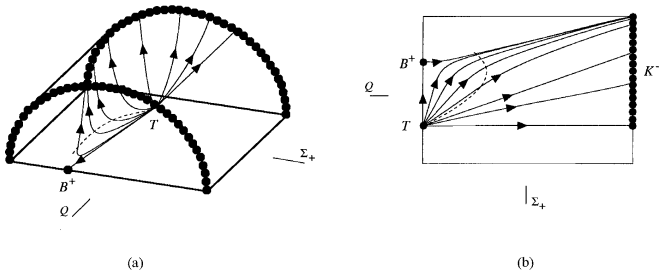

FIG. 4. The cylindrically symmetric submanifold for the case $\eta=1$, the incompressible fluid, as (a) a surface of solutions in the full phase space and (b) the surface of solutions projected onto $\Sigma_-=0$. The dashed line indicates where orbits pass through the $p=0$ surface.

this condition puts no extra limits on the choice of $\epsilon_\Sigma/\epsilon_Q$. If the direction of an orbit is close to the boundary $Q=1$, i.e., $\epsilon_Q \ll \epsilon_\Sigma$, it follows from Eq. (37) above that (excluding the case $\eta=1$)

$$\left.\frac{p}{\mu}\right|_{\xi\to-\infty} \approx \frac{1}{\eta-1}, \tag{38}$$

which is an aspect of the previously discussed asymptotic scale invariance of the source.

It follows that models with an initial positive pressure will reach a point where $p=0$ at a finite value of the independent variable. At this point they can then be matched with a cylindrically symmetric vacuum solutions as described in the previous section.

VII. THE INCOMPRESSIBLE FLUID

The incompressible fluid is characterized by a constant energy density $\mu=\mu_0$ which is equivalent to setting $\eta=1$. The cylindrically symmetric submanifold is a 2-dimensional surface in phase space spanned by the two eigenvectors corresponding to the non-zero eigenvalues of the point T. The structure of this submanifold is given in Fig. 4(a). The assumption of positivity of the pressure near the center of the cylinder yields the restriction $\epsilon_\Sigma/\epsilon_Q > 1/4\sqrt{3}$. The conditions $p/\mu \leqslant 1$ and $p>0$ at the center of the cylinder (i.e., the conditions Stela and Kramer[3] use), lead to $1/4\sqrt{3} < \epsilon_\Sigma/\epsilon_Q \leqslant 1/2\sqrt{3}$. This means that we cannot choose initial data such that the resulting orbit shadows the scale invariant boundary $Q=1$ (with the exception of the point T). This will be true for all models with $1 \leqslant \eta < 2$. In Fig. 4(b) the surface of solutions is projected onto the $\Sigma_-=0$ plane. The curve where the cylindrically symmetric submanifold passes through the surface $p=0$ is indicated in Fig. 4 by the dashed line.

VIII. GENERAL FEATURES

If we allow η to vary, it is clear that models with $1<\eta<2$ have the same qualitative behavior as the incompressible fluid, e.g., the point B^+ describes the intermediate state of the model for many of the orbits. As the point B^+ leaves the phase space at $\eta=2$, this intermediate behavior is instead described by a point on the line of equilibrium points K^+ whose values of the phase space variables is a function of η. As an example of this generic behavior for models with $\eta \geqslant 2$, the cylindrically symmetric submanifold for the case $\eta=4$ ($\gamma=\frac{4}{3}$) is shown in Fig. 5. For $\eta \geqslant 2$ we can always choose initial data so that the resulting solution ''shadows'' the scale invariant boundary $Q=1$, independently of the specific value of η.

We remark finally on models with $\eta=6$, which implies an equation of state of the form $\mu=\mu_0+5p$. In this case the equations simplify and can be solved exactly. The solution is characterized by the integrals

$$\frac{(1+Q)(1-2\Sigma_+)}{(1-Q)(1+2\Sigma_+)}=\text{const.}, \quad \Sigma_-=\text{const.} \tag{39}$$

All the dynamics of the cylindrically symmetric models therefore will take place in the plane $\Sigma_-=\sqrt{3}/2$. This feature, that the cylindrically symmetric submanifold coincides with a known

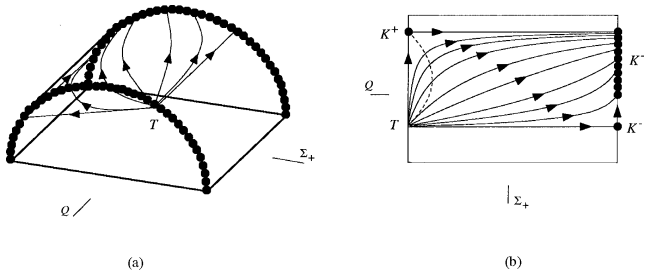

FIG. 5. The cylindrically symmetric submanifold for the case $\eta=4$, the incompressible fluid, as (a) a surface of solutions in the full phase space and (b) the surface of solutions projected onto $\Sigma_-=0$. The dashed line in (b) indicates where orbits pass through the $p=0$ surface. Different points on the line of equilibrium points K^- are denoted by simply K^- as their exact value is of lesser interest.

submanifold of the system of differential equations, exists only for the case $\eta=6$. The phase portrait is given in Fig. 6. The corresponding exact solution was found by Evans.[6]

A numerical study shows that the mass parameter m of the Levi-Civita line element takes values in the range $0<m<1$ although the exact values depend on the choice of η.

IX. DISCUSSION

In this paper we have shown that by an appropriate choice of variables the phase space of the field equations for static Bianchi type I perfect fluid models with an equation of state $\mu=\mu_0+(\eta-1)p$ can be made compact and everywhere regular. This together with the existence of a monotone function indicate that the global structure of the phase space can be well understood. We have also shown that these models are asymptotically scale invariant.

Even though the source is not scale invariant the dynamical systems approach has turned out to be a very useful tool. It constitutes a very simple way of implementing the otherwise rather cumbersome cylindrically symmetric constraint on the more general static Bianchi type I models and thus finding the cylindrically symmetric submanifold. All orbits describing cylindrically symmetric solutions start from a single point in phase space and can be parametrized by the quotient $\epsilon_\Sigma/\epsilon_Q$, where ϵ_Σ and ϵ_Q are the parameters of the local linearized solution of the field equations near the center of the cylinder. By following only orbits corresponding to positive pressure near the center of the cylinder, the physically reasonable interior solutions are easily singled out. These orbits are then stopped as the pressure of the fluid vanishes and it is shown that the corresponding solution with $\mu_0\neq 0$ can be matched at a finite radius with a Levi-Civita vacuum solution. This is in stark contrast with the case $\mu_0=0$ see Ref. 8, where such matching is impossible.

The present approach can easily be applied to static spherically symmetric models with the same equation of state. As this class of models have been studied extensively in the literature over the years it is unlikely that the dynamical systems approach would yield any new results. However, there are some other natural generalizations of the models in this paper. If the fluid is allowed to rotate without shear, i.e., making the models stationary rigidly rotating, the generali-

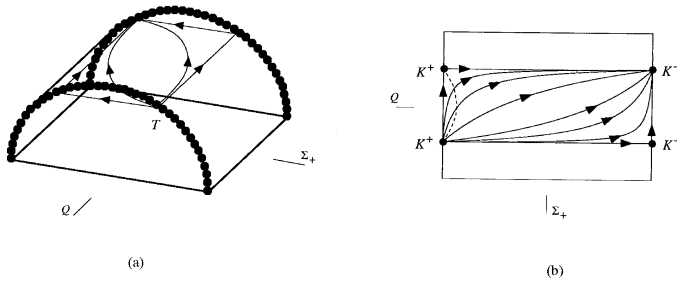

FIG. 6. The cylindrically symmetric submanifold for the case $\eta=6$ as (a) a surface of solutions in the full phase space and (b) the surface of solutions projected onto $\Sigma_-=0$. In (b) the dashed line indicates where the orbits pass the surface $p=0$.

zation is straightforward. The addition of rotation adds one dimension to the phase space, making the visualization somewhat more complicated, but the approach to proving the existence of interior solutions is still valid. A further generalization is to let the rotation be non-rigid. This is equivalent to a non-vanishing fluid shear. These models will contain an arbitrary function, e.g. the fluid shear, making the system of differential equations undetermined. Therefore they are not well-suited for the dynamical systems approach. One can of course define the functional dependence of the fluid shear, thus making the system determined again and then study that particular class of models. Another way of generalizing the models is by introducing a more general equation of state, $p=p(\mu)$. For these models the choice of variables in this paper will not be appropriate as they will not result in a compact phase space. But by making a slightly different choice and imposing the weak energy condition on the fluid, a compact phase space can be found. The question of existence of interior solutions with, e.g., a polytropic perfect fluid, can then be studied.

APPENDIX A: THE LEVI-CIVITA SOLUTION

The general static cylindrically symmetric solution is the Levi-Civita solution,[7]

$$ds^2=-r^{2m}dt^2+r^{-2m}(r^{2m^2}(dr^2+dz^2)+r^2d\varphi^2), \tag{40}$$

where m is a constant of referred to as the ''mass parameter.'' Introducing a new radial variable, $\rho=\rho(r)$, the above metric can be written in the form of (2),

$$ds^2=-\rho^{2m/A}dt^2+d\rho^2+\rho^{2(1-m)/A}d\varphi^2+\rho^{2m(m-1)/A}dz^2, \tag{41}$$

where $A=m(m-1)+1$. The kinematical quantities are

$$\theta=1/\rho, \tag{42}$$

$$\sigma_+=\frac{m^2-4m+1}{2A\rho}, \tag{43}$$

$$\sigma_-=\frac{\sqrt{3}(1-m^2)}{2A\rho}. \tag{44}$$

APPENDIX B: COMPARISON WITH KRAMER

In Ref. 7 Kramer introduced the variables y and z which are used for the numerical calculation in Ref. 3. They are defined in terms of the variables used in this paper according to

$$y=\frac{2\theta-\sigma_+-\sqrt{3}\sigma_-}{\theta-2\sigma_+}=\frac{2Q-\Sigma_+-\sqrt{3}\Sigma_-}{Q-2\Sigma_+}, \tag{45}$$

$$z=\frac{2\theta-\sigma_++\sqrt{3}\sigma_-}{\theta-2\sigma_+}=\frac{2Q-\Sigma_++\sqrt{3}\Sigma_-}{Q-2\Sigma_+}. \tag{46}$$

We easily see that the variable z becomes infinite at the point T, i.e., at the center of the cylinder, while the linear solution, Eqs. (34)–(36), implies that

$$y=\frac{2-8\sqrt{3}(\epsilon_\Sigma/\epsilon_Q)}{1-2\sqrt{3}(2+\eta)(\epsilon_\Sigma/\epsilon_Q)}, \tag{47}$$

at the point T. At the boundary of the cylinder the equality $y=1/z$ holds which is equivalent to Eq. (32).

[1] A. Z. Wang, M. F. A. da Silva, and N. O. Santos, Class. Quantum Grav. **14**, 2417 (1997).
[2] T. G. Philbin, Class. Quantum Grav. **13**, 1217 (1996).
[3] J. Stela and D. Kramer, Acta Phys. Pol. B **21**, 843 (1990).
[4] D. Kramer, Class. Quantum Grav. **5**, 393 (1988).

[5] J. Wainwright and G. F. R. Ellis, ed., *Dynamical systems in cosmology* (Cambridge U. P., Cambridge, 1997).
[6] A. B. Evans, J. Phys. A **10**, 1303 (1977).
[7] D. Kramer, H. Stephani, M. A. H. MacCallum, and E. Herlt, *Exact solutions of Einstein's field equations* (Cambridge U. P., Cambridge, 1980).
[8] U. S. Nilsson and C. Uggla, Class. Quantum Grav. **14**, 2931 (1997).
[9] C. W. Misner, Phys. Rev. D **42**, 1068 (1969).
[10] R. T. Jantzen and C. Uggla, ''The Kinematical Role of Automorphisms in the Orthonormal Frame Approach to Bianchi Cosmology,'' submitted to Class. Quantum Grav.
[11] M. Bradley and A. Karlhede, Class. Quantum Grav. **7**, 449 (1990).
[12] M. Bradley and M. Marklund, Class. Quantum Grav. **13**, 3021 (1996).
[13] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time* (Cambridge U. P., Cambridge, 1973).
[14] U. S. Nilsson and C. Uggla, Class. Quantum Grav. **14**, 1965 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 6 JUNE 1998

Closed formula for the matrix elements of the volume operator in canonical quantum gravity

T. Thiemann[a)]
Physics Department, Harvard University, Cambridge, Massachusetts 02138

(Received 15 May 1997; accepted for publication 16 February 1998)

We derive a closed formula for the matrix elements of the volume operator for canonical Lorentzian quantum gravity in four space–time dimensions in the continuum in a spin-network basis. We also display a new technique of regularization which is state dependent but we are forced to it in order to maintain diffeomorphism covariance and in that sense it is natural. We arrive naturally at the expression for the volume operator as defined by Ashtekar and Lewandowski up to a state-independent factor. © *1998 American Institute of Physics.*
[S0022-2488(98)02306-8]

I. INTRODUCTION

The volume functional V for Lorentzian canonical quantum gravity plays a quite important role in the quantization process: if one follows the approach advertized in Ref. 1 and chooses a representation in which traces of the holonomy along closed loops for a certain SU(2) A_a^i connection can be promoted to basic configuration operators, then it becomes technically very hard to define physically interesting operators such as the Euclidean and Lorentzian Hamiltonian constraint operators, an operator corresponding to the length of a curve, an operator corresponding to the ADM (Arnowitt–Deser–Misner) energy, matter Hamiltonian operators, and so forth. The common reason is that all these operators involve only the co-triad e_a^i of the intrinsic metric of the initial data hypersurface Σ polynomially which, however, cannot be written as a polynomial in the momentum conjugate to A_a^i and therefore does not seem to make any sense as a quantum operator.

Fortunately, there is a trick available: as proved in Refs. 2 and 3, the total volume of Σ is the generating functional of the co-triad meaning that one gets e_a^i from V essentially by functional derivation. Therefore, if we write $\delta t \dot{s}^a e_a^i \propto \{h_s, V\} + o(\delta t^2)$, where s is an edge with parameter length δt, h_s is the holonomy along s, and $\{.,.\}$ denotes Poisson brackets, then it would be sufficient to quantize V in such a way that (gauge-invariant) functions of a finite number of holonomies form a dense domain. Such a satisfactory quantization of V which has precisely this property was indeed given already in the literature. The first paper on the quantization of the volume was Ref. 4. Another quantization of the volume was later mentioned for the first time in Ref. 1 and supplemented recently by a detailed paper (second reference in Ref. 5). After Refs. 1 and 4, another paper (first reference in Ref. 6) appeared which focuses on the quantization on a lattice while Refs. 1 and 4 deal with the continuum formulation.

Therefore, the strategy outlined above and in Refs. 2 and 3 can be implemented and, in particular, was successfully employed to construct various operators listed above.[2,3,7–9] However, the explicit computation of the spectrum of most of those operators was not performed so far. The reason is that the spectra of all these operators are clearly largely determined by the spectrum of the volume operator whose precise spectrum is fairly unknown except for the most simple situations. (In Ref. 10 a general four-valent case in the continuum is considered while in Ref. 6 a general six-valent case, the maximum valence of a vertex in a hyper-cubic lattice, is considered. In Ref. 7 the extension to non-gauge-invariant states of the volume operator published first in the present paper is applied to compute the spectrum of a length operator in a general four-valent case and in Ref. 11 the same extension is employed to compute the matrix elements of the operator

[a)]Electronic mail: thiemann@math.harvard.edu New address: Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, Schlaatzweg 1, 14473 Potsdam, Germany. Electronic mail: thiemann@aei-potsdam.mpg.de

 © 1998 American Institute of Physics

constructed in Refs. 2 and 3 for a general three-valent vertex. See below for the meaning of ''valence'' and ''vertex'' in the theories under consideration.)

The present paper is devoted to the derivation of the complete set of matrix elements of this operator in closed form in terms of known quantities from which the spectrum can be obtained, in principle, straightforwardly although only numerically in general. A similar formula has been derived earlier in Ref. 12 by De Pietri and Rovelli. The difference between Ref. 12 and the present paper is twofold: First of all, we work with the operator introduced by Ashtekar and Lewandowski[5] while Ref. 12 works with the operator introduced earlier by Rovelli and Smolin.[4] These two operators agree only on vertices of valence smaller than or equal to four. Second and more importantly, we prefer to employ the connection representation of quantum gravity which leads to the, to our mind, quite compact invariant vector field calculus whereas Ref. 12 employs the loop representation which leads to *a*, to our mind, quite cumbersome, partly graphical notation. While both representations were proved to be unitarily equivalent in Ref. 13 (see Ref. 14 for a later proof) the (inverse) loop transformation that maps them into each other is rather complicated and thus a full proof that both formulas as well as the methods to arrive at them are identical is lacking at the moment (a proof exists for the case of a single nonplanar four-valent vertex[10]).

In Sec. III, after introducing the notation in Sec. II, we develop a novel technique to regularize the volume operator which is state dependent very much in the same way as the regularization displayed in the second reference of Ref. 5. This part of the paper is a continuation of the work on the point-splitting regularizations of the volume and area operators as displayed in Refs. 1 and 5. The derivation is rather short and gives a diffeomorphism covariant end result. We are unambiguously led to the Ashtekar–Lewandowski volume operator defined in Ref. 15.

In Sec. IV we repeat the argument given in Refs. 5 and 10 which shows that the spectrum is discrete and that its computation reduces to linear algebra, in the sense that we just need to determine separately the eigenvalues of an infinite number of finite-dimensional, positive semidefinite and symmetric real-valued matrices.

Finally, in Sec. V we derive the explicit formula for an arbitrary matrix element in a spin-network basis and compute the eigenvalues of the associated matrices in a number of simple cases. It should be noted that the formulas obtained, albeit quite complicated, are the starting point for a numerical evaluation of the spectrum and are in particular useful to determine the large spin behavior of the volume operator.

II. NOTATION

Let us first introduce the notation. We consider the Hamiltonian formulation of the Palatini action of four-dimensional Lorentzian vacuum gravity.

Denote by q_{ab} the intrinsic metric on the initial data hypersurface Σ and let e_a^i be its co-triad where $i,j,k,\ldots$ denote su(2) indices. The densitized triad $E_i^a := \det((e_b^j))e_i^a$ multiplied by $1/\kappa$, κ being Newton's constant, is the momentum conjugate to $K_a^i := \mathrm{sgn}(\det((e_c^j)))K_{ab}e_i^b$ where K_{ab} is the extrinsic curvature of Σ. Notice that usually the term ''densitized triad'' is used for the quantity $\sqrt{\det(q)}e_i^a$ which differs (possibly) by a sign from the above definition.

Upon performing a canonical point transformation discovered by Ashtekar[16] on the phase space of general relativity (of which we use its real Lorentzian variant; the real variant for Euclidean signature was first described in Ref. 17 and for the Lorentzian one later in Ref. 18) coordinated by $(K_a^i, E_i^a/\kappa)$ we arrive at the chart $(A_a^i/\kappa := (\Gamma_a^i + K_a^i)/\kappa, E_i^a)$, where Γ_a^i is the spin connection of e_a^i. It is easy to see that the configuration variable A_a^i is an su(2) connection. The reason for modifying the definition of the densitized triad is that only with our choice the above transformation is a symplectomorphism as explained in more detail in Ref. 3.

The virtue of casting general relativity into a connection dynamics formulation initialized by Ashtekar is as follows.

(1) First of all, the phase space of general relativity is now embedded into that of an su(2) gauge theory, thereby opening access to a wealth of techniques that have been proving useful in quantizing those theories. Probably the most important technique is the use of traces of the holonomy along piecewise analytic loops in Σ, so-called Wilson-loop functionals, to coordi-

natize the space of smooth connections modulo gauge transformations $\mathcal{A}/\mathcal{G}$. For gauge theories that have been emphasized, for instance, in Refs. 19 and 20 and for quantum gravity in Ref. 21.

(2) Using those Wilson loop functionals one can construct[22–24] a quantum configuration space of generalized connection modulo gauge transformations $\overline{\mathcal{A}/\mathcal{G}}$ and find natural, σ-additive, diffeomorphism invariant, regular Borel probability measures μ_0 thereon which equips us with a Hilbert space structure $\mathcal{H}:=L(\overline{\mathcal{A}/\mathcal{G}},d\mu_0)$. Of course, for this to make sense we have to make the following assumption: Wilson loop functionals continue to make sense as operators in the quantum theory. We will adopt this viewpoint in the sequel in the hope of capturing one of the physically interesting phases of quantum general relativity. The interested reader is referred to Ref. 1 and references therein for further details.

The following paragraph is meant to equip the reader with a working knowledge of the tools associated with $\overline{\mathcal{A}/\mathcal{G}}$.

We begin by defining cylindrical functions. Since the gauge-invariant information contained in a connection is captured by finite linear sums of products of Wilson loop functionals we may label gauge-invariant functions by piecewise analytic graphs γ which are just the union of the loops involved. Such a graph consists of maximally analytic pieces, called the edges e of γ, and the edges meet in the vertices v of γ. A function f on $\overline{\mathcal{A}/\mathcal{G}}$ is said to be cylindrical with respect to a graph γ iff it can be written as $f=f_\gamma\circ p_\gamma$, where $p_\gamma(A)=(h_{e_1}(A),\ldots,h_{e_n}(A))$ and where $e_1,\ldots,e_n$ are the edges of γ. Here $h_e(A)$ is the holonomy along e evaluated at $A\in\overline{\mathcal{A}/\mathcal{G}}$ and f_γ is a complex-valued function on $\mathrm{SU}(2)^n$. Since a function cylindrical with respect to a graph γ is automatically cylindrical with respect to any graph bigger than γ, a cylindrical function is actually given by a whole equivalence class of functions f_γ. In the sequel we will not distinguish between this equivalence class and one of its representatives.

We say that a function cylindrical with respect to a graph γ is of class C^n if it is of class C^n with respect to the standard differentiable structure of $\mathrm{SU}(2)^n$ and denote this class of functions by $\mathrm{Cyl}^n_\gamma(\overline{\mathcal{A}/\mathcal{G}})$, $\mathrm{Cyl}^n(\overline{\mathcal{A}/\mathcal{G}}):=\cup_{\gamma\in\Gamma}\mathrm{Cyl}^n_\gamma(\overline{\mathcal{A}/\mathcal{G}})$, where Γ is the set of piecewise analytic graphs in Σ. Thus we have a differential calculus[15] on $\overline{\mathcal{A}/\mathcal{G}}$.

Finally, the measure μ_0 is the faithful projective limit[25,26] of the self-consistent projective family $(\mu_{0,\gamma})_{\gamma\in\Gamma}$ defined by

$$\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_0(A) f(A) := \int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_{0,\gamma}(A) f_\gamma(p_\gamma(A)) := \int_{SU(2)^n} d\mu_H(g_1)\cdots d\mu_H(g_n) f_\gamma(g_1,\ldots,g_n),$$

where $g_I:=h_{e_I}(A)$. This equips us with an integral calculus on $\overline{\mathcal{A}/\mathcal{G}}$.

III. REGULARIZATION OF THE VOLUME OPERATOR

In this section we present a new regularization of the volume operator which uses the methods derived by Abhay Ashtekar and Jurek Lewandowski in Refs. 5, 10, and 15. Most of the issues involved in the regularization as presented here are borrowed from their work.

The advantage of our procedure is that it is rather short and compact.

Let $R\subset\Sigma$ be an open, connected region of Σ. We have the identity

$$\frac{1}{3!}\,\epsilon_{abc}\epsilon^{ijk}E^a_iE^b_jE^c_k=\det((E^a_i))=\det((q_{ab}))=[\det((e^i_a))]^2=:\det(q)\geqslant 0 \tag{III.1}$$

and can thus write the volume of the region R as measured by the metric q_{ab} as follows:

$$V(R):=\int_R d^3x\sqrt{\det(q)}=\sqrt{\frac{1}{3!}\,\epsilon_{ijk}\epsilon^{abc}E^a_iE^b_jE^c_k}. \tag{III.2}$$

The next step is to smear the fields E^a_i. Let $\chi_\Delta(p,x)$ be the characteristic function in the coordinate x of a cube with center p spanned by the three vectors $\mathbf{\Delta}_i=\Delta_i\mathbf{n}_i(\Delta)$ where $\mathbf{n}_i$ is a normal vector in the frame under consideration and which has coordinate volume, vol

$=\Delta_1\Delta_2\Delta_3\ \det(\mathbf{n}_1,\mathbf{n}_2,\mathbf{n}_3)$ (we assume the three normal vectors to be right oriented). In other words, $\chi_\Delta(p,x)=\Pi_{i=1}^3\theta(\Delta_i-|\langle n_i,x-p\rangle|)$ where $\langle .,.\rangle$ is the standard Euclidean inner product and $\theta(y)=1$ for $y>0$ and zero otherwise.

We consider the smeared quantity

$$E(p,\Delta,\Delta',\Delta'')\coloneqq\frac{1}{\mathrm{vol}(\Delta)\mathrm{vol}(\Delta')\mathrm{vol}(\Delta'')}\int_\Sigma d^3x\int_\Sigma d^3y\int_\Sigma d^3z\ \chi_\Delta(p,x)\chi_{\Delta'}\left(p,\frac{x+y}{2}\right)$$
$$\times\chi_{\Delta''}\left(p,\frac{x+y+z}{3}\right)\frac{1}{3!}\,\epsilon_{abc}\epsilon^{ijk}E_i^a(x)E_j^b(y)E_k^c(z). \tag{III.3}$$

Notice that if we take the limits $\Delta_i,\Delta_i',\Delta_i''\to 0$ in any combination and in any rate with respect to each other then we get back to (III.1) evaluated at the point p. This holds for any choice of linearly independent normal vectors $\mathbf{n}_i,\mathbf{n}_i',\mathbf{n}_i''$. The strange argument $(x+y+z)/3$ will turn out to be very crucial in obtaining a manifestly diffeomorphism covariant result. We will see this in a moment.

Then it is easy to see that the classical identity

$$V(R)=\lim_{\Delta\to 0}\lim_{\Delta'\to 0}\lim_{\Delta''\to 0}\int_R d^3p\sqrt{|E(p,\Delta,\Delta',\Delta'')|} \tag{III.4}$$

holds. We could introduce absolute values because we have the classical identity $\det(q)=|\det(q)|$ and the modulus is necessary for the regulated quantity (III.3) is not non-negative anymore. This continues to hold in the quantum theory and it is an open question whether one should impose the anholonomic constraint $\det(q)\geqslant 0$ in the quantum theory.[27] Also, observe that (III.3) is not gauge invariant any longer in contrast to $V(R)$, however, we will be interested only in the limit of shrinking all Δ to points and, as it turns out, recover gauge invariance also in the quantum theory.

The virtue of introducing the quantities (III.3) is that they enable us to define operators corresponding to $V(R)$, in the limit that all Δ shrink to a point, and which have the dense domain $\mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$. To see this, notice that due to the canonical brackets $\{A_a^i(x),E_j^b(y)\}=\kappa\delta_a^b\delta_j^i\delta^{(3)}(x,y)$ we are naturally led to represent the operator corresponding to E_i^a by $\hat{E}_i^a(x)=-i\ell_p^2\delta/\delta A_a^i(x)$ where $\ell_p=\sqrt{\hbar\kappa}$ is the Planck length.

Now let a graph γ be given. In order to simplify the notation, we subdivide each edge e with endpoints v,v' which are vertices of γ into two segments s,s' where $e=s\circ(s')^{-1}$ and s has an orientation such that it is *outgoing* at v while s' has an orientation such that it is *outgoing* at v'. This introduces new vertices $s\cap s'$ which we will call pseudo-vertices because they are not points of nonanalyticity of the graph. Let $E(\gamma)$ be the set of these segments of γ but $V(\gamma)$ the set of true (as opposed to pseudo) vertices of γ. Let us now evaluate the action of $\hat{E}_i^a(p,\Delta)\coloneqq 1/\mathrm{vol}(\Delta)\int_\Sigma d^3x\chi_\Delta(p,x)\hat{E}_i^a(x)$ on a function $f=p_\gamma^* f_\gamma$ cylindrical with respect to γ. We find $\{e:[0,1]\to\Sigma;\ t\to e(t)$ being a parametrization of the edge $e\}$

$$\hat{E}_i^a(p,\Delta)f=-\frac{i\ell_p^2}{\mathrm{vol}(\Delta)}\sum_{e\in E(\gamma)}\int_0^1 dt\chi_\Delta(p,e(t))\dot{e}^a(t)\,\frac{1}{2}\,\mathrm{tr}\left(h_e(0,t)\tau_i h_e(t,1)\,\frac{\partial}{\partial h_e(0,1)}\right)f_\gamma. \tag{III.5}$$

Here we have used (1) the fact that a cylindrical function is already determined by its values on $\mathscr{A}/\mathscr{G}$ rather than $\overline{\mathscr{A}/\mathscr{G}}$ so that it makes sense to take the functional derivative, (2) the definition of the holonomy as the path-ordered exponential of $\int_e A$ with the smallest parameter value to the left, (3) $A=dx^a\,A_a^i\tau_i/2$ where $\mathrm{SU}(2)\ni\tau_j=-i\sigma_j$, σ_i being the usual Pauli matrices, so that $[\tau_i/2,\tau_j/2]=\epsilon_{ijk}\tau_k/2$, and (4) we have defined $\mathrm{tr}(h\partial/\partial g)=h_{AB}\partial/\partial g_{AB}$, $A,B,C,\ldots$ being SU(2) indices. The state that appears on the right-hand side of (III.5) is actually well defined, in the sense of functions of connections, only when A is smooth for otherwise the integral over t does not exist.[28] However, we will be interested only in quantities constructed from operators of the form (III.5) and for which the limit of shrinking $\Delta\to 0$ to a point has a meaning in the sense of $\mathscr{H}=L_2(\overline{\mathscr{A}/\mathscr{G}},d\mu_0)$ and therefore will not be concerned with the actual range of the operator (III.5).

We now wish to evaluate the whole operator $\hat{E}(p,\Delta,\Delta',\Delta'')$ on f. It is clear that we obtain three types of terms: the first type comes from all three functional derivatives acting on f only, the second type comes from two functional derivatives acting on f and the remaining one acting on the trace appearing in (III.5), and, finally, the third type comes from only one derivative acting on f_γ and the remaining two acting on the trace. Explicitly we find [we mean by $\theta(t,t')$ the theta function which is unity if $0<t<t'<1$ and zero otherwise and likewise $\theta(t,t',t'')$ is 1 if $0<t<t'<t''<1$ and zero otherwise]

$$
\begin{aligned}
&\hat{E}(p,\Delta,\Delta',\Delta'')f \\
&\quad= \frac{i\not{\ell}_p^6}{8\cdot 3!\ \mathrm{vol}(\Delta)\mathrm{vol}(\Delta')\mathrm{vol}(\Delta'')}\ \epsilon_{abc}\epsilon^{ijk}\int_{[0,1]^3} dt\ dt'\ dt'' \\
&\qquad\times\Bigg\{\sum_{e,e',e''\in E(\gamma)} \dot{e}(t)^a\dot{e}'(t')^b\dot{e}''(t'')^c\chi_\Delta(p,e(t))\chi_{\Delta'}\left(p,\ \frac{e(t)+e'(t')}{2}\right) \\
&\qquad\times\chi_{\Delta''}\left(p,\ \frac{e(t)+e'(t')+e''(t'')}{3}\right) \\
&\qquad\times\mathrm{tr}\left(h_{e''}(0,t'')\tau_k h_{e''}(t'',1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right)\mathrm{tr}\left(h_{e'}(0,t')\tau_j h_{e'}(t',1)\ \frac{\partial}{\partial h_{e'}(0,1)}\right) \\
&\qquad\times\mathrm{tr}\left(h_e(0,t)\tau_i h_e(t,1)\ \frac{\partial}{\partial h_e(0,1)}\right) \\
&\qquad+\sum_{e',e''\in E(\gamma)} \dot{e}''(t)^a\dot{e}'(t')^b\dot{e}''(t'')^c\chi_\Delta(p,e''(t))\chi_{\Delta'}\left(p,\ \frac{e''(t)+e'(t')}{2}\right) \\
&\qquad\times\chi_{\Delta''}\left(p,\ \frac{e''(t)+e'(t')+e''(t'')}{3}\right)\Bigg[\theta(t,t'')\mathrm{tr}\left(h_{e''}(0,t)\tau_i h_{e''}(t,t'')\tau_k h_{e''}(t'',1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right) \\
&\qquad+\theta(t'',t)\mathrm{tr}\left(h_{e''}(0,t'')\tau_k h_{e''}(t'',t)\tau_i h_{e''}(t,1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right)\Bigg]\mathrm{tr}\left(h_{e'}(0,t')\tau_j h_{e'}(t',1)\ \frac{\partial}{\partial h_{e'}(0,1)}\right) \\
&\qquad+\sum_{e',e''\in E(\gamma)} \dot{e}'(t)^a\dot{e}'(t')^b\dot{e}''(t'')^c\chi_\Delta(p,e'(t))\chi_{\Delta'}\left(p,\ \frac{e'(t)+e'(t')}{2}\right) \\
&\qquad\times\chi_{\Delta''}\left(p,\ \frac{e'(t)+e'(t')+e''(t'')}{3}\right)\mathrm{tr}\left(h_{e''}(0,t'')\tau_k h_{e''}(t'',1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right) \\
&\qquad\times\Bigg[\theta(t,t')\mathrm{tr}\left(h_{e'}(0,t)\tau_i h_{e'}(t,t')\tau_j h_{e'}(t',1)\ \frac{\partial}{\partial h_{e'}(0,1)}\right) \\
&\qquad+\theta(t',t)\mathrm{tr}\left(h_{e'}(0,t')\tau_j h_{e'}(t',t)\tau_i h_{e'}(t,1)\ \frac{\partial}{\partial h_{e'}(0,1)}\right)\Bigg] \\
&\qquad+\sum_{e,e''\in E(\gamma)} \dot{e}(t)^a\dot{e}''(t')^b\dot{e}''(t'')^c\chi_\Delta(p,e(t))\chi_{\Delta'}\left(p,\ \frac{e(t)+e''(t')}{2}\right) \\
&\qquad\times\chi_{\Delta''}\left(p,\ \frac{e(t)+e''(t')+e''(t'')}{3}\right) \\
&\qquad\times\Bigg[\theta(t',t'')\mathrm{tr}\left(h_{e''}(0,t')\tau_j h_{e''}(t',t'')\tau_k h_{e''}(t'',1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right) \\
&\qquad+\theta(t'',t')\mathrm{tr}\left(h_{e''}(0,t'')\tau_k h_{e''}(t'',t')\tau_j h_{e''}(t',1)\ \frac{\partial}{\partial h_{e''}(0,1)}\right)\Bigg]
\end{aligned}
$$

$$\times \mathrm{tr}\left(h_e(0,t)\,\tau_i h_e(t,1)\,\frac{\partial}{\partial h_e(0,1)}\right)$$

$$+\sum_{e''\in E(\gamma)} \dot{e}''(t)^a \dot{e}''(t')^b \dot{e}''(t'')^c \chi_\Delta(p,e''(t))\chi_{\Delta'}\left(p,\frac{e''(t)+e''(t')}{2}\right)$$

$$\times \chi_{\Delta''}\left(p,\frac{e''(t)+e''(t')+e''(t'')}{3}\right)$$

$$\times\Bigg[\theta(t,t',t'')\mathrm{tr}\left(h_{e''}(0,t)\,\tau_i h_{e''}(t,t')\,\tau_j h_{e''}(t',t'')\,\tau_k h_{e''}(t'',1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)$$

$$+\theta(t,t'',t')\mathrm{tr}\left(h_{e''}(0,t)\,\tau_i h_{e''}(t,t'')\,\tau_k h_{e''}(t'',t')\,\tau_j h_{e''}(t',1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)$$

$$+\theta(t',t'',t)\mathrm{tr}\left(h_{e''}(0,t')\,\tau_j h_{e''}(t',t'')\,\tau_k h_{e''}(t'',t)\,\tau_i h_{e''}(t,1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)$$

$$+\theta(t',t,t'')\mathrm{tr}\left(h_{e''}(0,t')\,\tau_j h_{e''}(t',t)\,\tau_i h_{e''}(t,t'')\,\tau_k h_{e''}(t'',1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)$$

$$+\theta(t'',t,t')\mathrm{tr}\left(h_{e''}(0,t'')\,\tau_k h_{e''}(t'',t)\,\tau_i h_{e''}(t,t')\,\tau_j h_{e''}(t',1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)$$

$$+\theta(t'',t',t)\mathrm{tr}\left(h_{e''}(0,t'')\,\tau_k h_{e''}(t'',t')\,\tau_j h_{e''}(t',t)\,\tau_i h_{e''}(t,1)\,\frac{\partial}{\partial h_{e''}(0,1)}\right)\Bigg]\Bigg\} f_\gamma$$

$$=:[\hat{O}_{1,2,3}+\hat{O}_{2,31}+\hat{O}_{12,3}+\hat{O}_{1,23}+\hat{O}_{123}]f_\gamma. \tag{III.6}$$

The fact that the integrand of the terms involved in $\hat{O}_{12,3}$, $\hat{O}_{1,23}$, $\hat{O}_{2,31}$, and $\hat{O}_{123}$ vanishes if either of the cases $0<t=t'<1$, $0<t'=t''<1$, $0<t=t'=t''<1$ occurs is due to the fact that in this case in $\hat{O}_{12,3}$, $\hat{O}_{1,23}$, $\hat{O}_{2,31}$, and $\hat{O}_{123}$ we get a trace which contains $\tau_{(i}\tau_{j)}$, $\tau_{(j}\tau_{k)}$, and $\tau_{(k}\tau_{i)}$ contracted with ϵ^{ijk} which vanishes. {To see this recall that the functional derivative is

$$\delta h_e(A)/\delta A^i_a(x)=\frac{1}{2}\int_0^1 dt\Bigg[\frac{1}{2}\,\delta^{(3)}(e(t+),x)\dot{e}(t+)^a h_e(0,t)\,\tau_i h_e(t,1)$$

$$+\frac{1}{2}\,\delta^{(3)}(e(t-),x)\dot{e}(t-)^a h_e(0,t)\,\tau_i h_e(t,1)\Bigg]$$

(one sided derivatives and δ distributions). This expression is also correct if x is an endpoint of e [in which case there is only one term which survives in (III.7)], which results in the case that we consider $h_{e_1}\tau_j h_{e_2}$ instead of h_e, $e=e_1\circ e_2$, $x=e_1\cap e_2$ a point of analyticity, in a term involving $\tau_{(i}\tau_{j)}$.}

Given a triple e,e',e'' of (not necessarily distinct) edges of γ, consider the functions

$$x_{ee'e''}(t,t',t''):=\frac{e(t)+e'(t')+e''(t'')}{3}. \tag{III.7}$$

This function has the interesting property that the Jacobian is given by

$$3^3\,\det\left(\frac{\partial(x^1_{ee'e''},x^2_{ee'e''},x^3_{ee'e''})(t,t',t'')}{\partial(t,t',t'')}\right)=\epsilon_{abc}\dot{e}(t)^a\dot{e}'(t')^b\dot{e}''(t'')^a, \tag{III.8}$$

which is precisely the form of the factor that enters all the integrals in (III.6). This is why we have introduced the strange argument $(x+y+z)/3$.

We now consider the limit $\Delta_i,\Delta_i',\Delta_i''\to 0$. The idea is that all quantities in (III.6) are meaningful in the sense of functions on smooth connections and thus limits of functions as $\Delta\to 0$ are to be understood with respect to any Sobolev topology. The miracle is that the final function is again cylindrical and thus the operator that results in the limit has an extension to all of $\overline{\mathscr{A}/\mathscr{G}}$.

Lemma III.1: For each triple of edges e,e',e'' there exists a choice of vectors $\mathbf{n}_i,\mathbf{n}_i',\mathbf{n}_i''$ and a way to guide the limit $\Delta_i,\Delta_i',\Delta_i''\to 0$ such that

$$\int_{[0,1]^3} \det\left(\frac{\partial(x^a_{ee'e''})}{\partial(t,t',t'')}\right)\chi_\Delta(p,e)\chi_{\Delta'}(p,(e+e')/2)\chi_{\Delta''}(p,(e+e'+e'')/3)\hat{O}_{ee'e''} \qquad \text{(III.9)}$$

vanishes

(a) *if e,e',e'' do not all intersect p or*
(b) $\det(\partial(x^a_{ee'e''})/\partial(t,t',t''))_p=0$ *(which is a diffeomorphism invariant statement).*

Otherwise it tends to $1/8\,\mathrm{sgn}(\det(\partial(x^a_{ee'e''})/\partial((t,t',t'')))_p\hat{O}_{e,e',e''}(p)\Pi_{i=1}^3\Delta_i''$. *Here we have denoted by $\hat{O}_{ee'e''}(t,t',t'')$ the trace(s) involved in the various terms of (III.6).*

Remark: The idea that one should adapt the regularization to each triple of edges was first communicated to the author by Jurek Lewandowski. Its justification relies on the fact that the classical expression does not depend on the way we regularize. What is new here is to introduce even more parameters than in Ref. 10 to be adapted to a triple which simplifies the proofs. What is also new is that we introduced the $(x+y+z)/3$ argument which makes the proof especially clear. It should be mentioned, however, that different regularization schemes result in different operators (see, e.g., Ref. 4); on the other hand, the requirement of consistency of the dynamics of the theory as defined in Refs. 2 and 3 rules out the operator of Ref. 4.

Proof: If at least one of e,e',e'' does not intersect p, then, if we choose Δ_i, etc. smaller than some finite number Δ_0, (III.9) vanishes identically since the support of the characteristic functions is in a neighborhood around p which shrinks to zero with the Δ_i, etc.

So let us assume that all of e,e',e'' intersect p at parameter value t_0,t_0',t_0'' (this value is unique because the edges are not self-intersecting). Then we can write $e(t)=p+c(t-t_0)$, where c is analytic and vanishes at $\tau=t-t_0=0$. We have the case subdivision:
Case I: $\det(\partial(x^a_{ee'e''})/\partial(t,t',t''))_{t_0}=0$.
Case I(a): All of $\dot{c}(0),\dot{c}'(0),\dot{c}''(0)$ are colinear.
Case I(b): Two of $\dot{c}(0),\dot{c}'(0),\dot{c}''(0)$ are colinear and the third is linearly independent of them.
Case I(c): No two of $\dot{c}(0),\dot{c}'(0),\dot{c}''(0)$ are colinear.
Case II: $\det(\partial(x^a_{ee'e''})/\partial(t,t',t''))_{t_0}\neq 0$.
Notice that all vectors $\dot{c}(0),\dot{c}'(0),\dot{c}''(0)$ are nonvanishing by the definition of a curve.

We consider first case I. We exclude the trivial case that all three curves lie in a coordinate plane or line such that the determinant already vanishes for all finite values of the Δ's. Therefore there exist linearly independent unit vectors u,v,w (not necessarily orthogonal) in terms of which we may express c,c',c''.

In case I(a) we have an expansion of the form

$$\begin{aligned}
c(t)&=au(t+o(t''))+bv(t^m+o(t^{m+1}))+cw(t^n+o(t^{n+1})),\\
c'(t)&=au(t+o(t^2))+b'v(t^{m'}+o(t^{m'+1}))+c'w(t^{n'}+o(t^{n'+1})),\\
c''(t)&=a''u(t+o(t^2))+b''v(t^{m''}+o(t^{m''+1}))+c''w(t^{n''}+o(t^{n''+1})),
\end{aligned} \qquad \text{(III.10)}$$

where $a,b,c,a',b',c',a'',b'',c''$ are real numbers with $aa'a''\neq 0$ and at least one of the b's and c's being different from zero (also not for instance $b=c=b'=c'=0$). Furthermore, $m,m',m'',n,n',n''\geqslant 2$. The characteristic functions have support in coordinate cubes spanned by the vectors $\mathbf{n}_i,\mathbf{n}_i',\mathbf{n}_i''$. Now, since u,v,w are linearly independent, we may simply *choose*, for instance, $\mathbf{n}_i:=u$, $\mathbf{n}_i':=v$, $\mathbf{n}_i'':=w$. It follows then and from the fact that $0\leqslant\chi\leqslant 1$ that

$$\begin{aligned}
&\chi_\Delta(p,e)\chi_{\Delta'}(p,(e+e')/2)\chi_{\Delta''}(p,(e+e'+e'')/3)\\
&\quad=\tilde{\chi}_\Delta(0,c)\tilde{\chi}_{\Delta'}(0,(c+c')/2)\tilde{\chi}_{\Delta''}(0,(c+c'+c'')/3)\\
&\quad\leqslant\theta_{\Delta_1}(\langle c,u\rangle)\,\theta_{\Delta_1'}(\langle(c+c')/2,v\rangle)\,\theta_{\Delta_1''}(\langle(c+c'+c'')/3,w\rangle).
\end{aligned}\tag{III.11}$$

From the explicit expansions of c,c',c'' we conclude that (III.11) has the bound

$$\theta_{\delta_1\Delta_1}(t)\,\theta_{\delta_1'\Delta_1'}(t')\,\theta_{\delta_1''\Delta_1''}(t'')\tag{III.12}$$

for some sufficiently large numbers $\delta_1,\delta_1',\delta_1''$. On the other hand we also see from the explicit expansion of $|\det(\partial(x^a_{ee'e''})/\partial(t,t',t''))|$ around t_0 that it is bounded by $M(|t|^k+|t'|^k+|t''|^k)$ where M is a positive number and where $k=\min(m+n',m+n'',m'+n,m'+n'',m''+n,m''+n')-2\geqslant 2$.

The prescription of how to guide the limit in case I(a) is then to synchronize $\Delta_1=\Delta_1'=\Delta_1''=\Delta$ and to take the limit $\Delta\to 0$ first. The integral is at least of order Δ^5 while we divide only by an order of Δ^3 so that the result vanishes.

In case I(b) we have an expansion of the form (let w.l.g. c,c' have collinear tangents)

$$\begin{aligned}
c(t)&=au(t+o(t^2))+bv(t^m+o(t^{m+1}))+cw(t^n+o(t^{n+1})),\\
c'(t)&=au(t+o(t^2))+b'v(t^{m'}+o(t^{m'+1}))+c'w(t^{n'}+o(t^{n'+1})),\\
c''(t)&=a''v(t+o(t^2))+b''u(t^{m''}+o(t^{m''+1}))+c''w(t^{n''}+o(t^{n''}+1)).
\end{aligned}\tag{III.13}$$

We now argue as above and find that the product of the characteristic functions can be estimated by

$$\theta_{\delta_1\Delta_1}(t)\,\theta_{\delta_1'\Delta_1'}(t')\,\theta_{\delta_2''\Delta_2''}(t'')$$

while the determinant can be estimated as above just that k is now given by $k=\min(m,m',m'',n,n',n'')-1\geqslant 1$.

The prescription is now $\Delta_1=\Delta_1'=\Delta_2'':=\Delta\to 0$ first and we conclude that the integral is at least of order Δ^4 while we divide again only by Δ^3 such that the limit vanishes.

In case I(c) finally we have an expansion of the form

$$\begin{aligned}
c(t)&=au(t+o(t^2))+bv(t^m+o(t^{m+1}))+cw(t^n+o(t^{n+1})),\\
c'(t)&=av(t+o(t^2))+b'v(t^{m'}+o(t^{m'+1}))+c'w(t^{n'}+o(t^{n'+1})),\\
c''(t)&=a''u(t+o(t^2))+b''v(t+o(t^2))+c''w(t^{n''}+o(t^{n''+1})).
\end{aligned}\tag{III.14}$$

This time we estimate the product of the characteristic functions for instance by

$$\theta_{\delta_1\Delta_1}(t)\,\theta_{\delta_2'\Delta_2'}(t')\,\theta_{\delta_2''\Delta_2''}(t''),$$

while the determinant can be estimated as above and k is given by $k=\min(m,m',n,n',n'')-1\geqslant 1$ so that we have actually the same situation as in case I(b) upon synchronizing this time $\Delta_1=\Delta_2'=:\Delta_2''=:\Delta\to 0$.

As for case II we observe that the nonvanishing of the functional determinant at p implies that the map $x_{ee'e''}$ is actually invertible in a neighborhood of p by the inverse function theorem. In other words, there is only one point (t_0,t_0',t_0'') such that $x_{ee'e''}(t_0,t_0',t_0'')=p$. Moreover, since the determinant is nonvanishing at p, all three edges must be distinct from each other. It follows now from our choice of edges that p must be a vertex $v=e\cup e'\cup e''$ of γ in order that the result is nonvanishing and thus from the choice of parametrization $t_0=t_0'=t_0''=0$.

Therefore, if we take the limit $\Delta_i''\to 0$ first in any order, then the condition $\chi_{\Delta''}(p,x_{ee'e''})=1$ will actually imply $\chi_\Delta(p,e)=\chi_{\Delta'}(p,(e+e')/2)=1$ for small enough Δ_i'' so that we can take

these characteristic functions out of the integral and replace them by 1 if p is a common vertex of all three edges. Also we can replace the operator $\hat{O}_{ee'e''}(t,t',t'')$ by $\hat{O}_{ee'e''}(v)$. This holds only if the triple intersects in p.

If not all of e,e',e'' intersect in p, then the limit will vanish anyway if we take a suitable limit of the Δ_i as we have shown before. We can account for that case by replacing $\chi_\Delta(p,e),\chi_{\Delta'}(p,(e+e')/2)$ by $\chi_\Delta(p,v)\chi_{\Delta'}(p,v)$. Here v is the common vertex at which the distinct e,e',e'' must be incident, otherwise they could not even pass through a small enough neighborhood of p. We can also assume that all three edges have linearly independent tangents at v and expand still around $t=0$. The remaining integral divided by $\Delta_1''\Delta_2''\Delta_3''$ then tends to

$$\int_{[0,1]^3} d^3t\ \det\left(\frac{\partial x_{ee'e''}}{\partial t}\right)\delta^{(3)}(p,x_{ee'e''})=s(e,e',e'')\int_{C_{ee'e''}} d^3x\ \delta^{(3)}(p,x)=\frac{1}{8}\,s(e,e',e''), \tag{III.15}$$

where

$$s(e,e',e'')_v:=\mathrm{sgn}(\det(\dot{e}(0),\dot{e}'(0),\dot{e}''(0))). \tag{III.16}$$

The factor $\frac{1}{8}$ is due to the fact that in the limit $\Delta''\to 0$ we obtain an integral over $C(e,e',e'')$, the cone based at p and spanned by $\dot{e}(0),\dot{e}'(0),\dot{e}''(0)$ where the orientation is taken to be positive. This integral just equals $\int_{\mathbb{R}^3_+} d^3t\ \delta(0,t)\,\delta(0,t')\,\delta(0,t'')=\frac{1}{8}$ as one can easily check. This furnishes the proof. □

We conclude that (III.6) reduces to (in particular, the operators $\hat{O}_{12,3}\hat{O}_{1,23}\hat{O}_{2,31}\hat{O}_{123}$ drop out)

$$\lim_{\Delta''\to 0}\hat{E}(p,\Delta,\Delta',\Delta'')f=\sum_{e,e',e''}\frac{il_p^6 s(e,e',e'')_v}{8\cdot 3!\ \mathrm{vol}(\Delta)\mathrm{vol}(\Delta')}\,\chi_\Delta(p,v)\chi_{\Delta'}(p,v)\hat{O}_{e,e',e''}(0,0,0),$$

where v on the right-hand side is the intersection point of the triple of edges and it is understood that we only sum over such triples of edges which are incident at a common vertex. There is no factor of 3^3 missing because it cancels against a similar factor in $\mathrm{vol}(\Delta'')$. Moreover,

$$\hat{O}_{e,e',e''}(0,0,0)=\epsilon_{ijk}X^i_{e''}X^j_{e'}X^k_e \quad\text{and}\quad X^i_e:=X^i(h_e(0,1)):=\mathrm{tr}\left(\tau_i h_e(0,1)\,\frac{\partial}{\partial h_e(0,1)}\right) \tag{III.17}$$

is a right invariant vector field in the τ_i direction of su(2), that is, $X(hg)=X(h)$ (This important observation, which is fundamental for all that follows, is due to Abhay Ashtekar and Jurek Lewandowski.) We also have extended the values of the sign function to include 0, which takes care of the possibility that one has triples of edges with linearly dependent tangents.

The final step consists in choosing $\Delta=\Delta'$ and taking the square root of the modulus. We replace the sum over all triples incident at a common vertex $\Sigma_{e,e',e''}$ by a sum over all vertices followed by a sum over all triples incident at the same vertex $\Sigma_{v\in V(\gamma)}\Sigma_{e\cap e'\cap e''=v}$. Now, for small enough Δ and given p, at most one vertex contributes, that is, at most one of $\chi_\Delta(v,p)\neq 0$ because all vertices have finite separation. Then we can take the relevant $\chi_\Delta(p,v)=\chi_\Delta(p,v)^2$ out of the square root and take the limit which results in

$$\begin{aligned}
\hat{V}(R)_\gamma&=\int_R d^3p\,\widehat{\sqrt{\det(q)(p)}}_\gamma=\int_R d^3p\ \hat{V}(p)_\gamma,\\
\hat{V}(p)_\gamma&=l_p^3\sum_{v\in V(\gamma)}\delta^{(3)}(p,v)\hat{V}_{v,\gamma},\\
\hat{V}_{v,\gamma}&=\sqrt{\left|\frac{i}{3!\cdot 8}\sum_{e,e',e''\in E(\gamma),e\cap e'\cap e''=v}s(e,e',e'')q_{ee'e''}\right|},\\
q_{ee'e''}&=\epsilon_{ijk}X^i_eX^j_{e'}X^k_{e''},
\end{aligned} \tag{III.18}$$

where we could switch the order of the X's because a triple contributes only if the corresponding edges are distinct and so the X's commute.

Expression (III.18) is the final expression for the volume operator and coincides precisely with the expression found in Refs. 5 and 10 while it is genuinely different from the one found in Ref. 4 as pointed out in Ref. 10. Note that the final expression is manifestly diffeomorphism covariant. Although the procedure of adapting the limiting to a given triple of edges is somewhat nonstandard, there is an argument in favor of such a procedure: the discussion in Lemma III.1 reveals that any other regularization which would result in a finite contribution for the case where $s(e,e'e'')$ is zero would necessarily depend on the higher-order intersection characteristics of a triple of edges. However, since such a quantity is not diffeomorphism covariant, which is inacceptable, the dependence must be trivial, that is, a constant [this is precisely what happens in (III.4) according to (III.11)]. While there is no *a priori kinematical* reason to prefer one operator over the other, there is a *dynamical* reason: the volume operator can be seen as an essential ingredient in the regularization of the Lorentzian Wheeler–DeWitt constraint operator.[2,3] The fact that (III.18) contains $s(e,e',e'')$ is the precise reason for this operator to be anomaly free! Therefore, the regularization leading to (III.18) is the one singled out by the dynamics of the theory!

IV. SPECTRAL ANALYSIS

In this section we summarize a number of results due to Abhay Ashtekar and Jurek Lewandowski obtained in their series of papers.[5,10,15] We have included them here for the sake of self-containedness of the present paper. More specifically, the cylindrical consistency of the family of volume operators $\hat{V}_\gamma$ as defined in Ref. 15 and in the present paper, the self-adjointness of the corresponding projective limit, and the fact that the spectrum of the volume operator is discrete were first observed in Ref. 10. The proof of these properties given in Ref. 10 uses methods developed in Refs. 1, 5, 10, and 15.

This section is subdivided into three parts. First we prove that the family of operators derived in (III.18) defines a linear unbounded operator on $\mathscr{H}$. Next we show that the operator is symmetric, positive semi-definite, and admits self-adjoint extensions, and finally we show that its spectrum is discrete and that the operator so defined is anomaly free.

A. Cylindrical consistency

What we have obtained in (III.18) is a family of operators $(\hat{V}(R)_\gamma, D_\gamma)_{\gamma\in\Gamma}$. That is not enough to show that this family of cylindrical projections ''comes from'' a linear operator on $\mathscr{H}$. As proved in Ref. 15 for this to be the case we need to check that whenever $\gamma\subset\gamma'$, then we have the following.

(1) $p^*_{\gamma\gamma'} D_\gamma \subset D_{\gamma'}$, where $p_{\gamma\gamma'}$ is the restriction from γ' to γ. This condition makes sure that the operator defined on bigger graphs can be applied to functions defined on smaller graphs.
(2) $(\hat{V}(R)_{\gamma'})_{|\gamma} = \hat{V}(R)_\gamma$. This is the condition of cylindrical consistency and says that the operator on bigger graphs equals the operator on smaller graphs when restricted to functions thereon.

A graph $\gamma\subset\gamma'$ can be obtained from a bigger graph γ' by a finite series of steps consisting of the following basic ones:

(i) remove an edge from γ',
(ii) join two edges e',e'', such that $e'\cap e''$ is a point of analyticity, to a new edge $e=e'\circ(e'')^{-1}$, and
(iii) reverse the orientation of an edge.

Clearly, a dense domain for $\hat{V}(R)_\gamma$ is given by $D_\gamma := \mathrm{Cyl}^3_\gamma(\overline{\mathscr{A}/\mathscr{G}})$. This choice trivially satisfies requirement (1) since functions which just do not depend on some arguments or only on special combinations $h_e = h_{e'} h_{e''}$, $h_{e'} = h_e^{-1}$ are still thrice continuously differentiable if the original function was [here we have used the fact that SU(2) is a Lie group, that is, group multiplication and taking inverses is an analytic map].

Next, let us check cylindrical consistency. Consider first the case (i) that γ does not depend on an edge e on which γ' does. Then clearly $X_e^i f_\gamma = 0$ for any function cylindrical with respect to γ and so in the sum over triples over vertices in (III.18) the terms involving e drop out.

Next consider the case (ii). If $e = e' \circ (e'')^{-1}$ is an edge of γ and e', e'' are edges of γ' where $v := e' \cap e''$ is a point of analyticity for γ while for γ' it is not, then while v is a vertex for γ', it is only a pseudo-vertex for γ, and so in $\hat{V}(D)_\gamma$ there is no term corresponding to v. On the other hand, since the vertex v is a pseudo vertex for γ, it is in particular only two-valent and so the corresponding term in $\hat{V}(D)_{\gamma'}$ drops out. Likewise, if v is a vertex for γ at which the outgoing edge e is incident, then from right invariance of the vector field we have $X_e = X_{e' \circ (e'')^{-1}} = X_{e'}$ and so at vertices that belong to both γ and γ' the corresponding vertex operators coincide.

Finally, case (iii) is actually excluded by our unambiguous choice of orientation.

We conclude that there exists an operator $(\hat{V}(R), D)$ on $\mathscr{H}$ which is densely defined on $D = \mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$.

B. Symmetry, positivity, and self-adjointness

Notice that the vector field iX_e is symmetric on $\mathscr{H}_\gamma$, the completion of $\mathrm{Cyl}^1_\gamma(\overline{\mathscr{A}/\mathscr{G}})$ with respect to $\mu_{0,\gamma}$, e an edge of γ, because the Haar measure is right invariant. It follows from the explicit expression (III.18) in terms of the iX_e that all the projections $\hat{V}(R)_\gamma$ are symmetric. In this special case (namely, the volume operator leaves the space D_γ invariant) this is enough to show that $\hat{V}(R)$ is symmetric on D.

Furthermore, all $\hat{V}(R)_\gamma$ are positive semi-definite by inspection so that $\hat{V}(R), D$ is a densely defined, positive semi-definite, and symmetric operator. It follows that it has self-adjoint extensions, for instance, its Friedrich extension.

C. Discreteness and anomaly-freeness

The operator $\hat{V}(D)$ has the important property that it leaves the dense subset $\mathrm{Cyl}^\infty_\gamma(\overline{\mathscr{A}/\mathscr{G}}) \subset \mathscr{H}$ invariant, separately for each $\gamma \in \Gamma$. Moreover, recall that $\mathscr{H}$ possesses a convenient basis, called the spin-network basis in the sequel.[29,30,13]

Let us recall the most important features of that basis. Given a graph γ, let us associate with each edge $e \in E(\gamma)$ a spin quantum number $j_e > 0$. Also, with each vertex $v \in V(\gamma)$ we associate a certain contraction matrix c_v as follows: if π_j is the irreducible representation of SU(2) corresponding to spin j, then form the tensor product $\otimes_{e \in E(\gamma)} \pi_{j_e}(h_e)$ and for each vertex v contract the indices A of the matrices $\pi_{j_e}(h_e)_{AB}$ for all e incident at v in a gauge-invariant fashion. We obtain a gauge-invariant function $T_{\gamma,\mathbf{j},\mathbf{c}}(A)$ called a spin-network function. It turns out that, given $\gamma, \mathbf{j}$, there are only a finite number of linearly independent $\mathbf{c}$ compatible with $\gamma, \mathbf{j}$. Moreover, one can show that spin-network functions defined on different graphs are orthogonal and even on the same graph they are orthogonal whenever the vectors $\mathbf{j}$ do not equal each other. It follows that there is an orthonormal basis

$$\langle T_{\gamma,\mathbf{j},\mathbf{c}}, T_{\gamma',\mathbf{j}',\mathbf{c}'} \rangle = \delta_{\gamma,\gamma'} \delta_{\mathbf{j},\mathbf{j}'} \delta_{\mathbf{c},\mathbf{c}'} .$$

Now it is obvious that the operator $\hat{V}(R)$ leaves the finite-dimensional vector space $U_{\gamma,\mathbf{j}}$ spanned by spin-network states compatible with $\gamma, \mathbf{j}$ invariant. The matrix

$$(V(R)_{\gamma\mathbf{j}})_{\mathbf{c},\mathbf{c}'} := \langle T_{\gamma,\mathbf{j},\mathbf{c}} | \hat{V}(R) | T_{\gamma,\mathbf{j},\mathbf{c}'} \rangle \tag{IV.1}$$

is therefore finite dimensional, positive semi-definite, and symmetric. The task of computing its eigenvalues therefore becomes a problem in linear algebra!

Next, since from (III.18)

$$\hat{V}(R)_\gamma = l_p^3 \sum_{v \in V(\gamma) \cap R} \hat{V}_{v,\gamma}, \tag{IV.2}$$

and since $\hat{V}_{v,\gamma}$ involves only those $e \in E(\gamma)$ with $v \in e$, we find that $\hat{V}_{v,\gamma}$ can only change the entry c_v in $\mathbf{c}$. In other words, $[\hat{V}_{v,\gamma}, \hat{V}_{v',\gamma}] = 0$ and each $\hat{V}_{v,\gamma}$ can be diagonalized separately.

Finally, since the spins j_e only take discrete values, it follows that $\mathscr{H}_\gamma$ has a countable basis and the spectrum that $\hat{V}(R)$ attains on D_γ is therefore pure point. Let us check whether this is the complete spectrum. Assume it was not and let $\hat{P}$ be the spectral projection on the rest of the spectrum [the existence of the spectral projections relies on the fact that $\hat{V}(R)$ is self-adjoint and not only symmetric]. It follows that $u=\hat{P}v$ is orthogonal to D_γ where v is any vector in $\mathscr{H}_\gamma$. But D_γ is dense in $\mathscr{H}_\gamma$ and so we find for every $\epsilon>0$ a $\phi\in D_\gamma$ with $\|u-\phi\|<\epsilon$. Now we have from orthogonality $\epsilon^2>\|u-\phi\|^2=\|u\|^2+\|\phi\|^2>\|u\|^2$ and so $u=0$. This shows that the complete spectrum is already attained on D_γ. It is purely discrete as well in the physical sense that it is attained on a countable basis so that the eigenvalues only comprise a countable set. In a mathematical sense one would need to check that there are no accumulation points and no eigenvalues of infinite multiplicity. To settle this question it would be enough to show that the formula for the eigenvalue is an unbounded function of $j:=\Sigma_{e\in E(\gamma)}j_e$ because every countably infinite set of j's corresponding to different choices of **j** diverges. This is one possible future application of the explicit matrix element formulas which we derive in the next section.

Last, we wish to show that the volume operators are anomaly free (given the fact that we have largely adapted our regularization to a graph, this statement is far from trivial). By this we mean the following: given any two open sets $R_1,R_2\subset\Sigma$ we have vanishing Poisson brackets $\{V(R_1),V(R_2)\}=0$ because the functionals $V(R)$ depend on the momentum variable $E^a_i(x)$ only. Now, given a function f cylindrical with respect to a graph γ, it is not at all obvious any more that $[\hat{V}(R_1),\hat{V}(R_2)]f=0$ for any such f. Fortunately, given the above characterization of the spectrum, the commutator can be easily proved to vanish on cylindrical functions. To see this, note that the above results imply that if we choose any region $R(\gamma)$ such that $\gamma\subset R(\gamma)$, then there exists an eigenbasis of $\mathscr{H}_\gamma$ of $\hat{V}(R(\gamma))$. Now consider any region R. Since all regions are open by construction, all regions fall into equivalence classes with respect to γ: R,R' are equivalent if they contain the same vertices of γ (any vertex either has a neighborhood which lies completely inside R or it lies outside). Therefore any two $\hat{V}(R),\hat{V}(R')$ differ at most by some of the $\hat{V}_{v,\gamma}$, all of which are contained in the expression for $\hat{V}(R(\gamma))$. Since the $\hat{V}_{v,\gamma}$ commute, the eigenbasis of $\hat{V}(R(\gamma)$ is a simultaneous eigenbasis of all $\hat{V}_{v,\gamma}$ for all $v\in V(\gamma)$ and so this eigenbasis is a simultaneous eigenbasis of all $\hat{V}(R)_\gamma$. Since all $\mathscr{H}_\gamma$ are orthogonal, we have a simultaneous eigenbasis for all $\hat{V}(R)$.

While it is in general not enough to verify that two self-adjoint, unbounded operators commute on a dense domain (rather, by definition, we have to check that the associated spectral projections commute), in our case we are done because the spectral projections *are* the projections on the various D_γ because the point spectrum is already the complete spectrum. Thus we have verified that the commutator algebra mirrors the classical Poisson algebra.

V. THE COMPLETE SET OF MATRIX ELEMENTS

In this final section we are going to compute (IV.1) in a spin-network basis for each $\gamma,\mathbf{j},\mathbf{c},\mathbf{c}',R$. The advantage of our approach as compared to the ones proposed already in the literature[6] will be obvious: in Ref. 6 one works with an overcomplete and nonorthogonal set of states. This prevents one from giving a unified treatment as it is done here and only allowed one to find a formula for the eigenvalues in a few simple cases. More generally, one has to check all the time the linear independence of the states and can compute the matrix elements only in a case-by-case analysis.

In the next subsection we show in more detail a result which follows immediately from Refs. 1, 5, and 10: namely, that the problem of computing the matrix elements of the volume operator reduces to that of computing the matrix elements of a homogeneous polynomial of degree three of spin operators for a spin system.

This opens access to powerful techniques well known from the quantum theory of angular momentum for spin systems. Then we will show how the actual computation can be done by viewing the problem as a task in spin recoupling theory and that all the matrix elements can be written as algebraic functions of $6j$-symbols for which a closed formula exists (Racah formula). Finally we will diagonalize the volume operator for an infinite number of (very special) graphs.

A. The left regular representation and the spin-network representation

Consider a spin-network state $T_{\gamma,\mathbf{j},\mathbf{c}}(A)$. This is a state for the connection representation $L_2(\overline{\mathscr{A}/\mathscr{G}},d\mu_0)$ in the sense that the (equivalence classes modulo gauge transformations of) connection operators are diagonal in this representation. Notice that a spin-network state is gauge invariant so that it can be viewed alternatively as a function of $[A]$, the gauge equivalence class of A. Here we will use the fact that $d\mu_0$ extends to $\overline{\mathscr{A}}$ (Ref. 26) and reduces to the measure indicated in the Introduction when integrating gauge-invariant functions. Consider the usual Dirac (generalized) states δ_A defined by $\overline{\delta_A}(A'):=\delta_{\mu_0}(A,A')$ where δ_{μ_0} is the δ distribution with respect to μ_0 (Ref. 13) and $A,A'\in\overline{\mathscr{A}}$. We define the *spin-network representation* by

$$\delta_A(T_{\gamma,\mathbf{j},\mathbf{c}}):=\int_{\overline{\mathscr{A}}} d\mu_0(A')\,\overline{\delta_A(A')}\,T_{\gamma,\mathbf{j},\mathbf{c}}(A')=T_{\gamma,\mathbf{j},\mathbf{c}}(A)=:\langle A|\gamma,\mathbf{j},\mathbf{c}\rangle. \tag{V.1}$$

So far this is only notation. The advantage of this notion will become transparent upon proving that the states $|\gamma,\mathbf{j},\mathbf{c}\rangle$ are nothing else than the usual angular momentum states associated with an abstract spin system.

To see this, let v be a vertex of γ and $e_1,\ldots,e_n$ be the edges of γ incident at v, and we have chosen orientations such that all of them are outgoing at v. The spin-network state $T_{\gamma,\mathbf{j},\mathbf{c}}(A)$ can be written

$$T_{\gamma,\mathbf{j},\mathbf{c}}(A)=(c_v)_{m_1\ldots m_n}\prod_{i=1}^{n}(\pi_{j_{e_i}}(h_{e_i}(A)))_{m_i m_i'}(M_v)_{m_1'\ldots m_n'}=\mathrm{tr}\left(c_v\cdot\bigotimes_{i=1}^{n}\pi_{j_{e_i}}(h_{e_i}(A))\cdot M_v\right), \tag{V.2}$$

where c_v is the vertex contractor corresponding to v which contracts the indices corresponding to the starting points of the e_i and M_v contracts the indices corresponding to the endpoints (recall that the holonomy is a path-ordered exponential with the smallest parameter value to the left). Note that c_v is $\mathbb{C}$-valued while M_v still depends on the rest of the graph.

Consider a local gauge transformation at v, that is, $h_{e_i}\to g h_{e_i}$ for some $g\in\mathrm{SU}(2)$. Then (V.2) becomes $\mathrm{tr}(c_v\cdot\otimes_{i=1}^{n}\pi_{j_{e_i}}(g)\cdot\otimes_{i=1}^{n}\pi_{j_{e_i}}(h_{e_i}(A))\cdot M_v)$. In order for this to be gauge invariant, notice that the tensor product representation $\otimes_i\pi_{j_{e_i}}(g)$ can be written as a direct sum of orthogonal irreducible representations $\pi_j(g)$, and so we just need to choose c_v to be proportional to one of the various equivalent (but orthogonal) trivial representations $\pi_0(g)=\pi_0(1)$ contained in that decomposition (see Ref. 13 for more details) in order for (V.2) to be gauge invariant. This is precisely what a spin-network state is.

The point is now that

$$\prod_{i=1}^{n}(\pi_{j_{e_i}}(h_{e_i}(A)))_{m_i m_i'}=:\langle A|(m_1,j_1),\ldots,(m_n,j_n);m_1',\ldots,m_n'\rangle \tag{V.3}$$

transforms *precisely* as the usual angular momentum eigenstates

$$|(m_1,j_1),\ldots,(m_n,j_n);k_1,\ldots,k_m\rangle:=\prod_{i=1}^{n}|m_i,j_i;k_1,\ldots,k_m\rangle$$

of an abstract spin system of n spins $j_1,\ldots,j_n$ under rotations (see, e.g., Ref. 31). This is because $|(m,j);k\rangle$ transforms as $\hat{U}(g)|(m,j);k\rangle=\pi_j(g)_{m,m'}|(m',j);k\rangle$ by definition, where $\hat{U}(g)$ is a unitary representation of the rotation group. Here $k_1,\ldots,k_m$ are some additional quantum numbers of observables commuting with the operators $\hat{J}_i$ which in our case coincide with the $m_1',\ldots,m_n'$. In other words, the left regular representation $L_g f(h_e)=f(gh_e)$ that we are dealing with in the connection representation can be seen to be equivalent with the representation $\hat{U}(g)$ of the spin-network representation.

Let us now couple the spins $j_1,\ldots,j_n$ to resulting spin $j=0$ and denote a particular state with zero angular momentum by $|0\rangle$. We claim that up to a constant

$$(c_v)_{m_1..m_n}=\langle (m_1,j_1),\ldots,(m_n,j_n);k_1,\ldots,k_m|0\rangle, \tag{V.4}$$

where the inner product is now to be understood in the angular momentum Hilbert space of the abstract spin system. In other words, the c_v are just usual Clebsh–Gordan coefficients. The proof is easy. We have

$$\begin{aligned}(c_v)_{m_1\cdots m_n}\prod_{i=1}^{n}\pi_{j_i}(g)_{m_i m_i'}(g)&=\langle \hat{U}(g)((m_1',j_1),\ldots,(m_n',j_n);k_1,\ldots,k_m)|0\rangle\\ &=\langle (m_1',j_1),\ldots,(m_n',j_n);k_1,\ldots,k_m|\hat{U}(g^{-1})0\rangle=(c_v)_{m_1'\cdots m_n'},\end{aligned}$$

where we have used unitarity of the group and of the representation and that $\hat{U}(g)|0\rangle=|0\rangle$ is rotation invariant. The factor of proportionality is seen to equal unity because $c_v^2=1$ and since $\Sigma_{m_1\cdots m_n}|(m_1,j_1)\cdots(m_n,j_n);k_1\cdots k_m\rangle\langle(m_1,j_1)\cdots(m_n,j_n);k_1\cdots k_m|=1$ due to completeness and the Wigner–Eckart theorem.[31]

Next, consider the self-adjoint right invariant vector fields $Y_e:=-i/2X_e$. We have $[X^i,X^j]=-2\epsilon^{ijk}X^k$ and so $[Y^i,Y^j]=i\epsilon^{ijk}Y^k$, which is the usual angular momentum algebra. It follows that [we use a dual interpretation of the SU(2) indices $A,B,C,\ldots$, which we may choose to take values $\pm\frac{1}{2}$ and so can be interpreted to be the eigenvalues of the three-component of angular momentum]

$$\begin{aligned}\langle j=\tfrac{1}{2},C|Y_e^j h_e(A)_{CD}&=\langle j=\tfrac{1}{2},C|(-i/2)(\tau_j h_e(A))_{CD}\\ &=(i/2)\langle(\tau_j)_{EC}(\tfrac{1}{2},C)|h_e(A))_{ED}\\ &=(i/2)\,\frac{d}{dt_{t=0}}\langle\hat{U}(\exp(t\tau_j))(\tfrac{1}{2},E)|h_e(A))_{ED}\\ &=\langle\tfrac{1}{2},E|J_j h_e(A)_{ED},\end{aligned} \tag{V.5}$$

and so by multilinearity of the tensor product representation we find due to self-adjointness of J_I (set $Y_I:=Y_{e_I}$)

$$\begin{aligned}Y_I^j T_{\gamma,\mathbf{j},\mathbf{c}}(A)&=\frac{i}{2}\,\frac{d}{dt_{t=0}}\langle A|\gamma,\mathbf{j},\mathbf{m};\mathbf{m}'\rangle\langle\mathbf{j},\mathbf{m}|\hat{U}(\exp(-t\tau_j))_I 0\rangle(M_v)_{m_1'\cdots m_n'}\\ &=\langle A|\gamma,\mathbf{j}_v,\mathbf{m};\mathbf{m}'\rangle\langle\mathbf{j}_v,\mathbf{m}|J_I^j 0\rangle(M_v)_{m_1'\cdots m_n'},\end{aligned} \tag{V.6}$$

where $\mathbf{j}_v$ comprises the components of $\mathbf{j}$ which correspond to the edges incident at v.

Since spin-network states are thus characterized by the various zero-angular momentum eigenstates $|0\rangle$ of the abstract spin system, there is a one-to-one correspondence between the possible choices of c_v and $|0\rangle$. So let $|0\rangle$ correspond to c_v and $|0'\rangle$ to c_v'. Let $p(\mathbf{Y})$ be any polynomial of the Y_I^i corresponding to the vertex v. Then we have for its matrix elements

$$\begin{aligned}\langle T_{\gamma,\mathbf{j},\mathbf{c}}|p(\mathbf{Y})|T_{\gamma,\mathbf{j},\mathbf{c}'}\rangle&=\int_{\mathcal{M}\mathcal{G}}d\mu_0(A)\overline{\langle\mathbf{j}_v,\mathbf{m}|0\rangle\langle A|\gamma,\mathbf{j}_v,\mathbf{m};\mathbf{m}'\rangle(M_v)_{m_1'\cdots m_n'}}\langle\mathbf{j}_v,\overline{\mathbf{m}}|p(\mathbf{J})0'\rangle\\ &\quad\times\langle A|\gamma,\mathbf{j}_v,\overline{\mathbf{m}};\overline{\mathbf{m}}'\rangle(M_v)_{\overline{m}_1'\cdots\overline{m}_n'}\\ &=\overline{\langle\mathbf{j}_v,\mathbf{m}|0\rangle}\langle\mathbf{j}_v,\mathbf{m}|p(\mathbf{J})^\dagger 0'\rangle\prod_{v'\neq v}\delta_{c_{v'},c_{v'}'}\\ &=\langle 0|p(\mathbf{J})^\dagger 0'\rangle\prod_{v'\neq v}\delta_{c_{v'},c_{v'}'}=\langle p(\mathbf{J})0|0'\rangle\prod_{v'\neq v}\delta_{c_{v'},c_{v'}'}.\end{aligned} \tag{V.7}$$

Here we have used (V.6) in the first step, the orthonormality of the spin-network states in the second step, and in the last step we have again recalled that $\Sigma_{\mathbf{m}}|\mathbf{j}_v,\mathbf{m}\rangle\langle\mathbf{j}_v,\mathbf{m}|=1$ is the identity on the abstract angular momentum Hilbert space labeled by the spins $\mathbf{j}_v$ [completeness, notice that we sum over $\mathbf{m}$ in (V.7)].

Expression (V.7) is a huge simplification: in order to compute the matrix elements of the volume operator it is enough to compute the matrix elements of the angular momentum operator polynomial $\epsilon_{ijk}J^i_e J^j_{e'} J^k_{e''}$ between states that have zero angular momentum coming from $\mathbf{j}_v$. The next subsection is devoted to this task.

B. Recoupling theory for *n* angular momenta

The first task is to label all the linearly independent states $|0\rangle$ which correspond to the coupling of n angular momenta $j_1,\ldots,j_n$ to zero angular momentum. This is a well-known problem in the quantum theory of angular momentum and runs under the name recoupling theory. We review here what we need for our computation.

The states $|\mathbf{j},\mathbf{m}\rangle$ are a complete basis for a spin system consisting of n angular momenta. In this basis the $2n$ mutually commuting operators $(J^i_I)^2, J^3_I$ are diagonal with eigenvalues $j_I(j_I+1), m_I$.

We are obviously interested in a basis in which the square of the operator of total angular momentum $J:=J_1+\cdots+J_n$ and its three-component are diagonal with eigenvalues $j(j+1), m$. In order to label that basis we need $2(n-1)$ more quantum numbers of mutually commuting operators and commuting with J^2, J_3. The n operators J^2_I are readily verified to satisfy these conditions but we need $n-2$ more. In order to motivate our choice notice that $\hat{V}_{v,\gamma}$ is the square root of the modulus of $i(\frac{3}{2})^3\frac{1}{4}\Sigma_{I<J<K}\epsilon(e_I,e_J,e_K)q_{IJK}$ where

$$\tfrac{1}{4}q_{IJK}:=\tfrac{1}{8}\epsilon_{ijk}X^i_I X^j_J X^k_K=[Y^i_I Y^i_J,\ Y^j_J Y^j_K]=\tfrac{1}{4}[(Y^i_{IJ})^2,(Y^j_{JK})^2],$$

where $Y_{IJ}:=Y_I+Y_J$. Here we have used total antisymmetry of $\epsilon(e_I,e_J,e_K)$, $\hat{q}_{IJK}$ and the fact that $[X_I,\ X_J]=0$, $I\neq J$, in order to get rid of 3! and in order to sum only over $I<J<K$. This observation {i.e., the fact that the third-order polynomial $\mathrm{tr}(X_I[X_J,\ X_K])$ in the X can be written as a commutator of two second-order polynomials $\mathrm{tr}(X_I X_J)$ which are gauge invariant, also made independently by Abhay Ashtekar and Jurek Lewandowski} urges us to study the operator

$$\hat{q}_{IJK}:=[J^2_{IJ},\ J^2_{JK}],\quad \text{where}\quad J_{IJ}=J_I+J_J, \tag{V.8}$$

and thus it would be convenient if we could work in a basis in which the J_{IJ} also were diagonal. Clearly we cannot achieve this for all $n(n-1)/2$ operators J_{IJ}, but we can work in different bases adapted to $\hat{q}_{IJK}$ for a particular triple $I<J<K$ and expand them in terms of each other.

We will develop a theory of recoupling schemes here which is somewhat different from the one encountered in the standard literature and we also consider arbitrary n. The following paragraphs will therefore be somewhat detailed.

Definition V.1: A coupling scheme based on a pair $1\leqslant I<J\leqslant n$ *is an orthonormal basis* $|\mathbf{g}(IJ),\mathbf{j},j,m\rangle$, *diagonalizing besides* J^2, J_3, J^2_I, $I=1,\ldots,n$ *the squares of the additional* $n-2$ *operators* $G_2,\ldots,G_{n-1}$, *where*

$$G_1:=J_I,\quad G_2:=G_1+J_J,\quad G_3:=G_2+J_1,\quad G_4:=G_3+J_2,\ldots,$$

$$G_{I+1}:=G_I+J_{I-1},\quad G_{I+2}:=G_{I+1}+J_{I+1},\quad G_{I+3}:=G_{I+2}+J_{I+2},\ldots,$$

$$G_J:=G_{J-1}+J_{J-1},\quad G_{J+1}:=G_J+J_{J+1},\quad G_{J+2}:=G_{J+1}+J_{J+2},\ldots,$$

$$G_n:=G_{n-1}+J_n=J.$$

The vector $\mathbf{g}(IJ):=(g_2(IJ),\ldots,g_{n-1}(IJ))$ *denotes the eigenvalues of the squares of* $G_2,\ldots,G_{n-1}$. *We will call the coupling scheme based on the pair (12) the standard basis.*

Let us check that the G_I satisfy the angular momentum algebra and that the squares of the operators J, J_I, G_I and the operator J^3 are mutually commuting: for this it will be enough to check that the operators for the standard basis are mutually commuting since there is a relabeling of indices such that any coupling scheme becomes the standard basis. Since $[J_I,\ J_J]=0$, $I\neq J$, we find $[G^i_I,\ G^j_I]=\Sigma^P_{J=1}[J^i_J,\ J^j_J]=i\epsilon_{ijk}\Sigma^I_{J=1}J_J=i\epsilon_{ijk}G^k_I$, which implies that G^2_I has spectrum $g_I(g_I+1)$ with g_I integral or half-integral.

Let $I<J$. Then we can write $G_J=G_I+(J_{I+1}+\cdots+J_J)=:G_I+G'_I$ so that $[G_I,\ G'_I]=0$ because G_I,G'_I involve disjoint sets of J_K's and so $[G_J^2,\ G_I^2]=[G_I^2+(G'_I)^2,\ G_I^2]+2(G'_I)^i\{G_I^j[G_I^i,G_I^j]+[G_I^i,G_I^j]G_I^j\}=2i(G'_I)\epsilon_{ijk}\{G_I^jG_I^k+G_I^kG_I^j\}=0$. The same argument can be used for the other commutators.

The virtue of this definition is the following: we have now a neat labeling of the vectors $|0\rangle$ referred to in the previous subsection. We can choose them to be the vectors of the standard basis corresponding to $j=0$, that is, they are just the vectors $|\mathbf{g}(12),\mathbf{j},j=0,m=0\rangle$ ($m=0$ is forced by $j=0$) and the different vertex contractors just correspond to the various possible values of the intermediate (or so-called recoupling) vectors $G_2,\ldots,G_{n-1}$. We will therefore compute the matrix elements of $\hat{q}_{IJK}$ in this basis. In what follows we will drop the indices $\mathbf{j},j=0$, $m=0$ since they are the same in every coupling scheme, and, since $\hat{q}_{IJK}$ commutes with $J^2,J_I^2,(J)^3$, matrix elements between vectors with different values of these operators therefore vanish anyway. We will also write $\mathbf{g}$ instead of $\mathbf{g}(12)$.

After these long preparations we are now in the position to complete the derivation. We use completeness of the bases $|\mathbf{g}(IJ),\mathbf{j},j,m\rangle$ and have

$$\begin{aligned}\langle\mathbf{g}|\hat{q}_{IJK}|\mathbf{g}'\rangle&=\sum_{\mathbf{g}(IJ)}g_2(IJ)(g_2(IJ)+1)[\langle\mathbf{g}|\mathbf{g}(IJ)\rangle\langle(\mathbf{g}_{IJ})|J_{JK}^2|\mathbf{g}'\rangle-\langle\mathbf{g}|J_{JK}^2|\mathbf{g}_{IJ}\rangle\langle\mathbf{g}(IJ)|\mathbf{g}'\rangle]\\&=\sum_{\mathbf{g}(IJ),\mathbf{g}(JK),\mathbf{g}''}g_2(IJ)(g_2(IJ)+1)g_2(JK)(g_2(JK)+1)\langle\mathbf{g}(IJ)|\mathbf{g}''\rangle\langle\mathbf{g}(JK)|\mathbf{g}''\rangle\\&\quad\times[\langle\mathbf{g}(IJ)|\mathbf{g}\rangle\langle\mathbf{g}(JK)|\mathbf{g}'\rangle-\langle\mathbf{g}(JK)|\mathbf{g}\rangle\langle\mathbf{g}(IJ)|\mathbf{g}'\rangle],\end{aligned}\tag{V.9}$$

where we have used that the so-called $3nj$-*symbols* $\langle\mathbf{g}(IJ)|\mathbf{g}(JK)\rangle$ are real [also $\langle\mathbf{g}(IJ)|\mathbf{g}(KL)\rangle$ for any $I<J,K<L$ are real]. To see this, insert a unit in terms of the basis $|\mathbf{j},\mathbf{m}\rangle$ which allows us to write the $3nj$-symbol in terms of Clebsh–Gordan coefficients which are real.[31] It is understood that in (V.9) and in what follows the summation extends only over the allowed values of $\mathbf{g}$, etc., which are purely fixed by the values of $j_1,\ldots,j_n$.

The task left is to write the $3nj$-symbols of the form $\langle\mathbf{g}(IJ)|\mathbf{g}'\rangle$ in terms of known quantities. We will derive below a formula which expresses it in terms of the $6j$-symbol for which a closed formula (the Racah formula[31]) exists.

We will denote $\mathbf{g}(IJ)=(g_2(j_I,j_J),g_3(g_2,j_1),\ldots,g_{I+1}(g_I,j_{I-1}),g_{I+2}(g_{I+1},j_{I+1}),\ldots,g_J(g_{J-1},j_{J-1}),g_{J+1}(g_J,j_{J+1}),\ldots,g_{n-1}(g_{n-2},j_{n-1}))$, where the notation $j''(j,j')$ means that we couple J,J' to resulting spin j'' of J'', thereby keeping explicit track of the coupling scheme. The following two elementary lemmata will be crucial in computing the $3nj$-symbols.

Lemma V.1:

$$\begin{aligned}\langle\mathbf{g}(IJ)|\mathbf{g}'\rangle=\langle g_2(j_I,j_J),g_3(g_2,j_1),\ldots,g_{I+1}(g_I,j_{I-1}),g_{I+2}(g_{I+1},j_{I+1}),\ldots,g_J(g_{J-1},j_{J-1})|\\\times|g'_2(j_1,j_2),g'_3(g'_2,j_3),\ldots,g'_J(g'_{J-1},j_J)\rangle\delta_{g,J,g'_J}\cdots\delta_{g_{n-1},g'_{n-1}},\end{aligned}$$

where on the right-hand side we consider a spin system consisting of spins $J_1,\ldots,J_J$ and total spin quantum number g_J.

Proof: Since $G_J(12)^2,\ldots,G_{n-1}(12)^2$ are diagonal on both vectors it is immediately clear that the $3nj$-symbol must be proportional to the Kronecker delta δ_{g_K,g'_K} for $K=J,\ldots,n-1$. To get the factor of proportionality we expand both vectors in terms of the basis $|\mathbf{j},\mathbf{m}\rangle$ and Clebsh–Gordan coefficients. To do this, note that by definition

$$\begin{aligned}&\langle g_2(j_1,j_2),\ldots,g_{n-2}(g_{n-3},j_{n-2});g_{n-1}m\\&\quad-m_n,j_nm_n|g_2(j_1,j_2),\ldots,g_{n-2}(g_{n-3},j_{n-2}),g_{n-1}(g_{n-2},j_{n-1}),j(g_{n-1},j_n),m\rangle\\&\quad=\langle g_{n-1}(g_{n-2},j_{n-1})m-m_n,j_nm_n|j,m;g_{n-1},j_n\rangle,\end{aligned}$$

where on the right-hand side we have the standard Clebsh–Gordan coefficient. Here

$$|g_2(j_1,j_2),\ldots,g_{n-2}(g_{n-3},j_{n-2});g_{n-1}m-m_n,j_nm_n\rangle$$
$$:=|g_2(j_1,j_2),\ldots,g_{n-2}(g_{n-3},j_{n-2}),g_{n-1}(g_{n-2},j_{n-1}),m-m_n\rangle\otimes|j_nm_n\rangle.$$

We can now iterate this procedure and obtain

$$|g'_2(j_1,j_2),\ldots,g'_{n-2}(g'_{n-3},j_{n-2}),g'_{n-1}(g'_{n-2},j_{n-1}),j(g'_{n-1},j_n),m\rangle$$
$$=\sum_{m_J+\cdots+m_n=m}\langle g'_{n-1}m-m_n,j_nm_n|jm\rangle\langle g'_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g'_{n-1}m-m_n\rangle\cdots$$
$$\langle g'_Jm_J,j_{J+1}m_{J+1}|g'_{J+1},m_J+m_{J+1}\rangle|g'_2(j_1,j_2),\ldots,g'_J(g'_{J-1},j_J),m_J\rangle\otimes|j_{J+1},m_{J+1}\rangle\otimes\cdots$$
$$\otimes|j_n,m_n\rangle.$$

Similarly we have

$$|g_2(j_I,j_J),\ldots,g_{n-2}(g_{n-3},j_{n-2}),g_{n-1}(g_{n-2},j_{n-1}),j(g_{n-1},j_n),m\rangle$$
$$=\sum_{m_J+\cdots+m_n=m}\langle g_{n-1}m-m_nj_nm_n|jm\rangle\langle g_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g_{n-1}m-m_n\rangle\cdots$$
$$\langle g_Jm_J,j_{J+1}m_{J+1}|g_{J+1},m_J+m_{J+1}\rangle|g_2(j_I,j_J),\ldots,g_J(g_{j-1},j_{J-1}),m_J\rangle$$
$$\otimes|j_{J+1},m_{J+1}\rangle\otimes\cdots\otimes|j_n,m_n\rangle.$$

Thus we have the inner product

$$\begin{aligned}\langle\mathbf{g}(IJ)|\mathbf{g}'\rangle=&\sum_{m_J+\cdots+m_n=m}\langle g_2(j_I,j_J),\ldots,g_J(g_{J-1},j_{J-1}),m_J|g'_2(j_1,j_2),\ldots,g'_J(g'_{J-1},j_J),m_J\rangle\\
&\times\langle g'_{n-1}m-m_n,j_nm_n|jm\rangle\langle g'_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g'_{n-1}m-m_n\rangle\cdots\\
&\langle g'_Jm_J,j_{J+1}m_{J+1}|g'_{J+1},m_J+m_{J+1}\rangle\\
&\times\langle g_{n-1}m-m_n,j_nm_n|jm\rangle\langle g_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g_{n-1}m-m_n\rangle\cdots\\
&\langle g_Jm_J,j_{J+1}m_{J+1}|g_{J+1},m_J+m_{J+1}\rangle\\
=&\langle g_2(j_I,j_J),\ldots,g_J(g_{J-1},j_{J-1}),\tilde{m}|g'_2(j_1,j_2),\ldots,g'_J(g'_{J-1},j_J),\tilde{m}\rangle\\
&\times\sum_{m_{J+1},\ldots,m_n}\langle g'_{n-1}m-m_n,j_nm_n|jm\rangle\\
&\times\langle g'_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g'_{n-1}m-m_n\rangle\cdots\\
&\langle g'_Jm_J,j_{J+1}m_{J+1}|g'_{J+1},m_J+m_{J+1}\rangle\langle g_{n-1}m-m_n,j_nm_n|jm\rangle\\
&\times\langle g_{n-2}m-m_n-m_{n-1},j_{n-1}m_{n-1}|g_{n-1}m-m_n\rangle\cdots\\
&\langle g_Jm_J,j_{J+1}m_{J+1}|g_{J+1},m_J+m_{J+1}\rangle\\
=&\langle g_2(j_I,j_J),\ldots,g_J(g_{J-1},j_{J-1}),\tilde{m}|g'_2(j_1,j_2),\ldots,g'_J(g'_{J-1},j_J),\tilde{m}\rangle\delta_{g_J,g'_J}\cdots\delta_{g_{n-1},g'_{n-1}},\end{aligned}$$

where in the second step we have used the fact that the $3nj$-symbols are independent of m. [This is an extension of the usual Wigner–Eckart theorem which we shortly outline here: Let us expand $|\mathbf{g}(IJ)\rangle=\Sigma_{\mathbf{g}(KL)}\langle\mathbf{g}(KL);j,m|\mathbf{g}(IJ);j,m\rangle|\mathbf{g}(KL);j,m\rangle$ where we assume that $m<j$ w.l.g. Now apply the ladder operator $J^\dagger$ to both sides of this equation. We find $J^\dagger|\mathbf{g}(MN);j,m\rangle=c(j,m)|\mathbf{g}(MN);j,m+1\rangle$ for any $K<L$ where the coefficient only depends on j,m. Comparing coefficients we find $\langle\mathbf{g}(KL);j,m|\mathbf{g}(IJ);j,m\rangle=\langle\mathbf{g}(KL);j,m+1|\mathbf{g}(IJ);j,m+1\rangle$, that is, the $3nj$-symbols are rotation invariant. □

Thus, $\tilde{m}$ is arbitrary and in the last step we have used the sum rule for the Clebsh–Gordan coefficients, that is, $\Sigma_{m_1+m_2=m}(\langle j_1m_1,j_2m_2|jm\rangle)^2=1$. □

Lemma V.2:

$$\langle g_2(j_1,j_2),\ldots,g_K(g_{K-1},j_K),g_{K+1}(g_K,j_{K+1},g_{K+2}(g_{K+1},j_{K+2});m|$$
$$\times|g_2(j_1,j_2),\ldots,g_K(g_{K-1},j_K),g'_{K+1}(g_K,j_{K+2}),g'_{K+2}(g'_{K+1},j_{K+1});m\rangle$$
$$=\langle g_{K+1}(g_K,j_{K+1}),g_{K+2}(g_{K+1},j_{K+2});m|g'_{K+1}(g_K,j_{K+2}),g_{K+2}(g'_{K+1},j_3);m\rangle.$$

Proof: Again we expand

$$|g_2,\ldots,g_{K=2};m\rangle=\sum_{m_{K+1},m_{K+2}}\langle g_K m-m_{K+1}-m_{K+2},j_{K+1}m_{K+1}|g_{K+1}m-m_{K+2}\rangle$$
$$\times\langle g_{K+2}m-m_{K+2},j_{K+2}m_{K+2}|g_{K+2}m-m_{K+2}\rangle$$
$$\times|g_2,\ldots,g_k;m-m_{K+1}-m_{K+2}\rangle\otimes|j_{K+1}m_{K+1},j_{K+2}m_{K+2}\rangle$$

and

$$|g_2,\ldots,g_K,g'_{K+1},g_{K+2};m\rangle=\sum_{m_{K+1},m_{K+2}}\langle g_K m-m_{K+1}-m_{K+2},j_{K+2}m_{K+2}|g'_{K+1}m-m_{K+1}\rangle$$
$$\times\langle g_{K+2}m-m_{K+1},j_{K+1}m_{K+1}|g_{K+2}m-m_{K+1}\rangle|g_2,\ldots,g_k;m$$
$$-m_{K+1}-m_{K+2}\rangle\otimes|j_{K+1}m_{K+1},j_{K+2}m_{K+2}\rangle.$$

We find for the inner product of these states due to the normalization just

$$\sum_{m_{K+1},m_{K+2}}\langle g_K m-m_{K+1}-m_{K+2},j_{K+1}m_{K+1}|g_{K+1}m-m_{K+2}\rangle\langle g_{K+2}m$$
$$-m_{K+2},j_{K+2}m_{K+2}|g_{K+2}m-m_{K+2}\rangle\langle g_K m-m_{K+1}-m_{K+2},j_{K+2}m_{K+2}|g'_{K+1}m-m_{K+1}\rangle$$
$$\times\langle g_{K+2}m-m_{K+1},j_{K+1}m_{K+1}|g_{K+2}m-m_{K+1}\rangle$$
$$\equiv\langle g_{K+1}(g_K,j_{K+1}),g_{K+2}(g_{K+1},j_{K+2});m|g'_{K+1}(g_K,j_{K+2}),g_{K+2}(g'_{K+1},j_3);m\rangle$$

as claimed. □

We are now ready to reduce out the general $3nj$-symbol. We assume that $g_K=g'_K$, $K=J,\ldots,n-1$, as otherwise we get zero by Lemma V.1. Upon repeatedly using the completeness relations of the coupling scheme bases we first shift j_J and then j_I to the right in the matrix element

$$\langle\mathbf{g}(IJ),\mathbf{g}'\rangle=_{V.1}\langle g_2(j_I,j_J),g_3(g_2,j_1),\ldots,g_{I+1}(g_I,j_{I-1}),g_{I+2}(g_{I+1},j_{I+1}),\ldots,g_J(g_{J-1},j_{J-1})|$$
$$\times|g'_2(j_1,j_2),g'_3(g'_2,j_3),\ldots,g'_{J-1}(g'_{J-2},j_{J-1}),g_J(g'_{J-1},j_J)\rangle$$

$$=_{V.1}\sum_{h_2}\langle g_2(j_I,j_J),g_3(g_2,j_1),g_4,\ldots,g_J|h_2(j_I,j_1),g_3(h_2,j_J),g_4,\ldots,g_J\rangle$$
$$\times\langle h_2(j_I,j_1),g_3(h_2,j_J),g_4,\ldots,g_J|g'_2(j_1,j_2),g'_3(g'_2,j_3),\ldots,g'_{J-1},g_J\rangle$$

$$=_{V.1}\sum_{h_2}\langle g_2(j_I,j_J),g_3(g_2,j_1)|h_2(j_I,j_1),g_3(h_2,j_J)\rangle$$
$$\times\sum_{h_3}\langle h_2(j_I,j_1),g_3(h_2,j_J),g_4(g_3,j_2),g_5,\ldots,g_J$$
$$\times|h_2(j_I,j_1),h_3(h_2,j_2),g_4(h_3,j_J),g_5,\ldots,g_J\rangle$$
$$\times\langle h_2(j_I,j_1),h_3(h_2,j_2),g_4(h_3,j_J),g_5,\ldots,g_J|g'_2(j_1,j_2),g'_3(g'_2,j_3),\ldots,g'_{J-1},g_J\rangle$$

$$=_{V.1,V.2}\sum_{h_2}\langle g_2(j_I,j_J),g_3(g_2,j_1)|h_2(j_I,j_1),g_3(h_2,j_J)\rangle$$

$$
\begin{aligned}
&\times\sum_{h_3}\langle g_3(h_2,j_J),g_4(g_3,j_2)|h_3(h_2,j_2),g_4(h_3,j_J)\rangle\\
&\times\langle h_2(j_I,j_1),h_3(h_2,j_2),g_4(h_3,j_J),g_5,\ldots,g_J|g_2'(j_1,j_2),g_3'(g_2',j_3),\ldots,g_{J-1}',g_J\rangle\\
=&\cdots\\
=&\sum_{h_2}\langle g_2(j_I,j_J),g_3(g_2,j_1)|h_2(j_I,j_1),g_3(h_2,j_J)\rangle\\
&\times\sum_{h_3}\langle g_3(h_2,j_J),g_4(g_3,j_2)|h_3(h_2,j_2),g_4(h_3,j_J)\rangle\cdots\\
&\times\sum_{h_I}\langle g_I(h_{I-1},j_J),g_{I+1}(g_I,j_{I-1})|h_I(h_{I-1},j_{I-1}),g_{I+1}(h_I,j_J)\rangle\\
&\times\sum_{h_{I+1}}\langle g_{I+1}(h_I,j_J),g_{I+2}(g_{I+1},j_{I+1})|h_{I+1}(h_I,j_{I+1}),g_{I+2}(h_{I+1},j_J)\rangle\\
&\times\sum_{h_{I+2}}\langle g_{I+2}(h_{I+1},j_J),g_{I+3}(g_{I+2},j_{I+2})|h_{I+2}(h_{I+1},j_{I+2}),g_{I+3}(h_{I+2},j_J)\rangle\cdots\\
&\times\sum_{h_{J-1}}\langle g_{J-1}(h_{J-2},j_J),g_J(g_{J-1},j_{J-1})|h_{J-1}(H_{J-2},j_{J-1}),g_J(h_{J-1},j_J)\rangle\\
&\times\langle h_2(j_I,j_1)h_3(h_2,j_2),\ldots,h_I(h_{I-1},j_{I-1}),h_{I+1}(h_I,j_{I+1}),\ldots,h_{J-1}(h_{J-2},j_{J-1}),\\
&\times g_J(h_{J-1},j_J)||g_2'(j_1,j_2),\ldots,g_I'(g_{I-1}',j_I),g_{I+1}'(g_I',j_{I+1}),\ldots,g_J'(g_{J-1}',j_J)\rangle.
\end{aligned}
\tag{V.10}
$$

Now the last matrix element in the last line of (V.10) is different from zero only for $h_K=g_K'$, $K=I,\ldots,J-1$ by Lemma V.1 (we have $g_J=g_J'$ already) and in that case reduces by the same reasoning to

$$
\begin{aligned}
&\langle h_2(j_I,j_1)h_3(h_2,j_2),\ldots,h_{I-1}(h_{I-2},j_{I-2}),g_I(h_{I-1},j_{I-1})|g_2'(j_1,j_2),\ldots,g_I'(g_{I-1}',j_I)\rangle\\
&=\sum_{k_2}\langle h_2(j_I,j_1),h_3(h_2,j_2)|k_2(j_1,j_2),h_3(k_2,j_I)\rangle\\
&\times\sum_{k_3}\langle h_3(k_2,j_I),h_4(h_3,j_3)|k_3(k_2,j_3),h_3(k_3,j_I)\rangle\cdots\\
&\times\sum_{k_{I-1}}\langle h_{I-1}(k_{I-2},j_I),h_I(h_{I-1},j_{I-1})|k_{I-1}(k_{I-2},j_{I-1}),g_I'(k_{I-1},j_I)\rangle\\
&\times\langle k_2(j_1,j_2),k_3(k_2,j_3),\ldots,k_{I-1}(k_{I-2},j_{I-1}),g_I'(k_{I-1},j_I)\\
&\times|g_2'(j_1,j_2),g_3'(g_2',j_3),\ldots,g_I'(g_{I-1}',j_I)\rangle
\end{aligned}
\tag{V.11}
$$

and the last matrix element in the last line of (V.11) is just given by $\Pi_{K=2}^{I-1}\delta_{k_K,g_K'}$ so that the summation drops out. So, altogether, there are only the $I-3$ summation variables $h_2,\ldots,h_{I-1}$.

Equations (V.10) and (V.11) demonstrate that we manage to reduce the $3nj$-symbol under investigation to a polynomial of matrix elements of the structure

$$
\begin{aligned}
&\text{type I: } \langle j_{12}(j_1,j_2),j(j_{12},j_3)|j_{13}(j_1,j_3),j(j_{13},j_2)\rangle \text{ or}\\
&\text{type II: } \langle j_{12}(j_1,j_2),j(j_{12},j_3)|j_{23}(j_2,j_3),j(j_{23},j_1)\rangle.
\end{aligned}
\tag{V.12}
$$

In fact, all matrix elements in (V.10) are of type I and all matrix elements in (V.11) are of type I, except for the first factor after the first equality which is of type II.

The point is now that both matrix elements are related to the $6j$-symbol.[31] The $6j$-symbol is defined through the type II matrix element, namely

$$\langle j_{12}(j_1,j_2),j(j_{12},j_3)|j_{23}(j_2,j_3),j(j_{23},j_1)\rangle =: \sqrt{(2j_{12}+1)(2j_{23}+1)}(-1)^{j_1+j_2+j_3+j}\begin{Bmatrix} j_1 & j_2 & j_{12} \\ j_3 & j & j_{23}\end{Bmatrix}. \tag{V.13}$$

In order to relate the type I matrix element to the $6j$-symbol we need to use the following identity for the CG (Clebsh–Gordan) coefficients

$$\langle j_1m_1,j_2m_2|jm\rangle=(-1)^{j_1+j_2-j}\langle j_2m_2,j_1m_1|jm\rangle, \tag{V.14}$$

which is a consequence of the choice of phases for the CG coefficients in order to have them real valued.[31] Let us now set $j_1':=j_2$, $j_2':=j_1$, $j_3':=j_3$, $j_{12}':=j_{12}$, $j_{23}':=j_{13}$, and $j':=j$. Then we obtain

$$\langle j_{12}(j_1,j_2),j(j_{12},j_3)|j_{13}(j_1,j_3),j(j_{13},j_2)\rangle=\langle j_{12}'(j_2',j_1'),j'(j_{12}',j_3')|j_{23}'(j_2',j_3'),j'(j_{23}',j_1')\rangle,$$

which is almost the standard form except for the wrong order of j_1',j_2' in the first entry. Using the expansion

$$|j_{12}'(j_2',j_1'),j'(j_{12}',j_3');m'\rangle \sum_{m_1+m_2+m_3=m'}\langle j_2'm_2,j_1'm_1|j_{12}'m_1+m_2\rangle\langle j_{12}'m_1+m_2,j_3'm_3|j'm'\rangle|j_1'm_1,\ldots,j_3'm_3'\rangle$$

and (V.14) we therefore find

$$\langle j_{12}'(j_2',j_1'),j'(j_{12}',j_3')|j_{23}'(j_2',j_3'),j'(j_{23}',j_1')\rangle =(-1)^{j_1'+j_2'-j_{12}'}\langle j_{12}'(j_1',j_2'),j'(j_{12}',j_3')|j_{23}'(j_2',j_3'),j'(j_{23}',j_1')\rangle$$

and so we have for the type I matrix element

$$\langle j_{12}(j_1,j_2),j(j_{12},j_3)|j_{13}(j_1,j_3),j(j_{13},j_2)\rangle =(-1)^{j_1+j_2-j_{12}}\sqrt{(2j_{12}+1)(2j_{13}+1)}(-1)^{j_1+j_2+j_3+j}\begin{Bmatrix} j_2 & j_1 & j_{12} \\ j_3 & j & j_{13}\end{Bmatrix}. \tag{V.15}$$

Finally, for the benefit of the reader, we note the Racah formula,[31]

$$\begin{Bmatrix} j_1 & j_2 & j_{12} \\ j_3 & j & j_{23}\end{Bmatrix}=\Delta(j_1,j_2,j_{12})\Delta(j_1,j,j_{23})\Delta(j_3,j_2,j_{23})\Delta(j_3,j,,j_{12})w\begin{Bmatrix} j_1 & j_2 & j_{12} \\ j_3 & j & j_{23}\end{Bmatrix},$$

$$\Delta(a,b,c)=\sqrt{\frac{(a+b-c)!(a-b+c)!(-a+b+c)!}{(a+b+c+1)!}}, \tag{V.16}$$

$$w\begin{Bmatrix} j_1 & j_2 & j_{12} \\ j_3 & j & j_{23}\end{Bmatrix}=\sum_n (-1)^n(n+1)! \times[(n-j_1-j_2-j_{12})!(n-j_1-j-j_{23})!(n-j_3-j_2-j_{23})!(n-j_3-j-j_{12})!]^{-1} \times[(j_1+j_2+j_3+j-n)!(j_2+j_{12}+j+j_{23}-n)!(j_{12}+j_1+j_{23}+j_3-n)!]^{-1},$$

and the sum extends over all positive integers such that no factorial in the denominator has a negative argument. In conclusion, formulas (V.9)–(V.11) provide us with the general expression for the matrix elements of the operator $\hat{q}_{IJK}$ in the standard basis where it is understood that all occurring matrix elements are known in terms of $j_1,\ldots,j_n$ via (V.13), (V.15), and (V.16).

Let us call $M_{v,IJK}(\mathbf{g},\mathbf{g}'):=\langle\mathbf{g}|\hat{q}_{IJK}|\mathbf{g}'\rangle$. Then we have the matrix

$$M_v(\mathbf{g},\mathbf{g}')=\left(\frac{3}{2}\right)^3\frac{i}{4}\sum_{I<J<K}\epsilon(I,J,K)M_{v,IJK}(\mathbf{g},\mathbf{g}'),\tag{V.17}$$

which by inspection of (V.9) is a Hermitian matrix of a special kind, namely, it is of the form of i times a real-valued skew matrix. Its eigenvalues therefore come in pairs $\pm\lambda,\lambda>0$ or are zero. The eigenvalues of the positive semidefinite matrix $M_v^\dagger M_v$ are therefore $|\lambda|^2$ or 0 where λ is an eigenvalue of M_v. We conclude that the eigenvalues of $\hat{V}_{v,\gamma}$ are given by $l_p^3\sqrt{|\lambda|}$ or 0 where both eigenvectors of M_v corresponding to eigenvalues $\pm\lambda$ produce the same eigenvalue $l_p^3\sqrt{|\lambda|}$ of $\hat{V}_{v,\gamma}$. We see that the problem of diagonalizing the volume operator becomes equivalent to diagonalizing the matrices M_v which can be done for each $v\in V(\gamma)$ separately.

In the next subsection we will derive the spectrum in the most simple cases.

Remark: We have nowhere made use of the fact that we are only interested in gauge-invariant functions. Therefore, our results carry over, word by word, to the case of non-gauge-invariant functions provided we use the notion of an extended spin-network state.

Definition V.2: An extended spin-network state is a state of the form $T_{\gamma,\mathbf{j},\mathbf{c}}$ as before just that the vertex contractors now are Clebsh–Gordan coefficients of the form $\langle j_1m_1,\ldots,j_nm_n|j_1,\ldots,j_n,g_2,\ldots,g_{n-1},j\rangle$, where j may be different from zero.

This is important for applications in quantum gravity[2,3] and for the length operator[7] where we need the action of the volume operator on non-gauge-invariant functions in an intermediate step.

C. Examples

Given a vertex v and a vector of spins $\mathbf{j}$ coloring the edges incident at v, let us denote by $d(\mathbf{j},v)$ the number of linearly independent vectors $|\mathbf{g}\rangle$ compatible with $\mathbf{j}$. This number is the dimension of a vector space of vertex contractors associated with v and it is given purely algebraically as the number of linearly independent trivial representations which appear in the decomposition into irreducibles of the tensor product representation $\pi_{j_1}\otimes\cdots\otimes\pi_{j_n}$. One can determine this number for each case at hand by repeated application of the usual Clebsh–Gordan theorem $\pi_j\otimes\pi_{j'}=\pi_{j+j'}\oplus\pi_{j+j'-1}\oplus\cdots\oplus\pi_{|j-j'|}$ and counting all the trivial representation of spin 0 that appear. Notice that the sum is direct so that all the representations are orthogonal and therefore linearly independent. What one would like to have is a closed formula for $d(\mathbf{j},v)$, but that seems to be a complicated combinatorical problem. To our knowledge, such a dimension formula is only known for the case of an n-fold tensor product of fundamental representations $j_1=\cdots=j_n=\frac{1}{2}$ for the groups GL(N), U(N), and SU(N).[32] We will not deal with this problem in the present paper.

First of all, whenever $d(\mathbf{j},v)=1$, then 0 is the only possible eigenvalue (this observation was made first in Ref. 6 for the case of a trivalent graph in which the vertex contractors span only a one-dimensional vector space). More generally, whenever $d(\mathbf{j},v)$ is odd, we know that 0 is an eigenvalue of multiplicity at least 1.

Since M_v is a $d\times d$ skew matrix its characteristic polynomial has the structure $t^k(t^2-\lambda_1^2)\cdots(t^2-\lambda_{(d-k)/2}^2)=t^kp_{(d-k)/2}(t^2)$, where $0\leqslant k\leqslant d$ for d even and $1\leqslant k\leqslant d$ for d odd. It follows from Galois theory that we can determine the spectrum by quadratures exactly in general whenever $d(\mathbf{j},v)\leqslant 9$ and in fortunate cases (whenever 0 has multiplicity at least $d-8$) up to arbitrarily high dimension. In the most general case, however, a numerical evaluation is the only possible approach.

We consider the first nontrivial cases $d(\mathbf{j},v)\leqslant 9$ for an arbitrary valence n of the vertex. We may label the spins in such a way that $0<j_1\leqslant\cdots\leqslant j_n$. We may parametrize the situation by the non-negative number $k:=j_1+\cdots+j_{n-1}-j_n$ (there is no contractor for $k<0$). If k is low valued, then it is clear from Clebsh–Gordan theory that $d(\mathbf{j},v)$ will be low valued. Specifically, if $k=0$, then the first $n-1$ spins have to add up to maximum spin and thus $d=1$; the spectrum is only the point 0 irrespective of the value of n. In the following we will assume that $2j_1\geqslant k$, otherwise the discussion requires a case division. If $k=1$, then one sees readily that $d=n-1$: the only irreducible representations contained in $j_1\otimes j_2$ that can possibly add up with $j_3,\ldots,j_{n-1}$ to j_n are j_1+j_2,j_1+j_2-1. Repeating the argument, the assertion follows.

Let in general $c_k(n)$ be the number of representations with weight $j_1+\cdots+j_n-k$ contained in $j_1\otimes\cdots\otimes j_n$ for $j_{n+1}=j_1+\cdots+j_n-k$ to couple to angular momentum 0 where $j_1\leqslant,\ldots,\leqslant j_n$ and where we assume that $j_2-j_1\leqslant j_1+j_2-k$, that is $k\leqslant 2j_1$. We wish to show that $c_k(n)=(n+k$

$-2,k)$ where (n,k) is the usual binomial coefficient. Since $c_0(n)=1$ as we showed above, the induction is started. Now suppose we know all coefficients up to k for each n and start deriving those for $k+1$. We can assume that we know them already up to $n-1$ since $c_k(2)=1$ for all k is trivial to see since we have made the assumption $2j_1 \geqslant k$. Then we have the recursion $c_{k+1}(n-1)+c_k(n)=c_{k+1}(n)$ which follows from the fact that if we have the tensor product $j_1 \otimes \cdots \otimes j_{n-1}$, then there are $c_l(n-1)$ representations with weight $j_1+\cdots+j_{n-1}-l, 0 \leqslant l \leqslant k+1$. However, if we now form the tensor product with all of them with j_n, then the number of times that $j_1+\cdots+j_n-k$ arises from those with weight $j_1+\cdots+j_{n-1}-l, 0 \leqslant l \leqslant k$, which is given by $c_k(n)$, is equal to the number of times that $j_1+\cdots+j_n-(k+1)$ arises from them. In addition we have $c_{k+1}(n-1)$ terms coming from the representation $j_1+\cdots+j_{n-1}-(k+1)$ itself. It is now readily verified that $c_k(n)$ solves the recursion.

Notice that $n+1$ is the valence of the vertex. So we get back the result that in the trivalent case we have $d=c_k(2)=(k,k)=1$. We can now ask for which combinations of k,n we have $c_k(n) \leqslant 9$. Notice that $c_k(3)=k+1$:

Case $k=0$: n arbitrary.

Case $k=1$: $n-1 \leqslant 9$ so $n \leqslant 10$.

Case $k=2$: $n(n-1)/2 \leqslant 9$ so $n \leqslant 4$.

Case $3 \leqslant k \leqslant 8$: $n \leqslant 3$.

Case $k>8$: $n \leqslant 2$.

Reverse question (only $n \geqslant 2$ makes sense since we need valence at least 3 for the volume operator not to vanish trivially):

Case $n=2$: k arbitrary.

Case $n=3$: $k+1 \leqslant 9$ so $k \leqslant 8$.

Case $n=4$: $(k+1)(k+2)/2 \leqslant 9$ so $k \leqslant 2$.

Case $5 \leqslant n \leqslant 10$: $k \leqslant 1$.

Case $n>10$: $k=0$.

So, for instance we can treat the general four-valent case under the restriction that $j_4=j_1+j_2+j_3-k, 0 \leqslant k \leqslant 8$, for $j_3 \geqslant j_2 \geqslant j_1 \geqslant k/2$ in which case $d=k+1$. As the formulas already get quite involved, we will only discuss the cases $k=1,2,3$ explicitly.

Notice first of all that due to gauge invariance $J_1+J_2+J_3+J_4|0\rangle=0$ so that we can substitute for J_4. Thus, for instance, $\hat{q}_{124}|0\rangle=-(\hat{q}_{121}+\hat{q}_{122}+\hat{q}_{123})|0\rangle$ and it is easily verified that the operator identity $\hat{q}_{122}+\hat{q}_{121}=0$ holds. It follows that

$$M_v=\left(\frac{3}{4}\right)^3 \frac{i}{4}\,[\epsilon(1,2,3)-\epsilon(1,2,4)+\epsilon(1,3,4)-\epsilon(2,3,4)]\hat{q}_{123}=:\sigma\hat{q}_{123},$$

which simplifies the discussion tremendously since we just need to compute the matrix elements of the single operator $\hat{q}_{123}$. We have to compute the $k(k+1)/2$ matrix elements

$$\langle j_{12}|\hat{q}_{123}|j'_{12}\rangle=[j_{12}(j_{12}+1)-j'_{12}(j'_{12}+1)]\sum_{j_{23}} j_{23}(j_{23}+1)\langle j_{12}|j_{23}\rangle\langle j'_{12}|j_{23}\rangle$$

for $j_{12}, j'_{12}=j_1+j_2-l$, $l=0,\ldots,k$, and the sum extends over $j_{23}=j_2+j_3-l$, $l=0,\ldots,k$, which requires the computation of $(k+1)^2$ $6j$-symbols for $j_{123}=j=j_1+j_2+j_3-k=j_4$. The computation is rather tedious. We will display only a few intermediate calculational results before giving the eigenvalue λ_k.

Let $j_{12}:=j_1+j_2, j_{23}:=j_2+j_3$ and $J=2(j_1+j_2+j_3)$.

Case $k=1$: The $6j$-symbols are

$$\begin{aligned}\langle j_{12}|j_{23}\rangle&=-\langle j_{12}-1|j_{23}-1\rangle=\sqrt{\frac{j_1 j_3}{j_{12}j_{23}}},\\ \langle j_{12}-1|j_{23}\rangle&=\langle j_{12}|j_{23}-1\rangle=\sqrt{\frac{j_2 J}{2j_{12}j_{23}}},\end{aligned} \tag{V.18}$$

and so we have the only nonvanishing matrix element

$$\langle j_{12}|\hat{q}_{123}|j_{12}-1\rangle = 4\sqrt{j_1 j_2 j_3 (j_1+j_2+j_3)}. \tag{V.19}$$

The corresponding eigenvalue of the volume operator is

$$\lambda_1 = 2 l_p^3 \sqrt{|\sigma|}\sqrt[4]{j_1 j_2 j_3 (j_1+j_2+j_3)}. \tag{V.20}$$

This result, formula (V.20), was obtained before by Jurek Lewandowski using an independent method (also using gauge invariance $J_1+J_2+J_2+J_3=0$).

Case $k=2$: The $6j$-symbols are

$$\begin{aligned}
\langle j_{12}|j_{23}\rangle &= \sqrt{\frac{j_1(2j_1-1)j_3(2j_3-1)}{j_{12}(2j_{12}-1)j_{23}(2j_{23}-1)}},\\
\langle j_{12}-1|j_{23}\rangle &= \sqrt{2\frac{(2j_1-1)j_2 j_3(J-1)}{j_{12}(2j_{12}-2)j_{23}(2j_{23}-1)}},\\
\langle j_{12}-2|j_{23}\rangle &= \sqrt{\frac{j_2(2j_2-1)(J-1)(J-2)}{(2j_{12}-1)(2j_{12}-2)j_{23}(2j_{23}-1)}},\\
\langle j_{12}|j_{23}-1\rangle &= \sqrt{2\frac{j_1 j_2(2j_3-1)(J-1)}{j_{12}(2j_{12}-1)j_{23}(2j_{23}-2)}},\\
\langle j_{12}|j_{23}-2\rangle &= \sqrt{\frac{j_2(2j_2-1)(J-1)(J-2)}{j_{12}(2j_{12}-1)(2j_{23}-1)(2j_{23}-2)}},\\
\langle j_{12}-1|j_{23}-1\rangle &= \frac{(2j_2-1)(J-1)-(2j_1-1)(2j_3-1)}{2\sqrt{j_{12}(2j_{12}-2)j_{23}(2j_{23}-2)}},\\
\langle j_{12}-2|j_{23}-1\rangle &= -\sqrt{2\frac{(2j_1-1)(2j_2-1)j_3(J-2)}{(2j_{12}-1)(2j_{12}-2)j_{23}(2j_{23}-2)}},\\
\langle j_{12}-1|j_{23}-2\rangle &= -\sqrt{2\frac{j_1(2j_2-1)(2j_3-1)(J-2)}{j_{12}(2j_{12}-2)(2j_{23}-1)(2j_{23}-2)}},\\
\langle j_{12}-2|j_{23}-2\rangle &= 2\sqrt{\frac{j_1(2j_1-1)j_3(2j_{23}-1)}{(2j_{12}-1)(2j_{12}-2)(2j_{23}-1)(2j_{23}-2)}},
\end{aligned} \tag{V.21}$$

and we find then the matrix elements

$$\begin{aligned}
\langle j_{12}|\hat{q}_{123}|j_{12}-1\rangle &= \sqrt{\frac{2j_1 2j_2(2j_3-1)(J-1)}{2(2j_{12}-1)(2j_{12}-2)}}\frac{1}{2(2j_{23}-1)}\,[(2j_1-1)[6j_3+2j_{23}-1]+(2j_2-1)\\
&\quad\times[3J+2j_{23}-7]],\\
\langle j_{12}|\hat{q}_{123}|j_{12}-2\rangle &= 0,\\
\langle j_{12}-1|\hat{q}_{123}|j_{12}-2\rangle &= \sqrt{\frac{(2j_1-1)(2j_2-1)2j_3(J-2)}{2(2j_{12})(2j_{12}-1)}}\frac{1}{2(2j_{23}-1)}\\
&\quad\times[(J-1)[6j_2+2j_{23}-1]+(2j_3-1)[6j_1-2j_{23}+1]],
\end{aligned} \tag{V.22}$$

and since the eigenvalues of any 3×3 skew matrix with entries a,b,c are given by 0, $\pm i\sqrt{a^2+b^2+c^2}$, we find the eigenvalues $0,\lambda_2$ where

$$\lambda_2 = l_p^3\sqrt{|\sigma|}\sqrt[4]{\langle j_{12}|\hat{q}_{123}|j_{12}-1\rangle^2+\langle j_{12}-1|\hat{q}_{123}|j_{12}-2\rangle^2}. \tag{V.23}$$

Case $k=3$: We will refrain from displaying the $6j$-symbols explicitly in terms of the j_I. Rather, we consider them as given through the formulas of the previous section. Now it is easy to check that the eigenvalues of a 4×4 skew matrix with entries a,b,c,d,e,f are given by i times

$$t=\pm\frac{1}{\sqrt{2}}\sqrt{a^2+b^2+c^2+d^2+e^2+f^2\pm\sqrt{[(a+f)^2+(c+d)^2+(e-f)^2]\times[(a-f)^2+(c-d)^2+(e+f)^2]}}$$

and notice that the argument of the inner square root can be written as $(a^2+b^2+c^2+d^2+e^2+f^2)-4(af+cd-eb)^2$, which shows that t is purely real. Substituting the matrix elements $\langle j_{12}-l|\hat{q}_{123}|j_{12}-l'\rangle, 0\leqslant l<l,\leqslant 3$ for a,b,c,d,e,f we obtain

$$\lambda_3=l_p^3\sqrt{|\sigma|}\sqrt{|t|}. \tag{V.24}$$

The reader should now have a feeling of how to compute the spectrum analytically. The formulas derived can serve for analytical estimates as well as the starting point for the implementation of an appropriate computer code for an algebraic manipulation program.

ACKNOWLEDGMENTS

The author is indebted to Abhay Ashtekar and Jurek Lewandowski for communicating their results on the volume operator prior to publication. This research project was supported in part by DOE Grant No. DE-FG02-94ER25228 to Harvard University.

[1] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, J. Math. Phys. **36**, 519 (1995).
[2] T. Thiemann, ''Anomaly-free Formulation of Non-Perturbative, Four-Dimensional Lorentzian Quantum Gravity,'' Harvard University preprint HUTMP-96/B-350, Phys. Lett. B **380**, 257 (1996).
[3] T. Thiemann, ''Quantum Spin Dynamics (QSD),'' Harvard University preprint HUTMP-96/B-359, gr-qc/9606089, Class. Quantum Grav. **15**, 839–73 (1998); T. Thiemann, ''Quantum Spin Dynamics (QSD) II: The Kernel of the Wheeler-DeWitt Constraint Operator,'' Harvard University preprint HUTMP-96/B-352, gr-qc/9606090, Class. Quantum Grav. **15**, 875–905 (1998).
[4] C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995); Erratum, Nucl. Phys. B **456**, 734 (1995).
[5] A. Ashtekar and J. Lewandowski, ''Quantum Theory of Geometry I: Area Operators,'' preprint CGPG-96/2-4, gr-qc/9602046, Class. Quantum Grav. **14**, A55–A81 (1997); ''Quantum Theory of Geometry II: Volume Operators,'' gr-qc/9711031.
[6] R. Loll, Phys. Rev. Lett. **75**, 3048 (1995); Nucl. Phys. B **460**, 143 (1996); Class. Quantum Grav. **14**, 1725 (1997); Nucl. Phys. B **500**, 405 (1997).
[7] T. Thiemann, ''A length operator in canonical quantum gravity,'' Harvard University preprint HUTMP-96/B-354, gr-qc/9606092, J. Math. Phys. 3372–3392 (1998). (Paper II).
[8] T. Thiemann, ''QSD VI: Quantum Poincaré Algebra and a Quantum Positivity of Energy Theorem for Canonical Quantum Gravity,'' Harvard University preprint HUTMP-96/B-356, gr-qc/9705020, Class. Quantum Gravity (in press).
[9] T. Thiemann, ''QSD V: Quantum Gravity as the natural Regulator of the Hamiltonian Constraint of Matter Quantum Field Theories,'' Harvard University preprint HUTMP-96/B-357, gr-qc/9705019, Class. Quantum Gravity (in press).
[10] J. Lewandowski, Class. Quantum Grav. **14**, 71 (1997).
[11] R. Borissov, R. De Pietri, and C. Rovelli, Class. Quantum Grav. **14**, 2793 (1997).
[12] R. De Pietri and C. Rovelli, Phys. Rev. D **54**, 2664 (1996).
[13] T. Thiemann, ''The inverse loop transform,'' Harvard University preprint HUTMP-95/B-346, hep-th/9601105, J. Math. Phys. **39**, 1236 (1998).
[14] R. De Pietri, ''On the relation between the connection and the loop representation of quantum gravity,'' gr-qc/9605064, Class. Quantum Grav. **14**, 53 (1997).
[15] A. Ashtekar and J. Lewandowski, J. Geom. Phys. **17**, 191 (1995).
[16] A. Ashtekar, Phys. Rev. D **36**, 1587 (1987); *Lectures on Nonperturbative Canonical Gravity* (World Scientific, Singapore, 1991).
[17] A. Ashtekar, in *Mathematics and General Relativity* (AMS, Providence, 1987).
[18] F. Barbero, Phys. Rev. D **51**, 5507 (1995).
[19] R. Giles, Phys. Rev. D **8**, 2160 (1981).
[20] R. Gambini and A. Trias, Nucl. Phys. B **278**, 436 (1986).
[21] C. Rovelli and L. Smolin, Phys. Rev. Lett. **61**, 1155 (1988).
[22] A. Ashtekar and C. J. Isham, Class. Quantum Grav. **9**, 1433 (1992).
[23] A. Ashtekar and J. Lewandowski, ''Representation theory of analytic holonomy $C^\star$ algebras,'' in *Knots and Quantum Gravity*,'' edited by J. Baez (Oxford U.P. Oxford, 1994).
[24] J. Baez, Lett. Math. Phys. **31**, 213 (1994); ''Diffeomorphism invariant generalized measures on the space of connections modulo gauge transformations,'' hep-th/9305045, in *Proceedings of the Conference on Quantum Topology*, edited by D. Yetter (World Scientific, Singapore 1994).
[25] D. Marolf and J. M. Mourão, Commun. Math. Phys. **170**, 583 (1995).

[26] A. Ashtekar and J. Lewandowski, J. Math. Phys. **36**, 2170 (1995).
[27] R. Loll, Phys. Lett. B **399**, 227 (1997).
[28] J. Mourão, T. Thiemann, and J. Velhinho, ''Physical properties of measures on infinite dimensional spaces,'' preprint HUTMP-97/B-365, hep-th/9711139.
[29] C. Rovelli and L. Smolin, Phys. Rev. D **53**, 5743 (1995).
[30] J. Baez, ''Spin network states in gauge theory,'' gr-qc/9411007, Adv. Math. (in press); ''Spin networks in non-perturbative quantum gravity,'' gr-qc/9504036.
[31] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1974).
[32] H. Boerner, *Darstellungen von Gruppen* (Springer-Verlag, Berlin, 1967).

A length operator for canonical quantum gravity

T. Thiemann[a)]
Physics Department, Harvard University, Cambridge, Massachusetts 02138

(Received 15 May 1997; accepted for publication 16 January 1998)

We construct an operator that measures the length of a curve in four-dimensional Lorentzian vacuum quantum gravity. We work in a representation in which an SU(2) connection is diagonal and it is therefore surprising that the operator obtained after regularization is densely defined, does not suffer from factor ordering singularities, and does not require any renormalization. We show that the length operator admits self-adjoint extensions and compute part of its spectrum which, like its companions, the volume and area operators already constructed in the literature, is purely discrete and roughly quantized in units of the Planck length. The length operator contains full and direct information about all the components of the metric tensor which facilitates the construction of so-called weave states which approximate a given classical three-geometry. © *1998 American Institute of Physics.* [S0022-2488(98)02406-2]

I. INTRODUCTION

It is by now widely believed that perturbative quantization of gravity is impossible. Among the nonperturbative approaches, the canonical approach[1] has the advantage of not requiring any extra structure beyond classical general relativity and the principles of quantum theory which is in contrast, for instance, to string theory.

Roughly speaking (for details see, e.g., Ref. 2), in the canonical approach one splits the four-dimensional, Lorentzian space–time manifold M into time and space, that is, M has the topology $M=\mathbb{R}\times\Sigma$, where Σ is a Riemannian three-manifold of arbitrary topology, and writes the Einstein–Hilbert action in Hamiltonian form with respect to this $3+1$ split. One finds that the pull-back q_{ab}, $a=1,2,3$, to Σ of the four-dimensional metric $g_{\mu\nu}$, $\mu,\nu=0,...,3$, is one of the canonical coordinates and it is therefore of particular interest to construct an operator in the quantum theory that measures $q_{ab}(x)$ for all $x\in\Sigma$.

A natural classical phase space functional of q_{ab} which contains full and direct information about this positive definite three-metric is, of course, the length of an arbitrary piecewise smooth curve c in Σ. The classical length of c is defined by $L(c):=\int_0^1 dt\sqrt{q_{ab}(c(t))\dot{c}^a(t)\dot{c}^b(t)}$, where $[0,1]\to c$ is a parametrization of the curve and it is understood that one sums over smooth segments of c. This object should also play an important role in the quantum theory due to its intuitive physical meaning.

Here a word of caution is in order:

Due to its four-dimensional diffeomorphism invariance, canonical general relativity is subject to four constraints corresponding to infinitesimal diffeomorphisms of M.[2] Obviously then, the length $L(c)$ is not a Dirac observable, i.e., it has nonvanishing Poisson brackets with the constraint functionals because it is, in particular, not invariant under diffeomorphisms of Σ. This means that part of the information contained in $L(c)$, as we vary c, is actually gauge and thus not physically relevant. On the other hand, it is mathematically difficult to extract the ''relevant'' information from $L(c)$ and therefore we will, in the present paper, consider all $L(c)$ since it is at least an overcomplete set of data from which q_{ab} can be reconstructed.

The problem in quantizing the length functional is, of course, that q_{ab} turns into an operator and experience from, say, scalar field theories, raises the worry that this operator must be ill defined in the quantum theory for it is smeared only in one direction: it is well known that a scalar

a)Electronic mail: thiemann@math.harvard.edu New address: Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, Schlaatzweg 1, 14473 Potsdam, Germany. Electronic mail: thiemann@aei-potsdam.mpg.de

 © 1998 American Institute of Physics

field $\phi(x)$, even for free-field theory, is an operator-valued distribution which turns into a densely defined operator on the Hilbert space (Fock space) only when smearing it in three dimensions.

If one was working in a representation of canonical quantum gravity for which the intrinsic metric q_{ab} of an initial data hypersurface Σ was the configuration variable, then the operator corresponding to the length of a curve would be fairly easy to construct because it would act by multiplication and so there would be no immediate problem, even if q_{ab} would take values in a set of distributions rather than in the set of smooth functions of Σ.

However, in this so-called geometrodynamic representation[2] one of the most important operators, the Wheeler–DeWitt constraint operator, adopts an algebraic form which is so difficult that almost half a century after its discovery it is still unknown how to rigorously define it and much less how to solve it (compare Ref. 1 for a detailed analysis).

Fortunately, there is an alternative, called the connection representation, in which the Wheeler–DeWitt operator adopts a polynomial form after multiplying it by $\sqrt{\det((q_{ab}))}$.[3] In the Lorentzian signature, if one insists on a polynomial Wheeler–DeWitt constraint, the underlying connection is complex valued,[3] which renders the task of incorporating the correct reality conditions on the classical phase space variables at the quantum level into a difficult one. However, these complications can be overcome in two different ways:

(a) The first possibility is to start with Euclidean gravity and real connection variables in which the Wheeler–DeWitt constraint is also a polynomial[4] and to Wick rotate the Euclidean theory into the Lorentzian regime[5] (see also Ref. 6).
(b) The second possibility arises from the recent discovery[7,8] that by employing a certain novel technique it is indeed possible to obtain a finite, densely defined (and symmetric) operator corresponding to the original Wheeler–DeWitt constraint directly in the Lorentzian regime in terms of real-valued su(2) connection variables A_a^i and without multiplying by a power of $\sqrt{\det(q)}$.

The idea to use real connection variables for *Lorentzian* gravity rather than for Euclidean gravity was clearly spelled out for the first time in Ref. 9; in particular, it was pointed out there that, despite the algebraically complicated structure of the Wheeler–DeWitt constraint, the techniques associated with $\overline{\mathcal{A}/\mathcal{G}}$ as discussed below can be employed. The technique advertised in Refs. 7 and 8 is, however, independent of the one described in Ref. 9 which was restricted to the classical case and unavoidably is doomed to produce a highly singular Hamiltonian constraint operator (compare Ref. 10 where this singularity is discussed in the restricted context of lattice quantum gravity). In contrast, the one proposed in Refs. 7 and 8 is perfectly finite everywhere on the Hilbert space. Also, it is independent of the technique used in Ref. 11 which suffers from other problems that have to do with taking the square root of a nonpositive and not self-adjoint operator besides being a quantization of the Euclidean constraint operator only.

Summarizing, it seems then that we are forced to quantize the length functional in this real connection representation which we will use in the sequel. The problem then is that the metric depends on the momentum E_i^a canonically conjugate to the connection only, and it depends on it in a nonpolynomial fashion. Specifically, $q_{ab}=\det((E_i^a))E_a^iE_b^i$ where E_a^i is the *inverse* of the matrix E_i^a. It is therefore much harder to define a well-defined operator corresponding to it. In fact, that is why only volume and area operators could be constructed so far.[12,13] It turns out that the same novel technique introduced in Refs. 7 and 8 can be used to derive a completely well-defined operator corresponding to the length of a curve which is the content of the present paper. The paper is organized as follows:

In Sec. II we fix the notation and define the real connection representation (see Ref. 14 for a more detailed review and references therein). We recall the necessary information about the underlying Hilbert space and the techniques that come with the space of (generalized) connections modulo gauge transformations $\overline{\mathcal{A}/\mathcal{G}}$. This Hilbert space can be roughly thought of as the completion of a space Φ which is spanned by finite sums of finite products of traces of the holonomy of the connection around closed loops, also called the Wilson loop functional. The completion is with respect to a measure which, roughly speaking, reduces on functions of n independent Wilson loop variables to the n-fold product Haar measure on $\mathrm{SU}(2)^n$. A genuine state in Φ can then roughly be thought of as a functional of the connection labeled by a collection of loops, that is, a graph.

In Sec. III we derive the length operator and show that it is an unbounded, non-negative and symmetric operator on the full Hilbert space with dense domain and that it has self-adjoint extensions. The construction involves several steps:

(a) First one has to regularize the naive expression for the length operator, that is, one derives an operator which depends on a specific parametrization and a specific partition into segments of the curve in question. The regulator is characterized by the maximal interval length of that partition.
(b) Next, one takes the regulator to zero and observes that on each state from a dense subset Φ of the Hilbert space there exists a value ϵ_0 of the regulator such that for any $\epsilon < \epsilon_0$ the operator remains the same. The result is a family of regulator-independent operators, one for each state $\psi \in \Phi$.
(c) Finally, one confirms that this family of operators qualifies as the set of projections to the one-dimensional spaces spanned by states from Φ of a state-independent operator on the Hilbert space which has domain and range in the dense set Φ.

In Sec. IV we derive several properties of its spectrum and compute it for some simple situations. In particular, it is shown that the spectrum is purely discrete. With the techniques that are introduced to establish this result it is actually possible to show that the length operator so obtained is an essentially self-adjoint operator on the Hilbert space so that its self-adjoint extension is unique. It turns out that the length operator is indeed diagonal on the ''spin-network states'' introduced by Penrose[15] (see Sec. II).

Finally, in Sec. V we comment on how one can construct ''weave states'' which approximate a given classical geometry as defined by the length operator. We need to consider weave states which are more general than most of the ones previously considered in the literature in the sense that they necessarily involve intersections and overlappings of the loops involved and should therefore be called ''lattice states.'' This section, in contrast to the previous ones, is less precise and contains several heuristic elements which is due to the following issue:

Surprisingly, the length operator cannot be directly employed in order to construct weave states: This has to do with the fact that a curve is a very singular object in the sense that the length operator acts on a state nontrivially only if the graph on which the state depends intersects the curve. This has the striking consequence that, given some region occupied by some genuine loop configuration, the quantum volume of that region is positive while the length of ''almost any'' curve that penetrates the same region is zero. The conclusion is that a one-dimensional curve is a mathematical abstraction of a physical curve which should have a finite ''thickness,'' that is, on a macroscopic scale the curve looks one-dimensional but microscopically it looks rather like a solid tube. This is, of course, motivated by the general rule in quantum field theory that the excitations of the quantum field cannot be measured in a point but only in a ''smeared'' region.

We construct, directly from the length operator, an associated ''tube operator'' which is nothing else than a certain sum of length operators for a congruence of curves whose union is (piecewise) a solid cylinder. One difficulty that has to be overcome is that one has to define what it means that a tube is ''thin,'' i.e., that it should look one-dimensional on a macroscopic scale. We define a notion of ''quantum thinness,'' again by means of the length operator, roughly by requiring that the thickness of the tube should be of the order of the Planck length. Of course, the tube will be thin only on certain states. The length of a tube is, roughly speaking, the sum of the lengths of the length operators from which it is constructed. A weave state that approximates a classical metric is then a state with respect to which the tube is thin and such that the length of the tube agrees up to quantum corrections with the length of the ''central curve'' of the tube as measured by the classical metric. Using the calculations from Sec. IV we explicitly construct a weave state that approximates the length of a tube as measured by any slowly varying classical three-metric.

II. THE REAL CONNECTION REPRESENTATION

Let q_{ab} be the intrinsic metric of an initial data hypersurface Σ and let K_{ab} be its extrinsic curvature. Introduce a triad field e^i_a which transforms like a one-form under diffeomorphisms of Σ and according to the adjoint representation of SU(2) so that $q_{ab} = \delta_{ij} e^i_a e^j_b$. Let e^a_i be its inverse

and define $K_a^i := \mathrm{sgn}(\det((e_b^j)))K_{ab}e_i^b$ and $E_i^a := \det((e_b^j))e_i^a$. Then one can show that $(K_a^i/\kappa, E_i^a)$ is a canonical pair for Lorentzian four-dimensional canonical gravity, where κ is Newton's constant.

Now consider the spin-connection Γ_a^i, that is, the unique connection which annihilates the triad e_a^i. One can define an SU(2) connection $A_a^i := \Gamma_a^i + K_a^i$ which has the correct dimension of an inverse length. Then one can show that the so-called real Ashtekar variables given by $A_a^i/\kappa, E_i^a$ define a canonical pair. In retrospect, since this real connection formulation does not simplify the Wheeler–DeWitt constraint too much while the complex formulation suffers from the other problems mentioned in the Introduction, the virtue of using a connection dynamics formulation rather than a geometrodynamical one is the following: one can use techniques normally employed in Yang–Mills theory. In particular, one can use Wilson loop variables which serve as good coordinates on $\mathcal{A}/\mathcal{G}$, the space of smooth connections modulo gauge transformations. The use of loops to probe connections is radical: those Wilson loop variables can become well-defined operators only if the excitations of geometry are stringlike rather than bubblelike. On the other hand, given that assumption, it is possible to explicitly characterize the quantum configuration space $\overline{\mathcal{A}/\mathcal{G}}$ of *generalized* connections modulo gauge transformations. This is the precise analog, in the connection representation, of ''Wheeler's superspace'' in the metric formulation which, to the best of our knowledge, was never specified precisely.

One can show that the elements of $\overline{\mathcal{A}/\mathcal{G}}$ are in one-to-one correspondence with the set of all homomorphisms from the group of holonomically equivalent loops in Σ into SU(2)/Ad.[16] Moreover, there is a σ-additive, faithful, and diffeomorphism-invariant probability measure μ_0 on $\overline{\mathcal{A}/\mathcal{G}}$ which equips us with a Hilbert space structure $\mathcal{H} := L_2(\overline{\mathcal{A}/\mathcal{G}}, d\mu_0)$. This measure is defined as follows:

In the sequel we will denote by γ a closed graph, that is, a collection of analytic edges e which intersect in vertices v such that each vertex is at least two-valent. A function f on $\overline{\mathcal{A}/\mathcal{G}}$ is said to be cylindrical with respect to γ iff it is of the form $f(A) = (f_\gamma \circ p_\gamma)(A) = f_\gamma(h_{e_1}(A), \ldots, h_{e_n}(A))$, where $e_1, \ldots, e_n$ are the edges of γ, h_e is the holonomy along e_i, and f_γ is a complex-valued function on $\mathrm{SU}(2)^n$ which is gauge invariant. So each cylindrical function is determined through a graph γ and such an f_γ and one says that f_γ and f'_γ are equivalent whenever their pullbacks agree, that is $p_\gamma^* f_\gamma = p_{\gamma'}^* f_{\gamma'}$. Let us denote by $\mathrm{Cyl}_\gamma(\overline{\mathcal{A}/\mathcal{G}})$ the collection of functions cylindrical with respect to γ modulo cylindrical equivalence and denote by $\mathrm{Cyl}(\overline{\mathcal{A}/\mathcal{G}}) := \cup_\gamma \mathrm{Cyl}_\gamma(\overline{\mathcal{A}/\mathcal{G}})$ the set of all cylindrical functions. The measure (μ_0) can now be seen as the σ-additive extension[17] of the following self-consistent family of measures $(\mu_{0,\gamma})_\gamma$.[18] Let $f \in \mathrm{Cyl}_\gamma(\overline{\mathcal{A}/\mathcal{G}})$. Then

$$\int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu(A) f(A) := \int_{\overline{\mathcal{A}/\mathcal{G}}} d\mu_{0,\gamma}(A) f(A) := \int_{SU(2)^n} d\mu_H(g_1) \cdots d\mu_H(g_n) f_\gamma(g_1, \ldots, g_n), \tag{II.1}$$

where μ_H is the Haar measure on SU(2) and $g_I = h_{e_I}(A)$. In other words, everything is reduced to finite-dimensional integrals over SU(2). One can show that $\mathrm{Cyl}(\overline{\mathcal{A}/\mathcal{G}})$ is dense in $\mathcal{H}$.

Thus integral calculus is introduced on $\overline{\mathcal{A}/\mathcal{G}}$. One can even develop differential calculus on $\overline{\mathcal{A}/\mathcal{G}}$:[19] a differentiable cylindrical function is simply a differentiable function on $\mathrm{SU}(2)^n$. Functional derivatives can be evaluated on differentiable elements of $\mathrm{Cyl}(\overline{\mathcal{A}/\mathcal{G}})$ because on a finite graph a distributional connection can be replaced by a smooth one.[16,18] One sees that differential and integral calculus is inherited from the one on $\mathrm{SU}(2)^n$. Finally, let us define the spaces $\mathrm{Cyl}_\gamma^n(\overline{\mathcal{A}/\mathcal{G}})$, $\mathrm{Cyl}^n(\overline{\mathcal{A}/\mathcal{G}}) = \cup_\gamma \mathrm{Cyl}_\gamma^n(\overline{\mathcal{A}/\mathcal{G}})$, of differentiable cylindrical functions of order $n = 0,1,\ldots,\infty$. Each of $\mathrm{Cyl}^n(\overline{\mathcal{A}/\mathcal{G}})$ is dense in $\mathcal{H}$.

III. THE LENGTH OPERATOR

The stage is now set to derive the length operator. Before we do that we wish to comment on why it was impossible up to now to define this operator rigorously on $\mathcal{H}$ and why it was possible to define volume and area operators.

Consider for instance the volume of a region $R \subset \Sigma$,

$$V(R) := \int_R d^3x \sqrt{|\det(q)|} = \int_R d^3x \sqrt{\left|\frac{1}{3!}\,\epsilon_{abc}\epsilon^{ijk}E^a_i E^b_j E^c_k\right|}, \tag{III.1}$$

where ϵ_{abc} carries density weight -1. We see that this functional involves the square root of a *polynomial* in E^a_i. This is important because, according to the canonical commutation relations, we are supposed to replace E^a_i by $\hat{E}^a_i = -i\ell_p^2\delta/\delta A^i_a$, where $\ell_p := \sqrt{k\hbar}$ is the Planck length. Now one tries to regularize the polynomial and defines its square root by its spectral resolution. This has been done successfully in the literature.[12,13]

We know how to regularize a polynomial in the basic operators A, E, but certainly we do not know how to define a nonpolynomial function. This is precisely the case for

$$q_{ab} = \epsilon_{acd}\epsilon_{bef}\epsilon^{ijk}\epsilon^{imn}\,\frac{E^c_j E^d_k E^e_m E^f_n}{4\det((E^g_l))^2}. \tag{III.2}$$

Even if one could define it, it is by now well known that the operator version of the denominator of (III.2) has a huge kernel so that it could not be defined on a dense domain. This is the reason why a quantization of the length of a piecewise differentiable curve $c:[0,1]\to\Sigma;\ t\to c(t)$ given classically by

$$L(c) := \int_{[0,1]} dt \sqrt{\dot{c}^a(t)\dot{c}^b(t)q_{ab}(c(t))} \tag{III.3}$$

has escaped its quantization in the representation $\mathscr{H}$ so far. Note that no absolute value signs are necessary in (III.3) under the square root since $\dot{c}^a\dot{c}^b q_{ab} = (\dot{c}^a e^i_a)(\dot{c}^b e^j_b)\,\delta_{ij} \geqslant 0$.

A. Regularization of the length operator

The regularization of (III.3) is based on two *key* observations:

Observation 1: The triad is integrable with generating functional $V := V(\Sigma)$, the total volume of Σ. More precisely we have $e := \mathrm{sgn}(\det((e^i_a)))$,

$$\frac{\delta V}{\delta E^a_i} = e\,\frac{e^i_a}{2} = \frac{1}{\kappa}\,\{A^i_a, V\}, \tag{III.4}$$

which can be verified immediately.

Observation 2: The total volume can be quantized in a mathematically rigorous way. Its action on sufficiently differentiable cylindrical functions is given by (we follow Ref. 13)

$$\hat{V}f = l_p^3\hat{v}f = \ell_p^3 \sum_{v\in V(\gamma)} \hat{V}_v f_\gamma = \ell_p^3 \sum_{v\in V(\gamma)} \sqrt{\left|\frac{i}{8\cdot 3!}\sum_{e_I\cup e_J\cup e_K = v} \epsilon(e_I,e_J,e_K)\epsilon_{ijk}X^i_I X^j_J X^k_K\right|} f_\gamma. \tag{III.5}$$

Here $g_I = h_{e_I}(A)$, $X_I = X(g_I)$, and $X^j(g) := \mathrm{tr}(g[-i\sigma_j/2]\partial_g)$ are the components of the right invariant vector field on SU(2) (σ_j are Pauli matrices) where we have assumed that the edges are outgoing at a vertex. Here $V(\gamma)$ is the set of vertices of γ and finally $\epsilon(e_I,e_J,e_K) = \mathrm{sgn}(\det(\dot{e}_I(0),\dot{e}_J(0),\dot{e}_K(0)))$. Note that $\hat{v}$ is a dimensionless operator. One can show that (III.5) has dense domain $\mathrm{Cyl}^3(\mathscr{A}/\mathscr{G})$ and is an essentially self-adjoint operator on $\mathscr{H}$ with a purely discrete spectrum (compare the second paper of Ref. 13).

These two observations motivate the following regularization strategy.

Choose the basis in su(2) given by $\tau_j = -i\sigma_j/2$ and write $e_a = e^i_a\tau_i$, $A_a = A^i_a\tau_i$ so that for smooth A we have $h_e(A) = 1 + A(e) + o(A(e)^2)$, where $A(e) = \int_e A$. Then, according to (III.4),

$$q_{ab} = -2\,\mathrm{tr}(e_a e_b) = -\frac{8}{\kappa^2}\,\mathrm{tr}(\{A_a,V\}\{A_b,V\}). \tag{III.6}$$

Clearly we now are going to replace the integral in (III.3) by a Riemann sum and take the limit. So, let $t_0=0<t_1<\cdots<t_n=1$ be a partition of $[0,1]$ such that points t of nondifferentiability of c are among the values t_i and let $\Delta_i:=t_i-t_{i-1}$. It is understood that in the limit $n\to\infty$ also $\delta:=\max((\Delta_i)_{i=1}^n)\to 0$. Consider then the following quantity:

$$L_n(c):=\frac{1}{\kappa}\sum_{i=1}^{n}\sqrt{2\,\mathrm{tr}(\{h_c(t_{i-1},t_i),V\}\{h_c(t_{i-1},t_i)^{-1},V\})}. \tag{III.7}$$

Here $h_c(s,t)$ denotes the holonomy of A along c from the parameter value $s\to t$. It is easy to see that for a classical (that is, smooth) connection we have $\{h_c(t_{i-1},t_i)^{\pm 1},V\}=\pm\Delta_i\dot{c}^a(\tilde{t}_i)\{A_a(c(\tilde{t}_i)),V\}+o(\Delta_i^2)$, where $\tilde{t}_i$ is some value of $t\in[t_{i-1},t_i]$. Therefore (III.7) converges, in the sense of smooth functions, classically to (III.3) in the limit $n\to\infty$. Note that the argument of the square root is manifestly gauge invariant. The motivation to replace the A's by holonomies of course comes from the fact that $\hat{V}$ has finite action on holonomies as seen from (III.5).

The final step is to replace V by $\hat{V}$ and Poisson brackets by commutators times $1/i\hbar$. The result is

$$\hat{L}_n(c):=\ell_p\sum_{i=1}^{n}\sqrt{-8\,\mathrm{tr}([h_c(t_{i-1},t_i),\hat{v}][h_c(t_{i-1},t_i)^{-1},\hat{v}])}. \tag{III.8}$$

To complete the definition of the length operator we now define for each thrice differentiable cylindrical function $f\in \mathrm{Cyl}^3(\overline{\mathcal{A}/\mathcal{G}})$

$$\hat{L}(c)f:=\lim_{n\to\infty}[\hat{L}_n(c)f]. \tag{III.9}$$

As it stands, (III.9) does not make sense yet because we have not shown that the limit exists and it may matter how to choose the partition $(t_i)_{i=0}^n$. Also, it is far from obvious that the square root is well defined because we did not show that its argument is a positive and self-adjoint operator. These issues will be settled in the next subsections.

B. Finiteness and choice of the partition adapted to a graph

Let f be a function cylindrical with respect to a graph γ and let s be one of the segments of c into which we have partitioned it. We wish to study the action of the operator

$$\hat{l}_s^2:=-8\,\mathrm{tr}([h_s,\hat{v}][h_s^{-1},\hat{v}]) \tag{III.10}$$

on f. By choosing n in (III.9) large enough we may assume without loss of generality that s and γ are either disjoint, intersect at, at most one point, or s is contained in γ [here we have made use of the piecewise analyticity of γ (Ref. 18)]. Further we adapt the partition to the graph in the following way: (a) if s and γ intersect in a point, then this point is a boundary point of s, and (b) if s is contained in γ, then it is contained in an edge of γ. This is a choice that we have to make in order to achieve independence of the partition. The resulting operator is different if one assumes that s and γ do not intersect at an endpoint of s.

We have in general $V(s\cup\gamma)=V(\gamma)\cup V(s)\cup V(\gamma\cap s)=V(\gamma)\cup V(\gamma\cap s)$. The volume operator applied to a graph is a sum of the $\hat{V}_v$ for each vertex of the graph and $\hat{V}_v$ only depends on those edges of the graph which are incident at v as follows from (III.5). Therefore $h_s\hat{V}_v=\hat{V}_v h_s$ if s is not incident at v. It follows that

$$[h_s^{-1},\hat{v}]f=\sum_{v\in V(\gamma)}h_s^{-1}\hat{V}_v f-\sum_{v\in V(\gamma\cup_s)}\hat{V}_v h_s^{-1}f=\sum_{v\in V(\gamma)\cap\partial s}[h_s^{-1},\hat{V}_v]f-\sum_{v\in V(\gamma\cap s)-V(\gamma)}\hat{V}_v h_s^{-1}f, \tag{III.11}$$

and both sums in the last line have at most one nonvanishing term corresponding to an endpoint of s intersecting γ.

Case 1: $\gamma\cap s=\varnothing$. Then $\hat{l}_s^2 f=0$ as is immediate from (III.11).

Case 2: $p \in \gamma \cap s$. Subcase (a): p is neither a vertex of γ nor a kink of c and so only the second term in (III.11) survives. This implies that p is a trivalent vertex with only two independent tangent directions of the edges of γ and the segment s incident at p. The properties of the volume operator now imply that the contribution $\hat{V}_p$ of $\hat{v}$ vanishes and therefore $\hat{l}_p^2 f = 0$.

Subcase (b): p is not a vertex of γ but a kink of c. Since s is only one of the segments of c incident at p this case is not different from (a).

Subcase (c): p is a vertex of γ but not a kink of c. Now the result can be nonvanishing. We have for a vertex $p=v$ of γ that $\hat{l}_s^2 f = -8\,\mathrm{tr}([h_s, \hat{V}_v][h_s^{-1}, \hat{V}_v])f$, that is, only the first term in (III.11) survives. That the volume operator in the second commutator also reduces to $\hat{V}_v$ follows by a similar argument.

Subcase (d): p is both a vertex of γ and a kink of s. Again, this case does not differ from (c) for the same reason as (b) was equal to (a).

So we conclude that for large enough n the operator (III.8) reduces to

$$\hat{L}_n(c) := \not{\ell}_p \sum_{v \in V(\gamma)} \sum_{v \in s_i} \sqrt{-8\,\mathrm{tr}([h_{s_i}, \hat{V}_v][h_{s_i}^{-1}, \hat{V}_v])}, \tag{III.12}$$

where we have denoted the segment $s_i := c([\mathbf{t}_{i-1}, t_i])$. This demonstrates that $\hat{L}_n(c)$ is a finite operator for each n because there are at most $|V(\gamma)|$ terms in the sum (III.12). The next question is whether the limit $n \to \infty$ exists.

C. Existence of the limit $n \to \infty$

We now show that (III.12) is actually independent of the s_i so that the limit $n \to \infty$ is already taken and the limit therefore exists trivially.

In fact, s_i can be an arbitrarily ''short'' (we put this term in quotation marks because there is no background metric available with respect to which we could measure the length of s_i) segment of c starting at one and only one vertex v of γ.

Let us first recall the notion of a spin-network state.[20–22] We use the notation of Ref. 22: Label the edges e of a graph γ with nontrivial irreducible representations of SU(2), that is, assign to each of them a spin quantum number $j_e > 0$. If π_j is an irreducible representation of SU(2) with weight (or spin) j, then the spin-network state depends on $\pi_{j_e}(h_e)$. Further, assign to each vertex v of γ a contractor matrix m_v which contracts the matrix elements of the tensor product (over the set of edges e incident at v) of the $\pi_{j_e}(h_e)$ in such a way that the resulting state is gauge invariant. We label these spin-network states by $T_{\gamma,\mathbf{j},\mathbf{m}}$. One can show that spin-network states form an orthonormal basis of $\mathscr{H}$. Consider first the case that γ and $s := s_i$ intersect at only one point. Since $\hat{l}_s^2$ is gauge invariant, the result of applying $\hat{l}_s^2$ to a gauge-invariant cylindrical function must be a gauge-invariant cylindrical function f' which depends on the graph $s \cup \gamma$. Let us decompose f' into a basis of spin-network states. Since f did not depend on s, the spin assigned to s for a particular term T in the spin-network decomposition of f' can only be $j = \pm 1, 0$. In the case $j = 0$ the state T does not depend on s at all. In case that $j = \pm 1$, then T would be a spin-network state which contains an univalent vertex, namely the endpoint of s distinct from v, and there is only one edge, namely s, incident at it and colored with $j \neq 0$. Such a state is not gauge invariant. Therefore the case $j \neq 0$ does not appear.

Now consider the case that $s := s_i$ is contained in an edge e of γ and starts at a vertex v of γ. By choosing n high enough we may assume that s is properly contained in e, that is, $e' := e - s \neq \emptyset$. Without loss of generality we may assume that f is a spin-network state and thus e carries spin $j > 0$. Then in the spin-network decomposition of f' a particular term depends on $\gamma = \gamma \cup s$ in which the spins of all the edges of γ distinct from e are unchanged while e' carries spin j and s carries spin $j' \in \{j, j \pm 1\}$. Consider the divalent ''vertex'' $p = s \cap e'$ of γ. In the case that $j' = j$, then T depends on all of e with the same spin, i.e., it does not depend on s and e' with different spins. The only gauge-invariant way to contract the corresponding matrix elements at p is such that the resulting state depends on s, e' only through their product $e = s \circ e'$. In the case that $j' = j \pm 1$, then p is a two-valent vertex at which two edges e', s with distinct spins $j, j' = j \pm 1$ are incident. There exists no vertex contractor that makes such a state gauge invariant; therefore this case cannot appear.

This furnishes the proof that for a spin-network state f the function $\hat{L}_n(c)f$ still depends on all of the edges of γ if f did and that there is no dependence on the segments of the partition of c, which is to be expected since the classical length of a curve is a functional of E^a_i only. Thus, the limit $n\to\infty$ is already taken in (III.12).

D. Cylindrical consistency

Since we have adapted our partition to the graph on which a cylindrical function depends, although the operator (III.12) was derived, it is not obvious any more that the family of operators $(\hat{L}_\gamma(c))$ constructed actually line up and qualify as the projections of an operator on $\mathscr{H}$. So let $\gamma\subset\gamma'$. We need to show that (1) $(\hat{L}_{\gamma'}(c))_{|\gamma}=\hat{L}_\gamma(c)$, that is, the restriction of a projection to a smaller graph actually coincides with the projection to the smaller graph, and (2) the domain of $\hat{L}_\gamma(c)$ is contained in that of $\hat{L}_{\gamma'}(c)$. Issue (2) follows trivially by inspection if we choose the domain of $\hat{L}_\gamma(c)$ to be $\mathrm{Cyl}^3_\gamma(\overline{\mathscr{A}/\mathscr{G}})$, which we can since the volume operator has that domain. Issue (1) follows immediately: Given a curve c, the actions of $\hat{L}_\gamma(c)$ and $\hat{L}_{\gamma'}(c)$ on a function f cylindrical with respect to γ differ if either (a) γ' has more vertices intersecting c than γ or (b) there are additional edges of γ' incident at a common vertex of γ and γ' intersecting c. This follows directly by inspection from (III.12). So first let v be a vertex of γ' but not of γ and s be a segment of c incident of v. Then $[h_s^{-1},\ \hat{V}_v]f=-f\hat{V}_v h_s^{-1}=0$ and so this term in the sum (III.12) does not contribute. Now if case (b) occurs, then the cylindrical consistency of the volume operator assures that the edges of $\gamma'-\gamma$ incident at v do not contribute.

This furnishes the proof that (III.12) is consistently defined.

E. Symmetry and positivity

We wish to show that (1) every projection $\hat{L}_\gamma(c)$ is a symmetric operator on $\mathscr{H}$ with domain $\mathrm{Cyl}^3_\gamma(\overline{\mathscr{A}/\mathscr{G}})$ and (2) the family of projections $((\hat{L}_\gamma(c))_\gamma)$ comes from a positive semi-definite symmetric operator on $\mathscr{H}$ with domain $\mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$.

Lemma III.1: Every projection $\hat{L}_\gamma(c)$ defines a symmetric and positive semi-definite operator on $\mathscr{H}$ with domain $\mathscr{D}_\gamma:=\mathrm{Cyl}^3_\gamma(\overline{\mathscr{A}/\mathscr{G}})$.

Proof: Let us begin with $\hat{l}_s^2$. Since $\hat{V}_v$ is a symmetric operator on $\mathscr{H}$ we find for the adjoint of $[h_s^{-1},\ \hat{V}_v]$ using the unitarity of SU(2)

$$[h_s,\ \hat{V}_v]^\dagger=[\hat{V}_v,\ \overline{h_s}]=-[(h_s^{-1})^T,\ \hat{V}_v], \tag{III.13}$$

where $(.)^T$ denotes the transpose of (.). It follows that

$$\hat{l}_s^2:=+8\sum_{A,B}\hat{K}_{AB}(s)^\dagger\hat{K}_{AB}(s),\quad \text{where}\ \ \hat{K}_{AB}(s)=[(h_s^{-1})_{AB},\ \hat{V}], \tag{III.14}$$

which shows that $\hat{l}_s^2$ is a symmetric and positive semi-definite operator on $\mathscr{H}$ and therefore possesses a square root $\hat{l}_s$. The statement about the domain is a consequence of the fact that $\hat{L}_\gamma(c)$ is defined in terms of the spectral projections of $\hat{V}_\gamma$ which has domain $\mathscr{D}_\gamma$. □

Let us derive a more convenient expression for (III.12):

By means of the basic SL(2,ℂ) formula $g^{-1}_{AB}=\epsilon_{AC}\epsilon_{BD}g_{DC}$ we readily derive that $\hat{l}_s=\hat{l}_{s^{-1}}$. This motivates the following notation:

Let there be given an oriented curve c and a vertex v of γ that intersects it. If v is an interior point of c, consider any two segments $s^\pm_v$ of c which start at v and are ''short'' enough as not to intersect γ at any other of its vertices. If v is an endpoint of c, we need only one such segment $s^+_v=s^-_v$ starting at v. In (III.12) the segments s_i of c starting at vertices of γ come with the orientation of c. However, as we just showed, the operator $\hat{l}_s$ is invariant under reversal of the orientation of s and the result of applying it to a gauge-invariant function is independent of the choice of s as long as it is ''short'' enough. Therefore, let

$$N(v,c)=\begin{cases}0, & v\notin c,\\ \frac{1}{2}, & v \text{ is an point of } c,\\ 1, & v \text{ is an interior point of } c,\end{cases} \tag{III.15}$$

and define $\hat{L}_v^{\pm}(c):=\hat{l}_{s_v^{\pm}}(c)$. Then the length operator can be conveniently written

$$\hat{L}_\gamma(c)=\sum_{v\in V(\gamma)} N(v,c)[\hat{L}_v^+(c)+\hat{L}_v^-(c)]. \tag{III.16}$$

Theorem III.1: *The family* $(\hat{L}_\gamma(c),\mathscr{D}_\gamma)$ *defines a positive semi-definite symmetric operator on* $\mathscr{D}:=\mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$.

Proof: So far we have demonstrated $\hat{L}_\gamma(c)$ is a symmetric and positive semi-definite operator for each γ with domain $\mathscr{D}_\gamma=\mathrm{Cyl}^3_\gamma(\overline{\mathscr{A}/\mathscr{G}})$. However, to qualify as a symmetric projection of a symmetric operator requires more:

Let $\hat{S}$ be a symmetric operator on $\mathscr{H}$ with symmetric projections $\hat{S}_\gamma$. Let $f=p^*_\gamma f_\gamma$ and $g=p^*_{\gamma'}g_{\gamma'}$ be cylindrical functions. Then

$$\langle g,\hat{S}f\rangle=\langle g,\hat{S}_\gamma f\rangle=\langle \hat{S}_\gamma g,f\rangle\overset{!}{=}\langle \hat{S}g,f\rangle=\langle \hat{S}_{\gamma'}g,f\rangle. \tag{III.17}$$

Condition (III.17) is necessary and sufficient for a family of self-consistent symmetric projections to come from a symmetric operator. Let us check that (III.17) is satisfied for $\hat{L}(c)$ for each curve c.

It is sufficient to check it for the case that f,g are spin-network states. Now since $\hat{L}_\gamma(c)$ does not change the graph γ on which f depends nontrivially, it follows that $\langle g,\hat{L}_\gamma(c)f\rangle$ is nonvanishing if and only if $\gamma=\gamma'$ because spin-network states depending on different graphs are orthogonal.[20–22] The same holds for $\langle \hat{L}_{\gamma'}(c)g,f\rangle$. Symmetry now follows from the cylindrical consistency and the statement about the domain is because $\hat{V}$ has this domain. □

F. Self-adjointness

Although the expression of $\hat{L}_\gamma(c)f$ does not depend on the $s_v^{\pm}$ such that there must be a way to write purely in terms of the right invariant vector fields X_I very much like in the case of the volume operator (III.5), it is not clear to us how to do that. We can therefore not employ the essential self-adjointness of iX_I with core $\mathrm{Cyl}^1(\overline{\mathscr{A}/\mathscr{G}})$ in order to show that $\hat{L}_\gamma(c)$ is essentially self-adjoint as well with core $\mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$.

An independent proof which demonstrates that there exist self-adjoint extensions goes as follows:

Theorem III.2: *The operator* $(\hat{L}(c),\mathscr{D})$ *admits self-adjoint extensions on* $\mathscr{H}$.

Proof: The expression of $\hat{L}(c)$ is purely real and symmetric. Therefore it commutes with the antiunitary operator of complex conjugation. The assertion now follows from von Neumann's theorem (Ref. 23, p. 143). □

A possible choice of extension is given by its Friedrichs extension.

IV. SPECTRAL ANALYSIS OF THE LENGTH OPERATOR

A. Discreteness

The important feature of $\hat{L}_\gamma(c)$ is that it changes neither the graph γ on which a cylindrical function depends nor the irreducible representations of the corresponding spin-network states into which a cylindrical function f can be decomposed. Since the number of spin-network states associated with a fixed graph and a fixed coloring of its edges by irreducible representations is finite dimensional,[22] it follows that the length operator on each such subspace of $\mathscr{D}$ is just a symmetric, positive semi-definite, finite-dimensional matrix. Its eigenvalues are real and are given by certain non-negative functions $\lambda(\{j_I\})$ of the spins j_I associated with the edges of γ. Since spin

networks span $\mathcal{D}_\gamma$ and provide a countable basis for its completion $\mathcal{H}_\gamma$ (with respect to $\mu_{0,\gamma}$) it follows that there exists a countable basis of eigenvectors for $\mathcal{H}_\gamma$. The corresponding eigenvalues therefore form a countable set and lie in the point spectrum of $\hat{L}_\gamma(c)$.

Recall that a point in the spectrum is said to be in the discrete spectrum if it is an isolated point and an eigenvalue of finite multiplicity (clearly 0 has infinite multiplicity; all functions cylindrical with respect to graphs not intersecting the curve c are annihilated). In order to show that the point spectrum that $\hat{L}_\gamma(c)$ attains on $\mathcal{D}_\gamma$ is discrete it would be sufficient to show that $\lambda(\{j_I\})$ diverges whenever $j:=j_1+\cdots+j_n$, n being the number of edges of γ, diverges. Namely, if there was a finite condensation point in the point spectrum, then an infinite number of different n-tuples would give an eigenvalue which lies in a finite neighborhood of that condensation point but necessarily the corresponding values of j must diverge which would be a contradiction. The same argument shows that there would not be an eigenvalue of infinite multiplicity.

Looking at the eigenvalue formulas derived below for some simple graphs, such a behavior of λ is plausible because $\hat{L}_\gamma(c)$ is an unbounded operator; however, a strict proof is missing at this point.

Discreteness of the spectrum of $\hat{L}_\gamma(c)$ would follow immediately if we could prove it for the volume operator. An elegant method of proof would be to show that $\hat{V}$ is (a positive root of) an elliptic operator. Since this operator acts on the compact manifold $\mathrm{SU}(2)^n$, standard results form harmonic analysis would imply that its spectrum is discrete. Unfortunately this does not work: the volume operator is easily seen to be the fourth root of the operator $\hat{q}^\dagger\hat{q}$, $\hat{q}=\Sigma_{IJK}\epsilon_{IJK}\hat{q}_{IJK}$ (compare Ref. 7) and so is a sixth-order homogenous polynomial in the derivative operators X_I. Its principal symbol can be seen to be a non-negative function with an at least two-dimensional kernel and therefore is far from being invertible. Indeed, it is by now well known that the volume operator has a kernel which includes all divalent and trivalent graphs.

Summarizing, we have shown that $\hat{L}_\gamma(c)$ has pure point spectrum when restricted to $\mathcal{D}_\gamma$. Now consider the complete spectrum of $\hat{L}_\gamma(c)$ on all of $\mathcal{H}_\gamma$. Since $\hat{L}_\gamma(c)$ is a self-adjoint (not only symmetric) operator, it has spectral projections which are orthogonal for disjoint subsets of its spectrum. It follows that if I is a subset of the real numbers not contained in the part of the spectrum that $\hat{L}_\gamma(c)$ attains on $\mathcal{D}_\gamma$, and if $\hat{P}_I$ is the corresponding spectral projection, then $\mathcal{H}'_\gamma:=\hat{P}_I\mathcal{H}_\gamma$ and $\mathcal{D}_\gamma$ are orthogonal. However, $\mathcal{D}_\gamma$ is dense in $\mathcal{H}_\gamma$, therefore for each $v\in\mathcal{H}'_\gamma$, $\epsilon>0$ we find $f\in\mathcal{D}_\gamma$ such that $\|v-f\|<\epsilon$. On the other hand, by orthogonality $\|v-f\|\geqslant\|v\|$, which is a contradiction unless $v=0$. Therefore the complete spectrum of $\hat{L}_\gamma(c)$ is already attained on $\mathcal{D}_\gamma$ and it has no continuous part very much like its companions, the area and volume operators.[12,13] Thus, the spectrum on $\mathcal{H}_\gamma$ is discrete in the sense that it does not have a continuous part.

Similarly, it follows that the spectrum of $\hat{L}(c)$ is attained on $\mathcal{D}$ and also does not have a continuous part (this property is shared by all three operators, length, area, and volume): Although the set of piecewise analytic graphs is uncountable, the matrix elements of the length operator in a spin-network basis do not depend on the graph, they only depend on the quantum numbers **j**, **m** *and are therefore diffeomorphism invariant*. Therefore, the spectrum attained on $\mathcal{D}_\gamma$ and $\mathcal{D}_{\varphi(\gamma)}$, $\varphi\in\mathrm{Diff}(\Sigma)$ an analyticity preserving diffeomorphism, is identical. Moreover, the spectrum does not depend on whether two curves touch each other in a C^m or C^n fashion for any $1\leqslant m<n$. It follows that the spectrum only depends on the C^∞ properties of the diffeomorphism class of a graph, that is, all that matters is whether two edges are distinct or coincide. In other words, all we need to know about a graph is

(a) the number N of its vertices,
(b) the valence $n(v)$ of each vertex v, and
(c) the topologically different ways of connecting edges incident at different vertices in a C^∞ manner which is a finite number.

Thus, since this characterization of a graph depends on discrete labels and the union of countable sets is a countable set, it follows that the spectrum is still discrete in the sense that it does not have a continuous part. On the other hand, we see that every eigenvalue of the length, area, or volume operators is of uncountably infinite multiplicity when we consider all of $\mathcal{H}$.

Given the results stated in this section we can now establish that the length operator is even essentially self-adjoint with core $\mathrm{Cyl}^3(\overline{\mathcal{A}/\mathcal{G}})$. To see this we note that the Hilbert space is the completion of an uncountable direct sum

$$\mathcal{H}=\overline{\oplus_{\gamma,\mathbf{j}}\mathcal{H}_{\gamma,\mathbf{j}}},$$

where the sum is over all graphs and all colorings of the edges of the graph. As said above, each $\mathcal{H}_{\gamma,\mathbf{j}}$ is finite dimensional and each $\hat{L}_\gamma$ is self-adjoint on $\mathcal{H}_{\gamma,\mathbf{j}}$ so that $\hat{L}_\gamma \pm i$ id has dense range $D_{\gamma,\mathbf{j}}=\mathcal{H}_{\gamma,\mathbf{j}}$. We now compute

$$[\hat{L}(c)\pm \mathrm{iid}](\oplus_{\gamma,\mathbf{j}}\mathcal{H}_{\gamma,\mathbf{j}})=\oplus_{\gamma,\mathbf{j}}[\hat{L}_\gamma(c)\pm \mathrm{iid}].\mathcal{H}_{\gamma,\mathbf{j}}=\oplus_{\gamma,\mathbf{j}}\mathcal{H}_{\gamma,\mathbf{j}}\subset[\hat{L}(c)\pm \mathrm{iid}].\mathcal{H}.$$

However, the left-hand side is dense in $\mathcal{H}$. So, $\hat{L}(c)$ satisfies the basic criterion for essential self-adjointness. Actually this result follows trivially from general results on cylindrically consistent families of essentially self-adjoint operators[19] where the cylindrical subspaces are now labeled by γ, $\mathbf{j}$.

B. Eigenvalues

By inspection, the task of giving a closed formula for the eigenvalues of the length operator requires having a closed formula for the spectrum of the volume operator at one's disposal which we lack, however, at the present stage (see, however, Refs. 24 and 25 for a closed formula for its matrix elements). We will therefore restrict ourselves here to computing the spectrum for some simple types of graphs and thereby obtain the quantum of length.

Specifically, it is comparatively simple to compute the spectrum of the length operator when restricted to at most trivalent graphs, thus including the classical spin networks which were introduced in Ref. 15 and whose vertices are all precisely trivalent. Since for trivalent graphs the vertex contractors of a spin network are unique and since the result of applying the length operator to a spin-network state is a spin-network state on the same graph and with the same spin, it follows that all spin-network states on trivalent graphs are eigenvectors of the length operator. We will see that the corresponding eigenvalues are nonvanishing in general. This is an astonishing feature because the volume operator is known to vanish on trivalent graphs. Now, classically the volume of a region is known if one can measure the lengths of arbitrary curves through that region and so nonvanishing lengths of curves result in nonvanishing volume. This indicates that trivalent graphs are rather special and are insufficient to construct weave states that approximate a given classical geometry.

The computations are largely governed by the properties of the $6j$-symbol of the recoupling theory of angular momentum which we recall in the Appendix. The way the recoupling theory of spin systems enters the stage is as follows: in our computations we evaluate the volume operator at a given vertex v on functions f which transform according to an irreducible representation j of SU(2) under gauge transformations at v. This j is nothing other than the resulting total angular momentum to which the angular momenta j_I (which color the edges e_I incident at v) couple. There are different ways to couple angular momenta j_I to resulting spin j, and this freedom is determined by the recoupling quantum numbers $j_{IJ}, j_{IJK},\ldots$ (compare the Appendix) and results in different contractors of the not necessarily gauge-invariant (or extended) spin networks. Here we mean by an extended spin network just any function that depends on the edges of a graph through the matrix elements of irreducible representations of SU(2) evaluated at the holonomy of the corresponding edge and transforms at each vertex according to an irreducible representation of SU(2) under gauge transformations. The point is now that the function f is easily seen to be in the left regular representation of $\mathrm{SU}(2)^n$, n being the valence of the vertex v, defined by $\hat{R}(g_1,\ldots,g_n)f(h_1,\ldots,h_n)=f(g_1h_1,\ldots,g_nh_n)$. On the other hand, the connection with the abstract angular momentum Hilbert space where SU(2) acts by the abstract unitary representation $\hat{U}(g)$ and which is spanned by the vectors of the form $|(j_1,m_1),\ldots,(j_n,m_m);k_1,\ldots,k_n\rangle=|j_1,m_1;k_1\rangle \otimes\cdots\otimes|j_n,m_n;k_n\rangle$, where k_I are certain additional quantum numbers, is made as follows: Consider the special functions $(\pi_j(h))_{m,m'}$ given by the matrix elements of the jth irreducible representation of SU(2), $m,m'\in\{-j,-j+1,\ldots,j\}$. Consider the Hilbert space $L_2(\mathrm{SU}(2),d\mu_H)$ and let $|g\rangle$ be the usual Dirac generalized eigenstates of the multiplication operator $\hat{g}$. Then

$(\pi_j(h))_{m,m'}=\langle h|j,m;m'\rangle$. The proof is by checking the representation property $\hat{R}(g)(\pi_j(h))_{m,m'}=(\pi_j(gh))_{m,m'}=(\pi_j(g))_{m,\tilde{m}}(\pi_j(h))_{\tilde{m},m'}=\langle h|\hat{U}(g)|j,m;m'\rangle$ by definition of the states $|j,m;m'\rangle$. The reader is referred to Ref. 25 for more details.

It follows from these considerations that instead of doing tedious computations within the left regular representation (spin-network states or traces of the holonomy around closed loops[10]), it is far easier to do them in the abstract representation which allows the use of the powerful Clebsh–Gordan theory of angular momentum. We will do that in what follows. In particular, it follows immediately that the right invariant vector field X_I is identified with $2iJ_I$, where J_I is the angular momentum operator of the spin associated with the Ith edge which is the self-adjoint generator of the unitary group $\hat{U}(g)$.

We are now in the position to compute the spectrum of the length operator on trivalent graphs γ. First we cast the volume operator in the more convenient form

$$\hat{V}=\not{\sum}_p^3\hat{v}f=\not{\sum}_p^3\sum_{v\in V(\gamma)}\hat{V}_v \quad \text{where} \quad \hat{V}_v:=\sqrt{\left|\frac{i}{32}\hat{q}_v\right|},$$
$$\hat{q}_v=\sum_{e_I\cup e_J\cup e_K=v,I<J<K}\epsilon(e_I,e_J,e_K)\hat{q}_{IJK}, \quad \text{where} \quad \hat{q}_{IJK}:=[J_{IJ}^2,J_{JK}^2]. \tag{IV.1}$$

To arrive at this expression one only has to use elementary angular momentum algebra. Expression (IV.1) captures a neat interpretation of the volume operator: it measures the difference between recoupling schemes of n angular momenta based on J_{IJ} and J_{JK}, respectively. This is why the recoupling theory is important and the matrix elements of the volume operator can be given purely in terms of polynomials of $6j$-symbols.[25]

Now let v be a vertex of γ which intersects c. From (III.16) we find $[h:=h_{s_v^\pm},\ \hat{L}_v(c):=\hat{L}_v^\pm(c)]$

$$\tfrac{1}{8}\hat{L}_v(c)^2=-[\mathrm{tr}(h^{-1}\hat{V}h)\hat{V}+\hat{V}\,\mathrm{tr}(h^{-1}\hat{V}h)]+2\hat{V}^2+\mathrm{tr}(h^{-1}\hat{V}^2h), \tag{IV.2}$$

and the simplification that occurs on trivalent graphs is that the first three terms vanish identically. We have two possibilities:

case A: s_v lies within an edge of γ incident at v or

case B: s_v is not contained in an edge of γ.

We will discuss both cases separately.

1. Case A

We may without loss of generality assume that s_v coincides with one, say e_3, of the three edges of γ incident at v. This is because (1) if $e=s_v\circ e'$, then for any irreducible representation π_j of SU(2) we have $\pi_j(h_e)=\pi_j(h_{s_v})\pi_j(h_{e'})$ so a spin-network state also depends on s_v through π_j and (2) for a right invariant vector field $X(h_{s_v})=X(h_{s_v}h_{e'})=X(h_e)$ by definition of right invariance. Therefore the volume operator (IV.1) contains only one term $\hat{q}_{123}$.

Consider a trivalent spin-network function $f=T_{\gamma,\mathbf{j}}$ with spins j_1,j_2,j_3 assigned to e_1,e_2,e_3 (we have suppressed the contractor matrices $\mathbf{m}$ since they are unique for trivalent spin networks). Then $h_{e_3}f$ can be decomposed into extended spin network functions with total angular momentum $j=\frac{1}{2}$ and edge spins $j_1'=j_1,j_2'=j_2,j_3'=j_3\pm\frac{1}{2}$ by usual Clebsh–Gordan representation theory. Let us determine the precise coefficients of that decomposition.

Lemma IV.1: Let $n=2j$ so that $\pi_j(g)_{A_1,\ldots,A_n;B_1,\ldots,B_n}=g_{(A_1,B_1}\cdots g_{A_n),B_n}$ where the round bracket means total symmetrization in the A indices. Then

$$g_{A_0,B_0}\pi_j(g)_{A_1,\ldots,A_n;B_1,\ldots,B_n}=\pi_{j+1/2}(g)_{A_0,\ldots,A_n;B_0,\ldots,B_n}$$
$$-\frac{n}{n+1}\epsilon_{A_0(A_1}\pi_{j-1/2}(g)_{A_2,\ldots,A_n);(B_2,\ldots,B_n}\epsilon_{B_1)B_0}. \tag{IV.3}$$

Proof: Elementary linear algebra. □

We now multiply each of the two terms in (IV.3) with $g^{-1}_{B_0,A_0}$ and sum over A_0, B_0. The result is

$$g^{-1}_{B_0,A_0}\pi_{j+1/2}(g)_{A_0;\ldots,A_n;B_0,\ldots,B_n}=\frac{n+2}{n+1}\,\pi_j(g)_{A_1,\ldots,A_n;B_1,\ldots,B_n},$$
$$g^{-1}_{B_0,A_0}\epsilon_{A_0(A_1}\pi_{j-1/2}(g)_{A_2,\ldots,A_n);(B_n,\ldots,B_n}\epsilon_{B_1)B_0}=-\pi_j(g)_{A_1,\ldots,A_n;B_1,\ldots,B_n}. \tag{IV.4}$$

Formulas (IV.4) illustrate the computational reason for why the edge s_v does not appear in a gauge invariant function f after evaluating $\hat{L}(c)f$ on it. We now can write $h_{e_3}f=f_+ +[n_3/(n_3+1)]f_-$ where, according to (IV.3), the vectors f_+ and f_- respectively are proportional to the vectors $|j_{12}=j_3,j=\frac{1}{2};j_1,j_2,j_3+\frac{1}{2}\rangle$ and $|j_{12}=j_3,j=\frac{1}{2};j_1,j_2,j_3-\frac{1}{2}\rangle$, respectively, on the abstract angular momentum Hilbert space. That in both vectors j_{12} still equals j_3 follows from the fact that we did not change the way the matrices are contracted in f in multiplying it by h_{e_3}.

The space of states V_+ with total angular momentum $j=\frac{1}{2}$ and spins $j_1,j_2,j_3'=j_3+\frac{1}{2}$ is two-dimensional: it is spanned by $|j_{12}=j_3,j=\frac{1}{2};j_1,j_2,j_3+\frac{1}{2}\rangle$ and $|j_{12}=j_3+1,j=\frac{1}{2};j_1,j_2,j_3+\frac{1}{2}\rangle$, respectively. Likewise, the span of the space of states V_- with total angular momentum $j=\frac{1}{2}$ and spins $j_1,j_2,j_3'=j_3-\frac{1}{2}$ is given by $|j_{12}=j_3,j=\frac{1}{2};j_1,j_2,j_3-\frac{1}{2}\rangle$ and, provided that $j_3\geqslant 1$, $|j_{12}=j_3-1,j=1/2;j_1,j_2,j_3-\frac{1}{2}\rangle$, respectively. The volume operator leaves these two spaces separately invariant and the operator $\hat{q}_{123}$, being antisymmetric on $\mathscr{H}$, reduces to an antisymmetric 2×2 matrix on V_+ and to an antisymmetric 2×2 matrix on V_- if $j_3\geqslant 1$ and to the zero matrix if $j_3<1$. Let the matrix elements different from zero of these matrices be denoted by $\pm\mu_+$, $\pm\mu_-$, respectively. Then, since we take the modulus of the square root of $\hat{q}_{123}$ it follows that $f_\pm$ are already eigenvectors of $\hat{V}$ with eigenvalues $\lambda_\pm:=\sqrt{|\mu_\pm|}/4\sqrt{2}$. It follows then from (IV.4) and $h^{-1}\hat{V}^2h=(h^{-1}\hat{V}h)^2$ that our spin-network state is an eigenfunction of the length operator $\hat{L}_v(c)$ with eigenvalue $(\ell_p/2)\sqrt{[1/(n_3+1)][(n_3+2)\lambda_+^2+n_3\lambda_-^2]}$.

It remains to compute the matrix elements $\mu_\pm$ of $\hat{q}_{123}=[J^2_{12},J^2_{23}]$. We have

$$\begin{aligned}\mu_+&:=\langle j_{12}'=j_3+1,j=\tfrac{1}{2};j_1,j_2,j_3'=j_3+\tfrac{1}{2}|\hat{q}_{123}|j_{12}=j_3,j=\tfrac{1}{2};j_1,j_2,j_3'=j_3+\tfrac{1}{2}\rangle\\
&=[(j_3+1)(j_3+2)-j_3(j_3+1)]\langle j_{12}'=j_3+1,j=\tfrac{1}{2}|J^2_{23}|j_{12}=j_3,j=\tfrac{1}{2}\rangle\\
&=2(j_3+1)\sum_{j_{23}=j_1\pm 1/2}j_{23}(j_{23}+1)\langle j_{12}'=j_3+1,j=\tfrac{1}{2}|j_{23},j=\tfrac{1}{2}\rangle\times\langle j_{23},j=\tfrac{1}{2}|j_{12}=j_3,j=\tfrac{1}{2}\rangle,\end{aligned} \tag{IV.5}$$

where we have suppressed $j_1,j_2,j_3'=j_3+\frac{1}{2}$. In the last step we have inserted a complete 1 in form of the coupling scheme $(j_2,j_3')\to j_{23}$, $(j_1,j_{23})\to j$ and realized that the only possible values for j_{23} in order to couple to spin $j=\frac{1}{2}$ with j_1 are the ones displayed. Completely analogously we find

$$\begin{aligned}\mu_-&:=\langle j_{12}'=j_3-1,j=\tfrac{1}{2};j_1,j_2,j_3'=j_3-\tfrac{1}{2}|\hat{q}_{123}|j_12=j_3,j=\tfrac{1}{2};j_1,j_2,j_3'=j_3-\tfrac{1}{2}\rangle\\
&=[(j_3-1)j_3-j_3(j_3+1)]\langle j_{12}'=j_3-1,j=\tfrac{1}{2}|J^2_{23}|j_{12}=j_3,j=\tfrac{1}{2}\rangle\\
&=-2j_3\sum_{j_{23}=j_1\pm 1/2}j_{23}(j_{23}+1)\langle j_{12}'=j_3-1,j=\tfrac{1}{2}|j_{23},j=\tfrac{1}{2}\rangle\langle j_{23},j=\tfrac{1}{2}|j_{12}=j_3,j=\tfrac{1}{2}\rangle.\end{aligned} \tag{IV.6}$$

We now use the formulas given in the Appendix to compute the eight remaining matrix elements in terms of $6j$-symbols and evaluate the latter by means of the Racah formula. The result is

$$\mu_+ = \sqrt{(a+b+c)(a+b-c)(b+c-a)(c+a-b)},$$
$$\text{where } a=j_1=\tfrac{1}{2}, b=j_2+\tfrac{1}{2}, \text{ and } c=j_3+1,$$
$$\mu_- = -\sqrt{(a+b+c)(a+b-c)(b+c-a)(c+a-b)}, \tag{IV.7}$$
$$\text{where } a=j_1+\tfrac{1}{2}, b=j_2+\tfrac{1}{2}, c=j_3.$$

The final result is thus given by the following theorem.

Theorem IV.1: *The eigenvalue* λ *of* $\hat{L}_v(c)$ *for a trivalent spin-network state* T_γ *with vertex at* v *such that* c *and* γ *share a segment incident at* v *is given by* ($j_3 \in \{j_1+j_2, j_1+j_2-1, \ldots, |j_1-j_2|\}$)

$$\lambda = \frac{l_p}{2\sqrt{j_3+1/2}} \sqrt{(j_3+1)\sqrt{(j_1+j_2+j_3+2)(j_1+j_2-j_3)(j_2+j_3-j_1+1)(j_3+j_1-j_2+1)}}$$
$$\overline{+ j_3\sqrt{(j_1+j_2+j_3+1)(j_1+j_2-j_3+1)(j_2+j_3-j_1)(j_3+j_1-j_2)}}, \tag{IV.8}$$

provided the edges e_1, e_2, e_3 *of* γ *incident at* v *do not have colinear tangents at* v, *otherwise* $\lambda = 0$.

The formula for λ simplifies if we take v to be a divalent vertex not sharing a segment with c. Then $j_3=0$ and necessarily $j_1=j_2=j_0$, and we find simply

$$\lambda(j_0) = l_p \sqrt[4]{j_0(j_0+1)}$$

which for large spin grows as $\sqrt{j_0}$. In other words, the warping of Σ when the gravitational field is in the spin-network state labelled by j_0 grows with the spin j_0 since one and the same curve c appears longer the more spin j_0 we have in a neighborhood of it.

The quantum of length is achieved for $j_0=\frac{1}{2}$ and is given by

$$\frac{1}{\sqrt{2}} \sqrt[4]{3} l_p .$$

The length can change only in packets of $\Delta L \approx \pm \frac{1}{2}\sqrt{1/j_0} l_p$, which approaches zero for large spin so that for high spin the length looks like a continuous operator.

2. Case B

As this is actually a four-valent problem, the computations will be much more elaborate than in the previous case since we now have to couple four angular momenta which will necessarily invoke the $9j$-symbol. Fortunately one can reduce the $3(n-1)j$-symbol to a polynomial in $6j$-symbols, n the valence of the vertex.

The segment $e_4 := s_v$ carries spin $j_4=\frac{1}{2}$. First we note that if again f is a spin-network state with vertex v intersecting c, then we have [$\epsilon_{IJK} := \epsilon(e_I, e_J, e_K)$]

$$\hat{q}_v h_{e_4} f = [\epsilon_{124}\hat{q}_{124} + \epsilon_{234}\hat{q}_{234} + \epsilon_{314}\hat{q}_{314}] h_{e_4} f + h_{e_4}\epsilon_{123}\hat{q}_{123} f = [\epsilon_{124}+\epsilon_{234}+\epsilon_{124}]\hat{q}_{124} h_{e_4} f, \tag{IV.9}$$

where in the second step we exploited gauge invariance of f, that is, in infinitesimal form $[X_1 + X_2 + X_3] f = 0$.

The extended spin-network function $h_{e_4} f$ is represented on the abstract angular momentum Hilbert space by a vector proportional to $\psi := |j_{12}=j_3, j_{123}=0, j=\frac{1}{2}; j_1, j_2, j_3, j_4=\frac{1}{2}\rangle$, where the notation means that the operator $[(J_1+J_2+J_3)^i]^2$ is diagonal as well, the coupling scheme being given by $(j_1, j_2) \to j_{12}$, $(j_{12}, j_3) \to j_{123}$, $(j_{123}, j_4) \to j$.

The space of states with given values $j_1, j_2, j_3, j_4 = j = \frac{1}{2}$ is easily seen to be three-dimensional and is spanned by ψ, ψ_+, ψ_-, where $\psi_\pm = |j_{12}=j_3 \pm 1, j_{123}=1, j=\frac{1}{2}; j_1, j_2, j_3, j_4=\frac{1}{2}\rangle$. This is because $j=1/2$ requires $j_{123}=0,1$ and $j_{123}=0$ enforces $j_{12}=j_3$ while $j_{123}=1$ enforces $j_{12}=j_3 \pm 1$ by

the usual Clebsh–Gordan decomposition into irreducibles $(j)\otimes(j')=(j+j')\oplus(j+j'-1)\oplus\cdots\oplus(|j-j'|)$. The task left to do is to compute the matrix elements of $\hat{q}_v$ between $\psi,\psi_\pm$. Notice that the simplification that occurred in the second step of (IV.9) does not apply to the states $\psi_\pm$ because $j_{123}\neq 0$, so we need to use the full expression of $\hat{q}_v$ for the matrix element between ψ_+,ψ_- while for the other two we just need $q_{124}=[J_{12}^2, J_{24}^2]$. Let us begin with $\hat{q}_{124}$. In the following computation we are going to insert a complete 1 in the form of the basis corresponding to the coupling scheme $(j_2,j_4)\to j_{24}$, $(j_1,j_{24})\to j_{124}$, $(j_{124},j_3)\to j$ in order to evaluate J_{24}^2. We abbreviate $|j_{12}=j_3,j_{123}=0,j=\frac{1}{2};j_1,j_2,j_3,j_4=\frac{1}{2}\rangle=:|j_{12},j_{123}\rangle$, etc., and have for any values of the various j's

$$\begin{aligned}\langle j'_{12},j'_{123}|\hat{q}_{124}|j_{12},j_{123}\rangle&=[j'_{12}(j'_{12}+1)-j_{12}(j_{12}+1)]\langle j'_{12},j'_{123}|J_{24}^2|j_{12},j_{123}\rangle\\&=[j'_{12}(j'_{12}+1)-j_{12}(j_{12}+1)]\\&\quad\times\sum_{j_{24},j_{124}} j_{24}(j_{24}+1)\langle j'_{12},j'_{123}|j_{24},j_{124}\rangle\langle j_{24},j_{124}|j_{12},j_{123}\rangle,\end{aligned}\tag{IV.10}$$

where the allowed values for j_{24},j_{124} are $j_{124}=j_3\pm\frac{1}{2}$ in order to couple with j_3 the resulting $j=\frac{1}{2}$, and $j_{24}=j_2\pm\frac{1}{2}$ are the possible irreducible representations contained in $(j_2)\otimes(j_4)$. The next step is to reduce the matrix elements to $6j$-symbols. Note that the matrix elements are not of the standard textbook form $\langle j_{12},j_{34}|j_{13},j_{24}\rangle$ so that we need to invent a new reduction of the $9j$-symbol as a product of $6j$-symbols. Upon inserting a complete 1 corresponding to a basis of yet another coupling scheme we have

$$\begin{aligned}\langle j_{12},j_{123}|j_{24},j_{124}\rangle&\equiv\langle j_{12}(j_1,j_2),j_{123}(j_{12},j_3),j(j_{123},j_4)|j_{24}(j_2,j_4),j_{124}(j_1,j_{24}),j(j_{124},j_3)\rangle\\&=\sum_{j'_{124}}\langle j_{12}(j_1,j_2),j_{123}(j_{12},j_3),j(j_{123},j_4)\|j_{12}(j_1,j_2),j'_{124}(j_{12},j_4),j(j_{124},j_3)\rangle\\&\quad\times\langle j_{12}(j_1,j_2),j'_{124}(j_{12},j_4),j(j_{124},j_3)|j_{24}(j_2,j_4),j_{124}(j_1,j_{24}),j(j_{124},j_3)\rangle\\&=\langle j_{123}(j_{12},j_3),j(j_{123},j_4)\|j_{124}(j_{12},j_4),j(j_{124},j_3)\rangle\\&\quad\times\langle j_{12}(j_1,j_2),j_{124}(j_{12},j_4)|j_{24}(j_2,j_4),j_{124}(j_1,j_{24})\rangle.\end{aligned}\tag{IV.11}$$

In going from the third equality to the fourth equality, two things happened: (a) Because j_{124} is diagonal on the vectors involved in the second scalar product factor in the third line, the sum reduces to only one term $j'_{124}=j_{124}$, and (b) we used the Wigner–Eckart theorem to reduce the matrix elements,[26] that is, we could get rid of the first entry of the vectors involved in the first scalar product factor and of the last entry in the second, both in the third line. Equation (IV.11) is the form that allows us to express everything in terms of $6j$-symbols. Namely we have

$$\begin{aligned}&\langle j_{123}(j_{12},j_3),j(j_{123},j_4)|j_{124}(j_{12},j_4),j(j_{124},j_3)\rangle\\&\quad=(-1)^{j_1+j_2+j_4+j_{124}}\sqrt{(2j_{12}+1)(2j_{124}+1)}\begin{Bmatrix}j_1 & j_2 & j_{12}\\ j_4 & j_{124} & j_{24}\end{Bmatrix},\end{aligned}\tag{IV.12}$$

$$\begin{aligned}&\langle j_{12}(j_1,j_2),j_{124}(j_{12},j_4)|j_{24}(j_2,j_4),j_{124}(j_1,j_{24})\rangle\\&\quad=(-1)^{2(j_3+j_{12})+j_4+j-j_{123}}\sqrt{(2j_{123}+1)(2j_{124}+1)}\begin{Bmatrix}j_3 & j_{12} & j_{123}\\ j_4 & j & j_{124}\end{Bmatrix}.\end{aligned}\tag{IV.13}$$

In order to see this we do the following: In (IV.12) set $j'_1:=j_1$, $j'_2:=j_2$, $j'_3:=j_4$, $j':=j_{124}$, $j'_{12}:=j_{12}$, $j'_{23}:=j_{24}$ and use the standard formula (A1) for the $6j$-symbol given in the Appendix in terms of the primed j's. In (IV.13) we first set $j'_1:=j_3$, $j'_2:=j_{12}$, $j'_3:=j_4$, $j':=j$, $j'_{12}:=j_{123}$, $j'_{23}:=j_{124}$ and secondly recall the identity $|j_{12}(j_2,j_1),j\rangle=(-1)^{j_1+j_2-j_{12}}|j_{12}(j_1,j_2),j\rangle$ in order to apply the standard formula in terms of the primed j's.

As the whole expression becomes rather lengthy we refrain from writing the matrix elements (IV.10) out explicitly in terms of j_1,j_2,j_3. Rather, we consider them as known through (IV.10)–(IV.14) (and similar relations for the other $\hat{q}_{IJK}$) and define

$$a:=\langle\psi|\hat{q}_v|\psi_+\rangle,\quad b:=\langle\psi|\hat{q}_v|\psi_-\rangle,\quad c:=\langle\psi_+|\hat{q}_v|\psi_-\rangle. \tag{IV.14}$$

So, $\hat{q}_v$ reduces to an antisymmetric 3×3 matrix with nonvanishing off-diagonal entries $\pm a$, $\pm b,\pm c$, eigenvalues $0,\pm i\mu$ where $\mu=\sqrt{a^2+b^2+c^2}$ and eigenvectors $\psi^0,\psi^\pm$ which are chosen to be

$$\psi^0=c\psi-b\psi_++a\psi_-,\quad \psi^\pm=[\mp ib\mu-ac]\psi+[-\mp ic\mu+ab]\psi_++[b^2+c^2]\psi_-, \tag{IV.15}$$

and inverted:

$$\psi_-=\frac{1}{2\mu^2}[\psi^++\psi^-+2a\psi^0],$$

$$-[b^2+c^2]\psi_+=\frac{b}{\mu^2}[(b^2+c^2)\psi^0-(\psi^++\psi^-)/2]+\frac{ib}{2\mu}[\psi^+-\psi^-], \tag{IV.16}$$

$$+[b^2+c^2]\psi=\frac{c}{\mu^2}[(b^2+c^2)\psi^0-(\psi^++\psi^-)/2]-\frac{ic}{2\mu}[\psi^+-\psi^-].$$

Now let $\sigma:=\sqrt{\mu/32}$, σ being the eigenvalue of $\hat{V}$ on both $\psi^\pm$ while it vanishes on ψ^0. It is then straightforward to check that

$$\hat{V}\psi=\sigma\left[\psi-\frac{c}{\mu^2}\,\psi^0\right]=\sigma\,\frac{a^2+b^2}{\mu^2}\,\psi+\text{terms}\propto\psi_\pm. \tag{IV.17}$$

We now translate this result back into the representation $L_2(\overline{\mathscr{A}/\mathscr{G}},d\mu_0)$, multiply by $h_{e_4}^{-1}$, and take the trace. The result must be a gauge-invariant vector with no dependence on e_4, that is, $j=j_4=0$. It follows that $\mathrm{tr}(h_{e_4}\psi_\pm)=0$ because in writing $h_{e_4}\psi_\pm$ in terms of extended spin-networks we cannot get $j=0$ since the contractor corresponding to $j_{123}=1$ is unchanged. Therefore, we have the following.

Theorem IV.2: *The eigenvalue of $\hat{L}_v(c)$ on a trivalent spin-network state T_γ with vertex v such that c and γ do not share a segment incident at v is given by*

$$\hat{L}_v(c)T_\gamma=l_p\sqrt{\mu}\,\frac{a^2+b^2}{2\mu^2}\,T_\gamma,\quad \mu:=\sqrt{a^2+b^2+c^2}, \tag{IV.18}$$

where a,b,c are given through (IV.10)–(IV.14).

V. TUBE OPERATOR AND WEAVES

The length operator has a strange feature unshared by the area and volume operators: Ultimately, in a semiclassical approximation, one wishes to construct states which approximate a fixed classical three-geometry (Σ,q_{ab}). Such eigenstates have been called ''weaves'' in the literature (see, for instance, Ref. 27). Most of these eigenstates typically only involve linked, rather than intersecting, loops, and since the length and volume operators automatically annihilate those states, we need to construct more general weave states for which the name ''lattice states'' is more appropriate.

It is clear that a state that approximates such a geometry has to be defined on a (infinite, in case Σ is not compact) graph filling all of Σ in the following sense: consider a parameter ϵ which we may think of as a lattice spacing and which characterizes the average distance (as measured by q_{ab}) between neighboring vertices of the graph. In order to serve as a good approximation, the parameter needs to be small as compared to the scales of macroscopic objects. Consider any

macroscopic volume or area in Σ. It follows that the graph necessarily intersects these objects provided that ϵ is small enough and it intersects it the more often the smaller ϵ is. Each intersection makes a contribution no matter whether the intersection point is a vertex or not. On the other hand, even for the most random distribution of vertices with mean separation ϵ, even if ϵ is much smaller than the length of a given curve, the curve genuinely almost never intersects the graph at a vertex and so the quantum length of the curve would be always much smaller than its classical length.

This is obviously not what we want and so the length operator constructed cannot be directly associated with the length of a given classical curve. Following the old physical argument that an object always has to have an extension of at least one Planck volume in order not to form a black hole, the conclusion is that a one-dimensional curve cannot correspond to a physical object. This motivates the following definition.

Definition V.1: A tube C with a curve c as center is a two-parameter congruence of curves $C:(r,s)\in[0,1]\times[-\pi.\pi]\to C_{r,s}$ *such that* $C_{0,s}=c\forall s\in[-\pi,\ \pi]$.

So, the picture is that we have a congruence of curves, all of which look like c and which fill a three-dimensional region in Σ homeomorphic to a solid cylinder.

We are now ready to define a tube operator.

Definition V.2: The tube operator is given by

$$\hat{L}(C):=\sum_{(r,s)\in[0,1]\times[-\pi,\pi]}\hat{L}(C_{r,s}). \tag{V.1}$$

Notice that there is an uncountable sum involved in (V.1).

We would like to assign a length to a tube C that is associated somehow with the length of its central curve c. The basic problem is that it makes sense to speak of a length of a three-dimensional object only if it is in some sense one-dimensional. In terms of a tube, one would like to say that it is ''thin'' in the sense that the curves transversal to the central curve have small length as compared to the central one. Obviously, the concept of thinness then depends on the state, and while the tube operator is state independently defined its expectation value will be a good estimate of the length of the central curve only for states for which it is thin.

Let us make precise what we have just said.

Definition V.3: (i) Given a tube C and a state ψ *we define a real-valued function* $L(C,\psi,.)$ *on* $[0,1]$ *given by*

$$L(C,\psi,t):=\sum_{r,s}\ \frac{\langle\psi,\hat{L}(C_{r,s}([0,t]))\psi\rangle}{\|\psi\|^2},$$

where $C_{r,s}([0,t])$ *is the curve* $C_{r,s}$ *followed only until parameter value* t.

(ii) The function $L(C,\psi,.)$ *has discontinuities at a finite number n of points* $0\leqslant t_1<t_2<\cdots<t_n\leqslant 1$ *corresponding to the fact that new vertices of the graph* γ *underlying* ψ *enter the region filled by the* $C_{r,s}([0,t])$ *for all* r,s *as we increase* t. *The number n is finite as otherwise the state* ψ *will not be in the domain of the volume operator. Upon varying the parametrization of the tube we can arrange that at each* t_k *precisely one new vertex enters the tube chopped at value* t_k *and so n equals the number of vertices of* γ *that lie in the tube. Denote by* $p=\{p_i\}_{i=0}^n$ *a partition adapted to these discontinuities in the sense that* $0=p_0\leqslant t_1<p_1<t_2<p_2<\cdots<p_{n-1}<t_n\leqslant p_n=1$.

(iii) Given p, consider the family of n cylinders

$$C_k(C,\psi,p):=C([0,1]\times[-\pi,\pi]\times[p_{k-1},p_k]),\quad k=1,\ldots,n,$$

and let $S_k(C,\psi,p)$ *be the set of smooth curves e satisfying*

(a) *e is not self-intersecting,*
(b) *e lies in the cylinder* C_k, *that is,* $e\subset C_k(C,\psi,p)$, *and*
(c) *e starts and ends on the boundary of the cylinder* C_k, *that is,* $\partial e\subset\partial C_k(C,\psi,p)$.

(iv) Finally we define the following quantities,

$$\Delta_k(C,\psi) := \sup_p \sup_{e \in S_k(C,\psi,p)} \frac{\langle \psi, \hat{L}(e)\psi\rangle}{\|\psi\|^2},$$

$$\delta(C,\psi) := \inf_k \Delta_k(C,\psi), \tag{V.2}$$

$$\Delta(C,\psi) := \sup_k \Delta_k(C,\psi),$$

where the supremum in the first line of (V.2) is taken with respect to all partitions adapted to $t_1,..,t_n$ *which depend only on* C, ψ.

Let us explain the meaning of the quantities (V.2):

The partition p chops the tube into n cylinders C_k, each of which contains precisely one vertex of the graph γ underlying ψ. The number Δ_k is the maximal length of any curve that lies in C_k and starts and ends on the boundary of C_k and for any of the possible choppings. We restrict the curves e to be non-self-intersecting such that they cannot pass through vertices more than once. Finally, the numbers δ, Δ are just the minimum and maximum of the Δ_k.

The idea is to use δ, Δ to decide whether a tube is thin or not for a given state.

Definition V.4: A tube C *is said to be thin in the state* ψ *provided that*

(1) $0 < \delta(C,\psi) < o(l_p)$,
(2) $\delta(C,\psi)/\Delta(C,\psi) = o(1)$, *and*
(3) $n \gg 1$.

Condition (1) says that the cylinder of shortest length has nonzero length and is also not too long as compared to the Planck length. Condition (2) means that the longest cylinder is not much longer than the shortest one. Since the length Δ_k is in particular a measure for the transversal length of the tube, (1) and (2) imply that the tube has a transversal extension of order l_p. Moreover, it means that the regions C_k have approximately the same length. Finally, (3) means that the quantum length of the tube as defined below is macroscopic.

Definition V.5: (i) The length of the tube C *in the state* ψ *is defined by*

$$L(C,\psi) := \sum_{r,s} \frac{\langle \psi, \hat{L}(C_{r,s})\psi\rangle}{\|\psi\|^2}. \tag{V.3}$$

(ii) The physical length $L_{\rm phys}(c,\psi)$ *of a curve* c *in the state* ψ *is given by the length of* $L(C,\psi)$ *where* C *is any tube with central curve* c *and which is thin for* ψ.

Note that this definition of quantum length of a curve c in a state ψ does not refer to any background structure; it involves only the length operator.

Let us see whether that this makes mathematical sense and does give the correct answer for semi-classical considerations.

Since the tube is a *congruence* of curves $C_{r,s}$, each vertex of γ lies on at most one of the curves $C_{r,s}$ and, for a genuine distribution of vertices, each curve intersects at most one vertex. It follows that only a finite number of curves contribute for a finite number of times in (V.1) and so the tube operator is densely defined on $\mathscr{D}$ and it is trivial to see that it is positive semi-definite, symmetric, and possesses self-adjoint extensions. Since the length operators $\hat{L}(s_v)$ commute for distinct vertices v, it follows that all contributing length operators can be simultaneously diagonalized.

Next, let c be a smooth curve and q_{ab} some classical metric which is slowly varying in an open region containing c. We wish to construct a weave state ψ_q which approximates the classical length of c as measured by q_{ab}. For simplicity we consider a weave state ψ defined by a trivalent lattice so that all length operators are automatically diagonal. To simplify life further, let us assume that all edges have equal spin j_0, which implies that j_0 is integral. Let $\lambda(j_0)$ be the corresponding eigenvalue of (IV.18) (almost never will the lattice have a segment in common with one of the contributing curves so that we need to consider only case B of Sec. IV B). Let C be a tube thin for this weave with central curve c. Then we have $\delta(C,\psi) = \Delta(C,\psi) = \lambda(j_0) l_p$. If

length of the central curve of the tube and mean separation between vertices of the state as measured by q_{ab} are denoted as $L(c)$ and ϵ, respectively, then it follows that we obtain for the expectation value of the tube operator.

$$L(C)=l_p\,\frac{L(c)}{\epsilon}\,\lambda(j_0)(1+o(\epsilon/L(c))) \tag{V.4}$$

because, since C is thin for ψ, the number of vertices contained in the region filled by the tube is given by $n\approx L(c)/\epsilon$ (here we used that q_{ab} is slowly varying). It follows that in order for this quantity to equal $L(c)$ we must choose $\epsilon=\lambda(j_0)l_p$. As a consistency check, this means that the tube is also classically considered thin with respect to q_{ab} because the fact that there is only one vertex per unit $\delta(C,\psi)=\lambda(j_0)l_p:=\epsilon$ contained in the tube means that its transversal extension is also of order ϵ.

A few remarks are in order:

(1) Notice that λ is a monotonously increasing function of j_0 so that the lattice can be chosen coarser the more spin it carries in order to still approximate the same geometry [since then $\epsilon\propto\lambda(j_0)$].
(2) The limit $\epsilon\to 0$ blows up. It is physically meaningless to consider lattices with mean distance of vertices smaller than Planck scale which, as already stated in the literature, hints at a discrete structure.
(3) The limit $\epsilon\to\infty$ gives a zero eigenvalue because Σ becomes more and more empty.
(4) We see that there are weave states $\Psi(q_{ab},j_0;\delta(j_0))$ which approximate q_{ab} where we have considered j_0 as a free parameter. Notice that there is a large number of weave states corresponding to the various values of j_0 such that the same geometry is approximated [we cannot, however, let $j_0\to\infty$ because of the condition $\delta(C,\psi)=o(l_p)$ above]. Also observe that we use here weave states which necessarily involve intersecting and overlapping ($j_0>\frac{1}{2}$), rather than only linked, loops.[27]
(5) On the other hand, we can fix ϵ, vary j_0, and ask which classical geometry is being approximated. Since one and the same curve appears longer, according to (V.3), the more spin the lattice carries, we see that the corresponding q_{ab} must be warped as compared to δ_{ab}. This gives rise to the speculation that the spin of the lattice must have something to do with the curvature which in turn has to do with the local energy distribution of the gravitational field. Indeed, the spin characterizes the eigenvalue of the ADM Hamiltonian.[8,28]
(6) To summarize, the tube operator is state-independently defined; however, its expectation value for a given state can be associated with the length of its central curve only if it is thin, which is a state-dependent notion. On the other hand, the definition of ''thin'' involves only the length operator, so it is really quantum thinness that we are talking about; that is, rather than taking a weave that is supposed to approximate a geometry q_{ab} and saying that the transversal length of the tube should be of order l_p as measured by q_{ab}, we take it to be of order l_p as measured by the length operator for the given state. In fact, the notion of thinness applies to any state, not only to a weave state.

VI. SUMMARY

(i) A satisfactory quantization of the length of a piecewise smooth curve was given. The length operator leaves the piecewise analytic graph of a cylindrical function invariant (in particular, analytic), therefore it is proper to allow the curve to be only piecewise smooth rather than piecewise analytic. It is gauge invariant and diffeomorphism covariantly defined.
(ii) The length operator is an unbounded, self-adjoint, positive semi-definite operator with domain $\mathrm{Cyl}^3(\overline{\mathscr{A}/\mathscr{G}})$.
(iii) The spectral analysis turns out to be tedious but straightforward. In particular, the methods displayed here reveal that the complete spectrum can be found by writing a suitable computer code. This also is true for the volume operator.[25]
(iv) The spectrum is entirely discrete, the quantum of the length being given by $\sqrt[4]{3}l_p/\sqrt{2}$. In particular, all spin-network states on graphs with valence not bigger than 3 are eigenvectors of the length operator with, in general, nonvanishing eigenvalue.

(v) In order to approximate a three-geometry a tube operator has to be constructed. One can build intersecting and overlapping weave states which are such that they reproduce the classical length of the tubes given a classical geometry.

(vi) An interesting issue is the following: We could try to define the length of a tube in an alternative way which does not refer to a length operator at all: As a tube is a three-dimensional object it has a quantum volume and its $t=\text{const}$ discs do have a quantum area. We could say that a tube is thin for a state provided that each such disc has an area which is bounded from below and above by its infimum $a>0$ and supremum A as t varies and define $L=2V/(A+a)$ where V is the expectation value of the volume. Interestingly, these two notions of length do not seem to be closely related as the latter would be zero for any state based on a trivalent graph while the operator defined here would have a nonzero expectation value. This is not unexpected for two reasons: First of all, the ''formula'' length times area equals volume only holds in Euclidean space and, second, our length operator is at best a very different operator ordering of that ''formula'' as it involves commutators with holonomies which are absent in the definition of area and volume operators.

ACKNOWLEDGMENTS

This research project was supported in part by DOE Grant No. DE-FG02-94ER25228 to Harvard University.

APPENDIX: THE 6*j*-SYMBOL

The following can be found in any textbook on the recoupling theory of angular momenta (for instance, Ref. 26).

Consider the coupling of three angular momenta of some spin system consisting of three independent subsystems with angular momenta j_I, $I=1,2,3$. Since the three systems are independent of each other, the operators J_I^i, J_J^j, $I\neq J$, mutually commute. We are interested in states that are labeled by the quantum numbers of a maximal set of mutually commuting observables which includes the square of the total angular momentum $J^i=J_1^i+J_2^i+J_3^i$ of this system. It is easy to see that the set consisting of j_1,j_2,j_3,j,m (where m is the eigenvalue of J^3) is insufficient because j_i,m_i is a set consisting of six rather than five quantum numbers. The missing quantum number is any choice of J_{IJ}, $I<J$, which labels the eigenvalue of the square of $J_{IJ}:=J_I+J_J$. It is easy to see that J,J_{IJ} satisfy the angular momentum algebra.

Denote an orthonormal basis of states so constructed by $|j_{IJ},j;j_1,j_2,j_3\rangle$ where it is understood that one first couples j_I,j_J to resulting spin j_{IJ} and then j_{IJ},j_K to the resulting total spin j. Any choice of such a recoupling scheme leads to an orthonormal basis and thus the transformation between these bases must be unitary. The matrix elements of this transformation were explicitly computed by Racah. In particular we have

$$\langle j_{12},j;j_1,j_2,j_3|j_{23},j;j_1,j_2,j_3\rangle:=(-1)^{j_1+j_2+j_3+j}\sqrt{(2j_{12}+1)(2j_{23}+1)}\begin{Bmatrix} j_1 & j_2 & j_{12} \\ j_3 & j & j_{23}\end{Bmatrix}, \tag{A1}$$

where the 2×3 matrix on the right-hand side is the so-called $6j$-symbol for which a closed formula exists.[26]

There exists a choice of phases for the basis vectors such that all the $6j$-symbols are real. With this choice they enjoy a large number of symmetries of which we just need two:

(1) it is invariant under an arbitrary permutation of its columns and
(2) it is crossing symmetric, that is,

$$\begin{Bmatrix} j_1 & j_2 & j_3 \\ j_4 & j_5 & j_6\end{Bmatrix}=\begin{Bmatrix} j_1 & j_6 & j_5 \\ j_4 & j_3 & j_2\end{Bmatrix}. \tag{A2}$$

For the purposes of this paper it is sufficient to table the value of the $6j$-symbol for the special case that one of j_I,j_{IJ},j takes the value $\frac{1}{2}$. Then we have the special values

$$\left\{\begin{array}{ccc} a & b & c \\ \frac{1}{2} & c-\frac{1}{2} & b+\frac{1}{2} \end{array}\right\} = (-1)^{a+b+c} \sqrt{\frac{(a+c-b)(a+b-c+1)}{(2b+1)(2b+2)(2c)(2c+1)}}, \tag{A3}$$

$$\left\{\begin{array}{ccc} a & b & c \\ \frac{1}{2} & c-\frac{1}{2} & b-\frac{1}{2} \end{array}\right\} = (-1)^{a+b+c} \sqrt{\frac{(a+b+c+1)(b+c-a)}{(2b)(2b+1)(2c)(2c+1)}}. \tag{A4}$$

This is all one needs to know in order to verify the eigenvalue calculations of the present paper. How to generalize the recoupling theory for any number of spins is clear.

[1] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967); **162**, 1195 (1967); **162**, 1239 (1967).
[2] R. M. Wald, *General Relativity* (Univ. Chicago, Chicago, 1989).
[3] A. Ashtekar, Phys. Rev. Lett. **57**, 2244 (1986); Phys. Rev. D **36**, 1587 (1987).
[4] A. Ashtekar, in *Mathematics and General Relativity* (AMS, Providence, 1987).
[5] T. Thiemann, Acta Cosmologica **21**, 145 (1995); Class. Quantum Grav. **13**, 1383 (1996).
[6] A. Ashtekar, Phys. Rev. D **53**, 2865 (1996).
[7] T. Thiemann, Phys. Lett. B **380**, 257 (1996).
[8] T. Thiemann, ''Quantum Spin Dynamics (QSD),'' Class. Quantum. Grav. **15**, 839–873 (1998); Harvard University preprint HUTMP-96/B-359, gr-qc/9606089; ''Quantum Spin Dynamics (QSD) II: The Kernel of the Wheeler-DeWitt Constraint Operator,'' Class. Quantum. Grav. **15**, 875–905 (1998); Harvard University preprint HUTMP-96/B-352, gr-qc/9606090.
[9] F. Barbero, Phys. Rev. D **51**, 5507 (1995).
[10] R. Loll, Acta Cosmologica **21**, 132 (1995); Phys. Rev. D **54**, 5381 (1996).
[11] C. Rovelli and L. Smolin, Phys. Rev. Lett. **72**, 446 (1994).
[12] C. Rovelli and L. Smolin, Nucl. Phys. B **442**, 593 (1995); Erratum, Nucl. Phys. B **456**, 734 (1995).
[13] A. Ashtekar and J. Lewandowski, Class. Quantum Grav. **14**, A55–A81 (1997); ''Quantum Theory of Geometry II: Volume Operators,'' Adv. Theor. Math. **1**, 388–429 (1997); gr-qc/9711031.
[14] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão, and T. Thiemann, J. Math. Phys. **36**, 6456 (1995).
[15] R. Penrose, *Quantum Theory and Beyond*, edited by T. Bastin (Cambridge U.P., Cambridge, 1971).
[16] A. Ashtekar and C. J. Isham, Class. Quantum Grav. **9**, 1433 (1992); A. Ashtekar and J. Lewandowski, ''Representation theory of analytic holonomy $C^\star$ algebras,'' in *Knots and Quantum Gravity*, edited by J. Baez (Oxford U.P., Oxford, 1994).
[17] D. Marolf and J. M. Mourão, Commun. Math. Phys. **170**, 583 (1995).
[18] A. Ashtekar and J. Lewandowski, J. Math. Phys. **36**, 2170 (1995).
[19] A. Ashtekar and J. Lewandowski, J. Geom. Phys. **17**, 191 (1995).
[20] C. Rovelli and L. Smolin, Phys. Rev. D **53**, 5743 (1995).
[21] J. Baez, ''Spin network states in gauge theory,'' Adv. Math. (in press); ''Spin networks in non-perturbative quantum gravity,'' gr-qc/9504036 Adv. Theo. Math. **1**, 388–429 (1997).
[22] T. Thiemann, ''The inverse loop transform,'' Harvard University preprint HUTMP-95/B-346, hep-th/9601105, J. Math. Phys. **39**, 1236–48 (1998).
[23] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol. II* (Academic, New York, 1970).
[24] R. De Pietri and C. Rovelli, Phys. Rev. D **54**, 2664 (1996).
[25] T. Thiemann, ''Closed formula for the matrix elements of the volume operator for canonical quantum gravity,'' Harvard University preprint HUTMP-96/B-353, gr-qc/9606091; J. Math. Phys. **39**, 3347–3371 (1998). (Paper I)
[26] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1974).
[27] A. Ashtekar, C. Rovelli, and L. Smolin, Phys. Rev. Lett. **69**, 237 (1992).
[28] T. Thiemann, ''QSD VI: Quantum Poincaré Algebra and a Quantum Positivity of Energy Theorem for Canonical Quantum Gravity,'' Harvard University preprint HUTMP-96/B-356, gr-qc/9705020.

Nonsingular Hankel functions as a new basis for electronic structure calculations

E. Bott
Institut für Physikalische Chemie der Technischen Universität Darmstadt, D-64287 Darmstadt, Germany

M. Methfessel
Institute for Semiconductor Physics, Walter-Korsing Str. 2, D-15230 Frankfurt (Oder), Germany

W. Krabs
Fachbereich Mathematik der Technischen Universität Darmstadt, D-64289 Darmstadt, Germany

P. C. Schmidt
Institut für Physikalische Chemie der Technischen Universität Darmstadt, D-64287 Darmstadt, Germany

(Received 23 December 1997; accepted for publication 12 February 1998)

As a basis for electronic structure calculations, Gaussians are inconvenient because they show unsuitable behavior at larger distances, while Hankel functions are singular at the origin. This paper discusses a new set of special functions which combine many of the advantageous features of both families. At large distances from the origin, these ''smoothed Hankel functions'' resemble the standard Hankel functions and therefore show behavior similar to that of an electronic wave function. Near the origin, the functions are smooth and analytical. Analytical expressions are derived for two-center integrals for the overlap, the kinetic energy, and the electrostatic energy between two such functions. We also show how to expand such a function around some point in space and discuss how to evaluate the potential matrix elements efficiently by numerical integration. This supplies the elements needed for a practical application in an electronic structure calculation. © *1998 American Institute of Physics.* [S0022-2488(98)02206-3]

I. INTRODUCTION

The calculation of the electronic structure of molecules and crystals is becoming increasingly important as a means to understand phenomena on a microscopic level and to supply data which are inaccessible to experiment.[1] The approaches range from Hartree–Fock and configuration-interaction methods to density-functional calculations and the Car–Parrinello method for quantum-mechanical molecular dynamics. In all these approaches, the wave functions must be represented in an accurate but easily handled manner and several central quantities, such as the expectation values for a potential of general shape, must be evaluated. The choice of the basis set used in the representation has a large influence on the calculational efficiency, the numerical stability, and the effort of implementation. In this paper, we present a new set of analytical functions which were designed to optimize these aims.

In general, most applications use one of the following three approaches: (1) an expansion in plane waves, equivalent to a numerical representation on a real-space mesh, (2) an expansion in some kind of analytical basis set centered on the atoms, or (3) an expansion in numerical atomic functions such as the orbitals of a free atom. The expansion in plane waves is extremely popular, thanks to the simple formulation, general flexibility, and numerical robustness. This approach is often used for norm-conserving[2] or ultrasoft[3] pseudopotentials as well as in the linear augmented plane-wave (LAPW) method.[4] These successful methods rely heavily on fast-Fourier-transform algorithms for efficient execution. Unfortunately, plane waves treat all parts of space in the same way, even though more detail is needed close to the atoms. Depending on the system, a large

 © 1998 American Institute of Physics

number of basis functions are needed, leading to slow program execution and large storage requirements.

In contrast, representations using some kind of atom-centered local functions generally involve considerably fewer basis functions, with the potential of large savings in computer time and storage. Whether these gains are realized depends on the type of function used. For example, a basis of Gaussian functions leads to simple algorithms since most operations can be done analytically. Unfortunately, the behavior of the true wave functions is closer to exponential than Gaussian. Consequently several basis functions must be contracted to form chemical basis sets and numerical stability becomes an issue. These problems are avoided by Hankel or Slater functions, for example, in the linear-muffin-tin orbital (LMTO) method[4] which uses a basis of augmented Hankel functions centered on the atoms. The cost for this type of approach is that the mathematical operations are more complicated, making it difficult to implement essential features such as the forces on the atoms. Furthermore, due to the inherent simplicity of the plane-wave methods, their execution speeds are quite difficult to match in practice despite the smaller function set. Finally, Hankel functions are singular near the nucleus. This restricts their use to augmentation methods which replace each function by a numerical representation near the atom.

The new functions presented here are intended to combine as far as possible the positive features of Hankel functions and Gaussians. In essence, they are Hankel functions from which the singularities have been removed, leading to smooth functions which are analytical in all of space. They have the following advantages. Since their behavior at larger radii is that of Hankel functions, they are similar to the true wave functions and avoid the need for basis contractions and stability problems. The degree of smoothing can be tuned, making it possible to optimize the basis so that fewer functions are needed for an accurate representation. Because these functions are analytical, numerical operations on a real-space mesh can be done reasonably efficiently. In addition, many important operations (for example, all relevant two-center integrals) can be done analytically. Finally, since the functions are nonsingular and smooth over all of space they seem predestined for an implementation of pseudopotential methods using local orbitals.

An implementation of density-functional program, based on these functions and using augmentation to supply local atomic detail, has recently been completed.[5] In this paper, we present the mathematical properties of these new ''smoothed Hankel functions.'' Our aim is to supply those properties and equations which are needed for an electronic structure calculation. Ultimately, for any selected basis set $\{\varphi_i(\mathbf{r})\}$ the tasks are to evaluate the overlap matrix

$$S_{i,j}=\int \varphi_i(\mathbf{r})^*\varphi_j(\mathbf{r})d^3r, \tag{1.1}$$

the matrix of the kinetic energy

$$T_{i,j}=\int \varphi_i(\mathbf{r})^*(-\Delta)\varphi_j(\mathbf{r})d^3r, \tag{1.2}$$

and the matrix of the potential energy

$$V_{i,j}=\int \varphi_i(\mathbf{r})^*V(\mathbf{r})\varphi_j(\mathbf{r})d^3r \tag{1.3}$$

to high accuracy and reasonably efficiently, and to assemble the charge density out of the products

$$\rho_{ij}(\mathbf{r})=\varphi_i(\mathbf{r})^*\varphi_j(\mathbf{r}) \tag{1.4}$$

in a form which can be used in the further steps of the calculation. More specifically, we would like to evaluate the simpler terms such as S_{ij} and T_{ij} analytically. If the potential matrix element is to be calculated numerically, the functions must be easy to generate on a mesh. For realistic descriptions involving either augmentation or nonlocal pseudopotentials, it is necessary to expand each function locally in the vicinity of an atom. Finally, for the application to crystals, all these steps must be done for Bloch sums of the original functions. These points are addressed in the following.

The rest of this paper is organized as follows. In the following Secs. II and III we define some useful differential operators and collect some results concerning Fourier transforms which are needed in the following sections. In Sec. IV, we present a family of functions $F_L(\mathbf{r})$ of which the smoothed Hankel functions $H_L(\mathbf{r})$ and the related generalized Gaussians $G_{pL}(\mathbf{r})$ are special cases. Following this we introduce the functions $G_{pL}(\mathbf{r})$ and $H_L(\mathbf{r})$ in Secs. V and VI.

One pleasant feature of the smoothed Hankels is that the matrix elements Eqs. (1.1), (1.2) can be expressed analytically, using the functions $H_L(\mathbf{r})$ themselves, their energy derivatives $\dot{H}_L(\mathbf{r})$, and related functions $W_{pL}(\mathbf{r})$. These are discussed in Sec. VII. Sections VIII–X give expressions for the matrix elements in the nonperiodic case, show how to evaluate the Bloch sums of the $H_L(\mathbf{r})$, and how to obtain the matrix elements for the Bloch-periodic case.

In Secs. XI and XII we show how to evaluate Coulomb integrals involving the $H_L(\mathbf{r})$ (Sec. XI) and how to expand the functions around an arbitrary site, using a convenient technique which can be used for any type of smooth function (Sec. XII). Finally, in the Appendix, we give an integral representation of the $H_L(\mathbf{r})$ and elucidate the connection to the well-known Ewald summation method for lattice sums.

II. SPHERICAL HARMONICS $Y_L(\hat{\mathbf{r}})$ AND $\mathscr{Y}_L(\mathbf{r})$

In the following sections we shall be working with real valued spherical harmonics, labeled with the compound index $L=(\ell,m_\ell)$:

$$Y_L(\hat{\mathbf{r}}) := Y^{m}_{\ell}(\hat{\mathbf{r}}) \tag{2.1}$$

as given in (59:14:8) of Ref. 6 and with the spherical harmonic polynomials

$$\mathscr{Y}_L(\mathbf{r}) := r^{\ell} Y_L(\hat{\mathbf{r}}). \tag{2.2}$$

The $\mathscr{Y}_L(\mathbf{r})$ are real polynomials in the Cartesian coordinates x,y,z of $\mathbf{r}$. Therefore we can use expressions such as

$$\mathscr{Y}_L(i\mathbf{r}) \quad \text{or} \quad \mathscr{Y}_L(-\nabla) \tag{2.3}$$

obtained by substituting ix, iy, and iz and $-\partial/\partial x$, $-\partial/\partial y$, and $-\partial/\partial z$, respectively, for x, y, and z in the polynomial $\mathscr{Y}_L(\mathbf{r})$. Furthermore each additive term of the polynomial $\mathscr{Y}_L$ is a product of a real valued coefficient and exactly ℓ variables. Hence we obtain in particular

$$\mathscr{Y}_L(a\mathbf{g}) = a^{\ell}\mathscr{Y}_L(\mathbf{g}), \tag{2.4}$$

where a is any complex number and $\mathbf{g}$ is any possible argument for $\mathscr{Y}_L$ with three components which commute with the complex numbers.

Let S denote the unit sphere surface. The $Y_L(\hat{\mathbf{r}})$ form a complete orthonormal basis of $\mathscr{L}_2(S)$, i.e., of the set of all complex valued functions $f(\hat{\mathbf{r}})$ with

$$\int_0^{2\pi}\int_0^{\pi} |f(\hat{\mathbf{r}})|^2 \sin\theta d\theta d\varphi < \infty. \tag{2.5}$$

Since the $Y_L(\hat{\mathbf{r}})$ are bounded on the unit sphere surface, we can consequently write

$$Y_K(\hat{\mathbf{r}})Y_L(\hat{\mathbf{r}}) = \sum_M a_M Y_M(\hat{\mathbf{r}}) \tag{2.6}$$

and using the orthonormality of the $Y_L(\hat{\mathbf{r}})$ we see

$$a_M = C_{KLM}, \tag{2.7}$$

where the C_{KLM} denote the real Gaunt integrals (or Clebsch–Gordan coefficients)

$$C_{KLM}=\int_0^{2\pi}\int_0^{\pi} Y_K(\hat{\mathbf{r}})Y_L(\hat{\mathbf{r}})Y_M(\hat{\mathbf{r}})\sin\,\theta d\theta d\varphi. \tag{2.8}$$

By multiplying both sides of (2.6) with $r^{k+\ell}$ and using (2.2) we obtain the corresponding expression for the spherical harmonic polynomials:

$$\mathscr{Y}_K(\mathbf{r})\mathscr{Y}_L(\mathbf{r})=\sum_M C_{KLM}(x^2+y^2+z^2)^{(k+\ell-m)/2}\mathscr{Y}_M(\mathbf{r}), \tag{2.9}$$

where $K=(k,m_k)$, $L=(\ell,m_\ell)$, and $M=(m,m_m)$. According to the selection rules for the Gaunt integrals, $k+l-m$ is a non-negative even integer when C_{KLM} is nonzero.

III. FOURIER TRANSFORMS

Here we list some results from the theory of Fourier transformation and Fourier series expansion which we require later on. We shall apply Fourier transformation to elements of the class $\mathscr{S}(\mathscr{R}^3)$ of *rapidly decreasing functions* on $\mathscr{R}^3$ only, where $\mathscr{S}(\mathscr{R}^3)$ is defined by

$$\mathscr{S}(\mathscr{R}^3):=\left\{\sigma:\mathscr{R}^3\to\mathscr{C}\,\middle|\,\begin{array}{c}\sigma\in C^\infty(\mathscr{R}^3)\ \text{ and }\ \sup_{\mathbf{r}}|\mathbf{r}^\alpha D^\beta\sigma(\mathbf{r})|<\infty\ \text{ for all multi-indices}\\ \alpha=(\alpha_1,\alpha_2,\alpha_3),\beta=(\beta_1,\beta_2,\beta_3)\in\mathscr{N}_0^3\end{array}\right\}, \tag{3.1}$$

$$\mathbf{r}^\alpha:=x^{\alpha_1}y^{\alpha_2}z^{\alpha_3}, \tag{3.2}$$

$$D^\beta\sigma(\mathbf{r}):=\frac{\partial^{|\beta|}}{\partial x^{\beta_1}\partial y^{\beta_2}\partial z^{\beta_3}}\sigma(\mathbf{r}), \tag{3.3}$$

$$|\beta|:=\beta_1+\beta_2+\beta_3 \tag{3.4}$$

(see Definition 17.1 in Ref. 7), i.e., $\mathscr{S}(\mathscr{R}^3)$ is the class of all functions $\sigma:\mathscr{R}^3\to\mathscr{C}$ with the properties that σ is arbitrarily often differentiable, is continuous together with all its derivatives, and that σ and all its derivatives go to zero faster than the inverse of any polynomial as $|\mathbf{r}|\to\infty$. Obviously $\mathscr{S}(\mathscr{R}^3)$ is a linear space over the complex numbers $\mathscr{C}$.

We denote the Fourier transform of a rapidly decreasing function $\sigma(\mathbf{r})\in\mathscr{S}(\mathscr{R}^3)$ by $\hat{\sigma}(\mathbf{q})$ and define it by

$$\hat{\sigma}(\mathbf{q}):=\int_{\mathscr{R}^3} e^{-i\mathbf{q}\cdot\mathbf{r}}\sigma(\mathbf{r})d^3r. \tag{3.5}$$

In the following we shall simply write $\int$ for $\int_{\mathscr{R}^3}$.

A central theorem ensures that $\hat{\sigma}$ belongs to $\mathscr{S}(\mathscr{R}^3)$ for all $\sigma\in\mathscr{S}(\mathscr{R}^3)$ and that the Fourier transformation $\sigma\mapsto\hat{\sigma}$ is an isomorphism from $\mathscr{S}(\mathscr{R}^3)$ onto $\mathscr{S}(\mathscr{R}^3)$ (cf. Theorem 17.3 in Ref. 7). One can reconstruct $\sigma(\mathbf{r})$ from $\hat{\sigma}(\mathbf{q})$ by

$$\sigma(\mathbf{r})=\frac{1}{(2\pi)^3}\int e^{i\mathbf{q}\cdot\mathbf{r}}\hat{\sigma}(\mathbf{q})d^3q \tag{3.6}$$

(cf. Theorem 17.2 in Ref. 7).

Concerning the interrelationship between Fourier transformation and derivation, one has the following general results for $\sigma(\mathbf{r})\in\mathscr{S}(\mathscr{R}^3)$:

$$D_{\mathbf{q}}^\beta\hat{\sigma}(\mathbf{q})=\widehat{((-i\mathbf{r})^\beta\sigma(\mathbf{r}))}(\mathbf{q}), \tag{3.7}$$

$$\widehat{(D_{\mathbf{r}}^\beta\sigma)}(\mathbf{q})=(i\mathbf{q})^\beta\hat{\sigma}(\mathbf{q}) \tag{3.8}$$

(cf. pp. 146, 142 of Ref. 7) whereby $D_{\mathbf{q}}$ and $D_{\mathbf{r}}$ denote the differentiation with respect to $\mathbf{q}$ and $\mathbf{r}$, respectively.

If we consider a function $\sigma(\mathbf{r}) \in \mathscr{S}(\mathscr{R}^3)$ which can be separated into a radial and an angular part according to

$$\sigma(\mathbf{r}) = \frac{u(r)}{r} Y_L(\hat{\mathbf{r}}), \tag{3.9}$$

then the corresponding Fourier transform $\hat{\sigma}(\mathbf{q})$ is given by

$$\hat{\sigma}(\mathbf{q}) = \frac{v(q)}{q} Y_L(\hat{\mathbf{q}}), \tag{3.10}$$

where

$$v(q) = 4\pi(-i)^{\ell} \int_0^{\infty} u(r) j_{\ell}(qr) qr dr, \tag{3.11}$$

$$u(r) = \frac{4\pi i^{\ell}}{(2\pi)^3} \int_0^{\infty} v(q) j_{\ell}(qr) qr dq \tag{3.12}$$

(cf. pp. 35/36 of Ref. 8), whereby the j_{ℓ} denote the spherical Bessel functions (see 10.1.1 in Ref. 9).

Further useful relationships are (cf. Eq. (17.11) in Ref. 7)

$$\widehat{(e^{-a^2 r^2})}(\mathbf{q}) = \left(\frac{\pi}{a^2}\right)^{3/2} e^{-q^2/4a^2} \tag{3.13}$$

and for any $\sigma, \rho \in \mathscr{S}(\mathscr{R}^3)$:

$$\widehat{(\sigma(\mathbf{r}-\mathbf{R}))}(\mathbf{q}) = \hat{\sigma}(\mathbf{q}) e^{-i\mathbf{q}\cdot\mathbf{R}} \tag{3.14}$$

(obvious from (3.5)), and

$$\int \sigma(\mathbf{r})^* \rho(\mathbf{r}) d^3 r = \frac{1}{(2\pi)^3} \int \hat{\sigma}(\mathbf{q})^* \hat{\rho}(\mathbf{q}) d^3 q \tag{3.15}$$

(cf. Exercise 1 on p. 153 in Ref. 7).

Next, consider a Bloch sum

$$K^b(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} K(\mathbf{r}-\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{r}} \sum_{\mathbf{R}} e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} K(\mathbf{r}-\mathbf{R}) \tag{3.16}$$

for some rapidly decreasing function $K(\mathbf{r}) \in \mathscr{S}(\mathscr{R}^3)$ where $\mathbf{R}$ runs over the Bravais vectors of some lattice. Since the function $\mathbf{r} \mapsto \Sigma_{\mathbf{R}} e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} K(\mathbf{r}-\mathbf{R})$ is periodic in direct space and has the translational symmetry of the Bravais lattice, it can be expanded as a Fourier series in terms of the plane waves $e^{i\mathbf{G}\cdot\mathbf{r}}$ associated with the reciprocal lattice vectors $\mathbf{G}$:

$$\sum_{\mathbf{R}} e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} K(\mathbf{r}-\mathbf{R}) = \frac{1}{\Omega} \sum_{\mathbf{G}} \hat{K}(\mathbf{k}+\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \tag{3.17}$$

(cf. Appendix D in Ref. 10) where Ω is the volume of a direct lattice unit cell. From (3.16) we therefore obtain

$$K^b(\mathbf{r}) = \frac{1}{\Omega} \sum_{\mathbf{G}} \hat{K}(\mathbf{k}+\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \tag{3.18}$$

Starting from a second Bloch sum for some function $L(\mathbf{r}) \in \mathscr{S}(\mathscr{R}^3)$ with the same vector $\mathbf{k}$ as in the case of $K^b(\mathbf{r})$,

$$L^b(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} L(\mathbf{r}-\mathbf{R}) \tag{3.19}$$

we obtain from (3.18) and

$$\int_{\mathrm{UC}} e^{i(\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}} d^3r = \Omega\, \delta_{\mathbf{G}\mathbf{G}'} \tag{3.20}$$

(cf. Appendix D (D3)–(D5) in Ref. 10) the expression

$$\int_{\mathrm{UC}} K^b(\mathbf{r})^* L^b(\mathbf{r}) d^3r = \frac{1}{\Omega} \sum_{\mathbf{G}} \hat{K}(\mathbf{k}+\mathbf{G})^* \hat{L}(\mathbf{k}+\mathbf{G}). \tag{3.21}$$

In these equations, $\int_{\mathrm{UC}}$ denotes the integration over a unit cell.

IV. PROPERTIES OF THE FUNCTIONS $F_L(\mathbf{r}) := \mathscr{Y}_L(-\nabla) f(r)$

In this section we introduce a general method for creating families of functions (cf. Ref. 11) and present some properties which all these families have in common. Among others, the families G_L of generalized Gaussians and H_L of smoothed Hankels to be introduced in the Secs. V and VI are specific instances of such families.

Let $f(r) = f(|\mathbf{r}|)$ be a function defined in three-dimensional space which depends only on $r = |\mathbf{r}|$ and is an element of the linear space $\mathscr{S}(\mathscr{R}^3)$ of Eq. (3.1). In particular $f(r)$ is arbitrarily often differentiable and we can define a family of functions $F_L(\mathbf{r}) \in \mathscr{S}(\mathscr{R}^3)$ by

$$F_L(\mathbf{r}) := \mathscr{Y}_L(-\nabla) f(r). \tag{4.1}$$

Applying (3.8), (2.4), and (2.2) we see that the Fourier transform of F_L is given by

$$\widehat{F_L}(\mathbf{q}) = \mathscr{Y}_L(-i\mathbf{q})\widehat{f(r)}(\mathbf{q}) = (-i)^\ell q^\ell \widehat{f(r)}(\mathbf{q}) Y_L(\hat{\mathbf{q}}). \tag{4.2}$$

Since $f(r)$ depends on $\mathbf{r}$ only through $|\mathbf{r}| = r$, we obtain from Corollary 20.6 in Ref. 7 that $\widehat{f(r)}(\mathbf{q})$ depends on $\mathbf{q}$ only through $|\mathbf{q}| = q$. Therefore, $\widehat{F_L}(\mathbf{q})$ is of the form (3.10) and from (3.9)–(3.12) we can conclude that

$$F_L(\mathbf{r}) = f_\ell(r) Y_L(\hat{\mathbf{r}}), \tag{4.3}$$

where

$$f_\ell(r) = \frac{4\pi}{(2\pi)^3} \int_0^\infty q^{\ell+2} \widehat{f(r)}(\mathbf{q}) j_\ell(qr) dq. \tag{4.4}$$

By noting that $\mathscr{Y}_0(\mathbf{r}) = Y_0(\hat{\mathbf{r}}) = 1/\sqrt{4\pi}$ and comparing (4.3) with (4.1) we obtain

$$f_0(r) = f(r). \tag{4.5}$$

From (4.4) we see that $f_\ell(r)$ depends on r only through the spherical Bessel function j_ℓ. From the recurrence relation

$$j_\ell(qr) = \frac{1}{q}\left[-\frac{\partial}{\partial r} + \frac{\ell-1}{r}\right] j_{\ell-1}(qr) \quad \text{for } \ell \geqslant 1 \tag{4.6}$$

for the j_ℓ (cf. 10.1.22 in Ref. 9) we consequently get a recurrence relation for the f_ℓ:

$$f_\ell(r)=\left[-\frac{\partial}{\partial r}+\frac{\ell-1}{r}\right]f_{\ell-1}(r) \quad \text{for } \ell\geqslant 1. \tag{4.7}$$

Using (4.7) and (4.5) one can prove by induction over ℓ

$$f_\ell(r)=r^\ell\left(-\frac{1}{r}\frac{\partial}{\partial r}\right)^\ell f(r) \quad \text{for all } \ell\in\mathcal{N}_0 \tag{4.8}$$

(see p. 39 in Ref. 8). We see from (4.8) that we can write $f_\ell(r)$ in the form

$$f_\ell(r)=r^\ell\chi_\ell(r), \tag{4.9}$$

where $\chi_\ell(r)$ is given by

$$\chi_0(r)=f_0(r)=f(r), \tag{4.10}$$

$$\chi_\ell(r)=\left(-\frac{1}{r}\frac{\partial}{\partial r}\right)^\ell f(r) \quad \text{for all } \ell\in\mathcal{N}_0, \tag{4.11}$$

$$\chi_\ell(r)=-\frac{1}{r}\frac{\partial}{\partial r}\chi_{\ell-1}(r) \quad \text{for } \ell\geqslant 1. \tag{4.12}$$

To conclude this section, we derive a recurrence relation for the radial functions $\chi_\ell(r)$ for the special case that $f(r)$ fulfills

$$\Delta f(r)=p(\mathbf{r}) \tag{4.13}$$

for some known function $p(\mathbf{r})$.

For any $f(r)\in\mathcal{S}(\mathbb{R}^3)$ we have $\Delta f(r)=p(\mathbf{r})\in\mathcal{S}(\mathbb{R}^3)$. By writing the Laplacian operator Δ in polar coordinates (see, e.g., p. 32 in Ref. 12) we see that the function $\Delta f(r)$ also depends only on $r=|\mathbf{r}|$, i.e.,

$$p(\mathbf{r})=p(r). \tag{4.14}$$

By applying the operator $\mathcal{Y}_L(-\nabla)$ to both sides of (4.13) we obtain

$$\Delta F_L(\mathbf{r})=P_L(\mathbf{r}), \tag{4.15}$$

where

$$P_L(\mathbf{r})=\mathcal{Y}_L(-\nabla)p(r)=p_\ell(r)Y_L(\hat{\mathbf{r}}). \tag{4.16}$$

Next, we write the Laplacian in polar coordinates in the form

$$\Delta=\frac{1}{r}\frac{\partial^2}{\partial r^2}(r\cdot)-\frac{\widehat{\ell^2}}{r^2\hbar^2}, \tag{4.17}$$

where

$$\widehat{\ell^2}Y_L(\hat{\mathbf{r}})=\ell(\ell+1)\hbar^2Y_L(\hat{\mathbf{r}}) \tag{4.18}$$

(cf. pp. 32, 150, 151 in Ref. 12). This lets us derive from (4.15) the following relationship between $f_\ell(r)$ and $p_\ell(r)$:

$$\frac{\partial^2}{\partial r^2}f_\ell(r)+\frac{2}{r}\frac{\partial}{\partial r}f_\ell(r)-\frac{\ell(\ell+1)}{r^2}f_\ell(r)=p_\ell(r). \tag{4.19}$$

Applying (4.8) we can obtain from this equation

$$-p_{\ell-1}(r)=\left[\frac{\partial}{\partial r}+\frac{\ell+1}{r}\right]f_\ell(r) \quad \text{for } \ell\geqslant 1. \tag{4.20}$$

Together with (4.7) this leads to

$$f_{\ell+1}(r)=\frac{2\ell+1}{r}f_\ell(r)+p_{\ell-1}(r) \quad \text{for } \ell\geqslant 1, \tag{4.21}$$

which by virtue of (4.9) is equivalent to

$$r^2\chi_{\ell+1}(r)=(2\ell+1)\chi_\ell(r)+r^{-(\ell-1)}p_{\ell-1}(r) \quad \text{for } \ell\geqslant 1. \tag{4.22}$$

Since $\chi_0(r)$ equals $f(r)$, relationship (4.22) can be used to calculate the $\chi_\ell(r)$ for higher angular momenta by recursion, assuming that the $p_\ell(r)$ are known and that $\chi_1(r)$ has already been obtained. For example, $\chi_1(r)=-(1/r)(\partial/\partial r)\chi_0(r)$ can be used to do this (cf. (4.12)). The functions $f_\ell(r)=r^\ell\chi_\ell(r)$ are then also supplied, giving a method to calculate the functions of the family F_L in practical applications. Note, however, that a power series expansion is more stable numerically for small values of r.

V. THE GENERALIZED GAUSSIANS

In this section, we introduce functions G_L and G_{pL} which will be needed in the succeeding sections as part of recurrence relations, for the evaluation of the Hamiltonian and overlap matrix elements between smoothed Hankel functions, for the evaluation of Coulomb integrals, and in the expansion of smoothed Hankel functions around a chosen site.

The functions G_L are generalized Gaussians used in standard quantum mechanical molecular orbital approaches[13] defined for some energy parameter $\varepsilon=-\kappa^2\leqslant 0$ $(\kappa\geqslant 0)$ and an additional parameter $a>0$ by

$$G_L(\mathbf{r}):=\mathscr{Y}_L(-\nabla)g(r), \tag{5.1}$$

where

$$g(r):=\left(\frac{a^2}{\pi}\right)^{3/2}e^{\varepsilon/4a^2}e^{-a^2r^2}=\left(\frac{a^2}{\pi}\right)^{3/2}e^{\gamma\varepsilon}e^{-a^2r^2}, \tag{5.2}$$

$$\gamma:=\frac{1}{4a^2}. \tag{5.3}$$

The root function $g(r)$ is a normalized Gaussian of width $1/a$, multiplied by the factor $e^{\gamma\epsilon}$ for convenience in later expressions. Obviously $g(r)$ is an element of $\mathscr{S}(\mathscr{R}^3)$ and hence

$$G_L(\mathbf{r})\in\mathscr{S}(\mathscr{R}^3) \quad \text{for all } L. \tag{5.4}$$

From our general results of the last section it follows that

$$G_L(\mathbf{r})=g_\ell(r)Y_L(\hat{\mathbf{r}})=r^\ell\mu_\ell(r)Y_L(\hat{\mathbf{r}}) \tag{5.5}$$

with

$$\mu_\ell(r)=\left(-\frac{1}{r}\frac{\partial}{\partial r}\right)^\ell g(r)=(2a^2)^\ell g(r)=\left(\frac{a^2}{\pi}\right)^{3/2}e^{\gamma\varepsilon}(2a^2)^\ell e^{-a^2r^2}. \tag{5.6}$$

Consequently we can write in closed form

$$G_L(\mathbf{r})=\left(\frac{a^2}{\pi}\right)^{3/2}e^{\gamma\varepsilon}(2a^2)^\ell e^{-a^2r^2}r^\ell Y_L(\hat{\mathbf{r}}). \tag{5.7}$$

The following sections also require the Fourier transform $\widehat{G_L}$ of G_L which by (5.1), (3.8), (5.2), (5.3) and (3.13) can easily be determined to be

$$\widehat{G_L}(\mathbf{q}) = \mathscr{Y}_L(-i\mathbf{q}) e^{\gamma(\varepsilon - q^2)}. \tag{5.8}$$

The second set of functions to be introduced in this section are the functions G_{pL}, obtained by repeatedly applying the Laplace operator to the G_L defined above:

$$G_{pL}(\mathbf{r}) := \Delta^p G_L(\mathbf{r}) = \Delta^p(\mathscr{Y}_L(-\nabla) g(r)) = \mathscr{Y}_L(-\nabla)(\Delta^p g(r)). \tag{5.9}$$

To evaluate these functions in practice, a recurrence relation (in p) for the G_{pL} can be derived as follows. Due to

$$g(r) \in \mathscr{S}(\mathscr{R}^3) \Rightarrow \Delta^p g(r) \in \mathscr{S}(\mathscr{R}^3) \tag{5.10}$$

we can apply our general results of Sec. IV [cf. (4.3)] and introduce radial functions $g_{p\ell}(r)$, $\mu_{p\ell}(r)$, and $\varphi_{p\ell}(r)$ such that

$$G_{pL}(\mathbf{r}) = g_{p\ell}(r) Y_L(\hat{\mathbf{r}}) = \mu_{p\ell}(r) r^\ell Y_L(\hat{\mathbf{r}}) = \varphi_{p\ell}(r) \left(\frac{a^2}{\pi}\right)^{3/2} e^{\gamma\varepsilon} e^{-a^2 r^2} r^\ell Y_L(\hat{\mathbf{r}}). \tag{5.11}$$

We proceed by deriving a recurrence relation in p for the $\varphi_{p\ell}(r)$. Due to $G_{p+1,L}(\mathbf{r}) = \Delta G_{pL}(\mathbf{r})$ and the definition (5.11) of the $\varphi_{p\ell}(r)$ we have

$$\varphi_{p+1,\ell}(r) \left(\frac{a^2}{\pi}\right)^{3/2} e^{\gamma\varepsilon} e^{-a^2 r^2} r^\ell Y_L(\hat{\mathbf{r}}) = \Delta\left[\varphi_{p\ell}(r) \left(\frac{a^2}{\pi}\right)^{3/2} e^{\gamma\varepsilon} e^{-a^2 r^2} r^\ell Y_L(\hat{\mathbf{r}})\right]. \tag{5.12}$$

Using the Laplacian Δ in polar coordinates (4.17)/(4.18), this leads to

$$\varphi_{p+1,\ell}(r) = \frac{\partial^2}{\partial r^2}\varphi_{p\ell}(r) + \left[\frac{2(\ell+1)}{r} - 4a^2 r\right]\frac{\partial}{\partial r}\varphi_{p\ell}(r) + [4a^4 r^2 - 2a^2(2\ell+3)]\varphi_{p\ell}(r). \tag{5.13}$$

A comparison of the definition (5.11) of $\varphi_{p\ell}(r)$ with the explicit expression (5.7) for $G_L(\mathbf{r}) = G_{0L}(\mathbf{r})$ shows that

$$\varphi_{0\ell}(r) = (2a^2)^\ell \quad \text{for all } \ell \in \mathscr{N}_0. \tag{5.14}$$

From (5.13) we then obtain

$$\varphi_{1\ell}(r) = (2a^2)^{\ell+1}(2a^2 r^2 - (2\ell+3)) \quad \text{for all } \ell \in \mathscr{N}_0 \tag{5.15}$$

and that the $\varphi_{p\ell}(r)$ are of the form

$$\varphi_{p\ell}(r) = a_0 + a_2 r^2 + \ldots + a_{2p} r^{2p}. \tag{5.16}$$

Following (5.9), (5.8), and (3.8), the Fourier transform of $G_{pL}(\mathbf{r})$ is given by

$$\widehat{G_{pL}}(\mathbf{q}) = (-q^2)^p \mathscr{Y}_L(-i\mathbf{q}) e^{\gamma(\varepsilon - q^2)}. \tag{5.17}$$

Using this and relation (3.7), one can prove the following recurrence relation for the $G_{pL}(\mathbf{r})$ (by showing the equality of the Fourier transforms of the right- and left-hand sides):

$$G_{pL}(\mathbf{r}) = 4a^4 r^2 G_{p-1,L}(\mathbf{r}) - 2a^2(4p + 2\ell - 1) G_{p-1,L}(\mathbf{r}) - 8a^4(p-1)(2p+2\ell-1) G_{p-2,L}(\mathbf{r}) \tag{5.18}$$

for $p \geqslant 2$. Finally, from (5.18) and the definition (5.11) of $\varphi_{p\ell}(r)$ we immediately obtain the desired recurrence relation, which enables us to calculate the $\varphi_{p\ell}(r)$ for $p \geqslant 2$:

$$\varphi_{p\ell}(r)=4a^4r^2\varphi_{p-1,\ell}(r)-2a^2(4p+2\ell-1)\varphi_{p-1,\ell}(r)-8a^4(p-1)(2p+2\ell-1)\varphi_{p-2,\ell}(r). \tag{5.19}$$

Together with the explicit expressions for $\varphi_{0\ell}(r)$ and $\varphi_{1\ell}(r)$ given above, all required functions $\varphi_{p\ell}(r)$ can now be evaluated.

VI. UNSMOOTHED AND SMOOTHED HANKEL FUNCTIONS

In this section we present our convention for the (ordinary) Hankel functions $\check{H}_L$, define the smoothed Hankel functions H_L, derive a recurrence relation that allows the evaluation of the $H_L(\mathbf{r})$, and compare the smoothed Hankels H_L with the unsmoothed Hankels $\check{H}_L$. Next, we introduce the functions $H_{pL}(\mathbf{r}):=\Delta^p H_L(\mathbf{r})$ which are needed in recurrence relations, in the evaluation of the Hamiltonian and overlap matrix elements, and for the calculation of Coulomb integrals. Finally we give a recurrence relation for the calculation of the $H_{pL}(\mathbf{r})$ and obtain their Fourier transform $\widehat{H_{pL}}(\mathbf{q})$.

We start with the definition of the unsmoothed Hankels $\check{H}_L$ for the energy parameter $\varepsilon=-\kappa^2<0$ $(\kappa>0)$. They are singular at $\mathbf{r}=\mathbf{0}$ and are generated from a root function as discussed above:

$$\check{H}_L(\mathbf{r}):=\mathscr{Y}_L(-\nabla)\check{h}(r), \tag{6.1}$$

where

$$\check{h}(r):=\frac{e^{-\kappa r}}{r}. \tag{6.2}$$

The usual spherical Hankel functions of the first kind, $h^{(1)}_\ell(r)$, (see 10.1.1 in Ref. 9) and the $\check{H}_L$ are related by

$$\check{H}_L(\mathbf{r})=-i^\ell\kappa^{\ell+1}h^{(1)}_\ell(i\kappa r)Y_L(\hat{\mathbf{r}}). \tag{6.3}$$

The smoothed Hankel functions $H_L(\mathbf{r})$ with energy parameter $\varepsilon=-\kappa^2<0$ $(\kappa>0)$ and an additional ''smoothing'' parameter $a>0$ are defined in a similar way, starting from a modified root function $h(r)$:

$$H_L(\mathbf{r}):=\mathscr{Y}_L(-\nabla)h(r), \tag{6.4}$$

where

$$h(r):=\frac{1}{2r}(u_+(r)-u_-(r)), \tag{6.5}$$

$$u_\pm(r):=e^{\mp\kappa r}\left[1-\mathrm{erf}\left(\frac{\kappa}{2a}\mp ar\right)\right], \tag{6.6}$$

where erf(z) is the error function (cf. 7.1.1 in Ref. 9). It will be seen below that $h(r)$ is similar to $\check{h}(r)=\exp(-\kappa r)/r$ but is nonsingular. The function $h(r)$ is arbitrarily often continuously differentiable at every point in $\mathscr{R}^3$ including zero, i.e.,

$$h(r)\in C^\infty(\mathscr{R}^3), \tag{6.7}$$

and can be represented by a power series

$$\sum_{j=0}^{\infty}a_jr^{2j} \tag{6.8}$$

containing only even powers of r (cf. pp. 48/49 in Ref. 8). Using Eqs. (6.5)/(6.6) together with the property

$$\lim_{x\to\infty} \operatorname{erf}(\pm x) = \pm 1 \tag{6.9}$$

and the rule of de l'Hospital to determine $\lim_{r\to\infty} u_-(r)$ one can see:

$$h(r) \text{ behaves like } \frac{e^{-\kappa r}}{r} \text{ for large } r. \tag{6.10}$$

This fact together with (6.7) allows the conclusion

$$h(r) \in \mathscr{S}(\mathscr{R}^3) \tag{6.11}$$

[see (3.1) for the definition of $\mathscr{S}(\mathscr{R}^3)$].

Using the Laplacian Δ written in polar coordinates [Eq. (4.17)], one can prove by explicit calculation that $h(r)$ satisfies the following differential equation:

$$(\Delta+\varepsilon)h(a,\varepsilon;r) = -4\pi g(a,\varepsilon;r), \tag{6.12}$$

where we have noted the parameters a and ε of $h(r)$ [Eq. (6.5)] and of $g(r)$ [Eq. (5.2)] explicitly. Applying the operator $\mathscr{Y}_L(-\nabla)$ to both sides of (6.12) we obtain by means of (6.7), (6.4), and (5.1) the relation

$$(\Delta+\varepsilon)H_L(a,\varepsilon;\mathbf{r}) = -4\pi G_L(a,\varepsilon;\mathbf{r}). \tag{6.13}$$

To understand the relation between the smoothed and unsmoothed Hankel functions, it is useful to digress briefly into a formulation which involves distributions (avoided in the rest of this paper). It can be shown[11] that the corresponding differential equation for the unsmoothed Hankel functions is

$$(\Delta+\varepsilon)\check{H}_L(\varepsilon;\mathbf{r}) = -4\pi D_L(\mathbf{r}), \tag{6.14}$$

where

$$D_L(\mathbf{r}) = \mathscr{Y}_L(-\nabla)\,\delta(\mathbf{r}). \tag{6.15}$$

Comparing this with (6.13), we interpret the result as follows. The distribution $D_L(\mathbf{r})$ is a point multipole centered at the origin. The standard Hankel function $\check{H}_L(r)$ is the response of the Helmholtz operator $\Delta+\varepsilon$ to this singular source term. Our smoothed Hankel function $H_L(\mathbf{r})$ is the response of the same operator to a smeared-out version of the point multipole. This smooth multipole is the Gaussian $G_L(a,\varepsilon;\mathbf{r})$ which converges to $D_L(\mathbf{r})$ as $a\to\infty$. Since the source term is now smooth and analytical, the same is true for the function H_L itself. At large distances, the smooth Gaussians and the point multipoles are indistinguishable and the two types of functions coincide, by virtue of our definition of G_L involving an energy-dependent factor $e^{\gamma\varepsilon}$.

Still using distributions for now, the equations above are especially clear in Fourier space. Differential equation (6.13) translates to

$$(-q^2+\varepsilon)\widehat{H_L}(\mathbf{q}) = -4\pi\mathscr{Y}_L(-i\mathbf{q})e^{\gamma(\varepsilon-q^2)} \tag{6.16}$$

from which we deduce that the Fourier transform of the smoothed Hankel function is

$$\widehat{H_L}(\mathbf{q}) = \frac{-4\pi}{\varepsilon-q^2}\,\mathscr{Y}_L(-i\mathbf{q})e^{\gamma(\varepsilon-q^2)}. \tag{6.17}$$

These two equations apply to the standard Hankel function $\check{H}_L$ if γ is set to zero. The right-hand side of (6.16) then is $-4\pi\mathscr{Y}_L(-i\mathbf{q})$, which is the Fourier transform of the distribution $D_L(\mathbf{r})$ giving the point multipole.

Next, we derive a recurrence relation for the smoothed Hankel functions to permit their evaluation in practice. Due to (6.11) and $\Delta h(r) = -\varepsilon h(r) - 4\pi g(r)$ we can apply the results (4.3), (4.9), (4.10), (4.12), and (4.13)–(4.22) from Sec. IV. Using in addition the explicit formulas (6.5)/(6.6) for $h(r)$ and (5.6) for $\mu_\ell(r)$ we obtain

$$H_L(\mathbf{r}) = h_\ell(r) Y_L(\hat{\mathbf{r}}) = \chi_\ell(r) r^\ell Y_L(\hat{\mathbf{r}}), \tag{6.18}$$

$$\chi_0(r) = \frac{1}{2r}(u_+(r) - u_-(r)), \tag{6.19}$$

$$\chi_1(r) = \frac{1}{2r^3}(u_+(r) - u_-(r)) + \frac{\kappa}{2r^2}(u_+(r) + u_-(r)) - \frac{2a}{\sqrt{\pi} r^2} e^{\varepsilon/4a^2} e^{-a^2 r^2}, \tag{6.20}$$

$$\chi_{\ell+1}(r) = \frac{2\ell+1}{r^2}\chi_\ell(r) - \frac{\varepsilon}{r^2}\chi_{\ell-1}(r) - \frac{4\pi}{r^2}\left(\frac{a^2}{\pi}\right)^{3/2} e^{\varepsilon/4a^2}(2a^2)^{\ell-1} e^{-a^2 r^2} \quad \text{for } \ell \geqslant 1. \tag{6.21}$$

Equations (6.19)–(6.21) cannot directly be applied for $\mathbf{r} = \mathbf{0}$. For this point, using the rule of de l'Hospital we obtain

$$\chi_0(0) = -\kappa\left[1 - \operatorname{erf}\left(\frac{\kappa}{2a}\right)\right] + \frac{2a}{\sqrt{\pi}} e^{\varepsilon/4a^2} \tag{6.22}$$

and hence from (6.18) and $Y_0(\hat{\mathbf{r}}) = 1/\sqrt{4\pi}$,

$$H_0(\mathbf{0}) = -\frac{\kappa}{\sqrt{4\pi}}\left[1 - \operatorname{erf}\left(\frac{\kappa}{2a}\right)\right] + \frac{a}{\pi} e^{\varepsilon/4a^2}. \tag{6.23}$$

As we shall see below, the χ_ℓ are nonsingular for $r \to 0$ and therefore we have from (6.18)

$$H_L(\mathbf{0}) = 0 \quad \text{for all } \ell \neq 0. \tag{6.24}$$

In the following, we compare the behavior of the smoothed Hankel functions H_L with that of the unsmoothed Hankel functions $\check{H}_L$ in more detail. Considering first the radial parts of the functions near $r = 0$, note that we can represent $h(r)$ by a power series in r containing only even powers of r [see (6.8)]. Therefore we conclude from (4.10) and (4.12) that this is also true for all the $\chi_\ell(r)$. Furthermore, since $h_\ell(r)$ equals $\chi_\ell(r) r^\ell$ we can represent $h_\ell(r)$ by a power series in r which starts with a term proportional to r^ℓ. Hence the radial part $h_\ell(r)$ of $H_L(\mathbf{r})$ behaves as r^ℓ for small r, whereas the radial part of $\check{H}_L(\mathbf{r})$ behaves as $1/r^{\ell+1}$ (see p. 7127 in Ref. 14). In summary, $h_\ell(r)$ bends over near the origin and approaches 0 as r^ℓ whereas the radial part of $\check{H}_L(\mathbf{r})$ diverges when 0 is approached.

In Fig. 1 this behavior can be seen. For large values of r, the radial parts of H_L and $\check{H}_L$ become identical since from (6.10), (6.2), (6.4), and (6.1) we have

$$\lim_{|\mathbf{r}| \to \infty} H_L(a, \varepsilon; \mathbf{r}) = \check{H}_L(\varepsilon; \mathbf{r}). \tag{6.25}$$

Furthermore, from the definition (6.5)/(6.6) of $h(r)$ one can conclude

$$\lim_{a \to \infty} h(r) = \frac{e^{-\kappa r}}{r}. \tag{6.26}$$

Therefore we obtain from (6.2), (6.4), and (6.1) the relationship

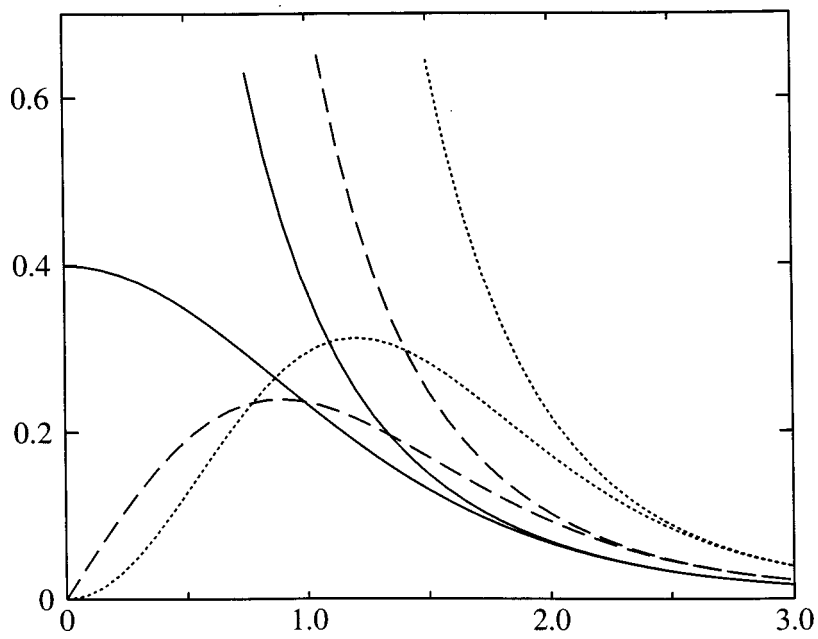

FIG. 1. Comparison of the radial parts of smoothed and unsmoothed Hankel functions for $\ell=0$ (continuous lines), $\ell=1$ (dashed lines) and $\ell=2$ (dotted lines). The energy parameter ε equals -1 in all cases and the smoothing parameter a for the smoothed Hankels equals 1.

$$\lim_{a\to\infty} H_L(a,\varepsilon;\mathbf{r})=\check{H}_L(\varepsilon;\mathbf{r}). \tag{6.27}$$

Consequently, as the parameter a is increased, the function $H_L(a,\varepsilon;\mathbf{r})$ becomes less and less ''smooth'' near the origin and begins to resemble the unsmoothed Hankel function. This motivates the name ''smoothing'' parameter for a. For large radii, the Gaussian source term in (6.13) is negligible and the differential equations (6.13) and (6.14) for the two types of functions H_L and $\check{H}_L$, respectively, are essentially the same.

Let us point out one important motivation of using the smoothed Hankel functions as basis functions in place of the normal unsmoothed functions. In the interstitial region close to muffin-tin spheres, the smoothed Hankel functions H_L show more resemblance to true wave functions than do the $\check{H}_L$. It follows that one can restrict oneself to substantially smaller basis sets with a corresponding gain in efficiency. To understand this, note that a true wave function ψ is a solution of Schrödinger's equation (here presented in atomic units):

$$\Delta\psi(\mathbf{r})=(-E+V(\mathbf{r}))\psi(\mathbf{r}). \tag{6.28}$$

The unsmoothed Hankel function $\check{H}_L(\varepsilon)$ with energy parameter $\varepsilon<0$ satisfies for $\mathbf{r}\neq\mathbf{0}$ the differential equation

$$\Delta\check{H}_L(\varepsilon;\mathbf{r})=-\varepsilon\check{H}_L(\varepsilon;\mathbf{r}) \tag{6.29}$$

(cf. p. 7130 in Ref. 14), i.e., it is a solution of Schrödinger's equation for energy $\varepsilon<0$ and $V(\mathbf{r})\equiv 0$. In comparison, according to (6.13) the smoothed Hankel function $H_L(a,\varepsilon)$ satisfies the differential equation

$$\Delta H_L(a,\varepsilon;\mathbf{r})=\left(-\varepsilon-4\pi\frac{G_L(a,\varepsilon;\mathbf{r})}{H_L(a,\varepsilon;\mathbf{r})}\right)H_L(a,\varepsilon;\mathbf{r}), \tag{6.30}$$

which is Schrödinger's equation for energy $\varepsilon<0$ and the potential

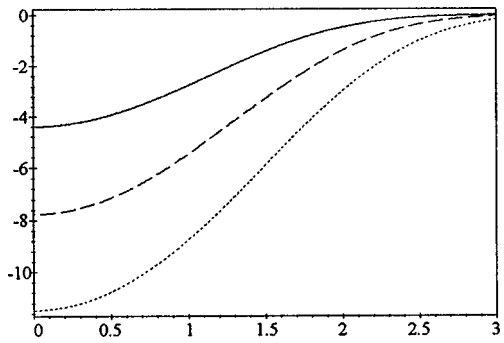

FIG. 2. The function $-4\pi G_L(a,\varepsilon;\mathbf{r})/H_L(a,\varepsilon;\mathbf{r})$ which plays the role of the potential when the differential equation (6.30) is interpreted as the Schrödinger equation, plotted along an arbitrary straight line starting from the origin for the case $\varepsilon=-1$, $a=1$ for $\ell=0$ (continuous line), $\ell=1$ (dashed) and $\ell=2$ (dotted). At intermediate distances outside the muffin-tin sphere, the functions can be adjusted to mimic the shape of the true potential which decreases as the Coulomb attraction of the screened nucleus becomes noticeable. This leads to a substantial reduction in the required basis size.

$$V(\mathbf{r})=-4\pi\frac{G_L(a,\varepsilon;\mathbf{r})}{H_L(a,\varepsilon;\mathbf{r})}. \tag{6.31}$$

This quantity is similar to a realistic potential at intermediate distances since it mimics the decrease of the screened Coulomb potential as the nucleus is approached (Fig. 2). In other words, a true wave function ''bends over'' as r becomes smaller and the potential drops below the absolute value of the energy because then the curvature (i.e., the second derivative) becomes negative. Such behavior is built into the smoothed Hankel functions but is not displayed by the normal Hankel functions. In practice, the shape of a wave function can be reasonably accurately represented by a single smoothed Hankel function outside the muffin-tin sphere or the core region if the two parameters ε and a are chosen appropriately, while a single standard Hankel function is generally inadequate.

To conclude this section, later on we will require the functions

$$H_{pL}(\mathbf{r}):=\Delta^p H_L(\mathbf{r}) \tag{6.32}$$

for which one can prove [by induction over p and using (6.13)] the following recurrence relation:

$$H_{0L}(a,\varepsilon;\mathbf{r})=H_L(a,\varepsilon;\mathbf{r}), \tag{6.33}$$

$$H_{p+1,L}(a,\varepsilon;\mathbf{r})=-\varepsilon H_{pL}(a,\varepsilon;\mathbf{r})-4\pi G_{pL}(a,\varepsilon;\mathbf{r}) \quad \text{for all } p\geqslant 0. \tag{6.34}$$

These equations enable one to calculate the $H_{pL}(\mathbf{r})$ in practice, whereby the $H_L(a,\varepsilon;\mathbf{r})$ are obtained from (6.18) to (6.24) and the $G_{pL}(a,\varepsilon;\mathbf{r})$ are obtained from (5.11), (5.14), (5.15), and (5.19). The Fourier transforms $\widehat{H_{pL}}$ follow from (6.32), (3.8), (6.13), and (5.8) and are given by

$$\widehat{H_{pL}}(\mathbf{q})=-\frac{4\pi}{\varepsilon-q^2}\,\mathscr{Y}_L(-i\mathbf{q})(-q^2)^p e^{\gamma(\varepsilon-q^2)}. \tag{6.35}$$

When p is set to zero, this expression gives the Fourier transform for the smoothed Hankel function $H_L(\mathbf{r})$. Since a product in reciprocal space corresponds to a convolution of the original functions, the following interesting interpretation arises. The smoothed Hankel function $H_L(\mathbf{r})$ can be seen as the convolution of the standard singular Hankel function $\check{H}_L$ [with Fourier transform $-4\pi\mathscr{Y}_L(-i\mathbf{q})/(\varepsilon-q^2)$] and a Gaussian (with Fourier transform $e^{-\gamma q^2}$), multiplied by our chosen normalization factor $e^{\gamma\varepsilon}$. Alternatively, it can up to a constant factor $1/\sqrt{4\pi}$ be considered the convolution of the generalized Gaussian G_L with the standard Hankel function of angular momentum zero. The two relevant length scales are the ''smoothing radius'' $R_{\mathrm{sm}}=1/a$ and the decay length $D=1/\kappa=1/\sqrt{-\varepsilon}$. These specify the width of the Gaussian and the Hankel functions,

respectively. If $R_{\mathrm{sm}} \ll D$, the smoothed Hankel function approaches the standard Hankel function, while a Gaussian is reproduced if $R_{\mathrm{sm}} \gg D$. Thus the smoothed Hankel functions interpolate between the two families and combine the features of both.

VII. THE FUNCTIONS $\dot{H}_L(\mathbf{r})$ AND $W_{pL}(\mathbf{r})$

To evaluate the elements of the kinetic energy and the overlap matrix we will need a further type of related functions, namely, the W_{pL} defined below. To calculate these explicitly we require the energy derivatives of the H_L, $\dot{H}_L$, which we consider first.

We denote the energy derivative of an arbitrary function f which depends on the energy parameter ε by $\dot{f}$ and we are interested here in the specific case

$$\dot{H}_L(\mathbf{r}) := \frac{\partial H_L(\mathbf{r})}{\partial \varepsilon}. \tag{7.1}$$

From (6.18) we have

$$\dot{H}_L(\mathbf{r}) = \dot{h}_\ell(r) Y_L(\hat{\mathbf{r}}) = \dot{\chi}_\ell(r) r^\ell Y_L(\hat{\mathbf{r}}). \tag{7.2}$$

To calculate $\dot{\chi}_\ell(r)$ we can derive a simple expression. Due to $\varepsilon = -\kappa^2$ we have

$$\frac{\partial}{\partial \varepsilon} = -\frac{1}{2\kappa}\frac{\partial}{\partial \kappa} \tag{7.3}$$

and from this we can obtain using (6.19)–(6.21) and (6.6)

$$\dot{\chi}_0(r) = \frac{1}{4\kappa}(u_+(r) + u_-(r)). \tag{7.4}$$

By induction over $\ell \in \dot{\mathcal{N}}_0$ it can be shown that

$$\dot{\chi}_{\ell+1}(r) = \tfrac{1}{2}\chi_\ell(r) \quad \text{for all } \ell \in \dot{\mathcal{N}}_0 \tag{7.5}$$

(cf. pp. 56/57 in Ref. 8). Therefore, we can calculate $\dot{H}_L(\mathbf{r})$ in the form (7.2) by calculating $\dot{\chi}_0(r)$ directly from (7.4) and obtaining the $\dot{\chi}_{\ell+1}(r)$ for $\ell \geqslant 0$ from (7.5). Hereby the $\chi_\ell(r)$ are obtained recursively from (6.19) to (6.21). For the special case $\mathbf{r} = \mathbf{0}$ we use

$$\dot{H}_0(\mathbf{0}) = \dot{\chi}_0(0)\frac{1}{\sqrt{4\pi}}, \tag{7.6}$$

$$\dot{H}_L(\mathbf{0}) = \lim_{\mathbf{r}\to\mathbf{0}} \dot{\chi}_\ell(r) r^\ell Y_L(\hat{\mathbf{r}}) = 0 \quad \text{for } \ell > 0. \tag{7.7}$$

Next, we turn to the $W_{pL}(\mathbf{r})$ mentioned above. As usual, they are generated from a root function $w(r)$ as

$$W_{pL}(\mathbf{r}) = \Delta^p \mathscr{Y}_L(-\nabla) w(r). \tag{7.8}$$

Here $w(r) \in \mathscr{S}(\mathscr{R}^3)$ with parameters $\varepsilon = -\kappa^2 < 0$ $(\kappa > 0)$ and $a > 0$ is defined by its Fourier transform $\widehat{w(r)}(\mathbf{q})$:

$$\widehat{w(r)}(\mathbf{q}) := \frac{4\pi e^{\gamma(\varepsilon - q^2)}}{(\varepsilon - q^2)^2}, \tag{7.9}$$

where γ equals $1/4a^2$ as before. According to (3.8) the Fourier transform of $W_{pL}(\mathbf{r})$ is

$$\widehat{W_{pL}}(\mathbf{q}) = \frac{4\pi}{(\varepsilon - q^2)^2}\, \mathscr{Y}_L(-i\mathbf{q})(-q^2)^p e^{\gamma(\varepsilon - q^2)}. \tag{7.10}$$

Comparison with (6.35) reveals that the derivative of W_{pL} with respect to γ is the negative of H_{pL}. Using (6.35) to write $H_L(\mathbf{r})$ in the form (3.6) we can get

$$\dot{H}_L(\mathbf{r}) = \frac{1}{(2\pi)^3} \int e^{i\mathbf{q}\cdot\mathbf{r}} \left[\frac{4\pi}{(\varepsilon - q^2)^2}\, \mathscr{Y}_L(-i\mathbf{q}) e^{\gamma(\varepsilon - q^2)} + \gamma \frac{-4\pi}{\varepsilon - q^2}\, \mathscr{Y}_L(-i\mathbf{q}) e^{\gamma(\varepsilon - q^2)} \right] d^3q. \tag{7.11}$$

By comparing this with (7.10) and (6.35) we obtain

$$W_{0L}(\mathbf{r}) = \dot{H}_L(\mathbf{r}) - \gamma H_L(\mathbf{r}). \tag{7.12}$$

Furthermore, one can prove the relation

$$W_{p+1,L}(\mathbf{r}) = -\varepsilon W_{pL}(\mathbf{r}) - H_{pL}(\mathbf{r}) \quad \text{for } p \geqslant 0 \tag{7.13}$$

by using (6.35) and (7.10) to show that the Fourier transforms of the right- and left-hand sides are equal. In sum, the $W_{pL}(\mathbf{r})$ can be obtained recursively from (7.12) and (7.13) if the $\dot{H}_L(\mathbf{r})$ and $H_{pL}(\mathbf{r})$ are calculated from (7.2) to (7.7) and (6.33)/(6.34), respectively.

VIII. OVERLAP AND HAMILTONIAN INTEGRALS IN THE NONPERIODIC CASE

In this section we discuss for the nonperiodic case the calculation of the overlap, kinetic energy, and potential integrals which are needed to evaluate the overlap and Hamiltonian matrix elements for the smoothed Hankel functions. As mentioned in Sec. I, these are the critical quantities when using the smoothed Hankel functions as basis set in any type of electronic structure calculation.

As preparation, we note the following important relationship. Let $P(\mathbf{r})$ and $Q(\mathbf{r})$ be two functions belonging to $\mathscr{S}(\mathscr{R}^3)$. Using Green's second identity (see, e.g., p. 350 in Ref. 15) for the case of a sphere around the origin and letting the radius of the sphere go to infinity shows

$$0 = \int [P(\mathbf{r})\Delta Q(\mathbf{r}) - Q(\mathbf{r})\Delta P(\mathbf{r})] d^3r. \tag{8.1}$$

In the following we treat the H_L and G_L as complex functions although they are actually real. We use this notation in order to be able to treat periodic structures involving Bloch sums of H_L and G_L in an analogous way as in the nonperiodic case. Moreover, where it is helpful we explicitly note the parameters $\gamma = 1/4a^2$ and $\varepsilon = -\kappa^2$ which define the Hankels and Gaussians.

First we consider the overlap integral $\int H_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r} - \mathbf{R}_1)^* H_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r} - \mathbf{R}_2) d^3r$ where H_{L_1} and H_{L_2} are smoothed Hankel functions centered at $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively, and $\varepsilon_1 \neq \varepsilon_2$. Due to (6.11) and (6.4), both $H_L(\mathbf{r})$ and $H_L(\mathbf{r} - \mathbf{R})$ belong to $\mathscr{S}(\mathscr{R}^3)$. By applying (8.1) and using (6.13) we obtain

$$\begin{aligned} &\int H_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r} - \mathbf{R}_1)^* H_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r} - \mathbf{R}_2) d^3r \\ &\quad = \frac{4\pi}{\varepsilon_1 - \varepsilon_2} \left\{ \begin{array}{l} \displaystyle\int H_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r} - \mathbf{R}_1)^* G_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r} - \mathbf{R}_2) d^3r \\ \displaystyle -\int G_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r} - \mathbf{R}_1)^* H_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r} - \mathbf{R}_2) d^3r \end{array} \right\}. \end{aligned} \tag{8.2}$$

The integral over a smoothed Hankel function times a Gaussian can be evaluated using (3.15), (3.14), (6.35), (5.8), (2.4), (2.9), (3.6) and again (6.35). The idea is to perform the integral in reciprocal space, by rewriting the product of the Fourier transforms of the two factors as a sum of Fourier transforms of smoothed Hankel functions. The phase factors for the positions $\mathbf{R}_1$ and $\mathbf{R}_2$ combine and we are left with back transforms at the vector $\mathbf{R}_1 - \mathbf{R}_2$. In detail:

$$\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* G_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= \frac{1}{(2\pi)^3}\int \frac{-4\pi}{\varepsilon_1-q^2}(-1)^{\ell_1}\mathcal{Y}_{L_1}(-i\mathbf{q})e^{\gamma_1(\varepsilon_1-q^2)}e^{i\mathbf{q}\cdot\mathbf{R}_1}\mathcal{Y}_{L_2}(-i\mathbf{q})e^{\gamma_2(\varepsilon_2-q^2)}e^{-i\mathbf{q}\cdot\mathbf{R}_2}d^3q$$

$$= e^{\gamma_2(\varepsilon_2-\varepsilon_1)}(-1)^{\ell_1}\frac{1}{(2\pi)^3}\int \frac{-4\pi}{\varepsilon_1-q^2}e^{(\gamma_1+\gamma_2)(\varepsilon_1-q^2)}e^{i\mathbf{q}\cdot(\mathbf{R}_1-\mathbf{R}_2)}$$

$$\times\sum_M C_{L_1L_2M}(-q^2)^{(\ell_1+\ell_2-m)/2}\mathcal{Y}_M(-i\mathbf{q})d^3q$$

$$= e^{\gamma_2(\varepsilon_2-\varepsilon_1)}(-1)^{\ell_1}\sum_M C_{L_1L_2M}H_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_1;\mathbf{R}_1-\mathbf{R}_2) \tag{8.3}$$

and similarly

$$\int G_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= e^{\gamma_1(\varepsilon_1-\varepsilon_2)}(-1)^{\ell_1}\sum_M C_{L_1L_2M}H_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_2;\mathbf{R}_1-\mathbf{R}_2). \tag{8.4}$$

These two relations also hold for $\varepsilon_1=\varepsilon_2$. From (8.2) to (8.4) we obtain the final result that the overlap integral for $\varepsilon_1\neq\varepsilon_2$ is given in closed form as

$$\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= 4\pi(-1)^{\ell_1}\sum_M C_{L_1L_2M}\left\{\begin{array}{l} \dfrac{e^{\gamma_2(\varepsilon_2-\varepsilon_1)}}{\varepsilon_1-\varepsilon_2}H_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_1;\mathbf{R}_1-\mathbf{R}_2) \\ +\dfrac{e^{\gamma_1(\varepsilon_1-\varepsilon_2)}}{\varepsilon_2-\varepsilon_1}H_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_2;\mathbf{R}_1-\mathbf{R}_2)\end{array}\right\}. \tag{8.5}$$

The values of the $H_{(\ell_1+\ell_2-m)/2,M}$ can be calculated from (6.33) and (6.34) because

$$C_{L_1L_2M} \text{ vanishes unless } \frac{\ell_1+\ell_2-m}{2}\in\mathcal{N}_0. \tag{8.6}$$

Equations (8.3), (8.4), and (8.5) are examples of an important general property: *Two-center integrals involving any combination of smoothed Hankel functions and Gaussians can be expressed analytically using functions of the same families, evaluated for the connecting vector between the two centers.* This is still the case when the kinetic energy operator or a Coulomb kernel is included in the integral. Furthermore, the analytical expressions are equally valid for Bloch-summed functions. These points are discussed below and in Sec. XI.

We still need the overlap integral in the case $\varepsilon_1=\varepsilon_2=:\varepsilon$ which involves, not surprisingly, the energy derivatives. It can be obtained by using (3.15), (3.14), (6.35), (2.4), (2.9), (3.6), and (7.10):

$$\int H_{L_1}(\gamma_1,\varepsilon;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= \frac{1}{(2\pi)^3}\int \frac{(4\pi)^2}{(\varepsilon-q^2)^2}e^{(\gamma_1+\gamma_2)(\varepsilon-q^2)}(-1)^{\ell_1}\mathcal{Y}_{L_1}(-i\mathbf{q})e^{i\mathbf{q}\cdot\mathbf{R}_1}\mathcal{Y}_{L_2}(-i\mathbf{q})e^{-i\mathbf{q}\cdot\mathbf{R}_2}d^3q$$

$$= (-1)^{\ell_1}\frac{4\pi}{(2\pi)^3}\int \frac{4\pi e^{(\gamma_1+\gamma_2)(\varepsilon-q^2)}}{(\varepsilon-q^2)^2}e^{i\mathbf{q}\cdot(\mathbf{R}_1-\mathbf{R}_2)}\sum_M C_{L_1L_2M}(-q^2)^{(\ell_1+\ell_2-m)/2}\mathcal{Y}_M(-i\mathbf{q})d^3q$$

$$= 4\pi(-1)^{\ell_1}\sum_M C_{L_1L_2M}W_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon;\mathbf{R}_1-\mathbf{R}_2). \tag{8.7}$$

The functions $W_{(\ell_1+\ell_2-m)/2,M}$ can be calculated from (7.12) and (7.13).

Next, the kinetic energy integrals follow almost immediately from our previous results. Using (6.13) we obtain

$$\begin{aligned}\int & H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*(-\Delta)H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \\ &= \varepsilon_2\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r + 4\pi \\ &\times\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*G_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \end{aligned} \tag{8.8}$$

as well as an equivalent expression obtained when the Laplace operator is moved to the first function in the integrand. This reduces the integral to cases which have already been handled in (8.2)–(8.7).

Finally, we discuss the potential integrals $\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*V(\mathbf{r})H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$ where $V(\mathbf{r})$ is a smooth potential. In an augmentation method, this will be the interstitial potential, extended smoothly through the muffin-tin spheres. In a pseudopotential method, it might be a smooth local potential defined in all of space to which nonlocal atomic terms are added later on. The evaluation of these integrals cannot be discussed before the representation of the potential has been chosen, but in any case will require careful thought. One sensible approach is to represent the smooth potential on a real-space mesh and to evaluate the integrals numerically. This exploits the advantage that realistic basis functions are often strongly smoothed, so that a relatively coarse mesh is adequate. For crystals, this approach has been successfully implemented as described in Ref. 5. A key step is to evaluate the smoothed Hankel functions efficiently on the real-space mesh. This is done by setting up the Fourier transforms (6.35) and applying a fast Fourier transform routine. In a similar manner, the output density can be calculated.

IX. EVALUATION OF BLOCH SUMS

In this section, we discuss the evaluation of the Bloch sums

$$H^b_{pL}(\gamma,\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}H_{pL}(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}), \tag{9.1}$$

$$W^b_{pL}(\gamma,\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}W_{pL}(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}), \tag{9.2}$$

where H_{pL} and W_{pL} have been introduced in the Secs. VI and VII, respectively, and $\mathbf{r}$ is some point in $\mathscr{R}^3$. These Bloch sums will be needed in Sec. X for the evaluation of the overlap and Hamiltonian matrix elements involving Bloch versions of smoothed Hankel functions.

As we will see below, for the calculation of the $H^b_{pL}(\gamma,\varepsilon;\mathbf{r})$ we also need the Gaussian Bloch sums

$$G^b_{pL}(\gamma,\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}G_{pL}(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}), \tag{9.3}$$

where the G_{pL} were defined in Sec. V. These can be evaluated by direct summation, whereby the $G_{pL}(\gamma,\varepsilon;\mathbf{r}-\mathbf{R})$ are obtained from (5.11), (5.14), (5.15), and (5.19). The sums converge rapidly because the G_{pL} are well-localized functions and therefore only a few terms in the sum (9.3) must be included.

The basic idea for the evaluation of the more slowly convergent sums is Ewald summation: Each term is split into a localized and a smooth part, and the corresponding sums are done in real

and reciprocal space, respectively. In the present case, the sum over the smoothed Hankels H_L [Eq. (9.1) for the case $p=0$] can by virtue of (3.18) and (6.35) be written as

$$H_L^b(\gamma,\varepsilon;\mathbf{r}) = H_L^b(\gamma',\varepsilon;\mathbf{r}) + (H_L^b(\gamma,\varepsilon;\mathbf{r}) - H_L^b(\gamma',\varepsilon;\mathbf{r}))$$
$$= \frac{1}{\Omega}\sum_{\mathbf{G}} \frac{-4\pi}{\varepsilon-|\mathbf{k}+\mathbf{G}|^2}\mathscr{Y}_L(-i(\mathbf{k}+\mathbf{G}))e^{\gamma'(\varepsilon-|\mathbf{k}+\mathbf{G}|^2)}e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}$$
$$+\sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}(H_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}) - H_L(\gamma',\varepsilon;\mathbf{r}-\mathbf{R})). \tag{9.4}$$

Here $\gamma'>\gamma$ is chosen to enable the rapid convergence of both $\Sigma_{\mathbf{G}}$ and $\Sigma_{\mathbf{R}}$ simultaneously. The first sum converges rapidly if γ' is large enough for a fast decay of the factor $e^{\gamma'(\varepsilon-|\mathbf{k}+\mathbf{G}|^2)}$ in reciprocal space. Convergence of the second sum comes about because the function $H_L(\gamma,\varepsilon;\mathbf{r})-H_L(\gamma',\varepsilon;\mathbf{r})$ is localized, since $H_L(\gamma,\varepsilon;\mathbf{r})$ and $H_L(\gamma',\varepsilon;\mathbf{r})$ coincide as soon as $|\mathbf{r}|$ becomes moderately large. The $H_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R})$ and $H_L(\gamma',\varepsilon;\mathbf{r}-\mathbf{R})$ in (9.4) are evaluated using (6.18)–(6.24). The values of the $H_{pL}^b(\gamma,\varepsilon;\mathbf{r})$ [Eq. (9.1)] for all $p\in\mathscr{N}_0$ can then be obtained from the recurrence relation

$$H_{0L}^b(\gamma,\varepsilon;\mathbf{r}) = H_L^b(\gamma,\varepsilon;\mathbf{r}), \tag{9.5}$$

$$H_{p+1,L}^b(\gamma,\varepsilon;\mathbf{r}) = -\varepsilon H_{pL}^b(\gamma,\varepsilon;\mathbf{r}) - 4\pi G_{pL}^b(\gamma,\varepsilon;\mathbf{r}) \tag{9.6}$$

following directly from the recurrence relation (6.33)/(6.34) for the H_{pL}, where the G_{pL}^b are obtained from (9.3).

In order to calculate W_{pL}^b of Eq. (9.2), we will require Bloch sums of the energy derivative:

$$\dot{H}_L^b(\gamma,\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\dot{H}_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}) \tag{9.7}$$

[see (7.1)] which can be evaluated by choosing a suitable parameter $\gamma'>\gamma$ in the Ewald sum

$$\dot{H}_L^b(\gamma,\varepsilon;\mathbf{r}) = \frac{1}{\Omega}\sum_{\mathbf{G}} \frac{4\pi}{\varepsilon-|\mathbf{k}+\mathbf{G}|^2}\left(\frac{1}{\varepsilon-|\mathbf{k}+\mathbf{G}|^2}-\gamma'\right)\mathscr{Y}_L(-i(\mathbf{k}+\mathbf{G}))e^{\gamma'(\varepsilon-|\mathbf{k}+\mathbf{G}|^2)}e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}$$
$$+\sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}(\dot{H}_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}) - \dot{H}_L(\gamma',\varepsilon;\mathbf{r}-\mathbf{R})). \tag{9.8}$$

This formula can be obtained by taking the derivative of formula (9.4) with respect to the energy parameter ε. Again, if γ' is chosen suitably, both sums in (9.8) converge rapidly. The energy derivatives $\dot{H}_L(\mathbf{r}-\mathbf{R})$ are evaluated according to (7.2)–(7.7).

After this preparation, we obtain W_{pL}^b from the recurrence relation

$$W_{0L}^b(\gamma,\varepsilon;\mathbf{r}) = \dot{H}_L^b(\gamma,\varepsilon;\mathbf{r}) - \gamma H_L^b(\gamma,\varepsilon;\mathbf{r}), \tag{9.9}$$

$$W_{p+1,L}^b(\gamma,\varepsilon;\mathbf{r}) = -\varepsilon W_{pL}^b(\gamma,\varepsilon;\mathbf{r}) - H_{pL}^b(\gamma,\varepsilon;\mathbf{r}), \tag{9.10}$$

which follows from (7.12)/(7.13), whereby the $H_{pL}^b(\gamma,\varepsilon;\mathbf{r})$ are calculated using (9.4)–(9.6).

X. OVERLAP AND HAMILTONIAN INTEGRALS IN THE PERIODIC CASE

In this section we show how to calculate the overlap and kinetic energy integrals for Bloch sums H_L^b of smoothed Hankel functions. As was discussed in Sec. VIII, the evaluation of the potential matrix elements is a separate problem which can, for example, be treated reasonably efficiently by numerical integration. This will not be considered further here. Exactly as in the nonperiodic case, two-center integrals can be expressed analytically using functions evaluated for the connecting vector between the centers. We recover the same expressions as before except that all appearing functions are now Bloch sums with the same wave vector **k**.

We require some preliminary considerations: Let

$$K^b(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} K(\mathbf{r}-\mathbf{R}) \tag{10.1}$$

and

$$L^b(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} L(\mathbf{r}-\mathbf{R}) \tag{10.2}$$

be Bloch sums for the same wave vector $\mathbf{k}$ where $K(\mathbf{r})$ and $L(\mathbf{r}) \in \mathscr{S}(\mathscr{R}^3)$. According to Green's second identity (see, e.g., p. 350 of Ref. 15) we have

$$\int_{\mathrm{UC}} (K^b(\mathbf{r})^* \Delta L^b(\mathbf{r}) - L^b(\mathbf{r}) \Delta K^b(\mathbf{r})^*) d^3r = \int_{\partial(\mathrm{UC})} \left(K^b(\mathbf{r})^* \frac{\partial}{\partial n} L^b(\mathbf{r}) - L^b(\mathbf{r}) \frac{\partial}{\partial n} K^b(\mathbf{r})^* \right) d^2r, \tag{10.3}$$

where the first integral is taken over a unit cell of the direct lattice and the second integral is taken over its surface, and where $\partial/\partial n$ denotes derivation in the direction of the outer normal on the unit cell surface. To every point $\mathbf{r}_0$ on the unit cell surface (with the exception of a finite number of vertex points) there corresponds some point

$$\mathbf{r}' = \mathbf{r}_0 + \mathbf{R}' \tag{10.4}$$

on the unit cell surface for a suitable direct lattice vector $\mathbf{R}' \neq \mathbf{0}$, such that the outer normal at $\mathbf{r}'$ is the negative of that in $\mathbf{r}_0$. This is a consequence of the fact that a unit cell, when translated through all direct lattice vectors, fills all of the space without either overlapping periodic images of itself or leaving voids. From (10.1) and (10.2) we have the Bloch property

$$K^b(\mathbf{r}+\mathbf{R}') = e^{i\mathbf{k}\cdot\mathbf{R}'} K^b(\mathbf{r}), \tag{10.5}$$

$$L^b(\mathbf{r}+\mathbf{R}') = e^{i\mathbf{k}\cdot\mathbf{R}'} L^b(\mathbf{r}) \tag{10.6}$$

for all $\mathbf{r} \in \mathscr{R}^3$. Consequently the contributions from opposite parts of the unit cell surface cancel in the right-hand side of (10.3), giving

$$\int_{\mathrm{UC}} K^b(\mathbf{r})^* \Delta L^b(\mathbf{r}) d^3r = \int_{\mathrm{UC}} L^b(\mathbf{r}) \Delta K^b(\mathbf{r})^* d^3r. \tag{10.7}$$

We are now able to proceed in straightforward analogy to the nonperiodic case [see (8.2)–(8.8)] whereby we use (10.7) in place of (8.1), (3.21) in place of (3.15), and (3.18) in place of (3.6). The two functions in question are again centered at the points $\mathbf{R}_1$ and $\mathbf{R}_2$ [not to be confused with the lattice vectors $\mathbf{R}$ used to write out Bloch sums as in (10.1)]. We obtain

$$\int_{\mathrm{UC}} H^b_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r}-\mathbf{R}_1)^* G^b_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r}-\mathbf{R}_2) d^3r$$

$$= (-1)^{\ell_1} e^{\gamma_2(\varepsilon_2-\varepsilon_1)} \sum_M C_{L_1 L_2 M} H^b_{(\ell_1+\ell_2-m)/2, M}(\gamma_1+\gamma_2, \varepsilon_1; \mathbf{R}_1-\mathbf{R}_2), \tag{10.8}$$

$$\int_{\mathrm{UC}} G^b_{L_1}(\gamma_1, \varepsilon_1; \mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2, \varepsilon_2; \mathbf{r}-\mathbf{R}_2) d^3r$$

$$= (-1)^{\ell_1} e^{\gamma_1(\varepsilon_1-\varepsilon_2)} \sum_M C_{L_1 L_2 M} H^b_{(\ell_1+\ell_2-m)/2, M}(\gamma_1+\gamma_2, \varepsilon_2; \mathbf{R}_1-\mathbf{R}_2), \tag{10.9}$$

giving the overlap integral in the case $\varepsilon_1 \neq \varepsilon_2$:

$$\int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r$$

$$= 4\pi(-1)^{\ell_1}\sum_M C_{L_1L_2M}\left(\begin{array}{l} \dfrac{e^{\gamma_2(\varepsilon_2-\varepsilon_1)}}{\varepsilon_1-\varepsilon_2} H^b_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_1;\mathbf{R}_1-\mathbf{R}_2) \\ + \dfrac{e^{\gamma_1(\varepsilon_1-\varepsilon_2)}}{\varepsilon_2-\varepsilon_1} H^b_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon_2;\mathbf{R}_1-\mathbf{R}_2) \end{array}\right). \tag{10.10}$$

The overlap integral in the case $\varepsilon_1=\varepsilon_2=:\varepsilon$ is given by:

$$\int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon;\mathbf{r}-\mathbf{R}_2)\,d^3r$$

$$= 4\pi(-1)^{\ell_1}\sum_M C_{L_1L_2M} W^b_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2,\varepsilon;\mathbf{R}_1-\mathbf{R}_2). \tag{10.11}$$

The required Bloch sums H^b_{pM} and W^b_{pM} can be evaluated as described in the preceding section. Finally, the kinetic energy integral is

$$\int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*(-\Delta) H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r$$

$$= \varepsilon_2 \int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r$$

$$+ 4\pi \int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* G^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r. \tag{10.12}$$

XI. COULOMB INTEGRALS

Up to now, the smoothed Hankel functions were mainly considered as a basis to represent the wave function. Once a localized representation is available, these functions are also useful to describe other quantities. For example, the implementation of Ref. 5 uses a combination of smoothed Hankel functions and Gaussians to represent the charge density of an atomic core. One important task is then to calculate the associated Coulomb integrals. That is, we need the integrals $\int\int P(\mathbf{r})Q(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|\,d^3r\,d^3r'$ where $P(\mathbf{r})Q(\mathbf{r}')$ is any combination of Gaussians and smoothed Hankels. We first discuss this problem for the nonperiodic case, then for periodic functions. The latter case is not fundamentally different but somewhat more complicated because of the long-range nature of the Coulomb interaction.

We are interested in Coulomb integrals of the following general form (written in atomic Rydberg units):

$$U = \int\int \frac{n_1(\mathbf{r})n_2(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d^3r\,d^3r' \tag{11.1}$$

where n_1 and n_2 are two charge densities. We can rewrite U as

$$U = \frac{1}{2}\int n_1(\mathbf{r})\phi_2(\mathbf{r})\,d^3r = \frac{1}{2}\int \phi_1(\mathbf{r})n_2(\mathbf{r})\,d^3r, \tag{11.2}$$

where $\phi_j(\mathbf{r})$ is the potential generated at point $\mathbf{r}$ by the charge distribution n_j, given by Coulomb's law as

$$\phi_j(\mathbf{r}) = 2\int \frac{n_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d^3r'. \tag{11.3}$$

This potential satisfies Poisson's equation

$$\Delta \phi_j(\mathbf{r}) = -8\pi n_j(\mathbf{r}). \tag{11.4}$$

Turning to the nonperiodic case first, our aim is to calculate the integral

$$U_{PQ} = \int\int \frac{P_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* Q_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}'-\mathbf{R}_2)}{|\mathbf{r}-\mathbf{r}'|} d^3r d^3r' \tag{11.5}$$

for all combinations $P,Q \in \{G,H\}$. We will do this using the potential generated by one of the two functions as in (11.2). To determine this potential, we first note that according to (6.13) and (5.7) we have

$$\Delta H_L(\gamma,\varepsilon;\mathbf{r}) = -4\pi e^{\gamma\varepsilon} G_L(\gamma,0;\mathbf{r}) - \varepsilon H_L(\gamma,\varepsilon;\mathbf{r}), \tag{11.6}$$

$$\Delta H_L(\gamma,0;\mathbf{r}) = -4\pi G_L(\gamma,0;\mathbf{r}). \tag{11.7}$$

Here $G_L(\gamma,0;\mathbf{r})$ denotes the function $G_L(\gamma,\varepsilon;\mathbf{r})$ for $\varepsilon=0$ and $H_L(\gamma,0;\mathbf{r})$ denotes the function $\lim_{\varepsilon\to 0} H_L(\gamma,\varepsilon;\mathbf{r})$. We see that the two potentials of interest are

$$\phi_{G_L}(\gamma,\varepsilon;\mathbf{r}) = 2e^{\gamma\varepsilon} H_L(\gamma,0;\mathbf{r}), \tag{11.8}$$

$$\phi_{H_L}(\gamma,\varepsilon;\mathbf{r}) = \frac{8\pi}{\varepsilon}(H_L(\gamma,\varepsilon;\mathbf{r}) - e^{\gamma\varepsilon} H_L(\gamma,0;\mathbf{r})) \tag{11.9}$$

since these functions satisfy Poisson's equation for the corresponding densities and go to zero at infinity. Inserting these expressions in (11.2) yields

$$U_{GG} = e^{\gamma_1\varepsilon_1} \int H_{L_1}(\gamma_1,0;\mathbf{r}-\mathbf{R}_1)^* G_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r, \tag{11.10}$$

$$U_{GH} = e^{\gamma_1\varepsilon_1} \int H_{L_1}(\gamma_1,0;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r, \tag{11.11}$$

$$U_{HH} = \frac{4\pi}{\varepsilon_1}\left\{ \begin{array}{l} \displaystyle\int H_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r \\ \displaystyle -e^{\gamma_1\varepsilon_1}\int H_{L_1}(\gamma_1,0;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r \end{array} \right\}. \tag{11.12}$$

This reduces all three Coulomb integrals to the integrals over products of Gaussians and smoothed Hankels for which we can use the expressions (8.3), (8.7), and (8.5), whereby the necessary values of the type $H_{pM}(\gamma,0;\mathbf{r}) = \lim_{\varepsilon\to 0} H_{pM}(\gamma,\varepsilon;\mathbf{r})$ occurring therein can be evaluated from:

$$H_{0M}(\gamma,0;\mathbf{r}) = H_M(\gamma,0;\mathbf{r}) = \left(\lim_{\varepsilon\to 0} \chi_m(\gamma,\varepsilon;r)\right) r^m Y_M(\hat{\mathbf{r}}), \tag{11.13}$$

$$\lim_{\varepsilon\to 0} \chi_0(\gamma,\varepsilon;r) = \frac{\mathrm{erf}(ar)}{r}, \tag{11.14}$$

$$\lim_{\varepsilon\to 0} \chi_1(\gamma,\varepsilon;r) = \frac{\mathrm{erf}(ar)}{r^3} - \frac{2a}{\sqrt{\pi} r^2} e^{-a^2r^2}, \tag{11.15}$$

$$\lim_{\varepsilon\to 0} \chi_{m+1}(\gamma,\varepsilon;r) = \frac{2m+1}{r^2} \lim_{\varepsilon\to 0} \chi_m(\gamma,\varepsilon;r) - \frac{4\pi}{r^2}\left(\frac{a^2}{\pi}\right)^{3/2} (2a^2)^{m-1} e^{-a^2r^2} \quad \text{for } m \geqslant 1, \tag{11.16}$$

for $\mathbf{r}\neq\mathbf{0}$,

$$H_M(\gamma,0;\mathbf{0}) = \frac{a}{\pi}\delta_{M0} \tag{11.17}$$

[cf. (6.18)–(6.21), (6.23), (6.24)] and

$$H_{p+1,M}(\gamma,0;\mathbf{r}) = -4\pi G_{pM}(\gamma,0;\mathbf{r}) \quad \text{for } p\geqslant 0 \tag{11.18}$$

[cf. (6.34)].

In the remainder of this section, we discuss the Coulomb integrals involving smoothed Hankels and Gaussians for the periodic case. Because of the long-ranged nature of the electrostatic potential, we must be careful about the way in which the limit $\varepsilon\to 0$ is done. We introduce new functions $F^b_{pL}(\gamma;\mathbf{r})$ by subtracting a homogeneous background from the Bloch function before taking the limit:

$$F^b_L(\gamma;\mathbf{r}) := \lim_{\varepsilon\to 0}\{H^b_L(\gamma,\varepsilon;\mathbf{r}) - \overline{H}^b_L(\gamma,\varepsilon)\}, \tag{11.19}$$

$$F^b_{pL}(\gamma;\mathbf{r}) := \Delta^p F^b_L(\gamma;\mathbf{r}), \tag{11.20}$$

where H^b_L is the previously defined Bloch sum of smoothed Hankels for the case $\mathbf{k}=\mathbf{0}$ and $\varepsilon<0$:

$$H^b_L(\gamma,\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} H_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}) \tag{11.21}$$

and the constant $\overline{H}^b_L$ is the average value of this Bloch sum over a unit cell:

$$\overline{H}^b_L(\gamma,\varepsilon) := \frac{1}{\Omega}\int_{\mathrm{UC}} H^b_L(\gamma,\varepsilon;\mathbf{r})d^3r. \tag{11.22}$$

In other words, we define the function $F^b_L(\gamma;\mathbf{r})$ by subtracting the average value from the Bloch sum over the smooth Hankels when $\varepsilon<0$ and then taking the limit for $\varepsilon\to 0$. In a representation as a Fourier sum over the reciprocal lattice, this is equivalent to simply leaving away the divergent term for $\mathbf{G}=\mathbf{0}$ and setting ε to zero in all others: Following (11.21), (3.18), and (6.35) the Fourier series for H^b_L is

$$H^b_L(\gamma,\varepsilon;\mathbf{r}) = \frac{1}{\Omega}\sum_{\mathbf{G}} \frac{-4\pi}{\varepsilon-G^2}\mathscr{Y}_L(-i\mathbf{G})e^{\gamma(\varepsilon-G^2)}e^{i\mathbf{G}\cdot\mathbf{r}} \tag{11.23}$$

and hence we obtain from (11.22) and (3.20) the average value

$$\overline{H}^b_L(\gamma,\varepsilon) = \frac{1}{\Omega}\widehat{H_L}(\mathbf{0}) = \frac{-4\pi}{\Omega}\frac{e^{\gamma\varepsilon}}{\varepsilon}Y_0\delta_{L0} = \frac{-\sqrt{4\pi}}{\Omega}\frac{e^{\gamma\varepsilon}}{\varepsilon}\delta_{L0}. \tag{11.24}$$

The factor δ_{L0} arises from $\mathscr{Y}_L(-i\mathbf{G}) = (-i)^{\ell}G^{\ell}Y_L(\mathbf{G})$, which is zero whenever $\mathbf{G}=\mathbf{0}$ and $L\neq 0$. From (11.19), (3.18), (11.24), and (6.35) we obtain the Fourier series for F^b_L:

$$F^b_L(\gamma;\mathbf{r}) = \frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}} \frac{4\pi}{G^2}\mathscr{Y}_L(-i\mathbf{G})e^{-\gamma G^2}e^{i\mathbf{G}\cdot\mathbf{r}}. \tag{11.25}$$

We see from (11.19), (11.21), (11.24), (11.25) that we can evaluate $F^b_L(\gamma;\mathbf{r})$ by Ewald summation in practice, using the formula

$$F_L^b(\gamma;\mathbf{r}) = F_L^b(\gamma';\mathbf{r}) + (F_L^b(\gamma;\mathbf{r}) - F_L^b(\gamma';\mathbf{r}))$$
$$= \frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}} \frac{4\pi}{G^2}\,\mathscr{Y}_L(-i\mathbf{G})e^{-\gamma' G^2}e^{i\mathbf{G}\cdot\mathbf{r}}$$
$$+\sum_{\mathbf{R}} (H_L(\gamma,0;\mathbf{r}-\mathbf{R}) - H_L(\gamma',0;\mathbf{r}-\mathbf{R})) + \frac{\sqrt{4\pi}}{\Omega}(\gamma-\gamma')\,\delta_{L0} \tag{11.26}$$

with some suitable $\gamma' > \gamma$. Here the last term is obtained from $\lim_{\varepsilon\to 0}\{\bar{H}_L^b(\gamma',\varepsilon) - \bar{H}_L^b(\gamma,\varepsilon)\}$ [cf. (11.24)] using the rule of de l'Hospital. The preceding expression involves terms of the type $H_L(\gamma,0,\mathbf{r}) = \lim_{\varepsilon\to 0} H_L(\gamma,\varepsilon,\mathbf{r})$, to be calculated according to (11.13)–(11.17).

Finally, to obtain the $F_{pL}^b(\gamma;\mathbf{r})$ for $p>0$ we first note that due to (11.20), (11.19), (6.13) and (11.23) we have

$$F_{1L}^b(\gamma;\mathbf{r}) = \lim_{\varepsilon\to 0}\{-\varepsilon H_L^b(\gamma,\varepsilon;\mathbf{r}) - 4\pi G_L^b(\gamma,\varepsilon;\mathbf{r})\} = \frac{\sqrt{4\pi}}{\Omega}\delta_{L0} - 4\pi G_L^b(\gamma,0;\mathbf{r}), \tag{11.27}$$

where G_L^b denotes the Bloch sum of the G_L for $\mathbf{k}=\mathbf{0}$. Using (11.27) we immediately get the result

$$F_{p+1,L}^b(\gamma,\mathbf{r}) = -4\pi G_{pL}^b(\gamma,0;\mathbf{r}) \quad \text{for } p\geqslant 1. \tag{11.28}$$

The $G_{pL}^b(\gamma,0;\mathbf{r})$ for $p\in\mathscr{N}_0$ are evaluated by direct summation as described in (9.3) with $\mathbf{k}$ set to zero. Using (11.26), (11.27), and (11.28), the $F_{pL}^b(\gamma;\mathbf{r})$ for all $p\in\mathscr{N}_0$ can be calculated.

After this somewhat tedious excursion to define the functions F_{pL}^b and to supply a method to calculate them, we can return to the Coulomb integrals in the case of periodic systems. For their evaluation we will need the integrals

$$\int_{\mathrm{UC}} F_{L_1}^b(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* G_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r, \tag{11.29}$$

$$\int_{\mathrm{UC}} G_{L_1}^b(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* F_{L_2}^b(\gamma_2;\mathbf{r}-\mathbf{R}_2)\,d^3r, \tag{11.30}$$

and

$$\int_{\mathrm{UC}} F_{L_1}^b(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* H_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r, \tag{11.31}$$

which are discussed next. For the evaluation of (11.29) and (11.30) we use Eq. (11.25) for F_L^b, proceed in the same manner as in Sec. X to evaluate the integrals (10.8) and (10.9), and obtain the first two integrals in closed form:

$$\int_{\mathrm{UC}} F_{L_1}^b(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* G_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\,d^3r$$
$$= e^{\gamma_2\varepsilon_2}(-1)^{\ell_1}\sum_M C_{L_1L_2M} F^b_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2;\mathbf{R}_1-\mathbf{R}_2), \tag{11.32}$$

$$\int_{\mathrm{UC}} G_{L_1}^b(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* F_{L_2}^b(\gamma_2;\mathbf{r}-\mathbf{R}_2)\,d^3r$$
$$= e^{\gamma_1\varepsilon_1}(-1)^{\ell_1}\sum_M C_{L_1L_2M} F^b_{(\ell_1+\ell_2-m)/2,M}(\gamma_1+\gamma_2;\mathbf{R}_1-\mathbf{R}_2). \tag{11.33}$$

Since H_L^b and F_L^b are Bloch functions for the same wave vector $\mathbf{k}=\mathbf{0}$ we have following (10.7) the relation

$$\int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* \Delta H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r = \int_{\mathrm{UC}} H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)\Delta F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* d^3r. \tag{11.34}$$

As before, we use this expression to reduce the third required integral (11.31) to the previous cases. Applying (6.13) and (11.27) gives

$$\int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= \frac{1}{\varepsilon_2}\left\{ \begin{array}{c} 4\pi \int_{\mathrm{UC}} G^b_{L_1}(\gamma_1,0;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \\ -4\pi \int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* G^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \\ -(\sqrt{4\pi}/\Omega)\,\delta_{L_1 0}\int_{\mathrm{UC}} H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \end{array} \right\}. \tag{11.35}$$

From (3.18), (3.20), (6.35) we see

$$\int_{\mathrm{UC}} H^b_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R})d^3r = \widehat{H_L}(\mathbf{0}) = -\sqrt{4\pi}\,\frac{e^{\gamma\varepsilon}}{\varepsilon}\,\delta_{L0} \tag{11.36}$$

giving the final expression for the integral (11.31) in terms of the previous cases (10.9) and (11.32):

$$\int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r$$

$$= \frac{4\pi}{\varepsilon_2}\left\{ \begin{array}{c} \int_{\mathrm{UC}} G^b_{L_1}(\gamma_1,0;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \\ -\int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* G^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r \\ +\frac{1}{\Omega}\,\frac{e^{\gamma_2\varepsilon_2}}{\varepsilon_2}\,\delta_{L_1 0}\delta_{L_2 0} \end{array} \right\}. \tag{11.37}$$

Finally, the two-center Coulomb integrals for the periodic case can be formulated. As before, we want to calculate the potential for one of the factors, then integrate over the product with the other factor. However, if $P(\mathbf{r})$ is any function representing a periodic charge density, then we can find a *periodic* potential ϕ_P satisfying $\Delta\phi_P = -8\pi P$ only if the neutrality condition is satisfied:

$$\int_{\mathrm{UC}} P(\mathbf{r})d^3r = 0. \tag{11.38}$$

Therefore, a sensible aim is to derive expressions for the integrals

$$\int_{\mathrm{UC}}\int_{\mathrm{UC}} \frac{\tilde{P}_{L_1}(\mathbf{r}-\mathbf{R}_1)^* \tilde{Q}_{L_2}(\mathbf{r}'-\mathbf{R}_2)}{|\mathbf{r}-\mathbf{r}'|} d^3r d^3r', \tag{11.39}$$

where $P(\mathbf{r})$ and $Q(\mathbf{r})$ are Bloch sums of smoothed Hankels or Gaussians, and $\tilde{P}$ and $\tilde{Q}$ are the corresponding functions with a possible constant term subtracted:

$$\tilde{P}_L(\mathbf{r}) := P_L(\mathbf{r}) - \frac{1}{\Omega}\int_{\mathrm{UC}} P_L(\mathbf{r})d^3r, \tag{11.40}$$

$$\tilde{Q}_L(\mathbf{r}) := Q_L(\mathbf{r}) - \frac{1}{\Omega}\int_{\mathrm{UC}} Q_L(\mathbf{r})d^3r. \tag{11.41}$$

In a properly formulated application of the resulting expressions, the various additional homogeneous terms will cancel in the final result. Our plan is therefore to calculate the integral

$$U = \frac{1}{2}\int_{\mathrm{UC}} \phi_{\tilde{P}_{L_1}}(\mathbf{r}-\mathbf{R}_1)^* \tilde{Q}_{L_2}(\mathbf{r}-\mathbf{R}_2)d^3r, \tag{11.42}$$

where $\phi_{\tilde{P}_L}(\mathbf{r})$ satisfies Poisson's equation

$$\Delta\phi_{\tilde{P}_L}(\mathbf{r}) = -8\pi\tilde{P}_L(\mathbf{r}). \tag{11.43}$$

Of course, the potential $\phi_{\tilde{P}_L}$ is determined only up to an additive constant. But since $\tilde{Q}_{L_2}(\mathbf{r})$ satisfies the neutrality condition by construction, adding an arbitrary constant to $\phi_{\tilde{P}_L}$ will not change the value of the integral (11.42). We therefore need not impose any additional condition besides (11.43) on $\phi_{\tilde{P}_L}$ in order to obtain the correct value of U.

First, we subtract the average values from G_L^b and H_L^b to make functions which satisfy the neutrality condition [cf. (3.18), (3.20), (5.8) and (11.36)]:

$$\tilde{G}_L^b(\gamma,\varepsilon;\mathbf{r}) = G_L^b(\gamma,\varepsilon;\mathbf{r}) - \frac{1}{\Omega}\frac{e^{\gamma\varepsilon}}{\sqrt{4\pi}}\delta_{L0} \tag{11.44}$$

and

$$\tilde{H}_L^b(\gamma,\varepsilon;\mathbf{r}) = H_L^b(\gamma,\varepsilon;\mathbf{r}) + \frac{\sqrt{4\pi}}{\Omega}\frac{e^{\gamma\varepsilon}}{\varepsilon}\delta_{L0}. \tag{11.45}$$

Next, we obtain two potentials which satisfy the Poisson equation for G_L^b and H_L^b, using (11.27), (5.7), (11.44) for the first and (6.13), (11.27), (5.7), (11.45) for the second:

$$\phi_{\tilde{G}_L^b}(\gamma,\varepsilon;\mathbf{r}) = 2e^{\gamma\varepsilon}F_L^b(\gamma;\mathbf{r}), \tag{11.46}$$

$$\phi_{\tilde{H}_L^b}(\gamma,\varepsilon;\mathbf{r}) = \frac{8\pi}{\varepsilon}(H_L^b(\gamma,\varepsilon;\mathbf{r}) - e^{\gamma\varepsilon}F_L^b(\gamma;\mathbf{r})). \tag{11.47}$$

Finally, we exploit that $F_L^b(\gamma;\mathbf{r})$ satisfies the neutrality condition according to the definition (11.19) and assemble the final expressions for the two-center Coulomb integrals by means of (11.42), the previous equations, and (11.36):

$$\begin{aligned} U_{GG} &= \int_{\mathrm{UC}}\int_{\mathrm{UC}} \frac{\tilde{G}_{L_1}^b(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*\tilde{G}_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}'-\mathbf{R}_2)}{|\mathbf{r}-\mathbf{r}'|}d^3r\,d^3r' \\ &= \frac{1}{2}\int_{\mathrm{UC}} \phi_{\tilde{G}_{L_1}^b}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*\left(G_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) - \frac{1}{\Omega}\frac{e^{\gamma_2\varepsilon_2}}{\sqrt{4\pi}}\delta_{L_20}\right)d^3r \\ &= e^{\gamma_1\varepsilon_1}\int_{\mathrm{UC}} F_{L_1}^b(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* G_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2)d^3r, \end{aligned} \tag{11.48}$$

$$U_{GH} = \int_{\mathrm{UC}}\int_{\mathrm{UC}} \frac{\tilde{G}_{L_1}^b(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^*\tilde{H}_{L_2}^b(\gamma_2,\varepsilon_2;\mathbf{r}'-\mathbf{R}_2)}{|\mathbf{r}-\mathbf{r}'|}d^3r\,d^3r'$$

$$= \frac{1}{2} \int_{\mathrm{UC}} \phi_{\tilde{G}^b_{L_1}}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* \left(H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) + \frac{\sqrt{4\pi}}{\Omega} \frac{e^{\gamma_2\varepsilon_2}}{\varepsilon_2} \delta_{L_2 0} \right) d^3r$$

$$= e^{\gamma_1\varepsilon_1} \int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r, \tag{11.49}$$

$$U_{HH} = \int_{\mathrm{UC}} \int_{\mathrm{UC}} \frac{\tilde{H}^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* \tilde{H}^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}'-\mathbf{R}_2)}{|\mathbf{r}-\mathbf{r}'|} d^3r d^3r'$$

$$= \frac{1}{2} \int_{\mathrm{UC}} \phi_{\tilde{H}^b_{L_1}}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* \left(H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) + \frac{\sqrt{4\pi}}{\Omega} \frac{e^{\gamma_2\varepsilon_2}}{\varepsilon_2} \delta_{L_2 0} \right) d^3r$$

$$= \frac{4\pi}{\varepsilon_1} \left\{ \begin{array}{c} \displaystyle\int_{\mathrm{UC}} H^b_{L_1}(\gamma_1,\varepsilon_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r \\ \displaystyle -e^{\gamma_1\varepsilon_1} \int_{\mathrm{UC}} F^b_{L_1}(\gamma_1;\mathbf{r}-\mathbf{R}_1)^* H^b_{L_2}(\gamma_2,\varepsilon_2;\mathbf{r}-\mathbf{R}_2) d^3r \end{array} \right\}$$

$$- \frac{\Omega}{\varepsilon_1} \frac{4\pi e^{\gamma_1\varepsilon_1}}{\Omega\varepsilon_1} \frac{4\pi e^{\gamma_2\varepsilon_2}}{\Omega\varepsilon_2} \delta_{L_1 0} \delta_{L_2 0}. \tag{11.50}$$

As usual, the integrals in these expressions can be calculated using functions evaluated for the connecting vector $\mathbf{R}_1 - \mathbf{R}_2$. The relevant equations are (11.32), (11.37), (10.10), and (10.11).

XII. EXPANSION OF A SMOOTH FUNCTION AROUND A SITE

Modern electronic structure calculations often use either augmentation or nonlocal pseudopotentials to introduce atomic detail at the atomic sites. In either approach it is neccessary to expand the basis functions as a sum of angular-momentum components around the various atomic sites. In this section, we present a method to do this, based on biorthogonal polynomials to the Gaussian family $G_{kL}(\mathbf{r})$. This approach can be applied to any smooth function, in particular to the smoothed Hankel functions and to plane waves. The result will be to approximate a given smooth function $F(\mathbf{r})$ by a finite expansion of the form

$$F(\mathbf{r}) \sim \sum_{kL} a_{kL} P_{kL}(\mathbf{r}). \tag{12.1}$$

The expansion functions P_{kL} are polynomials in x, y, and z which will be defined using the radial functions $\varphi_{k\ell}$ of the generalized Gaussians (5.11). The advantage over a Taylor series is that the expansion (12.1) converges evenly over some chosen range, making it significantly more useful in practical applications.

In Sec. V, the Gaussians $G_{kL}(\gamma,\varepsilon;\mathbf{r})$ were defined with an energy-dependent factor $e^{\gamma\varepsilon}$ in order to simplify expressions involving Gaussians and smoothed Hankel functions. Throughout this section, ε is zero and will be suppressed in the notation. Gaussians are therefore specified by one parameter γ which gives the decay of the functions as $e^{-a^2r^2}$ with $a^2 = 1/(4\gamma)$. Equivalently, we speak of the smoothing radius $R_{\mathrm{sm}} = 1/a$ which is the width of the Gaussian. As preparation, we prove two relationships for the functions G_{kL}.

(a) The radial functions $\varphi_{k\ell}(r)$ which define the generalized Gaussians as in (5.11) by

$$G_{kL}(\gamma;\mathbf{r}) = \varphi_{k\ell}(r) \left(\frac{a^2}{\pi} \right)^{3/2} e^{-a^2r^2} r^\ell Y_L(\hat{\mathbf{r}}) \tag{12.2}$$

are related to the generalized Laguerre polynomials $L_n^{(\alpha)}(x)$ (see Chapter 22 in Ref. 9) by

$$\varphi_{k\ell}(r) = (-1)^k (2a^2)^{k+\ell} 2^k k! L_k^{(\ell+1/2)}(a^2r^2). \tag{12.3}$$

Proof: We proceed by induction over $k \in \mathcal{N}_0$. For $k \leqslant 1$, Eq. (12.3) can be confirmed by comparing the explicit expressions for $L_0^{(\ell+1/2)}(x)$ and $L_1^{(\ell+1/2)}(x)$ (see 22.3.9 in Ref. 9) with expressions (5.14) and (5.15) for $\varphi_{0\ell}(r)$ and $\varphi_{1\ell}(r)$, respectively. Assume we have already shown (12.3) for all $k<k_0$ where $k_0 \geqslant 2$. By applying the recurrence relation for $L_n^{(\alpha)}(a^2r^2)$ (22.7.12 in Ref. 9) and the induction hypothesis, the right-hand side of (12.3) for $k=k_0$ is transformed into the right-hand side of the recurrence relation for the $\varphi_{p\ell}(r)$ (Eq. (5.19)) with p replaced by k_0. This completes the proof for all $k \in \mathcal{N}_0$. ◇

(b) The functions G_{kL} are orthogonal in both k and L in the following sense:

$$\int G_{kL}(\gamma;\mathbf{r}) G_{k'L'}(\gamma;\mathbf{r}) e^{a^2r^2} d^3r = \frac{1}{4\pi}\left(\frac{a^2}{\pi}\right)^{3/2} (2a^2)^{2k+\ell} 2^k k!(2k+2\ell+1)!!\,\delta_{kk'}\delta_{LL'}. \tag{12.4}$$

Proof: In the case $L \neq L'$ the orthogonality follows immediately from (12.2) and the orthogonality of the spherical harmonics $Y_L(\hat{\mathbf{r}})$. For the case $L=L'$, the result is obtained using (12.2), Eq. (2.9) for the case $K=L$, the orthogonality of the $Y_M(\hat{\mathbf{r}})$, $C_{LL0}=1/\sqrt{4\pi}$, (12.3) and the orthogonality relation of the generalized Laguerre polynomials

$$\int_0^\infty e^{-u} u^{\ell+1/2} L_k^{(\ell+1/2)}(u) L_{k'}^{(\ell+1/2)}(u)\,du = \frac{\Gamma(k+\ell+3/2)}{k!}\,\delta_{kk'} \tag{12.5}$$

(cf. 22.2.12 in Ref. 9) with the substitution $u=a^2r^2$. ◇

Returning to the desired expansion (12.1), we define the polynomials $P_{kL}(\gamma;\mathbf{r})$ by scaling $\varphi_{k\ell}(r)$ in a convenient way:

$$P_{kL}(\gamma;\mathbf{r}) = p_{k\ell}(\gamma;r) Y_L(\hat{\mathbf{r}}), \tag{12.6}$$

$$p_{k\ell}(\gamma;r) = \varphi_{k\ell}(r)\,\frac{a^\ell (2\ell+1)!!}{(2a^2)^{k+\ell}(2k+2\ell+1)!!}\,r^\ell. \tag{12.7}$$

The radial functions $\varphi_{k\ell}$ [see (5.16)] as well as the $r^\ell Y_L(\hat{\mathbf{r}}) = \mathscr{Y}_L(\mathbf{r})$ are polynomials in x,y,z. Therefore, we see that the P_{kL} are indeed polynomials in the Cartesian coordinates of $\mathbf{r}$. The $P_{kL}(\gamma;\mathbf{r})$ are biorthogonal to the G_{kL} as follows from Eqs. (12.6)/(12.7), (12.2), and (12.4):

$$\int G_{kL}(\gamma;\mathbf{r}) P_{k'L'}(\gamma;\mathbf{r})\,d^3r = \frac{1}{4\pi}(2a^2)^k a^\ell 2^k k!(2\ell+1)!!\,\delta_{kk'}\delta_{LL'}. \tag{12.8}$$

The significance of the biorthogonality is that, in order to expand a function as a linear combination of the P_{kL}, we can use the Gaussians G_{kL} to project out the coefficients. Consider some function $F(\mathbf{r})$ and assume that

$$\int_0^{2\pi}\int_0^{\pi} |F(r,\theta,\varphi)|^2 \sin\theta\, d\theta\, d\varphi < \infty \quad \text{for all } r \geqslant 0. \tag{12.9}$$

Such a function can be written in the form

$$F(\mathbf{r}) = \sum_L f_L(r) r^\ell Y_L(\hat{\mathbf{r}}) \tag{12.10}$$

because the $Y_L(\hat{\mathbf{r}})$ form a complete basis of $\mathscr{L}_2(S)$ [cf.(2.5)]. Now, if $f_L(r)$ satisfies the condition

$$\int_0^\infty e^{-u} u^{\ell+1/2}\left[f_L\left(\frac{\sqrt{u}}{a}\right)\right]^2 du < \infty$$

$$\text{where } u=a^2r^2, \tag{12.11}$$

we can expand $f_L(r)$ in terms of the generalized Laguerre polynomials $L_k^{(\ell+1/2)}(u)$ by using (12.5):

$$f_L\left(\frac{\sqrt{u}}{a}\right)\sim\sum_{k=0}^{\infty} c_{kL}(\gamma)L_k^{(\ell+1/2)}(u), \tag{12.12}$$

$$c_{kL}(\gamma)=\frac{k!}{\Gamma(k+\ell+3/2)}\int_0^\infty e^{-u}u^{\ell+1/2}f_L\left(\frac{\sqrt{u}}{a}\right)L_k^{(\ell+1/2)}(u)du. \tag{12.13}$$

For a finite approximation obtained by truncating the sum at some $n\in\mathcal{N}_0$ we have the property

$$\int_0^\infty e^{-u}u^{\ell+1/2}\left[f_L\left(\frac{\sqrt{u}}{a}\right)-\sum_{k=0}^{n} c_{kL}(\gamma)L_k^{(\ell+1/2)}(u)\right]^2du=\min_{\zeta\in\Pi_n}\int_0^\infty e^{-u}u^{\ell+1/2}\left[f_L\left(\frac{\sqrt{u}}{a}\right)-\zeta(u)\right]^2du, \tag{12.14}$$

where Π_n is the set of all polynomials of degree $\leqslant n$ (see, e.g., pp. 98–101 in Ref. 16). Using (12.3) and (12.7), expansion (12.12) translates to a linear combination of the radial functions $p_{k\ell}(r)$:

$$f_L(r)r^\ell\sim\sum_{k=0}^{\infty} a_{kL}(\gamma)p_{k\ell}(\gamma;r) \tag{12.15}$$

for suitable coefficients $a_{kL}(\gamma)$. By multiplying through with $Y_L(\hat{\mathbf{r}})$ we obtain the desired expansion (12.1) of $F(\mathbf{r})$, which is valid provided (12.11) holds for all angular-momentum components L. Because of the biorthogonality relation (12.8), the expansion coefficients $a_{kL}(\gamma)$ can be obtained by integration:

$$a_{kL}(\gamma)=\frac{4\pi}{(2a^2)^k a^\ell 2^k k!(2\ell+1)!!}\int G_{kL}(\gamma;\mathbf{r})F(\mathbf{r})d^3r. \tag{12.16}$$

As a specific instance, the aim might be to expand the smoothed Hankel function $H_{L'}(\gamma',\varepsilon';\mathbf{r}-\mathbf{R}')$ centered at $\mathbf{R}'$ around some other site $\mathbf{R}$. We obtain an expansion in the polynomials $P_{kL}(\gamma;\mathbf{r}-\mathbf{R})$ centered at $\mathbf{R}$:

$$H_{L'}(\gamma',\varepsilon';\mathbf{r}-\mathbf{R}')\sim\sum_{k,L} a_{kL}(\gamma,\mathbf{R}-\mathbf{R}')P_{kL}(\gamma;\mathbf{r}-\mathbf{R}) \tag{12.17}$$

whereby the coefficients are

$$\begin{aligned}a_{kL}(\gamma,\mathbf{R}-\mathbf{R}')&=\frac{4\pi}{(2a^2)^k a^\ell 2^k k!(2\ell+1)!!}\int G_{kL}(\gamma;\mathbf{r}-\mathbf{R})H_{L'}(\gamma',\varepsilon';\mathbf{r}-\mathbf{R}')d^3r\\&=\frac{4\pi}{(2a^2)^k a^\ell 2^k k!(2\ell+1)!!}e^{-\gamma\varepsilon'}(-1)^\ell\\&\quad\times\sum_M C_{LL'M}H_{(\ell+\ell'-m)/2+k,M}(\gamma+\gamma',\varepsilon';\mathbf{R}-\mathbf{R}')\end{aligned} \tag{12.18}$$

[compare (8.4)], i.e., the involved integral reduces to a sum of smoothed Hankel functions of the connecting vector $\mathbf{R}-\mathbf{R}'$. Altogether we obtain an analytical expansion theorem for the smoothed Hankel functions around an arbitrary point in space. In contrast to the well-known structure-constant expansion[17] for the standard Hankel functions, the present expression can be applied even when point $\mathbf{R}$ lies close to or is identical with the center of the function at $\mathbf{R}'$. Furthermore, analogous expressions can be used to expand a Bloch-summed function $H^b_{L'}(\gamma',\varepsilon';\mathbf{r}-\mathbf{R}')$. Hereby the integral over all space as in (12.18) is reduced to an integral over the unit cell which

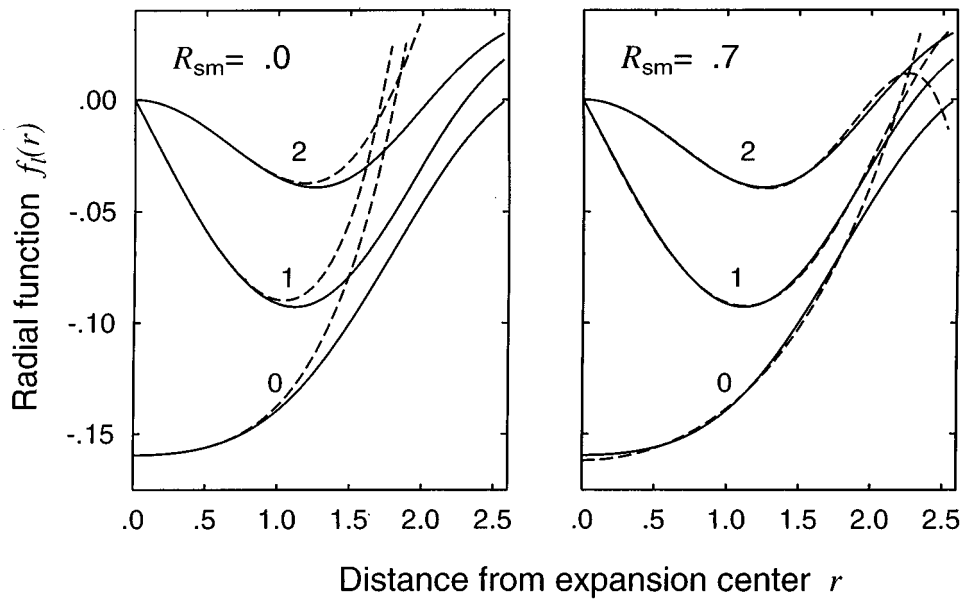

FIG. 3. Expansion of a smoothed Hankel function centered at point $\mathbf{R}'=(0,0,2)$ as a sum of the polynomials P_{kL} around the origin. The Hankel function has angular momentum $(\ell',m_{\ell'})=(1,0)$, energy $\varepsilon'=-1$, and smoothing radius $R'_{\rm sm}=1/a'=1$. Shown are the radial components for $(\ell,m_\ell)=(0,0)$, (1,0), and (2,0) for two choices of the smoothing radius $R_{\rm sm}=1/a$ when expanded up to $k_{\max}=2$. For $R_{\rm sm}=0$, the expansion reduces to a Taylor series, which is accurate near the center but deviates in an uncontrolled manner already at moderate distances. For $R_{\rm sm}=0.7$ the error in the expansion is spread evenly over the range where the Gaussian is large, here approximately to $r=1.5$.

is then evaluated analytically following (10.9). On the other hand, if the aim is to expand a plane wave, the coefficients are given by integrals over the plane wave times $G_{kL}(\gamma;\mathbf{r}-\mathbf{R})$. This simply picks out the corresponding Fourier coefficient of G_{kL}.

For a discussion of the convergence of the expansion (12.1), the central quantity is the width $R_{\rm sm}$ of the underlying Gaussian $g_0(r)=e^{-a^2r^2}=e^{-r^2/R_{\rm sm}^2}$ which defines the family $G_{kL}(\mathbf{r})$. By calculating the expansion coefficients according to (12.16), we are projecting out information about the function $F(\mathbf{r})$ in that part of volume where g_0 is non-negligible. Correspondingly, we obtain a representation which is accurate there but deviates where g_0 is small. When the extent of g_0 is made larger by increasing $R_{\rm sm}$, the resulting expansion is valid over a larger part of space but is less accurate overall for a fixed level of truncation. This is in contrast to a Taylor series, which places all emphasis on the (comparatively irrelevant) point at the center of expansion. Figure 3 illustrates the dependence of the convergence on the smoothing radius.

In order to formalize the discussion of convergence, we have chosen the normalization of the functions P_{kL} (12.6)/(12.7) such that

$$|p_{k\ell}(\gamma;r)|\leqslant 1 \quad \text{for all} \quad r\leqslant R_{\rm sm}$$

and

$$|p_{0\ell}(\gamma;R_{\rm sm})|=1. \tag{12.19}$$

Typical radial functions $p_{k\ell}(r)$ are shown in Fig. 4. The characteristic behavior is that the polynomials oscillate in a well-behaved manner up to a range of approximately $R_{\rm sm}$ and somewhat beyond, then quickly increase to large values. If we truncate (12.15) in order to approximate $f_L(r)r^\ell$ by the finite sum $\Sigma_{k\leqslant k_{\max}}a_{kL}(\gamma)p_{k\ell}(\gamma;r)$, then an upper bound for the approximation error on $r\leqslant R_{\rm sm}$ is given by

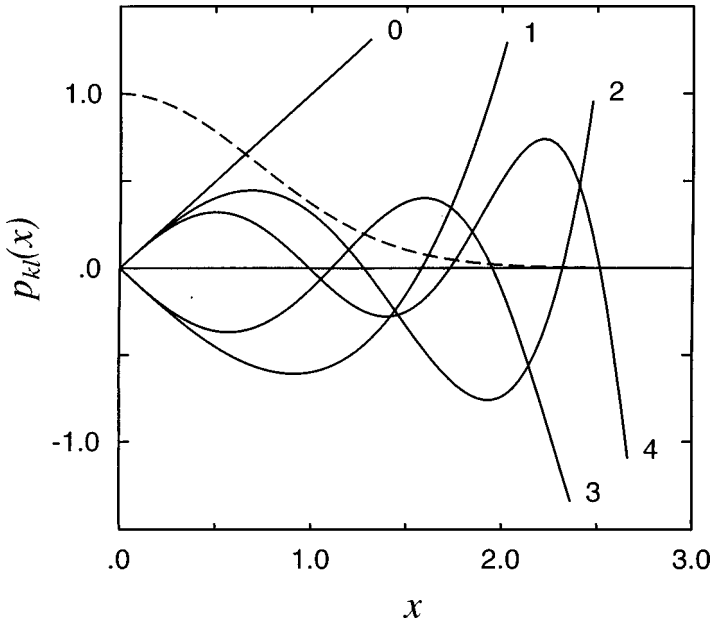

FIG. 4. Radial polynomials $p_{k\ell}(r)$ for $\ell=1$ and $R_{\rm sm}=1/a=1.0$, plotted for $0\leqslant k\leqslant 4$. The Gaussian $g_0(r)=\exp(-r^2/R_{\rm sm}^2)$ is also shown (dashed line). Note that the polynomials are well-behaved where g_0 is large, then increase rapidly outside this range.

$$\left| f_L(r) r^{\ell} - \sum_{k\leqslant k_{\max}} a_{kL}(\gamma) p_{k\ell}(\gamma;r) \right| = \left| \sum_{k>k_{\max}} a_{kL}(\gamma) p_{k\ell}(\gamma;r) \right| \leqslant \sum_{k>k_{\max}} |a_{kL}(\gamma)|. \quad (12.20)$$

In close analogy to the expansion of a function over a finite interval using Chebyshev poynomials,[18] the absolute values of the expansion coefficients generally decrease rapidly by orders of magnitude. The approximation error on $r\leqslant R_{\rm sm}$ is then bounded by the coefficient of the first neglected term, i.e.,

$$\left| f_L(r) r^{\ell} - \sum_{k\leqslant k_{\max}} a_{kL}(\gamma) p_{k\ell}(\gamma;r) \right| \leqslant |a_{k_{\max}+1,L}(\gamma)|. \quad (12.21)$$

XIII. SUMMARY

In this paper, we have presented a set of functions which combine many of the best features of Hankel functions and Gaussians as a basis for electronic structure calculations. These ''smoothed Hankel functions'' equal the standard Hankel functions at large distances, showing the same exponential decay. At smaller radii, the functions bend over smoothly until they approach the origin as $r^{\ell} Y_L(\hat{\mathbf{r}})$. The resulting object is nonsingular and analytical in all parts of space. The shape of each function is characterized by two separate parameters. The first is an energy which defines the decay at large distances, exactly as in the case of the standard Hankel functions. The second parameter is a smoothing radius which determines how strongly the singularity has been smoothed out, i.e., at what distance the smoothed Hankel function starts to deviate significantly from the standard function. A major advantage is that by adjusting these two parameters, basis functions can be made which closely resemble the true wave function outside the core regime. This is an improvement over both Gaussians and standard Hankel functions. A second advantage is that the resulting basis set is reasonably well suited to a direct numerical evaluation of the unavoidable three-center integrals, since a coarse mesh can be used. Furthermore, Bloch sums of the functions can be handled in a very similar way, leading to a unified description for crystals and molecules. In Ref. 5, a successful implementation of a density-functional crystal program based on these features was presented.

To use a given set of functions as a basis set, a number of calculational steps must be performed. To provide the neccessary formalism for the smoothed Hankel functions, we have first

introduced some families of auxiliary functions. These as well as the smoothed Hankel functions themselves are generated in compact form from simple root functions by differential operators. Using these results, we have derived analytical expressions for overlap and kinetic-energy two-center matrix elements as well as Coulomb integrals involving two such functions. To evaluate the expressions in practice, suitable recurrence relations were found. Finally, we have presented a general procedure by which an arbitrary smooth function can be expanded as a sum of polynomials around some point in space, whereby the approximation converges evenly over a chosen range as higher terms are included. Such a local representation is a necessary step when the functions are used in the context of augmentation methods or norm-conserving pseudopotentials. In sum, we believe the functions presented here have a high potential as a local basis set in electronic structure calculations and have derived expressions to perform the various calculational steps needed for such an application.

ACKNOWLEDGMENTS

We thank the Fonds der Chemischen Industrie for financial support.

APPENDIX A: DERIVATION OF AN EXPLICIT INTEGRAL REPRESENTATION FOR THE $\chi_\ell(r)$

Here, we make the connection to an explicit integral representation for the radial functions $\chi_\ell(r)$, related to the smoothed Hankel functions by $H_L(\mathbf{r})=\chi_\ell(r)r^\ell Y_L(\hat{\mathbf{r}})$ [see Eq. (6.18)]. Such representations are used in standard formulations of Ewald summation[19] and in Ref. 20. This representation is mainly of theoretical interest; for the calculation of the χ_ℓ it is more efficient to use the recurrence relations (6.19)–(6.21).

For $\chi_0(r)$ we obtain from (6.19) and (6.6) the representation

$$\chi_0(r)=\frac{1}{2r}\left[e^{-\kappa r}\frac{2}{\sqrt{\pi}}\int_{(\kappa/2a)-ar}^{\infty}e^{-t^2}dt-e^{\kappa r}\frac{2}{\sqrt{\pi}}\int_{(\kappa/2a)+ar}^{\infty}e^{-t^2}dt\right]. \tag{A1}$$

By making the substitutions

$$\xi=\frac{-t+\sqrt{t^2+2\kappa r}}{2r},\quad t=\frac{\kappa}{2\xi}-r\xi,\quad \frac{dt}{d\xi}=-\frac{\kappa}{2\xi^2}-r, \tag{A2}$$

and

$$\xi=\frac{t-\sqrt{t^2-2\kappa r}}{2r},\quad t=\frac{\kappa}{2\xi}+r\xi,\quad \frac{dt}{d\xi}=-\frac{\kappa}{2\xi^2}+r \tag{A3}$$

in the first and second integral, respectively, we obtain

$$\begin{aligned}\chi_0(r)&=\frac{1}{2r}\frac{2}{\sqrt{\pi}}\left[\int_a^0 e^{-\kappa^2/4\xi^2}e^{-r^2\xi^2}\left(-\frac{\kappa}{2\xi^2}-r\right)d\xi-\int_a^0 e^{-\kappa^2/4\xi^2}e^{-r^2\xi^2}\left(-\frac{\kappa}{2\xi^2}+r\right)d\xi\right]\\&=\frac{2}{\sqrt{\pi}}\int_0^a e^{\varepsilon/4\xi^2}e^{-r^2\xi^2}d\xi.\end{aligned} \tag{A4}$$

Finally, by applying (4.10) and (4.11) we see that

$$\chi_\ell(r)=\frac{2}{\sqrt{\pi}}2^\ell\int_0^a \xi^{2\ell}e^{\varepsilon/4\xi^2}e^{-r^2\xi^2}d\xi\quad\text{for all }\ell\in\mathcal{N}_0. \tag{A5}$$

APPENDIX B: EVALUATION OF BLOCH SUMS OF ORDINARY (UNSMOOTHED) HANKEL FUNCTIONS

The standard Ewald method[19] for evaluating lattice sums of Hankel functions is equivalent to splitting off a smoothed-Hankel part which is summed in reciprocal space. Here, we show how to do this in our formalism, whereby the inconvenient integrals to represent $\chi_\ell(r)$ are avoided (see Appendix A).

The aim is the evaluation of the Bloch sum

$$\check{H}_L^b(\varepsilon;\mathbf{r}) := \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \check{H}_L(\varepsilon;\mathbf{r}-\mathbf{R}) \tag{B1}$$

of unsmoothed Hankel functions at some point $\mathbf{r}$. This point cannot be a direct lattice point $\mathbf{R}'$ since $\check{H}_L^b(\mathbf{r})$ is undefined there. For a suitable parameter $\gamma>0$, we split (B1) up into two parts, which are evaluated as sums over reciprocal and direct space lattice vectors, respectively:

$$\begin{aligned}\check{H}_L^b(\varepsilon;\mathbf{r}) &= H_L^b(\gamma,\varepsilon;\mathbf{r}) + (\check{H}_L^b(\varepsilon;\mathbf{r}) - H_L^b(\gamma,\varepsilon;\mathbf{r})) \\ &= \frac{1}{\Omega}\sum_{\mathbf{G}} \frac{-4\pi}{\varepsilon-|\mathbf{k}+\mathbf{G}|^2}\, \mathscr{Y}_L(-i(\mathbf{k}+\mathbf{G}))\, e^{\gamma(\varepsilon-|\mathbf{k}+\mathbf{G}|^2)} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \\ &\quad + \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} (\check{H}_L(\varepsilon;\mathbf{r}-\mathbf{R}) - H_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R}))\end{aligned} \tag{B2}$$

[cf. (3.18), (6.35)]. In the real-space sum, the values $H_L(\gamma,\varepsilon;\mathbf{r}-\mathbf{R})$ can be calculated from (6.18) to (6.24). Furthermore, the values $\check{H}_L(\varepsilon;\mathbf{r}-\mathbf{R})$ can be obtained as follows in view of (6.27) and (6.18)–(6.24):

$$\check{H}_L(\mathbf{r}) = \check{\chi}_\ell(r) r^\ell Y_L(\hat{\mathbf{r}}), \tag{B3}$$

$$\check{\chi}_0(r) = \frac{e^{-\kappa r}}{r}, \tag{B4}$$

$$\check{\chi}_1(r) = \frac{e^{-\kappa r}}{r^3} + \frac{\kappa e^{-\kappa r}}{r^2}, \tag{B5}$$

$$\check{\chi}_{\ell+1}(r) = \frac{2\ell+1}{r^2}\,\check{\chi}_\ell(r) - \frac{\varepsilon}{r^2}\,\check{\chi}_{\ell-1}(r) \quad \text{for } \ell \geqslant 1. \tag{B6}$$

[1] R. O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).
[2] G. B. Bachelet, D. R. Haman, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).
[3] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
[4] O. K. Andersen, Phys. Rev. B **12**, 3060 (1975).
[5] M. Methfessel and M. van Schilfgaarde (unpublished); M. Methfessel, *NFP Manual* [Institute for Semiconductor Physics, Frankfurt (Oder), 1997].
[6] J. Spanier and K. B. Oldham, *An Atlas of Functions* (Springer, Berlin, 1987).
[7] Z. Szmydt, *Fourier Transformation and Linear Differential Equations* (PNW, Warszawa, 1977).
[8] E. Bott, Diplomarbeit, Technische Universität Darmstadt, 1997.
[9] M. Abramowitz and I. A. Stegun, *Pocketbook of Mathematical Functions* (Harri Deutsch, Thun, 1984).
[10] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt-Saunders, Philadelphia, 1976).
[11] M. Methfessel, Ph.D. thesis, Katholieke Universiteit de Nijmegen, 1986, pp. 22, 118–125.
[12] P. C. Schmidt and K. G. Weil, *Atom- und Molekülbau* (Thieme, Stuttgart, 1982).
[13] R. Daudel, G. Leroy, D. Peeters, and M. Sana, *Quantum Chemistry* (Wiley, Chichester, 1983), p. 160.
[14] M. Springborg and O. K. Andersen, J. Chem. Phys. **87**, 7125 (1987).
[15] I. N. Bronstein and K. A. Semendjajew, *Taschenbuch der Mathematik* (Harri Deutsch, Thun, 1985), 22. Aufl.
[16] H. Werner and R. Schaback, *Praktische Mathematik II Methoden der Analysis* (Springer, Berlin, 1972).
[17] R. J. Nozawa, J. Math. Phys. **7**, 1841 (1966).
[18] W. H. Press *et al.*, *Numerical Recipes in FORTRAN; the Art of Scientific Computing*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[19] F. S. Ham and B. Segall, Phys. Rev. **124**, 1786 (1961).
[20] S. Y. Savrasov, Phys. Rev. B **54**, 16470 (1996).

Differential calculus on the *h*-superplane

Salih Çelik
Department of Mathematics, Mimar Sinan University, 80690 Besiktas, Istanbul, Turkey

Sultan A. Çelik
Department of Mathematics, Yildiz Technical University, Sisli, Istanbul, Turkey and
Department of Mathematics, Mimar Sinan University, 80690 Besiktas, Istanbul, Turkey

Metin Arik
Department of Physics, Bogazici University, Bebek, Istanbul, Turkey

(Received 27 January 1997; accepted for publication 9 February 1998)

A noncommutative differential calculus on the h-superplane is presented via a contraction of the q-superplane. An R-matrix which satisfies both ungraded and graded Yang–Baxter equations is obtained and a new deformation of the $(1+1)$-dimensional classical phase space (the super-Heisenberg algebra) is introduced. © *1998 American Institute of Physics.* [S0022-2488(98)02305-6]

I. INTRODUCTION

A possible approach to quantum groups is obtained by deforming the coordinates of a linear space to be noncommuting objects.[1] In this scheme the quantum group structure appears if one considers linear transformations which preserve the algebraic properties of the algebra of coordinates. A natural and physically interesting question is whether one can define differentials and derivatives corresponding to these noncommuting variables. A general answer to this question was given by Connes,[2] who considered the differential algebra for non-commutative algebras. Differential geometry of Lie groups and supergroups plays an important role in the mathematical modeling of physical theories. Since a (graded) Hopf algebra or quantum (super) group can be regarded as a generalization of the notion of a (super) group, it is tempting to also generalize the corresponding notions of differential geometry.[3]

Recently Wess and Zumino[4] have shown that a consistent quantum deformation of the differential calculus is satisfied by an R-matrix which can be any solution of the quantum Yang–Baxter equation. The noncommutative plane has been studied in the course of the last few years[5] in that the geometry underlying the above-mentioned plane has a deep connection with the Yang–Baxter equation, which is important in two-dimensional exactly soluble statistical models. The quantum (super) plane was generalized to the supersymmetric quantum (super) plane[6,7] and it was shown that the quantum superplane is related to the graded Yang–Baxter equation.

The h-deformation of the supergroup $\mathrm{GL}(1|1)$ is given by Dabrowski and Parashar,[8] via a contraction of $\mathrm{GL}_q(1|1)$. They have also introduced a differential calculus on the h-superplane which is related to $\mathrm{GL}_h(1|1)$ via the Wess–Zumino formulas.

This paper considers an alternative approach where instead of adopting the Wess–Zumino calculus from the start, the R-matrix is obtained using the consistency conditions. This leads to a consistent exterior derivative. We start by introducing a noncommutative differential calculus on the h-superplane via a contraction of the q-superplane. We define derivatives and differentials on the h-superplane of noncommuting coordinates and give their commutation rules. We note the role of the graded Yang–Baxter equation and the connection to the quantum supergroup. We give a new deformation of the $(1+1)$-dimensional classical phase space. Finally we show that the q-deformed super-oscillator algebra satisfies the undeformed (classical) super-oscillator algebra when objects are transformed into new objects such that they are singular for certain values of the deformation parameter.

II. A DIFFERENTIAL CALCULUS ON *h*-SUPERPLANE

In this work we denote q-deformed objects by primed quantities. Unprimed quantities represent transformed coordinates. As usual, we assume that even (bosonic) objects commute with everything and odd (Grassmann) objects anticommute among themselves.

 © 1998 American Institute of Physics

A. Quantum *h*-superplane

We begin by considering the quantum superplane which is defined by Manin.[9] The commutation relation between the even coordinate x' and the odd (Grassmann) coordinate θ' of the quantum superplane is in the form

$$x'\theta' - q\theta' x' = 0, \tag{1}$$

where q is a complex deformation parameter.

We now introduce new coordinates x and θ, in terms of x' and θ' as

$$x = x', \quad \theta = \theta' - \frac{h}{q-1}x', \tag{2}$$

as in Ref. 10. This transformation is singular in the $q \to 1$ limit. Using relation (1), it is easy to verify that

$$x\theta = q\theta x + hx^2, \tag{3}$$

where the new deformation parameter h commutes with the coordinate x. Also, since the Grassmann coordinate θ' satisfies

$$\theta'^2 = 0$$

one obtains

$$\theta^2 = -h\theta x, \tag{4}$$

where h anti-commutes with θ and

$$h^2 = 0, \tag{5}$$

that is, the new deformation parameter h is a Grassmann number.[8] Taking the $q \to 1$ limit we obtain the following relations which define the h-superplane:

$$x\theta = \theta x + hx^2, \quad \theta^2 = -h\theta x. \tag{6}$$

B. Relations of coordinates and differentials

To establish a noncommutative differential calculus on the quantum h-superplane, we assume that the commutation relations between the coordinates and their differentials are in the following form:

$$\begin{aligned} x'\mathrm{d}x' &= A\mathrm{d}x'x' \\ x'\mathrm{d}\theta' &= F_{11}\mathrm{d}\theta' x' + F_{12}\mathrm{d}x'\theta', \\ \theta'\mathrm{d}x' &= F_{21}\mathrm{d}x'\theta' + F_{22}\mathrm{d}\theta' x', \\ \theta'\mathrm{d}\theta' &= B\mathrm{d}\theta'\theta'. \end{aligned} \tag{7}$$

We demand an exterior differential d obeying the condition:

$$\mathrm{d}^2 = 0, \tag{8}$$

and the graded Leibniz rule

$$\mathrm{d}(fg) = (\mathrm{d}f)g + (-1)^{\hat{f}} f(\mathrm{d}g), \tag{9}$$

where $\hat{f}$ is the Grassmann degree of f (recall that d should be odd), that is, $\hat{f}=0$ for even variables and $\hat{f}=1$ for odd variables. Considering the differential of a function and differentiating (2) we have

$$\mathrm{d}x=\mathrm{d}x', \quad \mathrm{d}\theta=\mathrm{d}\theta'+\frac{h}{q-1}\mathrm{d}x'. \tag{10}$$

Substituting (2) and (10) into (7) one has

$$\begin{aligned} x\mathrm{d}x &= A\mathrm{d}xx, \\ x\mathrm{d}\theta &= F_{11}\mathrm{d}\theta x+F_{12}\mathrm{d}x\theta+\frac{h}{q-1}(A-F_{11}-F_{12})\mathrm{d}xx, \\ \theta\mathrm{d}x &= F_{21}\mathrm{d}x\theta+F_{22}\mathrm{d}\theta x-\frac{h}{q-1}(A+F_{21}+F_{22})\mathrm{d}xx, \\ \theta\mathrm{d}\theta &= B\mathrm{d}\theta\theta-\frac{h}{q-1}[(B+F_{12}+F_{21})\mathrm{d}x\theta-(B-F_{11}-F_{22})\mathrm{d}\theta x]. \end{aligned} \tag{11}$$

These relations are slightly different from the results of Dabrowski and Parashar.[8] The reason for this difference is that in Ref. 8, the commutation relations among the matrix elements of a matrix belonging to $\mathrm{GL}_h(1|1)$ were obtained via the use of commutation relations of the dual *exterior* superplane. On the other hand we use the commutation relations of the dual superplane as in Ref. 3. The commutation relations among the matrix elements are the same in both approaches. However the commutation relations involving the differentials turn out not to be the same.

We know that the quantum supermatrices of the quantum $\mathrm{GL}_q(1|1)$ supergroup can be defined as linear transformations of the variables x' and θ', which preserve the commutation relation (1) and their duals,[9] that is, the quantum supergroup $\mathrm{GL}_q(1|1)$ acts as a linear transformation on the quantum superplane, which preserves (1) and the relations

$$\varphi'^2=0, \quad \varphi' y'-q^{-1}y'\varphi'=0. \tag{12}$$

In extending this property of covariance under the coaction of $\mathrm{GL}_q(1|1)$, from the superplane to its calculus, it will be assumed that the deformed group structure implies and is implied by invariance of the intermediary relations (7) under linear transformations of the quantum superplane. In the present work, this will be applied to the h-deformed superplane.

The coefficients A, B and F_{ij} given in (11) can be related to q by the consistency of calculus. Thus we apply the exterior derivative d to the relation (3). From the consistency condition

$$\mathrm{d}(x\theta-q\theta x-hx^2)=0 \tag{13}$$

we find

$$F_{11}=q(1-F_{22}), \quad F_{12}=-(1+qF_{21}). \tag{14a}$$

Applying the exterior derivative d on the second and third relations of (11) and using the definitions (8), (9) one finds

$$F_{12}=qF_{11}-1, \quad F_{21}=q(F_{22}-1). \tag{14b}$$

Similarly from the last relation of (11) we get

$$F_{12}+F_{21}=q(F_{11}+F_{22})-(1+q)B. \tag{14c}$$

On the other hand, using relation (4) we obtain

$$\theta\mathrm{d}\theta=\mathrm{d}\theta\theta+hF_{21}\mathrm{d}x\theta-h(1-F_{22})\mathrm{d}\theta x \tag{15}$$

$$= \mathrm{d}\theta\theta - \frac{h}{q}[F_{11}\mathrm{d}\theta x + (1+F_{12})\mathrm{d}x\theta].$$

Thus we have

$$\begin{gathered} B=1,\ \ F_{12}+F_{21}+1=(1-q)F_{21}, \\ 1-F_{11}-F_{22}=(1-q)(1-F_{22}). \end{gathered} \tag{14d}$$

To find the commutation relation between differentials, say $\mathrm{d}x$ and $\mathrm{d}\theta$, we apply the exterior derivative d on the first two relations of (11) and use the nilpotency of d [Eq. (8)]. Then it is easy to see that

$$\mathrm{d}x\mathrm{d}\theta = \frac{1}{q}\mathrm{d}\theta\mathrm{d}x, \quad (\mathrm{d}x)^2 = 0. \tag{16}$$

Solving the system (14) we now get

$$\begin{gathered} A \ \text{ undetermined}, \ \ F_{11}=q, \\ F_{12}=q^2-1, \ \ F_{21}=-q, \ \ F_{22}=0. \end{gathered} \tag{17}$$

We choose A equal to q^2, since this leads to the standard R-matrix [Eq. (38)] in the $h\to 0$ limit. We are thus led to the following deformed relations containing q and h

$$\begin{gathered} x\mathrm{d}x = q^2\mathrm{d}xx, \\ x\mathrm{d}\theta = q\mathrm{d}\theta x - h\mathrm{d}xx + (q^2-1)\mathrm{d}x\theta, \\ \theta\mathrm{d}x = -q\mathrm{d}x\theta - qh\mathrm{d}xx, \\ \theta\mathrm{d}\theta = \mathrm{d}\theta\theta - h(q\mathrm{d}x\theta + \mathrm{d}\theta x). \end{gathered} \tag{18}$$

Note that although in the $q\to 1$ limit the transformations (2) and (10) are ill behaved, the resulting commutation relations are well defined. We shall not use the limit process, yet.

C. Relations of derivatives and variables

First, we wish to construct the curl of any one-form $w(x',\theta')$. To this end let us denote the partial derivatives with respect to x' and θ' as

$$\partial_{x'} = \frac{\partial}{\partial x'}, \quad \partial_{\theta'} = \frac{\partial}{\partial \theta'}, \tag{19}$$

respectively, and introduce the super-gradient operator in vector notation

$$\nabla = (\partial_{x'}, \partial_{\theta'}). \tag{20}$$

We define the vectors $X'=(x',\theta')$ and $\mathrm{d}X'=(\mathrm{d}x',\mathrm{d}\theta')$. Then we can write the differential d as

$$\mathrm{d} = \mathrm{d}x'\partial_{x'} + \mathrm{d}\theta'\partial_{\theta'} = \mathrm{d}X'.\nabla, \tag{21}$$

where the dot denotes the inner product. If we write $w(x',\theta')$ in the basis $\mathrm{d}X'$ as

$$w(x',\theta') = \mathrm{d}x' w_1(x',\theta') + \mathrm{d}\theta' w_2(x',\theta') \tag{22}$$

where w_1 and w_2 are smooth functions of the variables then we get

$$\mathrm{d}w(x',\theta') = \mathrm{d}x'\mathrm{d}\theta'[q\partial_{x'}w_2(x',\theta') - \partial_{\theta'}w_1(x',\theta')]. \tag{23}$$

Thus the curl of the one-form $w(x',\theta')=(w_1(x',\theta'),w_2(x',\theta'))$ is given by

$$\nabla\times w(x',\theta')=q\partial_{x'}w_2(x',\theta')-\partial_{\theta'}w_1(x',\theta'). \tag{24}$$

Now we can find the commutation rules of the derivatives, once we obtain the derivatives ∂_x and ∂_θ. For this if we demand the chain rule on the expressions (2) we find

$$\partial_x=\partial_{x'}+\frac{h}{q-1}\partial_{\theta'},\quad \partial_\theta=\partial_{\theta'}. \tag{25}$$

It is easy to see that in the case of (25) the differential d given in (21) preserves its form

$$\mathrm{d}=\mathrm{d}x\partial_x+\mathrm{d}\theta\partial_\theta. \tag{26}$$

If we now put $w(x,\theta)=\mathrm{d}f(x,\theta)$ then we conclude that ∂_x and ∂_θ generate a non-commutative algebra with the commutation relations

$$\partial_\theta\partial_x=q\partial_x\partial_\theta,\quad \partial_\theta^2=0. \tag{27}$$

From (26)

$$\mathrm{d}f(x,\theta)=\mathrm{d}x\partial_x f+\mathrm{d}\theta\partial_\theta f, \tag{28}$$

so that replacing f with xf and θf we arrive at the following commutation relations between derivatives and variables:

$$\begin{aligned}
\partial_x x&=1+Ax\partial_x+F_{12}\theta\partial_\theta-\frac{h}{q-1}(A-F_{11}-F_{12})x\partial_\theta,\\
\partial_x\theta&=-F_{21}\theta\partial_x-\frac{h}{q-1}[(A+F_{21}+F_{22})x\partial_x+(1+F_{12}+F_{21})\theta\partial_\theta],\\
\partial_\theta x&=F_{11}x\partial_\theta,\\
\partial_\theta\theta&=1-\theta\partial_\theta-F_{22}x\partial_x-\frac{h}{q-1}(1-F_{11}-F_{22})x\partial_\theta.
\end{aligned} \tag{29}$$

These commutation relations are well defined in the limit $q\to 1$. This can be checked using (17).

The covariant differential structure which is obtained so far is a concrete example of non-commutative differential geometry. The complete framework of the differential calculus requires commutation relations of the exterior differentials with derivatives.

D. Relations of differentials with derivatives

Finally we shall find the commutation relations between differentials and derivatives. We assume that they have the following form in terms of primed quantities

$$\begin{aligned}
\partial_{x'}\mathrm{d}x'&=A_{11}\mathrm{d}x'\partial_{x'}+A_{12}\mathrm{d}\theta'\partial_{\theta'},\\
\partial_{x'}\mathrm{d}\theta'&=A_{21}\mathrm{d}\theta'\partial_{x'}+A_{22}\mathrm{d}x'\partial_{\theta'},\\
\partial_{\theta'}\mathrm{d}x'&=B_{11}\mathrm{d}x'\partial_{\theta'}+B_{12}\mathrm{d}\theta'\partial_{x'},\\
\partial_{\theta'}\mathrm{d}\theta'&=B_{21}\mathrm{d}\theta'\partial_{\theta'}+B_{22}\mathrm{d}x'\partial_{x'}.
\end{aligned} \tag{30}$$

Using (25) and (10), these commutation rules can be written as

$$\begin{aligned}
\partial_x \mathrm{d}x &= A_{11}\mathrm{d}x\partial_x + \frac{h}{q-1}(A_{11}-A_{12}+B_{11})\mathrm{d}x\partial_\theta + A_{12}\mathrm{d}\theta\partial_\theta + \frac{h}{q-1}B_{12}\mathrm{d}\theta\partial_x,\\
\partial_x \mathrm{d}\theta &= A_{21}\mathrm{d}\theta\partial_x + \frac{h}{q-1}(A_{11}-A_{21}+B_{22})\mathrm{d}x\partial_x + \frac{h}{q-1}(A_{12}-A_{21}+B_{12})\mathrm{d}\theta\partial_\theta,\\
\partial_\theta \mathrm{d}\theta &= B_{21}\mathrm{d}\theta\partial_\theta + \frac{h}{q-1}(B_{22}-B_{21}-B_{11})\mathrm{d}x\partial_\theta + B_{22}\mathrm{d}x\partial_x - \frac{h}{q-1}B_{12}\mathrm{d}\theta\partial_x,\\
\partial_\theta \mathrm{d}x &= B_{11}\mathrm{d}x\partial_\theta + B_{12}\mathrm{d}\theta\partial_x - \frac{h}{q-1}B_{12}(\mathrm{d}\theta\partial_\theta + \mathrm{d}x\partial_x).
\end{aligned} \tag{31}$$

Finding the coefficients A_{ij} and B_{ij} from these equations we get

$$\begin{aligned}
\partial_x \mathrm{d}x &= q^2\mathrm{d}x\partial_x - h\mathrm{d}x\partial_\theta + (q^2-1)\mathrm{d}\theta\partial_\theta,\\
\partial_x \mathrm{d}\theta &= q\mathrm{d}\theta\partial_x + qh(\mathrm{d}x\partial_x + \mathrm{d}\theta\partial_\theta),\\
\partial_\theta \mathrm{d}x &= -q\mathrm{d}x\partial_\theta,\\
\partial_\theta \mathrm{d}\theta &= \mathrm{d}\theta\partial_\theta + h\mathrm{d}x\partial_\theta.
\end{aligned} \tag{32}$$

Here, in order to obtain these relations we used that the exterior differential d (anti-) commutes with the differentials, that is,

$$\mathrm{d}(\mathrm{d}x) = -(\mathrm{d}x)\mathrm{d}, \quad \mathrm{d}(\mathrm{d}\theta) = (\mathrm{d}\theta)\mathrm{d} \tag{33}$$

and the relation

$$\partial_i(X^j \mathrm{d}X^k) = \delta^i{}_j \delta^k{}_l \mathrm{d}X^k, \tag{34}$$

where $\partial_1 = \partial_x$, $\partial_2 = \partial_\theta$, $X^1 = x$, and $X^2 = \theta$. Again, the relations (32) are well defined in the limit $q \to 1$.

E. The R-matrix formalism

We now shall obtain the R-matrix satisfying the graded Yang–Baxter equations

$$R_{12}R_{13}R_{23} = R_{23}R_{13}R_{12}, \tag{35}$$

$$\hat{R}_{12}\hat{R}_{23}\hat{R}_{12} = \hat{R}_{23}\hat{R}_{12}\hat{R}_{23}. \tag{36}$$

For this we define the commutation relations between variables and their differentials [see Eq. (18)] in the following form

$$X^i \mathrm{d}X^j = q(-1)^{\hat{i}(\hat{j}+1)} K^{ji}{}_{kl} \mathrm{d}X^k X^l \tag{37}$$

where $K \in End(\mathscr{C} \otimes \mathscr{C})$. Comparing (37) with (18) we have

$$K_{h,q} = \begin{pmatrix} q & 0 & 0 & 0 \\ h & 1 & 0 & 0 \\ -q^{-1}h & q-q^{-1} & 1 & 0 \\ 0 & -h & -q^{-1}h & q^{-1} \end{pmatrix} = (K^{ij}{}_{kl}). \tag{38}$$

If we define

$$K_h = \lim_{q \to 1} K_{h,q}, \tag{39}$$

$$\hat{K}_h = \lim_{q\to 1}(K_{h,q}P), \tag{40}$$

where P is the super permutation matrix, that is,

$$P^{ij}{}_{kl} = (-1)^{\hat{i}\hat{j}}\delta^i{}_l\delta^j{}_k,$$

we have

$$K_h = \begin{pmatrix} 1 & 0 & 0 & 0 \\ h & 1 & 0 & 0 \\ -h & 0 & 1 & 0 \\ 0 & -h & -h & 1 \end{pmatrix}, \quad \hat{K}_h = \begin{pmatrix} 1 & 0 & 0 & 0 \\ h & 0 & 1 & 0 \\ -h & 1 & 0 & 0 \\ 0 & -h & -h & -1 \end{pmatrix}. \tag{41}$$

Here the matrix $\hat{K}_h$ coincides with the $\hat{R}_h$ matrix of Ref. 8. We know from Ref. 8 that the matrix $\hat{K}_h$ satisfies Equation (36) with the grading

$$(\hat{K}_{12})^{abc}{}_{def} = \hat{K}^{ab}{}_{de}\delta^c{}_f,$$

$$(\hat{K}_{13})^{abc}{}_{def} = (-1)^{b(c+f)}\hat{K}^{ac}{}_{df}\delta^b{}_e, \tag{42}$$

$$(\hat{K}_{23})^{abc}{}_{def} = (-1)^{a(b+c+e+f)}\hat{K}^{bc}{}_{ef}\delta^a{}_d.$$

Also, the R-matrix

$$R_h = P\hat{K}_h \tag{43}$$

obeys both the ungraded and the graded Yang–Baxter equations with the grading again given by (42). This is due to the odd character of h. As a consequence,

$$K_h = R_h^{-1} \tag{44}$$

so that K_h has the same properties as R_h.

We note that the equation

$$\hat{K}_h T_1 T_2 = T_1 T_2 \hat{K}_h \tag{45}$$

is satisfied for the h-deformed supergroup GL(1|1). Here T is a supermatrix in $\mathrm{GL}_h(1|1)$ and $T_1 = T\otimes I$, $T_2 = I\otimes T$. It is assumed that the tensor product is graded, that is,

$$(T_1)^{ij}{}_{kl} = T^i{}_k\delta^j{}_l, \quad (T_2)^{ij}{}_{kl} = (-1)^{\hat{i}(\hat{j}+\hat{l})}T^j{}_l\delta^i{}_k.$$

It is well-known that a supermatrix has the form

$$T = \begin{pmatrix} a & \beta \\ \gamma & d \end{pmatrix}$$

with two even (Latin letters) and two Grassmann (Greek letters) matrix elements. Equation (45) explicitly reads

$$a\beta=\beta a,\quad a\gamma=\gamma a+h(a^2+\gamma\beta-ad),$$

$$d\beta=\beta d,\quad d\gamma=\gamma d-h(d^2-\gamma\beta-da),$$

$$\beta^2=0,\quad \gamma^2=h\gamma(d-a), \tag{46}$$

$$\beta\gamma=-\gamma\beta+h\beta(d-a),$$

$$ad=da+h\beta(a-d).$$

Using the K_h matrix, we now formulate the differential calculus on the h-superplane. The commutation relations between variables and their differentials are

$$X^i\mathrm{d}X^j=(-1)^{\hat{i}(\hat{j}+1)}K^{ji}{}_{kl}\mathrm{d}X^kX^l. \tag{47}$$

The commutation relations between variables and derivatives are

$$\partial_jX^i=\delta^i{}_j+(-1)^{\hat{i}\hat{j}}K^{ik}{}_{lj}X^l\partial_k, \tag{48}$$

and the relations between differentials and derivatives are

$$\partial_j\mathrm{d}X^i=(-1)^{\hat{i}(\hat{j}+1)}(K^{-1})^{ik}{}_{lj}\,\mathrm{d}X^l\partial_k. \tag{49}$$

Note that the commutation relations between variables can be expressed using the $\hat{K}$ matrix as

$$X^iX^j=\hat{K}^{ij}{}_{kl}X^kX^l, \tag{50}$$

and the relations between derivatives as

$$\partial_i\partial_j=\hat{K}^{kl}{}_{ji}\partial_l\partial_k. \tag{51}$$

F. The commutation relations

We would like to discuss the meaning of covariance in a graded version of non-commutative differential calculus of Wess–Zumino.[4] Before proceeding, we define the dual quantum h-superplane. For this we interpret the differentials $\mathrm{d}x$ and $\mathrm{d}\theta$, as the coordinates of the dual superplane as follows [see Eq. (12)]

$$\mathrm{d}x=\varphi,\quad \mathrm{d}\theta=y. \tag{52}$$

We now formulate the differential calculus on the h-superplane as follows.

The commutation relations of variables and their differentials are

$$x\theta=\theta x+hx^2,\quad \theta^2=-h\theta x,$$

$$\varphi y=y\varphi,\quad \varphi^2=0. \tag{53a}$$

Note that if we assume that the relations (53a) have to be covariant under the coaction

$$\delta(x)=a\otimes x+\beta\otimes\theta,\quad \delta(\theta)=\gamma\otimes x+d\otimes\theta,$$

and that β, γ anticommute with θ, φ and h we get anew the relations (46).

The commutation relations among the derivatives are

$$\partial_x\partial_\theta=\partial_\theta\partial_x,\quad \partial_\theta^2=0, \tag{53b}$$

and those between variables and derivatives are

$$\partial_x x = 1 + x\partial_x + hx\partial_\theta, \quad \partial_\theta x = x\partial_\theta,$$
$$\partial_x \theta = \theta\partial_x - h(x\partial_x + \theta\partial_\theta), \quad \partial_\theta \theta = 1 - \theta\partial_\theta + hx\partial_\theta. \tag{53c}$$

The commutation relations of variables with their differentials are

$$x\mathrm{d}x = \mathrm{d}x x, \quad \theta\mathrm{d}\theta = \mathrm{d}\theta\theta - h(\mathrm{d}x\theta + \mathrm{d}\theta x),$$
$$x\mathrm{d}\theta = \mathrm{d}\theta x - h\mathrm{d}x x, \quad \theta\mathrm{d}x = -\mathrm{d}x\theta - h\mathrm{d}x x. \tag{54}$$

The commutation relations between derivatives and differentials are

$$\partial_x \mathrm{d}x = \mathrm{d}x\partial_x - h\mathrm{d}x\partial_\theta, \quad \partial_\theta \mathrm{d}\theta = \mathrm{d}\theta\partial_\theta + h\mathrm{d}x\partial_\theta,$$
$$\partial_x \mathrm{d}\theta = \mathrm{d}\theta\partial_x + h(\mathrm{d}x\partial_x + \mathrm{d}\theta\partial_\theta), \quad \partial_\theta \mathrm{d}x = -\mathrm{d}x\partial_\theta. \tag{55}$$

Note that the simple relations involving the differentials φ and y in (53) are not obtained using the R-matrix $\hat{K}_h$. Instead the $q \to 1$ limit of (16) has been used.

The exterior differential

$$\mathrm{d} = \varphi\partial_x + y\partial_\theta \tag{56}$$

satisfies the usual properties such that

$$\mathrm{d}x - x\mathrm{d} = \varphi, \quad \mathrm{d}\theta + \theta\mathrm{d} = y, \tag{57a}$$

as expected. The relations of the exterior differential d with ∂_x and ∂_θ are

$$\mathrm{d}\partial_x = \partial_x \mathrm{d}, \quad \mathrm{d}\partial_\theta = -\partial_\theta \mathrm{d} \tag{57b}$$

and those with differentials φ and y

$$\mathrm{d}\varphi = -\varphi\mathrm{d}, \quad \mathrm{d}y = y\mathrm{d}. \tag{57c}$$

Thus the basic requirement for the exterior derivative is quite consistent:

$$\begin{aligned} \mathrm{d}^2 &= \mathrm{d}(\varphi\partial_x + y\partial_\theta) \\ &= -\varphi\mathrm{d}\partial_x + y\mathrm{d}\partial_\theta \\ &= -(\varphi\partial_x + y\partial_\theta)\mathrm{d} = -\mathrm{d}^2 \end{aligned}$$

so that d^2 must vanish.

As we understand from the language of Wess–Zumino,[4] covariance here means that all the relations between coordinates x, θ, differentials $\mathrm{d}x$, $\mathrm{d}\theta$ and derivatives ∂_x, ∂_θ, etc., must preserve their form when one changes the coordinates by

$$x \to ax + \beta\theta,$$
$$\theta \to \gamma x + d\theta, \tag{58a}$$

where the matrix $T = (T^i_j)$ is an element of the quantum supergroup acting on the quantum superspace. Therefore we change the differentials by

$$\mathrm{d}x \to a\mathrm{d}x - \beta\mathrm{d}\theta,$$
$$\mathrm{d}\theta \to -\gamma\mathrm{d}x + d\mathrm{d}\theta. \tag{58b}$$

This is consistent since the exterior differential d anticommutes with the Grassmann variables as mentioned before. Covariance can be maintained if one defines the transformation law of the partial derivatives as follows:

$$\partial_x \to (a^{-1} - a^{-1}\gamma d^{-1}\beta a^{-1})\partial_x - a^{-1}\gamma d^{-1}\partial_\theta,$$
$$\partial_\theta \to (d^{-1} - d^{-1}\beta a^{-1}\gamma d^{-1})\partial_\theta + d^{-1}\beta a^{-1}\partial_x. \tag{58c}$$

Hence the differentials transform under the action of supertranspose of supertranspose of T, $(T^{st})^{st}$, whereas the derivatives transform under the inverse of the supertranspose of T, $(T^{st})^{-1}$.

III. A NEW DEFORMATION OF CLASSICAL PHASE SPACE

We now shall obtain a new deformation of the $(1+1)$-dimensional super-Heisenberg algebra (the classical phase space). We denote the algebra (53) generated by coordinates x,θ and the derivatives ∂_x and ∂_θ by $\mathscr{B}_h$. It is interesting to note that simply identifying ∂_x and ∂_θ with ip_x and p_θ is not compatible with the hermiticity of coordinates and momenta. In other words, the algebra $\mathscr{B}_h$ cannot be interpreted as a deformation of the $(1+1)$-dimensional super-Heisenberg algebra. In order to identify ∂_x and ∂_θ with the momenta ip_x and p_θ, one must take care of Hermiticity of coordinates and momenta. To this end, let us define the Hermitian conjugation of the coordinates x and θ, respectively, as

$$x^+ = x, \quad \theta^+ = \theta + 2hx. \tag{59}$$

It is then easy to see that the deformation parameter of the algebra (53) becomes a pure imaginary parameter:

$$\overline{h} = -h, \tag{60}$$

where the bar denotes complex conjugation. In this case, the Hermitian conjugation of the derivatives ∂_x and ∂_θ are

$$\partial_x^+ = -\partial_x + 2h\partial_\theta, \quad \partial_\theta^+ = \partial_\theta. \tag{61}$$

Note that the definitions (59) and (61) for the classical case are obtained in the $h \to 0$ limit.

The relations (53) are now invariant under definitions (59)–(61). The above involution allows us to define the Hermitian operators

$$\hat{x} = x, \quad \hat{\theta} = \theta + hx, \tag{62}$$

and

$$\hat{p}_x = i(\partial_x - h\partial_\theta), \quad \hat{p}_\theta = \partial_\theta. \tag{63}$$

The final form of the h-deformed super-Heisenberg algebra is

$$\hat{x}\hat{\theta} = \hat{\theta}\hat{x} + h\hat{x}^2, \quad \hat{\theta}^2 = -h\hat{\theta}\hat{x},$$
$$\hat{p}_x\hat{p}_\theta = \hat{p}_\theta\hat{p}_x, \quad \hat{p}_\theta^2 = 0,$$
$$\hat{p}_x\hat{x} = \hat{x}\hat{p}_x + i(1 + h\hat{x}\hat{p}_\theta), \quad \hat{p}_\theta\hat{x} = \hat{x}\hat{p}_\theta, \tag{64}$$
$$\hat{p}_x\hat{\theta} = \hat{\theta}\hat{p}_x - h(\hat{x}\hat{p}_x + i\hat{\theta}\hat{p}_\theta), \quad \hat{p}_\theta\hat{\theta} = 1 - \hat{\theta}\hat{p}_\theta + h\hat{x}\hat{p}_\theta.$$

IV. A COMMENT ON SUPER-OSCILLATORS

We know that introducing one ''bosonic'' and one ''fermionic'' oscillator, A and B, respectively, and making the usual identification

$$x' \leftrightarrow A^+, \quad \theta' \leftrightarrow B^+,$$
$$\partial_{x'} \leftrightarrow A, \quad \partial_{\theta'} \leftrightarrow B, \tag{65}$$

one constructs the quantum super-oscillator algebra which is covariant under the quantum supergroup $\mathrm{GL}_q(1|1)$. Under identification (25) and (2) give

$$x \leftrightarrow A^+, \quad \partial_x \leftrightarrow A + \frac{h}{q-1} B,$$
$$\partial_\theta \leftrightarrow B, \quad \theta \leftrightarrow B^+ - \frac{h}{q-1} A^+, \tag{66}$$

where q is a real number. Substituting (66) into (29) and using (17), surprisingly all h-dependence cancels and one obtains the usual q-deformed super-oscillator algebra[11]

$$AA^+ = 1 + q^2 A^+ A + (q^2 - 1) B^+ B,$$
$$BB^+ = 1 - B^+ B, \quad B^2 = 0 = B^{+2}, \tag{67}$$
$$AB^+ = q B^+ A, \quad AB = q^{-1} BA.$$

In the $q \to 1$ limit, we get undeformed super-oscillator algebra.

An interesting problem is the construction of a differential calculus on the (h_1, h_2)-superplane using the methods of this paper. A differential calculus on this superplane using the Wess–Zumino formulas has been given in Ref. 12.

ACKNOWLEDGMENTS

This work was supported in part by T. B. T. A. K. the Turkish Scientific and Technical Research Council. We would like to express our deep gratitude to the referee for critical comments on the manuscript.

[1] V. G. Drinfeld, Proceedings of the IMS, Berkeley, CA, 1986 (unpublished); Yu I. Manin, *Quantum Groups and Non-Commutative Geometry* (CRM, Montreal University Press, 1988).
[2] A. Connes, *Non-Commutative Geometry* (Academic, London, 1994).
[3] B. Schmidke, S. Vokos, and B. Zumino, Z. Phys. C **48**, 249 (1990); E. Corrigan, B. Fairlie, P. Fletcher, and R. Sasaki, J. Math. Phys. **31**, 776 (1990).
[4] J. Wess and B. Zumino, Nucl. Phys. B, Proc. Suppl. **18**, 302 (1990).
[5] F. Mueller-Hoisen, J. Phys. A **25**, 1703 (1992).
[6] S. Soni, J. Phys. A **24**, L459 (1991); **24**, 619 (1991).
[7] W. S. Chung, J. Math. Phys. **35**, 2485 (1994).
[8] L. Dabrowski and P. Parashar, Lett. Math. Phys. **38** (1996).
[9] Yu. I. Manin, Commun. Math. Phys. **123**, 163 (1989).
[10] A. Aghamohammadi, M. Khorrami, and A. Shariati, J. Phys. A **28**, L225 (1995); V. Karimipour, Lett. Math. Phys. **30**, 87 (1994).
[11] M. Chaichian, P. Kulish, and J. Lukierski, Phys. Lett. B **262**, 43 (1991).
[12] S. Çelik, Lett. Math. Phys. **42**, 299 (1997).

Deforming maps for Lie group covariant creation and annihilation operators

Gaetano Fiore[a)]
Sektion Physik der Ludwig-Maximilians-Universität München, Theoretische Physik—Lehrstuhl Professor Wess, Theresienstrasse 37, 80333 München, Germany

(Received 21 April 1997; accepted for publication 1 December 1997)

Any deformation of a Weyl or Clifford algebra $\mathscr{A}$ can be realized through a ''deforming map,'' i.e., a formal change of generators in $\mathscr{A}$. This is true in particular if $\mathscr{A}$ is covariant under a Lie algebra $\mathbf{g}$ and its deformation is induced by some triangular deformation $U_h\mathbf{g}$ of the Hopf algebra $U\mathbf{g}$. We propose a systematic method to construct all the corresponding deforming maps, together with the corresponding realizations of the action of $U_h\mathbf{g}$. The method is then generalized and explicitly applied to the case that $U_h\mathbf{g}$ is the quantum group $\mathscr{U}_h\mathrm{sl}(2)$. A preliminary study of the status of deforming maps at the representation level shows in particular that ''deformed'' Fock representations induced by a compact $U_h\mathbf{g}$ can be interpreted as standard ''undeformed'' Fock representations describing particles with ordinary Bose or Fermi statistics. © *1998 American Institute of Physics.* [S0022-2488(98)02105-7]

I. INTRODUCTION

In recent years the idea of non-cocommutative Hopf algebras[1] (in particular quantum groups[2]) as candidates for generalized symmetry transformations in quantum physics has raised an increasing interest. One way to implement this idea in quantum field theory or condensed matter physics would be to deform the canonical commutation relations (CCR) of some system of mode creators/annihilators, covariant under the action of a Lie algebra $\mathbf{g}$, in such a way that they become covariant under the action of a non-cocommutative deformation $U_h\mathbf{g}$ (with deformation parameter h) of the cocommutative Hopf algebra $U\mathbf{g}$, as it has been done, e.g., in Refs. 3 and 4 for the $\mathscr{U}_h\mathrm{sl}(N)$ covariant Weyl algebra in N dimensions.

As a toy model for these deformations one can consider the deformed Weyl algebra $\mathscr{A}_h$ in one dimension[5] with generators fulfilling the ''quantum'' commutation relation (QCR)

$$\tilde{A}\tilde{A}^+ = \mathbf{1} + q^2\tilde{A}^+\tilde{A} \tag{I.1}$$

with $q=e^h$. When $q=1$ the above reduces to the classical Weyl algebra $\mathscr{A}$,

$$aa^+ = \mathbf{1} + a^+a. \tag{I.2}$$

If we define $n := a^+a$, $(x)_z=(z^x-1)/(z-1)$, and[6]

$$A := a\sqrt{\frac{(n)_{q^2}}{n}}, \quad A^+ := \sqrt{\frac{(n)_{q^2}}{n}}\,a^+, \tag{I.3}$$

we find out that A, A^+ fulfill the QCR (I.1); hence we can define an algebra homomorphism $f:\mathscr{A}_h \to \mathscr{A}[[h]]$, or ''deforming map'' (in the terminology of Refs. 7 and 8), starting from

$$f(\tilde{A}) = A, \quad f(\tilde{A}) = A^+. \tag{I.4}$$

The right-hand sides of (I.3) have to be understood as formal power series in the deformation parameter h.

a)Electronic mail: Gaetano.Fiore@physik.uni-muenchen.de

 © 1998 American Institute of Physics

We are interested in deformed multidimensional Weyl or Clifford algebras A_h where the QCR:

(1) keep a quadratic structure as in Eq. (I.1), so that one can represent the generators as creation or annihilation operators; and
(2) are covariant under the action $\widetilde{\triangleright}_h$ of $U_h\mathbf{g}$.

More precisely, the generators $\tilde{A}_i^+$, $\tilde{A}^i$ should transform linearly under the action of $\widetilde{\triangleright}_h$,

$$x\widetilde{\triangleright}_h\tilde{A}_i^+=\tilde{\rho}\,^j_i(x)A_j^+, \tag{I.5}$$

$$x\widetilde{\triangleright}_h\tilde{A}^i=\tilde{\rho}^{\vee\,i}_j(x)A^j, \tag{I.6}$$

with $\tilde{\rho}^{\vee}$ being the contragradient representation of the representation $\tilde{\rho}$ of $U_h\mathbf{g}$ ($\tilde{\rho}^{\vee}=\tilde{\rho}^{T}\circ S_h$, where S_h is the antipode of $U_h\mathbf{g}$ and ^{T} is the operation of matrix transposition).

In this work we essentially stick to the case that $U_h\mathbf{g}$ is *triangular*; we treat the general quasitriangular case in Ref. 9. In the former case one can show easily that, for arbitrary $\tilde{\rho}$, $U_h\mathbf{g}$-covariant QCR are given by

$$\tilde{A}^i\tilde{A}_j=\pm R^{ij}_{vu}\tilde{A}^u\tilde{A}^v, \tag{I.7}$$

$$\tilde{A}_i^+\tilde{A}_j^+=\pm R^{vu}_{ij}\tilde{A}_u^+\tilde{A}_v^+, \tag{I.8}$$

$$\tilde{A}^i\tilde{A}_j^+=\delta^i_j\mathbf{1}_{\mathscr{A}}\pm R^{ui}_{jv}\tilde{A}_u^+\tilde{A}^v; \tag{I.9}$$

here the sign $\pm$ refers to the Weyl/Clifford case, respectively, and R is the corresponding ''R-matrix'' of $U_h\mathbf{g}$. [One just has to note that $R=(\tilde{\rho}\otimes\tilde{\rho})\mathscr{R}$, where $\mathscr{R}$ is the universal triangular structure of H_h, and that $\tau\circ\Delta_h(x)=\mathscr{R}\Delta_h(x)\mathscr{R}^{-1}$ (τ denotes the flip operator).]

Here $\mathscr{A}_h$ is a left-module algebra of $U_h\mathbf{g}$: the ''quantum'' action $\widetilde{\triangleright}_h$ is extended to products of the generators as a left-module algebra map $\widetilde{\triangleright}_h: U_h\times\mathscr{A}_h\to\mathscr{A}_h$ (i.e., consistently with the QCR) using the coproduct $\Delta_h(x)=\Sigma_\mu\, x^\mu_{(1)}\otimes x^\mu_{(2)}$ of $U_h\mathbf{g}$,

$$x\widetilde{\triangleright}_h(a\cdot b)=\sum_\mu\,(x^\mu_{(1)}\widetilde{\triangleright}_h a)\cdot(x^\mu_{(2)}\widetilde{\triangleright}_h b), \tag{I.10}$$

because the (I.8) are covariant under (i.e., compatible with) $\widetilde{\triangleright}_h$.

The existence of deforming maps for arbitrary (i.e., not necessarily of the kind described above) deformations of Weyl (or Clifford) algebras is a consequence[10] of a theorem[11] asserting the triviality of the cohomology groups of the latter. (See Refs. 12 and 9 for effective and concise presentations of these results. See also Ref. 13, where the problem of stability of quantum mechanics under deformations was addressed for the first time.) However, no general method for their explicit construction is available. Actually, using cohomological arguments, one can also easily show that deforming maps are unique up to an inner automorphism,

$$f\to f_\alpha:=\alpha f(\cdot)\alpha^{-1},\quad \alpha=\mathbf{1}_{\mathscr{A}}+O(h); \tag{I.11}$$

therefore it is enough to construct one to find all of them.

In this work we present a general method which allows us, given a triangular Hopf algebra $U_h\mathbf{g}$ and any $U_h\mathbf{g}$-covariant deformed Weyl or Clifford algebra, to explicitly construct the corresponding deforming maps f and the corresponding realizations $\triangleright_h$ of $\widetilde{\triangleright}_h$ [$\triangleright_h$ is defined by $\triangleright_h:=(\mathrm{id}\otimes f)\circ\widetilde{\triangleright}_h\circ(\mathrm{id}\otimes f^{-1})$]. In a first attempt to generalize our construction procedure to *quasitriangular* $U_h\mathbf{g}$, we also generalize the construction to the case that $U_h\mathbf{g}$ is the quantum group $\mathscr{U}_h\mathrm{sl}(2)$ and $\tilde{\rho}$ is its fundamental representation. Finally we investigate on the status of deforming maps at the representation-theoretic level.

The construction method is based (Sec. II A) on use of the Drinfel'd–Reshetikhin twist $\mathscr{F}$,[14,15] intertwining between the coproducts of $U\mathbf{g}$ and $U_h\mathbf{g}$, and on the fact that within $\mathscr{A}[[h]]$ one can

realize both the action $\triangleright$ of $U_h\mathbf{g}$ (Sec. II A) and the action $\triangleright_h$ in an ''adjointlike'' way. We show first (Sec. III) that $\mathscr{F}$ can be used in a *universal* way to construct, within $\mathscr{A}_h$, $U_h\mathbf{g}$-tensors out of $U\mathbf{g}$-tensors, and in particular out of a_i^+, a^i objects A_i^+, A^i that transform under $\triangleright_h$ as in formula (I.6). Then (Sec. IV) we verify that the objects A_i^+, A^j really satisfy the QCR (I.8). In Sec. V we generalize our construction (by means of the Drinfel'd twist[16]) to the case of deformed Weyl and Clifford algebras with generators belonging to the fundamental representation of the quantum group $\mathscr{U}_h\mathrm{sl}(2)$; the deforming map is again completely explicit thanks to the semiuniversal expression[8] for $\mathscr{F}$. We compare our deforming map with the one previously found in Ref. 17. At the representation-theoretic level it would be natural to interpret deforming maps as ''operator maps,'' in other words as intertwiners between the representations of $\mathscr{A}$ and $\mathscr{A}_h$. However, we have to expect that, in the role of intertwiners, deforming maps may become singular at $h=0$, because the representation theories of $\mathscr{A}$, $\mathscr{A}_h$ are in general rather different. In Sec. VI we show that there is always a $*$-representation of $\mathscr{A}_h$ which is intertwined by f with the Fock representation of $\mathscr{A}$; this allows us to interpret $\tilde{A}^i, \tilde{A}_i^+$ as ''composite'' operators on a classical Fock space describing ordinary bosons and fermions. We also explicitly show that f^{-1} is ill defined as an intertwiner from the remaining (if any) unitarily inequivalent $*$-representations of $\mathscr{A}_h$.

On the basis of the above result we conclude that also at the quantization-of-field level noncommutative Hopf algebra symmetries are not necessarily incompatible with Bose or Fermi statistics (contrary to what is often claimed). We arrived at the same conclusion at the first-quantization level in Refs. 18 and 19, where the initial motivation for the present work originated. The connection between the two approaches through second quantization will be described elsewhere.

II. PRELIMINARIES AND NOTATION

A. Twisting groups into quantum groups

Let $H=(U\mathbf{g}, m, \Delta, \varepsilon, S)$ be the cocommutative Hopf algebra associated to the universal enveloping (UE) algebra $U\mathbf{g}$ of a Lie algebra $\mathbf{g}$. The symbol m denotes the multiplication [in the sequel it will be dropped in the obvious way $m(a\otimes b)\equiv ab$, unless explicitly required], whereas Δ, ε, and S are the comultiplication, counit, and antipode, respectively.

Let $\mathscr{F}\in U\mathbf{g}[[h]]\otimes U\mathbf{g}[[h]]$ (we will write $\mathscr{F}=\mathscr{F}^{(1)}\otimes\mathscr{F}^{(2)}$, in a Sweedler's notation with *upper* indices; on the rhs a sum $\Sigma_i \mathscr{F}_i^{(1)}\otimes\mathscr{F}_i^{(2)}$ of many terms is implicitly understood) be a ''twist,'' i.e., an element satisfying the relations

$$(\varepsilon\otimes\mathrm{id})\mathscr{F}=\mathbf{1}=(\mathrm{id}\otimes\varepsilon)\mathscr{F}, \tag{II.1}$$

$$\mathscr{F}=\mathbf{1}\otimes\mathbf{1}+O(h) \tag{II.2}$$

($h\in\mathbf{C}$ is the ''deformation parameter,'' and $\mathbf{1}$ the unit in $U\mathbf{g}$; from the second condition it follows that $\mathscr{F}$ is invertible as a power series). It is well known[14] that if $\mathscr{F}$ also satisfies the relation

$$(\mathscr{F}\otimes\mathbf{1})[(\Delta\otimes\mathrm{id})(\mathscr{F})]=(\mathbf{1}\otimes\mathscr{F})[(\mathrm{id}\otimes\Delta)(\mathscr{F})], \tag{II.3}$$

and $(U_h\mathbf{g}, m_h)$ is an algebra isomorphic to $U\mathbf{g}[[h]]$ with isomorphism, say, $\varphi_h: U_h\mathbf{g}\to U\mathbf{g}[[h]]$ (in particular, if $U_h\mathbf{g}=U\mathbf{g}[[h]]$ and $\varphi_h=\mathrm{id}\ (\mathrm{mod}\ h)$, or even $\varphi_h=\mathrm{id}$), then one can construct a triangular non-cocommutative Hopf algebra $H_h=(U_h\mathbf{g}, m_h, \Delta_h, \varepsilon_h, S_h, \mathscr{R})$ having an isomorphic algebra structure $[m_h=\varphi_h^{-1}\circ m\circ(\varphi_h\otimes\varphi_h)]$, an isomorphic counit $\varepsilon_h:=\varepsilon\circ\varphi_h^{-1}$, comultiplication and antipode defined by

$$\Delta_h(a)=(\varphi_h^{-1}\otimes\varphi_h^{-1})\{\mathscr{F}\Delta[\varphi_h(a)]\mathscr{F}^{-1}\},\quad S_h(a)=\varphi_h^{-1}\{\gamma^{-1}S[\varphi_h(a)]\gamma\}, \tag{II.4}$$

where

$$\gamma:=S\mathscr{F}^{-1(1)}\cdot\mathscr{F}^{-1(2)},\quad \gamma^{-1}=\mathscr{F}^{(1)}\cdot S\mathscr{F}^{(2)}, \tag{II.5}$$

and (triangular) universal R-matrix

$$\mathscr{R}:=[\varphi_h^{-1}\otimes\varphi_h^{-1}](\mathscr{F}_{21}\mathscr{F}^{-1}),\quad \mathscr{F}_{21}:=\mathscr{F}^{(2)}\otimes\mathscr{F}^{(1)}. \tag{II.6}$$

Condition (II.3) ensures that Δ_h is coassociative as Δ. The inverse of S_h is given by $S_h^{-1}(a) = \varphi_h^{-1}\{\gamma' S[\varphi_h(a)]\gamma'^{-1}\}$, where

$$\gamma' := \mathscr{F}^{(2)} \cdot S\mathscr{F}^{(1)}, \quad \gamma'^{-1} = S\mathscr{F}^{-1(2)} \cdot \mathscr{F}^{-1(1)}; \tag{II.7}$$

$\gamma^{-1}\gamma' \in \mathrm{Centre}(U\mathbf{g})$, and $S\gamma = \gamma'^{-1}$.

Conversely, given an h-deformation $H_h = (U_h\mathbf{g}, m_h, \Delta_h, \varepsilon_h, S_h, \mathscr{R})$ of H in the form of a triangular Hopf algebra, one can find[14] an isomorphism $\varphi_h : U_h \to U\mathbf{g}[[h]]$ and an invertible $\mathscr{F}$ satisfying conditions (II.1)–(II.3) such that H_h can be obtained from H through formulas (II.4), (II.5), and (II.7).

Examples of $\mathscr{F}$'s satisfying conditions (II.3), (II.1), and (II.2) are provided, e.g., by the so-called ''Reshetikhin twists'' [15]

$$\mathscr{F} := e^{h\omega_{ij} h_i \otimes h_j}, \tag{II.8}$$

where $\{h_i\}$ is a basis of the Cartan subalgebra of $\mathbf{g}$ and $\omega_{ij} = -\omega_{ji} \in \mathbf{C}$. A less obvious example is, for instance, the ''Jordanian'' deformation of Ref. 20.

A similar result to the above holds for genuine quantum groups. A well-known theorem by Drinfel'd, Proposition 3.16 in Ref. 16, proves, for any quasitriangular deformation $H_h = (U_h\mathbf{g}, m_h, \Delta_h, \varepsilon_h, S_h, \mathscr{R})$[2,21] of $U\mathbf{g}$, with $\mathbf{g}$ a simple finite-dimensional Lie algebra, the existence of an algebra isomorphism $\varphi_h : U_h\mathbf{g} \to U\mathbf{g}[[h]]$ and an invertible $\mathscr{F}$ satisfying condition (II.1) such that H_h can be obtained from H through formulas (II.4), (II.5), and (II.7), as well. This $\mathscr{F}$ does not satisfy condition (II.18); however, the (nontrivial) coassociator

$$\phi := [(\Delta \otimes \mathrm{id})(\mathscr{F}^{-1})](\mathscr{F}^{-1} \otimes \mathbf{1})(\mathbf{1} \otimes \mathscr{F})[(\mathrm{id} \otimes \Delta)(\mathscr{F})] \tag{II.9}$$

still commutes with $\Delta^{(2)}(U\mathbf{g})$,

$$[\phi, \Delta^{(2)}(U\mathbf{g})] = 0, \tag{II.10}$$

thus explaining why Δ_h is coassociative in this case, too. The corresponding universal (quasitriangular) R-matrix $\mathscr{R}$ is related to $\mathscr{F}$ by

$$\mathscr{R} = [\varphi_h^{-1} \otimes \varphi_h^{-1}](\mathscr{F}_{21} q^{t/2} \mathscr{F}^{-1}), \tag{II.11}$$

where $t := \Delta(\mathscr{C}) - \mathbf{1} \otimes \mathscr{C} - \mathscr{C} \otimes \mathbf{1}$ is the canonical invariant element in $U\mathbf{g} \otimes U\mathbf{g}$ ($\mathscr{C}$ is the quadratic Casimir). The twist $\mathscr{F}$ is defined (and unique) up to the transformation

$$\mathscr{F} \to \mathscr{F} T, \tag{II.12}$$

where T is a $\mathbf{g}$-invariant [i.e., commuting with $\Delta(U\mathbf{g})$] element of $U\mathbf{g}[[h]]^{\otimes^2}$ such that

$$T = \mathbf{1} \otimes \mathbf{1} + O(h), \quad (\varepsilon \otimes \mathrm{id})T = \mathbf{1} = (\mathrm{id} \otimes \varepsilon)T. \tag{II.13}$$

A function

$$T = T(\mathbf{1} \otimes \mathscr{C}_i, \mathscr{C}_i \otimes \mathbf{1}, \Delta(\mathscr{C}_i)) \tag{II.14}$$

of the Casimirs $\mathscr{C}_i \in U\mathbf{g}$ of $U\mathbf{g}$ and of their coproducts clearly is $\mathbf{g}$ invariant.

In general, as a consequence of the existence of an isomorphism φ_h, representations $\tilde{\rho}$, ρ of deformed and undeformed algebras are in one-to-one correspondence (except for special values of h making it singular) through

$$\rho = \tilde{\rho} \circ \varphi_h. \tag{II.15}$$

A special case of interest is when $U\mathbf{g}$ is a $*$-Hopf algebra and $\mathscr{F}$ is unitary,

$$\mathscr{F}^{*\otimes *} = \mathscr{F}^{-1}; \tag{II.16}$$

note that in this case

$$\gamma' = \gamma^*. \tag{II.17}$$

One can show[22] that $\mathscr{F}$ can always be made unitary if $\mathbf{g}$ is compact.

We will often use a ''tensor notation'' for our formulas: Eq. (II.3) will read

$$\mathscr{F}_{12}\mathscr{F}_{12,3} = \mathscr{F}_{23}\mathscr{F}_{1,23}, \tag{II.18}$$

and definition (II.9) will read $\phi \equiv \phi_{123} = \mathscr{F}_{12,3}^{-1}\mathscr{F}_{12}^{-1}\mathscr{F}_{23}\mathscr{F}_{1,23}$, for instance; the comma separates the tensor factors *not* stemming from the coproduct. For practical purposes it will be often convenient in the sequel to use the Sweedler's notation with *lower* indices $\Delta(x) \equiv x_{(1)} \otimes x_{(2)}$ for the cocommutative coproduct (on the rhs a sum $\Sigma_i x^i_{(1)} \otimes x^i_{(2)}$ of many terms is implicitly understood); similarly, we will use the Sweedler's notation $\Delta^{(n-1)}(x) \equiv x_{(1)} \otimes \cdots \otimes x_{(n)}$ for the $(n-1)$-fold coproduct. For the non-cocommutative coproducts Δ_h, instead, we will use a Sweedler's notation with barred indices: $\Delta_h(x) \equiv x_{(\bar{1})} \otimes x_{(\bar{2})}$.

B. Classical *U*g-covariant creators and annihilators

Let $\mathscr{A}$ be the unital algebra generated by $\mathbf{1}_{\mathscr{A}}$ and elements $\{a_i^+\}_{i \in I}$ and $\{a^j\}_{j \in I}$ satisfying the (anti)commutation relations

$$[a^i,\ a^j]_\pm = 0, \quad [a_i^+,\ a_j^+]_\pm = 0, \tag{II.19}$$

$$[a^i,\ a_j^+]_\pm = \delta^i_j \mathbf{1}_{\mathscr{A}}$$

(the $\pm$ sign denotes commutators and anticommutators, respectively), belonging respectively to some representation ρ and to its contragradient $\rho^\vee = \rho^T \circ S$ of H (^{T} is the transpose):

$$\begin{aligned} x \triangleright a_i^+ &= \rho(x)^l_i a_l^+, \\ x \triangleright a^i &= \rho(Sx)^i_l a^l, \end{aligned} \quad x \in U\mathbf{g}, \quad \rho(x)^i_j \in \mathbf{C}. \tag{II.20}$$

Equivalently, one says that a_i^+, a^i are ''covariant'' under $\triangleright$, or that they span two (left) modules of $U\mathbf{g}$:

$$(xy) \triangleright a = x \triangleright (y \triangleright a), \quad x,y \in U\mathbf{g}, \tag{II.21}$$

with either $a = a^i$ or a_i^+.

Here $\mathscr{A}$ is a (left) module algebra of $(H, \triangleright)$, if the action $\triangleright$ is extended on the whole $\mathscr{A}$ by means of the (cocommutative) coproduct

$$x \triangleright (ab) = (x_{(1)} \triangleright a)(x_{(2)} \triangleright b). \tag{II.22}$$

Then property (II.21) holds for all $a \in \mathscr{A}$.

Setting

$$\sigma(X) := \rho(X)^i_j a_i^+ a^j \tag{II.23}$$

for all $X \in \mathbf{g}$, one finds that σ: $\mathbf{g} \to \mathscr{A}$ is a Lie algebra homomorphism, so that σ can be extended to all of $U\mathbf{g}$ as an algebra homomorphism σ: $U\mathbf{g} \to \mathscr{A}$; on the unit element we set $\sigma(\mathbf{1}_{U\mathbf{g}}) := \mathbf{1}_{\mathscr{A}}$. σ can be seen as the generalization of the Jordan–Schwinger realization of $\mathbf{g} = \mathrm{su}(2)$ [formula (V.8)].[23]

Then it is easy to check the following.

Proposition 1: The (left) action $\triangleright: U\mathbf{g} \times \mathscr{A} \to \mathscr{A}$ can be realized in an ''adjointlike'' way:

$$x \triangleright a = \sigma(x_{(1)}) a \sigma(Sx_{(2)}), \quad x \in U\mathbf{g}, \quad a \in \mathscr{A}. \tag{II.24}$$

In the specially interesting case of a compact section $\mathbf{g}$ (with $*$-structure ''$*$'') one can introduce in $\mathscr{A}$ a $*$-structure, the ''Hermitian conjugation'' (which we will denote by $\star$), such that

$$(a^i)^\star = a_i^+. \tag{II.25}$$

Correspondingly, ρ is a $*$-representation ($\rho\circ\star=*\circ\rho^T$) and σ becomes a $*$-homomorphism, i.e., $\sigma\circ *=\star\circ\sigma$.

III. QUANTUM COVARIANT CREATORS AND ANNIHILATORS

Let H_h and φ_h be as in Sec. II A. Clearly, $\sigma_{\varphi_h}:=\sigma\circ\varphi_h$ is an algebra homomorphism σ_{φ_h}: $U_h\mathbf{g}\to\mathscr{A}[[h]]$. Inspired by Proposition 3 we are led to define

$$x\triangleright_h a := \sigma_{\varphi_h}(x_{(\bar{1})})\,a\,\sigma_{\varphi_h}(S_h x_{(\bar{2})}). \tag{III.1}$$

Using the Hopf algebra axioms it is straightforward to prove the relations [cf. relations (I.10)]

$$\begin{aligned}&(xy)\triangleright_h a = x\triangleright_h(y\triangleright_h a),\\ &x\triangleright_h(ab)=(x_{(\bar{1})}\triangleright_h a)(x_{(\bar{2})}\triangleright_h b),\end{aligned}\qquad \forall x,y\in U\mathbf{g}[[h]],\quad \forall a,b\in\mathscr{A}[[h]]. \tag{III.2}$$

In other words, we have the following.

Proposition 2. The definition (III.1) realizes $\triangleright_h$ *(the left action of* H_h*) on the algebra* $\mathscr{A}[[h]]$.

However, a_i^+, a^j are *not* covariant w.r.t. to $\triangleright_h$. One may ask whether there exist some objects A_i^+, $A^j\in\mathscr{A}$ that *are* covariant under $\triangleright_h$ and transform as in Eq. (I.8).

The answer comes from the following crucial proposition.

Proposition 3: The elements

$$A_i^+ := \sigma(\mathscr{F}^{(1)})\,a_i^+\,\sigma(S\mathscr{F}^{(2)}\gamma)\in\mathscr{A}[[h]], \tag{III.3}$$

$$A^i := \sigma(\gamma' S\mathscr{F}^{-1(2)})\,a^i\,\sigma(\mathscr{F}^{-1(1)})\in\mathscr{A}[[h]] \tag{III.4}$$

are "covariant" under $\triangleright_h$*, more precisely belong respectively to the representation* $\tilde{\rho}$ *and to its quantum contragredient* $\tilde{\rho}^\vee=\tilde{\rho}^T\circ S_h$ *of* $U_h\mathbf{g}$ *acting through* $\triangleright_h$:

$$x\triangleright_h A_i^+=\tilde{\rho}(x)_i^l A_l^+,\quad x\triangleright_h A^i=\tilde{\rho}(S_h x)_m^i A^m. \tag{III.5}$$

Proof: Due to relation (II.4), $\mathscr{F}$ is an intertwiner between Δ_h and Δ (in this proof we drop the symbol φ_h):

$$x_{(\bar{1}')}\mathscr{F}^{(1)}\otimes x_{(\bar{2}')}\mathscr{F}^{(2)}=\mathscr{F}^{(1)}x_{(1')}\otimes\mathscr{F}^{(2)}x_{(2')}. \tag{III.6}$$

Applying $\mathrm{id}\otimes S$ on both sides of the equation and multiplying the result by $\mathbf{1}\otimes\gamma$ from the right we find [with the help of relation (II.5)]

$$x_{(\bar{1}')}\mathscr{F}^{(1)}\otimes(S\mathscr{F}^{(2)})\,\gamma S_h x_{(\bar{2}')}=\mathscr{F}^{(1)}x_{(1')}\otimes(Sx_{(2')})(S\mathscr{F}^{(2)})\,\gamma. \tag{III.7}$$

Applying $\sigma\otimes\sigma$ to both sides and sandwiching a_i^+ between the two tensor factors we find

$$\sigma(x_{(\bar{1}')})A_i^+\,\sigma(S_h x_{(\bar{2}')})=\sigma(\mathscr{F}^{(1)})\,\sigma(x_{(1')})\,a_i^+\,\sigma(Sx_{(2')})\,\sigma[(S\mathscr{F}^{(2)})\,\gamma], \tag{III.8}$$

which, in view of formula (III.1), proves the first relation.

To prove the second relation, let us note that relation (II.4) implies an analogous relation

$$\Delta_h(a)\tilde{\mathscr{F}}=\tilde{\mathscr{F}}\Delta(a),\quad \text{with } \tilde{\mathscr{F}}:=[\gamma' S\mathscr{F}^{-1\ (2)}\otimes\gamma' S\mathscr{F}^{-1\ (1)}]\Delta(S\gamma). \tag{III.9}$$

This can be shown by applying in order the following operations to both sides of Eq. (II.4): multiplying by $\mathscr{F}^{-1}$ from the left and from the right, applying $S\otimes S$, multiplying by $\gamma'\otimes\gamma'$ from the left and by $\Delta(S\gamma)$ from the right, replacing $a\to S_h x$, and using the properties (II.4) and $(S_h\otimes S_h)\circ\Delta_h=\tau\circ\Delta_h\circ S_h$. Next, we observe that A^i can be rewritten as

$$A^i=\sigma[\tilde{\mathscr{F}}^{(1)}S(\gamma^{-1})_{(1)}]\,a^i\,\sigma[(\gamma^{-1})_{(2)}S\tilde{\mathscr{F}}^{(2)}\gamma]=\sigma(\tilde{\mathscr{F}}^{(1)})\,a^l\,\sigma(S\tilde{\mathscr{F}}^{(2)}\gamma)\,\rho(\gamma^{-1})_l^i; \tag{III.10}$$

where, reasoning as for the first relation,

$$\sigma(x_{(1')})A^i\sigma(S_h x_{(2')}) \overset{(III.10)}{=} \sigma(\tilde{\mathscr{F}}^{(1)})\sigma(x_{(1')})a^l\sigma(Sx_{(2')})\sigma[(S.\tilde{\mathscr{F}}^{(2)})\gamma]\rho(\gamma^{-1})^i_l$$

$$\overset{(II.24)}{=} \sigma(\tilde{\mathscr{F}}^{(1)})a^l\sigma[(S.\tilde{\mathscr{F}}^{(2)})\gamma]\rho(\gamma^{-1}Sx)^i_l$$

$$\overset{(II.4)}{=} \sigma(\tilde{\mathscr{F}}^{(1)})a^l\sigma[(S.\tilde{\mathscr{F}}^{(2)})\gamma]\rho(S_h x\cdot\gamma^{-1})^i_l$$

$$\overset{(III.10)}{=} \rho(S_h x)^i_l A^l,$$

which proves the second relation. □

Remark 1: The proposition clearly holds even if one chooses in formulas (III.3) and (III.4) two $\mathscr{F}$'s differing by a tranformation (II.12). We shall exploit this freedom when $U_h\mathbf{g}$ is genuinely quasitriangular.

Remark 2: Note that in the $*$-Hopf algebra case Eqs. (II.25), (II.16), and (II.17) imply

$$(A^i)^\star = A^+_i . \tag{III.11}$$

Remark 3: Under the right action $\triangleleft_h$ ($a\triangleleft_h x := (S_h^{-1}x)\triangleright_h a$ with $a\in\mathscr{A}$, $x\in U\mathbf{g}$) the covariance properties of A^i, A^+_i read

$$A^i\triangleleft_h x = \rho(x)^i_l A^l, \quad A^+_i\triangleleft_h x = \rho(S_h^{-1}x)^m_i A^+_m . \tag{III.12}$$

Remark 4: σ_{φ_h} is not the only algebra homomorphism $U_h\mathbf{g}\to\mathscr{A}[[h]]$. For any $\alpha = \mathbf{1}_{\mathscr{A}} + O(h)$ we find a new one by setting

$$\sigma_{\varphi_h,\alpha}(x) = \alpha\sigma_{\varphi_h}(x)\alpha^{-1}; \tag{III.13}$$

correspondingly, we can define a new realization of $\tilde{\triangleright}_h$ by

$$x\triangleright_{h,\alpha} a := \sigma_{\varphi_h,\alpha}(x_{(1)})a\sigma_{\varphi_h,\alpha}(S_h x_{(2)}). \tag{III.14}$$

Covariant objects under $\triangleright_{h\alpha}$ will be given by

$$A^+_{i,\alpha} := \alpha A^+_i\alpha^{-1}, \quad A^{i,\alpha} := \alpha A^i\alpha^{-1}. \tag{III.15}$$

If relations (III.11) hold, we can preserve them by choosing $\alpha^\star = \alpha^{-1}$.

To conclude this section, let us give useful alternative expressions for A^+_i, A^i by ''moving'' to the right/left past a^+_i, a^i the expressions $\sigma(\cdot)$ lying at their left/right in definitions (III.3) and (III.4).

Lemma 1: If $\mathscr{T}\in U\mathbf{g}[[h]]^{\otimes^3}$ *is* $\mathbf{g}$ *invariant* (*i.e.*, $[\mathscr{T}, U\mathbf{g}[[h]]^{\otimes^3}]=0$), *then* $m_{ij}S_i.\mathscr{T}$, $m_{ij}S_j.\mathscr{T}$ ($i,j=1,2,3$, $i\neq j$) *are* $\mathbf{g}$ *invariants belonging to* $U\mathbf{g}[[h]]^{\otimes^2}$.

(Here S_i denotes S acting on the ith tensor factor, and m_{ij} multiplication of the ith tensor factor by the jth from the right.)

Proof: For instance,

$$\mathscr{T}^{(1)}x_{(1)}\otimes.\mathscr{T}^{(2)}x_{(2)}\otimes.\mathscr{T}^{(3)}x_{(3)}\otimes x_{(4)} = x_{(1)}.\mathscr{T}^{(1)}\otimes x_{(2)}.\mathscr{T}^{(2)}\otimes x_{(3)}.\mathscr{T}^{(3)}\otimes x_{(4)}$$

$$\Downarrow (m_{23})^2\circ S_3$$

$$\mathscr{T}^{(1)}x_{(1)}\otimes.\mathscr{T}^{(2)}x_{(2)}Sx_{(3)}S.\mathscr{T}^{(3)}x_{(4)} = x_{(1)}.\mathscr{T}^{(1)}\otimes x_{(2)}.\mathscr{T}^{(2)}S.\mathscr{T}^{(3)}Sx_{(3)}x_{(4)}$$

for any $x\in U\mathbf{g}[[h]]$, whence (because of $a_{(1)}Sa_{(2)} = \varepsilon(a) = Sa_{(1)}a_{(2)}$)

$$\mathscr{T}^{(1)}x_{(1)}\otimes.\mathscr{T}^{(2)}S.\mathscr{T}^{(3)}x_{(2)} = x_{(1)}.\mathscr{T}^{(1)}\otimes x_{(2)}.\mathscr{T}^{(2)}S.\mathscr{T}^{(3)}, \tag{III.16}$$

so that $.\mathscr{T}^{(1)}\otimes.\mathscr{T}^{(2)}S.\mathscr{T}^{(3)}\in U\mathbf{g}[[h]]^{\otimes^2}$ is $\mathbf{g}$ invariant. □

Proposition 4: If H_h is triangular, the definitions (III.3) and (III.4) amount to

$$A_i^+=a_l^+\sigma(.\mathscr{T}^{-1(2)})\rho(.\mathscr{T}^{-1(1)})_i^l, \tag{III.17a}$$

$$A_i^+=\rho(S.\mathscr{T}^{(1)}\gamma'^{-1})_i^l\sigma(.\mathscr{T}^{(2)})a_l^+, \tag{III.17b}$$

$$A^{\ i}=\rho(.\mathscr{T}^{(1)})_l^i\sigma(.\mathscr{T}^{(2)})a^l, \tag{III.17c}$$

$$A^i=a^l\sigma(.\mathscr{T}^{-1(2)})\rho(\gamma^{-1}S.\mathscr{T}^{-1(1)})_l^i. \tag{III.17d}$$

If H_h is quasitriangular, the same formulas hold with in general four different $\mathscr{T}$'s [related to each other by transformations (II.12)].

Proof: Observing that

$$\sigma(x)a=\sigma(x_{(1)})a\,\sigma(Sx_{(2)}\cdot x_{(3)}), \tag{III.18}$$

$$a\,\sigma(x)=\sigma(x_{(3)}Sx_{(2)})a\,\sigma(x_{(1)}), \tag{III.19}$$

for all $x\in U\mathbf{g}$, $a\in\mathscr{A}$, we find from relations (III.3), (III.4), and (II.20)

$$A_i^+=a_l^+\sigma(.\mathscr{T}^{(1)}_{(2)}S.\mathscr{T}^{(2)}\gamma)\rho(.\mathscr{T}^{(1)}_{(1)})_i^l, \tag{III.20}$$

$$A^i=\sigma(\gamma'S.\mathscr{T}^{-1(2)}\cdot.\mathscr{T}^{-1(1)}_{(2)})\rho(.\mathscr{T}^{-1(1)}_{(1)})_l^iA^l. \tag{III.21}$$

On the other hand, applying the previous lemma to $.\mathscr{T}=\phi$ [formula (II.9)] we find that

$$T:=m_{23}(\mathrm{id}\otimes\mathrm{id}\otimes S)\phi\overset{(II.5)}{=}(.\mathscr{T}^{-1(1)}_{(1)}\otimes.\mathscr{T}^{-1(1)}_{(2)}).\mathscr{T}^{-1}(\mathbf{1}\otimes\gamma^{-1}S.\mathscr{T}^{-1(2)})$$

is $\mathbf{g}$ invariant, where one easily finds the relation

$$.\mathscr{T}^{(1)}_{(1)}\otimes.\mathscr{T}^{(1)}_{(2)}S(.\mathscr{T}^{(2)})\gamma=T^{-1}.\mathscr{T}^{-1}\equiv.\mathscr{T}'^{-1}, \tag{III.22}$$

after noting that $[T,.\mathscr{T}^{\pm1(1)}_{(1)}\otimes.\mathscr{T}^{\pm1(1)}_{(2)}]=0$. Replacing in Eq. (III.20) one finds relation (III.17a). In the triangular case $\phi\equiv\mathbf{1}^{\otimes^3}$, implying $T\equiv\mathbf{1}^{\otimes^2}$ and $\mathscr{T}'=\mathscr{T}$. Similarly one proves the other relations. Relations (III.17c) and (III.17d), can be found also more directly starting from relations (III.17a) and (III.17b) by observing that in the unitary-$\mathscr{T}$ case they follow from the latter two by applying the $*$-conjugation. □

IV. QUANTUM COMMUTATION RELATIONS: THE TRIANGULAR CASE

Theorem 1: *If the non-cocommutative Hopf algebra H_h is triangular [i.e., the twist $\mathscr{T}$ satisfies Eq. (II.1)], then A^i,A_j^+ close the quadratic commutation relations*

$$A^iA_j^+=\delta_j^i\mathbf{1}_{\mathscr{A}}\pm R_{jv}^{ui}A_u^+A^v, \tag{IV.1}$$

$$A^iA^j=\pm R_{vu}^{ij}A^uA^v, \tag{IV.2}$$

$$A_i^+A_j^+=\pm R_{ij}^{vu}A_u^+A_v^+, \tag{IV.3}$$

where R is the (numerical) quantum R-matrix of $U\mathbf{g}$ in the representation ρ,

$$R_{hk}^{ij}:=[(\rho\otimes\rho)(.\mathscr{R})]_{hk}^{ij}. \tag{IV.4}$$

Proof: Beside Eqs. (III.17), we will need their ''inverse'' relations:

$$a_l^+=A_i^+\sigma(.\mathscr{T}^{(2)})\rho(.\mathscr{T}^{(1)})_l^i,$$

$$a_l^+ = \rho(\gamma S.\mathscr{F}^{-1(1)})_l^i \sigma(.\mathscr{F}^{-1(2)}) A_i^+ ,$$
$$a^l = \rho(.\mathscr{F}^{-1(1)})_i^l \sigma(.\mathscr{F}^{-1(2)}) A^i,$$
$$a^l = A^i \sigma(.\mathscr{F}^{(2)}) \rho(S.\mathscr{F}^{(1)} \cdot \gamma)_i^l . \tag{IV.5}$$

Using Eqs. (III.17),

$$A^i A_j^+ \overset{(III.17)}{=} \rho(.\mathscr{F}^{(1)})_l^i \sigma(.\mathscr{F}^{(2)}) a^l a_m^+ \sigma(.\mathscr{F}^{-1(2')}) \rho(.\mathscr{F}^{-1(1')})_j^m$$

$$\overset{(II.19)}{=} \rho(.\mathscr{F}^{(1)})_l^i \sigma(.\mathscr{F}^{(2)}) [\delta_m^l \pm a_m^+ a^l] \sigma(.\mathscr{F}^{-1(2')}) \rho(.\mathscr{F}^{-1(1')})_j^m$$

$$\overset{(III.19)(II.20)}{=} (\rho_j^i \otimes \sigma)[.\mathscr{F}\mathscr{F}^{-1}] \pm \rho(.\mathscr{F}^{(1)}_{(1')} \mathscr{F}^{-1(2')})_l^i a_m^+ \sigma(.\mathscr{F}^{(2)}_{(2)} \mathscr{F}^{-1(2')}_{(2')}) a^l \rho(.\mathscr{F}^{(2)}_{(1)} \mathscr{F}^{-1(1')})_j^m$$

$$\overset{(IV.5)}{=} \delta_j^i \mathbf{1}_{\mathscr{A}} \pm \rho(G^{(1)})_j^m A_m^+ \sigma(G^{(2)}) A^l \rho(G^{(3)})_l^i , \tag{IV.6}$$

where $G := .\mathscr{F}_{12}.\mathscr{F}_{23,1}.\mathscr{F}^{-1}_{1,23}.\mathscr{F}^{-1}_{32}$. On the other hand, applying to Eq. (II.18) the permutations $\tau_{23} \circ \tau_{12}$ and τ_{23} we obtain, respectively,

$$.\mathscr{F}_{12}.\mathscr{F}_{23,1} = .\mathscr{F}_{31}(.\mathscr{F}^{(1)}_{(1)} \otimes .\mathscr{F}^{(2)} \otimes .\mathscr{F}^{(1)}_{(2)}), \tag{IV.7}$$

$$.\mathscr{F}^{-1}_{1,23}.\mathscr{F}^{-1}_{32} = (.\mathscr{F}^{(1)}_{(1)} \otimes .\mathscr{F}^{(2)} \otimes .\mathscr{F}^{(1)}_{(2)})^{-1} .\mathscr{F}^{-1}_{13}, \tag{IV.8}$$

which, replaced in the definition of G [using the definition (II.6)], gives $G = \mathscr{R}_{13}$; this proves Eq. (IV.1).

As for relation (IV.2),

$$A^i A^j \overset{(III.17)}{=} \rho(.\mathscr{F}^{(1)})_l^i \sigma(.\mathscr{F}^{(2)}) a^l a^m \sigma(.\mathscr{F}^{-1(2')}) \rho(\gamma^{-1} S.\mathscr{F}^{-1(1')})_m^j$$

$$\overset{(II.19)}{=} \pm \rho(.\mathscr{F}^{(1)})_l^i \sigma(.\mathscr{F}^{(2)}) a^m a^l \sigma(.\mathscr{F}^{-1(2')}) \rho(\gamma^{-1} S.\mathscr{F}^{-1(1')})_m^j$$

$$\overset{(II.20)}{=} \pm \rho(.\mathscr{F}^{(1)}.\mathscr{F}^{-1(2')}_{(2')})_l^i a^m \sigma(.\mathscr{F}^{(2)}_{(1)} \mathscr{F}^{-1(2')}_{(1')}) a^l \rho(\gamma^{-1} S.\mathscr{F}^{-1(1')} \cdot S.\mathscr{F}^{(2)}_{(2)})_m^j$$

$$\overset{(IV.5)}{=} \pm \rho(.\mathscr{F}^{(1)}.\mathscr{F}^{-1(2')}_{(2')}.\mathscr{F}^{-1(\bar{1})})_l^i A^m \sigma(.\mathscr{F}^{(2'')}.\mathscr{F}^{(2)}_{(1)}.\mathscr{F}^{-1(2')}_{(1')}.\mathscr{F}^{-1(\bar{2})}) \times A^l \rho[\gamma^{-1} S(.\mathscr{F}^{(1'')}.\mathscr{F}^{(2)}_{(2)}.\mathscr{F}^{-1(1')}) \gamma]_m^j$$

$$\overset{(II.4)}{=} \pm \rho(G^{-1(1)})_l^i A^m \sigma(G^{-1(2)}) A^l \rho[S_h(G^{-1(3)}]_m^j . \tag{IV.9}$$

However, we have already shown that $G = .\mathscr{R}_{13}$; by recalling that $(\mathrm{id} \otimes S_h).\mathscr{R}^{-1} = .\mathscr{R}$, relation (IV.2) follows.

Similarly one can prove relation (IV.3), which can be found also more directly by observing that in the unitary-$\mathscr{F}$ case it follows from the previous one by applying the $*$-conjugation and by noting that $\bar{R} = R_{21}$. □

Remark 5: It is interesting to ask how the invariants which can be constructed from A^i, A_j^+ are related to the ones which can be constructed from a^i, a_j^+. It is straightforward to prove that, e.g., any invariant of the form $I_h^n := A_{i_1}^+ \ldots A_{i_n}^+ A^{i_n} \ldots A^{i_1}$, coincides with the invariant of the form

$I^n := a^+_{i_1}\dots a^+_{i_n}, a^{i_n}\dots a^{i_1}$ (in particular, so is $I^2 \equiv N := A^+_i A^i = n$). One can show[9] the equality $P(A^i, A^+_j) = P(a^i, a^+_j)$ for all polynomial invariants $P(\tilde{A}^i, \tilde{A}^+_j)$. This is no longer true if H_h is genuinely quasitriangular.

V. QUANTUM COMMUTATION RELATIONS: THE $\mathscr{U}_h$sl(2) CASE

It is now tempting to consider the quasitriangular case and ask whether a transformation such as in Remark 1 can map a^i, a^+_j into $U_h\mathbf{g}$-covariant A^i, A^+_j satisfying relations of the type found in Refs. 3 and 4. We will stick here to the case $\mathbf{g} = \mathrm{sl}(2,\mathbf{C})$, $\rho \equiv$ fundamental representation, directing the reader to Ref. 9 for the general case.

We fix our conventions as follows. As ''classical'' generators of Usl(2) we choose $j_0, j_+, j_- \in \mathbf{g}$,

$$[j_0,\, j_\pm] = \pm j_\pm, \quad [j_+,\, j_-] = 2j_0, \tag{V.1}$$

$$\Delta(x) = \mathbf{1}\otimes x + x\otimes \mathbf{1}, \quad x = j_0, j_+, j_- . \tag{V.2}$$

As generators of $\mathscr{U}_h$sl(2) we choose J_0, J^+, J^- satisfying

$$[J_0,\, J_\pm] = \pm J_\pm, \quad [J_+,\, J_-] = \frac{q^{2J_0} - q^{-2J_0}}{q - q^{-1}}, \tag{V.3}$$

$$\Delta_h(J_0) = \mathbf{1}\otimes J_0 + J_0\otimes \mathbf{1}, \quad \Delta_h(J_\pm) = J_\pm \otimes q^{-J_0} + q^{J_0}\otimes J_\pm . \tag{V.4}$$

{Let us recall incidentally that the mapping $\varphi_h : \mathscr{U}_h\mathrm{sl}(2) \to U\mathrm{sl}(2)[[h]]$, up to an inner automorphism $\varphi_h \to \varphi_{h,v} := v\,\varphi_h(\cdot)v^{-1}$ is given by[7]

$$\varphi_h(J_0) = j_0, \quad \varphi_h(J_\pm) = \sqrt{\frac{[j \pm j_0]_q [1 + j \mp j_0]_q}{(j \pm j_0)(1 + j \mp j_0)}}\, j_\pm , \tag{V.5}$$

where j is the positive root of the equation $j(j+1) - \mathscr{C} = 0$, $\mathscr{C} = j_- j_+ + j_0(j_0+1)$ is the Casimir, $q = e^h$, and $[x]_q := (q^x - q^{-x})/(q - q^{-1})$.}

At the coalgebra level, the universal $\mathscr{F}$ [connecting Δ to Δ_h in formula (II.4)] is not explicitly known; however the Usl(2)[[h]]-valued matrix $F := (\rho\otimes \mathrm{id})\mathscr{F}$ has been determined in Ref. 8 and reads

$$F = \begin{Vmatrix} a(j,j_0) & b(j,j_0) j_- \\ -j_+ b(j,j_0) & a(j,j_0 - 1) \end{Vmatrix}, \quad F^{-1} = \begin{Vmatrix} a(j,j_0) & -b(j,j_0) j_- \\ j_+ b(j,j_0) & a(j,j_0 - 1) \end{Vmatrix}, \tag{V.6}$$

where

$$a(j,j_0) := \frac{q^{(j-j_0)/q}}{\sqrt{(1+2j)[1+2j]_q}} \left[\sqrt{(1 + j + j_0)[1 + j + j_0]_q} + q^{-(1+2j)/2} \sqrt{(j - j_0)[j - j_0]_q} \right],$$
$$b(j,j_0) := \frac{q^{(j-j_0)/2}}{\sqrt{(1+2j)[1+2j]_q}} \left[\sqrt{\frac{[1 + j + j_0]_q}{1 + j + j_0}} - q^{-(1+2j)/2} \sqrt{\frac{[j - j_0]_q}{j - j_0}} \right]; \tag{V.7}$$

the matrix elements in Eq. (V.6) are w.r.t. the orthonormal basis $\{|\uparrow\rangle, |\downarrow\rangle\}$ (spin up, spin down) of eigenstates of j_0 with eigenvalues $\pm\frac{1}{2}$. $a^+_\uparrow|0\rangle = |\uparrow\rangle$, etc. All indices in the sequel will run over $\{\uparrow,\downarrow\}$. F is unitary w.r.t. the su(2) $*$-structure $j_0^* = j_0$, $(j_+)^* = j_-$.

[See formulas (3.1) and (3.30) in Ref. 8. To match our conventions with theirs, one has to rescale $j_\pm$ by $\sqrt{2}$ and note that the right correspondence between our notation and theirs is $\mathscr{F} \equiv \mathscr{F}_q \leftrightarrow U_{q^{-1}}$, what is needed to match the coproducts.]

The homorphism σ in this case coincides with the well-known Jordan–Schwinger realization of sl(2) (Ref. 23) and reads

$$\sigma(j_+)=a_\uparrow^+a^\downarrow,\quad \sigma(j_-)=a_\downarrow^+a^\uparrow,\quad \sigma(j_0)=\tfrac{1}{2}(a_\uparrow^+a^\uparrow-a_\downarrow^+a^\downarrow), \tag{V.8}$$

implying in particular

$$\sigma(j)=n/2, \tag{V.9}$$

where $n:=a_i^+a^i$ is the ''classical number of particles'' operator. The $\mathscr{U}_h$su(N)-covarient Weyl algebra (with $N\geqslant 2$, $q=e^h$ real and positive) was introduced by Pusz and Woronowicz;[3] independently, Wess and Zumino[4] introduced its $\mathscr{U}_h$sl(N)-covariant generalization (arbitrary complex q) in R-matrix notation. One can consider also its Clifford version.[24] In the R-matrix notation[4] the QCR of the generators (which we will denote here by $\tilde{A}^i,\tilde{A}_j^+$) read

$$\tilde{A}^i\tilde{A}^j=\pm q^{\mp 1}R^{ij}_{vu}\tilde{A}^u\tilde{A}^v, \tag{V.10}$$

$$\tilde{A}_i^+\tilde{A}_j^+=\pm q^{\mp 1}R^{vu}_{ij}\tilde{A}_u^+\tilde{A}_v^+, \tag{V.11}$$

$$\tilde{A}^i\tilde{A}_j^+=\mathbf{1}_{\mathscr{A}}\delta_j^+\pm q^{\pm 1}R^{ui}_{jv}\tilde{A}_u^+\tilde{A}^v, \tag{V.12}$$

where $i,\ j=1,...,N$, $R=(\rho_d\otimes\rho_d)\mathscr{R}$ is the R-matrix of $\mathscr{U}_h$sl(N) in the defining representation ρ_d, and the sign $\pm$ refers to Weyl and Clifford, respectively. [The ''braiding'' (V.12) between $\tilde{A}^i,\tilde{A}_j^+$ could be also replaced by the inverse one: $\tilde{A}^i\tilde{A}_j^+=\mathbf{1}_{\mathscr{A}}\delta_j^i\pm q^{\mp}R^{-1\,ui}_{\ \ jv}\tilde{A}_u^+\tilde{A}^v$. For a comprehensive introduction to braiding, see Ref. 25.] Both have been treated subsequently by many other authors. They are related, but should not be confused, with the celebrated Biedenharn–Macfarlane–Hayashi q-oscillator (super)algebras.[5,26] [The generators α^i,α_j^+ of the latter fulfill ordinary (anti)commutation relations, except for the q-(anti)commutation relations $\alpha^i\alpha_i^+\mp q^2\alpha_i^+\alpha^i$ and are *not* $U_h\mathbf{g}$ covariant (in spite of the fact that they are usually used to construct a generalized Jordan–Schwinger realization of $U_h\mathbf{g}$). It is of interest to note that, however, the generators α^i,α_j^+ can be typically realized as algebraic ''functions'' of $\tilde{A}^i,\tilde{A}_i^+$,[17] whereas the generators A^i,A_i^+ can be typically realized only as *trascendental* ''functions'' of $\tilde{A}^i,\tilde{A}_i^+$.] When $q>0$ the QCR of the generators are compatible with the $*$-structure[3] of annihilators and creators,

$$A_i^+=(A^i)^\star. \tag{V.13}$$

When $N=2$ we will pick up $i,j,\ldots=\uparrow,\downarrow$; the R-matrix will read

$$R\equiv\|R^{ij}_{hk}\|:=\begin{Vmatrix} q & & & \\ & 1 & & \\ & (q-q^{-1}) & 1 & \\ & & & q \end{Vmatrix}, \tag{V.14}$$

where the row and columns of the matrix (V.14) are ordered in the usual way: $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, $\downarrow\downarrow$ from left to right and from up to down.

More explicitly, the Weyl QCR (V.10)–(V.12) read

$$\begin{aligned}
\tilde{A}^\uparrow\tilde{A}_\uparrow^+&=\mathbf{1}_{\mathscr{A}}+q^2\tilde{A}_\uparrow^+\tilde{A}^\uparrow+(q^2-1)\tilde{A}_\downarrow^+\tilde{A}^\downarrow,\\
\tilde{A}^\downarrow\tilde{A}_\downarrow^+&=\mathbf{1}_{\mathscr{A}}+q^2\tilde{A}_\downarrow^+A^\downarrow,\\
\tilde{A}^\uparrow\tilde{A}_\downarrow^+&=q\tilde{A}_\downarrow^+\tilde{A}^\uparrow,\quad \tilde{A}^\downarrow\tilde{A}_\uparrow^+=qA_\uparrow^+\tilde{A}^\downarrow,
\end{aligned} \tag{V.15}$$

$$\tilde{A}^\downarrow\tilde{A}^\uparrow=q\tilde{A}^\uparrow\tilde{A}^\downarrow,\quad \tilde{A}_\uparrow^+\tilde{A}_\downarrow^+=q\tilde{A}_\downarrow^+\tilde{A}_\uparrow^+ \tag{V.16}$$

and the Clifford

$$\tilde{A}^\uparrow\tilde{A}_\uparrow^+=\mathbf{1}_{\mathscr{A}}-\tilde{A}_\uparrow^+\tilde{A}^\uparrow+(q^{-2}-1)\tilde{A}_\downarrow^+\tilde{A}^\downarrow,$$

$$\tilde{A}^{\downarrow}\tilde{A}^{+}_{\downarrow}=\mathbf{1}_{\mathscr{A}}-\tilde{A}^{+}_{\downarrow}\tilde{A}^{\downarrow},\quad \tilde{A}^{\uparrow}\tilde{A}^{+}_{\downarrow}=-q^{-1}\tilde{A}^{+}_{\downarrow}\tilde{A}^{\uparrow},\tag{V.17}$$

$$\tilde{A}^{\downarrow}\tilde{A}^{+}_{\uparrow}=-q^{-1}\tilde{A}^{+}_{\uparrow}\tilde{A}^{\downarrow},$$

$$\tilde{A}^{\downarrow}\tilde{A}^{\uparrow}=-q^{-1}\tilde{A}^{\uparrow}\tilde{A}^{\downarrow},\quad \tilde{A}^{\uparrow}\tilde{A}^{\uparrow}=0,\quad \tilde{A}^{\downarrow}\tilde{A}^{\downarrow}=0;\tag{V.18}$$

$$\tilde{A}^{+}_{\uparrow}\tilde{A}^{+}_{\downarrow}=-q^{-1}\tilde{A}^{+}_{\downarrow}\tilde{A}^{+}_{\uparrow},\quad \tilde{A}^{+}_{\uparrow}\tilde{A}^{+}_{\uparrow}=0,\quad \tilde{A}^{+}_{\downarrow}\tilde{A}^{+}_{\downarrow}=0.\tag{V.19}$$

Now we try to construct the A^{+}_{i},A^{i}. The ansatz of Proposition 4 is equivalent to

$$A^{+}_{i}=a^{+}_{l}\sigma(\mathscr{F}^{-1(2)})\rho(\mathscr{F}^{-1(1)})^{l}_{i}f(n),$$
$$A^{i}=g(n)\rho(\mathscr{F}^{(1)})^{i}_{l}\sigma(\mathscr{F}^{(2)})a^{l},\tag{V.20}$$

with the same $\mathscr{F}$ and two invertible functions f,g. {In fact, in $T=K[\mathscr{C}\otimes\mathbf{1},\mathbf{1}\otimes\mathscr{C},\Delta(\mathscr{C})]$ the dependence on the last argument results only in a numerical factor, because of Eqs. (II.20) and (II.24). The $\mathscr{C}\otimes\mathbf{1}$ and $\mathbf{1}\otimes\mathscr{C}$ dependences can be replaced by the n dependence, since it is easy to prove that $\sigma(\mathscr{C})=(n/2)(n/2+1)$; the latter can be concentrated either at the left or at the right of a^{i},a^{+}_{j}, upon using the commutation relations $[n,\,a^{+}_{i}]=a^{+}_{i}$, $[n,\,a^{i}]=-a^{i}$.} The product fg is determined by requiring that the commutation relation between $N:=A^{+}_{i}A^{i}$ and A^{i},A^{+}_{j} are the same as those between $\tilde{N}:=\tilde{A}^{+}_{i}\tilde{A}^{i}$ and $\tilde{A}^{i}$,$\tilde{A}^{+}_{j}$:

$$NA^{+}_{i}=A^{+}_{i}(1+q^{\pm 2}N),\quad NA^{i}=A^{i}q^{\mp 2}(N-1);\tag{V.21}$$

one easily finds

$$gf(x)=\sqrt{\frac{(x+1)_{q^{\pm 2}}}{(x+1)}}.\tag{V.22}$$

If $q\in\mathbf{R}^{+}$ and we wish that $A^{+}_{i}=(A^{i})^{\star}$, we need to choose

$$g(x)=f(x)=\sqrt{\frac{(x+1)_{q^{\pm 2}}}{(x+1)}}.\tag{V.23}$$

This is nothing else but the already encountered function (I.3)[6] needed to transform the classical creation/annihilation operators in one dimension a^{+},a into the quantum ones A^{+},A.

If we are interested in Fock space representations, $A^{+}_{i}|0\rangle=a^{+}_{i}|0\rangle$, stating that the quantum and classical one-particle states coincide, is automatically satisfied, because $f(0)=1$.

Now one can express the rhs of (V.20) thoroughly in terms of a^{i},a^{+}_{j}. It is convenient to introduce the up–down ''number of particle'' operators $n^{\uparrow}$,$n^{\downarrow}$ ($n^{\uparrow}+n^{\downarrow}=n$) by

$$n^{\uparrow}:=a^{+}_{\uparrow}a^{\uparrow},\quad n^{\downarrow}:=a^{+}_{\downarrow}a^{\downarrow}.\tag{V.24}$$

Using Eqs. (V.6), (V.7), (V.8), and (V.20), it is easy algebra to prove the following. (In the fermionic case one has to fully exploit the nilpotency of a^{i},a^{+}_{j}.)

Proposition 5: Equations (V.20) amount, in the Weyl case, to

$$A^{+}_{\uparrow}=\sqrt{\frac{(n^{\uparrow})_{q^2}}{n^{\uparrow}}}\,q^{n^{\downarrow}}a^{+}_{\uparrow},\quad A^{+}_{\downarrow}=\sqrt{\frac{(n^{\downarrow})_{q^2}}{n^{\downarrow}}}\,a^{+}_{\downarrow},$$
$$A^{\uparrow}=a^{\uparrow}\sqrt{\frac{(n^{\uparrow})_{q^2}}{n^{\uparrow}}}\,q^{n^{\downarrow}},\quad A^{\downarrow}=a^{\downarrow}\sqrt{\frac{(n^{\downarrow})_{q^2}}{n^{\downarrow}}},\tag{V.25}$$

and, in the Clifford one, to

$$A^{+}_{\uparrow}=q^{-n^{\downarrow}}a^{+}_{\uparrow},\quad A^{+}_{\downarrow}=a^{+}_{\downarrow},\quad A^{\uparrow}=a^{\uparrow}q^{-n^{\downarrow}},\quad A^{\downarrow}=a^{\downarrow}.\tag{V.26}$$

We are ready for the main theorem of this section (the proof is a straightforward computation).

Theorem 2: *The elements $A^iA_j^+\in\mathscr{A}$ ($\mathscr{A}$ Weyl or Clifford) defined in formulas (V.20) and (V.23) satisfy the QCR (V.12) and (V.11).*

Let us compare the map (V.25) (for the Weyl algebra) with the one found in Ref. 17. That map, in our notation, would read

$$A_\uparrow^+=q^{n^\uparrow}a_\uparrow^+,\quad A_\downarrow^+=a_\downarrow^+,$$
$$A^\uparrow=a^\uparrow\,\frac{(n^\uparrow)_{q^2}}{n^\uparrow}\,q^{n^\downarrow},\quad A^\downarrow=a^\downarrow\,\frac{(n^\downarrow)_{q^2}}{n^\downarrow} \tag{V.27}$$

[clearly, it is not compatible with the $*$-structure $A_{i,\alpha}^+=(A^{i,\alpha})^\star$]. It is straightforward to check that the element $\alpha\in\mathscr{A}[[h]]$ needed to transform A_i^+,A^i into $A_{i,\alpha}^+,A^{i,\alpha}$ [formulas (III.15)] is

$$\alpha:=\sqrt{\frac{\Gamma(n^\uparrow+1)\Gamma(n^\downarrow+1)}{\Gamma_{q^2}(n^\uparrow+1)\Gamma_{q^2}(n^\downarrow+1)}}, \tag{V.28}$$

where Γ is the Euler Γ-function and Γ_{q^2} its q-deformation[27] satisfying the property

$$\Gamma_q(a+1)=(a)_q\Gamma_q(a). \tag{V.29}$$

The corresponding realization of $\tilde{\triangleright}_h$ is obtained through relation (III.14).

VI. REPRESENTATION THEORY

In this section we compare representations of $\mathscr{A}$ with representations of $\mathscr{A}_h$ and investigate whether the deforming maps found in the preceding sections can be interpreted as intertwiners between them.

We start with a general remark. At least perturbatively in h, we expect that, for any given representation π of $\mathscr{A}$ on some space V, the objects $\pi(A_{i,\alpha}^+),\pi(A^{i,\alpha})$ are well-defined operators on V, since $A_{i,\alpha}^+=a_i^++O(h)$, $A^{i,\alpha}=a^i+O(h)$; consequently, $\pi\circ f_\alpha$ is a representation of $\mathscr{A}_h$ on V. Let us prove this statement more rigorously for some specific kind of representations.

If **g** is compact, we can choose ρ to be a unitary representation; the **g**-covariant CCR will admit the $*$-relations (II.25). Assuming the latter, the Stone–Von Neumann theorem (or its Clifford counterpart) applies: there exists a ''$*$-representation'' π of Eqs. (II.19) and (II.25) on a separable Hilbert space $\mathscr{H}$, i.e., a π fulfilling the following properties:

(1) $\pi(a^i),\pi(a_i^+)$ are closed;
(2) $\pi(a_i^+)\subset[\pi(a^i)]^\dagger$;
(3) there exists a dense linear subset $\mathscr{D}\subset\mathscr{H}$ contained in the domain of the product of any two operators $\pi(a^i),\pi(a_i^+)$;
(4) the CCR (II.19) hold on $\mathscr{D}$;
(5) the non-negative definite operator $\pi(n):=\pi(a_i^+a^i)$ is essentially self-adjoint on $\mathscr{D}$.

Moreover, the irreducible components $(\pi_l,\mathscr{H}_l)$ of $(\pi,\mathscr{H})$ are, up to a unitary transformation $U_l:\mathscr{H}\to\mathscr{H}$, Fock space representations, i.e., there exist ''ground states'' $|0\rangle_l\in\mathscr{H}_l$:

$$\pi(a^i)|0\rangle_l=\mathbf{0}. \tag{VI.1}$$

One can choose the dense set $\mathscr{D}$ as the linear span of all the analytic vectors of the form $\pi(a_{i_1}^+\cdots a_{i_k}^+)|0\rangle_l$.

In the case under consideration, choosing $\mathscr{F}$ unitary [so that relation (III.11) holds] and setting $\Pi:=\pi\circ f$, one can easily realize that Π is a $*$-representation of the QCR [(I.8) or (V.10)–(V.12)] on $\mathscr{H}$, i.e., it fulfills conditions analogous to (1)–(5), with $\tilde{A}^i,\tilde{A}_i^+$ in the place of a^i,a_i^+.

The reason is essentially the following. Assume first that π and ρ are irreducible, and let $\mathscr{H}^{(k)}:=\mathrm{Span}_{\mathbf{C}}\{\pi(a_{i_1}^+\cdots a_{i_k}^+)|0\rangle\}$; $\mathscr{H}^{(k)}$ is a finite-dimensional eigenspace of $\pi(n)$, because ρ is finite-dimensional and, therefore, $\mathscr{A}$ has a finite number of generators. From Eqs. (III.17) and the

relation $[\sigma(U\mathbf{g}),\, n]=0$ it follows that $\pi(A^i),\pi(A_i^+)$ differ from $\pi(a^i),\pi(a_i^+)$ just by operators mapping each $\mathscr{H}^{(k)}$ into itself. [Incidentally, from relation (II.1) it follows in particular $\pi(A^i)|0\rangle=\mathbf{0}$ and $\pi(A_i^+)|0\rangle=\pi(a_i^+)|0\rangle$.] Therefore all the properties (1)–(5) are inherited by $\pi(A^i),\pi(A_i^+)$ as well. The same result holds for $\Pi_\alpha:=\pi\circ f_\alpha$ if $\alpha^\star=\alpha^{-1}$, because $\pi(\alpha)$ can be absorbed in the unitary transformation U. Finally, the result extends by linearity and orthogonality also to the case that π and/or ρ are reducible.

Summing up, f_α is an intertwiner between the $*$-representation of the CCR and a $*$-representation of the QCR. In other words, we can represent the objects $\tilde{A}^i,\tilde{A}_i^+$ [fulfilling the QCR and the $*$-relations (III.11)] as composite operators acting on the Fock space $\mathscr{H}$, as anticipated in the Introduction. The latter can be used to describe ordinary bosons and fermions. This disproves the quite common belief that non-cocommutative Hopf algebra symmetries are necessarily incompatible with ordinary Bose and Fermi statistics.

Let us analyze now a different situation. Let $\mathbf{g}=\mathrm{sl}(2)$, $\mathscr{A}$ be the corresponding two-dimensional Weyl algebra, and (π,V) be the Fock space representation of the latter (with ground state $|0\rangle$). Let $\mathscr{A}_h$ be the deformation of $\mathscr{A}$ considered in Sec. V with $q^2=e^{2h}$ a root of unity, i.e., $q^{2p}=1$, $q^{2k}\neq 1$ with $p,k\in\mathbf{N}$ and $k<p$. It is easy to realize that $\pi[(n_i)_{q^2}(a_i^+)^{mp}(a_{-i}^+)^l]|0\rangle=\mathbf{0}$ for $l,m=0,1,...$; consequently, $A^\uparrow,A^\downarrow$ annihilate all the vectors of the form $|m,n\rangle := \pi[(a_\uparrow^+)^{mp}(a_\downarrow^+)^{np}|0\rangle$ $(n,m=0,1,...)$. Although (π,V) was irreducile as a representation of $\mathscr{A}$, it is reducile as a representation of the subalgebra of $\mathscr{A}$ generated by A^i,A_j^+; the irreducible components $V_{m,n}$ are isomorphic and p^2-dimensional, and are obtained by applying A_i^+'s to the cyclic vectors $|m,n\rangle$. $V_{m,n}$ is also a (reducible) representation of $\mathscr{U}_h\mathrm{sl}(2)$ and may be called (with an abuse of terminology, since we have not introduced any spatial degrees of freedom) an ''anyonic space.'' Thus, f can be seen as an intertwiner from the classical bosonic Fock representation onto a direct sum of ''anyonic'' representations of $\mathscr{A}_h$.

Whenever some class P of representations of $\mathscr{A}_h$ is ''larger'' than the corresponding class p of representations of the CCR, then f_α^{-1} can of course be well defined (as an intertwiner between P and p) only on some proper subset of P; on its complement it must be singular.

It is instructive to see whether and how this phenomenon occurs in some concrete example, e.g., for the class of $*$-representations considered above. Both a deformed algebra $\mathscr{A}_h$ covariant under some triangular Hopf algebra H_h of the type (II.8) (with $\mathscr{F}$ unitary) and the deformed Clifford algebra of Sec. V are not good for this purpose, because P is as large as p. {For the latter algebra this was shown in Ref. 24. For the former this can be understood as follows. In the present case $\mathscr{R}=\mathscr{F}^{-2}$. Given any representation ρ of $U\mathbf{g}$ and the corresponding representation $\tilde{\rho}$ of $U_h\mathbf{g}$, let us choose a basis of eigenvectors $|l\rangle$ of the Cartan subalgebra. In this basis $R:=(\tilde{\rho}\otimes\tilde{\rho})\mathscr{R}$ will be diagonal, and, in particular, it will be $R_{ll}^{ll}=1$ as a consequence of the antisymmetry of $\omega_{ij}(h_i\otimes h_j)$. The QCR (I.8) will imply in particular

$$\tilde{A}^i\tilde{A}_i^+=\mathbf{1}_{\mathscr{A}}\pm\tilde{A}_i^+\tilde{A}^i \quad (\text{no sum over } i), \tag{VI.2}$$

and, setting $\tilde{N}_i := A_i^+\tilde{A}^i$ (no sum over i),

$$[\tilde{N}_i,\,\tilde{N}_j]=0,\quad [\tilde{N}_i,\,\tilde{A}_j^+]=\delta_{ij}\tilde{A}_j^+,\quad [\tilde{N}_i,\,\tilde{A}^j]=-\delta_i^j\tilde{A}^j. \tag{VI.3}$$

Let us denote by $\mathscr{A}_h^{(i)}$ the subalgebra generated by $\tilde{A}^i,\tilde{A}_i^+$. Each $\mathscr{A}_h^{(i)}$ separately is isomorphic to a classical one-dimensional Weyl/Clifford algebra, and therefore admits, up to unitary equivalences, a unique $*$-representation in the form of a Fock representation with level N_i. Because of the relations (VI.3), we can choose $\{N_i\}_{i\in I}$ as a complete set of commuting observables, where the uniqueness, up to unitary equivalences, of the $*$-representation of $\mathscr{A}_h$ follows immediately.} On the contrary, it was proved in Ref. 3 that there are many unitarily nonequivalent $*$-representations on separable Hilbert spaces of the $\mathscr{U}_h\mathrm{su}(2)$-covariant $(q\in\mathbf{R}^+)$ deformed Weyl algebra. [Incidentally, even one-dimensional deformed Heisenberg algebras may have more unitarily inequivalent representations.[28] Moreover, within each representation one has still some freedom in the ''physical'' interpretation of the observables, e.g., what are the ''right'' momentum/position observables (see, e.g., Ref. 29).]

Sticking to the case $0<q<1$, one can parametrize the Woronowicz–Pusz[3] unitarily inequivalent irreducible $*$-representations of this algebra by three parameters E, r, s, where $q^2<E<1$, and r, s are non-negative integers with $r+s\leqslant 2$. We shall denote the corresponding Hilbert space by $\mathscr{H}_{E,r,s}$. We divide them in the following classes for clarity:

(1) In the representation $s=2$ (and $r=0$, the value E is irrelevant) one parametrizes the vectors of an orthonormal basis of $\mathscr{H}_{E,0,2}$ by $\{|q^{2m_1}/(q^2-1),q^{2m_2}/(q^2-1)\rangle\}$.
(2) In the representations with $s=1=r$, one parametrizes the vectors of an orthonormal basis of $\mathscr{H}_{E,1,1}$ by $\{|q^{2n_1}E,q^{2m_1}/(q^2-1)\rangle\}$.
(3) In the representations with $s=1$ and $r=0$, one parametrizes the vectors of an orthonormal basis of $\mathscr{H}_{E,0,1}$ by $\{|0,q^{2m_1}/(q^2-1)\rangle\}$ (the value E is irrelevant).
(4) In the representations with $s=0$ and $r=2$, one parametrizes the vectors of an orthonormal basis of $\mathscr{H}_{E,2,0}$ by $\{|q^{2n_1}E,q^{2n_2}E\rangle\}$.
(5) Finally, in the representations with $s=0$ and $r=1$, one parametrizes the vectors of an orthonormal basis of $\mathscr{H}_{E,1,0}$ by $\{|q^{2n_1}E,0\rangle\}$.

Here n_1,n_2; m_1,m_2 denote integers, with $m_1\geqslant m_2\geqslant 0$, $n_1<n_2$. Only representations (1) and (3) have a ground state; but representation (3) is degenerate.

In the rest of this section we drop the symbols Π to avoid a too heavy notation. On the vectors $|\eta_\uparrow,\eta_\downarrow\rangle$ of the above basis of any $\mathscr{H}_{E,r,s}$, $\tilde{A}^i,\tilde{A}_i^+$, $i=\uparrow,\downarrow$, are defined (modulo a possible but here irrelevant phase in the case $r+s<2$) by

$$\tilde{A}^\uparrow|\eta,\eta_\downarrow\rangle=\sqrt{\eta-\eta_\downarrow}|q^{-2}\eta,\eta_\downarrow\rangle,\quad \tilde{A}^\downarrow|\eta,\eta_\downarrow\rangle=\sqrt{\eta_\downarrow-\frac{1}{q^2-1}}\,|q^{-2}\eta,q^{-2}\eta_\downarrow\rangle,$$
$$\tilde{A}_\uparrow^+|\eta,\eta_\downarrow\rangle=\sqrt{q^2\eta-\eta_\downarrow}|q^2\eta,\eta_\downarrow\rangle,\quad \tilde{A}_\downarrow^+|\eta,\eta_\downarrow\rangle=\sqrt{q^2n_\downarrow-\frac{1}{q^2-1}}\,|q^2\eta,q^2\eta_\downarrow\rangle. \tag{VI.4}$$

Hence, setting $\tilde{N}^\uparrow:=\tilde{A}_\uparrow^+\tilde{A}^\uparrow$, $\tilde{N}^\downarrow:=\tilde{A}_\downarrow^+\tilde{A}^\downarrow$, and $\tilde{N}:=\tilde{N}^\uparrow+\tilde{N}^\downarrow$, we find that $|\eta,\eta_\downarrow\rangle$ are eigenvectors of the following operators:

$$\tilde{N}^\uparrow|\eta,\eta_\downarrow\rangle=(\eta-\eta_\downarrow)|\eta,\eta_\downarrow\rangle,$$
$$\tilde{N}^\downarrow|\eta,\eta_\downarrow\rangle=\left(\eta_\downarrow-\frac{1}{q^2-1}\right)|\eta,\eta_\downarrow\rangle,$$
$$[1+(q^2-1)\tilde{N}^\downarrow]|\eta,\eta_\downarrow\rangle=(q^2-1)\eta_\downarrow|\eta,\eta_\downarrow\rangle, \tag{VI.5}$$
$$[1+(q^2-1)\tilde{N}]|\eta,\eta_\downarrow\rangle=(q^2-1)\eta|\eta,\eta_\downarrow\rangle.$$

Formally, the inverse of the transformation (V.25) reads

$$f^{-1}(\tilde{A}_\downarrow^+):=\tilde{a}_\downarrow^+=\sqrt{\frac{\log[1+(q^2-1)\tilde{N}^\downarrow]}{2\tilde{N}^\downarrow\log q}}\,\tilde{A}_\downarrow^+,$$
$$f^{-1}(\tilde{A}_\uparrow^+):=\tilde{a}_\uparrow^+=\sqrt{\frac{1+(q^2-1)\tilde{N}^\downarrow}{2\tilde{N}^\uparrow\log q}\log\left[\frac{1+(q^2-1)\tilde{N}}{1+(q^2-1)\tilde{N}^\downarrow}\right]}\tilde{A}_\uparrow^+, \tag{VI.6}$$
$$f^{-1}(\tilde{A}^\downarrow):=\tilde{a}^\downarrow=\tilde{A}^\downarrow\sqrt{\frac{\log[1+(q^2-1)\tilde{N}^\downarrow]}{2\tilde{N}^\downarrow\log q}},$$
$$f^{-1}(\tilde{A}^\uparrow):=\tilde{a}^\uparrow=\tilde{A}^\uparrow\sqrt{\frac{1+(q^2-1)\tilde{N}^\downarrow}{2\tilde{N}^\uparrow\log q}\log\left[\frac{1+(q^2-1)\tilde{N}}{1+(q^2-1)\tilde{N}^\downarrow}\right]}.$$

A glance at formula (VI.5) shows that the arguments of both logarithms in the inverse transformation (VI.6) are positive definite on representation (1), whereas at least one of the arguments of the two logarithms is negative on any representation (2), (3), (4), or (5) [in the representation (3) the other argument vanishes]. This makes f^{-1} ill-defined on all representations (2), (3), (4), or (5).

We conclude that representation (1) is the one intertwined by f^{-1} to the standard bosonic Fock representation of the su(2)-covariant Weyl algebra $\mathscr{A}$, whereas the representations of the classes (2), (4), and (5) have no classical analog, and representation (3) reduces to the representation of a one-dimensional Weyl algebra.

ACKNOWLEDGMENTS

I would like to thank P. Schupp for fruitful discussions on the subject. I am grateful to J. Wess for the warm hospitality at his institute. This work was financially supported by a TMR fellowship granted by the European Commission, Dir. Gen. XII for Science, Research and Development, under Contract No. ERBFMBICT960921.

[1] E. Abe, *Hopf Algebras*, Cambridge Tracts in Mathematics 74 (Cambridge U.P., Cambridge, 1980).
[2] V. G. Drinfeld, ''Quantum Groups,'' edited by A. Gleason, in Proceedings of the International Congress of Mathematicians, Berkeley, 1986 (American Mathematical Society, Providence, 1987), Vol. 1, p. 798.
[3] W. Pusz and S. L. Woronowicz, Rep. Math. Phys. **27**, 231 (1989).
[4] J. Wess and B. Zumino, Nucl. Phys. B, Proc. Suppl. **18**, 302 (1991).
[5] L. C. Biedenharn, J. Phys. A **22**, L873 (1989); A. J. Macfarlane, J. Phys. A **22**, 4581 (1989).
[6] S. Vokos and C. Zachos, Mod. Phys. Lett. A **9**, 1 (1994) and references therein; see also J. Seifert, Ph.D. thesis, Munich University (1996) (in German).
[7] T. L. Curtright and C. K. Zachos, Phys. Lett. B **243**, 237 (1990).
[8] T. L. Curtright, G. I. Ghandour, and C. K. Zachos, J. Math. Phys. **32**, 676 (1991).
[9] G. Fiore, ''Drinfel'd Twist and q-deforming Maps for Lie Groups Covariant Heisenberg Algebras,'' Preprint LMU-TPW 97-07 and e-print q-alg/9708017.
[10] M. Gerstenhaber, Ann. Math. **79**, 59 (1964).
[11] F. du Cloux, Asterisque (Soc. Math. France) **124–125**, 129 (1985).
[12] M. Pillin, Commun. Math. Phys. **180**, 23 (1996).
[13] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. (N.Y.) **111**, 61 (1978); **111**, 111 (1978).
[14] V. G. Drinfeld, Dokl. Akad. Nauk SSSR **273**, 531 (1983); Translation: Soviet Math. Dokl. **28**, (1983).
[15] N. Yu. Reshetikhin Lett. Math. Phys. **20**, 331 (1990).
[16] V. G. Drinfeld, Leningrad Math. J. **1**, 1419 (1990).
[17] O. Ogievetsky, Lett. Math. Phys. **24**, 245 (1992).
[18] G. Fiore and P. Schupp, Nucl. Phys. B **470**, 211 (1996); ''Statistics and Quantum Group Symmetries, e-print hep-th/9605133, Banach Center Publications, Vol. 40, Inst. of Mathematics, Polish Academy of Sciences, Warszawa, edited by P. Budzyski, W. Pusz, and S. Zakrweski.
[19] G. Fiore, ''On Bose-Fermi Statistics, Quantum Group Symmetry, and Second Quantization,'' Preprint LMU-TPW 96-17 and e-print hep-th/9611144. To appear in the proceedings of ''Group21.''
[20] C. Ohn, Lett. Math. Phys. **166**, 63 (1994).
[21] L. D. Faddeev, N. Y. Reshetikhin, and L. A. Takhtajan, Algebra i Analysis **1**, 178 (1989); translation: Leningrad Math. J. **1**, 193 (1990).
[22] B. Jurco, Commun. Math. Phys. **166**, 63 (1994).
[23] See, e.g., L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics, Encyclopedia of Mathematics and its Applications*, Vol. 8 (Addison–Wesley Reading, MA, 1981); V. Manko, G. Marmo, P. Vitale, and F. Zaccaria, Int. J. Mod. Phys. A **9**, 5541 (1994).
[24] W. Pusz, Rep. Math. Phys. **27**, 349 (1989).
[25] S. Majid, *Foundations of Quantum Groups* (Cambridge U.P., Cambridge, 1995).
[26] T. Hayashi, Commun. Math. Phys. **127**, 129 (1990).
[27] G. Gasper and M. Rahman, *Basic Hypergeometric Series, Encyclopedia of Mathematics and its Applications*, Vol. 35 (Cambridge U.P., Cambridge, 1990).
[28] See, for instance, A. Lorek, A. Ruffing, and J. Wess, ''A q-deformation of the Harmonic Oscillator,'' preprint MPI-PHT-96-26 and e-print hep-th/9605161, and references therein.
[29] A. Kempf, J. Math. Phys. **35**, 4483 (1994).

$d=2$, $N=2$ superconformally covariant operators and super W-algebras

François Gieres[a)]
Institut de Physique Nucléaire de Lyon, IN2P3/CNRS, Université Claude Bernard, 43, boulevard du 11 novembre 1918, F-69622-Villeurbanne Cedex, France, and Max-Planck-Institut für Physik—Werner Heisenberg Institut—Föhringer Ring 6, 80805 Munich, Germany

Stéphane Gourmelen
Institut de Physique Nucléaire de Lyon, IN2P3/CNRS, Université Claude Bernard, 43, boulevard du 11 novembre 1918, F-69622-Villeurbanne Cedex, France

(Received 27 August 1997; accepted for publication 3 March 1998)

We construct and classify superconformally covariant differential operators defined on $N=2$ super Riemann surfaces. By contrast to the $N=1$ theory, these operators give rise to partial rather than ordinary differential equations which leads to novel features for their matrix representation. The latter is applied to the derivation of $N=2$ super W-algebras in terms of $N=2$ superfields. © *1998 American Institute of Physics.* [S0022-2488(98)01606-5]

I. INTRODUCTION

In the context of two-dimensional conformal field theories and integrable models, an important role is played by conformally covariant differential operators in one complex variable:[1] these linear, n-th order operators have the general form $L^{(n)}=\partial^n+a_1^{(n)}\partial^{n-1}+\cdots+a_n^{(n)}$ where $\partial \equiv \partial/\partial z$ denotes the derivative with respect to local complex coordinates z parametrizing a compact Riemann surface and conformal covariance means that the coefficient functions $a_k^{(n)}(z)$ transform in such a way under local conformal changes of coordinates $z\to z'(z)$ that $L^{(n)}$ maps conformal fields to conformal fields. Any such differential operator can be reparametrized in the following way in terms of a projective connection and $n-2$ conformal fields:

$$L^{(n)}=L_n+M_{w_3}^{(n)}+\cdots+M_{w_n}^{(n)}.$$

Here, L_n is the so-called *Bol operator* of order n,[1] i.e., the conformally covariant version of the differential operator $\partial^n=(\partial/\partial z)^n$. The simplest examples are given by

$$L_0=1, \qquad L_1=\partial, \qquad L_2=\partial^2+\tfrac{1}{2}R, \tag{1}$$

$$L_3=\partial^3+2R\partial+(\partial R),$$

$$L_4=\partial^4+5R\partial^2+5(\partial R)\partial+\tfrac{3}{2}[(\partial^2 R)+\tfrac{3}{2}R^2],$$

where $R=R_{zz}(z)$ denotes a projective connection, i.e., a locally holomorphic field transforming with the Schwarzian derivative under a conformal change of coordinates $z\to z'(z)$:

$$R'(z')=(\partial z')^{-2}[R(z)-S(z',z)], \quad \text{where } S(z',z)=\partial^2 \ln \partial z'-\tfrac{1}{2}(\partial \ln \partial z')^2. \tag{2}$$

Each of the lower order operators $M_{w_k}^{(n)}$ $(k=3,\ldots,n)$ is also conformally covariant and depends linearly on a conformal field w_k of weight k, e.g.,

[a)]Alexander von Humboldt Fellow on leave of absence from the Université Claude Bernard.

 © 1998 American Institute of Physics

$$M^{(n)}_{w_n}=w_n, \quad M^{(n)}_{w_{n-1}}=w_{n-1}\partial+\tfrac{1}{2}(\partial w_{n-1}),$$
$$M^{(n)}_{w_{n-2}}=w_{n-2}\left[\partial^2-\frac{1-n}{2}R\right]+(\partial w_{n-2})\partial+\frac{n-1}{2(2n-3)}[\partial^2-(n-2)R]w_{n-2}. \tag{3}$$

The operators L_n and $M^{(n)}_{w_k}$ can both be obtained from Möbius covariant operators and admit a matrix representation which is related to the principal embedding of $sl(2)$ into $sl(n)$.[2] As we already mentioned, all of these covariant operators admit numerous applications[1] to integrable models (Lax pairs, Poisson operators) and to conformal models (conformal Ward identity, diffeomorphism anomaly, Krichever-Novikov algebra, Wess-Zumino-Witten model), in particular to the theory of W-algebras. Their $N=1$ supersymmetric generalization has been worked out in references[1,3] (see also Ref. 4) and the present paper is devoted to their $N=2$ supersymmetric generalization.

There is a new feature in $N=2$ superspace geometry which makes this theory considerably richer and more complicated: the ''square root'' of the translation generator ∂ is not given by a single odd operator as in $N=1$ supersymmetry ($D^2=\partial$), but it involves two odd operators ($\{D,\overline{D}\}=\partial$). Therefore, one has to deal with partial differential equations (involving D and $\overline{D}$) rather than ordinary differential equations (only involving D). Another aspect of the algebra $\{D,\overline{D}\}=\partial$ consists of the fact that it introduces a $U(1)$ symmetry into the theory: after projection from the super Riemann surface to the underlying ordinary Riemann surface, one thereby recovers $U(1)$-transformations in addition to the familiar conformal transformations. Henceforth, the Bol operators (1) acting on $U(1)$-neutral fields are to be generalized to conformally covariant operators acting on $U(1)$-charged fields. The latter as well as the original operators (1) arise from different types of $N=2$ super Bol operators which are constructed and classified in section IV. In particular, one is led to introduce operators which relate the chiral and anti-chiral subspaces of superconformal fields (section IV B). It is only for these so-called ''sandwich operators'' that we will be able to give a matrix representation (section VI). The latter provides a connection field with values in the Lie superalgebra $sl(n+1|n)$ and the associated $N=2$ super W_n-algebra[5] can be constructed by imposing zero curvature conditions on this connection (section VII).

To keep supersymmetry manifest, all considerations will be carried out in superspace while the projection to ordinary space will be indicated in each case. The passage from the $N=2$ to the $N=1$ superfield formalism will also be discussed so as to make contact with previous work which mostly uses the latter description. Applications of our results will be mentioned in the text and in the concluding remarks.

II. GEOMETRIC FRAMEWORK

In this section, we recall some definitions (following Ref. 6 and references therein) and we introduce some tools which are needed in the sequel.

A. Notation and basic relations

We will work on a compact $N=2$ super Riemann surface (SRS) locally parametrized by coordinates

$$(\mathbf{z};\overline{\mathbf{z}})\equiv(z,\theta,\overline{\theta};\overline{z},\theta^-,\overline{\theta^-})\equiv(x^{++},\theta^+,\overline{\theta^+};x^{--},\theta^-,\overline{\theta^-}), \tag{4}$$

with $z,\overline{z}$ even and $\theta,\overline{\theta},\theta^-,\overline{\theta^-}$ odd. The variables are complex and related by complex conjugation (denoted by *):

$$z^*=\overline{z}, \quad (\theta^+)^*=\theta^-, \quad (\overline{\theta^+})^*=\overline{\theta^-}.$$

As indicated in (4), we will omit the plus-indices of θ^+ and $\overline{\theta^+}$ to simplify the notation.

The canonical basis of the tangent space is defined by $(\partial,D,\overline{D};\overline{\partial},D_-,\overline{D}_-)$ with

$$\partial=\frac{\partial}{\partial z},\quad D=\frac{\partial}{\partial\theta}+\frac{1}{2}\overline{\theta}\partial,\quad \overline{D}=\frac{\partial}{\partial\overline{\theta}}+\frac{1}{2}\theta\partial$$
$$\overline{\partial}=\frac{\partial}{\partial\overline{z}},\quad D_-=\frac{\partial}{\partial\theta^-}+\frac{1}{2}\overline{\theta^-}\overline{\partial},\quad \overline{D}_-=\frac{\partial}{\partial\overline{\theta^-}}+\frac{1}{2}\theta^-\overline{\partial}. \tag{5}$$

The graded Lie brackets between these vector fields are given by

$$\{D,\overline{D}\}=\partial,\quad \{D_-,\overline{D}_-\}=\overline{\partial}, \tag{6}$$

all others brackets being zero, in particular,

$$D^2=0=\overline{D}^2,\quad (D_-)^2=0=(\overline{D}_-)^2. \tag{7}$$

For later reference, we note that this set of equations implies

$$[D,\overline{D}]^2=\partial^2,\quad [D_-,\overline{D}_-]^2=\overline{\partial}^2. \tag{8}$$

B. Superconformal transformations

By definition of the SRS, any two sets of local coordinates, say $\mathbf{z}$ and $\mathbf{z}'$, are related by a *superconformal transformation*, i.e., a smooth mapping satisfying three conditions:

(i) $\mathbf{z}'=\mathbf{z}'(\mathbf{z})\Leftrightarrow D_-\mathbf{z}'=0=\overline{D}_-\mathbf{z}'$,

$\overline{\mathbf{z}'}=\overline{\mathbf{z}}'(\overline{\mathbf{z}})\Leftrightarrow D\overline{\mathbf{z}'}=0=\overline{D}\overline{\mathbf{z}'}$;

(ii) $D\overline{\theta'}=0=\overline{D}\,\theta'$; (9)

(iii) $Dz'=\frac{1}{2}\overline{\theta'}D\theta',\quad \overline{D}z'=\frac{1}{2}\theta'\overline{D\theta'}$,

$D_-\overline{z'}=\frac{1}{2}\overline{\theta^-}'D_-\theta^-{}',\quad \overline{D}_-\overline{z'}=\frac{1}{2}\theta^-{}'\overline{D}_-\overline{\theta^-}'$.

These relations imply that D and $\overline{D}$ separately transform into themselves,

$$D'=e^{w}D,\quad \overline{D}'=e^{\overline{w}}\overline{D}, \tag{10}$$
$$\partial'=\{D',\overline{D}'\}=e^{w+\overline{w}}[\partial+(\overline{D}w)D+(D\overline{w})\overline{D}],$$

where

$$e^{-w}\equiv D\theta',\quad Dw=0,\quad e^{-\overline{w}}\equiv\overline{D\theta'},\quad \overline{Dw}=0, \tag{11}$$

and

$$e^{-w-\overline{w}}=\partial z'+\tfrac{1}{2}\,\overline{\theta}'\partial\theta'+\tfrac{1}{2}\theta'\partial\overline{\theta}'. \tag{12}$$

Note that eqs. (10), (11) imply $(D\theta')(D'\theta)=1=(\overline{D\theta'})(\overline{D}'\overline{\theta})$ and that analogous equations hold in the $\overline{z}$-sector.

C. $U(1)$ symmetry

Equation (12) determines the product of $D\theta'$ and $\overline{D\theta'}$, but not their quotient which is related to the $U(1)$ automorphism group of the $N=2$ supersymmetry algebra. To introduce $U(1)$-*transformations* parametrized by a superanalytic superfield K, we decompose eq. (12) in a symmetric way,

$$D\theta' = e^{+K/2}(\partial z' + \tfrac{1}{2}\overline{\theta'}\partial\theta' + \tfrac{1}{2}\theta'\partial\overline{\theta'})^{1/2},$$
$$\overline{D\theta'} = e^{-K/2}(\partial z' + \tfrac{1}{2}\overline{\theta'}\partial\theta' + \tfrac{1}{2}\theta'\partial\overline{\theta'})^{1/2}, \tag{13}$$

which implies $K=\ln(D\theta'/\overline{D\theta'})$.

D. Superconformal fields

In the following, we will consider *superconformal fields* $C_{p,q}(\mathbf{z},\overline{\mathbf{z}})$ transforming like

$$C'_{p,q} = e^{pw+q\overline{w}}C_{p,q} \quad (p,q\in\mathbf{Z}/2, \quad p+q\in\mathbf{Z}), \tag{14}$$

and which are supposed to have Grassmann parity $(-)^{p+q}$. The space of these fields will be denoted by $\mathcal{F}_{p,q}$. The pair (p,q) will be called the *superconformal weight* of $C_{p,q}$ while $\frac{1}{2}(p+q)$ and $\frac{1}{2}(p-q)$ will be referred to as its *conformal* and $U(1)$ *weight* (or *charge*), respectively—see eq. (18) below. Thus, $C_{p,q}$ is *neutral* with respect to $U(1)$ if $p=q$. Fields with another index structure (transforming also with $w^-,\overline{w^-}$) can be defined in an analogous way.

From the transformation properties of derivatives and fields, it follows that the following chirality conditions are superconformally covariant:

$$DC_{p,0}=0 \quad \text{or} \quad \overline{D}C_{0,p}=0. \tag{15}$$

The space of these so-called *chiral* and *anti-chiral* fields will be denoted by $\mathcal{F}^c_{p,0}$ and $\mathcal{F}^a_{0,p}$, respectively.

Note that the quadratic constraints $D\overline{D}C_{0,-1}=0$ and $\overline{D}DC_{-1,0}=0$ are also covariant.

E. Projection to component fields

A generic $N=2$ superfield admits the θ-expansion,

$$C(\mathbf{z},\overline{\mathbf{z}}) = c+\theta\psi+\overline{\theta\eta}+\theta\overline{\theta}d+\theta^-[\gamma+\theta e+\overline{\theta}f+\theta\overline{\theta}\delta]+\overline{\theta}^-[\epsilon+\theta g+\overline{\theta}h+\theta\overline{\theta}\tau]$$
$$+\theta^-\overline{\theta}^-[m+\theta\varphi+\overline{\theta\chi}+\theta\overline{\theta}n] \tag{16}$$

where the component fields $c,\ldots,n$ depend on z and $\overline{z}$. Equivalently, these space-time fields can be introduced by means of projection,

$$C|=c, \quad DC|=\psi, \quad \overline{D}C|=\overline{\eta}, \quad [D,\overline{D}]C|=-2d,\ldots,$$

where the bar denotes the projection onto the lowest component of the corresponding superfield.

Conformal and $U(1)$-transformations are related to the expressions

$$(w+\overline{w})| = -\ln\,\partial z',$$
$$(w-\overline{w})| = -\ln\left|\frac{D\theta'}{\overline{D\theta'}}\right| = -K|\equiv -k. \tag{17}$$

Here and in all following space-time equations, z' stands for the lowest component of the superfield $z'(\mathbf{z})$ and the *fermionic contributions are never spelled out.*

If $C=C_{p,q}$ belongs to $\mathcal{F}_{p,q}$, then its transformation law (14) can be projected down to ordinary space by taking into account eqs. (13), (17):

$$c'_{p,q} = (\partial z')^{-(1/2)(p+q)}e^{-(1/2)(p-q)k}c_{p,q}. \tag{18}$$

Obviously, $\frac{1}{2}(p+q)$ [resp., $\frac{1}{2}(p-q)$] is the conformal [resp., $U(1)$] weight of the ordinary conformal field $c_{p,q}(z,\overline{z})$.

Whereas $C_{p,q}|$ transforms homogeneously under conformal and $U(1)$-transformations, the (bosonic) space-time field $[D,\overline{D}]C_{p,q}|$ does not unless $p=0=q$. In fact, $[D,\overline{D}]C_{p,q}$ is not a superconformal field. The remedy to this problem consists of adding an appropriate term to the projection operator $[D,\overline{D}]$ which eliminates the unwanted contribution in the transformation law of $[D,\overline{D}]C_{p,q}|$—see section III C.

F. From $N=2$ to $N=1$ formalism

Since $N=2$ theories are often formulated in terms of $N=1$ superfields, we will outline how the latter are extracted from $N=2$ superfields.

A generic $N=2$ superfield (16) can be written in terms of the odd coordinates,

$$\theta_1=(\theta+\overline{\theta})/2, \quad \theta_2=(\theta-\overline{\theta})/2, \tag{19}$$

and the complex conjugate variables $\overline{\theta_1}=(\theta_1)^*, \overline{\theta_2}=(\theta_2)^*$ which give rise to the $N=1$ derivative

$$D_1\equiv\frac{\partial}{\partial\theta_1}+\theta_1\partial=D+\overline{D}, \quad (D_1)^2=\partial \quad \text{(and c.c.)}. \tag{20}$$

Indeed, one finds

$$C(z,\overline{z},\theta_1,\overline{\theta_1},\theta_2,\overline{\theta_2})=a+\theta_2\alpha+\overline{\theta_2}\beta+\theta_2\overline{\theta_2}b, \tag{21}$$

where $a,b(\alpha,\beta)$ represent even (odd) $N=1$ superfields depending on the variables $z,\overline{z},\theta_1,\overline{\theta_1}$.

III. DERIVATIVES, CONNECTIONS AND COVARIANT OPERATORS

Along the lines of the $N=1$ theory,[1] we now introduce the necessary ingredients for the construction and description of covariant operators.

A. Schwarzian derivative

The *super Schwarzian derivative*[7,1] associated to a superconformal change of coordinates $\mathbf{z}\to\mathbf{z}'(\mathbf{z})$ is defined by

$$\begin{aligned}-\mathcal{S}(\mathbf{z}',\mathbf{z})&=2e^{-(1/2)(w+\overline{w})}[D,\overline{D}]e^{(1/2)(w+\overline{w})}\\&=[D,\overline{D}](w+\overline{w})+D(w+\overline{w})\overline{D}(w+\overline{w})\\&=\partial(w-\overline{w})+(D\overline{w})(\overline{D}w)\\&=\frac{\partial\overline{D\theta'}}{\overline{D\theta'}}-\frac{\partial D\theta'}{D\theta'}+\frac{\partial\overline{\theta'}}{\overline{D\theta'}}\frac{\partial\theta'}{D\theta'}.\end{aligned} \tag{22}$$

It satisfies the ''chain rule,''

$$\mathcal{S}(\mathbf{z}'',\mathbf{z})=e^{-(w+\overline{w})}\mathcal{S}(\mathbf{z}'',\mathbf{z}')+\mathcal{S}(\mathbf{z}',\mathbf{z}). \tag{23}$$

From eqs. (22), (12), we can extract the bosonic component fields of $\mathcal{S}$:

$$\begin{aligned}\mathcal{S}(\mathbf{z}',\mathbf{z})|&=\partial k,\\[\mathcal{S}(\mathbf{z}',\mathbf{z})|&=S(z',z)+\tfrac{1}{2}(\partial k)^2.\end{aligned} \tag{24}$$

Here, $S(z',z)$ denotes the ordinary Schwarzian derivative (2): it represents the ''conformal part'' of $\mathcal{S}$ while ∂k represents its ''$U(1)$ part.''

B. Projective coordinates

A *superprojective (super Möbius) mapping* is a superconformal change of coordinates $\mathbf{Z}=(Z,\Theta,\overline{\Theta})\to\mathbf{Z}'(\mathbf{Z})$ satisfying $\mathcal{S}(\mathbf{Z}',\mathbf{Z})=0$. By virtue of (22), this condition is equivalent to the relation

$$(D_\Theta\Theta')(D_{\overline{\Theta}}\overline{\Theta}')=(cZ+d+\Theta\overline{\gamma}+\overline{\Theta}\gamma)^{-2}, \tag{25}$$

where c and d (γ and $\overline{\gamma}$) are Grassmann even (odd) constants. This condition can be solved so as to obtain the general expression of a superprojective mapping $\mathbf{Z}\to\mathbf{Z}'$: by taking into account the conditions $D_\Theta\overline{\Theta}'=0=D_{\overline{\Theta}}\Theta'$ and $D_\Theta Z'=\frac{1}{2}\overline{\Theta}'D_\Theta\Theta'$, $D_{\overline{\Theta}}Z'=\frac{1}{2}\Theta' D_{\overline{\Theta}}\overline{\Theta}'$, one finds

$$\begin{aligned}
\Theta'(\mathbf{Z})&=e^{+h}\left\{-2\frac{\alpha Z+\beta}{cZ+d}+\frac{\Theta}{cZ+d}\left[1+2\frac{\overline{\alpha}Z+\overline{\beta}}{cZ+d}\gamma\right]-\frac{\Theta\overline{\Theta}}{(cZ+d)^2}\gamma\right\},\\
\overline{\Theta}'(\mathbf{Z})&=e^{-\overline{h}}\left\{-2\frac{\overline{\alpha}Z+\overline{\beta}}{cZ+d}+\frac{\overline{\Theta}}{cZ+d}\left[1-2\frac{\alpha Z+\beta}{cZ+d}\overline{\gamma}\right]+\frac{\Theta\overline{\Theta}}{(cZ+d)^2}\overline{\gamma}\right\},\\
Z'(\mathbf{Z})&=\frac{aZ+b}{cZ+d}[1+2(\alpha\overline{\beta}+\overline{\alpha}\beta)+8\alpha\overline{\beta}\alpha\overline{\beta}]-\frac{\Theta(\overline{\alpha}Z+\overline{\beta})+\overline{\Theta}(\alpha Z+\beta)}{(cZ+d)^2}\\
&\quad\times[1+2(\alpha\overline{\beta}+\overline{\alpha}\beta)]+\Theta\overline{\Theta}\partial\left[\frac{\alpha Z+\beta}{cZ+d}\frac{\overline{\alpha}Z+\overline{\beta}}{cZ+d}\right].
\end{aligned} \tag{26}$$

Here $h,\overline{h},a,b$ ($\alpha,\beta,\overline{\alpha},\overline{\beta}$) are Grassmann even (odd) constants which satisfy the conditions

$$\begin{aligned}
\alpha d-\beta c=\gamma,\quad ad-bc=1,\\
\overline{\alpha}d-\overline{\beta}c=\overline{\gamma},\quad 2(\alpha\overline{\beta}+\overline{\alpha}\beta)=h-\overline{h}.
\end{aligned} \tag{27}$$

The lowest order component of Z' represents an ordinary projective transformation: $Z'|=(aZ|+b)(cZ|+d)^{-1}$.

A *superprojective structure* is an atlas of local coordinates for which all changes of charts $\mathbf{Z}\to\mathbf{Z}'$ are superprojective mappings.

A *quasi-primary superfield* $\mathcal{C}_{p,q}(\mathbf{Z},\overline{\mathbf{Z}})$ of superconformal weight (p,q) transforms according to the rule

$$\mathcal{C}'_{p,q}(\mathbf{Z}',\overline{\mathbf{Z}}')=(cZ+d+\Theta\overline{\gamma}+\overline{\Theta}\gamma)^{p+q}e^{-(p-q)H}\mathcal{C}_{p,q}(\mathbf{Z},\overline{\mathbf{Z}}), \tag{28}$$

with

$$H(\mathbf{Z})=h+\frac{1}{cZ+d}\left[2(\overline{\alpha}Z+\overline{\beta})+\Theta\overline{\gamma}-\overline{\Theta}\gamma+\frac{1}{2}\Theta\overline{\Theta}c\right]$$

under a superprojective change of coordinates $\mathbf{Z}\to\mathbf{Z}'$.

C. Projective connection

A *superprojective (super Schwarzian) connection* is a collection $\mathcal{R}\equiv\mathcal{R}_{\theta\overline{\theta}}$ of superfields (one for each coordinate patch) which are locally superanalytic (i.e., $D_-\mathcal{R}=0=\overline{D}_-\mathcal{R}$) and which transform under a superconformal change of coordinates according to

$$\mathcal{R}'(\mathbf{z}')=e^{w+\overline{w}}[\mathcal{R}(\mathbf{z})-\mathcal{S}(\mathbf{z}',\mathbf{z})]. \tag{29}$$

The existence of such connections on compact SRS's of arbitrary genus can be proven along the lines of the $N=0$ and $N=1$ theories.[1]

If $\mathbf{Z}$ belongs to a projective atlas, then the super Schwarzian derivative of the superconformal mapping $\mathbf{z}\to\mathbf{Z}$ represents a projective connection,

$$\mathcal{R}(\mathbf{z})=\mathcal{S}(\mathbf{Z},\mathbf{z}). \tag{30}$$

In fact, eq. (23) then implies the transformation law (29) and it ensures that the expression (30) is inert under a projective change of coordinates $\mathbf{Z}\to\mathbf{Z}'$. The relation (30) establishes the equivalence between projective structures and projective connections.

The superfield $\mathcal{R}$ admits the θ-expansion

$$\mathcal{R}_{\theta\bar{\theta}}(\mathbf{z})=\rho_z+\theta\eta_{\theta z}+\overline{\theta\eta_{\bar{\theta}z}}+\theta\bar{\theta}[-\tfrac{1}{2}r_{zz}].$$

The component fields $\rho=\mathcal{R}|,\ldots,r=[D,\bar{D}]\mathcal{R}|$ are locally holomorphic ($0=\bar{\partial}\rho=\cdots=\bar{\partial}r$) and the bosonic components have the transformation laws

$$\begin{aligned}\rho'&=(\partial z')^{-1}[\rho-\partial k],\\ r'&=(\partial z')^{-2}[r-S-(\rho-\tfrac{1}{2}\partial k)\partial k].\end{aligned} \tag{31}$$

From these equations, we conclude that the quantity $R=r-\frac{1}{2}\rho^2$ transforms like (2), i.e., like an ordinary projective connection while ρ represents a $U(1)$-connection.

Using a superprojective connection, we can define the second order operator

$$\mathcal{D}_{p,q}=p\bar{D}D-qD\bar{D}+pq\mathcal{R}, \tag{32}$$

which projects out an ordinary conformal field from the superconformal field $C_{p,q}$:

$$(\mathcal{D}_{p,q}C_{p,q})'|=e^{-(k/2)(p-q)}(\partial z')^{-(1/2)(p+q+2)}(\mathcal{D}_{p,q}C_{p,q})|.$$

D. Affine connection, covariant derivative and Miura transformation

A *superaffine connection* is a collection $B\equiv B_\theta,\bar{B}\equiv\bar{B}_{\bar{\theta}}$ of superfields (one for each coordinate patch) which are locally superanalytic, which satisfy the chirality conditions $DB=0=\overline{DB}$ and which transform under a superconformal change of coordinates according to

$$\begin{aligned}B'(\mathbf{z}')&=e^{w}[B(\mathbf{z})+D\bar{w}],\\ \bar{B}'(\mathbf{z}')&=e^{\bar{w}}[\bar{B}(\mathbf{z})+\bar{D}w].\end{aligned} \tag{33}$$

The only compact SRS's which admit a globally defined affine connection are those of genus one.[1] Nevertheless, these quantities can always be introduced locally and they represent extremely useful computational tools for deriving globally well-defined results.

Using an affine connection we can introduce *supercovariant derivatives*,

$$\begin{aligned}\nabla_{p,q}&=D-qB:\mathcal{F}_{p,q}\to\mathcal{F}_{p+1,q},\\ \bar{\nabla}_{p,q}&=\bar{D}-p\bar{B}:\mathcal{F}_{p,q}\to\mathcal{F}_{p,q+1}.\end{aligned} \tag{34}$$

Products of covariant derivatives are defined by taking into account the weights, e.g.,

$$\nabla\bar{\nabla}\nabla C_{p,q}=\nabla_{p+1,q+1}\bar{\nabla}_{p+1,q}\nabla_{p,q}C_{p,q}.$$

From the very definitions, it then follows that

$$\nabla^2=0=\bar{\nabla}^2. \tag{35}$$

Locally, we can express B and $\bar{B}$ as

$$B=DQ, \quad \overline{B}=\overline{D}Q, \tag{36}$$

with

$$Q=\ln(D\Theta\overline{D}\overline{\Theta})=\ln\Big(\partial Z+\tfrac{1}{2}\,\Theta\partial\overline{\Theta}+\tfrac{1}{2}\,\overline{\Theta}\partial\Theta\Big), \tag{37}$$

where Θ and $\overline{\Theta}$ belong to an atlas of superprojective coordinates $\mathbf{Z}$. Then, we have the operatorial expressions

$$\begin{aligned}\nabla_{p,q}&=(\overline{D}\overline{\Theta})^q\cdot D\cdot(\overline{D}\overline{\Theta})^{-q},\\ \overline{\nabla}_{p,q}&=(D\Theta)^p\cdot\overline{D}\cdot(D\Theta)^{-p}.\end{aligned} \tag{38}$$

Affine and projective connections are related by the *super Miura transformation*,

$$\mathcal{R}=D\overline{B}-\overline{D}B-B\overline{B}=[D,\overline{D}]Q-(DQ)(\overline{D}Q). \tag{39}$$

This formula follows from the expression of the projective connection in terms of the Schwarzian derivative [eq. (30)] or, equivalently, by comparing the expressions equations (53) and (52) below for the conformally covariant operator $\mathcal{L}_1$.

From eq. (39) we can determine explicit expressions for the component fields $[D,\overline{D}]\mathcal{R}|=r=R+\frac{1}{2}\rho^2$ and $\mathcal{R}|=\rho$. In fact, by taking into account (37), one recovers the ordinary Miura transformation for the projective connection R and an explicit expression for ρ:

$$\begin{aligned}R&=\partial^2\,\ln\,\partial Z|-\tfrac{1}{2}(\partial\,\ln\,\partial Z|)^2,\\ \rho&=([D,\overline{D}]\ln\,\partial Z)|.\end{aligned} \tag{40}$$

Furthermore, from eqs. (36) and (37), we conclude that

$$\begin{aligned}b\equiv\overline{D}B|&=\tfrac{1}{2}(\partial\,\ln\,\partial Z|-\rho),\\ \overline{b}\equiv D\overline{B}|&=\tfrac{1}{2}(\partial\,\ln\,\partial Z|+\rho).\end{aligned} \tag{41}$$

From these expressions, we obtain the *conformally and $U(1)$-covariant space-time derivative*,

$$\{\nabla,\overline{\nabla}\}C_{p,q}|=\left[\tilde{\partial}_{p,q}-\frac{p+q}{2}(\partial\,\ln\,\partial Z|)\right]c_{p,q},\ \text{ with }\quad \tilde{\partial}_{p,q}\equiv\partial-\frac{p-q}{2}\rho. \tag{42}$$

Here, $\tilde{\partial}_{p,q}$ denotes the $U(1)$-covariant derivative associated to the $U(1)$-connection ρ [cf. eq. (31)] and the local expression $\partial\,\ln\,\partial Z|$ transforms like an affine connection with respect to a conformal change of coordinates $z\to z'(z)$.[1]

E. Determination of conformally covariant operators

We are interested in superconformally covariant differential operators which are globally defined on compact SRS's of any genus. In order to construct these quantities, we start from an operator $\mathcal{L}$ which is a differential polynomial in the covariant derivatives $\nabla,\overline{\nabla}$ and require that it only depends on the affine connections $B,\overline{B}$ through the particular combination $\mathcal{R}=D\overline{B}-\overline{D}B-B\overline{B}$ (which is globally defined): using a variational argument,[8,2] one imposes $\delta\mathcal{L}=0$ while varying $B,\overline{B}$ subject to the condition that $\mathcal{R}$ is fixed,

$$0=\delta\mathcal{R}=D\,\delta\overline{B}-\overline{D}\,\delta B-\delta B\overline{B}-B\,\delta\overline{B}.$$

The latter relation can be rewritten as

$$\nabla\delta\overline{B}=\overline{\nabla}\delta B, \tag{43}$$

and the nilpotency of $\nabla,\overline{\nabla}$ then yields

$$\overline{\nabla}\nabla\delta\overline{B}=0=\nabla\overline{\nabla}\delta B, \tag{44}$$

while the chirality properties of $B,\overline{B}$ imply

$$\nabla\delta B=0=\overline{\nabla}\delta\overline{B}. \tag{45}$$

There are two basic classes of covariant differential operators: besides the super Bol operators (which only depend on a projective structure or equivalently on a projective connection), one can introduce operators which also depend linearly on a conformal field. These two classes will be determined in sections IV and V, respectively, and at the end of section V, we will show that the most general covariant operator can be written as a sum of these.

IV. SUPER BOL OPERATORS

We successively discuss the cases where the differential operator $\mathcal{L}$ contains an even and odd number of derivatives $\nabla,\overline{\nabla}$. All operators discussed in this section are supposed to be *normalized* in the sense that the coefficients of the highest order derivatives are constant.

A. Even number of derivatives

The general form of a normalized covariant operator involving an even number of derivatives is

$$\mathcal{L}_n^{even}=\alpha(\nabla\overline{\nabla})^n+\beta(\overline{\nabla}\nabla)^n \quad [n\in\mathbf{N}^*,(\alpha,\beta)\neq(0,0)], \tag{46}$$

and such that

$$\mathcal{L}_n^{even}:\mathcal{F}_{p,q}\to\mathcal{F}_{p+n,q+n}, \quad \text{for some } (p,q). \tag{47}$$

Thus, the operator does not modify the $U(1)$ charge of the fields. In order to simplify the notation, we do not spell out the dependence of $\mathcal{L}$ on p and q.

Imposing $\delta\mathcal{L}_n^{even}=0$ and using eqs. (43)–(45), one finds that p,q and n are related: the general solution reads as

$$\begin{aligned}\mathcal{L}_n^{even}&=q(\nabla\overline{\nabla})^n-p(\overline{\nabla}\nabla)^n, \quad \text{with } p+q=-n,\\ &=\{\nabla,\overline{\nabla}\}^{n-1}(q\nabla\overline{\nabla}-p\overline{\nabla}\nabla).\end{aligned} \tag{48}$$

Here and in the following, the constant overall factor of $\mathcal{L}$ has been chosen in a convenient way. For $n=1$ and $n=2$, we have the explicit expressions

$$\begin{aligned}\mathcal{L}_1^{even}&=qD\overline{D}-p\overline{D}D-pq\mathcal{R}, \quad \text{with } p+q=-1,\\ &=\frac{p+q}{2}\mathcal{L}_1^{sym}-\frac{p-q}{2}\left(\partial-\frac{p-q}{2}\mathcal{R}\right),\end{aligned}$$

$$\mathcal{L}_2^{even}=\frac{p+q}{2}\mathcal{L}_2^{sym}-\frac{p-q}{2}\Big(\partial^2-\mathcal{R}[D,\overline{D}]-2\frac{p-q}{2}\mathcal{R}\partial+(\overline{D}\mathcal{R})D$$
$$-(D\mathcal{R})\overline{D}-\frac{p-q}{2}\partial\mathcal{R}-pq\mathcal{R}^2\Big),\quad \text{with}\quad p+q=-2, \tag{49}$$

where $\mathcal{L}_1^{sym}$ and $\mathcal{L}_2^{sym}$ denote the symmetric solution ($p=q$); see eq. (53) below. Thus, the covariant operator $\mathcal{L}_n^{even}$ has a conformal part (a multiple of $\mathcal{L}_n^{sym}$) and a $U(1)$ part, each of which is proportional to the corresponding weight.

Projection to components is achieved by virtue of the operator $\mathcal{D}_{p,q}$ introduced in eq. (32):

$$(\mathcal{D}_{p+n,q+n}\mathcal{L}_n^{even}C_{p,q})|=pqL_{n+1}^{p,q}c_{p,q}\,. \tag{50}$$

Here, $L_n^{p,q}$ denotes the *generalization of the usual Bol operator* L_n (depending on the ordinary projective connection $R=r-\frac{1}{2}\rho^2$) *to charged conformal fields*: it amounts to replacing the ordinary partial derivative (acting on $c_{p,q}$) by the $U(1)$-covariant derivative $\tilde{\partial}_{p,q}$ introduced in eq. (42), e.g.,

$$L_2^{p,q}=\tilde{\partial}_{p,q}^2+\tfrac{1}{2}R,\quad \text{with}\ \ p+q=-1\,,$$
$$L_3^{p,q}=\tilde{\partial}_{p,q}^3+2R\tilde{\partial}_{p,q}+(\partial R),\quad \text{with}\ \ p+q=-2. \tag{51}$$

Obviously, the general solution (48) involves two different classes of solutions: the *symmetric solution* ($p=q$) for which the operator acts on neutral fields and the *asymmetric solutions* ($p\neq q$) among which we find as particular cases the *chiral* ($p=0$) and *anti-chiral* ($q=0$) *solutions*.

1. Symmetric solution

The symmetric solution $\mathcal{L}_n^{sym}:\mathcal{F}_{-n/2,-n/2}\to\mathcal{F}_{n/2,n/2}$ has the general form

$$\mathcal{L}_n^{sym}=(\nabla\overline{\nabla})^n-(\overline{\nabla}\nabla)^n=\partial^{n-1}[D,\overline{D}]+\frac{n}{2}\mathcal{R}\partial^{n-1}+\cdots, \tag{52}$$

and the simplest examples are

$$\mathcal{L}_1^{sym}=[D,\overline{D}]+\tfrac{1}{2}\mathcal{R},$$
$$\mathcal{L}_2^{sym}=\partial[D,\overline{D}]+\mathcal{R}\partial-(D\mathcal{R})\overline{D}-(\overline{D}\mathcal{R})D+(\partial\mathcal{R}), \tag{53}$$
$$\mathcal{L}_3^{sym}=\partial^2[D,\overline{D}]+\tfrac{3}{2}\mathcal{R}\partial^2-3(D\mathcal{R})\partial\overline{D}-3(\overline{D}\mathcal{R})\partial D+3(\partial\mathcal{R})\partial+(\tfrac{1}{2}[D,\overline{D}]\mathcal{R}-\tfrac{1}{4}\mathcal{R}^2)[D,\overline{D}]$$
$$-(2\partial D\mathcal{R}+\tfrac{1}{4}D\mathcal{R}^2)\overline{D}-(2\partial\overline{D}\mathcal{R}-\tfrac{1}{4}\overline{D}\mathcal{R}^2)D$$
$$+\tfrac{3}{2}(\partial^2\mathcal{R}+\tfrac{1}{2}\mathcal{R}[D,\overline{D}]\mathcal{R}-\tfrac{1}{4}\mathcal{R}^3-2(D\mathcal{R})(\overline{D}\mathcal{R})),$$

$$\mathcal{L}_4^{sym}=\partial^3[D,\overline{D}]+2\mathcal{R}\partial^3-6(\overline{D}\mathcal{R})\partial^2 D-6(D\mathcal{R})\partial^2\overline{D}+6\partial\mathcal{R}\partial^2+(2([D,\overline{D}]\mathcal{R})-\mathcal{R}^2)\partial[D,\overline{D}]$$
$$-(8\partial\overline{D}\mathcal{R}-2\mathcal{R}\overline{D}\mathcal{R})\partial D-(8\partial D\mathcal{R}+2\mathcal{R}D\mathcal{R})\partial\overline{D}+(6\partial^2\mathcal{R}+4\mathcal{R}[D,\overline{D}]\mathcal{R}-15D\mathcal{R}\overline{D}\mathcal{R}$$
$$-2\mathcal{R}^3)\partial+(\partial[D,\overline{D}]\mathcal{R}-\mathcal{R}\partial\mathcal{R})[D,\overline{D}]-(3\partial^2\overline{D}\mathcal{R}-\mathcal{R}\partial\overline{D}\mathcal{R}-\tfrac{3}{2}\overline{D}\mathcal{R}\partial\mathcal{R}+\tfrac{9}{2}\overline{D}\mathcal{R}[D,\overline{D}]\mathcal{R}$$
$$-2\mathcal{R}^2\overline{D}\mathcal{R})D-(3\partial^2 D\mathcal{R}+\mathcal{R}\partial D\mathcal{R}+\tfrac{3}{2}D\mathcal{R}\partial\mathcal{R}+\tfrac{9}{2}D\mathcal{R}[D,\overline{D}]\mathcal{R}-2\mathcal{R}^2 D\mathcal{R})\overline{D}$$
$$+2\mathcal{R}\partial[D,\overline{D}]\mathcal{R}-10D\mathcal{R}\partial\overline{D}\mathcal{R}+10\overline{D}\mathcal{R}\partial D\mathcal{R}+4\partial\mathcal{R}[D,\overline{D}]\mathcal{R}-4\mathcal{R}^2\partial\mathcal{R}+2\partial^3\mathcal{R},$$

where the derivatives only act on the field to their immediate right. Note that the result for $\mathcal{L}_1^{sym}$ is nothing but the Miura transformation (39). The operator $\mathcal{L}_n^{sym}$ has the hermiticity property

$$\int dzd\theta d\overline{\theta} f\mathcal{L}_n^{sym}g=(-1)^{n+1}\int dzd\theta d\overline{\theta}(\mathcal{L}_n^{sym}f)g, \quad \text{for } f,g\in\mathcal{F}_{-n/2,-n/2}.$$

The operators (53) play the role of Poisson (Hamiltonian) operators in the description of super-integrable models—see section VII.

2. Chiral and anti-chiral solutions

The anti-chiral ($q=0$) and chiral ($p=0$) solutions read as

$$\mathcal{L}_n^{anti}=(\overline{\nabla}\overline{\nabla})^n=\overline{D}(\nabla\overline{\nabla})^{n-1}D:\mathcal{F}_{-n,0}\to\mathcal{F}_{0,n}^a,$$
$$\mathcal{L}_n^{chir}=(\nabla\overline{\nabla})^n=D(\overline{\nabla}\overline{\nabla})^{n-1}\overline{D}:\mathcal{F}_{0,-n}\to\mathcal{F}_{n,0}^c. \tag{54}$$

For these operators, the target spaces are the anti-chiral and chiral subspaces, respectively, since $\overline{D}\mathcal{L}_n^{anti}=0$ and $D\mathcal{L}_n^{chir}=0$. Both classes of operators are related by Hermitian conjugation:

$$\int dzd\theta d\overline{\theta} f\mathcal{L}_n^{chir}g=\int dzd\theta d\overline{\theta}(\mathcal{L}_n^{anti}f)g, \quad \text{for } f\in\mathcal{F}_{-n,0}, g\in\mathcal{F}_{0,-n}.$$

Differential operators of this type have been introduced in reference 9: they occur as Lax operators in the $N=2$ supersymmetric KdV and KP hierarchies[10,9,11] —see section V.

Since q and p vanish for $\mathcal{L}_n^{anti}$ and $\mathcal{L}_n^{chir}$, respectively, the r.h.s. of eq. (50) vanishes for both of these operators: we will discuss their relation with ordinary Bol operators in the next section.

B. Sandwich operators

Before considering operators involving an odd number of derivatives, it is worthwhile to have another look at the chiral and anti-chiral solutions.

The anti-chiral solution (54) has the form

$$\mathcal{L}_n^{anti}C_{-n,0}=\overline{D}(\nabla\overline{\nabla})^{n-1}DC_{-n,0}\equiv\overline{D}\mathcal{K}_{n-1}DC_{-n,0}, \tag{55}$$

where $\mathcal{K}_n$ will be referred to as a ''*sandwich operator*'' for obvious reasons. Since $DC_{-n,0}\equiv\Phi_{-n+1,0}$ represents a chiral field, we have obtained a new conformally covariant operator,

$$\overline{D}\mathcal{K}_{n-1}=\overline{D}(\nabla\overline{\nabla})^{n-1}:\mathcal{F}_{-n+1,0}^c\to\mathcal{F}_{0,n}^a. \tag{56}$$

Although $\mathcal{K}_{n-1}$ formally coincides with $\mathcal{L}_{n-1}^{chir}=(\nabla\overline{\nabla})^{n-1}$, it does not represent the same operator, because it does not act on the same space. We have the operatorial relations

$$\overline{D}\mathcal{K}_1=\overline{D}[D\overline{D}+\mathcal{R}]=\overline{D}[\partial+\mathcal{R}],$$
$$\overline{D}\mathcal{K}_2=\overline{D}[\partial^2+3\mathcal{R}\partial+(\overline{D}D\mathcal{R})+2(D\overline{D}\mathcal{R})+2\mathcal{R}^2]. \tag{57}$$

Note that the operator $\mathcal{K}_n$ can be written in different forms since it is only defined as ''sandwiched'' between $\overline{D}$ and D.

The projection to component field expressions is simply done by applying D to $\overline{D}\mathcal{K}_n$:

$$(D\overline{D}\mathcal{K}_n C)|=L_{n+1}^{-n,0}c, \quad \text{for } C\in\mathcal{F}^c_{-n,0}. \tag{58}$$

Here, $C|=c$ and $L_n^{p,q}$ denotes the generalization of the Bol operator L_n to charged conformal fields; see eq. (51). This shows that $\mathcal{L}_n^{anti}$ also represents a supersymmetric generalization of the Bol operators, though the latter have to be projected out according to the procedure (58) rather than (50).

Of course, one can apply the same line of reasoning to the chiral solution for which

$$\mathcal{L}_n^{chir}C_{0,-n}=D(\overline{\nabla}\nabla)^{n-1}\overline{D}C_{0,-n}\equiv D\overline{\mathcal{K}}_{n-1}\overline{D}C_{0,-n} \tag{59}$$

and

$$D\overline{\mathcal{K}}_{n-1}:\mathcal{F}^a_{0,-n+1}\to\mathcal{F}^c_{n,0}. \tag{60}$$

C. Odd number of derivatives

The general form of a normalized covariant operator involving an odd number of derivatives is

$$\mathcal{L}_n^{odd}=\overline{\nabla}(\nabla\overline{\nabla})^n,$$

or

$$\overline{\mathcal{L}}_n^{odd}=\nabla(\overline{\nabla}\nabla)^n, \quad \text{with } n=0,1,2,\ldots, \tag{61}$$

acting on $\mathcal{F}_{p,q}$. These operators modify the $U(1)$ charge of the fields.

For $n=0$, the condition of covariance $\delta\mathcal{L}_n=0$ leads to the explicit results

$$\begin{aligned}\mathcal{L}_0^{odd}C_{0,q}&=\overline{\nabla}C_{0,q}=\overline{D}C_{0,q}\\ \overline{\mathcal{L}}_0^{odd}C_{p,0}&=\nabla C_{p,0}=DC_{p,0}.\end{aligned} \tag{62}$$

For $n\geqslant 1$, we find

$$\mathcal{L}_n^{odd}=\overline{D}(\nabla\overline{\nabla})^n:\{C\in\mathcal{F}_{-n,0}/\mathcal{L}_n^{anti}C=0\}\to\mathcal{F}^{\,a}_{0,n+1}, \tag{63}$$

the simplest examples being

$$\begin{aligned}\mathcal{L}_1^{odd}C_{-1,0}&=\overline{D}[\partial+\mathcal{R}]C_{-1,0},\\ \mathcal{L}_2^{odd}C_{-2,0}&=\overline{D}[\partial^2+3\mathcal{R}\partial-(\overline{D}\mathcal{R})D+(\overline{D}D\mathcal{R})+2(D\overline{D}\mathcal{R})+2\mathcal{R}^2]C_{-2,0},\end{aligned} \tag{64}$$

with

$$\begin{aligned}0&=\mathcal{L}_1^{anti}C_{-1,0}=\overline{D}DC_{-1,0},\\ 0&=\mathcal{L}_2^{anti}C_{-2,0}=\overline{D}[\partial+\mathcal{R}]DC_{-2,0}.\end{aligned} \tag{65}$$

If $C_{-n,0}$ is a chiral field, it automatically satisfies $\mathcal{L}_n^{anti}C_{-n,0}=0$, henceforth

$$\mathcal{F}^{\,c}_{-n,0}\subset\{C\in\mathcal{F}_{-n,0}/\mathcal{L}_n^{anti}C=0\}.$$

By comparing eqs. (56) and (63), we conclude that $\mathcal{L}_n^{odd}=\overline{D}(\nabla\overline{\nabla})^n$ is an extension (in the functional analytic sense[12]) of the operator $\overline{D}\mathcal{K}_n=\overline{D}(\nabla\overline{\nabla})^n$: both operators differ by their domain of definition and act in the same way on the smaller domain, i.e., on $\mathcal{F}^{\,c}_{-n,0}$. The Bol operator $L^{-n,0}_{n+1}$ acting on $c=C|$ is extracted in the same manner from $\mathcal{L}_n^{odd}C$ and $\overline{D}\mathcal{K}_nC$, i.e., according to eq. (58).

D. Comments

1. Super Bol operators in projective coordinates

The Bol operator L_n is most directly obtained from the Möbius covariant operator ∂_Z^n (acting on quasi-primary fields of appropriate weight) by going over from the projective coordinates Z to a generic system of conformal coordinates z by a conformal transformation (see Ref. 1 and references therein). Proceeding in the same way in the $N=2$ supersymmetric case, we start from the differential operator $\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]=(D_\Theta D_{\overline{\Theta}})^n-(D_{\overline{\Theta}}D_\Theta)^n$ which transforms homogeneously under a superprojective change of coordinates $\mathbf{Z}\to\mathbf{Z}'$ when acting on quasi-primary fields of appropriate weight [see eq. (28)]:

$$(\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]\mathcal{C}_{-n/2,-n/2})'=(cZ+d+\Theta\overline{\gamma}+\overline{\Theta}\gamma)^n\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]\mathcal{C}_{-n/2,-n/2}. \tag{66}$$

By passing over from the projective coordinates $\mathbf{Z}$ to generic conformal coordinates $\mathbf{z}$ by a superconformal transformation, the operator $\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]$ becomes the super Bol operator $\mathcal{L}_n^{sym}$: we have

$$\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]\mathcal{C}_{-n/2,-n/2}\equiv(D\Theta\overline{D}\overline{\Theta})^{-n/2}\mathcal{L}_n^{sym}C_{-n/2,-n/2}, \tag{67}$$

where $\mathcal{C}$ and C are related by

$$\mathcal{C}_{-n/2,-n/2}(\mathbf{Z},\overline{\mathbf{Z}})\equiv(D\Theta\overline{D}\overline{\Theta})^{n/2}C_{-n/2,-n/2}(\mathbf{z},\overline{\mathbf{z}}). \tag{68}$$

In operatorial form, eq (67) reads as

$$\mathcal{L}_n^{sym}=(D\Theta\overline{D}\overline{\Theta})^{n/2}\cdot\partial_Z^{n-1}[D_\Theta,D_{\overline{\Theta}}]\cdot(D\Theta\overline{D}\overline{\Theta})^{n/2}, \tag{69}$$

which means that $\mathcal{L}_n^{sym}$ represents the superconformally covariant version of the differential operator $\partial^{n-1}[D,\overline{D}]$ acting on $U(1)$-neutral superfields.

The same line of arguments applies to the sandwich operator $\mathcal{K}_n$ acting on chiral fields. In fact, $DC_{-n,0}=0$ is equivalent to $D_\Theta\mathcal{C}_{-n,0}=0$ where $\mathcal{C}_{-n,0}\equiv(D\Theta)^nC_{-n,0}$ and for such fields one has

$$D_{\overline{\Theta}}\partial_Z^n\mathcal{C}_{-n,0}=(\overline{D\Theta})^{-(n+1)}\overline{D}\mathcal{K}_nC_{-n,0}. \tag{70}$$

2. Recursion and factorization relations

The super Bol operators can be determined by using a *recursion relation*,[1] e.g., if $\psi\in\mathcal{F}_{-1/2,-1/2}$ is assumed to be Grassmann even, then

$$\mathcal{L}_1^{sym}\psi=0\Rightarrow\mathcal{L}_n^{sym}\psi^n=0. \tag{71}$$

Thus, one can determine the explicit form of $\mathcal{L}_2^{sym}$ by applying $\partial[D,\overline{D}]$ to ψ^2 while using $\mathcal{L}_1^{sym}\psi=0$ (i.e., $[D,\overline{D}]\psi=-\frac{1}{2}\mathcal{R}\psi$) and so on for $\mathcal{L}_3^{sym},\ldots$. The proof of (71) consists of going over to superprojective coordinates.

The operators $\overline{D}\mathcal{K}_n$ can be determined in the same way: if $\psi\in\mathcal{F}^{c}_{-1,0}$ is Grassmann even, then

$$\overline{D}\mathcal{K}_1\psi=0\Rightarrow\overline{D}\mathcal{K}_n\psi^n=0. \tag{72}$$

If the super Bol operators are defined by virtue of Möbius covariant operators as in eq. (69), their expressions in terms of covariant derivatives [eq. (52)] represent *factorization formulae*. The latter may then be proven by going over to projective coordinates for which ∇ and $\overline{\nabla}$ reduce to D_Θ and $D_{\bar{\Theta}}$, respectively.[1]

3. From N=2 to N=1 operators

An $N=2$ superfield F can be projected to an $N=1$ superfield $F|_{N=1}$ by evaluating it at $\theta=\overline{\theta},\theta^-=\overline{\theta^-}$. By virtue of this projection, one finds that the quantity $\mathcal{R}_1\equiv\frac{1}{2}(D-\overline{D})\mathcal{R}|_{N=1}$ transforms like an $N=1$ superprojective connection, i.e., according to $\mathcal{R}_1'=e^{3W}(\mathcal{R}_1-\mathcal{S}_1)$ where $W\equiv w|_{N=1}=\overline{w}|_{N=1}$ and where $\mathcal{S}_1\equiv\mathcal{S}|_{N=1}$ denotes the $N=1$ super Schwarzian derivative.

Since sandwich operators act on chiral fields C, we have $\overline{D}C=(D+\overline{D})C\equiv D_1C$ where $D_1=D+\overline{D}$ denotes the basic $N=1$ derivative introduced in section II F. For the simplest sandwich operator, we thus obtain the expression

$$(\overline{D}\mathcal{K}_1C)|_{N=1}=([D_1^3+(\overline{D}\mathcal{R})+\mathcal{R}D_1]C)|_{N=1},$$

which represents a $U(1)$-covariant version of the $N=1$ super Bol operator $D_1^3+\mathcal{R}_1$.

V. COVARIANT OPERATORS INVOLVING CONFORMAL FIELDS

In this section, we introduce covariant operators $\mathcal{M}^{(n)}_{\mathcal{W}_{p,q}}$ which depend on a projective connection $\mathcal{R}$ and, in a linear way, on a superconformal field $\mathcal{W}_{p,q}$. We restrict our attention to the case of sandwich operators since these are the only ones for which we will be able to give a matrix representation.

Thus, we try to construct a covariant operator which acts between the same spaces as $\mathcal{K}_n$ [cf. eq. (56)]:

$$\overline{D}\mathcal{M}^{(n)}_{\mathcal{W}_{p,q}}:\mathcal{F}^{\ c}_{-n,0}\rightarrow\mathcal{F}^{\ a}_{0,n+1}.$$

To do so, we proceed as before and first write $\mathcal{M}^{(n)}_{\mathcal{W}_{p,q}}$ as a differential polynomial in the nilpotent operators ∇ and $\overline{\nabla}$. It turns out that a covariant operator can only be obtained in the case of neutral superfields $\mathcal{W}_{k,k}$ of conformal weight $k\in\mathbf{N}$ with $1\leqslant k\leqslant n$: the corresponding operator reads as

$$\mathcal{M}^{(n)}_{\mathcal{W}_{k,k}}=\sum_{l=0}^{n-k}\{a^{(n)}_{kl}[(\nabla\overline{\nabla})^l\mathcal{W}_{k,k}]+b^{(n)}_{kl}[(\overline{\nabla}\nabla)^l\mathcal{W}_{k,k}]\}(\nabla\overline{\nabla})^{n-k-l},\tag{73}$$

with constant coefficients $a^{(n)}_{kl}$ and $b^{(n)}_{kl}$ (with $l=0,...,k$) to be determined by imposing the condition $\delta\overline{D}\mathcal{M}^{(n)}_{\mathcal{W}_{k,k}}=0$ and using eqs. (43)–(45). This leads to the following result (which is unique up to a global factor):

$$a^{(n)}_{kl}=\frac{\binom{n-k}{l}\binom{k+l}{l}}{\binom{2k+l}{l}}\quad,\qquad b^{(n)}_{kl}=\frac{\binom{n-k}{l}\binom{k+l-1}{l}}{\binom{2k+l}{l}},\tag{74}$$

for $l=1,\ldots,n-k$ and $a^{(n)}_{k0}+b^{(n)}_{k0}=1$. For instance,

$$\mathcal{M}^{(n)}_{\mathcal{W}_{n,n}}=\mathcal{W}_{n,n}$$

$$\mathcal{M}^{(n)}_{\mathcal{W}_{n-1,n-1}}=\mathcal{W}_{n-1,n-1}\partial+\frac{n-1}{2n-1}(D\mathcal{W}_{n-1,n-1})\overline{D}+\frac{n}{2n-1}(D\overline{D}\mathcal{W}_{n-1,n-1})$$
$$+\frac{n^2}{2n-1}\mathcal{R}\mathcal{W}_{n-1,n-1},\tag{75}$$

where we recall that sandwich operators can be cast into different forms since they are only defined as sandwiched between $\overline{D}$ and D. Component field results are recovered by applying the derivative D and subsequently projecting to the lowest component:

$$(D\overline{D}\mathcal{M}^{(n)}_{\mathcal{W}_{k,k}}C_{-n,0})|=\left[\widetilde{M}^{(n+1)}_{w_k}-\frac{n+k+1}{2k(2k+1)}\,\widetilde{M}^{(n+1)}_{v_{k+1}}\right]c_{-n,0}, \tag{76}$$

with $w_k=\mathcal{W}_{k,k}|$ and $v_{k+1}=(\mathcal{D}_{k,k}\mathcal{W}_{k,k})|$ where $\mathcal{D}_{k,k}$ denotes the second order operator (32). Here, $\widetilde{M}^{(n)}_{w_k}$ is the $U(1)$-covariant generalization of the operator $M^{(n)}_{w_k}$ mentioned in the Introduction—see eq. (3).

The *most general* normalized superconformally covariant sandwich operator of order n has the form

$$\begin{aligned}&\mathcal{L}^{(n)}:\mathcal{F}^{\ c}_{-n,0}\to\mathcal{F}^{\ a}_{0,n+1},\\ &C_{-n,0}\mapsto\mathcal{L}^{(n)}C_{-n,0}=\overline{D}[\partial^n+a^{(n)}_1\partial^{n-1}+a^{(n)}_2\partial^{n-2}+\ldots+a^{(n)}_n]C_{-n,0},\end{aligned} \tag{77}$$

with even coefficient functions $a^{(n)}_k(\mathbf{z})$ transforming in an appropriate way under superconformal changes of coordinates—see Refs. 2,1 and 3 for the $N=0$ and $N=1$ supersymmetric theories, respectively. It can be expressed as a sum of the previously introduced operators,

$$\mathcal{L}^{(n)}=\overline{D}\left[\mathcal{K}_n+\sum_{k=2}^{n}\mathcal{M}^{(n)}_{\mathcal{W}_{k,k}}\right]=\overline{D}\left[\partial^n+\frac{n(n+1)}{2}\,\mathcal{R}\partial^{n-1}+\ldots\right], \tag{78}$$

i.e., it can be parametrized in terms of a projective connection $\mathcal{R}$ and $n-1$ neutral superconformal fields $\mathcal{W}_{2,2},\ldots,\mathcal{W}_{n,n}$. The relation between these fields and the coefficients $a^{(n)}_k$ is invertible and given by differential polynomials.

Operators of the form (77) with $C_{-n,0}=D\Phi$ occur as Lax operators, e.g., $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ give rise to the $N=2$ super KdV and Boussinesq equations, respectively.[9,11]

VI. MATRIX REPRESENTATION OF COVARIANT OPERATORS

In this section, we discuss the matrix representation of the sandwich operators (78) while starting with the particular case of the super Bol operator,

$$\overline{D}\mathcal{K}_n=\overline{D}(\nabla\overline{\nabla})^n=\overline{D}[D-nB][\overline{D}+\overline{B}][D-(n-1)B][\overline{D}+2\overline{B}]\cdots[D-B][\overline{D}+n\overline{B}]. \tag{79}$$

The two conformally covariant differential equations,

$$Df=0,\quad \overline{D}\mathcal{K}_nf=0\quad (f\in\mathcal{F}_{-n,0}), \tag{80}$$

are equivalent to two systems of first order differential equations which can be cast into matrix form:

$$Q^{fac}_n\vec{f}=\vec{0},\quad \overline{Q}^{fac}_n\vec{f}=\vec{0}. \tag{81}$$

Here, $\vec{f}=(f_1,f_2,\ldots,f_{2n},f)^t$ and $Q^{fac}_n,\overline{Q}^{fac}_n$ denote the $(2n+1)\times(2n+1)$ matrix operators

$$Q^{fac}_n=-\mu_-+D\mathbf{1}-BH,\quad \overline{Q}^{fac}_n=-\overline{\mu}_-+\overline{D}\mathbf{1}-\overline{BH}, \tag{82}$$

where $\mu_-,\overline{\mu}_-,H,\overline{H}$ belong to the superprincipal embedding $sl(2|1)_{\text{pal}}\subset sl(n+1|n)$ of the Lie superalgebra $sl(2|1)$ into the Lie superalgebra $sl(n+1|n)$—see eqs. (A3) of Appendix A. We note that the operators (82) can be rewritten in the form

$$Q^{fac}_n=D\mathbf{1}-A^{fac}_\theta,\quad \overline{Q}^{fac}_n=\overline{D}\mathbf{1}-A^{fac}_{\overline{\theta}}, \tag{83}$$

where A_θ^{fac} and $\overline{A}_\theta^{fac}$ are to be interpreted as components of a $sl(2|1)_{\text{pal}}$-valued connection—a viewpoint which will be further developed in the next section.

While Q_n^{fac} and $\overline{Q}_n^{fac}$ describe the *factorized form* (79) *of the super Bol operator*, one obtains an equivalent matrix representation for this operator by conjugation with a group element $N \in SL(2|1)_{\text{pal}} \subset SL(n+1|n)$,

$$Q_n = \hat{N}^{-1} \cdot Q_n^{fac} \cdot N, \quad \overline{Q}_n = \hat{N}^{-1} \cdot \overline{Q}_n^{fac} \cdot N, \tag{84}$$

where N is an upper triangular matrix with entries 1 on the diagonal.[13,3] [The matrix equations (81) then become $Q_n \vec{f}\,' = \vec{0}, \overline{Q}_n \vec{f}\,' = \vec{0}$ with $\vec{f}\,' \equiv N^{-1}\vec{f} = (f_1', \ldots, f_{2n}', f)$, i.e., the component f of $\vec{f}$ is not modified which explains why Q_n and $\overline{Q}_n$ still represent the scalar equations (80).] The hatted group element $\hat{N} = \exp \hat{M}$ follows from $N = \exp M$ by changing the sign of the anticommuting part of the Lie superalgebra element M.[14,3]

By using the gauge freedom (84), we find a form of Q_n and $\overline{Q}_n$ which makes the dependence on the superprojective structure manifest:

$$\begin{aligned} Q_n &= -\mu_- + D\mathbf{1} - \tfrac{1}{2}[\mathcal{R}\overline{\mu}_+ + (D\mathcal{R})E_{++}], \\ \overline{Q}_n &= -\overline{\mu}_- + \overline{D}\mathbf{1} + \tfrac{1}{2}[\mathcal{R}\mu_+ + (\overline{D}\mathcal{R})E_{++}]. \end{aligned} \tag{85}$$

This matrix representation of $\overline{D}\mathcal{K}_n$ in which $\mu_\pm, \overline{\mu}_\pm$ and $E_{++} \equiv -\{\mu_+, \overline{\mu}_+\}$ belong to the Lie superalgebra $sl(2|1)_{\text{pal}}$ exhibits most clearly the underlying algebraic structure which is due to the covariance with respect to super Möbius transformations. By unraveling the matrix equations $Q_n \vec{f}\,' = \vec{0}, \overline{Q}_n \vec{f}\,' = \vec{0}$, one recovers our results (57) for the operator $\overline{D}\mathcal{K}_n$.

The matrix representation (85) can be extended to the sandwich operators involving superconformal fields that we discussed in section V:

$$\begin{aligned} Q_n &= -\mu_- + D\mathbf{1} - \frac{1}{2}\sum_{i=1}^{n} [i\mathcal{V}_{i,i}\overline{\mu}_+ + (D\mathcal{V}_{i,i})E_{++}]E_{++}^{i-1}, \\ \overline{Q}_n &= -\overline{\mu}_- + \overline{D}\mathbf{1} + \frac{1}{2}\sum_{i=1}^{n} [i\mathcal{V}_{i,i}\mu_+ + (\overline{D}\mathcal{V}_{i,i})E_{++}]E_{++}^{i-1}. \end{aligned} \tag{86}$$

Here, $\mathcal{V}_{1,1} \equiv \mathcal{R}$ and $\mathcal{V}_{i,i}$ is a superconformal field for $i \geqslant 2$. Then $\mathcal{V}_{i,i}$ $(i = 1, \ldots, n)$ can be expressed in terms of the projective connection $\mathcal{R}$ and the conformal fields $\mathcal{W}_{k,k}$ $(k = 2, \ldots, n)$ appearing in the expression (78) by means of invertible relations involving differential polynomials. Henceforth, the expressions (86) provide a matrix representation for the most general sandwich operator. Thanks to this result, explicit expressions for covariant operators can be derived in a straightforward way.

If the sandwich operators are parametrized in the form (77), their matrix realization reads as

$$\begin{aligned} Q_n^{hor} &= -\mu_- + D\mathbf{1}, \\ \overline{Q}_n^{hor} &= -\overline{\mu}_- + \overline{D}\mathbf{1} + \sum_{i=1}^{n} [\lambda_i a_i^{(n)} E_{1,2i} + \mu_i (\overline{D} a_i^{(n)}) E_{1,2i+1}], \end{aligned} \tag{87}$$

where E_{ij} are the matrices with entries $(E_{ij})_{kl} = \delta_{ik}\delta_{jl}$ and

$$\lambda_i = \frac{(n-i)!}{n!(i-1)!}, \quad \mu_i = \frac{(n-i)!}{n!\,i!}.$$

This gauge choice is referred to as the *horizontal gauge*[15] while (85), (86) is known as the $osp(1|2)$ *highest weight gauge*. In fact, as we will now verify, the combination $Q_n + \overline{Q}_n$ leads to

the matrix representation of $N=1$ superconformally covariant differential operators in terms of highest weight generators of $osp(1|2)_{\rm pal}\subset sl(2|1)_{\rm pal}$ (see Appendix A for the algebraic details).

By adding the expressions (86), we find

$$Q_n+\overline{Q}_n=-\kappa_-+(D+\overline{D})\mathbf{1}+\frac{1}{2}\sum_{i=1}^{n}[i\mathcal{V}_{i,i}M_{2i-1}-(D-\overline{D})\mathcal{V}_{i,i}M_{2i}].$$

Here, $\kappa_-\equiv\mu_-+\overline{\mu}_-\in osp(1|2)_{\rm pal}$ and $M_k\equiv(\mu_+-\overline{\mu}_+)^k$ are highest weight generators of $osp(1|2)_{\rm pal}$ while $D+\overline{D}=D_1$ denotes the basic $N=1$ derivative. By projecting from $N=2$ to $N=1$ superfields along the lines of section IV D 3, we get

$$Q_n^{(N=1)}\equiv(Q_n+\overline{Q}_n)|_{N=1}=-\kappa_-+D_1\mathbf{1}+\sum_{k=1}^{2n}V_{k+1}M_k, \tag{88}$$

with

$$V_{2i}=\frac{i}{2}\,\mathcal{V}_{i,i}\Big|_{N=1},\quad V_{2i+1}=-\frac{1}{2}(D-\overline{D})\mathcal{V}_{i,i}\Big|_{N=1}\qquad(i=1,\ldots,n).$$

The expression (88) coincides with the $N=1$ matrix representation found in Refs. 3,16 up to the fact that it involves a different (though equivalent) realization of the superprincipal embedding of $osp(1|2)_{\rm pal}$ in $sl(n+1|n)$. The same realization can be obtained by choosing Q_n and $\overline{Q}_n$ along the lines of Ref. 15:

$$\begin{aligned}Q_n'&=-\tilde{\mu}_++D\mathbf{1}+\frac{1}{2}\sum_{i=1}^{n}[(D\mathcal{V}_{i,i})\tilde{E}^i_{--}-\mathcal{V}_{i,i}[\tilde{\mu}_+,\tilde{E}^i_{--}]],\\ \overline{Q}_n'&=+\tilde{\overline{\mu}}_++\overline{D}\mathbf{1}-\frac{1}{2}\sum_{i=1}^{n}[(\overline{D}\mathcal{V}_{i,i})\tilde{E}^i_{--}+\mathcal{V}_{i,i}[\tilde{\overline{\mu}}_+,\tilde{E}^i_{--}]].\end{aligned} \tag{89}$$

Here, the tilde denotes the transposed matrix and the common structure of the results (86) and (89) becomes apparent if one takes into account the identities

$$[\mu_-,E^i_{++}]=-i\overline{\mu}_+E^{i-1}_{++},\quad [\overline{\mu}_-,E^i_{++}]=-i\mu_+E^{i-1}_{++}. \tag{90}$$

VII. WARD IDENTITIES AND *W*-ALGEBRAS FROM A ZERO CURVATURE CONDITION

The highest weight matrix representation (86) of covariant sandwich operators has the form

$$Q_n=D\mathbf{1}-A_\theta,\quad \overline{Q}_n=\overline{D}\mathbf{1}-A_{\bar\theta}, \tag{91}$$

where A_θ and $A_{\bar\theta}$ can be interpreted as the spinorial components of a connection with values in the Lie superalgebra $sl(2|1)_{\rm pal}\subset sl(n+1|n)$. The spinorial components can be supplemented with spatial components A_z and $A_{\bar z}$ represented by generic elements of $sl(n+1|n)$. Altogether these fields represent the fundamental variables for the superspace formulation of a (2,0)-supersymmetric gauge theory based on graded Lie algebras. (For the general framework of such theories, we refer to Ref. 16.)

Within the $N=1$ superfield formalism, it has been shown[16] that classical super W-algebras can be constructed in the following way: in the z-sector, one considers the highest weight gauge for the spinorial components of the connection [i.e., eqs. (86) in the present case], all entries of the connection components $A_z,A_{\bar z},A_\theta,\ldots$, are supposed to be smooth superfields and zero curvature conditions are imposed on the connection. We will now apply this procedure to the $N=2$ theory.

The integrability conditions for the system of differential equations $\vec{0}=Q_n\vec{f}=\overline{Q}_n\vec{f}=(\overline{\partial}-A_{\bar z})\vec{f}$ are the *zero curvature conditions*

$$
\begin{aligned}
&0 = DA_\theta - \hat{A}_\theta A_\theta, \quad 0 = \overline{D}A_{\bar{\theta}} - \hat{A}_{\bar{\theta}} A_{\bar{\theta}}, \\
&A_z = DA_{\bar{\theta}} + \overline{D}A_\theta - \hat{A}_\theta A_{\bar{\theta}} - \hat{A}_{\bar{\theta}} A_\theta, \\
&0 = DA_{\bar{z}} - \overline{\partial}A_\theta - A_\theta A_{\bar{z}} + \hat{A}_{\bar{z}} A_\theta, \\
&0 = \overline{D}A_{\bar{z}} - \overline{\partial}A_{\bar{\theta}} - A_{\bar{\theta}} A_{\bar{z}} + \hat{A}_{\bar{z}} A_{\bar{\theta}},
\end{aligned}
\tag{92}
$$

where the hat again denotes the automorphism of the Lie superalgebra which reverses the signs of all odd elements. As may be verified explicitly, the expressions (86) (involving smooth superfields $\mathcal{V}_{i,i}$) verify the constraints of the first line. The second line is simply a redefinition constraint which determines the component A_z in terms of the variables A_θ and $A_{\bar{\theta}}$. Finally, A_θ and $A_{\bar{\theta}}$ having the form (86), the last two conditions represent constraints for the entries of the matrix $A_{\bar{z}}$. Actually, most of these constraints are algebraic equations which allow us to express the entries of $A_{\bar{z}}$ in terms of other fields while the other ones represent partial differential equations for the independent fields. These differential equations are the Ward identities for the super W_n-algebras.

In the simplest case ($n=1$), the matrix $A_{\bar{z}}$ only contains one independent superfield H satisfying the differential equation

$$
\mathcal{L}_2^{sym} H = \overline{\partial}\mathcal{R}, \tag{93}
$$

where $\mathcal{L}_2^{sym}$ denotes the symmetric super Bol operator (53). This relation is nothing but the (2,0) *superconformal Ward identity*,[6] the superfield $H \equiv H_{\bar{z}}^{\ z}$ being interpreted as the Beltrami superfield (in the so-called restricted geometry) and the superprojective connection $\mathcal{R}$ as the stress-energy tensor. It expresses the superdiffeomorphism invariance of the generating functional $Z_c[H]$ in superconformal field theory.

For $n=2$, $A_{\bar{z}}$ contains two independent superfields $H \equiv H_{\bar{z}}^{\ z}$ and $G \equiv G_{\bar{z}}^{\ zz}$ which are associated to $\mathcal{R}$ and $\mathcal{V} \equiv \mathcal{V}_{2,2}$, respectively. These variables satisfy the system of differential equations

$$
\begin{aligned}
\overline{\partial}\mathcal{R} &= \mathcal{L}_2^{sym} H + [2\mathcal{V}\partial - (\overline{D}\mathcal{V})D - (D\mathcal{V})\overline{D} + 2(\partial\mathcal{V})]G, \\
\overline{\partial}\mathcal{V} &= \frac{-1}{16}[\mathcal{L}_4^{sym} + 4\mathcal{M}_{\mathcal{V}}^{(4)sym}]G + [2\mathcal{V}\partial - (\overline{D}\mathcal{V})D - (D\mathcal{V})\overline{D} + (\partial\mathcal{V})]H,
\end{aligned}
\tag{94}
$$

where $\mathcal{M}_{\mathcal{V}}^{(4)sym}$ is the following superconformally covariant operator. (This operator is an example of a covariant operator involving a conformal superfield and generalizing the symmetric Bol operators of section IV A 1. These operators can be determined in a systematic way along the lines of section V.)

$$
\begin{aligned}
\mathcal{M}_{\mathcal{V}}^{(4)sym} = {} & 5\mathcal{V}\partial[D,\overline{D}] + 5D\mathcal{V}\partial\overline{D} - 5\overline{D}\mathcal{V}\partial D + \tfrac{5}{2}\partial\mathcal{V}\,[D,\overline{D}] \\
& + (\tfrac{3}{2}[D,\overline{D}]\mathcal{V} + 7\mathcal{R}\mathcal{V})\partial + (3\partial D\mathcal{V} + \mathcal{R}D\mathcal{V} - 9D\mathcal{R}\mathcal{V})\overline{D} - (3\partial\overline{D}\mathcal{V} - \mathcal{R}\overline{D}\mathcal{V} + 9\overline{D}\mathcal{R}\mathcal{V})D \\
& + \partial[D,\overline{D}]\mathcal{V} + 3\mathcal{R}\partial\mathcal{V} - 5\overline{D}\mathcal{R}D\mathcal{V} - 5D\mathcal{R}\overline{D}\mathcal{V} + 8\partial\mathcal{R}\mathcal{V}.
\end{aligned}
$$

The relations (94) generalize the Ward identity (93) which is recovered for $\mathcal{V}=0=G$. They represent the *Ward identities associated to the $N=2$ super W_3-algebra* and they are manifestly covariant since all differential operators occurring on their r.h.s. are superconformally covariant when acting on the given fields.

As a matter of fact, the Ward identities (94) are equivalent to the *Poisson bracket algebra* (commutation relations)

$$
\begin{aligned}
[\mathcal{R}_2, \mathcal{R}_1] &= (\mathcal{L}_2^{sym})_1 \delta^{(3)}(\mathbf{z}_2, \mathbf{z}_1), \\
[\mathcal{R}_2, \mathcal{V}_1] &= (2\mathcal{V}\partial - (\overline{D}\mathcal{V})D - (D\mathcal{V})\overline{D} + (\partial\mathcal{V}))_1 \delta^{(3)}(\mathbf{z}_2, \mathbf{z}_1),
\end{aligned}
\tag{95}
$$

$$[\mathcal{V}_2,\mathcal{V}_1]=-\tfrac{1}{16}\,(\mathcal{L}_4^{sym}+4\mathcal{M}_{\mathcal{V}}^{(4)sym})_1\delta^{(3)}(\mathbf{z}_2,\mathbf{z}_1),$$

for which we used the notation $\mathcal{R}_i\equiv\mathcal{R}(\mathbf{z}_i)$ and $\mathcal{V}_i\equiv\mathcal{V}(\mathbf{z}_i)$. By spelling out all expressions on the r.h.s. of eqs. (95), one finds that this result exactly coincides with the one found in Ref. 17 by applying quite different methods (see also Refs. 9,18 as well as Ref. 11 for a general formula). The introduction of covariant operators not only allows us to cast the results in a compact form, it also helps to exhibit the conformal symmetry which plays a crucial role in the formulation of integrable models. In fact, the covariant operators occurring in eqs. (94), (95) yield the second Hamiltonian structure for the $N=2$ super Boussinesq equation.[17–19] We note that the algebra (95) can be rewritten in terms of $N=1$ superfields[18] in which case all $N=2$ covariant operators reduce to $N=1$ covariant ones (see Ref. 3 for the latter operators and the corresponding algebra).

By integrating the Ward identities (94), one can deduce the operator product expansions (OPE) of the spin 1 and spin 2 supercurrents $\mathcal{R}$ and $\mathcal{V}$. To do so, one proceeds along the lines of Ref. 16 and defines $\mathcal{R}$ and $\mathcal{V}$ as derivatives of a generating functional $Z_c=Z_c[H,G]$,

$$\mathcal{R}=-\left.\frac{\delta Z_c}{\delta H}\right|_{H=0,G=0},\qquad \mathcal{V}=-\left.\frac{\delta Z_c}{\delta G}\right|_{H=0,G=0}. \tag{96}$$

By using eqs. (B1) and (B4) of Appendix B, one obtains the (singular parts of the) *OPE of the classical $N=2$ super W_3-algebra*:[19] in terms of the notation $\mathcal{R}_i\equiv\mathcal{R}(\mathbf{z}_i)$ and $\mathcal{V}_i\equiv\mathcal{V}(\mathbf{z}_i)$, we have

$$\mathcal{R}_2\mathcal{R}_1=\frac{2}{z_{12}^2}+\left[\frac{\overline{\theta}_{12}\theta_{12}}{z_{12}^2}+\frac{\overline{\theta}_{12}}{z_{12}}\overline{D}-\frac{\theta_{12}}{z_{12}}D-\frac{\overline{\theta}_{12}\theta_{12}}{z_{12}}\partial\right]\mathcal{R}_1,$$

$$\mathcal{R}_2\mathcal{V}_1=\left[2\frac{\overline{\theta}_{12}\theta_{12}}{z_{12}^2}+\frac{\overline{\theta}_{12}}{z_{12}}\overline{D}-\frac{\theta_{12}}{z_{12}}D-\frac{\overline{\theta}_{12}\theta_{12}}{z_{12}}\partial\right]\mathcal{V}_1, \tag{97}$$

$$\begin{aligned}-16\mathcal{V}_2\mathcal{V}_1=&\frac{12}{z_{12}^4}+12\,\frac{\overline{\theta}_{12}\theta_{12}\mathcal{R}_1}{z_{12}^4}+12\,\frac{(\overline{\theta}_{12}\overline{D}-\theta_{12}D-\overline{\theta}_{12}\theta_{12}\partial)\mathcal{R}_1}{z_{12}^3}\\&+\frac{2(20\mathcal{V}_1+2[D,\overline{D}]\mathcal{R}_1-\mathcal{R}_1^2)-(\overline{\theta}_{12}\overline{D}+\theta_{12}D)(20\mathcal{V}_1+8[D,\overline{D}]\mathcal{R}_1-\mathcal{R}_1^2)}{z_{12}^2}\\&+\frac{\overline{\theta}_{12}\theta_{12}(6[D,\overline{D}]\mathcal{V}_1+6\partial^2\mathcal{R}_1+U_1)}{z_{12}^2}-\frac{\partial(20\mathcal{V}_1+2[D,\overline{D}]\mathcal{R}_1-\mathcal{R}_1^2)}{z_{12}}\\&+\frac{\overline{\theta}_{12}(12\partial\overline{D}\mathcal{V}_1+3\partial^2\overline{D}\mathcal{R}_1+\overline{\psi}_1)+\theta_{12}(12\partial D\mathcal{V}_1-3\partial^2 D\mathcal{R}_1+\psi_1)}{z_{12}}\\&-\frac{1}{4}\,\overline{\theta}_{12}\theta_{12}\frac{16\partial[D,\overline{D}]\mathcal{V}_1+8\partial^3\mathcal{R}_1+\partial[D,\overline{D}]\mathcal{R}_1^2+2(\partial\mathcal{U}_1-\overline{D}\psi_1+D\overline{\psi}_1)}{z_{12}},\end{aligned}$$

which expressions involve the composite supercurrents $\mathcal{R}^2$ and $\mathcal{U},\overline{\psi},\psi$ defined by

$$\mathcal{U}=28\mathcal{R}\mathcal{V}+4\mathcal{R}[D,\overline{D}]\mathcal{R}-15D\mathcal{R}\overline{D}\mathcal{R}-2\mathcal{R}^3,$$

$$\overline{\psi}=-4\mathcal{R}\overline{D}\mathcal{V}+36\overline{D}\mathcal{R}\mathcal{V}-\mathcal{R}\partial\overline{D}\mathcal{R}-\tfrac{3}{2}\,\partial\mathcal{R}\overline{D}\mathcal{R}+\tfrac{9}{2}[D,\overline{D}]\mathcal{R}\overline{D}\mathcal{R}-2\mathcal{R}^2\overline{D}\mathcal{R},$$

$$\psi=4\mathcal{R}D\mathcal{V}-36D\mathcal{R}\mathcal{V}-\mathcal{R}\partial D\mathcal{R}-\tfrac{3}{2}\,\partial\mathcal{R}D\mathcal{R}-\tfrac{9}{2}[D,\overline{D}]\mathcal{R}D\mathcal{R}+2\mathcal{R}^2D\mathcal{R}.$$

In conclusion, we note that one can explicitly derive the Poisson brackets (95) from the OPE (97) by applying the supersymmetric version of Cauchy's theorem (Appendix B).[16]

VIII. CONCLUDING REMARKS

Interestingly enough, the zero curvature conditions provide a link between two different classes of $N=2$ superconformally covariant operators: by starting from the matrix representation of sandwich operators, one obtains Ward identities involving symmetric covariant operators.

Concerning the matrix representation of covariant operators which are not of sandwich type we remark the following: if one disregards the covariance properties and considers $\mathcal{L}_n^{sym}$ as sandwiched between $\overline{D}$ and D, it has the form (77) and equation (87) with properly chosen coefficients $a_i^{(n)}$ then provides a matrix representation. However, this result is neither satisfactory nor complete, since it does not properly take into account conformal symmetry and since the scalar operator $\mathcal{L}_n^{sym}$ can only be recovered from this matrix representation up to contributions $\mathcal{L}_a$ satisfying $\overline{D}\mathcal{L}_a D=0$.

The expression $\mathcal{M}_{\mathcal{W}}^{(n)}C$ may be viewed as the result of a bilinear and covariant map J, i.e., $\mathcal{M}_{\mathcal{W}}^{(n)}C=J(\mathcal{W},C)$ which is known as the super Gordan transvectant.[1–3] Such bilinear as well as trilinear covariant operators[3] can be constructed along the lines of section III E. These operators occur in particular in the Poisson brackets of super W-algebras.[3]

There is a remarkable relationship between the matrix representation of conformally covariant operators and the general formula determining the singular vectors of the Virasoro algebra—see the second of Refs. 2. This relationship generalizes to the $N=1$ supersymmetric theory[3] and our results concerning $N=2$ covariant operators should also allow us to draw close parallels to the recent study of singular vectors of the $N=2$ superconformal algebra.[20]

Apart from the Gelfand-Dickey derivation of (super) W-algebras and the zero curvature construction that we discussed here, various other approaches have been considered in the literature—see for instance Refs. 21,22. The tools developed in the present work should provide the necessary ingredients for generalizing these constructions to the $N=2$ theory.

Finally, we note that our results can be extended—along the lines of Refs. 4 and 23—to the case of pseudo-differential operators which appear for instance as Lax operators in superintegrable models.[24]

Note added in proof: As we realized after completion of the present work, a first attempt to construct $N=2$ superconformally covariant operators has been made some time ago by Mathieu:[25] in particular, this author found the form of the lowest order, symmetric super Bol operators in a nonchiral basis of $N=2$ superspace.

ACKNOWLEDGMENTS

We wish to thank François Delduc for informing us about their results prior to publication, for numerous stimulating discussions and precious advice. F.G. acknowledges financial support by the Alexander von Humboldt-Foundation and expresses his gratitude to Stefan Theisen for the hospitality extended to him in Munich at the final stage of this work.

APPENDIX A: NONSTANDARD MATRIX REALIZATION OF $sl(n+1|n)$

$sl(n+1|n)$ is the graded Lie algebra of $(2n+1)\times(2n+1)$ matrices with a vanishing supertrace. It admits a purely *fermionic* system of simple roots.

We will consider a nonstandard matrix realization of $sl(n+1|n)$ which is referred to as the *diagonal representation*.[3,16] It consists of assigning a $\mathbf{Z}_2$-grading $i+j\,(\mathrm{mod}\ 2)$ to a matrix element M_{ij}, defining the supertrace by the alternating sum

$$\mathrm{str}\ M=\sum_{i=1}^{2n+1}(-1)^{i+1}M_{ii},$$

and the graded commutator by

$$[M,N\}_{ik}=\sum_{j=1}^{2n+1}(M_{ij}N_{jk}-(-1)^{(i+j)(j+k)}N_{ij}M_{jk}).$$

In the Serre-Chevalley basis, the basic commutation relations of $sl(n+1|n)$ read as

$$[h_i,h_j]=0, \quad [h_i,e_j]=+a_{ij}e_j,$$

$$\{e_i,f_j\}=\delta_{ij}h_j, \quad [h_i,f_j]=-a_{ij}f_j, \quad \text{for } i,j\in\{1,..,2n\}.$$

Here, the h_i belong to the Cartan subalgebra, the e_i denote the fermionic simple roots, f_i the associated negative roots and the a_{ij} are the elements of the Cartan matrix:

$$a_{ij}=\delta_{i+1,j}-\delta_{i,j+1}. \tag{A1}$$

In the diagonal representation, the Serre-Chevalley generators can be represented by the matrices

$$h_i=E_{i,i}+E_{i+1,i+1}, \quad e_i=E_{i,i+1}, \quad f_i=E_{i+1,i}, \tag{A2}$$

where E_{ij} denotes the $(2n+1)\times(2n+1)$ matrix with entries $(E_{ij})_{kl}=\delta_{ik}\delta_{jl}$.

The superprincipal embedding of $sl(2|1)$ in $sl(n+1|n)$ is denoted by $sl(2|1)_{\text{pal}}\subset sl(n+1|n)$ and is defined as follows:[15]

$$H=\sum_{i=1}^{n}(n-i+1)h_{2i-1}, \quad \bar{H}=-\sum_{i=1}^{n} ih_{2i},$$

$$\mu_+=\sum_{i=1}^{n} e_{2i-1}, \quad \bar{\mu}_+=-\sum_{i=1}^{n} e_{2i}, \tag{A3}$$

$$\mu_-=\sum_{i=1}^{n}(n-i+1)f_{2i-1}, \quad \bar{\mu}_-=\sum_{i=1}^{n} if_{2i}.$$

Here, $H,\bar{H}$ represent the Cartan generators of $sl(2|1)$, $\mu_+,\bar{\mu}_+$ the fermionic simple roots and $\mu_-,\bar{\mu}_-$ the associated negative roots. The nonvanishing commutators of these generators read as

$$[H,\bar{\mu}_\pm]=\pm\bar{\mu}_\pm, \quad [\bar{H},\mu_\pm]=\pm\mu_\pm,$$

$$\{\mu_+,\mu_-\}=H, \quad \{\bar{\mu}_+,\bar{\mu}_-\}=\bar{H}, \tag{A4}$$

$$\{\mu_+,\bar{\mu}_+\}\equiv -E_{++}, \quad \{\mu_-,\bar{\mu}_-\}\equiv -E_{--},$$

where E_{++} and E_{--} correspond to the bosonic roots of $sl(2|1)$.

The Lie superalgebra $sl(2|1)_{\text{pal}}$ contains the superprincipal embedding of $osp(1|2)$ into $sl(n+1|n)$, i.e., $osp(1|2)_{\text{pal}}\subset sl(2|1)_{\text{pal}}\subset sl(n+1|n)$: this embedding is defined by the combinations

$$J_0=H+\bar{H}, \quad \kappa_+=\mu_++\bar{\mu}_+, \quad \kappa_-=\mu_-+\bar{\mu}_-,$$
$$J_+=E_{++}, \quad J_-=-E_{--}, \tag{A5}$$

which satisfy the commutation relations

$$[J_0,\kappa_\pm]=\pm\kappa_\pm, \quad \{\kappa_+,\kappa_-\}=J_0, \quad \{\kappa_\pm,\kappa_\pm\}=\pm 2J_\pm. \tag{A6}$$

Here, J_0 is the Cartan generator and κ_+ the fermionic simple root. The highest weight generators M_k of $osp(1|2)_{\text{pal}}$ are defined by the relations

$$[J_0,M_k]=kM_k, \quad [\kappa_+,M_k\}=0 \quad (k=1,2,\ldots,2n), \tag{A7}$$

which are solved by

$$M_k=M_1^k, \quad \text{with } M_1=\mu_+-\bar{\mu}_+. \tag{A8}$$

APPENDIX B: SOME DISTRIBUTIONAL RELATIONS

In this appendix, we gather some distributional relations which hold in $d=2, N=2$ superspace and we use them to determine the correspondence rules between supercurrent OPE's and commutation relations.

Let $(\mathbf{z}_i,\overline{\mathbf{z}_i})\equiv(z_i,\theta_i,\overline{\theta_i};\overline{z_i},\theta_i^-,\overline{\theta_i^-})$ with $i=1,2$ denote the local coordinates of two points on a $N=2$ SRS. We then define the relative coordinates,

$$\theta_{12}=\theta_1-\theta_2,\quad \overline{\theta}_{12}=\overline{\theta_1}-\overline{\theta_2},\quad z_{12}=z_1-z_2-\tfrac{1}{2}(\theta_1\overline{\theta_2}+\overline{\theta_1}\theta_2). \tag{B1}$$

For a superanalytic superfield F (i.e., a field depending only on $\mathbf{z}$), one has the following Cauchy formulas[26,9] ($n\in\mathbf{N}$):

$$\int d^3\mathbf{z}_1(\overline{\theta}_{12}\theta_{12},\theta_{12},\overline{\theta}_{12},1)\frac{F(\mathbf{z}_1)}{z_{12}^{n+1}}=\left(1,-\overline{D},D,\frac{1}{2}[D,\overline{D}]\right)\frac{\partial^n F(\mathbf{z}_2)}{n!}. \tag{B2}$$

Here, $d^3\mathbf{z}$ denotes the measure $(2\pi i)^{-1}dz\,d\theta\,d\overline{\theta}$ with respect to which the Dirac distribution takes the form

$$\delta^{(3)}(\mathbf{z}_1,\mathbf{z}_2)=\delta^{(1)}(z_1-z_2)(\overline{\theta_1}-\overline{\theta_2})(\theta_1-\theta_2)=\frac{1}{z_{12}}\,\overline{\theta}_{12}\theta_{12}. \tag{B3}$$

The integration of Ward identities is performed by means of the Dirac distribution with respect to the measure $d^4\mathbf{z}=(2\pi i)^{-1}d\overline{z}dz\,d\theta\,d\overline{\theta}$:

$$\delta^{(4)}(\mathbf{z}_1,\mathbf{z}_2)=\delta^{(2)}(z_1-z_2,\overline{z_1}-\overline{z_2})(\overline{\theta_1}-\overline{\theta_2})(\theta_1-\theta_2)=\partial_{z_1}^-\frac{1}{z_{12}}\,\overline{\theta}_{12}\theta_{12}. \tag{B4}$$

A general OPE of two supercurrents $A(\mathbf{z}_2)$ and $B(\mathbf{z}_1)$ can be written in the form

$$A(\mathbf{z}_2)B(\mathbf{z}_1)=\sum_{n\in\mathbf{N}}\frac{a_n(\mathbf{z}_1)+\overline{\theta}_{12}\alpha_n(\mathbf{z}_1)+\theta_{12}\beta_n(\mathbf{z}_1)+\overline{\theta}_{12}\theta_{12}b_n(\mathbf{z}_1)}{z_{21}^{n+1}}+\text{regular terms}, \tag{B5}$$

and by virtue of the relations (B2), this OPE is equivalent to the commutation relation

$$[A(\mathbf{z}_2),B(\mathbf{z}_1)]=\sum_{n\in\mathbf{N}}\frac{1}{n!}\left(\frac{a_n(\mathbf{z}_1)}{2}[D,\overline{D}]+\alpha_n(\mathbf{z}_1)D-\beta_n(\mathbf{z}_1)\overline{D}+b_n(\mathbf{z}_1)\right)\partial_1^n\delta^{(3)}(\mathbf{z}_2,\mathbf{z}_1). \tag{B6}$$

[1] F. Gieres, Int. J. Mod. Phys. A **8**, 1 (1993).
[2] P. Di Francesco, C. Itzykson, and J.-B. Zuber, Commun. Math. Phys. **140**, 543 (1991); C. Itzykson, Int. J. Mod. Phys. B **6**, 1969 (1992).
[3] F. Gieres and S. Theisen, J. Math. Phys. **34**, 5964 (1993).
[4] W.-J. Huang, J. Math. Phys. **35**, 2570 (1994).
[5] H. Lu, C. N. Pope, L. J. Romans, X. Shen, and X.-J. Wang, Phys. Lett. B **264**, 91 (1991); J. M. Figueroa-O'Farill and E. Ramos, *ibid.* **262**, 265 (1991); J. M. Figueroa-O'Farill and E. Ramos, Nucl. Phys. B **368**, 361 (1992); T. Inami and H. Kanno, J. Phys. A **25**, 3729 (1992).
[6] F. Delduc, F. Gieres, and S. Gourmelen, Class. Quantum Grav. **14**, 1623 (1997).
[7] J. D. Cohn, Nucl. Phys. B **284**, 349 (1987); H. Aratyn, E. Nissimov, S. Pacheva, and A. H. Zimerman, Phys. Lett. B **240**, 127 (1990).
[8] W. Scherer, Lett. Math. Phys. **17**, 45 (1989).
[9] Z. Popowicz, Int. J. Mod. Phys. A **9**, 2001 (1994); Z. Popowicz, Phys. Lett. B **319**, 478 (1993).
[10] C.-A. Laberge and P. Mathieu, Phys. Lett. B **215**, 718 (1988); P. Labelle and P. Mathieu, J. Math. Phys. **32**, 923 (1991).
[11] F. Delduc and L. Gallot, Commun. Math. Phys. **190**, 395 (1997).
[12] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis* (Academic, New York, 1980).
[13] V. G. Drinfel'd and V. V. Sokolov, J. Sov. Math **30**, 1975 (1985).
[14] F. Delduc, E. Ragoucy, and P. Sorba, Commun. Math. Phys. **146**, 403 (1992).
[15] F. Delduc and M. Magro, J. Phys. A **29**, 4987 (1996); M. Magro, ''Modèles de Wess-Zumino-Novikov-Witten et de Toda supersymétriques $N=2$,'' Thèse de doctorat, ENS Lyon, February 1996.

[16] F. Gieres and S. Theisen, Int. J. Mod. Phys. A **9**, 383 (1994).
[17] S. Bellucci, E. Ivanov, S. Krivonos, and A. Pichugin, Phys. Lett. B **312**, 463 (1993).
[18] C. M. Yung, Phys. Lett. B **309**, 75 (1993).
[19] E. Ivanov and S. Krivonos, Phys. Lett. B **291**, 63 (1992).
[20] A. M. Semikhatov, ''Verma modules, extremal vectors and singular vectors on the non-critical $N=2$ string worldsheet,'' hep-th/9610084.
[21] R. Zucchini, Class. Quantum Grav. **10**, 253 (1993); J.-P. Ader, F. Biet, and Y. Noirot, Nucl. Phys. B **466**, 285 (1996).
[22] G. Akemann and R. Grimm, J. Math. Phys. **34**, 818 (1993); D. Garajeu, R. Grimm, and S. Lazzarini, *ibid.* **36**, 7043 (1995).
[23] W.-J. Huang, J. Math. Phys. **35**, 993 (1994); W.-J. Huang, J. C. Shaw, and H. C. Yen, J. Phys. A **28**, 165 (1995).
[24] S. Gourmelen, work in progress.
[25] P. Mathieu, ''Super Miura transformations, super Schwarzian derivatives and super Hill operators,'' in *Integrable and Superintegrable Systems*, edited by B. A. Kupershmidt (World Scientific, Singapore, 1990).
[26] P. West, *Introduction to Supersymmetry and Supergravity*, 2nd ed. (World Scientific, Singapore, 1990).

Anti-isospectral transformations, orthogonal polynomials, and quasi-exactly solvable problems

Avinash Khare[a] and Bhabani Prasad Mandal[b]
Institute of Physics, Sachivalaya Marg, Bhubaneswar-751005, India

(Received 1 December 1997; accepted for publication 23 December 1997)

We consider the double sinh-Gordon potential which is a quasi-exactly solvable problem and show that in this case one has two sets of Bender–Dunne orthogonal polynomials. We study in some detail the various properties of these polynomials and the corresponding quotient polynomials. In particular, we show that the weight functions for these polynomials are *not* always positive. We also study the orthogonal polynomials of the double sine-Gordon potential which is related to the double sinh-Gordon case by an anti-isospectral transformation. Finally we discover a new quasi-exactly solvable problem by making use of the anti-isospectral transformation. © *1998 American Institute of Physics.* [S0022-2488(98)03205-8]

I. INTRODUCTION

Recently, in an interesting paper, Krajewska *et al.*[1] have introduced an anti-isospectral transformation (which they also called duality transformation). In particular, using this transformation they relate the spectra of quasi-exactly solvable (QES) potentials V_1 and V_2 given by

$$V_1(x)=x^2(ax^2+b)^2-\hbar a(2M+3)x^2, \tag{1.1}$$

$$V_2(x)=x^2(ax^2-b)^2-\hbar a(2M+3)x^2, \tag{1.2}$$

where $a,b>0$ and M is non-negative integer. It may be noted here that the duality property does not hold for other (exactly non-calculable) levels. It is clearly of great interest to understand this new transformation in some detail and explore its various consequences. For example, one would like to know whether one can discover new QES problems by using this transformation. Second, are the QES levels of the dual potentials related even if they are valid over different domain? Further, whether the number of QES levels in the two cases are identical or not.

Another recent development is the work of Bender–Dunne[2] where they have shown that the eigenfunctions of the Schrödinger equation for a quasi-exactly solvable (QES) problem is the generating function for a set of orthogonal polynomials $\{P_n(E)\}$ in the energy variable E. It was further shown in one specific example that these polynomials satisfy the three-term recursion relation

$$P_n(E)=EP_{n-1}(E)+C_nP_{n-2}(E), \tag{1.3}$$

where C_n is E independent quantity. Using the well known theorem,[3,4] ''the *necessary and sufficient* condition for a family of polynomials $\{P_n\}$ (with degree $P_n=n$) to form an orthogonal polynomial system is that $\{P_n\}$ satisfy a three-term recursion relation of the form

$$P_n(E)=(A_nE+B_n)P_{n-1}(E)+C_nP_{n-2}(E), \quad n\geqslant 1, \tag{1.4}$$

where the coefficients A_n, B_n, and C_n are independent of E, $A_n\neq 0$, $C_1=0$, $C_n\neq 0$ for $n\geqslant 1$,'' it then followed that $\{P_n(E)\}$ for this problem forms an orthogonal set of polynomials with respect to some weight function, $w(E)$. Recently several authors have studied the Bender–Dunne polynomials in detail.[1,4–6] In fact it has been claimed that the Bender–Dunne construction is quite universal and valid for any quasi-exactly solvable model in both one as well as multi-dimensions.

[a]Electronic-mail: khare@iopb.stpbh.soft.net
[b]Present address: Theory Group, Saha Institute of Nuclear Physics, I/AF, Bidhannagar, Calcutta-64, India

 © 1998 American Institute of Physics

However, in a recent note[7] we have discussed three QES problems for all of which the Schrödinger equation can be transformed to Heun's equation and further in all these cases the three-term recursion relation satisfied by the Bender–Dunne polynomials is not of the form as given by Eq. (1.4) and hence does not form an orthogonal set. It is worth pointing out that in all these cases the Hamiltonian cannot be written in terms of quadratic generator of the Sl(2) algebra. We suspect that this may be the reason why the polynomials in these examples do not form an orthogonal set.

The question which we would like to raise and study in this paper is regarding the properties of the Bender–Dunne polynomials in case they form an orthogonal set. For example, in the example discussed by Bender–Dunne (as well as in most other QES examples discussed so far), for a given potential, the QES states have either even or odd number of nodes. The question is what happens in a QES problem if one can obtain states with even as well as odd number of nodes. In particular, do the polynomials corresponding to the even and odd number of nodes together form an orthogonal set or if the polynomials corresponding to even number of nodes form one orthogonal set and polynomials with odd number of nodes form a separate orthogonal set? Second, how universal are the properties of Bender–Dunne orthogonal polynomials in such cases? For example, are the weight functions always positive? Do the moments of weight function have pure power growth?[4]

The purpose of this note is to explore in some detail the various issues raised above. In particular, in Sec. II, we discuss the double sinh-Gordon (DSHG) model which is an QES problem but where states with odd as well as even number of nodes are known for a given potential. We show that in this case there are two sets of orthogonal polynomials with polynomials corresponding to even number of nodes forming one orthogonal set while the polynomials corresponding to odd number of nodes forming another orthogonal set. We study the properties of these polynomials as well as the corresponding quotient polynomials in detail and show that in many respect they are similar to the Bender–Dunne polynomials. However, unlike the Bender–Dunne case, the weight functions are not always positive for either of the orthogonal polynomial sets. In Sec. III, we discuss several consequences of the duality symmetry in the context of the QES problems. In particular, we study two QES problems, in both of which the domain of validity of the potential and it's dual are different. As a result we show that in the case of the double sine-Gordon (DSG) potential, the number of QES levels are almost half of those in the double sinh-Gordon case (which is dual to it). However, even then many of the predictions about their corresponding Bender–Dunne polynomials still go through. For example, the QES levels of DSG are still related to the corresponding levels of DSGH case. Further, the weight functions here are also not always positive. Finally using duality, we discover a new QES problem which has so far not been discussed in the literature. Section IV is reserved for discussions.

II. DSHG SYSTEM AND ORTHOGONAL POLYNOMIALS

The DSHG system is one of the few double well problems in quantum mechanics which is quasi-exactly solvable.[8,9] This double well system has found application in several different branches of physics starting from the theory of diffusion in bistable field,[10] quantum theory of instantons,[11,12] quantum theory of molecules and also as a model for nonlinear coherent structure.[13,14] The Hamiltonian for this case is given by (we shall assume $\hbar=2m=1$ throughout this paper)

$$H=-\frac{d^2}{dx^2}+(\zeta\cosh 2x-M)^2, \tag{2.1}$$

where ζ is a positive parameter. The value of M is not restricted in principle but it has been shown[8] that the solutions for first M levels are exactly known in case M is a positive integer. Note further that for $M>\zeta$, this potential has a double well structure with the two minima lying at $\cosh 2x_0=M/\zeta$. Let us now derive the three-term recursion relation in this case. On substituting

$$\psi(x)=e^{-(\zeta/2)\cosh 2x}\phi(x) \tag{2.2}$$

in the Schrödinger equation $H\psi=E\psi$ with H as given by Eq. (2.1) we obtain

$$\phi''(x)-2\zeta\ \sinh 2x\ \phi'(x)+[(E-M^2-\zeta^2)+2(M-1)\zeta\ \cosh 2x]\phi(x)=0. \tag{2.3}$$

On further substituting

$$z=\cosh 2x-1; \quad \phi=z^s\sum_{n=0}^{\infty}\ \frac{R_n(E)}{n!}\left(\frac{z+2}{2}\right)^{n/2} \tag{2.4}$$

we obtain the three-term recursion relation ($n\geqslant 0$).

$$R_{n+2}(E)-[n^2+4(s+\zeta)n+4s^2+E-M^2-\zeta^2-2(M-1)\zeta]R_n(E)$$
$$-4\zeta[M+1-2s-n]n(n-1)R_{n-2}(E)=0 \tag{2.5}$$

provided $2s^2=s$, i.e., either $s=0$ or $s=\frac{1}{2}$. Thus we have two sets of independent solutions: the even states (i.e., states with even number of nodes) for $s=0$ and the odd states for $s=\frac{1}{2}$. Note that unlike the Bender–Dunne (or most other QES) cases, s is not contained in the potential and this is perhaps related to the fact that for any integer M ($\geqslant 2$) the QES solutions corresponding to both even and odd states are obtained. From Eq. (2.5) we observe that the even and odd polynomials $R_n(E)$ do not mix with each other and hence we have two separate three-term recursion relations depending on whether n is odd or even. In particular, it is easily shown that the three-term recursion relations corresponding to the even and odd n cases, respectively, are given by ($n\geqslant 1$)

$$P_n(E)-[4n^2+8n(s+\zeta-1)+4s^2-8s+4-6\zeta+E-(M+\zeta)^2]P_{n-1}(E)$$
$$-8\zeta(n-1)(2n-3)(M+3-2s-2n)P_{n-2}=0, \tag{2.6}$$

$$Q_n(E)-[4n^2+4n(2s+2\zeta-1)+4s^2-4s+1-2\zeta+E-(M+\zeta)^2]Q_{n-1}(E)$$
$$-8\zeta(n-1)(2n-1)(M+2-2s-2n)Q_{n-2}=0, \tag{2.7}$$

with $P_0(E)=1$, $Q_0(E)=1$. These recursion relations generate a set of monic polynomials, of which the first few are

$$P_0(E)=1, \quad P_1(E)=\mathscr{E}+4s^2+2\zeta, \tag{2.8}$$

$$P_2(E)=\mathscr{E}^2+\mathscr{E}[8s^2+8s+12\zeta+4]+(4s^2+4s+10\zeta+4)(4s^2+2\zeta)+8\zeta(M-2s-1),$$

and

$$Q_0(E)=1, \quad Q_1(E)=\mathscr{E}+4s^2+4s+1+6\zeta, \tag{2.9}$$

$$Q_2(E)=\mathscr{E}^2+\mathscr{E}[8s^2+8s+20\zeta+10]+(4s^2+4s+6\zeta+1)(4s^2+4s+14\zeta+9)$$
$$+24\zeta(M-2s-2),$$

where

$$\mathscr{E}\equiv E-(M+\zeta)^2. \tag{2.10}$$

Note that unlike the Bender–Dunne case, the polynomials $P_n(E)$ and $Q_n(E)$ are not eigenfunctions of parity. It is easily seen that when M is a positive integer, exact solutions for first M levels are obtained. In particular, if M is odd (even) integer, then solutions with even number of nodes ($s=0$) are obtained when the coefficient of $P_{n-2}(Q_{n-2})$ vanishes. Similarly if M is odd (even) integer, the solutions with odd number of nodes ($s=\frac{1}{2}$) are obtained when the coefficient of $Q_{n-2}(P_{n-2})$ vanishes. Further for M even (say $2k+2$, $k=0,1,2,...$), half the levels, i.e., $(k+1)$ levels are obtained each from the zeros of the orthogonal polynomials $P_{k+1}(E)$ and $Q_{k+1}(E)$. On the other hand, when M is odd (say $2k+1$, $k=0,1,...$) then $k+1$ and k levels are obtained from the zeros of the orthogonal polynomials $P_{k+1}(E)$ and $Q_k(E)$, respectively.

We have studied the properties of the polynomial sets $\{P_n(E)\}$ and $\{Q_n(E)\}$ and we find that many of their properties are almost the same and in turn they are similar to the Bender–Dunne polynomials. In particular, since in both the cases the recursion relations are similar to those given by Eq. (1.4), hence for all values of the parameters M and s, they form an orthogonal set. Second, the wave function $\psi(x,E)$ is the generating function for the polynomials $P_n(E)$ as well as $Q_n(E)$. Third, when M is a positive integer, both of these polynomials exhibit factorization property whose precise form depends on whether M is even or odd. In particular, if $M=2k+1$, $k=0,1,2,...$ then one has $(n\geqslant 0)$

$$P_{k+1+n}(E)=P_{k+1}(E)\overline{P}_n(E),$$
$$Q_{k+n}(E)=Q_k(E)\overline{Q}_n(E). \tag{2.11}$$

On the other hand, if $M=2k+2$, $k=0,1,2,...$, then

$$P_{k+1+n}(E)=P_{k+1}(E)\overline{R}_n(E),$$
$$Q_{k+1+n}(E)=Q_{k+1}(E)\overline{S}_n(E), \tag{2.12}$$

where $\overline{P}_0(E)=\overline{Q}_0(E)=\overline{R}_0(E)=\overline{S}_0(E)=1$. Following Ref. 5, it is easily shown that the polynomial sets $\{P_n(E)\}$, $\{Q_n(E)\}$, $\{R_n(E)\}$ and $\{S_n(E)\}$ correspond to the non-exact spectrum for this problem.

To illustrate this factorization we list in factored form, the first few polynomials of both the types in case $M=3$ and $M=4$.

$$M=3:$$
$$P_0(E)=1,$$
$$P_1(E)=\mathscr{E}+2\zeta,$$
$$P_2(E)=\mathscr{E}^2+\mathscr{E}[12\zeta+4]+20\zeta^2+24\zeta, \tag{2.13}$$
$$P_3(E)=[\mathscr{E}+18\zeta+16]P_2(E),$$
$$P_4(E)=[\mathscr{E}^2+\mathscr{E}(46\zeta+52)+(36+28\zeta)(16+18\zeta)-240\zeta]P_2(E).$$

$$Q_0(E)=1,$$
$$Q_1(E)=\mathscr{E}+6\zeta+4,$$
$$Q_2(E)=\left[\mathscr{E}+14\zeta+\frac{35}{2}\right]Q_1(E), \tag{2.14}$$
$$Q_3(E)=[\mathscr{E}^2+\mathscr{E}(36\zeta+52)+(36+22\zeta)(16+14\zeta)-160\zeta]Q_1(E),$$

where $\mathscr{E}$ is given by Eq. (2.6). Let us notice that in this case $P_2(E)$, is a common factor of $P_n(E)$ for $n\geqslant 2$. The zeros of $P_2(E)$ are at

$$E_0=7+\zeta^2-2\sqrt{1+4\zeta^2},$$
$$E_2=7+\zeta^2+2\sqrt{1+4\zeta^2}, \tag{2.15}$$

which give the ground and second excited state eigenvalues for the potential when $M=3$. On the other hand $Q_1(E)$ is a common factor of $Q_n(E)$ for $n\geqslant 1$. The zeros of $Q_1(E)$ are at

$$E_1=\zeta^2+5, \tag{2.16}$$

which give the first excited state eigenvalue for the same potential (i.e., when $M=3$). In this way for $M=3$ one obtains exact energy eigenvalues for the first three levels.

$$M=4:$$

$$P_0(E)=1,$$

$$P_1(E)=\mathscr{E}+2\zeta+1,$$

$$P_2(E)=\mathscr{E}^2+\mathscr{E}[12\zeta+10]+20\zeta^2+44\zeta+9, \tag{2.17}$$

$$P_3(E)=[\mathscr{E}+18\zeta+25]P_2(E),$$

$$P_4(E)=[\mathscr{E}^2+\mathscr{E}(46\zeta+64)+(28\zeta+49)(18\zeta+25)-240\zeta]P_2(E).$$

$$Q_0(E)=1,$$

$$Q_1(E)=\mathscr{E}+6\zeta+1,$$

$$Q_2(E)=\mathscr{E}^2+\mathscr{E}[20\zeta+10]+84\zeta^2+116\zeta+9, \tag{2.18}$$

$$Q_3(E)=[\mathscr{E}+22\zeta+25]Q_2(E),$$

$$Q_4(E)=[\mathscr{E}^2+\mathscr{E}(52\zeta+74)+(30\zeta+49)(22\zeta+25)-336\zeta]Q_2(E).$$

In this case $P_2(Q_2)$ is a common factor of $P_n(Q_n)$ for $n\geqslant 2$. The zeros of $P_2(E)$ are at

$$\begin{aligned} E_1&=\zeta^2+2\zeta+11-4\sqrt{\zeta^2+\zeta+1},\\ E_3&=\zeta^2+2\zeta+11+4\sqrt{\zeta^2+\zeta+1}, \end{aligned} \tag{2.19}$$

which give the first and the third excited state energies for the potential when $M=4$. On the other hand the zeros of $Q_2(E)$ are at

$$\begin{aligned} E_0&=\zeta^2-2\zeta+11-4\sqrt{\zeta^2-\zeta+1},\\ E_2&=\zeta^2-2\zeta+11+4\sqrt{\zeta^2-\zeta+1}, \end{aligned} \tag{2.20}$$

which give the ground and the second excited state energies for the same potential ($M=4$). The corresponding eigenfunctions can be easily obtained by evaluating $\psi(x)$ in Eq. (2.4) at these values of E and not surprisingly they are the same as given by Razavy[8] (with appropriate change of parameters).

The norms (squared) of the orthogonal set of polynomials $P_n(E)$ and $Q_n(E)$ can be easily determined by using the recursion relations and we find (assuming $\gamma_0^{(p)}=\gamma_0^{(q)}=1$)

$$\begin{aligned} \gamma_n^{(p)}&=(-8\zeta)^n n!\prod_{k=1}^{n}[(2k-1)(M-2s-2k+1)],\\ \gamma_n^{(q)}&=(-8\zeta)^n n!\prod_{k=1}^{n}[(2k+1)(M-2s-2k)]. \end{aligned} \tag{2.21}$$

We observe from here that $\gamma_n^{(p)}$ vanishes for $2n\geqslant M-2s+1$ while $\gamma_n^{(q)}$ vanishes for $2n\geqslant M-2s$ provided M is a positive integer and as remarked by Bender–Dunne, this vanishing of norm is an alternative characterization of the QES problem. It may be noted that unlike the Bender–Dunne case, the norms are alternative in sign for $2n<M-2s+1$ or $2n<M-2s$ depending on the set $\{P_n(E)\}$, or $\{Q_n(E)\}$, respectively.

Using the factorization property as given by Eqs. (2.11) and (2.12) and using the recursion relations (2.6) and (2.7) satisfied by P_n and Q_n, it is easily shown that $P_n Q_n R_n$ and S_n satisfy the following recursion relations:

$$\begin{aligned}
\overline{P}_n(E) &= [(M+1+2n)^2+4(M+1+2n)(s+\zeta-1)+4s^2-8s+4-6\zeta+\mathscr{E}]\overline{P}_{n-1}(E) \\
&\quad +4\zeta(M-1+2n)(M+2n-2)[2-2s-2n]\overline{P}_{n-2}(E), \\
\overline{Q}_n(E) &= [(M-1+2n)^2+2(M-1+2n)(2s+2\zeta-1)+4s^2-4s+1-2\zeta+\mathscr{E}]\overline{Q}_{n-1}(E) \\
&\quad +4\zeta(M-3+2n)(M+2n-2)[3-2s-2n]\overline{Q}_{n-2}(E), \\
\overline{R}_n(E) &= [(M+2n)^2+4(M+2n)(s+\zeta-1)+4s^2-8s+4-6\zeta+\mathscr{E}]\overline{R}_{n-1}(E)+4\zeta(M-1+2n) \\
&\quad \times(M+2n-3)[3-2s-2n]\overline{R}_{n-2}(E), \\
\overline{S}_n(E) &= [(M+2n)^2+2(M+2n)(2s+2\zeta-1)+4s^2-4s+1-2\zeta+\mathscr{E}]\overline{S}_{n-1}(E) \\
&\quad +4\zeta(M-2+2n)(M+2n-1)[2-2s-2n]\overline{S}_{n-2}(E).
\end{aligned} \tag{2.22}$$

Using these recursion relations, it is straightforward to obtain the norms of theses polynomials and show that they are all positive. In particular, we obtain

$$\begin{aligned}
\overline{\gamma}_n^P &= (4\zeta)^n \prod_{k=1}^{n} [(M+2k+1)(M+2k)(2k+2s)], \\
\overline{\gamma}_n^Q &= (4\zeta)^n \prod_{k=1}^{n} [(M+2k)(M+2k-1)(2k+2s-1)], \\
\overline{\gamma}_n^R &= (4\zeta)^n \prod_{k=1}^{n} [(M+2k+1)(M+2k-1)(2k+2s-1)], \\
\overline{\gamma}_n^S &= (4\zeta)^n \prod_{k=1}^{n} [(M+2k+1)(M+2k)(2k+2s)].
\end{aligned} \tag{2.23}$$

Note that in Eqs. (2.22) and (2.23) M and s take only specific values. For example, in the case of $\overline{P}_n(E)M$ is odd and $s=0$, while in the case of $Q_n(E)$ M is odd and $s=\frac{1}{2}$. On the other hand, in the case of $\overline{R}_n(E)$ M is even and $s=\frac{1}{2}$ while in the case of $S_n(E)$ M is even but $s=0$.

Finally, we can also obtain the weight function $w(E)$ for our polynomials $\{P_n(E)\}$ and $\{Q_n(E)\}$ by using the expression as derived by Krajewska *et al.*[5] In particular, the weight function $w^{(p)}(E)$ can be written as

$$w^{(p)}(E) = \sum_{k=1}^{M} w_k^{(p)} \delta(E-E_k), \tag{2.24}$$

where the numbers $w_k^{(p)}$ satisfy the algebraic equation

$$\sum_{k=1}^{M} P_n(E_k) w_k^{(p)} = \delta_{n0}. \tag{2.25}$$

A similar relation also exists for the weight function of the Q polynomials.

As an illustration we have computed the weight functions $w_n^{(p)}$ and $w_n^{(q)}$ in case $M=3$ and 4. For example, when $M=3$, using Eqs. (2.24) to (2.25) and (2.13) we find that

$$w_0^{(p)}=\frac{1}{2}-\frac{2\zeta+1}{2\sqrt{1+4\zeta^2}},\quad w_2^{(p)}=\frac{1}{2}+\frac{2\zeta+1}{2\sqrt{1+4\zeta^2}},\quad w_1^{(q)}=1. \tag{2.26}$$

On the other hand, when $M=4$, using Eqs. (2.24), (2.25), (2.17), and (2.18) we find

$$\begin{aligned} w_0^{(q)}&=\frac{1}{2}-\frac{\zeta+1}{2\sqrt{\zeta^2-\zeta+1}},\quad w_1^{(p)}=\frac{1}{2}-\frac{\zeta+1}{2\sqrt{\zeta^2+\zeta+1}},\\ w_2^{(q)}&=\frac{1}{2}+\frac{\zeta+1}{2\sqrt{\zeta^2-\zeta+1}},\quad w_3^{(p)}=\frac{1}{2}+\frac{\zeta+1}{2\sqrt{\zeta^2+\zeta+1}}. \end{aligned} \tag{2.27}$$

As a cross check on our calculation of weight functions, we have calculated the norm of the polynomials by using the basic relations

$$\gamma_n^{(p)}=\int dE w^{(p)}(E)[P_n(E)]^2;\quad \gamma_n^{(q)}=\int dE w^{(q)}(E)[Q_n(E)]^2 \tag{2.28}$$

and for $M=3$ as well as 4 we have verified that we get the same answer as given by Eq. (2.21). It is worth pointing out that unlike the Bender–Dunne example, in none of our cases the weight functions are always positive. Actually this is not all that surprising. As has been proved by Finkel *et al.*,[4] if the three-term recursion relation is of the form

$$\hat{P}_{k+1}=(E-b_k)\hat{P}_k-a_k\hat{P}_{k-1},\quad k\geqslant 0 \tag{2.29}$$

with $a_0=0$ and $a_{n+1}=0$, then the weight functions are all positive if b_k is real for $0\leqslant k\leqslant n$ and $a_k>0$ for $1\leqslant k\leqslant n$. On comparing Eq. (2.29) with our recursion relation (2.6) and (2.7) we find that in both of our cases, a_k is in fact <0 for $1\leqslant k\leqslant n$.

Once the weight function is known, then one can also calculate the moments of $w(E)$, defined by

$$\mu_n=\int dE\ w(E)E^n. \tag{2.30}$$

Since in our case, the polynomials $\{P_n(E)\}$ and $\{Q_n(E)\}$ are not eigenfunctions of parity, hence unlike the Bender–Dunne case, odd moments do not vanish in our case. For example, the first few moments in our case for $M=3$ are (we choose $\mu_0^{(p)}=\mu_0^{(q)}=1$ without any loss of generality)

$$\begin{aligned} \mu_1^{(p)}&=(3+\zeta)^2-2\zeta;\quad \mu_2^{(p)}=-16\zeta+[(3+\zeta)^2-2\zeta]^2;\\ \mu_1^{(q)}&=(3+\zeta)^2-6\zeta-4;\quad \mu_2^{(q)}=[(3+\zeta)^2-6\zeta-4]^2. \end{aligned} \tag{2.31}$$

Following Finkel *et al.*,[4] we can also calculate the growth rate of the moments. In particular, since in both of our cases, $a_k\neq 0$ for $1\leqslant k<n$ where a_k as given by Eqs. (2.29), (2.6) and (2.7), hence following their discussion it is easily shown that for large n, to leading order

$$\mu_n^{(p)},\quad \mu_n^{(q)}\sim(M+\zeta)^{2n}. \tag{2.32}$$

Thus in our case the moments have a pure power growth.

III. NEW QES POTENTIALS FROM ANTI-ISOSPECTRAL TRANSFORMATIONS

In an interesting paper, Krajewska *et al.*[1] have recently discussed the consequences of the anti-isospectral transformation (also termed as duality transformation) in the context of the QES problems. In particular, they have shown that under the transformations $x\rightarrow ix=y$, if a potential $V(x)$ goes to $\overline{V}(y)$, i.e.,

$$V(x)\rightarrow\overline{V}(y) \tag{3.1}$$

then the QES levels of the two are also related. In particular, they have shown that if M levels of the potential $V(x)$ are QES levels with energy eigenvalues and eigenfunctions $E_k(k=0,1,...,M-1)$ and $\psi_k(x)$ respectively then the energy eigenfunctions of $\bar{V}(y)$ are given by

$$\bar{E}_k = -E_{M-1-k}, \quad \bar{\psi}_k(y) = \psi_{M-1-k}(ix). \tag{3.2}$$

As an illustration, these authors have discussed the x^6 potentials as given by Eqs. (1.1) and (1.2) and explained in details how the eigen spectra of these two dual potentials are related. However, in this particular case, the domain of x and y are the same.

The purpose of this section is to explore the consequences of this symmetry when the domain of validity of $V(x)$ and $\bar{V}(y)$ are different and to see how many of the results derived in Ref. 1 go through in this case. In particular, do $V(x)$ and $\bar{V}(y)$, hold the same number of bound states? Are the bound states of the two potentials still related by Eq. (3.2)? To begin with, we would also like to comment that, the duality symmetry is only useful in case the potential $V(x)$ is symmetric in x as otherwise the dual potential $\bar{V}(y)$ will be a complex potential.

A. DSG example

Let us consider the potential corresponding to the DSHG case, and try to explore the consequences of the duality symmetry in this case. On applying the duality transformation $x \to i\theta$ to the Schrödinger equation $H\psi = E\psi$ with H as given by Eq. (2.1) we obtain the following Schrödinger equation for the DSG equation:

$$\left[-\frac{d^2}{d\theta^2} - (\zeta \cos 2\theta - M)^2 \right] \psi(\theta) = \hat{E}\psi(\theta), \quad \hat{E} = -E. \tag{3.3}$$

Note that whereas the domain of x in DSHG equation is $-\infty \leqslant x \leqslant \infty$ in the DSG case, the domain is $0 \leqslant \theta \leqslant 2\pi$ provided we are solving the problem on a circle as we do here. As a result, unlike the DSHG case, in the DSG case due to the periodicity constraint, one has to demand that the wave function $\psi(\theta)$ must be invariant under $\theta \to \theta + \pi$, i.e.,

$$\psi(\theta + \pi) = \psi(\theta). \tag{3.4}$$

Looking at the QES eigenfunctions of the DSHG case, it is easy to see that when M is even (i.e., $M = 2,4,...$) then the eigenfunctions of the DSG case do not remain invariant under $\theta \to \theta + \pi$ but they change sign. As a result, the QES levels of DSG when M is even must be rejected. Thus on physical ground, we find that in case the domain of $V(x)$ and $\bar{V}(y)$ are different then the number of QES levels of $V(x)$ and $\bar{V}(y)$ need not be identical. In particular, boundary conditions may forbid some wave functions to be eigenfunctions. For example, for $M = 2$, naively one would have two exact energy eigenstates of DSG equation as given by

$$\begin{aligned} \psi_0^{\mathrm{DSG}}(\theta) &= \sin\ \theta e^{-(\zeta/2)\cos 2\theta}, \\ \psi_2^{\mathrm{DSG}}(\theta) &= \cos\ \theta e^{-(\zeta/2)\cos 2\theta}. \end{aligned} \tag{3.5}$$

However, both of these are unacceptable eigenfunctions of DSG, being odd under $\theta \to \theta + \pi$. Thus we have shown that even though DSG and DSHG are dual system, DSG equation has approximately only half the number of QES levels compared to the DSHG case. In particular, when M is an odd integer ($M = 1,3,5,...$) only then the DSG is a QES system and in that case the first M levels are exactly known, and in fact they can be immediately obtained from the corresponding DSHG case by making use of the duality relation (3.2). For example, for $M = 1$ the ground state of the above DSG is given by

$$\hat{E}_0 = -(1 + \zeta^2), \quad \psi_0 = e^{-(\zeta/2)\cos 2\theta}. \tag{3.6}$$

On the other hand, for $M = 3$, the first 3 levels of the DSG equation are are given by

$$\hat{E}_0 = -7 - \zeta^2 - 2\sqrt{1 + 4\zeta^2}; \quad \psi_0 = [2\zeta - \{\sqrt{1 + 4\zeta^2} - 1\}\cos 2\theta] e^{-(\zeta/2)\cos 2\theta};$$

$$\hat{E}_1 = -\zeta^2 - 5; \quad \psi_1 = \sin 2\theta e^{-(\zeta/2)\cos 2\theta}; \tag{3.7}$$

$$\hat{E}_2 = -7 - \zeta^2 + 2\sqrt{1+4\zeta^2}; \quad \psi_2 = [2\zeta + \{\sqrt{1+4\zeta^2}+1\}\cos 2\theta] e^{-(\zeta/2)\cos 2\theta}.$$

The energy eigenstates of the DSG equation have recently been obtained by Habib *et al.*[9] and not surprisingly our results are the same as obtained by them.

One can now study the Bender–Dunne polynomials of the DSG equation and it is easy to see that most of the discussion of the last section (for the DSHG case) also goes through in this case except that now only odd values of M give us a QES system. For example, the recursion relations for P_n and Q_n as given in the last section are also valid in this case provided we replace E by $-E$. Thus the three-term recursion relation for P_n as given by Eq. (2.6) with E changed to $-E$ will give an exact solution for the DSG case only if $s=0$ while the three-term recursion relation for Q_n as given by Eq. (2.7) (with E changed to $-E$) will give an exact solution only if $s=\frac{1}{2}$. Most of the other properties in the last section continue to be valid in the DSG case also (with the obvious replacement of E by $-E$) except now only $\bar{P}_n(E)$ and $\bar{Q}_n(E)$ exist while $\bar{R}_n(E)$ and $\bar{S}_n(E)$ as given by Eq. (2.22) do not exist (Note that P_n and Q_n do not give QES solution if M is even integer.) In particular, we want to emphasize that the norms of the polynomial sets $\{P_n(E)\}$, $\{Q_n(E)\}$, $\{\bar{P}_n(E)\}$ and $\{\bar{Q}_n(E)\}$ are unchanged from those of the DSHG case. However, the weight functions get interchanged. For example for $M=3$ the weight functions for the DSG case are

$$w_0^{(p)} = \frac{1}{2} + \frac{2\zeta+1}{2\sqrt{1+4\zeta^2}}, \quad w_2^{(p)} = \frac{1}{2} - \frac{2\zeta+1}{2\sqrt{1+4\zeta^2}}, \quad w_1^{(q)} = 1. \tag{3.8}$$

Thus even in this case, the weight function are *not* always positive. Finally, it is easily shown that the moments for the DSG and DSHG cases are related by

$$\mu_n^{(p,q)}|_{\mathrm{DSG}} = (-1)^n \mu_n^{(p,q)}|_{\mathrm{DSHG}} \tag{3.9}$$

so that for large n, to leading order, the DSG moments also have a pure power growth rate as given by

$$\mu_n \sim (-1)^n (M+\zeta)^{2n}. \tag{3.10}$$

B. New QES potential

We shall now point out a nontrivial application of the duality symmetry. In particular, using this, we obtain an entirely new QES potential which has so far not been discussed in the literature.[15]

Consider the following potential

$$V(x) = \frac{\mu^2[8\sinh^4(\mu x/2) - 4(5/\epsilon^2 - 1)\sinh^2(\mu x/2) + 2(1/\epsilon^4 - 1/\epsilon^2 - 2)]}{8[1 + (1/\epsilon^2) + \sinh^2(\mu x/2)]^2} \tag{3.11}$$

which is obtained in the context of the stability analysis of the ϕ^6-kink solution in $1+1$ dimensions.[12] As has been shown before, in this case the Schrödinger equation can be converted to Heun's equation with four regular singular points. Further, for any ϵ, the ground stat of the Schrödinger equation corresponding to this potential is given by ($\hbar = 2m = 1$)

$$\psi_0 = N_0 \left[\frac{\epsilon^2+1}{\epsilon^2+1+\epsilon^2\sinh^2(\mu x/2)}\right]\left[\frac{\epsilon^2\sinh^2(\mu x/2)}{\epsilon^2+1+\epsilon^2\sinh^2(\mu x/2)} + \frac{1}{2}\right]^{1/2}, \quad E_0 = 0. \tag{3.12}$$

Further for the special case of, $\epsilon^2 = \frac{1}{2}$, the second excited state is also analytically known and given by

$$\psi_2 = N_2\left[\frac{3}{3+\sinh^2(\mu x/2)}\right]\left[\frac{\sinh^2(\mu x/2)}{3+\sinh^2(\mu x/2)} - \frac{1}{4}\right], \quad E_2 = \frac{3}{4}\,\mu^2. \tag{3.13}$$

Thus for $\epsilon^2 = \frac{1}{2}$ this is a QES system. It may be noted that even though it is a QES system in one dimension, the Hamiltonian can not be written in this case in terms of the quadratic generators of Sl(2).[16] Further, as shown by us recently,[7] in this case the Bender–Dunne polynomials do not form an orthogonal set. We shall now apply the duality transformation and obtain a new QES system which has not been discussed before.

On considering the duality transformation $x \to i\theta$ in the Schrödinger equation $H\psi = E\psi$ corresponding to the potential (3.11), we find that we have a new periodic potential

$$\overline{V}(\theta) = -\frac{\mu^2[8\sin^4(\mu\theta/2) + 4(5/\epsilon^2 - 1)\sin^2(\mu\theta/2) + 2(1/\epsilon^4 - 1/\epsilon^2 - 2)]}{8[1 + 1/\epsilon^2 - \sin^2(\mu\theta/2)]^2}. \tag{3.14}$$

Notice that the domain of $V(x)$ and $\overline{V}(\theta)$ are very different, i.e., whereas $-\infty \leqslant \mu x \leqslant \infty$, $0 \leqslant \mu\theta \leqslant 2\pi$. However, the predictions of duality symmetry as given by Eq. (3.2) are still valid. In particular, using Eq. (3.2) we predict that the potential $V(\theta)$ as given above is a QES system in case $\epsilon^2 = \frac{1}{2}$ and it's ground and second excited states energy eigenvalues are given by

$$\begin{aligned} E_0 &= -\frac{3}{4}\,\mu^2, \quad \psi_0 = N_3\left[\frac{3}{3-\sin^2(\mu\theta/2)}\right]\left[\frac{1}{4} + \frac{\sin^2(\mu\theta/2)}{3-\sin^2(\mu\theta/2)}\right], \\ E_2 &= 0, \quad \psi_2 = N_4\left[\frac{\epsilon^2+1}{\epsilon^2+1-\epsilon^2\,\sin^2(\mu\theta/2)}\right]\left[\frac{1}{2} - \frac{\epsilon^2\,\sin^2(\mu\theta/2)}{\epsilon^2+1-\epsilon^2\,\sin^2(\mu\theta/2)}\right]. \end{aligned} \tag{3.15}$$

In particular, we would like to emphasize that for the QES potential (3.14), the second excited state ψ_2 is known at all real values of ϵ, while ψ_0 is only known at $\epsilon^2 = \frac{1}{2}$. One can solve the Schrödinger equation explicitly for the potential (3.14) and check that indeed these are the energy eigenstates of the system. In particular, note that both ψ_0 and ψ_2 are invariant under $\mu\theta \to \mu\theta + 2\pi$, i.e., $\psi(\mu\theta + 2\pi) = \psi(\mu\theta)$. It is really remarkable that using duality we are able to obtain a new QES system. In this case one can also obtain the three-term recursion relation for the Bender–Dunne polynomials[2] and show that they do not form an orthogonal set. This is related to the fact that the corresponding Schrödinger equation can be converted to Heun's equation with four regular singular points.

It may be worthwhile to look at all known QES systems and see if the corresponding dual systems have already been discussed in the literature or not. In this way one may discover some more QES systems.

IV. DISCUSSIONS

In this paper we have analyzed in some detail a QES system for which the energy eigenstates for levels with odd as well as even number of nodes are known for a given potential. We have seen that in this case one obtains two independent sets of orthogonal polynomials. Further, we have also seen that the weight functions in this case are not necessarily positive. It will be interesting to study a few other examples of the same type and enquire if in those cases, too, one obtains two sets of orthogonal polynomials or not and further, whether the weight functions in such cases are always positive or not. We have also compared the spectrum of two dual potentials in case their domain of validity are quite different. We have seen in one case that because of different boundary conditions, the QES spectrum of the dual potentials is not the same. It will be interesting to study several such QES dual potentials and see if some general conclusions can be obtained in these cases.

Finally, it would be really interesting if one can discover some new QES potentials by making use of the duality symmetry.

[1] A. Krajewska, A. Ushveridze, and Z. Walczak, Mod. Phys. Lett. A **12**, 1225 (1997).
[2] C. M. Bender and G. V. Dunne, J. Math. Phys. **37**, 6 (1996).
[3] T. S. Chihara, *An Introduction to Orthogonal Polynomials* (Gordon and Breach, New York, 1975); A. Erdelyi, W. Maguus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. II.
[4] F. Finkel, A. Gonzalez-Lopez, and M. A. Rodriguez, J. Math. Phys. **37**, 3954 (1996).
[5] A. Krajewska, A. Ushveridze, and Z. Walczak, Mod. Phys. Lett. A **12**, 1131 (1997).
[6] C. M. Bender, G. V. Dunne, and M. Moshe, Phys. Rev. A **55**, 2625 (1997).
[7] A. Khare and B. P. Mandal, Phys. Lett. A **239**, 197 (1998).
[8] M. Razavy, Am. J. Phys. **48**, 285 (1980).
[9] S. Habib, A. Khare, and A. Saxena, Physica D, in press.
[10] N. G. van Kampen, J. Stat. Phys. **17**, 71 (1977).
[11] A. M. Polyakov, Nucl. Phys. B **120**, 429 (1977).
[12] N. H. Christ and T. D. Lee, Phys. Rev. D **12**, 1606 (1975).
[13] A. Khare, S. Habib, and A. Saxena, Phys. Rev. Lett. **79**, 3797 (1997).
[14] S. N. Behera and A. Khare, Journal de Physique **C6**, 314 (1981).
[15] A. Ushveridze, *Quasi-Exactly Solvable Models in Quantum Mechanics* (Institute of Physics Publishing, Bristol, 1994).
[16] D. P. Jatkar, C. Nagaraja Kumar, and A. Khare, Phys. Lett. A **142**, 200 (1989).

Generalized Lie algebras of type A_n

Volodymyr Lyubashenko[a] and Anthony Sudbery
Department of Mathematics, University of York, Heslington, York, England YO1 5DD, United Kingdom

(Received 9 April 1997; accepted for publication 10 December 1997)

It is shown that the quantized enveloping algebra of $\mathfrak{sl}(n)$ contains a generalized Lie algebra, defined by means of axioms similar to Woronowicz's. This gives rise to Lie algebra-like generators and relations for the locally finite part of the quantized enveloping algebra, and suggests a canonical Poincaré–Birkhoff–Witt basis.
© *1998 American Institute of Physics.* [S0022-2488(98)01705-8]

I. INTRODUCTION

There have been several proposals[1–3] for a definition of a ''quantum Lie algebra,'' something which should bear the same relation to the quantized enveloping algebra $U_q(\mathfrak{g})$ as the classical Lie algebra $\mathfrak{g}$ bears to its enveloping algebra. Properties of the classical Lie algebra which are retained in one or more of these proposals include the following.

(1) The quantum Lie algebra $\mathfrak{g}_q$ should be a vector space of the same dimension as the classical Lie algebra $\mathfrak{g}$.
(2) It should be a deformation of $\mathfrak{g}$, i.e., the classical Lie algebra should be recovered in the limit $q\to 1$.
(3) It should have a binary operation, the *quantum Lie bracket*.
(4) The quantum Lie bracket should be a bilinear map from $\mathfrak{g}\times\mathfrak{g}$ to $\mathfrak{g}$.
(5) It should satisfy a Jacobi identity and an antisymmetry axiom.
(6) These axioms should be stated in terms of the quantum Lie bracket alone (i.e., without reference to a pre-existing enveloping algebra).
(7) The quantum Lie algebra should appear as a subspace of the quantized enveloping algebra, the bracket $[x,y]$ being given by the Hopf algebra adjoint action of x on y.
(8) It should be possible to define the relations of the enveloping algebra in terms of the quantum Lie bracket, by interpreting the bracket as a quadratic expression like a commutator.
(9) There should be a Poincaré–Birkhoff–Witt theorem stating that the quantized enveloping algebra has a basis consisting of ordered monomials in the Lie algebra elements.

It seems to be impossible to meet all these requirements simultaneously. We might expect to have to modify them, indeed some of them can be seen to be unrealistic, but none of the proposed definitions satisfies them all, even in modified form. In this paper we describe a form of quantum Lie algebra having a subset of the above properties which is sufficient for the formulation of a gauge theory based on a quantum group.[4] To avoid confusion with the other senses of quantum Lie algebra defined in the literature, we call our system a ''generalized Lie algebra'' (we are grateful to Delius for this suggestion).

The essential properties of our definition, its relation to other definitions and the ways in which it fails to meet the specifications set out above are as follows.

1. Our generalized Lie algebra is a vector space with a bilinear bracket operation satisfying versions of anticommutativity and the Jacobi identity which are formulated with the aid of an auxiliary structure, a generalized antisymmetrizer. This structure is close to that found by Woronowicz[5] in his theory of non-commutative differential geometry on quantum groups. However, Woronowicz's Lie algebra has the additional property that its generalized antisymmetrizer

[a]Current address: Department of Mathematics, Kansas State University, 137 Cardwell Hall, Manhattan, Kansas 66506-2602.

 © 1998 American Institute of Physics

satisfies the braid relation. Ours does not have this property. It may be that some property similar to this is needed for a fully satisfactory theory, but we have not yet succeeded in identifying it. The quantum Lie algebra of Delius *et al.*[2,3] does not satisfy a Jacobi identity, and the antisymmetry of their Lie bracket can only be defined if the field of scalars is considered to be the field of rational functions in an indeterminate q; it is not a property of the algebra for a specific value of q. The Jacobi identity for Majid's quantum Lie algebra is formulated with reference to the enveloping algebra which contains it, not intrinsically; and he has no antisymmetry axiom.

2. With the help of the generalized antisymmetrizer we can define the enveloping algebra of a generalized Lie algebra, which is not possible for a general algebra. Thus we start with the generalized Lie algebra and construct the enveloping algebra around it, unlike Delius *et al.* and Majid who start with the quantized enveloping algebra and find the quantum Lie algebra inside it. The relations of our enveloping algebra, as in the classical case, are quadratic-linear (generalized commutator=Lie bracket). However, this enveloping algebra cannot be directly identified with the quantized enveloping algebra which is (roughly speaking—see the next paragraph) generated by the generalized Lie algebra but with homogeneous quadratic relations, the right-hand side containing an extra central element as a factor. To put it another way, the structure constants in the quantized enveloping algebra are multiples of a function of the Casimirs. This is an interesting feature; it means that the PBW theorem leads to a description of the vector space structure of the enveloping algebra as a space of polynomial functions not on a flat space (namely the Lie algebra) but on a hypersurface of the same dimension. Thus quantization gives rise to curvature of the enveloping algebra.

3. The algebra generated by our generalized Lie algebra cannot be expected to be the same size as the quantized enveloping algebra, since this differs from the classical enveloping algebra in containing infinite power series (specifically, exponentials) in the generators. On the other hand, if, as is common in the mathematical literature, one takes the generators of the quantized enveloping algebra to be q^{H_i} in place of the Cartan subalgebra elements H_i, then one loses the polynomials in the H_i. It is a remarkable fact[6] that the quantized enveloping algebra so defined contains a subalgebra, its locally finite part, which has the same structure as the classical enveloping algebra and is a large part of the quantized enveloping algebra. It seems likely that the enveloping algebra of the generalized Lie algebra, as defined in this paper, coincides with the locally finite part of the quantized enveloping algebra. Here we prove this for the case of $\mathfrak{sl}(2)$.

4. The construction of a generalized Lie algebra given here makes sense for any simple Lie algebra, but it is only for the Lie algebras of type A_n that it yields a system with the same dimension as the classical Lie algebra. Thus it is only in these cases that a PBW theorem is likely to hold. Even here, we have no general proof of the theorem; it can be checked by an explicit calculation for A_1 and A_2.

On this question of dimension our construction is similar to that of Majid, which yields (or can be adapted to yield) a quantum Lie algebra of dimension n^2-1 where n is the dimension of any representation of the classical Lie algebra. Woronowicz's theory yields a satisfactory deformation of the classical Lie algebra only in the case of general linear groups;[7–9] in all other cases (see Refs. 10–13) the generalized Lie algebra has dimension n^2 where n is the dimension of a particular representation.

5. As indicated above, the commutator-like relations obtained from the generalized Lie algebra are not the only relations in the quantized enveloping algebra; there is also a relation between central elements, giving the extra (non-classical) central generator in terms of more familiar Casimir elements which are polynomials in Lie algebra elements. The existence of this extra relation is at present only a conjecture in all cases except $\mathfrak{sl}(2)$, for which it can be explicitly demonstrated.

The organization of the paper is as follows. The next section contains the definition of a generalized Lie algebra and the proof that such a system can be found inside the quantized enveloping algebra. This relies heavily on the theorems of Joseph and Letzter[6] concerning the locally finite part of a quantized enveloping algebra. In section III we complete the theory for $\mathfrak{sl}(2)_q$, proving the Poincaré–Birkhoff–Witt theorem, determining the extra relation between central elements, and proving that the algebra defined by the generalized Lie algebra relations, together with the extra relation, is the locally finite part of the quantized enveloping algebra.

Section IV contains some material on the generalized Lie algebra $\mathfrak{sl}(3)_q$, with formulae for the Lie brackets and a discussion of the rather surprising symmetry properties of the enveloping algebra.

Notation: We use the usual notation for coproducts in a Hopf algebra:

$$\Delta(x)=\sum x_{(1)}\otimes x_{(2)},$$

$$(\Delta\otimes\mathrm{id})\circ\Delta(x)=\sum x_{(1)}\otimes x_{(2)}\otimes x_{(3)},\ \text{etc.},$$

and for the q-analogue function:

$$[x]_q=\frac{q^x-q^{-x}}{q-q^{-1}}.$$

II. GENERALIZED LIE ALGEBRAS

Definition 2.1: A *generalized Lie algebra* over a field $\mathbb{K}$ is a $\mathbb{K}$-vector space L together with linear maps $\beta: L\otimes L\to L$ (the generalized Lie bracket, written as $x\otimes y\mapsto[x,y]$) and $\gamma: L\otimes L\to L\otimes L$ (the generalized antisymmetrizer) satisfying the following.

1. Generalized antisymmetry: For $t\in L\otimes L$,

$$\gamma(t)=0\Rightarrow\beta(t)=0. \tag{2.1}$$

2. The generalized Jacobi identity: For $x\in L$, define $\mathrm{ad}x: L\to L$ by $\mathrm{ad}x(y)=[x,y]$. Then

$$\mathrm{ad}[x,y]=m\circ(\mathrm{ad}\otimes\mathrm{ad})\circ\gamma(x\otimes y), \tag{2.2}$$

where $m:\mathrm{End}\,L\otimes\mathrm{End}\,L\to\mathrm{End}\,L$ denotes multiplication of linear maps on L.

Definition 2.2: A generalized Lie algebra is *balanced* if it satisfies a second Jacobi identity.

3. The right generalized Jacobi identity: For $x\in L$, define $\mathrm{rad}x: L\to L$ by $\mathrm{rad}x(y)=[y,x]$. Then

$$\mathrm{rad}[x,y]=\tilde{m}\circ(\mathrm{rad}\otimes\mathrm{rad})\circ\gamma(x\otimes y), \tag{2.3}$$

where $\tilde{m}:\mathrm{End}\,L\otimes\mathrm{End}\,L\to\mathrm{End}\,L$ denotes the opposite multiplication of linear maps: $\tilde{m}(A\otimes B)=BA$.

If γ is the usual antisymmetrization map, $\gamma(x\otimes y)=x\otimes y-y\otimes x$, the above become the usual antisymmetry and Jacobi axioms for a Lie algebra. These generalized axioms were found by Woronowicz to hold for the geometrical Lie bracket of left-invariant vector fields on a quantum Lie group;[5] he also found that in this case the generalized antisymmetrizer was of the form $\gamma=1-\sigma$ where σ satisfies the braid relation. This is not true of the generalized Lie algebras described in this paper; we do not know if there is a weaker axiom of this type that they satisfy.

Woronowicz included only one Jacobi identity among his axioms; the geometrical construction does not obviously yield a balanced generalized Lie algebra. This is also, at first sight, true of the construction from universal enveloping algebras to be described in this paper. However, inspection of the generalized Lie algebra $\mathfrak{sl}(2)_q$ reveals that it is in fact balanced, and it seems likely that this is true more generally.

Definition 2.3: A *representation* of a generalized Lie algebra (L,β,γ) is a linear map $\rho: L\to\mathrm{End}\,V$ for some vector space V, such that

$$\rho\circ\beta=m_V\circ(\rho\otimes\rho)\circ\gamma: L\otimes L\to\mathrm{End}\,V, \tag{2.4}$$

where m_V denotes multiplication of linear operators on V.

Then the Jacobi identity (2.3) states that ad is a representation of L in $\mathrm{End}_{\mathbb{K}}V$, as with classical Lie algebras.

Definition 2.4: The *universal enveloping algebra* of a generalized Lie algebra (L,β,γ) is the quotient of the tensor algebra of L by the ideal generated by $(\beta-\gamma)(V\otimes V)$. Then every representation of L automatically generates a representation of its universal enveloping algebra.

Let $x_1,\ldots,x_n$ be a basis for L, and let γ_{ij}^{kl} be the matrix elements of γ with respect to the basis $x_i\otimes x_j$ of $L\otimes L$; in the classical case,

$$\gamma_{ij}^{kl}=\delta_i^k\delta_j^l-\delta_i^l\delta_j^k.$$

Then the Jacobi identity can be written as

$$[[x_i,x_j],x_k]=\gamma_{ij}^{lm}[x_l,[x_m,x_k]],$$

the right Jacobi identity as

$$[x_i,[x_j,x_k]]=\gamma_{jk}^{lm}[[x_i,x_l],x_m],$$

the representation property as

$$\rho([x_i,x_j])=\gamma_{ij}^{kl}\rho(x_k)\rho(x_l),$$

and the relations in the universal enveloping algebra as

$$\gamma_{ij}^{kl}x_kx_l=[x_i,x_j]=\beta_{ij}^kx_k,$$

where β_{ij}^k are the structure constants of L (i.e., the matrix elements of β).

Let q be a fixed element of the ground field $\mathbb{K}$, $\mathfrak{g}$ a simple Lie algebra over $\mathbb{K}$, and let $\mathcal{U}=\overline{U}_q(\mathfrak{g})$ be the simply-connected quantized enveloping algebra of $\mathfrak{g}$, which is defined as follows. $\mathcal{U}$ contains a copy of the group algebra of the weight lattice of $\mathfrak{g}$ (with basis elements denoted by q^λ where λ is an integral weight of $\mathfrak{g}$ with respect to a Cartan subalgebra $\mathfrak{h}$) and generators $E_1,\ldots,E_r,F_1,\ldots,F_r$ corresponding to fundamental roots $H_1,\ldots,H_r$ of $\mathfrak{g}$, with relations

$$q^\lambda E_iq^{-\lambda}=q^{\langle H_i,\lambda\rangle}E_i,$$

$$q^\lambda F_iq^{-\lambda}=q^{-\langle H_i,\lambda\rangle}F_i,$$

$$E_iF_j-F_jE_i=\delta_{ij}\left(\frac{q^{2H_i}-q^{-2H_i}}{q_i-q_i^{-1}}\right),$$

$$[E_i,[E_i,\ldots,[E_i,E_j]_{q_i^{-n+1}}]_{q_i^{-n+3}}\cdots]_{q_i^{n-1}}=0,$$

$$[F_i,[F_i,\ldots,[F_i,F_j]_{q_i^{-n+1}}]_{q_i^{-n+3}}\cdots]_{q_i^{n-1}}=0,$$

where $\langle,\rangle$ denotes the Killing form in the Cartan subalgebra $\mathfrak{h}$ (which we do not distinguish from its dual $\mathfrak{h}^*$),

$$q_i=q^{\langle H_i,H_i\rangle},$$

$$n=\frac{2\langle H_i,H_j\rangle}{\langle H_i,H_i\rangle}=\text{number of roots } H_j+kH_i \text{ with } k>0,$$

and

$$[X,Y]_p=pXY-p^{-1}YX.$$

Then $\mathcal{U}$ is a Hopf algebra with comultiplication

$$\Delta(E_i)=E_i\otimes q^{-H_i}+q^{H_i}\otimes E_i,$$

$$\Delta(F_i)=F_i\otimes q^{-H_i}+q^{H_i}\otimes F_i,$$

$$\Delta(q^\lambda)=q^\lambda\otimes q^\lambda,$$

and antipodes

$$S(E_i)=-q_i^{-1}E_i,\quad S(F_i)=-q_iF_i,\quad S(q^\lambda)=q^{-\lambda}.$$

As in any Hopf algebra, we have the adjoint representation of $\mathcal{U}$ on itself, given by $x\mapsto \mathrm{ad}\ x\in \mathrm{End}_{\mathbb{K}}\ \mathcal{U}$ where

$$\mathrm{ad}\ x(y)=\sum\ x_{(1)}yS(x_{(2)}),\quad \text{if}\ \ \Delta(x)=\sum\ x_{(1)}\otimes x_{(2)}.$$

This is a representation:

$$\mathrm{ad}(xy)=\mathrm{ad}\ x\cdot\mathrm{ad}\ y, \tag{2.5}$$

and each ad x is a generalized derivation of $\mathcal{U}$, in the sense that

$$\mathrm{ad}\ x\ \ (yz)=\sum\ \ \mathrm{ad}\ x_{(1)}(y)\cdot\mathrm{ad}\ x_{(2)}(z). \tag{2.6}$$

The adjoint action of the generators of $\mathcal{U}$ is given by

$$\begin{aligned}\mathrm{ad}\ E_i(x)&=E_ixq^{H_i}-q^{H_i-1}xE_i,\\ \mathrm{ad}\ F_i(x)&=F_ixq^{H_i}-q^{H_i+1}xF_i,\\ \mathrm{ad}\ q^\lambda(x)&=q^\lambda xq^{-\lambda}.\end{aligned} \tag{2.7}$$

We use the adjoint representation to define a bracket on $\overline{U}_q(\mathfrak{g})$, defining

$$[x,y]=\mathrm{ad}\ x(y). \tag{2.8}$$

Each ad x is also a generalized derivation of this bracket:

$$[x,[y,z]]=\sum\ [[x_{(1)},y],[x_{(2)},z]]. \tag{2.9}$$

This is a kind of Jacobi identity for the adjoint bracket. The representation property gives another kind of Jacobi identity:

$$[[x,y],z]=\sum\ [x_{(1)},[y,[x_{(2)},z]]]. \tag{2.10}$$

Both of these are valid in any Hopf algebra.[1] However, they are not suitable as replacements for the Jacobi identity in defining a notion of generalized Lie algebra, as the ad-invariant subspaces of quantized enveloping algebras which are candidates for generalized Lie algebras are not usually subcoalgebras, and so the coproduct which is required to formulate the above Jacobi identities is not an intrinsic notion. That is why we are interested in the third kind of Jacobi identity, of the form (2.3), which is specific to quantized enveloping algebras.

The vector space $\overline{U}_q(\mathfrak{g})$ does not split up nicely into irreducible ad-invariant subspaces; strictly speaking it is not a direct sum of such spaces, for there are elements with components in an infinite number of them. The *locally finite part* of $\mathcal{U}=\overline{U}_q(\mathfrak{g})$ is the set of elements for which this is not true, i.e.,

$$\mathcal{F}=\mathrm{LF}\overline{U}_q(\mathfrak{g})=\{x\in\mathcal{U}:\ \text{ad}\,\mathcal{U}(x)\ \text{ is finite-dimensional}\}. \tag{2.11}$$

The discovery of Joseph and Letzter was that $\mathcal{F}$ is easily described, and has the same composition as the classical enveloping algebra $U(\mathfrak{g})$.

Theorem 2.1 (Joseph and Letzter[6])**:** *Let* λ *be a dominant integral weight of* $\mathfrak{g}$, V_λ *the simple* $\mathfrak{g}$*-module with highest weight* λ; V_λ *can also be regarded as a simple* $\mathcal{U}$*-module where* $\mathcal{U}=\overline{U}_q(\mathfrak{g})$, *and* $\mathrm{End}_{\mathbb{K}}\,V_\lambda$ *is a* $\mathcal{U}$*-module by conjugation [i.e., if* $x\in\mathcal{U}$ *acts on* V_λ *by* $\rho(x)$, *it acts on* $\mathrm{End}_{\mathbb{K}}\,V_\lambda$ *by* $T\mapsto\Sigma\rho(x_{(1)})T\rho(Sx_{(2)})$*]. Then we have the following.*

(1) $E_\lambda=\text{ad}\,\mathcal{U}(q^{-4\lambda})$ *is a finite-dimensional* ad*-invariant subspace of* $\mathcal{U}$, *isomorphic to* $\mathrm{End}_{\mathbb{K}}\,V_\lambda$ *as a* $\mathcal{U}$*-module.*

(2) *The locally finite part of* $\mathcal{U}$ is

$$\mathcal{F}=\sum_\lambda E_\lambda, \tag{2.12}$$

where the sum extends over all dominant integral weights of $\mathfrak{g}$.

For a dominant integral weight λ, we write $N_\lambda=\dim_{\mathbb{K}}V_\lambda$. Then our result on the existence of generalized Lie algebras in $\overline{U}_q(\mathfrak{g})$ is the following.

Theorem 2.2: *For each dominant integral weight* λ, $\overline{U}_q(\mathfrak{g})$ *contains an* (N_λ^2-1)*-dimensional* ad*-invariant subspace* L_λ, *a linear map* $\sigma:L_\lambda\otimes L_\lambda\to L_\lambda\otimes L_\lambda$, *and a central element* $C_\lambda\in\overline{U}_q(\mathfrak{g})$ *such that*

$$xy-m\circ\sigma(x\otimes y)=C_\lambda[x,y]\quad(x,y\in L_\lambda), \tag{2.13}$$

i.e.,

$$x_ix_j-\sigma_{ij}^{kl}x_kx_l=C_\lambda[x_i,x_j], \tag{2.14}$$

where m *denotes multiplication in* $\overline{U}_q(\mathfrak{g})$, $[,]$ *is the adjoint bracket, and* $\{x_i\}$ *is a basis of* L_λ.

Proof: Writing $\mathcal{U}=\overline{U}_q(\mathfrak{g})$, let $\overline{L}=\text{ad}\,\mathcal{U}(q^{-4\lambda})$. Then $\overline{L}$ is invariant under ad $\mathcal{U}$, and by Theorem 2.1 it is isomorphic as a $\mathcal{U}$-module to $\mathrm{End}_{\mathbb{K}}\,V_\lambda$. This contains a one-dimensional submodule (spanned by the identity of $\mathrm{End}_{\mathbb{K}}\,V_\lambda$) and so by complete reducibility of representations of $\overline{U}_q(\mathfrak{g})$ (Ref. 14, p. 324) we can write

$$\overline{L}=\mathbb{K}C_\lambda\oplus L,$$

as a $\mathcal{U}$-module, for some $C_\lambda\in\overline{L}$. Then C_λ carries the trivial representation of $\mathcal{U}$:

$$\text{ad}\,x(C_\lambda)=\varepsilon(x)C_\lambda,\quad\forall x\in\mathcal{U}.$$

Hence C_λ is a central element of $\mathcal{U}$, for

$$xC_\lambda=\sum x_{(1)}C_\lambda Sx_{(2)}x_{(3)}=\sum\ \text{ad}\,x_{(1)}(C_\lambda)x_{(2)}=\sum\varepsilon(x_{(1)})C_\lambda x_{(2)}=C_\lambda x.$$

Now $\overline{L}$ is a left coideal of $\mathcal{U}$:

$$\begin{aligned}x\in\overline{L}&\Rightarrow x=\text{ad}\,u(q^{-4\lambda}),\quad\text{for some}\ u\in\mathcal{U}\\&\Rightarrow\Delta(x)=\sum\Delta(u_{(1)})\Delta(q^{-4\lambda})\Delta(Su_{(2)})\\&=\sum(u_{(1)}\otimes u_{(2)})(q^{-4\lambda}\otimes q^{-4\lambda})(Su_{(4)}\otimes Su_{(3)})\\&\in\mathcal{U}\otimes\overline{L}.\end{aligned} \tag{2.15}$$

Hence for any $x \in L$ we can write

$$\Delta(x) = x_0 \otimes C_\lambda + \sum u' \otimes x',$$

with $x_0, u' \in \mathcal{U}$ and $x' \in L$. In fact we can choose C_λ so that $x_0 = x$: for if $q^{-4\lambda} = C_\lambda + w$ with $w \in L$ and if $x = \mathrm{ad}\, u(q^{-4\lambda})$, then from (2.15),

$$\begin{aligned}\Delta(x) &= \sum u_{(1)} q^{-4\lambda} S u_{(3)} \otimes \mathrm{ad}\, u_{(2)}(C_\lambda + w) \\ &= \sum u_{(1)} q^{-4\lambda} S u_{(3)} \otimes (\varepsilon(u_{(2)}) C_\lambda + \mathrm{ad}\, u_{(2)}(w)) \\ &= u_{(1)} q^{-4\lambda} S u_{(2)} \otimes C_\lambda + \sum u_{(1)} q^{-4\lambda} S u_{(3)} \otimes \mathrm{ad}\, u_{(2)}(w),\end{aligned}$$

but $\mathrm{ad}\, u_{(2)}(w) \in L$ since L is invariant under $\mathrm{ad}\, \mathcal{U}$, and the first term is $x \otimes C_\lambda$.

In L we have the bracket $[x,y] = \mathrm{ad}\, x(y)$. Define $\sigma: L \otimes L \to L \otimes L$ by

$$\sigma(x \otimes y) = \sum \mathrm{ad}\, u'(y) \otimes x', \quad \text{where } \Delta(x) = x \otimes C_\lambda + \sum u' \otimes x', \tag{2.16}$$

i.e.,

$$\sigma(x \otimes y) = \sum \mathrm{ad}\, x_{(1)}(y) \otimes x_{(2)} - [x,y] \otimes C_\lambda. \tag{2.17}$$

Then

$$m \circ \sigma(x \otimes y) = \sum x_{(1)} y(S x_{(2)}) x_{(3)} - [x,y] C_\lambda = xy - [x,y] C_\lambda.$$

Since C_λ is central, this establishes the theorem. □

The existence of a generalized Lie algebra in $U_q(\mathfrak{sl}(n))$ follows immediately.

Theorem 2.3: *If q is not a $2n$th root of unity nor a primitive twelfth root of unity, $U_q(\mathfrak{sl}(n))$ contains an (n^2-1)-dimensional generalized Lie algebra with the adjoint bracket.*

Proof: In Theorem 2, take $\mathfrak{g} = \mathfrak{sl}(n)$ and λ the highest weight of the n-dimensional (defining) representation. Then L_λ carries an irreducible representation of $U_q(\mathfrak{g})$ [a deformation of the adjoint representation of $\mathfrak{sl}(n)$]. Since C_λ is central, $\mathrm{ad}\, C_\lambda$ acts on L_λ as a multiple of the identity; in a lemma we will show that this multiple is $q^2 - 1 + q^{-2}$ (which vanishes only if q is a primitive 12th root of unity). Assuming this for the moment, we apply ad to (2.14) and restrict to $L = L_\lambda$ to obtain

$$\mathrm{ad}_L[x_i, x_j] = \gamma_{ij}^{kl}\, \mathrm{ad}_L\, x_k \cdot \mathrm{ad}_L\, x_l, \tag{2.18}$$

where

$$\gamma = \frac{1 - \sigma}{q^2 - 1 + q^{-2}}.$$

Thus the generalized Jacobi identity is satisfied in L by the adjoint bracket. Also, from (2.13),

$$\gamma\left(\sum x \otimes y\right) = 0 \Rightarrow \sigma\left(\sum x \otimes y\right) = \sum x \otimes y \Rightarrow C_\lambda \sum [x,y] = 0 \Rightarrow \sum [x,y] = 0,$$

since $\overline{U}_q(\mathfrak{g})$ contains no zero divisors.[6] Thus the adjoint bracket restricted to L_λ has the antisymmetry property with respect to the generalized antisymmetrizer γ.

Completion of the proof now needs the following lemmas.

Lemma 2.1: The central element in $\overline{L}$ is given by

$$C_\lambda = \sum_{r=0}^{n-1} (-1)^r \frac{[n-r]_q}{[n]_q} K_r, \tag{2.19}$$

where K_r is defined recursively by

$$K_r = \mathrm{ad}(F_r E_r) K_{r-1}, \quad K_0 = q^{-4\lambda}. \tag{2.20}$$

Lemma 2.2: The adjoint action of the central element on L is

$$\mathrm{ad}\ C_\lambda(x) = (q^2 - 1 + q^{-2})x, \quad \text{for } x \in L. \tag{2.21}$$

The proofs, both unenlightening calculations, are relegated to an appendix. □

Although the antisymmetrizer γ in this Lie algebra is not of the Woronowicz form $\gamma = 1 - \sigma$ where σ satisfies the braid relation, it is of a related form. According to (2.18) and (2.17), γ is a scalar multiple of $1 - \sigma$ where

$$\sigma(x \otimes y) = \overline{\sigma}(x \otimes y) - [x,y] \otimes C_\lambda,$$

and $\overline{\sigma}: \overline{L} \otimes \overline{L} \to \overline{L} \otimes \overline{L}$ is given by

$$\overline{\sigma}(x \otimes y) = \sum \mathrm{ad}\ x_{(1)}(y) \otimes x_{(2)}.$$

It is a straightforward exercise in Hopf algebra manipulation[15] to show that $\overline{\sigma}$ satisfies the braid relation.

III. THE GENERALIZED LIE ALGEBRA $\mathfrak{sl}(2)_q$

In the case of $\mathfrak{g} = \mathfrak{sl}(2)$, the simply connected quantized enveloping algebra is generated by E, F, and $q^{\pm(1/2)H}$ with relations

$$q^H E q^{-H} = qE,$$

$$q^H F q^{-H} = q^{-1} F,$$

$$EF - FE = \frac{q^{2H} - q^{-2H}}{q - q^{-1}}.$$

The highest weight of the fundamental two-dimensional representation is (i.e., corresponds via the Killing form to) $\lambda = \frac{1}{2}H$, so in the notation of Lemma 2.1 we have

$$K_0 = q^{-4\lambda} = q^{-2H},$$

$$K_1 = \mathrm{ad}(FE) K_0 = -(q - q^{-1})(qEF - q^{-1}FE).$$

Hence, according to Lemma 2.1, the central element is

$$C = q^{-2H} + \left(\frac{q - q^{-1}}{q + q^{-1}}\right)(qEF - q^{-1}FE). \tag{3.1}$$

A basis for the generalized Lie algebra is $(X_\pm, X_0)$ where

$$X_+ = \frac{\operatorname{ad} E(q^{-2H})}{q-q^{-1}} = q^{-H}E,$$

$$X_- = -\frac{\operatorname{ad} F(q^{-2H})}{q-q^{-1}} = q^{-H}F, \tag{3.2}$$

$$X_0 = -\frac{\operatorname{ad}(FE)q^{-2H}}{q^2-q^{-2}} = \frac{qEF-q^{-1}FE}{q+q^{-1}} = \frac{C-q^{-2H}}{q-q^{-1}}.$$

The adjoint brackets of these basis elements can be calculated to be

$$[X_+,X_+]=0, \quad [X_+,X_0]=-q^{-1}X_+, \quad [X_+,X_-]=(q+q^{-1})X_0,$$

$$[X_0,X_+]=qX_+, \quad [X_0,X_0]=(q-q^{-1})X_0, \quad [X_0,X_-]=-q^{-1}X_-, \tag{3.3}$$

$$[X_-,X_+]=-(q+q^{-1})X_0, \quad [X_-,X_0]=qX_-, \quad [X_-,X_-]=0.$$

The generalized Lie algebra $\mathfrak{sl}(2)_q$ is defined by these brackets together with the generalized antisymmetrizer $\gamma=(q^2-1+q^{-2})^{-1}\gamma'$ where

$$\gamma'(X_\pm\otimes X_\pm)=0,$$

$$\gamma'(X_\pm\otimes X_0)=q^{\mp 2}X_\pm\otimes X_0-X_0\otimes X_\pm,$$

$$\gamma'(X_0\otimes X_\pm)=q^{\pm 2}X_0\otimes X_\pm-X_\pm\otimes X_0,$$

$$\gamma'(X_+\otimes X_-)=X_+\otimes X_--X_-\otimes X_++(q^2-q^{-2})X_0\otimes X_0, \tag{3.4}$$

$$\gamma'(X_-\otimes X_+)=-\gamma'(X_+\otimes X_-),$$

$$\gamma'(X_0\otimes X_0)=(q-q^{-1})^2X_0\otimes X_0+\left(\frac{q-q^{-1}}{q+q^{-1}}\right)(X_+\otimes X_--X_-\otimes X_+)=\left(\frac{q-q^{-1}}{q+q^{-1}}\right)\gamma'(X_+\otimes X_-).$$

We note the following properties of $\mathfrak{sl}(2)_q$.

Proposition 1: 1. $\mathfrak{sl}(2)_q$ *is a balanced generalized Lie algebra.*

2. The generalized antisymmetrizer γ *is essentially idempotent:*

$$\gamma^2=\left(\frac{q^2+q^{-2}}{q^2-1+q^{-2}}\right)\gamma.$$

Proof: By calculation. □

The second of these properties is peculiar to $\mathfrak{sl}(2)_q$, reflecting the simple structure of its representations, as will become apparent when we consider $\mathfrak{sl}(3)_q$. The first, however, may be more general.

The enveloping algebra of the generalized Lie algebra $U(\mathfrak{sl}(2)_q)$—not to be confused with the quantized enveloping algebra $U_q(\mathfrak{sl}(2))$—is defined by means of the brackets (3.3) and the antisymmetrizer (3.4). By redefining the generators to absorb the factor q^2-1+q^{-2}, it can be presented as the associative algebra generated by three elements $Y_0, Y_\pm$ with relations

$$qY_0Y_+-q^{-1}Y_+Y_0=Y_+,$$

$$q^{-1}Y_0Y_--qY_-Y_0=-Y_-, \tag{3.5}$$

$$Y_+Y_--Y_-Y_++(q^2-q^{-2})Y_0^2=(q+q^{-1})Y_0.$$

On the other hand, as elements of the quantized enveloping algebra $U_q(\mathfrak{sl}(2))$ the generators $X_\pm$, X_0 and the central element C satisfy (by Theorem 2.2)

$$q^2X_0X_+ - X_+X_0 = qCX_+ ,$$

$$q^{-2}X_0X_- - X_-X_0 = -q^{-1}CX_- , \tag{3.6}$$

$$X_+X_- - X_-X_+ + (q^2 - q^{-2})X_0^2 = (q + q^{-1})CX_0 ,$$

and

$$CX_m = X_mC \quad (m = 0, \pm 1).$$

However, the central element C is not independent of the other generators X_m: there is a further relation,

$$C^2 = (q - q^{-1})^2\left(X_0^2 + \frac{qX_-X_+ + q^{-1}X_+X_-}{q + q^{-1}} \right) + 1, \tag{3.7}$$

as can be verified by direct calculation.

The main results of this section are that the locally finite part of $\overline{U}_q(\mathfrak{g})$ is isomorphic, as an algebra, to the algebra with generators X_0, $X_\pm$, C, and relations (3.6), and as a vector space to the algebra of polynomials in four commuting variables X_0, $X_\pm$, C subject to the relation (3.7). We use the following notation for the various algebras:

$\mathcal{A}$ = enveloping algebra of the generalized Lie algebra $\mathfrak{sl}(2)_q$
= algebra generated by $Y_\pm$, Y_0 with relations (3.5),

$\mathcal{B}$ = algebra generated by $X_\pm$, X_0, C with relations (3.6),

$\mathcal{C}$ = algebra generated by $X_\pm$, X_0, C with relations (3.6) and (3.7),

$\mathcal{F}$ = locally finite part of $\overline{U}_q(\mathfrak{g})$.

Then $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are related by homomorphisms

$$\varphi: \mathcal{B} \to \mathcal{A} \quad \text{with} \quad \varphi(C) = 1,$$

$$\psi: \mathcal{B} \to \mathcal{C} \quad \text{enforcing} \quad (3.7).$$

Lemma 3.1: The algebra $\mathcal{A}$ has a basis of ordered monomials $Y_-^l Y_0^m Y_+^n$ where l,m,n are non-negative integers.

Proof: It is evident that the commutation relations (3.5) allow any product of Y's to be expressed as a combination of ordered monomials. We have only to show that these monomials are independent. According to the diamond lemma,[16] this will follow if it can be shown that the two methods of reducing $Y_+Y_0Y_-$ to a combination of ordered monomials both lead to the same result. This is easily checked using (3.5). □

Lemma 3.2: The algebra $\mathcal{B}$ has a basis of ordered monomials $C^kX_-^lX_0^mX_+^n$.

Proof: $\mathcal{A} \otimes \mathbb{K}[C]$ contains an isomorphic copy of $\mathcal{B}$ generated by C and $X_m = CY_m$. This is spanned by monomials $C^kX_-^lX_0^mX_+^n = C^{k+l+m+n}Y_-^lY_0^mY_+^n$ and by Lemma 3.1 these are independent. □

Lemma 3.3: The algebra $\mathcal{C}$ is isomorphic as a vector space to the space of polynomials in four commuting variables C, X_0, X_+, X_- subject to the relation (3.7).

Proof: Let

$$C_2 = (q + q^{-1})X_0^2 + qX_-X_+ + q^{-1}X_+X_- .$$

Using (3.6), it can be checked that C_2 commutes with X_0, X_+, and X_-. Hence the two-sided ideal I generated by $(q + q^{-1})(C^2 - 1) - (q - q^{-1})^2C_2$ [i.e., the kernel of the relation (3.7)] is the same as the left ideal generated by it. It follows that the monomials $X_-^lX_0^mX_+^n$ and $CX_-^lX_0^mX_+^n$ are independent modulo I, and therefore form a basis of $\mathcal{C}$. But these monomials also form a basis of the commutative algebra described in the theorem. □

Theorem 3.1: *The locally finite part of $\overline{U}_q(\mathfrak{sl}(2))$ is isomorphic to $\mathcal{C}$.*

Proof: By Theorem 2.1, the locally finite part of $\mathcal{U} = \overline{U}_q(\mathfrak{g})$ is

$$\mathcal{F}=\sum_{n=0}^{\infty} V_n, \quad \text{where } V_n=\mathrm{ad}\,\mathcal{U}(q^{-2nH}).$$

Since $\frac{1}{2}nH$ is the highest weight of the $(n+1)$-dimensional representation of $\mathcal{U}$, $\dim V_n=(n+1)^2$. By the derivation property (2.6) of the adjoint action,

$$V_{m+n}\subseteq V_m V_n.$$

Hence

$$V_n\subseteq V_1^n.$$

Let $\mathcal{U}'$ be the subalgebra of $\mathcal{U}$ generated by V_1, i.e., by C, X_-, X_+, and X_0. Since the relations (3.6) and (3.7) hold in $\mathcal{U}$, $\mathcal{U}'$ is a quotient of $\mathcal{C}$ and therefore, by Lemma 3.3, is spanned by monomials $X_-^k X_0^l X_+^m$ and $CX_-^k X_0^l X_+^m$. Let W_n be the subspace spanned by monomials $X_-^k X_0^l X_+^m$ with $k+l+m=n$. Then

$$V_1^{n+2}=V_1^n+CW_{n+1}+W_{n+2}, \tag{3.8}$$

as can be proved by induction, using

$$V_1 W_r=CW_r+W_{r+1}$$

and

$$C^2W_r+W_{r+2}=W_r+W_{r+2},$$

which follows from (3.7). But

$$\dim W_r\leqslant\frac{1}{2}(r+1)(r+2), \tag{3.9}$$

so

$$\dim(V_1^{n+2})\leqslant\dim(V_1^n)+(n+3)^2.$$

We can now prove by induction that

$$V_1^n=V_n\oplus V_{n-2}\oplus\cdots\oplus(V_0 \text{ or } V_1).$$

For if this is true, then from (3.8) we have

$$V_1^{n+2}\supseteq V_{n+2}\oplus V_n\oplus\cdots,$$

since the V_n are known to be complementary in $\mathcal{U}$.[6] By dimensions the inclusion must be equality, so the sum in (3.8) is direct and also (3.9) must be equality. It follows that the monomials $X_-^k X_0^l X_+^m$ and $CX_-^k X_0^l X_+^m$ are independent and form a basis of $\Sigma V_n=\mathcal{F}$. By Lemma 3.3, $\mathcal{F}$ is isomorphic to $\mathcal{C}$. □

In one sense, the locally finite part of the quantized enveloping algebra is the same size as the classical enveloping algebra of $\mathfrak{sl}(2)$. Both can be written as $Z\otimes\mathcal{U}_0$ where Z, the center, is a polynomial algebra in one variable and $\mathcal{U}_0$ has the same decomposition into simple modules under the adjoint action in both cases. However, the above shows that it is more appropriate to regard the locally finite part of $\overline{U}_q(\mathfrak{sl}(2))$ as twice as big as the classical enveloping algebra. In the classical case the center is generated by a Casimir element which is a polynomial in the Lie algebra generators; in the generalized case the center is generated by a square root of such a polynomial, tending to 1 in the classical limit.

IV. THE GENERALIZED LIE ALGEBRA $\mathfrak{sl}(3)_q$

This section contains some details of and observations on the generalized Lie algebra of type A_2. All were established by mindless calculation.

We take λ to be the highest weight of the triplet representation,

$$\lambda=\frac{1}{3}(2H_1+H_2)$$

(identifying weights with elements of the Cartan subalgebra by means of the Killing form, and taking H_1 and H_2 to have unit length). A convenient basis for the generalized Lie algebra $\mathfrak{sl}(3)_q$ is the following, consisting of two elements T_1,T_2 (scalar multiples of the K_1,K_2 defined in Lemma 2.1) and elements $X_\alpha=X_{\pm1},X_{\pm2},X_{\pm12}$ corresponding to the roots $\alpha=\pm H_1,\pm H_2,\pm(H_1+H_2)$ of $\mathfrak{sl}(3)$:

$$T_1=\frac{\mathrm{ad}(F_1E_1)q^{-4\lambda}}{q-q^{-1}}=q^{-(2/3)(H_1+2H_2)}(qE_1F_1-q^{-1}F_1E_1),$$

$$T_2=\mathrm{ad}(F_2E_2)T_1,$$

$$X_1=\frac{\mathrm{ad}\ E_1(q^{-4\lambda})}{q-q^{-1}}=q^{-(1/3)(5H_1+4H_2)}E_1,$$

$$X_{12}=\mathrm{ad}\ E_2(X_1)=q^{-(1/3)(5H_1+H_2)}(E_2E_1-q^{-1}E_1E_2),$$

$$X_2=\mathrm{ad}\ F_1(X_{12}),$$

$$X_{-1}=\frac{\mathrm{ad}\ F_1(q^{-4\lambda})}{q-q^{-1}}=-q^{-(1/3)(5H_1+4H_2)}F_1,$$

$$X_{-12}=\mathrm{ad}\ F_2(X_{-1})=q^{-(1/3)(5H_1+H_2)}(qF_1F_2-F_2F_1),$$

$$X_{-2}=-\mathrm{ad}\ E_1(X_{-12}).$$

The adjoint action of $\overline{U_q}(\mathfrak{g})$ on these basis elements is

$$\mathrm{ad}(q^{H_i})T_j=T_j;$$

$$\mathrm{ad}(q^{H_i})X_{\pm j}=\begin{cases}q^{\pm1}X_{\pm j}, & \text{if } i=j,\\ q^{\mp1/2}X_{\pm j}, & \text{if } i\neq j;\end{cases}$$

$$\mathrm{ad}(q^{H_i})X_{\pm12}=q^{\pm1/2}X_{\pm12};$$

$$\mathrm{ad}\ E_i(X_\alpha)=\begin{cases}X_{\alpha+H_i}, & \text{if } \alpha+H_i \text{ is a root},\\ T_i, & \text{if } \alpha=-H_i,\\ 0, & \text{otherwise};\end{cases}$$

$$\mathrm{ad}\ F_i(X_\alpha)=\begin{cases}X_{\alpha-H_i}, & \text{if } \alpha-H_i \text{ is a root},\\ T_i, & \text{if } \alpha=H_i,\\ 0, & \text{otherwise};\end{cases}$$

$$\mathrm{ad}\ E_i(T_j)=\begin{cases}(q+q^{-1})X_i, & \text{if } i=j,\\ X_{12}, & \text{if } i\neq j;\end{cases}$$

	T_1	T_2	X_1	X_{-1}
T_1	$-(q^2-q^{-2})T_1$	$-(q-q^{-1})T_1$	$-q(q+q^{-1})$	$q^{-1}(q+q^{-1})$
T_2	$-(q-q^{-1})T_1$		$-q$	q^{-1}
X_1	$q^{-1}(q+q^{-1})$	q^{-1}	0	T_1
X_{-1}	$-q(q+q^{-1})$	$-q$	$-T_1$	0
X_2	$-q^2$	$-q^{-1}(q^2+q^{-2})$	$-q^{3/2}$	0
X_{-2}	q^{-2}	$q(q^2+q^{-2})$	0	$q^{-3/2}$
X_{12}	1	$-q^{-3}$	0	$q^{1/2}$
X_{-12}	-1	q^3	$-q^{-1/2}$	0

	X_2	X_{-2}	X_{12}	X_{-12}
T_1	q^{-2}	$-q^2$	-1	1
T_2	$q(q^2+q^{-2})$	$-q^{-1}(q^2+q^{-2})$	q^3	$-q^{-3}$
X_1	$q^{-3/2}$	0	0	$q^{1/2}$
X_{-1}	0	$-q^{3/2}$	$-q^{-1/2}$	0
X_2	0		0	$-q^{-5/2}$
X_{-2}		0	$q^{5/2}$	0
X_{12}	0	$-q^{-5/2}$	0	$-q^{-1}T_1+T_2$
X_{-12}	$q^{5/2}$	0	qT_1-T_2	0

FIG. 1. The brackets of $\mathfrak{sl}(3)_q$.

$$\text{ad } F_i(T_j)=\begin{cases}(q+q^{-1})X_{-i}, & \text{if } i=j,\\ X_{-12}, & \text{if } i\neq j.\end{cases}$$

From this the adjoint brackets between the basis elements can be calculated; the result is given in Fig. 1, in which $[X,Y]$ is given in the row labelled by X and the column labelled by Y. If $\alpha+\beta\neq 0$, $[X_\alpha,X_\beta]=c_{\alpha\beta}X_{\alpha+\beta}$ is represented by the coefficient $c_{\alpha\beta}$; similarly, $[T_i,X_\alpha]=b_{i\alpha}X_\alpha$ is represented by $b_{i\alpha}$.

The entries which are too long to fit into the table are

$$[T_2,T_2]=-(q^2-q^{-2})T_1+(q^3-q^{-3})T_2,$$

$$[X_2,X_{-2}]=q^{-1}(q-q^{-1})T_1-qT_2,$$

$$[X_{-2},X_2]=q(q-q^{-1})T_1+q^{-1}T_2.$$

Notice that if L is regarded as a module over $\mathbb{K}[q,q^{-1}]$ then, with this basis, the brackets exhibit the simple antisymmetry which was also found by Delius and Hüffmann. Define a map of *q-conjugation* on L, $X\to X^{\blacktriangledown}$, by the requirements that it leaves the above basis elements fixed and that $\{f(q)X\}^{\blacktriangledown}=f(q^{-1})X^{\blacktriangledown}$. Then it is apparent from the table that

$$[Y,X]=-[X,Y]^{\blacktriangledown}, \quad \text{for all } X,Y\in L. \tag{4.1}$$

The generalized antisymmetrizer of $\overline{U}_q(\mathfrak{g})$ can be described as follows. The generalized Lie algebra $L=\mathfrak{sl}(3)_q$ carries an irreducible representation of $\overline{U}_q(\mathfrak{g})$; as in the classical case, $L\otimes L$ splits into irreducible representations as

$$L\otimes L=M_{27}\oplus M_{10}\oplus M_{10}^{*}\oplus M_8^{\mathrm{a}}\oplus M_8^{\mathrm{s}}\oplus M_1,$$

where the numerical labels 27, 10, 8, 1 indicate the dimensions of the simple modules. These are completely defined by their highest-weight elements:

$$\begin{aligned}
W_{27}&=X_{12}\otimes X_{12},\\
W_{10}&=q^{1/2}X_1\otimes X_{12}-q^{1/2}X_{12}\otimes X_1,\\
W_{10}^{*}&=q^{1/2}X_2\otimes X_{12}-q^{-1/2}X_{12}\otimes X_2,\\
W_8^{\mathrm{s}}&=\{q^{5/2}(q+q^{-1})+q^{-5/2}\}X_1\otimes X_2+\{q^{-5/2}(q+q^{-1})+q^{5/2}\}X_2\otimes X_1\\
&\quad-(q^4T_1+q^{-1}T_2)\otimes X_2-X_2\otimes(q^{-4}T_1+qT_2),\\
W_8^{\mathrm{a}}&=q^{3/2}X_1\otimes X_2-q^{-3/2}X_2\otimes X_1-q(qT_1-T_2)\otimes X_{12}+q^{-1}X_{12}\otimes(q^{-1}T_1-T_2),
\end{aligned} \tag{4.2}$$

$$\begin{aligned}
W_1&=T_1\otimes T_2+T_2\otimes T_1-(q+q^{-1})(T_1\otimes T_1+T_2\otimes T_2)+(q^2+1+q^{-2})(q^{-1}X_1\otimes X_{-1}+qX_{-1}\otimes X_1\\
&\quad+q^{-1}X_2\otimes X_{-2}+qX_{-2}\otimes X_2-q^{-2}X_{12}\otimes X_{-12}-q^2X_{-12}\otimes X_{12}).
\end{aligned}$$

Let Π_{27}, Π_{10}, Π_{10}^{*}, Π_8^{s}, Π_8^{a}, and Π_1 denote the projectors onto these simple modules. Then we have the following.

Lemma 4.1: The generalized antisymmetrizer of $\overline{U}_q(\mathfrak{g})$ *is*

$$\gamma=\frac{(q^2+1)\Pi_{10}^{*}+(q^2+q^{-2})\Pi_8^{a}+(1+q^{-2})\Pi_{10}}{q^2-1+q^{-2}}. \tag{4.3}$$

Proof: Straightforward Hopf-algebra manipulations show that γ commutes with the diagonal adjoint action of $\mathcal{U}=U_q(\mathfrak{sl}(3))$ on $L\otimes L$, i.e., with $(\mathrm{ad}\otimes\mathrm{ad})\circ\Delta(u)$ for all $u\in\mathcal{U}$. Thus it is only necessary to calculate the action of γ on the highest-weight elements listed above. This is not too laborious for the cases of W_{10}, W_{10}^{*}, and W_{27}. For the octets the relevant highest-weight elements (eigenvectors of γ) and the eigenvalues can be determined from the Jacobi identity: in the two-dimensional space of highest-weight octet vectors [elements of $L\otimes L$ with weight H_1+H_2 which are annihilated by $(\mathrm{ad}\otimes\mathrm{ad})\circ\Delta(E_1)$ and $(\mathrm{ad}\otimes\mathrm{ad})\circ\Delta(E_2)$], the eigenvectors of γ are the relative eigenvectors of $\mathrm{ad}\circ\beta$ and $m\circ(\mathrm{ad}\otimes\mathrm{ad})$.

The eigenvalue 0 for the singlet can be verified in a similar way. We find that $\beta(W_1)=0$ but $m\circ(\mathrm{ad}\otimes\mathrm{ad})(W_1)\neq 0$. Since $\mathrm{ad}\circ\beta(W_1)=m\circ(\mathrm{ad}\otimes\mathrm{ad})\circ\gamma(W_1)$ and $\gamma(W_1)$ must be a multiple of W_1, it follows that $\gamma(W_1)=0$. □

The quantized enveloping algebra $\mathcal{U}=\overline{U}_q(\mathfrak{sl}(3))$ has a symmetry between (E_1,F_1,H_1) and (E_2,F_2,H_2) corresponding to the outer automorphism of $\mathfrak{sl}(3)$ which comes from the symmetry of the Dynkin diagram ∘—∘. The generalized Lie algebra $\mathcal{L}=\mathfrak{sl}(3)_q=\mathrm{ad}\mathcal{U}(q^{-4\lambda})$ is not invariant under this automorphism of $\mathcal{U}$. Indeed, the weight diagram of the 3-dimensional representation is taken by this symmetry to that of the conjugate representation, with highest weight $\lambda^{*}\neq\lambda$. This gives another generalized Lie algebra $L^{*}\subset\mathrm{ad}\mathcal{U}(q^{-4\lambda^{*}})$ with $L^{*}\neq L$. This is another octet (under

the adjoint representation) which also generates the locally finite part of $\mathcal{U}$. It is interesting to express the elements of L^* in terms of those of L and to compare this situation with the classical one.

In the classical enveloping algebra the octets are the Lie algebra itself, one which is quadratic in the generators, and multiples of these by functions of the Casimirs. The outer automorphism acts linearly on the generators. In the locally finite part of the quantized enveloping algebra we have the same general structure;[6] the quadratic octet has highest-weight element obtained by multiplying the factors of the tensor W_8^s in (4.2). The highest-weight element of the octet L is X_{12}. Let X_{12}^* be the highest-weight element of L^*, and let $Y_{12}=m(W_8^s)$. Then a calculation gives

$$X_{12}^*=\frac{q^{1/2}(q+q^{-1})CX_{12}+q(q-q^{-1})Y_{12}}{(q^{1/2}+q^{-1/2})(q^2+1+q^{-2})}. \tag{4.4}$$

Thus the elements of L^* are quadratic in the generators L, and the quantum counterpart of the outer automorphism is a nonlinear symmetry of $\overline{U}_q(\mathfrak{sl}(3))$.

APPENDIX: PROOFS OF LEMMAS 2.1 AND 2.2

Lemma 2.1: The central element in $\overline{L}$ is given by

$$C_\lambda=\sum_{r=0}^{n-1}(-1)^r\frac{[n-r]_q}{[n]_q}K_r, \tag{2.19}$$

where K_r is defined recursively by

$$K_r=\mathrm{ad}(F_rE_r)K_{r-1},\quad K_0=q^{-4\lambda}. \tag{2.20}$$

Proof: Let C be defined by (2.19). Since this has zero weight for the Cartan algebra generated by $H_1,\ldots,H_{n-1}$ [i.e., ad $q^{H_i}(C)=C$], in order to show that it is central it is sufficient to show that ad $E_i(C)=0$ for $i=1,\ldots,n-1$. We calculate ad $E_i(K_j)$.

Since ad $\overline{U}_q(\mathfrak{g})(q^{-4\lambda})$ is the direct sum of the q-analogues of the scalar and the adjoint modules of $\mathfrak{sl}(n)$, the structure of the roots of $\mathfrak{sl}(n)$ gives

$$\mathrm{ad}\,E_j\ \mathrm{ad}\,E_i\ \mathrm{ad}\,E_{i-1}\cdots\mathrm{ad}\,E_1(q^{-4\lambda})=0\quad\text{unless}\quad j=i+1. \tag{A1}$$

Hence

$$\mathrm{ad}\,E_j(K_i)=0\quad\text{if}\quad j>i+1\quad\text{and}\quad i>0.$$

This also holds if $i=0$ since $q^{-4\lambda}$ commutes with E_j for $j>1$.

For $i=j$ we have

$$\mathrm{ad}\,E_i(K_i)=\mathrm{ad}(F_iE_i+[2H_i]_q)\mathrm{ad}\,E_i(K_{i-1})=[\langle 2H_i,H_i\rangle]_q\ \mathrm{ad}\,E_i(K_{i-1})=[2]_q\ \mathrm{ad}\,E_i(K_{i-1}),$$

since

$$\begin{aligned}(\mathrm{ad}\,E_i)^2K_{i-1}&=\mathrm{ad}\,F_{i-1}\cdots\mathrm{ad}\,F_1(\mathrm{ad}\,E_i)^2\ \mathrm{ad}\,E_{i-1}\cdots\mathrm{ad}\,E_1(q^{-4\lambda})\\&=0\quad\text{by (A1).}\end{aligned}$$

For $j=i-1$ we have

$$\begin{aligned}\mathrm{ad}\,E_j(K_{j+1})&=\mathrm{ad}\,F_{j+1}\ \mathrm{ad}(F_jE_j+[2H_j])\mathrm{ad}\,E_{j+1}\ \mathrm{ad}\,E_j(K_{j-1})\\&=\mathrm{ad}([2H_j-1]_q)\mathrm{ad}\,F_{j+1}\ \mathrm{ad}\,E_{j+1}\ \mathrm{ad}\,E_j(K_{j-1})\\&=\mathrm{ad}(F_{j+1}E_{j+1}E_j)K_{j-1}\ \text{ since }\ [2\langle H_j,H_j\rangle-1]_q=1\\&=\mathrm{ad}[(E_{j+1}F_{j+1}-[2H_{j+1}]_q)E_j]K_{j-1}\end{aligned}$$

$$= \operatorname{ad} E_j(K_{j-1}), \tag{A2}$$

since $\operatorname{ad} F_{j+1} \operatorname{ad} E_j(K_{j-1}) = 0$ by the structure of the roots of $\mathfrak{sl}(n)$.

For $j < i-1$,

$$\begin{aligned}
\operatorname{ad} E_j(K_i) &= \operatorname{ad}(F_i \cdots F_{j+1} E_j F_j \cdots F_1 E_i \cdots E_1) q^{-4\lambda} \\
&= \operatorname{ad}[F_i \cdots F_{j+1}(F_j E_j + [2H_j]_q) F_{j-1} \cdots F_1 E_i \cdots E_1] q^{4\lambda} \\
&= [\langle 2H_j, H_{j+1} + \cdots + H_i \rangle]_q \operatorname{ad}(F_i E_i \cdots F_{j+1} E_{j+1} E_j) K_{j-1} \quad \text{by (A1)} \\
&= -\operatorname{ad}(F_i E_i \cdots F_{j+2} E_{j+2} E_j) K_{j-1} \quad \text{by (A2)} \\
&= 0 \quad \text{by (5.1)}.
\end{aligned}$$

To summarize,

$$\operatorname{ad} E_j(K_i) = 0 \quad \text{if } i < j-1 \text{ or } i > j+1,$$

$$\operatorname{ad} E_j(K_{j+1}) = \operatorname{ad} E_j(K_{j-1}), \tag{A3}$$

and

$$\operatorname{ad} E_j(K_j) = [2]_q \operatorname{ad} E_j(K_{j-1}).$$

Hence for $j = 1, \ldots, n-2$,

$$\begin{aligned}
\operatorname{ad} E_j(C) &= \frac{(-1)^{j-1}}{[n]_q}([n-j-1]_q - [2]_q[n-j]_q + [n-j+1]_q) \operatorname{ad} E_j(K_{j-1}) \\
&= 0,
\end{aligned}$$

while

$$\operatorname{ad} E_{n-1}(C) = \frac{(-1)^{n-1}}{[n]_q}([2]_q - [2]_q[1]_q) = 0.$$

Thus C is a highest-weight vector in the module $\overline{L}$. Since its weight is zero, it follows that C is an invariant element [$\operatorname{ad} x(C) = \varepsilon(x)C$ for all $x \in \mathcal{U}$] and therefore central. It is therefore a multiple of C_λ. But $C = q^{-4\lambda} + w$ with $w \in L$; hence $C = C_\lambda$. □

Lemma 2.2: The adjoint action of the central element on L is

$$\operatorname{ad} C_\lambda(x) = (q^2 - 1 + q^{-2})x, \quad \text{for } x \in L. \tag{A4}$$

Proof: It follows from Lemma 2.1 by Schur's lemma that $\operatorname{ad} C_\lambda$ acts as a multiple of the identity on the irreducible component L. To find the multiple we calculate $\operatorname{ad} C_\lambda(X_1)$ for $X_1 = \operatorname{ad} E_1(q^{-4\lambda})$. We use

$$\operatorname{ad} E_r(X_1) = 0 \quad \text{unless } r = 2,$$

$$\operatorname{ad} F_r(X_1) = 0 \quad \text{unless } r = 1,$$

$$\operatorname{ad} F_1(X_1) = K_1.$$

Then we have

$$\operatorname{ad} K_0(X_1) = q^{-\langle 4\lambda, H_1 \rangle} X_1 = q^{-2} X_1.$$

Now

$$\operatorname{ad} E_1(q^{-4\lambda}) = (q - q^{-1}) q^{-4\lambda + H_1} E_1,$$

so

$$\begin{aligned} K_1 &= (q-q^{-1})\,\mathrm{ad}\,F_1(q^{-4\lambda+H_1}E_1) \\ &= -(q-q^{-1})q^{-4\lambda+2H_1}(qE_1F_1-q^{-1}F_1E_1) \\ &= (q-q^{-1})^2 q^{-4\lambda+2H_1}F_1E_1 - q.q^{-4\lambda}(q^{4H_1-1}). \end{aligned} \tag{A5}$$

Therefore

$$\mathrm{ad}\,K_1(X_1) = -q(q^{\langle -4\lambda+4H_1,H_1\rangle}-q^{\langle -4\lambda,H_1\rangle})X_1 = -q(q^2-q^{-2})X_1. \tag{A6}$$

To calculate ad $K_2(X_1)$, note that

$$\mathrm{ad}\,E_2(K_1) = E_2K_1q^{H_2}-q^{H_2-1}K_1E_2 = q^{H_2-1}(E_2K_1-K_1E_2).$$

Since ad $F_2(X_1)=0$, we have

$$\mathrm{ad}\,K_2(X_1) = \mathrm{ad}[F_2\cdot \mathrm{ad}\,E_2(K_1)\cdot q^{H_2}]X_1 = \mathrm{ad}[q^{2H_2}F_2(E_2K_1-K_1E_2)]X_1.$$

Using (A6) and the fact that ad $F_2(X_1)=0$,

$$\mathrm{ad}[q^{2H_2}F_2K_1E_2]X_1 = -q\ \mathrm{ad}[q^{2H_2}(E_2F_2-[2H_2]_q)]X_1 = -(q^2-q^{-2})X_1;$$

using (A5) and the fact that $\mathrm{ad}(E_1E_2)X_1=0$,

$$\begin{aligned} \mathrm{ad}[q^{2H_2}F_2K_1E_2]X_1 &= -q\ \mathrm{ad}[q^{2H_2}F_2q^{-4\lambda}(q^{4H_1}-1)E_2]X_1 \\ &= -q\ \mathrm{ad}[q^{2H_2-4\lambda}(q^{4H_1-2}-1)(E_2F_2-[2H_2]_q)]X_1 \\ &= -q^{-2}(q^2-1)X_1; \end{aligned}$$

$$\text{therefore}\ \ \mathrm{ad}\,K_2(X_1) = -(q^2-1)X_1.$$

We now have that K_3 is a sum of products of ad E_3 and ad F_3, both of which annihilate X_1, and operators of which X_1 is an eigenvector; it follows that ad $K_3(X_1)=0$. Similarly ad $K_r(X_1)=0$ for $r>3$; hence

$$\mathrm{ad}\,C_\lambda(X_1) = \left(q^{-2}+q(q-q^{-1})\,\frac{(q+q^{-1})[n-1]_q-[n-2]_q}{[n]_q}\right)X_1 = (q^2-1+q^{-2})X_1.$$

□

[1] S. Majid, ''Quantum and braided Lie algebras,'' J. Geom. Phys. **13**, 307–356 (1994).
[2] G. W. Delius and A. Hüffmann, ''On quantum Lie algebras and quantum root systems,'' J. Phys. A **29**, 1703–1722 (1996).
[3] G. W. Delius, A. Hüffmann, M. D. Gould, and Y.-Z. Zhang, ''Quantum Lie algebras associated to $U_q(gl_n)$ and $U_q(sl_n)$,'' J. Phys. A **29**, 5611–5617 (1996).
[4] A. Sudbery, ''$SU_q(n)$ gauge theory,'' Phys. Lett. B **375**, 75–80 (1996).
[5] S. L. Woronowicz, ''Differential calculus on compact matrix pseudogroups (quantum groups),'' Commun. Math. Phys. **122**, 125–170 (1989).
[6] A. Joseph and G. Letzter, ''Separation of variables for quantized enveloping algebras,'' Am. J. Math. **116**, 127–177 (1994).
[7] A. Sudbery, ''The algebra of differential forms on a full matric bialgebra,'' Math. Proc. Camb. Philos. Soc. **114**, 111–130 (1993).
[8] P. Schupp, P. Watts, and B. Zumino, ''Differential geometry on linear quantum groups,'' Lett. Math. Phys. **25**, 139–147 (1992).
[9] K. Schmüdgen and A. Schüler, ''Covariant differential calculi on quantum spaces and on quantum groups,'' C. R. Acad. Sci. Paris **316**, 1155–1160 (1993).
[10] B. Jurčo, ''Differential calculus on quantized simple Lie groups,'' Lett. Math. Phys. **22**, 177–186 (1991).
[11] A. Sudbery, ''The quantum orthogonal mystery,'' *Quantum Groups: Formalism and Applications*, edited by J. Lukierski, Z. Popowicz, and J. Sobczyk (Polish Scientific PWN, Warsaw, 1995), pp. 303–316.
[12] U. Carow-Watamura, M. Schlieker, S. Watamura, and W. Weich, ''Bicovariant differential calculus on quantum groups $SU_q(N)$ and $SO_q(N)$,'' Commun. Math. Phys. **142**, 605–641 (1991).

[13] K. Schmüdgen and A. Schüler, ''Classification of bicovariant differential calculi on quantum groups,'' Commun. Math. Phys. **170**, 315–335 (1995).
[14] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994).
[15] S. L. Woronowicz, ''Solutions of the braid equation related to a Hopf algebra,'' Lett. Math. Phys. **23**, 143–145 (1991).
[16] G. M. Bergman, ''The diamond lemma for ring theory,'' Adv. Math. **29**, 178–218 (1978).

Symmetry classification and exact solutions of the Kramers equation

Stanislav Spichak and Valerii Stognii
Institute of Mathematics of the National Academy of Sciences of Ukraine, 3 Tereshchenkivs'ka Street, Kyiv 4, Ukraine, and Kyiv Polytechnical Institute, 37 Pobedy Avenue, Kyiv, Ukraine

(Received 15 April 1997; accepted for publication 16 September 1997)

The complete classification of Lie symmetries for the Kramers equation is described. Group invariance under infinitesimal transformations is used to generate a wide class of exact solutions. Some of the known solutions are obtained as particular cases. The general results can also be used to solve some types of moving-boundary problems. We researched some Q-conditional symmetry properties of this equation. © *1998 American Institute of Physics.* [S0022-2488(98)01505-9]

I. INTRODUCTION

The one-dimensional Fokker–Planck equation has the form (see Ref. 1)

$$\frac{\partial u(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x,t)u(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[B(x,t)u(x,t)],$$

where $u=u(t,x)$ is a probability density and A and B are differential functions. This is the basic equation in the theory of continuous Markovian process and serves as a mathematical model for a number of problems in physical and biological sciences (see Refs. 1–10).

In the case of several space variables, the Fokker–Planck equation describes many more different complex processes then the one-dimensional equation. In the general case, the Fokker–Planck equation can be written in the form

$$\frac{\partial u(x,t)}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}[A_i(x,t)u(x,t)] + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}[B_{ij}u(x,t)],$$

where $x=(x_1,\ldots,x_n)$, the coefficient of drift $A(x,t)$, and coefficient of diffusion $B(x,t)$ are defined as the vector and matrix,

$$A(x,t)=(A_1(x,t),\ldots,A_n(x,t)), \quad B(x,t)=\|B_{ij}(x,t)\|, \quad i,j=1,\ldots,n, \tag{1}$$

correspondingly. Now consider the two-dimensional Fokker–Planck equation that describes the motion of a particle in a fluctuating medium (see Ref. 1):

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}(yu) + \frac{\partial}{\partial y}(V'(x)u) + \gamma\frac{\partial}{\partial y}\left(yu + \frac{\partial u}{\partial y}\right), \tag{2}$$

where $u=u(t,x,y)$ is a probability density, γ is a constant, and $V(x)$ is an external potential. The equation (2) is known as Kramers equation. The coefficients of drift and diffusion (1) can be written in the form

$$A(t,x,y)=(y,-V'(x)-\gamma y), \quad B(t,x,y)=\begin{pmatrix}0 & 0\\ 0 & 2\gamma\end{pmatrix}.$$

II. LIE'S SYMMETRY ANALYSIS

We investigate symmetry properties of the equation (2) for different potentials under the infinitesimal local operators,

 © 1998 American Institute of Physics

$$Q=\xi^0(t,x,y,u)\,\frac{\partial}{\partial t}+\xi^1(t,x,y,u)\,\frac{\partial}{\partial x}+\xi^2(t,x,y,u)\,\frac{\partial}{\partial y}+\eta(t,x,y,u)\,\frac{\partial}{\partial u}. \tag{3}$$

Theorem 1: *The maximal invariance local group of the free Kramers equation (2) [with* $V'(x)=0$*] is the six-dimensional Lie group generated by the following operators:*

$$\begin{gathered}P_0=\partial_t=\frac{\partial}{\partial t},\quad P_1=\partial_x=\frac{\partial}{\partial x},\quad I,G_1=t\partial_x+\partial_y+\frac{1}{2}\,(y+\gamma x),\\ S_1=\exp(\gamma t)\left(\frac{1}{\gamma}\,\partial_x+\partial_y+y\right),\quad T_1=\exp(-\gamma t)\left(\frac{1}{\gamma}\,\partial_x-\partial_y\right).\end{gathered} \tag{4}$$

Proof: Applying the Lie's algorithm (see Refs. 11 and 12) to the equation (2), we find that

$$\begin{gathered}\xi^0=F(t),\quad \xi^1=\frac{3}{2}\,xF'(t)+G(t),\quad \xi^2=\frac{1}{2}\,yF'+\frac{3}{2}\,xF''+G',\\ \eta=-\frac{1}{2\gamma}\left[(2F''+\gamma F')\,\frac{y^2}{2}+y\left(\frac{3}{2}\,x(F'''+\gamma F'')+G''+\gamma G'+\frac{3}{2}\,xF'V''+V''G+\frac{1}{2}\,V'F'\right)\right.\\ \left.+\Theta(t,x),\right.\end{gathered} \tag{5}$$

with $F(t)$, $G(t)$, and $\Theta(t,x)$ satisfying the equations

$$\begin{gathered}5F'''-2\gamma^2F'+4V''F'+3xV'''F'+2V'''G=0,\\ \Theta_x+\tfrac{3}{2}\,x(-F''''+\gamma^2F'')-\tfrac{3}{2}\,xV''F''+\tfrac{3}{2}\,\gamma xV''F'+\tfrac{3}{2}\,V'F''+\tfrac{3}{2}\,\gamma V'F'-G'''+\gamma^2G'-V''G'\\ +\gamma V''G=0,\\ \Theta_t+V'(\tfrac{3}{2}\,x(F'''+\gamma F'')+(G''+\gamma G')+\tfrac{3}{2}\,xV''F'+V''G+\tfrac{1}{2}\,V'F')+2\gamma F''-\gamma^2F'=0.\end{gathered} \tag{6}$$

Let us put in (5), (6), $V'(x)=0$. Solving this equation we obtain from (3) and (5) the operators (4).

Remark 1: The change of variables,

$$u=w(\tau,\xi,\eta),\quad \tau=t,\quad \xi=x-\left(\frac{c}{\gamma}\right)t,\quad \eta=y-\frac{c}{\gamma},$$

reduces the equation (2) with $V'(x)=C=\text{const}$ to the free Kramers equation.

Remark 2: The change of variables,

$$u=w(\tau,\xi,\eta),\quad \tau=t,\quad \xi=x+\frac{c}{k},\quad \eta=y,\quad k\neq 0,$$

reduces the equation (2) with $V'(x)=kx+C$ to the equation (2) with $V'(x)=kx$.

Theorem 2: *The maximal invariance local group of the equation (2) with* $V'(x)=kx$ *when* $k\neq-\frac{3}{4}\gamma^2$, $\frac{3}{16}\gamma^2$ *is the six-dimensional Lie group.*

(a) *In the case* $k<\gamma^2/4$, *the correspondent symmetry operators are*

$$P_0,I,X_i=\exp(\lambda_i t)\left(\frac{\partial}{\partial x}+\lambda_i\,\frac{\partial}{\partial y}+\frac{1}{2\gamma}\,[(\lambda_i^2+\gamma\lambda_i+k)y+(-\lambda_i^3+\gamma^2\lambda_i-k\lambda_i+\gamma k)x]\right), \tag{7}$$

$i=1,2,3,4$,
where λ_i *are the roots of the equation*

$$\lambda^4+(2k-\gamma^2)\lambda^2+k^2=0. \tag{8}$$

(b) *In the case* $k>\gamma^2/4$ *the corresponding operators are*

$$P_0, I, Y_i, \quad k=1,2,3,4, \tag{9}$$

where $Y_1=\mathrm{Re}\, X_1$, $Y_2=\mathrm{Im}\, X_1$, $Y_3=\mathrm{Re}\, X_3$, $Y_4=\mathrm{Im}\, X_3$, X_i, *and the operators* X_i *are those in (7);* λ_i *are complex roots of the equation (8).*

(c) *In the case* $k=\gamma^2/4$, *the corresponding operators are*

$$\begin{gathered} P_0, I, X_i=\exp(\lambda_i t)\left(\frac{\partial}{\partial x}+\lambda_i\frac{\partial}{\partial y}+\frac{\gamma y}{2}+\frac{\gamma^2}{4}x\right), \quad i=1,3, \\ X_i=\exp(\lambda_i t)\left(t\frac{\partial}{\partial x}+(1+\lambda_i t)\frac{\partial}{\partial y}+\left(\frac{\gamma t}{2}+1\right)y+\frac{\gamma^2}{2}tx\right), \quad i=2,4, \end{gathered} \tag{10}$$

where $\lambda_2=\lambda_1$, $\lambda_3=\lambda_4=-\lambda_1$, λ_1 *are the roots of the equation (8).*

Theorem 3: *The maximal invariance local group of the equation (2) with* $V'(x)=kx$ *when* $k=-\frac{3}{4}\gamma^2$ *and* $k=\frac{3}{16}\gamma^2$ *is the eight-dimensional Lie group generated by the operators (7) and in addition the following operators: (a) in the case* $k=-\frac{3}{4}\gamma^2$,

$$\begin{gathered} N_1=\exp(\gamma t)\left(\partial_t+\frac{3}{2}\gamma x\partial_x+\frac{\gamma}{2}(3\gamma x+y)\partial_y+\frac{3}{4}\gamma y^2+\frac{3}{4}\gamma^2 xy-\frac{9}{16}\gamma^3x^2+\frac{\gamma}{2}\right), \\ N_2=\exp(-\gamma t)\left(\partial_t+\frac{3}{2}\gamma x\partial_x+\frac{\gamma}{2}(3\gamma x-y)\partial_y+\frac{1}{4}\gamma y^2+\frac{3}{4}\gamma^2 xy+\frac{9}{16}\gamma^3x^2-\frac{3}{2}\gamma\right); \end{gathered} \tag{11}$$

(b) in the case $k=\frac{3}{16}\gamma^2$,

$$\begin{gathered} N_1=\exp\left(\frac{1}{2}\gamma t\right)\left(\partial_t+\frac{3}{4}\gamma x\partial_x+\frac{\gamma}{4}\left(\frac{3}{2}\gamma x+y\right)\partial_y+\frac{1}{4}\gamma y^2+\frac{3}{8}\gamma^2 xy+\frac{9}{64}\gamma^3x^2\right), \\ N_2=\exp\left(-\frac{1}{2}\gamma t\right)\left(\partial_t-\frac{3}{4}\gamma x\partial_x+\frac{\gamma}{4}\left(\frac{3}{2}\gamma x-y\right)\partial_y+\gamma\right). \end{gathered} \tag{12}$$

Theorem 4: *The maximal invariance local group of the equation (2) with* $V'(x)\neq kx+c$ *is a two-dimensional Lie group generated by the operators* $P_0=\partial t, I$.

The proofs of Theorems 2, 3, and 4 follow from the solutions of the equations (5),(6) for correspondent potentials.

III. *Q*-CONDITIONAL SYMMETRY

In this section we present some results on Q-conditional symmetry of the Kramers equation. The term and the concept of Q-conditional symmetry of an equation was introduced in Refs. 12, 13, and 14.

Let us consider the operator (3) and the additional condition

$$Qu=0. \tag{13}$$

We say that the equation (2) is Q-conditionally invariant if the system (2), (13) is invariant under the operator (3). In particular, the Lie symmetry operator Q is a Q-conditional symmetry. The Q-conditional symmetry operator always leads to a reduction of an initial equation as well as a Lie symmetry one. So the possibility of finding exact solutions is expanded.

For example, let us consider the equation (2) with the potential

$$V'=kx^{-1/3}+\tfrac{3}{16}\gamma^2x, \quad k\neq 0. \tag{14}$$

The operator,

$$Q=\partial_t-\tfrac{3}{4}\gamma x\partial_x+\left(\tfrac{3}{8}\gamma^2-\tfrac{1}{4}\gamma y\right)\partial_y+\gamma u\partial_u, \tag{15}$$

is not a Lie symmetry operator of the equations (2), (14), as follows from Theorem 4. However, the operator (15) gives the invariant solution (ansatz)

$$u(t,x,y)=\exp(\gamma t)\varphi(\omega_1,\omega_2),\quad \omega_1=\exp(\tfrac{3}{4}\gamma t)x,\quad \omega_2=x^{-1/3}y+\tfrac{3}{4}\gamma x^{2/3},\tag{16}$$

which reduces the equation. Indeed, substituting (16) into (2), (14) we obtain the reduced equation,

$$\omega_2\varphi_1-\left(\frac{\omega_2^2}{2}+k\right)\omega_1^{-1}\varphi_2-\gamma\omega_1^{-1}\varphi_{22}=0,\quad \varphi_i=\frac{\partial\varphi}{\partial\omega_i},\quad \varphi_{22}=\frac{\partial^2\varphi}{\partial\omega_2\partial\omega_2}.$$

This equation may be integrated, in particular, when $\varphi=\varphi(\omega_2)$.

Remark 3: If Q is a Q-conditional operator for a PDE, then the equation is $\overline{Q}$-conditionally invariant under $\overline{Q}=f(x,u)Q$, where $f(x,u)$ is an arbitrary function of the independent and dependent variables. We say that Q and $\overline{Q}$ are equivalent operators. Then we may consider $\xi^0=1$ in (2) if $\xi^0\not\equiv 0$.

Theorem 5: *All Q-conditional operators of the equation (2) with $V'\neq kx^{-1/3}+\frac{3}{16}\gamma^2 x$, kx are equivalent to ∂_t, I.*

Theorem 6: *The equations (2) and (14) have the following Q-conditional symmetry operators*:

$$Q=\partial_t+F(t)x\partial_x+(\tfrac{1}{3}F(t)y+F'(t)x+\tfrac{2}{3}F^2x)\partial_y+fu\partial_u,$$

where $F(t)$ is an arbitrary solution of the equation

$$F''+2FF'+\tfrac{4}{9}F^3-\tfrac{1}{4}\gamma^2F=0,$$

$$2\gamma f=-y^2\left(\frac{2}{3}F'+\frac{4}{9}F^2+\frac{1}{3}\gamma F\right)-yx\left(\gamma F'+\frac{2}{3}\gamma F^2+\frac{1}{2}\gamma^2F\right)-x^2\left(\frac{3}{8}\gamma^2F'+\frac{1}{4}\gamma^2F^2+\frac{3}{16}\gamma^3F\right)-x^{2/3}k\left(2F'+\frac{4}{3}F^2+\gamma F\right)+C\exp\left\{-\frac{2}{3}\int F\,dt\right\}-\frac{4}{3}\gamma F+\gamma^2,$$

*where $C=$*const.

Theorem 7: *All Q-conditional symmetry operators $(\xi^0=1)$ of the equation (2) with $V'=kx$,* and $k\neq-\frac{1}{4}\gamma^2$, $\frac{3}{16}\gamma^2$ *are equivalent to the Lie symmetry operators given in (7), (9), or (10).*

Theorem 8: *All Q-conditional symmetry operators $(\xi^0=1)$ of the equation (2) with $V'=kx$, $k=-\frac{1}{4}\gamma^2$, $\frac{3}{16}\gamma^2$ (including those equivalent to Lie symmetry operators) have the following form*:

$$Q=\partial_t+(F(t)x+G(t))\partial_x+[\tfrac{1}{3}F(t)y+(F'(t)+\tfrac{2}{3}F(t)^2)x+G'(t)+\tfrac{2}{3}F(t)G(t)]\partial_y+f(t,x,y)u\partial_u.$$

Here the functions $F(t)$, $G(t)$ satisfy the following equations:

$$F''+2FF'+\tfrac{4}{9}F^3+(\tfrac{4}{5}k-\tfrac{2}{5}\gamma^2)F=0,$$

$$G''+\tfrac{4}{3}FG'+\tfrac{2}{3}F'G+\gamma G'+\tfrac{2}{3}\gamma FG+\tfrac{4}{9}F^2G+kG=h(t),$$

where $h(t)$ satisfies the equation

$$h''+(\tfrac{4}{3}F-\gamma)h'+(k+\tfrac{2}{3}F'+\tfrac{4}{9}F^2-\tfrac{2}{3}\gamma F)h=0.$$

The function $f(t,x,y)$ is

(a) $k=-\frac{3}{4}\gamma^2$, $2\gamma f=-y^2(\frac{2}{3}F'+\frac{4}{9}F^2+\frac{1}{3}\gamma F)-y[x(\gamma F'+\frac{2}{3}\gamma F^2)+h]$

$$+\tfrac{3}{4}\gamma^3x^2F+x(h'-\gamma h+\tfrac{2}{3}Fh)+s(t),$$

(b) $k=\tfrac{3}{16}\gamma^2$, $2\gamma f=-y^2(\tfrac{2}{3}F'+\tfrac{4}{9}F^2+\tfrac{1}{1}\gamma F)-y[x(\gamma F'+\tfrac{2}{3}\gamma F^2+\tfrac{1}{2})+h]$

$$-x^2(\tfrac{3}{16}\gamma^3F+\tfrac{3}{8}\gamma^2F'+\tfrac{1}{4}\gamma^2F^2)+x(h'-\gamma h+\tfrac{2}{3}Fh)+s(t),$$

where in both (a) and (b) the function $s(t)$ satisfies the equation

$$s'+\tfrac{2}{3}Fs=-2\gamma(\tfrac{2}{3}F'+\tfrac{4}{9}F^2+\tfrac{1}{3}\gamma F)+\tfrac{4}{3}\gamma^2F.$$

In conclusion, we consider the example Q_1-, Q_2-conditional symmetry of the Kramers equation with an arbitrary potential. It is easy to show that this equation is invariant under the two operators,

$$Q_1=\partial_x+V'(x),\quad Q_2=\partial_y+y,$$

under the additional conditions $Q_1u=Q_2u=0$. It turned out that the solution,

$$u=\exp\left\{-\frac{y^2}{2}-V(x)\right\}$$

[the well-known Boltzmann solution (see Ref. 1)], of the stationary Kramers equation (2) ($u_t=0$) with an arbitrary potential, is invariant under these operators.

From the above examples we see that further work on finding Q- and Q_1-, Q_2-conditional symmetry operators is of great interest.

IV. SOME EXACT SOLUTIONS

Now we consider the algorithms (see Refs. 11, 15, 16, and 17) of the construction of exact solutions of the Kramers equation (2).

Let us write down the final transformations generated by operators (7) ($k<\gamma^2/4$). The operator P_0 generates translation on variable t, I generates the identical transformation, X_i generates

$$t'=t,\quad x'=x+a_i\exp\{\lambda_i t\},\quad y'=y+a_i\lambda_i\exp\{\lambda_i t\},$$

$$u'(t',x',y')=\exp\left(-\frac{e^{\lambda_i t}}{2\gamma}\left[(\lambda_i^2+\gamma\lambda_i+k)ya_i+\frac{a_i^2}{2}\lambda_ie^{\lambda_i t}+(-\lambda_i^3+\gamma^2\lambda_i-k\lambda_i+\gamma k)xa_i+\frac{a_i^2}{2}e^{\lambda_i t}\right]\right)u(t,x,y), \tag{17}$$

where a_i are group parameters, and λ_i are roots of the equation (8).

It is appropriate to write here the corresponding formulas of generating solutions that follow from (17):

$$u_{\mathrm{II}}(t,x,y)=\exp\left\{\frac{e^{\lambda_i t}}{2\gamma}\left[(\lambda_i^2+\gamma\lambda_i+k)ya_i+\frac{a_i^2}{2}\lambda_ie^{\lambda_i t}+(-\lambda_i^3+\gamma^2\lambda_i-k\lambda_i+\gamma k)xa_i+\frac{a_i^2}{2}e^{\lambda_i t}\right]\right\}u_{\mathrm{I}}(t',x',y'), \tag{18}$$

where t',x',y' are given in (17), respectively.

A well-known solution of the stationary Kramers equation (2) is the Boltzman distribution,

$$u(x,y)=N\exp\{-V(x)-\tfrac{1}{2}y^2\} \tag{19}$$

(N is a normalization constant). Applying formulas (18) to (19) with $V'(x)=kx$, $k<\gamma^2/4$, one can easily obtain a nonstationary solution of the equation (2).

Now consider the free Kramers equation $[V'(x)=0]$ (see Ref. 18). Let us consider the operator from (4),

$$S_1=\exp(\gamma t)\left(\frac{1}{\gamma}\,\partial_x+\partial_y+y\right).$$

We find the invariants

$$\omega_1=t,\quad \omega_2=\gamma x-y,\quad \omega_3+\frac{y^2}{2}+\ln u,$$

and build anzatz

$$u(t,x,y)=\exp\left\{-\frac{y^2}{2}\right\}\varphi(\omega_1,\omega_2),\quad \omega_1=t,\quad \omega_2=\gamma x-y. \tag{20}$$

Substitution of (20) into the free Kramers equation gives rise to the heat equation,

$$\frac{\partial\varphi}{\partial\omega_1}-\gamma\,\frac{\partial^2\varphi}{\partial\omega_2^2}=0.$$

It is clear that by using solutions of the heat equation and the anzatz (20) one can construct many partial solutions.

A great number of partial solutions of the equation (2) can be found by means of the method described in Ref. 15. So, if $u_0(t,x,y)$ is a solution of the equation, then the functions Qu_0, $Q^2u_0,\ldots$ will also be solutions with any symmetry operator Q. For example, by means of the operators X_i (7) for the equation (2) with $V'(x)=kx$, $(k<\gamma^2/4)$, we find exact solutions beginning with $u_0=e^{\gamma t}$.

ACKNOWLEDGMENT

One of the authors (S.S.) is grateful to the DFFD of Ukraine (Project No. 1.4/356) for financial support.

[1] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 1985).
[2] G. A. Nariboli, *Stochastic Processes and their Applications 5* (North-Holland, Amsterdam, 1977), pp. 157–171.
[3] H. Risken, *The Fokker–Planck Equation* (Springer, Berlin, 1989).
[4] G. W. Bluman and I. D. Cole, *Similarity Methods for Differential Equations* (Springer, Berlin, 1974).
[5] G. Cicogna and D. Vitali, J. Phys. A **22**, 453 (1989).
[6] R. Ferrando, R. Spadacini, and G. Tommei, Phys. Rev. E **48**, 2437 (1993).
[7] T. Miyazawa, J. Math. Phys. **36**, 5643 (1995).
[8] C. C. A. Sastri and K. A. Dunn, J. Math. Phys. **26**, 3042 (1985).
[9] R. J. Wiltshire, J. Phys. A **27**, 7821 (1994).
[10] F. Wolf, J. Math. Phys. **29**, 305 (1988).
[11] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[12] P. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1986).
[13] G. W. Bluman and I. D. Cole, J. Math. Mech. **18**, 1025 (1969).
[14] W. I. Fushchich and I. M. Tsifra, J. Phys. A **20**, 45 (1987).
[15] W. Fushchych, W. Shtelen, and N. Serov, *Symmetry Analysis and Exact Solutions of Equations of Nonlinear Mathematical Physics* (Kluwer Academic, Dordrecht, 1993).
[16] N. H. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Reidel, Dordrecht, 1985).
[17] P. Winternitz, A. M. Grundland, and J. A. Tuszynski, J. Math. Phys. **28**, 2194 (1987).
[18] W. M. Shtelen and V. I. Stogny, J. Phys. A **22**, 539 (1989).

Erratum: ''Exact solutions of the Einstein equations with sources from linearized solutions'' [J. Math. Phys. 37, 4584–4589 (1996)]

G. F. Torres del Castillo
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, 72000 Puebla, Pue., México

(Received 22 January 1997; accepted for publication 4 February 1997)

[S0022-2488(98)01706-X]

By following the steps indicated in Ref. 1 after Eq. (11), one actually finds that $F_{\mu\nu}$ also satisfies the source-free Maxwell equations in the background metric $\tilde{g}_{\mu\nu}$ without any further condition. Therefore, Eq. (12) should not appear. This fact does not affect the results of the paper.

[1] G. F. Torres del Castillo, ''Exact solutions of the Einstein equations with sources from linearized solutions,'' J. Math. Phys. **37**, 4584–4589 (1996).

 © 1998 American Institute of Physics

Analytic regularization of the $(\bar{\psi}_i\psi_i)^2_D$ model at finite temperature and density

G. N. J. Añaños, A. P. C. Malbouisson, and N. F. Svaiter[a)]
Centro Brasileiro de Pesquisas Fisicas-CBPF, Rua Dr. Xavier Sigaud 150, Rio de Janeiro, RJ 22290-180 Brazil

(Received 17 November 1997; accepted for publication 3 March 1998)

Simultaneously using the dimensional and analytic regularization methods we reexamine the Gross–Neveu model at finite temperature and density (chemical potential) in a D-dimensional space–time. The regularized effective potential is presented at the $1/N$ (large N) approximation. It is found that in $D=3$ the effective potential is finite. © *1998 American Institute of Physics.*
[S0022-2488(98)02507-9]

I. INTRODUCTION

Besides computer simulations on the lattice, there have been many attempts over the years to go beyond the limitations of perturbative renormalizability in quantum field theory in the framework of continuous space–time formulation. For instance, much work has been done in the context of constructive field theory.[1] For scalar models, 't Hooft and Rivasseau[2] presented a construction of the ultraviolet limit for a massive Euclidean planar $(\lambda\varphi^4)_4$ model with the ''wrong'' sign of the coupling constant (yielding the loss of physical positivity). Since the number of planar graphs does not grow factorially with the perturbation order in λ (just a power behavior) the construction in this case can be done in a purely perturbative way. Only Feynman amplitudes matter for estimations. This result has been extended to the nonplanar case in the infrared region. Feldman *et al.* and Gawedzki and Kupianen[3] treated the model with a fixed ultraviolet cutoff and with the coupling constant with the ''good'' sign (leading to infrared asymptotic freedom), and showed that the model exists in the case of vanishing renormalized mass. There, the use of nonperturbative techniques such as polymer expansions and constructive renormalization have been applied to surmount the perturbative lack of convergence generated by ''instanton'' and ''renormalon'' singularities. These are, respectively, the factorial growth of the number of Feynman graphs and the factorial growth of a single renormalized amplitude due to renormalization of divergent subgraphs.

Also if an N-component (''colors'') field with a quartic coupling of intensity λ is considered, the introduction of an ultralocal intermediate field leads to the possibility of the summation of an infinite number of Feynman diagrams with an effective three-body coupling constant $\lambda/\sqrt{N}$ in the large-N limit. In this case, some of the more divergent diagrams are suppressed and an effective complete propagator is obtained. This is the basic idea underlying the so-called $1/N$ expansion. A well-known example where this idea is implemented is the four-fermion self-interaction model $(\bar{\psi}_i\psi_i)^2$, the Gross–Neveu model.[4] Although the Gross–Neveu model is perturbatively nonrenormalizable in $D=3$, at leading order in $1/N$, the resulting effective theory shares some of the basic requirements for a constructively renormalizable theory. This is due to the remarkable property of asymptotic safeness.[5] In other words, the effective coupling constant obtained from the effective propagator does not decrease to zero at high momenta: the model is not asymptotically free. Instead, it presents a nontrivial stable ultraviolet fixed point. Then by choosing N large enough de Calan *et al.*[5] have been able to show that there exist an N-independent λ^* such that $\lim_{N\to\infty}\lambda^*(N)=\lambda^*$. Based on this property it is possible to perform *convergent* phase space expansions for the Schwinger functions of a genuine nonrenormalizable model, although at the present time nothing can be said about whether or not these Schwinger functions verify the Osterwalder–Schrader axioms. From the point of view of the renormalization group, the model

a)Electronic mail: nfuxsvai@lafex.cbpf.br

 © 1998 American Institute of Physics

has been studied by many authors.[6] In this case it may be noted that, in general, associating an effective coupling constant which vanishes in some region of momentum space, asymptotic freedom prevents renormalon-like singularities to be present in the interesting part of the Borel plane. This means that the general summations of contributions defining the perturbative expansions can be done and the construction program can be carried out and, sometimes, Borel summability can be achieved giving *a posteriori* sense to perturbation methods.

In $D=2$, the massless model is asymptotically free and in any space–time dimension $D<4$ a chiral symmetry may be spontaneously broken with dynamical generation of mass. In the absence of chemical potential and at finite temperature the spontaneous broken symmetry may be restored via a second-order phase transition. For $D=2$ the condensation of kinks means that the phase transition may occur at any nonzero temperature. The question that is raised is whether the presence of the chemical potential changes the order of the phase transition.[7] In the literature there is no agreement on the subject since some results from lattice calculation indicate that for nonzero chemical potential the transition is of first order, while it is of second order for vanishing chemical potential.[8] Wolf obtained a different result in $D=2$, with a tricritical point separating second-order phase transition from first-order one. Recently, the four fermion model was investigated by Barducci *et al.*[9] where an explicitly broken chiral symmetry introduced by a fermionic mass term in the Lagrangian density was analyzed. These authors suggest that, depending on the value of the fermion mass, the transition could be of first or of second order. Also, indications for the existence of a tricritical phenomenon were found.

The main goal of this note is to investigate at the $(1/N)$ (large N) approximation the finite temperature behavior of the effective potential of the Gross–Neveu model. For more generality, we treat the model in arbitrary space–time dimension D, and to exhibit the analytic structure of the effective potential we use a mix between dimensional and zeta function analytic regularization methods. We determine that the regularized effective potential is free of any ultraviolet singularity in odd space–time dimensions. In the case of the Yukawa model the same kind of mathematical phenomenon was found in Ref. 10. This somewhat surprising result is to be compared to those obtained in the case of the calculation of the Casimir energy using similar regularization methods. As was suggested in Ref. 10, we also generalize in this note the analytical regularization procedure to the case of nonzero chemical potential. Recently Daicic and Frankel[11] investigated the magnetic properties of a Bose gas using very similar techniques. They investigated a Bose gas in the relativistic and nonrelativistic regime using a Mellin integral transform method to analyze the analytic structure of the thermodynamic potential. These authors claim that in this approach it is possible to present new results for a broad range of physical parameters such as temperature, field strength, etc., not revealed using another approach. We use the same methods to investigate the model in the case of nonzero chemical potential.

The outline of the paper is as follows: In Sec. II we briefly review the formalism of the effective potential, and the effective potential is obtained at zero and nonzero chemical potential. Conclusions are given in Sec. III. In this paper we use $\hbar=k_B=c=1$.

II. THE 1/*N* (LARGE *N*) EFFECTIVE POTENTIAL OF THE GROSS–NEVEU MODEL AT ZERO AND FINITE TEMPERATURE

Let us consider a system consisting of fermion fields in thermal equilibrium with a reservoir at temperature β^{-1}. They are defined on a $D=d+1$ dimensional flat space–time with trivial topology of the spacelike sections. The Gross–Neveu model is described in terms of a U(N) symmetric action for a set of N massless fermions

$$S(\bar{\psi},\psi)=\int d^Dx\left(\bar{\psi}(i\not{\partial})\psi+\frac{\lambda}{2N}(\bar{\psi}\psi)^2\right). \tag{1}$$

The symmetry which prevents the fermions from acquiring a mass in perturbation theory is

$$\psi\rightarrow\gamma^5\psi. \tag{2}$$

In an even dimension space–time the above transformation defines a discrete chiral symmetry. In an odd dimensional space–time a fermionic mass violates space parity.

The Green's functions of the model are generated by the generating functional

$$Z[\eta,\bar{\eta}]=c\int \mathscr{D}\psi\mathscr{D}\bar{\psi}\,\exp\left\{\int d^Dx(\mathscr{L}+\bar{\psi}\eta+\bar{\eta}\psi)\right\}, \tag{3}$$

where $\bar{\psi}(x)$, $\psi(x)$, $\bar{\eta}(x)$, and $\eta(x)$ are elements of a Grassmann algebra. The analysis of the symmetry behavior of the model is simplified by the introduction of a constraint field $\varphi(x)$ in the Lagrangian density, which becomes

$$\mathscr{L}'(\varphi,\psi,\bar{\psi})=\mathscr{L}(\psi,\bar{\psi})-\frac{1}{2}\left(\varphi+\left(\frac{\lambda}{N}\right)^{1/2}\bar{\psi}\psi\right)^2=\bar{\psi}(i\partial\!\!\!/)\psi-\left(\frac{\lambda}{N}\right)^{1/2}\varphi\bar{\psi}\psi-\frac{1}{2}\varphi^2. \tag{4}$$

Then we may write the generating functional given by Eq. (3) equivalently as

$$Z(\eta,\bar{\eta})=c'\int \mathscr{D}\psi\mathscr{D}\bar{\psi}\mathscr{D}\varphi\ \exp\, i\int d^Dx(\mathscr{L}'+\bar{\psi}\eta+\bar{\eta}\psi), \tag{5}$$

where $\mathscr{L}'(\varphi,\psi,\bar{\psi})$ is given by Eq. (4). To determine the equilibrium value of $\varphi(x)$ we define the generating functional of the connected Green functions $W(J)$ by

$$e^{iW(J)}=\int \mathscr{D}\psi\mathscr{D}\bar{\psi}\mathscr{D}\varphi\ \exp\left\{i\int d^Dx(\mathscr{L}'+J\varphi\right\}, \tag{6}$$

from which we get by a Legendre transform the effective action $\Gamma(\varphi_0)$,

$$\Gamma(\varphi_0)=\int d^Dx(\varphi_0(x)J(x)-W(J)) \tag{7}$$

where

$$\varphi_0(x)=\frac{\delta W(J)}{\delta J(x)}. \tag{8}$$

If we assume translational invariance the effective potential is given by

$$V(\varphi_0)=-\sum_n^{\infty}\ \frac{1}{n!}\,\tilde{\Gamma}^{(n)}(0,0,..0)(\varphi_0-\varphi_{\min})^n \tag{9}$$

or

$$V(\varphi_0)=\frac{1}{2}\,\varphi_0^2-iN\sum_{s=1}^{\infty}\int\ \frac{d^Dk}{(2\pi)^D}\,\frac{1}{s}\left(\frac{\lambda\varphi_0^2}{Nk^2}\right)^s. \tag{10}$$

As is well known, at zero temperature the minimum of the effective potential is not at the origin, there is a spontaneous symmetry breaking with a dynamical generation of mass associated to a bound state. To allow comparison with the results that we will get in the case of nonzero chemical potential, let us briefly review some basic features of the model at zero chemical potential.

If we assume that the system is in thermal equilibrium with a reservoir at temperature β^{-1} we may use the Matsubara formalism. In this case we have to perform the replacements $\omega\to\omega_n=2\pi/\beta(n+\frac{1}{2})$ and $(1/2\pi)\int dq_E^0=(1/\beta)\Sigma_n$. Defining $(\lambda/N)^{1/2}=g$, and using dimensional regularization to deal with ultraviolet divergences associated with the continuous Euclidean momenta, it is possible to obtain the effective potential in the form of a sum of Hurwitz zeta functions. In this case the effective potential is given by

$$V(\varphi_0,\beta)=\frac{1}{2}\,\varphi_0^2+2N\sum_{s=1}^{\infty}\,p(D,s)(g\varphi_0)^{2s}\beta^{2s-D}\zeta\left(2s-d,\,\frac{1}{2}\right), \tag{11}$$

where $p(D,s)$ is given by

$$p(D,s)=\frac{\pi^{d/2}}{\Gamma(s+1)}\,\Gamma\left(s-\frac{d}{2}\right)\,\frac{(-1)^s}{(2\pi)^{2s}}, \tag{12}$$

and the Hurwitz zeta function being defined by

$$\zeta(z,q)=\sum_{n=0}^{\infty}\,\frac{1}{(n+q^z)}, \tag{13}$$

which is analytic for $\mathrm{Re}(z)>1$.

Note that the effective potential is not yet well defined since the Hurwitz zeta function is defined only in an open connected set of points in the s complex plane. By a standard procedure it is possible to find the analytic extension of the Hurwitz zeta function to the whole s plane. To perform the analytic continuation of the Hurwitz zeta function we go along the following steps: first we use the Euler representation for the gamma function to express the Hurwitz zeta function as

$$\zeta(z,q)=\frac{1}{\Gamma(z)}\,\int_0^{\infty}dt\;t^{z-1}\,\frac{e^{-tq}}{1-e^{-t}}. \tag{14}$$

Next we split the integral from zero to infinity in two integrals from zero to one and from one to infinity. The second one is an analytic function, the divergences being associated with the zero limit of the first integration. Then using a Bernoulli representation for the integrand it is possible to get the following expression to the analytic extension of $\zeta(z,\frac{1}{2})$:

$$\zeta\left(z,\,\frac{1}{2}\right)=\frac{1}{\Gamma(z)}\,g_1\left(z,\,\frac{1}{2}\right)+\frac{1}{\Gamma(z)}\,\sum_{n=0}^{\infty}\,\frac{B_n(\frac{1}{2})}{n!}\,\frac{1}{(z+n-1)}, \tag{15}$$

where $g_1(z,\frac{1}{2})$ is given by

$$g_1\left(z,\,\frac{1}{2}\right)=\int_1^{\infty}dt\;t^{z-1}\,\frac{e^{t/2}}{e^t-1}, \tag{16}$$

and $B_n(x)$ are the Bernoulli polynomials. Taking $z=2s-d$ and defining $c_n=(B_n/n!)((1/2^{n-1})-1)$ we replace this analytic extension in the effective potential $V(\varphi_0,\beta)$. A straightforward calculation gives

$$V(\varphi_0,\beta)=\frac{1}{2}\,\varphi_0^2+2N\sum_{s=1}^{\infty}\,q(D,s)\left(g_1\left(2s-d,\,\frac{1}{2}\right)+\sum_{n=0}^{\infty}\,\frac{c_n}{2s-D+n}\right)(g\varphi_0)^{2s}\beta^{2s-D}, \tag{17}$$

where $q(D,s)$ is given by

$$q(D,s)=\frac{(-1)^s}{\Gamma(s+1)}\,\frac{1}{\Gamma(s-D/2+1)}\,\frac{1}{(4\pi)^{2s}-(D/2)}. \tag{18}$$

A direct consequence of the regularization method employed is the following: we see from the above equation that $V(\varphi_0,\beta)$ has no singularities for odd D and integer s. This means that for odd D at the $1/N$ (large N) approximation the regularized effective potential is finite. No renormalization procedure is needed. It is interesting to note that this fact is analogous from a mathematical point of view to similar situations in the Casimir effect and for anomalies in unbounded odd dimensional space–times.[12] For even D, the regularized effective potential has poles, as we

expected. On more physical grounds we remember that to exhibit the singularity structure of the effective potential at finite temperature is important for several reasons, including dimensional reduction. As explicitly pointed out by Landsman,[13] the possibility of dimensionally reducing a thermal field theory depends on whether we are able to find thermal counterterms that cancel the divergent contributions as $\beta\to 0$ comes from the integration of nonzero Matsubara frequencies. These temperature-dependent counterterms should be controlled by a thermal renormalization group and fixed by the imposition of renormalization conditions on Feynman diagrams computed with the aid of some regularization scheme. In spite of regularization being purely an intermediary tool in the process of renormalizing a quantum field theory, we know that dimensional reduction is scheme dependent,[14,15] being optimal in BPHZ (Bogoliubov–Panasiuk–Hepp–Zimmermann) subtractions at zero momenta and temperature β^{-1}. By analytical continuation of the Hurwitz zeta function we have the exact form of the polar terms that appear in the expansion of the effective potential in the framework of dimensional regularization. These will command the counterterms needed for renormalization. It would be interesting to see how dimensional reduction works in this case, as compared to the behavior obtained from other regularization methods. For a recent discussion of these subjects see, for example, Ref. 16.

Now we perform the same analysis in the case of nonzero chemical potential. Thus, let us introduce a chemical potential μ and define

$$\sigma=\frac{\mu\beta}{2\pi}. \tag{19}$$

It is well known that the effect of the chemical potential is to shift the energy levels by μ. Then following the same steps as before it is possible to show that the only change in Eq. (11) is that instead of $\zeta(z,\frac{1}{2})$, we have to analytically extend $\zeta(z,\frac{1}{2}+i\sigma)$. Let us first write an integral representation of the Hurwitz zeta function as follows:

$$\zeta\left(z,\frac{1}{2}+i\sigma\right)=\frac{1}{\Gamma(z)}\int_0^\infty dt\ t^{z-1}\frac{e^{-t(1/2+i\sigma)}}{1-e^{-t}}. \tag{20}$$

Proceeding as we have done to get Eq. (15) above, the integrand of Eq. (20) may be expressed in terms of the Bernoulli polynomials. Using the following expansion,

$$B_n(x+h)=\sum_{k=0}^{n} C_n^k B_k(x)h^{n-k}, \tag{21}$$

we have

$$B_n\left(\frac{1}{2}-i\sigma\right)=\sum_{k=0}^{n} A_{nk}\sigma^{n-k} \tag{22}$$

where

$$A_{nk}=-(-i)^{n-k}C_n^k(1-2^{1-k})B_k, \tag{23}$$

and B_k are the Bernoulli numbers.[17] Substituting Eqs. (21)–(23) in Eq. (20) we have for the analytic extension

$$\left(2s-d,\frac{1}{2}+i\sigma\right)=\frac{1}{\Gamma(2s-d)}g_2\left(2s-d,\frac{1}{2}-i\sigma\right)+\frac{1}{\Gamma(2s-d)}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\frac{A_{nk}}{(2s-D+n)}\sigma^{n-k} \tag{24}$$

where $g_2(2s-d,\frac{1}{2}+i\sigma)$ is given by

$$g_2\left(2s-d,\frac{1}{2}+i\sigma\right)=\int_1^\infty dt\ t^{2s-D}\frac{e^{t(1/2-i\sigma)}}{e^t-1}. \tag{25}$$

The effective potential is then expressed as

$$V(\varphi_0,\beta)=\frac{1}{2}\,\varphi_0^2+2N\sum_{s=1}^{\infty}p(D,s)\left(\zeta\left(2s-d,\,\frac{1}{2}+i\sigma\right)\right)+\left(\zeta\left(2s-d,\,\frac{1}{2}-i\sigma\right)\right)(g\varphi_0)^{2s}\beta^{2s-D}, \tag{26}$$

where $p(D,s)$ is given by Eq. (12) and the Hurwitz zeta function is given by Eq. (24). Some comments are in order. It is known that a dimensional regularization method fails for massless fields. Note we are using this method only in a $D-1$ dimensional integral, and the Matsubara frequencies play the role of masses. Consequently we must restrict ourselves to nonzero temperatures. Also since Eq. (26) is a Laurent series in β^{-1} our expansion is suitable for high temperatures. It may be seen from Eq. (26) that since we have the sum of two Hurwitz zeta functions (whose arguments have chemical potentials with opposite signs) the regularized effective potential remains real in the case of nonzero chemical potential. This means that at least at the one-loop approximation there are no indications for a first-order phase transition.

III. CONCLUSION

We obtained in this paper that the $1/N$ (large N) approximation effective potential is finite in any odd dimensional space–time, in particular in $D=3$. There are no ultraviolet divergences. In the Yukawa model the same mathematical phenomenon has been shown.[10] This situation is very similar to the one encountered in the calculation of the renormalized vacuum energy of scalar fields confined in boxes (Casimir energy).[18] Dolan and Nash used the zeta function analytic regularization method to obtain the Casimir energy of conformally coupled scalar field confined in odd and even dimensional spheres.[19] They obtained that for odd dimensional spheres (even space–time dimension) there is a pole in the point of interest, being necessary for the introduction of a counterterm, while for even dimensional spheres (odd dimensional space–time) the result obtained is naturally finite. No renormalization is needed. The question that we have to answer is the link between the one-loop effective energy and the Casimir energy. This problem has been studied by some authors where it was proved that the effective energy and the Casimir energy differ, but this is an inherent renormalization-scheme ambiguity.[20] We conjecture that the absence of singularities for the thermal regularized effective potential in odd-dimensional space–time is an artifact of our approximation. In any case we hope that this note could help further investigations on the subject.

ACKNOWLEDGMENT

This paper was supported by Conselho Nacional de Desenvolvimento Cientifico e Tecnologico do Brazil (CNPq).

[1] V. Rivasseau, *From Perturbative to Constructive Renormalization* (Princeton University Press, Princeton, NJ, 1991).
[2] G. 't Hooft, Commun. Math. Phys. **88**, 1 (1983); V. Rivasseau, *ibid.* **95**, 437 (1984).
[3] J. Feldman, J. Magnen, V. Rivesseau, and R. Seneor, Commun. Math. Phys. **99**, 437 (1987); K. Gawedzki and A. Kupianen, *ibid.* **99**, 197 (1985).
[4] D. Gross and A. Neveu, Phys. Rev. D **10**, 3235 (1974); L. Jacobs, *ibid.* **10**, 3956 (1974); B. J. Harrington, *ibid.* **11**, 779 (1975); R. F. Dashen, S. K. Ma, and R. Rajaraman, *ibid.* **11**, 1499 (1975).
[5] C. de Calan, P. A. Faria da Veiga, J. Magnen, and R. Sénéor, Phys. Rev. Lett. **66**, 3233 (1991).
[6] A. B. Zamolodchikov, Phys. Lett. **72B**, 481 (1978); W. Wentzel, *ibid.* **153B**, 297 (1985); M. G. Mitchand, A. C. Davies, and A. J. Macfarlane, Nucl. Phys. B **325**, 470 (1989); B. Rosenstein and B. J. Warr, *ibid.* **335**, 288 (1990); J. Zinn-Justin, *ibid.* **367**, 105 (1991); J. A. Gracey, *ibid.* **367**, 657 (1991); W. Chen, Y. Makenko, and G. W. Semenoff, Ann. Phys. (N.Y.) **228**, 341 (1993); W. Chen, Nucl. Phys. B **4335**, 673 (1995).
[7] U. Wolf, Phys. Lett. **157B**, 303 (1985); T. Treml, Phys. Rev. D **39**, 679 (1989).
[8] Y. Cohen, S. Elitzur, and E. Rabinovici, Nucl. Phys. B **220**, 102 (1983); F. Karsch, J. Kogut, and H. W. Wyld, *ibid.* **280**, 289 (1987).
[9] A. Barducci, R. Casalbouni, M. Modugna, G. Pettini, and R. Gatte, Phys. Rev. D **51**, 3042 (1995).
[10] A. P. C. Malbouisson, B. F. Svaiter, and N. F. Svaiter, J. Math. Phys. **38**, 2210 (1997).
[11] J. Daicic and N. E. Frankel, Phys. Rev. D **53**, 5745 (1996).
[12] A. P. C. Malbouisson, Nuovo Cimento B **108**, 987 (1993); B. F. Svaiter and N. F. Svaiter, J. Math. Phys. **35**, 1840 (1994).
[13] N. P. Landsman, Nucl. Phys. B **322**, 498 (1989).
[14] T. Appelquist and J. Carazone, Phys. Rev. D **11**, 2856 (1975).
[15] J. Ambjorn, Commun. Math. Phys. **67**, 109 (1979).
[16] A. P. C. Malbouisson, M. B. Silva-Neto, and N. F. Svaiter, Physica A **250**, 423 (1998).

[17] *Handbook of Mathematical Fuctions*, edited by M. Abramowitz and A. Stegun (Dover, New York, 1965).
[18] B. F. Svaiter and N. F. Svaiter, J. Math. Phys. **32**, 175 (1991); N. F. Svaiter and B. F. Svaiter, J. Phys. A **25**, 979 (1992).
[19] B. P. Dolan and C. Nash, Commun. Math. Phys. **148**, 139 (1992).
[20] B. Allen, Phys. Rev. D **33**, 3640 (1986); S. K. Blau and M. Viser, Nucl. Phys. B **310**, 103 (1989).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 7 JULY 1998

The doublet representation of non-Hilbert eigenstates of the Hamiltonian. II

M. Castagnino[a)]
Instituto de Astronomía y Física del Espacio, C.C. 67 Suc. 28, 1428 Bs. As., Argentina and Departamento de Física, FCEyN, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, 1428 Bs. As., Argentina

G. Domenech,[a)] M. Levinas,[a),b)] and N. Umérez[a)]
Instituto de Astronomía y Física del Espacio, C.C. 67 Suc. 28, 1428 Bs. As., Argentina

(Received 21 January 1997; accepted for publication 25 February 1998)

Applying the doublet representation we analyze the solutions of a Hamiltonian system which has eigenstates with complex eigenvalues. The example of the Friedrichs model allows us to show how the appearance of solutions with non-Hilbert initial conditions is linked to the energy degeneration of the Hamiltonian spectrum. We discuss the difficulties of giving a physical meaning to the growing or decaying non-Hilbert solutions. We also suggest a way to circumvent the problem of the anomalous probabilities related to both complex energy eigenvalues and degeneration of the spectrum. © *1998 American Institute of Physics.* [S0022-2488(98)02307-X]

I. INTRODUCTION

The doublet representation[1] offers a useful way to study the dynamical evolution of both Hilbert and non-Hilbert solutions of a Hamiltonian system. In the scheme of this formalism any physical system is represented, not only by its wave functions $\varphi(\omega)$ but also by a partner $\varphi^{\star}(\omega)$. These two functions allow us to construct an invariant scalar that, in case of Hilbert states, becomes the probability of the appearance of the corresponding state. For non-Hilbert solutions, the product of the wave function and its partner remains an invariant scalar, but its probabilistic interpretation is under discussion. Precisely in this work we apply the doublet representation to show the difficulties of giving a physical interpretation to non-Hilbert solutions and we analyze the relation between energy degeneration and the appearance of complex eigenvalues. The explicit computations performed in the frame of the Friedrichs model allow us to study the conditions that the interaction Hamiltonian must satisfy to keep the spectrum in the real domain. In the other case we show how to circumvent the problem of badly behaved probabilities by means of defining a ''reduced'' space that solves the degeneration problem.

The paper is organized as follows: in Sec. II we make a brief review of Ref. 1. In Sec. III we analyze which initial conditions correspond to states in Hilbert space $\mathscr{H}$ and which ones correspond to non-Hilbert states. Section IV is devoted to the study of the anomalous behavior of probabilities and mean values for non-Hilbert states. The conditions that the interaction must satisfy in order to obtain a real spectrum are studied in Sec. V. In Sec. VI we find a basis of the reduced space where the energy degeneration is removed. Finally, in Sec. VII we draw our main conclusions.

II. THE DOUBLET REPRESENTATION

Hamiltonian equations have, in general, well-defined continuous solutions $\varphi(\omega)$ that may or may not belong to $\mathscr{H}$. But in both cases it is possible to define a ''partner'' of the wave function, namely $\varphi^{\star}(\omega)$, that is not necessarily the complex conjugate of $\varphi(\omega)$.[1] Then, solutions of the Hamiltonian equations will be represented by the ''doublet'' $(\varphi,\varphi^{\star})$. In the particular case where

a)Fellow of CONICET, Argentina.
b)Electronic mail: levinas@iafe.uba.ar

 © 1998 American Institute of Physics

$\varphi \in \mathscr{H}$, solutions satisfy $\varphi^\star = \varphi^*$, where φ^* is the complex conjugate of φ. In Ref. 1 we have applied this formalism to study the Friedrichs Hamiltonian and we have obtained that the field equations in the doublet representation read as

$$\omega_0 \varphi_1 + \lambda \int_0^\infty g(\omega)\varphi(\omega)d\omega = i\dot{\varphi}_1, \tag{1}$$

$$\omega\varphi(\omega) + \lambda g(\omega)\varphi_1 = i\dot{\varphi}(\omega), \tag{2}$$

$$\omega_0 \varphi_1^\star + \lambda \int_0^\infty g(\omega)\varphi^\star(\omega)d\omega = -i\dot{\varphi}_1^\star, \tag{3}$$

$$\omega\varphi^\star(\omega) + \lambda g(\omega)\varphi_1^\star = -i\dot{\varphi}^\star(\omega), \tag{4}$$

where $\varphi_1, \varphi_1^\star, \varphi(\omega), \varphi^\star(\omega)$ are the wave functions of the doublet representation (the first two for the discrete mode), $\lambda \in \Re$ is the coupling constant, $\omega_0 \in \Re^+$ is the discrete eigenvalue of the free Hamiltonian, $0 \leqslant \omega < \infty$ is the continuous spectrum, and $g(\omega) = g^*(\omega)$ stands for the interaction function.

We demand that wave functions satisfy the following natural conditions:

$$\varphi_1\varphi_1^\star + \int_0^\infty \varphi(\omega)\varphi^\star(\omega)d\omega < \infty, \tag{5}$$

$$(\varphi_1^\star)^\star = \varphi_1, \quad [\varphi^\star(\omega)]^\star = \varphi(\omega). \tag{6}$$

With these conditions, the solution to system (1)–(4) is

$$\varphi_1(t) = \frac{1}{\sqrt{\alpha'(z_0)}}\,\tilde{\varphi}_1 e^{-iz_0 t} + \int_0^\infty \frac{\lambda g(\tilde{\omega})}{\alpha(\tilde{\omega})}\,\tilde{\varphi}(\tilde{\omega})e^{-i\omega t}d\tilde{\omega}, \tag{7}$$

$$\varphi_1^\star(t) = \frac{1}{\sqrt{\alpha'(z_0)}}\,\tilde{\varphi}_1^\star e^{+iz_0 t} + \int_0^\infty \frac{\lambda g(\tilde{\omega})}{\alpha(\tilde{\omega})}\,\tilde{\varphi}^\star(\tilde{\omega})e^{+i\omega t}d\tilde{\omega}, \tag{8}$$

$$\varphi(\omega,t) = \frac{1}{\sqrt{\alpha'(z_0)}}\,\frac{\lambda g(\omega)}{z_0-\omega}\,\tilde{\varphi}_1 e^{-iz_0 t} + \tilde{\varphi}(\omega)e^{-i\omega t} + \int_0^\infty \frac{\lambda^2 g(\omega)g(\tilde{\omega})}{(\tilde{\omega}-\omega)\alpha(\tilde{\omega})}\,\tilde{\varphi}(\tilde{\omega})e^{-i\tilde{\omega}t}d\tilde{\omega}, \tag{9}$$

$$\varphi^\star(\omega,t) = \frac{1}{\sqrt{\alpha'(z_0)}}\,\frac{\lambda g(\omega)}{z_0-\omega}\,\tilde{\varphi}_1^\star e^{+iz_0 t} + \tilde{\varphi}^\star(\omega)e^{+i\omega t} + \int_0^\infty \frac{\lambda^2 g(\omega)g(\tilde{\omega})}{(\tilde{\omega}-\omega)\alpha(\tilde{\omega})}\,\tilde{\varphi}^\star(\tilde{\omega})e^{+i\tilde{\omega}t}d\tilde{\omega}, \tag{10}$$

where

$$z_0(\lambda) = \omega_0 + \lambda^2 \int_0^\infty \frac{g^2(\omega)}{(z_0(\lambda)-\omega)}\,d\omega, \tag{11}$$

$$\alpha(z) = \omega_0 - z - \lambda^2 \int_0^\infty \frac{g^2(\omega)}{z-\omega}\,d\omega, \tag{12}$$

$\alpha' = d\alpha/dz$, and $(\tilde{\varphi}_1, \tilde{\varphi}(\tilde{\omega}))$, $(\tilde{\varphi}_1^\star, \tilde{\varphi}^\star(\tilde{\omega}))$ are the eigenfunctions of the free plus the interaction Hamiltonians. The singularities in $\tilde{\alpha}(\tilde{\omega})$ and $(\tilde{\omega}-\omega)^{-1}$ must be avoided making the shift $\pm i\epsilon$. We do not write it explicitly in order not to complicate the notation. A main characteristic of the solutions (7)–(10) is that they behave in a continuous way when the coupling constant λ goes to zero.

When $(\varphi_1, \varphi(\omega))$ belongs to $\mathscr{H}$, we have

$$\varphi_1^\star = \varphi_1^* \quad \text{and} \quad \varphi^\star(\omega) = \varphi^*(\omega).$$

In this particular case, Eq. (5) implies that

$$0 \leqslant \varphi_1 \varphi_1^\star \leqslant 1,$$

$$0 \leqslant \int_{\omega_1}^{\omega_2} \varphi(\omega)\varphi^*(\omega)d\omega \leqslant 1.$$

If $[\omega_1, \omega_2]$ is any nonempty interval, the last integral corresponds to a well-defined probability.

III. NON-HILBERT INITIAL CONDITIONS

From Eqs. (7)–(10) we can see that $\varphi^\star$ equals φ^* when $z_0 \in \Re$. In this case, the solutions belong to $\mathscr{H}$. But as we can obtain Hilbert solutions, even for those Hamiltonians that have complex eigenvalues, we are now interested in determining, for nonreal z_0, which initial conditions correspond to those states.

Initial conditions in $\mathscr{H}$ imply that $\varphi_1^\star(t=0) = \varphi_1^*(t=0)$ and $\varphi^\star(\omega, t=0) = \varphi^*(\omega, t=0)$. When analyzing the solutions (7)–(10) to the equations of motion, we see that this relation of conjugation holds true for any time value if the first terms of the rhs in (7)–(10) cancel with the corresponding residue evaluated in z_0. It is easy to see that these four conditions reduce to the following two:

$$2\pi i \ \mathrm{Res}\left[\frac{\lambda g(\omega)}{\alpha(\omega)} \ \tilde{\varphi}(\omega), z_0\right] = -\frac{1}{\sqrt{\alpha'(z_0)}} \ \tilde{\varphi}_1, \tag{13}$$

$$2\pi i \ \mathrm{Res}\left[\frac{\lambda g(\omega)}{\alpha(\omega)} \ \tilde{\varphi}^\star(\omega), z_0\right] = -\frac{1}{\sqrt{\alpha'(z_0)}} \ \tilde{\varphi}_1^\star. \tag{14}$$

An illustrative example with initial conditions in $\mathscr{H}$ that satisfy (13) and (14) is the general state whose initial conditions coincide with the discrete eigenstate of the free Hamiltonian H_0:

$$\varphi_1(t=0) = \varphi_1^\star(t=0) = 1, \quad \varphi(\omega, t=0) = \varphi^\star(\omega, t=0) = 0,$$

and evolve with the complete Hamiltonian H. Indeed, performing the change of basis (7)–(10), we obtain that

$$\tilde{\varphi}(\omega) = \frac{\lambda g(\omega)}{\alpha(\omega)},$$

and replacing it in Eq. (13), we see that

$$\begin{aligned} 2\pi i \ \mathrm{Res}\left[\frac{\lambda g(\omega)}{\alpha(\omega)} \ \tilde{\varphi}(\omega), z_0\right] &= 2\pi i \ \mathrm{Res}\left[\frac{\lambda^2 g^2(\tilde{z})}{\alpha^2(\tilde{z})}, \tilde{z} = z_0\right] \\ &= 2\pi i \ \mathrm{Res}\left[\frac{-i}{2\pi} \frac{1}{\alpha(\tilde{z})}, \tilde{z} = z_0\right] \\ &= -\frac{1}{\alpha'(z_0)} = -\frac{1}{\sqrt{\alpha'(z_0)}} \ \tilde{\varphi}_1, \end{aligned} \tag{15}$$

which is precisely that condition. This is also true for condition (14). Here we have used

$$\alpha_+(\omega) - \alpha_-(\omega) = -2\pi i \lambda^2 g^2(\omega).$$

On the other hand, one state whose initial values satisfy neither condition (13) nor condition (14) is, for example, the discrete eigenstate of the Hamiltonian,

$$\tilde{\varphi}_1 = \tilde{\varphi}_1^\star = 1 \quad \text{and} \quad \tilde{\varphi}(\omega) = \tilde{\varphi}^\star(\omega) = 0,$$

that it is obviously out of $\mathscr{H}$.

Finally, we note that there are initial conditions which satisfy either (13) or (14) but not both, and other ones which satisfy neither, but are not eigenstates of H. Nevertheless, all of them obey the ''normalization'' condition (5).

IV. PROBABILITY BEHAVIOR OUT OF HILBERT

To characterize the physical behavior of non-$\mathscr{H}$ states, we should not consider the temporal evolution of the wave functions only, since each one alone does not provide information either about transition probabilities or about the mean values. For example, the discrete eigenstate $(\tilde{\varphi}_1, \tilde{\varphi}_1^\star)$ whose components evolve with $e^{-iz_0 t}$ and $e^{iz_0 t}$, respectively, must not be interpreted as two decaying and growing states, because, in spite of the fact that they seem to be so, they are one single stationary state satisfying

$$\tilde{\varphi}_1(t)\tilde{\varphi}_1^\star(t) = \tilde{\varphi}_1(0)\tilde{\varphi}_1^\star(0) = 1,$$

$$\tilde{\varphi}(\tilde{\omega},t)\tilde{\varphi}^\star(\tilde{\omega},t) = \tilde{\varphi}(\tilde{\omega},0)\tilde{\varphi}^\star(\tilde{\omega},0) = 0,$$

as it may be seen from Eqs. (7)–(10).

Nevertheless, when evaluating probabilities for a general non-Hilbert state, including stationary ones, we see that—even though total ''probability'' is normalized to unity [see Eq. (5)]—the partial probability of finding the state in a bounded interval of energy may be out of the interval [0,1] or be a complex number. To investigate the nature of this problem let us consider an example. From Eqs. (7)–(10), the ''probability'' of being in the discrete eigenstate of H is

$$\varphi_1 \varphi_1^\star = \frac{1}{\alpha'}, \tag{16}$$

and the ''probability density'' of finding the system with continuous energy ω is

$$\varphi(\omega)\varphi^\star(\omega) = \frac{1}{\alpha'} \frac{\lambda^2 g^2(\omega)}{(z_0 - \omega)^2}. \tag{17}$$

Expanding the probability of finding the system in the discrete state up to the second order in the interaction constant λ, we find

$$\varphi_1 \varphi_1^\star = 1 - \lambda^2 \int_0^\infty \frac{g^2(\omega)}{(\omega_0 - \omega)^2}\, d\omega, \tag{18}$$

whereas for the probability density of finding it with a continuous energy ω, we obtain

$$\varphi(\omega)\varphi^\star(\omega) = \lambda^2 \frac{g^2(\omega)}{(\omega_0 - \omega)^2}. \tag{19}$$

So in this approximation we have

$$\varphi_1 \varphi_1^\star + \int_0^\infty \varphi(\omega)\varphi^\star(\omega) d\omega = 1.$$

We want to point out some features from the previous expressions

(1) Probability densities obey the condition expressed in (5).
(2) The probability density (19) has a pole in $\omega = \omega_0$.
(3) This pole produces complex ''probabilities'' in general and the ''probability density'' may be complex or greater than unity in its neighborhood.

Statements (1), (2), and (3) show us the nature of the problem and suggest a way to circumvent it. In fact, when we find ω_0 as the result of a measurement of the energy of the system, we do not know whether the corresponding state is the discrete eigenvector, the continuous vector with the same eigenvalue, or a linear combination of both. Then, it seems that we are not allowed to treat them as separated states assigning different probabilities to them because we cannot distinguish these states through energy measurements. But if we compute the probability to obtain ω_0 as the probability of the system to be in the discrete state plus the probability of the system to be in the continuum in a neighborhood of ω_0, the problem gets solved by itself because the infinite terms involving ω_0 cancel each other in the λ^2 approximation, i.e.,

$$\begin{aligned}\mathscr{P}[\omega_0-\Delta,\omega_0+\Delta] &= \varphi_1\varphi_1^\star + \int_{\omega_0-\Delta}^{\omega_0+\Delta} \varphi(\omega_0)\varphi^\star(\omega_0)\,d\omega \\ &= 1-\lambda^2\int_0^{\omega_0-\Delta} \frac{g^2(\omega)}{(\omega_0-\omega)^2}\,d\omega - \lambda^2\int_{\omega_0+\Delta}^{\infty} \frac{g^2(\omega)}{(\omega_0-\omega)^2}\,d\omega,\end{aligned}$$

does not have the problems mentioned in the above statements (2) and (3).

Nevertheless, this is only a second-order solution. It can be seen after a straightforward computation that in the next step of the expansion (fourth-order) complex contributions reappear from each term of the probability densities and do not cancel each other. This is so because, as we will see, this is not the right way to solve the problem.

Due to the bad behavior of probabilities out of $\mathscr{H}$, we have studied the behavior of another meaningful magnitude: the mean value. In the frame of the doublet representation we can generalize the expression of the free Hamiltonian mean value in any state as

$$\begin{aligned}\langle H_0(t)\rangle &= \omega_0\varphi_1(t)\varphi_1^\star(t) + \int_0^\infty \omega\varphi(\omega,t)\varphi^\star(\omega,t)\,d\omega \\ &= \omega_0\Bigg[\frac{1}{\sqrt{\alpha'(z_0)}}\,\tilde{\varphi}_1 e^{-iz_0t} + \int_0^\infty \frac{\lambda g(\tilde{\omega})}{\alpha_+(\tilde{\omega})}\,\tilde{\varphi}(\tilde{\omega})e^{-i\tilde{\omega}t}d\tilde{\omega}\Bigg] \\ &\quad\cdot\Bigg[\frac{1}{\sqrt{\alpha'(z_0)}}\,\tilde{\varphi}_1^\star e^{+iz_0t} + \int_0^\infty \frac{\lambda g(\tilde{\omega})}{\alpha_-(\tilde{\omega})}\,\tilde{\varphi}^\star(\tilde{\omega})e^{+i\tilde{\omega}t}d\tilde{\omega}\Bigg] \\ &\quad+\int_0^\infty\Bigg[\frac{1}{\sqrt{\alpha'(z_0)}}\,\frac{\lambda g(\omega)}{z_0-\omega}\,\tilde{\varphi}_1 e^{-iz_0t} + \lambda^2 g(\omega) \\ &\quad\times\int_0^\infty \frac{g(\tilde{\omega})}{\alpha_+(\tilde{\omega})(\tilde{\omega}-\omega)}\,\tilde{\varphi}(\tilde{\omega})e^{-i\tilde{\omega}t}d\tilde{\omega} + \tilde{\varphi}(\omega)e^{-i\omega t}\Bigg] \\ &\quad\times\Bigg[\frac{1}{\sqrt{\alpha'(z_0)}}\,\frac{\lambda g(\omega)}{z_0-\omega}\,\tilde{\varphi}_1^\star e^{+iz_0t} + \lambda^2 g(\omega) \\ &\quad\times\int_0^\infty \frac{g(\tilde{\omega})}{\alpha_-(\tilde{\omega})(\tilde{\omega}-\omega)}\,\tilde{\varphi}^\star(\tilde{\omega})e^{+i\tilde{\omega}t}d\tilde{\omega} + \tilde{\varphi}^\star(\omega)e^{+i\omega t}\Bigg]\omega\,d\omega,\end{aligned}$$

whose instabilities come from the real contributions of exponentials $e^{\pm iz_0t}$. That is to say, the mean value has oscillating terms (which correspond to real frequencies ω) and growing or decaying waves coming from the imaginary contribution of the ''frequency'' z_0. These instabilities do not cause problems to Hilbert states because they satisfy conditions (13) and (14), so the complex exponentials cancel each other and do not produce vanishing or indefinitely growing terms. On the other hand, non-$\mathscr{H}$ states will have anomalous terms in the mean values of the Hamiltonian which diverge for growing time values, except when they are eigenstates of H. So here we also have a problem for the mean values evaluated in non-$\mathscr{H}$ states.

In conclusion, we point out that, in spite of the fact that it is possible to find the dynamical evolution for any doublet, it is difficult to assign probabilities or mean values (i.e., to give physical meaning) to them. In the next section we show that this problem is linked to the fact that the probability of measuring energy ω_0 has contributions coming from the discrete state plus contributions from any interval around ω_0, belonging to the continuum.

V. INTERACTIONS THAT GENERATE REAL z_0

In the previous sections we have identified the appearance of anomalous probabilities with z_0 being a nonreal number. Now we will show that the problems created by this complex quantity arise from the fact that the improper integrals around ω_0 yield a complex eigenvalue of the complete Hamiltonian. To show the explicit limit between complex and real eigenvalues, we make use of the solutions (7)–(10) which are continuous in λ (see Ref. 1). Thus, we can suppose that z_0 in expression (11) can be expanded as a power series of λ, taking $z_0|_{\lambda=0}=\omega_0$,

$$z_0=\omega_0+\lambda^2\int_0^\infty \frac{g^2(\omega)}{(\omega_0-\omega)}\,d\omega-\frac{2}{3}\lambda^4\int_0^\infty \frac{g^2(\omega)}{(\omega_0-\omega)}\,d\omega\int_0^\infty \frac{g^2(\omega')}{(\omega_0-\omega')^2}\,d\omega'+\cdots . \tag{20}$$

As we can see, improper integrals appear in the expansion of z_0. A generic coefficient in the series always contains a factor like

$$\frac{g^2(\omega)}{(\omega_0-\omega)^n},$$

implying the appearance of poles in the function under integration which—when computed in the complex plane—give imaginary contributions to z_0. These divergences point out the influence of the degeneration in the free Hamiltonian spectrum: when the eigenstate that corresponds to the discrete eigenvalue of the free Hamiltonian interacts with the eigenstate of the continuum with energy ω_0 and its neighborhood, complex eigenvalues of the complete Hamiltonian appear. As it has been shown in Ref. 2, the reality of the discrete eigenvalue z_0 is related to $g(\omega)$ evaluated in the eigenvalue.

Expression (20) tells us that, if we confine ourselves to the second-order expansion, z_0 is real when $g(\omega_0)=0$. This is in accordance with well-known results that establish that

$$g(\omega_0)=0\Rightarrow\Gamma^{(2)}=\pi\lambda^2 g^2(\omega_0)=0,$$

where $\Gamma^{(2)}$ is the second order in λ contribution to half the inverse of time life.[3–5] But this is not true for higher orders in the λ expansion because the condition $g^2(\omega_0)=0$ is not sufficient to guarantee that the whole integral be nondivergent.

From the previous results, we conclude that the condition that $g(\omega)$ must satisfy in order to obtain a real spectrum is

$$\lim_{\lambda\to 0}\frac{g^2(\omega)}{(z_0(\lambda)-\omega)^n}<\infty,\quad \forall n\in\mathcal{N}, \tag{21}$$

for any ω belonging to an interval that contains ω_0 because, when performing the expansion (20), condition (21) guarantees the reality of the expansion. This is equivalent to requesting that

$$\lim_{\lambda\to 0} z(\lambda)=\omega_0$$

does not belong to the support of $g(\omega)$.

A trivial example of interaction satisfying (21) is the Ohmic interaction $g(\omega)=\omega$ with a cutoff that leaves ω_0 out of the integration interval. In fact, any interaction that vanishes over a finite interval around ω_0 is a trivial example.

A nontrivial interaction satisfying (21) is

$$g(\omega)=\begin{cases}\exp(\omega_0-\omega)^{-1}, & \text{if } \omega_0<\omega;\\ 0, & \text{if } \omega_0=\omega;\\ \exp-(\omega_0-\omega)^{-1}, & \text{if } \omega_0>\omega,\end{cases} \tag{22}$$

which removes, in a continuous way, the interaction between the degenerate states of the free Hamiltonian.

For interactions satisfying condition (21), the spectrum of the complete Hamiltonian is real, no pathologies appear in the evaluation of probabilities, and the whole problem can be formulated in Hilbert space with no need of enlarging the state space to the doublet one in order to diagonalize H.

VI. A REDUCED SPACE

In general cases in which $g(\omega)$ does not satisfy condition (21) and z_0 is not necessarily a real number, we may remove the degeneration that causes anomalous probabilities defining a new basis with the following procedure (which we will develop in Dirac notation because it is clearer in this case).

To exclude the continuous eigenvalue ω_0 from the continuous spectrum (thus removing the degeneration) we define the set

$$\mathscr{C}=\{\omega/0\leqslant\omega\leqslant\omega_0-a \quad \text{or} \quad \omega_0+a\leqslant\omega\}, \tag{23}$$

with $a\in\Re$, $a>0$. Next, we define the orthogonal basis:

$$|\Lambda\rangle=\Lambda_1|1\rangle+\int_{\omega_0-a}^{\omega_0+a}\lambda(\omega)|\omega\rangle d\omega\,\epsilon\mathscr{H}, \tag{24}$$

$$\{|\omega_r\rangle\}=\{|\omega\rangle/\omega\in\mathscr{C}\}, \tag{25}$$

where the basis vectors satisfy

$$\begin{aligned}&\langle\Lambda|\Lambda\rangle=\Lambda_1\Lambda_1^*+\int_{\omega_0-a}^{\omega_0+a}\Lambda(\omega)\Lambda^*(\omega)d\omega=1,\\ &\langle\omega_r|\Lambda\rangle=\langle\Lambda|\omega_r\rangle=0 \quad \text{and} \quad \langle\omega_r|\omega_r'\rangle=\delta(\omega_r-\omega_r').\end{aligned} \tag{26}$$

In this basis the matrix elements of H are

$$\begin{aligned}\langle\Lambda|H|\Lambda\rangle&=\omega_0\Lambda_1\Lambda_1^*+\lambda\int_{\omega_0-a}^{\omega_0+a}g(\omega)[\Lambda_1^*\Lambda(\omega)+\Lambda_1\Lambda^*(\omega)]d\omega\\ &+\int_{\omega_0-a}^{\omega_0+a}\omega\Lambda(\omega)\Lambda^*(\omega)d\omega\equiv\Lambda(\lambda),\end{aligned} \tag{27}$$

$$\langle\omega_r'|H|\omega_r\rangle=\omega_r\delta(\omega_r'-\omega_r), \tag{28}$$

$$\langle\omega_r|H|\Lambda\rangle=\lambda\Lambda_1 g(\omega_r). \tag{29}$$

With them, we can define the Hamiltonian $\tilde{H}$ which acts on the ''reduced'' space as

$$\tilde{H}=\Lambda(\lambda)|\Lambda\rangle\langle\Lambda|+\int_{\mathscr{C}}\omega|\omega\rangle\langle\omega|d\omega+\int_{\mathscr{C}}\lambda g(\omega)[\Lambda_1|\Lambda\rangle\langle\omega|+\Lambda_1^*|\omega\rangle\langle\Lambda|]d\omega. \tag{30}$$

Diagonalizing $\tilde{H}$, we obtain that its discrete eigenvalue satisfies the following recursive equation:

$$\tilde{\Lambda}(\lambda)=\Lambda(\lambda)+\lambda^2\int_{\mathscr{C}}\frac{\Lambda_1\Lambda_1^* g^2(\omega)}{(\tilde{\Lambda}(\lambda)-\omega)}\,d\omega. \tag{31}$$

Taking into account the results of Sec. V, we have that $\tilde{\Lambda}(\lambda)$ will be real (and the probability well defined) if $\lim_{\lambda\to 0}\tilde{\Lambda}(\lambda)$ does not belong to the support of $g(\omega)$, i.e., if

$$\lim_{\lambda\to 0}\tilde{\Lambda}(\lambda)=\lim_{\lambda\to 0}\Lambda(\lambda)=\omega_0+\int_{\omega_0-a}^{\omega_0+a}(\omega-\omega_0)\Lambda(\omega)\Lambda^*(\omega)d\omega \tag{32}$$

does not belong to the continuous set $\{\omega_r\}$ [we have used the normalization condition (26)]. As the function under integration in (32) is small enough near ω_0, it will be always possible to define some real number a near ω_0 to satisfy

$$\omega_0-a\leqslant\Lambda(\lambda=0)\leqslant\omega_0+a.$$

Then, we conclude that H has a real spectrum, the wave functions belong to $\mathscr{H}$, and the probabilities are well defined in the reduced space spanned by the basis (24)–(25).

We also notice that the reality of $\tilde{\Lambda}(\lambda)$ in (31) is guaranteed from a sufficiently small value of (ω_0-a) and it remains in the $a\to 0$ limit. In this case

$$\lim_{a\to 0}\tilde{\Lambda}\equiv W_0\in\Re,$$

and the spectrum remains nondegenerated because the continuous ω_0 energy value has been removed from the continuous spectrum. So, the eigenfunction which evolves with $\exp(iz_0t)$ becomes the Hilbert eigenfunction which evolves with $\exp(iW_0t)$ defined in the reduced space.

VII. CONCLUSIONS

We have shown that the states which do not satisfy conditions (13) and (14) do not belong to $\mathscr{H}$. So, even though we can predict their temporal evolution, it is impossible to define either well-behaved probabilities or nondivergent mean values for them. Non-Hilbert states are useful to describe the dynamical evolution of physical wave functions, but they cannot be considered as having the same physical nature as those of Hilbert space (the same as plane waves do not have the same physical nature as ordinary Hilbert states).

Taking into account that the existence of non-Hilbert states is linked to the appearance of nonreal eigenvalues and that these eigenvalues are related to the degeneration of the Hamiltonian spectrum, we have shown how to solve the problem. To do so we define a reduced basis that generates wave functions which represent states with well-defined probabilities and mean values.

ACKNOWLEDGMENTS

This work was partially supported by the following grants: Grant No. OI1*-CT94-0004 of the European Community, CONICET PID Grant No. 3183/93, and University of Buenos Aires Grant No. EX-198.

[1] M. Castagnino, G. Domenech, M. Levinas, and N. Umérez, J. Math. Phys. **37**, 2107 (1996).
[2] G. Parravicini, V. Gorini, and E. C. G. Sudarsham, J. Math. Phys. **21**, 2208 (1980).
[3] E. C. G. Sudarsham and Ch. V. Chiu, Phys. Rev. D **47**, 2602 (1993).
[4] C. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics* (Hermann, Paris, 1977).
[5] P. Exner, *Open Quantum Systems and Feynman Integrals* (Reidel, Amsterdam, 1985).

Multicomponent Wentzel–Kramers–Brillouin approximation on arbitrary symplectic manifolds: A star product approach

C. Emmrich and H. Römer
Fakultät für Physik, Universität Freiburg, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany

(Received 20 February 1997; accepted for publication 28 May 1997)

It is known that in the Wentzel–Kramers–Brillouin approximation of multicomponent systems like the Dirac equation or Born–Oppenheimer approximation, an additional phase appears apart from the Berry phase. So far, this phase was only examined in special cases, or under certain restrictive assumptions, namely, that the eigenspaces of the matrix or endomorphism valued symbol of the Hamiltonian form trivial bundles. We give a completely global derivation of this phase which does not depend on any choice of local trivializing sections. This is achieved using a star product approach to quantization. Furthermore, we give a systematic and global approach to a reduction of the problem to a problem defined completely on the different ''polarizations.'' Finally, we discuss to what extent it is actually possible to reduce the problem to a really scalar one, and make some comments on obstructions to the existence of global quasiclassical states. © *1998 American Institute of Physics.* [S0022-2488(98)03106-5]

I. INTRODUCTION

Whereas the semiclassical and Wentzel–Kramers–Brillouin (WKB) approximation of systems with Hamiltonians whose symbol is a scalar function is very well understood both on a local and a global level,[1–4] this is not the case to the same extent for systems with Hamiltonians whose symbols are matrix valued functions on a symplectic manifold, or sections in the endomorphism bundle of a vector bundle over a symplectic manifold M.

Such Hamiltonians appear in many places in physics: Examples are the Dirac equation, multicomponent wave equations like electrodynamics in media and Yang–Mills theories, and the Born–Oppenheimer approximation in molecular physics.

In the scalar case, a nice geometric interpretation of the semiclassical states as half densities on Lagrangian submanifolds invariant under the Hamiltonian flow exists, and discrete spectra may be computed using the Bohr–Sommerfeld condition, taking into account the Maslov correction.

In the multicomponent case, the analogous structures are not known: There exist many results on a local level, but on a global level results are only known either for special examples, or under the very restrictive assumption that the eigenspaces of the symbol of the Hamiltonian form trivial bundles over the symplectic base manifold M.[1,5–7].

Locally, the WKB ansatz for multicomponent systems still is

$$\Psi(x)=a(x,\hbar)\exp(iS(x)/\hbar),$$

with $a(x,\hbar)=a_0(x)+a_1(x)\hbar+\cdots$. Here, the $a_i(x)$ are now vector valued functions.[8,9,1]

In lowest order in $\hbar$ the time independent Schrödinger equations yields

$$[H_0(x,dS(x))-E]a_0(x)=0,$$

i.e., $a_0(x)$ has to be an eigenvector of the matrix valued principal symbol $H_0(x,p)$ at the point $p=dS(x)$. Denoting by $\lambda_\alpha(x,p)$ the α-th eigenvalue of the matrix $H_0(x,p)$ at the point (x,p), the phase $S(x)$ has to satisfy the Hamilton–Jacobi-equation for one of the eigenvalue functions λ_α: $\lambda_\alpha(x,dS)-E=0$, and one obtains a separate phase function for each ''polarization'' α.

 © 1998 American Institute of Physics

The equation of order $\hbar$ has a more complicated structure than in the scalar case: The absolute value of the amplitude still fulfills the usual transport equation, but the vector a_0 satisfies a more complicated equation which contains a Berry phase term,[10] and an additional term without an obvious geometric meaning.[11,7,9]

There exist some functional analytically rigorous results on the WKB approximation for special systems in the literature.[1,6] However, no analog of the Bohr–Sommerfeld quantization condition is known in general, and all those results have the common feature that they are based on the assumption that the bundles of eigenvectors of the symbol of the Hamiltonian are trivial, at least on the ''classically allowed region'' $\lambda_\alpha < \tilde{E}$ for an $\tilde{E} > E$. This assumption is not always fulfilled: Nontrivial eigenspace bundles do in fact occur in many physical situations. One immediate example is Dirac's equation for a field of spin 1/2 on a nontrivial space time manifold M, either Riemannian or pseudo-Riemannian, where the eigenbundles depend on the global structure of M (and the spin structure.) But even for the Born–Oppenheimer approximation the phenomenon of nontrivial eigenspace bundles may arise due to the phenomenon of level crossing, i.e., the possibility of two eigenvalues becoming equal in certain points on phase space: The degeneracy increases in those points. For the WKB approximation one has to cut out those points and to treat them separately. This may lead to the appearance of nontrivial bundles, and as a matter of fact, it is well known that nontrivial bundles play a role in real molecular systems.

In Ref. 7 a Bohr–Sommerfeld condition could be derived for systems whose symbol has nondegenerate eigenvalues, using the Moyal product on matrix valued symbols in order to reduce this case to the WKB approximation for a scalar system. To this end, a formally unitary $U = U_0 + U_1\hbar + \cdots$ was introduced such that

$$U^\dagger * H * U = \Lambda,$$

for some diagonal matrix valued symbol Λ. This approach has been further refined in Ref. 12.

However, this method has several disadvantages: First, considering this equation at zeroth order, one concludes that the columns of U_0 are eigenvectors of H_0. Hence, this method may again be applied globally only if the eigenvector bundles are trivial. Moreover, it is not possible to generalize this method to the case of degenerate eigenvalues with constant degeneracies. In this case, it is in general not possible to diagonalize the Hamiltonian globally (even if the eigenvector bundles are trivial), but only to block-diagonalize it. Finally, this approach is not natural from a geometric point of view: The Moyal product is defined only on trivial bundles over $\mathbb{R}^{2n}$, hence it is not possible to stick with the Moyal product if one wants to reduce the problem in a global way to a problem defined on the respective eigenbundles. In particular, it will become obvious that this method, if generalized to the degenerate case, will correspond to a quantization based on a noncanonical and unnatural choice of a connection: The choice of U_0 implies, as mentioned above, a choice of trivializing sections of the eigenbundles. In a sense to become clear in section IV, the quantization achieved by this process corresponds to a connection with respect to which the chosen set of trivializing sections is covariantly constant. Such a connection is obviously geometrically meaningless, since it depends on the choice of those sections.

Our goal in this article is to reduce the problem of WKB for endomorphism valued symbols ''as far as possible'' to the scalar case, using only globally defined methods: Apart from being an important progress in a geometric understanding of a multicomponent WKB approximation, this is an important step towards a Bohr–Sommerfeld like condition, since the latter is necessarily global in nature.

For this reduction of the problem, the notion of star products on arbitrary endomorphism bundles proves to be particularly useful, since we have to deal with nontrivial bundles. Such star products are known to exist on arbitrary endomorphism bundles over a symplectic manifold, as has been shown by Fedosov.[13] His method is particularly adapted to our purposes, since his proof of existence is constructive in nature, and there exist explicit formulas for the isomorphisms of different star products. Although we will restrict ourselves to formal star products, i.e., to formal power series in $\hbar$, the method can in principle be translated into operator language using the notion of asymptotic operator representations, which are, in a suitable sense, representations up to order $O(\hbar^\infty)$ of the star product algebra. Such representations are known to exist if certain quantization condition are satisfied.[14]

The main idea of the paper is to give a ''basis independent and geometrically natural version of the diagonalization procedure of.[7] '' What is meant by this becomes more apparent if one observes that essentially all additional complications in the WKB approximation of multicomponent systems are due to the fact that the projection operators onto the eigenspaces of the principal symbol of H are in general not constant. If one could replace ordinary derivatives by covariant derivatives with respect to which the projections are covariantly constant, the problem would be considerably simplified. However, as we assume that we are studying a given definite quantum mechanical system, we cannot simply replace ordinary derivatives by covariant derivatives ''by hand,'' as we do not want to change the physics of the system.

Here, the use of Fedosov's star product proves to be particularly powerful: His construction of a star product depends on the choice of a symplectic connection on the base space and a connection on the endomorphism bundle. Nevertheless, the different star products are isomorphic. Hence, if we interpret the choice of the connection as defining a definite quantization procedure, then we may change this connection without changing physics if we apply at the same time the corresponding isomorphism. This means that we are allowed to change the connection in order to simplify the problem, if we consistently correct the symbols of the Hamiltonian and all interesting observables at the same time.

Indeed, this additional freedom (which is completely missing if one restricts oneself to the Moyal product) may be used to simplify the problem considerably: One may find an $\hbar$-dependent connection, $\hbar$-dependent projections, and a corrected symbol such that all these structures are compatible in an obvious sense. An important feature of Fedosov's product is that the star product constructed from these data is compatible with the projections as well, i.e., if A and B are observables such that there corrected symbols commute with the projections, then the corresponding star product $A*B$ commutes with the projections as well. Hence, the different blocks defined by the projections are decoupled for a suitable set of observables.

Using these techniques, we will be lead to a geometrical interpretation of the additional phases appearing in the WKB approximation at order $\hbar$ as formed from the Poisson-curvature and an analog of the second fundamental form of a Berry-type connection. We will show that this even holds if one chooses a different connection for computing the correction. Hence, the Berry-type connection is distinguished among the set of all compatible connections.

The paper is organized as follows: In the next section we give a short overview over Fedosov's construction of star products on arbitrary endomorphism bundles to the extent needed for our applications. The main fact will be that a modification of the connection on the endomorphism bundle at some order $O(\hbar^k)$ induces an isomorphism of the corresponding star product algebras which implies corrections only at higher order $O(\hbar^{k+1})$. This will allow the use of inductive methods in the main part of the paper, and to compute the corrections iteratively order by order in $\hbar$.

In Section III, we show how one may use the freedom of choosing the connection in order to block-diagonalize the system, i.e., to find a decomposition of the vector bundle such that the Hamiltonian and a large class of observables preserve this decomposition at the quantum mechanical level. Furthermore, we will show that for a suitable class of observables, a formal Heisenberg equation of motion may be defined.

In Section IV we compute the correction of the symbol of H at order $\hbar$ for all compatible connections, thus reproducing in particular the additional phase observed in special cases before. We give a geometric interpretation of these terms, and show that a certain adapted connection which is of ''Berry type'' is distinguished among all adapted connections.

In Section V we discuss some physical aspects of the assumptions and results of the previous analysis.

Finally, in Section VI we examine to what extent it is actually possible to reduce the problem to a scalar one, and discuss some aspects connected to a Bohr–Sommerfeld like quantization condition.

II. STAR PRODUCTS ON ENDOMORPHISM BUNDLES

In this section we give a short introduction to Fedosov's beautiful construction of star products on arbitrary symplectic manifolds,[13] as needed for our purposes, and we prove three lemmas needed in Section III.

We assume the reader to be familiar with the general ideas of deformation quantization. For the deformation quantization program, we refer to Ref. 15. For an alternative, less explicit proof of existence of star products on arbitrary symplectic manifolds, see Ref. 16.

We denote the Moyal product, which is the star product on $\mathbb{R}^{2n}$ corresponding to Weyl ordering, by $\diamondsuit$:

$$(a \diamondsuit b)(x) = (e^{(i\hbar/2)\omega^{kl}(\partial/\partial x^k)(\partial/\partial y^l)} a(x)b(y))|_{x=y},$$

where ω^{kl} are the components of the Poisson tensor. The algebra $\mathcal{F}(\mathbb{R}^{2n})\otimes\mathbb{C}[[\hbar]]$ with the product $\diamondsuit$ is called the Weyl algebra over $\mathbb{R}^{2n}$. [Here, $\mathcal{F}(\mathbb{R}^{2n})$ denotes the algebra of smooth functions on $\mathbb{R}^{2n}$.]

Let M be a symplectic manifold with symplectic form ω and a symplectic connection ∂_s (such a connection always exists), and let V be a Hermitian vector bundle over M with Hermitian connection $\mathbf{\nabla}$. We consider the bundle $V[[\hbar]] = V\otimes\mathbb{C}[[\hbar]]$ which is a fiber bundle over M whose fiber over a point $q\in M$ is the $\mathbb{C}[[\hbar]]$ module $V_q[[\hbar]] = V_q\otimes\mathbb{C}[[\hbar]]$. Let $E[[\hbar]] = \cup_{q\in M} \mathrm{End}(V_q)\otimes\mathbb{C}[[\hbar]]$ be the bundle of endomorphisms of the fibers of $V[[\hbar]]$.

We may define a star product on the endomorphism $\Gamma(E[[\hbar]])$ of the bundle $V[[\hbar]]$ by Fedosov's construction.[13] To this end one considers the algebra,

$$\mathcal{A} = \Gamma(E[[\hbar]]\otimes W),$$

where W is the Weyl algebra bundle over M (i.e., the bundle over M whose fiber over q consists of Weyl algebra on the symplectic vector space T_qM).

$\mathcal{A}$ is graded as a vector space, with a degree defined by

$$\deg(\hbar) = 2, \quad \deg(y^i) = 1,$$

where the (y^i) are coordinates on the fiber T_qM. Defining $\mathcal{A}_i$ as the set of all $\hat{X}\in\mathcal{A}$ which are sums of terms of degree larger or equal i, $\mathcal{A}$ becomes a filtered algebra:

$$\mathcal{A} = \mathcal{A}_0 \supset \mathcal{A}_1 \supset \cdots .$$

The Hermitian connection $\mathbf{\nabla}$ on V induces a connection on $E[[\hbar]]$ which we denote by ∇ as well. Together with the symplectic connection ∂_s on M this defines a connection $\partial = 1\otimes\partial_s + \mathbf{\nabla}\otimes 1$ on $E[[\hbar]]\otimes W$.

One may construct by an iterative procedure, starting with ∂, an exterior covariant derivative D on $\mathcal{A}\otimes\Omega(M)$, the forms on M with values in $\mathcal{A}$, such that we have the following

(1) $D^2 = 0$.
(2) The space of covariant constant elements in $\mathcal{A} \hat{=} \mathcal{A}\otimes\Omega_0(M) \subset \mathcal{A}\otimes\Omega(M)$ is isomorphic as a vector space to $E[[\hbar]]$.
(3) D acts as a derivation on $\mathcal{A}\otimes\Omega(M)$ (with the product induced by the Moyal product on $\mathcal{A}$).

These properties guarantee that the covariant constant sections form a subalgebra and allow us to use the vector space isomorphism of 2 between the covariant constant elements in $\mathcal{A}$ and $E[[\hbar]]$ to define the star product on $\Gamma(E[[\hbar]])$ by pull back of the product on $\mathcal{A}$. More precisely, if we denote the isomorphism from $\Gamma(E[[\hbar]])$ to the covariantly constant elements in $\mathcal{A}$ by $\mathcal{Q}$, and its inverse by σ, the star product on $E[[\hbar]]$ is given by

$$a * b = \sigma(\mathcal{Q}(a) \diamondsuit \mathcal{Q}(b)),$$

for $a,b\in\Gamma(E[[\hbar]])$.

To give an explicit expression for this flat derivative, we introduce operators δ and δ^{-1} on $\mathcal{A}\otimes\Omega(M)$:

$$\delta = dx^i \wedge \frac{\partial}{\partial y^i}$$

and

$$\delta^{-1}a=\frac{1}{p+q}\,i_{y^i(\partial/\partial x^i)}a,$$

if a is a q-form that is homogeneous of degree p in the fiber coordinate y, $p+q\neq 0$, and $\delta^{-1}a=0$ if $p=q=0$. These operators have the important property that for any $\hat{X}\in\Omega(M)\otimes\mathcal{A}$:

$$\hat{X}=\hat{X}_{00}+(\delta^{-1}\delta+\delta\delta^{-1})\hat{X}.$$

Here, $\hat{X}_{00}$ denotes the part of $\hat{X}$ that is a zero form of degree 0 in y.

The ansatz,

$$D=-\delta+\partial+\left[\frac{i}{\hbar}\,r,\ \ \right],$$

for a one form r with values in $\mathcal{A}$, the flatness condition $D^2=0$ together with a ''normalization condition'' $\delta^{-1}r=0$ lead to the condition

$$r=\delta^{-1}\tilde{R}+\delta^{-1}\left(\partial r+\frac{i}{\hbar}\,r^2\right)-\delta^{-1}\Omega, \tag{1}$$

for a power series $\Omega=\Sigma_{i=1}^{\infty}\hbar^i\omega_i$ with coefficients ω^i in the closed 2-forms on M, by which r is uniquely defined and may be computed iteratively. Here, $\tilde{R}$ is the curvature of the connection ∂, considered as a two form with values in $\mathcal{A}$, acting on $\mathcal{A}$ as $(i/\hbar)[\tilde{R},\cdot\,]$: If R^E is the curvature of ∇ and R^s is the curvature of ∂_s, then $\tilde{R}$ is locally given by

$$\tilde{R}=-i\,\frac{\hbar}{2}\,R^E_{ij}dx^i\wedge dx^j+1\otimes\frac{1}{4}\,\omega_{ij}(R^s)^i_{klm}y^jy^kdx^l\wedge dx^m.$$

The freedom of choosing Ω comes from the fact that the curvature of D is not necessarily zero, but it is sufficient if it is a central 2-form, i.e., one commuting with all elements in $\mathcal{A}\otimes\Omega(M)$. Whereas the zeroth order term in $\hbar$ is always the symplectic form, the higher order terms may be chosen as arbitrary closed 2-forms. It is known that different choices of Ω lead in general to inequivalent star products, i.e., inequivalent quantizations.[17,18] Two star products are equivalent iff the Ω's differ by a power series with values in the exact 2-forms. Thus, the equivalence classes are parametrized by power series with coefficients in the second de Rham cohomology class of M.

For our considerations, this simply means that we have to stick with a fixed choice of Ω: All the statements below are true, as long as we do not change Ω (or rather its cohmology class.) Different choices of Hermitian connections will lead to equivalent star products, as long as Ω is not changed, and all formulas derived below will hold. We do not claim that any of the possible choices is *a priori* preferable. Rather, our approach is to assume that we are given a definite physical system, including the physically correct quantization, and we do not try to find a distinguished quantization on *a priori* reasons. Thus, we assume that we are given a definite Ω from the very beginning, and we will never change it throughout our analysis, in order not to change physics.

We note that in spite of the factor $1/\hbar$ in (1) the solution of this equation is a power series in $\hbar$, containing no negative powers.

With r determined, the covariantly constant continuation $\mathcal{Q}(a)$ of a in $\mathcal{A}$ is given by the unique solution to the equation,

$$\mathcal{Q}(a)=a_0+\delta^{-1}\left(\partial\mathcal{Q}(a)+\left[\frac{i}{\hbar}\,r,\mathcal{Q}(a)\right]\right), \tag{2}$$

which may again be solved iteratively.

Different choices of ∇ will yield different star products. However, they are equivalent in the following sense.

Given two Hermitian connections $\nabla^{(1)},\nabla^{(2)}$ and two symplectic connections $\partial_s^{(1)},\partial_s^{(2)}$ with corresponding derivatives $D^{(1)},D^{(2)}$, there is a formally unitary,

$$U_{(12)}=\exp_\diamond\left(\frac{i}{\hbar}A_{(12)}\right),$$

for some $A\in\mathcal{A}$ such that

$$D^{(1)}\hat{X}=0\Rightarrow D^{(2)}(U_{(12)}\diamond\hat{X}\diamond U_{(12)}^{-1})=0,$$

for any $\hat{X}\in\Gamma(\mathcal{A})$. (Here, $\exp_\diamond$ denotes the exponential with respect to the fiberwise Moyal product $\diamond$.)

Furthermore, letting $\mathcal{Q}^{(i)}(X)$ denote the covariant constant continuation of $X\in\Gamma(E[[\hbar]])$ with respect to $D^{(i)}$, and $\sigma^{(i)}$ denote the inverse of $\mathcal{Q}^{(i)}$ $(i=1,2)$, we have an isomorphism of the different star products given by

$$\phi(X)\overset{\text{def}}{=}\sigma^{(2)}(U_{(12)}\diamond\mathcal{Q}^{(1)}(X)\diamond U_{(12)}^{-1}).$$

To check this statement one simply has to compute

$$\begin{aligned}\phi(X*_1Y)&=\sigma^{(2)}(U_{(12)}\diamond\mathcal{Q}^{(1)}(X*_1Y)\diamond U_{(12)}^{-1})\\&=\sigma^{(2)}(U_{(12)}\diamond\mathcal{Q}^{(1)}(X)\diamond\mathcal{Q}^{(1)}(Y)\diamond U_{(12)}^{-1})\\&=\sigma^{(2)}((U_{(12)}\diamond\mathcal{Q}^{(1)}(X)\diamond U_{(12)}^{-1})\diamond(U_{(12)}\diamond\mathcal{Q}^{(1)}(Y)\diamond U_{(12)}^{-1}))\\&=\sigma^{(2)}(\mathcal{Q}^{(2)}(\phi(X))\diamond\mathcal{Q}^{(2)}(\phi(Y)))=\phi(X)*_2\phi(Y).\end{aligned}$$

We note that, in general, $\phi(X)\neq X$. This is precisely what was meant in the introduction by the statement that we can change the quantization procedure without changing physics, if we consistently correct the symbols of the Hamiltonian and the relevant observables at the same time.

We may express $U_{(12)}\diamond\hat{X}\diamond U_{(12)}^{-1}$ directly in terms of $A_{(12)}$ as

$$U_{(12)}\diamond\hat{X}\diamond U_{(12)}^{-1}=\sum_{k=0}^{\infty}\left(\frac{i}{\hbar}\right)^k\frac{1}{k!}[A_{(12)},[\underbrace{A_{(12)},\ldots}_{k\text{-times}}[A_{(12)},\hat{X}]\ldots]]. \tag{3}$$

For $A_{(12)}$, one may derive the following equation which defines it iteratively:[13]

$$A_{(12)}=\delta^{-1}\left(\frac{ad\left(\frac{i}{\hbar}A_{(12)}\right)}{\exp_\diamond\left(ad\left(\frac{i}{\hbar}A_{(12)}\right)\right)-1}(\Delta\gamma)_{(12)}+(D+\delta)A_{(12)}\right). \tag{4}$$

Here, $(\Delta\gamma)_{(12)}$ denotes the one form on M with values in $\mathcal{A}$ (acting by $(i/\hbar)[(\Delta\gamma)_{(12)}\cdot]$) which is induced by the difference of the two connection forms corresponding to $\partial^{(i)}=1\otimes\partial_s^{(i)}+\nabla^{(i)}\otimes 1$, $i=1,2$ on $E[[\hbar]]\otimes W$. (The difference of two connection forms is tensorial and hence defines a one form on M.) If $\nabla^{(i)}$ is locally given by $\partial/\partial x^i+\Gamma_i$, and the Christoffel symbols corresponding to $\partial_s^{(i)}$ are given by $(\Gamma_s^{(i)})_{lm}^k$; then $(\Delta\gamma)_{(12)}$ is locally given by

$$(\Delta\gamma)_{12}=-i\hbar(\Gamma_i^{(2)}-\Gamma_i^{(1)})dx^i+1\otimes\tfrac{1}{2}\,\omega_{ij}((\Gamma_s^{(2)})_{kl}^i-(\Gamma_s^{(1)})_{kl}^i)y^jy^kdx^l. \tag{5}$$

In particular, it is obvious from this expression that $A_{(12)}$ does not contain any negative powers of $\hbar$.

Remarks:

(1) Though U does contain inverse powers of $\hbar$, $U_{(12)} \diamond \hat{X} \diamond U_{(12)}^{-1}$ is again a power series in $\hbar$ for all $\hat{X} \in \Gamma(\mathcal{A})$ with $D^{(1)}\hat{X}=0$. This is shown in Ref. 13 by showing first that $D^{(2)} \times (U_{(12)} \diamond \hat{X} \diamond U_{(12)}^{-1})=0$ in an enlarged algebra W^+ whose elements may contain negative powers of $\hbar$ in some specified way, and then showing that any $Y \in W^+$ such that $D^{(2)}Y=0$ and such that $\sigma^{(2)}(Y)$ is a power series in $\hbar$, lies in the subalgebra W itself.

(2) The mappings $\sigma^{(i)}$ may be considered as the restriction of one mapping σ from $\mathcal{A}$ to $\Gamma(E[[\hbar]])$. This mapping is simply given by setting y to zero, i.e., it picks out the part of $\hat{X} \in \mathcal{A}$ not depending on y. We will use the notions of Ref. 13 and call $\sigma(\hat{X})$ the symbol of $\hat{X} \in \mathcal{A}$.

From now on we will restrict ourselves to one fixed symplectic connection ∂_s on M and will only modify the connection $\boldsymbol{\nabla}$ on V. We will prove two lemmas which will be essential in the following sections.

Lemma 1: Let $\boldsymbol{\nabla}^{(1)}$ and $\boldsymbol{\nabla}^{(2)}$ be connections on $E[[\hbar]]$, ∂_s a fixed symplectic connection, and let $D^{(1)}, D^{(2)}$ denote the corresponding flat Fedosov connections, and $U_{(12)}$ the corresponding automorphism. Assume that the difference of the two connections $\boldsymbol{\nabla}^{(1)}$ and $\boldsymbol{\nabla}^{(2)}$ is of order $O(\hbar^m)$ for some m. Then

$$\sigma(U_{(12)} \diamond \hat{X} \diamond U_{(12)}^{-1}) = \sigma(\hat{X}) + O(\hbar^{(m+1)}),$$

for any $\hat{X} \in \mathcal{A}$ which is covariantly constant with respect to $D^{(1)}$.

Proof: By (5), $A_{(12)}$ is contained in $\mathcal{A}_{2m+3}$. Hence, using (3), we may conclude that $U_{(12)} \diamond \hat{X} \diamond U_{(12)}^{-1} - \hat{X} \in \mathcal{A}_{2m+1}$, and the symbols $\sigma(U_{(12)} \diamond \hat{X} \diamond U_{(12)}^{-1}), \sigma(\hat{X})$ may differ only by a term of order $O(\hbar^{m+1})$. Q.E.D.

If we consider three different connections, $U_{(13)}$ is not necessarily the same as the composition $U_{(23)} \diamond U_{(12)}$ of $U_{(12)}$ and $U_{(23)}$, but $U_{(13)}^{-1} \diamond U_{(23)} \diamond U_{(12)}$ defines an automorphism of the elements in $\mathcal{A}$ covariantly constant with respect to $D^{(1)}$. Although this automorphism in general is not simply the identity, we have the following lemma.

Lemma 2: Let $\boldsymbol{\nabla}^{(1)}$, $\boldsymbol{\nabla}^{(2)}$, and $\boldsymbol{\nabla}^{(3)}$ be connections on $E[[\hbar]]$, ∂_s a fixed symplectic connection, and let $D^{(i)}$ denote the corresponding flat Fedosov connections and $U_{(ik)}$ the corresponding automorphisms of $\mathcal{A}$ mapping covariant constant sections with respect to $D^{(i)}$ to covariant constant sections with respect to $D^{(k)}$. Assume that the difference of the two connections $\boldsymbol{\nabla}^{(2)}$ and $\boldsymbol{\nabla}^{(3)}$ is of order $O(\hbar^m)$ for some m. Then

$$\sigma(U_{(13)} \diamond \hat{X} \diamond U_{(13)}^{-1}) = \sigma((U_{(23)} \diamond U_{(12)}) \diamond \hat{X} \diamond (U_{(23)} \diamond U_{(12)})^{-1}) + O(\hbar^{(m+2)}),$$

for any $\hat{X} \in \mathcal{A}$ which is covariantly constant with respect to $D^{(1)}$.

Proof: By the Baker–Campbell–Hausdorff formula, $U_{(23)} \diamond U_{(12)}$ may be written as $\exp_\diamond((i/\hbar)A)$ with $A = A_{(12)} + A_{(23)} + C$, where C is the sum of multiple commutators of $(1/\hbar)A_{(12)}$ and $(1/\hbar)A_{(23)}$ multiplied by $\hbar$. From formulas (4) and (5) it is obvious that A does not contain negative powers of $\hbar$.

Now, by assumption and (4), (5), $A_{(23)}$ is contained in $\mathcal{A}_{2m+3}$, and is of the special form

$$A_{23} = \hbar^{m+1} T^1_{(23)} + \hbar^m T^2_{(23)} + T^3_{(23)},$$

where $T^1_{(23)}$ contains only terms at least of degree 1 in the fiber coordinate y, $T^2_{(23)}$ contains only terms at least of degree 3 in y, and $T^3_{(23)} \in \mathcal{A}_{2m+4}$.

Similarly, $A_{12} = \hbar T^1_{12} + T^2_{12} + T^3_{12}$ with $T^1_{(12)}$ containing only terms at least of degree 1 in the fiber coordinate y, $T^2_{(12)}$ containing only terms at least of degree 3 in y, and $T^3_{(12)} \in \mathcal{A}_4$. Now, since all other terms in the Baker–Cambell–Hausdorff formula will be contained in $\mathcal{A}_{2m+5}$, we may conclude that

$$C = \frac{1}{\hbar}[\hbar T^1_{12} + T^2_{12}, \hbar^{m+1} T^1_{12} + \hbar^m T^2_{12}] + \cdots,$$

where the missing terms denoted by dots are in $\mathcal{A}_{2m+5}$. An inspection of each term, using the fact that T^2_{12}, T^2_{23} are scalar, finally shows that C is of the form

$$C=\hbar^m C^1+\hbar^{m+1}C^2+\hbar^{m+2}C^3+C^4,$$

where C^1 only contains terms at least quadratic in y, C^2 only terms at least of degree 1 in y, and $C^4\in\mathcal{A}_{2m+5}$. Since C^1 and C^2 contain no terms independent of y, they may contribute to the symbol of $\sigma((U_{(23)}\diamond U_{(12)})\diamond\hat{X}\diamond(U_{(23)}\diamond U_{(12)})^{-1})$ only through higher order terms of the Moyal product. Hence, the whole contribution of C to the symbol is of order $\hbar^{m+2}$.

On the other hand, we see from (4) and $(\Delta\gamma)_{(13)}=(\Delta\gamma)_{(12)}+(\Delta\gamma)_{(23)}$ that $A_{(13)}=A_{(12)}+A_{(13)}+\tilde{C}$, where $\tilde{C}\in\mathcal{A}_{2m+4}$ as well. A close inspection of the possible term in $\tilde{C}$ shows again that it contributes to the symbol of $U_{(13)}\diamond\hat{X}\diamond(U_{(13)})^{-1}$ only at order $\hbar^{m+2}$. Thus, the difference of the two symbols is of order $\hbar^{m+2}$. Q.E.D.

Let $\tilde{\nabla}$ be an arbitrary (possibly $\hbar$-dependent) connection on $V[[\hbar]]$, and π_α a set of projection valued sections in $E[[\hbar]]$ with $\pi_\alpha\pi_\beta=\delta_{\alpha\beta}\pi_\alpha$, $\Sigma_\alpha\pi_\alpha=id$. Define $\nabla=\Sigma_\alpha\pi_\alpha\circ\tilde{\nabla}\circ\pi_\alpha$ as a ''Berry-type connection.'' Then $\nabla\pi_\alpha=0$, and we have the following lemma.

Lemma 3: With the notations above, and a fixed symplectic connection ∂_s, let $*$ denote the Fedosov star product corresponding to ∇. Then, π_α are ''quantum projections,'' i.e., $\pi_\alpha*\pi_\beta=\delta_{\alpha\beta}\pi_\alpha$.

Proof: We will show $\pi_\alpha*\pi_\beta=\pi_\alpha\cdot\pi_\beta$. We first note that, since $\nabla\pi_\alpha=0$, $\forall\alpha$, it follows from (1) that $[r,\pi_\alpha]=0$, $\forall\alpha$. (Since π_α does not depend on y, there is no difference between $\diamond$-commutator and ordinary matrix commutator.) Then, from (2), $\mathcal{Q}(\pi_\alpha)=\pi_\alpha$ for all α. But then we have $\pi_\alpha*\pi_\beta=\pi_\alpha\cdot\pi_\beta=\delta_{\alpha\beta}\pi_\alpha$. Q.E.D.

III. BERRY CONNECTIONS AND ''BLOCK DIAGONALIZATION''

After the technical preliminaries of the last section, we are now in the position to address the actual problem of this paper.

We consider a Hamiltonian which is a section in the bundle $E[[\hbar]]$:

$$H=H_0+\hbar H_1+\cdots.$$

We assume that the eigenvalues $\lambda_\alpha(x)$ of $H_0(x)$ $(\alpha=1,...,r)$, have constant multiplicities m_α throughout phase space. We denote by $\pi_\alpha^{(0)}(x)$ the projections of V_x onto the α-th eigenspace of $H_0(x)$. Under the regularity assumptions above, these eigenspaces form smooth vector bundles over M.

We assume that we are given a Hermitian connection on V and a symplectic connection on M, which together define a star product; in many applications there is a distinguished choice of connections and quantization. For example, in the case of the Born-Oppenheimer approximation, one may start with a symplectic manifold which is just the nuclear phase space $\mathbb{R}^{2N}$ where N is three times the number of nuclei in the molecule. In this case, there is a distinguished quantization prescription, namely Weyl ordering, which corresponds to the choice of the trivial connection on the trivial bundle V.

Hence, we assume that we are given a definite quantization. As we allow the symbols of the observables to depend on $\hbar$, one may alternatively interpret this as the choice of a definite symbol calculus.

Nevertheless, we still have the freedom of changing the connection defining our star product without changing the physics, if we apply the corresponding isomorphism of the resulting algebras at the same time, i.e., we have to correct the symbol H by applying the isomorphism ϕ introduced above. We can use this freedom to reduce the problem to a simpler one, which is closer to the scalar case. That this is possible indeed is the content of the next theorem.

Theorem 4: *There exists a formally orthogonal decomposition $V[[\hbar]]=\oplus_{\alpha=1}^r V_\alpha$, $\dim(V_\alpha)=m_\alpha$ with corresponding quantum projections $\pi_\alpha^{(\infty)}$ (i.e., $\pi_\alpha^{(\infty)}*_\infty\pi_\beta^{(\infty)}=\delta_{\alpha\beta}\pi_\alpha^{(\infty)}$), and a compatible connection $\nabla^{(\infty)}$ [i.e., $\nabla^{(\infty)}\pi_\alpha^{(\infty)}=0$, $\forall\alpha$] such that the corrected symbol H_{corr} preserves the decomposition: $[H_{corr},\pi_\alpha^{(\infty)}]=0$ for $\alpha=1,...,r$. (Here, [,] denotes the $*_\infty$ commutator.)*

Proof: We show the theorem by induction on the order in $\hbar$.

For the connections $\nabla^{(k)}$ to be constructed below we denote by ϕ_k the isomorphism of the algebras $\Gamma(E[[\hbar]])$ with star products $*,*_k$ constructed from ∇ and $\nabla^{(k)}$, respectively: $\phi_k(X*Y)=\phi_k(X)*_k\phi_k(Y)$.

Let $\pi_\alpha^{(1)}(x) \overset{\text{def}}{=} \pi_\alpha^{(0)}(x)$ denote the projection onto the i-th eigenspace of $H_0(x)$ (i.e., of the zeroth order term of the original Hamiltonian H), and define $\nabla^{(1)}$ as the Berry-type connection:

$$\nabla^{(1)}\psi \overset{\text{def}}{=} \sum_\alpha \pi_\alpha^{(1)} \nabla(\pi_\alpha^{(1)}\psi).$$

The difference between the corrected Hamiltonian $H^{(1)} = \phi_1(H)$, which is obtained from H by applying the isomorphism corresponding to the change of the connection from ∇ to $\nabla^{(1)}$ is of order $O(\hbar)$ by Lemma 1. Hence, $\nabla^{(1)}\pi_\alpha^{(1)} \equiv 0$ by construction, and $[H^{(1)}, \pi_\alpha^{(1)}] = O(\hbar)$ for all α.

Now assume that we have found (formally orthogonal) projections $\pi_\alpha^{(k)}$ with $\pi_\alpha^{(k)} = \pi_\alpha^{(0)} + O(\hbar)$ with a corresponding Berry-type connection,

$$\nabla^{(k)}\psi \overset{\text{def}}{=} \sum_\alpha \pi_\alpha^{(k)} \nabla(\pi_\alpha^{(k)}\psi),$$

such that the corrected Hamiltonian $H^{(k)} = \phi_k(H)$, satisfies $[H^{(k)}, \pi_\alpha^{(k)}] = O(\hbar^k)$. We will construct new orthogonal projections $\pi_\alpha^{(k+1)}$ satisfying the requirement

$$[H^{(k)}, \pi_\alpha^{(k+1)}] = O(\hbar^{k+1}), \quad \forall \alpha. \tag{6}$$

To this end, we make the ansatz:

$$\pi_\alpha^{(k+1)} = \exp(i\hbar^k A)\, \pi_\alpha^{(k)}\, \exp(-i\hbar^k A), \tag{7}$$

for a fixed A independent of α. If we can find such an A, equation (7) automatically defines new formally orthogonal projections (i.e., $\pi_\alpha^{(k+1)} *_{k+1} \pi_\beta^{(k+1)} = \pi_\alpha^{(k+1)} \pi_\beta^{(k+1)} = \delta_{\alpha\beta}\pi_\alpha^{(k+1)}$, according to lemma 3). With this ansatz $\pi_\alpha^{(k+1)}$ is given up to order $O(\hbar^{k+1})$ by

$$\pi_\alpha^{(k+1)} = \pi_\alpha^{(k)} + i\hbar^k [A, \pi_\alpha^{(k)}] + O(\hbar^{k+1}).$$

Hence, equation (6) yields up to order $O(\hbar^{k+1})$ the condition

$$[H^{(k)}, \pi_\alpha^{(k)}] + i\hbar^k [H_0, [A, \pi_\alpha^{(0)}]] = O(\hbar^{k+1}), \quad \forall \alpha.$$

As the $\pi_\alpha^{(k)}$ are projections, we may write

$$H^{(k)} = \sum_\alpha \pi_\alpha^{(k)} H^{(k)} \pi_\alpha^{(k)} + \hbar^k W,$$

where the first term on the right hand side commutes with $\pi_\alpha^{(k)}$ and

$$\hbar^k W = \sum_{\beta \neq \gamma} \pi_\beta^{(k)} H^{(k)} \pi_\gamma^{(k)} = \sum_{\beta \neq \gamma} \pi_\beta^{(k)} [H^{(k)}, \pi_\gamma^{(k)}] \pi_\gamma^{(k)}.$$

As the commutator on the right hand side is of order $O(\hbar^k)$ the left hand side is of the same order. Hence, we have to solve

$$\hbar^k [W, \pi_\alpha^{(0)}] + i\hbar^k [H_0, [A, \pi_\alpha^{(0)}]] = O(\hbar^{k+1}), \quad \forall \alpha.$$

Note that in this equation only the projections $\pi_\alpha^{(0)}$ $(\alpha = 1, \ldots, r)$ appear. A solution to this equation is given by

$$A = i \sum_{\alpha \neq \beta} \frac{\pi_\alpha^{(k)} W \pi_\beta^{(k)}}{\lambda_\alpha - \lambda_\beta},$$

as may be easily verified using $\pi_\alpha^k W \pi_\alpha^k = 0$ for all α by construction of W.

Using the new projections we may define a new Berry-type connection $\nabla^{(k+1)}$ which preserves the decomposition defined by the projections even to arbitrary order in $\hbar$. As the difference of the projections is of order $O(\hbar^k)$, the difference of the connections $\nabla^{(k)}$ and $\nabla^{(k+1)}$ is of the same order, and hence, by Lemmas 1 and 2, the corrected symbol $H^{(k+1)}=\phi_{k+1}(H)$ differs from $H^{(k)}$ only at order $O(\hbar^{k+1})$ Thus, by (6), $H^{(k+1)}$ is compatible with the decomposition up to $O(\hbar^{k+1})$: $[H^{(k+1)},\pi_\alpha^{(k+1)}]=O(\hbar^{k+1})$.

As $\nabla^{(k)}$ and $\nabla^{(k+1)}$ differ only at order $O(\hbar^k)$, the connections $\nabla^{(k)}$ have a well defined limit ∇^∞, which by induction fulfills the assertions of the theorem. Q.E.D.

The importance of theorem 4 lies in the following two corollaries, which are a kind of generalization to our much more general setting of the results of the paper,[19] which deals only with the Dirac equation on flat space-time.

Corollary 5: Let $X,Y\in\Gamma E[[\hbar]]$ such that $[X,\pi_\alpha^{(\infty)}]=[Y,\pi_\alpha^{(\infty)}]=0$ for all α. Let $*_\infty$ denote the star product constructed from $\nabla^{(\infty)}$ by Fedosov's construction. Then $[X*_\infty Y,\pi_\alpha^{(\infty)}]=0$ for all α.

Proof: As $\nabla^{(\infty)}\pi_\alpha^{(\infty)}=0$, the curvature of $\nabla^{(\infty)}$ commutes with $\pi_\alpha^{(\infty)}$. As δ^{-1} only acts on the second factor of $\mathcal{A}=\Gamma(E[[\hbar]]\otimes W)$ and the Moyal product $\diamond$ only involves derivatives with respect to the fiber coordinates y, we may conclude from (1) by induction on the total degree that r commutes with $\pi_\alpha^{(\infty)}$. (Observe that we may write the term $r^2=r\diamond r$ as $1/2[r,r]_\diamond$.)

Now, the covariant constant continuation $\hat{X}=\mathcal{Q}^{(\infty)}(X)$ of X is given by[13]

$$\hat{X}=X+\delta^{-1}\left(\partial\hat{X}+\left[\frac{i}{\hbar}\,r,\hat{X}\right]_\diamond\right), \tag{8}$$

and we may again conclude inductively, $[\hat{X},\pi_\alpha^{(\infty)}]=0$, and similarly for $\hat{Y}$.

Using again the fact that the Moyal product $\diamond$ on $\mathcal{A}$ only involves derivatives with respect to y, it follows that $\hat{X}\diamond\hat{Y}$ commutes with $\pi_\alpha^{(\infty)}$ for all α. Hence, by taking the symbol, i.e., by setting $y=0$, we get the assertion of the corollary. Q.E.D.

Corollary 6: Denote by ϕ_∞ the isomorphism of the algebras $\Gamma(E[[\hbar]])$ with products $*$ and $*_\infty$. Let

$$\mathcal{O}\overset{\text{def}}{=}\{\phi_\infty^{-1}(\tilde{X})\,|\,\tilde{X}\in\Gamma(E[[\hbar]]),[\tilde{X},\pi_\alpha^{(\infty)}]=0,\forall\,\alpha\}.$$

Then, $\mathcal{O}$ is a subalgebra of $\Gamma(E[[\hbar]])$ for which the Heisenberg time evolution,

$$\dot{X}=\frac{1}{i\hbar}\,[H,X]_*,$$

is a well defined differential equation.

Proof: $\mathcal{O}$ is a subalgebra of $\Gamma(E[[\hbar]])$ by the preceding corollary. We only have to show that $[H,X]_*=O(\hbar)$, since then there will be no negative powers of $\hbar$ in the Heisenberg time evolution.

Denoting again $\phi_\infty(H)$ by H_{corr}, we know that H_{corr} is of the form $\Sigma_\alpha\lambda_\alpha\pi_\alpha^{(\infty)}+O(\hbar)$.

Now, by (1), r is of the form $r=1\otimes\hat{r}+O(\hbar)$ for a scalar one-form $\hat{r}$. From this one may conclude that the covariantly constant continuation $\mathcal{Q}^{(\infty)}(H_{corr})$ of H_{corr} is of the form $\Sigma_\alpha\rho_\alpha\pi_\alpha^{(\infty)}+O(\hbar)$ for some scalar ρ_α. Thus, since $\tilde{X}$ commutes with $\pi_\alpha^{(\infty)}$, we may conclude that the classical matrix commutator does not contribute to $[H_{corr},\tilde{X}]_{*_\infty}$ at zeroth order in $\hbar$. Hence, the commutator is of order $\hbar$. Applying ϕ_∞^{-1}, the corollary follows. Q.E.D.

IV. CORRECTION AT ORDER $\hbar$

In the previous section we have achieved a complete reduction of the multicomponent WKB problem to a set of different polarizations. In this section, we will explicitly compute the corrections to the symbol of the Hamiltonian at order $\hbar$. Thus, this section will generalize the results of Ref. 7, and give a completely geometric derivation of the results derived there for trivial bundles only. We achieve this by first using the Berry-type connection as in the proof of theorem 4 in the

previous section. However, the connection is not uniquely defined by the compatibility conditions: Hence, in the second subsection, we will investigate the dependence of the structure of this phase on the choice of a compatible connection (in the sense of theorem 4). It will turn out that it is unique up to a Berry-type phase, and that the Berry-type connection considered so far is geometrically distinguished among all compatible connections.

A. Berry-type connection

To compute the correction of the symbol H at order $\hbar$, we only have to carry out the first two steps of the iterative procedure in the proof of the previous theorem. Actually, it is not necessary to compute $H^{(2)}$, but it is sufficient to compute the ''block diagonal'' part of $H^{(1)}$ since

$$\begin{aligned} H_{corr} &= H^{(2)} + O(\hbar^2) \\ &= \sum_\alpha \pi_\alpha^{(2)} H^{(1)} \pi_\alpha^{(2)} + O(\hbar^2) \\ &= \sum_\alpha \exp(i\hbar A)\,\pi_\alpha^{(1)} \exp(-i\hbar A) H^{(1)} \exp(i\hbar A)\,\pi_\alpha^{(1)} \exp(-i\hbar A) \\ &= \sum_\alpha \exp(i\hbar A)\,\pi_\alpha^{(1)} (H^{(1)} + i\hbar [H^{(1)}, A])\,\pi_\alpha^{(1)} \exp(-i\hbar A) + O(\hbar^2) \\ &= \sum_\alpha \exp(i\hbar A)\,\pi_\alpha^{(0)} H^{(1)} \pi_\alpha^{(0)} \exp(-i\hbar A) + O(\hbar^2), \end{aligned}$$

with A as defined in the proof of theorem 4, and $\pi_\alpha^{(1)} = \pi_\alpha^{(0)}$.

Here, we used the fact that the zeroth degree part $H_0^{(1)}$ of $H^{(1)}$ is just $H_0^{(0)}$, and that $\pi^{(0)}$ is the projection onto the eigenspaces of $H_0^{(0)}$. Hence, in order to determine the correction of the symbol at order $\hbar$ up to a unitary rotation, we only have to compute $\pi_\alpha^{(0)} H^{(1)} \pi_\alpha^{(0)}$ for all α.

Defining $\gamma_l \overset{\text{def}}{=} \Sigma_\alpha \pi_\alpha^{(0)} (\nabla_l \pi_\alpha^{(0)})$ we get by formulas (3), (4):

$$\begin{aligned} \pi_\alpha^{(0)} H^{(1)} \pi_\alpha^{(0)} &= \pi_\alpha^{(0)} (H^{(0)} + \sigma([\gamma_l y^l, \nabla_\alpha H_0^{(0)} y^i]_\diamond) + \tfrac{1}{2}\,\sigma([\gamma_l y^l, [\gamma_m y^m, H_0^{(0)}]_{\mathrm{cl}}]_\diamond)\,\pi_\alpha^{(0)} + O(\hbar^2) \\ &= \pi_\alpha^{(0)} H^{(0)} \pi_\alpha^{(0)} + \pi_\alpha^{(0)} \left(\omega^{il} \frac{i\hbar}{2} (\gamma_i \nabla_l H_0^{(0)} - \nabla_i H_0^{(0)} \gamma_l) \right) \pi_\alpha^{(0)} \\ &\quad + \pi_\alpha^{(0)} \frac{i\hbar}{4} \omega^{il} (\gamma_i [\gamma_l, H_0^{(0)}]_{\mathrm{cl}} - [\gamma_i, H_0^{(0)}]_{\mathrm{cl}} \gamma_l)\,\pi_\alpha^{(0)} + O(\hbar^2). \end{aligned} \tag{9}$$

Here, $[\cdot,\cdot]_{\mathrm{cl}}$ denotes the usual matrix commutator, and $[\cdot,\cdot]_\diamond$ the Moyal product commutator.

In the computation, the following relations (following trivially from the projection property) prove to be useful:

$$\pi_\alpha^{(0)} (\nabla \pi_\beta^{(0)})\,\pi_\alpha^{(0)} = 0, \quad \forall\, \alpha, \beta,$$

$$\pi_\alpha^{(0)} (\nabla \pi_\beta^{(0)})\,\pi_\beta^{(0)} = \pi_\alpha^{(0)} (\nabla \pi_\beta^{(0)}), \quad \pi_\beta^{(0)} (\nabla \pi_\beta^{(0)})\,\pi_\alpha^{(0)} = (\nabla \pi_\beta^{(0)})\,\pi_\alpha^{(0)}, \quad \forall\, \alpha \neq \beta.$$

Using $H_0^{(0)}(x) = \Sigma_\alpha \lambda_\alpha(x)\,\pi_\alpha^{(0)}(x)$, we get for the second term on the right hand side of the last equation in (9):

$$\frac{i\hbar}{2}\,\omega^{il}\Bigg(\sum_{\beta}\ \pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\nabla_l(\lambda_\beta\pi_\beta^{(0)})\pi_\alpha^{(0)}-\sum_{\beta,\gamma}\ \pi_\alpha^{(0)}\nabla_i(\lambda_\beta\pi_\beta^{(0)})\pi_\gamma^{(0)}(\nabla_l\pi_\gamma^{(0)})\pi_\alpha^{(0)}\Bigg)$$

$$=\frac{i\hbar}{2}\,\omega^{il}\Bigg(\lambda_\alpha\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})(\nabla_l\pi_\alpha^{(0)})-\sum_{\beta}\ \lambda_\beta\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\pi_\beta^{(0)}(\nabla_l\pi_\alpha^{(0)})$$

$$-\sum_{\beta}\ \lambda_\beta\pi_\alpha^{(0)}(\nabla_i\pi_\beta^{(0)})\pi_\alpha^{(0)}(\nabla_l\pi_\alpha^{(0)})+\sum_{\beta,\gamma}\ \lambda_\beta\pi_\alpha^{(0)}\nabla_i(\pi_\beta^{(0)})\pi_\gamma^{(0)}(\nabla_l\pi_\alpha^{(0)})\Bigg)$$

$$=i\hbar\,\omega^{il}\sum_{\beta}\ \lambda_\beta\pi_\alpha^{(0)}(\nabla_i\pi_\beta^{(0)})(\nabla_l\pi_\alpha^{(0)}).$$

For the last term in (9) we get

$$\frac{i\hbar}{4}\,\omega^{il}\Bigg(\sum_{\beta,\gamma}\ \pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})[\pi_\beta^{(0)}\nabla_l\pi_\beta^{(0)},\lambda_\gamma\pi_\gamma^{(0)}]_{\mathrm{cl}}\pi_\alpha^{(0)}$$

$$-\sum_{\beta,\gamma,\rho}\ \pi_\alpha^{(0)}[\pi_\beta^{(0)}(\nabla_i\pi_\beta^{(0)}),\lambda_\gamma\pi_\gamma^{(0)}]_{\mathrm{cl}}\pi_\rho^{(0)}(\nabla_l\pi_\rho^{(0)})\pi_\alpha^{(0)}\Bigg)$$

$$=\frac{i\hbar}{4}\,\omega^{il}\Bigg(\lambda_\alpha\sum_{\beta}\ \pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\pi_\beta^{(0)}(\nabla_l\pi_\beta^{(0)})\pi_\alpha^{(0)}-\sum_{\gamma}\ \lambda_\gamma\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\pi_\gamma^{(0)}(\nabla_l\pi_\gamma^{(0)})\pi_\alpha^{(0)}$$

$$-\sum_{\gamma}\ \lambda_\gamma\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\pi_\gamma^{(0)}(\nabla_l\pi_\gamma^{(0)})\pi_\alpha^{(0)}+\sum_{\rho}\ \lambda_\alpha\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})\pi_\rho^{(0)}(\nabla_l\pi_\rho^{(0)})\pi_\alpha^{(0)}\Bigg)$$

$$=-\frac{i\hbar}{2}\,\omega^{il}\sum_{\gamma}\ \lambda_\gamma\pi_\alpha^{(0)}(\nabla_i\pi_\gamma^{(0)})(\nabla_l\pi_\alpha^{(0)}).$$

Thus, we finally get

$$\pi_\alpha^{(0)}H^{(1)}\pi_\alpha^{(0)}=\pi_\alpha^{(0)}H^{(0)}\pi_\alpha^{(0)}+\frac{i\hbar}{2}\,\omega^{il}\sum_{\gamma}\ \lambda_\gamma\pi_\alpha^{(0)}(\nabla_i\pi_\gamma^{(0)})(\nabla_l\pi_\alpha^{(0)})$$

$$=\pi_\alpha^{(0)}H^{(0)}\pi_\alpha^{(0)}+\frac{i\hbar}{2}\,\omega^{il}\lambda_\alpha\pi_\alpha^{(0)}(\nabla_i\pi_\alpha^{(0)})(\nabla_l\pi_\alpha^{(0)})$$

$$+\frac{i\hbar}{2}\,\omega^{il}\sum_{\gamma\neq\alpha}\ \lambda_\gamma\pi_\alpha^{(0)}(\nabla_i\pi_\gamma^{(0)})\pi_\gamma^{(0)}(\nabla_l\pi_\alpha^{(0)}).$$

In order to give a geometric interpretation of the correction terms we compute the curvature of the Berry connection:

$$F^B\psi=\sum_{\alpha}\ \pi_\alpha^{(0)}\circ\nabla\circ\pi_o^{(0)}\nabla\,\pi_\alpha^{(0)}\psi$$

$$=\sum_{\alpha}\ \pi_\alpha^{(0)}\nabla^2\pi_\alpha^{(0)}+\sum_{\alpha}\ \pi_\alpha^{(0)}(\nabla\,\pi_\alpha^{(0)})(\nabla\,\pi_\alpha^{(0)})$$

$$=\sum_{\alpha}\ \pi_\alpha^{(0)}F^o\pi_\alpha^{(0)}+\sum\ \pi_\alpha^{(0)}(\nabla\,\pi_\alpha^{(0)})\wedge(\nabla\,\pi_\alpha^{(0)}),$$

where F^o denotes the curvature of the original connection ∇, and the covariant derivatives have to be interpreted as covariant exterior derivatives, acting on forms.

Hence, the coefficient of λ_α of the α-th block in the correction at order $\hbar$ is $\omega^{-1}(\pi_\alpha^{(0)}F^B\pi_\alpha^{(0)}-\pi_\alpha^{(0)}F^0\pi_\alpha^{(0)})$. Here, ω^{-1} is the Poisson tensor. The curvature of the Berry connection is block diagonal by construction, the effect of the projection operators on both sides is just to pick out the right block, corresponding to α. In particular, if the original connection is flat, the coefficient of λ_α on the α-th block of the correction is exactly the Poisson curvature of the Berry connection, i.e., the contraction of the curvature with the Poisson tensor.

In order to interpret the remaining terms we define a vector bundle analog of the second fundamental form for embedded submanifolds:

For arbitrary $\beta\neq\alpha$ set

$$\widetilde{S}^{\beta\alpha}:\quad \Gamma(V_\alpha^{(1)})\to\Gamma(V_\beta^{(1)})\otimes\Gamma(\Lambda^1(M)),$$

$$\psi\mapsto\pi_\beta^{(1)}\circ\nabla\circ\pi_\alpha^{(1)}\psi=\pi_\beta^{(1)}(\nabla\pi_\alpha^{(1)})\psi.$$

$\widetilde{S}^{\beta\alpha}$ obviously is tensorial for $\beta\neq\alpha$; hence it induces a mapping $S^{\alpha\beta}:V_\alpha^{(1)}\to V_\beta^{(1)}\otimes\Lambda^1(M)$ which is a vector bundle analog of the second fundamental form: It measures the extent to which parallel transport with respect to ∇ tends to rotate a vector from a given eigenspace into the others.

For fixed $\alpha\neq\beta$, we may form $S^{\beta\alpha}\wedge S^{\alpha\beta}$, which is a two-form with values in the endomorphims of V_α, and hence may be contracted with the Poisson tensor to define an endomorphism of V_α, which is just the coefficient of λ_β in the α-th block of the first order correction of H.

B. Arbitrary compatible connection

For an arbitrary connection $\widetilde{\nabla}$ which is compatible in the sense of theorem 4 up to some order $\hbar^k$, we first consider the correction at order $\hbar$. By lemma 2, we may compute the correction at order $\hbar$ by first going over to the Berry connection, and then in a second step to the connection $\widetilde{\nabla}$, as the the corrected symbol obtained in this way differs from the one obtained by going over directly from ∇ to $\widetilde{\nabla}$ is of order $O(\hbar^2)$.

As both $\nabla^{(1)}$ and $\widetilde{\nabla}$ are compatible with the projections $\pi_\alpha^{(0)}$ up to $O(\hbar)$, the difference $\Delta\gamma$ of the respective connection forms commutes with $\pi_\alpha^{(0)}$ and $H^{(0)}$ up to order $O(\hbar)$. Hence, denoting the corrected symbol corresponding to the new connection $\widetilde{\nabla}$ by $\widetilde{H}$ we get by formulas (3) and (4),

$$\begin{aligned}\pi_\alpha^{(0)}\widetilde{H}\pi_\alpha^{(0)}&=\pi_\alpha^{(0)}(H^{(1)}+\sigma([(\Delta\gamma)_l y^l,\nabla_i^{(1)}H_0^{(0)}y^i]_\diamond)+\tfrac{1}{2}\,\sigma([(\Delta\gamma)_l y^l,[(\Delta\gamma)_m y^m,H_0^{(0)}]_{\mathrm{cl}}]_\diamond)\pi_\alpha^{(0)}\\&\quad+O(\hbar^2)\\&=\pi_\alpha^{(0)}H^{(1)}\pi_\alpha^{(0)}+\pi_\alpha^{(0)}\omega^{il}\frac{i\hbar}{2}((\Delta\gamma)_i\nabla_l H_0^{(0)}-\nabla_i H_0^{(0)}(\Delta\gamma)_l)\pi_\alpha^{(0)}+O(\hbar^2)\\&=\pi_\alpha^{(0)}H^{(1)}\pi_\alpha^{(0)}+\pi_\alpha^{(0)}\omega^{il}i\hbar((\Delta\gamma)_i(\partial_l\lambda_\alpha))\pi_\alpha^{(0)}+O(\hbar^2)\\&=\pi_\alpha^{(0)}H^{(1)}\pi_\alpha^{(0)}+\pi_\alpha^{(0)}i\hbar(\Delta\gamma)(X_{\lambda_\alpha})\pi_\alpha^{(0)}+O(\hbar^2).\end{aligned}$$

Here, X_{λ_α} denotes the Hamiltonian vector field corresponding to λ_α.

Hence, we see that for an arbitrary connection compatible with $\pi_\alpha^{(0)}$ the lowest order corrections differ just by a Berry phase term. The additional geometric term involving the eigenvalues themselves and not their derivatives are unique and always contain the Poisson curvature and the second fundamental form of the Berry connection and not the Poisson curvature of the chosen connection. Hence, the Berry connection is distinguished among all connections satisfying theorem 4.

V. PHYSICAL DISCUSSION

The effects described in the preceding sections owe their existence to the space time dependence of the eigenspaces of the symbol H associated to the wave equation under consideration. The elements of the eigenspaces are physically interpreted as generalized polarization states.

One example is polarization effects for light propagation in nonisotropic and nonhomogeneous media. Such effects may survive even in the classical limit, as evident from the existence of precession effects for a top in an external gravitational field. The details of the effect depend on the concrete wave equation and its symbol, or, stated differently, on the quantization. We do not know about any general principle dictating in every case the ''correct'' choice of one out of the infinity of *a priori* possible quantizations.

Sometimes, however, the situation is more favorable: For the motion of a particle of spin 1/2 in an external gravitational and/or gauge field, the correct quantization is believed to be given by Dirac's equation. Also, for a molecular system in the Born–Oppenheimer approximation, one normally has a unique quantization procedure.

When no unique evident quantization procedure is available, one would like to trace the influence of a change of quantization. This can be done rather efficiently in the global star product approach adopted in this work. The main tool is Fedosov's star product, and one essential piece of data entering into its construction, and hence also in the determination of the quantization, is a Hermitian connection ∇ which, in addition, also allows to measure the space-time dependence of eigenvectors and eigenspaces. In particular, nonvanishing of any characteristic class which can be constructed in terms of the curvature of the Hermitian connection will imply nontriviality of eigenbundles.

The global star product approach is already quite convenient for studying different quantizations for contractable phase space $\mathbb{R}^{2n}$ when all eigenbundles are necessarily trivial. Globalization becomes indispensible for nontrivial space time and eigenbundles. Examples for such a situation have already been mentioned in the introduction.

Even for known prescribed quantization (and Hermitian connection ∇) it turned out to be advantageous to perform a transformation from ∇ to a new connection ∇' leaving the eigenspaces of the symbol fixed. Supplementing the resulting change of the star product quantization by the corresponding isomorphism of observables leaves physics unchanged, but casts it into a form which can be interpreted more readily.

Among the Hermitian connections ∇' compatible with the eigenbundle decomposition, the Berry connection ∇'_B is distinguished in the following way:

For a section ϕ of an eigenbundle, $\nabla'_B\phi$ is obtained by simply projecting $\nabla\phi$ back onto the eigenspace, without additional rotation. The curvature and second fundamental forms thus arising enjoy a privileged status.

Of course, a global approach can always be replaced by a local procedure, supplemented by rules for changing coordinate patches, but conceptually, a genuinely global scheme is preferable.

Nontrivial base manifolds for the quantization bundles can also effectively arise if one considers parametrized systems of linear differential operators with a parameter λ varying in some manifold Λ. In this case, one considers the projection $p: M\times\Lambda\to M$, and the discussion of parametrized systems is performed in terms of a single differential operator on the pulled back bundle of the quantum bundle to $M\times\Lambda$ by means of p. The ordinary Berry phase, for instance, can be obtained by taking $\Lambda = S^1$.

Considering the additional phase appearing in addition to the Berry-type phase, no general intuitive interpretation is known to us. However, this phase does have physical consequences, e.g., the eigenvalues for WKB states are modified by this additional phase. In Ref. 11 a simple example with a 2×2-matrix valued symbol has been analyzed in detail, using more elementary techniques. This model could both be solved analytically, and using the WKB approximation. The same additional phase was observed there for this special example, and it could be shown that this additional phase does change the eigenvalues of the Hamiltonian in the WKB approximation, and that it is needed to get the correct eigenvalues in the WKB approximation.

Let us finally mention that the obstructions against the existence of global WKB amplitudes to be discussed in the next section are of global character and are inaccessable to local considerations.

VI. OUTLOOK

In the preceding sections, we have achieved a complete block diagonalization of the considered problem: We have found a decomposition of $V[[\hbar]]$ into subspaces, such that the symbol of

the Hamiltonian and the connection are compatible with this decomposition. Furthermore, we have found a maximal set of observables which is compatible with this decomposition as well in that sense, that their time evolution does not contain any negative powers of $\hbar$. Hence, for the purposes of WKB approximation, one may treat each block separately.

However, the problem of multicomponent WKB has not been completely reduced to a scalar problem. In this section, we will show that it can still be reduced closer to a purely scalar one, but only at the price of noncanonical choices of a connection. Nevertheless, this further reduction proves to be useful for a discussion of obstructions to a complete reduction to a scalar problem. Furthermore, for this discussion, only parallel transport along the Hamiltonian vector field will play a role, which does not depend on the choices made in the connection.

We have seen in the last section that the choice of a different connection leads at order $\hbar$ to the addition of a Berry-phase term $\gamma(X_{\lambda_\alpha})$ to the α-th block of the corrected symbol. We may use the Berry-phase term to simplify the problem by choosing the connection such that this term just compensates the remaining terms at order $\hbar$, so that the Hamiltonian is reduced to a set of scalar Hamiltonians up to order $O(\hbar^2)$ on the α-th block. This is always possible as long as the eigenvalue functions have no critical points, where their Hamiltonian vector fields vanish. In practice this means that one has to cut out those points. However, this is not a real problem if one has applications to WKB in mind, as the Lagrangian submanifolds relevant to the WKB approximation lie in level sets of the eigenvalue functions and hence will generically not intersect the critical points. It is possible to iterate the procedure above, which gives the following theorem.

Theorem 7: *Let $\tilde{M}$ be any open submanifold of M not containing critical points of the eigenvalue functions λ_α. Denote the embedding of $\tilde{M}$ by i, and let $\tilde{V}$ be the pull back bundle $i^*(V)$, i.e., the subbundle of V obtained by restricting V to $\tilde{M}$. Then there is an ($\hbar$ dependent) connection $\tilde{\boldsymbol{\nabla}}^\infty$ and a decomposition $\tilde{V}[[\hbar]]=\oplus_{\alpha=1}^m \tilde{V}_\alpha$ fulfilling theorem 4 such that the corrected Hamiltonian restricted to $\tilde{V}_\alpha$ is scalar, namely, the eigenvalue function λ_α of $H^{(0)}$ for each α.*

As stated before, the connection defined in this way is far from unique: Only the partial connection for each block, giving covariant derivatives in the direction of the respective Hamiltonian vector field, is uniquely defined by the iteration procedure above. However, we may always find a continuation of this partial connection to a connection on the whole bundle.

The result of theorems 4 and 7 is optimal in the following sense: We started from a decomposition of V into the bundles of eigenvectors of H_0. However, due to the higher order corrections, the eigenvalues of the quantum mechanical Hamiltonian operator will in general have lower degeneracy than the eigenvalues of the principal symbol H_0. One might hope to find a finer decomposition of the vector bundle $V[[\hbar]]$, so that one ends up with one dimensional subbundles and ''formal $U(1)$-connections.'' However, this is not possible in general, not even at order $O(\hbar^2)$: The connection constructed in the previous theorem has in general a $U(m_1)\times\cdots\times U(m_r)$ holonomy and cannot be replaced by a connection with a $U(1)\times\cdots\times U(1)$ holonomy without destroying the block-diagonality of the corrected symbol of the Hamiltonian.

To understand this fact better, we first consider the case of a two-dimensional phase space as base space.[20] In this special case it is always possible to reduce the problem to one-dimensional subbundles, at least in a tubular neighborhood of any energy level set $\lambda_\alpha^{-1}(E)$ for a noncritical value E:

To show the existence of such a decomposition and a compatible connection we start with a connection $\tilde{\boldsymbol{\nabla}}^\infty$ for which the Hamiltonian is scalar on each block, as in theorem 7. As the connection to be defined will respect the block structure, we may restrict ourselves to one block α and only need to consider the corresponding eigenvalue function λ_α.

If $\lambda_\alpha^{-1}(E)$ is diffeomorphic to $\mathbb{R}$, we may choose a transversal manifold S at some point $p\in\lambda_\alpha^{-1}(E)$. Choosing an arbitrary decomposition into one-dimensional subbundles on S, we may define a decomposition on a whole tubular neighborhood of $\lambda_\alpha^{-1}(E)$ by parallel transport along the Hamiltonian flow of λ_α. Denoting by $\tilde{\pi}_k$ the corresponding one-dimensional projections, we define a new connection $\hat{\boldsymbol{\nabla}}$ as $\Sigma_k \pi_k\circ\tilde{\boldsymbol{\nabla}}^\infty\circ\pi_k$. Since we have a set of trivializing sections covariantly constant along the Hamiltonian flow both for $\tilde{\boldsymbol{\nabla}}^\infty$ and for $\hat{\boldsymbol{\nabla}}$, it follows that

$$\hat{\boldsymbol{\nabla}}_{X_{\lambda_\alpha}} u=\tilde{\boldsymbol{\nabla}}^\infty_{X_{\lambda_\alpha}} u,$$

for any section u in the the α-th block, and hence the corrected symbol is not modified at order $\hbar$ by going over to the new connection. Hence, the Hamiltonian is still scalar [at least up to order $O(\hbar^2)$], and we have found the connection and decomposition we are looking for up to the same order. This procedure can be iterated again, and we can thus reduce the problem to a purely scalar one.

If $\lambda_\alpha^{-1}(E)$ is diffeomorphic to S^1, then the situation is a bit more involved: Due to the periodicity of the flow, we cannot simply define a decomposition into subbundles on a complete tubular neighborhood of the level set by parallel transport of an arbitrary decomposition on a transversal submanifold S. Instead, defining $U(p)$ as the holonomy for $\tilde{\nabla}^\infty$ along one period of the Hamiltonian flow line starting at the point $p \in S$, we choose the complex eigenspaces of $U(p)$ (which always exist as U is unitary) on S and parallel transport these along the Hamiltonian flow of λ_α. (If the eigenspaces of U are not one-dimensional then we have to choose one-dimensional subspaces in a smooth way along S.) We may define $\tilde{\nabla}$ as before and get a connection with the desired properties.

We note that we can even modify the connection further by adding a $u(1)\times u(1)$ valued one-form to the connection form such that the holonomy along the Hamiltonian flow vanishes and the symbol of the Hamiltonian is modified without destroying diagonality: The corresponding phase is moved in this way from the connection to the symbol.

These observations explain the special role of two-dimensional phase spaces for a multicomponent WKB approximation. Now, an analogous construction obviously does not work generally if the underlying phase space is more than two-dimensional: Even if the Hamiltonian system defined by λ_α is integrable, the flow will generically be only quasiperiodic. This does not only imply that the above construction does not work, but even that such a decomposition into one-dimensional subbundles with a compatible connection and scalar symbols do not exist.

Namely, assume such a decomposition would exist. Let γ denote the integral curve of X_{λ_α} through a fixed point p. Then (pulling back the respective bundles with the inclusion map $i:\gamma \to M$) we first claim that the decomposition of $i^*V[[\hbar]]$ into one-dimensional subbundles is uniquely determined by the decomposition at the point p and the requirement that the corrected symbol is diagonal with respect to the decomposition. To see this, we first note that given any such decomposition with compatible connection, it is determined by the decomposition at P through parallel transport. Hence, if there are several such decompositions, they must correspond to different connections.

However, we may modify any such connection without changing the decomposition by adding a $u(1)\times\cdots\times u(1)$ valued one-form to the connection form. By a proper choice, we may achieve that the corrected symbol on the α-th block is simply λ_α. Now assume that we are given two different decompositions with corresponding (properly modified) connections which coincide at the point p. Then, as the corrected symbols simply is λ_α in both cases, the difference of the connection forms of the two connections (modified as above) must vanish on X_{λ_α}, hence parallel transport along γ coincides for both connections. As the decomposition has to be invariant under parallel transport, the two decomposition must coincide over γ.

Hence, there is at most one such decomposition, and it is determined by parallel transport with the connection in theorem 7. Now, due to the quasiperiodicity, the curve γ comes arbitrarily close to p for certain arbitrarily large times T_i. However, there will generically be no one-dimensional subspace W_p in the fiber V_p over p which, when parallel transported along γ, will be in a suitable sense close to W_p for all the times T_i. Hence, there cannot exist a continuous decomposition into one-dimensional subspaces with the desired properties which is defined in a whole neighborhood of γ.

The difference to the situation in perturbation theory comes from the Berry-phase term. Hence, if one tries to completely diagonalize the corrected symbol, one has to solve a differential equation along the flow of λ_α, and not just simply to diagonalize a matrix. Thus, although it is always possible to completely diagonalize the symbol locally, this is not possible in general globally.

[1] V. P. Maslov and M. V. Fedoriuk, *Semi-Classical Approximation in Quantum Mechanics* (Reidel, Dordrecht, 1981).
[2] S. M. Bates and A. Weinstein, ''Lectures on the geometry of quantization,'' Berkeley Mathematics Lecture Notes, Vol. 8 (1995), Berkeley, CA, 1994.

[3] J. Duistermaat, ''Fourier integral operators,'' Courant Institute, NYU, New York, 1973.
[4] L. Hörmander, ''Fourier integral operators I,'' Acta Math. **127**, 79–183 (1971).
[5] N. Dencker, ''On the propagation of polarization sets for systems of real principal type,'' J. Funct. Anal. **46**, 351–372 (1982).
[6] M. Klein, A. Martinez, and X. P. Wang, ''On the Born-Oppenheimer approximation of wave operators in molecular scattering theory,'' Commun. Math. Phys. **152**, 73–95 (1993).
[7] R. G. Littlejohn and W. G. Flynn, ''Geometric phases in the asymptotic theory of coupled wave equations,'' Phys. Rev. A **44**, 5239–5256 (1991).
[8] V. Guillemin and S. Sternberg, ''Geometric asymptotics,'' *Math. Surveys* (American Mathematical Society, Providence, 1977), Vol. 14.
[9] M. V. Karasev and V. P. Maslov, ''Operators with general commutation relations and their applications. II. Unitary-nonlinear operator equations,'' J. Sov. Math. **15**, 273–368 (1981). Translation from Itogi Nauki Tekh., Ser. Sovrem. Probl. Math **13**, 144–267 (1979).
[10] M. V. Berry, ''Quantal phase factors accompanying adiabatic changes,'' Proc. R. Soc. London, Ser. A **392**, 45–57 (1984).
[11] K. Yabana and H. Horiuchi, ''Adiabatic viewpoint for the WKB treatment of coupled channel systems,'' Prog. Theor. Phys. **75**, 592–618 (1986).
[12] C. Emmrich and A. Weinstein, ''Geometry of the transport equation in multicomponent WKB approximation,'' Commun. Math. Phys. **176**, 701–711 (1996).
[13] B. Fedosov, ''A simple geometrical construction of deformation quantization,'' J. Diff. Geom. **40**, 213–238 (1994).
[14] B. Fedosov, ''Reduction and eigenstates in deformation quantization,'' in *Pseudo-Differential Calculus and Mathematical Physics*, Advances in Partial Differential Equations (Akademie-Verlag, Berlin, 1994), pp. 227–297.
[15] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, ''Deformation theory and quantization,'' Ann. Phys. **111**, 61–110 (1978); part II: **111**, 111–151 (1978).
[16] M. DeWilde and P. B. A. Lecomte, ''Existence of star-products and of formal deformations of the Poisson Lie Algebra of arbitrary symplectic manifolds,'' Lett. Math. Phys. **7**, 487–496 (1983).
[17] R. Nest and B. Tsygan, ''Algebraic index theorem for families,'' Adv. Math. **113**, 151–205 (1995).
[18] M. Bertelson, M. Cahen, and S. Gutt, ''Equivalence of star products,'' Univ. Libre Bruxelles, Travaux Math. Fascicule **1**, 1–15 (1996).
[19] H. O. Cordes, ''A pseudo-algebra of observables for the Dirac equation,'' Manuscr. Math. **45**, 77–105 (1983).
[20] M-V. Karasev, ''New global asymptotics and anomalies for the problem of quantization of the adiabatic invariant,'' Funct. Anal. Appl. **24**, 104–114 (1990).

QED in external field with space–time uniform invariants: Exact solutions

S. P. Gavrilov,[a] D. M. Gitman,[b] and A. E. Gonçalves[c]
Departamento de Física, CCET, Universidade Federal de Sergipe, 49000-000 Aracaju, SE, Brazil

(Received 21 October 1997; accepted for publication 28 January 1998)

We study exact solutions of Dirac and Klein–Gordon equations and Green functions in d-dimensional QED and in an external electromagnetic field with constant and homogeneous field invariants. The cases of even and odd dimensions are considered separately; they are essentially different. We direct attention to the asymmetry of the quasienergy spectrum, which appears in odd dimensions. The in and out classification of the exact solutions as well as the completeness and orthogonality relations is strictly proven. Different Green functions in the form of sums over the exact solutions are constructed. The Fock–Schwinger proper time integral representations of these Green functions are found. As physical applications, we consider the calculations of different quantum effects related to the vacuum instability in the external field. For example, we present mean values of particles created from the vacuum, the probability of the vacuum remaining a vacuum, the effective action, and the expectation values of the current and energy-momentum tensor. © *1998 American Institute of Physics.* [S0022-2488(98)02607-3]

I. INTRODUCTION

Exact solutions of the relativistic wave equations (Klein–Gordon or Dirac equation) in an external electromagnetic field as well as Green functions of those equations are very important in QED with such a field. Special complete sets of the exact solutions and special kinds of Green functions allow one to calculate different quantum effects, for example, particle scattering, pair creation and so on, in zero order with respect to the radiative corrections, but taking the interaction with the external field into account exactly.[1–3] They may serve also as a basis for the perturbation theory with respect to the radiative interaction, in which the external field is taken into account exactly (so-called generalized Furry picture).[4,5] In four-dimensional space–time the above exact solutions were studied for different configurations of the external field in numerous papers and books (see, for example, Refs. 3 and 6). Among all the configurations of the external electromagnetic fields, which admit the exact solutions in $d=4$, one is especially important due to the fact that it corresponds to a wide class of real physical situations. It is a combination of constant uniform electric and magnetic fields and a plane wave field. The exact solutions can be found if there exists a reference frame in which the constant electric and magnetic fields, and the wave 3-vector are collinear. First, the exact solutions of Klein–Gordon and Dirac equation in such a configuration were found in Ref. 7. Some of the Green functions of scalar field in the same configuration were studied in other papers.[8] Using those exact solutions and Green functions different quantum effects were calculated.[9] Calculation of Green functions of spinor field was not present in the literature. Moreover, lately, field theoretical models in dimensions different from $d=4$ attract attention due to various reasons. One can mention here models in $2+1$ dimensions which probably describe planar physical phenomenon, and models in $d>4$ dimensions in connection with Kaluza–Klein ideas (see, e.g., Ref. 10). That is why in the present article we are going to study exact solutions and Green functions (both scalar and spinor case) in d-dimensional QED

[a]On leave from Tomsk Pedagogical University, 634041 Tomsk, Russia; present electronic mail: gavrilov@sergipe.ufs.br
[b]Instituto de Física, Universidade de São Paulo, Caixa Postal 66318, 05315-970-São Paulo, SP, Brazil; electronic mail: gitman@fmal.if.usp.br
[c]Dept. Física, CCE, Universidade Estadual de Londrina, CP 6001. CEP 86051-990, Londrina, PR, Brasil; electronic mail: goncalve@fisica.uel.br

 © 1998 American Institute of Physics

with an external electromagnetic field, which may be considered as a generalization of the above mentioned special configuration in four dimensions. At the same time we also present new results in four dimensions. A general definition of such a configuration valid for any dimensions of space–time is the following: an external electromagnetic field with constant and homogeneous field invariants. Also, an important motivation to consider in exact solutions and Green functions in such a configuration of the external electromagnetic field is the fact that formally problems in some configurations of an external gravitational field lead to the same kind of equations, so that the former exact solutions can be used in the latter case after some simple identifications. Such a reduction may be done, for example, for De Sitter and FRW metrics,[11,12] for four fermionic models in an external field and so on.

In a sense, the paper can be considered as a continuation and generalization of Ref. 13, where only an electriclike external field was considered. The latter paper can help the reader to better understand the logic of the exact solutions constructed in this more complicated present case. Thus, here we sometimes omit technical details and explanations which can be extracted from the above mentioned previous paper.

The paper is organized as follows: In Sec. II we describe first the external electromagnetic field under consideration. Then, we present the so-called in and out complete sets of exact solutions of Dirac and Klein–Gordon equations in the external field. The cases of even and odd dimensions have to be considered separately; they are essentially different. We attract attention to the asymmetry of the quasienergy spectrum, which appears in odd dimensions. We spend enough time to prove strictly the in and out classification of the exact solutions as well as to prove the completeness and orthogonality relations. Then in Sec. III we construct different Green functions in the form of sums over the exact solutions and present them in the form of contour integrals over the Fock–Schwinger proper time. Among them are the causal Green function, in–in and out–out Green functions, the commutator function and so on. One ought to say that the spinor case is treated first, even in $d=4$ dimensions. As physical applications, we consider in Sec. III the calculations of different quantum effects related to the vacuum instability in the external field. For example, we present mean values of particles created from the vacuum, the probability for the vacuum remaining a vacuum, the effective action, the expectation values of the current and energy-momentum tensor. We calculate the latter quantities by means of both in and out sets of the exact solutions.

II. COMPLETE SETS OF EXACT SOLUTIONS

The Dirac equation in an external electromagnetic field with potentials $A_\mu(x)$ in d dimensions has the form ($\hbar=c=1$)

$$(\mathscr{P}_\mu\gamma^\mu-m)\psi(x)=0, \quad \mathscr{P}_\mu=i\partial_\mu-qA_\mu(x), \tag{1}$$

where $\psi(x)$ is a $2^{[d/2]}$-component column, γ^μ are γ-matrices in d dimensions,[14]

$$[\gamma^\mu,\gamma^\nu]_+=2\eta^{\mu\nu}, \quad \eta^{\mu\nu}=\mathrm{diag}(\underbrace{1,-1,-1,...}_{d}), \quad d=D+1,$$

and $x=(x^\mu)=(x^0,\mathbf{x})$, $\mathbf{x}=(x^i)$, $\mu=0,1,...,D$, $i=1,...,D$. The time-independent scalar product of the solutions of the Dirac equation may be chosen in the conventional form

$$(\psi,\psi')=\int \bar{\psi}(x)\gamma^0\psi'(x)d\mathbf{x}. \tag{2}$$

As usual, it is convenient to present $\psi(x)$ in the form

$$\psi(x)=(\mathscr{P}_\mu\gamma^\mu+m)\phi(x). \tag{3}$$

Then the functions ϕ have to obey the squared Dirac equation in d dimensions,

$$\left(\mathscr{P}^2-m^2-\frac{q}{2}\sigma^{\mu\nu}\mathscr{F}_{\mu\nu}\right)\phi(x)=0, \quad \mathscr{F}_{\mu\nu}=\partial_\mu A_\nu(x)-\partial_\nu A_\mu(x), \quad \sigma^{\mu\nu}=\frac{i}{2}[\gamma^\mu,\gamma^\nu]. \tag{4}$$

To construct the above mentioned generalized Furry picture in QED with an external field, one has to find special sets of exact solutions of Eq. (1), namely, two complete and orthonormal sets of solution: $\{{}_{\pm}\psi_{\{n\}}(x)\}$ which describes particles $(+)$ and antiparticles $(-)$ in the initial time instant $(x^0\to-\infty)$, and $\{{}^{\pm}\psi_{\{n\}}(x)\}$ which describes particles $(+)$ and antiparticles $(-)$ in the final time instant $(x^0\to+\infty)$. According to the general approach,[5] such solutions obey the following asymptotic conditions:

$$H_{\text{o.p.}}(x^0)\,{}_{\zeta}\psi_{\{n\}}(x)={}_{\zeta}\epsilon\,{}_{\zeta}\psi_{\{n\}}(x),\quad \operatorname{sgn}{}_{\zeta}\epsilon=\zeta,\quad x^0\to-\infty,$$
$$H_{\text{o.p.}}(x^0)\,{}^{\zeta}\psi_{\{l\}}(x)={}^{\zeta}\epsilon\,{}^{\zeta}\psi_{\{l\}}(x),\quad \operatorname{sgn}{}^{\zeta}\epsilon=\zeta,\quad x^0\to+\infty, \tag{5}$$

where $\zeta,\{n\}$ and $\zeta,\{l\}$ are complete sets of quantum numbers which characterize solutions ${}_{\zeta}\psi_{\{n\}}(x)$ and ${}^{\zeta}\psi_{\{l\}}(x)$, respectively, $H_{\text{o.p.}}=\gamma^0(m-\gamma^i\mathscr{P}_i)$ is a one-particle Dirac Hamiltonian in convenient external field gauge $A_0(x)=0$; ${}^{+}\epsilon$, ${}_{+}\epsilon$ are particle quasienergies and $|{}^{-}\epsilon|$ and $|{}_{-}\epsilon|$ are antiparticles quasienergies. All the information about the processes of particle scattering and creation by an external field (in zeroth order with respect to the radiative corrections) can be extracted from the decomposition coefficients (matrices) $G({}_{\zeta}|{}^{\zeta'})$,

$${}^{\zeta}\psi(x)={}_{+}\psi(x)G({}_{+}|{}^{\zeta})+{}_{-}\psi(x)G({}_{-}|{}^{\zeta}). \tag{6}$$

The matrices $G({}_{\zeta}|{}^{\zeta'})$ obey the following relations:

$$G({}_{\zeta}|{}^{+})G({}_{\zeta}|{}^{+})^{\dagger}+G({}_{\zeta}|{}^{-})G({}_{\zeta}|{}^{-})^{\dagger}=\mathbf{I},$$
$$G({}_{+}|{}^{+})G({}_{-}|{}^{+})^{\dagger}+G({}_{+}|{}^{-})G({}_{-}|{}^{-})^{\dagger}=0, \tag{7}$$

where $\mathbf{I}$ is the identity matrix.

Let us describe the external electromagnetic field in which we are going to construct the exact solutions. Such a field is a generalization of the corresponding field in $d=4$, which is a combination of the constant uniform field with the plane wave field and which admits there exact solutions. One can describe such a field in $d=4$ in arbitrary reference frame saying that both its field invariants do not depend on space–time coordinates. In $d>4$ there exist, generally speaking, more independent field invariants. Besides $I_1=1/2\mathscr{F}_{\mu\nu}\mathscr{F}^{\mu\nu}$, those are all possible independent invariant combinations which may be constructed from the field tensor $\mathscr{F}_{\mu\nu}$ and Levi-Civita tensor. It is easy to see that there exist $[d/2]$ such invariants. We define the external field under consideration in arbitrary dimensions in the same manner; all its field invariants have to be constant and uniform. One can see that such a field is a combination of a constant uniform field $F_{\mu\nu}$ and a plane-wave field $f_{\mu\nu}(nx)$

$$\mathscr{F}_{\mu\nu}=F_{\mu\nu}+f_{\mu\nu}(nx), \tag{8}$$

where n_μ is a real isotropic vector, $n_\mu n^\mu=0$. It is an eigenvector of the tensor $F_{\mu\nu}$, $F_{\mu\nu}n^\nu=\mathscr{E}n_\mu$, and $f_{\mu\nu}(nx)$ is a transverse field with respect to n_μ, $n^\mu f_{\mu\nu}(nx)=f_{\mu\nu}(nx)n^\nu=0$. If all of the invariants are equal to zero (it is only possible for $d>2$), only a plane-wave field configuration is possible. If some of the invariants are not equal to zero, then a constant, uniform component $F_{\mu\nu}$, is not zero. The field (8) is free. Such a field is of special interest due to the fact that QED with a free classical field can be treated as exact QED (without external fields) but with some special (coherent) initial photon states.[15,3]

If the eigenvalue $\mathscr{E}$ is not zero, the external field violates the vacuum stability (creating particles). One can find an inertial frame where the matrix $\mathscr{F}_{\mu\nu}$ has a simple form

$$F_{\mu\nu}=F^{\perp}_{\mu\nu}+F^{\|}_{\mu\nu},$$

$$F^{\|}_{\mu\nu}=E(\delta^0_\mu\delta^D_\nu-\delta^D_\mu\delta^0_\nu),$$

$$f_{\mu\nu}(nx)=\sum_{k=1}^{D-1}(n_\mu\delta^k_\nu-n_\nu\delta^k_\mu)\dot f_k(x_-),\quad \dot f_k(x_-)=\frac{df_k(x_-)}{dx_-},\tag{9}$$

$$n_\mu=\delta^0_\mu-\delta^D_\mu,\quad x_-=nx=x^0-x^D.$$

If $d>2$ is even then

$$F^{\perp}_{\mu\nu}=\sum_{j=1}^{(d-2)/2}H_j(\delta^{2j}_\mu\delta^{2j-1}_\nu-\delta^{2j}_\nu\delta^{2j-1}_\mu),$$

and if $d=2$ the fields $F^{\perp}_{\mu\nu}$ and $f_{\mu\nu}(nx)$ are absent. In the case of d is odd, and $\mathscr{E}\neq 0$,

$$F^{\perp}_{\mu\nu}=\sum_{j=1}^{(d-3)/2}H_j(\delta^{2j}_\mu\delta^{2j-1}_\nu-\delta^{2j}_\nu\delta^{2j-1}_\mu)\quad\text{if } d>3,\ \text{and } F^{\perp}_{\mu\nu}=0\quad\text{if } d=3.$$

The same formula is valued if $\mathscr{E}=0$, but at least one of the imaginary eigenvalues of $F_{\mu\nu}$ is zero. Here $f_k(x_-)$ are arbitrary functions of x_-. If $\mathscr{E}=0$, and all imaginary eigenvalues of $F_{\mu\nu}$ are not equal to zero (all field invariants are not equal to zero) we have

$$E=0,\quad F^{\perp}_{\mu\nu}=\sum_{j=1}^{(d-1)/2}H_j(\delta^{2j}_\mu\delta^{2j-1}_\nu-\delta^{2j}_\nu\delta^{2j-1}_\mu),\quad f_k(x_-)=0\ \text{ for all }\ k.$$

In the reference frame under consideration $\mathscr{E}=E$.

To realize the external electromagnetic field of the above form, we select the following potentials:

$$A_\mu(x)=A^E_\mu(x)+A^H_\mu(x)+f_\mu(x_-),$$

$$A^E_\mu(x)=Ex^0\delta^D_\mu,\quad A^H_i=-H_jx_{i+1}\delta_{i,2j-1},\quad j=1,...,[d/2]-1,\quad i=1,...,D-1,\tag{10}$$

$$f_\mu(x_-)=0\quad\text{if }\ \mu=0,D.$$

Below we present linearly independent sets of solutions of the squared Dirac equation, which correspond to the particles in the initial time instance and to the antiparticles in the final time instant:

$$\overset{-}{+}\phi_{p_-,n,\xi,r}(x)=\overset{-}{+}\phi_{p_-,n,r}(x^0,x^D)\overset{-}{+}u_{\xi,r}(x_-)\phi_{n,r}(x_\perp),$$

$$\phi_{n,r}(x_\perp)=\phi_{p_1,n_1}(x_1,x_2)\phi_{p_3,n_2}(x_3,x_4)\dots\phi_{p_{d-3},n_{(d-2)/2}}(x_{D-2},x_{D-1}),\quad\text{if } d \text{ is even},$$

$$\phi_{n,r}(x_\perp)=\phi_{p_1,n_1}(x_1,x_2)\phi_{p_3,n_2}(x_3,x_4)\dots\phi_{p_{d-2},n_{(d-3)/2}}(x_{D-3},x_{D-2})(2\pi)^{-1/2}$$
$$\times\exp(-ip_{D-1}x^{D-1})\quad\text{if } d \text{ is odd and } H_{(d-1)/2}=0,$$

$$\overset{-}{+}\phi_{p_-,n,r}(x^0,x^D)=(4\pi)^{-1/2}$$
$$\times\exp\left\{\frac{i}{2}(qE(x^2_-/2-x^2_D)-p_-x_++\lambda\ \ln(\mp\tilde\pi_-))+\overset{-}{+}J(x_-)-\overset{-}{+}K^\mu(x_-)\pi_{\perp\mu}\right\},\tag{11}$$

$$
{}^{-}_{+}J(x_-)=-\frac{1}{2qE}\int_{\mp\infty}^{\pi_-}qf\left(\frac{p_--\tau}{qE}\right)\left[qf\left(\frac{p_--\tau}{qE}\right)+qF{}^{-}_{+}K\left(\frac{p_--\tau}{qE}\right)\right]\tau^{-1}d\tau,
$$

$$
{}^{-}_{+}K(x_-)=-\frac{1}{qE}\int_{\mp\infty}^{\pi_-}\exp\left\{-\frac{F}{E}\ln\frac{\pi_-}{\tau}\right\}qf\left(\frac{p_--\tau}{qE}\right)\tau^{-1}d\tau,
$$

$$
{}^{-}_{+}u_{\xi,r}(x_-)=\mp\left(\frac{1}{2\pi_-}(1+\gamma^0\gamma^D)+\frac{1}{2}(1-\gamma^0\gamma^D)+\frac{1}{2\pi_-}(\gamma^0-\gamma^D)\gamma qf(x_-)\right)v_{\xi,r},
$$

$$
x_+=x^0+x^D,\quad \tilde{\pi}_-=\pi_-/\sqrt{qE},
$$

where p_-, $n=(n_1,n_2,\ldots,n_{[d/2]-1};p_1,p_3,\ldots,p_{2[(d-1)/2]-1})$, ξ and $r=(r_1,r_2,\ldots,r_{[d/2]-1})$ is a complete set of quantum numbers. Among them p_-, p_j are momenta of the continuous spectrum, n_j are integer quantum numbers, $\xi=\pm1$ and $r_j=\pm1$ are spin quantum numbers; the momentum p_- is the eigenvalue of the operator $2i(\partial/\partial x_+)$; if $qE>0$ is chosen then the signs $\mp$ assigned to the functions ${}^{-}_{+}\phi_{p_-,n,\xi,r}(x)$ are matched with those of the kinetic momentum $\pi_-=p_--qEx_-$ at $x_-\to\pm\infty$, and

$$
\begin{gathered}
x^\mu_\perp=0\quad\text{if}\quad\mu=0,D,\quad x^\mu_\perp=x^\mu\quad\text{if}\quad\mu=1,\ldots,D-1,\\
\pi_{\perp\mu}=0\quad\text{if}\quad\mu=0,D,\quad \pi_{\perp\mu}=i\frac{\partial}{\partial x^\mu}-qA^H_\mu(x)\quad\text{if}\quad\mu=1,\ldots,D-1,\\
qE\lambda=m^2+\sum_{j=1}^{[d/2]-1}\omega_j+\omega_0,\quad \omega_0=\begin{cases}0, & d\ \text{is even},\\ p^2_{d-2}, & d\ \text{is odd},\end{cases}\\
\omega_j=\begin{cases}|qH_j|(2n_j+1-r_j), & H_j\neq0,\\ p^2_{2j-1}+p^2_{2j}, & H_j=0.\end{cases}
\end{gathered}
\tag{12}
$$

Each function $\phi_{p_{2j-1},n_j}(x_{2j-1},x_{2j})$ obeys the following equations:

$$
(\pi^2_{\perp 2j-1}+\pi^2_{\perp 2j}-\omega_j)\phi_{p_{2j-1},n_j}(x_{2j-1},x_{2j})=0,
$$

$$
(\pi_{\perp 2j-1}-p_{2j-1})\phi_{p_{2j-1},n_j}(x_{2j-1},x_{2j})=0.
$$

If $H_j\neq0$, a solution of these equations is

$$
\begin{aligned}
\phi_{p_{2j-1},n_j}(x_{2j-1},x_{2j})=&\left(\frac{\sqrt{|qH_j|}}{2^{n_j+1}\pi^{3/2}n_j!}\right)^{1/2}\exp\left\{-ip_{2j-1}x^{2j-1}-\frac{|qH_j|}{2}\left(x^{2j}+\frac{p^{2j-1}}{qH_j}\right)^2\right\}\\
&\times\mathscr{H}_{n_j}\left[\sqrt{|qH_j|}\left(x^{2j}+\frac{p^{2j-1}}{qH_j}\right)\right],
\end{aligned}
$$

where $\mathscr{H}_{n_j}(x)$ is the Hermite polynomial with integer $n_j=0,1,\cdots$. If $H_j=0$, the discrete quantum numbers n_j have to be replaced by the momenta p_{2j}, and the corresponding function has the form

$$
\phi_{p_{2j-1},p_{2j}}(x_{2j-1},x_{2j})=(2\pi)^{-1}\exp\{-i(p_{2j-1}x^{2j-1}+p_{2j}x^{2j})\}.
$$

The symbol $\ln(\mp\tilde{\pi}_-)$ means the principal branch of the logarithm, $\ln(\mp\tilde{\pi}_-)=\ln|\tilde{\pi}_-|+i\pi\Theta(\pm\tilde{\pi}_-)$, while the integration paths in the τ-plane, as well as the arguments π_-, are shown in Figs. 1 and 2 for the functions ${}_{+}K(x_-)$, ${}_{+}J(x_-)$ and ${}^{-}K(x_-)$, ${}^{-}J(x_-)$, respectively.

One assumes the functions $f_k(x_-)$ obey the requirement that ${}^{-}_{+}K(x_-)$ and ${}^{-}_{+}J(x_-)$ are analytic functions, and in particular, the functions $f_k(x_-)$ vanish at $x_-\to\pm\infty$ quite rapidly.

The solution has a different form if d is odd, $\mathscr{E}=0$, and all the imaginary eigenvalues of $F_{\mu\nu}$ are not equal to zero (in this case a plane-wave field is absent),

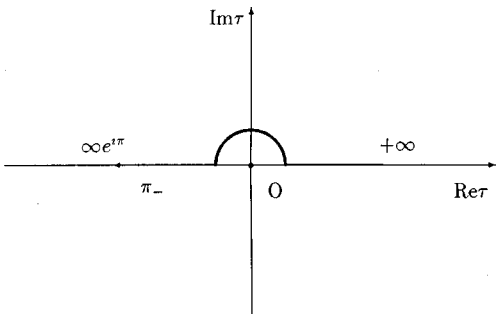

FIG. 1. Path of integration in ${}_+K(x_-)$ and ${}_+J(x_-)$ integrals.

$$
\begin{aligned}
&{}^-_+\phi_{n,\xi,r}(x)={}^-_+\phi_{n,r}(x^0)\,\phi_{n,r}(\mathbf{x})\,v_{\xi,r}\,,\\
&\phi_{n,r}(\mathbf{x})=\phi_{p_1,n_1}(x_1,x_2)\,\phi_{p_3,n_2}(x_3,x_4)\ldots\phi_{p_{d-2},n_{(d-1)/2}}(x_{D-1},x_D),\\
&{}^-_+\phi_{n,r}(x^0)=c\,\exp(\pm i|{}_\mp\epsilon_{nr}|x^0),\quad |{}_\mp\epsilon_{nr}|=\sqrt{m^2+\sum_{j=1}^{(d-1)/2}\omega_j},
\end{aligned}
\tag{13}
$$

where c is a normalization constant.

Here $v_{\xi,\tau}$ are some constant orthonormal spinors, $v^\dagger_{\xi,r}v_{\xi,r'}=\delta_{r,r'}$. Equation (4) allows one to subject these spinors to some supplementary conditions,

$$
\Xi_\pm v_{\mp 1,r}=0,\ \ \Xi_\pm=\tfrac{1}{2}(1\pm\gamma^0\gamma^D),\quad \text{rank}\ \ \Xi_\pm=J_{(d)}=2^{[d/2]-1};
\tag{14}
$$

$$
R_j v_{\xi,r}=r_j v_{\xi,r},\quad d\geqslant 4,\quad R_j=i\gamma^{2j-1}\gamma^{2j}.
\tag{15}
$$

If $d=2,3$, the independent quantum number r does not appear.

One can verify the solutions of the Dirac equation with different ξ, namely, $(\mathscr{P}_\mu\gamma^\mu+m)_+\phi_{p_-,n,+1,r}(x)$ and $(\mathscr{P}_\mu\gamma^\gamma+m)_+\phi_{p_-,n,-1,r}(x)$, or $(\mathscr{P}_\mu\gamma^\gamma+m)^-\phi_{p_-,n,+1,r}(x)$, and $(\mathscr{P}_\mu\gamma^\mu+m)^-\phi_{p_-,n,-1,r}(x)$, are linearly dependent for each sign ''+'' or ''−''. It means, in fact, that the spin projections of a particle $(+)$ and an antiparticle $(-)$ can take on only $J_{(d)}$ values. Thus to construct the complete sets, it is sufficient to use only the following solutions:

$$
{}^-_+\psi_{p_-,n,r}(x)=(\mathscr{P}_\mu\gamma^\mu+m)\,{}^-_+\phi_{p_-,n,+1,r}(x).
\tag{16}
$$

In particular, if $d=2,3$, there is only one spin projection, with only one spinor $v_{+1,r}=(1/\sqrt{2})\binom{1}{i}$. To get the case where $E=0$, one has to consider a limit in the solution ${}^-_+\psi_{p_-,n,r}(x)$, such that $-(qE)^{-1}\ln(\mp\tilde{\pi}_-)\to x_-/p_-$, and $p_->0$ for $\{{}_+\psi_{p_-,n,r}(x)\}$, whereas $p_-<0$ for $\{{}^-\psi_{p_-,n,r}(x)\}$.

Using solutions (13) we may construct the Dirac spinors in a slightly different way,

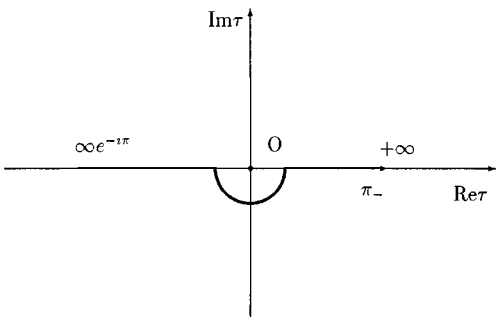

FIG. 2. Path of integration in ${}^-K(x_-)$ and ${}^-J(x_-)$ integrals.

$$
{}^{-}_{+}\psi_{n,\bar{r}}(x)=(\mathscr{P}_{\mu}\gamma^{\mu}+m){}^{-}_{+}\phi_{n,\bar{r}}(x),
$$

$$
{}^{-}_{+}\phi_{n,\bar{r}}(x)=({}^{-}_{+}\phi_{n,+1,r}(x)+\mathrm{sgn}(qH_{(d-1)/2}){}^{-}_{+}\phi_{n,-1,r}(x)), \tag{17}
$$

$$
\bar{r}=(r_1,r_2,\ldots,r_{(d-3)/2},\quad \mathrm{sgn}(qH_{(d-1)/2})),
$$

where

$$
R_j{}^{-}_{+}\phi_{n,\bar{r}}(x)=\bar{r}_j{}^{-}_{+}\phi_{n,\bar{r}}(x),\quad j=1,2,\ldots,(d-1)/2,
$$

and ${}^{-}_{+}\phi_{n,\bar{r}}(x)$ are eigenvectors of the time-independent operator $m^2-\mathscr{P}^i\mathscr{P}_i+(q/2)\sigma^{\mu\nu}F_{\mu\nu}$.

The solutions ${}_{+}\psi_{n,\bar{r}}(x)$ from (17) describe particles $(+)$ and ${}^{-}\psi_{n,\bar{r}}(x)$ describe antiparticles $(-)$ in any time instant. They form a complete and orthonormal set of solutions. There is an interesting asymmetry in energy spectrums of such particles and antiparticles in odd dimensions. If all quantum numbers $n_j=0$, and $r_j=\mathrm{sgn}(qH_j)$, then $\gamma\mathbf{P}{}^{-}_{+}\psi_{n,\bar{r}}(x)=0$, and the ground state of the particle has energy ${}_{+}\epsilon_0=m$, but the energy of the ground state of the antiparticle is dependent of the magnetic field and different: $|{}^{-}\epsilon_0|=\sqrt{m^2+2\min|qH_j|}$.

In the $d=4$ case, the solutions similar to (16) were found first in Ref. 7. In the case $E=0$, they coincide with ones in Ref. 16, and if $H=0$ they coincide with well-known Wolkov form.[17] In the case $E=0$, the solutions form a complete and orthonormal set, moreover ${}_{+}\psi_{p_-,n,r}(x)$ describe particles and ${}^{-}\psi_{p_-,n,r}(x)$ describe antiparticles in any time instant. It is no longer true if $E\neq 0$. However, solutions (16) can be used to construct new complete sets which can be classified by the correct way.

One can form two complete and orthonormal sets of the solutions: $\{{}_{\pm}\psi_{p_-,n,r}(x)\}$ and ${}^{\pm}\psi_{p_-,n,r}(x)$ using (16) and additional sets

$$
\begin{aligned}
{}^{+}\psi_{p_-,n,r}(x)&=\Theta(\pi_-)({}_{+}\psi(x)G({}_{+}|{}^{+}))_{p_-,n,r},\\
{}_{-}\psi_{p_-,n,r}(x)&=\Theta(-\pi_-)({}^{-}\psi(x)G({}_{-}|{}^{-})^{\dagger})_{p_-,n,r},
\end{aligned} \tag{18}
$$

where $G({}_{\zeta}|{}^{\zeta'})$ obeys relations (7) and, in particular, $G({}_{+}|{}^{-})$ are decomposition coefficients of ${}^{-}\psi_{p_-,n,r}(x)$ solutions in ${}_{+}\psi_{p_-,n,r}(x)$ solutions,

$$
G({}_{+}|{}^{-})_{p_-,n,r,p'_-,n',r'}=({}_{+}\psi_{p_-,n,r},{}^{-}\psi_{p'_-,n',r'}). \tag{19}
$$

The corresponding Klein–Gordon solutions follow from (11) by the replacement $u_{\xi,r}(x_-)=\exp\{-(1/2)\ln(\mp\tilde{\pi}_-)\}$.

The orthonormality of all the solutions can be verified by using the following integral transformations:

$$
{}_{\pm}\psi_{p_-,n,r}(x)=(2\pi qE)^{-1/2}\int_{-\infty}^{+\infty}M^*(p_D,p_-){}_{\pm}\psi_{p_D,n,r}(x)dp_D, \tag{20}
$$

$$
{}_{\pm}\psi_{p_D,n,r}(x)=(2\pi qE)^{-1/2}\int_{-\infty}^{+\infty}M(p_D,p_-){}_{\pm}\psi_{p_-,n,r}(x)dp_-,
$$

$$
M(p_D,p_-)=\exp\left\{-\frac{i}{4qE}\left((p_--2p_D)^2-2(p_D)^2\right)\right\}, \tag{21}
$$

$$
\int_{-\infty}^{+\infty}M^*(p_D,p'_-)M(p_D,p_-)dp_D=2\pi qE\,\delta(p_--p'_-).
$$

The same relation is valid for functions with $(\pm)$ indices above.

The saddle points $\pi_-=-2(qEx_0-p_D)$ give the main contribution to integrals (21) with ${}^{-}_{+}\psi_{p_-,n,r}(x)$ functions at $x_0\to\pm\infty$ (this was first found in Ref. 18 for the $d=4$ case without a

magnetic field). Since the plane wave vanishes in ${}^{-}_{+}\psi_{p_-,n,r,}(x)$ at $\pi_- \to \pm\infty$, relations (21) reduce at $x_0 \to \pm\infty$ to the space–time uniform field case and can be presented by use of formulas from Ref. 19 as follows:

$$
\begin{gathered}
{}^{-}_{+}\psi_{p_D,n,r}(x) = (\mathscr{P}_\mu \gamma^\mu + m)\,{}^{-}_{+}\phi_{p_D,n,r}(x), \quad x_0 \to \pm\infty, \\
{}^{-}_{+}\phi_{p_D,n,r}(x) = {}^{-}_{+}\phi_{n,r}(x^0)\exp(-ip_D x^D)\,\phi_{n,r}(x_\perp)v_{+1,r}, \\
{}^{-}_{+}\phi_{n,r}(x^0) = CD_{\nu-1}[\pm(1-i)\xi], \\
\xi = (qEx^0 - p_D)/\sqrt{qE}, \quad \nu = i\lambda/2, \quad C = (4\pi qE)^{-1/2}\exp\{(-\pi/2 + i\,\ln 2)\lambda/4\}(-i),
\end{gathered}
\tag{22}
$$

where the function $\phi_{n,r}(x_\perp)$ was defined in (11), and, $D_\nu(z)$ is the Weber parabolic cylinder function.[20] Such solutions were discussed in Ref. 13. One can now verify the orthonormality relations of the sets $\{{}_{+}\psi_{p_D,n,r}(x)\}$ and $\{{}^{-}\psi_{p_D,n,r}(x)\}$. Using transformation (20) one gets the orthonormality relations of sets $\{{}_{+}\psi_{p_-,n,r}(x)\}$ and $\{{}^{-}\psi_{p_-,n,r}(x)\}$ as well. Using the explicit forms of solutions, (16), (18), and relations (7) one can derive the following relations:

$$
\begin{gathered}
{}^{+}\psi(x)G({}_{+}|{}^{+})^\dagger = {}_{+}\psi(x) - {}^{-}\psi(x)G({}_{+}|{}^{-})^\dagger, \\
{}_{-}\psi(x)G({}_{-}|{}^{-}) = {}^{-}\psi(x) - {}_{+}\psi(x)G({}_{+}|{}^{-}).
\end{gathered}
\tag{23}
$$

By means of the latter one can get the orthonormality relations for both sets of the solutions,

$$
\begin{gathered}
({}_{\zeta}\psi_{p_-,n,r},\ {}_{\zeta'}\psi_{p'_-,n',r'}) = \delta_{\zeta\zeta'}\delta_{rr'}\delta_{nn'}\delta(p_- - p'_-), \\
({}^{\zeta}\psi_{p_-,n,r},\ {}^{\zeta'}\psi_{p'_-,n',r'}) = \delta_{\zeta\zeta'}\delta_{rr'}\delta_{nn'}\delta(p_- - p'_-),
\end{gathered}
\tag{24}
$$

where $\zeta, \zeta' = \pm$, $\delta_{nn'}$ is the Kronecker symbol for the discrete spectrum and the δ-function for the continuous one. Here $r = r' = +1$ if $d = 2,3$.

To solve a problem of the $(\pm)$ classification of the solutions, one needs to study their asymptotic behavior at $x_0 \to \pm\infty$. From the asymptotic forms (22) it follows[13] that the asymptotic of the quasienergies of these solutions, ${}_{+}\epsilon = -qEx^0$, is positive and ${}^{-}\epsilon = -qEx^0$ is negative. Thus the solutions ${}_{+}\psi_{p_D,n,r}(x)$ describe particles at $x_0 \to -\infty$, and the solutions ${}^{-}\psi_{p_D,n,r}(x)$ describe antiparticles at $x_0 \to +\infty$. Since the solutions ${}_{-}\psi_{p_-,n,r}(x)$ are orthogonal to ${}_{+}\psi_{p_-,n,r}(x)$, they describe antiparticles at $x_0 \to -\infty$, and solutions ${}^{+}\psi_{p_-,n,r}(x)$ describe particles at $x_0 \to +\infty$ since they are orthogonal to ${}^{-}\psi_{p_-,n,r}(x)$. One can verify this by taking into account that the main contribution to integrals (21) at $x_0 \to \pm\infty$ for ${}^{+}_{-}\psi_{p,r}(x)$ is given by point $\pi_- = 0$. In this limit the contribution of the plane wave does not depend on x and the results of transformation (21) are proportional to a superposition of the solutions in a constant uniform field,[20]

$$
\begin{gathered}
{}^{+}_{-}\psi_{p_D,n,r}(x) = \sum_{n,r} a_{n'r'}(\mathscr{P}_\mu \gamma^\mu + m)\,{}^{+}_{-}\phi_{p_D,n',r'}(x), \quad x_0 \to \pm\infty, \\
{}^{+}_{-}\phi_{p_D,n,r}(x) = {}^{+}_{-}\phi_{n,r}(x^0)\exp(-ip_D x^D)\,\phi_{n,r}(x_\perp)v_{+1r}, \\
{}^{+}_{-}\phi_{n,r}(x^0) = CD_{-\nu}[\pm(1+i)\xi],
\end{gathered}
\tag{25}
$$

where a_{nr} are some coefficients dependent on the plane-wave form. From the asymptotic representations (25), one can see[13] that the asymptotic of the quasienergies of these solutions ${}^{+}\epsilon = qEx^0$ is positive and ${}_{-}\epsilon = qEx^0$ is negative.

Both sets $\{{}_{\pm}\psi_{p_-,n,r}(x)\}$ and $\{{}^{\pm}\psi_{p_-,n,r}(x)\}$ are orthonormal and complete at any time instant. The form of the commutation function, which will be present in the next section, can serve as

direct proof of the last statement. In $d=4$ such complete sets of the solutions, were first found in Ref. 8. If the plane wave is absent and $E=0$, one can get the solutions with defined energies from ${}^{-}_{+}\psi_{p_D,n,r}(x)$ (22) considering the next limit in ${}^{-}_{+}\phi_{n,r}(x^0)$:

$$ {}^{-}_{+}\phi_{n,r}(x^0)\to[{}_{\mp}\epsilon_{nr}({}_{\mp}\epsilon_{nr}+p_D)]^{-1}\ \exp(\pm i|{}_{\mp}\epsilon_{nr}|x^0), $$

$$ |{}_{\mp}\epsilon_{nr}|=\sqrt{m^2+\sum_{j=1}^{[(d-1)/2]}\omega_j+\omega_0'},\quad \omega_0'=\begin{cases}p_D^2, & d \text{ is even},\\ 0, & d \text{ is odd},\end{cases} $$

where ω_j is defined in (12). In this case one can choose ${}^{+}\psi_{p_D,n,r}(x)={}_{+}\psi_{p_D,n,r}(x)$ and ${}_{-}\psi_{p_D,n,r}(x)={}^{-}\psi_{p_D,n,r}(x)$.

Let us select some important properties of the solution. One can see that the matrix elements $G({}_{\zeta}|{}^{\zeta'})$ are diagonal with respect to continuous quantum numbers and spin quantum numbers,

$$ G({}_{\zeta}|{}^{\zeta'})_{p_-,n,r,p'_-,n',r'}=\delta_{rr'}\delta(p_--p'_-)\delta(p_1-p'_1)\cdots\delta(p_{2[(d-1)/2]-1}-p'_{2[(d-1)/2]-1})g({}_{\zeta}|{}^{\zeta'})_{nn'}. \tag{26} $$

The solutions ${}^{-}_{+}\psi_{p_-,n,r}(x)$ satisfy the orthonormality conditions on the null-plane,

$$ \int\ {}^{-}_{+}\overline{\psi}_{p_-,n,r}(x)\Xi_-\ {}^{-}_{+}\psi_{p'_-,n',r'}(x)dx_+\ dx^1\cdots dx^{D-1}=\delta_{rr'}\delta_{nn'}\delta(p_--p'_-),\quad \mp\pi_->0. \tag{27} $$

Since ${}^{+}\psi_{p_-,n,r}(x)=0$ for $\pi_-<0$ and relations (23) take place, one gets the following representations for $G({}_{+}|{}^{-})G({}_{+}|{}^{-})^{\dagger}$ and $G({}_{+}|{}^{-})G^{\dagger}G({}_{+}|{}^{-})$ matrices:

$$ \begin{aligned}(G({}_{+}|{}^{-})G({}_{+}|{}^{-})^{\dagger})_{p_-,n,r,p'_-,n',r'}&=\int\ {}_{+}\overline{\psi}_{p_-,n,r}(x)\Xi_-\ {}_{+}\psi_{p'_-,n',r'}(x)dx_+dx^1\cdots dx^{D-1},\\ &\pi_-<0,\\ (G({}_{+}|{}^{-})^{\dagger}G({}_{+}|{}^{-}))_{p_-,n,r,p'_-,n',r'}&=\int\ {}^{-}\overline{\psi}_{p_-,n,r}(x)\Xi_-\ {}^{-}\psi_{p'_-,n',r'}(x)dx_+dx^1\cdots dx^{D-1},\\ &\pi_->0.\end{aligned} \tag{28} $$

Thus one can calculate ${}_{+}\mathscr{D}=g({}_{+}|{}^{-})g({}_{+}|{}^{-})^{\dagger}$ and ${}^{-}\mathscr{D}=g({}_{+}|{}^{-})^{\dagger}g({}_{+}|{}^{-})$ matrices using the following integrals with the solutions of the squared Dirac equation:

$$ {}^{-}_{+}\mathscr{D}_{nn'}=\int\ {}^{-}_{+}\phi^{\dagger}_{p_-,n,+1,r}(x){}^{-}_{+}\phi_{p_-,n',+1,r}(x)dx^2\ dx^4\cdots dx^{2[(d-1)/2]},\quad \mp\pi_-<0. \tag{29} $$

III. GREEN FUNCTIONS

The perturbation theory with respect to the radiative interaction for the matrix elements of the processes also has the usual Feynman structure in an external field creating pairs.[5,21,3] The Feynman diagrams have to be calculated by means of the causal propagator

$$ S^c(x,x')=c_v^{-1}i\langle 0,out|T\psi(x)\overline{\psi}(x')|0,in\rangle,\quad c_v=\langle 0,out|0,in\rangle, \tag{30} $$

where $\psi(x)$ is quantum spinor field in the generalized Furry representation, satisfying the Dirac equation (1), $|0,in\rangle$ and $|0,out\rangle$ are the initial and the final vacuum in this representation, and c_v is the vacuum to vacuum transition amplitude. The propagator $S^c(x,x')$ obeys the equation

$$ (\mathscr{P}_\mu\gamma^\mu-m)S^c(x,x')=-\delta^{(d)}(x-x'), \tag{31} $$

and is a Green function of the equation. Another important singular function is the commutation function

$$S(x,x')=i[\psi(x),\bar{\psi}(x')]_+ . \tag{32}$$

It obeys the homogeneous Dirac equation (1) and the initial condition

$$S(x,x')|_{x_0=x_0'}=i\gamma^0\delta(\mathbf{x}-\mathbf{x}'). \tag{33}$$

The commutation function $S(x,x')$ is, at the same time, the propagation function of the Dirac equation, i.e., it connects solutions of the equation in two different time instants.

QED with unstable vacuum has a number of peculiarities. Thus, for instance, in the calculation of the expectation values and the total probabilities Green functions of different types from (30) appear:[5,21,3]

$$\begin{aligned}
S^c_{in}(x,x')&=i\langle 0,in|T\psi(x)\bar{\psi}(x')|0,in\rangle,\\
S^-_{in}(x,x')&=i\langle 0,in|\psi(x)\bar{\psi}(x')|0,in\rangle,\\
S^+_{in}(x,x')&=i\langle 0,in|\bar{\psi}(x')\psi(x)|0,in\rangle,\\
S^{\bar{c}}_{in}(x,x')&=i\langle 0,in|\psi(x)\bar{\psi}(x')T|0,in\rangle,\\
S^c_{out}(x,x')&=i\langle 0,out|T\psi(x)\bar{\psi}(x')|0,out\rangle,
\end{aligned} \tag{34}$$

where the symbol of the T-product acts on both sides: It orders the field operators to the right and antiorders them to the left. The function $S^c_{\rm in}(x,x')$, $S^c_{\rm out}(x,x')$ obeys Eq. (31), $S^{\mp}(x,x')$ satisfies Eq. (1), and $S^{\bar{c}}_{\rm in}(x,x')$ obeys the equation

$$(\mathscr{D}_\mu\gamma^\mu-m)S^{\bar{c}}_{in}(x,x')=\delta^{(d)}(x-x'). \tag{35}$$

As well, all these different kinds of the Green functions are used to represent various matrix elements of operators of the current and energy-momentum tensor, and effective action beginning with zeroth order with respect to radiative interaction.

Solutions (16) and (18) with quantum number p_- are especially adapted to calculate all the Green functions. One can express the Green functions via solutions (16) and (18)[5,21,3]

$$S^c(x,x')=\theta(x_0-x_0')S^-(x,x')-\theta(x_0'-x_0)S^+(x,x'), \tag{36}$$

$$S(x,x')=S^-(x,x')+S^+(x,x'), \tag{37}$$

$$\begin{aligned}
S^-(x,x')&=i\int_{-\infty}^{+\infty}dp_-\sum_{nr\{n_j'\}}{}^+\psi_{p_-,n,r}(x)g({}_+|^+)^{-1}_{nn'}\,{}^+\bar{\psi}_{p_-,n',r}(x'),\\
S^+(x,x')&=i\int_{-\infty}^{+\infty}dp_-\sum_{nr\{n_j'\}}{}^-\psi_{p_-,n,r}(x)[g({}_-|^-)^{-1}]^*_{nn'}\,{}^-\bar{\psi}_{p_-,n',r}(x'),
\end{aligned} \tag{38}$$

$$S^c_{in}(x,x')=\theta(x_0-x_0')S^-_{in}(x,x')-\theta(x_0'-x_0)S^+_{in}(x,x'), \tag{39}$$

$$S^{\bar{c}}_{in}(x,x')=\theta(x_0'-x_0)S^-_{in}(x,x')-\theta(x_0-x_0')S^+_{in}(x,x'), \tag{40}$$

$$S^{\mp}_{in}(x,x')=i\int_{-\infty}^{+\infty}dp_-\sum_{nr}{}^{\pm}\psi_{p_-,n,r}(x)\,{}^{\pm}\bar{\psi}_{p_-,n,r}(x'), \tag{41}$$

$$S^c_{out}(x,x')=\theta(x_0-x_0')S^-_{out}(x,x')-\theta(x_0'-x_0)S^+_{out}(x,x'), \tag{42}$$

$$S^{\mp}_{out}(x,x')=i\int_{-\infty}^{+\infty}dp_-\sum_{nr}{}^{\pm}\psi_{p_-,n,r}(x)\,{}^{\pm}\bar{\psi}_{p_-,n,r}(x'), \tag{43}$$

where all $p_j'=p_j$, the symbol Σ_{nr} means the summation over all discrete quantum numbers n_j, r_j and the integration over all continuous p_j, and the symbol $\Sigma_{\{n_j\}}$ means the summation over all discrete quantum numbers n_j only. Using the relations between the Green functions and between the matrices $G(_\zeta|^{\zeta'})$ one can present the functions $S^{\mp}$, $S_{in}^{\mp}$ and $S_{out}^{\mp}$ as follows:

$$\pm S^{\mp}(x,x')=S^c(x,x')\pm\theta(\mp(x_0-x_0'))S(x,x'), \tag{44}$$

$$\pm S_{in}^{\mp}(x,x')=S_{in}^c(x,x')\pm\theta(\mp(x_0-x_0'))S(x,x'), \tag{45}$$

$$\pm S_{out}^{\mp}(x,x')=S_{out}^c(x,x')\pm\theta(\mp(x_0-x_0'))S(x,x'), \tag{46}$$

$$S_{in}^c(x,x')=S^c(x,x')-S^a(x,x'), \tag{47}$$

$$S_{out}^c(x,x')=S^c(x,x')-S^p(x,x'), \tag{48}$$

$$S^a(x,x')=-i\int_{-\infty}^{+\infty}dp_-\sum_{nr\{n_j'\}}{}_-\psi_{p_-,n,r}(x)[g(_+|^-)g(_-|^-)^{-1}]^{\dagger}_{nn'}{}_+\overline{\psi}_{p_-,n',r}(x'), \tag{49}$$

$$S^p(x,x')=i\int_{-\infty}^{+\infty}dp_-\sum_{nr\{n_j'\}}{}^+\psi_{p_-,n,r}(x)[g(_+|^+)^{-1}g(_+|^-)]_{nn'}{}^-\overline{\psi}_{p_-,n',r}(x'). \tag{50}$$

To calculate all the types of Green functions it is sufficient to take sums in $S^{\pm}(x,x')$ and $S^{a,p}(x,x')$ only. That will be done below.

It follows from (18)

$$S^{\pm}(x,x')=\int_{-\infty}^{+\infty}\Theta(\mp\pi_-){}_+^-Y(x,x',p_-)dp_-, \tag{51}$$

$${}_+^-Y(x,x',p_-)=i\sum_{nr}{}_+^-\psi_{p_-,n,r}(x){}_+^-\overline{\psi}_{p_-,n,r}(x'). \tag{52}$$

Also taking into account Eqs. (23) one gets

$$\begin{aligned}S^a(x,x')&=-\int_{-\infty}^{+\infty}\Theta(-\pi_-){}_+Y(x,x',p_-)dp_-,\\ S^p(x,x')&=\int_{-\infty}^{+\infty}\Theta(\pi_-){}^-Y(x,x',p_-)dp_-.\end{aligned} \tag{53}$$

Due to the fact that ${}_+\psi_{p_-,n,r}(x)$ and ${}^-\psi_{p_-,n,r}(x)$ have similar form (16), the sums in (52) can be taken in the same way. Since

$$\sum_r v_{+1,r}v^+_{+1,r}=\Xi_+ \quad \text{if } d>3, \text{ and } v_{+1}v^+_{+1}=\Xi_+ \quad \text{if } d=2,3,$$

the summation over the spin quantum numbers can be done in (52) to get

$$\begin{aligned}{}_+^-Y(x,x',p_-)=&\left[\gamma^0+(\gamma_\perp\mathscr{P}_\perp+m)\frac{1}{\pi_-}\right]\Xi_+\exp\left(-i\frac{q}{2}\sigma^{\mu\nu}F^\perp_{\mu\nu}{}_+^-a\right)\\ &\times\left[\gamma^0+(-\gamma_\perp\mathscr{P}_\perp'{}^*+m)\frac{1}{\pi_-'}\right]\gamma^0\exp\left\{\frac{1}{2}[\ln(\mp\tilde{\pi}_-)+(\ln(\mp\tilde{\pi}_-'))^*]\right\}\\ &\cdot\sqrt{qE}{}_+^-\tilde{f}^{(0)}(x,x',p_-),\end{aligned} \tag{54}$$

$$ {}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-)=i\sum_n {}^{-}_{+}\varphi_{p_-,n,r}(x)\,{}^{-}_{+}\varphi^{*}_{p_-,n,r}(x'), \tag{55} $$

$$ {}^{-}_{+}\varphi_{p_-,n,r}(x)=\exp\{-\tfrac{1}{2}\ln(\mp\tilde{\pi}_-)\}\,{}^{-}_{+}\phi_{p_-,n,r}(x^0,x^D)\,\phi_{n,r}(x_\perp), $$

$$ {}^{-}_{+}a=(2qE)^{-1}[(\ln(\mp\tilde{\pi}'_-))^*-\ln(\mp\tilde{\pi}_-)], \tag{56} $$

$$ \tilde{\pi}'_-=\tilde{\pi}_-+\sqrt{qE}\,y_-,\ \ y_\mu=x_\mu-x'_\mu, $$

$$ \mathscr{P}_{\perp\mu}=0 \quad \text{if } \mu=0,D,\ \ \mathscr{P}_{\perp\mu}=\mathscr{P}_\mu \quad \text{if } k=1,\ldots,D-1, $$

$$ \mathscr{P}'^{*}_{\mu}=-i\,\frac{\partial}{\partial x'^{\mu}}-qA_\mu(x'). $$

It is convenient to make the replacement

$$ u=(2qE)^{-1}[(\ln(\mp\tilde{\pi}'_-))^*-\ln(\mp\tau)] $$

in the functions ${}^{-}_{+}K(x_-)$ and ${}^{-}_{+}J(x_-)$, and the one

$$ u=(2qE)^{-1}[(\ln(\tilde{\pi}'_-))^*-\ln\,\tau] $$

in the functions ${}^{-}_{+}K^*(x'_-)$ and ${}^{-}_{+}J^*(x'_-)$ [from ${}^{-}_{+}\phi_{p_-,n,r}(x^0,x^D)$ and ${}^{-}_{+}\phi^{*}_{p_-,n,r}(x'^0,x'^D)$ defined in (11)]. We do not discuss the integration paths over the variable u, since it is natural to consider the plane-wave potential $f(x_-)$ to be an entire function on x_-, in that case the integrals ${}^{-}_{+}K(x_-)$, ${}^{-}_{+}J(x_-)$, ${}^{-}_{+}K^*(x'_-)$ and ${}^{-}_{+}J^*(x'_-)$ are well defined by the integration limits. For different forms of the plane-wave potential, the integration paths over u are no longer arbitrary, however, they easily are extracted from the integration paths shown in Figs. 1 and 2. Then, one can present some combinations involved in (55) in the following way:

$$ {}^{-}_{+}K(x_-)-{}^{-}_{+}K^*(x'_-)=l({}^{-}_{+}a)+2(e^{2qF\,{}^{-}_{+}a}-1)\int_{{}^{-}_{+}b}^{0}e^{-2qFu}qf(x_-(u))du, $$

$$ {}^{-}_{+}K(x_-)+{}^{-}_{+}K^*(x'_-)=l({}^{-}_{+}a)+2(e^{2qF\,{}^{-}_{+}a}+1)\int_{{}^{-}_{+}b}^{0}e^{-2qFu}qf(x_-(u))du, \tag{57} $$

$$ {}^{-}_{+}J(x_-)-{}^{-}_{+}J^*(x'_-)=\Phi({}^{-}_{+}a)+2\int_0^{{}^{-}_{+}a}qf(x_-(u))e^{2qFu}\,du\ qF\int_{{}^{-}_{+}b}^{0}e^{-2qFu'}qf(x_-(u'))du', $$

where

$$ \Phi({}^{-}_{+}a)=\int_0^{{}^{-}_{+}a}qf(x_-(u))[qf(x_-(u))+qFl(u)]du, $$

$$ l(u)=2\int_0^u e^{2qF(u-u')}qf(x_-(u'))du', \tag{58} $$

$$ {}^{-}_{+}b=-\infty-i\pi/(2qE)\Theta(\mp\pi'_-),\ \ x_-(u)=x'_-+y_-\,\frac{1-\exp\{-2qEu\}}{1-\exp\{-2qE\,{}^{-}_{+}a\}}. $$

The summation over n_j in (55) can be performed using the Möller formula.[20] After integration over p_j in (55), applying the operators $\exp\{-i\,{}^{-}_{+}K(x_-)\pi_\perp\}$ and $\exp\{-i\,{}^{-}_{+}K(x_-)\pi'_\perp\}^*$, and using relations (57), we obtain

$$ {}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-)=\exp\left\{ iq\Lambda - im^2 {}^{-}_{+}a - \frac{1}{2}\left[\ln(\mp\tilde{\pi}_-)+(\ln(\mp\tilde{\pi}'_-))^*\right] - i\,\frac{\pi_- + \pi'_-}{4}\,y_+ \right\} h({}^{-}_{+}a), \tag{59} $$

$$ h({}^{-}_{+}a)=\exp\left\{ i\Phi({}^{-}_{+}a) - i\,\frac{1}{4}\,(y+l({}^{-}_{+}a))qF^{\perp}\,\coth(qF^{\perp}\,{}^{-}_{+}a)(y+l({}^{-}_{+}a)) + \frac{i}{2}\,yqF^{\perp}l(s) \right\} Z_{(d)}, \tag{60} $$

where

$$ \begin{aligned} Z_{(d)} &= c_d \prod_{j=1}^{(d-2)/2} \left(\frac{qH_j}{\sin(qH_j s)} \right), \quad d \text{ is even}, \\ Z_{(d)} &= c_d s^{-1/2} \prod_{j=1}^{(d-3)/2} \left(\frac{qH_j}{\sin(qH_j s)} \right), \quad d \text{ is odd}, \\ c_d &= (4\pi)^{-d/2} \exp\{-i\pi(d-4)/4\}, \end{aligned} \tag{61} $$

and

$$ \Lambda = -\int_{x'}^{x} (A^E_\mu + A^H_\mu)\,dx^\mu. \tag{62} $$

Here $A^E_\mu + A^H_\mu$ is a potential of the constant uniform field $F_{\mu\nu}$, and the integral is taken along the line.

Let us remark that the functions ${}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-)$ are special solutions of the Klein–Gordon equation,

$$ \left(\pi_- 2i\,\frac{\partial}{\partial x_-} - iqE + \mathscr{P}^2_\perp - m^2 \right) {}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-)=0. \tag{63} $$

Taking into account the relations

$$ \begin{aligned} \pi_\perp e^{iq\Lambda} &= e^{iq\Lambda}\left(i\,\frac{\partial}{\partial x_\perp} + \frac{1}{2}\,qFy_\perp \right), \\ \pi'^{*}_\perp e^{iq\Lambda} &= e^{iq\Lambda}\left(-i\,\frac{\partial}{\partial x'_\perp} - \frac{1}{2}\,qFy_\perp \right), \end{aligned} \tag{64} $$

$$ \begin{aligned} \exp(-iq\sigma^{0D}F_{0D}\,{}^{-}_{+}a) &= \cosh(qE\,{}^{-}_{+}a) - i\sigma^{0D}\,\sinh(qE\,{}^{-}_{+}a), \\ \exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F^{(j)}_{\mu\nu}\,{}^{-}_{+}a \right) &= \cos(qH_j\,{}^{-}_{+}a) + iR_j\,\sin(qH_j\,{}^{-}_{+}a), \end{aligned} \tag{65} $$

where

$$ F^{(j)}_{\mu\nu} = H_j(\delta^{2j}_\mu \delta^{2j-1}_\nu - \delta^{2j}_\nu \delta^{2j-1}_\mu), $$

and matrix R_j is defined by (15), one can get relation

$$ \begin{aligned} &\exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F^{\perp}_{\mu\nu}\,{}^{-}_{+}a \right) \gamma\pi'^{*}_\perp\, {}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-) \\ &\quad = (\gamma\pi_\perp + \gamma qF^{\perp}l({}^{-}_{+}a))\exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F^{\perp}_{\mu\nu}\,{}^{-}_{+}a \right) {}^{-}_{+}\tilde{f}^{(0)}(x,x',p_-). \end{aligned} \tag{66} $$

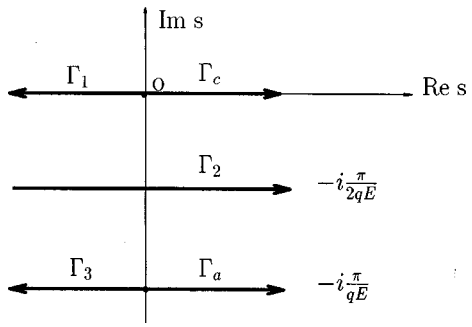

FIG. 3. Contours of integration $\Gamma_1,\Gamma_2,\Gamma_3,\Gamma_c,\Gamma_a$.

By using formulas (65), (66) and Eq. (63) one can transform expression (54) to the following form:

$$\bar{}_{+}Y(x,x',p_-)=\mp(\gamma\mathscr{P}+m)\bar{}_{+}\tilde{f}(x,x',p_-), \tag{67}$$

$$\bar{}_{+}\tilde{f}(x,x',p_-)=\left[\exp\left(-i\frac{q}{2}\sigma^{\mu\nu}F_{\mu\nu}\bar{}_{+}a\right)\mp\frac{1}{2}(\gamma^0-\gamma^D)\gamma\int_0^{\bar{}_{+}a}e^{qF(\bar{}_{+}a-2u)}q\frac{df(x_-(u))}{du}du\right.$$
$$\left.\times\exp\left\{-\frac{1}{2}[\ln(\mp\tilde{\pi}_-)+(\ln(\mp\pi'_-))^*]\right\}\right]\bar{}_{+}\tilde{f}^{(0)}(x,x',p_-). \tag{68}$$

In the external field under consideration the real proper time S is a function of π_- and π'_- because the classical equation of motion has a form $\pi_-^{-1}dx_-=m^{-1}dS$. Thus if $y_-\neq 0$, one can transform the p_- integration in Green functions into integration over the Fock–Schwinger proper time by making the change of the variable,

$$s=\bar{}_{+}a. \tag{69}$$

Then, one gets the following representations for the Green functions:

$$S^{\mp,a,p}(x,x')=(\gamma\mathscr{P}+m)\Delta^{\mp,a,p}(x,x'),$$

$$\mp\Delta^{\pm}(x,x')=\int_{\Gamma_c}f(x,x',s)ds-\Theta(\pm y_-)\int_{\Gamma_c-\Gamma_2-\Gamma_1}f(x,x',s)ds, \tag{70}$$

$$\Delta^a(x,x')=\int_{\Gamma_a}f(x,x',s)ds+\Theta(y_-)\int_{\Gamma_2+\Gamma_3-\Gamma_a}f(x,x',s)ds, \tag{71}$$

$$\Delta^p(x,x')=\int_{\Gamma_a}f(x,x',s)ds+\Theta(-y_-)\int_{\Gamma_1^a}f(x,x',s)ds. \tag{72}$$

All the contours of the integrals are shown in Fig. 3.
The contours Γ_c and Γ_1 are placed below the singular points on the real axis everywhere outside of the origin, and

$$f(x,x',s)=\left[\exp\left(-i\frac{q}{2}\sigma^{\mu\nu}F_{\mu\nu}s\right)+(n\gamma)\gamma\right.$$
$$\left.\times\int_0^s e^{qF(s-2u)}\frac{df((nx(u)))}{du}du\,\frac{\sinh(qEs)}{E(ny)}\right]f^{(0)}(x,x',s), \tag{73}$$

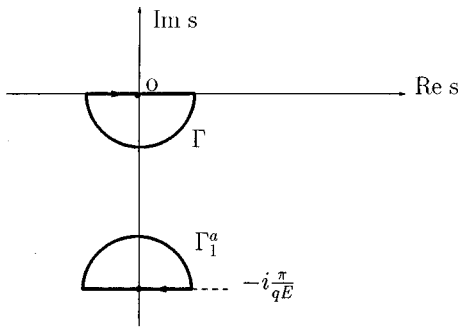

FIG. 4. Contours of integration Γ, Γ_1^a.

$$f^{(0)}(x,x',s)=\exp\{iq\Lambda\}\,\frac{qE}{\sinh(qEs)}$$
$$\times\exp\left\{-im^2s+i\Phi(s)-i\,\frac{1}{4}\,(y+l(s))qF\,\coth(qFs)(y+l(s))+\frac{i}{2}\,yqFl(s)\right\}Z_{(d)}, \tag{74}$$

where $Z_{(d)}$ is defined in (61). The function $f^{(0)}(x,x',s)$ has two singular points on the complex region between the contours $\Gamma_c-\Gamma_1$ and $\Gamma_a-\Gamma_3$. They are situated at the imaginary axis: $s_0=0$, and $qEs_1=-i\pi$. One can transform the contours $\Gamma_c-\Gamma_2-\Gamma_1$ into Γ and $\Gamma_2+\Gamma_3-\Gamma_a$ into Γ_1^a (see Fig. 4) with radii tending to zero. Since

$$\int_\Gamma f(x,x',s)ds=0 \quad \text{if } y_\mu y^\mu<0, \quad \int_{\Gamma_1^a} f(x,x',s)ds=0 \quad \text{if } y_0^2>y_D^2,$$

one can rewrite (70), (71) and (72) as follows:

$$S^{\mp}(x,x')=(\gamma\mathscr{P}+m)\Delta^{\mp}(x,x'), \tag{75}$$

$$\mp\Delta^{\pm}(x,x')=\int_{\Gamma^c} f(x,x',s)ds-\Theta(\pm y_0)\int_\Gamma f(x,x',s)ds,$$

$$S^{a,p}(x,x')=(\gamma\mathscr{P}+m)\Delta^{a,p}(x,x'), \tag{76}$$

$$\Delta^a(x,x')=\int_{\Gamma^a} f(x,x',s)ds+\Theta(-y^D)\int_{\Gamma_1^a} f(x,x',s)ds,$$

$$\Delta^p(x,x')=\int_{\Gamma^a} f(x,x',s)ds+\Theta(y^D)\int_{\Gamma_1^a} f(x,x',s)ds,$$

one can check that these expressions are valid for arbitrary x and x'. To prove this, one needs to verify that the functions $S^{\pm}(x,x')$ and $S^{a,p}(x,x')$, which are presented via integrals (75) and (76), and by means of representations (51) and (53), are the same solutions of the Dirac equation for any x and x'. Thus it is enough to check first that expressions (75) and (76) obey the Dirac equation for any x and x'. Then, one has to prove that the Cauchy conditions for distributions (75) and (76) coincide at $x_0=x_0'$ with (51) and (53), respectively. As to $S^{\pm}(x,x')$, one can use the proper time representations for $S^c(x,x')$ and $S(x,x')$ functions, which follow from (75), (36) and (37),

$$S^c(x,x')=(\gamma\mathscr{P}+m)\Delta^c(x,x'),$$

$$\Delta^c(x,x')=\int_{\Gamma_c} f(x,x',s)ds, \tag{77}$$

$$S(x,x')=(\gamma\mathscr{P}+m)\Delta(x,x'),$$

$$\Delta(x,x')=\mathrm{sign}(x_0-x_0')\int_{\Gamma} f(x,x',s)ds. \tag{78}$$

One can see that $f(x,x',s)$ obeys the following equations,

$$-i\frac{d}{ds}f(x,x',s)=\left(\mathscr{P}^2-m^2-\frac{q}{2}\sigma^{\mu\nu}\mathscr{F}_{\mu\nu}\right)f(x,x',s), \tag{79}$$

$$\lim_{s\to+0} f(x,x',s)=i\delta^{(d)}(x-x'). \tag{80}$$

Thus $f(x,x',s)$ is the Fock–Schwinger function.[22,23] So, representation (77) of causal Green function has the well-known Schwinger form.[22] The concrete representation (78) of the commutation function has the same form as the general representation.[24] Similar to Ref. 24, one can select all singularities of this function and see that it is continuous at x_0-x_0'. In the $d=4$ case one can transform this representation to the Fock form.[23] The analysis of space-time singularities in integrals (76) over the contour Γ_1^a can be done in a similar manner to Ref. 24. In the $d=4$ the representations, (75)–(78) coincide with the ones found in Ref. 8. To find the proper time representation we used the special behavior of solutions (16). But one has the different form of solutions (17) if d is odd, $E=0$ and all the imaginary eigenvalues H_j of the field tensor are not equal to zero. In this case since solutions (17) can represent a special case of the $d+1$ general form (16), one does not need to calculate Green functions of the case independently and representations (75), (77) and (78) are valid if one puts $E=0$, $f((nx))=0$ and replaces $qE/\sinh(qEs)$ $\to qH_{(d-1)/2}/\sin(qH_{(d-1)/2})$ in formulas (73) and (74) [the functions $S^{a,p}(x,x')$ are equal to zero].

IV. SOME PHYSICAL APPLICATIONS

All the information about the processes of particle creation, annihilation, and scattering in an external field (without radiative corrections) can be extracted from the matrices $G(_{\zeta}|^{\zeta'})$ (6). These matrices define a canonical transformation between in and out creation and annihilation operators in the generalized Furry representation,[5,3]

$$\begin{aligned} a^\dagger(out)&=a^\dagger(in)G(_+|^+)+b(in)G(_-|^+),\\ b(out)&=a^\dagger(in)G(_+|^-)+b(in)G(_-|^-). \end{aligned} \tag{81}$$

Here $a^\dagger_{\{n\}}(in)$, $b^\dagger_{\{n\}}(in)$, $a_{\{n\}}(in)$, $b_{\{n\}}(in)$ are creation and annihilation operators of in-particles and antiparticles, respectively, and $a^\dagger_{\{n\}}(out)$, $b^\dagger_{\{n\}}(out)$, $a_{\{n\}}(out)$, $b_{\{n\}}(out)$ are ones of out-particles and antiparticles; $\{n\}$ are possible quantum numbers. For example, let us calculate the mean numbers of antiparticles created (which are also equal to the number of pairs created) by the external field from the in-vacuum $|0,in\rangle$ with a given quantum number p_D,n,r. By using relations (81) and (21) one finds representation of this quantity

$$N_{p_D,n,r}=\langle 0,in|b^\dagger_{p_D,n,r}(out)b_{p_D,n,r}(out)|0,in\rangle=(G(_+|^-)^\dagger G(_+|^-))_{p_D,n,r,p_D',n',r}|_{p_D=p_D',n=n'}, \tag{82}$$

$$(G(_+|^-)^\dagger G(_+|^-))_{p_D,n,r,p_D',n',r}$$

$$= \int_{-\infty}^{+\infty} (G(_+|^-)G(_+|^-)^\dagger)_{p_-,n,r,p_-,n',r} M^*(p_D,p_-)M(p_D',p_-)\,\frac{dp_-}{2\pi qE}, \tag{83}$$

where the standard space coordinate volume regularization was used, so that $\delta(p_j-p_j')\to\delta_{p_j,p_j'}$. Then, by using formulas (28) and (29) one gets

$$N_{p_D,n,r} = \int_{-\infty}^{+\infty} {}^-\mathscr{D}_{nn} M^*(p_D,p_-)M(p_D',p_-)\frac{dp_-}{2\pi qE}\bigg|_{p_D=p_D'}. \tag{84}$$

If L_D is the length of the correspondent edge of the space box, then the maximum wave length of the plane wave is $2L_D$. Thus using the Fourier-series expansion of ${}^-\mathscr{D}$ one gets

$$N_{p_D,n,r} = \frac{1}{L_D}\int_0^{L_D} {}^-\mathscr{D}_{nn}\, dp_-. \tag{85}$$

One can calculate the quantities ${}^-_+\mathscr{D}_{nn'}$, given by Eq. (29), taking into account that the operator $\pi_\perp$ is Hermitian, and

$$e^{iK^*\pi_\perp}e^{-iK\pi_\perp} = \exp\left\{-\frac{i}{2}K^*qFK - i(K-K^*)\pi_\perp\right\}.$$

Then all the integrals over x^j in form (29) can be expressed in terms of the Laguerre polynomials.[25] In the special case when $n=n'$ one has

$${}^-\mathscr{D}_{nn} = \exp\left\{-\pi\lambda - \mathrm{Im}\ {}^-K(x_-)(qf(p_-/qE)+qF\ \mathrm{Re}\ {}^-K(x_-)) - \sum_{j=1}^{[d/2]-1} h_j\right\}\prod_{j=1}^{[d/2]-1} L_{n_j}(2h_j), \tag{86}$$

$$h_j = -|qH_j|((\mathrm{Im}\ {}^-K_{2j-1}(x_-))^2+(\mathrm{Im}\ {}^-K_{2j}(x_-))^2), \quad \pi_-\gt 0.$$

Remember that we are discussing the case in which the constant electric field acts for an infinite time. However, one can analyze the problem in finite times $T=x^0_{out}-x^0_{in}$, acting similar to Ref. 13. Then the mean number of pairs created by the external field $N_{p_D,n,r}$ is the same if time T is large enough: $\sqrt{qE}T\gg 1$, $\sqrt{qE}T\gg\lambda$ and $qET\gg|p_D|$.

Summing over the quantum numbers in (85), one can find the total number of pairs created from the vacuum. Using standard regularization with respect to the $(d-1)$-dimensional spatial volume $V_{(d-1)}$ and special regularization with respect to time T of acting of a constant electric field,[13] where $\int dp_D\, N_{p_D,n,r} = qETN_{p_D,n,r}$, one gets

$$N = V_{(d-2)}n^{cr},$$

$$n^{cr} = J_{(d)}\,\frac{Tm^2\beta(1)}{2^{(d-1)}\pi^{d/2}}\,\frac{E}{E_c}\exp\left\{-\pi\,\frac{E_c}{E}\right\}, \tag{87}$$

where n^{cr} is the number density of the created pairs for time T, and the coefficient $\beta(1)$ is defined by the next formula as a special case on $l=1$

$$\beta(l) = \prod_{j=1}^{(d-2)/2}\{qH_j\coth(l\pi H_j/E)\}, \quad d \text{ is even},$$

$$\beta(l) = \left(\frac{m^2E}{n\pi E_c}\right)^{1/2}\prod_{j=1}^{(d-3)/2}\{qH_j\coth(l\pi H_j/E)\}, \quad d \text{ is odd}. \tag{88}$$

Here, $E_c=m^2/|q|$ is the characteristic value of a constant electric field strength. This quantity N does not depend on the parameters of plane-wave field and is the same as the number of pairs created in a constant and uniform field.[13] The corresponding formulas for the $d=4$ case were first written in Ref. 26, and, in fact, can be derived easily from the Schwinger formulas.[22]

By using proper-time kernel $f(x,x',s)$ (73), one can construct the d-dimensional form of the Schwinger $out-in$ effective action[22]

$$\Gamma_{out-in}=\frac{1}{2}\,\mathrm{tr}\left\{\int dx\int_0^\infty s^{-1}f(x,x,s)ds\right\}. \tag{89}$$

It is not dependent on the parameters of the plane-wave field. And it is not amazing because the effective action is a function of the field invariants only which do not depend on the plane wave. Taking into account formulas (65) one can find the trace in (89) representation

$$\rho(s)=\mathrm{tr}\left\{\exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F_{\mu\nu}s\right)\right\}=2^{[d/2]}\cosh(qEs)\prod_{j=1}^{[(d-2)/2]}\cos(qH_js). \tag{90}$$

Then, one calculates the probability for a vacuum to remain a vacuum by using the Schwinger method,[22]

$$\begin{gathered}P_v=\exp(-2\,\mathrm{Im}\,\Gamma_{out-in})=\exp\{-\mu N\},\\ \mu=\sum_{l=0}^{\infty}\frac{\beta(l+1)}{(l+1)\beta(1)}\exp\left\{-l\pi\,\frac{E_c}{E}\right\},\end{gathered} \tag{91}$$

where $\beta(l)$ is defined by (88). This result coincides with the result from Ref. 13, and in $d=4$ the results coincide with Refs. 22 and 26.

Let the operator of the current of the Dirac field operator $\psi(x)$ have the form

$$j_\mu=\frac{q}{2}\,[\bar{\psi}(x),\gamma_\mu\psi(x)], \tag{92}$$

and operator of metric energy-momentum tensor (EMT) of the Dirac field operator have the form

$$\begin{gathered}T_{\mu\nu}=\tfrac{1}{2}\,(T^{\mathrm{can}}_{\mu\nu}+T^{\mathrm{can}}_{\nu\mu}),\\ T^{\mathrm{can}}_{\mu\nu}=\tfrac{1}{4}\,\{[\bar{\psi}(x),\gamma_\mu\mathscr{P}_\nu\psi(x)]+[\mathscr{P}^*_\nu\bar{\psi}(x),\gamma_\mu\psi(x)]\},\end{gathered} \tag{93}$$

where $T^{\mathrm{can}}_{\mu\nu}$ is the canonical EMT operator. We are going to discuss the following matrix elements with these operators:

$$\langle j_\mu\rangle^c=\langle 0,out|j_\mu|0,in\rangle c_v^{-1}, \tag{94}$$

$$\langle T_{\mu\nu}\rangle^c=\langle 0,out|T_{\mu\nu}|0,in\rangle c_v^{-1}, \tag{95}$$

$$\langle j_\mu\rangle^{in}=\langle 0,in|j_\mu|0,in\rangle, \tag{96}$$

$$\langle T_{\mu\nu}\rangle^{in}=\langle 0,in|T_{\mu\nu}|0,in\rangle, \tag{97}$$

$$\langle j_\mu\rangle^{out}=\langle 0,out|j_\mu|0,out\rangle, \tag{98}$$

$$\langle T_{\mu\nu}\rangle^{out}=\langle 0,out|T_{\mu\nu}|0,out\rangle. \tag{99}$$

Using the Green functions which were found before, one can present these matrix elements in the following form:

$$\langle j_\mu\rangle^c=iq\;\mathrm{tr}\{\gamma_\mu S^c(x,x)\}=iq\;\;\mathrm{tr}\{\gamma_\mu\gamma^k\mathscr{P}_\kappa\Delta^c(x,x')\}|_{x=x'}, \tag{100}$$

$$\langle T_{\mu\nu}\rangle^c = i/4\ \mathrm{tr}\{(\gamma_\mu(\mathscr{P}_\nu+\mathscr{P}'^*_\nu)+\gamma_\nu(\mathscr{P}_\mu+\mathscr{P}'^*_\mu))S^c(x,x')\}|_{x=x'} = i\ \mathrm{tr}\{B_{\mu\nu}\Delta^c(x,x')\}|_{x=x'}, \tag{101}$$

$$\langle j_\mu\rangle^{in} = \langle j_\mu\rangle^c - \langle j_\mu\rangle^a, \tag{102}$$

$$\langle j_\mu\rangle^{out} = \langle j_\mu\rangle^c - \langle j_\mu\rangle^p, \tag{103}$$

$$\langle T_{\mu\nu}\rangle^{in} = \langle T_{\mu\nu}\rangle^c - \langle T_{\mu\nu}\rangle^a, \tag{104}$$

$$\langle T_{\mu\nu}\rangle^{out} = \langle T_{\mu\nu}\rangle^c - \langle T_{\mu\nu}\rangle^p, \tag{105}$$

$$\langle j_\mu\rangle^{(a,p)} = iq\ \mathrm{tr}\{\gamma_\mu\gamma^\kappa\mathscr{P}_\kappa\Delta^{a,p}(x,x')\}|_{x=x'}, \tag{106}$$

$$\langle T_{\mu\nu}\rangle^{(a,p)} = i\ \mathrm{tr}\{B_{\mu\nu}\Delta^{a,p}(x,x')\}|_{x=x'}, \qquad B_{\mu\nu} = 1/4\{\gamma_\mu(\mathscr{P}_\nu+\mathscr{P}'^*_\nu)+\gamma_\nu(\mathscr{P}_\mu+\mathscr{P}'^*_\mu)\}\gamma^\kappa\mathscr{P}_\kappa, \tag{107}$$

where the Green functions are given by Eqs. (76), (77) and the relation

$$\Delta^c(x,x) = \tfrac{1}{2}[\Delta^-(x,x)-\Delta^+(x,x)]$$

is used. It is convenient to represent $\langle j_\mu\rangle^{a,p}$ and $\langle T_{\mu\nu}\rangle^{a,p}$ as follows:

$$-\langle j_\mu\rangle^a = \langle j_\mu\rangle^{(1)} + \langle j_\mu\rangle^{(2)}, \tag{108}$$

$$-\langle j_\mu\rangle^p = \langle j_\mu\rangle^{(1)} - \langle j_\mu\rangle^{(2)}, \tag{109}$$

$$-\langle T_{\mu\nu}\rangle^a = \langle T_{\mu\nu}\rangle^{(1)} + \langle T_{\mu\nu}\rangle^{(2)}, \tag{110}$$

$$-\langle T_{\mu\nu}\rangle^p = \langle T_{\mu\nu}\rangle^{(1)} - \langle T_{\mu\nu}\rangle^{(2)}, \tag{111}$$

where

$$\langle j_\mu\rangle^{(1)} = iq\ \mathrm{tr}\{\gamma_\mu\gamma^\kappa\mathscr{P}_\kappa\Delta^{(1)}(x,x')\}|_{x=x'}, \tag{112}$$

$$\langle T_{\mu\nu}\rangle^{(1)} = i\ \mathrm{tr}\{B_{\mu\nu}\Delta^{(1)}(x,x')\}|_{x=x'}, \tag{113}$$

$$\Delta^{(1)}(x,x') = -\frac{1}{2}\int_{\Gamma_3+\Gamma_2+\Gamma_a} f(x,x',s)ds, \tag{114}$$

and all contributions with derivatives of $\Theta(\pm y^D)$ functions, which are formally divergent, are included in terms $\langle j_\mu\rangle^{(2)}$ and $\langle T_{\mu\nu}\rangle^{(2)}$. The nature of such divergences is connected with infinite time T of acting of a constant electric field and was discussed in Ref. 13 (see also below).

The components $\langle j_\mu\rangle^{(1,2)}$ and $\langle T_{\mu\nu}\rangle^{(1,2)}$ cannot be calculated in the framework of the perturbation theory with respect to the external background or in the framework of the WKB method. Among them only the term $\langle j_\mu\rangle^{in}$ was calculated before in $d=4$.[3] Only expression (101) for $\langle T_{\mu\nu}\rangle^c$ has to be regularized and renormalized because of the ultraviolet divergences. Expression (100) for the term $\langle j_\mu\rangle^c$ is finite after the regularization lifting and equal to zero. The terms $\langle j_\mu\rangle^{(1)}$ and $\langle T_{\mu\nu}\rangle^{(1)}$ are also finite, and $\langle j_\mu\rangle^{(1)}=0$. The terms $\langle j_\mu\rangle^{(2)}$ and $\langle T_{\mu\nu}\rangle^{(2)}$ have to be regularizated with respect to time T of acting of a constant electric field[13] and do not have standard ultraviolet divergences. That is consistent with the fact that the ultraviolet divergences have a local nature and result (as in the theory without an external field) from the leading local terms at $s\to+0$. The nonzero contributions to the expressions $\langle j_\mu\rangle^{(2)}$ and $\langle T_{\mu\nu}\rangle^{(1,2)}$ are related to global features of the theory and indicate the vacuum instability.

At asymptotic region $x_0=T/2\to+\infty$, the densities of current and the EMT of particles created

are

$$j_\mu^{\rm cr}=\frac{\int d\mathbf{x}(\langle j_\mu\rangle^{in}-\langle j_\mu\rangle^{out})}{\int d\mathbf{x}}, \tag{115}$$

$$T_{\mu\nu}^{\rm cr}=\frac{\int d\mathbf{x}(\langle T_{\mu\nu}\rangle^{in}-\langle T_{\mu\nu}\rangle^{out})}{\int d\mathbf{x}}, \tag{116}$$

according to definitions (96)–(99).

Then, using representations (102)–(107), and taking into account terms $\langle j_\mu\rangle^{(2)}$ and $\langle T_{\mu\nu}\rangle^{(2)}$ which are uniform, one gets from (115) and (116),

$$j_\mu^{\rm cr}=2\langle j_\mu\rangle^{(2)}, \tag{117}$$

$$T_{\mu\nu}^{\rm cr}=2\langle T_{\mu\nu}\rangle^{(2)}. \tag{118}$$

Using special regularization with respect to time T of acting of a constant electric field,[13] one can interpret these divergent terms and get

$$\langle j_\mu\rangle^{\rm cr}=2|q|n^{\rm cr}\delta_\mu^D, \tag{119}$$

$$\langle T_{00}\rangle^{\rm cr}=\langle T_{33}\rangle^{\rm cr}=qETn^{\rm cr}, \tag{120}$$

and other components of $\langle T_{\mu\nu}\rangle^{\rm cr}$ are equal to zero.

To study the backreaction of particles created on the electromagnetic field and metrics one needs the expressions $\langle j_\mu\rangle^{in}$ and $\langle T_{\mu\nu}\rangle^{in}$. We found the form of $\langle j_\mu\rangle^{in}=\langle j_\mu\rangle^{(2)}=1/2\langle j_\mu\rangle^{\rm cr}$ and $\langle T_{\mu\nu}\rangle^{(2)}=1/2\langle T_{\mu\nu}\rangle^{\rm cr}$. Let us calculate other terms in expressions (104). By using formulas (65) one can find traces in (101) and (107) representations. One has the next nonzero traces: one presented by expression (90) and

$$\begin{aligned}&{\rm tr}\left\{\gamma^0\gamma^D\,\exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F_{\mu\nu}s\right)\right\}=\tanh(qEs)\rho(s),\\ &{\rm tr}\left\{\gamma^{2j}\gamma^{2j-1}\,\exp\left(-i\,\frac{q}{2}\,\sigma^{\mu\nu}F_{\mu\nu}s\right)\right\}=-\tan(qH_js)\rho(s).\end{aligned} \tag{121}$$

Then

$$\langle T_{\mu\nu}\rangle^c=\int_{\Gamma_c}\tau_{\mu\nu}(s)ds, \tag{122}$$

$$\begin{aligned}&\langle T_{\mu\nu}\rangle^{(1)}=-\frac{1}{2}\int_{\Gamma_3+\Gamma_2+\Gamma_a}\tau_{\mu\nu}(s)ds,\\ &\tau_{\mu\nu}(s)=b_\mu(s)\rho(s)f^{(0)}(x,x,s)\ \ \text{if}\ \ \mu=\nu,\\ &\tau_{0D}(s)=-b_D(s)\tanh(qEs)\rho(s)f^{(0)}(x,x,s),\\ &\tau_{\mu\nu}(s)=-b_{2j}(s)\tan(qH_js)\rho(s)f^{(0)}(x,x,s)\quad \text{if}\ \ \mu=2j,\quad \nu=2j-1,\quad j=1,2,...,[d/2]-1,\\ &b_D(s)=-b_0(s)=\frac{qE}{2}\coth(qEs),\quad b_{2j}(s)=b_{2j-1}(s)=\frac{qH_j}{2}\cot(qH_js).\end{aligned} \tag{123}$$

ACKNOWLEDGMENTS

D.M.G. thanks the Brazilian foundations CNPq for support. S.P.G. thanks the Brazilian foundation CAPES for support, and the Department of Physics of UEL and Department of Mathematical Physics of USP for their hospitality.

[1] W. Greiner, B. Müller, and J. Rafelski, *Quantum Electrodynamics of Strong Fields* (Springer-Verlag, Berlin, 1985).
[2] A. A. Grib, S. G. Mamaev, and V. M. Mostepanenko, *Vacuum Quantum Effects in Strong Fields* (Atomizdat, Moscow, 1988).
[3] E. S. Fradkin, D. M. Gitman, and S. M. Shvartsman, *Quantum Electrodynamics with Unstable Vacuum* (Springer-Verlag, Berlin, 1991).
[4] W. H. Furry, Phys. Rev. **81**, 115 (1951).
[5] D. M. Gitman, J. Phys. A **10**, 2007 (1977); D. M. Gitman and S. P. Gavrilov, Sov. Phys. J. **1**, 94 (1977); S. P. Gavrilov and D. M. Gitman, Sov. Phys. J. **6**, 491 (1980).
[6] V. G. Bagrov and D. M. Gitman, *Exact Solutions of Relativistic Wave Equations* (Kluwer Academic, Amsterdam, 1990).
[7] V. G. Bagrov, D. M. Gitman, and A. V. Jushin, Phys. Rev. D **12**, 3200 (1975); V. G. Bagrov, D. M. Gitman, and V. N. Shapovalov, J. Math. Phys. **23**, 2558 (1982).
[8] S. P. Gavrilov, D. M. Gitman, and Sh. M. Shvartsman, Sov. J. Nucl. Phys. **29**, 567 (1979); **29**, 715 (1979); Kratk. Soob. Fiz. (P. N. Lebedev Inst.) **2**, 22 (1979).
[9] S. P. Gavrilov and D. M. Gitman, Sov. Phys. J. **4**, 364 (1983); S. P. Gavrilov and D. M. Gitman, *ibid.* **9**, 775 (1982); S. P. Gavrilov and D. M. Gitman, Izw. VUZov Fizika (Sov. Phys. J.) **5**, 108 (1981); Yu. Yu. Volfengaut, S. P. Gavrilov, D. M. Gitman, and Sh. M. Shvartsman, Sov. J. Nucl. Phys. **33**, 386 (1981); D. M. Gitman, M. D. Noskov, and Sh. M. Shvartsman, Int. J. Mod. Phys. A **6**, 4437 (1991).
[10] I. L. Buchbinder, S. D. Odintsov, and I. L. Shapiro, *Effective Action in Quantum Gravity* (IOP, Bristol, 1992).
[11] G. Schäfer and H. Dehnen, J. Phys. A **13**, 517 (1980); I. L. Buchbinder and S. D. Odintsov, Sov. Phys. J. **5**, 12 (1982).
[12] S. P. Gavrilov, D. M. Gitman, and S. D. Odintsov, Int. J. Mod. Phys. A **12**, 4837 (1997).
[13] S. P. Gavrilov and D. M. Gitman, Phys. Rev. D **53**, 7162 (1996).
[14] R. Brauer and H. Weyl, Am. J. Math. **57**, 425 (1935).
[15] V. G. Bagrov, D. M. Gitman, and V. A. Kuchin, in *Actual Problems of Theoretical Physics* (MGU, Moscow 1976), p. 334; S. P. Gavrilov and D. M. Gitman, Sov. J. Nucl. Phys. **51**, 1040 (1990).
[16] P. F. Redmond, J. Math. Phys. **6**, 1163 (1965).
[17] D. M. Wolkov, Z. Phys. **94**, 250 (1935).
[18] N. B. Narozhny and A. I. Nikishov, Theor. Math. Phys. **26**, 9 (1976); N. B. Narozhny and A. I. Nikishov, in *Problems of Intense Field Quantum Electrodynamics*, Proc. P. N. Lebedev Phys. Inst. **168**, 175 (Nauka, Moscow, 1986).
[19] *Tables of Integral Transformations* (Bateman Manuscript Project), edited by A. Erdelyi *et al.* (McGraw-Hill, New York, 1954), Vol. 2.
[20] *Higher Transcendental Functions* (Bateman Manuscript Project), edited by A. Erdelyi *et al.* (McGraw-Hill, New York, 1953), Vol. 2.
[21] E. S. Fradkin and D. M. Gitman, Preprint MIT (1978), pp. 1–58; Preprint KFKI-1979-83 (Institute of Physics, Budapest), pp. 1–105; Preprint PhIAN (P. N. Lebedev Institute) **106** (1979), pp. 1–62; **107** (1979), pp. 1–40; Fortschr. Phys. **29**, 381 (1981).
[22] J. Schwinger, Phys. Rev. **82**, 664 (1951).
[23] V. Fock, Phys. Z. Sowjetunion **12**, 404 (1937).
[24] S. P. Gavrilov and D. M. Gitman, J. Math. Phys. **37**, 3118 (1996).
[25] I. S. Gradshtein and I. M. Rizhik, *Tables of Integrals, Sums, Series and Productions* (Nauka, Moscow, 1971).
[26] A. I. Nikishov, Zh. Eksp. Teor. Fiz. **57**, 1210 (1969) [Sov. Phys. JETP **30**, 660 (1970)]; in *Quantum Electrodynamics of Phenomena in Intense Fields,* Proc. P. N. Lebedev Phys. Inst. **111**, 153 (Nauka, Moscow, 1979).

Exponential stability for the energy exchange dynamics of the Holstein model

D. Hennig
Freie Universität Berlin, Fachbereich Physik, Institut für Theoretische Physik, Arnimallee 14, 14195 Berlin, Germany

(Received 30 April 1997; accepted for publication 15 April 1998)

We investigate the dynamics of the energy exchange between excitonic and vibrational degrees of freedom in the context of the Holstein model. Using Nekhoroshev-type arguments we derive bounds on the energy exchange for weak exciton–vibron coupling. It is shown that for large differences of the excitonic and vibronic frequencies the energy exchange is suppressed up to times growing exponentially with the ratio of the frequencies. Although the excitonic and vibronic energies are each for itself conserved there is equipartition of the excitonic energy among all lattice sites entailing an extended exciton. However, in the limit of large exciton–vibron coupling the excitonic actions are frozen separately, which leads to exciton localization. The action freezing is reflected in breather solutions. © *1998 American Institute of Physics.* [S0022-2488(98)01907-0]

I. INTRODUCTION

In their pioneering studies Fermi, Pasta, and Ulam (FPU)[1] showed that in the presence of nonlinearity the energy exchange between coupled systems may be suppressed, a behavior which is in contrast to the sharing of energy in linear models. In particular for systems of harmonic oscillators interacting via nonlinear springs they demonstrated numerically that the dispersion of the excitation energy can be inhibited, an effect which is also called self-trapping. The question of whether there appears equipartition of excitation energy is still an actual topic in various fields of physical significance. We quote here the energy exchange between excitonic and vibronic parts in the context of the Holstein model[2] and note that self-trapping of excitation energy plays an important role for optical storage patterns in nonlinear fibers, condensed matter physics, and biophysics.

With regard to the Holstein model we remark that weak exciton–vibron interactions allow a perturbational treatment of the corresponding system of coupled equations and that one can use ideas and techniques from classical mechanics. In the case of a finite number of degrees of freedom (DOF) application of perturbational methods lead to an understanding of stability phenomena in the energy exchange dynamics. Preservation of tori for all times provided the initial data lie in resonance free regions of phase space is assured by the Kolmogorov–Arnold–Moser theorem.[3] With the help of Nekhoroshev's arguments for nearly integrable systems, the stability of motion up to finite times growing exponentially with some power of the inverse of the perturbation strength is guaranteed[3–5] and the result holds for open sets of initial data. The energy exchange problem in the Holstein model can be formulated as a stability question in the sense of Nekhoroshev and one can derive estimates of the time interval during which the exciton–vibron energy exchange remains small. On the other hand, one has to be aware that the estimates of Nekhoroshev contain a dependence on the number of DOF such that the analytical stability results cannot be maintained in the thermodynamic limit when the number of DOF approaches infinity.[5,6] Bambusi and Giorgilli developed a general perturbational scheme making the extension of the Nekhoroshev statement to infinite lattices possible when the unperturbed motion has only a finite number of frequencies.[7] They applied their theory to a modified FPU chain and a system of weakly coupled rotators. The modified FPU lattice consists of particles having alternating light and heavy masses. The spectrum of the system is divided into a high frequency optical branch and an acoustic branch of low frequencies. The uncoupled system is decomposed into two sets of harmonic oscillators with a clear frequency separation between them. In the weak coupling limit

 © 1998 American Institute of Physics

it has been proven that there is little energy exchange between the optical and acoustical subsystems up to times growing exponentially with the frequency ratio. Suppression of the energy exchange does not mean that the individual actions associated with the two subsystems do not vary in time. However, such a situation seems to be unlikely since the optical modes are almost completely resonant. In the case of the lattice model of infinitely many weakly coupled rotators it has been proven that for localized initial conditions, i.e., if only a finite number of rotators is excited, such states are stable up to times which increase exponentially with a power of the inverse of the perturbation strength. In this case each action is separately frozen.[7]

In the present paper we investigate for the Holstein model the energy exchange between the excitonic and vibrational subsystems, respectively. Different energy sharing phenomena resulting from pronounced exchange, inhibition of exchange, and even strong quasiparticle localization are observed upon changes of the parameters of the system. In particular, when the uncoupled excitonic system and vibrational system evolve on distinct time scales we prove exponential inhibition of the energy exchange for weak exciton–vibron coupling and give quantitative statements on the stability properties of the coupled motion. Increasing the strength of the exciton–vibron interaction a transition to self-trapping of the excitonic energy appears for which the excitonic actions become separately frozen. In the extreme case even strong localization for which virtually the total excitonic energy is stored at a single lattice site is possible. The localization effect can be interpreted in terms of the discrete self-trapping (DST) model which arises after elimination of the vibronic DOF from the original Holstein system. The DST system exhibits a breather solution for which the action variables deviate only by a small amount from their initial values. In the case of static breathers the action variables are completely frozen and thus always stay at their initial values. Some aspects have already been discussed in Ref. 8.

II. THE HOLSTEIN MODEL

The Holstein model describes the movement of a single exciton along a one-dimensional molecular chain while interacting with intramolecular Einstein-type oscillators.[2] The Hamiltonian is given by

$$H = H_{\mathrm{exc}} + H_{\mathrm{vib}} + H_{\mathrm{int}}, \tag{1}$$

with

$$H_{\mathrm{exc}} = \epsilon \sum_n |c_n|^2 + V \sum_n |c_n - c_{n-1}|^2, \tag{2}$$

$$H_{\mathrm{vib}} = \sum_n \left(\frac{1}{2} p_n^2 + \frac{1}{2} \omega^2 q_n^2 \right), \tag{3}$$

$$H_{\mathrm{int}} = \alpha \sum_n q_n |c_n|^2. \tag{4}$$

The complex variable c_n is the excitonic occupation amplitude at a molecular site with index n, ϵ is the on-site energy, and V is the transfer matrix element due to dipole–dipole interaction. The vibrational part H_{vib} is given by the energy of a set of local harmonic oscillators corresponding to the intramolecular vibrations of the molecules where q_n and p_n are the coordinate and momentum, respectively. The exciton–vibron coupling is linear in the vibronic coordinate q_n and the latter couples to the excitonic occupation $|c_n|^2$. The strength of the coupling is regulated by α. The exciton–vibron interaction can be thought of as a modulation of the excitonic energy by the intramolecular vibrations.

The Holstein system can be cast into the form of a FPU model. To this end we perform a transformation from the excitonic variables c_n to pairs of momentum and coordinate variables (y_n, x_n), respectively, given by

$$c_n=\frac{1}{\sqrt{2}}\,(x_n+iy_n). \tag{5}$$

The transformed excitonic and interaction parts of the Hamiltonian then read as

$$H_{\mathrm{exc}}=\frac{1}{2}\,\epsilon\sum_n\,[x_n^2+y_n^2]+\frac{1}{2}\,V\sum_n\,[(x_n-x_{n-1})^2+(y_n-y_{n-1})^2], \tag{6}$$

$$H_{\mathrm{int}}=\frac{1}{2}\,\alpha\sum_n\,q_n(x_n^2+y_n^2). \tag{7}$$

For vanishing exciton–vibron coupling, i.e., $\alpha=0$, the system of (3), (6), and (7) decouples into two integrable systems, namely, the set of uncoupled intramolecular harmonic oscillators and the linear excitonic chain where both H_{vib} and H_{exc} are conserved quantities. When $\alpha\neq 0$ energy exchange between the two subsystems is possible due to the cubic interaction. It is the cubic coupling term which brings nonlinearity into play and in this sense the Holstein model represents a FPU system. We remark that the excitonic part H_{exc} can be brought into the form of a set of uncoupled harmonic oscillators through a transformation to normal coordinates. However, when dealing with the excitonic transfer it is more appropriate to use the coupled (nondiagonal) form of (6).

III. EXPONENTIAL BOUND ON THE ENERGY EXCHANGE

In this section we prove that for weak coupling $\alpha\ll 1$ and small ratio V/ω the energy exchange between the excitonic and vibronic DOF, respectively, is bounded up to exponentially long times. The proof is due to a Nekhoroshev-type argument and relies on a classical perturbation scheme developed by Bambusi and Giorgilli.[7] To be precise, we state the following theorem.

Theorem: Consider the Cauchy problem for the Hamiltonian (3), (6), (7) with initial data $\{p(0),q(0)\}$ and $\{x(0),y(0)\}$. The total energy $E(0)$ of the unperturbed (excitonic and vibronic) parts is given by

$$E(0)=\frac{1}{2}\,\epsilon\sum_n\,[(x_n(0))^2+(y_n(0))^2]+\sum_n\,\left(\frac{1}{2}\,(p_n(0))^2+\frac{1}{2}\,\omega^2(q_n(0))^2\right). \tag{8}$$

If it holds that

$$E(0)\leqslant\frac{1}{2^6}\left(\frac{\omega\epsilon}{\alpha}\right)^2, \tag{9}$$

and if the dimensionless parameter

$$\mu=2^9 3^2\left(\frac{V}{\omega}+\frac{\alpha}{\omega}\,\sqrt{E(0)}\left[\frac{1}{\omega^2}+\frac{1}{4\epsilon^2}\right]^{1/2}\right) \tag{10}$$

satisfies $\mu<1$ then the following estimate is true

$$\frac{|H_{\mathrm{vib}}(t)-H_{\mathrm{vib}}(0)|}{E(0)}\leqslant\mu, \tag{11}$$

for all times with

$$t\leqslant\frac{1}{\omega\mu}\,\exp\left(\frac{1}{\mu}\right). \tag{12}$$

Proof: The proof of the above theorem relies on the normal form results of Ref. 7. Before proceeding we set up the phase space, a symplectic form on it, and a norm for points in phase space. The phase space $\mathscr{P}:=l^2\times l^2\times l^2\times l^2$ with l^2 being the space of square summable sequences. The symplectic form Ω is defined as

$$\Omega(((p^1,q^1),(y^1,x^1)),((p^2,q^2),(y^2,x^2)))=\sum_k [(p_k^1 q_k^2-p_k^2 q_k^1)]+[(y_k^1 x_k^2-y_k^2 x_k^1)]. \tag{13}$$

The norm of a point $\xi=(p,q,x,y)\in\mathscr{P}$ is conveniently taken as the sum of the excitonic and vibronic partial energies, respectively:

$$\|\xi\|_{\mathscr{P}}^2=H_{\mathrm{exc}}(\xi)+H_{\mathrm{vib}}(\xi)=\frac{1}{2}\,\epsilon\sum_n [|x_n|^2+|y_n|^2]+\sum_n \left(\frac{1}{2}\,|p_n|^2+\frac{1}{2}\,\omega^2|q_n|^2\right). \tag{14}$$

In Ref. 7 an abstract scheme was introduced to develop a normal form theory for Hamiltonian systems defined on a generic Banach space. In particular, all the quantitative estimates obtained in the framework of this theory are then independent of the number of DOF. Here we apply this perturbation theory to bring the system (15) into normal form up to an exponentially small remainder.

We consider the real Hamiltonian function $H:\mathscr{G}\to R$ and decompose H as follows:

$$H=H_{\mathrm{vib}}+H_{\mathrm{exc}}+H_{\mathrm{int}}, \tag{15}$$

with H_{exc} and H_{int} being of class $C^\infty(\mathscr{P})$ and their symplectic gradients are defined for all $\xi\in\mathscr{P}$. For quantitative perturbational computation we consider the domain $\mathscr{G}=\{0\}\subset\mathscr{P}$ and its complex extension $\mathscr{G}_R=\mathscr{P}^{\mathscr{C}}$ is defined via the positive parameters R and $d<1$, yielding the complex ball B centered at the origin of phase space:

$$\mathscr{G}_{R,d}=B(\xi=0,R(1-d)). \tag{16}$$

(The parameter R will be fixed later.) Apparently, H_{vib} can be extended to a complex analytic function on $\mathscr{G}_{R,d}$. H_{vib} is considered as the unperturbed system. The main point is that the system H_{vib} generates a periodic continuous flow Φ_t on $\mathscr{P}^{\mathscr{C}}$ with period $T=2\pi/\omega$. It is readily seen that for every $d<1$ the complex domain $\mathscr{G}_{R,d}$ is invariant for Φ_t. With regard to the smoothness of Φ_t one has that the map Φ_t is of class $C^1(\mathscr{P}^{\mathscr{C}},\mathscr{P}^{\mathscr{C}})$ for every fixed t and the time derivative $d\Phi_t(\xi)/dt$ exists for every $\xi\in D$, where D is a flow-invariant dense subset of $\mathscr{P}^{\mathscr{C}}$. Finally, it holds $\forall t\in R$ that $\|\Phi_t'\|_{\mathscr{P},\mathscr{P}}<K$ for some $K>0$. With these smoothness conditions for the flow Φ_t related to H_{vib} the perturbation theory of Ref. 7 can be applied.

H_{vib} and H_{exc} represent two integrable subsystems if $\alpha=0$. For $\alpha\neq 0$ they interact through H_{int}. Moreover, the systems H_{vib} and H_{ext} commute, that is $\{H_{\mathrm{vib}},H_{\mathrm{exc}}\}=0$. Thus H_{exc} is already in normal form with respect to H_{vib}. Further we assume that $\nabla^\Omega H_{\mathrm{exc}}$ and $\nabla^\Omega H_{\mathrm{int}}$ can be extended to complex analytic functions from $\mathscr{G}_{R,0}$ to $\mathscr{P}^{\mathscr{C}}$ fulfilling the inequalities

$$\omega_0\geqslant\frac{1}{R}\sup\|\nabla^\Omega H_{\mathrm{exc}}(\xi)\|,\quad \omega_1\geqslant\frac{1}{R}\sup\|\nabla^\Omega H_{\mathrm{int}}(\xi)\|, \tag{17}$$

for $\xi\in\mathscr{G}_{R,0}$ and constants $\omega_{0,1}$. Provided the number μ given by

$$\mu=\frac{24}{d}\,e\pi\left(\frac{\omega_0+\omega_1}{\omega}\right), \tag{18}$$

satisfies $\mu<1$ for $0<d\leqslant 1/4$ then there exists an analytical canonical transformation $\mathscr{T}:\mathscr{G}_{R,2d}\to\mathscr{G}_{R,d}$ such that $H\circ\mathscr{T}$ takes the form

$$H(\mathscr{T}(\xi))=H_{\mathrm{vib}}(\xi)+H_{\mathrm{exc}}(\xi)+\mathscr{Z}(\xi)+\mathscr{R}(\xi). \tag{19}$$

Moreover, the following bound is valid for the symplectic gradient of $\mathscr{R}$:

$$\frac{1}{R}\sup_{\xi\in\mathcal{G}_{R,2d}}\|\nabla^{\Omega}\mathcal{R}(\xi)\|\leqslant 3(e+1)\omega_1\mu\,\exp\left(-\frac{1}{\mu}\right). \tag{20}$$

Furthermore, $\mathcal{Z}$ is in normal form with respect to H_{vib}, i.e., $\{H_{\mathrm{vib}},\mathcal{Z}\}=0$. Finally, for any analytic function $g:\mathcal{G}_{R,0}\to\mathcal{B}$, with $\mathcal{B}$ a generic Banach space, one obtains

$$\sup_{\xi\in\mathcal{G}_{R,2d}}\|g(\xi)-g(\mathcal{T}(\xi)\|_{\mathcal{B}}\leqslant\frac{2\pi(e+1)}{d}\,\frac{\omega_1}{\omega}\sup_{\xi\in\mathcal{G}_{R,d}}\|g(\xi)\|_{\mathcal{B}}. \tag{21}$$

With the help of quantitative estimates resulting from the normal form computations one derives an upper bound on the variation of the harmonic energy H_{vib} up to exponentially long times (See corollary 4.2 of Ref. 7). One infers that along the solutions of the Cauchy problem of (15) it holds that

$$|H_{\mathrm{vib}}(t)-H_{\mathrm{vib}}(0)|\leqslant\frac{6\pi(e+1)}{d}\,\frac{\omega_1}{\omega}\sup_{\xi\in\mathcal{G}_{R,d}}|H_{\mathrm{vib}}(\xi)|, \tag{22}$$

The inequality (22) is valid for all times $t\leqslant\min(T_0,T_*)$, where T_0 is the time it takes for the solution to escape from the domain $\mathcal{G}_{R,2d}$, and

$$T_*=\frac{2\pi}{3}\,\frac{1}{\omega\mu}\,\exp\left(\frac{1}{\mu}\right). \tag{23}$$

The proof of the normal form statement is based on a recursive algorithm which is equivalent to the Lie transform method used to perform consecutive canonical changes to give a (perturbed) Hamiltonian system a normal form. In this case one seeks to find normal forms in the neighborhood of an unperturbed periodic orbit. The scheme of the proof is presented in Ref. 7. In finite dimension the procedure goes essentially along the lines given in Ref. 9. In achieving the normal form one has to define a suitable norm for functions in order to obtain estimates for the Poisson brackets and the homological equation in terms of this norm. The extension of the method for the normal form computation to infinite dimensional systems requires a proper implementation of the the averaging procedure to derive estimates for the solution of the homological equation. According to Ref. 10 the averaging is performed over a periodic orbit of the unperturbed system. That the treatment of the periodic case does not impose restrictions on the analysis and involves the quasiperiodic orbits as well can be seen by noting that (in finite dimensions) all the frequencies are in a sense close to a resonance. In fact, according to Lochak any quasiperiodic motion can be approximated by a periodic one exploiting a theorem of Dirichlet.[11]

In order to verify the statement of the theorem we need the estimates of $\omega_{0,1}$ for $\xi\in\mathcal{G}_{R,0}$. Computation of ω_1 requires the expressions for the components of the symplectic gradient of H_{int}, which are given by

$$(\nabla^{\Omega}H_{\mathrm{int}})_{p_k}=-\frac{\partial H_{\mathrm{int}}}{\partial q_k}=-\frac{1}{2}\,\alpha(x_k^2+y_k^2), \tag{24}$$

$$(\nabla^{\Omega}H_{\mathrm{int}})_{q_k}=\frac{\partial H_{\mathrm{int}}}{\partial p_k}=0, \tag{25}$$

$$(\nabla^{\Omega}H_{\mathrm{int}})_{y_k}=-\frac{\partial H_{\mathrm{int}}}{\partial x_k}=-\alpha q_k x_k\,, \tag{26}$$

$$(\nabla^{\Omega}H_{\mathrm{int}})_{x_k}=\frac{\partial H_{\mathrm{int}}}{\partial y_k}=\alpha q_k y_k\,. \tag{27}$$

We derive

$$\|(\nabla^{\Omega}H_{\mathrm{int}})_p\|^2=\frac{1}{8}\,\alpha^2\sum_n\,(|x_n|^2+|y_n|^2)^2\leqslant\frac{1}{8}\,\alpha^2\left(\frac{2}{\epsilon}\right)^2\|\xi\|^4. \tag{28}$$

Furthermore, we get

$$\|(\nabla^{\Omega}H_{\mathrm{int}})_x\|^2+\|(\nabla^{\Omega}H_{\mathrm{int}})_y\|^2\leqslant\alpha^2\,\frac{2}{\omega^2}\,\|\xi\|^4. \tag{29}$$

Finally, we obtain

$$\omega_1=\alpha\left[\frac{2}{\omega^2}+\frac{1}{2\epsilon^2}\right]^{1/2}R. \tag{30}$$

With concern to ω_0 we need the gradients

$$(\nabla^{\Omega}H_{\mathrm{exc}})_{y_k}=-\frac{\partial H_{\mathrm{exc}}}{\partial x_k}=-V(x_k-x_{k-1})+V(x_{k+1}-x_k), \tag{31}$$

$$(\nabla^{\Omega}H_{\mathrm{exc}})_{x_k}=\frac{\partial H_{\mathrm{exc}}}{\partial y_k}=V(y_k-y_{k-1})-V(y_{k+1}-y_k). \tag{32}$$

We then estimate

$$\|(\nabla^{\Omega}H_{\mathrm{exc}})_x\|^2+\|(\nabla^{\Omega}H_{\mathrm{exc}})_y\|^2\leqslant\frac{1}{2}\,\epsilon\sum_n\,4V^2[|x_n-x_{n-1}|^2+|y_n-y_{n-1}|^2]\leqslant16V^2\|\xi\|^2, \tag{33}$$

and obtain eventually,

$$\omega_0=4V. \tag{34}$$

To arrive at the bound of the variation of $|H_{\mathrm{vib}}(t)-H_{\mathrm{vib}}(0)|$ given in the theorem above we have to assure that the orbit does not escape from the domain $\mathcal{G}_{R,2d=(1/2)}$ where we choose $d=1/4$. Exploiting the conservation of energy we can take R such that the escape time actually becomes infinite.[7] With the choice of the norm of a point ξ as the sum of the partial energies $H_{\mathrm{vib}}(t)+H_{\mathrm{exc}}(t)\equiv E(t)$ one has

$$\|\xi\|^2=E(t)\leqslant E(0)+|H_{\mathrm{int}}(0)|+|H_{\mathrm{int}}(t)|. \tag{35}$$

The size of the interaction can be bounded by $|H_{\mathrm{int}}|\leqslant\sqrt{2}\alpha/(\omega\epsilon)R^3$. If the right-hand side of (35) is less than or equal to $R^2/4$ then the solutions remain trapped inside the domain $\mathcal{G}_{R,(1/2)}$. From the latter condition follows the inequality

$$E(0)+2\alpha\,\frac{\sqrt{2}}{\omega\epsilon}\,R^3-\frac{R^2}{4}\leqslant0, \tag{36}$$

which can be satisfied for $R=\sqrt{8E(0)}$ provided

$$E(0)\leqslant\frac{1}{2^6}\left(\frac{\omega\epsilon}{\alpha}\right)^2, \tag{37}$$

yielding the inequality (9) from the theorem. Combining (30) and (34) readily gives μ as stated in the theorem which completes the proof. □

The above theorem tells us that if the (uncoupled) excitonic and vibronic partial dynamics of the Holstein lattice evolve on different time scales then there will be little energy exchange between the two subsystem up to times growing exponentially with the frequency separation between them. The frequency separation is determined by the ratio of ω/V, and thus for fixed V the vibronic frequency ω has to be large enough. It is worth mentioning that all the estimates

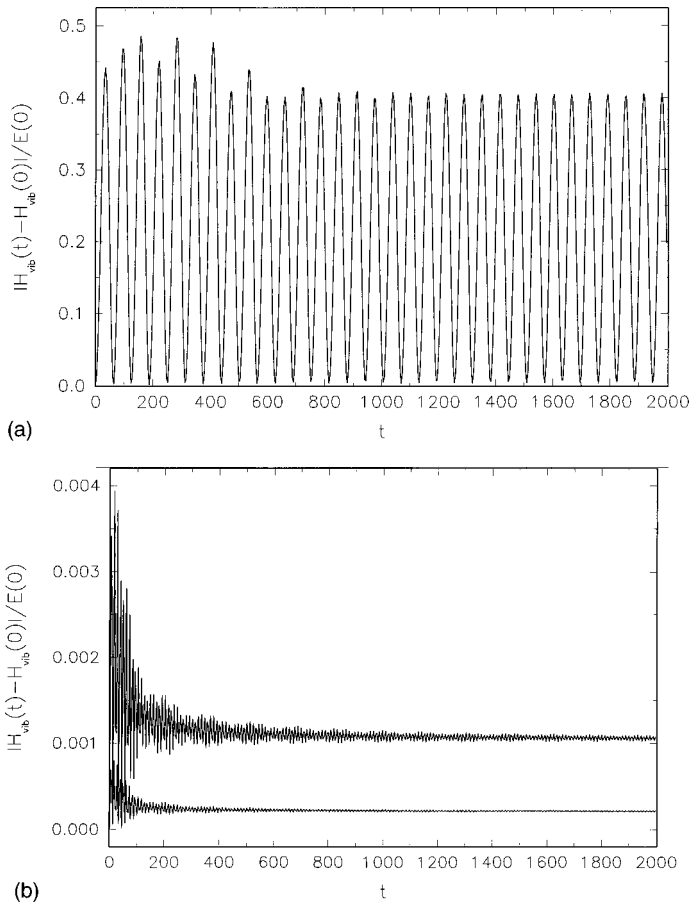

FIG. 1. Dynamics of the exchange between excitonic energy and vibronic energy in the Holstein model. (a) Time evolution of the energy ratio $|H_{\mathrm{vib}}(t)-H_{\mathrm{vib}}(0)|/E(0)$ for a chain consisting of $N=500$ sites. The excitonic energy is initially localized at the three central lattice sites, i.e., $x_{248\,250}(0)=y_{248\,250}(0)=0.5/3$, and $x_{249}(0)=y_{249}(0)=0.5$ and the remaining sites of $n\neq 248$, 249, 250 are unexcited. Parameters: $\omega^2=0.01$, $V=0.01$, and $\alpha=0.1$. (b) The same as in (a) but for increased frequency $\omega=1$. Upper curve $\epsilon=1$ and lower curve $\epsilon=5$.

entering the above theorem on stability are independent of the chain length N and thus the results on the bound of the energy exchange can be extended to the case of $N\to\infty$. In the present study the Nekhoroshev theorem does not separately control as usual the variations of the individual actions; it rather provides a bound on the energy exchange between the two subsystems of the Holstein lattice, namely the excitonic and vibronic parts, respectively. Nevertheless, the freezing of the separate excitonic actions is of considerable physical interest since it could lead to localization of the quasiparticle if the excitonic energy becomes stored at only a few lattice sites or even at a single site.

In order to illustrate the stability phenomenon expressed by the above theorem we show in Figs. 1(a) and 1(b) the temporal evolution of the energy exchange rate $|H_{\mathrm{vib}}(t)-H_{\mathrm{vib}}(0)|/E(0)$. For a lattice of length $N=500$ we used localized initial conditions for which only the central three sites contain solely excitonic energy and the remainder of the chain is kept unexcited. The duration of the integration interval is up to 10^{10} times the period $T=2\pi/\omega$ of an (uncoupled) vibronic oscillator. Figure 1(a) shows the evolution of the energy ratio for a frequency $\omega^2=0.01$. Obviously there is a pronounced energy exchange between the two subsystems. Due to the statement of the theorem increasing the frequency to $\omega^2=1$ results in a suppression of the energy exchange and each partial energy becomes virtually conserved as seen in Fig. 1(b). The upper and lower curves belong to an on-site energy $\epsilon=1$ and $\epsilon=5$, respectively. According to the estimate (11) it holds that the larger the on-site energy ϵ the more diminished the energy exchange, which is confirmed in the numerical investigations. However, the conservation of the excitonic energy does not lead necessarily to a freezing of the exchange among the various excitonic DOF in the sense that all

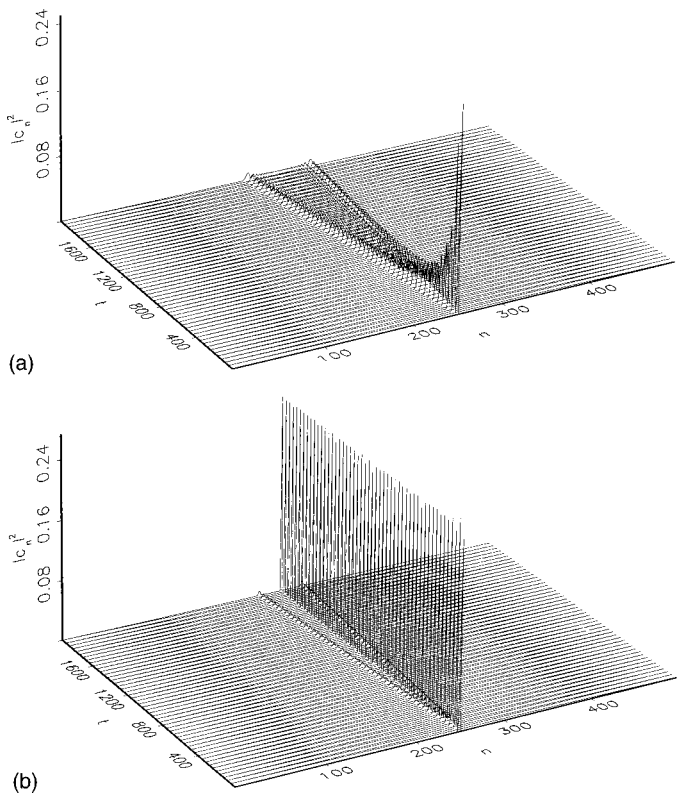

FIG. 2. Amplitude profile of the excitonic occupations (actions) $|c_n(t)|^2$ for the Holstein system. Initial data as in Fig. 1. (a) Equipartition of the excitonic energy among the molecular lattice sites. Parameters $\omega^2=0.5$, $\epsilon=1$, $V=0.01$, and $\alpha=0.1$. (b) Localization of the excitonic energy. Parameters as in (a) except for enlarged coupling $\alpha=1$. An equal picture of excitonic localization results from the DST system with $\gamma\equiv\alpha^2/\omega^2=2$.

individual excitonic occupations (actions) $|c_n(t)|^2$ stay close to their initial values $|c_n(0)|^2$. The latter behavior becomes apparent from Fig. 2(a) where the spreading of the excitonic actions $|c_n(t)|^2$ over the lattice in course of time is shown. Interestingly, when the exciton–vibron coupling is enlarged then the excitonic energy becomes highly localized around a single lattice site as Fig. 2(b) reveals. In Sec. IV we characterize this localized excitation pattern as a single site breather solution.

IV. THE DST SYSTEM, BREATHER SOLUTIONS, AND FREEZING OF THE ACTIONS

The results of Sec. III have shown that for weak coupling α (and by fixed V as well as ϵ) the exciton–vibron energy exchange is suppressed up to times growing exponentially with the vibron frequency ω. On the other hand between the individual DOF of the exciton subsystem we observe an energy exchange manifested in equipartition of the excitonic energy among the molecular lattice sites.

However, when α is increased one notices that not only the energy exchange between the two subsystems is suppressed but also the excitonic actions $|c_n|^2$ stay close to their initial values and exhibit less variations from constancy. In other words the exciton actions become (almost) conserved, which is also called freezing of the actions.[5]

The latter freezing effect can be understood in a heuristic way by considering the equations of motion for the coupled exciton–vibron system:

$$\ddot{q}_n=-\omega^2 q_n-\alpha|c_n|^2, \tag{38}$$

$$i\dot{c}_n=V[c_{n+1}+c_{n-1}-2c_n]+\epsilon c_n+\alpha q_n c_n\,. \tag{39}$$

In the limit of large frequency ω as well as large coupling α the terms on the right-hand side of (38) dominate the second derivative of $q_n(t)$ on the left-hand side giving $q_n=-\alpha/\omega^2|c_n|^2$

$\equiv \tilde{q}_n$, implying that the q coordinate is enslaved to $q_n = \tilde{q}_n$ as long as the ratio ω^2/α is of the order one. Inserting the instantaneous shift $q_n = \tilde{q}_n$ into Eq. (39) one obtains a closed equation in the excitonic occupations c_n in the form of a nonlinear discrete Schrödinger equation,

$$i\dot{c}_n = V[c_{n+1} + c_{n-1} - 2c_n] + \epsilon c_n - \gamma |c_n|^2 c_n , \tag{40}$$

which is called the discrete self-trapping (DST) equation.[12] Concerning the localization properties we remark that the DST system (with nonlinearity strength $\gamma = \alpha/\omega^2$) and the Holstein system (with α and ω) produce equal excitonic localization patterns $|c_n(t)|^2$. Thus the DST system serves to explain the localization behavior of the Holstein system from which the vibronic DOF are eliminated. In the following we briefly review the self-trapping properties of the DST system related to breather solutions. The stationary exciton localization pattern illustrated in Fig. 2(b) resembles a DST breather solution of the form $c_n = \phi_n \exp(i\Omega t)$. This solution is time independent in a rotating frame of frequency Ω and its spatial amplitudes ϕ_n decay exponentially with increasing distance from the central excitation peak at site $n = 250$. A characteristic feature of the single frequency DST breathers is that the excitonic occupations (actions) are completely frozen, i.e., $|c_n(t)|^2 = |c_n(0)|^2$. The linear stability of such an excitation pattern strongly localized around a single lattice site is readily demonstrated (see Ref. 13).

In Ref. 14 the existence of breather solutions for general networks of (weakly) coupled anharmonic oscillators has been proven by means of the anti-integrable limit. These authors suggested also the application of the anti-integrable limit to prove the existence of breather solutions for the DST system (cf. Sec. 9 of Ref. 14). The question then is whether such localized breather solutions are stable in the sense of Nekhoroshev, i.e., if one starts μ-close (the small parameter μ measures the deviation from integrability) within such localized periodic solutions does one stay then within $\mathcal{O}(\mu)$ at least for a time interval of the order of $C_1 \exp(-C_2/\mu^k)$ for some positive constants $C_{1,2}$ and k? That such solutions cannot be Lyapunov stable becomes clear by the fact that the Arnold web permeates the phase space densely. Recently Bambusi assured stability of breather solutions of general Hamiltonian networks for exponentially long time scales.[15] The DST system represents such a Hamiltonian lattice and the exponential stability of its localized periodic solutions corresponding to localized initial conditions can be proven with a Nekhoroshev-type argument using the theory developed in Ref. 15. In this way one can show that the situation of initially localized excitations changes little in course of time yielding exponentially stable breather states.

ACKNOWLEDGMENT

This work was supported by the Deutsche Forschungsgemeinschaft via Sonderforschungsbereich 337.

[1] E. Fermi, J. Pasta, and S. Ulam, Los Alamos Report No. LA-1940 (1955), later published in Lect. Appl. Math. **15**, 143 (1974).
[2] T. D. Holstein, Ann. Phys. (N.Y.) **8**, 325 (1959); **8**, 343 (1959).
[3] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978); *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer, New York, 1988).
[4] N. N. Nekhoroshev, Russ. Math. Surv. **32**, 1 (1977); Trudy Sem. Petrovs. **5**, 5 (1979).
[5] L. Galgani, A. Giorgilli, A. Martinoli, and S. Vanzini, Physica D **59**, 334 (1992), and references therein.
[6] G. Benettin and G. Gallavotti, J. Stat. Phys. **44**, 293 (1986); C. E. Wayne, Commun. Math. Phys. **103**, 351 (1986), **104**, 21 (1986), G. Benettin, L. Galgani, and A. Giorgilli, *ibid.* **113**, 87 (1987); G. Benettin, J. Fröhlich, and A. Giorgilli, *ibid.* **119**, 95 (1988).
[7] D. Bambusi and A. Giorgilli, J. Stat. Phys. **71**, 569 (1993).
[8] D. Hennig, in Proceedings of the Nonlinearity, Disorder and Fluctuations Conference, Crete 1996 [Physica D **113**, 196 (1998)].
[9] G. Benettin, L. Galgani, and A. Giorgilli, Commun. Math. Phys. **121**, 557 (1989).
[10] J. Moser, Commun. Pure Appl. Math. **23**, 609 (1970).
[11] P. Lochak, Russ. Math. Surv. **47**, 57 (1992).
[12] J. C. Eilbeck, P. S. Lomdahl, and A. C. Scott, Physica D **16**, 318 (1985).
[13] D. Hennig, K. Ø. Rasmussen, H. Gabriel, and A. Bülow, Phys. Rev. E **54**, 5788 (1996).
[14] R. S. MacKay and S. Aubry, Nonlinearity **7**, 1623 (1994).
[15] D. Bambusi, Nonlinearity **9**, 433 (1996).

Adler–Kostant–Symes scheme for face and Calogero–Moser–Sutherland-type models

Branislav Jurčo[a)]
CRM, Université de Montréal, Montréal, Quebec H3C 3J7, Canada

Peter Schupp[b)]
Department of Physics, Princeton University, Princeton, New Jersey 08544-0708

(Received 20 October 1997; accepted for publication 23 February 1998)

We give the construction of quantum Lax equations for IRF models and the difference version of the Calogero–Moser–Sutherland model introduced by Ruijsenaars. We solve the equations using factorization properties of the underlying face Hopf algebras/elliptic quantum groups. This construction is in the spirit of the Adler–Kostant–Symes method and generalizes our previous work to the case of face Hopf algebras/elliptic quantum groups with dynamical R matrices. © *1998 American Institute of Physics.* [S0022-2488(98)03306-4]

I. INTRODUCTION

Face Hopf algebras[1] have been found to be the algebraic structure that underlies some particularly interesting integrable models of statistical physics. They generalize Hopf algebras and quantum groups. Another closely related generalization of quantum groups, the so-called elliptic quantum groups, were introduced by Felder[2] in the context of IRF (face) models.[3] Face Hopf algebras and elliptic quantum groups play the same role for face models as quantum groups do for vertex models. The R matrices are now replaced by dynamical R matrices, which first appeared in the context of Liouville string field theory;[4] they can be understood as a reformulation of the Boltzmann weights in Baxters solutions of the face-type Yang–Baxter equation. A partial classification of the dynamical R matrices is given in Ref. 5. Recently, another type of integrable quantum systems, known as Ruijsenaars models,[6] and their various limiting cases have been shown to be connected to quantum groups and elliptic quantum groups[7–11] through different approaches.

The Calogero–Moser–Sutherland class of integrable models describes the motion of particles on a one-dimensional line or circle interacting via pairwise potentials that are given by Weierstrass elliptic functions and their various degenerations. The simplest case is an inverse r^2 potential. The Ruijsenaars–Schneider model is a relativistic generalization, whose quantum mechanical version, the Ruijsenaars model, is the model that we are interested in here.

The Hamiltonian of the Ruijsenaars model for two particles with coordinates x_1 and x_2 has the form

$$\mathcal{H}=\left\{\frac{\theta\big(c\,\eta/2-\lambda\big)}{\theta(-\lambda)}\,t_1^{(\lambda)}+\frac{\theta\big(c\,\eta/2+\lambda\big)}{\theta(\lambda)}\,t_2^{(\lambda)}\right\},$$

where $\lambda=x_1-x_2$. Here $c\in\mathbb{C}$ is the coupling constant, η is the relativistic deformation parameter, the θ function is given in (27); we have set $\hbar=1$. The Hamiltonian acts on a wave function as

$$\mathcal{H}\psi(\lambda)=\frac{\theta(c\,\eta/2-\lambda)}{\theta(-\lambda)}\,\psi(\lambda-\eta)+\frac{\theta(c\,\eta/2+\lambda)}{\theta(\lambda)}\,\psi(\lambda+\eta);$$

a)Electronic mail: jurco@CRM.UMontreal.CA
b)Electronic mail: schupp@princeton.edu

 © 1998 American Institute of Physics

the $t_i^{(\lambda)}$ that appear in the Hamiltonian are hence shift operators in the variable λ; in the present case of two particles they generate a one-dimensional graph:

Relative to a fixed vertex λ the vertices of this graph are at points $\eta \cdot \mathbb{Z} \in \mathbb{R}$, the (ordered) paths connect neighboring vertices and are hence intervals in $\mathbb{R}$. There are two paths per interval, one in the positive and one in the negative direction.

Now consider an N-particle Ruijsenaars model: The graph relative to a fixed λ is then a $N-1$-dimensional hypercubic lattice. Relative to $\lambda \in \mathbb{R}^{N-1}$, the vertices are at points $(\eta \cdot \mathbb{Z})^{\times(N-1)}$. This picture can obviously be generalized, and in the following we shall consider an arbitrary ordered graph.

Remark: For the Ruijsenaars system λ can take any value in $\mathbb{R}^{N-1}$, so we are *a priori* let to a huge graph that consists of a continuous family of disconnected graphs. Since the graphs are disconnected, it is sufficient to consider one. Later, we will, in fact, write all expressions with respect to one fixed vertex λ. This will lead to equations containing explicit shifts on the graph and thus to dynamical R-matrices. Apart from the physical interpretation there is *a priori* no reason to restrict the values of particle coordinates at the level of the shift operator to $\mathbb{R}$. We can and shall everywhere in the following take $\mathbb{C}$ instead of $\mathbb{R}$.

A. Face algebras

The Hilbert space of the model will be built from vector spaces on paths of fixed length. As we shall argue in the following, the operators on this Hilbert space can be chosen to be elements of a Face Algebra F[1] (or weak C^*-Hopf algebras,[12,13] which is essentially the same with a*-structure).

There are two commuting projection operators e^i, $e_i \in F$ for each vertex of the graph (projectors onto bra's and ket's corresponding to vertex i):

$$e_i e_j = \delta_{ij} e_i, \quad e^i e^j = \delta_{ij} e^i, \quad \Sigma e_i = \Sigma e^i = 1. \tag{1}$$

F shall be equipped with a coalgebra structure such that the combination $e^i_j \equiv e^i e_j = e_j e^i$ is a corepresentation:

$$\Delta(e^i_j) = \Sigma_k e^i_k \otimes e^k_j, \quad \epsilon(e^i_j) = \delta_{ij}. \tag{2}$$

It follows that $\Delta(1) = \Sigma_k e_k \otimes e^k \neq 1 \otimes 1$ (unless the graph has only a single vertex)—this is a key feature of face algebras (and weak C^*-Hopf algebras).

So far we have considered matrices with indices that are vertices, i.e., paths of length zero. In the given setting it is natural to also allow paths of fixed length on a finite oriented graph $\mathcal{G}$ as matrix indices. To illustrate this, here are some paths of length 3 on a graph that is a square lattice, e.g., on part of a graph corresponding to a particular Ruijsenaars system:

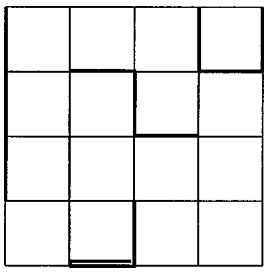

We shall use capital letters to label paths. A path P has an origin (source) $\cdot P$, an end (range) $P \cdot$, and a length $\sharp P$. Two paths Q, P can be concatenated to form a new path $Q \cdot P$, if the end of the first path coincides with the start of the second path, i.e., if $Q \cdot = \cdot P$ (this explains our choice of notation).

The important point is, that the symbols T^A_B, where $\sharp A=\sharp B\geqslant 0$, with relations

$$\Delta(T^A_B)=\sum_{A'} T^A_{A'}\otimes T^{A'}_B \quad (\sharp A=\sharp A'=\sharp B), \tag{3}$$

$$\epsilon(T^A_B)=\delta_{AB}, \tag{4}$$

$$T^A_B T^C_D=\delta_{A\cdot,\cdot C}\delta_{B\cdot,\cdot D}T^{A\cdot C}_{B\cdot D}, \tag{5}$$

span an object that obeys the axioms of a face algebra. Relations (3) and (4) make T^A_B a corepresentation; (5) is the rule for combining representations. The axioms of a face algebra can be found in Refs. 1, 14.

Pictorial representation:

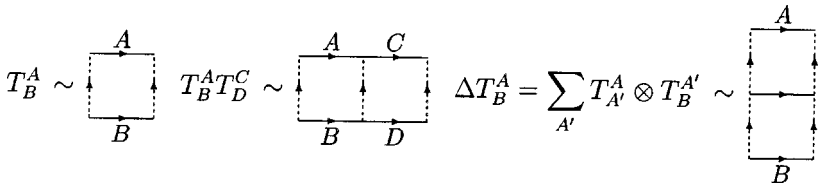

The dashed paths indicate the F space(s); their orientation is from the lower to upper ''index.'' Inner paths are summed over.

Often it is convenient not to consider a particular representation, i.e., T matrices corresponding to paths of a given fixed length but rather an abstract, universal T. Heuristically one can think of the universal T as an abstract matrix or group element, but it does, in fact, simply provide an alternative to the usual Hopf algebra notation: T, $T_1\otimes T_1$, T_1T_2, e.g., correspond to the identity map, coproduct map, and multiplication map of F, respectively. (More details are given in the Appendix.) Keeping this in mind we shall—without loss of generality—nethertheless, use a formal notation that treats T as if it was the canonical element T_{12} of $U\otimes F$, where U is the dual of F via the pairing $\langle,\rangle$

$$\langle T,f\rangle=f, \quad \langle T_1\otimes T_1,f\rangle=\Delta f, \quad \langle T_1T_2,f\otimes g\rangle=fg; \quad f,g\in F.$$

Here and in the following, we shall frequently suppress the the second index of T; it corresponds to the F-space. The displayed expressions are short for $\langle T_{12},f\otimes \mathrm{id}\rangle=f\in F$, $\langle T_{12}T_{13},f\otimes \mathrm{id}\otimes \mathrm{id}\rangle=\Delta f\in F\otimes F$, and $\langle T_{13}T_{23},f\otimes g\otimes \mathrm{id}\rangle=fg\in F$.

A face *Hopf* algebra has an antialgebra and anticoalgebra endomorphism, called the antipode and denoted by S or—in the universal tensor formalism—by $\tilde{T}$: $\langle\tilde{T},f\rangle=S(f)$. The antipode satisfies some compatibility conditions with the coproduct that is given in the Appendix.

Remark: In the limit of a graph with a single vertex a face Hopf algebra is the same as a Hopf algebra. Ordinary matrix indices correspond to closed loops in that case.

B. Boltzmann weights

By dualization we can describe a coquasitriangular structure of F by giving a quasitriangular structure for U. The axioms[1] for a quasitriangular face algebra are similar to those of a quasitriangular Hopf algebra; there is a universal $R\in U\otimes U$ that controls the noncocommutativity of the coproduct in U and the noncommutativity of the product in F,

$$RT_1T_2=T_2T_1R, \quad \tilde{R}T_2T_1=T_1T_2\tilde{R}, \quad \tilde{R}\equiv(S\otimes \mathrm{id})(R), \tag{6}$$

however, the antipode of R is no longer the inverse of R, but rather

$$\tilde{R}R=\Delta(1), \quad R\tilde{R}=\Delta'(1). \tag{7}$$

The numerical ''R matrix'' obtained by contracting R with two face corepresentations is given by the face Boltzmann weight W:

$$\langle \mathcal{R}, T^A_B \otimes T^C_D \rangle = R^{AC}_{BD} \equiv W\left(C \begin{smallmatrix} B \\ A \end{smallmatrix} D\right) \sim$$

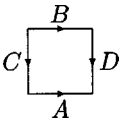

The pictorial representation makes sense since the Boltzmann weight is zero unless $C \cdot A$ and $B \cdot D$ are valid paths with a common source and range, as will be discussed in more detail below. Also note that $T^A_B \mapsto \langle \mathcal{R}, T^A_B \otimes T^C_D \rangle$ is a *representation* of the matrix elements of (T^A_B) while $T^C_D \mapsto \langle \mathcal{R}, T^A_B \otimes T^C_D \rangle$ is an *antirepresentation*. Consistent with our pictorial representation for the T matrices, we see that the orientation of the paths in F-space remain the same for the first case but are reversed for the latter:

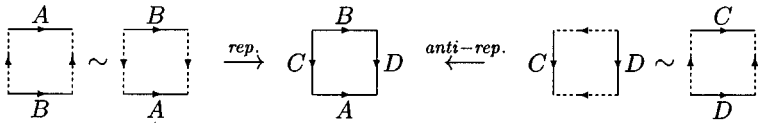

Definition: For $f \in F$ we can define two algebra homomorphisms $F \to U$:

$$R^+(f) = \langle R, f \otimes \mathrm{id} \rangle, \quad R^-(f) = \langle \tilde{R}, \mathrm{id} \otimes f \rangle. \tag{8}$$

Yang–Baxter Equation: As a consequence of the axioms of a quasitriangular face Hopf algebra, R satisfies the Yang–Baxter Equation,

$$R_{12} R_{13} R_{23} = R_{23} R_{13} R_{12} \in U \otimes U \otimes U. \tag{9}$$

Contracted with $T^A_B \otimes T^C_D \otimes T^E_F$, this expression yields a numerical Yang–Baxter equation with the following pictorial representation:[3]

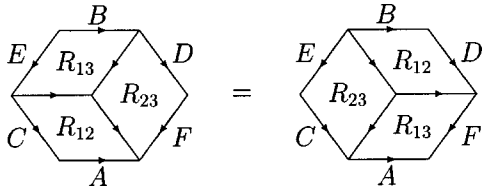

The inner edges are paths that are summed over. Moving along the outer edges of the hexagon, we will later read of the shifts in the Yang–Baxter equation for dynamical R matrices.

So far we have argued heuristically that the Ruijsenaars system naturally leads to graphs and face algebras. The formal relation between face Hopf algebras and oriented graphs, it is established by a lemma of Hayashi (Lemma 3.1 of Ref. 14), which says that any right or left comodule M of a face Hopf algebra F decomposes in a natural way as a linear space to a direct sum $M = \oplus_{i,j} M_{ij}$, with indices i, j running over all vertices. So we can naturally associate paths from j to i to any pair of indices, such that $M_{ij} \neq \varnothing$. So we may speak (and we really do in the paper) of the vectors of a comodule or a module as of paths. As the dual object U to a face Hopf algebra is again a face Hopf algebra characterized by the same set of vertices,[14] the same applies to its comodules. It is convenient to choose the orientation of paths appearing in the decomposition of a comodule of U (and hence in the module of F) opposite to the convention that one uses in the case of F.

Particularly, we have for any matrix corepresentation T^A_B of F with symbols A, B used to label some linear basis in M, the linear span $\langle T^A_B \rangle$ of all T^A_B decomposes as linear space (bicomodule of the face Hopf algebra) as a direct sum,

$$\bigoplus_{i,j,k,l} \langle e^i e_j T^A_B e^k e_l \rangle = \langle T^A_B \rangle,$$

i.e., a sum over paths with fixed starting and ending vertices. The upper indices i and k fix the beginning and the end of the path A and the lower indices j and l fix the beginning and the end of the path B.

Let us assume that the matrix elements T^A_B of a corepresentation of F act in a module of paths that we shall label by greek characters α, β, etc. The definition of the dual face Hopf algebra implies that the matrix element $(T^A_B)^\alpha_\beta$ is nonzero only if $\cdot\alpha=\cdot B$, $\cdot A=\alpha\cdot$, $B\cdot=\cdot\beta$, and $A\cdot=\beta\cdot$, i.e., if paths $\alpha\cdot A$ and $B\cdot\beta$ have common starting and end points. This justifies the pictorial representation used in the paper. It also follows immediately that in the case of a coquasitriangular Face Hopf algebra $\langle\mathcal{R},T^A_B\otimes T^C_D\rangle=R^{AC}_{BD}\equiv W(\,{}^B_{C\,D}{}_A)$ is zero unless $\cdot C=\cdot B$, $B\cdot=\cdot D$, $C\cdot=\cdot A$, and $A\cdot=D\cdot$.

To make contact with the Ruijsenaars type of models, we have to assume that the (coquasitriangular) face Hopf algebra F is generated by the matrix elements of some fundamental corepresentation of it. We shall postulate the paths of the corresponding corepresentation to be of length 1. The paths belonging to the n-fold tensor product of the fundamental corepresentation are then, by definition, of length n. Taking an infinite tensor product of the fundamental corepresentation, we get a graph that corresponds to the one generated by the shift operators of the related integrable model.

In the next section we are going to formulate a quantum version of the so-called main theorem, which gives the solution by factorization of the Heisenberg equations of motion. For this construction F needs to have a coquasitriangular structure—this will also fix its algebra structure.

II. QUANTUM FACTORIZATION

The cocommutative functions in F are of particular interest to us since they form a set of mutually commutative operators. We shall pick a Hamiltonian from this set. Cocommutative means that the result of an application of comultiplication Δ is invariant under exchange of the two resulting factors. The typical example is a trace of the T matrix. The following theorem gives for the case of face Hopf algebras what has become known as the ''main theorem'' for the solution by factorization of the equations of motion.[15–19] The following theorem is a direct generalization of our previous results for Hopf algebras/Quantum Groups.[20,21]

Theorem 0.1 (the main theorem for face algebras)**:**

(i) The set of cocommutative functions, denoted I, is a commutative subalgebra of F.

(ii) The Heisenberg equations of motion defined by a Hamiltonian $\mathcal{H}\in I$ are of Lax form,

$$i\,\frac{dT}{dt}=[M^\pm,T], \tag{10}$$

with $M^\pm=1\otimes\mathcal{H}-m_\pm\in U_\pm\otimes F$, $m_\pm=R^\pm(\mathcal{H}_{(2)})\otimes\mathcal{H}_{(1)}$; see (8).

(iii) Let $g_\pm(t)\in U_\pm\otimes F$ be the solutions to the factorization problem,

$$g_-^{-1}(t)g_+(t)=\exp(it(m_+-m_-))\in U\otimes F; \tag{11}$$

then

$$T(t)=g_\pm(t)T(0)g_\pm(t)^{-1} \tag{12}$$

solves the Lax equation (10); $g_\pm(t)$ are given by

$$g_\pm(t)=\exp(-it(1\otimes h))\exp(it(1\otimes h-M^\pm(0))), \tag{13}$$

and are the solutions to the differential equation

$$i\,\frac{d}{dt}\,g_\pm(t)=M^\pm(t)g_\pm(t),\quad g_\pm(0)=1. \tag{14}$$

Proof: The proof is similar to the one given in Ref. 21 for factorizable quasitriangular Hopf algebras. Here we shall only emphasize the points that are different because we are now dealing with face algebras. An important relation that we shall use several times in the proof is

$$\Delta(1)T_1T_2=T_1T_2=T_1T_2\Delta(1). \tag{15}$$

[Note that the generalization is not trivial since now $\Delta(1)\neq1\otimes1$ in general.]

(i) Let $f,g\in I\subset F$; then $fg\in I$. Let us show that f and g commute:

$$fg=\langle T_1T_2,f\otimes g\rangle=\langle\Delta(1)T_1T_2,f\otimes g\rangle=\langle\tilde{R}RT_1T_2,f\otimes g\rangle=\langle\tilde{R}T_2T_1R,f\otimes g\rangle=\langle T_2T_1R\tilde{R},f\otimes g\rangle$$
$$=\langle T_2T_1\Delta'(1),f\otimes g\rangle=\langle T_2T_1,f\otimes g\rangle=gf.$$

[$\tilde{R}\equiv(S\otimes\mathrm{id})(R)$ can be commuted with T_2T_1R in the fifth step because f and g are both cocommutative.]

(ii) We need to show that $[R^{\pm}(\mathcal{H}_{(2)})\otimes\mathcal{H}_{(1)},T]=0$. This follows from the cocommutativity of $\mathcal{H}$ and (6):

$$\langle T_1R^{\pm}_{21}T_2,\mathcal{H}\otimes\mathrm{id}\rangle=\langle R^{\pm}_{21}T_1T_2,\mathcal{H}\otimes\mathrm{id}\rangle=\langle T_2T_1R^{\pm}_{21},\mathcal{H}\otimes\mathrm{id}\rangle.$$

(iii) We need to show that $[m_+,m_-]=0$; then the proof of Ref. 21 applies,

$$m_+m_-=\langle T_1R_{13}T_2\tilde{R}_{32},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle\equiv\langle T_1T_2R_{13}\tilde{R}_{32},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle=\langle\tilde{R}_{12}T_2T_1R_{12}R_{13}\tilde{R}_{32},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle$$
$$=\langle\tilde{R}_{12}T_2T_1\tilde{R}_{32}R_{13}R_{12},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle=\langle R_{12}\tilde{R}_{12}T_2T_1\tilde{R}_{32}R_{13},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle$$
$$=\langle\Delta'(1)T_2T_1\tilde{R}_{32}R_{13},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle=\langle T_2\tilde{R}_{32}T_1R_{13},\mathcal{H}\otimes\mathcal{H}\otimes\mathrm{id}\rangle=m_-m_+.$$

Remark: The objects in this theorem $[M^{\pm},m_{\pm},g_{\pm}(t),T(t)]$ can be interpreted (a) as elements of $U\otimes F$, (b) as maps $F\to F$, or (c) when a representation of U is considered, as matrices with F-valued matrix elements.

III. DYNAMICAL OPERATORS

In the main theorem we dealt with expressions that live in $U\otimes F$ (and should be understood as maps from F into itself). In this section we want to write expressions with respect to one fixed vertex. Like we mentioned in the Introduction, we are particularly interested in the action of the Hamiltonian with respect to a fixed vertex. Since the Hamiltonian is an element of F, we shall initially fix the vertex in this space; due to the definition of the dual face Hopf algebra U this will also fix a corresponding vertex in that space. We shall proceed as follows: We will fix a vertex in the T-matrix with the help of $e^{\lambda},e_{\lambda}\in F$ and will also introduce the corresponding universal $T(\lambda)$. Next, we will construct dynamical R matrices $R^{\pm}(\lambda)$ as $R^{\pm}(T(\lambda))$—this is an algebra homomorphism—and will give the Yang–Baxter and RTT equations with shifts as an illustration. Finally, we shall plug everything into the main theorem.

Convention for corepresentation T with respect to a fixed vertex: $T(\lambda)^A_B$ is zero unless the range (end) of path A is equal to the fixed vertex λ. Such a T will map the vector space spanned by vectors v^A with $A\cdot=\lambda$ fixed to itself. With the help of the projection operator $e^{\lambda}\in F$ we can give the following explicit expression:

$$T(\lambda)^A_B \;=\; T^A_B\,e^{\lambda} \quad\sim$$

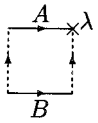

(16)

The universal $T(\lambda)$ is an abstraction of this and is defined analogously,

$$T_{12}(\lambda) \;=\; T_{12}\,(e^{\lambda})_2 \quad\sim$$

(17)

Coproducts of T: Expressions for the coproducts $\Delta_1T(\lambda)$ and $\Delta_2T(\lambda)$ follow either directly from the definition of $T(\lambda)$ or can be read of the corresponding pictorial representations.

(i) Coproduct in F-space:[2]

$$\Delta_2 T(\lambda) \;=\; T_{12}(\lambda)T_{13}(\lambda-h_2) \quad\sim\quad$$

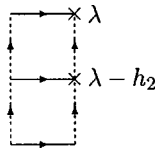

(18)

The shift operator h in F-space that appears here is

$$h_{(F)}=\sum_{\eta,\mu}\,(\mu-\eta)e^{\mu}_{\eta}\in F, \tag{19}$$

where we assume some appropriate (local) embedding of the vertices of the underlying graph in $\mathbb{C}^n$ so that the difference of vertices makes sense.

Proof: $\Delta_2 T_{12}(e^{\lambda})_2=\Sigma_{\eta}T_{12}T_{13}(e^{\lambda}_{\eta})_2(e^{\eta})_3=\Sigma_{\eta,\mu}T_{12}(e^{\lambda}e^{\mu})_2(e_{\eta})_2T_{13}(e^{\eta})_3=\Sigma_{\eta,\mu}T_{12}(\lambda)\times(e^{\mu}_{\eta})_2T_{13}(\lambda+\eta-\mu)=T_{12}(\lambda)T_{13}(\lambda-h_2)$. In the second and third step we used $e^{\lambda}e^{\mu}\propto\delta_{\lambda,\mu}$.

(ii) Coproduct in U-space (gives multiplication in F):

$$\Delta_1 T(\lambda) \;=\; T_{13}(\lambda-h_2)T_{23}(\lambda) \quad\sim\quad$$

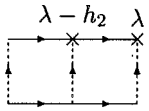

(20)

This time we have a shift operator in U-space:

$$h_{(U)}=\sum_{\eta,\mu}\,(\mu-\eta)E_{\mu}E^{\eta}\in U. \tag{21}$$

The proof is a little more involved than the one given for the F-case above and uses $(e^{\eta})_2T_{12}=(E^{\eta})_1T_{12}$ and $T_{12}(e^{\mu})_2=(E_{\mu})_1T_{12}$, which follow from $T\tilde{T}T=T$, $\tilde{T}T\tilde{T}=\tilde{T}$ and $T\tilde{T}=\Sigma_{\xi}E^{\xi}\otimes e^{\xi}$.

Dynamical R-matrix: Using the fact that $f\mapsto R^{+}(f)\equiv\langle R,f\otimes\mathrm{id}\rangle$ is an algebra homomorphism, we define

$$R_{12}(\lambda)\equiv\langle R,T_1(\lambda)\otimes T_2\rangle. \tag{22}$$

The numerical R-matrix is defined analogously: $R(\lambda)^{AC}_{BD}=\langle R,T(\lambda)^{A}_{B}\otimes T^{C}_{D}\rangle$. In the pictorial representation this fixes the one vertex of R that is only the end point to paths.

Dynamical RTT equation: From the coproduct $\Delta_1 T$ and $R\Delta(x)=\Delta'(x)R$ for all $x\in U$ follows:[2]

$$R_{12}(\lambda)T_1(\lambda-h_2)T_2(\lambda)=T_2(\lambda-h_1)T_1(\lambda)R_{12}(\lambda-h_3). \tag{23}$$

Pictorially,

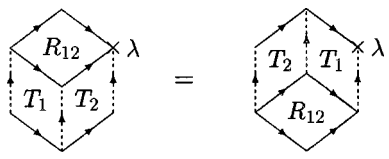

Shifts h_1,h_2 are in U-space, shift h_3 is in F-space. Twice contracted with R, the dynamical RTT-equation yields the dynamical Yang–Baxter equation:

Dynamical Yang–Baxter equation:[4]

$$R_{12}(\lambda)R_{13}(\lambda-h_2)R_{23}(\lambda)=R_{23}(\lambda-h_1)R_{13}(\lambda)R_{12}(\lambda-h_3). \tag{24}$$

Pictorially,

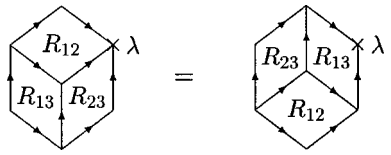

Hamiltonian and Lax operators in dynamical setting: The Hamiltonian $\mathcal{H}$ should be a cocommutative element of F. In the case of the Ruijsenaars model, it can be chosen to be the trace of a T-matrix, i.e., the U-trace of T in an appropriate representation ρ: $\mathcal{H}=\mathrm{tr}_1^{(\rho)}T_1$. This can be written as a sum over vertices λ of operators that act in the respective subspaces corresponding to paths ending in the vertex λ:

$$\mathcal{H}=\sum_{Q,\#Q\text{fixed}} T_Q^Q=\sum_{\lambda}\mathcal{H}(\lambda), \tag{25}$$

with $\mathcal{H}(\lambda)=\mathcal{H}e^{\lambda}=\Sigma T(\lambda)_Q^Q$. The pictorial representation of the Hamiltonian is two closed dashed paths (F-space) connected by paths Q of fixed length that are summed over. In $\mathcal{H}(\lambda)$, the end of path Q is fixed. When we look at a representation on Hilbert space, the paths Q with end point λ that appear in the component $\mathcal{H}(\lambda)$ of the Hamiltonian $\mathcal{H}$ will shift the argument of a state $\psi(\lambda)$ corresponding to the vertex λ to a new vertex, corresponding to the starting point of the path Q. In the next section we will see in detail how this construction is applied to the Ruijsenaars system. For this we will have to take a representation of the face Hopf algebra F. We shall then denote the resulting Lax operator by L. The Hamiltonian will contain a sum (coming from the trace) over shift operators.

For the Lax operators it is convenient to fix two vertices, λ and μ, corresponding to paths from μ to λ. (We did not need to do this for the Hamiltonian since it acts only on closed paths.) The T that appears in the Lax equation (10) becomes

$$T_{12}(\lambda,\mu)=T_{12}(e_\mu e^{\lambda})_2=(E_\lambda)_1T_{12}(E_\mu)_1;$$

it should be taken in some representation of F on the appropriate Hilbert space. On space U we are interested in a finite-dimensional representation that is going to give us matrices. Both is done in the next section and we shall call the resulting operator $L(\lambda,\mu)$. Similarly, we proceed with the other two Lax operators $M^{\pm}$:

$$M^{\pm}(\lambda,\mu)\equiv(E_\lambda\otimes 1)M^{\pm}(E_\mu\otimes 1)=E_\lambda\delta_{\lambda,\mu}\otimes\mathcal{H}-m_{\pm}(\lambda,\mu).$$

(Recall that $E_\lambda E_\mu=E_\lambda\delta_{\lambda,\mu}$.) If we define $M_{012}^{\pm}=T_{02}(1-R_{01}^{\pm})$, then $M_{12}^{\pm}=\langle M_{012}^{\pm},\mathcal{H}\otimes\mathrm{id}^2\rangle$ and we can write dynamical Lax equations in an obvious notation as

$$i\frac{dT(\lambda,\mu)}{dt}=\langle M_{012}^{\pm}(\lambda,\lambda-h_0)T_{12}(\lambda-h_0,\mu)-T_{12}(\lambda,\mu-h_0)M_{012}^{\pm}(\mu-h_0,\mu),\mathcal{H}\otimes\mathrm{id}^2\rangle, \tag{26}$$

where $h_0=\Sigma_{\alpha,\beta}(\alpha-\beta)E_\alpha^\beta$ is the shift operator in the space contracted by the Hamiltonian. If the Hamiltonian is a trace we shall find a sum over shifts.

Remark: There is another possible choice of conventions for the fixed vertex. We could have worked with $T_{12}\{\nu\}=(e_\nu)_2T_{12}$ instead of $T_{12}(\lambda)=T_{12}(e^{\lambda})_2$. This would have fixed the lower left vertex in the pictorial representation of T_B^A and the vertex that is *starting* point for all paths in $R\{\nu\}$. The dynamical equations would, of course, look a little different from the ones that we have given.

IV. THE LAX PAIR

Here we give as an example the Lax pair for the case of the N-particle quantum Ruijsenaars model. We may think of the face Hopf algebra F as of the elliptic quantum group associated to $sl(N)$ introduced by Felder.[2] We will continue to use the symbol F for it. In that case there is an additional spectral parameter entering all relations in the same way as it is in the case of ordinary quantum groups.

Let h be the Cartan subalgebra of $sl(N)$ and h^* its dual. The actual graph related to the elliptic quantum group F is $h^* \sim \mathbb{C}^{N-1}$. This is the huge graph of the remark made in the Introduction. However, it decomposes into a continuous family of disconnected graphs, each one isomorphic to $-\eta.\Lambda$, the $-\eta \in \mathbb{C}$ multiple of the weight lattice Λ of $sl(N)$, and we can restrict ourselves to this one component for simplicity. Correspondingly, the shifts in all formulas are rescaled by a factor $-\eta$. The space $h^* \sim \mathbb{C}^{N-1}$ itself will be considered as the orthogonal complement of $\mathbb{C}^N = \oplus_{i=1,\ldots,N} \mathbb{C}\varepsilon_i$, $\langle \varepsilon_i, \varepsilon_j \rangle = \delta_{ij}$ with respect to $\Sigma_{i=1,\ldots,N}\varepsilon_i$. We write the orthogonal projection $\epsilon_i = \varepsilon_i - 1/N \Sigma_k \varepsilon_k$ for the generator of h^*; $\langle \epsilon_i, \epsilon_j \rangle = \delta_{ij} - 1/N$. The points of the lattice $\eta.\Lambda$ will be denoted by greek characters λ, μ, etc.

The elliptic quantum group is defined by the matrix elements of the ''fundamental corepresentation.'' This is described with the help of the paths of length 1 in the following way: Let us associate a one-dimensional linear space $V_{\rho,\lambda} \sim \mathbb{C}\eta.\epsilon_k$ and a corresponding path of length 1 to any pair $\lambda, \rho \in \eta.\Lambda$, such that $\rho - \lambda = \eta\epsilon_k$, for some $k = 1,\ldots,N$. We let $V_{\rho,\lambda} = \varnothing$ for all other pairs of vertices. The vector space V of the fundamental corepresentation is formed by all paths of length 1,

$$V = \bigoplus_{\lambda,\rho} V_{\rho,\lambda} \quad \sim \quad \bigoplus_{\lambda,\rho} \rho \longleftarrow \lambda \quad = \quad \bigoplus_{\lambda,i} \lambda + \eta\epsilon_i \longleftarrow \lambda$$

As all spaces $V_{\rho,\lambda}$ are at most one dimensional, we can characterize the numerical R-matrix R^{AC}_{BD} in the fundamental corepresentation by just four indices referring to the vertices of the ''square'' defined by paths A,B,C,D in the case of nonzero matrix element R^{AC}_{BD}. Let us set $\cdot B = \cdot C = \nu$, $\cdot D = B\cdot = \mu$, $D\cdot = A\cdot = \lambda$ and $\cdot A = C\cdot = \rho$, and also

$$R^{AC}_{BD} = W\left(C \begin{matrix} B \\ A \end{matrix} D\right) \equiv W\begin{pmatrix} \nu & \mu \\ \rho & \lambda \end{pmatrix} \quad \sim \quad \begin{array}{ccc} \nu & \longrightarrow & \mu \\ \uparrow & & \downarrow \\ \rho & \longrightarrow & \lambda \end{array}.$$

Then the nonzero Boltzmann weights as given by Ref. 22 are ($i \neq j$)

$$W\begin{pmatrix} \lambda + 2\eta\epsilon_i & \lambda + \eta\epsilon_i \\ \lambda + \eta\epsilon_i & \lambda \end{pmatrix} \Big| u = 1 \quad \sim \quad \Longrightarrow \lambda,$$

$$W\begin{pmatrix} \lambda + \eta(\epsilon_i + \epsilon_j) & \lambda + \eta\epsilon_i \\ \lambda + \eta\epsilon_i & \lambda \end{pmatrix} \Big| u = \frac{\theta(\eta)\theta(-u + \lambda_{ij})}{\theta(u + \eta)\theta(\lambda_{ij})} \quad \sim \quad \begin{array}{cc} \Downarrow & \\ \Longrightarrow & \lambda \end{array},$$

and

$$W\begin{pmatrix} \lambda + \eta(\epsilon_i + \epsilon_j) & \lambda + \eta\epsilon_j \\ \lambda + \eta\epsilon_i & \lambda \end{pmatrix} \Big| u = \frac{\theta(u)\theta(\eta + \lambda_{ij})}{\theta(u + \eta)\theta(\lambda_{ij})} \quad \sim \quad \begin{array}{cc} \rightarrow & \\ \square & \\ & \lambda \end{array}.$$

Here $\lambda_{ij} \equiv \lambda_i - \lambda_j = \langle \lambda, \epsilon_i - \epsilon_j \rangle$, $\theta(u)$ is the Jacobi theta function,

$$\theta(u) = \sum_{j \in \mathbb{Z}} e^{\pi i (j+1/2)^2 \tau + 2\pi i (j+1/2)(u+1/2)}, \tag{27}$$

$u \in \mathbb{C}$ is the spectral parameter, and τ is the elliptic modulus parameter. The matrix element $T^A_B(\lambda|u)$ (now depending also on the spectral parameter u, which also enters all previous expressions in the standard way) in the fundamental corepresentation, is also uniquely determined by the value of the vertices that fix the length-1 paths A and B. Let us use the following notation for it

$(i,j=1,...,N)$:

$$T_B^A(\lambda|u)=\sum_\mu L_j^i(\mu,\lambda|u)e_\mu \quad \sim$$

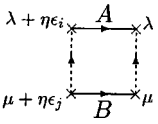

where $\mu=B\cdot$, $\lambda=A\cdot$, $\mu+\eta\epsilon_j=\cdot B$, $\lambda+\eta\epsilon_i=\cdot A$.

As the lattice $\eta.\Lambda$ that we consider is just one connected component of a continuous family of disconnected graphs, the vertices $\lambda,\mu,...$, are allowed to take any values in $\mathbb{C}^{(N-1)}$. This will be assumed implicitly in the rest of this section. Now we need to specify the appropriate representation of F that can be read from Ref. 7. We can characterize it by its path decomposition. To any pair of vertices λ, μ we associate a one-dimensional vector space $\tilde{V}_{\lambda,\mu}\sim\mathbb{C}$ (the path from μ to λ). The representation space $\tilde{V}$ is then

$$\tilde{V}=\bigoplus_{\lambda,\mu\in h^*}\tilde{V}_{\lambda,\mu}.$$

The matrix element $T_B^A(\lambda)\equiv L_j^i(\lambda,\mu|u)$ for fixed A, B and hence also with fixed i,j,λ,μ is obviously nonzero only if restricted to act from $\tilde{V}_{\lambda,\mu}$ to $\tilde{V}_{\lambda+\eta\epsilon_i,\mu+\eta\epsilon_j}$, in which case it acts as multiplication by

$$L_j^i(\lambda,\mu|u)=\frac{\theta\big(c\,\eta/N+u+\lambda_i-\mu_j\big)}{\theta(u)}\prod_{k\neq i}\frac{\theta\big(c\,\eta/N+\lambda_k-\mu_j\big)}{\theta(\lambda_k-\lambda_i)}. \tag{28}$$

Here we used notation $\lambda_i=\langle\lambda,\epsilon_i\rangle$ for $\lambda\in h^*$. $c\in\mathbb{C}$ will play the role of the coupling constant.

The Hamiltonian is chosen in accordance with Sec. III as $\mathcal{H}=\Sigma_P T_P^P$, i.e., the trace in the ''fundamental'' corepresentation. In accordance with the discussion of the previous sections, it is nonzero only when acting on the diagonal subspace (closed paths),

$$H=\bigoplus_{\lambda\in h^*}H_\lambda\equiv\bigoplus_{\lambda\in h^*}\tilde{V}_{\lambda,\lambda},$$

of $\tilde{V}$. So this is the actual state space of the integrable system under consideration. The Lax operator $M^\pm$ acts from $\tilde{V}_{\lambda,\mu}$ to $\tilde{V}_{\lambda+\eta\epsilon_i,\mu+\eta\epsilon_j}$ as multiplication by $M_k^{\pm l}(\lambda,\mu|u,v)=(1\otimes\mathcal{H})_k^l(\lambda,\mu|v)-m_{\pm k}^l(\lambda,\mu|u,v)$ with

$$(1\otimes\mathcal{H})_k^l(\lambda,\mu|v)=\delta_{\lambda,\mu}\delta_k^l\frac{\theta\big(c\,\eta/N+v\big)}{\theta(v)}\prod_{j'\neq i}\frac{\theta\big(c\,\eta/N+\lambda_{j',i}\big)}{\theta(\lambda_{j',i})}, \tag{29}$$

$$m_k^{+l}(\lambda,\mu|u,v)=\delta_{\epsilon_i+\epsilon_k,\epsilon_j+\epsilon_l}\delta_{\lambda,\mu+\eta\epsilon_j}\sum_{i,j}L_j^i(\mu+\eta\epsilon_k,\ \mu|v)W\left(\begin{matrix}\lambda+\eta\epsilon_l & \lambda\\ \mu+\eta\epsilon_k & \mu\end{matrix}\middle|v-u\right) \tag{30}$$

$$\sim$$

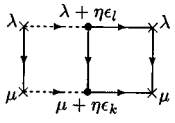

m^- is given by a similar formula with the inverse Boltzmann weight.

Dynamical Lax Equation:

$$i\frac{dL_k^i(\lambda,\mu|u)}{dt}=\sum_{j,\nu}M_j^{\pm i}(\lambda,\nu|u,v)L_k^j(\nu,\mu|u)-L_j^i(\lambda,\nu|u)M_k^{\pm j}(\nu,\mu|u,v).$$

The Hamiltonian $\mathcal{H}$ maps the component $H_\lambda \sim \mathbb{C}.|\lambda\rangle$ into the component $H_{\lambda+\eta\epsilon_i} \sim \mathbb{C}.|\lambda + \eta\epsilon_i\rangle$. Obviously, $\tilde{V}$ can be understood as the complex vector space of all functions in λ, μ as well as H can be understood as the complex vector space of all functions in λ. In that case the Hamiltonian $\mathcal{H}$ given by (29) is proportional to a difference operator in variable $\lambda \in h^*$,

$$\mathcal{H} \propto \sum_i t_i^{(\lambda)} \prod_{j \neq i} \frac{\theta(c\,\eta/N + \lambda_{j,i})}{\theta(\lambda_{j,i})}, \tag{31}$$

where $t_i^{(\lambda)}$ has an obvious meaning of the shift operator by $-\eta\epsilon_i$ in the variable λ. This is equivalent[23] to the Ruijsenaars Hamiltonian.[6] It follows from Ref. 7 that in the same way we can obtain the higher-order Hamiltonians concerning traces in properly fused ''fundamental representations.''

ACKNOWLEDGMENTS

We would like to thank Pavel Winternitz for constant interest and support, Koji Hasegawa and Jan Felipe van Diejen for interesting discussions, and Elliott Lieb for hospitality.

APPENDIX: UNIVERSAL *T*

For many reasons it is very convenient to use a formalism based on the so-called universal T. The expressions formally resemble those in a matrix representation but give, nevertheless, general face Hopf algebra statements. This greatly simplifies notation but also the interpretation and application of the resulting expressions. When we are dealing with quantizations of the functions on a group, we need to keep track both of the noncommutativity of the quantized functions and the residue of the underlying group structure. Both these structures can be encoded in algebraic relations for the universal T, which easily allows us to control two noncommutative structures. In fact, T can be regarded as a universal group element. Universal tensor expressions can *formally* be read in two ways: Either as ''group'' operations (or rather operations in U) or as the corresponding pull-back maps in the dual space. This simplifies the heuristics of ''dualizing and reversing arrows'' and allows us to keep track of the classical limit. Example: Multiplication in U, $x \otimes y \mapsto xy$: The corresponding pull-back map in the dual space F is the coproduct Δ. Both operations are summarized in the same universal expression $T_{12}T_{13}$.

Sometimes T can be realized as the canonical element $U \otimes F$, but we do not need to limit ourselves to these cases and will instead define T as the identity map from F to itself and will use the same symbol for the identity map $U \to U$.

For the application of this identity map to an element of F we shall nevertheless use the same bracket notation as we would for a true canonical element in the finite case, i.e., $f \equiv \langle \mathrm{id}, f\rangle = \langle T_{12}, f \otimes \mathrm{id}\rangle$. This notation is very convenient—like inserting the unity in quantum mechanics. The identity map on a product is given by $ab \mapsto a \cdot b$ or

$$\langle T_{12}, ab \otimes \mathrm{id}\rangle = \langle T_{13}T_{23}, a \otimes b \otimes \mathrm{id}\rangle = a \cdot b.$$

We shall write $\Delta_1 T_{12} = T_{13}T_{23}$ to express this fact—this is hence the universal-T notation for the multiplication map in F (and the coproduct map in U). The coproduct map in F (multiplication map in U) is $T_{12}T_{13}$:

$$f \mapsto \langle T_{12}T_{13}, f \otimes \mathrm{id} \otimes \mathrm{id}\rangle = \Delta(f).$$

The antipode map is $\tilde{T}: f \mapsto S(f)$, the contraction map $x \otimes 1$ with $x \in U = F^*$ maps $f \in F$ to $1\langle x, f\rangle$, the counit map is $1 \otimes 1$, etc. For face algebras the counit is not an algebra homomorphism, but rather $\Delta 1_F = \Sigma_k e_k \otimes e^k$ and $\Delta 1_U = \Sigma_k E_k \otimes E^k$, where $\langle E_k, f\rangle = \epsilon(fe_k)$ and $\langle E^k, f\rangle = \epsilon(e^k f)$. Therefore $\epsilon(ab) = \Sigma_k \epsilon(ae_k)\epsilon(e^k b)$. This is one of the face algebra axioms. It differs from the corresponding ordinary Hopf algebra axiom [there: $\epsilon(ab) = \epsilon(a)\epsilon(b)$]. Some important relations involving the coproduct and antipode maps are

$$\tilde{T}T = \sum_i E_i \otimes e_i, \quad T\tilde{T} = \sum_i E^i \otimes e^i, \quad \tilde{T}T\tilde{T} = \tilde{T}, \quad T\tilde{T}T = T.$$

In the pictorial representation these relations imply that the vertices of the paths in U and F that form T match to give a closed ''square.'' They also give us two ways to fix the four vertices of T:

$$(1\otimes e^i_k)T(1\otimes e^j_l) \;=\; (E^i_j\otimes 1)T(E^k_l\otimes 1) \quad \sim$$

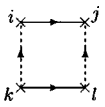

Further useful relations can be obtained from this by summing over some of the indices and using $\Sigma_i e^i_j = e_j$, $\Sigma_j e_j = 1$, etc.

[1] T. Hayashi, ''Quantum group symmetry of partition functions of IRF models and its application to Jones' index theory,'' Commun. Math. Phys. **157**, 331 (1993).
[2] G. Felder, ''Elliptic quantum groups,'' in *Proceedings of the XIth International Congress of Mathematical Physics*, edited by D. Iagolnitzer (International Press, Boston, 1995).
[3] R. J. Baxter, *Exactly Solved Models in Statistical Physics* (Academic, New York, 1982).
[4] J. L. Gervais and A. Neveu, ''Novel triangle relation and absence of tachyons in Liouville string field theory,'' Nucl. Phys. B **238**, 125 (1984).
[5] P. Etingof and A. Varchenko, ''Solutions of the quantum dynamical Yang–Baxter equation and dynamical quantum groups,'' q-alg/9708015.
[6] S. N. M. Ruijsenaars, ''Complete integrability of relativistic Calogero–Moser systems and elliptic functions identities,'' Commun. Math. Phys. **110**, 191 (1987).
[7] K. Hasegawa, ''Ruijsenaars' commuting difference operators as commuting transfer matrices,'' Commun. Math. Phys. **187**, 289 (1997).
[8] G. Felder and A. Varchenko, ''Elliptic quantum groups and Ruijsenaars models,'' q-alg/9704005.
[9] J. Avan, O. Babelon, and E. Billey, ''The Gervais–Neveu–Felder equation and the quantum Calogero–Moser model,'' Commun. Math. Phys. **178**, 281 (1996).
[10] G. E. Arutyunov, L. O. Chekhov, and S. A. Frolov, ''R-matrix quantization of the elliptic Ruijsenaars-Schneider model,'' q-alg/9612032.
[11] P. I. Etingof and A. A. Kirillov, ''Macdonald's polynomials and representations of quantum groups,'' Math. Res. Lett. **1**, 279 (1994).
[12] G. Mack and V. Schomerus, ''Quasi Hopf quantum symmetry in quantum theory,'' Nucl. Phys. B **370**, 185 (1992).
[13] G. Böhm and K. Szlachányi, ''A coassociative C^*-quantum group with nonintegral dimensions,'' Lett. Math. Phys. **35**, 437 (1996).
[14] T. Hayashi, ''Compact quantum groups of face type,'' Publ. Res. Inst. Math. Sci. **32**, 351 (1996).
[15] M. Adler, ''On a trace formula for pseudo-differential operators and the symplectic structure of the KdV-type equations,'' Invent. Math. **50**, 219 (1979).
[16] B. Kostant, ''The solution to a generalized Toda lattice and representation theory,'' Adv. Math. **34**, 195 (1979).
[17] W. Symes, ''System of Toda type, inverse spectral problems, and representation theory,'' Invent. Math. **159**, 13 (1980).
[18] A. G. Reyman and M. A. Semenov-Tian-Shansky, ''Group theoretical methods in the theory of finite-dimensional integrable systems,'' in *Encyclopedia of Mathematical Sciences*, Dynamical systems VII (Springer-Verlag, Berlin, 1993).
[19] N. Reshetikhin, ''Integrable discrete systems,'' in *Quantum Groups and their Applications in Physics*, International School of Physics ''Enrico Fermi,'' Varenna, 1994, edited by L. Castellani and J. Wess (Società Italiana di Fisica, Bologna, 1995).
[20] B. Jurčo and M. Schlieker, ''Quantized Lax equations and their solutions,'' Commun. Math. Phys. **185**, 397 (1997).
[21] B. Jurčo and P. Schupp, ''Twisted quantum Lax equations,'' Int. J. Mod. Phys. A **32**, 5735 (1997).
[22] M. Jimbo, T. Miwa, and M. Okado, ''Solvable lattice models whose states are dominant weights are dominant integral weights of $A^{(1)}_{n-1}$,'' Lett. Math. Phys. **14**, 123 (1987); ''Local state probabilities of solvable lattice models: An $A^{(1)}_n$ family,'' Nucl. Phys. B **300**, 74 (1988).
[23] J. F. van Diejen, ''Integrability of difference Calogero–Moser systems,'' J. Math. Phys. **35**, 2983 (1994).

On infravacua and superselection theory

Walter Kunhardt[a)]
Institut für Theoretische Physik der Universität Göttingen, Bunsenstrasse 9, 37073 Göttingen, Germany

(Received 9 September 1997; accepted for publication 18 March 1998)

In the Doplicher, Haag, and Roberts theory of superselection sectors, one usually considers states which are local excitations of some vacuum state. Here, we extend this analysis to local excitations of a class of ''infravacuum'' states appearing in models with massless particles. We show that the corresponding superselection structure, the statistics of superselection sectors and the energy-momentum spectrum are the same as with respect to the vacuum state. (The latter result is obtained with a novel method of expressing the shape of the spectrum in terms of properties of local charge transfer cocycles.) These findings provide evidence to the effect that infravacua are a natural starting point for the analysis of the superselection structure in theories with long-range forces. © *1998 American Institute of Physics.* [S0022-2488(98)01107-4]

I. INTRODUCTION

The description of the superselection structure of theories with long-range interactions such as quantum electrodynamics (QED) is a long-standing problem in the algebraic approach to quantum field theory. As a characteristic feature, such theories contain charges which obey Gauss's law and thus do not fit into the framework of superselection sectors as described by Doplicher, Haag, and Roberts.[1]

Focusing (mainly for definiteness) on QED, Gauss's law has the following three consequences: First, since every charge is accompanied by an electromagnetic field, an operator ψ_c which describes the creation of an electric charge in a gauge-invariant way in the ''physical'' Hilbert space, cannot be local. There is, however, the possibility of localizing the charge and (the Cauchy data of) its field in a spacelike cone $\mathscr{C}$ (as a simple example of a region extending to spacelike infinity). In this case, the charge cannot be detected by measurements in the spacelike complement $\mathscr{C}'$ of that cone.

Second, the configuration of the asymptotic electric flux at spacelike infinity is a classical observable whose values label an uncountable number of superselection sectors in the physical Hilbert space. Among these is the vacuum sector, but also the sector generated from the vacuum by the operator ψ_c. The former is identified with the vacuum representation π_0 of the observables, the latter with the representation $\pi_0 \circ \gamma_{\mathscr{C}}$, where $\gamma_{\mathscr{C}} \equiv \mathrm{Ad}\psi_c^*$. Obviously, the spacelike direction of $\mathscr{C}$ can be determined by measurements of the asymptotic flux distribution in the sector $\pi_0 \circ \gamma_{\mathscr{C}}$. This implies that the representations π_0 and $\pi_0 \circ \gamma_{\mathscr{C}}$, equal by construction when restricted to $\mathscr{C}'$, will not be equivalent in restriction to the spacelike complement $\mathscr{C}_1'$ of some arbitrary spacelike cone $\mathscr{C}_1$.

Third, Gauss's law lies at the origin of what is called the infraparticle problem: electrons are infraparticles, i.e., they do not correspond to eigenvalues of the mass operator. The physical picture for this is that every electron is accompanied by a cloud of low-energy photons. Related to this is the emission of bremsstrahlung in collisions of charged particles. The creation of infinitely many photons in such a process manifests itself on the mathematical side in the fact that the representations of the incoming and the outgoing free photon fields cannot both be Fock representations. However, this does not preclude that both representations can be equivalent, as is indeed the case for certain representations constructed by Kraus, Polley, and Reents.[2] Therefore these so-called KPR representations are interpreted as a background radiation field which is

a)Electronic mail: kunhardt@theorie.physik.uni-goettingen.de

 © 1998 American Institute of Physics

sufficiently chaotic in the sense that the addition of bremsstrahlung only amounts to a slight perturbation of them.

Motivated by this picture, Buchholz has pointed out in Ref. 3 that in front of such a background the spacelike direction of the localization cone $\mathscr{C}$ of ψ_c should not be detectable any more. In mathematical terms, the representations π_I and $\pi_I \circ \gamma_{\mathscr{C}}$ (where π_I corresponds to the KPR-like background) are expected to satisfy the so-called BF (Buchholz–Fredenhagen) criterion, i.e., they should become unitarily equivalent in restriction to the spacelike complement of *any* spacelike cone $\mathscr{C}_1$, in contrast to the situation prevailing in front of the vacuum.

Thus, trading the vacuum π_0 for a background π_I is expected to lead to an important improvement of the localizability properties of the charged states and should therefore permit one to carry through the BF variant of the Doplicher, Haag, and Roberts (DHR) analysis[4,5] in this case too.

In this perspective, the aim of the present contribution is to provide a consistency check of these ideas for the case of (massless) theories with DHR-like charges. We will show that in such theories, a DHR analysis can be based on a wide class of background representations π_I and that the main results of this analysis are invariant with respect to the choice of that background. More importantly, also the sets of sectors satisfying Borchers's criterion are invariant.

These results provide a justification for viewing such representations as a background and corroborate the hope that, in theories with long-range interactions, the superselection structure may indeed be described more easily in relation to such a background rather than to the vacuum. Recent results[6] obtained in the framework of the simple model of Ref. 7 also point into this direction.

We end this introduction by explaining the main assumptions and some notation. The starting point of the following analysis will be a quantum field theory given in its vacuum representation. More specifically, let $(\mathscr{H}_0, U_0, \Omega_0)$ be a vacuum Hilbert space, that is a Hilbert space $\mathscr{H}_0$ which carries a strongly continuous unitary representation U_0 of the translation group $\mathbf{R}^{1+s}$ (of $(1+s)$-dimensional Minkowski space, also denoted by $\mathbf{R}^{1+s}$) whose spectrum is contained in the forward lightcone $\overline{\mathbf{V}_+} \subset \mathbf{R}^{1+s}$ and such that the translation invariant subspace of $\mathscr{H}_0$ is spanned by a single unit vector Ω_0. Assume further that we are given a Haag–Kastler net

$$\mathscr{O} \mapsto \mathscr{A}(\mathscr{O})$$

of von Neumann algebras on $\mathscr{H}_0$, i.e., an isotonous mapping from the bounded regions $\mathscr{O} \subset \mathbf{R}^{1+s}$ to the von Neumann subalgebras of $\mathscr{B}(\mathscr{H}_0)$ which satisfies locality [i.e., $\mathscr{A}(\mathscr{O}_1) \subset \mathscr{A}(\mathscr{O}_2)'$ if $\mathscr{O}_1 \subset \mathscr{O}_2'$, in standard notation], translation covariance with respect to $\alpha_x := \mathrm{Ad}U_0(x)$ [i.e., $\alpha_x(\mathscr{A}(\mathscr{O})) = \mathscr{A}(\mathscr{O}+x)$] and weak additivity. The quasilocal algebra $\overline{\cup_{\mathscr{O}} \mathscr{A}(\mathscr{O})}^{\|\cdot\|}$ is also denoted by $\mathscr{A}$ and is assumed to act irreducibly. It will be of crucial importance later on that one can conclude from these assumptions that the net $\mathscr{A}$ has Borchers's property B and that the C^*-algebra $\mathscr{A}$ is simple.[8] (These conclusions do not, in fact, rely on the existence and/or uniqueness of the vacuum vector Ω_0.)

It is customary to introduce the following notations at this point: the identical representation $\pi_0 : \mathscr{A} \to \mathscr{B}(\mathscr{H}_0)$ is called the vacuum representation; it is the Gel'fand–Naïmark–Segal (GNS) representation of the vacuum state $\omega_0 = (\Omega_0, \cdot\, \Omega_0)$. The spectral family associated with the translations U_0 is denoted by E_0.

II. ENERGY COMPONENTS

In this section, we will present certain folia of states, namely energy components of positive energy representations. This notion has been introduced by Borchers and Buchholz[9] and discussed further by Wanzenberg.[10] We begin this section by recalling some details which are relevant in the present context.

For the whole of this section, let $(\mathscr{H}, \pi)$ be a positive energy representation of $(\mathscr{A}, \alpha)$. This means that there exists, on the Hilbert space $\mathscr{H}$, a strongly continuous representation U_π of the translations such that

$$\pi \circ \alpha_x = \mathrm{Ad}U_\pi(x) \circ \pi, \quad \mathrm{sp}U_\pi \subset \overline{\mathbf{V}_+}.$$

It is then known[11] that the operators $U_\pi(x)$ can be chosen such that $U_\pi(x) \in \pi(\mathcal{A})''$, and it is a remarkable fact[9] that there is a unique choice for which, on any subspace $\mathcal{H}' \subset \mathcal{H}$ which is invariant under $\pi(\mathcal{A})$, the spectrum $\mathrm{sp}U_\pi|_{\mathcal{H}'}$ has Lorentz-invariant lower boundary. From now on, U_π will always denote this canonical implementation of the translations and E_π the associated spectral family.

The notion of energy contents of π-normal states can now be introduced. To this end, denote by

$$\mathcal{S}_\pi(\Delta) := \{\mathrm{tr}\rho\pi(\cdot) | \rho \in \mathcal{T}_1(\mathcal{H}), \quad \rho \geqslant 0, \quad \mathrm{tr}\rho = 1, \quad \rho E_\pi(\Delta) = \rho\}$$

the set of all π-normal states which have energy-momentum in Δ. [In the present and in the next section, Δ always stands for a compact subset of $\overline{\mathbf{V}_+}$. Moreover, we will use the notation $\Delta_q := \overline{\mathbf{V}_+} \cap (q - \overline{\mathbf{V}_+})$ for the double cone in momentum space with apices 0 and $q \in \mathbf{V}_+$.] Then $\mathcal{S}_\pi = \cup_\Delta \mathcal{S}_\pi(\Delta)^{\|\cdot\|}$ is just the folium of π. Another set $\tilde{\mathcal{S}}_\pi$ of states, called the *energy component* of π, is defined by

$$\tilde{\mathcal{S}}_\pi := \overline{\cup_\Delta \tilde{\mathcal{S}}_\pi(\Delta)}^{\|\cdot\|},$$

where $\tilde{\mathcal{S}}_\pi(\Delta)$ denotes the set of all locally π-normal states in the weak closure of $\mathcal{S}_\pi(\Delta)$. Of course, $\mathcal{S}_\pi \subset \tilde{\mathcal{S}}_\pi$, and one can show that $\tilde{\mathcal{S}}_\pi$ is a folium, too. Physically, it is interpreted as the set of states which can be reached from the folium of π by operations requiring only a finite amount of energy. Some important properties of the elements of $\tilde{\mathcal{S}}_\pi$ are collected in the following Lemma, parts 2 and 3 of which are due to Wanzenberg.[10]

Lemma 2.1: Let $\omega \in \tilde{\mathcal{S}}_\pi$ and denote by $(\mathcal{H}_\omega, \pi_\omega, \Omega_\omega)$ its GNS triple.

(1) Then $(\mathcal{H}_\omega, \pi_\omega)$ is a locally π-normal positive energy representation of $(\mathcal{A}, \alpha)$.
(2) Moreover, if $\omega \in \tilde{\mathcal{S}}_\pi(\Delta_q)$ for some $q \in \mathbf{V}_+$, then one has $\Omega_\omega \in E_{\pi_\omega}(\Delta_q)\mathcal{H}_\omega$.
(3) In the situation of 2, one has $\tilde{\mathcal{S}}_{\pi_\omega}(\Delta) \subset \tilde{\mathcal{S}}_\pi(\Delta + \Delta_q - \Delta_q)$ for any Δ.

Proof: Using the positivity of the energy in the representation π, part 1 follows from the fact that ω is locally π-normal by arguments similar to those in Ref. 12. (In Ref. 12, these arguments are only applied to states $\omega \in \tilde{\mathcal{S}}_\pi(\Delta_q)$, but they carry over to norm limits of such states as well.)

For simplicity, part 2 will only be proved in the case that π_ω is factorial. For the general case, see Ref. 10. Let Δ_q be given. Then the idea is to show that Ω_ω does not have momentum outside Δ_q. To this end, fix some $p \in \overline{\mathbf{V}_+} \setminus \Delta_q$ and choose a neighborhood $\mathcal{N}_p \subset \overline{\mathbf{V}_+} \setminus \Delta_q$ of p and an open set $\mathcal{N} \subset \mathrm{sp}U_{\pi_\omega}$ such that

$$(\Delta_q + \mathcal{N} - \mathcal{N}_p) \cap \overline{\mathbf{V}_+} = \varnothing.$$

(Such a choice is always possible because the lower boundary of $\mathrm{sp}U_{\pi_\omega}$ is Lorentz invariant.) Now choose a test function f satisfying $\mathrm{supp}\, \tilde{f} \subset \mathcal{N} - \mathcal{N}_p$ and take an arbitrary $A \in \mathcal{A}$. Then, $A(f) := \int dx f(x) \alpha_x(A)$ is an element of $\mathcal{A}$ and satisfies $\pi(A(f))E_\pi(\Delta_q)\mathcal{H}_\pi = \{0\}$. Since $\omega \in \tilde{\mathcal{S}}_\pi(\Delta_q)$, this implies $\omega(A(f)^* A(f)) = 0$, hence $\pi_\omega(A(f))\Omega_\omega = 0$. This means that Ω_ω is orthogonal to

$$\mathcal{D} := \mathrm{span}\{\pi_\omega(A(f))^* \Psi | \Psi \in \mathcal{H}_\omega, A \in \mathcal{A}, \mathrm{supp}\, \tilde{f} \subset \mathcal{N} - \mathcal{N}_p\}.$$

Since π_ω is factorial and $\mathcal{N} - \mathcal{N}_p$ is open, it follows by an argument explained in Ref. 4 (see, in particular, the proof of Prop. 2.2 therein) that the closure of $\mathcal{D}$ equals $E_{\pi_\omega}(\mathrm{sp}U_{\pi_\omega} + \mathcal{N}_p - \mathcal{N})\mathcal{H}_\omega$. Thus, $\Omega_\omega \in \mathcal{D}^\perp$ yields

$$\{\Omega_\omega\}^\perp \supset \mathcal{D}^{\perp\perp} = \overline{\mathcal{D}} = E_{\pi_\omega}(\mathrm{sp}U_{\pi_\omega} + \mathcal{N}_p - \mathcal{N})\mathcal{H}_\omega \supset E_{\pi_\omega}(\mathcal{N}_p)\mathcal{H}_\omega,$$

where the last inclusion holds because $\mathrm{sp}U_{\pi_\omega} - \mathcal{N} \ni 0$. From this, we get $E_{\pi_\omega}(\mathcal{N}_p)\Omega_\omega = 0$ or, as $p \in \overline{\mathbf{V}_+} \setminus \Delta_q$ was arbitrary, $\Omega_\omega \in E_{\pi_\omega}(\Delta_q)\mathcal{H}_\omega$.

To prove part 3, let $\Psi \in E_{\pi_\omega}(\Delta)\mathscr{H}_\omega$. Part 2 and the cyclicity of Ω_ω imply that there exists in $\mathscr{A}$ a sequence $(A_n)_{n\in\mathbf{N}}$ of operators with energy-momentum support in $\Delta-\Delta_q$ such that $\Psi = \lim_{n\to\infty}\pi_\omega(A_n)\Omega_\omega$. From $\omega \in \widetilde{\mathscr{S}}_\pi(\Delta_q)$ follows $\omega(A_n^*\cdot A_n)\in\widetilde{\mathscr{S}}_\pi(\Delta_q+(\Delta-\Delta_q))=\widetilde{\mathscr{S}}_\pi(\Delta+\Delta_q-\Delta_q)$, i.e.,

$$\langle\Psi,\pi_\omega(\cdot)\Psi\rangle = \lim_{n\to\infty}\omega(A_n^*\cdot A_n)\in\widetilde{\mathscr{S}}_\pi(\Delta+\Delta_q-\Delta_q).$$

Thus any vector state from $\mathscr{S}_{\pi_\omega}(\Delta)$ lies in $\widetilde{\mathscr{S}}_\pi(\Delta+\Delta_q-\Delta_q)$. One now gets the assertion by taking convex combinations, norm limits and locally normal w^* limits. ■

The preceding lemma has shown that positivity of the energy is a property which ''survives'' the process of going from the representation π to the GNS representation of a state in $\widetilde{\mathscr{S}}_\pi$. Other properties survive as well, as for instance the compactness condition $C_\sharp$ of Fredenhagen and Hertel[13,14] which can be formulated as follows:

Definition: Condition $C_\sharp$ is said to be satisfied in the (positive energy) representation π if, for any Δ and any bounded region $\mathscr{O}\subset\mathbf{R}^{1+s}$, the set

$$\mathscr{S}_\pi(\Delta)|_{\mathscr{A}(\mathscr{O})}\equiv\{\omega|_{\mathscr{A}(\mathscr{O})}|\,\omega\in\mathscr{S}_\pi(\Delta)\}$$

is contained in a $\|\cdot\|$-compact subset of $\mathscr{A}(\mathscr{O})^*$.

This condition controls the infrared properties of the model under consideration, cf. Ref. 14. It has been established in the theory of a massive free particle (in any space-time dimension) and in the theory of a massless (scalar or vector) particle in at least $1+3$ space-time dimensions[15,14] and is believed to hold in QED as well. In the present context, it will play a technical role in the proof of Theorem 3.2 since it allows (by part 1 of the following lemma) a simplification in the definition of $\widetilde{\mathscr{S}}_\pi(\Delta)$ given above.

Lemma 2.2: Let $C_\sharp$ be satisfied in the representation π and let ω be a state in the weak closure of $\mathscr{S}_\pi(\Delta)$. Then

(1) ω is locally π-normal, i.e., $\omega\in\widetilde{\mathscr{S}}_\pi(\Delta)$;

(2) $C_\sharp$ is satisfied in the GNS representation of ω.

Proof: Both parts follow from the fact that, in restriction to $\mathscr{A}(\mathscr{O})$, any w^* limit of $\mathscr{S}_\pi(\Delta)$ is, as a consequence of $C_\sharp$, even a $\|\cdot\|$-limit of $\mathscr{S}_\pi(\Delta)$. Part 1 now follows directly. For part 2, we note that the above fact implies $\widetilde{\mathscr{S}}_\pi(\Delta)|_{\mathscr{A}(\mathscr{O})}\subset\overline{\mathscr{S}_\pi(\Delta)|_{\mathscr{A}(\mathscr{O})}}^{\|\cdot\|}$, which, in view of Lemma 2.1(3) yields the assertion. ■

III. INFRAVACUUM REPRESENTATIONS

Intuitively, an infrared cloud (in QED) can be split, given any energy threshold $\epsilon>0$, into a finite number N_ϵ of ''hard'' photons and infinitely many ''soft'' photons whose total energy is less than ϵ. As the creation of hard photons can be described by a quasilocal operation, this picture motivates the following mathematical notion.

Definition: An irreducible representation $(\mathscr{H}_I,\pi_I)$ of $\mathscr{A}$ is called an *infravacuum representation* if, for any $q\in\mathbf{V}_+$, the set $\widetilde{\mathscr{S}}_{\pi_0}(\Delta_q)$ contains some π_I-normal state.

With any arbitrarily small amount of energy, one can thus create from the vacuum representation some state in an infravacuum representation. In the terms of Ref. 10, this means that the transition energy between the sectors π_0 and π_I vanishes (cf. Theorem 3.2 below). As an example for infravacuum representations, we mention the KPR representations, i.e., a class of non-Fock representations (of the free asymptotic electromagnetic field) devised by Kraus, Polley and Reents,[2] cf. also Ref. 16, so as to be stable (up to unitary equivalence) under the bremsstrahlung produced in typical collision processes of charged particles. We note that with the above definition, the vacuum π_0 itself is an infravacuum representation, which will be convenient in the sequel. However, the notion of infravacuum representations is tailored to theories with massless particles in the sense that in a purely massive theory, π_0 would be the only infravacuum representation.

The following lemma collects some basic properties of any infravacuum representation.

Lemma 3.1: Let π_I be an infravacuum representation. Then

(1) for any $q \in \mathbf{V}_+$, $\tilde{\mathscr{S}}_{\pi_0}(\Delta_q)$ contains a *pure* π_I-normal state;

(2) π_I is a locally normal positive energy representation of $(\mathscr{A},\alpha)$;

(3) for any bounded $\mathscr{O}$, the restriction of $\pi_I : \mathscr{A}(\mathscr{O}) \to \pi_I(\mathscr{A}(\mathscr{O}))$ to uniformly bounded subsets of $\mathscr{A}(\mathscr{O})$ is a homeomorphism with respect to the weak operator topologies.

Proof: Choose, for some $q \in \mathbf{V}_+$, a state $\omega \in \mathscr{S}_{\pi_I} \cap \tilde{\mathscr{S}}_{\pi_0}(\Delta_q)$ and let $(\mathscr{H},\pi,\Omega)$ be its GNS triple. Since $\omega \in \mathscr{S}_{\pi_I}$, π is quasi-equivalent to the irreducible representation π_I. This implies that $\pi(\mathscr{A})'$ is a type I factor and thus contains some minimal projection P. Obviously, $\pi|_{P\mathscr{H}} \cong \pi_I$, which implies assertion 2 by Lemma 2.1(1). Moreover, recalling $\Omega \in E_\pi(\Delta_q)\mathscr{H}$, it also implies that the pure π_I-normal state $\omega_P := (P\Omega, \pi(\cdot)P\Omega)/\|P\Omega\|^2$ fulfills

$$\omega_P \in \mathscr{S}_{\pi_I}(\Delta_q) \subset \mathscr{S}_{\pi}(\Delta_q) \subset \tilde{\mathscr{S}}_{\pi}(\Delta_q) \subset \tilde{\mathscr{S}}_{\pi_0}(2\Delta_q).$$

Since q was arbitrary, this proves part 1. Finally, part 3 follows by applying Corollary 7.1.16. of Ref. 17 to the maps $\pi_I : \mathscr{A}(\mathscr{O}) \to \pi_I(\mathscr{A}(\mathscr{O}))$ which are isomorphisms of von Neumann algebras since π_I is faithful and locally normal. ■

Let us remark that one can even establish local unitary equivalence between π_0 and π_I under the additional assumption that π_I satisfies weak additivity. [If additivity and not merely weak additivity were assumed for π_0, additivity and hence weak additivity would follow for π_I by local normality.] The reasoning for this goes along the following line: According to the arguments collected nicely in Ref. 8, there exist Reeh–Schlieder vectors in $\mathscr{H}_0$ and $\mathscr{H}_I$, i.e., the von Neumann algebras $\mathscr{A}(\mathscr{O})$ and $\pi_I(\mathscr{A}(\mathscr{O}))$ possess cyclic and separating vectors. From this, one obtains local unitary equivalence by applying Theorem 7.2.9. of Ref. 17. We recall, however, that an attempt to establish unitary implementability of π_I on an algebra pertaining to an unbounded region (such as $\mathscr{O}'$) would fail because π_I need not be normal on such a region. Indeed, the interesting case will be the situation in which π_I does *not* fulfill the DHR criterion with respect to π_0.

To conclude this section about general properties of infravacuum representations, we want to compare the nets $\tilde{\mathscr{S}}_{\pi_0}$ and $\tilde{\mathscr{S}}_{\pi_I}$ of states. In the next theorem, we will have to make the additional assumption that the vacuum state is unique in the sense that, for some $q \in \mathbf{V}_+$, ω_0 is the only vacuum state in $\tilde{\mathscr{S}}_{\pi_0}(\Delta_q)$.

Theorem 3.2: Let π_I be an infravacuum representation. Then we have, for any Δ and any $q \in \mathbf{V}_+$:

$$\tilde{\mathscr{S}}_{\pi_I}(\Delta) \subset \tilde{\mathscr{S}}_{\pi_0}(\Delta + \Delta_q - \Delta_q).$$

If, moreover, the defining representation π_0 satisfies $C_\sharp$ and if the vacuum state is unique in the sense explained above, then also the converse is true, namely

$$\tilde{\mathscr{S}}_{\pi_0}(\Delta) \subset \tilde{\mathscr{S}}_{\pi_I}(\Delta + \Delta_q - \Delta_q).$$

Proof: Using Lemma 3.1(1), the first statement follows immediately from Lemma 2.1(3). To prove the second statement, we will first use the two additional assumptions to show $\omega_0 \in \tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$ for any Δ_q. This is done as follows. Take some $\omega \in \mathscr{S}_{\pi_I} \cap \tilde{\mathscr{S}}_{\pi_0}(\Delta_q)$. It is easy to see that $\omega \in \tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$. Now consider the family $(\omega_L)_{L>0}$ of states defined by $\omega_L := (1/|V_L|)\int_{V_L} d^{1+s}x\, \omega \circ \alpha_x$, where $V_L := [-L,L]^{(1+s)} \subset \mathbf{R}^{1+s}$. Since $\omega_L \in \tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$, any w^* limit $\tilde{\omega}$ of this family is a w^* limit of $\tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$. Now π_I satisfies $C_\sharp$ by Lemma 2.2(2), so Lemma 2.2(1) can be applied to $\tilde{\omega}$ and yields $\tilde{\omega} \in \tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$. In particular, $\tilde{\omega}$ has positive energy. On the other hand, $\tilde{\omega}$ is translation invariant by construction, i.e., it is a vacuum state. Moreover, by the first part of the present theorem, $\tilde{\omega} \in \tilde{\mathscr{S}}_{\pi_0}(2\Delta_q)$. Now if q is sufficiently small, the uniqueness assumption yields $\tilde{\omega} = \omega_0$ whence $\omega_0 \in \tilde{\mathscr{S}}_{\pi_I}(\Delta_q)$. Trivially, this conclusion remains true for any q. With this information, the second statement follows again from Lemma 2.1(3). ■

We remark, in addition, that the previous theorem does not, in general, entail that one of the sets $\tilde{\mathscr{S}}_{\pi_0}(\Delta)$ and $\tilde{\mathscr{S}}_{\pi_I}(\Delta)$ is contained in the other because the nets $\Delta \mapsto \tilde{\mathscr{S}}_{\pi}(\Delta)$ ($\pi = \pi_0$ or π_I)

need not be regular from the outside [which would mean $\tilde{\mathscr{S}}_\pi(\Delta)=\tilde{\tilde{\mathscr{S}}}_\pi(\Delta):=\cap_q\tilde{\mathscr{S}}_\pi(\Delta+\Delta_q)$], but the outer regularized nets $\tilde{\tilde{\mathscr{S}}}_\pi$ coincide. We note, as an aside, that the nets $\mathscr{S}_{\pi_0}$ and $\mathscr{S}_{\pi_I}$, in contrast, are regular from the outside, as follows from the continuity of the spectral families E_0 and E_{π_I}.

IV. THE DHR CRITERION

In the spirit of the theory of superselection sectors of Doplicher, Haag and Roberts,[1] it is natural to consider, for any infravacuum representation π_I, the set of representations

$$\mathrm{DHR}(\pi_I):=\{\pi|\,\pi|_{\mathscr{A}(\mathscr{O}')}\cong\pi_I|_{\mathscr{A}(\mathscr{O}')},\forall\mathscr{O}\in\mathscr{K}\}$$

(where $\mathscr{K}$ denotes the set of all double cones in $\mathbf{R}^{1+s}$). The well-known DHR analysis can be carried through under one additional assumption, namely that of Haag duality. Thus, we will restrict our attention in the sequel to those infravacuum representations which have this property, i.e.,

$$\pi_I(\mathscr{A}(\mathscr{O}'))'=\pi_I(\mathscr{A}(\mathscr{O})),\quad\forall\mathscr{O}\in\mathscr{K}.$$

We mention as an aside that Haag duality has been established for KPR representations (cf. Sec. III) if the net of observables fulfills an additional condition corresponding to an algebraic version of Gauss's law.[16,18]

Instead of $\mathrm{DHR}(\pi_I)$, one can now study the set

$$\Delta_{\pi_I,t}:=\bigcup_{\mathscr{O}\in\mathscr{K}}\Delta_{\pi_I,t}(\mathscr{O})$$

of transportable localized endomorphisms of the C^*-algebra $\pi_I(\mathscr{A})$, where $\Delta_{\pi_I,t}(\mathscr{O})$ denotes the subset of all such endomorphisms which act trivially on the subalgebra $\pi_I(\mathscr{A}(\mathscr{O}'))$. Transportability means that, for any $x\in\mathbf{R}^{1+s}$, the endomorphisms ρ_I and $\rho_{I,x}$ are unitarily equivalent in $\mathscr{H}_{\pi_I}$, where $\rho_{I,x}$ is given by the action of the translations

$$\Delta_{\pi_I,t}(\mathscr{O})\to\Delta_{\pi_I,t}(\mathscr{O}+x):\rho_I\mapsto\rho_{I,x}:=\mathrm{Ad}\,U_{\pi_I}(x)\circ\rho_I\circ\mathrm{Ad}\,U_{\pi_I}(x)^*.$$

We recall that $\Delta_{\pi_I,t}$ can be viewed as the set of objects of a tensor C^*-category with subobjects and finite direct sums,[19] also denoted by $\Delta_{\pi_I,t}$, whose morphisms $T\in I_{\pi_I}(\sigma_I,\rho_I)$ from σ_I to ρ_I are the intertwining operators $T\in\mathscr{B}(\mathscr{H}_{\pi_I})$ from $\sigma_I\circ\pi_I$ to $\rho_I\circ\pi_I$. It is a crucial consequence of Haag duality that the intertwiners are again local observables:

$$I_{\pi_I}(\sigma_I,\rho_I)\subset\pi_I(\mathscr{A}(\mathscr{O}))\quad\text{if }\sigma_I,\rho_I\in\Delta_{\pi_I,t}(\mathscr{O}).$$

[Comment on the notation: Viewing the index I as a label which distinguishes different infravacuum representations, we let for clarity all objects of $\Delta_{\pi_I,t}$ carry an index I. In particular, I may take on the value 0, then referring to the vacuum.] One eventually has to pass from $\Delta_{\pi_I,t}$ to the full subcategory $\Delta_{\pi_I,f}$ of objects with finite statistical dimension, which is a tensor C^*-category with conjugates. Its importance resides in the fact that it is possible to recover from it (by a deep result of Doplicher and Roberts[5]) a compact gauge group and a field net with normal commutation relations whose gauge invariant part contains all finite-dimensional DHR sectors of $\pi_I(\mathscr{A})$.

Though not explicitly stated in the literature, it is well-known that the above-mentioned notions do not depend on the choice of the infravacuum representation π_I. In fact, a slightly more general result is true:

Theorem 4.1: Let π_I be an irreducible, locally normal representation of $\mathscr{A}$ and assume that π_0 and π_I satisfy Haag duality. Then, a bijective functor

$$F:\Delta_{\pi_0,t}\to\Delta_{\pi_I,t}$$

is defined by the actions (on objects respectively, on morphisms)

$$F(\rho_0) := \pi_I \circ \pi_0^{-1} \circ \rho_0 \circ \pi_0 \circ \pi_I^{-1}, \quad \rho_0 \in \Delta_{\pi_0,t},$$

$$F(T) := \pi_I \circ \pi_0^{-1}(T), \quad T \in I_{\pi_0}.$$

This functor restricts to an isomorphism $F:\Delta_{\pi_0,f} \to \Delta_{\pi_I,f}$ of the corresponding tensor C^*-categories with conjugates.

Proof: As the algebra $\mathscr{A}$ is simple, any representation is faithful, so $\pi_I \circ \pi_0^{-1}: \pi_0(\mathscr{A}) \to \pi_I(\mathscr{A})$ is bijective. Disregarding for a moment the question of transportability, this means that F, as defined above, is well-defined on the objects ρ_0 and remembering $I_{\pi_0} \subset \pi_0(\mathscr{A})$, also on the morphisms. Obviously, F is bijective. One checks that $F(\rho_0)$ is localized in $\mathscr{O}$ iff ρ_0 is, which implies that $F(\rho_0)$ is transportable iff ρ_0 is. Since the statistical dimension of an object is a purely algebraic notion,[20] F clearly restricts to the subcategories of finite statistical dimension. As any bijective functor, F preserves all algebraic structures involved, such as subobjects, direct sums, conjugates and the symmetries. ■

Obviously, what lies at the heart of this proof is that $\pi_I \circ \pi_0^{-1}$ is a net isomorphism. The additional property of this isomorphism of being bicontinuous (with respect to the weak operator topologies) in restriction to norm-bounded subsets of local algebras [cf. Lemma 3.1(3)] has not been exploited yet but will play an essential role in Sec. VI.

We should emphasize that the above theorem and its proof remain valid if π_0 also is replaced by an irreducible, locally normal representation of $\mathscr{A}$ which satisfies Haag duality. In particular, the spectral properties of this representation do not enter into the argument.

As straightforward as the foregoing theorem may be, as important is its interpretation: It shows that replacing the vacuum π_0 by an infravacuum π_I does not affect the superselection structure of the theory. Not only the charges of the theory remain the same, but also their fusion structure, their statistics, their localization and their transportability properties. In more pictorial terms, an infravacuum is suited as well as the ''empty'' vacuum as a background in front of which DHR theory can take place.

V. COCYCLES AND THEIR SPECTRAL PROPERTIES

We have seen in the previous section that the DHR superselection structure of a theory does not depend on the infravacuum with respect to which it is defined. In elementary particle physics, physically meaningful (DHR) representations should, however, also fulfill what is called Borchers's criterion, i.e., have positive energy. In the next two sections, we want to convince ourselves that this additional property too is independent of the infravacuum representation chosen.

It is well-known[1] that Borchers's criterion is fulfilled for all $\rho_0 \in \Delta_{\pi_0,f}$ since the subcategory of finite-dimensional objects is closed under conjugates. However, it is important to note that the proof of this theorem, in relying on the additivity of the energy, makes crucial use of the fact that π_0 is a vacuum representation and thus fails when π_0 is replaced by π_I. Yet, any $\rho_I \in \Delta_{\pi_I,f}$ actually *does* fulfill Borchers's criterion, as follows from Theorem 6.3 below.

Leaving DHR theory for a moment, we first discuss a mathematical notion whose relevance will soon become clear. Thus, let $(\mathscr{H}_I, U_I)$ be a pair consisting of a Hilbert space $\mathscr{H}_I$ and a strongly continuous unitary representation U_I of $\mathbf{R}^{1+s}$ whose spectral family will be denoted by E_I. Note that no condition is imposed on its spectrum at this stage.

Definition: A strongly continuous function $\Gamma:\mathbf{R}^{1+s} \to \mathscr{U}(\mathscr{H}_I)$ is called a *cocycle over* $(\mathscr{H}_I, U_I)$ if it fulfills the cocycle equation

$$\Gamma(x)\mathrm{Ad}\, U_I(x)(\Gamma(y)) = \Gamma(x+y).$$

A cocycle is called *C-spectral*, where C is some closed subset of $\mathbf{R}^{1+s}$, if it complies with the following condition: for any compact $\Delta \subset \mathbf{R}^{1+s}$, one has

$$\mathrm{sp}(\Gamma(\cdot)E_I(\Delta)) \subset C - \Delta.$$

Here, the spectrum sp A of a uniformly bounded, strongly continuous operator-valued function $A:\mathbf{R}^{1+s} \to \mathscr{B}(\mathscr{H}_I)$ is, by definition, the support of its Fourier transform in the sense of operator-

valued distributions. In particular, one has $\int dx f(x)A(x)=0$ for any test function f with supp $\tilde{f}\cap \mathrm{sp}\, A=\varnothing$. Of course, if $A=U:\mathbf{R}^{1+s}\to\mathscr{U}(\mathscr{H}_I)$ is a group homomorphism, sp $A=$ sp U coincides with the common spectrum of the generators of the group.

The notion of spectral cocycles does not seem to have appeared in the literature. The idea of applying it to the above-mentioned problem and the main part of Theorem 5.2 are due to Buchholz.

The following basic properties of the spectrum will be important in the sequel:

Lemma 5.1: Let A, A_1 and A_2 be uniformly bounded, strongly continuous operator-valued functions on $\mathbf{R}^{1+s}$. Then, one has

(1) sp $A^*=-$ sp A, where $A^*(x)=(A(x))^*$;

(2) $\mathrm{sp}(A_1+A_2)\subset \mathrm{sp}\, A_1\cup \mathrm{sp}\, A_2$;

(3) $\mathrm{sp}(A_1A_2)\subset \mathrm{sp}\, A_1+\mathrm{sp}\, A_2$ if sp A_1 or sp A_2 is bounded.

Proof: Parts 1 and 2 are elementary, whereas part 3 can be obtained, by considering matrix elements, from the corresponding property of scalar-valued functions which in turn is the contents of Secs. 1,2 and Théorème II of Ref. 21, Chap. IV. ■

Theorem 5.2: A bijective correspondence between strongly continuous unitary representations $V:\mathbf{R}^{1+s}\to\mathscr{U}(\mathscr{H}_I)$ and cocycles Γ over $(\mathscr{H}_I,U_I)$ is given by

$$\Gamma(x)=V(x)U_I(x)^*.$$

Moreover, sp $V\subset C$ iff Γ is C-spectral.

Proof: The first assertion is obvious. As to the second one, let V satisfy sp $V\subset C$. Making use of the previous Lemma, one then has, for any compact $\Delta\subset\mathbf{R}^{1+s}$,

$$\mathrm{sp}(\Gamma(\cdot)E_I(\Delta))=\mathrm{sp}(V(\cdot)U_I^*(\cdot)E_I(\Delta))\subset \mathrm{sp}\, V-\mathrm{sp}(E_I(\Delta)U_I(\cdot))\subset C-\Delta$$

which shows that Γ is C-spectral. To prove the converse, let Γ be C-spectral. Let $\Delta\subset\mathbf{R}^{1+s}$ be compact and choose a cover of Δ by a finite number of pairwise disjoint Borel sets $(\Delta_j)_{j=1,\ldots,N}$. Then, again by Lemma 5.1, one gets

$$\mathrm{sp}(V(\cdot)E_I(\Delta))\subset \mathrm{sp}\left(V(\cdot)\sum_j E_I(\Delta_j)\right)=\mathrm{sp}\sum_j (\Gamma(\cdot)E_I(\Delta_j)E_I(\Delta_j)U_I(\cdot))$$

$$\subset \bigcup_j (\mathrm{sp}(\Gamma(\cdot)E_I(\Delta_j))+\mathrm{sp}(E_I(\Delta_j)U_I(\cdot)))\subset \bigcup_j (C-\Delta_j+\Delta_j)$$

$$=C+\bigcup_j(\Delta_j-\Delta_j).$$

Since the cover may be chosen such that the maximal diameter of the sets Δ_j is arbitrarily small, this implies $\mathrm{sp}(V(\cdot)E_I(\Delta))\subset C$ for any compact $\Delta\subset\mathbf{R}^{1+s}$, which, in view of $s-\lim_{\Delta\nearrow\mathbf{R}^{1+s}} E_I(\Delta)=\mathbf{1}_{\mathscr{H}_I}$, finally yields sp $V(\cdot)\subset C$. ■

VI. BORCHERS'S CRITERION AND INFRAVACUA

Let us now return to DHR theory. If π_I is an infravacuum representation of $\mathscr{A}$, the role of what was denoted by $(\mathscr{H}_I,U_I)$ in the previous section is of course taken over by $(\mathscr{H}_{\pi_I},U_{\pi_I})$. For any object $\rho_I\in\Delta_{\pi_I,t}$, the following set Z^{ρ_I} of cocycles is of interest:

$$Z^{\rho_I}:=\{\Gamma^{\rho_I}|\Gamma^{\rho_I}\ \text{is a cocycle over}\ (\mathscr{H}_{\pi_I},U_{\pi_I})\ \ \text{and}\ \Gamma^{\rho_I}(x)\in I_{\pi_I}(\rho_{I,x},\rho_I)\}.$$

An object ρ_I is called covariant if there exists a strongly continuous unitary representation $V_{\rho_I}:\mathbf{R}^{1+s}\to\mathscr{U}(\mathscr{H}_{\pi_I})$ which implements the translations in the representation $\rho_I\circ\pi_I$. [Such a representation will be called an *implementation* and denoted with the symbol V_{ρ_I} in order to avoid confusion with the unique *canonical* implementation U_π introduced in Sec. II for the positive energy representations π of $\mathscr{A}$.]

By relating V_{ρ_I} and Γ^{ρ_I} as in Theorem 5.2, we easily obtain the following result:

Lemma 6.1: Z^{ρ_I} is nonempty iff $\rho_I \in \Delta_{\pi_I,t}$ is covariant.

In the rest of this section, we will study the behavior of the cocycles introduced above under the bifunctor $F:\Delta_{\pi_0,t}\to\Delta_{\pi_I,t}$ described in Theorem 4.1. We now have cocycles $\Gamma^{\rho_0}\in Z^{\rho_0}$ and $\Gamma^{\rho_I}\in Z^{\rho_I}$ which take on values in the groups of local and unitary elements of $\pi_0(\mathscr{A})$ respectively $\pi_I(\mathscr{A})$. As a consequence, not only they are in the domain of F (which equals $\pi_I\circ\pi_0^{-1}$ on intertwiners) respectively F^{-1}, but we even have because of the continuity properties of F established in Lemma 3.1(3) (whose proof only relies on local normality—cf. also the remark after Theorem 4.1.):

Lemma 6.2: Let $\rho_0\in\Delta_{\pi_0,t}$, $\rho_I:=F(\rho_0)$. Then the functor F maps Z^{ρ_0} onto Z^{ρ_I}.

In other words, ρ_0 is covariant iff ρ_I is. [We will stick to the notation $\rho_I:=F(\rho_0)$ in the sequel.] Moreover, an elementary calculation shows that, in this case, implementations V_{ρ_0}, respectively, V_{ρ_I} of the translations for $\rho_0\circ\pi_0$, respectively, $\rho_I\circ\pi_I$ can be obtained from each other by the formula

$$V_{\rho_I}(x)=\pi_I\circ\pi_0^{-1}(V_{\rho_0}(x)U_{\pi_0}(x)^*)U_{\pi_I}(x). \tag{1}$$

We will now investigate in what extent the functor F also respects spectral properties. The inclusions obtained in Theorem 3.2 are taken as a starting point, but we will comment on more general situations at the end of this section.

Theorem 6.3: Let π_I be an irreducible, locally normal positive energy representation of $\mathscr{A}$ and assume that π_0 and π_I satisfy Haag duality. Denote by

$$F:\Delta_{\pi_0,t}\to\Delta_{\pi_I,t}$$

the bifunctor of Theorem 4.1. If both inclusions of Theorem 3.2 are valid, then one obtains for any covariant object $\rho_0\in\Delta_{\pi_0,t}$:

(1) F maps the C-spectral cocycles in Z^{ρ_0} onto the C-spectral ones in Z^{ρ_I};

(2) if V_{ρ_0} and V_{ρ_I} are related by (1) then $\mathrm{sp}\ V_{\rho_0}=\mathrm{sp}\ V_{\rho_I}$;

(3) ρ_0 has positive energy iff ρ_I has;

(4) if ρ_0 and ρ_I have positive energy and are finite direct sums of irreducibles, then the respective canonical implementations $U_{\rho_0\circ\pi_0}$ and $U_{\rho_I\circ\pi_I}$ of the translations are related by Eq. (1); in particular, their spectra coincide:

$$\mathrm{sp}\ U_{\rho_0\circ\pi_0}=\mathrm{sp}\ U_{\rho_I\circ\pi_I}.$$

Proof: 1. Let $\Gamma^{\rho_0}\in Z^{\rho_0}$ be C-spectral and let $\Gamma^{\rho_I}(x):=F(\Gamma^{\rho_0}(x))$. It will be shown in the next lemma that this entails, for any compact set Δ,

$$\mathrm{sp}(\Gamma^{\rho_I}(\cdot)E_I(\Delta))\subset\bigcap_{\Delta_0}\{C-\Delta_0|\tilde{\mathscr{S}}_{\pi_I}(\Delta)\subset\tilde{\mathscr{S}}_{\pi_0}(\Delta_0)\}.$$

As a consequence of the first inclusion of Theorem 3.2, Δ_0 can be chosen to be any neighborhood of Δ. This implies that the right-hand side reduces to $C-\Delta$, i.e., Γ^{ρ_I} is C-spectral. The converse is proved along the same lines, now using the second inclusion of Theorem 3.2.

2. Due to (1), part 1 applies to the cocycles $\Gamma^{\rho_0}(x):=V_{\rho_0}(x)U_0(x)^*$ and $\Gamma^{\rho_I}(x):=V_{\rho_I}(x)U_I(x)^*$. Using Theorem 5.2, this yields for any closed $C\subset\mathbf{R}^{1+s}$:

$$\mathrm{sp}\ V_{\rho_0}\subset C\ \text{iff}\ \mathrm{sp}\ V_{\rho_I}\subset C,$$

which immediately gives the assertion. From this, part 3 follows directly.

As to part 4, we first consider the special case when ρ_0 (and therefore also ρ_I) is irreducible. Then all implementations of the translations coincide up to a one-dimensional one, i.e., a phase factor. Together with part 2, this yields the following equivalences:

$$V_{\rho_0}=U_{\rho_0\circ\pi_0}\Leftrightarrow\ \mathrm{sp}\ V_{\rho_0}\ \text{has Lilb}\Leftrightarrow\ \mathrm{sp}\ V_{\rho_I}\ \text{has Lilb}\Leftrightarrow V_{\rho_I}=U_{\rho_I\circ\pi_I}$$

(where ''Lilb'' means ''Lorentz-invariant lower boundary''). Applying part 2 again gives all of the assertion in this case. In the general case of reducible ρ_0, we prove $V_{\rho_0}=U_{\rho_0\circ\pi_0}\Leftrightarrow V_{\rho_I}=U_{\rho_I\circ\pi_I}$ as follows: if $W_j\in I_{\pi_0}(\rho_{0,j},\rho_0)$ denote intertwiners which perform the decomposition of ρ_0 into irreducibles $\rho_{0,j}$, then $F(W_j)$ decompose ρ_I into irreducibles $\rho_{I,j}=F(\rho_{0,j})$. The implementations V_{ρ_0} and V_{ρ_I} decompose into subrepresentations $W_j^*V_{\rho_0}(x)W_j$ and $F(W_j^*)V_{\rho_I}(x)F(W_j)$ on the corresponding subspaces and these subrepresentations can be seen to be related by (1) iff V_{ρ_0} and V_{ρ_I} are. Since the latter are, by definition, canonical iff all their subrepresentations are, the assertion follows from the special case discussed first. ■

Lemma 6.4: Let Γ^{ρ_0} be C-spectral. Then we have for any compact Δ:

$$\mathrm{sp}(\Gamma^{\rho_I}(\cdot)E_I(\Delta))\subset\bigcap_{\Delta_0}\{C-\Delta_0\,|\,\tilde{\mathscr{S}}_{\pi_I}(\Delta)\subset\tilde{\mathscr{S}}_{\pi_0}(\Delta_0)\}.$$

Proof: Let Δ_0 be such that $\tilde{\mathscr{S}}_{\pi_I}(\Delta)\subset\tilde{\mathscr{S}}_{\pi_0}(\Delta_0)$ and let f be a test function satisfying $\mathrm{supp}\,\tilde{f}\cap(C-\Delta_0)=\varnothing$. Setting $\Gamma^{\rho_0}(f):=\int dx f(x)\Gamma^{\rho_0}(x)$, we obtain $\Gamma^{\rho_0}(f)E_0(\Delta_0)=\int dx f(x)\Gamma^{\rho_0}(x)E_0(\Delta_0)=0$ since Γ^{ρ_0} is C-spectral. This means $\omega(\pi_0^{-1}(\Gamma^{\rho_0}(f)^*\Gamma^{\rho_0}(f)))=0$ for any vector state $\omega\in\mathscr{S}_{\pi_0}(\Delta_0)$, hence (by weak continuity) for any $\omega\in\tilde{\mathscr{S}}_{\pi_0}(\Delta_0)$ and in particular for any $\omega\in\tilde{\mathscr{S}}_{\pi_I}(\Delta)$. We thus have $\pi_I\circ\pi_0^{-1}(\Gamma^{\rho_0}(f))E_I(\Delta)=0$ and hence by local normality $\int dx f(x)(\pi_I\circ\pi_0^{-1}(\Gamma^{\rho_0}(x))E_I(\Delta)=0$. [For this argument, one has to realize that f can be approximated in $L^1(\mathbf{R}^{1+s})$ by test functions with compact support.] As this holds for any f with $\mathrm{supp}\,\tilde{f}\cap(C-\Delta_0)=\varnothing$, it follows that $\mathrm{sp}(\pi_I\circ\pi_0^{-1}(\Gamma^{\rho_0}(\cdot))E_I(\Delta))\subset C-\Delta_0$. This yields the assertion. ■

The inclusions of Theorem 3.2 played a crucial role in the previous theorem. However, partial results remain valid if less information about the relation between $\tilde{\mathscr{S}}_{\pi_I}$ and $\tilde{\mathscr{S}}_{\pi_0}$ is available. For instance, let us merely assume (instead of these inclusions) that there exists a compact neighborhood $\mathscr{N}$ of 0 such that [cf. Lemma 2.1(3)]

$$\tilde{\mathscr{S}}_{\pi_I}(\Delta)\subset\tilde{\mathscr{S}}_{\pi_0}(\Delta+\mathscr{N})\quad\text{for all compact sets }\Delta.$$

In this case, we obtain with the same arguments as above that the cocycle $\Gamma^{\rho_I}:=F(\Gamma^{\rho_0}(\cdot))$ is $(C-\mathscr{N})$-spectral if Γ^{ρ_0} is C-spectral. In terms of V_{ρ_0} and V_{ρ_I}, this means

$$\mathrm{sp}\,V_{\rho_I}\subset\mathrm{sp}\,V_{\rho_0}-\mathscr{N}.$$

In general, all detailed information on the shape of sp V_{ρ_I} (such as the size of possible mass gaps) is lost, but what can easily be seen is that ρ_I has positive energy if ρ_0 has. The canonical implementations of the translations need, in general, not be related by (1) any more. A similar reasoning applies if the roles of π_0 and π_I are exchanged.

VII. CONCLUSION AND OUTLOOK

We have seen that, as far as DHR theory is concerned, the role of the vacuum representation π_0 can be taken over by any locally normal representation π_I satisfying Haag duality. Moreover, we have given sufficient conditions on π_I which ensure that the class of representations fulfilling Borchers's criterion is independent of π_I. These representations may be interpreted as background fields whose interaction with the charged particles of the model is weak.

Natural candidates for states describing such a background were elements of the energy component of the vacuum. The latter notion, introduced in Ref. 9, deserves some more interest on its own, and one might ask in the spirit of Sec. III under which circumstances other properties such as, e.g., Haag duality carry over from π_0 to π_I.

It seems thus that for most purposes, the characteristic feature of a vacuum representation, namely the existence of a translation invariant vector in the representation Hilbert space, is not an essential property. However, we recall that this property plays a crucial role for proving that the energy-momentum spectra are additive under the fusion of covariant sectors:

$$\mathrm{sp}\ U_{\rho_1\circ\rho_2\circ\pi_0}\supset \mathrm{sp}\ U_{\rho_1\circ\pi_0}+\mathrm{sp}\ U_{\rho_2\circ\pi_0}.$$

Trivially, using the results of Theorem 6.3, this yields

$$\mathrm{sp}\ U_{F(\rho_1\circ\rho_2)\circ\pi_I}\supset \mathrm{sp}\ U_{F(\rho_1)\circ\pi_I}+\mathrm{sp}\ U_{F(\rho_2)\circ\pi_I},$$

but it would of course be interesting to know under which assumptions the latter inclusion can be derived without relying on the vacuum sector.

Finally, returning to Buchholz's proposal of reconciling QED with superselection theory, several interesting questions arise. First, it has to be clarified which parts of the present work (in particular of Secs. IV and VI) can be generalized from DHR-like localized charges to charges localized in spacelike cones. Second, it will be important to check in specific models whether there exist backgrounds which permit better localization properties of charges of electric type than the vacuum does. In fact, this question could be answered affirmatively[6] for a simple model[7] recently introduced by Buchholz *et al.* Another candidate for such a model with the advantage of being closer to QED, might be the one proposed by Herdegen.[22]

ACKNOWLEDGMENTS

I am deeply indebted to Professor D. Buchholz for numerous helpful discussions and constant interest in this work. Financial support from the Deutsche Forschungsgemeinschaft (''Graduiertenkolleg Theoretische Elementarteilchenphysik'' at Hamburg University) is also gratefully acknowledged.

[1] S. Doplicher, R. Haag, and J. E. Roberts, Commun. Math. Phys. **23**, 199 (1971); **35**, 49 (1974).
[2] K. Kraus, L. Polley, and G. Reents, Ann. Inst. Henri Poincaré, Sect. A **26**, 109 (1977).
[3] D. Buchholz, Commun. Math. Phys. **85**, 49 (1982).
[4] D. Buchholz and K. Fredenhagen, Commun. Math. Phys. **84**, 1 (1982).
[5] S. Doplicher and J. E. Roberts, Commun. Math. Phys. **131**, 51 (1990).
[6] W. Kunhardt, ''On Infravacua and the Localization of Sectors'' (in preparation).
[7] D. Buchholz, S. Doplicher, G. Morchio, J. E. Roberts, and F. Strocchi, in *Operator Algebras and Quantum Field Theory*, edited by S. Doplicher, R. Longo, J. E. Roberts, and L. Szido (International Press, Cambridge, 1997), p. 647.
[8] C. d'Antoni, in *The Algebraic Theory of Superselection Sectors. Introduction and Recent Results*, edited by D. Kastler (World Scientific, Singapore, 1989), p. 248.
[9] H.-J. Borchers and D. Buchholz, Commun. Math. Phys. **97**, 169 (1985).
[10] R. Wanzenberg, Diploma thesis, Hamburg University, 1987.
[11] H.-J. Borchers, Commun. Math. Phys. **96**, 1 (1984).
[12] D. Buchholz and S. Doplicher, Ann. Inst. Henri Poincaré, Phys. Theor. **40**, 175 (1984).
[13] K. Fredenhagen and J. Hertel (unpublished).
[14] D. Buchholz and M. Porrmann, Ann. Inst. Henri Poincaré, Phys. Theor. **52**, 237 (1990).
[15] D. Buchholz and P. Jacobi, Lett. Math. Phys. **13**, 313 (1987).
[16] F. Hars, Diploma thesis, Hamburg University, 1993.
[17] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras* (Academic, New York, 1986), Vol. 2.
[18] W. Driessler, Commun. Math. Phys. **70**, 213 (1979).
[19] S. Doplicher and J. E. Roberts, Invent. Math. **98**, 157 (1989).
[20] R. Longo and J. E. Roberts, K-Theory **11**, 103 (1997).
[21] L. Schwartz, *Théorie des Distributions* (Hermann, Paris, 1966).
[22] A. Herdegen, J. Math. Phys. **37**, 100 (1996).

An increasing entropy for a free quantum particle

Clasine van Winter
Department of Mathematics and Department of Physics and Astronomy, University of Kentucky, Lexington, Kentucky 40506

(Received 14 November 1995; accepted for publication 24 April 1998)

For a quantum-mechanical particle without interaction, a linear manifold of states is identified that has a preferred time direction pointing to the future. States with the broken time-reversal symmetry are singled out by their behavior under dilations. At positive times such states can be described in terms of a density operator with the property that the trace of its square decreases as the time increases. This density operator determines an entropy that approaches its least upper bound when the time tends to infinity. The expectation value of the dilation operator was negative in the distant past and will be positive in the remote future. This is the irreversible aspect of the time evolution that causes the entropy to increase, although there is no approach to equilibrium. The density operator with increasing entropy is obtained from the usual density operator by an invertible transformation that is compared with the Λ-transformation in the Prigogine theory of irreversible behavior in K-systems and large Hamiltonian systems with many resonances. © *1998 American Institute of Physics.* [S0022-2488(98)03307-6]

I. INTRODUCTION

Let $\rho(t)$ be the density operator of a quantum system with Hamiltonian H. Following Von Neumann, one defines the entropy[1–3]

$$S[\rho(t)] = -\operatorname{Tr} \rho(t) \ln \rho(t). \tag{1.1}$$

If H is self-adjoint, the entropy $S[\rho(t)]$ is constant in time. The interpretation is that there is no irreversible behavior at the microscopic level.

The present paper examines one particle without interaction. For this case an invertible operator L is identified such that any ρ in the domain of L gives rise to a density operator $\sigma(t) := L\rho(t)$ $(t \geqslant 0)$ with the property that $S[\sigma(t)]$ increases between times $t=0$ and $t \to \infty$.

Units are chosen so that the time evolution takes the form

$$\rho(t) = \exp(-iH_0 t)\rho \exp(iH_0 t)$$

with $\rho = \rho(0)$. In the momentum representation the free Hamiltonian H_0 acts on $L^2(\mathbb{R}^3)$ as multiplication by k^2.

Let D be the dilation operator. This is the self-adjoint operator which acts on C_0^∞-functions $f(\mathbf{k}) \in L^2(\mathbb{R}^3)$ as

$$D = \frac{i}{2} (\mathbf{k}\cdot\boldsymbol{\nabla}_k + \boldsymbol{\nabla}_k \cdot \mathbf{k}). \tag{1.2}$$

Define $J := \exp(-\pi D/2)$. The operator J is self-adjoint but not bounded. Suppose that ρ is of the form

$$\rho = (1+J^2)^{-1/2} R (1+J^2)^{-1/2} \tag{1.3}$$

with some positive operator R in the trace class. If this condition is satisfied, the closure of $J\rho(t)J$ belongs to the trace class if $t \geqslant 0$, but not necessarily if $t < 0$. Hence states ρ satisfying Eq. (1.3) have a preferred time direction pointing to the future. For $t \geqslant 0$ only, define

 © 1998 American Institute of Physics

$$\sigma(t) := \rho(t) + J\rho(t)J. \tag{1.4}$$

According to Theorem 6.1

$$\text{Tr } \sigma(t) = \text{Tr } \sigma. \tag{1.5}$$

By general entropy inequalities[4]

$$S[\sigma(t)] \leqslant S[\rho(t)] + S[J\rho(t)J], \tag{1.6}$$

with equality if and only if

$$\rho(t)J\rho(t)J = 0, \quad J\rho(t)J\rho(t) = 0. \tag{1.7}$$

Typically Eq. (1.7) is not satisfied, but the relation becomes true as $t \to \infty$. By Lemma 7.2 the operators in Eq. (1.7) tend to 0 in the trace norm as $t \to \infty$. The right-hand side of Eq. (1.6) is equal to its value at $t=0$ by Lemma 7.1. Theorem 7.5 states that $S[\sigma(t)]$ tends to $S(\rho)+S(J\rho J)$ as $t \to \infty$. Hence the entropy increases between times $t=0$ and $t \to \infty$. Except in case ρ is of rank 1, it is an open problem whether the entropy increase is monotone.

Let P_+ and P_- be the spectral projections onto the subspaces of $L^2(\mathbb{R}^3)$ in which the dilation operator D is positive and negative, respectively. The proof that $S[\sigma(t)]$ tends to its least upper bound as $t \to \infty$ is based on the relations[5,6]

$$\begin{aligned} &\lim_{t\to\infty} \| P_- \exp(-iH_0 t) f \| = 0, \\ &\lim_{t\to-\infty} \| P_+ \exp(-iH_0 t) f \| = 0. \end{aligned} \tag{1.8}$$

As Eq. (1.2) indicates, D is the quantum analog of the scalar product $\mathbf{x}\cdot\mathbf{k}$, with $\mathbf{x}$ the position and $\mathbf{k}$ proportional to the momentum of the particle being considered. Intuitively, one expects that $\mathbf{x}\cdot\mathbf{k}$ was negative in the distant past and will be positive in the remote future. This notion is confirmed by Eq. (1.8). It represents an irreversible aspect of the time evolution, although there is no approach to equilibrium.

Due to Eq. (1.8), any $\rho(t)$ in the trace class tends in the trace norm to $P_+\rho(t)P_+$ as $t \to \infty$. Since JP_+ is bounded,

$$(1+J^2)^{1/2} P_+ \rho(t) P_+ (1+J^2)^{1/2}$$

belongs to the trace class. This means that $P_+\rho(t)P_+$ has the form of Eq. (1.3). Hence any $\rho(t)$ tends to a limit to which this paper applies.

Equation (1.4) identifies the operator L taking $\rho(t)$ into $\sigma(t)$. Theorem 6.3 states that L has an inverse such that $L^{-1}\sigma(t) = \rho(t)$. Let ρ be normalized such that Tr $\rho = 1$. Let A be a bounded operator on $L^2(\mathbb{R}^3)$. By Theorem 6.4 there exists an operator AL^{-1} taking the range of L into the trace class on $L^2(\mathbb{R}^3)$ in such a way that the expectation value of A at time $t \geqslant 0$ is given by

$$\langle A(t) \rangle = \text{Tr } A\rho(t) = \text{Tr}(AL^{-1})\sigma(t). \tag{1.9}$$

Hence $\sigma(t)$ can be used as a density operator to evaluate expectation values.

The notion of a transformation taking the usual density operator into a density operator with increasing entropy, has long been advocated by Prigogine and coauthors,[7,8] although not for a single quantum particle without interaction. Section II gives a brief overview of some of their work. The dilation operator and the operator J are discussed in Sec. III. It is of practical importance that functions in the domain of J are boundary values of analytic functions. This aspect of the dilation operator is the subject of Sec. IV. Section V is devoted to lemmas that characterize the behavior of $J\rho(t)J$ and related quantities as functions of t. This prepares for Sec. VI in which it is shown that Tr $\sigma(t)$ is constant and Tr $\sigma^2(t)$ decreasing. The construction of the operator L^{-1} is also in Sec. VI. Section VII is devoted to the entropy, including the proof that $S[\sigma(t)]$ tends to its least upper bound as $t \to \infty$. If ρ is of rank 1, the entropy increase is monotone by Sec. VIII.

Since ρ is supposed to be positive, σ defined by Eq. (1.4) is positive, but it is not true that every positive operator in the trace class is a σ generated by a positive ρ. This follows from examples in the Appendix. It is one of the reasons why it is an open problem whether the entropy increase is monotone in case the rank of ρ is larger than 1, as explained in Sec. VIII.

II. THE PRIGOGINE ENTROPY OPERATOR

The operator L has a role similar to the Λ-transformation in numerous papers by Prigogine and co-workers.[7–9] Their objective is a unified formulation of dynamics and thermodynamics, including a description of irreversible behavior at the microscopic level. The procedure is to find a transformation Λ which breaks the time-reversal symmetry in the sense that $\Lambda\rho(t)$ is defined only for $t\geqslant 0$ and evolves in time according to a semigroup. In quantum mechanics the transformed probability density $\Lambda\rho(t)$ has to be such that the quantity

$$\Omega(t) := \mathrm{Tr}\ \rho^*(t)\Lambda^*\Lambda\rho(t) \quad (t\geqslant 0) \tag{2.1}$$

is a monotone decreasing function of t. The operator $\Lambda^*\Lambda$ is called the microscopic entropy operator. The classical counterpart of the function $\Omega(t)$ is obtained if the trace in Eq. (2.1) is replaced by integration over the phase space. In agreement with Eq. (2.1), Theorem 6.2 states that the operator $\sigma(t)$ defined by Eq. (1.4) does have the property that $\mathrm{Tr}\ \sigma^2(t)$ is decreasing on $t \geqslant 0$.

Systems investigated by the Prigogine group divide into two categories calling for entirely different methods of analysis: classical dynamical systems[10] and large Hamiltonian systems, including quantum systems, with a continuous energy spectrum and infinitely many resonances.[11] We do not know of ideas in the literature that could be used to find an entropy operator for a single quantum particle without interaction.

In the context of classical dynamical systems, entropy operators were constructed explicitly for the baker map[12,13] and other Bernoulli systems,[14] as well as for K-flows.[15] This work has inspired many papers.[16–27] For an entropy operator $\Lambda^*\Lambda$ in a classical dynamical system to exist, it is necessary[28] that the time evolution of nonequilibrium states is generated by an operator whose spectrum is absolutely continuous, of uniform multiplicity, and running from $-\infty$ to ∞. Due to this condition, constructions for classical dynamical systems cannot be used in quantum mechanics.

In a quantum system with bound states, one could imagine an entropy that is defined only on the absolutely continuous subspace of the Hamiltonian, but this does not solve the problem that the spectrum of the Hamiltonian is bounded below. As outlined at the end of Sec. VIII, our operator $\sigma(t)$ can be viewed as a partial trace of a density operator $\tilde{\rho}(t)$ that changes in time according to a semigroup of isometries $\{U(t)|0\leqslant t<\infty\}$ with a generator whose spectrum consists of the lower half-plane plus the entire real axis.

Historically the Λ-transformation was first proposed for large Hamiltonian systems with many resonances.[7] The first example of this kind[7,29–32] was the Friedrichs model,[33] in which a discrete state embedded in a continuum disappears as the result of a perturbation. While the Friedrichs model may be solved explicitly, typical large Hamiltonian systems exhibit a high level of complexity. The existence of an operator Λ is closely related to the validity of kinetic equations that govern the relaxation of the system to equilibrium. To derive a kinetic theory, the Prigogine group uses the theory of subdynamics[7,34–39] which aims at constructing a complete set of projection operators that commute with the Liouville operator. This corresponds to identifying a complete set of subspaces in which time evolutions are mutually independent, hence the name subdynamics. It is an essential point in this theory that the Hamiltonian be decomposed into an unperturbed term plus a perturbation. The Λ-operator transforms unperturbed into perturbed projections.

One difficulty is that the perturbed subspaces do not belong to the unperturbed Hilbert space. Recent papers[32,40,41] therefore use Gel'fand triplets, as in the theory of rigged Hilbert spaces.[42–44] This amounts to including resonances in the spectral decomposition.[45] Closely related is an approach which treats decaying states as generalized eigenvectors with complex eigenvalues.[46] Either way, the time evolution is governed by unstable states. In the subdynamics theory irreversible behavior is due to resonances. A recent alternative[47–49] to the subdynamics theory is formulated entirely in terms of density operators because the dynamics cannot be implemented by wave functions. How this is related to the Λ-transformation is described in the original papers.[47,49]

Systems investigated by the Prigogine group are entirely different from the free particle in this paper, calling for entirely different methods of investigation. Yet a fundamental observation by Prigogine and coauthors[8,50] remains unchanged. They pointed out that any quantum operator $\Lambda^*\Lambda\rho$ cannot be written in the form $\Lambda^*\Lambda\rho = A_1\rho A_2$, where A_1 and A_2 are operators acting on the same Hilbert space as ρ. If $\Lambda^*\Lambda\rho$ did have this form, the transformation Λ would take wave functions into wave functions, hence pure states would go into pure states. In other words, states satisfying $\rho = \rho^2$ would go into states satisfying $\Lambda\rho = (\Lambda\rho)^2$, so that $\Omega(t)$ in Eq. (2.1) would not decrease as required. It follows that any operator Λ needs to act on operators and cannot act on wave functions. Our operator L taking ρ into σ agrees with this prediction.

III. THE DILATION OPERATOR

Recall that the Hamiltonian H_0 acts on $f(\mathbf{k}) \in L^2(\mathbb{R}^3)$ as multiplication by k^2. We write $|\mathbf{k}| = k$ and introduce two spherical polar coordinates ω. Identifying $f(\mathbf{k})$ and $f(k,\omega)$, we write L^2 instead of $L^2(\mathbb{R}^3)$.

The Mellin transform

$$Mf(k,\omega) = f\#(u,\omega)$$

defined by

$$f\#(u,\omega) = (2\pi)^{-1/2} \int_0^\infty k f(k,\omega) k^{iu-1/2} dk \tag{3.1}$$

is a unitary map taking L^2 onto the space $L^2(\mathbb{R}\times S^2)$ consisting of functions $f\#(u,\omega)$ with inner product

$$(f\#, g\#) = \int_{S^2} \int_{-\infty}^{\infty} f\#(u,\omega) \overline{g\#}(u,\omega) du\, d\omega.$$

We denote this space by ML^2. The inverse $M^{-1} f\#(u,\omega)$ is determined by

$$k f(k,\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} f\#(u,\omega) k^{-iu-1/2} du. \tag{3.2}$$

If we define $z := \ln k$ and

$$F(z,\omega) := f(e^z,\omega) e^{3z/2},$$

it follows from Eq. (3.2) that

$$F(z,\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} f\#(u,\omega) e^{-izu} du. \tag{3.3}$$

Hence many properties of the Mellin transform can be derived from properties of the Fourier transform. For example, if $f\# \in C_0^\infty(\mathbb{R}\times S^2)$, then F is in the domain of $i\partial/\partial z$. Applying $i\partial/\partial z$ to F corresponds to multiplying $f\#$ by u. Since $\partial/\partial z = k\partial/\partial k$, it is easy to verify that

$$i \frac{\partial}{\partial z} F(z,\omega) = k^{3/2} \left[ik \frac{\partial}{\partial k} + 3i/2 \right] f(k,\omega). \tag{3.4}$$

The operator in square brackets on the right is equal to the dilation operator D introduced by Eq. (1.2). Multiplying Eq. (3.4) by $k^{-1/2}$ and using Eq. (3.3) therefore gives

$$kDf(k,\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} u f\#(u,\omega) k^{-iu-1/2} du. \tag{3.5}$$

Up to this point, multiplication by u is restricted to C_0^∞ and D acts as a differential operator. We now extend multiplication by u from C_0^∞ to the set $\mathrm{Dom}(u) \in ML^2$ consisting of all $f\#(u,\omega) \in ML^2$ with the property that $uf\#(u,\omega) \in ML^2$. The operator u so defined is self-adjoint and has an absolutely continuous spectrum from $-\infty$ to ∞. As a result, D extends to a self-adjoint operator on L^2 with domain $\mathrm{Dom}(D) = M^{-1}\mathrm{Dom}(u)$ and the same spectrum as u. It follows from Eq. (3.5) that $D = M^{-1}uM$.

Once the self-adjoint operator D is available, we can define the positive operator $e^{\phi D}$ on L^2. Its domain $\mathrm{Dom}(e^{\phi D})$ is given by $M^{-1}\mathrm{Dom}(e^{\phi u})$, the inverse Mellin transform of the set of all $f\#(u,\omega) \in ML^2$ with the property that $e^{\phi u} f\#(u,\omega) \in ML^2$. Taking $\phi = -\pi/2$, we define

$$J := \exp(-\pi D/2) = M^{-1} \exp(-\pi u/2) M. \tag{3.6}$$

The set of all $f \in L^2$ with the property that $Jf \in L^2$ is denoted by K^2.

IV. ANALYTIC FUNCTIONS

For a description of the set K^2 we examine the class of all functions $f(ke^{i\phi},\omega)$ which are analytic in the sector $-\pi/2 < \phi < 0$ for almost every $\omega \in S^2$ and have the property that the integral

$$\int_{S^2} \int_0^\infty |ke^{i\phi} f(ke^{i\phi},\omega)|^2 dk d\omega$$

is bounded uniformly in the sector. It is known[51–53] that functions in this class have mean-square boundary values $f(k,\omega)$ and $f(ke^{-i\pi/2},\omega)$. Under the inner product

$$\langle f,g\rangle = \int_{S^2} \int_0^\infty [f(k,\omega)\overline{g}(k,\omega) + f(ke^{-i\pi/2},\omega)\overline{g}(ke^{-i\pi/2},\omega)]k^2 dk d\omega$$

the class is a Hilbert space which we denote by G^2.

For $f \in G^2$, we consider the Mellin transform

$$f\#(u,\omega,\phi) = (2\pi)^{-1/2} \int_0^\infty ke^{i\phi} f(ke^{i\phi},\omega) k^{iu-1/2} dk$$

and its inverse

$$ke^{i\phi} f(ke^{i\phi},\omega) = (2\pi)^{-1/2} \int_{-\infty}^\infty f\#(u,\omega,\phi) k^{-iu-1/2} du. \tag{4.1}$$

It is known[51–53] that a function f belongs to G^2 if and only if its Mellin transform is of the form

$$f\#(u,\omega,\phi) = e^{\phi u - i\phi/2} f\#(u,\omega,0) \tag{4.2}$$

with some function $f\#(u,\omega,0)$ satisfying

$$\int_{S^2} \int_{-\infty}^\infty (1+e^{-\pi u}) |f\#(u,\omega,0)|^2 du d\omega < \infty. \tag{4.3}$$

Given $f\#(u,\omega,0)$ satisfying Eq. (4.3), we can construct $f\#(u,\omega,\phi)$ according to Eq. (4.2). Since $f\#(u,\omega,0)$ is the same as the Mellin transform $f\#(u,\omega)$ of $f(k,\omega)$, it follows that $f(k,\omega)$ has an analytic continuation $f(ke^{i\phi},\omega) \in G^2$ if and only if its Mellin transform $f\#(u,\omega)$ satisfies Eq. (4.3). The condition expressed by Eq. (4.3) is that $f(k,\omega)$ and $Jf(k,\omega)$ belong to L^2. Hence $f(k,\omega) \in K^2$ if and only if $f(k,\omega)$ is the boundary value of a function $f(ke^{i\phi},\omega) \in G^2$.

Comparing Mellin transforms with the help of Eqs. (4.1) and (4.2) shows that

$$e^{\phi D} f(k,\omega) = e^{3i\phi/2} f(ke^{i\phi},\omega). \tag{4.4}$$

Taking $\phi = -\pi/2$ gives

$$Jf(k,\omega)=e^{-3i\pi/4}f(ke^{-i\pi/2},\omega). \tag{4.5}$$

If $-\pi/2<\phi<0$, then $\exp(-ik^2e^{2i\phi}t)$ is bounded if and only if $t\geqslant 0$. Hence, if $f\in G^2$, then $\exp(-ik^2e^{2i\phi}t)f\in G^2$ if $t\geqslant 0$ but not necessarily if $t<0$. In terms of boundary values at $\phi=0$, it follows that $f\in K^2$ has the property that $\exp(-ik^2t)f\in K^2$ if $t\geqslant 0$ but not necessarily if $t<0$. The set K^2 has a broken time-reversal symmetry. For $f\in K^2$, it follows from Eq. (4.5) that

$$J\ \exp(-ik^2t)f(k,\omega)=\exp(ik^2t)Jf(k,\omega)\quad (t\geqslant 0). \tag{4.6}$$

Unless stated otherwise, t is assumed to be non-negative in all of the following.

V. THE DENSITY OPERATOR ρ

This section uses Eq. (4.6) to derive identities satisfied by $J\rho(t)J$ and related operators. The results are used in Secs. VI and VII to study the density operator $\sigma(t)$ and the entropy $S[\sigma(t)]$ as functions of t.

Lemma 5.1: The density operator ρ is of the form (1.3) if and only if ρ and $J\rho$ belong to the trace class and ρJ and $J\rho J$ have closures in the trace class.

Proof: Suppose ρ has the form of Eq. (1.3). Since R is in the trace class by assumption and $(1+J^2)^{-1/2}$ and $J(1+J^2)^{-1/2}$ are bounded, ρ and $J\rho$ belong to the trace class. By the same token

$$(1+J^2)^{-1/2}RJ(1+J^2)^{-1/2}\ \text{ and }\ J(1+J^2)^{-1/2}RJ(1+J^2)^{-1/2} \tag{5.1}$$

belong to the trace class. On $f\in K^2$, these operators act as ρJ and $J\rho J$, respectively. Since K^2 is dense in L^2, it follows that ρJ and $J\rho J$ have closures given by expression (5.1). These closures belong to the trace class.

Now suppose that ρ, $J\rho$, and the closures of ρJ and $J\rho J$ belong to the trace class. Define

$$T:=(1+J)\rho(1+J). \tag{5.2}$$

This operator has domain K^2. By assumption, its closure $\overline{T}$ belongs to the trace class. Since J is positive, $(1+J^2)^{1/2}(1+J)^{-1}$ is bounded. Hence

$$R:=(1+J^2)^{1/2}(1+J)^{-1}\overline{T}(1+J^2)^{1/2}(1+J)^{-1}$$

belongs to the trace class. Solving for $\overline{T}$ gives

$$\overline{T}=(1+J)(1+J^2)^{-1/2}R(1+J)(1+J^2)^{-1/2}.$$

Hence

$$\rho=(1+J)^{-1}\overline{T}(1+J)^{-1}=(1+J^2)^{-1/2}R(1+J)(1+J^2)^{-1/2}(1+J)^{-1}=(1+J^2)^{-1/2}R(1+J^2)^{-1/2}. \tag{5.3}$$

This shows that ρ has the desired form, with R in the trace class.

If R is positive, then ρ is positive. Now suppose that ρ is positive. By Eq. (5.3) $(Rg,g)\geqslant 0$ whenever g is of the form $(1+J^2)^{-1/2}f$ with $f\in L^2$. Since the set of functions g is dense in L^2, it follows that R is positive. □

In case there is no danger of confusion, we henceforth write ρJ, $J\rho J$, and $J\rho(1+J^2)^{1/2}$ when we actually mean closures.

If ρ satisfies Eq. (1.3),

$$\begin{aligned}\mathrm{Tr}(1+J^2)^{1/2}\rho(1+J^2)^{1/2}&=\mathrm{Tr}(1+J^2)^{-1/2}\rho(1+J^2)^{1/2}+\mathrm{Tr}\ J^2(1+J^2)^{-1/2}\rho(1+J^2)^{1/2}\\&=\mathrm{Tr}\ \rho+\mathrm{Tr}\ J\rho J.\end{aligned}$$

Lemma 5.2: If ρ is of the form (1.3) and $t\geqslant 0$,

$$J\rho(t)=\exp(iH_0t)J\rho\ \exp(iH_0t). \tag{5.4}$$

If closures of all operators are understood,

$$\rho(t)J=\exp(-iH_0t)\rho J\ \exp(-iH_0t), \tag{5.5}$$

$$J\rho(t)J=\exp(iH_0t)J\rho J\ \exp(-iH_0t). \tag{5.6}$$

Proof: According to Lemma 5.1, ρ and $J\rho$ are bounded. Hence the range of ρ belongs to K^2. Equation (5.4) follows from Eq. (4.6). For any $f\in K^2$,

$$\rho(t)Jf=\exp(-iH_0t)\rho J\ \exp(-iH_0t)f.$$

The right-hand side can be extended to all $f\in L^2$ if ρJ is replaced by its closure identified in expression (5.1). It follows that $\rho(t)J$ is closable with closure in the trace class, the closure satisfying Eq. (5.5). To prove Eq. (5.6) we consider

$$J\rho(t)Jf=\exp(iH_0t)J\rho\ \exp(iH_0t)Jf=\exp(iH_0t)J\rho J\ \exp(-iH_0t)f.$$

According to expression (5.1) the operator on the right has a closure in the trace class. Hence, so does $J\rho(t)J$, and the two closures are equal. □

Corollary 5.3: If ρ is of the form (1.3) and $t>0$, there exists an operator $R(t)$ in the trace class such that $\rho(t)$ satisfies Eq. (1.3) with $R(t)$ instead of R.

Proof: Due to Lemma 5.2, $\rho(t)$, $J\rho(t)$, $\rho(t)J$, and $J\rho(t)J$ have closures in the trace class. Hence $R(t)$ in the trace class can be constructed following the second part of the proof of Lemma 5.1. □

Lemma 5.4: If ρ is of the form (1.3), $J^{1/2}\rho$ belongs to the trace class. The operators $\rho J^{1/2}$ and $J^{1/2}\rho J^{1/2}$ have closures in the trace class. If $t\geqslant 0$ and closures of all operators are understood

$$J^{1/2}\rho(t)=\exp(-H_0t)J^{1/2}\rho\ \exp(iH_0t), \tag{5.7}$$

$$\rho(t)J^{1/2}=\exp(-iH_0t)\rho J^{1/2}\ \exp(-H_0t), \tag{5.8}$$

$$J^{1/2}\rho(t)J^{1/2}=\exp(-H_0t)J^{1/2}\rho J^{1/2}\ \exp(-H_0t). \tag{5.9}$$

Proof: Because $J^{1/2}(1+J^2)^{-1/2}$ is bounded, the first part of the lemma can be proved in much the same way as the first part of Lemma 5.1. Taking $\phi=-\pi/4$ in Eq. (4.4) gives

$$J^{1/2}\ \exp(-ik^2t)f(k,\omega)=\exp(-k^2t)J^{1/2}f(k,\omega)\quad (t\geqslant 0) \tag{5.10}$$

for every $f\in K^2$. Since the range of ρ belongs to K^2 by Lemma 5.1, Eq. (5.7) follows.

Applying $J^{1/2}$ to Eq. (5.10) and using Eq. (4.6) gives

$$\exp(ik^2t)Jf(k,\omega)=J\ \exp(-ik^2t)f(k,\omega)=J^{1/2}\ \exp(-k^2t)J^{1/2}f(k,\omega),$$

hence

$$\exp(iH_0t)J^{1/2}g=J^{1/2}\ \exp(-H_0t)g \tag{5.11}$$

for every g of the form $g=J^{1/2}f$, with $f\in K^2$.

We claim that the set of functions g being considered in Eq. (5.11) is dense in L^2. Indeed, for f to belong to K^2, it is sufficient that the Mellin transform $f\#$ is of the form $(1+e^{-\pi u/2})^{-1}h\#$ with $h\#\in ML^2$. Hence it is sufficient that $g\#$ is of the form

$$g\#=e^{-\pi u/4}(1+e^{-\pi u/2})^{-1}h\#=[2\cosh(\pi u/4)]^{-1}h\#. \tag{5.12}$$

Suppose N exists such that $g\#(u,\omega)=0$ if $|u|>N$. Solving Eq. (5.12) for $h\#$ gives $h\#\in ML^2$, hence $g\#$ is of the desired form. Since the set of all $g\#$ with compact support is dense in ML^2, it follows that Eq. (5.11) is true on a dense set, as claimed.

Applying $\exp(-iH_0t)\rho$ to Eq. (5.11) gives

$$\rho(t)J^{1/2}g=\exp(-iH_0t)\rho J^{1/2}\ \exp(-H_0t)g. \tag{5.13}$$

Replacing $\rho J^{1/2}$ by its closure constructed as in the proof of Lemma 5.1, we can extend the right-hand side of Eq. (5.13) to all of L^2. Hence $\rho(t)J^{1/2}$ is closable with closure in the trace class satisfying Eq. (5.8). The proof of Eq. (5.9) is similar. □

Lemma 5.5: If ρ is of the form (1.3), the operator $J\rho J^{1/2}$ is closable. Its closure belongs to the trace class. If $t \geqslant 0$ and closures of operators are understood

$$J\rho(t)J\rho(t) = \exp(iH_0t)J\rho J^{1/2} \exp(-2H_0t)J^{1/2}\rho \exp(iH_0t). \tag{5.14}$$

Proof: Given the proofs of Lemmas 5.1 and 5.4, the statement about $J\rho J^{1/2}$ is easy to verify. Thus, the closures of $\rho J^{1/2}$ and $J\rho J^{1/2}$ are bounded. This means that the range of $\rho J^{1/2}$ belongs to K^2. Hence

$$J \exp(-iH_0t)\rho J^{1/2} = \exp(iH_0t)J\rho J^{1/2}.$$

Applying $J \exp(-iH_0t)\rho$ to Eq. (5.11) gives

$$J\rho(t)J^{1/2}g = J \exp(-iH_0t)\rho J^{1/2} \exp(-H_0t)g = \exp(iH_0t)J\rho J^{1/2} \exp(-H_0t)g \tag{5.15}$$

on a dense set of functions g. As in Lemmas 5.2 and 5.4, the operators in Eq. (5.15) have closures that are equal. Applying these to $J^{1/2}\rho(t)$ and using Eq. (5.7) gives the desired relation (5.14).□

VI. THE DENSITY OPERATOR σ

Theorem 6.1: *If ρ is of the form (1.3), the operator $\sigma(t)$ defined by Eq. (1.4) is positive and belongs to the trace class. Its trace satisfies Eq. (1.5).*

Proof: The operator $\sigma(t)$ belongs to the trace class due to Lemma 5.1 and Eq. (5.6). By the same token, Tr $\sigma(t)$ is constant on $t \geqslant 0$. That $\sigma(t)$ is positive follows from the assumption that R is positive. □

The next theorem states that Tr $\sigma^2(t)$ is decreasing, like the entropy operator (2.1) in the Prigogine theory.

Theorem 6.2: *If ρ is of the form (1.3), the operator $\sigma(t)$ defined by Eq. (1.4) has the property that*

$$\text{Tr } \sigma^2(s) > \text{Tr } \sigma^2(t) \quad \text{if } 0 \leqslant s < t, \tag{6.1}$$

$$\lim_{t\to\infty} \text{Tr } \sigma^2(t) = \text{Tr } \rho^2 + \text{Tr}(J\rho J)(J\rho J). \tag{6.2}$$

Proof: Due to Lemmas 5.2, 5.4, and 5.5,

$$\begin{aligned}
\text{Tr } \rho(t)[J\rho(t)J] &= \text{Tr}[J\rho(t)J]\rho(t) \\
&= \text{Tr}[J\rho(t)J^{1/2}][J^{1/2}\rho(t)] \\
&= \text{Tr}[J^{1/2}\rho(t)][J\rho(t)J^{1/2}] \\
&= \text{Tr}[J^{1/2}\rho(t)J^{1/2}][J^{1/2}\rho(t)J^{1/2}] \\
&= \text{Tr}[\exp(-H_0t)J^{1/2}\rho J^{1/2} \exp(-2H_0t)J^{1/2}\rho J^{1/2} \exp(-H_0t)].
\end{aligned} \tag{6.3}$$

Let T be the closure of $J^{1/2}\rho J^{1/2}$. Then

$$\begin{aligned}
&\text{Tr}[\exp(-H_0s)T \exp(-2H_0s)T \exp(-H_0s)] - \text{Tr}[\exp(-H_0t)T \exp(-2H_0t)T \exp(-H_0t)] \\
&\quad = \text{Tr}\{\exp(-H_0s)T[\exp(-2H_0s) - \exp(-2H_0t)]T \exp(-H_0s)\} \\
&\qquad + \text{Tr}\{\exp(-H_0t)T[\exp(-2H_0s) - \exp(-2H_0t)]T \exp(-H_0t)\}.
\end{aligned} \tag{6.4}$$

If $0 \leqslant s < t$, the operator $[\exp(-2H_0s) - \exp(-2H_0t)]$ is strictly positive. The trace is faithful. Hence the right-hand side of Eq. (6.4) is the sum of two positive traces. It follows that

$$\text{Tr } \rho(s)J\rho(s)J > \text{Tr } \rho(t)J\rho(t)J \quad \text{if } 0 \leqslant s < t.$$

By the definition of $\sigma(t)$ and Eq. (5.6)

$$\mathrm{Tr}\ \sigma^2(t)=\mathrm{Tr}\ \rho^2+\mathrm{Tr}\ \rho(t)J\rho(t)J+\mathrm{Tr}\ J\rho(t)J\rho(t)+\mathrm{Tr}(J\rho J)(J\rho J). \tag{6.5}$$

Terms one and four on the right do not depend on t while terms two and three are decreasing. This proves Eq. (6.1).

Recall that the operator T belongs to the trace class. Since $\exp(-H_0t)$ is bounded and tends to 0 strongly as $t\to\infty$, the operator $\exp(-H_0t)T\exp(-H_0t)$ tends to 0 in the trace norm as $t\to\infty$. It follows with Eq. (6.3) that $\mathrm{Tr}\ \rho(t)J\rho(t)J$ tends to 0 as $t\to\infty$. Equation (6.2) now follows from Eq. (6.5). □

We proceed to show how $\rho(t)$ can be recovered if Eq. (1.3) is true and $\sigma(t)$ is known. To simplify the notation, we first examine ρ and σ at $t=0$.

Since the trace class is a subset of the Schmidt class, its members are integral operators. The integral kernel of ρ is a $L^2\times L^2$-function $\rho(k,\omega;k',\omega')$. The Mellin transform $\rho\#:=M\rho M^{-1}$ is an integral operator with $ML^2\times ML^2$-kernel $\rho\#(u,\omega;v,\omega')$. By Eq. (3.6) MJM^{-1} acts as multiplication by $\exp(-\pi u/2)$. Hence $MJ\rho JM^{-1}$ has integral kernel

$$\exp(-\pi u/2)\rho\#(u,\omega;v,\omega')\exp(-\pi v/2).$$

Theorem 6.3: *Let L be a linear operator whose domain consists of all operators ρ of the form (1.3). Let L act according to*

$$L\rho=\rho+J\rho J.$$

Define $\rho\#:=M\rho M^{-1}$. Let $L\#$ act according to

$$L\#\rho\#=\rho\#+MJ\rho JM^{-1},$$

where the domain of $L\#$ consists of all operators $\rho\#$ with the property that $M^{-1}\rho\#M$ is of the form (1.3). If ρ is of the form (1.3), σ defined by Eq. (1.4) equals $L\rho$. Its Mellin transform $\sigma\#:=M\sigma M^{-1}$ equals $L\#\rho\#$. The operator $L\#$ transforms integral kernels according to

$$\sigma\#(u,\omega;v,\omega')=(L\#\rho\#)(u,\omega;v,\omega')=\{1+\exp[-\pi(u+v)/2]\}\rho\#(u,\omega;v,\omega'). \tag{6.6}$$

There is an inverse $(L\#)^{-1}$ defined on the range of $L\#$. It recovers the integral kernel of $\rho\#$ according to

$$\rho\#(u,\omega;v,\omega')=[(L\#)^{-1}\sigma\#](u,\omega;v,\omega')=\{1+\exp[-\pi(u+v)/2]\}^{-1}\sigma\#(u,\omega;v,\omega'). \tag{6.7}$$

The relations (6.6) and (6.7) remain true when ρ and σ are replaced by $\rho(t)$ and $\sigma(t)$, respectively.

Proof: Lemma 5.1 states that Eq. (1.3) is sufficient for ρ and $J\rho J$ to belong to the trace class. Since M is unitary, $M\rho M^{-1}$ and $MJ\rho JM^{-1}$ belong to the trace class. Hence L and $L\#$ are well defined. It is easy to see that σ equals $L\rho$ and $\sigma\#$ equals $L\#\rho\#$. These operators belong to the trace class. This leads to Eq. (6.6) as explained above. Since $\{1+\exp[-\pi(u+v)/2]\}^{-1}$ is a bounded function of u and v, Eq. (6.7) follows immediately.

If ρ is of the form (1.3), then $\rho(t)$ is of this form for all $t>0$. This is the statement of Corollary 5.3. Since L and the Mellin transform do not depend on t, the reasoning that proved Eqs. (6.6) and (6.7) can be repeated for every $t>0$. □

The next theorem shows that the operator $\sigma(t)$ can be used to calculate expectation values of observables.

Theorem 6.4: *Let A be a bounded operator on $L^2(\mathbb{R}^3)$. Under the assumptions of Theorem 6.3 there is an operator AL^{-1} defined on the range of L. If $\sigma(t)$ belongs to the range of L, then $(AL^{-1})\sigma(t)$ is in the trace class and Eq. (1.9) is satisfied.*

Proof: Let $A\#$ be the Mellin transform of A and define

$$(AL^{-1})\#:=A\#\{1+\exp[-\pi(u+v)/2]\}^{-1}.$$

Due to Eq. (6.6) $(AL^{-1})\#\sigma\#=A\#\rho\#$, hence $M^{-1}(AL^{-1})\#M$ satisfies the conditions of the theorem. □

Although ρ can be recovered from σ and σ can be used to calculate expectation values of observables, it is clear that ρ and σ are not equivalent as density operators. Indeed, our objective is to show that $\sigma(t)$ gives rise to an increasing entropy, while $\rho(t)$ does not. Theorem 6.2 states that Tr $\sigma^2(t)$ is decreasing while Tr $\rho^2(t)$ is not. Theorem 6.4 implies another distinction. To represent an observable, A must be self-adjoint on $L^2(\mathbb{R}^3)$. The operator AL^{-1} does not act on wave functions in $L^2(\mathbb{R}^3)$ but on operators in the range of L. For AL^{-1} to represent an observable, the obvious condition is that $(AL^{-1})L$ be self-adjoint on $L^2(\mathbb{R}^3)$.

VII. THE ENTROPY

Let A be a positive compact operator. We denote by $\gamma_n(A)$ the nth largest eigenvalue of A, counting multiplicity. If ρ satisfies Eq. (1.3), ρ and $J\rho J$ belong to the trace class. Hence

$$\sum_n \gamma_n(\rho)<\infty, \quad \sum_n \gamma_n(J\rho J)<\infty. \tag{7.1}$$

Referring to Eq. (1.1) we define

$$s(\gamma):=\begin{cases} -\gamma \ln \gamma & \text{if } \gamma>0 \\ 0 & \text{if } \gamma=0 \end{cases}$$

and assume

$$\begin{aligned} S(\rho)&:=\sum_n s[\gamma_n(\rho)]<\infty, \\ S(J\rho J)&:=\sum_n s[\gamma_n(J\rho J)]<\infty. \end{aligned} \tag{7.2}$$

Lemma 7.1: If ρ is of the form (1.3), Eq. (7.2) is satisfied, and $t\geqslant 0$, then

$$S[\rho(t)]=S(\rho), \quad S[J\rho(t)J]=S(J\rho J).$$

Proof: This follows from Eq. (5.6) and the fact that $\exp(-iH_0t)$ is unitary. □

Due to Eq. (7.1), at most a finite number of eigenvalues γ_n may exceed 1. Hence the sums in Eq. (7.2) converge absolutely. With Eq. (1.6) and Lemma 7.1, $S[\sigma(t)]$ is bounded above. To show that $S[\sigma(t)]$ is bounded below, we define

$$\nu(t):=\sigma(t)/\text{Tr } \sigma(t). \tag{7.3}$$

This gives

$$S[\sigma(t)]=[\text{Tr } \sigma(t)]\{S[\nu(t)]-\ln \text{Tr } \sigma(t)\}. \tag{7.4}$$

Since $\nu(t)$ is positive with Tr $\nu(t)=1$, all eigenvalues of $\nu(t)$ lie in the interval $[0,1]$, hence $S[\nu(t)]\geqslant 0$. By Theorem 6.1 Tr $\sigma(t)$ is a finite positive constant Tr σ. Hence

$$S[\sigma(t)]\geqslant -(\text{Tr } \sigma)\ln \text{Tr } \sigma.$$

We know from the proof of Theorem 6.2 that Tr $\rho(t)\ J\rho(t)J$ tends to 0 as $t\to\infty$. The next result is significant in view of the necessary and sufficient condition (1.7) for the entropy $S[\sigma(t)]$ to take its maximum possible value.

Lemma 7.2: If ρ is of the form (1.3), $\rho(t)\ J\rho(t)J$ and $J\rho(t)\ J\rho(t)$ tend to 0 in the trace norm as $t\to\infty$.

Proof: By Lemma 5.4 $J^{1/2}\rho$ belongs to the trace class. The operator $\exp(-2H_0t)$ is bounded if $t\geqslant 0$ and tends to 0 strongly as $t\to\infty$. Hence $\exp(-2H_0t)\ J^{1/2}\rho$ tends to 0 in the trace norm. With $J\rho J^{1/2}$ in the trace class by Lemma 5.5, $J\rho(t)\ J\rho(t)$ tends to 0 in the trace norm by Eq. (5.14). So does its adjoint $\rho(t)\ J\rho(t)J$. □

We proceed to prove two lemmas that will be used to show that the entropy $S[\sigma(t)]$ tends to its least upper bound $S(\rho)+S(J\rho J)$ as $t\to\infty$.

Lemma 7.3: Define

$$\pi_+(t):=P_+\rho(t)P_+, \quad \pi_-(t):=P_-J\rho(t)JP_-,$$

where P_+ and P_- are the spectral projections onto the subspaces of L^2 in which the dilation operator D is positive and negative, respectively. If ρ is of the form (1.3), the set of all nonvanishing eigenvalues $\gamma_k[\pi_+(t)+\pi_-(t)]$ $(k=1,2,...)$ is equal to the union of the sets of all nonvanishing $\gamma_n[\pi_+(t)]$ and $\gamma_m[\pi_-(t)]$ $(m,n=1,2,...)$.

Proof: To simplify the notation, we give the proof for $t=0$. The proof for $t>0$ is similar.

Let ψ_+ be an eigenvector of π_+ with eigenvalue $\gamma\neq 0$. Then $\psi_+=P_+\psi_+$, hence

$$(\pi_++\pi_-)\psi_+=\pi_+\psi_+=\gamma\psi_+.$$

Similarly, an eigenvector ψ_- of π_- with eigenvalue $\gamma\neq 0$ is an eigenvector of $\pi_++\pi_-$ with the same eigenvalue. If γ has multiplicity M_+ as eigenvalue of π_+ and multiplicity M_- as eigenvalue of π_-, the eigenvectors ψ_+ and ψ_- yield a set of M_++M_- linearly independent eigenvectors of $\pi_++\pi_-$. Hence γ is an eigenvalue of $\pi_++\pi_-$ with multiplicity $M\geqslant M_++M_-$.

Now suppose ψ is an eigenvector of $\pi_++\pi_-$ with eigenvalue γ. Then

$$\pi_+P_+\psi=(\pi_++\pi_-)P_+\psi=P_+(\pi_++\pi_-)\psi=\gamma P_+\psi. \tag{7.5}$$

Hence $P_+\psi$ and $P_-\psi$ are eigenvectors of $\pi_++\pi_-$. Let γ have multiplicity M as eigenvalue of $\pi_++\pi_-$ and suppose there are m linearly independent eigenvectors $P_+\psi$. Then $0\leqslant m\leqslant M$ and $\pi_++\pi_-$ has $M-m$ eigenvectors $P_-\psi$. Due to Eq. (7.5) γ is an eigenvalue of π_+ with multiplicity $M_+\geqslant m$. By the same token, γ is an eigenvalue of π_- with multiplicity $M_-\geqslant M-m$. Hence $M_++M_-\geqslant M$. By the first part of the proof $M\geqslant M_++M_-$. Hence $M=M_++M_-$. There is a one-to-one correspondence, counting multiplicity, between nonvanishing eigenvalues $\gamma_k[\pi_++\pi_-]$ on the one hand and nonvanishing γ in the union of all $\gamma_n[\pi_+]$ and $\gamma_m[\pi_-]$ on the other. □

Lemma 7.4: Let ρ be of the form (1.3). Let λ_k $(k=1,2,...)$ be the kth largest number, counting multiplicity, in the set consisting of all nonvanishing eigenvalues $\gamma_n(\rho)$ and $\gamma_m(J\rho J)$ $(m,n=1,2,...)$. In terms of the notation introduced in Lemma 7.3

$$\lim_{t\to\infty}\gamma_k[\pi_+(t)+\pi_-(t)]=\lambda_k. \tag{7.6}$$

Proof: If A and B are compact and positive[54]

$$|\gamma_n(A)-\gamma_n(B)|\leqslant\|A-B\|. \tag{7.7}$$

Since $P_-\exp(-iH_0t)$ tends to 0 strongly as $t\to\infty$, by Eq. (1.8), $P_-\rho(t)$ tends to 0 in the trace norm. The trace norms of $P_-\rho(t)$ and its adjoint are equal. Hence $\rho(t)P_-$ tends to 0 in the trace norm. Denoting the trace norm of an operator C in the trace class by $\|C\|_1$, we find

$$\lim_{t\to\infty}\|\pi_+(t)-\rho(t)\|_1=0. \tag{7.8}$$

Hence, given $\epsilon>0$, there exists $T>0$, not depending on n, such that

$$|\gamma_n[\pi_+(t)]-\gamma_n(\rho)|<\epsilon \quad \text{if} \quad t>T. \tag{7.9}$$

By a similar argument, T may be chosen so large that also

$$|\gamma_m[\pi_-(t)]-\gamma_m(J\rho J)|<\epsilon \quad \text{if} \quad t>T. \tag{7.10}$$

According to Lemma 7.3, $\gamma\neq 0$ is an eigenvalue of $\pi_+(t)+\pi_-(t)$ if and only if γ is an eigenvalue of $\pi_+(t)$ and/or $\pi_-(t)$. This suggests that each $\gamma[\pi_+(t)+\pi_-(t)]\neq 0$ tends to some λ_k as $t\to\infty$. Since Lemma 7.3 states that multiplicities match, λ_k with multiplicity M should be

the limit of M eigenvalues $\gamma[\pi_+(t)+\pi_-(t)]$. It is a complication, however, that graphs of eigenvalues $\gamma_n[\pi_+(t)]$ and $\gamma_m[\pi_-(t)]$ may intersect. Thus, for instance, it could happen that $\gamma_1[\pi_+(t)+\pi_-(t)]$ equals $\gamma_1[\pi_+(t)]$ at time t_1, but $\gamma_1[\pi_-(t)]$ at time t_2. To cope with this difficulty, we choose any integer l with the property that $\lambda_{l+1}<\lambda_l$. The sequence $\lambda_1,\lambda_2,\ldots,\lambda_l$ will contain i eigenvalues $\gamma_n(\rho)$ and j eigenvalues $\gamma_m(J\rho J)$, with $i+j=l$. These eigenvalues are the limits of i eigenvalues $\gamma_n[\pi_+(t)]$ and j eigenvalues $\gamma_m[\pi_-(t)]$. We claim that T may be chosen so large that the graphs of this set of $i+j$ eigenvalues do not intersect at times $t>T$, except, possibly, for intersections of graphs that tend to the same limit λ_k $(k=1,2,\ldots,l)$.

To prove this, we choose ϵ so small that the difference between successive λ's in the sequence $\lambda_1,\lambda_2,\ldots,\lambda_{l+1}$ exceeds 2ϵ. By Eq. (7.9) T exists such that at $t>T$ each $\gamma_n[\pi_+(t)]$ is within ϵ from some λ_q, and similarly for each $\gamma_m[\pi_-(t)]$ $(n,m=1,2,\ldots)$. In particular, the eigenvalues $\gamma_n[\pi_+(t)]$ and $\gamma_m[\pi_-(t)]$ that tend to λ_q with $q=l+1,\ l+2,\ldots$ do not exceed $\lambda_{l+1}+\epsilon<\lambda_l-\epsilon$ when $t>T$. This leaves precisely l eigenvalues in the intervals $(\lambda_k-\epsilon,\ \lambda_k+\epsilon)$ with $k=1,2,\ldots,l$. By the choice of ϵ, the ϵ-intervals centered at distinct λ_k $(k=1,2,\ldots,l)$ are disjoint. Hence the l eigenvalues being considered do not change intervals when $t>T$. Graphs of eigenvalues that tend to different limits λ_k $(k=1,2,\ldots,l)$ do not intersect when $t>T$. If λ_k is simple, the ϵ-interval centered at λ_k contains precisely one eigenvalue $\gamma_n[\pi_+(t)]$ or $\gamma_m[\pi_-(t)]$. By Lemma 7.3, this is a $\gamma_p[\pi_+(t)+\pi_-(t)]$. In this case p equals k. If λ_k is repeated M times, $\lambda_k=\lambda_{k+1}=\cdots=\lambda_{k+M-1}$, the ϵ-interval centered at λ_k contains M eigenvalues $\gamma_p[\pi_+(t)+\pi_-(t)]$. If at any time $t>T$ the M eigenvalues are ordered, they receive labels $k,\ k+1,\ldots,k+M-1$. In the course of time the ordering within the group of M may change, but it remains true that

$$|\gamma_{k+r}[\pi_+(t)+\pi_-(t)]-\lambda_{k+r}|<\epsilon \quad \text{if} \quad t>T, \quad r=0,1,\ldots,M-1.$$

Summarizing, choose l with the property that $\lambda_{l+1}<\lambda_l$. Choose ϵ so small that the difference between successive distinct λ's in the sequence $\lambda_1,\lambda_2,\ldots,\lambda_{l+1}$ exceeds 2ϵ. Now T exists such that Eqs. (7.9) and (7.10) are true when $t>T$. By what we have shown

$$|\gamma_k[\pi_+(t)+\pi_-(t)]-\lambda_k|<\epsilon \quad \text{if} \quad t>T, \quad k=1,2,\ldots,l.$$

Hence Eq. (7.6) is true. Since ϵ depends on l, the convergence is not uniform in k. It is uniform on any finite set $k=1,2,\ldots,l$. □

Theorem 7.5: *Let $\sigma(t)$ be defined by Eq. (1.4). Let $S[\sigma(t)]$ be defined as in Eq. (1.1). If ρ is of the form (1.3) and Eq. (7.2) is satisfied*

$$\lim_{t\to\infty} S[\sigma(t)]=S(\rho)+S(J\rho J). \tag{7.11}$$

Proof: By Eq. (7.8) and the corresponding equation with $\pi_-(t)$ and $J\rho(t)J$

$$\lim_{t\to\infty}\|\pi_+(t)+\pi_-(t)-\sigma(t)\|_1=0.$$

Due to Eqs. (7.6) and (7.7)

$$\lim_{t\to\infty}\gamma_k[\sigma(t)]=\lim_{t\to\infty}\gamma_k[\pi_+(t)+\pi_-(t)]=\lambda_k.$$

By Eq. (7.2) and the definition of λ_k in Lemma 7.4

$$\sum_k s(\lambda_k)=S(\rho)+S(J\rho J).$$

To prove Eq. (7.11), let us adopt the convention that $0\ln 0=0$. Then it is sufficient to show that

$$\lim_{t\to\infty}\sum_k \gamma_k[\sigma(t)]\ln\gamma_k[\sigma(t)]=\sum_k \lambda_k\ln\lambda_k, \tag{7.12}$$

hence that on the left the sum over k and the limit with respect to t may be interchanged. Our proof uses the dominated convergence theorem for sums. This requires a suitable estimate on $\gamma_k[\sigma(t)]$.

If A and B are compact and positive

$$\gamma_{i+j+1}(A+B) \leqslant \gamma_{i+1}(A) + \gamma_{j+1}(B) \quad (i,j=0,1,2,...).$$

This is a special case of an inequality due to Fan.[54,55] Take $A=\rho(t)$, $B=J\rho(t)J$, so that $A+B=\sigma(t)$. Take $i+j=k$. Suppose the sequence $\lambda_1,\ldots,\lambda_k$ contains i eigenvalues $\gamma_n(\rho)$ and j eigenvalues $\gamma_m(J\rho J)$. Then $\gamma_{i+1}(\rho) \leqslant \lambda_k$ and $\gamma_{j+1}(J\rho J) \leqslant \lambda_k$, so that $\gamma_{k+1}[\sigma(t)] \leqslant 2\lambda_k$. Since $\Sigma_k\, \lambda_k$ converges by Eq. (7.1), there is a finite N such that $\lambda_k \leqslant 1/2e$ when $k \geqslant N$. Since the function $-x\, \ln x$ is nondecreasing on $[0,1/e]$, it follows that

$$0 \leqslant -\gamma_{k+1}[\sigma(t)] \ln\, \gamma_{k+1}[\sigma(t)] \leqslant -2\lambda_k\, \ln 2\lambda_k \quad (k \geqslant N).$$

By Eq. (7.1) and assumption (7.2) the sum $-\Sigma_k\, \lambda_k \ln 2\lambda_k$ converges. Dominated convergence therefore assures that

$$\lim_{t\to\infty} \sum_{k=N}^{\infty} = \sum_{k=N}^{\infty} \lim_{t\to\infty}$$

on the left in Eq. (7.12). Adding the limits of the terms with $k=1,2,...,N-1$ gives the desired result (7.12). This completes the proof of Eq. (7.11). □

Theorem 7.6: *If the data are as in Theorem 7.5, the entropy $S[\sigma(t)]$ tends to its maximum possible value as $t\to\infty$.*

Proof: The proposition follows from Theorem 7.5 with Eq. (1.6) and Lemma 7.1. □

VIII. DISCUSSION

Let σ be equal to $L\rho$ as in Theorem 6.3. If ρ is a positive operator, so is σ. Given $\sigma\# = L\#\rho\#$, the operator $(L\#)^{-1}$ correctly recovers the positive operator $\rho\#$.

Now choose a positive operator σ in the trace class and construct

$$\tau\#(u,\omega;v,\omega') = \{1+\exp[-\pi(u+v)/2]\}^{-1}\sigma\#(u,\omega;v,\omega'). \tag{8.1}$$

This is the integral kernel of a bounded operator $\tau\#$. The Appendix gives examples in which the operator $\tau\#$ is not positive. Thus there is a distinction between ρ and σ not unlike a distinction between ρ and $\Lambda\rho$ in the Prigogine theory. Among other requirements, Λ must be positivity preserving, in order that densities are transformed into densities. For classical dynamical systems on a measure space with an invariant measure, it was shown that Λ^{-1} is not positivity preserving.[56]

It will be understood that this paper is meant only for operators σ of the form $L\rho$ with positive ρ satisfying Eq. (1.3). Let us call this the allowed class. If c is a positive constant and σ belongs to the allowed class, so does $c\sigma$. If σ_1 and σ_2 belong to the allowed class, so does $\sigma_1+\sigma_2$. Hence the allowed class is a cone in the trace class.

In published examples of entropy operators for classical dynamical systems,[12–15] the phase space is compact and there is a uniform equilibrium density ρ_0. The Λ-operator has the property that $\|\Lambda\rho(t)-\rho_0\|$ is strictly monotonically decreasing to 0 as $t\to\infty$, for all initial densities $\rho(0) \neq \rho_0$. The authors referred to strong Markov semigroups[12,13] or monotonic Markov semigroups.[14,15] These semigroups give rise to nondecreasing entropies[57–60] that tend to their maximal values of zero[61] as $t\to\infty$.

That an entropy may exhibit fluctuations is illustrated by computer calculations to determine the Boltzmann H-function for a system of two-dimensional hard disks.[7,8,62] For the entropy to increase, the H-function has to decrease, but under suitable conditions it was found to increase over limited periods of time.

Another example[59,60] is provided by the quadratic map $f(x)=4x(1-x)$ with phase space $[0,1]$. The Perron–Frobenius operator P for this map can be determined in closed form. Taking $\rho(x)=1$ as initial density, one finds $P\rho(x)=[4(1-x)]^{-1/2}$ with entropy

$$S(P\rho) = -\int_0^1 [P\rho(x)]\ln[P\rho(x)]dx = -1 + \ln 2.$$

This is less than the initial entropy $S(\rho)=0$. The equilibrium density in this example is given by $\rho_0(x)=[\pi^2 x(1-x)]^{-1/2}$. Essential for the example is that ρ_0 is not the uniform density and that the map is not invertible.[59,60]

As for the entropy in the present paper, we do not know whether $S[\sigma(t)]$ is a monotone function of t. First suppose that $\rho(t)$ is of rank 1, so that $\sigma(t)$ is of rank 2. Denote the eigenvalues of $\sigma(t)$ by $\lambda_1(t)$ and $\lambda_2(t)$. By Theorems 6.1 and 6.2, $\lambda_1(t)+\lambda_2(t)$ is constant while $\lambda_1^2(t)+\lambda_2^2(t)$ is decreasing. This requires that $[\lambda_1(t)-\lambda_2(t)]^2$ be decreasing. One implication is, that $\lambda_1(t)$ must not be equal to $\lambda_2(t)$ at any finite time. Thus, an operator of rank 2 with equal eigenvalues does not belong to the allowed class. Suppressing the variable t, we label the eigenvalues so that $\lambda_1>\lambda_2$. In terms of $\lambda:=(\lambda_1+\lambda_2)/2$ and $\mu:=(\lambda_1-\lambda_2)/2$, the entropy takes the form

$$S[\sigma(t)] = -(\lambda+\mu)\ln(\lambda+\mu) - (\lambda-\mu)\ln(\lambda-\mu).$$

Since λ is constant and μ decreasing, the entropy is an increasing function of t on $t\geqslant 0$.

It is an open problem whether the foregoing result generalizes to density operators of higher rank. Suppose $0\leqslant s\leqslant t$ and compare $\sigma(s)$ with $\sigma(t)$. There exist several conditions[63] that are necessary and sufficient in order that

$$\mathrm{Tr}\ f[\sigma(t)] \geqslant \mathrm{Tr}\ f[\sigma(s)]$$

for every concave function f. These conditions are difficult to verify, however. They may be stronger than required for the entropy to be non-decreasing.

Theorems 6.1 and 6.2 state that the trace norm of $\sigma(t)$ is constant in time while the Schmidt norm is decreasing. The normalized operator $\nu(t)$ defined in Eq. (7.3) has the same properties. Suppose we could extend the two theorems to all operators of finite rank. Then we could use the Calderón–Lions theorem to interpolate between the trace and the Schmidt norms,[64] showing that $\mathrm{Tr}[\nu(\tau)]^\alpha$ is a nonincreasing function of $t\geqslant 0$ for fixed α with $1\leqslant\alpha\leqslant 2$. As a result the Rényi entropies, often called α-entropies,[65]

$$S_\alpha[\nu(t)] := (1-\alpha)^{-1}\ \ln\ \mathrm{Tr}[\nu(t)]^\alpha\ (1<\alpha\leqslant 2)$$

would be nondecreasing. Since $S_\alpha[\nu(t)]$ tends to $S[\nu(t)]$ as α tends to 1, the entropy $S[\nu(t)]$ would be nondecreasing. By Eq. (7.4), this result would carry over to $S[\sigma(t)]$. In a time-independent setting a similar reasoning has been used to study the change in entropy when a measurement is performed on a quantum system.[66]

While the Rényi entropy $S_2[\sigma(t)]$ is increasing by Theorem 6.2, the interpolation method does not apply in this paper because Theorem 6.1 does not generalize well beyond the allowed class of operators σ determined by positive ρ. To see what may happen, consider a spherically symmetric situation and take

$$\sigma\#(u,v) = \exp(-3\pi|u|/4)\exp(-3\pi|v|/4).$$

This is the kernel of a positive operator. By Example (A2) in the appendix, $\sigma\#$ determines an operator $\tau\#$, according to Eq. (8.1), which is not positive. It is easy to verify that τ, $J\tau$, τJ, and $J\tau J$ belong to the Schmidt class. Whether these operators belong to the trace class would be difficult to determine. There is a sufficient condition for a function to be the integral kernel of an operator in the trace class, but it requires that the operator be positive.[67,68] In any case, there is no difficulty constructing $\tau(t)$ and

$$J\tau(t)J = \exp(iH_0t)J\tau J\ \exp(-iH_0t).$$

These operators determine $\sigma(t)$ according to

$$\sigma(t) = \tau(t) + J\tau(t)J.$$

There are spectral decompositions

$$\tau=\tau_+ + \tau_-, \quad J\tau J=(J\tau J)_+ +(J\tau J)_-$$

in terms of positive and negative parts in the Schmidt class.[69] To make σ positive, $(J\tau J)_+$ must cancel τ_- while τ_+ must cancel $(J\tau J)_-$. Since τ and $J\tau J$ have different time evolutions, however, their negative parts are unlikely to be canceled at later times. Hence $\sigma(t)$ is unlikely to be positive. Assuming that τ and $J\tau J$ have finite traces, Tr$\sigma(t)$ is equal to Tr σ, but if $\sigma(t)$ is not positive, Tr $\sigma(t)$ is not equal to the trace norm $\|\sigma(t)\|_1$ required to interpolate between trace and Schmidt norms.

Recall that the allowed class of operators σ is a cone in the trace class. There is an abstract interpolation theory for cones,[70,71] but its interpolating norms are difficult to identify. We do not know whether this is an avenue for further progress.

Summarizing, $S[\sigma(t)]$ is increasing if $\rho(t)$ is of rank 1. For $\rho(t)$ of higher rank it is an open problem whether $S[\sigma(t)]$ may exhibit fluctuations. An example to this effect would be very interesting. In any case $S[\sigma(t)]$ tends to its least upper bound as $t\to\infty$.

To complete this section, we summarize what made us introduce the density operator $\sigma(t)$. Let f belong to the set K^2 described in Sec. IV. Define the two-component quantity $\tilde{f}:=\{f,Jf\}$ with inner product

$$\langle\tilde{f},\tilde{g}\rangle := (f,g)+(Jf,Jg). \tag{8.2}$$

It can be shown that the set of all $\tilde{f}$ determined by $f\in K^2$ is complete under the inner product (8.2), hence a Hilbert space. We denote the Hilbert space by $\tilde{K}^2$. Given a complete orthonormal set $\{\phi_n\}$ $(n=1,2,...)$ in L^2, define

$$\psi_n:=(1+J^2)^{-1/2}\phi_n, \quad \tilde{\psi}_n:=\{\psi_n, J\psi_n\}.$$

This determines a complete orthonormal set $\{\tilde{\psi}_n\}$ $(n=1,2,...)$ in $\tilde{K}^2$.

A positive trace class operator $\tilde{\rho}$ on $\tilde{K}^2$ has to act according to

$$\tilde{\rho}\tilde{f}=\sum_n \mu_n\tilde{\psi}_n\langle\tilde{f},\tilde{\psi}_n\rangle$$

with some orthonormal set $\{\tilde{\psi}_n\}$ and positive eigenvalues μ_n satisfying $\Sigma_n\ \mu_n<\infty$. If $\tilde{f}$ is written as a column with elements $\tilde{f}_0=f$ and $\tilde{f}_1=Jf$, then $\tilde{\rho}$ takes the form of a 2×2 matrix with elements

$$\tilde{\rho}_{00}=\rho, \quad \tilde{\rho}_{01}=\rho J, \quad \tilde{\rho}_{10}=J\rho, \quad \tilde{\rho}_{11}=J\rho J.$$

Here ρ acts as

$$\rho f=\sum_n \mu_n\psi_n(f,\psi_n)=\sum_n \mu_n(1+J^2)^{-1/2}\phi_n(f,[1+J^2]^{-1/2}\phi_n).$$

Hence ρ is of the form (1.3) with Tr $R=\Sigma_n\ \mu_n$.

The transformation $U(t)$ defined by

$$U(t)\tilde{f}:=\{\exp(-iH_0t)f, \quad J\ \exp(-iH_0t)f\} \quad (t\geqslant 0)$$

is an isometry on $\tilde{K}^2$. The family $\{U(t)|0\leqslant t<\infty\}$ is a contraction semigroup. Denote its generator by T and let λ be a complex number. The operator T is closed and symmetric. In the momentum representation its resolvent acts according to

$$(T-\lambda)^{-1}\tilde{f}=\{(k^2-\lambda)^{-1}f(k,\omega), \quad J(k^2-\lambda)^{-1}f(k,\omega)\}.$$

By Sec. IV, $f(k,\omega)\in K^2$ if and only if $f(k,\omega)$ is the boundary value of a function $f\in G^2$. Hence the operator $(T-\lambda)^{-1}$ on $\tilde{K}^2$ is bounded if and only if $(k^2e^{2i\phi}-\lambda)^{-1}f(ke^{i\phi},\omega)\in G^2$ whenever $f(ke^{i\phi},\omega)\in G^2$. Thus $(T-\lambda)^{-1}$ is bounded if and only if Im $\lambda>0$. In other words, while the time evolution on L^2 is generated by the operator H_0 with spectrum $[0,\infty)$, on $\tilde{K}^2$ we have a generator whose spectrum includes the entire real axis $(-\infty,\infty)$. This is interesting because for a classical dynamical system to have a Prigogine entropy operator, it is necessary[28] that the spectrum of the generator of the time evolution runs from $-\infty$ to ∞.

Let $U^*(t)$ be the adjoint of $U(t)$ on $\tilde{K}^2$ and let $\tilde{\rho}(t)$ denote $U(t)\tilde{\rho}U^*(t)$. Since $U(t)$ is an isometry, the entropy $S[\tilde{\rho}(t)]$ is constant in time on $t\geqslant 0$. Notice, however, that if $\tilde{\rho}(t)$ is written as a 2×2 matrix, the sum of its diagonal elements is

$$\tilde{\rho}_{00}(t)+\tilde{\rho}_{11}(t)=\rho(t)+J\rho(t)J=\sigma(t).$$

The motivation to focus on $S[\sigma(t)]$ as an entropy came from publications by Mackey[59,72] in which he showed that taking a trace of a classical dynamical system with constant entropy may give rise to an entropy that increases. As examples he referred to the constructions of $\Lambda\rho(t)$ for the baker map[13] and for K-flows.[15] Recall that $\Lambda\rho(t)$ in these cases evolves in time according to a strong[12,13] (or monotonic[14,15]) Markov semigroup. Mackey[59,72] quoted a theorem due to Rohlin,[73] which essentially says that every such Markov semigroup is a trace of a K-system. Besides coarse graining, taking a trace appears to be one of very few procedures that may generate a system with increasing entropy.

APPENDIX: COUNTEREXAMPLES

Let $\sigma\#$ be a positive operator in the trace class on ML^2. Transformation (8.1) takes $\sigma\#$ into $\tau\#$. The integral kernel of $\tau\#$ is square-integrable. Hence $\tau\#$ belongs to the Schmidt class. This appendix gives examples in which the operator $\tau\#$ is not positive. The implication is that the underlying operator σ is not of the form $\rho+J\rho J$ with positive ρ. Hence σ does not belong to the allowed class to which this paper applies.

For the sake of simplicity the examples consider operators σ of rank 1 or 2 with integral kernels that do not depend on angular variables. There are many possibilities for generalizations. With a slight change of notation, it is sufficient to construct operators $\sigma\#$ in the trace class on $L^2(\mathbb{R})$ such that

$$I:=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\overline{f}(u)(1+e^{-u-v})^{-1}\sigma\#(u,v)f(v)\,du\,dv<0 \tag{A1}$$

for some $f(u)\in L^2(\mathbb{R})$.

Example A1: Suppose $\sigma\#(u,v)=g(u)\overline{g}(v)$ and

$$\overline{f}(u)g(u)=\begin{cases}-\alpha e^{-u} & \text{if } u>0\\ e^{u} & \text{if } u<0,\end{cases}$$

with some real constant α to be determined later. In the first quadrant of the uv-plane we introduce new integration variables $x=e^{-u}$, $y=e^{-v}$, in the second quadrant $x=e^{u}$, $y=e^{-v}$, and so on. This gives

$$I=1-\alpha+(\alpha^2-1)\int_0^1\int_0^1(1+xy)^{-1}dx\,dy.$$

Now

$$\int_0^1(1+xy)^{-1}dx=y^{-1}\ln(1+y)$$

and

$$\int_0^1 y^{-1}\ln(1+y)dy=\sum_{n=1}^{\infty}(-1)^{n-1}n^{-2}=\zeta(2)/2=\pi^2/12.$$

Hence consider

$$I_1(\alpha)=1-\alpha+(\alpha^2-1)(\pi^2/12).$$

This function has zeros $\alpha_1=12\pi^{-2}-1$ and $\alpha_2=1$. It is negative when $\alpha_1<\alpha<\alpha_2$. Given $\alpha\in(\alpha_1,\alpha_2)$, there are many ways in which $f(u)g(u)$ can be factorized into $f(u)$ and $g(u)$. Hence there are many operators $\sigma\#$ of rank 1 resulting in operators $\tau\#$ that are not positive. □

Example A2: For an example in which $\sigma\#$ has rank 2, we first examine $\sigma\#(u,v)=h(u)h(v)$ with

$$\overline{f}(u)h(u)=\begin{cases}-\beta e^{-2u} & \text{if } u>0\\ e^{2u} & \text{if } u<0.\end{cases}$$

Evaluating the integral in Eq. (A1) with the help of the integration variables x and y of Example A1 gives

$$I_2(\beta)=\frac{1}{4}-\frac{1}{4}\beta+(\beta^2-1)\int_0^1\int_0^1 xy(1+xy)^{-1}dxdy=\frac{1}{4}-\frac{1}{4}\beta+\beta^2-1-(\beta^2-1)(\pi^2/12).$$

This function has zeros $\beta_1=(\pi^2-9)(12-\pi^2)^{-1}$ and $\beta_2=1$. It is negative when $\beta_1<\beta<\beta_2$. With a view to the following we note that β_1 is close to 0.4 and α_1 close to 0.2. In any case $\alpha_1<\beta_1$.

We now pick positive numbers a and b, then choose

$$g(u)=\begin{cases}ae^{-u/2},\\ ae^{u/2},\end{cases}\quad h(u)=\begin{cases}be^{-3u/2} & \text{if } u>0\\ be^{3u/2} & \text{if } u<0.\end{cases}\tag{A2}$$

This makes

$$\sigma\#(u,v)=g(u)\overline{g}(v)+h(u)\overline{h}(v)$$

the kernel of a positive operator of rank 2. Notice that $\sigma\#(u,v)$ is well behaved in the sense of being continuous and positive as a function of u, v. However, if $\beta_1<\gamma<1$ and

$$f(u)=\begin{cases}-\gamma e^{-u/2} & \text{if } u>0\\ e^{u/2} & \text{if } u<0,\end{cases}\tag{A3}$$

then the integral I of Eq. (A1) is negative. □

Example A3: Returning to $\sigma\#(u,v)=g(u)\overline{g}(v)$ of rank 1, with $g(u)$ as in Eq. (A2), we let $f_1(u)$ be equal to $f(u)$ defined in Eq. (A3). As noted before, this makes I negative. We also choose

$$f_2(u)=\begin{cases}-\gamma e^{-3u/2} & \text{if } u>0\\ e^{3u/2} & \text{if } u<0.\end{cases}$$

Again I is negative. Hence we have identified linearly independent functions f_1 and f_2 for which I is negative. □

[1] E. H. Lieb and M. B. Ruskai, ''A fundamental property of quantum-mechanical entropy,'' Phys. Rev. Lett. **30**, 434–436 (1973).
[2] E. H. Lieb and M. B. Ruskai, ''Proof of the strong subadditivity of quantum-mechanical entropy,'' J. Math. Phys. **14**, 1938–1941 (1973).
[3] A. Wehrl, ''General properties of entropy,'' Rev. Mod. Phys. **50**, 221–260 (1978).
[4] G. Lindblad, ''An entropy inequality for quantum measurements,'' Commun. Math. Phys. **28**, 245–249 (1972).
[5] V. Enss, ''Asymptotic observables on scattering states,'' Commun. Math. Phys. **89**, 245–268 (1983).

[6] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators with Application to Quantum Mechanics and Global Geometry* (Springer, Berlin, 1987), Sec. 5.5.
[7] I. Prigogine, C. George, F. Henin, and L. Rosenfeld, ''A unified formulation of dynamics and thermodynamics,'' Chem. Scr. **4**, 5–32 (1973).
[8] I. Prigogine, *From Being to Becoming* (Freeman, New York, 1980).
[9] B. Misra and I. Prigogine, ''Time, probability, and dynamics,'' in *Long-Time Prediction in Dynamics*, edited by C. W. Horton, Jr., L. E. Reichl, and V. G. Szebehely (Wiley, New York, 1983), pp. 21–43.
[10] Reference 8, Appendix A.
[11] Reference 8, Appendix B.
[12] B. Misra, I. Prigogine, and M. Courbage, ''From deterministic dynamics to probabilistic descriptions,'' Proc. Natl. Acad. Sci. USA **76**, 3607–3611 (1979).
[13] B. Misra, I. Prigogine, and M. Courbage, ''From deterministic dynamics to probabilistic descriptions,'' Physica A **98**, 1–26 (1979).
[14] M. Courbage and B. Misra, ''On the equivalence between Bernoulli dynamical systems and stochastic Markov processes,'' Physica A **104**, 359–377 (1980).
[15] S. Goldstein, B. Misra, and M. Courbage, ''On intrinsic randomness of dynamical systems,'' J. Stat. Phys. **25**, 111–126 (1981).
[16] R. de la Llave, ''Rates of convergence to equilibrium in the Prigogine–Misra–Courbage theory of irreversibility,'' J. Stat. Phys. **29**, 17–31 (1982).
[17] M. Courbage and I. Prigogine, ''Intrinsic randomness and intrinsic irreversibility in classical dynamical systems,'' Proc. Natl. Acad. Sci. USA **80**, 2412–2416 (1983).
[18] M. Courbage, ''Intrinsic irreversibility of Kolmogorov dynamical systems,'' Physica A **122**, 459–482 (1983).
[19] S. Martinez and E. Tirapegui, ''Description of a class of Markov processes 'equivalent' to *K*-shifts,'' J. Stat. Phys. **37**, 173–186 (1984).
[20] S. Martinez and E. Tirapegui, ''A possible physical interpretation of the Λ operator in the Prigogine theory of irreversibility,'' Phys. Lett. A **110**, 81–83 (1985).
[21] G. Braunss, ''Intrinsic stochasticity of dynamical systems,'' Acta Applicandae Math. **3**, 1–21 (1985).
[22] Y. Elskens and I. Prigogine, ''From instability to irreversibility,'' Proc. Natl. Acad. Sci. USA **83**, 5756–5760 (1986).
[23] J. Kumičák and E. Brändas, ''Some aspects of approach to equilibrium in classical and quantum systems,'' Int. J. Quantum Chem. **32**, 669–683 (1987).
[24] Z. Suchanecki, A. Weron, M. Rybaczuk, and K. Weron, ''An explicit approach to the Λ-operator and the *H*-theorem in the Prigogine theory of irreversibility,'' Physica A **155**, 385–400 (1989).
[25] Z. Suchanecki and A. Weron, ''Characterizations of intrinsically random dynamical systems,'' Physica A **166**, 220–228 (1990).
[26] Z. Suchanecki, ''On lambda and internal time operators,'' Physica A **187**, 249–266 (1992).
[27] I. E. Antoniou and K. E. Gustafson, ''From probabilistic descriptions to deterministic dynamics,'' Physica A **197**, 153–166 (1993).
[28] B. Misra, ''Nonequilibrium entropy, Lyapounov variables, and ergodic properties of classical systems,'' Proc. Natl. Acad. Sci. USA **75**, 1627–1631 (1978).
[29] I. Prigogine, F. Mayné, C. George, and M. de Haan, ''Microscopic theory of irreversible processes,'' Proc. Natl. Acad. Sci. USA **74**, 4152–4156 (1977).
[30] M. de Haan, C. George, and F. Mayné, ''Liapounov function for the Friedrichs model,'' Physica A **92**, 584–598 (1978).
[31] T. Petrosky, I. Prigogine, and S. Tasaki, ''Quantum theory of non-integrable systems,'' Physica A **173**, 175–242 (1991).
[32] I. E. Antoniou and I. Prigogine, ''Intrinsic irreversibility and integrability of dynamics,'' Physica A **192**, 443–464 (1993).
[33] K. O. Friedrichs, ''On the perturbation of continuous spectra,'' Commun. Appl. Math. **1**, 361–406 (1948).
[34] I. Prigogine, C. George, and F. Henin, ''Dynamical and statistical descriptions of *N*-body systems,'' Physica (Amsterdam) **45**, 418–434 (1969).
[35] C. George, ''Subdynamics and correlations,'' Physica (Amsterdam) **65**, 277–302 (1973).
[36] A. Grecos, T. Guo, and W. Guo, ''Some formal aspects of subdynamics,'' Physica A **80**, 421–446 (1975).
[37] I. Prigogine and C. George, ''The second law as a selection principle: The microscopic theory of dissipative processes in quantum systems,'' Proc. Natl. Acad. Sci. USA **80**, 4590–4594 (1983).
[38] C. George, F. Mayné, and I. Prigogine, ''Scattering in superspace,'' in *Advances in Chemical Physics*, edited by I. Prigogine and S. A. Rice (Wiley, New York, 1985), Vol. 61, pp. 223–299.
[39] T. Y. Petrosky and H. Hasegawa, ''Subdynamics and nonintegrable systems,'' Physica A **160**, 351–385 (1989).
[40] T. Petrosky and I. Prigogine, ''Alternative formulation of classical and quantum dynamics for non-integrable systems,'' Physica A **175**, 146–209 (1991).
[41] I. Prigogine, ''Dissipative processes in quantum theory,'' Phys. Rep. **219**, 93–108 (1992).
[42] A. Böhm, ''Resonance poles and Gamow vectors in the rigged Hilbert space formulation of quantum mechanics,'' J. Math. Phys. **22**, 2813–2823 (1981).
[43] M. Gadella, ''A rigged Hilbert space of Hardy-class functions: Applications to resonances,'' J. Math. Phys. **24**, 1462–1469 (1983).
[44] A. Bohm and M. Gadella, *Dirac Kets, Gamow Vectors and Gel'fand Triplets* (Springer, Berlin, 1989).
[45] I. Antoniou and S. Tasaki, ''Generalized spectral decompositions of mixing dynamical systems,'' Int. J. Quantum Chem. **46**, 425–474 (1993).
[46] E. C. G. Sudarshan, C. B. Chiu, and V. Gorini, ''Decaying states as complex energy eigenvectors in generalized quantum mechanics,'' Phys. Rev. D **18**, 2914–2929 (1978).
[47] T. Petrosky and I. Prigogine, ''Quantum chaos, complex spectral representations and time-symmetry breaking,'' Chaos Solitons Fractals **4**, 311–359 (1994).
[48] T. Petrosky and I. Prigogine, ''Complex spectral representations and quantum chaos,'' in *Chaotic Dynamics and Trans-*

port in Fluids and Plasmas, edited by I. Prigogine (American Institute of Physics, New York, 1993), pp. 80–108.
[49] I. Prigogine and T. Petrosky, ''Poincaré resonances and the extension of classical and quantum mechanics,'' in *Nonlinear, Deformed and Irreversible Quantum Systems*, edited by H.-D. Doebner, V. K. Dobrev, and P. Natterman (World Scientific, Singapore, 1995), pp. 3–21.
[50] B. Misra, I. Prigogine, and M. Courbage, ''Lyapounov variable: Entropy and measurement in quantum mechanics,'' Proc. Natl. Acad. Sci. USA **76**, 4768–4772 (1979).
[51] C. van Winter, ''Fredholm equations on a Hilbert space of analytic functions,'' Trans. Am. Math. Soc. **162**, 103–139 (1971).
[52] C. van Winter, ''Complex dynamical variables for multiparticle systems with analytic interactions. I,'' J. Math. Anal. Appl. **47**, 633–670 (1974).
[53] M. Rosenblum and J. Rovnyak, ''Restrictions of analytic functions. II,'' Proc. Am. Math. Soc. **51**, 335–343 (1975).
[54] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators* (American Mathematical Society, Providence, 1969), Chap. II, Sec. 2.
[55] K. Fan, ''Maximum properties and inequalities for the eigenvalues of completely continuous operators,'' Proc. Natl. Acad. Sci. USA **37**, 760–766 (1951).
[56] K. Goodrich, K. Gustafson, and B. Misra, ''On converse to Koopman's lemma,'' Physica A **102**, 379–388 (1980).
[57] K. Yosida, *Functional Analysis* (Springer, Berlin, 1966), Chap. XIII, Sec. 3.
[58] J. Voigt, ''Stochastic operators, information, and entropy,'' Commun. Math. Phys. **81**, 31–38 (1981).
[59] M. C. Mackey, ''The dynamic origin of increasing entropy,'' Rev. Mod. Phys. **61**, 981–1061 (1989).
[60] M. C. Mackey, *Time's Arrow: The Origins of Thermodynamic Behavior* (Springer, New York, 1992), Chap. 3.
[61] Reference 60, Corollary 7.8.
[62] J. Orban and A. Bellemans, ''Velocity-inversion and irreversibility in a dilute gas of hard disks,'' Phys. Lett. A **24**, 620–621 (1967).
[63] W. Thirring, *A Course in Mathematical Physics 4; Quantum Mechanics of Large Systems* (Springer, New York, 1983), Proposition 2.1.18.
[64] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-Adjointness* (Academic, New York, 1975), Theorem IX.22.
[65] Reference 63, Definition 2.2.2; Remark 2.2.5.
[66] G. Lindblad, ''Entropy, information and quantum measurements,'' Commun. Math. Phys. **33**, 305–322 (1973), Proposition 4.
[67] Reference 54, Chap. III, Sec. 10.
[68] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. III. Scattering Theory* (Academic, New York, 1979), pp. 65, 66.
[69] Reference 54, Chap. III, Sec. 17.
[70] Y. Sagher, ''An application of interpolation theory to Fourier series,'' Studia Math. **41**, 169–181 (1972).
[71] Y. Sagher, ''Some remarks on interpolation of operators and Fourier coefficients,'' Studia Math. **44**, 239–252 (1972).
[72] Reference 60, Chap. 9.
[73] V. A. Rohlin, ''Exact endomorphisms of a Lebesgue space,'' Izv. Akad. Nauk SSSR Ser. Mat. **25**, 499–530 (1961) [Am. Math. Soc. Transl. Ser. 2 **39**, 1–36 (1964)].

Kodaira-Spencer deformation of complex structures and Lagrangian field theory

Giuseppe Bandelloni[a)]
Istituto di Fisica della Facoltà di Ingegneria, Università degli Studi di Genova, P.le Kennedy, I-16129 Genova, Italy
and Istituto Nazionale di Fisica Nucleare, INFN, Sezione di Genova, via Dodecaneso 33, I-16146 Genova, Italy

Serge Lazzarini[b)]
Institute for Theoretical Physics, University of Amsterdam, Valckenierstraat 65, NL-1018 XE Amsterdam, The Netherlands

(Received 15 October 1997; accepted for publication 20 January 1998)

Similar to the Beltrami parametrization of complex structures on a (compact) Riemann surface, we use in this paper the Kodaira-Spencer deformation theory of complex structures on a (compact) complex manifold of higher dimension. According to the Newlander-Nirenberg theorem, a smooth change of local complex coordinates is implemented with respect to an integrable complex structure parametrize by a Beltrami differential. The question of constructing a local field theory on a complex compact manifold is addressed and the action of smooth diffeomorphisms is studied in the BRS algebraic approach. The BRS cohomology for the diffeomorphisms gives generalized Gel'fand-Fuchs cocycles provided that the Kodaira-Spencer integrability condition is satisfied. The diffeomorphism anomaly is computed and turns out to be holomorphically split as in the bidimensional Lagrangian conformal models. Moreover, its algebraic structure is much more complicated than the one proposed in a recent paper [Losev *et al.* Nuc. Phys. B **484**, 196 (1997)]. © *1998 American Institute of Physics.* [S0022-2488(98)00807-X]

I. INTRODUCTION

A class of bidimensional conformal models, thanks to the equivalence between conformal and complex structures on a 2-D manifold, has been suitably incorporated in the framework of the local field theory upon using the Beltrami parametrization of complex structures. Moreover, in this parametrization the holomorphic factorization of the partition functions of these Lagrangian conformal models (free bosonic string, bc system possibly coupled to Yang-Mills gauge fields) has been proved.[2–4]

In a real even dimensions larger than two, the equivalence between conformal and complex structures is no longer valid. However, complex structures on a complex manifold are still parametrized by a vector-valued (0,1)-form fulfilling both an integrability and algebraic conditions. This is the Kodaira-Spencer deformation theory of complex structures on a manifold thus generalizing the Beltrami parametrization to larger dimensions. While in two-(real) dimensions, the Beltrami differential has been used in the study of some bidimensional Lagrangian conformal models on a (compact) Riemann surface, the Kodaira-Spencer deformation of complex structures has occurred in theoretical physics in the context of mirror symmetry in string field models with values in some target Kähler manifold.[5–8]

Quite recently, the holomorphicity of the partition function for chiral lagrangian of higher dimensional bc systems on a Kähler manifold has been considered in Ref. 1, but only with respect to the so-called ''chiral diffeomorphisms.'' It is a restriction which does not represent the whole problem of symmetry under arbitrary reparametrization. Moreover, this *a priori* holomorphicity in

a)Electronic-mail: beppe@genova.infn.it
b)Present address: Centre de Physique Théorique, CNRS-Luminy, Case 907, F-13288 Marseille Cedex 9. Electronic-mail: sel@cpt.univ-mrs.fr

 © 1998 American Institute of Physics

the Beltrami differential defining an (integrable) complex structure is put by hand, and it is not straightforward to decide whether a higher dimensional analog of the *bc* system actually enjoys the holomorphic factorization property. Moreover, the proposed ''chiral anomaly'' in Ref. 1 turns out to be much more complicated as it will be shown in this work.

In the present work, we simply address the question whether this parametrization might give rise to interesting features for possible local field theories on a compact complex manifold, for instance, whether the Beltrami differential may play a role at the field level. In particular, we have in mind the four-(real) dimensional interesting case and possible links with the recent work.[9] For other noteworthy implications of the holomorphicity in the Beltrami differential the reader is referred to Ref. 1.

Following closely the bidimensional approach where smooth diffeomorphisms connected to the identity map make up the ''gauge group,'' the action of smooth diffeomorphisms (and not chiral diffeomorphisms as considered in Ref. 1) is studied in the BRS formulation. Furthermore, the BRS cohomology for diffeomorphisms is analyzed along the lines developed in Refs. 10–12, and allows one to compute the local expression of a diffeomorphism anomaly. A few possible Lagrangian models will be investigated. The first part of the present paper is devoted to the notation and, according to the book[13] by Kodaira, we will introduce the parametrization of complex structures on a compact manifold by vector-valued (0,1)-forms, the higher dimensional analogs of the Beltrami differentials. We shall concentrate on the generalized matter fields considered as usual tensors.

The main part of the paper is based on the study of the diffeomorphism cohomology and the finding of anomalies under the locality principle. The cohomology can be carried out for the BRS operator of the diffeomorphisms which is nilpotent under the restriction that the complex structure is integrable.

II. BASIC IDEA FOR THE PERTURBATION OF COMPLEX STRUCTURES

Consider a complex compact manifold of complex dimension n according to a finite complex analytic covering $\{(U_\alpha, z_\alpha)\}$ with $M=\cup_\alpha U_\alpha$ and $z_\alpha: U_\alpha \to \mathbf{C}^n$, the local complex coordinate mapping, $z\equiv(z^1,\ldots,z^n)=(z^k)$, with biholomorphic transition functions

$$\text{in } U_\alpha \cap U_\beta \neq \varnothing: \quad z_\alpha = f_{\alpha\beta}(z_\beta), \tag{II.1}$$

for any point in the intersection. When the underlying smooth structure is considered (thus M is viewed as a real smooth manifold of dimension $2n$), the complex conjugate coordinates $\overline{z}\equiv(\overline{z^1},\ldots,\overline{z^n})=(\overline{z^k})$ is regarded as independent with respect to z.

Partial derivatives will be shortly denoted as $\partial/\partial z^k \equiv \partial_k$ and $\partial/\partial \overline{z^j} \equiv \partial_{\overline{J}}$. We shall denote by $T^{1,0}$ the holomorphic tangent bundle of M, and ∂ will stand for the natural frame on $T^{1,0}$ associated with the local background complex coordinates $(z,\overline{z})$. Similarly, dz will be the natural dual basis in the holomorphic cotangent space $(T^{1,0})^*$,

$$\partial \equiv \begin{pmatrix} \partial_1 \\ \vdots \\ \partial_n \end{pmatrix}, \quad \overline{\partial} \equiv \begin{pmatrix} \partial_{\overline{1}} \\ \vdots \\ \partial_{\overline{n}} \end{pmatrix}, \quad dz \equiv (dz^1,\ldots,dz^n), \quad d\overline{z} \equiv (d\overline{z^1},\ldots,d\overline{z^n}), \tag{II.2}$$

while we have for the external derivative (denoted here by a boldface symbol)

$$d = \boldsymbol{\partial} + \overline{\boldsymbol{\partial}} = dz\cdot\partial + d\overline{z}\cdot\overline{\partial} = dz^k\partial_k + d\overline{z^j}\partial_{\overline{J}}, \quad d^2 = \boldsymbol{\partial}^2 = \overline{\boldsymbol{\partial}}^2 = \{\boldsymbol{\partial}, \overline{\boldsymbol{\partial}}\} = 0. \tag{II.3}$$

This given system of local complex coordinates (which is fixed once and for all) will serve as a background for the perturbation implemented by modifying transition functions, according to the Kodaira-Spencer's approach, as follows. On any overlapping $U_\alpha \cap U_\beta \neq \varnothing$, of the finite covering, the background transition function $f_{\alpha\beta}(z_\beta)$ is replaced by another function $f_{\alpha\beta}(z_\beta,t)$, such that $f_{\alpha\beta}(z_\beta,0)=f_{\alpha\beta}(z_\beta)$, and where t is a complex parameter belonging to some domain of $\mathbf{C}^m$ containing the origin. One can thus get a complex analytic family parametrized by t of complex compact manifolds. It turns out that the infinitesimal deformation of M can be represented by an

element of the cohomology group $H^1(M,\Theta)$ of M with coefficients in the sheaf Θ of germs of holomorphic vector fields on M. In most cases, the Kodaira-Spencer's conjecture holds, dim $H^1(M,\Theta)=m$ would give the dimension of the moduli space of M when the number of moduli can be defined. A question of primary importance is to know whether it is always possible to deform a complex manifold. A partial answer is given, for instance, in Ref. 13: It is always possible to deform the complex structure on any complex compact manifold. Furthermore, all the complex compact manifolds belonging to the same complex analytic family are all diffeomorphic since they are subordinated to the same smooth structure underlying the background complex structure. This situation allows one to tackle the deformation problem by parametrizing the complex structures by a vector-valued (0,1)-form on M which extends to the higher dimensional case the well-known Beltrami parametrization of complex structures on a Riemann surface.

A. The Newlander-Nirenberg theorem and the Kodaira-Spencer integrability condition

Let $\boldsymbol{\mu}$ be a vector-valued (0,1)-form, a smooth section of $T^{1,0}\otimes(T^{0,1})^*$ and then locally represented as

$$\boldsymbol{\mu}(z,\bar{z})=d\bar{z}\cdot\mu(z,\bar{z})\cdot\partial=\mu^k_J(z,\bar{z})d\bar{z}^J\otimes\partial_k,\quad \mu=(\mu^k_J). \tag{II.4}$$

This object generalizes the usual Beltrami differential but will be subject to an integrability condition as stated below.

Theorem: (Newlander-Nirenberg[13]) For a given smooth vector-valued (0,1)-form, $\boldsymbol{\mu}$, viewed as a differential operator of first order and locally defined on M by Eq. (II.4), consider the following partial differential operator of first order:

$$\boldsymbol{\mathcal{L}}=\bar{\boldsymbol{\partial}}-\boldsymbol{\mu}=d\bar{z}\cdot(\bar{\partial}-\mu\cdot\partial)=d\bar{z}^i\mathcal{L}_{\bar{\imath}}=d\bar{z}^i(\partial_{\bar{\imath}}-\mu^k_{\bar{\imath}}\partial_k). \tag{II.5}$$

Next suppose that $\boldsymbol{\mu}$ obeys the following two conditions:

$$\det(I-\mu\bar{\mu})\neq 0,\quad \text{and}\quad \boldsymbol{\mathcal{L}}\boldsymbol{\mu}=\bar{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=\bar{\boldsymbol{\partial}}\boldsymbol{\mu}-\tfrac{1}{2}[\boldsymbol{\mu},\boldsymbol{\mu}]=0, \tag{II.6}$$

where the graded bracket reflects the bracket of vector fields.

Hence, the differential operators $\mathcal{L}_{\bar{1}},\ldots,\mathcal{L}_{\bar{n}},\bar{\mathcal{L}}_1,\ldots,\bar{\mathcal{L}}_n$ are linearly independent and they fulfill the condition

$$\boldsymbol{\mathcal{L}}^2=0\Leftrightarrow[\mathcal{L}_{\bar{\imath}},\mathcal{L}_{\bar{J}}]=0,\quad \bar{\imath},\bar{J}=1,\ldots,n. \tag{II.7}$$

Moreover, the system of n PDE's

$$\boldsymbol{\mathcal{L}}f=0\Leftrightarrow\mathcal{L}_{\bar{\imath}}f=0,\quad \bar{\imath}=1,\ldots,n, \tag{II.8}$$

has locally n linearly independent smooth solutions $Z=(Z^1(z,\bar{z}),\ldots,Z^n(z,\bar{z}))$, i.e., the Jacobian never vanishes,

$$\mathcal{J}\equiv\frac{\partial(Z,\bar{Z})}{\partial(z,\bar{z})}=|\det\lambda|^2\det(I-\mu\bar{\mu})\neq 0\Rightarrow\det\lambda\neq 0, \tag{II.9}$$

where $\lambda=(\partial_i Z^k)$ is the (half)-Jacobian matrix.

B. The smooth change of local complex coordinates

Furthermore, the solution Z provides a new system of complex coordinates on M, that is a new integrable complex structure parametrized by μ. The new complex structure $(Z,\bar{Z})$ on M can be seen as a deformation of the background complex structure $(z,\bar{z})$, and each such μ determines an integrable complex structure on M.

Remark: The coordinates Z local solutions of system (II.8) are defined up to an holomorphic function $F(Z)$. Also, when $\mu=(\mu_1^k)\equiv 0$ the change of local coordinates becomes holomorphic so that $\mu\equiv 0$ corresponds to the complex structure given by z. Moreover, one has $\det \lambda \neq 0$.

The Newlander-Nirenberg theorem infers a *smooth* change of local complex coordinates in $\mathbf{C}^n$, $(z,\bar{z})\to(Z,\bar{Z})$. These complex coordinates Z can be reobtained by the following row-vector of 1-forms:

$$dZ=\boldsymbol{\partial} Z+\boldsymbol{\mu} Z=(dz+d\bar{z}\cdot\mu)\cdot\lambda. \tag{II.10}$$

One accordingly gets the following integrability conditions, on the one hand $\lambda=(\partial_i Z^k)$, $\det \lambda \neq 0$,

$$\bar{\boldsymbol{\partial}}\boldsymbol{\partial} Z+\boldsymbol{\partial}(\boldsymbol{\mu} Z)=0\Leftrightarrow\bar{\partial}\lambda-\partial(\mu\cdot\lambda)=0, \tag{II.11}$$

and the other hand, the necessary integrability condition (II.6) on $\boldsymbol{\mu}$. Therefore, the defining system of n PDE's (II.8) for Z, namely,

$$\boldsymbol{\mathcal{L}} Z=(\bar{\boldsymbol{\partial}}-\boldsymbol{\mu})Z=0, \quad \boxed{\bar{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0,}$$

is the analog of the Beltrami equation but with in addition an integrability condition. The latter looks like the vanishing of a curvature. One may also assume that $\det(I-\mu\bar{\mu})>0$ and $\det \lambda >0$, when orientation-preserving coordinate transformations are in order.

According to definition (II.10) and its complex conjugate, one can write in matrix notation

$$(dZ,d\bar{Z})=(dz,d\bar{z})\begin{pmatrix} I & \bar{\mu} \\ \mu & I \end{pmatrix}\begin{pmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{pmatrix}, \tag{II.13}$$

so that by duality

$$\begin{pmatrix} \partial_Z \\ \partial_{\bar{Z}} \end{pmatrix}=\begin{pmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{pmatrix}^{-1}\begin{pmatrix} I & \bar{\mu} \\ \mu & I \end{pmatrix}^{-1}\begin{pmatrix} \partial_z \\ \partial_{\bar{z}} \end{pmatrix}. \tag{II.14}$$

One explicitly finds

$$\begin{pmatrix} \partial_Z \\ \partial_{\bar{Z}} \end{pmatrix}=\begin{pmatrix} \lambda^{-1}\cdot(I-\bar{\mu}\cdot\mu)^{-1}\cdot(\partial_z-\bar{\mu}\cdot\partial_{\bar{z}}) \\ \bar{\lambda}^{-1}\cdot(I-\mu\cdot\bar{\mu})^{-1}\cdot(\partial_{\bar{z}}-\mu\cdot\partial_z) \end{pmatrix}\equiv\begin{pmatrix} \lambda^{-1}\cdot\nabla \\ \bar{\lambda}^{-1}\cdot\bar{\nabla} \end{pmatrix}. \tag{II.15}$$

III. THE DIFFEOMORPHISM ACTION

By computing the pull-back of the 1-form (II.10) under any smooth diffeomorphism φ, $\varphi^* dZ=d\varphi^* Z=(d\varphi^* w+d\varphi^*\bar{w})\cdot(\varphi^*\mu))\cdot(\varphi^*\lambda)=(dz+d\bar{z}\cdot(\mu^\varphi))\cdot(\lambda^\varphi)$, where $(z,\bar{z})$ are coordinates in the source chart of φ and $(w,\bar{w})=\varphi(z,\bar{z})$ in the target chart, the finite diffeomorphism transformation of the matrices of basic geometric objects reads

$$\begin{aligned} \lambda^\varphi&=(\partial\varphi+\partial\bar{\varphi}\cdot(\mu\circ\varphi))\cdot(\lambda\circ\varphi), \\ \mu^\varphi&=(\bar{\partial}\varphi+\overline{\partial\varphi}\cdot(\mu\circ\varphi))\cdot(\partial\varphi+\partial\bar{\varphi}\cdot(\mu\circ\varphi))^{-1}. \end{aligned} \tag{III.1}$$

Thus, as a vector-valued $(0,1)$-form,

$$\boldsymbol{\mu}^\varphi=d\bar{z}\cdot\mu^\varphi\cdot\partial=(\bar{\boldsymbol{\partial}}\varphi+\overline{\boldsymbol{\partial}\varphi}\cdot(\mu\circ\varphi))\cdot(\partial\varphi+\partial\bar{\varphi}\cdot(\mu\circ\varphi))^{-1}\cdot\partial. \tag{III.2}$$

We arrive at the equivalence problem between integrable complex structures. Consider two copies of the compact complex manifold M with local background complex coordinates $(z,\bar{z})$ and let φ be any smooth diffeomorphism homotopic to the identity map. Given a vector-valued

(0,1)-form, $\boldsymbol{\mu}$, submitted to the condition $\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0$, let $(Z,\overline{Z})$ be the local complex coordinates at the point $(w,\overline{w})=\varphi(z,\overline{z})$ solving the system of n PDE's (II.12), then one can merely show that the integrability condition is preserved by diffeomorphisms (as will be explicitly proved in Appendix A and also below for the infinitesimal version)

$$\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0\Rightarrow\overline{\boldsymbol{\partial}}\boldsymbol{\mu}^{\varphi}-(\boldsymbol{\mu}^{\varphi})^2=0, \tag{III.3}$$

and thanks to the Newlander-Nirenberg theorem there are local complex coordinates Z^{φ} pertaining to the integrable complex structure parametrized by $\boldsymbol{\mu}^{\varphi}$; hence one has

$$(\overline{\boldsymbol{\partial}}-\boldsymbol{\mu}^{\varphi})Z^{\varphi}=0. \tag{III.4}$$

The problem is to see whether the mapping $(Z^{\varphi},\overline{Z^{\varphi}})\rightarrow(Z,\overline{Z})$ is indeed biholomorphic. If it is so, then $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{\varphi}$ define equivalent integrable complex structures which thus correspond to the same point t in the moduli space. The situation is the same as for Riemann surfaces, that is the mapping $(Z^{\varphi},\overline{Z^{\varphi}})\rightarrow(Z,\overline{Z})$ is actually biholomorphic. $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{\varphi}$ defining equivalent integrable complex structures. The reader is referred to Appendix B for the details.

The infinitesimal action given by the (graded) Lie derivative, $s\equiv L_{\mathbf{c}}=i_{\mathbf{c}}d-di_{\mathbf{c}}$ and, thanks to (II.10), writes locally in each open set U,

$$sZ=L_{\mathbf{c}}Z=i_{\mathbf{c}}dZ=(c+\overline{c}\cdot\mu)\cdot\lambda\equiv C\cdot\lambda\equiv\Upsilon, \tag{III.5}$$

where $\boldsymbol{c}=c\cdot\partial+\overline{c}\cdot\overline{\partial}=\Upsilon\cdot\partial_Z+\overline{\Upsilon}\cdot\partial_{\overline{Z}}$ is the (smooth) Faddeev-Popov ghost vector field expressed, respectively, in the holomorphic coordinates z and Z, and C reflects a change of basis in the Lie algebra of diffeomorphisms leading to the holomorphic factorization property of bidimensional conformal models.[14,15,3,4] Here, in matrix notation, we have

$$(C,\overline{C})=(c,\overline{c})\begin{pmatrix} I & \overline{\mu} \\ \mu & I \end{pmatrix},\quad (\Upsilon,\overline{\Upsilon})=(C,\overline{C})\begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix}, \tag{III.6}$$

the latter being the BRS counterpart of Eq. (II.13).

On the one hand, the infinitesimal version of (III.1) can be expressed in terms of the nilpotent BRS operation, $s^2=0$, as

$$\begin{aligned} s\lambda&=\boldsymbol{c}\lambda+(\partial c+\partial\overline{c}\cdot\mu)\cdot\lambda,\\ s\mu&=\boldsymbol{c}\mu+(\overline{\partial}-\mu\cdot\partial)c+(\overline{\partial}-\mu\cdot\partial)\overline{c}\cdot\mu,\\ s\boldsymbol{c}&=\tfrac{1}{2}[\boldsymbol{c},\boldsymbol{c}],\quad \boldsymbol{c}=c\cdot\partial+\overline{c}\cdot\overline{\partial}, \end{aligned} \tag{III.7}$$

where the graded bracket is that of vector fields. Note that the variation of μ is nonlinear in μ. Moreover, we know that the integrability condition (II.6) is s-invariant thanks to Appendix A.

On the other hand, with respect to the integrable complex structure defined by μ, the infinitesimal action of diffeomorphisms on the complex coordinate Z writes locally in U, with $s^2=0$,

$$\begin{aligned} &sZ\equiv\boldsymbol{c}Z=\Upsilon,\quad s\Upsilon=0,\\ &s\boldsymbol{d}+\boldsymbol{d}s=0\Rightarrow[s,\partial_Z]=-\partial_Z\boldsymbol{c}. \end{aligned} \tag{III.8}$$

Accordingly, one has the following remarkable identities, together with their complex conjugate:

$$[s-\boldsymbol{c},\partial_Z]=0,\quad s\boldsymbol{\partial}+\boldsymbol{\partial}s=\boldsymbol{\partial}\overline{\Upsilon}\cdot\partial_{\overline{Z}}-\overline{\boldsymbol{\partial}}\,\Upsilon\cdot\partial_Z. \tag{III.9}$$

By computing directly $s\ dZ=-ds\ Z$, one finds,

$$s\lambda=\partial(C\cdot\lambda),$$

$$s\boldsymbol{\mu}=-\overline{\boldsymbol{\partial}}(C\cdot\partial)+[\boldsymbol{\mu},C\cdot\partial],\quad \boldsymbol{\mu}=d\overline{z}\cdot\mu\cdot\partial, \tag{III.10}$$

$$s(C\cdot\partial)=\tfrac{1}{2}[C\cdot\partial,C\cdot\partial]=(C\cdot\partial C)\cdot\partial,$$

where the component C of the vector field $C\cdot\partial$, a smooth section of the holomorphic tangent bundle $T^{1,0}$,[13] is defined by (III.6).

The reader's attention is drawn to the fact that, these two systems of infinitesimal variations for $\boldsymbol{\mu}$ and λ, (III.7) and (III.10), are equivalent only upon the use of both the integrability conditions (II.6) and (II.11). The former represents the true infinitesimal action of diffeomorphisms, while the latter, which holds only in the case of an integrable complex structure, will govern the holomorphic factorization property. In the integrable case, the system (III.10) will also represent the infinitesimal action of diffeomorphisms, and the integrability condition (II.6) is seen to be stable under diffeomorphisms by

$$\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0\Rightarrow s(\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2)=[C\cdot\partial,\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2]=0. \tag{III.11}$$

IV. FIELDS AND LAGRANGIANS

According to what has been done in the two dimensional case,[16,17,3,4,11] one can take advantage of the complex coordinate system Z pertaining to an integrable complex structure μ, in order to exhibit some simple Lagrangians built on a n-dimensional complex manifold.

Before going further, one presumes that a classical action Γ_0^{Cl} has to be well defined on the complex manifold M, i.e.,

$$\Gamma_0^{\mathrm{Cl}}=\int_M dZ^{A_n}\wedge d\overline{Z}^{\overline{A}_n}L_{A_n\overline{A}_n}(Z,\overline{Z}), \tag{IV.1}$$

where in the integrand a multi-index notation has been used; see Eq. (V.26) below, with $|A_n|=|\overline{A}_n|=n$. This means that the Lagrangian density is a skew-symmetric covariant tensorial density under holomorphic change of local coordinates $Z\to Z'(Z)$,

$$L_{A'_n\overline{A}'_n}(Z',\overline{Z}')=\left|\frac{\partial(Z'^1,\ldots,Z'^n)}{\partial(Z^1,\ldots,Z^n)}\right|^{-2}L_{A_n\overline{A}_n}(Z,\overline{Z}),\quad |A_n|=|A'_n|=|\overline{A}_n|=|\overline{A}'_n|=n. \tag{IV.2}$$

For instance, one may ask whether analog Lagrangians of the bc-systems as well as a generalization of the bosonic string can be obtained over such a higher dimensional complex manifold, and what is the role of the complex structure.

Our purpose is first to use the holomorphic coordinates Z and then to make explicit the coupling of matter fields with the integrable complex structure μ.

The simplest higher-dimensional analog of the bidimensional bc-system[3,4] is constructed by considering a holomorphic vector bundle E over M and the fields Ψ as a smooth section of $T^{k,0*}\otimes E$, $0\leqslant k\leqslant n-1$, and Ψ^* a smooth section of $T^{n-k,n-1*}\otimes E^*$. Our choice for the bc fields is rather different from the one written in Ref. 1 and reduces to the usual bc-system for $n=1$. One has following well-defined action:

$$S_{bc}=\int_M \Psi^*\wedge(\overline{\boldsymbol{\partial}}+\overline{\boldsymbol{A}})\Psi, \tag{IV.3}$$

where the (0,1)-connection form $\overline{\boldsymbol{A}}$ parametrizes the holomorphic structure on E provided the integrability condition, $\overline{\boldsymbol{\partial}\boldsymbol{A}}+\overline{\boldsymbol{A}}^2=0$, is satisfied.[18]

The natural generalization of the free bosonic string action reads

$$S(\Phi)=\int_M \overline{\boldsymbol{\partial}}\Phi\wedge\overline{(\overline{\boldsymbol{\partial}}\Phi)}, \tag{IV.4}$$

where in a multi-index notation $\Phi=\Phi_I\,dZ^I$ is a smooth complex-valued $(n-1,0)$-form which can be chosen to be real only in the case $n=1$.

One can see that it is not possible to consider scalar fields in such a complex model unless there exists a Hermitian metric g on M.

In any case, the important point is that the above Lagrangian densities depend locally on the integrable complex structure μ. One can directly check that the two above actions are actually local in μ. Indeed, defining the following change of field variables $(\lambda^I_J)\Phi_I=\phi_J$, with $(\lambda^I_J)=\lambda^{i_1}_{j_1}\cdots\lambda^{i_{n-1}}_{j_{n-1}}$, according to the form degree, then action (IV.4) reads

$$S(\phi,\mu,\overline{\mu})=\int_M \det(I-\mu\cdot\overline{\mu})(\overline{\nabla}\phi+\imath(\boldsymbol{\tau})\phi)\wedge\overline{(\overline{\nabla}\phi+\imath(\boldsymbol{\tau})\phi)}, \tag{IV.5}$$

where $\nabla=dz\cdot\nabla$ [∇ defined in (II.15)] and $\imath(\boldsymbol{\tau})\phi=(n-1)\tau^r\wedge\phi_r$ is a $(n-1,1)$-form given by the inner product on the $(n-1,0)$-form $\phi=\phi_I\,dz^I$ with respect to the vector-valued $(1,1)$-form $\boldsymbol{\tau}=d\overline{z}\cdot(I-\mu\overline{\mu})^{-1}\wedge\boldsymbol{\partial}\mu\cdot\partial$. In the course of the calculation, condition (II.11) on the integrating factor has been used.

Similarly, for the bc-system the action takes the following local expression in μ:

$$S_{bc}(\psi,\psi^*,\mu,\overline{\mu})=\int_M \det(I-\mu\cdot\overline{\mu})\,\psi^*\wedge(\overline{\nabla}+\imath(\boldsymbol{\tau})+d\overline{z}\cdot(I-\mu\overline{\mu})^{-1}\cdot\overline{a})\psi, \tag{IV.6}$$

with $\imath(\boldsymbol{\tau})\psi=k\tau^r\wedge\psi_r$ is a $(k,1)$-form with values (as well ψ) in a copy of the holomorphic vector bundle E with respect to the complex analytic coordinates z, and where we have set for the components of the $(0,1)$-type connection $A_{\overline{Z}}=\lambda^{-1}\cdot(I-\mu\overline{\mu})^{-1}\cdot\overline{a}$. In this framework, (integrable) complex structures on E are parametrized by μ and $\overline{a}$ upon the integrability conditions $\overline{\boldsymbol{\mathcal{L}}}\boldsymbol{\mu}=0$ and $\overline{\boldsymbol{\mathcal{L}}}\overline{\boldsymbol{a}}+\overline{\boldsymbol{a}^2}=0$, respectively. However, the obvious holomorphic factorization property in μ for $n=1$, $(\psi^*\wedge(\overline{\boldsymbol{\partial}}-\boldsymbol{\mu}+\overline{\boldsymbol{a}})\psi)$ (Ref. 4) is spoiled, and the action S_{bc} strongly differs from the one proposed in) Ref. 1 which is assumed to enjoy that property in any dimension.

The action of diffeomorphisms gives rise, for instance, to the following local (in μ) transformation laws on the $(n-1,0)$-form ϕ and on the bc fields ψ and ψ^*:

$$s\phi=\boldsymbol{c}\phi+\imath(\tilde{\boldsymbol{\Omega}})\phi,\quad s\psi=\boldsymbol{c}\psi+\imath(\tilde{\boldsymbol{\Omega}})\psi,\quad s\psi^*=\boldsymbol{c}\psi^*+\imath(\tilde{\boldsymbol{\Omega}})\psi^*+\imath(\overline{\tilde{\boldsymbol{\Omega}}})\psi^*, \tag{IV.7}$$

where we have introduced the following vector-valued $(1,0)$-form carrying ghost grading one $\tilde{\boldsymbol{\Omega}}=dz\cdot(\partial c+\partial\overline{c}\cdot\mu)\cdot\partial$ which governs the parallel transport along the ghost vector field $\boldsymbol{c}$.

V. STUDY OF POSSIBLE DIFFEOMORPHISM ANOMALIES

From now on, since matter fields have not been specified, they will be collectively denoted by ϕ, and together with the integrable complex structure μ, one introduces the short-hand notation $f_0\equiv\{\phi,\psi,\psi^*,\mu^k_{\overline{i}},\overline{\mu^j_l}\}$, and for the diffeomorphism ghost fields as well $f_1\equiv\{c^i,\overline{c^l}\}$, and $f\equiv f_0\cup f_1$. In order to suppress the index summation and the integration over the manifold M as well, we also introduce the dual contraction by the bracket $\langle,\rangle$, defined as

$$\langle\tilde{f},f\rangle\equiv\int_M dz^n\wedge d\overline{z}^n\langle\tilde{f}(z,\overline{z}),f(z,\overline{z})\rangle. \tag{V.1}$$

The diffeomorphism symmetry (III.7) can be encoded in the following nilpotent BRS functional operator:

$$\delta_0=\sum_f\left\langle sf,\frac{\delta}{\delta f}\right\rangle,\quad sf(z,\overline{z})=\boldsymbol{c}f(z,\overline{z})+V(f,c,\overline{c})(z,\overline{z}), \tag{V.2}$$

where V is a local differential polynomial. The diffeomorphism invariance of the classical theory is stated in a Ward identity, $\delta_0\Gamma_0^{\mathrm{Cl}}[f_0]=0$, together with the integrability condition on the μ's, which may be seen as a local constraint,

$$[\mathcal{L}_{\bar{\imath}},\mathcal{L}_{\bar{J}}]=(\mathcal{L}_{\bar{\imath}}\mu_{\bar{J}}^{k}-\mathcal{L}_{\bar{J}}\mu_{\bar{\imath}}^{k})\partial_k\equiv F_{\bar{\imath}\bar{J}}^{k}(\mu)\partial_k,\quad \text{with}\quad F_{\bar{\imath}\bar{J}}^{k}(\mu)=0, \tag{V.3}$$

with of course the complex conjugate expression. According to the diffeomorphism invariance of the integrability condition, one has

$$\delta_0 F_{\bar{\imath}\bar{J}}^{k}(\mu)=0=\delta_0\bar{F}_{ij}^{\bar{l}}(\bar{\mu}). \tag{V.4}$$

Following the spirit of Ref. 15, in the vacuum sector ($\phi\equiv 0$), we can rephrase the Ward identity in terms of local Ward operators with respect to the ghost (parameter) fields, c's, as

$$\begin{aligned}W_k(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]\equiv&\left(\partial_k\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{r}}-\partial_{\bar{\imath}}\frac{\delta}{\delta\mu_{\bar{\imath}}^{k}}+\partial_r\left(\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{k}}\right)\right.\\&\left.+\bar{\mu}_k^{\bar{s}}\left(\partial_{\bar{s}}\bar{\mu}_{\bar{l}}^{\bar{J}}\frac{\delta}{\delta\bar{\mu}_{\bar{l}}^{\bar{J}}}+\partial_{\bar{J}}\left(\bar{\mu}_{\bar{l}}^{\bar{J}}\frac{\delta}{\delta\bar{\mu}_{\bar{l}}^{\bar{s}}}\right)\right)+\bar{F}_{kl}^{\bar{J}}(\bar{\mu})\frac{\delta}{\delta\bar{\mu}_{\bar{l}}^{\bar{J}}}\right)(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]=0,\end{aligned} \tag{V.5}$$

together with the complex conjugate expression $\bar{W}_{\bar{J}}\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]=0$. One can directly check that the combination

$$\begin{aligned}(W_k-\bar{\mu}_k^{\bar{l}}\bar{W}_{\bar{l}})(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]=&\left(\bar{F}_{kl}^{\bar{J}}(\bar{\mu})\frac{\delta}{\delta\bar{\mu}_{\bar{l}}^{\bar{J}}}-\bar{\mu}_k^{\bar{l}}F_{\bar{l}\bar{\imath}}^{r}(\mu)\frac{\delta}{\delta\mu_{\bar{\imath}}^{r}}+(I-\bar{\mu}\cdot\mu)_k^{t}\right.\\&\left.\times\left(\partial_r\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{t}}-\partial_{\bar{\imath}}\frac{\delta}{\delta\mu_{\bar{\imath}}^{t}}+\mu_{\bar{\imath}}^{r}\partial_r\frac{\delta}{\delta\mu_{\bar{\imath}}^{t}}+\partial_t\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{r}}\right)\right)(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]\\=&0,\end{aligned} \tag{V6}$$

no longer depends on the $\bar{\mu}$'s if the integrability conditions $F(\mu(z,\bar{z}))=\bar{F}(\bar{\mu}(z,\bar{z}))=0$ are taken into account. With these constraints one obtains a new local Ward operator

$$\begin{aligned}\mathcal{W}_k(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]\equiv&([(I-\bar{\mu}\cdot\mu)^{-1}]_k^{t}(W_t-\bar{\mu}_t^{\bar{l}}\bar{W}_{\bar{l}}))(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}]\\=&\left(\partial_r\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{k}}-\partial_{\bar{\imath}}\frac{\delta}{\delta\mu_{\bar{\imath}}^{k}}+\mu_{\bar{\imath}}^{r}\partial_r\frac{\delta}{\delta\mu_{\bar{\imath}}^{k}}+\partial_k\mu_{\bar{\imath}}^{r}\frac{\delta}{\delta\mu_{\bar{\imath}}^{r}}\right)(z,\bar{z})\Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}],\end{aligned} \tag{V.7}$$

which reflects the linear change of parameter $C^k=c^k+\mu_{\bar{\imath}}^{k}\bar{c}^{\bar{\imath}}$ corresponding to the variation (III.10) of μ. Recall that the latter is equivalent to the Lie derivative of μ only if the integrability condition is considered.

On the physical side, the meaning of the geometrical quantity

$$\Theta_k^{\bar{l}}(z,\bar{z})\equiv\frac{\delta\Gamma_0^{\mathrm{Cl}}}{\delta\mu_{\bar{\imath}}^{k}(z,\bar{z})}$$

is still lacking. It does not correspond to the energy-momentum tensor as in the bidimensional case since the equivalence between conformal and integrable complex structures no longer holds in higher dimensions. Moreover, the μ's which serve as classical source for the components of Θ are not all independent of each other due to the integrability condition. The question whether the latter induces some constraints on the theory is not yet under control. Nevertheless, one may write down the classical Ward identity (V.7) in the vacuum sector with respect to the ghost C, as

$$\left(\mathcal{L}_i^{\bar{}}(z,\bar{z}) \frac{\delta}{\delta \mu_i^k(z,\bar{z})} - \partial_r \mu_i^r(z,\bar{z}) \frac{\delta}{\delta \mu_i^k(z,\bar{z})} - \partial_k \mu_i^r(z,\bar{z}) \frac{\delta}{\delta \mu_i^r(z,\bar{z})} \right) \Gamma_0^{\mathrm{Cl}}[\mu,\bar{\mu}] = 0. \quad \text{(V.8)}$$

Geometrically, when passing to the $(Z,\bar{Z})$ complex coordinates, Θ behaves like a $(T^{1,0})^* \otimes (T^{0,1})$-valued density according to the following transformation law:

$$\vartheta(Z,\bar{Z}) \equiv \frac{1}{\mathcal{J}} \lambda^{-1} \cdot \Theta(z,\bar{z}) \cdot (I - \mu \cdot \bar{\mu}) \cdot \bar{\lambda}, \quad \text{(V.9)}$$

where $\mathcal{J}$ is the Jacobian defined in (II.9). In terms of the $(Z,\bar{Z})$ complex coordinates, the classical Ward identities (V.8) translate to the vanishing of the following divergences (see Appendix C for some computational details):

$$\partial_{Z^r} \bar{\vartheta}_k^r = 0. \quad \text{(V.10)}$$

Note that for $n=1$, (V.9) reduces to the transformation law of a quadratic differential.[17]

We now introduce external fields $\beta \equiv \beta_{-1} \cup \beta_{-2} \equiv \{\gamma, \eta_k^{\bar{}}, \bar{\eta}_J^{\bar{}}\} \cup \{\zeta_i, \bar{\zeta}_i^{\bar{}}\}$, with an assigned (negative) ghost grading, respectively, coupled to the BRS variation of the fields $f \equiv \{\phi, \mu_i^k, \bar{\mu}_l^{\bar{}}, c^i, \bar{c}^{i\bar{}}\}$, in order to have a Φ-Π neutral action, written in a shorthand notation as

$$\Gamma_1^{\mathrm{Cl}}[\beta,f] = \sum_f \langle \beta, sf \rangle. \quad \text{(V.11)}$$

One has $\delta_0 \Gamma_1^{\mathrm{Cl}}[\beta,f] = 0$ thanks to $s^2 = 0$. Within the following enlarged classical action, $\Gamma^{\mathrm{Cl}}[\beta,f] = \Gamma_0^{\mathrm{Cl}}[f_0] + \Gamma_1^{\mathrm{Cl}}[\beta,f]$, the BRS functional operator writes:

$$\delta_0 = \sum_f \left\langle \frac{\delta \Gamma^{\mathrm{Cl}}}{\delta \beta}, \frac{\delta}{\delta f} \right\rangle \quad \text{(V.12)}$$

and the diffeomorphism invariance of Γ^{Cl} simplifies into the so-called Slavnov identity[19]

$$\delta_0 \Gamma^{\mathrm{Cl}}[\beta, f] = 0. \quad \text{(V.13)}$$

In order to preserve the classical diffeomorphism invariance on the functional Γ^{Cl}, the BRS operation can be extended to the classical sources β, either trivially by saying that these external fields are s-invariant, or by defining

$$\begin{aligned} s\beta_{-1}(z,\bar{z}) &= \frac{\delta \Gamma^{\mathrm{Cl}}[\beta,f]}{\delta f_0(z,\bar{z})} = \frac{\delta \Gamma_0^{\mathrm{Cl}}[f_0]}{\delta f_0(z,\bar{z})} + \boldsymbol{c}\beta_{-1}(z,\bar{z}) + \tilde{V}(\beta_{-1}, f_1)(z,\bar{z}), \\ s\beta_{-2}(z,\bar{z}) &= \frac{\delta \Gamma^{\mathrm{Cl}}[\beta,f]}{\delta f_1(z,\bar{z})} = \boldsymbol{c}\beta_{-2}(z,\bar{z}) + \tilde{V}(\beta_{-2}, f_1)(z,\bar{z}) + X(f_0, \beta_{-1})(z,\bar{z}), \end{aligned} \quad \text{(V.14)}$$

where $\tilde{V}$ is the formal dual of the differential operator V given in (V.2), and X is the local differential polynomial

$$X_i(f_0, \beta_{-1})(z,\bar{z}) = \frac{\delta}{\delta c^i(z,\bar{z})} \sum_{f_0} \langle \beta_{-1}, sf_0 \rangle, \quad \text{(V.15)}$$

together with the complex conjugate expression. With these variations we can define the following nilpotent functional differential operator:

$$\delta = \delta_0 + \sum_\beta \left\langle s\beta, \frac{\delta}{\delta \beta} \right\rangle, \quad \text{(V.16)}$$

which corresponds to the linearized Slavnov operator acting on the space of local functionals in both the fields f and the sources β. Finding a quantum extension $\Gamma[\beta,f]$ of the classical functional Γ^{Cl} such that $\delta\Gamma=0$ amounts to study the δ_0-cohomology in the space of local functionals with respect to the ghost grading (the $\Phi\Pi$ charge). In the Lagrangian field theory language, a quantum extension will exist if the cohomology is trivial at the ghost grading one, i.e., there is no anomaly. It is now customary[20,10] to translate this cohomological problem into the space of local polynomials in the fields and sources and their derivatives spanned by the infinite set of local independent variables of the type

$$\partial_I\partial_{\overline{J}}\chi(z,\overline{z}),\quad \chi\equiv\{\phi,\mu,\overline{\mu},c,\overline{c},\gamma,\eta,\overline{\eta},\zeta,\overline{\zeta}\},\quad |I|\geqslant 0,\ |\overline{J}|\geqslant 0,$$
$$\partial_I\equiv(\partial_1)^{i_1}\ldots(\partial_n)^{i_n},\quad I=(i_1,\ldots,i_n),\quad |I|=i_1+\cdots+i_n, \tag{V.17}$$

$$\partial_{\overline{J}}\equiv(\overline{\partial}_1)^{j_1}\ldots(\overline{\partial}_n)^{j_n},\quad \overline{J}=(j_1,\ldots,j_n),\quad |\overline{J}|=j_1+\cdots+j_n,\delta_I^{M+N}\frac{I!}{M!N!}\equiv\prod_{k=1}^{n}\delta_{i_k}^{m_k+n_k}\frac{i_k!}{m_k!n_k!},$$

$$I+1_l=(i_1,\ldots,i_l+1,\ldots,i_n),\quad \overline{J}+1_{\overline{l}}=(j_1,\ldots,j_l+1,\ldots,j_n),$$

where we have conveniently used a multi-index notation and $|I|$ denotes the length of the multi-index I. The space of local polynomials in the fields and their derivatives can be endowed with the structure of a Fock space. The interchange between the two systems of local complex coordinates $(z,\overline{z})$ and $(Z,\overline{Z})$, can be taken into account. According to a previous work,[11] by considering the coordinates Z and $\overline{Z}$ as independent variables but non local in μ, we enlarge the set $\chi\equiv f\cup\beta$ of fields. The local cohomology will be that of the following differential operator

$$\delta=\left\langle sZ(z,\overline{z}),\frac{\partial}{\partial Z(z,\overline{z})}\right\rangle+\left\langle s\overline{Z}(z,\overline{z}),\frac{\partial}{\partial\overline{Z}(z,\overline{z})}\right\rangle+\sum_{|I|,|\overline{J}|\geqslant 0}\left(\sum_{\chi}\left\langle\partial_I\partial_{\overline{J}}s\chi(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\chi(z,\overline{z})}\right\rangle\right.$$
$$\left.+\left\langle\partial_I\partial_{\overline{J}}s\lambda(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\lambda(z,\overline{z})}\right\rangle+\left\langle\partial_I\partial_{\overline{J}}s\overline{\lambda}(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\overline{\lambda}(z,\overline{z})}\right\rangle\right), \tag{V.18}$$

which turns out to be nilpotent $\delta^2=0$ upon using both the integrability conditions (II.11) and (II.6). Note that the non local part in the μ's has been isolated.

For a given Φ-Π charge p, solutions will be obtained modulo total derivatives, and we get the following descent equations:

$$\begin{cases}\delta\Delta_{2n}^{p}(z,\overline{z})+d\Delta_{2n-1}^{p+1}(z,\overline{z})=0\\ \delta\Delta_{2n-1}^{p+1}(z,\overline{z})+d\Delta_{2n-2}^{p+2}(z,\overline{z})=0\\ \quad\vdots\\ \delta\Delta_{2n-r}^{p+r}(z,\overline{z})+d\Delta_{2n-r-1}^{p+r+1}(z,\overline{z})=0\\ \quad\vdots\\ \delta\Delta_{1}^{p+2n-1}(z,\overline{z})+d\Delta_{0}^{p+2n}(z,\overline{z})=0\\ \delta\Delta_{0}^{p+2n}(z,\overline{z})=0\end{cases} \tag{V.19}$$

where the lower index will label the Poincaré form degree and the upper index is with respect to the ghost grading. The last of these equation has the general solution

$$\Delta_0^{p+2n}(z,\overline{z})=\Delta_0^{p+2n,\natural}(z,\overline{z})+\delta\hat{\Delta}_0^{p+2n-1}(z,\overline{z}), \tag{V.20}$$

where $\Delta_0^{p+2n,\natural}(z,\overline{z})$ is an element of the δ-cohomology space in the space of local functions. It is readily seen[10,11] that

$$\boldsymbol{d}=dz^k\left\{\delta,\frac{\partial}{\partial c^k(z,\overline{z})}\right\}+d\overline{z}^k\left\{\delta,\frac{\partial}{\partial \overline{c}^k(z,\overline{z})}\right\},\tag{V.21}$$

which allows one to define the derivative operator on the local polynomials as a formal derivative

$$\begin{aligned}\partial_i&=\left\{\delta,\frac{\partial}{\partial c^i(z,\overline{z})}\right\}\\&=\left\langle\lambda_i(z,\overline{z}),\frac{\partial}{\partial Z(z,\overline{z})}\right\rangle+\left\langle(\overline{\mu}\cdot\overline{\lambda})_i(z,\overline{z}),\frac{\partial}{\partial\overline{Z}(z,\overline{z})}\right\rangle+\sum_{|I|,|J|\geqslant 0}\left(\left\langle\partial_{I+1_i}\partial_{\overline{J}}\lambda(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\lambda(z,\overline{z})}\right\rangle\right.\\&\quad\left.+\left\langle\partial_{I+1_i}\partial_{\overline{J}}\overline{\lambda}(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\overline{\lambda}(z,\overline{z})}\right\rangle+\sum_{\chi}\left\langle\partial_{I+1_i}\partial_{\overline{J}}\chi(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\chi(z,\overline{z})}\right\rangle\right),\end{aligned}\tag{V.22}$$

and one gets $\overline{\partial}_i$ by complex conjugation.

Substituting the solution Eq. (V.20) from the bottom of the system of descent equations (V.19) we successively get

$$\delta\left[\Delta_1^{p+2n-1}(z,\overline{z})+\frac{\partial\Delta_0^{p+2n,\natural}(z,\overline{z})}{\partial c^i(z,\overline{z})}dz^i+\frac{\partial\Delta_0^{p+2n,\natural}(z,\overline{z})}{\partial\overline{c}^i(z,\overline{z})}d\overline{z}^i+\boldsymbol{d}\hat{\Delta}_0^{p+2n-1}(z,\overline{z})\right]=0,\tag{V.23}$$

which implies that

$$\begin{aligned}&\Delta_1^{p+2n-1}(z,\overline{z})+\frac{\partial\Delta_0^{p+2n,\natural}(z,\overline{z})}{\partial c^i(z,\overline{z})}dz^i+\frac{\partial\Delta_0^{p+2n,\natural}(z,\overline{z})}{\partial\overline{c}^i(z,\overline{z})}d\overline{z}^i+\boldsymbol{d}\hat{\Delta}_0^{p+2n-1}(z,\overline{z})\\&\quad=\Delta_1^{p+2n-1,\natural}(z,\overline{z})+\delta\hat{\Delta}_1^{p+2n-2}(z,\overline{z})\end{aligned}\tag{V.24}$$

and by substituting step by step, we reach the top equation of (V.19) which gives the general solution of the δ-cohomology in the sector of ghost grading p[10,11]

$$\begin{aligned}\Delta_{2n}^{p}(z,\overline{z})&=\sum_{r,s=1}^{n}\frac{\partial^r}{\partial c^{A_r}(z,\overline{z})}\frac{\partial^s}{\partial\overline{c}^{A_s}(z,\overline{z})}\Delta_{2n-r-s}^{p+r+s,\natural}(z,\overline{z})dz^{A_r}\wedge d\overline{z}^{A_s}\\&\quad+\Delta_{2n}^{p,\natural}(z,\overline{z})+\delta\hat{\Delta}_{2n}^{p-1}(z,\overline{z})+\boldsymbol{d}\hat{\Delta}_{2n-1}^{p}(z,\overline{z}),\end{aligned}\tag{V.25}$$

where for the sake of compactness of writing increasing ordered multi-index notation has been used,

$$A_r=(a_1\ldots,,a_r),\quad a_1<\cdots<a_r,\quad dz^{A_r}=dz^{a_1}\wedge\cdots\wedge dz^{a_r},\quad\frac{\partial^r}{\partial c^{A_r}}=\frac{\partial^r}{\partial c^{a_1}\cdots\partial c^{a_r}},\tag{V.26}$$

with of course the complex conjugate expressions. Equation (V.25) links the elements of the diff-mod d cohomology to those of the local functional cohomology, once the $\Delta_{2n-r-s}^{p+r+s,\natural}$, $r,s=0,\ldots,n$ are known as δ-cocycles.

We are interested in finding possible anomalies in the quantum field theoretic sense, that is, $p=1$ in Eq. (V.25). According to our previous analysis in Ref. 11, it turns out that the only relevant tensorial index $2n-r-s$ is of scalar type, namely, for $r=s=n$. The problem reduces into solving the cocycle equation

$$\delta\Delta_0^{2n+1,\natural}(z,\overline{z})=0.\tag{V.27}$$

Following the lines of Refs. 10–12, we decompose the general $\Delta_0^{p,\natural}$, for $p \geqslant 2n$, with respect to its underived ghosts content, namely,

$$\Delta_0^{p,\natural}(z,\bar{z}) = \sum_{r,s=0}^{n} c^{A_r}(z,\bar{z})\bar{c}^{\bar{A}_s}(z,\bar{z})\mathcal{D}^{p-r-s}_{A_r\bar{A}_s}(z,\bar{z}), \tag{V.28}$$

where according to the summation on all possible ordered multi-indices (V.26), the $\mathcal{D}$'s are independent skew-symmetric tensors containing only derivatives of the ghost fields. This canonical decomposition identifies independent sectors according to the ghost grading. Inserting the decomposition (V.28) into the cocycle condition (V.27), introducing the following nilpotent operator:

$$\hat{\delta} = \delta - (c^i(z,\bar{z})\partial_i + \bar{c}^{\bar{\imath}}(z,\bar{z})\partial_{\bar{\imath}}), \quad \hat{\delta}^2 = 0, \tag{V.29}$$

we can canonically decompose with respect to underived ghost fields onto each ghost sector, and find the following chain of 2^{2n} equations:

$$\begin{cases}
\hat{\delta}\mathcal{D}^p(z,\bar{z}) = 0 \\
-\hat{\delta}\mathcal{D}^{p-1}_{\bar{\imath}}(z,\bar{z}) + \partial_{\bar{\imath}}\bar{c}^{\bar{J}}\mathcal{D}^{p-1}_{J}(z,\bar{z}) + \partial_{\bar{\imath}}c^j\mathcal{D}^{p-1}_{j}(z,\bar{z}) + \partial_{\bar{\imath}}\mathcal{D}^p(z,\bar{z}) = 0 \\
-\hat{\delta}\mathcal{D}^{p-1}_{i}(z,\bar{z}) + \partial_i c^j\mathcal{D}^{p-1}_{j}(z,\bar{z}) + \partial_i\bar{c}^{\bar{J}}\mathcal{D}^{p-1}_{J}(z,\bar{z}) + \partial_i\mathcal{D}^p(z,\bar{z}) = 0 \\
\vdots \\
(-)^{k+l}\hat{\delta}\mathcal{D}^{p-\underline{k}-l}_{A_k\bar{A}_l}(z,\bar{z}) + \epsilon^{A_r i}_{A_k}\Bigg(\sum_{t=1}^{k}(-)^{t-1+l}\partial_i c^{a_t}\mathcal{D}^{p-k-\underline{l}}_{A_{r+1_t}\bar{A}_l}(z,\bar{z}) \\
\quad + \sum_{t=1}^{l+1}(-)^{k+l+t}\partial_i\bar{c}^{\bar{a}_t}\delta^{\bar{A}_s}_{\bar{A}_l}\mathcal{D}^{p-\underline{k}-l}_{A_r\bar{A}_{s+1_t}}(z,\bar{z}) + (-)^{k-1}\partial_i\mathcal{D}^{p-\underline{k}-l-1}_{A_r\bar{A}_s}(z,\bar{z})\Bigg) \\
\quad + \epsilon^{\bar{A}_s\bar{\imath}}_{\bar{A}_l}\Bigg(\sum_{t=1}^{k+1}(-)^{t-1}\partial_{\bar{\imath}}c^{a_t}\delta^{A_r}_{A_k}\mathcal{D}^{p-k-\underline{l}}_{A_{r+1_t}\bar{A}_s}(z,\bar{z}) \\
\quad + \sum_{t=1}^{l}(-)^{k+t-1}\partial_{\bar{\imath}}\bar{c}^{\bar{a}_t}\mathcal{D}^{p-\underline{k}-l}_{A_k\bar{A}_{s+1_t}}(z,\bar{z}) + (-)^{k+l-1}\partial_{\bar{\imath}}\mathcal{D}^{p-\underline{k}-l-1}_{A_k\bar{A}_s}(z,\bar{z})\Bigg) = 0 \\
\vdots \\
(\hat{\delta} - \partial_i c^i(z,\bar{z}) - \partial_{\bar{\imath}}\bar{c}^{\bar{\imath}}(z,\bar{z}))\mathcal{D}^{p-2n}_{A_n\bar{A}_n}(z,\bar{z}) + (-1)^{n-1}\epsilon^{A_{n-1}r}_{A_n}\partial_r\mathcal{D}^{p-2n+1}_{A_{n-1}\bar{A}_n}(z,\bar{z}) \\
\quad - \epsilon^{\bar{A}_{n-1}\bar{J}}_{\bar{A}_n}\partial_{\bar{J}}\mathcal{D}^{p-2n+1}_{A_n\bar{A}_{n-1}}(z,\bar{z}) = 0,
\end{cases} \tag{V.30}$$

where $\epsilon^{A_r i}_{A_k}$ is the generalized Kronecker skew-symmetric tensor with respect to the multi-indices, and A_{r+1_t} means $a_1 < \cdots < a_t < \cdots < a_r$ for a given A_r. In the above formulas summation on repeated multi-indices must be performed and $\delta^{\bar{A}_s}_{\bar{A}_l}$ equals 1 if the ordered multi-indices are identical and 0 otherwise.

Since the δ-cohomology is translated in terms of the $\hat{\delta}$-cohomology, the first step is to solve the top equation of the system (V.30). Also, note that the $\hat{\delta}$ operator will act on the space of local polynomials in the following extended set of fields $\chi \equiv f \cup \beta \cup \{\lambda, \bar{\lambda}\}$. Solving the $\hat{\delta}$-cohomology is the technical part of the paper and in Appendix D the following theorem will be established.

Theorem: *Define in matrix notation the following n^2 local expressions of ghost grading one:*

$$\tilde{\Omega}(z,\bar{z}) = (\partial c + \partial\bar{c}\cdot\mu)(z,\bar{z}) = (\hat{\delta}\lambda\cdot\lambda^{-1})(z,\bar{z}), \tag{V.31}$$

and note the $\hat{\delta}$-coboundary

$$\mathrm{tr}\ \tilde{\Omega}(z,\bar{z})=\hat{\delta}\ \ln\det\lambda(z,\bar{z}). \tag{V.32}$$

Then, in the scalar sector of the space of analytic (local) functions in the fields and their derivatives, the ghost sectors with grading p of the $\hat{\delta}$-cohomology are generated by the following nontrivial cocycles

$$\text{for even } p:\quad \mathrm{tr}\ \tilde{\Omega}^{2r+1}(z,\bar{z})\mathrm{tr}\ \tilde{\bar{\Omega}}^{2s+1}(z,\bar{z}),\quad \text{for odd } p,:\quad \mathrm{tr}\ \tilde{\Omega}^{2k+1}(z,\bar{z}), \tag{V.33}$$

while the cocycle $\mathrm{tr}\ \tilde{\Omega}$ *can be reabsorbed by completing the set of generators with* $\ln\det\lambda$ *seen as an independent variable. The same statements hold true for the complex conjugate expressions.*

A nontrivial *local* anomaly modulo $\boldsymbol{d}$ will then be given by [cf. Eq. (V.25)]

$$\Delta^1_{2n}(z,\bar{z})=\frac{\partial^n}{\partial c^{A_n}(z,\bar{z})}\frac{\partial^n}{\partial \bar{c}^{A_n}(z,\bar{z})}\Delta_0^{2n+1,\natural}(z,\bar{z})dz^{A_n}\wedge d\bar{z}^{\bar{A}_n}, \tag{V.34}$$

with the *local* δ-cocycle $\Delta_0^{2n+1,\natural}$ given by the decomposition (V.28). It is also subject to the reality condition $\overline{\Delta^1_{2n}}(z,\bar{z})=\Delta^1_{2n}(z,\bar{z})$.

According to the definition of both $\hat{\delta}$ and (V.31), one considers in matrix notation the n^2 variables of ghost grading one,

$$\Omega(z,\bar{z})=(\delta\lambda\cdot\lambda^{-1})(z,\bar{z})=\tilde{\Omega}(z,\bar{z})+(c\lambda\cdot\lambda^{-1})(z,\bar{z}),\mathrm{tr}\ \Omega(z,\bar{z})=\delta\ \ln\det\lambda(z,\bar{z}).$$

Unfortunately, Ω is not local in the μ fields due to the presence of the λ terms. One has $\delta\Omega(z,\bar{z})=\Omega^2(z,\bar{z})$. Moreover, setting $\Lambda_i\equiv\partial_i\lambda\cdot\lambda^{-1}$, it is easily shown that

$$\begin{aligned}\Omega(z,\bar{z})&=(C^r\Lambda_r)(z,\bar{z})+\partial C(z,\bar{z}),\\ \delta\ \ln\det\lambda(z,\bar{z})&=(C^r\partial_r\ \ln\det\lambda)(z,\bar{z})+\mathrm{tr}(\partial C)(z,\bar{z}),\end{aligned} \tag{V.35}$$

after the use of Eq. (II.11) and the definition (III.6) of the factorized ghosts C. Note that the above variations contain a local part in the μ fields.

In turns out after some rather lengthy analysis based on the above theorem that the possible δ-cocycles $\Delta_0^{2n+1,\natural}$ are not local in the μ external fields due to the presence of the non-local λ terms. The former are listed hereafter

$$\begin{aligned}&\mathrm{tr}\ \Omega^n\ \mathrm{tr}\ \bar{\Omega}^n\ \mathrm{tr}\ \Omega=\delta(\mathrm{tr}\ \Omega^n\ \mathrm{tr}\ \bar{\Omega}^n\ \ln\det\lambda),\quad \text{for odd } n,\\ &\mathrm{tr}\ \Omega^{2r+1}\ \mathrm{tr}\ \bar{\Omega}^{2s+1}\ \mathrm{tr}\ \Omega=\delta(\mathrm{tr}\ \Omega^{2r+1}\ \mathrm{tr}\ \bar{\Omega}^{2s+1}\ \ln\det\lambda),\quad \text{for even } n,\quad n=r+s+1,\\ &\mathrm{tr}\ \Omega^{2n+1}+\mathrm{tr}\ \bar{\Omega}^{2n+1}\neq\delta(\cdots),\quad n\geqslant 3.\end{aligned} \tag{V.36}$$

The first two cocycles are trivial while the last nontrivial δ-cocycle can only produce a local cocycle of ghost grading $2n$ and not $2n+1$ as required. This shows that the local δ-cohomology modulo $\boldsymbol{d}$ is not trivial in that sector.

Therefore, if there were an anomaly, the latter is a nontrivial cocycle of the δ-cohomology modulo $\boldsymbol{d}$. So it is not surprising that the remaining possibility to be considered is that the *local* δ-cocycle $\Delta_0^{2n+1,\natural}$ matching with the principle of locality will appear as the variation of a nonlocal expression in the μ fields according to,

$$\Delta_0^{2n+1,\natural}=\delta\hat{\Delta}_0^{2n}, \tag{V.37}$$

where $\hat{\Delta}_0^{2n}$ does not belong to the nonlocal δ-cohomology modulo $\boldsymbol{d}$ and will provide nonlocal counterterms in the μ fields which are defined up to a total derivative. Since the nonlocal genera-

tors are the λ terms and that the variations (V.35) contain a local part, $\hat{\Delta}_0^{2n}$ can be formally expanded in terms of λ and its (independent) derivatives. By using once more Eq. (II.11), it assumes in the scalar sector the following finite expansion:

$$\hat{\Delta}_0^{2n}=\sum_{|I|=0}^{2n}(T^I\partial_I \ln\det\lambda+\mathrm{tr}(\mathcal{T}^{I+1_i}\cdot\partial_I\Lambda_i)+\text{c.c.}), \tag{V.38}$$

where in multi-index notation (V.17) both the scalar valued, $T^I=T^{\overbrace{1\ldots1}^{i_1}\ldots\overbrace{n\ldots n}^{i_n}}$ and matrix valued, $\mathcal{T}^{I+1_i}$, symmetric tensors are local expressions in the μ external fields with $\dim T^I=\dim \mathcal{T}^{I+1_i}+1=2n-|I|$, and $Q_{\Phi\Pi}(T^I)=Q_{\Phi\Pi}(\mathcal{T}^{I+1_i})=2n$. Recalling that in a Lagrangian formulation counterterms are defined up to a total derivative, so it is for those generated by $\hat{\Delta}_0^{2n}$. Thus, discarding the explicit form of the divergences one can write

$$\begin{aligned}\hat{\Delta}_0^{2n}&=\left(\sum_{|I|=0}^{2n}(-1)^{|I|}\partial_I T^I\right)\ln\det\lambda+\mathrm{tr}\left(\left(\sum_{|I|=0}^{2n}(-1)^{|I|}\partial_I\mathcal{T}^{I+1_i}\right)\cdot\Lambda_i\right)+\partial_i\Gamma^i+\text{c.c.}\\&\equiv R\ln\det\lambda+\mathrm{tr}(\mathcal{R}^{,i}\cdot\Lambda_i)+\partial_i\Gamma^i+\text{c.c.}\end{aligned} \tag{V.39}$$

with $\dim R=\dim\mathcal{R}^{,i}+1=2n=Q_{\Phi\Pi}(R)=Q_{\Phi\Pi}(\mathcal{R}^{,i})$. Substitution of $\hat{\Delta}_0^{2n}$ in Eq. (V.34), together with $[\partial/\partial c^\alpha(z,\bar{z}),\partial_\beta]=0$, provides the anomaly up to a total derivative, one can thus discard the divergences. The δ-variation of $\hat{\Delta}_0^{2n}$ writes

$$\begin{aligned}\delta\hat{\Delta}_0^{2n}=&\underbrace{R\ \mathrm{tr}\partial C+\mathrm{tr}(\mathcal{R}^{,i}\partial_i\cdot\partial C)}_{\text{local part}}+\delta R\ \ln\det\lambda+RC^r\partial_r\ln\det\lambda\\&+\mathrm{tr}(\delta\mathcal{R}^{,i}\cdot\Lambda_i+\mathcal{R}^{,i}\cdot(C^r\partial_r\Lambda_i+\partial_iC^r\Lambda_r+[\partial C,\Lambda_i]))+\text{c.c.},\end{aligned} \tag{V.40}$$

and by requiring the vanishing of the nonlocal part one gets

$$\text{(a)}\ \begin{cases}RC^r=0,\\ \mathcal{R}^{,i}C^r=0,\end{cases}\forall r=1,\ldots,n,\quad\text{(b)}\ \begin{cases}\delta R=0,\\ \delta\mathcal{R}_l^{k,i}=\mathcal{R}_s^{k,i}\partial_lC^s-\mathcal{R}_l^{k,s}\partial_sC^i-\mathcal{R}_l^{s,i}\partial_sC^k.\end{cases} \tag{V.41}$$

Note that $\mathcal{R}_l^{k,i}$ transforms in a rigid manner. Equations (V.41)(a) are solved by

$$\text{(a)}\ \begin{cases}R=\Xi_{A_n}C^{A_n},\\ \mathcal{R}^{,i}=\Sigma^{,i}_{A_n}C^{A_n},\end{cases}\quad\text{(b)}\ \frac{\partial\Xi_{A_n}}{\partial C^r}=\frac{\partial\Sigma^{,i}_{A_n}}{\partial C^r}=0,\quad\forall r=1,\ldots,n, \tag{V.42}$$

with $C^{A_n}\equiv C^1C^2\ldots C^n$, $\delta C^{A_n}=-\mathrm{tr}(\partial C)C^{A_n}$, and $Q_{\Phi\Pi}(\Xi_{A_n})=Q_{\Phi\Pi}(\Sigma^{,i}_{A_n})=n$. The local cocycle (V.37) then reads

$$\Delta_0^{2n+1,\natural}=C^{A_n}(\Xi_{A_n}\partial_rC^r+\Sigma^{k,i}_{l,A_n}\partial_i\partial_kC^l)+\partial_r\Gamma^r_{\mathrm{loc}}+\text{c.c.} \tag{V.43}$$

A direct substitution of (V.42)(a) into Eqs. (V.41)(b) gives, on the other hand

$$(\delta\Xi_{A_n}-\partial_rC^r\Xi_{A_n})C^{A_n}=0\Rightarrow\delta\Xi_{A_n}=\mathrm{tr}(\partial C)\Xi_{A_n}+C^r\xi_{r,A_n}, \tag{V.44}$$

with $Q_{\Phi\Pi}(\xi_{r,A_n})=n$, and on the other hand

$$\delta\Sigma^{k,i}_{l,A_n}=\mathrm{tr}(\partial C)\Sigma^{k,i}_{l,A_n}+\partial_lC^s\Sigma^{k,i}_{s,A_n}-\partial_sC^i\Sigma^{k,s}_{l,A_n}-\partial_sC^k\Sigma^{s,i}_{l,A_n}+C^r\sigma^{k,i}_{r,l,A_n}, \tag{V.45}$$

with $Q_{\Phi\Pi}(\sigma^{k,i}_{r,l,A_n})=n$. Obviously, due to (V.42)(b), the C dependence in both Eqs. (V.44) and (V.45) comes from the underived ghost content of the BRS operator and thus

$$\frac{\partial\xi_{A_n}}{\partial C^r}=\frac{\partial\sigma^{k,i}_{l,A_n}}{\partial C^r}=0,\quad\forall r=1,\ldots,n. \tag{V.46}$$

The terms Ξ_{A_n}, ξ_{r,A_n}, $\Sigma^{k,i}_{l,A_n}$ and $\sigma^{k,i}_{r,l,A_n}$ are local in the fields $\{\phi,\mu,\overline{\mu},c,\overline{c}\}$ and their explicit form has to be separately discussed according to the presence of the ϕ matter fields. The same considerations hold for the complex conjugate expressions.

A. Vacuum anomalies

In the vacuum sector, it is readily seen that $\partial_r=\{\partial/\partial C^r,\delta\}$. This easily implies together with Eqs. (V.42)(b), and (V.46) that $\xi_{r,A_n}=\partial_r\Xi_{A_n}$, and $\sigma^{k,i}_{r,l,A_n}=\partial_r\Sigma^{k,i}_{l,A_n}$. Equations (V.44) and (V.45) rewrite, respectively, as

$$\delta\Xi_{A_n}=\partial_r(C^r\Xi_{A_n}),\quad \delta\Sigma^{k,i}_{l,A_n}=\partial_r(C^r\Sigma^{k,i}_{l,A_n})+\partial_l C^s\Sigma^{k,i}_{s,A_n}-\partial_s C^i\Sigma^{k,s}_{l,A_n}-\partial_s C^k\Sigma^{s,i}_{l,A_n}, \quad \text{(V.47)}$$

showing that Ξ_{A_n} and $\Sigma^{k,i}_{l,A_n}$ are, respectively, a scalar and a tensorial densities. The inhomogeneous term in the δ-variation of μ cuts out the possible μ dependence of both Ξ_{A_n} and $\Sigma^{k,i}_{s,A_n}$; one has to figure out the latter as local expressions in the derivated ghosts C's only. According to both power counting index $2n$ and ghost grading n, it turns out that the most general expressions are labeled by the set of permutations π of the integers $\{1,\ldots,n\}$. Thus there are $n!$ scalar densities of the following type:

$$\Xi_{A_n,(\pi)}=\epsilon^{a_1\cdots a_n}\partial_{a_1}\partial_{i_1}C^{l_1}\cdots\partial_{a_n}\partial_{i_n}C^{l_n}\delta^{i_1}_{l_{\pi(1)}}\cdots\delta^{i_n}_{l_{\pi(n)}}, \quad \text{(V.48)}$$

the π's cover all the possible contractions. Saturation with the product C^{A_n} of the C's provides the solutions (V.42). The latter are numerated as the decompositions of the integer n into integer summands without regard to the order. Let $p(n)$ denote such a decomposition. Independent solutions of (V.42) can be seen as skew symmetrizations of already skew symmetric tensors of maximal rank,

$$R_{(p(n))}=\frac{(-1)^{n(n-1)/2}}{k_1!\cdots k_n!}\mathrm{tr}(V^{k_1})\cdots\mathrm{tr}(V^{k_n}), \quad \text{(V.49)}$$

for a given partition $p(n)$ of the integer n with $k_1+\cdots+k_n=n$, $k_1\leqslant\cdots\leqslant k_n$ and where in matrix notation the n^2 ghost graded 2 quantities $V^l_i\equiv C^r\partial_r\partial_i C^l$ rigidly transforms as

$$\delta V=[\partial C,V],\quad V\equiv C\cdot\partial(\partial C), \quad \text{(V.50)}$$

a transformation law very similar to that of the adjoint of $\mathrm{gl}(n,\boldsymbol{C})$ with matrix parameter ∂C. The first contributions $R_{(p(n))}\partial_r C^r$ to the cocycle (V.43) are developed with respect to the monomial $\partial_r C^r$ while the second are developed with respect to the monomials $\partial_i\partial_k C^l$. It is easy to see that $\Sigma^{k,i}_{s,A_n}\partial_k\partial_i C^s$ is a scalar density of power counting index $2n+1$ and ghost grading $n+1$, so that $\Sigma^{k,i}_{s,A_n}\partial_k\partial_i C^s=\Xi_{A_n}\mathrm{tr}(\partial C)$. One has, as diffeomorphism cocycles labeled by all the partitions of n,

$$R_{p(n)}\mathrm{tr}(\partial C)=\mathcal{R}^{k,i}_{l,p(n)}\partial_k\partial_i C^l. \quad \text{(V.51)}$$

Thus in the vacuum sector, the δ-cocycle modulo $\boldsymbol{d}$ (V.43) reduces to the following linear combination:

$$\Delta_0^{2n+1,\natural}=\sum_{p(n)}A_{p(n)}R_{(p(n))}\mathrm{tr}\ \partial C, \quad \text{(V.52)}$$

where the $A_{p(n)}$'s are complex numbers depending on the field content of the model. The cocycles (V.51) are higher dimensional analogs of the well-known Gel'fand-Fuchs cocycle.[21] For complex dimension $n=1$, one recovers the Gel'fand-Fuchs cocycle $C\partial C\partial^2 C$.[11]

Inserting the cocycle (V.52) into the computational formulas (V.34), one finds the local expression of (U-V part) of the vacuum anomaly modulo $\boldsymbol{d}$

$$\Delta^1_{2n}=(-1)^{n(n-1)/2}\sum_{p(n)} A_{p(n)}\frac{1}{k_1!\cdots k_n!}[\mathrm{tr}(\partial C)\mathrm{tr}((\partial\mathcal{V})^{k_1})\cdots\mathrm{tr}((\partial\mathcal{V})^{k_n})$$

$$+\mathrm{tr}(\partial\mathcal{V})\sum_{l=1}^{n} k_l\mathrm{tr}((\partial\mathcal{V})^{k_1})\cdots\mathrm{tr}((\partial\mathcal{V})^{k_l-1}\partial C)\cdots\mathrm{tr}((\partial\mathcal{V})^{k_n})], \tag{V.53}$$

where the summand is performed over all the partitions $p(n)$ of the integer n, $k_1+\cdots+k_n=n$, with $k_1\leqslant\cdots\leqslant k_n$, and where we have introduced the n^2 vector-valued (1,1)-forms

$$\mathcal{V}\equiv\boldsymbol{\partial}(d\bar{z}\cdot\mu)=\left(dz\cdot\frac{\partial}{\partial c}\right)\left(d\bar{z}\cdot\frac{\partial}{\partial\bar{c}}\right)V, \tag{V.54}$$

fulfilling the following Bianchi-like identities

$$\boldsymbol{\partial}\mathcal{V}=0,\quad \bar{\boldsymbol{\partial}}\mathcal{V}+[\mathcal{V},\boldsymbol{\mu}]=0, \tag{V.55}$$

thanks to the integrability condition (II.6). In the $n=1$ case, the well-known holomorphically split anomaly $\partial C\partial^2\mu$[14,15] is recovered.

Note that there is a very particular linear combination of the anomaly (V.53) given by the component of degree $2n+2$ of the Todd class according to formula (5.19) of Ref. 1. However, formula (1.11) or (8.5) of the quoted reference does not match as a diffeomorphism anomaly in the spirit of our construction in the sense that it is not a diffeomorphism δ-cocycle whereas each term in the summand (V.53) is.

In concluding this section, since the same argument holds for the complex conjugate part and according to the Ward id's (V.8), one can say that similarly to the bidimensional case, at the quantum level, the vacuum functional together with the anomaly (V.53) turns out to be holomorphically factorized in the Beltrami differential parametrizing an integrable complex structure. The anomalous Ward id's correspond to the loss of the vanishing of the divergence (V.10) at the quantum level, a similar phenomenon to the bidimensional case.[17]

A nonlocal functional $\Gamma[\mu]$ analogous to the Wess-Zumino-Polyakov action remains to be found out as functional in the integrating factor λ.

B. Matter field anomalies

Having in the previous subsection extracted all the C ghost dependence of both Ξ_{A_n} and $\Sigma^{k,i}_{l,A_n}$, their matter field dependence under the constraints (V.44) and (V.45), respectively, remain to be studied. Due to the transformation laws of matter fields under diffeomorphisms the ghost grading is this time carried by the $\bar{C}$ ghost fields. One finds according to the power counting, on the other hand for Ξ

$$\Xi_{A_n}=\bar{C}^1\cdots\bar{C}^nK_{A_n,\bar{A}_n}\equiv\bar{C}^{\bar{A}_n}K_{A_n,\bar{A}_n}, \tag{V.56}$$

where $K_{A_n,\bar{A}_n}$ is a differential polynomial in both the μ and matter fields with $Q_{\Phi\Pi}(K_{A_n,\bar{A}_n})=0$ and $\dim(K_{A_n,\bar{A}_n})=2n$. Due to the reality condition the cocycle (V.43) writes in the matter sector as

$$\Delta_0^{2n+1,\natural}=C^{A_n}\bar{C}^{\bar{A}_n}K_{A_n,\bar{A}_n}\mathrm{tr}(\partial C+\overline{\partial C})=c^{A_n}\bar{c}^{\bar{A}_n}\det(I-\mu\cdot\bar{\mu})K_{A_n,\bar{A}_n}\mathrm{tr}(\partial c+\mu\cdot\partial\bar{c}+\overline{\partial c}+\bar{\mu}\cdot\overline{\partial c}). \tag{V.57}$$

where it is more useful to rewrite the variations with respect to the true ghost fields. The cocycle condition (V.41)(b) is equivalent to

$$\delta(c^{A_n}\bar{c}^{\bar{A}_n}\det(I-\mu\cdot\bar{\mu})K_{A_n,\bar{A}_n})=0 \tag{V.58}$$

and yields

$$(\delta-\mathrm{tr}(\partial c+\overline{\partial c}))(\det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n})=c^r\kappa_{r,A_n,\overline{A}_n}+\overline{c^s}\,\overline{\kappa}_{\overline{s},A_n,\overline{A}_n}. \tag{V.59}$$

Recalling the derivative operator (V.22) combined the equation above, one finds

$$\kappa_{r,A_n,\overline{A}_n}=\partial_r(\det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n}),\quad \overline{\kappa}_{\overline{s},A_n,\overline{A}_n}=\overline{\partial_s}(\det(I-\mu\overline{\mu})K_{A_n,\overline{A}_n}), \tag{V.60}$$

which implies

$$\delta(\det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n})=\partial_r(c^r\ \det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n})+\overline{\partial_s}(\overline{c^s}\ \det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n}). \tag{V.61}$$

The latter shows that $\det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n}$ belongs to the uncharged scalar density sector of the diff-mod $\boldsymbol{d}$ cohomology. Given the following variation:

$$\delta\ \ln\det(I-\mu\cdot\overline{\mu})=(c\cdot\partial+\overline{c}\cdot\overline{\partial})\ln\det(I-\mu\cdot\overline{\mu})-\mathrm{tr}(\mu\cdot\partial\overline{c}+\overline{\mu}\cdot\overline{\partial c}) \tag{V.62}$$

together with Eq. (V.61) allows to express (V.57) as

$$\begin{aligned}\Delta_0^{2n+1,\natural}=&c^{A_n}\overline{c^{A_n}}\ \det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n}\ \mathrm{tr}(\partial c+\overline{\partial c})\\&-\delta(c^{A_n}\overline{c^{A_n}}\ \det(I-\mu\cdot\overline{\mu})K_{A_n,\overline{A}_n}\ \ln\det(I-\mu\cdot\overline{\mu})).\end{aligned} \tag{V.63}$$

On the other hand, the part coming from Σ remains under consideration. Rewriting the term

$$C^{A_n}\Sigma^{k,i}_{l,A_n}\partial_i\partial_k C^l\equiv\mathrm{tr}(VS), \tag{V.64}$$

inside the cocycle (V.43), the δ-cocycle condition yields $\delta S=[\partial C,S]$, which shows that the $2n-1$ ghost graded quantity S depends only on the C ghost fields and cannot depend on matter fields. So no contribution arises from the Σ term in the matter sector.

Summing up, the insertion of the cocycle (V.63) in the constructive equation (V.34) gives rise to the well-known trace anomaly which breaks the holomorphic splitting of the partition function when matter fields are involved.

VI. CONCLUDING REMARKS

Even if the physical motivation in studying higher complex dimensional manifolds is not well stated, the previous consideration must be regarded as an exercise. The link between Beltrami differentials as sources of relevant physical tensors in higher dimension is not yet known. The integrability condition on the μ's is taken into account and ensures the nilpotency of the BRS operator for the diffeomorphims. In fact, the integrability condition and the nilpotency are actually equivalent. In other words, demanding an integrable complex structure gives the nilpotency. Moreover, the integrability condition is preserved under diffeomorphims. The computation of the generalized Gel'fand-Fuchs unintegrated δ-cocycles (V.51) (modulo $\boldsymbol{d}$) is valid upon the use of the integrability condition on the Beltrami differential because of the hidden μ dependence in the C ghost fields and their diffeomorphism variation δC. However, if one wants to directly check that the anomaly (V.53) (which explicitly depends on μ) is a δ-cocycle then the use of the integrability condition is explicitly required.

Once more we emphasize the role of the holomorphic sector and the locality principle which is at the origin of a holomorphically split diffeomorphism anomalies in the vacuum sector with respect to an integrable complex structure. Contrary to the anomaly given in Ref. 1 for the so-called chiral diffeomorphims, the anomaly computed in the present work is more general with respect to the ''true'' diffeomorphisms. However, only the local part of the anomaly is provided, which is the ''universal'' or ultraviolet contribution, independent from the matter fields. The construction was guided by the bidimensional case and relies on the use of an intergrand factor which is non-local in the Beltrami differential. While in the bidimensional situation the conformal (Weyl) anomaly is equivalent to the well-known holomorphically factorized anomaly,[2] higher

dimensions in which diffeomorphic invariant quantities are equivalent to the factorized anomalies remain to be studied. This amounts to finding a globally defined version of the anomalies. Moreover, in the presence of matter fields, one recovers the usual trace anomaly.

Finally, a further careful analysis is needed, based on the local index theorem of Bismut-Gillet and Soulé, as pioneered in Refs. 3 and 4 in the bidimensional case or the attempt made in Ref. 1.

ACKNOWLEDGMENTS

The authors are grateful to the INFN for some financial support. One of us (S.L.) wishes to thank the INFN Section of the University of Genova for its warm hospitality as well as the Theoretical Physics Group of the University of Amsterdam where this work has been partly carried out under European financial support by Contract No. ERBCHBICT930301.

APPENDIX A: DIFFEOMORPHISM INVARIANCE OF THE INTEGRABILITY CONDITION

In this Appendix, the diffeomorphism invariance of the integrability condition on μ is proved. With $\boldsymbol{\mu}^{\varphi}$ defined by Eq. (III.2) of the main text, let us compute directly the integrability condition and similarly to the first order differential operator $\boldsymbol{\mathcal{L}}$ defined by Eq. (II.5), one set $\boldsymbol{\mathcal{L}}^{\varphi}=\overline{\boldsymbol{\partial}}-\boldsymbol{\mu}^{\varphi}$. The (0,2)-form with (1,0)-vector values reads

$$\begin{aligned}\overline{\boldsymbol{\partial}}\boldsymbol{\mu}^{\varphi}-(\boldsymbol{\mu}^{\varphi})^2=\boldsymbol{\mathcal{L}}^{\varphi}\boldsymbol{\mu}^{\varphi}&=-(\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot(\boldsymbol{\mathcal{L}}^{\varphi}(\mu\circ\varphi))\cdot(\partial\varphi+\partial\overline{\varphi}\cdot(\mu\circ\varphi))^{-1}\cdot\partial\\&=-(\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot((\boldsymbol{\mathcal{L}}^{\varphi}\varphi)\cdot\partial\mu+(\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot\overline{\partial}\mu)\cdot(\partial\varphi+\partial\overline{\varphi}\cdot(\mu\circ\varphi))^{-1}\cdot\partial\\&=-(\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot((\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot\mathcal{L}\mu)\cdot(\partial\varphi+\partial\overline{\varphi}\cdot(\mu\circ\varphi))^{-1}\cdot\partial,\end{aligned}\tag{A1}$$

where the third equality is obtained by using $\partial=\partial\varphi\cdot\partial+\partial\overline{\varphi}\cdot\overline{\partial}$ and $\overline{\boldsymbol{\partial}}=\overline{\boldsymbol{\partial}}\varphi\cdot\partial+\overline{\boldsymbol{\partial}\varphi}\cdot\overline{\partial}$, while the fourth one comes from the identity $\boldsymbol{\mathcal{L}}^{\varphi}\varphi=-(\boldsymbol{\mathcal{L}}^{\varphi}\overline{\varphi})\cdot(\mu\circ\varphi)$. Now, the integrability condition for μ writes

$$\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0\Leftrightarrow\mathcal{L}_{\bar{\imath}}\mu^{k}_{J}=\mathcal{L}_{\bar{J}}\mu^{k}_{\bar{\imath}},\tag{A2}$$

and shows that $\mathcal{L}\mu$ is a mixed tensor, symmetric in the lower indices. This symmetry implies the vanishing of the last expression (A1),

$$\overline{\boldsymbol{\partial}}\boldsymbol{\mu}-\boldsymbol{\mu}^2=0\Rightarrow\overline{\boldsymbol{\partial}}\boldsymbol{\mu}^{\varphi}-(\boldsymbol{\mu}^{\varphi})^2=0.\tag{A3}$$

APPENDIX B: ON THE EQUIVALENCE BETWEEN COMPLEX STRUCTURES

Any diffeomorphism φ when expressed in terms of its local representative will induce a smooth change of local coordinates. This change of variables can be written in matrix notation (with matrix entries), as well the inverse change of variables, as

$$\begin{pmatrix}\partial\varphi & \partial\overline{\varphi}\\ \overline{\partial\varphi} & \overline{\partial\varphi}\end{pmatrix},\quad\begin{pmatrix}\mathrm{A} & \mathrm{B}\\ \mathrm{C} & \mathrm{D}\end{pmatrix},\tag{B1}$$

with the following relations:

$$\begin{aligned}&\mathrm{A}\cdot\partial\varphi+\mathrm{B}\cdot\overline{\partial\varphi}=I_w,\quad&&\partial\varphi\cdot\mathrm{A}+\partial\overline{\varphi}\cdot\mathrm{C}=I_z,\\&\mathrm{A}\cdot\partial\overline{\varphi}+\mathrm{B}\cdot\overline{\partial\varphi}=0,\quad&&\partial\varphi\cdot\mathrm{B}+\partial\overline{\varphi}\cdot\mathrm{D}=0,\\&\mathrm{C}\cdot\partial\varphi+\mathrm{D}\cdot\overline{\partial\varphi}=0,\quad&&\overline{\partial\varphi}\cdot\mathrm{A}+\overline{\partial\varphi}\cdot\mathrm{C}=0,\\&\mathrm{C}\cdot\partial\overline{\varphi}+\mathrm{D}\cdot\overline{\partial\varphi}=I_{\bar{w}},\quad&&\overline{\partial\varphi}\cdot\mathrm{B}+\overline{\partial\varphi}\cdot\mathrm{D}=I_{\bar{z}}.\end{aligned}\tag{B2}$$

Starting with the equation in the target chart with respect to φ, $(w,\overline{w})=\varphi(z,\overline{z})$,

$$(\partial_{\bar{w}} - (\mu \circ \varphi) \cdot \partial_w) Z = 0, \tag{B3}$$

and by using the two last identities of the right column of the above equations (B2), one gets in the source chart of φ,

$$(\partial_{\bar{z}} - (\bar{\partial}\varphi + \overline{\partial\varphi} \cdot (\mu \circ \varphi)) \cdot (\mathrm{B} \cdot \partial_{\bar{z}} + \mathrm{A} \cdot \partial_Z))(Z \circ \varphi) = 0, \tag{B4}$$

and by Eq. (III.1) of the main text combined with the two first identities of the right column of (B2), one has

$$(\partial_{\bar{z}} - (\mu^\varphi) \cdot \partial_Z + (\mu^\varphi) \cdot \bar{\partial}\overline{\varphi} \cdot [(\mathrm{C} \cdot \partial_Z + \mathrm{D} \cdot \partial_{\bar{z}}) - (\mu \circ \varphi) \cdot (\mathrm{A} \cdot \partial_Z + \mathrm{B} \cdot \partial_{\bar{z}})])(Z \circ \varphi) = 0. \tag{B5}$$

The terms involving $\partial\overline{\varphi}$ is nothing but proportional to Eq. (B3) due to the laws for the change of variables, and thus the last parenthesis vanishes. The remaining

$$(\partial_{\bar{z}} - (\mu^\varphi) \cdot \partial_Z)(Z \circ \varphi) = 0, \tag{B6}$$

shows that $(Z \circ \varphi)$ are also n linear independent solutions of the Beltrami equations on the source of the diffeomorphism φ, see Eq. (III.4) in the main text. Recall the latter

$$(\partial_{\bar{z}} - (\mu^\varphi) \cdot \partial_Z) Z^\varphi = 0. \tag{B7}$$

Thus the mapping $(Z^\varphi, \overline{Z}^\varphi) \to (Z, \overline{Z})$ is indeed biholomorphic; see Theorem 5.3 in Ref. 13. In other words, this result states the equivalence between the complex structures parametrized by $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^\varphi$.

APPENDIX C: TRUE HOLOMORPHIC DIVERGENCE

In the following the computation of the divergence (V.10) is performed. In doing so, several results are required and are listed hereafter.

First, the computation of the inverse matrix

$$\begin{pmatrix} I & \overline{\mu} \\ \mu & I \end{pmatrix}^{-1} = \begin{pmatrix} (I - \overline{\mu} \cdot \mu)^{-1} & -(I - \overline{\mu} \cdot \mu)^{-1} \cdot \overline{\mu} \\ -(I - \mu \cdot \overline{\mu})^{-1} \cdot \mu & (I - \mu \cdot \overline{\mu})^{-1} \end{pmatrix}, \tag{C1}$$

yields in particular the following two identities:

$$(I - \overline{\mu} \cdot \mu)^{-1} - \overline{\mu} \cdot (I - \mu \cdot \overline{\mu})^{-1} \cdot \mu \equiv I, \quad (I - \overline{\mu} \cdot \mu)^{-1} \cdot \overline{\mu} - \overline{\mu} \cdot (I - \mu \cdot \overline{\mu})^{-1} \equiv 0. \tag{C2}$$

Second, one will use the useful identities

$$\partial_{Z^k}\left(\frac{1}{\mathcal{J}} \lambda_i^k\right) + \partial_{\bar{Z}^l}\left(\frac{1}{\mathcal{J}} \overline{\mu_i^r} \overline{\lambda_r^l}\right) \equiv 0, \quad \partial_{Z^k}\left(\frac{1}{\mathcal{J}} \mu_l^r \lambda_r^k\right) + \partial_{\bar{Z}^l}\left(\frac{1}{\mathcal{J}} \overline{\lambda_l^l}\right) \equiv 0, \tag{C3}$$

coming from the Jacobi multipliers device.

We are now in a position to compute the divergences (V.10) of the main text. Taking into account (II.11) and (II.15) one can show, on the one hand,

$$\frac{(I - \mu \cdot \overline{\mu})_l^J \overline{\lambda}_J^r}{\det(\overline{\lambda}(I - \mu \cdot \overline{\mu}))} \partial_{\bar{Z}^r}\left(\frac{1}{\det \lambda} (\lambda^{-1})_k^s \overline{\Theta_s^l}\right) = \frac{1}{\mathcal{J}} (\lambda^{-1})_k^s (\mathcal{L}_l \overline{\Theta_s^l} - \partial_r \mu_l^r \overline{\Theta_s^l} - \partial_s \mu_l^r \overline{\Theta_r^l}) \tag{C4}$$

which relates to the Ward id's (V.8) by virtue of the non-singularity of λ. On the other hand it is straightforward to show that

$$\partial_{Z^r}\left(\frac{1}{\det(\overline{\lambda}(I-\mu\cdot\overline{\mu}))}(I-\mu\cdot\overline{\mu})\cdot\overline{\lambda^r}\right)\equiv 0, \tag{C5}$$

thanks to the identities (C3) and (C2).

APPENDIX D: THE $\hat{\delta}$ COHOMOLOGY

The $\hat{\delta}$ operator, given by formulas (V.29) in the text, is defined on the space of local functions seen, according to the power counting, as differential polynomials in both the matter and the Φ-Π charged fields, and as analytic in the components of the Beltrami differential. The following set of fields:

$$\chi\equiv f_0\cup f_1\cup\beta\cup\{\lambda,\overline{\lambda}\},\quad f_0\equiv\{\phi,\mu,\overline{\mu}\},\quad f_1\equiv\{c,\overline{c}\},\quad \beta\equiv\{\gamma,\eta,\overline{\eta},\zeta,\overline{\zeta}\},$$

will serve as set of generators for our space of local functions. Thanks to the definition of $\hat{\delta}$, these local functions do not contain any underived ghosts c's.

Let us recall that the fields and their derivatives will be considered as independent coordinates, and in practice, will play the role of creation operators, while the annihilators will be the formal derivatives with respect to these coordinates, both acting on the Fock space structure enjoyed by the space of local functions. This Fock space is graded according to the ghost grading. For any operator, its adjoint will be given by the formal replacement of the derivative with respect to some coordinate by the formal multiplication with respect the same coordinate and vice versa.

According to this, let us now introduce the following self-adjoint operator:

$$\nu=\sum_{|I|,|J|\geqslant 0}(|I|+|\overline{J}|)\left(\sum_{f_1}<\partial_I\partial_{\overline{J}}f_1(z,\overline{z})\frac{\partial}{\partial\partial_I\partial_{\overline{J}}f_1(z,\overline{z})}\right) \tag{D1}$$

whose eigenvalues will provide the order of the derivatives of the ghost fields. The operator ν will decompose the space of local functions into a direct sum of subspaces according to its eigenvalues while the operator $\hat{\delta}$ will be filtered according to

$$[\nu,\hat{\delta}]=\sum_{|I|,|J|\geqslant 0}(|I|+|\overline{J}|)\hat{\delta}(|I|+|\overline{J}|),\quad \hat{\delta}=\sum_{|I|,|J|\geqslant 0}\hat{\delta}(|I|+|\overline{J}|). \tag{D2}$$

The general spectral method[20,10] insures that the $\hat{\delta}$-cohomology is isomorphic to the solutions $\tilde{\Delta}$ of the system

$$\begin{cases}\hat{\delta}(|I|+|\overline{J}|)\tilde{\Delta}(z,\overline{z})=0\\ \hat{\delta}^{\dagger}(|I|+|\overline{J}|)\tilde{\Delta}(z,\overline{z})=0\end{cases},\quad |I|+|\overline{J}|\geqslant 0. \tag{D3}$$

In other words, $\tilde{\Delta}$ is in the kernel of the Laplacians $\{\hat{\delta}(|I|+|\overline{J}|),\hat{\delta}^{\dagger}(|I|+|\overline{J}|)\}$.

In the following, we present in full detail the solution of the cohomology by the spectral sequence device. The zero eigenvalue of ν will select the part of $\hat{\delta}$ in which no ghost field is involved, namely,

$$\begin{aligned}\hat{\delta}(0)=\sum_{|I|,|\overline{J}|\geqslant 0}\Bigg(&\sum_{\beta_{-1}}\left\langle\partial_I\partial_{\overline{J}}\frac{\delta\Gamma_0^{\mathrm{Cl}}}{\delta f_0(z,\overline{z})},\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\beta_{-1}(z,\overline{z})}\right\rangle\\&+\sum_{\beta_{-2}}\left\langle\partial_I\partial_{\overline{J}}X(f_0,\beta_1)(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\beta_{-2}(z,\overline{z})}\right\rangle\Bigg).\end{aligned} \tag{D4}$$

Note that the latter is a nilpotent operator and is an annihilator with respect to the external sources coupled to the BRS variations. One will prove that the $\hat{\delta}$-cohomology does not actually depend on the external sources coupled to the BRS variations by performing a double-filtration. One first filtrates according to the counting operator

$$\mathcal{N}(\beta)=\sum_{|I|,|\overline{J}|\geqslant 0}\left\langle \partial_I\partial_{\overline{J}}\beta(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\beta(z,\overline{z})}\right\rangle, \tag{D5}$$

with respect to *one* of the external sources $\beta\equiv\{\gamma,\eta^{\overline{l}}_k,\overline{\eta^l_J},\zeta_i,\overline{\zeta_{\overline{i}}}\}$. Next, the second step of filtration is performed with respect to the full counting operator

$$\mathcal{N}=\sum_{\chi}\sum_{|I|,|\overline{J}|\geqslant 0}\left\langle \partial_I\partial_{\overline{J}}\chi(z,\overline{z}),\frac{\partial}{\partial\partial_I\partial_{\overline{J}}\chi(z,\overline{z})}\right\rangle. \tag{D6}$$

By virtue of the isomorphism (D3), the Laplacian corresponding to the given operator in the null subsector of the double filtration $[\mathcal{N},[\mathcal{N}(\beta),\hat{\delta}(0)]]$, shows that the $\hat{\delta}$-cohomology does not depend on any negatively ghost graded external sources β, thus restricts the set of generators into $\chi\equiv f\cup\{\lambda,\overline{\lambda}\}$, $f\equiv f_0\cup f_1$ on which one will now work with.

The second operator $\hat{\delta}(1)$ in the filtration (D2) will necessarily contain the ghost part, $\hat{\delta}_{\mathrm{Gh}}$ of $\hat{\delta}$ since $[\nu,\hat{\delta}_{\mathrm{Gh}}]=\hat{\delta}_{\mathrm{Gh}}$. The latter will play a central role in the computation of the cohomology. For both this reason and for the sake of compactness in the notation let us write it down as

$$\hat{\delta}_{\mathrm{Gh}}=\sum_{|I|\geqslant 0}\delta_I^{M+N}\frac{(I+1_\alpha)!}{(M+1_\alpha)!N!}(\partial_{M+1_\alpha}c^\beta\partial_{N+1_\beta}c^\rho)(z,\overline{z})\frac{\partial}{\partial\partial_{I+1_\alpha}c^\rho(z,\overline{z})}, \tag{D7}$$

$$\alpha,\beta,\rho=1,\ldots,n,\ldots,2n,\quad \alpha\equiv\begin{cases} i, & \text{for } \alpha=1,\ldots,n,\\ \overline{\imath}, & \text{for } \alpha=n+1,\ldots,2n;\end{cases}$$

$$I=(i_1,\ldots,i_{2n}),\quad i_\alpha=\begin{cases} i_k, & \text{for } \alpha=k=1,\ldots,n;\\ \overline{\imath_l}, & \text{for } \alpha=n+1,\ldots,2n,\quad l=\alpha-n\end{cases}$$

$$c^\alpha=\begin{cases} c^i, & \text{for } \alpha=1,\ldots,n;\\ \overline{c^i}, & \text{for } \alpha=n+1,\ldots,2n.\end{cases}$$

The former can be written as

$$\hat{\delta}(1)\equiv\partial_\alpha c^\beta(z,\overline{z})U^\alpha_\beta+R(1), \tag{D8}$$

where, according to the ghost grading, one can make the following decomposition for the operator $U^\alpha_\beta\equiv(U_0)^\alpha_\beta+(U_1)^\alpha_\beta$, where $(U_0)^\alpha_\beta$ is given by the Lie derivatives of the fields described as components of geometrical objects and $(U_1)^\alpha_\beta$ comes from $\hat{\delta}_{\mathrm{Gh}}$. For instance, in the case of matter fields $U^i_j\equiv S^i_j$ yields

$$(S_0)^i_j=\sum_{|M|,|N|\geqslant 0}\Bigg((m_i+1)\sum_{\chi\equiv f_0\cup\{\lambda,\overline{\lambda}\}}\left\langle \partial_{M+1_j}\partial_{\overline{N}}\chi(z,\overline{z}),\frac{\partial}{\partial\partial_{M+1_i}\partial_{\overline{N}}\chi(z,\overline{z})}\right\rangle$$
$$\times\sum_{A_{n-1}}\delta^{A_{r+1_j}}_{A_{n-1}}\left\langle \partial_M\partial_{\overline{N}}\phi_{A_{r+1_j}}(z,\overline{z}),\frac{\partial}{\partial\partial_M\partial_{\overline{N}}\phi_{A_{r+1_i}}(z,\overline{z})}\right\rangle$$
$$+\partial_M\partial_{\overline{N}}\lambda^r_j(z,\overline{z})\frac{\partial}{\partial\partial_M\partial_{\overline{N}}\lambda^r_i(z,\overline{z})}+\partial_M\partial_{\overline{N}}\overline{\mu^s_j}(z,\overline{z})\frac{\partial}{\partial\partial_M\partial_{\overline{N}}\overline{\mu s_i}(z,\overline{z})}$$
$$-\partial_M\partial_{\overline{N}}\mu^i_s(z,\overline{z})\frac{\partial}{\partial\partial_M\partial_{\overline{N}}\overline{\mu^j_s}(z,\overline{z})}\Bigg);\tag{D9}$$

$$(S_1)^i_j=\sum_{|M|\geqslant 0}(m_i+1)\partial_{M+1_j}c^\alpha(z,\overline{z})\frac{\partial}{\partial\partial_{M+1_i}c^\alpha(z,\overline{z})}-\sum_{|M|\geqslant 1}\partial_M c^i(z,\overline{z})\frac{\partial}{\partial\partial_M c^j(z,\overline{z})}.\tag{D10}$$

Note in addition that the trace $S^i_i=\mathrm{tr}\,S$ is nothing else but the ''little z'' indices counting operator

$$\mathrm{tr}\,S=N_\partial+N_\lambda+N_{\overline{\mu}}+N_\phi-N_\mu-N_c\equiv N_z(\downarrow)-N_z(\uparrow).\tag{D11}$$

One has $S^{\dagger j}_i=S^{\sim j}_i\equiv S^i_j$ and $\overline{S}\equiv N_{\overline{z}}(\downarrow)-N_{\overline{z}}(\uparrow)$ with $\overline{S^\dagger}=\overline{S}^\sim$. In full generality, one has for the ghost contribution

$$(U_1)^\alpha_\beta=\sum_{|I|\geqslant 0}\left((i_\alpha+1)\partial_{I+1_\beta}c^\rho(z,\overline{z})\frac{\partial}{\partial\partial_{I+1_\alpha}c^\rho(z,\overline{z})}-\partial_{I+1_\rho+1_\sigma}c^\alpha\frac{\partial}{\partial\partial_{I+1_\rho+1_\sigma}c^\beta(z,\overline{z})}\right),\tag{D12}$$

and the remaining piece $R(1)$ concerns only the ghost derivatives of order greater or equal to three,

$$R(1)=\sum_{|M|\geqslant 0}\partial_{M+1_\alpha+1_\beta}c^\rho(z,\overline{z})R^{M+1_\alpha+1_\beta}_\rho(z,\overline{z}),R^{M+1_\alpha+1_\beta}_\rho(z,\overline{z})$$
$$=\sum_{|N|\geqslant 0}\frac{(M+N+1_\alpha+1_\beta+1_\gamma)!}{(M+1_\alpha+1_\beta)!(N+1_\gamma)!}\partial_{N+1_\rho+1_\gamma}c^\sigma(z,\overline{z})\frac{\partial}{\partial\partial_{M+N+1_\alpha+1_\beta+1_\gamma}c^\sigma(z,\overline{z})}.\tag{D13}$$

Furthermore, the nilpotency, $\hat{\delta}(1)^2=0$, allows one to apply once more the spectral sequence analysis to the operator: $\hat{\delta}(1)$. The filtration will be performed with respect to the following counting operator:

$$\nu'=1+\partial_\alpha c^\beta(z,\overline{z})\frac{\partial}{\partial\partial_\alpha c^\beta(z,\overline{z})}.\tag{D14}$$

Remarkably, the cohomology reduces to the following finite filtration:

$$[\nu',\hat{\delta}(1)]=\sum_{a=1}^{2}a\,\hat{\delta}'(a),\quad \text{where}\begin{cases}\hat{\delta}'(1)=R(1),\\ \hat{\delta}'(2)=\partial_\alpha c^\beta(z,\overline{z})U^\alpha_\beta,\end{cases}\tag{D15}$$

and $\hat{\delta}(1)=\hat{\delta}'(1)+\hat{\delta}'(2)$. According to (D3), one has to solve the system

$$\begin{cases}\hat{\delta}'(a)\tilde{\Delta}(z,\overline{z})=0\\ \hat{\delta}'^\dagger(a)\tilde{\Delta}(z,\overline{z})=0\end{cases},\quad \text{for } a=1,2,\tag{D16}$$

which turns out to be equivalent to

$$\left\langle \tilde{\Delta}(z,\overline{z})|\{\hat{\delta}'^{\dagger}(a),\hat{\delta}'(a)\}|\tilde{\Delta}(z,\overline{z})\right\rangle = ||\hat{\delta}(a)\tilde{\Delta}(z,\overline{z})||^2 + ||\hat{\delta}'^{\dagger}(a)\tilde{\Delta}(z,\overline{z})||^2 \stackrel{!}{=} 0, \tag{D17}$$

where the scalar product is that the local functions induced by the definition of the adjoint operator. For $a=1$, the lowest order of formal derivatives with respect to the ghost derivatives in the Laplacian involved in Eq. (D17) can be seen of order three. So, the cocycle $\tilde{\Delta}$ has to be dependent on the first and second order derivatives of the ghost fields. The second filtration $a=2$, will select the scalar sector, and eliminates the second order ghost derivatives. It is then useful to perform the following change of variables in the first order derivatives of the ghost fields. Define in matrix notation the n^2 variables of ghost grading one,

$$\tilde{\Omega}(z,\overline{z}) = (\partial c + \partial\overline{c}\cdot\mu)(z,\overline{z}) = (\hat{\delta}\lambda\cdot\lambda^{-1})(z,\overline{z}), \tag{D18}$$

with $(\mathrm{tr}\ S)\tilde{\Omega}=0$, and note the $\hat{\delta}$-coboundary

$$\mathrm{tr}\ \tilde{\Omega}(z,\overline{z}) = \hat{\delta}\ \ln\det\lambda(z,\overline{z}). \tag{D19}$$

One has $\hat{\delta}\tilde{\Omega}(z,\overline{z}) = \hat{\delta}'(2)\tilde{\Omega}(z,\overline{z}) = \tilde{\Omega}^2(z,\overline{z})$, and

$$\hat{\delta}\tilde{\Omega}^{2k+1}(z,\overline{z}) = \tilde{\Omega}^{2k+2}(z,\overline{z}), \tag{D20}$$

showing that the $\tilde{\Omega}^{2k+2}$ are $\hat{\delta}$-coboundaries. By taking the trace in order to be in the scalar sector, one readily checks that

$$\hat{\delta}\ \mathrm{tr}\ \tilde{\Omega}^{2k+1}(z,\overline{z}) = \mathrm{tr}\ \tilde{\Omega}^{2k+2}(z,\overline{z}) = -\mathrm{tr}\ \tilde{\Omega}^{2k+2}(z,\overline{z}) = 0. \tag{D21}$$

This yields the following theorem which is the main result of this Appendix.

Theorem: *In the scalar sector of the space of analytic functions in the fields and their derivatives, the ghost sectors with grading p of the $\hat{\delta}$-cohomology are generated by the following nontrivial cocycles*:

$$\text{for even } p: \quad \mathrm{tr}\ \tilde{\Omega}^{2r+1}(z,\overline{z})\mathrm{tr}\ \overline{\tilde{\Omega}}^{2s+1}(z,\overline{z}),$$

$$\text{for odd } p: \quad \mathrm{tr}\ \tilde{\Omega}^{2k+1}(z,\overline{z}),$$

while the cocycle $\mathrm{tr}\ \tilde{\Omega}$ *can be reabsorbed by completing the set of generators with* $\ln\det\lambda$ *seen as an independent variable. The same statements hold true for the complex conjugate expressions.*

[1] A. Losev, G. Moore, N. Nekrasov, and S. Shatashvili, ''Chiral Lagrangians, anomalies, supersymmetry and holomorphy,'' Nucl. Phys. B **484**, 196 (1997), hep-th/9606082.

[2] M. Knecht, S. Lazzarini, and F. Thuillier, ''Shifting the Weyl anomaly to the chirally split diffeomorphism anomaly in two dimensions,'' Phys. Lett. B **251**, 279–283 (1990).

[3] M. Knecht, S. Lazzarini, and R. Stora, ''On holomorphic factorization for free conformal fields,'' Phys. Lett. B **262**, 25–31 (1991).

[4] M. Knecht, S. Lazzarini, and R. Stora, ''On holomorphic factorization for free conformal fields II, '' Phys. Lett. B **273**, 63–66 (1991).

[5] C. Vafa, ''Mirror transform and string theory,'' Lecture given at the Geometry and Topology Conference, April '93, Havard, in honor of R. Bott, hep-th/940315, March 1994.

[6] M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa, ''Kodaira-spencer theory of gravity and exact results for quantum string amplitudes,'' Commun. Math. Phys. **165**, 311–428 (1994).

[7] M. Bershadsky and V. Sadov, ''Theory of Kähler gravity,'' hep-th/9410011, 1994.

[8] J. M. F. Labastida and M. Marino (Santiago de Compostela U.), ''Type B topological matter, Kodaira-Spencer theory, and mirror symmetry,'' Phys. Lett. B **333**, 386–395 (1994).

[9] R. Zucchini, ''The quaternionic structure of 4d conformal field theory,'' gr-qc/9707048, July 1997.

[10] G. Bandelloni, ''Diffeomorphism cohomology in quantum field theory models,'' Phys. Rev. D **38**, 1156 (1988).

[11] G. Bandelloni and S. Lazzarini, ''Diffeomorphism cohomology in Beltrami parametrization,'' J. Math. Phys. **34**, 5413 (1993).

[12] G. Bandelloni and S. Lazzarini, ''Diffeomorphism cohomology in Beltrami parametrization II: the 1-forms,'' J. Math. Phys. **36**, 1–29 (1995).
[13] K. Kodaira, *Complex Manifolds and Deformation of Complex Structures*, Comprehensive Studies in Mathematics (Springer-Verlag, New-York, 1986).
[14] L. Baulieu and M. Bellon, ''Beltrami parametrization in string theory,'' Phys. Lett. B **196**, 142 (1987).
[15] C. Becchi, ''On the covariant quantization of the free string: the conformal structure,'' Nucl. Phys. B **304**, 513 (1988).
[16] R. Stora, ''The role of locality in string theory,'' in *Nonperturbative Quantum Field Theory*, edited by G. 't Hooft *et al.*, NATO ASI Ser. B, Vol. 185 (Plenum, New York, 1988).
[17] S. Lazzarini, ''*Sur les modèles conformes lagrangiens bidimensionnels,*'' PhD thesis, Université de Savoie, April 1990, unpublished.
[18] J.-L. Koszul, *Lectures on Fibre Bundles and Differential Geometry*, Re-issued 1965 (Tata Institute of Fundamental Research, Bombay, 1960).
[19] O. Piguet and A. Rouet, ''Symmetries in perturbative quantum field theory,'' Phys. Rev. B **76**, 1–77 (1981).
[20] J. Dixon, *Cohomology and Renormalization of Gauge Theories I-II-III* (unpublished, 1976-1979).
[21] D. B. Fuchs, *Cohomology of Infinite Dimensional Algebra* (Consultant Bureau, New York, 1986).

Time-dependent plane wave and multipole expansions of the electromagnetic field

Edwin A. Marengo[a] and Anthony J. Devaney
Center for Electromagnetics Research, Northeastern University, Boston, Massachusetts 02115

(Received 24 November 1997; accepted for publication 8 April 1998)

A new and conceptually simple derivation is presented of the time-dependent multipole expansion of the electromagnetic field radiated by a time-varying, localized, volume charge-current distribution. The analysis is based on a new time-dependent plane wave representation of the electromagnetic field (i.e., an ''angular spectrum expansion'' in the time domain), also derived in the paper. Expressions are given for the time-dependent plane wave spectra in terms of a fourfold Radon transform representation of the transverse current distribution. Two alternative expressions for the time-dependent multipole moments are derived; the first gives them in terms of the spectral amplitude vectors of the corresponding time-dependent plane wave representation while the second gives them in terms of a weighted radial-temporal average of the current distribution. Thus the analysis also sheds light on the relationship between the time-dependent plane wave and multipole expansions in their common domains of validity. © *1998 American Institute of Physics.*
[S0022-2488(98)01007-X]

I. INTRODUCTION

In describing transient wave phenomena, descriptions directly in space-time domain (where the sources and fields are well-localized) are usually more efficient and insightful than their conventional, frequency domain counterparts. Examples include the pulsed beams (PBs) and the focus wave modes (FWMs), among others.[1,2] This investigation is concerned with the time domain analogs of more traditional (frequency-domain-based) representations, namely, the plane wave and multipole expansions. Thus, the electromagnetic field radiated by a time-varying (e.g., pulsed), localized, volume charge-current distribution is expressed as a sum of time-dependent spherical (multipole) wave functions. We also obtain, via analytic signal representation, a new time-dependent plane wave representation of the electromagnetic field (i.e., an ''angular spectrum expansion'' in the time domain). The relationship between the expansion coefficients (moments) of the time-dependent plane wave and multipole expansions in their common domains of validity is also examined. The work reported here is an extension to the electromagnetic field case of the time domain multipole theory for scalar fields developed in Ref. 3. It is also an extension to the time domain of the frequency domain plane wave representations and multipole expansions of the electromagnetic field developed in Ref. 4.

Time-dependent multipolelike (spherical wave) expansions of electromagnetic radiation fields are not new. Early work in this area is associated with the names of Granzow,[5] Davidon[6] and Campbell *et al.*[7] More recent contributions include those of Heyman and Devaney,[3] where a scalar time-dependent multipole theory for volume source distributions is developed and Hansen,[8] where time-dependent multipole fields are applied to address theoretical aspects of spherical near field scanning. However, even though work in this field is not new, both the procedure we use to derive our results as well as many of the results are new. For instance, the time domain multipole theories in Refs. 6 and 7 originate from the Debye-potential formalism[4] while ours originates from the time-dependent plane wave representation, also derived in the paper. Granzow's[5] and Hansen's[8] focus is on the diffraction problem for prescribed field data over a boundary while ours is on the radiation problem for volume current distributions. Furthermore, our formulation is conceptually

[a]Electronic mail: emarengo@cdsp.neu.edu

 © 1998 American Institute of Physics

simple and insightful in describing the relationship between the time-dependent expansion coefficients (moments) of the time-dependent plane wave and multipole expansions in their common domains of validity. Aspects of our time-dependent plane wave representations of the electromagnetic field have also appeared before in connection with time domain planar near field scanning.[9] Our formulation, however, is given in terms of charge-current distributions (i.e., radiation problem), unlike those in Ref. 9, which are given in terms of fields prescribed over a boundary (i.e., diffraction problem). More importantly, our analysis not only corroborates those of previous contributions but also yields new, alternative expressions for both types of expansions.

Frequency domain plane wave and multipole expansions of scalar and electromagnetic fields have played a key role in classical and quantum-mechanical theories of wave radiation, propagation and diffraction and in defining fundamental limits in antenna theory.[10,11] Thus, in analogy to their frequency domain counterparts, the time-dependent expansion schemes considered here could be useful in treating canonical radiation and scattering problems directly in the time domain. They could also be of interest in diffraction problems formulated in terms of near or far field data (e.g., time domain near field scanning[8] and inverse diffraction from far field data[4]). Furthermore, they could also help in elucidating transient radiation phenomena (e.g., fundamental limitations of antennas driven by short pulses) and—by doing so—also in defining relevant figures of merit for transient radiation.

The remainder of the paper is organized as follows. Section II reviews known results on plane wave and multipole expansions for scalar fields. We examine, in analytic signal representation, the Weyl expansion (or angular spectrum expansion) of the scalar field generated by a localized scalar source distribution that is turned on at some initial time (say $t=0$). We then derive, again in analytic signal representation, a time domain multipole expansion of the scalar field based on the corresponding time-dependent plane wave representation. Methodologically, we Fourier invert to the time domain the corresponding frequency domain multipole expansion. The final form of the time-dependent multipole expansion is given explicitly in terms of real signals using the real spherical harmonics $S^{(1)}_{\ell,m}(\theta,\phi)$ and $S^{(2)}_{\ell,m}(\theta,\phi)$ (also known as tesseral harmonics[12]), defined, respectively, by the real and imaginary parts of the ordinary spherical harmonic of degree ℓ and order m, $Y_{\ell,m}(\theta,\phi)$. An alternative form of the time-dependent multipole expansion in which the time-dependent multipole moments are expressed as a weighted radial-temporal average of the source distribution is also derived.

The main contribution of the paper is contained in Sec. III. There we extend the theory developed in Sec. II for scalar fields to the electromagnetic field case, with the vector spherical harmonics in the latter case playing a role analogous to that of the ordinary spherical harmonics in the scalar case. In Sec. III A we derive, in analytic signal representation, a new time-dependent plane wave representation of the electromagnetic field generated by a time-varying, localized, volume charge-current distribution. The time-dependent plane wave spectra are given in terms of a fourfold Radon transform representation of the transverse current distribution. Later, we derive, following lines analogous to those of the scalar field treatment in Sec. II, a new form of time-dependent multipole expansion of the electromagnetic field; the spherical-wave expansion functions used are the vector analogs of the real spherical harmonics. However, unlike in Sec. II, our electromagnetic field case manipulations in Sec. III are carried out in the time domain. Two alternative expressions for the time-dependent multipole moments (i.e., the expansion coefficients) are derived; the first gives them in terms of the spectral amplitude vectors of the corresponding time-dependent plane wave representation while the second gives them in terms of a weighted radial-temporal average of the current distribution. Emphasis is placed on real spherical harmonics, hence our expansion involves only real time-dependent multipole wave functions and moments.

Notation: We define the temporal and spatial-temporal Fourier transforms, $q(\mathbf{r},\omega)$ and $\tilde{q}(\mathbf{K},\omega)$, respectively, of a scalar source $Q(\mathbf{r},t)$ of finite energy [i.e., square-integrable in $(\mathbf{r},t)$] via

$$q(\mathbf{r},\omega)=\int_{-\infty}^{\infty} dt' e^{i\omega t'} Q(\mathbf{r},t') \tag{1}$$

where

$$Q(\mathbf{r},t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega' e^{-i\omega' t}q(\mathbf{r},\omega'), \tag{2}$$

and

$$\tilde{q}(\mathbf{K},\omega)=\int_{-\infty}^{\infty}dt' e^{i\omega t'}\int d\mathbf{r}' e^{-i\mathbf{K}\cdot\mathbf{r}'}Q(\mathbf{r}',t') \tag{3}$$

where

$$Q(\mathbf{r},t)=\frac{1}{(2\pi)^4}\int_{-\infty}^{\infty}d\omega' e^{-i\omega' t}\int d\mathbf{K}' e^{i\mathbf{K}'\cdot\mathbf{r}}\tilde{q}(\mathbf{K}',\omega'). \tag{4}$$

Thus the spatial-temporal Fourier transform $\tilde{q}(\mathbf{K},\omega)$ of $Q(\mathbf{r},t)$ is simply the spatial Fourier transform of $q(\mathbf{r},\omega)$. We shall refer to $q(\mathbf{r},\omega)$ as being the source spectrum.

Unless otherwise stated, we shall subsequently use lowercase and uppercase letters to denote frequency and time domain quantities, respectively. Unless otherwise stated, a tilde ($\tilde{}$) over a constituent will denote a spatial-frequency domain constituent. A caret ($\hat{}$) over a vector will denote a unit vector, e.g., $\hat{\mathbf{r}}\equiv\mathbf{r}/|\mathbf{r}|$.

We shall denote the Cartesian coordinates of $\mathbf{r}$ (configuration space) and $\mathbf{K}$ (momentum space) as (x,y,z) and (K_x,K_y,K_z), respectively. The corresponding spherical coordinates will be denoted as (r,θ,ϕ) and (K,α,β), respectively. We also define the $\mathbf{K}$-space unit vector $\mathbf{s}\equiv\mathbf{K}/K$.

In order to simplify the manipulations time-dependent quantities will be manipulated first in analytic signal representation. For example, we define the analytic source distribution $Q^+(\mathbf{r},t)$ corresponding to a real, square-integrable source distribution $Q(\mathbf{r},t)$ by

$$Q^+(\mathbf{r},t)=\frac{1}{2\pi}\int_{0}^{\infty}d\omega' e^{-i\omega' t}q(\mathbf{r},\omega') \tag{5}$$

so that, for real t,

$$Q(\mathbf{r},t)=2\Re\{Q^+(\mathbf{r},t)\}, \tag{6}$$

where $\Re$ denotes the real part. The quantity $Q^+(\mathbf{r},t)$ defined by Eq. (5) is an analytic function in the lower-half plane of the complex t plane.[13] Of particular interest to us is the real-t limit of $Q^+(\mathbf{r},t)$, which defines the real, physical source distribution $Q(\mathbf{r},t)$ via Eq. (6). Subsequently a plus sign $(+)$ in the upper-right of a constituent will denote its analytic extension.

To simplify notation we also define a temporal integration operator ∂_t^{-1} such that

$$\partial_t^{-1}Q(\mathbf{r},t)\equiv\int_{-\infty}^{t}dt' Q(\mathbf{r},t'). \tag{7}$$

The corresponding nth-order temporal integration will be denoted as ∂_t^{-n}.

II. REVIEW OF THE THEORY FOR SCALAR FIELDS

This section reviews time-dependent plane wave and multipole expansions for scalar fields $U(\mathbf{r},t)$ radiated by scalar sources $Q(\mathbf{r},t)$ to the inhomogeneous wave equation

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)U(\mathbf{r},t)=-4\pi Q(\mathbf{r},t). \tag{8}$$

We will assume that the source $Q(\mathbf{r},t)$ is a square-integrable, continuous and continuously differentiable function of $(\mathbf{r},t)$, is confined to within a spherical volume about the origin [such that $Q(\mathbf{r},t)=0$ for $r>a$], and is turned on at some initial time, say $t=0$ [such that $Q(\mathbf{r},t)=0$ for $t<0$]. The radiated field $U(\mathbf{r},t)$ is the causal solution to Eq. (8), given by[14]

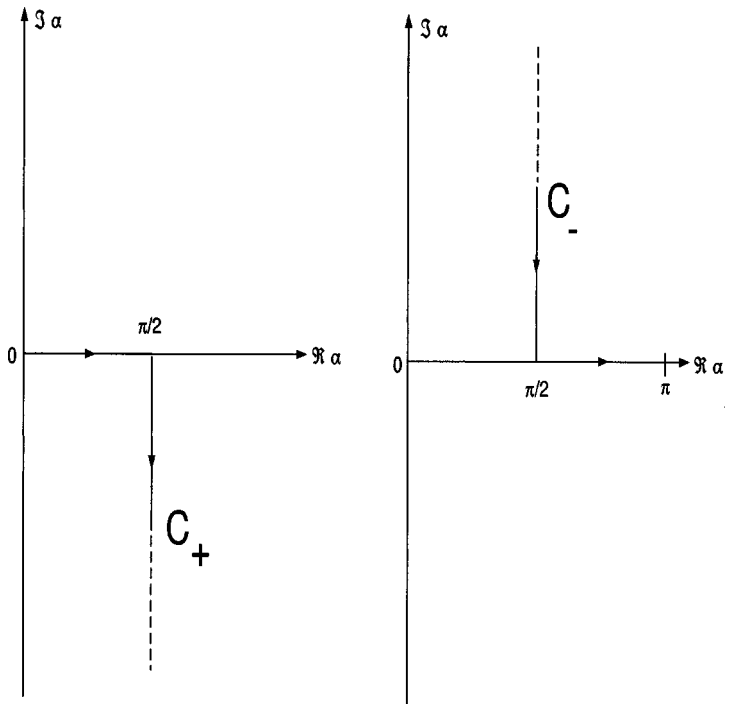

FIG. 1. The α-contours of integration C_+ and C_- in complex α space.

$$U(\mathbf{r},t)=\int_0^\infty dt' \int_{r'\leqslant a} d\mathbf{r}' \frac{Q(\mathbf{r}',t')\,\delta\left(t'-t+\dfrac{|\mathbf{r}-\mathbf{r}'|}{c}\right)}{|\mathbf{r}-\mathbf{r}'|}, \tag{9}$$

where $\delta(\cdot)$ is Dirac's delta function.

A. Plane wave expansion of scalar fields in the time domain

It is well known that the radiated field $U(\mathbf{r},t)$ at points $\mathbf{r}$ within any half-space $|z|\geqslant a$ can be written as a superposition of monochromatic homogeneous and inhomogeneous (evanescent) plane waves; in particular,[15]

$$U(\mathbf{r},t)=2\Re\{U^+(\mathbf{r},t)\}=2\Re\left\{\frac{ic}{(2\pi)^2}\int_0^\infty \frac{\omega}{c}\,d\left(\frac{\omega}{c}\right)\int_{-\pi}^{\pi} d\beta \int_{C_\pm} d\alpha\,\sin\alpha f(\mathbf{s},\omega)e^{[i(\omega/c)(\mathbf{s}\cdot\mathbf{r}-ci)]}\right\} \tag{10}$$

where $\mathbf{s}$ is the unit vector having Cartesian components (in general complex) $(\sin\alpha\cos\beta,\sin\alpha\sin\beta,\cos\alpha)$ while the contours of integration $C_\pm$ in the complex α plane are those shown in Fig. 1, where C_+ is used if $z\geqslant a$ and C_- if $z\leqslant -a$. The real-axis portion of the contours $C_\pm$ contributes to the propagating spectrum, describing homogeneous plane waves propagating into the half-space $z\geqslant a$ or $z\leqslant -a$. The complex portion of the contours $C_\pm$ corresponds, on the other hand, to the evanescent spectrum, describing evanescent plane waves that decay exponentially in amplitude along the z axis (i.e., in $\hat{\mathbf{z}}$ if $z\geqslant a$ and in $-\hat{\mathbf{z}}$ if $z\leqslant -a$). The spectral amplitude function $f(\mathbf{s},\omega)$ in (10) is related to the source spectrum $q(\mathbf{r},\omega)$ by[15]

$$f(\mathbf{s},\omega)=\tilde{q}(\mathbf{K},\omega)|_{\mathbf{K}=(\omega/c)\mathbf{s}}=\int_{r'\leqslant a} d\mathbf{r}'\,e^{-i(\omega/c)\mathbf{s}\cdot\mathbf{r}'}q(\mathbf{r}',\omega). \tag{11}$$

It is noted that since we require $q(\mathbf{r},\omega)$ to be continuous and compactly supported in space, it follows that $\tilde{q}(\mathbf{K},\omega)$ is, for fixed frequency ω, the boundary value, on the real K_x,K_y,K_z axes, of an entire analytic function of three complex variables K_x,K_y,K_z. Thus, for fixed frequency ω,

$f(\mathbf{s},\omega)$ is given by the value of the spatial Fourier transform $\tilde{q}(\mathbf{K},\omega)$, and its analytic continuation into the complex $\mathbf{K}$ plane, of the source spectrum $q(\mathbf{r},\omega)$ evaluated at $\mathbf{K}=(\omega/c)\mathbf{s}$.

By performing the ω integration in (10) and using (5) we obtain

$$U^+(\mathbf{r},t)=-\frac{1}{2\pi c}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{C_\pm}d\alpha\,\sin\,\alpha F^+(\mathbf{s},t-\mathbf{s}\cdot\mathbf{r}/c), \tag{12}$$

where from Eq. (11)

$$F^+(\mathbf{s},t)\equiv\frac{1}{2\pi}\int_0^\infty d\omega' e^{-i\omega' t}f(\mathbf{s},\omega')=\int_{r'\leqslant a}d\mathbf{r}' Q^+(\mathbf{r}',t+\mathbf{s}\cdot\mathbf{r}'/c). \tag{13}$$

Equations (12) and (13) define the time-dependent plane wave representation of the analytic signal $U^+(\mathbf{r},t)$ corresponding to the real, physical field $U(\mathbf{r},t)$.

Equation (13) is the time domain analog of (11). The operation defined by the integral on the right-hand side of Eq. (13), sometimes referred to as ''slant-stack transform'' of the analytic source distribution $Q^+(\mathbf{r},t)$, has been studied in recent years by Heyman.[16] It involves spatial Radon projection of $Q^+(\mathbf{r},t)$ evaluated at planes normal to the spectral propagation direction $\mathbf{s}$ and stacking them with a progressive time delay $\mathbf{s}\cdot\mathbf{r}/c$ associated with a wave front propagating at the speed of light along this direction. Mathematically, $F^+(\mathbf{s},t)$ is given by the value of the fourfold Radon transform, and its analytic continuation for complex $\mathbf{s}$,[17] of the analytic source distribution evaluated at the hyperplane $ct-\mathbf{s}\cdot\mathbf{r}=0$ in a four-dimensional space-time domain $(ct,\mathbf{r})$.

The real, physical field $U(\mathbf{r},t)$ is recovered from $U^+(\mathbf{r},t)$ via Eqs. (6) and (5). In the far radiation zone[14]

$$U(r\hat{\mathbf{r}},t)\big|_{r\to\infty}\sim\frac{1}{r}F(\mathbf{s},\tau)\big|_{\mathbf{s}=\hat{\mathbf{r}}} \tag{14}$$

with $\tau=t-r/c$. Thus the time domain radiation pattern is from Eq. (14), the propagating portion of the time-dependent plane-wave spectrum. The evanescent spectrum can be recovered by analytic continuation of the time domain radiation pattern to complex-valued observation directions.[17] The individual contributions of the propagating and evanescent spectra to the total field have been examined in detail by Heyman,[16] who has shown that the contribution from the evanescent spectrum vanishes after a certain time t_E that may be determined *a priori* for a given field point; in regions of space where t_E occurs prior to the causal arrival time of the field, the latter is describable only by its propagating spectrum within the causal time window while noncausal contributions due to both propagating and evanescent spectra cancel out prior to the causal arrival time.

B. Multipole theory in the frequency domain

The multipole expansion of the radiated field spectrum $u(\mathbf{r},\omega)$ is readily obtained by expanding the spectral amplitudes $f(\mathbf{s},\omega)$ of the (frequency domain) angular spectrum expansion

$$u(\mathbf{r},\omega)=\frac{i\omega/c}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_\pm}d\alpha\,\sin\,\alpha f(\mathbf{s},\omega)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}} \tag{15}$$

into a series of spherical harmonics $Y_{\ell,m}(\alpha,\beta)$ (see Appendix A for a review) so that

$$f(\mathbf{s},\omega)=\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^\ell\tilde{a}_{\ell,m}(\omega)Y_{\ell,m}(\alpha,\beta) \tag{16}$$

where

$$\tilde{a}_{\ell,m}(\omega)=i^\ell\int_{-\pi}^{\pi}d\beta\int_0^\pi d\alpha\,\sin\,\alpha f(\mathbf{s},\omega)Y^*_{\ell,m}(\alpha,\beta). \tag{17}$$

By substituting from Eqs. (16) and (17) into Eq. (15) while interchanging orders of integration and summation one readily finds

$$u(\mathbf{r},\omega)=\frac{i\omega}{c}\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}\tilde{a}_{\ell,m}(\omega)\Pi_{\ell,m}(\mathbf{r},\omega), \tag{18}$$

where

$$\Pi_{\ell,m}(\mathbf{r},\omega)=(-i)^{\ell}\frac{1}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha Y_{\ell,m}(\alpha,\beta)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}} \tag{19}$$

is precisely the angular spectrum representation of the scalar multipole field of degree ℓ and order m, i.e.,[4]

$$\Pi_{\ell,m}(\mathbf{r},\omega)=h_{\ell}^{(1)}\left(\frac{\omega}{c}r\right)Y_{\ell,m}(\theta,\phi). \tag{20}$$

In Eq. (20), $h_{\ell}^{(1)}(\cdot)$ is the spherical Hankel function of the first kind of order ℓ (as defined in Ref. 14). Even though Eqs. (18) and (20) were derived for $\omega>0$, they apply to all real ω. This fact is implicit in our time-dependent multipole treatment (Sec. II C). For example, the operator $\mathscr{L}_{\ell}^{(r,t)}$ in Eq. (24) applies to both analytic and real signals.

The multipole expansion Eq. (18) is valid for $r\geqslant a$ while the plane wave expansion Eq. (15) is valid for $|z|\geqslant a$. However, it is obvious that as long as $\mathbf{r}$ is outside the source, one can always decompose the radiated field spectrum $u(\mathbf{r},\omega)$ in the form Eq. (15) by a suitable choice of coordinates.[16]

An alternative expression for the multipole moments $\tilde{a}_{\ell,m}(\omega)$ is obtained by substituting (11) on (17) and interchanging orders of integration so as to yield[4]

$$\tilde{a}_{\ell,m}(\omega)=4\pi\int_{r'\leqslant a}d\mathbf{r}'q(\mathbf{r}',\omega)\Lambda_{\ell,m}^{*}(\mathbf{r}',\omega) \tag{21}$$

with

$$\Lambda_{\ell,m}(\mathbf{r},\omega)\equiv j_{\ell}\left(\frac{\omega}{c}r\right)Y_{\ell,m}(\theta,\phi)=(-i)^{\ell}\frac{1}{4\pi}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\,\sin\alpha Y_{\ell,m}(\alpha,\beta)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}, \tag{22}$$

where $j_{\ell}(\cdot)$ is the spherical Bessel function of the first kind and order ℓ (as defined in Ref. 4).

The multipole expansion enjoys several advantages over the angular spectrum expansion. First of all, we see from Eq. (17) that the multipole moments $\tilde{a}_{\ell,m}(\omega)$ are defined only by the real $\mathbf{s}$ (propagating) spectral amplitudes $f(\mathbf{s},\omega)$. Thus, in contrast to the angular spectrum expansion—which requires analytic continuation—the multipole expansion requires no analytic continuation of $f(\mathbf{s},\omega)$ for complex values of the polar angle α. In addition, unlike the plane waves $e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}$, the multipole fields $\Pi_{\ell,m}(\mathbf{r},\omega)$ obey Sommerfeld's radiation condition. Hence, the multipole expansion is, unlike the angular spectrum expansion, a true mode expansion of the radiation field spectrum. This can be of interest if one wishes to approximate the radiated field either by truncating the series (18) or by limiting the (α,β) region of integration in (15). In the former case the approximation will obey Sommerfeld's radiation condition while in the latter it will not. These advantages are obtained, however, at the expense of the increased complexity of the spherical wave fields $\Pi_{\ell,m}(\mathbf{r},\omega)$ as compared to the plane waves $e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}$.

C. Time-dependent multipole expansion of the scalar field

Next we obtain a time domain multipole expansion for the scalar field $U(\mathbf{r},t)$ following the lines of Ref. 3. Our starting point is the result [see Ref. 18, Eq. (10.1.16)]

$$\frac{1}{2\pi}\int_0^\infty d\omega e^{-i\omega t}h_\ell^{(1)}\left(\frac{\omega}{c}r\right)\tilde{a}_{\ell,m}(\omega)=-(-i)^\ell\frac{c}{r}\partial_t^{-1}\mathscr{L}_\ell^{(r,t)}a_{\ell,m}^+(t), \tag{23}$$

where

$$a_{\ell,m}^+(t)=\frac{1}{2\pi}\int_0^\infty d\omega e^{-i\omega t}\tilde{a}_{\ell,m}(\omega)$$

and

$$\mathscr{L}_\ell^{(r,t)}a_{\ell,m}^+(t)=\sum_{n=0}^{\ell}\frac{(\ell+n)!}{n!(\ell-n)!}\left(\frac{2r}{c}\right)^{-n}\partial_t^{-n}a_{\ell,m}^+(\tau). \tag{24}$$

In Eqs. (23) and (24) $\tau=t-r/c$ (as before) and ∂_t^{-n} denotes nth-order temporal integration [recall (7)].

By Fourier inverting Eq. (18) over positive frequencies while using Eqs. (20) and (23) we obtain

$$U^+(\mathbf{r},t)=\frac{1}{r}\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^\ell Y_{\ell,m}(\theta,\phi)\mathscr{L}_\ell^{(r,t)}a_{\ell,m}^+(t). \tag{25}$$

The analytic time-dependent multipole moments $a_{\ell,m}^+(t)$ are obtained by Fourier inverting over positive frequencies both sides of Eq. (17), thus giving

$$a_{\ell,m}^+(t)=i^\ell\int_{-\pi}^{\pi}d\beta\int_0^\pi d\alpha\,\sin\alpha F^+(\mathbf{s},t)Y_{\ell,m}^*(\alpha,\beta). \tag{26}$$

Thus, for fixed t, the analytic time-dependent multipole moments $a_{\ell,m}^+(t)$ are, apart from the factor i^ℓ, given by projections of the analytic time-dependent plane wave spectrum $F^+(\mathbf{s},t)$ [see Eq. (13)] onto the set of spherical harmonics $Y_{\ell,m}(\alpha,\beta)$. Note, however, that—as expected from our discussion in Eq. (17)—only the real-$\mathbf{s}$, propagating plane wave spectral components enter into play in Eq. (26).

An alternative form of Eq. (25) due to Hansen and Norris[19] is obtained using (see Ref. 19 and references therein)

$$h_\ell^{(1)}\left(\frac{\omega}{c}r\right)Y_{\ell,m}(\theta,\phi)=i^{-\ell}\tilde{P}_{\ell m}^{\omega}\frac{e^{i(\omega/c)r}}{i(\omega/c)r}, \tag{27}$$

where

$$\tilde{P}_{\ell m}^{\omega}=(-1)^m\sqrt{\frac{(2\ell+1)(\ell-m)!}{4\pi(\ell+m)!}}\left(\frac{1}{i(\omega/c)}\frac{\partial}{\partial x}+i\frac{1}{i\dfrac{\omega}{c}}\frac{\partial}{\partial y}\right)^m P_\ell^{(m)}\left(\frac{1}{i\dfrac{\omega}{c}}\frac{\partial}{\partial z}\right),\quad m\geqslant 0, \tag{28}$$

where $P_\ell^{(m)}(\cdot)$ is the mth derivative of the Legendre polynomial and for $m<0$, $\tilde{P}_{\ell,-m}^{\omega}=(-1)^m\tilde{P}_{\ell m}^{\omega *}$. On substituting from Eqs. (27) and (28) into Eqs. (18) and (20) while using the complex-point source field expansion of the real-point source field [see Eq. (16) in Ref. 19]

$$\frac{e^{i(\omega/c)r}}{4\pi r}=\frac{1}{4\pi j_0(i(\omega/c)\rho)}\int_{-\pi}^{\pi}d\phi'\int_0^\pi d\theta'\,\sin\theta'\frac{e^{i(\omega/c)\xi(\theta',\phi')}}{4\pi\xi(\theta',\phi')}, \tag{29}$$

where $\xi=|\mathbf{r}-i\rho\hat{\mathbf{r}}'|=\xi_r+i\xi_i$ is the complex length having real and imaginary parts ξ_r and ξ_i, respectively, where $\rho\leqslant a$ is a positive constant (such that $\xi_r\geqslant 0$ and $-\rho\leqslant\xi_i\leqslant\rho$), and transforming to the time domain one obtains

$$U^{+}(\mathbf{r},t)=\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^{\ell}\int_{-\pi}^{\pi}d\phi'\int_{0}^{\pi}d\theta'\ \sin\ \theta'\widetilde{L_0}\widetilde{P}^{t}_{\ell m}\ \frac{a^{+}_{\ell,m}(t-\xi(\theta',\phi')/c)}{4\pi\xi(\theta',\phi')},\qquad(30)$$

where $\widetilde{L_0}$ is a time domain operator whose frequency domain representation is $1/j_0((i\omega/c)\rho)$ [see Eq. (37) in Ref. 19] while $\widetilde{P}^{t}_{\ell m}$ is the time domain analog of $\widetilde{P}^{\omega}_{\ell m}$. Equation (30) is the result Eq. (59) in Ref. 19; it is an analytic field expansion in terms of time-dependent multipoles with complex-source locations. Unlike in Eq. (25), in Eq. (30) the basis functions are directionally localized; hence they are of interest in well-collimated wave radiation and propagation (see, e.g., Heyman and Felsen[1]). However, the real-signal counterpart of the analytic complex-source multipole expansion Eq. (30) is complicated and involves convolutions; in contrast, the real-signal version of Eq. (25) is straightforward and can be written in terms of the operator $\mathscr{L}^{(r,t)}_{\ell}$. This is done below.

We use (25) to recover the real, physical field $U(\mathbf{r},t)=2\Re\{U^{+}(\mathbf{r},t)\}$. In order to evaluate the real part of expression (25) we consider the real spherical harmonics $S^{(j)}_{\ell,m}$, with $\ell=0,1,...,$ $m=0,1,...,\ell$ (see Appendix A). They obey the following orthogonality and completeness properties:[3]

$$\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\ \alpha S^{(j)}_{\ell,m}(\alpha,\beta)S^{(j')}_{\ell',m'}(\alpha,\beta)=\delta_{\ell,\ell'}\delta_{m,m'}\delta_{j,j'}\ \frac{1}{c_m},\qquad(31)$$

where $c_m=1$ or 2 for $m=0$ or $m\geqslant 1$, respectively, and

$$\sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m S^{(j)}_{\ell,m}(\alpha,\beta)S^{(j)}_{\ell,m}(\alpha',\beta')=\frac{1}{\sin\ \alpha}\ \delta(\beta-\beta')\delta(\alpha-\alpha')\equiv\delta(\mathbf{s}-\mathbf{s}').\qquad(32)$$

[In Eq. (32), α' and β' are, respectively, the polar and azimuthal coordinates of the unit vector $\mathbf{s}'$.]

By taking the real part to both sides of Eq. (25) while using Eqs. (26), (32), and (31) one obtains (see derivation in Appendix B)

$$U(\mathbf{r},t)=\frac{1}{r}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m S^{(j)}_{\ell,m}(\theta,\phi)\mathscr{L}^{(r,t)}_{\ell}q^{(j)}_{\ell,m}(t),\qquad(33)$$

where we have introduced the new (real) multipole moments $q^{(j)}_{\ell,m}(t)$, defined via

$$q^{(j)}_{\ell,m}(t)=\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\ \alpha F(\mathbf{s},t)S^{(j)}_{\ell,m}(\alpha,\beta).\qquad(34)$$

On comparing Eqs. (14) and (33) with $\mathscr{L}^{(r,t)}_{\ell}$ given by Eq. (24), the far field is seen to correspond to the lowest order term (i.e., $n=0$) in the series expansion Eq. (24) corresponding to $\mathscr{L}^{(r,t)}_{\ell}$. Thus from Eq. (24) $\mathscr{L}^{(r,t)}_{\ell}q^{(j)}_{\ell,m}(t)|_{r\to\infty}\sim q^{(j)}_{\ell,m}(\tau)$. Using this result in Eq. (33) yields

$$U(\mathbf{r},t)|_{r\to\infty}\sim\frac{1}{r}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m S^{(j)}_{\ell,m}(\theta,\phi)q^{(j)}_{\ell,m}(\tau).\qquad(35)$$

Also, from (14) and (35), the time domain radiation pattern reduces to

$$F(\hat{\mathbf{r}},t)=\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m S^{(j)}_{\ell,m}(\theta,\phi)q^{(j)}_{\ell,m}(t).\qquad(36)$$

Since the multipole moments $q^{(j)}_{\ell,m}(t)$ define via Eq. (33) the field everywhere outside the source, it follows from Eq. (34) that the time domain radiation pattern $F(\hat{\mathbf{r}},t)$ also determines the field everywhere outside the source (problem of inverse diffraction from far field data[3,4]).

Equation (34) defines the real time-dependent multipole moments $q^{(j)}_{\ell,m}(t)$ as projections of the time-dependent plane wave spectrum onto the set of real spherical harmonics. An alternative

formula for $q^{(j)}_{\ell,m}(t)$ based on a radially dependent weighted time average of $Q(\mathbf{r},t)$ follows readily from the time domain analog of Eq. (21). The remainder of this section is based on Eq. (21) and the following result:[3]

$$j_\ell(z)=\frac{(z/2)^\ell}{2\ell!}\int_{-1}^{1}d\eta(1-\eta^2)^\ell e^{iz\eta}. \tag{37}$$

We begin by substituting from Eq. (37) into Eq. (21), thus obtaining

$$\tilde{a}_{\ell,m}(\omega)=\frac{4\pi c^{-\ell}\omega^\ell}{\ell!2^{\ell+1}}\int_{r\leqslant a}d\mathbf{r}q(\mathbf{r},\omega)Y^*_{\ell,m}(\theta,\phi)r^\ell\int_{-1}^{1}d\eta(1-\eta^2)^\ell e^{-i\omega r\eta/c}. \tag{38}$$

The analytic time-dependent multipole moments $a^+_{\ell,m}(t)$ are obtained by Fourier inverting $\tilde{a}_{\ell,m}(\omega)$ as given by Eq. (38) over positive frequencies. This procedure yields

$$a^+_{\ell,m}(t)=\frac{1}{2\pi}\int_0^\infty d\omega e^{-i\omega t}\tilde{a}_{\ell,m}(\omega)=i^\ell\int_{-\pi}^{\pi}d\phi\int_0^\pi d\theta\,\sin\,\theta Y^*_{\ell,m}(\theta,\phi)F^+_\ell(\hat{\mathbf{r}},t), \tag{39}$$

where

$$F^+_\ell(\hat{\mathbf{r}},t)=\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\int_{r\leqslant a}drr^\ell\int_{-1}^{1}d\eta(1-\eta^2)^\ell\frac{\partial^\ell}{\partial t^\ell}Q^+(\mathbf{r},t+r\eta/c). \tag{40}$$

The real time-dependent multipole moments $q^{(j)}_{\ell,m}(t)$ follow from Eqs. (39) and (40) with $q^+_{\ell,m}(t)=(-i)^\ell a^+_{\ell,m}(t)$:

$$q^{(j)}_{\ell,m}(t)=\int_{-\pi}^{\pi}d\phi\int_0^\pi d\theta\,\sin\,\theta S^{(j)}_{\ell,m}(\theta,\phi)F_\ell(\hat{\mathbf{r}},t), \tag{41}$$

where

$$F_\ell(\hat{\mathbf{r}},t)=\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\int_{r\leqslant a}drr^\ell\int_{-1}^{1}d\eta(1-\eta^2)^\ell\frac{\partial^\ell}{\partial t^\ell}Q(\mathbf{r},t+r\eta/c). \tag{42}$$

The time-dependent real multipole moments $q^{(j)}_{\ell,m}(t)$ are now given from Eqs. (41) and (42) by the projection of an ℓ-dependent (effective) time domain radiation pattern $F_\ell(\hat{\mathbf{r}},t)$ onto the set of real spherical harmonics. This latter quantity is computed using a two-step procedure described below.

The first step is that of η integration in (42). It involves radially dependent weighted time averaging of the ℓth time derivative of the source distribution, $(\partial^\ell/\partial t^\ell)Q(\mathbf{r},t)$ (η integration plays the role of temporal integration). The term $(1-\eta^2)^\ell$ in Eq. (42) acts as weighting kernel. The η-integration limits in Eq. (42) define the time window $t\pm r/c$ within which weighted time averaging of $(\partial^\ell/\partial t^\ell)Q(\mathbf{r},t)$ is carried out. This first step acts on both the temporal and radial dependence of the source. The temporal nature of the source enters into play here through both time derivation and time averaging. On the other hand, both ℓ and r (and, therefore, also the source's radius a) control the effective time averaging window.

The second step is that of radial integration. It acts only on the radial dependence of the source. Thus it plays a role in controlling the relative relevance of the outer versus the inner source's structure in defining $q^{(j)}_{\ell,m}(t)$.

III. TIME-DEPENDENT PLANE WAVE AND MULTIPOLE EXPANSIONS OF ELECTROMAGNETIC FIELDS

In Gaussian systems of units, the time-dependent electric and magnetic field vectors $\mathscr{E}(\mathbf{r},t)$ and $\mathscr{H}(\mathbf{r},t)$, respectively, generated in free space by a time-dependent current distribution $\mathscr{J}(\mathbf{r},t)$, are the causal solutions to the vector wave equations[14]

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\mathscr{E}(\mathbf{r},t)=-4\pi\left[-\frac{1}{c^2}\frac{\partial}{\partial t}+\partial_t^{-1}\nabla\nabla\cdot\right]\mathscr{T}(\mathbf{r},t),$$
$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\mathscr{H}(\mathbf{r},t)=-4\pi\,\frac{1}{c}\,\nabla\times\mathscr{T}(\mathbf{r},t), \tag{43}$$

where, again, ∂_t^{-1} denotes integration over time [see Eq. (7)]. Here and henceforth it is assumed that $\mathscr{T}(\mathbf{r},t)$ is a square-integrable, continuous and continuously differentiable function of $(\mathbf{r},t)$, is confined to a finite region of radius a around the origin, and is turned on at some initial time $t=0$. In Secs. III A and III B we derive new time-dependent plane wave and multipole expansions for $\mathscr{E}(\mathbf{r},t)$ and $\mathscr{H}(\mathbf{r},t)$ in terms of $\mathscr{T}(\mathbf{r},t)$.

A. Plane wave expansion of electromagnetic fields in the time domain

It follows from (43) that each Cartesian component of the electromagnetic field vectors $\mathscr{E}(\mathbf{r},t)$, $\mathscr{H}(\mathbf{r},t)$ satisfies an inhomogeneous scalar wave equation of the form Eq. (8). Therefore, the results of Sec. II are applicable to each Cartesian component of $\mathscr{E}(\mathbf{r},t)$ and $\mathscr{H}(\mathbf{r},t)$. Thus, we introduce the analytic field vectors $\mathscr{E}^+(\mathbf{r},t)=(1/2\pi)\int_0^\infty d\omega e^{-i\omega t}\mathbf{E}(\mathbf{r},\omega)$, $\mathscr{H}^+(\mathbf{r},t)=(1/2\pi)\int_0^\infty d\omega e^{-i\omega t}\mathbf{H}(\mathbf{r},\omega)$, where $\mathbf{E}(\mathbf{r},\omega)$ and $\mathbf{H}(\mathbf{r},\omega)$ are, respectively, the electric and magnetic field spectrum. They are related to the analytic current distribution $\mathscr{T}^+(\mathbf{r},t)=(1/2\pi)\int_0^\infty d\omega e^{-i\omega t}\mathbf{J}(\mathbf{r},\omega)$ by Eq. (43) with $\mathscr{E}$, $\mathscr{H}$ and $\mathscr{T}$ substituted by $\mathscr{E}^+$, $\mathscr{H}^+$ and $\mathscr{T}^+$, respectively.

By applying to Eq. (43) the scalar case result Eq. (12) one obtains

$$\mathscr{E}^+(\mathbf{r},t)=-\frac{1}{2\pi c}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{C_\pm}d\alpha\,\sin\alpha\,\mathbf{F}_E^+(\mathbf{s},t-\mathbf{s}\cdot\mathbf{r}/c),$$
$$\mathscr{H}^+(\mathbf{r},t)=-\frac{1}{2\pi c}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{C_\pm}d\alpha\,\sin\alpha\,\mathbf{F}_H^+(\mathbf{s},t-\mathbf{s}\cdot\mathbf{r}/c), \tag{44}$$

where from Eq. (13)

$$\mathbf{F}_E^+(\mathbf{s},t)=\int_{r'\leqslant a}d\mathbf{r}'\left[-\frac{1}{c^2}\frac{\partial}{\partial t}+\partial_t^{-1}\nabla'\nabla'\cdot\right]\mathscr{T}^+(\mathbf{r}',t+\mathbf{s}\cdot\mathbf{r}'/c),$$
$$\mathbf{F}_H^+(\mathbf{s},t)=\frac{1}{c}\int_{r'\leqslant a}d\mathbf{r}'\nabla'\times\mathscr{T}^+(\mathbf{r}',t+\mathbf{s}\cdot\mathbf{r}'/c), \tag{45}$$

with ∇' operating on $\mathbf{r}'$ (∇' does not act on $t+\mathbf{s}\cdot\mathbf{r}'/c$). The results Eqs. (44) and (45) above define a new time-dependent plane wave representation for electromagnetic fields whose interpretation is analogous to that of its scalar counterpart Eqs. (12) and (13).

Next we would like to express Eq. (45) more compactly in terms of the slant-stack transform

$$\hat{\mathscr{T}}^+(\mathbf{s},t)=\int_{r'\leqslant a}d\mathbf{r}'\mathscr{T}^+(\mathbf{r}',t+\mathbf{s}\cdot\mathbf{r}'/c) \tag{46}$$

of the analytic time-dependent current distribution $\mathscr{T}^+$. By manipulating the integrals of Eq. (45) while using (46) we obtain (see derivation in Appendix C)

$$\mathbf{F}_E^+(\mathbf{s},t)=\frac{1}{c^2}\frac{\partial}{\partial t}\{\mathbf{s}\times[\mathbf{s}\times\hat{\mathscr{T}}^+(\mathbf{s},t)]\},$$
$$\mathbf{F}_H^+(\mathbf{s},t)=-\frac{1}{c^2}\frac{\partial}{\partial t}[\mathbf{s}\times\hat{\mathscr{T}}^+(\mathbf{s},t)]. \tag{47}$$

We see from Eq. (47) that $\mathbf{F}_E^+(\mathbf{s},t)$ and $\mathbf{F}_H^+(\mathbf{s},t)$ are defined by the transverse part (relative to $\mathbf{s}$)—and its analytic continuation to complex $\mathbf{s}$—of $\hat{\mathscr{T}}^+(\mathbf{s},t)$. We see also from Eq. (47) that

$$\mathbf{s}\cdot\mathbf{F}_E^+(\mathbf{s},t)=\mathbf{s}\cdot\mathbf{F}_H^+(\mathbf{s},t)=0,$$
$$\mathbf{F}_E^+(\mathbf{s},t)=-\mathbf{s}\times\mathbf{F}_H^+(\mathbf{s},t). \tag{48}$$

Hence, the terms in the integrands of (44) represent—for each $\mathbf{s}$—electromagnetic plane waves, as expected.

We conclude this section by applying the scalar case result Eq. (14) to each Cartesian component of the field vectors, thus obtaining from (47) the far zone approximations

$$\mathscr{E}(r\hat{\mathbf{r}},t)|_{r\to\infty}\sim\frac{1}{r}\mathbf{F}_E(\hat{\mathbf{r}},\tau)|_{r\to\infty}=\frac{1}{c^2 r}\frac{\partial}{\partial t}\{\hat{\mathbf{r}}\times[\hat{\mathbf{r}}\times\hat{\mathscr{T}}(\hat{\mathbf{r}},\tau)]\},$$
$$\mathscr{H}(r\hat{\mathbf{r}},t)|_{r\to\infty}\sim\frac{1}{r}\mathbf{F}_H(\hat{\mathbf{r}},\tau)|_{r\to\infty}=-\frac{1}{c^2 r}\frac{\partial}{\partial t}[\hat{\mathbf{r}}\times\hat{\mathscr{T}}(\hat{\mathbf{r}},\tau)] \tag{49}$$

where, as before, $\tau=t-r/c$. Expressions (49) have been derived before in connection with questions of realizability of time domain radiation patterns[17] and in time domain antenna characterization.[20]

B. Time-dependent multipole expansion of the electromagnetic field

Next we will obtain an expansion of the analytic time-dependent electric and magnetic field vectors $\mathscr{E}^+(\mathbf{r},t)$ and $\mathscr{H}^+(\mathbf{r},t)$ in terms of vector spherical wave functions. Methodologically, first we will expand the analytic time-dependent plane wave spectra [$\mathbf{F}_E^+(\mathbf{s},t)$ and $\mathbf{F}_H^+(\mathbf{s},t)$] into a series of the two sets of vector functions $\mathbf{Y}_{\ell,m}(\alpha,\beta)$ and $\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)$; later we will substitute the vector spherical harmonics expansions for $\mathbf{F}_E^+(\mathbf{s},t)$ and $\mathbf{F}_H^+(\mathbf{s},t)$ into the time-dependent plane wave expansions—Eq. (44). The first step above yields

$$\mathbf{F}_E^+(\mathbf{s},t)=\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^{\ell}[a_{\ell,m}^+(t)\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)+b_{\ell,m}^+(t)\mathbf{Y}_{\ell,m}(\alpha,\beta)],$$
$$\mathbf{F}_{\mathbf{H}}^+(\mathbf{s},t)=\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^{\ell}[-a_{\ell,m}^+(t)\mathbf{Y}_{\ell,m}(\alpha,\beta)+b_{\ell,m}^+(t)\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)]. \tag{50}$$

It is noted that the forms of the expansions above are consistent with Eq. (48).

By making use of the orthogonality condition of the vector spherical harmonics, Eq. (A8), we obtain from Eq. (50)

$$a_{\ell,m}^+(t)=-\frac{i^{\ell}}{\ell(\ell+1)}\int_{-\pi}^{\pi}d\beta\int_0^{\pi}d\alpha\,\sin\alpha\mathbf{F}_H^+(\mathbf{s},t)\cdot\mathbf{Y}_{\ell,m}^*(\alpha,\beta),$$
$$b_{\ell,m}^+(t)=\frac{i^{\ell}}{\ell(\ell+1)}\int_{-\pi}^{\pi}d\beta\int_0^{\pi}d\alpha\,\sin\alpha\mathbf{F}_{\mathbf{E}}^+(\mathbf{s},t)\cdot\mathbf{Y}_{\ell,m}^*(\alpha,\beta). \tag{51}$$

Thus the analytic time-dependent multipole moments $a_{\ell,m}^+(t)$ and $b_{\ell,m}^+(t)$ are defined by the projections of $\mathbf{F}_H^+(\mathbf{s},t)$ and $\mathbf{F}_E^+(\mathbf{s},t)$, respectively, onto the set of vector spherical harmonics.

The time-dependent multipoles can also be expressed directly in terms of the slant-stack transform of the current distribution by substituting from Eq. (47) into Eq. (51), thus giving

$$a^{+}_{\ell,m}(t)=\frac{i^{\ell}c^{-2}}{\ell(\ell+1)}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha[\mathbf{s}\times\hat{\mathscr{T}}^{+}(\mathbf{s},t)]\cdot\mathbf{Y}^{*}_{\ell,m}(\alpha,\beta),$$
$$b^{+}_{\ell,m}(t)=\frac{i^{\ell}c^{-2}}{\ell(\ell+1)}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\{\mathbf{s}\times[\mathbf{s}\times\hat{\mathscr{T}}^{+}(\mathbf{s},t)]\}\cdot\mathbf{Y}^{*}_{\ell,m}(\alpha,\beta). \tag{52}$$

Next we insert the series expansions Eq. (50) into the time-dependent plane wave expansions of the field vectors $\mathscr{E}(\mathbf{r},t)$ and $\mathscr{H}(\mathbf{r},t)$ [refer to Eq. (44)] and interchange orders of summation and integration, thus finding

$$\mathscr{E}^{+}(\mathbf{r},t)=-\frac{1}{2\pi c}\frac{\partial}{\partial t}\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha(-i)^{\ell}[a^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)$$
$$+b^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{Y}_{\ell,m}(\alpha,\beta)],$$
$$\mathscr{H}^{+}(\mathbf{r},t)=-\frac{1}{2\pi c}\frac{\partial}{\partial t}\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha(-i)^{\ell}[-a^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{Y}_{\ell,m}(\alpha,\beta)$$
$$+b^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)]. \tag{53}$$

Next we shall solve the integrals over $\mathbf{s}$ above so as to obtain the sought-after time-dependent multipole expansion of the electromagnetic field. For this purpose we will use the vector counterpart of Eqs. (19) and (20), given by[4]

$$\frac{(-i)^{\ell}}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha\mathbf{Y}_{\ell,m}(\alpha,\beta)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}=h^{(1)}_{\ell}\left(\frac{\omega}{c}r\right)\mathbf{Y}_{\ell,m}(\theta,\phi). \tag{54}$$

Furthermore, by applying the curl operator $\nabla\times$ to both sides of (54) while using $\nabla e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}=i(\omega/c)\mathbf{s}e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}$ and $\nabla\times\mathbf{Y}_{\ell,m}(\alpha,\beta)=0$, one readily obtains[4]

$$\frac{(-i)^{\ell}}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}=\frac{-ic}{\omega}\nabla\times\left[h^{(1)}_{\ell}\left(\frac{\omega}{c}r\right)\mathbf{Y}_{\ell,m}(\theta,\phi)\right]. \tag{55}$$

The following results follow readily from Eqs. (54) and (55) and Eqs. (23) and (24):

$$\frac{(-i)^{\ell}}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha a^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{Y}_{\ell,m}(\alpha,\beta)$$
$$=\frac{1}{2\pi}\int_{0}^{\infty}d\omega e^{-i\omega t}\tilde{a}_{\ell,m}(\omega)h^{(1)}_{\ell}\left(\frac{\omega}{c}r\right)\mathbf{Y}_{\ell,m}(\theta,\phi)$$
$$=-(-i)^{\ell}\frac{c}{r}\mathbf{Y}_{\ell,m}(\theta,\phi)\partial_t^{-1}\mathscr{L}^{(r,t)}_{\ell}a^{+}_{\ell,m}(t) \tag{56}$$

and, similarly,

$$\frac{(-i)^{\ell}}{2\pi}\int_{-\pi}^{\pi}d\beta\int_{C_{\pm}}d\alpha\ \sin\alpha a^{+}_{\ell,m}(t-\mathbf{s}\cdot\mathbf{r}/c)\mathbf{s}\times\mathbf{Y}_{\ell,m}(\alpha,\beta)=\frac{1}{2\pi}\int_{0}^{\infty}d\omega e^{-i\omega t}\tilde{a}_{\ell,m}(\omega)\left\{\frac{-ic}{\omega}\right.$$
$$\left.\times\nabla\left[h^{(1)}_{\ell}\left(\frac{\omega}{c}r\right)\mathbf{Y}_{\ell,m}(\theta,\phi)\right]\right\}=(-i)^{\ell}\frac{c^2}{r}\partial_t^{-2}\nabla\times[\mathbf{Y}_{\ell,m}(\theta,\phi)\mathscr{L}^{(r,t)}_{\ell}a^{+}_{\ell,m}(t)]. \tag{57}$$

By substituting from Eqs. (56) and (57) into Eq. (53) we obtain

$$\mathscr{E}^{+}(\mathbf{r},t)=\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}\frac{1}{r}(-i)^{\ell}\{-c\partial_t^{-1}\nabla\times[\mathbf{Y}_{\ell,m}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}a_{\ell,m}^{+}(t)]+\mathbf{Y}_{\ell,m}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}b_{\ell,m}^{+}(t)\}$$
$$\mathscr{H}^{+}(\mathbf{r},t)=\sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell}\frac{1}{r}(-i)^{\ell}\{-\mathbf{Y}_{\ell,m}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}a_{\ell,m}^{+}(t)-c\partial_t^{-1}\nabla[\mathbf{Y}_{\ell,m}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}b_{\ell,m}^{+}(t)]\}. \tag{58}$$

Furthermore, we can now express the real electric and magnetic fields using the real vector spherical harmonics (refer to Appendix A for a review) following lines analogous to those used for scalar fields in Sec. II C. By analogy with our treatment of the scalar field we begin by introducing the new (real) multipole moments

$$c_{\ell,m}^{(j)}(t)=-\frac{1}{\ell(\ell+1)}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\mathbf{F}_H(\mathbf{s},t)\cdot\mathbf{S}_{\ell,m}^{(j)}(\alpha,\beta),$$
$$d_{\ell,m}^{(j)}(t)=\frac{1}{\ell(\ell+1)}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\mathbf{F}_E(\mathbf{s},t)\cdot\mathbf{S}_{\ell,m}^{(j)}(\alpha,\beta). \tag{59}$$

We can also express $c_{\ell,m}^{(j)}(t)$ and $d_{\ell,m}^{(j)}(t)$ in terms of the slant-stack transform of the current distribution by substituting from Eq. (47) into Eq. (59). This procedure yields

$$c_{\ell,m}^{(j)}(t)=\frac{c^{-2}}{\ell(\ell+1)}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha[\mathbf{s}\times\hat{\mathscr{T}}(\mathbf{s},t)]\cdot\mathbf{S}_{\ell,m}^{(j)}(\alpha,\beta),$$
$$d_{\ell,m}^{(j)}(t)=\frac{c^{-2}}{\ell(\ell+1)}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\{\mathbf{s}\times[\mathbf{s}\times\hat{\mathscr{T}}(\mathbf{s},t)]\}\cdot\mathbf{S}_{\ell,m}^{(j)}(\alpha,\beta), \tag{60}$$

in analogy to its analytic signal counterpart Eq. (52). Both forms of time-dependent multipole moments defined above [i.e., Eqs. (59) and (60)] appear to be new.

By employing the vector counterpart of the procedure used to derive Eq. (33) from Eq. (25) (see Appendix B), we obtain from Eq. (58) the sought-after time-dependent multipole representation for the real, physical fields:

$$\mathscr{E}(\mathbf{r},t)=\sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m\frac{1}{r}\{-c\partial_t^{-1}\nabla\times[\mathbf{S}_{\ell,m}^{(j)}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}c_{\ell,m}^{(j)}(t)]+\mathbf{S}_{\ell,m}^{(j)}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}d_{\ell,m}^{(j)}(t)\},$$
$$\mathscr{H}(\mathbf{r},t)=\sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m\frac{1}{r}\{-\mathbf{S}_{\ell,m}^{(j)}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}c_{\ell,m}^{(j)}-c\partial_t^{-1}\nabla\times[\mathbf{S}_{\ell,m}^{(j)}(\theta,\phi)\mathscr{L}_{\ell}^{(r,t)}d_{\ell,m}^{(j)}(t)]\}, \tag{61}$$

where, as before, $c_m=1$ or 2 if $m=0$ or $m\geqslant 1$, respectively. Equation (61) is, to the authors' knowledge, also new. Note also that Eq. (59) gives the real electric and magnetic multipole moments in terms of the far electric and magnetic fields while Eq. (60) gives them in terms of the transverse part of the slant-stack transform of the current distribution. Thus, Eq. (61) yields solutions to the radiation problem as well as to the inverse diffraction problem from time domain far field data. The theory above also corroborates, in the time domain, the well-known result that the radiation pattern uniquely determines the radiated field everywhere outside the source, as is evident from Eqs. (61) and (59).

In analogy to our treatment of the scalar field in Sec. II, we conclude this section with alternative expressions for the real time-dependent multipole moments $c_{\ell,m}^{(j)}(t)$ and $d_{\ell,m}^{(j)}(t)$. The corresponding derivation follows by analogy with the procedure used in Sec. II.

We conveniently rewrite $a_{\ell,m}^{+}(t)$ as defined in Eq. (52) as

$$a^{+}_{\ell,m}(t)=-\frac{i^{\ell}c^{-1}}{\ell(\ell+1)}\int_{r\leqslant a}d\mathbf{r}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha[\nabla\times\mathscr{T}^{+}(\mathbf{r},t+\mathbf{s}\cdot\mathbf{r}/c)]\cdot\mathbf{Y}^{*}_{\ell,m}(\alpha,\beta). \quad (62)$$

The result above was obtained by using (see Appendix C)

$$\mathbf{s}\times\hat{\mathscr{T}}^{+}(\mathbf{s},t)=-c\,\partial_t^{-1}\int d\mathbf{r}\nabla\times\mathscr{T}^{+}(\mathbf{r},t+\mathbf{s}\cdot\mathbf{r}/c). \quad (63)$$

A convenient expression for $b^{+}_{\ell,m}(t)$ is obtained, on the other hand, by substituting from Eq. (46) into Eq. (52) and interchanging orders of **r** and **s** integration, thus giving

$$b^{+}_{\ell,m}(t)=-\frac{i^{\ell}c^{-2}}{\ell(\ell+1)}\frac{\partial}{\partial t}\int_{r\leqslant a}d\mathbf{r}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\,\mathscr{T}^{+}(\mathbf{r},t+\mathbf{s}\cdot\mathbf{r}/c)\cdot\mathbf{Y}_{\ell,m}(\alpha,\beta), \quad (64)$$

where we have made use of $\mathbf{s}\times\mathbf{s}\times\mathbf{S}^{(j)}_{\ell,m}(\alpha,\beta)=-\mathbf{S}^{(j)}_{\ell,m}(\alpha,\beta)$ and the well-known vector identity $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})=\mathbf{b}\cdot(\mathbf{c}\times\mathbf{a})=\mathbf{c}\cdot(\mathbf{a}\times\mathbf{b})$.

Next we make use of the vector counterpart of Eq. (22):[4]

$$\frac{(-i)^{\ell}}{4\pi}\int_{-\pi}^{\pi}d\beta\int_{0}^{\pi}d\alpha\ \sin\alpha\,\mathbf{Y}_{\ell,m}(\alpha,\beta)e^{i(\omega/c)\mathbf{s}\cdot\mathbf{r}}=j_{\ell}\left(\frac{\omega}{c}r\right)\mathbf{Y}_{\ell,m}(\theta,\phi). \quad (65)$$

By substituting from (65) into Eqs. (63) and (64) we obtain after some manipulations

$$a^{+}_{\ell,m}(t)=-\frac{i^{\ell}c^{-1}}{\ell(\ell+1)}\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\theta\mathbf{Y}^{*}_{\ell,m}(\theta,\phi)\cdot\left\{\int_{r\leqslant a}drr^{\ell}[\nabla\times\mathbf{I}^{+}(\mathbf{r},t)]\right\}, \quad (66)$$

$$b^{+}_{\ell,m}(t)=-\frac{i^{\ell}c^{-2}}{\ell(\ell+1)}\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\theta\mathbf{Y}^{*}_{\ell,m}(\theta,\phi)\cdot\left[\int_{r\leqslant a}drr^{\ell}\mathbf{I}^{+}(\mathbf{r},t)\right],$$

with

$$\mathbf{I}^{+}(\mathbf{r},t)=\int_{-1}^{1}d\eta(1-\eta^2)^{\ell}\frac{\partial^{\ell}}{\partial t^{\ell}}\mathscr{T}^{+}(\mathbf{r},t+r\eta/c). \quad (67)$$

Equation (66) has been derived before from the Debye representation;[6,7] here it has been obtained from the time-dependent plane wave representation.

Finally, the real time-dependent electric and magnetic multipole moments $c^{(j)}_{\ell,m}(t)$ and $d^{(j)}_{\ell,m}(t)$ are readily obtained from Eqs. (66) and (67) by recalling that $c^{+}_{\ell,m}(t)=(-i)^{\ell}a^{+}_{\ell,m}(t)$ and $d^{+}_{\ell,m}(t)=(-i)^{\ell}b^{+}_{\ell,m}(t)$. We obtain

$$c^{(j)}_{\ell,m}(t)=-\frac{c^{-1}}{\ell(\ell+1)}\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\theta\mathbf{S}^{(j)}_{\ell,m}(\theta,\phi)\cdot\left\{\int_{r\leqslant a}drr^{\ell}[\nabla\times\mathbf{I}(\mathbf{r},t)]\right\}, \quad (68)$$

$$d^{(j)}_{\ell,m}(t)=-\frac{c^{-2}}{\ell(\ell+1)}\frac{4\pi c^{-\ell}}{\ell!2^{\ell+1}}\frac{\partial}{\partial t}\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\theta\mathbf{S}^{(j)}_{\ell,m}(\theta,\phi)\cdot\left[\int_{r\leqslant a}drr^{\ell}\mathbf{I}(\mathbf{r},t)\right],$$

with

$$\mathbf{I}(\mathbf{r},t)=\int_{-1}^{1}d\eta(1-\eta^2)^{\ell}\frac{\partial^{\ell}}{\partial t^{\ell}}\mathscr{T}(\mathbf{r},t+r\eta/c). \quad (69)$$

Equations (68) and (69) are the electromagnetic case counterparts of Eqs. (41) and (42) and their interpretation is analogous to that of Eqs. (41) and (42).

ACKNOWLEDGMENTS

The authors acknowledge support from the U.S. Air Force Office of Scientific Research, under Grant No. F49620-93-1-0093 and the Center for Electromagnetics Research at Northeastern University.

APPENDIX A: ORDINARY AND VECTOR SPHERICAL HARMONICS

We use the standard definition of the spherical harmonics:[14]

$$Y_{\ell,m}(\theta,\phi)=y_{\ell,m}P_{\ell}^{m}(\cos\ \theta)e^{im\phi} \tag{A1}$$

with

$$y_{\ell,m}=\sqrt{\frac{(2\ell+1)(\ell-m)!}{4\pi(\ell+m)!}}, \tag{A2}$$

where $\ell=0,1,2,...$, $m=0,\pm1,...,\pm\ell$, and $P_{\ell}^{m}(x)$ are the associated Legendre polynomials, defined by means of the Rodrigues formula

$$P_{\ell}^{m}(x)=\frac{(-1)^m}{2^{\ell}\ell!}\,(1-x^2)^{m/2}\left(\frac{d}{dx}\right)^{\ell+m}(x^2-1)^{\ell}. \tag{A3}$$

The spherical harmonics obey the orthogonality and closure properties[14]

$$\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\ \theta Y_{\ell,m}^{*}(\theta,\phi)Y_{\ell',m'}(\theta,\phi)=\delta_{\ell,\ell'}\delta_{m,m'}, \tag{A4}$$

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}Y_{\ell,m}(\theta,\phi)Y_{\ell,m}^{*}(\theta',\phi')=\frac{1}{\sin\ \theta}\,\delta(\theta-\theta')\,\delta(\phi-\phi')\equiv\delta(\hat{\mathbf{r}}-\hat{\mathbf{r}}'). \tag{A5}$$

The real spherical harmonics $S_{\ell,m}^{(1)}(\theta,\phi)\equiv\Re\{Y_{\ell,m}(\theta,\phi)\}$ and $S_{\ell,m}^{(2)}(\theta,\phi)\equiv\Im\{Y_{\ell,m}(\theta,\phi)\}$, where $\ell=0,1,2,...$, and $m=0.1,...,\ell$, obey, on the other hand, the orthogonality and closure properties of Eqs. (31) and (32).[3]

The vector spherical harmonic $\mathbf{Y}_{\ell,m}(\theta,\phi)$ is defined by[14,21]

$$\mathbf{Y}_{\ell,m}(\theta,\phi)=\mathbf{L}Y_{\ell,m}(\theta,\phi), \tag{A6}$$

where

$$\mathbf{L}=-i\mathbf{r}\times\nabla=-i\left(\mathbf{u}_{\phi}\frac{\partial}{\partial\theta}-\frac{1}{\sin\ \theta}\,\mathbf{u}_{\theta}\frac{\partial}{\partial\phi}\right). \tag{A7}$$

In Eq. (A7), $\mathbf{u}_{\theta}$ and $\mathbf{u}_{\phi}$ are, respectively, the unit vectors in the positive θ and ϕ directions. The $\mathbf{Y}_{\ell,m}(\theta,\phi)$'s are defined for $\ell=1,2,...$, and $m=0,\pm1,...,\pm\ell$. The value $\ell=0$ is excluded since $Y_{0,0}=1/\sqrt{4\pi}$, hence from Eqs. (A6) and (A7) $\mathbf{Y}_{0,0}=0$.

The vector spherical harmonics $\mathbf{Y}_{\ell,m}(\theta,\phi)$ are everywhere tangential to the unit sphere [such that $\hat{\mathbf{r}}\cdot\mathbf{Y}_{\ell,m}(\theta,\phi)=0$]. They satisfy the orthogonality condition[14,21]

$$\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\ \sin\ \theta\mathbf{Y}_{\ell,m}^{*}(\theta,\phi)\cdot\mathbf{Y}_{\ell',m'}(\theta,\phi)=\ell(\ell+1)\,\delta_{\ell,\ell'}\delta_{m,m'}. \tag{A8}$$

They also form, together with the associated vector functions $\hat{\mathbf{r}}\times\mathbf{Y}_{\ell,m}(\theta,\phi)$, a complete orthogonal set in terms of which all well-behaved vector functions $\mathbf{F}(\hat{\mathbf{r}})$ with $\hat{\mathbf{r}}\cdot\mathbf{F}(\hat{\mathbf{r}})=0$ can be expanded.[14,21]

We define the real vector spherical harmonics $\mathbf{S}_{\ell,m}^{(j)}$ via

$$\mathbf{S}^{(1)}_{\ell,m}(\theta,\phi)=\mathcal{F}\{\mathbf{Y}_{\ell,m}(\theta,\phi)\}=-i\mathbf{L}S^{(1)}_{\ell,m},$$
$$\mathbf{S}^{(2)}_{\ell,m}(\theta,\phi)=\Re\{\mathbf{Y}_{\ell,m}(\theta,\phi)\}=i\mathbf{L}S^{(2)}_{\ell,m}, \tag{A9}$$

so that, from (31) and $L^2Y_{\ell,m}=\ell(\ell+1)Y_{\ell,m}$ (see Refs. 14 and 21),

$$\int_{-\pi}^{\pi}d\phi\int_{0}^{\pi}d\theta\,\sin\theta\mathbf{S}^{(j)}_{\ell,m}(\theta,\phi)\cdot\mathbf{S}^{(j')}_{\ell',m'}(\theta,\phi)=\ell(\ell+1)\delta_{\ell,\ell'}\delta_{m,m'}\delta_{j,j'}\frac{1}{c_m}. \tag{A10}$$

They also form, together with the vector functions $\hat{\mathbf{r}}\times\mathbf{S}^{(j)}_{\ell,m}(\theta,\phi)$ a complete orthogonal set in terms of which all well-behaved vector functions $\mathbf{F}(\hat{\mathbf{r}})$ with $\hat{\mathbf{r}}\cdot\mathbf{F}(\hat{\mathbf{r}})=0$ can be expanded.

APPENDIX B: DERIVATION OF EQ. (33)

A key step in deriving (33) will consist in showing that

$$2\Re\left\{\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}(-i)^{\ell}Y^{m}_{\ell}(\theta,\phi)\mathscr{L}^{(r,t)}_{\ell}a^{+}_{\ell,m}(t)\right\}=\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_mS^{(j)}_{\ell,m}(\theta,\phi)\mathscr{L}^{(r,t)}_{\ell}q^{(j)}_{\ell,m}(t) \tag{B1}$$

with $a^{+}_{\ell,m}(t)$ and $q^{(j)}_{\ell,m}(t)$ given by (26) and (34), respectively, and where $c_m=1$ or 2 for $m=0$ or $m\geqslant 1$, respectively. To simplify the manipulations we will use Dirac bra-ket notation.[21] Thus we represent the spherical and real spherical harmonics $Y_{\ell,m}$ and $S^{(j)}_{\ell,m}$, respectively, by the vectors (kets) $|\ell,m\rangle$ and $|\ell,m,j\rangle$, where[21]

$$Y_{\ell,m}(\theta,\phi)\equiv\langle\mathbf{r}|\ell,m\rangle, \tag{B2}$$

$$S^{(j)}_{\ell,m}(\theta,\phi)\equiv\langle\mathbf{r}|\ell,m,j\rangle. \tag{B3}$$

In Dirac bra-ket notation, the orthogonality and closure properties defined in Eqs. (A4) and (A5) and Eqs. (31) and (32) translate into

$$\langle\ell,m|\ell',m'\rangle=\delta_{\ell,\ell'}\delta_{m,m'}, \tag{B4}$$

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}|\ell,m\rangle\langle\ell,m|=I, \tag{B5}$$

$$\langle\ell,m,j|\ell',m',j'\rangle=\frac{1}{c_m}\delta_{\ell,\ell'}\delta_{m,m'}\delta_{j,j'}, \tag{B6}$$

$$\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\sum_{j=1,2}c_m|\ell,m,j\rangle\langle\ell,m,j|=I, \tag{B7}$$

where I is the identity operator. We introduce, conveniently, the new analytic multipole moments

$$q^{+}_{\ell,m}(t)\equiv(-i)^{\ell}a^{+}_{\ell,m}(t), \tag{B8}$$

which are the ordinary spherical harmonic counterpart of $q^{(j)}_{\ell,m}(t)$. By substituting from (B8) into (25) and (26) while using Dirac bra-ket notation we obtain

$$U(\mathbf{r},t)=\frac{1}{r}2\Re\left\{\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}\langle\mathbf{r}|\ell,m\rangle\mathscr{L}^{(r,t)}_{\ell}\langle\ell,m|F^{+}(t)\rangle\right\}. \tag{B9}$$

By inserting the identity operator right after the term $\langle\ell,m|$ in Eq. (B9) while making use of Eqs. (B5) and (B7) and interchanging orders of summation one readily finds

$$U(\mathbf{r},t)=\frac{1}{r}\sum_{\ell'=0}^{\infty}\sum_{m'=0}^{\ell'}\sum_{j'=1,2}c_{m'}\langle\mathbf{r}|\ell',m',j'\rangle\mathscr{L}_{\ell'}^{(r,t)}\langle\ell',m',j'|F(t)\rangle. \tag{B10}$$

Finally, we complete our derivation of Eq. (33) by substituting expression (B3) and $q_{\ell',m'}^{(j')}(t)=\langle\ell',m',j'|F(t)\rangle$ [this follows from Eq. (34)] in Eq. (B10).

APPENDIX C: DERIVATION OF EQS. (47)

Consider a four-dimensional (4D) space-time with Cartesian coordinates ($x_0=ct$, $x_1=x$, $x_2=y$, $x_3=z$) and denote the coordinate vector in space-time as $\mathbf{x}$. The Radon transform $R\mathbf{f}(\mathbf{v},\xi)$ of a rather arbitrary vector function $\mathbf{f}(\mathbf{x})$ is defined as

$$R\mathbf{f}(\mathbf{v},\xi)=\int d\mathbf{x}\mathbf{f}(\mathbf{x})\,\delta(\xi-\mathbf{v}\cdot\mathbf{x}), \tag{C1}$$

where $\mathbf{v}$ is a unit vector in 4D Radon domain and ξ is a scalar. We will use the property [see Ref. 22, Eq. (6.7)]

$$R(\mathscr{H}\mathbf{f})=\mathscr{H}\left(v_1\frac{\partial}{\partial\xi},\ldots,v_n\frac{\partial}{\partial\xi}\right)R\mathbf{f}, \tag{C2}$$

where x_k is the kth Cartesian component of $\mathbf{x}$, v_k is the kth Cartesian component of the unit vector $\mathbf{v}$ and $\mathscr{H}(\partial/\partial x_1,\ldots,\partial/\partial x_n)$ is a linear operator with constant coefficients.

By putting $\mathbf{f}^+(\mathbf{x})\equiv\mathscr{T}^+(\mathbf{r},t=x_0/c)$, Eq. (46) and the various definitions above yield

$$\hat{\mathscr{T}}^+(\mathbf{s},t)=\frac{1}{\sqrt{2}}R\mathbf{f}^+(\mathbf{v}_\mathbf{s},ct/\sqrt{2}), \tag{C3}$$

where $\mathbf{v}_\mathbf{s}=(1/\sqrt{2})(1,-\mathbf{s})$ is the unit vector whose Cartesian components are ($x_0=(1/\sqrt{2})$, $x_1=-(1/\sqrt{2})$, $x_2=-(1/\sqrt{2})$, $x_3=-(1/\sqrt{2})$).

By substituting into Eq. (45) from Eqs. (C1)–(C3) we obtain

$$\mathbf{F}_E^+(\mathbf{s},t)=\frac{1}{\sqrt{2}}R\left\{\left[-\frac{1}{c^2}\frac{\partial}{\partial t}+\partial_t^{-1}\nabla\nabla\cdot\right]\mathbf{f}^+\right\}(\mathbf{v}_\mathbf{s},ct/\sqrt{2}) \tag{C4}$$

and

$$\mathbf{F}_H^+(\mathbf{s},t)=\frac{1}{c\sqrt{2}}R\{\nabla\times\mathbf{f}^+\}(\mathbf{v}_\mathbf{s},ct/\sqrt{2}). \tag{C5}$$

Using Eqs. (C2) and (C3) one can easily show that

$$\begin{aligned}R\{\nabla\nabla\cdot\mathbf{f}\}(\mathbf{v}_\mathbf{s},ct/\sqrt{2})&=\frac{1}{c^2}\frac{\partial^2}{\partial t^2}[\mathbf{ss}\cdot R\mathbf{f}(\mathbf{v}_\mathbf{s},ct/\sqrt{2})],\\ R\{\nabla\times\mathbf{f}\}(\mathbf{v}_\mathbf{s},ct/\sqrt{2})&=-\frac{1}{c}\frac{\partial}{\partial t}[\mathbf{s}\times R\mathbf{f}(\mathbf{v}_\mathbf{s},ct/\sqrt{2})].\end{aligned} \tag{C6}$$

The final step consists of applying the results (C6)–(C3) to Eqs. (C4)–(C5) and substituting into the result thus obtained from Eq. (C3).

[1] E. Heyman and L. B. Felsen, ''Complex-Source Pulsed Beam Fields,'' J. Opt. Soc. Am. A **6**, 806–817 (1989).
[2] R. W. Ziolkowski, I. M. Besieris, and A. M. Shaarawi, ''Localized Wave Representations of Acoustic and Electromagnetic Radiation,'' Proc. IEEE **79**, 1371–1377 (1991).
[3] E. Heyman and A. J. Devaney, ''Time-Dependent Multipoles and Their Application for Radiation from Volume Source Distributions,'' J. Math. Phys. **37**, 682–692 (1996).

[4] A. J. Devaney and E. Wolf, ''Multipole Expansions and Plane-Wave Representations of the Electromagnetic Field,'' J. Math. Phys. **15**, 234–244 (1974).
[5] K. D. Granzow, ''Multipole Theory in the Time Domain,'' J. Math. Phys. **7**, 634–640 (1966).
[6] W. C. Davidon, ''Time-Dependent Multipole Analysis,'' J. Phys. A **6**, 1635–1646 (1973).
[7] W. B. Campbell, J. Macek, and T. A. Morgan, ''Relativistic Time-Dependent Multipole Analysis for Scalar, Electromagnetic and Gravitational Fields,'' Phys. Rev. D **15**, 2156–2164 (1977).
[8] T. B. Hansen, ''Formulation of Spherical Near-Field Scanning for Electromagnetic Fields in the Time Domain,'' IEEE Trans. Antennas Propag. **45**, 620–630 (1997).
[9] T. B. Hansen and A. D. Yaghjian, ''Planar Near-Field Scanning in the Time Domain. I. Formulation,'' IEEE Trans. Antennas Propag. **42**, 1268–1279 (1994).
[10] R. F. Harrington, ''On the Gain and Beamwidth of Directional Antennas,'' IRE Trans. Antennas Propag. **6**, 219–225 (1958).
[11] D. R. Rhodes, ''On a Fundamental Principle in the Theory of Planar Antennas,'' Proc. IEEE **52**, 1013–1021 (1964).
[12] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).
[13] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, New York, 1995).
[14] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962).
[15] A. J. Devaney and G. C. Sherman, ''Plane-Wave Representations for Scalar Wave Fields,'' SIAM (Soc. Ind. Appl. Math.) Rev. **15**, 765–786 (1973).
[16] E. Heyman, ''Time-Dependent Plane-Wave Spectrum Representations for Radiation from Volume Source Distributions,'' J. Math. Phys. **37**, 658–681 (1996).
[17] A. D. Yaghjian and T. B. Hansen, ''Time-Domain Far Fields,'' J. Appl. Phys. **79**, 2822–2830 (1996).
[18] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, 9th ed. (Dover, New York, 1970).
[19] T. B. Hansen and A. N. Norris, ''Exact Complex Source Representations of Transient Radiation,'' Wave Motion **26**, 101–115 (1997).
[20] A. Shlivinski, E. Heyman, and R. Kastner, ''Antenna Characterization in the Time Domain,'' IEEE Trans. Antennas Propag. **45**, 1140–1149 (1997).
[21] C. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics* (Wiley, New York, 1997).
[22] S. R. Deans, *The Radon Transform and Some of its Applications* (Wiley, New York, 1983).

Arnol'd's transformation—An example of a generalized canonical transformation

Daniel R. Stump
Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824-1116

(Received 23 December 1997; accepted for publication 26 March 1998)

The duality transformation described by Arnol'd, which relates the orbits for different central potentials through a change of variables, is shown to be an example of a generalized canonical transformation, in an extended phase space that includes time as a dynamical variable. © *1998 American Institute of Physics.* [S0022-2488(98)00307-7]

I. INTRODUCTION

Classical mechanics is an old subject, but new discoveries are still being made about the mathematical structure of the theory. Arnol'd has described a remarkable and beautiful class of duality transformations[1–3] in classical mechanics, that relate the orbits of two different central potentials, which he calls dual potentials. In the most striking example, the elliptical orbits of an isotropic Hooke's law potential $[V_2(r)=1/2kr^2]$ transform into the Keplerian elliptical orbits of an inverse-square force $[V_{-1}(r)=-\gamma/r]$. The interesting fact that these two, most basic examples in mechanics—the harmonic oscillator and the inverse-square force—are connected by a change of variables, deserves to be more widely known.[4]

The purpose of this paper is to show that Arnol'd's transformation may be expressed as a new kind of canonical transformation. The duality transformation is no ordinary canonical transformation in phase space, but rather a generalized canonical transformation in an extended phase space that includes time as one of the generalized coordinates. Thus, this duality transformation, in addition to providing an interesting insight into the Kepler problem, also serves as a novel example of a canonical transformation.

In Secs. II and V, Arnol'd's transformation is described. In Sec. III we define the generalized canonical transformation. In Sec. IV we show that Arnol'd's transformation is an example of a generalized canonical transformation.

II. ARNOL'D'S DUALITY TRANSFORMATION

We begin by considering motion of a mass m in a two-dimensional orbit plane with Cartesian coordinates ξ and η, under the influence of a central potential $V_{\mathrm{H}}(\rho)$, where $\rho=\sqrt{\xi^2+\eta^2}$ is the radial coordinate. The time coordinate for the $\xi\eta$ orbit is denoted by τ. To define the duality transformation it is useful to combine ξ and η in a complex variable ζ, as $\zeta=\xi+i\eta$.

The duality transformation is a transformation of space and time variables $(\xi,\eta,\tau)\rightarrow(x,y,t)$ that transforms orbits of $V_{\mathrm{H}}(\rho)$ into dual orbits in the space–time with coordinates (x,y,t); the dual orbits correspond to some other central force. Let $z=x+iy$. The transformation equations are[1]

$$z=\zeta^{\alpha}, \tag{1}$$

$$\frac{dt}{d\tau}=|\zeta|^{2(\alpha-1)}, \tag{2}$$

where α is a free parameter. We may choose α as appropriate to make the transformation interesting. For example, the transformation appropriate for the duality of Hooke and Kepler elliptical orbits has $\alpha=2$.

 © 1998 American Institute of Physics

The reason for the transformation of time (2) is that then the law of equal areas holds in the dual (xy) space. The equal-area law holds in the $\xi\eta$ plane, because $V_{\rm H}(\rho)$ is central: $|\zeta|^2 d\phi/dt$ is constant, where $\phi=\arg\zeta$ is the angular coordinate and $|\zeta|\equiv\rho$ is the radial coordinate, in the $\xi\eta$ plane. For (1)–(2) the equal-area law also holds in the xy plane; i.e., $|z|^2 dF/dt$ is constant where $F=\arg z=\alpha\phi$ is the angular coordinate in the xy plane, because

$$|z|^2\,\frac{dF}{dt}=\alpha|\zeta|^2\,\frac{d\phi}{d\tau}.$$

Before constructing the generalized canonical transformation, it is useful to see how the equation of motion changes under the transformation (1)–(2). First, the equation of motion in the $\xi\eta$ plane is

$$\frac{d^2\zeta}{d\tau^2}=-\frac{1}{m}\left(\frac{\zeta}{\rho}\right)\frac{dV_{\rm H}}{d\rho},$$

and the energy of the $\xi\eta$ orbit is

$$E_{\rm H}=\frac{m}{2}\left|\frac{d\zeta}{d\tau}\right|^2+V_{\rm H}(\rho).$$

The equation of motion of the dual orbit that results from transforming the $\xi\eta$ orbit by the change of variables (1)–(2) is

$$\frac{d^2z}{dt^2}=-\frac{\alpha^2}{m}\left(\frac{z}{r}\right)\frac{d}{dr}\left[V_{\rm K}(r)-\frac{E_{\rm H}}{r^{2(\alpha-1)/\alpha}}\right];\tag{3}$$

in this equation $r\equiv|z|$ is the radial coordinate in the xy plane, and $V_{\rm K}(r)$ is the function

$$V_{\rm K}(r)=\frac{V_{\rm H}(r^{1/\alpha})}{r^{2(\alpha-1)/\alpha}}.\tag{4}$$

A general proof that $z(t)$ satisfies (3) is a straightforward exercise in calculus with complex numbers.[1–3] As a specific example, suppose $V_{\rm H}(\rho)$ is an isotropic Hooke's law potential, $V_{\rm H}(\rho)=1/2k\rho^2$, and let $\alpha=2$ in the transformation; then $V_{\rm K}(r)$ is a constant $(k/2)$, and the equation of motion for the dual orbit is

$$\frac{d^2z}{dt^2}=-\frac{4E_{\rm H}}{m}\frac{z}{r^3},$$

which is the equation for a Keplerian orbit. We shall derive (3) from the generalized canonical transformation in Sec. IV.

The duality transformation described so far is a transformation between *orbits*: We have only considered the transformation in space, not in phase space. In Sec. IV we shall construct the transformation between *systems*, and show that it is a generalized canonical transformation.

III. THE GENERALIZED CANONICAL TRANSFORMATION

Canonical transformations—transformations in phase space (q,p)—are a standard, but advanced, topic in classical mechanics.[5,6] It will be useful to review briefly the canonical transformation in general terms.

The canonical equations of motion are

$$\dot q=\frac{\partial H}{\partial p}\quad\text{and}\quad\dot p=-\frac{\partial H}{\partial q},\tag{5}$$

where H is the Hamiltonian. The variables q and p have n components. An overdot ($\dot{}$) denotes the derivative with respect to time τ, the independent evolution variable. These equations of motion may be derived from a variational principle in phase space: $\delta I=0$, where I is the integral

$$I=\int_{\tau_1}^{\tau_2}(p\dot{q}-H(q,p))d\tau.$$

The variation δ is over trajectories in phase space with given, fixed end points. A transformation in phase space $(q,p,H)\rightarrow(Q,P,K)$ is canonical if

$$pdq-Hd\tau=PdQ-Kd\tau+dF, \tag{6}$$

where dF is an arbitrary complete differential, and K is the transformed Hamiltonian. If (6) holds, then the variational principle $\delta I=0$ in the transformed representation gives the canonical equations of motion with Hamiltonian $K(Q,P)$,

$$\dot{Q}=\frac{\partial K}{\partial P} \quad \text{and} \quad \dot{P}=-\frac{\partial K}{\partial Q}. \tag{7}$$

The contribution to I from the complete differential, $\int dF=F(\tau_2)-F(\tau_1)$, has variation 0 because the end points are fixed. To be even more flexible, we could also allow the right side of (6) to be multiplied by a constant, but that would only be a scale transformation.

The Arnol'd transformation (1)–(2) is not an ordinary canonical transformation, because (2) is a position-dependent transformation of time. But we shall see that the Arnol'd transformation is a generalized canonical transformation is an extended phase space that includes time as a coordinate.

To define the *generalized* canonical transformations, we extend the $2n$-dimensional phase space $(q;p)$ to $2n+2$ dimensions $(q,\tau;p,\mathscr{E})$ where the time τ is now treated as another coordinate and $\mathscr{E}$ is the momentum conjugate to τ; and we take the Hamiltonian for this extended system to be $\tilde{H}$, given by

$$\tilde{H}(q,\tau;p,\mathscr{E})=H(q;p)+\mathscr{E}. \tag{8}$$

Then the canonical equations, with an evolution parameter s that parametrizes the position on the trajectory in the extended phase space, are

$$\frac{dq}{ds}=\frac{\partial\tilde{H}}{\partial p}=\frac{\partial H}{\partial p} \quad \text{and} \quad \frac{d\tau}{ds}=\frac{\partial\tilde{H}}{\partial\mathscr{E}}=1, \tag{9}$$

$$\frac{dp}{ds}=-\frac{\partial\tilde{H}}{\partial q}=-\frac{\partial H}{\partial q} \quad \text{and} \quad \frac{d\mathscr{E}}{ds}=\frac{\partial\tilde{H}}{\partial\tau}=0. \tag{10}$$

Thus, for (8) we have $\tau(s)=s$, and $\mathscr{E}$ is a constant because τ is an ignorable coordinate. Also, q and p satisfy the original equations of motion because $\tau=s$. The extension of the phase space has had no real effect on dynamics. But in this extended phase space we can make more general canonical transformations, including transformation of time, i.e., more general than an ordinary canonical transformation. The capability to make more general transformations may reveal new aspects of the theory, as in Arnol'd's duality relations.

A generalized canonical transformation is a canonical transformation in the extended phase space,

$$(q,\tau;p,\mathscr{E};\tilde{H})\rightarrow(Q,t;P,E;\tilde{K}),$$

such that the equations of motion in the new representation, *for the same evolution parameter* s, are the canonical equations with transformed Hamiltonian $\tilde{K}$. The condition for the transformation to be canonical is, analogous to (6), that

$$p\,dq+\mathscr{E}\,d\tau-\tilde{H}\,ds=P\,dQ+E\,dt-\tilde{K}\,ds. \tag{11}$$

[We could also add a complete differential to the right side of (11), but that is not necessary in this paper.] It is important to realize that the evolution parameter (s) must be the same for the $(q,\tau;p,\mathscr{E})$ and $(Q,t;P,E)$ systems, so that the same variational principle $\delta I=0$ applies to both systems.

In Sec. IV we apply this formalism, i.e., the condition (11), to the Arnol'd transformation.

IV. THE DUALITY TRANSFORMATION AS A GENERALIZED CANONICAL TRANSFORMATION

In Sec. I we defined a system with coordinates (ξ,η), or $\zeta=\xi+i\,\eta$, and time τ. The extended phase space for that system is the six-dimensional space with coordinates (ξ,η,τ) and conjugate momenta $(p_\xi,p_\eta,\mathscr{E})$. The Hamiltonian for this extended system is, by (8),

$$\tilde{H}=\frac{1}{2m}\,(p_\xi^2+p_\eta^2)+V_{\mathrm{H}}(\sqrt{\xi^2+\eta^2})+\mathscr{E}. \tag{12}$$

Again, we denote by s the independent evolution variable for trajectories in the six-dimensional (6D) phase space, and let an overdot $(\dot{})$ denote d/ds. The canonical equations are

$$(\dot{\xi},\dot{\eta},\dot{\tau})=\left(\frac{p_\xi}{m},\ \frac{p_\eta}{m},\ 1\right),$$

$$(\dot{p}_\xi,\dot{p}_\eta,\dot{\mathscr{E}})=\left(-\frac{\partial V_{\mathrm{H}}}{\partial\xi},-\frac{\partial V_{\mathrm{H}}}{\partial\eta},\ 0\right).$$

The projection of the trajectory to the $(\xi,\eta;p_\xi,p_\eta)$ subspace of the 6-D extended space is the trajectory of the original problem.

Now we apply the duality transformation (1)–(2) to this system. To construct a generalized canonical transformation we must supplement the rules (1)–(2) by transformation rules for the conjugate momenta. That is, we seek a generalized canonical transformation,

$$(\zeta,\tau;p_\zeta,\mathscr{E};\tilde{H})\rightarrow(z,t;P_z,E;\tilde{K}),$$

such that the coordinate part of the transformation is the Arnol'd duality transformation (1)–(2),

$$(z,t)=\left(\zeta^\alpha,\int|\zeta|^{2(\alpha-1)}d\tau\right), \tag{13}$$

where $z=x+iy$. The momentum $p_\zeta\equiv p_\xi+ip_\eta$ and the transformed momentum $P_z\equiv P_x+iP_y$ are required to be $p_\zeta=m\,d\zeta/d\tau$ and $P_z=M\,dz/dt$, where M is a parameter to be determined. M is the mass of the particle in the dual problem. The desired transformed momentum P_z conjugate to z is[3]

$$P_z=\frac{\alpha M}{m}\,\frac{z\zeta^*}{|z|^2}\,p_\zeta. \tag{14}$$

The transformed momentum E conjugate to t can be any function of $\mathscr{E}$, because both of these quantities turn out to be constants of the motion; we will leave E unspecified.

For this transformation to be canonical, the transformation rules must satisfy the condition (11). For (13) and (14) we have

$$P_x\,dx+P_y\,dy=\mathrm{Re}\ P_z^*\,dz=\frac{\alpha^2M}{m}\,(p_\xi\,d\xi+p_\eta\,d\eta).$$

At this point we choose to fix the value of M by

$$M=\frac{m}{\alpha^2}.$$

[We could instead leave M unspecified, and multiply the right side of (11) by the constant $m/(\alpha^2 M)$, a rescaling.] Then the condition (11) determines that the transformed Hamiltonian must be

$$\tilde{K}=\tilde{H}-\mathscr{E}\frac{d\tau}{ds}+E\frac{dt}{ds}.$$

We are free to set the evolution parameter s equal to τ, the time in the original problem, which is, after all, required by the equation of motion in the original 6-D phase space; then the transformed Hamiltonian becomes

$$\tilde{K}=\frac{1}{2m}(p_\xi^2+p_\eta^2)+V_{\mathrm{H}}(\sqrt{\xi^2+\eta^2})+E(\xi^2+\eta^2)^{(\alpha-1)}. \tag{15}$$

Finally, we rewrite $\tilde{K}$ in terms of the new coordinates and momenta, from the relations

$$\xi^2+\eta^2=(x^2+y^2)^{1/\alpha},$$

$$p_\xi^2+p_\eta^2=\alpha^2(P_x^2+P_y^2)(x^2+y^2)^{(\alpha-1)/\alpha}.$$

The result is that the transformed Hamiltonian is

$$\tilde{K}=\left(\frac{1}{2M}(P_x^2+P_y^2)+V_{\mathrm{K}}(\sqrt{x^2+y^2})+E\right)(x^2+y^2)^{(\alpha-1)/\alpha}, \tag{16}$$

where $V_{\mathrm{K}}(r)$ is determined from the original potential by Eq. (4).

Because the transformation is canonical in the 6-D phase space, the equations of motion in the (x,y,t) representation are canonical, with Hamiltonian $\tilde{K}$. The independent evolution parameter is τ (or s), and the equations of motion are

$$\frac{dx}{d\tau}=\frac{\partial\tilde{K}}{\partial P_x}=\frac{P_x}{M}r^{2(\alpha-1)/\alpha}, \tag{17}$$

$$\frac{dy}{d\tau}=\frac{\partial\tilde{K}}{\partial P_y}=\frac{P_y}{M}r^{2(\alpha-1)/\alpha}, \tag{18}$$

$$\frac{dt}{d\tau}=\frac{\partial\tilde{K}}{\partial E}=r^{2(\alpha-1)/\alpha}, \tag{19}$$

where $r=\sqrt{x^2+y^2}$; and similarly,

$$\frac{dP_x}{d\tau}=-\frac{\partial\tilde{K}}{\partial x}=-\frac{x}{r}\left(\frac{2(\alpha-1)\tilde{K}}{\alpha r^{3-2/\alpha}}+\frac{dV_{\mathrm{K}}}{dr}\right)r^{2(\alpha-1)/\alpha}, \tag{20}$$

$$\frac{dP_y}{d\tau}=-\frac{\partial\tilde{K}}{\partial y}=-\frac{y}{r}\left(\frac{2(\alpha-1)\tilde{K}}{\alpha r^{3-2/\alpha}}+\frac{dV_{\mathrm{K}}}{dr}\right)r^{2(\alpha-1)/\alpha}, \tag{21}$$

$$\frac{dE}{d\tau}=-\frac{\partial\tilde{K}}{\partial t}=0. \tag{22}$$

It should be emphasized that τ is not the time coordinate of the transformed system. The time coordinate is t and τ (or s) is the independent variable that parametrizes the position on the trajectory in the 6-D phase space $(x,y,t;P_x,P_y,E)$. The equations of motion for $(x,y;P_x,P_y)$ as functions of the time t of the transformed system are

$$\frac{dx}{dt}=\frac{dx/d\tau}{dt/d\tau}=\frac{P_x}{M}, \tag{23}$$

$$\frac{dy}{dt}=\frac{dy/d\tau}{dt/d\tau}=\frac{P_y}{M}; \tag{24}$$

and similarly,

$$\frac{dP_x}{dt}=-\frac{x}{r}\left(\frac{2(\alpha-1)\tilde{K}}{\alpha r^{3-2/\alpha}}+\frac{dV_{\mathrm{K}}}{dr}\right), \tag{25}$$

$$\frac{dP_y}{dt}=-\frac{y}{r}\left(\frac{2(\alpha-1)\tilde{K}}{\alpha r^{3-2/\alpha}}+\frac{dV_{\mathrm{K}}}{dr}\right). \tag{26}$$

These t-dependent equations are equivalent to (3). It is possible to derive the equation of motion (3) directly, without the formalism of the generalized canonical transformation; but then the full structure of the transformation is not revealed.

The canonical equations (17)–(22) have an unfamiliar form, because x and y appear in the velocities, and P_x and P_y appear in the forces. But the equations of motion (23)–(26) do reduce to a familiar form. First, $\tilde{K}$ is a constant of motion, as may be proven directly, or, more simply, from the fact that $\tilde{K}$ is the Hamiltonian for a system with translation invariance in the evolution parameter (τ). Therefore, for any particular trajectory we may write

$$\tilde{K}=\gamma,$$

where γ is a constant. Then the equations (23)–(26) are isomorphic to the motion of mass M in the central potential,

$$V_{\mathrm{eff}}(r;\gamma)=\frac{-\gamma}{r^{2(\alpha-1)/\alpha}}+V_{\mathrm{K}}(r).$$

Equations (23)–(26) and (16) are not simply the equations for motion in a fixed potential. Any particular trajectory is the same as motion in a potential, but the potential depends on the trajectory. For example, different initial values $(x_0,y_0,P_{0x},P_{0y},E_0)$ will have different values of γ.

The constant of the motion E, which is the momentum conjugate to the time t, may be determined from (16). For a trajectory with $\tilde{K}=\gamma$ we have

$$-E=\frac{1}{2M}(P_x^2+P_y^2)-\frac{\gamma}{r^{2(\alpha-1)/\alpha}}+V_{\mathrm{K}}(r);$$

that is, $-E$ equals the total energy of the effective potential problem in the xy plane.

Finally, we return to the duality of orbits: What is the transform of an orbit of the original (ξ,η,τ) system? We may answer this question from the generalized canonical transformation. Let E_{H} be the energy (value of H) of the $\xi\eta$ orbit. Then by (15) the value of $\tilde{K}$ is $E_{\mathrm{H}}+E(\xi^2+\eta^2)^{(\alpha-1)}$. Because $\tilde{K}$ is a constant of the motion, the value of E is 0. Thus the dual orbit is a trajectory of $V_{\mathrm{eff}}(r;\gamma)$ with $\gamma=E_{\mathrm{H}}$, and total energy $-E=0$.

Solutions of the (x,y,t) *system with* $E\neq 0$. The transform of any orbit of the original (ξ,η,τ) dynamical system is an orbit of the (x,y,t) dynamical system, and that (x,y,t) orbit has $E=0$. But the (x,y,t) system (17)–(22) also has orbits with $E\neq 0$. A trajectory with $E\neq 0$ is not the transform of any solution of the original (ξ,η,τ) system. So, what is the connection between the (x,y,t) trajectories with $E\neq 0$, and the original (ξ,η,τ) system?

The answer to this question emerges from the observation that t is not a local function of (ξ,η,τ). The relation (2) that defines t is a differential relation, so $t=\int(\xi^2+\eta^2)^{\alpha-1}d\tau$ is a *path-dependent* function of the (ξ,η,τ) variables. Thus the transformation $(\xi,\eta,\tau)\rightarrow(x,y,t)$ is a one-to-one correspondence between *paths* in (ξ,η,τ) and (x,y,t), rather than a one-to-one correspondence between *points* in these spaces.

The variational principle $\delta I=0$ is different for the (ξ,η,τ) and (x,y,t) systems. The integrand of the action integral I is the same in both cases, by (11). But the conditions at the end points are different. For the (ξ,η,τ) system, I is stationary with respect to variations (δ) that leave the end points fixed in (ξ,η,τ); then $\delta I=0$ implies the canonical equations with Hamiltonian $\tilde{H}$. For the (x,y,t) system, I is stationary with respect to variations (δ') that leave the end points fixed in (x,y,t), and we denote this variation by δ'; then $\delta' I=0$ implies the canonical equations (17)–(22) with Hamiltonian $\tilde{K}$. The variation δ' also leaves the end points fixed in (ξ,η,τ): The end points of (ξ,η) are fixed because the transformation $(\xi,\eta)\rightarrow(x,y)$ is just a point transformation (1), and the end points of τ are fixed because $\tau=s$ and the integral I is over a fixed interval of s.

The variations δ and δ' are different, because t is a path-dependent function of (ξ,η,τ). The set of paths with fixed end points in (ξ,η,τ) do not all have fixed end points in (x,y,t), because the end points of t depend on the path in (ξ,η,τ).

The orbits of $\tilde{H}$ satisfy $\delta I=0$, and therefore also $\delta' I=0$: The meaning of $\delta I=0$ is that I is stationary for any variation that leaves the end points of the path fixed in (ξ,η,τ); then, *a fortiori*, I is stationary for variations that leave the end points fixed in both (ξ,η,τ) and (x,y,t), i.e., $\delta' I=0$. Thus any orbit of $\tilde{H}$ transforms to an orbit of $\tilde{K}$. But the converse is not true, because $\delta' I=0$ does not imply $\delta I=0$; so there exist orbits of $\tilde{K}$ that are not the transforms of orbits of $\tilde{H}$.

The dynamical solutions of the (x,y,t) system that have $E=0$, have corresponding trajectories in (ξ,η,τ) for which $\tilde{H}$ is a constant of the motion, by (15). These are the orbits of $\tilde{K}$ that are transforms of orbits of $\tilde{H}$.

A dynamical solution of the (x,y,t) system that has $E\neq 0$, has a corresponding trajectory in (ξ,η,τ) for which $\tilde{H}$ is not a constant of the motion; thus, the trajectory in (ξ,η,τ) is not a solution of the original problem. The equations of motion satisfied by this (ξ,η,τ) trajectory derive not from $\delta I=0$, but rather from $\delta' I=0$. That is, I is stationary with respect to constrained variations, subject to the constraint that the end points t_1 and t_2 are fixed.

We may implement this constrained variation by the method of the undetermined Lagrange multiplier. The constrained variational principle $\delta' I=0$ is equivalent to the unconstrained variational principle,

$$\delta\left[I-\lambda\int_{s_1}^{s_2}(\xi^2+\eta^2)^{\alpha-1}\left(\frac{d\tau}{ds}\right)ds+\lambda(t_2-t_1)\right]=0, \tag{27}$$

where λ is the Lagrange multiplier; λ would be determined by the given value of t_2-t_1. It can be shown that the trajectory in (ξ,η,τ), that corresponds to an (x,y,t) trajectory for which the momentum conjugate to t has the value E, is a solution of the constrained equations with Lagrange multiplier $\lambda=E$. That is, E is the Lagrange multiplier for the constrained equations in the (ξ,η,τ) space. For the case $\lambda=0$, the constrained variational principle (27) is equivalent to $\delta I=0$, which is the original problem whose solutions are orbits of $\tilde{H}$.

The final conclusion is thus that the transform of an orbit of $\tilde{H}$, i.e., of an orbit that obeys $\delta I=0$, is an orbit of $\tilde{K}$ with $E=0$; and the transform of a trajectory that obeys (27), the variational principle for constrained variations, is an orbit of $\tilde{K}$ with $E=\lambda$.

V. DUAL POWER-LAW POTENTIALS

We have seen that the duality transformation (1)–(2) transforms an orbit of $V_{\rm H}(\rho)$ in the $\xi\eta$ plane into a dual orbit of the potential $V_{\rm eff}(r;\gamma)$ in the xy plane. If the energy of the $\xi\eta$ orbit is $E_{\rm H}$, then the parameter in $V_{\rm eff}(r;\gamma)$ is $\gamma=E_{\rm H}$. As a class of examples of this transformation, we now consider power-law potentials.

The result to be proven is that if $V_{\rm H}(\rho)$ is a power-law potential,

$$V_{\rm H}(\rho)=\tfrac{1}{2}k\rho^{\nu},$$

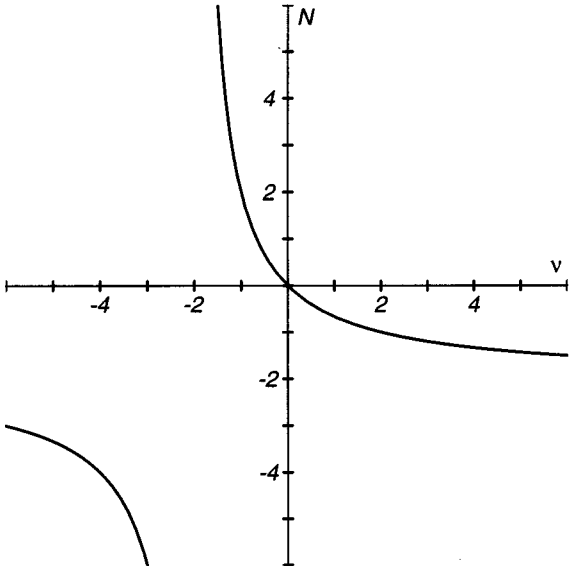

FIG. 1. The exponents ν and N of the dual power-law potentials.

then there exists a value of α such that the dual potential $V_{\text{eff}}(r;\gamma)$ is also a power law, for any orbit energy E_{H}. We construct this dual power-law potential by making $V_{\text{K}}(r)$ a constant. By (4),

$$V_{\text{K}}(r)=\frac{k}{2}\,r^{(\nu+2-2\alpha)/\alpha},$$

so the appropriate choice of α is

$$\alpha=\frac{\nu+2}{2};$$

with this choice the dual potential is

$$V_{\text{eff}}(r;\gamma)=-\gamma r^{N}+\frac{k}{2},$$

where

$$N=\frac{-2\nu}{\nu+2}.$$

In a more symmetric form, the relation between the dual exponents is

$$(\nu+2)(N+2)=4.$$

The constant term $k/2$ in $V_{\text{eff}}(r;\gamma)$ does not affect the dynamics in the dual problem, although it does contribute to the total energy $-E$. Thus, the dual orbit is an orbit for the power-law potential $-\gamma r^{N}$. For example, the dual potential of the isotropic Hooke's law potential ($\nu=2$) is the potential of an inverse-square force ($N=-1$); and *vice versa*. As another example, if $\nu=-4$ then $N=-4$; the potential $\propto\rho^{-4}$ is self-dual. Figure 1 shows the relation between the dual powers N and ν.

For the special case $\nu=2$ and $N=-1$ it is interesting to consider a specific orbit.[3] A particular elliptical orbit of $V_2(\rho)=\frac{1}{2}k\rho^2$ may be written as

$$(\xi(\tau),\eta(\tau))=(A\,\cos\,\omega\tau,B\,\sin\,\omega\tau),$$

where $\omega=\sqrt{k/m}$. This Hooke orbit is an ellipse centered at the origin. The dual orbit is[3] in parametric form with independent parameter τ,

$$x(\tau)=a(\cos 2\omega\tau-e) \tag{28}$$

$$y(\tau)=a\sqrt{1-e^2}\,\sin 2\omega\tau, \tag{29}$$

$$t(\tau)=\frac{T}{2\pi}\,(2\omega\tau-e\,\sin 2\omega\tau). \tag{30}$$

This Kepler orbit is an ellipse with a focal point at the origin, with semimajor axis $a=(B^2-A^2)/2$ and eccentricity $e=(B^2-A^2)/(B^2+A^2)$. The period of revolution is T, given by $T=\sqrt{4\pi^2a^3M/\gamma}$, where the strength of the Newtonian potential is $\gamma=1/2k(A^2+B^2)$.

Equations (28)–(30) are a standard parametrization of a Keplerian orbit. The variable $2\omega\tau$, which parametrizes the position and time on the orbit, is called the *eccentric anomaly* in celestial mechanics.[5,7] A nice aspect of the Arnol'd transformation is that it gives a new meaning to the eccentric anomaly: The eccentric anomaly is twice the phase angle of the dual Hooke ellipse.

Arnol'd's duality transformation reveals this interesting connection between two basic examples in mechanics—the harmonic oscillator and the inverse-square force. The duality is extended to other power laws by Fig. 1. The duality is even more generally between $V_{\rm H}(\rho)$ and $V_{\rm eff}(r;\gamma)$. The transformation is an example of a generalized canonical transformation, in an extended space that includes both space and time variables. We are left with the question of whether there are other interesting applications of this generalized kind of canonical transformation.

[1] V. I. Arnol'd, *Huygens and Barrow, Newton and Hooke* (Birkhauser-Verlag, Basel, 1990).
[2] T. Needham, Am. Math. Monthly **100**, 119 (1993); T. Needham, *Visual Complex Analysis* (Oxford University Press, Oxford, 1997).
[3] D. R. Stump, ''On the duality of elliptical orbits of Hooke and Kepler,'' submitted to Am. J. Phys.
[4] Arnol'd (see Ref. 1) credits the discovery of such a transformation between elliptical orbits to K. Bohlin, Bull. Astron. **28**, 144 (1911); Needham (see Ref. 2) also cites similar earlier work by Kasner.
[5] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison-Wesley, Reading, 1980).
[6] I. Percival and D. Richards, *Introduction to Dynamics* (Cambridge University Press, Cambridge, 1982).
[7] D. T. Greenwood, *Principles of Dynamics* (Prentice–Hall, Englewood Cliffs, 1965).

Lattice approximation of quantum statistical traces at a complex temperature

J. Lukkarinen[a)]
Helsinki Institute of Physics, P.O. Box 9, 00014 University of Helsinki, Finland

(Received 23 October 1997; accepted for publication 26 March 1998)

We show that the addition of an imaginary part to the temperature in the Gibbs ensemble will not destroy the convergence of the standard lattice approximation scheme for quantum mechanical systems with Hamiltonians of the type $\hat{H} = \hat{p}^2/2m + V(\hat{x})$, provided the potential function V is real and bounded from below and it satisfies the condition $\int d^n\mathbf{x}\, \exp(-tV(\mathbf{x})) < \infty$ for all $t > 0$. As a by-product we obtain an explicit bound for the real-temperature lattice kernels and a simple condition for the convergence of the real-temperature lattice expectation values of observables given by polynomially bounded functions. © *1998 American Institute of Physics.* [S0022-2488(98)00407-1]

I. INTRODUCTION

Ever since the introduction of the path-integrals by Feynman, the path-integral formulas with a complex ''temperature'' have presented an elusive mathematical problem—for instance, the Wiener measure, which is so useful for positive temperature statistics, cannot be extended to the complex case.[1] In fact, since real time processes and real temperature statistics can be identified with each other via analytic continuation, the present numerical nonperturbative methods in quantum field theories rely mostly on computations with lattice approximations to the *imaginary time*, i.e., real temperature, path-space measures.

There have already been proposals for numerical evaluation of the complex path-integrals,[2,3] but only now can we establish rigorously the assumed relations between the complex lattice integrals and the continuum traces. Also, the result of Theorem III.5 has been used[4] as an assumption in the derivation of a lattice approximation to the quantum microcanonical ensemble and the theorem now yields a wide range of potentials for which this approximation is valid.

For an introduction to the mathematics of operators on a Hilbert space, see, e.g., the second part of the comprehensive series[5] by Reed and Simon. The path-space measures and their applications are discussed in detail in another book[6] by Simon and the proofs of the statements made in the following section can be found from Sec. 6 of that book.

II. NOTATIONS AND DEFINITIONS

We will now consider quantum mechanics on the separable Hilbert space $L^2(\mathbb{R}^n)$; in the following, n will always denote the number of dimensions of the parameter space and we will use boldface letters for vectors in $\mathbb{R}^n$. In the same spirit, the n-dimensional Lebesque measure with respect to variable $\mathbf{x}$ is denoted by $d^n\mathbf{x}$. Operators will be distinguished by a ''hat''—especially the multiplication operator defined by a measurable function F will be denoted by $\hat{F}$.

The free Hamiltonian is defined in the standard way,

$$\hat{H}_0 := \frac{\hat{p}^2}{2m} = -\frac{\hbar^2}{2m}\nabla^2,$$

with the usual domain for this operator. To simplify the following discussion we will from now on

a)Electronic mail: jani.lukkarinen@helsinki.fi

 © 1998 American Institute of Physics

use the natural units $\hbar=1$ and, in addition, consider only the case $m=1$. The results for the situation $m\neq 1$ are easily reproduced from the following formulas by simply replacing t by t/m and the potential V by mV.

It is well-known that $\hat{H}_0$ generates an analytic semigroup on the right half-plane, where $e^{-\beta\hat{H}_0}$ are bounded integral operators for Re $\beta>0$ and they are given by the integral kernels

$$P(\mathbf{a},\mathbf{b};\beta):=\frac{1}{(2\pi\beta)^{n/2}}\exp\left(-\frac{1}{2\beta}|\mathbf{a}-\mathbf{b}|^2\right).$$

It is also evident that these integral kernels satisfy the semigroup property exactly, i.e., that for all $\mathbf{a}$, $\mathbf{b}\in\mathbb{R}^n$ and Re β_1, Re $\beta_2>0$,

$$\int d^n\mathbf{c}P(\mathbf{a},\mathbf{c};\beta_1)P(\mathbf{c},\mathbf{b};\beta_2)=P(\mathbf{a},\mathbf{b};\beta_1+\beta_2).$$

The potential V is assumed to be locally in $L^1(\mathbb{R}^n)$ and positive, but the results easily generalize for all real L^1_{loc} potentials bounded from below [just replace V by $V+\operatorname{ess\,sup}(-V)$]. The Hamiltonian $\hat{H}=\hat{H}_0+\hat{V}$ can now be defined as a sum of quadratic forms, when it is self-adjoint and bounded from below and, therefore, $e^{-\beta\hat{H}}$ (as well as $e^{-\beta\hat{V}}$) is an analytic semigroup on the right half-plane. In addition, we shall also require that

$$\int d^n\mathbf{a}e^{-tV(\mathbf{a})}<\infty,\quad \text{for all}\quad t>0, \tag{1}$$

so that $e^{-t\hat{H}}$ is trace-class for $t>0$ by the Golden, Thompson, Symanzik-inequality.[7]

Under these assumptions $e^{-t\hat{H}}$ is an integral operator for $t>0$ and it has an integral kernel K_∞, which can be obtained with aid of the conditional Wiener measure μ: for all $\mathbf{a}$, $\mathbf{b}\in\mathbb{R}^n$ and for all $t>0$ define

$$K_\infty(\mathbf{a},\mathbf{b};t)=\int d\mu_{\mathbf{a},\mathbf{b};t}(\boldsymbol{\omega})e^{-\int_0^t ds V(\boldsymbol{\omega}(s))}.$$

In addition to (1) we shall now require the potential V to be such that the Wiener integrals can be almost everywhere obtained from the lattice kernels

$$K_N(\mathbf{a},\mathbf{b};t):=\int\prod_{k=1}^{N-1}d^n\mathbf{x}_k\left(\frac{N}{2\pi t}\right)^{Nn/2}\exp\left[-\frac{N}{2t}\sum_{k=1}^{N}|\mathbf{x}_{k-1}-\mathbf{x}_k|^2-\frac{t}{N}\sum_{k=1}^{N}V(\mathbf{x}_k)\right]_{\mathbf{x}_N=\mathbf{b}}^{\mathbf{x}_0=\mathbf{a}},$$

i.e., that

$$K_\infty(\mathbf{a},\mathbf{b};t)=\lim_{N\to\infty}K_N(\mathbf{a},\mathbf{b};t),\quad \text{for all}\quad t>0\quad \text{and for almost all}\quad \mathbf{a},\mathbf{b}\in\mathbb{R}^n. \tag{2}$$

This requirement is clearly satisfied for *all* $\mathbf{a}$ and $\mathbf{b}$ if V is continuous, since then $\int ds V(\boldsymbol{\omega}(s))$ can be approximated by suitable Riemann sums leading to (2) by using the definition of the Wiener measure.

We shall also need the operator

$$\hat{T}(\beta):=e^{-\beta\hat{H}_0}e^{-\beta\hat{V}},$$

which is clearly a bounded integral operator for Re $\beta>0$, since then $\|\hat{T}(\beta)\|\leqslant\|e^{-\beta\hat{H}_0}\|\|e^{-\beta\hat{V}}\|\leqslant 1$. Notice that $K_N(\cdot,\cdot;\beta)$ is nothing but the integral kernel of the operator $\hat{T}(\beta/N)^N$.

III. TWO LATTICE THEOREMS

In this section we will prove two theorems concerning the convergence of traces of lattice operators. The first theorem will show that also certain unbounded multiplication operators are

measurable using the (real-temperature) Gibbs statistics if the potential increases sufficiently fast at infinity. This will be proven by using the explicit bounds given by the following lemma and its two corollaries. The rather technical proofs of the lemma and the corollaries are included in Sec. IV—the lemma itself is an extension of a result derived by Symanzik[8] for the continuum kernels.

Lemma III.1: For all $\mathbf{a}$, $\mathbf{b}$ *in* $\mathbb{R}^n$, *for all* $t>0$ *and for all positive integers* N,

$$K_N(\mathbf{a},\mathbf{b};t)\leqslant C_n t^{(n-1)/2}\int d^n\mathbf{c}P(\mathbf{a},\mathbf{c};t)e^{-tV(\mathbf{c})}P(\mathbf{c},\mathbf{b};t)\left(\frac{1}{|\mathbf{a}-\mathbf{c}|^{n-1}}+\frac{1}{|\mathbf{c}-\mathbf{b}|^{n-1}}\right)+\frac{1}{N}e^{-tV(\mathbf{b})}P(\mathbf{a},\mathbf{b};t), \tag{3}$$

where C_n *depends only on the dimension* n.

Taking the limit $N\to\infty$ in the previous lemma shows that *(3) is valid for almost all* $\mathbf{a}$, $\mathbf{b}$ *and for all* $t>0$ *also when* $N=\infty$.

Corollary III.2: Assume that in addition to (1) the potential satisfies

$$\int d^n\mathbf{a}|\mathbf{a}|^r e^{-tV(\mathbf{a})}<\infty,$$

for some $r\geqslant 0$ *and for all* $t>0$. *Then there exists functions* F_r *and* G_r, *continuous on* $(0,\infty)$, *such that for all* $t>0$ *and for all* $N<\infty$,

$$\int d^n\mathbf{a}|\mathbf{a}|^r K_N(\mathbf{a},\mathbf{a};t)\leqslant F_r(t)$$

and for all N, *including* ∞,

$$\int d^n\mathbf{a}d^n\mathbf{b}|\mathbf{a}|^r K_N(\mathbf{a},\mathbf{b};t)\leqslant G_r(t).$$

If the above assumption on the potential is true for 0 and r then it is clearly true for all values s in $(0,r)$ by the inequality $|\mathbf{a}|^s\leqslant 1+|\mathbf{a}|^r$. If the potential has no singularity in a neighborhood of the origin, it is enough to find one $r\geqslant 0$ for the statement to be true for all s with $0\leqslant s\leqslant r$.

Corollary III.3: For all N, *including* ∞, *and for all* $t>0$, $\operatorname{ess\,sup}_{\mathbf{a},\mathbf{b}\in\mathbb{R}^n}K_N(\mathbf{a},\mathbf{b};t)\leqslant C_n' t^{-n/2}$.

The convergence properties of the lattice traces are easily derived from these results; we only have to be careful that the limit of the traces of the integral kernels does give the trace of the continuum integral operator. This leads to the appearance of the factor of 2 in Eq. (4), which will be discussed more after the proof of the theorem.

Theorem III.4: Let A *be a measurable function bounded by a degree* $r\geqslant 0$ *polynomial and assume that the potential additionally satisfies*

$$\int d^n\mathbf{a}|\mathbf{a}|^{2r}e^{-tV(\mathbf{a})}<\infty,\ \text{for all}\ t>0.$$

Then for all $t>0$ *and* $N\geqslant 2$ *the operators* $\hat{A}e^{-t\hat{H}}$ *and* $\hat{A}\hat{T}(t/N)^N$ *are both Hilbert-Schmidt and trace-class and*

$$\operatorname{Tr}(\hat{A}e^{-t\hat{H}})=\lim_{k\to\infty}\operatorname{Tr}\left(\hat{A}\hat{T}\left(\frac{t}{2k}\right)^{2k}\right), \tag{4}$$

$$\operatorname{Tr}(\hat{A}\hat{T}(t/N)^N)=\int d^n\mathbf{a}\,A(\mathbf{a})K_N(\mathbf{a},\mathbf{a};t). \tag{5}$$

Proof: The requirement for A is equivalent to finding a $C\geqslant 0$ such that $|A(\mathbf{a})|\leqslant C(1+|\mathbf{a}|^r)$ for all $\mathbf{a}$. It follows now from Corollaries III.2 and III.3 that for $1\leqslant N\leqslant\infty$,

$$\int d^n\mathbf{a}d^n\mathbf{b}|A(\mathbf{a})K_N(\mathbf{a},\mathbf{b};t)|^2<\infty.$$

The operators $\hat{A}e^{-t\hat{H}}$ and $\hat{A}\hat{T}(t/N)^N$ are given by the integral kernels $A(\mathbf{a})K_N(\mathbf{a},\mathbf{b};t)$ and, therefore, they are Hilbert-Schmidt.

Since $\hat{A}e^{-t\hat{H}}$ is a product of the two Hilbert-Schmidt operators $\hat{A}e^{-(t/2)\hat{H}}$ and $e^{-(t/2)\hat{H}}$, it is trace-class. Moreover, the trace can be expressed in terms of the integral kernels of these Hilbert-Schmidt operators:

$$\mathrm{Tr}(\hat{A}e^{-t\hat{H}})=\int d^n\mathbf{a}d^n\mathbf{b}A(\mathbf{a})K_\infty\left(\mathbf{a},\mathbf{b};\frac{t}{2}\right)K_\infty\left(\mathbf{b},\mathbf{a};\frac{t}{2}\right).$$

Using Lemma III.1 and Corollary III.3 to give an upper bound for the Lebesque dominated convergence theorem shows now that

$$\begin{aligned}\int d^n\mathbf{a}d^n\mathbf{b}A(\mathbf{a})K_\infty\left(\mathbf{a},\mathbf{b};\frac{t}{2}\right)K_\infty\left(\mathbf{b},\mathbf{a};\frac{t}{2}\right)&=\lim_{N\to\infty}\int d^n\mathbf{a}d^n\mathbf{b}A(\mathbf{a})K_N\left(\mathbf{a},\mathbf{b};\frac{t}{2}\right)K_N\left(\mathbf{b},\mathbf{a};\frac{t}{2}\right)\\&=\lim_{N\to\infty}\int d^n\mathbf{a}\,A(\mathbf{a})K_{2N}(\mathbf{a},\mathbf{a};t),\end{aligned}$$

where the second equality follows from Fubini's theorem and a careful inspection of the definition of K_{2N}. Therefore,

$$\mathrm{Tr}(\hat{A}e^{-t\hat{H}})=\lim_{N\to\infty}\int d^n\mathbf{a}\,A(\mathbf{a})K_{2N}(\mathbf{a},\mathbf{a};t). \tag{6}$$

It has already been proven that both $\hat{A}\hat{T}(t)$ and $\hat{T}(t)^{N-1}$ are Hilbert-Schmidt for all $t>0$ and thus $\hat{A}\hat{T}(t/N)^N$ is trace-class and its trace can be expressed in terms of the kernels of these operators. It is then a matter of inspection to conclude

$$\mathrm{Tr}(\hat{A}\hat{T}(t/N)^N)=\int d^n\mathbf{a}\,A(\mathbf{a})K_N(\mathbf{a},\mathbf{a};t) \tag{7}$$

and the theorem follows from Eqs. (6) and (7). □

If $\lim_N K_N(\mathbf{a},\mathbf{a};t)$ exists for almost all $\mathbf{a}$, then an application of the dominated convergence theorem in (6) shows that $\mathrm{Tr}(\hat{A}e^{-t\hat{H}})=\lim_N\int d^n\mathbf{a}A(\mathbf{a})K_N(\mathbf{a},\mathbf{a};t)$. Therefore, also odd lattice sizes can be included in the limit, for instance, if V is continuous (then K_∞ is continuous and thus finite on the diagonal).

In the second theorem we will work with temperatures having an imaginary part. The following proof relies on a simple property of the Hilbert-Schmidt operators given as Lemma IV.1 in Sec. IV. To make the theorem easier to use, we will now state all the assumptions made on the potential explicitly.

Theorem III.5: Assume that A is an essentially bounded measurable function and $V\in L^1_{\mathrm{loc}}$ is a potential bounded from below for which (2) holds and

$$\int d^n\mathbf{a}e^{-tV(\mathbf{a})}<\infty,\quad\text{for all}\quad t>0.$$

Then for all complex β with $\mathrm{Re}\,\beta>0$, $\hat{A}e^{-\beta\hat{H}}$ *is trace-class and*

$$\mathrm{Tr}(\hat{A}e^{-\beta\hat{H}})=\lim_{N\to\infty}\int d^n\mathbf{a}\,A(\mathbf{a})K_{2N}(\mathbf{a},\mathbf{a};\beta).$$

Proof: Let $f_N(\beta)=\int d^n\mathbf{a}\, A(\mathbf{a})K_{2N}(\mathbf{a},\mathbf{a};\beta)$ and $g(\beta)=\mathrm{Tr}(\hat{A}e^{-\beta\hat{H}})$. Both are then well-defined on the right half-plane: f_N is an $2N$-fold integral over an absolutely integrable function and, by the Schwarz inequality, self-adjointness of $\hat{H}$ and Theorem III.4,

$$|g(\beta)|^2\leqslant \mathrm{Tr}(|\hat{A}|^2e^{-\mathrm{Re}\,\beta\hat{H}})\mathrm{Tr}(e^{-\mathrm{Re}\,\beta\hat{H}})<\infty.$$

In addition, g and all f_N are analytic on the right half-plane, as can be seen, e.g., by a standard application of Morera's theorem—note that

$$g(\beta)=\sum_n \langle\psi_n|\hat{A}\psi_n\rangle e^{-\beta E_n},$$

where $\{E_n\}$ are the eigenvalues of $\hat{H}$ and $\{\psi_n\}$ is a corresponding orthonormal basis.

Since Theorem III.4 states that $\lim_{N\to\infty} f_N(\beta)=g(\beta)$ for positive β, the theorem will follow from the Vitali convergence theorem[9] if we can only prove that the sequence (f_N) is uniformly bounded on every compact subset of the half-plane.

An argument similar to one in the proof of Theorem III.4 shows that $f_N(\beta)=\mathrm{Tr}(\hat{A}\hat{T}(\beta/2N)^{2N})$. Therefore, Lemma IV.1 establishes the bound $|f_N(\beta)|\leqslant\|\hat{A}\|\mathrm{Tr}(\hat{T}^*\hat{T})^N$, where $\hat{T}=\hat{T}(\beta/2N)$. However, as $\hat{T}^*\hat{T}=e^{-(\beta^*/2N)\hat{V}}e^{-(\mathrm{Re}\,\beta/N)\hat{H}_0}e^{-(\beta/2N)\hat{V}}$,

$$\mathrm{Tr}(\hat{T}^*\hat{T})^N=\mathrm{Tr}(\hat{T}(\mathrm{Re}\ \beta/N)^N)\leqslant F_0(\mathrm{Re}\ \beta)<\infty$$

by Corollary III.2 and Theorem III.4 If S is a compact set in the right half-plane, it follows that $F_0(\mathrm{Re}\,S)$ is bounded since F_0 was continuous. This proves that (f_N) is uniformly bounded on every compact subset of the right half-plane and hence completes the proof. □

If V is continuous, then the limit can be taken by using all values of N: Define $h_N(\beta)=\int d^n\mathbf{a}\, A(\mathbf{a})K_N(\mathbf{a},\mathbf{a};\beta)$. By the remark after Theorem III.4, $\lim_N h_N=g$ on the positive real axis. Also,

$$|h_{2N+1}(\beta)|=\left|\mathrm{Tr}\left(\hat{A}\hat{T}\left(\frac{\beta}{2N+1}\right)^{2N+1}\right)\right|\leqslant\|\hat{A}\|\|\hat{T}\|\mathrm{Tr}(\hat{T}^*\hat{T})^N\leqslant\|\hat{A}\|F_0\left(\mathrm{Re}\ \beta\frac{2N}{2N+1}\right),$$

and thus the sequence (h_N) is uniformly bounded on the right half-plane.

IV. PROOFS OF THE LEMMAS

Proof of Lemma III.1: By the relation between geometric and arithmetic means, $\exp(-(t/N)\sum_{k=1}^N V(\mathbf{x}_k))\leqslant(1/N)\sum_{k=1}^N\exp(-tV(\mathbf{x}_k))$. Using this and the semigroup property of the integral kernels P, we arrive at the inequality

$$K_N(\mathbf{a},\mathbf{b};t)\leqslant\frac{1}{N}e^{-tV(\mathbf{b})}P(\mathbf{a},\mathbf{b};t)+\int d^n\mathbf{c}e^{-tV(\mathbf{c})}\frac{1}{N}\sum_{k=1}^{N-1}P\left(\mathbf{a},\mathbf{c};k\frac{t}{N}\right)P\left(\mathbf{c},\mathbf{b};(N-k)\frac{t}{N}\right).$$

The terms in the last sum are explicitly

$$\begin{aligned}&\left(\frac{1}{2\pi t}\frac{N}{\sqrt{k(N-k)}}\right)^n\exp\left[-\frac{1}{2t}\left(\frac{N}{k}|\mathbf{a}-\mathbf{c}|^2+\frac{N}{N-k}|\mathbf{c}-\mathbf{b}|^2\right)\right]\\&\quad=P(\mathbf{a},\mathbf{c};t)P(\mathbf{c},\mathbf{b};t)\frac{N}{\sqrt{k(N-k)}}\left(\frac{N-k}{k}\right)^{(n-1)/2}\exp\left[-\frac{N-k}{k}\frac{|\mathbf{a}-\mathbf{c}|^2}{2t}\right]\\&\qquad\times\left(\frac{N}{N-k}\right)^{n-1}\exp\left[-\frac{k}{N-k}\frac{|\mathbf{c}-\mathbf{b}|^2}{2t}\right]\end{aligned}\tag{8}$$

$$= P(\mathbf{a},\mathbf{c};t)P(\mathbf{c},\mathbf{b};t)\frac{N}{\sqrt{k(N-k)}}\left(\frac{k}{N-k}\right)^{(n-1)/2}\exp\left[-\frac{k}{N-k}\frac{|\mathbf{c}-\mathbf{b}|^2}{2t}\right]$$
$$\times\left(\frac{N}{k}\right)^{n-1}\exp\left[-\frac{N-k}{k}\frac{|\mathbf{a}-\mathbf{c}|^2}{2t}\right]. \tag{9}$$

Assume now that $n>1$ and consider first the case $1\leqslant k\leqslant N/2$. Since, $t^s e^{-tr}\leqslant(s/re)^s$ for all positive r, s and t, the expression (8) has an upper bound

$$P(\mathbf{a},\mathbf{c};t)P(\mathbf{c},\mathbf{b};t)\frac{N}{\sqrt{k(N-k)}}2^{n-1}\left(\frac{(n-1)t}{e|\mathbf{a}-\mathbf{c}|^2}\right)^{(n-1)/2}.$$

On the other hand, if $N/2\leqslant k\leqslant N-1$, then the same reasoning can be applied to (9) leading to the same bound apart from the change of $|\mathbf{a}-\mathbf{c}|$ to $|\mathbf{c}-\mathbf{b}|$. Since in both cases these upper bounds are positive, the terms under inspection are for all $1\leqslant k\leqslant N-1$ bounded by

$$P(\mathbf{a},\mathbf{c};t)P(\mathbf{c},\mathbf{b};t)\left(\frac{4(n-1)}{e}\right)^{(n-1)/2}t^{(n-1)/2}\left(\frac{1}{|\mathbf{a}-\mathbf{c}|^{n-1}}+\frac{1}{|\mathbf{c}-\mathbf{b}|^{n-1}}\right)\frac{N}{\sqrt{k(N-k)}}.$$

Now $(1/N)\sum_{k=1}^{N-1}[k/N(1-(k/N))]^{-1/2}$ is a lower Riemann sum for the integral $\int_0^1 ds\,[1/\sqrt{s(1-s)}]=\pi$ and, therefore for all N,

$$\frac{1}{N}\sum_{k=1}^{N-1}\frac{N}{\sqrt{k(N-k)}}\leqslant\pi \tag{10}$$

and the Lemma follows if we choose $C_n=\pi 2^{n-1}(n-1)^{(n-1)/2}e^{-(n-1)/2}$.

The case $n=1$ is even simpler to prove, as then both of the exponentials in (8) can be majorized by one and the Lemma, with the choice $C_1=\pi$, follows from (10). □

Proofs of Corollaries III.2 and III.3: Define the following two functions on $(0,\infty)$

$$I_s(t)\coloneqq\int d^n\mathbf{a}|\mathbf{a}|^s e^{-tV(\mathbf{a})},\quad J_s(t)\coloneqq\int d^n\mathbf{a}|\mathbf{a}|^s P(\mathbf{0},\mathbf{a};t),$$

where s is a real parameter. J_s is finite for $s>-n$ and, in fact, it is then continuous, since a change of variables and relying on n-dimensional spherical coordinates shows that

$$J_s(t)=(2t)^{s/2}\frac{\Gamma((n+s)/2)}{\Gamma(n/2)}. \tag{11}$$

Under the assumptions of the theorem, also I_0 and I_r are everywhere finite. They are then also continuous, since for every converging sequence (t_k) in $(0,\infty)$ dominated convergence can be used to prove $\lim_{k\to\infty}I_s(t_k)=I_s(\lim t_k)$.

The rest is just a matter of bookkeeping when integrating both sides of the inequality (3). Using Jensen's inequality for $r\geqslant 1$ and the obvious relation $(1+|\mathbf{a}|)^r-|\mathbf{a}|^r\leqslant 1$ for $0\leqslant r<1$ it is easy to prove that for $r\geqslant 0$ and $\mathbf{a}$, $\mathbf{b}\in\mathbb{R}^n$

$$|\mathbf{a}|^r\leqslant 2^r(|\mathbf{a}-\mathbf{b}|^r+|\mathbf{b}|^r).$$

Applying this result, Eq. (11) and a little algebra shows that

$$\int d^n\mathbf{a}|\mathbf{a}|^r K_N(\mathbf{a},\mathbf{a};t)\leqslant C^{(1)}t^{-n/2}I_r(t)+C^{(2)}t^{(r-n)/2}I_0(t),$$

$$\int d^n\mathbf{a}d^n\mathbf{b}|\mathbf{a}|^r K_N(\mathbf{a},\mathbf{b};t)\leqslant C^{(3)}I_r(t)+C^{(4)}t^{r/2}I_0(t),$$

where the constants $C^{(k)}$ depend only on n and r. The functions on the right hand sides are continuous and can therefore be chosen as F_r and G_r.

Since $e^{-tV(\mathbf{c})}P(\mathbf{c},\mathbf{b};t)\leqslant(2\pi t)^{-n/2}$, Corollary III.3 is an immediate consequence of (11) and Lemma III.1. □

Lemma IV.1: If $\hat{A}$ is bounded, $\hat{T}$ is Hilbert-Schmidt and $N\geqslant 1$, then

$$|\mathrm{Tr}(\hat{A}\hat{T}^{2N})|\leqslant\|\hat{A}\|\mathrm{Tr}(\hat{T}^*\hat{T})^N.$$

Proof: By the Schwarz inequality,

$$|\mathrm{Tr}(\hat{A}\hat{T}^{2N})|^2\leqslant\mathrm{Tr}[(\hat{A}\hat{T}^N)^*(\hat{A}\hat{T}^N)]\mathrm{Tr}[(\hat{T}^N)^*\hat{T}^N]$$

and since $\mathrm{Tr}[(\hat{A}\hat{T}^N)^*(\hat{A}\hat{T}^N)]=\mathrm{Tr}[|\hat{A}|^2\hat{T}^N(\hat{T}^N)^*]\leqslant\|\hat{A}\|^2\mathrm{Tr}[(\hat{T}^N)^*\hat{T}^N]$, we get

$$|\mathrm{Tr}(\hat{A}\hat{T}^{2N})|\leqslant\|\hat{A}\|\mathrm{Tr}[(\hat{T}^N)^*\hat{T}^N].$$

Using a method similar to the so called min-max principle, Fan has shown[10] that for any finite-dimensional operator (i.e., matrix) $\hat{B}$, $\mathrm{Tr}[(\hat{B}^N)^*\hat{B}^N]\leqslant\mathrm{Tr}(\hat{B}^*\hat{B})^N$. The finite-dimensional operators are dense in $\mathscr{T}_2=$the Hilbert space of Hilbert–Schmidt operators with the inner product $(\cdot,\cdot)_{\mathrm{HS}}$. Since the product, the inner product and the norm are continuous on $\mathscr{T}_2$ and $\mathrm{Tr}[(\hat{B}^N)^*\hat{B}^N]=\|\hat{B}^N\|_{\mathrm{HS}}^2$, $\mathrm{Tr}(\hat{B}^*\hat{B})^N=(\hat{B}^*\hat{B},(\hat{B}^*\hat{B})^{N-1})_{\mathrm{HS}}$, it is obvious that the result of Fan is valid for every Hilbert−Schmidt operator. □

V. DISCUSSION

The main results of this paper are the two theorems, III.4 and III.5. The first one justifies the faith of the lattice community that also expectation values of polynomials are measurable with the lattice Monte Carlo methods if the potential increases sufficiently fast at infinity. The second one establishes the possibility of using the lattice integrals for computations of the expectation values even when the temperature has an imaginary part.

There are a few shortcomings of the present results. First, the exact class of potentials for which the integral kernels K_N can be used was not determined. This is not a serious problem since, by accepting the necessity for a double limit, statistics given by any potential in L^1_{loc} can be approached with the lattice integrals by replacing V by a sequence $V_k\in C^\infty$ which converges to it pointwise almost everywhere. The second shortcoming is the class of observables in Theorem III.5: the present proof cannot be easily extended to include also polynomial observables for polynomial potentials, although it is very natural to expect the integrals to converge to the right limit by Theorem III.4. Finally, the necessity of using only an even number of lattice points has no natural explanation and as such it is expected to be an artifact of the present approach. Indeed, as has been noted earlier, the requirement of even lattice sizes can be dropped for continuous potentials.

ACKNOWLEDGMENTS

I would like to thank A. Kupiainen for enlightening discussions and comments and for suggesting several useful references. I am also grateful to M. Chaichian and A. Demichev for their comments.

[1] R. Cameron, J. Anal. Math. **10**, 287 (1962/63).
[2] J. R. Klauder, Phys. Rev. A **29**, 2036 (1984).
[3] G. Parisi, Phys. Lett. **B131**, 393 (1983).
[4] J. Lukkarinen, ''Lattice Simulations of the Quantum Microcanonical Ensemble,'' HIP preprint 1997-15/TH, also appears in hep-th/9704103.
[5] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).
[6] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[7] Theorem 9.2 of Simon's book (Ref. 6), gives a path-integral exposition of the Golden-Thompson-Symanzik formula. The original references are S. Golden, Phys. Rev. **137**, B1127 (1965); C. J. Thompson, J. Math. Phys. **6**, 1812 (1965); K. Symanzik, *ibid.* **6**, 1155 (1965).

[8] K. Symanzik, J. Math. Phys. **6**, 1155 (1965).
[9] E. C. Titchmarsh, *The Theory of Functions*, 2nd ed. (Oxford University Press, London, 1939), pp. 168–170. The form in which we have used the theorem can be found from M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis* (Academic, New York, 1972), problem 33 on page 35.
[10] K. Fan, Proc. Natl. Acad. Sci. USA **35**, 652 (1949). See also H. Weyl, *ibid.* **35**, 408 (1949).

Generalized plasma dispersion functions

D. J. Percival[a] and P. A. Robinson[b]
School of Physics, The University of Sydney, New South Wales 2006, Australia

(Received 29 August 1997; accepted for publication 12 December 1997)

Generalized plasma dispersion functions have recently arisen in the derivation of an exact expression for the quadratic response tensor for a plasma with constituent species having Maxwellian velocity distributions. A range of mathematical properties satisfied by the generalized plasma dispersion functions are derived here, including recursion relations, series expansions, and approximate analytic forms, including Padé approximants. These properties are of central importance to the application of the exact quadratic response tensor to descriptions of second-order processes in weak and strong turbulence theory, such as three-wave interactions in warm plasmas. © *1998 American Institute of Physics.* [S0022-2488(98)02807-2]

I. INTRODUCTION

The plasma dispersion function[1] is widely used in descriptions of linear wave processes in collisionless warm plasmas, where constituent species satisfy a Maxwellian velocity distribution. The plasma dispersion function $Z(u)$ is usually defined for complex u by the integral expression

$$Z(u)=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\frac{dt\ e^{-t^2}}{t-u}, \tag{1}$$

for $\mathrm{Im}(u)=0$ and by the analytic continuation of (1) for $\mathrm{Im}(u)<0$. $Z(u)$ is closely related to the error function erf, with

$$Z(u)=i\sqrt{\pi}e^{-u^2}[1+\mathrm{erf}(iu)], \tag{2}$$

for all complex u. The linear response tensor, describing the linear plasma response to an electromagnetic disturbance, can be expressed in terms of $Z(u)$. The properties of the plasma dispersion function have been widely studied in the literature,[1–3] and are often applied in analytic treatments of wave processes in both space and laboratory plasmas (e.g., Refs. 4, 5 and the references cited therein).

Recently, an exact expression for the quadratic response tensor for a collisionless Maxwellian plasma was derived.[6] The quadratic response tensor describes all second-order wave processes (i.e., interactions between triads of waves, including coalescence and decay), and its exact representation permits a complete analytic description of processes previously only treated with relatively simple approximations.[7–10] An important point is that the validity of widely used approximate forms can now be assessed by comparison with the exact expression.

An important second-order wave process is the coalescence of two waves to give a third (product) wave; the reverse interaction describes a three-wave decay process. If two waves with frequencies ω_1, ω_2 and wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$ coalesce, then the product wave has a frequency ω and wave vector $\mathbf{k}$ which satisfies the wave matching conditions

$$\omega_1+\omega_2=\omega, \quad \mathbf{k}_1+\mathbf{k}_2=\mathbf{k}, \tag{3}$$

[a]Present address: Wide Area Surveillance Division, Defence Science and Technology Organisation, P.O. Box 1500, Salisbury S.A. 5108, Australia. Electronic mail: John.Percival@dsto.defence.gov.au
[b]Electronic mail: robinson@physics.usyd.edu.au

 © 1998 American Institute of Physics

imposed by the conservation of wave energy and momentum, respectively. The strength of wave coupling is described by the quadratic response tensor, which is a function of the frequencies and wave vectors of the interacting waves.

A key step in deriving the exact quadratic response tensor for a Maxwellian plasma was the introduction of the *generalized* plasma dispersion function $Y_{m,n}$, $m,n \geqslant 0$, defined by[6]

$$Y_{m,n}(a,b,c) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} dx\ e^{-x^2} \frac{x^m}{(x-c)^n} Z(a+bx), \quad m,n \geqslant 0, \tag{4}$$

for complex arguments a, b, c. This function is a generalization of the familiar plasma dispersion function $Z(u)$ of Ref. 1, in the sense that it also involves resonant integrals over velocity distributions; moreover, the lowest-order function $Y_{0,0}(a,b,c) = Y_{0,0}(a,b)$ reduces to $Z(a)$ in the limit $b \to 0$. The exact quadratic response tensor is given by a linear combination of $Z(u)$ and generalized plasma dispersion functions $Y_{m,n}$ of orders $0 \leqslant m,n \leqslant 3$. The derivation of mathematical properties satisfied by the $Y_{m,n}$ is the subject of this paper; these properties are of central importance in the application of the exact quadratic response tensor to particular physical processes.

In the treatment of a three-wave process satisfying (3) using the exact quadratic response tensor, the arguments a, b, and c of the generalized plasma dispersion functions $Y_{m,n}(a,b,c)$ have physical meanings. In particular, the magnitudes of the parameters a and c are determined by the ratio of the wave phase speed to plasma thermal speed for a coalescing wave and a product wave, respectively, while $b = -\cot\theta$ where θ is the angle between the coalescing wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$. Approximate forms of the $Y_{m,n}(a,b,c)$ suitable for application to particular interacting waves and geometries can be derived from the exact quadratic response tensor.

In this paper we derive a range of properties of the generalized plasma dispersion functions, as well as discussing approximations and the stability of numerical evaluations. Useful recursive properties of the generalized plasma dispersion functions are derived in Sec. II, and some partial differential equations satisfied by the $Y_{m,n}$ are written down. In Sec. III, the $Y_{m,n}$ are evaluated as the sum of terms involving Hermite polynomials and derivatives of the plasma dispersion function $Z(u)$. The real and imaginary parts of the generalized plasma dispersion functions for real arguments are identified; the latter are useful in discussions of plasma instabilities, for example. Some new results involving the nth derivative of the plasma dispersion function $Z^{(n)}$ are derived in Sec. IV. The results of a numerical investigation of the $Y_{0,n}$ for $n \leqslant 3$ and real arguments are also presented. In Sec. V, several useful analytical approximations, including Padé approximants for the lower-order $Y_{m,n}(a,b,c)$ are developed.

II. SYMMETRIES AND DERIVATIVE IDENTITIES

In this section, we derive recursive properties of the $Y_{m,n}$, as well as some differential equations which they obey. The generalized plasma dispersion functions (4) satisfy several symmetry properties and derivative identities, analogous with those satisfied by the plasma dispersion function of Ref. 1. To begin, since

$$\frac{x^m}{(x-c)^n} = \frac{x^{m-1}}{(x-c)^{n-1}} + \frac{cx^{m-1}}{(x-c)^n}, \quad m,n \geqslant 1, \tag{5}$$

it follows from the definition (4) that the recursion relation

$$Y_{m,n}(a,b,c) = Y_{m-1,n-1}(a,b,c) + cY_{m-1,n}(a,b,c), \quad m,n \geqslant 1, \tag{6}$$

is satisfied by $Y_{m,n}$. Therefore, only functions of the form $Y_{0,n}$, with $n \geqslant 0$, require explicit evaluation. The quadratic response tensor derived in Ref. 6 can then be expressed using $Z(u)$ and the generalized plasma dispersion functions $Y_{0,n}$, with $0 \leqslant n \leqslant 3$.

Using the symmetry[2,3]

$$Z^{(n)}(u^*) = (-1)^{n+1}[Z^{(n)}(-u)]^*, \quad n \geqslant 0, \tag{7}$$

for the nth derivative of $Z(u)$, denoted $Z^{(n)}(u)$, where it is understood that $Z^{(0)}(u)=Z(u)$, it can be shown by substitution into the definition (4) that

$$Y_{0,n}(a^*,b^*,c^*)=(-1)^{n+1}[Y_{0,n}(-a,b,-c)]^*, \quad n\geqslant 0. \tag{8}$$

The general result,

$$Y_{m,n}(a^*,b^*,c^*)=(-1)^{m+n+1}[Y_{m,n}(-a,b,-c)]^*, \quad m,n\geqslant 0, \tag{9}$$

can be proved by induction using (8) and the recursion relation (6). Other symmetry properties which can be proved using (7) in the definition (4) are

$$Y_{m,n}(a,b,c)=(-1)^{m+n}Y_{m,n}(a,-b,-c), \quad m,n\geqslant 0, \tag{10}$$

and

$$Y_{m,n}(a^*,b^*,c^*)=-[Y_{m,n}(-a,-b,c)]^*, \quad m,n\geqslant 0. \tag{11}$$

It is straightforward to show by direct partial differentiation of the definition (4) that the generalized plasma dispersion function satisfies the first-order partial differential equations

$$\frac{\partial Y_{m,n}}{\partial a}=\frac{\partial Y_{m-1,n}}{\partial b}, \quad m\geqslant 1, \quad n\geqslant 0, \tag{12}$$

and

$$\frac{\partial Y_{m,n}}{\partial c}=nY_{m,n+1}, \quad m,n\geqslant 0, \tag{13}$$

where the arguments of $Y_{m,n}(a,b,c)$ have been omitted for brevity. Using (12) and (13), together with the recursion relation (6) and the ordinary differential equation

$$\frac{d^2Z(u)}{du^2}+2u\,\frac{dZ(u)}{du}+2Z(u)=0, \tag{14}$$

satisfied by the plasma dispersion function,[1] it can be shown that $Y_{m,n}(a,b,c)$ is a solution of the partial differential equations

$$\frac{\partial^2 Y_{m,n}}{\partial a^2}+2a\,\frac{\partial Y_{m,n}}{\partial a}+2b\,\frac{\partial Y_{m,n}}{\partial b}+2Y_{m,n}=0, \tag{15}$$

and

$$\frac{\partial^2 Y_{m,n}}{\partial b\,\partial c}=n\,\frac{\partial Y_{m,n}}{\partial a}+c\,\frac{\partial^2 Y_{m,n}}{\partial a\,\partial c}. \tag{16}$$

Partial derivatives of the generalized plasma dispersion functions appear in the exact quadratic response tensor derived in Ref. 6, and are easily evaluated using (12).

III. EVALUATION OF $Y_{0,n}(a,b,c)$

In this section, we express the $Y_{0,n}$, $n\geqslant 0$ in terms of series involving the Hermite polynomials and derivatives of the standard plasma dispersion function Z. From the recursion relation (6), it is straightforward to show that the generalized plasma dispersion functions of arbitrary order can be expressed as linear combinations of the lowest-order functions of the form $Y_{0,n}$, $n\geqslant 0$. Thus only the functions $Y_{0,n}$ require explicit evaluation.

A. Evaluation of $Y_{0,0}(a,b)$

The function $Y_{0,0}(a,b,c)=Y_{0,0}(a,b)$ is independent of c and may be evaluated by substituting the plasma dispersion function in the form (1) into the definition (4) with $m=n=0$ to give

$$Y_{0,0}(a,b)=\frac{1}{\pi}\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dt\ e^{-x^2-t^2}\ \frac{1}{t-a-bx}. \tag{17}$$

Transforming to polar coordinates with $x=r\cos\theta$, $t=r\sin\theta$, and $dx\,dt=r\,dr\,d\theta$, the θ integration is of an elementary type and can be performed immediately. The remaining integral over radius r can be rewritten in terms of the plasma dispersion function to give

$$Y_{0,0}(a,b)=\frac{1}{\sqrt{1+b^2}}\,Z\left(\frac{a}{\sqrt{1+b^2}}\right),\quad \mathrm{Im}(a)\geqslant 0, \tag{18}$$

and the analytic continuation of (18) when $\mathrm{Im}(a)<0$.

B. Series expressions for $Y_{0,n}$, $n\geqslant 1$

The function $Y_{0,1}(a,b,c)$ can be evaluated by expanding the plasma dispersion function in the definition (4) in a power series about $x=c$ to give

$$Y_{0,1}(a,b,c)=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}dx\ e^{-x^2}\sum_{n=0}^{\infty}\ \frac{b^n}{n!}\,(x-c)^{n-1}Z^{(n)}(a+bc), \tag{19}$$

where $Z^{(n)}$ denotes the nth derivative of the plasma dispersion function. Taking the integral inside the summation and using a standard integral [Ref. 11, Eq. (3.462.4)] gives

$$\int_{-\infty}^{\infty}dx(x-c)^n e^{-x^2}=\sqrt{\pi}(2i)^{-n}H_n(-ic),\quad n\geqslant 0, \tag{20}$$

where H_n is a Hermite polynomial of integer order n; then

$$Y_{0,1}(a,b,c)=Z(c)Z(a+bc)+\sum_{n=1}^{\infty}\ \frac{b^n}{n!(2i)^{n-1}}\,H_{n-1}(-ic)Z^{(n)}(a+bc). \tag{21}$$

In writing (21), the leading term $n=0$ has been written explicitly, identifying $Z(c)$ using (1). The evaluation of $Y_{0,2}(a,b,c)$ proceeds by analogy to give

$$Y_{0,2}(a,b,c)=bZ(c)Z^{(1)}(a+bc)+Z^{(1)}(c)Z(a+bc)+\sum_{n=2}^{\infty}\ \frac{b^n}{n!(2i)^{n-2}}\,H_{n-2}(-ic)Z^{(n)}(a+bc). \tag{22}$$

In general, if $b\neq 0$, then

$$Y_{0,n}(a,b,c)=\sum_{k=0}^{n-1}\ \frac{b^k}{k!(n-k-1)!}\,Z^{(n-k-1)}(c)Z^{(k)}(a+bc)+\sum_{k=n}^{\infty}\ \frac{b^k}{k!(2i)^{k-n}}$$
$$\times H_{k-n}(-ic)Z^{(k)}(a+bc),\quad n\geqslant 1. \tag{23}$$

Otherwise,

$$Y_{0,n}(a,0,c)=\frac{1}{(n-1)!}\,Z^{(n-1)}(c)Z(a),\quad n\geqslant 1. \tag{24}$$

In writing (23) and (24), the definition

$$Z^{(n)}(u)=\frac{n!}{\sqrt{\pi}}\int_{-\infty}^{\infty}\frac{dx\ e^{-x^2}}{(x-u)^{n+1}},\quad \mathrm{Im}(u)>0, \tag{25}$$

and its analytic continuation for $\mathrm{Im}(u)\leqslant 0$, is required. To our knowledge, (25) has not appeared explicitly in the literature, although it is easily established by repeated differentiation of (1).

The infinite series in (23) is convergent for all a and c provided $|b|<1$. This result can be established by expressing the $Z^{(n)}$ in the series using the alternative integral expression[2]

$$Z^{(n)}(u)=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}dx\ \frac{(-1)^n H_n(x)e^{-x^2}}{x-u}, \tag{26}$$

so that the infinite series in (23) becomes

$$\sum_{k=n}^{\infty}\frac{b^k}{k!(2i)^{k-n}}H_{k-n}(-ic)Z^{(k)}(a+bc)=\frac{(-b)^n}{\sqrt{\pi}}\int_{-\infty}^{\infty}\frac{dx\ e^{-x^2}}{x-a-bc}\sum_{k=0}^{\infty}\frac{1}{(k+n)!}\left(\frac{-b}{2i}\right)^k \times H_k(-ic)H_{k+n}(x),\quad n\geqslant 1. \tag{27}$$

In writing (27), the integral and summation on the right side have been interchanged, and the sum index redefined to begin at $k=0$. From Ref. 12 [Eq. (5.12.2.1)], the sum

$$\sum_{k=0}^{\infty}\frac{t^k}{k!}H_k(u)H_{k+n}(v)=(1-4t^2)^{-(n+1)/2} \times\exp\left[\frac{4tuv-4t^2(u^2+v^2)}{1-4t^2}\right]H_n\left(\frac{v-2tu}{\sqrt{1-4t^2}}\right),\quad |t|<1/2, \tag{28}$$

can be derived. The absolute value of the nth term in the convergent series (28) with $t=-b/2i$, $u=-ic$, $v=x$ is an upper bound for the corresponding term in (27). Therefore the series in (23) converges absolutely whenever $|b|<1$.

When $|b|>1$, an alternative convergent series expression for the $Y_{0,n}$ is required. Substituting the plasma dispersion function definition (1) into (4) for the case $m=0$, $n=1$ gives

$$Y_{0,1}(a,b,c)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{dx\ e^{-x^2}}{x-c}\int_{-\infty}^{\infty}\frac{dy\ e^{-y^2}}{y-(a+bx)}, \tag{29}$$

which may be written as

$$Y_{0,1}(a,b,c)=\frac{1}{\pi}\int dy\ \frac{e^{-y^2}}{y-(a+bc)}\int dx\ e^{-x^2}\left[\frac{1}{x-c}-\frac{1}{x-(y-a)/b}\right]. \tag{30}$$

The x integration may be performed in terms of plasma dispersion functions; the y integration follows from the definition (4) to give the identity

$$Y_{0,1}(a,b,c)=Z(c)Z(a+bc)-Y_{0,1}(-a/b,1/b,a+bc). \tag{31}$$

Comparing (31) with the series expression (21), and making the transformations $-a/b\rightarrow a$, $1/b\rightarrow b$, and $a+bc\rightarrow c$, leads to the alternative series expression

$$Y_{0,1}(a,b,c)=-\sum_{n=1}^{\infty}\frac{b^{-n}}{n!(2i)^{n-1}}H_{n-1}(-i(a+bc))Z^{(n)}(c). \tag{32}$$

Again using a term-by-term comparison test with the convergent series (28), it can be shown that (32) is convergent whenever $|b|>1$. By analogy, one can derive for the general case

$$Y_{0,n}(a,b,c) = -b^{n-1}\sum_{k=n}^{\infty} \frac{b^{-k}}{k!(2i)^{k-n}} H_{k-n}(-i(a+bc))Z^{(k)}(c), \quad n \geqslant 1. \tag{33}$$

C. Real and imaginary parts of $Y_{0,n}(a,b,c)$

It is possible to separate the generalized plasma dispersion function $Y_{0,n}(a,b,c)$ explicitly into real and imaginary parts when its arguments are real. This is analogous to the well-known separation into real and imaginary parts of the plasma dispersion function $Z(u)$ for real u, given by[1]

$$Z(u) = \text{Re } Z(u) + i\sqrt{\pi}e^{-u^2}, \tag{34}$$

with

$$\text{Re } Z(u) = -2e^{-u^2}\int_0^u dt\ e^{t^2}. \tag{35}$$

From the definition (4) of the generalized plasma dispersion function with $m=0$, $n=1$, $Y_{0,1}(a,b,c)$ can be written as

$$Y_{0,1}(a,b,c) = \frac{1}{\sqrt{\pi}} \int dx\ e^{-x^2}[\mathscr{P}[1/(x-c)] + i\pi\delta(x-c)][\text{Re } Z(a+bx) + i\sqrt{\pi}e^{-(a+bx)^2}], \tag{36}$$

where $\mathscr{P}[1/(x\text{-}c)]$ denotes the Cauchy principal value of the integral at the pole $x=c$, and where the real and imaginary parts of the plasma dispersion function have been written explicitly using (34). Multiplying out the integrand in (36) and performing the two integrals over the delta function leaves two integrals to be evaluated. The integral $\mathscr{P}\int dx\ e^{-x^2}[1/(x-c)]e^{-(a+bx)^2}$ can be evaluated in terms of Re $Z(a+bx)$ using the substitution $y = \sqrt{1+b^2}[x + ab/(1+b^2)]$. The resulting separation gives

$$\text{Re } Y_{0,1}(a,b,c) = \frac{1}{\sqrt{\pi}}\mathscr{P}\int dx\ e^{-x^2}\frac{\text{Re } Z(a+bx)}{x-c} - \text{Im } Z(c)\text{Im } Z(a+bc), \tag{37}$$

and

$$\text{Im } Y_{0,1}(a,b,c) = \text{Im } Z\left(\frac{a}{\sqrt{1+b^2}}\right)\text{Re } Z\left(\frac{b(a+bc)+c}{\sqrt{1+b^2}}\right) + \text{Im } Z(c)\text{Re } Z(a+bc). \tag{38}$$

The above result can be generalized to $Y_{0,n}(a,b,c)$ with $n \geqslant 2$ by reexpressing the $Z^{(n)}$ via repeated differentiation of (34), to give

$$\text{Re } Y_{0,n}(a,b,c) = \frac{1}{\sqrt{\pi}}\mathscr{P}\int dx\ e^{-x^2}\frac{\text{Re } Z(a+bx)}{(x-c)^n} - \frac{1}{(n-1)!}[\text{Im } Z(a+bc)\text{Im } Z^{(n-1)}(c) + b^{n-1}\ \text{Im } Z(c)\text{Im } Z^{(n-1)}(a+bc)], \tag{39}$$

and

$$\text{Im } Y_{0,n}(a,b,c) = \frac{1}{(n-1)!}\left[\text{Im } Z^{(n-1)}(c)\text{Re } Z(a+bc) + b^{n-1}\ \text{Im } Z(c)\text{Re } Z^{(n-1)}(a+bc) + (\sqrt{1+b^2})^{n-1}\ \text{Im } Z\left(\frac{a}{\sqrt{1+b^2}}\right)\text{Re } Z^{(n-1)}\left(\frac{b(a+bc)+c}{\sqrt{1+b^2}}\right)\right], \tag{40}$$

for $n \geqslant 2$.

It is instructive to compare Im $Y_{0,n}(a,b,c)$ given by (38) and (40) with the series expressions (23) and (33) for $Y_{0,n}(a,b,c)$ calculated in the previous section. It is clear that for real arguments a, b, c, the imaginary part of the series solution in (23) and (33) has been explicitly summed. Therefore only the real part of the generalized plasma dispersion function needs to be approximated analytically or numerically to be useful in applications to nonlinear wave interactions involving weakly damped wave modes.

IV. DERIVATIVES OF THE PLASMA DISPERSION FUNCTION

Some analytic properties of the nth derivative of the plasma dispersion function $Z^{(n)}(u)$ have been considered in Refs. 2 and 3. The motivation for that work was the development of a systematic computational tool for the calculation of plasma dispersion by expanding linear dispersion integrals in terms of a set of orthogonal polynomials. The expansion of the dispersion integrals for a Maxwellian plasma in terms of Hermite polynomials led to a consideration of $Z^{(n)}$. The derivation given in Ref. 6 of the exact quadratic response tensor in terms of generalized plasma dispersion functions $Y_{m,n,}$ which may themselves be expressed as series in $Z^{(n)}$, motivates us to reexamine the derivatives of the plasma dispersion function. In this section, some new useful analytic results are presented. The numerical computation of $Z^{(n)}(u)$ for complex u is considered in Appendix A.

Repeated differentiation of $Z(u)$ expressed in the form (2) gives

$$Z^{(n)}(u)=(-1)^n H_n(u)Z(u)+2(-1)^n\sum_{k=1}^{n}\left\langle \begin{matrix} n \\ k \end{matrix}\right\rangle i^{k+1}H_{n-k}(u)H_{k+1}(iu). \tag{41}$$

To our knowledge, (41) has not appeared in the literature before, and is obtained by using Rodrigues's formula,

$$H_n(u)=(-1)^n e^{u^2}\frac{d^n}{du^n}\,e^{-u^2}, \tag{42}$$

and the identity [Ref. 13, Eq. (7.1.19)]

$$\frac{d^{n+1}}{du^{n+1}}\,\mathrm{erf}(u)=(-1)^n\,\frac{2}{\sqrt{\pi}}\,H_n(u)e^{-u^2}. \tag{43}$$

The identity (41) makes explicit the close relationship between the $Z^{(n)}$ and the Hermite polynomials; their respective three-term recursion relations differ only by a sign in one term. For real argument u, all terms in the finite sum in (41) are real, since $H_k(u)$ when k is an even or odd integer, is a linear combination of only even or odd powers of u, respectively, so that $i^k H_k(iu)$ is always real for real u. It follows that

$$\mathrm{Im}\ Z^{(n)}(u)=(-1)^n H_n(u)\mathrm{Im}\ Z(u)=i\sqrt{\pi}e^{-u^2}(-1)^n H_n(u), \tag{44}$$

and that only the real part of $Z^{(n)}(u)$ presents computational difficulties for real u.

An alternative expression for $Z^{(n)}$ in terms of the repeated integral of the complementary error function $\mathrm{erfc}(u)=[1-\mathrm{erf}(u)]$, defined by

$$I^n\ \mathrm{erfc}(x)=\int_x^\infty I^{n-1}\ \mathrm{erfc}(x')dx';\quad I^0\ \mathrm{erfc}(x)=\mathrm{erfc}(x), \tag{45}$$

is given by[2]

$$Z^{(n)}(u)=i\sqrt{\pi}(2i)^n n!e^{-u^2}I^n\ \mathrm{erfc}(-iu), \tag{46}$$

which can be obtained using Ref. 13 [Eq. (7.2.9)] to rewrite the nth derivative of (2). Note that $I^n \operatorname{erfc}(x)$ may be rewritten in terms of confluent hypergeometric functions or parabolic cylinder functions,[13] so the nth derivative of the plasma dispersion function is also related to these special functions.

The power series and asymptotic expansions of $Z^{(n)}$ can be derived by taking the nth derivative of the corresponding expansions of the plasma dispersion function.[1] This yields

$$Z^{(n)}(u)=(-1)^n H_n(u) i\sqrt{\pi} e^{-u^2} - u^{1-n} \sum_{j=j_{\min}}^{\infty} \frac{(-1)^j \sqrt{\pi}\Gamma(2j+2)u^{2j}}{\Gamma(j+3/2)\Gamma(2j+2-n)}, \quad |u| \ll 1, \tag{47}$$

with

$$j_{\min}=\begin{cases} (n-1)/2, & n \text{ odd}, \\ n/2, & n \text{ even}, \end{cases} \tag{48}$$

and

$$Z^{(n)}(u)=(-1)^n H_n(u) \sigma i\sqrt{\pi} e^{-u^2} - \frac{(-1)^n}{u^{n+1}} \sum_{j=0}^{\infty} \frac{\Gamma(j+\frac{1}{2})\Gamma(2j+n+1)}{\sqrt{\pi}\Gamma(2j+1)u^{2j}}, \quad |u| \gg 1, \tag{49}$$

with

$$\sigma=\begin{cases} 0, & \operatorname{Im}(u)>0, \\ 1, & \operatorname{Im}(u)=0, \\ 2, & \operatorname{Im}(u)<0. \end{cases} \tag{50}$$

In deriving the imaginary parts of (47) and (49), Rodrigues's formula (42) is used. The asymptotic and power series expansions are useful in practice to obtain approximate forms for the generalized plasma dispersion functions $Y_{m,n}$.

Simple expressions for $Z^{(n)}(0)$ may be written down using (41) and the zero-argument properties of the Hermite polynomials. These give

$$Z^{(n)}(0)=\begin{cases} i\sqrt{\pi}(-1)^{2k+1}\dfrac{(2k)!}{k!}, & n=2k, \\ (-1)^{k+1}2^{2k+1}k!, & n=2k+1. \end{cases} \tag{51}$$

From (51) it follows that there are $(k+1)$ zero crossings of $Z^{(n)}(u)$ with $u \geqslant 0$ when $n=2k$ or $n=2k+1$.

A. Numerical results

The real and imaginary parts of the lowest-order generalized plasma dispersion function $Y_{0,0}(a,b)$ are graphed in Fig. 1 for real a with $0 \leqslant a \leqslant 10$ for $b=0.1,1,10$. The minimum in $\operatorname{Re} Y_{0,0}(a,b)$ becomes shallow and shifts to higher a with increasing b. The magnitude of the maximum in $\operatorname{Im} Y_{0,0}(a,b)$ diminishes with increasing b such that the area under the curve is constant and independent of b. This may be proved by evaluating the integral $\int_{-\infty}^{\infty} \operatorname{Im} Y_{0,0}(a,b)da$ using (18) with (34) and showing that the result is independent of b. For $b \leqslant 0.1$, $Y_{0,0}(a,b)$ is indistinguishable from $Z(a)$ on the scale of the graphs in Fig. 1.

The real parts of $Y_{0,1}(a,b,c)$, $Y_{0,2}(a,b,c)$, and $Y_{0,3}(a,b,c)$ are graphed in Figs. 2, 3, and 4, respectively, for a number of parameter values. In graphing the generalized dispersion functions using the series expressions (23) and (33), numerical evaluation of the derivatives of the plasma dispersion function is required; this task is addressed in the Appendix. Padé approximants for the generalized plasma dispersion functions are also illustrated in Figs. 1 and 2; Padé approximants are discussed in Sec. V B.

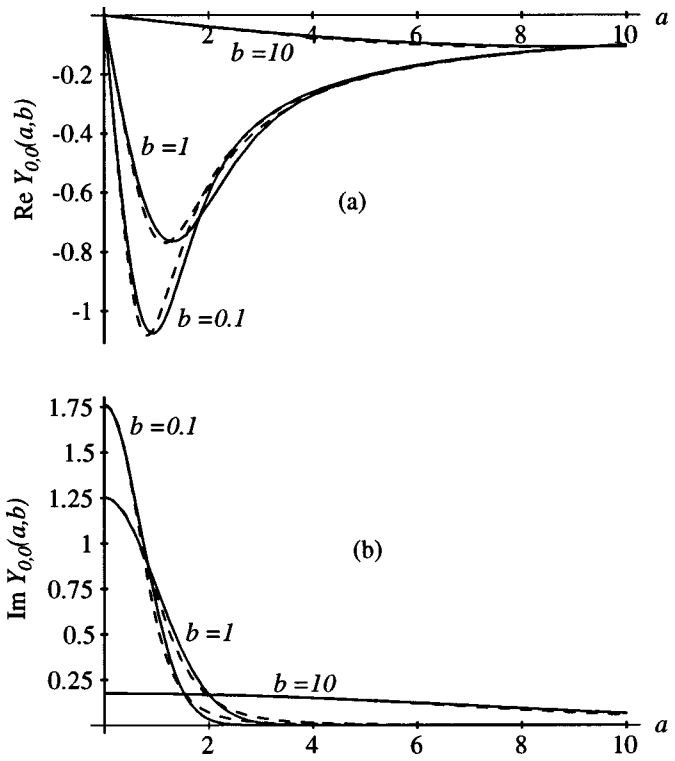

FIG. 1. Graphs of the (a) real and (b) imaginary parts of $Y_{0,0}(a,b)$ (solid curves) and their respective Padé approximants (dashed curves), for $0 \leqslant a \leqslant 10$ and $|b| = 0.1,1,10$, as labeled. When $|b| = 0.1$, the real and imaginary parts of $Y_{0,0}(a,b)$ are indistinguishable from those of $Z(a)$ on the scale of the given graphs.

V. APPROXIMATIONS FOR $Y_{0,n}(a,b,c)$

In the application of the quadratic response tensor to the treatment of specific nonlinear wave processes, such as the three-wave process satisfying the matching conditions (3), approximate analytic forms for the generalized plasma dispersion functions are required. As noted in the Introduction, the arguments of $Y_{m,n}(a,b,c)$ then have physical meaning, with the magnitudes of a and c given by the ratio of the wave phase speed to the plasma thermal speed for two of the interacting waves, and with $b = -\cot\theta$, where θ is the angle between the coalescing wave vectors. From these identifications, two classes of approximation are possible. *Geometric approximations* may be made for some interaction geometries; for example, the limit $|b| \ll 1$ holds for the coalescence of near perpendicularly propagating waves, while $|b| \gg 1$ for near parallel propagating or counterpropagating interacting waves. *Phase speed approximations* depend on the ratio of the phase speed of the interacting waves to the thermal speed of the plasma. The limiting cases $|a| \ll 1$, $|c| \ll 1$ and $|a| \gg 1$, $|c| \gg 1$ should then give useful approximate forms. In this section, analytic approximations and Padé approximants for the generalized plasma dispersion function are derived and discussed.

A. Analytic approximations

The approximation of the lowest-order generalized plasma dispersion function $Y_{0,0}(a,b)$ in various limits is straightforward. From (18), we have immediately

$$Y_{0,0}(a,b) \approx \begin{cases} Z(a), & |b|^2 \ll 1, \\ \dfrac{1}{b} Z\left(\dfrac{a}{b}\right), & |b|^2 \gg 1. \end{cases} \tag{52}$$

Approximate forms for the higher-order generalized plasma dispersion functions may be derived from the series expansions (23) and (33) for parameter ranges where the series rapidly converge, so that only leading terms need be retained. The power series and asymptotic expan-

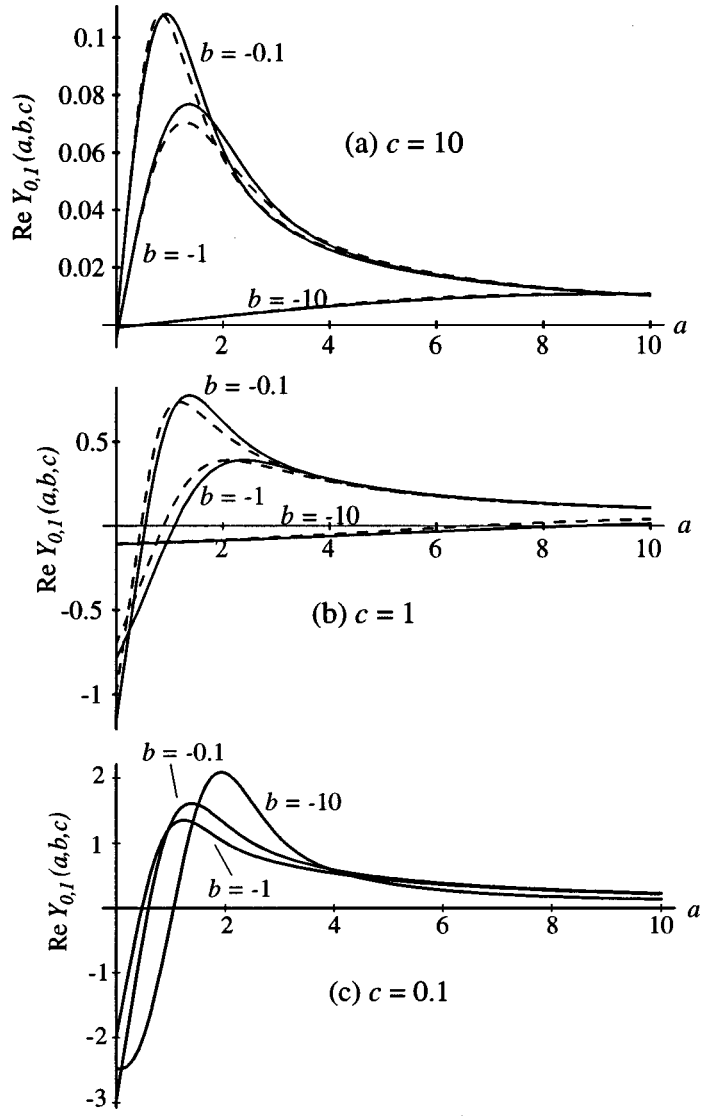

FIG. 2. The graph of Re $Y_{0,1}(a,b,c)$ (solid lines) with $0 \leqslant a \leqslant 10$ and (a) $c=10$, (b) $c=1$, and (c) $c=0.1$ for the physically relevant cases, with b taking the values -0.1, -1, and -10, as labeled. The Padé approximants for Re $Y_{0,1}$ are also shown (dashed line) for comparison, except for the case $c=0.1$, for which the simple approximant fails. Note the changes of scale on the vertical axes.

sions for the derivatives of the plasma dispersion function given by (47) and (49), respectively, can then be used to approximate the remaining terms in the series. This procedure can also be adopted to give geometric approximations of the quadratic response tensor for specific interaction geometries whenever $|b| \ll 1$ or $|b| \gg 1$ applies.

Simple phase-speed approximations for $Y_{0,1}(a,b,c)$ can be derived directly from the definition (4). If $|a| \gg 1$, then $Z(a+bx)$ may be approximated by the first term in its asymptotic expansion $-1/(a+bx)$ for small $|x|$, where the contribution of the integrand to the integral is significant (in view of the e^{-x^2} factor appearing in the integral). Then

$$Y_{0,1}(a,b,c) \approx \frac{1}{\sqrt{\pi}} \int \frac{dx\ e^{-x^2}}{(c-x)(a+bx)} \tag{53}$$

can be evaluated in terms of the plasma dispersion function, via partial fractions, to give

$$Y_{0,1}(a,b,c) \approx Z(a+bc)[Z(c)-Z(-a/b)], \tag{54}$$

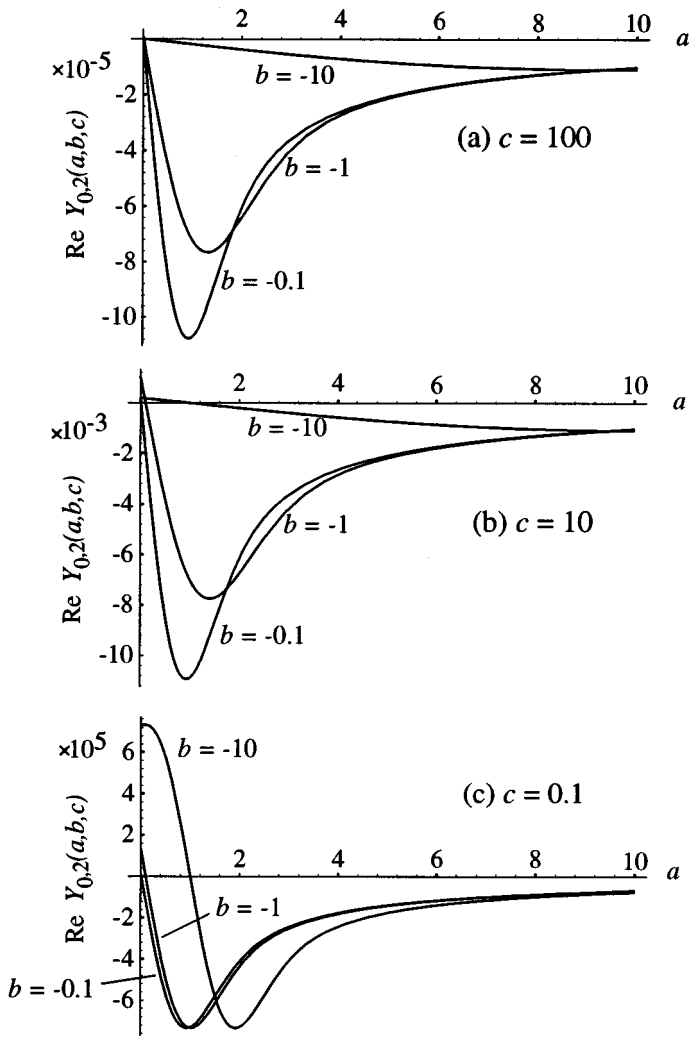

FIG. 3. The graph of Re $Y_{0,2}(a,b,c)$ with $0 \leqslant a \leqslant 10$ and (a) $c=100$, (b) $c=10$, and (c) $c=0.1$ for the physically relevant cases, with b taking the values -0.1, -1, and -10 as labeled.

where $|a+bc| \gg 1$ also holds. Comparisons with numerical evaluations suggest this approximation is valid for a wider range of parameters, failing only when $|a| \leqslant 1$, $|c| \leqslant 1$ or when $a+bc \approx 0$. An analogous approximation can be derived for the higher-order generalized plasma dispersion functions. In particular,

$$Y_{0,2}(a,b,c) \approx bZ^{(1)}(a+bc)[Z(c)-Z(-a/b)]+Z^{(1)}(c)Z(a+bc), \tag{55}$$

and

$$Y_{0,3}(a,b,c) \approx \frac{b^2}{2} Z^{(2)}(a+bc)[Z(c)-Z(-a/b)]+\tfrac{1}{2}Z(a+bc)Z^{(2)}(c)+bZ^{(1)}(a+bc)Z^{(1)}(c). \tag{56}$$

Repeated use of the recursion relation (6) generates approximate forms for $Y_{1,1}$, $Y_{1,2}$, and $Y_{2,2}$ which also appear in the quadratic response tensor for a thermal plasma. The power series and asymptotic expansions for the $Z^{(n)}$ may then be used to obtain simple analytic forms for the $Y_{m,n}$.

A more useful phase speed approximation for the generalized plasma dispersion function $Y_{0,n}(a,b,c)$ holds when $|c| \gg 1$ such that the resonant denominator $(x-c)^n$ in the definition (4) can be approximated by $(-c)^n$. It follows that

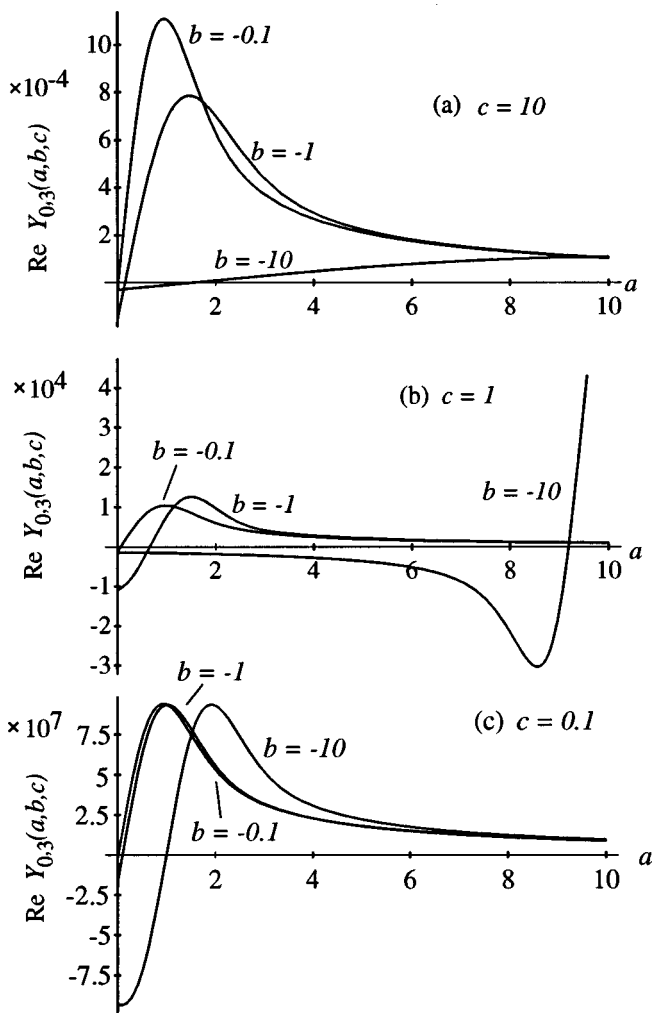

FIG. 4. The graph of Re $Y_{0,3}(a,b,c)$ with $0 \leqslant a \leqslant 10$ and (a) $c=10$, (b) $c=1$, and (c) $c=0.1$ for the physically relevant cases with b taking the values -0.1, -1 and -10, as labeled. Note the changes of scale on the vertical axes.

$$Y_{0,n}(a,b,c) \approx Y_{0,0}(a,b)/(-c)^n, \quad |c| \gg 1. \tag{57}$$

Substitution of (57) into the recursion relation (6) gives $Y_{m,n} \approx 0$ whenever $m \neq 0$. Taken together with (57), we find

$$Y_{m,n}(a,b,c) \approx \begin{cases} \dfrac{1}{(-c)^n \sqrt{1+b^2}}\, Z\left(\dfrac{a}{\sqrt{1+b^2}}\right), & m=0, \\ 0, & m \neq 0. \end{cases} \tag{58}$$

The validity of (58) is demonstrated in a particular case by comparing Re $Y_{0,1}$ with $c=10$ graphed in Fig. 2(a), respectively, with Re $Y_{0,0}$ shown in Fig. 1(a). Physically, the limit $|c| \gg 1$ corresponds to a product wave phase speed much greater than the plasma thermal speed in a three-wave coalescence process.

B. Padé approximants

The use of (modified) Padé approximants for the plasma dispersion function $Z(u)$ is a widely used technique in the analytic treatment of the dispersive properties of a Maxwellian plasma.[14–19]

This approach can be extended to the description of second-order (three-wave) wave processes in a Maxwellian plasma by deriving Padé approximants for the generalized plasma dispersion functions $Y_{0,n}$.

The Padé approximant $\tilde{Z}(u)$ for the plasma dispersion function $Z(u)$ may be written in the multipole form

$$\tilde{Z}(u)=\sum_{j=1}^{N}\left(\frac{\alpha_j}{u-\beta_j}+\frac{\alpha_j^*}{u+\beta_j^*}\right), \tag{59}$$

where the coefficients α_j, β_j are chosen so that the approximant fits a total of $2N$ leading terms selected jointly from the power series and the asymptotic expansion of $Z(u)$ given by (47) and (49) with $n=0$. The general form (59) satisfies the symmetry (7) with $n=0$. The $N=2$ Padé approximant which fits the first two terms of both the power series and asymptotic expansions of the plasma dispersion function is[16]

$$\tilde{Z}(u)=\frac{i\sqrt{\pi}+(\pi-2)u}{1-i\sqrt{\pi}u-(\pi-2)u^2}, \tag{60}$$

which may be numerically expressed in the two-pole form

$$\tilde{Z}(u)=\frac{\alpha}{u-\beta}+\frac{\alpha^*}{u+\beta^*}, \tag{61}$$

with

$$\alpha=-\frac{1}{2}-\frac{i\sqrt{\pi}}{2\sqrt{3\pi-8}}, \quad \beta=\frac{\sqrt{3\pi-8}-i\sqrt{\pi}}{2(\pi-2)}. \tag{62}$$

In Ref. 19, it is argued on physical grounds that the two-pole approximant (61) is the most useful form since it satisfies the correct symmetries, the requirement of non-negative wave damping in a Maxwellian plasma, has a monotonically decreasing imaginary part, and does not lead to spurious wave growth.

Substitution of the approximant (61) into the definition (4) of the generalized plasma dispersion function with $m=0$ and $n=0,1$ and performing the integrations using partial fractions gives the two lowest-order Padé approximants,

$$\tilde{Y}_{0,0}(a,b)=\frac{\alpha}{a-\sqrt{1+b^2}\beta}+\frac{\alpha^*}{a+\sqrt{1+b^2}\beta^*}, \tag{63}$$

and

$$\tilde{Y}_{0,1}(a,b,c)=\tilde{Z}(c)\tilde{Z}(a+bc)-\left[\frac{\alpha}{a+bc-\beta}\,\tilde{Z}\left(\frac{\beta-a}{b}\right)+\frac{\alpha^*}{a+bc+\beta^*}\,\tilde{Z}\left(\frac{-\beta^*-a}{b}\right)\right], \tag{64}$$

where the tilde again denotes a Padé approximated quantity. Padé approximants for higher-order generalized plasma dispersion functions may be derived from (63) and (64) using the recursion relation (6), together with the derivative identities (12). A graphical comparison of the real and imaginary parts of the Padé approximant $\tilde{Y}_{0,0}$ with those of $Y_{0,0}$ for particular real arguments is shown in Fig. 1. In Fig. 2, the Padé approximant for Re $Y_{0,1}(a,b,c)$ is compared with the exact expression when $c\geqslant 1$. For $c<1$, the Padé approximant for the generalized dispersion functions is poor because the effect of the resonant denominator $(x-c)^n$ in the defining integral (4) of $Y_{m,n}$ is inadequately approximated.

Padé approximants for the remaining generalized dispersion functions $Y_{1,1}$, $Y_{1,2}$, and $Y_{2,2}$ appearing in the exact quadratic response tensor may be written down using the recursion relation (6). Taking these results together, a simple rational approximation for the quadratic response tensor can then be written down. However, care must be taken with such a procedure. In general,

the two-pole Padé approximant is inadequate when the pole at $x=c$ in (4) is close to the origin, i.e., when $|c|<1$. The physical significance of this case is in the treatment of three-wave interactions when the product wave phase speed is slow in comparison with the plasma thermal speed.

VI. SUMMARY

A wide variety of mathematical properties of the generalized plasma dispersion functions $Y_{m,n}$, $m,n\geqslant 0$ have been examined here, including recursion relations, partial differential equations, and series expansions in terms of products of Hermite polynomials and derivatives of the standard plasma dispersion function $Z^{(n)}$, $n\geqslant 0$. Analytic and Padé approximate forms for the generalized plasma dispersion functions have also been discussed, and the ranges of validity for these approximations have been identified. Since the evaluation of the $Y_{m,n}$ requires the evaluation of the $Z^{(n)}$, these derivatives have been examined here, with some new analytic results obtained for them. The numerical stability of the recursive computation of the $Z^{(n)}$ has also been treated. Collectively, these results find a wide physical application in the treatment of second-order wave processes in warm collisionless plasmas with constituent species having Maxwellian velocity distributions, since it was shown in Ref. 6 that the exact quadratic response tensor which describes such processes can be written in terms of the $Y_{m,n}$ with $0\leqslant m,n\leqslant 3$. Such processes include many of those relevant to the theory and application of weak turbulence and parametric three-wave interactions.

The summability of the series expressions (23) and (33) for the generalized plasma dispersion functions remains an open question. Although numerous summable series involving Hermite polynomials are reported in the literature,[12] we have not been able to manipulate the series (23) or (33) into any known summable form. Further attention to this issue in the future would be desirable, motivated by the utility of a finite series expression for the exact quadratic response tensor for application to specific second-order plasma processes.

APPENDIX: NUMERICAL COMPUTATION OF $Z^{(n)}$

The numerical evaluation of the generalized plasma dispersion functions using the series expansions derived in Sec. III requires the computation of a sequence of $Z^{(n)}$. It is straightforward to derive the recursion relation[2,3]

$$Z^{(n+1)}(u)=-2uZ^{(n)}(u)-2nZ^{(n-1)}(u), \quad n\geqslant 1, \tag{A1}$$

by repeated application of the differential equation $Z^{(1)}(u)=-2-2uZ(u)$ satisfied by the plasma dispersion function. However, analysis of the stability of this recursion relation shows it to be unstable to upward iteration for large $|u|$. The argument is as follows: Assuming that $Z^{(n+1)}(u)\approx 0$ in (A1), the approximate recursion relation may be solved to give

$$Z^{(n)}(u)\approx\left(\frac{-1}{u}\right)^n n!\,Z(u). \tag{A2}$$

The initial assumption may be justified *a posteriori* only for fixed n and for $|u|\to\infty$ such that the right side of (A1) tends to zero. A comparison of (A2) with the first term in the asymptotic expansion (49) with $Z(u)\approx -1/u$ demonstrates that (A2) is the required form of solution for large $|u|$. However, a second growing solution of the recursion relation (51) may be identified by assuming $Z^{(n-1)}(u)\approx 0$ and then solving (A2) to give

$$Z^{(n)}(u)=(-2u)^n Z(u), \tag{A3}$$

which grows indefinitely for fixed n and $|u|\to\infty$, or for fixed $|u|>\frac{1}{2}$ and $n\to\infty$, again justifying the initial assumption *a posteriori* in these regimes. In any upward iteration scheme using (A1), inevitable numerical roundoff error will cause this unwanted second solution to grow to dominate the required solution in the regime of large $|u|$. The remedy is to perform *downward* iteration for large $|u|$, starting with $Z^{(\nu)}(u)=0$, $Z^{(\nu-1)}(u)=1$ for some large integer ν, and then normalizing

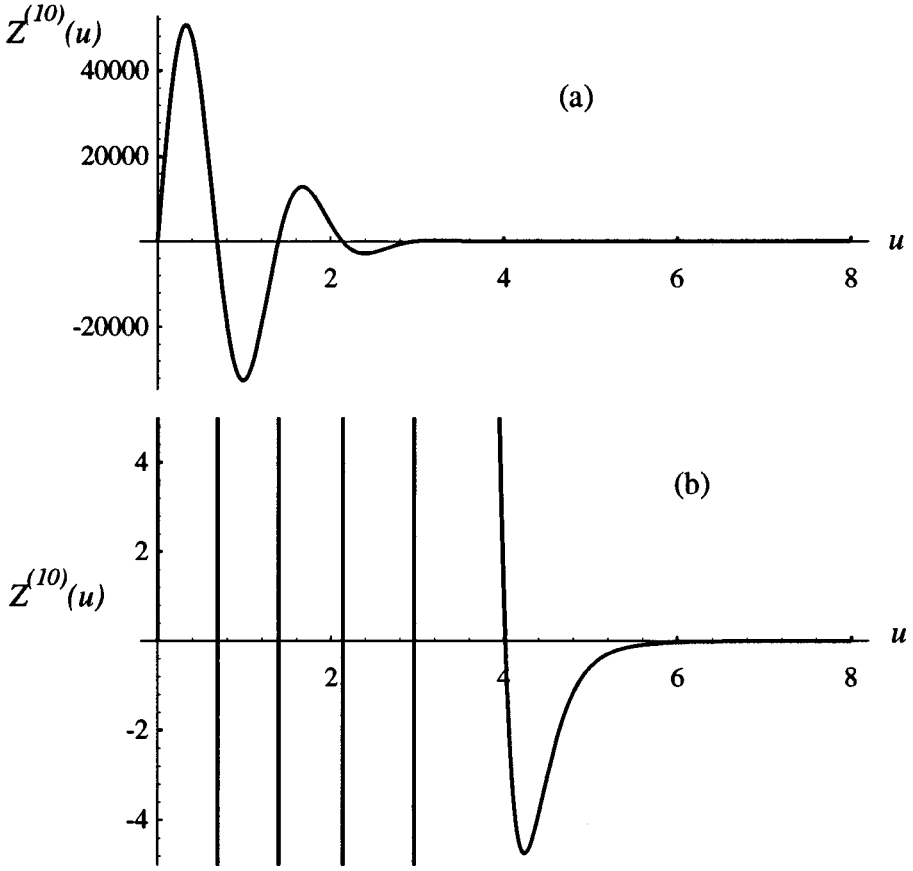

FIG. 5. Numerical evaluation of the real part of $Z^{(10)}(u)$ for real u in the domain $0 \leqslant u \leqslant 8$. For clarity, the vertical axis of (a) near the origin is magnified and illustrated in (b). Upward iteration using (A1) is stable in the oscillating regime; downward iteration is stable in the decaying regime for large u. The boundary between the two regimes is at $u \approx 3.2$.

the sequence of $Z^{(n)}(u)$, $n \leqslant \nu$ so obtained using the known form of $Z^{(0)}(u)$. The numerical solution of second-order difference equations and the stability of downward iteration techniques were reviewed in Ref. 20.

It is instructive to consider the form of $\mathrm{Re}\, Z^{(n)}(u)$ for large n, sketched in Fig. 5 for real argument u. There are two distinct regimes: a rapidly oscillating regime with $0 \leqslant u \leqslant \sqrt{n}$ and a decaying regime $u \geqslant \sqrt{n}$, where $Z^{(n)}(u)$ tends to zero. Upward iteration using the recursion relation (A1) is stable in the oscillating regime; downward recursion is required in the decaying regime. The expression (41) for the $Z^{(n)}(u)$ involving a finite sum of Hermite polynomials is useful for numerical computation only in the oscillating regime; catastrophic cancellation in the summation is a problem in the decaying regime for large n. The boundary $u = \sqrt{n}$ between the two regimes has been obtained empirically. We postulate that there are k zeros of $\mathrm{Re}\, Z^{(2k)}(u)$ and $\mathrm{Re}\, Z^{(2k+1)}(u)$ with $u < \sqrt{n}$ and only one zero crossing for $u > \sqrt{n}$. This result has been directly demonstrated numerically for $n \leqslant 100$ but not yet proved for all n.

A possible alternative scheme[2,3] for the computation of a sequence of $Z^{(n)}$ is to compute repeated integrals of the complementary error function and relate them to $Z^{(n)}$ using (46). An analysis of the recursive computation of $I^n \operatorname{erfc}(x)$ for real x using downward iteration has been given in Ref. 21. A generalization of this analysis to a complex argument is required here. For complex u in the upper half-plane with $\mathrm{Im}\, u > 0$, a sequence $Z(u)$, $Z^{(1)}(u), \ldots, Z^{(N)}(u)$ can be calculated using downward recursion to a relative error not exceeding 10^{-p} if the starting ν satisfies

$$\nu \geqslant \left(\sqrt{N} + \frac{p \ln 10 + \ln 2}{2\sqrt{2}\ \mathrm{Im}(u)} \right)^2 . \tag{A4}$$

From (A4), it is clear that as u approaches the real axis, the lower bound of ν grows indefinitely. Therefore on or near the real axis, the evaluation of the $Z^{(n)}(u)$ via the recursive computation of repeated integrals of the error function is inappropriate, and direct computation using (41) is required.

[1] B. D. Fried and S. D. Conte, *The Plasma Dispersion Function* (Academic, New York, 1961).
[2] P. A. Robinson, J. Math. Phys. **30**, 2484 (1989).
[3] P. A. Robinson, J. Comput. Phys. **88**, 381 (1990).
[4] D. B. Melrose, *Instabilities in Space and Laboratory Plasmas* (Cambridge University Press, Cambridge, 1986).
[5] A. I. Akhiezer, I. A. Akhiezer, R. V. Polovin, A. G. Sitenko, and K. N. Stepanov, *Plasma Electrodynamics: Vol. 1, Linear Theory* (Pergamon, Oxford, 1975).
[6] D. J. Percival and P. A. Robinson, Phys. Plasmas **5**, 1279 (1998).
[7] V. N. Tsytovich, *An Introduction to the Theory of Plasma Turbulence* (Pergamon Press, Oxford, 1972).
[8] R. C. Davidson, *Methods in Nonlinear Plasma Theory* (Academic, New York, 1972).
[9] A. I. Akhiezer, I. A. Akhiezer, R. V. Polovin, A. G. Sitenko, and K. N. Stepanov, *Plasma Electrodynamics: Vol. 2, Nonlinear Theory* (Pergamon, Oxford, 1975).
[10] A. G. Sitenko, *Fluctuations and Nonlinear Wave Interactions in Plasmas* (Pergamon, Oxford, 1982).
[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, San Diego, 1980).
[12] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series* (Gordon and Breach, New York, 1986), Vol. 2.
[13] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1980).
[14] B. D. Fried, C. L. Hedrick, and J. McCune, Phys. Fluids **11**, 249 (1968).
[15] P. Martin and A. González, Phys. Fluids **22**, 1413 (1979).
[16] P. Martin, G. Donoso, and J. Zamudio-Cristi, J. Math. Phys. **21**, 280 (1980).
[17] G. Németh, Á. Ág, and G. Páris, J. Math. Phys. **22**, 1192 (1981).
[18] K. G. Rönnmark, Plasma Phys. **25**, 669 (1983).
[19] P. A. Robinson and D. L. Newman, J. Plasma Phys. **40**, 553 (1988).
[20] W. Gautschi, SIAM (Soc. Ind. Appl. Math.) Rev. **9**, 24 (1967).
[21] W. Gautschi, Math. Comput. **15**, 227 (1961).

Advection–diffusion around a curved obstacle

Daljit S. Ahluwalia
Department of Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, New Jersey 07102

Joseph B. Keller
Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, California 94305-2125

Charles Knessl
Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607-7045

(Received 9 February 1998; accepted for publication 5 March 1998)

Advection and diffusion of a substance around a curved obstacle is analyzed when the advection velocity is large compared to the diffusion velocity, i.e., when the Peclet number is large. Asymptotic expressions for the concentration are obtained by the use of boundary layer theory, matched asymptotic expansions, etc. The results supplement and extend previous ones for straight obstacles. They apply to electrophoresis, the flow of ground water, chromatography, sedimentation, etc. © *1998 American Institute of Physics.* [S0022-2488(98)00707-5]

I. INTRODUCTION

Advection–diffusion processes occur in many different physical contexts, such as electrophoresis, ground water flow, chromatography, sedimentation, etc. Their quantitative description leads to the advection–diffusion equation for a concentration, together with appropriate boundary conditions. Our objective is to present a method for solving the resulting boundary value problems for the steady state in the difficult case in which the advection velocity is large compared to the relevant diffusion velocity, i.e., in the case of a large Peclet number. Previously, we treated the cases in which the flow is blocked by an impermeable infinite strip of finite width, either normal to the advection velocity[1] or oblique to it.[2] Now we consider blocking by curved cylindrical sheets, such as those shown in Figs. 1 and 2, and by axially symmetric bodies.

The flux $\mathbf{j}$ of a substance with concentration p is $\mathbf{j}=-D\boldsymbol{\nabla}p+v\mathbf{i}p$, where D is the diffusion coefficient and v is the advection velocity in the direction of the unit vector $\mathbf{i}$. In the steady state, conservation of mass implies that $\boldsymbol{\nabla}\cdot\mathbf{j}=0$, which we write as

$$\Delta p-2cp_x=0. \tag{1.1}$$

The Peclet number is $c=vL/2D$, where L is the width of an obstacle or some other reference length. On the surface B of an impermeable obstacle, the normal flux must vanish, i.e., $\mathbf{j}\cdot\mathbf{n}=0$. This condition can be written as

$$\partial_n p-(2c\ \cos\ \beta)p=0, \quad \mathbf{x}\ \text{ on }\ B. \tag{1.2}$$

Here $\mathbf{n}$ is the unit normal to B and $\cos\beta=\mathbf{n}\cdot\mathbf{i}$, so that β is the angle between the x axis and $\mathbf{n}$. In the exterior of a bounded obstacle, we suppose that $p=1$ at infinity:

$$p\to 1, \quad \text{as}\ \ |\mathbf{x}|\to\infty. \tag{1.3}$$

An exact solution of (1.1) that satisfies (1.2) on any surface B is, with H a constant,

$$p(\mathbf{x})=He^{2cx}. \tag{1.4}$$

 © 1998 American Institute of Physics

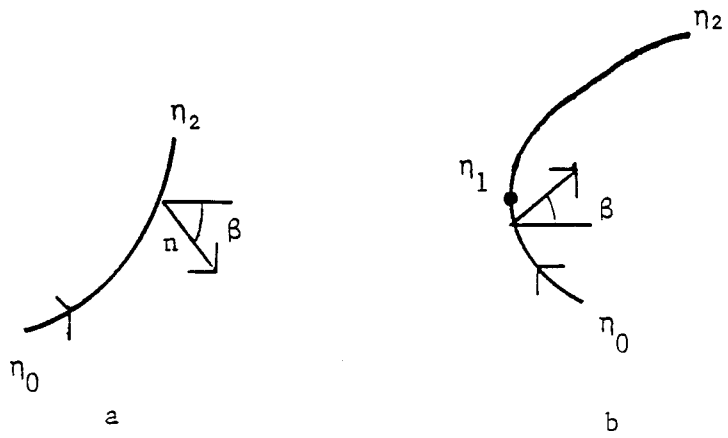

FIG. 1. Two examples of the normal cross section, in the x,y plane, of a nontrapping cylindrical obstacle B with generators parallel to the z axis. In each case the arclength η increases upward from the value η_0 at the lower end point, in the direction of the arrow. The unit tangent vector **t** points in this direction and the unit vector **n** points to the right. The angle β between the positive x axis and **n** is shown. In (a), $\beta<0$ and the curvature $\kappa(\eta)>0$. In (b), $\beta>0$ for $\eta_0<\eta<\eta_1$, $\beta<0$ for $\eta>\eta_1$, and $\beta(\eta_1)=0$ while $\kappa(\eta)<0$ for all η.

When B is a closed surface, (1.4) holds in the bounded domain inside B and H determines the total quantity of the substance in the domain. Since this solution does not satisfy (1.3), it does not hold in an exterior domain.

The flow upstream of a nontrapping boundary B, such as one of those in Fig. 1, is nearly parallel to the x axis except close to B. For c large, the flowing particles are concentrated in a thin diffusive boundary layer on the upstream or front side of B. They flow along B in this layer under the combined influence of advection and diffusion. At the downstream edge or edges of B, they leave the surface and flow away parallel to the x axis, creating a shadow behind B. This shadow is gradually filled as a result of diffraction and diffusion, so it ultimately disappears.

Many features of this flow are similar to those that occur when short waves hit a body, and we have analyzed those features before.[1,2] However, the concentration of particles in the diffusive boundary layer on the front side of B is a special feature of advection–diffusion, which does not occur in wave problems. This effect is enhanced when the front surface of B is concave to the left, as in Fig. 2, so that it can trap particles. Then the concentration becomes exponentially large there, and we shall show that it is given asymptotically by (1.4), with a particular value of H.

To analyze the concentration phenomenon, we shall find the first terms in the expansion of p within the diffusive boundary layer on the front side of B. In Sec. II, we treat nontrapping cylindrical boundaries, excluding the edges. In Sec. III, we complete that treatment by considering the special expansions that are needed near the edges. In Sec. IV, we analyze trapping cylindrical boundaries by finding the diffusive boundary layer expansion, the special expansion near the

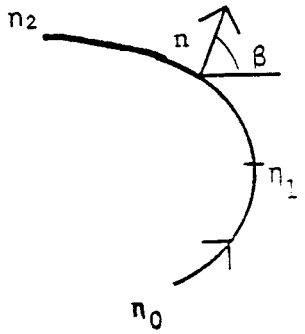

FIG. 2. The normal cross section of a trapping cylindrical obstacle with $\beta<0$ for $\eta_0<\eta<\eta_1$, $\beta>0$ for $\eta>\eta_1$, and $\beta(\eta_1)=0$, while $\kappa(\eta)>0$ for all η.

upstream edge, and the special expansion near a point where the boundary is normal to the advection velocity. We match the latter expansion to the diffusive boundary layer expansion in Sec. V. Then we complete this analysis by finding the special expansion valid near the downstream edge. In Sec. VII, we treat a special case. Finally, in Sec. VIII, we consider axially symmetric nontrapping boundaries.

II. DIFFUSIVE BOUNDARY LAYER ON A NONTRAPPING BOUNDARY

We consider the two-dimensional case in which B is a nontrapping cylindrical obstacle with generators parallel to the z axis. Then the cross section of B in the plane $z=0$ is a curve $x_0(\eta)$, $y_0(\eta)$, where η is arclength along it. When the obstacle is a thin sheet, the curve B is an open arc, with end points at η_0 and η_2, as is shown in Fig. 1. To determine p on the front or upstream side of B, we introduce the distance σ from B along the unit normal $\mathbf{n}$ to B, and write $(x,y)=[x_0(\eta),y_0(\eta)]+\sigma\mathbf{n}$. Then η, σ are curvilinear coordinates in the neighborhood of B. In terms of them,

$$\nabla p = p_\sigma\, \mathbf{n}(\eta) + [1+\sigma\kappa(\eta)]^{-1} p_\eta\, \mathbf{t}(\eta). \tag{2.1}$$

Here $\mathbf{n}(\eta)$ and $\mathbf{t}(\eta)$ are, respectively, the unit normal vector and the unit tangent vector to B while $\kappa(\eta)$ is the curvature of B. When η increases in the direction of increasing y, the y component of $\mathbf{t}$ is positive, and we choose the direction of $\mathbf{n}$ so that its x component is positive. Then we define the angle $\beta(\eta)$ between n and the positive x axis such that

$$\mathbf{n}\cdot\mathbf{i}=\cos\beta, \quad \mathbf{t}\cdot\mathbf{i}=-\sin\beta, \quad -\pi/2\leqslant\beta\leqslant\pi/2. \tag{2.2}$$

With these definitions, $\sigma<0$ on the upstream side of B. The curvature $\kappa(\eta)=\beta'(\eta)$ is positive when B is concave to the left, as in Fig. 1(a), and κ is negative when B is concave to the right, as in Fig. 1(b).

In anticipation of the boundary layer nature of p on the front side of B, we introduce the stretched coordinate $\xi=c\sigma$ and consider $p(\xi,\eta)$. Then we rewrite (1.1)–(1.3) as

$$p_{\xi\xi}-2\cos\beta\, p_\xi + c^{-1}\kappa(1+c^{-1}\kappa\xi)^{-1}p_\xi + 2c^{-1}(1+c^{-1}\kappa\xi)^{-1}\sin\beta\, p_\eta + c^{-2}(1+c^{-1}\kappa\xi)^{-2}p_{\eta\eta} - c^{-3}\kappa'(1+c^{-1}\kappa\xi)^{-3}\xi p_\eta = 0, \tag{2.3}$$

$$p_\xi - 2\cos\beta\, p = 0, \quad \text{at } \xi=0, \tag{2.4}$$

$$p\to 1, \quad \text{as } \xi\to-\infty. \tag{2.5}$$

To solve (2.3)–(2.5) near B for c large and $\xi\leqslant 0$, we seek an asymptotic expansion for p of the form

$$p = 1 + cQ^{(1)} + Q^{(0)} + c^{-1}Q^{(-1)} + O(c^{-2}), \quad \xi\leqslant 0. \tag{2.6}$$

We substitute (2.6) into (2.3)–(2.5) and equate coefficients of c^j, $j=1,0,-1$ in each equation, to obtain

$$Q^{(1)}_{\xi\xi} - 2\cos\beta\, Q^{(1)}_\xi = 0, \tag{2.7}$$

$$Q^{(0)}_{\xi\xi} - 2\cos\beta\, Q^{(0)}_\xi = -\kappa Q^{(1)}_\xi - 2\sin\beta\, Q^{(1)}_\eta, \tag{2.8}$$

$$Q^{(-1)}_{\xi\xi} - 2\cos\beta\, Q^{(-1)}_\xi = \kappa Q^{(0)}_\xi - 2\sin\beta\, Q^{(0)}_\eta + \xi\kappa^2 Q^{(1)}_\xi + 2\sin\beta\, \xi\kappa Q^{(1)}_\eta - Q^{(1)}_{\eta\eta}, \tag{2.9}$$

$$Q^{(j)}_\xi - 2\cos\beta\, Q^{(j)} = 2\delta_{j0}\cos\beta \quad \text{at } \xi=0, \quad j=1,0,-1, \tag{2.10}$$

$$Q^{(j)}\to 0, \quad \text{as } \xi\to-\infty. \tag{2.11}$$

The solution of (2.7) that satisfies (2.10) and (2.11) is

$$Q^{(1)}(\xi,\eta)=e^{2\xi\cos\beta}A_1(\eta). \tag{2.12}$$

To determine A_1 we use (2.12) for $Q^{(1)}$ in (2.8), integrate the resulting equation with respect to ξ from $-\infty$ to ξ, and use (2.11). In this way we get

$$Q^{(0)}_{\xi}-2\cos\beta\,Q^{(0)}=-\left[\kappa A_1+\sin\beta\left(\frac{A_1}{\cos\beta}\right)'-\frac{2\sin^2\beta\,\beta'}{\cos\beta}\,\xi A_1\right]e^{2\xi\cos\beta}. \tag{2.13}$$

Now we set $\xi=0$ in (2.13) and note that the left side equals $2\cos\beta$ in view of (2.10). Then since $\kappa=\beta'$, we can write (2.13) at $\xi=0$ as

$$\sin\beta\left(\frac{A_1}{\cos\beta}\right)'+\beta' A_1=-2\cos\beta. \tag{2.14}$$

The solution of (2.14) is

$$A_1(\eta)=-2\cot\beta(\eta)\left(\int_{\eta_0}^{\eta}\cos\beta(\eta')\,d\eta'+a_1\right)=-2\cot\beta(\eta)[y_0(\eta)-y_0(\eta_0)+a_1]. \tag{2.15}$$

Except for the undetermined constant a_1, (2.15) determines $A_1(\eta)$ and then (2.12) gives $Q^{(1)}$.

Now we use (2.14) in (2.13), solve for $Q^{(0)}$ and impose the condition (2.11) to get

$$Q^{(0)}(\xi,\eta)=\left[2\xi\cos\beta+\xi^2\frac{\sin^2\beta}{\cos\beta}\,\kappa A_1+A_0(\eta)\right]e^{2\xi\cos\beta}. \tag{2.16}$$

The function $A_0(\eta)$ is undetermined. To find it we use (2.12) for $Q^{(1)}$ and (2.16) for $Q^{(0)}$ in (2.9), integrate (2.9) with respect to ξ from $-\infty$ to 0, and use (2.10) and (2.11). In this way, we obtain

$$\begin{aligned}\sin\beta\left(\frac{A_0(\eta)}{\cos\beta}\right)'+\beta' A_0(\eta)=\sin\beta\left(\frac{1}{\cos\beta}-\frac{\sin^2\beta}{2\cos^4\beta}\,\beta' A_1\right)'-\frac{1}{2}\left(\frac{A_1}{\cos\beta}\right)''\\ -\frac{1}{2}\beta'\sin\beta\left(\frac{A_1}{\cos^2\beta}\right)'-\frac{\beta'^2 A_1}{2\cos\beta}.\end{aligned} \tag{2.17}$$

The solution of (2.17) for $A_0(\eta)$ can be written as follows, with a_0' an arbitrary constant:

$$\begin{aligned}A_0(\eta)=\cot\beta\Bigg[a_0'-\beta+\frac{1}{\sin\beta\cos\beta}+\frac{\cos^4\beta-\sin^2\beta}{2\cos^4\beta\sin\beta}\,\kappa A_1\\ +\int_{\eta_0}^{\eta}\frac{\beta'^2\sin^2\beta+\beta''\sin\beta\cos\beta}{2\cos^3\beta}A_1(\eta')\,d\eta'\Bigg].\end{aligned} \tag{2.18}$$

We can simplify this formula by using in the integral the first expression for $A_1(\eta)$ in (2.15). Then, by integrating by parts once we can evaluate the integral. Upon combining terms, simplifying, and setting $a_0=a_0'-\beta(\eta_0)+\beta'(\eta_0)a_1/\cos\beta(\eta_0)$, we get

$$A_0(\eta)=\cot\beta\left[a_0+\frac{1}{\sin\beta\cos\beta}+\beta'(\eta)A_1(\eta)\frac{\cot 2\beta}{\cos^3\beta}\right]. \tag{2.19}$$

Finally, we write the diffusive boundary layer expansion (2.6) as follows, by using (2.12) for $Q^{(1)}$ and (2.16) for $Q^{(0)}$:

$$p(\xi,\eta)=1+e^{2\xi\cos\beta}\left[cA_1(\eta)+2\xi\cos\beta+\xi^2\beta'\frac{\sin^2\beta}{\cos\beta}A_1(\eta)+A_0(\eta)\right]+O(c^{-1}). \tag{2.20}$$

The functions $A_1(\eta)$ and $A_0(\eta)$ are given by (2.15) and (2.19), respectively, which contain the arbitrary constants a_1 and a_0.

The expansion (2.20) is valid at interior points of nontrapping boundaries B, such as those shown in Figs. 1(a) and 1(b). At the end points η_0 and η_2, it must be replaced by the edge boundary layers that have been determined in Ref. 2. In the next section, we shall show how matching the upstream edge boundary layer expansion to (2.20) determines a_1 and a_0 for B of the form in Fig. 1(a). We shall also show how to determine a_1 and a_0 for B of the form in Fig. 1(b) in order to make $A_1(\eta)$ and $A_0(\eta)$ finite at the point η_1 where $\beta(\eta_1)=0$. In a later section, we shall consider trapping boundaries, like that in Fig. 2.

III. DETERMINING THE CONSTANTS IN THE EXPANSION FOR A NONTRAPPING BOUNDARY

First, we shall complete the determination of the diffusive boundary layer expansion (2.20) for a nontrapping boundary B, such as that shown in Fig. 1(a). Since $-\pi/2<\beta(\eta)<0$, the expansion (2.20) is not singular for any point in the interval $\eta_0<\eta<\eta_2$. To find the constants a_1 and a_0, we consider $p(\xi,\eta)$ near the leading or upstream edge $(0,\eta_0)$. We can represent p near that edge by another expansion, the edge boundary layer expansion. This involves the stretched variable $u=c(\eta-\eta_0)$ and the consideration of p as a function of ξ and u in the form

$$p=1+F_0(\xi,u)+O(c^{-1}). \tag{3.1}$$

Previously we determined this expansion when the boundary B was a strip of width $2L$ with its normal making an angle $\beta_0<0$ with the x axis, Ref. 2, Result 3 (iii). Then we expanded it for large values of u and matched it to the diffusive boundary layer expansion, which was a special case of (2.20). In that way, we determined the constants a_1 and a_0 and found for $p(\xi,\eta)$ the expansion

$$p\sim 1-e^{2\xi\cos\beta_0}\left[\frac{2c\cos^2\beta_0}{\sin\beta_0}(\eta-\eta_0)-2\xi\cos\beta_0+\frac{1}{\sin\beta_0}\right]. \tag{3.2}$$

We get exactly this same result even when B is curved—the curvature enters at a higher order.

The result (3.2) is valid when $\eta\to\eta_0$, $c\to\infty$, and $u=c(\eta-\eta_0)\to\infty$. Therefore, for η near η_0, it must agree with (2.20). When (2.20) is expanded for η near η_0, it becomes

$$\begin{aligned}p\sim 1-e^{2\xi\cos\beta_0}&\left[\frac{2c\cos\beta_0}{\sin\beta_0}([\eta-\eta_0]\cos\beta_0+a_1)-2\xi\cos\beta_0-\cot\beta_0\left(a_0+\frac{1}{\sin\beta_0\cos\beta_0}\right)\right]\\&+O\left[\frac{1}{c},\eta-\eta_0(\eta-\eta_0)^2\right]+a_1 f(\xi,\eta-\eta_0,c).\end{aligned} \tag{3.3}$$

Here $\beta_0=\beta(\eta_0)$ and the function f is $O(1)$ as $c\to\infty$. Equations (3.2) and (3.3) do indeed match, provided that a_1 and a_0 have the values

$$a_1=0,\quad a_0=-\frac{1+\sin\beta_0}{\sin\beta_0\cos\beta_0}. \tag{3.4}$$

With the values (3.4), which apply to B of the form in Fig. 1(a), the expansion (2.20) for p is completely determined, and it can be written as follows in the open interval $\eta_0<\eta<\eta_2$:

$$\begin{aligned}p(\xi,\eta)\sim 1+e^{2\xi}\cos\beta&\left[\left(-2c\cot\beta-2\xi^2\beta'\sin\beta-\frac{2\beta'\cot 2\beta}{\sin^2\beta\cos\beta}\right)[y_0(\eta)-y_0(\eta_0)]\right.\\&\left.+2\xi\cos\beta+\frac{1}{\sin^2\beta}-\cot\beta\frac{1+\sin\beta_0}{\sin\beta_0\cos\beta_0}\right],\quad \eta_0<\eta<\eta_2.\end{aligned} \tag{3.5}$$

Near η_0 the leading edge boundary layer expansion (3.1) applies, and it is given in Ref. 2. Near η_2 the trailing edge boundary layer applies, and it is also given in Ref. 2. The expansion (3.5) and these two edge expansions determine p all along the illuminated or front side of a boundary such as that in Fig. 1(a).

Next we shall consider a nontrapping boundary of the form shown in Fig. 1(b), which has a single point η_1 at which $\beta(\eta_1)=0$, and two downstream edges at η_0 and η_2. The leading term $cA_1(\eta)$ in (2.20) can be made regular at η_1 by choosing $a_1=y_0(\eta_0)-y_0(\eta_1)$. Then (2.15) becomes

$$A_1(\eta)=-2\cot\beta(\eta)\int_{\eta_1}^{\eta}\cos\beta(\eta')d\eta'=-2\cot\beta(\eta)[y_0(\eta)-y_0(\eta_1)]. \tag{3.6}$$

By considering the limit of the integral as η tends to η_1, we find that

$$A_1(\eta_1)=-\frac{2}{\kappa(\eta_1)}. \tag{3.7}$$

Here $\kappa(\eta_1)=\beta'(\eta_1)$ is the curvature of B at η_1, which is assumed to be negative. Then $A_1(\eta_1)$ is finite and positive, so the leading term in (2.20) is finite and positive at η_1.

At the next order, c^0, (2.20) contains three terms. The first two are obviously finite at $\eta=\eta_1$. The third term, $A_0(\eta)$, given by (2.19), is proportional to $\cot\beta(\eta)$, which is singular at η_1. We expand its coefficient as follows for η near η_1:

$$\begin{aligned}a_0+\frac{1}{\sin\beta\cos\beta}+\beta'A_1\frac{\cot 2\beta}{\cos^3\beta}&=a_0+\left(\frac{1}{\beta}+O(\eta-\eta_1)\right)-\left(\frac{1}{\beta}+\frac{\beta''(\eta_1)}{2[\beta'(\eta_1)]^2}+O(\eta-\eta_1)\right)\\&=a_0-\frac{\beta''(\eta_1)}{2[\beta'(\eta_1)]^2}+O(\eta-\eta_1).\end{aligned} \tag{3.8}$$

This coefficient will vanish at $\eta=\eta_1$ if a_0 has the value

$$a_0=\frac{\beta''(\eta_1)}{2[\beta'(\eta_1)]^2}=\frac{\kappa'(\eta_1)}{2\kappa^2(\eta_1)}. \tag{3.9}$$

Then $A_0(\eta)$ has a finite limit at η_1.

By using (2.19) for A_0 with a_0 given by (3.9), and (3.6) for A_1, we can write (2.20) as follows:

$$\begin{aligned}p(\xi,\eta)\sim 1+e^{2\xi\cos\beta}\Bigg[&\left(-2c\cot\beta-2\xi^2\beta'\sin\beta-\frac{2\beta'\cot 2\beta}{\sin^2\beta\cos\beta}\right)[y_0(\eta)-y_0(\eta_1)]\\&+2\xi\cos\beta+\frac{1}{\sin^2\beta}+\frac{\kappa'(\eta_1)}{2\kappa^2(\eta_1)}\cot\beta\Bigg],\quad \eta_0<\eta<\eta_2.\end{aligned} \tag{3.10}$$

This expression is finite at η_1. Near the downstream end points η_0 and η_2, p is given by trailing edge boundary layer expansions like that given in Ref. 2. These expansions and (3.10) determine p all along the front side of a boundary such as that in Fig. 1(b), for which $\beta(\eta)$ vanishes at just one point η_1, where $\kappa(\eta_1)<0$. From (3.7) and (3.10) we find that at this point

$$p(0,\eta_1)-1\sim-\frac{2c}{\kappa(\eta_1)}. \tag{3.11}$$

IV. TRAPPING BOUNDARIES

When the boundary B is concave to the left, as in Fig. 2, it can trap particles in the region to the left of B and to the right of $x_0(\eta_0)$. We assume that $x_0(\eta_0)>x_2(\eta_0)$. Therefore, we expect p to be exponentially large in this region, just as it is within a closed boundary. Thus, we assume

that p contains a term of the form (1.4). In addition, we expect there to be a diffusive boundary layer along the upstream side of B. Thus we seek an asymptotic expansion for p to the left of B of the form

$$p\sim e^{2c[x-x_0(\eta_0)]}(ch_1+h_0)+1+e^{2\xi\cos\beta}\left\{cA_1(\eta)+2\xi\cos\beta+\xi^2\beta'\frac{\sin^2\beta}{\cos\beta}A_1(\eta)+A_0(\eta)\right\},$$

$$\eta_2>\eta>\eta_0,\quad \eta\neq\eta_1. \tag{4.1}$$

In (4.1), h_1 and h_0 are constants, so the first term satisfies the differential equation (1.1), satisfies the boundary condition (1.2), and is exponentially small for $x<x_0(\eta_0)$. Consequently, the other terms in (4.1) must satisfy (1.1), (1.2) and tend to 1 at $x=-\infty$. These are just the conditions used to determine the diffusive boundary layer expansion (2.20). Therefore the functions $A_1(\eta)$ and $A_0(\eta)$ are again given by (2.15) and (2.19), respectively.

We shall write (2.15) for A_1 with η_2 instead of η_0 as the end point of integration and we shall keep (2.19) for A_0 unchanged:

$$A_1(\eta)=2\cot\beta(\eta)\left(\int_\eta^{\eta_2}\cos\beta(\eta')d\eta'+a_1\right),\quad \eta\neq\eta_1, \tag{4.2}$$

$$A_0(\eta)=\cot\beta(\eta)\left[a_0+\frac{1}{\sin\beta\cos\beta}+\frac{\beta' A_1\cot 2\beta}{\cos^3\beta}\right],\quad \eta\neq\eta_1. \tag{4.3}$$

At η_1, where $\beta(\eta_1)=0$, these expressions are not valid, and we shall see below that they are singular there. As a consequence, the constants a_1 and a_0 may have different values for $\eta>\eta_1$ and $\eta<\eta_1$. We shall allow this possibility, but at the end of Sec. V we shall see that, in fact, they have the same values in both regions.

Now, as in Sec. III, we represent p near the farthest upstream edge at η_2 by an edge boundary layer expansion, and we match it to (4.1) for η near η_2. The first term in (4.1) is exponentially small at $x_0(\eta_2)$, so it plays no role in the matching. Therefore the analysis is exactly the same as that in Sec. III when the difference in the signs of $\beta(\eta_2)$ and $\beta(\eta_0)$ are taken into account. The result of the matching is

$$a_1=0,\quad a_0=-\frac{1-\sin\beta_2}{\sin\beta_2\cos\beta_2},\quad \beta_2=\beta(\eta_2). \tag{4.4}$$

Thus, A_1 and A_0 are completely determined in the interval $\eta_2>\eta>\eta_1$, and they are both singular at η_1. Therefore we next introduce a new boundary layer expansion in the neighborhood of η_1.

We define the boundary layer variable τ by

$$\tau=c^{1/2}(\eta-\eta_1). \tag{4.5}$$

Then we write the boundary layer expansion of p as

$$p\sim e^{2c[x-x_0(\eta_0)]}(ch_1+h_0)+1+\sum_{j=-1}^{3}c^{j/2}q^{(j)}(\xi,\tau). \tag{4.6}$$

The first term is regular at η_1, so it has been left unchanged. Only the diffusive boundary layer expansion has been modified. The differential equation satisfied by this expansion, which we call q, is (2.3), with ∂_η replaced by $c^{1/2}\partial_\tau$, while the boundary conditions (2.4) and (2.5) are unchanged. In (2.3), we write $\beta(\eta)=\beta(\eta_1+c^{-1/2}\tau)$, and then we expand $\cos\beta(\eta)$, $\sin\beta(\eta)$, $\kappa(\eta)$, and $\kappa'(\eta)$ in powers of $c^{-1/2}$, retaining terms up to c^{-2}.

The resulting equation for q is

$$q_{\xi\xi}-2q_{\xi}+c^{-1}[(\kappa_1^2\tau^2+\kappa_1)q_{\xi}+2\kappa_1\tau q_{\tau}+q_{\tau\tau}]+c^{-3/2}[(\kappa_1\kappa_1'\tau^3+\kappa_1'\tau)q_{\xi}+\kappa_1'\tau^2 q_{\tau}]$$
$$+c^{-2}\left[\tau^4\left(\frac{\kappa_1\kappa_1''}{3}+\frac{(\kappa_1')^2}{4}-\frac{\kappa_1^4}{12}\right)q_{\xi}+\left(\frac{\kappa_1''\tau^2}{2}-\xi\kappa_1^2\right)q_{\xi}+(\kappa_1''-\kappa_1^3)\,\frac{\tau^3}{3}\,q_{\tau}-2\xi\kappa_1^2\tau q_{\tau}\right.$$
$$\left.-2\xi\kappa_1 q_{\tau\tau}\right]+O(c^{-5/2})=0. \tag{4.7}$$

Here $\kappa_1=\kappa(\eta_1)$, $\kappa_1'=\kappa'(\eta_1)$, and $\kappa_1''=\kappa''(\eta_1)$. Similarly we expand $\cos\beta(\eta)$ in the boundary condition (2.4), and we get

$$q_{\xi}-2q+\left[c^{-1}\kappa_1^2\tau^2+c^{-3/2}\kappa_1\kappa_1'\tau^3+c^{-2}\tau^4\left(-\frac{\kappa_1^4}{12}+\frac{\kappa_1\kappa_1''}{3}+\frac{(\kappa_1')^2}{4}\right)+O(c^{-5/2})\right]q$$
$$=2+O(c^{-1}),\quad \text{at}\quad \xi=0. \tag{4.8}$$

Now we substitute the expansion for q, given by the final sum in (4.6), into (4.7) and (4.8). Then in each equation we equate coefficients of the first five powers of $c^{-1/2}$. From (4.7) we obtain, with $L=\partial_{\xi}^2-2\partial_{\xi}$,

$$Lq^{(3)}=0, \tag{4.9}$$

$$Lq^{(2)}=0, \tag{4.10}$$

$$Lq^{(1)}=-(\kappa_1^2\tau^2+\kappa_1)q_{\xi}^{(3)}-2\kappa_1\tau q_{\tau}^{(3)}-q_{\tau\tau}^{(3)}, \tag{4.11}$$

$$Lq^{(0)}=-[\kappa_1\kappa_1'\tau^3q_{\xi}^{(3)}+\kappa_1'\tau q_{\xi}^{(3)}+\kappa_1'\tau^2q_{\tau}^{(3)}]-[\kappa_1^2\tau^2q_{\xi}^{(2)}+\kappa_1q_{\xi}^{(2)}+2\kappa_1\tau q_{\tau}^{(2)}+q_{\tau\tau}^{(2)}] \tag{4.12}$$

$$Lq^{(-1)}=-\left[\tau^4\left(\frac{\kappa_1'\kappa_1''}{3}+\frac{(\kappa_1')^2}{4}-\frac{\kappa_1^4}{12}\right)q_{\xi}^{(3)}+\left(\frac{1}{2}\,\kappa_1''\tau^2-\xi\kappa_1^2\right)q_{\xi}^{(3)}+\left(\frac{\kappa_1''}{3}-\frac{\kappa_1^3}{3}\right)\tau^3q_{\tau}^{(3)}-2\xi(\kappa_1^2\tau q_{\tau}^{(3)}\right.$$
$$\left.+\kappa_1q_{\tau\tau}^{(3)})\right]-[\kappa_1\kappa_1'\tau^3q_{\xi}^{(2)}+\kappa_1'\tau q_{\xi}^{(2)}+\kappa_1'\tau^2q_{\tau}^{(2)}]-[\kappa_1^2\tau^2q_{\xi}^{(1)}+\kappa_1q_{\xi}^{(1)}+2\kappa_1\tau q_{\tau}^{(1)}+q_{\tau\tau}^{(1)}]. \tag{4.13}$$

From (4.8) we get at $\xi=0$, with $L_1=\partial_{\xi}-2$,

$$L_1q^{(3)}=0, \tag{4.14}$$

$$L_1q^{(2)}=0, \tag{4.15}$$

$$L_1q^{(1)}+\kappa_1^2\tau^2q^{(3)}=0, \tag{4.16}$$

$$L_1q^{(0)}+\kappa_1^2\tau^2q^{(2)}+\kappa_1\kappa_1'\tau^3q^{(3)}=2, \tag{4.17}$$

$$L_1q^{(-1)}+\kappa_1^2\tau^2q^{(1)}+\kappa_1\kappa_1'\tau^3q^{(2)}+\tau^4\left[\frac{1}{4}\,(\kappa_1')^2+\frac{\kappa_1\kappa_1''}{3}-\frac{\kappa_1^2}{12}\right]q^{(3)}=0. \tag{4.18}$$

In addition, (2.5) yields

$$q^{(j)}\to 0,\quad \text{as}\quad \xi\to-\infty. \tag{4.19}$$

The solution $q^{(3)}$ of (4.9) satisfying (4.14) and (4.19) is, with $B_3(\tau)$ arbitrary,

$$q^{(3)}(\xi,\tau)=B_3(\tau)e^{2\xi}. \tag{4.20}$$

We use (4.20) in (4.11), integrate from $\xi=-\infty$ to $\xi=0$, and use (4.16) to obtain the solvability condition

$$\tfrac{1}{2}B_3''+\kappa_1\tau B_3'+\kappa_1 B_3=0. \tag{4.21}$$

The solution of (4.21) is, with b_3 and d_3 arbitrary,

$$B_3(\tau)=b_3e^{-\kappa_1\tau^2}\int_0^\tau e^{\kappa_1u^2}du+d_3e^{-\kappa_1\tau^2}. \tag{4.22}$$

Since $d_3e^{-\kappa_1\tau^2}$ is exponentially small as $\tau\to\pm\infty$, it will play no role in the matching. This term may be viewed as an exponentially small correction to the first term in (4.6), and can thus be neglected.

Now that the solvability condition (4.21) is satisfied, (4.11) has a solution satisfying (4.16) and (4.19). To find it we use (4.20)–(4.22) to write (4.11) in the form

$$q^{(1)}_{\xi\xi}-2q^{(1)}_{\xi}=-2\kappa_1^2\tau^2\beta_3(\tau)e^{2\xi}. \tag{4.23}$$

The solution of (4.23) and the boundary conditions is, with $B_1(\tau)$ arbitrary,

$$q^{(1)}(\xi,\tau)=[-\kappa_1^2\tau^2B_3(\tau)\xi+B_1(\tau)]e^{2\xi}. \tag{4.24}$$

Similarly, the solution of (4.10), (4.15), and (4.19) is

$$q^{(2)}(\xi,\tau)=B_2(\tau)e^{2\xi}. \tag{4.25}$$

By using (4.20) and (4.25) in (4.12), integrating from $\xi=-\infty$ to $\xi=0$, and using (4.17), we get the solvability condition

$$\frac{1}{2}B_2''+\kappa_1\tau B_2'+\kappa_1B_2=-2-\kappa_1'\left(\frac{\tau^2}{2}B_3\right)'. \tag{4.26}$$

The solution $B_2(\tau)$ of (4.26) is, when we neglect the term $d_3e^{-\kappa_1\tau^2}$ in (4.22) for B_3,

$$B_2(\tau)=-\frac{2}{\kappa_1}+b_2e^{-\kappa_1\tau^2}\int_0^\tau e^{\kappa_1u^2}du-b_3\frac{\kappa_1'}{3}e^{-\kappa_1\tau^2}\int_0^\tau(\tau^3-u^3)e^{\kappa_1u^2}du+d_2e^{-\kappa_1\tau^2}. \tag{4.27}$$

Matching the boundary layer expansion with (4.1) requires that $b_2=0$, as we will verify in Sec. V. Thus, we set $b_2=0$, and we also take $d_2=0$ because it represents an exponentially small correction to the first term in (4.6), just like the d_3 term in (4.22),

$$B_2(\tau)=-\frac{2}{\kappa_1}-b_3\frac{\kappa_1'}{3}e^{-\kappa_1\tau^2}\int_0^\tau(\tau^3-u^3)e^{\kappa_1u^2}du. \tag{4.28}$$

Finally to determine $B_1(\tau)$ in (4.24), we use (4.28), (4.24), and (4.25) in (4.13), integrate from $\xi=-\infty$ to $\xi=0$, and use (4.18). After some calculation, we get the solvability condition

$$B_1'+2\kappa_1\tau B_1=-\frac{b_3}{2}\kappa_1^2\tau^2-\kappa_1'\tau^2B_2+(\kappa_1^3-\kappa_1'')\frac{\tau^3}{3}B_3. \tag{4.29}$$

The solution of (4.29) is

$$B_1(\tau)=b_1\int_0^\tau e^{\kappa_1(u^2-\tau^2)}du+\left(\frac{2\kappa_1'}{\kappa_1}-\frac{b_3}{2}\kappa_1^2\right)\int_0^\tau u^2e^{\kappa_1(u^2-\tau^2)}du+\frac{b_3}{12}(\kappa_1^3-\kappa_1'')$$
$$\times\int_0^\tau(\tau^4-u^4)e^{\kappa_1(u^2-\tau^2)}du+\frac{b_3}{18}(\kappa_1')^2\int_0^\tau(\tau^3-u^3)^2e^{\kappa_1(u^2-\tau^2)}du. \tag{4.30}$$

Now we have found the first three terms $q^{(3)}$, $q^{(2)}$, and $q^{(1)}$ in the boundary layer expansion (4.6) about the point η_1. They are given by (4.20), (4.24), and (4.25). In these results, B_3 is given by (4.22) with $d_3=0$ and b_3 undetermined, B_2 is given by (4.28), and B_1 by (4.30) with b_1 undetermined. In Sec. V, we will match (4.6) with (4.1) and thereby determine b_3 and b_1, as well as the values of a_1 and a_0 for $\eta<\eta_1$.

V. MATCHING THE EXPANSION ABOUT η_1 TO THE DIFFUSIVE BOUNDARY LAYER EXPANSION

The two expansions (4.1) and (4.6) both represent p on and near the front side of the boundary B. Therefore they must be asymptotically equal near η_1, where both of them are valid. Since the exponential function and the constant 1 are the same in both expansions, we need consider only the final terms, which describe the diffusive boundary layer. We begin by expanding the boundary layer terms in (4.1) for η near η_1. We set $\eta=\eta_1+c^{-1/2}\tau$, use this in (4.1), and expand in powers of $c^{-1/2}$. From (4.2), with $a_1=0$, we obtain

$$A_1(\eta)=A_1(\eta_1+c^{-1/2}\tau)\sim\frac{2sc^{1/2}}{\kappa_1\tau}-\left(\frac{2}{\kappa_1}+\frac{s\kappa_1'}{\kappa_1^2}\right)$$
$$+c^{-1/2}\tau\left(\frac{\kappa_1'}{\kappa_1^2}-\frac{s\kappa_1''}{3\kappa_1^2}-\frac{2s\kappa_1}{3}+\frac{s}{2}\frac{(\kappa_1')^2}{\kappa_1^3}\right)+O(c^{-1}). \tag{5.1}$$

Here

$$s=\int_{\eta_1}^{\eta_2}\cos\beta(u)du=y_0(\eta_2)-y_0(\eta_1). \tag{5.2}$$

Next, from (4.3) we get

$$A_0(\eta)=\frac{c^{3/2}}{\tau^3}\cdot\frac{s}{\kappa_1^2}-\frac{c}{\tau^2}\cdot\frac{s\kappa_1'}{2\kappa_1^3}+\frac{c^{1/2}}{\tau}\left(-\frac{s}{2}-\frac{\kappa_1'}{2\kappa_1^3}+\frac{a_0}{\kappa_1}\right)+O(1). \tag{5.3}$$

We also find that

$$e^{2\xi\cos\beta(\eta)}=e^{2\xi}(1-c^{-1}\xi\kappa_1^2\tau^2+O(c^{-2})), \tag{5.4}$$

$$2\xi\cos\beta=O(1), \tag{5.5}$$

$$\xi^2\beta'\sin^2\beta\,A_1(\eta)/\cos\beta=O(c^{-1/2}). \tag{5.6}$$

Now we combine (5.1), (5.3) and (5.4)–(5.6) to evaluate the boundary layer terms in (4.1):

$$e^{2\xi\cos\beta}\{cA_1+2\xi\cos\beta+\xi^2\beta'\sin^2\beta\,A_1/\cos\beta+A_0\}$$
$$\sim c^{3/2}e^{2\xi}s\left(\frac{2}{\kappa_1\tau}+\frac{1}{\kappa_1^2\tau^3}\right)$$
$$+ce^{2\xi}\left(-\frac{2}{\kappa_1}-\frac{s\kappa_1'}{\kappa_1^2}-\frac{s\kappa_1'}{2\kappa_1^3\tau^2}\right)+c^{1/2}e^{2\xi}\left[-s\xi\left(2\kappa_1\tau+\frac{1}{\tau}\right)+\frac{\kappa_1'\tau}{\kappa_1^2}\right.$$
$$\left.-s\tau\left(\frac{\kappa_1''}{3\kappa_1^2}+\frac{2\kappa_1}{3}-\frac{(\kappa_1')^2}{2\kappa_1^3}\right)+\frac{a_0}{\kappa_1\tau}-\frac{1}{2\tau}\left(s+\frac{\kappa_1'}{\kappa_1^3}\right)\right]. \tag{5.7}$$

This must match the sum $c^{3/2}q^{(3)}(\xi,\tau)+cq^{(2)}(\xi,\tau)+c^{1/2}q^{(1)}(\xi,\tau)$ in (4.6) when this sum is expanded for τ large and positive. Upon comparing powers of c we see that $q^{(3)}$, $q^{(2)}$, and $q^{(1)}$ must match the coefficients of $c^{3/2}$, c, and $c^{1/2}$ in (5.7), respectively. Since each $q^{(j)}$ is proportional to $e^{2\xi}$, this factor cancels out. Then, by using (4.20), (4.25), and (4.24), we can write the matching conditions for $\tau>0$ as follows:

$$B_3(\tau)\sim s\left(\frac{2}{\kappa_1\tau}+\frac{1}{\kappa_1^2\tau^3}\right),\quad \tau\to+\infty, \tag{5.8}$$

$$B_2(\tau)\sim-\frac{2}{\kappa_1}-\frac{s\kappa_1'}{\kappa_1^2}-\frac{s\kappa_1'}{2\kappa_1^3\tau^2}, \tag{5.9}$$

$$-\kappa_1^2\tau^2B_3(\tau)\xi+B_1(\tau)\sim\left[-s\xi\left(2\kappa_1\tau+\frac{1}{\tau}\right)+\frac{\kappa_1'\tau}{\kappa_1^2}-s\tau\left(\frac{\kappa_1''}{3\kappa_1^2}+\frac{2\kappa_1}{3}-\frac{(\kappa_1')^2}{2\kappa_1^3}\right)\right.$$
$$\left.+\frac{a_0}{\kappa_1\tau}-\frac{1}{2\tau}\left(s+\frac{\kappa_1'}{\kappa_1^3}\right)\right]. \tag{5.10}$$

To use (5.8), we expand the integral for B_3 in (4.22) for large τ to obtain

$$B_3(\tau)\sim b_3\left(\frac{1}{2\kappa_1\tau}+\frac{1}{4\kappa_1^2\tau^3}\right). \tag{5.11}$$

Upon using (5.11) in (5.8) we see that the two sides become identical when b_3 has the value

$$b_3=4s\equiv4\int_{\eta_1}^{\eta_2}\cos\,\beta(u)du. \tag{5.12}$$

Next we expand the integral in (4.28) for large τ to get

$$B_2(\tau)\sim-\frac{2}{\kappa_1}-b_3\kappa_1'\left(\frac{1}{4\kappa_1^2}+\frac{1}{8\tau^2\kappa_1^3}\right). \tag{5.13}$$

When b_3 has the value $4s$ given by (5.12), the expression (5.13) for B_2 matches exactly the right side of (5.9). This agreement also confirms that setting $b_2=0$ was correct.

Next we consider (5.10), and use (5.11) for B_3 on the left side with $b_3=4s$. Then the terms linear in ξ on the left side agree exactly with those linear in ξ on the right side. Now we expand $B_1(\tau)$, given by (4.30), for $\tau\to+\infty$. After a long calculation we get

$$B_1(\tau)\sim\tau\left[\frac{\kappa_1'}{\kappa_1^2}+b_3\left(-\frac{\kappa_1}{6}-\frac{\kappa_1''}{12\kappa_1^2}+\frac{(\kappa_1')^2}{8\kappa_1^3}\right)\right]+\frac{1}{\tau}\left[\frac{2b_1}{\kappa_1}+\frac{b_3}{8}-\frac{\kappa_1'}{2\kappa_1^3}\right]. \tag{5.14}$$

We now use (5.14) in (5.10) and set $b_3=4s$. The coefficients of τ match exactly, and those of τ^{-1} match when b_1 has the value

$$b_1=\frac{a_0-\kappa_1 s}{2}. \tag{5.15}$$

Here a_0 is given by (4.4), and s by (5.2). Thus b_1 is known, and the boundary layer about η_1 is completely determined through the terms in (4.6) with positive powers of $c^{1/2}$, i.e., $q^{(3)}$, $q^{(2)}$, and $q^{(1)}$ are determined.

Now for $\tau<0$, we must match this boundary layer to the boundary layer expansion in (4.1). The expansions of B_3, B_2, and B_1 that we have obtained for $\tau>0$ are also valid for $\tau<0$ because B_3 and B_1 are odd functions of τ and B_2 is an even function. The expansion (5.7) is also valid for both positive and negative τ, provided that $a_1=0$. Since the boundary layer expansion matched

(5.7) for $\tau>0$, it also matches for $\tau<0$, provided that a_1 and a_0 have the same values for $\tau<0$ as they do for $\tau>0$. Thus, these constants must also have the values given in (4.4) for $\tau<0$, i.e., for $\eta<\eta_1$.

Now we have completed the determination of the diffusive boundary layer expansion in (4.1). The functions A_1 and A_0 in it are given by (4.2) and (4.3), with a_1 and a_0 given by (4.4). Thus,

$$A_1(\eta)=2\cot\beta(\eta)\cdot\int_{\eta}^{\eta_2}\cos\beta(u)du=2\cot\beta(\eta)\cdot[y_0(\eta_2)-y_0(\eta)],\quad \eta\neq\eta_1, \tag{5.16}$$

$$A_0(\eta)=\cot\beta(\eta)\left[\frac{\sin\beta_2-1}{\sin\beta_2\cos\beta_2}+\frac{1}{\sin\beta\cos\beta}+\frac{\beta' A_1\cot 2\beta}{\cos^3\beta}\right],\quad \eta\neq\eta_1. \tag{5.17}$$

The expansion of p around η_1 is given by (4.6). We can rewrite it more fully through terms of order $c^{1/2}$, by using (4.20)–(4.28) and (5.12):

$$p\sim e^{2c[x-x_0(\eta_0)]}(ch_1+h_0)+1+e^{2\xi}\Bigg[(c^{3/2}-c^{1/2}\kappa_1^2\tau^2\xi)4se^{-\kappa_1\tau^2}$$
$$\times\int_0^{\tau}e^{\kappa_1 u^2}du-c\left(\frac{2}{\kappa_1}+\frac{4s\kappa_1'}{3}e^{-\kappa_1\tau^2}\int_0^{\tau}(\tau^3-u^3)e^{\kappa_1 u^2}du\right)+c^{1/2}B_1(\tau)\Bigg]. \tag{5.18}$$

The constant s is defined in (5.2) and B_1 is given by (4.30), with b_1 determined by (5.15).

The constants h_1 and h_0 are still undetermined.

VI. LOWER EDGE BOUNDARY LAYER AND DETERMINATION OF h_1

We now determine a boundary layer expansion near the lower edge at η_0, where the previous expansions are not valid. By matching it to (4.11), we shall determine the constant h_1 in (4.1). The constant h_0 can be found in the same way by continuing the procedure to the next order. As in Sec. III, we introduce $u=c(\eta-\eta_0)$, but instead of (3.1) we write

$$p(\xi,u)=1+cF_1(\xi,u)+F_0(\xi,u)+O(c^{-1}). \tag{6.1}$$

The term proportional to c is needed to match the term $ch_1e^{2c[x-x_0(\eta_0)]}$ in (4.1). Next, we change variables from η to u in (2.3) and (2.4), replacing ∂_η by $c\partial_u$ and writing $\beta(\eta)=\beta(\eta_0+c^{-1}u)$ $=\beta(\eta_0)+\beta'(\eta_0)c^{-1}u+O(c^{-2})$. Then we can write (2.3) and (2.4) as follows:

$$[\partial_\xi^2+\partial_u^2-2\cos\beta_0\partial_\xi+2\sin\beta_0\partial_u]p+c^{-1}[2u\kappa_0\sin\beta_0\partial_\xi-2\xi\kappa_0\sin\beta_0\partial_u+2\kappa_0 u\cos\beta_0\partial_u$$
$$+\kappa_0\partial_\xi-2\xi\kappa_0\partial_u^2]p+O(c^{-2})p=0, \tag{6.2}$$

$$[\partial_\xi-2\cos\beta_0+c^{-1}2\kappa_0 u\sin\beta_0+O(c^{-2})]p=0,\quad \text{at } \xi=0,\quad u>0. \tag{6.3}$$

Substituting (6.1) for p into (6.2) and (6.3), and equating coefficients of c in each equation, yields

$$[\partial_\xi^2+\partial_u^2-2\cos\beta_0\partial_\xi+2\sin\beta_0\partial_u]F_1(\xi,u)=0, \tag{6.4}$$

$$[\partial_\xi-2\cos\beta_0]F_1=0,\quad \text{at } \xi=0,\quad u>0. \tag{6.5}$$

The coefficients of c^0 give

$$[\partial_\xi^2+\partial_u^2-2\cos\beta_0\partial_\xi+2\sin\beta_0\partial_u]F_0(\xi,u)$$
$$=[2\xi\kappa_0\partial_u^2+2\xi\kappa_0\sin\beta_0\partial_u-2\kappa_0 u\cos\beta_0\partial_u-\kappa_0\partial_\xi-2\kappa_0 u\sin\beta_0\partial_\xi]F_1(\xi,u), \tag{6.6}$$

$$[\partial_\xi - 2\cos\beta_0]F_0 = 2\cos\beta_0 - 2\kappa_0 u \sin\beta_0 \cdot F_1, \quad \text{at } \xi=0, \quad u>0. \tag{6.7}$$

Now we expand p given by (4.1) for η near η_0 by writing $\eta=\eta_0+c^{-1}u$ and expanding in power of c^{-1}:

$$p-1\sim e^{2\xi\cos\beta_0}\left(1-\frac{2u\xi\kappa_0\sin\beta_0}{c}\right)\left\{e^{-2u\sin\beta_0}(ch_1+h_0)\left(1-\frac{u^2}{c}\kappa_0\cos\beta_0\right)+cA_1(\eta_0)\right.$$
$$\left.+uA_1'(\eta_0)+2\xi\cos\beta_0+\frac{\xi^2\kappa_0}{\cos\beta_0}\sin^2\beta_0 A_1(\eta_0)+A_0(\eta_0)\right\}. \tag{6.8}$$

We require that p given by (6.1) be asymptotically equal to p given by (6.8). Then, from the coefficients of c and c^0 we obtain

$$F_1(\xi,u)\sim e^{2\xi\cos\beta_0}\{h_1e^{-2u\sin\beta_0}+A_1(\eta_0)\}, \quad \text{as } u\to+\infty, \tag{6.9}$$

$$F_0(\xi,u)\sim e^{2\xi\cos\beta_0}\left\{-2u\xi\kappa_0\sin\beta_0(h_1e^{-2u\sin\beta_0}+A_1(\eta_0))+e^{-2u\sin\beta_0}(h_0-h_1u^2\kappa_0\cos\beta_0)\right.$$
$$\left.+A_1'(\eta_0)u+2\xi\cos\beta_0+\frac{\xi^2\kappa_0\sin^2\beta_0}{\cos\beta_0}A_1(\eta_0)+A_0(\eta_0)\right\}, \quad \text{as } u\to+\infty. \tag{6.10}$$

To find F_1 we write

$$F_1(\xi,u)=e^{\xi\cos\beta_0-u\sin\beta_0}G_1(\xi,u). \tag{6.11}$$

Then from (6.4) and (6.5) we get the following conditions on G_1:

$$(\partial_\xi^2+\partial_u^2-1)G_1=0, \tag{6.12}$$

$$(\partial_\xi-\cos\beta_0)G_1=0, \quad \text{at } \xi=0, \quad u>0. \tag{6.13}$$

A solution of (6.12) can be expressed in terms of K_0, the modified Bessel function, in the form

$$G_1(\xi,u)=(\partial_\xi+\cos\beta_0)\int_0^\infty \mathscr{G}_1(\omega)K_0([(u-\omega)^2+\xi^2]^{1/2})d\omega. \tag{6.14}$$

The function $\mathscr{G}_1(\omega)$ is arbitrary so far.

Substitution of (6.14) into the boundary condition (6.13) and the use of (6.12) yields

$$\left(-\frac{d^2}{du^2}+\sin^2\beta_0\right)\int_0^\infty \mathscr{G}_1(\omega)K_0(|\omega-u|)d\omega=0, \quad u>0,$$
$$\equiv\phi(u), \quad u<0. \tag{6.15}$$

The boundary condition holds for $u>0$, and we define $\phi(u)$ to be the value of the left side of (6.15) for $u<0$. The problem (6.15) for $\mathscr{G}_1(\omega)$ is of the Wiener–Hopf type, so to solve it we introduce the Fourier transforms,

$$\tilde{\mathscr{G}}_1(\alpha)=\int_0^\infty e^{i\alpha\omega}\mathscr{G}_1(\omega)d\omega, \quad \phi(\alpha)=\int_{-\infty}^0 e^{i\alpha\omega}\phi(\omega)d\omega. \tag{6.16}$$

Then we Fourier transform (6.15) to obtain

$$(\alpha^2+\sin^2\beta_0)\tilde{\mathscr{G}}_1(\alpha)\frac{\pi}{\sqrt{\alpha^2+1}}=\tilde{\phi}(\alpha). \tag{6.17}$$

From (6.16) we assume that $\tilde{\mathscr{G}}_1(\alpha)$ is analytic for Im $\alpha \geqslant \varepsilon_1$ and that $\tilde{\phi}(\alpha)$ is analytic for Im $\alpha \leqslant \varepsilon_2$, for some positive ε_1 and ε_2. Now we factor the denominator of (6.17) and rewrite (6.17) as

$$\frac{\pi(\alpha^2+\sin^2\beta_0)}{\sqrt{\alpha+i}}\,\tilde{\mathscr{G}}_1(\alpha)=\sqrt{\alpha-i}\,\tilde{\phi}(\alpha). \tag{6.18}$$

The right side of (6.18) is analytic for Im $\alpha \leqslant 1$ and the left side is analytic for Im $\alpha \geqslant |\sin\beta_0|$, so both sides are analytic for all α and are therefore equal to a constant D_1. Thus

$$\tilde{\mathscr{G}}_1(\alpha)=\frac{D_1\sqrt{\alpha+i}}{\pi(\alpha^2+\sin^2\beta_0)}. \tag{6.19}$$

Inverting the Fourier transform (6.16) then yields

$$\mathscr{G}_1(u)=\frac{D_1}{2\pi^2}\int_{\mathscr{C}}\frac{\sqrt{\alpha+i}}{(\alpha^2+\sin^2\beta_0)}\,e^{-i\alpha u}d\alpha. \tag{6.20}$$

We choose the contour $\mathscr{C}$ to be a horizontal line above the branch point at $\alpha=-i$ and above the two poles of the integrand, so on $\mathscr{C}$, Im $\alpha>|\sin\beta_0|$.

Now we use the solution (6.20) in (6.14) to get

$$G_1(\xi,u)=(\partial_\xi+\cos\beta_0)\,\frac{D_1}{2\pi^2}\int_{\mathscr{C}}\frac{\sqrt{\alpha+i}}{(\alpha^2+\sin^2\beta_0)}\int_0^\infty K_0([\xi^2+(u-\omega)^2]^{1/2})e^{-i\alpha\omega}d\omega d\alpha. \tag{6.21}$$

For $\omega<0$ we can evaluate the integral with respect to α by closing the contour $\mathscr{C}$ in the upper half-plane. Because of the regularity of the integrand above $\mathscr{C}$, we obtain

$$\int_{\mathscr{C}}e^{-i\alpha\omega}\,\frac{\sqrt{\alpha+i}}{\alpha^2+\sin^2\beta_0}\,d\alpha=0. \tag{6.22}$$

Therefore we can extend the ω integration to $-\infty$ in (6.21) without changing the value of the right side. Then we can evaluate the ω integral by setting $\omega=u-\zeta$. First we find that for $|\text{Im }\alpha|<1$,

$$\begin{aligned}\int_{-\infty}^{\infty}K_0([\xi^2+(u-\omega)^2]^{1/2})e^{-i\alpha\omega}d\omega&=\int_{-\infty}^{\infty}K_0([\xi^2+\zeta^2]^{1/2})e^{i\alpha\zeta}d\zeta\cdot e^{-i\alpha u}\\&=\frac{\pi e^{-|\xi|\sqrt{1+\alpha^2}}}{\sqrt{1+\alpha^2}}\cdot e^{-i\alpha u},\quad |\text{Im }\alpha|<1.\end{aligned} \tag{6.23}$$

By using (6.23) in (6.21), and choosing $\mathscr{C}$ so that Im $\alpha<1$ and also Im $\alpha>|\sin\beta_0|$, we get

$$G_1(\xi,u)=(\partial_\xi+\cos\beta_0)\,\frac{D_1}{2\pi}\int_{\mathscr{C}}\frac{e^{-i\alpha u-|\xi|\sqrt{1+\alpha^2}}}{(\alpha^2+\sin^2\beta_0)\sqrt{\alpha-i}}\,d\alpha,\quad |\sin\beta_0|<\text{Im }\alpha<1. \tag{6.24}$$

Finally, (6.11) and (6.24) yield the solution $F_1(\xi,u)$ explicitly, except for the undetermined constant D_1.

In order to use the matching condition (6.9) we must evaluate the integral in (6.24) for $u\to+\infty$. To do so we consider a parallel contour $\mathscr{C}'$ on which $-1<\text{Im }\alpha<-|\sin\beta_0|$. This contour lies below the two poles of the integrand, but above the branch point at $-i$. On it the factor $e^{-i\alpha u}$ has the absolute value $e^{-u|\text{Im }\alpha|}$, so the integral along $\mathscr{C}'$ tends to zero exponentially as $u\to+\infty$. Therefore to evaluate the integral along $\mathscr{C}$ in (6.24) we shift the contour to $\mathscr{C}'$ and evaluate the residues at the poles:

$$\int_{\mathscr{C}} \cdots d\alpha = \int_{\mathscr{C}'} \cdots d\alpha - 2\pi i[\mathrm{res}(i \sin \beta_0) + \mathrm{res}(-i \sin \beta_0)]$$

$$= \int_{\mathscr{C}'} \cdots d\alpha - \pi e^{i\pi/4} \frac{e^{\xi \cos \beta_0}}{\sin \beta_0} \left[\frac{e^{u \sin \beta_0}}{\sqrt{1-\sin \beta_0}} - \frac{e^{-u \sin \beta_0}}{\sqrt{1+\sin \beta_0}} \right]. \tag{6.25}$$

We substitute (6.25) for the integral in (6.24) and we use the result in (6.11) for F_1. For $u \to +\infty$ we can neglect the integral over $\mathscr{C}'$, and thus we obtain

$$F_1(\xi,u) \sim \cot \beta_0 (-D_1 e^{i\pi/4}) e^{2\xi \cos \beta_0} \left[\frac{1}{\sqrt{1-\sin \beta_0}} - \frac{e^{-2u \sin \beta_0}}{\sqrt{1+\sin \beta_0}} \right], \quad u \to +\infty. \tag{6.26}$$

Now we use (6.26) in the matching condition (6.9). The factor $e^{2\xi \cos \beta_0}$ cancels, and the remaining factors yield the equation

$$\cot \beta_0 (D_1 e^{i\pi/4}) \left[\frac{e^{-2u \sin \beta_0}}{\sqrt{1+\sin \beta_0}} - \frac{1}{\sqrt{1-\sin \beta_0}} \right] = h_1 e^{-2u \sin \beta_0} + A_1(\eta_0). \tag{6.27}$$

Equating the terms independent of u determines D_1 in terms of A_1, which is given by (5.16):

$$D_1 = -A_1(\eta_0) e^{-i\pi/4} \tan \beta_0 \sqrt{1-\sin \beta_0} = -2e^{-i\pi/4} \sqrt{1-\sin \beta_0} \int_{\eta_0}^{\eta_2} \cos \beta(\eta) d\eta. \tag{6.28}$$

Then equating the terms involving $e^{-2u \sin \beta_0}$ determines h_1:

$$h_1 = -2 \cot \beta_0 \sqrt{\frac{1-\sin \beta_0}{1+\sin \beta_0}} \int_{\eta_0}^{\eta_2} \cos \beta(\eta) d\eta = -2 \cot \beta_0 \sqrt{\frac{1-\sin \beta_0}{1+\sin \beta_0}} [y_0(\eta_2) - y_0(\eta_0)]. \tag{6.29}$$

Since $\beta_0 < 0$ we see that $h_1 > 0$.

This completes the determination of the leading terms in the edge boundary layer expansion, and in the coefficient of the exponential term in (4.1). It shows that h_1 is proportional to the width of the boundary B. The next term F_0 in the edge boundary layer expansion can be found by solving (6.6) and (6.7). Then the matching condition (6.10) will determine an arbitrary constant D_0 in the expression for F_0, and the constant h_0 in (4.1).

VII. THE SPECIAL CASE $x_0(\eta_0) = x_0(\eta_2)$

In the special case in which $x_0(\eta_0) = x_0(\eta_2)$, the preceding analysis does not apply. This is because we assumed that $x_0(\eta_0) > x_0(\eta_2)$ in treating the boundary layer at the upper edge. Then the first term in (4.1) was exponentially small at η_2, and played no role in the matching. When $x_0(\eta_0) = x_0(\eta_2)$ this term must be included in the matching, just as it was included at the lower edge η_0 in (6.8), which led to the matching condition (6.9) and ultimately to (6.27). When this term is included in the matching condition at the upper edge η_2, the following equation, analogous to (6.27), is obtained:

$$-\cot \beta_2 (D_1^* e^{-i\pi/4}) \left[\frac{e^{2w \sin \beta_2}}{\sqrt{1-\sin \beta_2}} - \frac{1}{\sqrt{1+\sin \beta_2}} \right] = h_1 e^{2w \sin \beta_2} + A_1(\eta_2). \tag{7.1}$$

Here D_1^* is the arbitrary constant multiplier in F_1 in the upper edge boundary layer expansion, analogous to D_1 in (6.26), and $w = c(\eta_2 - \eta)$.

We now equate terms independent of w in (7.1) and use (4.2) for A_1 to get

$$D_1^* = A_1(\eta_2) e^{i\pi/4} \tan \beta_2 \sqrt{1+\sin \beta_2} = 2e^{i\pi/4} a_1 \sqrt{1+\sin \beta_2}. \tag{7.2}$$

Then the terms proportional to $e^{2w \sin \beta_2}$ yield

$$h_1 = -D_1^* e^{-i\pi/4} \cot \beta_2 / \sqrt{1-\sin \beta_2} = -2a_1 \cot \beta_2 \sqrt{\frac{1+\sin \beta_2}{1-\sin \beta_2}}. \tag{7.3}$$

The constant a_1 in these equations is not yet determined. To determine it, and the constant D_1, we use (6.27).

Equating the terms independent of u in (6.27) yields the first equality in (6.28). Now $A_1(\eta_2)$ is given by (4.2) with $a_1 \neq 0$, so

$$D_1 = -2e^{-i\pi/4} \sqrt{1-\sin \beta_0} \left(\int_{\eta_0}^{\eta_2} \cos \beta(\eta) d\eta + a_1 \right). \tag{7.4}$$

Then the terms in (6.27) independent of u yield h_1 in terms of D_1. The use of (7.4) gives

$$h_1 = -2 \cot \beta_0 \sqrt{\frac{1-\sin \beta_0}{1+\sin \beta_0}} \left(\int_{\eta_0}^{\eta_2} \cos \beta(\eta) d\eta + a_1 \right). \tag{7.5}$$

Equating the two expressions (7.5) and (7.3) for h_1 gives an equation for a_1 that has the solution

$$a_1 = -\int_{\eta_0}^{\eta_2} \cos \beta(\eta) d\eta \left[1 - \frac{\tan \beta_0}{\tan \beta_2} \left(\frac{1+\sin \beta_2}{1-\sin \beta_2} \right)^{1/2} \left(\frac{1+\sin \beta_0}{1-\sin \beta_0} \right)^{1/2} \right]^{-1}. \tag{7.6}$$

By using (7.6) in (7.7) we get

$$h_1 = 2 \cot \beta_2 \left(\frac{1+\sin \beta_2}{1-\sin \beta_2} \right)^{1/2} \left[1 - \frac{\tan \beta_0}{\tan \beta_2} \left(\frac{1+\sin \beta_2}{1-\sin \beta_2} \right)^{1/2} \left(\frac{1+\sin \beta_0}{1-\sin \beta_0} \right)^{1/2} \right]^{-1} [y_0(\eta_2) - y_0(\eta_0)]. \tag{7.7}$$

With D_1^* given by (7.2) and D_1 by (7.4), the leading terms in the asymptotic expansion for p in the edge boundary layers are determined. The constant h_1 given by (7.7) completes the determination of the diffusive boundary layer in (4.1). The boundary layer around η_1, constructed in Sec. IV, was matched to the diffusive boundary layer in Sec. V with $a_1 = 0$. Since $a_1 \neq 0$ in the present case, it must be taken into account in the matching. Its effect is to change the value of s from that given in (5.2) to

$$s = \int_{\eta_1}^{\eta_2} \cos \beta(\eta) d\eta + a_1 = y_0(\eta_2) - y_0(\eta_0) + a_1. \tag{7.8}$$

Then p in that boundary layer is given by (5.18) with s determined by (7.8).

VIII. AXIALLY SYMMETRIC OBSTACLES

We consider an obstacle axially symmetric about the x axis, which is the direction of the advection velocity. We introduce cylindrical coordinates (y, φ, x), where y is radial distance from the x axis, which is the polar axis, and φ is the angle about the x axis. Because $p = 1$ at infinity and the obstacle B is symmetric about the x axis, the concentration p is independent of φ. Then (1.1) and (1.2) simplify to

$$p_{yy} + y^{-1} p_y + p_{xx} - 2c p_x = 0, \tag{8.1}$$

$$\partial_n p - 2c \cos \beta p = 0, \quad \text{on } B. \tag{8.2}$$

As before, β is the angle between the normal to the obstacle and the positive x axis.

Now we consider the curve of intersection of the obstacle with a plane containing the x axis, and we call it B. We write the equation of B as $[x_0(\eta), y_0(\eta)]$, where η is the arclength along B, measured from the x axis. We define σ and $\xi = c\sigma$ as in Sec. II, and we define $\beta(\eta)$ by (2.2). In these variables, (8.1) and (8.2) become

$$p_{\xi\xi}+[y_0(\eta)+c^{-1}\xi\sin\beta(\eta)]^{-1}[c^{-1}\sin\beta\, p_\xi+c^{-2}(1+c^{-1}\xi\kappa)^{-1}\cos\beta\, p_\eta]$$
$$+c^{-1}\kappa(1+c^{-1}\xi\kappa)^{-1}p_\xi+c^{-2}(1+c^{-1}\xi\kappa)^{-2}p_{\eta\eta}-c^{-3}\xi\kappa'(1+c^{-1}\xi\kappa)^{-3}p_\eta$$
$$-2[\cos\beta p_\xi-c^{-1}(1+c^{-1}\xi\kappa)^{-1}\sin\beta\, p_\eta]=0, \tag{8.3}$$

$$p_\xi-2\cos\beta\, p=0, \quad \text{at } \xi=0. \tag{8.4}$$

We seek a solution $p(\xi,\eta,c)$ of (8.3), (8.4), and (2.5). For c large and $\xi\leqslant 0$, we assume that p has an asymptotic expansion of the form (2.6). We substitute it into (8.3), (8.4), and (2.5) and equate coefficients of c and of c^0. The coefficients of c yield (2.7), (2.10), and (2.11) with $j=1$, so their solution for $Q^{(1)}(\xi,\eta)$ is (2.12). The coefficient of c^0 in (8.3) is

$$Q^{(0)}_{\xi\xi}-2\cos\beta\, Q^{(0)}_\xi=-\left(\kappa+\frac{\sin\beta}{y_0}\right)Q^{(1)}_\xi-2\sin\beta\, Q^{(1)}_\eta. \tag{8.5}$$

We use (2.12) for $Q^{(1)}$ in (8.5) and integrate the resulting equation from $\xi=-\infty$ to $\xi=0$. After making use of (2.10) and (2.11) with $j=0$, we obtain

$$A_1'+\left(\frac{\kappa}{\sin\beta\cos\beta}+\frac{1}{y_0\cos\beta}\right)A_1=-\frac{2\cos^2\beta}{\sin\beta}. \tag{8.6}$$

The solution of (8.6) is

$$A_1(\eta)=-2\cot\beta\left[\int_0^\eta\cos\beta(\eta')e^{\int_\eta^{\eta'}y_0^{-1}(\eta'')\sec\beta(\eta'')d\eta''}d\eta'+a_1e^{-\int^\eta y_0^{-1}(\eta')\sec\beta(\eta')d\eta'}\right]. \tag{8.7}$$

Substitution of (8.7) for $A_1(\eta)$ into (2.12) for $Q^{(1)}(\xi,\eta)$, and the use of (2.12) in (2.6) yields

$$p-1=ce^{2\xi\cos\beta}A_1(\eta)+O(1), \quad \xi\leqslant 0. \tag{8.8}$$

This is the leading-order term in the expansion of $p-1$, with $A_1(\eta)$ given by (8.7), in which the constant a_1 is undetermined. In general, the result (8.8) is not valid at $\eta=0$ or any other point η_1 where $\beta(\eta_1)=0$, since the factor $\cot\beta(\eta)$ in (8.7) becomes infinite there. However, if $\kappa(0)<0$, we can make (8.7) finite and positive there by setting $a_1=0$, and then (8.8) becomes

$$p(\xi,\eta)-1=-2ce^{2\xi\cos\beta(\eta)}\cot\beta(\eta)\int_0^\eta\cos\beta(\eta')e^{-\int_{\eta'}^\eta y_0^{-1}(\eta'')\sec\beta(\eta'')d\eta''}d\eta'+O(1), \quad \xi\leqslant 0. \tag{8.9}$$

At $\eta=0$ we find that $A_1(0)=-1/\kappa(0)$, so that

$$p(0,0)-1=-c/\kappa(0)+O(1). \tag{8.10}$$

The right side is positive when $\kappa(0)<0$, so (8.9) and (8.10) hold only in that case. The three-dimensional result (8.10), which gives $p-1$ on the axis, is smaller by a factor of $\frac{1}{2}$ than the two-dimensional result (3.11). It can be shown that in the axially symmetric case in N dimensions, $p-1$ on the axis is smaller by a factor of $1/(N-1)$ than the two-dimensional result (3.11).

When $\kappa(0)>0$, we must include exponential terms in the expansion and proceed as in Secs. IV–VI. We shall not carry out this calculation here.

ACKNOWLEDGMENTS

D.S.A. and J.B.K. were supported by the Air Force Office of Scientific Research, the National Science Foundation, and the Office of Naval Research. C.K. was supported by National Science Foundation Grant No. DMS-93-00136 and Department of Energy Grant No. DE-FG02-93ER25168.

[1] J. B. Keller and C. Knessl, ''Advection–diffusion past a strip. I. Normal incidence,'' J. Math. Phys. **38**, 267–282 (1997).
[2] J. B. Keller and C. Knessl, ''Advection–diffusion past a strip. II. Oblique incidence,'' J. Math. Phys. **38**, 902–925 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 7 JULY 1998

A linearizing transformation for the Korteweg–de Vries equation; generalizations to higher-dimensional nonlinear partial differential equations

H. J. S. Dorren[a)]
Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 12 September 1997; accepted for publication 10 April 1998)

It is shown that the Korteweg–de Vries (KdV) equation can be transformed into an ordinary linear partial differential equation in the wave number domain. Explicit solutions of the KdV equation can be obtained by subsequently solving this linear differential equation and by applying a cascade of (nonlinear) transformations to the solution of the linear differential equation. It is also shown that similar concepts apply to the nonlinear Schrödinger equation. The role of symmetry is discussed. Finally, the procedure which is followed in the one-dimensional cases is successfully applied to find special solutions of higher-dimensional nonlinear partial differential equations. © *1998 American Institute of Physics.*
[S0022-2488(98)01407-8]

I. INTRODUCTION

Inverse scattering transformations (IST) form a powerful tool to solve certain classes of nonlinear partial differential equations (NPDEs). However, the success of the applicability of the IST is, modulo a few exceptions, limited to one-dimensional NPDEs only. This important limitation is caused by the fact that the IST uses the inverse problem of the Schrödinger equation to generate the solutions of the differential equations which have to be solved. The inverse problem of the Schrödinger equation is a well-studied problem in one dimension. Higher-dimensional NPDEs are rarely solved using inverse scattering techniques. There are a few reasons for the restricted applicability of inverse scattering methods in dimensions higher than one. The first reason is that higher-dimensional inverse scattering algorithms, like, for example, the Newton–Marchenko method (the inverse scattering problem of the Schrödinger equation in three-dimensions), are so complicated that is nearly impossible to apply these methods to real data. An alternative to the three-dimensional inverse scattering problem is given by the so-called $\bar{\partial}$ approach, which is successfully generalized to N dimensions (for an overview of the applications we refer to the book by Ablowitz and Clarkson[1]). But in applying the $\bar{\partial}$ approach, we readily face a second important restriction to the application of higher-dimensional inverse scattering methods. This restriction deals with the fact that for higher-dimensional inverse scattering methods the existence of the obtained solutions is difficult to prove. In one-dimensional cases this problem does not occur, since both the scattering data and the potential function depend on one single coordinate. In the three-dimensional case, where the scattering data depend on a three-component wave vector measured at a unit sphere, the five-dimensional data are mapped onto a three-dimensional potential function. As a result of this, in the three-dimensional inverse scattering problem two variables are redundant. The redundancy problem puts strong constraints on the classes of potential functions to be reconstructed and introduces additional complications in the application of inverse scattering methods for solving higher-dimensional NPDEs.

Another interesting and powerful approach which was successfully applied to other classes of NPDEs was developed by Calogero and Eckhaus.[2,3] It was shown by these authors that large classes of NPDEs can be transformed into linear partial differential equations by applying a cascade of limiting procedures involving rescaling techniques and asymptotic expansions. From now on we take over the Calogero terminology by denoting NPDEs which can be linearized using

a)Electronic mail: H.J.S.Dorren@ele.tue.nl

 © 1998 American Institute of Physics

limiting procedures ''*C*-integrable NPDEs,'' whereas those NPDEs which can be solved using inverse scattering techniques are called ''*S* integrable.'' Calogero[2] suggested that because the limiting procedures mentioned above all preserve the integrability, perhaps one universal equation follows by limiting procedures form large classes (all?) NPDEs. Keeping this idea in mind, we can conclude that it is not clear how *S*-integrable NPDEs fit in this concept, since these equations can be transformed into a linear integral equation. There is another important point with respect to *C* integrability which should be mentioned. The concept of *C*-integrability is easily generalized to higher dimensions.[4] An implication of this generalization is that if it is possible to fit in the concept of *S* integrability into the concept of *C*-integrability, an effective method is obtained to find solutions of higher-dimensional NPDEs.

Without aiming to be general, it is shown in this paper that in one dimension *S*-integrable NPDEs are indeed *C*-integrable. We show how to find a cascade of transformations that transform *S*-integrable NPDEs into an ordinary linear differential equation. Moreover, it is shown that in one dimension, there exists a clear relationship between solutions of the linearized equation and the inverse scattering transformation. From the results obtained in this paper, it must be concluded that in one dimension *S* integrability is a special case of *C*-integrability. Moreover, it is shown that by generalizing the ordinary linear differential equation, large classes of other integrable NPDEs can be obtained. All the obtained NPDEs contain an amount of symmetry. This symmetry is also present in its linearization. The most simple example, the linearization of the KdV equation, is only invariant under Galileian transformations. The linearization of the nonlinear Schrödinger equation is also invariant under the SU(2) generators.

This paper has the following structure. In Sec. II we derive a linearization scheme for the KdV equation. We give explicitly the linearized partial differential equation and the cascade of transformation that leads to the KdV-equation. In Sec. III, we solve the linearized equations and derive explicit solutions of the KdV equation using the transformations discovered in Sec. II. As an explicit example, the soliton solutions are constructed. In Sec. IV, the relation of the linearization method and the IST is highlighted. It is shown that the IST can be regarded as a special case of the linearizing procedure described in Sec. II. In Sec. V, the relation between more general one-dimensional differential equations and a generalization of the linearization procedure of Sec. II is discussed. As an explicit example, the nonlinear Schrödinger equation is investigated. In Sec. VI the concepts of Sec. II are generalized to more dimensions. As an example the three-dimensional equivalent of the KdV equation is investigated. We conclude this paper with a discussion.

II. A LINEARIZING TRANSFORMATION FOR THE KORTEWEG-DE VRIES EQUATION

As a starting point we consider the KdV equation:

$$u_{xxx}+u_t=6u_xu, \qquad u(x,t=0)=u_0(x). \tag{1}$$

It is well known that the KdV equation can be transformed into a linear integral equation.[5,6] We can ask ourselves the question whether it is also possible to transform the KdV equation directly into a linear differential equation. Without loss of generality, we can decompose a solution $u(x,t)$ of Eq. (1) in an infinite series of functions:

$$u(x,t)=\sum_{n=1}^{\infty} f^{(n)}(x,t). \tag{2}$$

If it is possible to solve all the functions $f^{(n)}(x,t)$ which determine the solution $u(x,t)$ in Eq. (1), we have solved the KdV equation. In this section it is shown that all the functions $f^{(n)}(x,t)$ can be obtained by applying nonlinear transformations to the solution of a linear differential equation. In order to find this differential equation, we substitute Eq. (2) into Eq. (1). We then obtain the following result:

$$\sum_{n=1}^{\infty} f_{xxx}^{(n)}(x,t)+f_t^{(n)}(x,t)=6\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} f_x^{(m)}(x,t)f^{(n)}(x,t),$$
$$\sum_{n=1}^{\infty} f^{(n)}(x,t=0)=\sum_{n=1}^{\infty} f_0^{(n)}(x). \tag{3}$$

Before we continue, we first have to explain the main concept used throughout this paper. The left-hand side of Eq. (3) describes a linear partial differential operator. The difficulties in solving the KdV equation, however, are related to the nonlinear nature of the right-hand side of Eq. (1). The presence of the nonlinear term makes that the standard techniques of substitution of special (exponential) solutions breaks down. We can overcome this problem by performing a perturbation theory. Instead of substituting a regular perturbation series $u(x,t)=\Sigma_{n=1}^{\infty}\epsilon^{(n)}f^{(n)}(x,t)$ into Eq. (1), and equalizing the different orders of ϵ, we treat in this paper the order n in Eq. (2) as the perturbation parameter. We can then find solutions of the differential equation by equalizing the sum of the perturbation parameter on both sides of Eq. (3). To illustrate this idea we take as an example the following solution:

$$u(x,t)=\sum_{n=1}^{\infty} A_n e^{in(kx-\omega t)}. \tag{4}$$

If we substitute, for example, the $n=2$ term of Eq. (4) onto the left-hand side of Eq. (3), we find that this contribution can be compared to the $n=1$ contribution on the right-hand side. If we substitute the full perturbation series, Eq. (4) into Eq. (1), is clear that there are always two equal orders on the right-hand side and the left-hand side which can be compared.

To implement this simple idea in a more general fashion (which is not valid for exponential solutions only) we proceed by equalizing the different orders of $f^{(n)}(x,t)$ in Eq. (3). If we collect the terms $f^{(n)}(x,t)$ of equal order (the sum of the upper indices of the nonlinear terms equals the upper index on the left-hand side), we obtain the following equations:

$$f_{xxx}^{(1)}(x,t)+f_t^{(1)}(x,t)=0,$$
$$f_{xxx}^{(2)}(x,t)+f_t^{(2)}(x,t)=6f_x^{(1)}(x,t)f^{(1)}(x,t),$$
$$f_{xxx}^{(3)}(x,t)+f_t^{(3)}(x,t)=6f_x^{(1)}(x,t)f^{(2)}(x,t)+6f_x^{(2)}(x,t)f^{(1)}(x,t), \tag{5}$$
$$f_{xxx}^{(4)}(x,t)+f_t^{(4)}(x,t)=6f_x^{(1)}(x,t)f^{(3)}(x,t)+6f_x^{(2)}(x,t)f^{(2)}(x,t)+6f_x^{(3)}(x,t)f^{(1)}(x,t),$$
$$\vdots$$

It is clear that if all the orders on the left-hand side of Eq. (5) and the right-hand side of Eq. (5) are summed up, Eq. (3) is retained. To deal with the nonlinearity in Eq. (5), it is convenient to introduce the following ansatz:

$$6f_x^{(n)}(x,t)f^{(m)}(x,t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,t|x',t')f^{(n+m)}(x',t')dx'dt'. \tag{6}$$

The ansatz given in Eq. (6) introduces a constraint on the space of all the possible functions $f^{(n)}(x,t)$. In the following section it will be shown that the solutions we obtain using the linearization method are more general than the solutions obtained using the IST. If we can construct a satisfactory kernel $G(x,t|x',t')$, we can solve all the equations in Eq. (6) simultaneously. If we substitute Eq. (6) into Eq. (5), we find that Eq. (5) transforms into

$$f^{(1)}_{xxx}(x,t)+f^{(1)}_{t}(x,t)=0;$$

$$f^{(2)}_{xxx}(x,t)+f^{(2)}_{t}(x,t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,t|x',t')f^{(2)}(x',t')dx'dt';$$

$$f^{(3)}_{xxx}(x,t)+f^{(3)}_{t}(x,t)=2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,t|x',t')f^{(3)}(x',t')dx'dt';$$

$$\vdots \tag{7}$$

$$f^{(n)}_{xxx}(x,t)+f^{(n)}_{t}(x,t)=(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,t|x',t')f^{(n)}(x',t')dx'dt';$$

$$\vdots$$

From this result we can conclude that we have transformed the KdV equation (1), by using the ansatz (6) and doing some bookkeeping, into an infinite series of equations which all have a similar structure. In order to solve the integration kernel $G(x,t|x',t')$, it is convenient to perform the remaining part of the analysis in the Fourier domain. The Fourier transform of $f^{(n)}(x,t)$ is given by

$$f^{(n)}(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{f}^{(n)}(k)e^{ikz}dk, \quad z=x-\frac{\omega(k)}{k}t. \tag{8}$$

We first solve $f^{(1)}(x,t)$ because this is the solution of a linear partial differential equation. If Eq. (8) is applied on both sides of $f^{(1)}(x,t)$ in Eq. (7), we obtain the following relationship:

$$f^{(1)}_{xxx}(x,t)+f^{(1)}_{t}(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}i[\omega(k)-k^3]\tilde{f}^{(1)}(k)e^{ikz}dk=0. \tag{9}$$

It follows from Eq. (9) that nontrivial solutions $f^{(1)}(x,t)$ exist if the following dispersion relation is satisfied:

$$\omega(k)=-k^3. \tag{10}$$

It is remarkable to conclude that the function $f^{(1)}(x,t)$ constraints the dispersion relation of all the remaining equations given in Eq. (7). As a result of this, the Fourier transform of the left-hand side of all the equations is given in Eq. (7) can be computed simultaneously. We continue by computing the Fourier transform of the right-hand side of Eq. (7). To carry out this calculation, it is useful to assume that the integration kernel $G(x,t|x',t')$ is invariant under translations in time and space:

$$G(x,t|x',t')=G(x-x'|t-t'). \tag{11}$$

Our aim is to compute the Fourier transform $\mathcal{F}(\cdot)$ of the right-hand side of all the equations appearing in Eq. (7):

$$\mathcal{F}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x-x'|t-t')f^{(n)}(x',t')dx'dt'\right). \tag{12}$$

To perform this computation, we first define the Fourier transform $\tilde{G}(k)$ of the integration kernel $G(x-x'|t-t')$:

$$G(x-x'|t-t')=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{G}(k)e^{ik[z-z']}dk. \tag{13}$$

The variable z in Eq. (13) is specified by Eq. (8). If Eq. (8) and Eq. (13) are substituted into Eq. (12) and if the integrations over z' and k' are carried out, we obtain the following result:

$$\mathcal{F}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-x'|t-t')f^{(n)}(x',t')dx'dt' \right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{G}(k)\tilde{f}^{(n)}(k)e^{ikz}dk. \tag{14}$$

Once the Fourier transform of both the left-hand side and the right-hand side of Eq. (7) are computed, we find by combining Eq. (7), Eq. (9), and Eq. (14) that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} i[\omega(k)-k^3]\tilde{f}^{(n)}(k)e^{ikz}dk = (n-1)\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{G}(k)\tilde{f}^{(n)}(k)e^{ikz}dk. \tag{15}$$

If we define

$$g^{(n)}(z,k) = \tilde{f}^{(n)}(k)e^{ikz}, \tag{16}$$

we find that a solution of $g^{(n)}(z,k)$ is given by the following differential equation:

$$\frac{d}{dz}g^{(n)}(z,k) + ikq^{(n)}(z,k) = (n-1)S(k)g^{(n)}(z,k), \quad S(k) = -\frac{\tilde{G}(k)}{k^2}. \tag{17}$$

From this final result, it follows that the cascade of equations given by Eq. (7) in the Fourier domain is equivalent with

$$\begin{aligned}
&\frac{d}{dz}g^{(1)}(z,k) + ikg^{(1)}(z,k) = 0;\\
&\frac{d}{dz}g^{(2)}(z,k) + ikg^{(2)}(z,k) = S(k)g^{(2)}(z,k);\\
&\frac{d}{dz}g^{(3)}(z,k) + ikg^{(3)}(z,k) = 2S(k)g^{(3)}(z,k);\\
&\vdots\\
&\frac{d}{dz}g^{(n)}(z,k) + ikg^{(n)}(z,k) = (n-1)S(k)g^{(n)}(z,k);\\
&\vdots
\end{aligned} \tag{18}$$

From this simple derivation we can conclude that in the Fourier domain the KdV equation can be linearized into an ordinary linear differential equation. If this differential equation is solved we can retain a solution of the KdV equation by using Eq. (6). Explicit solutions of the KdV equation can be computed by solving the following problem:

$$\begin{aligned}
&\frac{d}{dz}g^{(n)}(z,k) + ikg^{(n)}(z,k) = (n-1)S(k)g^{(n)}(z,k),\\
&6f_x^{(n)}(x,t)f^{(m)}(x,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-x'|t-t')f^{(n+m)}(x',t')dx'dt'.
\end{aligned} \tag{19}$$

With this result we have shown that the KdV equation can be transformed into an ordinary linear differential equation. The second equation in Eq. (19) defines a constraint on the solutions $f^{(n)}(x,t)$.

In the following section the equations given above will be solved and solutions of the KdV equation will be computed. The solutions which are obtained by using the linearization method

described above are more general than the solutions obtained using an inverse scattering approach, because the method also applies to nonexponential functions $f^{(n)}(x,t)$. It follows from Eq. (3) that the method introduced in this paper can also be applied if the functions $f^{(n)}(x,t)$ are polynomials. In the following sections it will be shown explicitly that solutions obtained by the IST form a special subclass of the solutions which can be obtained by the linearization method. An important conclusion that already can be drawn from the results of this section is that the linearization method can easily be generalized to higher dimensions. This will be demonstrated later in this paper. It can be concluded from the approach followed in this section that there are two important factors which determine the transformations from the linear ordinary differential equation to the nonlinear partial differential equation. The first factor is the dispersion relation. The dispersion relation defines the linear part of the nonlinear partial differential equation. The second factor is the ansatz given in Eq. (6) which deals with the nonlinearity. By modifying the dispersion relation and the ansatz (6), it is possible to construct other NPDEs which can be linearized into a similar form as presented in Eq. (19). It is an interesting puzzle to find out whether there are other equations (not in the KdV hierarchy) which can be linearized using a similar approach as presented in this paper. In general, one can expect that the method presented in this paper can be used to find solutions of NPDEs for which the nonlinear part consists of a power series. For investigating the integrability of these equations, it is not necessary to find a suitable Lax pair. Explicit examples are given in following sections in which it is shown that generalizations of Eq. (19) can be transformed into other S-integrable NPDEs. For example, it is shown that the nonlinear Schrödinger equation can be linearized into a similar form as given in Eq. (19). Apparently symmetry plays an important role in the linearization process of NPDEs. The linearization of the KdV equation is invariant under Galileian transformations. This invariance is already known for a long time[7] but can be used here to consider the concept of integrability from a different point of view. The Galileian invariance can be regarded as the most simple symmetry in the hierarchy of equations which will follow. This Galileian invariance also explains the presence of solitons. If we introduce more symmetry in the linearized differential equation, we obtain that the linear differential equation can be transformed into more complicated NPDEs.

III. SOLUTIONS OF THE KDV EQUATION

As shown in the previous section, the KdV equation can be linearized into the ordinary linear differential equation (17). In this section we will derive explicit solutions of this equation:

$$\frac{d}{dz}\,g^{(n)}(z,k)+ikg^{(n)}(z,k)=(n-1)S(k)g^{(n)}(z,k). \tag{20}$$

Equation (20) is an ordinary linear first-order differential equation. The general solution of Eq. (20) is given by

$$g^{(n)}(z,k)=C_1e^{M(k,n)z};\quad M(k,n)=(n-1)S(k)-ik. \tag{21}$$

To verify the validity of the ansatz (6), it is convenient to define $S(k)$ in the complex plane. This enables us to solve Eq. (20) using the Green's function technique. If we formulate solutions of Eq. (20) using a Green's function technique, we can find a relationship between the linearization method described in this paper and an inverse scattering method. In order to find a suitable Green's function we first solve the following problem:

$$\frac{d}{d\hat{z}}\,\mathcal{G}(\hat{z},k)+ik\mathcal{G}(\hat{z},k)=\delta(\hat{z}),\quad \hat{z}=z-z', \tag{22}$$

Eq. (22) can be solved easily by applying the following inverse Fourier transform:

$$\tilde{\mathcal{G}}(k',k)=\int_{-\infty}^{\infty}\mathcal{G}(\hat{z},k)e^{-ik'\hat{z}}d\hat{z}. \tag{23}$$

If the Fourier transform (23) is substituted into Eq. (22) we obtain the following result:

$$ik'\tilde{\mathscr{K}}(k',k)+ik.\tilde{\mathscr{K}}(k',k)=1. \tag{24}$$

From Eq. (24), it follows that the solution of $\tilde{\mathscr{K}}(k',k)$ is given by

$$\tilde{\mathscr{K}}(k',k)=\frac{1}{i(k+k')}. \tag{25}$$

By applying the inverse transformation of Eq. (23) we can solve the Green's function $\mathscr{G}(\hat{z},k)$:

$$\mathscr{G}(\hat{z},k)=\lim_{\epsilon\to 0}\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{e^{ik'\hat{z}}}{k+k'+i\epsilon}\,dk'. \tag{26}$$

Once the Green's function is solved, we can formulate the general solution of Eq. (20) in the form of an integral equation:

$$g^{(n)}(z,k)=\lim_{\epsilon\to 0}\frac{(n-1)}{2\pi i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{ik'[z-z']}}{k+k'+i\epsilon}S(k)g^{(n)}(z',k)dz'dk', \quad n\neq 1. \tag{27}$$

From the function $g^{(n)}(z,k)$ given in Eq. (27), we can solve, using Eq. (16) and the Fourier transformation given in Eq. (8), the function $f^{(n)}(x,t)$:

$$f^{(n)}=(x,t)=\lim_{\epsilon\to 0}\frac{(n-1)}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(k-k')}{k+k'+i\epsilon}S(k)g^{(n)}(z,k')dkdk', \quad n\neq 1. \tag{28}$$

Before formulating solutions of the KdV equation, we first have to assure that a kernel $G(x-x'|t-t')$ indeed exists. If we compute the left-hand-side of Eq. (6) using the representation presented in Eq. (28) of the solutions $f^{(n)}(x,t)$, we obtain

$$\begin{aligned}6f_x^{(n)}(x,t)f^{(m)}(x,t)=&-\lim_{\epsilon\to 0}\frac{(n-1)(m-1)}{(2\pi)^4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\\&\times\frac{6ik'S(k)S(k'')}{k+k'+i\epsilon}\frac{\delta(k-k')\delta(k''-k''')}{k''+k'''+i\epsilon}g^{(n)}(z,k')\\&\times g^{(m)}(z,k''')dkdk'dk''dk''', \quad n,m\neq 1.\end{aligned} \tag{29}$$

The kernel $G(x-x'|t-t')$ can be represented in the complex plane according to the following Cauchy formula:

$$G(x-x'|t-t')=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{k'^2e^{ik'[z-z']}}{k+k'+i\epsilon}S(k')dk'dk. \tag{30}$$

In Eq. (30), it is used that $S(k)=k^2\tilde{G}(k)$. The expression obtained in Eq. (29) is, according to Eq. (6), equal to:

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x-x'|t-t')f^{(n+m)}(x',t')dx'dt'\\&=-\lim_{\epsilon\to 0}\frac{[n+m-1]}{(2\pi)^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{k'^2S(k')S(k'')}{k+k'+i\epsilon}\frac{\delta(k''-k''')\delta(k'''-k')}{k''+k'''+i\epsilon}\\&\times g^{(n+m)}(z,k''')dkdk'dk''dk'''.\end{aligned} \tag{31}$$

If follows from comparing Eq. (29) to Eq. (31), that the ansatz is satisfied if the following condition is fulfilled:

$$\frac{3i(n-1)(m-1)}{\pi[n+m-1]}\,S(k)\,\delta(k-k')g^{(n)}(z,k')g^{(m)}(z,k''')=k'S(k')\,\delta(k'''-k')g^{(n+m)}(z,k'''). \tag{32}$$

If Eq. (32) is satisfied, we can finally formulate solutions of the KdV equation:

$$u(x,t)=\sum_{n=1}^{\infty} f^{(n)}(x,t)$$

$$=Ae^{ikz}+\lim_{\epsilon\to 0}\sum_{n=2}^{\infty}\frac{(n-1)}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(k-k')}{k+k'+i\epsilon}\,S(k')g^{(n)}(z,k')dk'dk. \tag{33}$$

We can find explicit solutions of Eq. (33) by using the fact that Eq. (32) defines a recursion relationship for the functions $g^{(n)}(z,k)$. This recursion relation can be made explicit by verifying Eq. (6) if one of the functions $f^{(n)}(x,t)$ is equal to $f^{(1)}(x,t)$. If we choose for $f^{(1)}(x,t)$ the following representation:

$$f^{(1)}(x,t)=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{ik'e^{ik'z}}{k+k'+i\epsilon}\,S(k')dk'dk, \tag{34}$$

the inverse scattering transformation can be retained. From comparing Eq. (28) to Eq. (34), it follows that by giving $f^{(1)}(x,t)$ the form (34), we have implicitly chosen that $\tilde{f}^{(1)}(k')=ik'$. If we compute the left-hand side of Eq. (6), we find that

$$6f_x^{(1)}(x,t)f^{(n-1)}(x,t)=-\lim_{\epsilon\to 0}\frac{(n-2)}{(2\pi)^4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{6ik'S(k)(k'')}{k+k'+i\epsilon}\,\frac{\delta(k-k')\,\delta(k''-k''')}{k''+k'''+i\epsilon}$$

$$\times g^{(1)}(z,k')g^{(m)}(z,k''')dkdk'dk''dk'''. \tag{35}$$

According to the ansatz, Eq. (6), this has to be equal to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x-x'|t-t')f^{(n)}(x',t')dx'dt'$$

$$=-\lim_{\epsilon\to 0}\frac{[n-1]}{(2\pi)^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{k'^2S(k')S(k'')}{k+k'+i\epsilon}\,\frac{\delta(k''-k''')\,\delta(k'''-k')}{k''+k'''+i\epsilon}$$

$$\times g^{(n)}(z,k''')dkdk'dk''dk'''. \tag{36}$$

By comparing Eq. (35) to Eq. (36) we find that equality is obtained if the following relationship is satisfied:

$$\frac{3i(n-2)}{\pi(n-1)}\,S(k)\,\delta(k-k')g^{(1)}(z,k')g^{(n-1)}(z,k''')=k'S(k')\,\delta(k'''-k')g^{(n)}(z,k'''). \tag{37}$$

Equation (37) is a recursion relation which makes it possible to compute $g^{(n)}(z,k)$ from $g^{(n-1)}(z,k)$. In order to relate $g^{(2)}(z,k)$ to $g^{(1)}(z,k)$, we have to check explicitly the ansatz (6) for the $6f_x^{(1)}(x,t)f^{(1)}(x,t)$ case. This leads to the following result:

$$\frac{3i}{2\pi}\,S(k)\,\delta(k-k')g^{(1)}(z,k')g^{(1)}(z,k''')=k'S(K')\,\delta(k''-k''')g^{(2)}(z,k'''). \tag{38}$$

The result we have obtained indicates that we can solve the KdV equation by using the recursion relationship defined by Eq. (37). If we continue the process of iteration, we obtain the following solutions $f^{(n)}(x,t)$ which determine the solution of the KdV equation:

$$f^{(1)}(x,t)=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{ik'e^{ik'z}}{k+k'+i\epsilon}\hat{S}(k')dk'dk, \tag{39}$$

$$f^{(2)}(x,t)=\lim_{\epsilon\to 0}\frac{1}{2\pi(2\pi i)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\hat{S}(k')e^{ik'z}}{k+k'+i\epsilon}\frac{\hat{S}(k'')e^{ik''z}}{k'+k''+i\epsilon}i(k'+k'')dk''dk'dk, \tag{40}$$

$$\vdots$$

where we have used that $\hat{S}(k)=2/3iS(k)$. If the infinite series of solutions of which the first two are given by Eq. (39) and Eq. (40) are substituted into Eq. (33), we obtain a solution of the KdV equation of which the first two terms are given by

$$\begin{aligned}u(x,t)=&\lim_{\epsilon\to 0}\frac{1}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{ik'e^{ik'z}}{k+k'+i\epsilon}\hat{S}(k')dk'dk\\&+\lim_{\epsilon\to 0}\frac{1}{2\pi(2\pi i)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\hat{S}(k')e^{ik'z}}{k+k'+i\epsilon}\frac{\hat{S}(k'')e^{ik''z}}{k'+k''+i\epsilon}i(k'+k'')dk''dk'dk+\cdots .\end{aligned} \tag{41}$$

In the following section it is shown that Eq. (41) describes a generalization of the inverse scattering solution of the KdV equation. From Eq. (41), it is easy to derive the soliton solutions of the KdV equation. This can be done by choosing the following representation of $S(k)$ (see Ref. 8):

$$\hat{S}(k)=\frac{id}{k+2i\beta},\quad \beta>0,\quad d\in\mathbb{R}. \tag{42}$$

We find that the solution (41) reduces to

$$u(x,t)=4d\beta e^{-2(\beta x-4\beta^3 t)}+16d^2 e^{-4(\beta x-4\beta^3 t)}+24\frac{d^3}{\beta}e^{-6(\beta x-4\beta^3 t)}+\cdots . \tag{43}$$

By carrying out the summation in Eq. (43), we can reformulate the solution (43) more compactly:

$$u(x,t)=\frac{8d\beta e^{-2(\beta x-4\beta^3 t)}}{\left(1+\frac{d}{\beta}e^{-2(\beta x-4\beta^3 t)}\right)^2}. \tag{44}$$

Hence, if we set

$$\beta=\frac{1}{2}\sqrt{c},\quad x_0=-\frac{1}{\sqrt{c}}\log\left(-\frac{d}{\beta}\right),\quad d<0, \tag{45}$$

we can simplify Eq. (44) a step further to

$$u(x,t)=-\frac{c}{2}\operatorname{sech}^2\left\{\frac{1}{2}\sqrt{c}(x-ct+x_0)\right\}. \tag{46}$$

Equation (46) describes the well-known KdV soliton.

The results obtained in this section make it clear that the approach suggested in this paper can be used to generate solutions of the KdV equation. For instance, the KdV soliton can be retained as a special case of Eq. (33). In the following section, it is shown that the solutions obtained in this section are, in fact, the inverse scattering solutions. The two-soliton solutions can be obtained by choosing a function $\hat{S}(k)$ which has a double pole. It will be shown in the next section that the

solution (41) is somewhat more general than the inverse scattering solution because the function $S(k)$ is less restricted than the physical scattering data as used in the IST.

IV. RELATION TO THE INVERSE SCATTERING TRANSFORMATION

To show that there is a close relationship between the results obtained by the linearization method and the well-known inverse scattering results, we discuss the latter in this section in more detail. The IST uses the inverse problem of the Schrödinger equation to generate solutions of the KdV equation. The inverse problem of the Schrödinger equation is given by the Marchenko equation:

$$K(x,y,t)+A(x+y,t)+\int_x^\infty K(x,z,t)A(y+z,t)dz=0. \tag{47}$$

The Marchenko equation relates a data function $A(x,t)$ to an integration kernel $K(x,y,t)$. In Eq. (47), the data function $A(x,t)$ is given by

$$A(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} R(k)e^{ik[x+8k^2t]}dt+\sum_{p=1}^{N}\rho_p e^{-k_p[x+4k_p^2t]}. \tag{48}$$

In Eq. (48), the function $R(k)$ describes the physical reflection coefficient. The function $R(k)$ corresponds to the continuous part of the spectrum of the Schrödinger equation. The numbers ρ_p represent the discrete part of the spectrum of the Schrödinger equation corresponding to the discrete eigenvalues k_p. The discrete part of the spectrum of the Schrödinger equation is given by the residues of the transmission coefficient $T(k,t)$ on the positive imaginary axis. The solution $u(x,t)$ of the KdV equation can be solved from the integration kernel $K(x,y,t)$ by using the following relationship:

$$u(x,t)=-2\frac{d}{dx}K(x,x,t). \tag{49}$$

The relation between the linearization method and the IST becomes more transparent if the analysis is performed in the wave number domain. A convenient way to do this is to relate the kernel $K(x,y,t)$ to a function $F(x,k,t)$ by using the following Fourier transform:[9]

$$K(x,y,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{ik(y-x)}(F(x,k,t)-1). \tag{50}$$

It is well known that the function $F(x,k,t)$ is related to the Jost solutions of the Schrödinger equation.[9] If we substitute Eq. (50) into Eq. (47), we obtain the following relationship:

$$F(x,k,t)=1+\int_{-\infty}^{\infty} C(k,k',z)F(x,k',t)dk'. \tag{51}$$

For reasons of simplicity, we do not take the contribution of the transmission coefficient into account in the analysis which follows. It can be shown that the transmission coefficient can be included in the computations in a conceptually similar manner.[9] In the special case of a negligible transmission coefficient, we find that the kernel $C(k,k',t)$ in Eq. (51) is given by

$$C(k,k',z)=\frac{1}{2\pi i}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon};\quad z=x-8k^2t. \tag{52}$$

If Eq. (52) is substituted into Eq. (51), we find the following relationship:

$$F(x,k,t)-1=+\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{R(k')F(x,k',t)e^{2ikz}}{k+k'+i\epsilon}dk'. \tag{53}$$

To find a relationship between the linearization technique and the IST, we expand Eq. (52) in a series:

$$F(k,x,t)=1+\int_{-\infty}^{\infty}C(k,k',z)dk'+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C(k,k',z)C(k',k'',z)dk'dk''+\cdots . \tag{54}$$

If we substitute Eq. (52) into Eq. (54), this result becomes equal to

$$\begin{aligned}F(k,x,t)-1=&\lim_{\epsilon\to 0}\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon}dk'\\&+\lim_{\epsilon\to 0}\frac{1}{(2\pi i)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon}\frac{R(k'')e^{2ik''z}}{k'+k''+i\epsilon}dk'dk''+\cdots .\end{aligned} \tag{55}$$

By taking the Fourier transform Eq. (50) of both sides of Eq. (55) we can compute the integration kernel $K(x,x,t)$:

$$\begin{aligned}K(x,x,t)=&\lim_{\epsilon\to 0^+}\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon}dkdk'\\&+\lim_{\epsilon\to 0^+}\frac{2}{2\pi(2\pi i)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon}\frac{R(k'')e^{2ik''z}}{k'+k''+i\epsilon}dkdk'dk''+\cdots .\end{aligned} \tag{56}$$

From this result we obtain by setting $x=y$ and using the relationship (49) that solutions of the KdV equation are given by the following relationship:

$$\begin{aligned}u(x,t)=&\lim_{\epsilon\to 0}\frac{1}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{2ik'e^{2ik'z}}{k+k'+i\epsilon}R(k')dk'dk\\&+\lim_{\epsilon\to 0}\frac{1}{2\pi(2\pi i)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{R(k')e^{2ik'z}}{k+k'+i\epsilon}\frac{R(k'')e^{2ik''z}}{k'+k''+i\epsilon}2i(k'+k'')dk''dk'dk+\cdots .\end{aligned} \tag{57}$$

Modulo some trivial rescaling, the analytical structure of Eq. (58) is equal to the result reflected in Eq. (41). We can therefore conclude that the functions $g^{(n)}(k,z)$ are related to the Jost solution of an inverse scattering problem. This is not a great surprise if we realize that the kernel $K(x,y,t)$ in the Marchenko equation is a Green's function by itself.[9] The linearization method constructs the Green's function directly. However, we have to realize that the results obtained in Sec. III are far more general than the IST. It turns out that Eq. (41) is only a special case on of the linearization method, depending on the initial conditions. Yet, it should be realized that although Eq. (58) and Eq. (41) have a similar structure; Eq. (41) is more general. This is related to the precise definition of the scattering data in Eq. (58). In general, the scattering data given by the reflection coefficient $R(k,t)$ and the transmission coefficient $T(k,t)$ have to satisfy a number of restrictions to assure a unitary S matrix and uniqueness of the corresponding inverse problem. Moreover, additional conditions have to be satisfied so that the Deift–Trubowits condition is not violated (an overview of all the conditions can be found in, for instance, the book by Chadan and Sabatier[9]). These additional constraints do not have to be satisfied by the results obtained by the linearization method. We can therefore consider the results reflected in Eq. (41) to be a generalization of the IST.

V. THE NONLINEAR SCHRÖDINGER EQUATION

In Sec. II it is shown that the KdV equation can be be transformed into a linear partial differential equation. As concluded in Sec. II, this linear differential equation depends on the balances of the dispersion of the nonlinear problem to a function $S(k)$ representing the nonlinear-

ity. We can ask ourselves the question whether there are more NPDEs which can be linearized into an equation having a similar structure as Eq. (20). In this section it is shown that this is indeed the case if we introduce new symmetries in the linearized equation. The most simple one-dimensional generalization of Eq. (20) is given by

$$c_1\left[\frac{d}{dz}\,g_j^{(n)}(z,k)\right]+ikc_2g_j^{(n)}(z,k)=c_3Q(n)S(k)g_j^{(n)}(z,k). \tag{58}$$

In Eq. (58), we have replaced the function $g^{(n)}(z,k)$ of Eq. (20) by an N-dimensional vector function $g_j^{(n)}(z,k)$. Furthermore, in Eq. (58) the $N\times N$ matrices c_i are introduced. The matrices c_i are assumed to have constant entries. Similarly, as in Sec. II, the variable z takes its value in a dispersion relation so that Eq. (58) is invariant under Galileian transformations. As argued in Sec. II, we can expect that Eq. (58) can be transformed into a wide variety of nonlinear equations which all have solitonic solutions. An interesting special case can be considered if we choose $N=2$ and if the matrices c_i take the following representations:

$$c_1=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\quad c_2=c_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{59}$$

We will show in this section that if we choose the dispersion relation to be equal to

$$\omega(k)=\pm k^2, \tag{60}$$

and if the function $Q(n)$ is equal to

$$Q(n)=\tfrac{1}{2}(n-1)(n-2), \tag{61}$$

the linear equation (58) can be transformed into the nonlinear Schrödinger equation. To find a similar cascade of transformations and rescaling which transforms the nonlinear Schrödinger equation into Eq. (58), we take as a starting point the nonlinear Schrödinger equation:

$$iu_t+u_{xx}\pm 2|u|^2u=0. \tag{62}$$

It is useful to transform the nonlinear Schrödinger equation into the following set of coupled differential equations:

$$iu_t+u_{xx}-2u^2v=0,\quad iv_t-v_{xx}+2v^2u=0. \tag{63}$$

The two coupled partial differential equations presented in Eq. (63) are equivalent to the nonlinear Schrödinger equation if we set $v=\mp u^*$. In order to simplify the bookkeeping, we formulate the set of coupled differential equations (63) as the following matrix equation:

$$\mathbf{I}\mathbf{w}_t-i\sigma_3\mathbf{w}_{xx}=-2i\mathbf{A}\mathbf{w}. \tag{64}$$

In Eq. (64), $\mathbf{I}$ represents the unity matrix and the vector $\mathbf{w}$ is given by $\mathbf{w}=(u,v)^T$. Furthermore, the Pauli spin matrices σ_i are introduced:

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix},\quad \sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{65}$$

Finally, all the nonlinearities are contained in the matrix $\mathbf{A}$ which is defined by the following relationship:

$$\begin{pmatrix}uv & 0\\ 0 & -uv\end{pmatrix}\begin{pmatrix}u\\ v\end{pmatrix}=\mathbf{A}\mathbf{w}. \tag{66}$$

We carry on in a similar manner as in Sec. II by assuming that the functions $u(x,t)$ and $v(x,t)$ can be decomposed in the following infinite perturbation series:

$$\mathbf{w}(x,t)=\sum_{n=1}^{\infty}\mathbf{f}^{(n)}(x,t). \tag{67}$$

Every component $f_i^{(n)}(x,t)$ of $\mathbf{f}^{(n)}(x,t)$ expresses the expansion of the components u and v of $\mathbf{w}(x,t)$. Similarly, as in Sec. II, the nonlinear Schrödinger equation is solved if solutions of all the functions $f_i^{(n)}(x,t)$ are constructed. If Eq. (67) is substituted into Eq. (64), we obtain the following result:

$$\sum_{n=1}^{\infty}(\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(n)}(x,t)=-2i\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty}s^{(klm)}(x,t), \tag{68}$$

where the vector function $\mathbf{s}^{(klm)}(x,t)$ is given by

$$\mathbf{s}^{(klm)}(x,t)=\begin{pmatrix} f^{(k)}(x,t)f_{(l)}(x,t)g^{(m)}(x,t) \\ -g^{(k)}(x,t)g^{(l)}(x,t)f^{(m)}(x,t)\end{pmatrix}. \tag{69}$$

We proceed, as in Sec. II, by collecting all the terms of which the sum of the orders on the left-hand side equals the sum of the orders on the right-hand side. In order to have a better view on the structure of the infinite number of equations which are then obtained, we write out the first four explicitly:

$$\begin{aligned}(\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(1)}(x,t)&=0,\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(2)}(x,t)&=0,\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(3)}(x,t)&=\mathbf{s}^{(111)}(x,t),\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(4)}(x,t)&=\mathbf{s}^{(211)}(x,t)+\mathbf{s}^{(121)}(x,t)+\mathbf{s}^{(112)}(x,t),\end{aligned} \tag{70}$$

We can linearize the set of equations given in Eq. (70) in the Fourier domain by introducing kernel $G(x',t'|x,t)$:

$$-2i\mathbf{s}^{(klm)}(x,t)=\sigma_3\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,x'|t,t')\mathbf{f}^{(k+l+m)}(x',t')dx'dt'. \tag{71}$$

If Eq. (71) is substituted into Eq. (70), we can reformulate Eq. (70) in the following way:

$$\begin{aligned}(\mathbf{I}\partial^t-i\sigma_3\partial_{xx})\mathbf{f}^{(1)}(x,t)&=0,\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(2)}(x,t)&=0,\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(3)}(x,t)&=\sigma_3\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,x'|t,t')\mathbf{f}^{3}(x',t')dx'dt',\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(4)}(x,t)&=3\sigma_3\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,x'|t,t')\mathbf{f}^{(4)}(x',t')dx'dt',\\ &\vdots\\ (\mathbf{I}\partial_t-i\sigma_3\partial_{xx})\mathbf{f}^{(n)}(x,t)&=\sigma_3Q(n)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(x,x'|t,t')\mathbf{f}^{(n)}(x',t')dx'dt'.\end{aligned} \tag{72}$$

Similar to Sec. II, we proceed, continuing our analysis in the wave number domain, by searching for the following wave-package solutions:

$$\mathbf{f}^{(n)}(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{\mathbf{f}}^{(n)}(k)e^{ikz}dk,\quad z=x-\frac{\omega(k)}{k}t. \tag{73}$$

If Eq. (73) is substituted into Eq. (70) and if we solve $\tilde{\mathbf{f}}^{(1)}(k)$, we find that dispersion relation given by Eq. (60) has to be satisfied. If we define, furthermore, the Fourier transform of the kernel $G(x-x'|t-t')$ to be equal to

$$G(x-x'|t-t')=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{G}(k)e^{ik[z-z']}dk, \tag{74}$$

and if we define $S(k)=\tilde{G}(k)/k$, we obtain that the set of equations (72) in the Fourier domain are equal to

$$\begin{aligned}
&\mathbf{I}\frac{d}{dz}\tilde{\mathbf{g}}^{(1)}(z,k)+ik\sigma_3\tilde{\mathbf{g}}^{(1)}(z,k)=0,\\
&\mathbf{I}\frac{d}{dz}\tilde{\mathbf{g}}^{(2)}(z,k)+ik\sigma_3\tilde{\mathbf{g}}^{(2)}(z,k)=0,\\
&\mathbf{I}\frac{d}{dz}\tilde{\mathbf{g}}^{(3)}(z,k)+ik\sigma_3\tilde{\mathbf{g}}^{(3)}(z,k)=\sigma_3S(k)\tilde{\mathbf{g}}^{(3)}(z,k),\\
&\vdots
\end{aligned} \tag{75}$$

In Eq. (75), we have defined

$$\tilde{\mathbf{g}}^{(n)}(z,k)=\tilde{\mathbf{f}}^{(n)}(k)e^{ikz}\quad S(k)=\frac{G(k)}{k}. \tag{76}$$

Resuming, we can conclude that the nonlinear Schrödinger equation can be linearized into

$$\mathbf{I}\frac{d}{dz}\tilde{\mathbf{g}}^{(n)}(z,k)+ik\sigma_3\tilde{\mathbf{g}}^{(n)}(z,k)=Q(n)\sigma_3S(k)\tilde{\mathbf{g}}^{(n)}(z,k). \tag{77}$$

In this section it is shown that the nonlinear Schrödinger equation can be linearized into a generalization of Eq. (20). It can be shown that the ansatz (71) leads to conditions on the solutions of the nonlinear Schrödinger equation.[10] Explicit solutions of the nonlinear Schrödinger equation computed by using the linearization method are given in this reference. It has to be remarked that the linearized equation is invariant under transformations of the SU(2) generators σ_i. This reflects that besides the Galileian invariance (which takes its value as the dispersion relation) another symmetry is present in the linearization of the nonlinear Schrödinger equation.

VI. GENERALIZATION TO HIGHER DIMENSIONS

Inspired by the results obtained in the previous sections, we show that a similar procedure can be applied to find solutions of higher-dimensional NPDEs. We choose as a starting point the most simple three-dimensional generalization of Eq. (20):

$$\sum_{i=1}^{3}\partial_{z_i}g^{(n)}(\mathbf{z},\mathbf{k})+i\sum_{i=1}^{3}k_ig^{(n)}(\mathbf{z},\mathbf{k})=(n-1)S(\mathbf{k})g^{(n)}(\mathbf{z},\mathbf{k}). \tag{78}$$

In Eq. (78), the solution $g^{(n)}(\mathbf{z},\mathbf{k})$ of the linearized problem depends now on a three-dimensional vector $\mathbf{z}$ having components $(z_1,z_2,z_3)^T$ and a three-dimensional wave number $\mathbf{k}$ with elements $(k_1,k_2,k_3)^T$. We derive a three-dimensional generalization of the KdV equation by following the reverse track of Sec. II. We first define a function $\hat{G}(\mathbf{k})$ so that

$$S(\mathbf{k})=\frac{\hat{G}(\mathbf{k})}{\mathbf{k}\cdot\mathbf{k}}. \tag{79}$$

If Eq. (79) is substituted into Eq. (78), we obtain the following result:

$$\mathbf{k}\cdot\mathbf{k}\sum_{i=1}^{3}\{\partial_{z_i}g^{(n)}(\mathbf{z},\mathbf{k})+ik_ig^{(n)}(\mathbf{z},\mathbf{k})\}=(n-1)\hat{G}(\mathbf{k})g^{(n)}(\mathbf{z},\mathbf{k}). \tag{80}$$

We have to take into account that Eq. (80) has to be solved for all n in the range from unity to infinity. Following the recipe derived in Sec. II, we find, if we write out the lowest orders of n in Eq. (80) explicitly, that

$$\begin{aligned}
&\mathbf{k}\cdot\mathbf{k}\sum_{i=1}^{3}\{\partial_{z_i}g^{(1)}(\mathbf{z},\mathbf{k})+ik_ig^{(1)}(\mathbf{z},\mathbf{k})\}=0,\\
&\mathbf{k}\cdot\mathbf{k}\sum_{i=1}^{3}\{\partial_{z_i}g^{(2)}(\mathbf{z},\mathbf{k})+ik_ig^{(2)}(\mathbf{z},\mathbf{k})\}=\hat{G}(\mathbf{k})g^{(2)}(\mathbf{z},\mathbf{k}),\\
&\mathbf{k}\cdot\mathbf{k}\sum_{i=1}^{3}\{\partial_{z_i}g^{(3)}(\mathbf{z},\mathbf{k})+ik_ig^{(3)}(\mathbf{z},\mathbf{k})\}=2\hat{G}(\mathbf{k})g^{(3)}(\mathbf{z},\mathbf{k}),
\end{aligned} \tag{81}$$

We proceed by defining a function $f^{(n)}(\mathbf{x},t)$. The functions $g^{(n)}(\mathbf{z},\mathbf{k})$ in Eq. (81) are defined to be related to functions $f^{(n)}(\mathbf{x},t)$ by the following relationship:

$$f^{(n)}(\mathbf{x},t)=\frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\tilde{f}^{(n)}(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{z}}d^3k;\quad g^{(n)}(\mathbf{z},\mathbf{k})=\tilde{f}^{(n)}(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{z}};\quad z_i=x_i-\frac{\omega(\mathbf{k})}{k_i}\,t. \tag{82}$$

If we carry out the inverse Fourier transform of Eq. (82), we find that the set of equations given in Eq. (81) is transformed into the following set of equations:

$$\begin{aligned}
&\sum_{i=1}^{3}\partial_{x_i}^3f^{(1)}(\mathbf{x},t)+\partial_tf^{(1)}(\mathbf{x},t)=0,\\
&\sum_{i=1}^{3}\partial_{x_i}^3\,f^{(2)}(\mathbf{x},t)+\partial_tf^{(2)}(\mathbf{x},t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\mathbf{x},t|\mathbf{x}',t')f^{(2)}(\mathbf{x}',t')d\mathbf{x}'dt',\\
&\sum_{i=1}^{3}\partial_{x_i}^3\,f^{(3)}(\mathbf{x},t)+\partial_tf^{(3)}(\mathbf{x},t)=2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\mathbf{x},t|\mathbf{x}',t')f^{(3)}(\mathbf{x}',t')d\mathbf{x}'dt',\\
&\qquad\vdots
\end{aligned} \tag{83}$$

The kernel $G(\mathbf{x},t|\mathbf{x}',t')$ in Eq. (83) is related to the function $\hat{G}(\mathbf{k})$ by means of the following Fourier transform:

$$G(\mathbf{x},t|\mathbf{x}',t)=\frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\hat{G}(\mathbf{k})e^{i\mathbf{k}\cdot[\mathbf{z}-\mathbf{z}']}d^3k. \tag{84}$$

We continue by defining the three-dimensional generalization of the ansatz (6) which relates the functions $f^{(n)}(\mathbf{x},t)$ to the integration kernel $G(\mathbf{x},t|\mathbf{x}',t')$:

$$6\left[\sum_{i=1}^{3}\partial_{x_i}\,f^{(n)}(\mathbf{x},t)\right]f^{(m)}(\mathbf{x},t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\mathbf{x},t'|\mathbf{x}',t')f^{(n+m)}(\mathbf{x}',t')d\mathbf{x}'dt'. \tag{85}$$

If we substitute Eq. (85) into Eq. (83), we obtain that the set of Eqs. (83) can be uniquely decomposed as

$$\sum_{i=1}^{3} \partial_{x_i}^3 f^{(1)}(\mathbf{x},t) + \partial_t f^{(1)}(\mathbf{x},t) = 0,$$

$$\sum_{i=1}^{3} \partial_{x_i}^3 f^{(2)}(\mathbf{x},t) + \partial_t f^{(2)}(\mathbf{x},t) = 6\left[\sum_{i=1}^{3} \partial_{x_i} f^{(1)}(\mathbf{x},t)\right] f^{(1)}(\mathbf{x},t),$$

$$\sum_{i=1}^{3} \partial_{x_i}^3 f^{(3)}(\mathbf{x},t) + \partial_t f^{(3)}(\mathbf{x},t) = 6\sum_{i=1}^{3} \left([\partial_{x_i} f^{(1)}(\mathbf{x},t)] f^{(2)}(\mathbf{x},t) + [\partial_{x_i} f^{(2)}(\mathbf{x},t)] f^{(1)}(\mathbf{xt})\right), \quad (86)$$

$$\sum_{i=1}^{3} \partial_{x_i}^3 f^{(4)}(\mathbf{x},t) + \partial_t f^{(4)}(\mathbf{x},t) = 6\sum_{i=1}^{3} \left([\partial_{x_i} f^{(1)}(\mathbf{x},t)] f^{(3)}(\mathbf{x},t) + [\partial_{x_i} f^{(2)}(x,t)] f^{(2)}(\mathbf{x},t) + [\partial_{x_i} f^{(3)}(\mathbf{x},t)] f^{(1)}(\mathbf{x},t)\right),$$

If we take the sums of both the left-hand sides and the right-hand sides of all the equations (86), we obtain, after some bookkeeping, the following result:

$$\sum_{n=1}^{\infty} \left[\sum_{i=1}^{3} \partial_{x_i}^3 f^{(n)}(\mathbf{x},t)\right] + \sum_{i=1}^{\infty} \partial_t f^{(n)}(\mathbf{x},t) = 6\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \left[\sum_{i=1}^{3} \partial_{x_i} f^{(n)}(\mathbf{x},t)\right] f^{(m)}(\mathbf{x},t). \quad (87)$$

Finally, if we define, for notational convenience,

$$u(\mathbf{x},t) = \sum_{n=1}^{\infty} f^{(n)}(\mathbf{x},t), \quad (88)$$

we find that Eq. (87) is equivalent with

$$\sum_{i=1}^{3} \partial_{x_i}^3 u(\mathbf{x},t) + \partial_t u(\mathbf{x},t) = 6\left[\sum_{i=1}^{3} \partial_{x_i} u(\mathbf{x},t)\right] u(\mathbf{x},t). \quad (89)$$

Equation (89) is a three-dimensional partial differential equation which can be linearized into Eq. (78) by applying a similar cascade of transformations and rescaling. Equation (89) is the three-dimensional generalization of the KdV equation. We can solve Eq. (89) by solving Eq. (78) and transforming the obtained solutions $g^{(n)}(\mathbf{x},\mathbf{k})$ into the functions $f^{(n)}(\mathbf{x},t)$. Before we can compute the solution, Eq. (78), we first have to find an adequate Green's function to solve the integration kernel $G(\mathbf{x},t|\mathbf{x}',t')$. We therefore first solve

$$\sum_{i=1}^{3} (\partial_{x_i} + ik_i)\mathcal{G}(\hat{\mathbf{z}},\mathbf{k}) = \delta(\hat{\mathbf{z}}), \quad \hat{\mathbf{z}} = \mathbf{z} - \mathbf{z}'. \quad (90)$$

The Fourier transform $\mathcal{H}(\mathbf{k}',\mathbf{k})$ of $\mathcal{G}(\hat{\mathbf{z}},\mathbf{k})$ is given by the following relationship:

$$\mathcal{H}(\mathbf{k}',\mathbf{k}) = \int_{-\infty}^{\infty} \mathcal{G}(\hat{\mathbf{z}},\mathbf{k}) e^{-i\mathbf{k}'\cdot\hat{\mathbf{z}}} d^3\hat{z}. \quad (91)$$

If Eq. (91) is substituted into Eq. (90), we obtain the following solution of $\mathcal{H}(\mathbf{k}',\mathbf{k})$:

$$\mathcal{H}(\mathbf{k}',\mathbf{k}) = \frac{1}{i\Sigma_{j=1}^{3}(p_j + k_j)}. \quad (92)$$

If we apply the inverse transform of Eq. (91) to Eq. (92), we find the following solution of the Green's function $\mathcal{G}(\hat{\mathbf{z}},\mathbf{k})$:

$$\mathcal{G}(\hat{\mathbf{z}},\mathbf{k})=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^3 i}\int_{-\infty}^{\infty}\frac{e^{i\mathbf{k}'\cdot\hat{\mathbf{z}}}}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\,d^3k'. \tag{93}$$

Once the Green's function is known, we can compute the general solution of the linear differential equation, Eq. (78):

$$g^{(n)}(\mathbf{z},\mathbf{k})=\lim_{\epsilon\to 0}\frac{(n-1)}{(2\pi)^3 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i\mathbf{k}'\cdot[\mathbf{z}-\mathbf{z}']}}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\,S(\mathbf{k})g^{(n)}(\mathbf{z}',\mathbf{k})d^3k'd^3z',\quad n\neq 1. \tag{94}$$

If we apply the Fourier transform given in Eqs. (82)–(94), we find that solutions of the functions $f^{(n)}(\mathbf{x},t)$ are given by

$$f^{(n)}(\mathbf{x},t)=\lim_{\epsilon\to 0}\frac{(n-1)}{(2\pi)^6 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(\mathbf{k}-\mathbf{k}')}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\,S(\mathbf{k})g^{(n)}(\mathbf{z},\mathbf{k}')d^3kd^3k',\quad n\neq 1. \tag{95}$$

In a similar manner as pointed out in Sec. III, it can be shown that if the integration kernel $G(\mathbf{x}-\mathbf{x}'|t-t')$ has the following structure:

$$G(\mathbf{x}-\mathbf{x}'|t-t')=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^6 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{[\mathbf{k}'\cdot\mathbf{k}']e^{i\mathbf{k}'\cdot[\mathbf{z}-\mathbf{z}']}}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\,S(\mathbf{k}')d^3kd^3k'. \tag{96}$$

A general form of the solutions $u(\mathbf{x},t)$ of the three-dimensional KdV equation is given by the following relationship:

$$u(\mathbf{x},t)=\sum_{n=1}^{\infty}f^{(n)}f^{(n)}(\mathbf{x},t)=Ae^{i\mathbf{k}\cdot\mathbf{x}}+\lim_{\epsilon\to 0}\sum_{n=2}^{\infty}\frac{(n-1)}{(2\pi)^6 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(\mathbf{k}-\mathbf{k}')}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\times S(\mathbf{k}')g^{(n)}(\mathbf{z},\mathbf{k}')d^3kd^3k'. \tag{97}$$

It can be verified by redoing the computations which are carried out in Sec. III, that if the ansatz (85), is satisfied, a recursion relation for the solutions $f^{(n)}(x,t)$ can be obtained. The solutions $f^{(n)}(x,t)$ which lead to a solution of the three-dimensional KdV equation are given by

$$f^{(1)}(\mathbf{x},t)=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^6 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i\mathbf{k}'\cdot\mathbf{z}}}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}\,S(\mathbf{k}')\left[\sum_{m=1}^{3}ik_m'\right]d^3kd^3k', \tag{98}$$

$$f^{(2)}(\mathbf{x},t)=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^9 i^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i(\mathbf{k}'+\mathbf{k}'')\cdot\mathbf{z}}S(\mathbf{k}')S(\mathbf{k}'')}{\Sigma_{l,m=1}^{3}(k_l+k_l'+i\epsilon)(k_m'+k_m''+\epsilon)}\times\left[i\sum_{m=1}^{3}(k_m'+k_m'')\right]d^3kd^3k'd^3k', \tag{99}$$

Once the functions $f^{(n)}(\mathbf{x},t)$ are computed, we construct the solution $u(\mathbf{x},t)$ of the three-dimensional KdV equation by summation of the functions $f^{(n)}(\mathbf{x},t)$:

$$u(\mathbf{x},t)=\lim_{\epsilon\to 0}\frac{1}{(2\pi)^6 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i\mathbf{k}'\cdot\mathbf{z}}}{\Sigma_{j=1}^{3}(k_j+k_j'+i\epsilon)}S(\mathbf{k}')\left[\sum_{m=1}^{3}ik_m'\right]d^3kd^3k'$$
$$+\lim_{\epsilon\to 0}\frac{1}{(2\pi)^9 i^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i(\mathbf{k}'+\mathbf{k}'')\cdot\mathbf{z}}S(\mathbf{k}')S(\mathbf{k}'')}{\Sigma_{l,m=1}^{3}(k_l+k_l'+i\epsilon)(k_m'+k_m''+i\epsilon)}$$
$$\times\left[i\sum_{m=1}^{3}(k_m'+k_m'')\right]d^3kd^3k'd^3k''+\cdots . \tag{100}$$

We can conclude that the approach followed in Sec. II for the one-dimensional KdV equation, can also be used to construct special solutions for the three-dimensional KdV equation. By generalizing Eq. (78) and by introducing more symmetry, we can, in principle, find special solutions of other higher-dimensional NPDEs. For proving the integrability, it would be necessary to prove that the set of solutions which are derived is sufficient for solving a reasonable Cauchy problem. It is also still unclear what the relationship between the procedure followed in this paper and the three-dimensional inverse scattering problem exactly is. In one dimension, this issue is clarified in Sec. IV, but in three dimensions the analysis is more complicated. It is not unthinkable that if this relationship is revealed, answers can be given with respect to the integrability of three-dimensional NPDEs.

VII. DISCUSSION

The following novel results are obtained in this paper. First, a new method is presented to obtain solutions of one-dimensional *S*-integrable differential equations. Whereas the IST solves *S*-integrable differential equations by using inversion techniques, the method presented in this paper solves these equations by using a direct method. The *S*-integrable nonlinear partial differential equation is linearized into an ordinary linear differential equation. There are several advantages of using a linearization method. The first advantage is that the method given is fairly general. In principle, the machinery applies to all NPDEs which have a nonlinear part consisting of a power series. Second, the method we have developed does not need Lax pairs for systematically investigating the integrability of NPDEs. The method can be generalized to higher-dimensional NPDEs. This is demonstrated in Sec. VI for the three-dimensional KdV equation. This result demonstrates that at least in one dimension, *S*-integrable NPDEs are also *C*-integrable.

As shown in Sec. II we can conclude that the linearized equation depends on two factors. The first factor is the dispersion relation which defines the space coordinate. The latter is invariant under Galileian transformations. The second factor is a generalization of the scattering data which is related by the ansatz used in this paper to the nonlinearity. By modifying the dispersion relation and the ansatz, Eq. (6), large classes of *S*-integrable NPDEs can be constructed and solved. It is remarked that the Galileian invariance of the space coordinate can be associated with a minimum symmetry property of the linearized differential equation. If the linearized equation, Eq. (17), is generalized, for instance, by making it invariant for other symmetries, more complex NPDEs can be constructed. It is shown in Sec. V that the nonlinear Schrödinger equation follows from a generalization of Eq. (17) which is invariant under the SU(2) generators. The mathematical reason for this invariance is related to the fact that the zero-curvature condition which is responsible for an infinite number of conservation laws, is invariant under gauge transformations, taking their value in a Lie group. As a result of this, the corresponding linearized problem is invariant under similar transformations in the corresponding Lie algebra.

ACKNOWLEDGMENTS

The author would like to thank Professor F. Calogero for raising his interest in the concept of *C*-integrability. Professor P. C. Sabatier has made valuable comments on the manuscript. The numerous discussions about the topic with Hans van den Heuvel were greatly appreciated. I would also like to thank H. M. M. ten Eikelder and Roel Snieder for constructively commenting on the manuscript.

[1] M. J. Ablowitz and P. A. Clarkson, Solitons, *Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[2] F. Calogero, ''Why are certain nonlinear PDE both widely applicable and integrable,'' in *What is Integrability*, edited by V. E. Zakharov (Springer-Verlag, Berlin, 1990), pp. 1–62.
[3] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescaling model PDEs and their integrability. I & II,'' Inverse Probl. **3**, 229–262 (1987); **4**, 11–33 (1988).
[4] F. Calogero and Ji Xiaoda, ''C-integrable nonlinear partial differentiation equations 1,'' J. Math. Phys. **32**, (1991).
[5] C. S. Gardner and G. K. Morikawa, ''Similarity in the asymptotic behavior of collision free hydrodynamic waves and water waves,'' Courant Institute of Mathematical Science Research Report No. NYO-9082, New York University, New York, 1960.
[6] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, ''Method for solving the Korteweg-de Vries equation,'' Phys. Rev. Lett. **19**, 1095–1097 (1967).
[7] A. V. Mikhailov, A. B. Shabat, and V. V. Sokolov, ''The symmetry approach to classification of integrable equations,'' in Ref. 2, pp. 117–184.
[8] H. J. S. Dorren, ''Singular and unstable solutions of the Korteweg-de Vries hierarchy,'' J. Math. Phys. **37**, 1858–1870 (1996).
[9] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer-Verlag, New York, 1989).
[10] H. J. S. Dorren and J. J. B. van den Heuvel, ''An exact analysis of the dynamics of solitons in randomly birefringent optical fibers'' (in preparation).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 7 JULY 1998

N-extended superelliptic integrable systems

Raúl Felipe
ICIMAF, calle E esq. a 15, Vedado, Havana, Cuba,
and Universidad de Antioquia, Medellín, Colombia

Fausto Ongay
CIMAT, Apdo. Postal 402, Guanajuato, Gto., 36000, Mexico

(Received 8 August 1997; accepted for publication 18 March 1998)

A supersymmetric extension for an arbitrary number of odd parameters of the stationary KdV equation is studied; we call this an N-extended super KdV. It is shown that it is related to superelliptic curves, and thereafter the complete ring of superdifferential operators commuting with the associated Schrödinger operator is constructed. © *1998 American Institute of Physics.* [S0022-2488(98)03706-2]

I. INTRODUCTION

The construction of commutative rings of differential operators is a classical problem of differential equations, dating back to the work of Burchnall and Chaundy[1] in the early part of the century. Their results were forgotten for almost 50 years, but rediscovered for the purposes of the theory of finite-zone potentials for the integration of soliton-type equations. These results were beautifully complemented by Krichever's theory,[2] relating the construction of solutions to the Korteweg–de Vries (KdV) hierarchy to the study of line bundles over elliptic curves.

On the other hand, although it is now recognized that supersymmetry may not fulfill all its expected promises,[3] the supersymmetrization of a given construct has nevertheless become a very important theoretical tool, both for physicists and mathematicians. For instance, on the physical side, there is an ongoing active search for supersymmetric extensions of the Toda lattice model,[4] while on the mathematical side, supergeometry is an important special (and tractable) case of noncommutative geometry.[5]

In particular, starting with the works of Manin–Radul,[6] Mulase,[7] and Rabin–Freund[8] among others, there is by now a considerable amount of work devoted to supersymmetric extensions of the KdV equation and associated hierarchy. Usually this proceeds by exploiting, either the property of the KdV of being a completely integrable system, and/or its surprising relation with the geometrical theory of elliptic curves (of course, both sides being united by the theory of Krichever), and making the necessary adaptations to the ''super'' setting. Up to now, however, most of the theory developed concentrates on the cases $N=1$ or $N=2$, obtained by adjoining to the standard theory 1 or 2 odd (fermionic) degrees of freedom.

In this work we adopt mainly the rather algebraic point of view of integrable systems to introduce a new supersymmetric extension to an arbitrary number of odd variables of the stationary KdV equation: such an approach,[9,10] although certainly less intuitive (and probably less appealing) than the geometrical approach, enables us to deal with the case of an arbitrary number of fermionic variables in a rather straightforward manner, and thereby extend to this case some results of Manin–Radul,[6] Rabin,[11] and Melzer.[12] In particular, we are able to construct a complete ring of superdifferential operators commuting with the Schrödinger operator on N odd parameters arising from a superelliptic curve.

The paper is organized as follows: In section II we discuss the algebraic setting of this extension, and obtain the stationary super KdV equation as a necessary condition for commutativity of the even generators of the ring; next, in section III a set of odd generators is obtained by use of the Akhiezer–Baker superfunction. Then, in section IV, we prove that we obtained indeed this way the full ring of commutative superoperators. Finally, there is a short section of the discussion of our results.

 © 1998 American Institute of Physics

II. *N*-EXTENDED SUPER KdV EQUATIONS

First we describe the setting for our constructions and fix some notational conventions. The material in the first part of this section is for the most part well known, and we refer the reader to Mulase[10] for a clear and detailed exposition.

Let $\mathbb{C}[[x]]$ denote the algebra of formal power series in the (complex) variable x, Λ a Grassmann algebra in the generators $\theta_1,\theta_2,\ldots,\theta_N$, and consider the tensor product $\mathscr{A}=\mathbb{C}[[x]]\otimes\Lambda$, whose elements are the superfunctions; hence, we will be working with a formal version of the supermanifold denoted $\mathbb{C}^{1,N}$.

We introduce the following notation: $\partial=\partial_x$, $\partial_\nu=\partial_{\theta_\nu}$, $\partial_{\mu\nu}=\partial_\mu\partial_\nu$, etc. Also if f is a superfunction, we denote $\partial f=f'$; $\partial_\nu f=f_\nu$, and so on.

The ground algebra for us is the algebra $\mathscr{B}$ of superdifferential operators generated by the odd derivatives ∂_ν, with coefficients in $\mathscr{A}$; it is an associative algebra under the composition of operators (denoted by $\circ$), although it is neither commutative nor supercommutative. Its elements will be denoted by the capital letters $U,V,\ldots$, so that a typical element looks like

$$U=u+u^\nu\partial_\nu+u^{\mu\nu}\partial_{\mu\nu}+\ldots\ .$$

(Here and in what follows, to alleviate the notations we will use the summation convention over the Greek indices, running from 1 to N; however, as other types of summations will occur, to avoid confusion we will write these out explicitly).

$\mathscr{B}$ is a differential algebra with the derivation ∂, acting on the coefficients of the elements $U\in\mathscr{B}$; we will write this action as $\partial U=U'$. Observe that $\partial U\neq\partial\circ U$; indeed, $\partial U=\partial\circ U-U\circ\partial=[\partial,U]$. Furthermore, the action of the odd superderivations also extends to $\mathscr{B}$, and we will write it as $\partial_\nu U=U_\nu$, etc. In what follows we restrict our discussion to even operators $U\in\mathscr{B}$, although of course some odd operators, such as U_ν, will also occur.

One readily proves the following commutation formulas:

$$[\partial^k,U]=\sum_{i=1}^{k}\binom{k}{i}U^{(i)}\circ\partial^{(k-i)}, \tag{1}$$

$$[\partial^k,U\circ\partial]=[\partial^k,U]\circ\partial;\quad [U,U\circ\partial]=-U'\circ U;\quad [\partial_\nu,U]=U_\nu. \tag{2}$$

Finally, we extend $\mathscr{B}$ by the formal ''antiderivation'' operator ∂^{-1}, to obtain the algebra $\mathscr{B}((\partial^{-1}))$ of super pseudo-differential operators, whose elements are formal expressions of the form $\Sigma_{i=-\infty}^{k}U_i\circ\partial^i$. The operator ∂^{-1} is a formal inverse of ∂, that commutes with the derivations ∂_ν, and satisfies the following commutation relation with the operator of multiplication by a superfunction f:

$$\partial^{-1}\circ f=\sum_{i=1}^{\infty}(-1)^i f^{(i)}\partial^{-(i+1)}.$$

The action of the superderivations also extends to the algebra of super pseudo-differential operators in the obvious way.

Let us now consider the super Schrödinger operator $L=L_U=\partial^2+U$, as the super Hamiltonian of a system.

If we look for operators that commute with such an operator L, then as in the nonsuper case, the even operators with this property are to be found among the fractional powers of L, $L^{k/2}$, which are in general only pseudo-differential. In particular, for reasons that are quite classic but will also be clarified in section III, we are interested in $L^{3/2}$, which using the above formulas is readily found to be

$$L^{3/2}=\partial^3+\tfrac{3}{2}\,U\circ\partial+\tfrac{3}{4}\,U'+O(\partial^{-1}),$$

where $O(\partial^{-1})$ means terms involving only negative powers of ∂. Still more specifically, we look for conditions on its differential operator part,

$$L_+^{3/2}=\partial^3+\tfrac{3}{2}\,U\circ\partial+\tfrac{3}{4}\,U',$$

that will ensure that it commutes with L. Using again the commutation formulas (1) and (2) one readily finds

$$[L,L_+^{3/2}]=-\tfrac{1}{4}\,U'''-\tfrac{3}{2}\,U\circ U'+\tfrac{3}{4}\,[U,U']=-\tfrac{1}{4}\,U'''-\tfrac{3}{4}\,(U\circ U'+U'\circ U), \tag{3}$$

which is an operator of order 0. Therefore, the condition for commutation becomes

$$U'''+6U\circ U'-3[U,U']=0, \tag{4}$$

which is a supersymmetric analog of the stationary KdV equation with N odd variables.

Remark 1: Although for actual computations formula (3) is more practical, the statement of the super KdV equation in equation (4) bears out an interesting point; namely, that the noncommutativity of the ground ring is reflected in the appearance of the commutator $[U,U']$*, which of course does not appear in the ordinary KdV. This may also be compared to the parallel found between classical electromagnetism and Yang–Mills theories.*

On the other hand, the general form of the super KdV equation we have just written suggests that the relationship with superelliptic curves is preserved in this case, a point that will be developed in the remainder of this paper.

Indeed, in the case $N=2$ one can easily work out the explicit equations arising from equation (4), and we get the following system of equations extending the stationary super KdV equations obtained by Manin–Radul[6] and Rabin:[13]

$$\begin{gathered}u'''+3(uu+u^{\nu}u_{\nu}+u^{12}u_{12})'=0,\\ (u^{12})'''+3(2(uu^{12}-u^1u^2)+u^{\nu}u_{\nu}^{12}+u^{12}(u_{\nu}^{\nu}+u_{12}^{12}))'=0,\end{gathered} \tag{5}$$

$$\begin{gathered}(u^1)'''+3(2uu^1+u^{\nu}u_{\nu}^1-u^{12}(u_2-u_{12}^1))'=0,\\ (u^2)'''+3(2uu^2+u^{\nu}u_{\nu}^2+u^{12}(u_1+u_{12}^2))'=0.\end{gathered} \tag{6}$$

Remark 2: This system is a direct generalization to the case of two odd variables of those of Rabin and Manin–Radul, both of which are obtained from even operators. There are, however, other possible extensions in the literature, and we now discuss the relationship of our equations to one of these.

Two rather general $N=2$ extensions of the KdV may be found in Laberge and Mathieu,[14] who proposed two types of extension, arising from operators of the type

$$\mathscr{L}=\partial^2+2\Phi D_1D_2-(D_2\Phi)D_1+(D_1\Phi)D_2$$

or

$$\mathscr{L}=\partial^2-2\Phi D_1D_2+(D_2\Phi)D_1-(D_1\Phi)D_2+(D_1D_2\Phi)-\Phi^2,$$

where $D_{\nu}=\theta_{\nu}\partial+\partial_{\nu}$; $\nu=1,2$, and Φ is an arbitrary superfunction.

Therefore, in their formulation $\mathscr{L}$ need not be of definite parity, and our system is related to the second type, which allows for an independent term. However, even if one considers only even superfunctions, say $\Phi=v+\theta_1\theta_2w$, our system does not in general coincide with theirs: In fact, in our notation, the operator of Laberge and Mathieu takes the form

$$\mathscr{L}=(1-2\,\theta_1\theta_2w)\,\partial^2-2w(\theta_1\partial_2-\theta_2\partial_1)\circ\partial+U,$$

where U is an operator of the same type as those used in this work. Now, while the first part can be easily transformed into ∂^2 via the change of coordinates $x\mapsto y-g(y)\,\theta_1\theta_2$, with $g=2/3u'$, the second term only changes by the determinant of the transformation of the odd parameters, and thus cannot in general disappear.

Remark 3: On the other hand, our approach could be modified to include even operators of this type by allowing extensions of the form

$$L_{U,V}=\partial^2+U+V\circ\partial.$$

However, the required calculations become much more involved: As an example, the third fractional power becomes in this case

$$(L_{U,V})_+^{3/2}=\partial^3+\tfrac{3}{2}\,U\circ\partial+\tfrac{3}{4}\,U'+(\tfrac{3}{4}\,V'+\tfrac{3}{8}\,V^2)\circ\partial-\tfrac{1}{4}\,(U\circ V+V\circ U)+\tfrac{1}{8}\,(V''+V^2);$$

so we will not pursue this line of study in this note.

We point out that in their work Laberge and Mathieu also considered extensions of the KdV that are bi-Hamiltonian, and their results imply that our equations do not posses this property in general.

Observe that the four equations in (5), (6) share the common property of having a part where no odd derivatives are involved; namely, the first two terms: For instance, for the first of the equations, this part is simply $u'''+3(uu)'=0$, which has exactly the form of the usual KdV (although recall that u is here in general a superfunction); indeed, modulo a change of variables, the reduced part of this equation is just the ordinary KdV.

This remark also gives the clue as to the construction of special solutions, with a nontrivial odd part, to these equations. Indeed, if we let $u=u(x-\theta_\nu\delta^\nu;\tau)$, where τ is an even parameter, and the δ^ν are odd parameters,[15] and set $u^\nu=\delta^\nu\,\dot{u}$, $u^{12}=\delta^1\delta^2\,\ddot{u}$, where a dot means the derivative with respect to the even parameter τ, then if we calculate the Taylor expansion of u we get

$$u(x-\theta_\nu\delta^\nu)=u(x)-\theta_\nu\delta^\nu u'(x)+\theta_1\theta_2\delta^1\delta^2u''(x);$$

then, since $u_\mu=\partial_\mu u=-\delta^\mu u'$, every term involving odd derivatives vanishes. A straightforward computation then shows that every solution of $u'''+3(uu)'=0$ solves the whole set of equations. Since the classical theory of elliptic functions tells us that the solutions to the reduced equation are given in terms of the elliptic function of Weierstrass, $\wp$, it is not surprising that solutions to this system are given by setting $u=-2\wp(x-\theta_\nu\delta^\nu;\tau)$.

For arbitrary N, the equations become rather cumbersome; nevertheless, the core of this argument remains true: for instance for $N=3$, assuming now $u=u(x-\theta_\mu\delta^\mu;\tau)$, and $u^\nu=\delta^\nu\dot{u}$, $u^{\mu\nu}=\delta^\mu\delta^\nu\ddot{u}$, $u^{123}=\delta^1\delta^2\delta^3\dddot{u}$, the super KdV [equation (4)] reduces to

$$u'''+3(uu)'=0;\quad u'''+6(u\dot{u})'=0,$$

$$\ddot{u}'''+6(u\ddot{u}+\dot{u}\dot{u})'=0;\quad \dddot{u}'''+6(u\dddot{u}+3\dot{u}\ddot{u})'=0,$$

and again, the first equation implies the remaining ones. Therefore, a solution is obtained by setting $u=-2\wp(x-\theta_\mu\delta^\mu;\tau)$, as before, and this assertion holds for general N.

III. THE AKHIEZER–BAKER FUNCTION AND SUPERELLIPTIC FUNCTIONS

In the classical case, the systematic construction of commuting operators may be achieved by use of the so-called Akhiezer–Baker functions. This construction also carries over to the super case.

The basic idea is to bring the operator L to a ''standard form'' which, for a second order operator such as Schrödinger's operator L, is ∂^2. This is possible because L is already in normal form (meaning no first order term appears), and is done by conjugating L with a so-called dressing operator Γ, defined by the condition

$$\Gamma\circ\partial\circ\Gamma^{-1}=L^{1/2},$$

from which immediately follows $\Gamma\circ\partial^k\circ\Gamma^{-1}=L^{k/2}$ for all k. It is also not difficult to show that

$$\Gamma=\left(1+\sum_{i\geqslant 1}V_i\circ\partial^{-i}\right).$$

In general, dressing operators are unique modulo monic operators (i.e., with leading coefficient 1) with constant coefficients.

This construction gives isospectral deformations of $L^{k/2}$, meaning that the spectra of the conjugated operators are the same; in particular, the eigenfunctions of each operator are bijectively mapped onto those of the other.

In particular, since the eigenfunctions of ∂ are simply exponentials, this can be explicitly achieved by the Akhiezer–Baker function, defined by a condition of the type $W=\Gamma e^{\xi}$, as follows.

Fix the parameters τ and δ^{ν}, and consider Schrödinger's operator $L=L_U$, as introduced in the previous section; then set $\xi=\kappa x+\theta_{\nu}\delta^{\nu}$, where κ is a complex parameter. Then, the Akhiezer–Baker function defined in the above way does indeed give the eigenfunctions of any of the even operators $L^{k/2}$ commuting with L, because of the relation

$$L^{k/2}W=\Gamma\circ\partial^{k}\circ\Gamma^{-1}\circ\Gamma e^{\xi}=\kappa^{k}\Gamma e^{\xi}=\kappa^{k}W.$$

Furthermore, it is well known in the cases $N=1$ and $N=2$ that Akhiezer–Baker functions are closely related to line bundles over superelliptic Riemann surfaces; this explains the occurrence within the theory of (super)elliptic functions. As we have seen, the condition for commutativity of L and $L_{+}^{3/2}$ shows this is also true in this N-extended case: roughly speaking, the parameters τ and δ^{ν} define a supertorus as a quotient of our base supermanifold, $\mathbb{C}^{1,N}$, by the transformations

$$x\mapsto x+1;\theta_{\nu}\mapsto\theta_{\nu},$$

$$x\mapsto x+\tau+\theta_{\nu}\delta^{\nu};\theta_{\nu}\mapsto\theta_{\nu}+\delta^{\nu};$$

the Akhiezer–Baker function then appears as the transition function of a line bundle from the chart about the point at infinity to the other chart required to cover the underlying elliptic curve (see, e.g., Rabin[13] for a very clear and detailed account of this for the case $N=1$); we now further discuss these facts.

First of all, recall the definition of the super Weierstrass function.[11,12] It is clear that for the case of N odd parameters it should be defined by

$$\mathscr{R}(x,\tau,\theta,\delta)=\wp\left(x,\tau+\frac{1}{2\pi i}(\theta_{\nu}\delta^{\nu})\right), \tag{7}$$

where $\wp$ denotes the classical Weierstrass elliptic function; again by computing its Taylor expansion one readily finds the explicit expression

$$\mathscr{R}=\wp+\frac{1}{2\pi i}\theta_{\nu}\delta^{\nu}\dot{\wp}+\frac{1}{2!}\left(\frac{1}{2\pi i}\theta_{\nu}\delta^{\nu}\right)^{2}\ddot{\wp}+\frac{1}{3!}\left(\frac{1}{2\pi i}\theta_{\nu}\delta^{\nu}\right)^{3}\dddot{\wp}+\frac{1}{4!}\left(\frac{1}{2\pi i}\theta_{\nu}\delta^{\nu}\right)^{4}\ddddot{\wp}\ldots, \tag{8}$$

where again a dot means derivative with respect to the modular parameter τ, and the terms run up to N-th order derivatives.

The superelliptic function $\mathscr{R}$ gives in fact the eigenvalue of L associated to W. To see this, set $W=e^{\xi}(1+\Sigma_{i\geqslant 1}a_i\kappa^{-i})$, and define $h(\kappa)=\mathscr{R}(1/\kappa)$; then one calculates on the one hand that $V_i(e^{\xi})=e^{\xi}a_i$, and $2V_1'+U=0$, and

$$\partial^{2}W=e^{\xi}(\kappa^{2}+a_1\kappa+a_2++2a_1'+O(\kappa^{-1})),$$

where $O(\kappa^{-1})$ means only negative powers of κ appear, and also

$$UW=U(e^{\xi})+O(\kappa^{-1})=-2V_1'(e^{\xi}+O(\kappa^{-1}));$$

and on the other,

$$hW=e^{\xi}(\kappa^{2}+a_1\kappa+a_2+O(\kappa^{-1})),$$

so we get

$$LW = hW + e^{\xi} O(\kappa^{-1}).$$

Hence, because the orders of the zeroes of both sections, LW and hW, are the same, and furthermore, they differ only by terms ''analytic at infinity,'' we can conclude that written in the chart where W describes the transition function of the bundle, h does indeed describe the eigenvalue of W.

(Also observe that $\mathscr{R}$ has a second order pole, and is associated to a second order differential operator; as in the classical theory, since $\partial e^{\xi} = \kappa e^{\xi}$, this fact clearly holds in general: Under the substitution $x \mapsto \kappa = x^{-1}$, a function having a single pole of order k will be associated to an operator also of order k.)

Now, just as in the purely bosonic, nongraded case, besides $\mathscr{R}$, its derivative $\mathscr{R}'$, which has a pole of order 3, is another generator of the even superelliptic functions; and this explains why one is interested in $L_+^{3/2}$. Furthermore, these are all the even generators needed; this is because every superelliptic function should have an elliptic function as its reduction (i.e., its body); since these are generated by $\wp$ and $\wp'$, by a suitable rational transformation, an even superelliptic function with a nontrivial body must differ from either $\mathscr{R}$ or $\mathscr{R}'$ only by higher order nilpotent terms; these can then be modified by coordinate changes to bring the superelliptic function to one of these standard forms.

For our purposes, an important feature of the superelliptic functions is that, as was the case for the coefficients of U;[16] they depend upon the odd parameters only through the combination $(1/2\pi i)(\theta_\nu \delta^\nu)$ [something we already used; cf. equation (7)], and in fact, the factor $1/2\pi i$, which is important for the geometric reasons we have described, may be ignored; roughly speaking, this amounts to work ''upstairs'' on the covering space $\mathbb{C}^{1,N}$. This is in fact essentially what we did in section II, and to simplify the writing, we will also assume this in what follows, and hence that superelliptic functions depend upon θ and δ through the combination $\theta_\nu \delta^\nu$.

Now, besides these even generators, the ring of superfunctions holomorphic off infinity must contain enough odd generators; these may be constructed from the odd superfunctions,

$$\sigma_\nu(x, \tau; \theta_\nu, \delta^\nu) = \delta^\nu (1 + \log \Theta(x, \tau))',$$

where Θ is an ordinary theta function, since the transformation: $\theta_\nu \mapsto \theta_\nu + \sigma_\nu$ is a valid change of the odd generators of the algebra $\mathscr{A}$ [this follows from the fact that $(\log \Theta(x,\tau))'$ only vanishes at the lattice points].

IV. THE FULL RING OF COMMUTATIVE SUPEROPERATORS

Taking all of the above into account, we are now in a position to explicitly construct the ring of differential operators commuting with Schrödinger's operator $L = L_U$. Since we already know the even generators, it remains essentially to determine the odd operators. From our discussion on superelliptic functions, it is clear that we should consider as candidates those associated with the odd superfunctions σ_ν. Let us now carry out the necessary calculations, fixing as above a super Weierstrass function determined by the modular parameters τ and δ, i.e., fixing a supertorus.

First of all, let us verify that indeed $\mathscr{R}'$ is associated to $L_+^{3/2}$. As before, define $g(\kappa) = \mathscr{R}'(\kappa^{-1})$, and let us compute gW and $L_+^{3/2} W$: Then one has

$$gW = e^{\xi}(\kappa^3 + a_1 \kappa^2 + a_2 \kappa + a_3 + O(\kappa^{-1})),$$

and, since one also has the relation $V_1'' + 2V_2' + U \circ V_1 = 0$, and therefore

$$\partial^3 W = e^{\xi}(\kappa^3 + a_1 \kappa^2 + a_2 \kappa + a_3 + 3(a_1' \kappa + a_1'' + a_2') + O(\kappa^{-1})),$$

$$(U \circ \partial) W = U(\partial e^{\xi}) + U(a_1 e^{\xi}) + e^{\xi} O(\kappa^{-1}) = e^{\xi}(-2a_1' \kappa - a_1'' - 2a_2' + O(\kappa^{-1})),$$

$$U' W = e^{\xi}(-2a_1'' + O(\kappa^{-1}));$$

by combining all these equations we get

$$L_+^{3/2} W = gW + O(\kappa^{-1}),$$

as desired.

Next, we define $h_\nu(\kappa)=\sigma_\nu(\kappa^{-1})$, and we need to show that these give the eigenvalues of the first order operators $D_\nu=\partial_\nu+\delta^\nu\partial$. But, since we can write $W=e^\xi g$, we get

$$(\partial_\nu+\delta^\nu\partial)W=\delta^\nu\left(1+\kappa+\frac{g'}{g}\right)W.$$

Now, since $g=1+\Sigma a_i\kappa^{-i}$, g'/g may be written as

$$\frac{g'}{g}=\frac{b_1}{\kappa}+\frac{b_2}{\kappa^2}+\frac{b_3}{\kappa^3}+\ldots,$$

with $b_1=a_1'$, $b_2=a_2-(a_1^2)'$, etc. Since also

$$(\log\Theta)'\left(\frac{1}{\kappa}\right)=\kappa+\frac{r_1}{\kappa}+\frac{r_2}{\kappa^2}+\ldots,$$

we can rewrite

$$1+\kappa+\frac{g'}{g}=1+(\log\Theta)'\left(\frac{1}{\kappa}\right)+O\left(\frac{1}{\kappa}\right),$$

and so

$$(\partial_\nu+\delta^\nu\partial)W=h_\nu W+e^\xi O\left(\frac{1}{\kappa}\right).$$

In this case it is equally easy to directly compute the conditions for D_ν to commute with L, and $L_+^{3/2}$; indeed, they are

$$[L,D_\nu]=[U,D_\nu]=-(U_\nu+\delta^\nu U')$$

and

$$[L_+^{3/2},D_\nu]=-\left(\frac{3}{2}\right)(U_\nu+\delta^\nu U')\circ\partial-\left(\frac{3}{4}\right)(U_\nu+\delta^\nu U')'.$$

Now, for our fixed supertorus, and from the fact that the potential U only depends upon $x-\theta_\nu\delta^\nu$, the vanishing of $U_\nu+\delta^\nu U'$ follows immediately.

To conclude, let us gather our results as a theorem.

Theorem 1: *For each N, a solution to the stationary super KdV equation, equation (4), is given in terms of elliptic functions by setting $U=u+u^\nu\partial_\nu+u^{\mu\nu}\partial_{\mu\nu}+\ldots$, where*

$$u=-2\wp(x-\theta_\mu\delta^\mu;\tau);\quad u^\mu=\delta^\mu u;\quad u^{\mu\nu}=\delta^\mu\delta^\nu\ddot{u},\ \textit{etc.}$$

Furthermore, a complete ring of even and odd superoperators, commuting with the associated Schrödinger operator $L=L_U=\partial^2+U$, is generated by

$$L;\quad L_+^{3/2};\quad D_\nu=\partial_\nu+\delta^\nu\partial,\quad \nu=1,\ldots,N.$$

Via the super Akhiezer–Baker function, $W=e^\xi(1+\Sigma_{i\geqslant 1}a_i\kappa^{-1})$, these are associated to the superelliptic functions $\mathscr{R},\mathscr{R}'$, σ_ν, where

$$\sigma_\nu=\delta^\nu(1+\log\Theta(x,\tau))'.$$

V. DISCUSSION

In this work we introduced, for arbitrary n, an N-extended supersymmetric extension of the theory of Burchnall and Chaundy in the elliptic case, obtaining the full $\mathbb{Z}_2$-graded ring of super-

differential operators commuting with the associated Schrödinger operator. The key step in the explicit construction of the ring is the introduction of a general bosonic Akhiezer–Baker function, which clearly displays the required generators as superelliptic functions; and, in spite of the generality of the setting, it was a pleasant outcome that the resulting equations and operators, and in particular the fermionic generators of the ring of operators, have a particularly simple form. This is witnessed by the example of the conditions for the commutativity of L and $L_+^{3/2}$ obtained for the case $N=2$, extending the original equations of Burchnall and Chaundy[1] ($N=0$), and also those of Manin–Radul,[6] and Mulase[9] and Rabin[11] ($N=1$).

Nevertheless, many aspects of these extensions remain to be studied. For instance, the original problem of Burchnall and Chaundy of writing algebraic expressions for a general pair of commuting operators is still open in the super case.

Another interesting point would be the analysis of the representation given by these operators. Also, it would certainly be interesting to carry out a detailed study of the N-extended superelliptic curves associated to the extension, as is done for the case $N=1$, in Rabin.[13] Furthermore, a Krichever functor[17] should be constructed for this case.

We hope to tackle these geometrical aspects in forthcoming works, but we would like to point out here that, on the other hand, by varying N, we have also embedded the KdV equation in a new ''hierarchy'' of equations, and we think it would be interesting to study the properties of this hierarchy and understand the possible geometrical consequences of this embedding. Furthermore (although this is just speculative) since we have been dealing here with an arbitrary number of odd coordinates, we find it tempting to think that there may be a link between this hierarchy, and the program set forth by Manin,[18] whose purpose is to obtain even geometry as ''a collective effect of odd geometry in an ∞-dimensional odd geometry.''

ACKNOWLEDGMENTS

This work was partially supported by CONACYT (Mexico), Grants No. 0329P-E9506 and Mexico-Cuba 1996 8.2, and COLCIENCIAS (Colombia), through Grant No. 1115-05-339-96, and the Consejo para el Desarrollo de la Investigación (CODI), of the University of Antioquia, Medellín, Colombia. The authors also wish to thank the referee for his (or her) accurate and valuable comments.

[1] J. L. Burchnall and T. W. Chaundy, Proc. London Math. Soc. **211**, 420 (1923); Proc. R. Soc. London, Ser. A **118**, 557 (1928).
[2] I. M. Krichever, Funct. Anal. Appl. **12**, 12 (1977).
[3] P. Labelle and P. Mathieu, J. Math. Phys. **32**, 923 (1991).
[4] J. M. Evans and J. O. Madsen, ''Integrability vs supersymmetry,'' CERN preprint, CERN-TH/96-224.
[5] Yu. I. Manin, *Gauge Field Theory and Complex Geometry* (Springer-Verlag, Berlin, 1988); see also B. A. Kupershmidt, *Elements of Superintegrable Systems* (Reidel, Dordrecht, 1987).
[6] Yu. I. Manin and A. O. Radul, Commun. Math. Phys. **98**, 65 (1985).
[7] M. Mulase, Invent. Math. **92**, 1 (1988).
[8] J. M. Rabin and P. G. Freund, Commun. Math. Phys. **114**, 131 (1988).
[9] M. Mulase, J. Diff. Geom. **19**, 403 (1984).
[10] M. Mulase, ''Geometric classification of $\mathbb{Z}_2$-commutative...,'' MSRI preprint 06929, 1991.
[11] J. M. Rabin, Commun. Math. Phys. **137**, 533 (1991).
[12] E. Melzer, J. Math. Phys. **29**, 1555 (1988).
[13] J. M. Rabin, J. Geom. Phys. **15**, 252 (1995).
[14] C. A. Laberge and P. Mathieu, Phys. Lett. B **215**, 718 (1988).
[15] More precisely, we tensor $\mathscr{B}$ with another copy of the Grassmann algebra $\mathscr{A}$, where the parameters live in, and we pull the relevant objects back to this new algebra.
[16] Remark, however, that the coefficients of U are not in general superelliptic functions.
[17] M. Mulase, Int. J. Math. **1**, 293 (1990).
[18] Yu. I. Manin, Lect. Notes Math. **1111**, 59 (1985); see also V. G. Pestov in ''Differential geometric methods in theoretical physics,'' Adv. Appl. Clifford Alg. **4**, 157 (1994).

Matrix formulation of Hamiltonian structures of constrained KP hierarchy

Wen-Jui Huang[a)]
Department of Physics, National Changhua University of Education, Changhua, Taiwan

Jiin-Chang Shaw[b)]
Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan

Ming-Hsien Tu[c)]
Department of Physics, National Tsing Hua University, Hsinchu, Taiwan

(Received 27 January 1998; accepted for publication 6 April 1998)

We give a matrix formulation of the Hamiltonian structures of constrained KP hierarchy. First, we derive from the matrix formulation the Hamiltonian structure of the one-constraint KP hierarchy, which was originally obtained by Oevel and Strampp. We then generalize the derivation to the multiconstraint case and show that the resulting bracket is actually the second Gelfand–Dickey bracket associated with the corresponding Lax operator. The matrix formulation of the Hamiltonian structure of the one-constraint KP hierarchy in the form introduced in the study of matrix model is also discussed. © *1998 American Institute of Physics.* [S0022-2488(98)01207-9]

I. INTRODUCTION

The Gelfand–Dickey (GD) hierarchy is defined by[1]

$$l_n=\partial^n+u_{n-1}\partial^{n-1}+\cdots+u_0, \tag{1.1}$$

which satisfies the hierarchy equations

$$\frac{\partial l_n}{\partial t_k}=[(l_n)_+^{k/n},l_n]. \tag{1.2}$$

Here $A_\pm$ denote the differential part and the integral part of the pseudodifferential operator A. The second Hamiltonian structure of (1.1) is described by the GD bracket which in operator form can be written as

$$\begin{aligned}\Theta_2^{\mathrm{GD}}\left(\frac{\delta H}{\delta l_n}\right)&\equiv\{l_n,H\}\equiv\{u_{n-1},H\}\partial^{n-1}+\{u_{n-2},H\}\partial^{n-2}+\cdots+\{u_0,H\}\\&=\left(l_n\frac{\delta H}{\delta l_n}\right)_+l_n-l_n\left(\frac{\delta H}{\delta l_n}l_n\right)_+,\end{aligned} \tag{1.3}$$

where

$$\frac{\delta H}{\delta l_n}\equiv\partial^{-1}\frac{\delta H}{\delta u_0}+\cdots+\partial^{-n}\frac{\delta H}{\delta u_{n-1}}. \tag{1.4}$$

Note that, from (1.2), the equation for u_1 is trivial; thus we can set $u_1=0$. However, imposing such a constraint leads to a modification of the GD bracket (1.3) due to the Dirac reduction. The modified bracket is found to be

[a)]Electronic mail address: wjhuang@phys.nthu.edu.tw
[b)]Electronic mail address: shaw@math.nctu.edu.tw
[c)]Electronic mail address: mhtu@phys.nthu.edu.tw

 © 1998 American Institute of Physics

$$\overline{\Theta}_2^{\mathrm{GD}}\left(\frac{\delta H}{\delta l_n}\right)=\left(l_n\frac{\delta H}{\delta l_n}\right)_+ l_n-l_n\left(\frac{\delta H}{\delta l_n}l_n\right)_++\frac{1}{n}\left[l_n,\int^x \mathrm{res}\left[l_n,\frac{\delta H}{\delta l_n}\right]\right] \tag{1.5}$$

where u_1 in l_n and in $\delta H/\delta l_n$ are both set to zero and $\mathrm{res}(\Sigma_i a_i\partial^i)\equiv a_{-1}$.

The above description of the GD hierarchy relied on the use of fractional-power pseudodifferential operators[2] associated with the scalar Lax operator l_n. However, one can put the formalism on a more general setting by considering the GD hierarchy as a reduction of a system of n first-order equations. Following Dickey,[3] one can substitute for the scalar Lax operator l_n a $n\times n$ first-order differential operator,

$$\mathscr{L}=I\partial+U,\quad \partial\equiv\partial/\partial x, \tag{1.6}$$

$$U=\begin{pmatrix} 0 & -1 & 0 & \cdots & \cdots \\ \cdots & \cdots & -1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 & -1 \\ u_0 & u_1 & \cdots & \cdots & u_{n-1} \end{pmatrix}. \tag{1.7}$$

This is the basis of the extension to the more general situation when the matrix U belongs to a semisimple Lie algebra, a task that was accomplished by Drinfeld and Sokolov.[4] Using matrix notation, the hierarchy equations (1.2) can be expressed as

$$\partial_t\mathscr{L}=[Q,\mathscr{L}], \tag{1.8}$$

where Q is another $n\times n$ matrix which cannot be arbitrarily chosen. We have to choose properly the matrix Q such that $[Q,\mathscr{L}]$ is consistent with the form of $\partial_t\mathscr{L}$. In fact, in view of (1.8), if the last column of Q is given then the rest of the elements of Q can be fixed. Following this strategy, it has been shown[3] that the Hamiltonian structure of the GD hierarchy can be extracted from (1.8) and turns out to be the GD bracket (1.3).

Recently, there has been much interest in the so-called constrained KP hierarchy[5–13] which has a pseudodifferential Lax operator of the form

$$L_{(n,m)}=\partial^n+u_{n-1}\partial^{n-1}+\cdots+u_0+\sum_{k=1}^{m}\phi_k\partial^{-1}\psi_k, \tag{1.9}$$

and satisfies the evolution equation

$$\frac{\partial L_{(n,m)}}{\partial t_k}=[(L_{(n,m)})_+^{k/n},L_{(n,m)}]. \tag{1.10}$$

From (1.10), it can be shown that each ϕ_i (each ψ_i) is an eigenfunction (adjoint eigenfunction) of the constrained KP hierarchy, i.e.,

$$\frac{\partial\phi_i}{\partial t_k}=((L_{(n,m)})_+^{k/n}\phi_i),\quad \frac{\partial\psi_i}{\partial t_k}=-((L^*_{(n,m)})_+^{k/n}\psi_i). \tag{1.11}$$

The bi-Hamiltonian structure of (1.10) has been constructed by Oevel and Strampp[8] (see also Ref. 9). The matrix formulation of the cKP hierarchy and its Hamiltonian structure have been discussed in Refs. 13 and 14. In contrast to the affine Lie algebraic approach,[13,14] in this paper we shall follow Dickey's approach[3] to give an elementary derivation of the Hamiltonian structure associated with the Lax operator $L_{(n,m)}$. In our approach, the Hamiltonian structures obtained by Oevel and Strampp[8] come out quite naturally. Since the first Hamiltonian structure can be obtained from the second Hamiltonian structure by replacing the Lax operator $L_{(n,m)}$ by $L_{(n,m)}+\lambda$, where λ is called the spectral parameter, we shall focus only on the second structure.

Our paper is organized as follows: In Sec. II, we follow the approach close to that of Ref. 3 to build up the matrix equation (1.8) associated with the Lax operator $L_{(n,1)}$ and then derive the Hamiltonian structure which was obtained by Oevel and Strampp. In Sec. III, we generalize this formulation to the Lax operator $L_{(n,m)}$ and obtain its Hamiltonian structure. We further show that this Hamiltonian structure is, in fact, the GD bracket (1.3) defined by the Lax operator $L_{(n,m)}$. In Sec. IV, we work out a few simple examples explicitly. Concluding remarks are presented in Sec. V.

II. MATRIX FORMULATION OF THE OEVEL AND STRAMPP HAMILTONIAN STRUCTURE

Now we consider the matrix formulation corresponding to the Lax operator for the one-constraint KP hierarchy,

$$L_{(n,1)}=\partial^n+u_{n-1}\partial^{n-1}+\cdots+u_0+\phi\partial^{-1}\psi. \tag{2.1}$$

A matrix representation can be easily found by expressing the constraint equation $L_{(n,1)}\varphi=0$ in a matrix form as $\mathscr{L}_{(n,1)}\Phi=0$, where $\mathscr{L}_{(n,1)}$ is a square matrix and Φ is a column matrix. An ansatz is given by

$$\mathscr{L}_{(n,1)}=I\partial+U, \quad \partial\equiv\partial/\partial x, \tag{2.2}$$

$$U=\begin{pmatrix} 0 & -\psi & 0 & 0 & \cdots & 0 \\ \cdots & 0 & -1 & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & -1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 & -1 \\ \phi & u_0 & u_1 & \cdots & \cdots & u_{n-1} \end{pmatrix}, \tag{2.3}$$

where U is an $(n+1)\times(n+1)$ matrix and the numeration of rows and columns will be $i,j=-1,0,1,2,\ldots,n-1$.

We introduce another $(n+1)\times(n+1)$ matrix Q and ask the commutator $[Q,\mathscr{L}_{(n,1)}]\equiv N$ to be consistent with the form of $\partial_t\mathscr{L}_{(n,1)}$. Let us now compute the matrix elements of the commutator: (1) $i=-1$, or $j=-1$,

$$N_{-1,-1}=-Q'_{-1,-1}+\phi Q_{-1,n-1}+\psi Q_{0,-1}, \tag{2.4}$$

$$N_{-1,i}=-Q'_{-1,i}-\epsilon_i Q_{-1,i-1}+\psi Q_{0,i}+u_i Q_{-1,n-1} \quad (\epsilon_0=\psi,\epsilon_1=\cdots=\epsilon_{n-2}=1), \tag{2.5}$$

$$N_{i,-1}=-Q'_{i,-1}+\phi Q_{i,n-1}+Q_{i+1,-1} \quad (i=0,1,2,\ldots,n-2), \tag{2.6}$$

$$N_{n-1,-1}=-Q'_{n-1,-1}+\phi Q_{n-1,n-1}-\phi Q_{-1,-1}-\sum_{j=0}^{n-1} u_j Q_{j,-1}; \tag{2.7}$$

(2) $i\neq-1$ and $j\neq-1$,

$$N_{i,0}=-Q'_{i,0}-\psi Q_{i,-1}+u_0 Q_{i,n-1}+Q_{i+1,0}, \tag{2.8}$$

$$N_{n-1,0}=-Q'_{n-1,0}-\psi Q_{n-1,-1}+u_0 Q_{n-1,n-1}-\phi Q_{-1,0}-\sum_{j=0}^{n-1} u_j Q_{j,0}, \tag{2.9}$$

$$N_{i,j}=-Q'_{i,j}-Q_{i,j-1}+u_j Q_{i,n-1}+Q_{i+1,j}, \tag{2.10}$$

$$N_{n-1,j}=-Q'_{n-1,j}-Q_{n-1,j-1}+u_j Q_{n-1,n-1}-\phi Q_{-1,j}-\sum_{k=0}^{n-1} u_k Q_{k,j}, \tag{2.11}$$

where $i=0,1,...,\ n-2$ and $j=1,2,...,n-2$.

In order to make N of the form $\partial_t \mathscr{L}_{(n,1)}$, we write $N=\delta\mathscr{L}_{(n,1)}$; i.e.,

$$N_{-1,0}=-\delta\psi, \tag{2.12}$$

$$N_{n-1,-1}=\delta\phi, \tag{2.13}$$

$$N_{n-1,j}=\delta u_j \quad (j=0,1,...,n-1), \tag{2.14}$$

and $N_{i,j}=0$ for other (i,j). In the matrix formulation the equation $\delta\mathscr{L}_{(n,1)}=[Q,\mathscr{L}_{(n,1)}]$ would serve as a definition of the Hamiltonian structure. In other words, $\delta\psi$, $\delta\phi$, and δu_j are to be identified as $\{\psi,H\}$, $\{\phi,H\}$, and $\{u_j,H\}$, respectively. Of course, certain matrix elements of Q must be identified as the components of the gradient of H. However, we shall do such an identification later.

We shall see that one can solve $Q_{i,j}$ in terms of $Q_{0,-1}$ and $Q_{i,n-1}$ $(i=-1,0,...,n-1)$. To this end we introduce

$$\hat{Q}_i\equiv\sum_{k=0}^{n-1} Q_{ik}\partial^k, \quad \hat{Q}^j\equiv\sum_{k=0}^{n-1}\partial^{-1-k}Q_{kj}, \tag{2.15}$$

$$\hat{N}_i\equiv\sum_{k=0}^{n-1} N_{ik}\partial^k; \tag{2.16}$$

then (2.8) and (2.10) are equivalent to

$$\hat{N}_i=-\partial\hat{Q}_i-\psi Q_{i,-1}+\hat{Q}_{i+1}+Q_{i,n-1}(L_{(n,1)})_+ \quad (i=0,1,...,n-2), \tag{2.17}$$

while (2.9) and (2.11) give

$$\hat{N}_{n-1}=-\partial\hat{Q}_{n-1}-\psi Q_{n-1,-1}-\phi\hat{Q}_{-1}+Q_{n-1,n-1}(L_{(n,1)})_+-\sum_{k=0}^{n-1} u_k\hat{Q}_k. \tag{2.18}$$

Since $\hat{N}_i=0$ $(i=0,1,...,n-2)$ we deduce a recursion relation for $\hat{Q}_i$ from (2.17),

$$\hat{Q}_{i+1}=\partial\hat{Q}_i-Q_{i,n-1}(L_{(n,1)})_+ +\psi Q_{i,-1} \quad (i=0,1,...,n-2). \tag{2.19}$$

Define

$$\hat{Q}_n\equiv-\sum_{k=0}^{n-1} u_k\hat{Q}_k-\hat{N}_{n-1}-\phi\hat{Q}_{-1}. \tag{2.20}$$

Then (2.19) also holds for $i=n-1$. Using the fact that $\hat{Q}_n$ is a pure differential operator we can solve (2.19) recursively to express $\hat{Q}_i$ in terms of $\hat{Q}^{n-1}$ and $\hat{Q}^{-1}$. The results are (see Appendix A for a proof)

$$\hat{Q}_0=(X(L_{(n,1)})_+)_+=(XL_{(n,1)})_+, \tag{2.21}$$

$$\hat{Q}_i=\partial^i(XL_{(n,1)})_+-(\partial^i X)_+(L_{(n,1)})_+ +(\partial^i\hat{Q}^{-1})_+\psi, \tag{2.22}$$

where we have defined $X\equiv\Sigma_{k=0}^{n-1}\partial^{-1-k}Q_{k,n-1}=\hat{Q}^{n-1}$.

Now from (2.20) we get

$$\hat{N}_{n-1} = -\partial^n (XL_{(n,1)})_+ + (\partial^n X)_+ (L_{(n,1)})_+ - (\partial^n \hat{Q}^{-1})_+ \psi - \phi \hat{Q}_{-1} - \sum_{k=0}^{n-1} u_k [\partial^k (XL_{(n,1)})_+ - (\partial^k X)_+ + (\partial^k \hat{Q}^{-1})_+ \psi]$$

$$= -(L_{(n,1)})_+ (XL_{(n,1)})_+ + (L_{(n,1)}X)_+ (L_{(n,1)})_+ - (L_{(n,1)}\hat{Q}^{-1})_+ \psi - \phi \hat{Q}_{-1}. \quad (2.23)$$

In view of $\delta u_j = N_{n-1,j}$ we have

$$\delta (L_{(n,1)})_+ = (L_{(n,1)}X)_+ (L_{(n,1)})_+ - (L_{(n,1)})_+ (XL_{(n,1)})_+ - (L_{(n,1)}\hat{Q}^{-1})_+ \psi - \phi \hat{Q}_{-1}. \quad (2.24)$$

It remains to characterize $\hat{Q}_{-1}$ and $\hat{Q}^{-1}$. Since $N_{i,-1} = 0$ $(i = -1,0,...,n-2)$, (2.4) and (2.6) give

$$Q_{-1,-1} = \int^x (\phi Q_{-1,n-1} + \psi Q_{0,-1}), \quad (2.25)$$

$$Q_{i+1,-1} = Q'_{i,-1} - \phi Q_{i,n-1} \quad (i = 0,1,...,n-2). \quad (2.26)$$

If we define

$$Q_{n,-1} \equiv -\phi Q_{-1,-1} - \sum_{k=0}^{n-1} u_k Q_{k,-1} - N_{n-1,-1}, \quad (2.27)$$

(2.7) becomes (2.26) with $i = n-1$. Combining (2.26) with (2.5) we find (see Appendix B)

$$\hat{Q}^{-1} = [(Q_{0,-1} - X\phi)\partial^{-1}]_{\geqslant -n}, \quad (2.28)$$

$$\hat{Q}_{-1} = (\hat{Q}_{-1})_+ = (\partial^{-1}\psi (XL_{(n,1)})_+ + \partial^{-1} Q_{-1,n-1} L_{(n,1)})_+. \quad (2.29)$$

Now putting, (2.28) and (2.29) into (2.24), we obtain

$$\delta (L_{(n,1)})_+ = (L_{(n,1)}X)_+ (L_{(n,1)})_+ - (L_{(n,1)})_+ (XL_{(n,1)})_+ - [L_{(n,1)}(Q_{0,-1} - X\phi)\partial^{-1}]_+ \psi - \phi(\partial^{-1}\psi (XL_{(n,1)})_+ + \partial^{-1} Q_{-1,n-1} L_{(n,1)})_+. \quad (2.30)$$

We observe that if we identify

$$X = \frac{\delta H}{\delta (L_{(n,1)})_+} = \partial^{-1} \frac{\delta H}{\delta u_0} + \cdots + \partial^{-n} \frac{\delta H}{\delta u_{n-1}},$$

$$Q_{-1,n-1} = \frac{\delta H}{\delta \phi}, \quad Q_{0,-1} = -\frac{\delta H}{\delta \psi}, \quad (2.31)$$

then (2.30) becomes

$$\delta (L_{(n,1)})_+ = (L_{(n,1)}X)_+ (L_{(n,1)})_+ - (L_{(n,1)})_+ (XL_{(n,1)})_+ + \left[\left((L_{(n,1)}X)_+ \phi + L_{(n,1)} \frac{\delta H}{\delta \psi} \right) \partial^{-1} \psi \right]_+ - \left[\phi \partial^{-1} \psi (XL_{(n,1)})_+ + \phi \partial^{-1} \frac{\delta H}{\delta \phi} L_{(n,1)} \right]_+$$

$$= [(L_{(n,1)}X)_+ L_{(n,1)} - L_{(n,1)}(XL_{(n,1)})_+]_+ + \left[L_{(n,1)} \frac{\delta H}{\delta \psi} \partial^{-1} \psi - \phi \partial^{-1} \frac{\delta H}{\delta \phi} L_{(n,1)} \right]_+, \quad (2.32)$$

which is one of the Hamiltonian flow derived by Oevel and Strampp.[8]

The identification (2.31), which is a generalization of the one used for the GD hierarchy (1.2),[1] is motivated by the following relation:

$$\delta H=\int\left(\frac{\delta H}{\delta u_0}\,\delta u_0+\cdots+\frac{\delta H}{\delta u_{n-1}}\,\delta u_{n-1}+\frac{\delta H}{\delta\phi}\,\delta\phi+\frac{\delta H}{\delta\psi}\,\delta\psi\right)=\mathrm{Tr}(Q\,\delta\mathscr{L}_{(n,1)}),\tag{2.33}$$

where $\mathrm{Tr}(A)=\int\mathrm{tr}(A)$ and $\mathrm{tr}(A)$ denotes the ordinary trace of a square matrix. The relation (2.33) is actually quite common in the matrix formulation of an integrable hierarchy.[4] It is related to the fact that the operation $\mathrm{Tr}(A)$ provides a natural scalar product between the cotangent space (to which Q belongs) and the tangent space (to which $\delta\mathscr{L}_{(n,1)}$ belongs) of the phase space manifold. In the case of GD hierarchy, the identification of this sort can be derived from the associated linear system.[15] However, we have not devised a similar proof for the present case. Here, the validity of (2.31) will be simply justified by the final result.

Now we turn to $\delta\psi$ and $\delta\phi$. First we combine (2.5), (2.25), and (2.31) to get

$$\begin{aligned}\delta\psi&=-N_{-1,0}=Q'_{-1,0}-\psi Q_{0,0}-u_0Q_{-1,n-1}+\psi Q_{-1,-1}\\&=-((L_{(n,1)})^*_+Q_{-1,n-1}+\hat{Q}_0^*\psi)_0+\psi\int^x(\phi Q_{-1,n-1}+\psi Q_{0,-1})\\&=-((L_{(n,1)})^*X^*\psi)_0-\left((L_{(n,1)})^*\,\frac{\delta H}{\delta\phi}\right)_0+\psi\int^x\left(\phi\,\frac{\delta H}{\delta\phi}-\psi\,\frac{\delta H}{\delta\psi}\right).\end{aligned}\tag{2.34}$$

Next, we note that (2.7) and (2.13) together imply

$$\delta\phi=N_{n-1,-1}=-Q'_{n-1,-1}+\phi Q_{n-1,n-1}-\phi Q_{-1,-1}-\sum_{k=0}^{n-1}u_kQ_{k,-1}.\tag{2.35}$$

Using the following identities (which are proved in Appendix C):

$$Q_{j,-1}=[\partial^j(Q_{0,-1}-X\phi)]_0\quad(j=0,1,\ldots,n-1),\tag{2.36}$$

$$Q'_{n-1,-1}=[\partial^n(Q_{0,-1}-X\phi)]_0+\phi Q_{n-1,n-1},\tag{2.37}$$

and the identification (2.31), we rewrite (2.35) as

$$\begin{aligned}\delta\phi&=-[\partial^n(Q_{0,-1}-X\phi)]_0-\sum_{k=0}^{n-1}u_j[\partial^j(Q_{0,-1}-X\phi)]_0-\phi Q_{-1,-1}\\&=\left[(L_{(n,1)})_+\left(X\phi+\frac{\delta H}{\delta\psi}\right)\right]_0-\phi\int^x\left(\phi\,\frac{\delta H}{\delta\phi}-\psi\,\frac{\delta H}{\delta\psi}\right)\\&=(L_{(n,1)}X\phi)_0+\left(L_{(n,1)}\,\frac{\delta H}{\delta\psi}\right)_0-\phi\int^x\left(\phi\,\frac{\delta H}{\delta\phi}-\psi\,\frac{\delta H}{\delta\psi}\right).\end{aligned}\tag{2.38}$$

Once again, (2.34) and (2.38) agree with the Hamiltonian flows reported in Ref. 8.

In conclusion, we have shown that the bracket of Oevel and Strampp[8] comes out naturally from the matrix formulation (which, of course, is basically the AKNS scheme).

III. GENERALIZATIONS

Having derived the Hamiltonian structure of the one-constraint KP hierarchy from a matrix formulation, we now come to the multiconstraint case. The Lax operator for this case is defined by

$$L_{(n,m)}=\partial^n+u_{n-1}\partial^{n-1}+\cdots+u_0+\sum_{j=1}^{m}\phi_j\partial^{-1}\psi_j,\tag{3.1}$$

which has matrix representation of the form

$$\mathscr{L}_{(n,m)}=I\partial+U,\tag{3.2}$$

$$U=\begin{pmatrix} 0 & \cdots & 0 & -\psi_m & \cdots & \cdots & \cdots & 0 \\ & \ddots & \ddots & \vdots & & & & \\ & & 0 & -\psi_1 & & & & \\ & & & 0 & -1 & & & \\ & & & & \ddots & -1 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & 0 & -1 \\ \phi_m & \cdots & \phi_1 & u_0 & u_1 & \cdots & \cdots & u_{n-1} \end{pmatrix}, \tag{3.3}$$

where U is a $(n+m)\times(n+m)$ matrix and the numeration of rows and columns will be $i,j=-m,-m+1,...,n-1$. Our matrix is connected by the Miura transformation, within the generalized Wilson–Drinfeld–Sokolov method, with the matrix used in Ref. 14. We also introduce another $(n+m)\times(n+m)$ matrix Q such that $[Q,\mathscr{L}_{(n,m)}]$ consistent with the form of $\delta\mathscr{L}_{(n,m)}$. We can compute the matrix elements $N_{ij}\equiv[Q,\mathscr{L}_{(n,m)}]_{ij}$ and make it of the form $\delta\mathscr{L}_{(n,m)}$. The result is the following: (1) $i<0$ or $j<0$,

$$\begin{aligned} N_{-\alpha,-\beta}&=-Q'_{-\alpha,-\beta}+\phi_\beta Q_{-\alpha,n-1}+\psi_\alpha Q_{0,-\beta}, \\ N_{-\alpha,0}&=-Q'_{-\alpha,0}+u_0 Q_{-\alpha,n-1}+\psi_\alpha Q_{0,0}-\sum_{k=1}^{m}\psi_k Q_{-\alpha,-k}, \\ N_{-\alpha,i}&=-Q'_{-\alpha,i}-Q_{-\alpha,i-1}+u_i Q_{-\alpha,n-1}+\psi_\alpha Q_{0,i} \quad (i=1,2,...,n-1), \\ N_{i,-\alpha}&=-Q'_{i,-\alpha}+\phi_\alpha Q_{i,n-1}+Q_{i+1,-\alpha} \quad (i=0,1,...,n-2), \\ N_{n-1,-\alpha}&=-Q'_{n-1,-\alpha}+\phi_\alpha Q_{n-1,n-1}-\sum_{k=1}^{m}\phi_k Q_{-k,-\alpha}-\sum_{k=0}^{n-1}u_k Q_{k,-\alpha}, \end{aligned} \tag{3.4}$$

where $\alpha,\beta=1,2,...,m$; (2) $i\geqslant 0$ and $j\geqslant 0$,

$$\begin{aligned} N_{i,0}&=-Q'_{i,0}+u_0 Q_{i,n-1}+Q_{i+1,0}-\sum_{k=1}^{m}\psi_k Q_{i,-k}, \quad (i=0,1,...,n-2), \\ N_{n-1,0}&=-Q'_{n-1,0}+u_0 Q_{n-1,n-1}-\sum_{k=1}^{m}\psi_k Q_{n-1,-k}-\sum_{k=1}^{m}\phi_k Q_{-k,0}-\sum_{k=0}^{n-1}u_k Q_{k,0}, \\ N_{i,j}&=-Q'_{i,j}-Q_{i,j-1}+u_j Q_{i,n-1}+Q_{i+1,j}, \quad -\begin{pmatrix} i=0,1,...,n-2 \\ j=1,2,...,n-1 \end{pmatrix} \\ N_{n-1,j}&=-Q'_{n-1,j}-Q_{n-1,j-1}+u_j Q_{n-1,n-1}-\sum_{k=1}^{m}\phi_k Q_{-k,j}-\sum_{k=0}^{n-1}Q_{k,j} \quad (j=1,2,...,n-1), \end{aligned} \tag{3.5}$$

with

$$\begin{aligned} &N_{-j,0}=-\delta\psi_j, \quad N_{n-1,-j}=\delta\phi_j, \quad j=1,2,...,m, \\ &N_{n-1,j}=\delta u_j, \quad j=0,1,...,n-1, \quad N_{i,j}=0, \quad \text{otherwise}. \end{aligned} \tag{3.6}$$

The definition of $\hat{Q}_i$, $\hat{Q}^i$, and $\hat{N}_i$ are still given by (2.15) and (2.16). Following the steps presented in the previous section and imposing the following identifications:

$$X \equiv \hat{Q}^{n-1} = \frac{\delta H}{\delta (L_{(n,m)})_+}, \quad Q_{-k,n-1} = \frac{\delta H}{\delta \phi_k},$$

$$Q_{0,-k} = -\frac{\delta H}{\delta \psi_k} \quad (k = 1,2,...,m), \tag{3.7}$$

we obtain, after performing a similar derivation the Hamiltonian structure associated with the Lax operator (3.1),

$$\delta (L_{(n,m)})_+ = [(L_{(n,m)} X)_+ L_{(n,m)} - L_{(n,m)} (X L_{(n,m)})_+]_+ + \sum_{k=1}^{m} \left[L_{(n,m)} \frac{\delta H}{\delta \psi_k} \partial^{-1} \psi_k - \phi_k \partial^{-1} \frac{\delta H}{\delta \phi_k} L_{(n,m)} \right]_+,$$

$$\delta \phi_j = (L_{(n,m)} X \phi_j)_0 + \left(L_{(n,m)} \frac{\delta H}{\delta \psi_j} \right)_0 - \sum_{k=1}^{m} \phi_k \int^x \left(\phi_j \frac{\delta H}{\delta \phi_k} - \psi_k \frac{\delta H}{\delta \psi_j} \right), \tag{3.8}$$

$$\delta \psi_j = -((L_{(n,m)})^* X^* \psi_j)_0 - \left((L_{(n,m)})^* \frac{\delta H}{\delta \phi_j} \right)_0 + \sum_{k=1}^{m} \psi_k \int^x \left(\phi_k \frac{\delta H}{\delta \phi_j} - \psi_j \frac{\delta H}{\delta \psi_k} \right).$$

We would like to remark that the Hamiltonian flow equations (3.8) are, in fact, coming from the second GD brackets defined by the Lax operator (3.1). To see this, let us denote $\delta H / \delta L_{(n,m)} = \delta H / \delta (L_{(n,m)})_+ + A$, where A is a differential operator. From the identity

$$\mathrm{tr}\left(\frac{\delta H}{\delta L_{(n,m)}} \delta L_{(n,m)} \right) = \mathrm{tr}\left(\frac{\delta H}{\delta (L_{(n,m)})_+} \delta (L_{(n,m)})_+ \right) + \sum_{j=1}^{m} \int \left(\frac{\delta H}{\delta \phi_j} \delta \phi_j + \frac{\delta H}{\delta \psi_j} \delta \psi_j \right), \tag{3.9}$$

we find that the differential operator A satisfies

$$(A \phi_j)_0 = \frac{\delta H}{\delta \psi_j}, \quad (A^* \psi_j)_0 = \frac{\delta H}{\delta \phi_j}. \tag{3.10}$$

Now, putting $L_{(n,m)}$ and $\delta H / \delta L_{(n,m)} = \delta H / \delta (L_{(n,m)})_+ + A$ into the second GD bracket (1.3) and using the relation (3.10), we obtain the Poisson brackets (3.8).

Note that when we set the next leading coefficient u_{n-1} to vanish we have to impose the constraint $\mathrm{res}[\delta H / \delta L_{(n,m)}, L_{(n,m)}] = 0$. Then the resulting bracket becomes the modified GD bracket (1.5) defined by $L_{(n,m)}$ with $u_{n-1} = 0$. From this modified bracket, it can be shown[16] that the Hamiltonian structures of the coupled AKNS hierarchy ($n = 1$, $m = 2$) and the coupled Yajima–Oikawa hierarchy[17] ($n = 2$, $m = 2$) come out as the special cases.

In the remaining part of this section, we would like to consider another reduction of the KP hierarchy, which has a Lax operator of the form

$$K = \partial^n + u_{n-1} \partial^{n-1} + \cdots + u_0 + a \frac{1}{\partial - s}. \tag{3.11}$$

Such a system was introduced in the study of the matrix model[18] and can be written in the form (2.1) if we equate with each other the integral parts of the two Lax operators:

$$\phi \partial^{-1} \psi = a \frac{1}{\partial - s}. \tag{3.12}$$

Then

$$a = \phi \psi, \quad s = -(\ln \psi)', \tag{3.13}$$

or, equivalently,

$$\phi = a e^{\int^x s}, \quad \psi = e^{-\int^x s}. \tag{3.14}$$

We can also formulate the Hamiltonian structure associated with K in matrix form by introducing

$$\mathscr{K} = \begin{pmatrix} \partial - s & -1 & 0 & 0 & \cdots & 0 \\ \cdots & \partial & -1 & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \partial & -1 \\ a & u_0 & u_1 & \cdots & \cdots & \partial + u_{n-1} \end{pmatrix}, \tag{3.15}$$

and choosing an $(n+1)\times(n+1)$ matrix R such that $[R,\mathscr{K}]$ consistent with $\delta\mathscr{K}$.

After equating the matrix elements of $[R,\mathscr{K}]$ with the matrix elements of $\delta\mathscr{K}$:

$$\delta\mathscr{K}_{-1,-1} = -\delta s, \quad \delta\mathscr{K}_{n-1,-1} = \delta a, \tag{3.16}$$

$$\delta\mathscr{K}_{n-1,j} = \delta u_j \quad (j = 0,1,...,n-1), \quad \delta\mathscr{K}_{i,j} = 0, \quad \text{otherwise}, \tag{3.17}$$

we can solve R_{ij} in terms of $R_{-1,-1}$ and $R_{i,n-1}$ from the constraint equation (3.17). Thus, the remaining task is to identify $R_{-1,-1}$ and $R_{i,n-1}$. However, the following naive identifications:

$$R_{-1,-1} = -\frac{\delta H}{\delta s}, \tag{3.18}$$

$$R_{-1,n-1} = \frac{\delta H}{\delta a}, \tag{3.19}$$

$$R_{i,n-1} = \frac{\delta H}{\delta u_i} \quad (i = 0,1,...,n-1), \tag{3.20}$$

do not give us the correct Hamiltonian structure. To get correct identifications we note that Eq. (3.13) can be expressed as a matrix equation:

$$\mathscr{K} = \Psi^{-1}\mathscr{L}_{(n,1)}\Psi, \tag{3.21}$$

where the matrix Ψ is defined by

$$\begin{pmatrix} \psi & \cdots & \cdots & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & \cdots & \cdots & 1 \end{pmatrix}. \tag{3.22}$$

From (3.21) and the Hamiltonian flow equation $\delta\mathscr{L}_{(n,1)} = [Q,\mathscr{L}_{(n,1)}]$, it is easy to show that $\delta\mathscr{K} = [-\Psi^{-1}\delta\Psi + \Psi^{-1}Q\Psi,\mathscr{K}]$. As a consequence, R is related to Q, determined in the previous section by

$$R = -\Psi^{-1}\delta\Psi + \Psi^{-1}Q\Psi. \tag{3.23}$$

From this relation we find that the correct identification of $R_{-1,-1}$ should be

$$R_{-1,-1} = -\frac{\delta H}{\delta s} + \partial_x^{-1}\delta s. \tag{3.24}$$

Substituting (3.19), (3.20), and (3.24) into the ''dynamical'' equations (3.16) involving δa, δs, and δL_+, we obtain the Hamiltonian structure,

$$\delta K_+ = (K_+X)_+K_+ - K_+(XK_+)_+ + \left[K_+\left(Xa + a\,\frac{\delta H}{\delta a} + \left(\frac{\delta H}{\delta s}\right)'\right)(\partial - s)^{-1}\right]_+ - a\left[(\partial - s)^{-1}\left(X + \frac{\delta H}{\delta a}\right)K_+\right]_+,$$

$$\delta a = \left[e^{-\int^x s}K_+\left(Xa + a\,\frac{\delta H}{\delta a} + \left(\frac{\delta H}{\delta s}\right)'\right)e^{\int^x s}\right]_0 - a\left[e^{\int^x s}K_+^*\left(X^* + \frac{\delta H}{\delta a}\right)e^{-\int^x s}\right]_0,$$

$$\delta s = \left(\frac{\delta H}{\delta s}\right)' + \left[e^{\int^x s}K_+^*\left(X^* + \frac{\delta H}{\delta a}\right)e^{-\int^x s}\right]_0', \tag{3.25}$$

where $X \equiv \delta H/\delta K_+$. One can show by explicit calculations that (3.25) is indeed the same as the second GD bracket defined by K.[15]

Rigorously to say, what we obtain here is not really a matrix formulation of the second Hamiltonian structure associated with K. This reason is that the identification of $R_{-1,-1}$, (3.24), contains the term $\partial_x^{-1}\delta s$ which can be determined only after δs is computed. In other words, R is not really the ''dual space'' of the matrix $\mathscr{K}$ (which contains all dynamical variables). The question of whether or not a matrix formulation for the second GD bracket defined by K remains open.

IV. EXAMPLES

In this section, we consider two explicit examples. The first is the cKP hierarchy associated with the Lax operator,

$$L_{(2,1)} = \partial^2 + u + \phi\partial^{-1}\psi. \tag{4.1}$$

It is easy to transform the corresponding eigenvalue equation $L_{(2,1)}\varphi = 0$ into a matrix form by taking

$$\mathscr{L}_{(2,1)} = \begin{pmatrix} \partial & -\psi & 0 \\ 0 & \partial & -1 \\ \phi & u & \partial \end{pmatrix}, \quad \Phi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{pmatrix}. \tag{4.2}$$

Then $\mathscr{L}_{(2,1)}\Phi = 0$ implies

$$\varphi_1' - \psi\varphi_2 = 0, \tag{4.3}$$

$$\varphi_2' - \varphi_3 = 0, \tag{4.4}$$

$$\phi\varphi_1 + u\varphi_2 + \varphi_3' = 0. \tag{4.5}$$

Substitutions of (4.3) and (4.4) into (4.5) gives

$$(\partial^2 + u + \phi\partial^{-1}\psi)\varphi_2 = 0, \tag{4.6}$$

which is precisely $L_{(2,1)}\varphi = 0$ once $\varphi = \varphi_2$ is imposed.

To consider the Hamiltonian structure associated with $\mathscr{L}_{(2,1)}$, let us consider the associated Lax equation,

$$\partial_t\mathscr{L}_{(2,1)} = [\mathscr{M}, \mathscr{L}_{(2,1)}]. \tag{4.7}$$

It is necessary to solve the form of $\mathscr{M}$ to make the above equation consistent.

Writing

$$\mathscr{M}=\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & -a-e \end{pmatrix}, \tag{4.8}$$

we have the following explicit expression for the commutator $[\mathscr{M},\mathscr{L}_{(2,1)}]$:

$$\mathscr{M}\mathscr{L}_{(2,1)}-\mathscr{L}_{(2,1)}\mathscr{M}=-\mathscr{M}' + \begin{pmatrix} c\phi+\psi d & -a\psi+cu+\psi e & -b+\psi f \\ f\phi+g & -d\psi+fu+h & -a-2e \\ -(2a+e)\phi-ud, & -g\psi-(a+2e)u-\phi b, & -h-c\phi-fu \end{pmatrix}. \tag{4.9}$$

Setting the matrix (4.9) equal to

$$\delta\mathscr{L}_{(2,1)}=\begin{pmatrix} 0 & -\delta\psi & 0 \\ 0 & 0 & 0 \\ \delta\phi & \delta u & 0 \end{pmatrix}, \tag{4.10}$$

we obtain five constraint equations,

$$\begin{gathered} -a'+c\phi+\psi d=0, \quad -c'-b+\psi f=0, \quad -d'+f\phi+g=0, \\ -e'-d\psi+fu+h=0, \quad -f'-2e-a=0. \end{gathered} \tag{4.11}$$

It is easy to see that a,b,e,g,h can be solved in terms of c,d,f and u,ϕ,ψ,

$$\begin{gathered} a=\partial^{-1}(c\phi+\psi d), \quad b=-c'+\psi f, \quad g=d'-f\phi, \\ e=-\tfrac{1}{2}(f'+\partial^{-1}(c\phi+\psi d)), \quad h=-\tfrac{1}{2}(f''+c\phi-d\psi)-fu. \end{gathered} \tag{4.12}$$

On the other hand, we have

$$\begin{aligned} \delta u&=-h'-g\psi-(a+2e)u-\phi b, \\ \delta\phi&=-g'-(2a+e)\phi-ud, \\ \delta\psi&=b'+(a-e)\psi-cu. \end{aligned} \tag{4.13}$$

Hence

$$\begin{aligned} \delta u&=(\tfrac{1}{2}\partial^3+\partial u+u\partial)f+\tfrac{1}{2}(\partial\phi+\phi\partial)c-\tfrac{1}{2}(\partial\psi+\psi\partial)d, \\ \delta\phi&=\tfrac{1}{2}(\partial\phi+\phi\partial)f-\tfrac{3}{2}(\phi\partial^{-1}\phi)c-(\partial^2+u+\tfrac{3}{2}\phi\partial^{-1}\psi)d, \\ \delta\psi&=\tfrac{1}{2}(\partial\psi+\psi\partial)f-(\partial^2+u-\tfrac{3}{2}\psi\partial^{-1}\phi)c+\tfrac{3}{2}(\psi\partial^{-1}\psi)d. \end{aligned} \tag{4.14}$$

If we identify

$$f=\frac{\delta H}{\delta u}, \quad c=\frac{\delta H}{\delta\phi}, \quad d=-\frac{\delta H}{\delta\psi}, \tag{4.15}$$

and regard δu, $\delta\phi$, and $\delta\psi$ as Hamiltonian flows, then we can read off the Poisson brackets,

$$\{u(x),u(y)\}=[\tfrac{1}{2}\partial_x^3+\partial_x u(x)+u(x)\partial_x]\delta(x-y),$$

$$\{u(x),\phi(y)\}=[\phi(x)\partial_x+\tfrac{1}{2}\phi'(x)]\delta(x-y),$$

$$\{u(x),\psi(y)\}=[\psi(x)\partial_x+\tfrac{1}{2}\psi'(x)]\delta(x-y),$$

$$\{\phi(x),\phi(y)\}=-\tfrac{3}{2}\phi(x)\epsilon(x-y)\phi(y),$$

$$\{\phi(x),\psi(y)\}=[\partial_x^2+u(x)]\delta(x-y)+\tfrac{3}{2}\phi(x)\epsilon(x-y)\psi(y),$$

$$\{\psi(x),\psi(y)\}=-\tfrac{3}{2}\psi(x)\epsilon(x-y)\psi(y), \tag{4.16}$$

where $\epsilon(x-y)\equiv\partial_x^{-1}\delta(x-y)$ is the antisymmetric step function. These are correct brackets,[8] which can be computed from the modified GD bracket (1.5) associated with $L=\partial^2+u+\phi\partial^{-1}\psi$.

The second example is to consider the Lax operator of the form

$$K=\partial^2+u+a\,\frac{1}{\partial-s}, \tag{4.17}$$

even though we would not get a genuine matrix formulation of the second GD bracket associated with K, as discussed in the previous section. However, we like to check explicitly for this simple case that our identifications (3.19), (3.20), and (3.24) indeed give us a correct Hamiltonian structure.

It is also easy to derive a matrix representation for K,

$$\mathscr{K}=\begin{pmatrix}\partial-s & -1 & 0\\ 0 & \partial & -1\\ a & u & \partial\end{pmatrix},\quad \Phi=\begin{pmatrix}\phi_1\\ \phi_2\\ \phi_3\end{pmatrix}. \tag{4.18}$$

Then, from $\mathscr{K}\Phi=0$ we have

$$(\partial-s)\phi_1-\phi_2=0,\quad \phi_2'-\phi_3=0,\quad a\phi_1+u\phi_2+\phi_3'=0, \tag{4.19}$$

which imply

$$[\partial^2+u\phi_2+a(\partial-s)^{-1}]\phi_2=0, \tag{4.20}$$

as desired. The associated Lax equation is given by

$$\partial_t\mathscr{K}=[\mathscr{N},\mathscr{K}]. \tag{4.21}$$

Writing

$$\mathscr{N}=\begin{pmatrix}n_1 & n_2 & n_3\\ n_4 & n_5 & n_6\\ n_7 & n_8 & n_9\end{pmatrix}, \tag{4.22}$$

then

$$[\mathscr{N},\mathscr{K}]=-\mathscr{N}'+\begin{pmatrix}an_3+n_4 & -n_1+un_3+sn_2+n_5 & -n_2+sn_3+n_6\\ -sn_4+an_6+n_7 & -n_4+un_6+n_8 & -n_5+n_9\\ -sn_7+an_9-an_1-un_4 & -n_7+un_9-an_2-un_5 & -n_8-an_3-un_6\end{pmatrix}, \tag{4.23}$$

which must be equal to

$$\delta\mathcal{K}=\begin{pmatrix} -\delta s & 0 & 0 \\ 0 & 0 & 0 \\ \delta a & \delta u & 0 \end{pmatrix}. \tag{4.24}$$

There are six constraint equations:

$$\begin{gathered} -n_2'-n_1+un_3+sn_2+n_5=0, \quad -n_3'-n_2+sn_3+n_6=0, \\ -n_4'-sn_4+an_6+n_7=0, \quad -n_5'-n_4+un_6+n_8=0, \\ -n_6'-n_5+n_9=0, \quad -n_9'-n_8-an_3-un_6=0, \end{gathered} \tag{4.25}$$

and three dynamical equations:

$$\begin{gathered} \delta s=n_1'-an_3-n_4, \\ \delta a=-n_7'-sn_7+an_9-an_1-un_4, \\ \delta u=-n_8'-n_7+un_9-an_2-un_5. \end{gathered} \tag{4.26}$$

We can solve n_2, n_4, n_5, n_7, n_8, and n_9 from (4.25) in terms of n_1, n_3, and n_6 as follows:

$$\begin{gathered} n_2=-(\partial-s)n_3+n_6, \\ n_4=(2\partial(\partial-s)^2+2\partial u-a)n_3-2\partial n_1-\partial(3\partial-2s)n_6, \\ n_5=-((\partial-s)^2+u)n_3+(\partial-s)n_6+n_1, \\ n_7=(2(\partial+s)\partial(\partial-s)^2+2(\partial+s)\partial u-(\partial+s)a)n_3-2(\partial+s)\partial n_1-((\partial+s)\partial(3\partial-2s)+a)n_6, \\ n_8=(\partial(\partial-s)^2+\partial u-a)n_3-(2\partial^2-\partial s+u)n_6-\partial n_1, \\ n_9=-((\partial-s)^2+u)n_3+(2\partial-s)n_6+n_1. \end{gathered} \tag{4.27}$$

Now, substituting (4.27) into the dynamical equations (4.26) and using the identifications

$$n_1=-\frac{\delta H}{\delta s}+\partial_x^{-1}\delta s, \quad n_3=\frac{\delta H}{\delta a}, \quad n_6=\frac{\delta H}{\delta u}, \tag{4.28}$$

we obtain the expected Poisson brackets,[15]

$$\begin{gathered} \{u(x),u(y)\}=[\tfrac{1}{2}\partial_x^3+2u(x)\partial_x+u'(x)]\delta(x-y), \\ \{u(x),a(y)\}=[3a(x)\partial_x+2a'(x)]\delta(x-y), \\ \{u(x),s(y)\}=[\tfrac{3}{2}\partial_x^2+s(x)\partial_x]\delta(x-y), \\ \{a(x),a(y)\}=[(4a(x)s(x)+2a'(x))\partial_x+a''(x)+2(a(x)s(x))']\delta(x-y), \\ \{a(x),s(y)\}=[(\partial_x+s(x))^2\partial_x+u(x)\partial_x]\delta(x-y), \\ \{s(x),s(y)\}=\tfrac{3}{2}\partial_x\delta(x-y). \end{gathered} \tag{4.29}$$

V. CONCLUDING REMARKS

We have studied the Hamiltonian structure of the cKP hierarchy in the matrix formulation. We showed that the procedure for the GD hierarchy can be gone through without difficulty for the

cKP hierarchy. We have not only reproduced the Oevel and Strampp result for the one-constraint case, but also generalized it to the multiconstraint case. We further showed that this Hamiltonian structure is nothing but the GD bracket. Hence, our result again confirms the fact that the GD bracket can be properly restricted to the Lax operator of the form (3.1). We have also discussed the matrix formulation of the Hamiltonian structure of the one-constraint KP hierarchy in the form given by (3.11). However, we did not obtain a genuine matrix formulation in the usual sense. Nevertheless, we found that one can still compute the Hamiltonian structure from our matrix formulation once a proper modification is made.

Finally, we want to remark that the constrained modified KP hierarchy[8] can be obtained from the Lax operator (2.1) via the gauge transformation

$$K=\phi^{-1}L_{(n,1)}\phi=\partial^n+v_{n-1}+\ldots+v_0+\partial^{-1}v_{-1}, \tag{5.1}$$

which satisfies the hierarchy equation

$$\frac{\partial K}{\partial t_k}=[K^n_{\geqslant 1},K]. \tag{5.2}$$

The bi-Hamiltonian structure associated with (5.2) has been obtained[8] from the lifted bracket of the cKP hierarchy by gauge transformation. It would be interesting to know whether or not a matrix formulation of this bi-Hamiltonian structure exists. Following the spirit of the present work we have worked out this structure in matrix formulation for $n=2$ and 3, but a general construction is yet to be given. We hope to report the results in this direction in the near future.

ACKNOWLEDGMENTS

The authors would like to thank Professor Aratyn for pointing out the gauge equivalence between our matrix representation (3.3) and the one used in Ref. 14. This work is supported by the National Science Council of Taiwan under Grants No. NSC-87-2112-M-018-008 (W.J.H) and No. NSC-87-2811-M-007-0025 (J.C.S and M.H.T). One of us (M.H.T) also wishes to thank the Center for Theoretical Sciences of National Science Council of Taiwan for partial support.

APPENDIX A: PROOFS FOR (2.21) AND (2.22)

From (2.19) we have

$$\hat{Q}_1=\partial\hat{Q}_0-Q_{0,n-1}(L_{(n,1)})_+ +\psi Q_{0,-1},$$

$$\begin{aligned}\hat{Q}_2&=\partial\hat{Q}_1-Q_{1,n-1}(L_{(n,1)})_+ +\psi Q_{1,-1}\\&=\partial^2\hat{Q}_0-\partial Q_{0,n-1}(L_{(n,1)})_+ -Q_{1,n-1}(L_{(n,1)})_+ +\partial\psi Q_{0,-1}+\psi Q_{1,-1},\end{aligned}$$

$$\ldots$$

$$\hat{Q}_i=\partial^i\left(\hat{Q}_0-\sum_{k=0}^{i-1}\partial^{-1-k}Q_{k,n-1}(L_{(n,1)})_+ +\sum_{k=0}^{i-1}\partial^{-1-k}Q_{k,-1}\psi\right) \tag{A1}$$

$$\ldots$$

$$\hat{Q}_n=\partial^n\left(\hat{Q}_0-\sum_{k=0}^{n-1}\partial^{-1-k}Q_{k,n-1}(L_{(n,1)})_+ +\sum_{k=0}^{n-1}\partial^{-1-k}Q_{k,-1}\psi\right). \tag{A2}$$

By the virtue of (2.20), $\hat{Q}_n$ must be a differential operator of order $\leqslant n-1$; it follows that

$$\left(\hat{Q}_0-\sum_{k=0}^{n-1}\partial^{-1-k}Q_{k,n-1}(L_{(n,1)})_+ +\sum_{k=0}^{n-1}\partial^{-1-k}Q_{k,-1}\psi\right)_+$$

$$=\left(\hat{Q}_0-\sum_{k=0}^{n-1}\partial^{-1-k}Q_{k,n-1}(L_{(n,1)})_+\right)_+=0. \tag{A3}$$

Define $X\equiv\Sigma_{k=0}^{n-1}\partial^{-1-k}Q_{k,n-1}=\hat{Q}^{n-1}$ then (A1) and (A3) are just (2.21) and (2.22), respectively.

APPENDIX B: PROOFS FOR (2.28) AND (2.29)

From (2.26) it follows that

$$(\partial^{-n}Q_{n,-1}+\cdots+\partial^{-1}Q_{1,-1})=(\hat{Q}^{-1})'-\hat{Q}^{n-1}\phi. \tag{B1}$$

Using $(\hat{Q}^{-1})'=[\partial,\hat{Q}^{-1}]$, we get

$$Q_{n,-1}=\partial^n(Q_{0,-1}-\hat{Q}^{-1}\partial-X\phi). \tag{B2}$$

Since $Q_{n,-1}$ is a scalar, it follows that

$$[\partial^n(Q_{0,-1}-\hat{Q}^{-1}\partial-X\phi)\partial^{-1}]_+=0. \tag{B3}$$

Moreover, since $\partial^n\hat{Q}^{-1}$ is a pure differential operator, we have

$$\partial^n\hat{Q}^{-1}=[\partial^n(Q_{0,-1}-X\phi)\partial^{-1}]_+, \tag{B4}$$

which leads to (2.28).

Next, from (2.5),

$$Q_{-1,i-1}=-Q'_{-1,i}+\psi Q_{0,i}+u_iQ_{-1,n-1}\quad(i=1,2,\ldots,n-1). \tag{B5}$$

Define $\tilde{Q}_{-1,-1}=-Q'_{-1,0}+\psi Q_{0,0}+u_0Q_{-1,n-1}$; then in view of (B5) we write

$$\tilde{Q}_{-1,-1}+Q_{-1,0}\partial+\cdots+Q_{-1,n-2}\partial^{n-1}+Q_{-1,n-1}\partial^n=-(\hat{Q}_{-1})'+\psi\hat{Q}_0+Q_{-1,n-1}(L_{(n,1)})_+, \tag{B6}$$

which implies

$$\tilde{Q}_{-1,-1}=-\partial\hat{Q}_{-1}+\psi\hat{Q}_0+Q_{-1,n-1}(L_{(n,1)})_+. \tag{B7}$$

Since a scalar is invariant under the adjoint operation; i.e., $f=f^*$, applying the adjoint operation to the both sides of (B7) gives

$$\tilde{Q}_{-1,-1}=(\hat{Q}^*_{-1}\partial+(L_{(n,1)})^*_+Q_{-1,n-1}+\hat{Q}^*_0\psi)_0=((L_{(n,1)})^*_+Q_{-1,n-1}+\hat{Q}^*_0\psi)_0. \tag{B8}$$

Now we treat $\tilde{Q}_{-1,-1}$ a little differently. An equivalent form of (B7) is

$$\partial^{-1}\tilde{Q}_{-1,-1}=-\hat{Q}_{-1}+\partial^{-1}(\psi\hat{Q}_0+Q_{-1,n-1}(L_{(n,1)})_+), \tag{B9}$$

which leads to (2.29), since $\hat{Q}_{-1}$ is a pure differential operator.

APPENDIX C: PROOFS FOR (2.36) AND (2.37)

Note that (2.26) gives

$$Q_{1,-1}=Q'_{0,-1}-\phi Q_{0,n-1},$$

$$Q_{2,-1}=Q''_{0,-1}-\phi Q_{1,n-1}-(\phi Q_{0,n-1})',$$

$$\cdots$$

$$Q_{j,-1}=Q^{(j)}_{0,-1}-\phi Q_{j-1,n-1}-(\phi Q_{j-2,n-1})'\cdots-(\phi Q_{0,n-1})^{(j-1)}, \tag{C1}$$

$$\cdots$$

$$Q_{n-1,-1}=Q^{(n-1)}_{0,-1}-\phi Q_{n-2,n-1}-(\phi Q_{n-3,n-1})'\cdots-(\phi Q_{0,n-1})^{(n-2)}.$$

In particular,

$$Q'_{n-1,-1}=Q^{(n)}_{0,-1}-\phi Q_{n-1,n-1}-(\phi Q_{n-2,n-1})'-\cdots-(\phi Q_{0,n-1})^{(n-1)}+\phi Q_{n-1,n-1}. \tag{C2}$$

The operator forms of (C1) and (C2) are (2.36) and (2.37), respectively.

[1] L. A. Dickey, *Soliton Equations and Hamiltonian Systems* (World Scientific, Singapore, 1991).
[2] I. M. Gelfand and L. A. Dickey, Funkt. Anal. Pril. **10**, 13 (1976).
[3] L. A. Dickey, Commun. Math. Phys. **82**, 345 (1981).
[4] V. G. Drinfeld and V. V. Sokolov, J. Sov. Math. **30**, 1975 (1984).
[5] Y. Cheng, J. Math. Phys. **33**, 3774 (1992).
[6] B. Xu and Y. Li, J. Phys. A **25**, 2957 (1992).
[7] J. Sidorenko and W. Strampp, Commun. Math. Phys. **157**, 1 (1993).
[8] W. Oevel and W. Strampp, Commun. Math. Phys. **157**, 51 (1993).
[9] Y. Cheng, Commun. Math. Phys. **171**, 661 (1995).
[10] F. Yu, Lett. Math. Phys. **29**, 175 (1993).
[11] L. Bonora and C. S. Xiong, J. Math. Phys. **35**, 5781 (1994).
[12] L. A. Dickey, Lett. Math. Phys. **34**, 379 (1995); *ibid.* **35**, 229 (1995).
[13] H. Aratyn, J. F. Gomes, and A. H. Zimerman, J. Math. Phys. **36**, 3419 (1995).
[14] H. Aratyn, L. A. Ferreira, J. F. Gomes, and A. H. Zimerman, J. Math. Phys. **38**, 1559 (1997).
[15] L. Bonora, Q. P. Liu, and C. S. Xiong, Commun. Math. Phys. **175**, 177 (1996).
[16] Q. P. Liu, J. Math. Phys. **37**, 2307 (1996).
[17] N. Yajima and M. Oikawa, Prog. Theor. Phys. **56**, 1719 (1976).
[18] L. Bonora and C. S. Xiong, Phys. Lett. B **285**, 191 (1992); *ibid.* **317**, 329 (1993).

Second harmonic generation: Hamiltonian structures and particular solutions

K. R. Khusnutdinova
Department of Mathematics, Bashkir State University, Frunze Str. 32, 450074 Ufa, Russia

H. Steudel[a)]
Institut für Physik der Humboldt—Universität, Arbeitsgruppe ''Nichtklassische Strahlung,''
Rudower Chaussee 5, D-12484 Berlin, Germany

(Received 3 February 1998; accepted for publication 24 March 1998)

For the equations describing second harmonic generation (SHG), we find a linear system different from one found by Kaup (but connected with it by a gauge transformation), which enables us to consider the SHG equations as related to the first negative flow in the coupled KdV hierarchy and to find therefore the Hamiltonian structures and the recursion operator. With the help of the recursion operator we write down the simplest commuting flows and consider their stationary points in order to find particular solutions of the SHG equations. We also investigate some particular invariant solutions of the SHG equations that can be of physical interest, and find their connections with invariant solutions of the Tzitzéica equation. The latter are either connected with Painlevé III or solvable in elliptic functions. © *1998 American Institute of Physics.* [S0022-2488(98)00507-6]

I. INTRODUCTION

The generation of the second harmonic wave from the red light of a ruby laser in a crystal of quartz in 1961[1] marked the starting point of nonlinear optics. The theory for the one-dimensional stationary process has soon been developed.[2,3] For short pulses, when the group-velocity mismatch between both frequency components becomes important, the process of second harmonic generation (SHG) is described by the equations (see, e.g., Refs. 4 and 5)

$$q_{1\chi}=-2q_2q_1^*, \quad q_{2\tau}=q_1^2, \tag{1}$$

where differentiation with respect to the independent variables χ and τ is indicated by subscripts; the asterisk denotes complex conjugation. Here q_1 and q_2 are slowly varying complex electric field amplitudes of both waves with carrier frequencies ω_1 and $\omega_2=2\omega_1$, respectively. It is assumed that the respective group velocities v_1 and v_2 do not coincide, but that the dispersion within either of the two waves is negligible. The characteristic variables χ and τ are connected with the laboratory space and time coordinates x, t as follows:

$$\chi=\nu\left(-t+\frac{x}{v_2}\right), \quad \tau=\nu\left(t-\frac{x}{v_1}\right),$$

where $\nu=(1/v_2-1/v_1)^{-1}$ is the group-velocity mismatch. (Only characteristic variables will be used below.)

When restricted to real amplitudes q_1, q_2 the SHG equations (1) reduce to the Liouville equation, as was noted in Ref. 6. For that ''C-integrable'' case, particular analytical solutions were studied in Ref. 7, and it was found in Ref. 5 that the *restricted Cauchy problem* (with arbitrary q_1 and $q_2\equiv 0$ given at $x=0$) leads to a Schrödinger or Hill's equation.

The general SHG equations (1) are shown to be ''*S*-integrable'' by Kaup.[8] The application of the inverse scattering method to this problem, however, meets some peculiarities and is not yet fully developed.

[a)]Electronic mail: steudel.photon.fta-berlin.de

 © 1998 American Institute of Physics

The group analysis of the SHG equations was given by Zhiber in Ref. 9, where it was found that the Lie–Bäcklund algebra of these equations consists of two parts: the nonlinear Schrödinger type and the Liouville type. The auto-Bäcklund transformation for the SHG equations was investigated in Ref. 10.

In the present paper we reveal the relationship of the SHG equations to the coupled KdV hierarchy, studied in Refs. 11–16 (see also the references therein) and discuss the Hamiltonian formalism. We also construct some exact solutions of the SHG equations and find connections of the self-similar and the traveling wave solutions with the corresponding solutions of the Tzitzéica equation.[17] The latter are either connected with Painlevé III[18] and studied by the method of isomonodromic deformations[19] or solvable in elliptic functions.[20]

II. LINEAR SYSTEM

Elimination of q_2 from Eqs. (1) implies the equation

$$q_{1\chi\tau}q_1^* - q_{1\tau}^* q_{1\chi} = -2(q_1 q_1^*)^2. \tag{2}$$

From (2) we obtain two scalar equations:

$$(\ln Q)_{\chi\tau} - \phi_\chi \phi_\tau = -Q, \quad (Q\phi_\tau)_\chi = 0, \tag{3}$$

where $q_1 = (\sqrt{Q}/2)\exp[i(\phi/2)]$, and we assume that $Q>0$.

The variable q_2 can be recovered as follows:

$$q_2 = -\tfrac{1}{4}((\ln Q)_\chi + i\phi_\chi)\exp(i\phi).$$

Equations (3) admit the following Lie point symmetries:

$$(1)\ \tilde{\tau} = \tau + a, \quad (2)\ \tilde{\chi} = \chi + a, \quad (3)\ \phi = \phi + a,$$

$$(4)\ \tilde{\tau} = e^a \tau, \quad \tilde{\chi} = e^{-a}\chi, \quad (5)\ \tilde{\tau} = f(\tau), \quad \tilde{Q} = \frac{Q}{f'(\tau)}. \tag{4}$$

Here $a=\text{const}$ (group parameter) and $f(\tau)$ is an arbitrary function ($f' \neq 0$). Applying the standard methods of group analysis of differential equations,[21,22] one can prove that all point symmetries are given by (4).

There exists also an obvious discrete symmetry: $\tau \to -\tau$, $\chi \to -\chi$.

Introducing the variables

$$s = Q^{-1}, \quad u = \phi_\tau, \quad v = (\ln s)_\chi, \quad w = \phi_\chi,$$

we may rewrite system (3) as follows:

$$s_\chi = sv, \quad u_\chi = uv = w_\tau, \quad v_\tau = s^{-1} - uw. \tag{5}$$

Specifying in (5) to $u=0$ one arrives at the Liouville equation $(\ln s)_{\chi\tau} = s^{-1}$,[6] which allowed us recently to investigate the initial value problem for an amplitude-modulated initial pulse.[5]

As a consequence of (5) we get

$$\left(\ln \frac{s}{u}\right)_\chi = 0,$$

so that $s = g(\tau)u$. Since in variables s, u, v, w the last symmetry in (4) takes the form

$$\tilde{\tau} = f(\tau), \quad \tilde{s} = sf'(\tau), \quad \tilde{u} = \frac{u}{f'(\tau)},$$

we can reduce $g(\tau)$ to 1 if $g(\tau)>0$ (i.e., $\phi_\tau > 0$). In this case $s=u$, and Eqs. (5) reduce to

$$u_\chi = uv = w_\tau, \quad v_\tau = u^{-1} - uw. \tag{6}$$

To get the case $s=-u$ [$g(\tau)<0$, i.e., $\phi_\tau<0$] we just have to change in (6), $u\to -u$, $w\to -w$.

Applying the Wahlquist–Estabrook method to system (5), we find out that these equations are the integrability conditions for the following linear system:

$$\Psi_\chi = U\Psi, \quad \Psi_\tau = V\Psi,$$

$$U=\begin{pmatrix} \frac{v}{2} & -\lambda \\ \frac{\alpha}{\lambda}(\alpha - w) & -\frac{v}{2} \end{pmatrix}, \quad V=\begin{pmatrix} 0 & \frac{\lambda}{2\alpha}u \\ \frac{1}{2\lambda}(s^{-1}-\alpha u) & 0 \end{pmatrix}, \tag{7}$$

where α and λ are parameters.

Let us take $\alpha=\lambda$. The linear system (7) transforms to

$$\Psi_\chi = \begin{pmatrix} \frac{v}{2} & -\lambda \\ \lambda - w & -\frac{v}{2} \end{pmatrix}\Psi, \quad \Psi_\tau = \begin{pmatrix} 0 & \frac{u}{2} \\ \frac{1}{2}(\lambda^{-1}s^{-1}-u) & 0 \end{pmatrix}\Psi. \tag{8}$$

It is remarkable that for the first component of the Ψ function we can derive from (8) the following equations:

$$-\Psi^1_{\chi\chi} + \left(\frac{v_\chi}{2} + \frac{v^2}{4} + \lambda w\right)\Psi^1 = \lambda^2\Psi^1, \quad \Psi^1_\tau = -\frac{1}{2\lambda}u\Psi^1_\chi + \frac{1}{4\lambda}uv\Psi^1. \tag{9}$$

The spectral problem (9) in the variables $v^0 = v_\chi/2 + v^2/4$, $v^1 = w$ turns out to be of the type considered in Refs. 11–16. As far as the τ derivative of the Ψ function is proportional to λ^{-1}, our equations (5) [and (6)] correspond to the first negative flow of the coupled KdV hierarchy (in variables v^0, v^1).

Let us note that the linear system (7) is connected with Kaup's linear system,[8]

$$\Psi_\chi = U_0\Psi, \quad \Psi_\tau = V_0\Psi,$$

$$U_0=\begin{pmatrix} i\alpha & 2q_2 \\ 2q_2^* & -i\alpha \end{pmatrix}, \quad V_0=\frac{i}{\alpha}\begin{pmatrix} q_1 q_1^* & q_1^2 \\ -q_1^{*2} & -q_1 q_1^* \end{pmatrix},$$

by the gauge transformation

$$U = C^{-1}(U_0 C - C_\chi), \quad V = C^{-1}(U_0 C - C_\tau),$$

where

$$C=\begin{pmatrix} ie^{i(\phi/2-\pi/4)} & -\frac{\lambda}{\alpha}e^{i(\phi/2-\pi/4)} \\ e^{-i(\phi/2-\pi/4)} & -\frac{i\lambda}{\alpha}e^{-i(\phi/2-\pi/4)} \end{pmatrix}.$$

III. HAMILTONIAN STRUCTURES AND RECURSION OPERATOR

A system of evolution equations admits Hamiltonian structure if it can be written in the form

$$u^i_\tau = \theta^{ij}\frac{\delta H}{\delta u^j},$$

where θ^{ij} is the Hamiltonian operator, i.e., a matrix operator, such that the associated Poisson bracket $\{F,G\}=\int(\delta F/\delta u^i)\,\theta^{ij}(\delta G/\delta u^j)d\chi$ is skew symmetric and satisfies the Jacobi identity. Here δ denotes the variational derivative. The Hamiltonian H is usually a function of the u^i and their χ derivatives, but nonlocal Hamiltonians may also make sense.

In order to discuss the Hamiltonian structures, let us rewrite system (6) in the form

$$p_{\chi\tau}=e^{-p}-e^{p}w, \quad w_\tau=e^{p}p_\chi, \tag{10}$$

where $u=e^p$.

One can easily verify that Eqs. (10) can be written in Hamiltonian form,

$$\begin{pmatrix} p \\ w \end{pmatrix}_\tau = A\begin{pmatrix} \dfrac{\delta H}{\delta p} \\ \dfrac{\delta H}{\delta w} \end{pmatrix}, \quad A=\begin{pmatrix} -\partial^{-1} & 0 \\ 0 & \partial \end{pmatrix}, \tag{11}$$

with the Hamiltonian $H=\int(e^{-p}+e^{p}w)d\chi$. We use the following notation: $\partial=d/d\chi$ is the total derivative and ∂^{-1} is its ''inverse.''

From the results by Antonowicz and Fordy,[14–16] it follows that equations from the coupled KdV hierarchy are tri-Hamiltonian. In the variables v^0,v^1 the Hamiltonian operators have the form

$$B^{(1)}=\begin{pmatrix} -I_1 & \partial \\ \partial & 0 \end{pmatrix}, \quad B^{(2)}=\begin{pmatrix} I_0 & 0 \\ 0 & \partial \end{pmatrix}, \quad B^{(3)}=\begin{pmatrix} 0 & I_0 \\ I_0 & I_1 \end{pmatrix},$$

where $I_0=-1/4\partial^3+1/2(v^0\partial+\partial v^0)$, $I_1=1/2(v^1\partial+\partial v^1)$, and $v^0=p_{\chi\chi}/2+p_\chi^2/4$, $v^1=w$. It was proved that all these Hamiltonian operators are compatible, and the recursion operator R_B (i.e., $B^{(n)}=R_B^{n-1}B^{(1)}$) was found:

$$R_B=\begin{pmatrix} 0 & I_0\partial^{-1} \\ 1 & I_1\partial^{-1} \end{pmatrix}.$$

It turns out that the operator A in (11) corresponds to $B^{(2)}$. Let us denote $A=A^{(2)}$ and $H=H^{(2)}$. Hamiltonian operator $B^{(1)}$ in our variables takes the form

$$A^{(1)}=-2\partial^{-1}\begin{pmatrix} e^{p}w\partial^{-1}e^{-p}+e^{-p}\partial^{-1}e^{p}w & -e^{-p}\partial^{-1}e^{p}\partial^2 \\ -\partial^2 e^{p}\partial^{-1}e^{-p} & 0 \end{pmatrix}\partial^{-1},$$

and is essentially nonlocal, while the third operator $B^{(3)}$ turns out to be of the Dubrovin–Novikov type:[23]

$$A^{(3)}=\frac{1}{2}\begin{pmatrix} 0 & -\partial+p_\chi \\ -\partial-p_\chi & 2w\partial+w_\chi \end{pmatrix}.$$

For all three Hamiltonian operators $A^{(i)}$, $i=1,2,3$, we have an infinite-dimensional analog of the Darboux theorem: there exist coordinates where we obtain a particularly simple expressions for $A^{(i)}$ with constant coefficients. The operator $A^{(3)}$ is of the Dubrovin–Novikov type:[23]

$$\theta^{ij}=g^{ij}\partial-g^{is}\Gamma^j_{sk}u^k_\chi,$$

with the metric

$$g_{ij}=-2\begin{pmatrix} 2w & 1 \\ 1 & 0 \end{pmatrix},$$

and $\Gamma^j_{sk}=-1/2g^{jl}(g_{sl,k}+g_{kl,s}-g_{sk,l})$ being the coefficients of the corresponding Levi–Cività connection. In the flat coordinates of the metric $u^1=e^{-p}+e^p w$, $u^2=e^{-p}-e^p w$, where the metric has the form

$$\tilde{g}_{ij}=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},$$

the operator $A^{(3)}$ takes the form $\tilde{A}^{(3)}=\tilde{g}^{ij}\partial$.

The operator $A^{(2)}$ has the local form

$$\tilde{A}^{(2)}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\partial,$$

in variables $u^1=p_\chi$ and $u^2=w$. The essentially nonlocal operator $A^{(1)}$ will take the simple local form

$$\tilde{A}^{(1)}=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\partial$$

in the coordinates $u^1=1/2(p_{\chi\chi}+w)+1/4(p_\chi^2+w^2)$, $u^2=1/2(p_{\chi\chi}-w)+1/4(p_\chi^2+w^2)$.

Let us note that nonlocal Hamiltonian operators arise naturally in applications, and in many cases they have a clear geometrical nature (see, e.g., Ref. 24).

The Hamiltonians corresponding to the operators $A^{(1)}$ and $A^{(3)}$ have the form

$$H^{(1)}=0,\quad H^{(3)}=-\int\left[\frac{1}{2}(e^{-p}+e^p w)[\partial^{-1}(e^{-p}-e^p w)]^2+e^p\right]d\chi.$$

The variational derivatives of the functional $F=\int f(p,w)[\partial^{-1}g(p,w)]^2 d\chi$ are

$$\frac{\delta H}{\delta p}=f_p[\partial^{-1}g]^2-2g_p\partial^{-1}f\partial^{-1}g,\quad \frac{\delta H}{\delta w}=f_w[\partial^{-1}g]^2-2g_w\partial^{-1}f\partial^{-1}g.$$

In variables w^1,w^2 such that $w^1_\chi=u^1=e^{-p}+e^p w$, $w^2_\chi=u^2=e^{-p}-e^p w$, Eqs. (10) take the form

$$\begin{pmatrix}w^1\\ w^2\end{pmatrix}=\begin{pmatrix}-\partial^{-1} & 0\\ 0 & \partial^{-1}\end{pmatrix}\begin{pmatrix}\dfrac{\delta H}{\partial w^1}\\ \dfrac{\delta H}{\delta w^2}\end{pmatrix},$$

with the local Hamiltonian

$$\tilde{H}^{(3)}=-\frac{1}{2}w^1_\chi(w^2)^2-\frac{2}{w^1_\chi+w^2_\chi}.$$

The recursion operator R_A [i.e., $A^{(n)}=R_A^{n-1}A^{(1)}$] takes the form

$$R_A=\frac{1}{2}\begin{pmatrix}0 & -1+p_\chi\partial^{-1}\\ \partial^2+p_\chi\partial & w+\partial w\partial^{-1}\end{pmatrix},$$

and is invertible:

$$R_A^{-1}=2\partial^{-1}\begin{pmatrix}e^p w\partial^{-1}e^{-p}+e^{-p}\partial^{-1}e^p w & e^{-p}\partial^{-1}e^p\\ -\partial^2 e^p\partial^{-1}e^{-p} & 0\end{pmatrix}.$$

The recursion operator generates commuting flows of our system (10). Their Hamiltonians give rise to higher conservation laws (in variables v^0,v^1 they have been discussed in Ref. 15).

The Lie–Bäcklund algebra for the full SHG equations (1) was found in Ref. 9 by Zhiber. It was shown that the algebra consists of two parts, namely the following (with the slight change in notation).

(i) The *nonlinear Schrödinger type*,

$$g_{n+1}=\frac{i}{2}\,\partial g_n-4iq_2\partial^{-1}(q_2g_n^*+q_2^*g_n),\quad f_{n+1}=\frac{(g_{n+1})_\tau}{2q_1}$$

(the functions f and g determine commuting flows as follows: $q_{1T}=f$, $q_{2T}=g$);

(ii) the *Liouville type*,

$$f=\left(q_1\,\frac{\partial}{\partial\tau}+2q_{1_\tau}\right)\Phi,\quad g=2q_1^2\Phi,$$

where $\Phi=\Phi(z,z_\tau,z_{\tau\tau},\ldots)$, $z=q_1^*q_{1_\tau}-q_{1_\tau}^*q_1$.

In our case [the reduction to (6)] the symmetries of the type (i) give rise to the symmetries generated by the recursion operator R_A, while the symmetries of type (ii) degenerate into the trivial symmetry $p_T=P_\tau$, $w_T=w_\tau$.

To the best of our knowledge, the Hamiltonian structures of the SHG equations have not been discussed in the literature so far.

IV. STATIONARY POINTS OF COMMUTING FLOWS

The system (10) possesses the trivial commuting flow (the shift along χ):

$$p_T=p_\chi,\quad w_T=w_\chi. \tag{12}$$

Operating with R_A on this flow we arrive at the new flow,

$$p_T=\tfrac{1}{2}(wp_\chi-w_\chi),\quad w_T=\tfrac{1}{2}(p_{\chi\chi}+\tfrac{1}{2}p_\chi^2+\tfrac{3}{2}w^2)_\chi, \tag{13}$$

commuting with our equations (10) and corresponding in variables v^0,v^1 to the second flow in the coupled KdV hierarchy:

$$v_T^0=-\tfrac{1}{4}v^1_{\chi\chi\chi}+v^0v^1_\chi+\tfrac{1}{2}v^0_\chi v^1,\quad v^1_T=v^0_\chi+\tfrac{3}{2}v^1v^1_\chi.$$

Continuing the procedure, we will obtain the coupled KdV equations and the whole coupled KdV hierarchy, represented in the variables p and w.

Let us consider the stationary points of the simplest commuting flows. For example, let us equate to zero the linear combination of the right-hand sides of (12) and (13), namely

$$\tfrac{1}{2}(wp_\chi-w_\chi)+ap_\chi=0,$$

$$\tfrac{1}{2}(p_{\chi\chi}+\tfrac{1}{2}p_\chi^2+\tfrac{3}{2}w^2)_\chi+aw_\chi=0,$$

which, after integrating, can be rewritten as follows:

$$p_{\chi\chi}+\tfrac{1}{2}p_\chi^2+\tfrac{3}{2}w^2+2aw=b,\quad p=\ln|w+2a|+f(\tau). \tag{14}$$

Here $a,b=\text{const}$, $f(\tau)$ is an arbitrary function, and it is assumed that $w\neq-2a$.

Thus, in order to get particular solutions, we have to consider the overdetermined system formed by initial equations (10) and differential constraints (14):

$$p_{\chi\tau}=e^{-p}-e^pw,\quad w_\tau=e^pp_\chi,$$

$$p_{\chi\chi}+\tfrac{1}{2}p_\chi^2+\tfrac{3}{2}w^2+2aw=b,$$

$$w=\epsilon e^{p-f(\tau)}-2a,\quad \epsilon=\pm1.$$

Eliminating the variable w with the help of the last relation, we arrive at the overdetermined system for the variable p and obtain the following expressions for the first derivatives:

$$p_\chi = \sqrt{A}, \quad p_\tau = \epsilon e^{f(\tau)}\sqrt{A} + f'(\tau),$$

$$A = -e^{2(p-f(\tau))} + 4\epsilon a e^{p-f(\tau)} - 2\epsilon e^{-p-f(\tau)} - 4a^2 + 2b.$$

Since the compatibility condition $p_{\chi\tau} = p_{\tau\chi}$ gives us $f'(\tau) = 0$, we see that $f(\tau) = \text{const}$. Let us denote $c = -\epsilon\epsilon^f$ so that $p_\tau = -cp_\chi$ and $w = -(1/c)e^p - 2a$. Then $p = p(x)$, $x = \chi - c\tau$ and p satisfies the equation

$$\frac{1}{2}p'^2 = -\frac{1}{2c^2}e^{2p} - \frac{2a}{c}e^p + \frac{1}{c}e^{-p} + \tilde{b}, \quad \tilde{b} = b - 2a^2.$$

Since the variable p is connected with the amplitude variable Q by formula $p = -\ln Q$, the latter is determined by the equation

$$Q'^2 = \frac{2}{c}Q^3 + 2\tilde{b}Q^2 - \frac{4a}{c}Q - \frac{1}{c^2},$$

which is solvable in elliptic functions:

$$Q(\chi - c\tau) = 2c\left[\wp(\chi - c\tau + \omega'; g_2, g_3) - \frac{\tilde{b}}{6}\right]. \tag{15}$$

Here $g_2 = \tilde{b}^2/3 + 2a/c^2$, $g_3 = 1/4c^4 - \tilde{b}^3/27 - a\tilde{b}/3c^2$ are the invariants of the Weierstrass $\wp$ function and ω' is a constant. To obtain the real and nonsingular periodic solution we have to impose the condition $\Delta = g_2^3 - 27g_3^2 > 0$ and to take ω' equal to the imaginary half-period of the Weierstrass $\wp$ function.

To obtain the phase of the wave we have to integrate the equations

$$\phi_\chi = w = -\frac{1}{c}Q^{-1} - 2a, \quad \phi_\tau = Q^{-1},$$

from which it follows that

$$\phi = \tilde{\phi}(\chi - c\tau) - 2a\chi, \tag{16}$$

and the function $\tilde{\phi}$ satisfies the equation

$$\tilde{\phi}'(x) = -\frac{1}{2c^2(\wp(x+\omega') - \wp(k))}, \quad \wp(k) = \frac{\tilde{b}}{6} = \text{const}.$$

This quadrature is given by means of the Weierstrass σ and ζ functions (see, e.g., Ref. 25) as follows:

$$\tilde{\phi}(x) = -\frac{1}{2c^2\wp'(k)}\left[\ln\frac{\sigma(x+\omega'-k)}{\sigma(x+\omega'+k)} + 2\zeta(k)x\right] + \phi_0, \quad \phi_0 = \text{const}. \tag{17}$$

The formulas (15)–(17) describe the four-parametric periodic solution of the SHG equations (one-phase solution). Solutions of that type can be reparametrized by means of the spectral parameter of the corresponding linear system and expressed in terms of Jacobi elliptic functions (see, e.g., the review[26]). The stationary points of the higher commuting flows will determine the finite-gap solutions.

If $a = 0$, we get from (15)–(17) the traveling wave solution, which is connected with the traveling wave solution of the Tzitzéica equation by point transformation. There also exists a connection between self-similar solutions, as will be shown below.

V. SIMILARITY REDUCTIONS OF THE SHG EQUATIONS AND PAINLEVÉ III

Similarity reductions of the three-wave interaction system were investigated in Refs. 27–31. In them were found the reductions to Painlevé III, IV, V, and VI. In this section we investigate some particular invariant solutions of the SHG equations (3) that can be of physical interest. (The group of Lie point symmetries is given in Sec. I). We find connection of these solutions with similarity solutions of the Tzitzéica equation.[17] The latter are either connected with solutions of Painlevé III[18,19] or solvable in elliptic functions.[20]

Let us consider first the traveling wave solution of (3). We seek a solution of (3) in the form

$$Q=Q(x), \quad \phi=\phi(x), \quad x=\chi-c\tau,$$

which yields the system

$$(\ln Q)''-(\phi')^2=\frac{Q}{c}, \quad (Q\phi')'=0. \tag{18}$$

Integrating the second equation of (18), we obtain $\phi'=\beta/Q$, $\beta=\text{const}$, and the single equation for the variable Q:

$$(\ln Q)''=\frac{Q}{c}+\frac{\beta^2}{Q^2}. \tag{19}$$

The case $\beta=0$ leads to solutions of the Liouville equation. Therefore we suppose that $\beta\neq 0$. The transformation

$$Q=(\beta^2|c|)^{1/3}\exp(v), \quad x=\left|\frac{c}{\beta}\right|^{1/3} y$$

transforms (19) to the form

$$v_{yy}=\epsilon e^{v}+e^{-2v}, \tag{20}$$

where $\epsilon=\text{sign}\, c$. Equation (20) coincides with the traveling wave reduction of the Tzitzéica equation,

$$v_{\xi\eta}=-\epsilon e^{v}-e^{-2v}, \tag{21}$$

with

$$v=v(y), \quad y=\frac{\xi-c\eta}{\sqrt{c}}, \quad c>0.$$

It turns out that another similarity reduction of the SHG equations, for self-similar solutions, also yields the similarity reduction of the Tzitzéica equation. Indeed, let us consider the solution of (3) invariant with respect to the scaling symmetry,

$$\tilde{\tau}=e^{a}\tau, \quad \tilde{\chi}=e^{na}\chi, \quad \tilde{Q}=e^{-(n+1)a}Q.$$

We may define the similarity variables

$$x=\frac{\chi}{\tau^n}, \quad F(x)=\tau\chi Q, \quad \phi(x), \tag{22}$$

and find the following system of nonlinear ordinary differential equations;

$$x^2(\ln F)''+x(\ln F)'-x^2(\phi')^2=\frac{F}{n}, \quad (F\phi')'=0. \tag{23}$$

The second equation of (23) gives $\phi' = \alpha/F$, $\alpha = \text{const}$. The case $\alpha = 0$ leads to solutions of the Liouville equation. Therefore we suppose that $\alpha \neq 0$. The variable $F(x)$ satisfies the single equation

$$x^2(\ln F)'' + x(\ln F)' = \frac{F}{n} + x^2 \frac{\alpha^2}{F^2}. \tag{24}$$

The change of variables

$$F = \frac{4|n|}{9} y \exp(v), \quad x = \frac{8|n|}{27|\alpha|} y^{3/2}$$

transforms Eq. (24) to the form

$$v_{yy} = -\frac{v_y}{y} + \frac{1}{y}[e^{-2v} + \epsilon e^{v}], \tag{25}$$

with $\epsilon = \text{sign } n$.

If $\epsilon = -1$ then Eq. (25) possesses the trivial solution $v = 0$. The corresponding solution of our equations (3) has the following nontrivial form:

$$Q(\chi, \tau) = -(n\alpha^2)^{1/3} \chi^{-1/3} \tau^{-(1+2n/3)},$$

$$\phi(\chi, \tau) = -3\left(\frac{\alpha}{n}\right)^{1/3} \chi^{1/3} \tau^{-n/3} + \phi_0, \quad \phi_0 = \text{const}.$$

Using the last symmetry in (4) we obtain a solution containing an arbitrary function $T(\tau)$:

$$Q(\chi, \tau) = \frac{1}{3} T(\tau) T'(\tau) \chi^{-1/3}, \quad \phi(\chi, \tau) = T(\tau) \chi^{1/3} + \phi_0.$$

Equation (25) coincides with the scaling reduction of Eq. (21) with $v = v(y)$, $y = -\xi\eta$ and is reducible using the transformation

$$W = \epsilon y^{1/4} e^{v}, \quad z = \frac{4}{3} y^{3/4},$$

to a special case of the Painlevé III equation,[18]

$$W'' = \frac{W'^2}{W} - \frac{W'}{z} + \frac{4}{3}\frac{W^2}{z} + \frac{1}{W}, \tag{26}$$

which was studied in Ref. 19 by the method of isomonodromic deformations.

In Ref. 19 this equation was studied with the help of an auxiliary 3×3 linear system, which can be deduced from the 3×3 linear problem of the Tzitzéica equation.[17]

The similarity reduction from the SHG equations implies the auxiliary 2×2 linear system. Scaling in (8) the independent and dependent variables are as follows:

$$s(\chi, \tau) = -zW^{-1}(z)/2, \quad z = (6\tau\chi)^{1/2},$$

$$u(\chi, \tau) = -3\chi z^{-1} W^{-1}(z), \quad \mu = (3\chi\lambda)^{-1},$$

$$v(\chi, \tau) = \chi^{-1}(1 - zW_z(z)W^{-1}(z))/2,$$

$$w(\chi, \tau) = -\chi z^{-1} W^{-1}(z)/2, \quad \Psi(\chi, \tau; \lambda) = \Phi(z; \mu),$$

we obtain the linear system

$$\Phi_z=\left[\mu\begin{pmatrix}0 & 0\\ -W & 0\end{pmatrix}+\frac{1}{2}\begin{pmatrix}0 & -W^{-1}\\ W^{-1} & 0\end{pmatrix}\right]\Phi,$$

$$\phi_\mu=\left[\begin{pmatrix}0 & 0\\ -\frac{zW}{2} & 0\end{pmatrix}+\frac{1}{4\mu}\begin{pmatrix}1-zW_zW^{-1} & -zW^{-1}\\ -zW^{-1} & 1-zW_zW^{-1}\end{pmatrix}+\frac{1}{3\mu^2}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\right]\Phi.$$

These equations are compatible provided that $W(z)$ satisfies (26).

It is interesting that two different reductions of the SHG equations and of the Tzitzéica equation are connected. One may expect that there exists some more general connections between these equations.

VI. CONCLUDING REMARKS

In the present paper we have discussed the relationship of Eqs. (6) to the coupled KdV hierarchy and the Hamiltonian formalism of these equations. Let us note that the full SHG equations (1) can be represented in Lagrangian form,

$$\frac{\delta L}{\delta q_i}=0,\quad \frac{\delta L}{\delta q_i^*}=0,\quad i=1,2,$$

with the Lagrangian

$$L=\int i\left[\frac{1}{2}\,(q_1q_{1\chi}^*-q_1^*q_{1\chi}+q_2q_{2\tau}^*-q_2^*q_{2\tau})-q_1^{*2}q_2+q_1^2q_2^*\right]d\chi,$$

analogously to the finite-dimensional case (see, e.g., Ref. 32), and in Hamiltonian form,

$$\begin{pmatrix}q_1\\ q_2\\ q_1^*\\ q_2^*\end{pmatrix}_T=\begin{pmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ -1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\end{pmatrix}\begin{pmatrix}\frac{\delta H}{\delta q_1}\\ \frac{\delta H}{\delta q_2}\\ \frac{\delta H}{\delta q_1^*}\\ \frac{\delta H}{\delta q_2^*}\end{pmatrix},\tag{27}$$

with the Hamiltonian

$$H=\int i\left[\frac{1}{2}\,(q_1q_{1X}^*-q_1^*q_{1X}-q_2q_{2X}^*+q_2^*q_{2X})-q_1^{*2}q_2+q_1^2q_2^*\right]dX,$$

and $T=\chi+\tau$, $X=\chi-\tau$. The variable T is nothing but the laboratory x coordinate used in the Introduction while X is a mixture of x and t. The other Hamiltonian operators, which can be obtained from (27) using a recursion operator, and their connections with those found in the present paper will be discussed elsewhere.

ACKNOWLEDGMENTS

We would like to thank A. I. Bobenko, E. V. Ferapontov, and A. M. Kamchatnov for valuable discussions and A. V. Zhiber for giving us the reprint.[9] K. R. Khusnutdinova acknowledges the support of Deutscher Akademischer Austauschdienst. H. Steudel wishes to acknowledge support by the Max Planck Society.

[1] P. A. Franken, A. E. Hill, C. W. Peters, and G. Weinreich, Phys. Rev. Lett. **7**, 118 (1961).
[2] J. A. Armstrong, N. Bloembergen, J. Ducuing, and P. S. Pershan, Phys. Rev. **127**, 1918 (1962).
[3] R. W. Terhune, P. D. Maker, and C. M. Savage, Appl. Phys. Lett. **2**, 54 (1963).

[4] S. A. Akhmanov, V. A. Vysloukh, and A. S. Chirkin, *Optics of Femtosecond Laser Pulses* (American Institute of Physics, New York, 1992).
[5] H. Steudel, C. Faria, M. Paris, A. M. Kamchatnov, and O. Steuernagel, to appear in Opt. Commun.
[6] F. G. Bass and V. G. Sinitsyn, Ukr. Fiz. J. **17**, 124 (1972).
[7] S. A. Akhmanov, A. S. Chirkin, K. N. Drabovich, A. I. Kovrigin, R. V. Khokhlov, and A. P. Sukhorukov, IEEE J. Quantum Electron. **QE-4**, 598 (1968).
[8] D. J. Kaup, Stud. Appl. Math. **59**, 25 (1978).
[9] A. V. Zhiber, in *Dinamika Neodnorodnoi Zhidkosti* (Institut Gidrodinamiki SO AN SSSR, Novosibirsk, 1980), Vol. 44, p. 3.
[10] H. Steudel and R. Meinel, Physica D **21**, 155 (1986).
[11] L. M. Alonso, J. Math. Phys. **21**, 2342 (1980).
[12] D. Levi, A. Sym, and S. Wojciechowski, J. Phys. A **16**, 2423 (1983).
[13] M. Boiti, P. J. Caudrey, and F. Pempinelli, Nuovo Cimento B **83**, 71 (1984).
[14] M. Antonowicz and A. P. Fordy, Phys. Lett. A **122**, 95 (1987).
[15] M. Antonowicz and A. P. Fordy, Physica D **28**, 345 (1987).
[16] M. Antonowicz and A. P. Fordy, Commun. Math. Phys. **124**, 465 (1989).
[17] M. Tzitzéica, Comptes, Rend. hebd. Seances l'Acad. Sciences Paris **150**, 955,1227 (1910).
[18] P. A. Clarkson, J. B. McLeod, A. Ramani, and P. J. Olver, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **17**, 789 (1986).
[19] A. V. Kitaev, Zap. Nauchn. Semin. LOMI **161**, 45 (1987).
[20] S. S. Safin and R. A. Sharipov, Teor. Mat. Fiz. **95**, 146 (1993).
[21] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[22] P. J. Olver, *Applications of the Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
[23] B. A. Dubrovin and S. P. Novikov, Usp. Mat. Nauk **44**, 29 (1989).
[24] E. V. Ferapontov, Am. Math. Soc. Trans. **170**, 33 (1995).
[25] I. S. Gradshtein and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1980).
[26] A. M. Kamchatnov, Phys. Rep. **286**, 199 (1997).
[27] A. S. Fokas, R. A. Leo, L. Martina, and G. Soliany, Phys. Lett. A **115**, 329 (1986).
[28] R. A. Leo, L. Martina, and G. Soliany, J. Math. Phys. **27**, 2623 (1986).
[29] R. A. Leo, L. Martina, G. Soliany, and G. Tondo, Prog. Theor. Phys. **76**, 739 (1986).
[30] L. Martina and P. Winternitz, Ann. Phys. (N.Y.) **196**, 231 (1989).
[31] A. V. Kitaev, J. Phys. A **23**, 3543 (1990).
[32] V. M. Agranovich, S. A. Darmanyan, A. M. Kamchatnov, T. A. Leskova, and A. D. Boardman, Phys. Rev. E **55**, 1894 (1997).

Motion of curves and surfaces and nonlinear evolution equations in (2+1) dimensions

M. Lakshmanan[a)]
Centre for Nonlinear Dynamics, Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, India

R. Myrzakulov[b)]
High Energy Physics Institute, National Academy of Sciences, 480082, Alma-Ata-82, Kazakstan, and Centre for Nonlinear Problems, PO Box 30, 480035, Alma-Ata-35, Kazakstan

S. Vijayalakshmi
Centre for Nonlinear Dynamics, Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, India

A. K. Danlybaeva
Centre for Nonlinear Problems, PO Box 30, 480035, Alma-Ata-35, Kazakstan

(Received 28 October 1997; accepted for publication 5 February 1998)

It is shown that a class of important integrable nonlinear evolution equations in (2+1) dimensions can be associated with the motion of space curves endowed with an extra spatial variable or equivalently, moving surfaces. Geometrical invariants then define topological conserved quantities. Underlying evolution equations are shown to be associated with a triad of linear equations. Our examples include Ishimori equation and Myrzakulov equations which are shown to be geometrically equivalent to Davey-Stewartson and Zakharov-Strachan (2+1) dimensional nonlinear Schrödinger equations, respectively. © *1998 American Institute of Physics.* [S0022-2488(98)02006-4]

The motion of curves in E^3 and the defining equations of surfaces have drawn wide interest in the past, especially since they give nice geometric interpretations of nonlinear evolution equations in (1+1) dimensions.[1–6] There are many physical situations where they play a natural role. Particularly they have important connections with soliton equations solvable by the Ablowitz, Kaup, Newell and Segur (AKNS) formalism.[1–6] For a recent renaissance of interest, see, for example, Ref. 7 and References therein. Other examples include dynamics of interfaces, surfaces and fronts, vortex filaments, supercoiled DNAs, magnetic fluxes, deformation of membranes, dynamics of proteins, propagation of flame fronts and so on.[8,9] Of special interest among these systems is the dynamics of isotropic Heisenberg ferromagnetic spin chains where an interesting equivalence with nonlinear Schrödinger family of equations arise in a natural and physical way.[10]

The question then arises as to whether nonlinear evolution equations in (2+1) dimensions of importance, especially spin equations, can be given a similar geometrical setting associated with motion of curves endowed with an additional coordinate or equivalent motion of surfaces. In this paper, we develop a general theory of obtaining nonlinear evolution equations in (2+1) dimensions as the compatibility requirements of the geometric equations defining the motion of curves and surfaces, incidentally equivalent to a triad of linear equations. Our analysis also brings out certain natural topological integral invariants associated with the resultant systems. We then consider a class of (2+1) dimensional spin systems including (2+1) dimensional generalized Heisenberg ferromagnatic spin systems and deduce their equivalent nonlinear Schrödinger family including Davey-Stewartson, Zakharov-Strachan and other equations. In this context, we also wish to add that in recent times a general approach (mainly due to Konopelchenko) has appeared in the literature[11,12] concerning the interpretation of hierarchies of integrable (2+1) dimensional systems

a)Electronic mail: lakshman@bdu.ernet.in
b)Electronic mail: cnlpmyra@satsun.sci.kz

 © 1998 American Institute of Physics

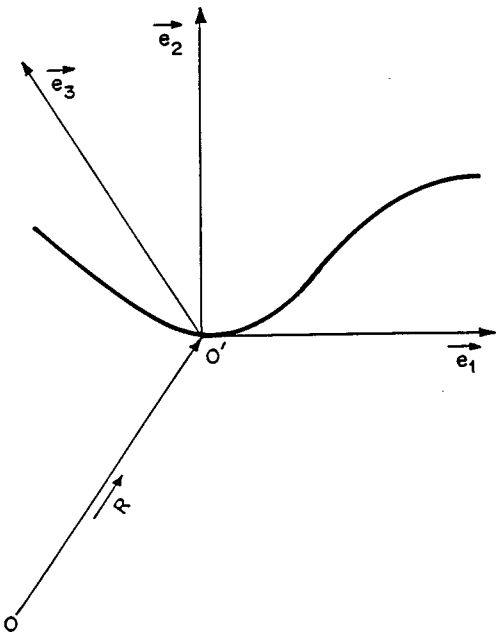

FIG. 1. Motion of a space curve in E^3 with the orthogonal trihedral $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$.

as special motions of surfaces in E^3 by inducing the surfaces in it via the solutions to two dimensional linear problems (2D LPs). But in the present paper, we are interested in extending the theory of moving space curve formalism in (1+1) dimensions to (2+1) dimensions.

We consider the space curve in E^3 as in Fig. 1, defined by the Serret-Frenet equations,[13]

$$\mathbf{e}_{jx}=\mathbf{D}\wedge\mathbf{e}_j, \quad \mathbf{D}=\tau\mathbf{e}_1+\kappa\mathbf{e}_3, \quad j=1,2,3, \tag{1}$$

where $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are the unit tangent, normal and binormal vectors, respectively, to the curve. Here x is the arclength parametrizing the curve. The curvature and torsion of the curve are defined, respectively, as

$$\begin{aligned}\kappa&=(\mathbf{e}_{1x}\cdot\mathbf{e}_{1x})^{1/2},\\ \tau&=\kappa^{-2}\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1xx}).\end{aligned} \tag{2}$$

The unit tangent vector $\mathbf{e}_1$ is given by $\mathbf{e}_1=\partial\mathbf{r}/\partial x=(1/\sqrt{g})(\partial\mathbf{r}/\partial\theta)$ where g is the metric $g=(\partial\mathbf{r}/\partial\theta)(\partial\mathbf{r}/\partial\theta)$ on the curve such that $x(\theta,t)=\int_0^\theta\sqrt{g(\theta',t)}d\theta'$. Here θ defines a smooth curve and $\mathbf{r}(\theta,t)$ is the position vector of a point on the curve at time t. Considering the motion of such a twisted curve, the time evolution of the orthogonal trihedral can be easily seen[3] to be

$$\begin{aligned}\mathbf{e}_{jt}&=\mathbf{\Omega}\wedge\mathbf{e}_j,\\ \mathbf{\Omega}&=\omega_1\mathbf{e}_1+\omega_2\mathbf{e}_2+\omega_3\mathbf{e}_3, \quad j=1,2,3,\end{aligned} \tag{3}$$

where $\omega_i, i=1,2,3$ are some functions of κ and τ and their derivatives. Then the compatibility of Eqs. (1) and (3) gives the evolution equations for κ and τ as[3]

$$\begin{aligned}\kappa_t&=\omega_{3x}+\tau\omega_2,\\ \tau_t&=\omega_{1x}-\kappa\omega_2,\\ \omega_{2x}&=\tau\omega_3-\kappa\omega_1.\end{aligned} \tag{4}$$

Now let us pass on to the subject of this paper, namely, to the theory of curves in (2+1) dimensions by endowing them with an additional spatial variable y. Alternately this could repre-

sent a surface specified by the trihedral $\mathbf{e}_j(x,y,t)$ which is set in motion. A starting point of our approach is the observation that the y-evolution of the trihedral $\mathbf{e}_j$ is determined by the following system of linear equations[14] (see also Refs. 15, 16);

$$\mathbf{e}_{jy}=\mathbf{\Gamma}\wedge\mathbf{e}_j,$$
$$\mathbf{\Gamma}=\gamma_1\mathbf{e}_1+\gamma_2\mathbf{e}_2+\gamma_3\mathbf{e}_3. \tag{5}$$

Here γ_j's are functions to be determined. Using the fact that $\mathbf{e}_i$'s, $i=1,2,3$ form an orthogonal trihedral and using (1) and (5) in the compatibility condition $\mathbf{e}_{jxy}=\mathbf{e}_{jyx}$, we obtain

$$\gamma_1=\partial_x^{-1}(\tau_y+\kappa\gamma_2),\quad \gamma_3=\partial_x^{-1}(\kappa_y-\tau\gamma_2), \tag{6a}$$

and $\gamma_2(x,y,t)$ is a solution of the following equation:

$$-\kappa\gamma_2=\tau_y-\left[\frac{\tau\partial_x^{-1}(\kappa_y-\tau\gamma_2)-\gamma_{2x}}{\kappa}\right]_x, \tag{6b}$$

where $\partial_x^{-1}\equiv\int_{-\infty}^{x}dx$. We also note that Eqs. (1) and (5) may be identified with the Gauss-Weingarten equations of surface theory written in terms of orthogonal coordinates and so Eqs. (6) with the Codazzi-Mainardi equations. On the other hand, the condition $\mathbf{e}_{jty}=\mathbf{e}_{jyt}$ gives rise to the set of equations

$$\gamma_{1t}=\omega_{1y}+\omega_3\gamma_2-\omega_2\gamma_3,$$
$$\gamma_{2t}=\omega_{2y}+\omega_1\gamma_3-\omega_3\gamma_1, \tag{7}$$
$$\gamma_{3t}=\omega_{3y}+\omega_2\gamma_1-\omega_1\gamma_2.$$

Choosing $\gamma_1,\gamma_2,\gamma_3$ consistently so that Eqs. (4) and (7) are compatible, one can obtain (2+1) dimensional evolution equations for κ and τ. Further one can easily show that the defining equations for the motion of the trihedral, namely (1), (3) and (5) are equivalent to a set of three Riccati equations,

$$z_{lx}=-i\tau z_l+\frac{\kappa}{2}[1+z_l^2], \tag{8a}$$

$$z_{ly}=-i\gamma_1 z_l+\tfrac{1}{2}[\gamma_3+i\gamma_2]z_l^2+\tfrac{1}{2}[\gamma_3-i\gamma_2], \tag{8b}$$

and

$$z_{lt}=-i\omega_1 z_l+\tfrac{1}{2}[\omega_3+i\omega_2]z_l^2+\tfrac{1}{2}[\omega_3-i\omega_2], \tag{8c}$$

where z_l, $l=1,2,3$ is a scalar variable obtained through an orthogonal rotation of the trihedral, $z_l=(e_{2l}+ie_{3l})/(1-e_{1l})$, $e_{1l}^2+e_{2l}^2+e_{3l}^2=1$. Introducing the transformation $z_l=v_2/v_1$, we obtain the triad of equivalent linear equations:

$$v_{1x}=\frac{i\tau}{2}v_1-\frac{\kappa}{2}v_2,\quad v_{2x}=\frac{\kappa}{2}v_1-\frac{i\tau}{2}v_2, \tag{9a}$$

$$v_{1y}=\frac{i\gamma_1}{2}v_1-\frac{1}{2}(\gamma_3+i\gamma_2)v_2,\quad v_{2y}=\frac{1}{2}(\gamma_3-i\gamma_2)v_1-\frac{i\gamma_1}{2}v_2, \tag{9b}$$

$$v_{1t}=\frac{i\omega_1}{2}v_1-\frac{1}{2}(\omega_3+i\omega_2)v_2,\quad v_{2t}=\frac{1}{2}(\omega_3-i\omega_2)v_1-\frac{i\omega_1}{2}v_2. \tag{9c}$$

The compatibility of these linear equations again gives rise to Eqs. (4), (6) and (7). Thus any nonlinear evolution equation obtained through the space curve formulation in (2+1) dimensions is equivalent to a triad of linear equations. Specific examples are given below.

The above formulation then leads us in a natural way to certain topological invariants as given in the following theorem.

Theorem: The above constructed evolution equations in (2+1) dimensions possess the following integrals of motion:[14]

$$K_1=\frac{1}{4\pi}\int\int(-\kappa\gamma_2)dxdy,$$
$$K_2=\frac{1}{4\pi}\int\int(-\tau\gamma_2)dxdy. \tag{10a}$$

Proof: The proof is straightforward and follows from the various relations discussed above. For example, from (6a) we get $(-\kappa\gamma_2)_t=(\tau_t)_y-(\gamma_{1t})_x$, or $(-\kappa\gamma_2)_t=(\omega_{1x}-\kappa\omega_2)_y-(\omega_{1y}+\omega_3\gamma_2-\omega_2\gamma_3)_x$. Hence follows the first statement of the theorem. Similarly the second one may also be proved. In terms of the unit vector $\mathbf{e}_1$ these integrals of motion take the form

$$K_1=\frac{1}{4\pi}\int\int\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1y})dxdy,$$

and

$$K_2=\frac{1}{4\pi}\int\int\frac{[\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1y})][\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1xx})]}{(\mathbf{e}_{1x}^2)^{3/2}}dxdy. \tag{10b}$$

So the above quantities K_j are the conserved integrals — invariants of the (2+1) dimensional evolution equations and they play an important role in the theory. In particular, curves can be classified by the value of the topological invariants K_j but we will not pursue this aspect further. We now present a few applications of the above presented formalism in finding the geometrically equivalent counterparts of some known (2+1) dimensional spin systems. To this end, first, following Ref. 10, we identify the tangent vector $\mathbf{e}_1$ with the unit spin vector $\mathbf{e}_1\equiv\mathbf{S}(x,y,t)$. Second, we introduce the following complex transformation which is a generalization of the one given in Ref. 10:

$$q(x,y,t)=a(x,y,t)\exp\, ib(x,y,t), \tag{11}$$

where $a(x,y,t)$ and $b(x,y,t)$ are functions of κ and τ, to be determined, and obtain the equation for the complex function $q(x,y,t)$ which turns out to be an interesting integrable (2+1) dimensional nonlinear evolution equation. We want to demonstrate our approach in the following examples of the (2+1) dimensional spin systems.

A. The Myrzakulov I (M-I) equation. This equation reads as[14,15]

$$\mathbf{e}_{1t}=(\mathbf{e}_1\wedge\mathbf{e}_{1y}+u\mathbf{e}_1)_x,\quad \mathbf{e}_1=\mathbf{S}(x,y,t), \tag{12a}$$

$$u_x=-\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1y}). \tag{12b}$$

In this case, we obtain

$$\gamma_2=\frac{u_x}{\kappa}, \tag{13a}$$

$$\mathbf{\Omega}=(\omega_1,\omega_2,\omega_3)=\left(\frac{\kappa_{xy}}{\kappa}-\tau\partial_x^{-1}\tau_y,-\kappa_y,-\kappa\partial_x^{-1}\tau_y\right). \tag{13b}$$

For the functions $a(x,y,t)$ and $b(x,y,t)$ in the transformation (11), we take the form

$$a=\frac{\kappa(x,y,t)}{2}, \quad b=-\partial_x^{-1}\tau(x,y,t). \tag{14}$$

Then the function q satisfies the following evolution equation:

$$iq_t(x,y,t)=q_{xy}+Vq, \quad V_x=2(|q|^2)_y, \tag{15}$$

an equation which belongs to the class of equations originally discovered by Calogero[17] and then discussed by Zakharov[18] and recently rederived by Strachan.[19] Many of its properties have received considerable attention recently.[20]

B. The Myrzakulov III (M-III) equation. This equation which is a generalization of Eq. (12) has the form[14]

$$\mathbf{e}_{1t}=(\mathbf{e}_1\wedge\mathbf{e}_{1y}+u\mathbf{e}_1)_x+2f(cf+d)\mathbf{e}_{1y}-4cv\mathbf{e}_{1x}, \tag{16a}$$

$$u_x=-\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1y}), \quad v_x=\frac{1}{4(2fc+d)^2}(\mathbf{e}_{1x}^2)_y, \tag{16b}$$

where f, c and d are constants. In this case, using the same form for γ_i's and ω_i's as in Eq. (13) and with the following choice for the functions $a(x,y,t)$ and $b(x,y,t)$,

$$a(x,y,t)=\frac{\kappa(x,y,t)}{2(2cf+d)}, \quad b(x,y,t)=2f(cf+d)x-\partial_x^{-1}\tau, \tag{17}$$

we obtain the evolution equation

$$iq_t(x,y,t)=q_{xy}-4ic(Vq)_x+2d^2Vq, \quad V_x=(|q|^2)_y. \tag{18}$$

This equation as well as the M-III equations (16) are integrable, that is, they have the Lax representations[14] and some properties of both equations were investigated in Refs. 14–16, 21, 22. Note that Eqs. (16) and (18) admit integrable reductions: when $c=0$ they pass on to the M-I and Zakharov equations, Eqs. (12) and (15), respectively. If we consider the case $d=0$, then they reduce to the M-II and Strachan equations, respectively.[23]

C. The Ishimori equation. This equation has the form[24]

$$\mathbf{e}_{1t}=\mathbf{e}_1\wedge(\mathbf{e}_{1xx}+\sigma^2\mathbf{e}_{yy})+u_x\mathbf{e}_{1y}+u_y\mathbf{e}_{1x}, \tag{19a}$$

$$u_{xx}-\sigma^2u_{yy}=-2\sigma^2\mathbf{e}_1\cdot(\mathbf{e}_{1x}\wedge\mathbf{e}_{1y}), \tag{19b}$$

where $\sigma^2=\pm1$. For the case $\sigma^2=-1$, we obtain

$$\gamma_2=\frac{u_{xx}+u_{yy}}{-2\kappa},$$

$$\omega_1=\frac{\tau\omega_3-\omega_{2x}}{\kappa}, \tag{20a}$$

$$\omega_2=u_x\gamma_2-\kappa_x+\gamma_{3y}+\gamma_1\gamma_2, \tag{20b}$$

$$\omega_3=-\kappa\tau+\kappa u_y+u_x\gamma_3-\gamma_{2y}+\gamma_1\gamma_3. \tag{20c}$$

If we choose $a(x,y,t)$ and $b(x,y,t)$ in the form

$$a=\tfrac{1}{2}[\kappa^2+\gamma_2^2+\gamma_3^2-2\kappa\gamma_2]^{1/2}, \tag{21a}$$

$$b=\partial_x^{-1}\left(\tau-\frac{u_y}{2}+\frac{\gamma_2\gamma_{3x}-\gamma_3\gamma_{2x}-\kappa\gamma_{3x}}{\kappa^2+\gamma_2^2+\gamma_3^2-2\kappa\gamma_2}\right), \tag{21b}$$

then the function q satisfies the following equation:

$$iq_t(x,y,t)+q_{xx}-q_{yy}-2q\phi=0, \tag{22a}$$

$$\phi_{xx}+\phi_{yy}+(|q|^2)_{xx}-(|q|^2)_{yy}=0, \tag{22b}$$

where ϕ is a function of x, y and t, which is nothing but the Davey-Stewartson equation II.[6] Note that Eqs. (19) and (22) are known to be gauge equivalent to each other.[25] Here their geometrical equivalence has been estabilished. Similarly equivalence can be established for $\sigma^2=1$.

D. The (2+1) dimensional isotropic Heisenberg ferromagnet model. This equation has the form[26]

$$\mathbf{e}_{1t}=\mathbf{e}_1\wedge(\mathbf{e}_{1xx}+\mathbf{e}_{1yy}). \tag{23}$$

In this case, we get

$$\omega_1=\frac{\tau\omega_3-\omega_{2x}}{\kappa},\quad \omega_2=-\kappa_x-\gamma_{3y}-\gamma_1\gamma_2,$$
$$\omega_3=\gamma_{2y}-\kappa\tau-\gamma_1\gamma_3. \tag{24}$$

However one finds that γ_i's can not be uniquely found due to the lack of an equation of the form (12b) or (16b) or (19b). So Eq. (6b) does not appear to be solvable, indicating that no straightforward linearization is available for the (2+1) dimensional isotropic Heisenberg spin equation without introducing constraints.

In conclusion, we have shown that a geometrical formulation can be developed for a class of (2+1) dimensional nonlinear evolution equations through moving space curve formalism. The formulation shows how a class of evolution equations can be associated with a triad of linear equations. Whether such a connection alone is sufficient to prove the integrability of the underlying (2+1) dimensional systems is a further intricate question. Even in (1+1) dimensions, there are some equations which though linearizable are not integrable; for example, the spherically symmetric Heisenberg ferromagnetic spin chain equation is equivalent to a nonlocal nonlinear Schrödinger equation which is of the nonPainlevé type and so is nonintegrable.[27] So also is the (1+1) dimensional isotropic Landau-Lifshitz equation with Gilbert damping[28] and so on. However the three examples considered in this paper are integrable namely, the M-I, M-III and Ishimori equations and their geometrical equivalents namely the Zakharov, Strachan and Davey-Stewartson equations, respectively, are also integrable. Their linearization can be performed through the set of equations (9) and the associated linear eigenvalue problems can be constructed. On the other hand the pure isotropic Heisenberg spin chain equation in (2+1) dimensions (example D above) seems not to be even linearizable by the space curve formalism without constraints, indicating the necessity of additional scalar fields or nonlocal terms for its linearization. It is hoped that other (2+1) dimensional equations of interest like the (2+1) dimensional Korteweg-de Vries, Nizhnik-Novikov-Veselov, breaking soliton, etc. equations and so on may have similar geometrical interpretations in terms of moving space curves or surfaces, for example with restrictions on the value of the curvature κ or torsion τ. These possibilities are being explored at present.

Finally it is of interest to consider the connection between our approach and to that of Konopelchenko who uses the (2+1) dimensional nonlinear PDEs to induce integrable dynamics (deformations) of the induced surfaces.[11] In our approach, we consider the compatibility of three linear problems, namely equations (1), (3) and (5). Two of them [equations (1) and (5)] can be considered as equivalent to general surfaces in orthogonal coordinates. In the approach of Konopelchenko, for a given 2D LP, the variable coordinates X^i $(i=1,2,3)$ of a surface in E^3 are defined as some integrals over certain bilinear combinations of solutions ψ of the 2D LP and solutions ψ^* of the adjoint 2D LP and thus surfaces are induced via the solutions of 2D LPs. Then the compatibilty of 2D LPs with time evolution leads to hierarchies of (2+1) dimensional nonlinear evolution equations both for the coefficients of the 2D LP and for the wave function ψ. This nonlinear evolution equation induces the corresponding evolution of the induced surface. In this sense, we believe that the two approaches differ from each other and give complementary approaches to integrable (2+1) dimensional evolution equations.

ACKNOWLEDGMENTS

This work of M.L. forms part of a Department of Science and Technology, Government of India sponsored research project. R.M. wishes to thank Bharathidasan University for hospitality during his visits to Tiruchirapalli. S.V. acknowledges the receipt of the award Junior Research Fellowship from the Council of Scientific and Industrial Research, India.

[1] G. L. Lamb, Jr., Phys. Rev. Lett. **37**, 235 (1976); J. Math. Phys. **18**, 1654 (1977).
[2] F. Lund and T. Regge, Phys. Rev. D **14**, 1524 (1976).
[3] M. Lakshmanan, Phys. Lett. A **64**, 353 (1978); J. Math. Phys. **20**, 1667 (1979).
[4] R. Sasaki, Nucl. Phys. B **154**, 343 (1979).
[5] A. Sym, *Springer Lecture Notes in Physics*, edited by R. Martini (Springer, Berlin, 1985), Vol 239, p. 154.
[6] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[7] A. Doliwa and P. M. Santini, Phys. Lett. A **185**, 373 (1994).
[8] A. R. Bishop, L. J. Cambell, and P. J. Channels, in *Fronts, Interfaces and Patterns* (North-Holland, New York, 1984).
[9] P. Pelcé, in *Dynamics of Curved Fronts* (Academic, New York, 1986).
[10] M. Lakshmanan, Phys. Lett. A **61**, 53 (1977); M. Lakshmanan, Th. W. Ruijgrok, and C. J. Thompson, Physica A **84**, 577 (1976).
[11] B. G. Konopelchenko, Stud. Appl. Math. **96**, 9 (1996).
[12] B. G. Konopelchenko, Phys. Lett. A **183**, 153 (1993); J. Math. Phys. **38**, 434 (1997).
[13] J. J. Stoker, *Differential Geometry* (Wiley-Interscience, New York, 1969).
[14] R. Myrzakulov, *On Some Integrable and Nonintegrable Soliton Equations of Magnets I-IV* (HEPI, Alma-Ata, 1987).
[15] R. Myrzakulov, S. Vijayalakshmi, G. N. Nugmanova, and M. Lakshmanan, Phys. Lett. A **233**, 391 (1997).
[16] R. Myrzakulov, A. K. Danlybaeva; R. Myrzakulov, and R. N. Syzdykova, preprints CNLP, Alma-Ata, 1994.
[17] F. Calogero, Lett. Nuovo Cimento **14**, 43 (1975).
[18] V. E. Zakharov, *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer-Verlag, Berlin, 1980).
[19] I. A. B. Strachan, Inverse Probl. **8**, L21 (1992).
[20] R. Radha and M. Lakshmanan, J. Phys. A **30**, 3229 (1997).
[21] R. Myrzakulov and G. N. Nugmanova, Izv. Nat. Akad. Sci. Rep. Kazakstan, Ser. Fiz. Mat. **6**, 32 (1995).
[22] G. N. Nugmanova, Ph. D. thesis, Kazak State University, Alma-Ata, 1992.
[23] I. A. B. Strachan, J. Math. Phys. **34**, 243 (1993).
[24] Y. Ishimori, Prog. Theor. Phys. **72**, 33 (1984).
[25] B. G. Konopelchenko, *Solitons in Multidimensions* (World Scientific, Singapore, 1993).
[26] M. Lakshmanan and M. Daniel, Physica A **107**, 533 (1981).
[27] K. Porsezian and M. Lakshmanan, J. Math. Phys. **32**, 2923 (1991).
[28] M. Daniel and M. Lakshmanan, Physica A **120**, 125 (1983).

The secular solutions of the linearized Korteweg–de Vries equation

H. Leblond
Laboratoire P.O.M.A., E.P.-C.N.R.S. 130, Université d'Angers, 2 Bd Lavoisier, 49045 Angers, Cedex 01, France

(Received 15 December 1997; accepted for publication 20 January 1998)

We study the inhomogeneous linearized Korteweg–de Vries (KdV) equation. It is solved by the inverse scattering transform method. The secular-producing terms on the right-hand side (rhs) are characterized in several ways: first we give a mathematical characterization as resonant terms. Second, the secular-producing terms are interpreted as conserved densities of the KdV equation. Third, it is checked that the removal of all linear terms from the rhs, polynomial in the solution of KdV, ensures the boundness of the solution of the linearized equation. Fourth, considering this solution itself as the rhs, we determine which part of it is secular producing, and which part is not. © *1998 American Institute of Physics.*
[S0022-2488(98)03006-0]

I. INTRODUCTION

In multiscale expansions, as in other perturbative methods, conditions are obtained at each step of the calculus, or at each perturbative order, by the requirement that all terms in the expansion are finite. Indeed, the equations derived at each step of the perturbative scheme can present unbounded solutions, called ''secular,'' especially when they grow linearly with time. The elimination of these unphysical solutions give relations to be satisfied by the quantities defined at the previous order.

Let us consider some multiscale expansion, leading to the Korteweg–de Vries (KdV) equation, satisfied by the main term f of the wave under consideration. The KdV equation reads:

$$\partial_t f - 6 f \partial_x f + \partial_x^3 f = 0. \tag{1}$$

Let us consider now some higher order correction g to f. g can be treated either by the renormalization formalism of Kodama and Taniuti,[1] or by the multi-times formalism of Kraenkel *et al.*[2,3] It will obey a KdV equation linearized about f:

$$\partial_t g - 6 \partial_x (f g) + \partial_x^3 g = w. \tag{2}$$

The right-hand side member (rhs) w is some polynomial expression, depending on f and its derivatives, the corrections to f of lower order than g, and on their derivatives.

It is well known that the KdV equation (1) is completely integrable, i.e., that the Cauchy problem for it can be solved by use of the inverse scattering transform (IST) method.[4–6] The IST method also gives explicit formulas for the resolution of the linearized KdV equation (2).[1] We recall this resolution in Sec. II. Unfortunately, the solution computed this way [formulas (15)–(17)] is not always bounded. If not, it is called ''secular.'' This phenomena occurs when a term on the rhs ''resonates,'' that is, from a mathematical point of view, when this term is a solution of the homogeneous equation (5). Physically, the resonance phenomenon occurs when each component of the inverse scattering transform of the source term evolves in time in the same way as the corresponding component in the transform of the main term. This gives rise to some mathematical characterization of the secular-producing terms that we give in Sec. III.

It can be read in the documentation[3,1] that the secular-producing terms are the linear ones. We discuss this assertion in detail in Sec. IV. It is seen that the secular-producing terms coincide with the derivatives of the conserved densities of the KdV equation, and that the removal of the linear terms, using the multiple time formalism of Kraenkel *et al.* and the KdV hierarchy, removes in

 © 1998 American Institute of Physics

fact all the secular-producing terms. Section V is devoted to the question of whether the solutions of the linearized KdV equation (2) are secular producing: indeed, the solution of the equation governing the second order of the perturbative expansion appears on the rhs of the equation that governs the evolution of the term of order three, and so on. It is thus necessary to answer this question, if we intend to prove that a given multiscale expansion, whose first term obeys KdV, can be solved up to any order, with all its terms bounded.

II. RESOLUTION OF THE LINEARIZED KDV EQUATION

The equation is linear, thus one could want to avoid using the heavy machinery of the IST method, and solve it in a simpler way, e.g., by use of a Fourier transform. In fact, here use of the IST method is necessary. Indeed, the solution f of the KdV equation intervenes in the present equation as an essential parameter, and in the general case it can be only expressed in terms of its inverse transform.

Let us recall some features related to the resolution of the KdV equation (1) through the inverse scattering transform (IST) method. Let us call ϕ_k and ψ_k the Jost functions, which are the solutions of the scattering equation:

$$\partial_x^2\psi+(k^2-f\,)\psi=0, \tag{3}$$

which satisfy the conditions

$$\phi_k \underset{x\to+\infty}{\sim} e^{+ikx}, \quad \psi_k \underset{x\to-\infty}{\sim} e^{-ikx}. \tag{4}$$

For each k real, or purely imaginary zero of the scattering coefficient $a(k)$, let us define

$$\Phi_k=\phi_k^2$$

and

$$\Psi_k=\partial_x(\psi_k^2).$$

The functions $\Phi_k e^{-8ik^3t}$ and $\Psi_k e^{-8ik^3t}$ are solutions of the homogeneous linearized KdV equation:

$$\partial_t g-6\partial_x(fg)+\partial_x^3 g=0. \tag{5}$$

Thus, the general solution of this equation is a linear combination (or integral) of these functions. Fortunately, there exists a closure formula for them (it is given, although with some typing mistakes, in Ref. 1, and can be proved in the same way as the closure formula for the Jost functions, which is detailed, e.g., in Ref. 6). It reads:

$$\begin{aligned} &i\int_{-\infty}^{+\infty}\frac{dk}{2\pi}\,\Phi_k(x,t)\,\frac{1}{ka^2(k)}\,\Psi_k(y,t)+\sum_{l=1}^{n}\frac{1}{i\kappa_l\dot a_l^2}\,\partial_k(\Phi_k(x,t)\Psi_k(y,t))|_{k=i\kappa_l}\\ &\quad+\sum_{l=1}^{n}\left(\frac{1}{\kappa_l^2\dot a_l^2}+\frac{i\ddot a_l}{\kappa_l\dot a_l^3}\right)\Phi_l(x,t)\Psi_l(y,t)=\delta(x-y). \end{aligned} \tag{6}$$

We use the notation: $\Phi_l=\Phi_k|_{k=i\kappa_l}$, and the analogous notation for Ψ_l. The scattering data are defined in such a way that (an asterisk denotes complex conjugation):

$$\begin{aligned} &\psi_k=a(k)\phi_k^*+b(k)\phi_k\\ &\dot a_l=\left.\frac{da(k)}{dk}\right|_{k=i\kappa_l}, \quad a(i\kappa_l)=0 \end{aligned} \tag{7}$$

Precisely, we want to solve Eq. (2), in which (1) f is some fixed solution of the KdV equation (1), corresponding to the Jost functions ϕ_k, ψ_k and the scattering data $\{\kappa_l\}_{l=1,\ldots,n}$, $a(k)$, $\dot{a}_l$, $\ddot{a}_l$; (2) w is some given function (real, smooth and vanishing quickly enough at infinity); (3) g is assumed to verify some initial condition, i.e., for $t=0$, g is equal to some given function g_0 (with the same regularity properties as w).

We define the following coefficients:

$$w(k,t)=\lambda(k)(\Phi_k,w)(t), \tag{8}$$

$$w_l^{(1)}(t)=\lambda_l(\Phi_l,w)(t)+\mu_l(\partial_k\Phi_l,w)(t), \tag{9}$$

$$w_l^{(2)}(t)=\mu_l(\Phi_l,w)(t). \tag{10}$$

The scalar product $(\cdot,\cdot)$ is defined by

$$(v_1,v_2)(t)=\int_{-\infty}^{+\infty} v_1(x,t)v_2(x,t)dx$$

and the constants by

$$\lambda(k)=\frac{i}{2\pi}\frac{1}{ka^2(k)}, \tag{11}$$

$$\lambda_l=\frac{1}{\kappa_l^2\dot{a}_l^2}+\frac{i\ddot{a}_l}{\kappa_l\dot{a}_l^3}, \tag{12}$$

$$\mu_l=\frac{1}{i\kappa_l\dot{a}_l^2}. \tag{13}$$

[Notice that the discrete eigenvalues κ_l and the scattering coefficients $a(k)$ do not depend on time.] We use the notation:

$$\partial_k\Phi_l=\partial_k\Phi_k|_{k=i\kappa_l}$$

and its analog for $\partial_k\Psi_l$. Consequently, the rhs w can be expanded in the following way:

$$w(x,t)=\int_{-\infty}^{+\infty} w(k,t)\Psi_k(x,t)dk+\sum_{l=1}^{n}[w_l^{(1)}(t)\Psi_l(x,t)+w_l^{(2)}(t)(\partial_k\Psi_l)(x,t)]. \tag{14}$$

We expand g_0 in the same way as w, with similar notations. Then the solution g of Eq. (2) reads:

$$g=g_1+g_2. \tag{15}$$

g_1 is a solution of the homogeneous equation (5) with the initial data g_0:

$$g_1(x,t)=\int_{-\infty}^{+\infty} g_0(k,t)\Psi_k(x,t)e^{-8ik^3t}dk+\sum_{l=1}^{n}[g_{0_l}^{(1)}(t)\Psi_l(x,t)$$
$$+g_{0_l}^{(2)}(t)[\partial_k\Psi_l(x,t)+24i\kappa_l^2t\Psi_l(x,t)]]e^{-8\kappa_l^3t}. \tag{16}$$

g_2 is a solution of the complete equation (2) with zero initial data:

$$g_2(x,t)=\int_{-\infty}^{+\infty}\Psi_k(x,t)\left[\int_0^t e^{8ik^3t'}w(k,t')dt'\right]e^{-8ik^3t}dk+\sum_{l=1}^{n}\left\{\Psi_l(x,t)\left[\int_0^t e^{8\kappa_l^3t'}w_l^{(1)}(t')dt'\right.\right.$$
$$\left.\left.+24i\kappa_l^2\int_0^t\int_0^{t'}e^{8\kappa_l^3t''}w_l^{(2)}(t'')dt'dt''\right]+\partial_k\Psi_l(x,t)\int_0^t e^{8\kappa_l^3t'}w_l^{(2)}(t')dt'\right\}e^{-8\kappa_l^3t}. \quad (17)$$

The expression of g is thus explicit, but explicit in terms of the squared Jost functions Φ_k, Ψ_k, that depend implicitly on the solution f of KdV.

III. THE SECULAR-PRODUCING TERMS ARE RESONANT

As stated in Sec. I, the solution g of the linearized KdV equation (2) computed previously is not bounded for any rhs w (I mean uniformly bounded in x as $t\to+\infty$). As an example, if the initial data for g is $g_0\equiv 0$, and the rhs $w=\partial_x f$, then

$$g=t\partial_x f. \quad (18)$$

This can be checked by explicit computation, either by transferring the solution (18) to Eq. (2), or by using formulas (8)–(17). This latter computation is lengthy, but it is important to see that it works. A solution like (18) is called *secular*, while it grows linearly with t. The corresponding rhs are called secular producing. They can be characterized as resonant terms, as follows. As written above, the resonance phenomenon occurs when the components of the inverse scattering transform of the source term (the rhs w) evolve in time in the same way as the corresponding components in the transform of the main term, i.e., if they oscillate with the same frequency $(8k^3)$, concerning the radiative part, or decrease with the same rate $(8\kappa_j^3)$, for the discrete spectrum part.

First, we can see that, if some given function w satisfies the condition

$$\partial_t(\Phi_k,w)=-8ik^3(\Phi_k,w) \quad (19)$$

for each k with $\mathrm{Im}\,k\geqslant 0$, then the corresponding solution g of the linearized KdV equation (2), with the rhs w and a zero initial condition is:

$$g=tw. \quad (20)$$

From (19), we see that

$$(\Phi_k w)(t)=(\Phi_k,w)(0)e^{-8ik^3t}, \quad (21)$$

$$(\Phi_j,w)(t)=(\Phi_j,w)(0)e^{-8\kappa_j^3t}, \quad (22)$$

$$(\partial_k\Phi_j,w)(t)=[(\partial_k\Phi_j,w)(0)+24i\kappa_j^2t(\Phi_j,w)(0)]e^{-8\kappa_j^3t}. \quad (23)$$

We compute the solution g of the linearized KdV equation (2) through expressions (15)–(17). We choose a zero initial condition $g_0=0$, thus g_1 is zero. The explicit t-dependency of $w(k,t)$, $w_j^{(1)}(t)$, $w_j^{(2)}(t)$ given by expressions (21)–(23) enables us to compute the integrals in formula (17). After factorization by a t factor, comparison with the expression of w obtained by substituting formulas (21)–(23) into (14) yields Eq. (20). The time evolution law (19) characterizes the secular-producing terms. It is intimately related with the time evolution of the scattering data. Thus we can understand the secular-producing terms as resonant, in the following sense: their components through the inverse scattering transform evolve in time in the same way as does the first-order solution f [see the expression of $\mathscr{A}_1=\frac{1}{2}f$, formula (47) in Sec. IV]. It is not necessary that relation (19) is satisfied for any k with $\mathrm{Im}\,k\geqslant 0$. The following condition is sufficient: (19) is satisfied for any real k and both this relation and its k derivative are satisfied for any $k=i\kappa_j$ of the discrete spectrum. The derivative of relation (19) reads:

$$\partial_t(\partial_k\Phi_j,w)=24i\kappa_j^2t(\Phi_k,w)-8\kappa_j^3(\partial_k\Phi_j,w). \quad (24)$$

An alternative way to describe this feature is the following mathematical property, that also characterizes the secular-producing terms: For a given smooth function w, and for any k with $\mathrm{Im}\, k \geqslant 0$, we define

$$\mathscr{M}_w(k) = \lim_{t \to +\infty} \frac{1}{t} \int_0^t e^{8ik^3 t'} (\Phi_k, w) dt'. \tag{25}$$

Then w is not secular producing if $\mathscr{M}_w(k) = 0$ for each k. Before we give some proof of this feature, we can show that the part of w corresponding to a nonzero $\mathscr{M}_w(k)$ is secular producing. Let us call this part w_m, and define it by

$$w_m(x,t) = \int_{-\infty}^{+\infty} w_m(k,t) \Psi_k(x,t) + \sum_{j=1}^{n} \left(w_{m,j}^{(1)}(t) \Psi_j(x,t) + w_{m,j}^{(2)}(t) \partial_k \Psi_j(x,t) \right) \tag{26}$$

with

$$w_m(k,t) = \lambda(k) \mathscr{M}_w(k) e^{-8ik^3 t}, \quad k \in \mathbb{R}, \tag{27}$$

$$w_{m,j}^{(1)}(t) = \left(\lambda_j \mathscr{M}_w(i\kappa_j) + \mu_j \left[\frac{d\mathscr{M}_w}{dk}(i\kappa_j) + 24 i \kappa_j^2 t \mathscr{M}_w(i\kappa_j) \right] \right) e^{-8\kappa_j^3 t}, \tag{28}$$

$$w_{m,j}^{(2)}(t) = \mu_j \mathscr{M}_w(i\kappa_j) e^{-8\kappa_j^3 t}. \tag{29}$$

If we compare the t dependency of the components of w_m, given by Eqs. (27)–(29), to those of the function w in Eqs. (21)–(23), we see that they evolve in the same way, and conclude that w_m satisfies condition (19), and is thus secular producing. Reciprocally, if w itself satisfies Eq. (19), the t dependency of (Φ_k, w) is known, thus $\mathscr{M}_w(k)$ can be computed explicitly, and it is easily seen that $w = w_m$. Furthermore, we show below that if $\mathscr{M}_w(k) = 0$ for each k (with the necessary assumptions on uniform limits), that is, when w_m is zero, then the solution g of the linearized KdV equation is small with regard to t as t is large, i.e., w is not secular producing. It then appears clearly that the secular-producing part of w is w_m.

Let us now prove that the condition $\mathscr{M}_w(k) = 0$ implies that g is not secular producing. For this, we make different reasonings for the radiation part and for the discrete spectrum part of g. Let us first consider the radiation part. We assume that

$$w(x,t) = \int_{-\infty}^{+\infty} w(k,t) \Psi_k(x,t) dk \tag{30}$$

and we write:

$$m(k,t) = \frac{1}{t} \int_0^t (\Phi_k, w)(t') e^{8ik^3 t'} dt'. \tag{31}$$

Thus

$$\mathscr{M}_w(k) = \lim_{t \to +\infty} m(k,t) \tag{32}$$

and the condition $\mathscr{M}_w(k) = 0$ can be written:

$$m(k,t) \xrightarrow[t \to +\infty]{} 0. \tag{33}$$

Then, using formula (17) to solve Eq. (2), we find that

$$g(x,t) = \int_{-\infty}^{+\infty} \lambda(k) e^{-8ik^3 t} \Psi_k(x,t) t m(k,t) dk. \tag{34}$$

Since k is real, Ψ_k bounded with time, if we assume that the limit (33) is reached uniformly [or at least in $\mathscr{L}_1(k)$] then:

$$g(x,t) \underset{t\to+\infty}{=} o(t). \tag{35}$$

This shows that g is not secular, for the part that comes from the continuous spectrum, if $\mathscr{M}_w(k)=0$ for each real k.

Let us prove the analogous result for the discrete spectrum. Consider first the evolution of the normalization coefficients of the Jost functions φ_j, ψ_j, for the discrete eigenvalues κ_j. We have

$$\int_{-\infty}^{+\infty}[\varphi_j(x,t)]^2dx=\frac{i\dot{a}_j(t)}{c_j(t)}=\frac{i\dot{a}_j(0)}{c_j(0)}\,e^{-8\kappa_j^3t} \tag{36}$$

and

$$\int_{-\infty}^{+\infty}[\psi_j(x,t)]^2dx=i\dot{a}_j(t)c_j(t)=i\dot{a}_j(0)c_j(0)e^{+8\kappa_j^3t}. \tag{37}$$

Thus

$$\Phi_j=\Phi_j^0e^{-8\kappa_j^3t}$$

and

$$\Psi_j=\Psi_j^0e^{+8\kappa_j^3t}, \tag{38}$$

where the functions Φ_j^0 and Ψ_j^0 are bounded with regard to t. The hypothesis reads precisely as follows:

$$\mathscr{M}_w(i\kappa_j)=0$$

and

$$\frac{d\mathscr{M}_w}{dk}(i\kappa_j)=0 \tag{39}$$

for each j with $j=i\kappa_j$ in the discrete spectrum. We define:

$$m_j^{(1)}(t)=\frac{1}{t}\int_0^t[(\partial_k\Phi_j,w)(t')-24i\kappa_j^2t'(\Phi_j,w)(t')]e^{8\kappa_j^3t'}dt', \tag{40}$$

$$m_j^{(2)}(t)=\frac{1}{t}\int_0^t(\Phi_j,w)(t')e^{8\kappa_j^3t'}dt'. \tag{41}$$

Thus condition (39) can be written:

$$m_j^{(1)}(t),m_j^{(2)}(t)\xrightarrow[t\to+\infty]{}0. \tag{42}$$

The nonradiative part g_d of the solution g of Eq. (2) is given by the discrete sum in Eq. (17). The third integral in this sum is straightforward evaluated as $t\mu_lm_l^{(2)}(t)$, while the first one reads:

$$\int_0^te^{8\kappa_l^3t'}w_l^{(1)}(t')dt'=t\mu_lm_l^{(1)}(t)+t\lambda_lm_l^{(2)}(t)+\mu_l\int_0^t24i\kappa_l^2t'(\Phi_l,w)(t')e^{8\kappa_l^3t'}dt'. \tag{43}$$

An integration by part of the latter integral yields a term integrated two times, that cancels the second integral in the expression of g_d. Relation (38) yields the following expression of $\partial_k\Psi_j$:

$$\partial_k \Psi_j = (\partial_k \Psi_j^0 - 24 i \kappa_j 2^t \Psi_j^0) e^{8\kappa_j^3 t}. \tag{44}$$

Making use of (38) and (44) in the expression of g_d yields

$$g_d(x,t) = t \sum_{j=1}^{n} \{\Psi_j^0(x,t)[\mu_j m_j^{(1)}(t) + \lambda_j m_j^{(2)}(t)] + \partial_k \Psi_j^0(x,t) \mu_j m_j^{(2)}(t)\}. \tag{45}$$

Because Ψ_j^0 and $\partial_k \Psi_j^0$ are bounded and $m_j^{(1)}(t)$, $m_j^{(2)}(t)$ tend to zero as $t \to +\infty$, we see that

$$g_d(x,t) \underset{t \to +\infty}{=} o(t). \tag{46}$$

Thus g_d is not secular, that completes the proof.

IV. LINEAR TERMS AND CONSERVED DENSITIES

A. Relation between the conserved densities and the secular-producing terms

A completely integrable equation has the very important property to present an infinite number of conserved densities. These are algebraically independent functionals $\mathscr{A}_j$ of the solution, whose integral over all space is constant in time. In the case of the KdV equation, various representations of these densities are given in many books. Reference 1 gives an explicit expression in the form of an integral and linear combination of the squares of the Jost functions. It reads as follows:

$$\mathscr{A}_j(x,t) = 2(-1)^j \sum_{l=1}^{n} \kappa_l^{2j-1} C_l(t) \psi_l^2(x,t) + \frac{i}{\pi} \int_{-\infty}^{+\infty} k^{2j-1} r(k,t) \psi_k^2(x,t) dk. \tag{47}$$

C_l and $r(kt)$ are, respectively, the normalization coefficient for bounded states and the reflection coefficient, and characterize the scattering data: $r(k,t) = -\overline{b}(k,t)/a(k)$, $C_l(t) = -i/\dot{a}_l c_l(t)$. With the notations of Ref. 6, they are, respectively, $R_-(k,t)$ and $D_{-,j}(t)$. Thus

$$\partial_{x'} \mathscr{A}_j(x,t) = \sum_{l=1}^{n} \lambda_{j,l} e^{-8\kappa_l^3 t} \Psi_l(x,t) + \int_{-\infty}^{+\infty} \lambda_j(k) e^{-8ik^3 t} \Psi_k(x,t) dk \tag{48}$$

with:

$$\lambda_{j,l} = 2(-1)^j \kappa_l^{2j-1} C_l(0), \quad \lambda_j(k) = \frac{i}{\pi} k^{2j-1} r(k,0). \tag{49}$$

We see that $\partial_{x'} \mathscr{A}_j$ is a linear combination and integral of the elementary solutions of the homogeneous linearized KdV equation (5), thus it is itself a solution of this equation. This can also be proved by mathematical induction and direct computation, starting from the recurrence formula for the $\mathscr{A}_j$ (given, e.g., in Ref. 1):

$$\partial_{x'} \mathscr{A}_{j+1} = \frac{-1}{4} (\partial_{x'}^3 \mathscr{A}_j - 4f \partial_{x'} \mathscr{A}_j - 2(\partial_x f) \mathscr{A}_j), \quad \mathscr{A}_0 = 1. \tag{50}$$

Thus, the derivatives of the conserved densities are secular-producing terms.

Reciprocally, a secular-producing term is a solution g of the homogeneous linearized KdV equation (5). Let us prove that $U = \int_{-\infty}^{x} g(x',t) dx'$ is a conserved density. Using (5), we can write:

$$\partial_t \int_{-\infty}^{+\infty} U dx = 6 \int_{-\infty}^{+\infty} f g dx - \int_{-\infty}^{+\infty} \partial_x^2 g dx. \tag{51}$$

Assuming that g vanishes at infinity with its derivatives, the second integral is zero. The first one is a constant. Indeed, using Eqs. (1) and (5):

$$\partial_t(fg)=6\partial_x(f^2g)-\partial_x(f(\partial_x^2g)+(\partial_x^2f\ )g-(\partial_x^2f\ )(\partial_x^2g)). \tag{52}$$

Thus the integral over all space is 0 and

$$\partial_t^2\int_{-\infty}^{+\infty}Udx=0. \tag{53}$$

Thus $\int_{-\infty}^{+\infty}Udx=pt+q$, where p and q are constants. If U is a polynomial in f and its derivatives, U is bounded with time, and hence $p=0$. Finally, $\int_{-\infty}^{+\infty}Udx=q$ is a constant. Together with the above assumption, that U is a polynomial in f, $\partial_x f,\ldots$, this means that U is a conserved density of the KdV equation.

Thus, if we consider only the terms that express explicitly as functions of f and its derivatives, the secular-producing terms are the derivatives of the conserved densities.

B. The KdV hierarchy and the linear terms

Kodama and Taniuti stated that the secular-producing terms are the linear ones.[1] These authors checked this assertion by direct computation of the first few orders, in the particular case of the one soliton solution. But, to my knowledge, no general proof of this feature can be found. Despite this, the multiple time formalism by Kraenkel *et al.* rests on this statement. Let us recall this procedure: Starting from some system of PDE's describing some physical situation, an adequate multiscale expansion leads to the KdV equation (1) at main order $f=f_1$, and to linearized KdV equations (2) for the following orders $g=f_2,f_3,f_4,\ldots$, with rhs. $w=w_2,w_3,w_4,\ldots$, polynomial in f and its derivatives. The rhs w_j also depends on the already computed higher order corrections $f_2,\ldots,f_{j-1}$, and on their derivatives. The secular-producing terms in w_j must be removed, so that the solution g_j is bounded. Therefore, higher-order time variables $t_2,t_3,\ldots$, are introduced, so that, for a given $j\geqslant 2$, the rhs w_j contains a term proportional to $\partial_{t_j}f$. The t_j dependence of f is not yet determined at this order. It is shown that the equation governing this dependence must be the jth equation of the KdV hierarchy, which reads[7]

$$\partial_{\tau_j}f=\partial_x\mathscr{L}^jf. \tag{54}$$

The variable τ_j is proportional to t_j, and the recurrence operator $\mathscr{L}$ is defined by

$$\mathscr{L}=-\frac{1}{4}\partial_x^2+f-\frac{1}{2}\int^x dx(\partial_x f\ ). \tag{55}$$

The only adjustable parameter is the constant α_j such as $\tau_j=\alpha_j t_j$. It is determined by the requirement that the linear term in the equation of the KdV hierarchy (54), $\alpha_j(-1/4)^j\partial_x^{2j+1}f$, cancels the linear term in w_j, that is also proportional to $\partial_x^{2j+1}f$, due to the homogeneity properties of the expansion. If the secular-producing terms are the linear ones, they are removed by this procedure.

But we have shown that the secular-producing terms are the x derivatives of the conserved densities. We prove that the previous procedure not only removes their linear part, but these terms totally. Indeed, integration by parts in expression (50), and immediate mathematical induction show that:

$$\mathscr{A}_j=\tfrac{1}{2}\mathscr{L}^{j-1}f \tag{56}$$

so that the equations of the KdV hierarchy (50) can be written:

$$\partial_{\tau_j}f=2\partial_x\mathscr{A}_{j+1}. \tag{57}$$

Thus the requirement that the linear terms are removed by using the KdV hierarchy, removes all terms of the $\partial_x\mathscr{A}_j$, which are exactly the secular-producing terms. Thus, the removal of the linear terms in the rhs w_j gives a correct result, even if the nonlinear part of the conserved densities is secular-producing.

C. Are the solutions of the linearized KdV equation secular producing?

This question is important because, as soon as we consider the third nonzero term in the perturbative expansion, the rhs w depends on the second term, which is some solution of the linearized KdV equation (2). The secular-producing terms coming from the leading order solution f have been studied in the previous section, the present section is devoted to the study of the effects of the terms of higher order.

Consider the solution $g=g_1+g_2$ for linearized KdV equation (2) described by formulas (15)–(17). The part g_1 of g that comes from the initial data is a solution of the homogeneous linearized KdV equation, and thus is secular producing. We will show that the part g_2 that comes from the rhs is not. To prove this, we shall compute the quantities $\mathscr{M}_{g_2}(k)=\lim_{t\to+\infty} m_{g_2}(k,t)$ and show that they are zero. In order to compute the quantities

$$m_{g_2}(k,t)=\frac{1}{t}\int_0^t e^{8ik^3t'}(\Phi_k,g_2)(t')dt' \tag{58}$$

we need the following formulas, that give the scalar products of the functions Φ_k, Ψ_k:

$$(\Phi_k,\Psi_{k'})=-2i\pi k a^2(k)\delta(k-k') \quad (k,k'\in\mathbb{R}), \tag{59}$$

$$(\Phi_l,\partial_k\Psi_m)=(\partial_k\Phi_l,\Psi_m)=i\kappa_l\dot{a}_l^2\delta_{l,m} \quad (l,m \text{ integer}), \tag{60}$$

$$(\Phi_l,\Psi_m)=(\Phi_k,\Psi_m)=0 \quad (k\in\mathbb{R},l,m \text{ integer}), \tag{61}$$

$$(\partial_k\Phi_l,\partial_k\Psi_m)=(\dot{a}_l^2+i\kappa_l\dot{a}_l\ddot{a}_l)\delta_{l,m} \quad (l,m \text{ integer}). \tag{62}$$

(These formulas are given in Ref. 1.) We find for a real k:

$$m_{g_2}(k,t)=\frac{1}{t}\int_0^t t' m_w(k,t')dt', \tag{63}$$

$m_w(k,t)$ is defined by Eq. (31), as $m_{g_2}(k,t)$, but for the rhs w of Eq. (2) instead of its solution g_2. The hypothesis is that w is not secular producing, thus:

$$m_w(k,t)\xrightarrow[t\to+\infty]{}0.$$

We see that

$$\mathscr{M}_{g_2}(k)=\lim_{t\to+\infty} m_{g_2}(k,t)=0. \tag{64}$$

For the discrete eigenvalues, we find

$$(\Phi_j,g_2)(t)=e^{-8\kappa_j^3 t}i\kappa_j\dot{a}_j^2\int_0^t e^{8\kappa_j^3 t'}w_j^{(2)}(t')dt'. \tag{65}$$

Thus, as previously,

$$m_{g_2}(i\kappa_j,t)=\frac{1}{t}\int_0^t t' m_w(i\kappa_j,t')dt'. \tag{66}$$

Thus $\mathscr{M}_{g_2}(i\kappa_j)=0$. It is also necessary to check that

$$\frac{d\mathscr{M}_{g_2}}{dk}(i\kappa_j)=0.$$

Taking the k derivative of Eq. (58), we obtain

$$\partial_k m_{g_2}(i\kappa_j,t)=\frac{1}{t}\int_0^t[(\partial_k\Phi_j,g_2)(t')-24i\kappa_j^2t'(\Phi_j,g_2)(t')]e^{8\kappa_j^3t'}dt'. \tag{67}$$

We obtain directly

$$(\Phi_j,g_2)(t)=e^{-8\kappa_j^3t}tm_w(i\kappa_j,t) \tag{68}$$

and, after some algebra

$$(\partial_k\Phi_j,g_2)(t)e^{8\kappa_j^3t}=\int_0^t e^{8\kappa_j^3t'}(\partial_k\Phi_j,w)(t')dt'+24i\kappa_j^2\int_0^t dt'\int_0^{t'}dt''e^{8\kappa_j^3t''}(\Phi_j,w)(t''). \tag{69}$$

Equation (67) reduces to

$$\partial_k m_{g_2}(i\kappa_j,t)=\frac{1}{t}\int_0^t dt'\left\{\int_0^{t'}e^{8\kappa_j^3t''}(\partial_k\Phi_j,w)(t'')dt''+24i\kappa_j^2 \times\left[\int_0^{t'}t''m_w(i\kappa_j,t'')dt''-t'^2m_w(i\kappa_j,t')\right]\right\}. \tag{70}$$

Making use of the expression of $\partial_k m_w(i\kappa_j,t)$, analogous to (67), and after an integration by parts, the expression of $\partial_k m_{g_2}(i\kappa_j,t)$, simplifies to

$$\partial_k m_{g_2}(i\kappa_j,t)=\frac{1}{t}\int_0^t t'\partial_k m_w(i\kappa_j,t')dt'. \tag{71}$$

While $\partial_k m_w(i\kappa_j,t)\to_{t\to+\infty}0$, we have

$$\partial_k m_{g_2}(i\kappa_j,t)\xrightarrow[t\to+\infty]{}0. \tag{72}$$

This completes the proof: g_2 is not secular producing.

Thus we have proved that the part of the solution of the linearized KdV equation that comes from the initial data is secular-producing, and the part that comes from the rhs (and would correspond to a zero initial data) is not.

V. CONCLUSION

The higher order terms in the perturbative expansion that describes KdV solitons obey the linearized KdV equation, with some rhs that expresses as a polynomial of the KdV solution. This linearized equation is solved by the inverse scattering transform method, in a quasiexplicit way. Unbounded or secular solutions exist for some values of the rhs. The terms on the rhs that produce these secular solutions have been characterized in several ways: First, a mathematical characterization, that does not take into account the dependency of the rhs on the leading order f, has been given. Second, the secular-producing terms have been interpreted as conserved densities of the KdV equation. Third, it has been checked that the removal of all terms linear in f, in some rhs w polynomial in f, ensures the boundness of the solution g of the linearized equation. Fourth, it has been determined which part of this solution g itself, considered as rhs, is secular producing, and which part is not.

All the conditions derived here are useful in the study of the higher order terms of the KdV-type multiscale expansions. They allow us to show the boundness of all terms in such an expansion. Boundness conditions for the higher order terms also fix the evolution of the main order term. In particular, we have justified an essential point of the multiple time expansions of Kraenkel *et al.*: The procedure that removes the linear terms in the rhs of the linearized KdV equation ensures the boundness of the solution.

ACKNOWLEDGMENTS

I express my gratitude to Dr. Isabelle Gruais (I.R.M.A.R., Université de Rennes, France), for helpful collaboration, and Professor Guy Métivier, for fruitful scientific discussions, and as Director of the P.D.E. lab. of the Institut de Recherche Mathématique de Rennes (I.R.M.A.R.), where I have begun the present study. I am also indebted to my former thesis director Dr. Miguel Manna (Laboratoire de Physique Mathématique, Université de Montpellier II, France), whose works on the KdV hierarchy have considerably influenced the present paper, and who gave me most of the bibliography.

[1] Y. Kodama and T. Taniuti, ''Higher order approximation in the reductive perturbation method. I. The weakly dispersive system,'' J. Phys. Soc. Jpn. **45**, 298 (1978).
[2] R. A. Kraenkel, M. A. Manna, and J. G. Pereira, ''The Korteweg-de Vries hierarchy and long water-waves,'' J. Math. Phys. **36**, 307 (1995).
[3] R. A. Kraenkel, M. A. Manna, and J. G. Pereira, ''The reductive perturbation method and the Korteweg–de Vries hierarchy,'' Acta Appl. Math. **39**, 389 (1995).
[4] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, ''Method for solving the Korteweg-de Vries equation,'' Phys. Rev. Lett. **19**, 1095 (1967).
[5] P. D. Lax, ''Integrals of nonlinear equations of evolution and solitary waves,'' Commun. Pure Appl. Math. **21**, 467 (1968).
[6] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations* (Academic, London, 1982).
[7] H. Flaschka, A. C. Newell, and M. Tabor, ''Integrability,'' in *What is Integrability?* edited by V. E. Zakharov (Springer, Berlin, 1991).

Hard chaos in magnetic billiards (on the hyperbolic plane)

Tamás Tasnádi
Research Group for Statistical Physics of the Hungarian Academy of Sciences, Department of Physics of Complex Systems, Eötvös University, 6-8. Múzeum krt., Budapest, 1088 Hungary

(Received 20 October 1997; accepted for publication 12 March 1998)

In this paper some results on the local and global stability analysis of magnetic billiard systems, established on two dimensional Riemannian manifolds of constant curvature are presented, with particular emphasis on the hyperbolic plane. For special billiards, possessing a discrete group of (rotational or translational) symmetry, a geometrical theorem, illustrated by numerical simulations, is given on the stability of trajectories with the same symmetry. We also present sufficient criteria for the global hyperbolicity of the dynamics (hard chaos), and give lower estimations for the Lyapunov exponent in terms of the shape of the billiard. © *1998 American Institute of Physics.* [S0022-2488(98)02606-1]

I. INTRODUCTION AND SUMMARY

This paper is devoted to a detailed local and global stability analysis of the classical dynamics of magnetic billiard systems established on two-dimensional Riemannian manifolds of constant curvature. In these systems a pointlike particle moves at unit speed along curves of constant curvature $\beta \in \mathbb{R}$ in a bounded region Q (with a piecewise two times differentiable boundary) of the given two dimensional manifold, and performs elastic reflections at the boundary of the domain Q. This motion can be interpreted as the frictionless motion of a billiard ball of unit charge and mass on the (curved) billiard table Q affected by a homogeneous magnetic field of magnitude β perpendicular to the billiard surface.

Though this model is somewhat artificial, it covers plenty of important and intensively investigated dynamical systems, as limit cases. If the magnetic field and the curvature of the manifold are zero, ordinary billiard systems are obtained, which are fundamental models in ergodic theory[1–4] and have still great importance in studies of quantum chaos.[5] Since the geodesic motion on manifolds of negative curvature is hyperbolic,[6–8] it is also worth investigating billiard systems on curved manifolds in classical[9] or quantum chaos theory.[10,5,11]

The presence of magnetic field changes drastically the dynamics of the billiard system; it breaks time reversal symmetry, diminishes or completely suppresses hyperbolicity in the bounceless segments, and strongly influences the effects of the bounces.[12–14] Previously there has been work done on classical[15–17] and quantum billiards[18] with a magnetic field, even on curved spaces.[19] Recently the study of planar magnetic billiards gains considerable inspiration from mesoscopic physics, since under certain conditions the motion of electrons in two dimensional mesoscopic systems can be approximated with high accuracy by classical or semiclassical methods, and the transport properties of these systems can also be investigated experimentally in the presence of a magnetic field as well.[20]

These are the main motivations of the studies presented in this work. In the rest of this section the main results are summarized along with the organization of the paper.

In the next two sections (in opposition to the main title) manifolds of curvature $+1$ (sphere), 0 (Euclidean plane), and -1 (hyperbolic plane) are treated simultaneously, since parallel results can be obtained by similar arguments in all three cases. In Section II the former results of Ref. 12 concerning the stability of the dynamics in the bounceless segments and at the bounces are briefly reviewed (for the sake of completeness), and they are adapted to our present necessities. Some basic notions and notations are also introduced, which will be used throughout the whole article.

In Section III we deal with the study of the joint stability (Jacobian) matrix of a bounceless flight and the succeeding bounce, which has also been investigated in Refs. 13, 15, 16, 17 in the Euclidean case. In general this linear mapping depends on several parameters (the flying time, the

 © 1998 American Institute of Physics

magnetic field, the angle of incidence, and the curvature of the billiard wall at the bouncing point), nevertheless, a simple geometric rule can be given for determining the stability character of the map in hand. This criterion is derived through a series of lemmas. Some of the subsidiary calculations are put in Appendices A and B.

From the fourth section on, our attention is restricted solely to the hyperbolic plane. In Section IV the result of the preceding section is applied to symmetric periodic orbits in special billiards with discrete symmetry. Numerical investigations are also presented for a regular triangular billiard, showing that the periodic orbit in question loses its stability via period doubling bifurcation, as the magnetic field decreases.

Besides Section III, Section V contains the main results of the paper. In this section we give essentially two types of geometric criteria for global hyperbolicity in weak magnetic fields: one of them applies to Sinai-like[1] (magnetic) billiards, i.e., billiards with a piecewise smooth, everywhere (sufficiently) dispersing boundary and the other concerns Bunimovich-like ones,[2,3,21,22] i.e., billiards with a piecewise smooth boundary containing also (strongly) focusing arcs lying sufficiently far from each other. Our method also enables us to give an efficient lower estimation for the Lyapunov exponent in certain cases; this is demonstrated in Appendix C, for a Bunimovich-like (magnetic) billiard on the hyperbolic plane.

Our main findings are summarized again in the Conclusions.

II. FORMER RESULTS, NOTATIONS

In this section the basic formulas indispensable in stability analysis are shortly reviewed. Further details and the derivation of these relations in a more general context (considering an inhomogeneous magnetic field and nonconstant curvature) can be found in Ref. 12. The notations introduced here will be used throughout the whole paper without explicit reference.

Let $(M,\langle\cdot,\cdot\rangle)$ be a two dimensional Riemannian manifold of constant (Gaussian) curvature $r\in\{-1,0,1\}$ with scalar product $\langle\cdot,\cdot\rangle$, volume form Ω, and Levi-Civita connection ∇. Suppose that the billiard Q is defined on M and let $\gamma_0:\mathbb{R}\to M$ denote a trajectory of unit speed, i.e., a continuous curve parametrized by an arc length which satisfies the equation of motion $\nabla_{\dot\gamma_0}\dot\gamma_0 = \beta\dot\gamma_0$ during the bounceless segments, and obeys the law of elastic reflection at the bounces. The stability of this individual trajectory can be analyzed with the help of *infinitesimal variations* or *Jacobi fields*[23,12] which are piecewise smooth vector fields along γ_0. They are defined as the (partial) derivative of a *variation* $\delta(t,\epsilon)$ of the unvaried trajectory γ_0 with respect to the variational parameter ϵ at $\epsilon=0$, where the *variation* $\delta:\mathbb{R}\times(-\epsilon_0,\epsilon_0)\to M$ is a piecewise smooth (more precisely everywhere C^∞ except at the reflections, where it is C^0) mapping for which $\delta(t,0) = \gamma_0$ holds and the curves $t\mapsto\delta(t,\epsilon)$ are possible trajectories for all $\epsilon\in(-\epsilon_0,\epsilon_0)$.

This latter condition on δ results, that the Jacobi fields (along γ_0) can be equivalently characterized as solutions of a certain second order ordinary linear differential equation during the bounceless segments, which (and the first derivative of which) have certain jumps at the reflection points depending only on geometric data.[12,13] This means that the Jacobi fields are completely determined by four scalar data, their value, and their first derivative at an arbitrary point. (In fact, they constitute a four dimensional vector space.)

Using the orthonormal coordinate frame $\{\dot\gamma_0,\Omega(\dot\gamma_0)\}$ along the trajectory, the Jacobi field $V(t)=a(t)\dot\gamma_0+b(t)\Omega(\dot\gamma_0)$ can be split into a *longitudinal* and a *transverse* component, where a is the *longitudinal*, b is the *transverse coordinate*. Restricting our attention to variations of unit speed (*special variations*) and considering only the transverse part of the infinitesimal variation, the number of relevant dimensions can be reduced from four to two. In terms of the transverse coordinate b and its derivative b' the effect of the reflections and also the evolution during the bounceless segments can be formulated with the help of a linear transformation $\mathbf{L}$, i.e.,

$$\begin{pmatrix} b_+ \\ b'_+ \end{pmatrix} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} b_- \\ b'_- \end{pmatrix},$$

where the subscripts ''+'' and ''−'' refer to the values of the coordinates after and before the transformation, respectively. Further we give the concrete form of the transformation $\mathbf{L}$ for different cases.

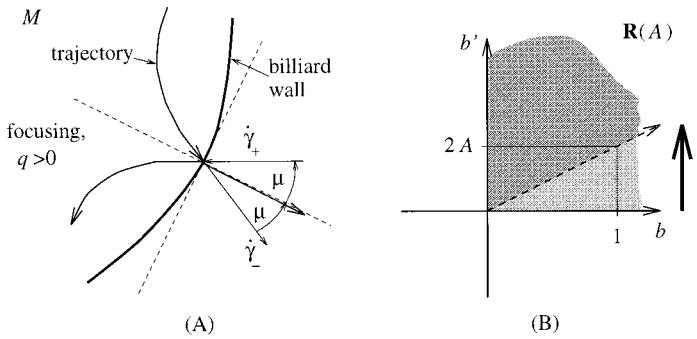

FIG. 1. The effect of the bounce.

A. The bounce

In this case the linear transformation denoted by $-\mathbf{R}(A)$ has the following matrix form:

$$-\mathbf{R}(A) := -\begin{pmatrix} 1 & 0 \\ 2A & 1 \end{pmatrix}, \quad A = -\frac{q+\beta\,\sin(\mu)}{\cos(\mu)}, \tag{1}$$

where q denotes the curvature of the billiard wall at the reflection point, β is the magnetic field, and μ is the angle of incidence. (The minus sign before the matrix, being irrelevant in most of the later calculations, will usually be omitted.) The sign conventions can be seen in figure 1/A; $q>0$ if the wall is focusing, and μ is measured from the incident velocity vector to the outward pointing normal vector. Figure 1/B shows that $\mathbf{R}(A)$ is a shearing in the (b,b') plane along the b' axis with parameter value $2A$.

(It should be noted, that in figure 1/A $\beta>0$, $q>0$ and $0<\mu<\pi/2$, i.e., $A<0$ while in figure 1/B $A>0$, so these figures do not correspond to the same situation.)

It has been derived in Ref. 12 that the transverse coordinate of the (special) Jacobi fields obeys the second order differential equation $b''=-(r+\beta^2)b$ during the bounceless segments, where r is the (constant) curvature of the manifold M and β is the (homogeneous) magnetic field. For describing the bounceless evolution of the Jacobi fields three different cases have to be considered separately according to the sign of $r+\beta^2$.

B. The elliptic case ($r+\beta^2>0$)

If $r=1$ or $r=0$ and $\beta\neq 0$ or $r=-1$ and $|\beta|>1$ then the bounceless segments are arcs of circles of radius R, and

$$\beta = \begin{cases} \cot(R), & \text{if } r=1 \text{ (on the sphere);} \\ 1/R, & \text{if } r=0 \text{ (on the Euclidean plane);} \\ \coth(R), & \text{if } r=-1 \text{ (on the hyperbolic plane).} \end{cases} \tag{2}$$

The evolution of the coordinates b, b' is given by the linear transformation $\mathbf{E}(t)$:

$$\mathbf{E}(t) := \begin{pmatrix} \cos(\chi t) & \dfrac{1}{\chi}\sin(\chi t) \\ -\chi\,\sin(\chi t) & \cos(\chi t) \end{pmatrix}, \quad \chi = \begin{cases} \sqrt{\beta^2+1}, & \text{if } r=1; \\ \beta, & \text{if } r=0; \\ \sqrt{\beta^2-1}, & \text{if } r=-1, \quad |\beta|<1. \end{cases} \tag{3}$$

The quantity χt appearing in the matrix has a geometric meaning; it is the angle at the center belonging to the arc of flight (figure 2/A). The matrix $\mathbf{E}(t)$ performs an (elliptic) rotation in the (b,b') plane (figure 2/B).

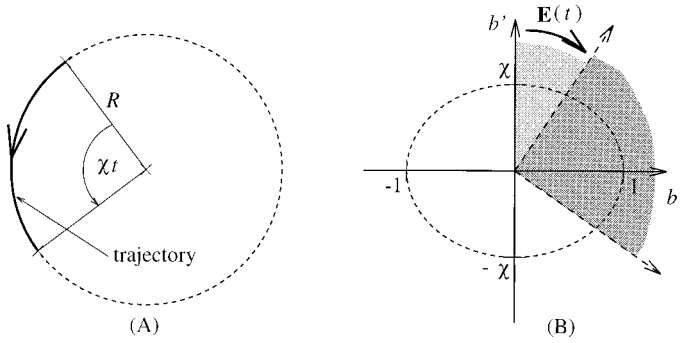

FIG. 2. The elliptic case.

C. The parabolic case ($r+\beta^2=0$)

It arises for $r=0$, $\beta=0$ or for $r=-1$, $|\beta|=1$. In the first case the particle moves along straight line segments on the Euclidean plane, while in the second case it follows segments of paracycles on the hyperbolic plane. In both cases the bounceless evolution of (b,b') is given by the matrix $\mathbf{P}(t)$:

$$\mathbf{P}(t):=\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}. \tag{4}$$

This transformation is a shearing in the (b,b') plane along the b axis with parameter value t, as shown in figure 3.

D. The hyperbolic case ($r+\beta^2<0$)

If $r=-1$ and $|\beta|<1$ then the bounceless trajectories are segments of hypercycles of altitude d (or specially geodesics, if $\beta=0$), and

$$\beta=\tanh(d). \tag{5}$$

In this case the evolution of the Jacobi field is described by the transformation $\mathbf{H}(t)$:

$$\mathbf{H}(t):=\begin{pmatrix} \cosh(\chi t) & \dfrac{1}{x}\sinh(\chi t) \\ \chi\sinh(\chi t) & \cosh(\chi t) \end{pmatrix}, \quad \chi=\sqrt{1-\beta^2}. \tag{6}$$

The quantity χt appearing in the matrix has again a geometric meaning; it is the length of the geodesic segment on the base line of the hypercycle which belongs to the arc of the bounceless

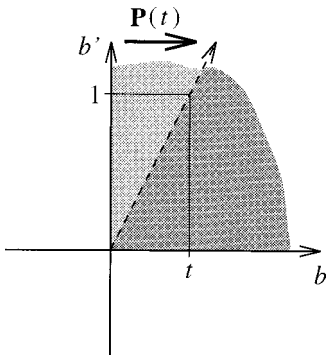

FIG. 3. The parabolic case.

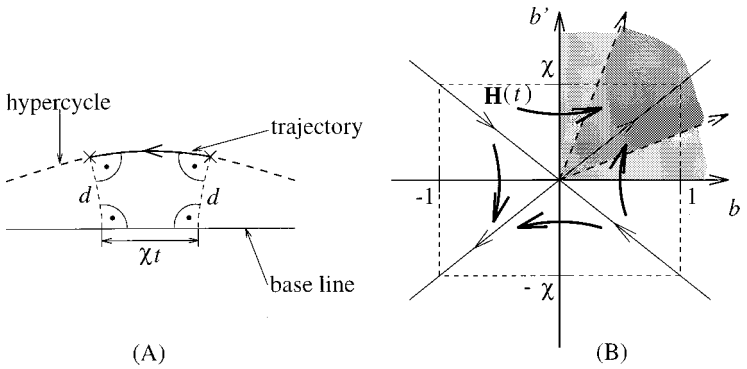

FIG. 4. The hyperbolic case.

flight, as shown in figure 4/A. The effect of the matrix $\mathbf{H}(t)$ is a hyperbolic transformation in the (b,b') plane (figure 4/B).

It is worth noticing that the transformations $\mathbf{R}(A)$, $\mathbf{E}(t)$, $\mathbf{P}(t)$, and $\mathbf{H}(t)$ appearing in the former paragraphs are all one parameter subgroups of the three dimensional Lie group $Sl(2;\mathbb{R})$ consisting of two by two real matrices of unit determinant. (The fact that these transformations are area preserving ones is due to the Hamiltonian nature of the original billiard problem.) The nontrivial elements (unequal to $\pm$identity) of this group can be divided into three disjoint classes; the *elliptic matrices* having trace strictly between -2 and 2 [like $\mathbf{E}(t)$, figure 2/B] perform rotations, and possess a pair of complex conjugate eigenvalues with no real eigenvectors; the *parabolic matrices* having trace ± 2 [like $\mathbf{R}(A)$ or $\mathbf{P}(t)$, figures 1/B, 3] perform shearings, and the 1 is a multiple eigenvalue of them with a single real eigenvector; while the *hyperbolic matrices* having trace strictly greater than 2 or less than -2 [like $\mathbf{H}(t)$, figure 4/B] describe a hyperbolic transformation with two real eigenvectors. Since every power of a matrix remains in the class of the original matrix, this classification carries over to the Lie algebra $sl(2;\mathbb{R})$ (consisting of matrices of zero trace) of the group in question.

The one parameter subgroups mentioned previously are generated by the following Lie algebra elements:

$$\mathbf{r}:=\left.\frac{d}{d(2A)}\right|_0 \mathbf{R}(A)=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},\quad \mathbf{e}:=\left.\frac{d}{dt}\right|_0 \mathbf{E}(t)=\begin{pmatrix}0 & 1\\ -\chi^2 & 0\end{pmatrix},$$

$$\mathbf{p}:=\left.\frac{d}{dt}\right|_0 \mathbf{P}(t)=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \mathbf{h}:=\left.\frac{d}{dt}\right|_0 \mathbf{H}(t)=\begin{pmatrix}0 & 1\\ \chi^2 & 0\end{pmatrix}. \tag{7}$$

Since

$$[\mathbf{r},\mathbf{e}]=[\mathbf{r},\mathbf{p}]=[\mathbf{r},\mathbf{h}]=\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}, \tag{8}$$

it can be seen, that $\mathbf{r}$ with either $\mathbf{e}$ or $\mathbf{p}$ or $\mathbf{h}$ generates the whole Lie algebra $sl(2;\mathbb{R})$, so in stability analysis for every fixed value of r and β the whole three dimensional Lie group $Sl(2;\mathbb{R})$ must be treated.

III. A GEOMETRICAL THEOREM ON LOCAL STABILITY

For determining the stability of an individual trajectory of the billiard system in the long time limit, an infinite product $\ldots\mathbf{R}_i\mathbf{L}_i\ldots\mathbf{R}_2\mathbf{L}_2\mathbf{R}_1\mathbf{L}_1$ of noncommuting matrices has to be investigated, where the matrices $\mathbf{R}_i$ correspond to the subsequent reflections and the transformations $\mathbf{L}_i$ belong to the bounceless segments. These latter ones equal either the elliptic ($\mathbf{E}(t_i)$), parabolic ($\mathbf{P}(t_i)$), or hyperbolic matrices ($\mathbf{H}(t_i)$) introduced in the previous section, depending on the sign of $r+\beta^2$. All the transformations in this product depend in a very complicated way on the shape of the

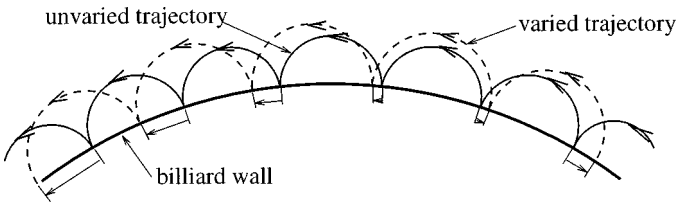

FIG. 5. The skipping orbit along a wall of constant curvature.

billiard domain (through the curvature of the boundary) and the geometry (i.e., flying times, angles of incidence) of the trajectory in question. For this reason the global stability analysis in billiard systems is a rather difficult problem; we turn to this question in Section V.

The stability analysis of individual periodic orbits is much simpler since in this case only a finite product of matrices must be considered. Simplifying the problem further, in this section we deal with a single product of two matrices of the form $\mathbf{R}(A)\mathbf{L}(t)$, which is sometimes called the *tangent map* or the *Jacobian of the bouncing map*.[15–17] It can be considered as the stability matrix of special trajectories having one bounce translational symmetry (examples are given in the next section), or it can also be regarded as the building stone of longer matrix products.

Beyond the non-Euclidean framework the main technical novelty of our approach, in contrast with the above cited works is the use of a coordinate frame fitted to the trajectory itself (and not to the billiard wall), which enables us to split the tangent map in a natural way into two components belonging to the bounceless flight and to the reflection. This decomposition plays a crucial role in the derivations of this section and of Section V.

In the rest of this section through a chain of lemmas and definitions a geometric theorem is derived, which states roughly that the tangent map depends only on the position of the intersection of the arc of the trajectory and the circle of curvature of the billiard wall (at the bounce point). We keep on treating the three cases for the curvature ($r=1,0,-1$) simultaneously.

Lemma 1: The stability of the ''skipping'' motion of the (charged) billiard particle along a billiard wall of constant curvature (on a manifold of constant curvature, in homogeneous magnetic field) is marginal, i.e., the corresponding **RL** matrix is parabolic (figure 5).

Proof: For proving this lemma we have to notice only that due to the (discrete) symmetry of the system, the arcs between the subsequent bounces are congruent, and the bouncing points of both the skipping orbit and its variation follow each other regularly in time and distance along the wall, so the components of the infinitesimal variation tangential to the billiard wall and tangential to the trajectory grow linearly with the number of reflections (i.e., in time). This means that the transverse component of the Jacobi field changes also linearly, so the tangent map **RL** is parabolic. □

The next lemma deals also with parabolic tangent maps, but in the special case of a geodesic billiard wall. For expressing it, the term of *opposite point* is needed. In order to handle extremities in a comprehensive way we equip the Euclidean plane with a single, infinitely far lying ideal point (which is a common intersection of every pair of straight lines), and the hyperbolic plane with an ''ideal line'' consisting of ideal points corresponding to different directions. (Geodesics and hypercycles have two different ideal points belonging to their two ''ends,'' while for paracycles these two points coincide.) In the following it is also advisable to consider a hypercycle as a curve consisting of two branches lying on opposite sides of its base line, and meeting each other in two ideal points.

Let γ be a curve of constant curvature β on a two dimensional manifold of constant curvature $r\in\{1,0,-1\}$, and let S denote an arbitrary point of γ.

Definition 2: The *point S' opposite to S (on γ)* is defined to be the intersection (distinct from S) of the curve γ and the geodesic line perpendicular to γ at S. If specially γ is a geodesic on a manifold of negative curvature (i.e., $\beta=0$, $r=-1$) then let S' be equal to S.

As figure 6 shows, if the curve γ is a circle, then the opposite point S' lies at the other end of the diameter starting at S (for arbitrary r); if γ is a paracycle, then the point S' is at infinity; in the case of hypercycles the opposite point S' lies on the other branch of γ, it is the reflection of the

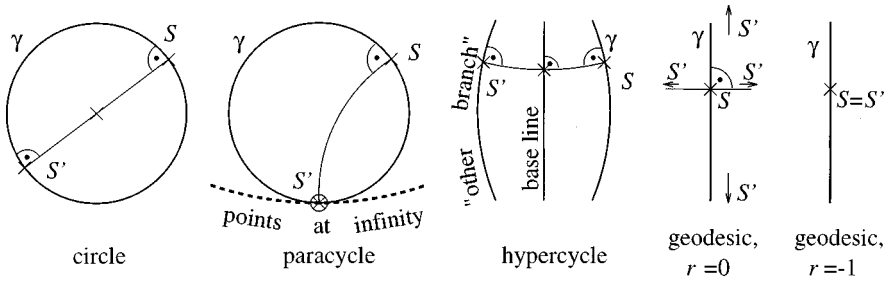

FIG. 6. The position of the point S' opposite to S.

point S in the base line of the hypercycle; while if γ is a geodesic on the Euclidean or hyperbolic plane, then S' is at infinity or coincides with S, respectively.

A better picture can be given for the definition of the opposite point in the case of hypercycles using the Poincaré models (either the P. disk or the P. half-plane model[10]) of the hyperbolic plane. In these models the curves representing the hypercycle γ and the geodesic perpendicular to γ at S do intersect each other in a (real or ideal) point S'' distinct from S, but this point S'' is outside of the disk or half-plane region which represents the points of the hyperbolic plane. Nevertheless, for both models there is a somehow natural involutory geometric transformation (an inversion for the disk model and a reflection for the half-plane model) which sends the points representing the hyperbolic plane out of this domain and vice versa. It can be easily seen that this transformation relates the point S'' to the above defined opposite point S'.

Let the arc SB of the curve γ of constant curvature β be a bounceless trajectory segment with starting point S, which reflects on a geodesic billiard wall η (on a two dimensional manifold of constant curvature $r\in\{1,0,-1\}$) at the bounce point B, and let S' denote the point on γ opposite to S.

Lemma 3: The tangent map belonging to the free flight SB and to the reflection at B is parabolic if and only if the billiard wall η goes through either the starting point S or the opposite point S'.

As figure 7 shows, in most cases this happens for two different positions of the billiard wall. However, if $S'=B$, i.e., the arc SB is a semicircle or if S' is the ideal point of the Euclidean plane, i.e., $r=\beta=0$ then for every possible position of the billiard wall $S'\in\eta$, so the tangent map is parabolic, while if $S=S'$, i.e., $\beta=0$ and $r=-1$ then the tangent map is never parabolic for geodesic billiard walls.

Proof: It follows from lemma 1 that if η goes through S then the tangent map is parabolic. For proving the same for S' different cases must be treated separately.

A. The trajectory γ is a circle

If the particle was started from S' instead of S, then the tangent map belonging to the bounce on the geodesic billiard wall going through S' would be parabolic, according to lemma 1. Since the matrix belonging to the semicircle $S'S$ (which is the difference between the arcs SB and $S'B$) is $-\mathbf{1}$ [equation (3)], the tangent map corresponding to the starting point S and bounce B on the billiard wall BS' is also parabolic.

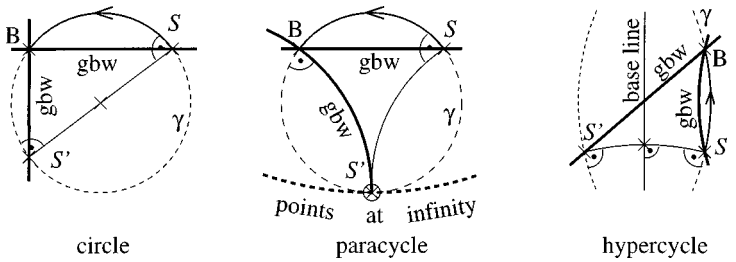

FIG. 7. Marginal cases for the tangent map with a geodesic billiard wall (gbw).

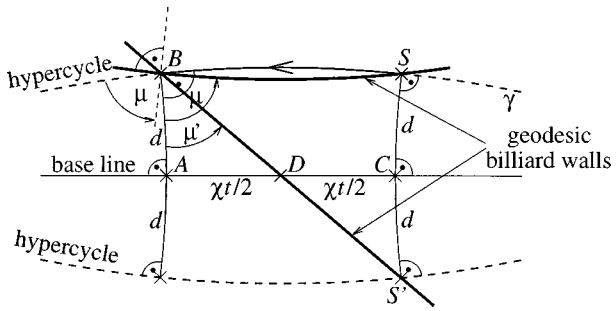

FIG. 8. Marginal positions of geodesic billiard walls in the hyperbolic case.

It can be easily verified by direct calculations [equation (13)], that if $S'=B$, i.e., the arc belonging to the bounceless flight is a semicircle then the tangent map is marginal for every position of the billiard wall.

B. The trajectory γ is a hypercycle

First we calculate the angle of incidence in terms of other geometric data and then compare it with the condition for the marginality of the tangent map. Let μ and μ' denote the angle of incidence belonging to the billiard wall going through S and S', respectively, according to the sign conventions indicated in Figure 1 (Figure 8).

Let d be the altitude of the hypercycle γ, let C and A denote the foot of the perpendicular geodesics dropped from the starting point S of the bounceless segment and from the bounce point B to meet the base line of the hypercycle, respectively, and furthermore let D be the mid-point of the segment AC. (Because of symmetry, D coincides with the mid-point of BS'.) It is also easy to see that $\mu=\angle ABS$ and $\mu'=\angle ABD=\angle CS'D$. According to the elementary connections summarized in Appendix A (A5), it is known that

$$\beta=\tanh(d), \quad \text{so} \quad \sinh(d)=\frac{\beta}{\chi}, \quad \cosh(d)=\frac{1}{\chi},$$

$$|\overline{AC}|=\chi t \quad \text{and} \quad \tan(\mu)=\frac{\cosh(a)}{\sinh(d)\sinh(a)}, \quad \text{where} \quad a=\frac{\chi t}{2}. \tag{9}$$

($|\beta|<1$, $\chi=\sqrt{1-\beta^2}$, t is the flying time on the arc SB, and $a=\chi t/2$ was introduced just to simplify the formulas.) Furthermore, it can be seen from the right triangle $DS'C\Delta$ that

$$\tan(\mu')=\frac{\tanh(a)}{\sinh(d)}=\frac{\sinh(a)}{\sinh(d)\cosh(a)}. \tag{10}$$

The tangent map is parabolic if and only if $\mathrm{Tr}(\mathbf{RH})=\pm 2$ which, having inserted $\mathbf{R}$ and $\mathbf{H}$ from (1) and (6), leads to the condition

$$2\cosh(2a)-\frac{2\beta}{\chi}\tan(\mu^*)\sinh(2a)=\pm 2, \tag{11}$$

where μ^* denotes the angle of incidence corresponding to the parabolic cases. Expressing $\tan(\mu^*)$, one gets

$$\tan(\mu^*)=\begin{cases}\dfrac{\cosh(a)}{\sinh(d)\sinh(a)}=\tan(\mu),\\ \dfrac{\sinh(a)}{\sinh(d)\cosh(a)}=\tan(\mu'),\end{cases} \tag{12}$$

which, comparing to (9) and (10) proves the lemma for the hyperbolic case.

C. The trajectory γ is a paracycle

Instead of getting again entangled in complicated calculations, it is enough to notice that the flying segment SB of a paracycle can be approached either by a series of hypercycle segments or by appropriate circular arcs, and the opposite points S' on the hypercycles or on the circles tend to the (infinitely far lying) opposite point on the paracycle, so by continuity arguments the statement of the lemma becomes evident also for paracycles.

D. The trajectory γ is a geodesic

If the curvature of the manifold r is positive, then the geodesics are at the same time (great) circles and the matrices belonging to the bounceless flights are elliptic; this case has already been treated before. If $r=0,-1$ then the direct calculation of the trace of the tangent map [equations (14), (15)] shows that the bouncing map is always marginal in the case of the Euclidean plane, while it is always hyperbolic for the hyperbolic plane. (Now $q=\beta=0$ so $A=0$.)

Finally, to complete the proof in the opposite direction we must show that in the $\beta\neq 0$ cases there can not be more than two different positions of the (geodesic) billiard wall having a marginal tangent map, since the following traces are strictly monotonic in μ [for geodesic walls $q=0$ so $A=-\beta\tan(\mu)$]:

$$\mathrm{Tr}(\mathbf{R}(A)\mathbf{E}(t))=2\cos(\chi t)+\frac{2A}{\chi}\sin(\chi t), \tag{13}$$

$$\mathrm{Tr}(\mathbf{R}(A)\mathbf{P}(t))=2+2At, \tag{14}$$

$$\mathrm{Tr}(\mathbf{R}(A)\mathbf{H}(t))=2\cosh(\chi t)+\frac{2A}{\chi}\sinh(\chi t). \tag{15}$$

[The quantities A and χ were defined in (1), (3), and (6)]. □

In the next lemma only the transformation **R** belonging to the reflection is investigated in the case of a curved billiard wall. For stating it the following definition is needed in which the concept of ideal point is used in the same sense as in Definition 2.

Let ξ and ξ' be two curves of constant curvature on a (simply connected) manifold of curvature $r\in\{1,0,-1\}$ intersecting each other transversely at the point K.

Definition 4: The *other intersection* (*meeting*) *point* K' *of the curves* ξ, ξ' (*corresponding to* K) is defined to be the common point of the two curves which is distinct from K and lies either on the same branch as K or on the branch opposite to K on both curves ξ and ξ'.

On the sphere and on the Euclidean plane all the curves of constant curvature have only one branch which, in the case of transverse crossing, intersect each other exactly in two points, K and K', so the definition is unambiguous. (The possibility for K' to be an ideal point is allowed.) On the hyperbolic plane, however, the hypercycles have two branches; thus the curves ξ and ξ' can have up to four meeting points; two of them are K and K'. If the branches containing K have two intersections then the one distinct from K is the other intersection point K' and in this case if both ξ and ξ' are hypercycles, their branches opposite to K do not intersect. On the other hand, if the branches containing K have no other intersections then the curves ξ and ξ' are necessarily hypercycles, and their other branches do meet each other in exactly one point which is the other meeting point K'.

Similarly to the definition of opposite point, the motivation for the nonobvious part (concerning the hypercycles) of this definition might also be better understood on special models of the hyperbolic plane. Either on the Poincaré disk or on the Poincaré half-plane model the circles representing the hypercycles ξ and ξ' do cross each other at a (unique) point K'' distinct from K, but this point K'' may well be out of the disk or half-plane region representing the hyperbolic plane. In this case the other meeting point K' can be obtained by affecting K'' by the (natural) involution (inversion or reflection) corresponding to the model; otherwise K' agrees with K''.

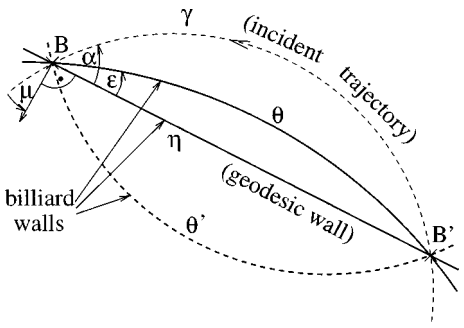

FIG. 9. Bounces having an identical **R** matrix.

Let us consider a particle which moves along the trajectory γ of constant curvature β (on a manifold of constant curvature $r \in \{-1,0,1\}$), and performs a reflection at the point $B \in \gamma$. Let θ, θ' be two possible billiard wall segments of constant curvature q, q', respectively, intersecting γ in B.

Lemma 5: The matrix **R** [equation (1)] belonging to the bounce on θ is equal to the matrix $\mathbf{R}'$ corresponding to the reflection on θ' if and only if the other intersection points of γ and θ as well as γ and θ' coincide (figure 9).

This means that for fixed trajectory γ and bouncing point B, the stability matrix **R** of the reflection depends only on the position of the other meeting point B' of the curves γ and θ.

Proof: The lemma is proven by straightforward calculations, using the polar equations of the curves of constant curvature, which are given in Appendix B. First the general case of $\beta \neq 0$ is considered.

Let η and θ denote the geodesic billiard wall and the billiard wall of constant curvature q, respectively, both of which pass through B and B'. Furthermore, let μ be the angle of incidence of the trajectory γ (of curvature β) in η, and let ϵ denote the angle between η and θ directed as it is shown in figure 9. The angle of incidence of γ in the billiard wall θ is $\mu+\epsilon$. For proving the lemma it is enough to investigate the conditions in which the bouncing effect of the geodesic wall η is identical to that of the curved wall θ. For this reason the coincidence of the only nontrivial element ($R_{2,1}$) of the matrix **R** has to be examined [see equation (1)]:

$$\beta \tan(\mu) = \frac{q+\beta \sin(\mu+\epsilon)}{\cos(\mu+\epsilon)}, \tag{16}$$

which leads to the connection

$$0 = q \cos(\mu) + \beta \sin(\epsilon). \tag{17}$$

Let the angle between the geodesic η and the trajectory γ be denoted by α (figure 9). Since $\alpha = \pi/2 - \mu$, equation (17) can also be written as

$$-\frac{\sin(\alpha)}{\beta} = \frac{\sin(\epsilon)}{q}, \tag{18}$$

which, according to lemma 13 of Appendix B, means that the other intersection points of the trajectory γ with the billiard walls η and θ coincide, which was to be proven. [The minus sign in equation (18) appears because of the sign conventions for the curvatures q and β.]

In the special case of $\beta = 0$ the proof is similar; let θ and θ' be two billiard walls of constant curvature q and q' belonging to the angle of incidence μ and μ', respectively. Requiring the equivalence of the corresponding bouncing maps we get

$$\frac{q}{\sin(\alpha)} = \frac{q'}{\sin(\alpha')}, \tag{19}$$

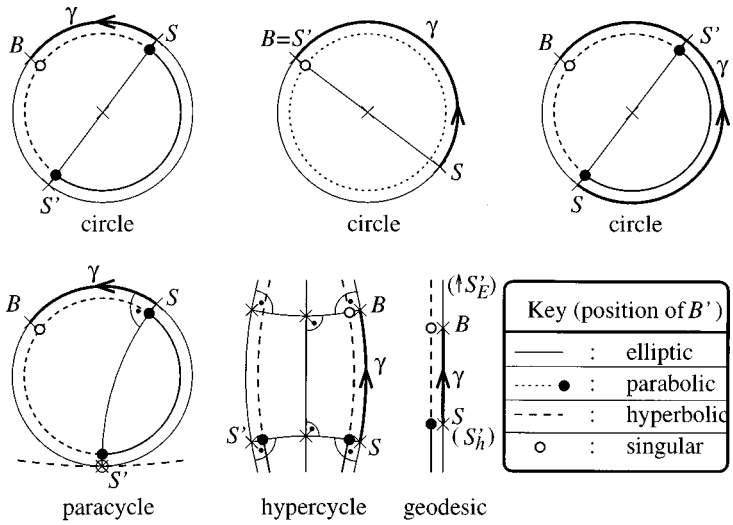

FIG. 10. The stability of the tangent map.

where $\alpha=\pi/2-\mu$ and $\alpha'=\pi/2-\mu'$ are the angles between the billiard walls θ, θ' and the geodesic trajectory γ. This means, using again lemma 13 of Appendix B, that θ and θ' meet each other at a point B' which lies on γ; this completes the proof of the lemma. □

Now, with the help of the previous lemmas we are ready to prove the main geometric theorem on the behavior of the tangent map **RL**.

Let the arc SB of the curve γ of constant curvature β be a bounceless trajectory segment with starting point S, which is reflected on the billiard wall θ of constant curvature q at the bouncing point B (on a two dimensional manifold of constant curvature $r\in\{1,0,-1\}$). (If the curvature of the billiard wall is not constant, then θ should be considered as its circle of curvature at B.) Let S' be the point on γ opposite to the starting point S, and let B' denote the other meeting point (different from B) of the two curves γ and θ.

Theorem 6: (On the stability of the tangent map). For fixed γ, S, and B on it, varying the billiard wall θ, the tangent map **RL** belonging to the bounceless flight SB and to the reflection at B depends only on the position of the other crossing point B'. This map is parabolic, if $B=S'$ (independently of the position of B') or if B' coincides with either S or S'; while **RL** is hyperbolic if ($B\neq S'$ and) B' lies on the same open arc SS' of γ where B is; otherwise the tangent map is elliptic (figure 10).

(Taking into consideration the ideal points, γ is topologically equivalent to a circle, and the points S and S' divide it into two parts.)

Proof: We remark that for geodesic trajectories ($\beta=0$) in the Euclidean case ($r=0$) the opposite point S'_E is the ideal point (so exactly the geodesic billiard walls pass through it), while on the hyperbolic plane $S'_h=S$.

The first statement of the theorem is a trivial consequence of lemma 5, since for fixed γ, S, and B the matrix **L** does not vary either. The positions of B' corresponding to parabolic mappings can also be obtained easily by joining the results of lemma 3 and lemma 5. The rest of the theorem can be proven by calculating Tr(**RL**) for the different cases [equations (13), (14), and (15)], and observing their monotonity properties as μ and q change. □

In the remaining parts of the paper our attention is restricted solely to the hyperbolic plane.

IV. APPLICATION: STABILITY OF TRAJECTORIES WITH DISCRETE SYMMETRY

In this section, applying the results of Theorem 6 of the previous section the stability of a special symmetric orbit, bouncing successively in the adjacent mid-points of the (geodesic) sides of a regular polygonal billiard on the hyperbolic plane is determined (the inset of Figure 11). Let the mid-points of the sides of the regular n-gon $V_1V_2\ldots V_n$ with center O be denoted by $M_1,M_2,\ldots,M_n$; let $R:=|OV_i|$ and $\rho:=|OM_i|$ be the circumradius and the inradius, respectively;

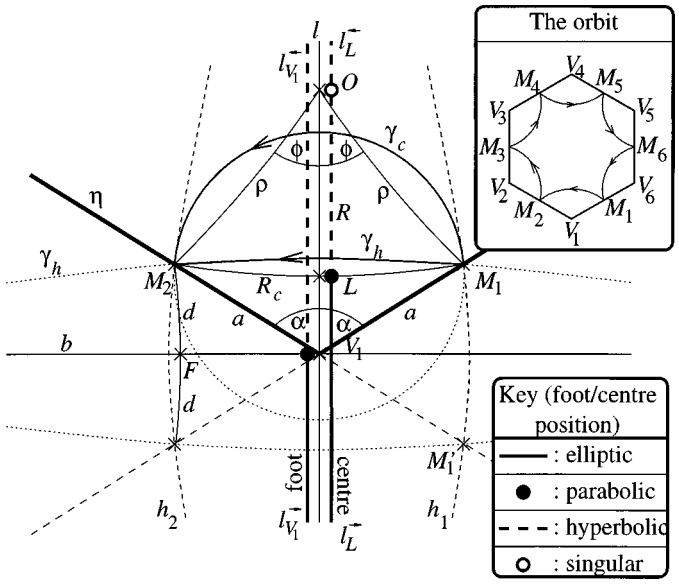

FIG. 11. The stability of the symmetric orbit in the regular polygon billiard.

and finally let $\alpha := \angle OV_iV_{i+1}$ and $a := |V_iM_i|$ denote the half of the interior angle and of the side of the polygon, respectively (Figure 11). These parameters are not independent; they all can be expressed with $\phi := \pi/n$ and R:

$$\sinh(a) = \sin(\phi)\sinh(R), \quad \cot(\alpha) = \tan(\phi)\cosh(R), \tag{20}$$

$$\tanh(\rho) = \cos(\phi)\tanh(R). \tag{21}$$

Figure 11 shows one segment with starting point M_1 and bouncing point M_2 of the investigated orbit in the polygon billiard. Let l be the geodesic line of symmetry passing through O and V_1, and let h_1, h_2 be the two branches of the hypercycle of base line l, going through M_1, M_2, respectively. It is obvious that varying the magnetic field β there is a whole series of trajectory segments γ of constant curvature β joining the points M_1 and M_2. It is also easy to see that if $|\beta| > 1$ (or $|\beta| = 1$) then these curves are circular arcs (or segments of paracycles), whose center lies on l (possibly at the ideal points of l), and the points opposite to the starting point M_1 on γ line up along the hypercycle branch h_2 (at the ideal points for $|\beta| = 1$); while if $|\beta| < 1$ then these curves are arcs of hypercycles, whose base line is perpendicular to l, so the points opposite to M_1 lie on the hypercycle branch h_1. Inversely, given any point L on the symmetry line l, there is exactly one circle/hypercycle segment (or arc of paracycle, if L is an ideal point) going from M_1 to M_2 inside the billiard domain, having a center/foot point L. (The term foot point of the hypercycle is used here to designate the intersection of the base line and l.) The only singular case is the circle of center O (i.e., the inscribed circle) which is tangential to the billiard walls at M_i.

Let L be the mid-point of the geodesic segment M_1M_2 and let us denote the open rays of the (geodesic) symmetry line l with origin L, V_1 directed inwards and outwards the billiard by $l_L^{\leftarrow}, l_{V_1}^{\leftarrow}$ and $l_L^{\rightarrow}, l_{V_1}^{\rightarrow}$, respectively.

Our aim in this section is to determine geometrically the stable, marginal, and unstable positions of the investigated symmetric orbits and the corresponding magnetic field regions. The following theorem answers this question, using the notations of figure 11, introduced before.

Theorem 7: The symmetric orbit bouncing successively in the mid-points M_i of the sides of the regular polygon $V_1V_2 \ldots V_n$ (on the hyperbolic plane) is stable, if the arc M_1M_2 of this orbit is either a circular segment with a center lying on the (open) ray $l_L^{\rightarrow}$ or it is a segment of a hypercycle with a foot point on $l_{V_1}^{\rightarrow}$; the stability of the orbit is marginal if the center of the circular arc is at L or the foot point of the hyperbolic segment is at V_1; while the orbit is unstable, if the center of the circle or the foot point of the hypercycle lies on the opposite rays $l_L^{\leftarrow}, l_{V_1}^{\leftarrow}$, respectively.

The magnetic field region belonging to the stable orbits is $\beta_{\min}<\beta<\beta_{\max}$, where

$$\beta_{\min}=\frac{\sin^2(\phi)\sinh(R)\cosh(R)}{\sqrt{1+\sin^2(\phi)\sinh^2(R)(1+\sin^2(\phi)\cosh^2(R))}}, \tag{22}$$

$$\beta_{\max}=\frac{\sqrt{1+\sin^2(\phi)\sinh^2(R)(1+\cos^2(\phi))}}{\sin(\phi)\cos(\phi)\sinh(R)}. \tag{23}$$

The stability regions for the hypercycles and circles are shown in Figure 11. (The center or foot points can also be ideal points in which case the arcs of the orbit are paracycles.)

Proof: The positions belonging to parabolic tangent maps can easily be found, since in this case, using lemma 3 the points opposite to M_1 are in the intersection of the (geodesic) line V_1M_2 (shortly denoted by η) and the hypercycle h_1, h_2. If the opposite point is $M_2=\eta\cap h_2$ then the corresponding orbit is the circle γ_c of center L, while if the opposite point is $M_1'\coloneqq\eta\cap h_1$ then the marginally stable symmetric trajectory is the hypercycle γ_h of foot point V_1 and base line b (figure 11).

It is quite plausible that the stable and unstable regions are between the two marginal cases. The rigorous proof of the statements of the lemma concerning the elliptic and hyperbolic cases can be accomplished with the use of theorem 6, taking systematically, one by one the different geometric arrangements depending on the position of the center/foot point of γ, the opposite point of M_1, and the other meeting point on η, belonging to M_2. (In theorem 6 the trajectory segment was fixed, and the position of the billiard wall could be varied, while in the present case the billiard wall η is fixed, and the orbit γ changes what makes the proof a bit cumbrous.)

To get the critical magnetic field $\beta_{\max}$ first the radius $R_c\coloneqq|LM_2|$ of the critical trajectory γ_c has to be determined from the right triangle $V_1LM_2\Delta$: $\sinh(R_c)=\sinh(a)\sin(\alpha)$. Having inserted here a and α from (20), and using equation (2) the formula (23) can easily be derived. If $\beta>\beta_{\max}$ then the arc γ between M_1 and M_2 no longer exists. In the $R\to\infty$ limit (23) simplifies to $\lim_{R\to\infty}\beta_{\max}=\sqrt{1+\cos^2(\phi)}/\cos(\phi)$.

For calculating the altitude d of the other critical trajectory γ_h let us consider the right triangle M_2FV_1, where F is the foot of the perpendicular dropped from M_2 to the base line b, which gives $\sinh(d)=\sinh(a)\cos(\alpha)$. The value of $\beta_{\min}$ [equation (22)] can be obtained by substituting here the values of a and α from the formula (20) and using the relationship (5). It is worth noticing that $\lim_{R\to\infty}\beta_{\min}=1$, which is also clear geometrically, since in this limit the vertices V_i tend to infinity, and the arc γ_h becomes a paracycle.

This completes the proof of the theorem. □

In the rest of this section the result of theorem 7 is demonstrated by numerical simulations made for an equilateral triangular billiard $V_1V_2V_3\Delta$ on the hyperbolic plane. The size of the billiard can be characterized by the half of the angle 2ψ subtended by the (whole geodesic) lines of the sides at the center O of the billiard (figure 12/A), which was chosen to be $\psi=75°$ in the investigated case. (For $\psi=60°$ the noncompact billiard with vertices at infinity—one third of which is the fundamental domain of the modular group $Sl(2;\mathbb{Z})$—has already been studied in references 5, 9, 24, 25, 11. Increasing ψ up to 90° the size of the triangle tends to zero.) Knowing this angle ψ the radius ρ of the inscribed circle can be calculated: $\tanh(\rho)=\cos(\psi)=(\sqrt{3}-1)/(2\sqrt{2})=0.2588...$, and using (21) and (22) the circumradius R and the critical magnetic field $\beta_{\min}$ can also be determined: $\tanh(R)=2\cos(\psi)=(\sqrt{3}-1)/\sqrt{2}=0{,}5176...$,

$$\beta_{\min}=\frac{3\cos(\psi)}{\sqrt{16\cos^4(\psi)-11\cos^2(\psi)+4}}=\frac{3(\sqrt{3}-1)\sqrt{22+5\sqrt{3}}}{\sqrt{2}\sqrt{409}}=0{,}4252... . \tag{24}$$

The results of the computer simulation can be seen in Figures 12, 13. The pairs of drawings and plots show the stable periodic orbits and the corresponding Poincaré sections of the phase space of the triangular billiard (i.e., the iterates of the bounce map from a few different initial points), for two magnetic field values $\beta_>=\sin(0{,}44\text{ rad})=0{,}4259...>\beta_{\min}$ (Figures 12(A), 13(A)) and $\beta_<=\sin(0{,}438\text{ rad})=0{,}4241...<\beta_{\min}$ (Figures 12(B), 13(B)) around the critical value $\beta_{\min}$.

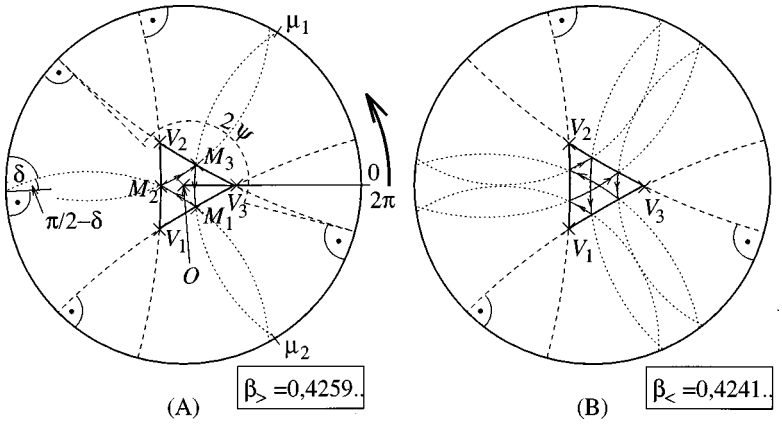

FIG. 12. The stable periodic orbit before (a) and after (b) the bifurcation in the triangular billiard (represented on the Poincaré disk model).

[Taking the $a\to\infty$ limit in the last equation of (A5), the relationship $\beta=\sin(\pi/2-\delta)$ can be obtained for the curvature of the hypercycle β and the angle $\pi/2-\delta$ between the hypercycle and its chord with foot points at infinity (Figure 12(A)). This angle is also meaningful on the Poincaré models of the hyperbolic plane, and it can be used as an alternative to β for characterizing the magnetic field. In our case $\arcsin(\beta_{\min})=0{,}4392...$ rad.]

The dots with coordinates (μ_1,μ_2) in the Poincaré maps (Figure 13) represent the consecutive arcs of the bounceless flights by the angles μ_1, μ_2 of their foot points on the border of the Poincaré disk (Figure 12(A)). (The plottings are enlarged, showing only one of the three ''island'' corresponding to the symmetric orbit of period three.)

It can be clearly seen that the stable periodic orbit belonging to the single ''island'' in figure 13/A loses its stability via period doubling bifurcation.

V. THEOREMS FOR GLOBAL INSTABILITY

The main aim of this section is to provide sufficient criteria for hard chaos in billiards on the hyperbolic plane with weak (i.e., $|\beta|<1$) magnetic field β. The notion *hard chaos* signifies here that there are no stable orbits in the billiard, i.e., every trajectory has a positive Lyapunov exponent. (Or practically, there are no ''islands of stability'' on the Poincaré section.) In most cases the theorems to be introduced also allow us to estimate from below the value of the Lyapunov exponent.

First Sinai-like billiards (i.e., billiards surrounded by dispersing arc segments) will be studied, and then, using the main theorem of Section III (Theorem 6) a more general criterion will be derived concerning Bunimovich-like domains, the boundary of which can contain (strongly) focusing segments, too, which lie sufficiently far from each other.

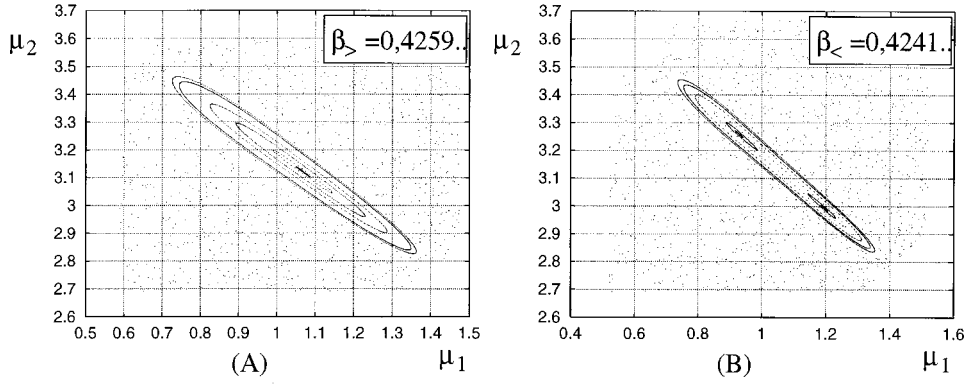

FIG. 13. Poincaré maps slightly above (a) and below (b) the critical field $\beta_{\min}=0,4252...$.

To gain a heuristic insight into the behavior of the product of elements of the $Sl(2;\mathbb{R})$ group (which describes the stability of the trajectories in the billiard system) it is advantageous to use the concept of invariant cone which was introduced and exhaustively studied in Refs. 26, 27.

Let us consider the group $Sl(2;\mathbb{R})$ (of two by two real matrices of unit determinant) with its canonical action on the $\mathbb{R}^2$ plane. For an element $L\in Sl(2;\mathbb{R})$ the (double) *cone* $C\subset\mathbb{R}^2$ (consisting of all the linear combinations of two linearly independent vectors with coefficients having the same sign) is *invariant* if $LC\subset C$. It is a trivial observation that an element of the group $Sl(2;\mathbb{R})$ (unequal to $\pm\mathbf{1}$) is elliptic if and only if it does not possess an invariant cone. (Indeed, if L had an invariant cone, then, by continuity arguments, there would be an invariant straight line, i.e., an eigenvector in this cone. Inversely, for hyperbolic and parabolic matrices an invariant cone can easily be constructed using the eigenvectors; e.g., the coordinate cone $\{(b,b')\in\mathbb{R}^2|bb'\geqslant 0\}$ is invariant in figures 1/B, 3, 4/B.)

The heuristic geometrical picture behind the proofs of the hard chaos theorems is based on the existence of a (fixed) cone in the (b,b') coordinate plane, which is shown to be left invariant by the elements of a suitably chosen partition (consisting of nonelliptic multiplicands) of the matrix product $\ldots\mathbf{R}_i\mathbf{H}_i\ldots\mathbf{R}_2\mathbf{H}_2\mathbf{R}_1\mathbf{H}_1$, which means that the whole product, as it leaves invariant the same cone, too, is also nonelliptic (in fact, it is mostly hyperbolic). (The linear mappings $\mathbf{R}_i$, $\mathbf{H}_i$ correspond to the subsequent reflections and bounceless flying segments, respectively.)

The appearance of a strictly positive Lyapunov exponent in a billiard system can be originated from two types of heuristical reasons; the billiard particle either moves in a domain of negative curvature, in weak magnetic field, so in the first place the bounceless segments are responsible for the strong hyperbolicity, or it performes elastic bounces on dispersing boundary segments sufficiently often, and the common effect of the reflections and the bounceless flights between them result in a net hyperbolic behavior. [As equation (1) shows, the effect of the reflection itself is marginal, but any of the three products $\mathbf{R}(A)\mathbf{H}(t)$, $\mathbf{R}(A)\mathbf{P}(t)$, $\mathbf{R}(A)\mathbf{E}(t)$ can be hyperbolic for sufficiently big parameter values A (i.e., for dispersing walls) and for small values of t in the elliptic case.] The estimations of this section are essentially guided by the first reasoning, since in the case investigated it is technically easier to carry out, and the conditions obtained are simpler and less restrictive on the shape of the billiard. (In Ref. 13 more refined estimations, taking into account also the dispersing effect of the reflections have been worked out and applied for planar magnetic billiards, where the bounceless motion is no more hyperbolic.)

It is easy to see that the previously mentioned $bb'\geqslant 0$ cone of the (b,b') coordinate plane is invariant for the matrices belonging to the bounces with positive parameter values $A>0$ [equation (1), Figure 1(B)] and for the parabolic and hyperbolic matrices corresponding to the bounceless segments with parameter values $t>0$ [equations (4), (6), Figures 3, 4(B)]; this motivates the following statement.

Lemma 8: For the products of the matrices $\mathbf{R}_i:=\mathbf{R}(A_i)$, $\mathbf{P}_i:=\mathbf{P}(t_i)$, and $\mathbf{H}_i:=\mathbf{H}(t_i)$ with positive parameter values $A_i>0$, $t_i>0$ [equations (1), (4), (6)] the following inequalities hold $(n=2,3,\ldots)$:

$$\mathrm{Tr}(\mathbf{R}_n\mathbf{H}_n\ldots\mathbf{R}_2\mathbf{H}_2\mathbf{R}_1\mathbf{H}_1)>\mathrm{Tr}(\mathbf{R}(\hat{A})\mathbf{H}(T))=2\cosh(\chi T)+\frac{2\hat{A}}{\chi}\sinh(\chi t), \tag{25}$$

$$\mathrm{Tr}(\mathbf{R}_n\mathbf{P}_n\ldots\mathbf{R}_2\mathbf{P}_2\mathbf{R}_1\mathbf{P}_1)>\mathrm{Tr}(\mathbf{R}(\hat{A})\mathbf{P}(T))=2+2\hat{A}T, \tag{26}$$

where $T=\Sigma_{i=1}^{n}t_i$, $\hat{A}=\Sigma_{i=1}^{n}A_i$, and χ is the function $\chi=\sqrt{1-\beta^2}$ of the magnetic field β.

Proof: Let us notice that all elements of the matrices $\mathbf{R}_i$, $\mathbf{P}_i$, and $\mathbf{H}_i$ are non-negative, and the matrix belonging to the reflection can be written in the form $\mathbf{R}_i=\mathbf{1}+\mathbf{A}_i$, where $\mathbf{1}$ is the identity matrix and

$$\mathbf{A}_i=\begin{pmatrix}0 & 0\\ A, & 0\end{pmatrix}.$$

Using this,

$$\mathrm{Tr}((\mathbf{1}+\mathbf{A}_n)\mathbf{H}_n\ldots(\mathbf{1}+\mathbf{A}_2)\mathbf{H}_2(\mathbf{1}+\mathbf{A}_1)\mathbf{H}_1)>\mathrm{Tr}(\mathbf{H}(T))+\sum_{i=1}^{n}\mathrm{Tr}(\mathbf{A}_i\mathbf{H}(T))=\mathrm{Tr}(\mathbf{R}(\hat{A})\mathbf{H}(T)), \tag{27}$$

where the trace of the last term equals the right side of the formula (25) of the lemma. [In the derivation (27) first the product of matrices under the trace has been expanded, and the obtained sum (of positive elements) has been reduced by omitting the terms containing more than one of the $\mathbf{A}_i$ factors. Then the result has been simplified using the one parameter subgroup property of the transformations $\mathbf{H}_i$ and the spacial form of the matrices $\mathbf{R}_i$.]

The second inequality (26) of the lemma can be proven along the same lines. □

Using this lemma the criterion for hard chaos in dispersing (magnetic) billiards can be given.

Theorem 9: (Hard chaos in dispersing billiards.) Let Q be a billiard domain in the hyperbolic plane with a piecewise C^2 boundary, the curvature q of which is everywhere $q\leqslant-|\beta|$ (sufficiently dispersing), where the magnetic field β satisfies the condition $|\beta|<1$. Then the magnetic billiard system Q demonstrates hard chaos; every (nonsingular) trajectory of the system has a Lyapunov exponent greater than or equal to $\chi=\sqrt{1-\beta^2}$.

(The attribute ''nonsingular'' means here that the trajectory does not run into a vertex of the billiard in the future.)

Proof: Because of the condition $q\leqslant-|\beta|$ of the theorem the parameters $A=[q+\beta\sin(\mu)]/\cos(\mu)$ are always positive in the matrices corresponding to the reflections [formula (1)], so lemma 8 can be applied for the matrix product describing the stability of the trajectory investigated. Since the ratio of the trace and the greatest eigenvalue of this matrix product tends to unity as $n\to\infty$, the Lyapunov exponent λ can be calculated from the trace of the Jacobian matrix, and it can be estimated from below with the help of the inequality (25):

$$\lambda=\limsup_{n\to\infty}\frac{\ln(\mathrm{Tr}(\mathbf{R}_n\mathbf{H}_n\ldots\mathbf{R}_2\mathbf{H}_2\mathbf{R}_1\mathbf{H}_1))}{T_n}\geqslant\lim_{n\to\infty}\left(\frac{1}{T_n}\ln\left(2\cosh(\chi T_n)+\frac{2\hat{A}_n}{\chi}\sinh(\chi T_n)\right)\right)=\chi, \tag{28}$$

where the matrices $\mathbf{H}_i$ and $\mathbf{R}_i$ correspond to the subsequent bounceless flying segments of time t_i and reflections with parameter A_i, and T_n, $\hat{A}_n$ denote the sums $T_n=\Sigma_{i=1}^{n}t_i$, $\hat{A}_n=\Sigma_{i=1}^{n}A_i$, respectively. (It was supposed, that $\hat{A}_n$ does not grow exponentially with n.) □

It is worth noticing that the result of the estimation obtained above does not contain any information about the reflections, it just tells us that if the dispersive curvature $-q$ of the billiard boundary is greater than the magnetic field β, then the magnetic billiard system is at least as much chaotic as it would be without the bounces. It is quite possible that for $-q-|\beta|>K>0$ (K is constant) stronger—but not so general—results could also be achieved for certain classes of billiards, taking into consideration the dispersive effects of the bounces.

Now let us turn to the investigation of billiards with focusing arcs in their boundary! For curvature values q in the regime $-|\beta|<q<|\beta|$ the sign of the parameter A in the matrices belonging to the bounces [formula (1)] depends on the angle of incidence μ, while for $q>|\beta|$ the shearing $\mathbf{R}(A)$ moves the positive b axis always downwards, in the direction of the $-b'$ axis; thus the cone $bb'\geqslant$ is not left invariant for any values of μ. However, as figure 14/A shows, it is possible that for sufficiently great parameter values τ_A and A, the transformation of the form $\mathbf{H}(\tau_A/2)\mathbf{R}(A)\mathbf{H}(\tau_A/2)$ does leave invariant the double cone mentioned above. (A similar assertion holds for the parabolic case.) This geometric observation is expressed in the following statement.

Lemma 10: If $A=-[q+\beta\sin(\mu)]/\cos(\mu)<-\chi<0$ holds for the shearing parameter A of the stability matrix $\mathbf{R}(A)$, where $\chi=\sqrt{1-\beta^2}$, $q>1$ (i.e., the billiard wall is strongly focusing) and $|\beta|<1$ (i.e., the magnetic field is weak); then there exists a time $\tau_A:=2\chi^{-1}\,\mathrm{artanh}(|\chi/A|)$ in the hyperbolic case, and $\tau_A':=|2/A|$ in the parabolic case, for which

$$\mathbf{H}\left(\frac{\tau_A}{2}\right)\mathbf{R}(A)\mathbf{H}\left(\frac{\tau_A}{2}\right)=-\mathbf{R}(-A), \tag{29}$$

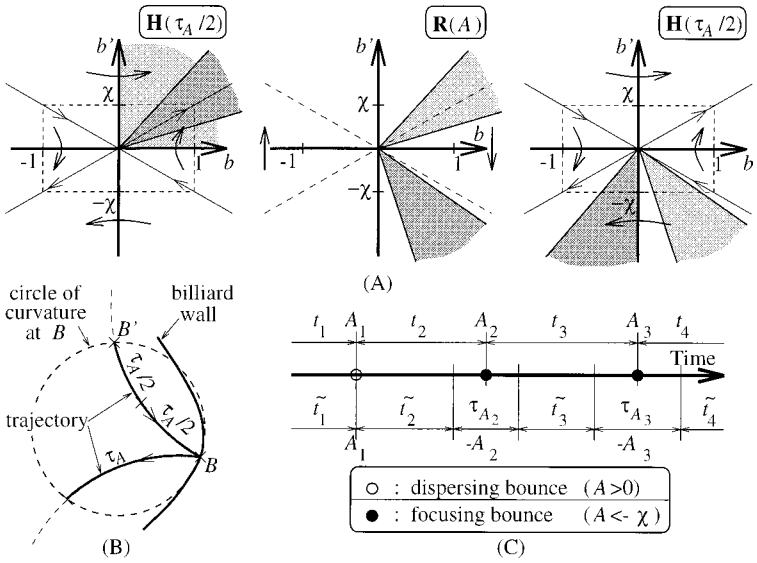

FIG. 14. The effect of the transformation $\mathbf{H}(\tau_A/2)\mathbf{R}(A)\mathbf{H}(\tau_A/2)$ for strongly focusing walls.

$$\mathbf{P}\left(\frac{\tau_A'}{2}\right)\mathbf{R}(A)\mathbf{P}\left(\frac{\tau_A'}{2}\right)=-\mathbf{R}(-A), \tag{30}$$

hold, respectively, and for $t\geqslant\tau_A$ (or for $t\geqslant\tau_A'$) values the double cone $bb'\geqslant 0$ is invariant under the effect of the transformation $\mathbf{H}(t/2)\mathbf{R}(A)\mathbf{H}(t/2)$ [or under that of $\mathbf{P}(t/2)\mathbf{R}(A)\mathbf{P}(t/2)$]. Moreover, the critical time τ_A (or τ_A') is equal to the flying time elapsed between the bouncing point B and the other meeting point B' of the outgoing (or equivalently, the incoming) trajectory segment and the circle of curvature of the billiard boundary.

The geometrical meaning of the critical time τ_A is demonstrated in figure 14/B, and **H**, **P**, and **R** are given by the formulas (1), (4), and (6).

Proof: First we remark, that if $q>1$ then $A+\chi=[-q-\beta\sin(\mu)+\chi\cos(\mu)]/\cos(\mu)=[\cos(\mu+\varphi)-q]/\cos(\mu)<0$, where $\varphi=\arctan(\beta/\chi)$, so the conditions of the lemma are in harmony.

We prove the statement of the lemma only for the hyperbolic case, since the parabolic case can be treated analogously. The validity of the relationship (29) can be shown by direct calculations, using the connection $\tanh(\chi\tau_A/2)=-\chi/A$:

$$\mathbf{H}\left(\frac{\tau_A}{2}\right)\mathbf{R}(A)\mathbf{H}\left(\frac{\tau_A}{2}\right)=\begin{pmatrix} C+\dfrac{AS}{\chi} & \dfrac{A(C-1)}{\chi^2}+\dfrac{S}{\chi} \\ AC+A+\chi S & C+\dfrac{AS}{\chi}\end{pmatrix}=-\begin{pmatrix}1 & 0\\ -2A & 1\end{pmatrix}, \tag{31}$$

where S stands for $S=\sinh(\chi\tau_A)$ and C for $C=\cosh(\chi\tau_A)$.

For $t\geqslant\tau_A$ values the double cone $bb'\geqslant 0$ is left invariant by each matrix in the last form of the product $\mathbf{H}(t/2)\mathbf{R}(A)\mathbf{H}(t/2)=-\mathbf{H}[(t-\tau_A)/2]\mathbf{R}(-A)\mathbf{H}[(t-\tau_A)/2]$.

Since $\mathrm{Tr}(\mathbf{H}(\tau_A/2)\mathbf{R}(A)\mathbf{H}(\tau_A/2))=\mathrm{Tr}(\mathbf{R}(A)\mathbf{H}(\tau_A))=-2$, by theorem 6 the last statement of the lemma is also true. □

This lemma enables us to ''dress up'' the matrices $\mathbf{R}(A)$ corresponding to the strongly focusing bounces ($q>1$) with hyperbolic matrices of the form $\mathbf{H}(\tau_A/2)$ and thus to trace back the stability analysis of certain billiards possessing focusing walls to the case of dispersing billiards. This procedure is shown graphically in figure 14/C. Before expressing the main theorem of this section we introduce a definition.

Let us characterize a trajectory (in a billiard with magnetic field $|\beta|<1$ on the hyperbolic plane) by the times t_i of the subsequent bounceless flying segments and the parameter values A_i [equation (1)] of the succeeding reflections, furthermore, assume, that the bounces take place either on dispersing boundary segments of curvature $q\leqslant-|\beta|$ or on focusing ones having curva-

ture greater than unity [i.e., either $A_i>0$ or $A_i<-\chi=-\sqrt{1-\beta^2}$]. Let τ_A, be zero in the case of $A_i>0$ and let it be $\tau_{A_i}:=2\chi^{-1}\,\mathrm{artanh}(|\chi/A_i|)$ (as in lemma 10), if $A_i<-\chi$.

Definition 11: The *reduced time* $\tilde{t}_i$ is defined to be $\tilde{t}_i:=t_i-(\tau_{A_{i-1}}+\tau_{A_i})/2$ (figure 14/C).

Theorem 12: (hard chaos in nondispersing billiards.) Let Q be a (magnetic) billiard domain (on the hyperbolic plane) with a piecewise C^2 boundary the curvature q of which is either sufficiently dispersing, $q\leqslant-|\beta|$, or strongly focusing, $q>1$, almost everywhere, where the magnetic field β is in the range $|\beta|<1$. Furthermore, suppose, that the focusing segments lie sufficiently far from each other, i.e., the reduced time of each bounceless segment on every trajectory is nonnegative, and along every (nonsingular) trajectory the inequality,

$$\liminf_{n\to\infty}\frac{\tilde{T}_n}{T_n}\geqslant K>0, \tag{32}$$

holds for the sums $\tilde{T}_n:=\Sigma_{i=1}^n\tilde{t}_i$, $T_n:=\Sigma_{i=1}^n t_i$ of the (reduced) flying times, with a positive constant K. In this case the Lyapunov exponent λ of every nonsingular trajectory satisfies the following estimation:

$$\lambda\geqslant K\chi=K\sqrt{1-\beta^2}. \tag{33}$$

Proof: The proof of the theorem is straightforward; first, using lemma 10 the shearings of negative parameters in the infinite matrix product describing the stability of the trajectories have to be ''dressed up'' and converted to shearings with positive parameters, and then theorem 9 should be applied with the estimation (32) to get the desired result:

$$\lambda=\limsup_{n\to\infty}\frac{\ln(\mathrm{Tr}(\mathbf{R}_n\mathbf{H}_n\ldots\mathbf{R}_2\mathbf{H}_2\mathbf{R}_1\mathbf{H}_1))}{T_n}=\limsup_{n\to\infty}\left(\frac{\tilde{T}_n}{T_n}\frac{\ln|\mathrm{Tr}(\tilde{\mathbf{R}}_n\tilde{\mathbf{H}}_n\ldots\tilde{\mathbf{R}}_2\tilde{\mathbf{H}}_2\tilde{\mathbf{R}}_1\tilde{\mathbf{H}}_1)|}{\tilde{T}_n}\right)\geqslant K\chi, \tag{34}$$

where $\mathbf{R}_i=\mathbf{R}(A_i)$, $\mathbf{H}_i=\mathbf{H}(t_i)$, $\tilde{\mathbf{R}}_i=\mathbf{R}(|A_i|)$, and $\tilde{\mathbf{H}}_i=\mathbf{H}(\tilde{t}_i)$. □

It should be noted that since a trajectory can perform more consecutive bounces on the same smooth (or C^2) focusing segment of the boundary, the nonnegativity condition for the reduced times also restricts the shape of the focusing segments, not only the distances between them. It is also clear from lemma 10 and definition 11 that for a circular (focusing) boundary segment the reduced times, corresponding to consecutive bounces on this arc, are zero, but possibly not only segments of constant curvature $q>1$ are allowed.

Wojtkowsky, investigating ordinary planar billiards, was also led to the same problem in a much simpler setup, without a magnetic field, on the Euclidean plane, in which case he could prove that exactly those focusing boundary segments are allowed, which satisfy the so called positive scattering property $R''(s)\leqslant 0$, where $R(s)$ is the radius of curvature of the billiard boundary as the function of arc length on the perimeter.[21] He also proved that this condition is robust under small perturbations of the billiard wall (at least when the inequality is strict). This suggests that the conditions of our theorem 12 are also robust under small perturbations, but we do not turn to the detailed analysis of this question.

In Appendix C an example is given for the application of theorem 12.

It is quite possible that one can go further and using the results of Oseledec[28] and Pesin,[29] suitably generalized for dynamical systems with singularities,[30,31] positive entropy theorems can also be formulated for the investigated billiard types, but these questions are out of the framework of the present study.

VI. CONCLUSIONS

In this work the local and global stability of magnetic billiard systems has been investigated on two dimensional Riemannian manifolds of constant curvature, with a special emphasis on the hyperbolic plane. For tracing the local behavior of the nearby trajectories the technique of Jacobi fields was used. A coordinate frame attached to the investigated trajectory was introduced, which allowed us to decompose the stability matrix into a product of elements of the group $Sl(2;\mathbb{R})$

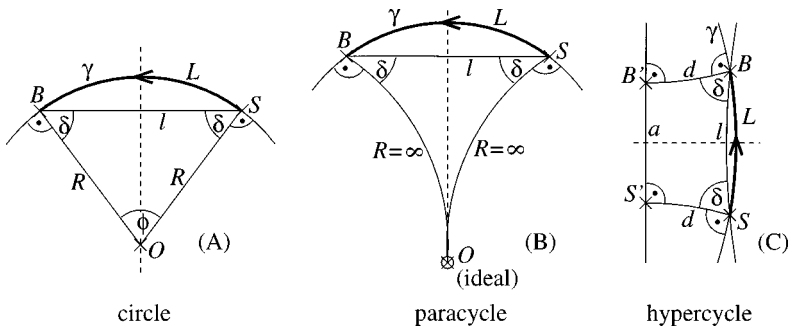

FIG. 15. The arcs, chords in the cycle γ.

related to the reflections and to the bounceless segments. (The foundation of this local analysis in a more general context can be found in Ref. 12, and its adaptation to our present purposes is in Section II of this work.)

Studying the local stability properties of the trajectories, it has been shown via a series of lemmas that on two dimensional manifolds of constant curvature (of arbitrary sign), in a homogeneous magnetic field the Jacobian of the bouncing map depends only on the position of the meeting point (different from the reflection point) of the incoming trajectory and the circle of curvature of the billiard wall at the bouncing point (theorem 6).

Applying this theorem, the stability region for a special symmetric orbit in regular polygon billiards on the hyperbolic plane has been determined, as the magnetic field changes (theorem 7). This result has also been supported by numerical simulations.

Finally, in Section V global criteria for hyperbolicity and lower estimations for the Lyapunov exponent have been given in magnetic billiards with a magnetic field less than unity on the hyperbolic plane. Theorem 9 treats dispersing billiards while theorem 12 deals with Bunimovich-like ones, and expresses a result similar to that of Wojtkowsky,[21] but generalized for magnetic billiards and for spaces of constant negative curvature, and proven with other methods. (See also lemma 10.)

ACKNOWLEDGMENTS

The author is indebted to Professor P. Szépfalusy for proposing this investigation, for many discussions, suggestions, and for critically reading the manuscript. The author is also thankful to Professor D. Szász for valuable discussions and information about the related mathematical literature.

This work has been supported by the Ministry for Culture and Education, Hungary, under Grant No. MKM 337.

APPENDIX A: BASIC RELATIONSHIPS FOR ARCS AND CHORDS OF CYCLES

In this appendix some useful connections are summarized, concerning the arc and chord lengths (and other specific parameters) of segments of curves of constant curvature on manifolds of curvature $r \in \{1,0,-1\}$. All of these formulas can be derived by straightforward, elementary calculations.

In figure 15 S and B denote the endpoints of the investigated segment of the cycle γ, L is the length of the arc SB while l is the length of the (geodesic) chord SB. From this point different cases are treated separately.

If γ is a *circle* of center O and radius R (i.e., $r+\beta^2>0$), then let ϕ be the central angle of its arc SB, and let δ denote the angle between the chord SB and the radii OS, OB (figure 15/A). Depending on the curvature r of the manifold the following relationships are valid.

On the sphere ($r=1$):

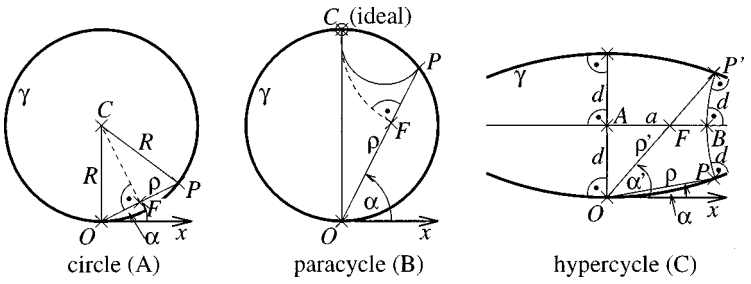

FIG. 16. Polar coordinates for curves of constant curvature.

$$L=\phi\ \sin(R),\quad \cos(l)=\cos^2(R)+\sin^2(R)\cos(\phi),$$

$$\tan(\delta)=\frac{1}{\cos(R)\tan(\phi/2)}=\frac{1+\cos(\phi)}{\sin(\phi)\cos(R)}. \tag{A1}$$

On the Euclidean plane ($r=0$):

$$L=\phi R,\quad l=2R\ \sin(\phi/2),\quad \delta=\frac{\pi-\phi}{2}. \tag{A2}$$

On the hyperbolic plane ($r=-1$, $|\beta|>1$):

$$L=\phi\ \sinh(R),\quad \cosh(l)=\cosh^2(R)-\sinh^2(R)\cos(\phi),$$

$$\tan(\delta)=\frac{1}{\cosh(R)\tan(\phi/2)}=\frac{1+\cos(\phi)}{\sin(\phi)\cosh(R)}. \tag{A3}$$

If γ is a *paracycle* (i.e., $r=-1$, $|\beta|=1$) then O is an ideal point, $R=\infty$, $\phi=0$ (figure 15/B), and the following equations are satisfied:

$$L=\sqrt{2(\cosh(l)-1)},\quad \tan(\delta)=\frac{1}{\sinh(l/2)}. \tag{A4}$$

Finally, if γ is a *hypercycle* (i.e., $r+\beta^2<0$) then the altitude d and the length $a=|S'B'|$ on the base line of the hypercycle are used to describe the geometry of the arc. δ denotes the angle between the (geodesic) chord SB and the altitudes $S'S$, $B'B$ at the endpoints of the chord (figure 15/C). The following connections are fulfilled:

$$L=a\ \cosh(d),\quad \cosh\ l=\cosh^2(d)\cosh(a)-\sinh^2(d),$$

$$\tan(\delta)=\frac{1}{\sinh(d)\tanh(a/2)}=\frac{1+\cosh(a)}{\sinh(d)\sinh(a)}. \tag{A5}$$

APPENDIX B: THE POLAR EQUATIONS OF CURVES OF CONSTANT CURVATURE

Let O be an arbitrary point of the manifold M of curvature $r\in\{1,0,-1\}$, and let x be a (geodesic) ray with origin O. The position of an arbitrary point $P\in M$ is characterized by its polar coordinates $(\rho,\alpha)\in\mathbb{R}_0^+\times[0,2\pi)$, where ρ is the (geodesic) distance from O to P, and α is the angle between the rays x and OP. In this appendix the polar equations of the cycles of M tangential to x at O are summarized (figure 16). (These formulas can be obtained using elementary trigonometric connections and the results of the previous appendix.)

The case of *circles* (for arbitrary r) and *paracycles* can be treated simultaneously; let γ be a circle (or paracycle) of center C, radius R, curvature q, which is tangential to x at O, and let P be an arbitrary point of its circumference with polar coordinates (ρ,α) (figures 16/A, B). (If γ is a

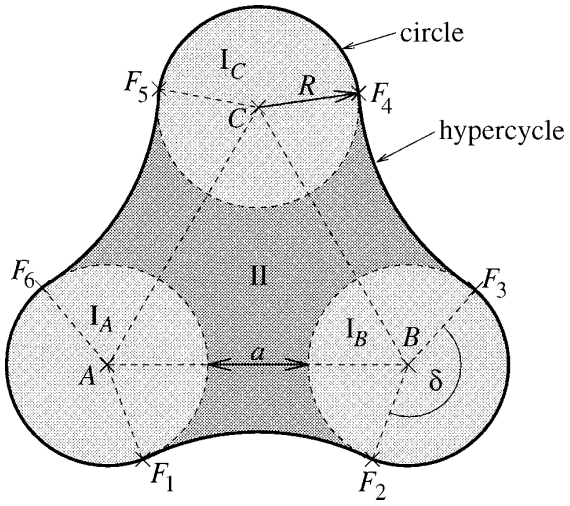

FIG. 17. The triangular billiard with focusing and dispersing boundary segments.

paracycle, then C is an ideal point, $R=\infty$, $q=1$.) Using the relations between the curvature q and the radius R [equation (2), with $\beta=q$], and considering the right triangle $OFC\Delta$ the following equations can be derived:

$$\frac{\sin(\alpha)}{q}=\begin{cases}\tan(\rho/2), & \text{if } r=1 \text{ (on the sphere);}\\ \rho/2, & \text{if } r=0 \text{ (on the Euclidean plane);}\\ \tanh(\rho/2), & \text{if } r=-1, \quad q\geqslant 1 \text{ (on the hyperbolic plane).}\end{cases} \tag{B1}$$

In the case of *hypercycles* the two branches of the curve have to be treated separately. Let γ be a hypercycle of altitude d, curvature $0<q<1$, which is tangential to the ray x at the origin O, and let P, $P'\in\gamma$ be arbitrary, opposite lying points with polar coordinates (ρ,α), (ρ',α'), respectively, on different branches of the hypercycle, as it is indicated in Figure 16(C). (The segment PP' meets the base line at right angle in B.) Using the formulas (A5) (with $l=\rho$, $\delta=\pi/2-\alpha$) and trigonometric relations for the right triangle $OAF\Delta$ (Figure 16(C)) the following equations can be derived:

$$\frac{\sin(\alpha)}{q}=\begin{cases}\tanh(\rho/2), & \text{if } r=-1, \quad |\sin(\alpha)/q|<1 \text{ (for the branch of } O);\\ \coth(\rho/2), & \text{if } r=-1, \quad |\sin(\alpha)/q|>1 \text{ (for the branch avoiding } O).\end{cases} \tag{B2}$$

(The primes on ρ, α have been omitted.)

The previous results can be summarized in the following statement.

Lemma 13: The equation (B1) (for $r+q^2\geqslant 0$) or (B2) (for $r+q^2<0$) holds for the length $\rho=|OO'|$ of the chord OO' of the curve γ of constant curvature q (on a manifold of curvature $r\in\{1,0,-1\}$), where α is the angle between the (geodesic) chord OO' and the curve γ.

APPENDIX C: AN APPLICATION OF THEOREM 12

In this appendix an example of the use of Theorem 12 (in Section V) is given, only for demonstrating purposes. The billiard domain, bounded by three focusing circular arcs and three dispersing segments of hypercycles can be considered as the triangular version of the Bunimovich stadium billiard on the hyperbolic plane (Figure 17). The centers of the circles of radius R lie at the vertices of the equilateral triangle $ABC\Delta$, and the hypercycle segments of curvature q are attached tangentially to the circles at the points $F_1,F_2,\ldots,F_6$ in such a way that the angle δ subtended by the circular boundary arcs at the centers of the circles are less than the straight angle. (Otherwise the particle bouncing back and forth along a diameter could be trapped in one circle for an infinitely long time.)

Let us divide the billiard domain into four regions, three of them being the disks I_A, I_B, I_C belonging to the circles, and the fourth is the rest, denoted by II. Let a be the minimal distance between two disks, and let $L:=\delta\sinh(R)$ be the length of the circular boundary arcs [equation (A3)]. It is evident that for magnetic fields $\beta<q$ the billiard particle (moving at unit speed) spends at most time L (in the limit case sliding along the boundary) inside a disk region of type I, and then passes at least time a (mostly) in region II, bouncing only on dispersive wall segments before entering again one of the disks and preforming a reflection on its focusing boundary. (It is possible that a bounceless trajectory segment intersects one of the disks, though both of its endpoints lie on a hypercycle segment. In this case the particle is not considered to have entered the disk region of type I.) Since the time elapsed outside the disks contributes entirely to the reduced time, theorem 12 can be applied with the constant $K=a/(a+L)$, and the estimation

$$\lambda\geqslant\frac{a}{a+L}\sqrt{1-\beta^2} \tag{C1}$$

can be obtained for the Lyapunov exponent λ, where the magnetic field $\beta<q<1$.

[1] Ya. G. Sinai, ''Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards,'' Russ. Math. Surv. **25**, 137–189 (1970); Usp. Mat. Nauk **25**, 141–192 (1970).
[2] L. A. Bunimovich, ''On the ergodic properties of some billiards,'' Funct. Anal. Appl. **8**, 73–74 (1974).
[3] L. A. Bunimovich, ''On the ergodic properties of nowhere dispersing billiards,'' Commun. Math. Phys. **65**, 295–312 (1979).
[4] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai, ''Ergodic theory,'' Vol. 245 of *Grundlehren der Matematischen Wissenschaften* (Springer-Verlag, Berlin, 1982).
[5] M. C. Gutzwiller, ''Chaos in classical and quantum mechanics,'' Vol. 1 of *Interdisciplinary Applied Mathematics* (Springer-Verlag, Berlin, 1990).
[6] J. Hadamard, ''Les surfaces à courbures opposées et leurs lignes geodésiques,'' J. Math. pure Appl. **5**, 27–73 (1898).
[7] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, 1978).
[8] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics* (W. Benjamin, New York, 1968), Chap. 3, pp. 53–79.
[9] E. Artin, ''Ein mechanisches System mit quasiergodischen Bahnen,'' Hamburger Math. Abh. **3**, 499–503 (1924); in *Collected Papers* (Addison-Wesley, Reading, MA, 1965), pp. 499–501.
[10] N. L. Balazs and A. Voros, ''Chaos on the pseudosphere,'' Phys. Rep. **143**, 109–240 (1986).
[11] A. Csordás, R. Graham, P. Szépfalusy, and G. Vattay, ''Transition from Poissonian to Gaussian-orthogonal-ensemble level statistics in a modified Artin's billiard,'' Phys. Rev. E **49**, 325–332 (1994).
[12] T. Tasnádi, ''The behaviour of nearby trajectories in magnetic billiards,'' J. Math. Phys. **37**, 5577–5598 (1996).
[13] T. Tasnádi, ''Hard chaos in magnetic billiards (On the Euclidean plane),'' Commun. Math. Phys. **187**, 597–621 (1997).
[14] W. Breymann, Z. Kovács, and T. Tél, ''Chaotic scattering in the presence of an external magnetic field,'' Phys. Rev. E **50**, 1994–2006 (1994).
[15] M. Robnik and M. V. Berry, ''Classical billiards in magnetic fields,'' J. Phys. A **18**, 1361–1378 (1985).
[16] O. Meplan, F. Brut, and C. Gignoux, ''Tangent map for classical billiards in magnetic fields,'' J. Phys. A **26**, 237–246 (1993).
[17] N. Berglund and H. Kunz, ''Integrability and ergodicity of classical billiards in a magnetic field,'' J. Stat. Phys. **83**, 81–126 (1996).
[18] G. M. Zaslavskiĭ and N. N. Filonenko, ''Statistical properties of the energy spectrum of 'gliding' electrons with mixed classical trajectories,'' Sov. Phys. JETP **38**, 317–323 (1974); Zh. Eksp. Teor. Fiz. **65**, 643–656 (1973).
[19] A. Comtet, ''On the Landau levels on the hyperbolic plane,'' Ann. Phys. (N.Y.) **173**, 185–209 (1987).
[20] G. Casati and H. A. Cerdeira, in *Chaos in Mesoscopic Systems*, Proceedings of the Miniworkshop on Nonlinearity: Chaos in Mesoscopic systems & the Adriatico Research Conference on Mesoscopic Systems and Chaos: A Novel Approach (World Scientific, Singapore, 1995).
[21] M. Wojtkowsky, ''Principles for the design of billiards with nonvanishing Lyapunov exponents,'' Commun. Math. Phys. **105**, 391–414 (1986).
[22] R. Markarian, ''Billiards with Pesin region of measure one,'' Commun. Math. Phys. **118**, 87–97 (1988).
[23] V. J. Donnay, ''Using integrability to produce chaos: Billiards with positive entropy,'' Commun. Math. Phys. **141**, 225–257 (1991).
[24] C. Series, ''The modular surface and continued fractions,'' J. London Math. Soc. **2**, 69–80 (1985).
[25] M-J. Giannoni and D. Ullmo, ''Coding chaotic billiards,'' Physica **41**, 371–390 (1990).
[26] E. Cornelis and M. Wojtkowsky, ''A criterion for the positivity of the Lyapunov characteristic exponent,'' Ergod. Theor. Dyn. Syst. **4**, 527–539 (1984).
[27] M. Wojtkowsky, ''Invariant families of cones and Lyapunov exponents,'' Ergod. Theor. Dyn. Syst. **5**, 145–161 (1985).
[28] V. I. Oseledec, ''The multiplicative ergodic theorem. The Lyapunov characteristic numbers of a dynamical system,'' Trans. Moscow Math. Soc. **19**, 197–231 (1968).
[29] Ya. B. Pesin, ''Lyapunov characteristic exponents and smooth ergodic theory,'' Russ. Math. Surveys **32**, 55–114 (1977).
[30] A. Katok and J.-M. Strelcyn, ''Invariant manifolds, entropy and billiards; smooth maps with singularities,'' Vol. 1222 of *Lecture Notes in Mathematics* (Springer-Verlag, Berlin, 1986), in collaboration with F. Ledrappier and F. Przytycki.
[31] D. Szász, ''On the *K*-property of some planar hyperbolic billiards,'' Commun. Math. Phys. **145**, 595–604 (1992).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 7 JULY 1998

Canonical general relativity: The diffeomorphism constraints and spatial frame transformations

M. A. Clayton[a)]
CERN–Theory Division, CH-1211, Geneva 23, Switzerland

(Received 30 July 1997; accepted for publication 3 April 1998)

Einstein's general relativity with both metric and vielbein treated as independent fields is considered, demonstrating the existence of a consistent variational principle and deriving a Hamiltonian formalism that treats the spatial metric and spatial vielbein as canonical coordinates. This results in a Hamiltonian in standard form that consists of Hamiltonian and momentum constraints as well as constraints that generate spatial frame transformations—all appearing as primary, first class constraints on phase space. The formalism encompasses the standard coordinate frame and vielbein approaches to general relativity, and the constraint algebra derived herein reproduces known results in either limit. © *1998 American Institute of Physics.* [S0022-2488(98)01507-2]

I. INTRODUCTION

A recent review of actions for general relativity (GR)[1] describes formulations in coordinate and (orthogonal) vielbein frames from variations of Einstein-Hilbert and Einstein-Palatini actions. Notably absent is an action that encompasses both of these cases. Such an action possesses the full general linear group invariance, as well as the usual diffeomorphism invariance, allowing the ''limit'' to orthogonal vielbein and coordinate frame approaches as different choices of gauge. This structure is represented on the full arbitrariness of the choice of vielbein (16 functions) as well as those of the spacetime metric (10 functions). That one can construct such an action is perhaps no surprise,[2] however we will also construct a Hamiltonian for the system in which the spatial $\mathrm{GL}(3,\mathbb{R})$ invariance is represented by infinitesimal generators on phase space. The resulting constraint algebra is derived, complete with conditions that determine surface contributions that guarantee that the field equations of GR follow properly from the Hamiltonian.[3]

As is always the case in such constructions, the Hamiltonian is far from unique. We will instead be focusing on the combined algebra of infinitesimal diffeomorphisms and $\mathrm{GL}(3,\mathbb{R})$ transformations which is uniquely determined once the parameterization of phase space and the action of the generations is fixed. In an earlier work[4] this algebra was derived through a generalization of the geometric argument of Teitelboim,[5,6] which led to a ''derivation'' of canonical GR.[7] This generalized algebra has appeared previously in the examination of (orthonormal) tetrad GR[8,9] and here we extend this type of analysis to the more general case, including an arbitrary tetrad as well as metric degrees of freedom.

It is noteworthy that we are not attempting to implement the full set of $\mathrm{GL}(4,\mathbb{R})$ generators,[9,10] nor the full set of spacetime diffeomorphisms,[11,12] instead we view the problem as purely geometrodynamic. Initial data consisting of the components of the spatial metric and vielbein that satisfies the constraints are given on the initial hypersurface, we determine the infinitesimal transformations of this data that lead to an equivalent physical problem (frame rotations and spatial diffeomorphisms), and then determine the conditions on the generator of time evolution that guarantee that the evolution of the system is consistent with spacetime diffeomorphism invariance.

Inevitably we end up with a system with a higher degree of redundancy (30 phase space degrees of freedom and 13 primary, first class constraints) and the possibility of new choices of gauge that are useful for numerical relativity exists; this will not be considered in this work. Ultimately we are interested in exploring the relationship between diffeomorphism invariance, the strong equivalence principle, and the evolution of quantum systems. In particular, whether it is

a)Electronic mail: Michael.A.Clayton@cern.ch

 © 1998 American Institute of Physics

possible to specify some type of quantum evolution that respects diffeomorphism invariance without the need to introduce an (approximate) timelike killing vector as is usually necessary.[13,14] We would also like to quantify the noninvariance of the conventional curved spacetime field theory, towards an understanding of which results may survive in a more fundamental theory based on the strong equivalence principle, and which depend strongly on the choice of frame.

II. THE GENERAL FRAME CONSTRUCTION

The natural setting for this work is the space of moving (vielbein) frames above a manifold **M**, that is, the space of smooth assignments of a frame of reference above each point, equivalent to smooth sections of the general linear frame bundle $GL\mathbf{M}$. Physical fields erected above **M** (the metric, curvature, matter fields, etc., ...) are then associated to this bundle through some representation of $\mathrm{GL}(n,\mathbb{R})$, and components may be defined in terms of said frame. Standard presentations of classical general relativity often adopt coordinate (or holonomic) frames from the outset, introducing moving (or vielbein) frames as a ''generalized'' concept in Riemannian geometry. Even when such frames are given equal footing with holonomic frames, in the variational principle coordinate components are once again given precedence. In order to demonstrate that this is unnecessary, we proceed to review some results from Riemannian geometry in a general linear frame.

A. Moving frames

A moving frame $\{e_A\}$ above a point in **M** will be written in terms of a particular coordinate frame as $e_A := E^\mu{}_A \partial_\mu$ and the coframe that is dual to it as $\theta^A := E^A{}_\mu dx^\mu$, where the duality relation $\theta^A[e_B] = \delta^A_B$ implies that the vielbeins satisfy $E^\mu{}_A E^A{}_\nu = \delta^\mu_\nu$ and $E^\mu{}_B E^A{}_\mu = \delta^A_B$. In this section we consider frames that are chosen to be a smooth section of the general linear frame bundle $GL\mathbf{M}$, however, in Section III we will consider spatial frames defined by $e_a := E^i{}_a \partial_i$, where ∂_i are the partial derivatives with respect to coordinates on a spatial hypersurface.

The volume form in a linear frame is given by[15]

$$\int_{\mathbf{M}} *\mathbf{1} = \int_{\mathbf{M}} \sqrt{|g|}\, \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3 = \int_{\mathbf{M}} d^4x \sqrt{|g|} E, \tag{1}$$

the last of which will be used here. Using compatibility and vanishing torsion, Gauss' law takes the form

$$\int_R dx E \sqrt{-g} \nabla_A [V]^A = \int_R dx \partial_\mu [\underline{E}^\mu{}_A \mathbf{V}^A] = \int_{\partial R} dS_\mu \underline{E}^\mu{}_A \mathbf{V}^A = \int_{\partial R} d\underline{S}_A \mathbf{V}^A, \tag{2}$$

where the surface measure is defined in terms of the normal vector n^A to the boundary ∂R of the region $R \subset \mathbf{M}$ by $S_A := *\mathbf{1}[n]_A$. We have also made the definition $\underline{T} := ET$ for tensors weighted by the vielbein density $E := \det[E^A{}_\mu]$, and $\mathbf{T} := \sqrt{|g|} T$ for tensors weighted by the spatial metric density $\sqrt{|g|} := \sqrt{|\det[g_{AB}]|}$.

Under the change of frame determined by $M^A{}_B \in \mathrm{GL}(n,\mathbb{R})$, the frame and coframe transform as $\theta^A \to \theta^B = M^B{}_A \theta^A$, $e_A \to e_B = e_A |M^{-1}|^A{}_B$, where $M^A{}_C |M^{-1}|^C{}_B = |M^{-1}|^A{}_C M^C{}_B = \delta^A_B$, and similarly for the components of tensors. The infinitesimal forms $M^A{}_B = \delta^A_B + \omega^A{}_B$, and $|M^{-1}|^A{}_B = \delta^A_B - \omega^A{}_B$ define the generators $\Delta^A{}_B$ of $\mathfrak{gl}(n,\mathbb{R})$, acting, for example, on vectors, covectors and both types of density as

$$\Delta_{\tilde{\omega}}[T]_A = -\omega^B{}_A T_B, \quad \Delta_{\tilde{\omega}}[T]^A = \omega^A{}_B T^B, \quad \Delta_{\tilde{\omega}}[\gamma] = -\omega^A{}_A \sqrt{\gamma}, \quad \Delta_{\tilde{\omega}}[E] = \omega^A{}_A E, \tag{3}$$

where $\tilde{\omega}$ represents the matrix $\omega^A_B \in \mathrm{GL}(n,\mathbb{R})$. These generators satisfy the Lie algebra of $\mathfrak{gl}(n,\mathbb{R})$, given by the matrix commutator[16]

$$[\Delta_{\tilde{\omega}_1}, \Delta_{\tilde{\omega}_2}] = -\Delta_{[\widetilde{\omega_1,\omega_2}]}, \tag{4}$$

where $[\omega_1,\omega_2]^A{}_B := \omega_1{}^A{}_C \omega_2{}^C{}_B - \omega_1{}^C{}_B \omega_2{}^A{}_C$.

Once one has a definition of parallel transport, the covariant derivative operator is defined. We will assume that it acts on the components of tensors as, for example, $\nabla_A[V]^B = e_A[V^B] + \Gamma^B_{AC} V^C$, and we may then define what it means for the components of the metric g_{AB} to be compatible with this connection:

$$\nabla_A[g]_{BC} = e_A[g_{BC}] - g_{DC}\Gamma^D_{AB} - g_{BD}\Gamma^D_{AC} = 0. \tag{5}$$

From the nonholonomic nature of the frame $\{e_A\}$ we have the structure constants[17]

$$[e_A, e_B] = C_{AB}{}^C e_C, \quad C_{AB}{}^C = E^C{}_\mu (e_A[E^\mu{}_B] - e_B[E^\mu{}_A]), \tag{6}$$

and the vanishing of the torsion tensor $T(X,Y) := \nabla_X[Y] - \nabla_Y[X] - [X,Y]$ ($\nabla_X := X^A \nabla_A$ and $[X,Y] := \mathscr{L}_X[Y]$) results in the relationship between $\Gamma^A_{[BC]}$ and the structure constants (6),

$$T^A_{BC} = 2\Gamma^A_{[BC]} - C_{BC}{}^A = 0. \tag{7}$$

Throughout we denote (anti-)symmetrization by [ ] and ( ), respectively, i.e., $T_{[AB]} := 1/2(T_{AB} - T_{BA})$ and $T_{(AB)} := 1/2(T_{AB} + T_{BA})$. This, combined with (5), allows the explicit solution of Γ^A_{BC} in terms of metric and vielbein components:

$$\Gamma^A_{BC} = \tfrac{1}{2} g^{AD}(e_B[g_{CD}] + e_C[g_{DB}] - e_D[g_{BC}]) + \tfrac{1}{2} g^{AD}(g_{BE} C_{DC}{}^E + g_{CE} C_{DB}{}^E) + \tfrac{1}{2} C_{BC}{}^A, \tag{8}$$

a combination of the standard coordinate frame and orthogonal frame results. We will also require the contraction of the Riemann curvature tensor: $R^A_{BCD} = e_C[\Gamma^A_{DB}] - e_D[\Gamma^A_{CB}] + \Gamma^E_{DB}\Gamma^A_{CE} - \Gamma^E_{CB}\Gamma^A_{DE} - C_{CD}{}^E \Gamma^A_{EB}$ to the Ricci tensor:

$$R_{AB} := e_C[\Gamma^C_{BA}] - e_B[\Gamma^C_{CA}] + \Gamma^D_{BA}\Gamma^C_{CD} - \Gamma^D_{CA}\Gamma^C_{BD} - C_{CB}{}^D \Gamma^C_{DA}. \tag{9}$$

B. The Einstein-Hilbert action in a general linear frame

A direct translation of the action for GR results in (we use $16\pi G = c = 1$ throughout)

$$S_{\mathrm{GR}} = -\int_{\mathbf{M}} d^4x E \sqrt{|g|} g^{AB} R_{AB}, \tag{10}$$

where the spacetime is now described by a section of $Riem_{\mathbf{M}} \times GL\mathbf{M}$, namely, a Riemannian metric and nondegenerate linear frame above $\mathbf{M}$.

In a variation of the connection coefficients in R_{AB} [i.e., not considering the variation of the structure constant in the definition (9)] we find that $\delta_\Gamma R_{AB} = \nabla_C[\delta\Gamma]^C_{BA} - \nabla_B[\delta\Gamma]^C_{CA}$, and therefore treating the components $\Gamma^A_{(BC)}$ as independent leads via a Palatini variation to the compatibility conditions (5),

$$\frac{\delta S_{\mathrm{GR}}}{\delta \Gamma^C_{(AB)}} = -E(\nabla_{e_C}[\mathbf{g}]^{BA} - \nabla_{e_D}[\mathbf{g}]^{(AD}\delta^{B)}_C) = 0. \tag{11}$$

Note that the action is not set up to generate the torsion-free conditions (7); the variations $\delta\Gamma^A_{[BC]}$ must be determined in terms of the vielbein degrees of freedom using (7) and the variation of (6):

$$\delta C_{AB}{}^C = \nabla_A[E^C{}_\mu \delta E^\mu{}_B] - \nabla_B[E^C{}_\mu \delta E^\mu{}_A] + \delta E^\mu{}_A E^D{}_\mu \Gamma^C_{DB} - \delta E^\mu{}_B E^D{}_\mu \Gamma^C_{DA}. \tag{12}$$

To perform the Einstein-Hilbert variation, requires the variation of (8),

$$\delta\Gamma^A_{BC} = \delta E^\mu{}_B E^D{}_\mu \Gamma^A_{DC} + \tfrac{1}{2}\nabla_B[g^{AD}\delta g_{CD} + E^A{}_\mu \delta E^\mu{}_C - g^{AD} g_{CE} E^E{}_\mu \delta E^\mu{}_D]$$
$$+ \tfrac{1}{2}\nabla_C[g^{AD}\delta g_{DB} - E^A{}_\mu \delta E^\mu{}_B - g^{AD} g_{BE} E^E{}_\mu \delta E^\mu{}_D]$$
$$- \tfrac{1}{2} g^{AD}\nabla_D[\delta g_{BC} - g_{BE} E^E{}_\mu \delta E^\mu{}_C - g_{CE} E^E{}_\mu \delta E^\mu{}_B], \tag{13}$$

where we note that the presence of the first term, although implying that $\delta\Gamma^A_{BC}$ is no longer a tensor, is precisely what guarantees that the variation of a covariant derivative will remain covariant.

The variation of (10) will be performed by treating the densitized components of the inverse metric $\mathbf{g}^{AB}$ and the frame degrees of freedom $\underset{\sim}{E}{}^\mu{}_A$ as independent variables, using

$$\delta\underset{\sim}{\mathbf{R}} = \underset{\sim}{R}_{AB}\delta\mathbf{g}^{AB} + \left[2\mathbf{R}^A{}_B E^B{}_\mu - \frac{1}{n-1}\mathbf{R}E^A{}_\mu\right]\delta\underset{\sim}{E}{}^\mu{}_A - E\nabla_A\nabla_B[\delta\mathbf{S}^{AB}], \tag{14}$$

where $\delta\mathbf{S}^{AB} := \delta\mathbf{g}^{AB} + g^{AB}\delta\sqrt{|g|} + 2\mathbf{g}^{AC}E^B{}_\mu\delta E^\mu{}_C - 2\mathbf{g}^{AB}E^C{}_\mu\delta E^\mu{}_C$. In computing this we have left the dimensionality n of spacetime arbitrary and noted that $\delta\sqrt{-g} = g_{AB}\delta\mathbf{g}^{AB}/(n-2)$ and $\delta E = E^A{}_\mu \delta\underset{\sim}{E}{}^\mu{}_A/(n-1)$. Discarding surface terms we find

$$\frac{\delta S_{\mathrm{GR}}}{\delta\mathbf{g}^{AB}} = \underset{\sim}{R}_{AB} = 0, \quad \frac{\delta S_{\mathrm{GR}}}{\delta\underset{\sim}{E}{}^\mu{}_A} = 2\mathbf{R}^A{}_B E^B{}_\mu - \frac{1}{n-1}\mathbf{R}E^A{}_\mu = 0, \tag{15}$$

which are Einstein's equations in empty spacetime. The equivalence of the variational results from the metric and vielbein degrees of freedom is expected algebraically due to an argument by Floreanini and Percacci,[10] and also since we know that Einstein's equations are covariant under frame transformations. This variational principle is actually a specific case of the more general formalism that includes affine frames, nonmetricity and torsion that appears in Ref. 2.

III. HAMILTONIAN FORMALISM

We turn now to the construction of a Hamiltonian formalism for the system, much of which is a straightforward application of the Bergmann-Dirac procedure for constrained systems described in detail in Ref. 18 (which we will follow). In order to consider the initial value problem, we must consider the embedding of a family of nonoverlapping, spacelike hypersurfaces Σ_t that foliate $\mathbf{M}$, labeled by some choice of time parameter. The geometry of this scenario has been considered extensively by Kuchař;[19] here we will give results that will be of some use in this work, ignoring global issues (other than the addition of surface contributions to the Hamiltonian) and assuming in all cases that a global section of the frame bundle $GL\mathbf{M}$ exists.

A. The surface-adapted basis

The spacetime metric (and inverse, respectively) may be put in surface-normal form (note the -2 signature of the spacetime metric),

$$g = \theta^\perp\otimes\theta^\perp - \gamma_{ab}\theta^a\otimes\theta^b, \quad g^{-1} = e_\perp\otimes e_\perp - \gamma^{ab}e_a\otimes e_b, \tag{16}$$

where the frame and its dual coframe are given by

$$e_\perp = \frac{1}{N}\partial_t - \frac{N^a}{N}e_a, \quad e_a = E^i{}_a\partial_i, \quad \theta^\perp = Ndt, \quad \theta^a = N^a dt + E^a{}_i dx^i. \tag{17}$$

The atlas fields N and N^a (the lapse function and shift vector, respectively) play the same geometric role as that assigned to them in coordinate frame work,[18,20] namely, the shift vector is the projection of the spacetime vector field that describes the ''flow of time'' parallel to Σ, and the lapse function the normal projection. (Note that they may either be thought of as a reparameterization of the spacetime metric, or as the $E^A{}_0$ components of the vielbein in a gauge where $E^0{}_i = 0$ in tetrad GR.) The only alteration of this conventional picture is that N^a are the components of the shift vector in the frame $\{e_a\}$ above Σ. It is straightforward to see that $E^{(4)}\sqrt{|g|} = NE\sqrt{\gamma}$

where we distinguish here between the four-dimensional determinant $E^{(4)}$ and from now on write $E \coloneqq \det E^a{}_i$ as representing the determinant of the spatial vielbein. We will also occasionally refer to a spatial vector as, for example $\vec{N} \coloneqq N^a e_a$.

For this surface-adapted frame we find the nonvanishing structure constants from (6),

$$[e_a, e_b] = C_{ab}{}^c e_c, \quad [e_\perp, e_a] = C_{\perp a}{}^\perp e_\perp + C_{\perp a}{}^b e_b, \quad C_{\perp a}{}^\perp = e_a[\ln N], \tag{18a}$$

$$C_{ab}{}^c = E^c{}_i (e_a[E^i{}_b] - e_b[E^i{}_a]), \quad C_{\perp a}{}^b = \frac{e_a[N^b]}{N} + \frac{N^c}{N} C_{ac}{}^b + \frac{1}{N} E^b{}_i \partial_t[E^i{}_a]. \tag{18b}$$

Taking perpendicular and parallel projections of the compatibility conditions (5) results in $\Gamma^\perp_{\perp\perp} = \Gamma^\perp_{a\perp} = 0$, $\Gamma^\perp_{\perp a} \coloneqq a_a$, $\Gamma^a_{\perp\perp} = a^a \coloneqq \gamma^{ab} a_b$, $\Gamma^\perp_{ab} \coloneqq K_{ab}$, $\Gamma^a_{b\perp} = K^a{}_b \coloneqq \gamma^{ac} K_{bc}$ and the surface compatibility conditions $\boldsymbol{\nabla}_a[\gamma]_{bc} = 0$, where $\boldsymbol{\nabla}_a$ is the covariant derivative operator defined on Σ. The remaining compatibility condition gives the relationship between the extrinsic curvature and the time derivatives of the metric and vielbein degrees of freedom,

$$\partial_t[\gamma_{ab}] - 2\gamma_{(ac} E^c{}_i \partial_t[E^i{}_{b)}] - 2NK_{ab} - \mathcal{L}_{\vec{N}}[\gamma]_{ab} = 0. \tag{19}$$

The projections of (7) ($T^\perp_{\perp a}$, $T^\perp_{ab}$, $T^a_{\perp b}$, and T^a_{bc}, respectively) result in $a_a = e_a[\ln N]$, $K_{[ab]} = 0$, $\Gamma^a_{\perp b} = K^a{}_b + C_{\perp b}{}^a$ and $\Gamma^a_{[bc]} = C_{bc}{}^a$, which completes the generalization of the results of Ref. 18.

The projected components of the Ricci tensor (9) that will appear in the surface reduction of (10) are (the operator $d_\perp$ is defined in the following section)

$$R^{(4)}_{\perp\perp} = -d_\perp[K] + \boldsymbol{\nabla}_a[a]^a + a^a a_a - K^a{}_b K^b{}_a, \tag{20a}$$

$$R^{(4)}_{ab} = R_{ab} + d_\perp[K]_{ab} - \boldsymbol{\nabla}_{(b}[a]_{a)} - a_a a_b + KK_{ab} - 2K_{ac} K^c{}_b, \tag{20b}$$

where $K \coloneqq K^a{}_a$ and we have written $R^{(4)}_{ab}$ to indicate the projected spacetime tensor and R_{ab} the intrinsically defined spatial Ricci tensor.

B. Surface-adapted derivatives

In order to consider the representation of diffeomorphisms in frames in which the metric is nondynamical (or prescribed in some manner), we need to consider in some detail the representation of spatial diffeomorphisms. The action of an infinitesimal diffeomorphism of Σ to itself generated by the vector field $\vec{X}$, may be written in an arbitrary frame as

$$\mathcal{L}_{\vec{X}}[T^{ab\cdots}_{mn\cdots}] \coloneqq \boldsymbol{\nabla}_{\vec{X}}[T]^{ab\cdots}_{mn\cdots} - \Delta_{\widetilde{\boldsymbol{\nabla}\vec{X}}}[T]^{ab\cdots}_{mn\cdots}, \tag{21a}$$

which is also applicable to densities: $\mathcal{L}_{\vec{X}}[\sqrt{\gamma}] = \boldsymbol{\nabla}_a[\mathbf{X}]^a$. In this form the outcome of the diffeomorphism is represented on the components of tensors as the sum of a covariant derivative $\boldsymbol{\nabla}_{\vec{X}} \coloneqq X^a \boldsymbol{\nabla}_a$ and a spatial frame rotation $[\widetilde{\boldsymbol{\nabla}\vec{X}}]^a_b \coloneqq \boldsymbol{\nabla}_b[X]^a$. The frame is unaffected: $\mathcal{L}_{\vec{X}}[e_a] = 0$.

This is not necessary though, and while appropriate for considering a fixed frame, is inappropriate for considering diffeomorphisms when constrained to an orthonormal frame since $\mathcal{L}_{\vec{X}}[\gamma]_{ab} = \boldsymbol{\nabla}_{(a}[X]_{b)} \neq 0$. Instead one may define a Lie action that operates on both the frame and tensor components as

$$\mathcal{L}'_{\vec{X}}[e_a] = \Delta_{\widetilde{\boldsymbol{\nabla}\vec{X}}}[e_a] = -\boldsymbol{\nabla}_a[X]^b e_b, \quad \mathcal{L}'_{\vec{X}}[T^{ab\cdots}_{mn\cdots}] = \boldsymbol{\nabla}_X[T]^{ab\cdots}_{mn\cdots}, \tag{21b}$$

so that when the action on a tensor (not just the components) is considered, one finds the same result as (21a). This representation is more appropriate for considering orthonormal frames since the action of $\mathcal{L}'$ on the components of the metric $\mathcal{L}'_{\vec{X}}[\gamma]_{ab} = \nabla_{\vec{X}}[\gamma]_{ab}$ vanishes due to compatibility. This is similar to how parallel transport in a principle bundle is transferred to associated bundles[17] since one could just as easily define the covariant derivative to act as $\boldsymbol{\nabla}_{\vec{X}}[T^{ab\cdots}_{mn\cdots}] = X^a e_a[T^{ab\cdots}_{mn\cdots}]$, and $\boldsymbol{\nabla}_{\vec{X}}[e_a] = X^b \Gamma^c_{ba} e_c$. This we do not pursue since the resulting action is not covariant with respect to general linear frame transformations when acting on frame or tensorial components alone.

Similarly, the definition of the derivative off of Σ that is surface-covariant is defined to act on the components of tensors as[18]

$$d_\perp[T^{ab\cdots}_{mn\cdots}] := e_\perp[T^{ab\cdots}_{mn\cdots}] + \Delta_{\tilde{C}_\perp}[T]^{ab\cdots}_{mn\cdots}, \tag{22}$$

where $\tilde{C}_\perp$ is the matrix with components $C_{\perp a}{}^b$ defined in (18). This operator defines the derivative normal to Σ that ''follows'' the vielbein (since $d_\perp[E^i{}_a]=0$ by definition), and describes the evolution of all quantities in terms of the original frame. (Note that the definition of $C_{\perp a}{}^b$ involves time derivatives of the frame.)

As with the case of the Lie derivative, we may also consider the opposite case, namely where the normal derivative is defined so that the metric does not ''evolve'' and the frame does. Let us introduce an operator that allows a more general evolution of the vielbein,

$$d'_\perp[E^i{}_a] := e_\perp[E^i{}_a] + \Delta_{\tilde{D}_\perp}[E^i{}_a] = e_\perp[E^i{}_a] - D_{\perp a}{}^b E^i{}_b, \tag{23}$$

where $\tilde{D}_\perp$ is determined by requiring that $d'_\perp[T^{ab\cdots}_{mn\cdots} e_a \otimes e_b \cdots \theta^m \otimes \theta^n \cdots] = d_\perp[T^{ab\cdots}_{mn\cdots} e_a \otimes e_b \cdots \theta^m \otimes \theta^n \cdots]$ (so that the operators act on tensors in an identical manner). We find that the action of $d'_\perp$ on the components of a tensor must be given by

$$d'_\perp[T^{ab\cdots}_{mn\cdots}] = e_\perp[T^{ab\cdots}_{mn\cdots}] + \Delta_{\tilde{C}'_\perp}[T^{ab\cdots}_{mn\cdots}], \quad \text{where} \quad C'_{\perp b}{}^a := C_{\perp b}{}^a - D_{\perp b}{}^a + E^a{}_i e_\perp[E^i{}_b]. \tag{24}$$

If we now require that $d'_\perp[\gamma]_{ab}=0$ (so that the normal derivative operator does not affect the components of the metric) then we may choose

$$D_{\perp b}{}^a = -\frac{1}{2N}\,\gamma^{ac}\partial_t[\gamma_{bc}] + \frac{1}{N}\,e_b[N^a] - \frac{1}{N}\,N^c E^a{}_i e_b[E^i{}_c], \tag{25}$$

and therefore $C'_{\perp b}{}^a = -1/2\gamma^{ac} e_\perp[\gamma_{bc}]$. (This is far from unique since any choice of $D_{\perp[ab]}$ will result in $d'_\perp[\gamma]_{ab}=0$, however, we have chosen what we consider to be a ''simplest choice'' which corresponds to the operator considered in Section III C as well as Ref. 4.) In terms of these operators, the compatibility condition (19) takes on either of the two forms

$$d_\perp[\gamma]_{ab} = 2K_{ab} \quad \text{or} \quad E^a{}_i d'_\perp[E^i{}_b] = -K^a{}_b. \tag{26}$$

Note that the total time derivative operators $d_t = N d_\perp + \mathscr{L}_{\vec{N}}$ and $d'_t = N d'_\perp + \mathscr{L}_{\vec{N}}$ correspond to the definition of the total derivative of the tensor, and are not equivalent to the time derivative of the components. For example, $d_t[X^a e_a] = d'_t[X^a e_a] = \partial_t[X^a e_a] \neq \partial_t[X^a] e_a$.

C. The Hamiltonian

From the results of the previous two sections, the GR action (10) is reduced to

$$S_{\mathrm{GR}} = \int dt \int_\Sigma d^3x \underline{\mathbf{N}}(d_\perp[K] + \gamma^{ab} d_\perp[K]_{ab} - 2(\boldsymbol{\nabla}_a[a]^a + a^a a_a) + R + K^2 - K^a{}_b K^b{}_a). \tag{27}$$

Using compatibility we find that $\boldsymbol{\nabla}_a[a]^a + a^a a_a = (\gamma^{ab}\boldsymbol{\nabla}_a\boldsymbol{\nabla}_b[N])/N$ is a surface term in the action which will be dropped, and furthermore that

$$\begin{aligned}\underline{\mathbf{N}}(d_\perp[K] + \gamma^{ab} d_\perp[K]_{ab}) = {} & 2\partial_t[\underline{\mathbf{K}}] - (\underline{K}_{ab} + \gamma_{ab}\underline{K})\partial_t[\boldsymbol{\gamma}^{ab}] - 2\mathbf{K}^a{}_b E^b{}_i \partial_t[\underline{E}^i{}_a] \\ & - 2\underline{\mathbf{N}}^a \boldsymbol{\nabla}_a[K] - 2E\boldsymbol{\nabla}_a[\mathbf{K}^a{}_b N^b] + 2\underline{\mathbf{N}}^a \boldsymbol{\nabla}_b[K]^b{}_a.\end{aligned} \tag{28}$$

Dropping the total time derivative and writing $S_{\mathrm{GR}} = \int dt L_{\mathrm{GR}}$, the Lagrangian is given by

$$\begin{aligned}L_{\mathrm{GR}} = \int_\Sigma d^3x [& -(\underline{K}_{ab} + \gamma_{ab}\underline{K})\partial_t[\boldsymbol{\gamma}^{ab}] - 2\mathbf{K}^a{}_b E^b{}_i \partial_t[\underline{E}^i{}_a] + \underline{\mathbf{N}}(R + K^2 - K^a{}_b K^b{}_a) \\ & + 2\underline{\mathbf{N}}^a(\boldsymbol{\nabla}_b[K]^b{}_a - \boldsymbol{\nabla}_a[K])].\end{aligned} \tag{29}$$

In this form the Lagrangian is most easily treated in Palatini form, that is, by considering the extrinsic curvature K_{ab} as a tensor of Lagrange multipliers that enforce (19). The configuration of the system at any instant in time is described by a section of $Riem_\Sigma \times GL\Sigma$, however, due to the form of (29) it is convenient to choose instead the densitized canonical coordinates $\boldsymbol{\gamma}^{ab}$ and $\underset{\sim}{E}{}^i{}_a$, and so we are considering sections of $\boldsymbol{Riem}_\Sigma \times \underline{GL\Sigma}$. Determining the conjugate momenta via

$$\underset{\sim}{\pi}_{ab} := \frac{\delta L_{\mathrm{GR}}}{\partial_t[\boldsymbol{\gamma}^{ab}]} = -(\underset{\sim}{K}_{ab} + \gamma_{ab}\underset{\sim}{K}), \quad \mathbf{p}^a{}_i := \frac{\delta L_{\mathrm{GR}}}{\partial_t[\underset{\sim}{E}{}^i{}_a]} = -2\mathbf{K}^a{}_b E^b{}_i, \tag{30}$$

the canonical coordinate-momentum pairs are

$$\{(Q^I, P_I)\} := \{(\boldsymbol{\gamma}^{ab}, \underset{\sim}{\pi}_{ab}), (\underset{\sim}{E}{}^i{}_a, \mathbf{p}^a{}_i)\}, \tag{31}$$

corresponding to the phase space $T^*(\boldsymbol{Riem}_\Sigma \times GL\Sigma)$. From the canonical symplectic form on this phase space comes the standard form of the Poisson bracket,

$$\{F, G\} := \int_\Sigma d^3x \sum_I \left(\frac{\delta F}{\delta Q^I(x)} \frac{\delta G}{\delta P_I(x)} - \frac{\delta F}{\delta P_I(x)} \frac{\delta G}{\delta Q^I(x)} \right), \tag{32}$$

where F and G are functions of the phase space variables. (The results herein could be globalized along the lines of Ref. 21, in which case we would introduce the related weak symplectic form on phase space, however, a local treatment is sufficient for this work.) In addition to the phase space coordinates we have the Lagrange multipliers $\{N, N^a, N^a{}_b, K_{ab}, \lambda^{ab}\}$ where the Hamiltonian determined from $H_{\mathrm{GR}} = \int_\Sigma dx \Sigma_I \dot{Q}^I P_I - L_{\mathrm{GR}}$ is written as

$$\begin{aligned} \mathscr{H}_{\mathrm{GR}} = {} & \underset{\sim}{\boldsymbol{\lambda}}^{ab}(K_{ab} - \overline{a}k_{ab} - (1-\overline{a})k'_{ab}) + N^b{}_a \cdot \underset{\sim}{\mathscr{T}}{}^a{}_b + \underset{\sim}{\mathbf{N}}(-R + K^a{}_b K^b{}_a - K^2) \\ & + 2\underset{\sim}{\mathbf{N}}^a \nabla_b[-K^b{}_a + \delta^b_a K], \end{aligned} \tag{33}$$

where we have defined

$$k_{ab} := -(\pi_{ab} - \tfrac{1}{4}\gamma_{ab}\pi), \quad k'_{ab} := -\tfrac{1}{2}\gamma_{(ac}p^c{}_i E^i{}_{b)}. \tag{34}$$

The tensor of constraints enforced by $N^a{}_b$ (defining the scalars $\pi := \gamma^{ab}\pi_{ab}$ and $p := p^a{}_i E^i{}_a$):

$$\cdot\underset{\sim}{\mathscr{T}}{}^a_b := 2\boldsymbol{\gamma}^{ac}\underset{\sim}{\pi}_{bc} - \underset{\sim}{\boldsymbol{\pi}}\delta^a_b - \mathbf{p}^a{}_i \underset{\sim}{E}{}^i{}_b + \underset{\sim}{\mathbf{p}}\delta^a_b, \tag{35}$$

guarantee the consistency of (30), and are uniquely specified by the requirement that they generate infinitesimal $\mathfrak{gl}(3,\mathbb{R})$ frame rotations on phase space, and it is straightforward to show that

$$\{\boldsymbol{\gamma}^{ab}, \cdot\underset{\sim}{\mathscr{T}}{}^c_d[\omega^d_c]\} = \Delta_{\tilde{\omega}}[\boldsymbol{\gamma}]^{ab}, \quad \{\underset{\sim}{\pi}_{ab}, \cdot\underset{\sim}{\mathscr{T}}{}^c_d[\omega^d_c]\} = \Delta_{\tilde{\omega}}[\underset{\sim}{\pi}]_{ab}, \tag{36a}$$

$$\{\underset{\sim}{E}{}^i{}_a, \cdot\underset{\sim}{\mathscr{T}}{}^b_c[\omega^c_b]\} = \Delta_{\tilde{\omega}}[\underset{\sim}{E}{}^i{}_a], \quad \{\mathbf{p}^a{}_i, \cdot\underset{\sim}{\mathscr{T}}{}^b_c[\omega^c_b]\} = \Delta_{\tilde{\omega}}[\mathbf{p}^a{}_i], \tag{36b}$$

where the infinitesimal $\mathrm{GL}(3,\mathbb{R})$ frame rotations $\Delta_{\tilde{\omega}}$ are defined in (3), and we have made use of the notation, for example $\cdot\underset{\sim}{\mathscr{T}}{}^b{}_a[\omega^a{}_b] := \int_\Sigma d^3x \omega^a{}_b \cdot\underset{\sim}{\mathscr{T}}{}^b{}_a$. These generators (the phase space representation of $\Delta^a{}_b$ of Ref. 4) satisfy the $\mathfrak{gl}(3,\mathbb{R})$ Lie algebra,

$$\{\cdot\underset{\sim}{\mathscr{T}}{}^a{}_b, \cdot\underset{\sim}{\mathscr{T}}{}^c{}_d[\omega^d{}_c]\} = \Delta_{\tilde{\omega}}[\cdot\underset{\sim}{\mathscr{T}}]^a{}_b, \tag{37}$$

strongly on phase space. This differs from (4) by a sign due to the fact that the operator Δ acts from the left while $\{\cdot, \mathscr{T}\}$ acts from the right. It is important to note that $\mathscr{T}^{ab}$ is not a symmetric tensor of constraints, in fact, it is precisely the antisymmetric components that generate SO(3) rotations in a local orthonormal frame.

The constraints imposed by λ^{ab} explicitly relate the extrinsic curvature to the conjugate momenta through

$$K_{ab} \approx \overline{a}k_{ab} + (1-\overline{a})k'_{ab}, \tag{38}$$

and in the form given, the variation of (33) with respect to K_{ab} would then determine λ^{ab} in terms of the other Lagrange multipliers as $\lambda_{ab} \approx -2N(K_{ab} - \gamma_{ab}K) - 2(\nabla_{(a}[N]_{b)} - \gamma_{ab}\nabla_c[N]^c$. It is rather inconvenient to carry around the Lagrange multipliers K_{ab} and λ^{ab} solely to deal with the ambiguity in determining K_{ab} from the conjugate momenta. This may be circumvented in a simple manner.

When passing to the Hamiltonian H_{GR}, we may replace each occurrence of the extrinsic curvature in the Hamiltonian with some combination of the canonical momenta consistent with (38) so that the only place K_{ab} will occur is in the constraint enforced by λ^{ab}. The variation of K_{ab} will then enforce $\lambda^{ab} \approx 0$, and neither K_{ab} nor λ^{ab} will play any further role and may therefore be dropped consistently. The Hamiltonian determined in this way is identical to that which would occur if one had replaced the extrinsic curvature by (19) in the Lagrangian (29). In this way we see that the GR Hamiltonian can always be written as

$$H_{\mathrm{GR}} = \int_\Sigma d^3x (N.\underline{\mathscr{H}} + N^a.\underline{\mathscr{H}}_a + N^a{}_b.\underline{\mathscr{T}}^b{}_a) + E_{\mathrm{GR}}, \tag{39a}$$

where

$$.\underline{\mathscr{H}} := -\underline{\mathbf{R}} + \mathbf{E}(K^a{}_b K^b{}_a - K^2), \quad .\underline{\mathscr{H}}_a := -2\mathbf{E}\nabla_b[K^b{}_a - \delta^b_a K], \tag{39b}$$

K_{ab} represents some linear combination of k_{ab} and k'_{ab} consistent with (38).

In order for these field equations to properly follow from the Hamiltonian, it was necessary to add the surface contributions E_{GR} to the Hamiltonian (39a), which are required to satisfy[3]

$$\delta E_{\mathrm{GR}} + S_R + S_k = 0, \tag{40a}$$

in order for the field equations to be properly recovered from the Hamiltonian, where

$$S_R := \int_{\partial\Sigma} d\underline{S}_a (N\boldsymbol{\nabla}_b[\delta \mathbf{S}^{ab}] - \boldsymbol{\nabla}_b[N]\delta \mathbf{S}^{ba}), \tag{40b}$$

comes from the variations of the surface Ricci curvature term, and ($H_{ab} := -2(K_{ab} - \gamma_{ab}K)$, and K_{ab} is again determined in terms of k_{ab} and k'_{ab})

$$\begin{aligned} S_k := & -\frac{1}{2}\int_{\partial\Sigma} d\underline{S}_a N^a \mathbf{H}^{bc}\delta\gamma_{bc} + \int_{\partial\Sigma} d\underline{S}_a N^b \delta \mathbf{H}^a{}_b \\ & + \int_{\partial\Sigma} d\underline{S}_a [N^a \mathbf{H}^c{}_b E^b{}_i \delta E^i{}_c - N^b \mathbf{H}^a{}_b E^c{}_i \delta E^i{}_c + N^b \mathbf{H}^c{}_b E^a{}_i \delta E^i{}_c + N_b \mathbf{H}^{ac} E^b{}_i \delta E^i{}_c], \end{aligned} \tag{40c}$$

comes from variations of the momentum constraint.

D. The constraint algebras

In this section we will consider two choices of the diffeomorphism constraints. First we will choose $.\underline{\mathscr{H}}$ and $.\underline{\mathscr{H}}_a$ to act on phase space as the operators $\mathscr{L}$ and $d_\perp$ from Sec. III B, and refer to the Hamiltonian constructed from these as H_{GR}. This corresponds to $\bar{a} = 1$ in (38), and K_{ab} is replaced everywhere by k_{ab} in (39) and (40). This choice will closely resemble the coordinate approach to canonical GR and has a straightforward coordinate frame gauge reduction. Since these constraints have a nontrivial action on the spatial metric, it is nontrivial to recover the limit where only orthonormal spatial vielbeins are considered. This is instead accomplished by considering a second set of generators $.\underline{\mathscr{H}}'$ and $.\underline{\mathscr{H}}'_a$ which act on phase space as $\mathscr{L}'$ and $d'_\perp$ from Sec. III B, and a second Hamiltonian H'_{GR} is defined by attaching primes to the constraints in (39). In this scheme K_{ab} is replaced everywhere by k'_{ab} (corresponding to $\bar{a} = 0$), and there is also mixing between the

GL(3,$\mathbb{R}$) constraints and the Momentum constraint appearing in (39). This second system will have a straightforward reduction to canonical vierbein GR, however the specialization to spatial coordinate frames is nontrivial.

The Hamiltonian H_{GR} is constructed from (35) and the operators that act on phase space as $d_\perp$ and $\mathscr{L}$, respectively,

$$\mathscr{H}[\underline{\ }] := -\underline{\mathbf{R}} + \mathbf{E}(k^a{}_b k^b{}_a - k^2), \tag{41a}$$

$$\mathscr{H}_a[\underline{\ }] := \mathbf{E}\nabla_b[2\gamma^{bc}\pi_{ac} - \delta^b_a\pi] = -2\mathbf{E}\nabla_b[k^b{}_a - \delta^b_a k]. \tag{41b}$$

From these we compute

$$\{\gamma^{ab}, \mathscr{H}_c[f^c]\} = \mathscr{L}_{\underline{f}}[\boldsymbol{\gamma}]^{ab}, \quad \{\underline{\pi}_{ab}, \mathscr{H}_c[f^c]\} = E\mathscr{L}_{\underline{f}}[\pi]_{ab}, \quad \{\underline{E}^i{}_a, \mathscr{H}_c[f^c]\} = 0, \tag{42a}$$

$$\{\mathbf{p}^a{}_i, \mathscr{H}_c[f^c]\} = -E^b{}_i(f^a\mathscr{H}_b - \tfrac{1}{2}\delta^a_b f^c\mathscr{H}_c) + E^b{}_i\mathscr{L}_{\underline{f}}[2\boldsymbol{\gamma}^{ac}\pi_{bc} - \tfrac{1}{2}\delta^a_b\boldsymbol{\pi}] \approx E^b{}_i\mathscr{L}_{\underline{f}}[\mathbf{p}^a{}_j E^j{}_b], \tag{42b}$$

and (defining $\nabla^2 := \gamma^{ab}\nabla_a\nabla_b$)

$$\{\gamma^{ab}, \mathscr{H}[f]\} = -\mathbf{f}(2k^{ab} - \gamma^{ab}k), \quad \{\underline{E}^i{}_a, \mathscr{H}[f]\} = 0, \tag{43a}$$

$$\{\underline{\pi}_{ab}, \mathscr{H}[f]\} = -E(\nabla_a\nabla_b[f] + \gamma_{ab}\nabla^w[f]) + f\underline{R}_{ab} - \underline{f}(2k_{ac}k^c{}_b + kk_{ab} - \gamma_{ab}k^2) + \underline{f}\gamma_{ab}(k^c{}_d k^d{}_c - k^2), \tag{43b}$$

$$\{\mathbf{p}^a{}_i, \mathscr{H}[f]\} = -2E^b{}_i\boldsymbol{\gamma}^{ac}\nabla_c\nabla_b[f] + fE^b{}_i\boldsymbol{\gamma}^{ac}(2R_{cb} - \tfrac{1}{2}\gamma_{cb}R) + \tfrac{1}{2}E^a{}_i\mathbf{f}(k^c{}_d k^d{}_c - k^2). \tag{43c}$$

It is straightforward to check from these that $\{k_{ab}, \mathscr{H}[N]\} \approx -NR_{ab} + 2Nk_{ac}k^c{}_b - Nkk_{ab} + \nabla_a\nabla_b[N]$, and we therefore reproduce the field equations (20b). We also find [using (43a)] that $\{\gamma_{ab}, \mathscr{H}[N]\} = 2Nk_{ab}$, reproducing the dynamical compatibility condition (19) or, equivalently, (26).

Using the above results we compute the constraint algebra

$$\{\mathscr{H}, \mathscr{T}^a{}_b[\omega^b{}_a]\} = 0, \quad \{\mathscr{H}_a, \mathscr{T}^b{}_c[\omega^c{}_b]\} = \Delta_{\tilde{\omega}}[\mathscr{H}]_a, \tag{44a}$$

$$\{\mathscr{H}[f], \mathscr{H}[g]\} = \int_\Sigma d^3x\,\gamma^{ab}(f\nabla_a[g] - g\nabla_a[f])\mathscr{H}_b, \tag{44b}$$

$$\{\mathscr{H}[f], \mathscr{H}_c[g^c]\} = \int_\Sigma d^3x\,\underline{f}\mathscr{L}_{\tilde{g}}[\mathscr{H}], \tag{44c}$$

$$\{\mathscr{H}_a[f^a], \mathscr{H}_a[g^a]\} = \int_\Sigma d^3x\,\underline{f}^a\mathscr{L}_{\tilde{g}}[\mathscr{H}]_a, \tag{44d}$$

which, in the limit of a spatial coordinate frame ($E^i{}_a = \delta^i_a$), agrees with the Dirac algebra.[5,6,7,21]

The Hamiltonian H'_{GR} is constructed from (35) and the operators that act on phase space as $d'_\perp$ and $\mathscr{L}'$, respectively,

$$\mathscr{H}'[\underline{\ }] = -\underline{\mathbf{R}} + \mathbf{E}(k'^a{}_b k'^b{}_a - k'^2) = \mathscr{H} + k'_{ab}\mathscr{T}^{ab} - \tfrac{1}{4}(\mathscr{T}^{(ab)}\mathscr{T}_{(ab)} - \tfrac{1}{2}\mathscr{T}^a{}_a\mathscr{T}^b{}_b), \tag{45a}$$

$$\mathscr{H}'_a[\underline{\ }] := \mathbf{E}\nabla_b[p^b{}_i E^i{}_a - \delta^b_a p] = -2\mathbf{E}\nabla_b[k'^b{}_a - \delta^b_a k'] + E\gamma_{ab}\nabla_c[\mathscr{T}]^{[bc]} = \mathscr{H}_a - E\nabla_b[\mathscr{T}]^b{}_a, \tag{45b}$$

where we have indicated the relationship between these primed constraints and the unprimed constraints in (41). Note that in (45b) there is a nontrivial mixing of the GL(3,$\mathbb{R}$) constraints and the momentum constraint as it appears in (39b). These act as

$$\{\boldsymbol{\gamma}^{ab}, \mathscr{H}'_c[f^c]\} = 0, \quad \{\underline{E}^i{}_a, \mathscr{H}'_c[f^c]\} = \Delta_{\tilde{\nabla}f}[\underline{E}^i{}_a], \tag{46a}$$

$$\{\underline{\pi}_{ab},\underline{\mathscr{H}}'_c[F^c]\}=E(V_{ab}[f]-\gamma_{ab}V^c{}_c[f])\approx\underline{\mathscr{L}}'_{\vec{f}}[\underline{\pi}]_{ab}, \tag{46b}$$

$$\{\mathbf{p}^a{}_i,\underline{\mathscr{H}}'_c[f^c]\}=\Delta_{\tilde{\nabla}f}[\mathbf{p}^a{}_i]+E^b{}_i\mathbf{U}^a{}_b[\vec{f}]-\tfrac{1}{2}E^a{}_i\mathbf{U}^c{}_c[\vec{f}]\approx\underline{\mathscr{L}}'_{\vec{f}}[\mathbf{p}^a{}_i], \tag{46c}$$

where $(H^a{}_b:=p^a{}_iE^i{}_b-\delta^a_bp)$

$$V_{ab}[f]:=\tfrac{1}{2}\nabla^c[f_aH_{cb}+f_cH_{ab}-f_aH_{bc}], \tag{47a}$$

$$\mathbf{U}^a{}_b[\vec{f}]:=\nabla_b[f]^c\mathbf{H}^a{}_c+\gamma_{bd}\nabla_c[f^c\mathbf{H}^{(ad)}]-\gamma_{bd}\nabla_c[f^a\mathbf{H}^{(cd)}]+\nabla_c[f_b\mathbf{H}^{[ca]}], \tag{47b}$$

and

$$\{\boldsymbol{\gamma}^{ab},\underline{\mathscr{H}}[f]\}=0,\quad\{\underline{E}^i{}_a,\underline{\mathscr{H}}'[f]\}=f\Delta_{\tilde{k}'}[\underline{E}^i{}_a], \tag{48a}$$

$$\{\underline{\pi}_{ab},\underline{\mathscr{H}}'[f]\}=-E(\nabla_a\nabla_b[f]+\gamma_{ab}\nabla^2[f])+f\underline{R}_{ab}+\underline{f}\,\gamma_{ab}(k'^c{}_dk'^d{}_c-k'^2), \tag{48b}$$

$$\{\mathbf{p}^a{}_i,\underline{\mathscr{H}}'[f]\}=-2E^b{}_i\boldsymbol{\gamma}^{ac}\nabla_c\nabla_b[f]+fE^b{}_i\boldsymbol{\gamma}^{ac}(2R_{cb}-\tfrac{1}{2}\gamma_{cb}R)+\mathbf{f}(k'^a{}_b-\delta^a_bk')p^b{}_i+\tfrac{1}{2}\mathbf{f}E^a{}_i(k'^c{}_dk'^d{}_c-k'^2). \tag{48c}$$

Here we find that (46a) reproduces (26) and that (20b) is satisfied when written in terms of the operator $d'_\perp$.

Using these results we compute the primed constraint algebra:

$$\{\underline{\mathscr{H}}',\underline{\mathscr{T}}^a{}_b[\omega^b{}_a]\}=0,\quad\{\underline{\mathscr{H}}'_a,\underline{\mathscr{T}}^b{}_c[\omega^c{}_b]\}=\Delta_{\tilde{\omega}}[\underline{\mathscr{H}}']_a, \tag{49a}$$

$$[\underline{\mathscr{H}}'[f],\underline{\mathscr{H}}'[g]\}=\int_\Sigma d^3x(f\nabla_a[g]-g\nabla_a[f])(\gamma^{ab}\underline{\mathscr{H}}'_b-E\nabla_b[\underline{\mathscr{T}}]^{[ab]}), \tag{49b}$$

$$\{\underline{\mathscr{H}}'[f],\underline{\mathscr{H}}'_c[g^c]\}=\int_\Sigma d^3x(\underline{f}\underline{\mathscr{L}}_{\vec{g}}[\underline{\mathscr{H}}']-fg^a\Delta_{\tilde{k}'}[\underline{\mathscr{H}}']_a+2g^a\nabla_c[fk'_{ab}]\underline{\mathscr{T}}^{[bc]}), \tag{49c}$$

$$\{\underline{\mathscr{H}}'_a[f^a],\underline{\mathscr{H}}'_a[g^a]\}=-\int_\Sigma d^3xf^ag^bR^c{}_{dab}\underline{\mathscr{T}}^d{}_c, \tag{49d}$$

which, when specialized to an orthonormal spatial frame ($\gamma_{ab}=\delta_{ab}$), is equivalent to that in Refs. 8 and 9.

These strong results agree with Ref. 4 up to the overall sign which is due to the fact that we are considering a right (Poisson) action in this work, whereas in Ref. 4 we derived the left action.[7] [Due to an unfortunate typesetting error, the left hand sides of equations (28b-e) of Ref. 4 should read as $[\Delta_{ax},\Delta_y]$, $[\delta_{\mathfrak{n}x},\Delta_y]$, $[\delta_{\mathfrak{n}x},\Delta_{ay}]$, and $[\delta_{ax},\Delta_y]$, respectively. Also note the rather obvious typo in the last line of (10), which should read as $\tilde{\nabla}_b[k]^a_c-g^{ad}\tilde{\nabla}_d[k]_{bc}$.] Either of the algebras (44) or (49) is a local representation of the Lie algebra $L\underline{\mathrm{Diff}}_0\mathbf{M}$ (the connected component since the exponential map is not onto) of the spacetime diffeomorphism group $\underline{\mathrm{Diff}}\mathbf{M}$. Note that although they were derived from a $(3+1)$ action, that algebras are valid in any dimension.

The Hamiltonian H_{GR} generates the weak evolution equations,

$$\partial_t[\boldsymbol{\gamma}^{ab}]=\{\boldsymbol{\gamma}^{ab},H_{\mathrm{GR}}\}\approx\Delta_{\tilde{N}}[\boldsymbol{\gamma}]^{ab}+\mathscr{L}_{\vec{N}}[\boldsymbol{\gamma}]^{ab}-N\Delta_{\tilde{k}}[\boldsymbol{\gamma}]^{ab}, \tag{50a}$$

$$\partial_t[\underline{E}^i{}_a]=\{\underline{E}^i{}_a,H_{\mathrm{GR}}\}\approx\Delta_{\tilde{N}}[\underline{E}^i{}_a], \tag{50b}$$

$$\partial_t[\underline{\pi}_{ab}]=\{\underline{\pi}^{ab},H_{\mathrm{GR}}\}\approx\Delta_{\tilde{N}}[\underline{\pi}]_{ab}+E\mathscr{L}_{\vec{N}}[\pi]_{ab}-E(\boldsymbol{\nabla}_a\boldsymbol{\nabla}_b[N]+\gamma_{ab}\nabla^2[N])+N\underline{R}_{ab}-NE(2k_{ac}k^c_b+kk_{ab})+NE\,\gamma_{ab}k^c_dk^d_c, \tag{50c}$$

$$\partial_t[\mathbf{p}^a{}_i]=\{\mathbf{p}^a{}_i,H_{\mathrm{GR}}\}\approx\Delta_{\tilde{N}}[\mathbf{p}^a{}_i]-2E^b{}_i\mathscr{L}_{\vec{N}}[\mathbf{k}^a_b]-2E^b{}_i\boldsymbol{\gamma}^{ac}\boldsymbol{\nabla}_b\boldsymbol{\nabla}_c[N]+2NE^b{}_i\boldsymbol{\gamma}^{ac}R_{cb}. \tag{50d}$$

Using compatibility (50a) may be written as

$$\Delta_{\tilde{\omega}}[\boldsymbol{\gamma}]^{ab}, \quad \text{where} \quad \omega^a{}_b = N^a{}_b - \boldsymbol{\nabla}_b[N]^a - Nk^a_b, \tag{51}$$

which is consistent with the easily verified result (noting that $\mathscr{T}_{(ab)} \approx 0$ is equivalent to $k_{ab} \approx k'_{ab}$) that the weak evolution equations for the primed system with Hamiltonian written as $H'_{\mathrm{GR}} = \int_\Sigma d^3x (N' \mathscr{H}' + N'^a \mathscr{H}'_a + N'^a{}_b \mathscr{T}'^b{}_a)$ is equivalent to inserting

$$N = N', \quad N^a = N'^a, \quad N^a{}_b = N'^a{}_b + \boldsymbol{\nabla}_b[N]^a + Nk^a_b, \tag{52}$$

in (50). In either of these formulations, the system is parameterized by the 30 fields (31), with 13 primary, first-class constraints ($\mathscr{T}^a_b$ with either $\mathscr{H}$ and $\mathscr{H}_a$, or $\mathscr{H}'$ and $\mathscr{H}'_a$), and the 13 associated Lagrange multipliers (N, N^a and $N^a{}_b$), and we find two configuration space degrees of freedom per spacetime point as expected.[18]

Beginning from the weak evolution equations (50), choosing $N^a{}_b \approx 0$ results in a system in which the spatial vielbeins do not evolve, and solving the GL(3,$\mathbb{R}$) constraints by choosing $p^a{}_i \approx E^b{}_i (2\gamma^{ac}\pi_{bc} - \frac{1}{2}\delta^a_b \pi)$ we can consider the evolution of the $\{\boldsymbol{\gamma}^{ab}, \pi_{ab}\}$ sector exclusively. Instead choosing $N'^a{}_b = 0$, we find a system in which $\boldsymbol{\gamma}^{ab}$ does not evolve and one may play the opposite game, solving the GL(3,$\mathbb{R}$) constraints by $\pi_{ab} \approx 1/2 \gamma_{ac}(p^c{}_i E^i{}_b + p\,\delta^c_b)$ and consider the evolution of $\{\mathbf{p}^a{}_i, E^i{}_a\}$. Note that from (52) one need only require that

$$N^{(ab)} = \gamma^{c(a} \boldsymbol{\nabla}_c [N]^{b)} - N\left(\pi^{ab} - \tfrac{1}{4}\gamma^{ab}\pi\right) = \gamma^{c(a} \boldsymbol{\nabla}_c [N]^{b)} + Nk^{ab}, \tag{53}$$

in order to find a system in which the metric degrees of freedom do not evolve, allowing other gauge choices in this case.

One may treat any choice of initial data in either of these two ways. By transforming the initial data by $\Delta_{\tilde{\omega}}$ defined so that $\Delta_{\tilde{\omega}}[E^i{}_a] = \delta^i_a$ (which is uniquely determined from the inverse as $[\tilde{\omega}]^a{}_i = E^a{}_i$), we transform the initial data to a physically equivalent set for which the frame is holonomic. If instead we make a choice of $\tilde{\omega}$ such that $\Delta_{\tilde{\omega}}[\gamma]^{ab} = \delta^{ab}$ ($\tilde{\omega}$ is defined only up to arbitrary spatial rotations which are generated by $\mathscr{T}^{[ab]}$) we may diagonalize the spatial metric.

IV. CONCLUSIONS

We have shown that by treating the components of the metric and the frame on an equal footing we not only have a consistent variational principle, but the generalized setting encompasses both the coordinate and orthonormal frame approaches via a choice of gauge. From the Einstein-Hilbert action for general relativity we have derived a Hamiltonian in the (extended) standard form $H_{\mathrm{GR}} = \int_\Sigma dx (N \mathscr{H} + N_a \mathscr{H}_a + N^a{}_b \mathscr{T}^b{}_a)$, where $\mathscr{H}$ and $\mathscr{H}_a$ are the generators (of the connected component) of spacetime diffeomorphisms, and $\mathscr{T}^a_b$ are the generators of $\mathfrak{gl}(3,\mathbb{R})$ which generate infinitesimal changes of frame on the spatial hypersurface Σ. The resulting constraint algebra was computed and found to be in agreement with the algebra derived by purely geometric means in Ref. 4. This generalized setting allows a straightforward treatment of fermions in a canonical setting (work in progress), where the primed constraints and algebra naturally appear.

ACKNOWLEDGMENTS

The author acknowledges support from the Natural Sciences and Engineering Research Council of Canada in the form of a postdoctoral fellowship.

[1] P. Peldán, ''Actions for gravity, with generalizations: A review,'' Class. Quantum Grav. **11**, 1087–1132 (1994).
[2] F. W. Hehl, J. Dermott McCrea, E. W. Mielke, and Y. Ne'eman, ''Metric affine gauge theory of gravity: Field equations, Noether identities, world spinors, and breaking of dilation invariance,'' Phys. Rep. **258**, 1–171 (1995).
[3] T. Regge and C. Teitelboim, ''Role of surface integrals in the Hamiltonian formulation of general relativity,'' Ann. Phys. (N.Y.) **88**, 286–318 (1974).
[4] M. A. Clayton, ''The tetrad frame constraint algebra,'' Class. Quantum Grav. **14**, 1851–1864 (1997).
[5] C. Teitelboim, ''How commutators of constraints reflect the spacetime structure,'' Ann. Phys. (N.Y.) **79**, 542–557 (1973).
[6] C. Teitelboim, ''The Hamiltonian structure of spacetime,'' in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 1.

[7] S. A. Hojman, K. V. Kuchař and C. Teitelboim, ''Geometrodynamics regained,'' Ann. Phys. (N.Y.) **96**, 88–135 (1976).
[8] M. Henneaux, ''Poisson brackets of the constraints in the Hamiltonian formulation of tetrad gravity,'' Phys. Rev. D **27**, 986–989 (1983).
[9] J. M. Charap, M. Henneaux, and J. E. Nelson, ''Explicit form of the constraint algebra in tetrad gravity,'' Class. Quantum Grav. **5**, 1405–1414 (1988).
[10] R. Floreanini and R. Percacci, ''Canonical algebra of GL(4)-invariant gravity,'' Class. Quantum Grav. **7**, 975–984 (1990).
[11] C. J. Isham and K. V. Kuchař, ''Representations of spacetime diffeomorphisms. I. Canonical parameterized field theories,'' Ann. Phys. (N.Y.) **164**, 288–315 (1985).
[12] C. J. Isham and K. V. Kuchař, ''Representations of spacetime diffeomorphisms. II. Canonical geometrodynamics,'' Ann. Phys. (N.Y.) **164**, 316–333 (1985).
[13] B. S. DeWitt, ''Quantum field theory in curved spacetime,'' Phys. Rep. **19**, 295–357 (1975).
[14] K. V. Kuchař, ''The problem of time in canonical quantization of relativistic systems,'' in *Conceptual Problems of Quantum Gravity*, edited by A. Ashtekar and J. Stachel (Birkhäuser, Boston, 1988), Vol. 2.
[15] R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago, 1984).
[16] Jerrold E. Marsden and Tudor S. Ratiu, *Introduction to Mechanics and Symmetry*, Vol. 17 of Texts in Applied Mathematics (Springer-Verlag, New York, 1994).
[17] M. Nakahara, *Geometry, Topology and Physics* (Hilger, New York, 1990).
[18] J. Isenberg and J. Nester, ''Canonical gravity,'' in Ref. 6.
[19] K. Kuchař, ''Geometry of hyperspace I,'' J. Math. Phys. **17**, 777–791 (1976).
[20] R. Arnowitt, S. Deser, and C. W. Misner, ''Dynamical structure and definition of energy in general relativity,'' Phys. Rev. **116**, 1322–1330 (1959).
[21] A. E. Fischer and J. E. Marsden, ''The initial value problem and the dynamical formulation of general relativity,'' in *General Relativity: An Einstein Centenary Survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, New York, 1979).

Geodesics in Lewis space–time

L. Herrera[a)]
Area de Física Teórica, Grupo de Relatividad, Universidad de Salamanca, Salamanca, Spain

N. O. Santos
Université Paris VI, CNRS/URA 769, Cosmologie et Gravitation Relativistes, Tour 22-12, 4ème étage, Boîte 142, 4 place Jussieu, 75005 Paris, France and Observatório Nacional CNPq, Departamento de Astrofísica, rua General José Cristino 77, 20921 Rio de Janeiro, Brazil

(Received 2 October 1997; accepted for publication 8 April 1998)

The geodesic equations are integrated for the Lewis metric and the effects of the different parameters appearing in the Weyl class on the motion of test particles are brought out. The appearance of a force parallel to the axial axis is without Newtonian analog and deserves particular attention. © *1998 American Institute of Physics.* [S0022-2488(98)02107-0]

I. INTRODUCTION

The Lewis metric[1] describes the vacuum space–time outside an axially symmetric unbounded, along the axial axis, source, endowed with angular momentum. Usually this metric is presented with four parameters[2] which may be real (Weyl class) or complex (Lewis class). In recent papers,[3–5] the physical and geometrical meanings of these parameters have been discussed for both classes. Thus for the Weyl class, it appears that one of the parameters is proportional to the energy per unit length (at least in the Newtonian limit), a second parameter is the arbitrary constant potential which is always present in the Newtonian solution, and the remaining two parameters are responsible for the nonstaticity of the metric, although they affect staticity in different ways.

It is the purpose of this work to delve more deeply into the physical meaning of the aforesaid parameters. Accordingly, we shall study the motion of test particles in the Lewis space–time.

The possible sources covering the axis of symmetry and matched to the Lewis metrics have been considered by various authors (see Refs. 3, 4), starting with the pioneering work of van Stockum.[6] Recently, the geodesic motion of test particles in van Stockum's source, representing a rigidly rotating dust cylinder with finite thickness, has been studied.[7] Remarkable results emerged from this work and in particular it was found that confinement occurs quite generally in the radial motion of the particles. It was pointed out that this confinement might provide a source for the extragalactic jet formation in active galaxies.[8] Hence it appears auspicious to study the geodesic motion in the exterior Lewis space–time.

As will be seen below, besides the usual dragging effect, to be expected whenever a rotating source is involved, another *mixed*, effect appears, which we call *topological frame dragging*, related simultaneously to the rotation of the source and to the topological defect associated with some metric parameters.

In the next section we shall present the Lewis metric; the geodesic equations are given in Sec. III; and the specific case of circular geodesics is considered in Sec. IV. The two next sections are devoted to the Weyl class. Two specific examples for circular geodesics are studied in Sec. V, exhibiting the difference in the role played by the two parameters responsible for the nonstaticity. Radial geodesics are considered for several different cases in Sec. VI. Furthermore, in this section, the geodesics along the axial axis show the appearance of a force parallel to the axis, without Newtonian analog. This force tends to damp the motion along the z axis, when the particle

[a)]On leave from Departamento de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela and Centro de Astrofísica Teórica, Merida, Venezuela.

 © 1998 American Institute of Physics

approaches this axis, whereas this tendency reverses when the particle moves away from it. Finally, conclusions are presented in the last section.

II. LEWIS SPACE–TIME

The general line element for a cylindrically symmetric stationary space–time is given by

$$ds^2=-f\,dt^2+2k\,dt\,d\phi+e^{\mu}(dr^2+dz^2)+l\,d\phi^2, \tag{1}$$

where f, k, μ, and l are functions only of r, and the ranges of the coordinates t, z, and ϕ are

$$-\infty<t<\infty, \quad -\infty<z<\infty, \quad 0\leqslant\phi\leqslant 2\pi, \tag{2}$$

the hypersurfaces $\phi=0$ and $\phi=2\pi$ being identified. The coordinates are numbered

$$x^0=t, \quad x^1=r, \quad x^2=z, \quad x^3=\phi. \tag{3}$$

Einstein's field equations for vacuum are

$$R_{\mu\nu}=0, \tag{4}$$

and with metric (1) the nonidentically null components become in van Stockum's[6] notation,

$$2e^{\mu}DR_0^0=\left(\frac{lf'+kk'}{D}\right)', \tag{5}$$

$$2e^{\mu}DR_0^3=\left(\frac{fk'-kf'}{D}\right)', \tag{6}$$

$$2e^{\mu}DR_3^0=\left(\frac{kl'-lk'}{D}\right)', \tag{7}$$

$$2e^{\mu}DR_3^3=\left(\frac{fl'+kk'}{D}\right)', \tag{8}$$

$$2R_1^1=-\mu''+\mu'\,\frac{D'}{D}-2\,\frac{D''}{D}+\frac{k'^2+f'l'}{D^2}, \tag{9}$$

$$2R_{22}=-\mu''-\mu'\,\frac{D'}{D}, \tag{10}$$

where the primes stand for differentiation with respect to r and

$$D^2=k^2+fl. \tag{11}$$

Equations (5)–(8) are not all independent; any one of them can be expressed in terms of the remaining three. The general solution of (4) for (1) is the stationary Lewis metric[1] given in the notation of[2]

$$f=ar^{-n+1}-\frac{c^2}{n^2a}\,r^{n+1}, \tag{12}$$

$$k=-Af, \tag{13}$$

$$e^{\mu}=r^{(n^2-1)/2}, \tag{14}$$

$$l=\frac{r^2}{f}-A^2f, \tag{15}$$

with

$$A=\frac{cr^{n+1}}{naf}+b. \tag{16}$$

The constants n, a, b, and c can be either real or complex, and the corresponding solutions belong to the Weyl or Lewis classes. For the Lewis class these constants are given by

$$n=im, \tag{17}$$

$$a=\tfrac{1}{2}(a_1^2-b_1^2)+ia_1b_1, \tag{18}$$

$$b=\frac{a_1a_2+b_1b_2}{a_1^2+b_1^2}+\frac{i}{a_1^2+b_1^2}, \tag{19}$$

$$c=\tfrac{1}{2}m(a_1^2+b_1^2), \tag{20}$$

where m, a_1, b_1, a_2, and b_2 are real constants and satisfy

$$a_1b_2-a_2b_1=1. \tag{21}$$

Equations (17)–(21) reveal that if the values of the parameters n and a, or n and b are known, we can obtain the parameter c. However, knowing n and c, we cannot obtain a and b. The metric coefficients (12)–(15) with (17)–(20) become[1]

$$f=(a^2-b_1^2)r\ \cos(m\ \ln r)+2a_1b_1r\ \sin(m\ \ln r), \tag{22}$$

$$k=-(a_1a_2-b_1b_2)r\ \cos(m\ \ln r)-(a_1b_2+a_2b_1)r\ \sin(m\ \ln r), \tag{23}$$

$$e^{\mu}=r^{-(m^2+1)/2}, \tag{24}$$

$$l=-(a_2^2-b_2^2)r\ \cos(m\ \ln r)-2a_2b_2r\ \sin(m\ \ln r). \tag{25}$$

III. GEODESICS

The equations governing geodesics can be derived from the Lagrangian

$$2\mathscr{L}=g_{\mu\nu}\frac{dx^{\mu}}{d\lambda}\frac{dx^{\nu}}{d\lambda}, \tag{26}$$

where λ is an affine parameter along the geodesics. For timelike geodesics λ is the proper time. From the extremal problem emerges the Euler–Lagrange equations

$$\frac{d}{d\lambda}\left(\frac{\partial\mathscr{L}}{\partial\dot{x}^{\alpha}}\right)-\frac{\partial\mathscr{L}}{\partial x^{\alpha}}=0, \tag{27}$$

and from them follow the geodesics given by

$$\ddot{x}^{\alpha}+\Gamma^{\alpha}_{\beta\gamma}\dot{x}^{\beta}\dot{x}^{\gamma}=0, \tag{28}$$

where the overdot stands for differentiation with respect to λ. For space–time (1) the Lagrangian (26) is

$$2\mathscr{L}=-f\dot{t}^2+2k\dot{t}\dot{\phi}+e^{\mu}(\dot{r}^2+\dot{z}^2)+l\dot{\phi}^2, \tag{29}$$

and the geodesic equations (28) are

$$D\ddot{t}+\frac{lf'+kk'}{D}\,\dot{t}\dot{r}+\frac{kl'-lk'}{D}\,\dot{r}\dot{\phi}=0, \tag{30}$$

$$2\ddot{r}+e^{-\mu}(f'\dot{t}^2-2k'\dot{t}\dot{\phi}-l'\dot{\phi}^2)+\mu'(\dot{r}^2-\dot{z}^2)=0, \tag{31}$$

$$\ddot{z}+\mu'\dot{r}\dot{z}=0, \tag{32}$$

$$D\ddot{\phi}+\frac{fk'-kf'}{D}\,\dot{f}\dot{r}+\frac{fl'+kk'}{D}\,\dot{r}\dot{\phi}=0. \tag{33}$$

The corresponding canonical momenta to (29) are

$$p_t=-\frac{\partial\mathscr{L}}{\partial\dot{t}}=f\dot{t}-k\dot{\phi}, \tag{34}$$

$$p_r=\frac{\partial\mathscr{L}}{\partial\dot{r}}=e^{\mu}\dot{r}, \tag{35}$$

$$p_z=\frac{\partial\mathscr{L}}{\partial\dot{z}}=e^{\mu}\dot{z}, \tag{36}$$

$$p_\phi=\frac{\partial\mathscr{L}}{\partial\dot{\phi}}=k\dot{t}+l\dot{\phi}. \tag{37}$$

The integrals of motion follow from the Euler–Lagrange equations (27)

$$\frac{dp_t}{d\lambda}=\frac{\partial\mathscr{L}}{\partial t}=0, \tag{38}$$

$$\frac{dp_z}{d\lambda}=\frac{\partial\mathscr{L}}{\partial z}=0, \tag{39}$$

$$\frac{dp_\phi}{d\lambda}=\frac{\partial\mathscr{L}}{\partial\phi}=0, \tag{40}$$

and the conserved quantities are

$$p_t=E,\quad p_z=P,\quad p_\phi=L, \tag{41}$$

where the constants E, P, and L represent, respectively, the total energy of the test particle, its momentum along the z axis, and its angular momentum about the z axis. From (34), (36), (37), (41) we can have

$$\dot{t}=\frac{1}{D^2}\,(Lk+El), \tag{42}$$

$$\dot{z}=Pe^{-\mu}, \tag{43}$$

$$\dot{\phi}=\frac{1}{D^2}\,(Lf-Ek), \tag{44}$$

and instead of integrating (31) we can use the line element (1) to obtain

$$-\epsilon=-f\dot{t}^2+2k\dot{t}\dot{\phi}+e^{\mu}(\dot{r}^2+\dot{z}^2)+l\dot{\phi}^2, \tag{45}$$

where $\epsilon=0$, 1 or -1 if the geodesics are, respectively, null, timelike, or spacelike.

IV. CIRCULAR GEODESICS

Now we restrict ourselves to the study of circular geodesics assuming

$$\dot{r}=\dot{z}=0, \quad \ddot{t}=0, \quad \ddot{\phi}=0, \tag{46}$$

then (31) becomes

$$f'\dot{t}^2-2k'\dot{t}\dot{\phi}-l'\dot{\phi}^2=0. \tag{47}$$

The angular velocity of the test particle is given by

$$\omega=\frac{\dot{\phi}}{\dot{t}}, \tag{48}$$

which becomes, using (47),

$$\omega=\frac{1}{l'}\left[-k'\pm(k'^2+f'l')^{1/2}\right]. \tag{49}$$

For a stationary space–time the normal velocity W^μ of the particle defined as the change in the displacement normal to $\tau^\mu=(1,0,0,0)$ relative to its displacement parallel to τ^μ, where τ^μ is a timelike Killing vector, is[9]

$$W^\mu=\left[(-g_{00})^{1/2}\left(dx^0+\frac{g_{0a}}{g_{00}}\,dx^a\right)\right]^{-1}V^\mu, \tag{50}$$

where

$$V^\mu=\left(-\frac{g_{0a}}{g_{00}}\,dx^a,dx^1,dx^2,dx^3\right), \tag{51}$$

and Latin indexes range from 1 to 3. Considering the metric (1) and the expression for ω, (48), then (50) becomes

$$W^\mu=\frac{(k\omega,0,0,f\omega)}{f^{1/2}(f-k\omega)}. \tag{52}$$

The modulus of W^μ, defined by $W^2=W^\mu W_\mu$, and using (52) is

$$W=\frac{D\omega}{f-k\omega}. \tag{53}$$

We can write (45) by considering (48)

$$\frac{\epsilon}{\dot{t}^2}=f-2k\omega-l\omega^2. \tag{54}$$

Substituting (53) into (54) we obtain

$$\frac{\epsilon}{\dot{t}^2}=\frac{D^2f}{(D+kW)^2}\,(1-W^2), \tag{55}$$

which shows that circular geodesics are timelike, null, or spacelike for, respectively, $W<1$, $W=1$, or $W>1$.

V. CIRCULAR GEODESICS IN THE WEYL CLASS

In this section we study circular geodesics in the Lewis space–time for the Weyl class, which means that the parameters n, a, b and c appearing in (12)–(15) are real.

The parameter n is associated with the Newtonian mass per unit length σ,

$$\sigma=\tfrac{1}{4}(1-n), \tag{56}$$

of a uniform line mass in the low density regime. The parameter a is connected to the constant arbitrary potential that exists in the corresponding Newtonian solution. The static and locally flat limit produces angular deficit δ, given by

$$\delta=2\pi\left(1-\frac{1}{a^{1/2}}\right). \tag{57}$$

The parameter b is associated, in the locally flat limit, with the angular momentum of a spinning string. Finally, the parameter c is related to the vorticity of the source when it is represented by a stationary completely anisotropic fluid. This parameter, together with the parameter b, is responsible for the nonstaticity of the metric. All these interpretations are restricted to the Weyl class only. For further details see Ref. 3.

We shall now discuss the expression for the angular velocity (49) and the tangential velocity (53). In order to exhibit more clearly the role played by different parameters, we shall consider two different cases, when $b=0$, $c\neq 0$ and $b\neq$, $c=0$.

A. Case $b=0$, $c\neq 0$

For this case substituting metric (12)–(15) into (49) we obtain

$$\omega=\frac{c}{n}\pm\omega_0=\frac{c}{1-4\sigma}\pm\omega_0, \tag{58}$$

where ω_0 is the angular velocity when the space–time is static given by Levi-Civita's metric,

$$\omega_0^2=\frac{1-n}{1+n}\,a^2r^{-2n}=\frac{2\sigma}{1-2\sigma}\,a^2r^{2(4\sigma-1)}, \tag{59}$$

and (56) has been used.

From (59) we can see that when the topological defect a decreases the total angle around the z axis, being $a>1$, it increases the angular velocity. This situation reverses for $a<1$.

The term $c/(1-4\sigma)$ represents an inertial frame dragging correction to the static case, similar to the one appearing in the case of the Kerr metric[10] or in the field of a massive charged magnetic dipole.[11] The presence of c in that term, becomes intelligible when we recall that this parameter measures the vorticity of the source when described by a rigidly rotating anisotropic cylinder.[3] It increases or diminishes the modulus of ω if the vorticity is in the same or opposite direction, respectively, of the rotation of the test particle. Furthermore, the dragging term is increased by the Newtonian mass per unit length σ.

Calculating the tangential velocity of the particle (53) with (12)–(15) and (56) we obtain

$$W=\left(\frac{c}{n}\pm\omega_0\right)\frac{r^n}{a}\left(1\pm\omega_0\,\frac{cr^{2n}}{na^2}\right)^{-1} \tag{60}$$

$$=\left[\frac{cr^{1-4\sigma}}{a(1-4\sigma)}\pm W_0\right]\left[1\pm W_0\,\frac{cr^{1-4\sigma}}{a(1-4\sigma)}\right]^{-1}, \tag{61}$$

where W_0 is the velocity in the static Levi-Civita space–time

$$W_0^2=\frac{1-n}{1+n}=\frac{2\sigma}{1-2\sigma}, \tag{62}$$

which can also be rewritten with (59)

$$W_0=\frac{r^n}{\alpha}\,\omega_0, \tag{63}$$

showing clearly the Newtonian limit $W_0=r\omega_0$. We see that the dragging term $(cr^{1-4\sigma})/[a(1-4\sigma)]$ for low energy densities and $a=1$ is proportional to the vorticity times the radius. However, it is worth noticing that this latter term decreases with a, and furthermore since $W_0<1$, W also decreases with a. Thus we have a dragging effect corrected by a topological defect. In order to understand why W decreases with a, unlike the angular velocity ω, which grows with a, it should be observed that the arc length Λ along a circular path in this case with $n=1$ becomes $d\Lambda=(r/a^{1/2})d\phi$. Therefore, for a given $d\phi$, $d\Lambda$ decreases with a, explaining thereby the decreasing of W. On the other hand, for a given Λ, $d\phi$ increases with a, which explains why the angular velocity also grows with a. This argument is not without its problems. Indeed, if $c=0$ the space–time reduces to the static Levi-Civita space–time, but the tangential velocity W_0 is unaffected by a contrary to the argument above. The same is true for the next case below, with respect to b.

We notice here too, in this case, that c plays no role as a topological defect, i.e., it does not appear in the metric term l, hence it is restricted to its frame dragging effect.

Considering (54) and (12)–(15) we obtain,

$$\frac{\epsilon}{\dot{t}^2}=\frac{r^{n+1}}{a}\left[a^2r^{-2n}-\left(\omega-\frac{c}{n}\right)^2\right], \tag{64}$$

and using (56), (58), (59) it becomes

$$\frac{\epsilon}{\dot{t}^2}=\frac{r^{n+1}}{a}\,(a^2r^{-2n}-\omega_0^2)=\frac{1-4\sigma}{1-2\sigma}\,ar^{4\sigma}. \tag{65}$$

We see from (65) that for circular geodesics the conditions for timelike, null and spacelike orbits, correspondingly $0\leqslant\sigma<1/4$, $\sigma=1/4$ and $1/4<\sigma\leqslant 1/2$ are unaffected by the vorticity of the source c.

B. Case $b\neq 0$, $c=0$

For this case substituting the metric (12)–(15) into (49) we obtain

$$\omega=\frac{\omega_0(b\omega_0\pm 1)}{1-b^2\omega_0^2}. \tag{66}$$

Now we have two different kinds of terms appearing in (66). On one hand, we have the angular velocity corresponding to the Levi-Civita space–time corrected by the factor $1-b^2\omega_0^2$, and on the other, we have the topological frame dragging term associated with b. We call it topological frame dragging since b also produces a topological defect and to distinguish from the frame dragging engendered by c that does not produce a topological defect in the previous case. The increasing of ω with b may be interpreted as follows. In the $n=1$ case the line element (1) takes the form

$$ds^2=-dt^2-2ba^{1/2}\,dt\,d\phi+dr^2+dz^2+\left(\frac{r^2}{a}-b^2a\right)d\phi^2. \tag{67}$$

Therefore, from (67), the arc length Λ along the circular path is given by

$$d\Lambda=\left(\frac{r^2}{a}-b^2a\right)^{1/2}d\phi. \tag{68}$$

Thus, for a given $d\Lambda$, larger b's lead to larger $d\phi$'s and thereby to larger angular velocities.

Considering (54) and (12)–(15) we obtain

$$\frac{\epsilon}{\dot{t}^2}=ar^{-n+1}(1+b\omega)^2-\frac{r^2}{f}\,\omega^2, \tag{69}$$

and using (56), (59), (66) it becomes

$$\frac{\epsilon}{\dot{t}^2}=\frac{1}{(1\mp b\omega_0)^2}\,\frac{r^{n+1}}{a}\,(a^2r^{-2n}-\omega_0^2)=\frac{1}{(1\mp b\omega_0)^2}\,\frac{1-4\sigma}{1-2\sigma}\,ar^{4\sigma}. \tag{70}$$

We see from (70) that for circular geodesics the condition for timelike, null, and spacelike orbits are the same as in cases $b=0$ and $c\neq 0$.

Calculating the tangential velocity of the particle (53) with (12)–(15) we obtain

$$W=\frac{r^n}{a}\,\omega_0=W_0. \tag{71}$$

From (71) we have that the particle velocity is not affected by b, in fact, its velocity is the same as the one in the static Levi-Civita space–time. We emphasize that b does not produce frame dragging upon W but topological frame dragging upon ω. While, in the previous case, we saw that c produces frame dragging upon W without producing topological defect.

In the general case, $b\neq 0$ and $c\neq 0$, both parameters contribute in a cumbersome way to frame dragging and topological effects.

VI. GEODESICS IN THE WEYL CLASS

Substituting (42)–(44) into (45) we have an expression for the radial speed $\dot{r}^2$ of the test particle,

$$\dot{r}^2=e^{-\mu}\left(E^2\,\frac{l}{D^2}+2EL\,\frac{k}{D^2}-\epsilon-L^2\,\frac{f}{D^2}-P^2e^{-\mu}\right). \tag{72}$$

We restrict the study of geodesics to $0<n<1$, which is the condition for circular timelike geodesics as expected in the Newtonian analog.[3]

From (14) and (43) we obtain

$$\dot{z}=Pr^{(1-n^2)/2}, \tag{73}$$

which means that if $P\neq 0$, particles increase their speed along z when distancing radially from the axis, while diminishing their axial speed when moving radially towards the axis. This result indicates that a force parallel to the z axis appears. In the flat case $n=1$ such an effect vanishes, bringing out its non-Newtonian nature.

From (14) and (32) it follows that

$$\ddot{z}=\frac{1-n^2}{2}\,\frac{\dot{r}\dot{z}}{r}, \tag{74}$$

therefore, this force tends to damp any motion along the z axis whenever the particle approaches that axis, and reverses this tendency in the opposite case. A similar result is obtained for the van Stockum space–time,[7] although in that case the effect is related to the vorticity of the source, whereas in our case, neither b nor c appear involved in (73). It is also worth noticing that non-Newtonian forces, parallel to the z axis, also appear in the field of axially symmetric rotating bodies.[12] However, as in the example of Ref. 7, the force parallel to z, in Ref. 12, unlike our case, is directly related to the spin of the source. In a stationary rotating universe a repulsive potential along the axis of rotation has also been observed.[13]

Now we study (72) for the two separate cases considered for circular geodesics and concentrate for null and timelike particles, i.e., $\epsilon=0$ and 1.

A. Case $b=0$, $c\neq 0$

For this case substituting (12)–(15) into (72) we obtain

$$\dot{r}^2 = r^{-(1-n)^2/2}[V_0 - V(r)], \tag{75}$$

where

$$V_0 = \frac{1}{a}\left(E - L\,\frac{c}{n}\right)^2 > 0, \tag{76}$$

$$V(r) = \epsilon r^{1-n} + P^2 r^{(1-n)(3+n)/2} + aL^2 r^{-2n} > 0. \tag{77}$$

The radial acceleration of the particle is from (75)

$$\ddot{r} = \tfrac{1}{4} r^{-1-(1-n)^2/2}[-(1-n)^2 V_0 - (1-n)(1+n)\epsilon r^{1-n} + (1+n)^2 aL^2 r^{-2n} - (1-n)(1+n)P^2 r^{(1-n)(3+n)/2}]. \tag{78}$$

From (77) we find

$$V'(r) = (1-n)\epsilon r^{-n} - 2naL^2 r^{-1-2n} + \tfrac{1}{2}(1-n)(3+n)P^2 r^{-1+(1-n)(3+n)/2}. \tag{79}$$

The equation $V'(r)=0$ has a solution, say $r=r_c$, which satisfies

$$2naL^2 r_c^{-1-2n} = (1-n)\epsilon r_c^{-n} + \tfrac{1}{2}(1-n)(3+n)P^2 r_c^{-1+(1-n)(3+n)/2}. \tag{80}$$

Since

$$V''(r) = \frac{1}{r}\left\{-n(1-n)\epsilon r^{-n} + 2n(2n+1)aL^2 r^{-1-2n} + \frac{1}{2}(1-n)(3+n)\left[\frac{1}{2}(1-n)(3+n) - 1\right]P^2 r^{-1+(1-n)(3+n)/2}\right\}, \tag{81}$$

and with (80) we have

$$V''(r_c) = \frac{1}{r_c}\{(1-n)(1+n)\epsilon r_c^{-n} + \tfrac{1}{2}(1-n)(3+n) \times [2n + \tfrac{1}{2}(1-n)(3+n)]P^2 r_c^{-1+(1-n)(3+n)/2}\}. \tag{82}$$

On the other hand, from (77), if $\epsilon\neq 0$ and/or $P\neq 0$, we have when $r\to 0$ or ∞, $V(r)\to\infty$. Furthermore, from (82) we have $V''(r_c)>0$, hence the equation $V'(r)=0$ has only one solution $r=r_c$ and is a minimum.

We see that in this case the parameter c only affects V_0, (76), by modifying the energy of the test particle, and leaving otherwise the geodesics indistinguishable from the static Levi-Civita space–time.

1. Case $\epsilon=0$, $P=0$

From (77) and (78) we have

$$V(r) = aL^2 r^{-2n}, \tag{83}$$

$$\ddot{r} = \tfrac{1}{4} r^{-1-(1-n)^2/2}[-(1-n)^2 V_0 + (1+n)^2 aL^2 r^{-2n}]. \tag{84}$$

A null particle with large r approaches z with decreasing negative acceleration, $\ddot{r}<0$, and increasing speed $\dot{r}$. Its speed attains a maximum at $\ddot{r}=0$ and then diminishes since the acceleration

becomes positive, $\ddot{r}>0$. The null particle arrives at a minimum distance $r=r_{\min}$ when $\dot{r}=0$ and $V_0=V(r)$. At $r=r_{\min}$, the null particle is reflected to infinity, $r\to\infty$, where $\dot{r}\to 0$.

2. Case $\epsilon\neq 0$, $P=0$

From (77) and (78) we have

$$V(r)=\epsilon r^{1-n}+aL^2r^{-2n}, \tag{85}$$

$$\ddot{r}=\tfrac{1}{4}r^{-1-(1-n)^2/2}[-(1-n)^2V_0-(1-n)(1+n)\epsilon r^{1-n}+(1+n)^2aL^2r^{-2n}]. \tag{86}$$

For this case $V_0=V(r)$ has two real roots, say $r_{\min}$ and $r_{\max}$. The timelike particle approaching z is reflected at $r=r_{\min}$ and moves outwards until it attains $\dot{r}=0$ at $r=r_{\max}$ repeating endlessly this trajectory. This kind of confinement of the test particle has also been found in the van Stockum space–time.[7] In this case if the motion is circular, $\dot{r}=0$, there are stable orbits, since $V(r)$ has a minimum.

3. Case $P\neq 0$

For this case, while the particles move along the z axis as described by (73), there is a confinement similar to the previous case in the $z=$constant plane. The effect of P, if $\epsilon\neq 0$, is to diminish the $r_{\min}$ and $r_{\max}$ attained by the particle. It is interesting to observe that even if $\epsilon=0$ the null particle has a similar trajectory in the $z=$constant plane which is not allowed if $P=0$, as described previously.

B. Case $b\neq 0$, $c=0$

For this case substituting (12)–(15) into (72) we obtain

$$\dot{r}^2=r^{-(1-n)^2/2}[V_0-V(r)], \tag{87}$$

where

$$V_0=\frac{1}{a}E^2>0, \tag{88}$$

$$V(r)=\epsilon r^{1-n}+P^2r^{(1-n)(3+n)/2}+a(L+bE)^2r^{-2n}>0. \tag{89}$$

We see that contrary to c, b affects $V(r)$ and not V_0, by modifying the angular momentum of the test particle. From (89) we see that even if $L=0$ there is an angular momentum term acting on the particle. This difference between the parameters b and c is expected from our analysis for the circular geodesics. For $L\neq 0$ and 0 and $\epsilon=0$, $P=0$; $\epsilon\neq 0$, $P=0$; and $P\neq 0$ the behavior of the geodesics is qualitatively similar to the case $b=0$, $c\neq 0$ with $L\neq 0$.

VII. CONCLUSIONS

As is apparent from the study of circular geodesics, two different types of dragging effects are present in Lewis space–time of the Weyl class. On one hand, there is a *typical* dragging effect, as shown by Eq. (58), related to the parameter c (which in turn is related to the vorticity of the source). On the other hand, there are corrections to the expression of the angular velocity of the test particle [Eq. (66)], as compared to the Levi-Civita case, which are related to the parameter b. Since this last parameter produces a topological defect, we call this effect topological frame dragging. Both effects are in principle observables through the precession of a gyroscope carried along a circular geodesic in Lewis space–time. From the analysis of the geodesic motion along the z axis we have discovered [Eq. (74)] the presence of a force, parallel to this axis and of non-Newtonian nature (it vanishes when $n=1$). Unlike other reported examples of similar forces,[7,12,13] in this case, the force is not related to the spin of the source. Finally, the effect of parameters b and c on the radial motion of a test particle is exhibited in different subcases. The confinement of the particle, appearing in some of these subcases, is also found in the van Stockum space–time[6] and deserves particular attention.

[1] T. Lewis, Proc. R. Soc. **136**, 176 (1932).
[2] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[3] M. F. A. da Silva, L. Herrera, F. M. Paiva, and N. O. Santos, Gen. Relativ. Gravit. **27**, 859 (1995).
[4] M. F. A. da Silva, L. Herrera, F. M. Paiva, and N. O. Santos, Class. Quantum Grav. **12**, 111 (1995).
[5] M. A. H. MacCallum and N. O. Santos, ''Stationary and static cylindrically symmetric Einstein spaces of the Lewis form,'' Class. Quantum Grav. (in press).
[6] W. J. van Stockum, Proc. R. Soc. Edinburgh **57**, 135 (1937).
[7] R. Opher, N. O. Santos, and A. Wang, J. Math. Phys. **37**, 1982 (1996).
[8] R. D. Blandford and M. J. Rees, Mon. Not. R. Astron. Soc. **169**, 395 (1974); Contemp. Phys. **16**, 1 (1975).
[9] J. L. Anderson, *Principles of Relativity* (Academic, New York, 1967), Sec. 10.6(a).
[10] W. Rindler and V. Perlick, Gen. Relativ. Gravit. **22**, 1067 (1990).
[11] W. B. Bonnor, Phys. Lett. A **158**, 23 (1991).
[12] W. B. Bonnor, J. Phys. A **10**, 1673 (1977).
[13] E. L. Schücking, *Atti del Convegno sulla Cosmologia* (G. Barbèra Editore, Florence, 1966).

A Poincaré–Birkhoff–Witt theorem for generalized Lie color algebras

César Bautista
Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Edif. 135, 14 sur y Av. San Claudio, Ciudad Universitaria, Puebla, Pue. C.P. 72570, México

(Received 13 August 1997; accepted for publication 18 February 1998)

A proof of the Poincaré–Birkhoff–Witt theorem is given for a class of generalized Lie algebras closely related to the Gurevich S-Lie algebras. As concrete examples, we construct the positive (negative) parts of the quantized universal enveloping algebras of type A_n and $M_{p,q,\epsilon}(n,\mathbb{K})$, which is a nonstandard quantum deformation of $GL(n)$. In particular, we get, for both algebras, a unified proof of the Poincaré–Birkhoff–Witt theorem, and we show that they are genuine universal enveloping algebras of certain generalized Lie algebras. © *1998 American Institute of Physics.* [S0022-2488(98)00406-X]

I. INTRODUCTION

Yamane[1] presented a proof of the Poincaré–Birkhoff–Witt (PBW) theorem for some class of quantum groups: Drinfeld–Jimbo quantum groups of type A_n. In his proof he did not use explicitly the Lie algebra theory concepts.

In this paper we show that Yamane used in an implicit manner some generalized Lie algebra. Such a generalized Lie algebra will be called *T-Lie algebra*.

The T-Lie algebras satisfy not only generalized antisymmetry and a Jacobi identity, but additional properties like multiplicativity (also generalized, in the same way as the Gurevich S-Lie algebras[2]). Such T-Lie algebras arise in a natural way embedded in the positive and negative parts of the Drinfeld–Jimbo quantum groups $U_q(sl_{n+1})$ of type A_n.

Our T-Lie algebras share some properties with the S-Lie algebras. But they are not equivalent, for example, T-Lie algebras satisfy a weaker multiplicativity condition. In particular, there are some T-Lie algebras that are not S-Lie algebras. However, classical Lie algebras,[3] Lie superalgebras,[4] and Scheunert generalized Lie algebras (Lie color algebras)[5] are all T-Lie algebras.

These T-Lie algebras are related to the problem of finding the appropriate definition of *a quantum Lie algebra*. There are already some generalized Lie algebras proposed to solve this problem: Majid braided Lie algebras,[6] Delius–Gould quantum Lie algebras,[7] new generalized Lie algebras of Gurevich–Rubstov,[8] generalized Lie algebras due to Lyubashenko–Sudbery,[9] among others. But the Delius–Gould definition as well as the Gurevich–Rubstov, depend on the associated universal enveloping algebra. This is not the case for the T-Lie algebras. Our axioms imply the properties of the universal enveloping algebra. In particular, we shall prove the PBW theorem.

The generalized Lie algebras axioms of Lyubashenko–Sudbery are not enough to obtain a PBW theorem (see Example IV.4). The main difference with the braided Lie algebras of Majid is that the symmetry of our T-Lie algebras is not a braid morphism. Only a part of such a symmetry is braided.

In particular, we get a T-Lie algebra $(sl^{\pm}_{n+1})_q$, which is a deformation of the Lie subalgebra of upper (lower) triangular matrices. Such a generalized Lie algebra meets almost all the requirements of a *quantum Lie algebra*, in the sense of Lyubashenko–Sudbery[9] [only it fails point seven; actually the universal enveloping algebra of $(sl^{\pm}_{n+1})_q$ has no Hopf algebra structure, although it seems possible to define a *braided* Hopf algebra on it, but, we do not attempt it in this paper]. Moreover, the universal enveloping of $(sl^{\pm}_{n+1})_q$ coincides $U^{\pm}_q(sl_{n+1})$, which is the positive part of the Drinfeld–Jimbo quantum group of type A_n; therefore the diagram in Fig. 1 commutes. This means that, relative to $U^{\pm}_q(sl_{n+1})$, the T-Lie algebra $(sl^{\pm}_{n+1})_q$ satisfies, in some sense, the quantum Lie algebra condition of Delius.[10]

 © 1998 American Institute of Physics

$$\begin{array}{ccc} sl_{n+1}^{\pm} & \longrightarrow & (sl_{n+1}^{\pm})_q \\ \downarrow & & \downarrow \\ U(sl_{n+1}^{\pm}) & \longrightarrow & U_q^{\pm}(sl_{n+1}) = U(sl_{n+1}^{\pm})_q \end{array}$$

FIG. 1. The classical $sl_{n+1}^{\pm}$ and the quantum $(sl_{n+1}^{\pm})_q$.

Some possible physical applications of the formalism of generalized Lie algebras are in the affine Toda theories,[11] quantum integrable systems,[11] and gauge theory.[9]

The paper is organized as follows. In Sec. II we shall define the T-Lie algebras. In Sec. III a list of classical and new Lie algebras is given. In Sec. IV we shall define the universal enveloping algebra of a T-Lie algebra, and we shall prove that expecting an analog at the PBW theorem for any such universal enveloping algebra constructed by means of commutators is too much, we have to restrict our generalized Lie algebras in an adequate way. However, in Sec. V we pursue the classical challenge: to prove the PBW theorem[3] by constructing a representation of the universal enveloping algebra on the symmetric algebra (with modifications inspired by Ref. 1). In Sec. VI the definition of a representation of T-Lie algebra is given. In Sec. VII we shall prove the PBW theorem for the universal enveloping of an adequate T-Lie algebra. Some remarks about braid morphisms are given in Sec. VIII. Section IX is devoted to explaining why we can apply the T-Lie algebra theory to a nonstandard quantum deformation algebra[12] of $GL(n)$. Similar explanations are given in Sec. X, but instead dealing with $U_q^{\pm}(sl_{n+1})$ the positive (negative) parts of the Drinfeld–Jimbo quantum groups of type A_n. In particular, in Sec. X we shall prove that $U_q^{\pm}(sl_{n+1})$ is a genuine universal enveloping algebra of a certain T-Lie algebra.

II. THE NOTION OF *T*-LIE ALGEBRA

Let k be a commutative unitary ring.

Definition II.1: A k-algebra A is strictly graded if there exist k submodules $(A_\eta)_{\eta\in\mathbb{N}}$ such that

$$A=\oplus_{\eta\in\mathbb{N}}A_\eta \quad and \quad A_{\eta_1}\cdot A_{\eta_2}\subseteq A_{\eta_1+\eta_2-1},$$

for all $\eta_1,\eta_2\in\mathbb{N}$. For $a\in A_\eta$, we shall put $\eta(a)=\eta$.

Remark II.1: For such graded algebras A we can induce a filtration of $A\otimes_k A$, given by

$$(A\otimes_k A)_\eta=\oplus_{\eta_1+\eta_2\leqslant\eta}A_{\eta_1}\otimes A_{\eta_2}.$$

Let L be a free k module with a given basis $\mathscr{B}$ totally ordered.

Definition II.2: Denote by L^n the k submodule of $L^{n\otimes}$ generated by

$$x_{i_1}\otimes\cdots\otimes x_{i_n},\quad x_{i_1}<\cdots<x_{i_n}\quad (x_{i_j}\in\mathscr{B}),$$

and by ${}^nL$ the k submodule generated by

$$x_{i_1}\otimes\cdots\otimes x_{i_n},\quad x_{i_1}>\cdots>x_{i_n}\quad (x_{i_j}\in\mathscr{B}).$$

Definition II.3: The module L with k morphisms,

$$S:L\otimes_k L\to L\otimes_k L\quad (presymmetry), \tag{1}$$

$$T:L\otimes_k L\to L\otimes_k L\quad (symmetry), \tag{2}$$

$$\langle,\rangle:L\otimes_k L\to L\otimes_k L\quad (pseudobracket), \tag{3}$$

$$[,]:L\otimes_k L\to L\quad (bracket), \tag{4}$$

$$\begin{array}{ccc} L_1 \otimes_k L_1 & \xrightarrow{\langle,\rangle_1} & L_1 \otimes_k L_1 \\ {\scriptstyle f\otimes f}\downarrow & & \downarrow{\scriptstyle f\otimes f} \\ L_2 \otimes_k L_2 & \xrightarrow{\langle,\rangle_2} & L_2 \otimes_k L_2 \end{array} \qquad \begin{array}{ccc} L_1 \otimes_k L_1 & \xrightarrow{S_1} & L_1 \otimes_k L_1 \\ {\scriptstyle f\otimes f}\downarrow & & \downarrow{\scriptstyle f\otimes f} \\ L_2 \otimes_k L_2 & \xrightarrow{S_2} & L_2 \otimes_k L_2 \end{array}$$

FIG. 2. A T-Lie algebra morphism.

is called T-Lie algebra with basis $\mathscr{B}$ *(or basic T-Lie algebra) if, for* $S_{12}=S\otimes_k Id_L$, $S_{23}=Id_L\otimes_k S$, *the following axioms are satisfied:*

(1) (a) $S^2=Id$.
  (b) $S(x\otimes y)=q_{x,y}y\otimes x$, *for certain* $q_{x,y}\in k$, $q_{x,x}=1$, $\forall x,y\in\mathscr{B}$.
  (c) *(Multiplicativity)*
    (i) $S(Id\otimes_k[,])|_{L^3}=([,]\otimes_k Id)S_{23}S_{12}|_{L^3}$.
    (ii) $S([,]\otimes_k Id)|_{L^3}=(Id\otimes_k[,])S_{12}S_{23}|_{L^3}$.
(2) *(Stability)*
  (a) *There exists a strict grading,*

$$L=\oplus_\eta L_\eta,$$

  of L relative to [,].
  (b)

$$\langle L_{\eta_1}\otimes L_{\eta_2}\rangle\subseteq(L\otimes_k L)_{\eta_1+\eta_2-1},$$

  for all L_{η_1}, L_{η_2}.
(3) $T=S+\langle,\rangle$.
(4) *(Antisymmetry)*
  (a) $[,]T=-[,]$.
  (b) $\langle,\rangle S=-\langle,\rangle$.
  (c) $[,]\langle,\rangle=0$.
(5) *(Jacobi identity)*

$$[,]((Id\otimes_k[,])S_{12}S_{23}-([,]\otimes_k Id)S_{23}S_{12}+(Id\otimes_k[,])S_{23}S_{12})|_{3_L}=0.$$

Multiplicativity conditions are used to control commutation relations in the universal enveloping algebra, whereas stability conditions are used to obtain a good gradation in the corresponding symmetric algebra.

Definition II.4: Let L_i *be a basic T-Lie algebra with bracket* $[,]_i$, *pseudobracket* $\langle,\rangle_i$ *and presymmetry* S_i, $i=1,2$. *A k morphism* $f:L_1\to L_2$ *is called a T-Lie morphism if f is a morphism of graded algebras relative to* $[,]_i$, $i=1,2$, *and the diagrams in Fig. 2 commute*.

III. EXAMPLES

In order to obtain a graduation in the stability conditions, it suffices to define a map $\eta:\mathscr{B}\to\mathbb{N}$ having properties [(2)(a)] and [(2)(b)] in the stability axiom. This remark will be used in the following examples.

A. Some common Lie algebras

Example III.1: Classical Lie algebras over fields are basic T-Lie algebras:

$$[,]\ \text{classical bracket},\quad \langle,\rangle=0,\quad T=S\ \text{usual switch},\quad \eta=1.$$

Example III.2: Lie superalgebras over fields[4] are basic T-Lie algebras.

Let $L=L_0\oplus L_1$ be a Lie superalgebra with bracket $[,]$ over a field k. Let $\mathscr{B}_\alpha$ be a basis of L_α, $\alpha=0,1$. Define $S:L\otimes_k L\to L\otimes_k L$ on the basis $\mathscr{B}=\mathscr{B}_0\cup\mathscr{B}_1$, by $S(x\otimes y)=(-1)^{\alpha\beta}y\otimes x$ if $x\in L_\alpha$, $y\in L_\beta$. Besides,

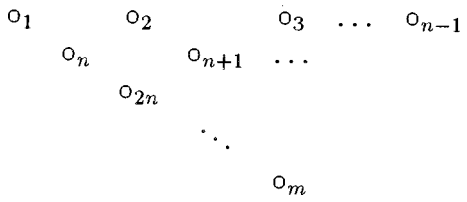

FIG. 3. The basic T-Lie algebra $(sl^+_{n+1})_q$.

$$\langle ,\rangle = 0, \quad T = S, \quad \eta = 1.$$

Example III.3 (Lie color algebras)*:* Let k be a field of characteristic zero. Let

$$L = \oplus_{\gamma\in\Gamma} L_\gamma$$

be a ϵ Lie algebra,[5] where Γ is an Abelian group and ϵ is a commutation factor on Γ. Let $[,]$ be a bracket of L. Put $\mathscr{B}_\gamma$ basis of L_γ for each $\gamma\in\Gamma$.

Define

$$S(x\otimes y) = \epsilon(\alpha,\beta) y\otimes x, \quad \text{if } x\in\mathscr{B}_\alpha, \quad y\in\mathscr{B}_\beta,$$

besides $\langle ,\rangle = 0$ and $\eta = 1$.

Multiplicativity conditions follow easily from the definition of the commutation factor. We conclude that every ϵ Lie algebra is a T-Lie algebra.

B. Linear *T*-Lie algebras

Example III.4: Let e_{ij}, $1\leqslant i, j\leqslant n$ be standard basis of gl_n matrices $n\times n$ over a field $\mathbb{K}$. Let $[,]$ be the usual bracket in gl_n, sl^+_n the Lie subalgebra of upper triangular matrices having trace zero. We put $x_i = e_{i,i+1}$, $i=1,\ldots,n-1$, $x_n=[x_1,x_2]$, $x_{n+1}=[x_2,x_3],\ldots,$ $x_{2n-3}=[x_{n-2},x_{n-1}]$, $x_{2n-2}=[x_1,x_{n+1}]\ldots,$ $x_{3n-6}=[x_{n-3},x_{2n-3}]$, $x_{3n-5}=[x_1,x_{2n-1}]$, $x_{3n-4}=[x_2,x_{2n}],\ldots,$ $x_m = [x_1,x_{m-1}]$, where $m=n(n-1)/2$, besides, we define $h_i=[x_i,x_i^t]$, $i=1,\ldots,m$ diagonal matrices in sl_n. Further, $q=\exp(t)\in\mathbb{K}[[t]]$, the formal series ring with indeterminate t and coefficients in $\mathbb{K}$, $k=\mathbb{K}[q,q^{-1}]$, $c_{i,j}\in\mathbb{Z}$, such that $[h_i,x_j]=c_{i,j}x_j$, $1\leqslant i, j\leqslant m$.

Let $(sl^+_n)_q$ be a free k module with basis $\mathscr{B}=\{x_i | 1\leqslant i\leqslant m\}$. We may define an order in $\mathscr{B}$ according to Fig. 3, from left to right and top to bottom. For example, $x_1<x_n<x_2<x_{2n}$. The first time that a diagram (Auslander–Reiten quiver of type A_{n-1}) of this type appears related to quantum groups, is in Ringel's work about the relationship between Poincaré–Birkhoff–Witt bases, quantum groups, and Hall algebras.[13]

Define

$$[x,y]_q = [x,y], \quad \text{if } x<y\in\mathscr{B},$$

$$\langle e_{ij},e_{uv}\rangle = \begin{cases} (q-q^{-1})e_{iv}\otimes e_{uj}, & \text{if } i<u<j, \quad u<j<v, \\ (q^{-1}-q)e_{uj}\otimes e_{iv}, & \text{if } u<i<v, \quad i<v<j, \\ 0, \text{ otherwise,} \end{cases}$$

$$S(x_i\otimes x_j) = q^{c_{i,j}} x_j\otimes x_i, \quad \text{if } x_i<x_j,$$

$$T = S + \langle ,\rangle.$$

Finally, we define η in such way that every basic element in Fig. 3 is in correspondence with a number belonging to Fig. 4; this yields, $\eta(e_{ij}) = i(j-i)$, $\forall i,j$.

The multiplicativity condition follows from the properties

$$x_i<x_j<x_l \Rightarrow \begin{cases} [x_i,x_j]<x_l, & \text{if } [x_i,x_j]\neq 0, \\ x_i<[x_j,x_l], & \text{if } [x_j,x_l]\neq 0, \end{cases}$$

$$\begin{array}{ccccccc} 1 & & 2 & & 3 & \ldots & n \\ & 2 & & 4 & & \ldots & 2(n-1) \\ & & 3 & & 6 & \ldots & 3(n-2) \\ & & & \ddots & & & \\ & & & & n & & \end{array}$$

FIG. 4. The graduation of $(sl^{+}_{n+1})_q$.

and

$$[h_i,[x_j,x_l]]=(c_{i,j}+c_{i,l})[x_j,x_l].$$

In the cases $n=2,3,4,5$, the Jacobi identity for $[,]_q$ can be verified by straightforward calculations. We get that $(sl^{+}_{n})_q$ with bracket $[,]_q$ is a basic T-Lie algebra, $n=2,3,4,5$.

Similarly, we can define $(sl^{-}_{n})_q$.

Remark III.1: We get

$$(sl^{\pm}_{n})_q|_{t=0}=sl^{\pm}_{n},$$

so in the cases $n=2,3,4,5$, $(sl^{\pm}_{n})_q$ is a deformation of $sl^{\pm}_{n}$ in the category of T-Lie algebras. Later, in Sec. X, such a property will be generalized for every n.

Example III.5: Starting from $(sl^{+}_{4})_q$ we are going to build a new basic T-Lie algebra, denoted $(\widetilde{sl^{+}_{4}})_q$. Its structure is

$$\widetilde{[,]}_q=[,]_q, \quad \widetilde{\langle,\rangle}=0, \quad \tilde{S}=S, \quad \widetilde{T}=S, \quad \tilde{\eta}=\eta.$$

Example III.6 [nonstandard quantum deformations[12] of $GL(n)$]*:* Let p,q be units in a commutative unitary ring k with $pq\neq 1$ and choose $n(n-1)/2$ discrete parameters ϵ_{ij}, $\epsilon_{ij}=\pm 1$, $1\leqslant i<j\leqslant n$, $\epsilon_{ii}=1$, $\epsilon_{ji}=\epsilon_{ij}$.

The k module $L_{p,q,\epsilon}(n,k)$ is then defined to be the free k module with basis

$$\mathscr{B}=\{Z^{j}_{i}|1\leqslant i,j\leqslant n\}.$$

We ordered $\mathscr{B}$ by putting $Z^{j}_{i}>Z^{v}_{u}$ if either $i>u$, or $i=u$ and $j>v$. Define η by $\eta(Z^{i}_{j})=j3^{i-1}$. Besides, we put $[,]=0$, and if $Z^{l}_{i}>Z^{v}_{u}$,

$$\langle Z^{l}_{i},Z^{v}_{u}\rangle=\begin{cases}\epsilon_{vl}(p-q^{-1})Z^{v}_{i}\otimes Z^{l}_{u}, & \text{if } i>u \text{ and } l>v,\\ 0, \text{ otherwise.}\end{cases} \tag{5}$$

$$S(Z^{l}_{i}\otimes Z^{v}_{u})=\begin{cases}\epsilon_{vl}pZ^{v}_{u}\otimes Z^{l}_{i}, & \text{if } i=u,\ l>v\\ \epsilon_{ui}qZ^{v}_{u}\otimes Z^{l}_{i}, & \text{if } l=v,\quad i>u,\\ \epsilon_{iu}\epsilon_{vl}p^{-1}qZ^{v}_{u}\otimes Z^{l}_{i}, & \text{if } i>u,\quad v>l,\\ \epsilon_{vl}\epsilon_{ui}Z^{v}_{u}\otimes Z^{l}_{i}, & \text{if } i>u,\quad l>v.\end{cases}$$

To prove that $L_{p,q,\epsilon}(n,k)$ is a basic T-Lie algebra, since (3.5) it suffices to check the stability condition [(2)(b)] for $Z^{l}_{i}>Z^{v}_{u}$, such that $i>u$ and $l>v$:

$$(i-u)3^{l-1}>(i-u)3^{v-1};$$

then

$$i3^{l-1}+u3^{v-1}>i3^{v-1}+u3^{l-1},$$

but this equation has a left side $\eta(Z^{l}_{i})+\eta(Z^{v}_{u})$, whereas the right side is $\eta(Z^{v}_{i})+\eta(Z^{l}_{u})$. This proves stability conditions.

IV. UNIVERSAL ENVELOPING ALGEBRAS

A. Construction of $U(L)$

Definition IV.1: Let L be a T-Lie algebra with basis $\mathscr{B}$, and $\otimes_k L$ the k-tensor algebra of the module L. The universal enveloping algebra $U(L)$ is the quotient

$$U(L)=\otimes_k L/J,$$

where J is the two-sided ideal generated by

$$x\otimes y-T(x\otimes y)-[x,y], \quad x,y\in\mathscr{B}.$$

Because of the stability axiom [(2)(b)], the algebra $U(L)$ have a similar structure to a *quadratic algebra with an ordering algorithm.*[14]

B. Examples

Example IV.2:

$$U(L_{p,q,\epsilon}(n,k))=M_{p,q,\epsilon}(n,k)$$

is a nonstandard quantum deformation[12] of $GL(n)$.

Example IV.3:

$$U(sl^{\pm}_{n+1})_q=U^{\pm}_q(sl_{n+1}),$$

the positive (negative) part of the Drinfeld–Jimbo quantum group of type A_n, $n=3,4$.

Example IV.4: In $U(\widetilde{sl^{+}_4})_q$, the equation $x_2x_6=0$ holds. Then $U(\widetilde{sl^{+}_4})_q$ is an enveloping algebra where the PBW theorem does not hold. So, if we want a good enveloping algebra, we have to add conditions to the T-Lie algebras.

Besides, if $\beta,\widetilde{S}$ denotes the bracket and the symmetry of $(\widetilde{sl^{+}_4})_q$, respectively, and the characteristic of the field $\mathbb{K}$ is zero, then for $\gamma=Id-\widetilde{S}$, where Id is the identity morphism on $(\widetilde{sl^{+}_4})^{2\otimes}_q$, the condition $\gamma(t)=0$ implies $\beta(t)=0$. Moreover, if $\mathscr{B}$ is the canonical basis of $(\widetilde{sl^{+}_4})_q$, and x,y,z are arbitrary elements in $\mathscr{B}$, straightforward calculations (using MATHEMATICA[15]) gives

$$\beta(\beta(x,y),z)=\beta(x,\beta(y,z))-q_{x,y}\beta(y,\beta(x,z))$$

$$\beta(z,\beta(x,y))=\beta(\beta(z,x),y)-q_{x,y}\beta(\beta(x,y),x)$$

where $\widetilde{S}(x\otimes y)=q_{x,y}y\otimes x$. This means that $(\widetilde{sl^{+}_4})_q$ has a structure of *balanced generalized Lie algebra*[9] and its universal enveloping algebra as such a generalized Lie algebra is the same as the T-Lie algebra. Therefore, the generalized Lie algebra axioms of Lyubashenko–Sudbery are not enough in order to obtain a PBW theorem.

Example IV.5: Let L be a basic T-Lie algebra. We are going to define a new T-Lie algebra L^0: $L^0=L$ in its structure of the k module, $[,]^0=0$, $\langle,\rangle^0=0$, $\eta^0=\eta$, $S^0=S$, and define

$$\mathscr{S}(L)=U(L^0);$$

$\mathscr{S}(L)$ is a free k module with a basis of the monomials formed by finite nondecreasing elements of $\mathscr{B}$. Such an object $\mathscr{S}(L)$ will be called the *q-symmetric algebra of L*.

V. THE RELATIONSHIP BETWEEN UNIVERSAL ENVELOPING ALGEBRAS AND SYMMETRIC ALGEBRAS

Let L be a T-Lie algebra with basis $\mathscr{B}$, $x_\lambda\in\mathscr{B}$, $\Sigma=(x_{\lambda_1},\ldots,x_{\lambda_u})$ finite nondecreasing sequence of elements of $\mathscr{B}$. We write $z_\lambda=x_\lambda\in\mathscr{S}(L)$, $z_\Sigma=z_{\lambda_1}\cdots z_{\lambda_u}\in\mathscr{S}(L)$, $z_\varnothing=1\in\mathscr{S}(L)$, $\eta(\lambda)=\eta(x_\lambda)$, $\eta(\Sigma)=\eta(z_\Sigma)=\eta(x_{\lambda_1})+\cdots+\eta(x_{\lambda_n})$. Besides, we put $x_\lambda\leqslant\Sigma$ if $x_\lambda\leqslant x_{\lambda_1}$.

Lemma V.1 (A–B)*:* Let L be a T-Lie algebra with basis $\mathscr{B}$. $\mathscr{P}=\mathscr{S}(L)$ q-symmetric algebra, $\mathscr{P}_p$ the k submodule generated by z_Σ such that $\eta(\Sigma)\leqslant p$. There is a k morphism,

$$_\cdot_ : L\otimes_k \mathscr{P}\to\mathscr{P}, \tag{6}$$

satisfying
(A) $x_\lambda\cdot z_\Sigma = z_\lambda z_\Sigma$, for $x_\lambda\leqslant\Sigma$;
(B) $x_\lambda\cdot z_\Sigma - z_\lambda z_\Sigma\in\mathscr{P}_{\eta(\lambda)+\eta(\Sigma)-1}$,

Proof: By induction on $\eta(\lambda)+\eta(\Sigma)$. If $\eta(\lambda)+\eta(\Sigma)=1$ then $\eta(\lambda)=1$ and $\Sigma=\varnothing$, it follows that $z_\varnothing=1$. Then define

$$x_\lambda\cdot 1 = z_\lambda,$$

so (A) and (B) holds. Assume the existence of $x_{\lambda'}\cdot z_{\Sigma'}$ for $\eta(\lambda')+\eta(\Sigma')<\eta(\lambda)+\eta(\Sigma)$ satisfying (A) and (B). We have to define $x_\lambda\cdot z_\Sigma$.

There are two cases: $\lambda\leqslant\Sigma$ or $\lambda\not\leqslant\Sigma$.

Case $\lambda\leqslant\Sigma$: Because (A):

$$x_\lambda\cdot z_\Sigma = z_\lambda z_\Sigma.$$

Case $\lambda\not\leqslant\Sigma$: We may write $\Sigma=(x_\mu,N)$ with $x_\mu\leqslant N$ and $x_\lambda>x_\mu$. Since $\eta(N)<\eta(\Sigma)$ and because at the induction hypothesis $x_\lambda\cdot z_N$ is already defined, and

$$w = x_\lambda\cdot z_N - z_\lambda z_N\in\mathscr{P}_{\eta(\lambda)+\eta(N)-1}.$$

Moreover, from $\eta(\mu)+\eta(\lambda)+\eta(N)-1<\eta(\mu)+\eta(\lambda)+\eta(N)=\eta(\lambda)+\eta(\Sigma)$ it follows that $x_\mu\cdot w$ is already defined.

We have

$$T(x_\lambda\otimes x_\mu) = q_{\lambda\mu}x_\mu\otimes x_\lambda + \sum_i \xi_i x_{\mu_i}\otimes x_{\lambda_i}, \tag{7}$$

and because at (B) and the induction hypothesis $x_{\lambda_i}\cdot z_N\in\mathscr{P}_{\eta(\lambda_i)+\eta(N)}$. As a consequence, $x_{\mu_i}\cdot(x_{\lambda_i}\cdot z_N)$ is already defined, because $\eta(\mu_i)+\eta(\lambda_i)+\eta(N)<\eta(\lambda)+\eta(\mu)+\eta(N)$ according to the stability axiom.

We may define

$$x_\lambda\cdot z_\Sigma = q_{\lambda\mu}z_\mu z_\lambda z_N + q_{\lambda\mu}x_\mu\cdot w + \sum_i \xi_i x_{\mu_i}\cdot(x_{\lambda_i}\cdot z_N) + [x_\lambda,x_\mu]\cdot z_N, \tag{8}$$

where $w=x_\lambda\cdot z_N - z_\lambda z_N$; $[x_\lambda,x_\mu]\cdot z_N$ is already defined because $\eta([x_\lambda,x_\mu])+\eta(N)<\eta(\lambda)+\eta(\mu)+\eta(N)=\eta(\lambda)+\eta(\Sigma)$.

Now it only remains to prove (B). From $z_\lambda z_\Sigma = q_{\lambda\mu}z_\lambda z_N$, we obtain

$$x_\lambda\cdot z_\Sigma - z_\lambda z_\Sigma = q_{\lambda\mu}x_\mu\cdot w + \sum_i \xi_i x_{\mu_i}\cdot x_{\lambda_i}\cdot z_N + [x_\lambda,x_\mu]\cdot z_N.$$

Moreover,

$$x_\mu\cdot w\in\mathscr{P}_{\eta(\mu)+\eta(w)}=\mathscr{P}_{\eta(\mu)+\eta(\lambda)+\eta(N)-1}=\mathscr{P}_{\eta(\lambda)+\eta(\Sigma)-1},$$

$$x_{\mu_i}\cdot x_{\lambda_i}\cdot z_N\in\mathscr{P}_{\eta(\mu_i)+\eta(\lambda_i)+\eta(N)}\subseteq\mathscr{P}_{\eta(\mu)+\eta(\lambda)-1+\eta(N)}=\mathscr{P}_{\eta(\lambda)+\eta(\Sigma)-1},$$

$$[x_\lambda,x_\mu]\cdot z_N\in\mathscr{P}_{\eta(\lambda)+\eta(\mu)-1+\eta(N)}=\mathscr{P}_{\eta(\lambda)+\eta(\Sigma)-1},$$

imply

$$x_\lambda\cdot z_\Sigma - z_\lambda z_\Sigma\in\mathscr{P}_{\eta(\lambda)+\eta(\Sigma)-1}.$$

□

Definition V.1: Let L be a T-Lie algebra with basis $\mathscr{B}$. We call L adequate if the morphism from the lemma (A–B) is such that the condition

$$x_{\lambda'}\cdot x_{\mu'}\cdot z_M - T(x_{\lambda'}\otimes x_{\mu'})\cdot z_M = [x_{\lambda'},x_{\mu'}]\cdot z_M, \tag{9}$$

for all $\eta(x_{\lambda'})+\eta(x_{\mu'})+\eta(M)\leqslant r$ *implies*

$$\begin{aligned}
&\langle x_\lambda,x_\mu\rangle\cdot x_\gamma\cdot z_N - q_{\lambda\mu}[x_\mu,[x_\lambda,x_\gamma]]\cdot z_N\\
&\quad = q_{\mu\gamma}q_{\lambda\gamma}\langle x_\gamma[x_\lambda,x_\mu]\rangle\cdot z_N + q_{\mu\gamma}q_{\lambda\gamma}x_\gamma\cdot\langle x_\lambda,x_\mu\rangle\cdot z_N + q_{\mu\gamma}\langle x_\lambda,x_\gamma\rangle\cdot x_\mu\cdot z_N\\
&\quad - q_{\lambda\mu}x_\mu\cdot\langle x_\lambda,x_\gamma\rangle\cdot z_N + q_{\mu\gamma}[x_\lambda,x_\gamma]\cdot x_\mu\cdot z_N - q_{\lambda\mu}x_\mu\cdot[x_\lambda,x_\gamma]\cdot z_N + x_\lambda\cdot\langle x_\mu,x_\gamma\rangle\cdot z_N\\
&\quad - q_{\lambda\mu}q_{\lambda\gamma}\langle x_\mu,x_\gamma\rangle\cdot x_\lambda\cdot z_N - q_{\lambda\gamma}q_{\lambda\mu}\langle[x_\mu,x_\gamma],x_\lambda\rangle\cdot z_N,
\end{aligned} \tag{10}$$

for every $x_\lambda>x_\mu>x_\gamma\in\mathscr{B}$, $x_\gamma\leqslant z_N$, *such that* $\eta(x_\lambda)+\eta(x_\mu)+\eta(x_\gamma)+\eta(N)\leqslant r+1$.

Lemma V.2 (C)*:* Let L be an adequate T-Lie algebra with basis $\mathscr{B}$, and $\mathscr{P}$ the related q-symmetric algebra. Then there exists a k morphism $_\cdot_: L\otimes_k\mathscr{P}\to\mathscr{P}$, such that

$$\text{(C)}\quad x_\lambda\cdot x_\mu\cdot z_N = T(x_\lambda\otimes x_\mu)\cdot z_N + [x_\lambda,x_\mu]\cdot z_N,\quad \forall z_N\in\mathscr{P},\quad \forall x_\lambda,\quad x_\mu\in\mathscr{B}.$$

Proof: Let $_\cdot_$ be the morphism from lemma (A–B). There are two cases:

1. $\mu\leqslant N$ or $\lambda\leqslant N$;
2. $\mu\not\leqslant N$ and $\lambda\not\leqslant N$.

(1) Assume $\mu\leqslant N$ and $\mu<\lambda$. Let $M=(\mu,N)$; then, by definition,

$$\begin{aligned}
x_\lambda\cdot x_\mu\cdot z_N &= x_\lambda z_M,\quad \text{where } \lambda\not\leqslant M,\\
&= z_\lambda\cdot z_M + q_{\lambda\mu}x_\mu\cdot w + \langle x_\lambda,x_\mu\rangle\cdot z_N + [x_\lambda,x_\mu]\cdot z_N.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
T(x_\lambda\otimes x_\mu)\cdot z_N + [x_\lambda,x_\mu]\cdot z_N &= q_{\lambda\mu}x_\mu\cdot x_\lambda\cdot z_N + \langle x_\lambda,x_\mu\rangle\cdot z_N + [x_\lambda,x_\mu]\cdot z_N\\
&= q_{\lambda\mu}x_\mu\cdot(z_\lambda z_N) + q_{\lambda\mu}x_\mu\cdot w + \langle x_\lambda,x_\mu\rangle\cdot z_N + [x_\lambda,x_\mu]\cdot z_N,
\end{aligned}$$

since $\mu<\lambda$ and $\mu\leqslant N$, it holds that $z_\lambda z_N = c z_{N'}$, where $\mu\leqslant N'$ and $c\in k$,

$$x_\mu\cdot(z_\lambda z_N) = c x_\mu\cdot z_{N'} = c z_\mu z_{N'} = z_\mu z_\lambda z_N,$$

so

$$q_{\lambda\mu}x_\mu\cdot(z_\lambda z_N) = q_{\lambda\mu}z_\mu z_\lambda z_N = z_\lambda z_\mu z_N = z_\lambda z_M;$$

therefore

$$T(x_\lambda\otimes x_\mu)\cdot z_N + [x_\lambda,x_\mu]\cdot z_N = z_\lambda z_M + q_{\lambda\mu}x_\mu\cdot w + \langle x_\lambda,x_\mu\rangle\cdot z_N + [x_\lambda,x_\mu]\cdot z_N = x_\lambda\cdot x_\mu\cdot z_N$$

[i.e., (C) holds for $\mu<\lambda$]. It follows, multiplying by $-q_{\mu\lambda}$:

$$-q_{\mu\lambda}x_\lambda\cdot x_\mu\cdot z_N = -x_\mu\otimes x_\lambda\cdot z_N - q_{\mu\lambda}\langle x_\lambda,x_\mu\rangle\cdot z_N - q_{\mu\lambda}[x_\lambda,x_\mu]\cdot z_N. \tag{11}$$

This implies, using antisymmetry,

$$x_\mu\cdot x_\lambda\cdot z_N - T(x_\mu\otimes x_\lambda)\cdot z_N = [x_\mu,x_\lambda]\cdot z_N, \tag{12}$$

and we conclude that (C) also holds for $\lambda<\mu$.

(2) Let $N=(\gamma,Q)$ where $\gamma\leqslant Q$, $\gamma<\lambda$, $\gamma<\mu$. We proceed by induction on $\eta(\lambda)+\eta(\mu)+\eta(N)$. Suppose that for each $\eta(\lambda')+\eta(\mu')+\eta(N')\leqslant r$, (C) holds. Then, for $\eta(\lambda)+\eta(\mu)+\eta(N)\leqslant r+1$, we have

$$x_\mu \cdot z_N = x_\mu \cdot (x_\gamma \cdot z_Q) = T(x_\mu \otimes x_\gamma) \cdot z_Q + [x_\mu, x_\gamma] \cdot z_Q, \tag{13}$$

because $\eta(\mu)+\eta(\gamma)+\eta(Q)=\eta(\mu)+\eta(N)\leqslant r$ and the induction hypothesis.

Now, $x_\mu \cdot z_Q = z_\mu z_Q + w$, where $w \in \mathscr{P}_{\eta(\mu)+\eta(Q)-1}$. We may apply (C) to $x_\lambda \cdot x_\gamma \cdot (z_\mu z_Q)$, since $z_\mu z_Q = c z_{Q'}$, where $c \in k$ and $\gamma \leqslant Q'$ because $\gamma \leqslant Q$, $\gamma < \mu$ and case (A).

Also, (C) applies to $x_\lambda \cdot x_\gamma \cdot w$ since

$$\eta(\lambda)+\eta(\gamma)+\eta(w) \leqslant \eta(\lambda)+\eta(\gamma)+\eta(\mu)+\eta(Q)-1 = \eta(\lambda)+\eta(\gamma)+\eta(N)-1 \leqslant r,$$

and the induction hypothesis.

The preceding remarks show that (C) applies to

$$x_\lambda \cdot x_\gamma \cdot x_\mu \cdot z_Q = x_\lambda \cdot x_\gamma \cdot (z_\mu z_Q) + x_\lambda \cdot x_\gamma \cdot w.$$

Using (13) and multiplying by x_λ,

$$\begin{aligned} x_\lambda \cdot x_\mu \cdot z_N &= x_\lambda \cdot T(x_\mu \otimes x_\gamma) \cdot z_Q + x_\lambda \cdot [x_\mu, x_\gamma] \cdot z_Q \\ &= q_{\mu\gamma} x_\lambda \cdot x_\gamma \cdot x_\mu \cdot z_Q + x_\lambda \cdot \langle x_\mu, x_\gamma \rangle \cdot z_Q + x_\lambda \cdot [x_\mu, x_\gamma] \cdot z_Q \\ &= q_{\mu\gamma} q_{\lambda\gamma} x_\gamma \cdot x_\lambda \cdot x_\mu \cdot z_Q + q_{\mu\gamma} \langle x_\lambda, x_\gamma \rangle \cdot x_\mu \cdot z_Q + q_{\mu\gamma} [x_\lambda, x_\gamma] \cdot x_\mu \cdot z_Q \\ &\quad + x_\lambda \cdot \langle x_\mu, x_\gamma \rangle \cdot z_Q + x_\lambda \cdot [x_\mu, x_\gamma] \cdot z_Q. \end{aligned}$$

Recall that λ, μ are interchangeable:

$$\begin{aligned} x_\mu \cdot x_\lambda \cdot z_N &= q_{\lambda\gamma} q_{\mu\gamma} x_\gamma \cdot x_\mu \cdot x_\lambda \cdot z_Q + q_{\lambda\gamma} \langle x_\mu, x_\gamma \rangle \cdot x_\lambda \cdot z_Q + q_{\lambda\gamma} [x_\mu, x_\gamma] \cdot x_\lambda \cdot z_Q \\ &\quad + x_\mu \cdot \langle x_\lambda, x_\gamma \rangle \cdot z_Q + x_\mu \cdot [x_\lambda, x_\gamma] \cdot z_Q. \end{aligned}$$

Now use $\eta(\lambda)+\eta(\mu)+\eta(Q)=\eta(\lambda)+\eta(N)\leqslant r$ to obtain

$$\begin{aligned} x_\lambda \cdot x_\mu \cdot z_N - q_{\lambda\mu} x_\mu \cdot x_\lambda \cdot z_N &= q_{\mu\gamma} q_{\lambda\gamma} x_\gamma \cdot ([x_\lambda, x_\mu] + \langle x_\lambda, x_\mu \rangle) \cdot z_Q + q_{\mu\gamma} \langle x_\lambda, x_\gamma \rangle \cdot x_\mu \cdot z_Q - q_{\lambda\mu} x_\mu \\ &\quad \cdot \langle x_\lambda, x_\gamma \rangle \cdot z_Q + q_{\mu\gamma} [x_\lambda, x_\gamma] \cdot x_\mu \cdot z_Q - q_{\lambda\mu} x_\mu \cdot [x_\lambda, x_\gamma] \cdot z_Q + x_\lambda \\ &\quad \cdot \langle x_\mu, x_\gamma \rangle \cdot z_Q - q_{\lambda\mu} q_{\lambda\gamma} \langle x_\mu, x_\gamma \rangle \cdot x_\lambda \cdot z_Q + x_\lambda \cdot [x_\mu, x_\gamma] \cdot z_Q \\ &\quad - q_{\lambda\mu} q_{\lambda\gamma} [x_\mu, x_\gamma] \cdot x_\lambda \cdot z_Q. \end{aligned} \tag{14}$$

Furthermore,

$$\begin{aligned} &x_\lambda \cdot [x_\mu, x_\gamma] \cdot z_Q - q_{\lambda\mu} q_{\lambda\gamma} [x_\mu, x_\gamma] \cdot x_\lambda \cdot z_Q \\ &\quad = -q_{\mu\gamma} x_\lambda \cdot [x_\gamma, x_\mu] + q_{\lambda\mu} q_{\lambda\gamma} q_{\mu\gamma} [x_\gamma, x_\mu] \cdot x_\lambda \cdot z_Q \\ &\quad = q_{\mu\gamma} q_{\lambda\gamma} q_{\lambda\mu} ([x_\gamma, x_\mu] \cdot x_\lambda \cdot z_Q - q_{\gamma\lambda} q_{\mu\lambda} x_\lambda \cdot [x_\gamma, x_\mu] \cdot z_Q). \end{aligned} \tag{15}$$

If we suppose that $x_\mu < x_\lambda$, then we can make use of the multiplicativity condition, and since $\eta([x_\gamma, x_\mu])+\eta(\lambda)+\eta(Q) < \eta(\lambda)+\eta(\mu)+\eta(\gamma)+\eta(Q) = \eta(\lambda)+\eta(\mu)+\eta(N)$, we obtain that (15) is equal to

$$\begin{aligned} &= q_{\mu\gamma} q_{\lambda\gamma} q_{\lambda\mu} [[x_\gamma, x_\mu], x_\lambda] \cdot z_Q + q_{\mu\gamma} q_{\lambda\gamma} q_{\lambda\mu} \langle [x_\gamma, x_\mu], x_\lambda \rangle \cdot z_Q \\ &= -q_{\lambda\gamma} q_{\lambda\mu} [[x_\mu, x_\gamma], x_\lambda] \cdot z_Q - q_{\lambda\gamma} q_{\lambda\mu} \langle [x_\mu, x_\gamma], x_\lambda \rangle \cdot z_Q. \end{aligned} \tag{16}$$

Using multiplicativity again and since $\eta(x_\gamma)+\eta([x_\lambda, x_\mu])+\eta(Q) < \eta(x_\lambda)+\eta(\mu)+\eta(\gamma)+\eta(Q) = \eta(x_\lambda)+\eta(x_\mu)+\eta(N)$, we may write

$$x_\gamma \cdot [x_\lambda, x_\mu] \cdot z_Q = q_{\gamma\mu} q_{\gamma\lambda} [x_\lambda, x_\mu] \cdot x_\gamma \cdot z_Q + [x_\gamma, [x_\lambda, x_\mu]] \cdot z_Q + \langle x_\gamma, [x_\lambda, x_\mu] \rangle \cdot z_Q. \tag{17}$$

Substitute (16) and (17) in (14),

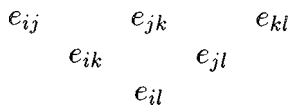

FIG. 5. A basic T-Lie algebra of type $(sl_4^+)_q$.

$$\begin{aligned} x_\lambda \cdot x_\mu \cdot z_N - q_{\mu\lambda} x_\mu \cdot x_\lambda \cdot z_N = {} & [x_\lambda, x_\mu] \cdot x_\gamma \cdot z_Q + q_{\mu\gamma} q_{\lambda\gamma} [x_\gamma, [x_\lambda, x_\mu]] \cdot z_Q + q_{\mu\gamma} q_{\lambda\gamma} \langle x_\gamma, [x_\lambda, x_\mu] \rangle \\ & \cdot z_Q + q_{\mu\gamma} q_{\lambda\gamma} x_\gamma \cdot \langle x_\lambda, x_\mu \rangle \cdot z_Q + q_{\mu\gamma} \langle x_\lambda, x_\gamma \rangle \cdot x_\mu \cdot z_Q - q_{\lambda\mu} x_\mu \langle x_\lambda, x_\gamma \rangle \\ & \cdot z_Q + q_{\mu\gamma} [x_\lambda, x_\gamma] \cdot x_\mu \cdot z_Q - q_{\lambda\mu} x_\mu [x_\lambda, x_\gamma] \cdot z_Q + x_\lambda \cdot \langle x_\mu, x_\gamma \rangle \cdot z_Q \\ & - q_{\lambda\mu} q_{\lambda\gamma} \langle x_\mu, x_\gamma \rangle \cdot x_\lambda \cdot z_Q - q_{\lambda\gamma} q_{\lambda\mu} [[x_\mu, x_\gamma], x_\lambda] \cdot z_Q \\ & - q_{\lambda\gamma} q_{\lambda\mu} \langle [x_\mu, x_\gamma], x_\lambda \rangle \cdot z_Q, \end{aligned}$$

since L is adequate,

$$\begin{aligned} = {} & [x_\lambda, x_\mu] \cdot x_\gamma \cdot z_Q + \langle x_\lambda, x_\mu \rangle \cdot x_\gamma \cdot z_Q + q_{\mu\gamma} q_{\lambda\gamma} [x_\gamma, [x_\lambda, x_\mu]] \cdot z_Q \\ & - q_{\lambda\gamma} q_{\lambda\mu} [[x_\mu, x_\gamma], x_\lambda] \cdot z_Q - q_{\lambda\mu} [x_\mu, [x_\lambda, x_\gamma]] \cdot z_Q. \end{aligned}$$

Thanks to the Jacobi identity and (A), we get

$$x_\lambda \cdot x_\mu \cdot z_N - q_{\lambda\mu} x_\mu \cdot x_\lambda \cdot z_N - \langle x_\lambda, x_\mu \rangle \cdot z_N = [x_\lambda, x_\mu] \cdot z_N, \tag{18}$$

if $x_\mu < x_\lambda$.

Multiplying both sides of (18) by $-q_{\lambda\mu}$ and using antisymmetry, we get

$$x_\mu \cdot x_\lambda \cdot z_N - q_{\lambda\mu} x_\lambda \cdot x_\mu \cdot z_N - \langle x_\mu, x_\lambda \rangle \cdot z_N = [x_\mu, x_\lambda] \cdot z_N,$$

so (18) also holds if $x_\lambda < x_\mu$. □

VI. REPRESENTATIONS

Definition VI.1: Let L be a basic T-Lie algebra and V a k module. A k morphism $_ \cdot _ : L \otimes_k V \to V$ is called a representation of L if it satisfies

$$x \cdot y \cdot v - T(x \otimes y) \cdot v = [x, y] \cdot v, \quad \forall x, y \in L, \quad \forall v \in V,$$

where $(a \otimes b) \cdot v$ means $a \cdot b \cdot v$.

Theorem VI.1: *If L is an adequate basic T-Lie algebra then L has a natural representation on its q-symmetric algebra $\mathcal{S}(L)$.*

Corollary VI.2: If L is an adequate basic T-Lie algebra, then its universal enveloping algebra $U(L)$ has a representation on the q-symmetric algebra $\mathcal{S}(L)$.

Inside $(sl_n^+)_q$, $n \geqslant 4$, the k submodules generated by the basic elements given in Fig. 5 have a structure of basic T-Lie algebra that looks like $(sl_4^+)_q$. But such an algebra has a graduation given by $\eta(e_{ab}) = a(b-a)$; this is not, in general, the graduation of $(sl_4^+)_q$. However, Fig. 5 is still a basic T-Lie algebra.

The basic T-Lie algebras given by Fig. 5 will be called *of type* $(sl_4^+)_q$. In a similar way we may define *basic T-Lie algebras of type* $(sl_n^\pm)_q$.

Example VI.3: Every basic T-Lie algebra of type $(sl_4^\pm)_q$ is adequate.

Proof: Suppose $x_\lambda > x_\mu > x_\gamma \in \mathcal{B}$, $x_\gamma \leqslant z_N$ such that $\eta(x_\lambda) + \eta(x_\mu) + \eta(x_\gamma) + \eta(N) \leqslant r + 1$. We have to prove that (10) holds.

Note that $\langle x_\lambda, x_\mu \rangle = 0$ for any x_λ, $x_\mu \in \mathcal{B}$, except e_{jk}, e_{il}, so each term in Eq. (10) vanishes or e_{jk}, e_{il} appears. This means that Eq. (10) holds trivially, except in the following cases:

$$e_{ij} < e_{jk} < e_{jl}, \quad e_{ij} < e_{ik} < e_{jl}, \quad e_{ik} < e_{jk} < e_{kl}, \quad e_{ik} < e_{jl} < e_{kl}.$$

Case: $e_{ij} < e_{jk} < e_{jl}$: The left side of (10) vanishes whereas the right side is

$$e_{jk}\cdot e_{il}\cdot z_N - q^2 e_{il}\cdot e_{jk}\cdot z_N + q\langle e_{ik},e_{jl}\rangle\cdot z_N = e_{jk}\cdot e_{il}\cdot z_N - q^2 e_{il}\cdot e_{jk}\cdot z_N + (q^2-1)e_{il}\cdot e_{jk}\cdot z_N = 0,$$

because $e_{il}\cdot e_{jk}\cdot z_N = e_{jk}\cdot e_{il}\cdot z_N$, since $\eta(e_{il})+\eta(e_{jk}) < \eta(e_{ij})+\eta(e_{jk})+\eta(e_{jl})$ and supposition (5.9).

Case: $e_{ij} < e_{ik} < e_{jl}$: Let be $d = \eta(e_{ij})+\eta(e_{ik})+\eta(e_{jl})$. The left side of (10) is

$$\begin{aligned}(q^{-1}-q)e_{il}\cdot e_{jk}\cdot e_{ij}\cdot z_N &= (q^{-1}-q)(qe_{il}\cdot e_{ij}\cdot e_{jk}\cdot z_N - qe_{il}\cdot e_{ik}\cdot z_N)\\ &= (q^{-1}-q)(e_{ij}\cdot e_{il}\cdot e_{jk}\cdot z_N) - (q^{-1}-q)qe_{il}\cdot e_{ik}\cdot z_N\end{aligned}$$

$[\eta(e_{il})+\eta(e_{ij})+\eta(e_{jk}) < d$ and (9)$]$

$$= (q^{-1}-q)(e_{ij}\cdot e_{il}\cdot e_{jk}\cdot z_N - e_{il}\cdot e_{ik}\cdot z_N + qe_{ik}\cdot e_{il}\cdot z_N)$$

$[\eta(e_{ik})+\eta(e_{il}) < d]$ and this is the right side of (10).

The remaining cases are similar. □

Example VI.4: Every basic T-Lie algebra of type $(sl_n^{\pm})_q$ is adequate, $n = 5,6$.

Proof: By similar calculations as in the previous example. □

Note that the symbol $\cdot z_N$ is redundant in calculations of Example VI.3. This remark leads to the following lemma.

Let $\otimes_k L$ be the tensorial k algebra and J_r the k submodule generated by

$$x_\alpha\otimes x_\beta\otimes x_\delta - T(x_\alpha\otimes x_\beta)\otimes x_\delta - [x_\alpha,x_\beta]\otimes x_\delta, \tag{19}$$

$$x_\alpha\otimes x_\beta\otimes x_\delta - x_\alpha\otimes T(x_\beta\otimes x_\delta) - x_\alpha\otimes[x_\beta,x_\delta], \tag{20}$$

for $\eta(\alpha)+\eta(\beta)+\eta(\delta)\leqslant r$, $\forall x_\alpha,x_\beta,x_\delta\in\mathscr{B}$.

Lemma VI.2: L is adequate if

$$\begin{aligned}\langle x_\lambda,x_\mu\rangle\otimes x_\gamma - q_{\lambda\mu}[x_\mu,[x_\lambda,x_\gamma]] \equiv{}& q_{\mu\gamma}q_{\lambda\gamma}\langle x_\gamma[x_\lambda,x_\mu]\rangle + q_{\mu\gamma}q_{\lambda\gamma}x_\gamma\otimes\langle x_\lambda,x_\mu\rangle + q_{\mu\gamma}\langle x_\lambda,x_\gamma\rangle\otimes x_\mu\\ &- q_{\lambda\mu}x_\mu\otimes\langle x_\lambda,x_\gamma\rangle + q_{\mu\gamma}[x_\lambda,x_\gamma]\otimes x_\mu - q_{\lambda\mu}x_\mu\otimes[x_\lambda,x_\gamma] + x_\lambda\\ &\otimes\langle x_\mu,x_\gamma\rangle - q_{\lambda\mu}q_{\lambda\gamma}\langle x_\mu,x_\gamma\rangle\otimes x_\lambda\\ &- q_{\lambda\gamma}q_{\lambda\mu}\langle[x_\mu,x_\gamma],x_\lambda\rangle \bmod J_r,\end{aligned} \tag{21}$$

for every $x_\lambda > x_\mu > x_\gamma\in\mathscr{B}$, such that $\eta(\lambda)+\eta(\mu)+\eta(\gamma)\leqslant r+1$.

VII. POINCARÉ–BIRKHOFF–WITT THEOREM

Let us define

$$\mathscr{T}^n = \underbrace{L\otimes_k\ldots\otimes_k L}_{n\text{-times}}.$$

For $u = x_{\lambda_1}\otimes\cdots\otimes x_{\lambda_n}\in\mathscr{T}^n$, define $\delta(u) = \eta(x_{\lambda_1})+\cdots+\eta(x_{\lambda_n})$, $D(u) = \#\{(x_{\lambda_i},x_{\lambda_j})\,|\,x_{\lambda_i} > x_{\lambda_j}$ and $i<j\}$. If $u\in\otimes_k L$ and $u = \Sigma_i\xi_i u_i$ with $u_i\in\mathscr{T}^i$, $\xi_i\in k$, $\forall i$; let us put

$$D(u) = \max\{D(u_i)\,|\,\xi_i\neq 0,i\}, \tag{22}$$

$$\delta(u) = \max\{\delta(u_i)\,|\,\xi_i\neq 0,i\}. \tag{23}$$

The number $D(u)$ is called *the disorder of* u.

Denote by $\mathscr{T}_p$ the k submodule generated by $u\in\otimes_k L$, such that $\delta(u)\leqslant p$.

Theorem VII.1 (Poincaré–Birkhoff–Witt)**:** *Let L be an adequate T-Lie algebra with a basis $\mathscr{B}$. The monomials formed by finite nondecreasing sequences of elements in $\mathscr{B}$ constitute a free k basis of the universal enveloping algebra $U(L)$.*

Proof: Let $P:\otimes_k L\to U(L)$ be the canonical k morphism, $\mathscr{M}$ the k submodule generated by the monomials described in the formulation of the theorem. We have to prove that $U(L)=\mathscr{M}$. Note that

$$U(L)=\sum_{p=1}^{\infty} P(\mathscr{T}_p).$$

If $p=1$, then $\mathscr{T}_p\subseteq L$; it follows that $P(\mathscr{T}_p)\subseteq\mathscr{M}$. Suppose that $P(\mathscr{T}_r)\subseteq\mathscr{M}$. It suffices to show that $P(\mathscr{T}_{r+1})\subseteq\mathscr{M}$.

Define $\mathscr{T}_r^u$ as the k submodule of $\mathscr{T}_r$ generated by elements with disorder $\leqslant u$, and proceed by a second induction on the disorder. We have $P(\mathscr{T}_{r+1}^0)\subset\mathscr{M}$. Suppose that $v=a\otimes x\otimes y\otimes b\in\mathscr{T}_{r+1}^u$, where $x>y\in\mathscr{B}$, and $a\in\mathscr{T}^n$, $b\in\mathscr{T}^m$ monomials formed by basic elements in $\mathscr{B}$. Then

$$P(v)=P(a\otimes q_{xy}y\otimes x\otimes b)+P(a\otimes\langle x,y\rangle\otimes b)+P(a\otimes[x,y]\otimes b)\equiv P(a\otimes q_{xy}y\otimes x\otimes b)\bmod\mathscr{T}_r,$$

but $P(a\otimes q_{xy}y\otimes x\otimes b)\in\mathscr{T}_{r+1}^{u-1}\subseteq\mathscr{M}$; hence $P(v)\in\mathscr{M}$. It follows that $P(\mathscr{T}_{r+1})\subseteq\mathscr{M}$.

It remains to prove linear independence. For a given sequence $\Sigma=(x_{\lambda_1},\ldots,x_{\lambda_n})$ of nondecreasing elements of $\mathscr{B}$; define $x_\Sigma=x_{\lambda_1}\cdots x_{\lambda_n}\in U(L)$. Suppose

$$\sum_i \xi_i x_{\Sigma_i}=0,$$

where each Σ_i is a sequence nondecreasing, and $\xi_i\in k$, $\forall i$. Using the representation of $U(L)$, we get from lemma (C),

$$0=\sum_i \xi_i x_{\Sigma_i}\cdot 1=\sum_i \xi_i z_{\Sigma_i},$$

and because the linear independence of the $z_{\Sigma_i}\in\mathscr{S}(L)$, it follows that $\xi_i=0$, $\forall i$. □

Corollary VII.1: $U(sl_n^{\pm})_q$ has a basis of the Poincaré–Birkhoff–Witt type, $n=2,3,4,5$.

VIII. BRAIDS

Proposition VIII.1: In $L=(sl_n^{\pm})_q$, $n=2,3,4$ the braid equation holds:

$$T_{12}T_{23}T_{12}|_{^3L}=T_{23}T_{12}T_{23}|_{^3L},$$

where $T_{12}=T\otimes_k Id_L$, $T_{23}=Id_L\otimes_k T$.

Proof: By straightforward calculations on the basic elements (using MATHEMATICA[15]). □

Proposition VIII.2: The presymmetry S of a T-Lie algebras holds for the braid equation,

$$S_{12}S_{23}S_{12}=S_{23}S_{12}S_{23}.$$

Proof: Let x,y,z be basic elements. Then

$$S_{12}S_{23}S_{12}(x\otimes y\otimes z)=q_{x,y}q_{x,z}q_{y,z}z\otimes y\otimes x=S_{23}S_{12}S_{23}(x\otimes y\otimes z).$$

□

Remark VIII.1: The symmetry of $\widetilde{(sl_4^+)_q}$ is a braid morphism, however, we have no PBW theorem for $U\widetilde{(sl_4^+)_q}$. As a consequence, the PBW theorem is independent from the braid equation.

IX. NONSTANDARD QUANTUM DEFORMATIONS OF $GL(n)$

Definition IX.1: Let p,q be units in a commutative unitary ring k with $pq\neq 1$ and choose $\alpha(\alpha-1)/2$ discrete parameters ϵ_{ij}, $\epsilon_{ij}=\pm 1$, $1\leqslant i<j\leqslant\alpha$, $\epsilon_{ii}=1$, $\epsilon_{ji}=\epsilon_{ij}$. Let m,n be positive integers such that $m,n\leqslant\alpha$.

The k module $L_{p,q,\epsilon}(n,m,k)$ is then defined to be the free k module with basis

$$\text{(a)}\quad \begin{matrix} Z_1^1 & Z_1^2 & \dots & Z_1^n \\ \vdots & \vdots & & \vdots \\ Z_m^1 & Z_m^2 & \dots & Z_m^n \end{matrix} \qquad \text{(b)}\quad \begin{matrix} \circ_i^j & \circ_i^{j+u} \\ \circ_{i+u}^j & \circ_{i+u}^{j+u} \end{matrix}$$

FIG. 6. (a) The basic T-Lie algebra $L_{p,q,\epsilon'}(\lambda,\mu,k)$. (b) Diagonal relationship.

$$\mathscr{B}=\{Z_i^j \mid 1\leqslant i\leqslant n, 1\leqslant j\leqslant m\}.$$

Now define an order on $\mathscr{B}$ and morphisms S,T, $\langle,\rangle$,[,], copying the structure of $L(n,k)$ in Example III.6.

Proposition IX.2: $L_{p,q,\epsilon}(n,m,k)$ has a structure of a basic T-Lie algebra.

In a similar way to the algebras of type $(sl_n^+)_q$ (see Sec. VI), we can define *algebras of type* $L_{p,q,\epsilon}(n,m,k)$.

Lemma IX.1: Every algebra of type $L_{p,q,\epsilon}(\lambda,\mu,k)$ is an adequate basic T-Lie algebra, where $\lambda,\mu\in\{2,3\}$.

Lemma IX.2: If $Z_u^v>Z_i^j>Z_a^b$, then there exists L being a T-Lie subalgebra of $L_{p,q,\epsilon}(n,m,k)$ and numbers $\lambda,\mu\in\{2,3\}$, such that L is of type $L_{p,q,\epsilon'}(\lambda,\mu,k)$ and $\{Z_u^v,Z_i^j,Z_a^b\}\subset L$.

Proof: Let us put the basic elements in a matrix array [Fig. 6(a)].

Note that for positive integers u,v the elements appearing in the pseudobracket definition are in a diagonal relationship [Fig. 6(b)], and they form a free basis of a T-Lie algebra of type $L_{p,q,\epsilon'}(2,2,k)$, where $\epsilon'=\{1,\epsilon_{i,i+u},\epsilon_{j,j+u}\}$.

For $Z_u^v>Z_i^j>Z_a^b$ there are several cases. The cases given by Figs. 7(a), 7(b), and 7(c), or they form a triangle that can be fitted, with vertices on the border, inside the rectangle of Fig. 7(d).

In the case given by Fig. 7(a) we may complete each triangle to a square and obtain $L_{p,q,\epsilon_0}(2,2,k)$. In the case given by Fig. 7(b), each triangle can be completed to a rectangle in the form of Fig. 7(e), and we get $L_{p,q,\epsilon_1}(3,2,k)$. Similarly, in the case given by Fig. 7(c), we get $L_{p,q,\epsilon_2}(2,3,k)$. Finally, in the case given by Fig. 7(d), we obtain $L_{p,q,\epsilon_3}(3,3,k)$. □

Theorem IX.3: *$L_{p,q,\epsilon}(n,m,k)$ is an adequate basic T-Lie algebra.*

Corollary IX.3: The monomials formed by nondecreasing finite sequences of elements in

$$\mathscr{B}=\{Z_i^j \mid 1\leqslant i\leqslant n, 1\leqslant j\leqslant m\}$$

constitute a free basis of the k module $M_{p,q,\epsilon}(n,m,k)=UL_{p,q,\epsilon}(n,m,k)$.

X. $U(sl_{n+1}^+)_q$. THE GENERAL CASE

Lemma X.1: Let e_{ab},e_{uv},e_{ij} be basic elements in $(sl_{n+1}^+)_q$ and [,] the usual bracket in sl_{n+1}.

(1) $e_{ab}<e_{uv}$ if and only if $a+b<i+j$ or $a+b=i+j$ and $b<j$.

(2) If $S(e_{ab}\otimes e_{uv})=q^{c_{ab,uv}}e_{uv}\otimes e_{ab}$ and $e_{ab}<e_{uv}$, then

$$c_{ab,uv}=-\delta_{v,a}+\delta_{v,b}+\delta_{u,a}-\delta_{u,b}.$$

(3) If $e_{ab}<e_{uv}<e_{ij}$, then

$$q^{c_{uv,ab}}[e_{uv},[e_{ab},e_{ij}]]=[e_{uv},[e_{ab},e_{ij}]]_q.$$

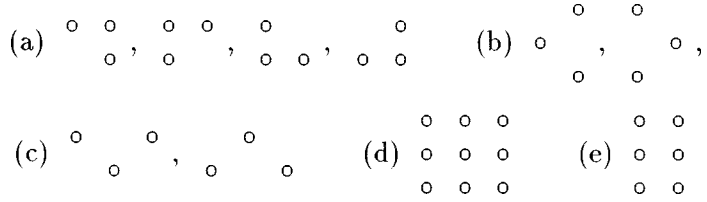

FIG. 7. Some cases in $L_{p,q,\epsilon}(n,m,k)$.

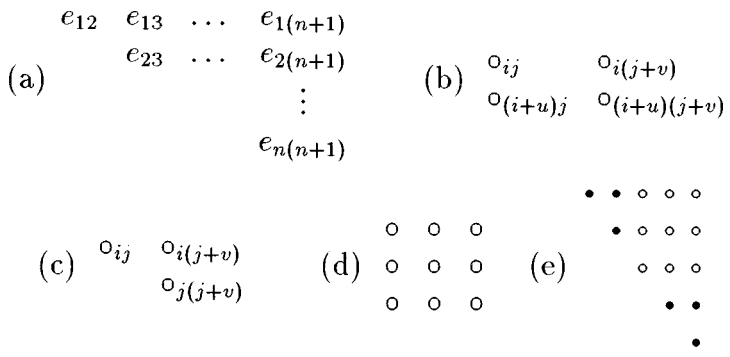

FIG. 8. Some cases in $(sl_{n+1}^{+})_q$.

Proof: (1) By the order definition.

(2) It follows from the formula $[e_{ij},e_{kl}]=\delta_{jk}e_{il}-\delta_{li}e_{kj}$ in the classical Lie algebra sl_n.

(3) Since $e_{ab}<e_{ij}$, we can suppose that $b=i$. We have to prove $q^{c_{uv,ab}}[e_{uv},e_{aj}]=[e_{uv},e_{aj}]_q$. There are two cases: (a) $e_{uv}<e_{aj}$; (b) $e_{uv}>e_{aj}$.

(3a) If $[e_{uv},e_{aj}]\neq 0$ then $v=a$ and $u<v=a<b$, it follows that $e_{uv}<e_{ab}$, since $u+v<a+b$ (a contradiction). Therefore $[e_{uv},e_{aj}]_q=[e_{uv},e_{aj}]=0$.

(3b) We have to prove

$$q_{uv,ab}[e_{aj},e_{uv}]=q_{uv,aj}[e_{aj},e_{uv}].$$

Both sides are zero, because if they are not, then $j=u$ and $i<j=u<v$; these imply $i+j<v+u$, and then $e_{ij}<e_{uv}$. Again, we have a contradiction. □

Theorem X.2: $(sl_{n+1}^{+})_q$ *is an adequate basic T-Lie algebra.*

Proof: Let $\mathscr{B}$ be the canonical basis of sl_{n+1}, and write the basic elements of $\mathscr{B}$ in the form e_{ij}. Now, we put this basic element in an upper triangular array [Fig. 8(a)]. Note that if $\langle e_{ij},e_{(i+u)(j+v)}\rangle\neq 0$, then the elements appearing in the pseudobracket definition are in a diagonal relationship [Fig. 8(b)], and if $[e_{ij},e_{(i+u)(j+v)}]_q\neq 0$ then $j=i+u$, and we get Fig. 8(c).

So, if we suppose that $e_{ij}>e_{uv}>e_{ab}$, then the elements appearing in the formulation of Lemma VI.2 (brackets and pseudobrackets) can be fitted inside a square of the form of Fig. 8(d), and such a square can be extended to an upper triangle [Fig. 8(e)], but this triangle gives a strictly graded algebra of type $(sl_6^{+})_q$. Since these algebras satisfy the condition of Lemma VI.2, in particular, the elements $e_{ij}>e_{uv}>e_{ab}$ satisfy this condition. Besides,

$$[,]_q((Id\otimes[,]_q)S_{12}S_{23}-([,]_q\otimes Id)S_{23}S_{12}+(Id\otimes[,]_q)S_{23}S_{12})(e_{ij}\otimes e_{uv}\otimes e_{ab})$$
$$=q^{c_{uv,ab}+c_{ij,ab}+c_{ij,uv}}([[e_{ab},e_{uv}],e_{ij}]-[e_{ab},[e_{uv},e_{ij}]]+[e_{uv},[e_{ab},e_{ij}]])=0,$$

since Lemma X.1 and the Jacobi identity in sl_n^{+}.

We conclude that $(sl_{n+1}^{+})_q$ is an adequate basic T-Lie algebra. □

Lemma X.3: Suppose that $e_{ij}<e_{ab}\in U_q^{+}(sl_{n+1})$. Then the following equations are satisfied in $U_q^{+}(sl_{n+1})$:

$$[e_{ij},e_{ab}]=\begin{cases} e_{ij}e_{ab}-qe_{ab}e_{ij}, & \text{if } i=a \quad \text{or} \quad j=b,\\ e_{ij}e_{ab}-e_{ab}e_{ij}-\langle e_{ij},e_{ab}\rangle, & \text{if } i\neq a, j\neq b, \quad \text{and} \quad j\neq a,\\ e_{ij}e_{ab}-q^{-1}e_{ab}e_{ij}, & \text{if } j=a. \end{cases} \tag{24}$$

Proof: By induction on n. For the cases $n=1,2,3,4,5$, the Eqs. (24) can be verified by straightforward calculations. So we may suppose $n>5$. Let us consider Fig. 8(a). Such a diagram can be thought of as formed by two overlapping triangles. The first one, a triangle T_1 with vertices $e_{12},e_{1n},e_{(n-1)n}$, and the second one, a triangle T_2 with vertices $e_{23},e_{2(n+1)},e_{n(n+1)}$.

The elements in T_i generate a k subalgebra isomorphic to $U_q^{+}(sl_n)$, $i=1,2$. Then, if e_{ij} and e_{ab} are both in T_1 or T_2, Eqs. (24) hold. As a consequence, we may suppose $i=1$ and $b=n+1$, and put $j\neq n+1$ and $a\neq 1$.

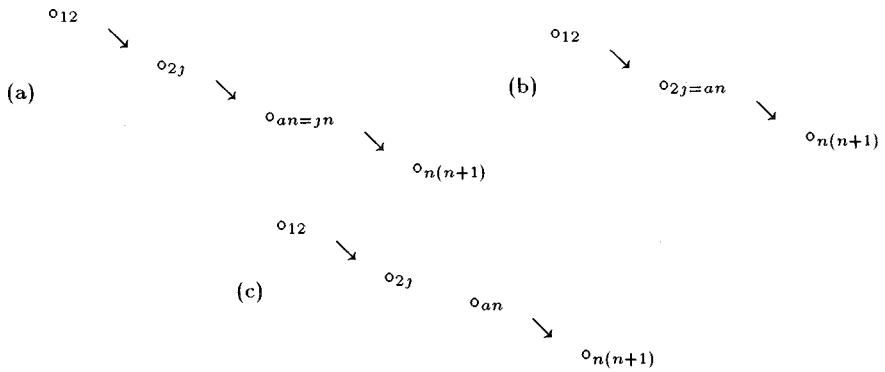

FIG. 9. (a) First case. (b) Second case. (c) Third case.

At Fig. 8(a), join the node rs with the node uv if $[e_{rs},e_{uv}]_q\neq 0$. We have several cases given by Fig. 9 [in the first and third cases, since e_{12},e_{2j},e_{jn} are in T_1 and the induction hypothesis, there is no arrow between 12 and an, whereas there is no arrow between $2j$ and $n(n+1)$ because $e_{2j},e_{an},e_{n(n+1)}$ are in T_2].

At the first case we get a graph of type A_4; then $e_{1j}=[e_{12},e_{2j}]_q$, $e_{a(n+1)}=[e_{an},e_{n(n+1)}]_q$ are in a subalgebra isomorphic to $U_q^+(sl_5)$; it follows that

$$[e_{1j},e_{j(n+1)}]_q=e_{1j}e_{j(n+1)}-q^{-1}e_{j(n+1)}e_{ij}.$$

In the second case we get a graph of type A_3; then $e_{1j}=[e_{12},e_{2n}]_q$, $e_{2(n+1)}=[e_{2n},e_{n(n+1)}]_q$ are in a subalgebra isomorphic to $U_q^+(sl_4)$, besides

$$[e_{1j},e_{a(n+1)}]_q=e_{1n}e_{2(n+1)}-e_{2(n+1)}e_{1n}-(q-q^{-1})e_{1(n+1)}e_{2n}.$$

In the third case we may insert the node ja in order to obtain

$$^\circ 12\longrightarrow{}^\circ 2j\longrightarrow{}^\circ ja\longrightarrow{}^\circ an\longrightarrow{}^\circ n(n+1);$$

this graph is of type A_5; then $e_{1j}=[e_{12},e_{2j}]_q$, $e_{a(n+1)}=[e_{an},e_{n(n+1)}]_q$ are in a subalgebra isomorphic to $U_q(sl_6^+)$ and

$$0=[e_{1j},e_{a(n+1)}]_q=e_{1j}e_{a(n+1)}-e_{a(n+1)}e_{1j}.$$

Now, only the cases $e_{1j}=e_{1(n+1)}$, $e_{a(n+1)}=e_{1(n+1)}$ remain. Suppose that $e_{1j}=e_{1(n+1)}$. Since $e_{12}e_{a(n+1)}=e_{a(n+1)}$, $e_{2(n+1)}e_{a(n+1)}=qe_{a(n+1)}e_{2(n+1)}$, and $e_{1(n+1)}=e_{12}e_{2(n+1)}-q^{-1}e_{2(n+1)}e_{12}$; it follows that

$$e_{1(n+1)}e_{a(n+1)}=qe_{a(n+1)}e_{1(n+1)}.$$

In a similar way, if $e_{a(n+1)}=e_{1(n+1)}$, we get

$$e_{1j}e_{a(n+1)}=qe_{a(n+1)}e_{1j}.$$

□

Theorem X.4: *There exists an isomorphism,*

$$U_q^+(sl_{n+1})\simeq U(sl_{n+1}^+)_q,$$

of k algebras.

Proof: Let us put $c_{ab,cd}=c_{uv}$, where $x_u=e_{ab}$, $x_v=e_{cd}$, and $x_u<x_v$, $1\leqslant a,b,c,d\leqslant n+1$. From

$$[e_{ij},e_{kl}]=\delta_{jk}e_{il}-\delta_{li}e_{kj},$$

it follows, if $e_{ab}<e_{cd}$,

$$c_{ab,cd}=\begin{cases}1, & \text{if } a=c \text{ or } b=d.\\ 0, & \text{if } a\neq c \text{ and } b\neq d\\ -1, & \text{if } b=a.\end{cases}$$

Now use Lemma X.3 in order to obtain the following equations in $U_q^+(sl_{n+1})$:

$$e_{ab}e_{cd}-q^{c_{ab,cd}}e_{cd}e_{ab}=[e_{ab},e_{cd}]_q+\langle e_{ab},e_{cd}\rangle,$$

for all $1\leqslant a,b,c,d\leqslant n+1$.

We conclude that $U_q^+(sl_{n+1})\simeq U(sl_{n+1}^+)_q$. □

Corollary X.1:

(1) *The monomials formed by nondecreasing finite sequences of elements in*

$$\mathscr{B}=\{e_{ij}|1\leqslant i\leqslant j\leqslant m\},$$

constitute a free basis of the k module $U_q^+(sl_{n+1})$, *where* $m=n(n+1)/2$.

(2) *We have*

$$(sl_{n+1}^+)_q|_{t=0}=sl_{n+1}^+,$$

and $(sl_{n+1}^+)_q$ *is a deformation of* sl_{n+1}^+ *in the category of T-Lie algebras.*

ACKNOWLEDGMENTS

I would like to thank M. Durdević and R. Bautista for their encouragement, C. Ramirez for giving me access to his papers collection, and A. Sudbery for drawing my attention to his own work about quantum algebras. I also thank J. García-Fernández, P. García, and L. Montgomery for the computer freeware.

The computer programs used in this work were developed by the author in collaboration with Yaocíhuatl E. Arroyo. This work has been partially supported by the Investigation Project IN106897 of DGAPA/UNAM.

[1] H. Yamane, ''A Poincaré–Birkhoff–Witt theorem for quantized universal enveloping algebras of type A_n,'' Publ. RIMS Kyoto Univ. **25**, 503–520 (1989).
[2] D. Gurevich, ''The Yang–Baxter equation and a generalization of formal Lie theory,'' Sov. Math. Dokl. **33**, 758–762 (1986).
[3] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Graduate Texts in Mathematics 9 (Springer-Verlag, New York, 1994).
[4] M. Scheunert, *The Theory of Lie Superalgebras*, Lecture Notes in Mathematics Vol. 716 (Springer-Verlag, Berlin, 1979).
[5] M. Scheunert, ''Generalized Lie algebras,'' J. Math. Phys. **20**, 712–720 (1979).
[6] S. Majid, ''Quantum and braided Lie algebras,'' J. Geom. Phys. **13**, 307–356 (1994).
[7] G. W. Delius and M. D. Gould, ''Quantum Lie algebras, their existence, uniqueness and q-antisymmetry,'' preprint q-alg/9605025.
[8] D. Gurevich and V. Rubstov, ''Quantization of Poisson pencils and generalized Lie algebras,'' preprint q-alg/9591997.
[9] V. Lyubashenko and A. Sudbery, ''Generalized Lie algebras of type A_n,'' preprint q-alg/9510004.
[10] G. W. Delius, ''Introduction to quantum Lie algebras,'' preprint q-alg/9605026.
[11] G. W. Delius and A. Huffmann, ''On quantum Lie algebras and quantum root systems,'' preprint q-alg/9506017.
[12] E. E. Demidov, Yu. I. Manin, E. E. Mukhin, D. V. Zhdanovich, ''Non-standard quantum deformations of $GL(n)$ and constant solutions of the Yang–Baxter equation,'' Prog. Theor. Phys. Suppl. **102**, 203–218 (1990).
[13] C. M. Ringel, ''PBW—Bases of quantum groups,'' preprint 93-062, sfb343, Diskrete Strukturen der Mathematik, U. Bielefeld, 1993.
[14] A. Sudbery, ''Matrix-element bialgebras determined by quadratic coordinate algebras,'' J. Alg. **158**, 375–399 (1993).
[15] S. Wolfram, MATHEMATICA. *A System for Doing Mathematics by Computer*, 2nd ed. (Addison-Wesley, Reading, MA, 1993).

Principal vertex operator representations for toroidal Lie algebras

Yuly Billig[a)]
Department of Mathematics and Statistics, University of New Brunswick, Fredericton, New Brunswick E3B 5A3, Canada

(Received 30 June 1997; accepted for publication 30 March 1998)

We introduce the principal vertex operator representations for the toroidal Lie algebras generalizing the construction for the affine Kac–Moody algebras. We also represent the derivations of the toroidal algebras and introduce analogs of the Sugawara operators. © *1998 American Institute of Physics.* [S0022-2488(98)02407-4]

I. INTRODUCTION

Vertex operators discovered by physicists in string theory have turned out to be important objects in mathematics. One can use vertex operators to construct various realizations of the irreducible highest weight representations for affine Kac–Moody algebras. Two of these, the principal and homogeneous realizations, are of particular interest. The principal vertex operator construction for the affine algebra $A_1^{(1)}$ allows one to construct soliton solutions of the Korteweg–de Vries (KdV) hierarchy of partial differential equations. On the other hand, the homogeneous realization is linked to the fundamental nonlinear Schrödinger equation.[1]

Eswara Rao and Moody[2] studied the homogeneous vertex operator construction for toroidal Lie algebras. The present paper is devoted to the principal realization. Here we construct the principal vertex operator representation for the toroidal Lie algebra $\hat{\mathfrak{g}}$ which is a universal central extension of $\tilde{\mathfrak{g}}=\dot{\mathfrak{g}}\otimes\mathbb{C}[t_0^{\pm},\ldots,t_n^{\pm}]$, in which $\dot{\mathfrak{g}}$ is a simply laced simple finite-dimensional Lie algebra over $\mathbb{C}$. This generalizes the principal vertex operator realization of the basic representations of affine Lie algebras.[3,4]

We add the Lie algebra $\mathscr{D}$ of vector fields on a torus to the toroidal Lie algebra $\hat{\mathfrak{g}}$ to form a larger algebra $\mathfrak{g}$. This is necessary in order to have a sufficiently large principal Heisenberg subalgebra. To construct a representation of $\mathfrak{g}$ we consider the standard module F for the principal Heisenberg subalgebra and then extend the action to all of $\mathfrak{g}$ by means of the vertex operators. The module F is irreducible as a module over $\mathfrak{g}$, and it is reducible over $\hat{\mathfrak{g}}$. It will be interesting to see how this representation fits in the framework of the Verma modules over the toroidal Lie algebras.[5]

In a subsequent article[6] we construct an extension of the KdV hierarchy using the representations introduced in the present paper.

The vertex operator representations for the toroidal Lie algebras constructed here may also be useful for quantum field theories in space–time of more than two dimensions.[7,8]

This paper is organized as follows. In Sec. II we recall the construction of the toroidal Lie algebra and in Sec. III we sketch the principal vertex operator realization of the basic representations of affine Lie algebras. In Sec. IV we present the principal Heisenberg subalgebra of $\mathfrak{g}$ and define its standard module. We also introduce there vertex operators that will allow us to extend this representation to $\mathfrak{g}$. At the end of Sec. IV we state the main theorem of the paper. The proof of the main theorem occupies Sec. V. In Sec. VI we construct the analogs of the Sugawara operators.

II. DEFINITIONS AND NOTATIONS

Let $\dot{\mathfrak{g}}$ be a simple finite-dimensional complex Lie algebra of type A_ℓ, D_ℓ, or E_ℓ (i.e., simply laced). The Killing form $(\cdot|\cdot)$ is nondegenerate on the Cartan subalgebra $\dot{\mathfrak{h}}$ of $\dot{\mathfrak{g}}$ and induces the map $\nu:\dot{\mathfrak{h}}\to\dot{\mathfrak{h}}^*$. Let $\dot{\Delta}$ be the root system of $\dot{\mathfrak{g}}$. Since $\dot{\mathfrak{g}}$ is simply laced then $(\alpha|\alpha)=2$ for all nonzero roots $\alpha\in\dot{\Delta}$.

a)Electronic mail: ybillig@unb.ca

 © 1998 American Institute of Physics

To construct a toroidal algebra let us consider a tensor product of $\dot{\mathfrak{g}}$ with the algebra of Laurent polynomials in $n+1$ variable:

$$\tilde{\mathfrak{g}}=\dot{\mathfrak{g}}\otimes\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}].$$

We define the toroidal Lie algebra corresponding to $\dot{\mathfrak{g}}$ as the universal central extension of $\tilde{\mathfrak{g}}$. The explicit construction of this extension[9,10] can be given as follows. Let $\dot{\mathscr{K}}$ be an $(n+1)$-dimensional space with the basis $\{K_0,K_1,\ldots,K_n\}$. Consider a derivation $d_p=t_p(d/dt_p)$ of the algebra $\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}]$ which can be naturally extended on the tensor product

$$\widetilde{\mathscr{K}}=\dot{\mathscr{K}}\otimes\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}],$$

The space of the universal central extension of $\tilde{\mathfrak{g}}$ is

$$\mathscr{K}=\widetilde{\mathscr{K}}/d\widetilde{\mathscr{K}},$$

where

$$d\widetilde{\mathscr{K}}=\left\{\sum_{p=0}^{n}K_p\otimes d_p(f)\,\middle|\,f\in\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}]\right\}\subset\widetilde{\mathscr{K}}.$$

We will denote the image of $K_p\otimes t_0^{r_0}\mathbf{t}^{\mathbf{r}}$ in $\mathscr{K}$ by $t_0^{r_0}\mathbf{t}^{\mathbf{r}}K_p$, where $\mathbf{r}=(r_1,\ldots,r_n)$ and $\mathbf{t}^{r}=t_1^{r_1}\ldots t_n^{r_n}$. Note that $\mathscr{K}$ has the defining relations

$$r_0t_0^{r_0}\mathbf{t}^{\mathbf{r}}K_0+r_1t_0^{r_0}\mathbf{t}^{\mathbf{r}}K_1+\cdots+r_nt_0^{r_0}\mathbf{t}^{\mathbf{r}}K_n=0.$$

The toroidal algebra is the Lie algebra

$$\hat{\mathfrak{g}}=\dot{\mathfrak{g}}\otimes\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}]\oplus\mathscr{K}$$

with the bracket

$$[g_1\otimes f_1(t_0\ldots t_n),g_2\otimes f_2(t_0\ldots t_n)]=[g_1,g_2]\otimes(f_1f_2)+(g_1|g_2)\sum_{p=0}^{n}d_p(f_1)f_2K_p \tag{2.1}$$

and

$$[\hat{\mathfrak{g}},\mathscr{K}]=0. \tag{2.2}$$

Finally, we shall add certain derivations to $\hat{\mathfrak{g}}$. Specifically, let $\mathscr{D}$ be the Lie algebra of derivations of $\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}]$:

$$\mathscr{D}=\left\{\sum_{p=0}^{n}f_p(t_0,\ldots,t_n)d_p\,\middle|\,f_0,\ldots,f_n\in\mathbb{C}[t_0^{\pm},t_1^{\pm},\ldots,t_n^{\pm}]\right\}.$$

It is a general fact, that a derivation acting on a Lie algebra can be lifted in a unique way to a derivation of the universal central extension of this Lie algebra.[11] Thus the natural action of $\mathscr{D}$ on $\tilde{\mathfrak{g}}$,

$$f_1(t_0,\ldots,t_n)d_p(g\otimes f_2(t_0,\ldots,t_n))=g\otimes f_1d_p(f_2), \tag{2.3}$$

has a unique extension to $\hat{\mathfrak{g}}$. We shall denote the lift of $f(t_0,\ldots,t_n)d_p$ by $f(t_0,\ldots,t_n)D_p$. Its action on the subspace $\tilde{\mathfrak{g}}$ is unchanged, while the action on $\mathscr{K}$ is given by[2]

$$f_1D_a(f_2K_b)=f_1d_a(f_2)K_b+\delta_{ab}\sum_{p=0}^{n}f_2d_p(f_1)K_p. \tag{2.4}$$

Consider a subalgebra $\mathscr{D}_+$ in $\mathscr{D}$:

$$\mathscr{D}_+ = \left\{ \sum_{p=1}^{n} f_p(t_0,\dots,t_n) D_p \,\middle|\, f_1,\dots,f_n \in \mathbb{C}[t_0^{\pm},t_1^{\pm},\dots,t_n^{\pm}] \right\}.$$

We will be working with the algebra $\mathfrak{g}$ which is a deformation of the semidirect product of $\hat{\mathfrak{g}}$ with $\mathscr{D}_+$:

$$\mathfrak{g} = \dot{\mathfrak{g}} \otimes \mathbb{C}[t_0^{\pm},t_1^{\pm},\dots,t_n^{\pm}] \oplus \mathscr{K} \oplus \mathscr{D}_+ .$$

Here multiplication in $\hat{\mathfrak{g}}$ is given by (2.1) and (2.2), the action of $\mathscr{D}_+$ on $\tilde{\mathfrak{g}}$ is given by (2.3) and on $\mathscr{K}$ by (2.4). The Lie bracket in $\mathscr{D}_+$ is the following:

$$[t_0^{r_0}\mathbf{t}^{\mathbf{r}} D_a, t_0^{m_0}\mathbf{t}^{\mathbf{m}} D_b] = m_a t_0^{r_0+m_0}\mathbf{t}^{\mathbf{r}+\mathbf{m}} D_b - r_b t_0^{r_0+m_0}\mathbf{t}^{\mathbf{r}+\mathbf{m}} D_a - m_a r_b \left\{ \sum_{p=0}^{n} r_p t_0^{r_0+m_0}\mathbf{t}^{\mathbf{r}+\mathbf{m}} K_p \right\}.$$

The last term in this formula is a 2-cocycle with values in a nontrivial $\mathscr{D}$-module $\mathscr{K}$. This abelian extension of the Lie algebra of vector fields on a torus is a generalization of the Virasoro algebra and was introduced by Rao and Moody.[2]

III. PRINCIPAL REALIZATION OF THE BASIC REPRESENTATION OF AFFINE LIE ALGEBRA

When $n=0$ the algebra $\mathfrak{g}$ constructed above yields the derived affine Kac–Moody algebra. We set $n=0$ throughout this section. In this case $\mathscr{K}$ is one-dimensional and is spanned by K_0, while $\mathscr{D}_+$ is trivial, so

$$\mathfrak{g} = \dot{\mathfrak{g}} \otimes \mathbb{C}[t_0,t_0^{-1}] \oplus \mathbb{C}K_0 .$$

Lepowsky and Wilson[3] constructed the principal realization of the basic representation of affine Kac–Moody algebra $A_1^{(1)}$, while the construction for a general affine algebra was given by Kac *et al.*[4] Let us recall this construction (see also Ref. 1 for details). Its starting point is the principal Heisenberg subalgebra in $\mathfrak{g}$.

Let $\{\alpha_1,\dots,\alpha_\ell\}$ be simple roots and θ be the highest positive root in $\dot{\Delta}$. We define the height for a root $\beta = \Sigma_{i=1}^{\ell} k_i\alpha_i$ as

$$ht(\beta) = \sum_{i=1}^{\ell} k_i .$$

Note that $ht(\theta) = h-1$, where h is a Coxeter number of $\dot{\mathfrak{g}}$.[12]

Consider the principle gradation of $\tilde{\mathfrak{g}}$ defined by

$$\deg(\dot{\mathfrak{g}}^{\alpha_1}\otimes 1) = \dots = \deg(\dot{\mathfrak{g}}^{\alpha_\ell}\otimes 1) = \deg(\dot{\mathfrak{g}}^{-\theta}\otimes t_0) = 1,$$

$$\deg(\dot{\mathfrak{g}}^{-\alpha_1}\otimes 1) = \dots = \deg(\dot{\mathfrak{g}}^{-\alpha_\ell}\otimes 1) = \deg(\dot{\mathfrak{g}}^{\theta}\otimes t_0^{-1}) = -1.$$

We choose nonzero root vectors $e_0 \in \dot{\mathfrak{g}}^{-\theta}\otimes t_0, e_1 \in \dot{\mathfrak{g}}^{\alpha_1}\otimes 1, \dots, e_\ell \in \dot{\mathfrak{g}}^{\alpha_\ell}$ and form the element $\overline{e} = \Sigma_{i=0}^{\ell} e_i \in \tilde{\mathfrak{g}}$. Since $\overline{e}$ is of degree 1 then its centralizer $\tilde{\mathfrak{s}}$ in $\tilde{g}$ is homogeneous with respect to the gradation.

Next, consider a projection

$$\pi: \quad \dot{\mathfrak{g}}\otimes\mathbb{C}[t_0,t_0^{-1}] \to \dot{\mathfrak{g}} \quad \text{by} \quad t_0 \mapsto 1.$$

Note that $\deg(\dot{\mathfrak{g}}^{-\theta}\otimes t_0) = 1$ and $\deg(\dot{\mathfrak{g}}^{-\theta}\otimes 1) = ht(-\theta) = -h+1$, while $\pi(\dot{\mathfrak{g}}^{-\theta}\otimes t_0) = \pi(\dot{\mathfrak{g}}^{-\theta}\otimes 1)$, and thus the principal $\mathbb{Z}$-gradation of $\tilde{\mathfrak{g}}$ induces the $\mathbb{Z}_h$-gradation of $\dot{\mathfrak{g}}$:

$$\dot{\mathfrak{g}} = \sum_{j\in\mathbb{Z}_h} \dot{\mathfrak{g}}_j .$$

The element $\pi(\overline{e})$ is regular in $\dot{\mathfrak{g}}$, hence its centralizer $\dot{\mathfrak{s}}=\pi(\tilde{\mathfrak{s}})$ in $\dot{\mathfrak{g}}$ is a Cartan subalgebra, which implies that both $\dot{\mathfrak{s}}$ and $\tilde{\mathfrak{s}}$ are abelian.

The preimage $\mathfrak{s}=\tilde{\mathfrak{s}}\oplus\mathbb{C}K_0$ of $\tilde{\mathfrak{s}}$ in $\mathfrak{g}$ is an infinite-dimensional Heisenberg algebra which is called the principal Heisenberg subalgebra of $\mathfrak{g}$. This subalgebra is the cornerstone of the construction of the vertex operator representation of $\mathfrak{g}$. The scheme of this construction is the following: one starts with the standard irreducible representation of $\mathfrak{s}$ and then magically one can lift this representation to the whole of $\mathfrak{g}$. Moreover the action of $\mathfrak{g}$ is prescribed and given by vertex operators. Finally one has only to check in one way or another that this construction works.

The irreducible representation of $\mathfrak{g}$ constructed in this way is of the highest weight Λ_0, where the linear functional Λ_0 is given (we consider the simply laced case) by

$$\Lambda_0(\nu^{-1}(\alpha_i)\otimes t_0^0)=0, \quad \Lambda_0(K_0)=1.$$

This highest weight representation is called the basic representation of $\mathfrak{g}$.

The Cartan subalgebra $\dot{\mathfrak{s}}$ is homogeneous with respect to the principle $\mathbb{Z}_h$-gradation:

$$\dot{\mathfrak{s}}=\sum_{j\in\mathbb{Z}_h}\dot{\mathfrak{s}}_j\,.$$

Since $(\dot{\mathfrak{g}}_i|\dot{\mathfrak{g}}_j)=0$ unless $i+j=0(\mathrm{mod}\ h)$ and $(\cdot|\cdot)$ is nondegenerate on $\dot{\mathfrak{s}}$ then one can choose a basis in $\dot{\mathfrak{s}}$: $\{T_1,T_2,\ldots,T_\ell\}$ such that $T_i\in\dot{\mathfrak{g}}_{m_i}$, where $1\leqslant m_1\leqslant m_2\leqslant\cdots\leqslant m_\ell<h$ and

$$(T_i|T_{\ell+1-j})=h\,\delta_{ij}\,.$$

The numbers $\{m_1,\ldots,m_\ell\}$ are called the exponents of $\dot{\mathfrak{g}}$.[12] Note that $\dot{\mathfrak{s}}\cap\dot{\mathfrak{h}}=(0)$, and therefore zero is not an exponent. Also, $m_{\ell+1-i}=h-m_i$.

It turns out that it is more convenient to work with a slightly different realization of $\mathfrak{g}$ based on the $\mathbb{Z}_h$-gradation of $\dot{\mathfrak{g}}$. Consider

$$\mathfrak{g}_s=\sum_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\oplus\mathbb{C}K_0\,,$$

with the Lie bracket

$$[g_1\otimes s^i,g_2\otimes s^j]=[g_1,g_2]\otimes s^{i+j}+\frac{i}{h}\,(g_1|g_2)\,\delta_{i,-j}K_0\,,$$

$$[K_0,\mathfrak{g}_s]=0.$$

The proof of the following lemma is straightforward, so we omit it.

Lemma 1: The Lie algebras $\mathfrak{g}$ *and* $\mathfrak{g}_s$ *are isomorphic and the isomorphism*

$$\psi:\quad \dot{\mathfrak{g}}\otimes\mathbb{C}[t_0,t_0^{-1}]\oplus\mathbb{C}K_0\to\sum_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\oplus\mathbb{C}K_0$$

is given by

$$\psi(e_\alpha\otimes t^i)=e_\alpha\otimes s^{ht(\alpha)+ih},$$

$$\psi(\nu^{-1}(\alpha)\otimes t^i)=\nu^{-1}(\alpha)\otimes s^{ih}+\delta_{i,0}\,\frac{ht(\alpha)}{h}\,K_0\,,$$

$$\psi(K_0)=K_0.$$

We shall identify $\mathfrak{g}$ and $\mathfrak{g}_s$ using this isomorphism.

In order to describe a basis in the principal Heisenberg subalgebra $\psi(\mathfrak{s})$ we need to introduce the sequence $\{b_i\}_{i\in\mathbb{Z}}$ such that $b_{i+j\ell}=m_i+jh$ for $j\in\mathbb{Z}$, $i=1,...,\ell$. Then $\psi(\mathfrak{s})$ is spanned by $\{T_i\otimes s^{b_i},K_0\}$, $i\in\mathbb{Z}$. Note that $b_{1-i}=-b_i$.

Let $\dot{\Delta}_s$ be the root system of $\dot{\mathfrak{g}}$ with respect to the Cartan subalgebra $\dot{\mathfrak{s}}$. For $\alpha\in\dot{\Delta}_s$ choose a root element

$$A^\alpha=\sum_{j\in\mathbb{Z}_h} A_j^\alpha.$$

Since $\pi(\overline{e})$ is a regular element of the Cartan subalgebra then $[\pi(\overline{e}),A^\alpha]=\lambda^\alpha A^\alpha$ with $\lambda^\alpha\neq 0$. Hence $A_j^\alpha=(ad\,\pi(\overline{e})/\lambda^\alpha)^j A_0^\alpha$ and all the components A_j^α are nonzero. We let the indices i,j in T_i and A_j^α run over $\mathbb{Z}$ by setting $T_{i_1}=T_{i_2}$ if $i_1\equiv i_2(\mathrm{mod}\,\ell)$ and $A_{j_1}^\alpha=A_{j_2}^\alpha$ if $j_1\equiv j_2(\mathrm{mod}\,h)$.

Define constants $\lambda_i^\alpha=\alpha(T_i)$. Then $[T_i,A^\alpha]=\lambda_i^\alpha A^\alpha$ and $[T_i,A_j^\alpha]=\lambda_i^\alpha A_{j+m_i}^\alpha$.

Consider an automorphism σ of $\dot{\mathfrak{g}}$ such that

$$\sigma=\zeta^j\mathrm{Id}\quad\text{on}\quad\dot{\mathfrak{g}}_j,$$

where $\zeta\in\mathbb{C}$ is a primitive h-root of 1. For a root element $A^\alpha=\Sigma_{j=1}^h A_j^\alpha$ we have

$$[T_i,\sigma(A^\alpha)]=\sigma([\sigma^{-1}(T_i),A^\alpha])=\sigma([\zeta^{-m_i}T_i,A^\alpha])=\zeta^{-m_i}\lambda_i^\alpha\sigma(A^\alpha).$$

Thus $\sigma(A^\alpha)=\Sigma_{j=1}^h \zeta^j A_j^\alpha$ is also a root element and σ induces an automorphism of $\dot{\Delta}_s$. Since all A_j^α are nonzero then the length of each orbit of σ in $\dot{\Delta}_s$ is h.

We may choose ℓ roots $\beta_1,\ldots,\beta_\ell\in\dot{\Delta}_s$ such that $A_0^{\beta_1},\ldots,A_0^{\beta_\ell}$ span $\dot{\mathfrak{g}}_0$. Note that if two root elements belong to the same orbit under the action of σ then their zero components are proportional. Therefore every orbit of σ contains exactly one of $\beta_1,\ldots,\beta_\ell$. Consequently, we may choose $\{\sigma^j(A^{\beta_i})\}$, $i=1,...,\ell$, $j=1,...,h$ as our family of the root elements. The set $\{A_j^{\beta_i}\otimes s^j,T_j\otimes s^{b_j},K_0\}$, $i=1,...,\ell,j\in\mathbb{Z}$ forms a basis of $\mathfrak{g}_s$.

Remark: The properties of the automorphism σ were studied by Kostant.[12] He proved, in particular, that σ acts on the root system $\dot{\Delta}_s$ as a Coxeter transformation. Remarkably, the complex coefficients λ_i^α could be interpreted as the orthogonal projections of the roots on a real plane invariant under the Coxeter transformation. These projections were introduced by Coxeter in order to visualize the regular polytopes of higher dimensions. In Chapters 12 and 13 of Ref. 13 and in Section 12.5 of Ref. 14 the case of the root system of type H_4 is treated, which also gives the answer for E_8.

Now we shall construct the standard representation (φ,F) of $\mathfrak{s}$. The space of this representation (called the Fock space) is the polynomial algebra in the infinitely many variables:

$$F=\mathbb{C}[x_1,x_2,x_3,\ldots].$$

For $i\geqslant 1$, $T_{1-i}\otimes s^{-b_i}$ is represented by a multiplication operator:

$$\varphi(T_{1-i}\otimes s^{-b_i})=b_i x_i,$$

$T_i\otimes s^{b_i}$ is represented by a differentiation operator:

$$\varphi(T_i\otimes s^{b_i})=\frac{\partial}{\partial x_i}$$

and $\varphi(K_0)=\mathrm{Id}$.

Indeed, the only relation to be checked is

$$[\varphi(T_i\otimes s^{b_i}),\varphi(T_{1-j}\otimes s^{b_{1-j}})]=\frac{1}{h}(T_i|T_{1-j})b_i\delta_{b_ib_j}\varphi(K_0).$$

But as it can be easily seen, both expressions are equal to $b_i\delta_{ij}$ Id.

One can lift this representation of $\mathfrak{s}$ to the whole $\mathfrak{g}$ using vertex operators. We consider the space $\mathfrak{g}[[z,z^{-1}]]$ of formal Laurent series in a variable z with coefficients in $\mathfrak{g}$ and the space $\mathrm{End}(F)[[z,z^{-1}]]$ of series with coefficients in $\mathrm{End}(F)$ (see Chapter 2 in Ref. 15 for details). The adjoint action of $\mathfrak{g}$ on $\mathfrak{g}[[z,z^{-1}]]$ is well-defined.

Every representation $\varphi:\mathfrak{g}\to\mathrm{End}(F)$ defines a homomorphism $\varphi:\mathfrak{g}[[z,z^{-1}]]\to \mathrm{End}(F)[[z,z^{-1}]]$.

The following theorem shows how to lift the standard representation of the principal Heisenberg subalgebra to $\mathfrak{g}$:

Theorem 2: (Ref. 1) *There exists a representation*

$$\varphi:\mathfrak{g}\to\mathrm{End}(\mathbb{C}[x_1,x_2,\ldots])$$

such that for $i\in\mathbb{N}$

$$\varphi(T_{1-i}\otimes s^{-b_i})=b_ix_i,\quad \varphi(T_i\otimes s^{b_i})=\frac{\partial}{\partial x_i},$$

$$\varphi(K_0)=\mathrm{Id},\quad \sum_{j\in\mathbb{Z}}\varphi(A_j^\alpha\otimes s^j)z^{-j}=A^\alpha(z),$$

where

$$A^\alpha(z)=\Lambda_0(A_0^\alpha\otimes s^0)\exp\left(\sum_{i=1}^{\infty}\lambda_i^\alpha z^{b_i}x_i\right)\exp\left(-\sum_{i=1}^{\infty}\lambda_{1-i}^\alpha\frac{z^{-b_i}}{b_i}\frac{\partial}{\partial x_i}\right).$$

Since $A_j^\alpha\otimes s^j$ together with $T_i\otimes s^{b_i}$ and K_0 span $\mathfrak{g}$ then the above formulas completely determine this representation.

We will use extensively the following Laurent series:

$$\delta(z)=\sum_{j\in\mathbb{Z}}z^j\quad\text{and}\quad D\delta(z)=\sum_{j\in\mathbb{Z}}jz^j.$$

The first of these is the formal analog of the delta function.

For an element $B=\Sigma_{j\in\mathbb{Z}_h}B_j\in\dot{\mathfrak{g}}$ denote by $B(z)$ the formal series $\Sigma_{j\in\mathbb{Z}}\varphi(B_j\otimes s^j)z^{-j}\in\mathrm{End}(F)[[z,z^{-1}]]$.

Corollary 3: Let $A=\Sigma_{j=1}^h A_j$, $B=\Sigma_{j=1}^h B_j\in\dot{\mathfrak{g}}$ *and let* $C^j=[A_j,B]\in\dot{\mathfrak{g}}$. *Then*

$$[A(z_1),B(z_2)]=\sum_{j=1}^{h}\left(C^j(z_2)+\frac{j}{h}(A_j|B_{-j})\right)\left(\frac{z_2}{z_1}\right)^j\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)+\sum_{j=1}^{h}(A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^jD\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

Proof of the Corollary:

$$\begin{aligned}[A(z_1),B(z_2)]&=\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[\varphi(A_j\otimes s^j),\varphi(B_i\otimes s^i)]z_1^{-j}z_2^{-i}\\&=\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}\left\{\varphi([A_j,B_i]\otimes s^{j+i})+\frac{1}{h}(A_j|B_i)j\,\delta_{j,-i}\varphi(K_0)\right\}z_1^{-j}z_2^{-i}\\&=\sum_{j\in\mathbb{Z}}\sum_{k=i+j\in\mathbb{Z}}\varphi([A_j,B_{k-j}]\otimes s^k)z_1^{-j}z_2^{j-k}+\frac{1}{h}\sum_{j\in\mathbb{Z}}(A_j|B_{-j})j\left(\frac{z_2}{z_1}\right)^j\\&=\sum_{j\in\mathbb{Z}}C^j(z_2)\left(\frac{z_2}{z_1}\right)^j+\frac{1}{h}\sum_{j_1=1}^{h}\sum_{\substack{j_2\in\mathbb{Z}\\ j=j_1+hj_2}}(A_{j_1}|B_{-j_1})(j_1+hj_2)\left(\frac{z_2}{z_1}\right)^{j_1+hj_2}\end{aligned}$$

$$= \sum_{j_1=1}^{h} \left(C^{j_1}(z_2) + \frac{j_1}{h} (A_{j_1}|B_{-j_1}) \right) \left(\frac{z_2}{z_1} \right)^{j_1} \delta\left(\left(\frac{z_2}{z_1} \right)^h \right)$$
$$+ \sum_{j_1=1}^{h} (A_{j_1}|B_{-j_1}) \left(\frac{z_2}{z_1} \right)^{j_1} D\delta\left(\left(\frac{z_2}{z_1} \right)^h \right).$$

IV. CONSTRUCTION OF THE VERTEX OPERATOR REPRESENTATION FOR THE TOROIDAL LIE ALGEBRA

In this section we construct a vertex operator representation for the toroidal Lie algebra that generalizes the principal realization of the basic representation of affine Lie algebra. Again it will be convenient to replace $\mathfrak{g}$ with

$$\mathfrak{g}_s = \sum_{j \in \mathbb{Z}} \dot{\mathfrak{g}}_j \otimes s^j \mathbb{C}[t_1^{\pm}, \ldots, t_n^{\pm}] \oplus \mathscr{K} \oplus \mathscr{D}_+,$$

with the Lie bracket

$$[g_1 \otimes f_1(s,\mathbf{t}), g_2 \otimes f_2(s,\mathbf{t})] = [g_1, g_2] \otimes (f_1 f_2) + (g_1|g_2) \left\{ \frac{1}{h} \left(s \frac{d}{ds} f_1 \right) f_2 K_0 + \sum_{p=1}^{n} d_p(f_1) f_2 K_p \right\},$$

$$[f_1(s,\mathbf{t}) D_p, g \otimes f_2(s,\mathbf{t})] = g \otimes f_1(d_p f_2)$$

while the multiplication in $\mathscr{D}_+$ and its action on $\mathscr{K}$ are the same as in $\mathfrak{g}$. Here and below we make identifications $s^{hr_0}\mathbf{t}^{\mathbf{r}} K_a = t_0^{r_0}\mathbf{t}^{\mathbf{r}} K_a$ and $s^{hr_0}\mathbf{t}^{\mathbf{r}} D_b = t_0^{r_0}\mathbf{t}^{\mathbf{r}} D_b$, $a = 0,1,\ldots,n$, $b = 1,\ldots,n$.

The following lemma is an immediate generalization of Lemma 1:

Lemma 4: The Lie algebras $\mathfrak{g}$ and $\mathfrak{g}_s$ are isomorphic and the isomorphism is given by

$$\psi(e_\alpha \otimes t_0^{r_0}\mathbf{t}^{\mathbf{r}}) = e_\alpha \otimes s^{ht(\alpha)+hr_0}\mathbf{t}^{\mathbf{r}},$$

$$\psi(\nu^{-1}(\alpha) \otimes t_0^{r_0}\mathbf{t}^{\mathbf{r}}) = \nu^{-1}(\alpha) \otimes s^{hr_0}\mathbf{t}^{\mathbf{r}} + \frac{ht(\alpha)}{h} s^{hr_0}\mathbf{t}^{\mathbf{r}} K_0,$$

$$\psi = \mathrm{Id} \quad \text{on} \quad \mathscr{K} \oplus \mathscr{D}_+.$$

We shall identify $\mathfrak{g}$ and $\mathfrak{g}_s$ using this isomorphism.

The subalgebra $\mathfrak{s}$ with the basis $\{T_i \otimes s^{b_i}, s^{ih} D_p, s^{ih} K_p, K_0\}$, $i \in \mathbb{Z}$, $p = 1,\ldots,n$ is the principal (degenerate) Heisenberg subalgebra of $\mathfrak{g}$.

Indeed, K_0 is its central element and the multiplication in $\mathfrak{s}$ is given by

$$[T_i \otimes s^{b_i}, T_{1-j} \otimes s^{-b_j}] = b_i \delta_{ij} K_0,$$

$$[s^{ih} D_a, s^{jh} K_b] = \delta_{ab} i s^{(i+j)h} K_0 = i \delta_{ab} \delta_{i,-j} K_0,$$

$$[s^{jh} D_p, T_i \otimes s^{b_i}] = 0, \quad [s^{ih} D_a, s^{jh} D_b] = 0,$$

$$[T_i \otimes s^{b_i}, s^{jh} K_p] = 0, \quad [s^{ih} K_a, s^{jh} K_b] = 0,$$

where $i,j \in \mathbb{Z}, a,b,p = 1,\ldots,n$. This subalgebra is degenerate as a Heisenberg algebra since $[D_p, K_p] = 0$.

The Heisenberg algebra $\mathfrak{s}$ can be represented on the space

$$F = \mathbb{C}[q_p^{\pm}, x_i, u_{pi}, v_{pi}]_{i \in \mathbb{N}}^{p=1,\ldots,n}$$

by differentiation and multiplication operators:

$$\varphi(T_i\otimes s^{b_i})=\frac{\partial}{\partial x_i},\quad \varphi(T_{1-i}\otimes s^{-b_i})=b_i x_i,$$

$$\varphi(s^{ih}D_p)=\frac{\partial}{\partial u_{pi}},\quad \varphi(s^{-ih}D_p)=iv_{pi},$$

$$\varphi(s^{ih}K_p)=\frac{\partial}{\partial v_{pi}},\quad \varphi(s^{-ih}K_p)=iu_{pi},$$

$$\varphi(D_p)=q_p\frac{\partial}{\partial q_p},\quad \varphi(K_p)=0,\quad \varphi(K_0)=\mathrm{Id},$$

where $i\geqslant 1$ and $p=1,...,n$. Our goal is to extend this representation of $\mathfrak{s}$ to $\mathfrak{g}$. Consider the following elements of $\mathfrak{g}[[z,z^{-1}]]$:

$$\sum_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh},\quad \sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}.$$

Note that

$$\left[T_i\otimes s^{b_i},\sum_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}\right]=0,\quad \left[s^{ih}K_p,\sum_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}\right]=0,$$

$$\left[s^{ih}D_p,\sum_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}\right]=\sum_{j\in\mathbb{Z}} r_p s^{(i+j)h}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}=\sum_{k=i+j\in\mathbb{Z}} r_p s^{kh}\mathbf{t}^{\mathbf{r}}K_0 z^{-kh+ih}$$

$$=r_p z^{ih}\sum_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}.$$

The theory of vertex operators suggests (see Lemma 14.5 in Ref. 1) that $\Sigma_{j\in\mathbb{Z}} s^{jh}\mathbf{t}^{\mathbf{r}}K_0 z^{-jh}$ should be represented by

$$K_0(z,\mathbf{r})=\mathbf{q}^{\mathbf{r}}\exp\left(\sum_{p=1}^{n} r_p\sum_{j\geqslant 1} z^{jh}u_{pj}\right)\exp\left(-\sum_{p=1}^{n} r_p\sum_{j\geqslant 1}\frac{z^{-jh}}{j}\frac{\partial}{\partial v_{pj}}\right).$$

Here $\mathbf{q}^{\mathbf{r}}=q_1^{r_1}\cdots q_n^{r_n}$.

In a similar way the commutator relations

$$\left[T_i\otimes s^{b_i},\sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}\right]=\sum_{j\in\mathbb{Z}}\lambda_i^{\alpha}A_{j+b_i}^{\alpha}\otimes s^{j+b_i}\mathbf{t}^{\mathbf{r}}z^{-j}=\lambda_i^{\alpha}\sum_{k=j+b_i\in\mathbb{Z}} A_k^{\alpha}\otimes s^k\mathbf{t}^{\mathbf{r}}z^{-k+b_i}$$

$$=\lambda_i^{\alpha}z^{b_i}\sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j},$$

$$\left[s^{ih}K_p,\sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}\right]=0,$$

$$\left[s^{ih}D_p,\sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}\right]=\sum_{j\in\mathbb{Z}} r_p A_j^{\alpha}\otimes s^{j+ih}\mathbf{t}^{\mathbf{r}}z^{-j}$$

$$=r_p\sum_{k=j+ih\in\mathbb{Z}} A_j^{\alpha}\otimes s^k\mathbf{t}^{\mathbf{r}}z^{-k+ih}=r_p z^{ih}\sum_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}$$

suggest that $\Sigma_{j\in\mathbb{Z}} A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}$ should be represented by

$$A^{\alpha}(z,\mathbf{r})=\mathbf{q}^{\mathbf{r}}\Lambda_0(A_0^{\alpha}\otimes s^0)\exp\left(\sum_{i=1}^{\infty}\lambda_i^{\alpha}z^{b_i}x_i\right)\exp\left(-\sum_{i=1}^{\infty}\lambda_{1-i}^{\alpha}\frac{z^{-b_i}}{b_i}\frac{\partial}{\partial x_i}\right)\exp\left(\sum_{p=1}^{n}r_p\sum_{j\in\mathbb{Z}}z^{jh}u_{pj}\right)$$
$$\times\exp\left(-\sum_{p=1}^{n}r_p\sum_{j\in\mathbb{Z}}\frac{z^{-jh}}{j}\frac{\partial}{\partial v_{pj}}\right)=A^{\alpha}(z)K_0(z,\mathbf{r}).$$

The last formula hints that for $B\in\dot{\mathfrak{g}}$ we may try to represent $\Sigma_{j\in\mathbb{Z}}B_j\otimes s^j\mathbf{t}^{\mathbf{r}}z^{-j}$ by

$$B(z,\mathbf{r})=B(z)K_0(z,\mathbf{r}).$$

Note that $B(z)$ is known from the affine case (Theorem 2).

In fact, we shall see that the same approach is valid even for $\mathscr{K}$ and $\mathscr{D}_+$, but for these we should first discuss the concept of the normal ordering.

Let $X(z)=\Sigma_{i\in\mathbb{Z}}X_iz^i$, $Y(z)=\Sigma_{j\in\mathbb{Z}}Y_jz^j\in\mathrm{End}(F)[[z,z^{-1}]]$. We say that the product $X(z)Y(z)$ exists if for every $k\in\mathbb{Z}$ and for every $v\in F$ the sum

$$\sum_{i\in\mathbb{Z}}X_iY_{k-i}v$$

has finitely many nonzero terms. If this is the case then

$$X(z)Y(z)=\sum_{k\in\mathbb{Z}}\left(\sum_{i\in\mathbb{Z}}X_iY_{k-i}\right)z^k.$$

It is possible that $X(z)Y(z)$ exists while $Y(z)X(z)$ does not. For example, this happens for

$$X(z)=\sum_{i\geqslant 1}x_iz^i,\quad Y(z)=\sum_{i\geqslant 1}\frac{\partial}{\partial x_i}z^{-i}.$$

To improve the situation we define the normal ordering for the product of operators on F. For a differential operator $P\in\mathrm{End}(F)$ the normal ordering $:x_iP:$ is defined as x_iP, while the normal ordering of the product $:(\partial/\partial x_i)P:$ is defined as $P(\partial/\partial x_i)$. Note that for $X(z)$ and $Y(z)$ from the above example $:X(z)Y(z):=:Y(z)X(z):=X(z)Y(z)$.

From the action of the principle Heisenberg subalgebra on F we see that the series $\Sigma_{j\in\mathbb{Z}}s^{jh}K_pz^{-jh}$ and $\Sigma_{j\in\mathbb{Z}}s^{jh}D_pz^{-jh}$ are represented by

$$K_p(z)=\sum_{i\geqslant 1}iu_{pi}z^{ih}+\sum_{i\geqslant 1}\frac{\partial}{\partial v_{pi}}z^{-ih}$$

and

$$D_p(z)=\sum_{i\geqslant 1}iv_{pi}z^{ih}+q_p\frac{\partial}{\partial q_p}+\sum_{i\geqslant 1}\frac{\partial}{\partial u_{pi}}z^{-ih}.$$

Using the analogy with $A^{\alpha}(z,\mathbf{r})$ we shall represent $\Sigma_{j\in\mathbb{Z}}s^{jh}\mathbf{t}^{\mathbf{r}}K_pz^{-jh}$ by

$$K_p(z,\mathbf{r})=K_p(z)K_0(z,\mathbf{r})$$

and $\Sigma_{j\in\mathbb{Z}}s^{jh}\mathbf{t}^{\mathbf{r}}D_pz^{-jh}$ by

$$D_p(z,\mathbf{r})=:D_p(z)K_0(z,\mathbf{r}):.$$

Note that we use the normal ordering in the case of $D_p(z,\mathbf{r})$ in order for the product to exist.

We summarize this discussion in our main theorem:

Theorem 5: *There exists a representation φ of the toroidal Lie algebra $\mathfrak{g}_s$ on the space*

$$F=\mathbb{C}[q_p^{\pm},x_i,u_{pi},v_{pi}]_{i\in\mathbb{N}}^{p=1,\ldots,n}$$

such that for $i=1,...,\ell$, $p=1,...,n$, $\alpha\in\dot{\Delta}_s$, $\mathbf{r}\in\mathbb{Z}^n$

$$\sum_{j\in\mathbb{Z}}\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_0)z^{-jh}=K_0(z,\mathbf{r})=\mathbf{q}^{\mathbf{r}}\exp\left(\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}z^{jh}u_{pj}\right)\exp\left(-\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}\frac{z^{-jh}}{j}\frac{\partial}{\partial v_{pj}}\right), \tag{4.1}$$

$$\sum_{j\in\mathbb{Z}}\varphi(T_i\otimes s^{m_i+jh}\mathbf{t}^{\mathbf{r}})z^{-m_i-jh}=T_i(z,\mathbf{r})=\left\{\sum_{j\geqslant 1}(jh-m_i)z^{jh-m_i}x_{j\ell+1-i}+\sum_{j\geqslant 0}z^{-jh-m_i}\frac{\partial}{\partial x_{j\ell+i}}\right\}K_0(z,\mathbf{r}), \tag{4.2}$$

$$\sum_{j\in\mathbb{Z}}\varphi(A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}})z^{-j}=A^{\alpha}(z,\mathbf{r})=\Lambda_0(A_0^{\alpha}\otimes s^0)\exp\left(\sum_{i=1}^{\infty}\lambda_i^{\alpha}z^{b_i}x_i\right)\times\exp\left(-\sum_{i=1}^{\infty}\lambda_{1-i}^{\alpha}\frac{z^{-b_i}}{b_i}\frac{\partial}{\partial x_i}\right)K_0(z,\mathbf{r}), \tag{4.3}$$

$$\sum_{j\in\mathbb{Z}}\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_p)z^{-jh}=K_p(z,\mathbf{r})=\left\{\sum_{i\geqslant 1}iu_{pi}z^{ih}+\sum_{i\geqslant 1}\frac{\partial}{\partial v_{pi}}z^{-ih}\right\}K_0(z,\mathbf{r}), \tag{4.4}$$

$$\sum_{j\in\mathbb{Z}}\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_p)z^{-jh}=D_p(z,\mathbf{r})=:\left\{\sum_{i\geqslant 1}iv_{pi}z^{ih}+q_p\frac{\partial}{\partial q_p}+\sum_{i\geqslant 1}\frac{\partial}{\partial u_{pi}}z^{-ih}\right\}K_0(z,\mathbf{r}):. \tag{4.5}$$

Remark: By choosing a specialization $q_p\in\mathbb{C}\setminus\{0\}$, $p=1,...,n$, we obtain a representation of the toroidal algebra $\hat{\mathfrak{g}}$ on the Fock space

$$\hat{F}=\mathbb{C}[x_i,u_{pi},v_{pi}]_{i\in\mathbb{N}}^{p=1,...,n}.$$

V. PROOF OF THE MAIN THEOREM

First of all, let us check that formulas (4.1)–(4.5) define a linear map φ: $\mathfrak{g}\to\mathrm{End}(F)$. The linear dependencies between the momenta at the left-hand side are of the form

$$\varphi(A_j^{\sigma^k(\alpha)}\otimes s^j\mathbf{t}^{\mathbf{r}})=\zeta^{kj}\varphi(A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}}).$$

and

$$j\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_0)+\sum_{p=1}^{n}r_p\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_p)=0.$$

These dependencies extend to the following relations for the corresponding series:

$$\sum_{j\in\mathbb{Z}}\varphi(A_j^{\sigma^k(\alpha)}\otimes s^j\mathbf{t}^{\mathbf{r}})z^{-j}=\sum_{j\in\mathbb{Z}}\varphi(A_j^{\alpha}\otimes s^j\mathbf{t}^{\mathbf{r}})(\zeta^{-k}z)^{-j}$$

and

$$-\frac{1}{h}D_z\sum_{j\in\mathbb{Z}}\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_0)z^{-jh}+\sum_{p=1}^{n}r_p\sum_{j\in\mathbb{Z}}\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}K_p)z^{-jh}=0,$$

where $D_z=z(\partial/\partial z)$. We have to show that the same relations hold for the right-hand sides of (4.1)–(4.5). Indeed, noting that $\lambda_i^{\sigma^k(\alpha)}=\zeta^{-km_i}\lambda_i^{\alpha}$ we obtain

$$A^{\sigma^k(\alpha)}(z,\mathbf{r})=\Lambda_0(A_0^{\sigma^k(\alpha)}\otimes s^0)\exp\left(\sum_{i=1}^{\infty}\lambda_i^{\alpha}\zeta^{-km_i}z^{b_i}x_i\right)\exp\left(-\sum_{i=1}^{\infty}\lambda_{1-i}^{\alpha}\zeta^{km_i}\frac{z^{-b_i}}{b_i}\frac{\partial}{\partial x_i}\right)K_0(z,\mathbf{r})$$

$$=\Lambda_0(A_0^{\alpha}\otimes s^0)\exp\left(\sum_{i=1}^{\infty}\lambda_i^{\alpha}(\zeta^{-k}z)^{b_i}x_i\right)\exp\left(-\sum_{i=1}^{\infty}\lambda_{1-i}^{\alpha}\frac{(\zeta^{-k}z)^{-b_i}}{b_i}\frac{\partial}{\partial x_i}\right)K_0(\zeta^{-k}z,\mathbf{r})$$

$$=A^{\alpha}(\zeta^{-k}z,\mathbf{r}).$$

The verification of the second relation is also straightforward:

$$D_zK_0(z,\mathbf{r})=D_z\mathbf{q}^{\mathbf{r}}\exp\left(\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}z^{jh}u_{pj}\right)\exp\left(-\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}\frac{z^{-jh}}{j}\frac{\partial}{\partial v_{pj}}\right)$$

$$=\sum_{p=1}^{n}r_p\left\{\sum_{j\geqslant 1}z^{jh}jhu_{pj}+\sum_{j\geqslant 1}z^{-jh}h\frac{\partial}{\partial v_{pj}}\right\}K_0(z,\mathbf{r})=h\sum_{p=1}^{n}r_pK_p(z,\mathbf{r}).$$

Observing that the momenta of the series on the left-hand sides of (4.1)–(4.5) span $\mathfrak{g}$ we conclude that the linear map φ: $\mathfrak{g}\to\mathrm{End}(F)$ is well-defined. We need to show now that φ is a homomorphism of Lie algebras.

The following Lemma and its Corollary which we state without proof are very useful for the computations with formal series.

Lemma 6: (cf. Proposition 2.2.2 in Ref. 15). *Let*

$$\delta(z)=\sum_{k\in\mathbb{Z}}z^k,\quad D\delta(z)=D_z\delta(z)=\sum_{k\in\mathbb{Z}}kz^k.$$

If the products in the left-hand sides exist then the following equalities hold:

$$(i)\quad X(z_1)\delta\left(\frac{z_2}{z_1}\right)=X(z_2)\delta\left(\frac{z_2}{z_1}\right),$$

$$(ii)\quad X(z_1)D\delta\left(\frac{z_2}{z_1}\right)=X(z_2)D\delta\left(\frac{z_2}{z_1}\right)+D_{z_2}(X(z_2))\,\delta\left(\frac{z_2}{z_1}\right).$$

Corollary 7: Let $X(z)=Y(z^h)$. The following equalities hold provided the products in the left-hand sides exist:

$$(i)\quad X(z_1)\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)=X(z_2)\delta\left(\left(\frac{z_2}{z_1}\right)^h\right),$$

$$(ii)\quad X(z_1)D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)=X(z_2)D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)+\frac{1}{h}D_{z_2}(X(z_2))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

Note that $A^{\alpha}(z,\mathbf{r})=A^{\alpha}(z)K_0(z,\mathbf{r})$ and $T_i(z,\mathbf{r})=T_i(z)K_0(z,\mathbf{r})$, hence for every $B\in\dot{\mathfrak{g}}$ we have $B(z,\mathbf{r})=B(z)K_0(z,\mathbf{r})$.

In order to verify the commutator relations between the elements of $\Sigma_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\times[\mathbb{C}[t_1^{\pm},\ldots,t_n^{\pm}]$ we have to show that for every $A,B\in\dot{\mathfrak{g}}$,

$$\sum_{j,i\in\mathbb{Z}}\varphi([A_j\otimes s^j\mathbf{t}^{\mathbf{r}},B_i\otimes s^i\mathbf{t}^{\mathbf{m}}])z_1^{-j}z_2^{-i}=\sum_{j,i\in\mathbb{Z}}[\varphi(A_j\otimes s^j\mathbf{t}^{\mathbf{r}}),\varphi(B_i\otimes s^i\mathbf{t}^{\mathbf{m}})]z_1^{-j}z_2^{-i}.\qquad(5.1)$$

Indeed, let $C^j=[A_j,B]$. Then

$$\sum_{j,i\in\mathbb{Z}} \varphi([A_j\otimes s^j\mathbf{t}^{\mathbf{r}},B_i\otimes s^i\mathbf{t}^{\mathbf{m}}])z_1^{-j}z_2^{-i}$$

$$=\sum_{j,i\in\mathbb{Z}} \varphi([A_j,B_i]\otimes s^{j+i}\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_1^{-j}z_2^{-i}+\sum_{j,i\in\mathbb{Z}}(A_j|B_i)\varphi\left(s^{j+i}\mathbf{t}^{\mathbf{r}+\mathbf{m}}\left(\frac{j}{h}K_0+\sum_{p=1}^{n} r_pK_p\right)\right)$$

$$=\sum_{j\in\mathbb{Z}}\sum_{k=j+i\in\mathbb{Z}} \varphi([A_j,B_{k-j}]\otimes s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_1^{-j}z_2^{j-k}$$

$$+\sum_{j_1=1}^{h}\sum_{\substack{j_2\in\mathbb{Z}\\ j=j_1+hj_2}}\sum_{\substack{i_2\in\mathbb{Z}\\ i=-j_1+hi_2}}(A_{j_1}|B_{-j_1})\frac{j_1+hj_2}{h}\varphi(s^{h(j_2+i_2)}\mathbf{t}^{\mathbf{r}+\mathbf{m}}K_0)z_1^{-j_1-hj_2}z_2^{j_1-hi_2}$$

$$+\sum_{j_1=1}^{h}\sum_{\substack{j_2\in\mathbb{Z}\\ j=j_1+hj_2}}\sum_{\substack{i_2\in\mathbb{Z}\\ i=-j_1+hi_2}}(A_{j_1}|B_{-j_1})\sum_{p=1}^{n} r_p\varphi(s^{h(j_2+i_2)}\mathbf{t}^{\mathbf{r}+\mathbf{m}}K_p)z_1^{-j_1-hj_2}z_2^{j_1-hi_2}$$

$$=\sum_{j_1=1}^{h}\sum_{j_2\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\varphi(C_k^{j_1}\otimes s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_2^{-k}\left(\frac{z_2}{z_1}\right)^{j_1}\left(\frac{z_2}{z_1}\right)^{hj_2}$$

$$+\sum_{j_1=1}^{h}\sum_{j_2\in\mathbb{Z}}\sum_{k=j_2+i_2\in\mathbb{Z}}(A_{j_1}|B_{-j_1})\frac{j_1+hj_2}{h}\varphi(s^{hk}\mathbf{t}^{\mathbf{r}+\mathbf{m}}K_0)z_1^{-j_1-hj_2}z_2^{j_1+hj_2-hk}$$

$$+\sum_{j_1=1}^{h}\sum_{j_2\in\mathbb{Z}}\sum_{k=j_2+i_2\in\mathbb{Z}}(A_{j_1}|B_{-j_1})\sum_{p=1}^{n} r_p\varphi(s^{hk}\mathbf{t}^{\mathbf{r}+\mathbf{m}}K_p)z_1^{-j_1-hj_2}z_2^{j_1+hj_2-hk}$$

$$=\sum_{j_1=1}^{h} C^{j_1}(z_2,\mathbf{r}+\mathbf{m})\left(\frac{z_2}{z_1}\right)^{j_1}\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)+\frac{1}{h}\sum_{j_1=1}^{h} j_1(A_{j_1}|B_{-j_1})K_0(z_2,\mathbf{r}+\mathbf{m})$$

$$\times\left(\frac{z_2}{z_1}\right)^{j_1}\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)+\sum_{p=1}^{n} r_p\sum_{j_1=1}^{h}(A_{j_1}|B_{-j_1})K_p(z_2,\mathbf{r}+\mathbf{m})\left(\frac{z_2}{z_1}\right)^{j_1}\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)$$

$$+\sum_{j_1=1}^{h}(A_{j_1}|B_{-j_1})K_0(z_2,\mathbf{r}+\mathbf{m})\left(\frac{z_2}{z_1}\right)^{j_1}D\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right).$$

For the right-hand side of (5.1) we use Corollary 3, Corollary 7, and the fact that $K_0(z,\mathbf{r})K_0(z,\mathbf{m})=K_0(z,\mathbf{r}+\mathbf{m})$:

$$\sum_{j,i\in\mathbb{Z}}[\varphi(A_j\otimes s^j\mathbf{t}^{\mathbf{r}}),\varphi(B_i\otimes s^i\mathbf{t}^{\mathbf{m}})]z_1^{-j}z_2^{-i}$$

$$=[A(z_1,\mathbf{r}),B(z_2,\mathbf{m})]$$

$$=[A(z_1)K_0(z_1,\mathbf{r}),B(z_2)K_0(z_2,\mathbf{m})]$$

$$=[A(z_1),B(z_2)]K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})$$

$$=\sum_{j=1}^{h} C^j(z_2)\left(\frac{z_2}{z_1}\right)^{j}\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})$$

$$+\frac{1}{h}\sum_{j=1}^{h} j(A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^{j}\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})$$

$$+\sum_{j=1}^{h}(A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^{j}D\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})$$

$$
\begin{aligned}
&=\sum_{j=1}^{h} C^j(z_2)\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r})K_0(z_2,\mathbf{m})\\
&\quad+\frac{1}{h}\sum_{j=1}^{h} j(A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r})K_0(z_2,\mathbf{m})\\
&\quad+\sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r})K_0(z_2,\mathbf{m})\\
&\quad+\sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \frac{1}{h} D_{z_2}(K_0(z_2,\mathbf{r}))K_0(z_2,\mathbf{m})\\
&=\sum_{j=1}^{h} C^j(z_2)\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m})+\frac{1}{h}\sum_{j=1}^{h} j(A_j|B_{-j})\\
&\quad\times\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m})+\sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m})\\
&\quad+\sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \sum_{p=1}^{n} r_p K_p(z_2)K_0(z_2,\mathbf{r})K_0(z_2,\mathbf{m})\\
&=\sum_{j=1}^{h} C^j(z_2)\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m})+\frac{1}{h}\sum_{j=1}^{h} j(A_j|B_{-j})\\
&\quad\times\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m})+\sum_{p=1}^{n} r_p \sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j \delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\\
&\quad\times K_p(z_2,\mathbf{r}+\mathbf{m})+\sum_{j=1}^{h} (A_j|B_{-j})\left(\frac{z_2}{z_1}\right)^j D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) K_0(z_2,\mathbf{r}+\mathbf{m}).
\end{aligned}
$$

Thus (5.1) holds.

It is easy to see that operators $\varphi(s^{jh}t^{\mathbf{r}}K_p)$, $p=0,1,...,n$ commute with $\varphi(\Sigma_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\times\mathbb{C}[t_1^{\pm},...,t_n^{\pm}]\oplus\mathscr{K})$. This follows from the obvious equalities:

$$[K_p(z_1,\mathbf{r}),A^{\alpha}(z_2,\mathbf{m})]=[K_p(z_1,\mathbf{r}),T_i(z_2,\mathbf{m})]=0,$$

$$[K_a(z_1,\mathbf{r}),K_b(z_2,\mathbf{m})]=0,\quad p,a,b=0,1,...,n.$$

In order to check the action of $\mathscr{D}_+$ on $\Sigma_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\mathbb{C}[t_1^{\pm},...,t_n^{\pm}]$ it is sufficient to verify the following equality for the generating series:

$$\sum_{j,i\in\mathbb{Z}}\varphi([s^{jh}\mathbf{t}^{\mathbf{r}}D_p,s^i\mathbf{t}^{\mathbf{m}}A_i^{\alpha}])z_1^{-jh}z_2^{-i}=\sum_{j,i\in\mathbb{Z}}[\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_p),\varphi(A_i^{\alpha}\otimes s^i\mathbf{t}^{\mathbf{m}})]z_1^{-jh}z_2^{-i}. \tag{5.2}$$

The corresponding equalities for $T_i\otimes s^{b_i}\mathbf{t}^{\mathbf{m}}$ and $s^{ih}\mathbf{t}^{\mathbf{m}}K_p$ will follow since $s^{jh}\mathbf{t}^{\mathbf{r}}D_p$ acts as a derivation and $s^i\mathbf{t}^{\mathbf{m}}A_i^{\alpha}$ generate $\Sigma_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\mathbb{C}[t_1^{\pm},...,t_n^{\pm}]\oplus\mathscr{K}$.

Let us prove that (5.2) holds.

$$
\begin{aligned}
\sum_{j,i\in\mathbb{Z}}\varphi([s^{jh}\mathbf{t}^{\mathbf{r}}D_p,s^i\mathbf{t}^{\mathbf{m}}A_i^{\alpha}])z_1^{-jh}z_2^{-i}&=\sum_{j,i\in\mathbb{Z}} m_p\varphi(s^{i+jh}\mathbf{t}^{\mathbf{r}+\mathbf{m}}A_i^{\alpha})z_1^{-jh}z_2^{-i}\\
&=\sum_{j\in\mathbb{Z}}\sum_{k=i+jh\in\mathbb{Z}} m_p\varphi(s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}}A_k^{\alpha})z_1^{-jh}z_2^{-k+jh}\\
&=m_pA^{\alpha}(z_2,\mathbf{r}+\mathbf{m})\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).
\end{aligned}
$$

Next, we will show that the right-hand side of (5.2) reduces to the same expression. In the following computations it will be convenient to denote $\varphi(s^{jh}D_p)=\partial/\partial u_{pj}$ by $\gamma_p(j)$, $\varphi(s^{-jh}D_p)=jv_{pj}$ by $\gamma_p(-j)$ and $\varphi(D_p)=q_p(\partial/\partial q_p)$ by $\gamma_p(0)$. Note that $[\gamma_p(j),K_0(z,\mathbf{m})]=m_p z^{jh}K_0(z,\mathbf{m})$ for $j\in\mathbb{Z}$.

$$\sum_{j,i\in\mathbb{Z}}[\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_p),\varphi(A_i^{\alpha}\otimes s^i\mathbf{t}^{\mathbf{m}})]z_1^{-jh}z_2^{-i}=[D_p(z_1,\mathbf{r}),A^{\alpha}(z_2,\mathbf{m})]$$

$$=\left[:\left\{\sum_{j\in\mathbb{Z}}\gamma_p(j)z_1^{-jh}\right\}K_0(z_1,\mathbf{r}):\ ,\ A_{\alpha}(z_2)K_0(z_2,\mathbf{m})\right]$$

$$=A_{\alpha}(z_2)\left[:\left\{\sum_{j\in\mathbb{Z}}\gamma_p(j)z_1^{-jh}\right\}K_0(z_1,\mathbf{r}):\ ,\ K_0(z_2,\mathbf{m})\right]$$

$$=A^{\alpha}(z_2)\left[\left\{\sum_{j<0}\gamma_p(j)z_1^{-jh}\right\}K_0(z_1,\mathbf{r}),\ K_0(z_2,\mathbf{m})\right]$$

$$+A^{\alpha}(z_2)\left[K_0(z_1,\mathbf{r})\left\{\sum_{j\geqslant 0}\gamma_p(j)z_1^{-jh}\right\},\ K_0(z_2,\mathbf{m})\right]$$

$$=A^{\alpha}(z_2)\left[\sum_{j<0}\gamma_p(j)z_1^{-jh},\ K_0(z_2,\mathbf{m})\right]K_0(z_1,\mathbf{r})$$

$$+A^{\alpha}(z_2)K_0(z_1,\mathbf{r})\left[\sum_{j\geqslant 0}\gamma_p(j)z_1^{-jh},\ K_0(z_2,\mathbf{m})\right]$$

$$=A^{\alpha}(z_2)\sum_{j<0}z_1^{-jh}z_2^{jh}m_pK_0(z_2,\mathbf{m})K_0(z_1,\mathbf{r})$$

$$+A^{\alpha}(z_2)K_0(z_1,\mathbf{r})\sum_{j\geqslant 0}z_1^{-jh}z_2^{jh}m_pK_0(z_2,\mathbf{m})$$

$$=m_pA^{\alpha}(z_2)K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)$$

$$=m_pA^{\alpha}(z_2)K_0(z_2,\mathbf{r})K_0(z_2,\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)=m_pA^{\alpha}(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

From the above computation we can see that

$$[D_p(z_1,\mathbf{r}),K_0(z_2,\mathbf{m})]=m_pK_0(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

To complete the proof of Theorem 5 we need to compute the commutators $[\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_a),\varphi(s^{ih}\mathbf{t}^{\mathbf{m}}D_b)]$. We will check the following equality for the generating series:

$$\sum_{j,i\in\mathbb{Z}}\varphi([s^{jh}\mathbf{t}^{\mathbf{r}}D_a,\quad s^{ih}\mathbf{t}^{\mathbf{m}}D_b])z_1^{-jh}z_2^{-ih}=\sum_{j,i\in\mathbb{Z}}[\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_a),\quad \varphi(s^{ih}\mathbf{t}^{\mathbf{m}}D_b)]z_1^{-jh}z_2^{-ih}. \quad (5.3)$$

To compute the left-hand side we use multiplication in $\mathfrak{g}$:

$$\sum_{j,i\in\mathbb{Z}}\varphi([s^{jh}\mathbf{t}^{\mathbf{r}}D_a,s^{ih}\mathbf{t}^{\mathbf{m}}D_b])z_1^{-jh}z_2^{-ih}$$

$$=\sum_{j,i\in\mathbb{Z}}\varphi(m_a s^{(j+i)h}\mathbf{t}^{\mathbf{r}+\mathbf{m}}D_b-r_b s^{(j+i)h}\mathbf{t}^{\mathbf{r}+\mathbf{m}}D_a)z_1^{-jh}z_2^{-ih}$$

$$-m_a r_b\varphi\left(s^{(j+i)h}\mathbf{t}^{\mathbf{r}+\mathbf{m}}\left\{jK_0+\sum_{p=1}^{n}r_pK_p\right\}\right)z_1^{-jh}z_2^{-ih}$$

$$=\sum_{j\in\mathbb{Z}}\sum_{k=j+i\in\mathbb{Z}}(m_a\varphi(s^{kh}\mathbf{t}^{\mathbf{r}+\mathbf{m}}D_b)-r_b\varphi(s^{kh}\mathbf{t}^{\mathbf{r}+\mathbf{m}}D_a))z_1^{-jh}z_2^{-kh+jh}$$

$$-\sum_{j\in\mathbb{Z}}\sum_{k=j+i\in\mathbb{Z}}m_a r_b\varphi\left(s^{kh}\mathbf{t}^{\mathbf{r}+\mathbf{m}}\left\{jK_0+\sum_{p=1}^{n}r_pK_p\right\}\right)\Bigg)z_1^{-jh}z_2^{-kh+jh}$$

$$=(m_aD_b(z_2,\mathbf{r}+\mathbf{m})-r_bD_a(z_2,\mathbf{r}+\mathbf{m}))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)$$

$$-m_a r_b\sum_{p=1}^{n}r_pK_p(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)-m_a r_bK_0(z_2,\mathbf{r}+\mathbf{m})D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

Observe that

$$[D_a(z,\mathbf{r}),\gamma_b(i)]=\left[:\left\{\sum_{j\in\mathbb{Z}}\gamma_a(j)z^{-jh}\right\}K_0(z,\mathbf{r}):,\gamma_b(i)\right]$$

$$=\left[\sum_{j<0}\gamma_a(j)z^{-jh}K_0(z,\mathbf{r}),\gamma_b(i)\right]+\left[K_0(z,\mathbf{r})\sum_{j\geqslant 0}\gamma_a(j)z^{-jh},\gamma_b(i)\right]$$

$$=\sum_{j<0}\gamma_a(j)z^{-jh}[K_0(z,\mathbf{r}),\gamma_b(i)]+[K_0(z,\mathbf{r}),\gamma_b(i)]\sum_{j\geqslant 0}\gamma_a(j)z^{-jh}$$

$$=-r_bz^{ih}\sum_{j<0}\gamma_a(j)z^{-jh}K_0(z,\mathbf{r})-r_bz^{ih}K_0(z,\mathbf{r})\sum_{j\geqslant 0}\gamma_a(j)z^{-jh}$$

$$=-r_bz^{ih}D_a(z,\mathbf{r}).$$

We use the last equality to compute the right-hand side of (5.3):

$$\sum_{j,i\in\mathbb{Z}}[\varphi(s^{jh}\mathbf{t}^{\mathbf{r}}D_a),\varphi(s^{ih}\mathbf{t}^{\mathbf{m}}D_b)]z_1^{-jh}z_2^{-ih}$$

$$=[D_a(z_1,\mathbf{r}),D_b(z_2,\mathbf{m})]$$

$$=\left[D_a(z_1,\mathbf{r}),\sum_{i<0}\gamma_b(i)z_2^{-ih}K_0(z_2,\mathbf{m})\right]+\left[D_a(z_1,\mathbf{r}),K_0(z_2,\mathbf{m})\sum_{i\geqslant 0}\gamma_b(i)z_2^{-ih}\right]$$

$$=\sum_{i<0}[D_a(z_1,\mathbf{r}),\gamma_b(i)]z_2^{-ih}K_0(z_2,\mathbf{m})+\sum_{i<0}\gamma_b(i)z_2^{-ih}[D_a(z_1,\mathbf{r}),K_0(z_2,\mathbf{m})]$$

$$+[D_a(z_1,\mathbf{r}),K_0(z_2,\mathbf{m})]\sum_{i\geqslant 0}\gamma_b(i)z_2^{-ih}+K_0(z_2,\mathbf{m})\sum_{i\geqslant 0}[D_a(z_1,\mathbf{r}),\gamma_b(i)]z_2^{-ih}$$

$$=-r_b\sum_{i<0}D_a(z_1,\mathbf{r})z_1^{ih}z_2^{-ih}K_0(z_2,\mathbf{m})+\sum_{i<0}\gamma_b(i)z_2^{-ih}m_aK_0(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)$$

$$+m_aK_0(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\sum_{i\geqslant 0}\gamma_b(i)z_2^{-ih}-r_bK_0(z_2,\mathbf{m})\sum_{i\geqslant 0}D_a(z_1,\mathbf{r})z_1^{ih}z_2^{-ih}$$

$$
\begin{aligned}
&= m_a D_b(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - r_b \sum_{i<0}\left(\frac{z_1}{z_2}\right)^{ih} \sum_{j<0} \gamma_a(j) z_1^{-jh} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \\
&\quad - r_b \sum_{i<0}\left(\frac{z_1}{z_2}\right)^{ih} K_0(z_1,\mathbf{r}) \sum_{j\geqslant 0} \gamma_a(j) z_1^{-jh} K_0(z_2,\mathbf{m}) - r_b K_0(z_2,\mathbf{m}) \sum_{i\geqslant 0} \sum_{j<0} \gamma_a(j) z_1^{-jh} K_0(z_1,\mathbf{r}) \\
&\quad \times \left(\frac{z_1}{z_2}\right)^{ih} - r_b K_0(z_2,\mathbf{m}) \sum_{i\geqslant 0} K_0(z_1,\mathbf{r}) \left(\frac{z_1}{z_2}\right)^{ih} \sum_{j\geqslant 0} \gamma_a(j) z_1^{-jh} \\
&= m_a D_b(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - r_b \sum_{i<0} \sum_{j<0} \left(\frac{z_1}{z_2}\right)^{ih} z_1^{-jh} \gamma_a(j) K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \\
&\quad - r_b \sum_{i<0} \sum_{j\geqslant 0} \left(\frac{z_1}{z_2}\right)^{ih} z_1^{-jh} K_0(z_1,\mathbf{r}) [\gamma_a(j), K_0(z_2,\mathbf{m})] \\
&\quad - r_b \sum_{i<0} \sum_{j\geqslant 0} \left(\frac{z_1}{z_2}\right)^{ih} z_1^{-jh} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \gamma_a(j) \\
&\quad - r_b \sum_{i\geqslant 0} \sum_{j<0} z_1^{-jh} \left(\frac{z_1}{z_2}\right)^{ih} [K_0(z_2,\mathbf{m}), \gamma_a(j)] K_0(z_1,\mathbf{r}) \\
&\quad - r_b \sum_{i\geqslant 0} \sum_{j<0} z_1^{-jh} \left(\frac{z_1}{z_2}\right)^{ih} \gamma_a(j) K_0(z_2,\mathbf{m}) K_0(z_1,\mathbf{r}) \\
&\quad - r_b \sum_{i\geqslant 0} \sum_{j\geqslant 0} \left(\frac{z_1}{z_2}\right)^{ih} z_1^{-jh} K_0(z_2,\mathbf{m}) K_0(z_1,\mathbf{r}) \gamma_a(j) \\
&= m_a D_b(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - r_b \delta\left(\left(\frac{z_2}{z_1}\right)^h\right) : \left\{ \sum_{j\in\mathbb{Z}} \gamma_a(j) z_2^{-jh} \right\} K_0(z_2,\mathbf{r}) K_0(z_2,\mathbf{m}) : \\
&\quad - m_a r_b \sum_{i>0} \sum_{j\geqslant 0} \left(\frac{z_2}{z_1}\right)^{(i+j)h} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) + m_a r_b \sum_{i\leqslant 0} \sum_{j<0} \left(\frac{z_2}{z_1}\right)^{(i+j)h} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m})
\end{aligned}
$$

$$
\begin{aligned}
&= m_a D_b(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - r_b D_a(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \\
&\quad - m_a r_b \sum_{k=i+j>0} \sum_{j=0}^{k-1} \left(\frac{z_2}{z_1}\right)^{kh} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \\
&\quad + m_a r_b \sum_{k=i+j<0} \sum_{j=k}^{-1} \left(\frac{z_2}{z_1}\right)^{kh} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \\
&= (m_a D_b(z_2,\mathbf{r}+\mathbf{m}) - r_b D_a(z_2,\mathbf{r}+\mathbf{m}))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - m_a r_b \sum_{k\in\mathbb{Z}} k \left(\frac{z_2}{z_1}\right)^{kh} K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) \\
&= (m_a D_b(z_2,\mathbf{r}+\mathbf{m}) - r_b D_a(z_2,\mathbf{r}+\mathbf{m}))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - m_a r_b K_0(z_1,\mathbf{r}) K_0(z_2,\mathbf{m}) D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \\
&= (m_a D_b(z_2,\mathbf{r}+\mathbf{m}) - r_b D_a(z_2,\mathbf{r}+\mathbf{m}))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \\
&\quad - \frac{1}{h} m_a r_b D_{z_2}(K_0(z_2,\mathbf{r})) K_0(z_2,\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - m_a r_b K_0(z_2,\mathbf{r}) K_0(z_2,\mathbf{m}) D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) \\
&= (m_a D_b(z_2,\mathbf{r}+\mathbf{m}) - r_b D_a(z_2,\mathbf{r}+\mathbf{m}))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right) - m_a r_b \sum_{p=1}^{n} r_p K_p(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)
\end{aligned}
$$

$$-m_a r_b K_0(z_2,\mathbf{r}+\mathbf{m}) D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right).$$

This establishes (5.3) and completes the proof of Theorem 5.

VI. SUGAWARA OPERATORS

In the representation we constructed the operators from $\mathscr{D}_+$ act as derivations on the algebra $\hat{\mathfrak{g}}=\dot{\mathfrak{g}}\otimes\mathbb{C}[t_0^{\pm},t_1^{\pm},\dots,t_n^{\pm}]\oplus\mathscr{K}$. It is possible to extend this action on $\mathscr{D}$, that is to represent $t_0^j\mathbf{t}^{\mathbf{r}}D_0$ by operators on F. For the affine case these are the Sugawara operators.

We will work with the toroidal algebra realized as $\Sigma_{j\in\mathbb{Z}}\,\dot{\mathfrak{g}}_j\otimes s^j\mathbb{C}[t_1^{\pm},\dots,t_n^{\pm}]\oplus\mathscr{K}$. For this realization

$$\psi(t_0^j\mathbf{t}^{\mathbf{r}}D_0)=\frac{1}{h}\,s^{jh}\mathbf{t}^{\mathbf{r}}D_s-\frac{1}{h}\,\nu^{-1}(\rho)\otimes s^{jh}\mathbf{t}^{\mathbf{r}} \tag{6.1}$$

where $\rho\in\dot{\mathfrak{h}}^*$ with $(\rho|\alpha_i)=1$ for every simple root $\alpha_i\in\dot{\Delta}$ and $D_s=s(d/ds)$.

Since for $\alpha\in\Delta$ the elements $A^{\alpha}\otimes t_0^i\mathbf{t}^{\mathbf{m}}$ generate $\hat{\mathfrak{g}}$ then we have to check that

$$\psi([t_0^j\mathbf{t}^{\mathbf{r}}D_0,A^{\alpha}\otimes t_0^i\mathbf{t}_{\mathbf{m}}])=[\psi(t_0^j\mathbf{t}^{\mathbf{r}}D_0),\psi(A^{\alpha}\otimes t_0^i\mathbf{t}^{\mathbf{m}})].$$

Note that $A^{\alpha}\in\dot{\mathfrak{g}}_{ht(\alpha)}$ and $(\rho|\alpha)=ht(\alpha)$.

We have

$$\psi([t_0^j\mathbf{t}^{\mathbf{r}}D_0,A^{\alpha}\otimes t_0^i\mathbf{t}^{\mathbf{m}}])=i\psi(A^{\alpha}\otimes t_0^{i+j}\mathbf{t}^{\mathbf{r}+\mathbf{m}})=iA^{\alpha}\otimes s^{ht(\alpha)+h(i+j)}\mathbf{t}^{\mathbf{r}+\mathbf{m}},$$

while

$$\begin{aligned}[\psi(t_0^j\mathbf{t}^{\mathbf{r}}D_0),\psi(A^{\alpha}\otimes t_i^0\mathbf{t}^{\mathbf{m}})]&=\left[\frac{1}{h}\,s^{jh}\mathbf{t}^{\mathbf{r}}D_s-\frac{1}{h}\,\nu^{-1}(\rho)\otimes s^{jh}\mathbf{t}^{\mathbf{r}},A_{\alpha}\otimes s^{ht(\alpha)+ih}\mathbf{t}^{\mathbf{m}}\right]\\&=\frac{1}{h}\,(ht(\alpha)+ih-(\rho|\alpha))A^{\alpha}\otimes s^{ht(\alpha)+h(i+j)}\mathbf{t}^{\mathbf{r}+\mathbf{m}}\\&=iA^{\alpha}\otimes s^{ht(\alpha)+h(i+j)}\mathbf{t}^{\mathbf{r}+\mathbf{m}}.\end{aligned}$$

This establishes (6.1). It is then sufficient to construct operators on F corresponding to $s^{jh}\mathbf{t}^{\mathbf{r}}D_s$.

It will be convenient to denote in this section $\varphi(T_i\otimes s^{b_i})$ by $\tau(i)$, $\varphi(s^{jh}D_p)$ by $\gamma_p(j)$, and $\varphi(s^{jh}K_p)$ by $\kappa_p(j)$.

In these notations

$$K_0(z,\mathbf{r})=\mathbf{q}^{\mathbf{r}}\exp\left(\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}\frac{\kappa_p(-j)}{j}\,z^{jh}\right)\exp\left(-\sum_{p=1}^{n}r_p\sum_{j\geqslant 1}\frac{\kappa_p(j)}{j}\,z^{-jh}\right),$$

$$A^{\alpha}(z)=\Lambda_0(A_0^{\alpha}\otimes s^0)\exp\left(\sum_{i\geqslant 1}\lambda_i^{\alpha}\,\frac{\tau(1-i)}{b_i}\,z^{b_i}\right)\exp\left(-\sum_{i\geqslant 1}\lambda_{1-i}^{\alpha}\,\frac{\tau(i)}{b_i}\,z^{-b_i}\right).$$

Nontrivial commutators in the principal Heisenberg subalgebra are

$$[\tau(i),\tau(1-i)]=b_i\quad\text{and}\quad[\gamma_p(j),\kappa_p(-j)]=j.$$

The action of the Heisenberg subalgebra on the vertex operators is determined by

$$[\tau(i),A^{\alpha}(z)]=\lambda_i^{\alpha}z^{b_i}A^{\alpha}(z),$$

$$[\gamma_p(j),K_0(z,\mathbf{r})]=r_pz^{jh}K_0(z,\mathbf{r}).$$

Consider the following operators on the space F:

$$L_\tau(j)=\frac{1}{2}\sum_{i\in\mathbb{Z}}:\tau(1-i)\tau(i+j\not{\ell}):,$$

$$L_{\gamma\kappa}(j)=\sum_{p=1}^{n}\sum_{i\in\mathbb{Z}}:\gamma_p(-i)\kappa_p(i+j):.$$

These are the analogs of the Sugawara operators.

Construct the formal generating series for these:

$$D_s(z)=-\sum_{j\in\mathbb{Z}}(L_\tau(j)+hL_{\gamma\kappa}(j))z^{-jh}.$$

Proposition 8: Formula (4.5) together with

$$\sum_{j\in\mathbb{Z}}\varphi(s^{ih}\mathbf{t}^{\mathbf{r}}D_s)z^{-ih}=D_s(z,\mathbf{r})=:D_s(z)K_0(z,\mathbf{r}):$$

defines the representation of the Lie algebra $\mathscr{D}$ on the space

$$\varphi(\hat{\mathfrak{g}})=\varphi\left(\sum_{j\in\mathbb{Z}}\dot{\mathfrak{g}}_j\otimes s^j\mathbb{C}[t_1^{\pm},\ldots,t_n^{\pm}]\oplus\mathscr{K}\right).$$

Proof: We have an action of $\mathscr{D}$ on $\hat{\mathfrak{g}}$ which is a unique extension of its natural action on $\tilde{\mathfrak{g}}$. Thus we need to prove that $\varphi([D,B])=[\varphi(D),\varphi(B)]$ for every $D\in\mathscr{D}$ and $B\in\hat{\mathfrak{g}}$. For subalgebra $\mathscr{D}_+$ this was proved in the course of Theorem 5, hence we need to consider only $D=s^{ih}\mathbf{t}^{\mathbf{r}}D_s$. Since $s^{ih}\mathbf{t}^{\mathbf{r}}D_s$ act on $\hat{\mathfrak{g}}$ as derivations and $\hat{\mathfrak{g}}$ is generated as an algebra by elements $A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}}$ then it is sufficient to prove that

$$\varphi([s^{ih}\mathbf{t}^{\mathbf{r}}D_s,A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}}])=[\varphi(s^{ih}\mathbf{t}^{\mathbf{r}}D_s),\varphi(A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}})].$$

Again we shall replace this with an equivalent identity for the corresponding series:

$$\varphi\left(\left[\sum_{i\in\mathbb{Z}}s^{ih}\mathbf{t}^{\mathbf{r}}D_sz_1^{-ih},\sum_{j\in\mathbb{Z}}A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}}z_2^{-j}\right]\right)=\left[\sum_{i\in\mathbb{Z}}\varphi(s^{ih}\mathbf{t}^{\mathbf{r}}D_s)z_1^{-ih},\sum_{j\in\mathbb{Z}}\varphi(A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}})z_2^{-j}\right]. \tag{6.2}$$

We have

$$\begin{aligned}
&\varphi\left(\left[\sum_{i\in\mathbb{Z}}s^{ih}\mathbf{t}^{\mathbf{r}}D_sz_1^{-ih},\sum_{j\in\mathbb{Z}}A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}}z_2^{-j}\right]\right)\\
&\quad=\sum_{i\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}j\varphi(A_j^\alpha\otimes s^{j+ih}\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_1^{-ih}z_2^{-j}\\
&\quad=\sum_{i\in\mathbb{Z}}\sum_{k=j+ih\in\mathbb{Z}}(k-ih)\varphi(A_k^\alpha\otimes s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_1^{-ih}z_2^{-k+ih}\\
&\quad=\sum_{k\in\mathbb{Z}}k\varphi(A_k^\alpha\otimes s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_2^{-k}\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)-h\sum_{k\in\mathbb{Z}}\varphi(A_k^\alpha\otimes s^k\mathbf{t}^{\mathbf{r}+\mathbf{m}})z_2^{-k}D\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\\
&\quad=-(D_{z_2}A^\alpha(z_2,\mathbf{r}+\mathbf{m}))\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)-A^\alpha(z_2,\mathbf{r}+\mathbf{m})\left(D_{z_2}\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right)\\
&\quad=-D_{z_2}\left(A^\alpha(z_2,\mathbf{r}+\mathbf{m})\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right),
\end{aligned}$$

where $D_{z_2}=z_2(\partial/\partial z_2)$.

In order to compute the right-hand side of (6.2), we will need two lemmas.

Lemma 9:

$$(i) \quad [L_\tau(j), A^\alpha(z)] = z^{jh}(D_z + jh)A^\alpha(z),$$

$$(ii) \quad [L_\tau(z_1), A^\alpha(z_2)] = D_{z_2}\left(A^\alpha(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right).$$

Proof: We will prove (i) assuming that $j \geqslant 0$. The case $j < 0$ is analogous.

$$\begin{aligned}
[L_\tau(j), A^\alpha(z)] &= \frac{1}{2}\left[\sum_{i\in\mathbb{Z}} :\tau(1-i)\tau(i+j\not{\prime}):, A^\alpha(z)\right] \\
&= \frac{1}{2}\left[\sum_{i>0} \tau(1-i)\tau(i+j\not{\prime}), A^\alpha(z)\right] + \frac{1}{2}\left[\sum_{i\leqslant 0} \tau(i+j\not{\prime})\tau(1-i), A^\alpha(z)\right] \\
&= \frac{1}{2}\sum_{i>0} \tau(1-i)[\tau(i+j\not{\prime}), A^\alpha(z)] + \frac{1}{2}\sum_{i>0} [\tau(1-i), A^\alpha(z)]\tau(i+j\not{\prime}) \\
&\quad + \frac{1}{2}\sum_{i\leqslant 0} \tau(i+j\not{\prime})[\tau(1-i), A^\alpha(z)] + \frac{1}{2}\sum_{i\leqslant 0} [\tau(i+j\not{\prime}), A^\alpha(z)]\tau(1-i) \\
&= \frac{1}{2}\sum_{i>0} \tau(1-i)\lambda^\alpha_i z^{b_i+jh}A^\alpha(z) + \frac{1}{2}\sum_{i>0} \lambda^\alpha_{1-i} z^{-b_i}A^\alpha(z)\tau(i+j\not{\prime}) \\
&\quad + \frac{1}{2}\sum_{i\leqslant 0} \tau(i+j\not{\prime})\lambda^\alpha_{1-i} z^{-b_i}A^\alpha(z) + \frac{1}{2}\sum_{i\leqslant 0} \lambda^\alpha_i z^{b_i+jh}A^\alpha(z)\tau(1-i) \\
&= \frac{1}{2}\sum_{i>0} \tau(1-i)\lambda^\alpha_i z^{b_i+jh}A^\alpha(z) + \frac{1}{2}\sum_{i>0} \lambda^\alpha_{1-i} z^{-b_i}A^\alpha(z)\tau(i+j\not{\prime}) \\
&\quad + \frac{1}{2}\sum_{i\leqslant -j\not{\prime}} \tau(i+j\not{\prime})\lambda^\alpha_{1-i} z^{-b_i}A^\alpha(z) + \frac{1}{2}\sum_{-j\not{\prime}<i\leqslant 0} \lambda^\alpha_{1-i} z^{-b_i}[\tau(i+j\not{\prime}), A^\alpha(z)] \\
&\quad + \frac{1}{2}\sum_{-j\not{\prime}<i\leqslant 0} \lambda^\alpha_{1-i} z^{-b_i}A^\alpha(z)\tau(i+j\not{\prime}) + \frac{1}{2}\sum_{i\leqslant 0} \lambda^\alpha_i z^{b_i+jh}A^\alpha(z)\tau(1-i) \\
&= \frac{1}{2}\sum_{k=i>0} \tau(1-k)\lambda^\alpha_k z^{b_k+jh}A^\alpha(z) + \frac{1}{2}\sum_{k=i+j\not{\prime}>j\not{\prime}} \lambda^\alpha_{1-k} z^{-b_k+jh}A^\alpha(z)\tau(k) \\
&\quad + \frac{1}{2}\sum_{k=1-i-j\not{\prime}>0} \lambda^\alpha_k z^{b_k+jh}\tau(1-k)A^\alpha(z) + \frac{1}{2}\sum_{-j\not{\prime}<i\leqslant 0} \lambda^\alpha_{1-i} z^{-b_i}\lambda^\alpha_i z^{b_i+jh}A^\alpha(z) \\
&\quad + \frac{1}{2}\sum_{0<k=i+j\not{\prime}\leqslant j\not{\prime}} \lambda^\alpha_{1-k} z^{-b_k+jh}A^\alpha(z)\tau(k) + \frac{1}{2}\sum_{k=1-i>0} \lambda^\alpha_{1-k} z^{-b_k+jh}A^\alpha(z)\tau(k) \\
&= z^{jh}D_z A^\alpha(z) + \frac{j}{2}\,z^{jh}\left(\sum_{k=1}^{\not{\prime}} \lambda^\alpha_k\lambda^\alpha_{1-k}\right)A^\alpha(z) = z^{jh}(D_z + hj)A^\alpha(z).
\end{aligned}$$

At the last step we used the equality

$$\frac{1}{2}\sum_{k=1}^{\not{\prime}} \lambda^\alpha_k\lambda^\alpha_{1-k} = \frac{1}{2}\sum_{k=1}^{\not{\prime}} (T_k|\nu^{-1}(\alpha))(T_{1-k}|\nu^{-1}(\alpha)) = \frac{h}{2}\,(\nu^{-1}(\alpha)|\nu^{-1}(\alpha)) = \frac{h}{2}\,(\alpha|\alpha) = h,$$

which follows from the fact that $\{1/h\ T_{1-k}\}$ is the dual basis for $\{T_k\}$.

Part (ii) is an immediate consequence of (i):

$$[L_\tau(z_1),A^\alpha(z_2)]=\sum_{j\in\mathbb{Z}} z_1^{-jh}[L_\tau(j),A^\alpha(z_2)]$$

$$=\sum_{j\in\mathbb{Z}} z_1^{-jh}z_2^{jh}(D_{z_2}+hj)A^\alpha(z_2)$$

$$=(D_{z_2}A^\alpha(z_2))\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)+A^\alpha(z_2)\left(D_{z_2}\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right)=D_{z_2}\left(A^\alpha(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right).$$

Lemma 10:

$$[:L_{\gamma\kappa}(z_1)K_0(z_1,\mathbf{r}):,K_0(z_2,\mathbf{m})]=\frac{1}{h}\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)D_{z_2}(K_0(z_2,\mathbf{m}))K_0(z_1,\mathbf{r}).$$

Proof:

$$[:L_{\gamma\kappa}(z_1)K_0(z_1,\mathbf{r}):,K_0(z_2,\mathbf{m})]$$

$$=\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}\sum_{i>0}[\gamma_p(-i)\kappa_p(i+j)K_0(z_1,\mathbf{r}),K_0(z_2,\mathbf{m})]$$

$$+\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}\sum_{i\leqslant 0}[\kappa_p(i+j)K_0(z_1,\mathbf{r})\gamma_p(-i),K_0(z_2,\mathbf{m})]$$

$$=\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}\sum_{i>0}[\gamma_p(-i),K_0(z_2,\mathbf{m})]\kappa_p(i+j)K_0(z_1,\mathbf{r})$$

$$+\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}\sum_{i\leqslant 0}\kappa_p(i+j)K_0(z_1,\mathbf{r})[\gamma_p(-i),K_0(z_2,\mathbf{m})]$$

$$=\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}m_p\sum_{i>0} z_2^{-ih}K_0(z_2,\mathbf{m})\kappa_p(i+j)K_0(z_1,\mathbf{r})$$

$$+\sum_{j\in\mathbb{Z}} z_1^{-jh}\sum_{p=1}^{n}m_p\sum_{i\leqslant 0}\kappa_p(i+j)K_0(z_1,\mathbf{r})z_2^{-ih}K_0(z_2,\mathbf{m})$$

$$=\sum_{j\in\mathbb{Z}}\sum_{k=i+j\in\mathbb{Z}}\sum_{p=1}^{n}m_p z_1^{-jh}z_2^{-kh+jh}\kappa_p(k)K_0(z_2,\mathbf{m})K_0(z_1,\mathbf{r})$$

$$=\frac{1}{h}\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)D_{z_2}(K_0(z_2,\mathbf{m}))K_0(z_1,\mathbf{r}).$$

Using these two lemmas we can handle the right-hand side of (6.2).

$$\left[\sum_{i\in\mathbb{Z}}\varphi(s^{ih}\mathbf{t}^{\mathbf{r}}D_s)z_1^{-ih},\ \sum_{j\in\mathbb{Z}}\varphi(A_j^\alpha\otimes s^j\mathbf{t}^{\mathbf{m}})z_2^{-j}\right]$$

$$=[D_s(z_1,\mathbf{r}),\ A^\alpha(z_2,\mathbf{m})]$$

$$=[:D_s(z_1)K_0(z_1,\mathbf{r}):,A^\alpha(z_2)K_0(z_2,\mathbf{m})]$$

$$=-[L_\tau(z_1)K_0(z_1,\mathbf{r}),A^\alpha(z_2)K_0(z_2,\mathbf{m})]-h[:L_{\gamma\kappa}(z_1)K_0(z_1,\mathbf{r}):,A^\alpha(z_2)K_0(z_2,\mathbf{m})]$$

$$=-[L_\tau(z_1),A^\alpha(z_2)]K_0(z_1,\mathbf{r})K_0(z_2,\mathbf{m})-hA^\alpha(z_2)[:L_{\gamma\kappa}(z_1)K_0(z_1,\mathbf{r}):,K_0(z_2,\mathbf{m})]$$

$$=-D_{z_2}\left(A^\alpha(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)\right)K_0(z_2,\mathbf{m})K_0(z_1,\mathbf{r})-A^\alpha(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^h\right)D_{z_2}(K_0(z_2,\mathbf{m}))K_0(z_1,\mathbf{r})$$

$$= -D_{z_2}\left(A^{\alpha}(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right) K_0(z_2,\mathbf{m})K_0(z_1,\mathbf{r})\right)$$

$$= -D_{z_2}\left(A^{\alpha}(z_2)\,\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right) K_0(z_2,\mathbf{m})K_0(z_2,\mathbf{r})\right) = -D_{z_2}\left(A^{\alpha}(z_2,\mathbf{r}+\mathbf{m})\,\delta\left(\left(\frac{z_2}{z_1}\right)^{h}\right)\right).$$

ACKNOWLEDGMENT

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

[1] V. G. Kac, *Infinite-Dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).
[2] S. Eswara Rao and R. V. Moody, ''Vertex representations for n-toroidal Lie algebras and a generalization of the Virasoro algebra,'' Commun. Math. Phys. **159**, 239–264 (1994).
[3] J. Lepowsky and R. L. Wilson, ''Construction of affine Lie algebra $A_1^{(1)}$,'' Commun. Math. Phys. **62**, 43–53 (1978).
[4] V. G. Kac, D. A. Kazhdan, J. Lepowsky, and R. L. Wilson, ''Realization of the basic representation of the Euclidean Lie algebras,'' Adv. Math. **42**, 83–112 (1981).
[5] S. Berman and B. Cox, ''Enveloping algebras and representations of toroidal Lie algebras,'' Pacific J. Math. **165**, 239–267 (1994).
[6] Y. Billig (unpublished).
[7] T. Inami, H. Kanno, T. Ueno, and C.-S. Xiong, ''Two-toroidal Lie algebra as current algebra of four-dimensional Kähler WZW model,'' Phys. Lett. B **399**, 97–104 (1997).
[8] T. Inami, H. Kanno, and T. Ueno, ''Higher-dimensional WZW model on Kähler manifold and toroidal Lie algebra,'' Phys. Lett. A **12**, 2757–2764 (1997).
[9] C. Kassel, ''Kähler differentials and coverings of complex simple Lie algebras extended over a commutative algebra,'' J. Pure Appl. Algebra **34**, 265–275 (1985).
[10] R. V. Moody, S. Eswara Rao, and T. Yokonuma, ''Toroidal Lie algebras and vertex representations,'' Geom. Ded. **35**, 283–307 (1990).
[11] G. M. Benkart and R. V. Moody, ''Derivations, central extensions, and affine Lie algebras,'' Algebras Groups Geom. **3**, 456–492 (1986).
[12] B. Kostant, ''The principal three-dimensional subgroup and the Betti numbers of a complex simple Lie group,'' Am. J. Math. **81**, 973–1032 (1959).
[13] H. S. M. Coxeter, *Regular Polytopes* (Methuen, London, 1948).
[14] H. S. M. Coxeter, *Regular Complex Polytopes* (Cambridge University Press, Cambridge, 1974).
[15] I. Frenkel, J. Lepowsky, and A. Meurman, *Vertex Operator Algebras and the Monster* (Academic, Boston, 1989).

Equivalence between chain categories of representations of affine $sl(2)$ and $N=2$ superconformal algebras

B. L. Feigin
Landau Institute for Theoretical Physics, Russian Academy of Sciences

A. M. Semikhatov and I. Yu. Tipunin
Tamm Theory Division, Lebedev Physics Institute, Russian Academy of Sciences, Moscow 117924, Russia

(Received 17 October 1997; accepted for publication 6 January 1998)

Highest-weight-type representation theories of the affine $sl(2)$ and $N=2$ superconformal algebras are shown to be equivalent modulo the respective spectral flows.
© *1998 American Institute of Physics.* [S0022-2488(98)03606-8]

I. INTRODUCTION

In this paper, we study highest-weight-type modules over the $N=2$ superconformal algebra and the affine $sl(2)$ algebra and demonstrate the equivalence of the two representation theories modulo the respective spectral flows. The motivation comes partly from the previous analysis of $\widehat{sl}(2)$ fusion rules,[1] which suggests the occurrence of more general modules than the usual Verma modules.[2] The $N=2$ superconformal algebra,[3] which is essential in a number of conformal field-theoretic/string-theoretic constructions is known to have various relations to $\widehat{sl}(2)$. A prerequisite for constructing complete $N=2$ fusion rules is a better understanding of modules over the $N=2$ algebra, and, in particular, in view of a recent progress in constructing $\widehat{sl}(2)$ conformal blocks,[4–6] of their relation to $\widehat{sl}(2)$ modules.

In what follows, we ''compare'' $\widehat{sl}(2)$ and $N=2$ representation theories and then find certain categories built out of representations of these algebras that are equivalent. In particular, we give a complete proof of the conjecture of Ref. 7 that singular vectors in the ''topological'' (≡ ''chiral'') Verma modules over the $N=2$ algebra are isomorphic to singular vectors in $\widehat{sl}(2)$ Verma modules. We extend this result along two directions. First, the relation between topological $N=2$ Verma modules and the standard $\widehat{sl}(2)$ Verma modules is extended to the equivalence of certain categories of representations of the two algebras, which are essentially the respective *twisted* $\mathcal{O}$ categories (see Ref. 8), in fact, the categories of chains of such modules related by the respective spectral flows. Hence, the importance of twisted (≡ spectral-flow-transformed) modules; on the $N=2$ side, such modules have been around for some time,[9,10] while for the $\widehat{sl}(2)$ algebra they do not seem to have previously been studied in any detail (see, however, Ref. 11). In general, the spectral flow transformations and the canonical involution constitute the automorphism group of $\widehat{sl}(2)$ as well as of the $N=2$ algebra; the action of *automorphisms* on representations can be defined in a natural way, hence the appearance of twisted modules.

Second, we extend the correspondence between $\widehat{sl}(2)$ and $N=2$ Verma modules to representations of a different highest-weight type. For the $N=2$ algebra, one should distinguish between two types of Verma-like objects: the topological Verma modules mentioned above, and the so-called *massive* $N=2$ Verma modules. The latter are often considered[12] as the ''standard $N=2$ Verma modules.'' Their $\widehat{sl}(2)$ counterpart, however, is not so standard: these are the *relaxed Verma modules*, which differ from the usual $\widehat{sl}(2)$ Verma modules by somewhat ''relaxed'' highest-weight conditions, and as a result possess infinitely many highest-weight vectors. On the $\widehat{sl}(2)$ side, the standard $\widehat{sl}(2)$ Verma modules may appear as submodules of the relaxed Verma modules. On the $N=2$ side, the situation is similar: a massive Verma module may contain submodules that are topological Verma modules (in fact, *twisted* topological Verma modules; such are, in particular, the submodules generated from the *charged* $N=2$ singular vectors).

It thus turns out that the modules looking ''standard'' on the $N=2$ side correspond to less

 © 1998 American Institute of Physics

standard $\widehat{sl}(2)$ modules, and vice versa, which is one of the reasons for the appearance of many types of modules that we consider in the following.

A proper identification of relaxed Verma modules may be important for constructing the complete $sl(2|1)_q$-symmetric, fusion rules. The $\widehat{sl}(2)$ modules appearing in the fusion constructions and in the constructions of conformal blocks[1,2,4–6] are related to the relaxed Verma modules introduced in this paper. The relaxed Verma modules can be defined as the induced modules from the representations of the *Lie* algebra $sl(2)$ of neither the highest-weight nor the lowest-weight type. In the modules used previously, however, there occur several ''Wakimoto-type'' effects when some singular vectors vanish and cosingular vectors appear instead.

The main results of this paper are (i) the structural theory of the relaxed $\widehat{sl}(2)$ Verma modules, and (ii) the equivalence of categories. As regards the relaxed Verma modules, we classify possible degenerations, explicitly construct singular vectors in the framework of the ''complex-power'' approach,[13] and describe the nature of the corresponding submodules. The structure of relaxed Verma modules, as regards the appearance of submodules, is identical to that of $N=2$ massive Verma modules; see, however, Ref. 14 for the analysis in intrinsic $N=2$ terms.

As to the equivalence of categories, we establish the pairwise equivalence of the following categories of $\widehat{sl}(2)$ modules on the one hand and $N=2$ modules on the other hand (the respective corners of the squares):

$$\begin{array}{ccc} \mathcal{CHW} & \rightsquigarrow & \mathcal{CRHW} \\ \updownarrow & & \updownarrow \\ \mathcal{CVER} & \rightsquigarrow & \mathcal{CRVER} \\ & \widehat{sl}(2) & \end{array} \quad \text{and} \quad \begin{array}{ccc} \mathcal{CTOP} & \rightsquigarrow & \mathcal{CMHW} \\ \updownarrow & & \updownarrow \\ \mathcal{CTVER} & \rightsquigarrow & \mathcal{CMVER} \\ & N=2 & \end{array} \tag{I.1}$$

In this diagram, ''$\mathcal{C}$'' always stands for *chains*, which we define below, of modules from the corresponding categories, while the categories themselves are as follows.

In the $\widehat{sl}(2)$ square, $\mathcal{VER}$ consists of the usual $\widehat{sl}(2)$ Verma modules together with all their images under the spectral flow (i.e., all the twisted Verma modules, which we also define below). Introducing the spectral-flow-transformed $\widehat{sl}(2)$ modules is essential for establishing the relation with $N=2$ modules, where the role of the spectral flow has been appreciated for some time.[15,16] $\mathcal{RVER}$ consists of (twisted) relaxed Verma modules, which differ from the modules from $\mathcal{VER}$ by one missing annihilation condition satisfied by the highest-weight vector: as a result, relaxed Verma modules acquire infinitely many ''almost-highest-weight'' vectors. Further, in the top line of the $\widehat{sl}(2)$ square we have the respective highest-weight-*type* modules, obtained from the bottom line by taking all possible quotient modules and gluing different modules to each other. These ''$\mathcal{O}$'' categories include twisted modules as well; in what follows, we give an intrinsic definition of these categories using a criterion that is invariant under twisting. (The standard definition of the $\mathcal{O}$ category is less convenient in the case involving twisted modules, because the separation of operators from the algebra into creation and annihilation ones depends on the twist.) We call $\mathcal{HW}$ and $\mathcal{RHW}$ the categories of (twisted) highest-weight- and relaxed-highest-weight-*type* modules. For the category $\mathcal{HW}$, therefore, the ''universal'' role of Verma modules in producing irreducible representations (by taking quotients) is played by the standard $\widehat{sl}(2)$ Verma modules $\mathcal{VER}$. As regards $\mathcal{RHW}$, on the other hand, distinguished in this way are the relaxed Verma modules $\mathcal{RVER}$. As a terminological remark, however, let us note that speaking of *Verma* modules in the $\widehat{sl}(2)$ context, we will always mean only the standard Verma modules $\mathcal{VER}$, as opposed to the *relaxed Verma* modules $\mathcal{RVER}$.

On the $N=2$ side, $\mathcal{TVER}$ are the topological Verma modules and all their spectral flow transforms, while $\mathcal{TOP}$ are the corresponding topological-highest-weight-type modules, constructed from modules from $\mathcal{TVER}$ by taking quotients and gluing. We will also give a spectral-flow-invariant criterion that characterizes modules from $\mathcal{TOP}$, irrespective of their twists. Further, $\mathcal{MVER}$ are the twisted massive Verma modules, those with a zero twist are commonly viewed

as the ''standard'' $N=2$ Verma modules. Finally, $\mathcal{MHW}$ is made up of modules of the same highest-weight type as the twisted massive Verma modules, but contains many more modules, all of which can be constructed out of modules from $\mathcal{MVER}$ by taking quotients and gluing. These, too, will be defined in ''intrinsic'' terms (using the condition of terminating fermionic chains).

The *chains* of the corresponding modules have to be introduced since the mapping that we use to establish equivalence of categories maps a given $\widehat{sl}(2)$ module into a sum involving $N=2$ modules twisted by all integers. Taking chains of modules makes the respective corners of the two squares in (I.1) equivalent. In application to Verma-like modules, a part of this statement amounts to the isomorphism between singular vectors in the respective modules. As another consequence, one can further derive identities between characters of certain $\widehat{sl}(2)$ and $N=2$ modules, as we will do in this paper for the ''Verma'' case, and carry over the fusion rules from $\widehat{sl}(2)$ to $N=2$ theories (which is left for the future).

The tools essential for the analysis of $\widehat{sl}(2)$ and $N=2$ modules include diagrams of *extremal vectors*. The idea to consider extremal vectors was put forward in Ref. 17, where it was observed that a number of representation-theoretic problems can naturally be reformulated, and quite often, solved, in terms of extremal vectors. An independent construction of Ref. 10 (and a similar one[18]) can be considered as a manifestation of this general observation.

As regards singular vectors, we borrow from Ref. 10 the construction of $N=2$ singular vectors in terms of continued fermionic operators. We also develop a parallel construction of singular vectors in the relaxed Verma modules over the affine $sl(2)$. The latter construction is formulated in quite different terms than the former, however, the two constructions turn out to be equivalent and, moreover, allow direct mappings between each other. An important point is that the isomorphism between singular vectors requires that one consider the $N=2$ singular vectors that satisfy *twisted* highest-weight conditions. Allowing the singular vectors to satisfy twisted highest-weight conditions makes it possible to choose those representatives among singular vectors with different twists that generate maximal submodules (see Ref. 14 as to how this ''eliminates'' the superfluous $N=2$ subsingular vectors).

In Sec. II, we introduce the $\widehat{sl}(2)$ and $N=2$ superconformal algebras, review some of their properties (in particular, the corresponding spectral flows), investigate the structure of the corresponding Verma modules, and define different categories of such modules. We also introduce auxiliary free-field systems. In Sec. III, we recall the simplest Kazama–Suzuki (KS) mapping[19,20] and construct another mapping, which is essentially inverse to KS, even though it acts between somewhat different spaces; it will be called the anti-KS mapping. These are used in Sec. IV to prove equivalence between the respective categories of $\widehat{sl}(2)$ and $N=2$ representations.

The paper is rather long, which may be partly excused by a large number of modules that we have to introduce in order to keep balance between the $\widehat{sl}(2)$ and $N=2$ sides. We omit some of the proofs.

II. THE $\widehat{sl}(2)$ AND $N=2$ ALGEBRAS, THEIR AUTOMORPHISMS, AND VERMA MODULES

In this section, we introduce the $\widehat{sl}(2)$ and the $N=2$ superconformal algebras. We recall their properties, define several types of modules, and give theorems describing the structure of singular vectors in these modules. We also introduce free-field theories that will be needed for a constructive realization of the mapping between representations of these two algebras.

A. $\widehat{sl}(2)$

Some of the facts given below about the affine $sl(2)$ algebra are well known, yet we hope to avoid being too boring. However, we try to give more details about twisted and ''relaxed'' modules. In Sec. II A 6, we describe singular vectors in relaxed Verma modules, where the situation is more involved than with the standard Verma modules, and give structural theorems about this class of $\widehat{sl}(2)$ modules.

1. $\widehat{sl}(2)$ Verma modules, spectral flow transform, and extremal diagrams

The level-k algebra is defined as

$$[J^0_m,J^\pm_n]=\pm J^\pm_{m+n},\quad [J^0_m,J^0_n]=\frac{K}{2}\,m\,\delta_{m+n,0},$$

$$[J^+_m,J^-_n]=Km\,\delta_{m+n,0}+2J^0_{m+n} \tag{II.1}$$

(where K is the central element, which we will not distinguish from its value k in representations). Commutation relations (II.1) correspond to the following operator product expansions of currents:

$$J^0(z)J^\pm(w)=\pm\frac{J^\pm}{z-w},\quad J^0(z)J^0(w)=\frac{k/2}{(z-w)^2},$$

$$J^+(z)J^-(w)=\frac{k}{(z-w)^2}+\frac{2J^0}{z-w}. \tag{II.2}$$

In terms of the currents, the Sugawara energy–momentum tensor has the form

$$T^{\mathrm{Sug}}=\frac{1}{k+2}\left(J^0J^0+\frac{1}{2}\,(J^+J^-+J^-J^+)\right). \tag{II.3}$$

Here and in what follows, we assume $k\neq -2$. The usual mode expansion of the energy–momentum tensors is $T^{\mathrm{Sug}}(z)=\Sigma_{n\in\mathbb{Z}}L^{\mathrm{Sug}}_n z^{-n-2}$, and similarly for the other energy–momentum tensors that we work with below.

We now define the Verma module $\mathcal{M}_{j,k}$ with the highest-weight vector $|j,k\rangle_{sl(2)}$ in the standard manner, by imposing the following annihilation conditions on the highest-weight vector:

$$\begin{aligned}&J^+_{\geqslant 0}|j,k\rangle_{sl(2)}=J^0_{\geqslant 1}|j,k\rangle_{sl(2)}=J^-_{\geqslant 1}|j,k\rangle_{sl(2)}=0,\\ &J^0_0|j,k\rangle_{sl(2)}=j|j,k\rangle_{sl(2)},\quad K|j,k\rangle_{sl(2)}=k|j,k\rangle_{sl(2)},\end{aligned}\qquad j,k\in\mathbb{C}, \tag{II.4}$$

The module is freely generated from $|j,k\rangle_{sl(2)}$ by the remaining modes of J^+, J^-, and J^0; thus the character of $\mathcal{M}_{j,k}$ is given by

$$\chi^{sl(2)}_{j,k}=(z,q)=\mathrm{Tr}_{\mathcal{M}_{j,k}}(q^{L^{\mathrm{Sug}}_0}z^{J^0_0})=\frac{z^jq^{\Delta_j}}{\Pi^\infty_{i=1}(1-zq^i)\Pi^\infty_{i=0}(1-z^{-1}q^i)\Pi^\infty_{i=1}(1-q^i)}, \tag{II.5}$$

where $\Delta_j=j(j+1)/(k+2)$ is the Sugawara dimension of the highest-weight state.

The $\widehat{sl}(2)$ Verma modules can conveniently be described with the help of the *extremal diagram*,

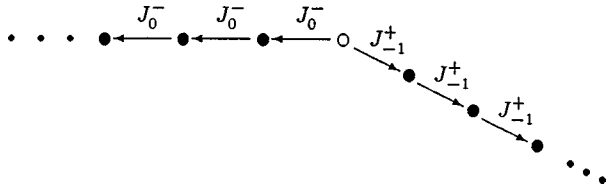

(II.6)

which expresses the fact that J^+_{-1} and J^-_0 are the highest-level operators that do not yet annihilate the highest-weight state $\circ$. All the other states of the module, according to their (charge, level) bigrading, lie in the interior of the angle in the diagram. In these conventions, e.g., J^0_{-1} is represented by a downward vertical arrow.

We will need automorphisms of the affine $sl(2)$ algebra. These are the canonical involution and the spectral flow transform,[11]

$$\mathcal{U}_\theta:\quad J^+_n\mapsto J^+_{n+\theta},\quad J^-_n\mapsto J^-_{n-\theta},\quad J^0_n\mapsto J^0_n+\frac{k}{2}\,\theta\delta_{n,0},\quad \theta\in\mathbb{Z}. \tag{II.7}$$

Spectral flow transformations with $\theta \in 2\mathbb{Z}$ are the continuation to the entire algebra of the affine Weyl group action. In general, $\widehat{sl}(2)$ modules are not invariant under the spectral flow and are mapped into ''twisted'' modules. (A notable exception is provided, for $k \in \mathbb{N}$, by the integrable representations. These are invariant under the action of spectral flow transformations with $\theta \in 2\mathbb{Z}$, while the spectral flow transformations with $\theta \in 2\mathbb{Z}+1$ map one integrable representation into another.)

Definition II.1: A twisted Verma module $\mathfrak{M}_{j,k,\theta}$ *is freely generated by* $J^+_{\leqslant \theta-1}$, $J^-_{\leqslant -\theta}$, *and* $J^0_{\leqslant -1}$ *from a twisted highest-weight vector* $|j,k;\theta\rangle_{sl(2)}$ *defined by the conditions*

$$J^+_{\geqslant \theta}|j,k;\theta\rangle_{sl(2)} = J^0_{\geqslant 1}|j,k;\theta\rangle_{sl(2)} = J^-_{\geqslant -\theta+1}|j,k;\theta\rangle_{sl(2)} = 0,$$
$$\left(J^0_0 + \frac{k}{2}\,\theta\right)|j,k;\theta\rangle_{sl(2)} = j|j,k;\theta\rangle_{sl(2)}. \tag{II.8}$$

The respective extremal diagrams are obvious ''rotations'' of (II.6). We identify $|j,k\rangle_{sl(2)} = |j,k;0\rangle_{sl(2)}$.

All possible (integral) twists of Verma modules constitute the category $\mathcal{VER}$. This category is very ''small:'' already in the untwisted case, many (if not all) interesting representations are *not* Verma modules, but rather can be obtained from Verma modules by taking quotients and ''gluing.'' This gives the category $\mathcal{O}$ (see Ref. 8), in which every irreducible representation is a quotient of a Verma module. An important feature, which, in fact, *characterizes* the modules from category $\mathcal{O}$ is that the action of all the annihilation operators on any vector from a given module spans out a finite-dimensional space (and the modules must be finite generated). In terms of extremal diagrams, this means simply that the annihilators act into the interior of the angle,

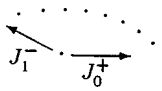

(II.9)

and therefore continuing to act with the annihilators on a state from a module from $\mathcal{O}$, one inevitably reaches the ''edge,'' at which the annihilators act by zero.

We need an $\mathcal{O}$-type category that would include the twisted modules. A remarkable fact is that it does not have to be defined by ''enumerating'' different twistings of $\mathcal{O}$, as $\cup_{\theta\in\mathbb{Z}}\mathcal{O}_\theta = \cup_{\theta\in\mathbb{Z}}\mathcal{U}_\theta\mathcal{O}$; instead, there exists an intrinsic definition. Consider first the following.

Definition II.2: Let $|X\rangle$ *be an element of a module over the affine* $sl(2)$ *algebra and let us fix an integer* θ. *For* J *being either* J^+ *or* J^-, *we say that the* J_θ *chain terminates on* $|X\rangle$, *and write* $(J_\theta)^{+\infty}|X\rangle = 0$ *if*

$$\exists N \in \mathbb{Z}, \quad n \geqslant N : (J_\theta)^n|X\rangle = 0.$$

Now, the property discussed before Eq. (II.9) is invariant under the twists. This underlies the next definition. Before giving it, let us assume once and for all that all the modules under consideration are finite generated (this requirement is often imposed in the form of the condition that the modules be graded with respect to an appropriate Virasoro generator L_0). Then, we have the following.

Definition II.3: Objects of the category $\mathcal{HW}$ *of* $\widehat{sl}(2)$ *twisted highest-weight-type representations are* $\widehat{sl}(2)$ *modules* $\mathfrak{U}$ *such that the following conditions are satisfied.*

(1) J^0_0 *is a diagonalizable operator, i.e.,* $\mathfrak{U}$ *can be decomposed into a direct sum of* J^0_0 *eigenspaces.*
(2) *The action with all the operators* J^0_p, $p \in \mathbb{N}$, *on any element* $|X\rangle \in \mathfrak{U}$ *spans a finite-dimensional space.*

(3) *For any element* $|X\rangle$ *of* $\mathfrak{U}$,

$$\forall n\in\mathbb{Z},\quad \begin{matrix} \textit{either} & (J_n^+)^{+\infty}|X\rangle=0,\\ \textit{or} & (J_{-n}^-)^{+\infty}|X\rangle=0. \end{matrix} \tag{II.10}$$

Morphisms are standard homomorphisms between $\widehat{sl}(2)$ *modules.*

Thus, conditions (II.10) do not distinguish between different θ twists of category $\mathcal{O}$. Heuristically, these conditions say that each straight line [in the conventions of (II.6)] of $\widehat{sl}(2)$ operators acting in a given module from the category intersects the boundary of the module, or simply that the extremal diagram, even though not of the simple form (II.6), is, nevertheless, angle shaped, the angle being *strictly less* than 180°.

2. Singular vectors in $\widehat{sl}(2)$ Verma modules

A singular vector in a Verma module $\mathcal{M}_{j,k}$ is a vector that is not proportional to the highest-weight vector and nevertheless satisfies the same annihilation conditions (II.4) as the highest-weight vector does.

For $\alpha\in\mathbb{C}$, we introduce the objects $(J_{-1}^+)^\alpha$ and $(J_0^-)^\alpha$ that implement the standard action of generators of the affine Weyl group on the space of highest weights; see Ref. 13 for details. These objects correspond to reflections with respect to two positive simple roots of $\widehat{sl}(2)$. In fact, $(J_{-1}^+)^\alpha$ and $(J_0^-)^\alpha$ define the following Weyl group action on the line $k=\text{const}$ in the kj plane of highest weights:

$$\begin{aligned} &(J_0^-)^{2j+1}{:}\,|j,k\rangle_{sl(2)}\to|-1-j,k\rangle_{sl(2)},\\ &(J_{-1}^+)^{k+1-2j}{:}\,|j,k\rangle_{sl(2)}\to|k+1-j,k\rangle_{sl(2)}. \end{aligned} \tag{II.11}$$

The action of $(J_{-1}^+)^\alpha$ and $(J_0^-)^\alpha$ can be extended from the highest-weight vectors to the Verma modules over these vectors. The extension is given by the following formulas:

$$\begin{aligned} (J_0^-)^\alpha J_m^+ &= -\alpha(\alpha-1)J_m^-(J_0^-)^{\alpha-2}-2\alpha J_m^0(J_0^-)^{\alpha-1}+J_m^+(J_0^-)^\alpha,\\ (J_0^-)^\alpha J_m^0 &= \alpha J_m^-(J_0^-)^{\alpha-1}+J_m^0(J_0^-)^\alpha,\\ (J_{-1}^+)^\alpha J_m^- &= -\alpha(\alpha-1)J_{m-2}^+(J_{-1}^+)^{\alpha-2}-k\alpha\delta_{m-1,0}(J_{-1}^+)^{\alpha-1}+2\alpha J_{m-1}^0(J_{-1}^+)^{\alpha-1}+J_m^-(J_{-1}^+)^\alpha,\\ (J_{-1}^+)^\alpha J_m^0 &= -\alpha J_{m-1}^+(J_{-1}^+)^{\alpha-1}+J_m^0(J_{-1}^+)^\alpha, \end{aligned} \tag{II.12}$$

which can be derived for α being a positive integer and then postulated for an arbitrary complex α. Then, we have the following theorem.

Theorem II.4:

(I) (Ref. 21) *A singular vector exists in the Verma module* $\mathcal{M}_{j,k}$ *over the affine* $sl(2)$ *algebra if and only if* $j=\mathrm{j}^+(r,s,k)$ *or* $j=\mathrm{j}^-(r,s,k)$, *where*

$$\left.\begin{aligned} \mathrm{j}^+(r,s,k)&=\frac{r-1}{2}-(k+2)\,\frac{s-1}{2}\\ \mathrm{j}^{-1}(r,s,k)&=-\frac{r+1}{2}+(k+2)\,\frac{s}{2} \end{aligned}\right\},\quad \begin{matrix} r,s\in\mathbb{N},\\ k\in\mathbb{C}. \end{matrix} \tag{II.13}$$

(II) (Ref. 13) *All singular vectors in the module* $\mathcal{M}_{j^\pm(r,s,k),k}$ *are given by the explicit constructions:*

$$
\begin{aligned}
|\mathrm{MFF}^+(r,s,k)\rangle &= (J_0^-)^{r+(s-1)(k+2)}(J_{-1}^+)^{r+(s-2)(k+2)}(J_0^-)^{r+(s-3)(k+2)}\cdots(J_{-1}^+)^{r-(s-2)(k+2)} \\
&\quad\times(J_0^-)^{r-(s-1)(k+2)}|\mathrm{j}^+(r,s,k),k\rangle_{sl(2)}, \\
|\mathrm{MFF}^-(r,s,k)\rangle &= (J_{-1}^+)^{r+(s-1)(k+2)}(J_0^-)^{r+(s-2)(k+2)}(J_{-1}^+)^{r+(s-3)(k+2)}\cdots(J_0^-)^{r-(s-2)(k+2)} \\
&\quad\times(J_{-1}^+)^{r-(s-1)(k+2)}|\mathrm{j}^-(r,s,k),k\rangle_{sl(2)}.
\end{aligned}
\tag{II.14}
$$

Remark II.5: As is well known,[13] the factors in (II.14) correspond to the affine Weyl group reflections with respect to the simple roots. When there are several singular vectors in a given Verma module, one can have different solutions to Eqs. (II.13) for (r,s), each such pair yielding a singular vector as given by Eqs. (II.14). Whenever we say ''all singular vectors,'' we mean, in accordance with Ref. 22, that the singular vectors are constructed either directly as in (II.14) or—which may be the case for rational k—by composing the factors from two such formulas (the latter being equivalent to truncating a formula from (II.14) as soon as the corresponding Weyl reflections produce a highest weight that differs from the original one by a negative integral multiple of a positive root).

Clearly, singular vectors in the twisted modules $\mathfrak{M}_{j,k;\theta}$ follow from the standard singular vectors (II.14) by applying the spectral flow transform.

3. Relaxed Verma modules

A different class of $\widehat{sl}(2)$ modules can be introduced by relaxing annihilation conditions (II.8).

Definition II.6: For $\theta\in\mathbb{Z}$, the relaxed twisted Verma module $\mathfrak{R}_{j,\Lambda,k;\theta}$ is generated from the state $|j,\Lambda,k;\theta\rangle_{sl(2)}$ that satisfies the annihilation conditions

$$
J^+_{\geqslant\theta+1}|j,\Lambda,k;\theta\rangle_{sl(2)} = J^0_{\geqslant 1}|j,\Lambda,k;\theta\rangle_{sl(2)} = J^-_{\geqslant -\theta+1}|j,\Lambda,k;\theta\rangle_{sl(2)} = 0, \tag{II.15}
$$

by a free action of the operators $J^+_{\leqslant\theta-1}$, $J^-_{\leqslant-\theta-1}$, and $J^0_{\leqslant-1}$ and by the action of operators J^+_θ and $J^-_{-\theta}$ subject to the constraint

$$
J^-_{-\theta}J^+_\theta|j,\Lambda,k;\theta\rangle_{sl(2)} = \Lambda|j,\Lambda,k;\theta\rangle_{sl(2)}. \tag{II.16}
$$

In addition, the highest-weight state $|j,\Lambda,k;\theta\rangle_{sl(2)}$ satisfies

$$
\left(J_0^0+\frac{k}{2}\,\theta\right)|j,\Lambda,k;\theta\rangle_{sl(2)} = j|j,\Lambda,k;\theta\rangle_{sl(2)}. \tag{II.17}
$$

The corresponding extremal diagram opens up to the straight angle; in the case of $\theta=0$ it thus becomes

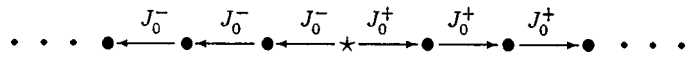

(II.18)

The state marked with ⋆ is the above $|j,\Lambda,k;\theta\rangle_{sl(2)}$, characterized, besides the level, by two quantum numbers j and Λ. The other states $|j,\Lambda,k;\theta|n\rangle_{sl(2)}$, $n\in\mathbb{Z}$, in the top floor are

$$
|j,\Lambda,k;\theta\rangle_{sl(2)} = \begin{cases} (J^-_{-\theta})^{-n}|j,\Lambda,k;\theta\rangle_{sl(2)}, & n<0, \\ (J^+_\theta)^n|j,\Lambda,k;\theta\rangle_{sl(2)}, & n>0, \end{cases} \tag{II.19}
$$

with $|j,\Lambda,k;\theta|0\rangle_{sl(2)} = |j,\Lambda,k;\theta\rangle_{sl(2)}$. We also set $|j,\Lambda,k|n\rangle_{sl(2)} = |j,\Lambda,k;0|n\rangle_{sl(2)}$. The Sugawara dimension of $|j,\Lambda,k;\theta|0\rangle_{sl(2)}$ is

$$
\Delta = \frac{j^2+j+\Lambda}{k+2} - \theta j + \frac{k}{4}\,\theta^2. \tag{II.20}
$$

Remark II.7: As we have already remarked, speaking of *Verma* modules in the $\widehat{sl}(2)$ context, we will always mean only the standard Verma modules, as opposed to the relaxed Verma modules; the same will apply to *Verma* highest-weight conditions, by which we will mean (possibly spectral-flow-transformed) highest-weight conditions in $\mathfrak{M}_{j,k;\theta}$, rather than Eqs. (II.15).

By the *relative charge* of a state $|v\rangle \in \mathcal{R}_{j,\Lambda,k}$ we understand $x=y-j$, where $J_0^0|v\rangle = y|v\rangle$ (and $J_0^0|j,\Lambda,k\rangle = j|j,\Lambda,k\rangle$). Thus, states with a negative (positive) relative charge are on the left (resp., on the right) of $\star = |j,\Lambda,k\rangle$.

In the generic case, one can travel both ways along the extremal diagram: when $\theta=0$, for example, ''untwisted'' diagram (II.18) acquires a ''fat'' form,

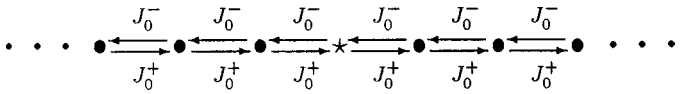

(II.21)

while the one with $\theta=-1$,

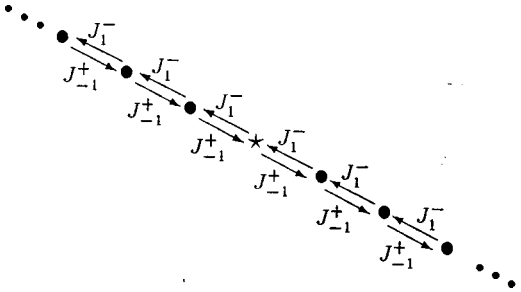

(II.22)

etc. Thus, all the points are equivalent, since the composition of the direct and the inverse arrows does in each case result only in a factor. However, this factor may vanish for some values of the parameters, which gives rise to the ordinary *Verma* modules. Namely, in the untwisted case (II.18), we have

$$\begin{aligned}
n\leqslant 0:\quad & J_0^-|j,\Lambda,k|n\rangle_{sl(2)} &&= |j,\Lambda,k|n-1\rangle_{sl(2)},\\
& J_0^+|j,\Lambda,k|n-1\rangle_{sl(2)} &&= (\Lambda-n(n-1)-2(n-1)j)|j,\Lambda,k|n\rangle_{sl(2)},\\
n\geqslant 0:\quad & J_0^+|j,\Lambda,k|n\rangle_{sl(2)} &&= |j,\Lambda,k|n+1\rangle_{sl(2)},\\
& J_0^-|j,\Lambda,k|n+1\rangle_{sl(2)} &&= (\Lambda-n(n+1)-2nj)|j,\Lambda,k|n\rangle_{sl(2)},
\end{aligned} \tag{II.23}$$

and similarly for nonzero twists $\theta\neq 0$. Whenever the parameters are such that, e.g., $J_0^+\approx 0$ at a certain step, one cannot come back to the $\star$ state by acting with J_0^+:

$$n\leqslant -1:\ \ \Lambda = n(n+1)+2nj \Rightarrow J_0^+|j,\Lambda,k|n\rangle_{sl(2)}=0. \tag{II.24}$$

One keeps on acting with J_{-1}^+ instead (one mode down); and thus the extremal diagram branches

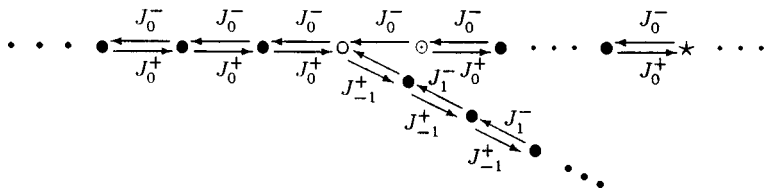

(II.25)

Therefore, any Verma module can be thought of as a *submodule* of a relaxed Verma module. Taking the quotient with respect to this Verma submodule [or, in the notation of (II.25), with

respect to $\circ$], we see that the state $\odot$ would satisfy Verma highest-weight conditions (II.8) with $\theta=1$. Hence, each Verma module is at the same time a *quotient module* of the appropriately twisted relaxed Verma module.

The situation is similar when $J_0^- \approx 0$ at a certain stage in diagram (II.21):

$$n \geqslant 1: \quad \Lambda = n(n-1)+2(n-1)j \Rightarrow J_0^- |j,\Lambda,k|n\rangle_{sl(2)} = 0. \tag{II.26}$$

Then the branching of the extremal diagram is a mirror image of (II.25), and the submodule is given by the spectral flow transform with $\theta=1$ of a standard Verma module. In both cases (II.24) and (II.26), we have

$$J_0^0 |j,\Lambda,k|n\rangle_{sl(2)} = (j+n)|j,\Lambda,k|n\rangle_{sl(2)}, \quad n \in \mathbb{Z}. \tag{II.27}$$

Hence, the Verma-module spin is $j_{\text{Verma}}=j+n$ for $n<0$ (where the Verma submodule in the relaxed module is untwisted), and $j_{\text{Verma}}=j+n+k/2$ for $n>0$ (where the Verma module is twisted by $\theta=1$).

By a straightforward application of the spectral flow, the above considerations translate into those for the *twisted* relaxed Verma modules; diagram (II.25) changes accordingly.

As in the Verma case, we now extend the category $\mathcal{RVER}$ of all twisted relaxed Verma modules to a larger category $\mathcal{RHW}$ of arbitrary (twisted) relaxed-highest-weight-type modules. We use the same method as in the case of the $\mathcal{HW}$ category. Now that the angle in the extremal diagrams has opened up to the straight angle, the extremal diagrams certainly contain straight lines that are infinite on both sides. However, a condition on the class of modules can still be given in the form of the requirement that, starting with any vector from a given module, the action with any of the bent $J^{\pm}$ chains,

terminates.

Definition II.8: The objects of category $\mathcal{RHW}$ are $\widehat{sl}(2)$ modules $\mathcal{U}$ such that the following conditions are satisfied:

(1) *J_0^0 is a diagonalizable operator.*

(2) *The action with all J_p^0, $p \in \mathbb{N}$, on any element $|X\rangle \in \mathfrak{U}$ spans a finite-dimensional space.*

(3) *For any element $|X\rangle$ of $\mathfrak{U}$,*

$$\forall\, \theta \in \mathbb{Z}, \quad \begin{array}{ll} \textit{either} & (J_\theta^+)^{+\infty}|X\rangle = 0, \\ \textit{or} & (J_{-\theta+1}^-)^{+\infty}|X\rangle = 0. \end{array} \tag{II.28}$$

Morphisms are standard homomorphisms between $\widehat{sl}(2)$ modules.

Again, this condition is invariant under the spectral flow, and thus selects all the twisted modules, none of which can be ''over-relaxed'' in the sense of its extremal diagram occupying *more* than half a plane.

4. "Charged" singular vectors

The ''Verma points'' encountered in the top floor of an extremal diagram can, of course, be viewed as singular vectors in the corresponding relaxed Verma module. The construction of these singular vectors is now obvious: we define

$$\Lambda_{\text{ch}}(p,j) = p(p+1)+2pj, \quad p \in \mathbb{Z}; \tag{II.29}$$

then the state

$$|C(p,j,k)\rangle_{sl(2)} = \begin{cases} (J_0^-)^{-p}|j,\Lambda_{\text{ch}}(p,j),k\rangle_{sl(2)}, & p \leqslant -1 \\ (J_0^+)^{p+1}|j,\Lambda_{\text{ch}}(p,j),k\rangle_{sl(2)}, & p \geqslant 0 \end{cases} \tag{II.30}$$

satisfies the Verma highest-weight conditions for $p \leqslant -1$ and the twisted Verma highest-weight conditions with the twist parameter $\theta=1$ for $p \geqslant 1$. We would like to stress that (II.30) follows

immediately from the analysis of extremal diagrams. In fact, these singular vectors may not even deserve a special name, and we mention them specifically in order to make the presentation parallel to the one for the $N=2$ algebra, where ''charged'' singular vectors are traditionally considered as a separate notion[12] [although they are given by an equivalently simple construction;[10] see (II.126)].

5. Relation with the previous constructions of some $\widehat{sl}(2)$ modules

It is of some interest to see how the $\widehat{sl}(2)$ modules constructed using the ''functional realization'' of $sl(2)$, which have appeared in the study of fusion, are related to relaxed Verma modules introduced in this paper. In the untwisted case, consider top-level states (II.18) and assume that Λ is expressed through a superficial variable μ as

$$\Lambda=\mu(\mu-1-2j)$$

(given Λ and j, we would, in general, have two different values of μ, but we do not discuss now how they can be distinguished). Then, *assuming all the relevant factors nonvanishing*, renormalize states (II.19) into

$$|j,\Lambda(\mu,j),k|n\rangle_{sl(2)}=\begin{cases} f_-(n,\mu)|j,\Lambda(\mu,j),k|n\rangle^*, & n\leqslant -1,\\ |j,\Lambda(\mu,j),k|0\rangle^*, & n=0,\\ f_+(n,j,\mu)|j,\Lambda(\mu,j),k|n\rangle^*, & n\geqslant 1,\end{cases} \tag{II.31}$$

where

$$f_-(n,\mu)=(-1)^n\,\frac{\Gamma(1-\mu-n)}{\Gamma(1-\mu)},\quad n\leqslant -1,$$

$$f_+(n,j,\mu)=(-1)^n\,\frac{\Gamma(2j+1-\mu+n)}{\Gamma(2j+1-\mu)},\quad n\geqslant 1. \tag{II.32}$$

Now, we parametrize μ and j as

$$\begin{aligned} j&=\tilde{\mu},\\ \mu&=\tilde{\mu}-s,\end{aligned} \tag{II.33}$$

and introduce a simplified notation,

$$|s,\tilde{\mu},k,n\rangle^{\#}\equiv|\tilde{\mu},(s-\tilde{\mu})(s+\tilde{\mu}+1),k|n\rangle^*. \tag{II.34}$$

Then Eqs. (II.23) are rewritten as

$$\begin{aligned} &J_0^+|s,\tilde{\mu},k,n\rangle^{\#}=-(\tilde{\mu}+n+s+1)|s,\tilde{\mu},k,n+1\rangle^{\#},\\ &J_0^0|s,\tilde{\mu},k,n\rangle^{\#}=(\tilde{\mu}+n)|s,\tilde{\mu},k,n\rangle^{\#},\qquad n\in\mathbb{Z}.\\ &J_0^-|s,\tilde{\mu},k,n\rangle^{\#}=(\tilde{\mu}+n-s-1)|s,\tilde{\mu},k,n-1\rangle^{\#}.\end{aligned} \tag{II.35}$$

We can consider the set of states $|s,\tilde{\mu},k,\cdot\rangle^{\#}$ as representing a state

$$\|s,\tilde{\mu},k\rangle\rangle^{\#}=\{|s,\tilde{\mu},k,\cdot\rangle^{\#}\},$$

which can be written in the ''x representation'' as

$$\langle x\|s,\tilde{\mu},k\rangle\rangle^{\#}=\sum_{n\in\mathbb{Z}}(-1)^n x^{\tilde{\mu}+s+n}|s,\tilde{\mu},k,n\rangle^{\#}, \tag{II.36}$$

so that the action of $J_0^{0,\pm}$ is represented by the right action of the generators,

$$S^+=\frac{\partial}{\partial x},\quad S^0=x\,\frac{\partial}{\partial x}-s,\quad S^-=-x^2\,\frac{\partial}{\partial x}+2sx. \tag{II.37}$$

Denote by $P_{s,\tilde{\mu},k}$ the module freely generated from the states $|s,\tilde{\mu},k,n\rangle^{\#}$, $n\in\mathbb{Z}$, by $J^{\pm,0}_{\leqslant -1}$.

We see from (II.24) and (II.26) that a Verma highest-weight state (a charged singular vector) occurs in the extremal diagram whenever

$$\mu=2j+p,\quad p\in\mathbb{Z},\quad \text{or}\quad \mu=p,\quad p\in\mathbb{Z}. \tag{II.38}$$

Precisely at these values of μ, one of the functions f_+ or f_- either has a pole or is identically zero, which invalidates relations (II.31). Then the charged singular vectors disappear and cosingular vectors appear instead.

6. Singular vectors in relaxed Verma modules

We are going to describe singular vectors in relaxed Verma modules only in the ''untwisted'' case $\theta=0$; similar results for modules with $\theta\neq 0$ can be obtained immediately by applying the spectral flow transform.

Our strategy is to reduce the relaxed case to the case of ordinary Verma modules. Given a relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$, we can arrive at an ordinary Verma module by the following trick. The states $(J_0^-)^{-\mu}|j,\Lambda,k\rangle$ and $(J_0^+)^{\mu+1}|j,\Lambda,k\rangle$ satisfy Verma highest-weight conditions (II.8) with $\theta=0$ or 1, respectively, whenever μ satisfies

$$\mu^2+(2j+1)\mu-\Lambda=0. \tag{II.39}$$

Therefore, the states

$$\begin{aligned}&|j_-,k;0\rangle_{sl(2)}=(J_0^-)^{-\mu}|j,\Lambda,k\rangle,\quad |j'_-,k;0\rangle_{sl(2)}=(J_0^-)^{-\mu'}|j,\Lambda,k\rangle,\quad \mu'=-\mu-2j-1,\\ &|j_+,k;1\rangle_{sl(2)}=(J_0^+)^{\mu+1}|j,\Lambda,k\rangle,\quad |j'_+,k;1\rangle_{sl(2)}=(J_0^+)^{\mu'+1}|j,\Lambda,k\rangle,\quad \mu'=-\mu-2j-1,\end{aligned} \tag{II.40}$$

which, in general, *do not belong to* $\mathcal{R}_{j,\Lambda,k}$, represent Verma highest-weight vectors, with the respective parameters

$$\begin{aligned}&j_-=j+\mu,\quad j'_-=-j-\mu-1,\\ &j_+=j+\mu+1+\frac{k}{2},\quad j'_+=-j-\mu+\frac{k}{2}.\end{aligned} \tag{II.41}$$

We now require that a singular vector exist in one of the ''continued'' modules with the highest-weight states from (II.40). Formulas (II.41) immediately give the following lemma.

Lemma II.9: The following conditions are equivalent:

$$\begin{aligned}&j_-=\mathrm{j}^{\pm}(r,s,k)\Leftrightarrow j'_+=\mathrm{j}^{\mp}(r,s,k)\Leftrightarrow j'_-=\mathrm{j}^{\mp}(r,s\mp 1,k),\\ &j'_-=\mathrm{j}^{\mp}(r,s,k)\Leftrightarrow j_+=\mathrm{j}^{\pm}(r,s,k)\Leftrightarrow j'_+=\mathrm{j}^{\mp}(r,s\pm 1,k).\end{aligned} \tag{II.42}$$

The existence of singular vectors in $\mathcal{R}_{j,\Lambda,k}$ is determined by whether and how many (if any) of the states (II.40) admit a singular vector and/or belong to $\mathcal{R}_{j,\Lambda,k}$. Whenever one of these states belongs to the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$—or, equivalently, μ or μ' is an integer of the appropriate sign—we arrive at precisely the charged singular vector considered above [observe that (II.29) is nothing but (II.39) with $\mu=p$]. As regards singular vectors that might exist on states (II.40), these must be the MFF vectors, which reduces the problem of explicitly constructing singular vectors in $\mathcal{R}_{j,\Lambda,k}$ to the MFF singular vectors.

We see from (II.41) and (II.39) that the condition for $\mathcal{R}_{j,\Lambda,k}$ to contain a relaxed Verma submodule reads as $\Lambda=\Lambda(r,s,j,k)$, where

$$\Lambda(r,s,j,k)=-(j-\mathrm{j}^-(r,s,k))(j-\mathrm{j}^+(r,s+1,k)). \tag{II.43}$$

Then, the state $(J_0^-)^{-\mu}|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}$, which satisfies the Verma highest-weight conditions, has the spin $\mathrm{j}^-(r,s,k)$. Therefore the singular vector constructed on this state reads as

$$\mathcal{MFF}^-(r,s,k)(J_0^-)^{-\mathrm{j}^-(r,s,k)+j}|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}, \tag{II.44}$$

where $\mathcal{MFF}^-$ is the singular vector *operator* [read off by dropping the highest-weight state in (II.14)]. This Verma-module singular vector has to be mapped back to the original relaxed Verma module $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$, where it will become a relaxed singular vector. In particular, no nonintegral powers of J_0^- should remain, which is achieved by acting on (II.44) with $(J_0^-)^{\mathrm{j}^-(r,s,k)-j+N}$, where N is an integer, and making use of (II.12). However, to be left with only *positive* integral powers after the rearrangements, the integer N has to be $\geqslant r+rs$. We thus define

$$|\Sigma^-(r,s,j,k)\rangle=(J_0^-)^{\mathrm{j}^-(r,s,k)-j+r+rs}\mathcal{MFF}^-(r,s,k)(J_0^-)^{-\mathrm{j}^-(r,s,k)+j}|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}, \tag{II.45}$$

as a ''canonical'' representative of the relaxed singular vector in $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$.

Similarly, the state $(J_0^+)^{\mu+1}|j,\Lambda(r,s,j,k),k\rangle$ with μ as in (II.41), i.e., $(J_0^+)^{1+\mathrm{j}^+(r,s+1,k)-j}\cdot|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}$, is a Verma highest-weight state twisted by the spectral flow transform with $\theta=1$. We then construct the corresponding MFF singular vector and finally map it back to the relaxed Verma module. In this way, the singular vector in the relaxed Verma module becomes

$$|\Sigma^+(r,s,j,k)\rangle=(J_0^+)^{-1-\mathrm{j}^+(r,s+1,k)+j+r+rs}\mathcal{MFF}^{+,1}(r,s,k)(J_0^+)^{1+\mathrm{j}^+(r,s+1,k)-j}|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}, \tag{II.46}$$

where $\mathcal{MFF}^{+,\theta}$ is the spectral flow transform (II.7) of the singular vector operator $\mathcal{MFF}^+$.

The construction can be illustrated in the following extremal diagram:

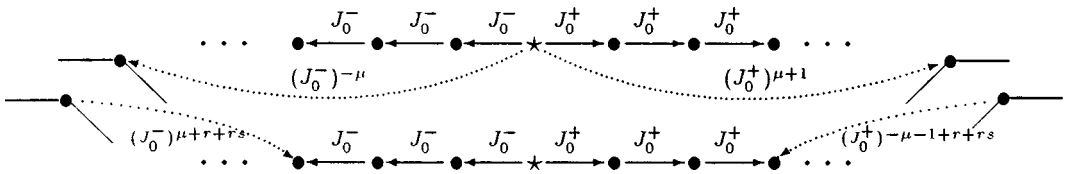

(II.47)

Every point in the lower floor is annihilated by the operators $J_{\geqslant 1}^{\pm,0}$ and, thus, satisfies the relaxed highest-weight conditions. The Verma module shown on the left contains a submodule. The Verma module and its submodule on the right are ''rotated'' by the spectral flow transform with $\theta=1$. The Verma modules are to be thought of as disconnected from the extremal diagram of the relaxed Verma module $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$, since they do not, in general, belong to $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$ (the mappings shown in dotted lines are given by complex powers of the generators).

Expressions (II.46), (II.45), and (II.30) give explicit constructions for singular vectors that can appear in relaxed Verma modules. In order to describe all degenerate cases with several singular vectors in $\mathcal{R}_{j,\Lambda,k}$ appearing simultaneously, we first classify the different degeneration patterns. Recall that, while the module $\mathcal{R}_{j,\Lambda,k}$ is irreducible for general j, Λ, and k, singular vectors appear in codimension 1, when these parameters are related by one equation. In the general case in codimension 1, j and k are still arbitrary. To describe degeneration patterns in higher codimensions, we divide the space of the highest-weights $O\equiv\{(j,\Lambda,k)\}$ into a union $\cup_i O_i=O$ of such O_i that $O_i\cap O_j=\varnothing$ and all modules $\mathcal{R}_{j,\Lambda,k}$ with $(j,\Lambda,k)\in O_i$ have the same type of degeneration for every i.

For brevity, we will say that a highest-weight state admits a singular vector whenever the Verma module generated from that state contains this singular vector and that a highest-weight state admits no singular vectors if no singular vectors exist in the Verma module. The sets $O_1,O_2,\ldots,O_n$ with the above properties can be singled out by requiring that a certain number of the states from (II.40) belong to the relaxed Verma module or/and admit singular vectors. The list of all possible cases is as follows.

codimension 0:

$O_\varnothing$, none of the states (II.40) belong to the module $\mathcal{R}_{j,\Lambda,k}$ and at least one of the states (II.40) admits no singular vectors.

codimension 1:

O_c, exactly one of the states (II.40) belongs to the module $\mathcal{R}_{j,\Lambda,k}$ and none of the states (II.40) admit a singular vector.

O_r, one of the states (II.40) admits precisely one singular vector; each of the other states (II.40) admits at least one singular vector and none of the states (II.40) belong to the module $\mathcal{R}_{j,\Lambda,k}$.

codimension 2:

O_{rr}, each of the states (II.40) admits at least two singular vectors and none of the states from (II.40) belong to $\mathcal{R}_{j,\Lambda,k}$.

O_{cc}, precisely one of the states from each line in (II.40) belongs to $\mathcal{R}_{j,\Lambda,k}$ and none of these two states admit a singular vector.

O_{rc}, one of the states from (II.40) belongs to $\mathcal{R}_{j,\Lambda,k}$ and admits precisely one singular vector; none of the states (II.40) admit two different singular vectors; no two states from different lines in (II.40) belong to $\mathcal{R}_{j,\Lambda,k}$.

codimension 3:

O_{rrc}, one of the states from (II.40) belongs to $\mathcal{R}_{j,\Lambda,k}$ and admits at least two singular vectors; no two states from different lines in (II.40) belong to $\mathcal{R}_{j,\Lambda,k}$.

O_{rcc}, precisely one of the states from each line in (II.40) belongs to $\mathcal{R}_{j,\Lambda,k}$; each of these two states admits at least one singular vector.

The sets O_i from this list can be singled out by systems of equations. Thus, the O_i of codimension 1 can be thought of as hyperplanes with certain lines eliminated, while the O_i of codimension 2 are lines with some (rational) points eliminated, and the O_i sets of codimension 3 are discrete collections of points.

We will describe degenerate cases in the order given in the above list. To begin with, observe that the completeness of the above list implies that *the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ is irreducible if and only if $(j,\Lambda,k)\in O_\varnothing$*. Further, the cases where the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ contains precisely one submodule are described in the following theorem. (Whenever we write $k+2=p/q$ or $k=p/q$ with $p\in\mathbb{Z}$, $q\in\mathbb{N}$, we always assume p and q to be coprime.)

Theorem II.10: (I) *The highest-weight (j,Λ,k) of the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ belongs to the set O_c if and only if $\Lambda=\Lambda_{\mathrm{ch}}(n,j)$ and*

$$(n,j,k)\in(\mathbb{Z}\times\mathbb{C}\times\mathbb{C})\backslash\{(n,\tfrac{1}{2}(r-2n-1)-\tfrac{1}{2}(k+2)s,k)\,|\,r,s\in\mathbb{Z},r\neq 0,r\cdot s\geqslant 0,n\in\mathbb{Z},k\in\mathbb{C}\}. \tag{II.48}$$

The highest weight of the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ belongs to the set O_r if and only if $\Lambda=\Lambda(r,s,j,k)$, with

$$(r,s,j,k)\in(\mathbb{N}\times\mathbb{N}\times\mathbb{C}\times(\mathbb{C}\backslash\mathbb{Q})\cup\mathbb{Y})\backslash\mathbb{X}(\mathbb{C}),$$

where

$$\mathbb{Y}=\left\{\left(r,s,j,-\frac{p}{q}-2\right)\,\middle|\,1\leqslant r\leqslant p,1\leqslant s\leqslant q,p,q\in\mathbb{N},j\in\mathbb{C}\right\}$$
$$\cup\left\{\left(r,1,j,-\frac{p}{q}-2\right)\,\middle|\,p+1\leqslant r\leqslant 2p,p,q\in\mathbb{N},j\in\mathbb{C}\right\}, \tag{II.49}$$

and, for any set $\mathcal{P}\subset\mathbb{C}$, we have defined

$$\mathbb{X}(\mathcal{P})=\{(r,s,\tfrac{1}{2}(\pm r-2n-1)\mp\tfrac{1}{2}(k+2)s,k)\,|\,n\in\mathbb{Z},r,s\in\mathbb{N},k\in\mathcal{P}\}. \tag{II.50}$$

(II) *In the O_c case, the module $\mathcal{R}_{j,\Lambda_{\mathrm{ch}}(n,j),k}$ contains precisely one submodule (which is isomorphic to a Verma module or a twisted Verma module) generated from singular vector (II.30).*

(III) *In the O_r case, the module $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$ contains precisely one submodule (which is isomorphic to a relaxed Verma module). Expressions (II.45) and (II.46) evaluate as elements of the relaxed Verma module $\mathcal{R}_{j,\Lambda(r,s,j,k),k}$ and satisfy the relaxed-highest-weight conditions*

$$J^+_{\geqslant 1}|\Sigma^\pm(r,s,j,k)\rangle=J^0_{\geqslant 1}|\Sigma^\pm(r,s,j,k)\rangle=J^-_{\geqslant 1}|\Sigma^\pm(r,s,j,k)\rangle=0,$$
$$J^-_0J^+_0|\Sigma^\pm(r,s,j,k)\rangle=\Lambda^\pm(r,s,j,k)|\Sigma^\pm(r,s,j,k)\rangle, \tag{II.51}$$
$$J^0_0|\Sigma^\pm(r,s,j,k)\rangle=(j\pm rs)|\Sigma^\pm(r,s,j,k)\rangle,$$

where

$$\Lambda^{\pm}(r,s,j,k)=\tfrac{1}{4}\,(r+(k+2)s-2rs\mp(1+2j))(r+(k+2)s+2rs\pm(1+2j)). \qquad \text{(II.52)}$$

Each of the singular vectors $\Sigma^{\pm}(r,s,j,k)$ generates the entire relaxed Verma submodule; in particular, $J_0^{\pm}$-descendants of (II.45) and (II.46) are on the same extremal subdiagram and coincide up to numerical factors whenever they are in the same grade:

$$\begin{aligned}
&c_+(i,j,k)(J_0^+)^i|\Sigma^-(r,s,j,k)\rangle=c_-(i,j,k)(J_0^-)^{2rs-i}|\Sigma^+(r,s,j,k)\rangle, \quad 0\leqslant i\leqslant 2rs,\\
&c_+(i,j,k)(J_0^+)^i|\Sigma^-(r,s,j,k)\rangle=c_-(i,j,k)(J_0^+)^{i-2rs}|\Sigma^+(r,s,j,k)\rangle, \quad i>2rs, \qquad \text{(II.53)}\\
&c_+(i,j,k)(J_0^-)^{-i}|\Sigma^-(r,s,j,k)\rangle=c_-(i,j,k)(J_0^-)^{-i+2rs}|\Sigma^+(r,s,j,k)\rangle, \quad i<0,
\end{aligned}$$

where the numerical coefficients $c_{\pm}(i,j,k)$ are (r-and s-dependent) polynomials in j and k.

Thus, (II.45) and (II.46) are, in general, different representatives of the same singular vector, since they generate the same submodule. This is not necessarily so in higher codimensions. To see what can happen there, consider again diagram (II.47). There, the Verma modules that contain singular vectors are shown as disconnected pieces, since they are related to the Verma module by mappings with complex powers of $J_0^{\pm}$. However, when one or both of the roots of (II.39) becomes an integer, one or both of these Verma modules are no longer disconnected from the extremal diagram of the relaxed Verma module. Then the Verma highest-weight conditions are encountered in the extremal diagram and restrict the possibilities to travel over the states in the lower floor in (II.47), since some of the states in (II.53) might vanish. For example, it may be impossible to obtain the Σ^+ representative of the singular vector as a J_0^+-descendant of Σ^-, or/and to obtain Σ^- as a J_0^--descendant of Σ^+. These cases in codimension $\geqslant 2$ describe the simultaneous occurrence of several singular vectors, and will be considered in what follows. We now proceed to codimension 2.

Codimension-2 cases: In codimension 2, we have three cases of a simultaneous occurrence of two singular vectors, which are described in the three following theorems. The following lemma governs the appearance of Verma submodules in a relaxed Verma module. If one puts $\Lambda^{\pm}(r,s,j,k)$ in (II.52) equal to $\Lambda_{\rm ch}(N,j\pm rs)$ [see Eq. (II.29)], one discovers that the condition $\Lambda^{\pm}(r,s,j,k)=\Lambda_{\rm ch}(N,j\pm rs)$ implies a similar relation for $\Lambda(r,s,j,k)$, and therefore charged singular vectors in the lower floor occur synchronously with charged singular vectors in the top floor of the extremal diagram. In this way, we arrive at the following.

Lemma II.11: Let $\mathcal{R}$ be the relaxed Verma module $\mathcal{R}\equiv\mathcal{R}_{j,\Lambda,k}$.

(1) *Let $\mathcal{R}\supset\mathcal{R}'$ and $\mathcal{R}\supset\mathcal{C}$, where $\mathcal{R}'$ is a relaxed Verma submodule and $\mathcal{C}$ is a (twisted) Verma module generated from a charged singular vector in $\mathcal{R}$, and let $\mathcal{C}$ be maximal in the sense that $\mathcal{R}\supset\mathcal{C}''\supset\mathcal{C}$, where $\mathcal{C}''$ is any other (twisted) Verma submodule, implies $\mathcal{C}''=\mathcal{C}$, Then $\mathcal{R}'\cap\mathcal{C}=\mathcal{C}'\neq\{0\}$, where $\mathcal{C}'$ is a (twisted) Verma module generated from a charged singular vector in $\mathcal{R}'$.*

(2) *Conversely, if $\mathcal{R}\supset\mathcal{R}'$, where $\mathcal{R}'$ is a relaxed Verma module and $\mathcal{R}'\supset\mathcal{C}'$, where $\mathcal{C}'$ is a submodule generated from a charged singular vector in $\mathcal{R}'$, then $\mathcal{R}\supset\mathcal{C}\supset\mathcal{C}'$, where $\mathcal{C}$ is a submodule generated from a charged singular vector in $\mathcal{R}$, and $\mathcal{R}\supset\mathcal{C}''\supset\mathcal{C}\Rightarrow\mathcal{C}''=\mathcal{C}$.*

(3) *If $\mathcal{R}\supset\mathcal{C}'$, where $\mathcal{C}'$ is a (twisted) Verma module, there exists a (twisted) Verma submodule $\mathcal{C}\subset\mathcal{R}$ such that the embedding is given by a charged singular vector and $\mathcal{R}\supset\mathcal{C}''\supset\mathcal{C}\Rightarrow\mathcal{C}''=\mathcal{C}$.*

This lemma is a necessary tool in the proof of the following three theorems. Consider first the occurrence of two different relaxed singular vectors.

We call a submodule *primitive* if it is not a sum of two other submodules.

Theorem II.12: *The highest weight of the module $\mathcal{R}_{j,\Lambda,k}$ belongs to the set O_{rr} if and only if $\Lambda=\Lambda(r,s,j,k)$, with*

$$(r,s,j,k)\in\mathbb{N}\times\mathbb{N}\times\mathbb{C}\times\mathbb{Q}\backslash(\mathbb{X}(\mathbb{Q})\cup\mathbb{Y}),$$

where $\mathbb{Y}$ and $\mathbb{X}(\mathbb{Q})$ are as in (II.49) and (II.50). Then we have the following.

(1) *Any primitive submodule in $\mathcal{R}_{j,\Lambda(r,s,j,p/q),p/q}$ is a relaxed Verma module. It is generated from $|\Sigma^-(r,s,j,p/q)\rangle$ or, equivalently, from $|\Sigma^+(r,s,j,p/q)\rangle$ with the same r and s.*

(2) *For any primitive submodule* $\mathcal{R}' \subset \mathcal{R}_{j,\Lambda(r,s,j,p/q),p/q}$, *there exists a singular vector* $e \equiv \mathcal{E}\cdot |\mathrm{j}^-(r,s,p/q),p/q\rangle_{sl(2)}$ *in the Verma module* $\mathcal{M}_{\mathrm{j}^-(r,s,p/q),p/q}$, *such that* $\mathcal{R}'$ *is generated from the relaxed singular vector*

$$(J_0^-)^{\mathrm{j}^-(r,s,k)-j+r+rs}\mathcal{E}(J_0^-)^{-\mathrm{j}^-(r,s,k)+j}|j,\Lambda(r,s,j,k),k\rangle_{sl(2)}. \tag{II.54}$$

(3) *Conversely, for any singular vector* e *in* $\mathcal{M}_{\mathrm{j}^-(r,s,p/q),p/q}$, *of the form* $e=|\mathrm{MFF}(r,s,p/q)\rangle^+$, $r\geqslant 1$, $s\geqslant 2$, *or* $e=|\mathrm{MFF}(r,s,p/q)\rangle^-$, $r,s\geqslant 1$, *there exists a relaxed Verma submodule in* $\mathcal{R}_{j,\Lambda(r,s,j,p/q),p/q}$ *generated by the relaxed singular vector (II.54).*

(4) *Two different singular vectors* e_1 *and* e_2 *in* $\mathcal{M}_{\mathrm{j}^-(r,s,p/q),p/q}$ *correspond in this way to the same relaxed Verma submodule in* $\mathcal{R}$ *if and only if one of the* e_i *is the* $\mathrm{MFF}^+(a\geqslant 1,1,p/q)$ *singular vector in the module generated from the other.*

In this case, therefore, the structure of the relaxed Verma module $\mathcal{R}_{j,\Lambda(r,s,j,p/q),p/q}$ is determined by the structure of the Verma module $\mathcal{M}_{\mathrm{j}^-(r,s,p/q),p/q}$.

Now, consider the case where both Verma modules shown in (II.47) actually belong to the relaxed extremal diagram. It is described in the following theorem.

Theorem II.13: *The highest weight of the relaxed Verma module* $\mathcal{R}_{j,\Lambda,k}$ *belongs to the set* O_{cc} *of the list of degeneration patterns if and only if* $j=j_{cc}(n,m)$, $\Lambda=\Lambda_{cc}(n,m)$, *where*

$$j_{cc}(n,m)=-\tfrac{1}{2}(1+m+n), \quad \Lambda_{cc}(n,m)=-mn, \tag{II.55}$$

and

$$(n,m,k)\in((-\mathbb{N})\times\mathbb{N}_0\times(\mathbb{C}\backslash\mathbb{Q}))\cup\left\{\left(n,m,-\frac{p}{q}-2\,\middle|\, n\in-\mathbb{N},m\in\mathbb{N}_0,p,q\in\mathbb{N},1\leqslant m-n\leqslant q\right)\right\}. \tag{II.56}$$

Then the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ *contains the Verma submodule* $\mathcal{C}_-\approx\mathcal{M}_{(1/2)(n-m-1),k}$ *generated from the charged singular vector* $|C(n,j,k)\rangle$ *and the twisted Verma submodule* $\mathcal{C}_+\approx\mathfrak{M}_{(1/2)(1+m-n+k),k;1}$ *generated from the charged singular vector* $|C(m,j,k)\rangle$. *The maximal submodule is* $\mathcal{C}_+\cup\mathcal{C}_-$, *and* $\mathcal{C}_+\cap\mathcal{C}_-=\varnothing$.

The third possibility in codimension 2, i.e., the case O_{rc} from the list, is when one of the states (II.40) belongs to the module $\mathcal{R}_{j,\Lambda,k}$ (hence this module contains a charged singular vector) and at the same time one of the states (II.40) admits a singular vector. This case is more involved and is analyzed in what follows.

According to Lemma II.11, the Verma submodule growing out of the charged singular vector on the top floor of the extremal diagram necessarily has a submodule generated from a singular vector $\mathrm{MFF}^{\pm}(r,s,k)$, and this submodule coincides with the submodule built on the charged singular vector in the lower floor of the extremal diagram, e.g.,

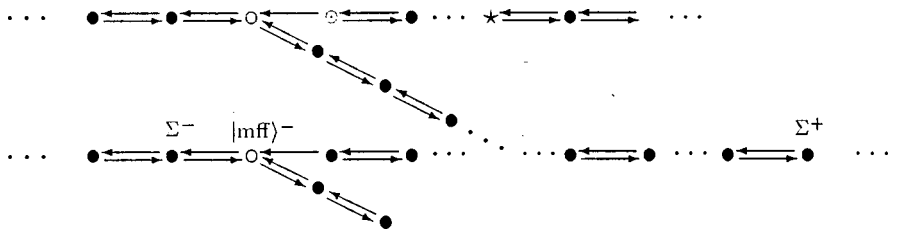

(II.57)

As we saw in (II.51), the Σ^- representative of the relaxed singular vector is at the distance of rs arrows on the left from the ''center'' (the $\star$), while Σ^+ is at the same distance on the right. It may happen that Σ^- would be inside the Verma submodule. In that case, the entire extremal diagram of the submodule cannot be generated by $J_0^{\pm}$-descendants of Σ^-. Unless we are in codimension 3, however, the entire diagram of the submodule would still be generated from Σ^+. In particular, following the arrows from Σ^+ to Σ^- into the Verma submodule makes the two representatives proportional. This is described more precisely in the following theorem.

Theorem II.14: *The highest weight of the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ belongs to the set O_{rc} of the list of degeneration patterns if and only if $j=j^{\sigma}_{rc}(r,s,n,k)$ and $\Lambda=\Lambda^{\sigma}_{rc}(r,s,n,k)$, where $\sigma\in\{+,-\}$,*

$$\begin{aligned} j^{\pm}_{rc}(r,s,n,k)&=-n-\tfrac{1}{2}\pm\tfrac{1}{2}\,(r-(k+2)s),\\ \Lambda^{\pm}_{rc}(r,s,n,k)&=-n^2\pm n(r-(k+2)s), \end{aligned} \tag{II.58}$$

and

$$(\sigma,r,s,n,k)\in(\{\pm\}\times\mathbb{N}\times\mathbb{N}\times\mathbb{Z}\times(\mathbb{C}\backslash\mathbb{Q}))\cup\mathbb{A}(\mathbb{C}\backslash\mathbb{Q})\cup\mathbb{B}, \tag{II.59}$$

where

$$\mathbb{A}(\mathcal{P})=\{(\{+\},r,0,n,k)|r\in\mathbb{N},n\in-\mathbb{N},k\in\mathcal{P}\}\cup\{(\{-\},r,0,n,k)|r\in\mathbb{N},n\in\mathbb{N}_0,k\in\mathcal{P}\},$$

$$\begin{aligned} \mathbb{B}=&\left\{\left(\{+\},r,s,n,-\frac{p}{q}-2\right)\middle|\, n\in-\mathbb{N},\ p,q\in\mathbb{N},\ 1\leqslant r\leqslant p,\ 0\leqslant s\leqslant q-1\right\}\\ &\cup\left\{\left(\{-\},r,s,n,-\frac{p}{q}-2\right)\middle|\, n\in\mathbb{N}_0,\ p,q\in\mathbb{N},\ 1\leqslant r\leqslant p,\ 0\leqslant s\leqslant q-1\right\}. \end{aligned} \tag{II.60}$$

Then, the relaxed Verma module $\mathcal{R}_{j,\Lambda,k}$ contains a relaxed Verma submodule $\mathcal{R}'$ generated from the singular vector $|\Sigma^{+}(r,s,j^{\pm}_{rc}(r,s,n,k),k)\rangle$ (if $n<0$) or from $|\Sigma^{-}(r,s,j^{\pm}_{rc}(r,s,n,k),k)\rangle$ (if $n\geqslant 0$); the module $\mathcal{R}_{j,\Lambda,k}$ contains also a Verma submodule $\mathcal{C}\approx\mathcal{M}_{\mathrm{j}^{\pm}(r,s,k),k}$ (if $n<0$) or a twisted Verma submodule $\mathcal{C}\approx\mathfrak{M}_{\mathrm{j}^{\pm}(r,s,k),k;1}$ (if $n\geqslant 0$) generated from the charged singular vector $|C(n,j,k)\rangle$ in either case. The maximal submodule in $\mathcal{R}_{j,\Lambda,k}$ is $\mathcal{C}\cup\mathcal{R}'$. The intersection $\mathcal{C}'\equiv\mathcal{C}\cap\mathcal{R}'$ is a Verma module (if $n<0$) or a twisted Verma module (if $n\geqslant 0$) generated from a charged singular vector $|c\rangle\in\mathcal{R}'$, which is also a singular vector in $\mathcal{C}$:

$$|c\rangle=\begin{cases} |\mathrm{mff}\rangle^{\pm}=\mathcal{MFF}^{\pm}(r,s+\frac{1}{2}\pm\frac{1}{2},k)|C(n,j^{\pm}_{rc}(r,s,n,k),k)\rangle, & n<0;\\ |\mathrm{mff}\rangle^{\mp,1}=\mathcal{MFF}^{\mp,1}(r,s+\frac{1}{2}\pm\frac{1}{2},k)|C(n,j^{\pm}_{rc}(r,s,n,k),k)\rangle, & n\geqslant 0. \end{cases} \tag{II.61}$$

Remark II.15: When the intersection $\mathcal{C}'=\mathcal{C}\cap\mathcal{R}'$ is generated by the singular vector $\mathcal{MFF}^{-}(r,s,k)|C(n,j^{-}_{rc}(r,s,n,k),k)\rangle$ with $n<0$ and $|n|\leqslant r(s+1)$, the relaxed singular vector $|\Sigma^{-}(r,s,j^{-}_{rc}(r,s,n,k),k)\rangle$ belongs to the Verma submodule $\mathcal{C}'$ and thus cannot generate the entire relaxed submodule $\mathcal{R}$:

$$|\Sigma^{-}(r,s,j^{-}_{rc}(r,s,n,k),k)\rangle=(J^{-}_{0})^{r+rs+n}|c\rangle. \tag{II.62}$$

The relative positions of Σ^{-} and Σ^{+} in the extremal diagram are as follows:[4]

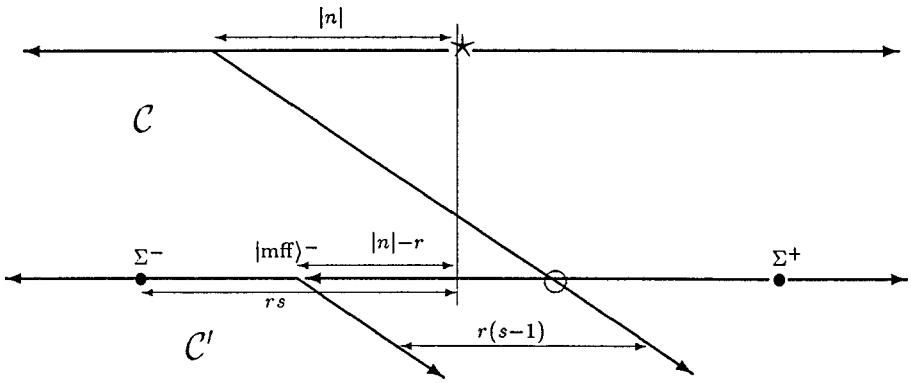

(II.63)

where length $|n|-r$ is shown as if positive. (At the point marked with a circle, two lines do not ''intersect,'' i.e., they lead to different vectors in the same grade. The same applies to the diagrams below, which will not be stated explicitly however.) Under the conditions of the Theorem, Σ^+ still has infinitely many descendants with respect to both J_0^- and J_0^+.

Going ''past the highest-weight'': The above concludes the analysis of codimension 2 cases. In the subsequent analysis of codimension 3, however, it will be very useful to know more subtle properties of the chosen representatives of singular vectors. As we saw in (II.63), descendants of Σ^- may not cover a part of the extremal subdiagram because of the *Verma* highest-weight conditions satisfied at $|\mathrm{mff}\rangle^-$. In applications, however, one needs to know whether or not a given *Verma* submodule in a relaxed Verma module is actually generated from a charged singular vector in some relaxed Verma submodule. (Both cases can occur, and the distinction between the two possibilities is crucial for the correct classification of degenerate cases, and, also, for the construction of embedding diagrams, etc.)

Thus, given a Verma highest-weight state $|X\rangle$, which satisfies $J_0^+|X\rangle=0$, we are interested in whether there exists a state $|Y\rangle$ *in the relaxed Verma module* such that $|X\rangle=J_0^-|Y\rangle$, or, to put it differently, whether $(J_0^-)^{-1}|X\rangle$ can be defined as an element of the relaxed Verma module.

We now answer this question for vectors (II.45) and (II.46). First of all, whenever Σ^- is in the Verma submodule $\mathcal{C}'$ generated by $\mathrm{MFF}^-(r,s,k)$, we consider instead of Σ^- the highest-weight vector of that submodule:

$$|\hat{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle\equiv|\mathrm{mff}\rangle^-=\mathcal{MFF}^-(r,s,k)|j_{rc}^-(r,s,n,k),\Lambda_{rc}^-(r,s,n,k),k|n\rangle_{sl(2)}, \tag{II.64}$$

of which Σ^- is a J_0^--descendant.

The vector $|\hat{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle$ is the farthest right one can reach by acting with J_0^+ on the original singular vector $|\Sigma^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle$. To map it farther right, to the outside of the Verma submodule $\mathcal{C}'$, one employs the operator $(J_0^-)^{-1}$, which is understood as one of the ''continued'' operators from (II.11)–(II.12):

$$|\tilde{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle=(J_0^-)^{-1}|\hat{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle. \tag{II.65}$$

This can be pictured as

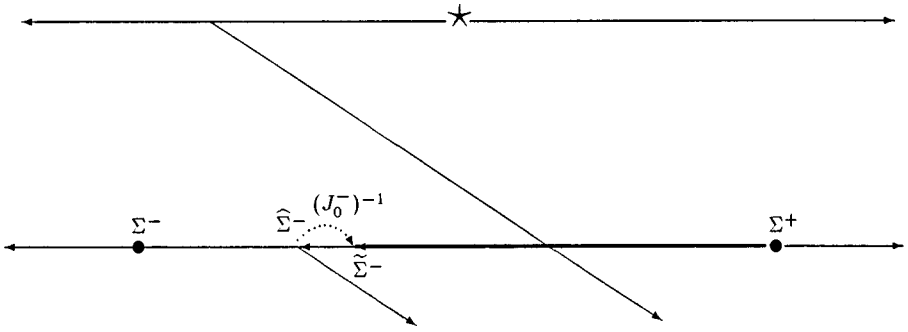

(II.66)

The conditions for $|\tilde{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle$ to exist are given by the following lemma. Let

$$f(r,s,n,k)=\begin{cases}\displaystyle\prod_{m=0}^{2r-n}(r-m-n-(k+2)s), & 1\leqslant n\leqslant 2r,\\ 1, & n\geqslant 2r+1.\end{cases} \tag{II.67}$$

Lemma II.16: In the relaxed Verma module $\mathcal{R}_{j_{rc}^-(r,s,n,k),\Lambda_{rc}^-(r,s,n,k),k}$ *with* $n<0$ *and* $|n|\leqslant r(s+1)$,

(I) The vector

$$|\tilde{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle=(J_0^-)^{-1}\mathcal{MFF}^-(r,s,k)|j_{rc}^-(r,s,n,k),\Lambda_{rc}^-(r,s,n,k),k|n\rangle_{sl(2)}, \tag{II.68}$$

exists if and only if $f(r,s,|n|,k)\neq 0$, *in which case it is a relaxed singular vector.*

(II) *Under the conditions of Theorem II.14, a* J_0^+*-descendant of* $|\Sigma^+(r,s,j_{rc}^-(r,s,n,k),k)\rangle$ *is proportional to vector (II.68):*

$$(J_0^-)^{rs-n-r-1}|\Sigma^+(r,s,j_{rc}^-(r,s,n,k),k)\rangle=(-1)^r a(r,s,n,k)|\tilde{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle, \tag{II.69}$$

where

$$\begin{aligned} a(r,s,n,k)&=\frac{\Pi_{t=0}^{r(s+1)-1}(\Lambda_{rc}^-(r,s,n,k)-i(i+1)-2ij_{rc}^-(r,s,n,k))}{\Pi_{i=0}^{2r-1}(n-i-r(s-1))}\\ &=\prod_{i=0}^{r(s+1)-1}(i-r-n+(k+2)s)\prod_{i=0}^{r(s-1)-1}(n-i). \end{aligned} \tag{II.70}$$

To prove the lemma, we observe that the action of $(J_0^-)^{-1}$ is given by (II.12) and by Eqs. (II.23); this can be rewritten as

$$\begin{aligned} (J_0^-)^{-1}|j,\Lambda,k|m\rangle_{sl(2)}&=|j,\Lambda,k|m+1\rangle_{sl(2)},\quad m\leqslant -1,\\ (J_0^{-1})^{-1}|j,\Lambda,k|m\rangle_{sl(2)}&=\frac{1}{\Lambda-m(m+1)-2mj}\,|j,\Lambda,k|m+1\rangle_{sl(2)},\quad m\geqslant 0. \end{aligned} \tag{II.71}$$

This applies unless the denominator in the last formula vanishes. A simple analysis of the relative charge of $|\hat{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle$ shows that $f(r,s,-n,k)$ is precisely the function responsible for the existence of $|\tilde{\Sigma}^-(r,s,j_{rc}^-(r,s,n,k),k)\rangle$: to evaluate (II.65), one uses Eqs. (II.12) and then, as negative powers of J_0^- reach the highest-weight state, one applies (II.71) the appropriate number of times; the relevant factors from the denominators are precisely the above $f(r,s,-n,k)$, whence the lemma follows.

Thus, whenever state (II.65) exists, it is proportional to a J_0^--descendant of Σ^+.

Codimension 3: The codimension-3 cases are more involved than the previous ones. Let us begin with the case where a further degeneration occurs in Theorem II.12.

Theorem II.17: *The highest weight of the relaxed Verma module* $\mathcal{R}_{j,\Lambda,k}$ *belongs to the set* O_{rrc} *of the list of degeneration patterns if and only if* $j=j_{rc}^{\sigma}(r,s,n,k)$ *and* $\Lambda=\Lambda_{rc}^{\sigma}(r,s,n,k)$ *as in (II.58), where*

$$(\sigma,r,s,n,k)\in(\{\pm\}\times\mathbb{N}\times\mathbb{N}\times\mathbb{Z}\times\mathbb{Q}\cup\mathbb{A}(\mathbb{Q}))\backslash\mathbb{B}\backslash\mathbb{D}, \tag{II.72}$$

with $\mathbb{A}(\mathcal{P})$ *and* $\mathbb{B}$ *as in (II.60) and*

$$\mathbb{D}=\left\{\left(\sigma,r,s,n,\frac{p}{q}\right)\middle|\, r,s,q\in\mathbb{N},\ \ p,n\in\mathbb{Z},\ \ \left|r-s\left(\frac{p}{q}+2\right)\right|\in\{|n|,|n|+1,|n|+2,\ldots\}\right\}. \tag{II.73}$$

Then we have the following.

(1) $\mathcal{R}_{j,\Lambda,k}$ *contains a submodule* $\mathcal{C}$ *generated from the charged singular vector* $|C(n,j,p/q)\rangle$; *if* $n<0$, *then* $\mathcal{C}\approx\mathcal{M}_{j+n,p/q}$, *where* $\mathcal{M}_{j+n,p/q}$ *is a Verma module with at least two singular vectors; if* $n\geqslant 0$, $\mathcal{C}\approx\mathcal{M}_{j+n+1+p/2q,p/q;1}$.

(2) *Any other primitive submodule in* $\mathcal{R}_{j,\Lambda,k}$ *is isomorphic either to a relaxed Verma module or, depending on whether* $n<0$ *or* $n\geqslant 0$, *to a Verma module or a twisted Verma module with the twist parameter* $\theta=1$, *respectively. Further, we have the following.*

(a) If $n<0$, *each Verma submodule* $\mathcal{C}'\subset\mathcal{R}_{j,\Lambda,k}$ *is a submodule in* $\mathcal{C}$ *and, moreover, there exists a relaxed Verma submodule* $\mathcal{R}'$ *such that* $\mathcal{C}'$ *is generated from a charged singular vector in* $\mathcal{R}'$.

(b) If $n\geqslant 0$, *each twisted Verma submodule* $\mathcal{C}'\subset\mathcal{R}_{j,\Lambda,k}$ *is a submodule in* $\mathcal{C}$ *and, moreover, there exists a relaxed Verma submodule* $\mathcal{R}'$ *such that* $\mathcal{C}'$ *is generated from a charged singular vector in* $\mathcal{R}'$.

(c) For any relaxed Verma submodule $\mathcal{R}'\subset\mathcal{R}_{j,\Lambda,k}$, the intersection $\mathcal{C}'=\mathcal{R}'\cap\mathcal{C}$ is nonempty. If, then, $|v\rangle$ is the highest-weight vector of $\mathcal{C}'$, there exists a vector $\widetilde{|v\rangle}\in\mathcal{R}_{j,\Lambda,k}$, defined as

$$\widetilde{|v\rangle}=\begin{cases}(J_0^+)^{-1}|v\rangle, & n\geqslant 0,\\ (J_0^-)^{-1}|v\rangle, & n<0,\end{cases} \tag{II.74}$$

that satisfies relaxed highest-weight conditions and generates the relaxed Verma submodule $\mathcal{R}'$.

(3) *For any primitive submodule $\mathcal{C}'\subset\mathcal{C}$, there exists a relaxed Verma submodule $\mathcal{R}'$ such that the following occurs.*

(a) $\mathcal{C}'\subset\mathcal{R}'\cap\mathcal{C}$.

(b) $\mathcal{R}'$ is generated from a vector $|\tilde v\rangle$, such that the vector

$$|v\rangle=\begin{cases}J_0^+|\tilde v\rangle, & n\geqslant 0,\\ J_0^-|\tilde v\rangle, & n<0,\end{cases} \tag{II.75}$$

generates a Verma module (if $n<0$) or a twisted Verma module (if $n\geqslant 0$) that either coincides with $\mathcal{C}'$ or contains $\mathcal{C}'$ as a submodule generated from the singular vector $|\mathrm{MFF}^{+}(a,1,k)\rangle$ (if $n<0$) or $|\mathrm{MFF}^{-,1}(a,1,k)\rangle$ (if $n\geqslant 0$), where $a\in\mathbb{N}$.

Whenever $s=0$, we have $\mathcal{R}'=\mathcal{R}_{j,\Lambda,k}$.

Thus, the structure of $\mathcal{R}_{j,\Lambda,k}$ is completely determined by the structure of its Verma submodule $\mathcal{C}=\mathcal{M}_{\mathfrak{j}-(r,s,p/q),p/q}\subset\mathcal{R}_{j,\Lambda,k}$ (if $n<0$) or $\mathcal{C}=\mathfrak{M}_{\mathfrak{j}^+(r,s,p/q),p/q;1}\subset\mathcal{R}_{j,\Lambda,k}$ (if $n\geqslant 0$).

The case O_{rcc} of the list of possible degenerations is the most involved one, and we describe it in the following theorem.

Theorem II.18: *The highest weight of the module $\mathcal{R}_{j,\Lambda,k}$ belongs to the set O_{rcc} of the list of degeneration patterns if and only if*

$$j=j_{cc}(n,m),\quad \Lambda=\Lambda_{cc}(n,m),\quad -n,m+1\in\mathbb{N},\quad k+2=\frac{r\pm(m-n)}{s},\quad r,s\in\mathbb{N}. \tag{II.76}$$

In this case, the module $\mathcal{R}_{j,\Lambda,k}$ contains a Verma submodule $\mathcal{C}_-$ generated from the charged singular vector $|C(n,j_{cc}(n,m),k)\rangle$ and a Verma submodule $\mathcal{C}_+$ (with the twist parameter $\theta=1$) generated from the charged singular vector $|C(m,j_{cc}(n,m),k)\rangle$. Each of the modules $\mathcal{C}_-$ and $\mathcal{C}_+$ admits at least one singular vector. Moreover, a singular vector $|\mathrm{MFF}^{\pm}(a,b,k)\rangle$ exists in the module $\mathcal{C}_-$ if and only if the singular vector $|\mathrm{MFF}^{\mp,1}(a,b,k)\rangle$ with the same $a,b\in\mathbb{N}$ exists in $\mathcal{C}_+$.

Any other primitive submodule $\mathcal{N}\subset\mathcal{R}_{j,\Lambda,k}$ is one of the following: (1) $\mathcal{N}$ is a Verma submodule, such that $\mathcal{N}\subset\mathcal{C}_-$; (2) $\mathcal{N}$ is a twisted Verma submodule with the twist parameter 1, such that $\mathcal{N}\subset\mathcal{C}_+$; (3) $\mathcal{N}$ is a relaxed Verma module, in which case the intersections $\mathcal{N}\cap\mathcal{C}_-=\mathcal{C}'_-$ and $\mathcal{N}\cap\mathcal{C}_+=\mathcal{C}'_+$ are nonempty. Whenever $\mathcal{C}'_+$ is generated from a singular vector $|\mathrm{MFF}^{\pm}(a,b,k)\rangle$, $\mathcal{C}'_-$ is generated from singular vector $|\mathrm{MFF}^{\mp}(a,b,k)\rangle$ with the same $a,b\in\mathbb{N}$ (and the above k).

Next, we give explicit constructions of singular vectors that generate submodules described in Theorem II.18.

Theorem II.19: *Under the conditions of Theorem II.18, we have the following.*

(1) *Whenever $k+2=(m-n+r)/s$, the module $\mathcal{R}_{j,\Lambda,k}$ contains a relaxed Verma submodule $\mathcal{R}'$ such that its intersections with $\mathcal{C}_-$ and $\mathcal{C}_+$ are generated from the following vectors:*

$$\begin{aligned}|\mathrm{mff}\rangle^{+}&=\mathcal{MFF}^{+}(r,s+1,k)(J_0^-)^{-n}|j_{cc}(n,m),\Lambda_{cc}(n,m),k\rangle_{sl(2)},\\ |\mathrm{mff}\rangle^{-,1}&=\mathcal{MFF}^{-,1}(r,s+1,k)(J_0^+)^{m+1}|j_{cc}(n,m),\Lambda_{cc}(n,m),k\rangle_{sl(2)},\end{aligned} \tag{II.77}$$

respectively, and $\mathcal{R}'$ itself is generated from any of the states,

$$(J_0^-)^{-1}|\mathrm{mff}\rangle^{+}\quad or\quad (J_0^+)^{-1}|\mathrm{mff}\rangle^{-,1}. \tag{II.78}$$

The states $|\mathrm{mff}\rangle^{+}$ *and* $|\mathrm{mff}\rangle^{-,1}$ *are charged singular vectors in* $\mathcal{R}'$*; after taking the quotient of* $\mathcal{R}'$ *with respect to* $|\mathrm{mff}\rangle^{+}$ *and* $|\mathrm{mff}\rangle^{-,1}$*, the states remaining on the top level furnish an irreducible representation of the Lie algebra* $sl(2)$ *of dimension* $m-n+2r$.

(2) *Whenever* $k+2=(-m+n+r)/s$ *and* $2r-m+n\leqslant -1$*, the module* $\mathcal{R}_{j,\Lambda,k}$ *contains a relaxed Verma submodule* $\mathcal{R}'$ *such that its intersections with* $\mathcal{C}_-$ *and* $\mathcal{C}_+$ *are generated from the vectors*

$$|\mathrm{mff}\rangle^{-}=\mathcal{MFF}^{-}(r,s,k)(J_0^-)^{-n}|j_{cc}(n,m),\Lambda_{cc}(n,m),k\rangle_{sl(2)},$$
$$|\mathrm{mff}\rangle^{+,1}=\mathcal{MFF}^{+,1}(r,s,k)(J_0^+)^{m+1}|j_{cc}(n,m),\Lambda_{cc}(n,m),k\rangle_{sl(2)}, \tag{II.79}$$

respectively, and $\mathcal{R}'$ *itself is generated from any of the states,*

$$(J_0^-)^{-1}|\mathrm{mff}\rangle^{-} \quad \textit{or} \quad (J_0^+)^{-1}|\mathrm{mff}\rangle^{+,1}. \tag{II.80}$$

The vectors $|\mathrm{mff}\rangle^{-}$ *and* $|\mathrm{mff}\rangle^{+,1}$ *are charged singular vectors in* $\mathcal{R}'$*; after taking the quotient with respect to* $|\mathrm{mff}\rangle^{+}$ *and* $|\mathrm{mff}\rangle^{-,1}$*, the states remaining on the top level of* $\mathcal{R}'$ *furnish an irreducible representation of the Lie algebra* $sl(2)$ *of dimension* $m-n-2r$.

(3) *Whenever* $k+2=(-m+n+r)/s$ *and* $2r-m+n\geqslant 0$*, the module* $\mathcal{R}_{j,\Lambda,k}$ *contains submodules* $\mathcal{C}'_-\subset\mathcal{C}_-$ *and* $\mathcal{C}'_+\subset\mathcal{C}_+$ *that are generated from singular vectors (II.79) respectively. There does not exist a relaxed Verma submodule* $\mathcal{R}'\subset\mathcal{R}_{j,\Lambda,k}$ *in which either* $|\mathrm{mff}\rangle^{-}$ *or* $|\mathrm{mff}\rangle^{+,1}$ *would be a charged singular vector.*

Moreover, whenever $2r-m+n\geqslant 1$*, the states* $(J_0^+)^i|\mathrm{mff}\rangle^{+,1}$ *and* $(J_0^-)^{2r-m+n-1-i}|\mathrm{mff}\rangle^{-}$*,* $i=0,\ldots,2r-m+n-1$*, satisfy the relaxed highest-weight conditions, are in the same grade for each* i *and are linearly independent. The module* $\mathcal{C}'_-$ *contains the singular vector* $|\mathrm{MFF}^{+}(2r-m+n,1,(-m+n+r)/s-2)\rangle$*, while the module* $\mathcal{C}'_+$ *contains the singular vector* $|\mathrm{MFF}^{-,1}(2r-m+n,1,(-m+n+r)/s-2)\rangle$*. After taking the quotients with respect to the submodules generated from these two singular vectors, the states remaining in the top level of the submodule* $\mathcal{C}'_-\cup\mathcal{C}'_+$ *arrange into a direct sum of two irreducible representations of the* $sl(2)$ *Lie algebra of the dimension* $2r-m+n$ *each.*

Parts 1 and 2 of Theorem II.19 follow from Lemma II.16. This can be illustrated as

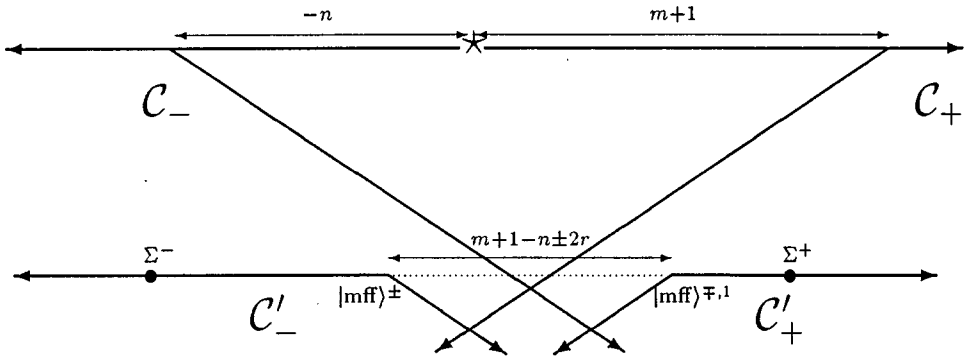

(II.81)

where the upper signs are to be taken in case 1 of the theorem, and the lower signs, in case 2.

In case 3 of the theorem, the appearance of singular vectors $|\mathrm{MFF}^{+}(2r-m+n,1,(n+r-m)/s-2)\rangle$ and $|\mathrm{MFF}^{-,1}(2r-m+n,1,(n+r-m)/s-2)\rangle$ in the modules $\mathcal{C}'_-$ and $\mathcal{C}'_+$, respectively, follows by a direct analysis of the relevant highest weights. This can be illustrated by the extremal diagram,

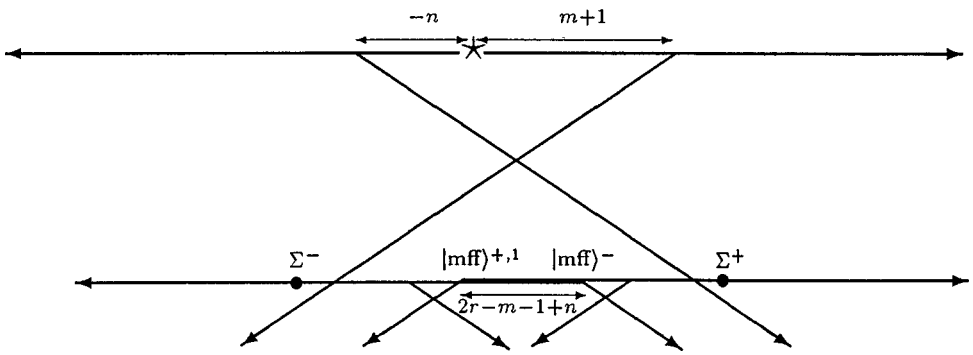

(II.82)

where $|\mathrm{mff}\rangle^{+,1}$ lies on the upper of the two overlapping lines, and $|\mathrm{mff}\rangle^{-}$, on the lower one.

Remark II.20: As we see from the theorem, whenever two different singular vectors appear in the same grade, these are necessarily descendants of singular vectors in the corresponding (twisted) *Verma* submodules. Thus, understanding by the relaxed singular vector proper the entire floor in the extremal diagram, we can say that two different relaxed singular vectors in the same grade never occur.

This concludes the list of our constructions on the $\widehat{sl}(2)$ side. The above formulation was given in such a way as to facilitate a comparison with the $N=2$ superconformal algebra; as we will see later, all of the above objects and constructions have appropriate $N=2$ analogs.

B. Auxiliaries

We now introduce two free-field systems that will be needed in Sec. III.

1. Fermions (ghosts)

We will need a fermionic system (the so-called ghosts, or a bc system), defined in terms of the operator products as

$$B(z)C(w)=\frac{1}{z-w}, \tag{II.83}$$

with the energy–momentum tensor,

$$T^{\mathrm{GH}}=-B\partial C. \tag{II.84}$$

We use the mode expansions $B(z)=\Sigma_{n\in\mathbb{Z}}B_n z^{-n-1}$ and $C(z)=\Sigma_{n\in\mathbb{Z}}C_n z^{-n}$.

Denote by Ω the module generated from the vacuum $|0\rangle_{\mathrm{GH}}$ defined by the conditions

$$C_{\geqslant 1}|0\rangle_{\mathrm{GH}}=B_{\leqslant 0}|0\rangle_{\mathrm{GH}}=0. \tag{II.85}$$

The thus defined vacuum is an sl_2-invariant state;[23] along with this one, we can choose other highest-weight states, $|\lambda\rangle_{\mathrm{GH}}$, which belong to the same module and are determined by

$$C_{\geqslant 1-\lambda}|\lambda\rangle_{\mathrm{GH}}=B_{\geqslant\lambda}|\lambda\rangle_{\mathrm{GH}}=0. \tag{II.86}$$

Then, as is easy to see,

$$(BC)_0|\lambda\rangle_{\mathrm{GH}}=-\lambda|\lambda\rangle_{\mathrm{GH}},$$

$$L_0^{\mathrm{GH}}|\lambda\rangle_{\mathrm{GH}}=(1/2)\lambda(\lambda-1)|\lambda\rangle_{\mathrm{GH}}. \tag{II.87}$$

The character of Ω is given by

$$\chi^{\mathrm{GH}}(z,q)=\mathrm{Tr}_{\Omega}(q^{L_0^{\mathrm{GH}}}z^{-(BC)_0})=\prod_{i=0}^{\infty}(1+zq^i)\prod_{i=1}^{\infty}(1+z^{-1}q^i). \tag{II.88}$$

We can connect the states $|\lambda\rangle_{\mathrm{GH}}$ with different λ by means of operators $c(\mu,\nu)$ and $b(\mu,\nu)$, which are the products of fermionic modes,

$$c(\mu,\nu)=\prod_{n=1}^{\nu-\mu+1} C_{\mu+n}, \quad b(\mu,\nu)=\prod_{n=1}^{\nu-\mu+1} B_{\nu+n}, \quad \nu-\mu+1\in\mathbb{N}. \tag{II.89}$$

These $c(\mu,\nu)$ and $b(\mu,\nu)$ map a vector $|\lambda\rangle_{\mathrm{GH}}$ as follows:

$$c(-\lambda-l+1,-\lambda):|\lambda\rangle_{\mathrm{GH}}\mapsto|\lambda+l\rangle_{\mathrm{GH}}, \quad b(-\lambda-l,\lambda-1):|\lambda\rangle_{\mathrm{GH}}\mapsto|\lambda-l\rangle_{\mathrm{GH}}, \quad l\in\mathbb{N}. \tag{II.90}$$

All the necessary properties of the objects $c(\mu,\nu)$ and $b(\mu,\nu)$ can be derived when $\nu-\mu+1$ is a positive integer and then generalized to the case of arbitrary μ and ν. For example,

$$c(\mu,\nu-1)c(\nu,\nu_1)=c(\mu,\nu_1), \quad b(\mu,\nu-1)b(\nu,\nu_1)=b(\mu,\nu_1). \tag{II.91}$$

2. The ''Liouville'' scalar

In this section we introduce a ''Liouville'' scalar that will be used to invert the KS mapping. This is just a free scalar, called ''Liouville'' for the signature of its operator product:

$$\phi(z)\phi(w)=-\ln(z-w). \tag{II.92}$$

We define vertex operators $\psi=e^{\phi}$ and $\psi^*=e^{-\phi}$, with the mode expansions

$$\psi(z)=\sum_{n=-\infty}^{\infty}\frac{\psi_n}{z^n}, \quad \psi^*(z)=\sum_{n=-\infty}^{\infty}\frac{\psi_n^*}{z^{n-1}}, \tag{II.93}$$

and

$$\partial\phi(z)=\sum_{n=-\infty}^{\infty}\frac{\phi_n}{z^{n+1}}, \quad \text{with } [\phi_n,\phi_m]=-n\,\delta_{n+m,0}. \tag{II.94}$$

We take the energy–momentum tensor to be

$$T_{\phi}=-\tfrac{1}{2}\,\partial\phi\partial\phi+\tfrac{1}{2}\,\partial^2\phi: \tag{II.95}$$

then ψ and ψ^* have the dimensions 0 and -1, respectively.

The ψ and ψ^* operators are naturally represented as intertwining operators on a direct sum,

$$\Xi=\bigoplus_{n\in\mathbb{Z}}\mathcal{H}_n, \tag{II.96}$$

where $\mathcal{H}_n$ is a Verma module over the Heisenberg algebra (II.94) with the highest-weight vector $|n\rangle_{\phi}$,

$$\begin{aligned}&\phi_m|n\rangle_{\phi}=0, \quad m\geqslant 1, \quad \psi_m|n\rangle_{\phi}=0, \quad m\geqslant n+1,\\ &\psi_m^*|n\rangle_{\phi}=0, \quad m\geqslant -n+2, \quad \phi_0|n\rangle_{\phi}=-n|n\rangle_{\phi}.\end{aligned} \tag{II.97}$$

Note also that the highest-weight vectors $|n\rangle_{\phi}$ can be realized as the operators $|n\rangle_{\phi}\sim e^{n\phi}(0)$.

C. $N=2$

In the description of our second main ingredient, the $N=2$ algebra, we follow Refs. 9 and 10, particularly as regards the singular vectors. Other (nonexhaustive) references on the $N=2$ superconformal algebra are Refs. 3, 12, 15, 16, and 24.

1. The $N=2$ superconformal algebra

The $N=2$ superconformal algebra is spanned by Virasoro generators L_m, two fermionic fields $G_r^{\pm}$, and a $U(1)$ current H_n (and a central element C). The nonvanishing commutation relations are

$$[L_n,L_m]=(n-m)L_{n+m}+\frac{\mathsf{C}}{12}(n^3-n)\delta_{n+m},\quad [H_n,H_m]=\frac{\mathsf{C}}{3}n\delta_{n+m},$$

$$[L_n,H_m]=-mH_{n+m},\quad [L_n,G_r^{\pm}]=(\tfrac{1}{2}n-r)G_{n+r}^{\pm},\quad [H_n,G_r^{\pm}]=\pm G_{n+r}^{\pm}, \tag{II.98}$$

$$\{G_r^{+},G_s^{-}\}=2L_{r+s}+(r-s)H_{r+s}+\frac{\mathsf{C}}{3}(r^2-\tfrac{1}{4})\delta_{r+s},\quad n,m\in\mathbb{Z},\quad r,s\in\mathbb{Z}+\tfrac{1}{2}.$$

We find it more convenient to work with the version of this algebra obtained by redefining the generators as

$$L_n=L_n|_{(\mathrm{II}.98)}+\tfrac{1}{2}(n+1)H_n,\quad G_r=G_{r+1/2}^{+},\quad Q_r=G_{r-1/2}^{-}. \tag{II.99}$$

This redefinition affects two things: the modding of the fermions and the choice of the Virasoro generators. The latter is due to the freedom of adding a derivative of the $U(1)$ current to the energy–momentum tensor, and is nothing but a change of basis in the algebra. As to the different moddings, they label different members of a family of isomorphic algebras related by the spectral flow[15,16] [see Eq. (II.106) below]. Thus, any statement found for one of these algebras does immediately translate to any other one. In order to be specific in our constructions below, we need to choose the concrete representative from that family as well as the modding of the generators. These will be

$$[L_m,L_n]=(m-n)L_{m+n},\quad [H_m,H_n]=\frac{\mathsf{C}}{3}m\delta_{m+n,0},$$

$$[L_m,G_n]=(m-n)G_{m+n},\quad [H_m,G_n]=G_{m+n},$$

$$[L_m,Q_n]=-nQ_{m+n},\quad [H_m,Q_n]=-Q_{m+n},\quad m,n\in\mathbb{Z}. \tag{II.100}$$

$$[L_m,H_n]=-nH_{m+n}+\frac{\mathsf{C}}{6}(m^2+m)\delta_{m+n,0},$$

$$\{G_m,Q_n\}=2L_{m+n}-2nH_{m+n}+\frac{\mathsf{C}}{3}(m^2+m)\delta_{m+n,0},$$

Denote this algebra as $\mathcal{A}$. Its generators L_m, Q_m, H_m, and G_m will be called the Virasoro generators, the ''BRST'' current, the $U(1)$ current, and the spin-2 fermionic current, respectively. [$Q(z)$ *is* the BRST current when the algebra (II.100) is realized on the bosonic string world sheet.[25,26]]

As a Lie superalgebra, $\mathcal{A}$ is generated by the following set of elements:

$$e_1=\tfrac{1}{2}Q_0,\quad e_2=G_1,\quad e_3=H_1,$$

$$h_1=L_0,\quad h_2=L_0+H_0+\frac{C}{3},\quad h_3=\frac{\mathsf{C}}{3}, \tag{II.101}$$

$$f_1=G_0,\quad f_2=\tfrac{1}{2}Q_{-1},\quad f_3=H_{-1}.$$

The elements $h_i, i=1,2,3$ generate the Cartan subalgebra $\mathfrak{h}$.

Now we define the *topological Verma modules* over the $N=2$ algebra [The name is inherited from the noncritical bosonic string, where matter vertices can be dressed into $N=2$ primaries that

satisfy the highest-weight conditions (II.102); in that context, the algebra (II.100) is viewed as a topological algebra.]; the highest-weight vector in any such module satisfies the annihilation and eigenvalue conditions,

$$Q_{\geqslant 0}|h,t\rangle_{\rm top}=G_{\geqslant 0}|h,t\rangle_{\rm top}=L_{\geqslant 1}|h,t\rangle_{\rm top}=H_{\geqslant 1}|h,t\rangle_{\rm top}=0, \tag{II.102}$$

$$H_0|h,t\rangle_{\rm top}=h|h,t\rangle_{\rm top}, \quad L_0|h,t\rangle_{\rm top}=0, \quad \mathsf{C}|h,t\rangle_{\rm top}=\frac{3(t-2)}{t}\,|h,t\rangle_{\rm top} \tag{II.103}$$

(the last equation being simply a parametrization of the central charge in terms of a new parameter, t; later this parameter will be identified with $k+2$, where k is the $\widehat{sl}(2)$ level). Accordingly, the topological Verma module $\mathcal{V}_{h,t}$ is freely generated by

$$L_{-m}, \quad m\in\mathbb{N}, \quad H_{-m}, \quad m\in\mathbb{N}, \quad Q_{-m}, \quad m\in\mathbb{N}, \quad G_{-m}, \quad m\in\mathbb{N}, \tag{II.104}$$

from the *topological highest-weight vector* $|h,t\rangle_{\rm top}$.

The extremal diagram of a topological Verma module is of the form

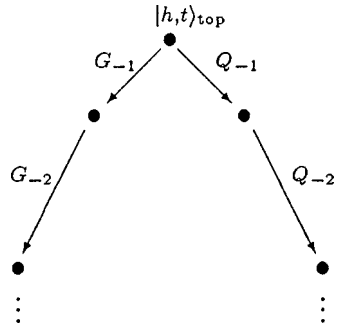

(II.105)

where the bigrading in the plane of the diagram is by (charge, level).

2. Automorphisms of the N=2 algebra

Spectral flow transform: When applied to the generators of the algebra $\mathcal{A}$, Eqs. (II.100), the spectral flow transformation $\mathcal{U}_\theta$ acts as

$$L_n\mapsto L_n+\theta H_n+\frac{\mathsf{C}}{6}\,(\theta^2+\theta)\,\delta_{n,0}, \quad H_n\mapsto H_n+\frac{\mathsf{C}}{3}\,\theta\delta_{n,0},$$
$$Q_n\mapsto Q_{n-\theta}, \quad G_n\mapsto G_{n+\theta}. \tag{II.106}$$

This gives an isomorphic algebra $\mathcal{A}_\theta$, whose generators L_n^θ, Q_n^θ, H_n^θ, and G_n^θ can be taken as the rhs's of (II.106). For $\theta\in\mathbb{Z}$, the spectral flow is an *auto*morphism of the algebra $\mathcal{A}$. For arbitrary θ, we obtain a family of algebras such that any two of its members, $\mathcal{A}_{\theta_1}$ and $\mathcal{A}_{\theta_2}$, are related by an isomorphism of the type of (II.106) with $\theta=\theta_1-\theta_2$. This family includes the Neveu–Schwarz and Ramond $N=2$ algebras, as well as the algebras in which the fermion modes range over $\pm\theta+\mathbb{Z}$, $\theta\in\mathbb{C}$.

The involutive automorphism: The $N=2$ algebra has the involutive automorphism

$$G_n\to Q_{n+1}, \quad Q_n\to G_{n-1},$$
$$H_n\to -H_n, \quad L_n\to L_n-(n+1)H_n. \tag{II.107}$$

Together with automorphisms (II.106) for $\theta\in\mathbb{Z}$, transformations (II.107) span the group $\mathbb{Z}_2\ltimes\mathbb{Z}$ isomorphic to the group generated by the canonical involution and the spectral flow of the affine $sl(2)$ algebra.

3. Twisted topological Verma modules

We now define *the twisted topological Verma modules* $\mathfrak{V}_{h,t;\theta}$. This will be a module over the algebra $\mathcal{A}_\theta$ [see (II.106)], with the *twisted topological highest-weight vector* $|h,t;\theta\rangle_{\mathrm{top}}$ defined by

$$\begin{array}{llll} L_m|h,t;\theta\rangle_{\mathrm{top}}=0, & m\geqslant 1, & Q_\lambda|h,t;\theta\rangle_{\mathrm{top}}=0, & \lambda\in-\theta+\mathbb{N}_0, \\ H_m|h,l;\theta\rangle_{\mathrm{top}}=0, & m\geqslant 1, & G_\nu|h,t;\theta\rangle_{\mathrm{top}}=0, & \nu\in\theta+\mathbb{N}_0, \end{array} \qquad \theta\in\mathbb{Z}; \tag{II.108}$$

in addition, the ''Cartan'' generators act on the state $|h,t;\theta\rangle_{\mathrm{top}}$ as

$$\left(H_0+\frac{\mathsf{C}}{3}\,\theta\right)|h,t;\theta\rangle_{\mathrm{top}}=h|h,t;\theta\rangle_{\mathrm{top}},$$

$$\left(L_0+\theta H_0+\frac{\mathsf{C}}{6}(\theta^2+\theta)\right)|h,t;\theta\rangle_{\mathrm{top}}=0, \tag{II.109}$$

$$\mathsf{C}|h,t;\theta\rangle_{\mathrm{top}}=\frac{3(t-2)}{t}\,|h,t;\theta\rangle_{\mathrm{top}}.$$

The $\theta=0$ case of (II.108) describes the ''ordinary'' topological Verma modules introduced above.

Next, we need the concept of terminating fermionic chains.

Definition II.21: Let F denote either Q or G (the fermionic $N=2$ generators), and $|X\rangle$ be an element of a module over the $N=2$ algebra. Fix also an integer n. We say that the fermionic F chain terminates on $|X\rangle$, and write ... $F_{n-3}F_{n-2}F_{n-1}F_n|X\rangle=0$ if

$$\exists N\in\mathbb{Z},\quad N\leqslant n:\ F_NF_{N+1}\cdots F_n|X\rangle=0.$$

Now we are in a position to give the following.

Definition II.22: An $N=2$ module $\mathfrak{U}$ is an object of the topological $N=2$ category $\mathcal{TOP}$ whenever the following conditions are satisfied.

(1) *H_0 and L_0 are diagonalizable operators, i.e., $\mathfrak{U}$ can be decomposed into a direct sum of intersections of H_0 and L_0 eigenspaces.*

(2) *The action with all H_p, $p\in\mathbb{N}$, and L_p, $p\in\mathbb{N}$, on any element $|X\rangle\in\mathfrak{U}$ spans a finite-dimensional space.*

(3) *For any element $|X\rangle$ of $\mathfrak{U}$,*

$$\forall n\in\mathbb{Z},\quad \begin{array}{ll} \textit{either} & \ldots Q_{n-3}Q_{n-2}Q_{n-1}Q_n|X\rangle=0, \\ \textit{or} & \ldots G_{-n-4}G_{-n-3}G_{-n-2}G_{-n-1}|X\rangle=0. \end{array} \tag{II.110}$$

Morphisms are standard homomorphisms between $N=2$ modules.

Condition (II.110) will be called the condition of terminating fermionic chains.

Note finally that the characters of twisted topological Verma modules are given by

$$\chi^{N=2}_{h,\theta}(z,q)=\mathrm{Tr}_{\mathfrak{V}_{h,t;\theta}}(q^{L_0}z^{H_0})=q^{-\theta h+\mathsf{C}(\theta^2-\theta)/6}z^{h-\mathsf{C}\theta/3}\,\frac{\prod_{i=-\theta+1}^{\infty}(1+zq^i)\prod_{i=\theta+1}^{\infty}(1+z^{-1}q^i)}{\prod_{i=1}^{\infty}(1-q^i)^2}. \tag{II.111}$$

4. Topological N=2 singular vectors

Here, we describe singular vectors in the topological Verma modules introduced above; later, we will see that these are isomorphic to the $\widehat{sl}(2)$ singular vectors, as suggested in Ref. 7. We introduce two operators $g(\mu,\nu)$ and $q(\mu,\nu)$, with $\mu,\nu\in\mathbb{C}$, that represent the action of two ''$N=2$ Weyl group'' generators r_{α_1} and r_{α_2} when μ and ν are special (see Refs. 9 and 10 for the details): the operators $g(\mu,\nu)$ and $q(\mu,\nu)$ act on the plane $t=\mathrm{const}$ as

$$g(ht+\theta-1,\theta-1):|h,t;\theta\rangle_{\rm top}\mapsto\left|\frac{2}{t}-h,t;ht+\theta-1\right\rangle_{\rm top},$$
$$q(-(h+1)t-\theta+1,-\theta-1):|h,t;\theta\rangle_{\rm top}\mapsto\left|\frac{2}{t}-2-h,t;(h+1)t+\theta-1\right\rangle_{\rm top}. \tag{II.112}$$

The action of operators $g(a,b)$ and $q(a,b)$ on the highest-weight vectors can be extended to the topological Verma module generated from these highest-weight vectors.[10] This allows us to construct singular vectors in topological Verma modules.

Theorem II.23: (I) *A topological singular vector exists in the topological Verma module* $\mathcal{V}_{h,t}$ *if and only if either* $h=\mathsf{h}^+(r,s,t)$ *or* $h=\mathsf{h}^-(r,s,t)$, *where*

$$\begin{aligned}\mathsf{h}^+(r,s,t)&=-\frac{r-1}{t}+s-1,\\ \mathsf{h}^-(r,s,t)&=\frac{r+1}{t}-s,\end{aligned}\qquad r,s\in\mathbb{N}. \tag{II.113}$$

(II) (Ref. 9) *All singular vectors in the topological Verma module* $\mathcal{V}_{\mathsf{h}^\pm(r,s,t),t}$ *are given by the explicit construction:*

$$\begin{aligned}|E(r,s,t)\rangle^+=&\,g(-r,(s-1)t-1)q((1-s)t,r-1-t)\cdots\\ &\times g((s-2)t-r,t-1)q(-t,r-1-(s-1)t)\\ &\times g((s-1)t-r,-1)|\mathsf{h}^+(r,s,t),t\rangle_{\rm top},\end{aligned} \tag{II.114}$$

$$\begin{aligned}|E(r,s,t)\rangle^-=&\,q(-r,(s-1)t-1)g((1-s)t,r-t-1)\cdots\\ &\times q((s-2)t-r,t-1)g(-t,r-1-(s-1)t)\\ &\times q((s-1)t-r,-1)|\mathsf{h}^-(r,s,t),t\rangle_{\rm top},\end{aligned} \tag{II.115}$$

where $r,s\in\mathbb{N}$ *and the factors in the first two lines of each formula are*

$$g(-r-t-mt+st,-1+mt)q(-mt,r-1+mt-st),\quad s-1\geqslant m\geqslant 1 \tag{II.116}$$

and

$$q(-r-t-mt+st,-1+mt)g(-mt,r-1+mt-st),\quad s-1\geqslant m\geqslant 1, \tag{II.117}$$

respectively.

Part (I) can be deduced from Ref. 12. Singular vectors (II.115), (II.114) satisfy highest-weight conditions (II.108) with $\theta=\mp r$, respectively.

5. Massive N=2 Verma modules

We now introduce $N=2$ Verma modules of a different type, in which the highest-weight states satisfy the following annihilation and eigenvalue conditions:

$$Q_{\geqslant 1}|h,l,t\rangle=G_{\geqslant 0}|h,l,t\rangle=L_{\geqslant 1}|h,l,t\rangle=H_{\geqslant 1}|h,l,t\rangle=0, \tag{II.118}$$

$$H_0|h,l,t\rangle=h|h,l,t\rangle,\quad L_0|h,l,t\rangle=l|h,l,t\rangle. \tag{II.119}$$

The module freely generated from $|h,l,t\rangle$ by the remaining $N=2$ generators will be denoted as $U_{h,l,t}$. Because of the property to have a dimension l generally different from zero, $|h,l,t\rangle$ are called the massive highest-weight states and $U_{h,l,t}$, accordingly, *the massive* $N=2$ *Verma modules*.

The twists of massive Verma modules are introduced by imposing the highest-weight conditions [as before, $\mathsf{c}=3(t-2)/t$]

$$\begin{aligned}&L_m|h,l,t;\theta\rangle=0,\quad m\geqslant 1,\quad Q_\lambda|h,l,t;\theta\rangle=0,\quad \lambda\in-\theta+\mathbb{N},\\ &H_m|h,l,t;\theta\rangle=0,\quad m\geqslant 1,\quad G_\nu|h,l,t;\theta\rangle=0,\quad \nu\in\theta+\mathbb{N}_0,\end{aligned} \tag{II.120}$$

$$\left(H_0+\frac{c}{3}\,\theta\right)|h,l,t;\theta\rangle=h|h,l,t;\theta\rangle,\quad \left(L_0+\theta H_0+\frac{c}{6}\,(\theta^2+\theta)\right)|h,l,t;\theta\rangle=l|h,l,t;\theta\rangle.$$

Extremal diagrams of massive Verma modules have the following ''fat'' form for $\theta=0$ (see Ref. 10 for the details)

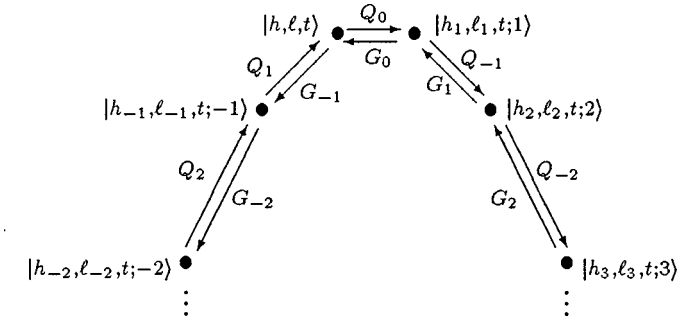

(II.121)

The states in the diagram are freely generated from $|h,l,t\rangle$ by the $G_{-1}G_{-2}\cdots$ and $Q_0Q_{-1}\cdots$ arrows, in accordance with the definition of massive Verma modules. The arrows mapping back toward $|h,l,t\rangle$ are described in the following lemma.

Lemma II.24: The mappings

$$\begin{aligned}&|h_\theta,l_\theta,t;\theta\rangle\mapsto|h_{\theta+1},l_{\theta+1},t;\theta+1\rangle,\quad \theta<0,\\ &|h_\theta,l_\theta,t;\theta\rangle\mapsto|h_{\theta-1},l_{\theta-1},t;\theta-1\rangle,\quad \theta>0\end{aligned}\tag{II.122}$$

are implemented by $Q_{-\theta}$ and $G_{\theta-1}$ respectively:

$$\begin{aligned}&Q_{-\theta}|h,l,t;\theta\rangle=2l\left|h-\frac{2}{t},l+h-\frac{2}{t},t;\theta+1\right\rangle,\quad \theta<0,\\ &G_{\theta-1}|h,l,t;\theta\rangle=2(l-h)\left|h+\frac{2}{t},l-h,t;\theta-1\right\rangle,\quad \theta>0.\end{aligned}\tag{II.123}$$

It is worth mentioning that, while $|h,l,t\rangle$ in (II.121) satisfies the untwisted highest-weight conditions [$\theta=0$ in (II.120)], the other states in the diagram satisfy the twisted massive highest-weight conditions (II.120), with θ running over all integers (from $-\infty$ to $+\infty$ as one moves from left to right).

It follows from (II.123) that, as soon as one of the factors on the right-hand sides vanishes, the respective state would satisfy the twisted *topological* highest-weight conditions (II.102). At these ''topological points,'' the extremal diagram branches, e.g.,

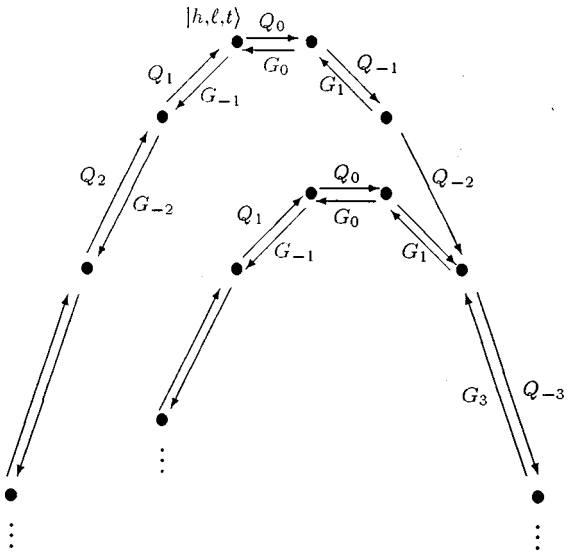

(II.124)

The inner diagram in (II.124) corresponds to an $N=2$ *subrepresentation*, which is identified with a singular vector. By iteratively applying formulas (II.123), we arrive at the following theorem.

Theorem II.25: *A massive Verma module $U_{h,l,t}$ contains a twisted topological Verma submodule if and only if $l=\mathsf{l}_{\mathrm{ch}}(r,h,t)$, where*

$$\mathsf{l}_{\mathrm{ch}}(r,h,t)=r\left(h+\frac{r-1}{t}\right),\quad r\in\mathbb{Z}; \tag{II.125}$$

The corresponding singular vector reads as

$$|E(r,h,t)\rangle_{\mathrm{ch}}=\begin{cases} Q_r\cdots Q_0|h,\mathsf{l}_{\mathrm{ch}}(r,h,t),t\rangle, & r\leqslant 0,\\ G_{-r}\cdots G_{-1}|h,\mathsf{l}_{\mathrm{ch}}(r,h,t),t\rangle, & r\geqslant 1.\end{cases} \tag{II.126}$$

It satisfies twisted topological highest-weight conditions (II.102) with $\theta=-r$.
(These are singular vectors of the ''charged'' series from Ref. 12.)

The remaining singular vectors in massive Verma modules can be constructed as follows (we only give the ''untwisted'' case; that of twisted massive modules follows by applying the spectral flow transform). As follows from the analysis of the Kač determinant,[12] the necessary condition for a massive singular vector to exist in the module $\mathcal{U}_{h,l,t}$ is $l=\mathsf{l}(r,s,h,t)$, where

$$\mathsf{l}(r,s,h,t)=\frac{1}{2}\,h-\frac{1}{2}\,rs-\frac{1}{4t}+\frac{r^2}{4t}-\frac{1}{4}\,h^2t+\frac{1}{4}\,s^2t,\quad r,s\in\mathbb{N}. \tag{II.127}$$

The necessary and sufficient condition as well as explicit formulas for the singular vectors are given in the following.

Theorem II.26: (I) (Ref. 14) For $\mathsf{c}\neq 3$, *a massive singular vector in $\mathcal{U}_{h,l,t}$ can be constructed if and only if $l=\mathsf{l}(r,s,h,t)$, where $(r,s,h,t)\in(\mathbb{N}\times\mathbb{N}\times\mathbb{C}\times\mathbb{C})\setminus\mathbb{X}$, with*

$$\mathbb{X}=\{(r,s,h,t)\,|\,st-r=m-n,ht-1=-m-n,r+st=0,n\in\mathbb{N},m\in-\mathbb{N}_0\}. \tag{II.128}$$

(II) (Ref. 10) *The massive singular vectors in $\mathcal{U}_{h,\mathsf{l}(r,s,h,t),t}$ are of the form*

$$
\begin{aligned}
|S(r,s,h,t)\rangle^- &= g(-rs,r+\theta^-(r,s,h,t)-1)\mathcal{E}^{-,\theta^-(r,s,h,t)}(r,s,t)\\
&\quad\cdot g(\theta^-(r,s,h,t),-1)|h,l(r,s,h,t),t\rangle,\\
|S(r,s,h,t)\rangle^+ &= q(1-rs,r-\theta^+(r,s,h,t)-1)\mathcal{E}^{+,\theta^+(r,s,h,t)}(r,s,t)\\
&\quad\cdot q(-\theta^+(r,s,h,t),0)|h,l(r,s,h,t),t\rangle,
\end{aligned}
\tag{II.129}
$$

where

$$
\begin{aligned}
\theta^-(r,s,h,t) &= \frac{t}{2}\,(h-\mathsf{h}^-(r,s,t)),\\
\theta^+(r,s,h,t) &= \frac{t}{2}\,(h-1-\mathsf{h}^+(r,s,t)),
\end{aligned}
\tag{II.130}
$$

and $\mathcal{E}^{\pm,\theta}(r,s,t)$ *denotes the spectral flow transform (II.106) of the topological singular vector operator* $\mathcal{E}^{\pm}(r,s,t)$. *The two expressions (II.129) in the general position belong to the same extremal diagram and generate the same massive submodule.*

This construction translates in the obvious way to the twisted case.

Degenerate cases corresponding to the simultaneous occurrence of several $N=2$ singular vectors can be described similarly to the $\widehat{sl}(2)$ case. An essential point is to consider on equal footing all states from each extremal diagram. Restricting oneself to only top-level representatives of extremal diagrams leads to additional complications in describing the structure of $N=2$ modules. In a number of degenerate cases, the entire $N=2$ submodule is *not* generated by the top-level representative of the extremal diagram of the submodule. This already happens in the case where a charged singular vector $|E(n,h,t)\rangle_{\mathrm{ch}}$ (see Theorem II.25) and a massive singular vector labeled by two integers r and s coexist in a massive $N=2$ module and $r>n$. Describing this case only in terms of top-level representatives of each of the singular vectors would lead one to assume that *sub*singular vectors exist in the module, since the chosen representative of the singular vector does not generate the entire submodule.

Let us describe in more detail the case where two charged singular vectors coexist with a massive singular vector. This is directly analogous to the case considered in part (3) of Theorem II.19:

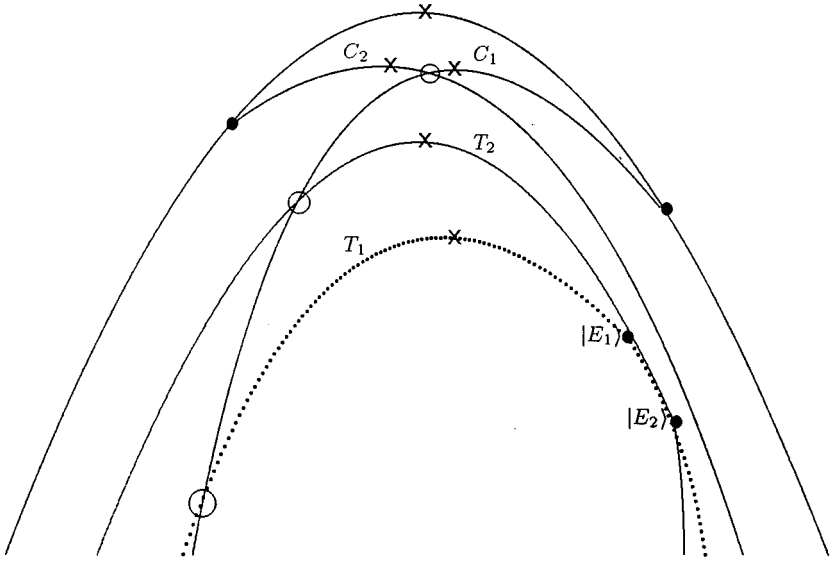

(II.131)

In this extremal diagram, the crosses denote the top-level representatives, the filled dots are twisted topological highest-weight states, and circles indicate that different curves do not intersect, i.e., correspond to linearly independent states at the overlapping points. C_1 and C_2 are extremal

diagrams of submodules built on the charged singular vectors, while T_1 and T_2 are topological submodules in the modules corresponding to C_1 and C_2, respectively. Submodules T_1 and T_2 are generated by topological singular vectors $|E_1\rangle$ and $|E_2\rangle$, respectively. The pairs of states lying on the section of the T_1 and T_2 curves between $|E_1\rangle$ and $|E_2\rangle$ give linearly independent singular vectors in the same grade. For the values of the parameters assumed in the diagram, choosing the top-level representative of T_1 and T_2 would conceal the fact that there exist linearly independent singular vectors in the same grade. By an analysis parallel to the one of Sec. II A 6, two linearly independent singular vectors in the same grade exist whenever $m-n\leqslant 2r$, where $m\geqslant 1$ and $n\leqslant -1$ label the two charged singular vectors and r is the first from the pair of integers (r,s) that label the massive singular vector. The top-level representatives would be inside the overlapping section of the two curves only when both m and $|n|$ are $\leqslant r$; however, these restrictions have nothing to do with the structure of submodules that applies whenever $m+|n|\leqslant 2r$, and is an artifact of the choice of top-level representatives. Let us note also that in the case where two linearly independent singular vectors exist in the same grade, each of these singular vectors belongs to a topological submodule; this is completely similar to the $\widehat{sl}(2)$ case discussed above [see part (3) of Theorem II.19]. For more details and a systematic account of degenerations of $N=2$ Verma modules, we refer the reader to Ref. 14.

The above condition for the fermionic chains to terminate is *not* satisfied for the (twisted) massive Verma modules. Instead, a fermionic chain terminates whenever there are two or more arrows along the same line, e.g.,

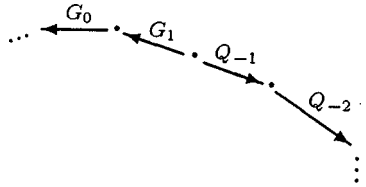

(II.132)

which is formalized in the definition of the category of (twisted) massive highest-weight-type modules.

Definition II.27: An $N=2$ module $\mathfrak{M}$ is an object of the category $\mathcal{MHW}$ whenever the following occurs.

(1) *H_0 and L_0 are diagonalizable operators, i.e., $\mathfrak{U}$ can be decomposed into a direct sum of intersections of H_0-L_0 eigenspaces.*
(2) *The action with all H_p, $p\in\mathbb{N}$, and L_p, $p\in\mathbb{N}$, on any element $|X\rangle\in\mathfrak{U}$ spans a finite-dimensional space.*
(3) *For any element $|X\rangle$ of $\mathfrak{U}$,*

$$\forall n\in\mathbb{Z},\quad \begin{matrix} either & \ldots Q_{n-3}Q_{n-2}Q_{n-1}Q_n|X\rangle=0, \\ or & \ldots G_{-n-3}G_{-n-2}G_{-n-1}G_{-n}|X\rangle=0. \end{matrix} \tag{II.133}$$

Morphisms are standard homomorphisms between $N=2$ modules.

III. KAZAMA–SUZUKI AND THE RELATED MAPPINGS

A. The Kazama–Suzuki mapping

We use the simplest KS construction associated with the coset $\widehat{sl}(2)_k\oplus u(1)/u(1)$. The KS ''numerator'' $u(1)$ algebra fermionizes into a couple of spin-1 ghosts, denoted in the following as BC, which allows us to build up the $N=2$ algebra generators in the standard way:[19,20] The odd generators Q and G are given by

$$Q=CJ^+,\quad G=\frac{2}{k+2}BJ^-, \tag{III.1}$$

while the $U(1)$ current and the energy–momentum tensor of the $N=2$ algebra take the form

$$H=\frac{k}{k+2}\,BC-\frac{2}{k+2}\,J^0, \tag{III.2}$$

$$T=\frac{1}{k+2}\,(J^+J^-)-\frac{k}{k+2}\,B\partial C-\frac{2}{k+2}\,BCJ^0. \tag{III.3}$$

Lemma III.1; The modes of currents (III.1), (III.2), and (III.3) satisfy the algebra (II.100). The eigenvalue of the central element in each representation is expressed through the eigenvalue of the $\widehat{sl}(2)$ *central element as*

$$\mathsf{c}=\frac{3k}{k+2}. \tag{III.4}$$

Thus, Eqs. (III.1), (III.2), and (III.3) determine a mapping,

$$F_{\mathrm{KS}}:\mathcal{A}\to\mathcal{U}\widehat{sl}(2)_k\otimes[BC], \tag{III.5}$$

where $\mathcal{A}$ *is the* $N=2$ *algebra (II.100) and* $\mathcal{U}$ *denotes the universal enveloping (and* $[BC]$ *is the free fermion theory).*

The KS mapping also produces a bosonic current,

$$I^+=\sqrt{\frac{2}{k+2}}\,(BC+J^0), \tag{III.6}$$

whose modes commute with the $N=2$ generators (III.1)–(III.3). Expanding as $I^+(z)=\Sigma_{n\in\mathbb{Z}}I^+_n z^{-n-1}$, we get a Heisenberg algebra,

$$[I^+_m,I^+_n]=m\,\delta_{m+n,0}. \tag{III.7}$$

The energy–momentum tensor of this $U(1)$ current is introduced as

$$T^+=\frac{1}{2}\,(I^+)^2-\frac{1}{\sqrt{2(k+2)}}\,\partial I^+. \tag{III.8}$$

We define a module $\mathcal{H}^+_p$ over the Heisenberg algebra by the highest-weight conditions

$$I_{\geqslant 1}|p\rangle^+=0,\quad I_0|p\rangle^+=p|p\rangle^+. \tag{III.9}$$

Its character is

$$\chi^+_{p,k}(z,q)=\mathrm{Tr}_{\mathcal{H}^+_p}(q^{L^+_0}z^{I^+_0})=z^p\,\frac{q^{(1/2)p(p+\sqrt{2/(k+2)})}}{\Pi_{i=1}^{\infty}(1-q^i)}. \tag{III.10}$$

Lemma III.2: Under the KS mapping, we have the identities

$$T_{\mathrm{Sug}}+T_{\mathrm{GH}}=T+T^+ \tag{III.11}$$

[where T is the energy–momentum tensor (III.3)], and

$$J^0-BC=-2H+\frac{k-2}{\sqrt{2(k+2)}}\,I^+.$$

B. The ''anti''-Kazama–Suzuki mapping

A mapping in the inverse direction to the KS mapping can be defined as follows. It is associated with the ''coset'' $\mathcal{A}\oplus u(1)/u(1)$, where the numerator $u(1)$ is represented in terms of the antifermionic system $\psi=e^{\phi}$, $\psi^*=e^{-\phi}$ from Sec. II B 2. Then we can identify the affine $sl(2)$ generators as

$$J^+=Q\psi, \quad J^-=\frac{3}{3-\mathsf{c}}\,G\psi^*,$$
$$J^0=-\frac{3}{3-\mathsf{c}}\,H+\frac{\mathsf{c}}{3-\mathsf{c}}\,\partial\phi. \tag{III.12}$$

Lemma III.3: For $\mathsf{c}\neq 3$, *generators (III.12) close to the algebra (II.2) of the level* $k=2c/(3-\mathsf{c})$, *where* c *is the* $N=2$ *central charge (the eigenvalue of the central element). Thus, Eqs. (III.12) determine a mapping,*

$$F_{\mathrm{KS}}^{-1}:\widehat{sl}(2)\to\mathcal{U}\mathcal{A}\otimes[\psi\psi^*]. \tag{III.13}$$

The Sugawara energy–momentum tensor (II.3) takes for the construction (III.12) the form

$$T^{\mathrm{Sug}}=T+\frac{1}{2}\,\partial H+\frac{k+2}{4}\,HH-\frac{k+2}{2}\,H\partial\phi+\frac{k}{4}\,\partial\phi\partial\phi. \tag{III.14}$$

We also have a free scalar, with signature -1, whose modes commute with the $\widehat{sl}(2)$ generators (III.12):

$$I^-=\sqrt{\frac{k+2}{2}}\,(H-\partial\phi). \tag{III.15}$$

The modes I_n^- introduced as $I^-(z)=\Sigma_{n=-\infty}^{\infty}I_n^-z^{-n-1}$ generate a Heisenberg algebra,

$$[I_n^-,I_m^-]=-n\,\delta_{m+n,0}. \tag{III.16}$$

This current has the energy–momentum tensor,

$$T^-=-\frac{1}{2}\,(I^-)^2-\frac{1}{\sqrt{2(k+2)}}\,\partial I^-. \tag{III.17}$$

The module $\mathcal{H}_q^-$ is defined as a Verma module over the Heisenberg algebra (III.16), with the highest-weight vector defined by

$$I_n^-|q\rangle^-=0, \quad n\geqslant 1, \quad I_0^-|q\rangle^-=q|q\rangle^-. \tag{III.18}$$

Lemma III.4: Under the anti-KS mapping, we have the identities

$$T+T_{\phi}=T^{\mathrm{Sug}}+T^- \tag{III.19}$$

[where T *is the energy–momentum tensor of the* $N=2$ *algebra and* T_{ϕ} *is the energy–momentum tensor of the Liouville system (II.95)], and*

$$-2H+\partial\phi=J^0+\left(\frac{k}{2}-1\right)\sqrt{\frac{k+2}{2}}\,I^-. \tag{III.20}$$

C. Composing the KS and anti-KS mappings

Let us consider compositions of the KS and anti-KS mappings.

Lemma III.5: The composition $F_{KS}^{-1}\circ F_{KS}$ *maps the* $\widehat{sl}(2)$ *algebra into the* $\widehat{sl}(2)$ *algebra in the tensor product* $\mathcal{U}\widehat{sl}(2)\otimes[BC]\otimes[\psi\psi^*]$ *represented by the currents*

$$\overline{J}^+ = J^+ e^{\phi} C, \quad \overline{J}^- = J^- e^{-\phi} B, \quad \overline{J}^0 = J^0 + \frac{k}{2}(\partial\phi - BC). \tag{III.21}$$

For the energy–momentum tensors, we then have

$$\overline{T}^{\mathrm{Sug}} = T^{\mathrm{Sug}} + \frac{k}{4}((\partial\phi)^2 - 2\partial\phi BC + \partial BC - B\partial C) + J^0(\partial\phi - BC). \tag{III.22}$$

A remarkable feature of these formulas is the appearance of a *null current* $BC-\partial\phi$; the fact that it is null is crucial for establishing that the generators on the rhs's close to the $\widehat{sl}(2)$ algebra.

Remark III.6: The formulas of the lemma can be given a coset-like interpretation by noticing that the image of $\widehat{sl}(2)$ under the above embedding is characterized by the fact that it commutes with (the modes of) the currents

$$I^+ = \sqrt{\frac{2}{k+2}}(J^0 + BC), \quad F^- = \sqrt{\frac{2}{k+2}}\left(J^0 - \frac{k}{2}BC + \frac{k+2}{2}\partial\phi\right) \tag{III.23}$$

(these currents do also commute with each other and are matter-like and Liouville-like, respectively).

The same happens with the $N=2$ algebra under the action of $F_{\mathrm{KS}}\circ F_{\mathrm{KS}}^{-1}$:

Lemma III.7: The composition $F_{KS}\circ F_{KS}^{-1}$ *maps the* $N=2$ *algebra into the* $N=2$ *algebra in the tensor product* $\mathcal{UA}\otimes[BC]\otimes[\psi\psi^*]$, *represented by the currents*

$$\overline{Q} = Qe^{\phi}C, \quad \overline{G} = Ge^{-\phi}B,$$

$$\overline{H} = H + \frac{k}{k+2}(BC - \partial\phi), \tag{III.24}$$

$$\overline{T} = T + H(BC - \partial\phi) + \frac{k}{2(k+2)}((\partial\phi)^2 - 2\partial\phi BC + \partial^2\phi - 2B\partial C).$$

Remark III.8: We also find two currents that commute with the ''target'' $N=2$ algebra (III.24):

$$I^- = \sqrt{\frac{k+2}{2}}(H - \partial\phi), \quad F^+ = \sqrt{\frac{k+2}{2}}\left(H - \frac{2}{k+2}BC - \frac{k}{k+2}\partial\phi\right) \tag{III.25}$$

(these currents are Liouville-like and matter-like respectively, and commute with each other).

D. KS and anti-KS mappings at the level of representations

1. The Verma module case

The following theorem is essentially a reformulation of Lemmas III.5 and III.7.

Theorem III.9: (I) *The composition of the KS and anti-KS mappings induces an isomorphism of* $\widehat{sl}(2)$ *representations,*

$$\mathfrak{M}_{j,k;\theta}\otimes\Omega\otimes\Xi \approx \bigoplus_{n,m\in\mathbb{Z}} \mathfrak{M}_{j,k;\theta+n-m}\otimes\mathcal{H}^+_{\sqrt{[2/(k+2)]}(j-(k/2)\theta-m)}\otimes\mathcal{H}^-_{-\sqrt{[2/(k+2)]}(j-(k/2)(\theta-m)-[(k+2)/2]n)}, \tag{III.26}$$

where on the lhs the $\widehat{sl}(2)$ *action is given by the rhs's of (III.21), while on the rhs the* $\widehat{sl}(2)$ *algebra acts naturally on the modules* $\mathfrak{M}_{j,k;\theta}$ *and trivially, on the Heisenberg modules.*

(II) *The composition of the anti-KS mapping and the direct KS mapping induces an isomorphism of* $N=2$ *representations,*

$$\mathfrak{V}_{h,t;\theta}\otimes\Xi\otimes\Omega\approx\bigoplus_{n,m\in\mathbb{Z}}\mathfrak{V}_{h,t;\theta+m-n}\otimes\mathcal{H}^{-}_{\sqrt{t/2}(h-[(t-2)/t]\theta+n)}\otimes\mathcal{H}^{+}_{-\sqrt{t/2}(h-[(t-2)/t](\theta-n)+(2/t)m)}, \tag{III.27}$$

where on the lhs the $N=2$ algebra acts by generators (III.24), while on the rhs we have the natural action of $N=2$ on the modules $\mathfrak{V}_{h,t;\theta}$ and the trivial action on the Heisenberg modules.

Remark III.10: The theorem remains valid when the $\widehat{sl}(2)$ or $N=2$ modules are no longer Verma modules but rather arbitrary modules from the category of highest-weight modules.

In fact, one has a stronger result, which follows from the analysis of each of the mappings F_{KS} and F_{KS}^{-1} separately.

Theorem III.11: (I) *The KS mapping induces an isomorphism of $N=2$ representations,*

$$\mathfrak{M}_{j,k;\theta}\otimes\Omega\approx\bigoplus_{m\in\mathbb{Z}}\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-(k/2)\theta-m)} \tag{III.28}$$

where on the lhs the $N=2$ algebra acts by generators (III.1)–(III.3), while on the rhs it acts naturally on the twisted topological Verma modules $\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}$.

(II) *The anti-KS mapping induces an isomorphism of $\widehat{sl}(2)$ representations,*

$$\mathfrak{V}_{h,t;\theta}\otimes\Xi\approx\bigoplus_{n\in\mathbb{Z}}\mathfrak{M}_{-(t/2)h,t-2;n-\theta}\otimes\mathcal{H}^{-}_{\sqrt{t/2}(h-(t-2)/t\theta+n)} \tag{III.29}$$

where on the lhs the $\widehat{sl}(2)$ algebra acts by generators (III.12), while on the rhs it acts naturally on the twisted Verma modules $\mathfrak{M}_{-(t/2)h,t-2;n-\theta}$.

Remark III.12: Observe that for $k\in\mathbb{Z}$, each of the formulas (III.28) and (III.29) translates the periodicity under $\theta\mapsto\theta+2$ on the $\widehat{sl}(2)$ side (which is known to hold for integrable representations) into the periodicity under $\theta\mapsto\theta+k+2$ on the $N=2$ side (which, too, is a well-known property of the corresponding $N=2$ representations and conformal models, where it is observed, for example, in the Landau–Ginzburg formulation).

Proof of the Theorem: It is easy to see that the lhs of (III.26) contains each state $|j,k\rangle_{sl(2)}\otimes|\theta\rangle_{\mathrm{GH}}$ with all $\theta\in\mathbb{Z}$. Further, we see that each state $|j,k\rangle_{sl(2)}\otimes|\theta\rangle_{\mathrm{GH}}$ satisfies the $N=2$ and Heisenberg annihilation conditions, Eqs. (II.102) and (III.9), respectively. Therefore, each state $|j,k\rangle_{sl(2)}\otimes|\theta\rangle_{\mathrm{GH}}$ can be rewritten in the form

$$|j,k\rangle_{sl(2)}\otimes|\theta\rangle_{\mathrm{GH}}=\left|\frac{-2j}{k+2},k+2;\theta\right\rangle_{\mathrm{top}}\otimes\left|\sqrt{\frac{2}{k+2}}\,(j-\theta)\right\rangle^{+},\quad\theta\in\mathbb{Z}. \tag{III.30}$$

Thus, as θ runs over the integers, every $\widehat{sl}(2)$ highest-weight state gives rise to infinitely many states satisfying twisted topological highest-weight conditions (II.108)–(II.109). From any of these states we can generate the twisted topological Verma module $\mathfrak{V}_{-2j/(k+2),k+2;\theta}$. All the states in every such module have a fixed value of the I_0^+ charge [see (III.6)], and this charge is different for modules with different θ. Thus the modules $\mathfrak{V}_{-2j/(k+2),k+2;\theta}$ do not overlap for different $\theta\in\mathbb{Z}$. Therefore, we have the mapping

$$\mathfrak{M}_{j,k;\theta}\otimes\Omega\leftarrow\bigoplus_{m\in\mathbb{Z}}\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-(k/2)\theta-m)}, \tag{III.31}$$

whose kernel may be different from zero only as a result of the vanishing of some singular vectors in the modules $\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}$. In a similar way, we arrive at the mapping

$$\mathfrak{V}_{h,t;\theta}\otimes\Xi\leftarrow\bigoplus_{n\in\mathbb{Z}}\mathfrak{M}_{-(t/2)h,t-2;n-\theta}\otimes\mathcal{H}^{-}_{\sqrt{t/2}(h-[(t-2)/t]\theta+n)}. \tag{III.32}$$

Now, we tensor (III.31) with Ξ and make use of (III.32). Note that for the modules under consideration, tensor products preserve exactness. Then the mapping (III.31) gives rise to

$$\mathfrak{M}_{j,k;\theta}\otimes\Omega\otimes\Xi\leftarrow\bigoplus_{m,n\in\mathbb{Z}}\mathfrak{M}_{j,k;\theta+n-m}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-(k/2)\theta-m)}\otimes\mathcal{H}^{-}_{-\sqrt{2/(k+2)}(j-(k/2)(\theta-m)-[(k+2)/2]n)}, \tag{III.33}$$

which is an isomorphism by Theorem III.9; hence both (III.31) and (III.32) are isomorphisms.□

Corollary 1: Singular vectors in the $\widehat{sl}(2)$ Verma modules and in the topological $N=2$ Verma modules occur (or do not occur) simultaneously. Thus, formulas (II.113) for positions of the topological singular vectors are rederived by translating the positions of the $\widehat{sl}(2)$ singular vectors under the KS mapping.

Corollary 2: Let there exist a singular vector with the J_0^0 charge equal to r in the $\widehat{sl}(2)$ Verma module on the lhs of (III.31); then each topological $N=2$ Verma module $\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}$ from the rhs of (III.31) contains a topological singular vector, which satisfies highest-weight conditions (II.108) with $\theta\to m-\theta+r$.

Proof: Let us consider the tensor product $|S(r)\rangle\otimes|\lambda\rangle_{\mathrm{GH}}$, where $|S(r)\rangle$ denotes a singular vector with the charge r in the $\widehat{sl}(2)$ Verma module from the lhs of (III.31), and $|\lambda\rangle_{\mathrm{GH}}$ is the ghost λ-vacuum (II.86). Let us determine which module $\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}$ from the rhs of (III.31) the state $|S(r)\rangle\otimes|\lambda\rangle_{\mathrm{GH}}$ belongs to. To this end, we calculate the eigenvalue of I_0^+ [see (III.6)] on the state $|S(r)\rangle\otimes|\lambda\rangle_{\mathrm{GH}}$:

$$I_0^+|S(r)\rangle\otimes|\lambda\rangle_{\mathrm{GH}}=\sqrt{\frac{2}{k+2}}\left(j+r-\frac{k}{2}-\lambda\right)|S(r)\rangle\otimes|\lambda\rangle_{\mathrm{GH}}.$$

Therefore, the state $|S(r)\rangle\otimes|m+r\rangle_{\mathrm{GH}}$ belongs to $\mathfrak{V}_{-2j/(k+2),k+2;m-\theta}$ and, as follows easily from formulas (II.86) and (III.1), satisfies highest-weight conditions (II.108) with $\theta\to m-\theta+r$. □
A constructive description of the correspondence between singular vectors is considered in Sec. IV A.

Corollary 3: The isomorphism established in the theorem implies an identity for the characters:

$$\begin{aligned}&\mathrm{Tr}_{\mathcal{M}_{j,k}\otimes\Omega}(q^{L_0^{\mathrm{Sug}}+L_0^{\mathrm{GH}}}z^{J_0^0-(BC)_0})\\&\quad=\mathrm{Tr}_{\oplus_{\theta=-\infty}^{\infty}\mathfrak{V}_{-2j/(k+2),k+2;\theta}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-\theta)}}(q^{L_0+L_0^+}z^{-2H_0+(k-2)/\sqrt{2(k+2)}I_0^+}).\end{aligned} \tag{III.34}$$

Inserting the characters, given above, of each of the modules involved, this can be rewritten as the following formal identity:

$$\begin{aligned}\frac{\prod_{i=0}^{\infty}(1+zq^i)\prod_{i=1}^{\infty}(1+z^{-1}q^i)}{\prod_{i=0}^{\infty}(1-z^{-1}q^i)\prod_{i=1}^{\infty}(1-zq^i)}&=\frac{\prod_{i=1}^{\infty}(1+z^2q^i)\prod_{i=1}^{\infty}(1+z^{-2}q^i)}{\prod_{i=1}^{\infty}(1-q^i)^2}\\&\quad\times\left(1+(1+z^2)\sum_{r=1}^{\infty}\left[\frac{z^{-r}}{1+z^2q^r}+\frac{q^rz^{r-2}}{1+z^{-2}q^r}\right]\right).\end{aligned} \tag{III.35}$$

2. The relaxed/massive Verma module case

Now, we give an extension of Theorems III.9–III.11 to the relaxed and massive Verma modules.

Theorem III.13: (I) *The composition of the KS and anti-KS mappings induces an isomorphism of* $\widehat{sl}(2)$ *representations,*

$$\mathfrak{R}_{j,\Lambda,k;\theta}\otimes\Omega\otimes\Xi\approx\bigoplus_{n,m\in\mathbb{Z}}\mathfrak{R}_{j,\Lambda,k;\theta+n-m}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-(k/2)\theta-m)}\otimes\mathcal{H}^{-}_{-\sqrt{2/(k+2)}(j-(k/2)(\theta-m)-[(k+2)/2]n)} \tag{III.36}$$

where on the lhs the $\widehat{sl}(2)$ action is given by the rhs's of (III.21), while on the rhs the $\widehat{sl}(2)$ algebra acts naturally on the modules $\mathfrak{R}_{j,\Lambda,k;\theta}$ and trivially on the Heisenberg modules.

(II) *The composition of the anti-KS mapping and the direct KS mapping induces an isomorphism of $N=2$ representations,*

$$\mathfrak{U}_{h,l,t;\theta}\otimes\Xi\otimes\Omega\approx\bigoplus_{n,m\in\mathbb{Z}}\mathfrak{U}_{h,t,l;\theta+m-n}\otimes\mathcal{H}^{-}_{\sqrt{t/2}(h-[(t-2)/t]\theta+n)}\otimes\mathcal{H}^{+}_{-\sqrt{t/2}(h-[(t-2)/t](\theta-n)+(2/t)m)}, \tag{III.37}$$

where on the lhs the $N=2$ algebra acts by generators (III.24), while on the rhs we have the natural action of $N=2$ on the modules $\mathfrak{M}_{h,l,t;\theta}$ and the trivial action on the Heisenberg modules.

This allows us to obtain, similarly to the above, the following theorem.

Theorem III.14: (I) *The KS mapping induces an isomorphism of $N=2$ representations,*

$$\mathfrak{R}_{j,\Lambda,k;\theta}\otimes\Omega\approx\bigoplus_{\lambda\in\mathbb{Z}}\mathfrak{U}_{-2j/(k+2),\Lambda/(k+2),k+2;\lambda-\theta}\otimes\mathcal{H}^{+}_{\sqrt{2/(k+2)}(j-(k/2)\theta-\lambda)}, \tag{III.38}$$

where on the lhs the $N=2$ algebra acts by generators (III.1)–(III.3), while on the rhs it acts naturally on the twisted massive Verma modules $\mathfrak{U}_{-2j/(k+2),l,k+2;\lambda-\theta}$.

(II) *The anti-KS mapping induces an isomorphism of $\widehat{sl}(2)$ representations,*

$$\mathfrak{U}_{h,l,t;\theta}\otimes\Xi\approx\bigoplus_{n\in\mathbb{Z}}\mathfrak{R}_{-(t/2)h,tl,t-2;n-\theta}\otimes\mathcal{H}^{-}_{\sqrt{t/2}(h-[(t-2)/t]\theta+n)}, \tag{III.39}$$

where on the lhs the $\widehat{sl}(2)$ algebra acts by generators (III.12), while on the rhs it acts naturally on the twisted relaxed Verma modules $\mathfrak{R}_{-(t/2)h,tl,t-2;n-\theta}$.

We could word by word repeat the proof of Theorems III.9–III.11 in the case of relaxed Verma modules. Instead, we illustrate the main points by the following diagram:

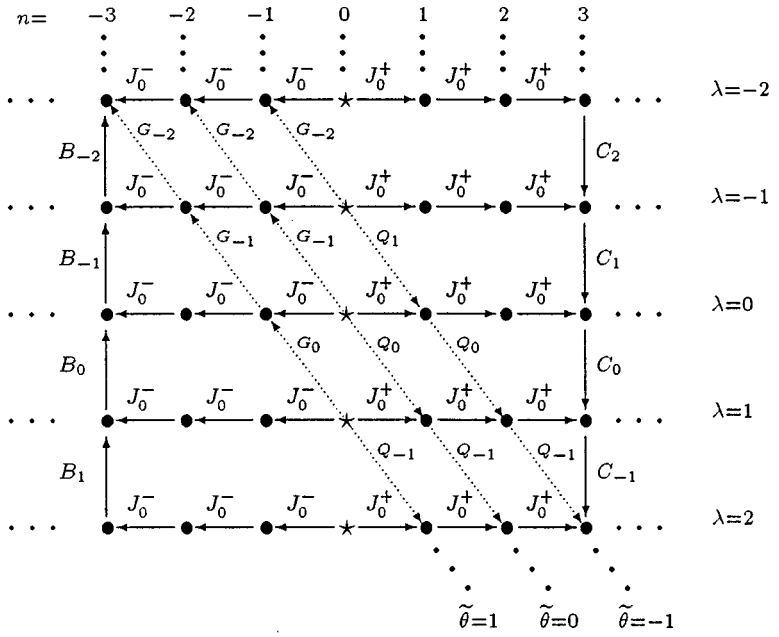

(III.40)

This describes the case of $\theta=0$ in (III.38) [which is the $\widehat{sl}(2)$ ''spectral'' parameter, not to be confused with the one labeling twisted $N=2$ modules that are also present in the diagram; the $N=2$ spectral parameter is denoted by $\tilde{\theta}$ here]. The diagram represents the tensor product of the relaxed Verma module extremal diagram (II.18) with a ghost extremal diagram. The latter consists of ghost vacua (II.86) taken in different pictures. As we have chosen $\theta=0$, the $\widehat{sl}(2)$ arrows are horizontal, as in (II.18), while the ghost ones are shown vertical. For simplicity, the ghost (B and

C) arrows are shown explicitly only in two columns. Since the different pictures (II.86) in the free-fermion system are all equivalent, all the vertical arrows are invertible, however, we have separated the B and C arrows in the diagram, trying to keep it readable. Thus the • dots denote $|j,\Lambda,k|n\rangle_{sl(2)}\otimes|\lambda\rangle_{\mathrm{GH}}$, the values of n being written in the upper row and those of λ, in the right column. The $\star$'s denote $|j,\Lambda,k\rangle_{sl(2)}\otimes|\lambda\rangle_{\mathrm{GH}}$. Now, the dotted lines are precisely the extremal diagrams (II.121) of massive Verma modules $\mathfrak{U}_{h,l,t;\tilde{\theta}}$ from the rhs of (III.38), viewed from above (the grading with respect to the level is in the direction orthogonal to the page).

Note also that the analogs of Corollaries 1 and 2 from the previous section are straightforward to formulate in the relaxed/massive case.

IV. EQUIVALENCE BETWEEN CATEGORIES

A. Equivalence of Verma chain categories

It follows from the above analysis that an equivalence between some categories of $\widehat{sl}(2)$ and $N=2$ modules can only be established for those categories that effectively allow for a factorization with respect to the spectral flow. These can be defined as follows.

Let us consider the objects that are infinite chains $(\mathfrak{M}_{j,k;\theta})_{\theta\in\mathbb{Z}}$, where $\mathfrak{M}_{j,k;\theta}$ are twisted Verma modules. As a morphism between $(\mathfrak{M}_{j,k;\theta})_{\theta\in\mathbb{Z}}$ and $(\mathfrak{M}_{j',k';\theta})_{\theta\in\mathbb{Z}}$, we take any Verma module morphisms $\mathfrak{M}_{j,k;\theta_1}\to\mathfrak{M}_{j',k';\theta_2}$. Let us call this category the $\widehat{sl}(2)$ Verma chain category $\mathcal{CVER}$.

On the $N=2$ side, the topological Verma chain category $\mathcal{CTVER}$ is defined similarly. Namely, one takes the chains of twisted topological Verma modules, every such chain comprising the modules that are the spectral flow transforms of a topological Verma module with all $\theta\in\mathbb{Z}$. Morphisms of the chains are defined similarly to the $\widehat{sl}(2)$ case. [It is important to note here that any submodule of a (twisted) topological Verma module is a twisted topological Verma module.]

The meaning of the definition of morphisms of chains is that, obviously, given a morphism between any two modules, one spreads it over the entire chains by spectral flow transforms.

To define a functor relating such chains, we first construct correspondences between individual modules in the chains.

Given a topological Verma module $\mathfrak{V}_{h,t;\theta}$, and an arbitrary $\theta'\in\mathbb{Z}$, take the Heisenberg modules $\mathcal{H}^{+}_{-\sqrt{(2/t)}((t/2)j+(t/2)\theta'-\theta)}$ and consider

$$\mathfrak{V}_{h,t;\theta}\otimes\mathcal{H}^{+}_{-\sqrt{(2/t)}((t/2)j+(t/2)\theta'-\theta)}\oplus\bigoplus_{m\in\mathbb{Z},m\neq 0}\mathfrak{V}_{h,t;\theta+m}\otimes\mathcal{H}^{+}_{-\sqrt{2/t}((t/2)j+(t/2)\theta'-\theta+m)}, \tag{IV.1}$$

where $\mathfrak{V}_{h,t;\theta+m}$ are the images of $\mathfrak{V}_{h,t;\theta}$ under the spectral flow. By Theorem III.9, (IV.1) is isomorphic to the tensor product of an $\widehat{sl}(2)$ Verma module $\mathfrak{M}_{-(t/2)h,t-2;\theta'}$ with a ghost module. We define the result of applying $F_{\mathrm{KS}}(\theta,\theta')$ to $\mathfrak{V}_{h,t;\theta}$ to be the module $\mathfrak{M}_{-(t/2)h,t-2;\theta'}$:

$$F_{\mathrm{KS}}(\theta,\theta'):\mathfrak{V}_{h,t;\theta}\rightsquigarrow\mathfrak{M}_{-(t/2)h,t-2;\theta'},\quad \theta,\theta'\in\mathbb{Z}. \tag{IV.2}$$

A similar definition can be given for the correspondence,

$$F^{-1}_{\mathrm{KS}}(\theta,\theta'):\mathfrak{M}_{j,k;\theta}\rightsquigarrow\mathfrak{V}_{-[2/(k+2)]j,k+2;\theta'},\quad \theta,\theta'\in\mathbb{Z}. \tag{IV.3}$$

Given a Verma module $\mathfrak{M}_{j,k;\theta}$ and $\theta'\in\mathbb{Z}$, we construct the sum

$$\mathfrak{M}_{j,k;\theta}\otimes\mathcal{H}^{-}_{-\sqrt{2/(k+2)}(j+\theta'-[(k+2)/2]\theta)}\oplus\bigoplus_{n\in\mathbb{Z},n\neq 0}\mathfrak{M}_{j+(k/2)n,k;\theta+n}\otimes\mathcal{H}^{-}_{-\sqrt{2/(k+2)}(j+\theta'-[(k+2)/2]\theta+n)}, \tag{IV.4}$$

which, by Theorem III.9, is isomorphic to the module $\mathfrak{V}_{-[2/(k+2)]j,k+2;\theta'}$ tensored with a module of antifermions. This twisted topological $N=2$ module is by definition the result of applying $F^{-1}_{\mathrm{KS}}(\theta,\theta')$ to $\mathfrak{M}_{j,k;\theta}$.

While the correspondences from (IV.2) and (IV.3) depend on the chosen θ and θ', the θ dependence disappears when applied to the elements of $\mathcal{CVER}$ and $\mathcal{CTVER}$; therefore $F_{\mathrm{KS}}(\cdot,\cdot)$ and $F^{-1}_{\mathrm{KS}}(\cdot,\cdot)$, which we denote again by F_{KS} and F^{-1}_{KS}, are candidates for the functors

$$F_{\mathrm{KS}} \;:\; \mathcal{CTVER} \;\rightsquigarrow\; \mathcal{CVER}$$
$$F_{\mathrm{KS}}^{-1} \;:\; \mathcal{CVER} \;\rightsquigarrow\; \mathcal{CTVER}$$

Evidently, the composition of F_{KS} and F_{KS}^{-1} maps each chain of twisted Verma modules into an isomorphic chain. Therefore, F_{KS} and F_{KS}^{-1} would be the direct and inverse functors, thus establishing the isomorphism of categories, once we define how F_{KS} and F_{KS}^{-1} act on morphisms.

Recall that in Verma-module categories, morphisms are naturally identified with singular vectors. As we see from the isomorphisms of Sec. III D, an $\widehat{sl}(2)$ singular vector exists in a twisted Verma module $\mathfrak{M}_{j,k;\theta}$ (hence in all those with $\theta \mapsto \theta+n$, $n\in\mathbb{Z}$) if and only if a topological singular vector exists in at least one (hence in each) twisted topological Verma module $\mathfrak{V}_{-2j/(k+2),k+2;m}$, $m\in\mathbb{Z}$, from the rhs of (III.28). The following lemma gives an explicit mapping between the ''building blocks'' of the corresponding singular vectors.

Lemma IV.1: The KS mapping induces a correspondence between the ''continued'' objects, $(J^-_{-\theta})^{\nu-\mu+1}$ and $(J^+_{\theta-1})^{\nu-\mu+1}$ on the one hand, and $g(\mu,\nu)$, $q(\mu,\nu)$ on the other hand, which act on the respective highest weights as shown in (II.11) and (II.112). The correspondence reads as

$$F_{\mathrm{KS}}(\theta',\theta):g(\mu,\nu)\mapsto(J^-_{-\theta})^{\nu-\mu+1}b(\mu,\nu), \qquad F_{\mathrm{KS}}(\theta',\theta):q(\mu,\nu)\mapsto(J^+_{\theta-1})^{\nu-\mu+1}c(\mu,\nu),$$
$$F_{\mathrm{KS}}^{-1}(\theta,\theta'):(J^-_{-\theta})^{\nu-\mu+1}\mapsto g(\mu,\nu)e^{-(\nu-\mu+1)\phi}, \qquad F_{\mathrm{KS}}^{-1}(\theta,\theta'):(J^+_{\theta-1})^{\nu-\mu+1}\mapsto q(\mu,\nu)e^{(\nu-\mu+1)\phi}.$$

This leads us to the following theorem.

Theorem IV.2: *The KS and anti-KS mappings give rise to identifications, which we denote again by $F_{\mathrm{KS}}(\theta,\theta')$ and $F_{\mathrm{KS}}^{-1}(\theta,\theta')$, respectively, between the twisted $\widehat{sl}(2)$ singular vectors (II.14) in the Verma module $\mathfrak{M}_{j,k;\theta}$ and the $N=2$ singular vectors (II.114) and (II.115) $|E(r,s,t)\rangle^{\pm,\theta}$ in the twisted topological Verma modules $\mathfrak{V}_{-[2/(k+2)]j,k+2;\theta}$:*

$$\begin{aligned} F_{\mathrm{KS}}(\theta,\theta'):&|E(r,s,k+2)\rangle^{\pm,\theta}\mapsto|S^{\mathrm{MFF}}_{\pm}(r,s,k)\rangle^{\theta'},\\ F_{\mathrm{KS}}^{-1}(\theta,\theta'):&|S^{\mathrm{MFF}}_{\pm}(r,s,k)\rangle^{\theta}\mapsto|E(r,s,k+2)\rangle^{\pm,\theta'}, \end{aligned} \tag{IV.5}$$

where $|E(r,s,t)\rangle^{\pm,\theta}$ and $|S^{\mathrm{MFF}}_{\pm}(r,s,k)\rangle^{\theta'}$ denote the singular vectors transformed by the corresponding spectral flows.

Evidently, $F_{\mathrm{KS}}(\cdot,\cdot)$ and $F_{\mathrm{KS}}^{-1}(\cdot,\cdot)$ applied to the chains of the respective Verma modules take morphisms (between chains) into morphisms. Thus we have made F_{KS} and F_{KS}^{-1} into functors. The results of this section can be summarized in the following theorem.

Theorem IV.3: *The functors F_{KS} and F_{KS}^{-1} are covariant functors that are inverse to each other and that therefore establish the equivalence between the Verma chain category $\mathcal{CVER}$ on the $\widehat{sl}(2)$ side and the topological Verma chain category $\mathcal{CTVER}$ on the $N=2$ side.*

B. Extending the equivalence to highest-weight-type categories

We have seen in the previous section that the $\widehat{sl}(2)$ Verma, and $N=2$ topological Verma chain categories are equivalent. Now we would like to extend the equivalence to larger categories. Let us define the highest-weight-type chain category.

Definition IV.4: The highest-weight-type chain category $\mathcal{CHW}$ is the category whose objects are chains of modules from the category $\mathcal{HW}$ of $\widehat{sl}(2)$ highest-weight-type representations twisted by the spectral flow with integer parameters. A morphism between two such chains $(\mathcal{N}_\theta)_{\theta\in\mathbb{Z}}$ and $(\mathcal{N}'_\theta)_{\theta\in\mathbb{Z}}$ is any morphism $\mathcal{N}_{\theta_1}\to\mathcal{N}'_{\theta_2}$ (which is understood to be spread over the entire chains by the spectral flow).

Remark IV.5: As an aside, note that the condition for any such chain to admit nontrivial automorphisms singles out the *minimal* $\widehat{sl}(2)$ models, and the same is true for the chains of $N=2$ modules. Then, taking the factors with respect to the respective spectral flows, we would have the same number of fields in the $\widehat{sl}(2)$ and $N=2$ minimal models. Thus, the minimal models of these algebras can be thought of as *periodical chains* of representations; we hope to demonstrate elsewhere the usefulness of such a viewpoint.

On the $N=2$ side, these twisted categories mix into one category because submodules of a (twisted) topological Verma module are the twisted topological Verma modules with different θ.

Definition IV.6: The topological chain category $\mathcal{CTOP}$ is the category whose objects are chains of modules from the topological category $\mathcal{TOP}$ twisted by the spectral flow with integer parameters. A morphism between two such chains $(\mathcal{U}_\theta)_{\theta\in\mathbb{Z}}$ and $(\mathcal{U}'_\theta)_{\theta\in\mathbb{Z}}$ is any morphism $\mathcal{U}_{\theta_1}\to\mathcal{U}'_{\theta_2}$ of $N=2$ modules.

Theorem IV.7: *The functors F_{KS} and F_{KS}^{-1} establish the equivalence between categories $\mathcal{CHW}$ and $\mathcal{CTOP}$.*

C. Equivalence of the relaxed and massive Verma chain categories

We define chains of relaxed Verma modules on the $\widehat{sl}(2)$ side and chains of massive Verma modules on the $N=2$ side, as this was done in Sec. IV A. Then, we formulate the theorem, whose proof is a consequence of Lemma IV.1.

Theorem IV.8: *The KS and anti-KS mappings give rise to identifications, which we denote again by $F_{\mathrm{KS}}(\theta,\theta')$ and $F_{\mathrm{KS}}^{-1}(\theta,\theta')$ respectively, between the following.*

(1) *Singular vectors (II.45) and (II.46) in the relaxed Verma module $\mathfrak{R}_{j,\Lambda,k;\theta}$ and $N=2$ singular vectors (II.129) in the twisted massive Verma modules $\mathfrak{U}_{-[2/(k+2)]j,\Lambda/(k+2),k+2;\theta}$:*

$$\begin{aligned} F_{\mathrm{KS}}(\theta,\theta'):|\Sigma^{\pm}(r,s,j,k)\rangle^{\theta}&\mapsto\left|S\left(r,s,\frac{-2j}{k+2},k+2\right)\right\rangle^{\mp,\theta'},\\ F_{\mathrm{KS}}^{-1}(\theta,\theta'):|S(r,s,h,t)\rangle^{\mp,\theta}&\mapsto\left|\Sigma^{\pm}\left(r,s,-\frac{t}{2}h,t-2\right)\right\rangle^{\theta'}, \end{aligned} \tag{IV.6}$$

and also between the following.

(2) *Charged $\widehat{sl}(2)$ singular vectors (II.30) and charged $N=2$ singular vectors (II.126):*

$$\begin{aligned} F_{\mathrm{KS}}(\theta,\theta'):|C(r,j,k)\rangle^{\theta}&\mapsto|E(r,-2j/(k+2),k+2)\rangle_{\mathrm{ch}}^{\theta'},\\ F_{\mathrm{KS}}^{-1}(\theta,\theta'):|E(r,h,t)\rangle_{\mathrm{ch}}^{\theta}&\mapsto|C(r,-ht/2,t-2)\rangle^{\theta'}, \end{aligned} \tag{IV.7}$$

where $|\Sigma^{\pm}(r,s,j,k)\rangle^{\theta}$, $|S(r,s,h,t)\rangle^{\mp,\theta'}$, etc., denote the respective singular vectors transformed by the corresponding spectral flows.

Again, as in Sec. IV A, we can formulate the following.

Theorem IV.9: *The functors F_{KS} and F_{KS}^{-1} are covariant functors that are inverse to each other and that therefore establish the equivalence of the relaxed Verma chain category $\mathcal{CRVER}$ on the $\widehat{sl}(2)$ side and the massive Verma chain category $\mathcal{CMVER}$ on the $N=2$ side.*

D. Extending the equivalence to the relaxed highest-weight-type categories

We now formulate an extension of the categorial equivalence F_{KS} and F_{KS}^{-1} to the case of the largest categories $\mathcal{CRHW}$ and $\mathcal{CMHW}$ in diagram (I.1). The definitions of the chain categories $\mathcal{CRHW}$ and $\mathcal{CMHW}$ can be obtained by means of a step-by-step repetition of Definitions IV.4 and IV.6 and are omitted. Then the analog of Theorem IV.7 reads as follows.

Theorem IV.10: *The functors F_{KS} and F_{KS}^{-1} establish the equivalence between categories $\mathcal{CRHW}$ and $\mathcal{CMHW}$.*

V. CONCLUDING REMARKS

We believe that the issues addressed in this paper will be essential in a number of problems in the $\widehat{sl}(2)$ and $N=2$ representation theories as well as in conformal field theory.

Comparing the $\widehat{sl}(2)$ and $N=2$ highest-weight representations, it may be observed that the occurrence of infinitely many highest-weight states in $\widehat{sl}(2)$ relaxed Verma modules is intuitively more ''obvious'' than in its $N=2$ counterpart; all of the $\widehat{sl}(2)$ relaxed-highest-weight states are on

the same floor in the extremal diagram, whereas in the $N=2$ case this is not so: extremal diagrams of untwisted massive $N=2$ modules contain only two vectors at the top level, and it may be tempting to assign to these a more fundamental status than to the other vectors from the extremal diagram. Working with the top-level representative of $N=2$ extremal diagrams conceals the ''topological'' (''twisted-chiral'') nature of some of submodules in massive $N=2$ Verma modules (see Ref. 14 for a detailed description of the structure of $N=2$ Verma modules).

As to the structure of relaxed-$\widehat{sl}(2)$ or massive $N=2$ Verma modules, we have seen that a significant difference from the familiar Virasoro or the usual-$\widehat{sl}(2)$ case is that (to use the $N=2$ language) a given massive $N=2$ Verma module may contain submodules of two types: the twisted topological and the massive ones. An important point is that every submodule can be *freely* generated from a twisted (topological) highest-weight vector [singular vectors (II.114) and (II.115) in the topological Verma modules and (II.126) and (II.129) in massive Verma modules]. From these results, it is not too difficult to derive the complete classification of embedding diagrams of the relaxed-$\widehat{sl}(2)$ and massive-$N=2$ Verma modules.[27] These follow (an extension of) the familiar I–II–III pattern; the embedding structure can be described without invoking subsingular vectors,[14] since the above singular vectors generate *maximal* submodules [that subsingular vectors in massive-$N=2$/relaxed-$\widehat{sl}(2)$ modules are superfluous can most clearly be seen in the $\widehat{sl}(2)$ language].

Another problem is the derivation of fusion rules. In order to completely define the fusion rules, one has to fix, among other things, the category of representations. On the $\widehat{sl}(2)$ side, the fusion rules have been obtained in Refs. 1 and 2 (see also Refs. 4 and 5). However, these do not include the twisted (spectral-flow-transformed) $\widehat{sl}(2)$ representations, whose contribution is to be expected on general grounds.[2] This is related to the fact that only the quantum group $osp(1|2)_q$ has so far been observed as a symmetry of $\widehat{sl}(2)$ fusion rules, while a larger quantum group $sl(2|1)_q$ is to be expected. This is an indication that some modules are still missing from the known fusion; we hope that we have described the representation-theoretic part of the desired more general construction. Then the complete $\widehat{sl}(2)$ fusion rules factored with respect to the spectral flow would be isomorphic to complete $N=2$ fusion rules factored over the $N=2$ spectral flow.

ACKNOWLEDGMENTS

We wish to thank F. Malikov and V. Sirota for useful remarks. AMS is also grateful to C. Preitschopf for discussions. The work of AMS and IYT was supported in part by RFFI Grant No. 96-01-00725, the work of IYT was supported in part by a Landau Foundation grant, and the work of AMS, by Grant No. 93-0633-ext from the European Community.

[1] H. Awata and Y. Yamada, ''Fusion rules for the fractional level sl(2) algebra,'' Mod. Phys. Lett. A **7**, 1185 (1992).
[2] B. Feigin and F. Malikov, ''Modular functor and representation theory of $\widehat{sl}_2$,'' Contemp. Math. **202**, ''Operads: Proceedings of Renaissance Conferences.''
[3] M. Ademollo, L. Brink, A. D'Adda, R. D'Auria, E. Napolitano, S. Sciuto, E. Del Guidice, P. Di Vecchia, S. Ferrara, F. Gliozzi, R. Musto, and R. Pettorino, ''Dual string with U(1) color symmetry,'' Phys. Lett. B **62**, 105 (1976); M. Ademollo, L. Brink, A. D'Adda, R. D'Auria, E. Napolitano, S. Sciuto, E. Del Guidice, P. Di Vecchia, S. Ferrara, F. Gliozzi, R. Musto, R. Pettorino, and J. H. Schwarz, ''Dual string models with nonabelian color and flavor symmetries,'' Nucl. Phys. B **111**, 77 (1976).
[4] O. Andreev, ''Operator algebra of the SL(2) conformal field theories,'' Phys. Lett. B **363**, 166 (1995).
[5] J. L. Petersen, J. Rasmussen, and M. Yu, ''Fusion, crossing and monodromy in conformal field theory based on SL(2) current algebra with fractional level,'' Nucl. Phys. B **481**, 577–624 (1996).
[6] P. Furlan, A. Ch. Ganchev, and V. B. Petkova, ''$A_1^{(1)}$ admissible representations—fusion transformations and local correlators,'' Nucl. Phys. B **491**, 635 (1997).
[7] A. M. Semikhatov, ''The MFF singular vectors in topological conformal theories,'' Mod. Phys. Lett. A **9**, 1867 (1994).
[8] V. G. Kac, *Infinite Dimensional Lie Algebras* (Cambridge University Press, Cambridge, 1990).
[9] A. M. Semikhatov and I. Yu. Tipunin, ''Singular vectors of the topological conformal algebra,'' Int. J. Mod. Phys. A **11**, 4597 (1996).
[10] A. M. Semikhatov and I. Yu. Tipunin, ''All singular vectors of the $N=2$ superconformal algebra via the algebraic continuation approach,'' hep-th/9604176.
[11] K. Bardakçi and M. B. Halpern, ''New dual quark models,'' Phys. Rev. D **3**, 2493 (1971).
[12] W. Boucher, D. Friedan, and A. Kent, ''Determinant formulae and unitarity for the $N=2$ superconformal algebras in two-dimensions or exact results on string compactification,'' Phys. Lett. B **172**, 316 (1986).

[13] F. G. Malikov, B. L. Feigin, and D. B. Fuchs, ''Singular vectors in Verma modules over Kač–Moody algebras,'' Funkt. Anal. Pril. **20**, 25 (1986).
[14] A. M. Semikhatov and I. Yu. Tipunin, ''The structure of Verma modules over the $N=2$ superconformal algebra,'' hep-th/9704111.
[15] A. Schwimmer and N. Seiberg, ''Comments on the $N=2$, $N=3$, $N=4$ superconformal algebras in two-dimensions,'' Phys. Lett. B **184**, 191 (1987).
[16] W. Lerche, C. Vafa, and N. P. Warner, ''Chiral rings in $N=2$ superconformal theories,'' Nucl. Phys. B **324**, 427 (1989).
[17] B. L. Feigin and A. V. Stoianovsky, ''Functional models of representations of current algebras and semi-infinite Schubert cells,'' Funkt. Anal. ego Pril. **28**, 68 (1994).
[18] A. M. Semikhatov, ''Verma modules, extremal vectors, and singular vectors on the non-critical $N=2$ string world-sheet,'' hep-th/9610084.
[19] P. Di Vecchia, J. L. Petersen, M. Yu, and H. B. Zheng, ''Explicit construction of unitary representations of the $N=2$ superconformal algebra,'' Phys. Lett. B **174**, 280 (1986).
[20] Y. Kazama and H. Suzuki, ''New $N=2$ superconformal field theories and superstring compactification,'' Nucl. Phys. B **321**, 232 (1989).
[21] V. G. Kač and D. A. Kazhdan, ''Structure of representations with highest weight of infinite dimensional Lie algebras,'' Adv. Math. **34**, 97 (1979).
[22] F. Malikov, ''Verma modules over rank-2 Kač–Moody algebras,'' Alg. Anal. **2**, 65 (1990).
[23] D. H. Friedan, E. J. Martinec, and S. H. Shenker, ''Conformal invariance, supersymmetry and string theory,'' Nucl. Phys. B **271**, 93 (1986).
[24] T. Eguchi, S. Hosono, and S.-K. Yang, ''Hidden fermionic symmetry in conformal topological field theories,'' Commun. Math. Phys. **140**, 159 (1991).
[25] B. Gato-Rivera and A. M. Semikhatov, ''$d\leqslant 1 \cup d\geqslant 25$ and W constraints from BRST-invariance in the $c\neq 3$ topological algebra,'' Phys. Lett. B **293**, 72 (1992).
[26] M. Bershadsky, W. Lerche, D. Nemeschansky, and N. P. Warner, ''Extended $N=2$ superconformal structure of gravity and W gravity coupled to matter,'' Nucl. Phys. B **401**, 304–347 (1993).
[27] A. M. Semikhatov and V. A. Sirota, ''Embedding diagrams of $N=2$ and relaxed-$\widehat{sl}(2)$ Verma modules,'' hep-th/9712102.

The Edmonds asymptotic formulas for the 3*j* and 6*j* symbols

James P. M. Flude[a)]
Department of Mathematics, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom

(Received 3 November 1997; accepted for publication 5 March 1998)

The purpose of this paper is to provide definitions for, and proofs of, the asymptotic formulas given by Edmonds, which relate the $3j$ and $6j$ symbols to rotation matrices. © *1998 American Institute of Physics.* [S0022-2488(98)00907-4]

I. INTRODUCTION

The purpose of this paper is to provide definitions for, and proofs of, the asymptotic formulas given by Edmonds,[1] which relate the $3j$ and $6j$ symbols to rotation matrices.

The study of the asymptotics of the $6j$ symbol dates back to Racah,[2] who stated an asymptotic formula relating $6j$ symbols and Legendre polynomials. Edmonds[1] later generalized Racah's formula by replacing the Legendre polynomial by a rotation matrix. For $a,b,c \gg f,m,n$, Edmonds stated,

$$\begin{Bmatrix} c & b & a \\ f & a+n & b+m \end{Bmatrix} \approx \frac{(-1)^{a+b+c+f+m}}{\sqrt{(2a+1)(2b+1)}}\, d^{f}_{nm}(\phi), \tag{1}$$

where

$$\cos\,\phi = \frac{a(a+1)-b(b+1)-c(c+1)}{2\sqrt{a(a+1)b(b+1)}}. \tag{2}$$

Edmonds also gave a further result which related the $3j$ symbol to a rotation matrix. For $a,c \gg b$, Edmonds stated,

$$\begin{pmatrix} b & a & c \\ \beta & \alpha & \gamma \end{pmatrix} \approx \frac{(-1)^{b-a-\gamma}}{\sqrt{2c+1}}\, d^{b}_{\delta,-\beta}(\theta), \tag{3}$$

where

$$\cos\,\theta\,\frac{\gamma}{\sqrt{c(c+1)}}. \tag{4}$$

Edmonds did not give a proof of either result, nor did he define how the angular momenta scaled or how the $3j$ and $6j$ symbols approached the rotation matrices.

Edmonds reasoned that, for sufficiently large c, (4) could be replaced by

$$\cos\,\theta = \frac{\gamma}{c}. \tag{5}$$

Brussaard and Tolhoek,[3] following a suggestion of Edmonds, argued heuristically, that if $a,c \gg b$ and θ is defined by (5) then (3) holds. Brussaard and Tolhoek did not define how a and c scaled, nor did they define how the $3j$ symbol approached the rotation matrix.

[a)]Electronic mail: james.flude@maths.nott.ac.uk

 © 1998 American Institute of Physics

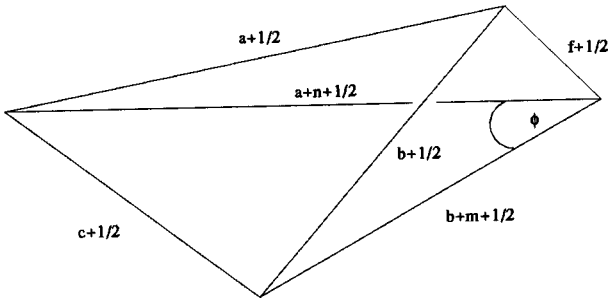

FIG. 1. Associated tetrahedron.

Some years later, Ponzano and Regge[4] discussed the geometry of the asymptotics of the $3j$ and $6j$ symbols. To each asymptotic $6j$ symbol, Ponzano and Regge associated a tetrahedron whose edges had length $j+\frac{1}{2}$, where j is an entry in the $6j$ symbol, and to each asymptotic $3j$ symbol they associated a triangle whose edges were labeled in the same manner. Thus the tetrahedron associated to the above $6j$ symbol is that given in Fig. 1, from which

$$\cos\,\phi=\frac{(a+n+(1/2))^2+(b+m+(1/2))^2-(c+(1/2))^2}{2(a+n+(1/2))(b+m+(1/2))}. \tag{6}$$

The triangle associated to the above $3j$ symbol is given in Fig. 2, from which

$$\cos\,\theta=\frac{\gamma}{c+(1/2)}. \tag{7}$$

Hence there is a discrepancy between the geometrical aspect, given by Ponzano and Regge, and the algebraic aspect, given by Edmonds, Brussaard, and Tolhoek. If we relabel the edges of the associated tetrahedron and triangle by $\sqrt{j(j+1)}$ instead of $j+\frac{1}{2}$ then ϕ and θ are defined by (2) and (4), respectively. In Ref. 4, Ponzano and Regge give an important asymptotic formula for the $6j$ symbol in which the associated tetrahedron is label as above. If we relabel the edges of the associated tetrahedron by $\sqrt{j(j+1)}$ and apply the relevant modifications to Ponzano and Regge's asymptotic formula, then this modified formula provides a worse approximation to the $6j$ symbol than Ponzano and Regge's original formula. This is discussed in the first part of Sec. IV. We define ϕ and θ by (6) and (7), respectively, rather than by (2) and (4) for the following two reasons. The first is that all the formulas are simpler with these definitions and the second is that these definitions are consistent with Ponzano and Regge which is important for the above reasons.

The purpose of this paper is to provide definitions for, and proofs of, the asymptotic formulas which relate the $3j$ symbol and rotation matrix of (3), and the $6j$ symbol and rotation matrix of

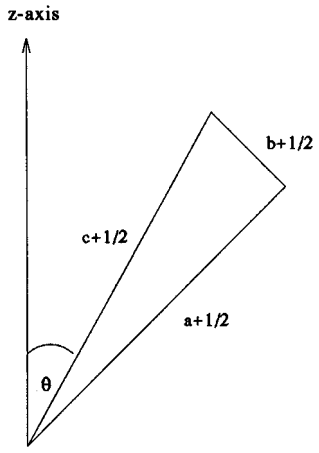

FIG. 2. Associated triangle.

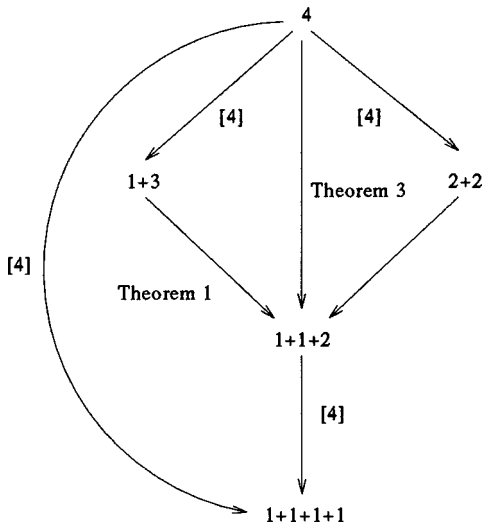

FIG. 3. Asymptotics of the $6j$ symbol.

(1). In both cases, we define how the angular momenta scale, and we also define how these particular $3j$ and $6j$ symbols approach the rotation matrices of (3) and (1), respectively. These definitions make precise $a,c \gg b$ and $a,b,c \gg f,m,n$ and $\approx$. With these definitions, in Sec. II, we prove (3), where θ is defined by (7), and in Sec. III, we prove (1), where ϕ is defined by (6). In the first part of Sec. IV, we discuss why ϕ and θ are defined by (6) and (7) rather than by (2) and (4). In the last part of Sec. IV, we discuss the relationship between the results proved in Secs. II and III. This entails using the result proved in Sec. II to arrive at an expression of a similar type to that proved in Sec. III.

The study of the asymptotics of $3j$ and $6j$ symbols is important in understanding the Ponzano–Regge model of three dimensional Euclidean quantum gravity.[4] In this approach, three dimensional spacetime is approximated by a simplicial manifold formed by gluing together tetrahedra. To each edge of a tetrahedron in this simplicial manifold Ponzano and Regge associate a half integer j, with the length of that edge defined as $j+\frac{1}{2}$, in such a way that the triangle inequality is satisfied on all the faces of that tetrahedron. They then thus associate to each such labeled tetrahedron a $6j$ symbol.

Any potential candidate for the quantum theory of gravity must reduce to general relativity in some suitable limit. By allowing various combinations of the edges in each tetrahedron in the simplicial manifold to become large, the Ponzano–Regge model reduces to three dimensional general relativity without a cosmological constant. Thus to understand the relationship between the Ponzano–Regge model of quantum gravity and general relativity the asymptotics of the $3j$ and $6j$ symbols is important. The reduction of the above asymptotic $6j$ symbol to the rotation matrix $d^{f}_{nm}(\theta)$ plays a crucial role in understanding the relationship between the Ponzano–Regge model of three dimensional quantum gravity and the construction of topological state sums from tensor categories.[5]

Figure 3[6] shows all the possible asymptotic limits of the $6j$ symbol. These limits correspond to moving various combinations of the four vertices of the associated tetrahedron. At the top of Fig. 3 is the unscaled $6j$ symbol in which the four vertices remain fixed. In the $1+3$ case, one vertex is taken to infinity while the remaining three remain fixed. This represents the reduction of a $6j$ symbol to a $3j$ symbol.[4] In the $1+1+2$ case, two of the vertices remain fixed and the remaining two move away from the fixed vertices as well as away from each other. This represents (1).

II. ROTATION MATRICES AND 3*j* SYMBOLS

Following Edmonds,[1] we first define the $3j$ symbol and then the rotation matrix.

Let a, b, and c be non-negative half integers which satisfy the triangle inequality, and let $|\alpha| \leqslant a$, $|\beta| \leqslant b$, and $|\gamma| \leqslant c$ with $\alpha+\beta+\gamma=0$, then the $3j$ symbol is defined by

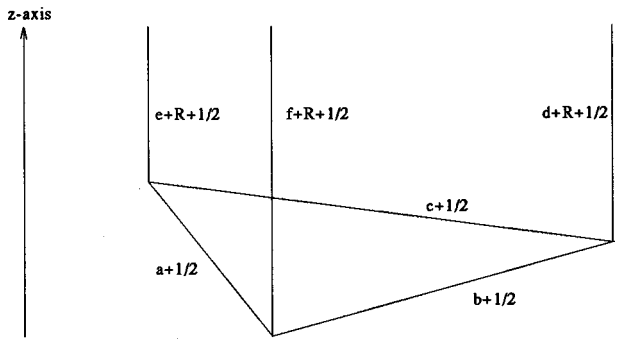

FIG. 4. Associated tetrahedron.

$$\begin{pmatrix} a & b & c \\ \alpha & \beta & \gamma \end{pmatrix} = [\Delta(abc)(a+\alpha)!(a-\alpha)!(b+\beta)!(b-\beta)!(c+\gamma)!(c-\gamma)!]^{1/2} \times \sum_z \frac{(-1)^{z+a-b-\gamma}}{(z!(a+b-c-z)!(c-b+\alpha+z)!(b+\beta-z)!(a-\alpha-z)!(c-a-\beta+z)!)}, \tag{8}$$

where the summation is over all integers z such that the factorial arguments are non-negative and

$$\Delta(abc) = \frac{(a+b-c)!(a-b+c)!(-a+b+c)!}{(a+b+c+1)!}.$$

Let j be a non-negative half integer and let $|m'|$, $|m| \leqslant j$, then the rotation matrices are defined by

$$d^j_{m'm}(\theta) = (-1)^{j-m'}[(j+m')!(j-m')!(j+m)!(j-m)!]^{1/2} \times \sum_k (-1)^k \frac{(\cos\,\theta/2)^{2k+m'+m}(\sin\,\theta/2)^{2j-2k-m'-m}}{k!(j-m'-k)!(j-m-k)!(m'+m+k)!},$$

where the summation is over all integers k such that the factorial arguments are non-negative. The rotation matrices are the matrix elements in the $2j+1$ dimensional representation of rotations about a fixed axis.

We have divided the discussion of the relationship between the $3j$ symbol and the rotation matrix into two cases depending upon γ. We discuss the case where $c-|\gamma| \to \infty$ in theorem 1 and the case where $c = \pm\gamma$ in theorem 2.

Theorem 1: *Fix b and β and let $a, c \to \infty$ in such a way that $c-a=\delta$ is fixed and let $c-|\gamma| \to \infty$, then*

$$\sqrt{a+b+c+1}\begin{pmatrix} b & a & c \\ \beta & \alpha & \gamma \end{pmatrix} - (-1)^{b-a-\gamma} d^b_{\delta,-\beta}(\theta) \to 0,$$

where

$$\cos\,\theta = \frac{\gamma}{c+(1/2)}. \tag{9}$$

The theorem is nontrivial as the right hand side does not converge to zero under this asymptotic limit.

Proof: Eliminating a and α from (8), as $c-a=\delta$ and $\alpha+\beta+\gamma=0$, and then expanding the Δ symbol gives

$$\begin{pmatrix} b & a & c \\ \beta & \alpha & \gamma \end{pmatrix} = (-1)^{b-a-\gamma}\left[\frac{(b+\beta)!(b-\beta)!(b-\delta)!(b+\delta)!}{2c+b-\delta+1}\right]^{1/2} \times \sum_z \frac{(-1)^z[\Omega_1\Omega_2\Omega_3]^{1/2}}{z!(b-\delta-z)!(b-\beta-z)!(\beta+\delta+z)!}, \tag{10}$$

where

$$\Omega_1 = \frac{(2c-\delta-b)!}{(2c-\delta+b)!}, \qquad \Omega_2 = \frac{(c+\gamma-\delta+\beta)!(c+\gamma)!}{[(c+\gamma-b+\beta+z)!]^2},$$

$$\Omega_3 = \frac{(c-\gamma-\delta-\beta)!(c-\gamma)!}{[(c-\gamma-\delta-\beta-z)!]^2}.$$

As a and c tend to infinity in the manner defined above, it is important that the range of summation in (10) remains finite. If this is not the case then the error generated by the approximations used to reduce the $3j$ symbol to a rotation matrix may not converge to zero. Brussaard and Tolhoek[4] did not discuss the effect the asymptotic limit had on the range of summation. The summation variable z is bounded above by

$$\min(c-\delta-\gamma-\beta, b-\delta, b-\beta)$$

and is bounded below by

$$\max(-c+b-\gamma-\beta, 0, -\beta-\delta).$$

By hypothesis, $c-|\gamma|\to\infty$ so that the first term in each of the above is redundant as b, β, and δ are all fixed. Thus the range of summation is fixed eventually.

Stirling's approximation[7] to $n!$ is

$$n! \sim \sqrt{2\pi} n^{n+(1/2)} e^{-n}$$

where $a \sim b$ means $a/b \to 1$ as n tends to infinity. Also,

$$n! \sim \sqrt{2\pi}(n+1)^{n+(1/2)} e^{-(n+1)} \tag{11}$$

and

$$n! \sim \sqrt{2\pi}(n+\tfrac{1}{2})^{n+(1/2)} e^{-(n+(1/2))}. \tag{12}$$

Applying (11) to the first factor in Ω and (12) to the rest gives

$$\Omega \sim \left(\sin\frac{\theta}{2}\right)^{2z+\beta+\delta}\left(\cos\frac{\theta}{2}\right)^{2b-2z-\beta-\delta}, \tag{13}$$

where θ defined by (9) and $\Omega = [\Omega_1\Omega_2\Omega_3]^{1/2}$. The constraint that the factorial arguments in (8) have to be non-negative implies that the above powers are also non-negative as

$$2z+\beta+\delta = z+(z+\beta+\delta),$$

$$2b-2z-\beta-\delta = (b-\beta-z)+(b-\delta-z).$$

The right-hand side of (13) is bounded above by 1 and so

$$\Omega - \left(\sin\frac{\theta}{2}\right)^{2z+\beta+\delta}\left(\cos\frac{\theta}{2}\right)^{2b-2z-\beta-\delta} \to 0.$$

As the range of summation in (10) is eventually fixed, the error generated by applying various approximations to the factorials in Ω_i, $i=1,2,3$, will converge to zero. Thus

$$\sqrt{a+b+c+1}\begin{pmatrix} b & a & c \\ \beta & \alpha & \gamma \end{pmatrix} - (-1)^{b-a-\gamma}(-1)^{\beta-b} d^b_{\beta,\delta}(\pi-\theta) \to 0$$

and as $d^b_{\beta,\delta}(\pi-\theta)=(-1)^{b-\beta} d^b_{\delta,-\beta}(\theta)$ the result follows. □

Theorem 2: *Fix b and β, let $\gamma=\pm c$ and let $a,c\to\infty$ so that $c-a=\delta$ is fixed, then*

$$\sqrt{a+b+c+1}\begin{pmatrix} b & a & c \\ \beta & \alpha & \gamma \end{pmatrix} - (-1)^{\xi}\delta_{\alpha,\mp a} \to 0,$$

where

$$\xi = \begin{cases} -a+b-c & \text{for } \gamma=c \\ 2b+\delta-\beta & \text{for } \gamma=-c. \end{cases}$$

Proof: Similar to that for theorem 1. □

III. ROTATION MATRICES AND 6*j* SYMBOLS

Following Edmonds,[1] we define the $6j$ symbol as follows. Let a, b, c, d, e, and f be non-negative half integers such that each of the triples $\{abc\}$, $\{cde\}$, $\{aef\}$, and $\{bdf\}$ satisfies the triangle inequality and the sum of the entries in each triple is an integer, then the $6j$ symbol is defined by

$$\begin{Bmatrix} a & b & c \\ d & e & f \end{Bmatrix} = [\Delta(abc)\Delta(cde)\Delta(aef)\Delta(bdf)]^{1/2} \times \sum_z \frac{(-1)^z (z+1)!}{\begin{pmatrix} (z-a-b-c)!(z-a-e-f)!(b+c+e+f-z)!(z-d-e-c)! \\ \times (z-b-d-f)!(a+b+d+e-z)!(a+c+f+d-z)! \end{pmatrix}},$$

where the summation is over all integers z such that the factorial arguments are non-negative and

$$\Delta(abc) = \frac{(a+b-c)!(a-b+c)!(-a+b+c)!}{(a+b+c+1)!}.$$

Theorem 3: *Fix f, m, n and let $a=\lambda a_0$, $b=\lambda b_0$ and $c=\lambda c_0$ with $\lambda\to\infty$, then*

$$\sqrt{\omega_a \omega_b}\begin{Bmatrix} c & b & a \\ f & a+n & b+m \end{Bmatrix} - (-1)^{\xi} d^f_{nm}(\phi) \to 0,$$

where

$$\omega_a = 2a+n+f+1, \quad \omega_b = 2b+m+f+1, \quad \xi = a+b+c+f+m$$

and

$$\cos\phi = \frac{(a+n+(1/2))^2+(b+m+(1/2))^2-(c+(1/2))^2}{2(a+n+(1/2))(b+m+(1/2))}.$$

The scaling used in theorem 3 is clearly different from that used in theorem 1. The reason for this is that as λ tends to infinity, ϕ is preserved, whereas, if we had scaled a_0, b_0, and c_0 as we did in theorem 1 then ϕ would not have been preserved. Theorem 3 is nontrivial as $d^f_{nm}(\phi)$ does not converge to zero under this asymptotic limit.

Proof: Changing the summation variable from z to t, where $z=a+b+c+m+n+t$, and then expanding the Δ symbols gives,

$$\sqrt{\omega_a\omega_b}\begin{Bmatrix} c & b & a \\ f & a+n & b+m \end{Bmatrix} = (-1)^{\nu}[(f+m)!(f-m)!(f+n)!(f-n)!]^{1/2} \times \sum_t \frac{(-1)^t[\Pi_{i=1}^6 \Phi_i]^{1/2}}{t!(f-m-t)!(f-n-t)!(m+n+t)1},$$

where $\nu = a+b+c+m+n$ and

$$\Phi_1 = \frac{(2a+n-f)!}{(2a+n+f)!}, \qquad \Phi_2 = \frac{(2b+m-f)!}{(2b+m+f)!},$$

$$\Phi_3 = \frac{(a+b-c)!(a+b-c+m+n)!}{[(a+b-c-t)!]^2},$$

$$\Phi_4 = \frac{(a-b+c)!(a-b+c+n-m)!}{[(a-b+c+n-f+t)!]^2},$$

$$\Phi_5 = \frac{(-a+b+c)!(-a+b+c+m-n)!}{[(-a+b+c+m-f+t)!]^2},$$

$$\Phi_6 = \frac{[(a+b+c+m+n+1+t)!]^2}{(a+b+c+1)!(a+b+c+m+n+1)!}.$$

As in the proof of theorem 1, the constraints on t ensure that as a, b, and c tend to infinity in the manner defined above, the range of summation remains finite so that the error generated by applying various approximations to the factorials in Φ_i, $i=1,2,...,6$, will converge to zero. By applying (11) to Φ_1 and Φ_2 and (12) to the rest and following the arguments used in the proof of theorem 1 the result follows. □

IV. DISCUSSION

In the first part of this section we discuss why ϕ and θ are defined by (6) and (7) rather than by (2) and (4). In the last part of this section we discuss the relationship between theorems 1 and 3 which entails using theorem 1 to arrive at an expression of a similar type to that in theorem 3.

In the Introduction we noted that there was a discrepancy between the geometrical and algebraic parts of the asymptotics of the $6j$ symbol. Edmonds defined ϕ and θ by (2) and (4), respectively, whereas the geometrical interpretation of the asymptotics, given by Ponzano and Regge, leads to ϕ and θ being defined by (6) and (7). Of course, (2) is asymptotic to (6), and (4) is asymptotic to (7), because, if we substitute the binomial expansion of $\sqrt{j(j+1)}$ into each of the terms in (2) and (4) then (2) will converge to (6), and (4) will converge to (7), but, the geometry which leads to ϕ and θ being defined as in (2) and (4), when applied to a certain asymptotic formula, leads to a worse approximation to the $6j$ symbol. In Ref. 4, Ponzano and Regge give an asymptotic formula for the $6j$ symbol in which all the entries all large. They label the edges of the tetrahedron associated to this asymptotic $6j$ symbol by $j+\frac{1}{2}$, where j is an entry in the $6j$ symbol. Suppose we now relabel the edges of the this tetrahedron by $\sqrt{j(j+1)}$ and apply the relevant modifications to the Ponzano–Regge asymptotic formula. Then as can be seen in Table I, this modified asymptotic formula provides a worse approximation to the $6j$ symbol than Ponzano and Regge's original formula. Thus the labeling of the edges of the associated tetrahedron by $j+\frac{1}{2}$ is critical in the Ponzano–Regge asymptotic formula and as (1) and (3) should give rise to an asymptotic formula of the Ponzano–Regge type under some further asymptotic limit, we shall define ϕ and θ as in (6) and (7). Also, by defining ϕ and θ by (6) and (7), all the formulas are simpler, which can be seen in the proofs of theorems 1 and 3.

The column labeled $6j$ contains the exact value of the $6j$ symbol

$$\begin{Bmatrix} a & b & c \\ d & e & f \end{Bmatrix}$$

TABLE I. Exact and approximate values for the $6j$ symbol.

a	b	c	d	e	f	$6j$	$j+\frac{1}{2}$	$\sqrt{j(j+1)}$
1	1	1	1	1	1	0.166 666 67	0.166 826 79	−0.078 256 88
$\frac{7}{2}$	7	$\frac{9}{2}$	$\frac{17}{2}$	5	$\frac{5}{2}$	−0.041 785 54	−0.041 520 25	−0.029 373 47
$\frac{17}{2}$	$\frac{15}{2}$	10	$\frac{15}{2}$	$\frac{15}{2}$	4	0.016 494 29	0.016 422 50	0.012 934 66
$\frac{13}{2}$	8	$\frac{9}{2}$	$\frac{13}{2}$	6	$\frac{15}{2}$	0.025 518 04	0.025 505 76	0.022 840 36
5	8	12	9	7	6	−0.022 441 18	−0.022 422 08	−0.019 444 64
9	9	9	9	9	9	−0.015 650 06	−0.015 640 09	−0.014 856 78
7	8	9	6	9	5	−0.013 703 61	−0.013 212 03	−0.008 720 52
7	8	9	6	9	14	0.001 663 74	0.001 663 76	0.012 944 91
13	15	24	$\frac{29}{2}$	$\frac{33}{2}$	$\frac{19}{2}$	−0.008 529 12	−0.008 561 89	−0.008 929 74
13	15	24	$\frac{29}{2}$	$\frac{33}{2}$	$\frac{29}{2}$	−0.005 622 96	−0.005 432 66	−0.004 712 80

as defined in Sec. III. The next column contains the Ponzano–Regge approximation to the $6j$ symbol in which all the entries are large,[4]

$$\begin{Bmatrix} a & b & c \\ d & e & f \end{Bmatrix} \simeq \frac{1}{\sqrt{12\pi V}} \cos\left(\sum_{h<k} j_{hk}\theta_{hk} + \frac{\pi}{4} \right),$$

where

$$j_{12}=a+\tfrac{1}{2}, \quad j_{13}=b+\tfrac{1}{2}, \quad j_{14}=c+\tfrac{1}{2},$$

$$j_{23}=d+\tfrac{1}{2}, \quad j_{24}=e+\tfrac{1}{2}, \quad \text{and } j_{34}=f+\tfrac{1}{2},$$

are the edge lengths of the associated tetrahedron and θ_{hk} is the exterior dihedral angle between the two planes on the associated tetrahedron which meet in the edge j_{hk}, and V is the volume of the associated tetrahedron. The last column contains the modified Ponzano–Regge formula in which $\sqrt{j(j+1)}$ has replaced $j+\frac{1}{2}$ everywhere, including in the calculation of the volume V and the exterior dihedral angles θ_{hk}.

Superficially, theorems 1 and 3 appear to be similar as they both involve the reduction of the primary object to a rotation matrix. It is thus natural to ask if it is possible to arrive at an expression of a similar type to that in theorem 3 by using theorem 1. This is indeed possible, but it is important to point out that the asymptotic limit in this result will be the limit used in theorem 1 which is not the same as the scaling used in theorem 3.

The following arguments are not intended to prove the link between theorems 1 and 3 but are to be regarded as a possible justification for there being a link.

To discuss the existence of such a link we shall use two limits, both of which are of the same type used in theorem 1. The first limit reduces the relevant $6j$ symbol to a $3j$ symbol,[4] and the second limit when applied to this $3j$ symbol reduces it, by theorem 1, to a rotation matrix. The overall result is an expression of the same type as in theorem 3 but with a different scaling of the non-negative half integers.

In Ref. 4, Ponzano and Regge discuss how the $6j$ symbol

$$\begin{Bmatrix} a & b & c \\ d+R & e+R & f+R \end{Bmatrix}$$

reduces to the $3j$ symbol

$$\begin{pmatrix} a & b & c \\ \alpha & \beta & \gamma \end{pmatrix}$$

as $R\to\infty$, where $\alpha=e-f$, $\beta=f-d$, and $\gamma=d-e$. If we expand the particular factorials in the above $6j$ symbol suggested by Ponzano and Regge, using (11) and (12), and then follow the proof of theorem 1, this shows that, as $R\to\infty$,

$$\sqrt{2R+1}\begin{Bmatrix} a & b & c \\ d+R & e+R & f+R \end{Bmatrix} - (-1)^{\chi}\begin{pmatrix} a & b & c \\ \alpha & \beta & \gamma \end{pmatrix} \to 0, \tag{14}$$

where $\chi = a+b+c+2(d+e+f)$. This represents a rigorous statement of Ponzano and Regge's work.

The geometrical interpretation of the above result is discussed by Ponzano and Regge in Ref. 4. As $R \to \infty$ in the above $6j$ symbol, three of the vertices of the associated tetrahedron remain fixed and the remaining one moves further and further away from the others. This causes the edges labeled by d, e, and f to become parallel to each other and it is these edges which define a z axis, with α, β, and γ being the projections of the edges labeled by a, b, and c onto this axis (see Fig. 4). The triangle associated to the above $3j$ symbol is the unscaled triangle in the associated tetrahedron formed by the three fixed vertices.

Now, replace a by b, b by $b+m$, c by f, d by $a+n$, e by a and f by c in (14) and apply the symmetries

$$\begin{Bmatrix} a & b & c \\ d & e & f \end{Bmatrix} = \begin{Bmatrix} f & a & e \\ c & d & b \end{Bmatrix}$$

and

$$\begin{pmatrix} a & b & c \\ \alpha & \beta & \gamma \end{pmatrix} = \begin{pmatrix} c & a & b \\ \gamma & \alpha & \beta \end{pmatrix},$$

which gives

$$\sqrt{\omega_1}\begin{Bmatrix} c+R & b & a+R \\ f & a+n+R & b+m \end{Bmatrix} - (-1)^{\xi}\begin{pmatrix} f & b & b+m \\ n & a-c & c-a-n \end{pmatrix} \to 0, \tag{15}$$

where

$$\omega_1 = 2R+1 \quad \text{and} \quad \xi = 2b+m+f+2n+2(c+R).$$

We now scale b by adding to it a real number R' which is tending to infinity. Thus, by theorem 1,

$$\sqrt{\omega_2}\begin{pmatrix} f & b+R' & b+m+R' \\ n & a-c & c-a-n \end{pmatrix} - (-1)^{\mu} d^{f}_{nm}(\phi) \to 0, \tag{16}$$

where

$$\omega_2 = 2(b+R')+m+f+1 \quad \text{and} \quad \mu = (a+R)-(b+R')-(c+R)+2f$$

and ϕ is the angle between the edges labeled by $a+n+R$ and $b+m+R'$ in the tetrahedron associated to the above $6j$ symbol. Equations (15) and (16) give

$$\sqrt{\omega_1\omega_2}\begin{Bmatrix} c+R & b+R' & a+R \\ f & a+n+R & b+m+R' \end{Bmatrix} - (-1)^{\tau} d^{f}_{nm}(\phi) \to 0,$$

where

$$\tau = (a+R)+(b+R')+(c+R)+f+m$$

and

$$\cos\phi = \frac{(a+n+R+(1/2))^2+(b+m+R'+(1/2))^2-(c+R+(1/2))^2}{2(a+n+R+(1/2))(b+m+R'+(1/2))}.$$

This result is of the same type as theorem 3 but with two scalings.

In the Introduction we had Fig. 3 which represented the different asymptotic limits of the $6j$ symbol in terms of the tetrahedron associated to that $6j$ symbol. The above result is represented by the arrow from 4 to $1+1+2$ via $1+3$. The arrow from 4 to $1+3$ is Ponzano and Regge's reduction of a $6j$ symbol to a $3j$ symbol in which one vertex of the associated tetrahedron is taken to infinity while the remaining three are fixed, and the arrow from $1+3$ to $1+1+2$ is where one of the three previously fixed vertices is taken to infinity. Hence the present paper has completed the triangle with vertices 4, $1+3$ and $1+1+2$.

ACKNOWLEDGMENTS

The author would like to thank John Barrett for several helpful discussions. Figure 3 is produced here courtesy of John Barrett. The author is supported by an EPSRC research studentship.

[1] A. R. Edmonds, *Angular Momentum in Quantum Mechanics*, CERN 55–26 Geneva, 1955.
[2] G. Racah, *Group Theory and Spectroscopy* (Institute of Advanced Study, Princeton, 1951), p. 74.
[3] P. J. Brussaard and H. A. Tolhoek, Physica (Amsterdam) **XXIII**, 955 (1957).
[4] G. Ponzano and T. Regge, in *Spectroscopic and Group Theoretical Methods in Physics*, edited by F. Block (North-Holland, Amsterdam, 1968), pp. 1–58.
[5] J. W. Barrett and L. Crane, Class. Quantum Grav. **14**, 2113 (1997).
[6] J. W. Barrett (private communication, 1997).
[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Wiley, New York, 1966).

On a class of Robin boundary value problems in physical geodesy

Stefan Ritter[a)]
Mathematisches Institut II, Universität Karlsruhe, D-76128 Karlsruhe, Germany

(Received 23 October 1997; accepted for publication 2 March 1998)

In physical geodesy one considers several (exterior) Robin boundary value problems for the Laplace equation in three dimensions. The ellipsoidal Stokes problem, which occurs in context of gravimetric determination of the geoid, belongs to this class. Up to now, this and related problems have been treated with high order series expansions of spherical and spheroidal harmonics. In this article we investigate the nullfield method for this class of boundary value problems. An integral equation formulation is achieved, and existence and uniqueness conditions are attained in view of the Fredholm alternative. For the case that the underlying surface is a triaxial ellipsoid or an oblate spheroid, explicit expressions for the eigenvalues and eigenfunctions for the boundary integral operator are provided. © *1998 American Institute of Physics.* [S0022-2488(98)02707-8]

I. INTRODUCTION

The ellipsoidal Stokes problem is an exterior Robin (third) boundary value problem (BVP) for the Laplace equation in three dimensions. It arises in physical geodesy as one possible form of linearized BVP in the context of gravimetric determination of the geoid and is therefore of considerable interest. The ellipsoidal approach to the geoid reflects the real situation much better than the spherical approach.

Let D denote an (open) triaxial ellipsoid with semi-axes $a>b>c$ and boundary ∂D with outer unit normal n (see the Appendix). The Stokes problem is to find a solution u on ∂D of

$$\Delta u = 0, \quad \text{in } \mathbb{R}^3\backslash\overline{D}, \tag{1}$$

$$\frac{\partial u}{\partial n} + \lambda p u = k, \quad \text{on } \partial D, \tag{2}$$

$$u(x) = O\left(\frac{1}{|x|}\right), \quad |x|\to\infty. \tag{3}$$

Here, $\lambda\in\mathbb{R}$, $p(x):=(n(x),x)$, $x\in\partial D$, is the distance of the tangent plane in $x\in\partial D$ from the origin, and $k\in C(\partial D)$. In the mathematical literature the function p is known as the *support function* of D.

The solution u has the physical meaning of the *anomalous potential*, i.e., the difference between the gravity potential of the Earth and a reference gravity potential. By λp we introduce a variety of boundary conditions. In applications, λ will be determined such that λp shall be a fairly good approximation of $-(1/\gamma)(\partial\gamma/\partial n)$, where γ is the *normal gravity*. The right-hand side k is the negative *gravity anomaly*.

There are several papers about this and related problems: in Ref. 1 solutions of the ellipsoidal Stokes problem are considered in terms of an infinite series; and in Refs. 2 and 3, Zagrebin's problem, which differs only slightly (in one term of the boundary condition) from our problem are studied. All of these papers refer to oblate spheroids.

For convenience we introduce the notations

a)Electronic mail: Stefan.Ritter@math.uni-karlsruhe.de

 © 1998 American Institute of Physics

$$u(x)|_{\pm} := \lim_{t\to 0\pm} u(x+t\ n(x)) \quad \text{and} \quad \left.\frac{\partial u(x)}{\partial n(x)}\right|_{\pm} := \lim_{t\to 0\pm} (n(x), \text{grad}\ u(x+t\ n(x))),$$

for $x \in \partial D$.

For $\lambda \leqslant 0$, the first Green theorem for the solution u of the homogeneous problem ($k=0$) yields

$$-\int_{\mathbb{R}^3\setminus\overline{D}} |\text{grad}\ u(y)|^2 dy = \int_{\partial D} u(y)|_{+} \left.\frac{\partial u(y)}{\partial n(y)}\right|_{+} ds(y) = -\lambda \int_{\partial D} p(y)|u(y)|_{+}|^2 ds(y),$$

and hence $u = \text{const}$ in $\mathbb{R}^3\setminus\overline{D}$ and with (3) it follows $u=0$ in $\mathbb{R}^3\setminus\overline{D}$. Unfortunately for $\lambda > 0$, which is the very interesting case in geodesy, uniqueness of u cannot be obtained.

A classical approach is to discuss a single-layer potential ansatz for (1), (2), (3). This leads to an integral equation for an unknown density which is of *mediate* interest, and, the final solution is obtained in an additional step by evaluating the potential on the boundary.

In this paper the nullfield method is applied. By use of suitable representation formulas for harmonic functions, the BVP is transformed to a Fredholm integral equation of the second kind. Since the accompanying integral operators are compact, the Fredholm alternative gives existence and uniqueness conditions for the solution of the integral equation and for the BVP itself. In contrast to the single-layer approach the nullfield method leads directly to an integral equation for the required boundary values.

The plan is as follows: in Sec. II, the nullfield method is introduced and the integral equation formulation for (1)–(3) is derived. In Sec. III, we study the integral equation for the case of the triaxial ellipsoid in detail and work out the consequences of Fredholm's alternative for the case $\lambda > 0$. Then we discuss the important (degenerate) case of ∂D being an oblate spheroid and we provide plots of the eigenvalues of the boundary integral operator. We finish the paper by presenting uniqueness-conditions for u.

II. THE NULLFIELD METHOD

We study the BVP (1)–(3) by an integral equation approach using the so-called *nullfield method* (see Ref. 4): Choose a suitable representation formula for harmonic functions in $\mathbb{R}^3\setminus\overline{D}$ that vanishes for $x \in D$. Then insert the given boundary data into this formula and take it as ansatz for the unknown boundary parameters. Assure for this ansatz the complementary nullfield condition which leads to an integral equation. Then the equivalence of the integral equation and the BVP is to be verified and the integral equation is to be solved.

In the first step we take the representation theorem for a function u which is (i) continuously differentiable in $\mathbb{R}^3\setminus\partial D$, (ii) harmonic in $\mathbb{R}^3 \cup \infty \setminus \partial D$, and (iii) converges uniformly to zero for $|x| \to \infty$ (see Ref. 5, p. 146),

$$u(x) = -\frac{1}{4\pi}\int_{\partial D} \left\{ \left(\left.\frac{\partial u(y)}{\partial n(y)}\right|_{+} - \left.\frac{\partial u(y)}{\partial n(y)}\right|_{-} \right) \frac{1}{|x-y|} - (u(y)|_{+} - u(y)|_{-}) \frac{\partial \frac{1}{|x-y|}}{\partial n(y)} \right\} ds(y),$$
$$x \in \mathbb{R}^3\setminus\partial D. \tag{4}$$

Then the nullfield condition $u=0$ in D yields $u(x)|_{-}=0$ on ∂D and hence $\partial u/\partial n\,|_{-}=0$ on ∂D. The ansatz $\mu := u|_{+}$ and (2) lead to

$$\left.\frac{\partial u(y)}{\partial n(y)}\right|_{+} = k(y) - \lambda p(y)\mu(y), \quad y \in \partial D, \tag{5}$$

and for $u(x)$ in $\mathbb{R}^3\setminus\overline{D}$ we obtain

$$u(x)=-\frac{1}{4\pi}\int_{\partial D}\frac{k(y)-\lambda p(y)\mu(y)}{|x-y|}ds(y)+\frac{1}{4\pi}\int_{\partial D}\mu(y)\frac{\partial\frac{1}{|x-y|}}{\partial n(y)}ds(y). \tag{6}$$

By straightforward use of the continuity of single-layer potentials and the limiting relation for the double-layer potentials (see, e.g., Ref. 4),

$$\lim_{D\ni x\to x_0\in\partial D}\frac{1}{4\pi}\int_{\partial D}\varepsilon(y)\frac{\partial\frac{1}{|x-y|}}{\partial n(y)}ds(y)=\frac{1}{4\pi}\int_{\partial D}\varepsilon(y)\frac{\partial\frac{1}{|x_0-y|}}{\partial n(y)}ds(y)-\frac{1}{2}\varepsilon(x_0), \tag{7}$$

for $\varepsilon\in C(\partial D)$, the nullfield condition $u(x)|_-=0$, $x\in\partial D$, leads to the Fredholm integral equation of the second kind,

$$\mu(x)-\mathbf{A}'\mu(x)=-\mathbf{S}k(x),\quad x\in\partial D, \tag{8}$$

with the integral operator

$$\mathbf{A}'\mu(x):=\frac{1}{2\pi}\int_{\partial D}\mu(y)\frac{\partial\frac{1}{|x-y|}}{\partial n(y)}ds(y)+\lambda\frac{1}{2\pi}\int_{\partial D}\frac{p(y)\mu(y)}{|x-y|}ds(y),\quad x\in\partial D, \tag{9}$$

and the single-layer potential operator,

$$\mathbf{S}k(x):=\frac{1}{2\pi}\int_{\partial D}\frac{k(y)}{|x-y|}ds(y),\quad x\in\partial D, \tag{10}$$

for $\mu,k\in C(\partial D)$. Conversely, by subsequent use of the limiting properties of single-layer and double-layer potentials (see, e.g., Ref. 4) we see that u from (6) with μ from (8) is a solution of the boundary value problem.

The nullfield condition $u=0$ in D, or, equivalently, $u=0$ on ∂D may seem artificial because it prescribes u in D which is the complementary domain of the boundary value problem. But, due to the representation formula, this is not a restriction for u in the domain of our interest. On the other hand this condition yields an integral equation of second kind for $u|_+$ in a direct way.

Integral equation (8) can be solved numerically in various ways, e.g., by boundary element techniques such as Galerkin- and collocation methods. Spectral methods are also appropriate because we obtain a complete system of eigenfunctions of the integral operators $\mathbf{A}'$ and $\mathbf{S}$.

In what follows we investigate the theory of (8), i.e., we study solvability and give uniqueness conditions for its solutions.

First we note that $\mathbf{A}'$ is a linear combination of two classical potential operators, namely the *adjoint electrostatic operator*,

$$\mathbf{K}'\mu(x):=\frac{1}{2\pi}\int_{\partial D}\mu(y)\frac{\partial\frac{1}{|x-y|}}{\partial n(y)}ds(y),\quad x\in\partial D, \tag{11}$$

and the single-layer operator,

$$\mathbf{T}'\mu(x):=\frac{1}{2\pi}\int_{\partial D}\frac{p(y)\mu(y)}{|x-y|}ds(y),\quad x\in\partial D, \tag{12}$$

say

$$\mathbf{A}'\mu=\mathbf{K}'\mu+\lambda\mathbf{T}'\mu. \tag{13}$$

Both $\mathbf{K}'$ and $\mathbf{T}'$ are well known to be compact in the Banach space $C(\partial D)$ of continuous complex valued functions on ∂D equipped with the maximum norm. Hence $\mathbf{A}'$ is compact in $C(\partial D)$. For $k \in C(\partial D)$ we have that $\mathbf{S}k$ is continuous on ∂D and therefore the right-hand side of (8) is continuous.

Let $\langle C(\partial D), C(\partial D)\rangle$ denote the dual system corresponding to the nondegenerate bilinear form

$$\langle u,v\rangle := \int_{\partial D} u(y)\overline{v(y)}ds(y). \tag{14}$$

By interchanging the order of integration, it is readily seen that the adjoint $\mathbf{A}$ of $\mathbf{A}'$ with respect to (14) is given by

$$\mathbf{A}\mu = \mathbf{K}\mu + \lambda \mathbf{T}\mu, \tag{15}$$

where $\mathbf{K}$ is the *electrostatic integral operator* and $\mathbf{T}$ is a single-layer operator, given by

$$\mathbf{K}\mu(x) := \frac{1}{2\pi}\int_{\partial D} \mu(y)\frac{\partial \frac{1}{|x-y|}}{\partial n(x)}ds(y), \quad x \in \partial D, \tag{16}$$

$$\mathbf{T}\mu(x) := \frac{1}{2\pi}p(x)\int_{\partial D}\frac{\mu(y)}{|x-y|}ds(y), \quad x \in \partial D. \tag{17}$$

Hence, $\langle \mathbf{A}\mu,\varepsilon\rangle = \langle \mu, \mathbf{A}'\varepsilon\rangle$, for $\mu,\varepsilon \in C(\partial D)$.

Because of the Fredholm alternative we have to study the homogeneous adjoint equation

$$\mu - \mathbf{A}\mu = 0, \quad \mu \in C(\partial D). \tag{18}$$

If (18) has only the trivial solution $\mu = 0$, then (8) is uniquely solvable for arbitrary $k \in C(\partial D)$. If there exist nontrivial solutions $\mu_1, \mu_2, \ldots, \mu_n$ of (18), integral equation (8) is solvable for k that satisfies

$$\langle -\mathbf{S}k, \mu_j\rangle = 0, \quad j = 1,2,\ldots,n. \tag{19}$$

In this case the solution of (8) is given as sum of a special (inhomogeneous) solution and an arbitrary linear combination of $\mu_1, \mu_2, \ldots, \mu_n$.

III. THE STOKES INTEGRAL OPERATOR FOR THE TRIAXIAL ELLIPSOID

Now we study the homogeneous equation (18) for the case of ∂D being a triaxial ellipsoid with semi-axes $a > b > c$. We derive both eigenvalues and eigenfunctions of $\mathbf{A}$ in terms of Lamé functions and surface ellipsoidal harmonics.

In ellipsoidal coordinates (ρ,μ,ν) (see the Appendix), ∂D is given as parametric surface $\rho = 0$. From Ref. 6, we have the following result.

The functions $p(x)\mathbb{M}_n^m(x)$, $x \in \partial D$, with $p(x) = (n(x),x)$ and $\mathbb{M}_n^m$ from (A9) are eigenfunctions of $\mathbf{K}$ with eigenvalues

$$k_n^m := \frac{2abc}{2n+1}[L_n^m]'(0)K_n^m(0) - 1, \tag{20}$$

say

$$\mathbf{K}(p\mathbb{M}_n^m) = k_n^m\, p\mathbb{M}_n^m, \tag{21}$$

for $n \in \mathbb{N}$, $m = 0,\ldots,2n$. Note that each k_n^m has multiplicity one.

Again, by Ref. 6, it is seen that the eigenvalues and eigenfunctions of $\mathbf{K}'$ are given by

$$\mathbf{K}'(\mathbb{M}_n^m)=k_n^m\ \mathbb{M}_n^m, \tag{22}$$

for $n\in\mathbb{N}$, $m=0,...,2n$, and k_n^m from (20). From the completeness of the $\mathbb{M}_n^m$ in $L^2(\partial D)$, it follows that there are no other eigenfunctions of $\mathbf{K}'$ and $\mathbf{K}$ in $C(\partial D)$.

For $\mathbf{T}$ from (17) we have the following.

Lemma 1: The functions $p\mathbb{M}_n^m$ are eigenfunctions of $\mathbf{T}$,

$$\mathbf{T}(p\mathbb{M}_n^m)=t_n^m p\mathbb{M}_n^m, \tag{23}$$

with eigenvalues

$$t_n^m:=\frac{abc}{2n+1}L_n^m(0)K_n^m(0)\neq 0, \tag{24}$$

for $n\in\mathbb{N}$, $m=0,...,2n$.

Proof: For given n,m consider

$$u(\rho,\mu,\nu):=\begin{cases} u_o(\rho,\mu,\nu)=L_n^m(0)\ \mathbb{K}_n^m(\rho,\mu,\nu), & \rho\leqslant 0,\\ u_i(\rho,\mu,\nu)=K_n^m(0)\ \mathbb{L}_n^m(\rho,\mu,\nu), & \rho>0,\end{cases}$$

with $\mathbb{K}_n^m,\mathbb{L}_n^m$ from (A5), (A7) which is continuous in $\mathbb{R}^3$ and harmonic in $\mathbb{R}^3\backslash\partial D$. Therefore u can be written as a single-layer potential,

$$u(x)=\frac{1}{4\pi}\int_{\partial D}\frac{\sigma(y)}{|x-y|}ds(y),\quad x\in\mathbb{R}^3,$$

with $\sigma\in C(\partial D)$. From the jump relation for the normal derivative of u on ∂D, we obtain by use of (A4)

$$-\sigma(x)=\frac{\partial u_o(x)}{\partial n(x)}-\frac{\partial u_i(x)}{\partial n(x)}=2p(0,\mu,\nu)\mathbb{M}_n^m(\mu,\nu)(L_n^m(0)[K_n^m]'(0)-K_n^m(0)[L_n^m]'(0)),$$

with $x=x(0,\mu,\nu)$. From the Wronskian

$$L_n^m(0)[K_n^m]'(0)-K_n^m(0)[L_n^m]'(0)=-\frac{2n+1}{abc},$$

we conclude

$$\sigma(\mu,\nu)=2(2n+1)\frac{p(0,\mu,\nu)}{abc}\mathbb{M}_n^m(\mu,\nu).$$

Regarding the definition of $\mathbf{T}$ (17), the assertion follows immediately. From (A8) we see that $t_n^m\neq 0$ for all n,m. □

As a consequence of Lemma 1 we observe that the eigenvalues and eigenfunctions of $\mathbf{T}'$ are given by

$$\mathbf{T}'(\mathbb{M}_n^m)=t_n^m\ \mathbb{M}_n^m, \tag{25}$$

for $n\in\mathbb{N}$, $m=0,\ldots,2n$, and t_n^m from (24).

A combination of the above results yields the spectrum of both $\mathbf{A}$ and $\mathbf{A}'$.

Theorem 1: Let $\lambda\in\mathbb{R}$. *The eigenvalues and eigenfunctions of* $\mathbf{A}=\mathbf{K}+\lambda\mathbf{T}$ *and of* $\mathbf{A}'=\mathbf{K}'+\lambda\mathbf{T}'$ *for the triaxial ellipsoid* $\partial D:\rho=0$ *are given by*

$$\mathbf{A}(p\mathbb{M}_n^m)=a_n^m\ p\mathbb{M}_n^m, \tag{26}$$

and

$$\mathbf{A}'(\mathbb{M}_n^m) = a_n^m\ \mathbb{M}_n^m, \tag{27}$$

with

$$a_n^m = k_n^m + \lambda t_n^m = \frac{2abc}{2n+1}[L_n^m]'(0)K_n^m(0) - 1 + \lambda \frac{abc}{2n+1} L_n^m(0)K_n^m(0), \tag{28}$$

for $n \in \mathbb{N}$, $m = 0, \ldots, 2n$.

Now we discuss the consequences of Theorem 1 for the homogeneous adjoint integral equation (18). The eigenvalues and eigenfunctions of $I - \mathbf{A}$ are given by

$$(I - \mathbf{A})(p\mathbb{M}_n^m) = (1 - a_n^m)p\mathbb{M}_n^m, \tag{29}$$

for $n \in \mathbb{N}$, $m = 0, \ldots, 2n$.

In order to guarantee solvability of (8) in view of Fredholm's alternative, λ has **either** to be such that

(i) $a_n^m \neq 1$ for all n, m,

or

(ii) there are some n, m with $a_n^m = 1$. For those n, m, the right-hand side of (8) has to be such that

$$\langle \mathbf{S}k, p\mathbb{M}_n^m \rangle = 0. \tag{30}$$

By use of Fubini's theorem and Lemma 1 [note that $\mathbf{S}k = (1/p)\mathbf{T}k$], this reads as

$$\langle \mathbf{S}k, p\mathbb{M}_n^m \rangle = \int_{\partial D} k(y) \frac{1}{2\pi} \int_{\partial D} \frac{p(x)\mathbb{M}_n^m(x)}{|x-y|} ds(x) ds(y) = t_n^m \int_{\partial D} k(y)\mathbb{M}_n^m(y) ds(y),$$

and because of $t_n^m \neq 0$ this is equivalent to

$$\langle k, \mathbb{M}_n^m \rangle = 0. \tag{31}$$

A very important (degenerate) case of the above analysis is that of D being an oblate spheroid since the Earth's surface can be approximated by a suitable spheroid.

In oblate spheroidal coordinates (ξ, η, ϕ) with parameter $d > 0$ (see the Appendix), ∂D is given as parametric surface $\xi = \xi_0 > 0$.

For the eigenvalues and eigenfunctions of $\mathbf{K}$ and $\mathbf{K}'$ we have

$$\mathbf{K}(p\mathbb{Y}_n^{\pm m}) = k_n^m\ p\mathbb{Y}_n^{\pm m} \quad \text{and} \quad \mathbf{K}'(\mathbb{Y}_n^{\pm m}) = k_n^m\ \mathbb{Y}_n^{\pm m}, \tag{32}$$

with eigenvalues

$$k_n^m := -1 - 2(-1)^m \frac{(n-m)!}{(n+m)!}(\xi_0^2 + 1) Q_n^m(i\xi_0)[P_n^m]'(i\xi_0), \tag{33}$$

for $n \in \mathbb{N}$, $m = 0, \ldots, n$ (see Ahner Refs. 7 and 6). The $\mathbb{Y}_n^m(\eta, \phi)$ are the surface spherical harmonics from (A16) and $p(\xi_0, \eta)$ is given by (A12). Analogously, we obtain

$$\mathbf{T}(p\mathbb{Y}_n^{\pm m}) = t_n^m p\mathbb{Y}_n^{\pm m} \quad \text{and} \quad \mathbf{T}'(\mathbb{Y}_n^{\pm m}) = t_n^m\ \mathbb{Y}_n^{\pm m} \tag{34}$$

with eigenvalues

$$t_n^m := (-1)^m \frac{(n-m)!}{(n+m)!} \frac{id^2}{2} \xi_0(\xi_0^2 + 1)\ Q_n^m(i\xi_0) P_n^m(i\xi_0). \tag{35}$$

Note that k_n^0, t_n^0 have multiplicity one while k_n^m, t_n^m, $m > 0$, have multiplicity two.

Hence

$$\mathbf{A}(p\mathbb{Y}_n^{\pm m}) = a_n^m\ p\mathbb{Y}_n^{\pm m} \quad \text{and} \quad \mathbf{A}'(\mathbb{Y}_n^{\pm m}) = a_n^m\ \mathbb{Y}_n^{\pm m}, \tag{36}$$

with

$$a_n^m = k_n^m + \lambda t_n^m$$
$$= -1 + (-1)^m \frac{(n-m)!}{(n+m)!}(\xi_0^2+1)\ Q_n^m(i\xi_0)\left\{ i\frac{\lambda d^2\xi_0}{2}\ P_n^m(i\xi_0) - 2[P_n^m]'(i\xi_0)\right\}, \qquad (37)$$

for $n \in \mathbb{N}$, $m = 0,...,n$.

The eigenvalues and eigenfunctions of $I - \mathbf{A}$ are given by

$$(I-\mathbf{A})(p\mathbb{Y}_n^{\pm m}) = (1-a_n^m)p\mathbb{Y}_n^{\pm m}, \qquad (38)$$

for $n \in \mathbb{N}$, $m = 0,...,n$.

The condition for λ and k in view of solvability of (8) is the following: λ has **either** to be such that

(i) $a_n^m \neq 1$ for all n,m, i.e.,

$$\lambda \neq -\frac{2i}{d^2\xi_0}\left\{ 2(-1)^m \frac{(n+m)!}{(n-m)!}\frac{1}{(\xi_0^2+1)Q_n^m(i\xi_0)} + 2[P_n^m]'(i\xi_0)\right\}, \qquad (39)$$

or

(ii) there are some n,m with $a_n^m = 1$. For those n,m, the right-hand side $-\mathbf{S}k$ has to be such that

$$\langle \mathbf{S}k, p\mathbb{Y}_n^{\pm m}\rangle = 0. \qquad (40)$$

This reads as

$$\langle \mathbf{S}k, p\mathbb{Y}_n^{\pm m}\rangle = \int_{\partial D} k(y)\frac{1}{2\pi}\int_{\partial D}\frac{p(x)\overline{\mathbb{Y}_n^{\pm m}}(x)}{|x-y|}ds(x)ds(y) = t_n^m \int_{\partial D} k(y)\overline{\mathbb{Y}_n^{\pm m}}(y)ds(y),$$

and because of $t_n^m \neq 0$ it is equivalent to

$$\langle k, \mathbb{Y}_n^{\pm m}\rangle = 0. \qquad (41)$$

IV. APPLICATION

The classical ellipsoidal Stokes problem which occurs in geoid height determination is BVP (1)–(3) for an oblate spheroid with

$$\lambda = \frac{8}{d^2\xi_0^2}. \qquad (42)$$

In the geodetic literature (see, e.g., Ref. 1) the boundary condition (2) is often written in the form

$$\frac{\partial u}{\partial \xi}(x) + \frac{2}{\xi_0}u(x) = -f(x), \quad x \in \partial D, \qquad (43)$$

with

$$f(x) = -\frac{d^2\xi_0}{4}\frac{k(x)}{p(x)}, \quad x \in \partial D. \qquad (44)$$

The eigenvalues of the integral operator $I - \mathbf{A}$ are given by

$$1 - a_n^m = 2 + (-1)^{m+1}\frac{(n-m)!}{(n+m)!}(\xi_0^2+1)\ Q_n^m(i\xi_0)\left\{\frac{4i}{\xi_0}P_n^m(i\xi_0) - 2[P_n^m]'(i\xi_0)\right\}.$$

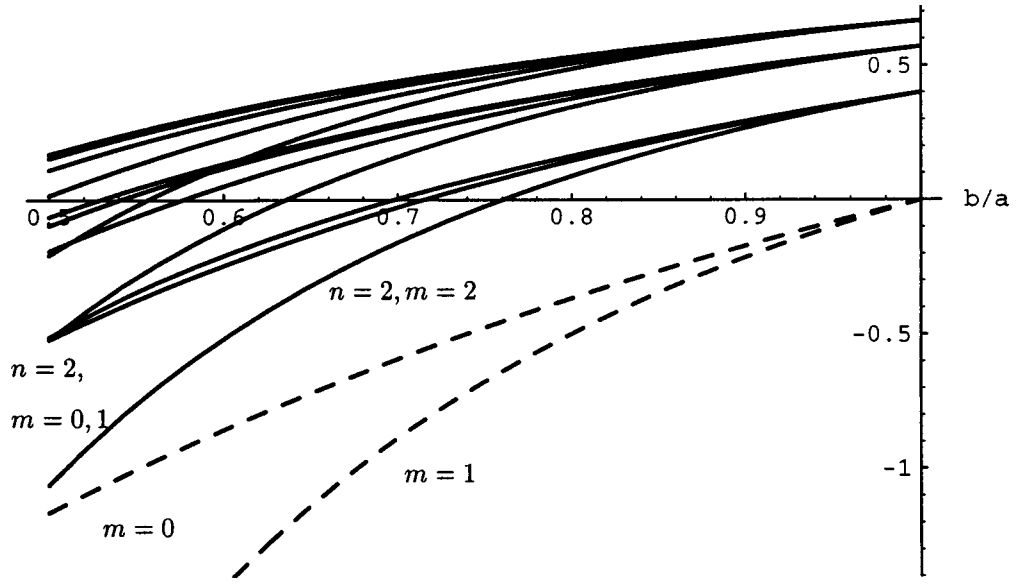

FIG. 1. Eigenvalues of $I-\mathbf{A}$ for $n=1$ (dashed) and $n=2,3,4$.

Figure 1 depicts some eigenvalues $1-a_n^m$. In view of the Stokes problem for a spheroid close to the Earth's surface we are interested in the behavior of the eigenvalues for $4/3<\xi_0<\infty$, i.e., for the axes-ratios $0{,}8<b/a<1$.

For the case $1-a_n^m=0$ for some n,m, the right-hand side f of (43), in view of (35), has to be such that

$$\begin{aligned}\left\langle -\frac{4}{d^2\xi_0}pf,\mathbb{Y}_n^{\pm m}\right\rangle &= -\frac{4}{d^2\xi_0}\int_{\partial D}p(y)f(y)\overline{\mathbb{Y}_n^{\pm m}}(y)ds(y)\\ &= -\frac{d(\xi_0^2+1)}{2}\int_0^{2\pi}\int_{-1}^{1}f(\eta,\phi)\overline{\mathbb{Y}_n^{\pm m}}(\eta,\phi)d\eta\,d\phi\\ &= -\frac{d(\xi_0^2+1)}{2}\langle f,\mathbb{Y}_n^{\pm m}\rangle_{S^2}=0;\end{aligned}$$

hence the solvability condition is

$$\langle f,\mathbb{Y}_n^{\pm m}\rangle_{S^2}=0. \tag{45}$$

Here $\langle\cdot,\cdot\rangle_{S^2}$ denotes the usual inner product in $L^2(S^2)$ for the unit sphere S^2. As indicated in Fig. 1, $1-a_n^m=0$ may be true for spheroids with axes-ratio $b/a<0.8$ for $n\geqslant 2$, and in fact is true for the sphere for $n=1$.

In addition, the condition

$$u(x)=\frac{c}{|x|}+O\left(\frac{1}{|x|^3}\right),\quad |x|\to\infty, \tag{46}$$

with arbitrary $c\in\mathbb{R}$ is imposed. This effect is that f has to satisfy

$$\langle f,\mathbb{Y}_1^j\rangle_{S^2}=0,\quad j=-1,0,1, \tag{47}$$

in order to guarantee solvability of the integral equation for μ.

V. CONCLUSIONS

An integral equation formulation for the ellipsoidal Stokes problem has been achieved by the nullfield method. From explicit expressions for the eigenvalues and eigenfunctions for $(I-\mathbf{A}')$ in the case of the oblate spheroid we have seen that, apart from (47), further assumptions (45) for the inhomogeneity have possibly to be imposed for the case that the axes-ratio is less than 0,8 in order to attain uniqueness of the solution and hence well-posedness of the problem. The results are in agreement with the numerical treatment presented in Ref. 1. Further work to be done is the numerical treatment of integral equation (8) for which the foundation has been presented. This can be done either by boundary element techniques such as Galerkin schemes or collocation methods. An alternative approach are spectral methods which are applicable for this problem since we have a complete set of eigenfunctions for the boundary integral operator.

APPENDIX: ELLIPSOIDAL AND OBLATE SPHEROIDAL COORDINATES

1. Ellipsoidal coordinates and Lamé-functions

The equation of a triaxial ellipsoid D with semi-axes $a>b>c$ is given by

$$D: \frac{x_1^2}{a^2}+\frac{x_2^2}{b^2}+\frac{x_3^2}{c^2}\leqslant 1. \tag{A1}$$

The *ellipsoidal coordinates* of $x(x_1,x_2,x_3)$ are defined as the three roots ρ, μ, ν of the cubic equation in θ,

$$\frac{x_1^2}{a^2+\theta}+\frac{x_2^2}{b^2+\theta}+\frac{x_3^2}{c^2+\theta}=1, \tag{A2}$$

for fixed values of x_1, x_2, x_3, where

$$\rho\in[-c^2,\infty), \quad \mu\in[-b^2,-c^2], \quad \nu\in[-a^2,-b^2]. \tag{A3}$$

The three families of surfaces $\rho=$const (ellipsoids), $\mu=$const (hyperboloids of one sheet) and $\nu=$const (hyperboloids of two sheets) form a triply-orthogonal coordinate system in $\mathbb{R}^3$. ∂D is given by $\rho=0$, points with $\rho<0$ lie in $\underline{D}$, those with $\rho>0$ lie in $\mathbb{R}^3\backslash D$. For a differentiable function u, the normal derivative on ∂D is given by

$$\frac{\partial u(0,\mu,\nu)}{\partial n(\mu,\nu)}=2\,p(0,\mu,\nu)\,\frac{\partial u(\rho,\mu,\nu)}{\partial \rho}\bigg|_{\rho=0}, \quad p(0,\mu,\nu)=\frac{abc}{\sqrt{\mu\nu}}, \tag{A4}$$

with the *support function* $p(x):=(n(x),x)$, $x\in\partial D$ of the ellipsoid.

The *interior ellipsoidal harmonics of order* $n\in\mathbb{N}$ are defined by

$$\mathbb{L}_n^m(\rho,\mu,\nu):=L_n^m(\rho)L_n^m(\mu)L_n^m(\nu), \quad m=0,\ldots,2n, \tag{A5}$$

with the Lamé-functions of the first kind L_n^m. In Cartesian coordinates the $\mathbb{L}_n^m$ turn out as harmonic polynomials of order n.

For $L_n^m(\rho)$, the corresponding function of the *second kind* is given by

$$K_n^m(\rho):=(2n+1)L_n^m(\rho)\int_\rho^\infty \frac{d\rho'}{[L_n^m(\rho')]^2\Delta(\rho')}, \quad \rho>-c^2, \tag{A6}$$

with

$$\Delta(\rho):=\sqrt{(\rho+a^2)(\rho+b^2)(\rho+c^2)}.$$

The second kind Lamé-functions generate the *exterior ellipsoidal harmonics of order* n,

$$\mathbb{K}_n^m(\rho,\mu,\nu) := K_n^m(\rho)L_n^m(\mu)L_n^m(\nu), \quad m=0,\ldots,2n, \tag{A7}$$

which are regular at infinity.

A Theorem due to Stieltjes implies that the zeros of $L_n^m(\lambda)$ are simple and contained in $(-a^2,-c^2)$ (see Ref. 8). Hence we have

$$L_n^m(0)K_n^m(0)\neq 0, \quad n\in\mathbb{N}, \quad m=0,\ldots,2n. \tag{A8}$$

The system of *ellipsoidal surface harmonics of order* $n\in\mathbb{N}$, defined by

$$\mathbb{M}_n^m(\mu,\nu) := L_n^m(\mu)L_n^m(\nu), \quad m=0,\ldots,2n, \tag{A9}$$

forms a complete orthogonal system in $L^2(\partial D)$.

2. Oblate spheroidal coordinates

For the oblate spheroid D with semi-axes $a>b$ we introduce the system of *oblate spheroidal coordinates* (ξ,η,ϕ) by

$$\begin{aligned} x_1 &= \frac{d}{2}[(\xi^2+1)(1-\eta^2)]^{1/2}\cos\phi, \\ x_2 &= \frac{d}{2}[(\xi^2+1)(1-\eta^2)]^{1/2}\sin\phi, \\ x_3 &= \frac{d}{2}\xi\eta, \end{aligned} \tag{A10}$$

with $\xi\in[0,\infty)$, $\eta\in[-1,1]$, $\phi\in[0,2\pi]$ and parameter $d>0$. The surfaces $\xi=\text{constant}>0$ represent oblate spheroids with semi-axes $a=(d/2)(\xi^2+1)^{1/2}$, $b=(d/2)\xi$. The degenerate case $\xi=0$ corresponds to the circular disk of radius $r=d/2$ which lies in the x_1/x_2-plane and is centered at the origin. Hence ∂D is given by

$$\xi=\xi_0=\left(\frac{a^2}{b^2}-1\right)^{-1/2}. \tag{A11}$$

For $p(x)=(n(x),x)$ we have

$$p(\xi,\eta)=\frac{d}{2}\xi\left(\frac{\xi^2+1}{\xi^2+\eta^2}\right)^{1/2}. \tag{A12}$$

The normal derivative of a differentiable function u on $\xi=\text{const}$ is given by

$$\frac{\partial u}{\partial n}=\frac{4}{d^2\xi}\,p(\xi,\eta)\,\frac{\partial u}{\partial \xi}, \tag{A13}$$

and for the surface element on $\xi=\text{const}$ we have

$$ds=\left(\frac{d}{2}\right)^3\xi(\xi^2+1)\frac{1}{p(\xi,\eta)}\,d\eta\,d\phi. \tag{A14}$$

The *interior* and *exterior spheroidal harmonics* are defined by

$$\mathbb{P}_n^m(\xi,\eta,\phi)=P_n^m(i\xi)\,\mathbb{Y}_n^m(\eta,\phi) \quad \text{and} \quad \mathbb{Q}_n^m(\xi,\eta,\phi)=Q_n^m(i\xi)\,\mathbb{Y}_n^m(\eta,\phi), \tag{A15}$$

where

$$\mathbb{Y}_n^m(\eta,\phi)=P_n^{|m|}(\eta)\ e^{im\phi} \tag{A16}$$

denote the *surface spherical harmonics* of order n and degree m,$n\in\mathbb{N}$, $m=-n,\ldots,n$. They form a complete orthogonal system in the Hilbert space $L^2(\partial D)$ with inner product

$$(u,v)=\int_{\partial D}p(y)\ u(y)\overline{v(y)}ds(y). \tag{A17}$$

[1] Z. Martinec and E. W. Grafarend, ''Solution to the Stokes boundary-value problem on an ellipsoid of revolution,'' Stud. Geoph. Geod. **41**, 103–129 (1997).
[2] J. Otero and J. Capdevila, ''A series solution for Zagrebin's problem,'' in *Geodetic Theory Today*, Int. Assoc. of Geodesy Symposia, Symposium No. 114, edited by F. Sanso (Springer-Verlag, Berlin, 1995), pp. 280–293.
[3] J. Otero, ''A uniqueness theorem for a Robin boundary value problem of physical geodesy,'' Quart. Appl. Math. (to appear).
[4] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory* (Wiley, New York, 1983).
[5] E. Martensen, *Potentialtheorie* (Teubner, Stuttgart, 1968).
[6] S. Ritter, ''The spectrum of the electrostatic integral operator for an ellipsoid,'' in *Inverse Scattering and Potential Problems in Mathematical Physics*, Methoden und Verfahren der Mathematischen Physik 40, edited by R. Kleinman, R. Kress, and E. Martensen (Peter Lang, Frankfurt a. M./Bern, 1994), pp. 157–167.
[7] J. F. Ahner, ''On the eigenvalues of the electrostatic integral operator II,'' J. Math. Anal. Appl. **181**, 328–334 (1994).
[8] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics* (Chelsea, New York, 1955).

Solutions of $(2+1)$-dimensional spin systems

Z. Thomova and P. Winternitz
Centre de Recherches Mathématiques, Université de Montréal, CP 6128, Succ. Centre Ville, Montréal, Quebec H3C 3J7, Canada

W. J. Zakrzewski
Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, United Kingdom

(Received 27 January 1997; accepted for publication 26 March 1998)

We use the methods of group theory to reduce the equations of motion of two spin systems in $(2+1)$ dimensions to sets of coupled ordinary differential equations. We present solutions of some classes of these sets and discuss their physical significance. © *1998 American Institute of Physics.* [S0022-2488(98)00207-2]

I. INTRODUCTION

In this paper we look for solutions of the equations of the Landau–Lifshitz model (with, perhaps, nonvanishing anisotropy) and of a nonlinear vector diffusion equation. The equations are given, respectively, by

$$\frac{\partial \boldsymbol{\phi}}{\partial t} = \boldsymbol{\phi} \times \mathbf{F} \tag{1.1}$$

and

$$\frac{\partial \boldsymbol{\phi}}{\partial t} = \mathbf{F} - \boldsymbol{\phi}(\boldsymbol{\phi} \cdot \mathbf{F}), \tag{1.2}$$

where $\mathbf{F}$ is given by

$$\mathbf{F} = \Delta \boldsymbol{\phi} + (A\phi_3 + B)\mathbf{e}_3, \tag{1.3}$$

where $\mathbf{e}_3$ is a unit vector in the third direction in the $\boldsymbol{\phi}$ space and $\boldsymbol{\phi}$ satisfies $\boldsymbol{\phi} \cdot \boldsymbol{\phi} = 1$. A and B are possible anisotropy coefficients.

The motivation for this work comes from the original observation made by Landau and Lifshitz[1] in their study of the ferromagnetic continuum. They pointed out that for phenomena for which substantial spatial variations occur only over a large number of lattice spacings, we can use the continuum approximation. They showed that a ferromagnetic medium is characterized by the magnetization vector $\mathbf{M}$ (like the vector $\boldsymbol{\phi}$ above) that precesses around the effective magnetic field and so obeys, what is now called the Landau–Lifshitz equation, namely (1.1). Since the original work of Landau and Lifshitz, many papers have been written on the subject,[2] and the equation has been modified by the inclusion of various additional terms to $\mathbf{F}$. It has been used to describe the dynamics of magnetic bubbles in a ferromagnetic continuum and also of vortices in He II or in a superconductor.[2] Various studies of the dynamics of such topological soliton-like structures have been performed both theoretically and experimentally,[3,4] and they have exhibited many interesting, and perhaps unexpected, phenomena—like the skew deflection of these structures under the influence of a magnetic field gradient that resembles the more familiar Hall motion of electrons in external magnetic and electric fields.[5]

A recent work of Papanicolaou and Tomaras,[2] as well as some earlier work of other people,[6] has shown that many experimentally observed facts can indeed be explained using the Landau–Lifshitz equation. Much of the work involved deriving various conserved quantities describing these structures and then using them to restrict the description of the dynamics. All this work has provided further evidence as to the relevance of the Landau–Lifshitz equation to the description of

 © 1998 American Institute of Physics

physical phenomena. However, as the Landau–Lifshitz equation is quite complicated, only some results were obtained in an analytical form. Most more recent studies[7] involved numerical simulations.

The vector nonlinear diffusion equation (1.2) has less obvious physical applications, but it has been used[8] in the study of phase ordering kinetics, where one investigates the time evolution of a system quenched from the disordered into an ordered phase. This topic has attracted considerable attention in recent years.[9] In fact, it has been shown that many features of phase ordering in systems supporting topologically stable defects [for example, in systems described by the $O(N)$ vector model in d dimensions with $d \leqslant N$[10,11]], or in two- and three-dimensional nematic liquid crystals[12] can be understood theoretically by investigating the dynamics of the numerous topological defects generated during the quench. A special and interesting case is that of the $O(3)$ model system in two spatial dimensions. It supports topologically stable, but nonsingular objects, which, in the condensed matter community language, are called topological textures. Such systems were studied numerically in Ref. 8.

Given the paucity of analytical results for both Eqs. (1.1) and (1.2) (especially involving the dynamics) one of our aims in this paper is to see what time-dependent solutions can be found using the group theoretical method of symmetry reduction.[13–15] This method exploits the symmetry of the original equations to find solutions invariant under some subgroup (the classic example one can give here involves seeking solutions in three dimensions that are rotationally invariant). The method puts all such attempts on a unified footing and it has been applied with success to many equations.[16] The method gives equations whose solutions represent specific solutions of the full equations; the solutions are determined locally and the method does not tell us whether these solutions are stable or not with respect to any perturbations.

In a recent paper,[17] two of us (PW and WJZ), together with Grundland, have applied this technique to looking for solutions of the relativistic CP^1 model.

In this paper we investigate solutions of (1.1) and (1.2). We are particularly interested in time-dependent solutions; all time-independent solutions of (1.1) and (1.2) (when there is no anisotropy) are also the time-independent solutions of the relativistic model and so can be found in Ref. 17.

Like in the relativistic CP^1 model studied before, in order to perform the symmetry reductions, we have to decide what variables to use. To avoid having to use the constrained variables ($\boldsymbol{\phi}$) it is convenient to use the W formulation of the model that involves the stereographic projection of the sphere $\boldsymbol{\phi}\cdot\boldsymbol{\phi}=1$ onto the complex plane. In this formulation instead of using the $\boldsymbol{\phi}$ fields, we express all the dependence on $\boldsymbol{\phi}$ in terms of their stereographic projection onto the complex plane W. The $\boldsymbol{\phi}$ fields are then related to W by

$$\phi_1=\frac{W+W^*}{1+|W|^2}, \quad \phi_2=i\,\frac{W-W^*}{1+|W|^2}, \quad \phi_3=\frac{1-|W|^2}{1+|W|^2}. \tag{1.4}$$

To perform our analysis it is convenient to use the polar version of the W variables; i.e., to put $W=R\,\exp iQ$ and then study the equations for R and Q. The advantage of this approach is that the equations become simple; the disadvantage comes from having to pay attention that R is real and Q should be periodic with a period of 2π. (If the period is not 2π then the solution may become multivalued.) Thus if we find solutions that do not obey these restrictions, then these solutions, however interesting they may be, cannot be treated as solutions of the original model.

In the case of the Landau–Lifshitz equation the equations for R and Q take the form

$$\partial_t R-2\,\frac{(1-R^2)}{(1+R^2)}\,(\partial_x Q\partial_x R+\partial_y Q\partial_y R)-R(\partial_{xx}Q+\partial_{yy}Q)=0 \tag{1.5}$$

and

$$\partial_t Q=B+A\,\frac{1-R^2}{1+R^2}-\frac{\partial_{xx}R+\partial_{yy}R}{R}+\frac{(1-R^2)}{(1+R^2)}\,((\partial_x Q)^2+(\partial_y Q)^2)+\frac{2}{(1+R^2)}\,((\partial_x R)^2+(\partial_y R)^2), \tag{1.6}$$

while for the diffusion case they are, respectively,

$$\partial_t Q - 2\frac{(1-R^2)}{R(1+R^2)}(\partial_x Q\partial_x R+\partial_y Q\partial_y R)-(\partial_{xx}Q+\partial_{yy}Q)=0 \tag{1.7}$$

and

$$\partial_t R+BR+AR\frac{1-R^2}{1+R^2}-\partial_{xx}R-\partial_{yy}R+\frac{(1-R^2)R}{(1+R^2)}((\partial_x Q)^2+(\partial_y Q)^2)$$

$$+\frac{2R}{(1+R^2)}((\partial_x R)^2+(\partial_y R)^2)=0. \tag{1.8}$$

Note, that, in the Landau–Lifshitz case, if we put $R=1$ the equations become $\Delta Q=0$ and $\partial_t Q=B$, which have a very simple solution, and in the diffusion case, we have to set $B=0$ and then we end up with $\partial_t Q-\Delta Q=0$ as the equation for Q. The latter case is the nonrelativistic analog of what was found in the relativistic case where $R=1$ reduced the equation for Q to the linear wave equation for the phase Q.

In the next section we determine the symmetry group of our equations (1.6), (1.5) and of (1.7) and (1.8). In the following sections we solve the derived equations and discuss their solutions.

II. THE SYMMETRY GROUP AND ITS TWO-DIMENSIONAL SUBGROUPS

The symmetry group of our systems of equations, respectively (1.6) and (1.5) and (1.7) and (1.8), can be calculated using the standard methods.[13–16] We actually made use of a MACSYMA package[18] that provides a simplified and partially solved set of determining equations.

Solving the determining equations we find that three different cases must be distinguished.

(1) $A=B=0$, i.e., the anisotropy is absent. The Landau–Lifshitz equation and the diffusion equation have isomorphic symmetry groups, consisting of translations in space and time directions, rotations in the x, y plane, dilations, and a group of $O(3)$ rotations between the components of the field $\boldsymbol{\phi}$. The corresponding Lie algebra L_1 has the structure of a direct sum,

$$L_1=s(2,1)\oplus o(3). \tag{2.1}$$

Bases for these two algebras are given by the following vector fields, acting on space–time and on the fields in the $\{R,Q\}$ realization of Eqs. (1.5)–(1.8):

$$\begin{aligned} s(2,1):\quad & P_0=\partial_t,\quad P_1=\partial_x,\quad P_2=\partial_y,\\ & L=-x\partial_y+y\partial_x,\quad D=2t\partial_t+x\partial_x+y\partial_y. \end{aligned} \tag{2.2}$$

$$\begin{aligned} o(3):\quad & X=\frac{1}{2}\left(\sin Q\left(R-\frac{1}{R}\right)\partial_Q+\cos Q(R^2+1)\partial_R\right),\\ & Y=\frac{1}{2}\left(\cos Q\left(R-\frac{1}{R}\right)\partial_Q-\sin Q(R^2+1)\partial_R\right),\\ & Z=\partial_Q. \end{aligned} \tag{2.3}$$

(2) $A\neq 0$.

The symmetry algebra for both equations is reduced to

$$L_2=\{P_0,P_1,P_2,L\}\oplus\{Z\}, \tag{2.4}$$

i.e., the dilations are absent and the only $\boldsymbol{\phi}$ rotations left are those around the third axis (i.e., around ϕ_3).

(3) $A=0$, $B\neq 0$.

The symmetry algebra for the dissipative equations (1.8) and (1.7) is still L_2, as in (2.4). That of the Landau–Lifshitz equation is

$$L_3=\{P_0,P_1,P_2,L,\tilde{D}\}\oplus\{X,Y,Z\}, \tag{2.5}$$

with

$$\tilde{D}=2t\partial_t+x\partial_x+y\partial_y+2Bt\partial_Q, \tag{2.6}$$

where Q is replaced by $Q-Bt$ in x and y, Eq. (2.3).

In order to perform symmetry reduction we need to classify the subalgebras of the symmetry algebras L_1, L_2, and L_3. We wish to reduce Eqs. (1.5)–(1.8) to ordinary differential equations. To do this, we will require that the solutions are invariant under a two-dimensional subgroup of the symmetry group. In order to do this systematically we need to derive a classification of the two-dimensional subalgebras of the symmetry algebra. Moreover, we can restrict ourselves to subalgebras, all elements of which act nontrivially on space–time, i.e., that do not contain any rotations in $\boldsymbol{\phi}$ space.

The subalgebra classification can be done in an algorithmic way;[14] the results are quite simple and we present them without a proof.

(1) $A=B=0$. Every two-dimensional subalgebra of L_1, each element of which acts nontrivially on space–time, is conjugate under the action of the group of inner automorphisms to one of the following ones:

$$\begin{aligned}
&A_{2,1}=\{P_1+aZ,P_2+bZ\},\\
&A_{2,2}=\{L+aZ,P_0+bZ\},\\
&A_{2,3}=\{P_0+aZ,P_2+bZ\},\\
&A_{2,4}=\{P_0-vP_1+aZ,P_2+bZ\},\quad v\neq 0,\\
&A_{2,5}=\{D+bL+aZ,P_0\},\quad b\neq 0,\\
&A_{2,6}=\{D+aZ,P_0\},\\
&A_{2,7}=\{D+aZ,L+bZ\},\\
&A_{2,8}=\{D+aZ,P_2\}.
\end{aligned} \tag{2.7}$$

The parameters a, b, and v are arbitrary real numbers. In some cases their ranges can be further constrained, but that is not important for our purposes.

(2) $A\neq 0$.
Every two-dimensional subalgebra of the considered type is conjugate to one listed above as $A_{2,1},\ldots,A_{2,4}$.

(3) $A=0$, $B\neq 0$.
For the dissipative equations (1.8) and (1.7) the subalgebra classes are represented by $A_{2,1},\ldots,A_{2,4}$. For the Landau–Lifshitz equations (1.5) and (1.6) they are represented by $A_{2,1},\ldots,A_{2,8}$ with D replaced by $\tilde{D}$ and P_0 replaced by $P_0=P_0+bZ$.

We can now proceed to perform various reductions. We are particularly interested in reductions that do not result in time independence as these were already studied in Ref. 17.

III. SOLUTIONS OF THE LANDAU–LIFSHITZ EQUATION

A. General procedure

Our aim is to solve the Landau–Lifshitz equations (1.5) and (1.6), using the method of symmetry reduction. This involves assuming that a solution is invariant under a subgroup G_0 of the symmetry group G, namely one of the two-dimensional groups corresponding to the algebras $A_{2,1},\ldots,A_{2,8}$ of (2.7). The assumption makes it possible to reduce the partial differential equations (1.5) and (1.6) to a pair of coupled ordinary differential equations. Whenever possible, we decouple them and find explicit solutions for the functions R and Q, hence for W, and finally for the vector $\vec{\phi}$ figuring in (1.1).

For all eight algebras in (2.7) the invariant solution will have the form

$$R(x,y,t)=R(\xi),\quad Q(x,y,t)=\alpha(\xi)+\beta(x,y,t), \tag{3.1}$$

where ξ and β are explicitly given and $R(\xi)$ and $\alpha(\xi)$ satisfied coupled ordinary differential equations obtained by substituting (3.1) into (1.5) and (1.6).

The reduced equation (1.5) is

$$(\nabla\xi)^2\alpha_{\xi\xi}+\left[2\frac{(1-R^2)}{R(1+R^2)}(\nabla\xi)^2R_\xi+\Delta\xi\right]\alpha_\xi=\frac{R_\xi}{R}\xi_t-2\frac{(1-R^2)}{R(1+R^2)}R_\xi(\nabla\xi,\nabla\beta)-\Delta\beta. \tag{3.2}$$

For algebra $A_{2,1}$ we have

$$\nabla\xi^2=\Delta\xi=\Delta\beta=0,\quad \xi_t=1, \tag{3.3}$$

and so (3.2) reduces to $R_\xi=0$.

In all other cases we have $(\nabla\xi)^2\neq 0$. Equation (3.2) is a first-order linear inhomogeneous equation for α_ξ. We can integrate it explicitly and obtain α_ξ in terms of R, whenever the functions ξ and β satisfy

$$\frac{d}{d\xi}\left(\frac{h}{(\nabla\xi)^2}\right)=0,\quad \frac{d}{d\xi}\left[h\frac{(\nabla\xi,\nabla\beta)}{(\nabla\xi)^2}\right]-\frac{h\Delta\beta}{(\nabla\xi)^2}=0, \tag{3.4}$$

where

$$h(\xi)=1,\quad \text{for } \Delta\xi=0,$$

$$\frac{h'(\xi)}{h(\xi)}=\frac{\Delta\xi}{(\nabla\xi)^2},\quad \text{for } \Delta\xi\neq 0. \tag{3.5}$$

Conditions (3.4) are always satisfied for the algebras $A_{2,2}\ldots,A_{2,6}$, not, however, for $A_{2,7}$ and $A_{2,8}$. When conditions (3.4) are satisfied, we can integrate Eq. (3.2) once to obtain

$$\alpha_\xi=\frac{S}{h}\frac{(1+R^2)^2}{R^2}+\mu\frac{1+R^2}{R^2}+\nu, \tag{3.6}$$

where S is an arbitrary real integration constant, and where we have

$$\begin{aligned}
&\mu=-\frac{v}{2}, && \nu=0, && \text{for } A_{2,4},\\
&\mu=0, && \nu=-\frac{a}{b^2+1}, && \text{for } A_{2,5},\\
&\mu=0, && \nu=\frac{a\xi}{1+\xi^2}, && \text{for } A_{2,6},\\
&\mu=0, && \nu=0, && \text{for } A_{2,2},A_{2,3}.
\end{aligned} \tag{3.7}$$

Equation (1.6) for algebras $A_{2,2},\ldots,A_{2,8}$ is reduced to a second-order differential equation for $R(\xi)$ that also involves $\alpha_\xi(\xi)$. For reductions corresponding to Lie algebras $A_{2,2},\ldots,A_{2,6}$, we can substitute α_ξ from (3.6), to obtain an ordinary differential equation for $R(\xi)$ alone. To transform this equation to a standard form we put

$$R(\xi)=\sqrt{-U(\eta)},\quad \eta=\int h^{-1}(\xi)d\xi. \tag{3.8}$$

The equation for $U(\eta)$ is then written as

$$U_{\eta\eta}=\left(\frac{1}{2U}+\frac{1}{U-1}\right)U_\eta^2-\frac{2S^2}{U}(U+1)(U-1)^3+p\frac{U(U+1)}{U-1}+qU+m(U-1)^2. \tag{3.9}$$

Equation (3.9) can be integrated in terms of elliptic functions if p, q, and m are constants. This is always the case for algebras $A_{2,3},\ldots,A_{2,6}$. In the case of algebra $A_{2,2}$ this is true if we set $A=0$, $B=b$.

Equation (3.9) has a first integral that we can write as

$$U_\eta^2=-4S^2U^4+K_1U^3+KU^2+K_2U+K_3, \tag{3.10}$$

where K is an integration constant, and the constants K_1, K_2, and K_3 are related to the coefficients S, p, q, and m in (3.9).

In this article we restrict ourselves to solutions of the Landau–Lifshitz equation that are obtained by solving (3.10).

We shall first discuss solutions of (3.10) in general, then run through algebras $A_{2,2},\ldots,A_{2,6}$ and specify the values of the coefficients in (3.10) in each case, as well as the independent variable η.

Algebra $A_{2,1}$ leading to a first-order equation, will be treated separately.

B. Solutions of the elliptic function equation

We shall call (3.10) the ''elliptic function equation.'' Its solutions are, of course, well known.[19] We shall, however, list those that are relevant in the context of solving (3.9), and more importantly, the Landau–Lifshitz equation.

Several comments are in order here.

(1) The functions $R(\eta)$ must be real (and non-negative); hence $U(\eta)$ must be real and nonpositive.
(2) For $S\neq 0$ the coefficient of the highest power of U in (3.10) is non-negative. This means that all real solutions of (3.10) are nonsingular.
(3) For $S=0$, $K_1\neq 0$ in (3.10) the real solutions of (3.10) can be singular. Since we are really interested in the fields ϕ_i we note that singular solutions of U will give regular functions ϕ_i.
(4) In general, Eq. (3.10) is solved in terms of Jacobi elliptic functions. However, these reduce to elementary functions whenever the polynomial on the right-hand side has multiple roots, or when $S=K_1=0$.

Let us run through individual cases.

I. $S\neq 0$

We rewrite (3.10) as

$$U_\eta^2=-4S^2(U-U_1)(U-U_2)(U-U_3)(U-U_4). \tag{3.11}$$

(1) $U_1\leqslant U\leqslant U_2=U_3=U_4<0$,

$$U(\eta)=U_2-\frac{U_2-U_1}{1+S^2(U_2-U_1)^2(\eta-\eta_0)^2}, \tag{3.12}$$

this is an algebraic solitary wave, equal to U_2 for $\eta\to\pm\infty$, and dipping down to U_1 for $\eta=\eta_0$.

(2) $U_1=U_2=U_3<U\leqslant U_4\leqslant 0$,

$$U(\eta)=U_1+\frac{U_4-U_1}{1+S^2(U_4-U_1)^2(\eta-\eta_0)^2}. \tag{3.13}$$

Also an algebraic solitary wave, rising to $U=U_4$ for $\eta=\eta_0$, equal to U_1 for $\eta\to\pm\infty$.

(3) $U_1\leqslant U<U_2=U_3<U_4$, $U_2\leqslant 0$,

$$U(\eta)=U_2-\frac{(U_4-U_2)(U_2-U_1)}{(U_4-U_1)\cosh^2\,\mu(\eta-\eta_0)-(U_2-U_1)},$$

$$\mu=S\sqrt{(U_4-U_2)(U_2-U_1)}. \tag{3.14}$$

(4) $U_1<U_2=U_3<U\leqslant U_4\leqslant 0$,

$$U(\eta)=U_3+\frac{(U_3-U_1)(U_4-U_3)}{(U_4-U_1)\cosh^2\ \mu(\eta-\eta_0)-(U_4-U_3)}, \tag{3.15}$$

with μ as in (3.14).

The last two solutions are solitons: the first one a well, the second a bump.

(5) $U_1\leqslant U\leqslant U_2<U_3=U_4,\ U_2\leqslant 0$,

$$U(\eta)=U_4-\frac{(U_4-U_2)(U_4-U_1)}{(U_2-U_1)\sin^2\ \mu(\eta-\eta_0)+U_4-U_2},$$

$$\mu=S\sqrt{(U_4-U_2)(U_4-U_1)}. \tag{3.16}$$

(6) $U_1=U_2<U_3\leqslant U\leqslant U_4\leqslant 0$,

$$U(\eta)=U_1+\frac{(U_4-U_1)(U_3-U_1)}{(U_4-U_3)\sin^2\ \mu(\eta-\eta_0)+U_3-U_1},$$

$$\mu=S\sqrt{(U_4-U_1)(U_3-U_1)}. \tag{3.17}$$

(7) $U_1\leqslant U\leqslant U_2<U_3<U_4,\ U_2\leqslant 0$,

$$U(\eta)=U_4-\frac{(U_4-U_2)(U_4-U_1)}{(U_2-U_1)sn^2(\mu(\eta-\eta_0),k)+U_4-U_2},$$

$$\mu=S\sqrt{(U_4-U_2)(U_3-U_1)},\quad k^2=\frac{(U_4-U_3)(U_2-U_1)}{(U_4-U_2)(U_3-U_1)}. \tag{3.18}$$

(8) $U_1<U_2<U_3\leqslant U\leqslant U_4<0$,

$$U(\eta)=U_1+\frac{(U_4-U_1)(U_3-U_1)}{(U_4-U_3)sn^2\ \mu(\eta-\eta_0)+U_3-U_1}, \tag{3.19}$$

with k^2 and μ as in (3.18).

(9) $U_1\leqslant U\leqslant U_2\leqslant 0$, $U_{3,4}=p\pm iq$, $q>0$,

$$U(\eta)=\frac{(MU_1-NU_2)cn(\mu(\eta-\eta_0),k)+MU_1+NU_2}{(M-N)cn(\mu(\eta-\eta_0),k)+M+N},$$

$$M^2=(U_2-p)^2+q^2,\quad N^2=(U_1-p)^2+q^2,$$

$$k^2=\frac{(U_2-U_1)^2-(M-N)^2}{4MN},\quad \mu=2S\sqrt{MN}. \tag{3.20}$$

Solutions (3.16)–(3.20) are periodic. All the elementary solutions can be viewed as limits of solutions (3.18), (3.19), and (3.20).

II. $S=0$, $K_1\neq 0$

Set

$$\mu=\tfrac{1}{2}\sqrt{|K_1|(U_3-U_1)}. \tag{3.21}$$

(1) $K_1<0$, $U_1=U_2<U\leqslant U_3\leqslant 0$,

$$U=U_3-(U_3-U_2)\tanh^2\ \mu(\eta-\eta_0). \tag{3.22}$$

(2) $K_1<0$, $U<U_1=U_2<U_3$, $U_1\leqslant 0$,

$$U=U_3-\frac{(U_3-U_1)}{\tanh^2\ \mu(\eta-\eta_0)}. \tag{3.23}$$

(3) $K_1<0,\ U<U_1=U_2=U_3\leqslant 0,$

$$U=U_1-\sqrt{\frac{2}{-K_1}}\frac{1}{(\eta-\eta_0)^2}. \tag{3.24}$$

(4) $K_1<0,\ U\leqslant U_1<U_2=U_3,\ U_1\leqslant 0,$

$$U=U_3-\frac{U_3-U_1}{\sin^2\ \mu(\eta-\eta_0)}. \tag{3.25}$$

(5) $K_1<0,\ U_1<U_2<U\leqslant U_3\leqslant 0,$

$$U=U_3-(U_3-U_2)sn^2(\mu(\eta-\eta_0),k),\quad k^2=\frac{U_3-U_2}{U_3-U_1}. \tag{3.26}$$

(6) $K_1<0,\ U\leqslant U_1<U_2<U_3,\ U_1\leqslant 0,$

$$U=U_3-\frac{U_3-U_1}{sn^2(\mu(\eta-\eta_0),k)}; \tag{3.27}$$

k as in (3.26)

(7) $K_1>0,\ U_1<U<U_2=U_3=0,$

$$U=U_1\ \frac{1}{\cosh^2\ \mu(\eta-\eta_0)}. \tag{3.28}$$

(8) $K_1>0,\ U_1<U<U_2<0<U_3,$

$$U=(U_2-U_1)sn^2(\mu(\eta-\eta_0),k),\quad k^2=\frac{U_2-U_1}{U_3-U_1}. \tag{3.29}$$

(9) $K_1<0,\ U_1\leqslant 0,\ U_{2,3}=p\pm iq,\ q>0,$

$$U=U_1+A-\frac{2A}{1-cn(\mu(\eta-\eta_0),k)},$$

$$A^2=(p-U_1)^2+q^2,\quad k^2=\frac{A-p+U_1}{2A},\quad \mu=\sqrt{|K_1|A}. \tag{3.30}$$

III. $S=0,\ K_1=0,\ K\neq 0$

(1) $K>0,\ U\leqslant U_1<0<U_2,$

$$U=U_1-(U_2-U_1)\sinh^2\ \frac{\sqrt{K}}{2}\ (\eta-\eta_0). \tag{3.31}$$

(2) $K>0,\ U<U_1=U_2=0,$

$$U=-\exp(-\sqrt{K}(\eta-\eta_0)). \tag{3.32}$$

(3) $K<0,\ U_1<U<U_2\leqslant 0,$

$$U=U_1+(U_2-U_1)\cos^2\ \frac{\sqrt{-K}}{2}\ (\eta-\eta_0). \tag{3.33}$$

IV. $S=K_1=K=0$, $K_2\neq 0$

$$U=-\frac{K_3}{K_2}+\frac{K_2}{4}(\eta-\eta_0)^2, \quad K_2<0, \quad K_3<0. \tag{3.34}$$

V. $S=K_1=K=K_2=0$

$$U=\sqrt{K_3}(\eta-\eta_0), \quad K_3>0. \tag{3.35}$$

C. Individual reductions

1. Algebra $A_{2,1}$

This is an exceptional case when (3.2) implies $R_\xi=0$. We find that the only solution for W of (1.4) is

$$W=R_0e^{iQ}, \quad Q=ax+by+\left(B+\frac{1-R_0^2}{1+R_0^2}(a^2+b^2+A)\right)t+\alpha_0, \tag{3.36}$$

where R_0 and α_0 are integration constants.

2. Algebra $A_{2,2}$

We find

$$W=R(\rho)\exp\, i[\alpha(\rho)+a\phi+bt], \quad \xi=\rho, \tag{3.37}$$

where ρ and ϕ are polar coordinates. The single-valuedness of W requires a to be an integer. The phase $\alpha(\rho)$ and variable η satisfy

$$\alpha_\rho(\rho)=S\frac{(1+R^2)^2}{\rho R^2}, \quad \eta=\ln\,\rho \tag{3.38}$$

[see (3.6)]. For the function $U(\eta)$ of (3.8) we obtain the elliptic function equation if and only if we set

$$A=0, \quad b=B \tag{3.39}$$

[A and B are defined in (1.6)].

We have

$$K_1=K_2=2a^2+4S^2-\frac{K}{2}, \quad K_3=-4S^2, \tag{3.40}$$

in (3.10).

For $S\neq 0$, Eq. (3.40) implies that we can have two negative and two positive roots in Eq. (3.11) or two negative roots and two complex conjugate ones. These cases lead to real solutions, namely (3.16), (3.18), and (3.20). Note that all of them are periodic. In particular, for $a=0$, Eq. (3.11) always has a double root, $U_3=U_4=1$, and reduces to

$$U_\eta^2=-4S^2\left[U^2+\left(1+\frac{K}{8S^2}\right)U+1\right]. \tag{3.41}$$

For $S=0$, $K\neq 4a^2$ we obtain the equation

$$\begin{gathered}U_\eta^2=2\left(a^2-\frac{K}{4}\right)U(U-U_1)(U-U_2),\\ U_1U_2=1, \quad U_1+U_2=\frac{2K}{K-4a^2}.\end{gathered} \tag{3.42}$$

The relevant solutions of (3.42) in this case are the following.
(1) $2a^2 \leqslant K < 4a^2$, $U_1 < U_2 < 0$; solution (3.29) (with $U_3 = 0$). For $K = 2a^2$ we have $U = -1$.
(2) $K > 4a^2$, $U_1 \leqslant U_2 < 0$. Solutions (3.26), (3.27), (3.22), and (3.23) (all with $U_3 = 0$).
(3) $K > 4a^2$, $0 < U_1 < U_2$.
Solutions (3.26) (with $U_3 \to U_2$, $U_2 \to U_1$, $U_1 = 0$).
(4) $K > 4a^2$, $U_{1,2} = p \pm iq$, $q > 0$; solution (3.30).

For $S = 0$, $K = 4a^2$ (3.10) reduces to an elementary one, and its solution is

$$R(\rho) = -R_0^2 \rho^{\pm 2a}, \tag{3.43}$$

where R_0 is an integration constant.

3. Algebra $A_{2,3}$ and $A_{2,4}$

The reduction formulas in both of these cases are

$$W = R(\xi) \exp i[\alpha(\xi) - at - by], \quad \xi = x + vt, \quad \eta = \xi, \tag{3.44}$$

with $v = 0$ and $v \neq 0$ for the algebras $A_{2,3}$ and $A_{2,4}$, respectively. Since the Landau–Lifshitz equation is not Galilei invariant, we cannot change the value of v by a group transformation. The transformation (3.8) leads to (3.10) with

$$K_1 = -\frac{K}{2} + 4S^2 + 2(A + b^2) - \frac{v^2}{2} + 2vS + 2(-B + a),$$

$$K_2 = -\frac{K}{2} + 4S^2 + 2(A + b^2) + \frac{3v^2}{2} - 6vS + 2(B - a), \tag{3.45}$$

$$K_3 = -(2S - v)^2.$$

Equation (3.6) in this case gives

$$\alpha_\xi = S\,\frac{(1 + R^2)^2}{R^2} - \frac{v}{2}\,\frac{1 + R^2}{R^2}. \tag{3.46}$$

For $S \neq 0$ we obtain (3.11) with the constraint

$$U_1 U_2 U_3 U_4 = \left(1 - \frac{v}{2S}\right)^2, \tag{3.47}$$

imposed on these roots. Hence, only an even number of roots can be negative (0, 2, or 4). This, however, means that all solutions (3.12)–(3.20) can occur, though in some cases we must impose $U_4 \leqslant 0$ ($U_4 = 0$ is allowed for $v = 2S$).

For $S = 0$, all solutions (3.22)–(3.30) can occur.

4. Algebra $A_{2,5}$

The reduction formula is

$$W = R(\xi) \exp i\left[\alpha(\xi) + \frac{a}{b}\,\phi + Bt\right], \quad \xi = \ln \rho + \frac{1}{b}\,\phi, \tag{3.48}$$

and we must set $A = 0$ in the Landau–Lifshitz equation. The phase $\alpha(\xi)$ satisfies

$$\alpha_\xi = S\,\frac{(1 + R^2)^2}{R^2} - \frac{a}{b^2 + 1} \tag{3.49}$$

[see (3.6)].

The function $U=-R^2(\xi)$ satisfies (3.10), and for $S\neq 0$ we have

$$K_1=K_2=-\frac{K}{2}+4S^2-\frac{2a^2}{(b^2+1)^2},\quad K_3=-4S^2. \tag{3.50}$$

The solutions that can occur in this case are (3.14)–(3.20). However, there are constraints between various parameters of the solution that follow from the requirement of single-valuedness of W.

For $a=0$, Eq. (3.11) again reduces to (3.41), and we only obtain the elementary periodic solutions (3.16).

For $S=0$, $K_1\neq 0$ we obtain the equation

$$U_\xi^2=K_1U(U-U_1)(U-U_2),$$

$$U_1U_2=1,\quad U_1+U_2=2\left[1+\frac{2a^2}{K_1(b^2+1)}\right]. \tag{3.51}$$

For $-a^2/(b^2+1)<K_1<0$ we have $U_1\leqslant U_2<0$, and solutions (3.22), (3.23), (3.26), and (3.27) are obtained.

For $K_1>0$ we have $0<U_1\leqslant U_2$ but we obtain no real solutions.

For $K_1<-a^2/(b^2+1)$ we obtain solutions (3.30).

Finally, for $S=0$, $K_1=0$ the solution is $U=-R_0^2\exp(\pm\sqrt{K}\xi)$, and hence

$$R=R_0\exp\left[\pm\frac{1}{2}\sqrt{K}\left(\ln\rho+\frac{1}{b}\phi\right)\right]. \tag{3.52}$$

5. Algebra $A_{2,6}$

The reduction formula is

$$W=R(\xi)\exp i[\alpha(\xi)+Bt+a\ln x],\quad \xi=\frac{y}{x},\quad \eta=\arctan\frac{y}{x}=\phi, \tag{3.53}$$

and α satisfies

$$\alpha_\xi=\frac{S}{1+\xi^2}\frac{(1+R^2)^2}{R^2}+\frac{a\xi}{1+\xi^2}, \tag{3.54}$$

and $U(\phi)=-R^2(\xi)$ satisfies (3.11) with $K_1=K_2=-K/2+4S^2+2a^2$, $K_3=-4S^2$. For $a=0$ the equation again reduces to (3.41).

A real solution is obtained only for $K>8S^2$, namely (3.16). It is periodic in ϕ and hence single-valued when μ is an integer.

For $S=0$ we get a real solution for $K_1<0$, namely solution (3.25), which in this case reduces to

$$U(\phi)=-\tan^2\tfrac{1}{2}\sqrt{|K_1|}(\phi-\phi_0). \tag{3.55}$$

This is a single-valued function whenever $\sqrt{|K_1|}$ is an integer.

6. Algebras $A_{2,7}$ and $A_{2,8}$

The corresponding reductions lead to equations that we cannot decouple without introducing higher derivatives, so we will not treat them here.

IV. SOLUTIONS OF THE NONLINEAR DIFFUSION EQUATION

A. General procedure

Let us consider the system (1.7) and (1.8), the NDLE for short. We impose that the solution be invariant under one of the Lie groups generated by the algebra in Eq. (2.7). The functions R and Q will then have the form (3.1) with ξ and β as in Sec. III (different for each subalgebra $A_{2,1},\ldots,A_{2,8}$).

As for the LL equation, algebra $A_{2,1}$ must be treated separately.

For $A_{2,2},\ldots,A_{2,8}$ we always have $(\nabla\xi)^2\neq 0$. Equations (1.7) and (1.8) reduce to

$$\alpha_{\xi\xi}=-2\,\frac{(1-R^2)}{1+R^2}\,\frac{R_\xi}{R}\,\alpha_\xi+\frac{f_\xi}{f}\,\alpha_\xi-2\,fm\,\frac{(1-R^2)}{1+R^2}\,\frac{R_\xi}{R}+hf, \tag{4.1}$$

$$R_{\xi\xi}=\frac{f_\xi}{f}\,R_\xi+\frac{2R}{1+R^2}\,R_\xi^2+\frac{1-R^2}{1+R^2}\,R[\alpha_\xi^2+2mf\alpha_\xi+g^2]+\frac{1}{(\nabla\xi)^2}\left[BR+AR\,\frac{1-R^2}{1+R^2}\right]. \tag{4.2}$$

The functions $f(\xi)$, $m(\xi)$, $h(\xi)$, and $g(\xi)$ are defined by the relations

$$\frac{f_\xi}{f}=\frac{\xi_t-\Delta\xi}{(\nabla\xi)^2},\quad m=\frac{(\nabla\xi,\nabla\beta)}{(\nabla\xi)^2 f},$$

$$h=\frac{\beta_t-\Delta\beta}{(\nabla\xi)^2 f},\quad g^2=\frac{(\nabla\beta)^2}{(\nabla\xi)^2}. \tag{4.3}$$

In order to decouple Eqs. (4.1) and (4.2), we impose a restriction on the functions defined in (4.3), namely

$$m_\xi+h=0. \tag{4.4}$$

Equation (4.1) can then be integrated once to give

$$\alpha_\xi=\left[S\,\frac{(1+R^2)^2}{R^2}-m\right]f(\xi), \tag{4.5}$$

where S is an integration constant. We substitute (4.5) into (4.2) and put

$$R(\xi)=\sqrt{-U(\eta)},\quad \eta=\int f(\xi)d\xi. \tag{4.6}$$

The equations are decoupled and the one for $U(\eta)$ is already in a standard form,[20] namely

$$U_{\eta\eta}=\left(\frac{1}{2U}+\frac{1}{U-1}\right)U_\eta^2+2S^2\,\frac{(1+U)(1-U)^3}{U}+M\,\frac{1+U}{1-U}\,U+NU, \tag{4.7}$$

with

$$M=\frac{2}{f^2}\left(g^2-m^2f^2+\frac{A}{(\nabla\xi)^2}\right),$$

$$N=\frac{2}{f^2}\,\frac{B}{(\nabla\xi)^2}. \tag{4.8}$$

We now make a further restriction, namely, that M and N, defined in (4.8) are constants. Equation (4.7) then has a first integral K and we obtain the elliptic equation (3.10) with

$$K_1=-\frac{K}{2}+4S^2+M+N,$$

$$K_2=-\frac{K}{2}+4S^2+M-N, \tag{4.9}$$

$$K_3=-4S^2.$$

In many cases we have $M=N=0$ and the polynomial on the right-hand side of (3.10) has a double root at $U_3=U_4=1$. The solution we obtain for $S\neq 0$ is (3.16) with $U_4=1$, i.e.,

$$U(\eta)=1-\frac{(1-U_2)(1-U_1)}{(U_2-U_1)\sin^2\ \mu(\eta-\eta_0)+1-U_2},$$
$$\mu=S\sqrt{(1-U_2)(1-U_1)},\quad U_1\leqslant U\leqslant U_2<0. \tag{4.10}$$

For $S=0$, $K>0$ we obtain solution (3.25), i.e.,

$$U=-\tan^2\sqrt{\frac{K}{8}}\,(\eta-\eta_0). \tag{4.11}$$

B. Individual reductions

1. Algebra $A_{2,1}$

We have

$$R=R(t),\quad Q=\alpha_0+ax+by, \tag{4.12}$$

where α_0 is a constant, and R_t satisfies

$$R_t=-\frac{R(1-R^2)}{1+R^2}\,(A+a^2+b^2)-BR. \tag{4.13}$$

Equation (4.13) can easily be integrated [differently depending on whether $(A+B+a^2+b^2)\times(A-B+a^2+b^2)$ vanishes, or not] and we obtain a transcendental equation for $R(t)$.

2. Algebra $A_{2,2}$

The reduction formula is (3.37). We have $m=0$ and (4.4) requires $b=0$, so the solutions are static ones. The variable η and constants involved satisfy

$$\eta=\ln\,\rho,\quad M=2a^2,$$
$$A=B=N=b=0. \tag{4.14}$$

Since we have $b=0$, the solutions are static ones. All solutions (3.12)–(3.30) can occur.

3. Algebra $A_{2,3}$

We have

$$\eta=x,\quad M=2(b^2+A),\quad N=2B,\quad a=0, \tag{4.15}$$

and again, the solutions are static ones, since the reduction formula is (3.44). All solutions of Sec. III B can occur.

4. Algebra $A_{2,4}$

The reduction formula is (3.44), and we have

$$\eta=\frac{1}{v}\,(e^{v(x+vt)}-1),\quad a=0,\quad A=-b^2,\quad B=0,$$

$$M=N=0,\quad v\neq 0. \tag{4.16}$$

The obtained solutions are (4.10) and (4.11) and they are t dependent.

5. Algebra $A_{2,5}$

We have Eq. (3.48) with

$$\eta=\xi=\ln\,\rho+\frac{1}{b}\,\phi,\quad A=B=N=0,$$

$$M=\frac{2a^2b^2}{(b^2+1)^2},\quad b\neq 0. \tag{4.17}$$

Since we have $M\geqslant 0$ we obtain the solutions (3.16), (3.18), (3.20), (3.22), (3.23), (3.26), (3.27), and (3.30). The parameters of these solutions must satisfy, however, certain constraints in order for the solutions to be single-valued. For $K=2M$ we obtain

$$U=-R_0^2\rho^{2|ab|/(b^2+1)}e^{[2|a|/(b^2+1)]\phi}, \tag{4.18}$$

a solution that is not single-valued.

6. Algebra $A_{2,6}$

The reduction formula is (3.54), and we have

$$\eta=\int\frac{d\xi}{1+\xi^2}=\phi,\quad A=B=a=M=N=0, \tag{4.19}$$

so the relevant solutions are (4.10) and (4.11). The solutions are static and they are single-valued for μ or $\sqrt{K/8}$ being integers.

7. Algebra $A_{2,7}$

We put

$$W=R(\xi)\exp\,i\left[\alpha(\xi)+\frac{a}{2}\ln\,t+b\phi\right],$$

$$\xi=\frac{x^2+y^2}{t},\quad \eta=\int\frac{1}{\xi}\,e^{-\xi/4}d\xi=Ei\left(-\frac{1}{4}\,\xi\right), \tag{4.20}$$

where $Ei(x)$ is the exponential integral function. Moreover, we have $b=a=A=B=M=N=0$ and the relevant solutions are (4.10) and (4.11).

8. Algebra $A_{2,8}$

We have

$$W=R(\xi)\exp\,i\left[\alpha(\xi)+\frac{a}{2}\ln\,t\right],\quad \xi=\frac{x}{\sqrt{t}}, \tag{4.21}$$

$$\eta = \int e^{-\xi^2/4} d\xi = \sqrt{\pi}\Phi(\xi),$$

where $\Phi(x)$ is the probability integral. We have $M=N=A=B=a=0$, and so we obtain solutions (4.10) and (4.11)

We see that time-dependent solutions are obtained for the algebras $A_{2,1}$, $A_{2,4}$, $A_{2,7}$, and $A_{2,8}$. For $A_{2,5}$ $A_{2,7}$, and $A_{2,8}$ the solutions are trigonometric ones.

The phases $\alpha(\xi)$ can be calculated by direct integration, since we have

$$\frac{d\alpha}{d\eta} = -S\,\frac{(1-U)^2}{U}, \tag{4.22}$$

and U is already known.

Thus, for U given by (4.10) we get

$$\alpha(\eta) = \sqrt{\frac{(U_1-1)(U_2-1)}{U_1U_2}}\,\arctan\left\{\sqrt{\frac{U_2(U_1-1)}{U_1(U_2-1)}}\,\cot\,\mu(\eta-\eta_0)\right\}$$
$$-\sqrt{(U_1-1)(U_2-1)}\,\arctan\left\{\sqrt{\frac{(U_1-1)}{(U_2-1)}}\,\cot\,\mu(\eta-\eta_0)\right\}+\alpha_0, \tag{4.23}$$

while for U given by (4.11) we have

$$\alpha = 4\sqrt{\frac{2}{K}}\,S\,\cot\{\sqrt{2K}(\eta-\eta_0)\}+\alpha_0. \tag{4.24}$$

V. CONCLUSIONS

The Landau–Lifshitz equation (1.1) has received quite a bit of previous attention, mainly in the context of continuum Heisenberg ferromagnetic spin systems.[21–27] The anisotropy coefficients A and B of Eq. (1.3) were usually set equal to zero. Use was made of the fact that Eq. (1.1) is integrable, at least in the one-dimensional case, or in the two-dimensional, spherically symmetric one.

A one-soliton solution has been obtained;[23,25] a radially symmetric one. In our variables R, Q of Eq. (1.5), (1.6), this solution corresponds to

$$R = \frac{4t+\alpha_1}{4t+\alpha_1+\cosh^2\left[\dfrac{4t+\alpha_1}{(4t+\alpha_1)^2+\alpha_2^2}\,(x^2+y^2)\right]}, \tag{5.1}$$

where α_1 and α_2 are arbitrary real constants. As noted by Lakshamanan and Porsezian[25] the soliton spreads in time.

The solution (5.1) is not among the invariant solutions obtained in this article, nor can it be obtained from such a solution by applying transformations from the symmetry group. As often happens,[28] the method of symmetry reduction that does not rely on integrability, provides different solutions for integrable equations, than the use of Lax pairs or Backlund transformations.

We note that Eq. (1.1) with $A\neq 0$ is not integrable.

In general, we have reduced the LL equation to the ordinary differential equation (3.9). We have integrated Eq. (3.9) in terms of elliptic functions whenever p, q, and m are constants. For the algebras $A_{2,3},\ldots,A_{2,6}$ this was always the case.

For algebra $A_{2,2}$ we obtained Eq. (3.10) only for $A=0$, $B=b$. Let us briefly consider the case when the anisotropy coefficient A does not vanish. We then return to the original variable $\xi=\rho=\sqrt{x^2+y^2}=\exp\,\eta$ and transform Eq. (3.9) into

$$U_{\rho\rho} = \left(\frac{1}{2U}+\frac{1}{U-1}\right)U_\rho^2-\frac{1}{\rho}\,U_\rho+\frac{2S^2}{\rho^2}\,(U-1)^2\left(-U+\frac{1}{U}\right)-\frac{2(a^2+A^2\rho^2)}{\rho^2}+2(B-b). \tag{5.2}$$

For $a=0$, $b=B$ this is the equation for the fifth Painlevé transcendent P_V.[20] However, for $B \neq b$, Eq. (5.2) does not have the Painlevé property. According to the Painlevé conjecture,[29,30] Eq. (1.1) is hence, in general, not integrable.

This has not stopped us from obtaining numerous solutions, both in integrable and nonintegrable cases. The algebra $A_{2,2}$ (cylindrical symmetry) for $A=0$, $B=b$ leads to periodic solutions, as discussed in Sec. III C. The periodicity is in the radial variable ρ. The time dependence is restricted to the phase Q, as is seen in Eq. (3.37). Moreover, the time dependence is entirely due to the presence of the external field B (we have $b=B$) that generates a rotation between the components ϕ_1 and ϕ_2 of the original vector $\boldsymbol{\phi}$.

Some elementary nonperiodic solutions that we can extract from Sec. III are

$$R^2=\frac{U_1-U_2S^2(U_2-U_1)^2(\ln \rho/\rho_0)^2}{1+S^2(U_2-U_1)^2(\ln \rho/\rho_0)^2}, \tag{5.3}$$

$$R^2=\frac{4R_0^2U_4(U_2-U_1)-U_2(U_4-U_1)[2R_0^2+\rho^{2\mu}+R_0^4\rho^{-2\mu}]}{(U_4-U_1)[2R_0^2+\rho^{2\mu}+R_0^4\rho^{-2\mu}]-4(U_2-U_1)R_0^2}, \tag{5.4}$$

with S, R_0, ρ_0, U_i constants and

$$Q=S\int \frac{(1+R^2)^2}{\rho R^2}\,d\rho+a\phi+Bt, \tag{5.5}$$

in both cases.

For $S=0$, we have, for instance,

$$R=\left[-U_1+\sqrt{\frac{2}{-K_1}}\left(\ln\frac{\rho}{\rho_0}\right)^{-2}\right]^{1/2}, \tag{5.6}$$

$$R=\frac{2\sqrt{-U_1}R_0}{\rho^{\mu}+R_0^2\rho^{-\mu}}, \tag{5.7}$$

with

$$Q=a\phi+Bt+Q_0. \tag{5.8}$$

For $A\neq 0$, as mentioned above, solutions are obtained in terms of $P_V(\rho)$. Their time dependence is again given by the term Bt in the phase Q.

For algebras $A_{2,3}$ and $A_{2,4}$ we obtain Eq. (3.10) and a multitude of explicit solutions for all values of a, b, A, and B. Note that for

$$S\neq 0, \quad B=a, \quad A=-b^2, \tag{5.9}$$

in particular, for the one-dimensional ($b=0$), static ($a=0$) with no external fields ($A=B=0$), two of the roots in Eq. (3.11) coincide, and the equation reduces to

$$U_\eta^2=-4S^2(U-1)^2\left[U^2+\left(-\frac{K_1}{4S^2}+2\right)U-\frac{K_3}{4S^2}\right]. \tag{5.10}$$

Equation (5.10) only allows elementary solutions like (3.12)–(3.17), not, however, the elliptic function ones. These occur when the fields A and B are such that (5.9) is not satisfied.

To our knowledge, the NLDE (1.2) has not been investigated from the point of view of its integrability, and we have no solutions to compare ours to.

We have derived many explicit exact solutions of both equations. Looking at them we note that most of them have infinite energy. They can describe coherent phenomena in various solid state and condensed matter applications.

Looking first at the solutions of the LL equation, we note that some of our solutions have finite energy. In particular, this is the case for (3.43). This solution is obtained from the familiar

static solution describing n solitons ''on top of each other.''[31] Its time dependence is given by the factor e^{iBt}, which thus describes a rotation of this static solution in the ϕ_1, ϕ_2 plane with the angular frequency given by the anisotropy B. The other solutions of this class correspond to the static elliptic solutions discussed in Ref. 17 again rotated by e^{iBt}.

The solutions corresponding to algebras $A_{2,1}$, $A_{2,3}$, and $A_{2,4}$ have infinite energies. As such, they describe various waves in the medium (generalizations of plane waves). These can, for instance, be spin waves; the energy per period is finite.

An interesting class of solutions are those corresponding to algebras $A_{2,5}$ and $A_{2,6}$. Given the choice of parameters, these solutions can be of finite energy; however, due to their dependence on the variable $\phi = \arctan(y/x)$, they may become singular when x and y vanish. They can be used to describe media with defects.

Most of the comments made above apply also to the solutions of the NLDE. The static solutions in both cases are, of course, the same. When we consider nonstatic solutions, the most interesting from the physical point of view, are perhaps solutions corresponding to algebras $A_{2,1}$, $A_{2,7}$, and $A_{2,8}$. All of them have infinite energies. The solution corresponding to $A_{2,1}$ describes a structure shrinking toward the origin (or expanding to infinity—depending on the values of the parameters). The other solutions describe field configurations evolving in time. They can be used in the description of some physical phenomena in condensed matter or solid state physics.

Among the questions that we plan to return to, we mention the study of ''partially invariant'' solutions[32–34] of Eqs. (1.1) and (1.2), and also ''conditionally invariant'' ones.[35] A study of solutions involving Painlevé transcendents is also warranted.

ACKNOWLEDGMENTS

Most of the work reported in this paper was performed when one of us (WJZ) visited the CRM, Université de Montréal, Canada. He wishes to thank the Centre de Recherches Mathématiques for its support and hospitality. The research of P.W. was partly supported by research grants from the NSERC of Canada and the FCAR du Québec.

[1] L. Landau and E. Lifshitz, Phys. Z. (Sowjetunion) **8**, 153 (1935).
[2] See, e.g., N. Papanicolaou and T. N. Tomaras, Nucl. Phys. B **360**, 425 (1991).
[3] A. P. Malozemoff and J. C. Slonczewski, *Magnetic Domain Walls in Bubble Materials* (Academic, New York, 1979).
[4] T. H. O'Dell, *Ferromagnetodynamics, the Dynamics of Magnetic Bubbles, Domains and Domain Walls* (Wiley, New York, 1981).
[5] V. G. Bar'yakhtar *et al.*, Sov. Phys. Usp. **20**, 298 (1977).
[6] F. D. M. Haldane, Phys. Rev. Lett. **57**, 1488 (1986); F. G. Mertens *et al.*, *Nonlinear Coherent Structures in Physics and Biology*, edited by K. H. Spatschek and F. G. Mertens (Plenum, New York, 1993).
[7] N. Papanicolaou, Physica D **74**, 107 (1994); N. Papanicolaou and W. J. Zakrzewski, *ibid.* **80**, 225 (1995); Phys. Lett. A **210**, 328 (1996).
[8] See, e.g., A. J. Bray and K. Humayan, J. Phys. A **23**, 5897 (1990); M. Zapotocky and W. J. Zakrzewski, Phys. Rev. E **51**, R5189 (1995).
[9] For a recent review of the theory of phase ordering, see, e.g., A. J. Bray, Adv. Phys. **43**, 357 (1994).
[10] A. J. Bray and K. Humayun, Phys. Rev. E **47**, 9 (1993).
[11] B. Yurke *et al.*, Phys. Rev. E **47**, 1525 (1993).
[12] M. Zapotocky, P. M. Goldbart, and N. Goldenfeld, Phys. Rev. E **51**, 1216 (1995).
[13] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
[14] P. Winternitz, ''Lie groups and solutions of nonlinear partial differential equations,'' in *Integrable Systems, Quantum Groups and Quantum Field Theories*, edited by A. Ibort and M. A. Rodriguez (Kluwer Academic, Dordrecht, 1992).
[15] G. Gaeta, *Nonlinear Symmetries and Nonlinear Equations* (Kluwer Academic, Dordrecht, 1994).
[16] C. Rogers and W. F. Ames, *Nonlinear Boundary Value Problems in Science and Engineering* (Academic, San Diego, 1989).
[17] A. M. Grundland, P. Winternitz, and W. J. Zakrzewski, J. Math. Phys. **37**, 1501 (1996).
[18] B. Champagne, W. Hereman, and P. Winternitz, Comput. Phys. Commun. **66**, 319 (1991).
[19] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists* (Springer-Verlag, Berlin, 1971).
[20] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[21] M. Lakshmanan, Phys. Lett. A **61**, 53 (1977).
[22] V. E. Zakharov and L. A. Takhtadzhyan, Teor. Mat. Fiz. **38**, 26 (1979).
[23] M. Lakshmanan and M. Daniel, Physica A **107**, 533 (1981).
[24] A. V. Mikhailov and A. I. Yaremchuk, JETP Lett. **36**, 78 (1982).
[25] M. Lakshmanan and K. Porsezian, Phys. Lett. A **146**, 329 (1990).
[26] K. Porsezian and M. Lakshmanan, J. Math. Phys. **32**, 2923 (1991).
[27] M. Daniel, K. Porsezian, and M. Lakshmanan, J. Math. Phys. **35**, 6498 (1994).
[28] D. Levi, C. R. Menyuk, and P. Winternitz, Phys. Rev. A **44**, 6057 (1991).

[29] M. J. Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 1006 (1980).
[30] M. J. Ablowitz and P. A. Clarkson, Solitons, *Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[31] See, e.g., B. Piette and W. J. Zakrzewski, ''Skyrmion scattering in $(2+1)$ dimensions,'' in *Solitions and Chaos* (Springer–Verlag, Berlin, 1991), p. 325.
[32] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[33] L. Martina and P. Winternitz, J. Math. Phys. **33**, 2718 (1992).
[34] L. Martina, G. Soliani, and P. Winternitz, J. Phys. A **25**, 4425 (1992).
[35] D. Levi and P. Winternitz, J. Phys. A **22**, 2915 (1989).

Introduction to the special issue on wavelet and time-frequency analysis

Bruno Torrésani
Centre de Physique Théorique, CNRS-Luminy, Case 907, 13288 Marseille Cedex 09, France and Laboratoire d'Analyse, Topologie et Probabilités, CMI, Université de Provence, 39 Rue Joliot-Curie, 13442 Marseille Cedex, France

(Received 7 April 1998; accepted for publication 7 April 1998)

[S0022-2488(98)01308-5]

I. INTRODUCTION

Many recent scientific developments in theoretical sciences (mathematics, theoretical physics, signal processing, etc.) are based on *efficient representations* of mathematical objects such as functions, signals, operators. Prototypes of such efficient representations are of course Fourier expansions, or more generally spectral decompositions for self-adjoint operators. However, the word *efficient* may be given several different interpretations. For example, representing efficiently a function may be understood as representing it with a minimal amount of data. Alternatively, an efficient representation may also be a representation from which mathematical or physical properties may be retrieved easily. Another alternative interpretation may be that an efficient representation gives a way to simulate, or generate easily the object under study.

In that respect, wavelet analysis and more generally, time-frequency analysis, may be considered as good candidates for providing efficient representations of functions or operators in a large class of different contexts (ranging from signal compression to mathematical analysis).

Wavelet analysis seems to be a relatively new scientific domain, since it was introduced in its ''modern form'' in the early eighties. However, it is worth pointing out that similar methods have been used for decades in different contexts in mathematics and mathematical physics, as well as signal and image processing. What is really more recent is the fact that since then, strong interactions have been developed between those different areas, which have led to significant advances.

At a very elementary level, wavelet analysis and time-frequency analysis consist in providing alternative representations of functions, from which information regarding time localization as well as frequency localization may be obtained more easily.

This may be done by decomposing those functions into more ''elementary functions,'' generated using simple rules. For example, classical wavelets are translates and dilates of a single function, while the so-called Gabor functions are obtained as translated and modulated copies of a single function. Such a methodology is far from original, and several origins may be quoted. For example, Littlewood–Paley analysis may be considered as one of the ''ancestors'' of wavelet analysis in the field of mathematical analysis (see Ref. 1 for a review), while mathematical physicists have been developing coherent states methods for a long time in the context of coherent optics, or quantization (see, e.g., Ref. 2). In more applied areas, one may mention as well similar developments going back to Gabor[3] in signal processing or Marr[4] in vision.

An alternative approach to time-frequency analysis consists of constructing different types of time-frequency representations for functions, which are merely generalizations of Wigner functions. Needless to say, again, such methods owe a lot to mathematical physics and quantization. Wigner functions have received a renewed interest in recent years, both in engineering (signal processing) and mathematics (for example, the analysis of nonstationary stochastic processes). It is now obvious that time-frequency signal processing and quantization methods have a lot in common, and that enormous progress can be made by developing the connections.

Even though mathematical physics may be considered one of the origins of wavelet and time-frequency analysis, it seems that the recent developments of the latter did not have a significant impact on the former (at least the impact was much less spectacular than the impact on mathematics or more applied sciences). It is one of the goals of this volume to help disseminate

 © 1998 American Institute of Physics

those new techniques in the mathematical physics community. Since there already exist many textbooks (more than 50 in fact) covering various aspects of wavelet theory (see, for example, Refs. 5–7, 1, 8, or 9 for an historical perspective), there is no need for an introductory article on wavelets. All the contributions to this special issue emphasize either specific methodologies that are expected to be useful in mathematical physics, or mathematical physics problems solved with wavelet or time-frequency techniques.

The contributions to this special issue have been organized in four sections: *Coherent states and semiclassical analysis; quantization and signal processing; Littlewood–Paley-type methods; fractals, multifractals and applications;* and *generalized wavelet and time-frequency bases*. Such an organization is only intended to serve as a reader's guide, and not as a clear-cut separation between scientific domains, which would be artificial.

II. COHERENT STATES APPROACHES

The connection of wavelet analysis with coherent states and the theory of square integrable group representations was recognized in a seminal paper by Grossmann, Morlet, and Paul.[10] Based on a famous theorem of Duflo and Moore on square-integrable representations, these authors expressed wavelet transform and Gabor transform as intertwiners between the (left or right) regular representation and a square-integrable representation, in the case of the $ax+b$ group and the Heisenberg group respectively. This work has been the starting point of many further investigations and extensions to more general groups (for example, general semidirect products $H \times \mathbf{R}^n$, with $H \subset GL(n,\mathbf{R})$ in Ref. 11), and representations that are not necessarily square integrable (a review may be found in Ref. 2). Several extensions of this theory are presented in four articles of this volume.

Ali extends the classical Duflo–Moore theorem on square integrable representations of locally compact groups so as to cover the case of vector coherent states. His treatment yields in particular coherent states incorporating spin degrees of freedom.

The case of semidirect products $H \times \mathbf{R}^n$ is considered in the contributions of Aniello *et al.* and Führ. The paper by Aniello *et al.* extends previous results of Bernier and Taylor[11] and Führ, focusing on representations realized as induced representations. They also discuss the existence of discrete frames associated with the continuous transforms.

The paper by Führ focuses on the case where the dilation group H is Abelian. He gives a complete characterization of those groups which yield a resolution of the identity, by establishing a correspondence with commutative associative algebra with unity.

Antoine and Vandergheynst address the problem of the construction of wavelet decompositions on spheres, a problem of great practical importance (for example, in geophysics or in astronomy, for obvious reasons). The main problem is to find an acceptable substitute for dilations on spheres. A natural "$ax+b$" group adapted to the geometry as the sphere is given by $SO(n) \times \mathbf{R}_*^+ \subset SO_o(n,1)$. The corresponding continuous wavelet transform on S^{n-1} is then obtained from a suitable unitary representation of $SO_o(n,1)$ acting in $L^2(S^{n-1}, d\mu)$, which is shown to be square-integrable when restricted to the quotient $SO_o(n,1)/\mathbf{R}^{n-1}$.

The paper by Paul and Uribe is of a different nature, and develops an application of coherent states rather than construction of such states. The main result is a new weighted trace formula for Schrödinger operators with a discrete spectrum. More precisely, the (squared modulus of) coherent states transform of the eigenfunctions of the Hamiltonian H is used to weigh the spectral density smeared out with a test function whose Fourier transform is compactly supported. The leading term in the $\hbar$ expansion of the latter quantity is given an expression which involves the complex trajectories infinitesimally close to closed hyperbolic orbits.

III. QUANTIZATION AND SIGNAL PROCESSING

The so-called time-frequency methods amount to realize functions in a different space, in which both time and frequency variables are present. Besides linear decomposition methods such as wavelet or Gabor transform, bilinear transformations such as (generalized) Wigner transforms have also been quite popular, in particular in the signal processing community. The prototype of such bilinear transforms is the celebrated Wigner function, known as the *Wigner–Ville* distribution in the signal processing community, after the engineer J. Ville who introduced it as a candidate for

local spectrum for nonstationary signals. The Wigner–Ville distribution (defined as a function on the time-frequency plane) has been generalized to yield families of time-frequency distributions, and also families of *time-scale distributions*, transforming functions into functions of one time and one scale variable (the scale being heuristically interpreted as the inverse of a frequency variable). See Ref. 12 for a review.

The article written by Flandrin examines in detail several of the generic properties of bilinear time-frequency distributions. The classical Wigner function is known to be positive and separable in the case of Gaussian functions, which are of minimum spread in the time-frequency plane (minimum uncertainty). The relationship between these properties is examined, as well as their possible generalizations to smoothed Wigner–Ville distributions. Similar questions are answered in the (more complex technically speaking) case of affine distributions.

The interrelations between time-frequency analysis, pseudodifferential calculus and quantization are developed in the papers by Matz and Hlawatsch, and Bertrand and Bertrand. The first article is more ''signal-processing oriented,'' and addresses the problem of time-varying system identification via time-frequency analysis. Time-varying systems (in other words, linear operators relating output data to input data, to be recovered from data) may in some cases be estimated thanks to time-frequency methods. Matz and Hlawatsch describe a class (the underspread systems) for which this may be done, and provide estimates for the accuracy of the method. The connections with the classical pseudodifferential calculus (e.g., associated with the Weyl–Heisenberg group) are also emphasized.

The article of Bertrand and Bertrand develop the symbolic calculus associated with the action of the ''$ax+b$'' group on a (generalized) Hardy space $\mathscr{H}$, and provide the mathematical foundations for a corresponding signal theory. Some of the questions addressed in the previous paper (Wigner functions, star products, etc.) are also studied in this context. The case of larger covariance groups containing the ''$ax+b$'' group is also examined.

IV. LITTLEWOOD–PALEY-TYPE APPROACHES

As mentioned earlier, Littlewood–Paley methods may be considered the ''ancestors'' of wavelet analysis in the harmonic analysis community. Reviews may be found for example in Refs. 1 and 13. Their applications include function space characterization, pseudodifferential calculus, etc.

Frick, Grossmann and Tchamitchian, using similar mathematical tools, address a problem of a different nature, of great practical interest: the analysis of functions which are defined on closed subsets of the real lines (such situations occur, for example, in astronomy, where some observations may be done only during the night). Missing data make spectral analysis quite unprecise, and the authors show how to modify the usual continuous wavelet analysis so as to minimize the effect of these missing data, by perturbing the wavelets near the boundaries of the gaps in order to restore vanishing moments. These techniques have found beautiful applications, for example, for the estimation of cycles in the sunspot data.

Lemarié's contribution is a review paper on recent results that have been obtained on the existence and uniqueness of global solutions for Navier–Stokes equations, which have recently enjoyed a renewed interest, in particular since the work of Cannone.[14] This paper reviews the existence results for global weak and mild solutions, as well as some more recent uniqueness results. It also provides a good sample of the techniques needed for this kind of analysis, in particular the use of Littlewood–Paley decompositions for deriving sharp estimates for the nonlinear terms in the equations.

The paper by Lewalle does not, strictly speaking, rely on Littlewood–Paley methods, but rather on formal analytical methods. He studies the Poisson equation directly in the wavelet domain, with specific choices for the analyzing wavelet (Hermitian wavelets). The motion in the position-scale-number space is actually Hamiltonian, which leads to interesting simplifications.

V. FRACTALS, MULTIFRACTALS AND APPLICATIONS

One of the main contributions of wavelet analysis lies in the development of a rigorous fractal and multifractal analysis for functions and signals. It was first developed in the context of turbulence, as a generalization of the multifractal formalism for measures, and eventually led to a

beautiful connection with microlocal analysis. The very nature of wavelets allow a very precise analysis of local regularity properties of functions, which was used in particular by Jaffard and co-workers (see in particular Refs. 15 and 16) to set up a rigorous multifractal formalism for functions.

Jaffard's paper is an extension of the latter formalism to the case of functions with oscillating singularities (i.e., functions whose local regularity is characterized not only by an usual Hölder exponent, but also by a second exponent which describes the speed at which oscillations accumulate at the singular point). Surprisingly enough, such oscillating singularities have been shown to occur in many apparently simple situations, which made necessary the development of a corresponding version of multifractal formalism.

A numerical version of the multifractal formalism was developed earlier by Arneodo and co-workers, and applied to many practical physical problems (turbulence, DLA aggregates, financial time series, etc.) The contribution of Arneodo, Bacry and Muzy provides statistical justifications for their algorithm, by developing a probabilistic random cascade model based on wavelet bases (also strongly influenced by the cascade models of turbulence). The multifractal formalism previously mentioned thus appears as a method for estimating the parameters of this model, which gives the whole program a beautiful consistency.

Another application of the fractal-based method is presented in the article by Guerin and Holschneider, who consider the problem of wave scattering by fractal measures. They develop the theory of time evolution for Schrödinger operators when the potential is defined as a measure (in the one-dimensional situation). When the potential is fractal, they show how some fractal characteristics of the potential (namely some wavelet-defined fractal dimensions) may be estimated from the reflection coefficient.

VI. GENERALIZED WAVELET AND TIME-FREQUENCY BASES

One of the origins of modern wavelet analysis lies on a famous no-go theorem, known as the Balian–Low theorem (see Ref. 5 for a precise statement and a discussion of consequences), which may be formulated as follows (in a weak form). Let $g \in L^2(\mathbf{R})$ be such that the collection of its time-frequency translates $g_{mn}(t) = \exp\{2i\pi mt\}g(t-n), m,n \in \mathbf{Z}$ form an orthonormal basis of $L^2(\mathbf{R})$. Then the function $g(x)$ cannot be simultaneously smooth and localized, in the following sense: either $\int t^2|g(t)|^2dt = \infty$ or $\int \omega^2|\hat{g}(\omega)|^2d\omega = \infty$. The paper by Zak exploits this result to show that the eigenfunctions of the Schrödinger operator of an electron in the plane, subject to the action of a constant perpendicular magnetic field cannot be localized. A particularly interesting aspect of this paper is that it describes an alternative version of Gabor analysis using magnetic translations.

Classical wavelet bases are associated with periodic tilings of the real axis in the one-dimensional case, and of $\mathbf{R}^n$ in higher dimensions (detailed accounts may be found in many textbooks, see e.g. Ref. 5). Thanks to such a translation invariance, they are particularly well adapted to problems in which translation invariance play a role (for example, they yield simple and efficient representations for large classes of differential and pseudodifferential operators).

There are problems in which such a translation invariance does not exist any more, and is replaced by some more subtle requirements. This is in particular the case for quasicrystals, which are naturally associated with aperiodic tilings of the plane. A natural question was whether it is possible to construct wavelet bases associated with such aperiodic tilings of the plane.

This question receives its first answer in the article by Gazeau, Patera and Pelantova. This paper shows how to construct on such tilings wavelet bases that mimic the celebrated Haar basis (each basis function is the difference of the characteristic functions of two adjacent tiles). In such a case, the dilation has, of course, to be replaced with a more complex operation that preserves the geometry of the tiling.

A complementary answer is also provided by Bernuau's paper. The wavelet bases constructed by Bernuau have a similar algebraic nature, but unlike the Haar-type functions, are smooth, and their smoothness may be controlled (in particular, wavelets with exponential decay are constructed). This article also studies the continuity of the wavelet family with respect to the considered quasicrystal.

The article by Wei and Chen is of a different nature, for it is not, strictly speaking, devoted to time-frequency analysis. The authors develop an elementary theory of square waves, as an alternative for Fourier analysis, useful for practical applications, in electronics for example. This contribution should be useful in particular because of the arguments developed for proving basic results such as orthogonality, completeness, etc.

[1] Y. Meyer and R. Coifman, *Wavelets and Operators* (Cambridge University Press, London, 1995).
[2] S. T. Ali, J.-P. Antoine, J.-P. Gazeau, and U. A. Mueller, ''Coherent states and their generalizations: A mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995).
[3] D. Gabor, ''Theory of communication,'' J. Inst. Elec. Eng. **93**, 429–657 (1946).
[4] D. Marr, *Vision* (Freeman, San Francisco, 1982).
[5] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Series in Applied Mathematics (SIAM, New York, 1992).
[6] M. Holschneider, *Wavelets: An Analysis Tool* (Oxford University Press, Oxford, 1995).
[7] S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, New York, 1998).
[8] B. Torrésani, *Analyse Continue par Ondelettes* (Inter Editions/Editions du CNRS, Paris, 1995). English translation to appear at SIAM (translated and revised by R. D. Ryan).
[9] B. B. Hubbard, *The World According to Wavelets*, edited by A. K. Peters (Wellesley, New York, 1996).
[10] A. Grossmann, J. Morlet, and T. Paul, ''Transforms associated to square integrable group representations. I. General results,'' J. Math. Phys. **26**, 2473–2479 (1985).
[11] D. Bernier and K. Taylor, ''Wavelets from square-integrable representations,'' SIAM (Soc. Ind. Appl. Math.), J. Math. Anal. **27**, 594–608 (1996).
[12] L. Cohen, *Time-Frequency Analysis* (Prentice–Hall, Englewood Cliffs, 1995).
[13] E. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals* (Princeton University Press, Princeton, 1996).
[14] M. Cannone, *Ondelettes, Paraproduits et Navier-Stokes* (Diderot, Editeur des arts et sciences, Paris, 1996).
[15] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy, *Ondelettes, Multifractales et Turbulence* (Diderot éditeur des Arts et Sciences, Paris, 1995).
[16] S. Jaffard, ''Wavelet methods for fast resolution of elliptic problems,'' SIAM (Soc. Ind. Appl. Math.), J. Math. Anal. **28**, 944–998 (1997).

A general theorem on square-integrability: Vector coherent states

S. Twareque Ali
Department of Mathematics and Statistics, Concordia University, Montréal, Québec, H4B 1R6 Canada

(Received 15 August 1997; accepted for publication 13 March 1998)

We derive a generalization of the well-known theorem for the square integrability of a unitary irreducible representation of a locally compact group. The generalization covers the case of representations admitting vector coherent states. The result is illustrated by an example drawn from the isochronous Galilei group. The construction yields a wide variety of coherent states, labeled by phase space points, which satisfy a resolution of the identity condition, and incorporate spin degrees of freedom. © *1998 American Institute of Physics.* [S0022-2488(98)00108-X]

I. INTRODUCTION

Let G be a locally compact (usually Lie) group and U a continuous, unitary irreducible representation (UIR) of G on some (separable, complex) Hilbert space $\mathfrak{H}$. A nonzero vector $\eta \in \mathfrak{H}$ is said to be *admissible* if

$$I(\eta) \equiv \int_G |\langle U(g)\eta|\eta\rangle|^2 \, d\mu(g) < \infty, \tag{I.1}$$

where μ denotes the left (or right) Haar measure on G. If such a vector exists, the representation U is said to be *square-integrable*. Square-integrable representations have been studied extensively in the literature (see, e.g., Ref. 1 for a general discussion and references) and one of the most important results in this connection is

Theorem I.1: *Let the UIR $g \mapsto U(g)$ of the locally compact group G be square integrable and $\eta \in \mathfrak{H}$ an admissible vector. Then, the mapping*

$$W_\eta : \mathfrak{H} \to L^2(G, d\mu), \quad (W_\eta \phi)(g) = [c(\eta)]^{-1/2} \langle \eta_g | \phi \rangle, \quad \phi \in \mathfrak{H}, \quad g \in G,$$

where

$$c(\eta) = \frac{I(\eta)}{\|\eta\|^2}, \quad \eta_g = U(g)\eta, \tag{I.2}$$

is a linear isometry onto a (closed) subspace $\mathfrak{H}_\eta$ of $L^2(G, d\mu)$. On $\mathfrak{H}$, the resolution of the identity

$$\frac{1}{c(\eta)} \int_G |\eta_g\rangle\langle\eta_g| \, d\mu(g) = I, \tag{I.3}$$

holds. The subspace $\mathfrak{H}_\eta = W_\eta(\mathfrak{H}) \subset L^2(G, d\mu)$ is a reproducing kernel Hilbert space, so that the corresponding projection operator

$$\mathbb{P}_\eta = W_\eta W_\eta^*, \quad \mathbb{P}_\eta L^2(G, d\mu) = \mathfrak{H}_\eta, \tag{I.4}$$

has the reproducing kernel K_η:

$$(\mathbb{P}_\eta \Phi)(g) = \int_G K_\eta(g, g') \Phi(g') d\mu(g'), \quad \Phi \in L^2(G, d\mu),$$

 © 1998 American Institute of Physics

$$K_\eta(g,g')=\frac{1}{c(\eta)}\langle \eta_g|\,\eta_{g'}\rangle, \tag{I.5}$$

as its integral kernel. Furthermore, W_η intertwines U and the left regular representation $U_\swarrow$,

$$W_\eta U(g)=U_\swarrow(g)W_\eta, \quad g\in G. \tag{I.6}$$

The proof of this theorem can, for example, be found in Ref. 2. Note that in the above, the reproducing kernel is a map $K_\eta: G\times G\to\mathbb{C}$, with the properties:

$$K_\eta(g,g')=\overline{K_\eta(g',g)}, \quad g,g'\in G,$$

$$K_\eta(g,g)>0, \quad g\in G, \tag{I.7}$$

$$\int_G K_\eta(g,g'')K_\eta(g'',g')\;\; d\mu(g'')=K_\eta(g,g'),$$

while the left regular representation $U_\swarrow$ acts on $L^2(G,d\mu)$ in the manner

$$(U_\swarrow(g)\Phi)(g')=\Phi(g^{-1}g'), \quad \Phi\in L^2(G,d\mu). \tag{I.8}$$

The vectors η_g, $g\in G$, are called *coherent states* (CS) of the representation U. The class of representations for which the above theorem is applicable is restricted to those of the *discrete series*. For purposes of physical applications, however, the above definitions of square-integrability and coherent states are unduly restrictive. One would like to work with a fairly wide class of UIR's of a locally compact group and still be able to define coherent states and resolutions of the identity—in a sense more general than (I.3). One possibility would be to look for generalizations of Theorem I.1. In the next section we obtain such a generalization in Theorem II.2, which is the main result of this paper. It characterizes square integrability properties of *vector coherent states (VCS)*,[3–5] giving a criterion for obtaining a resolution of the identity for such coherent states. Section III illustrates this generalization with a construction based on a representation of the isochronous Galilei group.

II. A GENERALIZATION TO VECTOR COHERENT STATES

Once again, let G be a locally compact group, H a closed subgroup of G and $X=G/H$ the corresponding left coset space. Assume that X carries the measure ν which is *invariant* under the action of G and for any $g\in G$, denote its coset gH by $[g]$. Let U be a UIR of G and $\mathfrak{K}$ a *finite dimensional* subspace of $\mathfrak{H}$ which is stable under U when the latter is restricted to H:

$$U(h)\,\eta\in\mathfrak{K}, \quad \text{for all} \quad \eta\in\mathfrak{K}, \quad h\in H. \tag{II.1}$$

Let $\mathbb{P}$ be the projection operator for the subspace $\mathfrak{K}$ and write

$$U(h)\mathbb{P}=V(h), \quad h\in H. \tag{II.2}$$

We assume that the representation $h\mapsto V(h)$, of the subgroup H, acts *irreducibly* on $\mathfrak{K}$.

Let $\{\eta^i\}_{i=1}^n$ (n = dimension of $\mathfrak{K}$) be an orthonormal basis for $\mathfrak{K}$ and $\sigma: X\to G$ a (global) Borel section of G. Then for each $[g]\in X$, the set $\{\eta^i_{\sigma([g])}=U(\sigma([g]))\,\eta^i\}_{i=1}^n$ are vector coherent states. Clearly the set of vectors

$$\mathfrak{S}_\sigma=\{\eta^i_{\sigma([g])}|\,i=1,2,\ldots,n, \quad [g]\in X\} \tag{II.3}$$

is total in $\mathfrak{H}$, and we are interested in conditions for their square integrability, i.e., conditions to ensure a resolution of the identity of the type:

$$\sum_{i=1}^n \int_X |\,\eta^i_{\sigma([g])}\rangle\langle\,\eta^i_{\sigma([g])}|\,d\nu([g])=\int_X U(\sigma([g]))\mathbb{P}U(\sigma[g]))^*=\lambda I, \quad \lambda>0. \tag{II.4}$$

When $\mathfrak{K}$ is one dimensional, the set of vectors $\mathfrak{S}_\sigma$ are the well-known *Gilmore–Perelomov coherent states*,[6,7] and a generalization of Theorem I.1 for this case has been worked out in Ref. 8. On the other hand, if H is the trivial subgroup $\{e\}$ consisting of the identity element alone, we are back within the framework of Theorem I.1.

Looking upon G as a principal bundle, fibered over X and with structure group H, we consider the G-homogeneous associated bundle $B=G\times_{V(H)}\mathfrak{K}$. This bundle, also fibered over X, is obtained by identifying elements $(g,\eta),(g',\eta')\in G\times\mathfrak{K}$, whenever $g'=gh$ and $\eta'=V(h)^{-1}\eta$ for some $h\in H$. Elements in B can therefore be put into one to one correspondence with vectors $U(g)\eta \in\mathfrak{H}$, for all $g\in G$ and $\eta\in\mathfrak{K}$. Following the terminology of Ref. 9, these vectors may be called a family of *coherent vectors*. The canonical projection $\pi:B\to X$ is then simply $\pi(U(g)\eta)=[g]$. Sections in the bundle B can be put into one to one correspondence with functions $f:G\to\mathfrak{K}$, which satisfy

$$f(gh)=V(h)^{-1}f(g),\quad h\in H. \tag{II.5}$$

The class of square integrable sections, with respect to the invariant measure ν, is the Hilbert space $L^2_{\mathfrak{K}}(X,d\nu)$ and on it the representation $g\mapsto{}^VU(g),\quad g\in G$,

$$({}^VU(g)f)(g')=f(g^{-1}g'),\quad g'\in G, \tag{II.6}$$

is unitary and is just the standard induced representation[10] of G (induced from the UIR V of the subgroup H).

Consider next the functions $\Phi:G\to\mathfrak{K}$, defined by

$$\Phi(g)=\mathbb{P}U(g)^{-1}\phi,\quad \phi\in\mathfrak{H}. \tag{II.7}$$

Clearly these functions satisfy (II.5) and hence also describe sections of B. We denote this class of sections (for all $\phi\in\mathfrak{H}$) by $\mathfrak{H}_K$, the reason for this subscript will become clear momentarily. A natural question which now arises is under what conditions $\mathfrak{H}_K$ forms a subspace of $L^2_{\mathfrak{K}}(X,d\nu)$. There are two reasons for wanting to know this: First, the linear map $W:\mathfrak{H}\to\mathfrak{H}_K$, such that

$$(W\phi)(g)=\Phi(g)=\mathbb{P}U(g)^{-1}\phi,\quad g\in G,\quad \phi\in\mathfrak{H}, \tag{II.8}$$

it is easily seen to intertwine the representation U (of G on $\mathfrak{H}$) with the induced representation ${}^VU$, and if $\mathfrak{H}_K$ could be identified with a closed subspace of $L^2_{\mathfrak{K}}(X,d\nu)$, then U would be a subrepresentation of an induced representation. Second, square integrability of the sections Φ would imply

$$\int_X\langle\phi|\mathbb{P}([g])\phi\rangle\ \ d\nu([g])<\infty, \tag{II.9}$$

for all $\phi\in\mathfrak{H}$, where we have written, with obvious justification,

$$\mathbb{P}([g])=U(g)\mathbb{P}U(g)^*. \tag{II.10}$$

Thus we would then have

$$\int_X\mathbb{P}([g])\ \ d\nu([g])=A \tag{II.11}$$

for some bounded positive operator A (the above integral converging weakly). On the other hand, from the invariance of the measure ν and (II.10), it would follow that $U(g)AU(g)^*=A$, for all $g\in G$, with the irreducibility of U then requiring that $A=\lambda I$, for some $\lambda>0$. This way we shall have obtained a square integrability condition of the type (II.4).

The set of sections $\mathfrak{H}_K$ has the natural structure of a reproducing kernel Hilbert space,[11,12] with reproducing kernel $K:G\times G\to\mathcal{L}(\mathfrak{K})$ (= bounded operators on $\mathfrak{K}$),

$$K(g,g')=\mathbb{P}U(g)^{-1}U(g')\mathbb{P}. \tag{II.12}$$

Indeed, since $K(g,g')=K(g',g)^*$ and as an operator on $\mathfrak{K}$, $K(g,g)$ is strictly positive, K is a strictly positive definite kernel and therefore the relationship

$$\langle \Phi_g^\eta | \Phi_{g'}^{\eta'} \rangle_K = \langle \eta | K(g,g')\,\eta' \rangle_{\mathfrak{H}} \tag{II.13}$$

can be used to define a scalar product $\langle \cdot | \cdot \rangle_K$ on the linear span of sections of the type Φ_g^η, for all $g \in G$ and $\eta \in \mathfrak{K}$, where

$$\Phi_g^\eta(g') = K(g',g)\,\eta = W(U(g)\,\eta)(g'). \tag{II.14}$$

Using standard analytic techniques,[11] it then follows that $\mathfrak{H}_K$ is complete in this scalar product, hence a Hilbert space, and the map W in (II.8) is unitary. The fact that $\mathfrak{H}_K$ is a reproducing kernel Hilbert space is reflected in the generalized evaluation condition for functions Φ in $\mathfrak{H}_K$ at points $g \in G$:

$$\langle \Phi_g^\eta | \Phi \rangle_K = \langle K(\cdot\ ,\ g)\,\eta | \Phi \rangle_K = \langle \eta | \Phi(g) \rangle_{\mathfrak{H}}, \quad \eta \in \mathfrak{H}, \quad g \in G. \tag{II.15}$$

If now $\mathfrak{H}_K \subset L^2_{\mathfrak{K}}(X, d\nu)$, then for any two sections $\Phi, \Phi' \in \mathfrak{H}_K$,

$$\int_X \langle \Phi(g) | \Phi'(g) \rangle_{\mathfrak{H}}\ \ d\nu([g]) < \infty, \tag{II.16}$$

and taking the particular sections $\Phi_g^\eta, \Phi_{g'}^{\eta'}$ we get,

$$\begin{aligned} \int_X \langle K(g'',g)\,\eta | K(g'',g')\,\eta' \rangle_{\mathfrak{H}}\ \ d\nu([g'']) &= \int_X \langle \eta | U(g^{-1})\mathbb{P}([g''])U(g)\,\eta' \rangle\ \ d\nu([g'']) \\ &= \int_X \langle \eta | \mathbb{P}([g''])\,\eta' \rangle\ \ d\nu([g''], \end{aligned} \tag{II.17}$$

the last equality following from the invariance of the measure ν. Thus square-integrability of the sections Φ implies

$$I(\eta) \equiv \int_X \langle \eta | \mathbb{P}([g])\,\eta \rangle\ \ d\nu([g]) < \infty, \tag{II.18}$$

for all $\eta \in \mathfrak{K}$. In fact, as asserted in Theorem II.2 below, the above condition, which is clearly a generalization of (I.1), is both necessary and sufficient.

Definition II.1: Let U be a UIR of G on the Hilbert space $\mathfrak{H}$ and H a closed subgroup of G. Let V be a finite dimensional UIR of H. A subspace $\mathfrak{K} \subset \mathfrak{H}$ is said to be V-admissible if $U(H)$ restricted to $\mathfrak{K}$ is unitarily equivalent to V and if there exists a nonzero vector $\eta \in \mathfrak{K}$ such that the condition (II.18) is satisfied. If such a subspace $\mathfrak{K}$ exists, the representation U is said to be square integrable mod(H).

With the conditions of the above definition being satisfied, let us set

$$c_V(\eta) = \frac{I(\eta)}{\|\eta\|^2}, \tag{II.19}$$

and denote by $\mathcal{D}$ the dense set in $\mathfrak{H}$ spanned by the vectors $U(g)\,\eta$, $g \in G$. Then, in view of the second equality in (II.17), the map $W_\eta : \mathcal{D} \to L^2_{\mathfrak{K}}(X, d\nu)$,

$$(W_\eta \phi)(g) = [c_V(\eta)]^{-1/2} \mathbb{P} U(g^{-1})\,\phi, \tag{II.20}$$

is well defined. The set $\mathcal{D}$ is stable under U and W_η [being a multiple of the map W in (II.8)]) intertwines U and the induced representation ${}^V U$. Retracing the lines of the proof of Theorem I.1

(as found, for example, in Ref. 2), we can show that W_η is closed. An application of an extended version of Schur's lemma (see Ref. 2) then shows that W_η is an isometry. Thus we have proved the following result:

Theorem II.2: *Let G be a locally compact group, H a closed subgroup of G, U a UIR of G on the Hilbert space $\mathfrak{H}$ and V a finite dimensional UIR of H. Let $\mathfrak{K}$ be a V-admissible subspace of $\mathfrak{H}$. Then, the mapping (II.20) extends to all of $\mathfrak{H}$ as a linear isometry. Its range coincides with the set $\mathfrak{H}_K$, which is thus also closed in the L^2-norm as a subspace of $L^2_{\mathfrak{K}}(X,d\nu)$, and which we then denote by $\mathfrak{H}_\eta$. On $\mathfrak{H}$, the resolution of the identity*

$$\frac{1}{c_V(\eta)}\int_X \mathbb{P}([g])\ \ d\nu([g]) = I \tag{II.21}$$

holds. The subspace $\mathfrak{H}_\eta = W_\eta(\mathfrak{H}) \subset L^2_{\mathfrak{K}}(X,d\nu)$ is a reproducing kernel Hilbert space, so that the corresponding projection operator

$$\mathbb{P}_\eta = W_\eta W_\eta^*, \quad \mathbb{P}_\eta L^2_{\mathfrak{K}}(X,d\nu) = \mathfrak{H}_\eta, \tag{II.22}$$

has the reproducing kernel K_η:

$$(\mathbb{P}_\eta \Phi)(g) = \int_X K_\eta(g,g')f(g')\ \ d\nu([g']), \quad f \in L^2_{\mathfrak{K}}(X,d\nu),$$
$$K_\eta(g,g') = \frac{1}{c(\eta)}\mathbb{P}U(g^{-1})U(g')\mathbb{P}, \tag{II.23}$$

as its integral kernel. Furthermore, W_η intertwines U and the induced representation ${}^V U$,

$$W_\eta U(g) = {}^V U(g) W_\eta, \quad g \in G. \tag{II.24}$$

Note: (i) As a consequence of this theorem, the quantity $c_V(\eta)$ is independent of the particular $\eta \in \mathfrak{K}$ chosen to evaluate it. (ii) The condition (II.23) translates into the property

$$\int_X K_\eta(g,g'')K_\eta(g'',g')\ \ d\nu([g'']) = K_\eta(g,g'). \tag{II.25}$$

(iii) The set of positive, bounded operators

$$a(\Delta) = \int_\Delta \mathbb{P}([g])\ \ d\nu([g]),$$

for arbitrary Borel sets $\Delta \subset X$, form a normalized positive operator-valued measure and $\{U,a\}$ is a system of covariance:[13]

$$U(g)a(\Delta)U(g)^* = a(g\Delta).$$

It is well known (see, for example, Ref. 14) that the existence of such a system of covariance implies that U is a subrepresentation of an induced representation, which in this case is just a restatement of (II.24).

III. AN EXAMPLE USING THE ISOCHRONOUS GALILEI GROUP

We illustrate Theorem II.2 with an example drawn from the *isochronous Galilei group*. The kinematics of a free nonrelativistic physical particle is governed by its invariance under the action of the Galilei group,[15,16] which is a ten parameter group $\mathcal{G}$ of transformations of Newtonian space-time. An element $g \in \mathcal{G}$ is of the form

$$g = (b,\mathbf{a},\mathbf{v},R), \quad \mathbf{a},\mathbf{v} \in \mathbb{R}^3, \quad R \in \mathrm{SO}(3), \tag{III.1}$$

where b is a time and $\mathbf{a}$ a spatial translation, $\mathbf{v}$ a velocity boost and R a spatial rotation. Actually, we shall work with the *central extension* of $\mathcal{G}$, and also, in order to accommodate particles with half-integral spins, it is necessary to replace the rotation group SO(3) by SU(2). Denoting this extended group by $\tilde{\mathcal{G}}$, its elements will be written as $g=(\theta,b,\mathbf{a},\mathbf{v},\rho)$, $\theta\in\mathbb{R}$, $\rho\in\mathrm{SU}(2)$. The group multiplication in $\tilde{\mathcal{G}}$ is

$$g_1g_2=(\theta_1+\theta_2+\xi(g_1,g_2),b_1+b_2,\mathbf{a}_1+R(\rho_1)\mathbf{a}_2+\mathbf{v}_1b_2,\mathbf{v}_1+R(\rho_1)\mathbf{v}_2,\rho_1\rho_2),$$
$$g_i=(\theta_i,b_i,\mathbf{a}_i,\mathbf{v}_i,\rho_i),\quad i=1,2, \tag{III.2}$$

where $\xi:\mathcal{G}\times\mathcal{G}\to\mathbb{R}$ is a *multiplier*, which up to equivalence (in the sense of cocycles), can be taken to be

$$\xi(g_1,g_2)=\mu[\tfrac{1}{2}\mathbf{v}_1^2b_2+\mathbf{v}_1\cdot R(\rho_1)\mathbf{a}_2],\quad \mu=\mathrm{const}>0. \tag{III.3}$$

A couple of subgroups of $\tilde{\mathcal{G}}$ are of particular interest to us here for the construction of CS. The subgroup $\tilde{\mathcal{G}}'$, having elements $(\theta,0,\mathbf{a},\mathbf{v},\rho)$, i.e., with the time part b set equal to zero, is called the *isochronous extended Galilei group*. Another useful subgroup of $\tilde{\mathcal{G}}$, one which is also a subgroup of $\tilde{\mathcal{G}}'$, consists of elements of the type $g=(\theta,0,\mathbf{a},\mathbf{v},\mathbb{I}_2)$, where $\mathbb{I}_2$) is the 2×2 unit matrix. We denote this group by $\tilde{\mathcal{G}}_{\mathrm{WH}}$ and note that it is isomorphic to the Weyl–Heisenberg group for three degrees of freedom. Finally, the subgroup Θ, with elements $(\theta,0,\mathbf{0},\mathbb{I}_2)$, is called the phase subgroup.

The particular representations of $\tilde{\mathcal{G}}$ which concern us here are the ones which describe free nonrelativistic quantum particles of mass $\mu>0$ and spin $j=0,1/2,1,3/2,\ldots$. These are characterized by μ, j and an additional parameter $E_0\in\mathbb{R}$ which is identified with the internal energy. The *momentum space* realizations of these representations (see, e.g., Ref. 16) are carried by the Hilbert spaces $\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$, where $\mathcal{K}^j$ is a $(2j+1)$-dimensional spinor space (which we identify with $\mathbb{C}^{2j+1}$), $\hat{\mathbb{R}}^3$ the dual of $\mathbb{R}^3$ (which we again identify with $\mathbb{R}^3$), and $d\mathbf{k}$ the Lebesgue measure on $\hat{\mathbb{R}}^3$. The spaces $\mathcal{K}^j$ carry the unitary irreducible representations of SU(2), denoted $\mathcal{D}^j(\rho)$, $\rho\in\mathrm{SU}(2)$. The representation of the Galilei group in question, $\tilde{U}^{E_0,j}$, acts on the Hilbert space $\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$ in the manner:

$$(\tilde{U}^{E_0,j}(\theta,b,\mathbf{a},\mathbf{v},\rho)\tilde{\boldsymbol{\phi}})(\mathbf{k})=\exp\left[i\left\{\theta+\left(E_0+\frac{\mathbf{k}^2}{2\mu}\right)b-\mathbf{k}\cdot\mathbf{a}\right\}\right]\ \mathcal{D}^j(\rho)\ \ \tilde{\boldsymbol{\phi}}(R(\rho^{-1})(\mathbf{k}-\mu\mathbf{v})), \tag{III.4}$$

for all $\tilde{\boldsymbol{\phi}}\in\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$. This is the so-called *momentum space representation* for a nonrelativistic system of mass μ, internal energy E_0 and spin-j. When dealing with a single system, the internal energy may be set equal to zero, simply by adjusting the ground state. We shall do this in what follows and write the representations as $\tilde{U}^j$.

Unitary irreducible representations of the isochronous Galilei group $\tilde{\mathcal{G}}'$ can now be read off from (III.4) after setting $b=0$. These representations are also carried by $\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$, and using the same notation as before (also taking $E_0=0$),

$$(\tilde{U}^j(\theta,\mathbf{a},\mathbf{v},\rho)\tilde{\boldsymbol{\phi}})(\mathbf{k})=\exp[i(\theta-\mathbf{k}\cdot\mathbf{a})]\ \mathcal{D}^j(\rho)\ \ \tilde{\boldsymbol{\phi}}(R(\rho^{-1})(\mathbf{k}-\mu\mathbf{v})),\quad (\theta,\mathbf{a},\mathbf{v},\rho)\in\tilde{\mathcal{G}}'. \tag{III.5}$$

We proceed to demonstrate the square integrability mod$(\Theta\otimes\mathrm{SU}(2))$ of these representations in the sense of Theorem II.2.

Let us first decompose the representation space $\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$ into rotationally invariant subspaces. For any $\tilde{\boldsymbol{\phi}}\in\mathcal{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$, and $\rho\in\mathrm{SU}(2)$,

$$(\tilde{U}^j(0,\mathbf{0},\mathbf{0},\rho)\tilde{\boldsymbol{\phi}}=\mathcal{D}^j(\rho)\tilde{\boldsymbol{\phi}}(R(\rho^{-1})\mathbf{k}). \tag{III.6}$$

The space $L^2(\hat{\mathbb{R}}^3, d\mathbf{k})$ can be written as the tensor product $L^2(S^2, d\Omega(\mathbf{k})) \otimes L^2(\mathbb{R}^+, r^2 dr)$, where S^2 is the unit sphere, $d\Omega\mathbf{k})$ the angular measure in the angles of $\mathbf{k}$ and $r = \|\mathbf{k}\|$. Let $\{e_n\}_{n=1}^{\infty}$ be an orthonormal basis of $L^2(\mathbb{R}^+, r^2 dr)$:

$$\int_{\mathbb{R}^+} \overline{e_n(r)} e_m(r) r^2 dr = \delta_{nm}. \tag{III.7}$$

For each $\ell = 0,1,2,\ldots$, the spherical harmonics $Y^{\ell m}(\mathbf{k})$ (in the angles of $\mathbf{k}$), $m = -\ell, -\ell + 1, \ldots, \ell - 1, \ell$, span a subspace of $L^2(S^2, d\Omega(\mathbf{k}))$ which is stable under rotations. Furthermore,

$$f_{n\ell m}(R^{-1}\mathbf{k}) = \sum_{m'} \mathcal{D}^{\ell}(R)_{m'm} f_{n\ell m'}(\mathbf{k}), \tag{III.8}$$

where

$$f_{n\ell m}(\mathbf{k}) = Y^{\ell m}(\mathbf{k}) \quad e_n(\|\mathbf{k}\|). \tag{III.9}$$

For fixed ℓ, n, let $\mathfrak{K}^{\ell n}$ be the subspace of $L^2(\hat{\mathbb{R}}^3, d\mathbf{k})$ spanned by the unit vectors $f_{n\ell m}$, $m = -\ell, -\ell + 1, \ldots, \ell - 1, \ell$. Then,

$$L^2(\hat{\mathbb{R}}^3, d\mathbf{k}) = \bigoplus_{n=1}^{\infty} \bigoplus_{\ell=0}^{\infty} \mathfrak{K}^{\ell n}, \tag{III.10}$$

and each subspace $\mathfrak{K}^{\ell n}$ carries a representation of SU(2) unitarily equivalent to the UIR $\mathcal{D}^{\ell}$.

Next, in view of the well-known decomposition (e.g., for angular momenta)

$$\mathcal{D}^j(\rho) \otimes \mathcal{D}^{\ell}(\rho) \simeq \bigoplus_{J=|j-\ell|}^{j+\ell} \mathcal{D}^J(\rho), \tag{III.11}$$

($\simeq$ implying unitary equivalence), the space $\mathfrak{K}^j \otimes L^2(\hat{\mathbb{R}}^3, d\mathbf{k})$ decomposes as

$$\mathfrak{K}^j \otimes L^2(\hat{\mathbb{R}}^3, d\mathbf{k}) = \bigoplus_{n=1}^{\infty} \bigoplus_{\ell=1}^{\infty} \bigoplus_{J=|j-\ell|}^{j+\ell} \mathfrak{K}^{nJ\ell}, \tag{III.12}$$

in which each $\mathfrak{K}^{nJ\ell}$ is a $(2J+1)$-dimensional space, carrying a UIR of SU(2), unitarily equivalent to $\mathcal{D}^J$. The vectors

$$\tilde{\boldsymbol{\eta}}^{nJM\ell}(\mathbf{k}) = \sum_m C(j,m;\ \ell, M-m|JM) \boldsymbol{\chi}^{jm} \quad Y^{\ell,\ M-m}(\mathbf{k}) e_n(\|\mathbf{k}\|), \tag{III.13}$$

where $\boldsymbol{\chi}^{jm}$ is the $2j+1$ component column vector with 1 in the mth position and zeroes elsewhere and the $C(j,m;\ \ell, M-m|JM)$ are SU(2) Clebsch–Gordan coefficients, form an orthonormal basis in $\mathfrak{K}^{nJ\ell}$ and satisfy

$$\tilde{U}^j(0,\mathbf{0},\mathbf{0},\rho)\tilde{\boldsymbol{\eta}}^{nJM\ell} = \mathcal{D}^J(\rho)\tilde{\boldsymbol{\eta}}^{nJM\ell}, \quad M = -J, -J+1, \ldots, J-1, J. \tag{III.14}$$

The sum over m in (III.13) runs through a subset of $-|j-\ell|, -|j-\ell|+1, \ldots, j+\ell$, determined by the constraints $-j \leq m \leq j$ and $-\ell \leq M-m \leq \ell$.

Let V^J denote the UIR of the subgroup $H = \Theta \otimes \mathrm{SU}(2)$ of $\tilde{\mathcal{G}}'$:

$$V^J(\theta,\rho) = e^{i\theta}\mathcal{D}^J(\rho). \tag{III.15}$$

Then (III.14) implies that for any vector $\tilde{\boldsymbol{\eta}} \in \mathfrak{K}^{nJ\ell}$

$$\tilde{U}^j(\theta,\mathbf{0},\mathbf{0},\rho)\tilde{\boldsymbol{\eta}} = V^J(\theta,\rho)\tilde{\boldsymbol{\eta}}. \tag{III.16}$$

This also means that the projection operator,

$$\mathbb{P}^{nJ\ell} = \sum_{M=-J}^{J} |\tilde{\boldsymbol{\eta}}^{nJM\ell}\rangle\langle\tilde{\boldsymbol{\eta}}^{nJM\ell}|, \tag{III.17}$$

for the subspace $\mathfrak{K}^{nJ\ell}$ is a minimal projector satisfying

$$\tilde{U}^j(0,\mathbf{0},\mathbf{0},\rho)\mathbb{P}^{nJ\ell}\tilde{U}^j(0,\mathbf{0},\mathbf{0},\rho)^* = \mathbb{P}^{nJ\ell}, \tag{III.18}$$

for all $\rho \in \mathrm{SU}(2)$ [minimality implying that there is no other projector $\mathbb{P}' \subset \mathbb{P}^{nJ\ell}$ also satisfying (III.18)]. The space $\Gamma = \mathcal{G}'/\Theta \otimes \mathrm{SU}(2)$ has a natural identification with the phase space of a free classical particle of mass μ, and hence is topologically isomorphic to $\mathbb{R}^6$, with global coordinates $(\mathbf{q},\mathbf{p})$, denoting position and momentum, respectively, and such that the measure $d\mathbf{q}d\mathbf{p}$ is $\mathcal{G}'$-invariant. Furthermore, a straightforward computation using properties of spherical harmonics shows that, for any section $\sigma:\Gamma \to \mathcal{G}'$,

$$\begin{aligned} c_{V^J}(\tilde{\boldsymbol{\eta}}^{nJM'\ell}) &= \sum_{M=-J}^{J} \int_{\Gamma} |\langle \tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})} | \tilde{\boldsymbol{\eta}}^{nJM'\ell}\rangle|^2 \; d\mathbf{q}d\mathbf{p} \\ &= (2\pi)^3 \sum_{M=-J}^{J} \sum_{m} |C(j,m;\ \ell, M-m|JM)|^2 |C(j,m;\ \ell, M'-m|JM')|^2. \end{aligned} \tag{III.19}$$

where of course, $\tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})} = U(\sigma(\mathbf{q},\mathbf{p}))\tilde{\boldsymbol{\eta}}^{nJM\ell}$. Thus the subspace $\mathfrak{K}^{nJ\ell}$ of $\mathfrak{K}^j \otimes L^2(\hat{\mathbb{R}}^3, d\mathbf{k})$ is V^J-admissible, and the representation $\tilde{U}^j$ is square integrable mod$(\Theta \otimes \mathrm{SU}(2))$, which also implies an interesting relation among Clebsch–Gordan coefficients, namely, that the sum on the right-hand side of (III.19) is independent of M', and indeed $c_{V^J}(\tilde{\boldsymbol{\eta}}^{nJM'\ell}) = c_{V^J}(\tilde{\boldsymbol{\eta}})$ for any vector $\tilde{\boldsymbol{\eta}} \in \mathfrak{K}^{nJ\ell}$.

Consequently, for each $n = 1,2,\ldots,$ and each $J = |j-\ell|, |j-\ell|+1, \ldots, j+\ell$, the map $W_{nJ\ell}: \mathfrak{K}^j \otimes L^2(\hat{\mathbb{R}}^3, d\mathbf{k}) \to \mathfrak{K}^J \otimes L^2(\Gamma, d\mathbf{q}d\mathbf{p})$:

$$(W_{nJ\ell}\tilde{\boldsymbol{\phi}})^M(\mathbf{q},\mathbf{p}) = [c_{V^J}]^{-1/2} \langle \tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})} | \tilde{\boldsymbol{\phi}}\rangle, \quad M = -J, -J+1, \ldots, J-1, J, \tag{III.20}$$

is an isometry. On $\mathfrak{K}^j \otimes L^2(\hat{\mathbb{R}}^3, d\mathbf{k})$ the resolution of the identity,

$$[c_{V^J}]^{-1} \sum_{M=-J}^{J} \int_{\Gamma} |\tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})}\rangle\langle \tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})}| \; d\mathbf{q}d\mathbf{p} = I, \tag{III.21}$$

is valid for the class of *Galilei coherent states*,

$$[c_{V^J}]^{-1/2} \; \tilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma(\mathbf{q},\mathbf{p})}, \quad (\mathbf{q},\mathbf{p}) \in \mathbb{R}^6, \quad M = -J, -J+1, \ldots, J-1, J. \tag{III.22}$$

Explicitly, using the section $\sigma = \sigma_0$, where

$$\sigma_0(\mathbf{q},\mathbf{p}) = \left(0, \mathbf{q}, \frac{\mathbf{p}}{\mu}, \mathbb{I}_2\right), \tag{III.23}$$

the CS are the functions,

$$\frac{1}{[c_{V^J}]^{1/2}}\ \widetilde{\boldsymbol{\eta}}^{nJM\ell}_{\sigma_0(\mathbf{q},\mathbf{p})}(\mathbf{k})$$

$$=\frac{1}{[c_{V^J}]^{1/2}}\ \left(\widetilde{U}^j\left(0,\mathbf{q},\frac{\mathbf{p}}{\mu},\mathbb{I}_2\right)\widetilde{\boldsymbol{\eta}}^{nJM\ell}\right)(\mathbf{k})$$

$$=\frac{1}{[c_{V^J}]^{1/2}}e^{-i\mathbf{k}\cdot\mathbf{q}}\ \widetilde{\boldsymbol{\eta}}^{nJM\ell}(\mathbf{k}-\mathbf{p})$$

$$=\frac{1}{[c_{V^J}]^{1/2}}e^{-i\mathbf{k}\cdot\mathbf{q}}\sum_m\ C(j,m;\ \ell,M-m|JM)\boldsymbol{\chi}^{jm}Y^{\ell,M-m}(\mathbf{k}-\mathbf{p})e_n(\|\mathbf{k}-\mathbf{p}\|), \tag{III.24}$$

$(M=-J,-J+1,\ldots,J-1,J)$ which in general are vector coherent states.

The image of $\mathfrak{K}^j\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$ in $\mathfrak{K}^J\otimes L^2(\Gamma,d\mathbf{q}d\mathbf{p})$ under the isometry $W_{nJ\ell}$ is a reproducing kernel Hilbert space $\mathfrak{H}_{nJ\ell}$. Its elements are vector-valued functions $\boldsymbol{\Phi}$ with components,

$$\Phi^M(\mathbf{q},\mathbf{p})=\frac{1}{[c_{V^J}]^{1/2}}\sum_m\ C(j,m;\ \ell,M-m|JM)\int_{\hat{\mathbb{R}}^3}e^{i\mathbf{k}\cdot\mathbf{q}}\ \overline{Y^{\ell,M-m}(\mathbf{k}-\mathbf{p})}\ \overline{e_n(\|\mathbf{k}-\mathbf{p}\|)}\widetilde{\phi}^m(\mathbf{k})\ \ d\mathbf{k},$$

$$M=-J,-J+1,\ldots,J-1,J. \tag{III.25}$$

The reproducing kernel for $\mathfrak{H}_{nJ\ell}$ is matrix-valued with matrix elements

$$[\mathbf{K}_{nJ\ell}(\mathbf{q},\mathbf{p};\ \mathbf{q}'\mathbf{p}')]_{MM'}=\frac{1}{[c_{V^J}]^{1/2}}\sum_m\ \overline{C(j,m;\ \ell,M-m|JM)}C(j,m;\ \ell,M'-m|JM')$$
$$\times\int_{\mathbb{R}^3}\exp[i\mathbf{k}\cdot(\mathbf{q}-\mathbf{q}')]\ \ \overline{Y^{\ell,M-m}(\mathbf{k}-\mathbf{p})}Y^{\ell,M'-m}(\mathbf{k}-\mathbf{p}')$$
$$\times\overline{e_n(\|\mathbf{k}-\mathbf{p}\|)}e_n(\|\mathbf{k}-\mathbf{p}'\|)\ \ d\mathbf{k}, \tag{III.26}$$

$M,M'=-J,-J+1,\ldots,J-1,J$. Finally, the image of $\widetilde{U}^j$ in $\mathfrak{H}_{nJ\ell}\subset\mathfrak{K}^J\otimes L^2(\Gamma,d\mathbf{q}d\mathbf{p})$ under $W_{nJ\ell}$ is

$$\mathbf{U}_{nJ\ell}(g)=W_{nJ\ell}\widetilde{U}^j(g)W^{-1}_{nJ\ell},\ \ g\in\widetilde{\mathcal{G}}',$$

$$(\mathbf{U}_{nJ\ell}(\theta,\mathbf{a},\mathbf{v},\rho)\boldsymbol{\Phi})(\mathbf{q},\mathbf{p})=\exp[i\{\theta+\mu\mathbf{v}\cdot(\mathbf{q}-\mathbf{a})\}]$$
$$\times\mathcal{D}^J(\rho)\boldsymbol{\Phi}(R(\rho^{-1})(\mathbf{q}-\mathbf{a}),\ \ R(\rho^{-1})(\mathbf{p}-\mu\mathbf{v})), \tag{III.27}$$

which again extends to the whole of $\mathfrak{K}^J\otimes L^2(\Gamma,d\mathbf{q}d\mathbf{p})$ as the representation induced from V^J.

IV. DISCUSSION

In conclusion, observe that since the isochronous Galilei group $\widetilde{\mathcal{G}}'$ is essentially the Weyl-Heisenberg group for three degrees of freedom together with rotations, the coherent states (III.22) are also the VCS for this latter group. Also, it should be noted that the coherent states (III.24) are all possible such states arising from the Galilei group, which are labeled by phase space points, which satisfy a resolution of the identity and incorporate spin degrees of freedom. The time evolved versions of these states obey the Schrödinger equation. The vectors $\widetilde{\boldsymbol{\eta}}^{nJM\ell}$, $M=-J,-J-1,\ldots,J$, from which the coherent states (III.22) are built, can be looked upon as quantum mechanical wave functions for a system with spin-j, orbital angular momentum ℓ and total angular momentum J. Thus the associated coherent states (III.22) should be useful in analyzing molecular or atomic systems with such spin and angular momentum content. Actually, the reso-

lution of the identity (III.21) is not so surprising (it is easy to check it by direct computation) and in fact a much more general expression can be obtained as follows: Let $\eta^i, \quad i=1,2,\ldots,n$, be any set of orthonormal vectors in $\mathbb{C}^{2j+1}\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$, spanning the subspace $\mathfrak{K}$ having projection operator $\mathbb{P}$. Then by direct computation it can be seen that

$$\frac{1}{n(2\pi)^3}\int_\Gamma \mathbb{P}(\mathbf{q},\mathbf{p})\ \ d\mathbf{q}d\mathbf{p}=I, \quad \mathbb{P}(\mathbf{q},\mathbf{p})=\tilde{U}^j(\sigma(\mathbf{q},\mathbf{p}))\mathbb{P}\tilde{U}^j(\sigma(\mathbf{q},\mathbf{p}))^*, \tag{IV.1}$$

for any one of the sections [obtained from the basic section (III.23)]:

$$\sigma(\mathbf{q},\mathbf{p})=\sigma_0(\mathbf{q},\mathbf{p})(\theta(\mathbf{q},\mathbf{p}),\mathbf{0},\mathbf{0},\rho), \tag{IV.2}$$

for fixed $\rho\in \mathrm{SU}(2)$, and where $\theta(\mathbf{q},\mathbf{p})$ is an arbitrary, real valued, measurable function of $(\mathbf{q},\mathbf{p})$. Taking the η^i to be n eigenstates of some observable (e.g., the Hamiltonian of some atomic system in an external potential), one could use the more general coherent states $[n(2\pi)^3]^{-1/2}\eta^i_{\sigma(\mathbf{q},\mathbf{p})}$, $i=1,2,\ldots,n$, $(\mathbf{q},\mathbf{p})\in\Gamma$, in atomic computations involving such n-level systems. Even more generally, if F is any trace-class operator on $\mathbb{C}^{2j+1}\otimes L^2(\hat{\mathbb{R}}^3,d\mathbf{k})$, then with σ as in (IV.2) and $c(F)=(2\pi)^3[\mathrm{Tr}\ \ F]$,

$$\frac{1}{c(F)}\int_\Gamma F(\mathbf{q},\mathbf{p})\ \ d\mathbf{q}d\mathbf{p}=I, \quad F(\mathbf{q},\mathbf{p})=\tilde{U}^j(\sigma(\mathbf{q},\mathbf{p}))F\tilde{U}^j(\sigma(\mathbf{q},\mathbf{p}))^*. \tag{IV.3}$$

From the point of view of quantization theory, Theorem II.2 could be used to extend the method of coherent states quantization (such as in Ref. 17) to include internal degrees of freedom as well. One could also exploit in this way the rich theory of Bergman type kernels associated to holomorphic, Hermitian vector bundles (as worked out, e.g., in Ref. 18). Finally, it would appear that a quantization using the coherent states (III.22) amounts to quantizing a particle with spin using the classical phase space of a spin-0 system. However, there is a great deal of flexibility attained in transferring the spin degrees of freedom to the fibers of a vector bundle rather than working with a classical phase space which artificially incorporates ''classical spin degrees of freedom.''

ACKNOWLEDGMENTS

Part of this work was done while the author was visiting the Arnold Sommerfeld Institut für Mathematische Physik, Technische Universität Clausthal, Clausthal, Germany, with support from the Alexander von Humboldt-Stiftung, Bonn. He would like to thank Professor H.-D. Doebner for hospitality at the Arnold Sommerfeld Institut and the Humboldt-Stiftung for its support. Thanks are due to S. Berceanu for comments and suggestions. The author is also grateful to Stephan De Bièvre for critical comments and extensive suggestions on an earlier draft of the manuscript, which helped greatly to improve the clarity and conciseness of the presentation.

[1] S. T. Ali, J.-P. Antoine, J.-P. Gazeau, and U.-A. Müller, ''Coherent states and their generalizations: A mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995).
[2] A. Grossmann, J. Morlet, and T. Paul, ''Integral transforms associated to square integrable representations. I. General results,'' J. Math. Phys. **26**, 2473–2479 (1985).
[3] C. Quesne, ''Generalized vector coherent state theory of $sp(2n,\mathbb{R})$ vector operators and of $sp(2n,\mathbb{R})\supset u(N)$ reduced Wigner coefficients,'' J. Phys. A: Math. Gen. **24**, 2697–2174 (1991).
[4] D. J. Rowe, G. Rosensteel, and R. Gilmore, ''Vector coherent state representation theory,'' J. Math. Phys. **26**, 2787–2791 (1985).
[5] D. J. Rowe and J. Repka, ''Vector-coherent-state theory as a theory of induced representations,'' J. Math. Phys. **32**, 2614–2634 (1991).
[6] R. Gilmore, ''Geometry of symmetrized states,'' Ann. Phys. (N.Y.) **74**, 391–463 (1972); ''On properties of coherent states,'' Rev. Mex. Fis. **23**, 143–187 (1974).
[7] A. Perelomov, *Generalized Coherent States and their Applications* (Springer-Verlag, Berlin, 1986).
[8] D. M. Healy, Jr., and F. E. Schroeck, Jr., ''On informational completeness of covariant localization observables and Wigner coefficients,'' J. Math. Phys. **36**, 453–507 (1995).
[9] J. H. Rawnsley, ''Coherent states and Kähler manifolds,'' Quart. J. Math., Oxford **28**, 403–415 (1977).
[10] G. W. Mackey, ''Imprimitivity for representations of locally compact groups I,'' Proc. Natl. Acad. Sci. USA **35**, 537–545 (1949).

[11] U. Cattaneo, ''Densities of covariant observables,'' J. Math. Phys. **23**, 659–664 (1982).
[12] H. Meschkowsky, *Hilbertsche Raüme mit Kernfunktionen* (Springer-Verlag, Berlin, 1962).
[13] S. T. Ali, J-P. Antoine, and J-P. Gazeau, ''Square integrability of group representations on homogeneous spaces. II. Coherent and quasi-coherent states. The case of the Poincaré group,'' Ann. Inst. Henri Poincaré: Phys. Theor. **55**, 857–890 (1991).
[14] H. Scutaru, ''Coherent states and induced representations,'' Lett. Math. Phys. **2**, 101–107 (1977).
[15] J-M. Lévy-Leblond, ''Galilei group and non-relativistic quantum mechanics,'' J. Math. Phys. **4**, 453–507 (1963).
[16] J-M. Lévy-Leblond, ''Galilei group and Galilean invariance,'' in *Group Theory and its Applications*, edited by Ernest M. Loebl (Academic, New York, 1971), Vol. II, pp. 221–299.
[17] A. Odzijewicz, ''Coherent states and geometric quantization,'' Commun. Math. Phys. **150**, 385–413 (1992).
[18] Z. Pasternak-Winiarski and J. Wojcieszynski, ''Bergman spaces and kernels for holomorphic vector bundles,'' Demonstratio Mathematica **30**, 199–214 (1997).

Wavelet transforms and discrete frames associated to semidirect products

Paolo Aniello[a] and Gianni Cassinelli[b]
Dipartimento di Fisica, Università di Genova, I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy

Ernesto De Vito[c]
Dipartimento di Matematica, Università di Modena, Via Campi 213/B, 41100 Modena, Italy and I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy

Alberto Levrero[d]
Dipartimento di Fisica, Università di Genova, I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy

(Received 11 August 1997; accepted for publication 4 February 1998)

We consider a semidirect product $G=\mathbb{R}^n\times' H$ and its unitary representations U of the form $\mathrm{Ind}_{G_0}^{G}(p_0 m)$ where Ind is the unitary induction, p_0 is in the dual group of $\mathbb{R}^n$, G_0 is the stability group of p_0, and m is a unitary representation of $G_0\cap H$. We give sufficient conditions such that U defines a wavelet transform and a discrete frame. © *1998 American Institute of Physics.* [S0022-2488(98)00208-4]

I. INTRODUCTION

In the framework of *wavelet analysis* it is well known that the groups that are semidirect products with normal Abelian factor play an essential role in the construction of both continuous wavelet transforms and of discrete frames.

The classical example, Ref. 1, is the group $ax+b$ and its quasi-regular representation U acting on $L^2(\mathbb{R})$ as

$$(U_{ab}f\ )(x)=a^{-1/2}f\left(\frac{x-b}{a}\right),\quad x\in\mathbb{R},\ \ f\in L^2(\mathbb{R}),\ \ a\in\mathbb{R}^+,\ \ b\in\mathbb{R}.$$

We refer for this to the treatise of Ref. 2 and to the review paper Ref. 3.

In recent years there has been an interest in extending these kind of results to other groups: on one side, for example, Refs. 4–6, obtaining results valid for a fairly general class of groups; on the other side, discussing specific cases. For example, in Ref. 7, the group $\mathbb{R}^n\times'(\mathbb{R}^+\times\mathrm{SO}(n))$ has been considered in order to define wavelet transforms on $L^2(\mathbb{R}^n)$.

In this paper we consider the case of groups G that are semidirect product of $\mathbb{R}^n$ and a topological group H and their representations U of the form $U=\mathrm{Ind}_{G_0}^{G}(p_0 m)$ where Ind is the unitary induction *à la* Mackey, $p_0\in\mathbb{P}^n$, the dual of $\mathbb{R}^n$, G_0 is the stability group of p_0 with respect to the dual action of G on $\mathbb{P}^n$, and m is a unitary representation of $G_0\cap H$. For such groups we consider two kinds of problems.

First we describe the conditions under which the representations $\mathrm{Ind}_{G_0}^{G}(p_0 m)$ are square-integrable and characterize for them the formal degree operator. Moreover we discuss a concrete example that extends the results of Ref. 7 to vector-valued functions.

Second we consider the problem of the existence of a countable set $(g_i)_{i\in I}$ of G and a vector $\psi\in\mathscr{H}$ such that $(U_{g_i}\psi)_{i\in I}$ is a *frame* in $\mathscr{H}$, denoting by $\mathscr{H}$ the Hilbert space where U acts. In particular we show that, if H is the direct product of $\mathbb{R}^+$ (the multiplicative group of positive numbers) and a semisimple connected Lie group, and m is a cyclic finite-dimensional represen-

[a]Electronic mail: aniello@genova.infn.it
[b]Electronic mail: cassinelli@genova.infn.it
[c]Electronic mail: devito@genova.infn.it
[d]Electronic mail: levrero@genova.infn.it

 © 1998 American Institute of Physics

tation, then there are discrete frames associated to $\mathrm{Ind}_{G_0}^{G}(p_0 m)$ provided that the stability group $H \cap G_0$ is compact and that the orbit $G[p_0]$ is open. These results extend the ones obtained by Refs. 4 and 6 since we consider wavelet transforms for vector-valued functions and we *do not* assume the existence either of a *fundamental domain*[6] or of a *frame generator*.[4]

II. PRELIMINARIES AND NOTATIONS

Let G be a *locally compact second countable* (lcsc) topological group. We denote by μ_G and Δ_G respectively a left Haar measure on G and its modular function.

We review the basic definition and properties of square integrable representations and of discrete frames. In this paper we will use the word *representation* to mean a continuous unitary representation of G acting in a complex separable Hilbert space $\mathscr{H}$.

A representation U of G in $\mathscr{H}$ is *square integrable* if it is irreducible and there exist two nonzero vectors $\phi, \psi \in \mathscr{H}$ such that the map $c_{\phi,\psi}$ from G to $\mathbb{C}$,

$$c_{\psi,\phi}(g) = \langle U_g \psi, \phi \rangle_{\mathscr{H}},$$

is in $L^2(G, \mu_G)$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product (linear in the second variable). We recall this well-known result of Duflo and Moore (Theorem 3 of Ref. 8).

Theorem 1: (Duflo and Moore) *Let U be a square-integrable representation of G acting in $\mathscr{H}$. There exists a unique positive self-adjoint operator K_U in $\mathscr{H}$ such that*

(1) *if $\phi, \psi \in \mathscr{H}$, $\phi \neq 0$, the map $c_{\psi,\phi}$ is in $L^2(G, \mu_G)$ if and only if $\psi \in \mathrm{Dom}\, K_U^{-1/2}$; and*
(2) *if $\phi, \phi' \in \mathscr{H}$ and if $\psi, \psi' \in \mathrm{Dom}\, K_U^{-1/2}$, then*

$$\langle c_{\psi,\phi}, c_{\psi',\phi'} \rangle_{L^2(G,\mu_G)} = \langle \phi, \phi' \rangle_{\mathscr{H}} \langle K_U^{-1/2} \psi', K_U^{-1/2} \psi \rangle_{\mathscr{H}}.$$

The operator K_U is called the *formal degree* of the representation and the vectors $\psi \in \mathrm{Dom}\, K_U^{-1/2}$ are called *admissible*.

Let $\mathscr{H}$ be a (complex separable) Hilbert space and $\{\psi_i : i \in I\}$ a countable family of nonzero vectors in $\mathscr{H}$. We will say that $\{\psi_i\}_{i \in I}$ is a *frame* in $\mathscr{H}$ if there exist two positive numbers α and β such that the following condition holds:

$$\alpha \|\phi\|^2 \leqslant \sum_{i \in I} |\langle \psi_i, \phi \rangle|^2 \leqslant \beta \|\phi\|^2 \quad \forall \phi \in \mathscr{H}.$$

The numbers α and β are called *frame bounds* of the frame $\{\psi_i\}_{i \in I}$.

If there exists a representation of a group G acting in $\mathscr{H}$, then it is of particular interest to determine frames of the following form:

$$\psi_i = U_{g_i} \psi,$$

where $\{g_i\}_{i \in I}$ is a countable family of elements of G and ψ is a fixed element in $\mathscr{H}$.

From now on let G be the semidirect product $\mathbb{R}^n \times' H$, where H is a lcsc topological group acting continuously on $\mathbb{R}^n$ (G is in a natural way a lcsc topological group). Explicitly, let Ω be the representation of H that defines the action of H on $\mathbb{R}^n$. Denoting $g = (x, h)$ the elements of G, with $x \in \mathbb{R}^n$ and $h \in H$, we have

$$(x,h)(x',h') = (x + \Omega(h)x', hh'),$$

and the inner action of G on $\mathbb{R}^n$ is

$$g[x] := gxg^{-1} = \Omega(h)x, \quad x \in \mathbb{R}^n, \quad g = (y,h) \in G.$$

We denote by $d^n x$ the Lebesgue measure on $\mathbb{R}^n$. Let $\mathbb{P}^n$ be the dual group of $\mathbb{R}^n$, which we identify with $\mathbb{R}^n$ in a natural way, and $d^n p$ be the corresponding Lebesgue measure. By duality, G acts continuously on $\mathbb{P}^n$, explicitly

$$g[p]\cdot x = p\cdot g^{-1}[x], \quad x\in\mathbb{R}^n,\ p\in\mathbb{P}^n,\ g\in G,$$

where $\cdot$ denotes the natural scalar product in $\mathbb{R}^n$. We fix an element p_0 of $\mathbb{P}^n$ and we denote by $\mathcal{O}$ the corresponding orbit

$$\mathcal{O}=\{p\in\mathbb{P}^n : p=g[p_0], g\in G\}$$

and by G_0 the stability subgroup of p_0,

$$G_0=\{g\in G : g[p_0]=p_0\}.$$

Let $H_0=G_0\cap H$, so that $G_0=\mathbb{R}^n\times' H_0$. We choose a representation m of H_0 acting in a complex separable Hilbert space $\mathcal{K}$ and we consider the representation p_0m of G_0 in $\mathcal{K}$:

$$(p_0m)(xh) := e^{ip_0\cdot x}m(h), \quad x\in\mathbb{R}^n, h\in H_0.$$

We denote by $\mathrm{Ind}_{G_0}^{G}(p_0m)$ the representation unitarily induced by p_0m from G_0 to G.[9]

III. SQUARE INTEGRABILITY

The following theorem solves the problem of the square integrability of the induced representations of G.

Theorem 2: *The representation* $\mathrm{Ind}_{G_0}^{G}(p_0m)$ *is square integrable if and only if* $\mathcal{O}$ *has positive Lebesgue measure in* $\mathbb{P}^n$ *and* m *is a square-integrable representation.*

This result was proven by Kleppner and Lipsman (Corollary 11.1 of Ref. 10) as a consequence of the characterization of the Plancherel measure for a locally compact group with a closed normal type I subgroup. In Ref. 11 we give an elementary proof of this theorem, based only on the properties of the commutative Fourier transform.

Let us assume that $\mathrm{Ind}_{G_0}^{G}(p_0m)$ is square integrable. Since the square integrability of a representation depends only on its equivalence class, we can freely choose the explicit form of $\mathrm{Ind}_{G_0}^{G}(p_0m)$. Due to Theorem 2 the measure of $\mathcal{O}$ has to be positive, hence the restriction of the Lebesgue measure d^np to $\mathcal{O}$ is a nonzero σ-finite measure defined on $\mathcal{O}$. In fact it is H-quasi-invariant, due to the following elementary lemma (the proof is given, for example, in Ref. 11). Let $\rho(h)=\Delta_H(h)/\Delta_G(h)$.

Lemma 1: For any $h\in H$ *and* $E\in\mathcal{B}(\mathbb{P}^n)$ *we have*

$$(d^np)(h[E])=\rho(h^{-1})(d^np)(E).$$

With the above notations, a suitable choice is the representation U acting in $L^2(\mathcal{O},\mathcal{K})$ as

$$(U_{xh}\phi)(p)=\rho(h)^{-1/2}e^{ip\cdot x}m(q(p)^{-1}hq(h^{-1}[p]))\phi(h^{-1}[p]), \tag{1}$$

where $p\in\mathcal{O}$, $x\in\mathbb{R}^n$, $h\in H$, $\phi\in L^2(\mathcal{O},\mathcal{K})$, and q is a measurable section for the action of H on $\mathcal{O}$, that is, a measurable map $q:\mathcal{O}\to H$ such that

$$q(p_0)=e \quad \text{where } e \text{ is the identity},$$

$$q(p)[p_0]=p, \quad p\in\mathcal{O}$$

(we recall that the orbit $\mathcal{O}$ is a Borel subset of $\mathbb{R}^n$).

With this choice the condition of admissibility is given by the following corollary (for the proof see Corollary 2 of Ref. 11). Let K_m be the formal degree of m.

Corollary 1: With the above assumptions, $\psi\in L^2(\mathcal{O},\mathcal{K})$ *is an admissible vector for* U *if and only if the following two conditions hold:*

$$\psi(p) \in \mathrm{Dom}\ K_m^{-1/2} \quad \text{for almost all } p \in \mathcal{O},$$

$$\int_{\mathcal{O}} \Delta_G(G(p))^{-1} \|K_m^{-1/2}\psi(p)\|_{\mathcal{K}}^2 \, d^n p < \infty.$$

If H_0 is compact, all its irreducible representations are square integrable, and one can easily show that $\mathrm{Dom}\ K_m^{-1/2} = \mathcal{K}$, K_m is a multiple of the identity, and $\Delta_G(q(p))$ is *independent* of the choice of the section q, precisely

$$\Delta_G(q(p)) = \Delta_G(g_p) \quad \forall g_p \in G \quad \text{such that } g_p[p_0] = p.$$

Going back to the general case, if U is square integrable, we can define a corresponding *wavelet transform*. Let $\psi \in L^2(\mathcal{O}, \mathcal{K})$ be an admissible vector normalized in a such way that

$$\int_{\mathcal{O}} \Delta_G(g(p))^{-1} \|K_m^{-1/2}\psi(p)\|_{\mathcal{K}}^2 \, d^n p = 1.$$

Let $\mathcal{W}_\psi$ be the map from $L^2(\mathcal{O}, \mathcal{K})$ to $L^2(G, \mu_G)$ defined by

$$(\mathcal{W}_\psi \phi)(g) = \langle U_g \psi, \phi \rangle L^2(\mathcal{O}, \mathcal{K}), \quad g \in G, \phi \in L^2(\mathcal{O}, \mathcal{K}).$$

As a consequence of Theorem 1, one can easily prove that $\mathcal{W}_\psi$ is an isometry and the projection on its range is a convolution operator, precisely

$$F \in \mathrm{Ran}\ \mathcal{W}_\psi \Leftrightarrow F * \mathcal{W}_\psi \psi = F,$$

where $*$ denotes the convolution in $L^2(G, \mu_G)$. Moreover, the inverse map of $\mathcal{W}_\psi$ from $\mathrm{Ran}(\mathcal{W}_\psi)$ onto $L^2(\mathcal{O}, \mathcal{K})$ is given by

$$F \mapsto \int_G F(g)(\mathcal{W}_\psi \psi)(g) \, d\mu_G(g),$$

where the integral is in the weak sense.

To present this kind of result, we consider a specific example. Let $G = \mathbb{R}^n \times' (\mathbb{R}^+ \times \mathrm{SO}(n))$, where $\mathrm{SO}(n)$ is the n-dimensional special orthogonal group. The direct product $H = \mathbb{R}^+ \times \mathrm{SO}(n)$ acts on $\mathbb{R}^n$ as

$$(aR)[x] = aRx,$$

where Rx is the usual matrix action and a acts as a scalar. The group H is unimodular and the Haar measure is $a^{-1} da dR$, where dR is the normalized Haar measure on $\mathrm{SO}(n)$. The modular function of G is $\Delta_G(xaR) = a^{-n}$. The dual action is

$$(aR)[p] = a^{-1} Rp.$$

With respect to this action, $\mathbb{P}^n$ is the union of two orbits $\{0\}$ and $\mathcal{O} = \mathbb{P}^n \backslash \{0\}$. The first orbit has zero Lebesgue measure, so that the corresponding representations are not square integrable. We consider the orbit $\mathcal{O}$ and we fix $p_0 = (0,...,0,1) \in \mathcal{O}$. The stability group is $G_0 = \mathbb{R}^n \times' \mathrm{SO}(n-1)$ where $\mathrm{SO}(n-1)$ is regarded as a subgroup of $\mathrm{SO}(n)$ in the following way:

$$\mathrm{SO}(n-1) \ni R \mapsto \begin{pmatrix} R & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{SO}(n).$$

Since $\mathrm{SO}(n-1)$ is compact, every representation of G induced on the orbit $\mathcal{O}$ is square integrable. The usual way to construct wavelets in $L^2(\mathbb{R}^n)$ is to choose m as the trivial representation. We suggest an alternative way in order to obtain vector-valued wavelets.

Let m be a representation of $\mathrm{SO}(n-1)$. As a consequence of the Frobenius reciprocity theorem and of the fact that $\mathrm{SO}(n)$ is compact, there exists an (in general nonunique) represen-

tation D of SO(n) such that m is a subrepresentation of D, when D is restricted to SO($n-1$). This allows us to use a *covariant realization* of $\mathrm{Ind}_{G_0}^{G}(p_0 m)$, instead of the formula (1). It is defined in the following way. Let D act in W and m act in the subspace $\mathscr{K}$ of W. For all $p \in \mathscr{O}$ let $\mathscr{K}_p = D(R_p)\mathscr{K}$ where $R_p \in \mathrm{SO}(n)$ is such that $R_p p_0 = p/|p|$. Define the Hilbert space $\mathscr{H}$

$$\mathscr{H} = \{\phi \in L^2(\mathbb{P}^n, W) : \phi(p) \in \mathscr{K}_p \quad \text{for almost all } p\}$$

and the representation V acting in $\mathscr{H}$ as

$$(V_{xaR}\phi)(p) = a^{n/2} e^{ip\cdot x} D(R)\phi(aR^{-1}p).$$

It is possible to show[12] that V is equivalent to $\mathrm{Ind}_{G_0}^{G}(p_0 m)$. Moreover, a vector $\psi \in \mathscr{H}$ is admissible if and only if

$$\int |p|^{-n} \|\psi(p)\|_W^2 \, d^n p < \infty.$$

The Fourier–Plancherel transform intertwines V with a *geometric* representation $\hat{V}$ acting on a closed subspace $\hat{\mathscr{H}}$ of $L^2(\mathbb{R}^n, W)$ as

$$(\hat{V}_{xaR} f\,)(y) = a^{-n/2} D(R) f\left(\frac{R^{-1}(y-x)}{a}\right).$$

To give an explicit description of the subspace $\hat{\mathscr{H}}$ let us choose, for example, $n=3$ and D the representation of SO(3),

$$D(R) = R,$$

acting in $W = \mathbb{C}^3$ in a natural way. Let $\mathscr{K} = \{(0,0,c) : c \in \mathbb{C}\}$, hence the restriction m of D to SO(2) leaves invariant $\mathscr{K}$ (m is the trivial representation). An easy calculation shows that

$$\hat{\mathscr{H}} = \{f \in L^2(\mathbb{R}^3, \mathbb{C}^3) : (\nabla \wedge f\,) = 0\},$$

where $\nabla\wedge$ is the curl operator (in the sense of distributions). Moreover, we observe that the orthogonal complement of $\hat{\mathscr{H}}$,

$$(\hat{\mathscr{H}})^{\perp} = \{f \in L^2(\mathbb{R}^3, \mathbb{C}^3) : \nabla \cdot f = 0\},$$

where $\nabla\cdot$ is the divergence operator (in the sense of distributions), is the direct sum of two closed subspaces where $\hat{V}$ acts irreducibly. These subspaces correspond to the choice of the subspaces of W,

$$\mathscr{K}_1 = \{(c, ic, 0) : c \in \mathbb{C}\}, \quad \mathscr{K}_2 = \{(c, -ic, 0) : c \in \mathbb{C}\},$$

and of the representations of SO(2),

$$m_1(\theta) = e^{i\theta}, \quad m_2(\theta) = e^{-i\theta},$$

where $\theta \in [0, 2\pi]$ is the usual parameter of SO(2).

IV. CONSTRUCTION OF DISCRETE FRAMES

In this section we consider the problem of the existence of discrete frames associated to the induced representations of $G = \mathbb{R}^n \times' H$, when $H = \mathbb{R}^+ \times S$ and S is a semisimple connected Lie group.

The action of H on $\mathbb{R}^n$ is

$$h[x]=(as)[x] := a(s[x]),$$

where $x\mapsto s[x]$ is a smooth action of S on $\mathbb{R}^n$. Let $\mathbb{P}^n$ denote the dual group of $\mathbb{R}^n$ and fix $p_0 \in \mathbb{P}^n$. Assume that the orbit $\mathcal{O}=G[p_0]$ for the dual action of G on $\mathbb{P}^n$ is *open* and that the subgroup $H_0=H\cap G_0$ is compact (notice that $\mathbb{R}^+\cap H_0=\varnothing$). In these hypotheses $\mathcal{O}$ is homeomorphic to G/G_0 and to H/H_0.

We will prove the existence of frames associated to the representation $\mathrm{Ind}_{G_0}^{G}(p_0 m)$ where m is any unitary finite dimensional *cyclic* representation of H_0.

We need to establish some preliminary topological results.

Lemma 2: Let Γ be a discrete subgroup of H. Then, for any compact subset C in $\mathcal{O}$, there exists a finite partition $\Gamma_1,\ldots,\Gamma_d$ of Γ such that

$$h',h''\in\Gamma_j \Rightarrow h'[C]\cap h''[C]=\varnothing.$$

Proof: Let $\pi:H\to\mathcal{O}$ be the canonical projection. First of all, we prove that $\pi^{-1}(C)$ is compact. By Theorem 5.11 of Ref. 12, there exists a section $q:\mathcal{O}\to H$ such that the set $q(C)$ is relatively compact in H. Since H_0 is compact also, $\pi^{-1}(C)=q(C)H_0$ is relatively compact. Since $\pi^{-1}(C)$ is closed, it is compact.

By lemma 3.3 of Ref. 5, there exists a finite partition $\Gamma_1,\ldots,\Gamma_d$ of Γ such that

$$h',h''\in\Gamma_j \Rightarrow h'\pi^{-1}(C)\cap h''\pi^{-1}(C)=\varnothing.$$

Hence

$$\pi^{-1}(h'[C]\cap h''[C])=\varnothing$$

and $h'[C]\cap h''[C]=\varnothing$, since π is surjective. □

Lemma 3: There exist a discrete subgroup Γ of H and a compact C in $\mathcal{O}$ such that

$$\mathcal{O}=\cup\{h[C]:h\in\Gamma\}.$$

Proof: Since H is the direct product of $\mathbb{R}^+$ and a semisimple connected Lie group, as a consequence of Theorem C of Ref. 13 H has a discrete subgroup Γ such that H/Γ is compact. Denote by π' the canonical projection of H onto H/Γ. There exists a compact $K'\subset H$ such that $\pi'(K')=H/\Gamma$, hence

$$H=\cup\{hK:h\in\Gamma\},$$

where $K=(K')^{-1}$ is compact in H. Define $C=\pi(K)$. Then C is compact in $\mathcal{O}$ and

$$\cup\{h[C]:h\in\Gamma\}=\cup\{\pi(hK):h\in\Gamma\}=\mathcal{O}.$$

□

Lemma 4: For any nonvoid open set $X\subset\mathcal{O}$ there exists a finite set $\{h_1,\ldots,h_L\}$ in H, such that

$$\cup\{hh_k[X]:h\in\Gamma,\ \ 1\leqslant k\leqslant L\}=\mathcal{O},$$

where Γ is the discrete subgroup of H given by the previous lemma.

Proof: The family of sets $\{h[X]:h\in H\}$ is an open covering of the compact subset C of $\mathcal{O}$ singled out by the previous lemma. Hence there exists a finite subcovering $\{h_k[X]:1\leqslant k\leqslant L\}$ and we have the thesis. □

Having established these results, let m be a unitary cyclic finite dimensional representation of the compact group H_0 acting in the Hilbert space $\mathcal{K}$ of dimension M. Let v be a cyclic vector for m. Then there exist M elements $s_1,\ldots,s_M$ of H_0 such that $\{m(s_i)v:i=1,\ldots,s_M\}$ is an algebraic basis of $\mathcal{K}$. Fix these elements. Let $c:\mathcal{O}\to H$ be a section for the action of H on $\mathcal{O}$. Since H is a Lie group acting smoothly on $\mathcal{O}$ we can assume that c is locally continuous around p_0. For all $h\in H$ and $p\in\mathcal{O}$ define $m(h,p) := m(c(p)^{-1}hc(h^{-1}[p]))$. Then we have the following.

Lemma 5: There exist a closed ball B with center p_0, contained in $\mathcal{O}$, and two positive numbers α, β such that

$$\alpha\|w\|^2_{\mathcal{K}} \leqslant \sum_{i=1}^{M} |\langle m(s_i,p)v,w\rangle_{\mathcal{K}}|^2 \leqslant \beta\|w\|^2_{\mathcal{K}}$$

for all $w \in \mathcal{K}$ and $p \in B$.

Proof: Let $\{e_i\}_{i=1}^{M}$ be an orthonormal basis of $\mathcal{K}$ and, for all $p \in \mathcal{O}$, $A(p)$ be the operator

$$A(p)w = \sum_{i=1}^{M} \langle m(s_i,p)v,w\rangle_{\mathcal{K}}\, e_i, \quad w \in \mathcal{K}.$$

Since the vectors $m(s_i)v = m(s_i,p_0)v$ form a basis of $\mathcal{K}$, $A(p_0)$ is invertible and

$$\|A(p_0)^{-1}\|^{-2}\|w\|^2_{\mathcal{K}} \leqslant \sum_{i=1}^{M} |\langle m(s_i,p)v,w\rangle_{\mathcal{K}}|^2 \leqslant \|A(p_0)\|^2\|w\|^2_{\mathcal{K}}.$$

By the properties of c there exists a neighborhood of p_0 where the function $p \mapsto A(p)$ is continuous. The thesis easily follows. □

We consider the explicit realization of $\mathrm{Ind}_{G_0}^{G}(p_0 m)$ acting in $L^2(\mathcal{O},\mathcal{K})$ as

$$(U_{xh}f\,)(p) = \rho(h)^{-1/2} e^{ix\cdot p} m(h,p) f(h^{-1}[p]).$$

We notice that, since S is semisimple and connected, ρ has the following expression

$$\rho(as) = a^n, \quad a \in \mathbb{R}^+, \quad s \in S.$$

Fix the closed ball B with center p_0 as in Lemma 5 and the discrete subgroup Γ of H as in Lemma 3. Let Y be an open subset of $\mathcal{O}$ such that $p_0 \in Y$ and $s_i[Y] \subset B$ for all $i = 1,...,M$.

The set $X := \cap\{s_i[Y]: 1 \leqslant i \leqslant M\}$ is not empty (since it contains p_0) and open, hence, by Lemma 4, there exists a finite family $\{h_1,\ldots,h_L\} \subset H$ such that

$$\cup\{hh_k[X]: h \in \Gamma, \quad 1 \leqslant k \leqslant L\} = \mathcal{O}.$$

Since the set $\cup\{h_k[B]: 1 \leqslant k \leqslant L\}$ is compact, due to Lemma 2, there exists a finite partition $\Gamma_1,\ldots,\Gamma_d$ of Γ such that

$$h',h'' \in \Gamma_j \Rightarrow h'\left[\bigcup_{k=1}^{L} h_k[B]\right] \cap h''\left[\bigcup_{k=1}^{L} h_k[B]\right] = \varnothing.$$

Let ν be an integer such that $B \subset R$ where

$$R := \{x \in \mathbb{R}^n: -\pi\nu \leqslant x_i \leqslant \pi\nu \;\; 1 \leqslant i \leqslant n\}$$

and, for all $l \in \mathbb{Z}^n$, $x_l = 2\pi l \in \mathbb{R}^n$.

We observe that the countable family of functions

$$\{e_l : l \in \mathbb{Z}^n\}$$

defined as

$$e_l(p) = (2\pi\nu)^{-n/2} e^{ix_l\cdot p}, \quad p \in R,$$

is a Hilbert basis of $L^2(R)$.

Now we are in a position to describe a covariant frame in $L^2(\mathcal{O},\mathcal{K})$. Let $\psi \in L^2(\mathcal{O},\mathcal{K})$ be the vector

$$\psi(p) := \chi_Y(p)v, \quad p \in \mathcal{O},$$

and consider the countable set of vectors

$$\psi_{h,k,i,l} := U_h U_{h_k} U_{s_i} U_{x_l} \psi,$$

where $h \in \Gamma$, $1 \leqslant k \leqslant L$, $1 \leqslant i \leqslant M$ and $l \in \mathbb{Z}^n$ (Γ is countable, since it is discrete). With these notations we have the following result.

Theorem 3: *The set* $\{\psi_{h,k,i,l}\}$ *is a frame in* $L^2(\mathscr{O},\mathscr{K})$.

Proof: Given $f \in L^2(\mathscr{O},\mathscr{K})$, we have

$$\begin{aligned}\langle \psi_{h,k,i,l}, f\rangle_{L^2(\mathscr{O},\mathscr{K})} &= \langle U_{x_l}\psi, U_{(hh_k s_i)^{-1}} f\rangle_{L^2(\mathscr{O},\mathscr{K})} = \int_{\mathscr{O}} e^{-ix_l\cdot p}\langle \psi(p), (U_{(hh_k s_i)^{-1}} f\,)(p)\rangle_{\mathscr{K}}\, d^n p \\ &= (2\pi\nu)^{n/2}\int_R \overline{e_l(p)}\langle \psi(p), (U_{(hh_k s_i)^{-1}} f\,) \\ &\quad \times (p)\rangle_{\mathscr{K}}\, d^n p.\end{aligned}$$

Since ψ is bounded, the function

$$\mathscr{O} \ni p \mapsto \langle \psi(p), (U_{(hh_k s_i)^{-1}} f\,)(p)\rangle_{\mathscr{K}}$$

is in $L^2(R)$, hence

$$\begin{aligned}\sum_{hkil} |\langle \psi_{h,k,i,l}, f\rangle_{L^2(\mathscr{O},\mathscr{K})}|^2 &= (2\pi\nu)^n \sum_{hki} \int_{\mathscr{O}} |\langle \psi(p), (U_{(hh_k s_i)^{-1}} f\,)(p)\rangle_{\mathscr{K}}|^2\, d^n p \\ &= (2\pi\nu)^n \sum_{hki} \int_{\mathscr{O}} |\langle m(s_i, s_i[p])\psi(p), (U_{(hh_k)^{-1}} f\,)(s_i[p])\rangle_{\mathscr{K}}|^2\, d^n p\end{aligned}$$

(changing variable $p \mapsto s_i^{-1}[p]$)

$$\begin{aligned}&= (2\pi\nu)^n \sum_{hki} \int_{\mathscr{O}} |\langle m(s_i, p)\psi(s_i^{-1}[p]), (U_{(hh_k)^{-1}} f\,)(p)\rangle_{\mathscr{K}}|^2\, d^n p \\ &= (2\pi\nu)^n \sum_{hki} \int_{s_i[Y]} |\langle m(s_i, p)v, (U_{(hh_k)^{-1}} f\,)(p)\rangle_{\mathscr{K}}|^2\, d^n p.\end{aligned}$$

Recalling that $X = \cap\{s_i[Y] : 1 \leqslant i \leqslant M\}$, for any positive function ϕ we have

$$\int_X \phi\, d^n p \leqslant \int_{s_i[Y]} \phi\, d^n p \leqslant \int_B \phi\, d^n p$$

for all $1 \leqslant i \leqslant M$. Hence, since the sum on the index i is finite,

$$\begin{aligned}&(2\pi\nu)^n \sum_{hki} \int_{s_i[Y]} |\langle m(s_i, p)v, (U_{(hh_k)^{-1}} f\,)(p)\rangle_{\mathscr{K}}|^2 d^n p \\ &\geqslant (2\pi\nu)^n \sum_{hk} \int_X \sum_i |\langle m(s_i, p)v, (U_{(hh_k)^{-1}} f\,)(p)\rangle_{\mathscr{K}}|^2\, d^n p \\ &\geqslant \alpha(2\pi\nu)^n \sum_{hk} \int_X \|(U_{(hh_k)^{-1}} f\,)(p)\|^2_{\mathscr{K}}\, d^n p \\ &= \alpha(2\pi\nu)^n \sum_{hk} \int_X \rho(hh_k)^{-1} \|f(hh_k[p])\|^2_{\mathscr{K}}\, d^n p\end{aligned}$$

(changing variable $p \mapsto (hh_k)^{-1}[p]$)

$$= \alpha(2\pi\nu)^n \sum_{hk} \int_{hh_k[X]} \|f(p)\|^2_{\mathscr{K}}\, d^n p \geqslant \alpha(2\pi\nu)^n \|f\|^2_{L^2(\mathscr{O},\mathscr{K})},$$

taking into account that

$$\cup\{hh_k[X]: h\in\Gamma,\ \ 1\leqslant k\leqslant L\} = \mathscr{O}.$$

On the other hand

$$(2\pi\nu)^n \sum_{hki} \int_{s_i[Y]} |\langle m(s_i,p)v, (U_{(hh_k)^{-1}} f\ )(p)\rangle_{\mathscr{K}}|^2\, d^n p$$

$$\leqslant (2\pi\nu)^n \sum_{hk} \int_B \sum_i\ |\langle m(s_i,p)v, (U_{(hh_k)^{-1}} f\ )(p)\rangle_{\mathscr{K}}|^2\, d^n p$$

(reasoning in the same way as before)

$$\leqslant \beta(2\pi\nu)^n \sum_{hk} \int_{hh_k[B]} \|f(p)\|^2_{\mathscr{K}}\, d^n p$$

[by property (2)]

$$= \beta(2\pi\nu)^n \sum_{j=1}^{d} \sum_{h\in\Gamma_j} \sum_{k=1}^{L} \int_{hh_k[B]} \|f(p)\|^2_{\mathscr{K}}\, d^n p$$

$$\leqslant d\beta(2\pi\nu)^n \int_{\mathscr{O}} \|f(p)\|^2_{\mathscr{K}}\, d^n p = d\beta(2\pi\nu)^n \|f\|^2_{L^2(\mathscr{O},\mathscr{K})}.$$

Summarizing we have

$$\alpha(2\pi\nu)^n \|f\|^2_{L^2(\mathscr{O},\mathscr{K})} \leqslant \sum_{hkil} |\langle \psi_{h,k,i,l}, f\rangle_{L^2(\mathscr{O},\mathscr{K})}|^2 \leqslant d\beta(2\pi\nu)^n \|f\|^2_{L^2(\mathscr{O},\mathscr{K})}.$$

□

[1] A. Grossmann and J. Morlet, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **15**, 723 (1984).
[2] See, for example, I. Daubechies, *Ten Lectures on Wavelets* (SIAM 1992) and references therein.
[3] C. E. Heil and D. F. Walnut, SIAM (Soc. Ind. Appl. Math.) Rev. **31**, 628 (1989).
[4] D. Bernier and K. F. Taylor, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **27**, 594 (1996).
[5] H. Feichtinger and K. Grochenig, J. Funct. Anal. **86**, 307 (1989).
[6] H. Führ, J. Math. Phys. **37**, 6353 (1996).
[7] R. Murenzi, in *Wavelets: Time-Frequency Methods and Phase Space (Proc. Marseille 1987)*, edited by J. M. Combes, A. Grossmann, and P. Tchamitchian (Springer-Verlag, New York, 1989).
[8] M. Duflo and C. C. Moore, J. Funct. Anal. **21**, 209 (1976).
[9] G. W. Mackey, Ann. Math. **55**, 101 (1952).
[10] A. Kleppner and R. L. Lipsman, Ann. Sci. École Norm. Sup. **5**, 459 (1972).
[11] P. Aniello, G. Cassinelli, E. De Vito, and A. Levrero, to appear in Rev. Math. Phys. **10** (1998).
[12] V. S. Varadarajan, *Geometry of Quantum Theory*, 2nd ed. (Springer-Verlag, New York, 1985).
[13] A. Borel, Topology **2**, 111 (1963).

Continuous wavelet transforms with Abelian dilation groups

Hartmut Führ[a)]
Zentrum Mathematik der Technischen Universität München, 80290 Munich, Germany

(Received 15 August 1997; accepted for publication 27 March 1998)

This paper is devoted to the construction of wavelet (or coherent state) systems arising from the action of certain semidirect products $G=\mathbb{R}^k \rtimes H$ on $L^2(\mathbb{R}^k)$. For this purpose previous results which guarantee the existence of inversion formulas are applied to the special case where H is Abelian. The questions of systematic construction and conjugacy of such groups are completely resolved by setting up a correspondence to unit groups of commutative associative algebras. As an application the numbers of conjugacy classes of possible Abelian groups are computed for $k=2,3,4$. For $k \geqslant 7$, there are uncountably many conjugacy classes. We then compute the admissibility conditions belonging to Abelian groups. The final section contains a characterization of Abelian matrix groups acting ergodically on some subset of $\mathbb{R}^k$. This result ensures that the approach via associative algebras yields all possible groups. © *1998 American Institute of Physics.*
[S0022-2488(98)00608-2]

I. INTRODUCTION

This paper is concerned with the construction of coherent state (or wavelet) systems arising from certain square integrable representations on $L^2(\mathbb{R}^k)$. Recall (e.g., from Ref. 1) that generally a system of coherent states on a Hilbert space $\mathcal{H}$ is given by a triple $(X,(\eta_x)_{x\in X},\mu)$, where X is a topological space, the coherent states $(\eta_x)_{x\in X}$ are a family of vectors in $\mathcal{H}$ (satisfying some continuity condition) and μ is a Borel measure on X yielding a **resolution of the identity**,

$$\mathrm{Id}_{\mathcal{H}}=\int_X |\eta_x\rangle\langle\eta_x| d\mu(x).$$

Here the integral is taken in the weak operator sense and $|\eta_x\rangle\langle\eta_x|$ denotes the rank-one-operator $\phi \mapsto \langle\phi,\eta_x\rangle\eta_x$. The topological assumptions are sometimes weakened to measure theoretic ones. The concept of coherent states arose in quantum optics and has by now spread to a vast range of (quantum) physical applications. The interested reader is referred to the collection Ref. 1, the review paper Ref. 2 and the extensive bibliographies therein.

In this paper we consider the special case where $\mathcal{H}$ is $L^2(\mathbb{R}^k)$ (or a closed subspace), the topological space $X=G=\mathbb{R}^k \rtimes H$ is a group of affine mappings acting on $\mathcal{H}$ via its canonical representation π, the coherent states are given by $\eta_x := \pi(x)\eta$ (η some suitably chosen *fiducial* or *admissible vector*) and finally μ is left Haar measure on G. Coherent states thus obtained from unitary actions have been frequently studied in mathematics and mathematical physics, especially for semidirect products of the type considered here (a small sample of the literature is given in Refs. 3–8). Readers familiar with wavelet theory will recognize the mapping $\phi \mapsto (\langle\phi,\pi(x)\eta\rangle)_{x\in G}$ as a continuous wavelet transform and the resolution of the identity as a reformulation of the wavelet inversion formula. In fact, the well known one-dimensional continuous wavelet transform is a special case of this construction.

The group-theoretic approach leads to the concept of **square integrable group representations**. In this paper we are only concerned with square integrability in the strict sense of Ref. 9: a representation is square integrable iff it is irreducible and has a (nontrivial) square integrable matrix coefficient. This property immediately yields a wavelet inversion formula of the desired

a)Electronic mail: fuehr@mathematik.tu-muenchen.de

 © 1998 American Institute of Physics

kind; for the one-dimensional wavelet transform this was first observed in Ref. 5. Both requirements have been weakened (irreducibility to cyclicity,[6,7] square integrability to square integrability on quotients[10,11,8]), at the cost of technical difficulties. The strong form of square integrability is better understood, especially for the type of affine groups considered in this paper.[12,13]

Thus, the construction depends on the choice of a matrix group $H<\mathrm{GL}(k,\mathbb{R})$, the so called *dilation group*. The canonical action of the group $\mathbb{R}^k \rtimes H$ on $\mathbb{R}^k$ gives rise to a strongly continuous unitary representation π on $\mathrm{L}^2(\mathbb{R}^k)$, and we call H **admissible** if π has a square integrable subrepresentation.

This paper deals with the case of Abelian dilation groups. We are going to give a full description of the admissible Abelian dilation groups in terms of commutative associative algebras, including the question of conjugacy of the dilation groups. The latter was already raised in Ref. 12, Remark (5), for the special class of groups considered there. It is of interest because replacing one dilation group by a conjugate group results in a simple change of coordinates in the corresponding wavelet transform; we may thus restrict our attention to conjugacy classes. The main tool for the classification may be sketched as follows: We generalize the construction from an algebraic point of view, substituting the affine group $\mathbb{R}\rtimes\mathbb{R}^\times$ underlying the one-dimensional transform by a (somewhat modified) affine group of a k-dimensional associative algebra with unity, $\mathcal{A}\rtimes_{\mathrm{T}}\mathcal{A}^\times$, where $\mathcal{A}$ is viewed as a vector group, $\mathcal{A}^\times$ denotes the group of invertible elements of $\mathcal{A}$ and $\rtimes_{\mathrm{T}}$ denotes a (group-theoretic) semidirect product, in which not the usual action of $\mathcal{A}^\times$ on $\mathcal{A}$ is used but a transposed action. Under the regular representation of the algebra, the action of $\mathcal{A}^\times$ on $\mathcal{A}$ becomes the action of a linear group on $\mathbb{R}^k$. The resulting linear group can then be shown to be admissible. The construction works for general associative algebras with unity. It is the main result of this paper that every admissible **Abelian** group is obtained in this fashion from a commutative associative algebra with unity, and that the question of conjugacy of admissible groups translates on the algebraic level to a question concerning algebra isomorphisms.

It should be noted, however, that while the algebraic viewpoint is extremely useful when it comes to the enumeration and classification of wavelet transforms with Abelian dilation groups, it is not clear how the algebraic structure relates to properties of the wavelet transform (if it does at all). The latter are connected to geometric properties of the action of the dilation group (the ''zooming''-effect, for instance). So far comparatively few groups have been considered, and the relationship between group structure and properties of the wavelet transform has rarely been explicitly investigated. In my opinion this question deserves further attention.

The paper is structured as follows: Section II contains a more detailed review of the group-theoretic approach to the construction of continuous wavelet transforms on $\mathrm{L}^2(\mathbb{R}^k)$. In particular, the admissibility criterion for groups established previously[13] is reviewed, which leads to considering linear groups with open orbits. In addition to these known results, we obtain some further properties of admissible groups.

In Sec. III we deal with Abelian matrix groups with open orbits. We establish a connection with certain maximal Abelian matrix algebras. In particular it is shown that the associative algebra generated by such a group coincides with its Lie algebra, and that the group is an open subgroup of the algebra's unit group. It is thus possible to obtain also the nonconnected Abelian dilation groups from these algebras (Theorem 11).

We then proceed to characterize the algebras in terms of abstract associative algebras (Theorem 13).

In Sec. IV we use the results of Sec. III to show how one can construct a full catalog of (conjugacy classes of) admissible Abelian groups from a (modulo isomorphism) complete system of commutative associative algebras (see Corollary 14). As an application, we compute the numbers of conjugacy classes of Abelian admissible groups in dimensions 2 through 4, and show how the structure theory of associative algebras can be utilized for this task. From dimension 7 on, there are uncountably many conjugacy classes of Abelian admissible groups (Corollary 15). We then turn to the sets of wavelets (or admissible functions) associated to each dilation group. It is shown that only a superficial knowledge of the algebra structure is needed to explicitly compute the admissibility condition.

Section V finally contains a characterization of Abelian matrix groups acting ergodically on some subset $U\subset\mathbb{R}^k$ of positive measure (Theorem 25). This result, together with Proposition 5, ensures that the list of admissible groups obtained via associative algebras is complete.

II. SQUARE INTEGRABLE REPRESENTATIONS OF AFFINE GROUPS

This section contains a review of results on the construction of wavelet transforms on $L^2(\mathbb{R}^k)$ from quasiregular representations of certain groups of affine mappings established previously,[12,13] and some additional results.

As mentioned in the Introduction, the construction depends on the choice of a dilation group H. In Refs. 12, 13, H was assumed to be closed, without any comments concerning the necessity of this assumption, but Proposition 5 below shows that it is, in fact, necessary for the success of the construction. The subgroup of the full affine group of $\mathbb{R}^k$ generated by H and the translations is the semidirect product $G:=\mathbb{R}^k \rtimes H$. Elements of G are denoted by (x,h), with $x \in \mathbb{R}^k$ and $h \in H$, and the group law is given by $(x_1,h_1)(x_2,h_2)=(x_1+h_1x_2,h_1h_2)$. From the canonic action of G on $\mathbb{R}^k$ arises a very natural unitary representation, the *quasiregular representation* π defined by $(\pi(x,h)f)(y):=|\det(h)|^{-1/2}f(h^{-1}(y-x))$, where f denotes an element of $L^2(\mathbb{R}^k)$.

It turns out that the question of admissibility of a matrix group H can be answered by looking at the action of H on the dual group $\widehat{\mathbb{R}^k}$. If we view elements of $\mathbb{R}^k$ as column vectors and elements of $\widehat{\mathbb{R}^k}$ as row vectors, the dual action of H becomes matrix multiplication on the right. There is a one-to-one correspondence between irreducible subrepresentations and subsets $U \subset \widehat{\mathbb{R}^k}$ of positive measure on which H acts ergodically (Ref. 13, Corollary 4). We denote the representation belonging to U by π_U. Now two cases can occur: The action of H on U may be *properly ergodic*, meaning that all H-orbits in U are null sets. To my knowledge, there exists no general criterion for square integrability of π_U in this situation (for partial results see Ref. 13, Remark 12). However, it will turn out that if H is closed and Abelian, then H cannot act ergodically on a subset of $\widehat{\mathbb{R}^k}$ unless this subset is already an open orbit; see Theorem 25 below. Hence, this case never occurs in the context of Abelian dilation groups.

Let us now turn to the second case; without loss of generality we may thus assume U to be an orbit of positive measure. By Ref. 14, Lemma 4, U is, in fact, open. We then have the following characterization of square integrability (see Ref. 13, Theorem 10).

Theorem 1: *π_U is square integrable iff for one (and hence every) $\gamma \in U$ the stabilizer H_γ in H is compact.*

We are now going to give some results on open orbits of matrix groups and their consequences for admissible dilation groups. A simple fact that will be frequently used is that if H is a matrix group such that Hx is open for some $x \in \mathbb{R}^k$, then so is $H'x$ for every matrix group H' containing H. At first we are going to give a simple but very useful characterization for matrix groups with open orbits involving their Lie algebras. We will view these Lie algebras canonically as matrix Lie algebras. In order to simplify the following discussion, we consider the action on column vectors even though we are, in fact, interested in the action on row vectors.

The above cited Lemma 4 of Ref. 14 was obtained by applying Sard's Theorem to the quotient mapping $p_x: H \ni h \mapsto hx$, for $x \in U$. As a further consequence of this we obtain the following characterization:

Lemma 2: Let $H < \mathrm{GL}(k,\mathbb{R})$ be closed, let $\mathfrak{h}$ be its Lie algebra. Then, for $x \in \mathbb{R}^k$ the orbit Hx is open iff the differential of the map $H \ni h \mapsto hx$ has full rank at the neutral element; or more explicitly, iff the condition

$$(\mathcal{O}) \quad \textit{The linear mapping } \mathfrak{h} \ni X \mapsto Xx \in \mathbb{R}^k \textit{ is onto,}$$

is fulfilled.

As a first application we note the following.

Theorem 3: *Let $H < \mathrm{GL}(k,\mathbb{R})$ be closed. Suppose there exists $x \in \mathbb{R}^k$ with Hx open. Then, up to a null set, $\mathbb{R}^k$ is the union of finitely many open H-orbits.*

Proof: The argument uses terminology and results from real algebraic geometry. Let O denote the union of open H-orbits. By condition $(\mathcal{O})$ of Lemma 2, $x \in O$ iff the differential of the quotient map belonging to x has maximal rank, which is the case iff some $k \times k$-subdeterminant of the differential does not vanish. Determinants being polynomial, it follows that O is a nonempty Zariski-open (particularly conull) subset of $\mathbb{R}^k$. By a standard trick, O is homeomorphic to a

Zariski-closed subset in $k+1$ dimensions, hence it has finitely many connected components, as a result of Whitney.[15] But each component is contained in an open orbit. □

Remarks 4: (1) From the requirement that H operate on an open orbit with compact fixed groups one can easily deduce that $k \leqslant \dim H \leqslant 1+k(k-1)/2$. Both bounds are sharp, as examples 15 and 16 of Ref. 13 show.

(2) Since H and the connected component H_0 of the neutral element in H have the same Lie algebra, it follows from Lemma 2 that Hx is open iff H_0x is open. It can be shown that H operates on Hx with compact stabilizers iff H/H_0 is finite and H_0 operates on H_0x with compact stabilizers. Here the sufficiency is clear; the proof for the necessity uses the finiteness of the open H_0-orbits. □

We end this section with the proof that only closed matrix groups are useful for the construction of wavelet transforms. Note that the proposition holds, in particular, for Lie group topologies.

Proposition 5: Let $H<\mathrm{GL}(k,\mathbb{R})$, endowed with a locally compact group topology $\mathcal{T}$. Suppose that the semidirect product $\mathbb{R}^k \rtimes H$ is a topological semidirect product, i.e., $\mathbb{R}^k \rtimes H$ endowed with the product topology ($\mathbb{R}^k$ topologized as usual) is a topological group. If the quasiregular representation of $\mathbb{R}^k \rtimes H$ on $\mathrm{L}^2(\mathbb{R}^k)$ has a square integrable subrepresentation, then H is closed and $\mathcal{T}$ is the relative topology.

Proof: In order that the semidirect product be topological, $\mathcal{T}$ must be finer than the relative topology. Now suppose that π has a square integrable subrepresentation, which is then equivalent to a subrepresentation of the regular representation of $\mathbb{R}^k \rtimes H$, by Ref. 9, Theorem 2. Since the matrix coefficients of the regular representation are C_0-functions,[16] π has a nontrivial C_0 matrix coefficient $V_{f,g}$. We may assume $V_{f,g}(1)=1$. Let $W_{f,g}$ denote the matrix coefficient of the quasiregular representation of $\mathbb{R}^k \rtimes \mathrm{GL}(k,\mathbb{R})$ corresponding to the same pair of functions f,g. Then $V_{f,g}$ is the restriction of $W_{f,g}$ to $\mathbb{R}^k \rtimes H$. Since $W_{f,g}$ is continuous, there exists a compact neighborhood U of 1 in $\mathrm{GL}(k,\mathbb{R})$ with $|W_{f,g}|>1/2$ on U. $U \cap H$ is $\mathcal{T}$-closed in H, since U is closed and $\mathcal{T}$ is finer than the relative topology. Since H is closed in $\mathbb{R}^k \rtimes H$, the restriction of $V_{f,g}$ to H vanishes at infinity. By choice of U, this implies that $U \cap H$ is contained in a $\mathcal{T}$-compact set, hence is $\mathcal{T}$-compact itself. But then $\mathcal{T}$ coincides on $U \cap H$ with the relative topology, and since $U \cap H$ is a neighborhood of 1 for both topologies, it follows that they coincide on H. In particular, the relative topology is locally compact, which means H is closed. □

III. ABELIAN MATRIX GROUPS WITH OPEN ORBITS

In this section we start with the investigation of Abelian matrix groups with open orbits. For this purpose the Lie algebras of these groups will play a crucial role. This may seem surprising, since Abelian Lie algebras have no particular structure. But the algebras we are interested in will turn out to be associative algebras as well and this associative structure is decisive.

The first result of this section is not particularly deep, but will prove to be extremely useful.

Lemma 6: Let $H<\mathrm{GL}(k,\mathbb{R})$ be Abelian and $x \in \mathbb{R}^k$. If Hx is open, then the stabilizer H_x of x is trivial.

Proof: Since the stabilizers H_y of $y \in Hx$ are conjugate in H to H_x and H is Abelian, we have $H_x=H_y$. Being open, Hx spans $\mathbb{R}^k$, whence we obtain that H_x fixes every $y \in \mathbb{R}^k$. But then H_x is trivial. □

The first consequence is that the compactness condition of Theorem 1 is trivially fulfilled when H is Abelian. As a second application we show the following.

Proposition 7: Let *$H<\mathrm{GL}(k,\mathbb{R})$ be Abelian. Suppose there exists $x \in \mathbb{R}^k$ with Hx open. Then H is a closed, k-dimensional subgroup.*

Proof: If Hx is open, then so is $\overline{Hx}$. Since the operation of $\overline{H}$ on the open orbit is free, the quotient mapping $\overline{H} \ni h \mapsto hx \in \overline{Hx}$ is a homeomorphism (in fact, a diffeomorphism), and the image of H under this map is open. Hence, H is open in $\overline{H}$, which entails $H=\overline{H}$. $\dim H=k$ is then clear. □

The next aim is to establish the connection between Abelian matrix groups having an open orbit and maximal Abelian matrix algebras. We start with some definitions.

Definition 8: An Abelian subgroup $H<\mathrm{GL}(k,\mathbb{R})$ is maximal Abelian if H is not properly contained in an Abelian $H'<\mathrm{GL}(k,\mathbb{R})$. It is called maximal connected Abelian if it is connected

and is not properly contained in a connected Abelian subgroup of $\mathrm{GL}(k,\mathbb{R})$. (*Note that being maximal connected Abelian is* not *the same as being maximal Abelian and connected.*)

By $\mathrm{gl}(k,\mathbb{R})$ *we denote the space of* $k\times k$*-matrices, either endowed with its structure as Lie algebra or as associative algebra. A Lie (respectively, associative) subalgebra of* $\mathrm{gl}(k,\mathbb{R})$ *is called maximal Abelian if it is Abelian and not properly contained in another Lie (respectively, associative) subalgebra.*

The following remarks collect the basic properties of maximal Abelian groups and algebras and their relation to Abelian groups with open orbit.

Remarks 9: (1) Clearly, maximal (connected) Abelian subgroups are closed. By Zorn's Lemma, every (connected) Abelian subgroup is contained in a maximal (connected) Abelian subgroup.

(2) Maximal connected Abelian subgroups are in one-to-one correspondence to maximal Abelian Lie-subalgebras of $\mathrm{gl}(k,\mathbb{R})$, via the exponential map. But the latter are in fact maximal Abelian *associative* subalgebras, since every Abelian Lie-subalgebra generates an Abelian associative subalgebra (and since every associative subalgebra is a Lie-subalgebra).

On the other hand, every maximal Abelian associative subalgebra is a maximal Abelian Lie-subalgebra as well.

(3) If H is an Abelian matrix groups having an open orbit, then its connected component H_0 is maximal connected Abelian: Every Abelian connected group H_1 containing H_0 has an open orbit [since H_0 has an open orbit, by (2) of Remark 4]; hence, by Proposition 7 $\dim H_0 = k = \dim H_1$, and thus $H_0 = H_1$.

(4) A trivial but useful observation is that every maximal Abelian subalgebra contains the identity. This fact has consequences for the wavelet transforms constructed from Abelian dilation groups: The group $\mathbb{R}^k \rtimes H$ is never unimodular, since H contains the positive multiples of the identity, and for such an element $a\cdot \mathrm{Id}\in H$ we have $\Delta_G(0,a\cdot \mathrm{Id}) = a^{-k}$. The nonunimodularity then implies that not every vector in $\mathrm{L}^2(\mathbb{R}^k)$ (or a subspace carrying an irreducible subrepresentation) is admissible (see Ref. 9, the remark after Theorem 3). □

Summarizing the results of our discussion so far, we see that we are interested in constructing k-dimensional maximal Abelian (associative or Lie-) subalgebras satisfying condition $(\mathcal{O})$ of Lemma 2. A complete description of these algebras immediately yields a complete description of at least the connected Abelian matrix groups with open orbit. But as we will soon see, it is also possible to obtain the nonconnected Abelian matrix groups with open orbits from these algebras, as open subgroups of their unit groups. Before we prove this result, we need some preliminary results on unit groups of associative subalgebras of $\mathrm{gl}(k,\mathbb{R})$. Generally speaking, for a given associative algebra $\mathcal{A}$ with unity element $1_{\mathcal{A}}$, we let $\mathcal{A}^\times$ denote the *unit group* or group of invertible elements of $\mathcal{A}$.

Proposition 10: Let $\mathcal{A}\subset \mathrm{gl}(k,\mathbb{R})$ *denote a* d*-dimensional associative subalgebra containing the identity. Then* $\mathcal{A}^\times$ *is almost connected and conull with respect to the standard measure on* $\mathcal{A}\cong\mathbb{R}^d$. *If* $H<\mathrm{GL}(k,\mathbb{R})$ *is maximal Abelian, then we have* $H=\mathcal{A}^\times$, *where* $\mathcal{A}=\mathrm{span}(H)$ *denotes the algebra generated by* H. *In particular* H *has only finitely many connected components.*

Proof: We start by observing that $\mathcal{A}^\times = \mathcal{A}\cap \mathrm{GL}(k,\mathbb{R})$: Indeed, if $A\in\mathcal{A}$ is an invertible matrix, then multiplication by A is one-to-one on $\mathcal{A}$, hence onto. This shows ''$\supset$'', and the other direction is trivial. Since det is a polynomial mapping, it follows that $\mathcal{A}^\times$ is a Zariski-open subset of $\mathcal{A}$, hence the same argument as in the proof of Theorem 3 yields that $\mathcal{A}^\times$ has finitely many connected components and is a conull subset. The statements concerning H are obvious. □

Theorem 11: *Let* $H<\mathrm{GL}(k,\mathbb{R})$ *be Abelian with open orbit in* $\mathbb{R}^k$. *Let* $\mathfrak{h}$ *denote its Lie algebra and* $\mathcal{A}:=\mathrm{span}(H)$, *the associative algebra generated by* H. *Then we have* $\mathfrak{h}=\mathcal{A}$, *and* H *is an open subgroup of* $\mathcal{A}^\times$ *with a finite index. There exists exactly one open (and hence conull)* $\mathcal{A}^\times$*-orbit in* $\mathbb{R}^k$.

Proof: We have $H<\mathcal{A}^\times$, hence both have an open orbit, thus both have dimension k by Proposition 7. Thus H is open in $\mathcal{A}^\times$, hence $\mathcal{A}^\times/H$ is finite, by Proposition 10. $\mathcal{A}^\times\subset\mathcal{A}$ is open, hence $\dim\mathcal{A}=k$. $\mathfrak{h}$ being an associative subalgebra containing the identity element, it contains its own image under the exponential mapping, i.e., $\mathfrak{h}\supset H_0$, where H_0 is the connected component of H. But $\mathrm{span}(H_0)$ is at least k-dimensional, hence $\mathfrak{h}=\mathrm{span}(H_0)=\mathcal{A}$.

The single open $\mathcal{A}^\times$-orbit is obtained as follows: Since $\mathfrak{h}=\mathcal{A}$ fulfills condition $(\mathcal{O})$, there exists $x\in\mathbb{R}^k$ such that the mapping $\mathcal{A}\ni X\mapsto Xx\in\mathbb{R}^k$ is a vector space isomorphism. Now

$\mathcal{A}^\times x$ is the image of the open conull subset $\mathcal{A}^\times$ under this isomorphism, so $\mathcal{A}^\times x$ is the desired orbit. □

The following proposition shows a simple method of constructing associative subalgebras satisfying condition $(\mathcal{O})$. It uses the regular representations of associative algebras with unity.

Proposition 12: Let $\mathcal{A}$ be a real k-dimensional associative (not necessarily Abelian) algebra with unity. The (left) regular representation ρ of $\mathcal{A}$ is defined by choosing a vector space isomorphism $\psi:\mathcal{A}\to\mathbb{R}^k$ and defining, for all $A,B\in\mathcal{A}$: $\rho(A)\psi(B):=\psi(AB)$. This defines an algebra-monomorphism of $\mathcal{A}$ into $\mathrm{gl}(k,\mathbb{R})$. *$\rho(\mathcal{A})$ satisfies condition $(\mathcal{O})$ and $\rho(\mathcal{A}^\times)$ operates freely on a single open orbit in $\mathbb{R}^k$. As a consequence, $\rho(\mathcal{A}^\times)^\mathsf{T}$ is an admissible group, in fact the quasiregular representation of $\mathbb{R}^k\rtimes\rho(\mathcal{A}^\times)^\mathsf{T}$ is square integrable.*

Proof: That ρ is an algebra monomorphism is verified easily. $\rho(\mathcal{A})$ satisfies condition $(\mathcal{O})$ because choosing $x:=\psi(1_\mathcal{A})$ we obtain $\rho(A)x=\psi(A)$, for all $A\in\mathcal{A}$. Hence the open orbit $\rho(\mathcal{A}^\times)\psi(1_\mathcal{A})$ is conull. The last statement of the theorem is then clear by Theorem 1. □

We have thus found a simple general procedure to construct admissible matrix groups. Note that nonisomorphic algebras yield nonconjugate groups, since an open subgroup of $\mathcal{A}^\times$ spans $\mathcal{A}$. Hence we see that the conjugacy problem for matrix groups operating freely on an open orbit contains as a subproblem the classification of real finite-dimensional associative algebras; to my knowledge the latter is far from being solved.

Returning to the Abelian case, the strategy from now on will be to show that every maximal Abelian subalgebra satisfying condition $(\mathcal{O})$ is conjugate to its own regular representation. This result answers both questions raised in the beginning of this section at once: Running through all regular representations of all (isomorphism classes of) k-dimensional commutative associative algebras, we obtain all candidates for $\mathfrak{h}$, and from the open subgroups of their unit groups all candidates for H. And: Two Abelian algebras satisfying condition $(\mathcal{O})$ are conjugate iff they are isomorphic as associative algebras.

One should note, however, that only in principle do these results give a full description of the conjugacy classes of Abelian dilation groups. As far as I know, the classification of Abelian algebras has only been achieved for low dimensions.

Now for the theorem. In its present form the proof has been pointed out to me by Professor Stein A. Strømme.

Theorem 13: *An Abelian associative subalgebra $\mathcal{A}\subset\mathrm{gl}(k,\mathbb{R})$ satisfies condition $(\mathcal{O})$ iff it is conjugate to its own regular representation.*

Proof: The ''if'' part is provided by the previous proposition. For the ''only-if'' part, let $x\in\mathbb{R}^k$ be such that the mapping ϕ: $\mathcal{A}\ni X\mapsto Xx$ is a vector space isomorphism. Let ψ: $\mathcal{A}\to\mathbb{R}^k$ denote the vector space isomorphism used to define ρ. Then the relation $\phi\psi^{-1}\rho(A)\psi\phi^{-1}=A$ is easily checked for each $A\in\mathcal{A}$, hence we are done. □

IV. APPLICATIONS AND EXAMPLES

In this section we give some applications of the characterization of admissible Abelian groups. The following corollary gives a summary of how a complete catalog of these groups is obtained from the isomorphism classes of real associative algebras of the same dimension. Given the previous results, the task left is mainly bookkeeping, except for one minor problem: If $H_1,H_2<\mathcal{A}^\times$ are two open subgroups, when are their images under the regular representation conjugate? The answer to this question is fairly obvious, but it somewhat complicates the notation.

Corollary 14: Let $\mathcal{R}_k$ be a system of representatives of the isomorphism classes of real commutative associative algebras with unity of dimension k. For $\mathcal{A}\in\mathcal{R}_k$ let

$$\mathcal{G}_\mathcal{A}\subset\{H<\mathcal{A}^\times : H\subset\mathcal{A}^\times \text{ open}\}$$

be a system of representatives with respect to the equivalence relation ''$\approx$,'' defined by $H_1\approx H_2:\Leftrightarrow$ There exists an algebra automorphism $\eta:\mathcal{A}\to\mathcal{A}$ such that $\eta(H_1)=H_2$.

Since $\mathcal{A}^\times$ has finitely many open subgroups, $\mathcal{G}_\mathcal{A}$ is a finite set. Let $\mathfrak{UD}(k)$ denote the set of conjugacy classes of Abelian groups with open orbits in $\mathbb{R}^k$. Then the following map is a bijection:

$$\dot{\bigcup}_{\mathcal{A}\in\mathcal{R}_k}\mathcal{G}_{\mathcal{A}}\to\mathfrak{UD}(k)$$

$$H\mapsto[\rho(H)].$$

Each algebra contributes exactly one conjugacy class of connected groups and one class of groups having a single open orbit. By transposing each $H\in[H]\in\mathfrak{UD}(k)$ *we obtain the representatives of all admissible Abelian dilation groups.*

Proof: Well-definedness as well as injectivity on each $\mathcal{G}_{\mathcal{A}}$ is proved by the equivalence

$$H_1\approx H_2\Leftrightarrow\rho(H_1)\ \text{ is conjugate to }\rho(H_2).$$

The proof for sufficiency is straightforward. For the necessity, recall that $\rho(H_1)$ and $\rho(H_2)$ both span $\rho(\mathcal{A})$, hence every matrix that conjugates $\rho(H_1)$ into $\rho(H_2)$ conjugates $\rho(\mathcal{A})$ into itself, yielding the desired algebra automorphism. By the same argument it is clear that nonisomorphic algebras yield nonconjugate groups with open orbits; hence, it remains to show that the map is onto. But this is immediate from Theorems 11 and 13 and the assumptions on $\mathcal{G}_{\mathcal{A}}$. The fact that all admissible Abelian groups are covered follows from Proposition 5 and Theorem 25. □

As a further corollary we note the following.

Corollary 15: For $k\geqslant 7$ *there are uncountably many nonconjugate Abelian subgroups of* $\mathrm{GL}(k,\mathbb{R})$ *with open orbits.*

Proof: By Theorem 16 from Ref. 17, Chapter 2, Section 2, there exists an uncountable family $(\mathcal{N}_\alpha)_{\alpha\in\Lambda}$ of nilpotent commutative associative algebras of dimension $k-1$, which are pairwise nonisomorphic. For nilpotent algebras $\mathcal{N}_1$ and $\mathcal{N}_2$, we have $\mathbb{R}\cdot 1+\mathcal{N}_1\cong\mathbb{R}\cdot 1+\mathcal{N}_2$ iff $\mathcal{N}_1\cong\mathcal{N}_2$. The sufficiency is clear, and for the necessity we observe that $\mathcal{N}_i$ is the Jacobson radical in $\mathbb{R}\cdot 1+\mathcal{N}_i$, hence every isomorphism of the bigger algebras maps the smaller algebras onto each other. Thus $(\mathbb{R}\cdot 1+\mathcal{N}_\alpha)_{\alpha\in\Lambda}$ is the desired family of nonisomorphic algebras with unity. □

We are now going to compute the numbers of conjugacy classes of admissible Abelian groups of dimension $\leqslant 4$. The first case is extremely simple.

16 The case $k=2$*:* Clearly, every two-dimensional real algebra is isomorphic to $\mathbb{R}[X]/(X^2+a)$, with $a\in\{-1,0,1\}$, and any two of these algebras are nonisomorphic. For $a=1$ the algebra is isomorphic to $\mathbb{C}$; for $a=-1$ it is isomorphic to the product algebra $\mathbb{R}\times\mathbb{R}$. For $a=1$, we have only one connected component of the unit group. For $a=0$ we observe that $b+cX+(X^2)$ is invertible iff $b\neq 0$, hence the unit group has two connected components, which means that this algebra contributes two conjugacy classes. Finally, $\mathbb{R}\times\mathbb{R}$ has 4 components of the unit groups. The isomorphism $(r,s)\mapsto(s,r)$ maps the open subgroups of index 2 onto each other, hence we obtain 3 conjugacy classes of admissible groups. Summarizing, we have 6 two-dimensional groups, 3 of which are connected. □

For the higher-dimensional cases we need a bit more theory. The following structure theorem is useful in two ways: It reduces the task to classification of nilpotent algebras, and it simplifies the calculation of the unit groups.

In the following, $\mathbb{K}$ denotes $\mathbb{R}$ or $\mathbb{C}$. $\mathbb{C}$-algebras are of course $\mathbb{R}$-algebras as well, and $\mathrm{gl}(m,\mathbb{C})$ is canonically embedded in $\mathrm{gl}(2m,\mathbb{R})$. Recall that an associative algebra $\mathcal{A}$ is called nilpotent if $\mathcal{A}^n=\{0\}$ for some $n\in\mathbb{N}$. The smallest such n is called the nilpotency class of $\mathcal{A}$.

Theorem 17: *Let* $\mathcal{A}$ *be a real finite-dimensional commutative associative algebra with unity. Then there exists a decomposition (unique up to order of summands),*

$$\mathcal{A}=\mathcal{A}_1\oplus\mathcal{A}_2\oplus\cdots\oplus\mathcal{A}_\ell,$$

where each algebra $\mathcal{A}_i$ *is irreducible, i.e. cannot be written as the direct product of proper subalgebras.* $\mathcal{A}$ *is irreducible iff* $\mathcal{A}=1_{\mathcal{A}}\cdot\mathbb{K}+\mathcal{N}$*, with* $\mathcal{N}$ *a nilpotent* $\mathbb{K}$*-algebra.*

Proof: Since I could not find a reference of the theorem in this form, I include a short proof, using results from Ref. 18. The unique decomposition is obtained from the decomposition $1_{\mathcal{A}}=e_1+e_2+\ldots+e_\ell$, where the e_i are the minimal idempotents of $\mathcal{A}$, by letting $\mathcal{A}_i:=e_i\mathcal{A}$ (by Theorem 2.22 of Ref. 18). In particular, $\mathcal{A}$ is irreducible iff $1_{\mathcal{A}}$ is its only idempotent; thus adjoining a unit element to a nilpotent algebra $\mathcal{N}$ yields an irreducible algebra. For the other direction, we conclude from Ref. 18, Theorem 2.14 that the radical $\mathcal{N}$ of an irreducible algebra $\mathcal{A}$

is a maximal ideal (any proper nonnilpotent ideal would contain idempotents), whence $\mathcal{A}/\mathcal{N}$ is a (finite) field extension of $\mathbb{R}$, i.e., either $\mathbb{R}$ or $\mathbb{C}$. Now the Principal Theorem of Wedderburn (Ref. 18, Theorem 3.23) yields the desired structure of $\mathcal{A}$. □

For the calculations carried out below the following definition is useful.

Definition 18: For a finite-dimensional associative algebra $\mathcal{A}$ let

$$\#_{OS}(\mathcal{A}) := |\{H < \mathcal{A}^\times : H \ \ open\}/\approx|,$$

where $|A|$ denotes the cardinality of the set A and $\approx$ is the equivalence relation from Theorem 14.

The fact that an algebra isomorphism maps minimal idempotents onto minimal idempotents accounts for the uniqueness of the decomposition; it also simplifies the computation of $\#_{OS}(\mathcal{A})$, as the following proposition shows.

Proposition 19: Let $\mathcal{A}$ be a real finite-dimensional commutative associative algebra with unity, and let

$$\mathcal{A} = \mathcal{B}_1 \oplus \mathcal{B}_2 \oplus \cdots \oplus \mathcal{B}_m$$

be the (unique) decomposition into algebras $\mathcal{B}_i$, where each $\mathcal{B}_i$ is the direct sum of isomorphic irreducible algebras, and $\mathcal{B}_i$ and $\mathcal{B}_j$ have no isomorphic irreducible subalgebras in common, for $i \neq j$. Such a decomposition is obtained simply by collecting the isomorphic irreducible components. Then we have

$$\#_{OS}(\mathcal{A}) = \#_{OS}(\mathcal{B}_1) \cdot \#_{OS}(\mathcal{B}_2) \cdot \ \cdots \ \cdot \#_{OS}(\mathcal{B}_m).$$

If $\mathcal{B}$ is the direct sum of ℓ pairwise isomorphic irreducible $\mathbb{K}$-algebras, we have

$$\#_{OS}(\mathcal{B}) = \begin{cases} 1 & : \ \mathbb{K} = \mathbb{C}, \\ \ell + 1 & : \ \mathbb{K} = \mathbb{R}. \end{cases}$$

Proof: The first equation is immediate from the fact that an automorphism of $\mathcal{A}$ maps each $\mathcal{B}_i$ into itself, and that every open subgroup of $\mathcal{A}^\times$ is the direct product of open subgroups of the $\mathcal{B}_i^\times$. For the second equation, observe that an element $\alpha 1_{\mathcal{A}_i} + n$ of an irreducible component $\mathcal{A}_i$ ($\alpha \in \mathbb{K}$ and n nilpotent) is invertible iff $\alpha \neq 0$. Hence, the unit group of such an algebra is connected in the complex case and has two connected components in the real case. Since the unit group of the whole algebra is the direct product of the unit groups of the components, we obtain 2^ℓ connected components of the unit group in the real case and 1 in the complex case, whence we obtain $\#_{OS}(\mathcal{A}) = 1$ for the latter. For the former, it is easily seen that two open subgroups having the same number of connected components can be mapped onto each other by an algebra automorphism. □

In the following we will give a complete list of irreducible algebras of real dimension up to 4. k will denote the $\mathbb{K}$-dimension of the algebra and n the nilpotency class of the nilpotent part. Since the algebras of highest and lowest possible nilpotency class are extremely simple to describe, only the case $k=4, n=3$ is not straightforward.

(i) $\mathbb{K}=\mathbb{R}, k=2, n=2$: $\mathcal{A} \cong \mathbb{R}[X]/(X^2)$.
(ii) $\mathbb{K}=\mathbb{R}, k=3, n=2$: $\mathcal{A} \cong \mathbb{R}[X,Y]/(X^2, Y^2, XY)$.
(iii) $\mathbb{K}=\mathbb{R}, k=3, n=3$: $\mathcal{A} \cong \mathbb{R}[X]/(X^3)$.
(iv) $\mathbb{K}=\mathbb{R}, k=4, n=2$: $\mathcal{A} \cong \mathbb{R}[X,Y,Z]/(X^2, Y^2, Z^2, XY, XZ, YZ)$.
(v) $\mathbb{K}=\mathbb{R}, k=4, n=3$: $\mathcal{A} \cong \mathbb{R}[X,Y]/(X^3, Y^2 - aX^2, XY)$, with $a \in \{-1,0,1\}$. For the proof, suppose we are given $\mathcal{A} = \mathbb{R} \cdot 1 + \mathcal{N}$, where $\mathcal{N}$ is a three-dimensional algebra of nilpotency class 3. Pick $X \in \mathcal{N}$ with $X^2 \neq 0$, then X, X^2 and some suitably chosen $Y \in \mathcal{N}$ are a basis of $\mathcal{N}$. From $X^2Y = XY^2 = Y^3 = 0$ we obtain the relations $XY = bX^2$ and $Y^2 = aX^2$ for some $a, b \in \mathbb{R}$. Hence $\mathcal{A} \cong \mathbb{R}[X,Y]/(X^3, Y^2 - aX^2, XY - bX^2)$. Substituting by scalar multiples of X and Y, we may reduce the general case to $(a,b) \in \{-1,0,1\} \times \{0,1\}$. The substitution $Y \mapsto Y - X$ yields an isomorphism $\mathbb{R}[X,Y]/(X^3, Y^2 - X^2, XY - X^2) \to \mathbb{R}[X,Y]/(X^3, Y^2, XY)$ as well as an isomorphism $\mathbb{R}[X,Y]/(X^3, Y^2, XY - X^2) \to \mathbb{R}[X,Y]/(X^3, Y^2 + X^2, XY)$. Moreover, $\mathbb{R}[X,Y]/(X^3, Y^2 + X^2, XY - X^2) \cong \mathbb{R}[X,Y]/(X^3, Y^2, XY - X^2)$, by the substitution $X \mapsto \sqrt{2}X$, $Y \mapsto (\sqrt{2} - 1)X + Y$. In short, we

may assume $b=0$. Among the three remaining cases, $a=0$ yields the only algebra in which the radical has a two-dimensional annihilator (as a short calculation shows). Finally, $a=1$ and $a=-1$ yield nonisomorphic algebras, because in the case $a=1$ every element Z with $Z^2=0$ is a scalar multiple of X^2 (another simple calculation), whereas in the case $a=-1$ we have $(X+Y)^2=0=(X^2)^2$, and $X+Y, X^2$ are linearly independent.

(vi) $\mathbb{K}=\mathbb{R}, k=4, n=4$: $\mathcal{A}\cong\mathbb{R}[X]/(X^4)$.

(vii) $\mathbb{K}=\mathbb{C}, k=2, n=2$: $\mathcal{A}\cong\mathbb{C}[X]/(X^2)$.

Now we can calculate the numbers of conjugacy classes of admissible Abelian groups.

20 The case k=3 Combining the irreducible algebras just given, we obtain the following complete list of algebras and corresponding numbers of conjugacy classes:

$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}\oplus\mathbb{R}$	$\#_{OS}(\mathcal{A})=4$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{C}$	$\#_{OS}(\mathcal{A})=2$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}[X]/(X^2)$	$\#_{OS}(\mathcal{A})=4$
$\mathcal{A}=\mathbb{R}[X,Y]/(X^2,Y^2,XY)$	$\#_{OS}(\mathcal{A})=2$
$\mathcal{A}=\mathbb{R}[X]/(X^3)$	$\#_{OS}(\mathcal{A})=2$
5 algebras	14 conjugacy classes

Finally, the four-dimensional groups. □

21 The case k=4: Analogous to the three-dimensional case, we obtain the following.

$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}\oplus\mathbb{R}\oplus\mathbb{R}$	$\#_{OS}(\mathcal{A})=5$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}\oplus\mathbb{C}$	$\#_{OS}(\mathcal{A})=3$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}\oplus\mathbb{R}[X]/(X^2)$	$\#_{OS}(\mathcal{A})=6$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}[X,Y]/(X^2,Y^2,XY)$	$\#_{OS}(\mathcal{A})=4$
$\mathcal{A}=\mathbb{R}\oplus\mathbb{R}[X]/(X^3)$	$\#_{OS}(\mathcal{A})=4$
$\mathcal{A}=\mathbb{C}\oplus\mathbb{R}[X]/(X^2)$	$\#_{OS}(\mathcal{A})=2$
$\mathcal{A}=\mathbb{C}\oplus\mathbb{C}$	$\#_{OS}(\mathcal{A})=1$
$\mathcal{A}=\mathbb{R}[X]/(X^2)\oplus\mathbb{R}[X]/(X^2)$	$\#_{OS}(\mathcal{A})=3$
$\mathcal{A}=\mathbb{C}[X]/(X^2)$	$\#_{OS}(\mathcal{A})=1$
$\mathcal{A}=\mathbb{R}[X,Y]/(X^3,Y^2-aX^2,XY)$ (with $a\in\{-1,0,1\}$)	$\#_{OS}(\mathcal{A})=2$
12 algebras	35 conjugacy classes

It should be noted that the actual construction of the admissible groups from the algebras is straightforward: Direct product decompositions yield block matrix structures, and the relations defining an irreducible algebra allow an almost immediate calculation of the regular representation of that algebra.

We are now going to take a closer look at the admissibility condition that has to be fulfilled by the functions serving as wavelets (or *fiducial vectors* in the terminology of mathematical physics) for the wavelet transform. In Ref. 13, Theorem 13, it was shown how this condition could be obtained from the modular function of $\mathbb{R}^k\rtimes H$. More precisely, $g\in\pi_U$ is admissible iff

$$\int_U |\hat{g}(\omega)|^2\Psi(\omega)d(\omega)<\infty,$$

where the *admissibility function* Ψ is obtained by transporting the modular function of $\mathbb{R}^k\rtimes H$ to the open orbit U, i.e. choosing any $\gamma\in U$, we let $\Psi(\gamma h):=\Delta_G(0,h)=\Delta_H(h)|\det(h)|^{-1}=|\det(h)|^{-1}$. ($H$ is Abelian, hence $\Delta_H\equiv 1$.) We are going to calculate Ψ, restricting our attention

mainly to indecomposable groups. For the general case, observe that if an admissible dilation group H is the direct product of two lower-dimensional admissible dilation groups H_1 and H_2, i.e.,

$$H=\left\{\begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}: h_i \in H_i, \text{ for } i=1,2\right\},$$

then the open H-orbits are of the form $U=U_1\times U_2$ (U_i an open H_i-orbit), and the admissibility function Ψ is immediately calculated as $\Psi(x_1,x_2)=\Psi_1(x_1)\Psi_2(x_2)$, where Ψ_i is the admissibility function corresponding to the open H_i-orbit U_i. Another rather trivial observation concerns the behavior of admissible vectors when we replace a dilation group by a conjugate group: Clearly, if H is an admissible group and $A\in \mathrm{GL}(k,\mathbb{R})$, then $A^{-1}HA$ is an admissible dilation group and $f\in \mathrm{L}^2(\mathbb{R}^k)$ is admissible for H iff $f\circ A$ is admissible for $A^{-1}HA$.

Hence, we may restrict our attention to particularly simple representatives of conjugacy classes of indecomposable Abelian dilation groups, and for these the admissibility condition has an extremely simple form.

Theorem 22: *Let $H<\mathrm{GL}(k,\mathbb{R})$ be an admissible Abelian matrix group. Suppose that $\mathcal{A}:=\mathrm{span}(H)$ is indecomposable. Without loss of generality we may assume that $H=\mathcal{A}^\times$. Then, after a suitable change of coordinates, the single open orbit in $\widehat{\mathbb{R}^k}$ is*

$$U=\begin{cases} \{(\omega_1,\omega_2,\ \dots\ ,\omega_k): \omega_1^2+\omega_2^2\neq 0\} & : \quad \mathcal{A} \text{ is a } \mathbb{C}\text{-algebra,} \\ \{(\omega_1,\omega_2,\ \dots\ ,\omega_k): \omega_1\neq 0\} & : \quad \text{otherwise.} \end{cases}$$

and the admissibility function Ψ of the square integrable representation $\pi=\pi_U$ is given by

$$\Psi(\omega)=\begin{cases} (\omega_1^2+\omega_2^2)^{-k/2} & : \quad \mathcal{A} \text{ is a } \mathbb{C}-\text{algebra,} \\ |\omega_1|^{-k} & : \quad \text{otherwise.} \end{cases}$$

Proof: We have $\mathcal{A}=\mathbb{K}\cdot 1_k+\mathcal{N}$, with $\mathcal{N}$ nilpotent, and $\mathcal{A}^\mathsf{T}$ satisfies condition $(\mathcal{O})$. Let m denote the $\mathbb{K}$-dimension of $\mathcal{A}$, i.e., $m=k$ when $\mathbb{K}=\mathbb{R}$, $m=k/2$ otherwise. Let $N_2,\ \dots\ ,N_m$ denote a $\mathbb{K}$-basis of $\mathcal{N}$, let $\psi:\mathcal{A}\to\mathbb{K}^m$ be the $\mathbb{K}$-vector space isomorphism with $\psi(1_k)=e_1=(1,0,\ \dots\ ,0)^\mathsf{T}$ and $\psi(N_i)=e_i$ $(2\leqslant i\leqslant m)$, where $e_1,\ \dots\ ,e_m$ denotes the standard basis of $\mathbb{K}^m$. Let ρ denote the corresponding regular representation of $\mathcal{A}$, by Theorem 13 we may consider $\mathcal{B}:=\rho(\mathcal{A})^\mathsf{T}$ instead of $\mathcal{A}$. Now, given $A=a\cdot 1_k+\Sigma_{i=2}{}^m b_i N_i\in\mathcal{A}$, we see that

$$\rho(A)^\mathsf{T}=\begin{pmatrix} a & \mathbf{b} \\ 0_{m,1} & c_{m-1,m-1} \end{pmatrix},$$

where $\mathbf{b}=(b_2,\ \dots\ ,b_m)$ and $c_{m-1,m-1}$ is some element of $\mathrm{gl}(m-1,\mathbb{K})$. Now $\rho(A)$ is invertible iff $a\neq 0$, hence, choosing $\gamma:=(1,0,0,\ \dots\ ,0)\in\widehat{\mathbb{R}^k}$ yields

$$\gamma\mathcal{B}^\times=\begin{cases} \{(\omega_1,\omega_2,\ \dots\ ,\omega_k) & : \ \omega_1^2+\omega_2^2\neq 0\} & \mathbb{K}=\mathbb{C}, \\ \{(\omega_1,\omega_2,\ \dots\ ,\omega_k) & : \ \omega_1\neq 0\} & \mathbb{K}=\mathbb{R}. \end{cases}$$

The absolute value of the *real* determinant of $\rho(a\cdot 1_k+N)=a\cdot 1_k+\rho(N)$ with $a\in\mathbb{K}$ and N nilpotent is easily seen to be $|a|^k$, hence, we obtain for $(\omega_1,\omega_2,\ \dots\ ,\omega_k)\in U$, that

$$\Psi(\omega)=\begin{cases} (\omega_1^2+\omega_2^2)^{k/2} & : \quad \mathbb{K}=\mathbb{C}, \\ |\omega_1|^k & : \quad \mathbb{K}=\mathbb{R}. \end{cases}$$

□

Using this result it is not hard to show that in every dimension there exist (up to a change of coordinates) only finitely many admissibility conditions belonging to Abelian dilation groups. We end this section by noting some representation-theoretic consequences of our results.

Theorem 23: *Let $H<\mathrm{GL}(k,\mathbb{R})$ be closed and Abelian. Then the following are equivalent.*

(a) *The quasiregular representation of $\mathbb{R}^k \rtimes H$ on $\mathrm{L}^2(\mathbb{R}^k)$ contains an irreducible subrepresentation.*

(b) *The quasiregular representation of $\mathbb{R}^k \rtimes H$ on $\mathrm{L}^2(\mathbb{R}^k)$ is a finite direct sum of square integrable representations.*

(c) *There exists an Abelian closed subgroup H' containing H with H'/H finite, such that the quasiregular representation of $\mathbb{R}^k \rtimes H'$ on $\mathrm{L}^2(\mathbb{R}^k)$ is square integrable.*

Any of these conditions implies that the regular representation of $\mathbb{R}^k \rtimes H$ is the direct sum of irreducible subrepresentations.

Proof: $(c)\Rightarrow(b)\Rightarrow(a)$ is obvious. Concerning $(c)\Rightarrow(a)$, it follows from Ref. 13, Corollary 6 and Theorem 25 below, that H has an open orbit in $\hat{\mathbb{R}}^k$, hence, $H<\mathrm{span}(H)^\times$ with finite index, by Theorem 11, and by the same result the larger group is as desired. The statement concerning the regular representation has already been made in Ref. 12, Remark 4 after Theorem 2; for more details see Ref. 19. □

It can even be shown that in the case that the equivalent conditions of the previous theorem are fulfilled the regular representation of $\mathbb{R}^k \rtimes H$ is, in fact, an infinite multiple of π; i.e., π contains *all* square integrable representations of $\mathbb{R}^k \rtimes H$.

V. ERGODICALLY ACTING ABELIAN MATRIX GROUPS

This section is devoted to proving a characterization of Abelian matrix groups acting ergodically on some subset of $\mathbb{R}^k$. For this purpose, a lemma is needed which is intuitively clear and probably well known; however, I could not find a reference for it.

Lemma 24: Let G be group acting on a σ-finite measure space (X,μ), such that the action of G respects measurable sets and μ is quasi-invariant under G. Let H be a subgroup of G acting ergodically on some measurable, H-invariant subset $A\subset X$ of positive measure. Then there exists a measurable G-invariant subset $B\subset X$ on which G acts ergodically.

Proof: We may assume μ to be finite. Then H operates ergodically on some subset of positive measure iff there exists an $\epsilon>0$ such that for each finite union $X=U_1\cup U_2\cup\cdots\cup U_n$ with measurable, H-invariant sets U_i there exists an i with $\lambda(U_i)\geqslant\epsilon$. But clearly this must hold for G also. □

Theorem 25: *Let $H<\mathrm{GL}(k,\mathbb{R})$ be Abelian. H operates ergodically on a subset $U\subset\mathbb{R}^k$ iff $\overline{H}$ has an open orbit in $\mathbb{R}^k$.*

Proof: Let $H<\mathrm{GL}(k,\mathbb{R})$ be an Abelian group acting ergodically on some subset of $\mathbb{R}^k$ of positive measure. Let $H_1\supset H$ be a maximal Abelian subgroup of $\mathrm{GL}(k,\mathbb{R})$; by the previous lemma H_1 acts ergodically on a subset of $\mathbb{R}^k$. H_1 is an algebraic subgroup of $\mathrm{GL}(k,\mathbb{R})$: by Proposition 10 it is the intersection of $\mathrm{GL}(k,\mathbb{R})$ with a subalgebra of $\mathrm{gl}(k,\mathbb{R})$, i.e., with a vector space. Then by Ref. 20, Theorem 3.1.1 the H_1-orbits are locally closed, and Glimm's Theorem (Ref. 20, 2.1.14) finally implies that there exists an H_1-orbit of positive measure. Hence H is contained in an Abelian matrix group with an open orbit.

By Theorem 3, $\mathbb{R}^k$ is (up to a null set) the union of finitely many H_1-orbits, hence, we may assume that H operates ergodically on $U\subset H_1x$ for some $x\in\mathbb{R}^k$ with H_1x open. By Lemma 6, H_1 operates freely on the open orbit, which enables us to move the situation to H_1. Standard Lebesgue measure and the image of Haar measure of H_1 under the projection map $h\mapsto hx$ are equivalent, both being quasi-invariant regular measures. The action of H on H_1x, when moved to H_1, becomes action by translation. Summarizing, we have obtained an action of H by translation on a subset of H_1, which is ergodic with respect to Haar measure. But this is well-known to be the case iff H is dense in an open subgroup of H_1 (see for instance Ref. 20, Lemma 2.2.13). Hence $\overline{H}$ has an open orbit, by (2) of Remark 4.

The other direction uses the result just cited: If H is dense in a group H_1 with open orbit, then H operates ergodically on H_1. Since the operation of H_1 on the open orbit is free, this yields an ergodic operation of H on this open orbit. □

VI. CONCLUDING REMARKS

The object of this paper was the construction and classification of Abelian admissible dilation groups. The restriction to Abelian groups was mainly motivated by their greater accessibility. Even in this setting we obtain a great variety of groups, each of which gives rise to a family of wavelet transforms (indexed by the admissible vectors), and we have found a convenient way of constructing these groups. For the exploitation of the transforms for (more or less) practical purposes, there is still a lot to be done. The construction of the groups ensures the existence of wavelet inversion formulas (i.e., a wavelet synthesis corresponding to the wavelet analysis given by the transform), but the interpretation of the wavelet transform has not been considered. Given the central role that the reconstruction plays in most of the applications of continuous wavelet transforms, it seems plausible that at least some of the various possible groups yield transforms which can be usefully interpreted. So far only comparatively few have been considered.[4,21,22,8] Correspondingly, the physical relevance of a system of coherent states depends not only on the resolution of the identity but also on additional features of the coherent states (as required by the application at hand), which were not under consideration here.

Finally, it should be said that virtually none of the results concerning admissible Abelian groups may be generalized to the non-Abelian case. To begin with, as the example $H=\mathbb{R}^+ \cdot \mathrm{SO}(k)$ studied previously[21,22] shows there exist admissible groups which operate with nontrivial stabilizers on an open orbit in $\widehat{\mathbb{R}^k}$. Even in the case where the dilation group operates freely on an open orbit in $\widehat{\mathbb{R}^k}$, the Lie algebra of the group is no longer an associative algebra (take, for instance, the family of non-Abelian groups constructed in Ref. 12), thus the whole approach fails. The only exception is the construction of admissible groups from the regular representation of an associative algebra given in Proposition 12, and the fact that nonisomorphic algebras yield nonconjugate groups.

The equivalence established in Theorem 23 does not hold in the general case: It is, of course, possible that the quasi-regular representation is irreducible but not square integrable [take $H=\mathrm{GL}(k,\mathbb{R})$, for instance]. But even in the case where π has a square integrable subrepresentation it need not decompose into a direct sum of square integrable subrepresentations. As an example consider the group $H=\mathbb{R}^+ \cdot \mathrm{SO}_0(1,k-1)^4$ (for $k \geqslant 3$). Using Theorem 1, it is easily seen that the quasi-regular representation is the direct sum of two square integrable and one non-square integrable representation.

[1] J. R. Klauder and B. Skagerstam, in *Coherent States* (World Scientific, Singapore, 1985).
[2] S. T. Ali, J.-P. Antoine, J.-P. Gazeau, and U. A. Mueller, ''Coherent states and their generalizations: A mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995).
[3] E. W. Aslaksen and J. R. Klauder, ''Continuous representation theory using the affine group,'' J. Math. Phys. **10**, 2267–2275 (1969).
[4] G. Bohnke, ''Treillis d'ondelettes aux groupes de Lorentz,'' Ann. Inst. Henri Poincaré **54**, 245–259 (1991).
[5] A. Grossmann, J. Morlet, and T. Paul, ''Transforms associated to square integrable group representations I: General results,'' J. Math. Phys. **26**, 2473–2479 (1985).
[6] C. J. Isham and J. R. Klauder, ''Coherent states for n-dimensional Euclidean groups $E(n)$ and their application,'' J. Math. Phys. **32**, 607–620 (1991).
[7] J. R. Klauder and R. F. Streater, ''A wavelet transform for the Poincaré group,'' J. Math. Phys. **32**, 1609–1611 (1991).
[8] B. Torrésani, ''Wavelets associated with representations of the affine Weyl-Heisenberg group,'' J. Math. Phys. **32**, 1273–1279 (1991).
[9] M. Duflo and C. C. Moore, ''On the regular representation of a nonunimodular locally compact group,'' J. Funct. Anal. **21**, 209–243 (1976).
[10] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, ''Square integrability of group representations on homogeneous spaces. I. Reproducing triples and frames,'' Ann. Inst. Henri Poincaré **55**, 829–856 (1991).
[11] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, ''Square integrability of group representations on homogeneous spaces. II. Coherent and quasi-coherent states. The case of the Poincaré group,'' Ann. Inst. Henri Poincaré **55**, 857–890 (1991).
[12] D. Bernier and K. Taylor, ''Wavelets from square-integrable representations,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **27**, 594–608 (1996).
[13] H. Führ, ''Wavelet frames and admissibility in higher dimensions,'' J. Math. Phys. **37**, 6353–6366 (1996).
[14] P. Green, ''Square-integrable representations and the dual topology,'' J. Funct. Anal. **35**, 279–294 (1980).
[15] Theorem 3 of H. Whitney, ''Elementary structure of a real algebraic variety,'' Ann. Math. **66**, 545–556 (1957).
[16] Theorem 20.16 of E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis I* (Springer Verlag, Berlin, 1963).
[17] D. A. Suprunenko and R. I. Tyshkevich, *Commutative Matrices* (Academic, New York, 1968).
[18] A. A. Albert, *Structure of Algebras* (American Mathematical Society Colloquium Publications, Providence, RI, 1961).

[19] L. Baggett and K. Taylor, ''Groups with completely reducible regular representations,'' Proc. Am. Math. Soc. **72**, 593–600 (1978).
[20] R. J. Zimmer, *Ergodic Theory and Semisimple Groups* (Birkhäuser, Boston, 1984).
[21] J.-P. Antoine and R. Murenzi, ''Two-dimensional directional wavelets and the scale-angle representation,'' Signal Process. **52**, 259–281 (1996).
[22] M. Duval-Destin and R. Murenzi, ''Spatio-temporal wavelets: Application to the analysis of moving patterns,'' in *Progress in Wavelet Analysis and Applications*, edited by Y. Meyer and S. Roques (Editions Frontière, Gif-sur-Yvette, 1993), pp. 399–408.

Wavelets on the *n*-sphere and related manifolds

J.-P. Antoine and P. Vandergheynst
Institut de Physique Théorique, Université Catholique de Louvain,
B - 1348 Louvain-la-Neuve, Belgium

(Received 15 December 1997; accepted for publication 31 March 1998)

We present a purely group-theoretical derivation of the continuous wavelet transform (CWT) on the $(n-1)$-sphere S^{n-1}, based on the construction of general coherent states associated to square integrable group representations. The parameter space of the CWT, $X \sim \mathrm{SO}(n) \times \mathbb{R}^+_*$, is embedded into the generalized Lorentz group $\mathrm{SO}_0(n,1)$ via the Iwasawa decomposition, so that $X \simeq \mathrm{SO}_0(n,1)/N$, where $N \simeq \mathbb{R}^{n-1}$. Then the CWT on S^{n-1} is derived from a suitable unitary representation of $\mathrm{SO}_0(n,1)$ acting in the space $L^2(S^{n-1}, d\mu)$ of finite energy signals on S^{n-1}, which turns out to be square integrable over X. We find a necessary condition for the admissibility of a wavelet, in the form of a zero mean condition, which entails all the usual filtering properties of the CWT. Next the Euclidean limit of this CWT on S^{n-1} is obtained by redoing the construction on a sphere of radius R and performing a group contraction for $R \to \infty$, from which one recovers the usual CWT on flat Euclidean space. Finally, we discuss the extension of this construction to the two-sheeted hyperboloid $H^{n-1} \sim \mathrm{SO}_0(n-1,1)/\mathrm{SO}(n-1)$ and some other Riemannian symmetric spaces. © *1998 American Institute of Physics.*
[S0022-2488(98)00308-9]

I. INTRODUCTION

In most cases of physical interest, experimental data are given on the line (signal processing), on the plane (image analysis), or occasionally in $\mathbb{R}^3$ (e.g., in fluid dynamics). However, there are instances where data are given on a *sphere*. Of course, geophysical data is the prime example (e.g., weather prediction or global circulation models), but others appear, for instance, in statistical problems, computer vision or medical imaging (see Ref. 1 for a list and precise references). In such cases, not only S^2, but also higher dimensional spheres or other manifolds may occur. The standard methods are based on Fourier analysis, but analyzing data with the continuous wavelet transform (CWT) is by now a well-established procedure (see Ref. 2 for a survey of applications in physics). So the question arises, how does one extend the CWT to the sphere or another manifold?

Let us first make that statement precise. We may speak of a genuine spherical CWT if (i) the signals and the wavelets live *on* the sphere; (ii) the transform involves (local) dilations of some kind; and (iii) for small scales, the spherical CWT reduces to the usual CWT on the (tangent) plane (Euclidean limit). The same criteria apply to other manifolds.

The problem has attracted a lot of interest in the last couple of years and many proposals have been made, but, in our opinion, none of them is fully satisfactory according to these criteria. In particular, local dilations are in general missing. For instance:

(i) A number of works extend to S^2 the discrete wavelet scheme based on a multiresolution analysis, using adapted interpolation methods and spline functions.[3–5] This approach is in general motivated by numerical purposes, but leads often to difficulties around the poles. A different technique is to use second generation wavelets.[6,7]

(ii) Others[8–10] exploit the geometry of the sphere, as encoded in the system of spherical harmonics, but as a result their analyzing functions are poorly localized. In fact they do not really resemble wavelets. A comprehensive survey of this approach may be found in the recent monograph of Freeden *et al.*[11]

(iii) One defines a WT on the tangent bundle of the sphere[12] or instead a Gabor transform on the sphere itself[13] (since no dilation is then involved).

 © 1998 American Institute of Physics

(iv) The most satisfactory approach is that of Holschneider,[14] who produces a CWT on S^2 that satisfies the three criteria above. However, the role of dilation is played by an abstract parameter that satisfies a number of *ad hoc* assumptions. The correct Euclidean limit is obtained, but it is essentially put by hand.

As can be seen from this brief description, none of the proposed solutions fully qualifies for a genuine CWT on S^2. In a previous paper,[15] we have presented a new approach to the CWT on the 2-sphere, entirely derived from group theory, following the formalism of general coherent states developed in Ref. 16. In particular, the Euclidean limit has also a precise group-theoretical formulation, in terms of *group contraction*.[17,18]

In the present paper, we shall extend our construction to spheres in higher dimensions (S^{n-1}) and discuss the application to other manifolds, for instance, a two-sheeted hyperboloid or similar Riemannian symmetric spaces. The paper is organized as follows. In Sec. II, we present a rapid survey of the coherent state construction based on square integrable group representations.[16] In Sec. III we describe the affine transformations of S^{n-1}, that is, motions and local dilations, and show that they belong to the conformal group of S^{n-1}, namely the Lorentz group $\mathrm{SO}_0(n,1)$. Then the parameter space of the CWT may be identified as $X \sim \mathrm{SO}(n)\cdot \mathbb{R}^+_* \sim \mathrm{SO}_0(n,1)/\mathbb{R}^{n-1}$, using the Iwasawa decomposition of $\mathrm{SO}_0(n,1)$. Then, in Sec. IV, we select a suitable unitary representation of $\mathrm{SO}_0(n,1)$, acting in the space $L^2(S^{n-1},d\mu)$ of finite energy signals on S^{n-1}, and show that it is square integrable over X. From there, we derive the CWT on S^{n-1}. Section V is devoted to the Euclidean limit, using the theory of contraction of group representations (the main definitions for the latter are collected in the Appendix). Finally, in Sec. VI, we discuss the extension of the construction to other manifolds such as, for instance, the two-sheeted hyperboloid $H^{n-1} \sim \mathrm{SO}_0(n-1,1)/\mathrm{SO}(n-1)$ and some Riemannian symmetric spaces.

As will be seen, the present results are rather similar to those obtained previously in the case of the 2-sphere.[15] As a byproduct, we also obtain a CWT on the circle, for $n=2$. The latter is quite different from that of Holschneider,[19] which is largely based on a periodization procedure applied to wavelets on the interval.

II. GENERAL SETUP: THE CWT ON A MANIFOLD

We begin by a brief sketch of the method of construction of coherent states (CS) associated to a group representation. Further details may be found, for instance, in the review paper, Ref. 16, and the references quoted therein.

Let Y be a manifold. For instance, Y could be space $\mathbb{R}^n$, the 2-sphere S^2, space-time $\mathbb{R}\times\mathbb{R}$ or $\mathbb{R}^2\times\mathbb{R}$, etc. In order to construct coherent states on Y, one typically needs two ingredients:

(1) the class of finite energy signals living on Y, i.e., the space $L^2(Y,d\mu)\equiv\mathcal{H}$.
(2) a (locally compact) group G of transformations acting (transitively) on Y, i.e., $y\mapsto g[y]$, with $g[g'[y]]=gg'[y]$, $e[y]=y$, and for any pair $y,y'\in Y$, there is at least one $g\in G$ such that $g[y]=y'$.

From this one obtains a natural unitary representation of G in the space $L^2(Y,d\mu)$, namely the left quasi-regular representation (for simplicity, we assume that μ is G-invariant, but quasi-invariance suffices):

$$[U(g)f](y)=f(g^{-1}y). \tag{2.1}$$

Then a system of CS on Y associated to G may be defined if U is a *square integrable* representation of G, that is, U is irreducible (cyclicity would be sufficient) and there exists a nonzero vector $\eta\in L^2(Y,d\mu)$, called *admissible*, such that the matrix element $\langle U(g)\,\eta|\,\eta\rangle$ is square integrable as a function on G, with respect to the (left or right) Haar measure on G. When this is the case, the corresponding CS, indexed by G, are obtained as the vectors in the orbit of the admissible vector η under U:

$$\eta_g=U(g)\,\eta, \quad g\in G. \tag{2.2}$$

Quite often, however, the representation U is not square integrable in the strict sense just described [it would be a discrete series representation, and many groups have no discrete series; a case in point is the Lorentz group $SO_0(n,1)$, for $n\neq 2$]. However, it may become square integrable when restricted to a homogeneous space $X=G/H$, for some closed subgroup H. By this we mean the following. Let $\sigma: X\to G$ be a Borel section. Then the nonzero vector $\eta\in L^2(Y,d\mu)$ is said to be *admissible* mod(H,σ), and the representation U *square integrable* mod(H,σ), if the following condition holds:

$$c_X(\eta,\phi)\equiv\int_X d\nu(x)|\langle U(\sigma(x))\,\eta|\phi\rangle|^2<\infty,\quad \forall\,\phi\in\mathcal{H} \tag{2.3}$$

(we assume that ν is a G-invariant measure on X; again this is not a restriction). Then CS indexed by X may be defined as

$$\eta_{\sigma(x)}=U(\sigma(x))\,\eta,\quad x\in X. \tag{2.4}$$

The condition (2.3) may also be rewritten as

$$0<\int_X d\nu(x)|\langle\,\eta_{\sigma(x)}|\phi\rangle|^2=\langle\phi|A_\sigma\phi\rangle<\infty,\quad \forall\,\phi\in\mathcal{H}, \tag{2.5}$$

where A_σ is a positive, bounded, invertible operator.[16] If the operator A_σ^{-1} is also bounded, the family $\mathcal{S}_\sigma=\{\eta_{\sigma(x)},x\in X\}$ is called a *frame*, and a *tight frame* if $A_\sigma=\lambda I$, for some $\lambda>0$. This terminology is familiar in the discrete case, for instance, in wavelet or Gabor analysis.[20,21] From now on, we shall normalize the admissible vector η by $c_X(\eta,\eta)=1$, and we shall assume that it generates a frame $\mathcal{S}_\sigma$ (otherwise domain problems arise).

Under these assumptions, the CS defined in (2.4) have the following properties:

(1) The set $\mathcal{S}_\sigma$ is total in $\mathcal{H}$: $(\mathcal{S}_\sigma)^\perp=\{0\}$.

(2) The range, $\mathcal{H}_\eta$, of the linear map $W_\eta:\mathcal{H}\to L^2(X,d\nu)$, given by

$$(W_\eta\phi)(x)=\langle\,\eta_{\sigma(x)}|\phi\rangle,\quad \phi\in\mathcal{H}, \tag{2.6}$$

is complete with respect to the scalar product $\langle\Phi|\Psi\rangle_\eta\equiv\langle\Phi|W_\eta A_\sigma^{-1}W_\eta^{-1}\Psi\rangle$ and W_η is unitary from $\mathcal{H}$ onto $\mathcal{H}_\eta$.

(3) The orthogonal projection from $L^2(X,d\nu)$ onto $\mathcal{H}_\eta$ is an integral operator K_σ, and $\mathcal{H}_\eta$ is a reproducing kernel Hilbert space of functions:

$$\Phi(x)=\int_X d\nu(y)K_\sigma(x,y)\Phi(y),\quad \forall\Phi\in\mathcal{H}_\eta. \tag{2.7}$$

The kernel is given explicitly by $K_\sigma(x,y)=\langle\,\eta_{\sigma(x)}|A_\sigma^{-1}\eta_{\sigma(y)}\rangle$, that is, K_σ is essentially the autocorrelation function of η.

(4) The map W_η may be inverted on its range by the adjoint operator, $W_\eta^{-1}=W_\eta^*$ on $\mathcal{H}_\eta$, to obtain the reconstruction formula:

$$W_\eta^{-1}\Phi=\int_X d\nu(x)\Phi(x)A_\sigma^{-1}\eta_{\sigma(x)},\quad \Phi\in\mathcal{H}_\eta. \tag{2.8}$$

In more familiar terms, properties (1) and (2) mean that the CS family $\mathcal{S}_\sigma$ yields a new representation of the system under consideration, unitary equivalent to the original one. By property (3), the Hilbert space of this CS representation is a space of (usually nice) functions. Finally, property (4) means that the vector $\phi=W_\eta^{-1}\Phi$ may be expanded in terms of (reconstructed from) CS $\eta_{\sigma(x)}$, in the same sense as an expansion in terms of (possibly generalized) eigenvectors of some self-adjoint operator, exactly as for Fourier series or integrals, or, for that matter, wavelet expansions.

A particular case of this construction is that considered, independently, by Gilmore,[22] and Perelomov,[23] namely the situation where the subgroup H is the subgroup H_η of G that leaves the admissible vector η invariant up to a phase:

$$U(h)\,\eta = e^{i\omega(h)}\,\eta, \quad h \in H_\eta, \tag{2.9}$$

where $\omega(h)$ is a real-valued, measurable function on H_η. Then the admissibility condition (2.3) is independent of the choice of the section σ, and the frame is automatically tight, with $A_\sigma = I$.

In the case where η is invariant under H (thus $H \subset H_\eta$), this construction is standard in noncommutative harmonic analysis (U is then called a representation of the relative discrete series of X). For a review and original references, see Ref. 24 [where the CS map (2.6) is called the Poisson transform].

Let us conclude by giving some familiar examples of this construction.

(1) The ''$ax+b$'' group acting on $\mathbb{R}$ yields the usual one-dimensional (1-D) continuous wavelets.
(2) The Weyl–Heisenberg group, also acting on $\mathbb{R}$, gives the windowed Fourier transform, or Gabor transform. Here all vectors are admissible.
(3) The similitude group of $\mathbb{R}^n$, $\mathrm{SIM}(n) = \mathbb{R}^n \rtimes (\mathbb{R}^+_* \times \mathrm{SO}(n))$, consisting of translations $b \in \mathbb{R}^n$, dilations $a > 0$ and rotations $R \in \mathrm{SO}(n)$, yields the n-dimensional wavelets ($\rtimes$ denotes a semidirect product). For an axisymmetric wavelet η, the isotropy group H is $\mathrm{SO}(n-1)$ and so $X = \mathbb{R}^n \cdot \mathbb{R}^+_* \cdot S^{n-1} \sim \mathbb{R}^{2n}$ (which may be interpreted as phase space), and we are in the Gilmore–Perelomov situation.
(4) Coherent states on the Galilei group or the Poincaré group, both inaccessible to the standard Gilmore–Perelomov method.[16]

III. AFFINE TRANSFORMATIONS ON THE SPHERE S^{n-1}

The usual CWT on the line is derived from a natural unitary representation of the ''$ax+b$'' group, namely the left quasi-regular representation (4.1) in the space of finite energy signals $L^2(\mathbb{R}, dx)$.[25,26] Similarly, in two dimensions, one starts from the quasi-regular representation of the similitude group of the plane in the space $L^2(\mathbb{R}^2, d^2x)$.[27,28] Note that in each case, the group acts transitively on the basis manifold, $\mathbb{R}$ or $\mathbb{R}^2$. The same scheme applies to the CWT on a general manifold, subject to the transitive action of some group of transformations that contains dilations.

Let us apply this method to the sphere S^{n-1} and consider the space of finite energy signals $\mathcal{H} = L^2(S^{n-1}, d\mu)$, where $d\mu(\omega)$ is the usual (rotation invariant) measure on S^{n-1}. The first step for constructing a CWT on S^{n-1} is to identify the appropriate affine transformations. For that purpose, we note first the isomorphism $S^{n-1} \simeq \mathrm{SO}(n)/\mathrm{SO}(n-1)$, from which we deduce immediately that $\mathrm{SO}(n)$ acts transitively on S^{n-1}; these are the so-called motions or displacements on the sphere. Correspondingly, the left quasi-regular representation (4.1), acting in $\mathcal{H} = L^2(S^{n-1}, d\mu)$, is easily computed using the matrix realization of $\mathrm{SO}(n)$ in $\mathbb{R}^n$. It decomposes into the direct sum of all unitary irreducible representations (UIRs) of $\mathrm{SO}(n)$, and this is simply the Fourier series on S^{n-1} (see Sec. IV).

In order to obtain dilations, and also for generalization to other manifolds, let us recall some well-known geometrical facts.[29,30] Let $\mathcal{M}$ be an n dimensional (pseudo-) Riemannian manifold with metric $g_{\mu\nu}$. A *motion* is a transformation $z^\mu = z^\mu(x^1, \ldots, x^n)$, that leaves the metric invariant in the sense that

$$g'_{\mu\nu} \equiv \frac{\partial z^\alpha}{\partial x^\mu} \cdot g_{\alpha\beta} \cdot \frac{\partial z^\beta}{\partial x^\nu} = g_{\mu\nu}. \tag{3.1}$$

For the cases $\mathcal{M} = \mathbb{R}^n, S^n$ and $\mathbb{R}^n_{(p,q)}$ [the pseudo-Euclidean space with metric (p,q)] the motion groups are, respectively, the Euclidean group $E(n) = \mathbb{R}^n \rtimes \mathrm{SO}(n)$, the linear orthogonal group $\mathrm{O}(n+1)$ and the pseudo-orthogonal group $\mathrm{O}(p,q)$ (note that, in the last two cases, reflections are irrelevant for our purposes).

The map $\{x^\mu\} \mapsto \{z^\nu\}$ is a *conformal transformation* if the transformed metric is locally proportional to the initial one:

$$g'_{\mu\nu}(x)=\lambda(x)g_{\mu\nu}(x). \tag{3.2}$$

In the case of the the pseudo-Euclidean space $\mathbb{R}^n_{(p,q)}$, obvious conformal transformations are motions from $O(p,q)$, dilations $x\mapsto a\cdot x,(a>0)$, and inversions. But in fact, there are no other ones, according to the theorem of Liouville: Every differentiable conformal transformation of a domain in the (pseudo-)Euclidean space $\mathbb{R}^n_{(p,q)}, n\geqslant 3$, can be expressed as the composition of a motion, a dilation and an inversion. Note that we consider here *global* conformal transformations. For $n=2$, any analytic function defines a local conformal transformation, but for $n\geqslant 3$, every local conformal transformation is the restriction of a global one, so that no ambiguity arises.

Writing now explicitly the infinitesimal generators of such transformations, as first order differential operators, one obtains the Lie algebra of the conformal group of $\mathbb{R}^n_{(p,q)}$, and it is isomorphic to the Lie algebra $\mathfrak{so}(p+1,q+1)$ of the pseudo-orthogonal group $SO_0(p+1,q+1)$. For the Euclidean case $\mathbb{R}^{n-1}$, the conformal group is the Lorentz group $SO_0(n,1)$. Now the sphere S^{n-1}, with its usual metric, is conformally equivalent to the Euclidean space $\mathbb{R}^{n-1}$. The explicit correspondence may be visualized as the stereographic projection from the South Pole to the tangent hyperplane at the North Pole. Therefore S^{n-1} and $\mathbb{R}^{n-1}$ have the same conformal group, that is, $SO_0(n,1)$. This is the group we are looking for. It remains to compute explicitly the action of $SO_0(n,1)$ on the sphere S^{n-1} in order to identify the dilations, and then to choose an appropriate reprepresentation of $SO_0(n,1)$ in $L^2(S^{n-1},d\mu)$.

First we have to parametrize the sphere S^{n-1}. Let $(x_0,x_1,x_2,\ldots,x_n)$ denote the Cartesian coordinates in Minkowski space $\mathbb{R}^{n+1}_{(1,n)}$ on which $SO_0(n,1)$ acts.[31] We choose the poles of the sphere along the x_n-axis, and introduce spherical coordinates by $\omega=(\theta_1,\theta_2,\ldots,\theta_{n-1})\in S^{n-1}$, in the usual way:

$$\begin{aligned} x_1 &= \sin\theta_{n-1}\cdots\sin\theta_2\sin\theta_1,\\ x_2 &= \sin\theta_{n-1}\cdots\sin\theta_2\cos\theta_1,\\ &\cdots\\ x_n &= \cos\theta_{n-1}, \end{aligned} \tag{3.3}$$

with $0\leqslant\theta_1<2\pi, 0\leqslant\theta_2,\ldots,\theta_{n-1}\leqslant\pi$ (the Euler angles). In these coordinates, the normalized $SO(n)$-invariant measure on S^{n-1} reads

$$d\mu(\omega)=\frac{\Gamma(n/2)}{2\pi^{n/2}}\sin\theta_2\sin^2\theta_3\cdots\sin^{n-2}\theta_{n-1}\,d\theta_1\,d\theta_2\cdots d\theta_{n-1}. \tag{3.4}$$

In order to compute the action of $SO_0(n,1)$ on S^{n-1}, we use the Iwasawa decomposition,[29,32] which gives the unique factorization

$$SO_0(n,1)=KAN, \tag{3.5}$$

where $K\sim SO(n)$ is the maximal compact subgroup, corresponding to rotations in the subspace $(x_1,x_2,\ldots,x_n)$; $A\sim SO_0(1,1)\sim\mathbb{R}^+_*\sim\mathbb{R}$ is the one-dimensional subgroup of Lorentz boosts along the x_n-axis; $N\sim\mathbb{R}^{n-1}$ is $(n-1)$-dimensional and Abelian, and corresponds under stereographic projection to translations in the tangent hyperplane at the North Pole.

In the natural $(n+1)\times(n+1)$ matrix representation, an element of A is given as

$$\hat{a}(t)=\begin{pmatrix}\cosh t & 0^t & \sinh t\\ 0 & \mathbb{I}_{n-1} & 0\\ \sinh t & 0^t & \cosh t\end{pmatrix}, \tag{3.6}$$

whereas elements of the subgroup N are indexed by vectors $\xi\in\mathbb{R}^{n-1}$ and read, with $\Delta=\|\xi\|_2^2$,

$$\hat{n}(\xi)=\begin{pmatrix} 1+\frac{\Delta}{2} & \xi^t & -\frac{\Delta}{2} \\ \xi & \mathbb{I}_{n-1} & -\xi \\ \frac{\Delta}{2} & \xi^t & 1-\frac{\Delta}{2} \end{pmatrix}. \tag{3.7}$$

Thus every $g\in \mathrm{SO}_0(n,1)$ can be written uniquely as $g=\gamma a n(\xi), \gamma\in \mathrm{SO}(n), a\in A, n\in N$.

Let M be the centralizer of A in K, that is, $M=\{\gamma\in K:\ \gamma a=a\gamma,\ \forall a\in A\}$ [since $\mathrm{SO}_0(n,1)$ is connected and has finiter center, we take the simpler definition of M[29,32]]. Then $P=MAN$ is a closed subgroup of G, called the minimal parabolic subgroup. In the present case, we see that $M=\mathrm{SO}(n-1)$, corresponding to rotations around the x_n-axis, so that the stability subgroup of the North Pole is precisely P. This subgroup P is not invariant, but $Y=G/P\simeq K/M$, which gives here $Y=\mathrm{SO}_0(n,1)/\mathrm{SO}(n-1)\cdot A\cdot N\simeq S^{n-1}$. This shows that $\mathrm{SO}_0(n,1)$ acts transitively on S^{n-1}. We have identified above $K=\mathrm{SO}(n)$ with motions on S^{n-1} and A with dilations, which constitute our basic operations. Thus the parameter space of our theory is the homogeneous space $X\equiv \mathrm{SO}_0(n,1)/N\simeq \mathrm{SO}(n)\cdot A$.

In the spherical coordinates (3.3), any $\gamma\in \mathrm{SO}(n)$ admits the following Euler decomposition

$$\gamma=m(\psi)u_{n-1}(\theta_{n-1})\cdots u_1(\theta_1), \tag{3.8}$$

where $m(\psi)\in \mathrm{SO}(n-1), \{\theta_1,\ldots,\theta_{n-1}\}$ are the Euler angles on S^{n-1} and $u_j(\theta_j)$ is a rotation in the 2-plane x_j, x_{j+1} (this expression is unique except when θ_j is equal to 0 or π, $3\leqslant j\leqslant n$). For completeness, we write the integration formula on $\mathrm{SO}_0(n,1)$:

$$\int_{\mathrm{SO}_0(n,1)} dg\ f(g)=\int_{\mathrm{SO}(n)} d\mu_n(\gamma)\int_A\int_N dt\ d^{n-1}\xi\ e^{(n-1)t} f(\gamma a(t)n(\xi)), \tag{3.9}$$

where $d\mu_n(\gamma)$ is the invariant measure on $\mathrm{SO}(n)$. In other words, our parameter space $X=\mathrm{SO}_0(n,1)/N\simeq \mathrm{SO}(n)\cdot A$ admits the $\mathrm{SO}_0(n,1)$-invariant measure

$$d\nu(x)=d\mu_n(\gamma)e^{(n-1)t}dt=\frac{d\mu_n(\gamma)da}{a^n}\ (a=e^{-t}, t\in\mathbb{R}). \tag{3.10}$$

We are now in a position to compute the action of $\mathrm{SO}_0(n,1)$ on S^{n-1}, using the Iwasawa decomposition of the latter. Indeed, for $g_0\in \mathrm{SO}_0(n,1)$, we get in this way

$$g_0 g=\gamma_{g_0} a_{g_0} n_{g_0}. \tag{3.11}$$

This induces a homeomorphism on the quotient, indeed the map $\gamma\mapsto\gamma_{g_0}$ does not depend on γ, but rather on its equivalence class in $K/M\simeq S^{n-1}$. In view of the Cartan decomposition $G=KAK$,[29,32] it is enough to show this for elements $g_0\equiv g_0(a)\in A$, since the action of K is trivial. Let us express $\gamma\in \mathrm{SO}(n)$ by its Euler decomposition (3.8), $\gamma=\gamma(\psi,\theta_1,\ldots,\theta_{n-1})$. By definition of $M\sim \mathrm{SO}(n-1)$, g_0 commutes with $m(\psi)$, so that

$$g_0\gamma=m(\psi)g_0 u_{n-1}(\theta_{n-1})\cdots u_1(\theta_1). \tag{3.12}$$

For a pure dilation, $g_0\equiv g_0(a)\in A$, this reduces to

$$g_0\gamma=m(\psi)u_{n-1}((\theta_{n-1})_{g_0})u_{n-2}(\theta_{n-2})\cdots u_1(\theta_1). \tag{3.13}$$

Indeed an explicit calculation shows that the action of a pure dilation on S^{n-1} is given by $\omega=\{\theta_j\}\mapsto\omega_a=\{(\theta_j)_a\}$, where

$$(\theta_j)_a=\theta_j, j=1,\ldots,n-2,$$
$$\tan\frac{(\theta_{n-1})_a}{2}=a\ \tan\frac{\theta_{n-1}}{2}. \tag{3.14}$$

Thus the transformation corresponding to a pure dilation is exactly the usual Euclidean dilation lifted on S^{n-1} by inverse stereographic projection.

IV. THE CONTINUOUS WAVELET TRANSFORM ON S^{n-1}

A. Principal series representations of the Lorentz group $SO_0(n,1)$

We have seen that the parameter space of the CWT on S^{n-1} is $X\equiv \mathrm{SO}_0(n,1)/N\simeq \mathrm{SO}(n)\cdot A$. According to our program, as outlined in Sec. II, the construction of spherical wavelets, that is, affine coherent states on S^{n-1}, proceeds in two steps. First, to select a realization of X by operators acting in the Hilbert space $L^2(S^{n-1},d\mu)$, that will come, of course, from a unitary irreducible representation U of $\mathrm{SO}_0(n,1)$ on the same space. Then to show that U is square integrable on X, that is, modulo N and a suitable section $\sigma: X\to \mathrm{SO}_0(n,1)$.

Let us look first at rotations only. A natural unitary representation of $\mathrm{SO}(n)$ is the quasi-regular representation,

$$(U_{qr}(\gamma)f)(\omega)=f(\gamma^{-1}\omega),\quad \gamma\in \mathrm{SO}(n),\quad f\in L^2(S^{n-1},d\mu). \tag{4.1}$$

U_{qr} is infinite dimensional and decomposes into the direct sum of all the irreducible unitary representations U_l, $l=0,1,2,...$, which are of finite dimension d_l. A convenient orthonormal basis in the d_l-dimensional representation space of U_l is that of the hyperspherical harmonics Y_l^M, where $M\equiv(m_1,m_2,\cdots,m_{n-2})$ runs over the set I_l^n of all integers m_j such that $l\geqslant m_{n-2}\geqslant\cdots\geqslant m_2\geqslant|m_1|\geqslant 0$. Thus the dimension of U_l is

$$d_l=\#I_l^n=\frac{(2l+n-2)(l+n-3)!}{(n-2)!\,l!}, \tag{4.2}$$

which equals $2l+1$ for $n=3$, $(l+1)^2$ for $n=4$, etc.[33,34] For $n=3$, $Y_l^M(\omega)$ reduces to the usual spherical harmonic $Y_l^{m_1}(\theta_2,\theta_1)$, and for $n\geqslant 4$, $Y_l^M(\theta_1,\ldots,\theta_{n-1})$ is the product of $Y_{m_2}^{m_1}(\theta_2,\theta_1)$ and $(n-3)$ factors $(\sin\theta_j)^{n-j}C_{q_j}^{p_j}(\cos\theta_j)$, for $j=3,\ldots,n-1$, where C_q^p is a Gegenbauer polynomial. In these notations, the expansion of an arbitrary function of $L^2(S^{n-1},d\mu)$ in this hyperspherical basis is just Fourier analysis on S^{n-1}:

$$f(\omega)=\sum_{l=0}^{\infty}\sum_{M\in I_l^n}\hat{f}(l,M)Y_l^M(\omega), \tag{4.3}$$

$$\|f\|^2=\sum_{l=0}^{\infty}\sum_{M\in I_l^n}|\hat{f}(l,M)|^2, \tag{4.4}$$

where $\hat{f}(l,M)=\langle Y_l^M|f\rangle$ denotes the Fourier coefficient of f. We will also make use of the familiar transformation formula for spherical harmonics:

$$[U_{qr}(\gamma)Y_l^M](\omega)=Y_l^M(\gamma^{-1}\omega)=\sum_{M'\in I_l^n}\mathcal{D}^l_{MM'}(\gamma)Y_l^{M'}(\omega),\quad \gamma\in \mathrm{SO}(n), \tag{4.5}$$

where $\mathcal{D}^l_{MM'}$ are Wigner D-functions.

In order to incorporate dilations into this scheme, we have to lift U_{qr} from $\mathrm{SO}(n)$ to $\mathrm{SO}_0(n,1)$. However, the measure $d\mu$ is not dilation invariant, so that a Radon-Nikodym derivative or cocycle $\lambda(g,\omega)$ will necessarily creep in, through the relation

$$d\mu(g^{-1}\omega)=\lambda(g,\omega)d\mu(\omega). \tag{4.6}$$

The function $\lambda(g,\omega)$ must satisfy the cocycle equation

$$\lambda(g_1 g_2,\omega)=\lambda(g_1,\omega)\cdot\lambda(g_2,{g_1}^{-1}\omega). \tag{4.7}$$

A good choice is given by the representations of the continuous principal series of $\mathrm{SO}_0(n,1)$. These representations, originally introduced by Harish-Chandra, are studied in detail the mathematical literature.[24,33,34] Among these are the so-called class I representations, for which there exists a vector in the carrier Hilbert space which is invariant under the action of the maximal compact subgroup $\mathrm{SO}(n)$.[32,33] Class I representations are induced by unitary irreducible representations of the minimal parabolic subgroup $P=MAN$, which are trivial on M. They are given by the operators:

$$[U^s(g)f](\omega)=\lambda(g,\omega)^{s/2}\chi(a)f(g^{-1}\omega),\quad g\in\mathrm{SO}_0(n,1),\quad s\in\mathbb{C},\quad f\in L^2(S^{n-1},d\mu), \tag{4.8}$$

where $g=\gamma a n$, the Iwasawa decomposition, χ is a character of A, and $\lambda(g,\omega)$ is a multiplier, satisfying the cocycle equation (4.7). In the sequel we will choose the trivial character $\chi(a)\equiv 1$. The main properties of these principal series representations are summarized by the following theorem.[33,34]

Theorem 4.1: *The representation*

$$[U^s(g)f](\omega)=\lambda(g,\omega)^{s/2}f(g^{-1}\omega), \tag{4.9}$$

is a strongly continuous representation of $\mathrm{SO}_0(n,1)$ *in* $L^2(S^{n-1},d\mu)$. *It is reducible if* $s=0,-1,-2,\ldots$ *and cyclic otherwise. It is unitary and irreducible if and only if* $\mathrm{Re}\ s=(n-1)/2$.

In the following, we will always set $s=(n-1)/2$ and write simply $U\equiv U^{(n-1)/2}$.

As already mentioned, we are only interested in the action of dilations and motions. We thus restrict ourselves to the corresponding homogeneous space $X\simeq\mathrm{SO}(n)\cdot A\equiv\mathrm{SO}_0(n,1)/N$ using a suitable section σ: $X\to\mathrm{SO}_0(n,1)$ in the principal fiber bundle defined by the Iwasawa decomposition. Thus we will concentrate on the reduced expression

$$[U(\sigma(x))f](\omega)=\lambda(\sigma(x),\omega)^{(n-1)/2}f(\sigma(x)^{-1}\omega). \tag{4.10}$$

We write points of the space X as pairs $x\equiv(\gamma,a)$, with $\gamma\in\mathrm{SO}(n)$ and $a\in A\simeq\mathrm{SO}(1,1)$, and choose the natural (Iwasawa) section

$$\sigma_I(\gamma,a)=\gamma a. \tag{4.11}$$

We have already computed the action of dilations in (3.14) and, by $\mathrm{SO}(n)$ invariance, it is easily seen that, with $\omega=(\theta_1,\theta_2,\ldots,\theta_{n-1})$,[35]

$$\lambda(\sigma_I(\gamma,a),\omega)\equiv\lambda(a,\theta_{n-1})=\frac{4a^2}{[(a^2-1)\cdot\cos\ \theta_{n-1}+(a^2+1)]^2}. \tag{4.12}$$

In addition, the choice of the section (4.11) implies

$$U(\sigma_I(\gamma,a))=U(\gamma a)=U(\gamma)U(a),$$

so that the representation (4.10) factorizes as

$$[U(\sigma_I(\gamma,a))f](\omega)=(U_{qr}(\gamma)\circ D^a f)(\omega), \tag{4.13}$$

where $U_{qr}(\gamma),\gamma\in\mathrm{SO}(n)$ is the quasi-regular representation (4.1) of $\mathrm{SO}(n)$ in $L^2(S^{n-1},d\mu)$, and $D^a,a\in\mathbb{R}^+_*$ is a pure dilation, that is

$$(D^a f)(\omega)=\lambda(a,\theta_{n-1})^{(n-1)/4}f(\omega_{1/a}),$$

where we have used the notation of Eqs. (3.14).

B. Lorentz coherent states as wavelets on the sphere S^{n-1}

We are now ready to build a system of coherent states for the Lorentz group, indexed by points of the homogeneous space $X=\mathrm{SO}_0(n,1)/N$. Since N is not the isotropy subgroup of a particular vector in the representation space $L^2(S^{n-1},d\mu)$, the resulting coherent states are not of the Gilmore–Perelomov type.[24]

Following the general approach outlined in Sec. II, the coherent states associated to the representation U are defined, as in (2.4), by the elements of the orbit of η under $\mathrm{SO}_0(n,1)$:

$$\eta_{\sigma(x)}(\omega)=[U(\sigma(x))\,\eta](\omega), \quad x\in X, \tag{4.14}$$

provided $\eta\in L^2(S^{n-1},d\mu)$ is admissible mod(N,σ).

The result is given by the following theorem.

Theorem 4.2: *The representation U of $\mathrm{SO}_0(n,1)$ is square integrable modulo (N,σ_I), that is, there exists a nonzero admissible vector $\eta\in L^2(S^{n-1},d\mu)$, which means that*

$$\frac{1}{d_l}\sum_{M\in I_l^n}\int\frac{da}{a^n}|\hat{\eta}_a(l,M)|^2<\infty, \quad \textit{uniformly in } l, \tag{4.15}$$

where

$$\eta_a(\omega)\equiv(D^a\eta)(\omega)=\lambda(a,\theta_{n-1})^{(n-1)/4}\eta(\omega_{1/a}). \tag{4.16}$$

Proof: Square integrability means there exists a vector $\eta\in L^2(S^{n-1},d\mu)$ such that

$$I=\int_0^\infty\frac{da}{a^n}\int_{\mathrm{SO}(n)}d\mu_n(\gamma)|\langle U(\sigma(\gamma,a))\,\eta|\phi\rangle|^2<\infty, \forall\,\phi\in L^2(S^{n-1},d\mu). \tag{4.17}$$

Using the factorization (4.13), this reduces to

$$I=\int\frac{da}{a^n}\int d\mu_n(\gamma)|\langle U_{qr}(\gamma)\,\eta_a|\phi\rangle|^2<\infty.$$

Expanding both terms in the scalar product in a Fourier series on the sphere and using orthogonality relations for hyperspherical harmonics, we get

$$\begin{aligned}\langle U_{qr}(\gamma)\,\eta_a|\phi\rangle&=\sum_{l,l'=0}^{\infty}\sum_{M,M'\in I_l^n}\overline{[U_{qr}(\gamma)\,\eta_a](l,M)}\,\hat{\phi}(l',M')\int_{S^{n-1}}d\mu(\omega)\overline{Y_l^M(\omega)}Y_{l'}^{M'}(\omega)\\&=\sum_{l=0}^{\infty}\sum_{M\in I_l^n}\overline{\widehat{[U_{qr}(\gamma)\,\eta_a]}(l,M)}\,\hat{\phi}(l,M).\end{aligned}$$

The addition theorem (4.5) gives

$$\widehat{[U_{qr}(\gamma)\,\eta_a]}(l,M)=\sum_{Q\in I_l^n}\mathcal{D}^l_{MQ}(\gamma)\,\hat{\eta}_a(l,Q),$$

and (4.17) becomes

$$I=\int_0^\infty \frac{da}{a^n}\Bigg\{ \sum_{\text{all indices}} \overline{\widehat{\eta_a}(l,Q)}\widehat{\eta_a}(l',Q')\overline{\hat{\phi}(l,M)\hat{\phi}(l',M')} \times \int_{SO(n)} d\mu_n(\gamma)\overline{\mathcal{D}^l_{MQ}(\gamma)}\mathcal{D}^{l'}_{M'Q'}(\gamma)\Bigg\}.$$

With the orthogonality relations for Wigner functions,

$$\int_{SO(n)} d\mu_n(\gamma)\overline{\mathcal{D}^l_{MQ}(\gamma)}\mathcal{D}^{l'}_{M'Q'}(\gamma)=\frac{1}{d_l}\delta_{ll'}\delta_{MM'}\delta_{QQ'},$$

this expression simplifies into

$$\sum_{l=0}^\infty \sum_{M\in I_l^n} |\hat{\phi}(l,M)|^2 \sum_{M'\in I_l^n}\frac{1}{d_l}\int_0^\infty \frac{da}{a^n}|\widehat{\eta_a}(l,M')|^2.$$

Putting

$$S_l=\sum_{M\in I_l^n}|\hat{\phi}(l,M)|^2,$$

$$G_l=\frac{1}{d_l}\sum_{M'\in I_l^n}\int_0^\infty \frac{da}{a^n}|\widehat{\eta_a}(l,M')|^2,$$

we see that $(S_l)\in l^1$, since $\Sigma_{l=0}^\infty |S_l|=\|\phi\|^2$. Then the admissibility condition (4.17) becomes

$$\sum_{l=0}^\infty S_l G_l<\infty, \quad \forall S_l\in l^1.$$

By Köthe duality,[36] this series converges absolutely if and only if $G_l\in l^\infty$. In conclusion, η is admissible if and only if

$$\frac{1}{d_l}\sum_{M\in I_l^n}\int \frac{da}{a^n}|\widehat{\eta_a}(l,M)|^2<c, \quad \text{uniformly in } l,$$

and there are obviously many functions η (in fact a dense set) that satisfy this condition. ■

Notice that, once η is admissible, (4.15) may be written as

$$I=\int_X d\nu(x)|\langle U(\sigma_I(x))\,\eta|\phi\rangle|^2=\int_X d\nu(x)|\langle \eta_{\sigma_I(x)}|\phi\rangle|^2\leqslant c\|\phi\|^2. \tag{4.18}$$

This means that the family $\{\eta_{\sigma_I(x)}, x\in X\}$ is a continuous family of CS, but not necessarily a frame, in the terminology of Sec. II and Ref. 16. For $n=3$ we have shown in Ref. 15 that the family indeed is a frame. The proof presumably extends to higher dimensions, but this remains to be seen. The next step is whether that frame is tight. For this, we would need an equality in (4.18), with $c=c(\eta)$ the analog of the admissibility constant of the vector η. Equivalently, we would need $G_l=G(\eta)$, independently of l, but we have been unable to prove it, even for $n=3$. In fact, we conjecture this is not true, but becomes so only in the Euclidean limit. In that limit indeed, the family $\{\eta_{\sigma_I(x)}, x\in X\}$ converges to the usual tight frame of the flat two-dimensional (2-D) CWT, as shown in Sec. V below.

Theorem 4.2 yields the basic ingredient for writing the CWT on S^{n-1}. Given an admissible vector (mother wavelet) $\psi\in L^2(S^{n-1},d\mu)$, our wavelets on the sphere are the functions $\psi_{\gamma,a}=U(\sigma_I(\gamma,a))\psi$, and the CWT reads, with $U_S(\gamma,a)\equiv U(\sigma_I(\gamma,a))$:

$$S(\gamma,a)=\langle U(\sigma_I(\gamma,a))\psi|s\rangle=\int_{S^{n-1}}d\mu(\omega)\overline{[U_S(\gamma,a)\psi](\omega)}s(\omega)=\int_{S^{n-1}}d\mu(\omega)\overline{\psi_a(\gamma^{-1}\omega)}s(\omega). \tag{4.19}$$

Now the condition (4.15) is necessary and sufficient for the admissibility of η, but it is somewhat complicated to use in practice, since it requires the evaluation of nontrivial Fourier coefficients. Instead, we have derived in our previous paper a simpler, although only necessary, condition.[15] The same condition is valid in the general case.

Proposition 4.3: An admissible vector $\eta\in L^2(S^{n-1},d\mu)$ *necessarily satisfies the condition*

$$\int_{S^{n-1}}d\mu(\omega)\frac{\eta(\omega)}{(1+\cos\ \theta_{n-1})^{(n-1)/2}}=0,\quad \omega=(\theta_1,\theta_2,\ldots,\theta_{n-1}), \tag{4.20}$$

where $d\mu(\omega)$ *is the invariant measure on* S^{n-1} *given in (3.4).*

Proof: The proof is entirely similar to that given for $n=3$ in Ref. 15, so we give only a brief sketch. In order to compute

$$\int_0^\infty\frac{da}{a^n}|\langle Y_l^M|D^a\eta\rangle|^2, \tag{4.21}$$

we choose for η an axisymmetric function, with support bounded away from the South Pole, that is

$$\eta(\omega)\equiv\eta(\theta_{n-1})\ \text{and}\ \eta(\theta_{n-1})=0\ \text{if}\ \theta_{n-1}>\tilde{\theta},\quad \text{for some}\ \tilde{\theta}<\pi.$$

Hence only the $M=0$ term contributes in $\langle Y_l^M|D^a\eta\rangle$. Then we split the integration range in (4.21) into three parts: $[0,\infty)=[0,\epsilon]\cup[\epsilon,1/\epsilon]\cup[1/\epsilon,\infty)$. In the first term, we make the change of variables $\theta'=(\theta_{n-1})_{1/a}$ and use the relation coming from the cocycle equation (4.7), namely

$$\lambda(a^{-1},\theta')\lambda(a,\theta'_a)=1.$$

Then, for $\epsilon\ll 1$, a is very small, since $0\leqslant a\leqslant\epsilon$, and one has

$$\theta'_a\simeq 0\ \text{and}\ \lambda(a^{-1},\theta')\simeq\frac{4a^2}{(1+\cos\ \theta')^2}.$$

Notice that, for large l, one has $Y_l^0(0)\sim l^{-(n-2)/2}$, which exactly compensates the asymptotic behavior $d_l\sim l^{n-2}$ for $l\gg 1$ in (4.15), so that the only divergence can possibly come from the integral over a. But, counting the powers of a from (4.16), we see that a logarithmic divergence remains in the integral over small scales ($a\leqslant\epsilon$), unless we impose the condition (4.20) (of course, the result must be independent of the support of η, so that we have to integrate over the full sphere S^{n-1}).

The second term poses no problem, since the integrand is bounded and continuous on $[\epsilon,1/\epsilon]$. Finally the last term also converges, since the only large scale divergence will never be reached because of the support property of η.

Now, if we drop the restriction on the support of η, the condition (4.20) is *a fortiori* necessary, which proves the statement. ■

This necessary condition is the exact equivalent of the usual necessary condition for Euclidean wavelets, $\int d^{n-1}x\,\psi(x)=0$. Furthermore, it reduces to the latter in the Euclidean limit, as we shall see in Sec. V. The interesting point is that (4.20) is a zero mean condition, as in the flat case. As such it will play the same role, namely it ensures that the CWT on S^{n-1} given in (4.19) acts as a local filter, a property crucial for applications. Thus our spherical CWT will have a comparable behavior. One should notice also that the poles do not play any particular role in this CWT, since the sphere S^{n-1} is a homogeneous space under SO(n): all the points of S^{n-1} are really equivalent, despite the appearance to the contrary given by (4.20). Practical examples given in Ref. 15 illustrate this statement.

One may also wonder what should be added to condition (4.20) to make it also sufficient. By analogy with the limiting flat case,[21] we expect that a slightly faster vanishing at the South Pole will do, but this remains to be proven.

A nice outcome of Proposition 4.3 is that it yields an easy way of building spherical wavelets. Indeed, given a smooth function $\phi \in L^2(S^{n-1}, d\mu)$, we define

$$\eta_\phi(\omega) = \phi(\omega) - \frac{1}{\alpha}(D^\alpha \phi)(\omega) \quad (\alpha > 1).$$

Then it is easily checked that η_ϕ satisfies the zero mean condition (4.20), that is, it is a spherical wavelet, similar to the ''difference wavelets'' commonly used in vision in the case $n=3$ (the DOG wavelet, for instance).[37] We have shown in Ref. 15 a Gaussian spherical wavelet of that kind, which is the inverse stereographic projection of an ordinary Gaussian in the tangent plane.

V. THE EUCLIDEAN LIMIT

According to Holschneider,[14] a good wavelet transform on the sphere should satisfy a geometrical constraint expressing its asymptotic Euclidean behavior. This means that, the sphere being locally flat, the associated wavelet transform should match the usual CWT in the plane at small scales or, what amounts to the same, for large values of the radius of curvature. In this section, we will give a precise mathematical meaning to this statement using the technique of group contractions. The main definitions concerning the latter are collected in the Appendix.

A. Contracting the Lorentz group and its homogeneous spaces

In the first step, we reformulate the theory described so far on a sphere of radius R and let $R \to \infty$. In this limit, the Lorentz group $\mathrm{SO}_0(n,1)$ is contracted along its minimal parabolic subgroup $P = \mathrm{SO}(n) \cdot \mathbb{R}_*^+ \cdot N$ into a semidirect product:

$$G_1 = \mathrm{SO}_0(n,1) \overset{R\to\infty}{\to} G_2 = \mathbb{R}^{n-1} \rtimes \mathrm{SIM}(n-1),$$

where $\mathrm{SIM}(n-1) = \mathbb{R}^{n-1} \rtimes (\mathbb{R}_*^+ \times \mathrm{SO}(n-1))$ is the similitude group of $\mathbb{R}^{n-1}$, that is, the invariance group of the Euclidean CWT. The detailed calculation runs as follows.

1. At the level of Lie algebras

We start with the Lorentz Lie algebra $\mathfrak{g} = \mathfrak{so}(n,1)$ and write its generators as X_{ij}, $0 \leqslant i < j \leqslant n$, with commutation relations

$$[X_{ij}, X_{mn}] = g_{im}X_{jn} - g_{jm}X_{in} - g_{in}X_{jm} + g_{jn}X_{im}, \tag{5.1}$$

where g_{ij} is the metric tensor $\mathrm{diag}(1,-1,\ldots,-1)$. We also write $X_i = X_{0i} - X_{in}$, $i = 1,\ldots,n-1$.

For this Lie algebra, the Cartan and the Iwasawa decomposition read, respectively (vector space direct sums):

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{l} \tag{5.2}$$

$$= \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}, \tag{5.3}$$

where $\mathfrak{k} = \mathfrak{so}(n)$ is the maximal compact subalgebra. Let $\mathfrak{p}$ be the minimal parabolic subalgebra, with the decomposition

$$\mathfrak{p} = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n},$$

where $\mathfrak{m} = \mathfrak{so}(n-1)$. In terms of the generators X_{ij} of $\mathfrak{so}(n,1)$, we have the following identifications:

$$\begin{aligned} &\mathfrak{k} = \mathrm{span}\{X_{ij},\ 1 \leqslant i < j \leqslant n\} \sim \mathfrak{so}(n), \qquad &&\mathfrak{l} = \mathrm{span}\{X_{0i}, i = 1,\ldots,n\}, \\ &\mathfrak{n} = \mathrm{span}\{X_i, i = 1,\ldots,n-1\} \sim \mathbb{R}^{n-1}, \quad &&\mathfrak{a} = \mathrm{span}\{X_{0n}\} \sim \mathbb{R}. \end{aligned} \tag{5.4}$$

One can check immediately that $\mathfrak{p}$ is isomorphic to $\mathfrak{sim}(n-1)=\mathfrak{so}(n-1)\oplus\mathbb{R}\oplus\mathbb{R}^{n-1}$, the Lie algebra of the similitude group SIM$(n-1)$.

Then we introduce the following contraction scheme

$$X_{ij}\mapsto R^{-1}X_{ij},\quad 1\leqslant i<j\leqslant n-1,$$

$$\phi_R: X_{in}\mapsto X_{in},\quad i=1,\ldots,n-1,$$

$$X_{0i}\mapsto X_{0i},\quad i=1,\ldots,n.$$

This contraction maps $\mathfrak{k}$ to $\mathfrak{e}(n-1)=\mathbb{R}^{n-1}\oplus\mathfrak{so}(n-1)$, the $(n-1)$-dimensional Euclidean algebra and preserves the parabolic subalgebra $\mathfrak{p}$. The contracted algebra is

$$\mathfrak{g}_2=\mathbb{R}^{n-1}\oplus\mathfrak{so}(n-1)\oplus\mathbb{R}\oplus\mathbb{R}^{n-1}. \tag{5.5}$$

Thus $\mathfrak{g}_2$ contains two subalgebras isomorphic to $\mathfrak{sim}(n-1)$. One is $\mathfrak{p}$, which is the sum of the last three terms in (5.5), and the other one corresponds to the sum of the first three terms. We choose the first possibility. In the notation of the Appendix, this gives $\mathfrak{s}=\mathfrak{p}$, which is preserved under the contraction, and $\mathfrak{v}_c=\text{span}\{X_{in}, i=1,\ldots,n-1\}$. Clearly $\{\mathfrak{s},\mathfrak{v}_c\}$ satisfy the commutation relations (A3).

Of course, the decomposition (A3) is not unique. One could have chosen instead for $\mathfrak{s}'$ the other $\mathfrak{sim}(n-1)$ subalgebra, and for $\mathfrak{v}'_c$ the complement, and obtained an isomorphic structure for $\mathfrak{g}_2$, except that $\mathfrak{s}'$ is no more preserved by contraction.

2. At the level of Lie groups

As a consequence, the group contraction procedure maps $G_1=\text{SO}_0(n,1)$ onto the group G_2 associated with $\mathfrak{g}_2$, namely the semidirect product

$$G_2\equiv V_c\rtimes S\equiv\mathbb{R}^{n-1}\rtimes\text{SIM}(n-1).$$

Hence the quotient G_2/V_c is isomorphic to the similitude group SIM$(n-1)$ of $\mathbb{R}^{n-1}$. The contraction preserves the minimal parabolic subgroup $P=MAN$, which is isomorphic to SIM$(n-1)$.

The contracting map $\Pi_R: G_2\to G_1$ is given in (A4). In order to compute it explicitly, we write $v=\Sigma_{i=1}^{n-1}v_iX_{in}\in\mathfrak{v}_c$ and denote by $s=(m(\psi),t,\xi)$ the elements of P, with $m(\psi)\in\text{SO}(n-1)$, $t\in\mathbb{R}$, $\xi\in N\sim\mathbb{R}^{n-1}$. Finally, we rewrite the Euler decomposition (3.8) as $\gamma=m(\psi)u_{n-1}(\theta_{n-1})u^{(n-2)}(\beta)$, with $\beta=(\theta_1,\cdots,\theta_{n-2})\in S^{n-2}$ and $u^{(n-2)}(\beta)=u_{n-2}(\theta_{n-2})\ldots u_1(\theta_1)$. Introducing polar coordinates (r,α) in $\mathbb{R}^{n-1}$, with $\alpha\in S^{n-2}$, we get $v=re_\alpha=ru^{(n-2)}(\alpha)e_0$, where e_α is the unit vector in $\mathbb{R}^{n-1}$ pointing to $\alpha\in S^{n-2}$, and so e_0 corresponds to the North Pole. Denoting by a hat the $(n+1)\times(n+1)$ matrix representation of the respective elements, as in (3.6) and (3.7), and expanding the matrix elements in powers of R^{-1}, the exponential factor in (A4) becomes

$$(\exp_{G_1}R^{-1}v)\hat{}=\begin{pmatrix}1 & 0^t & 0\\ 0 & \mathbb{I}_{n-1} & \dfrac{v}{R}\\ 0 & -\dfrac{v^t}{R} & 1\end{pmatrix}+\mathcal{O}\left(\frac{1}{R^3}\right).$$

Similarly $s=(m(\psi),t,\xi)\in P=MAN$ reads

$$\hat{s}=\hat{m}(\psi)\cdot\hat{a}(t)\cdot\hat{n}(\xi),\quad\text{with}\quad\hat{m}(\psi)=\begin{pmatrix}1 & 0 & 0\\ 0 & m(\psi) & 0\\ 0 & 0 & 1\end{pmatrix}.$$

Combining the two factors, according to (A4), we obtain, in the limit $r\ll R$:

$$\Pi_R(re_\alpha,(m(\psi),t,\xi))=\hat{m}(\psi)\cdot\hat{u}_{n-1}\left(\frac{r}{R}\right)\cdot\hat{u}^{(n-2)}(\alpha)\cdot\hat{a}(t)\cdot\hat{n}(\xi)\quad(r/R\ll 1). \tag{5.6}$$

This calculation shows that, in geometrical terms, the contraction amounts indeed do let the radius of the sphere go to infinity.

3. At the level of homogeneous spaces

The next step is to transfer the contraction process to the relevant homogeneous spaces. On one hand, the manifolds $S^{n-1}=G_1/MAN$ and $\mathbb{R}^{n-1}=G_2/MAN$, that carry the respective CWT, are related through contraction (notice that we use here the isomorphism between the two forms of the contracted group G_2, as discussed above). On the other hand, since the abelian subgroup N is preserved under the contraction, the parameter space $X=\mathrm{SO}_0(n,1)/N$ of the spherical CWT goes into that of the Euclidean CWT, namely $\mathrm{SIM}(n-1)=G_2/N$. Thus the geometrical picture is coherent.

Next, we have to formulate the contraction directly on the parameter spaces, that is, we have to restrict the contraction map (A4) to the respective homogeneous spaces $\mathrm{SIM}(n-1)$ and X. To achieve this, we introduce a section $\tilde{\sigma}:\mathrm{SIM}(n-1)\to N\rtimes\mathrm{SIM}(n-1)$ by the relation:

$$\tilde{\sigma}:(b,(a,m(\psi)))\mapsto(n(b),(b,(a,m(\psi)))), \tag{5.7}$$

$$b\in\mathbb{R}^{n-1},\quad a\in A,\quad m(\psi)\in\mathrm{SO}(n-1),\quad n(b)\in N\sim\mathbb{R}^{n-1}.$$

Combining this with the canonical projection of the Iwasawa bundle,

$$\Gamma:KAN\to X\simeq KA,$$

we may define the restricted contraction map $\tilde{\Pi}_R:\mathrm{SIM}(n-1)\to X$ by

$$\tilde{\Pi}_R(g)=\Gamma(\Pi_R(\tilde{\sigma}(g))),\quad g\in\mathrm{SIM}(n-1), \tag{5.8}$$

where $\Pi_R:G_2\to G_1$ is the contraction map (A4). Altogether we have the following diagram:

$$\begin{array}{ccc} G_2=\mathbb{R}^{n-1}\rtimes\mathrm{SIM}(n-1) & \overset{\Pi_R}{\to} & G_1=\mathrm{SO}_0(n,1) \\ \uparrow\tilde{\sigma} & & \downarrow\Gamma \\ G_2/N\simeq\mathrm{SIM}(n-1) & \overset{\tilde{\Pi}_R}{\to} & \mathrm{SO}_0(n,1)/N=X. \end{array} \tag{5.9}$$

Notice that the parameter space of the spherical CWT is $X=G_1/N\simeq\mathrm{SO}(n)\cdot A$, which is not a group (and this forced us to use the general formalism of Ref. 16). After contraction, we get $G_2/N\simeq\mathrm{SIM}(n-1)$, so that the missing group structure is restored by the contraction.

B. The Euclidean limit of the spherical CWT

Now we are ready to show that the CWT on the sphere S^{n-1} converges to the usual CWT on $\mathbb{R}^{n-1}$ in the geometrical limit of large radius. For that purpose, we have to prove that the associated series of square integrable representations of $\mathrm{SO}_0(n,1)$ modulo N contracts to the usual wavelet representation of $\mathrm{SIM}(n-1)$,[27,28] and then that the admissibility condition on the sphere carries over to that in the flat case. Here we use the precise definition of contraction introduced by Dooley.[38,39] Essentially the same result was obtained earlier by Mickelsson and Niederle.[40]

Let $\mathcal{H}_R=L^2(S_R^{n-1},d\mu_R)$ be the Hilbert space of square integrable functions on a sphere of radius R, with $d\mu_R(\omega)=R^2\,d\mu(\omega)$ and $\mathcal{H}=L^2(\mathbb{R}^{n-1},d^{n-1}x)$. In view of the geometry of the problem, it is clear that the contracting map $I_R:\mathcal{H}_R\to\mathcal{H}$ amounts essentially to a stereographic projection. As before, we write points on the sphere S^{n-1} as pairs (θ_{n-1},β) and introduce polar coordinates (r,β) in $\mathbb{R}^{n-1}$, with $\beta\in S^{n-2}$. Then, with a suitable convergence factor, we obtain an isometry $I_R:\mathcal{H}_R\to\mathcal{H}$, namely

$$(I_R f)(r,\beta)=\left(\frac{4R^2}{4R^2+r^2}\right)^{(n-1)/2} f\left(2\ \arctan\frac{r}{2R},\beta\right). \tag{5.10}$$

One checks indeed that $\|I_R f\|_{\mathcal{H}}=\|f\|_{\mathcal{H}_R}$, so that injectivity of I_R is ensured. The inverse map reads

$$(I_R^{-1} f)(\theta_{n-1},\beta)=\left(\frac{2}{1+\cos\theta_{n-1}}\right)^{(n-1)/2} f\left(2R\ \tan\frac{\theta_{n-1}}{2},\beta\right), \tag{5.11}$$

and a closer inspection shows that it is also an isometry. Thus I_R is unitary.

For each R, we choose $\mathcal{D}_R=\mathcal{D}=\mathcal{C}_0(\mathbb{R}^{n-1})$, the space of continuous functions of compact support, which is dense in $\mathcal{H}$. Now let U be usual wavelet representation of SIM$(n-1)$ in $\mathcal{H}$ and U_R the representation (4.10) of $\mathrm{SO}_0(n,1)$ realized in $\mathcal{H}_R$.

Theorem 5.4 (Euclidean limit): *The representation U of* SIM$(n-1)$ *is a contraction of the family of representations U_R of* $\mathrm{SO}_0(n,1)$ *as* $R\to\infty$.

Proof: For every $g\in\mathrm{SIM}(n-1)$, we have to verify the strong limit

$$\lim_{R\to\infty}\|I_R U_R(\tilde{\Pi}_R(g))I_R^{-1}\phi-U(g)\phi\|_{\mathcal{H}}=0. \tag{5.12}$$

We look first at pointwise convergence. The subgroups A and $M=\mathrm{SO}(n-1)$ are preserved by contraction, and it is easily seen that the associated operators commute with I_R. Thus it is enough to consider elements $g=(re_\alpha,(t,m(\psi)))\in\mathrm{SIM}(n-1)=\mathbb{R}^{n-1}\rtimes(\mathbb{R}\times\mathrm{SO}(n-1))$ of the form $g=(re_0,(0,\mathbb{I}))$, that is, $\alpha=0,t=0,\psi=0$. Then, according to (5.6) and (5.8), one has

$$\tilde{\Pi}_R(g)=\hat{u}\left(\frac{r}{R}\right),$$

from which one readily checks that, pointwise,

$$\lim_{R\to\infty}|I_R U_R(\tilde{\Pi}_R(g))I_R^{-1}\phi-U(g)\phi|=0.$$

Next, since I_R^{-1} is a unitary operator, we can compute the strong limit in $\mathcal{H}_R$. Define, for $\phi\in\mathcal{D}, C=\max_{x\in\mathbb{R}^{n-1}}\phi(x)$. Then, using the relation (5.11), we obtain the following bound, for all $\phi\in\mathcal{D}$, uniformly in $R\geqslant 1$:

$$\begin{aligned}|U_R(\tilde{\Pi}_R(g))I_R^{-1}\phi-I_R^{-1}U(g)\phi|&<|U_R(\tilde{\Pi}_R(g))I_R^{-1}\phi|+|I_R^{-1}U(g)\phi|\\&<2C\cdot\max\left(1+\tan^2\frac{\theta}{2}\right)^{(n-1)/2},\end{aligned}$$

where the maximum is taken over all θ such that $(2R\ \tan(\theta/2),\beta)\in\mathrm{supp}\ \phi$. Finally Lebesgue's dominated convergence theorem gives the result. ■

This theorem yields the expected result that local wavelet analysis on the sphere, as defined here, is equivalent to local wavelet analysis in flat space. Indeed the whole structure on the sphere S_R^{n-1} goes into the corresponding one in $\mathbb{R}^{n-1}$ as $R\to\infty$. In particular, since $U_R\to U$, the corresponding matrix elements also converge to one another, and therefore the square integrability condition (4.15) converges into the corresponding one for the CWT in $\mathbb{R}^{n-1}$,

$$\int_{\mathbb{R}^{n-1}} d^{n-1}k\frac{|\hat{\psi}(k)|^2}{|k|^{n-1}}<\infty.$$

Admissible wavelets on S^{n-1} converge to admissible wavelets on $\mathbb{R}^{n-1}$. Also, because the renormalizing factor in (5.11) is exactly the one that links the two invariant measures under stereographic projection, it follows that the necessary condition (4.20) also goes into the corresponding necessary condition for wavelets in the plane, $\int d^{n-1}x\,\psi(x)=0$.

VI. EXTENSION TO OTHER MANIFOLDS

Although the spheres S^{n-1}, and in particular S^2, are the manifolds on which a wavelet transform is most desirable for applications, the mathematical analysis made so far invites us to consider other manifolds with similar geometrical properties, and first other Riemannian symmetric spaces of constant curvature. In order to understand better the problems involved in the extension of the preceding construction, we will look first at the case $n=3$.

A. The two-dimensional hyperbolic space

Let us go back to the sphere S^2. In this case, the stereographic projection from the South Pole to the plane tangent at the North Pole is a bijection $\Phi: S^2 \to \mathbb{C} \cup \{\infty\}$, given by

$$\Phi(\omega) \equiv \zeta = 2 \tan \frac{\theta}{2} e^{i\varphi}, \ \omega \equiv (\theta, \varphi); \ 0 \leqslant \theta \leqslant \pi, \ 0 \leqslant \varphi < 2\pi$$

(here S^2 is taken as the Riemann sphere and the tangent plane at the North Pole as the complex plane $\mathbb{C}$; the extra factor 2 is due to our choice of the tangent plane rather than the equatorial one[30]). Then a dilation $\zeta \mapsto a\zeta$ in $\mathbb{C}$ corresponds to the mapping $\theta \mapsto \theta_a$ on S^2, where $\tan(\theta_a/2) = a \tan(\theta/2)$ [compare Eq. (3.14)]. The conformal factor linking the two metrics on S^2 and in the plane is

$$h(r) = \frac{2}{1+(r/2)^2} = 1 + \cos \theta, \quad r = |\zeta|.$$

Now $S^2 = \mathrm{SO}(3)/\mathrm{SO}(2)$ is a compact Riemannian symmetric space of constant curvature $\kappa = 1$, and it has a noncompact dual, $H^2 = \mathrm{SO}_0(2,1)/\mathrm{SO}(2)$, of constant curvature $\kappa = -1$.[29] H^2 is a two-sheeted hyperboloid, of equation $-x_1^2 - x_2^2 + x_3^2 = 1$, symmetric around the x_3-axis. Duality corresponds to the fact that $\mathrm{SO}(3)$ and $\mathrm{SO}_0(2,1)$ are the two real forms of the complex group $\mathrm{SO}(3)^{\mathbb{C}} \sim \mathrm{SL}(2,\mathbb{C})$.

As for the sphere, we can perform the stereographic projection Φ from the South Pole ($x_1 = x_2 = 0, x_3 = -1$) onto the equatorial plane $x_3 = 0$. Then Φ maps the upper sheet H_+^2 onto the interior $\mathcal{D}_+$ of the unit disk, and the lower sheet H_-^2 onto the exterior. The domain $\mathcal{D}_+$ is called the Lobachewskian disk.[41] Introducing hyperbolic coordinates on H_+^2,

$$\begin{aligned} x_1 &= \sinh \chi \cos \varphi, \\ x_2 &= \sinh \chi \sin \varphi \quad (\chi \in \mathbb{R}, 0 \leqslant \varphi < 2\pi), \\ x_3 &= \cosh \chi \end{aligned} \tag{6.1}$$

which give the metric $ds^2 = d\chi^2 + \sinh^2 \chi \, d\varphi^2$, one obtains for a point $\zeta \in \mathcal{D}_+$ (see Fig. 1):

$$\zeta = \tanh \frac{\chi}{2} e^{i\varphi}.$$

The Lobachewskian metric on $\mathcal{D}_+$ is ($\zeta = u + iv$):

$$dl^2 = \frac{4}{[1-(u^2+v^2)]^2}(du^2 + dv^2) = d\chi^2 + \sinh^2 \chi \, d\varphi^2.$$

This shows that $\mathcal{D}_+$ is conformally equivalent to H_+^2, with conformal factor $h(r) = 2/(1-r^2)$, where $r^2 = (u^2+v^2)$. For both manifolds, the isometry group is $\mathrm{SO}_0(2,1)$.

What about dilations? One way to identify them is to consider the Poincaré half-plane realization of the Lobachewski space,[41] $H_+ = \{z \in \mathbb{C}, \mathrm{Im}\, z > 0\}$, which corresponds to $\mathcal{D}_+$ via the Cayley transform, also a conformal map:

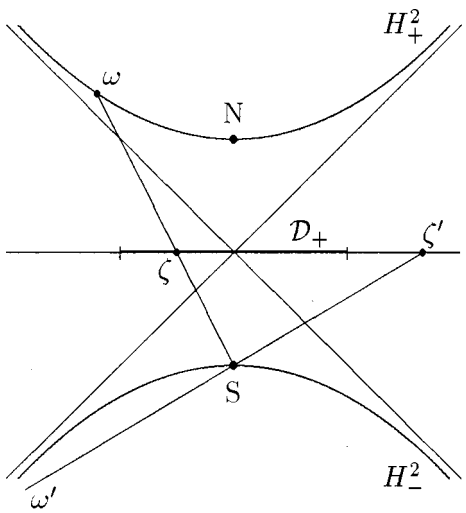

FIG. 1. Stereographic projection from the upper sheet H_+^2 of the hyperboloid onto the interior $\mathcal{D}_+$ of the unit disk.

$$H_+ \ni z \mapsto \zeta = -\frac{z-i}{z+i} \in \mathcal{D}_+ .$$

If one introduces in H_+ geodesic polar coordinates $z=(\chi,\varphi)$, it turns out the curves $\varphi=\text{const}$, which are geodesics, are half-circles centered on the real axis, thus orthogonal to it. In particular, χ is the Poincaré distance between z and i, measured along the geodesic that links the two points. In these coordinates, the Cayley transform reads

$$z=(\chi,\varphi) \mapsto \zeta = \tanh\frac{\chi}{2} e^{i\varphi} \in \mathcal{D}_+ .$$

In particular the point $z=i \in H_+$ goes into $\zeta=0$, and all the geodesics through i go into the diameters of $\mathcal{D}_+$. As for isometries, those of H_+ are matrices from $\mathrm{SL}(2,\mathbb{R})$, acting by homographic transformation, and they correspond to the similar action of SU(1,1) on $\mathcal{D}_+$. In the two cases, there is a simple transitive action of $\mathrm{SL}(2,\mathbb{R})/Z_2$, resp. $\mathrm{SU}(1,1)/Z_2$, both isomorphic to $\mathrm{SO}_0(2,1)$, as expected. In $\mathcal{D}_+$, a dilation should be radial, and leave both $\zeta=0$ and $\zeta=1$ invariant, but there is no such element in SU(1,1). As a way out, we may use as a natural notion of dilation in H_+ the map $\chi \mapsto a\chi, 0<a<\infty$. In $\mathcal{D}_+$, the corresponding map leads to

$$\zeta \mapsto \zeta_a = \tanh\frac{a\chi}{2} e^{i\varphi},$$

which then lifts to H_+^2 by inverse stereographic projection. This map has all the required properties for a dilation, but does not come directly from a linear group action. Thus it can only be used for constructing wavelets on H_+^2 if one puts it by hand, in the same way as Holschneider did for S^2.[14] Of course, it remains to obtain a suitable representation of the resulting set $\mathrm{SO}_0(2,1)\cdot\mathbb{R}_*^+$ in $L^2(H_+^2, d\mu)$, where $d\mu$ is the $\mathrm{SO}_0(2,1)$-invariant measure, and to show that it is square integrable in a suitable sense.

An alternative way of introducing dilations via a group action is based on analytic continuation from the spherical case, that is, using the familiar Weyl trick. We start from the Lie algebra $\mathfrak{so}(3,1)$, with commutation relations (5.1) and Iwasawa decomposition (5.3), for $n=3$. The duality $\mathfrak{so}(3) \mapsto \mathfrak{so}(2,1)$ results from Weyl's unitary trick applied to the compact subalgebra $\mathfrak{k}$:

$$\mathfrak{k} = \mathfrak{m} \oplus \mathfrak{r} \mapsto \mathfrak{k}^* = \mathfrak{m} \oplus i\mathfrak{r},$$

with $\mathfrak{m}=\mathrm{span}\{X_{12}\}$ and $\mathfrak{r}=\mathrm{span}\{X_{13},X_{23}\}$. Define the new generators as $\overline{X_{13}=iX_{13}}$, $\overline{X_{23}=iX_{23}}$. Then do the same for $\mathfrak{n} \mapsto \mathfrak{n}^* = \mathrm{span}\{\overline{X_1,X_2}\}$, $\overline{X_j=iX_j}$. This yields the decomposition

$$\mathfrak{g}^* = \mathfrak{k}^* \oplus \mathfrak{a} \oplus \mathfrak{n}^*, \tag{6.2}$$

and an explicit calculation shows that $\mathfrak{g}^* \sim \mathfrak{so}(1,3)$, that is, the original metric $(+---)$ associated to $\mathfrak{so}(3,1)$ is transformed into $(---+)$. Then $\mathfrak{k}^* = \text{span}\{X_{12}, \overline{X_{13}, X_{23}}\} \sim \mathfrak{so}(1,2)$ and we obtain the pseudo-Iwasawa decomposition

$$G^* = \mathrm{SO}_0(1,3) = \mathrm{SO}_0(1,2) \cdot A \cdot N^*.$$

The Lorentz boosts along H^2_+ are generated by $\overline{X_{13}, X_{23}}$. Hence the subgroup $A \sim \exp tX_{03}$, $t \in \mathbb{R}$, represents dilations, which commute with the subgroup $\mathrm{SO}(2)$ of rotations around the x_3-axis.

The next step is to compute the action of G^* on H^2_+. This may be done, as in the spherical case, using the Euler decomposition (3.8) of $\gamma^* \in \mathrm{SO}_0(1,2)$, corresponding to the hyperbolic coordinates (6.1):

$$\gamma^* = m(\psi) u^*(\chi) u(\varphi), \text{with } m(\psi), u(\varphi) \in \mathrm{SO}(2),\ u^*(\chi) \in \mathrm{SO}_0(1,1).$$

Then, as in Eqs. (3.12) and (3.13), we get:

$$g_0(a)\, \gamma^* = m(\psi) u^*(\chi_a) u(\varphi),$$

where the transformation $\chi \mapsto \chi_a$ may be obtained by a direct calculation.

At this stage, two questions remain to be answered:

(1) Does one have a unitary irreducible representation U of $\mathrm{SO}_0(1,3)$ in $L^2(H^2_+, d\mu)$? As in the spherical case, U should reduce to the quasi-regular representation of $\mathrm{SO}_0(1,2)$ upon restriction to the latter.
(2) Is the representation U, if any, square integrable modulo N^* and a suitable section?

We conjecture that the answer to both questions is positive, which would then yield the CWT on H^2_+. Work in this direction is in progress.

B. Higher dimensional Riemannian symmetric spaces

In higher dimensions, the situation is exactly the same:

$$S^{n-1} = \mathrm{SO}(n)/\mathrm{SO}(n-1) \quad \text{and } H^{n-1} = \mathrm{SO}_0(n-1,1)/\mathrm{SO}(n-1)$$

are dual Riemannian symmetric spaces, with constant curvature $\kappa = \pm 1$, respectively. Again $\mathrm{SO}(n)$ and $\mathrm{SO}_0(n,1)$ are two real forms of the complexified group $\mathrm{SO}(n)^{\mathbb{C}}$.

For the hyperbolic space H^{n-1}, the extension from $n=3$ to arbitrary $n>3$ proceeds exactly as for the spheres, in the way described in the previous section. For instance, stereographic projection $\omega \mapsto \zeta$ from the South Pole maps H^n_+ onto the interior of the unit ball of the equatorial hyperplane, and the two are conformally equivalent when the latter is equipped with the Lobachewskian metric. The reason is that the additional dimensions are compactified, so that nothing essential changes, in particular with respect to square integrability.

The same comment applies to the construction of dilations via the Weyl trick. The metric $(+-\cdots--)$ associated to $\mathfrak{so}(n,1)$ is transformed into $(--\cdots-+)$, and $\mathrm{SO}_0(1,n-1)$ is embedded into $\mathrm{SO}_0(1,n)$. This behavior is illustrated by comparing the present paper with the corresponding one for $n=3$.[15]

In addition, there are now additional noncompact real forms $\mathrm{SO}_0(p,q), p+q=n$, leading to pseudo-Riemannian symmetric spaces $H^{p,q} = \mathrm{SO}_0(p,q)/(\mathrm{SO}(p) \times \mathrm{SO}(q))$. These are homogeneous spaces of higher rank, and few explicit results are available.

How can one extend this machinery to other manifolds? In the case of the sphere S^{n-1}, the conformal group $\mathrm{SO}_0(n,1)$ is larger than the motion group $\mathrm{SO}(n)$, and thus could be used for incorporating dilations. But the sphere is the only compact manifold with that property. In the general case, an alternative way to proceed is to embed the motion group G of the manifold into the complexified group $G^{\mathbb{C}}$ in order to define dilations. This has been done for S^2, in fact, since $\mathrm{SO}(3)^{\mathbb{C}} \sim \mathrm{SL}(2,\mathbb{C})$, that is, the Lorentz group serves both as the conformal group of S^2 and as

complexification of its motion group. This remark may open the way to future extension of wavelets to higher dimensional manifolds. In particular, it could be interesting to consider the case of nontrivial phase spaces. Indeed, the sphere S^2 is the phase space of a rotator, and serves as parameter space for coherent states on SO(3). Other systems exist with ''exotic'' phase spaces, such as a torus $SO(2)\times SO(2)$ and similar manifolds.[42]

Another direction that looks promising is to consider the wavelet problem in the general framework of representation theory of semisimple groups. Let G be a noncompact connected semisimple Lie group, containing subgroups that can be interpreted as motions and dilations of a certain manifold Y. The existence of wavelets associated to G amounts to find an isometry from $L^2(X)$ to $L^2(Y)$, where X is either a subgroup G_0 of G or a quotient $X=G/H$. In the first case, the central tool is a square integrable or discrete series representation of G_0; in the second one, a representation of the relative discrete series of X. At this stage, no general theory is available, but results have been obtained in a number of cases:

(i) Let $G=KAN$ as before, with minimal parabolic subgroup $P=MAN$ and let $\overline{N}=\theta(N)$, where θ is the Cartan involution of G.[29,32] In this case, under specific conditions that involve only $\overline{N}$, a wavelet transform exists for any measurable subset S of MA, with $X=\overline{N}\times S$ and U a principal series representation of G (induced by P), acting in $L^2(\overline{N})$.[43]

(ii) A special case of the previous example is that of $G=SL(n+2,\mathbb{R})$, with $X=\overline{N}\times A_1\simeq H_n\times\mathbb{R}$, where H_n is the Weyl–Heisenberg group and $A_1\subset A$. Here also $Y=\overline{N}\simeq H_n$.[44]

(iii) Other wavelets on H_n have been obtained using closed subgroups of $SU(n+1,1)$, either the affine Weyl–Heisenberg group,[45] or the minimal parabolic subgroup P.[46] More precisely, let $U^{n+1}=\{(z,z_{n+1})\in\mathbb{C}^{n+1}:\text{Im } z_{n+1}>|z|^2\}$ be the Siegel upper half-plane. Then H_n is the boundary of U^{n+1} and P may be considered as the affine automorphism group of U^{n+1}, in the sense that it consists of operations that may be interpreted as translations, dilations and rotations on U^{n+1}. Here again P has a natural unitary representation in $L^2(H_n)$, which decomposes into a direct sum of square integrable representations, each of which thus generates a wavelet transform on H_n (one recovers the case of the affine Weyl–Heisenberg group simply by ignoring the rotations).

(iv) The previous result extends to the group of real symmetric matrices.[47] Let $\mathcal{H}_n=\{z=x+iy\in\mathbb{C}^{n\times n},\ z^t=z,y \text{ real and positive definite}\}$ be the tube domain over the cone of real positive $n\times n$ symmetric matrices, and B its Shilov boundary, that is, the space of $n\times n$ real symmetric matrices. The symplectic group $Sp(n,\mathbb{R})$ acts on $\mathcal{H}_n$, and it contains two automorphism groups of B, called P_0 and P. The elements of P_0 may be identified with translations and dilations, while P contains in addition rotations from $SO(n)$. Both P_0 and P have UIRs on $L^2(B)$, which are square integrable, and B essentially reduces to the nilpotent component in the Iwasawa decomposition of $Sp(n,\mathbb{R})$. Thus we are in a situation very similar to the previous one.

In all those examples, one gets a wavelet transform, either on the Weyl–Heisenberg group H_n, or on the space of real symmetric matrices, starting from various transitive group actions. According to the criteria stated in Sec. I, we ought to ask the question of the Euclidean limit. Formally, in the case where $X=G/K\simeq AN$, one recovers the Euclidean CWT by letting ρ, half the sum of the positive restricted roots of $\mathfrak{g}$, go to 0. However, it remains to be seen whether this limit can be given a rigorous meaning, in particular whether it can be interpreted in terms of group contraction. Should this be true, we would have a whole family of CWTs on nontrivial manifolds.

VII. CONCLUSION

The construction presented here fulfills all the requirements stated in the Introduction for a continuous wavelet transform on the sphere. It is entirely derived from group theory, following the formalism of general coherent states developed in Ref. 16. In addition, the Euclidean limit is valid, with a precise group-theoretical formulation. Thus the formula (4.19) yields a genuine CWT on the sphere, which has none of the defects of the other versions mentioned in the Introduction. Preliminary tests in the case $n=3$ are encouraging, although there remain computational problems, because the formula (4.19) requires a pointwise convolution on the sphere, which is very

time consuming. The construction extends to higher dimensional spheres in a straightforward way, with the same advantages and the same limitations. As for other manifolds, typically Riemannian symmetric spaces of noncompact type, the geometrical picture is similar, but it is not obvious how to define appropriate dilations by a pure group action, through the embedding of the motion group into a larger group, such as the conformal group of the manifold, if it is larger, or the complexified group.

APPENDIX CONTRACTIONS OF LIE GROUPS AND LIE ALGEBRAS

1. Contraction of Lie algebras and Lie groups

We begin by recalling some basic facts concerning the process of contraction, both for Lie algebras and Lie groups. Let $\mathfrak{g}_1=(V,[.,.]_1)$ and $\mathfrak{g}_2=(V,[.,.]_2)$ be two Lie algebras on the same vector space V. We say that $\mathfrak{g}_2$ is a *contraction* of $\mathfrak{g}_1$ if there exists a one parameter family of invertible linear mappings $\phi_R, R\in[1,\infty)$, from V to V such that

$$\lim_{R\to\infty} \phi_R^{-1}[\phi_R X,\phi_R Y]_1=[X,Y]_2, \quad \forall X,Y\in V. \tag{A1}$$

The limit (A1) defines a new Lie algebra structure on V, which is not isomorphic to the original one. In the case of a Inönü–Wigner contraction,[17] a particular subalgebra of $\mathfrak{g}_1$ is conserved throughout the process. More precisely, suppose there exists a subalgebra $\mathfrak{s}$ in $\mathfrak{g}_1$ and a vector subspace $\mathfrak{v}_c$, complement of $\mathfrak{s}$ in $\mathfrak{g}_1$, that is

$$\mathfrak{g}_1=\mathfrak{s}+\mathfrak{v}_c, \tag{A2}$$

such that

$$[\mathfrak{s},\mathfrak{s}]_2\subset\mathfrak{s}, \quad [\mathfrak{v}_c,\mathfrak{v}_c]_2=0, \quad [\mathfrak{s},\mathfrak{v}_c]\subset\mathfrak{v}_c. \tag{A3}$$

Using (A2) we can decompose any $X\in V$ as

$$X=X_{\mathfrak{s}}+X_c, \quad X_{\mathfrak{s}}\in\mathfrak{s}, \quad X_c\in\mathfrak{v}_c,$$

and define the contracting mappings

$$\phi_R(X)=X_{\mathfrak{s}}+\frac{1}{R}X_c.$$

Then applying (A1) does not affect the subalgebra $\mathfrak{s}$. We say in this case that we have a contraction of $\mathfrak{g}_1$ along $\mathfrak{s}$.

The contraction process may be lifted to the corresponding Lie groups.[38,39] Let again $\mathfrak{g}_1$ and $\mathfrak{g}_2$ be two Lie algebras such that $\mathfrak{g}_2$ is a contraction of $\mathfrak{g}_1$. Let G_1 be the simply connected Lie group with Lie algebra $\mathfrak{g}_1$. Let S be the subgroup of G_1 whose Lie algebra is $\mathfrak{s}$ in the decomposition (A2). Defining the semidirect product

$$G_2\equiv V_c\rtimes S, \quad V_c=\exp\,\mathfrak{v}_c\simeq\mathfrak{v}_c,$$

one easily checks that $\mathfrak{g}_2=(V,[.,.]_2)$ is the Lie algebra of G_2. Consider now the family of maps $\Pi_R: G_2\to G_1$ given by

$$\Pi_R:(v,s)\mapsto(\exp_{G_1} R^{-1}v)\cdot s. \tag{A4}$$

They play essentially the same role as the maps ϕ_R of (A1) at the level of the corresponding groups, namely:

$$\lim_{R\to\infty} \Pi_R^{-1}(\Pi_R(g)\overset{1}{\circ}\Pi_R(g'))=g\overset{2}{\circ}g', \forall g,g'\in G_2, \tag{A5}$$

where $\circ_1$, $\circ_2$ denote the product in G_1, G_2, respectively. Indeed one easily checks that $T_e\Pi_R = \phi_R$, where $T_e\Pi_R$ is the derivative of Π_R evaluated at the neutral element of G_1. It is easily seen on (A4) that the subgroup S is preserved during the contraction.

2. Contraction of Lie group representations

Whereas contractions of Lie algebras and Lie groups are relatively ancient and well-known,[17,18] the extension of the procedure to group representations is rather recent.[40] A rigorous version has been given by Dooley,[38,39] which we follow. The additional difficulty here is that the representation space itself varies during the procedure.

Let G_2 be a contraction of G_1, defined by the contraction map $\Pi_R: G_2 \to G_1$ and let U be a representation of G_2 in a Hilbert space $\mathcal{H}$. Suppose that, for each $R \in [1,\infty)$, we have a representation $\{\mathcal{H}_R, U_R\}$ of G_1, a dense subspace $\mathcal{D}_R$ of $\mathcal{H}$ and a linear injective map $I_R: \mathcal{H}_R \to \mathcal{D}_R$. Then one says that the representation U of G_2 is a contraction of the family of representations $\{U_R\}$ of G_1 if there exists a dense subspace $\mathcal{D}$ of $\mathcal{H}$ such that, for all $\phi \in \mathcal{D}$ and $g \in G_2$, one has: (i) for every R large enough, $\phi \in \mathcal{D}_R$ and $U_R(\Pi_R(g)) I_R^{-1}\phi \in I_R^{-1}(\mathcal{D}_R)$, (ii) $\lim_{R\to\infty} \| I_R U_R(\Pi_R g) I_R^{-1}\phi - U(g)\phi\|_{\mathcal{H}} = 0$.

ACKNOWLEDGMENTS

During the preparation of this work, we benefited from illuminating discussions with Bruno Torrésani, Peter Maass, Matthias Holschneider, Jean-Pierre Gazeau and many others. It is a pleasure to thank them all. In addition, we want to thank one of the anonymous referees for his thorough reading and his constructive suggestions. P.V. is a research fellow from F.R.I.A., Belgium.

[1] D. M. Healy, Jr., D. Rockmore, and S. S. B. Moore, ''FFTs for the 2-sphere—Improvements and variations,'' Technical Report PCS-TR96-292, Dartmouth College, Hanover, NH, 1996.

[2] *Wavelets in Physics*, edited by J. C. van den Berg (Cambridge University Press, Cambridge,1998) (to appear).

[3] S. Dahlke, W. Dahmen, E. Schmitt, and I. Weinreich, ''Multiresolution analysis and wavelets on S^2 and S^3,'' Numer. Funct. Anal. Optimiz. **16**, 19–41 (1995).

[4] D. Potts and M. Tasche, ''Interpolatory wavelets on the sphere,'' in *Approximation Theory VIII*, edited by C. K. Chui and L. L. Schumaker (World Scientific, Singapore, 1995), pp. 335–342.

[5] J. Göttelmann, ''Locally supported wavelets on the sphere,'' preprint J. Gutenberg-Univ. Mainz, 1996.

[6] P. Schröder and W. Sweldens, ''Spherical wavelets: Efficiently representing functions on the sphere,'' *Computer Graphics Proceedings (SIGGRAPH95)* (ACM Siggraph, Los Angeles, 1995), pp. 161–175.

[7] W. Sweldens, ''The lifting scheme: a custom-design construction of biorthogonal wavelets,'' Appl. Comput. Harm. Anal. **3**, 1186–1200 (1996).

[8] D. Potts, G. Steidl, and M. Tasche, ''Kernels of spherical harmonics and spherical frames,'' in *Advanced Topics in Multivariate Approximation*, edited by F. Fontanella, K. Jetter, and P-J. Laurent (World Scientific, Singapore, 1996), pp. 1–154.

[9] F. J. Narcowich and J. D. Ward, ''Nonstationary wavelets on the m-sphere for scattered data,'' Appl. Comput. Harm. Anal. **3**, 1324–1336 (1996).

[10] W. Freeden and U. Windheuser, ''Combined spherical harmonic and wavelet expansion—A future concept in Earth's gravitational determination,'' Appl. Comput. Harm. Anal. **4**, 1–37 (1997), and papers quoted therein.

[11] W. Freeden, M. Schreiner, and T. Gervens, *Constructive Approximation on the Sphere, with Applications to Geomathematics* (Clarendon Press, Oxford, 1997).

[12] S. Dahlke and P. Maass, ''Continuous wavelet transforms with applications to analyzing functions on spheres,'' J. Fourier Anal. Appl. **2**, 379–396 (1996).

[13] B. Torrésani, ''Phase space decompositions: Local Fourier analysis on spheres,'' preprint CPT-93/P.2878, Marseille, 1993; ''Position-frequency analysis for signals defined on spheres,'' Signal Process. **43**, 341–346 (1995).

[14] M. Holschneider, ''Continuous wavelet transforms on the sphere,'' J. Math. Phys. **37**, 4156–4165 (1996).

[15] J-P. Antoine and P. Vandergheynst, ''Wavelets on the 2-sphere: a group-theoretical approach,'' Appl. Comput. Harm. Anal. (to appear).

[16] S. T. Ali, J-P. Antoine, J-P. Gazeau, and U. A. Mueller, ''Coherent states and their generalizations: A mathematical overview,'' Rev. Math. Phys. **7**, 1013–1104 (1995), and references quoted therein.

[17] E. Inönü and E. P. Wigner, ''On the contraction of groups and their representations,'' Proc. Natl. Acad. Sci. USA **39**, 510–524 (1953).

[18] J. Saletan, ''Contraction of Lie groups,'' J. Math. Phys. **2**, 1–21 (1961).

[19] M. Holschneider, ''Wavelet analysis on the circle,'' J. Math. Phys. **31**, 39–44 (1990).

[20] I. Daubechies, A. Grossmann, and Y. Meyer, ''Painless nonorthogonal expansions,'' J. Math. Phys. **27**, 1271–1283 (1986).

[21] I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992).

[22] R. Gilmore, ''Geometry of symmetrized states,'' Ann. Phys. (N.Y.) **74**, 391–463 (1972); ''On properties of coherent states,'' Rev. Mex. Fis. **23**, 143–287 (1974).

[23] A. M. Perelomov, ''Coherent states for arbitrary Lie group,'' Commun. Math. Phys. **26**, 222–236 (1972); *Generalized Coherent States and their Applications* (Springer-Verlag, Berlin, 1986).
[24] V. F. Molchanov, ''Harmonic analysis on homogeneous spaces,'' *Representation Theory and Noncommutative Harmonic Analysis II*, edited by A. A. Kirillov (Springer-Verlag, Berlin, 1995).
[25] A. Grossmann, J. Morlet, and T. Paul, ''Integral transforms associated to square integrable representations. I. General results,'' J. Math. Phys. **26**, 2473–2479 (1985).
[26] A. Grossmann, J. Morlet and T. Paul, ''Integral transforms associated to square integrable representations. II. Examples,'' Ann. Inst. Henri Poincaré: Phys. Theor. **45**, 293–309 (1986).
[27] R. Murenzi, ''Ondelettes multidimensionnelles et applications à l'analyse d'images,'' Thèse de Doctorat, Univ. Cath. Louvain, Louvain-la-Neuve, 1990.
[28] J-P. Antoine, P. Carrette, R. Murenzi, and B. Piette, ''Image analysis with two-dimensional continuous wavelet transform,'' Signal Process. **31**, 241–272 (1993).
[29] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic, New York, 1978).
[30] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry—Methods and Applications. I* (Springer-Verlag, New York, 1984).
[31] It would be more consistent to call this group $SO_0(1,n)$, but we stick to the traditional physicists' notation.
[32] A. W. Knapp, *Lie Groups Beyond an Introduction* (Birkhäuser-Verlag, Basel, 1996).
[33] R. Takahashi, ''Sur les représentations unitaires des groupes de Lorentz généralisés,'' Bull. Soc. Math. France **91**, 289–433 (1963).
[34] N. Ja.Vilenkine, *Fonctions Spéciales et Théorie de la Représentation des Groupes* (Dunod, Paris, 1969).
[35] It would be more logical to denote this function by $\lambda(a^{-1},\theta_{n-1})$, but we conform to the traditional notation of Eq. (4.7).
[36] G. Köthe, *Topological Vector Spaces. I* (Springer-Verlag, Berlin, 1969).
[37] M. Duval-Destin, ''Analyse spatiale et spatio-temporelle de la stimulation visuelle à l'aide de la transformée en ondelettes,'' Thèse de Doctorat, Université d'Aix-Marseille II, 1991; J-P. Antoine, M. Duval-Destin, R. Murenzi, and B. Piette, ''Image analysis with 2D wavelet transform: detection of position, orientation and visual contrast of simple objects,'' in *Wavelets and their Applications (Proc. Marseille 1989)*, edited by Y. Meyer (Masson, Paris and Springer-Verlag, Berlin, 1991), pp. 144–159.
[38] A. H. Dooley, ''Contractions of Lie groups and applications to analysis,'' in *Topics in Modern Harmonic Analysis* (Istituto Nazionale di Alta Matematica Francesco Severi, Roma, 1983), Vol. I, pp. 483–515.
[39] A. H. Dooley and J. W. Rice, ''Contractions of rotation groups and their representations,'' Math. Proc. Camb. Philos. Soc. **94**, 509–517 (1983); ''On contractions of semisimple Lie groups,'' Trans. Am. Math. Soc. **289**, 185–202 (1985).
[40] J. Michelsson and J. Niederle, ''Contraction of representations of de Sitter groups,'' Commun. Math. Phys. **27**, 167–180 (1972).
[41] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications. I* (Springer-Verlag, Berlin, 1985).
[42] P. Leboeuf and M. Saraceno, ''Eigenfunctions of non-integrable systems in generalized phase spaces,'' J. Phys. A: Math. Gen **23**, 1745–1746 (1990).
[43] T. Kawazoe, ''Wavelet transforms associated to a principal series representation of semisimple Lie groups. I,'' Proc. Jpn. Acad., Ser. A: Math. Sci. **71**, 154–157 (1995); ''Wavelet transforms associated to a principal series representation of semisimple Lie groups. II,'' *ibid.* **71**, 158–160 (1995).
[44] T. Kawazoe, ''Wavelet transform associated to an induced representation of $SL(n+2,\mathbb{R})$,'' Ann. Inst. Henri Poincaré: Phys. Theor. **65**, 1–13 (1996).
[45] H. Liu and L. Peng, ''Admissible wavelets associated with the Heisenberg group,'' Pacific J. Math. **180**, 101–123 (1997).
[46] J. He and H. Liu, ''Admissible wavelets associated with the affine automorphism group of the Siegel upper half-plane,'' J. Math. Anal. Appl. **208**, 58–70 (1997).
[47] J. He and L. Peng, ''Wavelet transform on the symmetric matrix space,'' preprint, Beijing, 1997.

Weighted trace formula near a hyperbolic trajectory and complex orbits

T. Paul[a)]
CEREMADE, URA 749 CNRS, Université Paris-Dauphine, Place de Lattre de Tassigny, 75775, Paris, Cedex 16, France

A. Uribe
Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109

(Received 15 January 1998; accepted for publication 3 April 1998)

In this paper we consider a weighted trace formula for Schrödinger operators. More precisely, let $\psi_j^\hbar$ and $E_j^\hbar$ denote the eigenfunctions and eigenvalues of a Schrödinger-type operator $H_\hbar$ with a discrete spectrum. Let $\psi_{(x,\xi)}$ be a coherent state centered at a point (x,ξ) of a hyperbolic closed orbit γ. We show that, as $\hbar\to 0$, the leading term of $\Sigma_j \varphi\{[E_j(\hbar)-E]/\hbar\}|(\psi_{(x,\xi)},\psi_j^\hbar)|^2$ can be expressed in terms of the analytic continuation on the upper and lower half-planes of the positive and negative frequencies part of φ. The result is also related to complex trajectories surrounding γ. © *1998 American Institute of Physics.* [S0022-2488(98)01908-2]

I. INTRODUCTION

Consider a Schrödinger operator $H=-\hbar^2\Delta+V(x)$ with V smooth, on $\mathbb{R}^n$ (in which case we assume V tends to infinity at infinity and therefore H has a discrete spectrum) or on a compact Riemannian manifold, M.

The trace (Gutzwiller) formula (Ref. 1) expresses the smeared out spectral density over a Fourier compactly supported test function: let $\{E_j\}$ and $\{\varphi_j\}$ be the eigenvalues and eigenfunctions of H. Then, under certain hypothesis on the classical flow (for example, that the periodic trajectories of energy E are isolated),

$$\sum_j \varphi\left(\frac{E_j-E}{\hbar}\right)\sim\sum_{k=0}^{\infty} c_k(\varphi)\hbar^{-n+1+k}+\sum_{\gamma}\sum_{l=0}^{\infty} d_\gamma^l(\varphi)\hbar^l, \tag{1}$$

where c_k are distributions whose Fourier transform are supported on 0, the second sum is over the periodic trajectories γ of energy E, and d_γ^l have Fourier transforms supported on the set of periods of γ.

We remarked in Ref. 2 that one can isolate the contribution of a given periodic trajectory by ponderating the sum (1) by the so-called Husimi function of φ_j: let $\psi_{(x,\xi)}^a$ be a coherent state at $(x,\xi)\in\mathbb{R}^{2n}$ of ''vacuum'' $a\in\mathcal{S}(\mathbb{R}^n)$:

$$\psi_{(x,\xi)}^a(y)=\rho(y-x)(2\pi\hbar)^{-3n/4}2^{-n/4}e^{-ix\xi/2\hbar}e^{i\xi y/\hbar}a\left(\frac{y-x}{\sqrt{\hbar}}\right) \tag{2}$$

(here ρ is a compactly supported C^∞ function equal to 1 near zero). Let $(x,\xi)\in\gamma$, where γ is a periodic trajectory of the classical underlying flow of energy $E:=\xi^2+V(x)$. We showed that

$$\sum \varphi\left(\frac{E_j-E}{\hbar}\right)|(\psi_{(x,\xi)}^a,\varphi_j)|^2\sim\sum_{k=0}^{\infty}\nu_k(\varphi)\hbar^{-n+1/2+k}, \tag{3}$$

where the ν_k are distributions whose Fourier transforms are supported on the set of periods of (iterates of) γ.

[a)]Electronic mail: paulth@zin.ceremade.dauphine.fr

 © 1998 American Institute of Physics

Let us suppose for simplicity that the dimension $n=2$ (the results being immediately extended to the general case). We remarked in Ref. 2 that if γ is *elliptic*, i.e., if the Poincaré mapping of γ is a rotation of angle θ, the distribution ν_0 could be expressed as a Dirac measure on the energies of the quasimodes associated to γ:

$$\nu_0(\varphi)=\sum_{j,k\in\mathbb{Z}} c_k\varphi\left(\frac{1}{T_\gamma}\left(2\pi j+\left(k+\frac{1}{2}\right)\theta+\frac{S_\gamma}{\hbar}+\sigma_\gamma\right)\right) \tag{4}$$

(here T_γ, S_γ, and σ_γ are the period, action, and Maslov index of γ). On the contrary, if γ is *hyperbolic* the measure ν_0 is Lebesgue continuous.

In this paper we want to give another formulation of ν_0 that involves the complex trajectories of the linearized near γ, i.e., the complex trajectories infinitesimally close to γ (see Theorem 2.1 below). Before we state the results we would like to comment on the role of complex periodic trajectories in the study of the semiclassical spectrum of Schrödinger operators.

Usually the complex trajectories appear to give exponential corrections to the semiclassical expansions. Among the works dedicated to this problem let us mention the work of Balian and Bloch on the trace formula;[3] the resurgence method of Ecalle and Voros;[4–6] and more recently, the work by Fefferman and Secco[7] on the correction term of the number of negative eigenvalues of atomic systems.

More recently, Delande *et al.*[8] and Eckhardt *et al.*,[9] discussed the trace formula for a mapping depending of a parameter near a bifurcation.

A formal computation of the contribution of a given hyperbolic trajectory to the trace formula was given in Ref. 10, where it was remarked, however, that the contribution of different periodic trajectories should interfere destructively. The use of coherent states in the present paper will isolate a given trajectory by microlocalizing near a point in phase space. This makes only one trajectory contributing and the complexified linearized flow near this trajectory present already at the leading order.

The paper is organized as follows: the results are presented in the next section and proven in Sec. III. Section IV contains some concluding remarks.

II. THE RESULTS

Let H be as in the Introduction. Let γ be a hyperbolic closed trajectory of the Hamiltonian $\xi^2+V(x)$. Let μ be the Lyapounov exponent at $(x,\xi)\in\gamma$ [that is, let $e^{\pm\mu}$ be the two eigenvalues of the Poincaré mapping of γ at (x,ξ)] and $\varphi\in\mathscr{S}(\mathbb{R})$ be such that its Fourier transform $\hat{\varphi}$ is compactly supported. We will denote $\varphi^{\pm}$ the Hardy and anti-Hardy part of φ, namely,

$$\varphi^{\pm}(x):=\int_{\mathbb{R}^{\pm}}\hat{\varphi}(\xi)e^{-i\xi x}\,d\xi. \tag{5}$$

We will denote W_a the Wigner function of the symbol a (see the next section) and express W_a on the variables $(x_T,x_\perp,x_s,x_u)$, where x_T is tangent to γ, x_s and x_u are along the stable and unstable manifold of the Poincaré mapping, and $x_\perp$ is a transverse direction to the energy shell.

Then we have the following.

Theorem 2.1: *Let us suppose that*

$$\sum_{k=0}^{\infty}\left(2\frac{1-e^{-\mu}}{1+e^{-\mu}}\right)^k\int|W_a(x_T,x_\perp=0,x_s,x_u)|\frac{|x_sx_u|^k}{k!}dx_T\,dx_s\,dx_u\leqslant\infty. \tag{6}$$

Then as $\hbar\to 0$ along any sequence of the type

$$\hbar=\frac{S_\gamma}{2\pi k+\sigma_\gamma+\alpha},\quad \alpha\in[0,1[,\quad k\to\infty, \tag{7}$$

$$\sum_j\varphi\left(\frac{E_j(\hbar)-E}{\hbar}\right)|(\psi^a_{(x,\xi)},\psi^\hbar_j)|^2=\hbar^{-n+1/2}\sum_{l,m\in\mathbb{Z}}c_{lm}\varphi^{\mathrm{sign}(m)}(z_{lm})+O(\hbar^{-n+3/2}), \tag{8}$$

where $z_{lm} := (2\pi/T_\gamma)\ (l+i(m+\frac{1}{2})\mu+\alpha)$, $l,m\in\mathbb{Z}$. *Moreover, if* $m\geqslant 0$,

$$c_{lm}=\frac{1}{m!}\int x_u^m\ \partial_{x_s}^m W^a(x_T,x_\perp=0,x_s,x_u)e^{ix_sx_u}\,dx_T\,dx_s\,dx_u; \tag{9}$$

if $m<o$,

$$c_{lm}=\frac{1}{m!}\int x_s^m\ \partial_{x_u}^m W^a(x_T,x_\perp=0,x_s,x_u)e^{ix_sx_u}\,dx_T\,dx_s\,dx_u. \tag{10}$$

Remark: The hypothesis (6) is a condition of concentration in phase space of the symbol a near γ along the stable and unstable manifolds. An example of a symbol satisfying (6) is any Gaussian [since in this case the Wigner function is itself a Gaussian and so satisfies (6), as an easy computation shows].

The next result gives a dynamical interpretation of the numbers z_{lm}. In the elliptic case, namely when the linearized flow near γ is stable, one can think, as in Ref. 11 (see also Refs. 12 and 13), at the flow around γ as being integrable, namely, as seating on an (infinitesimal) torus around γ. There are then two ''actions'' arising: the one parametrizing the continuous family of periodic trajectories that γ belongs to (see Ref. 2), and the one coming from the Poincaré mapping. This leads to the following normal form:

$$E(A,B)=E_c(A)+\frac{\theta}{T_\gamma}B,\quad B\ \ small,\quad A\ \ near\ \ S_\gamma, \tag{11}$$

with

$$\frac{\partial E}{\partial A}(S_\gamma,B)=\frac{2\pi}{T_\gamma}. \tag{12}$$

Indeed, $E(A,B)$ induces on the angles φ_A,φ_B, conjugate to A,B a flow at time T_γ given by:

$$\varphi_A(T_\gamma)=\varphi_A(0)+2\pi,$$

$$\varphi_B(T_\gamma)=\varphi_B(0)+\theta.$$

The numbers $(1/T_\gamma)(2\pi j+(k+\frac{1}{2})\theta+(S_\gamma/\hbar)+\sigma_\gamma)$ appearing in (4) are precisely the quantities $[E(A_j,B_k)-E]/\hbar$, where A_j and B_k are the values of the actions quantized by Bohr–Sommerfeld conditions:

$$A_j=S_\gamma+(j+\sigma_\gamma)\hbar\quad\text{and}\quad B_k=(k+\tfrac{1}{2})\hbar. \tag{13}$$

We want to show that this situation is still valid in the hyperbolic case if we consider complex torus and complex actions.

Let P_γ be the Poincaré mapping of γ. Since γ is hyperbolic P_γ can be represented as a matrix of the form

$$P_\gamma=\begin{pmatrix}\cosh\mu & \sinh\mu\\ \sinh\mu & \cosh\mu\end{pmatrix}. \tag{14}$$

We will extend this mapping to the complex in the following way: consider the complex symplectic dilation $D_{\sqrt{i}}$ in $\mathbb{R}^2\otimes\mathbb{C}$:

$$\begin{pmatrix}x\\ \xi\end{pmatrix}\overset{D_{\sqrt{i}}}{\to}\begin{pmatrix}\sqrt{i}x\\ \dfrac{1}{\sqrt{i}}\xi\end{pmatrix}:=\begin{pmatrix}z_1\\ z_2\end{pmatrix}. \tag{15}$$

Under this transformation, the mapping P_γ becomes

$$D_{\sqrt{\hbar}}^{-1} P_\gamma D_{\sqrt{\hbar}} = \begin{pmatrix} \cosh(i\mu) & -i\sinh(i\mu) \\ i\sinh(i\mu) & \cosh(i\mu) \end{pmatrix} = \begin{pmatrix} \cos(i\mu) & -\sin(i\mu) \\ \sin(i\mu) & \cos(i\mu) \end{pmatrix} = R(i\mu), \tag{16}$$

where $R(\theta)$ is the rotation of angle θ.

This means that, thanks to this symplectic complex dilation, we are back to an ''elliptic'' situation *with complex time*. In particular, we find that there exist complex stable tori and a complex normal form $F(A,B)$ satisfying

$$F(A,B) = E_c(A) + i\frac{\mu}{T_\gamma} B, \tag{17}$$

which gives rise to quantized Bohr–Sommerfeld values of the energy $F(A_j, B_k)$ with A_j and B_k given by (13).

We just proved the following.

Proposition 2.2: Let z_{lm} as in Theorem 2.1; then

$$z_{lm} = F(A_l, B_m), \tag{18}$$

with

$$A_j = S_\gamma + (j+\sigma_\gamma)\hbar \quad \textit{and} \quad B_k = (k+\tfrac{1}{2})\hbar. \tag{19}$$

We will come back to the interpretation of this result in the final section of the paper.

III. PROOF OF THE THEOREM

In Ref. 2 we proved that the first coefficient $\nu_0(\varphi)$ in (3) can be written as

$$\nu_0(\varphi) = \sum_n e^{in[(S_\gamma/\hbar)+\sigma_\gamma]} \int_{-\infty}^{\infty} (a, Z(s(\dot{x},\dot{\xi}))U^n a)\, ds\ \hat{\varphi}(nT_\gamma); \tag{20}$$

here $(\dot{x},\dot{\xi})$ is the tangent vector to the flow at (x,ξ); Z is a Weyl operator defined by

$$Z(e,f)a(\eta) = e^{-i(ef/2)} e^{ie\eta} a(\eta - f), \tag{21}$$

and U is the metaplectic operator of the linearized low at time T_γ (see Ref. 2).

We also showed that one can find a symplectic mapping R such that $U = M(S(T_\gamma))$, where M is the metaplectic representation such that

$$R^{-1} S(T_\gamma) R = \begin{pmatrix} 1 & 0 & \alpha & 0 \\ 0 & e^{-\mu} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{\mu} \end{pmatrix}, \tag{22}$$

where $\alpha \in \mathbb{R}$ and μ is the local Lyapounov exponent of γ at (x,ξ).

Let us denote $a' := M(R)a$. Then one easily checks that, if $\hbar$ satisfies (6),

$$\begin{aligned} \nu_0(\varphi) &= \sum_n e^{in\alpha} \int \overline{a'(\eta)} e^{in(\alpha/2)\partial_{\eta_1}^2} e^{-n(\mu/2)} a'(\eta_1 - s, e^{-n\mu}\eta_2)\, d\eta\, ds \\ &= \sum_n e^{-n(\mu/2)} e^{in\alpha} \int \overline{a'(0,\eta_2)} a'(0, e^{-\mu}\eta_2)\, d\eta_2, \end{aligned} \tag{23}$$

where a' is the Fourier transform of a' with respect to η_1. Let us call $b(x) := a'(0,x)$ and let $W_b(x,\xi)$ be the Wigner function of b, namely,

$$W_b(x,\xi) := \int \overline{b(x-y)}b(x+y)e^{2i\xi y}\, dy. \tag{24}$$

Lemma 3.1: Let us suppose that

$$\sum_{k=0}^{\infty}\left(2\frac{1-e^{-\mu}}{1+e^{-\mu}}\right)^k \int |W_b(x,\xi)|\frac{|x\xi|^k}{k!}dx\ d\xi \leqslant \infty, \tag{25}$$

then

$$e^{-n(\mu/2)}\int \overline{a'(0,\eta_2)}a'(0,e^{-\mu}\eta_2)d\eta_2 = \sum_{k=0}^{\infty}\frac{\overline{b^{(k)}(0)}\hat{b}^{(k)}(0)}{k!}e^{-(k+1/2)\mu}, \tag{26}$$

where $b^{(k)}$ is the kth derivative of b.

Proof: We have, by straightforward computations,

$$\int \overline{a'(0,\eta_2)}a'(0,e^{-\mu}\eta_2)d\eta_2 = \int \overline{b(x)}b(e^{-\mu}x)dx$$

$$= \int W_b\left(\frac{x(1+e^{-\mu})}{2},\xi\right)e^{-i\xi x(1-e^{-\mu})}\, dx\ d\xi \tag{27}$$

$$= \frac{2}{1+e^{-\mu}}\int W_b(x,\xi)e^{-2i\xi x[(1-e^{-\mu}/1+e^{-\mu})]}dx\ d\xi; \tag{28}$$

(25) implies that the RHS of (27) is convergent. Therefore (26) is also convergent. □

Let us show now that the Lemma implies the formula (8). Since $\hat{\varphi}$ is compactly supported we can plug (20) in the expansion (26) and invert the summations. It now suffices to note that $e^{(-\xi)}\chi_{[0,+\infty[}$ is the Fourier transform of $2\pi/(i-x)$ and use the Cauchy and Poisson formulas to get

$$\nu_0(\varphi) = \sum_{l,m\in\mathbb{Z}} c_{lm}\varphi^{\mathrm{sign}(m)}(z_{lm}), \tag{29}$$

with

$$c_{lm} = \frac{\overline{b^{(k)}(0)}\hat{b}^{(k)}(0)}{k!}. \tag{30}$$

We need now to express the hypothesis (25) and the c_{lm} in terms of the Wigner function of a. To do so let us first remark that if S is a symplectic mapping and M the metaplectic representation, we have

$$W_{M(S)a}(x,\xi) = (W_a oS)(x,\xi). \tag{31}$$

Together with the fact that

$$\overline{\hat{a}(0)}\hat{a}(0) = \int W_a(0,\xi)d\xi, \tag{32}$$

we get easily that (25) is equivalent to

$$\sum_{k=0}^{\infty}\left(2\frac{1-e^{-\mu}}{1+e^{-\mu}}\right)^k \int |W_a(x_T, x_\perp = 0, x_s, x_u)|\frac{|x_s x_u|^k}{k!}dx_T\ dx_s\ dx_u \leqslant \infty, \tag{33}$$

and the expression (9) and (10) for the coefficients.

IV. COMMENTS

Link with ''top resonances'': although the operator H is elliptic with discrete spectrum, which implies that there are no ''resonances,'' the formula (8) suggests that the Hardy and anti-Hardy parts of the weighted spectral density $\rho(\lambda):=\Sigma\,\delta(\lambda-[E_j(\hbar)-E]/\hbar)|(\psi^a_{(x,\xi)},\psi^\hbar_j)|^2$ have poles, in the semiclassical limit. These poles are precisely located on the same lattice that in the case of the so-called ''top resonances'' (Refs. 14–16): if the potential V tends to zero at infinity and if there is an unstable fixed point or a manifold of unstable fixed points on the energy surface, then one can prove that H has resonances at a distance $\sim\hbar$ from the real axis. Our result gives a microlocalized version of this phenomenon.

Link with the ''analytic dilation'' method: let us look at the poles z_{l0} and the corresponding coefficient c_0,

$$c_{l0}=\int W^a(x_T,x_\perp=0,x_s,x_u)e^{ix_sx_u}\,dx_T\,dx_s\,dx_u\,. \tag{34}$$

Calling

$$\begin{pmatrix}x\\ \xi\end{pmatrix}:=\begin{pmatrix}\dfrac{x_s+x_u}{\sqrt{2}}\\[2ex] \dfrac{x_s-x_u}{\sqrt{2}}\end{pmatrix}, \tag{35}$$

one gets

$$c_{l0}=\int W^a(x_T,x_\perp=0,x,\xi)e^{i[(\xi^2-x^2)/2]}dx_T\,dx\,d\xi. \tag{36}$$

Using elementary properties of the Wigner function, one gets that

$$c_{l0}=(b,g^+)(g^-,b), \tag{37}$$

where b was defined in Sec. III and $g^\pm(x):=e^{\pm i(x^2/2)}$.

Moreover, the same computation in the case of γ elliptic gives rise to

$$c^{ell}_{l0}=(b,g)(g,b), \tag{38}$$

where $g(x):=e^{-x^2/2}$. Let $|g\rangle\langle g|$ be the orthogonal projector on the vector g in $L^2(\mathbb{R})$. If one calls D_α the operator of dilation by α, one sees that $|g^+\rangle\langle g^-|$ is the analytic continuation of $D_\alpha^{-1}|g\rangle\langle g|D_\alpha$ evaluated at $\alpha=\sqrt{i}$, and so is $c_{l0}:=(b,|g^+\rangle\langle g^-|b)$. This suggest that the poles of the weighted spectral measure can be obtained, as the usual resonances do, by analytic dilation.

Link with ''normal forms'': in Refs. 17 and 18, Guillemin introduced quantized normal forms near a closed trajectory. The result of this paper suggests that the complex dilation of this normal form gives rise to poles of the spectral density suitably microlocalized.

ACKNOWLEDGMENT

The research of A. Uribe was supported, in part, by National Science Foundation Grant No. DMS-9107600.

[1] T. Paul and A. Uribe, ''The semi-classical trace formula and propagation of wave packets,'' J. Funct. Anal. **132**, 192–249 (1995).
[2] T. Paul and A. Uribe, ''On the pointwise behaviour of semi-classical measures,'' Commun. Math. Phys. **175**, 229–258 (1996).
[3] R. Balian and C. Bloch, Ann. Phys. (N.Y.) **63**, 592–606 (1971); **64**, 271–307 (1971); **69**, 76–160 (1972); **85**, 514–545 (1974).
[4] J. Ecalle, ''Les fonctions résurgentes,'' Publications Mathématiques de l'Université Paris Sud, 81-05 and 81-06, 1981.
[5] A. Voros, ''The return of the quartic oscillator. The complex WKB method,'' Ann. Inst. Henri Poincaré, Sect. A **34**, 211–388 (1983).

[6] A. Voros and P. Cartier, in *Grothendieck Festschrift* (Birkhauser, Basel, 1990), Vol. 2.
[7] A. Córdoba, C. Feffermann, and L. Secco, *Weyl Sums and Atomic Energy Oscillations Revista Matemática Iberoamericana* (1995), Vol 11, No. 1. pp. 167–228.
[8] D. Delande, M. Kus, and F. Haake, ''Prebifurcation periodic ghost orbits in semi-classical quantization,'' Phys. Rev. Lett. **71**, 2167–2171 (1993).
[9] B. Eckhardt, M. Kus, and F. Haake, Z. Phys. B **65**, 221 (1993).
[10] A. Voros, ''Unstable periodic orbits and semi-classical quantization,'' J. Phys. A **21**, 685–692 (1988).
[11] A. Voros, ''The WKB–Maslov method for nonseparable systems Coll. Inst. CNRS n. 237,'' *Géométrie Symplectique et Physique Mathématique*, pp. 217–252.
[12] J. V. Ralston, ''On the construction of quasimodes associated with stable periodic orbits,'' Commun. Math. Phys. **51**, 219–242 (1976).
[13] Y. Colin de Verdière, ''Quasi-modes sur les varietes Riemanniennes,'' Invent. Math. **43**, 15–42 (1977).
[14] T. Paul, Thèse d'etat, Université d'Aix-Marseille, 1985.
[15] R. Briet, P. Duclos, and J. M. Combes, Commun. Part. Diff. Eq. **12**, 201 (1987).
[16] C. Gérard and J. Sjöstrand, Commun. Math. Phys. **108**, 391–421 (1987).
[17] V. Guillemin, ''Wave-trace invariants and a theorem of Zelditch,'' Bull. Am. Math. Soc. **12**, 303–308 (1993).
[18] V. Guillemin, ''Wave-trace invariants,'' Duke Math. J. **83**, 287–352 (1996).

Separability, positivity, and minimum uncertainty in time–frequency energy distributions

Patrick Flandrin[a)]
Ecole Normale Supérieure de Lyon, Laboratoire de Physique (URA 1325 CNRS), 46 allée d'Italie, 69364 Lyon Cedex 07, France

(Received 24 June 1997; accepted for publication 4 February 1998)

Gaussian signals play a very special role in classical time–frequency analysis because they are solutions of apparently unrelated problems such as minimum uncertainty, and positivity and separability of Wigner–Ville distributions. We investigate here some of the logical connections which exist between these different features, and we discuss some examples and counterexamples of their extension to more general joint distributions within Cohen's class and the affine class. © *1998 American Institute of Physics.* [S0022-2488(98)00508-8]

I. INTRODUCTION

In most cases, physical signals are defined as functions of time. Until the somewhat recent past, signal analysis has been mainly concerned with such a temporal description of signals, or with a dual description in the frequency domain, obtained from a classical Fourier transform. Much information can, of course, be gained from both descriptions, but it is also clear that time-only and frequency-only representations, being in some sense orthogonal to each other, are not best suited for displaying mixed information about time and frequency. In fact, there are many situations in which a Fourier representation, although mathematically correct, is not able to capture in a direct way time-dependent spectral features which may be intuitively expected. This is so, for instance, with music signals, for which our everyday experience suggests that the ''frequency of tones is continuously changing.'' In this case, Fourier analysis is clearly not well adapted and intuition rather calls for a joint time–frequency description, revealing not only the different frequencies occurring in a piece of music, but also their time of occurrence, their duration,..., i.e., the kind of information which is indeed coded on a musical score.

Although interpretation may be different, the search for a joint time–frequency description of signals has much to share with the problem of finding joint distributions of position and momentum in Quantum Mechanics. It is therefore quite natural that most of the tools which have been developed in either domain have indeed found applications in both. This is especially the case for all the Wigner-based distributions which have been proposed and extensively studied since the pioneering works of Wigner in quantum mechanics[1] and Ville in signal theory.[2] One can, however, remark that, in quantum mechanics, joint distributions $P_\psi(q,p)$ are mostly used as a computational tool, allowing us to write

$$(\psi,\mathbf{G}\psi)=\int\int_{-\infty}^{+\infty} g(q,p)P_\psi(q,p)dq\ dp,$$

where **G** stands for the operator associated with a given classical quantity $g(q,p)$ in the sense of some (nonunique) correspondence rule. This nonunicity—which is due to the fact that two dual variables connected by a Fourier transform relationship are associated with elementary operators which do not commute—carries over to the definition of $P_\psi(q,p)$, and the situation is similar for time and frequency. In signal theory, however, a joint distribution is basically considered as a (quasi-) density function of its variables, and the main issue is much more to get a readable time–frequency ''picture'' of a signal, which has therefore to be interpreted *per se* and not only through inner products with test functions.

[a)]Electronic mail: flandrin@physique.ens-lyon.fr

 © 1998 American Institute of Physics

An extensive body of literature has been devoted to time–frequency analysis with Wigner-type distributions, from the point of view of both theory and interpretation (for a survey, see, e.g., Ref. 3 or 4), and here we would like to focus further on some specific questions related to both aspects. More precisely, if we consider Gaussian signals of the form

$$g(t)=Ce^{-\alpha t^2}, \tag{1}$$

with $C\in\mathbf{C}$ and $\alpha>0$, we know that, besides the fact that their Fourier transform is also Gaussian, namely that

$$G(f)\equiv\int_{-\infty}^{+\infty}g(t)e^{-i2\pi ft}\,dt=C\sqrt{\frac{\pi}{\alpha}}\,e^{-\pi^2f^2/\alpha},$$

such signals happen to play a very special role in classical time–frequency analysis. (Throughout the paper, we will adopt the convention of using a lower case symbol for representing a signal in the time domain, and the corresponding upper case symbol for denoting its Fourier transform, in the frequency domain.) This is in fact so for at least three different reasons.

(1) Minimum uncertainty. Gaussians (1) are the only minimizers for the time–frequency uncertainty relation,[5]

$$\Delta t_x\Delta f_x\geqslant\frac{1}{4\pi}, \tag{2}$$

where

$$\Delta t_x^2=\frac{1}{E_x}\int_{-\infty}^{+\infty}t^2|x(t)|^2\,dt, \tag{3}$$

$$\Delta f_x^2=\frac{1}{E_x}\int_{-\infty}^{+\infty}f^2|X(f)|^2\,df, \tag{4}$$

and

$$E_x=\int_{-\infty}^{+\infty}|x(t)|^2dt.$$

(2) Positivity. Gaussians (1) are—up to linear and quadratic phase terms—the only signals for which the Wigner–Ville distribution

$$W_x(t,f)\equiv\int_{-\infty}^{+\infty}x\left(t+\frac{\tau}{2}\right)\overline{x\left(t-\frac{\tau}{2}\right)}e^{-i2\pi f\tau}\,d\tau=\int_{-\infty}^{+\infty}X\left(f+\frac{\nu}{2}\right)\overline{X\left(f-\frac{\nu}{2}\right)}e^{i2\pi\nu t}\,d\nu, \tag{5}$$

is everywhere non-negative (Hudson's theorem;[6] cf. Proposition 1 below).

(3) Separability. The Wigner–Ville distribution (5) of Gaussians (1) is separable, namely,

$$W_g(t,f)=|C|^2\sqrt{\frac{2\pi}{\alpha}}\,e^{-2\alpha t^2}e^{-(2\pi^2/\alpha)f^2}. \tag{6}$$

These three properties are important ones for attaching to Gaussian signals a specific status and for offering a nice physical interpretation of their time–frequency content by means of the Wigner–Ville distribution.

(1) Minimum uncertainty is often invoked for considering Gaussian signals as time–frequency ''atoms'' and for using them as elementary building blocks (the so-called ''Gabor logons'' [5]) of linear signal decompositions.

(2) Although it is known that no time–frequency distribution—be it positive or not—can be interpreted pointwise,[7] positivity improves the readability of a time–frequency diagram.

(3) Separability provides further insight in the probability picture of time–frequency distributions by corresponding to a notion of ''independence,'' according to which a logon is a state whose time and frequency behaviors are decoupled.[8]

Given these facts, an interesting question is to know whether the three mentioned properties (minimum uncertainty, positivity, and separability) have logical reasons to be related or, on the contrary, if their simultaneous occurrence is only fortuitous. Moreover, since many other joint distributions do exist besides the Wigner–Ville distribution, a further question which is worth investigating is how the results pertaining to the pair ''Gaussian/Wigner–Ville'' can be extended to other situations, especially in the important case of affine (scale-based) distributions.

The paper is devoted to these questions and is organized as follows. In Sec. II, the situation of the Wigner–Ville distribution is considered in some detail by reviewing and/or establishing a number of its specific properties connected in a direct way with the considered problem. In Sec. III we address the question of a possible extension of such results to the more general ''Cohen's class,'' whereas in Sec. IV we are concerned with similar questions within the ''affine class'' framework. Finally, some conclusions are drawn, together with the possibility of getting further extensions in more general situations.

II. WIGNER–VILLE DISTRIBUTION

In this section, we will review and/or establish basic properties—related to positivity, separability, and minimum uncertainty—which hold in the case of the Wigner–Ville distribution. As far as positivity is concerned, it has to be noted that many other important results exist (see, e.g., Refs. 9–12), even with extensions beyond the Wigner–Ville case. The purpose of this paper is not to review all of them, and only those results which are connected in some way with the two other issues of separability and minimum uncertainty will be considered here.

A. Positivity

Proposition 1 (Hudson's theorem[6]): The Wigner–Ville distribution is positive for signals of the type

$$g_{\alpha,\beta,\gamma}(t) \equiv e^{-(\alpha t^2+\beta t+\gamma)}, \tag{7}$$

with $(\alpha,\beta,\gamma)\in \mathbf{C}^3$ *and* $Re\{\alpha\}>0$, *and only for them.*

Proof: The fact that (generalized) Gaussian signals of the form (7) have a positive Wigner–Ville distribution follows from a direct calculation, according to which

$$W_{g_{\alpha,\beta,\gamma}}(t,f) = \sqrt{\frac{2\pi}{\mathrm{Re}\{\alpha\}}}\,\exp\left(-2\,\mathrm{Re}\{\alpha t^2+\beta t+\gamma\}-\frac{2\pi^2}{\mathrm{Re}\{\alpha\}}\,[f-f_i(t)]^2\right),$$

with

$$f_i(t) \equiv \frac{1}{2\pi}\,\mathrm{Im}\{2\alpha t+\beta\}.$$

The original proof of the converse was first stated by Hudson in Ref. 6. It consists of introducing—according to the definition (7)—the test signal $g_{1,z,\gamma}(t)$, for which the corresponding Wigner–Ville distribution is everywhere non-negative. Making use of Moyal's formula,[4] we have

$$\left|\int_{-\infty}^{+\infty} x(t)\overline{g_{1,z,\gamma}(t)}\,dt\right|^2 = \int\int_{-\infty}^{+\infty} W_x(t,f)\,W_{g_{1,z,\gamma}}(t,f)\,dt\,df,$$

and this quantity is guaranteed to be everywhere positive for those signals $x(t)$ whose Wigner–Ville distribution is non-negative. It follows that the function

$$F(z) = e^{\gamma}\int_{-\infty}^{+\infty} x(t)\overline{g_{1,z,\gamma}(t)}\,dt \tag{8}$$

is analytic, has no zeros, and furthermore satisfies

$$|F(z)|^2 \leqslant \sqrt{\frac{\pi}{2}}\, E_x \exp\left(\frac{1}{2}\,(\mathrm{Re}\{z\})^2\right),$$

as can be checked by applying the Cauchy–Schwarz inequality.

As a consequence, $F(z)$ is an entire function of order at most 2, without zeros. Hence, by Hadamard's factorization theorem, $F(z)$ is necessarily the exponential of some quadratic form in z. Letting $z=i2\pi f$ and using the fact that Fourier transforms of Gaussians are still Gaussians, we can deduce from (8) that $x(t)$ is itself the exponential of some quadratic form in t and the proof is complete.

Two remarks.

(1) In order to have finite energy signals—i.e., $x(t)\in L^2(\mathbf{R})$—we have assumed that $\mathrm{Re}\{\alpha\}>0$ and this, in turn, guarantees that the Wigner–Ville distribution is bounded since it is easy to show that $|W_x(t,f)|\leqslant 2E_x$. A degeneracy can, however, be observed whenever $\mathrm{Re}\{\alpha\}=\mathrm{Re}\{\beta\}=0$, a situation for which the Wigner–Ville distribution reads as

$$W_{g_{\alpha,\beta,\gamma}}(t,f) = \exp(-2\,\mathrm{Re}\{\gamma\})\,\delta(f-f_i(t)).$$

(2) Minimum uncertainty implies positivity of the Wigner–Ville distribution, but the class of signals (7) which have a positive Wigner–Ville distribution is larger than the class (1) of signals with minimum uncertainty. In fact, the former can be seen as a modulation of the latter by means of a ''chirp'' wave form whose phase is quadratic in time and, hence, whose instantaneous frequency is linear. Precisely, we get

$$\Delta t_{g_{\alpha,\beta,\gamma}}\Delta f_{g_{\alpha,\beta,\gamma}} = \frac{1}{4\pi}\left(1+\left(\frac{\mathrm{Im}\{\alpha\}}{\mathrm{Re}\{\alpha\}}\right)^2\right)^{1/2} \geqslant \frac{1}{4\pi},$$

with equality if and only if $\mathrm{Im}\{\alpha\}=0$, i.e., if the Gaussian is not modulated in frequency, thus reducing to (1).

B. Separability

Proposition 2: If a Wigner–Ville distribution is separable, it is positive and factorizable as

$$W_x(t,f) = \frac{|x(t)|^2|X(f)|^2}{E_x}. \tag{9}$$

Proof: Starting from the definition (5), it is well known (and easy to check) that a Wigner–Ville distribution always has correct marginals, i.e., that

$$\int_{-\infty}^{+\infty} W_x(t,f)\,dt = |X(f)|^2, \tag{10}$$

$$\int_{-\infty}^{+\infty} W_x(t,f)\,df = |x(t)|^2. \tag{11}$$

Assuming that separability holds, there exist two functions $p(t)$ and $Q(f)$ such that $W_x(t,f)=p(t)Q(f)$. Using (11) and (10), we get $P(0)Q(f)=|X(f)|^2\geqslant 0$ and $p(t)q(0)=|x(t)|^2\geqslant 0$. It follows that $P(0)$ and $Q(f)$ on one hand, and $p(t)$ and $q(0)$ on the other hand, are necessarily of the same sign. One more integration leads to $P(0)q(0)=E_x>0$, and this concludes the proof.

Remark: We know from Moyal's formula that, for any finite energy signal $x(t)$, we have

$$\int\int_{-\infty}^{+\infty} W_x^2(t,f)\,dt\,df = E_x^2.$$

Applying this result to the factorized distribution (9), we are led to the conclusion that signals with a separable Wigner–Ville distribution must necessarily be such that

$$\left(\int_{-\infty}^{+\infty} |x(t)|^4 \, dt \right)^{1/4} \left(\int_{-\infty}^{+\infty} |X(f)|^4 \, df \right)^{1/4} = E_x,$$

a property which holds for Gaussians. Gaussians are therefore solutions for separability. In fact, it can be shown that they are the only solutions, as claimed in the following.

Proposition 3: The Wigner–Ville distribution is separable for Gaussian signals only.

Proof: This result can be viewed as a Corollary to Propositions 2 and 1 (''separability⇒positivity⇒Gaussians''), but it can also be derived in a direct way as follows. Assuming that separability holds for a Wigner–Ville distribution, we can start from the factorization of (9) and, after taking an inverse Fourier transform over frequency, we get

$$\frac{1}{|x(t)|^2} x\left(t+\frac{\tau}{2}\right)\overline{x\left(t-\frac{\tau}{2}\right)} = \gamma_x(\tau), \tag{12}$$

with

$$\gamma_x(\tau) \equiv \frac{1}{E_x} \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right)\overline{x\left(t-\frac{\tau}{2}\right)} dt.$$

This defines a functional equation whose first implication is that its left-hand side must be a function of τ only. Writing $x(t)=|x(t)|\exp\{i \arg x(t)\} \equiv \exp\{\varphi(t)+i\psi(t)\}$, we end up with two necessary conditions upon a modulus term and a phase term, respectively. The phase condition (of independence upon the t variable) reads as

$$\frac{\partial}{\partial t}\left[\psi\left(t+\frac{\tau}{2}\right) - \psi\left(t-\frac{\tau}{2}\right) \right] = 0,$$

for any t and τ, with the consequence that the phase must necessarily be of the form $\psi(t)=at+b$, with $(a,b)\in\mathbf{R}^2$. After a change of variables, the companion condition on the modulus leads, for any v and w, to

$$\dot{\varphi}\left(\frac{v+w}{2}\right) = \frac{\dot{\varphi}(v)+\dot{\varphi}(w)}{2},$$

where the overdot indicates the (time) derivative.

We recognize in this relation *Jensen's functional equation*,[13] whose most general (continuous) solution in $\dot{\varphi}(t)$ is given by linear functions of the form $\dot{\varphi}(t)=a't+b'$ with $(a',b')\in\mathbf{R}^2$. It follows that $\varphi(t)=(a'/2)t^2+b't+c'$, with $c'\in\mathbf{R}$ and the requirement that $a'\leqslant 0$ in order to guarantee that $x(t)\in L^2(\mathbf{R})$.

Gaussian signals therefore appear as candidates for defining separable Wigner–Ville distributions. In order to be admissible solutions, we have to further check that the left-hand side of (12) is not only independent of t, but also proportional to the deterministic correlation function $\gamma_x(\tau)$, which an elementary calculation proves to be true.

C. Minimum uncertainty

Proposition 4 (Ref. 14): Given any $T>0$, the Wigner–Ville distribution satisfies the time–frequency uncertainty relation,

$$\Sigma_{tf}^2(W_x) \equiv \frac{1}{E_x} \int \int_{-\infty}^{+\infty} \left(\frac{t^2}{T^2} + T^2 f^2 \right) W_x(t,f)\, dt\, df \geqslant \frac{1}{2\pi}, \tag{13}$$

with equality if and only if $x(t)$ is a Gaussian signal.

Proof: This follows directly from the marginal properties (10)–(11) of the Wigner–Ville distribution and from the definitions (3)–(4), since we can write

$$\Sigma_{tf}^2(W_x)=\frac{1}{E_x}\left[\frac{1}{T^2}\int_{-\infty}^{+\infty}t^2|x(t)|^2dt+T^2\int_{-\infty}^{+\infty}f^2|X(f)|^2df\right]$$

$$=\frac{\Delta t_x^2}{T^2}+T^2\Delta f_x^2=\left(\frac{\Delta t_x}{T}-T\Delta f_x\right)^2+2\Delta t_x\Delta f_x\geqslant 2\Delta t_x\Delta f_x, \tag{14}$$

with equality if the arbitrary duration T is ''matched'' to the signal $x(t)$ according to $T=\sqrt{\Delta t_x/\Delta f_x}$. In this case, we know from (2) that the right-hand side of (14) is lower bounded too, with equality in the only case of Gaussians, whence the result.

Remark: In the above derivation, it has been implicitly assumed that the mean time and the mean frequency of the analyzed signal are both zero, i.e., that

$$\int_{-\infty}^{+\infty}t|x(t)|^2\,dt=\int_{-\infty}^{+\infty}f|X(f)|^2\,df=0.$$

As for the standard Gabor–Heisenberg uncertainty relation (2), this of course does not reduce the generality of the result, since the Wigner–Ville distribution being covariant with respect to shifts in time and frequency, uncertainty relations pertaining to signals with nonzero mean time and/or mean frequency can be established *mutatis mutandis* by introducing suitable shifts.

Because a Wigner–Ville distribution attains generally negative values, the interpretation of (13) as a measure of time–frequency energy spread can be questioned. A companion result can, however, be obtained for *squared* Wigner–Ville distributions, thus reinforcing the prominent role of Gaussians. This is given by the following.

Proposition 5 (Ref. 15): Given any $T>0$, the squared Wigner–Ville distribution satisfies the time–frequency uncertainty relation

$$\Sigma_{tf}^2(W_x^2)=\frac{1}{E_x^2}\int\int_{-\infty}^{+\infty}\left(\frac{t^2}{T^2}+T^2f^2\right)W_x^2(t,f)\,dt\,df\geqslant\frac{1}{4\pi},$$

with equality if and only if $x(t)$ is a Gaussian signal.

Proof: The proof proceeds as previously, except that we have to evaluate second-order moments with respect to $W_x^2(t,f)$. It is easy to show that

$$\int_{-\infty}^{+\infty}W_x^2(t,f)\,df=\int_{-\infty}^{+\infty}\left|x\left(t+\frac{\tau}{2}\right)\right|^2\left|x\left(t-\frac{\tau}{2}\right)\right|^2d\tau,$$

and, therefore,

$$\frac{1}{E_x^2}\int\int_{-\infty}^{+\infty}t^2W_x^2(t,f)\,dt\,df=\frac{1}{E_x^2}\int\int_{-\infty}^{+\infty}\left(\frac{v+w}{2}\right)^2|x(v)|^2|x(w)|^2dv\,dw$$

$$=\frac{1}{4E_x^2}\left[2E_x\int_{-\infty}^{+\infty}v^2|x(v)|^2dv+2\left(\int_{-\infty}^{+\infty}v|x(v)|^2dv\right)^2\right]\geqslant\frac{\Delta t_x^2}{2},$$

with equality for signals whose mean time is zero. It can be proved in a similar way that

$$\frac{1}{E_x^2}\int\int_{-\infty}^{+\infty}f^2W_x^2(t,f)\,dt\,df\geqslant\frac{\Delta f_x^2}{2},$$

with equality for signals whose mean frequency is zero.

Given these two properties, the proof is finished by reorganizing terms exactly as in the nonsquared case.

Remark: The fact that both the Wigner–Ville distribution $W_x(t,f)$ and its square $W_x^2(t,f)$ have a minimum time–frequency spread which is attained for Gaussian signals suggests a natural link between minimum uncertainty and positivity. For simplicity, let us assume that $x(t)$ is of unit energy so that both $W_x(t,f)$ and $W_x^2(t,f)$ integrate to 1 in the time–frequency plane. In fact, as

far as the spread related to $W_x(t,f)$ only is concerned, a decrease in the spread measure could *a priori* be obtained not only by concentrating positive values around the origin, but also by accepting negative values far from the origin. This, however, turns out to be in contradiction with the spread inequality related to $W_x^2(t,f)$ since a distribution with negative values would create ''holes''—in between positive contributions—when squared, thus increasing the spread measure as compared to a positive distribution.

It is possible to find many other ways of formulating uncertainty relations by means of the Wigner–Ville distribution, among which we can cite the following results.

Proposition 6: Any signal $x(t)$ of finite energy E_x satisfies the ''uncertainty relation''

$$\alpha_x \equiv \frac{1}{E_x}\int_{-\infty}^{+\infty} t^2|\dot{x}(t)|^2 dt \geqslant \frac{1}{4},$$

with

$$\lim_{\gamma\to 0_+} \alpha_{x_\gamma} = \tfrac{1}{4},$$

for signals of the form

$$x_\gamma(t) \equiv C\ \exp(-\beta|t|^\gamma), \tag{15}$$

with $C\in\mathbf{C}$ and $(\beta,\gamma)\in\mathbf{R}_+^2$.

Proof: For convenience, let us introduce the quantity

$$J \equiv \int_{-\infty}^{+\infty} t\dot{x}(t)\overline{x(t)}\,dt,$$

whose imaginary part is sometimes referred to as the ''covariance'' of the signal.[3] By applying the Cauchy–Schwarz inequality, we get $|J|^2 \leqslant \alpha_x E_x^2$, whereas an integration by parts shows that $\mathrm{Re}\{J\} = -E_x/2$, and the claimed inequality follows from the fact that we always have $|J|^2 \geqslant (\mathrm{Re}\{J\})^2$.

From the way the lower bound has been derived, it is not clear whether it can be effectively attained. However, it can be shown by a direct calculation that signals of the form (15) are such that $\alpha_{x_\gamma} = (\gamma+1)/4$, with the consequence that $\lim_{\gamma\to 0_+}\alpha_{x_\gamma} = 1/4$, thus completing the proof.

Three remarks.

(1) Although the lower bound $\alpha_x = \frac{1}{4}$ is asymptotically sharp for signals $x_\gamma(t)$ when $\gamma\to 0_+$, it has to be remarked that the bandwidth Δf_{x_γ} and, hence, the usual uncertainty measure remain finite only if $\gamma > \frac{1}{2}$.

(2) The ''alpha moment'' α_x (a quantity which has been introduced and studied by Titlebaum; see Ref. 16) is such that

$$\alpha_x = \frac{1}{E_x}\int_{-\infty}^{+\infty} t^2|\dot{x}(t)|^2 dt = \frac{1}{E_x}\int_{-\infty}^{+\infty} f^2|\dot{X}(f)|^2 df.$$

It follows that any property of α_x which holds for a given signal $x(t)$ also holds for its Fourier transform $X(f)$. In particular, the lower bound for α_x is also attained for signals whose spectrum reads as $X(f) = C\ \exp(-\beta|f|^\gamma)$, in the limit $\gamma\to 0_+$, with the consequence that the corresponding signals have a finite mean-square duration—and, hence, a finite usual uncertainty measure—only if $\gamma > \frac{1}{2}$.

(3) The uncertainty measure given by the ''alpha moment'' α_x is not minimum for Gaussian signals. In fact, expanding $x(t)\in L^2(\mathbf{R})$ as $x(t) = \Sigma_{n=0}^{+\infty} x_n \psi_n(t)$ on the orthonormal basis of Hermite functions $\{\psi_n(t); n\in\mathbf{N}\}$ (defined as in Ref. 17), we get

$$E_x\alpha_x = \sum_{n=0}^{+\infty}\left(\left[\frac{3}{4}+\frac{n(n+1)}{2}\right]|x_n|^2 - \frac{\sqrt{(n+1)(n+2)(n+3)(n+4)}}{2}\,\mathrm{Re}\{x_n x_{n+4}\}\right).$$

This proves that $\alpha_{\psi_n}=3/4+n(n+1)/2$ and thus that the Gaussian (associated to $n=0$) may be considered as a minimizer for α_x, but in the restricted class of Hermite functions.

Proposition 7: For any signal $x(t)$ of finite energy E_x and of mean time and mean frequency identically zero, the Wigner–Ville distribution satisfies the uncertainty relation

$$\sigma_{tf}^2(W_x^2)\equiv\frac{1}{E_x^2}\int\int_{-\infty}^{+\infty}t^2f^2W_x^2(t,f\,)dt\;df\geqslant\frac{1}{32\pi^2}\left(\alpha_x-\frac{1}{2}+4\pi^2\Delta t_x^2\Delta f_x^2\right),$$

with equality for signals whose ''covariance'',

$$\mathrm{cov}_{tf}(x)\equiv\mathrm{Im}\left(\int_{-\infty}^{+\infty}t\dot{x}(t)\overline{x(t)}dt\right),\tag{16}$$

is zero.

Proof: A way of proving this result is to make use of the so-called ''inner interference formula,'' [15] according to which

$$W_x^2(t,f\,)=\int\int_{-\infty}^{+\infty}W_x\left(t+\frac{\tau}{2},\;f+\frac{\nu}{2}\right)W_x\left(t-\frac{\tau}{2},\;f-\frac{\nu}{2}\right)d\tau\;d\nu.$$

By making a change of variable and reorganizing terms, this leads to

$$\sigma_{tf}^2(W_x^2)=\frac{1}{8E_x^2}\int\int\int\int_{-\infty}^{+\infty}K(v,w,\xi,\zeta)W_x(v,\xi)W_x(w,\zeta)dv\;dw\;d\xi\;d\zeta,$$

with

$$K(v,w,\xi,\zeta)=v^2\xi^2+v^2\zeta^2+2v^2\xi\zeta+2vw\xi^2+4vw\xi\zeta.$$

The contribution of the first term of $K(v,w,\xi,\zeta)$ to the integral reduces to

$$\int\int\int\int_{-\infty}^{+\infty}v^2\xi^2W_x(v,\xi)W_x(w,\zeta)dv\;dw\;d\xi\;d\zeta=E_x^2\sigma_{tf}^2(W_x),$$

and it can be shown[4] that

$$\sigma_{tf}^2(W_x)=\frac{1}{4\pi^2}\left(\alpha_x-\frac{1}{2}\right).$$

The second term contributes as

$$\int\int\int\int_{-\infty}^{+\infty}v^2\zeta^2W_x(v,\xi)W_x(w,\zeta)dv\;dw\;d\xi\;d\zeta$$

$$=\int\int_{-\infty}^{+\infty}v^2\zeta^2|x(v)|^2|X(\zeta)|^2dv\;d\zeta=E_x^2\Delta t_x^2\Delta f_x^2,$$

whereas the third and fourth ones provide a null contribution because of the assumptions of mean time and mean frequency equal to zero.

Finally, the fifth term is such that

$$\int\int\int\int_{-\infty}^{+\infty}vw\xi\zeta W_x(v,\xi)W_x(w,\zeta)dv\;dw\;d\xi\;d\zeta=\frac{1}{4\pi^2}\,\mathrm{cov}_{tf}^2(x)\geqslant0,$$

with (following Cohen[3]), the ''covariance'' of $x(t)$ defined by (16).

Remark: If $x(t)$ is chosen to be $x_\gamma(t)$ as defined in (15), we get explicitly

$$\sigma_{tf}^2(W_{x_\gamma}^2)=\frac{1}{128\pi^2}\left[(\gamma-1)+\gamma^2\,\frac{\Gamma(3/\gamma)\Gamma(2-1/\gamma)}{\Gamma^2(1/\gamma)}\right],$$

and a numerical estimate of the γ_0 minimizing the spread measure gives $\gamma_0\approx 1.33$, with $\sigma_{tf}^2(W_{x_{\gamma 0}}^2)\approx 0.77/64\pi^2$. This is less than in the Gaussian case ($\gamma=2$) for which $\sigma_{tf}^2(W_{x_2}^2)=1/64\pi^2$.

III. COHEN'S CLASS

Some results pertaining to Wigner–Ville distributions can be generalized to members of *Cohen's class*, the general class of quadratic time–frequency energy distributions which are covariant to time and frequency shifts. By definition, Cohen's class can be written as[4]

$$C_x(t,f;\varphi)=\int\int_{-\infty}^{+\infty}\varphi(\xi,\tau)A_x(\xi,\tau)e^{-i2\pi(t\xi+f\tau)}d\xi\,d\tau, \tag{17}$$

where $\varphi(\xi,\tau)$ is some arbitrary kernel function and $A_x(\xi,\tau)$ is the so-called ''ambiguity function,'' obtained as a two-dimensional Fourier transform of the Wigner–Ville distribution and defined explicitly by

$$A_x(\xi,\tau)=\int_{-\infty}^{+\infty}x\left(t+\frac{\tau}{2}\right)\overline{x\left(t-\frac{\tau}{2}\right)}e^{i2\pi\xi t}\,dt.$$

Within this more general framework, Gaussian signals still appear to be central with respect to separability properties, as illustrated by the following claims.

Proposition 8: Smoothed pseudo-Wigner–Ville distributions—characterized by a separable kernel—are the only members of Cohen's class which are separable when applied to Gaussian signals.

Proof: Assuming that the analyzed signal $x(t)$ is a Gaussian, we know that its Wigner–Ville distribution $W_x(t,f)$ is separable, a property which carries over to its two-dimensional Fourier transform $A_x(\xi,\tau)$. Using (17), we get

$$\varphi(\xi,\tau)=\frac{1}{A_x(\xi,\tau)}\int\int_{-\infty}^{+\infty}C_x(t,f;\varphi)e^{i2\pi(t\xi+f\tau)}dt\,df,$$

and, if we require $C_x(t,f;\varphi)$ to be separable, it follows that the kernel function $\varphi(\xi,\tau)$ must itself be separable, a situation corresponding to distributions referred to as ''smoothed pseudo-Wigner–Ville distributions.'' [4]

Proposition 9: Smoothed pseudo-Wigner–Ville distributions—characterized by a separable kernel—are separable for Gaussian signals only.

Proof: The proof is very similar to the previous one since, according to (17), we can write

$$W_x(t,f)=\int\int_{-\infty}^{+\infty}\left(\frac{1}{\varphi(\xi,\tau)}\int\int_{-\infty}^{+\infty}C_x(t',f';\varphi)e^{i2\pi(t'\xi+f'\tau)}dt'\,df'\right)e^{-i2\pi(t\xi+f\tau)}d\xi\,d\tau.$$

Assuming therefore that both the kernel function $\varphi(\xi,\tau)$ and the associated smoothed Wigner–Ville distribution $C_x(t,f;\varphi)$ are separable, we get that the Wigner–Ville distribution itself is separable, which in turn implies that $x(t)$ is necessarily a Gaussian.

Proposition 10: When applied to Gaussian signals, spectrograms are separable if and only if their short-time window is Gaussian.

Proof: Assuming that the analyzed signal $x(t)$ is a Gaussian, we know that both its Wigner–Ville distribution $W_x(t,f)$ and its ambiguity function $A_x(\xi,\tau)$ are separable. If we now consider a spectrogram,

$$S_x^{(h)}(t,f)\equiv\left|\int_{-\infty}^{+\infty}x(s)\overline{h(s-t)}e^{-i2\pi fs}\,ds\right|^2=\left|\int_{-\infty}^{+\infty}X(\xi)\overline{H(\xi-f)}e^{i2\pi\xi t}\,d\xi\right|^2$$

having a short-time window $h(t)$ which is Gaussian, we get that the associated kernel function[4] $\varphi(\xi,\tau)=\overline{A_h(\xi,\tau)}$ is separable, and it follows that the spectrogram—as an inverse Fourier transform of $\varphi(\xi,\tau)A_x(\xi,\tau)$—is itself separable.

Conversely, if we suppose that a spectrogram is separable when applied to a Gaussian signal $g(t)$, we get from (17) [with $\varphi(\xi,\tau)=\overline{A_h(\xi,\tau)}$] that

$$\overline{A_h(\xi,\tau)}=\frac{1}{A_g(\xi,\tau)}\int\int_{-\infty}^{+\infty}S_g^{(h)}(t',f')e^{i2\pi(t'\xi+f'\tau)}dt'\,df'$$

is separable, with the consequence that $W_h(t,f)$ itself—as an inverse Fourier transform of $A_x(\xi,\tau)$—is separable and, hence, that the window $h(t)$ must be Gaussian.

More generally, we can remark that imposing the requirement of separability to spectrograms leads to specific constraints which have to be jointly satisfied by the analyzed signal $x(t)$ and the analysis window $h(t)$. Starting from the definition of a spectrogram of window $h(t)$ and assuming that this spectrogram is separable according to $S_x^{(h)}(t,f)=p(t)Q(f)$, we get

$$p(t)q(0)=\int_{-\infty}^{+\infty}S_x^{(h)}(t,f)df=\int_{-\infty}^{+\infty}|x(s)|^2|h(s-t)|^2ds$$

and

$$P(0)Q(f)=\int_{-\infty}^{+\infty}S_x^{(h)}(t,f)dt=\int_{-\infty}^{+\infty}|X(\xi)|^2|H(\xi-f)|^2d\xi.$$

Since $P(0)q(0)=E_xE_h$, separable spectrograms happen to be necessarily of the form

$$S_x^{(h)}(t,f)=\int\int_{-\infty}^{+\infty}\frac{|x(s)|^2|X(\xi)|^2}{E_x}\,\frac{|h(s-t)|^2|H(\xi-f)|^2}{E_h}\,ds\,d\xi,$$

a form which has to be compared to the (always valid) expression

$$S_x^{(h)}(t,f)=\int\int_{-\infty}^{+\infty}W_x(s,\xi)W_h(s-t,\xi-f)ds\,d\xi,$$

or even to

$$S_x^{(h)}(t,f)=\int\int_{-\infty}^{+\infty}C_x(s,\xi;\varphi)\overline{C_h(s-t,\xi-f;\varphi)}ds\,d\xi,$$

with the requirement that the parametrization function $\varphi(\xi,\tau)$ be such that $|\varphi(\xi,\tau)|=1$ (see, e.g., Ref. 4).

Reasoning along the same lines in the more general case of any distribution within Cohen's class, we can end up with the result that a separable distribution is necessarily such that

$$\begin{aligned}C_x(t,f;\varphi)&=\int\int_{-\infty}^{+\infty}W_x(s,\xi)\Pi(s-t,\xi-f)ds\,d\xi\\&=\int\int_{-\infty}^{+\infty}\frac{|x(s)|^2|X(\xi)|^2}{E_x}\,\frac{F(s-t,0)\psi(0,\xi-f)}{\varphi(0,0)}\,ds\,d\xi,\end{aligned}\tag{18}$$

where $\Pi(t,f)$ stands for the two-dimensional Fourier transform of $\varphi(\xi,\tau)$, and $F(t,\tau)$ and $\psi(\xi,\nu)$ for one-dimensional partial Fourier transforms of the same quantity, over frequency and time, respectively. This leads to the following.

Proposition 11: If a Cohen's class distribution has correct marginals and is separable, it is positive and factorizable as

$$C_x(t,f;\varphi)=\frac{|x(t)|^2|X(f)|^2}{E_x}. \tag{19}$$

Proof: This can be proved either directly—as for Proposition 2—or by using (18), since we know[4] that having correct marginals implies that $\varphi(\xi,0)=\varphi(0,\tau)=1$.

Remark: Given *any* signal $x(t)$, one could imagine to formally obtain a distribution factorizable as in (19) by choosing for the parametrization function the (signal-dependent) quantity $\varphi(\xi,\tau)\equiv E_x\gamma_X(\xi)\gamma_x(\tau)/A_x(\xi,\tau)$, with the γ's defined as in Proposition 3. In the general case, this would unfortunately correspond to pathological φ's, with singularities at those points where the ambiguity function vanishes while the individual time and frequency correlation functions do not. Regular solutions can be guaranteed by imposing to the ambiguity function to be everywhere positive but, as mentioned in Ref. 18, imposing positivity of the ambiguity function is equivalent to imposing positivity of the Wigner–Ville distribution, with Gaussians as only solutions.

Proposition 12: The Wigner–Ville distribution is the only member of Cohen's class which both has correct marginals and is separable when applied to Gaussian signals.

Proof: By Fourier transformation, (19) becomes $\varphi(\xi,\tau)A_x(\xi,\tau)=E_x\gamma_X(\xi)\gamma_x(\tau)$. Assuming that $x(t)$ is a Gaussian signal, this simplifies to $\varphi(\xi,\tau)=1$, thus defining the Wigner–Ville distribution as the only solution.

Proposition 13 (Ref. 4): For any signal $x(t)$ whose mean time and mean frequency are zero, and any $T>0$, distributions of Cohen's class satisfy the time–frequency uncertainty relation

$$\Sigma_{tf}^2(C_x)\geqslant\frac{1}{2\pi}-\frac{1}{4\pi^2}\left[\frac{1}{T^2}\frac{\partial^2\varphi}{\partial\xi^2}(0,0)+T^2\frac{\partial^2\varphi}{\partial\tau^2}(0,0)\right],$$

with equality if and only if $x(t)$ is a Gaussian signal.

Proof: The proof is straightforward and follows directly from the fact that $C_x(t,f;\varphi)$ has for Fourier transform $\varphi(\xi,\tau)A_x(\xi,\tau)$. Defining

$$\Sigma_t^2(C_x)\equiv\frac{1}{E_x}\int_{-\infty}^{+\infty}t^2\left(\int_{-\infty}^{+\infty}C_x(t,f;\varphi)df\right)dt,$$

we get

$$\begin{aligned}\Sigma_t^2(C_x)&=\frac{1}{E_x}\int_{-\infty}^{+\infty}|x(s)|^2\left(\int_{-\infty}^{+\infty}t^2F(s-t,0)dt\right)ds\\&=\frac{1}{E_x}\int_{-\infty}^{+\infty}|x(s)|^2\left(\int_{-\infty}^{+\infty}(s^2-2s\theta+\theta^2)F(\theta,0)d\theta\right)ds=\varphi(0,0)\Delta t_x^2-\frac{1}{4\pi^2}\frac{\partial^2\varphi}{\partial\xi^2}(0,0),\end{aligned}$$

and, similarly,

$$\Sigma_f^2(C_x)\equiv\frac{1}{E_x}\int_{-\infty}^{+\infty}f^2\left(\int_{-\infty}^{+\infty}C_x(t,f;\varphi)dt\right)df=\varphi(0,0)\Delta f_x^2-\frac{1}{4\pi^2}\frac{\partial^2\varphi}{\partial\tau^2}(0,0).$$

Assuming that $C_x(t,f;\varphi)$ is a valid energy distribution [i.e., that $\varphi(0,0)=1$], we are led to

$$\Sigma_{tf}^2(C_x)=\frac{\Sigma_t^2(C_x)}{T^2}+T^2\Sigma_f^2(C_x)=\Sigma_{tf}^2(W_x)-\frac{1}{4\pi^2}\left[\frac{1}{T^2}\frac{\partial^2\varphi}{\partial\xi^2}(0,0)+T^2\frac{\partial^2\varphi}{\partial\tau^2}(0,0)\right],$$

and therefore to the claim, from the result established previously for $\Sigma_{tf}^2(W_x)$.

Proposition 14: For any signal $x(t)$ whose mean time and mean frequency are zero, and any $T>0$, spectrograms with a window $h(t)$ satisfy the time–frequency uncertainty relation,

$$\Sigma_{tf}^2(S_x^{(h)})=\Sigma_{tf}^2(W_x)+\Sigma_{tf}^2(W_h)\geqslant\frac{1}{\pi},$$

with equality, if and only if $x(t)$ and $h(t)$ are matched Gaussian signals.

Proof: This can be viewed as a Corollary to the previous Proposition since, in the case of spectrograms, we get

$$-\frac{1}{4\pi^2}\left[\frac{1}{T^2}\frac{\partial^2\varphi}{\partial\xi^2}(0,0)+T^2\frac{\partial^2\varphi}{\partial\tau^2}(0,0)\right]=\frac{1}{T^2}\frac{\partial^2\overline{A_h}}{\partial\xi^2}(0,0)+T^2\frac{\partial^2\overline{A_h}}{\partial\tau^2}(0,0),$$

$$=\Sigma_{tf}^2(W_h)$$

and hence

$$\Sigma_{tf}^2(S_x^{(h)})=\Sigma_{tf}^2(W_x)+\Sigma_{tf}^2(W_h).$$

The total time–frequency spread being the sum of two independent positive quantities, it is minimized when each of these quantities is itself minimized, which is obtained when both $x(t)$ and $h(t)$ are Gaussian signals (see Proposition 4). In order to ensure that the two individual minimizations are compatible, the same arbitrary T has to be chosen for both $\Sigma_{tf}^2(W_x)$ and $\Sigma_{tf}^2(W_h)$. This leads to

$$T=\sqrt{\frac{\Delta t_x}{\Delta f_x}}=\sqrt{\frac{\Delta t_h}{\Delta f_h}},$$

and, hence, to $\Delta t_h=\Delta t_x$ and $\Delta f_h=\Delta f_x$, since $x(t)$ and $h(t)$ being Gaussians, $\Delta t_x\Delta f_x=\Delta t_h\Delta f_h=1/4\pi$.

Remark: Although the derivation is different, this result is equivalent to the one given by Janssen in Ref. 19.

Proposition 15: For any signal $x(t)$ whose mean time and mean frequency are zero, and any $T>0$, spectrograms with a real-valued and even window $h(t)$ are such that

$$\sigma_{tf}^2(S_x^{(h)})=\frac{1}{4\pi^2}[(\alpha_x+\alpha_h-1)+4\pi^2(\Delta t_x^2\Delta f_h^2+\Delta t_h^2\Delta f_x^2)].$$

Proof: Let us introduce the time–frequency spread measure,

$$\sigma_{tf}^2(C_x)\equiv\frac{1}{E_x}\int\int_{-\infty}^{+\infty}t^2f^2C_x(t,f;\varphi)dt\,df, \tag{20}$$

where $C_x(t,f;\varphi)$ stands for any distribution within Cohen's class. This spread measure (20) can be equivalently expressed as

$$\sigma_{tf}^2(C_x)=\frac{1}{16\pi^4E_x}\frac{\partial^4}{\partial\xi^2\partial\tau^2}(\varphi(\xi,\tau)A_x(\xi,\tau))\Big|_{(0,0)},$$

with

$$\frac{\partial^4}{\partial\xi^2\,\partial\tau^2}(\varphi A_x)=\frac{\partial^4\varphi}{\partial\xi^2\,\partial\tau^2}A_x+\frac{\partial^2\varphi}{\partial\tau^2}\frac{\partial^2A_x}{\partial\xi^2}+\frac{\partial^2\varphi}{\partial\xi^2}\frac{\partial^2A_x}{\partial\tau^2}+f\frac{\partial^4A_x}{\partial\xi^2\,\partial\tau^2}$$
$$+2\left(\frac{\partial^3\varphi}{\partial\xi\,\partial\tau^2}\frac{\partial A_x}{\partial\xi}+\frac{\partial^3\varphi}{\partial\xi^2\,\partial\tau}\frac{\partial A_x}{\partial\tau}+\frac{\partial\varphi}{\partial\tau}\frac{\partial^3A_x}{\partial\xi^2\,\partial\tau}+\frac{\partial\varphi}{\partial\xi}\frac{\partial^3A_x}{\partial\tau\,\partial\tau^2}\right).$$

This holds for any distribution within Cohen's class and, in particular, for spectrograms $S_x^{(h)}(t,f)$ with a window $h(t)$. Given such a window, (normalized) spectrograms are known to correspond to the kernel function $\varphi(\xi,\tau)=\overline{A_h(\xi,\tau)}/E_h$. Assuming, furthermore, that the window is real valued and even, we can show that

$$\frac{\partial A_h}{\partial\xi}(0,0)=\frac{\partial A_h}{\partial\tau}(0,0)=0$$

and

$$\frac{\partial^3 A_h}{\partial\xi^2\,\partial\tau}(0,0)=\frac{\partial^3 A_h}{\partial\xi\,\partial\tau^2}(0,0)=0,$$

whereas, for any signal $x(t)$,

$$\frac{\partial^2 A_x}{\partial\xi^2}(0,0)=4\pi^2 E_x\Delta t_x^2,$$

$$\frac{\partial^2 A_x}{\partial\tau^2}(0,0)=4\pi^2 E_x\Delta f_x^2,$$

and

$$\frac{\partial^4 A_x}{\partial\xi^2\,\partial\tau^2}(0,0)=4\pi^2 E_x\left(\alpha_x-\frac{1}{2}\right).$$

It thus follows that, in the case of spectrograms, the overall spread measure (20) reads as announced.

Remark: Positivity of spectrograms guarantees that $\sigma_{tf}^2(S_x^{(h)})\geqslant 0$. In the special case where $h(t)=x(t)$ (matched spectrograms), we get $2\alpha_x-1+4\pi^2(2\Delta t_x^2\Delta f_x^2)\geqslant 0$ and therefore $\alpha_x\geqslant 1/2-4\pi^2\Delta t_x^2\Delta f_x^2\geqslant 1/4$, as already proved in Proposition 6.

IV. AFFINE TIME–FREQUENCY DISTRIBUTIONS

The Wigner–Ville distribution and, more generally, all members of Cohen's class, all belong to a family of time–frequency distributions which are covariant with respect to shifts in time and frequency. Moreover, assuming that an energy distribution depends quadratically on the signal, it can be shown that covariance with respect to shifts in time and frequency is sufficient to reduce the class of admissible solutions to Cohen's class. Following this line, it becomes therefore possible to generate other classes of distributions by modifying in a suitable manner the covariance requirement. Among the various solutions which have been derived this way, a prominent example is given by the class of the so-called *affine* time–frequency distributions, whose construction relies basically on the consideration of dilations in place of frequency shifts.

According to Ref. 20, the general class of affine time–frequency distributions $P_X^{(k)}(t,f)$ which are covariant with respect to each solvable three-parameter extension of the affine group is defined on $\mathbf{R}\times\mathbf{R}_+$ and can be parametrized by

$$P_X^{(k)}(t,f)=f^{2(r+1)-q}\int_{-\infty}^{+\infty}\mu_k(u)X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)}e^{-i2\pi tf\zeta_k(u)}\,du, \tag{21}$$

with $\zeta_k(u)=\lambda_k(u)-\lambda_k(-u)$, and where r and q are real-valued free parameters. In this expression, $X(f)$ stands as before for the Fourier spectrum of the analyzed signal, but only positive frequencies are considered. The parametrization function $\lambda_k(u)$ is given explicitly by

$$\lambda_k(u)=\left(k\,\frac{e^{-u}-1}{e^{-ku}-1}\right)^{1/(k-1)}, \tag{22}$$

with $k\in\mathbf{R}$, whereas $\mu_k(u)$ is a weight function whose value can be made dependent on $\lambda_k(u)$ so as to guarantee specific requirements for $P_X^{(k)}(t,f)$.[20]

From (21), it is apparent that affine distributions depend naturally on the product variable tf, a quantity which can be interpreted as a Mellin variable. To justify this point, let us recall that a definition of the Mellin transform $\underline{X}(s)$ of $X(f)$ can be given by[21]

$$\underline{X}(s)\equiv\int_0^{+\infty}X(f)f^{i2\pi s+r}\,df,$$

thus corresponding to an expansion of the spectrum $X(f)$ on ''hyperbolic chirps,''

$$C_0(f) \equiv f^r e^{-i2\pi s \log f}, \quad f \geqslant 0,$$

whose group delay is

$$t_{C_0}(f) = -\frac{1}{2\pi}\frac{d}{df}(-2\pi s \log f) = \frac{s}{f}.$$

In this picture, the Mellin variable s appears therefore as a *hyperbolic chirp rate* related to time and frequency by a relation of the type $s = tf$.

A companion interpretation amounts to consider the Mellin variable s as *a scale* variable.[22,3] This comes from the fact that, choosing, e.g., Weyl's correspondence rule, the operator $\mathcal{S}$ associated to $s = tf$ can be written as

$$\mathcal{S} = \frac{\mathcal{T}\mathcal{F} + \mathcal{F}\mathcal{T}}{2},$$

where $\mathcal{T}$ and $\mathcal{F}$ stand for the time and frequency operators, defined, respectively, by $(\mathcal{T}x)(t) = tx(t)$ and $(\mathcal{F}x)(t) = (-i/2\pi)\dot{x}(t)$. It thus follows that, for any $\lambda > 0$,

$$(e^{i2\pi\lambda S}X)(f) = e^{-\lambda/2}X(e^{-\lambda}f),$$

which allows us to interpret $\mathcal{S}$ as the infinitesimal generator of a scaling operator (whose eigenfunctions are precisely hyperbolic chirps).

Whatever the interpretation, and while $P_X^{(k)}(t,f)$ is usually considered as a time–frequency distribution, it can equally well be reparametrized as

$$\tilde{P}_X^{(k)}(s,f) \equiv P_X^{(k)}\left(\frac{s}{f}, f\right), \tag{23}$$

so that it becomes a function of (Mellin) scale and (Fourier) frequency variables.

A. Minimum uncertainty

Because the operators attached to scale and frequency do not commute, namely

$$[\mathcal{S},\mathcal{F}] \equiv \mathcal{S}\mathcal{F} - \mathcal{F}\mathcal{S} = \frac{i}{2\pi}\mathcal{F},$$

the corresponding variables obey a specific form of uncertainty relation. Precisely, let $X(f)$ be an analytic signal such that

$$E_X \equiv \int_0^{+\infty} |X(f)|^2 f^{2r+1} df < +\infty,$$

$$\frac{1}{E_X}\int_{-\infty}^{+\infty} s|\underline{X}(s)|^2 ds = s_0,$$

and

$$\frac{1}{E_X}\int_0^{+\infty} f|X(f)|^2 df = f_0.$$

If we denote by Δs_X and Δf_X its mean-square deviations in scale and frequency, defined, respectively, by

$$\Delta s_X^2 \equiv \frac{1}{E_X} \int_{-\infty}^{+\infty} (s-s_0)^2 |\underline{X}(s)|^2 \, ds \tag{24}$$

and

$$\Delta f_X^2 \equiv \frac{1}{E_X} \int_0^{+\infty} (f-f_0)^2 |X(f)|^2 f^{2r+1} \, df, \tag{25}$$

we can prove the following.[3,21–23]

Proposition 16: Scale and frequency satisfy the uncertainty relation

$$\Delta s_X \Delta f_X \geqslant \frac{f_0}{4\pi}, \tag{26}$$

with equality if and only if $X(f)$ *is a ''Klauder wavelet'' of the form*

$$X(f) = C f^{\alpha} e^{-\beta f} f^{i\gamma} U(f), \tag{27}$$

with $C \in \mathbf{C}$, $(\alpha, \beta, \gamma) \in \mathbf{R}_+ \times \mathbf{R}_+ \times \mathbf{R}$ *and* $U(\cdot)$ *the unit step function.*

Proof: In order to get simplified notations, let us introduce the quantities $\mathcal{S}_0 \equiv \mathcal{S} - s_0 \mathcal{I}$ and $\mathcal{F}_0 \equiv \mathcal{F} - f_0 \mathcal{I}$, with $\mathcal{I}$ the identity operator. It follows from the unitarity of the Mellin transform that

$$\Delta s_X^2 = \int_0^{+\infty} |(S_0 X)(f)|^2 f^{2r+1} \, df,$$

so that we can write, by using the Cauchy–Schwarz inequality,

$$\begin{aligned} E_X^2 \Delta s_X^2 \Delta f_X^2 &= \int_0^{+\infty} |(S_0 X)(f)|^2 f^{2r+1} \, df \int_0^{+\infty} |(\mathcal{F}_0 X)(f)|^2 f^{2r+1} \, df \\ &\geqslant \left| \int_0^{+\infty} \overline{(S_0 X)(f)} (\mathcal{F}_0 X)(f) f^{2r+1} \, df \right|^2 = \left| \int_0^{+\infty} \overline{X(f)} (S_0 \mathcal{F}_0 X)(f) f^{2r+1} \, df \right|^2. \end{aligned}$$

Defining the anticommutator of any two operators $\mathcal{A}$ and $\mathcal{B}$ as $[\mathcal{A}, \mathcal{B}]_+ \equiv \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}$ and remarking that $[\mathcal{S}_0, \mathcal{F}_0] = [\mathcal{S}, \mathcal{F}]$, we get, in the specific case of scale and frequency, that

$$\mathcal{S}_0 \mathcal{F}_0 = \tfrac{1}{2}([\mathcal{S}_0, \mathcal{F}_0] + [\mathcal{S}_0, \mathcal{F}_0]_+) = \tfrac{1}{2}\left(\frac{i}{2\pi} \mathcal{I} + [\mathcal{S}_0, \mathcal{F}_0]_+ \right),$$

and it follows that

$$\Delta s_X^2 \Delta f_X^2 \geqslant \frac{f_0^2}{16\pi^2},$$

with equality if

$$[\mathcal{S}_0, \mathcal{F}_0]_+ = 0, \tag{28}$$

and for those signals $X(f)$ which satisfy, for some $\lambda \in \mathbf{C}$,

$$(\mathcal{S}_0 X)(f) = \lambda (\mathcal{F}_0 X)(f). \tag{29}$$

Plugging (29) into (28) yields that λ is necessarily purely imaginary (namely that $\lambda^2 = -\Delta s_X^2 / \Delta f_X^2$). By developing explicitly Eq. (29) which characterizes the signals with minimum uncertainty in scale and frequency, we end up with the differential equation

$$2 f \dot{X}(f) = [4\pi \, \mathrm{Im}\{\lambda\}(f - f_0) - 1 - i 4\pi s_0] X(f),$$

whose solution is given by

$$X(f)=K\ \exp(2\pi\ \mathrm{Im}\{\lambda\}f-(2\pi\ \mathrm{Im}\{\lambda\}f_0+\tfrac{1}{2}+i2\pi s_0)\log f),$$

with $K\in\mathbf{C}$ and $\mathrm{Im}\{\lambda\}<0$, so that the signal is of finite energy. This solution, which is only defined for positive frequencies, is of the form (27), and this concludes the proof.

In order to get an analog of Proposition 4 in the case of Mellin–Fourier variables, it is natural to consider the *unitary* Bertrand distribution $B_X(t,f)\equiv P_X^{(0)}(t,f)$, which, in many respects, is the counterpart of the Wigner–Ville distribution in the Mellin–Fourier domain. By definition, the unitary Bertrand distribution (also referred to as the *tomographic* distribution[20]) is obtained from (21) by setting $k=0$ in (22), and it reads explicitly as

$$B_X(t,f)=f^{2(r+1)-q}\int_{-\infty}^{+\infty}\left(\frac{u/2}{\sinh(u/2)}\right)^{2(r+1)}X\left(\frac{(u/2)e^{u/2}f}{\sinh(u/2)}\right)\overline{X\left(\frac{(u/2)e^{-u/2}f}{\sinh(u/2)}\right)}e^{i2\pi ftu}\,du. \quad (30)$$

With this definition, we are led to the following.

Proposition 17: Given any $S>0$, the unitary Bertrand distribution satisfies the Mellin-frequency uncertainty relation,

$$\Sigma_{sf}^2(\tilde{B}_X)\equiv\frac{1}{E_X}\int_{-\infty}^{+\infty}\int_0^{+\infty}\left[\left(\frac{s-s_0}{S}\right)^2+S^2\left(\frac{f}{f_0}-1\right)^2\right]\tilde{B}_X(s,f)f^{q-1}ds\ df\geqslant\frac{1}{2\pi},$$

with equality, if and only if $X(f)$ is a Klauder wavelet.

Proof: The proof relies on marginal properties of the unitary Bertrand distribution, according to which

$$\int_{-\infty}^{+\infty}\tilde{B}_X(s,f)ds=f^{2(r+1)-q}|X(f)|^2$$

and

$$\int_0^{+\infty}\tilde{B}_X(s,f)f^{q-1}\,df=|\underline{X}(s)|^2.$$

Using the definitions (24) and (25), it follows directly from these relations that

$$\Sigma_{sf}^2(\tilde{B}_X)=\frac{\Delta s_X^2}{S^2}+S^2\,\frac{\Delta f_X^2}{f_0^2}=\left(\frac{\Delta s_X}{S}-S\,\frac{\Delta f_X}{f_0}\right)^2+2\,\frac{\Delta s_X\Delta f_X}{f_0}\geqslant 2\,\frac{\Delta s_X\Delta f_X}{f_0},$$

with equality if S is ''matched'' to the signal according to $S=\sqrt{f_0\Delta s_X/\Delta f_X}$. In this case, we know from (26) that the product $\Delta s_X\Delta f_X$ is lower bounded too, with equality in the only case of Klauder wavelets, whence the result.

B. Separability

The question we are interested in here can be stated as follows: is it possible to get affine distributions (23) which are separable in their Mellin and Fourier variables for some signal $X(f)$, i.e., such that they can be put in the form

$$\tilde{P}_X^{(k)}(s,f)=p(s)Q(f), \quad (31)$$

where $p(s)$ and $Q(f)$ are functions to be determined?

A first answer to this question can be looked after in the spirit of Proposition 3, which has been established thanks to the correct marginal properties of the Wigner–Ville distribution, in both time and frequency. In the specific case of the unitary Bertrand distribution (30), we have, in fact,[20] the two marginal properties

$$\int_{-\infty}^{+\infty} B_X(t,f)\,dt = f^{2r+1-q}|X(f)|^2 \tag{32}$$

and

$$\int_{-\infty}^{+\infty}\int_{0}^{+\infty} B_X(t,f)\,\delta\left(t-\frac{s}{f}\right) f^{q-1}\,dt\,df = |\underline{X}(s)|^2. \tag{33}$$

Within this new context, and considering first the frequency marginal property (32), it is easy to show that the separability condition (31) leads to

$$\begin{aligned}|X(f)|^2 &= f^{q-2r-1}\int_{-\infty}^{+\infty} B_X(t,f)\,dt \\ &= f^{q-2(r+1)}\int_{-\infty}^{+\infty}\widetilde{B}(s,f)\,\frac{ds}{f} = f^{q-2(r+1)}\int_{-\infty}^{+\infty} p(s)Q(f)\,\frac{ds}{f} = f^{q-2(r+1)}P(0)\,\frac{Q(f)}{f}\geqslant 0,\end{aligned}$$

whereas, by using the Mellin marginal (33), we obtain, in a similar way,

$$\begin{aligned}|\underline{X}(s)|^2 &= \int_{-\infty}^{+\infty}\int_{0}^{+\infty} B_X(t,f)\,\delta\left(t-\frac{s}{f}\right) f^{q-1}dt\,df \\ &= \int_{0}^{+\infty}\widetilde{B}_X(s,f)f^{q-1}\,df = p(s)\int_{0}^{+\infty} Q(f)f^{q-1}\,df \geqslant 0.\end{aligned}$$

Combining these two marginals, we get

$$P(0)\int_{0}^{+\infty} Q(f)f^{q-1}\,df = E_X \geqslant 0,$$

with the consequence that

$$\widetilde{B}_X(s,f) = \frac{|\underline{X}(s)|^2 f^{2(r+1)-q}|X(f)|^2}{E_X} \geqslant 0.$$

We could therefore think of having proved that, as for Wigner–Ville, separability implies positivity. The situation is, however, different since, unfortunately, the assumption of separability can be proved not to be admissible for unitary Bertrand distributions. This is evidenced by the following.

Proposition 18: The only affine distributions which are compatible with separability are characterized by (21) and (22) with $k=-1$, $1/2$, *or* 2.

Proof: Assuming that separability holds, we can take the Fourier transform of (31) with respect to s, and we get

$$\begin{aligned}P(\sigma)Q(f) &= \int_{-\infty}^{+\infty}\widetilde{P}_X^{(k)}(s,f)e^{-i2\pi\sigma s}\,ds \\ &= f^{2(r+1)-q}\int_{-\infty}^{+\infty}\mu_k(u)X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)}\,\delta(\sigma-\zeta_k(u))du \\ &= f^{2(r+1)-q}\mu_k(\zeta_k^{-1}(\sigma))w(\sigma)X(\lambda_k(\zeta_k^{-1}(\sigma))f)\overline{X(\lambda_k(-\zeta_k^{-1}(\sigma))f)},\end{aligned}$$

with $w(\sigma)=(d/d\sigma)\zeta_k^{-1}(\sigma)$. Setting $\zeta_k^{-1}(\sigma)=u$, we end up with

$$P(\zeta_k(u))Q(f) = f^{2(r+1)-q}\mu_k(u)w(\zeta_k(u))X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)},$$

a relation which implies that

$$X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)}=\frac{P(\zeta_k(u))}{\mu_k(u)w(\zeta_k(u))}\,\frac{Q(f)}{f^{2(r+1)-q}}\equiv\tilde{P}(u)\tilde{Q}(f). \tag{34}$$

We know from Ref. 20 that $\lambda_k(0)=1$. Imposing $u=0$ in (34), we get therefore $|X(f)|^2=\tilde{P}(0)\tilde{Q}(f)$ and, hence,

$$\frac{X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)}}{|X(f)|^2}=\frac{\tilde{P}(u)}{\tilde{P}(0)}. \tag{35}$$

This states that the left-hand side of (35) is necessarily a function of u only. In order to find solutions (and conditions for their existence) to the functional equation (35), we can factor $X(f)$ in a (non-negative) modulus term and a phase term according to $X(f)=|X(f)|\exp\{i\arg X(f)\}\equiv\exp\{\Phi(f)+i\Psi(f)\}$, thus leading to

$$\frac{X(\lambda_k(u)f)\overline{X(\lambda_k(-u)f)}}{|X(f)|^2}=e^{\{\Phi(\lambda_k(u)f)+\Phi(\lambda_k(-u)f)-2\Phi(f)]+i[\Psi(\lambda_k(u)f)-\Psi(\lambda_k(-u)f)\}}.$$

Within this formulation, imposing (35) to hold requires that both the real and imaginary parts of the above exponential do not depend upon f. The condition on the phase $\Psi(f)$ is

$$\frac{\partial}{\partial f}[\Psi(\lambda_k(u)f)-\Psi(\lambda_k(-u)f)]=0,$$

or, equivalently [by letting $\lambda_k(u)f=\xi$ and $\lambda_k(-u)f=\zeta$], $\xi\dot{\Psi}(\xi)=\zeta\dot{\Psi}(\zeta)$ for any $(\xi,\zeta)\in\mathbf{R}_+^2$. This does not impose any restriction on the admissible distributions but constrains the phase to be of the form

$$\Psi(f)=a\,\log f+b, \tag{36}$$

with $(a,b)\in\mathbf{R}^2$. Concerning the modulus, the condition can be written as

$$\frac{\partial}{\partial f}[\Phi(\lambda_k(u)f)+\Phi(\lambda_k(-u)f)-2\Phi(f)]=0,$$

or, equivalently,

$$\dot{\Phi}(f)=\frac{\lambda_k(u)\dot{\Phi}(\lambda_k(u)f)+\lambda_k(-u)\dot{\Phi}(\lambda_k(-u)f)}{2}.$$

This expression can be rewritten by introducing, as before, $\lambda_k(u)f=\xi$ and $\lambda_k(-u)f=\zeta$, thus defining f as the *Stolarsky's generalized mean* of ξ and ζ,[24] whose expression reads explicitly as $f=\Theta(\xi,\zeta)$, with

$$\Theta(\xi,\zeta)=\left(\frac{1}{k}\,\frac{\xi^k-\zeta^k}{\xi-\zeta}\right)^{1/(k-1)}. \tag{37}$$

What we get is $\Theta(\xi,\zeta)\dot{\Phi}(\Theta(\xi,\zeta))=(\xi\dot{\Phi}(\xi)+\zeta\dot{\Phi}(\zeta))/2$, and finally, if we let

$$m(f)\equiv f\dot{\Phi}(f), \tag{38}$$

we end up with $m(\Theta(\xi,\zeta))=(m(\xi)+m(\zeta))/2$ or, equivalently,

$$\Theta(\xi,\zeta)=m^{-1}\left(\frac{m(\xi)+m(\zeta)}{2}\right).$$

We recognize therefore that the Stolarsky's generalized mean (37) must necessarily be of the form of a *quasiarithmetic generalized mean* (in the sense of Kolmogorov and Nagumo), a situation we know to be true only if $k=-1$, 1/2, or 2, as proved in (Ref. 25, Appendix A). This concludes the proof.

It appears therefore that the unitary Bertrand distribution ($k=0$) is excluded from the class of affine distributions which may be separable in their Fourier and Mellin variables. The three cases mentioned in the Proposition can, however, be studied in further detail by specifying the corresponding generalized means and the associated distributions, and by identifying the signals which guarantee separability.

1. Case k=2

The generalized mean (37) reduces in this case to the arithmetic mean[25]

$$\Theta(\xi,\zeta)=\frac{\xi+\zeta}{2}\Leftrightarrow m(f)=af+b,$$

and, according to (38), $\dot{\Phi}(f)=a+b/f$, which leads to $\Phi(f)=af+b\log f+c$. Together with (36), the associated class of signals is therefore of the form

$$X_2(f)=Cf^{\alpha}e^{-\beta f}f^{i\gamma}U(f),$$

with $C\in\mathbf{C}$, $(\alpha,\beta,\gamma)\in\mathbf{R}^3$, and $U(\cdot)$ the unit step function, thus corresponding to the family of the ''Klauder wavelets'' (27) when α and β are both positive.

Assuming that $X(f)$ is analytic, i.e., vanishes for negative frequencies, it is known[20,25] that the case ($k=2$, $q=0$, $r=-1/2$) corresponds to the Wigner–Ville distribution,

$$W_X(t,f)=\int_{-2f}^{+2f}X\left(f+\frac{\nu}{2}\right)\overline{X\left(f-\frac{\nu}{2}\right)}e^{i2\pi\nu t}\,d\nu,$$

for a suitable choice of the weighting function $\mu_2(u)$, namely $\mu_2(u)=\cosh^{-2}(u/2)$. In this case, a direct calculation shows that

$$\widetilde{W}_{X_2}(s,f)=2|C|^2f^{2\alpha+1}e^{-2\beta f}\int_{-1}^{+1}(1-w^2)^{\alpha}\left(\frac{1+w}{1-w}\right)^{i\gamma}e^{i4\pi sw}\,dw\;U(f).$$

One can remark that, in the real case where $\gamma=0$, this result can be given a closed form expression which reads as

$$\widetilde{W}_{X_2}(s,f)=2|C|^2\sqrt{\pi}\Gamma(\alpha+1)f^{2\alpha+1}e^{-2\beta f}\,\frac{J_{\alpha+1/2}(4\pi|s|)}{(2\pi|s|)^{\alpha+1/2}}\,U(f),$$

where $J_{\nu}(\cdot)$ stands for the Bessel function of the first kind.[26]

Remark: Whereas in this special case, $\widetilde{W}_{X_2}(s,f)$ is separable and is of the form $\widetilde{W}_{X_2}(s,f)=k(s)f|X(f)|^2$, it can be checked that $k(s)$ is neither proportional to the Mellin density $|\underline{X}(s)|^2$ nor positive.

2. Case k=−1

The generalized mean (37) reduces in this case to the geometric mean,[25]

$$\Theta(\xi,\zeta)=\sqrt{\xi\zeta}\Leftrightarrow m(f)=a\,\log f+b,$$

and, according to (38), $\dot{\Phi}(f)=a(\log f)/f+b/f$, which leads to $\Phi(f)=(a/2)\log^2 f+b\log f+c$. Together with (36), the associated class of signals reads as

$$X_{-1}(f)=Cf^{\alpha}e^{-\beta\log^2 f}f^{i\gamma}U(f), \tag{39}$$

and it generalizes a family of wave forms defined by Altes[27] (the special case of the above expression with $\alpha=0$ and $\beta>0$).

It is known[20,25] that the case ($k=-1$, $q=0$, $r=-1/2$) corresponds to the so-called *Unterberger distribution*, which, in its ''active form'' [i.e., when setting $\mu(u)=\cosh(u/2)$[20]], reads as

$$\tilde{U}_X(s,f)=f\int_{-\infty}^{+\infty}\left(1+\frac{1}{\gamma^2}\right)X(\gamma f)\overline{X\left(\frac{f}{\gamma}\right)}e^{i2\pi(\gamma-1/\gamma)s}\,d\gamma. \tag{40}$$

A direct calculation shows that

$$\tilde{U}_{X_{-1}}(s,f)=|C|^2f^{2\alpha+1}e^{-2\beta(\log^2 f+\gamma^2/4)}\int_{-\infty}^{+\infty}e^{-2\beta(a\,\sinh w-i(\gamma/2))^2}e^{i4\pi sw}\,dw\;U(f).$$

Remark: The family of wave forms (39) has been introduced by Altes in a context of active sonar, such as encountered in bat echolocation. They are, in fact, solutions to the problem of ''Doppler tolerance,'' which consists in estimating at best (no bias and the highest signal-to-noise ratio) a time delay in the presence of some unknown Doppler shift.

3. Case k=1/2

The generalized mean (37) reduces in this case to the square-root mean[25]

$$\Theta(\xi,\zeta)=\left(\frac{\sqrt{\xi}+\sqrt{\zeta}}{2}\right)^2\Leftrightarrow m(f)=a\sqrt{f}+b,$$

and, according to (38), $\dot{\Phi}(f)=a/\sqrt{f}+b/f$, which leads to $\Phi(f)=2a\sqrt{f}+b\,\log f+c$. Together with (36), the associated class of signals reads as

$$X_{1/2}(f)=Cf^{\alpha}e^{-\beta\sqrt{f}}f^{i\gamma}U(f).$$

It is known[20,25] that the case ($k=1/2$, $q=0$, $r=-1/2$) comprises the so-called ''*D* distribution:''

$$\tilde{D}_X(s,f)=f\int_{-4}^{+4}\left[1-\left(\frac{\gamma}{4}\right)^2\right]X\left(\left(1+\frac{\gamma}{4}\right)^2f\right)\overline{X\left(\left(1-\frac{\gamma}{4}\right)^2f\right)}e^{i2\pi s\gamma}d\gamma,$$

and a direct calculation shows that

$$\tilde{D}_{X_{1/2}}(s,f)=4|C|^2f^{2\alpha+1}e^{-2\beta\sqrt{f}}\int_{-1}^{+1}(1-w^2)^{\alpha}\left(\frac{1+w}{1-w}\right)^{i\gamma}e^{i8\pi sw}\,dw\;U(f).$$

If $\gamma=0$, this result can be simplified to

$$\tilde{D}_{X_{1/2}}(s,f)=4|C|^2\sqrt{\pi}\Gamma(\alpha+1)f^{2\alpha+1}e^{-2\beta\sqrt{f}}\,\frac{J_{\alpha+1/2}(8\pi|s|)}{(4\pi|s|)^{\alpha+1/2}}\,U(f).$$

C. Positivity

It has already been mentioned that Wigner–Ville distributions attain generally negative values, but that they are everywhere positive when applied to a Gaussian signal and, more generally, to the exponential of a quadratic polynomial.[6] It is our purpose in this section to point out related results in the case of affine time–frequency distributions.

1. Unitary Bertrand distributions

We will first show that, in the case of the (unitary) Bertrand distribution (30), ''Klauder wavelets'' play a role which has much to share with the one played by Gaussians in the Wigner–Ville case. To this end, it is convenient to generalize and reparametrize the Klauder wavelet (27) as

$$X_2(f) \equiv C f^{-(r+1)} e^{2\pi\lambda(f_0 \log f - f)} e^{-i2\pi(\beta_0 \log f + \xi f)} U(f), \tag{41}$$

with $C \in \mathbf{C}$, and r, λ, f_0, β_0, and ξ real-valued parameters such that $\lambda \geqslant 0$ and $f_0 > 0$. This reformulation allows a simple physical interpretation of the parameters since the mean frequency f_m and the Q factor defined by $Q \equiv f_m/\Delta f$ are given, respectively, by $f_m = f_0$ and $Q^2 = 4\pi\lambda$. Moreover, the phase term is written so as to correspond to a modulation with a hyperbolic group delay $t_{X_2}(f) = \xi + \beta_0/f$.

Given these notations, we have the following.

Proposition 19: The unitary Bertrand distribution of the Klauder wavelet (41) is everywhere non-negative.

Proof: The proof relies on properties of characteristic functions (which are Fourier transforms of non-negative functions), and it proceeds as follows. We first start from the definition (41) of the Klauder wavelet and we plug it into the definition (30) of the Bertrand distribution. After some manipulations, the result can be expressed as

$$B_{X_2}(t,f) = |C|^2 f^{\alpha-q} \int_{-\infty}^{+\infty} M_{\alpha,\beta}(u) e^{-i2\pi\zeta u}\, du\; U(f), \tag{42}$$

with

$$M_{\alpha,\beta}(u) \equiv \left(\frac{u/2}{\sinh(u/2)} \right)^{\alpha} \exp\{-\beta u \coth(u/2)\} \tag{43}$$

and

$$\alpha \equiv 4\pi\lambda f_0; \quad \beta \equiv 2\pi\lambda f; \quad \zeta \equiv \beta_0 - (t-\xi) f. \tag{44}$$

In order to prove that (42) corresponds to a non-negative quantity, it is then sufficient to show that $M_{\alpha,\beta}(u)$ is—up to a positive constant—a characteristic function, a property which will obviously hold if each of the two factors of (43) is itself proportional to a characteristic function.[28] Using first the fact that[29]

$$\sinh x = x \prod_{k=1}^{\infty} \left(1 + \frac{x^2}{k^2\pi^2} \right),$$

we can write

$$\left(\frac{u/2}{\sinh(u/2)} \right)^{\alpha} = \prod_{k=1}^{\infty} \left(1 + i\, \frac{u}{2k\pi} \right)^{-\alpha} \left(1 - i\, \frac{u}{2k\pi} \right)^{-\alpha}.$$

This is clearly a characteristic function since, α being positive, each of the factors of this infinite product is the characteristic function of a Gamma distribution.[28] Using then the fact that[29]

$$\coth \pi x = \frac{1}{\pi x} + \frac{2x}{\pi} \sum_{k=1}^{\infty} \frac{1}{x^2 + k^2},$$

we can write.

$$\exp\{-\beta u \coth(u/2)\} = e^{-2\beta} \prod_{k=1}^{\infty} \exp\left\{ 4\beta \left(\frac{1}{1 + (u/2k\pi)^2} - 1 \right) \right\}. \tag{45}$$

Since $(1 + (u/2k\pi)^2)^{-1}$ is a characteristic function (namely the characteristic function of a Laplace distribution) and $\beta \geqslant 0$, it suffices to make use of a theorem by Lukacs[28]—stating that, if $g(u)$ is a characteristic function and γ a positive real number, $\exp\{\gamma(g(u)-1)\}$ is itself a characteristic function—for guaranteeing that each of the factors of (45) and, hence, (45) itself is a

characteristic function. The two factors of (43) each being a characteristic function, the Fourier transform of their product is everywhere non-negative and the proof is complete.

Two remarks.

(1) Assuming as we did that $\lambda>0$, we have $X_2(f)\in L^2(\mathbf{R}_+, f^{2r+1}\,df)$ and $B_{X_2}(t,f)$ is a bounded square-integrable function. A degeneracy is, however, observed in the case of the ''hyperbolic chirps'' $C_0(f)$ corresponding to $X_2(f)$ with $\lambda=0$, and for which we know[20] that

$$B_{C_0}(t,f)=|C|^2 f^{-q-1}\delta\left(t-\left(\xi+\frac{\beta_0}{f}\right)\right)U(f),$$

a result which also follows directly from (42), (43), and (44), since we have then $\alpha=\beta=0$, and hence $M_{\alpha,\beta}=1$.

(2) According to the definition (27), the nonzero frequency f_0 can be interpreted as the central frequency of the (envelope of the) Klauder wavelet. Within the narrow band assumption $Q^2=4\pi\lambda\gg 1$, we get the approximation

$$X_2(f)\sim C'\exp\left\{-\frac{\pi}{f_0}(\lambda-i\beta_0)\delta f^2-i2\pi\left(\xi+\frac{\beta_0}{f_0}\right)\delta f\right\},$$

where $\delta f\equiv f-f_0$ and C' is some suitable (complex-valued) constant. This means that narrow band Klauder wavelets reduce to Gaussians, in clear accordance with the fact that the Bertrand distribution reduces to the Wigner–Ville one for narrow band signals,[20] and that Wigner–Ville distributions are non-negative for Gaussians.

2. Unterberger distributions

A result similar to the one of the previous Proposition can be obtained in the case of the active Unterberger distribution (40), but it requires replacing the Klauder wavelet by the signal

$$X_{-1}(f)\equiv Cf^{-(r+1)}e^{-2\pi\lambda(\alpha_0/f+f)}e^{-i2\pi(\xi f-\alpha/f)}U(f),$$

with $\alpha_0>0$. More precisely, we can prove the following.

Proposition 20: The active Unterberger distribution of $X_{-1}(f)$ is everywhere non-negative.

Proof: The proof follows from a direct calculation according to which

$$U_{X_{-1}}(t,f)=2|C|^2 f^{-(2r+1)}\int_{-\infty}^{+\infty}e^{-4\pi\lambda(\alpha_0/f+f)\sqrt{1+(w/2)^2}}e^{i4\pi(tf-\xi f-\alpha/f)w}\,dw\;U(f),$$

a result which can be equivalently expressed as (see, e.g., Ref. 30)

$$U_{X_{-1}}(t,f)=4|C|^2 f^{-(2r+1)}G_\beta\left(4\pi\left(tf-\xi f-\frac{\alpha}{f}\right)\right)U(f),$$

with $\beta\equiv 4\pi\lambda(f+\alpha_0/f)$ and

$$G_\beta(y)\equiv\frac{2\beta}{\sqrt{\beta^2+y^2}}K_1(\beta^2+y^2), \tag{46}$$

$K_1(\cdot)$ being a modified Bessel function of the first kind.[26] Positivity of the distribution is therefore guaranteed by the simple fact that modified Bessel functions $K_\nu(z)$ are known to be positive if $\nu>-1$ and $z>0$.[26]

Unterberger distributions also exist under a ''passive'' form,[20] whose definition reads as

$$V_X(t,f)=f\int_{-\infty}^{+\infty}X(\gamma f)\overline{X\left(\frac{f}{\gamma}\right)}e^{i2\pi(\gamma-1/\gamma)tf}\,\frac{d\gamma}{\gamma}.$$

''Active'' Unterberger distributions are known to be perfectly localized for signals whose group delay is defined by squared hyperbolas of the form $\alpha+\gamma/f^2$, but they are also known not to

be unitary (in fact, it can be shown[20] that unitary Bertrand distributions are the only distributions guaranteeing both localization and unitarity). ''Passive'' Unterberger distributions share a lot of properties with ''active'' ones, at the notable exception of localization. Both forms are, in fact, ''dual'' in the sense that, whereas none has simultaneously properties similar to the unitary Bertrand distribution, they can be combined so as to satisfy the Moyal-type formula

$$\int_{-\infty}^{+\infty}\int_{0}^{+\infty} U_X(t,f)V_Y(t,f)\,dt\,\frac{df}{f}=\left|\int_{0}^{+\infty} X(f)\overline{Y(f)}\,\frac{df}{f}\right|^2.$$

With the above definition for $V_X(t,f)$, the following can be proved.

Proposition 21: The passive Unterberger distribution of $X_{-1}(f)$ is everywhere non-negative.

Proof: This can be checked by a direct calculation, according to which

$$V_X(t,f)=2|C|^2 f^{-(2r+1)}K_0\left(\sqrt{\beta^2+4\pi^2\left(tf-\xi f-\frac{\alpha}{f}\right)^2}\right)U(f),$$

which concludes the proof exactly as in the previous Proposition.

Remark: The closed form expression given above can be found almost exactly in Ref. 31, where the authors introduce a time-scale energy distribution whose definition has essentially the form of a passive Unterberger distribution.

In the case of the Wigner–Ville distribution, Hudson's theorem guarantees that Gaussian signals are the only ones ensuring positivity, whereas the question is left open to prove that Klauder wavelets would be the only signals with a (unitary) Bertrand distribution everywhere positive. In the case of the (active) Unterberger distribution, a class of signals ensuring positivity has been evidenced in Proposition 20, but it can be proved that this solution is not unique. Two counterexamples to unicity for positive Unterberger distributions are provided by the two following claims.

Proposition 22: The active Unterberger distribution of

$$Y_{-1}(f)\equiv Cf^{-(r+1)}e^{-2\pi\lambda f^2}e^{-i2\pi(\xi f-\alpha/f)}U(f),$$

is everywhere non-negative.

Proof: The active Unterberger distribution (40) admits the (time–frequency) equivalent form

$$U_X(t,f)=2f\int_{-\infty}^{+\infty}\cosh u\,X(e^u f)\overline{X(e^{-u}f)}e^{i4\pi tf\sinh u}\,du\,U(f).$$

Plugging the definition of $Y_{-1}(f)$ into this expression yields

$$U_{Y_{-1}}(t,f)=\frac{|C|^2}{\sqrt{2\lambda}}f^{-(2r+1)}e^{-4\pi\lambda f^2}e^{(-2\pi/\lambda)(t-\xi-\alpha/f^2)^2}U(f),$$

whence the result.

Remark: When applied to $Y_{-1}(f)$ with $\alpha=0$ (a constant group delay), the active Unterberger distribution is not only positive, but also separable in time and frequency.

Proposition 23: The active Unterberger distribution of

$$Z_{-1}(f)\equiv Cf^{-(r+1)}e^{-2\pi\lambda f}e^{-i2\pi(\xi f-\alpha/f)}U(f)$$

is everywhere non-negative.

Proof: The proof follows from a direct calculation, according to which

$$U_{Z_{-1}}(t,f)=2|C|^2 f^{-(2r+1)}G_{4\pi\lambda f}\left(4\pi\left(tf-\xi f-\frac{\alpha}{f}\right)\right)U(f), \tag{47}$$

with $G(\cdot)$ defined as in (46), and positivity stemming from the usual assumption $\lambda>0$.

V. CONCLUDING REMARKS

In this paper, a number of results have been given which—although still preliminary in many respects and far from exhausting the subject—are believed to clarify some of the links which exist between positivity, separability, and minimum uncertainty in time–frequency energy distributions. In fact, whereas in the case of the Wigner–Ville distribution, these features may admit order relations of the type ''separability⇒positivity⇒minimum uncertainty,'' it has been shown that the situation of other distributions is more intricate, with possible incompatibilities between the three different properties. Affine distributions (based on frequency and on a Mellin variable, in place of the usual time variable) have especially been considered in some detail, with two main consequences which make them depart from the Wigner–Ville case.

(1) In comparison with the known fact that minimizers of the classical time–frequency uncertainty—Gaussians with a linear phase—have a positive and separable Wigner–Ville distribution, signals with minimum frequency-scale uncertainty—namely, Klauder wavelets—do possess a positive (unitary) Bertrand distribution, but this latter is not separable.
(2) Whereas positivity is an exception in the Wigner–Ville case and can be observed with Gaussian signals only, it turns out that this situation of unicity is no longer true in the affine case, different classes of signals having been evidenced to lead to positive distributions.

As already mentioned, many questions are still left open, such as, e.g., the question of unicity for the positivity of *unitary* Bertrand distributions. From another perspective, it can be finally mentioned that other types of extensions (e.g., to the hyperbolic class[32] and, more generally, to ''warped'' classes[33]) could have been considered, but have not, since the corresponding results can be readily anticipated from those given here. In fact, such extensions being related to usual classes (Cohen of affine) by a warping operation, all the results obtained in one case can be transformed, *mutatis mutandis*, to get the corresponding results in the other case. For instance, the Altes-Marinovic distribution,[34]

$$Q_X(s,f)=f\int_{-\infty}^{+\infty} X(e^{u/2}f)\overline{X(e^{-u/2}f)}e^{i2\pi su}\,du,$$

which is known to belong to (and be a central member of) the hyperbolic class,[32] can be expressed as well in terms of the usual Wigner distribution as

$$Q_X(s,f)=W_{\tilde{X}}\left(\frac{s}{f_0},f_0\log\frac{f}{f_0}\right),$$

with $\tilde{X}(f)=e^{f/2f_0}X(e^{f/f_0})$ and $f_0>0$. It follows that any result pertaining to the Wigner–Ville distribution can be transposed to the Altes–Marinovic distribution, provided that it is applied to signals warped the suitable way. In particular, an analog of Hudson's theorem can be given as (no new proof is really necessary)

Proposition 24: The Altes–Marinovic distribution is positive for ''Altes signals'' of the type (39), and only for them.

ACKNOWLEDGMENTS

Thanks to Richard G. Baraniuk for many useful discussions and comments, and to the anonymous reviewers for their careful reading of the manuscript.

[1] E. P. Wigner, ''On the quantum correction for thermodynamic equilibrium,'' Phys. Rev. **40**, 749–759 (1932).
[2] J. Ville, ''Théorie et applications de la notion de signal analytique,'' Câbles Transm. **2ème A**, 61–74 (1948).
[3] L. Cohen, *Time–Frequency Analysis* (Prentice–Hall, Englewood Cliffs, 1995).
[4] P. Flandrin, *Temps–Fréquence* (Hermès, Paris, 1993).
[5] D. Gabor, ''Theory of communication,'' J. Inst. Electr. Eng. **93**, 429–457 (1946).
[6] R. L. Hudson, ''When is the Wigner quasi-probability density non-negative?,'' Rep. Math. Phys. **6**, 249–252 (1974).
[7] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, ''On the time–frequency discrimination of energy distributions—Can they look sharper than Heisenberg?,'' *IEEE-ICASSP'84*, San Diego, CA, 1984, pp. 41B7.1–41B7.4.
[8] K. Urbanik, ''Joint probability distributions of observables in quantum mechanics,'' Stud. Math. **XXI**, 117–133 (1961).
[9] A. J. E. M. Janssen, ''Positivity of weighted Wigner distributions,'' SIAM J. Math. Anal. **12**, 752–758 (1981).

[10] A. J. E. M. Janssen, ''Positivity properties of phase-plane distribution functions,'' J. Math. Phys. **25**, 2240–2252 (1984).
[11] A. J. E. M. Janssen, ''Bilinear phase-plane distribution functions and positivity,'' J. Math. Phys. **26**, 1986–1994 (1985).
[12] A. J. E. M. Janssen, ''Positivity of time–frequency distribution functions,'' Signal Process. **14**, 243–252 (1988).
[13] J. Aczél, *Lectures on Functional Equations and Their Applications* (Academic, New York, 1966).
[14] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, ''The Wigner distribution—A tool for time–frequency signal analysis. Part I: Continuous-time signals,'' Philips J. Res. **35**, 217–250 (1980).
[15] A. J. E. M. Janssen, ''On the locus and spread of pseudo-density functions in the time–frequency plane,'' Philips J. Res. **37**, 79–110 (1982).
[16] R. A. Altes, ''Methods of wide-band signal design for radar and sonar systems,'' Ph.D. dissertation, University of Rochester, Rochester, NY, 1970.
[17] N. J. De Bruijn, ''Uncertainty principles in Fourier analysis,'' in *Inequalities*, edited by O. Shisha (Academic, New York, 1967), pp. 55–71.
[18] E. H. Lieb, ''Integral bounds for radar ambiguity functions and Wigner distributions,'' J. Math. Phys. **31**, 594–599 (1990).
[19] A. J. E. M. Janssen, ''Optimality property of the Gaussian window spectrogram and an application to lattice expansions,'' IEEE Trans. Signal Process. **SP-39**, 202–204 (1991).
[20] J. Bertrand and P. Bertrand, ''A class of affine Wigner distributions with extended covariance properties,'' J. Math. Phys. **33**, 2515–2527 (1992).
[21] J. Ph. Ovarlez, ''La transformation de Mellin—Un outil pour l'analyse des signaux à large bande,'' Thèse de doctorat, Univ. Paris 6, 1992.
[22] L. Cohen, ''The scale representation,'' IEEE Trans. Signal Process. **41**, 3275–3292 (1993).
[23] J. R. Klauder, ''Path integrals for affine variables,'' in *Functional Integration: Theory and Applications*, edited by J.-P. Antoine and E. Tirapegui (Plenum, New York, 1980), pp. 101–119.
[24] K. B. Stolarsky, ''Generalizations of the logarithmic mean,'' Math. Mag. **48**, 87–92 (1975).
[25] P. Flandrin and P. Gonçalvès, ''Geometry of affine time–frequency distributions,'' Appl. Comp. Harm. Anal. **3**, 10–39 (1996).
[26] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (Dover, New York, 1972).
[27] R. A. Altes, ''Sonar for generalized target description and its similarity to animal echolocation systems,'' J. Acoust. Soc. Am. **59**, 97–105 (1976).
[28] E. Lukacs, *Characteristic Functions* (Griffin, London, 1970).
[29] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1980).
[30] F. Oberhettinger, *Fourier Transforms of Distributions and Their Inverses—A Collection of Tables* (Academic, New York, 1973).
[31] B. Escudié and A. Grossmann, ''Une représentation bilinéaire en temps et échelle,'' 13ème Coll. GRETSI, Juan-les-Pins, France, 1991, pp. 33–36.
[32] A. Papandreou, F. Hlawatsch, and G. F. Boudreaux-Bartels, ''The hyperbolic class of quadratic time–frequency representations. Part I: Constant-Q warping, the hyperbolic peradigm, properties and members,'' IEEE Trans. Signal Process. **SP-41**, 3425–3444 (1993).
[33] R. G. Baraniuk and D. L. Jones, ''Unitary equivalence: A new twist on signal processing,'' IEEE Trans. Signal Process. **SP-43**, 2269–2282 (1995).
[34] R. A. Altes, ''Wideband, proportional-bandwidth Wigner–Ville analysis,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-38**, 1005–1012 (1990).

Time-frequency transfer function calculus (symbolic calculus) of linear time-varying systems (linear operators) based on a generalized underspread theory[a)]

Gerald Matz and Franz Hlawatsch
Department of Communications and Radio-Frequency Engineering, Vienna University of Technology, Gusshausstrasse 25/389, A-1040 Wien, Austria

(Received 17 December 1997; accepted for publication 31 March 1998)

We introduce a generalized concept of underspread linear time-varying systems (linear operators) which contains a previous definition as a special case. We show that an existing approximate transfer function calculus (symbolic calculus) can be extended to this wider and practically more relevant class of underspread operators. As a mathematical underpinning of this calculus, we establish explicit bounds on various error quantities associated with it. The transfer function calculus provides a theoretical basis for various methods recently proposed for nonstationary signal processing, and it has important implications in the theory of time-varying power spectra. © *1998 American Institute of Physics.* [S0022-2488(98)00708-7]

I. INTRODUCTION

Quantum physics and time-frequency signal theory are strongly interrelated fields. In a formal sense, the phase space of physics (physical quantities position q and momentum p) plays a similar role in quantum mechanics as the time-frequency plane in signal theory (variables time t and frequency f). This formal analogy has led to numerous interactions between the two fields. Important examples are the Wigner distribution,[1–9] the uncertainty principle,[10–15] the connections between coherent states and the short-time Fourier transform or Gabor expansion,[9,12,15–24] and the relation of the Zak (or Weil–Brezin) transform with the polyphase decomposition.[15,20,25–30] However, notwithstanding their similarity on a formal mathematical level, the interpretation and scope of time-frequency signal analysis and semiclassical quantum mechanics are somewhat different.

(i) In the semiclassical approach to quantum mechanics, one desires a ''mock phase space'' [31] that is similar to the phase space of classical mechanics. In contrast, in signal analysis one desires joint time-frequency descriptions (of signals, systems, random processes,...) in order to combine pure time-domain and pure frequency-domain descriptions. More specifically, the phase-space descriptions of quantum observables (corresponding to self-adjoint operators on Hilbert spaces) are intended to preserve as much as possible the structural properties and conceptual power of the phase-space description of classical observables (functions in phase space). An example is the computation of expectations in phase space using quasi-probability densities. On the other hand, in signal analysis there is nothing comparable to classical observables (therefore, taking the classical limit $h \to 0$ does not make sense in signal analysis). Rather, one is interested in extending existing powerful concepts like the transfer function/frequency response of time-invariant linear systems and the power spectral density of stationary random processes to the time-varying and nonstationary cases, respectively. These time-frequency representations should retain as much as possible the conceptual power and intuitive interpretation of the transfer function (power spectral density).

(ii) The difficulties encountered when working with a quantum-mechanical mock phase space arise mainly from the fundamental mathematical difference between functions (classical observables) and operators (quantum-mechanical observables). In contrast, the main complications in the time-frequency description of time-varying systems (nonstationary random

[a)]Dedicated to Professor Wolfgang Mecklenbräuker on the occasion of his sixtieth birthday.

 © 1998 American Institute of Physics

processes) are due to their fundamental difference from time-invariant systems (stationary random processes) for which a pure frequency-domain description is sufficient and satisfactory.

(iii) Another difference between the viewpoints taken in quantum mechanics and in signal analysis concerns the noncommutativity of linear operators. In quantum mechanics, this noncommutativity is a fundamental fact that is part of our mathematical description of the physical world, whereas in signal analysis it can be considered simply as an inconvenience that prevents powerful, simple, and intuitive concepts and methods valid in the time-invariant/stationary case from being extended to the more general time-varying/nonstationary case. In a signal analysis context, therefore, it is reasonable and practically important to ask whether there exists a subclass of operators (time-varying systems) that is approximately commuting (cf. Subsections III C and IV A).

A. Summary of results and outline of paper

In this paper we consider a time-frequency description of linear operators or, equivalently, linear time-varying systems that are related to semiclassical quantum mechanics[2,15,32–36] and to the mathematical theory of pseudo-differential operators.[37–41] This time-frequency description of operators is an attempt to generalize the simple and efficient frequency-domain description of linear time-invariant systems—namely, the transfer function or frequency response—to the time-varying case. The result is a transfer function calculus (known as *symbolic calculus* in a quantum-mechanical context) that extends a previously introduced transfer function calculus[42–44] to a considerably wider and practically more relevant class of linear operators. This extended transfer function calculus provides a theoretical basis for several methods that have recently been proposed for nonstationary signal processing (specifically, to design and implement time-varying filters for signal enhancement, estimation, and detection[45–53]). Furthermore, it has important implications in the theory of time-varying power spectra for nonstationary random processes.[42,53–55]

The transfer function calculus presented is based on the generalized Weyl symbol which is discussed in Subsection I C. This discussion is preceded by a brief review of time-varying, time-invariant, and frequency-invariant linear operators in Subsection I B.

As a basis for our extended transfer function calculus, in Section II we develop a generalized concept of ''underspread'' operators that uses weighted integrals and moments of the operators' spreading function and does not require the spreading function to have compact support. Subsequently, in Section III we employ these integrals/moments to formulate explicit bounds on the errors incurred by the transfer function approximation. These bounds provide a rigorous mathematical underpinning for the extended, approximate transfer function calculus proposed. We note that several of these bounds and their proofs are analogous to similar bounds and proofs presented previously,[42–44] even though our formulation using weighted integrals and moments of the spreading function is new. Finally, in Section IV we show that underspread operators are approximately commuting and normal.

We note that symbolic calculi in a similar spirit have been obtained in the context of semiclassical quantum mechanics[15,33,39,56] and pseudo-differential operators,[15,37,38,40] with particular similarity to classical asymptotics of quantum mechanics[10,15,39,56–58] (like the WBK method[10] or Voros' h-expansions[39,56]). The difference is that these theories define specific symbol classes—whose mathematical sophistication has hitherto prevented their application to signal processing problems—directly in phase space whereas in our approach we formulate growth conditions in the (dual) spreading function domain (cf. Subsection II B). Also, whereas in quantization and pseudo-differential operator theory one studies the operators corresponding to a given symbol, we consider a given linear operator and investigate how close its generalized Weyl symbol comes to the intuitive engineering notion of a time-frequency transfer function.

B. Time-varying, time-invariant, and frequency-invariant linear operators

Linear time-varying systems/operators are useful models in a variety of engineering applications as diverse as speech production and mobile communications. (In what follows, we tacitly assume that all systems/operators are linear.) Except for the limiting cases of time-invariant and

frequency-invariant systems (see below), we will subsequently restrict our attention to Hilbert–Schmidt operators,[59] i.e., operators $\mathbf{H}$ with a finite Hilbert–Schmidt norm,

$$\|\mathbf{H}\|_2 \triangleq \sqrt{\sum_k \|\mathbf{H}u_k\|_2^2} < \infty,$$

where $\{u_k(t)\}$ is an arbitrary orthonormal basis of $L_2(\mathbb{R})$. Although in quantum mechanics many observables correspond to unbounded operators, our restriction to Hilbert–Schmidt operators (which are always bounded) is consistent with most engineering applications. The input-output relation for a Hilbert–Schmidt operator $\mathbf{H}$ can be written as (integrals are from $-\infty$ to ∞)

$$(\mathbf{H}x)(t) = \int_{t'} h(t,t')x(t')dt', \tag{1}$$

where $x(t)$ and $(\mathbf{H}x)(t)$ are the system's input and output signal, respectively, and $h(t,t') \in L_2(\mathbb{R}^2)$ is the kernel (impulse response) of $\mathbf{H}$. We note that $\|\mathbf{H}\|_2^2 = \int_t \int_{t'} |h(t,t')|^2 dt dt'$.

Unfortunately, time-varying systems/operators are much more difficult to analyze and characterize than time-invariant systems, i.e., operators that commute with time shifts and have a convolution-type kernel of the form $h(t,t') = g(t-t')$. For time-invariant systems the *transfer function* (frequency response),

$$G(f) = \int_\tau g(\tau) e^{-i2\pi f\tau} d\tau,$$

is an extremely simple and efficient system description. This is due to the following properties.

(i) The complex sinusoids $\{e^{i2\pi ft}\}$ (parametrized by frequency f) are the generalized[60] eigenfunctions of any time-invariant operator, with $G(f)$ the associated generalized eigenvalues, i.e., $(\mathbf{H}x)(t) = G(f)e^{i2\pi ft}$ for $x(t) = e^{i2\pi ft}$. This is also the reason why all linear time-invariant systems commute with each other.

(ii) As a consequence of the previous property, the Fourier transform of $(\mathbf{H}x)(t)$ equals the Fourier transform of $x(t)$ multiplied by $G(f)$. Hence, in the frequency (Fourier) domain, the input-output relation simplifies to the multiplication of two functions.

(iii) The transfer function of the series connection (product) of two linear time-invariant systems with transfer functions $G_1(f)$ and $G_2(f)$ equals $G_1(f)G_2(f)$.

(iv) The adjoint[59] $\mathbf{H}^+$ of a time-invariant system has impulse response $g^*(-\tau)$, and hence its transfer function is simply the complex conjugate of $G(f)$.

(v) For input signals $x(t) \in L_2(\mathbb{R})$, the maximum system gain is equal to the supremum of $|G(f)|$.

The dual of linear time-invariant systems is given by linear ''frequency-invariant'' systems (i.e., operators commuting with frequency shifts), which correspond to time-domain multiplications and have an impulse response of the form $h(t,t') = m(t)\delta(t-t')$. Here, the factor $m(t)$ plays the role of a ''temporal transfer function'' with properties similar to those of the spectral transfer function $G(f)$ of linear time-invariant systems. We note that the quantum-mechanical counterparts of time-invariant and frequency-invariant systems are observables that depend solely on momentum and solely on position, respectively.

Compared to time-invariant and frequency-invariant systems, general linear operators are much more difficult and less intuitive to work with. In particular, as opposed to time-invariant (frequency-invariant) operators which are always normal (i.e., which satisfy $\mathbf{H}^+\mathbf{H} = \mathbf{H}\mathbf{H}^+$) and which always commute, general operators are potentially non-normal and typically noncommuting. For normal operators, a generalization of the transfer function $G(f)$ or $m(t)$ is provided by the operator's eigenvalues[59] (in the non-normal case one has to recur to singular values[59,61]). For example, the adjoint operator and the operator norm (maximum system gain) correspond to the complex conjugate and supremum, respectively, of the eigenvalues. Unfortunately, for practical engineering applications this concept has certain drawbacks and limitations.

(i) The eigenfunctions of a general linear operator are not known *a priori*, and their computa-

tion is expensive. The lack of a simple structure of the eigenfunctions prevents the application of efficient algorithms like the fast Fourier transform that is useful in the time-invariant case.

(ii) The eigenfunctions and eigenvalues are not related to the frequency or time variable.

(iii) The eigenfunctions of two noncommuting operators are different, and the eigenvalues of the operators' product (series connection of systems) are not equal to the product of the operators' eigenvalues.

C. The generalized Weyl symbol

Since the eigenvalue-based transfer function concept does not yield the desired physical relevance and computational efficiency, an alternative ''transfer function-like'' characterization of linear operators is needed. A ''time-varying transfer function'' of time-varying systems was introduced by Zadeh[62] and later complemented by Bello's frequency-dependent modulation function[63] which is equivalent to the Kohn–Nirenberg symbol[15,37] introduced in the theory of pseudo-differential operators. These two definitions correspond to what is known in the physical literature as standard and anti-standard correspondence rules.[31,34–36] In a signal processing context, the Weyl symbol,[4,15,32,64,65] originally introduced in quantum mechanics,[66] was then recognized as a further definition of a ''transfer function.'' [67]

Finally, Kozek[42,44,68] placed all three definitions within a unifying framework by defining the *generalized Weyl symbol* as

$$L_{\mathbf{H}}^{(\alpha)}(t,f) \triangleq \int_{\tau} h^{(\alpha)}(t,\tau) e^{-i2\pi f\tau} d\tau, \tag{2}$$

with the following coordinate-transformed version of the impulse response $h(t,t')$ in (1),

$$h^{(\alpha)}(t,\tau) \triangleq h(t+(\tfrac{1}{2}-\alpha)\tau, t-(\tfrac{1}{2}+\alpha)\tau). \tag{3}$$

Here, α is a parameter for which an interpretation will be provided in Subsection II A. Although any real-valued number may be taken for α, the case $|\alpha| \leqslant 1/2$ is of particular importance since this condition guarantees that the generalized Weyl symbol satisfies finite support properties.[42] We note that the generalized Weyl symbol is a special case of a more general correspondence rule between classical and quantum observables.[31,34,35,69] It is a function of the physically relevant variables time t and frequency f (in the quantum mechanical context, these correspond to position q and momentum p). For $\alpha = 0$, 1/2, and $-1/2$, it reduces to, respectively, the Weyl symbol, Zadeh's time-varying transfer function, and Bello's frequency-dependent modulation function. For rank-one operators, i.e., operators $\mathbf{H}$ of the type $h(t,t') = s(t)s^*(t')$, it equals the generalized Wigner distribution[70,71] of $s(t)$. For $\alpha = 0$ the generalized Wigner distribution simplifies to the Wigner distribution[1–6,8,9] whereas for $\alpha = 1/2$ the Rihaczek distribution[72] is obtained.

The generalized Weyl symbol satisfies many desirable mathematical properties.[42,68] In particular, it preserves inner products and Hilbert–Schmidt norms of operators, its integral over the entire time-frequency plane yields the operator's trace, and it is ''covariant'' to time-frequency shifts (Galilean invariance) and time-frequency scalings of operators. Furthermore, for time-invariant and frequency-invariant operators it simplifies to the spectral transfer function $G(f)$ and the temporal transfer function $m(t)$, respectively (this is the signal analysis reformulation of the fact that any valid quantum-mechanical correspondence rule yields the position and momentum operators as counterparts of the classical observables position and momentum, respectively).

The Weyl symbol (obtained with $\alpha = 0$) possesses an inherent symmetry which leads to two further important properties.

(i) The Weyl symbol of the adjoint operator $\mathbf{H}^+$ equals the complex conjugate of the Weyl symbol of $\mathbf{H}$, i.e., $L_{\mathbf{H}^+}^{(0)}(t,f) = L_{\mathbf{H}}^{(0)*}(t,f)$. Hence, the Weyl symbol of a self-adjoint (Hermitian) operator, i.e., an operator satisfying $\mathbf{H}^+ = \mathbf{H}$, is real-valued.[15,31,32,36]

(ii) The Weyl symbol is covariant to operator transformations by metaplectic operators $\mathbf{U} \in \mathscr{M}$, i.e., unitary operators that belong to the metaplectic representation $\mathscr{M}$ of the symplectic group $Sp(2,\mathbb{R})$ consisting of measure-preserving, linear time-frequency coordinate

transforms $\binom{t}{f} \to \begin{pmatrix} a & b \\ c & d \end{pmatrix}\binom{t}{f}$ [with $\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc = 1$] such as rotations, shearings, and scalings:[4,15,40,41,64,73–75]

$$L^{(0)}_{\mathbf{UHU}^+}(t,f) = L^{(0)}_{\mathbf{H}}(at+bf, ct+df). \tag{4}$$

Unfortunately, in spite of its attractive mathematical properties the generalized Weyl symbol cannot, in general, be used as a transfer function in a manner similar to the time-invariant or frequency-invariant case. For example, the generalized Weyl symbol of an operator product does not generally equal the product of the individual symbols, and the supremum of the magnitude of the generalized Weyl symbol does not generally reflect the operator norm. Indeed, in a signal analysis context it has often been conjectured that the concept of a time-varying transfer function requires the system to be slowly time-varying.[48,63,76–80] However, it seems that until recently no quantitative formulation of this conjecture has been available.

In his pioneering work,[42–44] Kozek defined a class of ''underspread'' linear operators which even includes operators with fast time variations. He showed that this class allows an approximate transfer function calculus (symbolic calculus) based on the generalized Weyl symbol. Underspread operators were defined by the condition that the operator's spreading function (see Section II) has compact support contained within a rectangle of area $\ll 1$. Unfortunately, in practice this is typically not satisfied exactly but only effectively.

II. EXTENDED CONCEPT OF UNDERSPREAD SYSTEMS

Our extension of the generalized Weyl symbol based transfer function calculus (symbolic calculus) is based on a generalized concept of underspread operators. This concept is related to the time-frequency shifts introduced by linear operators and their characterization by means of the spreading function.

A. The spreading function

Contrary to time-invariant systems (which cause only time shifts) and frequency-invariant systems (which cause only frequency shifts), general linear operators shift the input signal with respect to both time and frequency. This behavior is reflected by the following reformulation of the input-output relation (1) as a weighted superposition of time-frequency-shifted versions of the input signal,[42–44,63,64,68,81]

$$(\mathbf{H}x)(t) = \int_\tau \int_\nu S^{(\alpha)}_{\mathbf{H}}(\tau,\nu) x^{(\alpha)}_{\tau,\nu}(t)\, d\tau\, d\nu. \tag{5}$$

Here, $x^{(\alpha)}_{\tau,\nu}(t) = x(t-\tau)e^{i2\pi\nu t}e^{i2\pi(\alpha-1/2)\tau\nu}$ is the signal $x(t)$ shifted by τ in time and by ν in frequency, with $\alpha \in \mathbb{R}$ expressing the freedom in ordering the time and frequency shifts.[32,42,68] This generalized time-frequency shift is equivalent to the Schrödinger representation of the three-dimensional (3-D) Heisenberg group[15,42,75,82] with parameters $(-\tau, \nu, \alpha\tau\nu)$. Relation (5) is, furthermore, equivalent to Weyl's original idea for establishing a correspondence rule between classical and quantum observables by Fourier-transforming the classical observable and then replacing the position and momentum variables in the exponent by the position and momentum operators.[15,31,32,34,36,83] The ''weighting function'' in (5), $S^{(\alpha)}_{\mathbf{H}}(\tau,\nu)$, is the *(generalized) spreading function* of $\mathbf{H}$ which is defined as[15,42,63,68,81]

$$S^{(\alpha)}_{\mathbf{H}}(\tau,\nu) \triangleq \int_t h^{(\alpha)}(t,\tau) e^{-i2\pi\nu t} dt,$$

with $h^{(\alpha)}(t,\tau)$ as in (3). For $\alpha = \pm 1/2$, the spreading function has first been introduced under the names of the ''delay-Doppler spread function'' and ''Doppler-delay spread function'' by Bello[63] who considered communication channels with multipath propagation and moving scatterers. Here, the received signal $(\mathbf{H}x)(t)$ consists of several delayed and Doppler-shifted versions of the original signal with the spreading function describing the ''reflectivity'' of the individual scatterers.

The spreading function is the 2-D (symplectic) Fourier transform of the generalized Weyl symbol in (2),

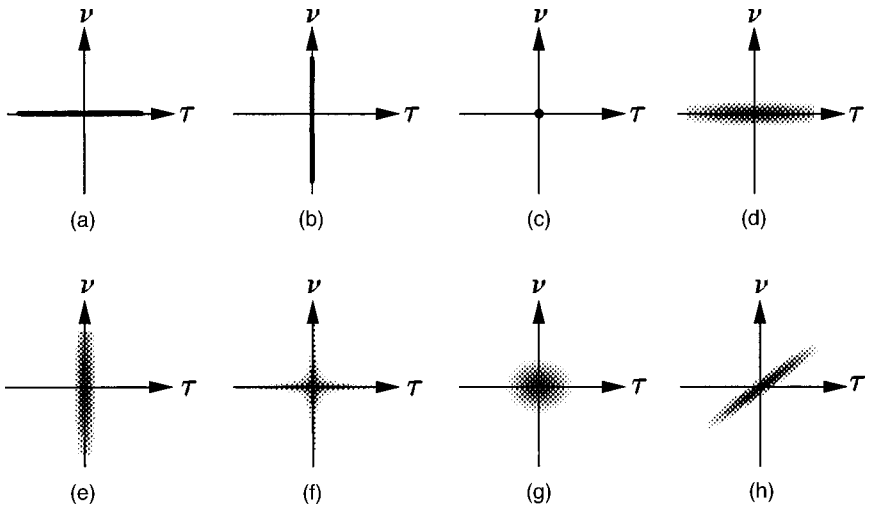

FIG. 1. Examples for the magnitude of the spreading function of some generic types of operators: (a) time-invariant system, (b) frequency-invariant system, (c) identity operator, (d) quasi-time-invariant system, (e) quasi-frequency-invariant system, (f)–(h) three different versions of underspread operators.

$$S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)=\int_t\int_f L_{\mathbf{H}}^{(\alpha)}(t,f\,)e^{-i2\pi(\nu t-\tau f\,)}dtdf. \tag{6}$$

Spreading functions with distinct α differ only by a phase factor,[42]

$$S_{\mathbf{H}}^{(\alpha_2)}(\tau,\nu)=S_{\mathbf{H}}^{(\alpha_1)}(\tau,\nu)e^{i2\pi(\alpha_1-\alpha_2)\tau\nu}, \tag{7}$$

so that $|S_{\mathbf{H}}^{(\alpha_2)}(\tau,\nu)|=|S_{\mathbf{H}}^{(\alpha_1)}(\tau,\nu)|$. Hence, we will write $|S_{\mathbf{H}}(\tau,\nu)|$ instead of $|S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)|$.

Like the generalized Weyl symbol, the spreading function preserves inner products and Hilbert–Schmidt norms of operators. Its value at the origin equals the operator's trace, and its magnitude is covariant to metaplectic transformations $\mathbf{U}\in\mathscr{M}$ (cf. Subsection I C),

$$|S_{\mathbf{UHU}^+}(\tau,\nu)|=|S_{\mathbf{H}}(a\tau+b\nu,c\tau+d\nu)|. \tag{8}$$

For a time-invariant system, we have $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)=g(\tau)\delta(\nu)$ which is perfectly concentrated along the τ axis. Similarly, for a frequency-invariant system, $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)=M(\nu)\delta(\tau)$ [with $M(\nu)$ the Fourier transform of $m(t)$] which is perfectly concentrated along the ν axis (cf. Fig. 1). This is in accordance with the fact that time-invariant systems cause only time shifts and frequency-invariant systems cause only frequency shifts.

B. Measures for the spread of the spreading function

The spread, or extension, of the spreading function about the origin of the (τ,ν) plane provides a global characterization of the time-frequency shifts introduced by the operator. [If the spreading function is concentrated about some point $(\tau_0,\nu_0)\neq(0,0)$, this corresponds to an offset time-frequency shift which can be split off from the operator, resulting in an operator whose spreading function is concentrated about the origin.] The extension in the τ direction characterizes the operator's ''length of memory'' whereas the extension in the ν direction determines the fastness of the operator's time-variations or fluctuations. The mathematical formulation of the concept of underspread operators (cf. Subsection II C) requires quantitative measures of the extension of the spreading function about the origin. For operators with a compactly supported spreading function, such a measure is provided by the area of the smallest rectangle containing the support of the spreading function.[42–44] In practice, however, the compact support condition is often not met. To circumvent this problem, we here propose to use the following weighted integrals of the spreading function $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)$ which are normalized by the L_1 or L_2 norm of $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)$ [These norms are defined as $\|S_{\mathbf{H}}\|_1=\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|d\tau d\nu$ (which is assumed to be finite) and $\|S_{\mathbf{H}}\|_2^2=\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|^2d\tau d\nu$. We note that $\|S_{\mathbf{H}}\|_2=\|\mathbf{H}\|_2$]:

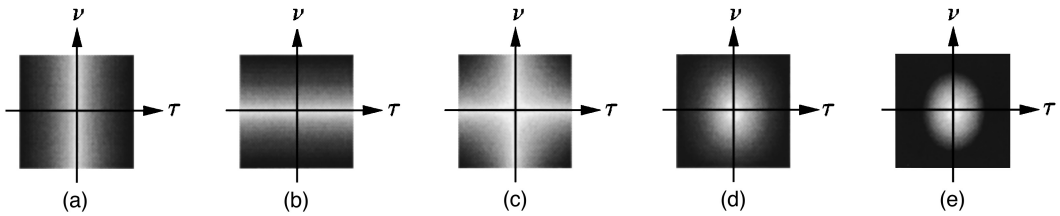

FIG. 2. Gray-scale plots (darker shades correspond to larger values) of some specific weighting functions: (a) $\phi(\tau,\nu)=|\tau|^k$, (b) $\phi(\tau,\nu)=|\nu|^k$, (c) $\phi(\tau,\nu)=|\tau\nu|^k$, (d) $\phi(\tau,\nu)=|1-A_s^{(0)}(\tau,\nu)|$, (e) $\phi(\tau,\nu)=|1-1/A_s^{(0)}(\tau,\nu)|$ with $A_s^{(0)}(\tau,\nu)$ the (symmetric) ambiguity function of a normalized Gaussian function (cf. Subsections III D and III H).

$$m_{\mathbf{H}}^{(\phi)} \triangleq \frac{1}{\|S_{\mathbf{H}}\|_1} \int_\tau \int_\nu \phi(\tau,\nu)|S_{\mathbf{H}}(\tau,\nu)|\,d\tau\,d\nu, \tag{9}$$

$$M_{\mathbf{H}}^{(\phi)} \triangleq \frac{1}{\|\mathbf{H}\|_2} \left[\int_\tau \int_\nu \phi^2(\tau,\nu)|S_{\mathbf{H}}(\tau,\nu)|^2\,d\tau\,d\nu \right]^{1/2}.$$

Here, $\phi(\tau,\nu)$ is a non-negative weighting function which satisfies $\phi(\tau,\nu)\geqslant\phi(0,0)=0$ and thus penalizes contributions of the spreading function lying away from the origin. We note that $m_{\mathbf{H}}^{(\phi)}$ and $M_{\mathbf{H}}^{(\phi)}$ are non-negative and do not depend on the parameter α. Since the spreading function of the adjoint operator $\mathbf{H}^+$ is given by $S_{\mathbf{H}^+}^{(\alpha)}(\tau,\nu)=S_{\mathbf{H}}^{(-\alpha)*}(-\tau,-\nu)$, we have $m_{\mathbf{H}^+}^{(\phi)}=m_{\mathbf{H}}^{(\phi)}$ and $M_{\mathbf{H}^+}^{(\phi)}=M_{\mathbf{H}}^{(\phi)}$ if the weighting function is even-symmetric, i.e., $\phi(-\tau,-\nu)=\phi(\tau,\nu)$. Figure 2 shows some specific weighting functions that will be used in Sections III and IV.

We also introduce the (absolute) moments $m_{\mathbf{H}}^{(k,l)}$ and $M_{\mathbf{H}}^{(k,l)}$ of the spreading function as a special case of the above weighted integrals using weighting functions $\phi(\tau,\nu)=|\tau|^k|\nu|^l$ with $k,l\in\mathbb{N}_0$:

$$m_{\mathbf{H}}^{(k,l)} \triangleq \frac{1}{\|S_{\mathbf{H}}\|_1} \int_\tau \int_\nu |\tau|^k|\nu|^l|S_{\mathbf{H}}(\tau,\nu)|\,d\tau\,d\nu,$$

$$M_{\mathbf{H}}^{(k,l)} \triangleq \frac{1}{\|\mathbf{H}\|_2} \left[\int_\tau \int_\nu \tau^{2k}\nu^{2l}|S_{\mathbf{H}}(\tau,\nu)|^2\,d\tau\,d\nu \right]^{1/2}.$$

We have $m_{\mathbf{H}^+}^{(k,l)}=m_{\mathbf{H}}^{(k,l)}$ and $M_{\mathbf{H}^+}^{(k,l)}=M_{\mathbf{H}}^{(k,l)}$ and, furthermore,

$$0\leqslant m_{\mathbf{H}}^{(k,l)}\leqslant\sqrt{m_{\mathbf{H}}^{(2k,0)}m_{\mathbf{H}}^{(0,2l)}}, \quad 0\leqslant M_{\mathbf{H}}^{(k,l)}\leqslant\sqrt{M_{\mathbf{H}}^{(2k,0)}M_{\mathbf{H}}^{(0,2l)}}. \tag{10}$$

Moments with $k=0$ or $l=0$ penalize mainly contributions of the spreading function that are located away from the τ axis or away from the ν axis, respectively. Thus, operators with a spreading function concentrated along the τ axis [i.e., quasi-time-invariant operators; see Fig. 1(d)] have small $m_{\mathbf{H}}^{(0,l)}$ and small $M_{\mathbf{H}}^{(0,l)}$ [see Fig. 2(b)], whereas operators with a spreading function concentrated along the ν axis [i.e., quasi-frequency-invariant operators; see Fig. 1(e)] have small $m_{\mathbf{H}}^{(k,0)}$ and small $M_{\mathbf{H}}^{(k,0)}$ [see Fig. 2(a)]. Moments with $k=l$ penalize mainly contributions of the spreading function that are located away from the τ *and* ν axes, i.e., that lie in oblique directions in the (τ,ν) plane. This is due to the fact that the corresponding weighting function is constant along the hyperbolas $|\tau\nu|=c$ [cf. Fig. 2(c)]. In particular, a superposition (i.e., parallel connection) of a (quasi-) time-invariant operator and a (quasi-) frequency-invariant operator has a spreading function concentrated along the τ and ν axes and will thus have small $m_{\mathbf{H}}^{(k,k)}$ and $M_{\mathbf{H}}^{(k,k)}$.

C. Underspread operators

Conceptually, a linear operator is called *underspread* if its spreading function is concentrated in a small region about the origin of the (τ,ν) plane, which indicates that the operator introduces only small time-frequency shifts τ,ν or, in other words, that the operator's memory is short and/or its time-variations are slow.

The concept of underspread operators was first used for random linear time-varying systems in the context of fading multipath channels[63,78,81,84,85] where physical considerations (path loss, limited velocities) naturally suggested limitations of delay (τ) and Doppler shifts (ν). Adopting the terminology of the random case, Kozek[42–44] introduced underspread deterministic linear operators by requiring that their spreading function be exactly zero outside a small rectangular support region about the origin. In practice, however, this condition is often not satisfied exactly but only effectively. This raises the question of how to choose the effective support region and how the modeling error resulting from a specific choice of effective support region affects the validity of the results obtained using the compact support model.

These problems and limitations can be circumvented by using the weighted integrals $m_{\mathbf{H}}^{(\phi)}$ and $M_{\mathbf{H}}^{(\phi)}$ and, specifically, the moments $m_{\mathbf{H}}^{(k,l)}$ and $M_{\mathbf{H}}^{(k,l)}$ introduced in Subsection II B. Hence, without being forced to assume that the spreading function has finite support, we can consider an operator $\mathbf{H}$ to be underspread if suitable integrals/moments of the spreading function are ''small.'' We note that this is not a clear-cut definition of underspread operators. Indeed, we will see in Section III that bounds on various error quantities associated with the transfer function approximation are formulated using different integrals/moments of the spreading function. Hence, having these approximation errors small calls for different integrals/moments being small. This is somewhat similar to the many different quantities that can be used to measure the effective bandwidth of a signal. Our concept of underspread operators is thus more complicated and diffuse, but also more flexible and practically relevant, than Kozek's definition which required the spreading function to have (small) compact support. It is precisely this flexibility which, in Section III, will allow us to establish bounds for the transfer function calculus (symbolic calculus) approximation without unnecessarily restrictive assumptions.

If the spreading function is exactly zero outside a finite region $\mathcal{S}$ around the origin, the weighted integrals are bounded as $m_{\mathbf{H}}^{(\phi)} \leqslant \phi_{\max}$ and $M_{\mathbf{H}}^{(\phi)} \leqslant \phi_{\max}$ with $\phi_{\max} = \max_{(\tau,\nu)\in\mathcal{S}}\{\phi(\tau,\nu)\}$. In particular, for rectangular regions $\mathcal{S}=[-\tau_{\max},\tau_{\max}]\times[-\nu_{\max},\nu_{\max}]$, one has $m_{\mathbf{H}}^{(k,l)} \leqslant \tau_{\max}^k \nu_{\max}^l$ and $M_{\mathbf{H}}^{(k,l)} \leqslant \tau_{\max}^k \nu_{\max}^l$. Thus, operators that are underspread in the sense of Kozek,[42–44] i.e., that have compact support contained within a rectangle of area $4\tau_{\max}\nu_{\max} \ll 1$, are a special case of our framework.

It should be noted that the concept of underspread operators is not equivalent to that of slowly time-varying (quasi-time-invariant) operators which requires $|S_{\mathbf{H}}(\tau,\nu)|$ to be narrow with respect to ν [i.e., small moments $m_{\mathbf{H}}^{(0,l)}$ and $M_{\mathbf{H}}^{(0,l)}$; see Fig. 1(d) and Fig. 2(b)]. In particular, a slowly time-varying operator may be overspread (i.e., not underspread) if its memory is very long, and an operator with fast time-variations may be underspread if its memory is short enough.

Due to the 2-D Fourier transform relation (6) between the spreading function and the generalized Weyl symbol, the generalized Weyl symbol of an underspread operator is a smooth (i.e., a 2-D lowpass-type) function. This can also be seen from the following relations between the partial derivatives of the generalized Weyl symbol and the moments of the spreading function:

$$\left|\frac{\partial^{k+l} L_{\mathbf{H}}^{(\alpha)}(t,f)}{\partial t^l \partial f^k}\right| \leqslant (2\pi)^{k+l} \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(k,l)}, \quad \left\|\frac{\partial^{k+l} L_{\mathbf{H}}^{(\alpha)}}{\partial t^l \partial f^k}\right\|_2 = (2\pi)^{k+l} \|\mathbf{H}\|_2 M_{\mathbf{H}}^{(k,l)}. \tag{11}$$

In particular, $|\partial L_{\mathbf{H}}^{(\alpha)}(t,f)/\partial t|$ will be small—implying smoothness of the generalized Weyl symbol in the time direction—if the moments $m_{\mathbf{H}}^{(0,1)}$ and $M_{\mathbf{H}}^{(0,1)}$ are small. Similarly, the generalized Weyl symbol will be smooth in the frequency direction (i.e., $|\partial L_{\mathbf{H}}^{(\alpha)}(t,f)/\partial f|$ will be small) if $m_{\mathbf{H}}^{(1,0)}$ and $M_{\mathbf{H}}^{(1,0)}$ are small. We note that the first relation (bound) in (11) bears some resemblance to the definition of specific symbol classes[15,37,38,40,41,56] of pseudo-differential operators, with the difference that our bound is uniform over the entire time-frequency plane whereas in pseudo-differential operator theory there is an additional term controlling the growth of the partial derivatives of the symbol.

III. TRANSFER FUNCTION CALCULUS (SYMBOLIC CALCULUS)

In this section, we show that for linear operators that are underspread in the extended sense of Subsection II C, the generalized Weyl symbol $L_{\mathbf{H}}^{(\alpha)}(t,f)$ is an approximate ''time-frequency transfer function'' that generalizes the spectral/temporal transfer function of time-invariant/frequency-

invariant linear systems. As a mathematical underpinning of this approximate time-frequency transfer function calculus (symbolic calculus), we establish explicit upper bounds on the errors associated with it. These bounds are formulated in terms of the integrals and/or moments of the spreading function introduced in Subsection II B; they do not require the spreading function to have compact support. Hence, our subsequent results will show that a transfer function calculus (symbolic calculus) is valid for a significantly wider class of systems than that considered previously.[42–44] This calculus leads to simple and numerically efficient time-frequency methods for specific nonstationary signal processing applications[45–53] and it yields important results on time-varying power spectra of nonstationary random processes.[42,54,55]

A. Approximate uniqueness of the generalized Weyl symbol

The dependence of the generalized Weyl symbol on the parameter α is a difference from the time-invariant and frequency-invariant cases where the transfer function of a given operator is uniquely defined. In a signal processing context, this dependence may be considered an inconvenience since in practice it is often not clear which α value should be selected. In particular, the case $\alpha=0$ features an important symplectic covariance [cf. (4)] whereas the cases $\alpha=\pm 1/2$ lend themselves to simpler discrete-time implementations. Fortunately, the following bounds on the generalized Weyl symbol's α-dependence (which extend existing bounds[42]) show that in the underspread case the choice of α is not critical.

Theorem III.1: *For any operator* $\mathbf{H}$*, the difference* $\Delta_1^{(\alpha_1,\alpha_2)}(t,f)\triangleq L_{\mathbf{H}}^{(\alpha_1)}(t,f)-L_{\mathbf{H}}^{(\alpha_2)}(t,f)$ *between two generalized Weyl symbols with parameters* α_1 *and* α_2 *is bounded as*

$$\frac{|\Delta_1^{(\alpha_1,\alpha_2)}(t,f)|}{\|S_{\mathbf{H}}\|_1}\leq 2\pi|\alpha_1-\alpha_2|m_{\mathbf{H}}^{(1,1)},\quad \frac{\|\Delta_1^{(\alpha_1,\alpha_2)}\|_2}{\|\mathbf{H}\|_2}\leq 2\pi|\alpha_1-\alpha_2|M_{\mathbf{H}}^{(1,1)}. \tag{12}$$

[Here, $\|\Delta_1^{(\alpha_1,\alpha_2)}\|_2$ is the L_2 norm of $\Delta_1^{(\alpha_1,\alpha_2)}(t,f)$ defined as $\|\Delta_1^{(\alpha_1,\alpha_2)}\|_2^2=\int_t\int_f|\Delta_1^{(\alpha_1,\alpha_2)}(t,f)|^2\,dtdf$. We consider bounds on the *normalized* error magnitude and L_2 norm. The magnitude is normalized by $\|S_{\mathbf{H}}\|_1$ since $|L_{\mathbf{H}}^{(\alpha)}(t,f)|\leq\|S_{\mathbf{H}}\|_1$, and the L_2 norm is normalized by $\|\mathbf{H}\|_2$ since $\|L_{\mathbf{H}}^{(\alpha)}\|_2=\|\mathbf{H}\|_2$. Furthermore, in this and subsequent bounds the integrals and moments involved are assumed to exist.]

Proof: With (6) and (7), the 2-D Fourier transform of $\Delta_1^{(\alpha_1,\alpha_2)}(t,f)$ is given by

$$\hat{\Delta}_1^{(\alpha_1,\alpha_2)}(\tau,\nu)=\int_t\int_f\Delta_1^{(\alpha_1,\alpha_2)}(t,f)e^{-i2\pi(\nu t-\tau f)}dtdf=S_{\mathbf{H}}^{(\alpha_1)}(\tau,\nu)[1-e^{i2\pi(\alpha_1-\alpha_2)\tau\nu}].$$

The first bound (L_∞ bound) in (12) is then shown as

$$\begin{aligned}|\Delta_1^{(\alpha_1,\alpha_2)}(t,f)|&=\left|\int_\tau\int_\nu\hat{\Delta}_1^{(\alpha_1,\alpha_2)}(\tau,\nu)e^{i2\pi(t\nu-f\tau)}d\tau d\nu\right|\leq\int_\tau\int_\nu|\hat{\Delta}_1^{(\alpha_1,\alpha_2)}(\tau,\nu)|d\tau d\nu\\&=\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)||1-e^{i2\pi(\alpha_1-\alpha_2)\tau\nu}|d\tau d\nu\\&=2\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)||\sin(\pi(\alpha_1-\alpha_2)\tau\nu)|d\tau d\nu\\&\leq 2\pi|\alpha_1-\alpha_2|\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)||\tau||\nu|d\tau d\nu\\&=2\pi|\alpha_1-\alpha_2|\|S_{\mathbf{H}}\|_1m_{\mathbf{H}}^{(1,1)},\end{aligned}$$

where we used $|\sin x|\leq|x|$. The second bound (L_2 bound) is shown by noting that

$$\|\Delta_1^{(\alpha_1,\alpha_2)}\|_2^2=\|\hat{\Delta}_1^{(\alpha_1,\alpha_2)}\|_2^2=\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|^2|1-e^{i2\pi(\alpha_1-\alpha_2)\tau\nu}|^2d\tau d\nu,$$

and proceeding similarly as above, using $\sin^2 x \leq x^2$. □

Note that the bounds in this theorem depend only on the difference $\alpha_1-\alpha_2$ and not on α_1, α_2 individually. For $|\alpha|\leq 1/2$, one obtains the coarser but α-independent bounds $2\pi m_{\mathbf{H}}^{(1,1)}$ and $2\pi M_{\mathbf{H}}^{(1,1)}$. For an *underspread* operator whose moments $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ are small, the bounds in (12) show that the generalized Weyl symbol is approximately independent of α, i.e.,

$$L_{\mathbf{H}}^{(\alpha_1)}(t,f)\approx L_{\mathbf{H}}^{(\alpha_2)}(t,f),$$

as long as $|\alpha_1-\alpha_2|$ is not too large. Hence, the time-frequency transfer function of an underspread operator is approximately unique. We note that small $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ requires the spreading function to be concentrated along the τ and/or ν axes (i.e., not oriented in oblique directions). Important examples are quasi-time-invariant and quasi-frequency-invariant systems and parallel connections (i.e., sums) thereof. Since small $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ do not require the operator's spreading function to be exactly supported within a rectangular region about the origin, the bounds in (12) are more general than previously derived bounds.[42] For example, $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ will be small for an operator whose spreading function is concentrated within the region $|\tau\nu|\leq c$ with small c [cf. Fig. 1(f)]. We also note that results in a similar spirit have been derived by Kohn and Nirenberg[37] for the special case $\alpha_1=1/2$ and $\alpha_2=-1/2$ and by Folland[15] for the special case $\alpha_1=0$ and $\alpha_2=1/2$. Finally, we recall from (10) that $m_{\mathbf{H}}^{(1,1)}\leq\sqrt{m_{\mathbf{H}}^{(2,0)}m_{\mathbf{H}}^{(0,2)}}$ and $M_{\mathbf{H}}^{(1,1)}\leq\sqrt{M_{\mathbf{H}}^{(2,0)}M_{\mathbf{H}}^{(0,2)}}$.

B. Adjoint operator

The kernel of the adjoint $\mathbf{H}^+$ of a linear operator $\mathbf{H}$ with kernel $h(t,t')$ is given by $h^*(t',t)$. In the special case of a time-invariant or frequency-invariant operator, the transfer function of the adjoint operator is obtained by complex conjugation of the original transfer function. In contrast, the generalized Weyl symbol of the adjoint of a general linear operator is easily shown to be $L_{\mathbf{H}^+}^{(\alpha)}(t,f)=L_{\mathbf{H}}^{(-\alpha)*}(t,f)$, which does not equal $L_{\mathbf{H}}^{(\alpha)*}(t,f)$ unless $\alpha=0$. However, the following bounds are obtained from Theorem III.1.

Corollary III.2: For any operator $\mathbf{H}$*, the difference* $\Delta_2^{(\alpha)}(t,f)\triangleq L_{\mathbf{H}^+}^{(\alpha)}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)$ *is bounded as*

$$\frac{|\Delta_2^{(\alpha)}(t,f)|}{\|S_{\mathbf{H}}\|_1}\leq 4\pi|\alpha|m_{\mathbf{H}}^{(1,1)},\quad \frac{\|\Delta_2^{(\alpha)}\|_2}{\|\mathbf{H}\|_2}\leq 4\pi|\alpha|M_{\mathbf{H}}^{(1,1)}. \tag{13}$$

Proof: With $L_{\mathbf{H}^+}^{(\alpha)}(t,f)=L_{\mathbf{H}}^{(-\alpha)*}(t,f)$, we have $\Delta_2^{(\alpha)}(t,f)=L_{\mathbf{H}}^{(-\alpha)*}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)$. Hence, the above bounds follow straightforwardly by applying (12) with $\alpha_1=-\alpha$, $\alpha_2=\alpha$. □

For $|\alpha|\leq 1/2$ one obtains the α-independent (coarser) bounds $2\pi m_{\mathbf{H}}^{(1,1)}$ and $2\pi M_{\mathbf{H}}^{(1,1)}$. For an underspread operator whose spreading function is concentrated along the τ and ν axes so that $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ are small, these bounds show that (for $|\alpha|$ not too large) the generalized Weyl symbol of the adjoint of an operator is approximately equal to the complex conjugate of the operator's generalized Weyl symbol,

$$L_{\mathbf{H}^+}^{(\alpha)}(t,f)\approx L_{\mathbf{H}}^{(\alpha)*}(t,f). \tag{14}$$

Due to the similarity of the bounds in Corollary III.2 and Theorem III.1, the above approximation is valid in the same situations where the generalized Weyl symbol is approximately unique. For the special case $\alpha=-1/2$, a result in a similar spirit has been derived by Kohn and Nirenberg.[37]

An important consequence of Corollary III.2 concerns self-adjoint operators (where $\mathbf{H}^+=\mathbf{H}$). Whereas in the time-invariant or frequency-invariant case the transfer function of a self-adjoint operator is real-valued, the generalized Weyl symbol of a self-adjoint general operator is not real-valued unless $\alpha=0$. However, Corollary III.2 implies that the imaginary part of the generalized Weyl symbol of a self-adjoint operator,

$$\mathcal{I}^{(\alpha)}(t,f)\triangleq\frac{1}{2i}\left[L_{\mathbf{H}}^{(\alpha)}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)\right]=\frac{1}{2i}\left[L_{\mathbf{H}^+}^{(\alpha)}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)\right]=\frac{\Delta_2^{(\alpha)}(t,f)}{2i},$$

is bounded as

$$\frac{|\mathscr{T}^{(\alpha)}(t,f)|}{\|S_{\mathbf{H}}\|_1} \leq 2\pi|\alpha| m_{\mathbf{H}}^{(1,1)} \leq \pi m_{\mathbf{H}}^{(1,1)}, \quad \frac{\|\mathscr{T}^{(\alpha)}\|_2}{\|\mathbf{H}\|_2} \leq 2\pi|\alpha| M_{\mathbf{H}}^{(1,1)} \leq \pi M_{\mathbf{H}}^{(1,1)},$$

the coarser bounds being valid for $|\alpha| \leq 1/2$. Hence, the generalized Weyl symbol of an underspread, self-adjoint operator is approximately real-valued even if $\alpha \neq 0$. This result is useful for a discrete-time implementation since here the generalized Weyl symbol with $\alpha = \pm 1/2$ is preferable. Furthermore, the above result shows the approximate real-valuedness of certain time-varying power spectra[86] in the case $\alpha \neq 0$.

C. Operator products

The transfer function of the product (composition, series connection) of two time-invariant linear operators $\mathbf{H}_1$ and $\mathbf{H}_2$ equals the product of the individual transfer functions, $G_1(f)G_2(f)$. Similarly, the temporal transfer function of the product of two frequency-invariant linear operators is given by $m_1(t)m_2(t)$. This composition property of conventional transfer functions is the cornerstone of many techniques for signal detection, estimation, filtering, channel equalization, etc.

Contrary to the time-invariant/frequency-invariant case, the generalized Weyl symbol of the product $\mathbf{H}_2\mathbf{H}_1$ of two general linear operators can not be obtained by multiplying the individual generalized Weyl symbols of $\mathbf{H}_1$ and $\mathbf{H}_2$. In a signal processing context, this is unfortunate since a multiplicative composition property of the generalized Weyl symbol (a multiplicative symbolic calculus) similar to the time-invariant/frequency-invariant case would allow very simple and intuitive formulations of (statistical) signal processing techniques in nonstationary environments.[42,47–54,87] In the context of quantum mechanics, such a symbolic calculus—i.e., the approximation of the star product of symbols by the ordinary product—would establish an asymptotic correspondence (in the classical limit $h \to 0$, where h denotes Planck's constant) between pointwise products and Jordan products as well as between Poisson brackets and commutators[15] (cf. Subsection IV A). Besides the case where $\mathbf{H}_1$ and $\mathbf{H}_2$ are either both time-invariant or both frequency-invariant operators (or, in the case $\alpha = 0$, any metaplectic transformation thereof), for Hilbert–Schmidt operators the relation $L_{\mathbf{H}_2\mathbf{H}_1}^{(\alpha)}(t,f) = L_{\mathbf{H}_1}^{(\alpha)}(t,f) L_{\mathbf{H}_2}^{(\alpha)}(t,f)$ is correct only in the following situations.

(i) For $\alpha = 1/2$, if $\mathbf{H}_1$ is a linear time-invariant operator with spectral transfer function $G_1(f)$ or $\mathbf{H}_2$ is a linear frequency-invariant operator with temporal transfer function $m_2(t)$, it can be shown that

$$L_{\mathbf{H}_2\mathbf{H}_1}^{(1/2)}(t,f) = L_{\mathbf{H}_1}^{(1/2)}(t,f) L_{\mathbf{H}_2}^{(1/2)}(t,f) = \begin{cases} G_1(f) L_{\mathbf{H}_2}^{(1/2)}(t,f), & \text{for } \mathbf{H}_1 \text{ time-invariant,} \\ L_{\mathbf{H}_1}^{(1/2)}(t,f) m_2(t), & \text{for } \mathbf{H}_2 \text{ frequency-invariant.} \end{cases} \tag{15}$$

(ii) In the dual case $\alpha = -1/2$, if $\mathbf{H}_1$ is linear frequency-invariant with temporal transfer function $m_1(t)$ or $\mathbf{H}_2$ is linear time-invariant with spectral transfer function $G_2(f)$, one obtains

$$L_{\mathbf{H}_2\mathbf{H}_1}^{(-1/2)}(t,f) = L_{\mathbf{H}_1}^{(-1/2)}(t,f) L_{\mathbf{H}_2}^{(-1/2)}(t,f) = \begin{cases} m_1(t) L_{\mathbf{H}_2}^{(-1/2)}(t,f), & \text{for } \mathbf{H}_1 \text{ frequency-invariant,} \\ L_{\mathbf{H}_1}^{(-1/2)}(t,f) G_2(f), & \text{for } \mathbf{H}_2 \text{ time-invariant.} \end{cases} \tag{16}$$

In general, however, the generalized Weyl symbol of operator products $\mathbf{H}_2\mathbf{H}_1$ is described by the so-called star (or twisted) product[15,31,33,38,40,41,56,69,88,89] $L_{\mathbf{H}_2}^{(\alpha)} \# L_{\mathbf{H}_1}^{(\alpha)}$, which (when generalized to arbitrary α) can be formulated as

$$(L^{(\alpha)}_{\mathbf{H}_2}\#L^{(\alpha)}_{\mathbf{H}_1})(t,f)\triangleq L^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(t,f)$$

$$=\sum_{k,l=0}^{\infty} c_{kl}\,\frac{\partial^{k+l}L^{(\alpha)}_{\mathbf{H}_2}(t,f)}{\partial t^l\partial f^k}\,\frac{\partial^{k+l}L^{(\alpha)}_{\mathbf{H}_1}(t,f)}{\partial t^k\partial f^l}$$

$$=L^{(\alpha)}_{\mathbf{H}_2}(t,f)L^{(\alpha)}_{\mathbf{H}_1}(t,f)+\sum_{k+l>0} c_{kl}\,\frac{\partial^{k+l}L^{(\alpha)}_{\mathbf{H}_2}(t,f)}{\partial t^l\partial f^k}\,\frac{\partial^{k+l}L^{(\alpha)}_{\mathbf{H}_1}(t,f)}{\partial t^k\partial f^l},$$

with $c_{kl}=(-1)^{k+l}(\alpha+1/2)^k(\alpha-1/2)^l/[(4\pi^2)^{k+l}k!l!]$. It is thus seen that the difference between $L^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(t,f)$ and $L^{(\alpha)}_{\mathbf{H}_2}(t,f)L^{(\alpha)}_{\mathbf{H}_1}(t,f)$ is essentially determined by the higher-order partial derivatives of the symbols. This difference will be small if these derivatives are small, i.e., if the symbols are smooth functions. The following subsections describe quantitative formulations of this qualitative discussion which are simple and relevant to signal processing applications.

1. Approximate multiplicative symbol calculus

A positive statement (in the spirit of earlier results obtained in the context of quantization,[14,15,39,56,88] pseudo-differential operators,[15,37,38,40,41,56] and transfer function calculus[42–44]) on the general validity of an approximate multiplicative transfer function (symbolic) calculus is provided by the following theorem.

Theorem III.3: *For any two operators* $\mathbf{H}_1$ *and* $\mathbf{H}_2$, *the difference* $\Delta^{(\alpha)}_3(t,f)\triangleq L^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(t,f)-L^{(\alpha)}_{\mathbf{H}_1}(t,f)L^{(\alpha)}_{\mathbf{H}_2}(t,f)$ *is bounded as*

$$\frac{|\Delta^{(\alpha)}_3(t,f)|}{\|S_{\mathbf{H}_1}\|_1\|S_{\mathbf{H}_2}\|_1}\leq 2\pi B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2},\quad \textit{with}\quad B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}\triangleq|\alpha+\tfrac{1}{2}|m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}+|\alpha-\tfrac{1}{2}|m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}. \quad (17)$$

Proof: The spreading function of the operator product $\mathbf{H}_2\mathbf{H}_1$ is given by the so-called twisted convolution,[15,38,42]

$$S^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(\tau,\nu)=(S^{(\alpha)}_{\mathbf{H}_2}\natural S^{(\alpha)}_{\mathbf{H}_1})(\tau,\nu)\triangleq\int_{\tau'}\int_{\nu'}S^{(\alpha)}_{\mathbf{H}_2}(\tau',\nu')S^{(\alpha)}_{\mathbf{H}_1}(\tau-\tau',\nu-\nu')e^{-i2\pi\phi_\alpha(\tau,\nu,\tau',\nu')}d\tau'd\nu', \quad (18)$$

with $\phi_\alpha(\tau,\nu,\tau',\nu')=(\alpha+1/2)\tau'(\nu-\nu')+(\alpha-1/2)(\tau-\tau')\nu'$ [note that the twisted convolution and the star (twisted) product are 2-D Fourier transform pairs]. Hence, the Fourier transform of $\Delta^{(\alpha)}_3(t,f)$ is obtained as

$$\hat{\Delta}^{(\alpha)}_3(\tau,\nu)=(S^{(\alpha)}_{\mathbf{H}_2}\natural S^{(\alpha)}_{\mathbf{H}_1})(\tau,\nu)-(S^{(\alpha)}_{\mathbf{H}_2}**S^{(\alpha)}_{\mathbf{H}_1})(\tau,\nu)$$

$$=\int_{\tau'}\int_{\nu'}S^{(\alpha)}_{\mathbf{H}_2}(\tau',\nu')S^{(\alpha)}_{\mathbf{H}_1}(\tau-\tau',\nu-\nu')[e^{-i2\pi\phi_\alpha(\tau,\nu,\tau',\nu')}-1]d\tau'd\nu',$$

where $**$ denotes ordinary 2-D convolution. Thus, we obtain

$$|\Delta^{(\alpha)}_3(t,f)|\leq\int_\tau\int_\nu|\hat{\Delta}^{(\alpha)}_3(\tau,\nu)|d\tau d\nu$$

$$\leq 2\int_\tau\int_\nu\int_{\tau'}\int_{\nu'}|S_{\mathbf{H}_2}(\tau',\nu')||S_{\mathbf{H}_1}(\tau-\tau',\nu-\nu')|$$

$$\times|\sin(\pi\phi_\alpha(\tau,\nu,\tau',\nu'))|d\tau d\nu d\tau' d\nu'.$$

Substituting $\tau_1=\tau-\tau'$, $\nu_1=\nu-\nu'$ and using $|\sin x|\leq|x|$ yields the bound (17) after straightforward manipulations. □

By virtue of its symmetry with respect to $\mathbf{H}_1$ and $\mathbf{H}_2$, the bound in (17) also applies to the differences $L^{(\alpha)}_{\mathbf{H}_1\mathbf{H}_2}(t,f)-L^{(\alpha)}_{\mathbf{H}_1}(t,f)L^{(\alpha)}_{\mathbf{H}_2}(t,f)$ and $L^{(\alpha)}_{\mathbf{H}_1\star\mathbf{H}_2}(t,f)-L^{(\alpha)}_{\mathbf{H}_1}(t,f)L^{(\alpha)}_{\mathbf{H}_2}(t,f)$, where $\mathbf{H}_1\star\mathbf{H}_2=(\mathbf{H}_1\mathbf{H}_2+\mathbf{H}_2\mathbf{H}_1)/2$ is the (commutative) Jordan product[15] of $\mathbf{H}_1$ and $\mathbf{H}_2$.

The bound shows that if $\mathbf{H}_1$ and $\mathbf{H}_2$ are ''jointly underspread'' such that $m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}$ and $m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}$ are both small, then we have an approximate multiplicative composition property,

$$L^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(t,f)\approx L^{(\alpha)}_{\mathbf{H}_1}(t,f)L^{(\alpha)}_{\mathbf{H}_2}(t,f). \tag{19}$$

This approximate relation provides a basis for several signal processing methods for nonstationary environments recently proposed for a variety of problems like equalization of mobile radio channels,[48,49] estimation of nonstationary random signals corrupted by noise,[42,47,50,51,53,87] and detection or classification of nonstationary signals.[52,53,87] Furthermore, the above approximation can be used to show that two important time-varying power spectra of nonstationary random processes—the Wigner–Ville spectrum[86,90] and the (generalized) evolutionary spectrum[54,91]—are basically equivalent for underspread processes.[42,54,55,87]

We note that (19) implies $L^{(\alpha)}_{\mathbf{H}_2\mathbf{H}_1}(t,f)\approx L^{(\alpha)}_{\mathbf{H}_1\mathbf{H}_2}(t,f)$, which suggests that two jointly underspread operators are approximately commuting, i.e., $\mathbf{H}_2\mathbf{H}_1\approx\mathbf{H}_1\mathbf{H}_2$. This issue will be investigated further in Subsection IV A.

In the following, we briefly discuss the approximation (19) for the cases $\alpha=\pm1/2$. The case $\alpha=0$ deserves special attention and will be considered in more detail afterwards.

For $\alpha=1/2$, $B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}$ simplifies to $B^{(1/2)}_{\mathbf{H}_1,\mathbf{H}_2}=m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}$, which will be small if $|S_{\mathbf{H}_1}(\tau,\nu)|$ is concentrated along the τ axis and/or $|S_{\mathbf{H}_2}(\tau,\nu)|$ is concentrated along the ν axis. Thus, for $\alpha=1/2$ the approximation (19) is good when $\mathbf{H}_1$ is a quasi-time-invariant system and/or $\mathbf{H}_2$ is a quasi-frequency-invariant system [note the consistency with the exact multiplicative calculus in (15)].

For $\alpha=-1/2$, we have $B^{(-1/2)}_{\mathbf{H}_1,\mathbf{H}_2}=m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}$. This will be small if $|S_{\mathbf{H}_1}(\tau,\nu)|$ is concentrated along the ν axis and/or $|S_{\mathbf{H}_2}(\tau,\nu)|$ is concentrated along the τ axis, i.e., if $\mathbf{H}_1$ is quasi-frequency-invariant and/or $\mathbf{H}_2$ is quasi-time-invariant [note the consistency with (16)].

For both $\alpha=1/2$ and $\alpha=-1/2$, the bound is large, and thus the approximation (19) is poor, if the operators' spreading functions are oriented in oblique directions. For any α with $|\alpha|\leq1/2$, it follows from the definition of $B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}$ in (17) that $B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}$ is a convex combination of the two extreme cases $B^{(1/2)}_{\mathbf{H}_1,\mathbf{H}_2}=m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}$ and $B^{(-1/2)}_{\mathbf{H}_1,\mathbf{H}_2}=m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}$, i.e.,

$$B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}=\gamma B^{(1/2)}_{\mathbf{H}_1,\mathbf{H}_2}+(1-\gamma)B^{(-1/2)}_{\mathbf{H}_1,\mathbf{H}_2},\quad\text{with}\quad 0\leq\gamma=\alpha+\tfrac{1}{2}\leq1.$$

In particular, this implies

$$\min\{B^{(1/2)}_{\mathbf{H}_1,\mathbf{H}_2},B^{(-1/2)}_{\mathbf{H}_1,\mathbf{H}_2}\}\leq B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}\leq\max\{B^{(1/2)}_{\mathbf{H}_1,\mathbf{H}_2},B^{(-1/2)}_{\mathbf{H}_1,\mathbf{H}_2}\},\quad|\alpha|\leq\tfrac{1}{2}.$$

2. Case $\alpha=0$

The case $\alpha=0$ is of particular importance. Here, $B^{(\alpha)}_{\mathbf{H}_1,\mathbf{H}_2}$ simplifies to

$$B^{(0)}_{\mathbf{H}_1,\mathbf{H}_2}=\tfrac{1}{2}\,[m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}+m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}],$$

which is symmetric with respect to $\mathbf{H}_1$ and $\mathbf{H}_2$. Thus, the approximation (19) will be good if both $m^{(0,1)}_{\mathbf{H}_1}m^{(1,0)}_{\mathbf{H}_2}$ and $m^{(1,0)}_{\mathbf{H}_1}m^{(0,1)}_{\mathbf{H}_2}$ are small. This requires that the spreading functions of $\mathbf{H}_1$ and $\mathbf{H}_2$ are both concentrated about the origin of the (τ,ν) plane, with *similar* orientation parallel to the τ or ν axis. However, this requirement is relaxed by a refinement of the bound $2\pi B^{(0)}_{\mathbf{H}_1,\mathbf{H}_2}$ that is stated in the next theorem. This improved bound can be attributed to the covariance of the Weyl symbol $L^{(0)}_{\mathbf{H}}(t,f)$ to metaplectic transformations $\mathbf{U}\in\mathscr{M}$ [cf. (4)].

Theorem III.4: *For any two operators* $\mathbf{H}_1$ *and* $\mathbf{H}_2$, *the difference* $\Delta_3^{(0)}(t,f)=L_{\mathbf{H}_2\mathbf{H}_1}^{(0)}(t,f)-L_{\mathbf{H}_1}^{(0)}(t,f)L_{\mathbf{H}_2}^{(0)}(t,f)$ *obeys the bound [that is generally tighter than (*17*)]*

$$\frac{|\Delta_3^{(0)}(t,f)|}{\|S_{\mathbf{H}_1}\|_1\|S_{\mathbf{H}_2}\|_1}\leqslant 2\pi\min_{\mathbf{U}\in\mathscr{M}} B_{\mathbf{U}\mathbf{H}_1\mathbf{U}^+,\mathbf{U}\mathbf{H}_2\mathbf{U}^+}^{(0)}. \tag{20}$$

Proof: Specializing the proof of Theorem III.3 to $\alpha=0$, it is seen that

$$|\Delta_3^{(0)}(t,f)|\leqslant\pi\int_{\tau_1}\int_{\nu_1}\int_{\tau_2}\int_{\nu_2}|S_{\mathbf{H}_2}(\tau_1,\nu_1)||S_{\mathbf{H}_1}(\tau_2,\nu_2)||\tau_1\nu_2-\tau_2\nu_1|d\tau_1 d\nu_1 d\tau_2 d\nu_2. \tag{21}$$

Performing a symplectic coordinate transform $\binom{\tau_i}{\nu_i}=\mathscr{A}\binom{\tau_i'}{\nu_i'}$, where $\mathscr{A}=\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)$ with $\det\mathscr{A}=ad-bc=1$, and using the invariance of the symplectic form,[15,73,75] $\tau_1\nu_2-\tau_2\nu_1=\tau_1'\nu_2'-\tau_2'\nu_1'$, the right-hand side of (21) becomes

$$\pi\int_{\tau_1'}\int_{\nu_1'}\int_{\tau_2'}\int_{\nu_2'}|S_{\mathbf{H}_2}(a\tau_1'+b\nu_1',c\tau_1'+d\nu_1')||S_{\mathbf{H}_1}(a\tau_2'+b\nu_2',c\tau_2'+d\nu_2')|$$
$$\times|\tau_1'\nu_2'-\tau_2'\nu_1'|d\tau_1' d\nu_1' d\tau_2' d\nu_2'.$$

Let $\tilde{\mathbf{H}}_i=\mathbf{U}\mathbf{H}_i\mathbf{U}^+$ where $\mathbf{U}\in\mathscr{M}$ is the unitary operator corresponding to $\mathscr{A}$ via the metaplectic representation.[15] Using the covariance property (8), the above expression becomes

$$\pi\int_{\tau_1'}\int_{\nu_1'}\int_{\tau_2'}\int_{\nu_2'}|S_{\tilde{\mathbf{H}}_2}(\tau_1',\nu_1')||S_{\tilde{\mathbf{H}}_1}(\tau_2',\nu_2')||\tau_1'\nu_2'-\tau_2'\nu_1'|d\tau_1' d\nu_1' d\tau_2' d\nu_2'.$$

Inserting $|\tau_1'\nu_2'-\tau_2'\nu_1'|\leqslant|\tau_1'\nu_2'|+|\tau_2'\nu_1'|$ and using $\|S_{\tilde{\mathbf{H}}_i}\|_1=\|S_{\mathbf{H}_i}\|_1$ (which holds since the spreading function magnitude of $\tilde{\mathbf{H}}_i$ is obtained from that of $\mathbf{H}_i$ via an area-preserving coordinate transform), we obtain

$$|\Delta_3^{(0)}(t,f)|\leqslant\pi\|S_{\mathbf{H}_1}\|_1\|S_{\mathbf{H}_2}\|_1[m_{\tilde{\mathbf{H}}_1}^{(0,1)}m_{\tilde{\mathbf{H}}_2}^{(1,0)}+m_{\tilde{\mathbf{H}}_1}^{(1,0)}m_{\tilde{\mathbf{H}}_2}^{(0,1)}].$$

As this bound is valid for all $\mathbf{U}\in\mathscr{M}$, we finally obtain the bound (20). □

Since the symplectic group contains time-frequency rotations and time-frequency shearings, this theorem shows that $\Delta_3^{(0)}(t,f)$ may be small even if the spreading functions of $\mathbf{H}_1$ and $\mathbf{H}_2$ are oriented in (similar) *oblique* directions. This is not true for $\alpha\neq0$, which once again shows the exceptional position of the Weyl symbol.

3. Composition of H *with* $\mathbf{H}^+$

In some applications, e.g., for an innovations system representation of random processes[54] or for a quadratic time-frequency analysis of linear operators,[92] the composition of $\mathbf{H}$ with $\mathbf{H}^+$ is of importance (cf. also the proofs of Theorems III.6 and III.12). For time-invariant or frequency-invariant operators, the transfer function of $\mathbf{H}^+\mathbf{H}$ is $|G(f)|^2$ or $|m(t)|^2$, respectively. For general operators, the following result can be obtained from Theorem III.3 and Corollary III.2.

Corollary III.5: For any operator $\mathbf{H}$, *the difference* $\Delta_4^{(\alpha)}(t,f)\triangleq L_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(t,f)-|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2$ *is bounded as*

$$\frac{|\Delta_4^{(\alpha)}(t,f)|}{\|S_{\mathbf{H}}\|_1^2}\leqslant 2\pi C_{\mathbf{H}}^{(\alpha)},\quad with\quad C_{\mathbf{H}}^{(\alpha)}\triangleq c_\alpha m_{\mathbf{H}}^{(0,1)}m_{\mathbf{H}}^{(1,0)}+2|\alpha|m_{\mathbf{H}}^{(1,1)}, \tag{22}$$

where $c_\alpha=|\alpha+1/2|+|\alpha-1/2|$.

Proof: Subtracting and adding $L_{\mathbf{H}}^{(\alpha)}(t,f)L_{\mathbf{H}^+}^{(\alpha)}(t,f)$ from/to $\Delta_4^{(\alpha)}(t,f)$, we obtain

$$|\Delta_4^{(\alpha)}(t,f)| \leq |L_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(t,f) - L_{\mathbf{H}}^{(\alpha)}(t,f) L_{\mathbf{H}^+}^{(\alpha)}(t,f)| + |L_{\mathbf{H}^+}^{(\alpha)}(t,f) - L_{\mathbf{H}}^{(\alpha)*}(t,f)||L_{\mathbf{H}}^{(\alpha)}(t,f)|$$
$$\leq 2\pi c_\alpha \|S_{\mathbf{H}}\|_1^2 m_{\mathbf{H}}^{(1,0)} m_{\mathbf{H}}^{(0,1)} + 4\pi|\alpha| m_{\mathbf{H}}^{(1,1)} \|S_{\mathbf{H}}\|_1 |L_{\mathbf{H}}^{(\alpha)}(t,f)|,$$

where we used (17) with $m_{\mathbf{H}^+}^{(k,l)} = m_{\mathbf{H}}^{(k,l)}$ and $\|S_{\mathbf{H}^+}\|_1 = \|S_{\mathbf{H}}\|_1$ as well as (13). From this, (22) follows with $|L_{\mathbf{H}}^{(\alpha)}(t,f)| \leq \|S_{\mathbf{H}}\|_1$. □

Hence, for an underspread operator with small $m_{\mathbf{H}}^{(0,1)} m_{\mathbf{H}}^{(1,0)}$ and small $m_{\mathbf{H}}^{(1,1)}$, we have

$$L_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(t,f) \approx |L_{\mathbf{H}}^{(\alpha)}(t,f)|^2,$$

i.e., ''squaring'' the operator is approximately equivalent to squaring its generalized Weyl symbol. We note that the bound (22), and thus the above approximation, remain valid if $\mathbf{H}^+\mathbf{H}$ is replaced by $\mathbf{H}\mathbf{H}^+$ or by $\mathbf{H} \star \mathbf{H}^+ = (\mathbf{H}\mathbf{H}^+ + \mathbf{H}^+\mathbf{H})/2$.

For $|\alpha| \leq 1/2$, we have $c_\alpha = 1$ and thus $C_{\mathbf{H}}^{(\alpha)} \leq m_{\mathbf{H}}^{(1,0)} m_{\mathbf{H}}^{(0,1)} + m_{\mathbf{H}}^{(1,1)}$. The bound $C_{\mathbf{H}}^{(\alpha)}$ is tightest for $\alpha = 0$, in which case $C_{\mathbf{H}}^{(0)} = m_{\mathbf{H}}^{(0,1)} m_{\mathbf{H}}^{(1,0)}$. This is due to the fact that the Weyl symbol of the adjoint is exactly obtained by complex conjugation. Furthermore, for $\alpha = 0$ the above result can be refined similarly to Theorem III.4, yielding the tighter bound

$$\frac{|\Delta_4^{(0)}(t,f)|}{\|S_{\mathbf{H}}\|_1^2} \leq 2\pi \min_{\mathbf{U} \in \mathcal{M}} C_{\mathbf{U}\mathbf{H}\mathbf{U}^+}^{(0)} = 2\pi \min_{\mathbf{U} \in \mathcal{M}} m_{\mathbf{U}\mathbf{H}\mathbf{U}^+}^{(0,1)} m_{\mathbf{U}\mathbf{H}\mathbf{U}^+}^{(1,0)}. \tag{23}$$

D. Approximate eigenvalues and eigenfunctions

The response of a time-invariant system to a complex sinusoid $e^{i2\pi f_0 t}$—i.e., a signal with perfect frequency concentration—is $G(f_0)e^{i2\pi f_0 t}$. Thus, the complex sinusoids $e^{i2\pi f_0 t}$ are the generalized[60] eigenfunctions of any time-invariant system, with $G(f_0)$ the associated generalized eigenvalues. Similarly, the response of a frequency-invariant system to a Dirac impulse $\delta(t - t_0)$—i.e., a signal perfectly localized in time—is given by $m(t_0)\delta(t - t_0)$, and hence the Dirac impulses $\delta(t - t_0)$ are the generalized eigenfunctions with $m(t_0)$ the associated generalized eigenvalues. Note that in the time-invariant and frequency-invariant cases the eigenfunctions are highly structured since they are obtained by frequency shifts of the unity function 1 and by time shifts of the Dirac impulse $\delta(t)$, respectively; moreover, the eigenvalues are given by the values of the transfer function.

On the other hand, the eigenfunctions (singular functions) of general linear operators are not localized and structured in any sense, and the eigenvalues (singular values) are not given by the transfer function values. However, we will now show that *underspread* operators have a well-structured set of time-frequency-localized ''approximate eigenfunctions,'' with the associated ''approximate eigenvalues'' given by the values of the generalized Weyl symbol. Note that it is not necessary to consider approximate singular functions since underspread operators are approximately normal—see Subsection IV B.

Let $s(t)$ be a normalized function that is well concentrated about the origin of the time-frequency plane (e.g., a Gaussian function). We consider the family of functions $s_{t_0,f_0}(t) = s(t - t_0)e^{i2\pi f_0 t}$ obtained by time-frequency-shifting $s(t)$ to the time-frequency point (t_0, f_0). By construction, $s_{t_0,f_0}(t)$ is then well time-frequency-concentrated about (t_0, f_0). The practical usefulness of the set $\{s_{t_0,f_0}(t)\}$ stems from the fact that it is highly structured and parameterized by physically meaningful variables. Such time-delayed and frequency-shifted signal sets are particularly appealing in communication applications, for example, in mobile radio and digital broadcasting.[93–96] We shall now state conditions under which the response of a linear operator $\mathbf{H}$ to $s_{t_0,f_0}(t)$ is approximately $L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)s_{t_0,f_0}(t)$, which implies that $s_{t_0,f_0}(t)$ is an ''approximate eigenfunction'' of $\mathbf{H}$ with $L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)$ the associated ''approximate eigenvalue.'' The next theorem is essentially an adaptation and extension of an existing result on approximate eigenpairs.[42,95] (Approximate eigenpairs and approximate operator diagonalization have previously been considered also at a mathematically more sophisticated level.[14,15,17,97–99]) In the following,

$$A_s^{(\alpha)}(\tau,\nu) \triangleq \int_t s(t+(\tfrac{1}{2}-\alpha)\tau)s^*(t-(\tfrac{1}{2}+\alpha)\tau)e^{-i2\pi\nu t}dt$$

denotes the (generalized) ambiguity function[100,101] of the function $s(t)$, which equals the spreading function of the rank-one operator with kernel $s(t)s^*(t')$. We note that $A_s^{(\alpha)}(0,0)=\|s\|_2^2=1$.

Theorem III.6: *For any operator* $\mathbf{H}$*, any time-frequency point* (t_0,f_0)*, and any normalized function* $s(t)$ *(i.e.,* $\|s\|_2=1$*), the difference* $\Delta_5^{(\alpha)}(t) \triangleq (\mathbf{H}s_{t_0,f_0})(t) - L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)s_{t_0,f_0}(t)$ *is bounded as*

$$\frac{\|\Delta_5^{(\alpha)}\|_2}{\|S_{\mathbf{H}}\|_1} \leq D_{\mathbf{H},s}^{(\alpha)}, \quad \text{with} \quad D_{\mathbf{H},s}^{(\alpha)} \triangleq \sqrt{2\pi C_{\mathbf{H}}^{(\alpha)} + m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)} + 2m_{\mathbf{H}}^{(\phi_s)}}, \tag{24}$$

with the weighting function $\phi_s(\tau,\nu)=|1-A_s^{(\alpha)}(\tau,\nu)|$ *and* $C_{\mathbf{H}}^{(\alpha)}$ *as defined in* (22).

Proof: First, we develop the square of $\|\Delta_5^{(\alpha)}\|_2$ as

$$\begin{aligned}
\|\Delta_5^{(\alpha)}\|_2^2 &= \|\mathbf{H}s_{t_0,f_0}\|_2^2 - 2\,\mathrm{Re}\{L_{\mathbf{H}}^{(\alpha)*}(t_0,f_0)\langle \mathbf{H}s_{t_0,f_0}, s_{t_0,f_0}\rangle\} + |L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|^2\|s_{t_0,f_0}\|_2^2 \\
&= \|\mathbf{H}s_{t_0,f_0}\|_2^2 - 2\,\mathrm{Re}\{L_{\mathbf{H}}^{(\alpha)*}(t_0,f_0)[\langle \mathbf{H}s_{t_0,f_0}, s_{t_0,f_0}\rangle - L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)]\} - |L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|^2 \\
&\leq \big|\|\mathbf{H}s_{t_0,f_0}\|_2^2 - |L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|^2\big| + 2|L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|\,|\langle \mathbf{H}s_{t_0,f_0}, s_{t_0,f_0}\rangle - L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|,
\end{aligned} \tag{25}$$

where we used $\|s_{t_0,f_0}\|_2=\|s\|_2=1$. Using

$$\langle \mathbf{G}s_{t_0,f_0}, s_{t_0,f_0}\rangle = \int_\tau\int_\nu S_{\mathbf{G}}^{(\alpha)}(\tau,\nu)A_s^{(\alpha)*}(\tau,\nu)e^{i2\pi(\nu t_0-\tau f_0)}d\tau d\nu, \tag{26}$$

with $\mathbf{G}=\mathbf{H}^+\mathbf{H}$, and recalling the definition of $\Delta_4^{(\alpha)}(t,f)$ in Corollary III.5, the first term in (25) becomes

$$\begin{aligned}
\big|\|\mathbf{H}s_{t_0,f_0}\|_2^2 - |L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|^2\big| &= |\langle \mathbf{H}^+\mathbf{H}s_{t_0,f_0}, s_{t_0,f_0}\rangle - L_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(t_0,f_0) + L_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(t_0,f_0) - |L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)|^2| \\
&= \left|\int_\tau\int_\nu S_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(\tau,\nu)[A_s^{(\alpha)*}(\tau,\nu)-1]e^{i2\pi(\nu t_0-\tau f_0)}d\tau d\nu + \Delta_4^{(\alpha)}(t_0,f_0)\right| \\
&\leq \int_\tau\int_\nu |S_{\mathbf{H}^+\mathbf{H}}(\tau,\nu)|\,|1-A_s^{(\alpha)}(\tau,\nu)|\,d\tau d\nu + |\Delta_4^{(\alpha)}(t_0,f_0)| \\
&\leq \|S_{\mathbf{H}^+\mathbf{H}}\|_1 m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)} + \|S_{\mathbf{H}}\|_1^2 2\pi C_{\mathbf{H}}^{(\alpha)} \leq \|S_{\mathbf{H}}\|_1^2[m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)} + 2\pi C_{\mathbf{H}}^{(\alpha)}],
\end{aligned}$$

where we used (22) and Young's inequality for twisted convolution,[15] $\|S_{\mathbf{H}^+\mathbf{H}}\|_1 \leq \|S_{\mathbf{H}}\|_1^2$. In a similar manner, the factor of the second term in (25) can be developed and bounded as

$$|\langle \mathbf{H}s_{t_0,f_0}, s_{t_0,f_0}\rangle - L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)| \leq \int_\tau\int_\nu |S_{\mathbf{H}}(\tau,\nu)|\,|1-A_s^{(\alpha)}(\tau,\nu)|\,d\tau d\nu = \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}.$$

Inserting these two bounds into (25) and using $|L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)| \leq \|S_{\mathbf{H}}\|_1$ yields the bound (24). □

For an underspread operator $\mathbf{H}$ where $D_{\mathbf{H},s}^{(\alpha)} = \sqrt{2\pi C_{\mathbf{H}}^{(\alpha)} + m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)} + 2m_{\mathbf{H}}^{(\phi_s)}}$ can be made small, we thus have the ''approximate eigenvalue relation,''

$$(\mathbf{H}s_{t_0,f_0})(t) \approx L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)s_{t_0,f_0}(t),$$

which implies that properly time-frequency-localized functions are approximate eigenfunctions and the generalized Weyl symbol $L_{\mathbf{H}}^{(\alpha)}(t,f)$ is an approximate eigenvalue distribution over the time-frequency plane. In particular, small $m_{\mathbf{H}}^{(\phi_s)}$ and $m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$ requires that, on the effective supports of $|S_{\mathbf{H}}(\tau,\nu)|$ and of $|S_{\mathbf{H}^+\mathbf{H}}(\tau,\nu)|$, there is $\phi_s(\tau,\nu)\approx 0$ [recall the definition (9)] and thus

$A_s^{(\alpha)}(\tau,\nu)\approx A_s^{(\alpha)}(0,0)\equiv 1$. The latter condition can be satisfied only if these effective supports are small, i.e., if $\mathbf{H}$ is underspread. [Recall from (18) that $S_{\mathbf{H}^+\mathbf{H}}^{(\alpha)}(\tau,\nu)=(S_{\mathbf{H}^+}^{(\alpha)}\natural S_{\mathbf{H}}^{(\alpha)})(\tau,\nu)$, and note that this twisted convolution enlarges the effective support of $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)$ maximally by a factor of four.[42,44]] Moreover, $A_s^{(\alpha)}(\tau,\nu)\approx A_s^{(\alpha)}(0,0)$ around the origin implies that the second-order derivatives of $A_s^{(\alpha)}(\tau,\nu)$ with respect to τ and ν are small at the origin. Since these second-order derivatives correspond to the effective duration and bandwidth[102,103] of $s(t)$, it is seen that the function $s(t)$ needs to have good time-frequency concentration.

The bound $D_{\mathbf{H},s}^{(\alpha)}$ is tightest for $\alpha=0$ since here $C_{\mathbf{H}}^{(\alpha)}$ is smallest. Furthermore, $C_{\mathbf{H}}^{(0)}$ can be replaced by $\min_{\mathbf{U}\in\mathscr{U}} C_{\mathbf{UHU}^+}^{(0)}$ which allows oblique orientation of the spreading function of $\mathbf{H}$. We note that $m_{\mathbf{H}}^{(\phi_s)}$ and $m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$ can be adapted to an oblique orientation of the spreading function by a proper choice of the function $s(t)$.

E. Input-output relation for deterministic signals

For time-invariant systems, due to the fact that the complex sinusoids are eigenfunctions, the Fourier transform of the output signal $(\mathbf{H}x)(t)$ is given by $G(f)X(f)$. Similarly, for frequency-invariant systems $(\mathbf{H}x)(t)$ equals $m(t)x(t)$. In the case of general operators, one may desire a similar ''input-output relation'' stating that a suitably defined time-frequency representation of $(\mathbf{H}x)(t)$ equals the time-frequency representation of $x(t)$ multiplied by the time-frequency transfer function $L_{\mathbf{H}}^{(\alpha)}(t,f)$. To derive such an input-output relation, we note that the input signal $x(t)$ can be expanded[15,17,21–23] into the set of time-frequency-shifted versions $s_{t_0,f_0}(t)=s(t-t_0)e^{i2\pi f_0 t}$ of a normalized function $s(t)$,

$$x(t)=\int_{t_0}\int_{f_0}\mathrm{STFT}_x^{(s)}(t_0,f_0)s_{t_0,f_0}(t)\,dt_0\,df_0, \tag{27}$$

where

$$\mathrm{STFT}_x^{(s)}(t,f)=\int_{t'}x(t')s^*(t'-t)e^{-i2\pi f t'}dt'=\langle x,s_{t,f}\rangle \tag{28}$$

is the short-time Fourier transform[9,21,22,24] of the signal $x(t)$ using $s(t)$ as an analysis window. For Gaussian $s(t)$, the short-time Fourier transform is equivalent to the coherent states expansion of quantum physics.[15–18,104] In what follows, we will adopt the short-time Fourier transform as a time-frequency signal representation. Assuming $s(t)$ to have good time-frequency concentration and $\mathbf{H}$ to be underspread, we know from Subsection III D that the $s_{t_0,f_0}(t)$ are approximate eigenfunctions, i.e., $(\mathbf{H}s_{t_0,f_0})(t)\approx L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)s_{t_0,f_0}(t)$. Hence, applying $\mathbf{H}$ to both sides of (27) we obtain

$$\begin{aligned}(\mathbf{H}x)(t)&=\int_{t_0}\int_{f_0}\mathrm{STFT}_x^{(s)}(t_0,f_0)(\mathbf{H}s_{t_0,f_0})(t)\,dt_0\,df_0\\&\approx\int_{t_0}\int_{f_0}\mathrm{STFT}_x^{(s)}(t_0,f_0)L_{\mathbf{H}}^{(\alpha)}(t_0,f_0)s_{t_0,f_0}(t)\,dt_0\,df_0.\end{aligned}$$

Comparing with (27) [with $x(t)$ replaced by $(\mathbf{H}x)(t)$], this suggests that $\mathrm{STFT}_{\mathbf{H}x}^{(s)}(t,f)\approx\mathrm{STFT}_x^{(s)}(t,f)L_{\mathbf{H}}^{(\alpha)}(t,f)$. Hence, the generalized Weyl symbol can be interpreted as a weighting function in the time-frequency (short-time Fourier transform) domain, thus motivating a design and implementation of time-varying filters based on the short-time Fourier transform (see Subsection III F). The next theorem gives a bound on the quality of the above approximation.

Theorem III.7: *For any operator* $\mathbf{H}$, *any signal* $x(t)$, *and any normalized function* $s(t)$, *the difference* $\Delta_6^{(\alpha)}(t,f)\triangleq\mathrm{STFT}_{\mathbf{H}x}^{(s)}(t,f)-L_{\mathbf{H}}^{(\alpha)}(t,f)\mathrm{STFT}_x^{(s)}(t,f)$ *is bounded as*

$$\frac{|\Delta_6^{(\alpha)}(t,f)|}{\|S_{\mathbf{H}}\|_1\|x\|_2}\leq D_{\mathbf{H}^+,s}^{(\alpha)}+4\pi|\alpha|m_{\mathbf{H}}^{(1,1)},\quad \frac{\|\Delta_6^{(\alpha)}\|_2}{\|\mathbf{H}\|_2\|x\|_2}\leq\sqrt{2}M_{\mathbf{H}}^{(\phi_s')}+4\pi|\alpha|M_{\mathbf{H}}^{(1,1)}, \tag{29}$$

with the weighting functions $\phi_s(\tau,\nu)=|1-A_s^{(\alpha)}(\tau,\nu)|$ *[used in* $D^{(\alpha)}_{\mathbf{H}^+,s}$*; see (24)] and* $\phi_s'(\tau,\nu)=\sqrt{1-\mathrm{Re}\{A_s^{(\alpha)}(\tau,\nu)\}}$ *(used in* $M^{(\phi_s')}_{\mathbf{H}}$*).*

Proof: We have

$$\begin{aligned}\Delta_6^{(\alpha)}(t,f)&=\langle \mathbf{H}x,s_{t,f}\rangle-L_{\mathbf{H}}^{(\alpha)}(t,f)\langle x,s_{t,f}\rangle\\&=\langle x,[\mathbf{H}^+-L_{\mathbf{H}}^{(\alpha)*}(t,f)\mathbf{I}]s_{t,f}\rangle\\&=\langle x,[\mathbf{H}^+-L_{\mathbf{H}^+}^{(\alpha)}(t,f)\mathbf{I}+L_{\mathbf{H}^+}^{(\alpha)}(t,f)\mathbf{I}-L_{\mathbf{H}}^{(\alpha)*}(t,f)\mathbf{I}]s_{t,f}\rangle\\&=\Delta_A^{(\alpha)}(t,f)+\Delta_B^{(\alpha)}(t,f),\end{aligned}\tag{30}$$

with

$$\Delta_A^{(\alpha)}(t,f)\triangleq\langle x,[\mathbf{H}^+-L_{\mathbf{H}^+}^{(\alpha)}(t,f)\mathbf{I}]s_{t,f}\rangle,\quad \Delta_B^{(\alpha)}(t,f)\triangleq[L_{\mathbf{H}^+}^{(\alpha)}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)]^*\langle x,s_{t,f}\rangle.$$

Using Schwarz' inequality, (24), (13), and $\|s_{t,f}\|_2=\|s\|_2=1$, we obtain the bounds

$$|\Delta_A^{(\alpha)}(t,f)|\leqslant\|x\|_2\|\mathbf{H}^+s_{t,f}-L_{\mathbf{H}^+}^{(\alpha)}(t,f)s_{t,f}\|_2\leqslant\|x\|_2\|S_{\mathbf{H}^+}\|_1D^{(\alpha)}_{\mathbf{H}^+,s}=\|x\|_2\|S_{\mathbf{H}}\|_1D^{(\alpha)}_{\mathbf{H}^+,s},\tag{31}$$

$$|\Delta_B^{(\alpha)}(t,f)|\leqslant|L_{\mathbf{H}^+}^{(\alpha)}(t,f)-L_{\mathbf{H}}^{(\alpha)*}(t,f)|\|x\|_2\|s_{t,f}\|_2\leqslant\|x\|_2\|S_{\mathbf{H}}\|_1 4\pi|\alpha|m_{\mathbf{H}}^{(1,1)}.\tag{32}$$

The first (L_∞) bound in (29) then follows upon inserting (31) and (32) into $|\Delta_6^{(\alpha)}(t,f)|\leqslant|\Delta_A^{(\alpha)}(t,f)|+|\Delta_B^{(\alpha)}(t,f)|$ [cf. (30)].

To derive the second (L_2) bound in (29), we note that

$$\begin{aligned}\|\Delta_A^{(\alpha)}\|_2^2&=\int_t\int_f|\langle x,\mathbf{H}^+s_{t,f}-L_{\mathbf{H}^+}^{(\alpha)}(t,f)s_{t,f}\rangle|^2dtdf\\&\leqslant\int_t\int_f\|x\|_2^2\|\mathbf{H}^+s_{t,f}-L_{\mathbf{H}^+}^{(\alpha)}(t,f)s_{t,f}\|_2^2dtdf\\&=\|x\|_2^2\int_t\int_f[\|\mathbf{H}^+s_{t,f}\|_2^2-2\,\mathrm{Re}\{L_{\mathbf{H}^+}^{(\alpha)*}(t,f)\langle\mathbf{H}^+s_{t,f},s_{t,f}\rangle\}+|L_{\mathbf{H}^+}^{(\alpha)}(t,f)|^2\|s_{t,f}\|_2^2]dtdf.\end{aligned}\tag{33}$$

Using (26), it can be shown that

$$\int_t\int_f\|\mathbf{H}^+s_{t,f}\|_2^2dtdf=\int_t\int_f\langle\mathbf{H}\mathbf{H}^+s_{t,f},s_{t,f}\rangle dtdf=S^{(\alpha)}_{\mathbf{H}\mathbf{H}^+}(0,0)=\mathrm{Tr}\{\mathbf{H}\mathbf{H}^+\}=\|S_{\mathbf{H}}\|_2^2$$

and

$$\int_t\int_f L_{\mathbf{H}^+}^{(\alpha)*}(t,f)\langle\mathbf{H}^+s_{t,f},s_{t,f}\rangle dtdf=\int_\tau\int_\nu|S_{\mathbf{H}^+}(\tau,\nu)|^2A_s^{(\alpha)*}(\tau,\nu)d\tau d\nu;$$

furthermore, with $\|s_{t,f}\|_2^2=\|s\|_2^2=1$ we have $\int_t\int_f|L_{\mathbf{H}^+}^{(\alpha)}(t,f)|^2\|s_{t,f}\|_2^2dt\,df=\|L_{\mathbf{H}}^{(\alpha)}\|_2^2=\|S_{\mathbf{H}}\|_2^2$. Inserting these three results into (33) yields

$$\begin{aligned}\|\Delta_A^{(\alpha)}\|_2^2&\leqslant\|x\|_2^2\left[2\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|^2d\tau d\nu-2\,\mathrm{Re}\left\{\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|^2A_s^{(\alpha)*}(\tau,\nu)d\tau d\nu\right\}\right]\\&=2\|x\|_2^2\int_\tau\int_\nu|S_{\mathbf{H}}(\tau,\nu)|^2[1-\mathrm{Re}\{A_s^{(\alpha)}(\tau,\nu)\}]d\tau d\nu=2\|x\|_2^2\|\mathbf{H}\|_2^2[M^{(\phi_s')}_{\mathbf{H}}]^2.\end{aligned}\tag{34}$$

The L_2 norm of $\Delta_B^{(\alpha)}(t,f)$ can be bounded using (13),

$$\|\Delta_B^{(\alpha)}\|_2^2 \leq \int_t \int_f |L_{\mathbf{H}^+}^{(\alpha)}(t,f) - L_{\mathbf{H}}^{(\alpha)*}(t,f)|^2 \|x\|_2^2 \|s_{t,f}\|_2^2 dt\,df \leq \|x\|_2^2 \|\mathbf{H}\|_2^2 [4\pi|\alpha| M_{\mathbf{H}}^{(1,1)}]^2. \quad (35)$$

The second (L_2) bound in (29) then follows by inserting (34) and (35) into $\|\Delta_6^{(\alpha)}\|_2 \leq \|\Delta_A^{(\alpha)}\|_2 + \|\Delta_B^{(\alpha)}\|_2$ [cf. (30)]. □

Thus, if $D_{\mathbf{H}^+,s}^{(\alpha)}$ (see Subsection III D) and $M_{\mathbf{H}}^{(\phi_s')}$ can be made small by a suitable choice of $s(t)$, and if $m_{\mathbf{H}}^{(1,1)}$ and $M_{\mathbf{H}}^{(1,1)}$ are small, we obtain the approximate input-output relation

$$\mathrm{STFT}_{\mathbf{H}x}^{(s)}(t,f) \approx L_{\mathbf{H}}^{(\alpha)}(t,f)\,\mathrm{STFT}_x^{(s)}(t,f).$$

In particular, small $M_{\mathbf{H}}^{(\phi_s')}$ requires that $\mathrm{Re}\{A_s^{(\alpha)}(\tau,\nu)\} \approx A_s^{(\alpha)}(0,0) \equiv 1$ on the effective support of $|S_{\mathbf{H}}(\tau,\nu)|$, thus implying that this effective support is small, i.e., that $\mathbf{H}$ is underspread. Note that the bounds in (29) simplify, and are tightest, in the case $\alpha = 0$.

F. Performance of the short-time Fourier transform filter

The relation $\mathrm{STFT}_{\mathbf{H}x}^{(s)}(t,f) \approx L_{\mathbf{H}}^{(\alpha)}(t,f)\mathrm{STFT}_x^{(s)}(t,f)$ valid in the underspread case suggests that the output signal $(\mathbf{H}x)(t)$ can approximately be obtained by applying an inverse short-time Fourier transform to $L_{\mathbf{H}}^{(\alpha)}(t,f)\mathrm{STFT}_x^{(s)}(t,f)$. This corresponds to an approximate time-frequency implementation of $\mathbf{H}$ that consists of calculating the short-time Fourier transform of the input signal, $\mathrm{STFT}_x^{(s)}(t,f)$, multiplying by the time-frequency transfer function $L_{\mathbf{H}}^{(\alpha)}(t,f)$, and applying the inverse short-time Fourier transform (27) to $L_{\mathbf{H}}^{(\alpha)}(t,f)\mathrm{STFT}_x^{(s)}(t,f)$. The resulting time-varying system will be called a *short-time Fourier transform filter* and denoted by $\hat{\mathbf{H}}^{(\alpha)}$. Using (27) and (28), the input-output relation of the short-time Fourier transform filter is

$$(\hat{\mathbf{H}}^{(\alpha)}x)(t) = \int_{t'} \int_{f'} \langle x, s_{t',f'}\rangle L_{\mathbf{H}}^{(\alpha)}(t',f') s_{t',f'}(t)\,dt'\,df'.$$

Such short-time Fourier transform filters [possibly with weighting functions other than $L_{\mathbf{H}}^{(\alpha)}(t,f)$] have been considered previously.[17,22–24,42,45,46,105,106] They are an intuitively appealing, practical way of implementing a time-varying filter. In a time-frequency-discretized form, they permit time-varying subband signal processing using the Gabor expansion[12,19,107,108] or, equivalently, DFT filter banks.[23,109–114] Furthermore, in a mathematical context short-time Fourier transform filters are a special case of Toeplitz operators.[42,98,115]

In the following, we first establish a relation and bound for the difference between the generalized Weyl symbol of the original operator $\mathbf{H}$ and that of the short-time Fourier transform filter $\hat{\mathbf{H}}^{(\alpha)}$; subsequently, we will bound the difference between the exact output signal $(\mathbf{H}x)(t)$ and the output signal $(\hat{\mathbf{H}}^{(\alpha)}x)(t)$ of the short-time Fourier transform filter. We note that the L_∞ bound in the next theorem is an extension and adaptation of an existing bound.[42,44]

Theorem III.8: *For any operator* $\mathbf{H}$, *the difference* $\Delta_7^{(\alpha)}(t,f) \triangleq L_{\mathbf{H}}^{(\alpha)}(t,f) - L_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(t,f)$ *satisfies*

$$\frac{|\Delta_7^{(\alpha)}(t,f)|}{\|S_{\mathbf{H}}\|_1} \leq m_{\mathbf{H}}^{(\phi_s)}, \quad \frac{\|\Delta_7^{(\alpha)}\|_2}{\|\mathbf{H}\|_2} = M_{\mathbf{H}}^{(\phi_s)},$$

with the weighting function $\phi_s(\tau,\nu) = |1 - A_s^{(\alpha)}(\tau,\nu)|$ *where* $s(t)$ *is the normalized window used in the short-time Fourier transform.*

Proof: The spreading function of $\hat{\mathbf{H}}^{(\alpha)}$ can be written as[15,42,46,64] $S_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(\tau,\nu) = S_{\mathbf{H}}^{(\alpha)}(\tau,\nu) \times A_s^{(\alpha)*}(\tau,\nu)$. The first ($L_\infty$) bound is then shown as

$$|\Delta_7^{(\alpha)}(t,f)| = \left| \int_\tau \int_\nu [S_{\mathbf{H}}^{(\alpha)}(\tau,\nu) - S_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(\tau,\nu)] e^{i2\pi(\nu t - \tau f)} d\tau\,d\nu \right|$$

$$\leq \int_\tau \int_\nu |S_{\mathbf{H}}(\tau,\nu)||1 - A_s^{(\alpha)*}(\tau,\nu)|\,d\tau\,d\nu = \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}.$$

The expression for the L_2 norm of $\Delta_7^{(\alpha)}(t,f)$ is shown by noting that

$$\|\Delta_7^{(\alpha)}\|_2^2=\|S_{\mathbf{H}}^{(\alpha)}-S_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}\|_2^2=\int_\tau\int_\nu |S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)[1-A_s^{(\alpha)*}(\tau,\nu)]|^2 d\tau d\nu,$$

and proceeding similarly. □

Hence, if $m_{\mathbf{H}}^{(\phi_s)}$ and $M_{\mathbf{H}}^{(\phi_s)}$ can be made small by a suitable choice of $s(t)$, then $L_{\mathbf{H}}^{(\alpha)}(t,f)\approx L_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(t,f)$. Furthermore, since $\|\Delta_7^{(\alpha)}\|_2$ equals the Hilbert–Schmidt norm $\|\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)}\|_2$ and since the operator norm[59] $\|\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)}\|$ is bounded from above by the Hilbert–Schmidt norm, we obtain

$$\|\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)}\|\leq\|\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)}\|_2=\|\mathbf{H}\|_2 M_{\mathbf{H}}^{(\phi_s)}. \tag{36}$$

Therefore, if $M_{\mathbf{H}}^{(\phi_s)}$ is small the short-time Fourier transform filter $\hat{\mathbf{H}}^{(\alpha)}$ will be close to $\mathbf{H}$, i.e., $\hat{\mathbf{H}}^{(\alpha)}\approx\mathbf{H}$. This is also shown by the following bounds on the difference of the respective output signals.

Corollary III.9: For any operator $\mathbf{H}$ *and any signal* $x(t)$, *the difference* $\Delta_8^{(\alpha)}(t)\triangleq(\mathbf{H}x)(t)-(\hat{\mathbf{H}}^{(\alpha)}x)(t)$ *is bounded as*

$$\frac{|\Delta_8^{(\alpha)}(t)|}{\|S_{\mathbf{H}}\|_1\|x\|_\infty}\leq m_{\mathbf{H}}^{(\phi_s)},\quad \frac{\|\Delta_8^{(\alpha)}\|_2}{\|\mathbf{H}\|_2\|x\|_2}\leq M_{\mathbf{H}}^{(\phi_s)},$$

with the weighting function $\phi_s(\tau,\nu)$ *as in Theorem III.8. [Here,* $\|x\|_\infty=\sup_t|x(t)|$ *denotes the* L_∞ (sup) *norm of* $x(t)$.*]*

Proof: Using (5), the relation $S_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(\tau,\nu)=S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)A_s^{(\alpha)*}(\tau,\nu)$, and $|x_{\tau,\nu}^{(\alpha)}(t)|\leq\|x_{\tau,\nu}^{(\alpha)}\|_\infty=\|x\|_\infty$, the first ($L_\infty$) bound is shown as

$$\begin{aligned}|\Delta_8^{(\alpha)}(t)|&=\left|\int_\tau\int_\nu [S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)-S_{\hat{\mathbf{H}}^{(\alpha)}}^{(\alpha)}(\tau,\nu)]x_{\tau,\nu}^{(\alpha)}(t)\,d\tau d\nu\right|\\&\leq\int_\tau\int_\nu |S_{\mathbf{H}}(\tau,\nu)||1-A_s^{(\alpha)*}(\tau,\nu)||x_{\tau,\nu}^{(\alpha)}(t)|\,d\tau d\nu\leq\|x\|_\infty\|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}.\end{aligned}$$

The second (L_2) bound is shown by noting that

$$\|\Delta_8^{(\alpha)}\|_2^2=\|(\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)})x\|_2^2\leq\|\mathbf{H}-\hat{\mathbf{H}}^{(\alpha)}\|^2\|x\|_2^2$$

and using (36). □

Thus, for an underspread operator $\mathbf{H}$ for which $m_{\mathbf{H}}^{(\phi_s)}$ and $M_{\mathbf{H}}^{(\phi_s)}$ can be made small by a suitable choice of the short-time Fourier transform window $s(t)$, we have $(\mathbf{H}x)(t)\approx(\hat{\mathbf{H}}^{(\alpha)}x)(t)$, i.e., the output of the short-time Fourier transform filter is a good approximation to the output of the given operator $\mathbf{H}$.

G. Input-output relation for random signals

When a wide-sense stationary random process $x(t)$ with power spectral density[116] $P_x(f)$ is passed through a time-invariant system $\mathbf{H}$ with frequency response $G(f)$, the output $(\mathbf{H}x)(t)$ is again wide-sense stationary with power spectral density $P_{\mathbf{H}x}(f)=|G(f)|^2P_x(f)$. Similarly, the response of a frequency-invariant system with temporal transfer function $m(t)$ to a (generally nonstationary) white process $x(t)$ with mean instantaneous intensity[116] $q_x(t)$ is again white, with mean instantaneous intensity $q_{\mathbf{H}x}(t)=|m(t)|^2q_x(t)$. We shall now investigate whether the simple multiplicative ''input-output relations'' $P_{\mathbf{H}x}(f)=|G(f)|^2P_x(f)$ and $q_{\mathbf{H}x}(t)=|m(t)|^2q_x(t)$, which relate the second-order statistics of $x(t)$ with those of $(\mathbf{H}x)(t)$, can be extended to the general time-varying/nonstationary case.

Let us consider an operator $\mathbf{H}$ whose input $x(t)$ is a general zero-mean, nonstationary random process. The correlation operator of $x(t)$ is defined as $\mathbf{R}_x\triangleq E\{x\otimes x^*\}$, where E denotes the expectation operator and $x\otimes x^*$ is the rank-one operator whose kernel is $x(t)x^*(t')$. Hence, the

kernel of $\mathbf{R}_x$ is the correlation function $r_x(t,t')=E\{x(t)x^*(t')\}$. The system output $(\mathbf{H}x)(t)$ is a zero-mean, nonstationary random process whose correlation operator is given by

$$\mathbf{R}_{\mathbf{H}x}=E\{(\mathbf{H}x)\otimes(\mathbf{H}x)^*\}=\mathbf{H}E\{x\otimes x^*\}\mathbf{H}^+=\mathbf{H}\mathbf{R}_x\mathbf{H}^+.$$

We now look for a time-frequency reformulation of the above input-output relation in terms of the generalized Weyl symbol. Specifically, if $\mathbf{H}$ and $\mathbf{R}_x$ are jointly underspread in the sense of Subsection III C, then (19) and (14) imply (similar results have been obtained previously[42,54])

$$L^{(\alpha)}_{\mathbf{R}_{\mathbf{H}x}}(t,f)=L^{(\alpha)}_{\mathbf{H}\mathbf{R}_x\mathbf{H}^+}(t,f)\approx L^{(\alpha)}_{\mathbf{H}}(t,f)L^{(\alpha)}_{\mathbf{R}_x}(t,f)L^{(\alpha)}_{\mathbf{H}^+}(t,f)\approx|L^{(\alpha)}_{\mathbf{H}}(t,f)|^2L^{(\alpha)}_{\mathbf{R}_x}(t,f),$$

or equivalently,

$$\overline{W}^{(\alpha)}_{\mathbf{H}x}(t,f)\approx|L^{(\alpha)}_{\mathbf{H}}(t,f)|^2\overline{W}^{(\alpha)}_{x}(t,f), \tag{37}$$

where $\overline{W}^{(\alpha)}_x(t,f)$ denotes the generalized Wigner–Ville spectrum[86,90,117] of a nonstationary process $x(t)$. The generalized Wigner–Ville spectrum is defined as the generalized Weyl symbol of the correlation operator,

$$\overline{W}^{(\alpha)}_x(t,f)\triangleq L^{(\alpha)}_{\mathbf{R}_x}(t,f)=\int_\tau r^{(\alpha)}_x(t,\tau)e^{-i2\pi f\tau}d\tau,$$

with $r^{(\alpha)}_x(t,\tau)$ obtained from $r_x(t,t')$ according to (3). It is a second-order time-frequency representation (''time-dependent power spectrum'') of nonstationary random processes. It reduces to the power spectral density $P_x(f)$ in the case of a wide-sense stationary process and to the mean instantaneous intensity $q_x(t)$ in the case of a (nonstationary) white process. The next theorem bounds the error of the approximation (37).

Theorem III.10: *For any operator* $\mathbf{H}$ *and any (generally nonstationary and nonwhite) random process* $\underline{x}(t)$ *with correlation operator* $\mathbf{R}_x$, *the difference* $\Delta^{(\alpha)}_9(t,f)\triangleq\overline{W}^{(\alpha)}_{\mathbf{H}x}(t,f)-|L^{(\alpha)}_{\mathbf{H}}(t,f)|^2\overline{W}^{(\alpha)}_x(t,f)$ *is bounded as*

$$\frac{|\Delta^{(\alpha)}_9(t,f)|}{\|S_{\mathbf{H}}\|_1^2\|S_{\mathbf{R}_x}\|_1}\leq 2\pi c_\alpha[m^{(0,1)}_{\mathbf{H}}m^{(1,0)}_{\mathbf{R}_x}+m^{(1,0)}_{\mathbf{H}}m^{(0,1)}_{\mathbf{R}_x}+m^{(0,1)}_{\mathbf{H}}m^{(1,0)}_{\mathbf{H}}]+4\pi|\alpha|m^{(1,1)}_{\mathbf{H}}, \tag{38}$$

with $c_\alpha=|\alpha+1/2|+|\alpha-1/2|$.

Proof: Subtracting and adding $L^{(\alpha)}_{\mathbf{H}}(t,f)L^{(\alpha)}_{\mathbf{R}_x}(t,f)L^{(\alpha)}_{\mathbf{H}^+}(t,f)$, we have

$$|\Delta^{(\alpha)}_9(t,f)|=|\Delta^{(\alpha)}_A(t,f)+\Delta^{(\alpha)}_B(t,f)|\leq|\Delta^{(\alpha)}_A(t,f)|+|\Delta^{(\alpha)}_B(t,f)|, \tag{39}$$

with

$$\Delta^{(\alpha)}_A(t,f)\triangleq L^{(\alpha)}_{\mathbf{H}\mathbf{R}_x\mathbf{H}^+}(t,f)-L^{(\alpha)}_{\mathbf{H}}(t,f)L^{(\alpha)}_{\mathbf{R}_x}(t,f)L^{(\alpha)}_{\mathbf{H}^+}(t,f), \tag{40}$$

$$\Delta^{(\alpha)}_B(t,f)\triangleq L^{(\alpha)}_{\mathbf{H}}(t,f)L^{(\alpha)}_{\mathbf{R}_x}(t,f)[L^{(\alpha)}_{\mathbf{H}^+}(t,f)-L^{(\alpha)}_{\mathbf{H}}{}^*(t,f)]. \tag{41}$$

Since according to (18) $S^{(\alpha)}_{\mathbf{H}\mathbf{R}_x\mathbf{H}^+}(\tau,\nu)=(S^{(\alpha)}_{\mathbf{H}}\natural S^{(\alpha)}_{\mathbf{R}_x}\natural S^{(\alpha)}_{\mathbf{H}^+})(\tau,\nu)$ and since the 2-D Fourier transform of $L^{(\alpha)}_{\mathbf{H}}(t,f)L^{(\alpha)}_{\mathbf{R}_x}(t,f)L^{(\alpha)}_{\mathbf{H}^+}(t,f)$ is given by $(S^{(\alpha)}_{\mathbf{H}}{**}S^{(\alpha)}_{\mathbf{R}_x}{**}S^{(\alpha)}_{\mathbf{H}^+})(\tau,\nu)$, the 2-D Fourier transform of $\Delta^{(\alpha)}_A(t,f)$ in (40) is obtained as

$$\begin{aligned}\hat{\Delta}^{(\alpha)}_A(\tau,\nu)&=(S^{(\alpha)}_{\mathbf{H}}\natural S^{(\alpha)}_{\mathbf{R}_x}\natural S^{(\alpha)}_{\mathbf{H}^+})(\tau,\nu)-(S^{(\alpha)}_{\mathbf{H}}{**}S^{(\alpha)}_{\mathbf{R}_x}{**}S^{(\alpha)}_{\mathbf{H}^+})(\tau,\nu)\\&=\int_{\tau_1}\int_{\nu_1}\int_{\tau_2}\int_{\nu_2}S^{(\alpha)}_{\mathbf{H}}(\tau_1,\nu_1)S^{(\alpha)}_{\mathbf{R}_x}(\tau_2,\nu_2)S^{(\alpha)}_{\mathbf{H}^+}(\tau-\tau_1-\tau_2,\nu-\nu_1-\nu_2)\\&\quad\cdot[e^{-i2\pi[\phi_\alpha(\tau-\tau_1,\nu-\nu_1,\tau_2,\nu_2)+\phi_\alpha(\tau,\nu,\tau_1,\nu_1)]}-1]d\tau_1d\nu_1d\tau_2d\nu_2.\end{aligned}$$

Using $|\sin x| \leq |x|$ and substituting $\tau_3 = \tau - \tau_1 - \tau_2$, $\nu_3 = \nu - \nu_1 - \nu_2$, the first term in (39) can then be bounded as

$$\begin{aligned}
|\Delta_A^{(\alpha)}(t,f)| &\leq \int_\tau \int_\nu |\hat{\Delta}_A^{(\alpha)}(\tau,\nu)| d\tau d\nu \\
&\leq 2\pi \int_{\tau_1} \int_{\nu_1} \int_{\tau_2} \int_{\nu_2} \int_{\tau_3} \int_{\nu_3} |S_{\mathbf{H}}(\tau_1,\nu_1)||S_{\mathbf{R}_x}(\tau_2,\nu_2)||S_{\mathbf{H}^+}(\tau_3,\nu_3)| \\
&\quad \cdot [|\alpha + \tfrac{1}{2}|[|\tau_1\nu_2| + |\tau_1\nu_3| + |\tau_2\nu_3|] + |\alpha - \tfrac{1}{2}|[|\tau_2\nu_1| + |\tau_3\nu_1| + |\tau_3\nu_2|]] \\
&\quad \cdot d\tau_1 d\nu_1 d\tau_2 d\nu_2 d\tau_3 d\nu_3 \\
&= 2\pi \|S_{\mathbf{H}}\|_1^2 \|S_{\mathbf{R}_x}\|_1 c_\alpha [m_{\mathbf{H}}^{(0,1)} m_{\mathbf{R}_x}^{(1,0)} + m_{\mathbf{H}}^{(1,0)} m_{\mathbf{R}_x}^{(0,1)} + m_{\mathbf{H}}^{(0,1)} m_{\mathbf{H}}^{(1,0)}], \qquad (42)
\end{aligned}$$

where the final expression is obtained by collecting corresponding terms in the integral and using $m_{\mathbf{H}^+}^{(k,l)} = m_{\mathbf{H}}^{(k,l)}$. A bound on the second term in (39) is obtained by using $|L_{\mathbf{H}}^{(\alpha)}(t,f)| \leq \|S_{\mathbf{H}}\|_1$ in (41) and applying (13),

$$|\Delta_B^{(\alpha)}(t,f)| \leq \|S_{\mathbf{H}}\|_1 \|S_{\mathbf{R}_x}\|_1 |L_{\mathbf{H}^+}^{(\alpha)}(t,f) - L_{\mathbf{H}}^{(\alpha)*}(t,f)| \leq \|S_{\mathbf{H}}\|_1^2 \|S_{\mathbf{R}_x}\|_1 4\pi |\alpha| m_{\mathbf{H}}^{(1,1)}. \qquad (43)$$

Inserting (42) and (43) in (39) yields the bound (38). □

The theorem shows that the approximate input-output relation (37) will be valid if $\mathbf{H}$ and $\mathbf{R}_x$ are jointly underspread such that the moments appearing in the bound (38) are small. This requires the spreading functions of $\mathbf{H}$ and $\mathbf{R}_x$ to be similarly concentrated along the τ or ν axis. The bound (38) is tightest for $\alpha = 0$, in which case also a refined bound similar to (20) and (23) can be obtained by using metaplectic transformations of $\mathbf{H}$ and $\mathbf{R}_x$; this allows the spreading functions of $\mathbf{H}$ and $\mathbf{R}_x$ to be oriented in (similar) oblique directions.

H. Infimum and supremum of the Weyl symbol

We have seen in Subsection III D that the generalized Weyl symbol of an underspread operator can be interpreted as an approximate eigenvalue distribution over the time-frequency plane. We now assume $\mathbf{H}$ to be self-adjoint, such that the eigenvalues $\lambda_{\mathbf{H},k}$ of $\mathbf{H}$ are real-valued. Furthermore, we consider the Weyl symbol $L_{\mathbf{H}}^{(0)}(t,f)$ of $\mathbf{H}$ which is real-valued as well. We shall investigate how close $\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f)$ and $\sup_{t,f} L_{\mathbf{H}}^{(0)}(t,f)$ are to $\lambda_{\mathbf{H}}^{\inf} \triangleq \inf_k \lambda_{\mathbf{H},k}$ and $\lambda_{\mathbf{H}}^{\sup} \triangleq \sup_k \lambda_{\mathbf{H},k}$, respectively. This issue will be seen to be of importance in Subsections III I and IV B. Related issues in the theories of quantization and pseudo-differential operators are the boundedness or positivity of operators corresponding to bounded or positive symbols, the results being theorems of the Calderón-Vaillancourt type[15,40,41,118] and Gårding inequalities,[14,15,40,97] respectively. The next theorem is an extension and adaptation of previous results.[42,44]

Theorem III.11: *For any self-adjoint operator* $\mathbf{H}$*, there is*

$$\frac{|\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f) - \lambda_{\mathbf{H}}^{\inf}|}{\|S_{\mathbf{H}}\|_1} \leq m_{\mathbf{H}}^{(\phi_s)}, \quad \frac{|\sup_{t,f} L_{\mathbf{H}}^{(0)}(t,f) - \lambda_{\mathbf{H}}^{\sup}|}{\|S_{\mathbf{H}}\|_1} \leq m_{\mathbf{H}}^{(\phi_s)}, \qquad (44)$$

with the weighting function $\phi_s(\tau,\nu) = |1 - 1/A_s^{(0)}(\tau,\nu)|$ *where* $s(t)$ *is an arbitrary normalized function.*

Proof: Let us define[15,42,44] the "lower symbol" $L_{\mathbf{H}}^{L}(t,f) \triangleq \int_\tau \int_\nu S_{\mathbf{H}}^{(0)}(\tau,\nu) \times A_s^{(0)*}(\tau,\nu) e^{i2\pi(t\nu - f\tau)} d\tau d\nu$ and the "upper symbol" $L_{\mathbf{H}}^{U}(t,f) \triangleq \int_\tau \int_\nu [S_{\mathbf{H}}^{(0)}(\tau,\nu)/A_s^{(0)}(\tau,\nu)] \times e^{i2\pi(t\nu - f\tau)} d\tau d\nu$, where $s(t)$ is an arbitrary normalized function. [For Gaussian $s(t)$, the lower and upper symbols can be related to the Wick and anti-Wick symbols of quantum mechanics[15,119] which can be defined using the normal and anti-normal correspondence rules.[34,36,83]] It is known[15,42] that for self-adjoint operators $\mathbf{H}$,

$$\inf_{t,f} L_{\mathbf{H}}^{U}(t,f) \leq \lambda_{\mathbf{H}}^{\inf} \leq \inf_{t,f} L_{\mathbf{H}}^{L}(t,f), \qquad (45)$$

$$\sup_{t,f} L_{\mathbf{H}}^{L}(t,f\,) \leq \lambda_{\mathbf{H}}^{\sup} \leq \sup_{t,f} L_{\mathbf{H}}^{U}(t,f\,). \tag{46}$$

We now have

$$\begin{aligned} |L_{\mathbf{H}}^{(0)}(t,f\,) - L_{\mathbf{H}}^{L}(t,f\,)| &= \left| \int_{\tau}\int_{\nu} S_{\mathbf{H}}^{(0)}(\tau,\nu)[1-A_s^{(0)*}(\tau,\nu)] e^{i2\pi(\tau f-\nu t)} d\tau d\nu \right| \\ &\leq \int_{\tau}\int_{\nu} |S_{\mathbf{H}}(\tau,\nu)||1-A_s^{(0)}(\tau,\nu)| d\tau d\nu = \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\tilde{\phi}_s)}, \end{aligned} \tag{47}$$

with $\tilde{\phi}_s(\tau,\nu) = |1-A_s^{(0)}(\tau,\nu)|$. Similarly, we can show

$$|L_{\mathbf{H}}^{(0)}(t,f\,) - L_{\mathbf{H}}^{U}(t,f\,)| \leq \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}, \tag{48}$$

with $\phi_s(\tau,\nu) = |1-1/A_s^{(0)}(\tau,\nu)|$. Furthermore, due to $|A_s^{(0)}(\tau,\nu)| \leq \|s\|_2^2 = 1$ there is

$$\begin{aligned} m_{\mathbf{H}}^{(\phi_s)} &= \frac{1}{\|S_{\mathbf{H}}\|_1} \int_{\tau}\int_{\nu} \frac{|1-A_s^{(0)}(\tau,\nu)|}{|A_s^{(0)}(\tau,\nu)|} |S_{\mathbf{H}}(\tau,\nu)| d\tau d\nu \\ &\geqslant \frac{1}{\|S_{\mathbf{H}}\|_1} \int_{\tau}\int_{\nu} |1-A_s^{(0)}(\tau,\nu)||S_{\mathbf{H}}(\tau,\nu)| d\tau d\nu = m_{\mathbf{H}}^{(\tilde{\phi}_s)}. \end{aligned} \tag{49}$$

It can be shown that (47) implies $|\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) - \inf_{t,f} L_{\mathbf{H}}^{L}(t,f\,)| \leq \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\tilde{\phi}_s)}$ and that (48) implies $|\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) - \inf_{t,f} L_{\mathbf{H}}^{U}(t,f\,)| \leq \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}$. Combining these two inequalities with (45) and (49), it can be shown that $|\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) - \lambda_{\mathbf{H}}^{\inf}| \leq \|S_{\mathbf{H}}\|_1 m_{\mathbf{H}}^{(\phi_s)}$. The bound on $|\sup_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) - \lambda_{\mathbf{H}}^{\sup}|$ follows similarly using (47), (48), (46), and (49). □

Hence, if $m_{\mathbf{H}}^{(\phi_s)}$ can be made small by a suitable choice of $s(t)$, we have

$$\inf_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) \approx \lambda_{\mathbf{H}}^{\inf}, \quad \sup_{t,f} L_{\mathbf{H}}^{(0)}(t,f\,) \approx \lambda_{\mathbf{H}}^{\sup}.$$

Small $m_{\mathbf{H}}^{(\phi_s)}$ requires $A_s(\tau,\nu) \approx A_s(0,0) \equiv 1$ on the effective support of $|S_{\mathbf{H}}(\tau,\nu)|$, thus implying that this effective support is small, i.e., that $\mathbf{H}$ is underspread. We note that $m_{\mathbf{H}}^{(\phi_s)}$ can be adapted to oblique orientation of $|S_{\mathbf{H}}(\tau,\nu)|$ by a proper choice of the normalized function $s(t)$.

I. Operator norm (maximum system gain)

We now return to a general (i.e., not necessarily self-adjoint) operator $\mathbf{H}$. The operator norm (maximum system gain) of $\mathbf{H}$ is defined as

$$\|\mathbf{H}\| \triangleq \sup_{\|x\|_2 = 1} \|\mathbf{H}x\|_2.$$

For time-invariant and frequency-invariant operators, $\|\mathbf{H}\|$ equals the supremum of the magnitude of the transfer function. We now ask whether for a general operator $\|\mathbf{H}\|$ is similarly related to the magnitude of $L_{\mathbf{H}}^{(\alpha)}(t,f\,)$.

Theorem III.12: *For any operator* $\mathbf{H}$, *the difference between the supremum of* $|L_{\mathbf{H}}^{(\alpha)}(t,f\,)|^2$ *and the squared operator norm is bounded as*

$$\frac{|\sup_{t,f} |L_{\mathbf{H}}^{(\alpha)}(t,f\,)|^2 - \|\mathbf{H}\|^2|}{\|S_{\mathbf{H}}\|_1^2} \leq 2\pi m_{\mathbf{H}}^{(0,1)} m_{\mathbf{H}}^{(1,0)} + 4\pi|\alpha| m_{\mathbf{H}}^{(1,1)} + m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}, \tag{50}$$

with the weighting function $\phi_s(\tau,\nu)$ *as in Theorem* III.11.

Proof: Subtracting and adding $\sup_{t,f} |L_{\mathbf{H}}^{(0)}(t,f\,)|^2$ and $\sup_{t,f} L_{\mathbf{H}^+\mathbf{H}}^{(0)}(t,f\,)$ from/to $\sup_{t,f} |L_{\mathbf{H}}^{(\alpha)}(t,f\,)|^2 - \|\mathbf{H}\|^2$, and using[59] $\|\mathbf{H}\|^2 = \lambda_{\mathbf{H}^+\mathbf{H}}^{\sup}$, we obtain

$$\left|\sup_{t,f}|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2-\|\mathbf{H}\|^2\right| \leq \left|\sup_{t,f}|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2-\sup_{t,f}|L_{\mathbf{H}}^{(0)}(t,f)|^2\right| + \left|\sup_{t,f}|L_{\mathbf{H}}^{(0)}(t,f)|^2-\sup_{t,f}L_{\mathbf{H}^+\mathbf{H}}^{(0)}(t,f)\right| + \left|\sup_{t,f}L_{\mathbf{H}^+\mathbf{H}}^{(0)}(t,f)-\lambda_{\mathbf{H}^+\mathbf{H}}^{\sup}\right|. \tag{51}$$

We next derive a bound on the difference $\Delta_{10}^{(\alpha)}(t,f)\triangleq|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2-|L_{\mathbf{H}}^{(0)}(t,f)|^2$ by noting that the 2-D Fourier transform of $|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2$ is given by $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)**S_{\mathbf{H}}^{(\alpha)*}(-\tau,-\nu)$ and by using $S_{\mathbf{H}}^{(\alpha)}(\tau,\nu)=S_{\mathbf{H}}^{(0)}(\tau,\nu)e^{-i2\pi\alpha\tau\nu}$,

$$\begin{aligned}
|\Delta_{10}^{(\alpha)}(t,f)| &\leq \int_\tau\int_\nu|\hat{\Delta}_{10}^{(\alpha)}(\tau,\nu)|\,d\tau\,d\nu\\
&= \int_\tau\int_\nu\left|\int_{\tau'}\int_{\nu'}S_{\mathbf{H}}^{(0)}(\tau',\nu')S_{\mathbf{H}}^{(0)*}(\tau'-\tau,\nu'-\nu)\right.\\
&\quad\left.\times[e^{-i2\pi\alpha[\tau'\nu'-(\tau'-\tau)(\nu'-\nu)]}-1]\,d\tau'\,d\nu'\right|d\tau\,d\nu\\
&\leq 2\int_{\tau_1}\int_{\nu_1}\int_{\tau'}\int_{\nu'}|S_{\mathbf{H}}(\tau',\nu')||S_{\mathbf{H}}(\tau_1,\nu_1)||\sin\,\pi\alpha(\tau'\nu'-\tau_1\nu_1)|\,d\tau'\,d\nu'\,d\tau_1\,d\nu_1\\
&\leq 2\pi|\alpha|\int_{\tau_1}\int_{\nu_1}\int_{\tau'}\int_{\nu'}|S_{\mathbf{H}}(\tau',\nu')||S_{\mathbf{H}}(\tau_1,\nu_1)|[|\tau'\nu'|+|\tau_1\nu_1|]\,d\tau'\,d\nu'\,d\tau_1\,d\nu_1\\
&= 4\pi|\alpha|\|S_{\mathbf{H}}\|_1^2m_{\mathbf{H}}^{(1,1)}.
\end{aligned}$$

It can be shown that this bound also implies that the first term in (51) is bounded as $|\sup_{t,f}|L_{\mathbf{H}}^{(\alpha)}(t,f)|^2-\sup_{t,f}|L_{\mathbf{H}}^{(0)}(t,f)|^2|\leq 4\pi|\alpha|\|S_{\mathbf{H}}\|_1^2m_{\mathbf{H}}^{(1,1)}$. Furthermore, it follows from (22) with $\alpha=0$ that the second term in (51) is bounded as $|\sup_{t,f}|L_{\mathbf{H}}^{(0)}(t,f)|^2-\sup_{t,f}L_{\mathbf{H}^+\mathbf{H}}^{(0)}(t,f)|\leq 2\pi\|S_{\mathbf{H}}\|_1^2m_{\mathbf{H}}^{(0,1)}m_{\mathbf{H}}^{(1,0)}$. Finally, since $\mathbf{H}^+\mathbf{H}$ is self-adjoint, the bounds in (44) hold for $\mathbf{H}^+\mathbf{H}$ and we obtain that the third term in (51) is bounded as $|\sup_{t,f}L_{\mathbf{H}^+\mathbf{H}}^{(0)}(t,f)-\lambda_{\mathbf{H}^+\mathbf{H}}^{\sup}|\leq\|S_{\mathbf{H}^+\mathbf{H}}\|_1m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}\leq\|S_{\mathbf{H}}\|_1^2m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$. Inserting these three bounds into (51) yields (50). □

Hence, if $m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$ can be made small by a suitable choice of the function $s(t)$, and if $m_{\mathbf{H}}^{(1,1)}$ and $m_{\mathbf{H}}^{(0,1)}m_{\mathbf{H}}^{(1,0)}$ are small too, then

$$\sup_{t,f}|L_{\mathbf{H}}^{(\alpha)}(t,f)|\approx\|\mathbf{H}\|.$$

In particular, small $m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$ requires that $A_s(\tau,\nu)\approx A_s(0,0)\equiv 1$ on the effective support of $|S_{\mathbf{H}^+\mathbf{H}}(\tau,\nu)|$, thus implying that $\mathbf{H}^+\mathbf{H}$ is underspread, which will be true if $\mathbf{H}$ is underspread.

The bound in (50) is tightest for $\alpha=0$ since here the bound's second term vanishes. Moreover, using (23) we obtain the tighter bound,

$$\frac{|\sup_{t,f}|L_{\mathbf{H}}^{(0)}(t,f)|^2-\|\mathbf{H}\|^2|}{\|S_{\mathbf{H}}\|_1^2}\leq 2\pi\min_{\mathbf{U}\in\mathcal{M}}\{m_{\mathbf{UHU}^+}^{(0,1)}+m_{\mathbf{UHU}^+}^{(1,0)}\}+m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}. \tag{52}$$

The first term in this refined bound may be small even if the spreading function is oriented in an oblique direction in the (τ,ν) plane. The second term, $m_{\mathbf{H}^+\mathbf{H}}^{(\phi_s)}$, can be adapted to oblique directions by a proper choice of $s(t)$.

IV. PROPERTIES OF UNDERSPREAD OPERATORS

In this section, we show that besides permitting a simple approximate transfer function (symbolic) calculus, underspread operators have other ''desirable'' properties that may simplify their analysis.

A. Approximate commutation

In contrast to the time-invariant or frequency-invariant case, two general linear operators $\mathbf{G}$ and $\mathbf{H}$ typically do not commute, i.e., $\mathbf{GH}\neq\mathbf{HG}$, or equivalently $[\mathbf{G},\mathbf{H}]\triangleq\mathbf{GH}-\mathbf{HG}\neq\mathbf{0}$. The following theorem shows that this situation is somewhat different for jointly underspread operators.

Theorem IV.1: *The operator norm of the commutator of two operators* $\mathbf{G}$ *and* $\mathbf{H}$ *is bounded as*

$$\frac{\|[\mathbf{G},\mathbf{H}]\|^2}{\|S_{\mathbf{G}}\|_1^2\|S_{\mathbf{H}}\|_1^2}\leq 16\pi^2\left[\min_{\mathbf{U}\in\mathcal{M}} B^{(0)}_{\mathbf{UGU}^+,\mathbf{UHU}^+}\right]^2+8\pi\min_{\mathbf{U}\in\mathcal{M}}\{[m^{(0,1)}_{\mathbf{UGU}^+}+m^{(0,1)}_{\mathbf{UHU}^+}][m^{(1,0)}_{\mathbf{UGU}^+}+m^{(1,0)}_{\mathbf{UHU}^+}]\}+4m^{(\phi_s)}_{[\mathbf{G},\mathbf{H}]^+[\mathbf{G},\mathbf{H}]}, \tag{53}$$

with $B^{(0)}_{\mathbf{G},\mathbf{H}}=(1/2)[m^{(0,1)}_{\mathbf{G}}m^{(1,0)}_{\mathbf{H}}+m^{(1,0)}_{\mathbf{G}}m^{(0,1)}_{\mathbf{H}}]$ *and with the weighting function* $\phi_s(\tau,\nu)$ *as in Theorem* III.11.

Proof: Applying (52) to $[\mathbf{G},\mathbf{H}]$ yields

$$\|[\mathbf{G},\mathbf{H}]\|^2\leq\sup_{t,f}|L^{(0)}_{[\mathbf{G},\mathbf{H}]}(t,f)|^2+\|S_{[\mathbf{G},\mathbf{H}]}\|_1^2\left[2\pi\min_{\mathbf{U}\in\mathcal{M}}\{m^{(0,1)}_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}m^{(1,0)}_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}\}+m^{(\phi_s)}_{[\mathbf{G},\mathbf{H}]^+[\mathbf{G},\mathbf{H}]}\right]. \tag{54}$$

In order to bound $\sup_{t,f}|L^{(0)}_{[\mathbf{G},\mathbf{H}]}(t,f)|^2$, we consider $L^{(0)}_{[\mathbf{G},\mathbf{H}]}(t,f)$. Subtracting and adding $L^{(0)}_{\mathbf{G}}(t,f)L^{(0)}_{\mathbf{H}}(t,f)$ and applying (20) twice, we obtain

$$\begin{aligned}|L^{(0)}_{[\mathbf{G},\mathbf{H}]}(t,f)|&=|L^{(0)}_{\mathbf{GH}}(t,f)-L^{(0)}_{\mathbf{HG}}(t,f)|\\&\leq|L^{(0)}_{\mathbf{GH}}(t,f)-L^{(0)}_{\mathbf{G}}(t,f)L^{(0)}_{\mathbf{H}}(t,f)|+|L^{(0)}_{\mathbf{G}}(t,f)L^{(0)}_{\mathbf{H}}(t,f)-L^{(0)}_{\mathbf{HG}}(t,f)|\\&\leq 2\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1 2\pi\min_{\mathbf{U}\in\mathcal{M}} B^{(0)}_{\mathbf{UGU}^+,\mathbf{UHU}^+},\end{aligned}$$

which gives the bound

$$\sup_{t,f}|L^{(0)}_{[\mathbf{G},\mathbf{H}]}(t,f)|^2\leq\|S_{\mathbf{G}}\|_1^2\|S_{\mathbf{H}}\|_1^2 16\pi^2\left[\min_{\mathbf{U}\in\mathcal{M}} B^{(0)}_{\mathbf{UGU}^+,\mathbf{UHU}^+}\right]^2. \tag{55}$$

With regard to the second term in (54), we have

$$\begin{aligned}\|S_{[\mathbf{G},\mathbf{H}]}\|_1 m^{(0,1)}_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}&=\|S_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}\|_1 m^{(0,1)}_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}\\&\leq\int_\tau\int_\nu|\nu|[|S_{\mathbf{UGU}^+\mathbf{UHU}^+}(\tau,\nu)|+|S_{\mathbf{UHU}^+\mathbf{UGU}^+}(\tau,\nu)|]d\tau d\nu\\&\leq\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1 2[m^{(0,1)}_{\mathbf{UGU}^+}+m^{(0,1)}_{\mathbf{UHU}^+}],\end{aligned}$$

where the last step can be shown using the twisted convolution (18). In a similar manner, one can show that $\|S_{[\mathbf{G},\mathbf{H}]}\|_1 m^{(1,0)}_{\mathbf{U}[\mathbf{G},\mathbf{H}]\mathbf{U}^+}\leq\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1 2[m^{(1,0)}_{\mathbf{UGU}^+}+m^{(1,0)}_{\mathbf{UHU}^+}]$. Inserting these bounds into (54) and using $\|S_{[\mathbf{G},\mathbf{H}]}\|_1\leq\|S_{\mathbf{GH}}\|_1+\|S_{\mathbf{HG}}\|_1\leq 2\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1$, the bound (53) follows. □

Hence, the operator norm of the commutator of two jointly underspread operators is small, which shows that two jointly underspread operators approximately commute, i.e., $\mathbf{GH}\approx\mathbf{HG}$. Since two operators commute if and only if they have a common set of eigenfunctions,[15,59] this result is consistent with Theorem III.6. While here we have considered the operator norm $\|[\mathbf{G},\mathbf{H}]\|$, we note that bounds on the Hilbert–Schmidt norm $\|[\mathbf{G},\mathbf{H}]\|_2$ have been formulated previously.[42]

Almost commuting operators may look peculiar to quantum physicists since in the classical limit ($h\to 0$) of the quantum-mechanical symbolic calculus the Weyl symbol of the commutator $[\mathbf{G},\mathbf{H}]$ reduces to the (generally nonzero) Poisson bracket of the symbols which is defined as[15]

$$\{L_{\mathbf{H}}^{(\alpha)},L_{\mathbf{G}}^{(\alpha)}\}\triangleq\frac{\partial L_{\mathbf{H}}^{(\alpha)}(t,f)}{\partial t}\frac{\partial L_{\mathbf{G}}^{(\alpha)}(t,f)}{\partial f}-\frac{\partial L_{\mathbf{G}}^{(\alpha)}(t,f)}{\partial t}\frac{\partial L_{\mathbf{H}}^{(\alpha)}(t,f)}{\partial f}.$$

Specifically, if the Weyl symbol is re-defined[15,34,36,56] to include Planck's constant h, then one can show[15,56] that $L_{[\mathbf{G},\mathbf{H}]}^{(0)}=-i(h/2\pi)\{L_{\mathbf{G}}^{(0)},L_{\mathbf{H}}^{(0)}\}+\mathcal{O}(h^2)$, i.e., the Weyl symbol of $[\mathbf{G},\mathbf{H}]$ is proportional to $\{L_{\mathbf{G}}^{(0)},L_{\mathbf{H}}^{(0)}\}$ up to errors of second order. However, the following bound shows that the Poisson bracket of two jointly underspread operators is close to zero.

Theorem IV.2: *The Poisson bracket of the generalized Weyl symbols of two operators* **G** *and* **H** *is bounded as*

$$\frac{|\{L_{\mathbf{G}}^{(\alpha)},L_{\mathbf{H}}^{(\alpha)}\}|}{\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1}\leq 8\pi^2\min_{\mathbf{U}\in\mathcal{M}}B_{\mathbf{UGU}^+,\mathbf{UHU}^+}^{(0)},\tag{56}$$

with $B_{\mathbf{G},\mathbf{H}}^{(0)}=(1/2)[m_{\mathbf{G}}^{(0,1)}m_{\mathbf{H}}^{(1,0)}+m_{\mathbf{G}}^{(1,0)}m_{\mathbf{H}}^{(0,1)}]$.

Proof: Since differentiations and multiplications in the time-frequency domain correspond to multiplications and convolutions in the (τ,ν) domain, it is easily shown that the 2-D Fourier transform of the Poisson bracket is given by

$$\begin{aligned}\int_t\int_f\{L_{\mathbf{G}}^{(\alpha)},L_{\mathbf{H}}^{(\alpha)}\}e^{-i2\pi(\nu t-\tau f)}dtdf=\int_{\tau'}\int_{\nu'}[&i2\pi\nu' S_{\mathbf{G}}^{(\alpha)}(\tau',\nu')i2\pi(\tau-\tau')S_{\mathbf{H}}^{(\alpha)}(\tau-\tau',\nu-\nu')\\&-i2\pi\tau' S_{\mathbf{G}}^{(\alpha)}(\tau',\nu')i2\pi(\nu-\nu')S_{\mathbf{H}}^{(\alpha)}(\tau-\tau',\nu-\nu')]d\tau' d\nu'.\end{aligned}$$

Furthermore, the magnitude of the Poisson bracket is bounded by the L_1 norm of its Fourier transform, and therefore

$$\begin{aligned}|\{L_{\mathbf{G}}^{(\alpha)},L_{\mathbf{H}}^{(\alpha)}\}|&\leq\int_\tau\int_\nu\left|4\pi^2\int_{\tau'}\int_{\nu'}[\nu'(\tau-\tau')-\tau'(\nu-\nu')]S_{\mathbf{G}}^{(\alpha)}(\tau',\nu')\right.\\&\qquad\left.\times S_{\mathbf{H}}^{(\alpha)}(\tau-\tau',\nu-\nu')d\tau' d\nu'\right|d\tau d\nu\\&\leq 4\pi^2\int_{\tau_1}\int_{\nu_1}\int_{\tau_2}\int_{\nu_2}|\nu_1\tau_2-\tau_1\nu_2||S_{\mathbf{G}}(\tau_1,\nu_1)||S_{\mathbf{H}}(\tau_2,\nu_2)|d\tau_1 d\nu_1 d\tau_2 d\nu_2\\&\leq 4\pi^2\|S_{\mathbf{G}}\|_1\|S_{\mathbf{H}}\|_1[m_{\mathbf{G}}^{(0,1)}m_{\mathbf{H}}^{(1,0)}+m_{\mathbf{G}}^{(1,0)}m_{\mathbf{H}}^{(0,1)}].\end{aligned}\tag{57}$$

Since $\nu_1\tau_2-\tau_1\nu_2$ in (57) is invariant to symplectic coordinate transforms, the above bound remains valid if **G** and **H** are replaced by $\mathbf{UGU}^+$ and $\mathbf{UHU}^+$, respectively, with $\mathbf{U}\in\mathcal{M}$. □

Hence, if **G** and **H** are jointly underspread such that $\min_{\mathbf{U}\in\mathcal{M}}B_{\mathbf{UGU}^+,\mathbf{UHU}^+}^{(0)}$ is small, then the Poisson bracket $\{L_{\mathbf{G}}^{(\alpha)},L_{\mathbf{H}}^{(\alpha)}\}$ is approximately zero, thereby confirming our result that the commutator of jointly underspread operators vanishes approximately. We note that the bound in (56) does not depend on α, and that it equals the bound in (20) up to a factor.

B. Approximate normality

Normal operators have the advantage of allowing an eigenvalue decomposition instead of a (numerically more expensive) singular value decomposition. In contrast to time-invariant or frequency-invariant systems, general operators may be non-normal, i.e., $\mathbf{HH}^+\neq\mathbf{H}^+\mathbf{H}$ or equivalently $[\mathbf{H},\mathbf{H}^+]=\mathbf{HH}^+-\mathbf{H}^+\mathbf{H}\neq\mathbf{0}$. A bound on $\|[\mathbf{H},\mathbf{H}^+]\|$ could be obtained as a special case of (53). However, the following theorem exploits the self-adjointness of $[\mathbf{H},\mathbf{H}^+]$ to yield a tighter and simpler bound.

Theorem IV.3: *The operator norm of the commutator of an operator* $\mathbf{H}$ *and its adjoint* $\mathbf{H}^+$ *is bounded as*

$$\frac{\|[\mathbf{H},\mathbf{H}^+]\|}{\|S_{\mathbf{H}}\|_1^2} \leq 4\pi \min_{\mathbf{U}\in\mathcal{M}} \{m^{(0,1)}_{\mathbf{UHU}^+} m^{(1,0)}_{\mathbf{UHU}^+}\} + m^{(\phi_s)}_{\mathbf{HH}^+} + m^{(\phi_s)}_{\mathbf{H}^+\mathbf{H}}, \tag{58}$$

with the weighting function $\phi_s(\tau,\nu)$ *as in Theorem* III.11.

Proof: It is known[59] that $\|[\mathbf{H},\mathbf{H}^+]\| = \max\{-\lambda^{\inf}_{[\mathbf{H},\mathbf{H}^+]}, \lambda^{\sup}_{[\mathbf{H},\mathbf{H}^+]}\} = \max\{\lambda^{\sup}_{-[\mathbf{H},\mathbf{H}^+]}, \lambda^{\sup}_{[\mathbf{H},\mathbf{H}^+]}\}$. First assume that $\|[\mathbf{H},\mathbf{H}^+]\| = \lambda^{\sup}_{[\mathbf{H},\mathbf{H}^+]}$. In that case, the second bound in (44) yields

$$\begin{aligned}\|[\mathbf{H},\mathbf{H}^+]\| = \lambda^{\sup}_{[\mathbf{H},\mathbf{H}^+]} &\leq \sup_{t,f} L^{(0)}_{[\mathbf{H},\mathbf{H}^+]}(t,f) + \|S_{[\mathbf{H},\mathbf{H}^+]}\|_1 m^{(\phi_s)}_{[\mathbf{H},\mathbf{H}^+]} \\ &\leq \sup_{t,f} |L^{(0)}_{[\mathbf{H},\mathbf{H}^+]}(t,f)| + \|S_{[\mathbf{H},\mathbf{H}^+]}\|_1 m^{(\phi_s)}_{[\mathbf{H},\mathbf{H}^+]}.\end{aligned} \tag{59}$$

Specializing (55), the first term in (59) can be bounded as

$$\sup_{t,f} |L^{(0)}_{[\mathbf{H},\mathbf{H}^+]}(t,f)| \leq \|S_{\mathbf{H}}\|_1^2 4\pi \min_{\mathbf{U}\in\mathcal{M}} m^{(0,1)}_{\mathbf{UHU}^+} m^{(1,0)}_{\mathbf{UHU}^+}, \tag{60}$$

where we used that $B^{(0)}_{\mathbf{H},\mathbf{H}^+} = m^{(0,1)}_{\mathbf{H}} m^{(1,0)}_{\mathbf{H}}$ due to $m^{(k,l)}_{\mathbf{H}^+} = m^{(k,l)}_{\mathbf{H}}$. The second term in (59) can be bounded as

$$\begin{aligned}\|S_{[\mathbf{H},\mathbf{H}^+]}\|_1 m^{(\phi_s)}_{[\mathbf{H},\mathbf{H}^+]} &\leq \int_\tau \int_\nu \phi_s(\tau,\nu)[|S_{\mathbf{HH}^+}(\tau,\nu)| + |S_{\mathbf{H}^+\mathbf{H}}(\tau,\nu)|]\,d\tau\,d\nu \\ &= \|S_{\mathbf{HH}^+}\|_1 m^{(\phi_s)}_{\mathbf{HH}^+} + \|S_{\mathbf{H}^+\mathbf{H}}\|_1 m^{(\phi_s)}_{\mathbf{H}^+\mathbf{H}} \leq \|S_{\mathbf{H}}\|_1^2 [m^{(\phi_s)}_{\mathbf{HH}^+} + m^{(\phi_s)}_{\mathbf{H}^+\mathbf{H}}].\end{aligned} \tag{61}$$

Inserting (60) and (61) into (59), the bound (58) follows. If $\|[\mathbf{H},\mathbf{H}^+]\| = -\lambda^{\inf}_{[\mathbf{H},\mathbf{H}^+]} = \lambda^{\sup}_{-[\mathbf{H},\mathbf{H}^+]}$, the bound is shown similarly by applying the second bound in (44) to $-[\mathbf{H},\mathbf{H}^+]$. □

Hence, an underspread operator $\mathbf{H}$ for which $\min_{\mathbf{U}\in\mathcal{M}}\{m^{(0,1)}_{\mathbf{UHU}^+} m^{(1,0)}_{\mathbf{UHU}^+}\}$ as well as $m^{(\phi_s)}_{\mathbf{HH}^+}$ and $m^{(\phi_s)}_{\mathbf{H}^+\mathbf{H}}$ are small satisfies $\mathbf{HH}^+ \approx \mathbf{H}^+\mathbf{H}$ and is thus approximately normal. We note that bounds on the Hilbert–Schmidt norm $\|[\mathbf{H},\mathbf{H}^+]\|_2$ have been formulated previously.[42]

V. CONCLUSION

We have introduced a new concept of ''underspread'' linear operators (linear time-varying systems) that does not require the support of the spreading function to be compact and thus extends a previous definition of underspread operators. While sufficiently general to be relevant to a broad variety of engineering applications, our underspread concept still is restrictive enough to permit the formulation of a simple and intuitive approximate transfer function calculus (symbolic calculus) that is based on the generalized Weyl symbol. Indeed, in the case of underspread operators the generalized Weyl symbol can be viewed as an approximate ''time-frequency transfer function'' that is similarly simple to use as the conventional transfer function of linear time-invariant systems. The quality and scope of validity of this transfer function approximation was assessed by providing explicit upper bounds on various approximation errors; these bounds are formulated in terms of weighted integrals and moments of the spreading function.

The transfer function calculus provides a foundation and justification of several methods which have been proposed to design and implement time-varying filters for signal enhancement, estimation, and detection.[45–53] Furthermore, the calculus can immediately be applied to the theory of time-varying power spectra of nonstationary random processes. Here, the operators are either correlation operators or innovation systems (cf. Subsection III G). This application leads to the important concept of *underspread random processes* for which several important definitions of time-varying power spectra can be shown to become approximately equivalent.[42,54,55]

ACKNOWLEDGMENTS

We are grateful to the anonymous referees for several useful comments and for pointing out relevant references. This work was supported by FWF Grants No. P10012-ÖPY and No. 11904-TEC.

[1] E. P. Wigner, Phys. Rev. **40**, 749 (1932).
[2] J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[3] E. P. Wigner, in *Perspectives in Quantum Theory*, edited by W. Yourgrau and A. van der Merwe (Dover, New York, 1971), Vol. 40, Chap. 4, pp. 25–36.
[4] N. G. de Bruijn, Nieuw Archief voor Wiskunde (3) **XXI**, 205 (1973).
[5] J. Ville, Cables Transm. **2eA**., No. 1, 61 (1948).
[6] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, Philips J. Res. **35**, 217 (1980).
[7] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, Philips J. Res. **35**, 372 (1980).
[8] *The Wigner Distribution—Theory and Applications in Signal Processing*, edited by W. Mecklenbräuker and F. Hlawatsch (Elsevier, Amsterdam, 1997).
[9] P. Flandrin, *Temps-Fréquence* (Hermès, Paris, 1993).
[10] A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1965), Vol. I.
[11] G. B. Folland and A. Sitaram, J. Fourier Anal. Appl. **3**, 207 (1997).
[12] D. Gabor, J. IEE **93**, 429 (1946).
[13] N. G. de Bruijn, in *Inequalities*, edited by O. Shisha (Academic, New York, 1967), pp. 57–71.
[14] C. L. Fefferman, Bull. Am. Math. Soc. **9**, 129 (1983).
[15] G. B. Folland, *Harmonic Analysis in Phase Space*, Vol. 122 of *Annals of Mathematics Studies* (Princeton University Press, Princeton, NJ, 1989).
[16] A. Perelomov, *Generalized Coherent States and Their Applications* (Springer-Verlag, Berlin, 1986).
[17] I. Daubechies, IEEE Trans. Inf. Theory **34**, 605 (1988).
[18] S. T. Ali, J.-P. Antoine, J.-P. Gazeau, and U. A. Müller, Rev. Math. Phys. **7/7**, 1013 (1995).
[19] *Gabor Analysis and Algorithms: Theory and Applications*, edited by H. G. Feichtinger and T. Strohmer (Birkhäuser, Boston, 1998).
[20] R. Tolimieri and M. An, *Time-Frequency Representations* (Birkhäuser, Boston, 1998).
[21] F. Hlawatsch and G. F. Boudreaux-Bartels, IEEE Signal Process. Mag. **9**, 21 (1992).
[22] S. H. Nawab and T. F. Quatieri, in *Advanced Topics in Signal Processing*, edited by J. S. Lim and A. V. Oppenheim (Prentice–Hall, Englewood Cliffs, NJ, 1988), Chap. 6, pp. 289–337.
[23] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* (Prentice–Hall, Englewood Cliffs, NJ, 1983).
[24] M. R. Portnoff, IEEE Trans. Acoust., Speech, Signal Process. **28**, 55 (1980).
[25] J. Zak, Phys. Rev. Lett. **19**, 1385 (1967).
[26] A. J. E. M. Janssen, Philips J. Res. **43**, 23 (1988).
[27] H. Bölcskei and F. Hlawatsch, IEEE Trans. Signal Process. **45**, 851 (1997).
[28] M. G. Bellanger, G. Bonnerot, and M. Coudreuse, IEEE Trans. Acoust., Speech, Signal Process. **24**, 109 (1976).
[29] P. P. Vaidyanathan, *Multirate Systems and Filter Banks* (Prentice–Hall, Englewood Cliffs, NJ, 1993).
[30] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding* (Prentice–Hall, Englewood Cliffs, NJ, 1995).
[31] N. L. Balazs and B. K. Jennings, Phys. Rep. **104**, 347 (1984).
[32] H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1950).
[33] J. C. T. Pool, J. Math. Phys. **7**, 66 (1966).
[34] L. Cohen, J. Math. Phys. **7**, 781 (1966).
[35] L. Cohen, J. Math. Phys. **17**, 1863 (1976).
[36] C. L. Mehta, J. Math. Phys. **5**, 677 (1964).
[37] J. J. Kohn and L. Nirenberg, Commun. Pure Appl. Math. **18**, 269 (1965).
[38] A. Grossmann, G. Loupias, and E. M. Stein, Ann. Inst. Fourier **18**, 343 (1968).
[39] A. Voros, Ann. Inst. Henri Poincaré **26**, 343 (1977).
[40] L. Hörmander, Commun. Pure Appl. Math. **32**, 359 (1979).
[41] R. Howe, J. Funct. Anal. **38**, 188 (1980).
[42] W. Kozek, Ph.D. thesis, Vienna University of Technology, 1997.
[43] W. Kozek, IEEE Trans. Signal Process. **45**, 219 (1997).
[44] W. Kozek, in *Gabor Analysis and Algorithms: Theory and Applications*, edited by H. G. Feichtinger and T. Strohmer (Birkhäuser, Boston, 1998), Chap. 10, pp. 323–352.
[45] R. Bourdier, J. F. Allard, and K. Trumpf, Signal Process. **15**, 193 (1988).
[46] W. Kozek and F. Hlawatsch, in *Proceedings of the IEEE–SP International Symposium on Time-Frequency and Time-Scale Analysis* (Victoria, B.C., Canada, 1992), pp. 163–166.
[47] A. M. Sayeed, P. Lander, and D. L. Jones, J. Electrocardiol. **28**, 53 (1995).
[48] G. W. Wornell, IEEE Trans. Inf. Theory **41**, 1418 (1995).
[49] G. W. Wornell, IEEE Trans. Inf. Theory **42**, 488 (1996).
[50] A. A. Beex and M. Xie, *Proceedings of the IEEE ICASSP-95*, Detroit, MI, 1995, pp. 1565–1568.
[51] H. Kirchauer, F. Hlawatsch, and W. Kozek, in Ref. 50, pp. 1549–1552.
[52] G. Matz and F. Hlawatsch, in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis* (Paris, France, 1996), pp. 213–216.
[53] J. A. Sills, Ph.D. thesis, Georgia Institute of Technology, Atlanta, 1995.
[54] G. Matz, F. Hlawatsch, and W. Kozek, IEEE Trans. Signal Process. **45**, 1520 (1997).
[55] W. Kozek, F. Hlawatsch, H. Kirchauer, and U. Trautwein, *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, 1994, pp. 417–420.
[56] A. Voros, J. Funct. Anal. **29**, 104 (1978).

[57] A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1965), Vol. II.
[58] T. Arai, J. Math. Phys. **36**, 622 (1995).
[59] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, 2nd ed. (Springer-Verlag, New York, 1982).
[60] I. M. Gelfand and G. E. Schilow, *Verallgemeinerte Funktionen (Distributionen)* (VEB, Berlin, 1960).
[61] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators* (American Mathematical Society, Providence, RI, 1969).
[62] L. A. Zadeh, Proc. IRE **76**, 291 (1950).
[63] P. A. Bello, IEEE Trans. Commun. Syst. **11**, 360 (1963).
[64] R. G. Shenoy and T. W. Parks, IEEE Trans. Signal Process. **42**, 318 (1994).
[65] A. J. E. M. Janssen, Philips J. Res. **44**, 7 (1989).
[66] H. Weyl, Z. Phys. **46**, 1 (1928).
[67] W. Kozek, Signal Process. **29**, 77 (1992).
[68] W. Kozek, in Ref. 46, pp. 167–170.
[69] P. Flandrin, B. Escudié, and J. Gréa, C. R. Acad. Sci. Paris, Ser. I **294**, 281 (1982).
[70] A. J. E. M. Janssen, Philips J. Res. **37**, 79 (1982).
[71] F. Hlawatsch and P. Flandrin, in *The Wigner Distribution—Theory and Applications in Signal Processing*, edited by W. Mecklenbräuker and F. Hlawatsch (Elsevier, Amsterdam, 1997), pp. 59–133.
[72] A. W. Rihaczek, IEEE Trans. Inf. Theory **14**, 369 (1968).
[73] R. Simon, E. C. G. Sudarshan, and N. Mukunda, Phys. Rev. A **36**, 3868 (1987).
[74] A. Weil, Acta Math. **111**, 143 (1964).
[75] W. Miller, Jr., in *Radar and Sonar, Part I*, edited by R. E. Blahut, W. Miller, Jr., and C. H. Wilcox (Springer-Verlag, New York, 1991), pp. 66–168.
[76] T. Kailath, IEEE Trans. Inf. Theory **9**, 233 (1963).
[77] T.-Y. Leou and J. K. Aggarwal, Proc. IEEE **72**, 980 (1984).
[78] J. G. Proakis, *Digital Communications*, 3rd ed. (McGraw-Hill, New York, 1995).
[79] J. A. Sills and E. W. Kamen, Circuits Syst. Signal Process. **15**, 609 (1996).
[80] S. G. Mallat, G. Papanicolaou, and Z. Zhang, Ann. Stat. **26**, 1 (1998).
[81] K. A. Sostrand, Proc. NATO Adv. Study Inst. Signal Process. Emphasis Underwater Acoust. **2**, 25.1 (1968).
[82] R. Howe, Bull. Am. Math. Soc. **3**, 821 (1980).
[83] K. E. Cahill and R. J. Glauber, Phys. Rev. **177**, 1882 (1969).
[84] R. S. Kennedy, *Fading Dispersive Communication Channels* (Wiley, New York, 1969).
[85] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part III: Radar-Sonar Signal Processing and Gaussian Signals in Noise* (Krieger, Malabar, 1992).
[86] P. Flandrin, in *Time and Frequency Representation of Signals and Systems*, edited by G. Longo and B. Picinbono (Springer, Wien, 1989), pp. 69–124.
[87] G. Matz, Ph.D. thesis, Vienna University of Technology, in preparation.
[88] R. Estrada, J. M. Gracia-Bondía, and J. C. Várilly, J. Math. Phys. **30**, 2789 (1989).
[89] F. Bayen, M. Flato, C. Fronsdal, *et al.*, Ann. Phys. **111**, 61 (1978).
[90] W. Martin and P. Flandrin, IEEE Trans. Acoust., Speech, Signal Process. **33**, 1461 (1985).
[91] M. B. Priestley, J. R. Stat. Soc. Ser. B **27**, 204 (1965).
[92] F. Hlawatsch, *Proceedings of the IEEE ISCAS-92*, San Diego, CA, 1992, pp. 1459–1462.
[93] M. Sandell, Ph.D. thesis, Lulea University of Technology, Lulea, Sweden, 1996.
[94] R. Haas and J. C. Belfiore, Wireless Personal Commun. **5**, 1 (1997).
[95] W. Kozek and A. F. Molisch, *Proceedings of the IEEE Workshop on Signal Processing Applications in Wireless Communication* (Paris, France, 1997), pp. 325–328.
[96] W. Kozek and A. F. Molisch, IEEE J. Sel. Areas Comm. (submitted).
[97] A. Córdoba and C. Fefferman, Comm. Partial Diff. Eqns. **3**, 979 (1978).
[98] R. Rochberg, Lecture Notes Pure Appl. Math. **51**, 425 (1990).
[99] Y. Meyer, in AMS *Proceedings of the Symposia in Applied Mathematics: Different Perspectives on Wavelets*, edited by I. Daubechies (AMS, Providence, RI, 1993), pp. 35–58.
[100] F. Hlawatsch, IEEE Trans. Signal Process. **39**, 1564 (1991).
[101] P. M. Woodward, *Probability and Information Theory with Application to Radar* (Pergamon, London, 1953).
[102] A. Papoulis, *Signal Analysis* (McGraw-Hill, Singapore, 1984).
[103] C. H. Wilcox, in *Radar and Sonar, Part I*, edited by R. E. Blahut, W. Miller, Jr., and C. H. Wilcox (Springer-Verlag, New York, 1991), pp. 229–260.
[104] R. J. Glauber, Phys. Rev. **131**, 2766 (1963).
[105] A. Dembo and D. Malah, IEEE Trans. Acoust., Speech, Signal Process. **36**, 168 (1988).
[106] W. Kozek, H. G. Feichtinger, and J. Scharinger, *Proceedings of the IEEE ISCAS-96*, Atlanta, GA, 1996, pp. 509–512.
[107] H. G. Feichtinger and K. Gröchenig, in *Wavelets—A Tutorial in Theory and Applications*, edited by C. Chui (Academic, Boston, 1992), pp. 359–397.
[108] S. Farkash and S. Raz, IEEE Trans. Signal Process. **42**, 611 (1994).
[109] K. Swaminathan and P. P. Vaidyanathan, IEEE Trans. Circuits Syst. **33**, 1170 (1986).
[110] M. Vetterli, IEEE Trans. Acoust., Speech, Signal Process. **35**, 356 (1987).
[111] M. Poize, M. Renaudin, and P. Venier, *Quatorzième Colloque GRETSI* (Juan-Les-Pins, France, 1993), pp. 351–354.
[112] Y. Lin and P. P. Vaidyanathan, *Proceedings of the APCCAS-94* (Taipei, Taiwan, 1994), pp. 254–259.
[113] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, in *Proceedings of the IEEE ICASSP-96*, Atlanta, GA, 1996, Vol. 3, pp. 1391–1394.
[114] H. Bölcskei and F. Hlawatsch, in *Gabor Analysis and Algorithms: Theory and Applications*, edited by H. G. Feichtinger and T. Strohmer (Birkhäuser, Boston, 1998), Chap. 9, pp. 295–322.
[115] K. Nowak, Colloq. Math. **LXV**, Fasc. 1, 149 (1993).

[116] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. (McGraw-Hill, New York, 1991).
[117] P. Flandrin and W. Martin, in *The Wigner Distribution—Theory and Applications in Signal Processing*, edited by W. Mecklenbräuker and F. Hlawatsch (Elsevier, Amsterdam, 1997), pp. 211–267.
[118] A. Calderón and R. Vaillancourt, J. Math. Soc. Jpn. **23**, 374 (1971).
[119] F. A. Berezin, Math. USSR Sb. **15**, 577 (1971).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 8 AUGUST 1998

Symbolic calculus on the time-frequency half-plane

J. Bertrand
LPTM, University Paris VII, 75251 Paris Cedex 05, France

P. Bertrand
ONERA, Chemin de la Hunière, F-91761, Palaiseau Cedex, France

(Received 5 August 1997; accepted for publication 9 February 1998)

This study concerns a special symbolic calculus of interest for signal analysis. This calculus associates functions on the time-frequency half-plane $f>0$ with linear operators defined on the positive-frequency signals. Full attention is given to its construction which is entirely based on the study of the affine group in a simple and direct way. The correspondence rule is detailed and the associated Wigner function is given. Formulas expressing the basic operation (star-bracket) of the Lie algebra of symbols, which is isomorphic to that of the operators, are obtained. In addition, it is shown that the resulting calculus is covariant under a three-parameter group which contains the affine group as subgroup. This observation is the starting point of an investigation leading to a whole class of symbolic calculi which can be considered as modifications of the original one. © *1998 American Institute of Physics.* [S0022-2488(98)00408-3]

I. INTRODUCTION

The notion of time-frequency analysis of signals received a theoretical basis when Gabor[1] proposed to interpret signals by operations very similar to those of quantum mechanics. In this approach, the real signals are replaced by their positive-frequency parts (the ''analytic'' signals) and the communication characteristics are introduced as the mean values of Hermitian operators defined on the corresponding Hilbert space. Actually, the analogy with quantum mechanics has rapidly suggested we recognize the time and the frequency as two conjugate variables and has led to the search for a phase space formulation of the theory. Various solutions have then been proposed which were essentially adaptations of solutions developed precedently in the quantum mechanical context.

From a mathematical point of view, the question of the phase space formulation of a quantum theory appears as a problem of symbolic calculus since each operator of the theory has to be replaced by a function (its symbol) of the phase space variables.[2] The construction of such a correspondence rule is determined by the class of operators to symbolize and by subsidiary constraints relative to the physical interpretation of the formalism. In signal theory, this last point concerns simply the formal invariance of the operations in changes of reference clocks. The corresponding constraint will be automatically ensured if the correspondence rule is covariant with respect to relevant representations of the affine group. The object of the following developments is to construct a symbolic calculus on the time-frequency half-plane along these lines.

In Sec. II we describe a study of the affine group and its subgroups, their representations in the Hilbert space $\mathcal{H}$ of signals and their actions in the time-frequency half-plane Γ. Decompositions of $\mathcal{H}$ and Γ into invariant subspaces are carried out in parallel and are shown to be related in a natural way. The correspondence rule is derived in Sec. III by exploiting the completeness of the invariant decompositions just performed. Some properties of this rule are given in Sec. IV and the associated Wigner function is determined. Section V is devoted to the derivation of the star product, which allows to symbolize products of operators and is the core of the symbolic calculus. In Sec. VI, the star bracket is introduced and shown to reduce to an ordinary Poisson bracket in some special cases. As an application of this remark, a Hamiltonian flow is defined in phase space. Finally, it is shown in Sec. VII that the method has possible extensions when working with larger groups containing the affine group.

 © 1998 American Institute of Physics

II. THE AFFINE GROUP IN SIGNAL THEORY

The signals to be considered are real-valued functions of time $s(t)$ which are interpreted independently of the choice of time origin and units. This means that the description of the signal processing operations must not change if time t is replaced by $at+b$ where b is real and a positive. The corresponding change on the signal can be readily written down. Notice that there is the freedom of rescaling the signal when the units are changed. The transformations on the signal that will be considered are written as

$$s(t)\rightarrow a^r s(a^{-1}(t-b)), \tag{II.1}$$

where r is a real exponent.

The set of transformations (II.1) forms the affine group, with parameters (a,b) and composition law expressed by

$$(a,b)(a',b')=(aa',b+ab'). \tag{II.2}$$

The group structure is the mathematical expression of the equivalence of the reference clocks. To stress the fact that only positive dilations are considered, the parameter u such that

$$a=e^u \tag{II.3}$$

will be used instead of a.

A. Representation in the Hilbert space of positive-frequency signals

The Fourier transform of the signal $s(t)$ will be introduced by

$$\hat{s}(f)\equiv\int_{\mathbb{R}} s(t)e^{-2i\pi ft}\,dt. \tag{II.4}$$

Since $s(t)$ is a real-valued function, it is characterized by its positive-frequency part only. In agreement with the usual practice in signal theory, this property will be used to substitute to $s(t)$ the so-called analytic signal whose Fourier transform is given by

$$S(f)=2Y(f)\hat{s}(f), \tag{II.5}$$

where $Y(f)$ is Heaviside's step function.

The counterpart of transformation (II.1) for the analytic signal $S(f)$ is easily obtained. Namely, by action of an element (u,b) of the affine group, signal $S(f)$ goes to a new function $[\mathbf{U}(u,b)S](f)$ defined by

$$\mathbf{U}(u,b)S(f)\equiv e^{(r+1)u}e^{-2i\pi fb}S(e^u f),\quad r\in\mathbb{R}. \tag{II.6}$$

The operators $\mathbf{U}(u,b)$ constitute an irreducible representation of the affine group that is unitary for the scalar product defined by

$$(S,S')\equiv\int_{\mathbb{R}^+} S(f)S'^*(f)f^{2r+1}df. \tag{II.7}$$

The Hilbert space of functions $S(f)$ on the half-line which are square integrable for the measure $f^{2r+1}\,df$ will be denoted by $\mathcal{H}$ (the reference to r being implicit). From a physical point of view, it may be necessary to distinguish between signals transforming according to (II.6) with different values of r. However, the corresponding representations are easily shown to be unitarily equivalent. In fact, it may be shown that the affine group has only two inequivalent unitary representations differing by the sign of the imaginary exponential in (II.6).[3]

The infinitesimal operators of representation (II.6) are introduced by the relations

$$\boldsymbol{\beta}S(f)\equiv-\frac{1}{2i\pi}\frac{d}{du}\mathbf{U}(u,b)S(f)\bigg|_{u=b=0}=-\frac{1}{2i\pi}\left(r+1+f\frac{d}{df}\right)S(f), \tag{II.8}$$

$$\mathbf{f}S(f)\equiv-\frac{1}{2i\pi}\frac{d}{db}\mathbf{U}(u,b)S(f)\Big|_{u=b=0}=fS(f). \tag{II.9}$$

The operators $\boldsymbol{\beta}$ and $\mathbf{f}$ thus obtained verify the commutation relation:

$$[\boldsymbol{\beta},\mathbf{f}]=-\frac{1}{2i\pi}\mathbf{f}. \tag{II.10}$$

The subgroups of the affine group and their representations will now be studied in detail. There are two kinds of subgroups consisting either of pure dilations or of pure translations. The translation subgroup is invariant by conjugation. On the contrary, since a dilation can be performed from an arbitrary time origin ξ, there is a whole family of conjugate dilation subgroups G_ξ which consist of the following set of elements:

$$G_\xi=\{(e^u,\xi(1-e^u))\}. \tag{II.11}$$

The restriction $\mathbf{U}_\xi$ of representation (II.6) to G_ξ is given by

$$\mathbf{U}_\xi(u)S(f)=e^{(r+1)u}e^{-2i\pi f\xi(1-e^u)}S(e^u f), \tag{II.12}$$

and the corresponding generator is

$$\boldsymbol{\beta}_\xi S(f)\equiv-\frac{1}{2i\pi}\frac{d}{du}\mathbf{U}_\xi(u)S(f)\Big|_{u=0}=(\boldsymbol{\beta}-\xi\mathbf{f})S(f), \tag{II.13}$$

where operators $\boldsymbol{\beta}$ and $\boldsymbol{f}$ are given by (II.8) and (II.9), respectively.

The operator of representation (II.12) can be written in terms of generator $\boldsymbol{\beta}_\xi$ as

$$\mathbf{U}_\xi=e^{-2i\pi u\boldsymbol{\beta}_\xi}. \tag{II.14}$$

For a given ξ, the eigenfunctions of the self-adjoint operator $\boldsymbol{\beta}_\xi$ are defined by the equation

$$\boldsymbol{\beta}_\xi\psi_\beta^\xi(f)=\beta\psi_\beta^\xi(f),\quad \beta\in\mathbb{R}, \tag{II.15}$$

whose solution is, up to a multiplicative constant,

$$\psi_\beta^\xi(f)\equiv f^{-2i\pi\beta-r-1}e^{-2i\pi\xi f}. \tag{II.16}$$

By construction, the functions $\psi_\beta^\xi(f)$ are also eigenfunctions of the unitary operator (II.14) and verify

$$\mathbf{U}_\xi(u)\psi_\beta^\xi(f)=e^{-2i\pi\beta u}\psi_\beta^\xi(f). \tag{II.17}$$

For a given ξ, functions (II.16) constitute an improper basis of the Hilbert space $\mathcal{H}$. In fact, one verifies that there is orthogonality of the functions corresponding to different β,

$$(\psi_\beta^\xi(f),\psi_{\beta'}^\xi(f))=\delta(\beta-\beta'), \tag{II.18}$$

where the scalar product has the form (II.7). The completeness of this basis is expressed by

$$\int_{\mathbb{R}}\psi_\beta^\xi(f)\psi_\beta^{\xi*}(f')d\beta=f^{-2r-1}\delta(f-f'). \tag{II.19}$$

For each ξ, the set of all the projectors on the functions $\psi_\beta^\xi(f)$ gives a decomposition of the Hilbert space $\mathcal{H}$. From (II.17) it results that each of these projectors is invariant by the transformations of the subgroup G_ξ.

More generally, the action of the whole affine group on functions $\psi_\beta^\xi(f)$ is given by

$$\mathbf{U}(u,b)\psi_{\beta}^{\xi}(f)=e^{-2i\pi u\beta}\psi_{\beta}^{\xi'}(f), \tag{II.20}$$

with $\xi'=e^{u}\xi+b$. This shows that the set of projectors associated with subgroup G_{ξ} is transformed as a whole into the set of projectors associated with subgroup $G_{e^{u}\xi+b}$. In other words, this means that the set of all decompositions of $\mathcal{H}$ by the various subgroups G_{ξ} is invariant by action of the whole group.

B. Action of the group in the time-frequency half-plane

The physical phase space for signal analysis is the time-frequency half-plane, $\Gamma=\{(t,f)\}$, t real, $f>0$. This space can also be obtained as an orbit of the coadjoint representation of the affine group whose action in variables (t,f) is given by

$$(u,b):(t,f)\rightarrow(e^{u}t+b,e^{-u}f), \tag{II.21}$$

where the pair (u,b) characterizes an element of the group.

This action induces a transformation on functions $\mathcal{A}(t,f)$ defined on Γ. This transformation has the form

$$(u,b):\mathcal{A}(t,f)\rightarrow\mathcal{A}(e^{-u}(t-b),e^{u}f). \tag{II.22}$$

According to definition (II.11), the restriction of representation (II.21) to subgroup G_{ξ} reads as

$$(t,f)\rightarrow(e^{u}t+\xi(1-e^{u}),e^{-u}f), \tag{II.23}$$

and the corresponding transformation of phase space functions is given by

$$\mathcal{A}(t,f)\rightarrow\mathcal{A}(e^{-u}(t-\xi)+\xi,e^{u}f). \tag{II.24}$$

The transformation (II.23) has the following invariant:

$$(t-\xi)f=\tilde{\beta},\quad \tilde{\beta}\in\mathbb{R}, \tag{II.25}$$

where $\tilde{\beta}$ is an arbitrary constant. For each ξ, this relation defines a family of curves in phase space labeled by $\tilde{\beta}$. Actually any point in Γ belongs to one, and only one, of these curves and a partition of phase space has thus been achieved for any value of ξ.

Uniform distributions on curves (II.25) are given by the Dirac distributions:

$$\delta_{\tilde{\beta}}^{\xi}\equiv\delta((t-\xi)f-\tilde{\beta}). \tag{II.26}$$

These distributions are invariant by action of the subgroup G_{ξ} and transform under the full group as follows:

$$(u,b):\delta_{\tilde{\beta}}^{\xi}=\delta((t-\xi)f-\tilde{\beta})\rightarrow\delta((e^{-u}(t-b)-\xi)e^{u}f-\tilde{\beta})=\delta_{\tilde{\beta}}^{\xi e^{u}+b}. \tag{II.27}$$

This result can be compared with (II.20). In fact, the full group connects the various phase space partitions associated with subgroups G_{ξ}.

C. Correspondence between invariant structures

The subgroup G_{ξ}, considered for any given value of ξ, has allowed us to perform decompositions of both $\mathcal{H}$ and Γ where the resulting partitions have been labeled by parameters denoted, respectively, by β and $\tilde{\beta}$. We will now show that a natural connection does exist between these two parameters. To develop this point, it is useful to introduce the notion of localized signals in a consistent way.

A signal $S_{t_0}(f)\in\mathcal{H}$ is said to be localized at time t_0 provided an affine transformation, defined by the element (u,b) of the affine group, sends it into a signal localized at $(e^{u}t_0+b)$. This is expressed by the following equation:

$$\mathbf{U}(u,b)S_{t_0}(f)\equiv e^{(r+1)u}e^{-2i\pi fb}S_{t_0}(e^u f)=S_{e^u t_0+b}(f), \tag{II.28}$$

whose solution is, up to a multiplicative constant,

$$S_{t_0}(f)=f^{-r-1}e^{-2i\pi ft_0}. \tag{II.29}$$

Notice the factor f^{-r-1} which is required by dilation invariance. In this way, the whole set of localized states S_{t_0} is stable by action of the affine group.

In the decomposition of $\mathcal{H}$ relative to the subgroup G_ξ, the β-subspace is generated by $\psi_\beta^\xi(f)$ of the form (II.16). Developing the phase of this function in the vicinity of a frequency f_0 leads to the approximate expression

$$\psi_\beta^\xi(f)\simeq e^{-2i\pi\beta(\ln f_0-1)}f^{-r-1}e^{-2i\pi f(\beta/f_0+\xi)}. \tag{II.30}$$

A comparison with (II.29) shows that, apart from a constant phase, the approximated function can be seen as a state localized at

$$t_0=\beta/f_0+\xi. \tag{II.31}$$

Thus, in a local study of functions (II.16), it is possible to associate with any frequency f a time t given by $t=\beta/f+\xi$. This relation defines precisely an orbit of subgroup G_ξ in phase space Γ [cf. (II.25)]. It leads naturally to identify the parameters of the connected elements by setting

$$\tilde{\beta}=\beta. \tag{II.32}$$

This identification, which is clearly invariant under the group action, will play a central role in the following construction.

III. GEOMETRIC CORRESPONDENCE BETWEEN SYMBOLS AND OPERATORS

The aim now is to set up a correspondence between operators on the Hilbert space $\mathcal{H}$ and functions (their symbols) on the time-frequency half-plane Γ. The approach will be essentially based on the parallel decompositions of $\mathcal{H}$ and Γ introduced in the preceding section. The derived correspondence will be said to be geometric to emphasize the fact that it is essentially founded on a study of the group and of the invariants associated with its subgroups.

A. Characterization of operators by their diagonal elements in the various ξ-bases

Consider an operator $\mathbf{A}$ defined by a kernel according to

$$[\boldsymbol{A}S](f_1)=\int_{\mathbb{R}^+}A(f_1,f_2)S(f_2)f_2^{2r+1}\,df_2. \tag{III.1}$$

Its diagonal matrix elements on the G_ξ-invariant basis defined by (II.16) are given by

$$I_{\mathcal{H}}(\xi,\beta)\equiv(\psi_\beta^\xi,\mathbf{A}\psi_\beta^\xi)=\int_0^\infty\int_0^\infty\psi_\beta^{\xi*}(f_1)A(f_1,f_2)\psi_\beta^\xi(f_2)(f_1f_2)^{2r+1}\,df_1\,df_2. \tag{III.2}$$

Quantities $I_{\mathcal{H}}(\xi,\beta)$ have several properties that will be detailed below. First, for any given value of ξ and β, they are invariant by action of the representation $\boldsymbol{U}_\xi$ on the operator $\mathbf{A}$. Moreover, the operator $\mathbf{A}$ can be recovered from the values of $I_{\mathcal{H}}(\xi,\beta)$ $(-\infty<\xi<\infty,-\infty<\beta<\infty)$, and a simple formula exists for the trace of operators. To be able to discriminate between functions $I_{\mathcal{H}}(\xi,\beta)$ associated with distinct operators, we will also use the notation $I_{\mathcal{H}}(\mathbf{A};\xi,\beta)$ for (III.2).

1. G_ξ-invariance of $I_{\mathcal{H}}(\xi,\beta)$

The matrix elements of the transformed operator $\mathbf{A}'=\mathbf{U}_\xi^{-1}\mathbf{A}\mathbf{U}_\xi$ are given by

$$(\psi_\beta^\xi,\mathbf{U}_\xi^{-1}\mathbf{A}\mathbf{U}_\xi\psi_\beta^\xi)=(\mathbf{U}_\xi\psi_\beta^\xi,\mathbf{A}\mathbf{U}_\xi\psi_\beta^\xi)=(\psi_\beta^\xi,\mathbf{A}\psi_\beta^\xi)$$

the last equality being a consequence of property (II.17).

2. Affine covariance

The action of the full group on $I_{\mathcal{H}}(\xi,\beta)$ is computed in the same way, using property (II.20). One obtains

$$I_{\mathcal{H}}(\mathbf{A}';\xi,\beta)=I_{\mathcal{H}}(\mathbf{A};e^{u}\xi+b,\beta), \tag{III.3}$$

where $\mathbf{A}'=\mathbf{U}^{-1}(u,b)\mathbf{A}\mathbf{U}(u,b)$.

This relation expresses the diagonal elements of $\mathbf{A}'$ in the ξ-basis in terms of the diagonal elements of $\mathbf{A}$ in the basis corresponding to the subgroup $G_{\xi'}$, conjugate of G_{ξ} by the transformation (u,b) of the affine group.

3. Reconstruction of the operator A

Defining the two-dimensional Fourier transform of $I_{\mathcal{H}}(\xi,\beta)$ by

$$\hat{I}_{\mathcal{H}}(u,v)\equiv\int_{\mathbb{R}^2}e^{-2i\pi(u\xi+v\beta)}I_{\mathcal{H}}(\xi,\beta)d\xi\,d\beta, \tag{III.4}$$

we can compute it from (III.2) and find

$$\hat{I}_{\mathcal{H}}(u,v)=A\left(\frac{ue^{(v/2)}}{2\sinh(v/2)},\frac{ue^{(-v/2)}}{2\sinh(v/2)}\right)\left(\frac{u}{2\sinh(v/2)}\right)^{2r}\frac{|u|}{4\sinh^2(v/2)}Y(u/v). \tag{III.5}$$

The inversion of this relation can be done directly by changing the variables.

The kernel of operator $\mathbf{A}$ is thus given by

$$A(f_1,f_2)=\hat{I}_{\mathcal{H}}(f_1-f_2,\ln(f_1/f_2))(f_1f_2)^{-r-1}|f_1-f_2|. \tag{III.6}$$

4. Trace of a product of two operators

The scalar product of two operators can be defined as

$$\mathrm{Tr}(\mathbf{A}\mathbf{B}^{\dagger})=\int_{\mathbb{R}^+\times\mathbb{R}^+}A(f_1,f_2)B^{*}(f_1,f_2)(f_1f_2)^{2r+1}df_1df_2, \tag{III.7}$$

where $\mathbf{B}^{\dagger}$ is the Hermitian conjugate of operator $\mathbf{B}$.

If $\hat{I}_{\mathcal{H}}(\mathbf{A};u,v)$ and $\hat{I}_{\mathcal{H}}(\mathbf{B};u,v)$ are the functions corresponding, respectively, to the operators $\mathbf{A}$ and $\mathbf{B}$, the computation of the trace of $\mathbf{A}\mathbf{B}^{\dagger}$ is performed using (III.6) and the result follows:

$$\mathrm{Tr}(\mathbf{A}\mathbf{B}^{\dagger})=\int_{\mathbb{R}^2}\hat{I}_{\mathcal{H}}(\mathbf{A};u,v)\hat{I}_{\mathcal{H}}(\mathbf{B}^{*};u,v)|u|du\,dv. \tag{III.8}$$

B. Characterization of symbols by their line integrals on ξ-orbits

The description of functions $\mathcal{A}(t,f)$ defined on phase space Γ can be carried out in a way which parallels that followed for the operators. The starting point is the association of a function $I_{\Gamma}(\xi,\beta)$ with each phase space function $\mathcal{A}(t,f)$ by the relation

$$I_{\Gamma}(\xi,\beta)\equiv\int_{\Gamma}\mathcal{A}(t,f)\,\delta((t-\xi)f-\beta)dt\,df. \tag{III.9}$$

For a given ξ, the function defined by (III.9) is interpreted as the integral of $\mathcal{A}(t,f)$ on the various orbits of the subgroup G_{ξ}. The study of the properties of function $I_{\Gamma}(\xi,\beta)$ follows the same steps as in Sec. III A. For practical reasons, the notation $I_{\Gamma}(\mathcal{A};\xi,\beta)$ will also be used for expression (III.9).

1. G_ξ-invariance of $I_\Gamma(\xi,\beta)$

The integral in (III.9) is clearly invariant under the action of G_ξ on $\mathcal{A}(t,f)$ defined by (II.24).

2. Affine covariance

When $\mathcal{A}(t,f)$ is transformed into $\mathcal{A}'(t,f)\equiv\mathcal{A}(e^u t+b,a^{-1}f)$, the function $I_\Gamma(\mathcal{A};\xi,\beta)$ becomes

$$I_\Gamma(\mathcal{A}';\xi,\beta)=I_\Gamma(\mathcal{A};e^u\xi+b,\beta). \tag{III.10}$$

This relation connects the integrals of the function $\mathcal{A}'$ on the G_ξ orbits to the integrals of $\mathcal{A}$ on the orbits of the subgroup $G_{\xi'}$, conjugate of G_ξ by the transformation (u,b) of the affine group.

3. Reconstruction of the function $\mathcal{A}(t,f)$

Relation (III.9) defines, in fact, the Radon transform of $\mathcal{A}(t,f)$ with respect to arrays of hyperbolas parametrized by ξ and β. To invert this transform, we introduce the two-dimensional Fourier transform of $I_\Gamma(\xi,\beta)$ in a manner analogous to (III.4) and compute

$$\hat{I}_\Gamma(u,v)=\int_{\mathbb{R}} e^{-2i\pi ut}\mathcal{A}(t,(u/v))\frac{dt}{|v|}. \tag{III.11}$$

The inversion of this formula gives

$$\mathcal{A}(t,f)=f\int_{\mathbb{R}} e^{2i\pi vft}\hat{I}_\Gamma(fv,v)|v|dv. \tag{III.12}$$

4. Scalar product of two symbols

Consider two functions $\mathcal{A}(t,f)$, $\mathcal{B}(t,f)$ and their corresponding $\hat{I}_\Gamma$ functions. The use of relation (III.12) and Parseval's formula yields

$$\int_\Gamma \mathcal{A}(t,f)\mathcal{B}^*(t,f)dt\,df=\int_{\mathbb{R}^2}\hat{I}_\Gamma(\mathcal{A};u,v)\hat{I}_\Gamma(\mathcal{B}^*;u,v)|u|du\,dv. \tag{III.13}$$

C. The geometric correspondence rule

So far, operators on the Hilbert space $\mathcal{H}$ and functions on the phase space Γ have been characterized, respectively, by functions $I_{\mathcal{H}}(\xi,\beta)$ or $I_\Gamma(\xi,\beta)$. Moreover, for each value of ξ, both functions $I_{\mathcal{H}}(\xi,\beta)$ and $I_\Gamma(\xi,\beta)$ are invariant by action of the subgroup G_ξ. Thus, it is possible to define the correspondence between operators and symbols by requiring that

$$I_{\mathcal{H}}(\xi,\beta)\equiv I_\Gamma(\xi,\beta), \tag{III.14}$$

and to denote the common function by $I(\xi,\beta)$. This relation is stable by the affine group as seen from properties (III.3) and (III.10). Actually, the consistency with the group action implies only the proportionality of the two members of (III.14). The interest of the strict identification which has been adopted is to ensure that the identity operator will be symbolized by the function one.

According to (III.5) and (III.11), the Fourier transform of relation (III.14) leads to

$$\hat{I}(u,v)=A\left(\frac{ue^{(v/2)}}{2\sinh(v/2)},\frac{ue^{(-v/2)}}{2\sinh(v/2)}\right)\left(\frac{u}{2\sinh(v/2)}\right)^{2r}\frac{|u|}{4\sinh^2(v/2)}Y(u/v)$$

$$=\int_{\mathbb{R}} e^{-2i\pi ut}\mathcal{A}(t,(u/v))\frac{dt}{|v|}.$$

The equality of the two right hand sides gives the relation between the kernel $A(f_1,f_2)$ and the symbol $\mathcal{A}(t,f)$. This is the correspondence rule that can be rewritten either as a formula giving the operator kernel in terms of the symbol (Weyl mapping) or as a formula giving the symbol in terms of the kernel (Wigner correspondence). The result is the following

Result 1: The operator **A** *corresponding to the function* $\mathcal{A}(t,f)$*, defined on the time-frequency half-plane* Γ*, is given by its kernel [defined in (III.1)] as*

$$A(f_1,f_2)=(f_1f_2)^{-r-1}\frac{f_1-f_2}{\ln(f_1/f_2)}\int_{\mathbb{R}}e^{-2i\pi(f_1-f_2)t}\mathcal{A}\left(t,\frac{f_1-f_2}{\ln(f_1/f_2)}\right)dt. \tag{III.15}$$

Conversely, the operator **A** *defined by its kernel* $A(f_1,f_2)$*, is represented on phase space by a symbol* $\mathcal{A}(t,f)$ *according to*

$$\mathcal{A}(t,f)=f^{2r+2}\int_{\mathbb{R}}e^{2i\pi vft}A\left(\frac{fve^{v/2}}{2\sinh(v/2)},\frac{fve^{-v/2}}{2\sinh(v/2)}\right)\left(\frac{v}{2\sinh(v/2)}\right)^{2r+2}dv. \tag{III.16}$$

In addition, the following unitarity property holds:

$$\int_{\Gamma}\mathcal{A}(t,f)\mathcal{B}^*(t,f)dt\ df=\mathrm{Tr}(\mathbf{AB}^{\dagger}). \tag{III.17}$$

Among the properties of the correspondence, it can be deduced from (III.16) that the identity operator with kernel

$$I(f_1,f_2)=(f_1f_2)^{-r-1/2}\delta(f_1-f_2), \tag{III.18}$$

has a symbol equal to one, and that operators such that

$$A(f_1,f_2)=A^*(f_2,f_1), \tag{III.19}$$

are in correspondence with real-valued functions $\mathcal{A}(t,f)$.

At this point, it would be desirable to characterize in a precise way classes of operators and functions that are put in correspondence by the above rule. Clearly, because of property (III.17), Hilbert-Schmidt operators can be associated with functions in the space $L^2(\mathbb{R}^+;f^{2r+1}\ df)$. In fact, it is possible to characterize other classes of operators by their symbols,[4,5] as is done in the usual Weyl case with bounded operators.[6,7] But this is outside the scope of the present paper.

Result 1 concerns only functions defined on the positive half-line and their corresponding operators. Some side comments concerning its possible extension are given in the following remark.

Remark 1: A direct extension of Result 1 can be obtained if a function $\mathcal{A}(t,f)$ defined on the whole plane is substituted in (III.15) and if negative values for f_1 and f_2 are allowed. However, the result will be a function $A(f_1,f_2)$ defined only in the first and third quadrant of the (f_1,f_2) plane. The regions where f_1 and f_2 have different signs are definitely forbidden in the above symbolic calculus.

IV. SOME SPECIAL CASES

A. Symbol of the projector on a β-subspace of $\mathcal{H}$

The β-subspace associated with subgroup G_ξ is characterized by the function $\psi^\xi_\beta(f)$ introduced in (II.16). The projector on this subspace is the linear operator defined by

$$S(f)\rightarrow(\psi^\xi_\beta,S)\psi^\xi_\beta(f), \tag{IV.1}$$

where the scalar product has the form (II.7). The corresponding kernel is given by [cf. (III.1)]

$$\Pi^\xi_\beta(f_1,f_2)=\psi^\xi_\beta(f_1)\psi^{\xi*}_\beta(f_2). \tag{IV.2}$$

A direct application of formula (III.16) shows that the symbol associated with (IV.2) is the distribution $\delta((t-\xi)f-\beta)$ with support in Γ localized on the β-orbit of subgroup G_ξ.

B. Affine Wigner function as a particular symbol

Each signal $S(f)$ defines a real linear functional on the set of Hermitian operators by

$$\langle \mathbf{A} \rangle = (\mathbf{A}S, S) = (S, \mathbf{A}S), \tag{IV.3}$$

where the scalar product is given by (II.7).

This functional can also be written as

$$\langle \mathbf{A} \rangle = \mathrm{Tr}(\mathbf{\Pi}_S \mathbf{A}), \tag{IV.4}$$

where the trace operation has been defined by (III.7) and where $\mathbf{\Pi}_S$ is the projector on signal $S(f)$ with kernel given by

$$\Pi_S(f_1, f_2) = S(f_1)S^*(f_2). \tag{IV.5}$$

From the reciprocal correspondence (III.15), (III.16) and its properties, it results that each real linear functional on the set of Hermitian operators can be expressed as a real linear functional of the real-valued functions on the time-frequency half-plane. The Wigner function for the signal $S(f)$ is defined as the function (or distribution) $\mathcal{P}(t,f)$ on the time-frequency half-plane which allows us to rewrite (IV.4) in the form

$$\langle \mathbf{A} \rangle = \int_\Gamma \mathcal{P}(t,f)\mathcal{A}(t,f)\,dt\,df, \tag{IV.6}$$

where $\mathcal{A}(t,f)$ is the symbol of operator $\mathbf{A}$. From this definition, it is clear that the form of the Wigner function is fundamentally dependent of the symbolic calculus which is adopted.

The equality of the right hand sides of (IV.4) and (IV.6) gives a relation of type (III.17) (unitarity property) provided the Wigner function $\mathcal{P}(t,f)$ is identified with the geometric symbol of the projector $\mathbf{\Pi}_S$. This makes it possible to obtain the relevant Wigner function by applying the correspondence rule (III.16) to the kernel (IV.5). The result is the affine Wigner function:[8]

$$\mathcal{P}(t,f) = f^{2r+2}\int_{\mathbb{R}} e^{2i\pi v f t} S\left(\frac{f v e^{v/2}}{2\sinh(v/2)}\right) S^*\left(\frac{f v e^{-v/2}}{2\sinh(v/2)}\right)\left(\frac{v}{2\sinh(v/2)}\right)^{2r+2} dv. \tag{IV.7}$$

This function can be introduced in various ways. In the above development, it arises as a consequence of the general correspondence defined by (III.15), (III.16).

Expression (IV.7) can also be obtained directly by writing the unitarity relation (III.17) with one symbol equal to $\mathcal{P}(t,f)$ and the other one equal to the localized distribution $\delta(t-t_0)\delta(f-f_0)$. This leads to the relation

$$\mathcal{P}(t_0, f_0) = \mathrm{Tr}(\mathbf{\Pi}_S \mathbf{\Delta}_{t_0,f_0}), \tag{IV.8}$$

where $\mathbf{\Delta}_{t_0,f_0}$ is the operator with symbol $\delta(t-t_0)\delta(f-f_0)$. The kernel of this operator is given by (III.15) and reads as

$$\Delta(f_1, f_2) = (f_1 f_2)^{-r-1}\frac{f_1 - f_2}{\ln f_1/f_2} e^{-2i\pi(f_1-f_2)t_0}\delta\left(\frac{f_1-f_2}{\ln f_1/f_2} - f_0\right). \tag{IV.9}$$

Function $\mathcal{P}(t,f)$ can then be obtained by (IV.8). These results are summarized by the following.

Result 2: The Wigner function of signal $S(f)$ associated with the geometric correspondence is given by (IV.7). It can be obtained either as the symbol of the projector $\mathbf{\Pi}_S$ on signal $S(f)$, or as the mean value (IV.8) of the operator $\mathbf{\Delta}_{t_0,f_0}$ whose symbol is $\delta(t-t_0)\delta(f-f_0)$.

The unitarity property (III.17) gives directly the formulas

$$\int_\Gamma \mathcal{P}(t,f)\,dt\,df = \|S\|^2,$$
$$\int_\Gamma \mathcal{P}_1(t,f)\mathcal{P}_2(t,f)\,dt\,df = \mathrm{Tr}(\mathbf{\Pi}_{S_1}\mathbf{\Pi}_{S_2}) = |(S_1,S_2)|^2, \tag{IV.10}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ stand for the Wigner functions corresponding to S_1 and S_2, respectively.

Many other properties could be mentioned which are related to the interpretation of the Wigner function as a time-frequency representation.[9,10]

C. Exponentials of generators and Weyl's formulation of the correspondence

We will now show that the symbol of the operator $\mathbf{E}_{u_0v_0} \equiv e^{-2i\pi(u_0\boldsymbol{\beta}+v_0\mathbf{f})}$, where $\boldsymbol{\beta}$ and $\mathbf{f}$ are the generators defined by (II.8) and (II.9), is given by the function $\mathcal{E}_{u_0v_0}(t,f) = e^{-2i\pi(u_0tf+v_0f)}$.

It follows from the developments of Sec. II A that the operator $\mathbf{E}_{u_0v_0}$ satisfies the following eigenvalue equation:

$$e^{-2i\pi(u_0\boldsymbol{\beta}+v_0\mathbf{f})}\psi_\beta^{-v_0/u_0}(f) = e^{-2i\pi\beta u_0}\psi_\beta^{-v_0/u_0}(f). \tag{IV.11}$$

The kernel of $\mathbf{E}_{u_0v_0}$ is obtained directly from the spectral decomposition:

$$E_{u_0v_0}(f_1,f_2) = \int_{\mathbb{R}} e^{-2i\pi\beta u_0}\psi_\beta^{-v_0/u_0}(f_1)\psi_\beta^{*-v_0/u_0}(f_2)\,d\beta, \tag{IV.12}$$

and has the explicit expression

$$E_{u_0v_0}(f_1,f_2) = e^{(r+1)u_0}\exp\{-2i\pi f_1(v_0/u_0)(e^{u_0}-1)\}\,\delta(f_1e^{u_0}-f_2)f_2^{-2r-1}. \tag{IV.13}$$

Substituting this result in (III.16), we can write the symbol of $\mathbf{E}_{u_0v_0}$ as

$$\mathcal{E}_{u_0v_0}(t,f) = \int_{\mathbb{R}} e^{2i\pi vft}e^{(r+1)u_0}\exp\{-2i\pi(f(v_0/u_0)(e^{u_0}-1)\lambda(v))\}$$
$$\times\lambda(-v)e^{(2r+2)(v/2)}\delta(e^{u_0}\lambda(v)-\lambda(-v))\,dv, \tag{IV.14}$$

where, by definition,

$$\lambda(v) = \frac{ve^{v/2}}{2\sinh(v/2)}. \tag{IV.15}$$

Since

$$\delta(e^{u_0}\lambda(v)-\lambda(-v)) = \frac{1}{\lambda(u_0)}\,\delta(v+u_0), \tag{IV.16}$$

the result of integration in (IV.14) is just

$$\mathcal{E}_{u_0v_0}(t,f) = e^{-2i\pi(u_0ft+v_0f)}. \tag{IV.17}$$

Thus, we have established the correspondence

$$\mathcal{E}_{u_0v_0}(t,f) \equiv e^{-2i\pi(u_0ft+v_0f)} \leftrightarrow \mathbf{E}_{u_0v_0} = e^{-2i\pi(u_0\boldsymbol{\beta}+v_0\mathbf{f})}. \tag{IV.18}$$

Remark that the variable $\beta = tf$ rather than t appears naturally. It will sometimes be convenient to use it by performing the change of variables according to

$$\tilde{\mathcal{A}}(\beta,f) = [\mathcal{A}(t,f)]_{t=\beta/f}. \tag{IV.19}$$

The correspondence (IV.18) allows us to write the operator $\mathbf{A}$ in terms of its symbol $\mathcal{A}(t,f)$ by the Weyl formula:

$$\mathbf{A}=\int \hat{\mathcal{A}}(u,v)e^{-2i\pi(u\boldsymbol{\beta}+v\boldsymbol{f})}du\ dv, \tag{IV.20}$$

where $\hat{\mathcal{A}}(u,v)$ is the Fourier transform of $\tilde{\mathcal{A}}(\beta,f)$ given by

$$\hat{\mathcal{A}}(u,v)\equiv\int_{\mathbb{R}\times\mathbb{R}^+} e^{2i\pi(u\beta+vf)}\tilde{\mathcal{A}}(\beta,f)d\beta\ df. \tag{IV.21}$$

The correspondence between functions and operators defined by (IV.20) and (IV.21) has a form analogous to Weyl's original rule. In fact, it is what Cohen calls Weyl's rule for operators $\boldsymbol{\beta}$ and $\boldsymbol{f}$.[11] It can also be obtained as a by-product of Kirillov's analysis[12] as shown by Unterberger[4] (see also Shenoy and Parks[13]).

Remark 2: The symbols of the self-adjoint operators $\boldsymbol{\beta}$ and $\boldsymbol{f}$ are immediately seen to be $\beta=tf$ and f respectively. The time operator, which is especially important in signal analysis, can be defined as the operator with symbol t. A direct computation leads to the result

$$\boldsymbol{t}=(1/2)(\boldsymbol{f}^{-1}\boldsymbol{\beta}+\boldsymbol{\beta}\boldsymbol{f}^{-1}), \tag{IV.22}$$

or

$$\boldsymbol{t}=-\frac{1}{2i\pi}\left(\frac{r+1/2}{f}+\frac{d}{df}\right). \tag{IV.23}$$

Such an operator is not essentially self-adjoint on the Hilbert space $L^2(\mathbb{R}^+;f^{2r+1}df)$ and hence has no conventional spectral decomposition. However, it is known from the deformation quantization theory on phase space[2] that such a difficulty can be easily overcome and does not affect the physical interpretation.

V. STAR PRODUCT ON THE SYMBOLS

Operators on the Hilbert space $\mathcal{H}$ have been mapped to functions on the time-frequency half-plane Γ by the rule (III.16). Since the algebra of operators is non-Abelian, the product of operators is mapped into a composition law of functions that cannot be the ordinary product. This situation is familiar from Weyl's calculus where the composition law of functions is known as Moyal's product and is invariant by Heisenberg's group. Analogous products, invariant by different groups, have been constructed by deformation of the usual product and are generally referred to as star products ($\star$ products).[2] In the present case, the composition law of symbols corresponding to the product of operators leads directly to a star product that is invariant by the affine group.

In the following, it will be convenient to use another notation for the operator $\exp\{-2i\pi(u\boldsymbol{\beta}+v\boldsymbol{f})\}$. In fact, according to (IV.11) and (IV.12), the action of that operator on a signal $S(f)$ can be written as

$$\exp\{-2i\pi(u\boldsymbol{\beta}+v\boldsymbol{f})\}S(f)=e^{(r+1)u}\exp\{-2i\pi f(v/u)(e^u-1)\}S(fe^u). \tag{V.1}$$

Comparing this expression with the definition (II.6) of representation $\mathbf{U}$, we are able to rewrite (V.1) as

$$\exp\{-2i\pi(u\boldsymbol{\beta}+v\boldsymbol{f})\}S(f)=\mathbf{U}\left(u,\frac{v}{u}(e^u-1)\right)S(f). \tag{V.2}$$

This result can also be obtained directly by noticing that, in the description of the affine group by 2×2 matrices, the exponentiation of the Lie algebra element defined by (u,v) is equal to

$$\exp\begin{vmatrix} u & v \\ 0 & 0 \end{vmatrix}=\begin{vmatrix} e^u & (v/u)(e^u-1) \\ 0 & 1 \end{vmatrix}.$$

The use of representation $\mathbf{U}(u,b)$ allows us to take advantage of the group law and to write the composition of operators as

$$\mathbf{U}(u,b)\mathbf{U}(u',b')=\mathbf{U}(u+u',b+e^{u}b'). \tag{V.3}$$

It shows directly that the product of two operators of type (V.1) is an operator of the same type. With notation (V.2), the form (IV.20) of the correspondence rule (III.15) between functions and operators can be written as

$$\mathbf{A}=\int_{\mathbb{R}^2}\hat{\mathcal{A}}(u,v)\mathbf{U}\left(u,v\,\frac{e^{u}-1}{u}\right)du\;dv, \tag{V.4}$$

where $\hat{\mathcal{A}}(u,v)$ is defined in (IV.21).

Let **A** and **B** be operators given in terms of their symbols $\tilde{\mathcal{A}}(\beta,f)$ and $\tilde{\mathcal{B}}(\beta,f)$ by rule (V.4), (IV.21). Their product is equal to

$$\mathbf{AB}=\int_{\mathbb{R}^4}\hat{\mathcal{A}}(u_1,v_1)\hat{\mathcal{B}}(u_2,v_2)\mathbf{U}\left(u_1,v_1\frac{e^{u_1}-1}{u_1}\right)\mathbf{U}\left(u_2,v_2\frac{e^{u_2}-1}{u_2}\right)du_1\;dv_1\;du_2\;dv_2,$$

or, using (V.3),

$$=\int_{\mathbb{R}^4}\hat{\mathcal{A}}(u_1,v_1)\hat{\mathcal{B}}(u_2,v_2)\mathbf{U}\left(u_1+u_2,v_1\frac{e^{u_1}-1}{u_1}+e^{u_1}v_2\frac{e^{u_2}-1}{u_2}\right)du_1\;dv_1\;du_2\;dv_2.$$

This product can be rewritten in the form (V.4) of the correspondence rule if variables (u,v) are introduced such that

$$u=u_1+u_2,\quad v\,\frac{e^{u}-1}{u}=v_1\frac{e^{u_1}-1}{u_1}+e^{u_1}v_2\frac{e^{u_2}-1}{u_2}. \tag{V.5}$$

The result is

$$\mathbf{AB}=\int_{\mathbb{R}^2}\hat{\mathcal{C}}(u,v)\mathbf{U}\left(u,v\,\frac{e^{u}-1}{u}\right)du\;dv, \tag{V.6}$$

with $\hat{\mathcal{C}}(u,v)$ defined by

$$\begin{aligned}\hat{\mathcal{C}}(u,v)=&\int_{\mathbb{R}^2\times\mathbb{R}^2}\hat{\mathcal{A}}(u_1,v_1)\hat{\mathcal{B}}(u_2,v_2)\,\delta(u-u_1-u_2)\\&\times\delta\left(v-\left(v_1\frac{e^{u_1}-1}{u_1}+e^{u_1}v_2\frac{e^{u_2}-1}{u_2}\right)\frac{u_1+u_2}{e^{u_1+u_2}-1}\right)du_1\;dv_1\;du_2\;dv_2.\end{aligned} \tag{V.7}$$

It follows from (V.6) that $\hat{\mathcal{C}}(u,v)$ is the Fourier transform of the symbol of the operator **AB**.

To obtain an explicit expression of the star product $\tilde{\mathcal{A}}(\beta,f)\star\tilde{\mathcal{B}}(\beta,f)$, it suffices now to invert the Fourier transform of $\hat{\mathcal{C}}(u,v)$ and to express the result in terms of $\tilde{\mathcal{A}}(\beta,f)$ and $\tilde{\mathcal{B}}(\beta,f)$. In a first step, we obtain

$$\begin{aligned}\tilde{\mathcal{A}}(\beta,f)\star\tilde{\mathcal{B}}(\beta,f)=&\int\hat{\mathcal{A}}(u_1,v_1)\hat{\mathcal{B}}(u_2,v_2)e^{2i\pi((u_1+u_2)\beta+(v_1+v_2)f)}\\&\times e^{-2i\pi f h(u_1,v_1,u_2,v_2)}du_1\;dv_1\;du_2\;dv_2,\end{aligned} \tag{V.8}$$

where the function h is defined by

$$h(u_1,v_1,u_2,v_2)\equiv(v_1+v_2)-\frac{u_1+u_2}{e^{u_1+u_2}-1}\left(v_1\frac{e^{u_1}-1}{u_1}+e^{u_1}v_2\frac{e^{u_2}-1}{u_2}\right), \tag{V.9}$$

or, after some manipulations,

$$h(u_1,v_1,u_2,v_2)=(v_1u_2-v_2u_1)\frac{u_1e^{u_1}(e^{u_2}-1)-u_2(e^{u_1}-1)}{u_1u_2(e^{u_1+u_2}-1)}. \tag{V.10}$$

The exponential of function $h(u_1,v_1,u_2,v_2)$ in (V.8) is what makes the difference with a classical product. It can be replaced by its series expansion so that the integrals in (V.8) can be performed. In the process, the following replacements take place:

$$v_i\to\frac{1}{2i\pi}\partial_{f_i},\quad u_i\to\frac{1}{2i\pi}\partial_{\beta_i},\quad i=1,2. \tag{V.11}$$

The development of the star product can thus be written as

$$\tilde{\mathcal{A}}(\beta,f)\star\tilde{\mathcal{B}}(\beta,f)=\tilde{\mathcal{A}}(\beta,f)\tilde{\mathcal{B}}(\beta,f)+\sum_{n=1}^{\infty}\frac{1}{n!}\left(\frac{f}{4i\pi}(\partial_{f_2}\partial_{\beta_1}-\partial_{f_1}\partial_{\beta_2})T(\partial_{\beta_1},\partial_{\beta_2})\right)^n\tilde{\mathcal{A}}(\beta_1,f_1)\tilde{\mathcal{B}}(\beta_2,f_2)\Big|_{\substack{f_1=f_2=f\\ \beta_1=\beta_2=\beta}}, \tag{V.12}$$

where the contribution of operator T will be discussed below.

If (V.12) is rewritten in terms of t rather than β, using the definition

$$\beta=tf,\quad \partial_\beta=(1/f)\partial_t; \tag{V.13}$$

the result is reminiscent of the Weyl product. Actually, the difference lies in the presence of the operator $T(\partial_{\beta_1},\partial_{\beta_2})$ which comes from the development in powers of u_1 and u_2 of

$$T(u_1,u_2)=2\frac{u_1e^{u_1}(e^{u_2}-1)-u_2(e^{u_1}-1)}{u_1u_2(e^{u_1+u_2}-1)}. \tag{V.14}$$

Operator T introduces derivatives of all orders with respect to β_1 and β_2. At the lowest orders, one has

$$T(\partial_{\beta_1},\partial_{\beta_2})=1+\frac{i}{12\pi}(\partial_{\beta_2}-\partial_{\beta_1})+\frac{1}{12(2\pi)^2}\partial_{\beta_1}\partial_{\beta_2}+\frac{i}{720(2\pi)^3}(\partial_{\beta_2}-\partial_{\beta_1})(\partial_{\beta_1}^2+\partial_{\beta_2}^2+5\partial_{\beta_1}\partial_{\beta_2})+\cdots. \tag{V.15}$$

As in Weyl's case, the first term in (V.12) is the usual product and the following terms introduce derivatives of the symbols. It can be observed that the star product of two real-valued functions is not real. This is a direct manifestation of the fact that the product of Hermitian operators is not Hermitian.

Remark 3: From Remark 1, it follows that in a product of operators, the positive and negative frequency parts are multiplied separately. Hence, the same property will hold for the star product if its expression is extended to symbols defined on the whole plane.

The star product simplifies when one of the symbols depends only on one variable. This will be illustrated by computing its expression when $\tilde{\mathcal{A}}(\beta,f)$ is equal either to β or to a function of f alone and when $\tilde{\mathcal{B}}(\beta,f)$ is the generic distribution:

$$\Delta_{\beta_0,f_0}(\beta,f)=\delta(\beta-\beta_0)\delta(f-f_0), \tag{V.16}$$

with Fourier transform (IV.21) given by

$$\hat{\Delta}_{\beta_0,f_0}(u,v)=e^{-2i\pi(u\beta_0+vf_0)}. \tag{V.17}$$

(i) Case where $\tilde{\mathcal{A}}(\beta,f)=\beta$. The Fourier transform of β is

$$\hat{\mathcal{A}}(u,v)=-\frac{1}{2i\pi}\delta'(u)\delta(v). \tag{V.18}$$

Computation of the star product from (V.8) and (V.10) gives

$$\beta\star\delta(\beta-\beta_0)\delta(f-f_0)=-\frac{1}{2i\pi}\int \delta'(u_1)e^{-2i\pi[u_2(\beta_0-\beta)+v_2f_0]} \times\exp\left\{2i\pi\left[u_1\beta+fv_2\frac{\lambda(u_1+u_2)}{\lambda(u_2)}\right]\right\}du_1\,du_2\,dv_2, \tag{V.19}$$

where $\lambda(u)$, defined by (IV.15), is such that

$$\lambda(-u)=e^{-u}\lambda(u). \tag{V.20}$$

Performing the u_1-integration, we can write (V.19) as

$$\beta\star\delta(\beta-\beta_0)\delta(f-f_0)=\beta\delta(\beta-\beta_0)\delta(f-f_0) +f\int_{\mathbb{R}^2}e^{2i\pi[u_2(\beta-\beta_0)+v_2(f-f_0)]}\left(\frac{d}{du_2}(\ln(\lambda(u_2)))\right)v_2\,du_2\,dv_2. \tag{V.21}$$

(ii) Case where $\tilde{\mathcal{A}}(\beta,f)$ *depends only on* f. The Fourier transform of $\tilde{\mathcal{A}}(\beta,f)=g(f)$, with $g(f)$ an arbitrary function, is equal to

$$\hat{\mathcal{A}}(u,v)=\delta(u)\int_{\mathbb{R}}g(f)e^{-2i\pi vf}\,df. \tag{V.22}$$

Substitution of the Fourier transforms (V.17) and (V.22) in the expression of the star product (V.8)–(V.10) gives

$$g(f)\star\delta(\beta-\beta_0)\delta(f-f_0)=\int_{\mathbb{R}^2}g(f\lambda(-u))e^{2i\pi[u(\beta-\beta_0)+v(f-f_0)]}\,du\,dv. \tag{V.23}$$

VI. STAR BRACKET OF SYMBOLS AND HAMILTONIAN FLOW

A. Definition of the star bracket

A Lie bracket is now defined on operators by the operation

$$-2\pi i[\mathbf{A},\mathbf{B}]=-2\pi i(\mathbf{AB}-\mathbf{BA}). \tag{VI.1}$$

This bracket defines a composition law on the subspace of Hermitian operators. When transformed by the correspondence rule, using the expression for the star product, this bracket gives rise to an operation on real-valued functions that will be called a star bracket and denoted by $\{,\}_\star$. Thus, the star bracket of symbols $\tilde{\mathcal{A}}(\beta,f)$ and $\tilde{\mathcal{B}}(\beta,f)$ is defined by

$$\{\tilde{\mathcal{A}},\tilde{\mathcal{B}}\}_\star(\beta,f)=-2\pi i(\tilde{\mathcal{A}}(\beta,f)\star\tilde{\mathcal{B}}(\beta,f)-\tilde{\mathcal{B}}(\beta,f)\star\tilde{\mathcal{A}}(\beta,f)). \tag{VI.2}$$

According to (V.12), the bracket can be written as

$$\{\tilde{\mathcal{A}}(\beta,f),\tilde{\mathcal{B}}(\beta,f)\}_\star=\Bigg[f(\partial_{f_1}\partial_{\beta_2}-\partial_{f_2}\partial_{\beta_1})\Bigg(1+\frac{1}{12(2\pi)^2}\partial_{\beta 1}\partial_{\beta_2}+\cdots\Bigg)$$
$$+\sum_{n=2}^{\infty}\frac{2i\pi}{n!}\Bigg(\frac{f}{4i\pi}(\partial_{f_1}\partial_{\beta_2}-\partial_{f_2}\partial_{\beta_1})\Bigg)^{n'}$$
$$\times[T^n(\partial_{\beta_2},\partial_{\beta_1})-(-1)^nT^n(\partial_{\beta_1},\partial_{\beta_2})]\Bigg]\tilde{\mathcal{A}}(\beta_1,f_1)\tilde{\mathcal{B}}(\beta_2,f_2)\Bigg|_{\substack{f_1=f_2=f\\ \beta_1=\beta_2=\beta}}. \tag{VI.3}$$

The first term in f is a Poisson bracket given by

$$\{\tilde{\mathcal{A}}(\beta,f),\tilde{\mathcal{B}}(\beta,f)\}_P=f\Bigg(\frac{\partial\tilde{\mathcal{A}}(\beta,f)}{\partial\beta}\frac{\partial\tilde{\mathcal{B}}(\beta,f)}{\partial f}-\frac{\partial\tilde{\mathcal{A}}(\beta,f)}{\partial f}\frac{\partial\tilde{\mathcal{B}}(\beta,f)}{\partial\beta}\Bigg). \tag{VI.4}$$

In fact, if the variable β is replaced by its definition (V.13), the expression (VI.4) becomes identical with the usual Poisson bracket on Γ. This is the same bracket as obtained by Kirillov's theory[12] when considering phase space as an orbit of the coadjoint representation.

The form (VI.3) of the star bracket is reminiscent of that obtained with the Weyl calculus. However, due to the fact that operator $T(\partial_{\beta_1},\partial_{\beta_2})$ has no symmetry, derivatives of arbitrary order will, in general, appear in the expression of the bracket.

Some symbols are such that their star bracket with any other function has the form of a Poisson bracket. In deformation quantization, such symbols are called *preferred observables*.[2,14,15] A typical example is the harmonic oscillator Hamiltonian on $\mathbb{R}^2$ with Weyl quantization.

In the present case, the preferred observable are represented by the symbols β, f, and $\ln f$. Explicitly, their star brackets with any symbol $\mathcal{X}(\beta,f)$ are given by

$$\{\beta,\mathcal{X}(\beta,f)\}_\star=f\frac{\partial}{\partial f}\mathcal{X}(\beta,f), \tag{VI.5}$$

$$\{f,\mathcal{X}(\beta,f)\}_\star=-f\frac{\partial}{\partial\beta}\mathcal{X}(\beta,f), \tag{VI.6}$$

$$\{\ln f,\mathcal{X}(\beta,f)\}_\star=-\frac{\partial}{\partial\beta}\mathcal{X}(\beta,f). \tag{VI.7}$$

The derivation of these formulas can be based on expressions (V.21), (V.23) and on their counterpart with reverse order in the star products. Final results follow from the relations

$$\lambda(u)-\lambda(-u)=u,\quad \lambda(-u)=e^{-u}\lambda(u), \tag{VI.8}$$

which are verified by the function (IV.15).

B. Extended covariance and Hamiltonian flows

Consider a one-parameter group of transformations on signals $S(f)$ defined by an operator $\mathbf{O}$ such that

$$S(f;\alpha)=e^{-2i\pi\alpha\mathbf{O}}S(f;0), \tag{VI.9}$$

with $S(f;0)=S_0(f)$ given. Signal $S(f;\alpha)$ verifies the equation

$$\frac{\partial S(f;\alpha)}{\partial\alpha}=-2i\pi\alpha\mathbf{O}S(f;\alpha), \tag{VI.10}$$

and the projector $\mathbf{\Pi}$ on $S(f;\alpha)$ verifies

$$\frac{\partial \mathbf{\Pi}}{\partial \alpha} = -2i\pi[\mathbf{O}, \mathbf{\Pi}]. \tag{VI.11}$$

Now, going to phase space, we know that the geometric symbol of $\mathbf{\Pi}$ is the affine Wigner function (IV.7) for signal $S(f;\alpha)$, and that the symbol of the bracket $-2i\pi[,]$ is the star bracket $\{,\}_\star$. Thus equation (VI.11) becomes the following equation in Γ:

$$\frac{\partial \mathcal{P}(t,f;\alpha)}{\partial \alpha} = \{\mathcal{O}(t,f), \mathcal{P}(t,f;\alpha)\}_\star, \tag{VI.12}$$

where $\mathcal{O}(t,f)$ is the symbol of operator $\mathbf{O}$.

Consider now the case where the symbol $\mathcal{O}(t,f)$ has the particular form

$$\mathcal{O}_H(t,f) = \mu t f + \nu f + \sigma \ln f, \tag{VI.13}$$

where μ, ν and σ are real constants. From the values of the star brackets (VI.5), (VI.6) and (VI.7), it follows that the bracket in (VI.12) reduces to Poisson's bracket. As a result, equation (VI.12) takes the form of a Liouville equation with Hamiltonian equal to $\mathcal{O}_H(t,f)$:

$$\frac{\partial \mathcal{P}(t,f;\alpha)}{\partial \alpha} = \{\mu t f + \nu f + \sigma \ln f, \mathcal{P}(t,f;\alpha)\}_P = \mu f \frac{\partial \mathcal{P}(t,f;\alpha)}{\partial f} - (\mu t + \nu + (\sigma/f)) \frac{\partial \mathcal{P}(t,f;\alpha)}{\partial t}. \tag{VI.14}$$

This means that the α-evolution of the Wigner function is identical to that of an incompressible fluid in the time-frequency half-plane. Integration of equation (VI.14) with the initial condition

$$\mathcal{P}(t,f;0) = \mathcal{P}_0(t,f) \tag{VI.15}$$

gives

$$\mathcal{P}(t,f;\alpha) = \mathcal{P}_0((-\nu\mu^{-1} + (t+\nu\mu^{-1})e^{-\mu\alpha} - \sigma f^{-1}\alpha), f e^{\mu\alpha}). \tag{VI.16}$$

The evolution of the Wigner function in terms of parameter α has a counterpart in Hilbert space that can be obtained using the symbolic calculus. The operator $\mathbf{O}_H$ for this evolution is found directly from (IV.20)–(IV.21) and has the expression

$$\mathbf{O}_H = \mu\boldsymbol{\beta} + \nu\mathbf{f} + \sigma \ln \mathbf{f}, \tag{VI.17}$$

where the operators $\boldsymbol{\beta}$ and $\mathbf{f}$ have been introduced in (II.8) and (II.9). The expression (VI.9) written for the operator $\mathbf{O}_H$ then gives the α-evolution in the Hilbert space which corresponds to the phase space evolution (VI.16). In fact, this leads to a special form of a general transformation which can be written as

$$(u,b,c): S(f) \to S'(f) = e^{(r+1)u} e^{-2i\pi b f} f^{-2i\pi c} S(e^u f), \quad u,b,c \in \mathbb{R}. \tag{VI.18}$$

This transformation is a projective unitary representation of a three-parameter group G_0 which has been studied in Ref. 16. It has been proved that transformations (VI.18) correspond to symplectic transformations in the phase space. This extended covariance of the calculus is at the basis of a generalization of the construction process that will now be considered.

VII. OTHER SYMBOLIZATIONS WITH GEOMETRICAL FEATURES

The method presented above to construct a symbolic calculus on the affine group can be embedded in a more general approach based on the systematic consideration of groups containing the affine group and having the time-frequency half-plane as a coadjoint orbit. Such groups belong to the class of simply connected Lie groups having a coadjoint orbit of dimension smaller than or equal to two.[17] In this class, the groups of interest are $SL(2,\mathbb{R})$ and the three-parameter solvable groups with Lie algebras defined by generators X_a, X_b and X_c satisfying[18]

$$[X_a,X_b]=X_b, \quad [X_a,X_c]=kX_c, \quad k\in\mathbb{R}, \tag{VII.1}$$

$$[X_a,X_b]=X_b, \quad [X_a,X_c]=X_c+X_b. \tag{VII.2}$$

Examples of symbolic calculi invariant by $SL(2,\mathbb{R})$ have been thoroughly studied in Refs. 19–20. Here, we will treat the cases of groups G_k with Lie algebras defined by relations (VII.1) for $k\neq 1$, and (VII.2) for $k=1$. They provide a natural generalization of the group G_0 considered so far in this paper and will be associated with new forms of symbolic calculus. All groups G_k have been introduced in Ref. 16 as a tool to find explicit forms of time-frequency pseudo-distributions. We refer to that paper for the details concerning the precise construction of their unitary representations (projective or not) according to the values of k. Here we need only recall the results corresponding to values of k different from 0 and 1.

The groups G_k, $k\neq 0,1$ consist of elements (u,b,c) with the composition law

$$gg'=(u+u',b+e^u b',c+e^{ku}c'). \tag{VII.3}$$

Their unitary representations in the Hilbert space $\mathcal{H}$ defined in Sec. II A are given by

$$\mathbf{U}_k(u,b,c)S(f)=e^{(r+1)u}e^{-2i\pi(bf+cf^k)}S(e^u f). \tag{VII.4}$$

The action of the groups in the time-frequency half-plane Γ is

$$(t,f)\to(e^u t+b+kce^{-(k-1)u}f^{k-1},e^{-u}f). \tag{VII.5}$$

It is clear from (VII.4) and (VII.5) that the cases $k=0$ and $k=1$ are degenerate and deserve a special treatment (see Ref. 16).

Following the same pattern as in Sec. II, we determine the subgroups of G_k that are conjugate to the dilation subgroup. They are labeled by two real parameters ξ,η and consist of

$$G_{\xi\eta}=\{(e^u,\xi(1-e^u),\eta(1-e^{ku})\}. \tag{VII.6}$$

The restriction of representations (VII.4) to subgroups $G_{\xi\eta}$ is

$$\mathbf{U}_{\xi\eta}S(f)=e^{(r+1)u}e^{-2i\pi(\xi(1-e^u)f+\eta(1-e^{ku})f^k)}S(fe^u). \tag{VII.7}$$

The eigenfunctions of $\mathbf{U}_{\xi\eta}$ are

$$\psi_\beta^{\xi\eta}(f)=f^{-2i\pi\beta-r-1}e^{-2i\pi(\xi f+\eta f^k)}. \tag{VII.8}$$

They transform under the full group G_k as

$$\mathbf{U}(u,b,c)\,\psi_\beta^{\xi\eta}(f)=e^{-2i\pi\beta u}\psi_\beta^{\xi e^u+b,\,\eta e^{ku}+c}(f). \tag{VII.9}$$

In phase space Γ, the curves invariant by action (VII.5) restricted to $G_{\xi\eta}$ are given by

$$(t-\xi)f-k\eta f^k=\tilde{\beta}. \tag{VII.10}$$

A local study of $\psi_\beta^{\xi\eta}(f)$ analogous to that performed in Sec. II C leads to the identification

$$\tilde{\beta}=\beta. \tag{VII.11}$$

As in Sec. II, we have achieved a correspondence between invariant structures in $\mathcal{H}$ and Γ. But we now have one extra parameter labeling the vectors $\psi_\beta^{\xi\eta}(f)$ and the curves (VII.10). To proceed with the construction, we must fix one of these parameters in such a way that the affine covariance of the procedure is preserved. Because of the transformation law (VII.9), the only possible choices are either $\eta=0$ or $\beta=\beta_0$ fixed. The case $\eta=0$ yields the calculus just developed in the preceding sections. The only possibility to obtain a new calculus is to fix the value of β and to let η free.

The same geometric construction as in Sec. III will be carried out. Consider an operator **A** on $\mathcal{H}$ and form its diagonal matrix elements on $\psi_{\beta_0}^{\xi\eta}(f)$:

$$I_{\mathcal{H}}^{\beta_0}(\xi,\eta)=(\psi_{\beta_0}^{\xi\eta},\mathbf{A}\psi_{\beta_0}^{\xi\eta}). \tag{VII.12}$$

Consider a function $\mathcal{A}(t,f)$ and form its integral with respect to curves (VII.10):

$$I_{\Gamma}^{\beta_0}(\xi,\eta)=\int_{\Gamma}\mathcal{A}(t,f)\,\delta((t-\xi)f-k\,\eta f^k-\beta_0)dt\;df. \tag{VII.13}$$

The study of the properties of functions $I_{\mathcal{H}}^{\beta_0}(\xi,\eta)$ and $I_{\Gamma}^{\beta_0}(\xi,\eta)$ follow the same steps as in Sec. III.

1. $G_{\xi\eta}$-invariance of $I_{\mathcal{H}}^{\beta_0}(\xi,\eta)$ and $I_{\Gamma}^{\beta_0}(\xi,\eta)$

It can be readily verified, using (VII.6), (VII.9) and the definitions of $I_{\mathcal{H}}^{\beta_0}(\xi,\eta)$ and $I_{\Gamma}^{\beta_0}(\xi,\eta)$.

2. G_k-covariance

If $\mathbf{A}'=\mathbf{U}_k^{-1}(u,b,c)\mathbf{A}\mathbf{U}_k(u,b,c)$ is the operator transformed from **A** by the group representation (VII.4), it follows from (VII.9) that

$$I_{\mathcal{H}}^{\beta_0}(\mathbf{A}';\xi,\eta)=I_{\mathcal{H}}^{\beta_0}(\mathbf{A};e^u\xi+b,e^{ku}\eta+c). \tag{VII.14}$$

On the other hand, if $\mathcal{A}'(t,f)=\mathcal{A}(e^u t+b+kce^{-(k-1)u}f^{k-1},a^{-1}f)$, a direct computation leads to

$$I_{\Gamma}^{\beta_0}(\mathcal{A}';\xi,\eta)=I_{\Gamma}^{\beta_0}(\mathcal{A};e^u\xi+b,e^{ku}\eta+c). \tag{VII.15}$$

3. Possibility of reconstruction of the operator **A** and of the function $\mathcal{A}(t,f)$

In spite of the fact that the functions $\psi_{\beta_0}^{\xi\eta}(f)$ do not form a basis, it can be shown that the operator **A** is completely characterized by its diagonal elements (VII.12).

The reconstruction of the function $\mathcal{A}(t,f)$ could be performed by inverting the Radon transform (VII.13).

These properties allow us to base the correspondence rule on the identification

$$I_{\mathcal{H}}^{\beta_0}(\xi,\eta)=I_{\Gamma}^{\beta_0}(\xi,\eta), \tag{VII.16}$$

which can be seen as an extension of (III.14). The resulting correspondence rule has the form

$$\mathcal{A}(t,f)=f^{2r+2}\int_{\mathbb{R}}e^{2i\pi u\beta_0}e^{2i\pi(tf-\beta_0)(\lambda_k(u)-\lambda_k(-u))}A(f\lambda_k(u),f\lambda_k(-u))(\lambda_k(u)\lambda_k(-u))^{r+1}\,du\,, \tag{VII.17}$$

where the functions $\lambda_k(u)$ are defined by

$$\lambda_k(u)=e^{u/2}\left(k\frac{\sinh(u/2)}{\sinh(ku/2)}\right)^{1/(k-1)}. \tag{VII.18}$$

Conversely, the kernel of the operator **A** can be expressed in terms of the symbol by

$$A(f_1,f_2)=\int e^{-2i\pi(f_1-f_2)(t-(\beta_0/f))}\delta(f_1-f_2-k^{-1}f^{1-k}(f_1^k-f_2^k))\times\left(\frac{f_1}{f_2}\right)^{-2i\pi\beta_0}(f_1f_2)^{-r}f^{-k}|f_2^{k-1}-f_1^{k-1}|\mathcal{A}(t,f)dt\;df\,. \tag{VII.19}$$

The operator **A** corresponding to this kernel can be written in Weyl's form as

$$\mathbf{A}=\int_{\mathbb{R}^2}\hat{\mathcal{A}}(k,\beta_0;u,v)e^{-2i\pi(u\boldsymbol{\beta}+v\boldsymbol{f})}\,du\,dv, \tag{VII.20}$$

where

$$\hat{\mathcal{A}}(k,\beta_0;u,v)=e^u\left|\frac{1-e^{(k-1)u}}{(1-k)u}\right|(\lambda_k(-u))^{k+1}e^{2i\pi\beta_0(u-\lambda_k(u)+\lambda_k(-u))}$$
$$\times\int_\Gamma e^{2i\pi tf(\lambda_k(u)-\lambda_k(-u))}e^{2i\pi f(v/u)\lambda_k(-u)(e^u-1)}\mathcal{A}(t,f)f\,dt\,df. \tag{VII.21}$$

It may be noticed that the limit for $k=0$ of function $\lambda_k(u)$ is identical to $\lambda(u)$ defined in (IV.15). In that case, the relation (VII.21) becomes a Fourier transform with respect to $\beta=tf$ and f and the rule (IV.20)–(IV.21) is recovered. In the general case, the Fourier transform (IV.21) is replaced by the k-dependent transform (VII.21) for which Parseval's formula does not hold. As a consequence, the unitarity property (III.17) cannot be obtained.

However, it is possible to recover an expression for the scalar product of operators in terms of symbols. This is accomplished by associating two symbols with any operator **A**: the symbol $\mathcal{A}(t,f)$ defined by (VII.17) and a dual symbol $\mathcal{A}_d(t,f)$ defined by

$$\mathcal{A}_d(t,f)=f^{2r+2}\int_{\mathbb{R}}e^{2i\pi u\beta_0}e^{2i\pi(tf-\beta_0)(\lambda_k(u)-\lambda_k(-u))}A(f\lambda_k(u),f\lambda_k(-u))\mu_d(u)du, \tag{VII.22}$$

where

$$\mu_d(u)=(\lambda_k(u)\lambda_k(-u))^{r+1}\left|\frac{d}{du}(\lambda_k(u)-\lambda_k(-u))\right|. \tag{VII.23}$$

It can be observed that it is only in the case $k=0$ that

$$\frac{d}{du}(\lambda_k(u)-\lambda_k(-u))=1, \tag{VII.24}$$

leading to

$$\mathcal{A}(t,f)=\mathcal{A}_d(t,f). \tag{VII.25}$$

For an arbitrary value of k, the unitarity formula is replaced by

$$\mathrm{Tr}(\mathbf{A}\mathbf{B}^\dagger)=\int_\Gamma\mathcal{A}(t,f)\mathcal{B}_d^*(t,f)dt\,df, \tag{VII.26}$$

where $\mathcal{A}(t,f)$ is the symbol of **A** and $\mathcal{B}_d(t,f)$ is the dual symbol of **B**.

These different correspondence rules may be of interest for the study of special operators. The case $k=-1$ where

$$\lambda_{-1}(u)=e^{u/2} \tag{VII.27}$$

stands out. The symbolic calculus corresponding to $k=-1$ and $\beta_0=0$ has been introduced directly by Unterberger[21] who applied it in a mathematical context.

A Wigner function $\mathcal{P}_k^{\beta_0}(t,f)$ can be associated with each k-calculus. Because of the form of relation (VII.26), the Wigner function of signal $S(f)$ is defined as the dual symbol of the projector on the signal. We thus associate a time-frequency distribution with each value of (k,β_0). When $\beta_0=0$, these distributions coincide with the generalized passive distributions introduced in Ref. 16. The case $k=1$, which has not been treated explicitly here, can be deduced by continuity.

VIII. CONCLUSION

A correspondence rule between functions on the time-frequency half-plane and operators on the Hilbert space of positive-frequency signals has been obtained. The rule, which is said to be geometric, is entirely based on the study of the affine group and more specially on the representations of its subgroups in the two domains. It is obtained, relatively to each subgroup, by identification of the decomposition in invariant subspaces with the corresponding tomographic decomposition in the time-frequency half-plane. The construction is very natural and the result can be considered as an affine version of the Weyl rule introduced in quantum mechanics.

In the correspondence, a Lie algebra of Hermitian operators is transformed into a Lie algebra of symbols, also called star algebra. The basic operations of this algebra are the star product and the star bracket of symbols. Formulas for these operations have been given. Special attention has been paid to the case where the star bracket reduces to a Poisson bracket.

The obtained symbolic calculus has been integrated into a family of calculi covariant by three-parameter groups G_k, $k \in \mathbb{R}$, containing the affine group. In this family, the geometric rule corresponds to the special value $k=0$ and is the only one to ensure the unitarity property.

ACKNOWLEDGMENT

The authors are grateful to D. Sternheimer for his constructive remarks on the manuscript.

[1] D. Gabor, ''Theory of communication,'' J. Inst. Electr. Eng. **93**, 429–457 (1946).
[2] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, ''Quantum mechanics as a deformation of classical mechanics,'' Lett. Math. Phys. **1**, 521–530 (1977); ''Deformation theory and quantization, I—Deformation of symplectic structures,'' Ann. Phys. (N.Y.) **111**, 61–110 (1978); ''Deformation theory and quantization, II—Physical applications,'' *ibid.* **111**, 111–151 (1978).
[3] N. Ya. Vilenkin, *Special Functions and the Theory of Group Representations* (American Mathematical Society, Providence, RI, 1968).
[4] A. Unterberger, ''Calcul de Weyl construit sur une autre relation de commutation que celle d'Heisenberg,'' Report, Université de Reims, France, 1984.
[5] D. Arnal and J.C. Cortet, ''Représentations star des groupes exponentiels,'' J. Funct. Anal. **92**, 103–135 (1985).
[6] J-B. Kammerer, ''Analysis of the Moyal product in a flat space,'' J. Math. Phys. **27**, 529–535 (1986).
[7] J.M. Maillard, ''On the twisted convolution product and the Weyl transform of tempered distributions,'' J. Geom. Phys. **3**, 231–261 (1986).
[8] J. Bertrand and P. Bertrand, ''Représentations temps-fréquence des signaux,'' C.R. Acad. Sc. Paris **299**, 635–638 (1984); ''Time-frequency representation of wide-band signals,'' Rech. Aerosp. **1985-5**, 277–283 (1985) (English and French versions).
[9] G. Boudreaux-Bartels, ''Mixed time-frequency signal transformations,'' in *The Transforms and Applications Handbook*, edited by A. Poularikas (CRC Press, Boca Raton, 1996).
[10] P. Flandrin, *Temps-Fréquence* (Hermès, Paris, 1993).
[11] L. Cohen, *Time-Frequency Analysis* (Prentice–Hall, Englewood Cliffs, NJ, 1995).
[12] A. A. Kirillov, *Elements of the Theory of Representations* (Springer-Verlag, Berlin, 1976).
[13] R. G. Shenoy and T. W. Parks, ''Wide-band ambiguity functions and affine Wigner distributions,'' Signal Process. **41**, 339–363 (1995).
[14] D. Arnal, J. C. Cortet, P. Molin, and G. Pinczon, ''Covariance and geometrical invariance in star-quantization,'' J. Math. Phys. **24**, 276–283 (1983).
[15] D. Sternheimer, ''Phase-space representations,'' in *Applications of Group Theory in Physics and Mathematical Physics, Lectures in Applied Mathematics*, edited by M. Flato, P. Sally, and G. Zuckerman (American Mathematical Society, Providence, 1985), pp. 255–267.
[16] J. Bertrand and P. Bertrand, ''A class of affine Wigner functions with extended covariance properties,'' J. Math. Phys. **33**, 2515–2527 (1992).
[17] D. Arnal, M. Cahen, and J. Ludwig, ''Lie groups whose coadjoint orbits are of dimension smaller or equal to two,'' Lett. Math. Phys. **33**, 183–186 (1995).
[18] See, for example, E. Cartan, Ann. Mat. Pura Appl. **11**, 17 (1932); N. Bourbaki, *Groupes et Algèbres de Lie* (Hermann, Paris, 1960), Chap. I, p. 133; N. Jacobson, *Lie algebras* (Interscience, New York, 1962). A detailed list is given in J. Patera, R. T. Sharp, P. Winternitz, and H. Zassenhaus, J. Math. Phys. **17**, 986 (1976).
[19] C. Fronsdal, ''Invariant $\star$-product quantization of the one-dimensional Kepler problem,'' J. Math. Phys. **20**, 2226–2232 (1979).
[20] A. Unterberger and J. Unterberger, ''La série discrète de $SL(2,R)$ et les opérateurs pseudo-différentiels sur une demi-droite,'' Ann. Sci. Ecole Norm. Sup. **17**, 83 (1984); ''Quantification et analyse pseudo-différentielle,'' *ibid.* **21**, 133–158 (1988).
[21] A. Unterberger, ''The calculus of pseudo-differential operators of Fuchs type,'' Comm. Partial Diff. Eqns. **9**, 1179–1236 (1984).

Wavelet analysis of signals with gaps

Peter Frick
Institute of Continuous Media Mechanics, Korolev 1, 614061, Perm, Russia

Alex Grossmann
Centre de Physique Théorique, CNRS, Luminy case 907, 13288 Marseille, France

Philippe Tchamitchian
LATP, CNRS (UMR 6632) et Faculté des Sciences et Techniques de Saint-Jérôme, 13397 Marseille Cedex 20, France

(Received 1 January 1998; accepted for publication 9 April 1998)

A recently introduced algorithm [Frick *et al.*, Astrophys. J. **483**, 426 (1997)] of spectral analysis of data with gaps via a modified continuous wavelet transform is developed and studied. This algorithm is based on a family of functions called ″gapped wavelets″ which fulfill the admissibility condition on the gapped support. The wavelet family is characterized by an additional parameter which should be calculated for every scale and position. Three theorems concerning the properties of gapped wavelet transform are formulated and proved. They affirm the global stability of the algorithm as well as its stability in both limits of large and small scales. These results are illustrated by some numerical examples, which show that the algorithm really attenuates the artifacts coming from gaps (and/or boundaries), and is particularly efficient at small and large scales. © *1998 American Institute of Physics.* [S0022-2488(98)01208-0]

I. INTRODUCTION

Finding periodicities in observed quantities has been a most important enterprise for a very long time. There are, however, many fields for which the data come with inevitable gaps due to the existence of time (or space) intervals at which measurements cannot be made. A prominent example is astronomy, where the gaps are caused by seasonal windows of observation, cloudy skies, telescope maintenance, etc. An illustrative example can be seen in Fig. 1.

The causes of the interruptions are in general not related to the quantities under study, and their effects are felt in several ways. The most direct one is, of course, the loss of useful information. A more insidious effect is the possible contamination of data with spurious periodicities arising from the spectral properties of the set of gaps.

A straightforward ''filling of the holes,'' either by interpolation or by arbitrary extension can have quite undesirable properties. A typical result of interpolation will be the smoothing of the signal, leading to the loss of information in high frequencies.

Fourier transform techniques have been adapted to such a situation: Welsh estimators[1] are perhaps the most commonly used. An extension of Fourier transform to an uneven data set is known in astronomy as the Lomb-Scargle periodogram.[2,3] The idea of this technique is to correct the basic functions $\cos(\omega t)$ and $\sin(\omega t)$ by a phase shift and a mean value subtraction to preserve their normalization conditions on a given set of observations.

In some instances, it is desirable to dispose of time-scale transforms. To this end, Foster[4] recently introduced an analogous algorithm for wavelet transform on an irregular data set. He considers the Morlet wavelets and proposes to reorthogonalize the three basic functions (the real and the imaginary parts of Morlet wavelet and a constant) by rotating the matrix of their scalar products.

We study here another technique for analyzing signals with gaps, introduced in Refs. 5,6 in connection with the spectral study of stellar chromospheric activity. Two examples of data records taken from Ref. 7 are given in Fig. 1. They show that the whole time of observations includes only few periods of main oscillation and that the total length of the gaps is comparable with the length of the record. The main idea of the proposed procedure is to consider the wavelet transform not as

 © 1998 American Institute of Physics

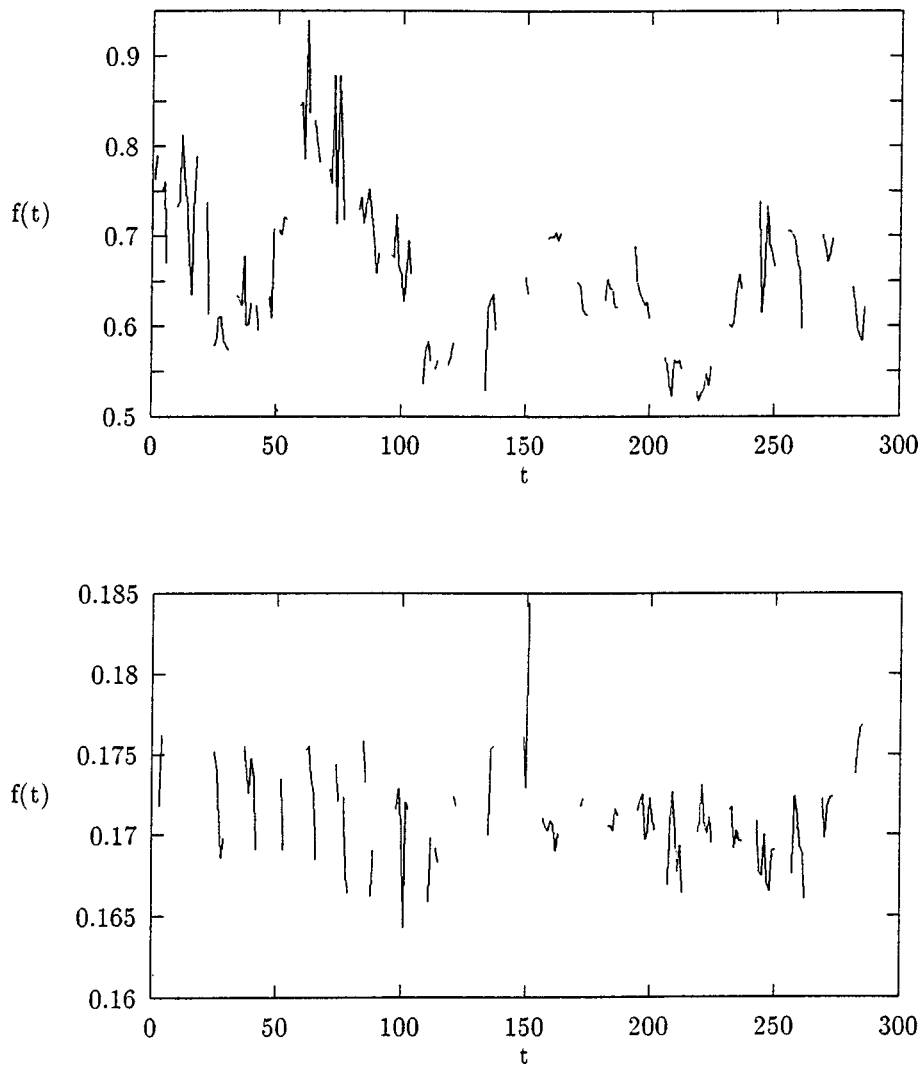

FIG. 1. Two typical examples of data with gaps which concern the stellar chromospheric activity of stars near the lower main sequence at the Mount Wilson Observatory. The figures are taken from Ref. 7.

a convolution of a signal with gaps and a given analyzing wavelet, but as a convolution of a signal and of a wavelet with gaps, thus transferring the gap problem from the unknown signal function to the known wavelet function. Punching holes in a wavelet will of course destroy its useful properties, and one tries to correct the broken wavelet by making it at least admissible.

The algorithm under consideration has been called in Refs. 5 and 6 ''adaptive wavelets'' to emphasize that the analyzing functions are modified by the gaps in the data. This name does not appear to be fortunate because of the extensive use of the word ''adaptive'' in wavelet analysis. This is why we shall write ''the gapped wavelet method'' or ''wavelets with gaps.''

The paper is organized as follows. In Sec. II the definition of wavelets with gaps is given. Some theorems concerning their properties are formulated. The possible developments of the method are also discussed. The properties of wavelet transform with gapped wavelets are illustrated by some examples in Sec. III. The proofs of theorems are given in Sec. IV.

II. DEFINITIONS AND MAIN STATEMENTS

For the sake of simplicity, we limit our study to the one-dimensional setting. The interested reader can verify that all the foregoing statements extend straightforwardly to the n-dimensional case.

Let $\Omega \subset \mathbb{R}$ be an open set. We want to develop a wavelet analysis for functions supported on Ω, which depends as little as possible on Ω. It turns out that the naive approach consisting in performing a standard wavelet analysis of f is not satisfactory. See Sec. III. The reason is the following. Let ψ be the mother wavelet, and denote as usual

$$\psi_{a,b}(x) = \frac{1}{a}\psi\left(\frac{x-b}{a}\right). \tag{1}$$

It is well known that ψ must satisfy the so-called *admissibility condition*, which essentially amounts to demand that the first moment of ψ vanishes. Obviously, when this condition is fulfilled, so it is for all the wavelets $\psi_{a,b}$.

Analyzing a given f supported in Ω means computing the coefficients $\langle f, \psi_{a,b}\rangle$. Hence, even if f is flat on Ω (for example, $f=\chi_\Omega$), this analysis will give significant coefficients for those values of b and a such that $\int_\Omega \psi_{a,b} \neq 0$. This is in contradiction with the very nature of wavelet transform. This is why it is desirable to modify the wavelets so as to enforce an adapted admissibility condition.

A. Gapped wavelets

We start by choosing a mother wavelet ψ satisfying the admissibility condition. We introduce next a non-negative envelope function ϕ such that $|\psi| \leqslant \phi$. Throughout the paper we assume ψ and ϕ to be rapidly decreasing at infinity, as well as their derivative. We define $\phi_{a,b}$ as in (1).

1. The family $\theta_{a,b}$

If $a>0$ and $b \in \mathbb{R}$ we define

$$\theta_{a,b} = \psi_{a,b}\chi_\Omega - C(a,b)\phi_{a,b}\chi_\Omega, \tag{2}$$

where

$$C(a,b) = \begin{cases} \dfrac{\int_\Omega \psi_{a,b}}{\int_\Omega \phi_{a,b}} & \text{if} \quad \displaystyle\int_\Omega \phi_{a,b} \neq 0, \\ 0 \quad \text{otherwise.} \end{cases} \tag{3}$$

This ensures that

$$\int_{\mathbb{R}^n} \theta_{a,b} = 0. \tag{4}$$

Let us note that as $|\psi| \leqslant \phi$, we have

$$|C(a,b)| \leqslant 1 \tag{5}$$

for all a,b. We emphasize that the analyzing function $\theta_{a.b}$ changes for every scale a and every position b if only a gap falls in its vicinity. Thus, strictly speaking, the family $\theta_{a,b}$ is not a family of wavelets because its self-similarity is broken.

2. The family $\theta^{(1)}_{a,b}$

This construction can be generalized by requiring the modified wavelets to have vanishing moments of higher orders. This can be achieved by introducing, for a given integer k, the functions

$$\theta^{(k)}_{a,b}(t) = \psi_{a,b}(t)\chi_\Omega(t) - \left(C_0(a,b) + C_1(a,b)\left(\frac{t-b}{a}\right) + \ldots + C_k(a,b)\left(\frac{t-b}{a}\right)^k\right)\phi_{a,b}(t)\chi_\Omega(t). \tag{6}$$

The constants C_j are readily obtained from the corresponding conditions

$$0 \leqslant j \leqslant k, \quad \int_\Omega \theta_{a,b}^{(k)}(t) \left(\frac{t-b}{a} \right)^j dt = 0. \tag{7}$$

For example, for $k=1$ one finds

$$C_0(a,b) = \frac{\int_\Omega \psi_{a,b} \int_\Omega ((t-b)/a)^2 \phi_{a,b} - \int_\Omega ((t-b)/a) \psi_{a,b} \int_\Omega ((t-b)/a) \phi_{a,b}}{\int_\Omega \phi_{a,b} \int_\Omega ((t-b)/a)^2 \phi_{a,b} - (\int_\Omega ((t-b)/a) \phi_{a,b})^2}, \tag{8}$$

$$C_1(a,b) = \frac{\int_\Omega \phi_{a,b} \int_\Omega ((t-b)/a) \psi_{a,b} - \int_\Omega \psi_{a,b} \int_\Omega ((t-b)/a) \phi_{a,b}}{\int_\Omega \phi_{a,b} \int_\Omega ((t-b)/a)^2 \phi_{a,b} - (\int_\Omega ((t-b)/a) \phi_{a,b})^2}. \tag{9}$$

Let us note that C_0 and C_1 are no longer necessarily bounded by 1. In numerical calculations, described below, we checked the values of these coefficients. They really can become larger than 1, but this only happens for large values of a, such that the (effective) support of the wavelet is at least of the same size as Ω. However, in such a case, the physical meaning of our analysis becomes doubtful.

B. Gapped wavelet transform: Definition and properties

In this subsection, we only consider the case $k=0$.

Definition 1: The gapped wavelet transform is the linear map which associates to any $f \in L^2(\Omega)$ *the set of scalars* $\langle f, \theta_{a,b} \rangle$, $b \in \mathbb{R}$, $a>0$.

As explained before, we may view f as the restriction to Ω of some square-integrable function globally defined on $\mathbb{R}$. Since only the values of f on Ω are relevant, there is no harm in not distinguishing between the various extensions of f in $\mathbb{R}$. For simplicity we will denote all of them by the same letter f. Therefore we have by construction

$$\langle f, \theta_{a,b} \rangle = \langle f \chi_\Omega, \theta_{a,b} \rangle. \tag{10}$$

Our first result states that this map is continuous from $L^2(\Omega)$ to $L^2(\mathbb{R} \times \mathbb{R}^+, (db\ da)/a)$, as for the usual wavelet transform. This is a stability result, in which the most remarkable point is that the stability constant A_0 (see the theorem below) does not depend on Ω. In other words, whatever the shape of Ω, the stability of the gapped wavelet transform is uniformly controlled.

Theorem 2: *There exists a constant* A_0, *independent of* Ω, *such that*

$$\int_0^{+\infty} \int_{\mathbb{R}} |\langle f, \theta_{a,b} \rangle|^2 \frac{da\ db}{a} \leqslant A_0 \int_\Omega |f|^2. \tag{11}$$

When Ω is of finite measure, we may assume that $m_\Omega f$, the mean value of f over Ω, vanishes, since $\int_\Omega \theta_{a,b} = 0$. Therefore, inequality (11) improves in

$$\int_0^{+\infty} \int_{\mathbb{R}} |\langle f, \theta_{a,b} \rangle|^2 \frac{da\ db}{a} \leqslant A_0 \int_\Omega |f - m_\Omega f|^2. \tag{12}$$

We now examine to what extent our gapped wavelet transform carries out some time-scale information. Let $(S_a)_{a>0}$ be a Littlewood–Paley approximation of identity, defined by a function Φ such that $\hat{\Phi}(\omega) = 1$ if $|\omega| \leqslant 1$, $\hat{\Phi}(\omega) = 0$ if $|\omega| \geqslant 2$. Denoting by Φ_a the function $(1/a)\, \Phi(./a)$ we write

$$S_a f = f * \Phi_a. \tag{13}$$

By construction, $S_a f$ is a smoothed version of f, at scale a. Since $\hat{\Phi}_a(\omega) = 1$ when $|\omega| \leqslant (1/a)$ and

$$\widehat{S_a f}(\omega) = \hat{f}(\omega) \hat{\Phi}_a(\omega), \tag{14}$$

the spectral content of f in the interval $|\omega| \leqslant 1/a$ is preserved under the action of S_a. Also, since $\hat{\Phi}_a(\omega)=0$ when $|\omega| \geqslant 2/a$, the spectral content of f in the domain $|\omega| \geqslant 2/a$ is totally erased.

Here of course, the constants 1 and 2 which we use when defining $\hat{\Phi}$ are arbitrary, and may be replaced by any other choice without changing the foregoing statements.

The next theorem describes how the coefficients $\langle f, \theta_{a,b} \rangle$, for fixed $a>0$, attenuate the low-frequency content of f.

Theorem 3: *There exists a constant $A_1>0$, independent of Ω, such that*

$$\int_{\mathbb{R}^n} |\langle S_{a'} f, \theta_{a,b} \rangle|^2 db \leqslant A_1 \left(\frac{a}{a'} \right)^2 \int_{\mathbb{R}} |S_{a'} f|^2, \tag{15}$$

for all $a' \geqslant a > 0$.

This means that $\langle f, \theta_{a,b} \rangle$ filters in f the frequencies lower in modulus than $1/a$. The exponent 2 in the term $(a/a')^2$ is directly linked to the vanishing of the first moment of the gapped wavelets. When using the functions $\theta_{a,b}^{(k)}$ for $k \geqslant 1$, one gets a term $(a/a')^{2k+2}$, provided the constants $C_i(a,b)$, $0 \leqslant i \leqslant k$, in (6) are uniformly controlled. This fact may be proved by a straightforward adaptation of the proof of the theorem (see Sec. IV below).

The situation is more subtle regarding the damping of the high-frequency content of f, because the geometry of Ω must be taken into consideration. This is the reason why we assume a mild regularity property on the boundary $\partial\Omega$ of Ω, quantified by an exponent $\alpha>0$, and defined as follows (for the sake of completeness, the definition is given in arbitrary dimension n).

Definition 4: Ω verifies the property (P_α), where $\alpha \in]0,1]$, when there exists a constant A_Ω, such that for all $b \in \mathbb{R}^n$ and for all $\rho, \varepsilon > 0$, $\varepsilon \leqslant \rho$ one has

$$\min \left\{ \begin{array}{l} |\{y \in B(b,\rho) \cap \Omega ; d(y, \partial\Omega) \leqslant \varepsilon\}| \\ |\{y \in B(b,\rho) \cap \Omega^c ; d(y, \partial\Omega) \leqslant \varepsilon\}| \end{array} \right\} \leqslant A_\Omega \left(\frac{\varepsilon}{\rho} \right)^\alpha |B(b,\rho)|. \tag{16}$$

Here $B(b,\rho)$ is the Euclidean ball of center b and radius ρ, while $|E|$ stands for the Lebesgue measure of E.

Such a property means that, at each scale ρ, the local density in Ω or in Ω^c of the strip around $\partial\Omega$ of width ε is uniformly controlled by $(\varepsilon/\rho)^\alpha$.

For example, if Ω is a finite union of open intervals in dimension 1, or if Ω is a regular domain (at least Lipschitz) in several dimensions, then (P_1) is true. Also, most classical examples of sets with fractal boundary, such as von Koch curve, do fulfill (P_α) for some $\alpha>0$.

Let us go back to our gapped wavelet transform, for an open set Ω fulfilling (P_α). The next theorem states that α describes the damping of the high-frequency content of f in the coefficients $\langle f, \theta_{a,b} \rangle$ for fixed $a>0$.

Theorem 5: *For any $\alpha \in]0,1]$, there exists a constant A_2, independent of Ω, such that if $\partial\Omega$ verifies the property (P_α)*

$$\int_{\mathbb{R}} |\langle f - S_{a'} f, \theta_{a,b} \rangle|^2 \, db \leqslant A_2 (1 + A_\Omega) \left(\frac{a'}{a} \right)^\alpha \int_{\mathbb{R}} |f - S_{a'} f|^2, \tag{17}$$

for all $a \geqslant a' > 0$.

The proof will show that this result does not rely on the cancellation properties of the gapped wavelets and thus is also valid for standard wavelet transforms.

III. ILLUSTRATIONS

The three theorems formulated above show that no dramatical changes (which lead to the growth of the spectral energy) are expected in the spectra of the signal calculated using the gapped wavelets $\theta_{a,b}$ and that the density of spectral energy which reflects the artifacts coming from the boundaries and gaps should be attenuated (at different rates) in both high and low frequencies.

No exact results have been obtained concerning the functions $\theta_{a,b}^{(1)}$, but in numerical simulations we used them as well as the functions $\theta_{a,b}$. We denote by $T0$ the results of the transform using the functions $\theta_{a,b}$, which are only admissible (the zeroth order moment vanishes), and as

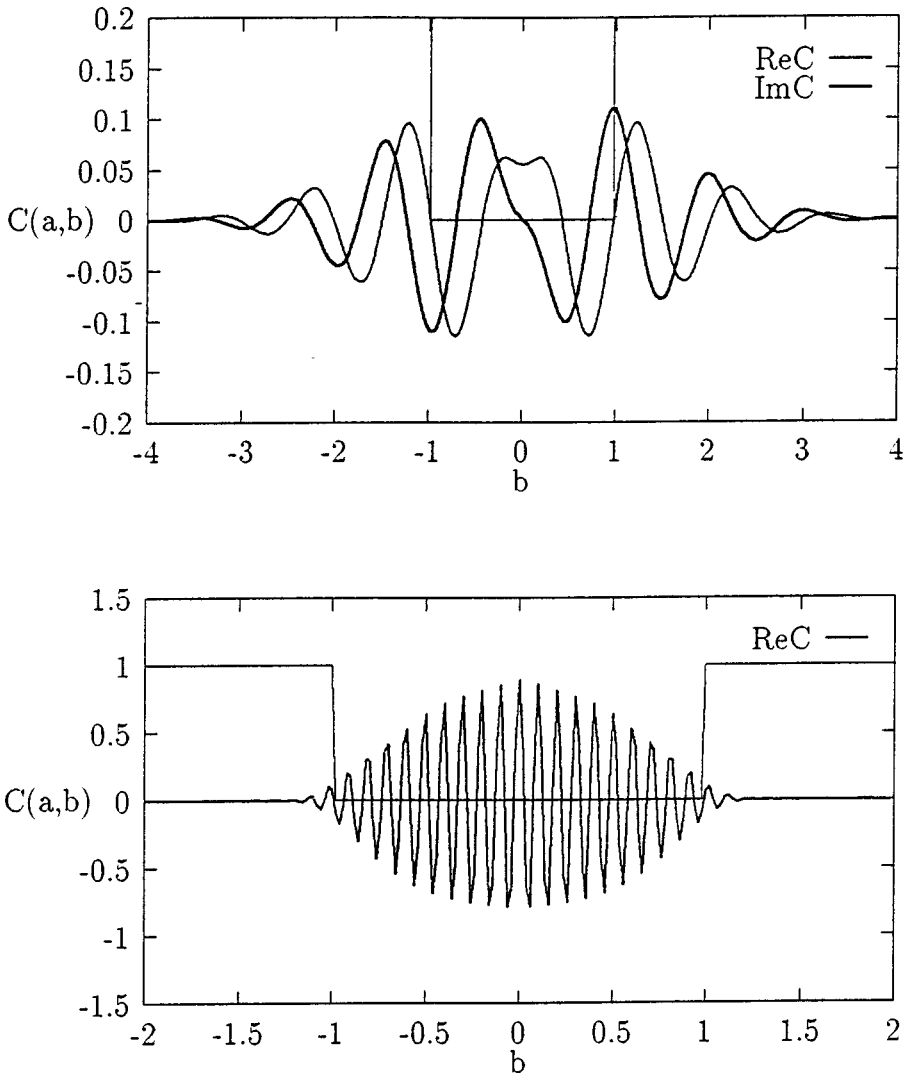

FIG. 2. Variations of parameter $C(a,b)$ in the vicinity of a gap $[-1,1]$ (the characteristic function is shown by the solid line). The real and the imaginary parts of $C(a,b)$ for $a=1$ are shown in the upper panel; the real part of $C(a,b)$ for $a=0.1$ is presented in the lower panel.

$T1$ the results of the transform using the functions $\theta_{a,b}^{(1)}$ for which the first order moment vanishes as well as the zeroth. We compare both with the standard wavelet transform with Morlet wavelet (denoted in the figures as M). Let us emphasize that in the latter case (taken as the reference) we also do not use any interpolation of signals in the gap, and we do not try to eliminate the mean value, the trend, etc.

In the examples to be discussed below, ψ is the Morlet wavelet

$$\psi(t)=e^{i\omega t}e^{-t^2/2} \tag{18}$$

with appropriate ω.

We show in Fig. 2 how the parameter $C(a,b)$ varies when the function $\theta_{a,b}$ passes through an isolated gap $[-1,1]$. In the upper panel of Fig. 2 is presented the graph for $a=1$. Both real and imaginary parts are shown. Notice that the modulus of C does not surpass the value 0.15. In the lower panel of the same figure we present the variations of the real part of $C(a,b)$ for $a=0.1$ (the imaginary one is not shown due to the high density of lines). Now, the maxima of C approach to one in the central part of the gap. Obviously, these large values of $C(a,b)$ correspond to strong

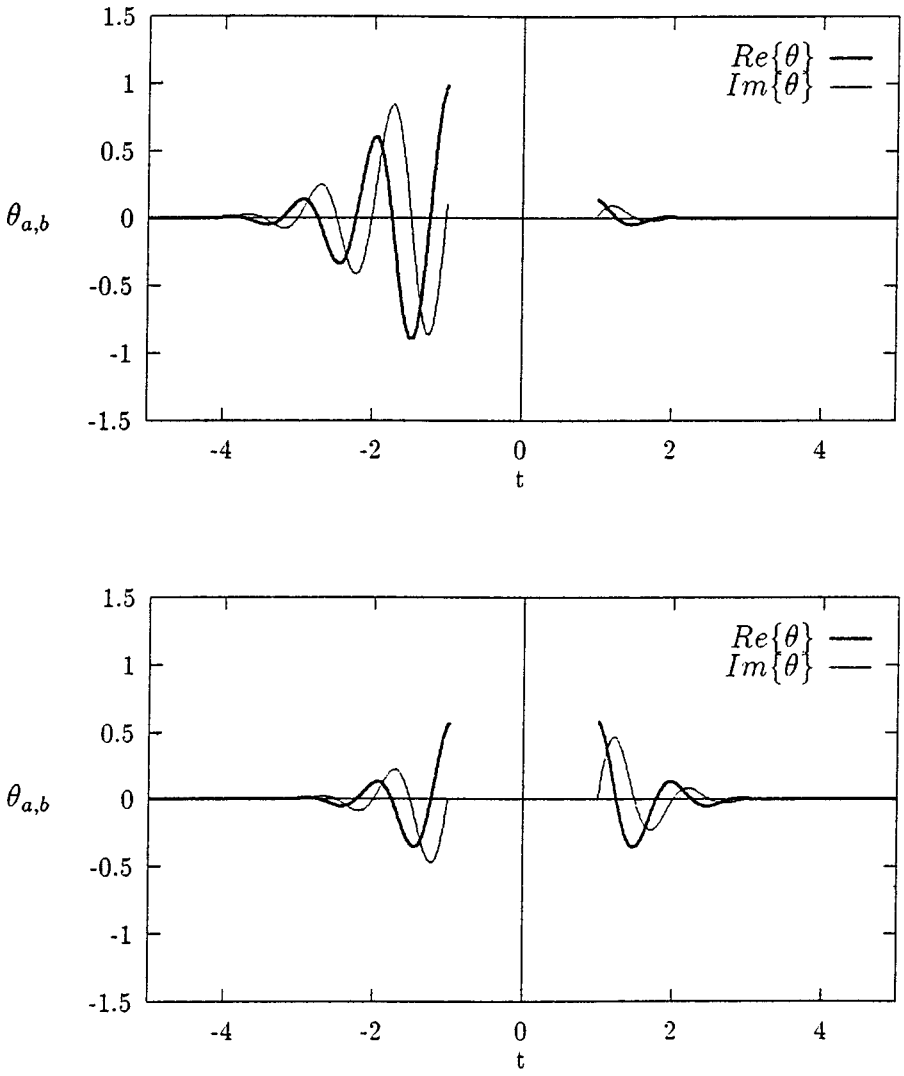

FIG. 3. Gapped wavelet $\theta_{a,b}(t)$ for $a=1$ when going through the gap $[-1,1]$: $b=-1$ (upper panel), $b=0$ (lower panel).

deformations of the analyzing wavelet, provided by the fact that only a far part of exponentially decreasing tail of the wavelet reaches the boundary of the gap. But this also implies that these strongly deformed functions do not give a considerable contribution.

Two examples of gapped wavelets $\theta_{a,b}(t)$ are shown on Fig. 3. We consider the same isolated gap $[-1,1]$ and the scale $a=1$. The functions $\theta_{a,b}(t)$ are plotted for two positions: $b=-1$, where the center of the wavelet is on the boundary of the gaps (upper panel of Fig. 3) and $b=0$, where the wavelet is centered in the gap. One can see that the analyzing wavelet does not undergo drastic changes in the domain lying away from the gap. The discontinuities of the gapped wavelet are on the boundary of the gap and do not introduce additional contributions in the wavelet transform.

We start by analyzing a function with a simple (gaussian) spectrum. The function $f(t)=1+t+\exp(-t^2/2)\cos(2\pi t)$ is a superposition of a Morlet wavelet and a linear trend. The function is given only in a single interval (window) $[-0.6,2.6]$ (Fig. 4.). The true spectrum of this function (the mean value of $|W(a,b)|^2$ for given a) is also shown on the same figure. One can see that the standard algorithm M gives a spectrum that increases with scale a, without any well pronounced peak at $a=1$. The algorithm $T0$ clearly displays the top of the Gaussian bump, but gives the wrong maximum at large scales. Algorithm $T1$ efficiently suppress the wrong power in the large scales, and also slightly improves the small scales part of the spectrum.

In Fig. 5 we show the wavelet spectrum for a harmonic signal superposed with a linear function and known on an interval with regular gaps. The basic spectrum M displays two wrong

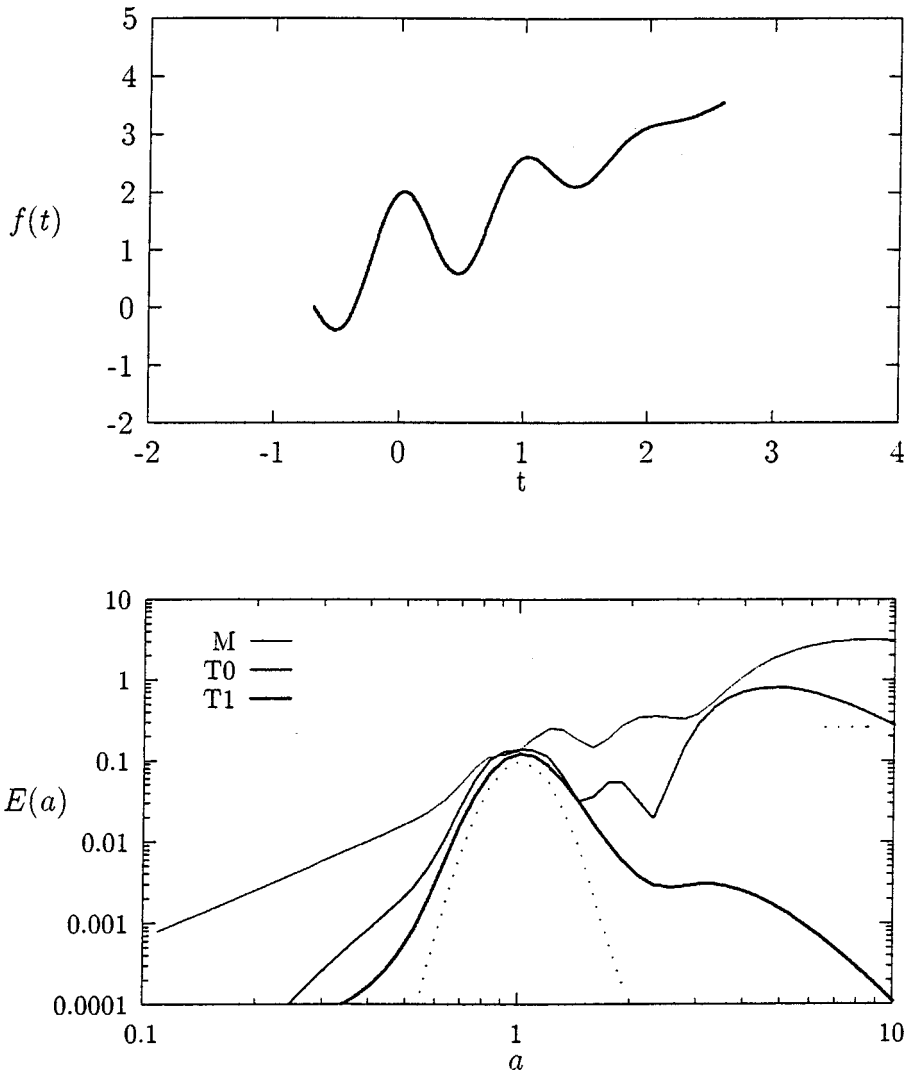

FIG. 4. The function $f(t)=1+t+\exp(-t^2/2)\cos(2\pi t)$ given on the interval $[-0.6,2.6]$ and its wavelet spectra M (line), $T0$(line) and $T1$ (line). The ''ideal'' spectrum is shown by a dotted line.

maxima with orders of magnitude comparable with the main one: in large scales and in a scale smaller than the main one. As indicated by the theorems proved above, the wrong maxima are suppressed in both large and small scales. The transform $T1$ is again essentially more effective in large scales. Let us note that the standard algorithm displays the main maximum at $a=0.74$, $T0$ gives $a=0.76$ and $T1$ shows the maximum at $a=0.79$ (the true value is $a=0.8$ and is marked in the figure by an arrow). This indicates that the gapped wavelet technique not only suppresses the low and high noise frequencies, but is also better at estimating the frequency of the signal.

The next example (Fig. 6) presents a signal with two frequencies in a linear background. In this case the standard technique produces an additional peak in intermediate scales. The gapped algorithm eliminates it. The large scale artifact is again better suppressed by the wavelets $T1$.

The last example (Fig. 7) presents a signal with two harmonic components without any background but with regular gaps. The results show that there is an unavoidable artifact (the left peak, which remains on the same magnitude in all three spectra). At the same time one sees that the standard techniques absolutely lost the high frequency, which is drowned in noise, in spite of being well observed by the naked eye. In the gapped wavelet method spectra this frequency gives a pronounced peak.

Surely, there are a lot of possibilities to improve the results of wavelet transform of signals

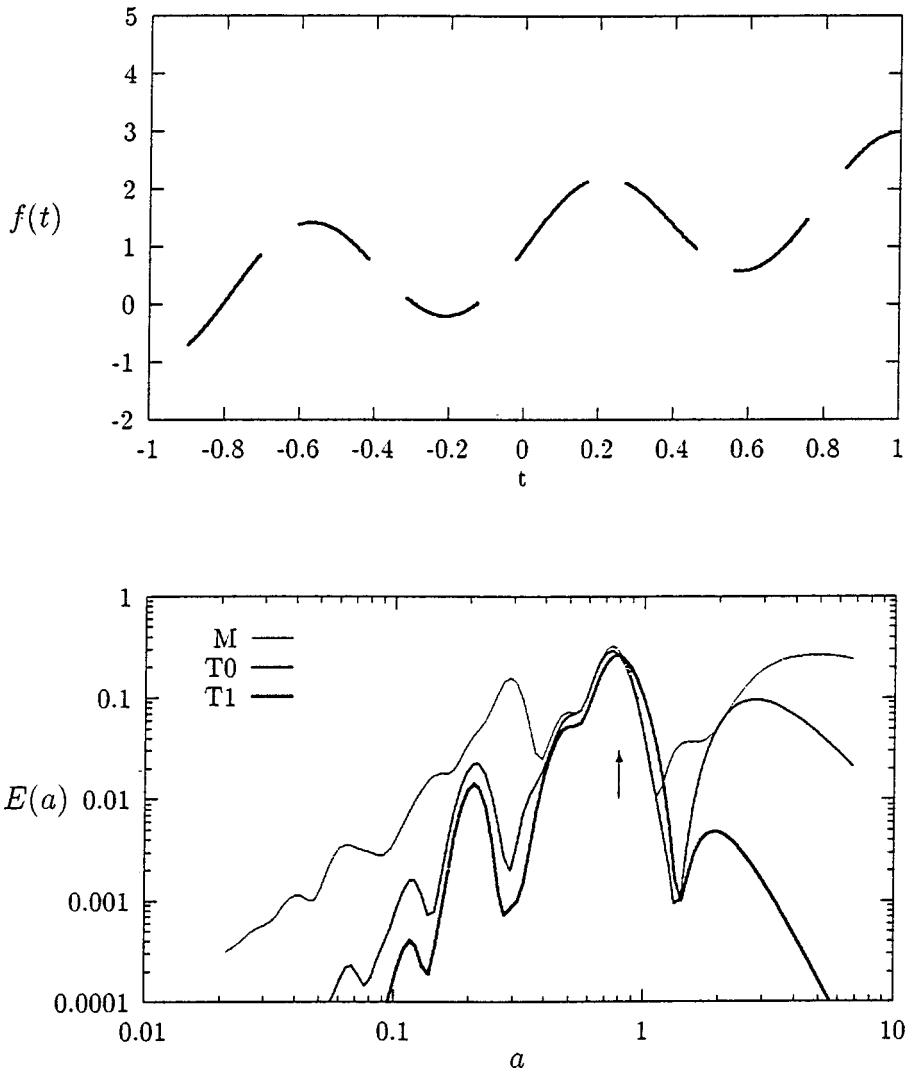

FIG. 5. The analyzed signal $f(t)=1+t+\sin((4\pi/5)\,t)$ with gaps (above) and its spectra (below), obtained by standard Morlet wavelet (M) and gapped wavelets θ ($T0$) and $\theta^{(1)}$ ($T1$). The true frequency is marked by the arrow.

with gaps, especially in a concrete situation when the expected character of the signal is known *a priori*.

Our examples illustrate the proved properties of gapped wavelets $\theta_{a,b}$ and show that the wavelets $\theta^{(1)}_{a,b}$ provide further improvement of spectral resolution of the wavelet transform.

IV. PROOFS

In what follows, the letter A denotes a constant, which may change from line to line.

A. Proof of Theorem 2

This proof is based on classical estimates in the theory of singular integral operators. One can show the following result:

Lemma 6: The matrix $(\langle \theta_{a,b}, \theta_{a',b'} \rangle)$, where $a,a'>0$, $b,b' \in \mathbb{R}$, *defines a continuous operator in* $L^2(\mathbb{R}^2_+, db\ da/a)$.

In this statement, the word ''matrix'' is used by analogy with the standard case where the parameters a,b belong to a discrete, or even finite, set.

Lemma 6 implies Theorem 2. Indeed, if we define an operator T by

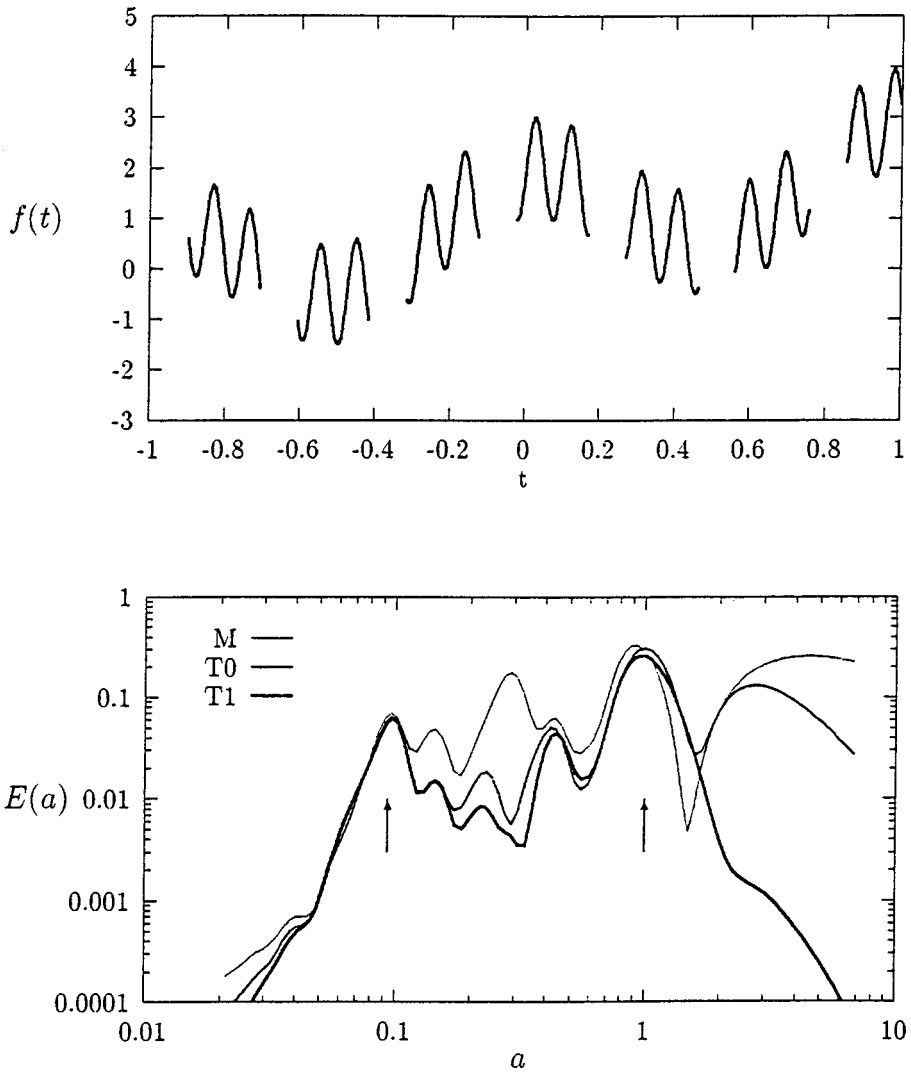

FIG. 6. The analyzed signal $f(t)=t+c+\cos(2\pi t)+\sin(21\pi t)$ with gaps and its spectra. Both frequencies are marked in the scales by arrows.

$$Tf=\int\int \langle f,\theta_{a,b}\rangle \theta_{a,b}\frac{da\ db}{a} \tag{19}$$

then we have

$$\langle Tf,f\rangle=\int\int |\langle f,\theta_{a,b}\rangle|^2\frac{da\ db}{a} \tag{20}$$

and

$$\langle Tf,Tf\rangle=\int\int\int\int \langle f,\theta_{a,b}\rangle\langle \theta_{a,b},\theta_{a',b'}\rangle\langle f,\theta_{a',b'}\rangle\frac{da\ db}{a}\frac{da'db'}{a'}. \tag{21}$$

Applying Lemma 6 to the r.h.s. of (21), we get

$$\langle Tf,Tf\rangle\leqslant A_0\langle Tf,f\rangle \tag{22}$$

which implies continuity of T with $\|T\|\leqslant A_0$. Here A_0 is the norm of the operator defined by $(\langle \theta_{a,b},\theta_{a',b'}\rangle)$. We have to show that it can be bounded uniformly in Ω.

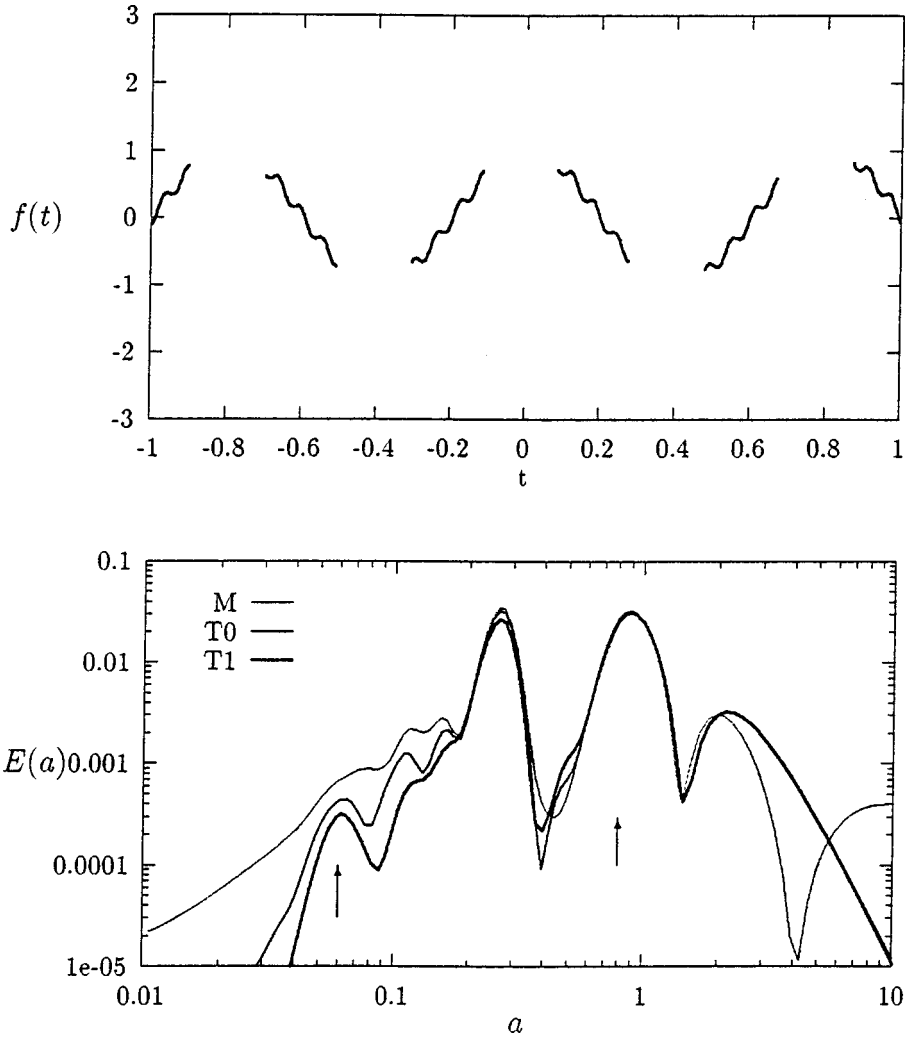

FIG. 7. The analyzed signal $f(t)=\cos(5\pi t/2)+0.1\,\sin(33\pi t)$ with gaps and its spectra.

To do so, we shall show that there exists a constant A, independent on Ω, such that

$$\int\int |\langle \theta_{a,b}, \theta_{a',b'}\rangle| \frac{da'\ db'}{a'} \leqslant A \tag{23}$$

for all $a>0$, $b\in\mathbb{R}$. This inequality implies that the matrix $(\langle \theta_{a,b}, \theta_{a',b'}\rangle)$ defines a continuous operator on $L^{\infty}(\mathbb{R}^2_+, da\ db/a)$. By symmetry one gets the continuity of the same operator on $L^1(\mathbb{R}^2_+, da\ db/a)$ and, by interpolation, on $L^2(\mathbb{R}^2_+, da\ db/a)$. In addition, the corresponding operator norms are controlled by the constant A of (23), which does not depend on Ω.

The proof of (23) depends on the following estimates. By assumption, there exists A such that the wavelet ψ and the function ϕ satisfy the estimates

$$|\psi(x)|+\phi(x)\leqslant A(1+|x|)^{-3}, \tag{24}$$

$$|\psi(x+h)-\psi(x)|+|\phi(x+h)-\phi(x)|\leqslant A|h|(1+|x|)^{-3}, \tag{25}$$

if $x,h\in\mathbb{R}$ and $|h|\leqslant |x|/2$. We set

$$\omega(x)=(1+|x|)^{-3}. \tag{26}$$

Lemma 7: There exists a constant A, independent of Ω, *such that*

$$|\langle \theta_{a,b}, \theta_{a',b'} \rangle| \leqslant A \frac{a'}{a^2} \omega\left(\frac{b'-b}{a} \right) \tag{27}$$

for all $a \geqslant a' > 0$ *and* $b, b' \in \mathbb{R}$.

If $a \geqslant a'$

$$\langle \theta_{a,b}, \theta_{a',b'} \rangle = \int [\psi_{a,b}(x) - C(a,b)\phi_{a,b}(x)] \overline{\theta_{a',b'}(x)} \tag{28}$$

as $G^2 = G$. Since the first moment of $\theta_{a',b'}$ is zero, one can write

$$\langle \theta_{a,b}, \theta_{a',b'} \rangle = \int [\psi_{a,b}(x) - \psi_{a,b}(b')] \overline{\theta_{a',b'}(x)} dx - C(a,b) \int [\phi_{a,b}(x) - \phi_{a,b}(b')] \overline{\theta_{a',b'}(x)} dx. \tag{29}$$

The first term on the r.h.s. is controlled by

$$A \int \frac{|x-b'|}{a^2} \left[\omega\left(\frac{x-b}{a} \right) + \omega\left(\frac{b'-b}{a} \right) \right] \frac{1}{a'} \omega\left(\frac{x-b'}{a'} \right) dx \tag{30}$$

by (5), (24), and (25). This integral can be evaluated as

$$A \frac{a'}{a^2} \omega\left(\frac{b-b'}{a} \right). \tag{31}$$

The second term on the r.h.s. of (29) is controlled in the same way. This completes the proof of Lemma 7.

One then deduces (23) by writing

$$\begin{aligned} &\int \int |\langle \theta_{a,b}, \theta_{a',b'} \rangle| \frac{db'\,da'}{a'} \\ &\leqslant A \int \int_{a' \geqslant a} \frac{a}{a'^2} \omega\left(\frac{b-b'}{a'} \right) \frac{db'\,da'}{a'} + A \int \int_{a' \leqslant a} \frac{a'}{a^2} \omega\left(\frac{b-b'}{a} \right) \frac{db'\ da'}{a'} \\ &\leqslant A \int_a^{+\infty} \frac{a}{a'} \frac{da'}{a'} + A \int_0^a \frac{a'}{a} \frac{da'}{a'} = 2A. \end{aligned} \tag{32}$$

Thus inequality (23) is proved and the proof of theorem 2 is completed.

B. Proof of Theorem 3

Theorem 3 is a consequence of estimates which are close analogs of the preceding ones. Denoting by $\tilde{\Phi}$ the function

$$\tilde{\Phi}(x) = 2\Phi(2x) \tag{33}$$

one has by construction $\tilde{\Phi} * \Phi = \Phi$. From this one gets the identity

$$\tilde{S}_a S_a = S_a, \tag{34}$$

where $\tilde{S}_a$ is the convolution operator with $\tilde{\Phi}_a$. Thus, if $0 < a \leqslant a'$, then

$$\langle S_{a'} f, \theta_{a,b} \rangle = \langle S_{a'} f, \tilde{S}_{a'} \theta_{a,b} \rangle. \tag{35}$$

Inequality (15) is a simple consequence of

Lemma 8: If $0<a\leqslant a'$ *then*

$$|\tilde{S}_{a'}\theta_{a,b}(x)|\leqslant A\frac{a}{a'^2}\omega\left(\frac{x-b}{a'}\right), \tag{36}$$

where A is independent of Ω.

We leave it to the reader to verify that (35) and (36) imply (15) and only explain why (36) is true. We write

$$\tilde{S}_{a'}\theta_{a,b}(x)=\int \tilde{\Phi}_{a'}(x-y)\,\theta_{a,b}(y)dy=\int [\tilde{\Phi}_{a'}(x-y)-\tilde{\Phi}_{a'}(x-b)]\theta_{a,b}(y)dy. \tag{37}$$

Thus one gets as in the proof of (27)

$$|\tilde{S}_{a'}\theta_{a,b}(x)|\leqslant A\int \frac{|y-b|}{a'^2}\left[\omega\left(\frac{x-y}{a'}\right)+\omega\left(\frac{x-b}{a'}\right)\right]\frac{1}{a}\omega\left(\frac{y-b}{a}\right)dy\leqslant A\frac{a}{a'^2}\omega\left(\frac{x-b}{a'}\right). \tag{38}$$

C. Proof of Theorem 5

Let Ψ be an even C^∞ function, such that Supp $\Psi\subset\{\omega\in\mathbb{R};1/4\leqslant|\omega|\leqslant 1\}$, and

$$\int_0^{+\infty}\hat{\Psi}(\tau\omega)\frac{d\tau}{\tau}=1 \tag{39}$$

for all $\omega\neq 0$.

Then, if Δ_τ is the convolution operator by Ψ_τ, where $\Psi_\tau(t)=(1/\tau)\,\Psi(t/\tau)$, the above identity means that for any $f\in L^2(\mathbb{R})$ one has

$$f=\int_0^{+\infty}\Delta_\tau f\frac{d\tau}{\tau}. \tag{40}$$

By construction, one has

$$\Delta_\tau(I-S_{a'})=0 \tag{41}$$

if $\tau\geqslant a'$. It follows that

$$(I-S_{a'})f=\int_0^{a'}\Delta_\tau(I-S_{a'})f\frac{d\tau}{\tau}. \tag{42}$$

This identity enables us to deduce Theorem 5 from the inequality

$$\int_{\mathbb{R}}|\langle\Delta_\tau f,\theta_{a,b}\rangle|^2db\leqslant A(1+A_\Omega)\left(\frac{\tau}{a}\right)^\alpha\int_{\mathbb{R}}|\Delta_\tau f|^2, \tag{43}$$

where $0<\tau\leqslant a$.

Indeed, if $a'\leqslant a$ one applies (42) and then the Minkowski inequality and (43). That gives

$$\left(\int_{\mathbb{R}}|\langle f-S_{a'}f,\theta_{a,b}\rangle|^2db\right)^{1/2}\leqslant A(1+A_\Omega)^{1/2}\int_0^{a'}\left(\frac{\tau}{a}\right)^{\alpha/2}\left(\int_{\mathbb{R}}|\Delta_\tau(I-S_{a'})f|^2\right)^{1/2}\frac{d\tau}{\tau}. \tag{44}$$

Using the Cauchy–Schwarz inequality one gets

$$\int_{\mathbb{R}}|\langle f-S_{a'}f,\theta_{a,b}\rangle|^2db\leqslant A(1+A_\Omega)\left(\int_0^{a'}\left(\frac{\tau}{a}\right)^\alpha\frac{d\tau}{\tau}\right)\left(\int_0^{+\infty}\|\Delta_\tau(I-S_{a'})f\|_2^2\frac{d\tau}{\tau}\right). \tag{45}$$

To deduce inequality (17) one computes

$$\int_0^{+\infty} \|\Delta_\tau (I - S_{a'}) f\|_2^2 \frac{d\tau}{\tau} = A \|(I - S_{a'}) f\|_2^2, \tag{46}$$

and

$$\int_0^{a'} \left(\frac{\tau}{a}\right)^\alpha \frac{d\tau}{\tau} = \frac{1}{\alpha}\left(\frac{a'}{a}\right)^\alpha. \tag{47}$$

It remains to prove (43), by using the geometric hypothesis (P_α). One starts by taking a function $\tilde{\Psi}$, even and C^∞, such that $\hat{\tilde{\Psi}}(\omega) = 1$ if $\omega \in \text{Supp } \hat{\Psi}$, with

$$\text{Supp } \hat{\tilde{\Psi}} \subset \left\{ \omega \in \mathbb{R}^n; \tfrac{1}{8} \leqslant |\omega| \leqslant 2 \right\}. \tag{48}$$

Denoting as usual by $\tilde{\Delta}_\tau$ the convolution operator by $\tilde{\Psi}_\tau$ one gets

$$\tilde{\Delta}_\tau \Delta_\tau = \Delta_\tau, \tag{49}$$

from which one obtains

$$\langle \Delta_\tau f, \theta_{a,b} \rangle = \langle \Delta_\tau f, \tilde{\Delta}_\tau \theta_{a,b} \rangle. \tag{50}$$

Lemma 9: There exists a constant A, independent of Ω, *such that, if* $\tau \leqslant a$, *then for any* $x \in \mathbb{R}$

$$\int_{\mathbb{R}} |\tilde{\Delta}_\tau \theta_{a,b}(x)| db \leqslant A, \tag{51}$$

and for any $b \in \mathbb{R}$

$$\int_{\mathbb{R}} |\tilde{\Delta}_\tau \theta_{a,b}(x)| dx \leqslant A(1 + A_\Omega)\left(\frac{\tau}{a}\right)^\alpha. \tag{52}$$

Inequality (51) means that the operator associating to any function g the function

$$b \mapsto \langle g, \tilde{\Delta}_\tau \theta_{a,b} \rangle \tag{53}$$

(where a and τ are two fixed parameters with $\tau \leqslant a$), is uniformly bounded on $L^1(\mathbb{R})$. Inequality (52) means that the same operator is bounded on $L^\infty(\mathbb{R})$ and that its norm is controlled by

$$A(1 + A_\Omega)\left(\frac{\tau}{a}\right)^\alpha. \tag{54}$$

Inequality (43) follows by interpolation.

Now, let us prove (51). It is enough to write

$$\tilde{\Delta}_\tau \theta_{a,b}(x) = \int_{\mathbb{R}} \tilde{\Psi}_\tau (x - y) \theta_{a,b}(y) dy, \tag{55}$$

and to note that by (5) for any y

$$\int_{\mathbb{R}} |\theta_{a,b}(y)| db \leqslant \|\psi\|_1 + \|\phi\|_1. \tag{56}$$

This yields (51).

The proof of (52) is more elaborate. We write

$$\tilde{\Delta}_\tau \theta_{a,b}(x) = \int_{\mathbb{R}} \tilde{\Psi}_\tau(x-y) G(y) [\psi_{a,b}(y) - C(a,b)\phi_{a,b}(y)] dy. \tag{57}$$

Let us start with

$$\int_{\mathbb{R}} \tilde{\Psi}_\tau(x-y) G(y) \phi_{a,b}(y) dy = \int_{\mathbb{R}} \tilde{\Psi}_\tau(x-y) G(y) [\phi_{a,b}(y) - \phi_{a,b}(x)] dy + \tilde{\Delta}_\tau G(x) \phi_{a,b}(x). \tag{58}$$

The first term on the right is controlled by

$$A \frac{\tau}{a^2} \omega\left(\frac{x-b}{a}\right), \tag{59}$$

as we have seen several times. Its integral with respect to x is dominated by $A\tau/a$.

The hypothesis (P_α) is used for the second term.

Lemma 10: There exists an absolute constant A such that

$$\left| \int_{\mathbb{R}} \tilde{\Delta}_\tau G(x) \phi_{a,b}(x) dx \right| \leqslant A A_\Omega \left(\frac{\tau}{a}\right)^\alpha, \tag{60}$$

for all $\tau \leqslant a$, $b \in \mathbb{R}$.

Admit this result for the moment. Then one has

$$\int_{\mathbb{R}} \left| \int_{\mathbb{R}} \tilde{\Psi}_\tau(x-y) G(y) \phi_{a,b}(y) dy \right| dx \leqslant A(1+A_\Omega) \left(\frac{\tau}{a}\right)^\alpha. \tag{61}$$

By the same argument, one obtains

$$\int_{\mathbb{R}} \left| \int_{\mathbb{R}} \tilde{\Psi}_\tau(x-y) G(y) \psi_{a,b}(y) dy \right| dx \leqslant A(1+A_\Omega) \left(\frac{\tau}{a}\right)^\alpha. \tag{62}$$

Lemma 9 is thus completely proved using (58) and (5) once more.

It remains to prove Lemma 10. There exist two rapidly decreasing sequences of numbers $(c_k)_{k\in\mathbb{Z}}$ and $(d_l)_{l\in\mathbb{Z}}$ and two sequences of functions $(\phi_k)_{k\in\mathbb{Z}}$ and $(\tilde{\Psi}_l)_{l\in\mathbb{Z}}$ satisfying the following properties:

$$\begin{cases} \phi(x) = \sum_{k\in\mathbb{Z}} c_k \phi_k(x), & \tilde{\Psi}(x) = \sum_{l\in\mathbb{Z}} d_l \tilde{\Psi}_l(x), \\ \text{Supp } \phi_k \subset [k-1, k+1], & \text{Supp } \tilde{\Psi}_l \subset [l-1, l+1], \\ \|\phi_k\|_\infty = \|\tilde{\Psi}_l\|_\infty = 1, & \\ \int_{\mathbb{R}} \tilde{\Psi}_l = 0. & \end{cases} \tag{63}$$

One constructs these objects with the help of a partition of the identity which is used to decompose ϕ on one hand, and the primitive of $\tilde{\Psi}$ vanishing at infinity on the other hand. This gives directly the c_k, ϕ_k, and by differentiation the d_l, $\tilde{\Psi}_l$. The sequences (c_k) and (d_l) are rapidly decreasing because so are ϕ and the primitive of $\tilde{\Psi}$, since $\int \tilde{\Psi} = 0$. Verifying the other properties in (63) is straightforward.

One can now write

$$\tilde{\Delta}_\tau G(x)\phi_{a,b}(x)=\sum_k \sum_l c_k d_l \int_{\mathbb{R}} \frac{1}{\tau}\tilde{\Psi}_l\left(\frac{x-y}{\tau}\right) G(y)dy \frac{1}{a}\phi_k\left(\frac{x-b}{a}\right). \tag{64}$$

Let us examine the integral

$$I_{kl}=\int_{\mathbb{R}}\left|\int_{\mathbb{R}}\frac{1}{\tau}\tilde{\Psi}_l\left(\frac{x-y}{\tau}\right)G(y)dy\right|\frac{1}{a}\phi_k\left(\frac{x-b}{a}\right)dx, \tag{65}$$

(we may suppose ϕ_k non-negative.) By support considerations, we may assume that $y\in\Omega$, $(x-y)/\tau\in[k-1,k+1]$ and $(x-b)/a\in[l-1,l+1]$. Moreover, since $\int\tilde{\Psi}_l=0$, the integral w.r.t. y does not vanish only when Supp $\tilde{\Psi}_l((x-)/\tau$ intersects Ω^c, which implies $d(y,\partial\Omega)\leqslant 2\tau$. We notice that

$$\left\{(x,y);\frac{x-b}{a}\in[l-1,l+1],\frac{x-y}{\tau}\in[k-1,k+1],y\in\Omega,d(y,\partial\Omega)\leqslant 2\tau\right\}\subset\mathscr{E}_{kl}, \tag{66}$$

where

$$\mathscr{E}_{kl}=\{(x,y);x\in[y+(k-1)\tau,y+(k+1)\tau],$$
$$y\in\Omega\cap[b-k\tau+a(l-2),b-k\tau+a(l+2)],d(y,\partial\Omega)\leqslant 2\tau\}. \tag{67}$$

We thus have, using (63)

$$I_{kl}\leqslant\frac{1}{\tau a}|\mathscr{E}_{kl}|\leqslant\frac{2}{a}|\mathscr{E}'_{kl}|, \tag{68}$$

where

$$\mathscr{E}'_{kl}=\{y\in\Omega\cap[b-k\tau+a(l-2),b-k\tau+a(l+2)],d(y,\partial\Omega)\leqslant 2\tau\}. \tag{69}$$

Using once more $\int\tilde{\Psi}_l=0$, we replace G by $1-G$ in I_{kl} without changing anything, which amounts to replacing Ω by Ω^c. Hence we also have

$$I_{kl}\leqslant\frac{2}{a}|\mathscr{E}''_{kl}|, \tag{70}$$

where

$$\mathscr{E}''_{kl}=\{y\in\Omega^c\cap[b-k\tau+a(l-2),b-k\tau+a(l+2)],d(y,\partial\Omega)\leqslant 2\tau\}. \tag{71}$$

By (P_α), we have

$$I_{kl}\leqslant 8A_\Omega\left(\frac{\tau}{a}\right)^\alpha. \tag{72}$$

Summing over $k,l\in\mathbb{Z}$ achieves the proof of Lemma 10, and the theorem follows.

V. FURTHER CONSIDERATIONS

The gapped wavelet transform is a simple alternative to the well-known construction of wavelet bases on an interval or a domain (e.g., Ref. 8), which could also be used to define spectral estimators (see Ref. 9).

It should be noted that the computation of the coefficients $\langle f,\theta_{a,b}\rangle$ basically involves standard wavelet and scaling function coefficients of the signal and of the characteristic function of its support. It is therefore possible in principle to use a pyramidal algorithm, following the ideas of Refs. 10,11.

Finally, let us point out that the existence of a reconstruction formula similar to the usual one is an open problem.

ACKNOWLEDGMENTS

Discussions with B. Torrésani have contributed very much to this paper. In particular, the idea of introducing the family $\theta^{(1)}_{a,b}$ is due to him. P.F. acknowledges partial financial support from CRDF under Grant 176100 and from RFBR under Grant 96-02-16252a and thanks CPT of University of Marseille–Luminy for hospitality.

[1] D. R. Brillinger *Time Series; Data Analysis and Theory* (Holden-Day, San Francisco, 1981).
[2] N. Lomb, ApSS **39**, 447 (1976).
[3] L. D. Scargle, Astrophys. J. **263**, 875 (1982).
[4] G. Foster, Astron. J. **112**, 1709 (1996).
[5] P. Frick, S. Baliunas, D. Galyagin, D. Sokoloff, and W. Soon, Astrophys. J. **483**, 426 (1997).
[6] D. Galyagin and P. Frick, ''Adaptive wavelets,'' in *Mathematical Modelling of Systems and Processes* (Perm Technical University Press, Perm, 1996), Vol. 4, p. 20.
[7] S. L. Baliunas, R. A. Donahue, W. H. Soon *et al.*, Astrophys. J. **438**, 269 (1995).
[8] A. Cohen, I. Daubechies, and P. Vial, Appl. Comp. Harmonic Anal. **1**, 54 (1994).
[9] H. Gao, PhD thesis, UC Berkeley, 1993.
[10] P. Abry and A. Aldroubi, Int. J. Fourier Anal. Appl. **2**, 135 (1996).
[11] M. A. Muschietti and B. Torrésani, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **26**, 925 (1995).

Some remarks on the Navier–Stokes equations in $\mathbb{R}^3$

Pierre Gilles Lemarié-Rieusset[a)]
Département de Mathématiques, Université d'Evry, Bd des Coquibus, 91025 Evry Cedex, France

(Received 20 January 1998; accepted for publication 9 April 1998)

We study various existence and uniqueness results for solutions of the Navier–Stokes equations in connection with function spaces related to real harmonic analysis. © *1998 American Institute of Physics.* [S0022-2488(98)01708-3]

I. INTRODUCTION

In this paper, we are mainly interested in reviewing recent results concerning weak solutions of the Navier–Stokes equations which belong to $\mathscr{C}([0,+\infty),L^3(\mathbb{R}^3))$. Such solutions were first studied by Kato[1] in 1984. L^3 estimates have been used as well in the study of Leray's weak solutions, leading to the uniqueness theorem of Sohr and Von Wahl.[2]

A new impetus has been given to the study of such solutions by the recent book of Cannone.[3] Following the idea of Federbush[4] that a good time-frequency analysis of the fluctuation term (see below) could give new existence theorems, Cannone uses the Littlewood–Paley decomposition instead of divergence-free wavelet bases (used in Ref. 4) to provide a precise analysis of the fluctuation. In Ref. 5, we show how this *frequential* analysis of the fluctuation, in terms of a *Besov norm*, could solve the uniqueness problem for solutions in $\mathscr{C}([0,+\infty),L^3(\mathbb{R}^3))$. This frequential approach is highlighted by a result of Le Jan and Sznitman,[6] which shows us how to deal with the Fourier transform of the Navier–Stokes equations. A short time after our result, Meyer proposed a new proof using only *spatial* estimates, in terms of a *Lorentz norm*.[7] We find very striking the fact that there is yet no actual *time-frequency* analysis of the Navier–Stokes equations.

We will show as well the relation between our result and the Sohr and Von Wahl theorem. This implies introducing energy estimates for L^3 solutions, a technique which is useful for proving the existence of weak solutions for an initial value with infinite energy.[8]

Besides, we will pay a few words on the self-similar solutions described by Cannone[3] and some related asymptotic results of Planchon.[9]

II. THE NAVIER–STOKES EQUATIONS IN $\mathbb{R}^3$

The classical *Navier–Stokes equations* describe the motion of a Newtonian fluid; we consider only the case when the fluid is viscous, homogeneous, incompressible and fills the entire space and is not submitted to external forces; then, the equations describing the evolution of the motion $\mathbf{u}(t,x)$ of the fluid element at time t and position x are given by

$$\rho\partial_t\mathbf{u}=\mu\Delta\mathbf{u}-\rho(\mathbf{u}\cdot\nabla)\mathbf{u}-\nabla p, \tag{1}$$

$$\nabla\cdot\mathbf{u}=0. \tag{2}$$

ρ is the (constant) *density* of the fluid, and we may assume with no loss of generality that $\rho=1$. μ is the *viscosity* coefficient, and we may assume as well that $\mu=1$. p is the (unknown) *pressure*, whose action is to maintain the divergence of $\mathbf{u}$ to be 0 (this *divergence free* condition expresses the incompressibility of the fluid). p can be expressed as a function of $\mathbf{u}$, provided that $\mathbf{u}$ vanishes at infinity; indeed, taking the divergence of (1), we obtain

$$\Delta p=-\nabla\cdot(\mathbf{u}\cdot\nabla)\mathbf{u}, \tag{3}$$

a)Electronic mail: lemarie@lami.univ-evry.fr

 © 1998 American Institute of Physics

and this formula allows one to eliminate p in many cases: indeed, formula (3) determines p up to a harmonic function and the requirement that $\mathbf{u}$ remains equal to 0 at infinity allows one to get rid of the contribution this harmonic term could have given to $\partial_t \mathbf{u}$.

Since $\nabla \cdot \mathbf{u} = 0$, the equation (1) can be rewritten as well as

$$\partial_t \mathbf{u} = \Delta \mathbf{u} - \nabla \cdot (\mathbf{u} \otimes \mathbf{u}) - \nabla p, \tag{4}$$

and we shall thus study solutions of (4) and (2) which vanish in a suitable sense at $x = \infty$. Moreover, we complement (4) and (2) (and the boundary condition $\mathbf{u} = 0$ at infinity) by an initial value $\mathbf{u} = \mathbf{u}_0$ at $t = 0$, and we study solutions for $t > 0$.

There is a huge amount of literature on the mathematical theory of the Navier–Stokes equations; classical references are the books by Temam,[10] Ladyzhenskaya[11] or Constantin and Foias;[12] a more recent reference is the book by Lions.[13]

A *classical solution* of (2) and (4) is a time-dependent vector field which has one bounded and continuous time derivative and two bounded and continuous space derivatives. The existence of such solutions was obtained first by Oseen in the beginning of the century.[14] Following the Picard iteration scheme, $\mathbf{u}$ is obtained as a fixed point of the associated integral transform: one considers first the linear problems (P1) and (P2):

$$\nabla \cdot \mathbf{u}_1 = 0, \ \partial_t \mathbf{u}_1 = \Delta \mathbf{u}_1 - \nabla p_1, \ \mathbf{u}_1(0,\cdot) = \mathbf{u}_0, \tag{P1}$$

$$\nabla \cdot \mathbf{u}_2 = 0, \ \partial_t \mathbf{u}_2 = \nabla \mathbf{u}_2 + \mathbf{f} - \nabla p_2, \ \mathbf{u}_2(0,\cdot) = 0. \tag{P2}$$

The first one is easily solved by $\mathbf{u}_1 = e^{t\Delta}\mathbf{u}_0 = W_t * \mathbf{u}_0$ where $W_t(x) = (4\pi t)^{-3/2} e^{-|x|^2/4t}$ is the heat kernel on $\mathbb{R}^3$. The second one is solved by the integral formula: $\mathbf{u}_2 = \int_0^t \mathbb{P} e^{(t-s)\Delta} \mathbf{f}(s)\,ds = \int_0^t O_{t-s} * \mathbf{f}(s)\,ds$ where $\mathbb{P}$, the *Leray–Hopf projection operator* on divergence free vector fields, is formally given as the operator matrix $Id_3 - (1/\Delta)\nabla \otimes \nabla$ and where the *Oseen kernel* $O_t(x) = t^{-3/2} O(x/\sqrt{t})$ is a matrix function given by $O_{k,l}{}^{(x)} = (2\pi)^{-3} \int_{\mathbb{R}^3} [(|\xi|^2 \delta_{k,l} - \xi_k \xi_l)/|\xi|^2] \times e^{-|\xi|^2} e^{ix\cdot\xi} d\xi$. The functions $O_{k,l}$ are analytical on $\mathbb{R}^3$ and satisfy for all $\alpha \in \mathbb{N}^3$ the estimate $|(\partial^{|\alpha|}/\partial x^\alpha) O_{k,l}(x)| \leq C_\alpha (1+|x|)^{-(3+|\alpha|)}$. Then, we seek a solution $\mathbf{u}$ of the Navier–Stokes equations as a solution of the integral equation:

$$\mathbf{u} = e^{t\Delta}\mathbf{u}_0 - \int_0^t (\nabla \otimes O_{t-s}) * (\mathbf{u} \otimes \mathbf{u})\,ds, \tag{5}$$

where the j-th component of $(\nabla \otimes O_{t-s}) * (\mathbf{u} \otimes \mathbf{u})$ is given by $\Sigma_k \Sigma_l [1/(t-s)^2] \times \partial_k O_{j,l}(x/\sqrt{t-s}) * (u_k(s,x) u_l(s,x))$. Following Federbush, we call $e^{t\Delta}\mathbf{u}_0$ the tendency and $\int_0^t (\nabla \otimes O_{t-s}) * (\mathbf{u} \otimes \mathbf{u})\,ds$ the *fluctuation*.

The existence of a classical solution, whatever regular the initial value may be, is guaranteed only for a finite interval of time. If we look for *global solutions*, i.e., solutions defined for every $t > 0$, we have to consider *weak solutions*, which were introduced in 1934 by Leray[15] for $\mathbf{u}_0 \in (L^2)^3$, or *mild solutions*, which were introduced by Kato in the early 60's for the equations on a bounded domain and an initial value in a Sobolev space[16] and in 1984[1] for the equations on the whole space and $\mathbf{u}_0 \in (L^3)^3$. Weak solutions satisfy equations (2) and (4) in the sense of distributions and mild solutions satisfy the integral equation (5). As a matter of fact, the notion of mild solutions refers to an abstract approach of evolution equations in a Banach space, but we shall stick to a more simple definition, because in the whole space the Oseen kernel is quite explicit and so the integral equation can be handled in a more direct way.

III. WEAK AND MILD SOLUTIONS

We begin by defining weak and mild solutions of the Navier–Stokes equations on the whole space.

Definition 1 (weak solutions)*:* A *weak solution* of the Navier–Stokes equations on $(0,T) \times \mathbb{R}^3$ is a distribution vector field $\mathbf{u}(t,x)$ in $(\mathscr{D}'((O,T) \times \mathbb{R}^3)^3$ such that:

(a) $\mathbf{u}$ is locally square integrable on $(0,T)\times\mathbb{R}^3$,
(b) $\nabla\cdot\mathbf{u}=0$,
(c) $\exists p\in\mathscr{D}'((0,T)\times\mathbb{R}^3)\ \partial_t\mathbf{u}=\Delta\mathbf{u}-\nabla\cdot(\mathbf{u}\otimes\mathbf{u})-\nabla p$.

For defining mild solutions, we consider only uniformly locally square integrable vector fields.

Definition 2: (uniformly locally integrable functions)*:*

For $1\leqslant p<\infty$, we define F_p of functions which are uniformly locally in L^p by: $f\in F_p$ if and only if $\sup_{x\in\mathbb{R}^3}\int_{|y-x|\leqslant 1}|f(y)|^p dy<+\infty$.

It is normed by

$$\|f\|_{F_p}=\sup_{x\in\mathbb{R}^3}\left(\int_{|y-x|\leqslant 1}|f(y)|^p dy\right)^{1/p}.$$

Moreover we define E_p as the closure in F_p of $\mathscr{D}(\mathbb{R}^3)$; this is the space of the elements in F_p which vanish at infinity and it is characterized by $\lim_{x\to\infty}\int_{|y-x|\leqslant 1}|f(y)|^p dy=0$.

Definition 3 (mild solutions)*:* A *mild solution* of the Navier–Stokes equations on $(0,T^*)\times\mathbb{R}^3$ with initial value $\mathbf{u}_0\in(\mathscr{S}'(\mathbb{R}^3))^3$, $\nabla\cdot\mathbf{u}_0=0$, is a vector field $\mathbf{u}(t,x)$ in $(\mathscr{D}'((0,T^*)\times\mathbb{R}^3)^3$ such that:
(a) $\mathbf{u}\in\cap_{T<T^*}L^2((0,T),(F_2)^3)$,
(b) $\mathbf{u}=e^{t\Delta}\mathbf{u}_0-\int_0^t(\nabla\otimes O_{t-s})*(\mathbf{u}\otimes\mathbf{u})ds$.

The point is that when $v\in F_1$ then $\partial_j v$ belongs to a Banach space Y of tempered distributions whose norm is shift invariant (so that convolution with functions in L^1 operate boundedly on Y) and on which the Riesz transforms $\partial_j/\sqrt{-\Delta}$ operate boundedly (for instance, Y can be defined as the space of distributions which split in a sum of a low-frequency term in $\dot{B}^{-1,\infty}_{\infty}$ and a high-frequency term in $\dot{B}^{-4,\infty}_{\infty}$). Thus, if $\mathbf{u}\in\cap_{T<T^*}L^2((0,T),(F_2)^3)$ we find that $\mathbf{w}=\int_0^t(\nabla\otimes O_{t-s})*(\mathbf{u}\otimes\mathbf{u})ds$ is well defined and belongs to $\cap_{T<T^*}L^\infty((0,T),(Y)^3)$; moreover it satisfies $\partial_t\mathbf{w}=\Delta\mathbf{w}+\mathbb{P}(\nabla\cdot(\mathbf{u}\otimes\mathbf{u}))$ and $\lim_{t\to 0}\|\mathbf{w}\|_Y=0$. See Ref. 5 for further details.

The two notions of weak and mild solutions are equivalent for solutions which vanish at infinity.[5]

Proposition 1: Let $\mathbf{u}\in\cap_{T<T^*}L^2((0,T),(F_2)^3)$. Then the following assertions are equivalent.

(A) $\mathbf{u}$ satisfies the following requirements:
(a) $\nabla\cdot\mathbf{u}=0$,
(b) $\partial_t\mathbf{u}=\Delta\mathbf{u}-\mathbb{P}(\nabla\cdot(\mathbf{u}\otimes\mathbf{u}))$,
where $\mathbb{P}$ is the Leray–Hopf projection operator on divergence free vector fields:

(B) $\mathbf{u}$ is a mild solution of the Navier–Stokes equations: $\exists\mathbf{u}_0\in(\mathscr{S}'(\mathbb{R}^3))^3$ such that
(a) $\nabla\cdot\mathbf{u}_0=0$,
(b) $\mathbf{u}=e^{t\Delta}\mathbf{u}_0-\int_0^t e^{(t-s)\Delta}\mathbb{P}(\nabla\cdot(\mathbf{u}\otimes\mathbf{u}))ds$.

If we assume that $\mathbf{u}$ vanishes at infinity [i.e., we assume more precisely that $\mathbf{u}\in\cap_{T<T^*}L^2((0,T),(E_2)^3)$] then (A) and (B) are equivalent to the following assertion.

(C) $\mathbf{u}$ is a weak solution of the Navier–Stokes equation:
(a) $\nabla\cdot\mathbf{u}=0$,
(b) $\exists p\in\mathscr{D}'((0,T^*)\times\mathbb{R}^3)\ \partial_t\mathbf{u}=\Delta\mathbf{u}-\nabla\cdot(\mathbf{u}\otimes\mathbf{u})-\nabla p$.

Sketch of the proof: Equivalence between (A) and (B) is a simple exercise on distributions [derivatives and integrals are weakly defined: apply equations to a test function in $\mathscr{C}^\infty_{comp}((0,T^*)\times\mathbb{R}^3)$]. For equivalence between (A) and (C), we have to prove that we may apply $\mathbb{P}$ to (C) and that it gives (A); we may apply $\mathbb{P}$ to high frequency components without any problem: let $\varphi\in\mathscr{D}(\mathbb{R}^3)$ be such that $|\xi|\leqslant 1/2\Rightarrow\varphi(\xi)=1$ and $|\xi|\geqslant 1\Rightarrow\varphi(\xi)=0$ and let S_j be defined as the Fourier multiplier $\mathscr{F}(S_j f)=\varphi(\xi/2^j)\mathscr{F}f$; then $\mathbb{P}(Id-S_j)$ operates boundedly on $\mathscr{S}^3$ and $(\mathscr{S}')^3$; thus (C) gives $\partial_t(Id-S_j)\mathbf{u}=\Delta(Id-S_j)\mathbf{u}-(Id-S_j)\mathbb{P}(\nabla\cdot(\mathbf{u}\otimes\mathbf{u}))$ and we may conclude since for $u\in E_2$ we have $S_j u\to 0$ weakly in $\mathscr{S}'$ when $j\to-\infty$ while $\|S_j\mathbb{P}(\nabla\cdot(\mathbf{u}\otimes\mathbf{u}))\|_\infty\leqslant C2^j\|\mathbf{u}\|^2_{F_2}$. ■

Almost all the solutions we shall consider in this paper are weak/mild solutions in the sense of Proposition 1: for instance, the Leray solutions [which are in $L^\infty((L^2)^3)$], the Kato solutions [which are in $L^\infty((L^p)^3)$ for some $p\in[3,\infty)$], the Cannone self-similar solutions [which satisfy for some $p\in(3,\infty)\ t^{1/2-3/2p}\mathbf{u}\in L^\infty((L^p)^3)$], the Meyer solutions in $L^\infty((L^{3,\infty})^3)$ and so on.

Since a mild solution is defined as a fixed point of an integral transform, one may expect some uniqueness results. The following one is quite obvious.

Definition 4 (regular mild solutions): A mild solution $\mathbf{u}$ of the Navier–Stokes equations on $(0,T^*)\times\mathbb{R}^3$ is called *regular* if it satisfies the following requirements:
(a) $\mathbf{u}\in\cap_{T<T^*}L^\infty((0,T),(F_2)^3)$,
(b) $\mathbf{u}\in\cap_{0<\epsilon<T<T^*}L^\infty((\epsilon,T),(L^\infty)^3)$,
(c) $\lim_{t\to 0}\sqrt{t}\|\mathbf{u}(t,\cdot)\|_\infty=0$.

Proposition 2: Let $\mathbf{u}_0\in(F_2)^3$, $\nabla\cdot\mathbf{u}_0=0$. Then there exists at most one regular mild solution with $\mathbf{u}_0$ as an initial value.

Proof: Define $B(\mathbf{f},\mathbf{g})=\int_0^t(\nabla\otimes O_{t-s})*(\mathbf{f}\otimes\mathbf{g})ds$. Let $\mathbf{u}$ and $\mathbf{v}$ be two regular mild solutions: $\mathbf{u}=e^{t\Delta}\mathbf{u}_0-B(\mathbf{u},\mathbf{u})$ and $\mathbf{v}=e^{t\Delta}\mathbf{u}_0-B(\mathbf{v},\mathbf{v})$ so that $\mathbf{w}=\mathbf{u}-\mathbf{v}$ is solution of the equation $\mathbf{w}=-B(\mathbf{w},\mathbf{v})-B(\mathbf{u},\mathbf{w})$. Now, we use the fact that the norm of F_2 is shift invariant and the fact that $\nabla\otimes O_{t-s}$ is a convolution matrix with functions whose L^1 norm is $O(1/\sqrt{t-s})$ to get

$$\|\mathbf{w}(t,\cdot)\|_{F_2}\leqslant C\int_0^t\frac{1}{\sqrt{t-s}}\frac{1}{\sqrt{s}}\,ds\ \sup_{0<s<t}\|\mathbf{w}(s,\cdot)\|_{F_2}\sup_{0<s<t}\sqrt{s}(\|\mathbf{u}(s,\cdot)\|_\infty+\|\mathbf{v}(s,\cdot)\|_\infty).$$

Since $\int_0^t(1/\sqrt{t-s})(1/\sqrt{s})ds=\pi$, we get local uniqueness [as long as $\pi C\sup_{0<s<t}\sqrt{s}(\|\mathbf{u}(s,\cdot)\|_\infty+\|\mathbf{v}(s,\cdot)\|_\infty)<1$]: indeed, we then have $\sup_{0<s<t}\|\mathbf{w}(s,\cdot)\|_{F_2}\leqslant\pi C\ \sup_{0<s<t}\sqrt{s}(\|\mathbf{u}(s,\cdot)\|_\infty+\|\mathbf{v}(s,\cdot)\|_\infty)\sup_{0<s<t}\|\mathbf{w}(s,\cdot)\|_{F_2}$, which gives $\sup_{0<s<t}\|\mathbf{w}(s,\cdot)\|_{F_2}=0$. Then, global uniqueness follows by looking at the largest t such that $\mathbf{u}=\mathbf{v}$ on $[0,t)$ and applying local uniqueness again on $[t,t+\delta t]$: since F_2 is a dual space and since $\mathbf{u}$ and $\mathbf{v}$ are weakly continuous into $(\mathscr{S}'(\mathbb{R}^3))^3$, we find $\mathbf{u}(t,\cdot)=\mathbf{v}(t,\cdot)\in(F_2)^3$ while $\mathbf{u}$ and $\mathbf{v}$ are mild solutions on (t,T^*) with $\mathbf{u}(t,\cdot)$ as the initial value; moreover, since $\mathbf{u}$ and $\mathbf{v}$ are bounded on (t,T) for all $T<T^*$, they are regular mild solutions. ■

IV. EXISTENCE OF REGULAR MILD SOLUTIONS

The existence of regular mild solutions has been proved for an initial value $\mathbf{u}_0$ in many spaces: Sobolev spaces H^s for $s\geqslant 1/2$,[16] Lebesgue spaces L^p for $p\in[3,\infty)$,[1] Besov spaces, Morrey-Campanato spaces and others. We will focus on the Morrey-Campanato space $M_{2,3}$.

Definition 5 (Morrey-Campanato space): For $1\leqslant p\leqslant q\leqslant\infty$, the Morrey-Campanato space $M_{p,q}$ is defined as the spaces of functions f which are locally in L^p and such that

$$\sup_{x\in\mathbb{R}^3,0<r\leqslant 1}\frac{1}{r^{3(1/p-1/q)}}\left(\int_{|x-y|\leqslant r}|f(y)|^p dy\right)^{1/p}<\infty, \tag{6}$$

where the left hand side of inequality (6) is the norm of f in $M_{p,q}$.

The homogeneous Morrey-Campanato space $\dot{M}_{p,q}$ is defined in the same way, by taking the supremum on all $r\in(0,\infty)$ instead of $r\in(0,1]$.

It is easy to check that $M_{p,p}=F_p$, $F_q\subset M_{p,q}$ and $M_{p,\infty}=L^\infty$. The existence of solutions in Morrey-Campanato spaces has been discussed by several authors (see for instance Refs. 3, 4, 5, 17–19). We shall specialize our attention to two subspaces of $M_{2,3}$.

Theorem 1: (mild solutions in $M_{2,3}$) Let $\mathbf{u}_0\in(M_{2,3})^3$, $\nabla\cdot\mathbf{u}_0=0$. Then we have the following.
(a) $\mathbf{U}=e^{t\Delta}\mathbf{u}_0$ satisfies $\sup_{t>0}\|\mathbf{U}\|_{M_{2,3}}+\sup_{0<t\leqslant 1}\sqrt{t}\|\mathbf{U}\|_\infty<C\|\mathbf{u}_0\|_{M_{2,3}}$.
(b) If $\lim_{t\to 0}\sqrt{t}\|e^{t\Delta}\mathbf{u}_0\|_\infty=0$, then there exists $T^*=T^*(\mathbf{u}_0)>0$ and a regular mild solution $\mathbf{u}$ on $(0,T^*)\times\mathbb{R}^3$ with initial value $\mathbf{u}_0$. Moreover, $\mathbf{u}\in\cap_{T<T^*}L^\infty((0,T),(M_{2,3})^3)$.
(c) There exists two constants $\epsilon_0,\epsilon_1>0$ such that if $\mathbf{u}_0\in(\dot{M}_{2,3})^3$ and $\|\mathbf{u}_0\|_{\dot{M}_{2,3}}<\epsilon_0$ then there exists a mild solution $\mathbf{u}$ on $(0,\infty)\times\mathbb{R}^3$ with initial value $\mathbf{u}_0$ such that $\sup_{t>0}\|\mathbf{u}\|_{\dot{M}_{2,3}}<\infty$ and $\sup_{t>0}\sqrt{t}\|\mathbf{u}\|_\infty<\epsilon_1$. Moreover such a solution is unique.
(d) Under the hypotheses of (c), if $\|\mathbf{u}_0\|_{\dot{M}_{2,3}}<\epsilon_0$ and $\mathbf{u}_0$ is homogeneous ($\forall\lambda>0\,\mathbf{u}_0(\lambda x)=1/\lambda\,\mathbf{u}_0(x)$) then $\mathbf{u}$ is self-similar: $\forall\lambda>0\,\mathbf{u}(\lambda^2 t,\lambda x)=(1/\lambda)\mathbf{u}(t,x)$.

Sketch of the proof: The proof is by now quite classical. It is enough to check that the operator B (defined as $B(\mathbf{f},\mathbf{g})=\int_0^t(\nabla\otimes O_{t-s})*(\mathbf{f}\otimes\mathbf{g})ds$) is bilinear continuous on E and on $\dot{E}$ where

$$E=\{\mathbf{f}\in L^\infty((0,T),(M_{2,3})^3/\ \sup_{0<t<T}\sqrt{t}\|\mathbf{f}\|_\infty<\infty,\lim_{t\to 0}\sqrt{t}\|\mathbf{f}\|_\infty=0\},$$

$$\dot{E}=\{\mathbf{f}\in L^\infty((0,\infty),(\dot{M}_{2,3})^3)/\sup_{0<t}\sqrt{t}\|\mathbf{f}\|_\infty<\infty\}.$$

Indeed, we have the inequalities

$$\|(\nabla\otimes O_{t-s})*(\mathbf{f}\otimes\mathbf{g})\|_{M_{2,3}}\leqslant C\ \frac{1}{\sqrt{t-s}}\ \|(\mathbf{f}\otimes\mathbf{g})\|_{M_{2,3}}\leqslant C\ \frac{1}{\sqrt{t-s}}\ \|\mathbf{f}\|_{M_{2,3}}\|\mathbf{g}\|_\infty$$

and

$$\|(\nabla\otimes O_{t-s})*(\mathbf{f}\otimes\mathbf{g})\|_\infty\leqslant C\inf\left(\frac{\|\mathbf{f}\otimes\mathbf{g}\|_\infty}{\sqrt{t-s}},\ \frac{\|\mathbf{f}\otimes\mathbf{g}\|_{M_{2,3}}}{t-s}\right)\leqslant C\inf\left(\frac{\|\mathbf{f}\|_\infty}{\sqrt{t-s}},\ \frac{\|\mathbf{f}\|_{M_{2,3}}}{t-s}\right)\|g\|_\infty,$$

from which we get $\|B(\mathbf{f},\mathbf{g})\|_E\leqslant C_0\|f\|_E\sup_{0<s<T}\sqrt{s}\|g(s,\cdot)\|_\infty$ where C_0 does not depend on T.

Then a contraction argument shows the existence and uniqueness of the solutions. We show how to prove point (b) [point (c) is shown in a very similar way and point (d) is obvious since the norm of $\dot{M}_{2,3}$ is invariant through the transformation $f(x)\to\lambda f(\lambda x)$]. We first notice that we have

$$\|B(\mathbf{f},\mathbf{f})-B(\mathbf{g},\mathbf{g})(t,x)\|_E\leqslant C_0\ \sup_{0<s<t}\ \|(\mathbf{f}-\mathbf{g})(s,x)\|_E\ \sup_{0<s<t}\ (\sqrt{s}\|\mathbf{f}(s,x)\|_\infty+\sqrt{s}\|\mathbf{g}(s,x)\|_\infty).$$

Moreover we have $\|B(\mathbf{f},\mathbf{f})\|_\infty\leqslant C_1\sup_{0<s<t}\|\mathbf{f}\|_{M_{2,3}}^{1/2}\ \sup_{0<s<t}\|\mathbf{f}\|_\infty^{3/2}$. Thus if $R_0=\|\mathbf{u}_0\|_{M_{2,3}}$, if $\epsilon_0>0$ is chosen such that $C_1\sqrt{2R_0\epsilon_0}<1$ and $2C_0\epsilon_0<1$, and if T^* is chosen such that $\sup_{0<s<T^*}\|e^{s\Delta}\mathbf{u}_0\|_\infty<\epsilon_0$, then the map $\mathbf{f}\to e^{t\Delta}\mathbf{u}_0-B(\mathbf{f},\mathbf{f})$ maps the closed subset of E defined by $\{\mathbf{u}/\sup_{0<s<T^*}\|\mathbf{u}\|_{M_{2,3}}<2R_0$ and $\sup_{0<s<T^*}\|\mathbf{u}\|_\infty<\epsilon_0\}$ into itself and is a contraction on it; thus we have a fixed point in this set, which can be obtained by iterating the map. ■

Remark: The smoothness of regular solutions in Morrey-Campanato spaces is discussed in detail by Kato in Ref. 17 and by Taylor in Ref. 19. There is some difficulty in the fact that these spaces are not separable, hence $e^{t\Delta}$ is not a C_0-semigroup on these spaces.

V. THE LITTLEWOOD–PALEY DECOMPOSITION

In the next section, we shall discuss the theory of Cannone about adapted spaces for the Navier–Stokes equations. It relies on the Littlewood–Paley decomposition of tempered distributions. In this section, we recall briefly some points of the Littlewood–Paley theory.

Definition 6 (Littlewood–Paley decomposition): Let $\varphi\in\mathscr{D}(\mathbb{R}^3)$ be such that $|\xi|\leqslant 1/2\Rightarrow\varphi(\xi)=1$ and $|\xi|\geqslant 1\Rightarrow\varphi(\xi)=0$. Let ψ be defined as $\psi(\xi)=\varphi(\xi/2)-\varphi(\xi)$. Let S_j and Δ_j be defined as the Fourier multipliers $\mathscr{F}(S_jf)=\varphi(\xi/2^j).\mathscr{F}f$ and $\mathscr{F}(\Delta_jf)=\psi(\xi/2^j).\mathscr{F}f$. Then for all $N\in\mathbb{Z}$ and all $f\in\mathscr{S}'(\mathbb{R}^3)$ we have $f=S_Nf+\Sigma_{j\geqslant N}\Delta_jf$ in $\mathscr{S}'(\mathbb{R}^3)$. This equality is called the *Littlewood–Paley decomposition* of the distribution f. If moreover $\lim_{N\to-\infty}S_Nf=0$ in $\mathscr{S}'$, then the equality $f=\Sigma_{j\in\mathbb{Z}}\Delta_jf$ is called the homogeneous Littlewood–Paley decomposition of f.

Δ_jf can be viewed as a filtering of f through a bandpass filter which selects the frequencies around $|\xi|\approx 2^j$.

A very basic tool in using the Littlewood–Paley decomposition is the so-called Bernstein inequalities which express how to control the L^p norm of the derivatives of the dyadic blocks:

$$\forall\alpha\in\mathbb{N}^3,\quad\exists C_\alpha,\quad\forall p\in[1,\infty],\quad\forall f\in L^p,\quad\|\partial^\alpha S_jf\|_p\leqslant C_\alpha 2^{j|\alpha|}\|S_jf\|_p,$$

$$\forall\alpha\in\mathbb{N}^3,\quad\exists C_\alpha,\quad\forall p\in[1,\infty],\quad\forall f\in L^p,\quad\|\partial^\alpha\Delta_jf\|_p\leqslant C_\alpha 2^{j|\alpha|}\|\Delta_jf\|_p,$$

$$\forall\alpha\in\mathbb{R},\quad\exists C_\alpha,\quad\forall p\in[1,\infty],\quad\forall f\in L^p,\|(\sqrt{-\Delta})^\alpha\Delta_jf\|_p\leqslant C_\alpha 2^{j|a|}\|\Delta_jf\|_p,$$

whereas we easily control the L^q norm of a dyadic block by its L^p norm if $q>p$:

$$\exists C, \quad \forall p \in [1,\infty], \quad \forall q \in [p,\infty], \quad \forall f \in L^p, \quad \|S_j f\|_q \leqslant C 2^{j(3/p-3/q)} \|S_j f\|_p .$$

Besov space $B_p^{s,q}$ may be easily defined through the Littlewood–Paley decomposition as the space of tempered distributions f such that for all $j \in \mathbb{N}$ we have $S_j f \in L^p$ and such that moreover $2^{js}\|\Delta_j f\|_p \in l^q(\mathbb{N})$. The space $B_\infty^{-1,\infty}$ plays a key role in finding regular mild solutions since $f \in B_\infty^{-1,\infty}$ if and only if $\forall T>0 \sup_{O<t<T} \sqrt{t}\|e^{t\Delta} f\|_\infty < \infty$.

For further information on the Littlewood–Paley decomposition, the reader may consult the classical book by Bergh and Löfström on interpolation spaces,[20] the nice booklet by Frazier, Jawerth and Weiss on Littlewood–Paley decomposition and wavelets,[21] the book by Meyer on Wavelets and Operators (Ref. 22, vol. 3) and the various treatises on paradifferential calculus.

VI. ADAPTED SPACES FOR THE NAVIER–STOKES EQUATIONS

Definition 7 (adapted spaces):[3] A Banach space X is *adapted* to the Navier–Stokes equations if the following assertions are satisfied.
(a) $\mathscr{S}(\mathbb{R}^3) \subset X \subset \mathscr{S}'(\mathbb{R}^3)$ (continuous imbeddings).
(b) The norm of X is shift invariant:

$$\forall f \in X, \quad \forall x_0 \in \mathbb{R}^3 \|f(x)\|_X = \|f(x-x_0)\|_X.$$

(c) The pointwise product between two elements of X is still well defined as a tempered distribution.
(d) There is a sequence of real numbers $\eta_j > 0$, $j \in \mathbb{Z}$, such that

$$\sum_{j \in \mathbb{Z}} 2^{-|j|} \eta_j < \infty,$$

and such that

$$\forall j \in \mathbb{Z}, \quad \forall f \in X, \quad \forall g \in X \|\Delta_j(fg)\|_X \leqslant \eta_j \|f\|_X \|g\|_X .$$

Theorem 2:[3] Let X be a Banach space adapted to the Navier–Stokes equations. Then, for all $\mathbf{u}_0 \in X^3$ such that $\boldsymbol{\nabla} \cdot \mathbf{u}_0 = 0$, there is a $T^* = T^*(\|\mathbf{u}_0\|_X)$ and a unique solution $\mathbf{u}$ of the Navier–Stokes equations on $(0,T^*)$ with initial value $\mathbf{u}_0$ such that $\mathbf{u} - e^{t\Delta}\mathbf{u}_0 \in \mathscr{C}([0,T^*),X^3)$. Moreover we have the inequality

$$\forall t \in (0,T^*) \|\mathbf{u} - e^{t\Delta}\mathbf{u}_0\|_X \leqslant C \left(t \sum_{4^j t \leqslant 1} 2^j \eta_j + \sum_{4^j t > 1} 2^{-j} \eta_j \right) \|\mathbf{u}_0\|_X^2.$$

Sketch of the proof: The proof lies again on the Picard iteration scheme. The main point is to prove that the bilinear transform $\mathbf{u},\mathbf{v} \to \mathbf{w} = \int_0^t e^{(t-s)\Delta}\mathbb{P}(\boldsymbol{\nabla} \cdot (\mathbf{u} \otimes \mathbf{v}))ds$ is continuous on $L^\infty(X^3)$. It is enough to notice the following properties of this operator: (α) $e^{(t-s)\Delta}\mathbb{P}(\boldsymbol{\nabla} \cdot)$ is a matrix of convolution operators; (β) since the norm of X is shift invariant, we have for $f \in X$ and $g \in L^1 \|f * g\|_X \leqslant \|f\|_X \|g\|_1$; γ) for all j we have $\Delta_j = (\Sigma_{k=j-2}^{k=j+2} \Delta_k)\Delta_j$; δ); we thus may conclude that

$$\|e^{(t-s)\Delta}\mathbb{P}(\boldsymbol{\nabla} \cdot \Delta_j(\mathbf{u} \otimes \mathbf{v}))\|_X \leqslant C \min(2^j, 4^{-j}(t-s)^{-3/2}) \|\Delta_j(\mathbf{u} \otimes \mathbf{v}))\|_X .$$

The continuity of the bilinear transform is then easily established, writing

$$\|\mathbf{w}\|_X \leqslant \sum_{j \in \mathbb{Z}} \|\Delta_j \mathbf{w}\|_X \leqslant \int_0^t \sum_{j \in \mathbb{Z}} \|e^{(t-s)\Delta}\mathbb{P}(\boldsymbol{\nabla} \cdot \Delta_j)(\mathbf{u} \otimes \mathbf{v})\|_X ds,$$

and estimating the integral $\int_0^t \Sigma_{j \in \mathbb{Z}} \min(2^j, 4^{-j}(t-s)^{-3/2}) \eta_j ds$. □

If we look precisely at the definition of an adapted space (and at the examples developed by Cannone), we may see that the theorem is not really a result about the *existence* of solutions but a result about *regularity* at $t=0$ of mild solutions. Indeed, let us assume that assertion (c) in the

definition is ensured by the embedding of X in L^2_{loc}, the space of locally square integrable functions, and that X is invariant under dilations as well as translations: $\forall \lambda>0$ $f\to f(\lambda x)$ is a continuous mapping in X. Then we have the following result.

Proposition 3: Let X be an adapted space for the Navier–Stokes equations. If X is continuously embedded in L^2_{loc} and if $\forall \lambda>0$, $f\to f(\lambda x)$ is a continuous mapping in X, then X is continuously embedded in $M_{2,3}$ and for all $f\in X$ we have $\lim_{t\to 0}\ \sqrt{t}\|e^{t\Delta}f\|_\infty=0$.

Proof: Since the norm of X is shift invariant, the embedding $X\subset L^2_{loc}$ is equivalent to $X\subset F_2$. Thus, we may define the following numbers:

$$\alpha(r)=\sup_{\|f\|_X\leqslant 1, x\in\mathbb{R}^3}\int_{|x-y|\leqslant r}|f(y)|^2\,dy.$$

We are going to estimate $\alpha(r)$ for $r\leqslant 1$. We choose a smooth non-negative function θ which is equal to 1 on $|x|\leqslant 1$ and to 0 for $|x|\geqslant 2$, and we write $f^2=S_0(f^2)+(\Sigma_{j\geqslant 0}\Delta_j(f^2))=g+h$. It is easy to see that g is bounded (since $f^2\in F_1$ and that S_0 maps F_1 into L^∞) so that $|\int g(y)\theta((y-x)/r)dy|\leqslant Cr^3\|f\|_X^2$. We may write h as $\Sigma_{1\leqslant j\leqslant 3}\partial_j h_j$ where $h_j\in X$ and $\|h_j\|_X\leqslant C\|f\|_X^2$; we then have $|\int h(y)\theta((y-x)/r)dy|\leqslant\Sigma_{1\leqslant j\leqslant 3}1/r|\int h_j(y)\partial_j\theta((y-x)/r)dy|\leqslant C\sqrt{r}\sqrt{\alpha(2r)}\|f\|_X^2$; since $\alpha(2r)\leqslant C\alpha(r)$ and $\int_{|x-y|\leqslant r}|f(y)|^2\,dy\leqslant\int(g(y)+h(y))\theta((y-x)/r)dy$, we obtain that for $r\leqslant 1$, $\alpha(r)\leqslant C(r^3+\sqrt{\alpha(r)}\sqrt{r})$ which gives $\alpha(r)\leqslant Cr$: thus X is embedded in the Morrey-Campanato space $M_{2,3}$.

This implies that $X\subset B^{-1,\infty}_\infty$ (and in particular that $\sup_{0\leqslant t\leqslant 1}\sqrt{t}\|e^{t\Delta}f\|_\infty\leqslant C\|f\|_X$). We are now going to show that we also have $\lim_{t\to 0}\|e^{t\Delta}f\|_\infty=0$. Indeed let us define

$$C_j=\sup_{\|f\|_X\leqslant 1}\|\Delta_j f\|_\infty.$$

Since X is invariant through dyadic dilations, we see that C_j is slowly varying: $\exists\,\gamma>1$, $\forall j\in\mathbb{Z}(1/\gamma)\leqslant C_j/C_{j+1}\leqslant\gamma$. If f_j satisfies $\|f_j\|_X=1$ and $|\Delta_j f_j(0)|\geqslant 1/2C_j$, then $f_j f_{j+3}=\Sigma_{j+1\leqslant k\leqslant j+5}\Delta_k(f_j f_{j+3})$, hence $1/4C_jC_{j+3}\leqslant(\Sigma_{j+1\leqslant k\leqslant j+5}C_k)(\Sigma_{j\leqslant l\leqslant j+6}\eta_l)$. Thus, we get $\lim_{j\to+\infty}2^{-j}C_j=0$. This is just what we needed to show: indeed, write for $f\in X$ $f=S_N f+\Sigma_{j\geqslant N}\Delta_j f=f_N+g_N$; then for $t\in(0,1]$ and $N\geqslant 1$ we have $\sqrt{t}\|e^{t\Delta}f\|_\infty\leqslant\sqrt{t}\|e^{t\Delta}f_N\|_\infty+\sqrt{t}\|e^{t\Delta}g_N\|_\infty\leqslant\sqrt{t}\|f_N\|_\infty+C\|g_N\|_{B^{-1,\infty}_\infty}\leqslant C(2^N\sqrt{t}+\sup_{j\geqslant N}2^{-j}C_j)\|f\|_X$. ■

Thus, we have shown that the spaces studied by Cannone are subspaces of one space, the space $M_{2,3}$. Moreover, for all the examples quoted in Ref. 3 [Sobolev spaces $H^s_p(p<3,s>3/p-1)$, Morrey-Campanato spaces $M^{2,p}(p>3)$, Lebesgue spaces L^p $(p>3)$], we have the property that $\forall f,g\in X\cap L^\infty\|fg\|_X\leqslant C(\|f\|_X\|g\|_\infty+\|g\|_X\|f\|_\infty)$ so that the regular mild solutions described in Theorem 1 are as well in $\cap_{T<T*}L^\infty((0,T),X^3)$ [and we may notice that conversely a mild solution in $\cap_{T<T*}L^\infty((0,T),X^3)$ where X is one of these spaces and is in fact a regular mild solution]. It means that we do not need a very precise analysis in frequency for proving the existence of a solution. Moreover, the tools used for exhibiting solutions in $M_{2,3}$ can be used as well for obtaining regularity results in the limit cases of Cannone's theorem, when $\eta_j=2^j$ (think of L^3: this is the same recipe that we use in limit cases[5]).

VII. SELF-SIMILAR SOLUTIONS

In view of theorem 1, we see clearly the strategy for exhibiting self-similar solutions to the Navier–Stokes equations. We take a space of tempered distributions X which contains nontrivial homogeneous distributions and try to get a theorem of global existence and uniqueness for solutions of the Navier–Stokes equations. Again, we assume that the norm of X is shift-invariant, and we assume, for the sake of homogeneity, that $\|f(\lambda x)\|_X=(1/\lambda)\|f\|_X$. This implies that $X\subset\dot{B}^{-1,\infty}_\infty$, or equivalently that for all $f\in X$ we have $\sup_{0<t}\sqrt{t}\|e^{t\Delta}f\|_\infty\leqslant C\|f\|_X$.

We write again $B(\mathbf{f},\mathbf{g})=\int_0^t(\nabla\otimes O_{t-s})*(\mathbf{f}\otimes\mathbf{g})ds$ and we have to check on which Banach space based on X and containing the tendencies $e^{t\Delta}\mathbf{u}_0$ for $\mathbf{u}_0\in X^3$ the bilinear transform B is continuous. The first space we can try is $E=L^\infty((0,\infty),X^3)$. Of course, the bilinear product uv should then be defined for elements of X; hence we should assume that X is embedded in L^2_{loc}; we then find that $X\subset\dot{M}_{2,3}$, so that this case has already been discussed. There are many instances of

spaces X which can be treated this way and provide self-similar solutions: homogeneous Besov spaces $\dot{B}^{s,\infty}_p$ where $p<3$ and $s=3/p-1$,[3,5,23] the Lorentz space $L^{3,\infty}$[7] and the space $\dot{B}^{2,\infty}_{PM}=\{f/|\xi|^2.\mathscr{F}f\in L^\infty\}$ used by Le Jan and Sznitman.[6]

We may use the smoothing effect of $e^{t\Delta}$ and try to start from a more singular initial value. Cannone[3] and Planchon[9] have shown that one could take $\mathbf{u}_0\in\dot{B}^{s,\infty}_p$ where $p\in(3,\infty)$ and $s=3/p-1$. Then $E=\{\mathbf{f}/\sup_{0<t}t^{1/2-3/2p}\|\mathbf{f}(t,.)\|_p<\infty\}$ is a good choice.

This latter example can even be generalized by replacing L^p by F_p. We thus may consider a $\mathbf{u}_0\in(\dot{B}^{-1,\infty}_\infty)^3$ such that $\sup_{0<t}t^{1/2-3/2p}\|e^{t\Delta}\mathbf{u}_0\|_{F_p}<\infty$; then if this quantity is small enough, we shall have a mild solution in $X_p=\{\mathbf{f}/\sup_{0<t}t^{1/2-3/2p}\|\mathbf{f}(t,\cdot)\|_{F_p}<\infty\}$. As a matter of fact, the first instance of self-similar solutions was constructed with the help of Morrey-Campanato spaces.[24] X_p is a kind of Besov space above a Morrey-Campanato space, as in Ref. 18.

Self-similar solutions are connected to the asymptotic behavior of solutions as t goes to ∞. Planchon[9] has studied the problem of solutions $\mathbf{u}(t,x)$ such that $\sup_{t>0}t^{1/2-3/2p}\|\mathbf{u}\|_p<\infty$ and $\lim_{t\to\infty}\sqrt{t}\mathbf{u}(t,\sqrt{t}x)$ exists in L^p where p is fixed in $(3,\infty)$. Then the initial value $\mathbf{u}_0$ [well defined since $\mathbf{u}\in\cap_{T>0}L^2((0,T),(F_2)^3)$] belongs to $\dot{B}^{3/p-1,\infty}_p$, the limit $\mathbf{V}=\lim_{t\to\infty}\sqrt{t}\mathbf{u}(t,\sqrt{t}x)$ satisfies that $(1/\sqrt{t})\mathbf{V}(x/\sqrt{t})$ is a self-similar solution of the Navier–Stokes equations and the initial value $\mathbf{v}_0\in\dot{B}^{3/p-1,\infty}_p$ of $(1/\sqrt{t})\mathbf{V}(x/\sqrt{t})$ satisfies $e^\Delta\mathbf{v}_0=\lim_{t\to\infty}\sqrt{t}\{e^{t\Delta}\mathbf{u}_0\}(t,\sqrt{t}x)$ in L^p. Notice that this proves that the asymptotic behavior is governed only by the lowest frequencies of the initial value $\mathbf{u}_0$, since for all $N\in\mathbb{Z}$ we have $\sqrt{t}\|\{e^{t\Delta}(Id-S_N)\mathbf{u}_0\}(t,\sqrt{t}x)\|_p\leqslant C(1/t)\|[(Id-S_N)/\Delta]\mathbf{u}_0\|_{\dot{B}^{3/p-1,\infty}_p}\leqslant C(1/2^Nt)\|(Id-S_N)\mathbf{u}_0\|_{\dot{B}^{3/p-1,\infty}_p}$ and thus $e^\Delta\mathbf{v}_0=\lim_{t\to\infty}\sqrt{t}\{e^{t\Delta}S_N\mathbf{u}_0\}(t,\sqrt{t}x)$.

VIII. UNIQUENESS IN LIMIT SPACES

In this section, we pay a few words to the uniqueness problems for mild solutions in $\mathscr{C}([0,T^*),(L^3)^3)$. As a matter of fact, we may not construct directly a solution in $\mathscr{C}([0,T^*),(L^3)^3)$, given $\mathbf{u}_0\in(L^3)^3$, because L^3 is a limit space and this is not clear whether B is continuous or not on $\mathscr{C}([0,T^*),(L^3)^3)$. The reader may find various constructions in Refs. 1, 3 or 9, and uniqueness was granted only in the subspaces of $\mathscr{C}([0,T^*),(L^3)^3)$ where the iteration algorithm converges (for instance, for regular mild solutions). The constructions aimed to prove that the fluctuation was in $\mathscr{C}([0,T^*),(L^3)^3)$, and even in a smaller subspace [such as $\mathscr{C}([0,T^*),(Y)^3)$ with $Y=\dot{B}^{0,1}_3$ (Refs. 3 and 5) or $Y=\dot{B}^{0,2}_3$ (Ref. 9) or, as one may easily check for a regular mild solution, $Y=L^{3,1}$]. The idea we had in Ref. 5 was that the problem of uniqueness could be treated in a *greater* space, since one did not bother to get an estimate especially in the L^3 norm, the solutions being already given.

The example of Le Jan and Sznitman[6] is particularly interesting. They consider the limit space $X=\dot{B}^{2,\infty}_{PM}$ and show in a very simple way that the operator B is well behaved on this space. As a matter of fact, since the Riesz transforms are obviously bounded on this space, we may just look at the scalar bilinear operator $A(u,v)=\int_0^t e^{(t-s)\Delta}\sqrt{-\Delta}(uv)ds$ operating on $L^\infty(X)$.

Now, they do not try to prove that $\int_0^t\|e^{(t-s)\Delta}\sqrt{-\Delta}(uv)\|_X ds$ can be controlled; this is too gross an estimate. But the estimate they use is even more direct: $(1/\sqrt{-\Delta})(uv)\in X$ since (taking the Fourier transform) it is enough to notice that $1/|\xi|^2*1/|\xi|^2\leqslant C/|\xi|$ (an obvious inequality due to radial invariance and homogeneity), and then one writes

$$|\mathscr{F}\{A(uv)(t,\cdot)\}(\xi)\leqslant\frac{C}{|\xi|^2}\sup_{0<s<t}\|u\|_X\sup_{0<s<t}\|v\|_X\int_0^t e^{-(t-s)|\xi|^2}|\xi|^2ds$$

$$\leqslant\frac{C}{|\xi|^2}\sup_{0<s<t}\|u\|_X\sup_{0<s<t}\|v\|_X\int_0^\infty e^{-\tau}d\tau.$$

These simple estimates suggest that one can deal with certain limit spaces, as far as their norm can be computed *frequency by frequency* or at least by frequency packets. This is the case of the Besov l^∞ spaces: if X is a Banach space of tempered distributions with a shift invariant norm and if $\dot{B}^{s,\infty}_X$ for $s>0$ is the space of distributions $f=\Sigma_{j\in\mathbb{Z}}\Delta_j f$ such that $2^{-js}\|\Delta_j f\|_X\in l^\infty(\mathbb{Z})$, then $\forall s\in(0,2)$, $\forall f\in L^\infty((0,\infty),\dot{B}^{s,\infty}_X)\int_0^\infty\tau\Delta e^{\tau\Delta}f(\tau,\cdot)(d\tau/\tau)\in B^{s,\infty}_X$.

Thus we see for instance that B operates boundedly on $\dot{B}_2^{1/2,\infty}$ and this implies readily the uniqueness of mild solutions in $\mathscr{C}((0,T^*),\dot{H}^{1/2})$. This argument does not work directly for the limit space L^3 since we then should work with the Besov space $\dot{B}_3^{0,\infty}$ which is not embedded in L^2_{loc} so that we would be bound again to use the smoothing effect of the heat kernel.

The next idea is to split the solutions in tendency and fluctuation. We consider two mild solutions $\mathbf{u}=e^{t\Delta}\mathbf{u}_0-B(\mathbf{u},\mathbf{u})=e^{t\Delta}\mathbf{u}_0-\mathbf{w}_1$ and $\mathbf{v}=e^{t\Delta}\mathbf{u}_0-B(\mathbf{v},\mathbf{v})=e^{t\Delta}\mathbf{u}_0-\mathbf{w}_2$ in $\mathscr{C}((0,T^*),(L^3)^3)$ and write $\mathbf{w}=\mathbf{u}-\mathbf{v}=\mathbf{w}_2-\mathbf{w}_1=-B(\mathbf{w},\mathbf{v})-B(\mathbf{u},\mathbf{w})$. Thus we see clearly the role of the fluctuations: they control the behavior of $\mathbf{w}$. Indeed, we now turn to the scalar operator A (since the Riesz transforms operate on all the spaces involved in our computations) and write first

$$A(u,v)=-(-\Delta)^{-1/4}\int_0^t(t-s)\Delta e^{(t-s)\Delta}(-\Delta)^{-1/4}(uv)\,\frac{ds}{t-s},$$

and find that A maps $\mathscr{C}([0,T],L^3)\times\mathscr{C}([0,T],L^3)$ into $L^\infty([0,T],(\dot{B}_2^{1/2,\infty})$ due to Sobolev inequalities. Then we use the continuity of u to split it in u_1+u_2 (and the same for v) where $\forall t<T^*\|u_1(t,\cdot)\|_3<\epsilon$ and $\sup_{0<s<t}\|u_2(s,\cdot)\|_4<\infty$. Now we notice that, due to Sobolev and Hölder inequalities, $\forall f\in L^p(-\Delta)^{-3/2p}M_f$ (where M_f is the operator of pointwise multiplication by f) is continuous on L^q for all $q\in(p/(p-1),\infty)$ and on $\dot{L}^{3/p}_q$ for all $q\in(1,p)$, hence by interpolation on $\dot{B}^{s,\infty}_q$ for all $q\in(p/(p-1),p)$ and all $s\in(0,3/p)$. Taking $p=3{,}4$, $q=2$, $s=1/2$, we easily get that

$$\forall T<T^*,\quad \forall t<T\|A(u,w)(t,\cdot)\|_{\dot{B}_2^{1/2,\infty}}\leqslant C(\epsilon+C(T,u)t^{1/4})\sup_{0<s<t}\|w(s,\cdot)\|_{\dot{B}_2^{1/2,\infty}},$$

which implies $w=0$ as far as $C(\epsilon+C(T,u)t^{1/4})<1$; hence we find that we have local uniqueness and this uniqueness propagates to the whole interval $[0,T^*)$ because of the continuity of $\mathbf{u}$ and $\mathbf{v}$.

Thus, a precise analysis of the frequencies involved in the fluctuation allows one to prove uniqueness in $\mathscr{C}([0,T^*),(L^3)^3)$. This proof works in a more general pattern.

Theorem 3:[5] Let $\dot{m}_{p,3}$ be the closure of the smooth compactly supported functions in $\dot{M}_{p,3}$. If $p>2$ and $\mathbf{u}$ and $\mathbf{v}$ are two weak solutions of the Navier–Stokes equations on $(0,T^*)\times\mathbb{R}^3$ such that $\mathbf{u}$ and $\mathbf{v}$ belong to $\mathscr{C}([0,T^*),(\dot{m}_{p,3})^3)$ and have the same initial value, then $\mathbf{u}=\mathbf{v}$.

Hence we have uniqueness in $\mathscr{C}([0,T^*),(Y)^3)$ for all limit space Y in which the test functions are dense and which is embedded into L^p_{loc} for some $p>2$.

But for L^3, there is now a much simpler way to get uniqueness: the space $\dot{B}_3^{0,\infty}$ was not good for proving uniqueness but it can be replaced by another l^∞ space: $L^{3,\infty}$.[7] Moreover this result is not really *new*: the fact that $f\to\int_0^t(t-s)\Delta e^{(t-s)\Delta}(1/\sqrt{-\Delta})f(s,\cdot)[ds/(t-s)]$ maps $L^\infty((0,T),L^{3/2,\infty})$ into $L^\infty((0,T),L^{3,\infty})$ can be considered as a dual result to the fact that $f\to\int_s^T(t-s)\Delta e^{(t-s)\Delta}(1/\sqrt{-\Delta})f(t,\cdot)[dt/(t-s)]$ maps $L^1((0,T),L^{3/2,1})$ into $L^1((0,T),L^{3,1})$; this latter result is proved in exactly the same way as the fact that $f\to\int_0^t(t-s)\Delta e^{(t-s)\Delta}(1/\sqrt{-\Delta})f(s,\cdot)\times[ds/(t-s)]$ maps $L^1((0,T),L^{3/2,1})$ into $L^1((0,T),L^{3,1})$ or the fact that $f\to\int_0^t e^{(t-s)\Delta}f(s,\cdot)ds$ maps $L^1((0,T),L^{3/2,1})$ into $L^1((0,T),\mathscr{C}_0)$ (remember that $1/\sqrt{-\Delta}$ maps $L^{3,1}$ into $\mathscr{C}_0$), a fact which is used by Lions in Ref. 13 to get the Foias-Guillopé-Temam theorem that the Leray weak solutions belong to $\cap_{T>0}L^1((0,T),\mathscr{C}_0)$!

IX. WEAK SOLUTIONS WITH INFINITE ENERGY

There is still another way to prove uniqueness in L^3. We split $\mathbf{u}_0$ in $\mathbf{v}_0+\mathbf{w}_0$ where $\mathbf{v}_0\in(L^2)^3$ and $\|\mathbf{w}_0\|_3<\epsilon$. Then Kato's theorem allows one to find a regular mild solution $\mathbf{w}$ on $(0,\infty)\times\mathbb{R}^3$ with initial value $\mathbf{w}_0$ and Leray's theory allows one to find a weak solution $\mathbf{v}\in L^\infty((0,\infty),(L^2)^3)\cap L^2((0,\infty),(\dot{H}^1)^3)$ of the problem

$$\mathbf{v}(0,\cdot)=\mathbf{v}_0,\quad \boldsymbol{\nabla}\cdot\mathbf{v}=0,\quad \partial_t\mathbf{v}=\Delta\mathbf{v}-\mathbb{P}(\boldsymbol{\nabla}\cdot[\mathbf{v}\otimes\mathbf{v}+\mathbf{v}\otimes\mathbf{w}+\mathbf{w}\otimes\mathbf{v}]),$$

with an energy estimate:

$$\forall t>0\|\mathbf{v}(t,\cdot)\|_2^2+2\int_0^t\|\boldsymbol{\nabla}\otimes\mathbf{v}(s,\cdot)\|_2^2\,ds\leqslant\|\mathbf{v}_0\|_2^2+2\int_0^t\boldsymbol{\nabla}\otimes\mathbf{v}\cdot\mathbf{v}\otimes\mathbf{w}ds.$$

Now, if $\mathbf{u}$ is a mild solution of Navier–Stokes in $\mathscr{C}([0,T^*),(L^3)^3)$ with initial value $\mathbf{u}_0$, then $\mathbf{z}=\mathbf{u}-\mathbf{w}$ is a solution for the same problem as $\mathbf{v}$ and belongs to $\mathscr{C}([0,T^*),(L^3)^3)$. If we can prove that $\mathbf{z}$ belongs to $\cap_{T<T^*}L^\infty((0,T),(L^2)^3)\cap L^2((0,T),(\dot{H}^1)^3)$, then we may apply the Sohr and Von Wahl uniqueness theorem[2] to get $\mathbf{u}=\mathbf{v}+\mathbf{w}$ on $(0,T^*)$. Since $\mathbf{v}+\mathbf{w}$ does not depend on $\mathbf{u}$, this proves the uniqueness of $\mathbf{u}$! This is exactly what is done in Ref. 8, and moreover this method provides a way to prove the existence of global suitable weak solutions for the Navier–Stokes equations with an initial value with infinite energy.

Definition 8 (suitable weak solutions)*:*[25] A *suitable weak solution* of the Navier–Stokes equations on $(0,T)$ is a distribution vector field $\mathbf{u}(t,x)$ in $(\mathscr{D}'((0,T)\times\mathbb{R}^3)^3$ such that:
(a) $\mathbf{u}$ and $\nabla\otimes\mathbf{u}$ are locally square integrable on $(0,T)\times\mathbb{R}^3$;
(b) $\nabla\cdot\mathbf{u}=0$;
(c) $\exists p\in L^{5/3}_{loc}((0,T)\times\mathbb{R}^3)\ \partial_t\mathbf{u}=\Delta\mathbf{u}-(\mathbf{u}\cdot\nabla)\mathbf{u}-\nabla p$;
(d) for all compact subsets K of $(0,T)\times\mathbb{R}^3$ we have $\sup_{t\in(0,T)}\int_{(t,x)\in K}|\mathbf{u}|^2dx<+\infty$;
(e) $\forall\varphi\in\mathscr{D}((0,T)\times\mathbb{R}^3)$ such that $\varphi\geqslant 0$; we have

$$2\int\int|\nabla\otimes\mathbf{u}|^2\varphi dxdt\leqslant\int\int|\mathbf{u}|^2(\partial_t\varphi+\Delta\varphi)dxdt+\int\int(|\mathbf{u}|^2+2p)(\mathbf{u}\cdot\nabla)\varphi dxdt.$$

We recall that suitable solutions were proved in Ref. 25 to be regular outside from a subset of $(0,T)\times\mathbb{R}^3$ whose one-dimensional Hausdorff measure is equal to 0. The theorem proved in Ref. 8 states the existence of global suitable weak solutions for an initial value $\mathbf{u}_0\in L^q(\mathbb{R}^3)$ where $2\leqslant q<\infty$.

Theorem 4 (suitable weak solutions with unbounded energy)**:** Let $q\in[2,+\infty)$ and let $\mathbf{u}_0\in L^q(\mathbb{R}^3)$ such that $\nabla\cdot\mathbf{u}_0=0$. Then there exists a suitable weak solution $\mathbf{u}$ of the Navier–Stokes equations on $(0,+\infty)$ such that $\mathbf{u}$ belongs to $\cap_{T>0}L^2((0,T),(E_2)^3)\cap L^r((0,T),(L^q)^3)$ (with $1/r=3/2(1/2-1/q)$ if $q\leqslant 6$ and $1/r=1-3/q$ if $q\geqslant 6$) and such that $\mathbf{u}(0,\cdot)=\mathbf{u}_0$.

ACKNOWLEDGMENTS

The author thanks G. Furioli and E. Terraneo for their useful comments on the first version of this paper. He also thanks them (and M. Cannone, F. Planchon and Y. Meyer) for many valuable discussions on the Navier–Stokes equations.

[1] T. Kato, ''Strong L^p solutions of the Navier–Stokes equations in $\mathbb{R}^m$ with applications to weak solutions,'' Math. Z. **187**, 471–480 (1984).
[2] W. Von Wahl, *The Equations of Navier–Stokes and Abstract Parabolic Equations* (Vieweg & Sohn, Wiesbaden, 1985).
[3] M. Cannone, *Ondelettes, Paraproduits et Navier–Stokes* (Diderot Editeur, Paris, 1995).
[4] P. Federbush, ''Navier and Stokes meet the wavelet,'' Commun. Math. Phys. **155**, 219–248 (1993).
[5] G. Furioli, P. G. Lemarié-Rieusset, and E. Terraneo ''Unicité dans $L^3(\mathbb{R}^3)$ et d'autres espaces limites pour Navier–Stokes,'' submitted to Revista Mat. Iberoamer.
[6] Y. Le Jan and A. S. Sznitman, ''Cascades aléatoires et 'equations de Navier–Stokes,'' C. R. Acad. Sci. Paris **324**, 823–826 (1997).
[7] Y. Meyer, ''Wavelets, paraproducts and Navier–Stokes equations,'' to appear in Mem. AMS.
[8] P. G. Lemarié-Rieusset, ''Solutions faibles d'énergie infinie pour les équations de Navier–Stokes dans $\mathbb{R}^3$,'' work in progress.
[9] F. Planchon, ''Solutions globales et comportement asymptotique pour les équations de Navier–Stokes,'' Thesis, Ecole Polytechnique, 1996.
[10] R. Temam, *Navier–Stokes Equations* (North-Holland, New York, 1977).
[11] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Fluids* (Gordon and Breach, New York, 1969).
[12] P. Constantin and C. Foias, *Navier–Stokes Equations* (University of Chicago Press, Chicago, IL, 1988).
[13] P. L. Lions, *Mathematical Topics in Fluid Mechanics* (Oxford Science, 1996), Vol. 1.
[14] C. W. Oseen, *Neuere Methoden und Ergebnisse in der Hydrodynamik* (Akademische Verlagsgesellschaft, Leipzig, 1927).
[15] J. Leray, ''Essai sur le mouvement d'un fluide visqueux emplissant l'espace,'' Acta Math. **63**, 193–248 (1934).
[16] H. Fujita and T. Kato, ''On the Navier–Stokes initial value problem, I,'' Arch. Ration. Mech. Anal. **16**, 269–315 (1964).
[17] T. Kato, ''Strong solutions of the Navier–Stokes equations in Morrey spaces,'' Bol. Soc. Brasil. Math. **22**, 127–155 (1992).
[18] H. Kozono and Y. Yamazaki, ''Semilinear heat equations and the Navier–Stokes equations with distributions in new function spaces as initial data,'' Comm. P. D. E. **19**, 959–1014 (1994).
[19] M. E. Taylor, ''Analysis on Morrey spaces and applications to Navier–Stokes equations and other evolution equations,'' Comm. P. D. E. **17**, 1407–1456 (1992).
[20] J. Bergh and J. Löfström, *Interpolation Spaces* (Springer-Verlag, Berlin, 1976).

[21] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood–Paley Theory and the Study of Function Spaces*, CBMS Regional Conference Series 79 (American Mathematical Society, Providence, RI, 1991).
[22] Y. Meyer, *Ondelettes et Opérateurs* (Hermann, Paris, 1990), Vols. 1–3.
[23] J. Y. Chemin, ''Sur l'unicité dans le système de Navier–Stokes tridimensionnel,'' preprint, 1997.
[24] Y. Giga and T. Miyakawa, ''Navier–Stokes flow in $\mathbb{R}^3$ with measures as initial vorticity and Morrey spaces,'' Comm. P. D. E. **14**, 577–618 (1989).
[25] L. Caffarelli, R. Kohn, and L. Nirenberg, ''Partial regularity of suitable weak solutions of the Navier–Stokes equations,'' Commun. Pure Appl. Math. **35**, 771–831 (1982).

Formal improvements in the solution of the wavelet-transformed Poisson and diffusion equations

Jacques Lewalle[a)]
Department of Mechanical, Aerospace, and Manufacturing Engineering, Syracuse University, Syracuse, New York 13244-1240

(Received 19 September 1997; accepted for publication 27 March 1998)

Hermitian wavelets' relation to the Laplace operator leads to a natural measure of the scale factor that emphasizes the largest component wave number. For the Poisson equation (e.g. the pressure equation in Navier–Stokes turbulence), the wavelet transform of the solution at a given location and scale depends on the wavelet transform of the source field at the same location and at nearby and larger scales. For the diffusion problem, the Hamiltonian formulation is simplified through a canonical transformation. © *1998 American Institute of Physics.*
[S0022-2488(98)01408-X]

I. INTRODUCTION

Navier–Stokes turbulence motivates the present research effort. The improvements in experimental and computational techniques have yielded a wealth of data, and the role of intermittency and near-singularities, among other important physical descriptors, is being actively documented. However, the consistency and importance of the associated scaling relations falls short of a theory, since a link to fundamental principles (mass and momentum conservation) relies on obtaining a solution (experimental or numerical) rather than on symbolic relations.

The closest thing to a theory of turbulence is the renormalization group approach,[1] which is formulated in Fourier space[2,3] and therefore conceals the effects of intermittency as iterative low-pass filtering is applied. Wavelet transforms[4,5] provide a compromise between spectral and spatial information, and could capture scale-dependent intermittency and its dynamics. Recent successes in data analysis[6] and computational simulation[7,8] attest to the relevance of the tool. In this paper, some incremental progress is reported on the analytical treatment of the pressure and viscous terms in Navier–Stokes turbulence. Any wavelet can potentially yield the desired result, but analytical manipulations involving the Laplacian are most effective with the continuous wavelets of the Hermitian family,[9,10] which are used exclusively in this paper. Ostensibly, we will assume an infinite spatial domain, which matches the Fourier version[2] but can be relaxed in simple geometries with the help of Green's functions.[10]

Notations and earlier results[9,10] are summarized below. Hermitian wavelets are the family of continuous wavelets obtained by taking derivatives of the Gaussian bell-shaped curve. In the one-dimensional case, let us define the mother wavelet g_n as

$$g_n(x) = -\frac{d}{dx} g_{n-1}(x),$$

where

$$g_0(x) = e^{-x^2/2}. \tag{1}$$

The family of g_n's obeys the Hermite recursion formula

$$g_n(x) = x g_{n-1}(x) - (n-1) g_{n-2}(x). \tag{2}$$

For $n>0$, the Hermitian wavelet transform at scale κ of a field $p'(x)$ is given by

a)Electronic mail: jlewalle@syr.edu

 © 1998 American Institute of Physics

$$p(\kappa,x)=\kappa^{1/2}\int_{-\infty}^{\infty}p'(y)g_n(\kappa(y-x))dy. \tag{3}$$

The wavelet number κ is a measure of inverse length scale, analogous to the wave number k of Fourier analysis, and equal to the inverse dilation factor a used in much of the wavelet literature. (κ,x) will be refered to as wavelet space, κ (>0) as the wavelet number. In this paper, we will use primed symbols for the spatial representation of fields, hatted symbols for the Fourier representation, and plain symbols for the wavelet representation.

Integration by parts and Eq. (2) yield the compatibility equation

$$\kappa^3\frac{\partial p}{\partial\kappa}+\frac{\partial^2 p}{\partial x^2}+\left(n+\frac{1}{2}\right)\kappa^2 p=0. \tag{4}$$

This condition, which must be satisfied by any Hermitian wavelet transform $p(\kappa,x)$, is the basis of all manipulations of the Laplace operator in the remainder of this paper.

These relations have been generalized[9] to N-dimensions if the transform is carried out in each Cartesian direction independently and if the same wavelet index n applies in each direction. Accordingly, the N-D wavelet transform of the pressure field $p(\mathbf{x})$ is

$$p(\kappa_j,\mathbf{x})=\int\cdots\int_{-\infty}^{\infty}p'(\mathbf{y})\prod_{j=1}^{N}[\kappa_j^{1/2}g_n(\kappa_j(y_j-x_j))dy_j]. \tag{5}$$

It can be noted that $\{\kappa_j\}$ is a set of positive scaling coefficients (wavelet-numbers), and is not a vector. This fact is reflected in the notations, where $\boldsymbol{\kappa}$ will be avoided, and where the summation convention does not apply to κ_j.

In Sec. II, we set out to simplify some results relative to the Poisson equation, with the specific goal of finding workable alternatives to the compact Fourier representation[1–3] of pressure in incompressible turbulence. A change in variables leads to a ''natural'' measure of the scale factor, which favors the largest scale component, in contrast to the Euclidean metric of Fourier space. Then, in Sec. III, we show that this change of variables is part of a canonical transformation that reformulates the Hamiltonian for diffusion[10] in action-angle variables, possibly the simplest expression of conservation in simple transport processes. The Poisson solution is then reformulated in these variables.

II. THE POISSON EQUATION REVISITED

In incompressible flows, pressure obeys the Poisson equation

$$\nabla^2 p'=S'(\mathbf{x}), \tag{6}$$

in which the source terms S' depend on nonlinear interactions. In the spatial representation of the solution

$$p'(\mathbf{x})=\int S'(\mathbf{y})G'(\mathbf{x},\mathbf{y})d\mathbf{y}, \tag{7}$$

the form of Green's function G' depends on the dimension N of the space (and of course on the configuration of the boundaries), but in all cases p' at one point will depend on sources at all other points in the field. This enriches the physics of any flow, and complicates its calculation or modeling.

In the Fourier representation

$$\hat{p}(\mathbf{k})=\int p'(\mathbf{x})e^{-i\mathbf{k}\cdot\mathbf{x}}d\mathbf{x}, \tag{8}$$

the spectral distribution of pressure is given by

$$\hat{p}(\mathbf{k}) = -\frac{\hat{S}(\mathbf{k})}{|\mathbf{k}|^2}. \tag{9}$$

This relation is algebraic, and is local in wave number space. This facilitates enormously the formal manipulations of the Navier–Stokes equations and their coding for spectral calculations; this is also reflected in a rich tradition of modeling[2,3] and in renormalization theory.[1] However, any effect of the localization of the sources, as would occur in an intermittent field, is distributed throughout the Fourier spectrum, and is generally eliminated from modeling considerations. Wavelet transforms are an obvious alternative to address this priority. We will treat successively the one-dimensional case to establish notations and procedures, then generalize the result to *N*-dimensions.

A. The one-dimensional case

Making use of the compatibility condition [Eq. (4)], the 1-D version of Eq. (6) transforms into

$$\kappa^3 \frac{\partial p}{\partial \kappa} + \left(n + \frac{1}{2}\right)\kappa^2 p = -S(\kappa, x). \tag{10}$$

This equation was solved[9] using the method of characteristics, and this section consists of a reformulation of the earlier result. After elimination of the characteristic parameter, the solution is

$$p(\kappa, x) = -\kappa^{-[n+(1/2)]} \int_0^{\kappa} S(\kappa', x)\kappa'^{[n-(5/2)]} d\kappa'. \tag{11}$$

This relation was interpreted as reflecting the asymptotic equilibrium for the corresponding diffusion problem originating in the remote past under the action of fixed sources. An alternative presentation is obtained by defining the relative scale factor

$$\xi = \kappa'/\kappa, \tag{12}$$

i.e., the ratio of the wavelet number for the source to that of the resulting pressure field. Then

$$p(\kappa, x) = -\kappa^{-2} \int_0^1 S(\xi\kappa, x)\xi^{n-\frac{5}{2}} d\xi. \tag{13}$$

It was noted[9] that all spectral contributions to the source terms are evaluated locally at the point x of interest. In order to capture the effect of a remote source, one must look at scales large enough (small κ) for the analyzing wavelet to cover both the source and the point of observation, which will affect the amplitude of the wavelet coefficient. Also noteworthy, the κ^{-2} prefactor is reminiscent of the Fourier solution [Eq. (9)]. Furthermore, it is apparent that only spectral contributions to the source terms for $\kappa' < \kappa$ are reflected in the pressure field. As n increases, we see in Fig. 1 that the weight of spectral contributions shifts closer to $\xi = 1$. The net result of the spatial/spectral compromise is that we retain a strictly local spatial distribution (traceable to the compatibility condition [Eq. (4)]), and a limited spectral localization of the effect of the source. A fuller discussion will benefit from the *N*-dimensional result.

B. The *N*-dimensional case

The properties of the transform used above can be generalized to *N*-dimensions. In particular, the compatibility equation becomes

$$\sum_j \kappa_j^3 \frac{\partial p}{\partial \kappa_j} + \left(n + \frac{1}{2}\right)\sum_j \kappa_j^2 p + \sum_j \frac{\partial^2 p}{\partial x_j^2} = 0. \tag{14}$$

Based on this relation, the transformed Poisson equation and its solution were derived previously.[9] In characteristic form, and with an arbitrary reference characteristic parameter s_0, the solution is

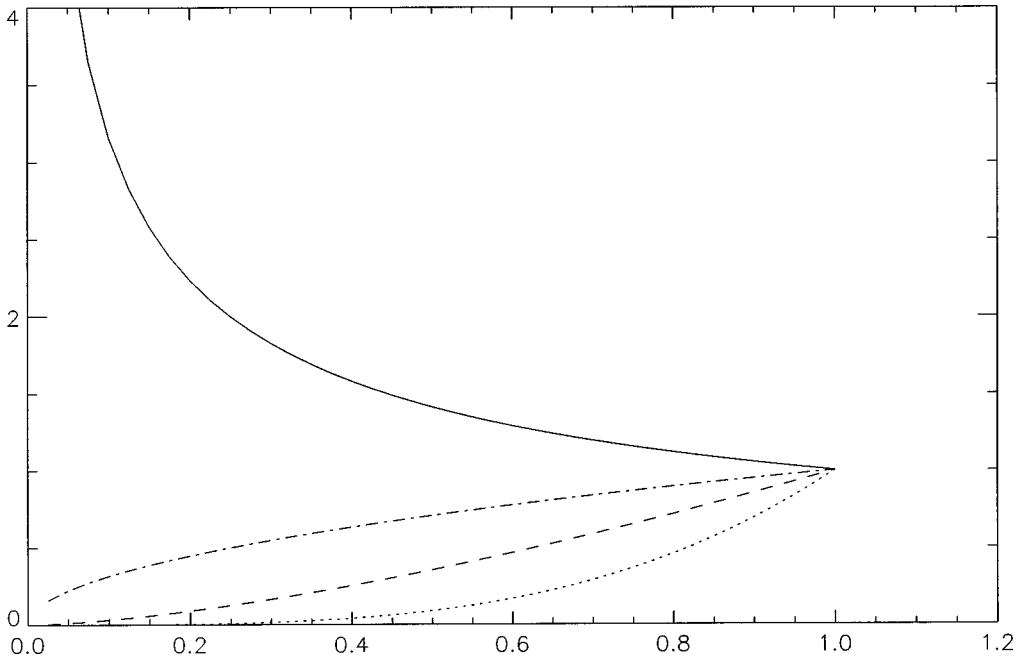

FIG. 1. Relative weight of contributions to the pressure at a given scale κ from sources at relative scale ξ, according to Eq. (13). Solid line: $n=2$; dash-dotted line: $n=3$; dashed line: $n=4$; dotted line: $n=6$.

$$\kappa_j(s)=\frac{\kappa_j(s_0)}{\sqrt{1-2(s-s_0)\kappa_j^2(s-s_0)}} \tag{15}$$

and

$$p(\kappa_j(s),\mathbf{x})=-\int_{-\infty}^{s} S(\kappa_j(\sigma),\mathbf{x})\prod_{i=1}^{N}\left[\frac{\kappa_i(\sigma)}{\kappa_i(s)}\right]^{[n+(1/2)]} d\sigma. \tag{16}$$

In this case, the elimination of the characteristic parameter s is not trivial, and it requires the introduction of a unique scale factor. Let

$$\frac{1}{\kappa^2}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\kappa_i^2}. \tag{17}$$

It is easy to show that

$$\frac{d\kappa}{\kappa^3}=ds=\frac{d\kappa_i}{\kappa_i^3}. \tag{18}$$

Simple manipulations yield a new expression for pressure

$$p(\kappa_j,\mathbf{x})=-\kappa^{-2}\int_0^1 S\left(\frac{\xi\kappa_j}{\sqrt{a_j}},\mathbf{x}\right)\xi^{N[n+(1/2)]-3}\prod_{i=1}^{N} a_i^{-(n/2)-(1/4)}d\xi. \tag{19}$$

We recognize, in order, the κ^{-2} prefactor, the wavelet-transformed source field, the relative scale factor as in Eq. (12), and a product of scale-dependent directional factors

$$a_i(\kappa_i,\kappa,\xi)=\xi^2+\frac{\kappa_i^2}{\kappa^2}(1-\xi^2). \tag{20}$$

Dominant and vanishing contributions can be identified from Eqs. (19) and (20). As $\xi\to 0$,

$$a_j(\kappa_j,\kappa,\xi)\to\frac{\kappa_j^2}{\kappa^2}, \tag{21}$$

and the integrand I in Eq. (19) behaves as

$$I\to S(\xi\kappa,\mathbf{x})\xi^{N[n+(1/2)]-3}\prod_{i=1}^{N}\left[\frac{\kappa_i}{\kappa}\right]^{-n-(1/2)}. \tag{22}$$

In this relation, the same scaling parameter $\xi\kappa$ applies in all directions for the source field. The exponent of ξ is increasingly dominant as N increases. This weighting factor is plotted in Fig. 2 for $N=2$ and 3. Next, let us examine the ''diagonal'' terms for which all κ_j's are equal. Then, $\kappa_j=\kappa$ and the integrand reduces to

$$I=S(\xi\kappa_j,\mathbf{x})\xi^{N[n+(1/2)]-3}. \tag{23}$$

Thus, we see that the factor $\xi^{N[n+(1/2))]-3}$ applies for the whole interval $0<\xi<1$ along this line, sharply selecting the contributions from the spectral range $\xi\sim 1$. Directional effects will manifest themselves when the κ_j's are not equal, but the smallest κ_j will always dominate the scalar factor κ, so that κ_j/κ will never be small. Further discussion will be facilitated by the change of variables suggested by Eq. (17).

III. ACTION-ANGLE VARIABLES

A. The diffusion equation

The manipulations from the previous section indicate that a change of variables from κ_j to something proportional to $1/\kappa_j^2$ might simplify all relations derived from the compatibility condition. For the execution of this idea, the broader context of the diffusion equation

$$\frac{\partial u'}{\partial t}=\nu\frac{\partial^2 u'}{\partial x_j^2} \tag{24}$$

is helpful. Following steps similar to those in Sec. II, the compatibility condition was used[9] to obtain the characteristic representation of diffusion. In the one-dimensional case, time t is the characteristic parameter, and we have

$$\frac{d\kappa}{dt}=\nu\kappa^3, \tag{25}$$

and

$$\frac{du}{dt}=-\nu\kappa^2\left(n+\frac{1}{2}\right)u. \tag{26}$$

These characteristic equations can also be derived as the canonical equations corresponding to the Hamiltonian

$$H(K,U,\kappa,u)=\nu\kappa^3K-(n+\tfrac{1}{2})\nu\kappa^2uU. \tag{27}$$

Here, U and K are the configuration variables canonically associated with the momenta u and κ, respectively. This result was extended to N-dimensional diffusion, and explicit solutions for the trajectories were presented. It was shown that trajectories initiating from a given scale escape to infinity in a finite time, with the smallest scales escaping first; in this sense, irreversibility is consistent with the Hamiltonian structure. It was also shown that a Hamiltonian can be written to include nonlinear terms, such as convection, reaction or pressure terms.

For the change in variables, we will consider the general case of fields with M components in an N-dimensional space. Introducing the change of variables

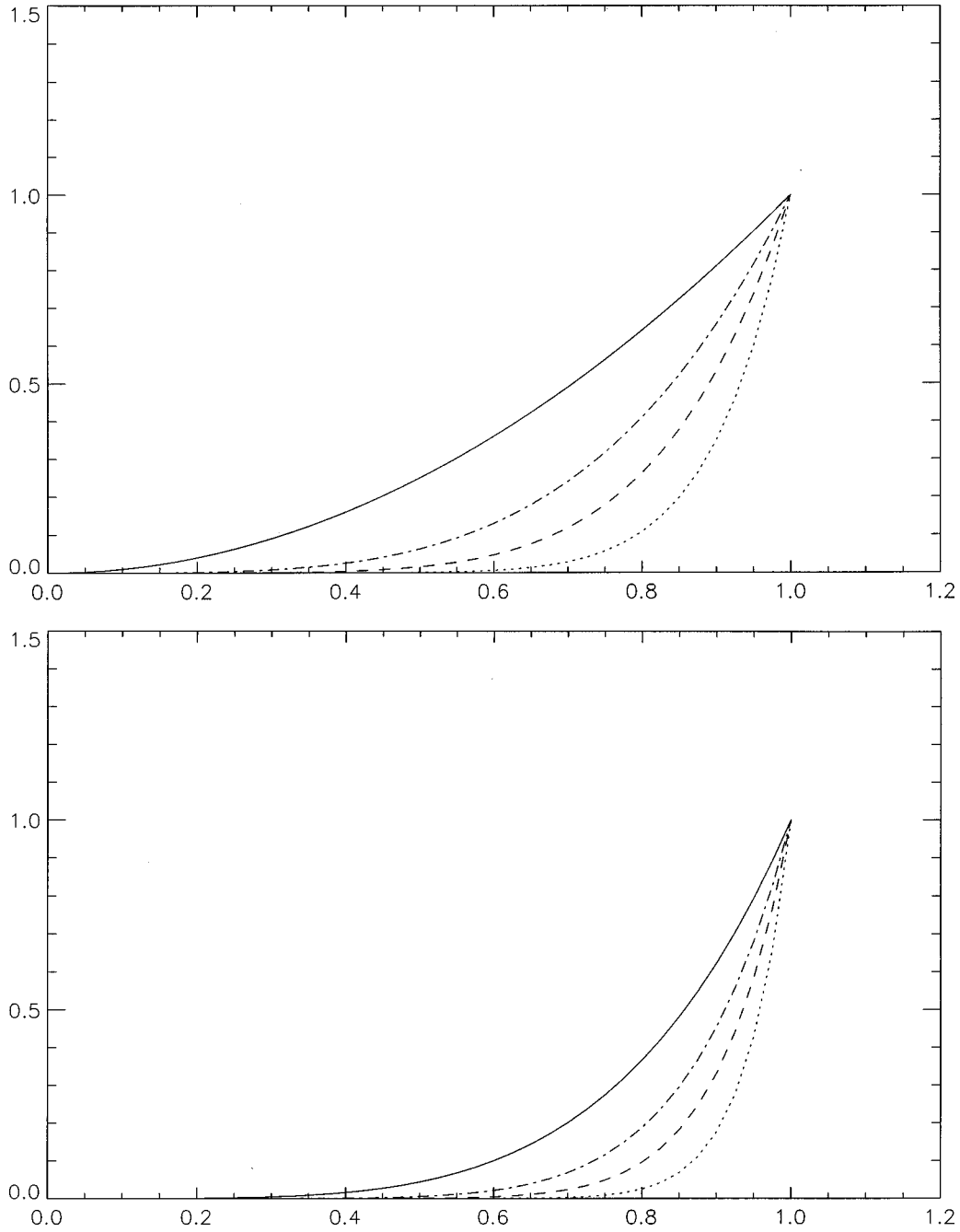

FIG. 2. Two-dimensional and three-dimensional case: relative weight of contributions to pressure at a given scale κ from sources at relative scale ξ, according to Eq. (22). Solid line: $n=2$; dash-dotted line: $n=3$; dashed line: $n=4$; dotted line: $n=6$.

$$\zeta_j = \frac{1}{2\kappa_j^2} \tag{28}$$

in the compatibility relation Eq. (14), elementary substitutions lead us to factor out the asymptotic slope (large κ or small ζ) from the transformed field:

$$u_i(\kappa_j, t; x_m) = f_i(\zeta_j, t; x_m) \prod_j \zeta_j^{(n/2)+(1/4)}. \tag{29}$$

Accordingly, a steep drop-off of f_i towards large scales can be expected. In fact, starting from the smallest scales ($\zeta\sim 0$), the asymptotic slope[11] of the wavelet-transformed field u_i is factored out to give f_i. With these variables, the compatibility equation becomes

$$\sum_{i=1}^{N} \frac{\partial^2 f_m}{\partial x_i^2} = \sum_{j=1}^{N} \frac{\partial f_m}{\partial \zeta_j}. \tag{30}$$

Since the compatibility condition is at the root of the characteristic equations for the diffusion and Poisson equations, considerable simplifications can be expected with these new variables.

The manipulations are best presented in the context of a canonical transformation.[12] We start from the configuration variables K_j and U_i, and their associate momenta κ_j and u_i. Then, let

$$\alpha = \frac{n}{2} + \frac{1}{4} \tag{31}$$

for short. Let us call Z_j and F_i the new configuration variables, and ζ_j and f_i the associated momenta as defined above. This change of variable will be a canonical transformation if there exists a generating function $G(\kappa_j, u_i, Z_j, F_i)$ such that

$$\zeta_j = \frac{\partial G}{\partial Z_j}, \quad f_i = \frac{\partial G}{\partial F_i}, \quad K_j = \frac{\partial G}{\partial \kappa_j}, \quad \text{and } U_i = \frac{\partial G}{\partial u_i}. \tag{32}$$

There is no difficulty in finding

$$G(\kappa_j, u_i, Z_j, F_i) = \sum_j \frac{1}{2\kappa_j^2} Z_j + 2^{\alpha N} \prod_j \kappa_j^{2\alpha} \sum_i u_i F_i. \tag{33}$$

The canonical transformation so defined yields the Hamiltonian

$$H(\zeta_j, f_i, Z_j, F_i) = -\nu \sum_j Z_j. \tag{34}$$

In these coordinates, the only active canonical equations are

$$\frac{d\zeta_j}{dt} = \frac{\partial H}{\partial Z_j} = -\nu, \tag{35}$$

which indicate that

$$\zeta_j(t) = \zeta_j(0) - \nu t. \tag{36}$$

This linear ''motion'' towards negative ζ_j's corresponds to the escape towards large κ_j's discussed elsewhere.[10] The solution for negative ζ_j's has no clear physical meaning and should be discarded.

The other variables remain unchanged under the action of diffusion. First and foremost, for the wavelet transform of the field, we have

$$\frac{df_i}{dt} = \frac{\partial H}{\partial F_i} = 0. \tag{37}$$

From the solution presented earlier,[10] we see that the factoring out of the asymptotic tail of u_i at large κ_j leads to a constant value of f_i in the characteristic representation. When we eliminate t between the two momenta, the initial local spectrum $f_i(\zeta_j, 0; x_m)$ drifts with invariant shape towards small scales (small ζ_j).

The configuration variables Z_j and F_i complete the canonical structure without affecting the diffusion itself. We have:

$$\frac{dF_i}{dt} = -\frac{\partial H}{\partial f_i} = 0 \tag{38}$$

and

$$\frac{dZ_j}{dt} = -\frac{\partial H}{\partial z_j} = 0. \tag{39}$$

The frozen coordinates Z_j, F_i and f_i are similar to those encountered in the ''action-angle'' formulation of some mechanical problems.[12]

B. Reformulation of the pressure solution

The solution of the Poisson equation is also simplified with these variables. Let us call f_p and f_S the action-angle transforms of the pressure field p' and its sources S', respectively. The Poisson equation (6) and the compatibility equation (30) give

$$\sum_i \frac{\partial f_p}{\partial \zeta_i} = f_s(\zeta_j, \mathbf{x}). \tag{40}$$

In characteristic form, this becomes

$$\frac{df_s(\zeta_j(s), \mathbf{x})}{ds} = f_s(\zeta_j, \mathbf{x}), \tag{41}$$

with

$$\frac{d\zeta_i}{ds} = -1. \tag{42}$$

The solution is easily obtained in the form

$$f_p(\zeta_j, \mathbf{x}) = -\int_s^\infty f_s(\zeta_j(\sigma), \mathbf{x}) d\sigma. \tag{43}$$

The elimination of the characteristic parameter is achieved with the introduction of the scalar scale factor

$$\zeta = \frac{1}{N}\sum_i \zeta_i \tag{44}$$

matching Eq. (17), and its relative value

$$X = \frac{\zeta(\sigma)}{\zeta(s)} = \frac{1}{\xi} \tag{45}$$

matching Eq. (12). It is easy to see that $\zeta/\zeta_i \geqslant 1/N$ for any given i. Substitution gives the solution of the Poisson equation in the form

$$f_p(\zeta_j, \mathbf{x}) = \zeta \int_1^\infty f_s\left(\left[1 + \frac{\zeta}{\zeta_j}(X-1)\right]\zeta_j, \mathbf{x}\right) dX. \tag{46}$$

In comparison with Eq. (19), we note the elimination of the scale and directional factors in the integrand. Contribution from the sources at the selected location and nearby scales retains a form comparable to that seen in Sec. II, including directional effects, but all other scale contributions are captured by the prefactor ζ, proportional to κ^{-2}.

The directional effects are also simpler to gage in this representation. With the ζ_j's equal in all directions (''diagonal'' terms), the sources are evaluated as $f_s(X\zeta_j)$, with the same factor in all

directions as in Sec. II B. It was noted in Sec. II A that the wavelet scale must be sufficient to cover both the source (in f_s) and the point of observation (in f_p). If we take this scale as reference, the source could have dominant components both smaller and larger. If smaller, the directional factor ζ/ζ_j is large, and is compounded by the factor $(X-1)$, so that the smallest scales sources are not observable as small scale pressure except in their immediate vicinity. On the other hand, for (directional) scales much larger than the distance to the source, we could have $\zeta/\zeta_j \sim \frac{1}{3}$ and contributions from source terms moderately smaller than the scale of observation. Therefore, the simple one-dimensional result is modified to include a finite vicinity around ζ_j, in addition to all larger values, in the contributions from the source to the integral.

Also, near-singularities of the source field will be somewhat smoothed out in the integration. Here the distinction[13] should be made between mathematical singularities (e.g., the x^{-2} associated with a line vortex, which is not integrable) and the near-singularities associated with arbitrarily large (but finite) Reynolds number. The balance between steepening gradients in near-singularities and the smoothing effects of viscous dissipation, described by the Navier–Stokes dynamics in wavelet space, will be addressed in the future.

IV. DISCUSSION

Wavelet tools show great promise in the understanding and numerical prediction of intermittent and nonhomogeneous phenomena. Particularly with the modern focus on computer simulation and on data storage and transmission, the emphasis in the literature is overwhelmingly on the use of orthonormal wavelets and wavelet packets, and the success of this line of work is undeniable.[6–8] The present work represents one formal step on a longer route toward numerical work, a route in which ''better'' wavelets are intended to facilitate primarily the physical formulation rather than the computer implementation. The underlying rationale is that a better formulation of the intermittent multiscale physics of turbulence might have a natural numerical counterpart.

On the surface, the continuous wavelet transforms are very uneconomical. The doubling of the number of independent variables makes a computational implementation of the equations[9] unrealistic for any problem of technical relevance. On the other hand, the physics of several phenomena intertwined in the theory and modeling of fluid turbulence are captured in promising ways by the Hermitian wavelets. Their unique properties affect directly the spatially localized contribution of source terms to the resulting pressure fluctuations and the spectral content of these contributions. Furthermore, within the Hermitian class, the order n is known to be associated with the cancellation of moments of the mother wavelet up to order $(n-1)$; to correspond to an asymptotic spectral slope $\kappa^{-[n+(1/2)]}$ that can be resolved with the given wavelet;[11] and to be related to the evaluation of structure functions[14,15] and to the resolution of the singularity spectrum of multifractal fields.[16]

This paper represents progress on two counts. First, a better compromise between the remarkable algebraic simplicity of the pressure terms in Fourier space and their nonlocal representation in configuration space is achieved, with the potential to include spatial and spectral intermittency in turbulence models. Second, the Hamiltonian representation is simplified by the same canonical transformation used for the pressure term, so that viscous behavior becomes computationally trivial. At a more general level, something is learned about the measure of scales. The Euclidean metric has been adopted in the spatial and Fourier domain, and is reflected in the Green's function for the former, in shell-related manipulations for the latter. Alternatives to Eq. (17) did not prove successful, and one of the surprises of this research is that the measure embodied in Eq. (17) is the natural one associated with the characteristic formulation. This observation applies to Poisson, Laplace, diffusion and second-order wave problems equally. Unlike the Euclidean metric of Fourier space, the wavelet number norm is dominated by the smallest wavelet number component.

The impact of the change of variables on the nonlinear terms (convective and pressure sources) and of a discrete modeling of the Hermitian wavelet coefficients is currently being studied. Toward this goal, the redundancy built in the continuous wavelet transforms is one untapped asset. Careful consolidation of the wavelet coefficients within a Hamiltonian framework has the potential to help both physical understanding and computational worth.

ACKNOWLEDGMENT

This work initiated from Jean Paul Bonnet's challenge that any alternative to Fourier would be judged by its handling of the pressure term.

[1] W. D. McComb, *The Physics of Fluid Turbulence* (Oxford University, Oxford, 1990).
[2] D. C. Leslie, *Developments in the Theory of Turbulence* (Clarendon, Oxford, 1973).
[3] M. Lesieur, *Turbulence in Fluids* (Martinus Nijhoff, Dordrecht, 1987).
[4] Daubechies, *Ten Lectures on Wavelets* (S.I.A.M., Philadelphia, 1992).
[5] M. Farge, ''Wavelet transforms and their applications to turbulence,'' Annu. Rev. Fluid Mech. **24**, 395–457 (1992).
[6] R. Camussi and G. Guj, ''Orthonormal wavelet decomposition of turbulent flows: Intermittency and coherent structures,'' J. Fluid Mech. **348**, 177–199 (1997).
[7] M. Farge, N. Kevlahan, V. Perrier, and E. Goirand, ''Wavelets and turbulence,'' Proc. IEEE **84**, 639–669 (1996).
[8] S. Grossman and M. Löhden, ''Wavelet analysis of Navier–Stokes flow,'' Z. Phys. B **100**, 137–147 (1996).
[9] J. Lewalle, ''Wavelet transforms of some equations of fluid mechanics,'' Acta Mech. **104**, 1–25 (1994).
[10] J. Lewalle, ''Hamiltonian formulation for the diffusion equation,'' Phys. Rev. E **55**, 1590–1599 (1997).
[11] V. Perrier, T. Philipovitch, and C. Basdevant, ''Wavelet spectra compared to Fourier spectra,'' J. Math. Phys. **36**, 1506–1519 (1995).
[12] I. C. Percival and D. Richards, *Introduction to Dynamics* (Cambridge University Press, Cambridge, 1982).
[13] N. K.-R. Kevlahan and J. C. Vassilicos, ''The space and scale dependencies of the self-similar structure of turbulence,'' Proc. R. Soc. London, Ser. A **447**, 341–363 (1994).
[14] V. Perrier and C. Basdevant, ''Besov norms in terms of the continuous wavelet transform: Application to structure function,'' Math. Methods Appl. Sci. **6**, 649–664 (1996).
[15] M. Vergassola and U. Frisch, ''Wavelet transforms of self-similar processes,'' Physica D **54**, 58–64 (1991).
[16] E. Bacry, J. F. Muzy, and A. Arneodo, ''Singularity spectrum of fractal signals from wavelet analysis: Exact results,'' J. Stat. Phys. **70**, 635–674 (1993).

Oscillation spaces: Properties and applications to fractal and multifractal functions

Stéphane Jaffard[a)]
Département de Mathematiques, Universite Paris XII, 61 Avenue du General de Gaulle, 94010 Creteil Cedex, France

(Received 25 January 1998; accepted for publication 3 April 1998)

We establish a wavelet characterization of the p oscillation, and give two applications: a wavelet formula for the fractal dimension of graphs and a multifractal formalism for ''chirp-type'' Hölder singularities. © *1998 American Institute of Physics.* [S0022-2488(98)01508-4]

I. INTRODUCTION

The global regularity of a function is usually measured either by determining which Sobolev spaces it belongs, or, when more precise information is needed, by determining to which Besov or Triebel–Lisorkin spaces it belongs. Another way to measure global regularity is to estimate the p variation. If one is only interested in determining the range of ps for which the p variation is bounded, both approaches are essentially identical, as shown by Bergh and Peetre (see Ref. 1) who proved the following sharp imbeddings between V_p (the space of functions with bounded p variation) and Besov spaces:

$$\dot{B}_p^{1/p,\inf(1,p)} \hookrightarrow V_p \hookrightarrow \dot{B}_p^{1/p,\infty}. \tag{1}$$

Nonetheless, in several situations one wishes to determine how the p variation increases when one uses finer and finer grids in its definition. This will lead us to new functional spaces that we will study and relate to other new spaces which are defined by a simple condition on the wavelet coefficients. We will consider two situations in which these spaces are needed: the problem of determining the fractal dimension of graphs of functions, and a multifractal formalism for ''chirp-type'' Hölder singularities.

Let λ be a dyadic cube included in $\mathbf{R}^d$; if f is a continuous function defined on $\mathbf{R}^d$; let

$$\operatorname{osc}(f,\lambda) = \sup_{x\in\lambda} f(x) - \inf_{x\in\lambda} f(x) \tag{2}$$

denote the oscillation of f on the cube λ. The p oscillation of f at scale j is defined by

$$\operatorname{Osc}_p(f,j) = \sum_k \left(\operatorname{osc}(f,\lambda_{j,k})\right)^p,$$

where $\lambda_{j,k}$ is the dyadic cube $k2^{-j}+2^{-j}[0,1]^d$. Note that when $p=1$ one uses the term ''oscillation'' instead of ''1-oscillation.'' For analyzing smooth functions, the definition of the p oscillation using (2), which takes into account only first-order differences, has to be modified: one uses instead nth-order differences, which are defined by induction as follows:

$$\Delta_f^1(t,h) = f(t+h) - f(t),$$

$$\Delta_f^n(t,h) = \Delta_f^{n-1}(t+h,h) - \Delta_f^{n-1}(t,h);$$

and, in the definition of the p oscillation, (2) is replaced by

a)Electronic mail: jaffard@cmla.ens-cachan.fr

 © 1998 American Institute of Physics

$$\mathrm{osc}(f,\lambda)=\sup_{[t,t+nh]\subset\lambda}|\Delta_f^n(t,h)|. \tag{3}$$

We will consider two problems where one has to estimate the order of magnitude of the p oscillation when $j\to+\infty$. This can be done by introducing the p-variation spaces $V^{s,p}$ defined by

$$f\in V^{s,p},\quad \text{if } \sup_{j\in\mathbf{Z}}2^{js}\left(\sum_k|\mathrm{osc}(f,\lambda_{j,k})|^p\right)^{1/p}\leqslant C \tag{4}$$

(one chooses here in the definition of the oscillation a fixed n larger than $[s]+1$); the $V^{s,p}$ seminorm is defined by the left-hand side of the inequality (these spaces are defined modulo polynomials of degree n). We can also consider the p-oscillation exponent,

$$\omega_p(f)=\liminf_{j\to\infty}\frac{\log(\mathrm{Osc}_p(f,j))}{\log(2^{-j})}.$$

Of course, $\omega_p(f)=\sup\{s:f\in V^{s/p,p}\}$.

These notions are closely related to the p variation: Recall that a function $f:[0,1]\to\mathbf{R}$ has a bounded p variation (we write $f\in V^p$) if

$$\inf\sum|f(t_{i+1})-f(t_i)|^p\leqslant C, \tag{5}$$

where the infimum is taken on all finite subdivisions of $[0,1]$. Clearly, if f belongs to V_p,

$$A_j=\sum_k\left|\sup_{\lambda_{j,k}}f(x)-\inf_{\lambda_{j,k}}f(x)\right|^p\leqslant C \tag{6}$$

(take for subdivision the points where the supremum and the infimum are attained in each dyadic interval of length 2^{-j}). Thus, if f belongs to V_p, then $f\in V^{0,p}$. Conversely, suppose that there exists a positive ϵ such that $f\in V^{\epsilon,p}$. It follows that for any j, $A_j\leqslant C2^{-\epsilon j}$; let $(t_i)_{i=1,\ldots,n}$ be a finite subdivision; let J be such that each interval of length 2^{-J} contains at most two points of the subdivision. The sum on all indexes i such that t_{i+1} and t_i belong to the same dyadic interval of length 2^{-J} is bounded by A_J. The sum on all indexes i such that t_{i+1} and t_i belong to the same dyadic interval of length 2^{-J-1} but do not belong to the same dyadic interval of length 2^{-J} is bounded by $A_{J-1,\ldots}$; thus, since $A_j\leqslant C2^{-\epsilon j}$, $f\in V_p$. Hence the following imbeddings:

$$\forall\epsilon>0,\quad V^{\epsilon,p}\hookrightarrow V_p\hookrightarrow V^{0,p}. \tag{7}$$

Theorem 1 will give a wavelet characterization of the p-oscillation exponent. We will use a family of 2^d-1 smooth wavelets $\psi^{(i)}$, such that the $\psi^{(i)}$ and their partial derivatives have fast decay. The $2^{dj/2}\psi^{(i)}(2^jx-k)$ $(i=1,\ldots,2^d-1,j\in\mathbf{Z},k\in\mathbf{Z}^d)$ form an orthonormal basis of $L^2(\mathbf{R}^d)$. We will use a L^∞ normalization for wavelets, so that we write

$$f(x)=\sum_{i,j,k}C_{j,k}^{(i)}\psi^{(i)}(2^jx-k), \tag{8}$$

where $C_{j,k}^{(i)}=2^{dj}\int f(t)\psi^{(i)}(2^jt-k)dt$. We will from now on use the following simpler notations; λ and λ' will denote, respectively, the cubes $\lambda_{j,k}=k2^{-j}+[0,2^{-j}]^d$ and $\lambda_{j',k'}=k'2^{-j'}+[0,2^{-j'}]^d$, C_λ will denote the coefficient $C_{j,k}^{(i)}$, and ψ_λ will denote the wavelet $\psi^{(i)}(2^jx-k)$ (note that we ''forget'' to write the index i of the wavelet, which is of no consequence).

The following functional spaces will be closely related to the spaces $V^{s,p}$, and thus will allow us to characterize p-oscillation exponents.

Definition 1: Let $p>0$, and s, $s'\in\mathbf{R}$; a function f belongs to $\mathcal{C}_p^{s,s'}(\mathbf{R}^d)$ if its wavelet coefficients satisfy

$$\sup_{j\in\mathbf{Z}} 2^{sj}\left(\sum_k \sup_{\lambda'\subset\lambda} |C_{\lambda'}2^{s'j'}|^p\right)^{1/p} < \infty. \tag{9}$$

The left-hand side defines the $\mathcal{O}_p^{s,s'}(\mathbf{R}^d)$ *seminorm.*

Proposition 1 will imply that this definition is independent of the wavelet basis chosen.

Theorem 1: *The following imbeddings hold:*

$$\forall \epsilon>0, \quad \forall \eta>0, \quad C^\epsilon \cap V^{s+\eta,p} \hookrightarrow \mathcal{O}_p^{s,0} \hookrightarrow V^{s,p}; \tag{10}$$

thus, if f: $\mathbf{R}^d\to\mathbf{R}$ *belongs to* $C^\epsilon(\mathbf{R}^d)$ *for* $\epsilon>0$, *the p-oscillation exponent of* f is *given by*

$$\omega_p(f) = \sup\{s: f\in \mathcal{O}_p^{s/p,0}\} = \lim_{j\to+\infty} \sup \frac{\log(\Sigma_{\lambda\in\Lambda_j}\sup_{\lambda'\subset\lambda}|C_{\lambda'}|^p)}{j\ \log 2}. \tag{11}$$

The spaces $\mathcal{O}_p^{s,s'}$ are defined by conditions on the wavelet coefficients; therefore the first point to check is that this definition is independent of the wavelet basis chosen. One way to do it, following Ref. 2, is to check that condition (9) is invariant under the action of infinite matrices which belong to the algebras $\mathcal{M}^\gamma$ for γ large enough, which is the purpose of Proposition 1 below; these algebras are defined as follows: an infinite matrix $A(\lambda,\lambda')$ indexed by the dyadic cubes belongs to $\mathcal{M}^\gamma$ if

$$|A(\lambda,\lambda')| \leqslant \frac{C2^{-(d/2+\gamma)(j-j')}}{(1+(j-j')^2)(1+2^{\inf(j,j')}\ \mathrm{dist}(\lambda,\lambda'))^{d+\gamma}}.$$

Matrices of operators which map a wavelet basis onto another belong to these algebras, and more generally matrices (on wavelet bases) of pseudodifferential operators of order 0, such as the Hilbert transform in dimension 1, or the Riesz transforms in higher dimensions; see Ref. 2. We denote by $\mathcal{O}p(\mathcal{M}^\gamma)$ the space of operators whose matrix on a wavelet basis belongs to $\mathcal{M}^\gamma$. The following proposition will be proved in Sec. VI.

Proposition 1: If $\gamma\geqslant\sup(|s|,|s'|)$, *the operators which belong to* $\mathcal{O}p(\mathcal{M}^\gamma)$ *are continuous on* $\mathcal{O}_p^{s,s'}(\mathbf{R}^d)$.

The fractional derivative or order α [which we denote by $(-\Delta)^{\alpha/2}$] can be written as the product of a diagonal operator on a wavelet basis (each ψ_λ is multiplied by $2^{\alpha j}$) and an operator in $\mathcal{O}p(\mathcal{M}^\gamma)$. Therefore, as a consequence of Proposition 1, $\|(-\Delta)^{\alpha/2}f\|_{\mathcal{O}_p^{s,0}}$ and $\|f\|_{\mathcal{O}_p^{s,\alpha}}$ are equivalent norms; it follows that

$$\omega_p((-\Delta)^{\alpha/2}f) = \sup\{s: f\in\mathcal{O}_p^{s/p,\alpha}\}. \tag{12}$$

Before proving Proposition 1 and Theorem 1, we first point out a consequence of Theorem 1 concerning the computation of the fractal dimension of graphs. Let K be a bounded subset of $\mathbf{R}^{d+1}$. Denote by $N(K,j)$ the number of dyadic cubes of size 2^{-j} necessary to cover K. The upper box dimension of K (also called the fractal dimension) is, by definition,

$$\overline{\dim}_b(K) = \lim_{j\to\infty} \sup \frac{\log N(K,j)}{j\ \log 2}.$$

If K is the graph of a compactly supported continuous function, $N(K,j)$ is related to the oscillation of f; indeed, see Ref. 3,

$$N(\mathrm{Graph}(f),j) \sim 2^{dj} + 2^j\ \mathrm{Osc}_1(f,j) \tag{13}$$

[where we use the standard notation $A\sim B$ if $A=\mathcal{O}(B)$ and $B=\mathcal{O}(A)$]. The following corollary is an immediate consequence of (13) and Theorem 1.

Corollary 1: Let f: $\mathbf{R}^d\to\mathbf{R}$ *be a compactly supported function which belongs to* $C^\epsilon(\mathbf{R}^d)$ *for an* $\epsilon>0$;

$$\overline{\dim_b}(\mathrm{Graph}(f))=\sup(d,1-\sup\{s:f\in\mathcal{O}_1^{s,0}\})=\sup\left(d,1+\limsup_{j\to\infty}\frac{\log(\Sigma_\lambda \sup_{\lambda'\subset\lambda}|C_{\lambda'}|)}{j\log 2}\right).$$

Note that Proposition 1 together with Corollary 1 imply that, if $f\in C^\epsilon(\mathbf{R}^d)$ for an $\epsilon>0$, the fractal dimension of the graph of f is left invariant by operators such as the Hilbert and Riesz transforms, or, more generally, by pseudodifferential operators of order 0 with smooth coefficients. This result shows that the fractal dimension of the graph of a function f is in some sense a ''robust'' global regularity criterium for f.

The wavelet characterization given by Corollary 1 should find applications in rugosimetry. Indeed the fractal dimension of a surface has been shown to be a pertinent way to model the notion of *rugosity*; see Ref. 4; note also that in Ref. 5 other generalizations of Sobolev spaces are introduced following a similar motivation. Therefore finding a numerically stable algorithm to measure this fractal dimension became an important issue. Numerical algorithms based on the oscillation are discussed in Ref. 4. Alternative algorithms based on the wavelet characterization of Corollary 1 might prove numerically more stable, since they would not be based directly on the pointwise value of the profile, but on wavelet coefficients, which are averaged quantities, and, as such, are less sensitive to noise.

The idea of relating the fractal dimension of graphs to wavelet coefficients can be traced back to Ref. 6 where Deliu and Jawerth proposed a formula relating the oscillation of f to the Besov spaces $B_1^{s,\infty}$ to which f belongs locally: If $\alpha=\sup\{s:f\in B_1^{s,\infty}\}$, they asserted that $\overline{\dim_b}(\mathrm{Graph}(f))$ is equal to $d+1-\alpha$ [the relationship with wavelet coefficients is a consequence of the wavelet characterization of the $B_1^{s,\infty}(\mathbf{R}^d)$ norm: $\sup_j\Sigma_k 2^{(s-d)j}|C_\lambda|$]. Kamont and Wolnik constructed a counterexample in Ref. 3; the comprehension of this counterexample is actually the starting point which led us to the definition of the spaces $\mathcal{O}_p^{s,s'}$.

In Sec. III, we will propose another use of the spaces $\mathcal{O}_p^{s,s'}$. We will show how a multifractal formalism for chirps can be based of the determination of the spaces $\mathcal{O}_p^{s,s'}$ to which the function f belongs. Theorem 1 will be proved in two steps. In Sec. IV we suppose that the wavelet basis we use is compactly supported, and we obtain an upper and lower bound of the p oscillation. In Sec. V, we prove that, if f satisfies a uniform regularity condition, the upper and lower bounds are of the same order of magnitude, and in Sec. VI we prove that the definition of the spaces $\mathcal{O}_p^{s,s'}$ is independent of the wavelet basis chosen. In Sec. VII we apply these results on the example of lacunary wavelet series already considered in Ref. 7. Indeed, they will supply new counterexamples to the formula proposed in Ref. 6, and the standard multifractal formalism will also be wrong in this case, whereas the use of the spaces $\mathcal{O}_p^{s,s'}$ will allow us to compute easily the fractal dimension of their graphs, and the multifractal formalism for chirps that we construct in Sec. III yield their correct spectrum of chirps, and, as a consequence, their correct spectrum of singularities.

II. IMBEDDINGS WITH BESOV SPACES

Let us now mention a few imbeddings which hold locally. First, clearly, if $s_1\leqslant s_2$ and $s_1'\leqslant s_2'$,

$$\mathcal{O}_p^{s_2,s_2'}\subset\mathcal{O}_p^{s_1,s_1'}.$$

Using the normalization we chose for wavelets, $f\in B_p^{s,q}$ if (see Ref. 2)

$$2^{-nj/p}\left(\sum_k |C_\lambda 2^{js}|^p\right)^{1/p}=\epsilon_j,\quad \text{with}\quad \epsilon_j\in l^q.$$

Since $\sup_{\lambda'\subset\lambda}|C_{\lambda'}2^{j's'}|\geqslant|C_\lambda 2^{js'}|$, it follows that

$$\mathcal{O}_p^{s,s'}\hookrightarrow B_p^{s'+s+n/p,\infty}.$$

If $j'\geqslant j$, for any positive α, $2^{(j'-j)\alpha}\geqslant 1$; therefore

$$\sup_j \sum_k \sup_{\lambda' \subset \lambda} |C_{\lambda'} 2^{j's'} 2^{sj}|^p \leqslant \sup_j \sum_k \sup_{\lambda' \subset \lambda} |C_{\lambda'} 2^{j'(s'+\alpha)} 2^{(s-\alpha)j}|^p;$$

hence the imbeddings,

$$\forall \alpha > 0, \quad \mathcal{O}_p^{s-\alpha, s'+\alpha} \hookrightarrow \mathcal{O}_p^{s,s'} \hookrightarrow \mathcal{O}_p^{s+\alpha, s'-\alpha} \hookrightarrow B_p^{s+s'+n/p,\infty}. \tag{14}$$

Conversely, if $f \in B_p^{s,p}$,

$$\sum_{k'} |C_{\lambda'} 2^{j'(s-n/p)}|^p = \epsilon_{j'}, \quad \text{with} \quad \epsilon_{j'} \in l^1,$$

thus for any j,

$$\sum_{j' \geqslant j} \sum_{k'} |C_{\lambda'} 2^{j'(s-n/p)}|^p \leqslant C.$$

In the sum over k', we group the terms which correspond to a dyadic cube λ' included in a given dyadic cube λ; thus

$$\sum_{\lambda} \sum_{j' \geqslant j} \sum_{\lambda' \subset \lambda} |C_{\lambda'} 2^{j'(s-n/p)}|^p \leqslant C,$$

hence, *a fortiori*,

$$\sum_{\lambda} \sup_{\lambda' \subset \lambda} |C_{\lambda'} 2^{j'(s-n/p)}|^p \leqslant C,$$

and thus $f \in \mathcal{O}_p^{0, s-n/p}$, hence the imbeddings $B_p^{s+n/p,p} \hookrightarrow \mathcal{O}_p^{0,s} \hookrightarrow B_p^{s+n/p,\infty}$. Using (14) we obtain more generally,

$$\forall s > 0, \quad \forall s' \in \mathbf{R}, \quad \forall p > 0, \quad B_p^{s+s'+n/p,p} \hookrightarrow \mathcal{O}_p^{s,s'} \hookrightarrow B_p^{s+s'+n/p,\infty}. \tag{15}$$

Let $H^{s,p}$ denote the space of functions which belong to L^p and whose fractional derivative of order s belongs to L^p. The Besov spaces $B_p^{s,q}$ or the Triebel–Lisorkin spaces $F_p^{s,q}$ can be seen as a slight variation on the definition of Sobolev spaces in the sense that

$$\forall \epsilon > 0, \quad \forall q, \quad H^{s+\epsilon,p} \hookrightarrow B_p^{s,q} \hookrightarrow H^{s-\epsilon,p}$$

and

$$\forall \epsilon > 0, \quad \forall q, \quad H^{s+\epsilon,p} \hookrightarrow F_p^{s,q} \hookrightarrow H^{s-\epsilon,p}.$$

Similarly, (1), together with the above imbeddings, shows that the spaces V_p are also a variation on the definition of Besov or Sobolev spaces, and (15) shows that, for $s \geqslant 0$, it is also the case for oscillation spaces.

On the contrary the spaces $\mathcal{O}_p^{s,s'}$ for $s < 0$ cannot be sharply imbedded between Sobolev spaces, and thus are new spaces of a really different nature. The use of these spaces proposed in the next section is an illustration of this fact: They allow one to capture oscillatory behaviors which are left undetected by Sobolev or Besov spaces.

Another striking difference between the spaces $\mathcal{O}_p^{s,s'}$ and the Besov spaces is shown by the fact that Besov spaces are invariant under a permutation of the wavelet coefficients at the same scale, which is clearly not the case for oscillation spaces. *Such permutations allow us to modify the Hölder singularities, hence the spectrum of singularities of the function, whereas the oscillation spaces take into account the geometric disposition of the wavelet coefficients.*

III. A MULTIFRACTAL FORMALISM FOR CHIRPS

We now consider a situation where the full range of oscillation spaces are needed. Let $f:\mathbf{R}^d \to \mathbf{R}$ be a bounded function and let $x_0 \in \mathbf{R}^d$; f is $C^\alpha(x_0)$ $(\alpha \geqslant 0)$ if there exists a polynomial P of degree at most $[\alpha]$ and a constant C such that, if $|x-x_0| \leqslant 1$, $|f(x)-P(x-x_0)| \leqslant C|x-x_0|^\alpha$. The *Hölder exponent* of f at x_0 [which will be denoted by $h_f(x_0)$] is, by definition, the supremum of all values of α such that f is $C^\alpha(x_0)$. Characterizing the pointwise regularity of a function with the sole Hölder exponent yields a rather poor information since it does not describe the more or less oscillatory behavior of the function near the point x_0. This oscillatory behavior is properly modeled with the help of the following definition, which was introduced by Meyer; see Ref. 8 or 9.

Definition 2: Let f be a function in $L^\infty_{loc}(\mathbf{R}^d)$; f is called a (h,β)-type chirp at x_0 if $\forall n \in \mathbf{N}$, f can be written as a finite sum of partial derivatives of order n of functions which belong to $C^{h+n(1+\beta)}(x_0)$.

The simplest example of a one-dimensional (h,β)-type chirp at x_0 is supplied by the function

$$|x-x_0|^h \sin\left(\frac{1}{|x-x_0|^\beta}\right). \tag{16}$$

The interior of the set of points (h,β) such that a function f is a (h,β)-type chirp at x_0 is always a domain of the form $h<h_f(x_0)$, $\beta<\beta_f(x_0)$; see Ref. 7. The non-negative real number $\beta_f(x_0)$ is called the *chirp exponent at* x_0.

Chirps are known to exist in a variety of signals, ranging from gravitational waves to ultrasounds emitted by bats. Their existence is an open problem in other signals, such as the velocity of a turbulent fluid. One isolated chirp is usually hard to detect, see Refs. 10 and 11, and establishing a multifractal formalism adapted to chirps could be a way to show their presence on a statistical ground, without having to find precisely the location of each of them.

Let us first recall a few basic points concerning multifractal functions and the standard multifractal formalism adapted to the sole Hölder exponents. When considering a multifractal function f, we expect the set of points where f has a given Hölder exponent h to be a fractal set; we have thus to deal with a whole collection of fractal sets indexed by the parameter h. A relevant mathematical notion is the *spectrum of singularities* $d(h)$ of the function f. For each $h \geqslant 0$, $d(h)$ is the Hausdorff dimension of the set of points x_0 where the Hölder exponent of f is h [we take, as usual, $\dim(\varnothing) = -\infty$.) It is clearly impossible to compute numerically the spectrum of a signal following the definition; indeed, it involves the successive determination of several intricate limits which would yield enormous, totally unstable computations. The only method is to derive it using only averaged quantities (which should be numerically stable) extracted from the signal. Therefore Frisch and Parisi, see Ref. 12, introduced a formula which allows us to compute the spectrum of singularities of a signal by a Legendre transform of some numerically easily computable function. An alternative formula based on the local maxima of the wavelet transform was proposed by Arneodo, Bacry, and Muzy; see Ref. 13. The *Structure Function* method of Frisch and Parisi requires first the computation of

$$S_p(l) = \int_{\mathbf{R}^d} |f(x+l)-f(x)|^p dx.$$

Supposing that the scaling $S_p(l) \sim |l|^{\tau(q)}$ holds, the spectrum of singularities is computed using the formula

$$d(h) = \inf_p (ph - \tau(p) + d). \tag{17}$$

Mathematically, this formula, and the wavelet-based formulas, can be shown to yield the same exponent $\tau(p)$ which can also be deduced from the functional, Besov-type spaces to which the function f belongs; see Ref. 14:

$$\tau(p) = \sup\{s : f \in B^{s/p,p}_\infty\}.$$

A careful study of the wavelet formulas shows that this formalism is only adapted to ''cusp-type'' singularities (i.e., singularities with oscillation exponent $\beta=0$); see Ref. 15.

If we want to study chirps located in a signal, we are naturally led to define a *Spectrum of oscillating singularities* as follows.

Definition 3: The Spectrum of oscillating singularities $d(h,\beta)$ of a function f is the Hausdorff dimension of the set of points where f has the chirp exponents (h,β).

We will now establish by thermodynamic arguments (similar to those advocated in Refs. 13 and 12) a multifractal formalism which yields $d(h,\beta)$ from the knowledge of the p-oscillation exponents of f and its partial derivatives, or alternatively from its wavelet coefficients. This formula is thus motivated by two reasons: To obtain a more complete information about the Hölder singularities of the signal; and, by taking into account the ''chirp-type'' behavior explicitly in the construction of the multifractal formalism, to eliminate one cause of failure of the standard multifractal formalism, and therefore to obtain a formula with a wider range of validity.

Let

$$\zeta(p,s')=\liminf_{j\to+\infty}\frac{\log(\Sigma_k \sup_{\lambda'\subset\lambda}|C_{\lambda'}|^p 2^{s'j'})}{\log 2^{-j}}=\sup\{s: f\in \mathcal{C}_p^{s/p,s'/p}\}=\omega_p((-\Delta)^{s'/2p}f\ ).$$

If f has a chirp of exponents (h,β) at x_0, its wavelet coefficients are of the order of magnitude of $|k2^{-j}-x_0|^h$ near the curve $2^{-j}\sim|k2^{-j}-x_0|^{1+\beta}$ and decay fast away from this curve (see Ref. 16 or 8 for a precise mathematical statement). Our purpose is to estimate for each (h,β) the contribution of the chirps of exponents (h,β) to the quantity

$$\sum_{\lambda\in\Lambda_j}\sup_{\lambda'\subset\lambda}|C_{\lambda'}|^p 2^{s'j'}. \tag{18}$$

Consider a cube λ of size 2^{-j} which contains a chirp of exponents (h,β). The wavelet coefficients $C_{\lambda'}$ for $\lambda'\subset\lambda$ are negligible as long as $2^{-j'}\geqslant(2^{-j})^{1+\beta}$, i.e., as long as $j'\leqslant j(1+\beta)$. When $j'\sim j(1+\beta)$, for some values of k',

$$|C_{\lambda'}|\sim(2^{-j})^h\sim 2^{-j'h/(1+\beta)},$$

so that

$$\sup_{\lambda'\subset\lambda}(|C_{\lambda'}|^p 2^{s'j'})\sim 2^{-j'(hp/(1+\beta)-s')}\sim 2^{-j(hp-(1+\beta)s')}$$

[as long as $s'\leqslant ph/(1+\beta)$, otherwise the supremum is infinite]. The contribution of the chirps of exponents (h,β) to (18) is thus

$$2^{d(h,\beta)j}2^{-j(hp-(1+\beta)s')}=2^{-j(hp-(1+\beta)s'-d(h,\beta))}.$$

When $j\to+\infty$, the main contribution is obtained for the couple (h,β), realizing the infimum of $hp-(1+\beta)s'-d(h,\beta)$; hence, the heuristic formula,

$$\zeta(s',p)=\inf_{h,\beta} hp-(1+\beta)s'-d(h,\beta).$$

If $d(h,\beta)$ is a convex function, it follows that

$$d(h,\beta)=\inf_{s',p} hp-(1+\beta)s'-\zeta(s',p). \tag{19}$$

In Sec. VI we will verify that this formula allows us to recover the increasing part of the spectrum in the case of a lacunary wavelet series. Obtaining the decreasing part of the spectrum would correspond to an infimum in (19) obtained for negative values of p, which cannot be expected

from a mathematical approach of this type (in the case of the standard multifractal formalism, this problem already appeared, and was solved numerically by the use of the ''Wavelet Maxima Method;'' see Ref. 13).

If one is interested only in obtaining the spectrum of singularities $d(h)$, it can be recovered from (19): For a fixed h, we expect the value of β which brings the largest contribution to (19) to yield the right dimension $d(h)$; hence we expect the spectrum of singularities to be obtained by the formula

$$d(h)=\sup_{\beta} d(h,\beta)=\sup_{\beta}\inf_{s',p}(1+\beta)s'+hp-\zeta(s',p). \tag{20}$$

Of course, this formula is by no means equivalent to (17), as will be shown in Sec. VI on the example of a lacunary wavelet series.

A mathematical investigation of the range of validity of these formulas, together with numerical algorithms, are now being worked out in collaboration with Arneodo, Bacry, and Muzy, and will be the subject of a forthcoming paper.

IV. UPPER AND LOWER BOUNDS FOR THE *p* OSCILLATION

In the next two sections we compare the spaces $V^{s,p}$ with $\mathcal{O}_p^{s,0}$ when the wavelets are compactly supported. Let D be the smallest integer such that the 2^d-1 functions $\psi^{(i)}$ are all supported in $[0,D]^d$.

For each j, let

$$Q_j(x)=\sum_k C_\lambda \psi(2^j x-k).$$

If $p\leqslant 1$,

$$(\mathrm{osc}(f,\lambda))^p\leqslant\sum_{j'}(\mathrm{osc}(Q_{j'},\lambda))^p,$$

and, if $p>1$, using the Hölder inequality,

$$(\mathrm{osc}(f,\lambda))^p\leqslant\sum_{j'}((1+|j-j'|)\mathrm{osc}(Q_{j'},\lambda))^p.$$

We will now estimate the terms of this sum. We suppose that $p\leqslant 1$. [The case $p>1$ is similar, after adding everywhere the extra factor $(1+|j-j'|)$.]

If $j'<j$, using the mean-value theorem,

$$\mathrm{osc}(Q_{j'},\lambda)=\sup_{\lambda}Q_{j'}(x)-\inf_{\lambda}Q_{j'}(x)\leqslant C2^{-j}\sup_{\lambda}|\nabla Q_{j'}(x)|\leqslant C2^{-j}2^{j'}\sup_{D\lambda'\cap\lambda\neq\varnothing}|C_{\lambda'}|,$$

where $D\lambda'$ is the cube $k'2^{-j'}+2^{-j'}[0,D]^d$ [which contains the support of the wavelet $\psi(2^{j'}x-k')$]. Note that, if we use nth-order differences in the definition of the oscillation, using Taylor's formula, the upper bound becomes

$$\mathrm{osc}(Q_{j'},\lambda)\leqslant C2^{-n(j-j')}\sup_{D\lambda'\cap\lambda\neq\varnothing}|C_{\lambda'}|. \tag{21}$$

If $j'\geqslant j$,

$$\mathrm{osc}(Q_{j'},\lambda)\leqslant C\sup_{D\lambda'\cap\lambda\neq\varnothing}|C_{\lambda'}|.$$

Thus

$$\mathrm{Osc}_p(f,j)\leqslant C\sum_k \sum_{j'<j} 2^{-np(j-j')} \sup_{D\lambda'\cap\lambda\neq\varnothing} |C_{\lambda'}|^p + C\sum_k \sum_{j'\geqslant j} \sup_{D\lambda'\cap\lambda\neq\varnothing} |C_{\lambda'}|^p. \tag{22}$$

We will first bound the second sum; we can suppose that all the suprema in (22) are finite, and, after perhaps changing the constant C, that they are attained. Let λ be fixed and $j'\geqslant j$. Since $\lambda'\subset\lambda$, $D\lambda'\subset D\lambda$; thus, either the supremum is attained for $\lambda'\subset\lambda$, or it is attained in one of the D^d cubes included in $D\lambda$, of size 2^{-j}, so that

$$\sum_k \sum_{j'\geqslant j} \sup_{D\lambda'\cap\lambda\neq\varnothing} |C_{\lambda'}|^p \leqslant CD^d \sum_k \sum_{j'\geqslant j} \sup_{\lambda'\subset\lambda} |C_{\lambda'}|^p.$$

The same argument applied to the first sum of (22) yields the upper bound

$$C'\sum_{j'<j}\sum_k 2^{np(j'-j)}|C_{\lambda'}|^p, \tag{23}$$

where λ' is the dyadic cube of side $2^{-j'}$ which contains λ. Each cube λ' contains $2^{d(j-j')}$ cubes λ, so that (23) is bounded by

$$C\sum_{j'<j}\sum_{k'} 2^{(d-np)(j-j')}|C_{\lambda'}|^p.$$

Thus

$$\mathrm{Osc}_p(f,j)\leqslant C\sum_{j'<j}\sum_{k'} 2^{(d-np)(j-j')}|C_{\lambda'}|^p + C\sum_k \sum_{j'\geqslant j} \sup_{\lambda'\subset\lambda} |C_{\lambda'}|^p. \tag{24}$$

We will now obtain a lower bound for the p oscillation of f; since the wavelets have a vanishing integral,

$$\begin{aligned} |C_{\lambda'}| &= \left| \int f(x) 2^{dj'} \psi(2^{j'}x-k')\,dx \right| \\ &= \left| \int (f(x)-f(k'2^{-j'}))2^{dj'}\psi(2^{j'}x-k')\,dx \right| \\ &\leqslant C\ \mathrm{osc}(f,D\lambda') \end{aligned}$$

(if the oscillation is defined using nth-order differences, the argument is the same, using wavelets with at least n vanishing moments). Let j be fixed and $j'\geqslant j$ and λ' be a dyadic cube. Either there exists λ such that $D\lambda'\subset\lambda$, in which case

$$|C_{\lambda'}|\leqslant C\ \mathrm{osc}(f,\lambda),$$

or $D\lambda'$ intersects, at most, D^d cubes λ and

$$|C_{\lambda'}|\leqslant C\sum \mathrm{osc}(f,\lambda),$$

where the sum is taken on these D^d cubes. Thus,

$$\forall j \sum_k \sup_{\lambda'\subset\lambda} |C_{\lambda'}|^p \leqslant CD^d \sum_k \mathrm{osc}(f,\lambda)^p = CD^d\ \mathrm{Osc}_p(f,j). \tag{25}$$

The second imbedding in (10) follows.

V. COMPARISON OF THE UPPER AND LOWER BOUNDS

The right-hand side of (24) and the left-hand side of (25) need not be of the same order of magnitude. We will nonetheless see that it is the case if $f \in C^{\epsilon}(\mathbf{R}^d)$ for an $\epsilon > 0$. This uniform regularity condition implies that (cf. Ref. 2)

$$\exists C > 0, \quad \forall j', k', \quad |C_{\lambda'}| \leqslant C 2^{-\epsilon j'}. \tag{26}$$

For λ fixed, denote

$$A(\lambda) = \sup_{\lambda' \subset \lambda} |C_{\lambda'}|^p, \tag{27}$$

$$B(\lambda) = \sum_{j' \geqslant j} \sup_{\lambda' \subset \lambda} |C_{\lambda'}|^p. \tag{28}$$

Estimate (25) can be rewritten as

$$\sum_k A(\lambda) \leqslant C \ \mathrm{Osc}_p(f, j), \tag{29}$$

and estimate (24) can be rewritten as

$$\mathrm{Osc}_p(f, j) \leqslant C \sum_{j' < j} \sum_{k'} 2^{(d - np)(j - j')} |C_{\lambda'}|^p + C \sum_k B(\lambda);$$

thus, if ω is such that

$$\forall j \sum_{\lambda \in \Lambda_j} \sup_{\lambda' \subset \lambda} |C_{\lambda'}|^p \leqslant 2^{-\omega j}, \tag{30}$$

$$\mathrm{Osc}_p(f, j) \leqslant C \sum_{j' < j} 2^{(d - np)(j - j')} 2^{-\omega j'} + C \sum_k B(\lambda). \tag{31}$$

Let us now compare $A(\lambda)$ and $B(\lambda)$; of course,

$$A(\lambda) \leqslant B(\lambda). \tag{32}$$

Let $\mu (= \mu(j,k))$ be defined by $A(\lambda) = 2^{-\mu j}$. In (28) we can bound $\sup_{\lambda' \subset \lambda} |C_{\lambda'}|^p$ either by $C 2^{-\epsilon p j'}$, or by $2^{-\mu j}$. Thus

$$B(\lambda) \leqslant \sum_{j' = j}^{[\mu j / \epsilon p]} 2^{-\mu j} + C \sum_{j' = [\mu j / \epsilon p] + 1}^{\infty} 2^{-\epsilon p j'} = \left(\left[\frac{\mu j}{\epsilon p} \right] - j \right) 2^{-\mu j} + C 2^{-\epsilon p [\mu / \epsilon p] j} \leqslant \frac{C \mu j}{\epsilon p} 2^{-\mu j},$$

so that

$$B(\lambda) \leqslant C A(\lambda) |\log A(\lambda)|. \tag{33}$$

We consider now the right-hand side terms of (31).

Let $R > 0$; in the last sum of (31), we separate the terms for which $B(\lambda) \leqslant 2^{-Rj}$ [and the corresponding sum is bounded by $\mathcal{O}(2^{dj} 2^{-Rj})$], and the other terms for which, because of (32) and (33),

$$|\log A(\lambda)| \leqslant 2 |\log B(\lambda)| \leqslant 2 R j \log 2;$$

so that, using (33), we finally get

$$\sum_k B(\lambda) \leqslant CRj \sum_k A(\lambda) + \mathcal{O}(2^{(d-Rp)j}).$$

Thus $\forall \eta > 0$ and $\forall R > 0$, $\mathrm{Osc}_p(f,j)$ is bounded by

$$C \sum_{j'<j} 2^{(d-np)(j-j')} 2^{(-\omega+\eta)j'} + CRj \sum_k A(\lambda) + \mathcal{O}(2^{(d-Rp)j}). \tag{34}$$

The first imbedding in (10) is then a direct consequence of (29), (30), and (34), by taking η arbitrarily small and R large enough; hence Theorem 1 holds for compactly supported wavelets.

VI. INDEPENDENCE OF THE WAVELET BASIS

We now prove Proposition 1. Recall that we write $\lambda = \lambda_{j,k}$, $\lambda' = \lambda_{j',k'}, \ldots$. Let $D_{\lambda'} = \Sigma_{\lambda''} A(\lambda', \lambda'') C_{\lambda''}$ and

$$\omega_{\lambda,\lambda'} = \frac{2^{-(d/2+\gamma)|j-j'|}}{(1+(j-j')^2)(1+2^{\inf(j,j')}\ \mathrm{dist}(\lambda,\lambda'))^{d+\gamma}}.$$

The $D_{\lambda'}$ satisfy $|D_{\lambda'}| \leqslant C \Sigma_{\lambda''} C \omega_{\lambda',\lambda''} |C\lambda''|$; assume that the C_λ satisfy (9); we want to bound

$$\sum_{|\lambda|=2^{-j}} \sup_{\lambda' \subset \lambda} 2^{s'pj'} |D_{\lambda'}|^p \leqslant \sum_{|\lambda|=2^{-j}} \sup_{\lambda' \subset \lambda} 2^{s'pj'} \left(\sum_{\lambda''} \omega_{\lambda',\lambda''} |C_{\lambda''}| \right)^p.$$

We split the sum over λ'' into two terms depending on whether $\lambda'' \subset 3\lambda$ or $\lambda'' \not\subset 3\lambda$.

First case: $\lambda'' \subset 3\lambda$. Then, if $\gamma > |s'|$ (see Ref. 2),

$$\sum_{\lambda'' \subset 3\lambda} \omega_{\lambda',\lambda''} |C_{\lambda''}| \leqslant \Big(\sup_{\lambda'' \subset 3\lambda} 2^{s'j''} |C_{\lambda''}| \Big) \sum_{\lambda''} 2^{-s'j''} \omega_{\lambda',\lambda''} \leqslant C \sup_{\lambda'' \subset 3\lambda} (|C_{\lambda''}| 2^{s'j''}) 2^{-s'j'};$$

therefore

$$\sum_{|\lambda|=2^{-j}} \sup_{\lambda' \subset \lambda} 2^{s'pj'} \left(\sum_{\lambda'' \subset 3\lambda} \omega_{\lambda',\lambda''} |C_{\lambda''}| \right)^p \leqslant C 2^{s'pj''} \sum_{|\lambda|=2^{-j}} \sup_{\lambda'' \subset 3\lambda} |C_{\lambda''}|^p$$

$$\leqslant C 3^d 2^{s'pj''} \sum_{|\lambda|=2^{-j+1}} \sup_{\lambda'' \subset \lambda} |C_{\lambda''}|^p.$$

Second case: $\lambda'' \not\subset 3\lambda$. Then $2^{s'j'} \omega_{\lambda',\lambda''} \leqslant C 2^{s'j''} \tilde{\omega}_{\lambda,\lambda''}$, where $\tilde{\omega}_{\lambda,\lambda''}$ is defined as $\omega_{\lambda,\lambda''}$, except that we pick a $\tilde{\gamma} = \gamma - s'$; so that

$$\sum_{|\lambda|=2^{-j}} \sup_{\lambda' \subset \lambda} 2^{s'pj'} \left(\sum_{\lambda'' \not\subset 3\lambda} \omega_{\lambda',\lambda''} |C_{\lambda''}| \right)^p \leqslant C \sum_{|\lambda|=2^{-j}} \left(\sum_{\lambda''} 2^{s'j''} \omega_{\lambda,\lambda''} |C_{\lambda''}| \right)^p$$

$$\leqslant C \sup_{j'} 2^{s'pj'} \sum_{|\lambda'|=2^{-j'}} |C_{\lambda'}|^p$$

[the last inequality holds by continuity of the operators in $\mathcal{O}p(M^\gamma)$ on the spaces $B_p^{s,\infty}$; see Ref. 2]; hence Proposition 1 follows.

VII. LACUNARY WAVELET SERIES

We first recall the model of the lacunary wavelet series introduced in Ref. 7, and the main properties of the sample paths of these processes. The process $\mathbf{F}_{\alpha,\eta}$ we consider is defined on $[0,1]$ and depends on two parameters: $\eta \in (0,1)$ and $\alpha > 0$. It is defined by its wavelet coefficients as follows. For $j \geqslant 0$ we pick at random and independently $[2^{nj}]$ positions $k \in \{0,\ldots,2^j - 1\}$, and the

corresponding wavelet coefficients will be $C_{j,k}=2^{-\alpha j}$; the other wavelet coefficients are set to 0. The parameter η characterizes the *lacunarity* of the wavelet series, whereas α is related to the uniform regularity; indeed, it follows directly from the uniform regularity criterium of Ref. 2 that, if α is not an integer, $\mathbf{F}_{\alpha,\eta}\in C^{\alpha}(\mathbf{R})$, and this regularity is the best possible uniform regularity. Nonetheless, $\mathbf{F}_{\alpha,\eta}$ has much smoother sample paths at most locations. The Hölder exponent of $\mathbf{F}_{\alpha,\eta}$ is almost surely almost everywhere α/η. The spectrum of chirps of $\mathbf{F}_{\alpha,\eta}$ is supported by the segment $h=\alpha(\beta+1)$, $h\in[\alpha,\alpha/\eta]$ and is equal to $\eta h/\alpha$ on this segment; see Ref. 7.

Let us first compute the fractal dimension of the graph of $\mathbf{F}_{\alpha,\eta}$. Let J be fixed. We will estimate the oscillation of $\mathbf{F}_{\alpha,\eta}$ on each interval λ of length 2^{-J}. Let $\epsilon>0$. If $2^{(\eta-\epsilon)}\geqslant 2^{J}$, at least one wavelet coefficient of generation j is located in each interval with a probability of at least $1-e^{-Cj}$. Therefore

$$\mathrm{Osc}(\mathbf{F}_{\alpha,\eta},J)\geqslant 2^{J}2^{-\alpha J},\quad \text{for } j=\frac{J}{\eta-\epsilon},$$

so that

$$\mathrm{Osc}(\mathbf{F}_{\alpha,\eta},J)\geqslant 2^{J(1-\alpha/(\eta-\epsilon))},\quad \forall\epsilon>0,$$

and $\overline{\dim_b}(\mathrm{Graph}(\mathbf{F}_{\alpha,\eta}))\geqslant 2-\alpha/\eta$.

On the other side, if $j\geqslant J$,

$$\mathrm{osc}(\Delta_j,J)\leqslant 2^{-\alpha j}2^{\eta j},\quad \text{for } j\leqslant\frac{J}{\eta}$$

and

$$\mathrm{osc}(\Delta_j,J)\leqslant 2^{-\alpha j}2^{J},\quad \text{for } j\geqslant\frac{J}{\eta};$$

thus $\mathrm{Osc}(\mathbf{F}_{\alpha,\eta},J)\leqslant C2^{J(1-\alpha/\eta)}$. Hence, the following theorem holds.

Theorem 2: *The fractal dimension of the graph of the lacunary wavelet series* $\mathbf{F}_{\alpha,\eta}$ *is*

$$\overline{\dim_b}(\mathrm{Graph}(\mathbf{F}_{\alpha,\eta}))=\sup\left\{1,2-\frac{\alpha}{\eta}\right\}.$$

We can now check what the formula proposed in Ref. 6 would yield on this example. The norm of $\mathbf{F}_{\alpha,\eta}$ in $B_1^{s,\infty}$ is

$$\sup_j\sum_k 2^{(s-1)j}|C_{j,k}|=\sup_j\sum_k 2^{(s-1)j}2^{-\alpha j}2^{\eta j},$$

which is finite if $s<\alpha-\eta+1$. Thus, the formula of Ref. 6 yields a dimension $2-(\alpha-\eta+1)=1-\alpha+\eta$, which is correct only when $\eta=1$, i.e., when the wavelet series is no longer lacunary. Note that the counterexample proposed by Kamont and Wolnik is very similar: it is a deterministic lacunary series on the Schauder basis.

Let us consider in this example the effect of a fractional integration (or derivation if s is negative) of order s [when $(s+\alpha)/\eta\leqslant 1$] on the fractal dimension of the graph: It amounts to replacing the wavelet coefficients $2^{-\alpha j}$ by $2^{-(\alpha+s)j}$ (and to replace the orthogonal wavelet basis by a biorthogonal one, which is of no consequence here, since the passage matrix belongs to the algebras M^{γ}). Note that, for this example, the imbeddings (14) together with Corollary 1 imply that the dimension of the graph is lessened by s/η and not by s as might be expected. *This happens because the computation of the fractal dimension of a graph takes strongly into account the contribution of the chirps, and their Hölder exponent changes by more than s after a fractional integration or derivation of order s*. Note, for instance, that the fractal dimension of the graph of a Weierstrass function $\Sigma 2^{-\alpha j}\sin(2^{j}x)$ changes by only s after a fractional integration or derivation

of order s, which is consistent with the fact that they have no chirps. It is the same for the Riemann function $\Sigma n^{-2}\sin(n^2x)$; see Ref. 17, which has a too small number of chirps (they are countable; see Ref. 8).

Let us now check formula (19) of the multifractal formalism for chirps in this case. Let J be fixed. For each interval of length 2^{-J} we consider the first nonvanishing wavelet coefficient in the interval. For $2^{\eta J}$ of them, the size of this coefficient will be $2^{-\alpha J}$. For $2^{\eta(J+1)}$ of them, it will be $2^{-\alpha(J+1)},\ldots$, until J', such that $2^{\eta J'}=2^J$ for which the size will be $2^{-\alpha J'}$. After that level, all intervals will contain one coefficient. Thus, with probability arbitrarily close to 1 for J large enough,

$$\sum_k \sup_{\lambda'\subset\lambda}(2^{s'j'}|C_{\lambda'}|^p)=\sum_{j'=J}^{J/\eta}2^{\eta j'}2^{s'j'}2^{-\alpha j'p},\quad \text{if } s'\leqslant\alpha p$$

$$=+\infty,\quad \text{otherwise.}$$

Therefore

$$\text{if } -s'+\alpha p\leqslant 0,\quad \zeta(s',p)=-\infty;$$

$$\text{if } 0\leqslant -s'+\alpha p\leqslant\eta,\quad \zeta(s',p)=(-s'+\alpha p-\eta)/\eta;$$

$$\text{if } -s'+\alpha p\geqslant\eta,\quad \zeta(s',p)=-s'+\alpha p-\eta.$$

A standard computation shows that the Legendre transform (19) of this function is supported by $h=\alpha(\beta+1)$, and is equal to $\eta h/\alpha$ for $h\in[\alpha,\alpha/\eta]$ and to 1 if $h\geqslant\alpha/\eta$. Therefore, the multifractal formalism allows us to recover correctly the increasing part of the spectrum.

The advantage of using the multifractal formalism for chirps is patent on this example: first, it yields more complete information about the Hölder singularities of the function, but, furthermore, the standard multifractal formalism is completely wrong here: The reader can easily check that, for the function $\mathbf{F}_{\alpha,\eta}$, $\tau(p)=1+\alpha p-\eta$ which would yield a spectrum supported only on $h=\alpha$, for which $d(\alpha)=\eta$, whereas the right spectrum is supported by the interval $[\alpha,\alpha/\eta]$ on which $d(h)=h\eta/\alpha$.

ACKNOWLEDGMENTS

The author is very indepted to A. Kamont and B. Wolnik for allowing him to use their unpublished results of Ref. 3.

[1] J. Bergh and J. Peetre, ''On the spaces V_p,'' Boll. U.M.I.(4) **10**, 632–648 (1974).
[2] Y. Meyer, *Ondelettes et Opérateurs* (Hermann, Paris, 1990), p. 1.
[3] A. Kamont and B. Wolnik, ''Wavelet expansions and fractal dimension,'' to appear in Constr. Approsc.
[4] B. Dubuc, S. W. Zucker, C. Tricot, J. F. Quiniou, and D. Wehbi, ''Evaluating the fractal dimension of surfaces,'' Proc. R. Soc. London, Ser. A **425**, 113–127 (1989).
[5] C. Tricot, ''Function norms and fractal dimension,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **28**, 189–212 (1997).
[6] A. Deliu and B. Jawerth, ''Geometrical dimension versus smoothness,'' Constr. Approsc. **8**, 211–222 (1992).
[7] S. Jaffard, ''On lacunary wavelet series,'' preprint, 1997.
[8] S. Jaffard and Y. Meyer, ''Wavelet methods for pointwise regularity and local oscillations of functions,'' Mem. Am. Math. Soc. **123**, 587 (1996).
[9] Y. Meyer, ''Wavelets, Vibrations and scalings,'' CRM Ser. AMS, Presses de l'Université de Montréal, 1998.
[10] E. Chassande-Mottin and P. Flandrin, ''On the time–frequency detection of chirps,'' preprint, 1997.
[11] J-M. Innocent and B. Torrésani, ''Wavelets and binary coalescence detection,'' Appl. Comput. Harmon. Anal. **4**, 113–116 (1997).
[12] U. Frisch and G. Parisi, *Fully Developed Turbulence and Intermittency*, Proceedings of the Internation Summer School of Physics Enrico Fermi (North Holland, New York, 1985), pp. 84–88.
[13] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, J-F. Muzy, ''Ondelettes, multifractales et turbulence,'' Diderot Editeur, Arts et Sciences, Paris (1995).
[14] S. Jaffard, ''Multifractal formalism for functions Part I and Part II,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **28**, 944–998 (1997).
[15] A. Arneodo, E. Bacry, S. Jaffard, and J-F. Muzy, ''Oscillating singularities on Cantor sets: A grandcanonical multifractal formalism,'' J. Stat. Phys. **87**, 179–209 (1997).
[16] Ph. Tchamitchian and B. Torrésani, *Ridge and Skeleton Extraction from the Wavelet Transform, Wavelets and their Applications*, edited by Mary Beth Ruskai (Jones and Bartlett, Boston, 1992).
[17] F. Chamizo and A. Córdoba, ''The fractal dimension of a family of Riemann's graphs,'' C.R. Hebd. Seances Acad. Sci. **317**, 455–460 (1993).

Random cascades on wavelet dyadic trees

A. Arneodo
Centre de Recherche Paul Pascal, Avenue Schweitzer, 33600 Pessac, France

E. Bacry
Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau, France

J. F. Muzy
Centre de Recherches Mathématiques, Université de Montréal, C.P.6128 S. Centre Ville Montréal, Canada

(Received 15 January 1998; accepted for publication 31 March 1998)

We introduce a new class of random fractal functions using the orthogonal wavelet transform. These functions are built recursively in the space-scale half-plane of the orthogonal wavelet transform, ''cascading'' from an arbitrary given large scale towards small scales. To each random fractal function corresponds a random cascading process (referred to as a *$\mathscr{W}$-cascade*) on the dyadic tree of its orthogonal wavelet coefficients. We discuss the convergence of these cascades and the regularity of the so-obtained random functions by studying the support of their singularity spectra. Then, we show that very different statistical quantities such as correlation functions on the wavelet coefficients or the wavelet-based multifractal formalism partition functions can be used to characterize very precisely the underlying cascading process. We illustrate all our results on various numerical examples. © *1998 American Institute of Physics.* [S0022-2488(98)01008-1]

I. INTRODUCTION

Fractal and multifractal concepts[1–3] are now widely used to characterize multiscale phenomena that occur in various situations in physics, chemistry, geology and biology.[4–13] The multifractal formalism has been originally established to account for the statistical scaling properties of singular measures.[2,3,14–21] This formalism lies upon the determination of the so-called $f(\alpha)$ singularity spectrum[2] which quantifies the relative contribution of each singularity of the measure: Let S_α be the subset of points x where the measure of an ϵ-box $B_x(\epsilon)$, centered at x, scales like $\mu(B_x(\epsilon))\sim\epsilon^\alpha$ in the limit $\epsilon\to 0^+$, then by definition, $f(\alpha)=\dim_H(S_\alpha)$ is the Hausdorff-dimension of S_α. Actually, there exists a deep analogy that links the multifractal formalism with that of statistical thermodynamics.[22–24] This analogy provides a natural connection between the $f(\alpha)$ spectrum and a directly observable spectrum $\tau(q)$ defined from the power-law behavior, in the limit $\epsilon\to 0^+$, of the partition function[2,25] $Z_q(\epsilon)=\sum_i \mu(B_i(\epsilon))^q\sim\epsilon^{\tau(q)}$, where the sum is taken over a partition of the support of the singular measure into boxes of size ϵ. The variables q and $\tau(q)$ play the same role as the inverse of temperature and the free energy in thermodynamics, while the Legendre transform $f(\alpha)=\min_q(q\alpha-\tau(q))$ indicates that, instead of energy and entropy, we have α and $f(\alpha)$ as the thermodynamical variables conjugate to q and $\tau(q)$, respectively.[2,16–18,26] Let us recall that this thermodynamic multifractal formalism has been worked out in mathematics in the context of dynamical system theory.[22–24] However, rigorous proof of the above connection has been made only on some restricted classes of singular measures, e.g., invariant measures of some expanding Markov maps (''cookie-cutter'' Cantor sets) on an interval or a circle,[17,21] the invariant measure associated to the dynamical systems for period-doubling and for critical circle mappings with golden rotation number.[17] It has been developed into a powerful technique accessible also to experimentalists. Successful applications have been reported for multifractal measures which appear beyond the scope of dynamical systems.[6,10] Although valid for deterministic multifractals only, this description has been mainly applied for characterizing stochastic systems. But there is no reason, *a priori*, that all the realizations of the same stochastic multifractal measure correspond to a unique $f(\alpha)$-curve. Each realization has its own unique distribution of singularities and one crucial issue is to relate these distributions to

 © 1998 American Institute of Physics

some averaged versions computed experimentally. As emphasized in Ref. 27, one can take further advantage of the analogy with the thermodynamic formalism by using methods created specifically to study disorder in spin-glass theory.[28] When carrying out replica averages of the random partition function associated with a stochastic measure, one gets multifractal spectra $\tau(q,n)$ that generally depend on the number of members n in the replica average chosen (let us note that $n=0$ and $n=1$ correspond, respectively, to commonly used quenched and annealed averaging).[27] Then by Legendre transforming $\tau(q,n)$, some type of averaged $f(\alpha)$ spectra are being found.[27] Some care is thus required when interpreting these average spectra in order to avoid some misunderstanding of the underlying physics.

Multiplicative cascade models have enjoyed increasing interest in recent years as the paradigm of multifractal objects.[1–3,27,29] The notion of cascade actually refers to a self-similar process whose properties are defined multiplicatively from coarse to fine scales. In that respect, it occupies a central place in the statistical theory of turbulence.[13,29] Since Richardson's famous poem,[31] the turbulent cascade picture has been often invoked to account for the intermittency phenomenon observed in fully developed turbulent flows:[29,30] Energy is transferred from large eddies down to small scales (where it is dissipated) through a cascade process in which the transfer rate at a given scale is not spatially homogeneous, as supposed in the theory developed by Kolmogorov in 1941,[32] but undergoes local intermittent fluctuations.[13] Over the past 30 years, refined models including the log-normal model of Kolmogorov[33] and Obukhov,[34] multiplicative hierarchical cascade models like the random β-model, the α-model, the p-model (for a review see Ref. 29), the log-stable models[35–37] and more recently the log-infinitely divisible cascade models[38–41] with the rather popular log-Poisson model advocated by She and Leveque,[42] have grown in the literature as reasonable models to mimic the energy cascading process in turbulent flows. On a very general ground, a self-similar cascade is defined by the way the scales are refined and the statistics of the multiplicative factors at each step of the process.[27,29,37] One can thus distinguish discrete cascades that involve discrete scale ratios leading to log-periodic corrections to scaling (discrete scale invariance[43]), from continuous cascades without preferable scale factors (continuous scale invariance). As far as the fragmentation process is concerned, one can specify whether some conservation laws are operating or not;[27] in particular one can discriminate between conservative (the measure is conserved at each cascade step) and nonconservative (only some fraction of the measure is transferred at each step) cascades. More fundamentally, there are two main classes of self-similar cascade processes: deterministic cascades that generally correspond to solvable models and random cascades that are likely to provide more realistic models but for which some theoretical care is required as far as their multifractal limit and some basic multifractal properties (including multifractal phase transitions) are concerned.[27] As a notable member of the later class, the independent random cascades, introduced by Mandelbrot (commonly called $\mathscr{M}$-cascades)[30,44] as a general model of random curdling in fully developed turbulence, have a special status since they are the main cascade model for which deep mathematical results have been obtained.[45,46]

However, in physics as well as in other applied sciences, fractals appear not only as singular measures, but also as singular functions.[1,4–13] In order to stay in the context of fully developed turbulence, directly observable quantities are the velocity field or the temperature field rather than the dissipation field.[13,47] A classical way of analyzing the intermittent character of turbulent velocity signals consists in calculating the moments $S_p(l)=\langle \delta v_l^p \rangle \sim l^{\zeta_p}$ of the probability density function of longitudinal velocity increments $\delta v_l(x)=v(x+l)-v(x)$ over inertial separation l.[13,47,48] As originally prompted by Frisch and Parisi,[49] by Legendre transforming the scaling exponents ζ_p of the structure functions S_p, one expects to get the Hausdorff dimension $D(h)=\min_p(ph-\zeta_p+1)$ of the subset of $\mathbf{R}$ for which the velocity increments behave as $\delta v_l \sim l^h$. In a more general context, $D(h)$ will be defined as the spectrum of Hölder exponents of the signal under study and thus will have a similar status than the $f(\alpha)$-singularity spectrum for singular measures. Unfortunately, as pointed out by Muzy *et al.*,[50] there are some fundamental limitations to the structure function approach which intrinsically fails to fully characterize the $D(h)$ singularity spectrum. In previous work,[51–54] we have shown that there exists a natural way of performing a multifractal analysis of fractal functions which consists of using the continuous wavelet transform.[55–57] By using wavelets instead of boxes, like in the classical multifractal formalism,[2] one can take advantage of the freedom in the choice of these ''generalized oscillating boxes'' to get rid of possible smooth behavior that could mask singularities or perturb the estimation of their

strength h.[52,53] The other fundamental advantage of using wavelets is that the skeleton defined by the wavelet transform modulus maxima (WTMM) provides an adaptative space-scale partitioning from which one can extract the $D(h)$ singularity spectrum via the scaling exponents $\tau(q)$ of some partition functions defined on the skeleton. The so-called WTMM method[50–54] therefore gives access to the entire $D(h)$ spectrum via the usual Legendre transform $D(h)=\min_q(qh-\tau(q))$. We refer the reader to Refs. 52,58 for rigorous mathematical results. Let us mention that for the same reasons previously raised for stochastic multifractal measures, the theoretical treatment of random multifractal functions requires special attention. Let us also note that in a more recent work,[59–61] we have further generalized the WTMM multifractal formalism in order to incorporate in this statistical description (which applies for cusp-like singularities only) the possible existence of oscillating singularities. This new ''grand canonical'' description allows us to get the singularity spectrum $D(h,\beta)$ which accounts for the statistical contribution of singularities of Hölder exponent h and oscillation exponent β (where β characterizes the local power-law divergence of the instantaneous frequency).

Beyond the multifractal description, there is, however, the practical issue of defining in any concrete way how to build a multifractal function. Schertzer and Lovejoy[35] suggested a simple power-law filtering (fractional integration) of singular cascade measure as a mean to stochastically simulate fields reminiscent of passive scalars in turbulence. In the same spirit, the bounded cascade model of Marshak *et al.*[62] consists in acting on the multiplicative weights during the cascade in physical space. In Ref. 63, the midpoint displacement technique for building fractional Brownian motions was generalized to generate deterministic or random multiaffine functions. The same goal was achieved in Refs. 52,53 by combining fractional or ordinary integration with signed measures obtained by recursive cascade like procedures. Several other attempts to simulate ''synthetic turbulence'' that shares the intermittency properties of turbulent velocity data have partially succeeded.[64–66] More recently, the concept of self-similar cascades leading to multifractal measures has been generalized to the construction of scale-invariant signals using orthonormal wavelet basis.[67–70] Instead of redistributing the measure over sub-intervals with multiplicative weights, one allocates the wavelet coefficients in a multiplicative way on the dyadic grid. This method allows us to generate multifractal functions from a given deterministic or probabilistic multiplicative process. The main goal of this paper is to provide some mathematical framework to random $\mathscr{W}$-cascades on wavelet dyadic trees.[67–70]

The paper is organized as follows. In Secs. II and III, we explain how the $\mathscr{W}$-cascades are built using a wavelet orthogonal basis and we characterize the regularity properties of their corresponding random fractal functions by studying the support of their singularity spectrum. This support is linked to the statistical spectrum obtained with the wavelet based multifractal formalism.[50–54] The self-similarity kernel[41,68–72] which, from a statistical point of view, characterizes the self-similarity properties of a cascade process (in a different way from the multifractal formalisms) is introduced in Sec. IV. In Sec. V, we compute explicitly the correlation function of two wavelet coefficients of a $\mathscr{W}$-cascade.[73] It is proved to follow a power-law behavior when varying the spatial distance of the two coefficients. The statistical spectrum, the self-similarity kernel as well as the correlation function are shown to be numerically well estimated directly on the fractal function, using its wavelet decomposition (continuous, orthogonal or its associated extrema representation) with an arbitrary analyzing wavelet. All these results are illustrated on various computer generated numerical signals.

II. INTRODUCING WAVELET RANDOM CASCADES

A. The periodic wavelet orthogonal decomposition

As mentionned in the introduction, a $\mathscr{W}$-cascade[67–70] is built recursively on the dyadic grid of the orthogonal wavelet transform,[55–57] involving only scales that range between a given large scale L and the scale 0 (excluded). Thus the corresponding fractal function $f(x)$ does not involve scales greater than L. We can thus consider, for the sake of simplicity, that $f(x)$ is a periodic function of period L. In the following we will choose $L=1$. The $\mathscr{W}$-cascade will then be defined using a periodic orthonormal wavelet basis[74] of $L^2_{\mathrm{per}}([0,1])$, i.e., the space of 1-periodic functions with finite energy.

Such a basis can be constructed using two functions $\phi(x)$ and $\psi(x)$ of $L^2_{\text{per}}([0,1])$ (ψ is referred to as the *analyzing wavelet*) by means of translations and dilations of $\psi(x)$

$$\psi_{j,k}=2^{j/2}\psi(2^j x-k), \quad j\geqslant 0, \quad 0\leqslant k<2^{j-1}. \tag{1}$$

One can prove[55–57,74] that the so-obtained family of functions $\{\phi(x),\{\psi_{j,k}\}_{j,k}\}$ is an orthonormal basis of $L^2_{\text{per}}([0,1])$ if ϕ and ψ satisfy some conditions. Among these conditions, $\psi(x)$ should be localized around 0 and have N_ψ ($\geqslant 1$) vanishing moments

$$\int_0^1 x^n \psi(x)dx=0, \quad \text{for all } n<N_\psi. \tag{2}$$

The *wavelet coefficients* $\{c_\phi,\{c_{j,k}\}_{j,k}\}$ of a function $f(x)$ are then defined (modulo a normalization factor) as the coefficients of f in the orthonormal wavelet basis

$$\begin{cases} c_\phi=\langle f,\phi\rangle=\displaystyle\int_0^1 f(x)\phi(x)dx, \\ c_{j,k}=2^{j/2}\langle f,\psi_{j,k}\rangle=2^{j/2}\displaystyle\int_0^1 f(x)\psi_{j,k}(x)dx. \end{cases} \tag{3}$$

Remark: Let us note that the usual definition of the wavelet coefficients does not involve any normalization factor.[55–57] However, as we will see in Sec. III, the normalization factor $2^{j/2}$ has been introduced so that the Lipschitz exponent can be directly deduced from the power-law behavior of the coefficients $\{c_{j,k}\}_{j,k}$.

Since $\{\phi(x),\{\psi_{j,k}\}_{j,k}\}$ is an orthonormal basis, one gets the reconstruction formula

$$f(x)=c_\phi\phi(x)+\sum_{j\geqslant 0} 2^{-j/2}\sum_{0\leqslant k<2^j} c_{j,k}\psi_{j,k}(x). \tag{4}$$

On the one hand, let us note that, since all the $\psi_{j,k}$ have at least one vanishing moment, c_ϕ essentially ''captures'' the mean value of f. This explains why it is often referred to as the *approximation coefficient*.[55–57] On the other hand, assuming that the scale 1 ''corresponds'' to $\psi(x)$, one can easily prove that $\psi_{j,k}(x)$ is localized around $x=x_{j,k}$ and corresponds to the scale a_j with

$$x_{j,k}=2^{-j}k \quad \text{and} \quad a_j=2^{-j}. \tag{5}$$

Therefore, $c_{j,k}$ essentially captures the details of $f(x)$ around the point $x_{j,k}$ and at the scale a_j. They will be referred to as the *detail coefficients*.[55–57] As displayed in Fig. 1, these coefficients lie on a *dyadic grid* in the space-scale half-plane.

B. Building a $\mathscr{W}$-cascade

In this section, we build a random function $f(x)$ by specifying its wavelet coefficients $\{c_{j,k}\}_{j,k}$ and c_ϕ. The coefficient c_ϕ is chosen to be an arbitrary random variable and the $\{c_{j,k}\}_{j,k}$ are defined recursively in the following way:[68–70]

$$\begin{cases} c_{0,0}=1, \\ c_{j,2k}=W^{(l)}_{j-1,k}c_{j-1,k}, \\ c_{j,2k+1}=W^{(r)}_{j-1,k}c_{j-1,k}, \end{cases} \tag{6}$$

for all j ($j\geqslant 1$) and k ($0\leqslant k<2^{j-1}$) and where the $W^{(\epsilon)}_{j,k}$ ($\epsilon=l$ or r) are independent identically distributed (i.i.d.) real valued random variables.

Notation 1: Since all the random variables $W^{(\epsilon)}_{j,k}$ are i.i.d., we will often omit the indexes j,k and (ϵ) and we will use W as the generic name for these variables.

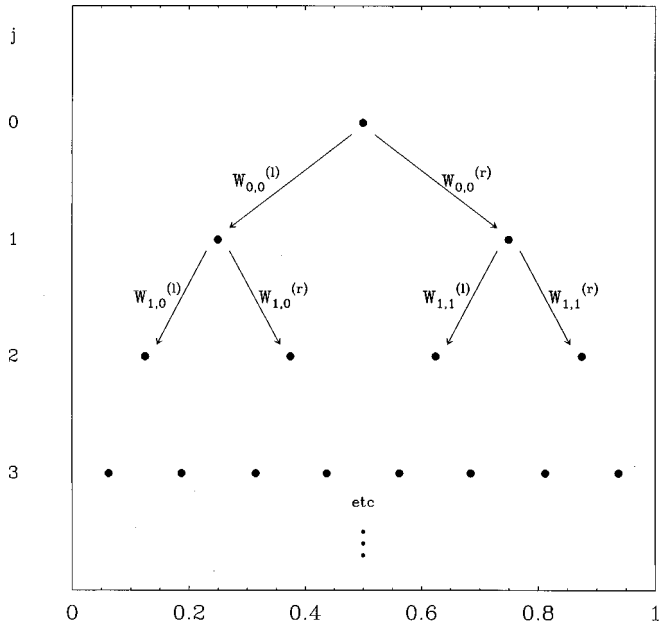

FIG. 1. Sketch of the construction rule of a $\mathscr{W}$-cascade. The wavelet coefficients $\{c_{j,k}\}_{j,k}$ lie on a dyadic grid. At each scale $a_j=2^{-j}$, the grid displays 2^j coefficients with abscissa $x_{j,k}=2^{-j}k$. The value of the wavelet coefficient $c_{j,2k}$ (resp. $c_{j,2k+1}$) is obtained from the value of the wavelet coefficient $c_{j-1,k}$ by multiplying it by $W_{j-1,k}^{(l)}$ (resp. $W_{j-1,k}^{(r)}$) as defined in Eq. (6).

As illustrated in Fig. 1, this recursive rule can be seen as a cascade process going from large scales (starting at scale 1) to smaller scales. It lies on a binary tree whose nodes are the wavelet coefficients and whose branches basically correspond (apart from the sign of the coefficients) to the same action of multiplying by W.

In the following, such a recursive rule will be referred to as a *$\mathscr{W}$-cascade* and $f(x)$ will be referred to as the function *corresponding* to the $\mathscr{W}$-cascade. Let us note that both a $\mathscr{W}$-cascade and its corresponding function are fully defined by the analyzing wavelet ψ, the laws of c_ϕ and W.

Let us note that the so-obtained function $f(x)$ [assuming that the infinite sum in Eq. (4) converges] is *self-similar* in the sense that the law of a wavelet coefficient $|c_{j_1,k}|$ at the scale 2^{-j_1} can be linked to the law of another wavelet coefficient $|c_{j_2,k'}|$ at scale $2^{-j_2}>2^{-j_1}$ using a multiplicative random variable depending only on the ratio of the two scales

$$|c_{j_1,k}| =_l |c_{j_2,k'}| X_{j_1-j_2},$$

where $=_l$ stands for the equality in law and where $X_n=|W_1 \ldots . W_n|$ (the W_i's are i.i.d real valued random variables with the same law as W). Thus, from a statistical point of view, the details of the function f at a scale a_1 are the same as the details at a scale a_2 up to a rescaling factor that depends only on a_1/a_2.

In this section, we just gave a ''theoretical'' description of a $\mathscr{W}$-cascade. Indeed, we have not proved that the sum in Eq. (4) does converge in some sense towards a random function $f(x)$. This will be the purpose of the next section. Actually, we will not only prove that, for almost all realizations of the $\mathscr{W}$-cascade, Eq. (4) does converge in $L^2_{\text{per}}([0,1])$, but we will also be able to characterize some regularity properties of the limit function.

Remark: A $\mathscr{W}$-cascade can be related to the $\mathscr{M}$-cascades previously introduced in Refs. 44–46. An $\mathscr{M}$-cascade is defined using the same recursive rule as a $\mathscr{W}$-cascade [Eq. (6)], but the $c_{j,k}$'s no longer correspond to wavelet detail coefficients: At step j of the recursion, a measure μ_j on $[0,1]$ is defined by $\mu_j[x_{j,k},x_{j,k}+a_j]=c_{j,k}$ $(\forall k, 0\leqslant k<2^{-j})$, where $x_{j,k}$ and a_j are defined as in Eq. (5). In Ref. 45, the authors proved that, under certain conditions (on W), μ_j converges towards a nondegenerated measure μ (when $j\to+\infty$). Thus the main difference between $\mathscr{M}$-cascades and $\mathscr{W}$-cascades is that $\mathscr{M}$-cascades are fractal measure models whereas $\mathscr{W}$-cascades are fractal function models. $\mathscr{M}$-cascades can be used, for instance, for modeling the energy dissipation in a

turbulent flow[29,30,33–42] whereas $\mathscr{W}$-cascades can be used directly for modeling the velocity signal of the same flow.[67–70] Moreover, as we will see in the next section, the underlying wavelet structure of a $\mathscr{W}$-cascade makes the proofs for convergence of the cascade and for the characterization of the corresponding fractal function much easier than for $\mathscr{M}$-cascades.

Remark: Let us note that Eq. (6) can be rewritten as

$$\begin{cases} \ln|c_{0,0}| = 0, \\ \ln|c_{j,2k}| = \ln|c_{j-1,k}| + \ln|W^{(l)}_{j-1,k}|, \\ \ln|c_{j,2k+1}| = \ln|c_{j-1,k}| + \ln|W^{(r)}_{j-1,k}|. \end{cases}$$

If $|W|$ is log-normal, these equations correspond to what one could call a *tree-autoregressive* process. This process is of order 1 in the sense that the regression involves only one term. Actually, we are currently working on higher order models. The notion of autoregressive models lying on a tree (including the orthonormal wavelet dyadic tree) has been introduced by Basseville and collaborators.[75] Let us emphasize that, besides the fact that the processes they study are autoregressive directly on the $c_{j,k}$ and not on their logarithm and that the processes we consider do not really correspond to autoregressive processes in the sense that they are not asymptotically stationnary (i.e., $\{\ln c_{j,k}\}_j$ for a fixed k is not stationary even when $j \to +\infty$), our approach is significantly different from theirs since we concentrate on the analysis of the fractal function $f(x)$ itself and not on the properties of the tree-process.

III. CONVERGENCE AND REGULARITY OF A $\mathscr{W}$-CASCADE

A. Convergence

In order to get the convergence of the sum in Eq. (4) for a given realization of the $\mathscr{W}$-cascade, we need to have upper bounds of the wavelet detail coefficients $\{c_{j,k}\}_{j,k}$. Actually, we are going to study the law of the maximum of the wavelet coefficients $\{c_{j,k}\}_{0 \leqslant k < 2^j}$ at a given scale 2^{-j}. This is the purpose of the following lemma which is proved in Appendix A.

Lemma 1: Let us consider the wavelet coefficients $\{c_\phi, \{c_{j,k}\}_{j,k}\}$ of a given $\mathscr{W}$-cascade associated to the random variable W [Eq. (6)]. Let $m_j = \max_k |c_{j,k}|$ ($m_0 = 1$) and Q_j^α the subset of the probability space Ω

$$Q_j^\alpha = \{\omega \in \Omega, m_j > 2^{-j\alpha}\}.$$

Let us set $q_j^\alpha = \mathrm{Prob}\{Q_j^\alpha\}$ and

$$p_j^\alpha = \mathrm{Prob}\{|W_1 \ldots W_j| > 2^{-j\alpha}\},$$

where the W_i are i.i.d random variables with the same law as W. Then

$$q_j^\alpha \leqslant 2^j p_j^\alpha.$$

Thus in order to get an upper bound to q_j^α, we just need an upper bound of p_j^α. Since p_j^α can be rewritten in the following way: $p_j^\alpha = \mathrm{Prob}\{\Sigma_{i=1}^{\ j} \log_2|W_i| \geqslant -j\alpha\}$, we can easily get an upper bound by using a large deviation property (see Appendix B for the proof of the following Lemma).[76]

Lemma 2: (Same notations as in Lemma 1.) If $\mathscr{E}(\log_2|W|) < +\infty$ then for all $\alpha < -\mathscr{E}(\log_2|W|)$, for all $\epsilon > 0$, there exists $J > 0$ such that for all $j > J$

$$p_j^\alpha < e^{j\epsilon} 2^{j(F(\alpha)-1)},$$

where $F(\alpha)$ is defined as the Legendre transform of the function $\tau(q)$

$$F(\alpha) = \inf_q (q\alpha - \tau(q)), \tag{7}$$

and where $\tau(q)$ is defined as

$$\tau(q) = -\log_2 \mathscr{E}(|W|^q) - 1, \quad \forall q \in \mathbf{R}. \tag{8}$$

Moreover, $F(\alpha)$ is a concave function, such that $F(-\mathcal{E}(\log_2|W|))=1$ and which is increasing on $]-\infty,-\mathcal{E}(\log_2|W|)]$ and decreasing on $[-\mathcal{E}(\log_2|W|),+\infty[$.

We are now ready to state the convergence proposition. For this purpose, we have to make two hypothesis on the law of W. Basically these hypothesis ensure that small values of W are much more probable than large values. This ''asymetry'' of the distribution function of W ensures that, for almost all realizations, the wavelet coefficients will ''fastly'' converge to 0 when the scale goes to 0 and thus that the sum in Eq. (4) will converge.

Proposition 1: (Convergence). *Let us consider a given $\mathcal{W}$-cascade associated with the random variable W [Eq. (6)]. If the law of W is such that*

(H1) $0<-\mathcal{E}(\log_2|W|)$ *and*
(H2) $\exists\,\eta>0,\ F|_{[0,\eta[}<0$ *[where F is defined in Eq. (7)],*
then for almost all realizations of the $\mathcal{W}$-cascade, the sum in Eq. (4) converges in $L^2_{\mathrm{per}}([0,1])$.

Proof: Let $\alpha=\eta/2$. Thus $0<\alpha<-\mathcal{E}(\log_2|W|))$ and $F(\alpha)<0$. By combining Lemma 1 and Lemma 2, one gets that (for ϵ arbitrarily small and j large enough)

$$q_j^\alpha \leqslant 2^j p_j^\alpha < e^{j\epsilon} 2^{jF(\alpha)}$$

and thus

$$\sum_j q_j^\alpha < \infty.$$

This last inequality can be rewritten as $\Sigma_j\ \mathrm{Prob}\{Q_j^\alpha\}<\infty$. By using the Borel-Cantelli lemma, one thus gets

$$\mathrm{Prob}\{Q_j^\alpha \ \text{infinitely often}\}=0,$$

which is equivalent to say that for almost all realizations of the $\mathcal{W}$-cascade, there exists J such that $m_j \leqslant 2^{-j\alpha}$ for $j \geqslant J$. This implies that

$$\sum_{j\geqslant J}\ \sum_{0\leqslant k<2^j} 2^{-j} c_{j,k}^2 \leqslant \sum_{j\geqslant J} m_j^2 \leqslant \sum_{j\geqslant J} 2^{-2\alpha j} < \infty.$$

Thus the sum in Eq. (4) converges in $L^2_{\mathrm{per}}([0,1])$. □

B. Regularity

The global regularity of a function is easily characterized by its orthogonal wavelet coefficients. Indeed, one can prove[77] that $f(x)$ is uniformly Lipschitz α (for $0<\alpha<N_\psi$) if and only if there exists a constant C such that $|c_{j,k}|<C2^{-j\alpha}$ for all j and k.

Let us recall that f is said to be uniformly Lipschitz α with $\alpha\in]0,1]$, if there exists a constant C such that $|f(x)-f(y)|<C|x-y|^\alpha$. Moreover, f is uniformly Lipschitz $\alpha\in]n,n+1]$, if $f^{(n)}$ is uniformly Lipschitz $\alpha-n$ (where $f^{(n)}$ is the nth derivative of f).

Remark: Let us note that the $2^{j/2}$ factor in Eq. (3) has been chosen so that the power-law behavior of $|c_{j,k}|$ when $j\to+\infty$ directly gives the Lipschitz regularity α (instead of $\alpha+1/2$ if there were no factor).

Thus, as for proving the convergence, in order to get the Lipschitz regularity of f, as long as N_ψ is large enough, one just needs to get upper bounds to the wavelet transform coefficients. All the work has already been done in the previous section. The following proposition is a direct application of Lemma 1 and Lemma 2.

Proposition 2: (Maximum global regularity). *Let us consider a given $\mathscr{W}$-cascade associated with the random variable W [Eq. (6)]. Let us suppose that the law of W is such that*

(H1) $0<-\mathscr{E}(\log_2|W|)$ *and*
(H2) $\exists\,\eta>0, F|_{[0,\eta[}<0$ *[where F is defined in Eq. (7)].*

Let us define

$$\alpha_{\min}=\sup\{h<-\mathscr{E}(\log_2|W|),F(h)<0\}. \tag{9}$$

If $\alpha<\alpha_{\min}$ $(0<\alpha<N_\psi)$ *then, for almost all realizations of the $\mathscr{W}$-cascade, $f(x)$ is uniformly Lipschitz α.*

Proof: The proof is very similar to the proof of the convergence proposition (Proposition 1). First, let us note that (according to that proposition) (H1) and (H2) ensure the convergence of the cascade. Let us choose $\alpha<\alpha_{\min}$ and $0<\alpha<N_\psi$. We thus have $0<\alpha<-\mathscr{E}(\log_2|W|)$ and $F(\alpha)<0$. Again, by combining Lemma 1 and Lemma 2 one gets that

$$\sum_j \text{Prob}\{Q_j^\alpha\}<\infty.$$

By using the Borel-Cantelli lemma, one deduces that $\text{Prob}\{Q_j^\alpha \text{ infinitely often}\}=0$. This means that for almost all realizations of the $\mathscr{W}$-cascade, there exists J such that $m_j\leqslant 2^{-j\alpha}$ for $j\geqslant J$, i.e., f is uniformly Lipschitz α. □

Remark: Let us note that, in most common cases, W is such that the left branch F_l of F (i.e., the branch corresponding to $F(\alpha)$ for $\alpha\in]-\infty,-\mathscr{E}(\log_2|W|)]$) is invertible. Thus $\alpha_{\min}$ just corresponds to the value

$$\alpha_{\min}=F_l^{-1}(0). \tag{10}$$

The local *Hölder exponent*[1,53] $h(x_0)$ measures the local regularity of a function f at a given point x_0. It is defined as the greatest exponent h such that there exists a constant C and a polynomial $P(x)$ such that

$$|f(x)-P(x)|<C|x-x_0|^h, \quad \text{for } x \text{ in a neighborhood of } x_0.$$

One can easily prove that $h(x_0)$ is greater than the ''maximum global regularity'' of f (i.e., the maximum α such that f is uniformly Lipschitz α). Thus, Proposition 2 can alternatively be seen as a ''minimum local regularity'' proposition:

Corollary 1: (Minimum local regularity). *Under the hypothesis (H1) and (H2) of Proposition 2, for almost all realizations of the $\mathscr{W}$-cascade, the local Hölder exponent of f at any point x is greater than or equal to $\alpha_{\min}$, i.e.,*

$$\forall x, \quad \alpha_{\min}\leqslant h(x).$$

All the arguments we used in the Lemmas 1 and 2 for deriving upper bounds to the absolute value of the wavelet coefficients can easily be inverted to get lower bounds. These new lemmas will lead to a ''maximum local regularity'' proposition. We are not going to give the full proof of this proposition since it is very close to the proof of the minimum local regularity proposition. We will just give the main steps of the proof.

Proposition 3: (Maximum local regularity). *Let us consider a given $\mathscr{W}$-cascade associated to the random variable W* [*Eq.* (6)]. *Let us suppose that the law of W is such that*

(H1) $0<-\mathscr{E}(\log_2|W|)$ *and*
(H2) $\exists\,\eta>0, F|_{[0,\eta[}<0$ *[where F is defined in Eq. (7)].*
Let us define

$$\alpha_{\max}=\inf\{h>-\mathscr{E}(\log_2|W|),F(h)<0\}. \tag{11}$$

Then, for almost all realizations of the $\mathscr{W}$-cascade, the local Hölder exponent $h(x)$ of $f(x)$ at any point x [as long as $h(x)<N_\psi$] is smaller than or equal to $\alpha_{\max}$, i.e.,

$$\forall x, \quad h(x)<N_\psi \Rightarrow h(x)\leqslant \alpha_{\max}.$$

Proof: We use the same notations as in Sec. III A except that we invert all the inequality signs. We define $\tilde{m}_j=\min_k|c_{j,k}|$, the subset $\tilde{Q}_j^\alpha$ of Ω

$$\tilde{Q}_j^\alpha=\{\omega\in\Omega,\tilde{m}_j<2^{-j\alpha}\},$$

and $\tilde{q}_j^\alpha=\mathrm{Prob}\{\tilde{Q}_j^\alpha\}$ its measure. We also set $\tilde{p}_j^\alpha=\mathrm{Prob}\{|W_1\ldots W_j|<2^{-j\alpha}\}$. Then, using exactly the same proof as for Lemma 1, one can show that

$$\tilde{q}_j^\alpha\leqslant 2^j\tilde{p}_j^\alpha.$$

If $\mathscr{E}(\log_2|W|)<+\infty$ and $\alpha>-\mathscr{E}(\log_2|W|)$ then a large deviation property[76] leads (as for Lemma 2) to

$$\forall\epsilon, \quad \exists J>0, \quad \forall j>J, \quad \tilde{p}_j^\alpha<e^{j\epsilon}2^{j(F(\alpha)-1)}.$$

Then we choose $\alpha>\alpha_{\max}$ with $0<\alpha<N_\psi$. We thus have $\alpha>-\mathscr{E}(\log_2|W|)$ and $F(\alpha)<0$. By using the Borel-Cantelli lemma in the same way as we did in Proposition 2, one gets that for almost all realizations of the $\mathscr{W}$-cascade, one has $\tilde{m}_j>2^{-j\alpha}$.

Moreover, one can prove[78–80] that if $h(x_0)$ $(<N_\psi)$ is the Hölder exponent of a function f at the point x_0, and if h is such that the wavelet coefficients of f verify

$$|c_{j,k}|\leqslant a_j^h+\frac{|x_{j,k}-x_0|^h}{|\ln|x_{j,k}-x_0||}, \tag{12}$$

where $a_j\to 0$ and $x_{j,k}\to x_0$ are defined in Eq. (5), then $h\geqslant h(x_0)$.

Since $c_{j,k}\geqslant\tilde{m}_j>2^{-j\alpha}$, one gets that $\alpha>h(x_0)$. Moreover, α can be chosen arbitrarily close to $\alpha_{\max}$, thus $\alpha_{\max}\geqslant h(x_0)$. □

Remark: In most common cases, W is such that the right branch F_r of F is invertible. In that case, in the same way we had Eq. (10), one gets

$$\alpha_{\max}=F_r^{-1}(0). \tag{13}$$

Generally, a good way to characterize the singular behavior of a function is to compute its *singularity spectrum* $D(h)$.[49–54] It is defined as

$$D(h)=\mathrm{Dim}_H(\{x\in\mathbf{R},h(x)=h\}), \tag{14}$$

where $h(x)$ is the Hölder exponent of the function at the point x and $\mathrm{Dim}_H(\mathscr{P})$ stands for the Hausdorff dimension of a set of points $\mathscr{P}$. Thus one can restate Corollary 1 and Proposition 3 in terms of the support of $D(h)$:

Corollary 2: (Support of the singularity spectrum). *Under the hypothesis (H1) and (H2) of Proposition 2, for almost all realizations of the $\mathscr{W}$-cascade, the support of the singularity spectrum $D(h)$ of f is included in the largest interval on which $F(\alpha)\geqslant 0$, i.e., the interval $[\alpha_{\min},\alpha_{\max}]$ [Eqs. (9) and (11)].*

Remark: Let us note that the Hölder exponents of a given function are fully characterized by its wavelet coefficients $\{c_{j,k}\}_{j,k}$ [through Eq. (12)] as long as these exponents are smaller than N_ψ. Thus the $D(h)$ singularity spectrum of the function $f(x)$ corresponding to a $\mathscr{W}$-cascade does not depend on the analyzing wavelet ψ that is chosen to build the cascade provided N_ψ is large enough. Particularly, as long as ψ satisfies $N_\psi>\alpha_{\max}$ [where $\alpha_{\max}$ defined in Eq. (11) depends only on the law of W], the cascade will correspond to the same singularity spectrum.

Remark: It was proved in Ref. 81 that for any function, the left branch $D_l(h)$ of the singularity spectrum of f is smaller than the left branch $F_l(\alpha)$ of the spectrum obtained with the multifractal formalism. Since the singularity spectrum obtained with the multifractal formalism leads, for almost all realizations of a given $\mathscr{W}$-cascade, to the function $F(\alpha)$ defined by Eq. (7), one can easily prove that for any $\mathscr{W}$-cascade we have

$$D_l(h=\alpha)\leqslant F_l(\alpha).$$

Definition 1: From now on, the function $F(\alpha)$ [Eq. (7)] will be referred to as the statistical spectrum *of the $\mathscr{W}$-cascade.*

Both spectra $D(h)$ and $F(\alpha)$ bring valuable information on the $\mathscr{W}$-cascade. The $D(h)$ spectrum has been initially introduced for characterizing the singular behavior of deterministic fractal signals. It was proved[52,58] that, for a large class of self-similar functions, the $D(h)$ spectrum can be obtained using the wavelet based multifractal formalism. In the case of random $\mathscr{W}$-cascades, we actually get two spectra: the spectrum $D(h)$ for each realization (which *a priori* depends on the realization) and the statistical spectrum $F(\alpha)$ that characterizes the probability that a given singular behavior appears in a realization of the cascade. Thus, for instance, the maximum value of $F(\alpha)$ corresponds to the most probable singular behavior in a realization of a $\mathscr{W}$-cascade. On the other hand, the negative values of $F(\alpha)$ correspond to ''rare'' events that one should not expect to observe in almost all realizations. In the next section, we will show that, in the case of $\mathscr{W}$-cascades, the wavelet based multifractal formalism[50–54] actually leads to a very reliable numerical estimation of the $F(\alpha)$ spectrum. Along with the *kernel* function[41,68–72] that we will introduce in Sec. IV and the correlation functions[73] in Sec. V, these statistical quantities allow a very accurate characterization of the random process.

Before moving on, let us first illustrate our purpose with some numerical simulations of $\mathscr{W}$-cascades corresponding to different laws for W.

C. Numerical simulations of $\mathscr{W}$-cascades

As stated in the Introduction, random cascade models have been introduced in the context of the phenomenological study of fully developed turbulence.[27,29,30,33–42] They were proposed to mimic, in some sense, the kinetic energy transfer from coarser scales to smaller ones. Log-normal statistics have been first guessed, 40 years ago, by Kolmogorov[33] and Obukhov[34] in order to account for the so-called ''intermittency phenomenon'' while the ''log-Poisson'' model[39] has been recently proposed by She and Lévèque[42] as a more accurate description of this phenomenon. Let us illustrate the results discussed above on these two models when extrapolated to multifractal functions.

1. Log-normal $\mathscr{W}$-cascades

Let us first start with W being a log-normal random variable. If μ and σ^2 are, respectively, the mean and the variance of $\ln|W|$ then a straightforward computation leads to

$$\tau(q)=-\frac{\sigma^2}{2\ln 2}q^2-\frac{\mu}{\ln 2}q-1,$$

and

$$F(\alpha)=-\frac{(\alpha+\mu/\ln 2)^2\ \ln 2}{2\sigma^2}+1.$$

Thus a log-normal $\mathscr{W}$-cascade is converging in $L^2_{\mathrm{per}}([0,1])$ if

$$\text{(H1) } \mu<0 \quad \text{and} \quad \text{(H2) } \frac{|\mu|}{\sigma}>\sqrt{2\ln 2}.$$

Moreover, by solving $F(0)=0$, one gets $\alpha_{\min}$ and $\alpha_{\max}$

$$\alpha_{\min}=-\frac{\sqrt{2}\sigma}{\sqrt{\ln 2}}-\frac{\mu}{\ln 2} \quad \text{and} \quad \alpha_{\max}=\frac{\sqrt{2}\sigma}{\sqrt{\ln 2}}-\frac{\mu}{\ln 2}. \tag{15}$$

In Fig. 2, we illustrate one realization of a ''very irregular'' ($\alpha_{\min}=0.13$) log-normal $\mathscr{W}$-cascade as well as one realization of a ''more regular'' one ($\alpha_{\min}=0.3$). The $\tau(q)$ and $F(\alpha)$ spectra corresponding to the irregular $\mathscr{W}$-cascade are displayed on the same figure.

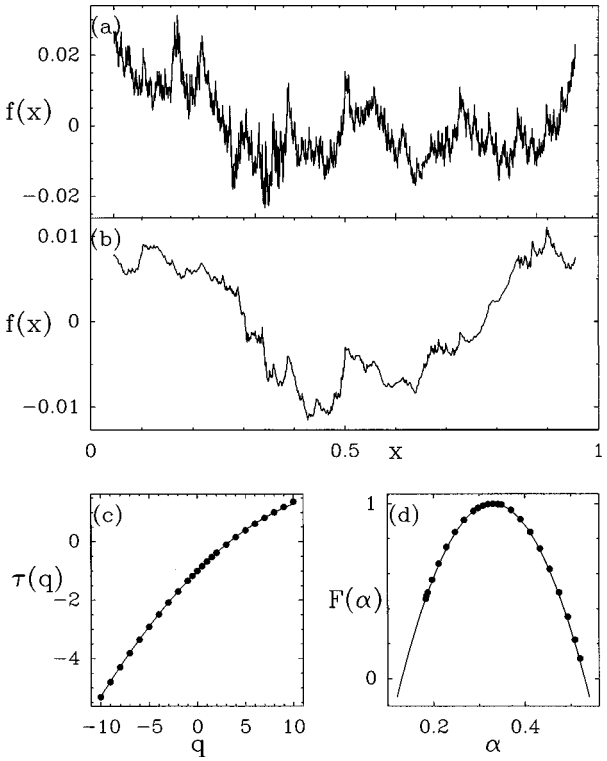

FIG. 2. Log-normal $\mathscr{W}$-cascades. (a) Realization of the random function corresponding to a log-normal $\mathscr{W}$-cascade using the ''Daubechies 5'' compactly supported orthonormal wavelet basis.[56] The law of $\ln|W|$ is Gaussian with mean $\mu=-0.33\ln 2$ and variance $\sigma^2=0.02\ln 2$. From Eq. (15) one gets that $\alpha_{\min}=0.13$ and $\alpha_{\max}=0.53$. (b) Realization of the random function corresponding to a log-normal $\mathscr{W}$-cascade using the ''Daubechies 5'' wavelet with the following parameter values: $\mu=-0.8\ln 2$ and $\sigma^2=0.125\ln 2$. From Eq. (15) one gets that $\alpha_{\min}=0.3$ and $\alpha_{\max}=1.3$. The fact that $\alpha_{\min}$ is greater for the cascade in (b) ($\alpha_{\min}=0.3$) than for the cascade in (a) ($\alpha_{\min}=0.13$) explains why the graph in (a) appears to be much more irregular than the graph in (b). (c) The $\tau(q)$ function [Eq. (8)] for the $\mathscr{W}$-cascade illustrated in (a). The symbols (●) correspond to the data computed using the WTMM method [Eq. (16)] with an order 2 spline wavelet on 1000 realizations of length 65536 of the $\mathscr{W}$-cascade. These numerical data are in remarkable agreement with the theoretical prediction (solid line); this illustrates the fact that the determination of the $\tau(q)$-spectrum can be performed using any analyzing wavelet (i.e., not necessarily the one that was used for building the cascade). (d) The $F(\alpha)$ statistical singularity spectrum [Eq. (7)] for the $\mathscr{W}$-cascade illustrated in (a). The numerical spectrum (●) was obtained by Legendre transforming the $\tau(q)$ data in (c). The theoretical spectrum (solid line) provides a remarkable fit of the data.

2. Log-Poisson $\mathscr{W}$-cascades

Let λ be the mean and the variance of the Poisson variable P. We consider that the law of $\ln|W|$ is the same as $P\ln\delta+\gamma$. A straightforward computation leads to

$$\tau(q)=\frac{\lambda(1-\delta^q)-\gamma q}{\ln 2}-1,$$

and

$$F(\alpha)=\left(\frac{\alpha}{\ln\delta}+\frac{\gamma}{\ln 2\ln\delta}\right)\left(\ln\left(\frac{\alpha\ln 2+\gamma}{-\lambda\ln\delta}\right)-1\right)+1-\frac{\lambda}{\ln 2}.$$

In Fig. 3, one realization of a log-Poisson $\mathscr{W}$-cascade is shown together with the corresponding $\tau(q)$ and $F(\alpha)$ spectra.

D. Computing the $F(\alpha)$ statistical spectrum using the multifractal formalism approach

Let us imagine that we have a large number of numerical signals which correspond to different realizations of the same random function $f(x)$. In order to characterize the self-similar behavior of the underlying cascade process, one could try to compute the $F(\alpha)$ statistical spectrum. In

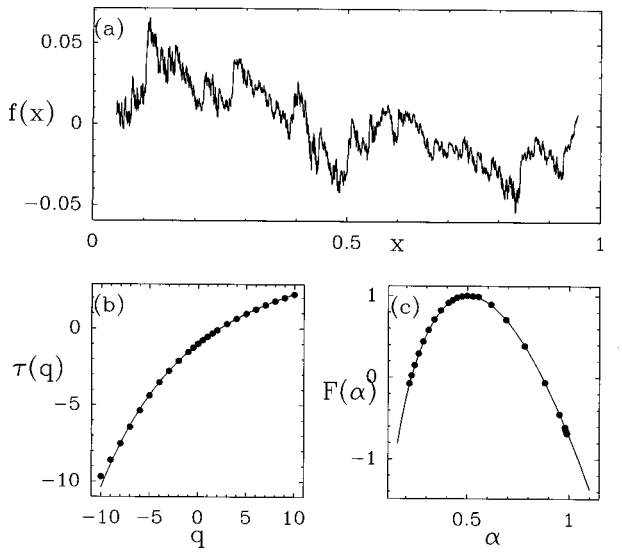

FIG. 3. Log-Poisson $\mathscr{W}$-cascades. (a) Realization of the random function corresponding to a log-Poisson $\mathscr{W}$-cascade using the ''Daubechies 5'' compactly supported orthonormal wavelet basis.[56] The mean of the Poisson variable P is $\lambda=2$ and the law of $\ln|W|$ is the same as $P\ln\delta+\gamma$ with $\delta=0.88$ and $\gamma=-0.11$. (b) The $\tau(q)$ spectrum [Eq. (8)] for the $\mathscr{W}$-cascade illustrated in (a). The data (●) computed using the WTMM method [Eq. (16)] with an order 2 spline analyzing wavelet on 1000 realizations of length 65535 of the $\mathscr{W}$-cascade, are in perfect agreement with the theoretical prediction (solid line). (c) The $F(\alpha)$ statistical singularity spectrum [Eq. (7)] for the $\mathscr{W}$-cascade illustrated in (a). The numerical spectrum (●) obtained by Legendre transforming the $\tau(q)$ data in (b) is compared to the theoretical spectrum (solid line).

order to do so, one can use the *wavelet based multifractal formalism* approach[50–54] which consists of computing a *partition function* $\hat{\mathscr{Z}}_j(q)$ corresponding, at each scale 2^{-j}, to the spatial average of the wavelet coefficients to the power q

$$\hat{\mathscr{Z}}_j(q)=\sum_{\text{realization }(r)}\sum_{0\leqslant k<2^j}|c_{j,k}^{(r)}|^q. \tag{16}$$

If the number of realizations is large enough, one can approximate realization averages by probability averages and get

$$\mathscr{Z}_j(q)=\sum_{0\leqslant k<2^j}\mathscr{E}(|c_{j,k}|^q)=2^j\mathscr{E}(|c_{j,k}|^q). \tag{17}$$

Since the law of $c_{j,k}$ is the same as the law of $W_1\ldots W_j$, one finally gets

$$\mathscr{Z}_j(q)=2^j\mathscr{E}(|W_1\ldots W_j|^q)=2^j\mathscr{E}(|W|^q)^j.$$

The $\tau(q)$ function [and consequently the $F(\alpha)$ spectrum] is obtained by analyzing the power-law scaling of $\mathscr{Z}_j(q)$ along the scales $a_j=2^{-j}$

$$\mathscr{Z}_j(q)=2^j\mathscr{E}(|W|^q)^j=a_j^{-\log_2\mathscr{E}(|W|^q)-1}\sim a_j^{\tau(q)}\quad a_j\to 0, \tag{18}$$

and

$$F(\alpha)=\inf_q(q\alpha-\tau(q)). \tag{19}$$

As long as the number of realizations is large, this approach leads to very precise estimations of $F(\alpha)$. However, from a practical point of view, we have made a major assumption: We have assumed that the realizations of the wavelet coefficients $\{c_{j,k}\}_{j,k}$ were known. This is clearly not the case since the only way to recover them from the realizations of f is to compute the scalar products of these realizations with the $\psi_{j,k}$'s; but we do not know what the analyzing wavelet ψ

is! What happens if we analyze the function f with a different analyzing wavelet ψ^1? Indeed, one can show[55–57] that the new wavelet coefficients $\{c^{(1)}_{j,k}\}$ can be expressed as a linear combination of the old ones

$$c^{(1)}_{j,k}=\sum_{k',j'} K_{\psi,\psi^1}(j-j',k-k')c_{j',k'}, \tag{20}$$

where K_{ψ,ψ^1} is a function that depends on ψ and ψ^1 only. This function basically corresponds to the scalar product of $\psi_{j,k}$ with $\psi^1_{j',k'}$. Let us note that the function K is localized in both variables and, from a numerical point of view, if ψ and ψ^1 are both well localized in Fourier and direct spaces, then the sum in Eq. (20) involves only a few terms. Before showing numerical applications, let us try to understand roughly how this affects the computation of $F(\alpha)$.

Let us note that the $F(\alpha)$ spectrum [and this is true also for any statistical quantity that is based on how the details of $f(x)$ change along scales] is not changed when changing the underlying $\mathscr{W}$-cascade independently on the scale. Indeed, for instance, if one multiplies each wavelet coefficient $c_{j,k}$ [defined recursively in Eq. (6)] by a random variable $X_{j,k}$ whose law does not depend on j or k, then the power-law behavior of $\mathscr{Z}_j(q)$ does not change along the scales

$$\mathscr{Z}_j(q)=2^j\mathscr{E}(|c_{j,k}|^q)=2^j\mathscr{E}(|W|^q)^j\mathscr{E}(|X|^q)\sim a_j^{\tau(q)}, \quad a_j\to 0.$$

Actually, this is exactly what happens if one replaces each coefficient $c_{j,k}$ by a linear combination of itself and its ''sons'' coefficients $c_{j+1,2k}$ and $c_{j+1,2k+1}$

$$c'_{j,k}=\lambda_1 c_{j,k}+\lambda^{(l)}c_{j+1,2k}+\lambda^{(r)}c_{j+1,2k+1}. \tag{21}$$

Indeed, it is easy to prove that the law of $c'_{j,k}$ is the same as the law of $W_1\ldots W_j(\lambda_1+\lambda^{(l)}W^{(l)}+\lambda^{(r)}W^{(r)})$ which can be rewritten as $c_{j,k}X_{j,k}$ with $X_{j,k}=\lambda_1+\lambda^{(l)}W^{(l)}+\lambda^{(r)}W^{(r)}$ (which does not depend on j).

It is somewhat more intricate when one performs a linear combination of the coefficients along the space axis

$$c'_{j,k}=\lambda_1 c_{j,k}+\lambda_2 c_{j,k+1}. \tag{22}$$

One can easily prove that, among the 2^j coefficients at the scale 2^{-j}, 2^l have, with their right neighbor, the first common ancestor at a scale 2^{-l}. Thus one can express the new partition function $\mathscr{Z}'_j(q)$ as

$$\mathscr{Z}'_j(q)=\sum_{l=0}^{j-1} 2^l\mathscr{E}(|W|^q)^l\mathscr{E}(|\lambda_1 W_1^{(1)}\ldots W_{j-l}^{(1)}+\lambda_2 W_1^{(2)}\ldots W_{j-l}^{(2)}|^q), \tag{23}$$

where the $W_i^{(1)}$ and the $W_i^{(2)}$ are i.i.d. random variables with the same law as W. Let T_l be the lth term in the latter sum, i.e.,

$$T_l=2^l\mathscr{E}(|W|^q)^l\mathscr{E}(|\lambda_1 W_1^{(1)}\ldots W_{j-l}^{(1)}+\lambda_2 W_1^{(2)}\ldots W_{j-l}^{(2)}|^q).$$

Let us note that the last term T_{j-1} behaves as $\mathscr{Z}_j(q)$:

$$2^{j-1}\mathscr{E}(|W^q|)^{j-1}\mathscr{E}(|\lambda_1 W_1^{(1)}+\lambda_2 W_1^{(2)}|^q)\sim C(q)2^j\mathscr{E}(|W^q|)^j=C(q)\mathscr{Z}_j(q), \tag{24}$$

where $C(q)$ does not depend on j. Thus one just needs to get an upperbound to all the other terms

$$\sum_{l=0}^{j-2} T_l\leqslant\sum_{l=0}^{j-2} 2^l\mathscr{E}(|W|^q)^l\mathscr{E}((|\lambda_1 W_1^{(1)}\ldots W_{j-l}^{(1)}|+|\lambda_2 W_1^{(2)}\ldots W_{j-l}^{(2)}|)^q),$$

$$\leqslant C_1(q)\sum_{l=0}^{j-2} 2^l\mathscr{E}(|W|^q)^l\mathscr{E}(|\lambda_1 W_1^{(1)}\ldots W_{j-l}^{(1)}|^q+|\lambda_2 W_1^{(2)}\ldots W_{j-l}^{(2)}|^q),$$

$$\leqslant C_2(q)\sum_{l=0}^{j-2} 2^l \mathscr{E}(|W|^q)^l \mathscr{E}(|W|^q)^{j-l},$$

$$\leqslant C_2(q)\sum_{l=0}^{j-2} 2^l \mathscr{E}(|W|^q)^j = C_3(q)2^j \mathscr{E}(|W|^q)^j = C_3(q)\mathscr{Z}_j(q),$$

where $C_1(q)$, $C_2(q)$ and $C_3(q)$ do not depend on j. From this last inequality and from Eq. (24), one deduces that the new partition function $\mathscr{Z}'_j(q)$ [Eq. (23)] behaves as expected

$$\mathscr{Z}'_j(q) \sim \mathscr{Z}_j(q), \quad j \to +\infty.$$

Thus as for Eq. (21), when one performs the linear combination on the wavelet coefficients corresponding to Eq. (22), the partition function displays exactly the same power-law behavior when the scale goes to 0. Actually, one can easily prove that this result still holds when we combine the two linear combinations (21) and (22).

Proposition 4: Let us consider a given $\mathscr{W}$-cascade associated to the random variable W [Eq. (6)]. Let $\mathscr{Z}_j(q)$ be the associated partition function defined by Eq. (17) *and let* λ_1, λ_2, $\lambda^{(l)}$ *and* $\lambda^{(r)}$ *in* **R**. *If we redefine the wavelet cascade coefficients* $c_{j,k}$ *in the following way:*

$$c'_{j,k} = \lambda_1 c_{j,k} + \lambda_2 c_{j,k+1} + \lambda^{(l)} c_{j+1,2k} + \lambda^{(r)} c_{j+1,2k+1},$$

then the newly obtained partition function $\mathscr{Z}'_j(q)$ behaves as the first one

$$\mathscr{Z}'_j(q) \sim \mathscr{Z}_j(q), \quad j \to +\infty.$$

Proof: The proof is straightforward and left to the reader. □

We thus expect the multifractal formalism to lead to a good determination of the $F(\alpha)$ statistical spectrum independently of the considered analyzing wavelet. Actually, it is likely to provide a good estimation of the left branch of $F(\alpha)$ $(q>0)$ but not of the right branch which corresponds to negative values of q. Indeed, the linear combination (20) might lead to null wavelet coefficients that would induce instabilities in the computation of $\hat{\mathscr{Z}}_j(q)$ [Eq. (16)] for $q<0$. In order to circumvent these instabilities, one should use the *Wavelet Transform Modulus Maxima* (WTMM) method introduced in Refs. 50–54. It basically consists in computing the partition function only on the *local modulus maxima* of the wavelet transform. Let us recall that the *modulus maxima* $\{x_i(a)\}_i$ of the continuous wavelet transform are defined at each scale a as the position of the local maxima of the absolute value of the wavelet transform.[80,82] These maxima lie on connected curves called *maxima lines*. The set of all the maxima lines existing at scale a will be denoted $\mathscr{L}(a)$. Then, the WTMM method consists in replacing the partition function $\hat{\mathscr{Z}}_j(q)$ by a new partition function which is stable for all q's in **R**

$$\hat{\mathscr{Z}}_j^{(\max)}(q) = \sum_{l_i \in \mathscr{L}(a)} \Big(\sup_{(x,a') \in l_i} |T_{\psi^1}(x,a')| \Big)^q, \quad \forall q \in \mathbf{R},$$

where $T_{\psi^1}(b,a)$ corresponds to the continuous wavelet transform of $f(x)$ at scale a and position b using the analyzing wavelet ψ^1.

As shown in Figs. 2 and 3, the WTMM method leads to a very good estimation of the $F(\alpha)$ statistical spectrum whatever the law of W and the analyzing wavelet ψ^1 are.

IV. THE SELF-SIMILARITY KERNEL

The function $f(x)$ associated to a $\mathscr{W}$-cascade is self-similar in the sense that the details of f at large scales are ''similar'' to its details at smaller scales up to a normalization factor. Let us look at how this property translates on the laws of the wavelet coefficients. Let P_j be the probability distribution function (p.d.f.) of the coefficients $|c_{j,k}|$ (P_j does not depend on k). Let $P_j^{(\log)}(x)$ be the p.d.f. of $\log|c_j|$

$$P_j^{(\log)}(x)=e^x P_j(e^x).$$

If $j_2>j_1$ then

$$\log|c_{j_2}|=\log|c_{j_1}|+\log|W_{j_1+1}|+\ \ldots\ +\log|W_{j_2}|. \tag{25}$$

This equation can be rewritten as

$$P_{j_2}^{(\log)}(x)=P_{j_1}^{(\log)}*G_{j_1,j_2}(x), \tag{26}$$

where $*$ denotes the convolution product and $G_{j_1,j_2}(x)=G\ *\ \ldots\ *\ G$, where $G(x)$ is the p.d.f. of $\log|W|$.[41,68–72] In the Fourier space, one gets

$$\hat{P}_{j_2}^{(\log)}(p)=\hat{P}_{j_1}^{(\log)}(p)\hat{G}^{s(j_1,j_2)}(p). \tag{27}$$

In the case of the $\mathscr{W}$-cascades, $s(j_1,j_2)=j_2-j_1$ represents the number of steps of the cascade from the scale 2^{-j_1} to the scale 2^{-j_2}. Of course, one cannot pick up any function s. It must satisfy the ''transitivity'' relation

$$s_{j_1,j_3}=s_{j_1,j_2}+s_{j_2,j_3}, \tag{28}$$

and the ''reflexivity'' relation

$$s_{j,j}=0. \tag{29}$$

Using any function s that satisfies both Eqs. (28) and (29), relation (27) can be seen as a *first order self-similarity property* that links the details at a scale 2^{-j_2} to the details at a larger scale 2^{-j_1}. The link is made through the *self-similarity kernel* $G_{j_1,j_2}(x)$ whose Fourier transform is of the form $\hat{G}^{s(j_1,j_2)}$. In the physical space, the kernel relation becomes[41,68–72,83–87]

$$P_{j_2}(e^x)=\int G_{j_1,j_2}(u)e^{-u}P_{j_1}(e^{x-u})du. \tag{30}$$

As we have seen, in the case of $\mathscr{W}$-cascades , the kernel function $G_{j_1,j_2}(x)$ depends only on j_2-j_1, i.e., only on the logarithm of the ratio of the two corresponding scales 2^{-j_1} and 2^{-j_2}. This can be seen as a ''scale-stationarity'' property of the self-similarity kernel.[38] In the following, a function that satisfies Eq. (27) with $s(j_1,j_2)=j_2-j_1$ will be referred to as a *scale-similar* function.

Let us note that, in the case of scale similar functions, the self-similarity kernel is directly linked to the statistical spectrum obtained by the multifractal formalism [Eq. (19)]. Indeed, the partition function [Eq. (17)] can be rewritten as

$$\mathscr{Z}_j(q)=2^j\mathscr{E}(|c_{j,k}|^q)=2^j\int e^{qu}P_j^{(\log)}(u)du=2^j\hat{P}_j^{(\log)}(iq).$$

Using relation (27) in the scale-similar case [i.e., $s(j_1,j_2)=j_2-j_1$] and relation (18), one gets

$$\mathscr{Z}_j(q)=2^j\hat{G}^{-j}(iq)\hat{P}_0^{(\log)}(iq)=2^j\hat{G}^{-j}(iq)\mathscr{Z}_0(q)\sim 2^{-j\tau(q)}.$$

The self-similarity kernel and the $\tau(q)$ spectrum are thus linked by the relation

$$\tau(q)=\log_2\ \hat{G}(iq)-1. \tag{31}$$

Remark: Let us note that if $f(x)$ corresponds to the Brownian motion, the orthogonal wavelet coefficients $\{c_{j,k}\}_{j,k}$ correspond to i.i.d white noises and thus the kernel $G_{j_1,j_2}(u)$ defined by Eq. (30) corresponds to the dirac distribution $\delta(u)$.[68,69]

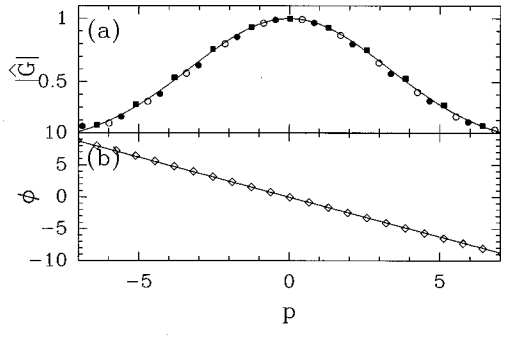

FIG. 4. Numerical computation of the self-similarity kernel of a log-normal $\mathscr{W}$-cascade with parameters $\mu=-0.37\ln 2$ and $\sigma^2=0.026\ln 2$. $\hat{G}_{a,a'}(p)=|\hat{G}_{a,a'}|e^{i\Phi_{a,a'}}$ as computed for $a/a'=5$, when using the Haar wavelet (●), an order 1 spline wavelet (○) and the complex Morlet wavelet (■).[56] (a) $|\hat{G}_{a,a'}|$ vs p; (b) $\Phi_{a,a'}$ vs p. The solid lines correspond to the theoretical predictions given by Eq. (32).

From a numerical point of view, one can compute the kernel of a $\mathscr{W}$-cascade by computing the wavelet coeffcients $\{c_{j,k}\}_{j,k}$ and the p.d.f. $P_j^{(\log)}$ at two different scales j_1 and j_2.[68,69] Then, from Eq. (27), $\hat{G}_{j_1,j_2}(p)=\hat{P}_{j_2}^{(\log)}(p)/\hat{P}_{j_1}^{(\log)}(p)$, i.e., the self-similarity kernel is obtained by performing the deconvolution of $P_{j_2}^{(\log)}$ by $P_{j_1}^{(\log)}$. As pointed out in the remark just below, this deconvolution requires some special care in order to avoid numerical instabilities.[68,69]

Remark: In order to compute $\hat{G}$, one has to perform a deconvolution. The deconvolution is performed in the Fourier space and thus consists basically in dividing $\hat{P}_{j_2}^{(\log)}$ by $\hat{P}_{j_1}^{(\log)}$. This division is unstable in the neighborhood of the (high) frequencies p for which $\hat{P}_{j_1}^{(\log)}(p)\simeq 0$. The smaller the support of $P_{j_1}^{(\log)}$, the slower the decay of $\hat{P}_{j_1}^{(log)}$ and thus the more stable the deconvolution. In order to decrease the support of $P_{j_1}^{(\log)}$, one could compute the p.d.f. $P_j^{(\log),(\max)}$ of the logarithm of the values of the modulus maxima of the continuous wavelet transform instead of computing the p.d.f. $P_j^{(\log)}$ of the logarithm of the orthogonal wavelet coefficients. Since, in the case of deterministic self-similar signals, the self-similarity properties are captured by the modulus maxima,[53,54,88] it is likely that, in the stochastic case, the self-similarity relation (26) still holds if $P_j^{(\log)}$ is replaced $P_j^{(\log),(\max)}$. Actually, from a numerical point of view, one can check[68–70] that it does hold with a very good precision for any scale a and not only for the dyadic scales $a_j=2^{-j}$. Since the support of $P_a^{(\log),(\max)}$ is much smaller than the one of $P_a^{(\log)}$, it gives a much more stable numerical method for performing the deconvolution and thus for computing the kernel $G_{a,a'}$.

Figures 4 and 5 report the results of the numerical computation of the self-similarity kernel of log-normal $\mathscr{W}$-cascades when using different analyzing wavelets ψ.[68–70] For the same reasons as

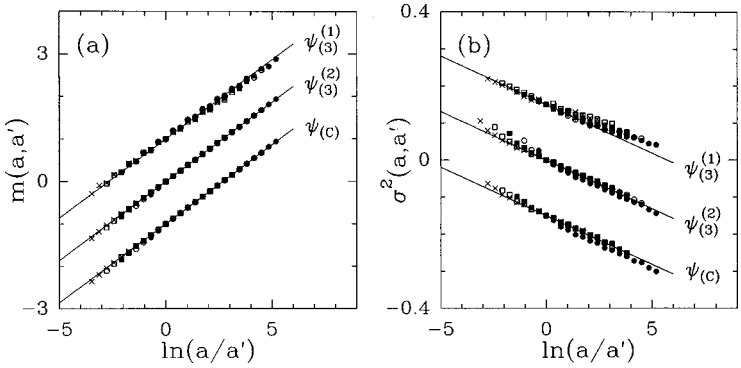

FIG. 5. Numerical computation of the scale dependence of the self-similarity kernel $\hat{G}_{a,a'}(p)$ of the log-normal $\mathscr{W}$-cascade studied in Fig. 4. (a) $m(a,a')=\partial\,\mathrm{Im}(\hat{G}_{a,a'})/\partial p|_{p=0}$ vs $\ln(a/a')$; (b) $\sigma^2(a,a')=-\partial^2(\ln|\hat{G}_{a,a'}|)/\partial p^2|_{p=0}$ vs $\ln(a/a')$. The symbols correspond to the following values of the reference scale $a'=2^5$ (●), 2^6 (○), 2^7 (■), 2^8 (□), 2^9 (×) and 2^{10} (△). The solid lines correspond to the theoretical predictions given by Eq. (33).

the ones we previously mentioned when estimating the $F(\alpha)$ statistical spectrum, since $\hat{G}$ depends only on $\tau(q)$ [Eq. (31)], its estimation should not depend upon the choice of ψ. This is clearly verified in the numerical simulations reported in Figs. 4 and 5. Moreover, the data shown in Fig. 4 are in remarkable agreement with the theoretical shape of a log-normal kernel

$$\hat{G}_{a,a'}(p)=e^{ipm(a,a')-p^2\sigma^2(a,a')/2}, \tag{32}$$

where

$$\begin{cases} m(a,a')=\mu\ \ln(a/a'), \\ \sigma^2(a,a')=-\sigma^2\ \ln(a/a'). \end{cases} \tag{33}$$

For values of $|p|\leqslant 7$, one does not see in the numerical data any significant departure from the Gaussian behavior of the kernel modulus $|G_{a,a'}|$, as well as from the linear behavior of its phase $\Phi_{a,a'}$.

As far as the scale dependence of the self-similarity kernel is concerned, we have plotted in Fig. 5, $m(a,a')=\partial\ \mathrm{Im}(\hat{G}_{a,a'})/\partial p|_{p=0}$ and $\sigma^2(a,a')=-\partial^2(\ln|\hat{G}_{a,a'}|)/\partial p^2|_{p=0}$ as functions of $\ln(a/a')$. One can see that, for different values of the reference scale a', all the points obtained when varying the scale a fall on a unique straight line which matches perfectly the theoretical predictions [Eq. (33)] and confirms the scale-similarity of the log-normal $\mathscr{W}$-cascade under study.

V. CORRELATION FUNCTIONS IN $\mathscr{W}$-CASCADES

The tree structure of a $\mathscr{W}$-cascade induces correlations between different details of the corresponding function $f(x)$.[73,88,89] These correlations can be characterized by computing the correlation between two wavelet coefficients at an arbitrary scale $a=2^{-j}$ and at a distance $\Delta x=2^{-j}\Delta k$. Since the wavelet coefficients $\{c_{j,k}\}_k$ at a given scale 2^{-j} are not stationary in k, we will compute an ''averaged version'' of the correlation function:[73]

Proposition 5: Let us consider a given $\mathscr{W}$-cascade associated with the random variable W [Eq. (6)*]. Let $a_j=2^{-j}$ and $\Delta x_{j,\Delta k}=2^{-j}\Delta k$. If we define $C(\Delta x_{j,\Delta k},a_j)$ as the correlation function*

$$C(\Delta x_{j,\Delta k},a_j)=\frac{1}{2^j}\sum_{k=0}^{k<2^j-\Delta k}\mathrm{Cov}(\log|c_{j,k}|,\log|c_{j,k+\Delta k}|), \tag{34}$$

where Cov *stands for the covariance. Then*

$$C(\Delta x_{j,2^p},a_j)=\sigma^2(j-p-2+2^{p-j+1}),\quad \forall p<j,$$

where σ^2 is the variance of $\log|W|$.

Proof: By definition one has

$$C(\Delta x_{j,2^p},a_j)=\frac{1}{2^j}\sum_{k=0}^{k<2^j-2^p}\mathrm{Cov}(\log|c_{j,k}|,\log|c_{j,k+2^p}|).$$

Let us fix k and set $k_1=k$, $k_2=k+\Delta k$. Let us suppose that the last common ancestor (on the binary tree of the $\mathscr{W}$-cascade) of c_{j,k_1} and c_{j,k_2} is at scale $2^{-d(j,k_1,k_2)}$ [in the following $d(j,k_1,k_2)$ will be referred to as the *$\mathscr{W}$-distance* between the two wavelet coefficients]. Then, one can write

$$|c_{j,k_1}|=|W_1|\ldots|W_{d(j,k_1,k_2)}||W^{(1)}_{d(j,k_1,k_2)+1}|\ldots|W_j|^{(1)},$$

and

$$|c_{j,k_2}|=|W_1|\ldots|W_{d(j,k_1,k_2)}||W^{(2)}_{d(j,k_1,k_2)+1}|\ldots|W_j|^{(2)},$$

where all the W_i, $W_i^{(1)}$ and $W_i^{(2)}$ are i.i.d. random variables with the same law as W. Then their covariance is

$$\mathrm{Cov}(c_{j,k_1},c_{j,k_2})=\sum_{i=1}^{i=d(j,k_1,k_2)}\mathrm{Cov}(\log|W_i|,\log|W_i|)=\sigma^2 d(j,k_1,k_2),$$

where σ^2 is the variance of $\log|W|$. Then

$$C(\Delta x_{j,2^p},a_j)=2^{-j}\sigma^2\sum_{d=0}^{d=j-1}N_{j,2^p}(d)d, \tag{35}$$

where $N_{j,2^p}(d)$ is the number of wavelet coefficients $c_{j,k}$ $(0\leqslant k<2^j-2^p)$ such that $c_{j,k}$ and $c_{j,k+2^p}$ are at a $\mathscr{W}$-distance d. It is clear that $N_{j,2^p}(d)=0$ for $d\geqslant j-p$. Moreover, one can easily show that

$$\forall d<j-p,\quad N_{j,2^p}(d)=2^pN_{j-p,1}(d).$$

Since $N_{j,1}(d)=2^d$, Eq. (35) becomes

$$\begin{aligned}C(\Delta x_{j,2^p},a_j)&=2^{-j}\sigma^2\sum_{d=0}^{d=j-p-1}2^pN_{j-2^p,1}(d)d,\\&=2^{p-j}\sigma^2\sum_{d=0}^{d=j-p-1}2^d d,\\&=2^{p-j}\sigma^2 2((j-p)2^{j-p-1}-2^{j-p}+1),\\&=\sigma^2(j-p-2+2^{p-j+1}).\end{aligned}$$

□

From this Proposition, one easily deduces the asymptotic behavior of the correlation function:[73]

Corollary 3: When Δx is small $(a<\Delta x\ll 1)$, the correlation function $C(\Delta x,a)$ [Eq. (34)] of a $\mathscr{W}$-cascade behaves as a logarithm function

$$C(\Delta x,a)=-\sigma^2\log_2(\Delta x)+o(\Delta x). \tag{36}$$

Thus, asymptotically, the correlation function does not depend on the scale a. From a numerical point of view, the cascade is constructed from the scale 1 $(j=0)$ down to a small scale (corresponding to the sampling rate of the numerical signal) 2^{-J}. If, on the contrary, we consider that the sampling rate is 1, then the signal has a total size $L=2^J$. Increasing J amounts in building a signal longer. The last corollary means that[73]

$$C(\Delta x,a)\sim\sigma^2\log_2(L/\Delta x), \tag{37}$$

when $a<\Delta x\ll L$.

Using the same kind of computations, one gets that the ''two-scale'' correlation function $C(\Delta x,a,a')$ between the coefficients at scale a and the coefficients at scale a' actually follows the same law as $C(\Delta x,a)$ as long as Δx is greater than the supremum of a and a'[73]

$$C(\Delta x,a,a')\sim\sigma^2\log_2(L/\Delta x), \tag{38}$$

when $\sup(a,a')<\Delta x\ll L$.

All these results are illustrated in Fig. 6 in the case of a log-normal $\mathscr{W}$-cascade. As seen in Figs. 6(a) and 6(b), the numerical computation of both the ''one-scale'' $C(\Delta x,a)$ and the ''two-scale'' $C(\Delta x,a,a')$ correlation functions are in very good agreement with the theoretical predictions given by Eqs. (37) and (38).

Remark: By the same kind of arguments as the ones used in Sec. III, one expects Eqs. (37) and (38) to hold even when computing the correlation functions using an analyzing wavelet ψ^1 which is different from the wavelet ψ used to build the $\mathscr{W}$-cascade.

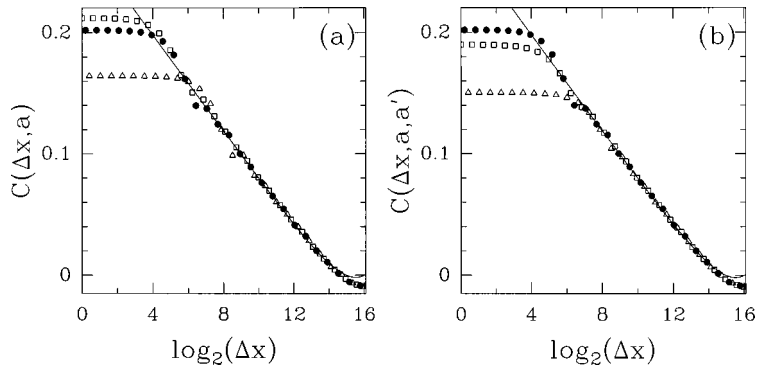

FIG. 6. Numerical computation of the wavelet coefficient correlation functions of a log-normal $\mathscr{W}$-cascade with variance $\sigma^2=0.03$. (a) ''One-scale'' correlation functions $C(\Delta x,a)$ for $a=4$ (□), 16 (●) and 64 (△). (b) ''Two-scale'' correlation functions $C(\Delta x,a,a')$ for $a=a'=16$ (●), $a=4$, $a'=16$ (□) and $a=16$, $a'=64$ (△). The data in (a) and (b) are in perfect agreement with the theoretical logarithm dependence (solid lines) given by Eqs. (37) and (38) ($L=2048$).

VI. CONCLUSION

To summarize, we have presented a first theoretical step towards a rigorous mathematical treatment of random cascading processes on the dyadic tree of their orthogonal wavelet coefficients. We have elaborated on the convergence of these $\mathscr{W}$-cascades and discussed the regularity of the limiting random functions by studying the support of their singularity spectra. We have shown mathematically and checked numerically on various computer synthetized signals, that very different statistical quantities such as the statistical spectrum, the self-similarity kernel and the correlation functions can be extracted directly from the fractal function using its wavelet decomposition (orthogonal, continuous or its associated modulus maxima) with an arbitrary analyzing wavelet. This mathematical study actually provides algorithms that are readily applicable to experimental situations. Recent applications of our methodology in the context of fully-developed turbulence[69,70,73] have revealed the existence of a (nonscale invariant) log-normal cascading process underlying the turbulent velocity fluctuations. More surprising are the results of a similar investigation of financial time series.[90] Underlying the fluctuations of the volatility (standard deviation) of the price variations, there exists a causal information cascade from large to small time scales that can be visualized with the wavelet representation. Let us emphasize that the fact that variations of prices over a one month scale influence in the future the daily price variations, is likely to be extraordinarily rich in consequences and this, not only for the fundamental understanding of the nature of financial markets, but also (and maybe more important) for practical applications. Indeed, the nature of the correlations across scales that are implied by this causal cascade has profound implications on the market risk, a problem of utmost concern for all financial institutions as well as individuals. These preliminary results are very promising as far as further experimental investigations of multiplicative cascade processes are concerned. There is no doubt in our minds that similar wavelet-based statistical analysis will lead to significant progress in fields other than hydrodynamic turbulence and finance.

ACKNOWLEDGMENTS

We are very grateful to S. Manneville and S. G. Roux for interesting discussions and technical assistance.

APPENDIX A: PROOF OF LEMMA 1

We want to prove the following lemma (Sec. III A)

Lemma 1: Let us consider the wavelet coefficients $\{c_\phi,\{c_{j,k}\}_{j,k}\}$ *of a given* $\mathscr{W}$*-cascade associated to the random variable W [Eq. (6)]. Let* $m_j=\max_k|c_{j,k}|$ $(m_0=1)$ *and* Q_j^α *the subset of the probability space* Ω

$$Q_j^\alpha=\{\omega\in\Omega, m_j>2^{-j\alpha}\}.$$

Let us set $q_j^\alpha = \mathrm{Prob}\{Q_j^\alpha\}$ *and*

$$p_j^\alpha = \mathrm{Prob}\{|W_1 \ldots W_j| > 2^{-j\alpha}\},$$

where the W_i *are i.i.d random variables with the same law as* W. *Then*

$$q_j^\alpha \leqslant 2^j p_j^\alpha .$$

Proof: By decomposing the binary tree into two binary subtrees, we obtain

$$\begin{aligned} q_j^\alpha &= \mathrm{Prob}\{m_j > 2^{-j\alpha}\}, \\ &= 1 - \mathrm{Prob}\{m_j \leqslant 2^{-j\alpha}\}, \\ &= 1 - \mathrm{Prob}\{m_{j-1}^{(l)}|W^{(l)}| \leqslant 2^{-j\alpha}, \text{ and } m_{j-1}^{(r)}|W^{(r)}| \leqslant 2^{-j\alpha}\}, \end{aligned}$$

where $W^{(l)}$ and $W^{(r)}$ are i.i.d. random variables with the same law as W and where $m_{j-1}^{(l)}$ and $m_{j-1}^{(r)}$ are i.i.d. random variables (independent from $W^{(l)}$ and $W^{(r)}$) with the same law as m_{j-1}. Since all the involved variables are independent, we get

$$q_j^\alpha = 1 - \mathrm{Prob}\{|W_1| m_{j-1} \leqslant 2^{-j\alpha}\}^2,$$

where W_1 is a random variable (independent from m_{j-1}) with the same law as W. By again decomposing each subtree into two subtrees, we get

$$q_j^\alpha = 1 - \mathrm{Prob}\{|W_1| m_{j-2}^{(l)}|W^{(l)}| \leqslant 2^{-j\alpha} \text{ and } |W_1| m_{j-2}^{(r)}|W^{(r)}| \leqslant 2^{-j\alpha}\}^2,$$

with the same notations as before. This time, since the same variable W_1 appears on both sides of the ''and,'' we cannot just split the two terms on each side of the ''and'' and keep the equality with q_j^α. Actually one can easily prove that for any independent random variables X_0, X_1, and Y, if X_0 and X_1 have the same law then

$$\mathrm{Prob}\{X_0 Y \leqslant a \text{ and } X_1 Y \leqslant a\} \geqslant \mathrm{Prob}\{XY \leqslant a\}^2.$$

By using this result, one easily gets from the last expression of q_j^α

$$q_j^\alpha \leqslant 1 - \mathrm{Prob}\{|W_1||W_2| m_{j-2} \leqslant 2^{-j\alpha}\}^4,$$

where W_2 is a random variable (independent from m_{j-2} and W_1) with the same law as W. By decomposing recursively each subtree into two subtrees and by using the fact that $m_0 = 1$, one finally obtains

$$q_j^\alpha \leqslant 1 - \mathrm{Prob}\{|W_1 W_2 \ldots W_j| \leqslant 2^{-j\alpha}\}^{2^j}.$$

Since $p_j^\alpha = 1 - \mathrm{Prob}\{|W_1 W_2 \ldots W_j| \leqslant 2^{-j\alpha}\}$, one gets

$$q_j^\alpha \leqslant 1 - (1 - p_j^\alpha)^{2^j}.$$

Moreover, since $p_j^\alpha \leqslant 1$ and $j \geqslant 0$, we have $(1 - p_j^\alpha)^{2^j} \geqslant 1 - 2^j p_j^\alpha$ and therefore

$$q_j^\alpha \leqslant 2^j p_j^\alpha . \qquad \square$$

APPENDIX B: PROOF OF LEMMA 2

We want to prove the following lemma (Sec. III A)

Lemma 2: (Same notations as in Lemma 1) If $\mathcal{E}(\log_2|W|) < +\infty$ *then for all* $\alpha < -\mathcal{E}(\log_2|W|)$, *for all* $\epsilon > 0$, *there exists* $J > 0$ *such that for all* $j > J$

$$p_j^\alpha < e^{j\epsilon} 2^{j(F(\alpha)-1)},$$

where $F(\alpha)$ *is defined as the Legendre transform of the function* $\tau(q)$

$$F(\alpha)=\inf_q(q\alpha-\tau(q)),$$

and where $\tau(q)$ *is defined as*

$$\tau(q)=-\log_2 \mathscr{E}(|W|^q)-1.$$

Moreover, $F(\alpha)$ *is a concave function, such that* $F(-\mathscr{E}(\log_2|W|))=1$ *and which is increasing on* $]-\infty,-\mathscr{E}(\log_2|W|)]$ *and decreasing on* $[-\mathscr{E}(\log_2|W|),+\infty[$.

Proof: Since $\alpha<-\mathscr{E}(\log_2|W|)\neq-\infty$, one has the following large deviation property:[76]

$$\limsup_{j\to\infty}\frac{1}{j}\ln p_j^\alpha\leqslant-\Phi(-\alpha), \tag{B1}$$

where Φ is the Cramér transform of $\log_2|W|$, i.e.,

$$\Phi(\alpha)=\sup_q(q\alpha-L(q)),$$

with

$$L(q)=\ln \mathscr{E}(e^{q\log_2 W})=\ln \mathscr{E}(|W|^{q/\ln 2}).$$

Equation (B1) is equivalent to

$$\forall\epsilon>0,\quad \exists J,\quad \forall j>J,\quad \frac{1}{j}\ln p_j^\alpha<-\Phi(-\alpha)+\epsilon,$$

which can be rewritten as

$$p_j^\alpha<e^{j\epsilon}e^{-j\Phi(-\alpha)}. \tag{B2}$$

Since $L(q)=-(\tau(q/\ln 2)+1)\ln 2$, we get

$$\begin{aligned}-\Phi(-\alpha)&=-\sup_q(-q\alpha-L(q)),\\&=-\sup_q(-q\alpha+(\tau(q/\ln 2)+1)\ln 2),\\&=(\inf_q(q\alpha-\tau(q))-1)\ln 2,\\&=(F(\alpha)-1)\ln 2.\end{aligned}$$

By replacing this expression in (B2), we get

$$p_j^\alpha<e^{j\epsilon}2^{j(F(\alpha)-1)}.$$

Moreover, the Cramér transform $\Phi(\alpha)$ is a convex function such that $\Phi(\mathscr{E}(\log_2|W|))=0$ and which is decreasing on $]-\infty,\mathscr{E}(\log_2|W|)]$ and increasing on $[\mathscr{E}(\log_2|W|),+\infty[$. Thus, since $F(\alpha)=1-\Phi(-\alpha)/\ln 2$, $F(\alpha)$ is a concave function such that $F(-\mathscr{E}(\log_2|W|))=1$ and which is increasing on $]-\infty,-\mathscr{E}(\log_2|W|)]$ and decreasing on $[-\mathscr{E}(\log_2|W|),+\infty[$. □

[1] B. B. Mandelbrot, *Fractals: Form, Chance and Dimension* (Freeman, San Francisco, 1977); *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
[2] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).
[3] G. Paladin and A. Vulpiani, Phys. Rep. **156**, 148 (1987).
[4] J. Feder, *Fractals* (Pergamon, New York, 1988).
[5] T. Vicsek, *Fractal Growth Phenomena* (World Scientific, Singapore, 1989).
[6] *Fractals in Physics, Essays in honor of B. B. Mandelbrot*, edited by A. Aharony and J. Feder, Physica D **38** (North-Holland, Amsterdam, 1989).
[7] B. J. West, *Fractal Physiology and Chaos in Medicine* (World Scientific, Singapore, 1990).

[8] *Fractals and Disordered Systems*, edited by A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1991).
[9] H. E. Stanley, S. V. Buldyrev, A. Goldberger, S. Havlin, S. M. Ossadnik, C. K. Peng, and M. Simmons, Fractals **1**, 283 (1993).
[10] *Fractals in Natural Sciences*, edited by T. Vicsek, M. Shlesinger, and M. Matsushita (World Scientific, Singapore, 1994).
[11] B. J. West and W. Deering, Phys. Rep. **254**, 1 (1994).
[12] *Fractals in Geoscience and Remote Sensing*, edited by G. G. Wilkinson, I. Kanellopoulos, and J. Mégier, Images Understanding Research Series, Vol. 1 (ECSC-EC-EAEC, Brussels, Luxemburg, 1995).
[13] U. Frisch, *Turbulence* (Cambridge University Press, Cambridge, 1995).
[14] R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, J. Phys. A **17**, 3521 (1984).
[15] E. B. Vul, Ya. G. Sinai, and K. M. Khanin, Usp. Mat. Nauk. **39**, 3 (1984) [J. Russ. Math. Surv. **39**, 1 (1984)].
[16] R. Badii, Thesis, University of Zurich, 1987.
[17] P. Collet, J. Lebowitz, and A. Porzio, J. Stat. Phys. **47**, 609 (1987).
[18] M. J. Feigenbaum, J. Stat. Phys. **46**, 919,925 (1987).
[19] M. H. Jensen, L. P. Kadanoff, and I. Procaccia, Phys. Rev. A **36**, 1409 (1987).
[20] B. B. Mandelbrot, *Fractals and Multifractals: Noise, Turbulence and Galaxies*, Selecta, Vol. 1 (Springer, New York, 1989).
[21] D. Rand, Ergod. Th. Dyn. Sys. **9**, 527 (1989).
[22] Ya. G. Sinai, J. Russ. Math. Surveys **27**, 21 (1972).
[23] R. Bowen, Lect. Notes Math. **470**, 1 (1975).
[24] D. Ruelle, *Thermodynamic Formalism* (Addison-Wesley, Reading, 1978).
[25] P. Grassberger, R. Badii, and A. Politi, J. Stat. Phys. **51**, 135 (1988).
[26] T. Bohr and T. Tèl, in *Direction in Chaos*, Vol. 2, edited by B. L. Hao (World Scientific, Singapore, 1988), p. 194.
[27] H. G. E. Hentschel, Phys. Rev. E **50**, 243 (1994).
[28] S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).
[29] C. Meneveau and K. R. Sreenivasan, J. Fluid Mech. **224**, 429 (1991).
[30] B. B. Mandelbrot, J. Fluid Mech. **62**, 331 (1974).
[31] L. Richardson, Proc. R. Soc. London, Ser. A **110**, 709 (1926).
[32] A. N. Kolmogrov, C. R. Acad. Sci. USSR **30**, 301 (1941).
[33] A. N. Kolmogorov, J. Fluid Mech. **13**, 82 (1962).
[34] A. M. Obukhov, J. Fluid Mech. **13**, 77 (1962).
[35] D. Schertzer and S. Lovejoy, J. Geophys. Res. **92**, 9693 (1987).
[36] S. Kida, J. Phys. Soc. Jpn. **60**, 5 (1990).
[37] D. Schertzer, S. Lovejoy, F. Schmitt, Y. Chigirinskaya, and D. Marsan, Fractals **5**, 427 (1997).
[38] E. A. Novikov, Physica A **2**, 814 (1990); Phys. Rev. E **50**, 3303 (1995).
[39] Z. S. She and E. C. Waymire, Phys. Rev. Lett. **74**, 262 (1995).
[40] B. Dubrulle, Phys. Rev. Lett. **73**, 959 (1994); J. Phys. II **6**, 1825 (1996).
[41] B. Castaing and B. Dubrulle, J. Phys. II **5**, 895 (1995).
[42] Z. S. She and E. Leveque, Phys. Rev. Lett. **72**, 336 (1994).
[43] D. Sornette, in *Scale Invariance and Beyond*, edited by B. Dubrulle, F. Graner, and D. Sornette (EDP Sciences, Les Ulis, 1997), pp. 235.
[44] B. B. Mandelbrot, C. R. Acad. Sci. Paris Ser. A **278**, 289,355 (1974).
[45] J. P. Kahane and J. Peyrière, Adv. Math. **22**, 131 (1976).
[46] G. M. Molchan, Commun. Math. Phys. **179**, 681 (1996).
[47] *Turbulence: A Tentative Dictionary*, edited by P. Tabeling and O. Cardoso (Plenum, New York, 1995).
[48] A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics*, Vol. II (MIT Press, Cambridge, MA, 1975).
[49] U. Frisch and G. Parisi, in Proc. of Int. School on *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, edited by M. Ghil, R. Benzi, and G. Parisi (North-Holland, Amsterdam, 1985), p. 84.
[50] J. F. Muzy, E. Bacry, and A. Arneodo, Phys. Rev. E **47**, 875 (1993).
[51] J. F. Muzy, E. Bacry, and A. Arneodo, Phys. Rev. Lett. **67**, 3515 (1991).
[52] E. Bacry, J. F. Muzy, and A. Arneodo, J. Stat. Phys. **70**, 635 (1993).
[53] J. F. Muzy, E. Bacry, and A. Arneodo, Int. J. Bifurcation Chaos **4**, 245 (1994).
[54] A. Arneodo, E. Bacry, and J. F. Muzy, Physica A **213**, 232 (1994).
[55] Y. Meyer, *Ondelettes* (Herman, Paris, 1990).
[56] I. Daubechies, *Ten Lectures on Wavelets* (S.I.A.M., Philadelphia, 1992).
[57] S. Mallat, *A Wavelet Tour in Signal Processing* (Academic Press, New York, 1998).
[58] S. Jaffard, ''Multifractal formalism for functions, Parts I and II,'' SIAM J. Math. Analysis (1997), to appear.
[59] A. Arneodo, E. Bacry, and J. F. Muzy, Phys. Rev. Lett. **74**, 4823 (1995).
[60] A. Arneodo, E. Bacry, S. Jaffard, and J. F. Muzy, J. Stat. Phys. **87**, 179 (1997).
[61] A. Arneodo, E. Bacry, S. Jaffard, and J. F. Muzy, ''Singularity spectrum of multifractal functions involving oscillating singularities,'' J. of Fourier Analysis and Applications (1997), to appear.
[62] A. Marshak, A. Davis, R. Cahalan, and W. Wiscombe, Phys. Rev. E **49**, 55 (1994).
[63] A. L. Barabási and T. Vicsek, Phys. Rev. A **44**, 2730 (1991).
[64] J. Eggers and S. Grossmann, Phys. Rev. A **45**, 2360 (1992).
[65] J. A. C. Humphrey, C. A. Schuler, and B. Rubinski, Fluid Dyn. Res. **9**, 81 (1992).
[66] A. Juneja, D. P. Lathrop, K. R. Sreenivasan, and G. Stolovitzky, Phys. Rev. E **49**, 5179 (1994).
[67] R. Benzi, L. Biferale, A. Crisanti, G. Paladin, M. Vergassola, and A. Vulpiani, Physica D **65**, 352 (1993).
[68] S. G. Roux, Thesis, University of Aix-Marseille II, 1996.
[69] A. Arneodo, J. F. Muzy, and S. G. Roux, J. Phys. II **7**, 363 (1997).
[70] A. Arneodo, S. Manneville, and J. F. Muzy, Eur. Phys. J. B **1**, 129 (1998).
[71] B. Castaing, Y. Gagne, and E. J. Hopfinger, Physica D **46**, 177 (1990).

[72] B. Castaing, in *Scale Invariance and Beyond*, edited by B. Dubrulle, F. Graner, and D. Sornette (EDP Sciences, Les Ulis, 1997), p. 225.
[73] A. Arneodo, E. Bacry, S. Manneville, and J. F. Muzy, Phys. Rev. Lett. **80**, 708 (1998).
[74] V. Perrier, Ph.D. thesis, University of Paris VI, 1991.
[75] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, IEEE Trans. Inf. Theory **38**, 766 (1992).
[76] D. W. Strook, *An Introduction to the Theory of Large Deviations* (Springer-Verlag, Berlin, 1984).
[77] Y. Meyer, *Wavelet and Operators*, Advanced Mathematics (Cambridge University Press, Cambridge, 1992).
[78] S. Jaffard, C.R. Acad. Sci. Paris **308**, serie I, 79 (1989); serie I, 19 (1992).
[79] M. Holschneider and P. Tchamitchian, in *Les Ondelettes*, edited by P. G. Lemarié (Springer-Verlag, Berlin, 1990), p. 102.
[80] S. Mallat and W. L. Hwang, IEEE Trans. Inf. Theory **38**, 617 (1992).
[81] S. Jaffard and Y. Meyer, Memoirs of the American Mathematical Society, no. 587, Vol. 123 (Sept. 1996).
[82] S. Mallat and S. Zhong, IEEE Trans. Pattern. Anal. Mach. Intell. **14**, 710 (1992).
[83] B. Castaing, Y. Gagne, and M. Marchand, Physica D **68**, 387 (1993).
[84] A. Naert, L. Puech, B. Chabaud, J. Peinke, B. Castaing, and B. Hebral, J. Phys. II **4**, 215 (1994).
[85] B. Chabaud, A. Naert, J. Peinke, F. Chillà, B. Castaing, and B. Hebral, Phys. Rev. Lett. **73**, 3227 (1994).
[86] B. Castaing, J. Phys. II **6**, 105 (1996).
[87] F. Chillà, J. Peinke, and B. Castaing, J. Phys. II **6**, 455 (1996).
[88] M. Greiner, J. Giesemann, P. Lipa, and P. Carruthers, Z. Phys. C **69**, 305 (1996).
[89] M. Greiner, J. Giesemann, and P. Lipa, Phys. Rev. E **56**, 4263 (1997).
[90] A. Arneodo, J.F. Muzy, and D. Sornette, Eur. Phys. J. B **2**, 277 (1998).

Time-dependent scattering on fractal measures

Charles-Antoine Guerin[a] and Matthias Holschneider[b]
Centre de Physique Théorique,[c] CNRS, Luminy, Case 907, F-13288 Marseille Cedex 9, France

(Received 1 January 1998; accepted for publication 3 April 1998)

In this paper we study the time evolution for the Schrödinger equation and the wave equation on the line when the interaction term is a fractal measure. First, we extend the usual one-dimensional potential scattering formalism to interactions defined as measures. Then we show how to retrieve information on the fractality of the interaction term from time-dependent scattering data. In the case of the Schrödinger equation we shall obtain the wavelet correlation dimension of the scatterer. For the wave equation the whole set of generalized multifractal dimensions can be recovered, provided the scatterer actually is fractal (nonsmooth). In this latter case, we also show how the reflected wave packets can be interpreted in terms of wavelet transform of the interaction. © *1998 American Institute of Physics.* [S0022-2488(98)01108-6]

I. INTRODUCTION AND MAIN RESULTS

In the last two decades, there has been a growing interest in the scattering of waves by fractal objects because of the many important applications in physics and engineering and also because of the mathematical challenge. By now, a huge amount of literature has been devoted to the subject; we refer to, e.g., Refs. 1–8 for some survey articles. The typical result is that the scattered intensity in the stationary regime can be simply related to some fractal dimension of the obstacle. However, the temporal aspect seems to have been neglected in the literature. Other than the paper by Berry and Blackwell,[9] who studied the time-delay of waves encountering multiscales surfaces, we do not know of other theoretical results relating time-dependent data to fractal properties of the scatterer.

The aim of this paper is to study the time evolution generated by Schrödinger operators on the line when the potential term is a fractal measure. We wish to show how the time-space behavior of the wave packets mirrors the ''fractality'' of the scatterer. The problem we want to deal with can be put in the following abstract form. Let $\mathscr{H}$ be some Hilbert space and H a self-adjoint operator on $\mathscr{H}$, with domain $D(H)$. Consider the Cauchy problem of first and second order in time:

$$i\partial_t \Psi_t - H\Psi_t = 0, \quad \Psi_{t=0} = f \in D(H), \tag{1.1}$$

$$\partial_{tt} \Phi_t + H\Phi_t = 0, \quad \Phi_{t=0} = f \in D(H), \quad \partial_t \Phi_{t=0} = g \in D(H). \tag{1.2}$$

The first equation is the Schrödinger equation, which describes the evolution of quantum wave packets under the action of the Hamiltonian H. Its solution is uniquely determined by the initial state:

$$\Psi_t = e^{-itH} f, \quad t \geqslant 0. \tag{1.3}$$

We will refer to the second equation as ''the wave equation.'' It can be solved by reduction to a first-order matricial equation. If H is non-negative, there exists a unique non-negative self-adjoint operator $H^{1/2}$ such that $(H^{1/2})^2 = H$ and $D(H)$ is dense in $D(H^{1/2})$. The unique solution of (1.2) is then given by (see, e.g., Ref. 10)

[a]Electronic mail: guerin@cpt.univ-mrs.fr
[b]Electronic mail: hols@cpt.univ-mrs.fr
[c]Unité Propre de Recherche 7061.

 © 1998 American Institute of Physics

$$\Phi_t = \cos(H^{1/2}t) f + H^{-1/2} \sin(H^{1/2}t) g, \quad t \geqslant 0, \tag{1.4}$$

where the functions of H are defined by the usual functional calculus. In this paper, we will take $\mathscr{H} = L^2(\mathbb{R})$, the Hilbert space of square integrable functions on $\mathbb{R}$; H will be a generalized Schrödinger operator formally given by $H = -\Delta + V$, where V is a fractal measure on $\mathbb{R}$. We want to analyze the scattering potential V using data that are available away from the scatterer. Instead of reconstructing V completely we are interested only in its fractal features and how they are mirrored in the scattering data. This type of analysis has been proposed in the stationary scattering approach.[11,12] It will turn out that fractal properties described through small scale properties of wavelet transforms play an important role.

A. The results

Let us show briefly the main results, which go in two directions. First, we propose an original approach to study singular Schrödinger operators on the line, by generalizing the tools that are used for smooth interactions (Jost functions, Lippmann–Schwinger solutions, etc.). In particular, we retrieve and strengthen some known results concerning the spectrum of the Hamiltonian and the wave operators and derive an explicit formula for the spectral measure. This makes it possible to study arbitrary evolution operators, in particular those given by (1.3) and (1.4).

The second aspect of this paper concerns the wavelet analysis in connection to the scattering problem. We shall show how to perform ''experimental'' wavelet transforms of the potential by means of time-dependent scattering data. We assume for the moment some familiarity with the usual concepts of wavelet analysis. The complete definitions can be found in Sec. II. A natural quantity to look at is the amount of energy scattered back from the scatterer. For the sake of simplicity suppose that V is compactly supported. We call Ω^- and Ω^+ the regions to the left and to the right of the support of V and $\mathbb{I}_{\Omega^\pm}$ the corresponding characteristic functions, $\mathbb{I}_{\Omega^\pm}(x) = 1$ if $x \in \Omega^\pm$, $\mathbb{I}_{\Omega^\pm}(x) = 0$ else. We use the same notation for the associated multiplication operator, i.e., $(\mathbb{I}_{\Omega^\pm} f)(x) = \mathbb{I}_{\Omega^\pm}(x) f(x)$. First consider the wave equation since here the geometry is simpler and the results are more complete. Let f be compactly supported in Ω^- and consider the solution $\Phi_t = \Phi_t[f]$ of the wave equation corresponding to the initial data $\Phi_0 = f$, $\partial_t \Phi_{t=0} = -\partial_x f$. This corresponds to sending the wave packet of shape f towards the interaction potential. It is natural to look at the part of the wave that will be scattered back into region Ω^-. Therefore for $q \geqslant 1$ we look at the asymptotic amount of q-energy present in region Ω^-.

$$p_q[f] = \lim_{t \to +\infty} \| \mathbb{I}_{\Omega^-} \Phi_t[f] \|_q^q. \tag{1.5}$$

We shall show that the limits actually exist for suitable initial data. As usual we have used the notation $\|s\|_q^q = \int |s|^q$. Now take a function f in $S_0(\mathbb{R})$ the closed subspace of $S(\mathbb{R})$, the class of Schwartz, consisting of those functions for which all moments vanish. Such functions are prototypes of wavelets (see the next section). Let $f_{b,a}$ stand for the scaled and translated version of f,

$$f_{b,a}(x) = \frac{1}{a} f\left(\frac{x-b}{a}\right).$$

We then explore the potential by sending ''test'' functions located at smaller and smaller scales towards the potential. We fix some $b \in \Omega^-$. To make the functions supported by Ω^- we take a compactly supported, smooth function χ which is identically 1 in some neighborhood of b. Let $f_{(b,a)} = \chi f_{b,a}$. We then consider the following integrated reflection probability:

$$\rho(a) = \int_a^1 \frac{d\alpha}{\alpha^{q+1}} p_q[f_{(b,\alpha)}].$$

As we shall show, we have

$$\lim_{a\to 0} \begin{matrix}\sup \\ \inf\end{matrix} \frac{\log \rho(a)}{\log a} = \kappa_q^{\pm}[V],$$

where $\kappa_q^{\pm}[V]$ are the upper and lower generalized wavelet dimensions of the potential, provided this latter is ''sufficiently'' singular (this will be made precise). Thus if the potential is truly fractal, in the sense that it has nontrivial wavelet dimensions, it may be read from the scattering data. The reason for this result is that the backscattered wave is essentially the wavelet transform of the scattering potential with a wavelet given by a primitive of the initial data, provided the latter is supported in Ω^- (Theorem 11.3). We therefore can recover the truly singular parts of the potential.

In the case of the Schrödinger equation the results are limited to the so-called wavelet correlation dimension $\kappa_2^{\pm}$. On the other side, the potential may be a signed measure, since one need not assume H to be positive as for the wave equation. More precisely, for $\Psi_t = e^{-itH} f$ we define

$$p^{\pm}[f] = \lim_{t\to\infty} \|\mathbb{I}_{\Omega^{\pm}} \Psi_t\|_2^2.$$

Again we shall prove that the limit actually exists. This quantity can be interpreted as follows. When f is normalized (i.e., $\|f\|=1$), $p^{\pm}[f]$ is the final probability of finding a particle described by the wave function Ψ_t in the left and right region, respectively. Consider now a function f in $S_+(\mathbb{R})$, the closed subspace of $S(\mathbb{R})$ having a Fourier transform supported by the positive frequencies only. Again we explore V by sending wavelets at smaller and smaller scales towards the scatterer. More precisely, fix again $b \in \Omega^-$ and consider the following integrated reflection probability

$$\rho(a) = \int_a^1 \frac{d\alpha}{\alpha^3} p^-[f_{b,\alpha}].$$

Then we shall show that

$$\lim_{a\to 0} \begin{matrix}\sup \\ \inf\end{matrix} \frac{\log \rho(a)}{\log a} = \min\{\kappa_2^{\pm}[V], 0\},$$

where $\kappa_2^{\pm}[V]$ are the upper and lower wavelet correlation dimensions of the scatterer. Therefore, again for truly fractal potentials, we may characterize their fractality through a scattering experiment.

The rest of the paper is organized as follows. In Sec. II, we recall the definition of fractal wavelet dimensions. In Secs. III–IX we extend the usual one-dimensional scattering formalism (for which we refer to Refs. 13–15) to singular potentials defined as measures. In Sec. X, we study the time evolution and in Sec. XI we show how to retrieve some specific fractal properties from the long time asymptotic. In Sec. XII we give some numerical illustrations of our results.

II. THE GENERALIZED FRACTAL WAVELET DIMENSIONS

As mentioned in the Introduction, we will be concerned in the scattering problem with potentials given by fractal measures on the line. By ''fractal'' we mean, loosely speaking, a measure which has a nontrivial fractal dimension. There are several definitions of the fractal dimension of a measure, such as the box dimension, the information, and the correlation dimension, to cite only a few. Historically, these dimensions have been introduced in the frame of dynamical systems and ergodic theory and characterize essentially probability measures. Recently,[16] a new set of fractal dimensions has been constructed by means of wavelet transform, namely the *fractal wavelet dimensions*. These dimensions are well adapted to characterize fractality of signed or complex measures. One of them, the so-called wavelet correlation dimension, is particularly suited for scattering problems (see Refs. 11 and 12).

A. Wavelet analysis

The wavelet transform can be seen as a generalization of the Fourier transform; while the latter consists in decomposing a function in pure frequencies, the former allows a decomposition on a basis of wavelets, which are basically functions well localized both in position and frequency and which are oscillating. A typical example would be a function g that is in $L^1(\mathbb{R})\cap L^\infty(\mathbb{R})$ and has zero mean, that is $\int g=0$. We will adopt the following notational convention. For all functions g, we write

$$g_a(x)=a^{-1}g(a^{-1}x), \quad g_{b,a}(x)=a^{-1}g(a^{-1}(x-b)), \quad \tilde{g}(x)=\overline{g(-x)}$$

for the dilated, resp. translated and dilated, resp. symmetric version. The wavelet transform of a function or more generally a distribution s with respect to a wavelet g is given by

$$\mathscr{W}_g s(b,a)=\int dt\ a^{-1}\overline{g(a^{-1}(t-b))}s(t)=\int dt\ \overline{g_{b,a}}(t)s(t). \tag{2.1}$$

This expression can also be seen as a convolution,

$$\mathscr{W}_g s(b,a)=(\tilde{g}_a * s)(b), \tag{2.2}$$

or in the Fourier space,

$$\mathscr{W}_g s(b,a)=\int dk\ \overline{\hat{g}(ak)}e^{ibk}\hat{s}(k). \tag{2.3}$$

Intuitively, the wavelet transform acts as a filter, selecting the details present in s at scale a and position b. It is therefore a tool to analyze the local regularity of functions. Convenient classes of wavelets are the following. Let $\mathscr{S}(\mathbb{R})$ be the Schwartz space of smooth functions which, together with their derivatives, decrease faster than any polynomial, that is

$$\max_{q,m\leqslant N}\ \sup_{x\in\mathbb{R}}|x^q\partial^m g(x)|<\infty, \quad N=0,1,2,\ldots .$$

Let $\mathscr{S}_+(\mathbb{R})$ be the closed subspace of Schwartz functions with no negative frequencies [that is $\hat{g}(\xi)=0$ if $\xi\leqslant 0$], $\mathscr{S}_-(\mathbb{R})$ the subspace of Schwartz functions with no positive frequencies, and $\mathscr{S}_0(\mathbb{R})=\mathscr{S}_+(\mathbb{R})\oplus\mathscr{S}_-(\mathbb{R})$. Functions in $\mathscr{S}_+(\mathbb{R})$, $\mathscr{S}_-(\mathbb{R})$, or $\mathscr{S}_0(\mathbb{R})$ have all moments vanishing, that is

$$\int dx\ x^n g(x)=0, \quad n=0,1,2,\ldots,$$

or, what amounts to the same, have a flat Fourier transform near zero,

$$\hat{g}(\xi)=O(\xi^\infty), \quad \xi\to 0,$$

where the convenient notation "$O(\xi^\infty)$" means $O(\xi^n)$ for all n.

B. The generalized wavelet dimension

The generalized wavelet dimensions are now constructed in the following way. Take some wavelet g in $\mathscr{S}_+(\mathbb{R})$ and set for $q\geqslant 1$

$$G_g(a,q)=\|\mathscr{W}_g\mu(\cdot,a)\|_q^q=\int db|\mathscr{W}_g\mu(\cdot,a)|^q$$

and

$$\Gamma_g(a,q)=\min\left\{\int_a^1 \frac{d\alpha}{\alpha}\,G_g(\alpha,q), \int_0^a \frac{d\alpha}{\alpha}\,G_g(\alpha,q)\right\}.$$

Then the (upper and lower) wavelet dimensions $\kappa_q^{\pm}(\mu)$ are defined by

$$\kappa_q^+(\mu)=\limsup_{a\to 0}\frac{\log \Gamma_g(a,q)}{\log a},\quad \kappa_q^-(\mu)=\liminf_{a\to 0}\frac{\log \Gamma_g(a,q)}{\log a},$$

and do not depend on the chosen wavelet g in $\mathscr{S}_+(\mathbb{R})$ provided $g\neq 0$. Clearly analogous definitions hold for wavelets $\mathscr{S}_-(\mathbb{R})$. For real-valued fractals we obtain the same dimensions. However, for complex values they may be different. Note that a wavelet in $\mathscr{S}_\pm(\mathbb{R})$ cannot have compact support (otherwise its Fourier transform would be analytic and hence could not vanish on a semi-axis without vanishing everywhere). However, admitting an arbitrarily small error on the dimensions $\kappa_q^\pm$, one always may suppose that the analyzing wavelet is compactly supported. This is useful in the applications.

Remark: The dimensions κ_q^- can be interpreted in terms of the limiting regularity in some Besov-space sense. Such an interpretation is not possible for the dimensions κ_q^+. Using the partition function G_g, it would be possible to define the analog of the dimensions κ_q^- independently of the wavelet in $\mathscr{S}_+(\mathbb{R})$. However, the analog definition of κ_q^+ may depend on the wavelet due to lacunarity of the scales. This is why the integrated partition function Γ_g has been introduced in Ref. 16.

A particular role is played by the index $q=2$. The so-called *wavelet correlation dimension* $\kappa_2^\pm$ of a signed measure governs the L^2 asymptotic of its Fourier transform, as shows the following:

Theorem 2.1:[16] *If μ is a finite complex measure, then either $\mu\notin L^2(\mathbb{R})$ and*

$$\lim_{p\to\infty}\sup_{\inf}\frac{\log \int_0^p d\xi\ |\hat{\mu}(\xi)|^2}{\log p}=-\kappa_2^\mp[\mu], \tag{2.4}$$

or $\mu\in L^2(\mathbb{R})$ and

$$\lim_{p\to\infty}\sup_{\inf}\frac{\log \int_p^{+\infty} d\xi\ |\hat{\mu}(\xi)|^2}{\log p}=-\kappa_2^\mp[\mu], \tag{2.5}$$

If μ is real, one can take indiscriminately a wavelet g in $\mathscr{S}_0(\mathbb{R})$, $\mathscr{S}_+(\mathbb{R})$, or $\mathscr{S}_-(\mathbb{R})$.

C. The correlation dimension

The wavelet correlation dimension is actually a generalization of the correlation dimension defined for positive measures as follows.[17] Consider the function

$$\Omega_\mu(r)=\int \mu(dx)\ \mu(B(x,r)),$$

where $B(x,r)$ is the ball of radius r around the point x. The quantities

$$D^+[\mu]=\limsup_{r\to 0}\frac{\log \Omega_\mu(r)}{\log r},\quad D^-[\mu]=\liminf_{r\to 0}\frac{\log \Omega_\mu(r)}{\log r},$$

are called respectively upper and lower correlation dimensions of μ. When the upper and lower dimensions coincide, their common value is simply called the correlation dimension. For positive finite measures, the correlation dimension has been related to the mean quadratic Fourier asymptotics:

Theorem 2.2:[18] *For any finite positive measure μ on $\mathbb{R}$,*

$$\lim_{p\to\infty}\sup_{\inf}\frac{\log \int_{|\xi|\leq p} d\xi\ |\hat{\mu}(\xi)|^2}{\log \xi}=1-D^\mp[\mu]. \tag{2.6}$$

Note that the classical correlation dimension and the wavelet correlation are essentially the same in the case of positive measures. Precisely,

$$-\kappa_2^\pm[\mu]=1-D^\pm[\mu]$$

provided $\mu \notin L^2(\mathbb{R})$, which is always the case for nontrivial singular measures.

For signed or complex measures, the correlation dimensions are not sufficient to characterize the Fourier asymptotic. The relevant dimensions in this case are the upper and lower wavelet dimensions. Intuitively, the reason is the following. To compute the ''classical'' fractal dimensions, one sooner or later tests the measures with boxes, whose size becomes smaller and smaller. However, it makes no sense to test signed measures with boxes because cancellations may appear which corrupt the information. The idea of wavelet dimensions is precisely to replace the boxes by functions which oscillate faster and faster.

III. PRELIMINARIES: SOME PROPERTIES OF BOREL MEASURES

We now expose the scattering formalism for singular potentials. (This whole part is taken from the thesis of one of us.[19] We reproduce it here for the convenience of the reader.) Before entering the core of the paper, we list some properties of Borel measures that are necessary for the following. We refer, e.g., to Ref. 20 for all classical results on measure theory and to e.g., Ref. 21 for the functional analysis. We denote by $C_0^k(\mathbb{R})$ the space of complex-valued k times continuously differentiable functions with compact support on $\mathbb{R}$. $\mathscr{D}=C_0^\infty(\mathbb{R})$ is the space of test functions and $\mathscr{D}'$, its dual, is the space of distributions on $\mathbb{R}$. We write (η, f) for the action of a distribution η on a function f. We use the notation $\eta_n \overset{\mathscr{D}'}{\to} \eta$ to say a sequence of distributions (η_n) converges in the weak topology of $\mathscr{D}'$ towards some other distribution η [that is $(\eta_n, s) \xrightarrow[n\to\infty]{} (\eta, s)$ for any test function s]. We denote by $C_b^0(\mathbb{R})$ the Banach space of continuous bounded functions on $\mathbb{R}$, equipped with the supremum norm $\|\cdot\|_\infty$. We call $\mathscr{M}$ the dual of $C_b^0(\mathbb{R})$, that is the space of finite Radon measures on $\mathbb{R}$. The action of a measure μ in $\mathscr{M}$ on a continuous function f in $C_b^0(\mathbb{R})$ is given by

$$(\mu, f) = \int f(x)\ \mu(dx),$$

where the right-hand side is the Lebuesgue integral of f with respect to μ. We use the notation $\mu_n \overset{\mathscr{M}}{\to} \mu$ to say a sequence of measures (μ_n) converges in the weak topology of $\mathscr{M}$ towards some other measure μ [that is $(\mu_n, f) \xrightarrow[n\to\infty]{} (\mu, f)$ for any $f \in C_b^0(\mathbb{R})$]. A strong topology on $\mathscr{M}$ can also be defined via the norm

$$\|\mu\| = \sup_{\substack{f \in C_b^0(\mathbb{R}) \\ \|f\|_\infty = 1}} \left| \int f\mu(dx) \right| = \int |\mu|(dx) = \text{total mass of the measure } \mu.$$

If μ is a Borel measure, we use the canonical notation μ_{ac}, μ_{sc}, and μ_{pp} for its absolutely continuous part, singular continuous part, and discrete part, respectively. In the following, we will always implicitly assume that the measures are σ-finite Borel measures. When $\mu(\mathbb{R}^-) < \infty$, one may define a ''primitive'' of μ by

$$L_\mu(x) = \mu(]-\infty, x[).$$

Here L_μ is a locally bounded measurable function, which is continuous to the left and admits at most a countable set of discontinuities located at the atomic points of μ [i.e., $\mu(\{x\}) \neq 0$]. Moreover, L_μ is almost everywhere (a.e.) differentiable in the classical sense with

$$L'_\mu(x) = f_\mu(x) \quad \text{a.e.},$$

where $f_\mu \in L^1_{\text{loc}}(\mathbb{R})$ is the Radon–Nykodim derivative of μ_{ac}, that is $\mu_{ac}(dx) = f_\mu(x)\ dx$. Unless explicitly stated, the word ''almost everywhere'' will always refer to the Lebesgue measure.

A function f is said to be absolutely continuous on a set K (possibly unbounded) if, for any $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\sum_{i=1}^{n} |b_i - a_i| < \delta \Rightarrow \sum_{i=1}^{n} |f(b_i) - f(a_i)| < \varepsilon$$

for any finite collection of disjoints intervals $[a_1, b_1],\ldots,[a_n, b_n]$ of K. Absolutely continuous functions can be thought of as primitives of absolutely continuous measures: a function is absolutely continuous on K if and only if it is the indefinite integral of some locally integrable function on K. To distinguish between the classical derivatives and the distributional derivatives of a function, say f, we will write $f', f'', \ldots, f^{(n)}$ for the former and $Df, \Delta f, \ldots, D^{(n)}f$ for the latter. This distinction is especially important for functions such as L_μ, for which $L'_\mu = f_\mu$ a.e. but $DL_\mu = \mu$, as can be seen from the following.

Lemma 3.1: If μ is a Borel measure such that $\mu(\mathbb{R}^-) < \infty$ and $s \in C_0^1(\mathbb{R})$, then we have the following ''integration by part'' formula:

$$\int dx\ s'(x) L_\mu(x) = -\int s(x)\mu(dx). \tag{3.1}$$

Hence, μ is the distributional derivative of L_μ in $\mathscr{D}$.

Proof: The proof is a standard exercise of functional analysis and we leave it to the reader.□

The two notions of derivation for L_μ coincide if μ is absolutely continuous. In fact, we have the following simple but convenient criterion for absolute continuity.

Lemma 3.2: Let $\eta \in \mathscr{D}'$. If $D\eta = f$, where $f \in L^1_{\text{loc}}(\mathbb{R})$ (locally integrable functions), then η is absolutely continuous on every compact.

Proof: Obvious. □

The differentiation rule for the product has to be adapted to the functions of type L_μ. If f and g are absolutely continuous functions, it is well known that fg is also absolutely continuous with $(fg)' = f'g + fg'$ a.e. An easy generalization of this is the following.

Lemma 3.3: Let μ be a Borel measure such that $\mu(\mathbb{R}^-) < \infty$ and g be an absolutely continuous function. Then the function $G = gL_\mu$ is a.e. differentiable and

$$G'(x) = g'(x)L_\mu(x) + g(x)f_\mu(x) \text{a.e.}, \tag{3.2}$$

where $\mu_{ac}(dx) = f_\mu(x)dx$.

Proof: Obvious. □

If μ is a Borel measure such that $\mu(\mathbb{R}^+) < \infty$, one can define a primitive by integrating to the right:

$$R_\mu(x) = \mu([x, +\infty[).$$

All the results stated for L_μ can be translated in an obvious way to R_μ. Here and everywhere, we will adopt the notational convention

$$\int_{-\infty}^{x} \mu(dt) := \mu(]-\infty, x[), \quad \int_{x}^{+\infty} \mu(dt) := \mu([x, +\infty[), \quad \int_{a}^{b} \mu(dt) := \mu([a, b[), \quad a < b.$$

Note that this notation preserves the chain rule for integrals.

We end this section with a useful identity. If $h \in L^1(\mathbb{R})$, it is well known that

$$\int_{x}^{+\infty} dx_1\ h(x_1) \int_{x_1}^{+\infty} dx_2 \cdots \int_{x_{p-1}}^{+\infty} dx_p\ h(x_p) = \frac{1}{p!} \left(\int_{x}^{+\infty} dy\ h(y) \right)^p,$$

and if, furthermore, $\int dx |x h(x)| < \infty$, then

$$\int_{x}^{+\infty} (x_1 - x)|h(x_1)|\ dx_1 \int_{x_1}^{+\infty} dx_2 \ldots \int_{x_{p-1}}^{+\infty} (x_p - x_{p-1})\ dx_p |h(x_p)| \leqslant \frac{1}{p!} \left(\int_{x}^{+\infty} dy (y - x)|h(y)| \right)^p.$$

The last inequality can be extended to measures as follows.

Lemma 3.4: If μ is a positive Borel measure such that $\int (1 + |x|)\mu(dx) < \infty$, then

$$\int_x^{+\infty}(x_1-x)\ \mu(dx_1)\cdots\int_{x_{p-1}}^{\infty}(x_p-x)\ \mu(dx_p)\leqslant\frac{1}{p!}\left(\int_x^{+\infty}dy(y-x)\ \mu(dy)\right)^p. \tag{3.3}$$

Proof: First suppose the measure μ is of compact support. Let us introduce the function

$$K(t)=\begin{cases}0 & \text{if } t>0,\\ -t & \text{if } t\leqslant 0.\end{cases}$$

Then

$$G^1_\mu(x)=K*\mu(x)=\int_x^{+\infty}(y-x)\ \mu(dy)$$

is a continuous bounded function. Now take some non-negative function $h\in C_0^\infty(\mathbb{R})$ such that $\int h=1$ and set $h_n(x)=nh(nx)$. Then h_n is an approximate identity. So if we define $\mu_n=\mu*h_n$ and $G^1_{\mu_n}=K*\mu_n=G^1_\mu*h_n$, then $\mu_n\overset{\mathscr{M}}{\to}\mu$ and $G^1_{\mu_n}(x)\to G^1_\mu(x)$ for all x. What is more, we claim that $\mu_nG^1_{\mu_n}\overset{\mathscr{M}}{\to}\mu G^1_\mu$. Indeed, for all $f\in C_b^0(\mathbb{R})$ we have

$$|(\mu_nG^1_{\mu_n},f\ )-(\mu G^1_\mu,f\ )|\leqslant\|\mu_n\|\sup_{x\in\Omega}|f(x)(G^1_{\mu_n}(x)-G^1_\mu(x))|+|(\mu_n-\mu,G^1_\mu s)|,$$

where Ω is some compact including all the supports of μ_n. The second term on the right-hand side converges to zero. On the other hand, the function $G^1_{\mu_n}$ is easily seen to be Lipschitz with a Lipschitz constant $2\|\mu_n\|$. Since the sequence μ_n is strongly bounded for it is weakly convergent, there is a common Lipschitz constant (precisely $\sup_n\|\mu_n\|$) for all the $G^1_{\mu_n}$ and thus this family of functions is equicontinuous. It follows that $G^1_{\mu_n}\to G^1_\mu$ uniformly on every compact and thus the first term converges to zero. Iterating this procedure, one can show that $\mu_nG^p_{\mu_n}\overset{\mathscr{M}}{\to}\mu G^p_\mu$, where G^p_μ and $G^p_{\mu_n}$ are defined recursively by

$$G^p_\mu(x)=K*(\mu G^{p-1}_\mu)(x),\quad G^p_{\mu_n}(x)=K*(\mu_nG^{p-1}_{\mu_n})(x).$$

It follows that

$$\begin{aligned}\int_x^{+\infty}(x_1-x)\ \mu_n(dx_1)\cdots\int_{x_{p-1}}^{\infty}(x_p-x)\ \mu_n(dx_p)&=K*(\mu_nG^p_{\mu_n})(x)\\ &\to K*(\mu G^p_\mu)(x)\\ &=\int_x^{+\infty}(x_1-x)\ \mu(dx_1)\cdots\int_{x_{p-1}}^{\infty}(x_p-x)\ \mu(dx_p)\end{aligned}$$

for all x. Since the formula (3.3) holds for μ_n, this proves the lemma for μ. If μ is not compactly supported, it suffices to take a sequence of truncations $\mu_n(dx)=(1-\theta_n(x))\ \mu(dx)$, where θ_n is the Heaviside function centered at n [$\theta_n(x)=1$ if $x\geqslant n$,0 otherwise] and to repeat the same arguments as previously. □

Remark: Surprisingly, the analog formula

$$\int_x^{+\infty}\mu(dx_1)\cdots\int_{x_{p-1}}^{\infty}\mu(dx_p)=\frac{1}{p!}\left(\int_x^{+\infty}\mu(dy)\right)^p$$

is false, as can be easily seen on the counterexample of the Dirac function. The difference is essentially due to the fact that G_μ is more regular than R_μ.

IV. THE HAMILTONIAN

Now we attack the scattering problem. We equip $L^2(\mathbb{R})$ with the usual scalar product $\langle f,g\rangle = \int fg$ and adopt the notation $\|f\|$ for the induced norm. Let H_0 be the self-adjoint realization of the negative Laplacian $-\Delta$ on $L^2(\mathbb{R})$, with domain $D(H_0)=W^2(\mathbb{R})$, where $W^p(\mathbb{R})$ is the Sobolev space of functions such that

$$\int dk(1+k^{2p})|\hat{f}(k)|<\infty.$$

Let V be a finite real Borel measure on $\mathbb{R}$. The operator sum $H=H_0+V$ is in general not defined on $D(H_0)$, but there exists a unique self-adjoint operator H with domain $D(H)\subset W^1(\mathbb{R})$ corresponding to this formal expression in the sense of quadratic forms (see, e.g., Ref. 22). The action of this operator is given by

$$\langle f,Hg\rangle = \int dx\ \overline{f'(x)}g'(x) + \int \overline{f(x)}g(x)\ V(dx)$$

for all $f\in W^1(\mathbb{R})$ and $g\in D(H)$. The domain $D(H)$ is the set of functions $f\in W^1(\mathbb{R})$ such that the distribution $(-\Delta+V)f\in L^2(\mathbb{R})$. The eigenvectors are those functions $f\in W^1(\mathbb{R})$ for which the distribution $(-\Delta+V-E)f=0$ for some real E.

Throughout this paper, we will assume that V is a real Borel measure satisfying the typical short-range condition

$$\int (1+|x|^2)\ |V|(dx)<\infty. \tag{4.1}$$

For Hamiltonians with such singular interactions, it is known that the spectral and scattering theory do not qualitatively differ from the regular case. The essential spectrum of H and H_0 coincide and there is no singular continuous spectrum (see Ref. 23 for the dimension one and Ref. 24 for the general case); the wave operators exist and are complete.[23,25]

In order to compute the time evolution generated by the Hamiltonian H, we need an explicit formula for the spectral measure and/or an eigenfunction expansion. This can be achieved by means of the so-called generalized eigenfunctions. We begin by introducing the usual definitions for smooth potentials. First suppose $V(x)$ is some continuous real function with compact support and V is the associated multiplication operator. Then the Hamiltonian $H=H_0+V$ is a self-adjoint operator on $D(H)=D(H_0)$. The generalized eigenfunctions associated to this operator are solutions of the stationary Schödinger equation or, what amounts to the same, the Helmholtz equation,

$$(\Delta+z^2)\varphi = V\varphi, \tag{4.2}$$

where $z>0$ and φ satisfies an appropriate asymptotic condition (radiation condition). For the spectral analysis of H, it is useful to let the parameter z run in the complex upper half-plane. We define the family of functions $\varphi_z^{\pm}$, with $\Im z\geqslant 0$, as the unique solutions of the Helmholtz equation (4.2) with asymptotic behavior

$$\varphi_z^{\pm}(x)\to\begin{cases} t_z^{\pm}e^{\pm izx}, & x\to\pm\infty,\\ e^{\pm izx}+r_z^{\pm}e^{\mp izx}, & x\to\mp\infty.\end{cases} \tag{4.3}$$

The coefficients t_z^+ and t_z^- are called transmission coefficients from the left and from the right, respectively, and the coefficients r_z^+ and r_z^- are the reflection coefficients from the left and from the right, respectively. The functions $\varphi_z^{\pm}$ are generalized functions in the sense that $H\varphi_z^{\pm} = z^2\varphi_z^{\pm}$, formally because $\varphi_z^{\pm}\notin L^2(\mathbb{R})$. For $z\neq 0$, the Helmholtz equation (4.2) together with asymptotic condition (4.3) is equivalent to the integral equation

$$\varphi_z^{\pm}(x) = e^{\pm izx} + \frac{1}{2iz}\int dy\ e^{iz|x-y|}V(y)\varphi_z^{\pm}(y). \tag{4.4}$$

Indeed, the above function verifies the Helmholtz equation and has the same asymptotic as $\varphi_z^{\pm}$, so by the uniqueness of the solution it is $\varphi_z^{\pm}$. This integral equation may be used to derive an expression of the transmission and reflection coefficients. Letting $x \to \pm\infty$ in (4.4) and comparing with (4.3) gives

$$t_z^{\pm} = 1 + \frac{1}{2iz} \int dy\ e^{\mp izy} V(y) \varphi_z^{\pm}(y), \tag{4.5}$$

$$r_z^{\pm} = \frac{1}{2iz} \int dy\ e^{\pm izy} V(y) \varphi_z^{\pm}(y). \tag{4.6}$$

When the potential is a measure, we cannot treat the Helmholtz equation as an ordinary differential equation any more. However, the integral equation (4.4) can still make sense if we understand $V(y)$ as a distribution. So, the idea is to seek a solution of (4.4) and verify *a posteriori* that it has the same properties as the usual generalized eigenfunctions for the Hamiltonian defined in the sense of quadratic forms. We define the generalized eigenfunctions of our problem as the continuous bounded solutions of

$$\varphi_z^{\pm}(x) = e^{\pm izx} + \frac{1}{2iz} \int e^{iz|x-y|}\ V(dy) \varphi_z^{\pm}(y), \quad \mathcal{I}z \geqslant 0. \tag{4.7}$$

We will study the existence and analycity properties of these functions, which will then allow us to compute explicitly the spectral measure and the wave operators and to establish an eigenfunction expansion.

V. THE JOST FUNCTIONS

Following the standard method used for regular potentials, we will first study a more convenient family of functions, namely the Jost functions $J_z^{\pm}$, that we define as the solutions of

$$J_z^{+}(x) = e^{izx} + \int_x^{+\infty} \frac{\sin(z(y-x))}{z} J_z^{+}(y)\ V(dy), \tag{5.1}$$

$$J_z^{-}(x) = e^{-izy} + \int_{-\infty}^{x} \frac{\sin(z(x-y))}{z} J_z^{-}(y)\ V(dy). \tag{5.2}$$

To remove the trivial oscillatory part, it is convenient to introduce the functions $j_z^{\pm}(x) = e^{\mp izx} J_z^{\pm}(x)$. These functions satisfy the modified integral equations

$$j_z^{+}(x) = 1 + \int_x^{+\infty} e^{iz(y-x)} \frac{\sin(z(y-x))}{z} j_z^{+}(y)\ V(dy), \tag{5.3}$$

$$j_z^{-}(x) = 1 + \int_{-\infty}^{x} e^{iz(x-y)} \frac{\sin(z(x-y))}{z} j_z^{-}(y)\ V(dy). \tag{5.4}$$

In order to solve these integral equations, we introduce the formal operators

$$G_z^{+} f(x) = \int_x^{+\infty} e^{iz(y-x)} \frac{\sin(z(y-x))}{z} f(y)\ V(dy), \tag{5.5}$$

$$G_z^{-} f(x) = \int_{-\infty}^{x} e^{iz(x-y)} \frac{\sin(z(x-y))}{z} f(y)\ V(dy). \tag{5.6}$$

Note that $G_z^{\pm}$ can also be defined in $z=0$. The properties of these operators are summarized in the following theorem.

Theorem 5.1: *Let V be a real Borel measure satisfying condition (4.1). Then, we have the following.*

(a) *For $\Im z \geqslant 0$, $G_z^{\pm}$ and $(1-G_z^{\pm})^{-1}$ are well-defined operators in $\mathscr{B}(C_b^0(\mathbb{R}))$, the space of bounded operators on $C_b^0(\mathbb{R})$. They are uniformly bounded in the upper half-plane $\Im z \geqslant 0$.*

(b) *The operator-valued functions $z \mapsto G_z^{\pm}$ and $z \mapsto (1-G_z^{\pm})^{-1}$ are analytic in $\Im z > 0$, continuous in $\Im z \geqslant 0$, and continuously differentiable on $\Im z \geqslant 0$, $z \neq 0$. Their derivatives are uniformly bounded in $\Im z \geqslant 0$.*

Proof: (a) Let us introduce the kernels $g_z^{\pm}(x) = \mp \theta_0(\mp x) e^{\mp izx} z^{-1} \sin(zx)$, where as before θ_0 is the Heaviside function centered at 0. For $\Im z \geqslant 0$, these functions satisfy the following estimates, which can be obtained by elementary but tedious computations (here the upper dot indicates the derivative with respect to the variable z):

$$|g_z^{\pm}(x)| \leqslant C\theta_0(\mp x)\frac{|x|}{1+|zx|} \leqslant C\theta_0(\mp x)|x|, \tag{5.7}$$

$$|\dot{g}_z^{\pm}(x)| \leqslant C\theta_0(\mp x)\left(|x| + \frac{|x|^2}{1+|zx|}\right) \leqslant C\theta_0(\mp x)(1+|x|^2), \tag{5.8}$$

$$|\ddot{g}_z^{\pm}(x)| \leqslant C\theta_0(\mp x)\left(|x| + |x|^2 + \frac{|x|^3}{1+|zx|}\right) \leqslant C\theta_0(\mp x)\left(1 + |x|^2 + \frac{|x|^2}{|z|}\right), \tag{5.9}$$

for some constant $C > 0$. Now, $G_z^{\pm} f$ can be rewritten as the convolution of a bounded continuous function with a finite measure, $G_z^{\pm} f = g_z^{\pm} * (Vf)$, and so $G_z^{\pm} f$ is itself continuous and bounded. Let us show that it is in fact uniformly bounded. For, note that

$$\sup_{x \in \mathbb{R}} \int_x^{+\infty} (y-x)\ |V|(dy) < \infty, \quad \sup_{x \in \mathbb{R}} \int_x^{+\infty} (y-x)^2\ |V|(dy) < \infty, \tag{5.10}$$

as soon as condition (4.1) is verified. Using the estimations (5.7) and (5.10) we obtain

$$|G_z^{\pm} f(x)| \leqslant C_1 \|f\|_{\infty} \int \theta_0(\pm(x-y))|x-y|\ |V|(dy) < C_2 \|f\|_{\infty}$$

for some constants C_1, $C_2 > 0$. Thus, $G_z^{\pm}$ is uniformly bounded on $\Im z \geqslant 0$. To prove the existence of $(1-G_z^{\pm})^{-1}$, we will show that its Neumann series is absolutely convergent. We detail the proof for $(1-G_z^{+})^{-1}$, the proof for $(1-G_z^{-})^{-1}$ being similar. Let us form the nth iterate

$$(G_z^{+})^n f(x) = \int_x^{+\infty} g_z^{+}(x_1 - x)\ V(dx_1) \int_{x_1}^{+\infty} \cdots \int_{x_{n-1}}^{+\infty} g_z^{+}(x_n - x_{n-1}) f(x_n)\ V(dx_n).$$

Since $|g_z^{\pm}(x)| \leqslant c|x|$ for some constant $c > 0$, we have

$$|(G_z^{+})^n f(x)| \leqslant c^n \|f\|_{\infty} \int_x^{+\infty} (x_1 - x)\ |V|(dx_1) \int_{x_1}^{+\infty} \cdots \int_{x_{n-1}}^{+\infty} (x_n - x_{n-1})\ |V|(dx_n),$$

and it follows from Lemma 3.4 that

$$|(G_z^{+})^n f(x)| \leqslant \|f\|_{\infty} \frac{c^n}{n!} \left(\int_x^{+\infty} (y-x)\ |V|(dy) \right)^n \leqslant \frac{C^n}{n!}$$

for some constant $C > c > 0$. Thus,

$$\|(G_z^{+})^n\| \leqslant \frac{C^n}{n!}$$

in the operator norm of $\mathscr{B}(C_b^0(\mathbb{R}))$. It follows that the Neumann series $\Sigma(G_z^+)^n$ is uniformly absolutely convergent and so the inverse $(1-G_z^+)^{-1}$ exists and its norm is uniformly bounded.

(b) Let us define the operator $\dot{G}_z^\pm$, $\Im z \geqslant 0$, by

$$\dot{G}_z^\pm f(x) = \int \dot{g}_z^\pm(x-y) f(y)\ V(dy).$$

Using (5.8) and (5.10), one can easily check that $\dot{G}_z^\pm$ is a well-defined operator on $\mathscr{B}(C_b^0(\mathbb{R}))$, uniformly bounded on $\Im z \geqslant 0$. We omit the details. For $\Im z > 0$, one can show by direct computation that $\dot{G}_z^\pm$ is the norm derivative of $G_z^\pm$ and $(1-G_z^\pm)^{-1}\dot{G}_z^\pm(1-G_z^\pm)^{-1}$ is the norm derivative of $(1-G_z^\pm)^{-1}$. So $G_z^\pm$ and $(1-G_z^\pm)^{-1}$ are norm differentiable on $\Im z > 0$, hence analytic there. Moreover, by virtue of (5.7) and (5.10), we have, for any $\varepsilon > 0$,

$$\begin{aligned}
|(G_{k+i\varepsilon}^\pm - G_k^\pm) f(x)| &\leqslant \|f\|_\infty \int |g_{k+i\varepsilon}^\pm(x-y) - g_k^\pm(x-y)|\ |V|(dy) \\
&\leqslant \varepsilon \|f\|_\infty \int \sup_{\Im z \geqslant 0} |\dot{g}_z^\pm(x-y)|\ |V|(dy) \\
&\leqslant \varepsilon C \|f\|_\infty \int \theta_0(\mp(x-y))(1+|x-y|^2)\ |V|(dy) \\
&\leqslant \varepsilon C' \|f\|_\infty
\end{aligned}$$

for some constants C', $C > 0$, and thus $G_z^\pm$ is continuous on $\Im z \geqslant 0$. That $(1-G_z^\pm)^{-1}$ is also continuous on $\Im z \geqslant 0$ follows from the second resolvent formula

$$\begin{aligned}
\|(1-G_{k+i\varepsilon})^{-1} - (1-G_k)^{-1}\| &= \|(1-G_{k+i\varepsilon})^{-1}(G_{k+i\varepsilon} - G_k)(1-G_k)^{-1}\| \\
&\leqslant \|(1-G_{k+i\varepsilon})^{-1}\| \|G_{k+i\varepsilon} - G_k\| \|(1-G_k)^{-1}\|.
\end{aligned}$$

In the same manner, one can prove using (5.9) and (5.10) that $\dot{G}_z^\pm$ is continuous on $\Im z \geqslant 0$, $z \neq 0$, as well as $(1-G_z^\pm)^{-1}\dot{G}_z^\pm(1-G_z^\pm)^{-1}$. This concludes the proof of Theorem 5.1. □

Corollary 5.2: For $\Im z \geqslant 0$, the equations (5.1) and (5.2) admit unique continuous solutions $J_z^\pm$. They are given by

$$J_z^\pm(x) = e^{\pm izx}[(1-G_z^\pm)^{-1} I](x),$$

where $I \equiv 1$ is the constant function one. The functions $J_z^\pm$ are analytic in $\Im z > 0$, continuous in $\Im z \geqslant 0$, and continuously differentiable on $\Im z \geqslant 0$, $z \neq 0$. Moreover, the functions $J_k^\pm(x)$ and $(1+|x|)^{-1}\dot{J}_k^\pm(x)$ are uniformly bounded on the real k axis, and we have the symmetry $\overline{J_k^\pm} = J_{-k}^\pm$.

Proof: The existence, analyticity, and boundedness properties of $J_z^\pm$ follow from Theorem 5.1. Moreover, since V is real we have, for real k, $\overline{G_k^\pm I} = G_{-k}^\pm I$, and also $\overline{(G_k^\pm)^n I} = (G_{-k}^\pm)^n I$ for all n. This property carries on the Neumann series, leading to $\overline{J_k^\pm} = J_{-k}^\pm$. □

Next we study the asymptotic properties of the Jost functions. Letting $x \to \pm\infty$ and using the short-range condition (4.1) together with (5.7), a straightforward computation shows that

$$J_z^\pm(x) = \begin{cases} e^{\pm izx} + O(x^{-2} e^{\mp \Im z x}), & x \to \pm\infty, \\ \alpha_z^\pm e^{\pm izx} + \beta_z^\pm e^{\mp izx} + O(x^{-2} e^{\pm \Im z x}), & x \to \mp\infty, \end{cases} \tag{5.11}$$

where

$$\alpha_z^\pm = 1 - \frac{1}{2iz} \int j_z^\pm(y)\ V(dy), \quad \beta_z^\pm = \int \frac{e^{\pm 2izy}}{2iz}\, j_z^\pm(y)\ V(dy). \tag{5.12}$$

The coefficients $\alpha_z^\pm$ and $\beta_z^\pm$ satisfy the following.

Proposition 5.3: For $\Im z \geqslant 0$, the coefficients $\alpha_z^\pm$ verify

$$\alpha_z^+ = \alpha_z^- =: \alpha_z, \tag{5.13}$$

and for real k we have the additional relations

$$\beta_k^+ = -\overline{\beta_k^-}, \quad |\alpha_k|^2 - |\beta_k^\pm|^2 = 1. \tag{5.14}$$

Proof: The classical method (see, e.g., Ref. 26) to prove (5.14) relies upon the properties of the Wronskian associated to the functions $J_z^\pm$. In our case the manipulations with the Wronskian become touchy because of the poor regularity of the involved functions. Therefore, we have to proceed more carefully. We start with the following.

Lemma 5.4: For $\Im z \geqslant 0$, the functions $J_z^\pm$ are absolutely continuous on every compact.

Proof: The functions $J_z^\pm$ satisfy the following estimates as $x \to \pm\infty$:

$$|J_z^\pm(x) - e^{\pm izx}| \leqslant \mathcal{O}(1/x^2), \quad |J_z^\pm(x)| \leqslant \mathcal{O}(1).$$

So we may define the functions

$$F_z^-(x) = \int_{-\infty}^{x} J_z^-(t)\ V(dt) - z^2 \int_{-\infty}^{x} dt (J_z^-(t) - e^{-izt}) - ize^{-izx},$$

$$F_z^+ = (x) = -\int_{x}^{+\infty} J_z^+(t)\ V(dt) + z^2 \int_{x}^{+\infty} dt (J_z^+(t) - e^{+izt}) + ize^{+izx}.$$

Using Lemma 3.1 we see that $DF_z^\pm = (V - z^2) J_z^\pm = \Delta J_z^\pm$. Hence $DJ_z^\pm = F_z^\pm + C^\pm$ for some constant $C^\pm$. To estimate the constants, take some sequence of test functions $s_n^\pm(x) = s(x \mp n)$, with $s \in C_0^\infty$. Then

$$(DJ_z^\pm, s_n^\pm) = -(J_z^\pm, s_n'^\pm) \xrightarrow{n \to \pm\infty} -(e^{\pm izx}, s_n'^\pm) = \pm iz (e^{\pm izx}, s_n^\pm).$$

On the other hand, $(F_z^\pm, s_n^\pm) \xrightarrow{n \to \pm\infty} \pm iz (e^{\pm izx}, s_n^\pm)$. Hence $C^\pm = 0$ and thus $DJ_z^\pm = F_z^\pm$. Now $F_z^\pm$ is locally bounded and measurable, hence locally integrable. It follows from Lemma 3.2 that $J_z^\pm$ is absolutely continuous. This concludes the proof of Lemma 5.4. □

The functions $F_z^\pm(x)$ are defined for all x and so we may define everywhere the ''Wronskian''

$$\mathcal{W}[J_z^+, J_z^-](x) = J_z^+(x) F_z^-(x) - J_z^-(x) F_z^+(x).$$

A priori, $\mathcal{W}[J_z^+, J_z^-]$ is only measurable with a countable number of discontinuities. However, a cancellation of the singularities makes the Wronskian more regular than expected.

Lemma 5.5: $\mathcal{W}[J_z^+, J_z^-](x)$ is absolutely continuous on every compact.

Proof: We keep the notations of Lemma 5.4. For any x we can rewrite

$$\mathcal{W}[J_z^+, J_z^-](x) = w_1(x) + w_2(x),$$

where

$$w_1(x) = J_z^+(x) \int_{-\infty}^{x} J_z^-(t)\ V(dt) + J_z^-(x) \int_{x}^{+\infty} J_z^+(t)\ V(dt)$$

and w_2 is an absolutely continuous function. So it suffices to prove the absolute continuity of w_1. After some splittings and rearrangements of integrals, we obtain, for any $b > a$,

$$w(b) - w(a) = [J_z^+(b) - J_z^+(a)] \int_{-\infty}^{a} J_z^-(t)\ V(dt) + \int_{a}^{b} [J_z^+(b) J_z^-(t) - J_z^+(t) J_z^-(a)] V(dt)$$

$$+ [J_z^-(b) - J_z^-(a)] \int_{b}^{+\infty} J_z^+(t)\ V(dt) = \mathrm{Term}_1 + \mathrm{Term}_2 + \mathrm{Term}_3.$$

We then estimate each term independently:

$$|\mathrm{Term}_1| \leqslant \|V\| e^{\Im z b} \|j_z^-\|_\infty |J_z^+(b) - J_z^+(a)|,$$

$$|\mathrm{Term}_2| \leqslant \|j_z^+\|_\infty e^{-\Im z a} \int_a^b |J_z^-(t) - J_z^-(a)|\,|V|(dt) + \|j_z^-\|_\infty e^{\Im z b} \int_a^b |J_z^+(t) - J_z^+(b)|\,|V|(dt),$$

$$|\mathrm{Term}_3| \leqslant \|V\| \|j_z^+\|_\infty e^{-\Im z a} |J_z^-(b) - J_z^-(a)|\,.$$

Now the functions $J_z^\pm$ are themselves absolutely continuous on every bounded interval $[a, b]$. So given $\varepsilon > 0$, there exists some $\delta > 0$ such that for any finite collection of disjoint segments $[a_1, b_1], [a_2, b_2], \ldots, [a_n, b_n]$ of $[a, b]$,

$$\sum |b_i - a_i| < \delta \Rightarrow \sum |J_z^\pm(b_i) - J_z^\pm(a_i)| < \varepsilon.$$

For such a_i and b_i we have, in view of the previous estimations,

$$\sum |w(b_i) - w(a_i)| \leqslant 2\|V\| (e^{\Im z b} \|j_z^-\|_\infty + e^{-\Im z a} \|j_z^+\|_\infty) \varepsilon,$$

which proves the absolute continuity of the function w_1 on the interval $[a, b]$. This concludes the proof of Lemma 5.5. □

Corrolary 5.6: The Wronskian is constant.

Proof: $J_z^\pm$ is absolutely continuous and $F_z^\pm$ is the sum of a smooth function and a function of type L_μ, so we may apply the differentiation rule of Lemma 3.3 to the Wronskian:

$$\mathscr{W}[J_z^+, J_z^-]'(x) = J_z^+(x) F_z^{-\prime}(x) + J_z^{+\prime}(x) F_z^-(x) - J_z^-(x) F_z^{+\prime}(x) + J_z^{-\prime}(x) F_z^+(x).$$

Now $F_z^{\pm\prime} = J_z^\pm f_V - z^2 J_z^\pm$ a.e. (where f_V is the density of V_{ac}) and $J_z^{\pm\prime} = F_z^\pm$ a.e. Using these identities, we obtain $\mathscr{W}[J_z^+, J_z^-]'(x) = 0$ a.e. Since $\mathscr{W}[J_z^+, J_z^-](x)$ is absolutely continuous on every compact, it follows that $\mathscr{W}[J_z^+, J_z^-](x)$ is constant. □

The value of the Wronskian can be computed by letting $x \to \pm\infty$. This yields

$$\mathscr{W}[J_z^+, J_z^-](x) = -2iz\alpha_z^+ = -2iz\alpha_z^-, \tag{5.15}$$

proving $\alpha_z^+ = \alpha_z^-$. For real z, one can derive other relations. Since the measure V is real, the functions $J_k^\pm$ and $\overline{J_k^\pm}$ satisfy for real k the same equation (6.4), so we may apply the same method (that is, the Wronskian method) to the pairs of solutions $\mathscr{W}[J_k^\pm, \overline{J_k^\pm}]$ and $\mathscr{W}[J_k^\pm, \overline{J_k^\mp}]$. This yields (5.14) and completes the proof of Proposition 5.3. □

We end the study of the Jost function with the usual characterization of bound states.

Proposition 5.7: α_z has a finite number of zeros in the upper half-plane. They lie on the positive imaginary axis, away from zero. If z_0 is a zero of α_z, then z_0^2 is an eigenvalue of the Hamiltonian H.

Proof: By the mere definition, we see that $\alpha_z = 0$ if and only if $\int j_z^\pm(t)\ V(dt) = 2iz$. Now, $j_z^\pm$ is uniformly bounded in the upper half-plane and thus $|z| = |\int j_z^\pm(t)\ V(dt)|/2 \leqslant M$ for some constant $M > 0$. It follows that the zeros of α_z lie within some semicircle $|z| \leqslant M$, $\Im z \geqslant 0$. Moreover, the condition $|\alpha_k|^2 - |\beta_k^\pm|^2 = 1$ when $k \in \mathbb{R}$ forces $|\alpha_k| \geqslant 1$. Since α_z is continuous on $\Im z \geqslant 0$, this means that the possible zeros are relegated to some compact region $|z| \leqslant M$, $\Im z \geqslant \varepsilon > 0$, in which α_z is analytic. Hence, there is only a finite number of zeros in the upper half-plane. Now suppose z_0 is a zero of α_z. Then $J_{z_0}^\pm$ is in $L^2(\mathbb{R})$. Since $J_z^\pm$ is also a solution of the weak Helmholtz equation (6.4), this implies that z_0^2 is an eigenvalue of H and $J_{z_0}^\pm$ is an associated eigenvector. Since H is self-adjoint, this implies $z_0^2 \in \mathbb{R}$. This concludes the proof of Proposition 5.7 and Theorem 6.1. □

VI. THE GENERALIZED EIGENFUNCTIONS

Let us now return to the generalized eigenfunctions $\varphi_z^\pm$. Set $\varphi_z^\pm = J_z^\pm / \alpha_z$ whenever $\alpha_z \neq 0$. Then

$$\alpha_z = 1 - \frac{1}{2iz}\int j_z^{\pm}(y)\ V(dy) = \left(1 + \frac{1}{2iz}\int e^{\mp izy}\varphi_z^{\pm}(y)\ V(dy)\right)^{-1} =: t_z^{-1}$$

and, using the integral equations (5.1)/(5.2), $\varphi_z^{\pm}$ is easily seen to satisfy (4.7). Conversely, if $\varphi_z^{\pm}$ is a continuous solution of (4.7), then the same computation shows that $t_z^{-1}\varphi_z^{\pm}$ verifies the integral equation (5.1)/(5.2) and thus coincides with $J_z^{\pm}$ by the uniqueness of the solution. Thus, the integral equation (4.7) admits a unique continuous solution $\varphi_z^{\pm}$ as soon as $\alpha_z \neq 0$. Summing up all the results we have obtained for $J_z^{\pm}$ and translating them to $\varphi_z^{\pm}$ we can state the following.

Theorem 6.1: *For $\Im z \geqslant 0$ and $\alpha_z \neq 0$, the functions $\varphi_z^{\pm}(x)$ have the following asymptotics,*

$$\varphi_z^{\pm}(x) \to \begin{cases} t_z e^{\pm izx}, & x \to \pm\infty, \\ e^{\pm izx} + r_z^{\pm} e^{\mp izx}, & x \to \mp\infty, \end{cases} \tag{6.1}$$

for complex numbers $t_z = \alpha_z^{-1}$ and $r_z^{\pm} = \beta_z^{\pm}/\alpha_z$ that we refer to as the transmission and reflection coefficients, respectively. If the potential V has compact support, this asymptotic becomes exact outside the support. For fixed x, the functions $\varphi_z^{\pm}(x)$ are meromorph in $\Im z > 0$ with a finite number of poles located on the imaginary axis, corresponding to the zeros of α_z. Moreover, the functions $\varphi_z^{\pm}(x)$ are continuous on the closed upper half-plane $\Im z \geqslant 0$ and continuously differentiable on $\Im z \geqslant 0$, $z \neq 0$. The same result holds for the transmission and reflection coefficients. On the real k axis, the functions $\varphi_k^{\pm}(x)$ and $(1+|x|)^{-1}\dot{\varphi}_k^{\pm}(x)$ are uniformly bounded and we have the additional properties

$$\overline{\varphi_k^{\pm}} = \varphi_{-k}^{\pm}, \quad \overline{t_k} = t_{-k}, \quad \overline{r_k^{\pm}} = r_{-k}^{\pm}, \tag{6.2}$$

and

$$r_k^{+}\overline{t_k} + t_k\overline{r_k^{-}} = 0, \quad |t_k|^2 + |r_k^{\pm}|^2 = 1. \tag{6.3}$$

Remark: As in the regular case, the relations (6.3) express the unitarity of the *scattering matrix*

$$S = \begin{pmatrix} t_k & r_k^{-} \\ r_k^{+} & t_k \end{pmatrix}.$$

The terminology ''generalized eigenfunctions'' for $\varphi_z^{\pm}$ is justified by the following fact.

Proposition 6.2: For $\Im z \geqslant 0$ and $\alpha_z \neq 0$, the functions $\varphi_z^{\pm}$ satisfy the weak Helmholtz equation

$$(\Delta + z^2)\varphi_z^{\pm} = V\varphi_z^{\pm} \quad in \quad \mathscr{D}'. \tag{6.4}$$

Proof: Let us define the kernel $K_z(x) = (2iz)^{-1}e^{iz|x|}$. When $\alpha_z \neq 0$, the integral equation (4.7) satisfied by $\varphi_z^{\pm}$ can be rewritten as $\varphi_z^{\pm}(x) = e^{\pm izx} + K_z * (V\varphi_z^{\pm})(x)$, where the convolution is well defined since

$$|K_z * (V\varphi_z^{\pm})(x)| = \left|\frac{1}{2iz\alpha_z}\int e^{iz|x-y|}e^{\pm iz(y-x)}j_z^{\pm}(y)\ V(dy)\right| \leqslant \frac{1}{2|z\alpha_z|}\,\|V\|\|j_z^{\pm}\|_{\infty}.$$

The function K_z is well known to be a fundamental solution in $\mathscr{D}'$ of $(\Delta + z^2)K_z = \delta_0$. Now take some test function $s \in \mathscr{D}$. Then

$$((\Delta + z^2)\varphi_z^{\pm}, s) = ((\Delta + z^2)(K_z * (V\varphi_z^{\pm})), s) = (K_z * (V\varphi_z^{\pm}), s'' + z^2 s).$$

The formula $(f * g, h) = (g, h * \tilde{f}\,)$, where $\tilde{f}(x) = f(-x)$, holds as soon as both sides are defined. Since $\tilde{K}_z = K_z$, it follows that

$$((\Delta + z^2)\varphi_z^{\pm}, s) = (V\varphi_z^{\pm}, K_z * (s'' + z^2 s)) = (V\varphi_z^{\pm}, s).$$

This proves the proposition. □

On the real z axis, one can state a uniqueness theorem for the Helmholtz equation.

Proposition 6.3: Let $f \in C_b^0(\mathbb{R})$ be a weak solution of

$$(\Delta + k^2) f = Vf, \quad k \in \mathbb{R}.$$

If $\lim_{|x|\to+\infty} |\varphi_k^+(x) - f(x)| = 0$, then $f = \varphi_k^+$. The analogous result holds for φ_k^- and any linear combination of φ_k^+ and φ_k^-.

Proof: The function $K_k * (Vf)$ is well defined in $C_b^0(\mathbb{R})$ because it is the convolution of a continuous bounded function K_k with a finite measure Vf. Now $(\Delta + k^2)(f - K_k * (Vf)) = 0$, and thus the general form of f is $f(x) = K_k * (Vf)(x) + ae^{ikx} + be^{-ikx}$. Hence, $\varphi_k^+(x) - f(x) = K_k * V(\varphi_k^+ - f) + (1-a)e^{ikx} - be^{-ikx}$. The condition on the asymptotic behavior imposes $a = 1$ and $b = 0$. Thus, $f(x) = K_k * (Vf)(x) + e^{+ikx}$. Since the solution of the above equation is unique [in $C_b^0(\mathbb{R})$], we have in fact, $f = \varphi_k^+$. This concludes the proof. □

We end the study of the generalized eigenfunctions with a useful approximation for the following. For real k, it follows from Eqs. (5.3)/(5.4) that $j_k^{\pm} = 1 + O(1/k)$, $k \to +\infty$, and consequently

$$r_k^{\pm} = \beta_k^{\pm} = -i\sqrt{\frac{\pi}{2}} \frac{\hat{V}(\mp 2ik)}{k} + O(1/k^2), \quad k \to \infty, \tag{6.5}$$

where $\hat{V}$ is the Fourier transform of V:

$$\hat{V}(\xi) = \frac{1}{\sqrt{2\pi}} \int e^{-i\xi x} \, V(dx).$$

The first term on the rhs of (6.5) is known as the *Born approximation* for the reflection coefficient.

VII. THE SPECTRUM

In this section, we will retrieve and strengthen some known results concerning the spectrum (see Refs. 23 and 24) and compute the spectral measure. Let $R(\zeta) = (H - \zeta)^{-1}$ be the resolvent of H. Let us look for a kernel of $R(\zeta)$, that is a function $\mathcal{G}(x, y, \zeta)$ such that $\mathcal{G}(x, \cdot, \zeta) \in L^2(\mathbb{R})$ and for all $f \in L^2(\mathbb{R})$,

$$R(\zeta) f = \int dy \; \mathcal{G}(\cdot, y, \zeta) f(y). \tag{7.1}$$

For continuous potentials V, the Green's function $\mathcal{G}(x, y, \zeta)$ may be easily computed by the standard method,

$$\mathcal{G}(x, y, \zeta) = \frac{J_z^+(\max(x,y)) J_z^-(\min(x,y))}{\mathcal{W}(J_z^+, J_z^-)}, \tag{7.2}$$

where z is the complex square root of ζ in the upper half-plane $z^2 = \zeta$, $\Im z > 0$. In fact, this expression of the kernel remains valid if V is a finite measure, as will be shown.

Theorem 7.1: *Let V be a finite real Borel measure satisfying (4.1) and $H = -\Delta + V$ be defined in the sense of quadratic forms. Then the resolvent $R(\zeta) = (H - \zeta)^{-1}$ is given by (7.1) and (7.2).*

We begin with the following.

Lemma 7.2: For all $y \in \mathbb{R}$, we have the distributional equality

$$(-\Delta + V - \zeta) \mathcal{G}(\cdot, y, \zeta) = \delta_y, \tag{7.3}$$

where δ_y is the Dirac distribution centered at y.

Proof: The function $\mathcal{G}(x, y, \zeta)$ can be rewritten as

$$\mathcal{G}(x,y,\zeta)=\frac{J_z^+(y)J_z^-(x)}{\mathcal{W}(J_z^+,J_z^-)}+\theta_y(x)\tilde{\mathcal{G}}_y(x),\quad\text{with}\quad\tilde{\mathcal{G}}_y(x)=\frac{J_z^+(y)J_z^-(x)-J_z^+(x)J_z^-(y)}{\mathcal{W}(J_z^+,J_z^-)},$$

and, as usual, θ_y is the Heaviside function centered at y. Since $J_z^\pm$ satisfy the weak Helmholtz equation (6.4) we have $(-\Delta+V-\zeta)\mathcal{G}(\cdot,y,\zeta)=(-\Delta+V-\zeta)\theta_y\tilde{\mathcal{G}}_y$. Let us first compute the term $\Delta(\theta_y\tilde{\mathcal{G}}_y)$. Take some positive function $h\in C_0^\infty$ such that $\int dx\ h(x)=1$ and set $h_n(x)=nh(nx)$ and $\theta_{y,n}(x)=\theta_y*h_n(x)$. The sequence h_n is an approximate identity and so $\theta_{y,n}\overset{\mathcal{D}'}{\to}\theta_y$. The same holds for the derivatives, that is, $D^{(k)}\theta_{y,n}\overset{\mathcal{D}'}{\to}D^{(k)}\theta_y$. Now the function $\theta_{y,n}$ is C^∞ and so we may apply the usual differentiation rule (see, e.g., Ref. 27) for the distribution $\theta_{y,n}\tilde{\mathcal{G}}_y$, that is

$$\begin{aligned}\Delta(\theta_{y,n}\tilde{\mathcal{G}}_y)&=D(\theta'_{y,n}\tilde{\mathcal{G}}_y+\theta_{y,n}D\tilde{\mathcal{G}}_y)=D(\theta'_{y,n}\tilde{\mathcal{G}}_y)+\theta'_{y,n}D\tilde{\mathcal{G}}_y+\theta_{y,n}\Delta\tilde{\mathcal{G}}_y\\&=D(\theta'_{y,n}\tilde{\mathcal{G}}_y)+\theta'_{y,n}D\tilde{\mathcal{G}}_y-\theta_{y,n}(V-\zeta)\tilde{\mathcal{G}}_y.\end{aligned}$$

On the other hand we have $\Delta(\theta_{y,n}\tilde{\mathcal{G}}_y)\overset{\mathcal{D}'}{\to}\Delta(\theta_y\tilde{\mathcal{G}}_y)$, $\theta'_{y,n}\tilde{\mathcal{G}}_y\overset{\mathcal{D}'}{\to}\delta_y\tilde{\mathcal{G}}_y(y)=0$ and $\theta_{y,n}(V-\zeta)\tilde{\mathcal{G}}_y\overset{\mathcal{D}'}{\to}\theta_y(V-\zeta)\tilde{\mathcal{G}}_y$. It remains the term $\theta'_{y,n}D\tilde{\mathcal{G}}_y$ which must be manipulated more carefully. Keeping the notations of Lemma 5.4, we have the distributional equality

$$D\tilde{\mathcal{G}}_y(x)=\frac{J_z^+(y)F_z^-(x)-F_z^+(x)J_z^-(y)}{\mathcal{W}(J_z^+,J_z^-)}.$$

The rhs is defined everywhere and can easily be shown to be continuous at y (although the functions $F_z^\pm$ are themselves not necessary continuous). Hence, $D\tilde{\mathcal{G}}_y$ coincides with a function which is continuous at y, and takes the value 1 there. It follows that $\theta'_{y,n}D\tilde{\mathcal{G}}_y\overset{\mathcal{D}'}{\to}\delta_y$. Putting all this together, we obtain (7.3). □

Proof of Theorem 7.1: Now call $R^\sharp(\zeta)f$ the rhs of (7.1). For $\zeta=z^2$, $\Im z>0$, $R^\sharp(\zeta)$ is a bounded operator (in fact Hilbert–Schmidt) on $L^2(\mathbb{R})$. Moreover, $R^\sharp(\zeta)f$ is easily seen to be a continuous bounded function for all $f\in\mathcal{D}$, so we may consider the distribution $(-\Delta+V-\zeta)R^\sharp(\zeta)f$. Now, for all $s\in\mathcal{D}$ we have

$$((-\Delta+V-\zeta)R^\sharp(\zeta)f,s)=\int dy\ f(y)\left[\int dx(-s''(x)-\zeta s(x))\mathcal{G}(x,y,\zeta)+\int V(dx)\ \mathcal{G}(x,y,\zeta)\right]$$

by Fubini's theorem. On the other hand,

$$\int dx(-s''(x)-\zeta s(x))\mathcal{G}(x,y,\zeta)+\int V(dx)\ \mathcal{G}(x,y,\zeta)=((-\Delta+V-\zeta)\mathcal{G}(\cdot,y,\zeta),s)=s(y).$$

Therefore, we have the distributional equality $(-\Delta+V-\zeta)R^\sharp(\zeta)f=f$. It follows that $R^\sharp(\zeta)f\in D(H)$ and $(H-\zeta)R^\sharp(\zeta)f=f$. Thus $R^\sharp(\zeta)f=R(\zeta)f$ holds on the dense set $\mathcal{D}$, hence on $L^2(\mathbb{R})$ and so $R^\sharp(\zeta)=R(\zeta)$. This completes the proof of Theorem 7.1. □

We now turn to the study of the spectrum. Let $\sigma_d(H)$, $\sigma_{sc}(H)$, and $\sigma_{ac}(H)$ stand for the discrete, singular continuous, and absolutely continuous spectrum of H, respectively, and let $\sigma_c(H)=\sigma_{ac}(H)+\sigma_{sc}(H)$ be the continuous spectrum of H. The spectrum has the same characteristics as for regular potentials, as will be shown.

Theorem 7.3: *The positive spectrum is purely absolutely continuous and the negative spectrum is composed of a finite number of negative eigenvalues,*

$$\sigma_{ac}(H)=[0,+\infty[,\quad \sigma_{sc}(H)=\varnothing,\quad \sigma_d(H)=\{-\lambda_1^2,\ldots,-\lambda_N^2\},\tag{7.4}$$

where $i\lambda_1,\ldots,i\lambda_N$ are the zeros of α_z in the upper half-plane. Let (E_λ) be the spectral family associated to H, i.e., formally $H=\int \lambda\, dE_\lambda$. For any states $f,g\in\mathscr{D}$, the associated spectral measure $\mu_{f,g}$, defined by $d\mu_{f,g}(\lambda)=(f,dE_\lambda g)$, is absolutely continuous on the positive axis with density:

$$\frac{d\mu_{f,g}(\lambda)}{d\lambda}=m_{f,g}(\lambda)=\int dx\, dy\, \overline{f(y)}g(x)p(x,y,\lambda), \tag{7.5}$$

where the kernel $p(x,y,\lambda)$ is given by

$$p(x,y,\lambda)=\frac{1}{2\pi}\,\Re\left\{\frac{\varphi_k^+(y)\varphi_k^-(x)}{kt_k}\right\}=\frac{1}{2\pi}\,\Re\left\{\frac{\varphi_k^+(x)\varphi_k^-(y)}{kt_k}\right\},\quad k^2=\lambda,k>0. \tag{7.6}$$

Proof: In view of Proposition 6.1, the kernel of the resolvent $\mathscr{G}(x,y,\zeta)$ is analytic on the negative real axis with only poles produced by the zeros of the Wronskian, which are also the zeros of α_k, that is, bound states (see Proposition 5.7). Hence, there is no continuous spectrum on the negative axis. Let us show that there are no imbedded eigenvalues in the positive spectrum. Suppose $E=k^2$, $k\geqslant 0$, is an eigenvalue. Then there exists some nonzero $f\in D(H)\subset C_b^0(\mathbb{R})\cap L^2(\mathbb{R})$ such that $(\Delta+k^2)f=Vf$. Now, as was explained in Proposition 6.3, any continuous bounded solution of the above equation can be written as $f(x)=K_k*(Vf)+ae^{ikx}+be^{-ikx}$ for some constants a and b. The supplementary condition $f\in L^2(\mathbb{R})$ forces $f(x)\to 0$ as $|x|\to+\infty$. The inspection of the asymptotic behavior of f then imposes $a=b=0$ and $f=0$. Hence E cannot be an eigenvalue. So far, we have proven $\sigma(H)\cap\,]-\infty,\,0[=\sigma_d(H)$, where $\sigma_d(H)$ as in (7.4). Let us use Stone's formula to derive the spectral projectors on the positive spectrum. If $E[\alpha,\,\beta]$ is the spectral projector on the interval $[\alpha,\beta]\subset\mathbb{R}^+$, then

$$E(\alpha,\beta)=\text{s.}\lim_{\varepsilon\to 0}\frac{1}{2i\pi}\int_\alpha^\beta d\zeta\; R(\zeta+i\varepsilon)-R(\zeta-i\varepsilon),\quad 0<\alpha<\beta.$$

The above operator is a projector with kernel

$$e_{\alpha,\beta}(x,y)=\frac{1}{2i\pi}\int_\alpha^\beta d\zeta\;\mathscr{G}(x,y,\zeta+io)-\mathscr{G}(x,y,\zeta-io). \tag{7.7}$$

The integrand can be computed more precisely since for $\zeta=k^2$ $(k>0)$ we have

$$\frac{1}{2i\pi}\{\mathscr{G}(x,y,\zeta+io)-\mathscr{G}(x,y,\zeta-io)\}=p(x,y,\zeta),$$

with

$$p(x,y,\zeta)=\frac{1}{2\pi}\,\Re\left\{\frac{J_k^+(y)J_k^-(x)}{k\alpha_k}\right\}=\frac{1}{2\pi}\,\Re\left\{\frac{\varphi_k^+(y)\varphi_k^-(x)}{kt_k}\right\},\quad x<y,$$

and $p(x,y,\zeta)=p(y,x,\zeta)$, $x>y$. The distinction $x<y$ and $x>y$ is actually unnecessary, as is shown in the following.

Lemma 7.4: For all $x,y,k\in\mathbb{R}$ we have

$$2\Re\left\{\frac{\varphi_k^+(y)\varphi_k^-(x)}{t_k}\right\}=\Re\{\overline{\varphi_k^+(x)}\varphi_k^+(y)\}+\Re\{\overline{\varphi_k^-(x)}\varphi_k^-(y)\}. \tag{7.8}$$

This implies in particular

$$\Re\left\{\frac{\varphi_k^+(y)\varphi_k^-(x)}{t_k}\right\}=\Re\left\{\frac{\varphi_k^+(x)\varphi_k^-(y)}{t_k}\right\}. \tag{7.9}$$

Proof: We may expand $\varphi^\pm(x,k)$ on the basis $\{\varphi^\mp(x,k),\overline{\varphi^\pm(x,k)}\}$:

$$\varphi_k^+(x)=\frac{1}{\overline{r_k^+}}\,\overline{\varphi_k^+(x)}+\frac{t_k}{r_k^-}\,\varphi_k^-(x), \tag{7.10}$$

$$\varphi_k^-(x)=\frac{1}{\overline{r_k^-}}\,\overline{\varphi_k^-(x)}+\frac{t_k}{r_k^+}\,\varphi_k^+(x). \tag{7.11}$$

Indeed, the functions on the left- and right-hand sides both verify the weak Helmholtz equation and satisfy the same asymptotic condition. They must therefore coincide, by Proposition 6.3. Now using the relations (6.3) between the coefficients, (7.8) is easily verified □

After the change of variables $\zeta=k^2$, we obtain

$$E[\alpha,\ \beta]f(x)=\int dy\ e_{\alpha,\beta}(x,y)f(y), \tag{7.12}$$

with

$$e_{\alpha,\beta}(x,y)=e_{\alpha,\beta}(y,x)=\frac{1}{\pi}\int_{\sqrt{\alpha}}^{\sqrt{\beta}}dk\ \Re\left\{\frac{\varphi_k^+(y)\varphi_k^-(x)}{t_k}\right\}. \tag{7.13}$$

This shows in particular that $\sigma_c=[0,+\infty[$. It remains to prove the absence of singular continuous spectrum. To this end, it suffices to show that the spectral measure $\mu_{f,g}$ is absolutely continuous on $(0,+\infty)$ for all $f,g\in\mathscr{D}$. Now, for all $\alpha>0$, a straightforward computation using (7.12) and the continuity properties of $\varphi_z^\pm$ stated in Theorem 6.1 shows that

$$\lim_{h\to 0}\frac{u_{f,g}([\alpha,\alpha+h])}{h}=\int dx\ dy\ \overline{f}(y)g(x)p(x,y,\alpha)=:m_{f,g}(\alpha).$$

Since $m_{f,g}(\alpha)$ is continuous, it follows that $\mu_{f,g}$ is absolutely continuous on $\mathbb{R}^+$ with Radon–Nikodym derivative $m_{f,g}$, and since this holds on a dense set of functions f,g, it implies $\sigma_{ac}=[0,+\infty[$. This completes the proof of Theorem 7.3. □

VIII. THE GENERALIZED FOURIER TRANSFORMS

Let $\mathscr{F}_0$ be the unitary Fourier transform on $L^2(\mathbb{R})$ with adjoint $\mathscr{F}_0^*$

$$\mathscr{F}_0 f(k)=\text{l.i.m.}\ \frac{1}{\sqrt{2\pi}}\int dx\ e^{-ikx}f(x), \tag{8.1}$$

$$\mathscr{F}_0^* g(x)=\text{l.i.m.}\ \frac{1}{\sqrt{2\pi}}\int dk\ e^{+ikx}g(k), \tag{8.2}$$

where l.i.m. $\int$ stands for the limit in the $L^2(\mathbb{R})$ norm of $\int_{-R}^{+R}$ as $R\to\infty$. To simplify the notation, we will sometimes set $\hat{f}(k)=\mathscr{F}_0 f(k)$. Let us now introduce a family of solutions of physical importance for the scattering problem. For $k\in\mathbb{R}$, we define the functions $\psi_k^\pm$ by

$$\psi_k^+=\begin{cases}\varphi_k^+ & \text{if } k\geqslant 0,\\ \overline{\varphi_k^-} & \text{if } k<0,\end{cases}\qquad \psi_k^-=\begin{cases}\varphi_k^- & \text{if } k\geqslant 0,\\ \overline{\varphi_k^+} & \text{if } k<0.\end{cases}$$

They are solutions of the integral equation

$$\psi_k^\pm(x)=e^{\pm ikx}+\int\frac{e^{i|k||x-y|}}{2i|k|}\,V(dy)\ \psi_k^\pm(y). \tag{8.3}$$

These functions write as the superposition of an ingoing free wave and an outgoing perturbed wave. For regular potentials, (8.3) is the well-known Lippmann–Schwinger equation. We define the linear maps $\mathscr{T}_\pm : L^1 \to L^\infty$ and $\mathscr{T}_\pm^* : L^1 \to L^\infty$ by

$$\mathscr{T}_+ f(k) = \frac{1}{\sqrt{2\pi}} \int dx\ \overline{\psi_k^+(x)} f(x) \quad \text{and} \quad \mathscr{T}_+^* g(x) = \frac{1}{\sqrt{2\pi}} \int dk\ \psi_k^+(x) g(k),$$

$$\mathscr{T}_- f(k) = \frac{1}{\sqrt{2\pi}} \int dx\ \psi_k^-(x) f(x) \quad \text{and} \quad \mathscr{T}_-^* g(x) = \frac{1}{\sqrt{2\pi}} \int dk\ \overline{\psi_k^-(x)} g(k).$$

Then we have the following proposition.

Proposition 8.1: $\mathscr{T}_\pm$ and $\mathscr{T}_\pm^*$ can be extended to continuous maps from $L^2(\mathbb{R})$ to $L^2(\mathbb{R})$. $\mathscr{T}_\pm^*$ is the adjoint of $\mathscr{T}_\pm$.

Proof: The proof is a straightforward consequence of the asymptotics of the functions $\psi_k^\pm$, which behave like planes waves $e^{\pm ikx}$ for large arguments. We omit the details. □

The operators $\mathscr{T}_+$ and $\mathscr{T}_-$ are usually called *generalized Fourier transform*. They satisfy the so-called *intertwining property*, which we state in the following form:

Proposition 8.2 (intertwining property): For all bounded measurable function η we have

$$\eta(H).\mathscr{T}_\pm^*.\mathscr{T}_0 = \mathscr{T}_\pm^*.\mathscr{T}_0\, \eta(H_0), \tag{8.4}$$

where $\eta(H)$ and $\eta(H_0)$ are defined according to the spectral theorem.

Proof: We will first show that $H.\mathscr{T}_+^*.\mathscr{T}_0 f = \mathscr{T}_+^*.\mathscr{T}_0 H_0 f$ holds on $D(H_0)$. For any x,y we have

$$\begin{aligned} |\psi_k^+(x) - \psi_k^+(y)| &\leqslant \left| \int \frac{e^{i|k||x-t|} - e^{i|k||y-t|}}{2i|k|}\ \psi_k^+(t)\ V(dt) \right| + |e^{ikx} - e^{iky}| \\ &\leqslant \|V\| \sup_{k\in\mathbb{R}} \|\psi_k^+\|_\infty |x-y| + |k||x-y|\,. \end{aligned}$$

It follows that $\mathscr{T}_+^*.\mathscr{T}_0 f$ is Lipschitz and *a fortiori* continuous for all $f \in D(H_0)$. So, we may consider the distribution $(-\Delta + V).\mathscr{T}_+^*.\mathscr{T}_0 f$ in $\mathscr{D}'$. Now for any $s \in \mathscr{D}$ we have

$$\begin{aligned} ((-\Delta+V).\mathscr{T}_+^*.\mathscr{T}_0 f, s) &= -(\mathscr{T}_+^*.\mathscr{T}_0 f, s'') + (V, \mathscr{T}_+^*.\mathscr{T}_0 f s) \\ &= -\int dx\ s''(x) \int dk\ \psi_k^+(x) \hat{f}(k) + \int V(dx)\ s(x) \int dk\ \psi_k^+(x) \hat{f}(k) \\ &= \int dk\ \hat{f}(k) \left[-\int dx s''(x) \psi_k^+(x) + \int V(dx)\ s(x) \psi_k^+(x) \right], \end{aligned}$$

where we have used Fubini's theorem to interchange the integrals. Now,

$$-\int dx\ s''(x) \psi_k^+(x) + \int V(dx)\ s(x) \psi_k^+(x) = k^2 \int dx\, s(x) \psi_k^+(x),$$

owing to the weak Helmholtz equation $(-\Delta + V)\psi_k^+ = k^2 \psi_k^+$. Hence, using again Fubini's theorem,

$$((-\Delta+V).\mathscr{T}_+^*.\mathscr{T}_0 f, s) = \int dx\ s(x) \int dk\ k^2 \hat{f}(k) \psi_k^+(x) = (\mathscr{T}_+^*.\mathscr{T}_0 H_0 f, s).$$

Now $\mathscr{T}_+^*.\mathscr{T}_0 H_0 f$ is in $L^2(\mathbb{R})$. So, the distribution $(-\Delta + V).\mathscr{T}_+^* f$ is in $L^2(\mathbb{R})$, which means that $\mathscr{T}_+^* f$ is in $D(H)$ and $H.\mathscr{T}_+^*.\mathscr{T}_0 f = \mathscr{T}_+^*.\mathscr{T}_0 H_0 f$. Also we have

$$(H-\zeta).\mathscr{T}_+^*.\mathscr{T}_0 f = \mathscr{T}_+^*.\mathscr{T}_0 (H_0 - \zeta) f$$

for any $\zeta \in \mathbb{C}\backslash\mathbb{R}$ and for all $f \in D(H_0)$. If we introduce the free resolvent $R_0(\zeta)=(H_0-\zeta)^{-1}$, it follows that $R(\zeta).\mathscr{F}_+^*.\mathscr{F}_0 g=.\mathscr{F}_+^*.\mathscr{F}_0 R_0(\zeta)g$ for all $g \in \mathrm{Ran}\,(H_0-\zeta)$ (Ran=range). H_0 being self-adjoint, we have $\mathrm{Ran}\,(H_0-\zeta)=L^2(\mathbb{R})$ and thus $R(\zeta).\mathscr{F}_+^*.\mathscr{F}_0=.\mathscr{F}_+^*.\mathscr{F}_0 R_0(\zeta)$ for all $\zeta \notin \mathbb{R}$. This implies $E(\Lambda).\mathscr{F}_+^*.\mathscr{F}_0=.\mathscr{F}_+^*.\mathscr{F}_0 E_0(\Lambda)$ for all Borel sets Λ, where $E_0(\cdot)$ is the spectral family associated to H_0 (the last identity can be obtained by means of Stone's formula). The relation (8.4) then follows from the spectral theorem. □

Corrolary 8.3: Let $\mathscr{H}_{ac}$ be the absolutely continuous subspace of $\mathscr{H}$,

$$\mathscr{H}_{ac}=\{f \in \mathscr{H} \mid \textit{the spectral measure associated to } f \textit{ is absolutely continuous}\},$$

and P_{ac} is the orthogonal projector on $\mathscr{H}_{ac}$. Then

$$\mathscr{F}_\pm P_{ac}=\mathscr{F}_\pm . \tag{8.5}$$

Proof: Let $\mathscr{H}_{pp}$ be the purely punctual subspace, that is the set of bound states. Then $\mathscr{H}_{pp} \oplus .\mathscr{H}_{ac}=\mathscr{H}$, so it suffices to prove that $\mathscr{F}_\pm$ vanishes on $\mathscr{H}_{pp}$. If $-\lambda^2 \in \sigma_d(H)$ is an eigenvalue and f an associated eigenfunction, then $Hf=-\lambda^2 f$. Now by taking the adjoint in Eq. (8.4), with $\eta(x)=(x-i)^{-1}$, we have $\mathscr{F}_\pm(H+i)^{-1}f=.\mathscr{F}_0(H_0+i)^{-1}.\mathscr{F}_0^*.\mathscr{F}f$, and thus $(-\lambda^2+i).\mathscr{F}_\pm f(k)=(k^2+i)^{-1}.\mathscr{F}_\pm f(k)$ for all k, which forces $\mathscr{F}_\pm f(k)=0$ for all k. □

IX. THE WAVE OPERATORS

We are now able to derive explicit formulas for the wave operators.

Theorem 9.1: *The (strong) wave operators $W_\pm$ exist and are complete. They are given by*

$$W_\pm = \mathrm{s.}\lim_{t\to\pm\infty} e^{itH}e^{-itH_0}=.\mathscr{F}_\mp^*.\mathscr{F}_0 . \tag{9.1}$$

Proof: The existence and completeness of the wave operators have already been proved in a more general setting.[25] Once the intertwining property and minimal regularity properties for $\varphi_k^\pm(x)$ [precisely, for fixed x, $\varphi_k^\pm(x)$ should be continuously differentiable on $\mathbb{R}\backslash 0$ with respect to k] have been established (Theorem 6.1 and Proposition 8.2), the proof of formula (9.1) is the same as for regular potentials (see, e.g., Ref. 28), so we omit it. The main ingredient is a stationary phase argument. □

The following is a straightforward consequence of the last theorem.

Corrolary 9.2: The generalized functions $\psi_k^\pm$ form a complete orthonormal set of $\mathscr{H}_{ac}$,

$$\mathscr{F}_\pm.\mathscr{F}_\pm^*=1, \quad \mathscr{F}_\pm^*.\mathscr{F}_\pm=\mathscr{P}_{ac} . \tag{9.2}$$

Proof: The wave operators are isometric, as a strong limit of unitary operators. So $W_\pm^* W_\pm = .\mathscr{F}_\pm.\mathscr{F}_\pm^*=1$, where $*$ stands for the Hermitian conjugation. This expresses the ''orthogonality'' of the generalized functions $\psi_k^\pm$, for it corresponds formally to $\langle \psi_k^\pm, \psi_{k'}^\pm \rangle = \delta(k-k')$. Furthermore, $W_\pm W_\pm^* = (W_\pm W_\pm^*)(W_\pm W_\pm^*)$ is a projector onto $\mathscr{H}_{ac}$, hence $W_\pm W_\pm^* = .\mathscr{F}_\pm^*.\mathscr{F}_\pm = P_{ac}$, which proves the completeness of the functions $\psi_k^\pm$. □

The wave operators provide a simple approximation of the evolution operator $e^{\pm itH}$ at large time:

$$\mathrm{s.}\lim_{t\to+\infty}(e^{\pm itH}-W_\pm e^{\pm itH_0})\mathscr{P}_{ac}=0. \tag{9.3}$$

In particular, for all initial states $f \in .\mathscr{H}_{ac}$ we have

$$\mathrm{s.}\lim_{t\to+\infty} \Psi_t-\tilde{\Psi}_t=0, \tag{9.4}$$

where $\Psi_t=e^{-itH}f$ and $\tilde{\Psi}_t=W_- e^{-itH_0}f$ is explicit,

$$\tilde{\Psi}_t(x)=\frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^0 dk\; e^{-itk^2}\overline{\varphi_k^-}(x)\hat{f}(k)+\int_0^{+\infty} dk\; e^{-itk^2}\varphi_k^+(x)\hat{f}(k)\right). \tag{9.5}$$

X. THE TIME EVOLUTION

We now wish to look at the time-space behavior of the wave packets Ψ_t and Φ_t defined in the Introduction. For sake of simplicity, we will now suppose V is a finite measure of compact support. In fact, the more general condition (4.1) also works, but brings in technical complications and carries no new essential idea. We call Ω^- and Ω^+ the regions to the left and to the right of the support of V and $\mathbb{I}_{\Omega^\pm}$ the corresponding characteristic functions, $\mathbb{I}_{\Omega^\pm}(x)=1$ if $x\in\Omega^\pm$, $\mathbb{I}_{\Omega^\pm}(x)=0$ else. We use the same notation for the associated multiplication operator, i.e., $(\mathbb{I}_{\Omega^\pm} f\,)(x)=\mathbb{I}_{\Omega^\pm}(x)f(x)$. Let us now define

$$p^\pm[f]=\lim_{t\to+\infty}\|\mathbb{I}_{\Omega^\pm}\Psi_t\|_2^2,\quad \Psi_t=e^{-iHt}f, \tag{10.1}$$

assuming for the moment that the limits exist. When f is normalized (i.e., $\|f\|=1$), these quantities can be interpreted as the final probability of finding a particle described by the wave function Ψ_t in the left, respectively right, region. Analogously, we define

$$p_q^\pm[f]=\lim_{t\to+\infty}\|\mathbb{I}_{\Omega^\pm}\Phi_t\|_q^q \tag{10.2}$$

with Φ_t the solution of the wave equation with initial data $(f,-f')$. It will be clear from the context which definition we use.

The probability of presence $p^\pm[f]$ can be related to the transmission and reflection coefficients as follows.

Theorem 10.1: (Schrödinger equation) *For any initial state $f\in\mathscr{H}_{ac}$ we have*

$$p^-[f]=\int_{-\infty}^0 dk\ |\overline{r_k^+}\hat f(-k)+\overline{t_k}\hat f(k)|^2,\quad p^+[f]=\int_0^{+\infty}dk\ |r_k^-\hat f(-k)+t_k\hat f(k)|^2. \tag{10.3}$$

This result is perfectly intelligible. Indeed, it merely states that the energy that can be found on the left of the potential are the positive frequencies traveling to the right that have been reflected by V and the negative frequencies traveling to the left that have been transmitted through the potential.

Proof: For the sake of simplicity we will only consider the quantity p^-. The case of p^+ is analogous. As before, define $\tilde\Psi_t=W_-e^{-itH_0}f$. In view of (9.4), we have $p^\pm[f]=\lim_{t\to+\infty}\|\mathbb{I}_{\Omega^\pm}\tilde\Psi_t\|^2$. Now, since outside of the potential support the functions $\varphi_k^\pm$ are given by their asymptotic expressions (6.1), the function $\tilde\Psi_t$ takes a simple form there and we have

$$\mathbb{I}_{\Omega^-}\tilde\Psi_t(x)=(\sqrt{2\pi})^{-1}\mathbb{I}_{\Omega^-}(x)\Bigg(\int_{-\infty}^0 dk\ e^{-itk^2}e^{ikx}\overline{t_k}\hat f(k)+\int_0^{+\infty}dk[e^{-itk^2}e^{ikx}\hat f(k)$$
$$+e^{-itk^2}e^{-ikx}r_k^+\hat f(k)]\Bigg), \tag{10.4}$$

At this point, let us introduce the following operators defined through Fourier multipliers:

$$T:\hat s(k)\mapsto\overline{t_k}\hat s(k),\quad R_\pm:\hat s(k)\mapsto r_k^\pm\hat s(k).$$

In addition we need the parity operator $\Pi:s(x)\mapsto s(-x)$ and the positive/negative frequencies projectors $P^\pm$:

$$\widehat{P^\pm f}(k)=\begin{cases}\hat f(k) & \text{if sign }(k)=\pm1,\\ 0 & \text{otherwise.}\end{cases}$$

Then the previous equations may be rewritten more intuitively as follows:

$$\mathbb{I}_{\Omega^-}\tilde\Psi_t=\mathbb{I}_{\Omega^-}e^{-itH_0}[P^+ +\Pi R_+P^+ +TP^-]f. \tag{10.5}$$

Note that the positive frequencies of f are reflected ($\Pi R_+ P^+ f$) while the negative frequencies are transmitted through the potential ($TP^- f$). Eventually only the waves with negatives frequencies—$(\Pi R_+ P^+ + TP^-)f$—will contribute to the wave function in Ω^-. Indeed, we have the following.

Lemma 10.2: For all $f \in L^2(\mathbb{R})$ we have

$$\lim_{t\to+\infty} \|\mathbb{I}_{\Omega^-} e^{-itH_0} f\|^2 = \int_{-\infty}^{0} dk |\hat{f}(k)|^2, \quad \lim_{t\to+\infty} \|\mathbb{I}_{\Omega^+} e^{-itH_0} f\|^2 = \int_{0}^{+\infty} dk |\hat{f}(k)|^2.$$

Proof: The proof follows from the well-known (see, e.g., Ref. 29, p. 60) asymptotics s. $\lim_{t\to+\infty} e^{-itH_0} - V_t = 0$, where $V_t f(x) = (\sqrt{2it})^{-1} e^{ix^2/4t} \hat{f}(x/2t)$. □

A direct application of the last lemma yields the formula (10.3). Thanks to the relations (5.3), it can be easily verified that $p^+[f] + p^-[f] = 1$; that is, no energy remains trapped in the region of interaction. □

Remark: Theorem 10.1 can be seen as the one-dimensional version of the famous Dollard's ''scattering into cones theorem'' (see Ref. 30 for a complete discussion) which asserts that the probability $P_C[f]$ for the wave function $e^{-iHt} f$ to emerge in a cone C in $\mathbb{R}^3$ with apex at the origin is given by

$$P_C[f] = \lim_{t\to+\infty} \|\mathbb{I}_C e^{-iHt} f\|^2 = \int_C dk |\widehat{\mathscr{S} f}(k)|^2,$$

where $\mathscr{S} = W_+^\star W_-$ is the scattering operator. In the one-dimensional case, there are only two possible directions (left and right), and the ''cones'' reduce to the left and right half-axes.

We now turn to the study of the wave equation. We will assume that the initial wave packet Φ_0 is prepared by sending an incident wave train from the remote left on the scatterer. Thus, for $t<0$, the wave function is simply a (free) solution of the wave equation propagating to the right, that is, of the form $\Phi_t(x) = s(x-t)$. Therefore, it is natural to impose the initial conditions

$$g = \partial_t \Phi_0 = -\partial_x \Phi_0 = -f'. \tag{10.6}$$

The temporal properties of the wave equation can now be summarized in the following theorem. We make the statements for initial states located in Ω^- and propagating to the right; they can be adapted in an obvious way to states with support in Ω^+ and propagating to the left.

Theorem 10.3: (wave equation) *Suppose V is compactly supported and such that $-\Delta + V$ is a positive operator. Let $f \in W^1(\mathbb{R})$ be a function with compact support in Ω^- and consider the solution of the wave equation with initial conditions (10.6), that is,*

$$\Phi_t = \cos(H^{1/2}t) f - H^{-1/2} \sin(H^{1/2}t) f'.$$

Then

$$\mathbb{I}_{\Omega^-} \Phi_t(x) = \mathbb{I}_{\Omega^-}(x)[f(x-t) + (\Pi R^+ f)(x+t)], \tag{10.7}$$

$$\mathbb{I}_{\Omega^+} \Phi_t(x) = \mathbb{I}_{\Omega^+}(x)(Tf)(x-t). \tag{10.8}$$

We will use the spectral density $m_{f,g}(\lambda)$ introduced in Sec. VI to compute the exact time-dependent solution of the wave equation. We begin with the following.

Lemma 10.4: For all bounded measurable function η and all $f \in \mathscr{H}_{ac}$ with support in Ω^- we have

$$\eta(H) f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dk \ \eta(k^2) \varphi_k^+(x) \hat{f}(k), \tag{10.9}$$

the equality holding in $L^2(\mathbb{R})$. If $f \in D(H)$, then both left- and right-hand sides are continuous functions and thus the equality holds pointwise.

Proof: For any $u \in \mathscr{D}$, we have

$$\langle u,\eta(H)f\rangle=\int_0^\infty d\lambda\ \eta(\lambda)m_{u,f}(\lambda)=\frac{1}{2\pi}\int_0^\infty d\lambda\ \eta(\lambda)\int dx\ dy\ \overline{u}(y)f(x)\Re\left\{\frac{\varphi^+_{\sqrt{\lambda}}(y)\varphi^-_{\sqrt{\lambda}}(x)}{\sqrt{\lambda}t_{\sqrt{\lambda}}}\right\}.$$

Interchanging the order of integration, which is justified since the inner integrals are absolutely convergent, we obtain

$$\langle u,\eta(H)f\rangle=\langle u,Af\rangle,$$

with

$$Af(y)=\frac{1}{2\pi}\int dx\ f(x)\int_0^\infty d\lambda\ \eta(\lambda)\Re\left\{\frac{\varphi^+_{\sqrt{\lambda}}(y)\varphi^-_{\sqrt{\lambda}}(x)}{\sqrt{\lambda}t_{\sqrt{\lambda}}}\right\}.$$

Making the change of variables $\lambda=k^2$, using the symmetry properties of $\varphi_k^\pm$, and interchanging again the integrals yields finally

$$Af(y)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}dk\ \eta(k^2)\varphi_k^+(y)\int dx\ f(x)t_k^{-1}\varphi_k^-(x).$$

Since f is supported on Ω^-, we have $f(x)t_k^{-1}\varphi_k^-(x)=f(x)e^{-ikx}$ and

$$Af(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}dk\ \eta(k^2)\varphi_k^+(y)\hat{f}(k). \tag{10.10}$$

Now A can be extended to a bounded operator in $L^2(\mathbb{R})$. Hence, by density, $\eta(H)f=Af$ holds on $L^2(\mathbb{R})$. Note that Af is a continuous function [since φ_k^+ is continuous and the integral in (10.10) is uniformly convergent in y]. If $f\in D(H)$, then $\eta(H)f\in D(H)\subset W^1(\mathbb{R})$ is also continuous and thus the equality $\eta(H)f=Af$ holds pointwise. This concludes the proof. □

Proof of Theorem 10.3: It follows from the last lemma that

$$\Phi_t(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}dk\ e^{-ikt}\varphi_k^+(x)\hat{f}(k).$$

Using the asymptotic expression (6.1) of φ_k^+ yields (10.7) and (10.8) and the theorem is proved. □

In order to gain a deeper understanding, we now present an alternative proof. Consider the following function,

$$\Phi_t(x)=\frac{1}{\sqrt{2\pi}}\int dk\ \hat{f}(k)\varphi_k^+(x)e^{-ikt}, \tag{10.11}$$

where $\varphi_k^+(x)$ are the generalized eigenfunctions. Clearly this presents a solution of the wave equation in the sense of distributions. Moreover, upon using the asymptotic expressions valid in $\Omega^\pm$, we recover the equations of the theorem. We only have to check that Φ_t corresponds to the initial data requested. For $t=0$ we have

$$\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int dk\ \hat{f}(k)e^{ikx}\varphi_k^+(x)e^{-ikx}. \tag{10.12}$$

Now for each x, $\varphi_k^+(x)e^{-ikx}$ is analytic and bounded in the upper complex k half-plane (see Theorem 6.1). Therefore, by the Paley–Wiener theorem, it follows that Φ_0 is supported by inside Ω^-. Now for $x\in\Omega^-$ we may use the asymptotic expressions for φ_k^+ and we have

$$\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int dk\ \hat{f}(k)(e^{ikx}+r_k^+e^{-ikx}). \tag{10.13}$$

Now again thanks to Paley–Wiener, the second part is a function supported outside Ω^-. Therefore by inverse Fourier transform, we have $\Phi_0 = f$. On the other hand, upon derivating under the integral sign, we have $\partial_t \Phi_0 = -f'$ and the theorem is proved. □

Remark: Lemma 10.4 can also be used to compute the exact time evolution in the Schrödinger case for initial states supported in $\Omega^\pm$ and thus provides an alternative proof of Theorem 10.1 without using the wave operators. However, the advantage of the first method is that the initial state need not be with support inside $\Omega^\pm$.

XI. RECOVERING THE FRACTAL DIMENSIONS

We shall now propose a method to retrieve wavelet dimensions of the potential from the asymptotic probabilities of presence $p^\pm[f]$ and $p_q[f]$ by choosing an appropriate family of initial states. As usual, we begin with the quantum evolution. Suppose we can prepare some initial state f in $\mathscr{S}_+(\mathbb{R})$. Then we have the following.

Theorem 11.1: (Schrödinger equation) *Let $f \in \mathscr{S}_+(\mathbb{R})$. For all $b \in \mathbb{R}$, we have*

$$p^-[f_{b,a}] = a^2 \| \mathscr{W}_g V(\cdot, a/2) \|^2 + O(a^2), \quad a \to 0, \tag{11.1}$$

where the wavelet $g \in \mathscr{S}_+(\mathbb{R})$ is essentially a primitive of f: $\hat{g}(k) = \sqrt{\pi/8}\hat{f}(k)/k$.

Corrolary 11.2: For all $b \notin \mathrm{supp}(V)$, we have

$$\lim_{a \to 0} \begin{matrix} \sup \\ \inf \end{matrix} \frac{\log \rho(b,a)}{\log a} = \min\{\kappa_2^\pm[V], 0\}, \quad \textit{where} \quad \rho(b,a) = \int_a^1 \frac{d\alpha}{\alpha^3} p^-[f_{b,\alpha}].$$

Proof of Theorem 11.1: First suppose $f \in \mathscr{H}_{ac}$. Then a straightforward computation using the Born approximation (6.5) and Eq. (10.3) yields

$$p^-[f_{b,a}] = a^2 \int_0^{+\infty} dk |\hat{V}(k)|^2 |\hat{g}(ak)|^2 + O(a^2) = a^2 \| \mathscr{W}_g V(\cdot, a) \|^2 + O(a^2),$$

which is the desired result. If $f \notin \mathscr{H}_{ac}$, then we may still apply formula (10.3) to its projection $P_{ac} f$ on $\mathscr{H}_{ac}$, that is,

$$p^-[P_{ac} f_{b,a}] = \int_{-\infty}^0 dk \, |\overline{r_k^+}(P_{ac} f_{b,a})\hat{}(-k) + \overline{t_k}(P_{ac} f_{b,a})\hat{}(k)|^2. \tag{11.2}$$

Now, take an orthonormal basis $(\eta_1, \ldots, \eta_n)$ of the space of bound states and call P_{bs} the orthogonal projector on this basis. The states $f_{b,a}$ can be decomposed as

$$f_{b,a} = P_{ac} f_{b,a} + P_{bs} f_{b,a} = P_{ac} f_{b,a} + \sum_{j=1}^n \langle f_{b,a}, \eta_j \rangle \eta_j .$$

The scalar products $\langle f_{b,a}, \eta_j \rangle$ are nothing but wavelet transforms of the eigenvectors η_j,

$$\langle f_{b,a}, \eta_j \rangle = \mathscr{W}_f \eta_j(b,a).$$

Now, outside of the support of V, the η_j are smooth and rapidly decreasing because they are merely exponential functions (as solutions of $\Delta \eta_j = \lambda_j^2 \eta_j$). Hence the wavelet coefficients in this region are rapidly decreasing at small scales (see, e.g., Ref. 31), that is, for fixed b,

$$|\langle f_{b,a}, \eta_j \rangle| \leqslant O(a^\infty). \tag{11.3}$$

Using (11.2) and (11.3), we obtain after some elementary estimates,

$$p^-[f_{b,a}] = p^-[P_{ac} f_{b,a}] + O(a^\infty) = \int_0^{+\infty} dk \, |r_k^+|^2 |\widehat{f_{b,a}}(k)|^2 + O(a^\infty),$$

and (11.1) follows again from the Born approximation. □

Thus, the behavior of $p^-[f]$ as one varies the characteristic scale of the initial state f is governed by the wavelet correlation dimension of the potential, provided the latter is sufficiently singular (at least nonsquare integrable). Analogous results can be stated for $p^+[f]$ in place of $p^-[f]$. It suffices to take a wavelet in $\mathscr{S}_-(\mathbb{R})$ (no positive frequencies) instead of $\mathscr{S}_+(\mathbb{R})$.

Remark: The assumption that the initial states belong to $\mathscr{S}_\pm(\mathbb{R})$ is somewhat unrealistic. Indeed, such states cannot have compact support and thus have to be prepared and controlled not only far from the target, but also in the region of interaction, which is practically impossible. However, as was explained previously, admitting an arbitrarily small error on the dimensions $\kappa_2^\pm$ we may compute the latter with wavelets having a finite number of moments vanishing. So, in fact, we may assume that the initial state has compact support outside the potential range.

For the wave equation we can say a little more. However, the initial data has to be chosen with some more care. Let us describe briefly the ''experiment'' we want to set. We want to prepare an initial state located at position b_{in}, then send the wave packet toward the target and observe the reflected part at position b_{out} in order to analyze the scatterer at position b. We proceed as follows. The time needed by the wave packet to travel from b_{in} to b and then back to b_{out} is $t=2(b-b_{\text{in}})+(b_{\text{in}}-b_{\text{out}})=2b-b_{\text{in}}-b_{\text{out}}$. By looking at the wave function at time t and position b_{out} we expect to recover the local structure of the potential at position b. Now fix some $f\in S_0(\mathbb{R})$, and some cutoff function χ in $C_0^\infty(\Omega^-)$ such that $\chi=1$ in a neighborhood of b_{in}. We make a smooth truncation of the initial state $f_{b,a}$ by considering the family of pseudo-wavelets $f_{(b_{\text{in}},a)}=\chi f_{b_{\text{in}},a}$, which have now compact support inside Ω^-. With this kind of initial data, the intuitive result is actually verified as shows the following.

Theorem 11.3: *Let $\Phi_{t,b,a}$ be the solution of the wave equation with initial data $(f_{(b,a)},f'_{(b,a)})$. Choose b_{in}, b, and b_{out} and set $t=2b-b_{\text{in}}-b_{\text{out}}$. We then have for b_{out} large enough*

$$\Phi_{t,b_{\text{in}},a}(b_{\text{out}})=a\mathscr{W}_h V(b,a/2)+O(a), \tag{11.4}$$

where h is essentially a primitive of f: $\hat{h}(k)=i\sqrt{\pi/8}\,\overline{\hat{f}(-k)}/k$.

Proof: A direct application of Theorem 10.3 allows us to conclude. We obtain, for all for t large enough [see Eq. (10.8)],

$$\begin{aligned}
\Phi_{t,b,a}(b_{\text{out}}) &= \int dk\ \widehat{f_{(b_{\text{in}},a)}}(k)\,r_k^+\,e^{-ik(2b-b_{\text{in}})}\\
&= f_{(b_{\text{in}},a)}*(\mathscr{T}_0^{-1}r_k^+)(2b-b_{\text{in}})\\
&= f_{b_{\text{in}},a}*(\mathscr{T}_0^{-1}r_k^+)(2b-b_{\text{in}})+O(a^\infty)\\
&= -i\sqrt{\pi}/2\int dk\ \hat{f}(ak)\,\frac{\hat{V}(-2k)}{k}\,e^{-2ikb}+O(a)\\
&= a\mathscr{W}_h V(b,a/2)+O(a).
\end{aligned}$$

Here we have used the Born approximation (6.5) and the fact that the wavelet transform with respect to f may be approximated by the $f_{(b,a)}$ in such that

$$|\mathscr{W}_f s(b,a)-(f_{(b,a)},s)|\leqslant O(a^\infty),\quad a\to 0.$$

□

Upon varying the observation point, we may compute the q-energy:

$$
\begin{aligned}
p_q^-[f_{(b_{\mathrm{in}},a)}] &= \lim_{t\to+\infty} \int_{\Omega^-} db_{\mathrm{out}} |\Phi_{t,b_{\mathrm{in}},a}(b_{\mathrm{out}})|^q \\
&= \lim_{b\to+\infty} \int_{\Omega^-} db_{\mathrm{out}} |\Phi_{2b-b_{\mathrm{in}},b_{\mathrm{in}},a}(b_{\mathrm{out}})|^q \\
&= \lim_{b\to+\infty} \int_{\Omega^-} db_{\mathrm{out}} |a\dot{\mathscr{W}}_h V(b+b_{\mathrm{out}}/2,a/2)+a\varepsilon(b+b_{\mathrm{out}}/2,a)|^q \\
&= 2\int db_{\mathrm{out}} |a\dot{\mathscr{W}}_h V(b_{\mathrm{out}},a/2)+a\varepsilon(b_{\mathrm{out}},a)|^q,
\end{aligned}
$$

where $|\varepsilon(b,a)|\leqslant O(1)$ is the Born error term. Thus if the ''wavelet term'' is preponderant with respect to the ''error term,'' one may compute the wavelet dimensions, as it is shown next.

Proposition 11.4: Suppose V is sufficiently singular in the sense that

$$\lim_{a\to 0} \|\dot{\mathscr{W}}_h V(\cdot,a)\|_q^q = \infty$$

for some (and hence all) $h\in\mathscr{S}_0(\mathbb{R})$. Then for the integrated q-energy of the wave packets

$$\rho_q(a) = \int_a^1 \frac{d\alpha}{\alpha^{q+1}}\, p_q^-[f_{(b_{\mathrm{in}},\alpha)}]$$

we have

$$\lim_{a\to 0} \sup_{\inf} \frac{\log \rho_q(a)}{\log a} = \kappa_q^{\pm}[V]. \tag{11.5}$$

Proof: In view of the inequalities

$$|\|f\|_q - \|g\|_q|^q \leqslant \|f+g\|_q^q \leqslant C(\|f\|_q+\|g\|_q)^q$$

we have

$$a^q|\|\dot{\mathscr{W}}_h V(\cdot,a/2)\|_q - \|\varepsilon(\cdot,a)\|_q|^q \leqslant p_q^-[f_{(b_{\mathrm{in}},a)}] \leqslant Ca^q(\|\dot{\mathscr{W}}_h V(\cdot,a/2)\|_q^q + \|\varepsilon(\cdot,a)\|_q^q).$$

Now the norm $\|\varepsilon(\cdot,a)\|_q$ remains bounded as $a\to 0$. This implies that $p_q^-[f_{(b_{\mathrm{in},a})}]$ and $a^q\|\dot{\mathscr{W}}_h V(\cdot,a)\|_q^q$ are equivalent at small scales and the conclusion follows. □

The scattering process performs a wavelet analysis of the potential. Therefore, in particular for singular potentials, its singularities may be analyzed with the help of remote wavelet techniques.

XII. NUMERICAL RESULTS

We wish to end this paper with a numerical illustration of the ''wavelet transform'' $\Phi_{t,b,a}$. Once the reflection coefficient r_k^+ is known for all k, Lemma 10.4 yields an exact expression of the evolution operators (hence the wave packets Φ_t) in the left region Ω^-, provided the initial state has its support in Ω^-. So we need only sample r_k^+ at different frequencies and estimate the integral (10.9) by a discrete sum. This shall be done on self-similar potentials.

A self-similar measure on $\mathbb{R}$ was defined by Hutchinson[32] to be a probability measure V satisfying

$$V = \sum_{i=1}^N p_i V\circ S_i^{-1},$$

where the p_i are a set of weights ($0<p_i<1$, $\Sigma_{i=1}^N p_i = 1$) and the S_i a set of similarities $S_i(x) = l_i x + b_i$, with $0<l_i<1$. The support of V is the unique nonempty compact K globally invariant

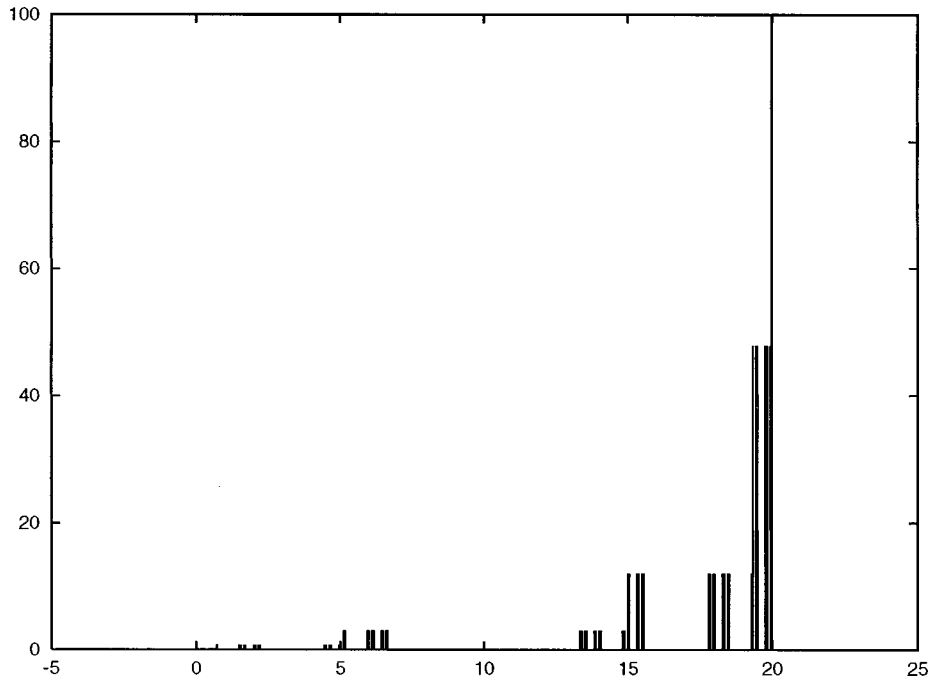

FIG. 1. Approximation of order 6 of the triadic Cantor measure on [0, 20] with weights $p_1=0.2$ and $p_2=0.8$.

under these contractions, i.e., $K=\cup_{i=1}^{N} S_i(K)$. This support may be of zero Lebesgue measure. This occurs when the so-called *strong open set condition*[33] is satisfied; that is, there exist a closed bounded set U such that $S_i(U)\cap S_j(U)=\varnothing$ for $i\neq j$. It can be shown that the set U at least contains K. When there are only two similitudes satisfying the strong open set condition with same contraction ratio $l_1=l_2=1/3$ and same weights $p_1=p_2=1/2$, V is the well-known triadic

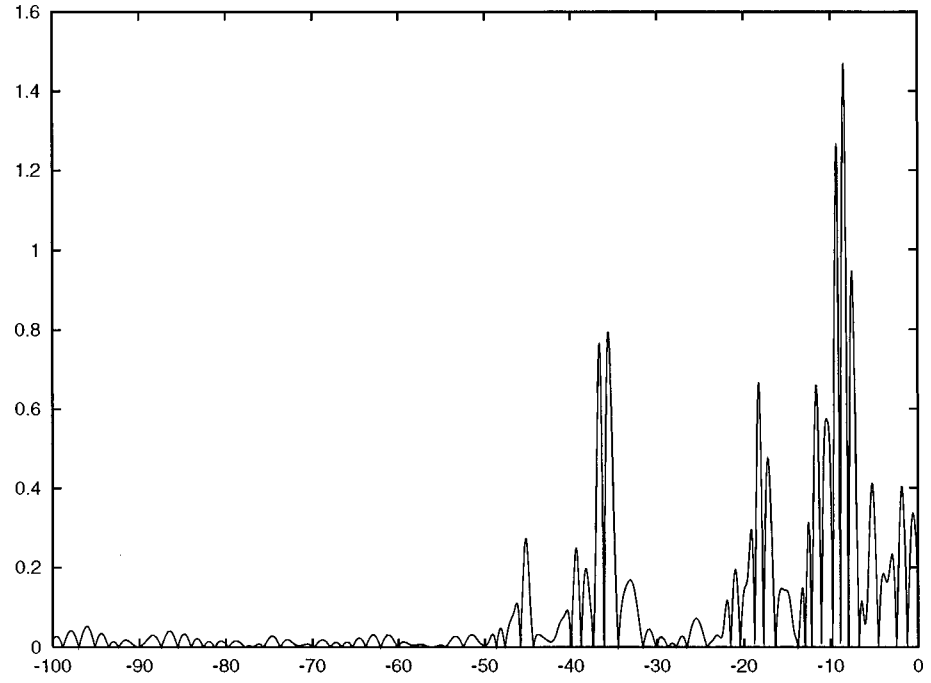

FIG. 2. Modulus of the wave function $\Phi_{t,b,a}(x)$ against x for $b=-30$, $a=0.1$, and $t=80$ and as initial state f a second derivative of the Gaussian. The potential lies to the right (between 0 and 20). Here only the reflected part of the wave is visible.

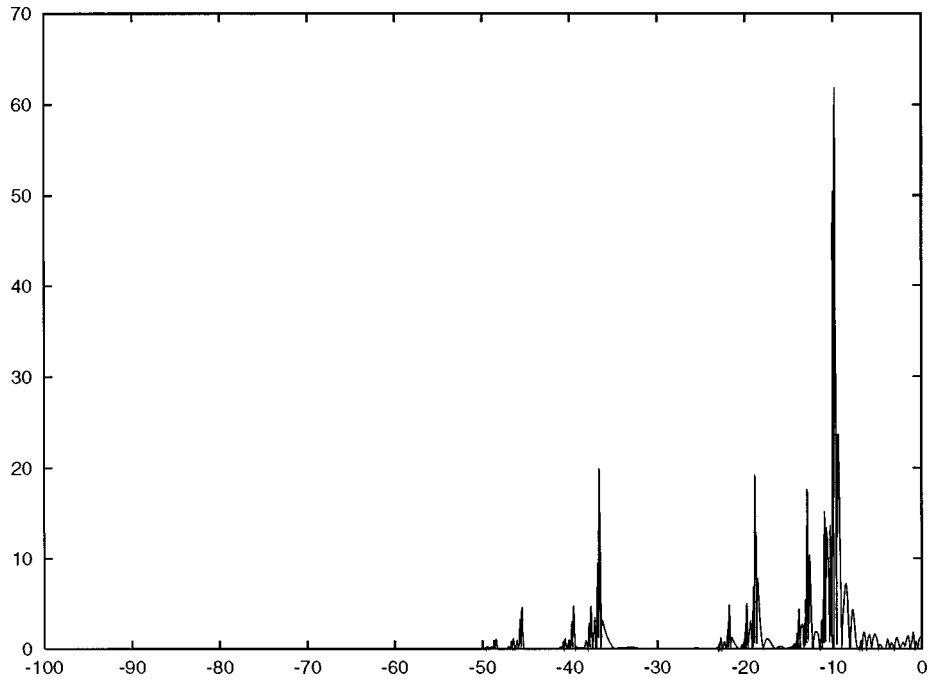

FIG. 3. Same as Fig. 2 with $a=0.05$ and $t=80$.

Cantor measure. Self-similar measures can be constructed by an iterative procedure. Starting from the characteristic function $\mathbb{I}_{[a,b]}$ of some interval $[a, b]$, the sequence $V_n=U^n\mathbb{I}_{[a,b]}$, where $Uf=\Sigma_{i=1}^{N}p_i f\circ S_i^{-1}$, is shown to converge toward V in the sense of distributions (see, e.g., Ref. 31). Thus, a self-similar measure can be approximated arbitrarily well by a finite sum of characteristic functions. Now for these approximations, the corresponding reflection coefficient $r_{k,n}^{+}$ can be computed by the transfer matrix method (see Ref. 11 for a description of the algorithm). Moreover, we have uniform convergence $r_{k,n}^{+}\rightarrow r_k^{+}$ on any finite range of frequencies.

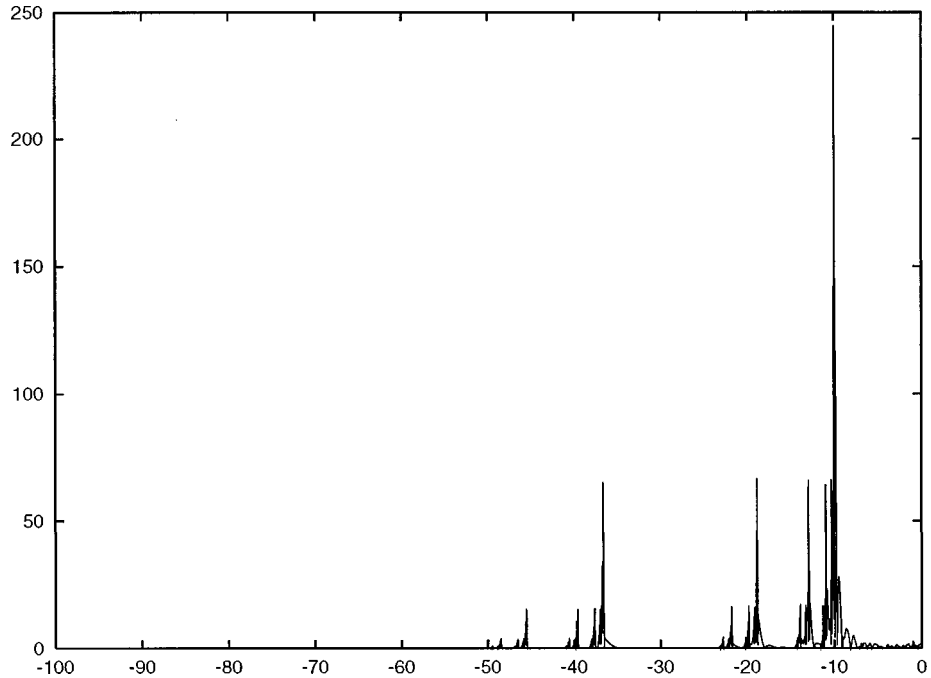

FIG. 4. Same as Fig. 2 with $a=0.01$ and $t=80$. The structure of the potential is recovered, but dilated by two.

We have computed the wave functions $\Phi_{t,b,a}$ when V is an approximation $(n=6)$ of the triadic Cantor measure on the interval $[0, 20]$ with $p_1=0.2$, $p_2=0.8$ (Fig. 1). We have taken a family of initial states $f_{b,a}$, where f is the second derivative (in order to have two vanishing moments) of a Gaussian, located to the left of the potential:

$$f(x)=\partial_{xx}\left(\frac{1}{\sqrt{2\pi}}\,e^{-|x|^2/2}\right). \tag{12.1}$$

Note that the effective support of this function is finite in the numerical application. In Figs. 2–4, we have plotted the modulus of the wave function $\Phi_{t,b,a}$ in the left region after a sufficiently long time (so that the incident part wave has completely traveled to the right and only the reflected part of the wave function is visible) for different dilation parameters. As the scale parameter a becomes smaller, the pattern of the wave function converges to the wavelet transform of the potential, but dilated by two (this is due to the fact that the interrogating wave, which is sent in and then reflected, interacts with the scatterer during a time which is twice the length of the support.)

ACKNOWLEDGMENTS

Many thanks to Adriaan Tip and Sandro Vaienti for helpful discussions.

[1] G. Korvin, *Fractal Models in the Earth Sciences* (Elsevier, Amsterdam, 1992).
[2] S. K. Sinha, ''Scattering from fractal structures,'' Physica D **38**, 310–314 (1989).
[3] E. Jakeman, ''Scattering by fractals,'' in *Fractals in Physics*, edited by L. Pietronero and E. Tosatti (Elsevier Sciences, Amsterdam, 1986), pp. 55–60.
[4] J. Teixeira, ''Experimental methods for studying fractal aggregates,'' in *On Growth and Form*, edited by H. E. Stanley and N. Ostrowsky (Nijhoff, Dordrecht, 1986).
[5] P. Pfeifer, *Chemistry and Physics of Solid Surfaces VII* (Springer Verlag, New York, 1988), pp. 283–303.
[6] P. W. Schmidt, ''Small-angle scattering studies of disordered, porous and fractal systems,'' J. Appl. Crystallogr. **24**, 414–435 (1991).
[7] T. Vicsek, *Fractal Growth Phenomena* (World Scientific, Singapore, 1989).
[8] D. L. Jaggard, ''Fractal electrodynamics: from superantennas to superlattices,'' in *Fractals in Engineering*, edited by J. Levy Véhel, E. Lutton, and C. Tricot (Springer Verlag, New York, 1997).
[9] M. V. Berry, ''Diffractal echoes,'' J. Phys. A **14**, 3101–3110 (1981).
[10] D. R. Yafaev, *Mathematical Scattering Theory*, Translations of Mathematical Monographs (American Mathematical Society, Providence, 1980), Vol. 105.
[11] C. A. Guérin and M. Holschneider, ''Scattering on fractal measures,'' J. Phys. A **29**, 7651–7667 (1996).
[12] C. A. Guérin, M. Holschneider, and M. Saillard, ''Electromagnetic scattering on multi-scales rough surfaces,'' Waves Random Media **7**(3), 331–349 (1997).
[13] L. D. Faddeev, ''Properties of the s-matrix of the one-dimensional equation,'' Am. Math. Soc. Trans. II **65**, 139–166 (1964).
[14] P. Deift and E. Trubovitz, ''Inverse scattering on the line,'' Commun. Pure Appl. Math. **32**, 121–251 (1979).
[15] R. J. Newton, ''Inverse scattering. I. One dimension,'' J. Math. Phys. **21**, 493–505 (1980).
[16] M. Holschneider, ''Fractal wavelet dimension and localization,'' Commun. Math. Phys. **160**, 457–473 (1994).
[17] Ya. B. Pesin, ''On rigorous mathematical definition of the correlation dimension and generalized spectrum for dimension,'' J. Stat. Phys. **71**(3/4), 529–547 (1993).
[18] C. A. Guérin and M. Holschneider, ''On equivalent definitions of the correlation dimension for a probability measure,'' J. Stat. Phys. **86**(3/4), 707–720 (1997).
[19] C. A. Guérin, ''Diffusion sur les mesures fractales,'' Ph.D. thesis, Université de Provence, Marseille, 1998.
[20] W. Rudin, *Real and Complex Analysis* (McGraw–Hill, New York, 1974).
[21] K. Yosida, *Functional Analysis* (Springer Verlag, Berlin, 1980), 6th ed.
[22] J. F. Brasche, ''Perturbation of Schrödinger Hamiltonian by measures,'' J. Math. Phys. **26**, 621–626 (1985).
[23] A. Tip, ''A class of form perturbations of the Laplacian on $l^2(r)$,'' J. Math. Phys. **31**, 308–315 (1990).
[24] J. F. Brasche, P. Exner, Yu. A. Kuperin, and P. Šeba, ''Schrödinger operators with singular interactions,'' J. Math. Anal. Appl. **184**(1), 112–139 (1994).
[25] J. Herczynki, ''On Schrödinger operators with distributional potentials,'' J. Operator Theory **21**, 273–293 (1989).
[26] M. S. Marinov and B. Segev, ''Analytical properties of scattering amplitudes in one-dimensional quantum theory,'' J. Phys. A **29**, 2839–2851 (1996).
[27] L. Hörmander, *The Analysis of Linear Partial Differential Operators 1*. (Springer Verlag, Berlin, 1990), 2nd ed.
[28] F. A. Berezin and M. A. Shubin, *The Schrödinger Equation*, Vol. 66 of *Mathematics and Its Applications* (Kluwer Academic, Dordrecht, 1991).
[29] M. Reed and B. Simon, *Fourier Analysis and Self-Adjointness* (Academic, New York, 1975).
[30] W. Amrein, J. Jauch, and K. Sinha, *Scattering theory in Quantum Mechanics* (Benjamin, New York, 1977).
[31] M. Holschneider, *Wavelets, an Analysis Tool* (Oxford U.P., Oxford, 1995).
[32] J. E. Hutchinson, ''Fractals and self similarity,'' Indiana Univ. Math. J. **30**(5), 713–746 (1981).
[33] R. S. Strichartz, ''Self-similar measures and their fourier transform 2,'' Trans. Am. Math. Soc. **336(1)**, 335–361 (1993).

Orthonormal sets of localized functions for a Landau level

J. Zak
Department of Physics, Technion—Israel Institute of Technology, Haifa, Israel 32000

(Received 15 January 1998; accepted for publication 27 March 1998)

Maximal sets of commuting magnetic translations are used for constructing a set of eigenfunctions for a Landau level on a von Neumann–Gabor lattice. Localization and orthogonality turn out to be two conflicting features of this set. It is shown how to construct complete orthonormal sets of optimally localized eigenfunctions on a von Neumann–Gabor lattice for each Landau level. By using the Balian–Low theorem it is pointed out that the uncertainties of the orbit center coordinates cannot both be made finite. © *1998 American Institute of Physics.*
[S0022-2488(98)00908-6]

Among the most elementary problems in early Quantum Mechanics is the motion of an electron in a magnetic field.[1] In a 2-dimensional case the energy spectrum of this problem consists of Landau levels which have been attracting much attention over the years. Each such Landau level is infinitely degenerate, which is a consequence of the magnetic translational symmetry,[2] the latter being a subgroup of the Weyl–Heisenberg group.[3] Maximal sets of commuting magnetic translations are closely linked to von Neumann lattices.[4,5] The latter is a rectangular lattice in the phase plane with a unit cell of area h. Such a lattice was first introduced by von Neumann[5] in the xp (coordinate-momentum)-plane and independently by Gabor[6] in the $t\nu$ (time-frequency)-plane.

The Balian[7]–Low[8] theorem relates to a set of states on a von Neumann lattice, and can be formulated in the following way. Given a square integrable function $\psi(x)$, one builds out of it by translations in the phase plane a von Neumann set $\psi_{mn}(x)$, where the indices m,n label the number of the unit cell on the von Neumann lattice in the phase plane $[\psi_{00}(x)\equiv\psi(x)]$. m and n assume all integer values $0,\pm1,\ldots$. In the theorem it is then claimed that if $\psi_{mn}(x)$ are orthogonal to $\psi_{m'n'}(x)$ for all $m,n\neq m',n'$ it follows that at least one of the two quantities $\langle x^2\rangle$ or $\langle p^2\rangle$ in the state $\psi(x)$ diverges (the triangular brackets denote the expectation value). Although discovered by physicists, the Balian–Low theorem has become of widespread interest in the Engineering literature of signal processing.[9]

In physics the von Neumann lattice has been known to appear in a natural way in the motion of an electron in a constant magnetic field $\boldsymbol{B}$ because of the magnetic translational symmetry of the problem.[2,10] In particular, this has to do with the construction of magnetic Wannier functions.[4,11] There is an interesting connection between the Balian–Low theorem and the properties of magnetic Wannier functions in the xy-plane for the case when $\boldsymbol{B}$ is parallel to the z-axis, despite the fact that the latter problem has two degrees of freedom. This can be seen in the following way. Instead of the pairs of the conjugate coordinates xp_x and yp_y one can choose two other pairs: the x- and y-components of the velocity v_x,v_y (which will later be connected to the operators P_W and W, respectively) and the components X,Y of the magnetic orbit center. The Hamiltonian does not depend on the XY-degree of freedom. They are constants of motion and they determine the infinitesimal magnetic translations.[2] Correspondingly, the commuting finite translations for X and Y create a von Neumann lattice in the phase plane of the orbit center.[10] These same translations when considered in the xy-plane are nothing else but the commuting magnetic translations (translations accompanied by gauge shifts),[2] which are used in the construction of magnetic Wannier functions.[12–14] As was already mentioned, when only a magnetic field is present and there is no periodic potential, the Hamiltonian does not depend on the coordinates X and Y of the orbit center.[15] Each Landau level is then infinitely degenerate with respect to the location of this center and one can therefore apply the Balian–Low theorem to it.

This paper is an extension of a publication by the author in Ref. 16, and it shows now that the Balian[7]–Low[8] theorem can be directly applied in the construction of orthonormalized sets of localized eigenfunctions for a single Landau level. When the motion is restricted to the xy-plane

 © 1998 American Institute of Physics

perpendicular to the magnetic field $\boldsymbol{B}$ the energy spectrum consists of discrete levels which carry the name of Landau.[1] The wave function for each Landau level l can be written either as $\psi_l(x,y)$ or $\phi_l(W,X)$, where W and X are defined in the following way ($[W,\ P_W]=[X,\ P_X]=i\hbar$):

$$W\equiv\frac{m\lambda^2}{\hbar}v_y=\frac{\lambda^2}{\hbar}\left(p_y+\frac{\lambda^2}{\hbar}x\right),\quad P_W\equiv mv_x=p_x, \tag{1}$$

$$X=\frac{\lambda^2}{\hbar}\left(p_x+\frac{\hbar}{\lambda^2}y\right),\quad P_X=p_y=\frac{\hbar}{\lambda^2}Y. \tag{2}$$

Here λ^2 is the square of the cyclotron radius, WP_W denotes the velocity degree of freedom and XP_X—the orbit center degree of freedom (as was already mentioned above). The relation between the wave function $\psi(x,y)$ in the xy-representation and the wave function $\phi(WX)$ in the WX-representation is[17]

$$\psi(x,y)=\frac{1}{2\pi\lambda^2}\int\int\exp\left[-\frac{i}{\lambda^2}(xy+WX-xX-yW)\right]\phi(W,X)dWdX. \tag{3}$$

Since the Hamiltonian for an electron in a magnetic field does not depend on the orbit center degree of freedom,[15] the wave function $\phi(WX)$ separates into a product function with one factor $\gamma_l(W)$ depending on W and the other one $\delta(X)$ on X only. The factor $\delta(X)$ in the wave function can be chosen completely arbitrarily and one can then apply to the XY-degree of freedom the Balian–Low theorem. This is done by applying to $\delta(X)$ the complete commuting set of translations in the XY-plane, which are also the commuting magnetic translations. They create the von Neumann set, and an explicit condition is given on $\delta(X)$ for the set to be orthogonal.[7] For this purpose the kq-representation[14] turns out to be convenient. Having carried everything out in the WX-coordinates one can then transform the results into the xy-plane [Eq. (3)]. An analysis of the Balian–Low theorem for an arbitrary Landau level is presented. It should be stressed that the magnetic field problem considered here contains two degrees of freedom, XP_X and WP_W. However, only the latter enters into the Hamiltonian [see Eq. (15)]. The von Neumann lattice is built in the phase plane of XP_X, which correspondingly leads to a lattice in the xy-plane.

We start by formulating the Balian–Low theorem in the XY-plane of the orbit center, or in the X and P_X couple of conjugate coordinates in Eq. (2). The shift operator $D(\alpha)$ is used,[18]

$$D(\alpha)=\exp(\alpha a^+-\alpha^* a), \tag{4}$$

where a is the annihilation operator,

$$a=\frac{1}{\lambda\sqrt{2}}\left(X+\frac{i}{\hbar}\lambda^2P_X\right), \tag{5}$$

and α is a number,

$$\alpha=\frac{1}{\lambda\sqrt{2}}\left(\overline{X}+\frac{i}{\hbar}\lambda^2\overline{P}_X\right). \tag{6}$$

Here $\overline{X}$ and $\overline{P}_X$ are expectation values of the operators X and P_X in the eigenstate of a. a^+ is the Hermitian conjugate of a and α' is the complex conjugate of α. A von Neumann set in the XP_X phase plane is[5]

$$\delta_{mn}(X)=D(\alpha_{mn})\,\delta(X), \tag{7}$$

where α_{mn} is obtained from α in Eq. (6) by setting $\overline{X}=md$, $\overline{P}_X=n\hbar\,(2\pi/d)$ with d a constant. $\delta(X)$ in Eq. (7) is an arbitrary square integrable function. The properties of the von Neumann set in Eq. (7) are determined by the initial function $\delta(X)$. The Balian–Low theorem states that if the

functions in Eq. (7) are orthogonal for $mn \neq m'm'$, then at least one of the uncertainties ΔX or ΔP_X in the $\delta(X)$-state diverges. An example of a normalized δ-function, that leads to an orthogonal von Neumann set in Eq. (7), is

$$\delta(X) = \begin{cases} \dfrac{1}{\sqrt{d}}, & |X| < d/2, \\ 0, & |X| > d/2. \end{cases} \tag{8}$$

For the δ in Eq. (8), $\Delta X = d/\sqrt{12}$ and $\Delta P_X = \infty$, in agreement with the Balian–Low theorem.

The condition on δ for an orthogonal von Neumann set in Eq. (7) was first formulated in the kq-representation[14,19] (see also Ref. 7). This condition is that the absolute value of the kq-function, $C(k,q)$, for δ is constant,[16]

$$|C(k,q)| = \text{const}, \tag{9}$$

where (a is an arbitrary constant) $C^{(a)}(k,q)$ and $\delta(X)$ are related in the following way:[12]

$$C^{(a)}(k,q) = \left(\frac{a}{2\pi}\right)^{1/2} \sum_n e^{ikan} \delta(q - na). \tag{10}$$

It is easy to check that the function $\delta(X)$ in Eq. (8) leads to $|C^{(d)}(k,q)| = \text{const}$ [$C^{(d)}(k,q)$ is the kq-function in Eq. (10) with a replaced by d]. A normalized kq-function that satisfies Eq. (9) is necessarily a pure phase factor,

$$C(k,q) = \frac{1}{\sqrt{2\pi}} \exp[i\varphi(k,q)], \tag{11}$$

where $\varphi(k,q)$ is real. It is to be pointed out that the kq-functions $C(k,q)$ satisfy the following Bloch-like boundary conditions:[14]

$$C^{(a)}(k,q) = C^{(a)}\left(k + \frac{2\pi}{a}, q\right) = \exp(-ika) C^{(a)}(k, q+a). \tag{12}$$

It is easy to check that a kq-function, $C(k,q)$, that satisfies Eqs. (12) and the condition in Eq. (9) [or Eq. (11)] is necessarily discontinuous in either k or q (or both). The Balian-Low theorem is a consequence of this discontinuity.[7]

In the X-representation the condition given by Eq. (9) [or Eq. (11)] assumes the following form [Eq. (10) is used]:

$$a \sum_n \delta^*(X - na)\, \delta(X - na - la) = \delta_{l0}. \tag{13}$$

Again it is easy to check that the function in Eq. (8) with $d = a$ satisfies Eq. (13).

Now let us connect the Balian–Low theorem with the dynamics of an electron in a magnetic field B. When the motion is in the xy-plane with the magnetic field $\boldsymbol{B}$ in the z-direction, Schrödinger's equation assumes the form [the Landau gauge $\boldsymbol{A} = (0, Bx)$ is chosen]

$$\left[\frac{p_x^2}{2m} + \frac{(p_y + (\hbar/\lambda^2)x)^2}{2m}\right] \psi(x,y) = E\psi(x,y), \quad \lambda^2 = \frac{\hbar c}{eB}. \tag{14}$$

Eq. (14) is written in the xy-representation. For the magnetic field problem it is convenient to work in the WX-representation according to the transformation given in Eqs. (1)–(3). In the WX-representation, Schrödinger's equation [Eq. (14)] assumes the form

$$\left[\frac{P_W^2}{2m} + \frac{mw^2W^2}{2}\right] \phi(W,X) = E\phi(W,X). \tag{15}$$

Since the Hamiltonian does not depend on the orbit center degree of freedom XP_X, the wave function $\phi(W,X)$ can therefore be chosen as a product function,

$$\phi_l(W,X)=\gamma_l(W)\,\delta(X), \tag{16}$$

where l labels the Landau levels. On the other hand, the magnetic translations depend on the orbit center operators X and P_X only.[2,14] In particular, the commuting finite magnetic translations in x- and y-directions can be written in the following way:[10]

$$\begin{aligned}\tau_x(d)&=\exp\left[\frac{i}{\hbar}\left(p_x+\frac{\hbar}{\lambda^2}y\right)Nd\right]\equiv\exp\left(iX\frac{2\pi}{d}\right),\\ \tau_y(Nd)&=\exp\left(\frac{i}{\hbar}p_y d\right)\equiv\exp\left(\frac{i}{\hbar}P_X d\right),\end{aligned} \tag{17}$$

where N is an integer which is defined by the following rationality condition:[2]

$$N=\frac{hc/e}{Bd^2}=\frac{2\pi\lambda^2}{d^2}. \tag{18}$$

This relation has a simple meaning of the ratio of the elementary fluxon hc/e to the flux of the magnetic field B through a unit cell of area d^2 [d is an arbitrary constant;[20] see α_{mn} in Eq. (7)]. As is seen from Eq. (17), the magnetic translations depend only on the XP_X degree of freedom of the orbit center. We can therefore use directly the results in Eqs. (4)–(7) for constructing a von Neumann set for a Landau level, by replacing $\delta(X)$ in Eq. (16) by $\delta_{mn}(X)$ in Eq. (7). We get

$$\phi_{lmn}(W,X)=\gamma_l(W)\,\delta_{mn}(X). \tag{19}$$

Now, when $\delta(X)$ satisfies the condition given by Eq. (9) or (11), the set of functions in Eq. (19) for each Landau level l is orthogonal. By using the transformation in Eq. (3) we get the orthonormal von Neumann set $\psi_{lmn}(x,y)$ in the xy-representation [the operators in Eqs. (1) and (2) are used],

$$\psi_{lmn}(x,y)=(-1)^{mn}\exp\left(\frac{i}{\lambda^2}yNdn\right)\psi_l(x+Ndn,y+dm). \tag{20}$$

The function $\psi_l(x,y)$ in Eq. (20) is found according to Eqs. (3) and (16), and can be given by the two alternative forms,

$$\begin{aligned}\psi_l(x,y)&=\frac{1}{\sqrt{2\pi}}\int\exp(iyk)\,\gamma_l(k\lambda^2+x)F_\delta(k)\,dk\\ &=\frac{\exp(-ixy/\lambda^2)}{(1/\sqrt{2\pi})\,\lambda^2}\int\exp\left(\frac{i}{\lambda^2}xz\right)F_\gamma\left(\frac{z-y}{\lambda^2}\right)\delta(z)\,dz,\end{aligned} \tag{21}$$

where F_δ and F_γ are the Fourier transforms of δ and γ, respectively [γ and δ are the functions in Eq. (16)]. It should be pointed out that the condition given by Eq. (9) [or Eq. (11)] is necessary and sufficient for the von Neumann set [Eqs. (19) or (20)] for each Laudau level l to be orthogonal.

By using the results for the single degree of freedom XP_X [$P_X=(\hbar/\lambda^2)\,Y$], the Balian–Low theorem implies that the coordinates X and Y of the orbit center in the state ψ_l [Eq. (21)] for a magnetic field cannot both be well localized, or in a more precise language, at least one of the two uncertainties ΔX or ΔY in the ψ_l-state [Eq. (21)] diverges. It is interesting to point that for obtaining the result there is no need to go to the configuration plane xy [see Eq. (3)]. However, since the coordinates x and y of the electron in a magnetic field are related to the operators WP_W and XP_X [$P_X=(\hbar/\lambda^2)\,Y$] in Eqs. (1) and (2), one has

$$x=W-Y, \quad y=X-\frac{\lambda^2}{\hbar}P_W. \tag{22}$$

Bearing in mind that the uncertainties ΔW and ΔP_W are finite in the state given by Eq. (16) [or alternatively, by Eq. (21)], we come to the conclusion [by using Eq. (22)] that at least one of the uncertainties Δx or Δy diverges in the state $\psi_l(x,y)$ of Eq. (21). This also means that the electron in a magnetic field cannot be well localized in both the x- and y- directions.

It is of interest to compare the consequences of the Balian–Low theorem with the known results for the Wannier function in a magnetic field.[11,21] In Ref. 11 it was shown that Wannier functions for a Bloch electron in the xy-plane, with the magnetic field B perpendicular to the plane, cannot fall off at infinity faster than r^{-2} ($r^2=x^2+y^2$). In Ref. 21 Wannier functions with a r^{-2} fall off were actually constructed for such a two-dimensional problem where only a magnetic field is present. One can see that the Balian-Low theorem for Landau levels is in good agreement with the results of Refs. 11 and 21. In order to see it, we notice that as a consequence of the Balian–Low theorem the quantity $\langle x^2+y^2\rangle=\langle r^2\rangle$ diverges (the triangular brackets denote the expectation value) because either $\langle x^2\rangle$ or $\langle y^2\rangle$ has to diverge. This divergence of $\langle r^2\rangle$ also follows when the function falls off as r^{-2} at infinity, like in Ref. 21.

It should be pointed out that the Balian–Low theorem is completely general in nature. For example, when $\delta(X)$ is given by Eq. (8), then $\psi_0(x,y)$ in Eq. (21) assumes the form (for the lowest Landau level, $l=0$)

$$\psi_0(x,y)=\left(\frac{1}{4\pi^3 d^2\lambda^6}\right)^{1/4}\exp\left(-\frac{ixy}{\lambda^2}\right)\int_{-d/2}^{d/2}\exp\left[\frac{i}{\lambda^2}xz-\frac{(y-z)^2}{2\lambda^2}\right]dz. \tag{23}$$

In deriving Eq. (23), use was made of the second line in Eq. (21). For large y, Eq. (23) becomes

$$\psi_0(x,y)\sim\left(\frac{d^2}{4\pi^3\lambda^6}\right)^{1/4}\exp\left(-\frac{ixy}{\lambda^2}-\frac{y^2}{2\lambda^2}\right)\sin\left(\frac{xd}{2\lambda^2}\right)\Big/\left(\frac{xd}{2\lambda^2}\right). \tag{24}$$

This function has a $1/x$ fall-off in the x-direction and is Gaussian in the y-direction.

Finally, it can be directly checked that the von Neumann set in Eq. (20) is orthonormal, when $\psi_l(x,y)$ is given by Eq. (21). In the particular case of the localized function in Eq. (23), the explicit orthonormal von Neumann set is [see Eq. (20)]

$$\psi_{0mn}(x,y)=\frac{(-1)^{mn}}{(4\pi^3 d^2\lambda^6)^{1/4}}\exp\left(-\frac{i}{\lambda^2}xy-\frac{i}{\lambda^2}xmd\right)$$
$$\times\int_{-d/2}^{d/2}\exp\left[\frac{i}{\lambda^2}(x+nNd)z-\frac{(y+md-z)^2}{2\lambda^2}\right]dz. \tag{25}$$

In conclusion, the Balian–Low theorem was applied to the motion of an electron in the xy-plane, with a magnetic field perpendicular to this plane. By using this theorem it was shown that orthonormality on a von Neumann lattice and localizability are incompatible features. In particular, it is shown that even for the best localized eigenfunctions for the Landau levels the uncertainties Δx and Δy can never both be made finite. On a qualitative level, incompatibility of orthonormality and localizability is not an unexpected feature for wave functions in quantum mechanics.[22] In this paper we derive complete orthonormal sets of localized functions for Landau level wave functions. Although the derivation was carried out in the Landau gauge, the result that Δx and Δy can never both be finite in a Landau level state $\psi_l(x,y)$ [see Eq. (20)], for any l, is gauge independent. This is seen from the expression of the wave function in the WX-representation [Eq. (19)]. The part of the wave function $\delta_{mn}(X)$ in Eq. (19) that leads to the von Neumann set does not depend on the gauge because the X-coordinate does not appear in the Hamiltonian of the problem [Eq. (14)].

As is well known, in the symmetric gauge the wave function for any Landau level can be chosen well localized in both x- and y-directions.[15] However, for such a well localized wave

function, the von Neumann set in Eq. (20) will *not be orthogonal*. As was shown in this paper the uncertainties Δx and Δy for the functions in the orthonormal set [Eq. (20)] cannot both be finite. This, however, does not limit the applicability of these functions in perturbation theory for well localized potentials. We should also point out that the idea of constructing orthonormalized sets of localized functions for the lowest Landau level was recently utilized in the framework of wavelets.[23] The orthonormal sets in Eq. (20) are built on a von Neumann lattice which is related to the Weyl–Heisenberg group. The orthonormal set in Ref. 23 is related to the group $ax+b$.

ACKNOWLEDGMENTS

This research was supported in part by the Fund for the Promotion of Research at Technion.

[1] L. D. Landau, Z. Phys. **64**, 629 (1930).
[2] E. Brown, Phys. Rev. **133**, A1038 (1964); J. Zak, *ibid.* **134**, A1602, A1607 (1964).
[3] J. Zak, Phys. Rev. B **39**, 694 (1989).
[4] I. Dana and J. Zak, Phys. Rev. B **28**, 811 (1983).
[5] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1955), pp. 405–407.
[6] D. Gabor, J. Inst. Electr. Eng. (London) **93**, 429 (1946).
[7] R. Balian. C. R. Acad. Sci. Paris **292**, 1357 (1981).
[8] F. Low, in *Passion for Physics—Essays in Honor of Geoffrey Chew*, edited by C. DeTar *et al.* (World Scientific, Singapore, 1985), p. 17.
[9] I. Daubechies and A. J. E. M. Janssen, IEEE Trans. Inf. Theory **39**, 3 (1993); J. J. Benedetto, C. Heil, and D. F. Walnut, J. Fourier Anal. Appl. **1**, 355 (1995).
[10] J. Zak, *Proceedings of the International Symposium on Coherent States: Past, Present and Future*, edited by D. H. Feng and J. R. Klauder (World Scientific, Singapore, 1994), pp. 553–560; J. Zak, Europhys. Lett. **17**, 443 (1992), and references therein.
[11] D. J. Thouless, J. Phys. C **17**, L325 (1984).
[12] R. Peierls, Z. Phys. **80**, 763 (1933).
[13] J. M. Luttinger, Phys. Rev. **84**, 814 (1951).
[14] J. Zak, in *Solid State Physics*, edited by F. Seitz, D. Turnbull, and H. Ehrenreich (Academic, New York, 1972), Vol. 27.
[15] M. H. Johnson and B. A. Lippman, Phys. Rev. **76**, 828 (1949).
[16] J. Zak, Phys. Rev. Lett. **79**, 533 (1997).
[17] M. Boon, in *Lecture Notes in Physics* (Springer-Verlag, Berlin, 1976), Vol. 50, p. 282.
[18] R. J. Glauber, Phys. Rev. **131**, 2766 (1963).
[19] H. Bacry, A. Grossmann, and J. Zak, Phys. Rev. B **12**, 118 (1975).
[20] The constant d is completely arbitrary when no periodic potential is involved. (In the presence of such a potential d can be the lattice constant.)
[21] E. I. Rashba, L. E. Zhúkov, and A. L. Efros, Phys. Rev. B **55**, 5306 (1997).
[22] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[23] J. P. Antoine and F. Bagarello, J. Phys. A **21**, 2471 (1994); F. Bagarello, *ibid.* **27**, 5583 (1994); and **29**, 565 (1996).

Tau-wavelets in the plane

Jean Pierre Gazeau[a)]
Laboratoire de Physique Théorique de la Matière Condensée,
Université Paris 7 Denis Diderot, 2 Place Jussieu, 75251 Paris Cedex 05, France

Jiří Patera[b)]
Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada

Edita Pelantová[c)]
Department of Mathematics, Faculty of Nuclear Science and Physical Engineering,
Czech Technical University, Prague, Czech Republic

(Received 18 December 1997; accepted for publication 31 March 1998)

First an iterative process is described for subdivision and for scaling-up of the Penrose–Robinson tiles. Then we build the τ-wavelets of Haar (i.e., suitably normalized characteristic functions) on the tiles so that the orthogonality and multiresolution conditions are satisfied. Here $\tau=\frac{1}{2}(1+\sqrt{5})$. © *1998 American Institute of Physics.* [S0022-2488(98)01608-9]

I. INTRODUCTION

This paper should be seen as a continuation (and also an improvement) of our recent construction of discrete wavelets based on an algebraic scaling factor.[1,2] We call these functions tau-wavelets of Haar. The mathematical idea backing our approach originated in the identity

$$\tau^j=\tau^{j-1}+\tau^{j-2}, \tag{1.1}$$

valid for any $j\in\mathbb{Z}$ and obeyed by the (ubiquitous) golden mean

$$\tau=\frac{1}{2}\,(1+\sqrt{5})=2\cos\frac{\pi}{5}=1.618\ldots\,. \tag{1.2}$$

Indeed, comparing (1.1), with the dyadic splitting,

$$2^j=2^{j-1}+2^{j-1}, \tag{1.3}$$

which underlies most of the wavelet constructions along multiresolution guidelines, it was natural to explore the ''τ-adic'' multiresolution analysis. With regard to their algebraic properties exemplified by (1.1) and (1.3), we can view 2 and τ as similar at a certain extent. This fact has deep repercussions on the relationship between the two kinds of wavelets. Whereas the dyadic wavelets (Ref. 3 and references therein) are the functions $\psi(x)\in L^2(\mathbb{R})$ such that the family of functions,

$$\psi_{j,k}:=2^{j/2}\psi(2^jx-k),\quad j,k\in\mathbb{Z}, \tag{1.4}$$

form an orthonormal basis in $L^2(\mathbb{R})$[6], the same holds (or can be defined so) for τ-adic wavelets.

By a τ-adic wavelet[1,2] we mean a vector $\psi\in L^2(\mathbb{R})$ such that the family

$$\tau^{j/2}\psi(\tau^jx-b),\quad j\in\mathbb{Z},$$

is an orthonormal basis in $L^2(\mathbb{R})$ when b assumes its values in an appropriate subset Λ of the set $\mathbb{Z}_\tau$ of τ-integers. The set $\mathbb{Z}_\tau$ is discrete and can be defined as the set of real numbers which have a zero fractional part in the numeration system[4] based on τ,

a)Electronic mail: gazeau@ccr.jussieu.fr
b)Electronic mail: patera@crm.umontreal.ca
c)Electronic mail: pelantova@km1.fjfi.cvut.cz

 © 1998 American Institute of Physics

$$\mathbb{Z}_\tau = -\mathbb{Z}_\tau. \tag{1.5}$$

It tiles the real line aperiodically with two tiles of length 1 and $1/\tau$. The subset Λ is given by

$$\Lambda = \tau\mathbb{Z}_\tau^+ \cup (-1-\tau\mathbb{Z}_\tau^+) \subset \mathbb{Z}_\tau,$$

where $\mathbb{Z}_\tau^+$ stands for the set of non-negative elements of $\mathbb{Z}_\tau$. Equivalently, one can define $\mathbb{Z}_\tau^+$ as the non-negative part of the 1-dimensional quasicrystal with acceptance window the open interval $(-1,\tau)$; see (2.11) and (2.12) below.

The extension of τ-wavelets of Haar to smooth, compactly supported, τ-wavelets is conceivable in spite of the current lack of algorithms like Daubechies construction.[3] Indeed, standard multiresolution involves integers for scaling parameters 2^j as well as for the translation parameter k in (1.4). The underlying additive group structure of the integers greatly facilitates any harmonic approach which makes Fourier series enter into the game. With the τ-integers we lose the ring structure. The next best property is the Meyer's,[5] $\mathbb{Z}_\tau + \mathbb{Z}_\tau \subset \mathbb{Z}_\tau + F$ and also[6] $\mathbb{Z}_\tau \times \mathbb{Z}_\tau \subset \mathbb{Z}_\tau + G$, where F and G are finite sets. These properties appear to pave the road to analysis of τ-wavelets based on almost periodic functions.[5,7,8] Another indication in favor of this direction are the summation formulas for Fourier exponentials in Refs. 2, 9.

It is natural to ask the question about the reason for extending the τ-wavelets to two dimensions. Our motivation stems mainly from quasicrystalline studies.[10–13] Quasicrystals are alloys like $Al_{0.86}Mn_{0.14}$ or $Al_xPd_yMn_z$, discovered in 1984.[14] Their x-ray diffraction patterns show local n-fold point symmetry, with $n=5,8,10,12$. Similar features are unveiled in quasicrystalline surfaces by scanning microscopy. Such symmetries are crystallographically ''forbidden'' for being incompatible with translational invariance. An observed diffraction pattern is a constellation of bright Bragg spots of unequal intensity displaying self-similarity with irrational scaling factors. Similar self-similarity is encountered in other types of quasicrystal experiments. The scaling parameters involve the following (algebraic) integers:

$$\begin{aligned} &\tfrac{1}{2}(1+\sqrt{5}), \quad \text{for } n=5,10; \\ &1+\sqrt{2}, \quad \text{for } n=8; \\ &2+\sqrt{3}, \quad \text{for } n=12. \end{aligned} \tag{1.6}$$

Only in the first case an exhaustive description of self-similarities in a large class of n-dimensional quasicrystals can be found in literature.[15]

In the study of quasicrystals, there appears to be at least three types of problems amenable to wavelet analysis. The first one concerns the diffraction patterns (see, for example, the papers of Hof, and Gähler, Klitzing in Ref. 13) where, at a given resolution, it is necessary to classify and label the Bragg spots according to their position and intensity. Concepts like the Meyer sets Λ and their star map, yielding the dual Λ^*, are crucial in this content.[16–18]

The second application of wavelet analysis is the treatment of patterns obtained by the tunnelling or atomic force microscopy of quasicrystals surfaces. The images faithfully reflect electronic densities of atomic sites. Two-dimensional wavelets based on irrationalities (1.6) appear to be particularly appropriate for the scanning analysis, especially where it is a matter of discerning defects, shearing or other breaking of scaling invariance or local symmetries.

The third application, possibly the most promising one, concerns the determination of electronic transport properties in quasicrystals. Interesting results already exist for 1-dimensional models of aperiodic Hamiltonians (see Ref. 19 for a recent review) but the situation for real quasicrystals is far from being clarified. The lack of periodicity hampers the efficiency of making use of Fourier analysis and related theorems (Floquet, Bloch,...) for determining transport wave functions and spectral characteristics (gap positions, integrated density of states,...). However, it seems natural to expand the wave function of an electron, submitted to a quasiperiodic potential whose minima are distributed according to the nodes of a quasicrystal Λ, in terms of a wavelet family corresponding to Λ such as the one constructed in this paper.

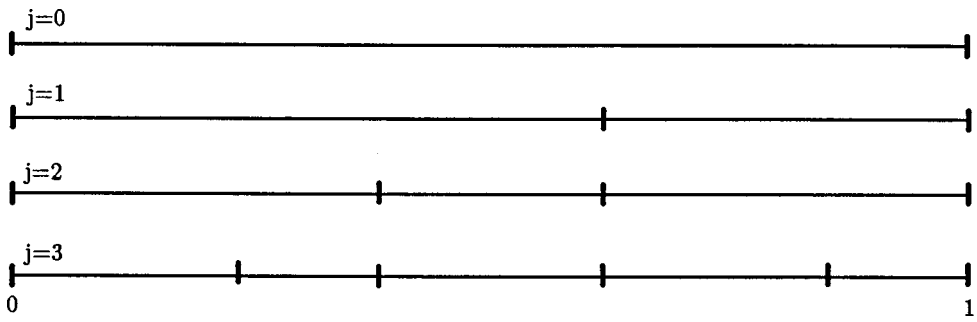

FIG. 1. The beginning of the iterative sequence of subdivisions of the unit interval.

There have been several attempts to involve wavelet analysis in the field of spectral problems (diffraction or Schödinger) related aperiodic (mostly self-similar) structures.[3,20] However, none makes use of the concept of algebraic discrete wavelet as the one presented in Ref. 2 and in this paper.

In the next section we recall the one-dimensional case in order to present in a simpler manner the algebraic τ-integers and the analytic τ-wavelet material. The two-dimensional case is considered in Section III. There we describe a nonunique process for subdividing the two types of Penrose–Robinson triangles. Section IV is devoted to the construction of τ-wavelets in the plane. Finally, Section V contains a definition of multiresolution analysis adapted to our present approach.

II. TAU-WAVELETS OF HAAR ON THE LINE

In this section we first recall the procedure followed in Ref. 2 during the construction of the Haar set based on scaling by the irrational factor $\tau=\frac{1}{2}(1+\sqrt{5})$. Later in the section we point out on this transparent one-dimensional case the strategy we use in two dimensions.

The algebraic nature of τ (it is a solution of the equation $x^2=x+1$) allows for the τ-adic property,

$$\frac{1}{\tau^j}=\frac{1}{\tau^{j+1}}+\frac{1}{\tau^{j+2}}, \quad j\in\mathbb{Z}.$$

This equality provides a subdivision of the unit interval into two parts:

$$A=[0,1]=[0,1/\tau]\cup[1/\tau,1]. \tag{2.1}$$

Equation (2.1) is the starting point of an iterative sequence of subdivisions of A into intervals (Fig. 1). At each step only the large intervals get divided. Thus at each stage we have two lengths of intervals:

$$A_{j,b}=\left[\frac{b}{\tau^j},\frac{b+1}{\tau^j}\right], \quad j\in\mathbb{N}, \quad A_{0,0}=A; \tag{2.2}$$

$$A_{j,b}=A_{j+1,\tau b}\cup A_{j+2,\tau^2 b+\tau}, \tag{2.3}$$

where b is what we called a τ-integer.[2] It is submitted to a certain boundedness condition.

Let us first explain the concept of the τ-integer.[21,1,2] It is derived from the system of numeration based on τ.[4] Each non-negative real number x can be expanded in powers of τ with coefficients of the expansion taking only two values: 0 or 1. More precisely,

$$x=\sum_{l=-\infty}^{j}\xi_l\tau^l=\xi_j\tau^j+\xi_{j-1}\tau^{j-1}+\ldots, \tag{2.4}$$

with

$$\xi_l \in \{0,1\}, \quad \xi_l \xi_{l-1} = 0. \tag{2.5}$$

The second condition in (2.5) means that no partial sum like $\tau^k + \tau^{k-1}$ occurs in (2.4), owing to the fact that it is equal to τ^{k+1}.

Coefficients ξ_k in (2.4) are computed using the so called Rényi algorithm: First, j in (2.4) is the integer satisfying

$$\tau^j \leqslant x < \tau^{j+1}, \quad \text{i.e.,} \quad 1 \leqslant \frac{x}{\tau^j} < \tau, \tag{2.6}$$

and so

$$\xi_j = \left[\frac{x}{\tau^j}\right] := \text{integer part of } \frac{x}{\tau^j} = 1.$$

Then we put

$$r_j = \left\{\frac{x}{\tau^j}\right\} := \text{fractional part of } \frac{x}{\tau^j},$$

and the other coefficients ξ_k, $k<j$ are determined recursively from (ξ_j, r_j):

$$\xi_k = [\tau r_{k+1}], \quad r_k = \{\tau r_{k+1}\}. \tag{2.7}$$

Equivalently we have

$$\xi_{j-k} = [\tau T_\tau^k (x/\tau^{j+1})], \tag{2.8}$$

where

$$T_\tau y := (\tau y) \pmod 1, \quad \text{for any } y \in [0,1].$$

In this system of numeration based on τ, the set of non-negative τ-integers is the strictly increasing sequence of real numbers having only non-negative powers of τ in their τ-expansion,

$$\mathbb{Z}_\tau^+ = \{0,1,\tau,\tau^2,\tau^2+1,\tau^3,\tau^3+1,\tau^3+\tau,\tau^4,\ldots\}. \tag{2.9}$$

The set $\mathbb{Z}_\tau$ of τ-integers is then defined by

$$\mathbb{Z}_\tau = \mathbb{Z}_\tau^+ \cup (-\mathbb{Z}_\tau^+). \tag{2.10}$$

It was pointed out in Ref. 22 that the numbers in $\mathbb{Z}_\tau^+$ can be introduced as the subset of all the non-negative points $a+b\tau$ of the 1-dimensional cut and project quasicrystal,

$$\Sigma((-1,\tau)) = \{a+b\tau \,|\, a,b \in \mathbb{Z}, a+b\tau' \in (-1,\tau)\}, \tag{2.11}$$

whose acceptance window is the open interval $(-1,\tau)$. Here τ and τ' are the two solutions of the quadratic equation $x^2 = x+1$. Note that $\Sigma((-1,\tau))$ has the reflection symmetry in the point $-\tau/2$. In order to have the desired reflection symmetry (1.5) in the origin, we use (2.10) together with

$$\mathbb{Z}_\tau^+ = \Sigma((-1,\tau))^+ = \{a+b\tau \geqslant 0 \,|\, a,b \in \mathbb{Z}, a+b\tau' \in (-1,\tau)\}. \tag{2.12}$$

Furthermore, from Ref. 22 we know all the inflation (or scaling) centers of the point set $\Sigma((-1,\tau))$. We say that x is an s-inflation center of $\Sigma((-1,\tau))$ provided

$$s(y-x)+x \in \Sigma((-1,\tau)), \quad \text{for all } y \in \Sigma((-1,\tau)). \tag{2.13}$$

That includes the centers which belong to $\Sigma((-1,\tau))$ as well as all those which lie outside of it. Consequently, we also know the scaling properties of the set $\mathbb{Z}_\tau^+$. In particular, we know not only that every point $x \in \mathbb{Z}_\tau^+$ is an inflation center of $\mathbb{Z}_\tau^+$, but also all its scaling factors s. It was shown

in Ref. 22 that the scaling factors s, appropriate for all points $x \in \Sigma((-1,\tau))$, form the one-dimensional quasicrystal $\Sigma([0,1])$ contained in $\Sigma((-1,\tau))$. For just one point $x \in \Sigma((-1,\tau))$, the admissible scaling factors form in general a larger x-dependent quasicrystal denoted by $\Sigma([-\mu(x),1])$ in Ref. 22. In all cases one has

$$\Sigma([0,1]) \subseteq \Sigma([-\mu(x),1]) \subset \Sigma((-1,\tau)).$$

We now come to the explicit construction of the τ-Haar basis in $L^2[0,1]$. For that we need a ''mother'' wavelet called here the τ-Haar wavelet.

Definition 2.1: The τ-Haar wavelet is the function defined on the real line $\mathbb{R}$ as

$$h^\tau(x) = \begin{cases} \tau^{-1/2}, & \text{for } x \in \left[0, \dfrac{1}{\tau}\right], \\ -\tau^{1/2}, & \text{for } x \in \left]\dfrac{1}{\tau}, 1\right], \\ 0, & \text{otherwise.} \end{cases} \tag{2.14}$$

One can verify directly that

$$\int_0^1 (h^\tau(x))^2 dx = \frac{1}{\tau^2} + \tau\left(1 - \frac{1}{\tau}\right) = 1, \tag{2.15}$$

$$\int_0^1 h^\tau(x)\,dx = \tau^{-1/2} \cdot \frac{1}{\tau} - \tau^{1/2} \cdot \frac{1}{\tau^2} = 0. \tag{2.16}$$

Thus $h^\tau(x)$ is orthogonal to $\chi_{[0,1]} \equiv \chi_A$, the characteristic function of the unit interval.

The following summarizes the construction.

Theorem 2.2: The system

$$\{\tau^{j/2} h^\tau(\tau^j x - b) \equiv h^\tau_{j,b}(x), \tau^{j/2} h^\tau(\tau^j x + b + 1)\}, \quad j \in \mathbb{Z}, \ b \in \tau\mathbb{Z}_\tau, \tag{2.17}$$

is an orthonormal basis in $L^2(\mathbb{R})$. It is called the τ-Haar basis of $L^2(\mathbb{R})$.

The following expansion formula is crucial for the proof of the theorem

$$\tau^{-1/2} \sum_{j<0} \tau^{j/2} h^\tau_{j,0}(x) = \chi_A(x). \tag{2.18}$$

Our task in this paper is to construct the τ-wavelets in the plane. The analogy with the one-dimensional Haar wavelets can be made very close. For that, however, we have to reformulate the one-dimensional problem in an equivalent way.

All Haar wavelets with support $[0,1]$ have the form

$$f_{j,k}(x) = 2^{j/2} f(2^j x - k), \quad j = 0,1,...; \quad k = 0,1,...,2^j - 1.$$

These functions can be generated from the step function f by means of the scale-up function $\varphi(x) = 2x$ and the displacement function $\alpha(x) = x - 1$. Let us describe the first steps of the generation process:

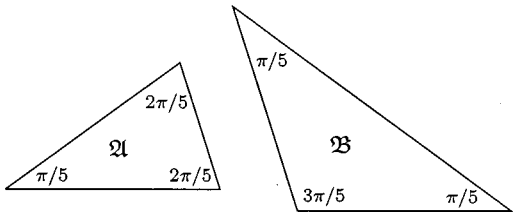

FIG. 2. The Penrose–Robinson triangles $\mathfrak{A}$ and $\mathfrak{B}$.

$$
\begin{aligned}
j=1:\quad & f_{1,0}=2^{1/2}f(2x)=2^{1/2}f(\varphi(x)),\\
& f_{1,1}=2^{1/2}f(2x-1)=2^{1/2}f(\alpha\varphi(x));\\
\\
j=2:\quad & f_{2,0}=2\,f(4x)=2\,f(\varphi\varphi(x)),\\
& f_{2,1}=2\,f(4x-1)=2\,f(\alpha\varphi\varphi(x)),\\
& f_{2,2}=2\,f(4x-2)=2\,f(\varphi\alpha\varphi(x)),\\
& f_{2,3}=2\,f(4x-3)=2\,f(\alpha\varphi\alpha\varphi(x));\\
& \vdots
\end{aligned}
$$

Thus by composing the mother function with ''admissible'' words formed from the letters α and φ, we are able to generate all functions. Here a word w is admissible provided it ends by φ and it does not contain adjacent α's; an empty word is also admissible. Let us denote by ''Admis'' the set of admissible words, and by $l(w)$ the multiplicity of occurrence of the letter φ in the word w. Then all the Haar wavelets with support $[0,1]$ are described as follows:

$$\{2^{l(w)/2}f(w(x))\,|\,w\in\text{Admis}\}.$$

The same discussion is also valid for τ-Haar wavelets.

III. SUBDIVISION OF PENROSE–ROBINSON TRIANGLES

Our construction of τ-wavelets in the plane will start out from successive subdivisions of the Penrose–Robinson triangles. The latter will play the role of the interval $[0,1]$ in the previous section. More precisely, there will be two of them to play this role, namely the triangles $\mathfrak{A}$ and $\mathfrak{B}$ shown on Figure 2.

Let us denote by ΔABC a triangle with vertices labeled by A,B,C counterclock-wise in $\mathbb{C}$. The length of the side with vertices A,B is denoted by $\overline{AB}$.

Let us define on the set of all triangles in $\mathbb{C}$, congruent to either $\mathfrak{A}$ or $\mathfrak{B}$, a splitting operator Ψ,

$$\Psi(\Delta ABC)=(\Psi_{\mathfrak{A}}(\Delta ABC),\Psi_{\mathfrak{B}}(\Delta ABC)),$$

in the following way.

(a) Let ΔABC be similar to $\mathfrak{A}$, where A is the vertex at the angle $\pi/5$. Denote by D the point on the side AC such that $\overline{AD}:\overline{DC}=\tau$ (Fig. 3). Then we define the two components of the splitting to

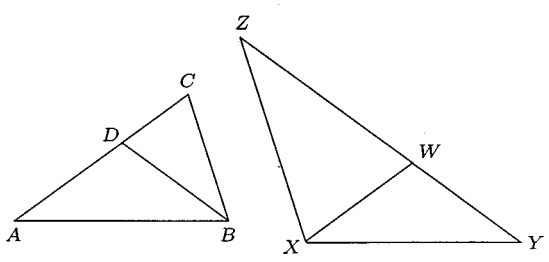

FIG. 3. The first subdivision of the triangles $\mathfrak{A}$ and $\mathfrak{B}$.

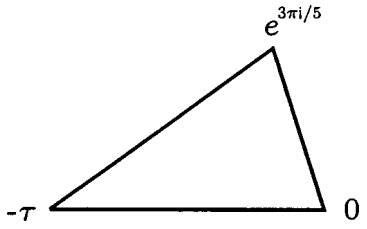

FIG. 4. The basic triangle T and the coordinates of its vertices in the complex plane.

be the following:

$$\Psi_{\mathfrak{U}}(\Delta ABC) := \Delta BCD, \quad \Psi_{\mathfrak{B}}(\Delta ABC) := \Delta ABD.$$

(b) Let ΔXYZ be similar to $\mathfrak{B}$, where X is the vertex at the angle $3\pi/5$. Denote by W the point on the side YZ such that $\overline{YW}:\overline{WZ} = 1/\tau$ (Fig. 3). Then we define the two components of the splitting to be the following:

$$\Psi_{\mathfrak{U}}(\Delta XYZ) := \Delta YWX, \quad \Psi_{\mathfrak{B}}(\Delta XYZ) := \Delta WZX.$$

It follows immediately that

• $\Delta = \Psi_{\mathfrak{U}}(\Delta) \cup \Psi_{\mathfrak{B}}(\Delta)$, where $\Psi_{\mathfrak{U}}(\Delta)$ is always similar to $\mathfrak{U}$ and $\Psi_{\mathfrak{B}}(\Delta)$ is always similar to $\mathfrak{B}$.

• If Δ is similar to $\mathfrak{U}$, then

$$\text{vol } \Psi_{\mathfrak{U}}(\Delta) : \text{vol } \Psi_{\mathfrak{B}}(\Delta) = 1/\tau.$$

• If Δ is similar to $\mathfrak{B}$, then

$$\text{vol } \Psi_{\mathfrak{U}}(\Delta) : \text{vol } \Psi_{\mathfrak{B}}(\Delta) = \tau.$$

Now we can start the wavelet construction. The basic area will be the triangle T given by its vertices $-\tau, 0, e^{3\pi i/5}$ in $\mathbb{C}$ (Fig. 4). Its area is given by vol $T = \frac{1}{2}\tau^2 \sin \pi/5$.

The idea of the subdivision is to proceed as follows. The triangle T is divided in the first step to $\Psi_{\mathfrak{U}}(T)$ and $\Psi_{\mathfrak{B}}(T)$. Subsequently at each step we have triangles of only two sizes. The division is accomplished by applying the operator Ψ to the triangles of larger volume (Fig. 5).

Let us now describe the subdivision in detail. Denote $\mathfrak{U}_n$ and $\mathfrak{B}_n$ the set of all triangles similar, respectively, to $\mathfrak{U}$ and $\mathfrak{B}$ which are obtained in the n-th step of the division process.

The sequences of division sets $(\mathfrak{U}_n)_0^\infty$ and $(\mathfrak{B}_n)_0^\infty$ can be described recursively as follows:

$$\begin{aligned}
&\mathfrak{U}_0 = \{T\}, && \mathfrak{B}_0 = \varnothing,\\
&\mathfrak{U}_1 = \{\Psi_{\mathfrak{U}}(T)\}, && \mathfrak{B}_1 = \{\Psi_{\mathfrak{B}}(T)\},\\
&\mathfrak{U}_2 = \{\Psi_{\mathfrak{U}}(T), \Psi_{\mathfrak{U}}\Psi_{\mathfrak{B}}(T)\}, && \mathfrak{B}_2 = \{\Psi_{\mathfrak{B}}\Psi_{\mathfrak{B}}(T)\},\\
&\vdots && \vdots\\
&\mathfrak{U}_{2n} = \mathfrak{U}_{2n-1} \cup \Psi_{\mathfrak{U}}(\mathfrak{B}_{2n-1}), && \mathfrak{B}_{2n} = \Psi_{\mathfrak{B}}(\mathfrak{B}_{2n-1}),\\
&\mathfrak{U}_{2n+1} = \Psi_{\mathfrak{U}}(\mathfrak{U}_{2n}), && \mathfrak{B}_{2n+1} = \mathfrak{B}_{2n} \cup \Psi_{\mathfrak{B}}(\mathfrak{U}_{2n}).
\end{aligned}$$

The number of similar triangles at each step can be readily determined using the following rule: Let n be even. Then the number of triangles similar to $\mathfrak{U}$ is equal to the term F_{n+1} of the

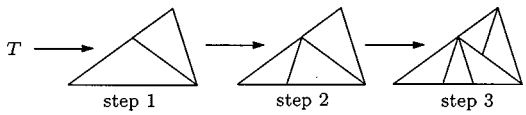

FIG. 5. The first three steps of the subdivision process.

Fibonacci series (which starts at $F_0=0$ and $F_1=1$) and the number of triangles similar to $\mathfrak{B}$ is equal to F_n. For n odd the numbers are interchanged for the triangles $\mathfrak{U}$ and $\mathfrak{B}$.

Note that all the triangles in $\mathfrak{U}_{2n}\cup\mathfrak{U}_{2n-1}$ have the volume vol T/τ^{2n} while all the triangles in $\mathfrak{B}_{2n}\cup\mathfrak{B}_{2n+1}$ have the volume vol T/τ^{2n+1}.

IV. TAU-WAVELETS IN THE PLANE

So far we have been subdividing the triangles. Now we define the system of orthonormal step functions on them.

For every triangle Δ which is being divided at the step n, we define the function $f_{\Delta,n}$ as follows.

(a) If $n=2k$, then for each $\Delta\in\mathfrak{U}_{2k}$ put

$$f_{\Delta,2k}(z)=\begin{cases}\tau^{k+1/2}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{U}}(\Delta),\\ -\tau^{k-1/2}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{B}}(\Delta),\\ 0, & \text{otherwise.}\end{cases}$$

(b) If $n=2k-1$, then for each $\Delta\in\mathfrak{B}_{2k-1}$ put

$$f_{\Delta,2k-1}(z)=\begin{cases}-\tau^{k-1}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{U}}(\Delta),\\ \tau^{k}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{B}}(\Delta),\\ 0, & \text{otherwise.}\end{cases}$$

Consider the example of $f_{T,0}$ and $f_{\Psi_{\mathfrak{B}}(T),1}$. These two functions will play the role of two ''mother'' wavelets:

$$f_{\mathfrak{U}}\equiv f_{T,0}(z)=\begin{cases}\tau^{1/2}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{U}}(T),\\ -\tau^{-1/2}(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{B}}(T),\\ 0, & \text{otherwise;}\end{cases}$$

$$f_{\mathfrak{B}}\equiv f_{\Psi_{\mathfrak{B}}(T),1}(z)=\begin{cases}-(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{U}}\Psi_{\mathfrak{B}}(T),\\ \tau(\mathrm{vol}T)^{-1/2}, & \text{for } z\in\Psi_{\mathfrak{B}}\Psi_{\mathfrak{B}}(T),\\ 0, & \text{otherwise.}\end{cases}$$

Let us point out that supp $f_{\Delta,n}=\Delta$, the powers of τ and the signs are chosen so that

$$\int_T f_{\Delta,n}^2(z)\,dz=1,$$

$$\int_T f_{\Delta,n}(z)\,dz=\int_\Delta f_{\Delta,n}(z)\,dz=0.$$

Note that the integral over Δ implies that $f_{\Delta,n}$ is orthogonal to the characteristic function of its support.

We obtained an orthonormal system of functions on T because for any $m\leqslant n$ and for any function $f_{\Delta',m}$ one has

(a) either the supp $f_{\Delta',m}=\Delta'$ and supp $f_{\Delta,n}=\Delta$ are disjoint;

(b) or $f_{\Delta',m}$ is constant on Δ.

We have attached to the triangles in $(\cup_{n=0}^{\infty}\mathfrak{U}_n)\cup(\cup_{n=0}^{\infty}\mathfrak{B}_n)$ step functions. The set of all these functions is denoted by B.

So far we have a system of orthogonal functions on the triangle T but not yet wavelets. For our system we need to find two mother wavelets because the supports are found on tiles of two shapes. We are now ready to show that we are able to generate the whole wavelet system starting from the two mother wavelets $f_{\mathfrak{U}}$ and $f_{\mathfrak{B}}$ by the scale-up function Φ, and two planar displacement-isometries α and β.

The functions Φ, α and β are defined by

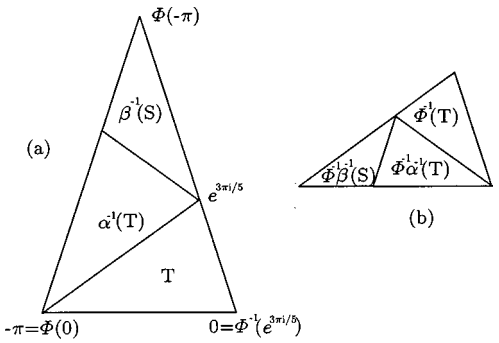

FIG. 6. The action of scaling-up Φ and of displacement isometries α and β. For simplicity of notation we set $S=\Psi_{\mathfrak{B}}(T)$.

$$\Phi(z)=\tau e^{-3\pi i/5}z-\tau, \tag{4.1}$$

$$\alpha(z)=e^{-\pi i/5}(z+\tau)-\tau, \tag{4.2}$$

$$\beta(z)=e^{-3\pi i/5}z-1-\tau. \tag{4.3}$$

For simplicity of notation we introduce $S=\Psi_{\mathfrak{B}}(T)$. That action of the functions α, β and Φ is illustrated in Fig. 6(a). For the scaled-up triangles T and S, we have

$$\begin{aligned}\Phi(T)&=T\cup\alpha^{-1}(T)\cup\beta^{-1}(S),\\ \Phi(S)&=\alpha^{-1}(T)\cup\beta^{-1}(S).\end{aligned} \tag{4.4}$$

Applying Φ^{-1} to (4.4), we get

$$\begin{aligned}T&=\Phi^{-1}T\cup\Phi^{-1}\alpha^{-1}(T)\cup\Phi^{-1}\beta^{-1}(S),\\ S&=\Phi^{-1}\alpha^{-1}(T)\cup\Phi^{-1}\beta^{-1}(S).\end{aligned} \tag{4.5}$$

The transformations are shown in Fig. 6(b).

In order to determine the requirements imposed on the words in α, β and Φ, we perform one more step of the iteration. Substitution of (4.5) into the right side of (4.4) followed by inversion yields

$$\begin{aligned}T=&\Phi^{-1}\Phi^{-1}(T)\cup\Phi^{-1}\Phi^{-1}\alpha^{-1}(T)\cup\Phi^{-1}\Phi^{-1}\beta^{-1}(S)\cup\Phi^{-1}\alpha^{-1}\Phi^{-1}(T)\\ &\cup\Phi^{-1}\alpha^{-1}\Phi^{-1}\alpha^{-1}(T)\cup\Phi^{-1}\alpha^{-1}\Phi^{-1}\beta^{-1}(S)\cup\Phi^{-1}\beta^{-1}\Phi^{-1}\alpha^{-1}(T)\\ &\cup\Phi^{-1}\beta^{-1}\Phi^{-1}\beta^{-1}(S).\end{aligned}$$

The admissibility rules which one sees so far are (i) $\alpha^{-1}\alpha^{-1},\alpha^{-1}\beta^{-1},\beta^{-1}\beta^{-1},\beta^{-1}\alpha^{-1}$ are not admissible, i.e., α^{-1},β^{-1} are always separated by Φ^{-1}; (ii) a word acting on S ends always by β^{-1}; (iii) the combination $\beta^{-1}\Phi^{-1}\Phi^{-1}$ is forbidden because by β^{-1} we map only $T-\Phi^{-1}\Phi^{-1}(T)=S$ and not the whole T.

Application of all the admissible words to T and S generates the system of triangles $\mathfrak{U}_n$ and $\mathfrak{B}_n$, respectively. We have now the rules for constructing the supports of all the step functions starting from the triangles T and S. Therefore we can describe the set of all orthonormal step functions by the two mother wavelets $f_{\mathfrak{U}}$ and $f_{\mathfrak{B}}$,

$$B=\{\tau^{l(w)}f_{\mathfrak{U}}(w(x))\,|\,w\in\mathrm{Admis}_{\mathfrak{U}}\}\cup\{\tau^{l(w)}f_{\mathfrak{B}}(w(x))\,|\,w\in\mathrm{Admis}_{\mathfrak{B}}\},$$

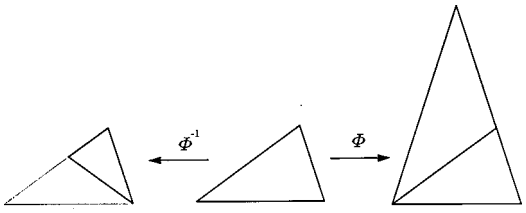

FIG. 7. The action of Φ and Φ^{-1} on the basic triangle T.

where $l(w)$ is the multiplicity of Φ in the word w. The sets $\mathrm{Admis}_{\mathfrak{U}}$, $\mathrm{Admis}_{\mathfrak{B}}$ consist of all words in α, β, and Φ which end by Φ and do not contain the strings $\alpha\alpha$, $\alpha\beta$, $\beta\beta$, $\beta\alpha$ and $\Phi\Phi\beta$. Furthermore the words in $\mathrm{Admis}_{\mathfrak{U}}$ never begin with β, and the words in $\mathrm{Admis}_{\mathfrak{B}}$ always begin with β.

Our next task is to scale-up the orthonormal system on the whole plane $\mathbb{C}$. For that we define

$$W_n=\{f_{\Delta,2n}|\Delta\in\mathfrak{U}_{2n}\}\cup\{f_{\Delta,2n+1}|\Delta\in\mathfrak{B}_{2n+1}\},\quad \text{for } n=0,1,2,\ldots .$$

The idea of the process is to proceed step-by-step each time enlarging a triangle of type $\mathfrak{U}$ to the next larger triangle of type $\mathfrak{U}$ by adjoining to it an appropriate triangle of type $\mathfrak{B}$. We want to exploit the property of Φ,

$$\forall z\in\mathbb{C},\quad \exists k\in\mathbb{N},\quad z\in\Phi^k(T),$$

which implies that

$$\mathbb{C}=\bigcup_{k\in N}\Phi^k(T)$$

On Fig. 7 we illustrate the action of Φ and Φ^{-1} on the triangle T. The self-similarity of the triangles (and functions as well) is guaranteed by the property $\Phi^{-1}(T)=\Psi_{\mathfrak{U}}(T)$. Indeed, we have

$$\Phi^{-1}(\mathfrak{U}_n)=\{\Delta\in\mathfrak{U}_{n+2}|\Delta\in\Psi_{\mathfrak{U}}(T)\},$$

$$\Phi^{-1}(\mathfrak{B}_n)=\{\Delta\in\mathfrak{B}_{n+2}|\Delta\in\Psi_{\mathfrak{U}}(T)\}.$$

For all $n\in\mathbb{N}$ and for all functions $f\in W_{n-1}$ one has

$$f(z)\in W_{n-1}\Rightarrow\tau f(\Phi(z))\in W_n . \tag{4.6}$$

In order to illustrate the scale-up process of the functions, let us first carry out the enlargement by one step: from T to $\Phi(T)$.

Putting

$$W_n^{(-1)}=\left\{\frac{1}{\tau}\,f_{\Delta,n}(\Phi^{-1}(z))\,\middle|\, f_{\Delta,n}\in W_n\right\},$$

we have taken an orthonormal system W_n of functions and by a linear mapping Φ we have stretched it. The scale factor of Φ is τ, therefore we had to multiply the result by $1/\tau$ in order to preserve the normalization. The old function set is contained in the new one, $W_n\subset W_{n+1}^{(-1)}$, as follows from (4.6). The system $W_n^{(-1)}$, $0\leqslant n<\infty$, is orthonormal on $\Phi(T)$. The orthonormal system in the plane $\mathbb{C}$ is defined by

$$\widetilde{W}_n=\bigcup_{k=0}^{\infty}W_{n+k}^{(-k)},\quad n\in\{0,1,\ldots\}, \tag{4.7}$$

where

$$W_n^{(-k)}=\left\{\frac{1}{\tau^k}\,f_{\Delta,n}(\Phi^{-k}(z))\,\middle|\,f_{\Delta,n}\in W_n\right\}.$$

The inversion of the implication in (4.6) was not possible until now because the support of a function in W_n could have been mapped out of T. It now becomes possible because we have the supports in the whole plane.

Our system (4.7) now has the multiresolution property

$$f(z)\in\tilde{W}_n\Leftrightarrow\tau f(\Phi(z))\in\tilde{W}_{n+1},\quad\text{for all } n=0,1,\ldots. \tag{4.8}$$

The relation (4.8) can be used now to define $\tilde{W}_n$ for negative values of the index n. Then (4.8) applies to all $n\in\mathbb{Z}$.

In this paper we have disregarded the standard matching rules of the tiling based on the algebraic coloring of the vertices. Their observation would lead to some technical complications but would not prevent the construction.

V. MULTIRESOLUTION ANALYSIS

Recall that the multiresolution analysis in $\mathbb{R}^2$ based on wavelets[23] is a decomposition of $L^2(\mathbb{R}^2)$ into a chain of subspaces,

$$\cdots\subset V_2\subset V_1\subset V_0\subset V_{-1}\subset V_{-2}\subset\cdots,\quad \cap_{j\in\mathbb{Z}}V_j=\{0\}, \tag{5.1}$$

such that $\cup_{j\in\mathbb{Z}}V_j$ is dense in $L^2(\mathbb{R}^2)$. In addition for each $f\in L^2(\mathbb{R}^2)$ one has

$$f(x)\in V_j\Leftrightarrow f(2x)\in V_{j+1}. \tag{5.2}$$

Furthermore there exists a mother function $f_0\in V_0$ such that the functions $2^j f_0(2^j x-k)$ form an orthonormal basis of V_0. It means that one can generate an orthonormal basis of V_0 by a scale-up function $\Phi(x_1,x_2)=(2x_1,2x_2)$ and by two displacement functions $\alpha(x_1,x_2)=(x_1-1,x_2)$ and $\beta(x_1,x_2)=(x_1,x_2-1)$.

Let us compare that with the properties of our system. Let us introduce the notation V_j for the linear span

$$V_j:=\left\{\bigcup_{k=j}^{\infty}\tilde{W}_k\right\}_{\text{linear span}},\quad \forall j\in\mathbb{Z}.$$

Our V_j's satisfy (5.1). The property (5.2) can be rewritten for our subspaces in the form

$$f(x)\in V_j\Leftrightarrow f(\Phi(x))\in V_{j+1},$$

where $\Phi(x)$ is defined in (4.1). We have found two mother functions $f_{\mathfrak{U}}$ and $f_{\mathfrak{B}}$ in V_0 such that we can generate orthonormal basis of V_0 by Φ, α and β from (4.1).

It remains to show that our system is dense in $L^2(\mathbb{R}^2)$. For that we define the linear span of the characteristic functions of all supports in our system,

$$S=\left\{\chi\Delta\,\middle|\,\exists f\in\bigcup_{n\in\mathbb{Z}}V_n,\quad\text{such that }\Delta=\sup f\right\}_{\text{linear span}}.$$

As in the one-dimensional case like (2.18) an expansion formula allows one to express the characteristic function of T in terms of the affine transforms of the mother wavelets $f_{\mathfrak{U}}$ and $f_{\mathfrak{B}}$. Then according to the Stone–Weierstrass theorem, every continuous function in $L^2(\mathbb{R}^2)$ can be approximated with any precision by a linear combination of functions from our system. Since continuous functions are dense in $L^2(\mathbb{R}^2)$, it follows that the closure of the linear span of our system equals $L^2(\mathbb{R}^2)$.

ACKNOWLEDGMENTS

The work of Jiří Patera was supported in part by the Natural Sciences and Engineering Research Council of Canada and by the Fonds FCAR of Quebec. The work of Edita Pelantová was supported partially by the Grant Agency of the Czech Republic.

[1] J. P. Gazeau and J. Patera, *Tau-Wavelet Analysis For Five-fold Quasicrystals, Proceedings of the 5th International Conference on Quasicrystals, Avignon, France, May 1995*, edited by C. Janot and R. Mosseri (World Scientific, Singapore, 1995), pp. 31–34.
[2] J. P. Gazeau and J. Patera, ''Tau wavelets of Haar,'' J. Phys. A **29**, 4549–4559 (1996).
[3] P. Wojtaszczyk, *A Mathematical Introduction to Wavelets* (Cambridge University Press, Cambridge, 1997).
[4] A. Rényi, ''Representations for real numbers and their ergodic properties,'' Acta Math. Acad. Sci. Hung. **8**, 477–493 (1957).
[5] Y. Meyer, *Diophantine Approximation and Algebraic Numbers, Proc. Les Houches, March 1994, Beyond Quasicrystals*, Les Editions de Physique, edited by F. Axel and D. Gratias (Springer-Verlag, Berlin, 1995), pp. 3–16.
[6] Č. Burdík, C. Frougny, J.-P. Gazeau, and R. Krejcar, ''β-integers as natural counting systems for quasicrystals,'' to appear in J. Phys. A (1998).
[7] Y. Meyer, *Nombres de Pisot, Nombres de Salem et Analyse Harmonique*, Lecture Notes in Mathematics 117 (Springer-Verlag, Berlin, 1970).
[8] Y. Meyer, *Algebraic Numbers and Harmonic Analysis*, Lecture Notes in Mathematics (North Holland, New York, 1972).
[9] J. P. Gazeau and V. Spiridonov, ''Toward discrete wavelets with irrational scaling factor,'' J. Math. Phys. **37**, 3001–3013 (1996).
[10] C. Janot, *Quasicrystals: A Primer* (Oxford University Press, Oxford, 1994).
[11] C. Janot and R. Mosseri, in *Quasicrystals* (Ref. 1).
[12] F. Axel and D. Gratias, in *Beyond Quasicrystals* (Ref. 5), pp. 3–16.
[13] R. V. Moody, in *Mathematics of Long Range Aperiodic Order, Proceedings of NATO ASI, Waterloo, 1995* (Kluwer Academic, New York, 1997).
[14] D. Schechtman, I. Blech, D. Gratias, and J. W. Cahn, Phys. Rev. Lett. **53**, 1951 (1984).
[15] Z. Masáková, J. Patera, and E. Pelantová, ''Inflation centers of the cut and project quasicrystals,'' J. Phys. A **31**, 1443–1453 (1998).
[16] R. V. Moody, ''Meyer sets and their duals,'' in Ref. 13.
[17] L. Chen, R. V. Moody, and J. Patera, ''Noncrystallographic root systems,'' in *Quasicrystals and Discrete Geometry*, Fields Institute Monograph Series, edited by J. Patera (American Mathematical Society, Providence, RI, 1998).
[18] J. Patera, ''Noncrystallographic root systems and quasicrystals,'' in Ref. 13.
[19] J. Bellissard, ''Introduction to the special issue on quantum problems in condensed matter physics,'' J. Math. Phys. **38**, 1753–1756 (1997).
[20] D. Barache, Doctoral dissertation, Université Paris 7—Denis Diderot, 1995.
[21] J. P. Gazeau, ''Quasicrystals and their symmetries,'' in *Symmetries and Structural Properties of Condensed Matter, Zajaczkowo 1994*, edited by T. Lulek, W. Florek, and S. Walcerz (World Scientific, Singapore, 1995).
[22] Z. Masáková, J. Patera, and E. Pelantová, ''Minimal distances in quasicrystals,'' J. Phys. A **31**, 1539–1552 (1998).
[23] I. Daubechies, ''Wavelet transforms and orthonormal wavelet bases,'' in *Different Perspectives on Wavelets, Proceedings of the Symposium of Applied Mathematics*, edited by I. Daubechies (American Mathematical Society, Providence, RI, 1993).

Wavelets bases adapted to a self-similar quasicrystal

Guillaume Bernuau[a)]
Ceremade, Université Paris IX Dauphine, Paris, France

(Received 15 January 1998; accepted for publication 3 April 1998)

Given any self-similar quasicrystal Λ in $\mathbb{R}^n$ with inflation $\theta>1$, we construct bases of $\mathbb{L}^2(\mathbb{R}^n)$ having the following structure: $\theta^{nj/2}\psi_\lambda(\theta^j x-\lambda)$, $\lambda\in\Lambda\backslash\theta\Lambda$, $j\in\mathbb{Z}$, where the mother wavelets ψ_λ, $\lambda\in\Lambda\backslash\theta\Lambda$, are smooth and with exponential decay or compact support. We also show that wavelets ψ_λ constitute a relatively compact set in some Sobolev space and that they depend continuously on λ when Λ is equipped with an appropriate topology. © *1998 American Institute of Physics.* [S0022-2488(98)01808-8]

I. INTRODUCTION

A. Physical motivations

Wavelets theory finds interesting applications in problems involving some self-similarity or fractal properties. Thus it may appear very useful in the study of those new physical structures named quasicrystals, which were first discovered in 1982.[1] Indeed quasicrystals are long range ordered physical structures exhibiting symmetries which are incompatible with translational invariance and their existence proves that order is not synonymous with periodicity. In particular, many observed quasicrystals revealed a scale invariance involving an irrational scaling factor. Precisely, in quasicrystal studies, adapted wavelets could be valuable in the analysis of the diffraction pattern of self-similar quasicrystals, in the analysis of quasicrystals surfaces by tunneling or atomic face microscopy, and in the study of the Schrödinger equation with quasiperiodic potential, appearing in the field of electronic transport in quasicrystals.

B. Mathematical motivations

Mathematical quasicrystals are discrete aperiodic point sets which enlarge lattices from the harmonic analysis point of view (see Refs. 2–5 for an overview). Self-similarity may exist for those sets and usually the involved scaling factor is an irrational number $\theta>1$.[2] Thus a self-similar mathematical quasicrystal Λ gives rise to an increasing sequence of discrete aperiodic grids $\Lambda_j=\theta^{-j}\Lambda$, $j\in\mathbb{Z}$, which approximate the whole space $\mathbb{R}^n$,

$$\cdots\subset\Lambda\subset\theta^{-1}\Lambda\subset\cdots\subset\theta^{-j}\Lambda\subset\cdots\subset\mathbb{R}^n.$$

Wavelets construction usually relies on the approximation of the whole space $\mathbb{R}^n$ by an increasing sequence of discrete grids $(\Gamma_j)_{j\in\mathbb{Z}}$. The main idea consists in introducing a sequence of embedded subspaces V_j of $\mathbb{L}^2(\mathbb{R}^n)$ which mimics the approximation $\Gamma_j\approx\mathbb{R}^n$. Usually one has no reason to get rid of translation invariance and hence chooses dyadic grids $\Gamma_j=2^{-j}\mathbb{Z}^n$. Meanwhile many problems, especially in the study of quasicrystalline structures, involve aperiodicity and irrational self-similarity. Therefore it is natural to try to construct wavelet bases where the grids $\Lambda_j=\theta^{-j}\Lambda$ play a key role and replace the usual dyadic grids $2^{-j}\mathbb{Z}^n$.

Another important (and purely mathematical) reason motivating our work is the construction of wavelets bases of $\mathbb{L}^2(\mathbb{R}^n)$ having the following structure:

$$\theta^{nj/2}\psi_\gamma(\theta^j x-\gamma),\quad j\in\mathbb{Z},\quad \gamma\in\Gamma, \tag{1}$$

where $\theta>1$ is some irrational factor, Γ is a denumerable subset of $\mathbb{R}^n$, and the set of mother wavelets $\{\psi_\gamma,\gamma\in\Gamma\}$ should have the lowest cardinality possible.

a)Electronic mail: Gbernuau@aol.com and bernuau@ceremade.dauphine.fr

 © 1998 American Institute of Physics

C. Related papers and objectives

The construction of wavelets associated with aperiodic grids or enjoying nonintegral scaling factor has been widely studied (see Refs. 6–9) but in these constructions either the involved self-similarity is a rational number (Refs. 9 and 7) or there is simply no self-similarity (Refs. 6 and 8). The first example, to our knowledge, of wavelets bases having the structure given in (1), with an irrational scaling factor θ, was obtained by Gazeau and Patera.[10] They prove the existence of an orthonormal Haar basis of $\mathbb{L}^2(\mathbb{R})$ having the following structure:

$$\tau^{j/2}h(\tau^j x-\lambda),\quad j\in\mathbb{Z},\quad \lambda\in\Lambda_\tau,$$

where h is compactly supported and piecewise constant, τ is the golden number $(1+\sqrt{5})/2$, and Λ_τ is a mathematical quasicrystal in $\mathbb{R}$. Then Gazeau and Spiridonov[11] gave two other explicit examples of a Haar basis of $\mathbb{L}^2(\mathbb{R})$ where the value of the scaling factor is $1+\sqrt{2}$ or $2+\sqrt{3}$. Recently, Gazeau, Patera, and Pelantova[12] have constructed a Haar basis of $\mathbb{L}^2(\mathbb{R}^2)$ adapted to Penrose–Robinson tiling and with scaling factor $(1+\sqrt{5})/2$.

In this paper, we present two constructions of wavelets bases of $\mathbb{L}^2(\mathbb{R}^n)$ adapted to any prescribed self-similar mathematical quasicrystal Λ in $\mathbb{R}^n$ with inflation $\theta>1$. By adapted wavelets bases, we mean Riesz bases of $\mathbb{L}^2(\mathbb{R}^n)$ having the following structure:

$$\theta^{nj/2}\psi_\lambda(\theta^j x-\lambda),\quad \lambda\in\Lambda\backslash\theta\Lambda,\quad j\in\mathbb{Z},$$

where the set of mother wavelets $\mathcal{M}=\{\psi_\lambda,\lambda\in\Lambda\backslash\theta\Lambda\}$ consists of arbitrarily regular functions with exponential decay in the multidimensional case and with compact support in the one-dimensional case. These constructions are based on the results of Lemarié-Rieusset[7] for the multidimensional case and may be connected, in the one-dimensional case, with the results of Michelli and Buhmann.[8]

We also focus on the study of the set $\mathcal{M}$. If Λ is a lattice, $\mathcal{M}$ is made of one single wavelet. If Λ is no longer a lattice, we show that $\mathcal{M}$ becomes a relatively compact set in some Sobolev space. Indeed we show that Λ can be equipped with a topology which makes it precompact and which makes the mapping $\lambda\in\Lambda\backslash\theta\Lambda\to\psi_\lambda\in H$ continuous (H is some Sobolev space). In one dimension we describe a construction of compactly supported wavelets bases and obtain a finite set $\mathcal{M}$ of compactly supported mother wavelets.

In Sec. II, we recall some definitions about Delaunay sets, quasicrystals, and algebraic numbers. In Sec. III, the multiresolution scheme into which wavelets will be derived is introduced and studied. In Sec. IV, wavelets with exponential decay are constructed. In Sec. V, the set $\mathcal{M}$ is studied. In Sec. VI, we restrict our attention to the one-dimensional case and construct compactly supported wavelets.

This work is part of a Ph.D. thesis[5] where we also construct orthonormal Haar bases adapted to any self-similar tiling of $\mathbb{R}^n$ and frames adapted to quasicrystals.

II. DELAUNAY SETS, QUASICRYSTALS, AND ALGEBRAIC NUMBERS

Let us recall some definitions and properties about self-similar aperiodic sets. A Delaunay set is a subset Λ of $\mathbb{R}^n$ such that there exist $0<R_1<R_2$ satisfying the following conditions:

(i) Λ is uniformly discrete: any ball of radius R_1 contains at most one point of Λ.
(ii) Λ is relatively dense: any ball of radius R_2 contains at least one point of Λ.

Λ is said to be self-similar if there exists $\theta>1$ such that $\theta\Lambda\subset\Lambda$. We are interested here in the study of very special Delaunay sets: quasicrystals. These special sets may be considered as mathematical models for physical quasicrystals (see Refs. 1, 4, 2, 5, 3). Here is their definition:

Definition 2.1: A Delaunay set Λ is a quasicrystal (in the mathematical sense) if there exists a finite set $F\subset\mathbb{R}^n$ such that $\Lambda-\Lambda\subset\Lambda+F$.

The terminology concerning mathematical quasicrystals is not unified yet. One can find a quasilattice, Meyer set, model, or simply a quasicrystal. In the following, ''quasicrystal'' will always denote the mathematical concept defined above. There are many equivalent definitions of

a quasicrystal including the famous cut and project method construction (see Ref. 5 for an overview). In this article, only two fundamental properties of quasicrystals will be needed. The first one deals with the local configurations of a quasicrystal:

Property 2.1: Let Λ *be a quasicrystal in* $\mathbb{R}^n$. *For any* $R>0$ *the local configuration of size R, around* $\lambda \in \Lambda$, *defined by*

$$C_{\lambda,R}=(\Lambda-\lambda)\cap B_{0,R}$$

describes a finite set as λ *goes through* Λ.

Here $B_{0,R}$ denotes the ball centered at 0 with radius *R*.

Proof: $C_{\lambda,R}\subset B_{0,R}\cap(\Lambda-\Lambda)\subset B_{0,R}\cap(\Lambda+F)$. Since F is a finite set and since Λ is a Delaunay set, $B_{0,R}\cap(\Lambda+F)$ is finite. □

The second property deals with self-similarity. We first have to recall the definitions of two classes of algebraic integers (see Ref. 13 for further details):

Definition 2.2: A Pisot number is an algebraic integer $\theta>1$ *whose algebraic conjugates lie in the complex open disk of radius* 1.

Definition 2.3: A Salem number is an algebraic integer $\theta>1$ *whose algebraic conjugates lie in the complex closed disk of radius* 1 *with at least one of them having modulus equal to* 1.

The fundamental result concerning quasicrystals' self-similarity is

Theorem 2.1 (Meyer): *Let* Λ *be a self-similar quasicrystal with inflation* $\theta>1$. *Then* θ *is either a Pisot number or a Salem number.*

One can refer to Ref. 2 for a proof.

Let us give some examples of self-similar quasicrystals:

Example 2.1: Let $\tau=(1+\sqrt{5})/2$ *be the golden number. Then the set*

$$\Lambda=\{p+q\tau,(p,q)\in\mathbb{Z}^2,|p-q/\tau|\leqslant 1\}$$

is a self-similar quasicrystal in $\mathbb{R}$ *with inflation* τ.

Proof:

(i) τ is a Pisot number since
$$X^2-X-1=(X-\tau)(X+\tau^{-1}).$$

(ii) One way to check that Λ is a quasicrystal is to use Lagarias's characterization.[14] A Delaunay set Λ is a quasicrystal if and only if $\Lambda-\Lambda$ is uniformly discrete. In our example, $\Lambda\cap\mathbb{R}_+$ is relatively dense in $\mathbb{R}_+$ because, if $\lambda\in\Lambda$, either $\lambda+1$ or $\lambda+\tau$ is also in Λ. Consequently Λ is relatively dense in $\mathbb{R}$ because it is symmetric about 0. Λ is uniformly discrete because the following conditions:
$$\begin{cases}|p+q\tau|\leqslant 1,\\ |p-q/\tau|\leqslant 2,\end{cases}$$
determine a finite set of integral solutions (p,q). Hence 0 is isolated in $\Lambda-\Lambda$. The same argument holds for $\Lambda-\Lambda$ replacing 2 by 4 in the second condition. Finally Λ and $\Lambda-\Lambda$ are both Delaunay sets: Λ is a quasicrystal.

(iii) $\tau\Lambda$ is a subset of Λ because if $\tau(p+q\theta)=p'+q'\tau$ then
$$|p'-q'\tau^{-1}|=|-\tau^{-1}(p-q\tau^{-1})|\leqslant 1.$$ □

Example 2.2: Let $\theta=(1+\sqrt{2}+\sqrt{2\sqrt{2}-1})/2$ *and* $\alpha=\arccos((1-\sqrt{2})/2)$. *Denote by* σ *and* σ' *the linear mappings defined by*

$$\sigma(p+q\theta+r\theta^2+s\theta^3)=p+q\theta^{-1}+r\theta^{-2}+s\theta^{-3},$$

$$\sigma'(p+q\theta+r\theta^2+s\theta^3)=p+qe^{i\alpha}+re^{2i\alpha}+se^{3i\alpha},$$

where $(p,q,r,s)\in\mathbb{Z}^4$. *Then the set*

$$\Lambda=\{(a,b)\in\mathbb{Z}[\theta]^2,|\sigma(a)|+|\sigma(b)|+|\sigma'(a)|+|\sigma'(b)|\leqslant 1\}$$

is a self-similar quasicrystal in $\mathbb{R}^2$ *with inflation* θ.

Proof:

(i) θ is a Salem number since

$$X^4-2X^3+X^2-2X+1=(X-\theta)(X-\theta^{-1})(X-e^{i\alpha})(X-e^{-i\alpha}).$$

(ii) One can show, using similar arguments as in the first example, that Λ and $\Lambda-\Lambda$ are both Delaunay sets in $\mathbb{R}^2$. Hence Λ is a quasicrystal in $\mathbb{R}^2$.

(iii) Let $\lambda=(a,b)$ in Λ. Then $\theta\lambda=(a',b')$ is in $\mathbb{Z}[\theta]^2$ and $\sigma(a')=\theta^{-1}\sigma(a)$, $\sigma'(a')=e^{i\alpha}\sigma'(a)$. The same relations hold between b and b'. Hence $|\sigma(a')|+|\sigma(b')|+|\sigma'(a')|+|\sigma'(b')|\leqslant 1$. Consequently $\theta\Lambda\subset\Lambda$. □

III. MULTIRESOLUTION ANALYSIS ADAPTED TO A SELF-SIMILAR QUASICRYSTAL

To achieve the construction of wavelets bases, one needs to introduce a multiresolution analysis (Ref. 15). In its most general form, it consists of an increasing sequence $(V_j)_{j\in\mathbb{Z}}$ of closed subspaces whose union is dense in $\mathbb{L}^2(\mathbb{R}^n)$ and whose intersection reduces to $\{0\}$. Moreover if, for some absolute constant $\theta>1$, V_j is linked to V_0 by a change of scale equal to θ^j, i.e.,

$$f(x)\in V_j \Leftrightarrow f(\theta^{-j})\in V_0,$$

then wavelets bases derived from the sequence $(V_j)_{j\in\mathbb{Z}}$ will have the desired following structure:

$$\theta^{nj/2}\psi_\gamma(\theta^j x-\gamma),\quad j\in\mathbb{Z},\quad \gamma\in\Gamma.$$

Therefore, given any self-similar quasicrystal Λ in $\mathbb{R}^n$ with inflation $\theta>1$, we wish to construct a multiresolution analysis where the inflation θ and the translations lying in Λ will play a key role. A possible answer to our requirement is implicitly given in Ref. 7 and we shall present it now.

Let s be an integer strictly greater than $n/2$. $H^s(\mathbb{R}^n)$ will be the Sobolev space of functions f in $\mathbb{L}^2(\mathbb{R}^n)$ such that each distribution $\partial^\alpha f$, with $|\alpha|\leqslant s$, is an element of $\mathbb{L}^2(\mathbb{R}^n)$. Let us denote by $\mathcal{V}_0^{(s)}(\Lambda)$ (or more simply $\mathcal{V}_0^{(s)}$) the subspace of $\mathbb{L}^2(\mathbb{R}^n)$ defined by

$$\mathcal{V}_0^{(s)}(\Lambda)=\left\{ f\in\mathbb{L}^2(\mathbb{R}^n),\quad \exists(a_\lambda)_{\lambda\in\Lambda},\quad \Delta^s f=\sum_{\lambda\in\Lambda} a_\lambda\delta_\lambda\right\}.$$

$\mathcal{V}_0^{(s)}$ is the space of splines of order s with knots in Λ. Then one makes a change of scale to define the subspace $\mathcal{V}_j^{(s)}$, $j\in\mathbb{Z}$:

$$\mathcal{V}_j^{(s)}=\left\{ f\in\mathbb{L}^2(\mathbb{R}^n),\quad \exists(a_\lambda)_{\lambda\in\Lambda},\quad \Delta^s f=\sum_{\lambda\in\Lambda} a_\lambda\delta_{\theta^{-j}\lambda}\right\}.$$

We are ready to introduce a multiresolution analysis adapted to Λ:

Property 3.1: The sequence of subspaces $(\mathcal{V}_j^{(s)})_{j\in\mathbb{Z}}$ is a multiresolution analysis of $\mathbb{L}^2(\mathbb{R}^n)$ i.e.,

(*i*) *for any $j\in\mathbb{Z}$, $\mathcal{V}_j^{(s)}$ is a closed subspace of $\mathbb{L}^2(\mathbb{R}^n)$,*
(*ii*) *for any $j\in\mathbb{Z}$, $\mathcal{V}_j^{(s)}\subset\mathcal{V}_{j+1}^{(s)}$,*
(*iii*) *the union of $\mathcal{V}_j^{(s)}$, $j\in\mathbb{Z}$, is dense in $\mathbb{L}^2(\mathbb{R}^n)$,*
(*iv*) *the intersection of $\mathcal{V}_j^{(s)}$, $j\in\mathbb{Z}$, reduces to $\{0\}$,*
(*v*) *$f\in\mathcal{V}_j^{(s)}$ if and only if $f(\theta^{-j}x)\in\mathcal{V}_0^{(s)}$.*

Proof:

(*i*) Let $(f_j)_{j\in\mathbb{Z}}$ be a convergent sequence in $\mathbb{L}^2(\mathbb{R}^n)$ with values in $\mathcal{V}_0^{(s)}$. Let f denote its limit. Then $\Delta^s f_j$ converges to $\Delta^s f$ in $\mathcal{D}(\mathbb{R}^n)$ the space of distributions on $\mathbb{R}^n$. Write $\Delta^s f_j=\Sigma_{\lambda\in\Lambda} a_\lambda^j\delta_\lambda$. Choosing $\varphi\in C_0^\infty(\mathbb{R}^n)$, $\varphi(0)=1$, with support sufficiently small around 0, we get $\lim_{j\to\infty} a_\lambda^j=\langle\Delta^s f,\varphi(x-\lambda)\rangle=a_\lambda$. Finally for any $\varphi\in C_0^\infty(\mathbb{R}^n)$, one has $\langle\Delta^s f,\varphi\rangle=\Sigma_{\lambda\in\Lambda} a_\lambda\varphi(\lambda)$.

(*ii*) Clear since $\theta\Lambda\subset\Lambda$.

(*iii*) We use the fundamental property 4.1 mentioned below: if $f \in (\cup_{j\in\mathbb{Z}} \mathcal{V}_j^{(s)})^{\perp} = \cap_{j\in\mathbb{Z}} \mathcal{V}_j^{(s)^{\perp}}$ then $f = \Delta^s h$ where $h \in H^{2s}(\mathbb{R}^n)$ vanishes on $\cup_j \theta^{-j}\Lambda$ which is dense in $\mathbb{R}^n$. Hence h, which is continuous, vanishes on the whole space and so does f.

(*iv*) If $f \in \cap_j \mathcal{V}_j^{(s)}$ then $\Delta^s f$ is a sum of Dirac masses located on $\cap_j \theta^{-j}\Lambda = \{0\}$. Then $\Delta^s f = \alpha\delta_0$ and, since $f \in \mathbb{L}^2(\mathbb{R}^n)$, $f = 0$.

(*v*) Straightforward using Fourier transform. □

The sequence $(\mathcal{V}_j^{(s)})_{j\in\mathbb{Z}}$ is said to be adapted to Λ. We now study the subspace $\mathcal{V}_0^{(s)}$.

A. Variational study of $\mathcal{V}_0^{(s)}$

All the results of this subsection can be found in Ref. 7. Some technical proofs will not be detailed and the reader may refer to the original work of Lemarié-Rieusset.

We are concerned with the problem of constructing a Riesz basis of $\mathcal{V}_0^{(s)}$, i.e., a family of vectors of $\mathcal{V}_0^{(s)}$, $\{g_\lambda, \lambda \in \Lambda\}$, whose linear combinations are dense in $\mathcal{V}_0^{(s)}$ and for which there exist $0 < c_1 < c_2$ such that for any real sequence $(a_\lambda)_{\lambda\in\Lambda}$ with finite support:

$$c_1 \sum_{\lambda\in\Lambda} |a_\lambda|^2 \leqslant \left\| \sum_{\lambda\in\Lambda} a_\lambda g_\lambda(x) \right\|_{\mathbb{L}^2}^2 \leqslant c_2 \sum_{\lambda\in\Lambda} |a_\lambda|^2.$$

In order to solve this problem one considers the following minimization problem:

> Given $F \in H^s(\mathbb{R}^n)$, find, among functions of $H^s(\mathbb{R}^n)$ equal to F on Λ, a function f which minimizes the functional:

$$J(f) = \int_{\mathbb{R}^n} |\Delta^{s/2} f|^2 dx.$$

f is a solution of this problem if and only if f minimizes, among the same set of functions, the following functional:

$$\tilde{J}(f) = J(f) + \sum_{\lambda\in\Lambda} |f(\lambda)|^2.$$

This functional is coercive on $H^s(\mathbb{R}^n)$ since we have

Property 3.2: Let n be a positive integer and $s > n/2$ a real number. Then, on Sobolev space $H^s(\mathbb{R}^n)$, $\|f\|_{H^s}$ and $|||f|||_s = (\Sigma_{\lambda\in\Lambda} |f(\lambda)|^2)^{1/2} + \|\Delta^{s/2} f\|_{\mathbb{L}^2}$ are equivalent norms.

One can refer to Ref. 7 for a proof of this property.

Consequently the initial minimization problem has a unique solution f. Then for any $h \in C_0^\infty(\mathbb{R}^n)$, vanishing on Λ, for any $\alpha \in \mathbb{R}$, one has

$$J(f) \leqslant J(f + \alpha h),$$

which leads to

$$\int_{\mathbb{R}^n} \Delta^{s/2} f \Delta^{s/2} h \, dx = 0 \tag{2}$$

and hence

$$\Delta^s f = \sum_{\lambda\in\Lambda} a_\lambda \delta_\lambda$$

for a sequence $(a_\lambda)_{\lambda\in\Lambda}$. Therefore f is in $\mathcal{V}_0^{(s)}$.

Conversely if $f \in \mathcal{V}_0^{(s)}$, f is the unique solution of the minimization problem associated with $F = f$. Hence $\mathcal{V}_0^{(s)}$ is exactly the set of all solutions obtained as F goes through $H^s(\mathbb{R}^n)$.

B. Lagrangian Riesz basis of $\mathcal{V}_0^{(s)}$

On the subspace $\mathcal{V}_0^{(s)}$, one has

$$\|f\|_{H^s} \sim \left(\sum_{\lambda \in \Lambda} |f(\lambda)|^2 \right)^{1/2}.$$

Indeed using property 3.2 we find

$$\|f\|_{H^s} \leqslant C \left(\left(\sum_{\lambda \in \Lambda} |f(\lambda)|^2 \right)^{1/2} + \|\Delta^{s/2} f\|_{\mathbb{L}^2} \right),$$

$$\left(\sum_{\lambda \in \Lambda} |f(\lambda)|^2 \right)^{1/2} \leqslant C' \|f\|_{H^s}.$$

Conversely let $K_s = \{h \in H^s(\mathbb{R}^n), \forall \lambda \in \Lambda, h(\lambda) = 0\}$. Using (2), $\Delta^{s/2} f$ is orthogonal to $\Delta^{s/2} K_s$. In particular, $\Delta^{s/2} f$ is orthogonal to $\Delta^{s/2}(F - f)$ for any function F in $H^s(\mathbb{R}^n)$ equal to f on Λ. Therefore the orthogonal projection of $\Delta^{s/2} F$ on $\Delta^{s/2} K_s$ is $\Delta^{s/2}(F - f)$ and one has $\|\Delta^{s/2} f\|_{\mathbb{L}^2} \leqslant \|\Delta^{s/2} F\|_{\mathbb{L}^2}$. One can choose $F(x) = \Sigma_{\lambda \in \Lambda} f(\lambda) \varphi(x - \lambda)$ where all $\varphi(x - \lambda)$, $\lambda \in \Lambda$, have disjoined supports. Hence

$$\|\Delta^{s/2} f\|_{\mathbb{L}^2} \leqslant C \left(\sum_{\lambda \in \Lambda} |f(\lambda)|^2 \right)^{1/2}.$$

We are now ready for the following result:

Theorem 3.1: *For any integer s greater than n/2, each element $f \in \mathcal{V}_0^{(s)}$ can be written, in $H^s(\mathbb{R}^n)$:*

$$f(x) = \sum_{\lambda \in \Lambda} f(\lambda) \mathcal{L}_\lambda^{(s)}(x),$$

where $\mathcal{L}_\lambda^{(s)}$ is the unique element of $\mathcal{V}_0^{(s)}$ equal to 1 at $\lambda \in \Lambda$ and vanishing on $\Lambda \backslash \{\lambda\}$. Moreover $\{\mathcal{L}_\lambda^{(s)}, \lambda \in \Lambda\}$ is a Riesz basis of $\mathcal{V}_0^{(s)}$.

This basis is called a Lagrangian interpolation basis and $\mathcal{L}_\lambda^{(s)}$ is called Lagrangian spline of order s associated with λ.

C. Regularity of functions in $\mathcal{V}_0^{(s)}$

Let f be in $\mathcal{V}_0^{(s)}$. f is in $\mathbb{L}^2(\mathbb{R}^n)$ and hence $\Delta^s f$ is in $H^{-2s}(\mathbb{R}^n)$. We also know that $\Delta^s f = \Sigma_{\lambda \in \Lambda} a_\lambda \delta_\lambda$ for some sequence $(a_\lambda)_{\lambda \in \Lambda}$. Hence, for any φ in $C_0^\infty(\mathbb{R}^n)$

$$\left| \sum_{\lambda \in \Lambda} a_\lambda \varphi(\lambda) \right| \leqslant \|\Delta^s f\|_{H^{-2s}} \|\varphi\|_{H^{2s}}.$$

Let $\Lambda_R = \Lambda \cap B_{0,R}$. Choosing $\varphi_R = \Sigma_{\lambda \in \Lambda_R} a_\lambda \theta(x - \lambda)$ where all $\theta(x - \lambda)$, $\lambda \in \Lambda$, have disjoined supports and where $\theta(0) = 1$, we get

$$\sum_{\lambda \in \Lambda_R} |a_\lambda|^2 \leqslant \|\Delta^s f\|_{H^{-2s}} C_s \left(\sum_{\lambda \in \Lambda_R} |a_\lambda|^2 \right)^{1/2}.$$

Hence $(a_\lambda)_{\lambda \in \Lambda}$ is in $\ell^2(\Lambda)$. But then, for any $\sigma > n/2$,

$$|\langle \Delta^s f, \varphi \rangle| = \left| \sum_{\lambda \in \Lambda} a_\lambda \varphi(\lambda) \right| \leqslant \left(\sum_{\lambda \in \Lambda} |a_\lambda|^2 \right)^{1/2} \left(\sum_{\lambda \in \Lambda} |\varphi(\lambda)|^2 \right)^{1/2} \leqslant \left(\sum_{\lambda \in \Lambda} |a_\lambda|^2 \right)^{1/2} C_\sigma \|\varphi\|_{H^\sigma}.$$

Therefore $\mathcal{V}_0^{(s)}$ is included in $H^{2s-\sigma}(\mathbb{R}^n)$ for any $\sigma > n/2$ which shows that elements of $\mathcal{V}_0^{(s)}$ are regular functions. Let us end this section with a fundamental result of Lemarié-Rieusset concerning the decay of Lagrangian splines $\mathcal{L}_\lambda^{(s)}$:

Theorem 3.2 (Lemarié-Rieusset): *Let Λ be a Delaunay set in $\mathbb{R}^n$. Then for any integer s greater than $n/2$, there exist two constants $A=A(\Lambda,s)>0$ and $\beta=\beta(\Lambda,s)>0$ such that for any $\lambda\in\Lambda$:*

$$|\mathcal{L}_\lambda^{(s)}(x)|\leqslant Ae^{-\beta|x-\lambda|}.$$

One can refer to Ref. 7 for this technical result.

Thus a basis of $\mathcal{V}_0^{(s)}$ has been constructed and it consists of regular and well located functions. We are now ready to construct wavelets bases.

IV. CONSTRUCTION OF WAVELETS BASES

A nontrivial element of $\mathcal{V}_0^{(s)}$ is a wavelet if it is orthogonal to $\mathcal{V}_{-1}^{(s)}$. Let $\mathcal{W}_{-1}^{(s)}$ be the orthogonal of $\mathcal{V}_{-1}^{(s)}$ in $\mathcal{V}_0^{(s)}$. We wish to construct a Riesz basis of $\mathcal{W}_{-1}^{(s)}$. The fundamental property which makes things possible is

Property 4.1: Let $s>n/2$ be an integer. Then the orthogonal subspace, in $\mathbb{L}^2(\mathbb{R}^n)$, of $\mathcal{V}_0^{(s)}$ is $\Delta^s K_{2s}=\{\Delta^s h, h\in H^{2s}(\mathbb{R}^n), \forall\lambda\in\Lambda, h(\lambda)=0\}$.

Proof: If $f\in\mathcal{V}_0^{(s)}$ and if $h=\Delta^s h_1$ with $h_1\in K_{2s}$ then $\int_{\mathbb{R}^n} fh=\langle\Delta^s f,h_1\rangle=0$. Hence $\mathcal{V}_0^{(s)}$ is included in the orthogonal of $\Delta^s K_{2s}$.

Conversely let $f\in\mathbb{L}^2(\mathbb{R}^n)$ be orthogonal to $\Delta^s K_{2s}$. We call φ a C^∞ function, equal to 1 at 0, and with support sufficiently small so that all $\lambda+\text{supp }\varphi$ are disjoined. Then for any $g\in C_0^\infty$, $g(x)-\Sigma_{\lambda\in\Lambda} g(\lambda)\varphi(x-\lambda)$ is in K_{2s}. Hence

$$\int_{\mathbb{R}^n} f(x)\Delta^s\left(g(x)-\sum_{\lambda\in\Lambda} g(\lambda)\varphi(x-\lambda)\right)dx=0$$

and

$$\langle\Delta^s f,g\rangle=\sum_{\lambda\in\Lambda}\langle f,\Delta^s\varphi(x-\lambda)\rangle g(\lambda)$$

which means that $f\in\mathcal{V}_0^{(s)}$. □

The construction of a basis of $\mathcal{W}_{-1}^{(s)}$ is straightforward:

Property 4.2: The wavelets family:

$$\{\Psi_\lambda^{(s)}(x)=\Delta^s\mathcal{L}_\lambda^{(2s)}(x),\ \lambda\in\Lambda\backslash\theta\Lambda\}$$

is a Riesz basis of $\mathcal{W}_{-1}^{(s)}$.

Proof: Consider $f\in\mathbb{L}^2(\mathbb{R}^n)$ orthogonal to $\mathcal{V}_{-1}^{(s)}$. Then there exists $h\in H^{2s}(\mathbb{R}^n)$ vanishing on $\theta\Lambda$ such that $f=\Delta^s h$. Moreover if $f\in\mathcal{V}_0^{(s)}$ then $h\in\mathcal{V}_0^{(2s)}$. We deduce that h can be written in $H^{2s}(\mathbb{R}^n)$:

$$h(x)=\sum_{\lambda\in\Lambda} h(\lambda)\mathcal{L}_\lambda^{(2s)}(x).$$

Since h vanishes on $\theta\Lambda$, we have:

$$h(x)=\sum_{\lambda\in\Lambda\backslash\theta\Lambda} h(\lambda)\mathcal{L}_\lambda^{(2s)}(x).$$

Hence f can be written in $\mathbb{L}^2(\mathbb{R}^n)$:

$$f(x)=\sum_{\lambda\in\Lambda\backslash\theta\Lambda} h(\lambda)\Delta^s\mathcal{L}_\lambda^{(2s)}(x).$$

Once again h vanishes on $\theta\Lambda$, hence, by property 3.2, its H^{2s} norm is equivalent to $\|\Delta^s h\|_{\mathbb{L}^2} = \|f\|_{\mathbb{L}^2}$. h is in $\mathcal{V}_0^{(2s)}$ hence $\|h\|_{H^{2s}}$ is equivalent to $(\Sigma_{\lambda\in\Lambda}|h(\lambda)|^2)^{1/2}$. Finally the $\mathbb{L}^2$ norm of $\Sigma_{\lambda\in\Lambda\setminus\theta\Lambda}\ a_\lambda \Delta^s \mathcal{L}_\lambda^{(2s)}(x)$ is equivalent to the l^2 norm $(\Sigma_{\lambda\in\Lambda\setminus\theta\Lambda}\ |a_\lambda|^2)^{1/2}$. □

As a corollary of the preceeding results, we obtain a Riesz basis of $\mathbb{L}^2(\mathbb{R}^n)$ composed with regular wavelets:

Theorem 4.1: *Let Λ be a self-similar quasicrystal in $\mathbb{R}^n$ with inflation $\theta>1$. Then for any integer $s>n/2$, there exists a Riesz basis of $\mathbb{L}^2(\mathbb{R}^n)$ having the following structure:*

$$\theta^{nj/2}\psi_\lambda^{(s)}(\theta^j x-\lambda),\quad \lambda\in\Lambda\setminus\theta\Lambda,\quad j\in\mathbb{Z}.$$

The wavelets $\psi_\lambda^{(s)}$ belong to $H^{2s-\sigma}(\mathbb{R}^n)$ for any $\sigma>n/2$. Moreover any $f\in\mathbb{L}^2(\mathbb{R}^n)$ can be expanded into a unique orthogonal series:

$$f=\sum_{j\in\mathbb{Z}} f_j,$$

where for each integer j:

$$f_j(x)=\sum_{\lambda\in\Lambda\setminus\theta\Lambda} a_{j,\lambda}\theta^{nj/2}\psi_\lambda^{(s)}(\theta^j x-\lambda).$$

Proof: Let $\psi_\lambda^{(s)}(x-\lambda)=\Psi_\lambda^{(s)}(x)=\Delta^s\mathcal{L}_\lambda^{(2s)}(x)$. We use the following results:
(*i*) $\mathbb{L}^2(\mathbb{R}^n)=\oplus_{j\in\mathbb{Z}}\mathcal{W}_j^{(s)}$ where the sum is orthogonal.
(*ii*) The collection of functions $\theta^{nj/2}\Psi_\lambda^{(s)}(\theta^j x)$, $\lambda\in\Lambda\setminus\theta\Lambda$, is a Riesz basis of $\mathcal{W}_j^{(s)}$. □

V. STUDY OF THE MOTHER WAVELETS

In this section, we shall study the set of mother wavelets $\mathcal{M}_s=\{\psi_\lambda^{(s)},\lambda\in\Lambda\setminus\theta\Lambda\}$. First, we will show that wavelets $\psi_\lambda^{(s)}$ have exponential decay in a strong sense. Second, we will evaluate how different wavelets $\psi_\lambda^{(s)}$ are one from the other. Indeed if Λ is a lattice, $\mathcal{M}_s$ consists of one single function thanks to translation invariance. If Λ is aperiodic, $\mathcal{M}_s$ will lose its finiteness and will become a relatively compact set in some Sobolev space.

Wavelets $\psi_\lambda^{(s)}$ are very well localized functions as stated in the following result:

Theorem 5.1: *Let $s>n/2$ be an integer and let Λ be a self-similar quasicrystal in $\mathbb{R}^n$ with inflation $\theta>1$. Then there exist two constants $A=A(\Lambda,s)>0$ and $\beta=\beta(\Lambda,s)>0$ such that, for any $\lambda\in\Lambda\setminus\theta\Lambda$ and for any multi-index α whose length is strictly less than $2s-n$, the following estimate holds:*

$$|\partial^\alpha\psi_\lambda^{(s)}(x)|\leqslant Ae^{-\beta|x|}.$$

The proof relies on a fundamental property which can be seen as a local Bernstein inequality:

Property 5.1: Let $s>n/2$ be an integer. Consider a Delaunay set Λ in $\mathbb{R}^n$ and a radius $r>0$ such that each open ball $B_{x,2r}$ contains at most one point of Λ. Then there exists $C=C(\Lambda,s,r)$ such that, for any f in $\mathcal{V}_0^{(s)}(\Lambda)$, for any $\sigma\in]0,2s-n/2[$ and for any open ball $B_{x,r}$ centered on a point of Λ or whose distance to Λ exceeds 0, the following estimate holds:

$$\|f\|_{H^\sigma(B_{x,r/2})}\leqslant C\|f\|_{\mathbb{L}^2(B_{x,r})}. \tag{3}$$

Here the norm $\|f\|_{H^\sigma(B)}$ is the ''restriction'' Sobolev norm defined by $\inf\|F\|_{H^\sigma(\mathbb{R}^n)}$ where $F\in H^\sigma(\mathbb{R}^n)$ and where $F=f$ on B.

Proof: Let f be in $\mathcal{V}_0^{(s)}$ and $\lambda\in\Lambda$. On $B_{\lambda,r}$, f satisfies $\Delta^s f=\gamma\delta_\lambda$ for some $\gamma\in\mathbb{R}$. Translating f if necessary, we can restrict our attention to the case where $\lambda=0$ and for the sake of simplicity we will assume $r=1$. Let $B=B_{0,1}$ and let us show the following preliminary inequality:

$$\|f\|_{H^\sigma(B/2)}\leqslant C|\gamma|+C'\|f\|_{\mathbb{L}^2(B)},\quad \sigma\in]0,2s-n/2[,$$

where C and C' depend only on s and $|B|$.

First case: $\gamma=0$

Define $\tilde{f}\in\mathbb{L}^2(\mathbb{R}^n)$ to coincide with f on B and to vanish outside B. Then $S=\Delta^s\tilde{f}$ is a distribution belonging to $H^{-2s}(\mathbb{R}^n)$ and whose support is included in $|t|=1$. Let E be the fundamental solution of Δ^s given by

$$E(t)=c(s,n)|t|^{2s-n}(\log|t|)^{\epsilon},$$

where $\epsilon=0$ if n is odd and $\epsilon=1$ if n is even (see, e.g., Ref. 16). Then

$$\Delta^s(\tilde{f}\star E)=\tilde{f}$$

and

$$\tilde{f}=S\star E.$$

If $|t|<1/2$, $y\to E(t-y)$ coincides, in a neighborhood of supp S, with a C^∞ function $y\to E'(t,y)$ vanishing outside $|y|\leqslant 2$ and the following uniform estimate holds

$$\sup_{|t|<1/2}\|E'(t,\cdot)\|_{H^{2s}(\mathbb{R}^n)}<C_s.$$

Thus, for $|\alpha|\leqslant 2s-[n/2]=q$, ($[x]$ denotes the integral part of x):

$$|\partial^\alpha f(t)|\leqslant C_s\|S\|_{H^{-2s}}\leqslant C_s\|f\|_{\mathbb{L}^2(B)}.$$

Integrating on $B/2$ gives

$$\|f\|_{H^q(B/2)}\leqslant C_s\|f\|_{\mathbb{L}^2(B)}.$$

Hence, by interpolation, the same inequality holds when replacing q by $\sigma\in]0,2s-n/2[$.

Second case: $\gamma=1$

f can be written $f=g+E$ where E is defined above and where g satisfies $\Delta^s g=0$. E is in $\mathbb{L}^2(B)$ and hence so is g. The preceeding estimate then holds for g,

$$\|g\|_{H^\sigma(B/2)}\leqslant C_s\|g\|_{\mathbb{L}^2(B)},\quad \sigma\in]0,2s-n/2[.$$

Then observe that $E\in H^\sigma(B)$, $\sigma\in]0,2s-n/2[$, since $E=f-g$. Therefore

$$\|f\|_{H^\sigma(B/2)}\leqslant C_s\|g\|_{\mathbb{L}^2(B)}+\|E\|_{H^\sigma(B/2)}\leqslant C_s\|f\|_{\mathbb{L}^2(B)}+(1+C_s)\|E\|_{H^\sigma(B/2)}\leqslant C_s\|f\|_{\mathbb{L}^2(B)}+C_s'.$$

One can deduce the general case $\gamma\in\mathbb{R}$ using a homogeneity argument. Let $\varphi\in C_0^\infty$ be a function with value 1 at 0 and vanishing for $|x|\geqslant r$. One has

$$\gamma=\langle\gamma\delta_0,\varphi\rangle=\langle\Delta^s f,\varphi\rangle=\langle f,\Delta^s\varphi\rangle$$

hence $|\gamma|\leqslant C\|f\|_{\mathbb{L}^2(B)}$ and 5.1 holds true when $B_{x,r}$ is centered on a point $\lambda\in\Lambda$. If the distance from $B_{x,r}$ to Λ exceeds 0 then things work just as in the preceeding case where $\gamma=0$. □

Proof of theorem 5.1: Write $\mathcal{L}_\lambda^{(s)}(x)=L_\lambda^{(s)}(x-\lambda)$. Let α be a multi-index with $|\alpha|<2s-n$. Choose $\sigma>n/2$ such $|\alpha|+\sigma<2s-n/2$. Then using Sobolev injections

$$|\partial^\alpha L_\lambda^{(s)}(x)|\leqslant C_{\sigma,r}\|\partial^\alpha L_\lambda^{(s)}\|_{H^\sigma(B_{x,r/2})}\leqslant C_{\sigma,r,\alpha}\|L_\lambda^{(s)}\|_{H^{\sigma+|\alpha|}(B_{x,r/2})}$$

since $\sigma+|\alpha|<2s-n/2$, property 5.1 applies and we obtain

$$|\partial^\alpha L_\lambda^{(s)}(x)|\leqslant C'_{\sigma,r,\alpha}\|L_\lambda^{(s)}\|_{\mathbb{L}^2(B_{x,r})}.$$

Finally using the exponential decay of $L_\lambda^{(s)}$ given in theorem 3.2, we get

$$|\partial^\alpha L_\lambda^{(s)}(x)|\leqslant C''_{\sigma,r,\alpha}e^{-\beta|x|}.$$

Consequently the same estimates hold true for $\psi_\lambda^{(s)}$ since $\psi_\lambda^{(s)}=\Delta^s L_\lambda^{(2s)}$. □

We, now, wish to study the deformation $\lambda \to \psi_\lambda^{(s)}$ in order to measure how different wavelets $\psi_\lambda^{(s)}$ are one point to another. To achieve this goal, we introduce a topology on Λ that will make it precompact.

Property 5.2: Let Λ be an aperiodic Delaunay set in $\mathbb{R}^n$. Define a mapping d on Λ^2 by the rule $d(\lambda,\lambda')=R^{-1}$ if R is the supremum of all $r>0$ such that $\Lambda-\lambda$ and $\Lambda-\lambda'$ coincide on $B_{0,r}$. Then d is a distance on Λ called local configurations distance. Moreover, if Λ is an aperiodic quasicrystal, (Λ,d) is precompact.

Proof: If $d(\lambda,\lambda')=0$ then $\Lambda-\lambda$ and $\Lambda-\lambda'$ coincide on the whole space. Hence $\lambda-\lambda'+\Lambda=\Lambda$ which forces $\lambda=\lambda'$ since Λ is aperiodic. Let λ, λ', and λ'' be three points in Λ. Put $d(\lambda,\lambda'')=a^{-1}$, $d(\lambda,\lambda')=b^{-1}$, $d(\lambda'',\lambda')=c^{-1}$ and $\rho\in]0,1[$. $\Lambda-\lambda$ and $\Lambda-\lambda'$ coincide on $B_{0,\rho b}$ whereas $\Lambda-\lambda''$ and $\Lambda-\lambda'$ coincide on $B_{0,\rho c}$. Therefore $\Lambda-\lambda$ and $\Lambda-\lambda''$ coincide on $B_{0,\rho\ \min(b,c)}$. Hence $a\geqslant\rho\ \min(b,c)$ and $a^{-1}\leqslant b^{-1}+c^{-1}$. Suppose Λ is an aperiodic quasicrystal and let $\epsilon>0$. Using property 2.1, we know that $(\Lambda-\lambda)\cap B_{0,\epsilon^{-1}}$ describes a finite set as λ goes through Λ. Hence Λ can be covered with a finite number of d-balls with radius ϵ. □

We are ready to state the second result concerning the set $\mathcal{M}_s$:

Theorem 5.2: *Let $s>n/2$ be an integer and let Λ be a self-similar quasicrystal in $\mathbb{R}^n$ with inflation $\theta>1$. Then the mapping*

$$\lambda\in\Lambda\backslash\theta\Lambda\to\psi_\lambda^{(s)}\in H^{2s-[n/2]-1}(\mathbb{R}^n)$$

is continuous when Λ is equipped with the local configurations distance.

Proof: Let $\epsilon>0$ and $\lambda\in\Lambda\backslash\theta\Lambda$. We use a technical result of Lemarié-Rieusset[7] ensuring that there exists a family of functions $g_{\lambda,j}$, $\lambda\in\Lambda$, $j\in\mathbb{N}$, such that

$$\lim_{j\to\infty}\ \sup_{\lambda\in\Lambda}\|L_\lambda^{(2s)}-g_{\lambda,j}\|_{H^{2s}}=0$$

and such that $g_{\lambda,j}$ depends only on the set $(\Lambda-\lambda)\cap B_{0,R_j}$ where $(R_j)_j$ is an unbounded increasing sequence. Therefore choose j such that $\sup_{\lambda\in\Lambda}\|L_\lambda^{(2s)}-g_{\lambda,j}\|_{H^{2s}}\leqslant\epsilon$ and λ' such that $d(\lambda,\lambda')\leqslant R_j^{-1}$. Then

$$\|L_\lambda^{(2s)}-L_{\lambda'}^{(2s)}\|_{H^{2s}}\leqslant\|L_\lambda^{(2s)}-g_{\lambda,j}\|_{H^{2s}}+\|g_{\lambda,j}-g_{\lambda',j}\|_{H^{2s}}+\|g_{\lambda',j}-L_{\lambda'}^{(2s)}\|_{H^{2s}}\leqslant 2\epsilon.$$

Since Λ is a quasicrystal, $\Lambda-\Lambda$ is a Delaunay set and each $L_\lambda^{(2s)}$ belongs to $\mathcal{V}_0^{(2s)}(\Lambda-\Lambda)$. We then apply property 5.1 to the Delaunay set $\Lambda-\Lambda$ and observe that inequalities (3) can be turned into global estimates which show that $\|f\|_{H^{2s}}$ and $\|f\|_{H^{4s-[n/2]-1}}$ are equivalent norms on $\mathcal{V}_0^{(2s)}(\Lambda-\Lambda)$. Hence

$$\|L_\lambda^{(2s)}-L_{\lambda'}^{(2s)}\|_{H^{4s-[n/2]-1}}\leqslant 2C\epsilon$$

and since $\psi_\lambda^{(s)}=\Delta^s L_\lambda^{(2s)}$:

$$\|\psi_\lambda^{(s)}-\psi_{\lambda'}^{(s)}\|_{H^{2s-[n/2]-1}}\leqslant 2C\epsilon.$$ □

Consequence 5.1: The set $\mathcal{M}_s$ is relatively compact in $H^{2s-[n/2]-1}(\mathbb{R}^n)$.

Proof: This follows from the precompacity of (Λ,d) and the continuity of $\Lambda\to\psi_\lambda^{(s)}$. □

VI. COMPACTLY SUPPORTED WAVELETS ADAPTED TO A SELF-SIMILAR QUASICRYSTAL IN $\mathbb{R}$

We now restrict our attention to the one-dimensional case and the quasicrystal Λ can be considered as a strictly increasing sequence $(\lambda_k)_{k\in\mathbb{Z}}$. We wish to construct a Riesz basis of $\mathbb{L}^2(\mathbb{R})$ composed with compactly supported wavelets adapted to Λ and to its self-similarity. This means that the basis should have the following structure:

$$\theta^{j/2}\psi_\lambda(\theta^j x-\lambda),\quad j\in\mathbb{Z},\quad \lambda\in\Lambda.$$

Once again we will introduce a multiresolution analysis $(V_j^{(s)})_{j\in\mathbb{Z}}$ and then obtain the wavelet basis.

Let s be an integer greater than 1. We call $V_0^{(s)}$ the closed subspace of $\mathbb{L}^2(\mathbb{R})$ of all functions whose derivative of order s (in the sense of distributions) is a sum of Dirac masses located on Λ,

$$\frac{d^s}{dx^s} f = \sum_{\lambda\in\Lambda} a_\lambda \delta_\lambda .$$

With the notations of Sec. III, $V_0^{(2s)} = \mathcal{V}_0^{(s)}$.

It is equivalent to defining $V_0^{(s)}$ as the subspace of functions in C^{s-2}, in $\mathbb{L}^2(\mathbb{R})$ and whose restriction to each $[\lambda_k,\lambda_{k+1}]$ is a polynomial of degree $s-1$. These spline spaces have been widely studied. Nevertheless the context of a self-similar quasicrystal is interesting because it involves irrational inflations (Pisot or Salem numbers) which is quite unusual in the field of wavelets.

A. Riesz basis of $V_0^{(s)}$ composed with translates of a finite number of compactly supported functions

One can find many references about one-dimensional splines with knots located on a Delaunay set (see e.g., Ref. 17 or 18) so only classical results will be recalled. The first one deals with a Riesz basis of $V_0^{(s)}$:

Theorem 6.1: *Let Λ be a Delaunay set in $\mathbb{R}$. For any integer $s\geqslant 2$, there exists a Riesz basis of $V_0^{(s)}$, $\{B_\lambda^{(s)}(x),\lambda\in\Lambda\}$, defined as follows:*
$B_\lambda^{(s)}$ is the unique function in $V_0^{(s)}$ satisfying
(i) supp $B_\lambda^{(s)} = [\lambda,\lambda']$ *where* $\lambda'\in\Lambda$.
(ii) $]\lambda,\lambda'[$ *contains exactly* $s-1$ *points of* Λ.
(iii) $\int_{\mathbb{R}} B_\lambda^{(s)}(x)dx = (\lambda'-\lambda)/s$.
One can refer to Ref. 17 for a proof.
Now if Λ is a quasicrystal, the B-spline basis has a very simple structure:

Corollary 6.1: If Λ is a quasicrystal in $\mathbb{R}$, the B-spline basis of $V_0^{(s)}$ can be written

$$\{\varphi_\lambda(x-\lambda),\ \ \lambda\in\Lambda\},$$

where the set $\{\varphi_\lambda(x),\lambda\in\Lambda\}$ is a finite set of compactly supported functions:

Proof: It is straightforward since the B-spline $B_{\lambda_k}^{(s)}$ depends only on the local structure $\Lambda\cap[\lambda_k,\lambda_{k+s}]$ and since the quasicrystal Λ has only a finite number of local configurations $[\lambda_k,\lambda_{k+s}]-\lambda_k$ when k goes through $\mathbb{Z}$. □

B. Compactly supported wavelets

As in the n-dimensional case, we define a multiresolution analysis involving Λ's scaling property:

$$f\in V_j^{(s)} \Leftrightarrow f(\theta^{-j}x)\in V_0^{(s)} .$$

Then once again the sequence $(V_j)_j$ is a multiresolution analysis and the problem is to construct a Riesz basis of $W_{-1}^{(s)}$, the orthogonal subspace of $V_{-1}^{(s)}$ in $V_0^{(s)}$:

$$V_0^{(s)} = V_{-1}^{(s)} \oplus^{\perp} W_{-1}^{(s)} .$$

Let us prove the following result:

Theorem 6.2: *Let Λ be a self-similar quasicrystal with inflation $\theta>1$. Then for each $s>1$, there exists a Riesz basis of $W_{-1}^{(s)}$ having the following structure:*

$$\psi_\lambda^{(s)}(x-\lambda),\ \ \lambda\in\Lambda\backslash\theta\Lambda,$$

where $\{\psi_\lambda^{(s)},\lambda\in\Lambda\backslash\theta\Lambda\}$ is a finite set of compactly supported wavelets.

This result should be connected with a method developed by Buhmann and Micchelli[8] to construct spline wavelets. Let us first recall the setting of the problem in Ref. 8:

Given two embedded Delaunay sets $\Lambda_{-1}\subset\Lambda_0$, does there exist a Riesz basis of the orthogonal supplement of $V_0^{(s)}(\Lambda_{-1})$ in $V_0^{(s)}(\Lambda_0)$ composed with compactly supported functions?

Our problem is a similar one since $\theta\Lambda$ is a Delaunay set included in Λ and since $V_{-1}^{(s)} = V_0^{(s)}(\theta\Lambda)$. But its originality comes from the finite number of mother wavelets and also from the fact that our wavelets satisfy some scaling equations involving an irrational scale.

Proof of theorem 6.2: We first observe that $W_{-1}^{(s)}$ has an explicit characterization just as it works in property 4.1: for any integer $s\geqslant 2$, $W_{-1}^{(s)}$ is the set of $f\in\mathbb{L}^2(\mathbb{R})$ which can be written

$$f=\frac{d^s}{dx^s}\,g,$$

where $g\in V_0^{(2s)}$ and where g vanishes on $\theta\Lambda$.

Therefore the idea is to study the subspace $\widetilde{W}_0^{(s)}$ of $V_0^{(s)}$ of functions which vanish on $\theta\Lambda$. We are going to construct a Riesz basis of $\widetilde{W}_0^{(2s)}$ composed with compactly supported functions. Then by differentiation, wavelets of $W_{-1}^{(s)}$ will be obtained. To achieve this goal, the main idea is to look for functions in $\widetilde{W}_0^{(s)}$ with minimal support. Precisely we denote by E the set of integers k such that $\lambda_{k+1}\notin\theta\Lambda$, and for each $k\in E$, define $\Psi_k^{(s)}$ as the unique element in $\widetilde{W}_0^{(s)}$ satisfying the following conditions:

(i) the support of $\Psi_k^{(s)}$ is compact and included in $[\lambda_k,+\infty)$,
(ii) $\Psi_k^{(s)}(\lambda_{k+1})=1$,
(iii) any function in $\widetilde{W}_0^{(s)}$ whose support is included in the support of $\Psi_k^{(s)}$ is proportional to $\Psi_k^{(s)}$.

Then we can state:

Theorem 6.3: *The collection $\{\Psi_k^{(s)},k\in E\}$ is a Riesz basis of $\widetilde{W}_0^{(s)}$.*

Sketch of the proof: This result can be found in detail in Ref. 5. We will only recall the main steps to the proof:

Proposition 6.1: For any $k\in E$, the support of $\Psi_k^{(s)}$ writes $[\lambda_k,\lambda_{k+N_k}]$ where N_k is the smallest integer greater than s such that:

$$N_k=s+\mathrm{card}(]\lambda_k,\lambda_{k+N_k}[\cap\theta\Lambda).$$

Proposition 6.2: There exists a finite set $\mathcal{G}_s$ of compactly supported functions such that for each $k\in E$, $\Psi_k^{(s)}(x+\lambda_k)\in\mathcal{G}_s$.

Proposition 6.3: Any function in $\widetilde{W}_0^{(s)}$ with compact support is a linear combination of $\Psi_k^{(s)},k\in E$.

Proposition 6.4: The set of all functions in $\widetilde{W}_0^{(s)}$ with compact support is dense in $\widetilde{W}_0^{(s)}$.

Proposition 6.5: There exist $0<c_1\leqslant c_2$ depending only on s and Λ such that for any real sequence $(\alpha_k)_{k\in E}$ with finite support

$$c_1\sum_{k\in E}|\alpha_k|^2\leqslant\left\|\sum_{k\in E}\alpha_k\Psi_k^{(s)}\right\|_{\mathbb{L}^2}^2\leqslant c_2\sum_{k\in E}|\alpha_k|^2.$$

Theorem 6.3 is straightforward using these technical properties. Therefore one can deduce theorem 6.2: using property 3.2, $\|g\|_{\mathbb{L}^2}$ and $\|(d^s/dx^s)g\|_{\mathbb{L}^2}$ are equivalent norms on $\widetilde{W}_0^{(2s)}$. Then $\{(d^s/dx^s)\Psi_k^{(2s)},k\in E\}$ is a Riesz basis of $W_{-1}^{(s)}$ with the desired structure. □

As a corollary of theorem 6.2 we can state:

Corollary 6.2: Let Λ be a self-similar quasicrystal in $\mathbb{R}$ with inflation $\theta>1$. Then, for any integer $s>1$, there exists a Riesz basis of $\mathbb{L}^2(\mathbb{R})$ having the following structure:

$$\theta^{j/2}\psi_\lambda^{(s)}(\theta^j x-\lambda),\quad \lambda\in\Lambda\setminus\theta\Lambda,\quad j\in\mathbb{Z},$$

where $\{\psi_\lambda^{(s)},\lambda\in\Lambda\setminus\theta\Lambda\}$ is a finite set of C^{s-1} compactly supported functions.

[1] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, ''Metallic phase with long-range orientational order and no translational symmetry,'' Phys. Rev. Lett. **53**, 1951–1953 (1984).
[2] Y. Meyer, ''Quasicrystals, diophantine approximations and algebraic numbers,'' in *Beyond Quasicrystals*, edited by F. Axel and D. Gratias, Proc. Les Houches, March 1994, Les Editions de Physique (Springer, New York, 1995).
[3] R. V. Moody, ''Meyer sets and their duals,'' *Mathematics of Long Range Aperiodic Order*, edited by R. V. Moody (Kluwer, Dordrecht, 1997).
[4] J.-P. Gazeau, ''Quasicrystals and their symmetries,'' *Symmetries and Structural Properties of Condensed Matter*, edited by W. Florek, T. Lulek, and S. Walcerz (World Scientific, Singapore, 1995).
[5] G. Bernuau, ''Propriétés spectrales et géométriques des quasicristaux. Ondelettes adaptées aux quasicristaux,'' Ph.D. thesis, Ceremade, Université Paris IX Dauphine, France, 1998.
[6] S. Jaffard, ''Construction et propriétés des bases d'ondelettes. Remarques sur la contrôlabilité exacte,'' Ph.D. thesis, Ecole Polytechnique, France, 1989.
[7] P.-G. Lemarié-Rieusset, ''Base d'ondelettes sur les groupes de Lie stratifiés,'' Bull. Soc. Math. Fr. **117**, 211–232 (1989).
[8] M. D. Buhmann and C. A. Micchelli, ''Spline prewavelets for nonuniform knots,'' Numer. Math. **61**, 455–474 (1992).
[9] P. Auscher, ''Wavelets bases for $\mathbb{L}^2(\mathbb{R})$ with rational dilation factor,'' in *Wavelets and Their Applications*, edited by M. B. Ruskai *et al.* (Jones and Bartlett, 1992), pp. 439–452.
[10] J.-P. Gazeau and J. Patera, ''Tau wavelets of Haar,'' J. Phys. A **29**, 4549–4559 (1996).
[11] J.-P. Gazeau and V. Spiridonov, ''Toward discrete wavelets with irrational scaling factor,'' J. Math. Phys. **37**, 3001–3013 (1996).
[12] J.-P. Gazeau, J. Patera, and E. Pelantova, ''Tau wavelets in the plane,'' J. Math. Phys. (to be published).
[13] R. Salem, *Algebraic Numbers and Fourier Analysis* (Heath, Boston, 1963).
[14] J. C. Lagarias, ''Meyer's concept of quasicrystal and quasiregular sets,'' Commun. Math. Phys. **179**, 365–376 (1996).
[15] Y. Meyer, *Ondelettes et Opérateurs, Tome I* (Hermann, Paris, 1990).
[16] L. Schwartz, *Théorie des Distributions* (Hermann, Paris, 1966).
[17] L. L. Schumaker, *Spline Functions: Basic Theory* (Wiley, New York, 1981).
[18] C. de Boor, ''Odd degree spline interpolation at a bi-infinite knot sequence,'' in *Approximation Theory*, edited by R. Schabak and K. Scherer, Springer Notes in Mathematics Vol. 556 (Springer, Bonn, 1976).

Square wave analysis

Yuchuan Wei[a)]
Institute of Applied Physics, Beijing University of Science and Technology, Beijing 100083, People's Republic of China

Nanxian Chen
Institute of Applied Physics, Beijing University of Science and Technology, Beijing 100083, People's Republic of China and CCAST(World Lab), P.O. Box 8730, Beijing 100080, People's Republic of China

(Received 17 November 1997; accepted for publication 13 March 1998)

This paper concerns square wave analysis, which is in no small part a generalization of the Fourier analysis based on sine and cosine functions. Following sine and cosine functions square waves become another frequently used and easily generated waveform in electronics, so it is an urgent practical problem to study the basic properties of square waves and the fundamental theory of square wave analysis. Like the sine-cosine function system, the square wave system, i.e., the function system of square waves with different frequencies (or periods), is linearly independent and complete in the real Hilbert spaces $L^2(-\pi,\pi)$, so square waves can approximate an arbitrary quadratically integrable function with a vanishing mean-square error. Since the square wave system is nonorthogonal, its biorthogonal functions are presented so that a function with a certain condition can be expanded as a square wave series easily. Meanwhile the orthogonalization of the square wave system is discussed so that a quadratically integrable function can be approximated best by a superposition of finite square waves. These results form the theory basis of the square wave analysis technique in modern electronics. © *1998 American Institute of Physics.* [S0022-2488(98)00808-1]

I. INTRODUCTION

About two hundred years ago, the French mathematician J. B. J. Fourier asserted that any function with period 2π can be expressed as a trigonometric series. This great idea has had an important influence upon many sciences including mathematics and physics. In electronics, Fourier analysis has been playing an important role, and a signal is often considered as a superposition of many sine and cosine waves with different frequencies. However, with the development of switching circuits and pulse technique, it becomes very easy to generate and process square waves. For example, it is quite convenient to obtain a system of square waves with different frequencies from a common high-frequency oscillation by means of counting circuits, whereas it is complicated to do the same thing with sine and cosine functions. Again, multipliers and summers for square waves are very simple. The following basic questions arise naturally:

(1) Can a signal be considered as a superposition of square waves with different frequencies (or periods)?
(2) How to decompose a signal into a square wave series?
(3) How to approximate a signal best by a superposition of finite square waves?

This problem is called square wave analysis, which is a natural generalization of Fourier's idea.

The aim of this paper is to answer these practical questions mathematically and lay a theory foundation for the square wave analysis technique in electronics. In terms of mathematics, we shall consider the following items one after another in $L^2(-\pi,\pi)$:

(1) Square wave system and its basic properties,

a)Corresponding author. Electronic mail: weiyuch@bltda.com.bta.net.cn

 © 1998 American Institute of Physics

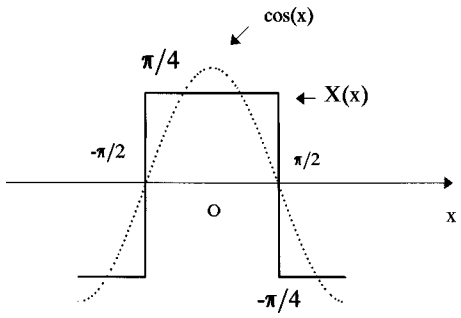

FIG. 1. The even square wave $X(x)$.

(2) biorthogonal functions and square wave series,
(3) orthogonalization and the best approximation.

In this paper $L^2(-\pi,\pi)$ denotes the Hilbert space of real quadratically integrable functions defined on $(-\pi,\pi)$, and $\langle,\rangle$ denotes the inner product in the space.

The results show that square wave analysis has a close relation with the Möbius function in the number theory unexpectedly.

II. SQUARE WAVE SYSTEM AND ITS BASIC PROPERTIES

Definition: The functions $X(x)$ and $Y(x)$ are called an even and odd square wave, respectively, which are defined on $(-\infty,+\infty)$ by

$$X(x)=\frac{\pi}{4}\operatorname{sgn}(\cos x), \tag{1}$$

and

$$Y(x)=\frac{\pi}{4}\operatorname{sgn}(\sin x). \tag{2}$$

Here sgn is the sign function:

$$\operatorname{sgn}(x)=\begin{cases} 1, & x>0 \\ 0, & x=0 \\ -1, & x<0. \end{cases} \tag{3}$$

For convenience let the amplitude of the square waves be $\pi/4$. See Figs. 1 and 2.

The Fourier series of the square waves are

$$X(x)=\cos x-\tfrac{1}{3}\cos(3x)+\tfrac{1}{5}\cos(5x)-\tfrac{1}{7}\cos(7x)+\ldots \tag{4}$$

$$=\sum_{n=1}^{\infty} A(n)\cos(nx), \tag{5}$$

and

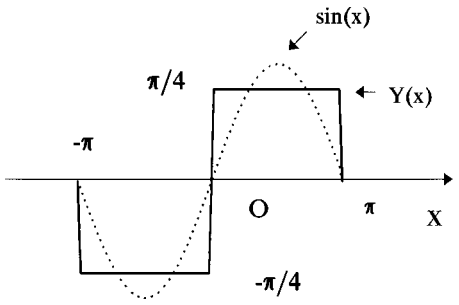

FIG. 2. The odd square wave $Y(x)$.

$$Y(x)=\sin x+\tfrac{1}{3}\sin(3x)+\tfrac{1}{5}\sin(5x)+\tfrac{1}{7}\sin(7x)+\cdots \tag{6}$$

$$=\sum_{n=1}^{\infty} B(n)\sin(nx). \tag{7}$$

Notice that the Fourier coefficients

$$A(n)=\begin{cases}(-1)^{(n-1)/2}\dfrac{1}{n}, & n=1,3,5,\ldots\\ 0, & n=2,4,6,\ldots\end{cases} \tag{8}$$

and

$$B(n)=\begin{cases}\dfrac{1}{n}, & n=1,3,5,\ldots\\ 0, & n=2,4,6,\ldots\end{cases} \tag{9}$$

have the property that

$$A(n)A(m)=A(mn) \tag{10}$$

and

$$B(n)B(m)=B(mn), \tag{11}$$

for any two positive integers m and n. We shall use this property many times.

Definition: We call the function system of square waves

$$1,X(x),Y(x),X(2x),Y(2x),\ldots,X(nx),Y(nx),\ldots \tag{12}$$

the square wave system.

By the way, please note that the square wave system defined here is quite different from the Walsh function system, which consists of very complicated rectangular waves (Ref. 1, Chap. 1). The Redemacher function system can be considered as a proper subset of the square wave system in essence, which is a well-known incomplete system (Ref. 1, Chap. 1).

In what follows we shall show that the square wave system is linearly independent, nonorthogonal and complete in $L^2(-\pi,\pi)$.

(1) Linear independence

Proposition 1: The function system of odd square waves

$$Y(x),Y(2x),Y(3x),\ldots,Y(nx),\ldots \tag{13}$$

is linearly independent, i.e., any finite subsystem of the function system is linearly independent.

Proof: We will prove that $\{Y(nx)\}_{n=1}^{N}$ is linearly independent for $N=2,3,4,\ldots$ by mathematical induction.

1^0. $Y(x)$ and $Y(2x)$ are linearly independent.

Notice that

$$c_1Y(x)+c_2Y(2x)=\begin{cases}\dfrac{\pi}{4}(c_1+c_2) & x\in\left(0,\dfrac{\pi}{2}\right)\\ \dfrac{\pi}{4}(c_1-c_2) & x\in\left(\dfrac{\pi}{2},\pi\right).\end{cases} \tag{14}$$

Therefore

$$\begin{aligned}&c_1Y(x)+c_2Y(2x)=0\\ \Rightarrow&\begin{cases}(c_1+c_2)=0\\(c_1-c_2)=0\end{cases}\\ \Rightarrow&c_1=c_2=0.\end{aligned}$$

2^0. If $\{Y(nx)\}_{n=1}^{k}$ is linearly independent, then so is $\{Y(nx)\}_{n=1}^{k+1}$.

Notice that

$$\begin{aligned}&c_1Y(x)+c_2Y(2x)+\ldots+c_kY(kx)+c_{k+1}Y((k+1)x)\\ =&\begin{cases}\dfrac{\pi}{4}(c_1+c_2+\ldots+c_k+c_{k+1}), & x\in\left(0,\dfrac{\pi}{k+1}\right)\\ \dfrac{\pi}{4}(c_1+c_2+\ldots+c_k-c_{k+1}), & x\in\left(\dfrac{\pi}{k+1},\dfrac{\pi}{k}\right).\end{cases}\end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned}&c_1Y(x)+c_2Y(2x)+\ldots+c_kY(kx)+c_{k+1}Y((k+1)x)=0\\ \Rightarrow&\begin{cases}c_1+c_2+\ldots+c_k+c_{k+1}=0\\c_1+c_2+\ldots+c_k-c_{k+1}=0\end{cases}\\ \Rightarrow&c_{k+1}=0\\ \Rightarrow&c_1Y(x)+c_2Y(2x)+\ldots+c_kY(kx)=0\\ \Rightarrow&c_1=c_2=\ldots=c_k=0\\ \Rightarrow&c_1=c_2=\ldots=c_k=c_{k+1}=0.\end{aligned}$$

From 1^0 and 2^0, it follows that the function system of odd square waves $\{Y(nx)\}_{n=1}^{\infty}$ is linearly independent. □

It can be proved in the same way that the function system of even square waves

$$1,X(x),X(2x),\ldots,X(nx),\ldots \tag{16}$$

is linearly independent as well. Considering the orthogonality of even functions and odd functions, we have

Proposition 2: The square wave system is linearly independent.

(2) Nonorthogonality

Though odd square waves and even square waves are orthogonal, the square wave system is not an orthogonal system. In fact, we have the following inner product formulas.

Formula: For any two positive integers m and n, we have

$$\langle X(mx),X(nx)\rangle=\frac{\pi^3}{8}A\left(\frac{mn}{(m,n)^2}\right), \tag{17}$$

$$\langle Y(mx),Y(nx)\rangle=\frac{\pi^3}{8}B\left(\frac{mn}{(m,n)^2}\right), \tag{18}$$

where (m,n) denotes the greatest common divisor of m and n.

Proof:

$$\begin{aligned}\langle Y(mx),Y(nx)\rangle&=\left\langle\sum_{l=1}^{\infty}B(l)\sin(mlx),\sum_{k=1}^{\infty}B(k)\sin(nkx)\right\rangle\\&=\pi\sum_{ml=nk}B(l)B(k)\\&=\pi\sum_{r=1}^{\infty}B\left(\frac{r[m,n]}{m}\right)B\left(\frac{r[m,n]}{n}\right)\\&=\pi B\left(\frac{[m,n]^2}{mn}\right)\sum_{r=1}^{\infty}B(r^2)\\&=\frac{\pi^3}{8}B\left(\frac{mn}{(m,n)^2}\right).\end{aligned}$$

Here $[m,n]$ denotes the least common multiple of m and n, and we have

$$mn=(m,n)[m,n]. \tag{19}$$

The other equality can be deduced similarly. □

(3) Completeness

We have two basic series

$$\sin x=\sum_{n=1}^{\infty}\mu(n)B(n)Y(nx) \tag{20}$$

and

$$\cos x=\sum_{n=1}^{\infty}\mu(n)A(n)X(nx) \tag{21}$$

in $L^2(-\pi,\pi)$.

Here, $\mu(n)$ denotes the Möbius function (Ref. 2, Chap. 6), which is defined by

$$\mu(n)=\begin{cases}1, & \text{if } n=1\\(-1)^r, & \text{if } n \text{ is the product of } r \text{ distinct primes}\\0, & \text{if } n \text{ is divisible by a prime square.}\end{cases} \tag{22}$$

Notice that $\mu(n)$ takes three values $1,-1$, and 0 only. The first ten values are $\mu(1)=1$, $\mu(2)=-1$, $\mu(3)=-1$, $\mu(4)=0$, $\mu(5)=-1$, $\mu(6)=1$, $\mu(7)=-1$, $\mu(8)=0$, $\mu(9)=0$, $\mu(10)=1$.

The Möbius function has a simple property that

$$\sum_{d|n}\mu(d)=\begin{cases}1 & n=1\\0 & n=2,3,4,\ldots,\end{cases} \tag{23}$$

where the summation goes over the positive factors of n including 1 and n, and $d|n$ denotes that d is a factor of n. For example, for $n=6$, we have

$$\sum_{d|6} \mu(d)=\mu(1)+\mu(2)+\mu(3)+\mu(6)=0.$$

Proof of the series (20) and (21).

$$\sum_{n=1}^{\infty} \mu(n)B(n)Y(nx) \tag{24}$$

$$=\sum_{n=1}^{\infty} \mu(n)B(n)\sum_{m=1}^{\infty} B(m)\sin(mnx) \tag{25}$$

$$=\sum_{k=1}^{\infty}\left(\sum_{d|k} \mu(d)\right)B(k)\sin kx \tag{26}$$

$$=\sin x. \tag{27}$$

From (25) to (26) the summation order was changed, but this does not matter for the series encountered here are unconditionally convergent in $L^2(-\pi,\pi)$. The reason for unconditional convergence is that the series

$$\sum_{k=1}^{\infty}\left(\sum_{d|k} |\mu(d)|\right)|B(k)|\ \sin kx \tag{28}$$

is convergent, the norm square of which is

$$\pi\sum_{k=1}^{\infty}\left(\sum_{d|k} |\mu(d)|\right)^2 B^2(k)\leqslant \pi\sum_{k=1}^{\infty} d^2(k)B^2(k)$$

$$=\pi\left[\frac{1}{1^2}+\frac{d^2(3)}{3^2}+\frac{d^2(5)}{5^2}+\ldots\right]$$

$$=\frac{3}{128}\pi^5. \quad \text{(Reference 2, Sec. 6.14.)}$$

Here $d(k)$ denotes the number of positive factors of the positive integer k (Ref. 2, Chap. 6), and we have

$$\sum_{d|k} |\mu(d)|\leqslant d(k). \tag{29}$$

The series about cos$x$ can be discussed similarly. □

The convergence process of the series (20) is shown in Fig. 3. The first few square wave components contained in sin$x$ are shown in Fig. 4.

Proposition 3: The square wave system is complete in $L^2(-\pi,\pi)$.

Proof: Since the sine and cosine function system is complete in $L^2(-\pi,\pi)$, and sine and cosine functions can be expressed as a square wave series, the square wave system is complete in the Hilbert space $L^2(-\pi,\pi)$. In other words, superpositons of square waves can approximate an arbitrary quadratically integrable function with a vanishing mean-square error. □

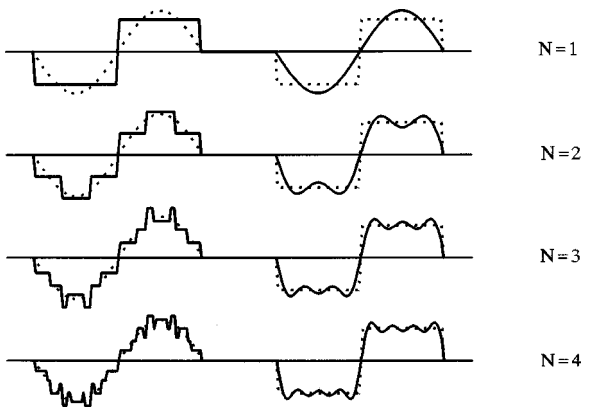

FIG. 3. The trigonometric polynomials $\Sigma_{k=1}^{N}(1/(2k-1))\sin((2k-1)x)$ approximate $Y(x)$, and the square wave polynomials $\Sigma_{k=1}^{N}[\mu(2k-1)/2k-1]\,Y((2k-1)x)$ approximate $\sin x$ as their term number N increase.

Remark: Since the square wave system is linearly independent and complete, one can synthesize any signal with many square waves with different frequencies by a summation circuit. See Fig. 5. However, due to its nonorthogonality, one needs a special method when decomposing a signal into square waves.

III. BIORTHOGONAL FUNCTIONS AND SQUARE WAVE SERIES

Proposition 4: If $f(x)\in L^2(-\pi,\pi)$ has a square wave series:

$$f(x)=C_0+\sum_{n=1}^{\infty}(C(n)X(nx)+D(n)Y(nx)), \tag{30}$$

then all the coefficients $C_0, C(n), D(n)$ are unique.

Proof: Suppose $f(x)$ has another square wave series:

$$f(x)=C_0'+\sum_{n=1}^{\infty}(C'(n)X(nx)+D'(n)Y(nx)). \tag{31}$$

So,

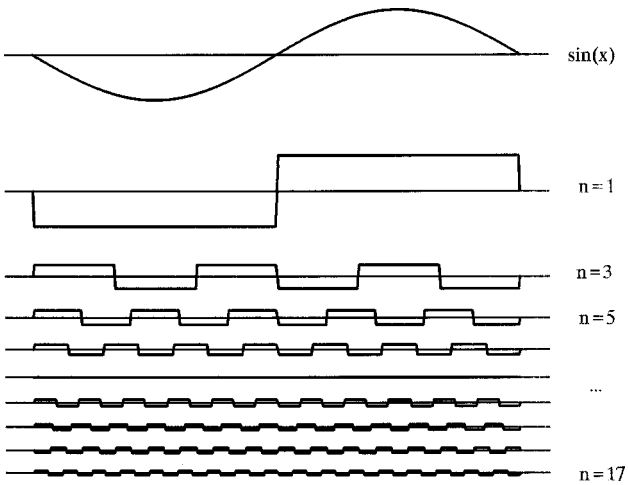

FIG. 4. The first few square wave components $\mu(n)B(n)Y(nx)$ contained in $\sin x$. ($n=1,3,5,7,9,11,13,15,17$.)

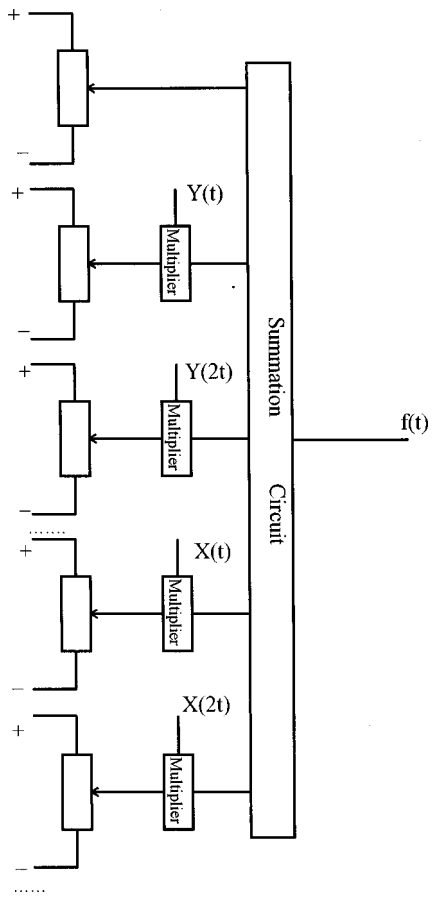

FIG. 5. Schematic diagram of synthesizing a time signal $f(t)$ by means of a series of square waves with different frequencies.

$$(C_0-C_0')+\sum_{n=1}^{\infty}((C(n)-C'(n))X(nx)+(D(n)-D'(n))Y(nx))=0.$$

Substituting their Fourier series for square waves $X(nx)$ and $Y(nx)$, one obtains that

$$a_0+\sum_{n=1}^{\infty}(a(n)\cos nx+b(n)\sin nx)=0, \tag{32}$$

where

$$a_0=C_0-C_0'=0,$$

$$a(n)=\sum_{d|n}A\left(\frac{n}{d}\right)(C(d)-C'(d))=0,$$

$$b(n)=\sum_{d|n}B\left(\frac{n}{d}\right)(D(d)-D'(d))=0.$$

Therefore (see Ref. 3, Chap, 2)

$$C_0-C_0'=0,$$

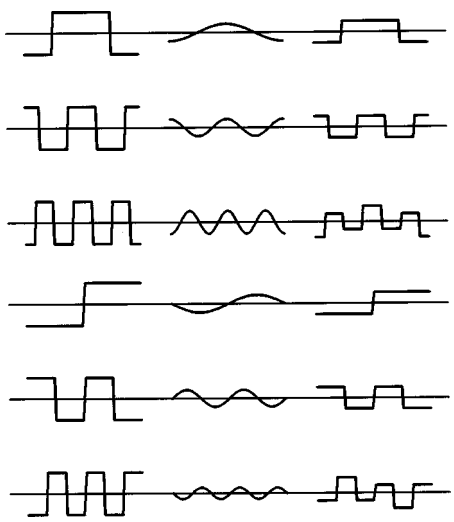

FIG. 6. Square waves, biorthogonal functions and orthonormalized functions.

$$C(n)-C'(n)=\sum_{d|n} 0\mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)=0,$$

$$D(n)-D'(n)=\sum_{d|n} 0\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)=0.$$

In other words, $C_0, C(n), D(n)$ are unique. □

Definition: For $n=1,2,3,\ldots$, the functions $h_n(x)$ and $g_n(x)$ are defined by

$$h_n(x)=\frac{1}{\pi}\sum_{d|n} \mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)\cos dx, \tag{33}$$

and

$$g_n(x)=\frac{1}{\pi}\sum_{d|n} \mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)\sin dx, \tag{34}$$

respectively.

Example: When $n=1,2,3$, we have

$$h_1(x)=\frac{1}{\pi}\cos x, \quad h_2(x)=\frac{1}{\pi}\cos 2x, \quad h_3(x)=\frac{1}{\pi}\left(\cos 3x+\frac{1}{3}\cos x\right),$$

and

$$g_1(x)=\frac{1}{\pi}\sin x, \quad g_2(x)=\frac{1}{\pi}\sin 2x, \quad g_3(x)=\frac{1}{\pi}\left(\sin 3x-\frac{1}{3}\sin x\right),$$

as shown in Fig. 6. In Fig. 6, the graphs of 18 functions are arranged in the following order:

$$\begin{pmatrix} X(x) & h_1(x) & E_1(x) \\ X(2x) & h_2(x) & E_2(x) \\ X(3x) & h_3(x) & E_3(x) \\ Y(x) & g_1(x) & e_1(x) \\ Y(2x) & g_2(x) & e_2(x) \\ Y(3x) & g_3(x) & e_3(x) \end{pmatrix}$$

Proposition 5: For any two positive integers m and n, we have

$$\langle X(mx), h_n(x)\rangle = \delta_{mn}, \tag{35}$$

$$\langle Y(mx), g_n(x)\rangle = \delta_{mn}, \tag{36}$$

where δ_{mn} is the Kronecher's delta symbol.

Proof:

$$\begin{aligned}
\langle Y(mx), g_n(x)\rangle &= \left\langle \sum_{k=1}^{\infty} B(k)\sin(mkx), \frac{1}{\pi}\sum_{d|n} B\left(\frac{n}{d}\right)\mu\left(\frac{n}{d}\right)\sin dx \right\rangle \\
&= \sum_{m|d,d|n} B\left(\frac{n}{d}\right)\mu\left(\frac{n}{d}\right)B\left(\frac{d}{m}\right) \\
&= \begin{cases} 0, & m \nmid n \\ B\left(\dfrac{n}{m}\right)\displaystyle\sum_{m|d,d|n}\mu\left(\frac{n}{d}\right) = \begin{cases} 0, & \dfrac{n}{m} = 2,3,\ldots \\ 1, & \dfrac{n}{m} = 1 \end{cases} \end{cases} \\
&= \begin{cases} 0, & m \neq n \\ 1, & m = n \end{cases} \\
&= \delta_{mn}.
\end{aligned}$$

Here $m \nmid n$ denotes that m is not a factor of n.

The other relation in this proposition can be proved similarly. □

Proposition 6: The square wave system

$$1, X(x), Y(x), X(2x), Y(2x), \ldots, X(nx), Y(nx), \ldots \tag{37}$$

and the function system

$$\frac{1}{2\pi}, h_1(x), g_1(x), h_2(x), g_2(x), \ldots, h_n(x), g_n(x), \ldots \tag{38}$$

are biorthogonal.

Proof: Due to Proposition 5 and the orthogonality of even functions and odd functions, this proposition is plain. □

Proposition 7: If a function $f(x) \in L^2(-\pi, \pi)$ has a square wave series:

$$f(x) = C_0 + \sum_{n=1}^{\infty} (C(n)X(nx) + D(n)Y(nx)), \tag{39}$$

then the coefficients in the square wave series can be determined by

$$C_0=\frac{1}{2\pi}\int_{-\pi}^{+\pi} f(x)dx, \tag{40}$$

$$C(n)=\int_{-\pi}^{+\pi} f(x)h_n(x)dx, \tag{41}$$

$$D(n)=\int_{-\pi}^{+\pi} f(x)g_n(x)dx. \tag{42}$$

Proof: By Proposition 6, this proposition is plain. □

Consequence of Proposition 7: If a function $f(x)\in L^2(-\pi,\pi)$ has a square wave series:

$$f(x)=C_0+\sum_{n=1}^{\infty}\,(C(n)X(nx)+D(n)Y(nx)), \tag{43}$$

and its Fourier series is written as

$$f(x)=a_0+\sum_{n=1}^{\infty}\,(a(n)\cos nx+b(n)\sin nx), \tag{44}$$

then we have

$$C_0=a_0,\quad C(n)=\sum_{d|n}\,\mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)a(d),\quad D(n)=\sum_{d|n}\,\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)b(d), \tag{45}$$

or inversely (Ref. 3, Chap. 2),

$$a_0=C_0,\quad a(n)=\sum_{d|n}\,A\left(\frac{n}{d}\right)C(d),\quad b(n)=\sum_{d|n}\,B\left(\frac{n}{d}\right)D(d). \tag{46}$$

Now a natural question is satisfying what conditions a function can be expanded as a square wave series. The following is a sufficient condition.

Proposition 8: Let the Fourier series of $f(x)\in L^2(-\pi,\pi)$ be

$$f(x)=a_0+\sum_{n=1}^{\infty}\,(a(n)\cos nx+b(n)\sin nx). \tag{47}$$

If the Fourier coefficients satisfy the condition

$$\sum_{n=1}^{\infty}\,a^2(n)d(n)<\infty, \tag{48}$$

$$\sum_{n=1}^{\infty}\,b^2(n)d(n)<\infty, \tag{49}$$

then we have

$$f(x)=C_0+\sum_{n=1}^{\infty}\,(C(n)X(nx)+D(n)Y(nx)), \tag{50}$$

where

$$C_0=a_0,\quad C(n)=\sum_{d|n}\,\mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)a(d),\quad D(n)=\sum_{d|n}\,\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)b(d). \tag{51}$$

Proof:

$$f(x)=a_0+\sum_{n=1}^{\infty}(a(n)\cos nx+b(n)\sin nx) \tag{52}$$

$$=a_0+\sum_{n=1}^{\infty}a(n)\sum_{l=1}^{\infty}A(l)\mu(l)X(nlx)+\sum_{n=1}^{\infty}b(n)\sum_{l=1}^{\infty}B(l)\mu(l)Y(nlx) \tag{53}$$

$$=a_0+\sum_{n=1}^{\infty}\left(\sum_{d|n}a(d)\mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)\right)X(nx)+\sum_{n=1}^{\infty}\left(\sum_{d|n}b(d)\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)\right)Y(nx) \tag{54}$$

$$=C_0+\sum_{n=1}^{\infty}(C(n)X(nx)+D(n)Y(nx)). \tag{55}$$

From (53) to (54), the summation orders were changed. But this does not matter because all the following series

$$\sum_{n=1}^{\infty}a(n)\sum_{l=1}^{\infty}A(l)\mu(l)X(nlx) \tag{56}$$

$$=\sum_{n=1}^{\infty}a(n)\sum_{l=1}^{\infty}A(l)\mu(l)\sum_{m=1}^{\infty}A(m)\cos(nlmx) \tag{57}$$

$$=\sum_{k=1}^{\infty}\left(\sum_{nlm=k}a(n)A(l)\mu(l)A(m)\right)\cos(kx) \tag{58}$$

and

$$\sum_{n=1}^{\infty}b(n)\sum_{l=1}^{\infty}B(l)\mu(l)Y(nlx) \tag{59}$$

$$=\sum_{n=1}^{\infty}b(n)\sum_{l=1}^{\infty}B(l)\mu(l)\sum_{m=1}^{\infty}B(m)\sin(nlmx) \tag{60}$$

$$=\sum_{k=1}^{\infty}\left(\sum_{nlm=k}b(n)B(l)\mu(l)B(m)\right)\sin(kx) \tag{61}$$

are unconditionally convergent in $L^2(-\pi,\pi)$, or one can say, they have sums independent of orders. The reason for the unconditional convergence is that the series

$$\sum_{k=1}^{\infty}\left(\sum_{nlm=k}|a(n)A(l)\mu(l)A(m)|\right)\cos(kx) \tag{62}$$

and

$$\sum_{k=1}^{\infty}\left(\sum_{nlm=k}|b(n)B(l)\mu(l)B(m)|\right)\sin(kx) \tag{63}$$

are convergent in $L^2(-\pi,\pi)$. As an example, the norm square of the series (62) is

$$\pi\sum_{k=1}^{\infty}\left(\sum_{nlm=k}|a(n)A(l)\mu(l)A(m)|\right)^2 \tag{64}$$

$$= \pi \sum_{k=1}^{\infty} \left(\sum_{nj=k} |a(n)||A(j)| \sum_{lm=j} |\mu(l)| \right)^2 \tag{65}$$

$$= \pi \sum_{k=1}^{\infty} \left(\sum_{nj=k} |a(n)|A'(j) \right)^2 \tag{66}$$

$$\leqslant \pi \sum_{k=1}^{\infty} d(k) \sum_{nj=k} a^2(n) A'^2(j) \tag{67}$$

$$\leqslant \pi \sum_{k=1}^{\infty} \sum_{nj=k} a^2(n) d(n) A'^2(j) d(j) \tag{68}$$

$$= \pi \sum_{n=1}^{\infty} a^2(n) d(n) \sum_{j=1}^{\infty} A'^2(j) d(j) \tag{69}$$

$$< \infty, \tag{70}$$

where

$$A'(j) = |A(j)| \sum_{l|j} |\mu(l)|. \tag{71}$$

From (67) to (68) we used (Ref. 2, Sec. 6.5)

$$d(nj) \leqslant d(n) d(j), \tag{72}$$

and from (69) to (70) we used

$$\sum_{j=1}^{\infty} A'^2(j) d(j) = \sum_{j=1}^{\infty} A^2(j) d(j) \left(\sum_{d|j} |\mu(d)| \right)^2$$

$$\leqslant \sum_{j=1}^{\infty} A^2(j) d^3(j) < \infty. \quad \text{(Ref. 2, Section 6.5)}$$

□

Definition: A subset K of $L^2(-\pi, \pi)$ is defined by

$$K = \left\{ f(x) \in L^2(-\pi, \pi) : f(x) = a_0 + \sum_{n=1}^{\infty} (a(n) \cos nx + b(n) \sin nx), \right.$$

$$\left. \text{where } \sum_{n=1}^{\infty} a^2(n) d(n) < \infty \text{ and } \sum_{n=1}^{\infty} b^2(n) d(n) < \infty \right\}.$$

Proposition 9: K is a linear subspace of $L^2(-\pi, \pi)$.
Proof: One can prove the following two items easily.
1^0. If $f(x) \in K$, then $cf(x) \in K$, where c is any real number.
2^0. If $f_1(x), f_2(x) \in K$, then $f_1(x) + f_2(x) \in K$. □
Proposition 10: The square wave system is a basis of the subspace K .
Proof: First of all, the square wave system is in K.
As an example, let us consider

$$Y(mx) = \sum_{l=1}^{\infty} B(l) \sin(lmx) = \sum_{n=1}^{\infty} b(n) \sin(nx). \tag{73}$$

It follows that

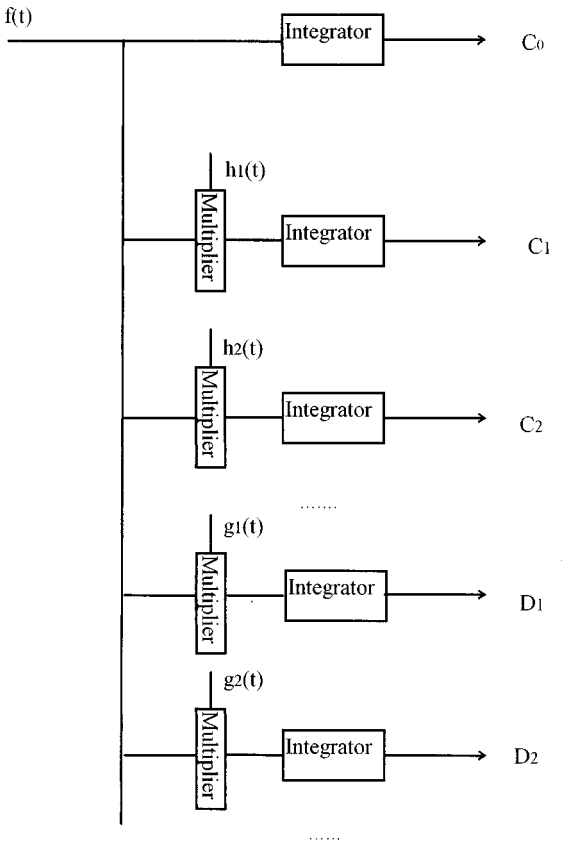

FIG. 7. Schematic diagram of decomposing a time signal into a series of square waves with different frequencies.

$$\sum_{n=1}^{\infty} b^2(n)d(n)=\sum_{l=1}^{\infty} B^2(l)d(ml)$$
$$\leqslant d(m)\sum_{l=1}^{\infty} B^2(l)d(l)$$
$$<\infty.$$

Second, by Proposition 4 and Proposition 8, the square wave system is a basis of the subspace K. □

Whether the square wave system is a basis of $L^2(-\pi,\pi)$ is still unknown.

Remark: A true signal always has little components with very high frequencies, so one can decompose any signal into the square wave series with the help of the biorthogonal functions, either by a computer program or by a specially designed device (see Fig. 7).

IV. ORTHOGONALIZATION AND THE BEST APPROXIMATION

Now that the square wave system is linearly independent, nonorthogonal and complete, one can get an orthonormal basis of $L^2(-\pi,\pi)$ by the Gram-Schmidt orthogonalization process.

Definition: The function $\beta(n)$ is defined on positive integers by

$$\beta(n)=\sum_{d|n} \mu(d)A^2(d)=\sum_{d|n} \mu(d)B^2(d). \tag{74}$$

For example, $\beta(1)=1$, $\beta(2)=1$, $\beta(3)=1-(\frac{1}{3})^2=\frac{8}{9}$.

Proposition 11: For $n=1,2,3,\ldots$, we have

$$\sum_{d|n} \beta(d)A^2\left(\frac{n}{d}\right)=\sum_{d|n} \beta(d)B^2\left(\frac{n}{d}\right)=1. \tag{75}$$

Proof:

$$\begin{aligned}\sum_{d|n} \beta(d)A^2\left(\frac{n}{d}\right)&=\sum_{d|n} A^2\left(\frac{n}{d}\right)\sum_{d'|d} \mu(d')A^2(d')\\&=\sum_{d|n}\sum_{d'|d} \mu(d')A^2\left(\frac{n}{d/d'}\right)\\&=\sum_{r|n} A^2\left(\frac{n}{r}\right)\sum_{d'|\frac{n}{r}} \mu(d')\\&=\sum_{r|n} A^2\left(\frac{n}{r}\right)\delta_{\frac{n}{r},1}\\&=A^2(1)=1.\end{aligned}$$

□

Definition: For $n=1,2,3,\ldots$, the functions $E_n(x)$ and $e_n(x)$ are defined by

$$E_n(x)=\frac{1}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n} \mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)X(dx), \tag{76}$$

and

$$e_n(x)=\frac{1}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n} \mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)Y(dx), \tag{77}$$

respectively.

Example: For $n=1,2,3$, we have

$$E_1(x)=\frac{1}{\sqrt{\pi^3/8}}X(x)\quad E_2(x)=\frac{1}{\sqrt{\pi^3/8}}X(2x)\quad E_3(x)=\frac{1}{\sqrt{\pi^3/9}}\left(X(3x)+\frac{1}{3}X(x)\right),$$

and

$$e_1(x)=\frac{1}{\sqrt{\pi^3/8}}Y(x)\quad e_2(x)=\frac{1}{\sqrt{\pi^3/8}}Y(2x)\quad e_3(x)=\frac{1}{\sqrt{\pi^3/9}}\left(Y(3x)-\frac{1}{3}Y(x)\right),$$

as shown in Fig. 6.

Proposition 12: Both $\{E_n(x)\}_{n=1}^{\infty}$ and $\{e_n(x)\}_{n=1}^{\infty}$ are orthonormal systems (ONS).

Proof:

$$\begin{aligned}&\langle E_m(x),E_n(x)\rangle\\&=\frac{1}{\pi^3/8}\frac{1}{\sqrt{\beta(m)}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|m} A\left(\frac{m}{d}\right)\mu\left(\frac{m}{d}\right)\sum_{d'|n} A\left(\frac{n}{d'}\right)\mu\left(\frac{n}{d'}\right)\langle X(dx),X(d'x)\rangle\\&=\frac{1}{\sqrt{\beta(m)}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|m} A\left(\frac{m}{d}\right)\mu\left(\frac{m}{d}\right)\sum_{d'|n} A\left(\frac{n}{d'}\right)\mu\left(\frac{n}{d'}\right)A\left(\frac{dd'}{(d,d')^2}\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{\beta(m)}} \frac{1}{\sqrt{\beta(n)}} \sum_{d|m} A\left(\frac{m}{d}\right) \mu\left(\frac{m}{d}\right) \sum_{d'|n} A\left(\frac{n}{d'}\right) \mu\left(\frac{n}{d'}\right) A\left(\frac{dd'}{(d,d')^2}\right) \sum_{l|(d,d')} \beta(l) A^2\left(\frac{(d,d')}{l}\right) \\
&= \frac{1}{\sqrt{\beta(m)}} \frac{1}{\sqrt{\beta(n)}} \sum_{d|m} A\left(\frac{m}{d}\right) \mu\left(\frac{m}{d}\right) \sum_{d'|n} A\left(\frac{n}{d'}\right) \mu\left(\frac{n}{d'}\right) \sum_{l|(d,d')} \beta(l) A\left(\frac{d}{l}\right) A\left(\frac{d'}{l}\right) \\
&= \frac{1}{\sqrt{\beta(m)}} \frac{1}{\sqrt{\beta(n)}} \sum_{d|m} A\left(\frac{m}{d}\right) \mu\left(\frac{m}{d}\right) \sum_{l|d} \beta(l) A\left(\frac{d}{l}\right) \sum_{l|d',d'|n} A\left(\frac{n}{d'}\right) A\left(\frac{d'}{l}\right) \mu\left(\frac{n}{d'}\right) \\
&= \frac{1}{\sqrt{\beta(m)}} \frac{1}{\sqrt{\beta(n)}} \sum_{d|m} A\left(\frac{m}{d}\right) \mu\left(\frac{m}{d}\right) \sum_{l|d} \beta(l) A\left(\frac{d}{l}\right) \delta_{nl} \\
&= \frac{1}{\sqrt{\beta(m)}} \frac{1}{\sqrt{\beta(n)}} \sum_{n|d,d|m} A\left(\frac{m}{d}\right) \mu\left(\frac{m}{d}\right) A\left(\frac{d}{n}\right) \beta(n) \\
&= \delta_{mn}.
\end{aligned}
$$

Similarly, one has

$$\langle e_m(x), e_n(x) \rangle = \delta_{mn}. \qquad \square$$

Proposition 13: The function system

$$\frac{1}{\sqrt{2\pi}}, E_1(x), e_1(x), E_2(x), e_2(x), \ldots, E_n(x), e_n(x), \ldots \tag{78}$$

is an orthonormal basis of $L^2(-\pi, \pi)$.

Proof: The square wave system is linearly independent and complete, so its orthonormalized function system is an orthonormal basis of $L^2(-\pi, \pi)$. $\square$

Consequence of Proposition 13: Any function $f(x) \in L^2(-\pi, \pi)$ can be written as the following series:

$$f(x) = r_0 + \sum_{n=1}^{\infty} (r(n) E_n(x) + q(n) e_n(x)), \tag{79}$$

where

$$r_0 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} f(x)\, dx, \tag{80}$$

$$r(n) = \int_{-\pi}^{+\pi} f(x) E_n(x)\, dx, \tag{81}$$

$$q(n) = \int_{-\pi}^{+\pi} f(x) e_n(x)\, dx. \tag{82}$$

Example:

$$\sin x = \frac{\pi}{\sqrt{\pi^3/8}} \sum_{n=1}^{\infty} \frac{1}{\sqrt{\beta(n)}} \mu(n) B(n) e_n(x), \tag{83}$$

$$\cos x=\frac{\pi}{\sqrt{\pi^3/8}}\sum_{n=1}^{\infty}\frac{1}{\sqrt{\beta(n)}}\mu(n)A(n)E_n(x). \tag{84}$$

Proof of example:

It is because

$$\langle \sin x,e_n(x)\rangle=\frac{1}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n}\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)\langle \sin x,Y(dx)\rangle$$

$$=\frac{\pi}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\mu(n)B(n)$$

and

$$\langle \cos x,E_n(x)\rangle=\frac{\pi}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\mu(n)A(n). \qquad \square$$

Next let us consider how to approximate a given function $f(x)\in L^2(-\pi,\pi)$ by a superposition of finite square waves.

Definition: Let N be a given positive integer. For $n=1,2,3,\ldots,N$, one defines

$$h_n^{(N)}(x)=\frac{1}{\pi^3/8}\sum_{m=1}^{N}\left(\sum_{m|k,n|k}^{1\leqslant k\leqslant N}\frac{1}{\beta(k)}A\left(\frac{k}{m}\right)A\left(\frac{k}{n}\right)\mu\left(\frac{k}{m}\right)\mu\left(\frac{k}{n}\right)\right)X(mx) \tag{85}$$

$$g_n^{(N)}(x)=\frac{1}{\pi^3/8}\sum_{m=1}^{N}\left(\sum_{m|k,n|k}^{1\leqslant k\leqslant N}\frac{1}{\beta(k)}B\left(\frac{k}{m}\right)B\left(\frac{k}{n}\right)\mu\left(\frac{k}{m}\right)\mu\left(\frac{k}{n}\right)\right)Y(mx). \tag{86}$$

Proposition 14: The best approximation of a function $f(x)\in L^2(-\pi,\pi)$ by a superposition of the finite square waves

$$1,X(x),Y(x),X(2x),Y(2x),\ldots,X(Nx),\quad \text{and } Y(Nx) \tag{87}$$

is

$$f^{(N)}(x)=C_0+\sum_{n=1}^{N}(C^{(N)}(n)X(nx)+D^{(N)}(n)Y(nx)), \tag{88}$$

where

$$C_0=\frac{1}{2\pi}\int_{-\pi}^{+\pi}f(x)dx, \tag{89}$$

$$C^{(N)}(n)=\int_{-\pi}^{+\pi}f(x)h_n^{(N)}(x)dx, \tag{90}$$

$$D^{(N)}(n)=\int_{-\pi}^{+\pi}f(x)g_n^{(N)}(x)dx. \tag{91}$$

Proof: $f(x)\in L^2(-\pi,\pi)$ can be written as

$$f(x)=r_0+\sum_{n=1}^{\infty}(r(n)E_n(x)+q(n)e_n(x)), \tag{92}$$

where

$$r_0=\frac{1}{2\pi}\int_{-\pi}^{+\pi}f(x)dx,$$

$$r(n)=\int_{-\pi}^{+\pi}f(x)E_n(x)dx, \tag{93}$$

$$q(n)=\int_{-\pi}^{+\pi}f(x)e_n(x)dx.$$

The best approximation to $f(x)$ out of the subspace spanned by

$$1,X(x),Y(x),X(2x),Y(2x),\ldots,X(Nx), \quad \text{and } Y(Nx)$$

is

$$\begin{aligned}
f^{(N)}(x)&=r_0+\sum_{n=1}^{N}(r(n)E_n(x)+q(n)e_n(x))\\
&=r_0+\sum_{n=1}^{N}\left[r(n)\frac{1}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n}\mu\left(\frac{n}{d}\right)A\left(\frac{n}{d}\right)X(dx)\right.\\
&\quad\left.+q(n)\frac{1}{\sqrt{\pi^3/8}}\frac{1}{\sqrt{\beta(n)}}\sum_{d|n}\mu\left(\frac{n}{d}\right)B\left(\frac{n}{d}\right)Y(dx)\right]\\
&=C_0+\sum_{n=1}^{N}(C^{(N)}(n)X(nx)+D^{(N)}(n)Y(nx)),
\end{aligned}$$

where

$$\begin{aligned}
C^{(N)}(n)&=\frac{1}{\sqrt{\pi^3/8}}\sum_{s=1}^{sn\leqslant N}\mu(s)A(s)r(sn)\frac{1}{\sqrt{\beta(sn)}}\\
&=\frac{1}{\sqrt{\pi^3/8}}\sum_{s=1}^{sn\leqslant N}\mu(s)A(s)\frac{1}{\sqrt{\beta(sn)}}\int_{-\pi}^{+\pi}f(x)E_{sn}(x)dx\\
&=\int_{-\pi}^{+\pi}f(x)\ \frac{1}{\sqrt{\pi^3/8}}\sum_{s=1}^{sn\leqslant N}\mu(s)A(s)\frac{1}{\sqrt{\beta(sn)}}E_{sn}(x)dx\\
&=\int_{-\pi}^{+\pi}f(x)\frac{1}{\pi^3/8}\sum_{s=1}^{sn\leqslant N}\mu(s)A(s)\frac{1}{\beta(sn)}\sum_{d|sn}\mu\left(\frac{sn}{d}\right)A\left(\frac{sn}{d}\right)X(dx)dx\\
&=\int_{-\pi}^{+\pi}f(x)\frac{1}{\pi^3/8}\sum_{m=1}^{N}\left(\sum_{m|k,n|k}^{1\leqslant k\leqslant N}\frac{1}{\beta(k)}A\left(\frac{k}{m}\right)A\left(\frac{k}{n}\right)\mu\left(\frac{k}{m}\right)\mu\left(\frac{k}{n}\right)\right)X(mx)dx\\
&=\int_{-\pi}^{+\pi}f(x)h_n^{(N)}(x)dx,
\end{aligned}$$

and similarly

$$D^{(N)}(n)=\int_{-\pi}^{+\pi}f(x)g_n^{(N)}(x)dx. \qquad \square$$

In fact, $\{h_n^{(N)}(x)\}_{n=1}^N$ are the biorthogonal functions of $\{X(nx)\}_{n=1}^N$ in the subspace spanned by $\{X(nx)\}_{n=1}^N$, and $\{g_n^{(N)}(x)\}_{n=1}^N$ are the biorthogonal functions of $\{Y(nx)\}_{n=1}^N$ in the subspace spanned by $\{Y(nx)\}_{n=1}^N$.

Example: When $N=3$, we have

$$h_1^{(3)}(x)=\frac{1}{\pi^3}(9X(x)+3X(3x)),\quad h_2^{(3)}(x)=\frac{1}{\pi^3}8X(2x),$$

$$h_3^{(3)}(x)=\frac{1}{\pi^3}(3X(x)+9X(3x)),\quad g_1^{(3)}(x)=\frac{1}{\pi^3}(9Y(x)-3Y(3x)),$$

$$g_2^{(3)}(x)=\frac{1}{\pi^3}8Y(2x),\quad g_3^{(3)}(x)=\frac{1}{\pi^3}(-3Y(x)+9Y(3x)).$$

The best approximations of sin$x$ and cos$x$ are

$$\begin{aligned}\sin^{(3)}x&=\frac{\pi}{\sqrt{\pi^3/8}}\sum_{n=1}^{3}\frac{1}{\sqrt{\beta(n)}}\mu(n)B(n)e_n(x)\\&=\frac{1}{\pi^2}(9Y(x)-3Y(3x)),\end{aligned}$$

$$\begin{aligned}\cos^{(3)}x&=\frac{\pi}{\sqrt{\pi^3/8}}\sum_{n=1}^{3}\frac{1}{\sqrt{\beta(n)}}\mu(n)A(n)E_n(x)\\&=\frac{1}{\pi^2}(9X(x)+3X(3x)).\end{aligned}$$

If square waves are generated from a common high-frequency pulse signal by frequency-division technique, we need consider a function $f(x)$'s best approximation by a superposition of the finite square waves 1, $\{X(dx)\}_{d|N_1}$and $\{X(d'x)\}_{d'|N_2}$, where N_1 and N_2 are two positive integers.

Proposition 15: Let N_1 and N_2 be two positive integers and $f(x)\in L^2(-\pi,\pi)$. Then $f(x)$'s best approximation by a linear combination of 1, $\{X(dx)\}_{d|N_1}$ and $\{X(d'x)\}_{d'|N_2}$, is

$$f^{[N_1,N_2]}(x)=C_0+\sum_{d|N_1}C^{[N_1]}(d)X(dx)+\sum_{d'|N_2}D^{[N_2]}(d')Y(d'x).\tag{94}$$

The coefficients in (94) are determined by

$$C_0=\frac{1}{2\pi}\int_{-\pi}^{+\pi}f(x)dx,\tag{95}$$

$$C^{[N_1]}(d)=\int_{-\pi}^{+\pi}f(x)h_d^{[N_1]}(x)dx,\tag{96}$$

$$D^{[N_2]}(d')=\int_{-\pi}^{+\pi}f(x)g_{d'}^{[N_2]}(x)dx,\tag{97}$$

where

$$h_d^{[N_1]}(x)=\frac{1}{\pi^3/8}\sum_{m|N_1}\left(\sum_{m|k,d|k}^{k|N_1}\frac{1}{\beta(k)}A\left(\frac{k}{m}\right)A\left(\frac{k}{d}\right)\mu\left(\frac{k}{m}\right)\mu\left(\frac{k}{d}\right)\right)X(mx),\tag{98}$$

$$g_{d'}^{[N_2]}(x)=\frac{1}{\pi^3/8}\sum_{m'|N_2}\left(\sum_{m'|k,d'|k}^{k|N_2}\frac{1}{\beta(k)}B\left(\frac{k}{m'}\right)B\left(\frac{k}{d'}\right)\mu\left(\frac{k}{m'}\right)\mu\left(\frac{k}{d'}\right)\right)Y(m'x). \tag{99}$$

Proof: It is plain that $f(x)$'s best approximation in this case is

$$f^{[N_1,N_2]}(x)=r_0+\sum_{d|N_1}r(d)E_d(x)+\sum_{d'|N_2}q(d')e_{d'}(x). \tag{100}$$

The rest of the proof is similar to that of proposition 14. □

Example: $N_1=N_2=6$.

The best approximations of sin$x$,cos$x$ by a superposition of the square waves 1, $Y(x),Y(2x),Y(3x),Y(6x)$ and $X(x),X(2x),X(3x),X(6x)$ are

$$\begin{aligned}\sin^{[6,6]}x&=\frac{\pi}{\sqrt{\pi^3/8}}\sum_{d|6}\frac{1}{\sqrt{\beta(d)}}\mu(d)B(d)e_d(x)\\&=\frac{1}{\pi^2}(9Y(x)-3Y(3x)),\end{aligned}$$

$$\begin{aligned}\cos^{[6,6]}x&=\frac{\pi}{\sqrt{\pi^3/8}}\sum_{d'|6}\frac{1}{\sqrt{\beta(d')}}\mu(d')A(d')E_{d'}(x)\\&=\frac{1}{\pi^2}(9X(x)+3X(3x)).\end{aligned}$$

Remark: Since the orthogonalized square waves are none other than combinations of several square waves, one can generate them easily as well, so they can be regarded as a new basic waveform in electronics (see Fig. 6). In applications we have to express a signal as a superposition of finite square waves rather than infinite square waves, since the number of square waves is finite after all, and therefore the best approximation is an important practical problem.

V. CONCLUSION

Since the square wave system is linearly independent and complete, one can synthesize any signal by a summation circuit of square waves with different frequencies (see Fig. 5). By means of biorthogonal functions one can decompose any signal into a square wave series with a computer or a specially designed device (see Fig. 7). With the orthogonalized square waves, a signal can be easily represented as a superposition of finite square waves with the least mean-square error. These results form the theory basis for the square wave analysis technique in modern electronics.

We will discuss other common waveforms later (Ref. 4).

ACKNOWLEDGMENTS

The authors would like to thank Hui Li at Yale University, and Wenxiang Cong at Beijing University of Science and Technology, for their help.

[1] K.G. Beauchamp, *Walsh Functions and their Applications* (Academic, London, 1975).
[2] L. K. Hua, *Introduction to Number Theory* (Springer-Verlag, Berlin, 1982).
[3] T. M. Apostol, *Introduction to Analytic Number Theory* (Springer-Verlag, New York, 1976).
[4] Y. Wei, ''Easily generated function analysis,'' Sci. D. thesis, Beijing University of Science and Technology (1988).

On the number of bound states for the one-dimensional Schrödinger equation

Tuncay Aktosun[a)]
Department of Mathematics, North Dakota State University, Fargo, North Dakota 58105

Martin Klaus[b)]
Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Cornelis van der Mee[c)]
Department of Mathematics, University of Cagliari, Cagliari, Italy

(Received 13 August 1997; accepted for publication 19 May 1998)

The number of bound states of the one-dimensional Schrödinger equation is analyzed in terms of the number of bound states corresponding to ''fragments'' of the potential. When the potential is integrable and has a finite first moment, the sharp inequalities $1-p+\Sigma_{j=1}^{p} N_j \leqslant N \leqslant \Sigma_{j=1}^{p} N_j$ are proved, where p is the number of fragments, N is the total number of bound states, and N_j is the number of bound states for the jth fragment. When $p=2$ the question of whether $N=N_1+N_2$ or $N=N_1+N_2-1$ is investigated in detail. An illustrative example is also provided. © *1998 American Institute of Physics.* [S0022-2488(98)03109-0]

I. INTRODUCTION

Consider the one-dimensional Schrödinger equation

$$\psi''(k,x)+k^2\psi(k,x)=V(x)\psi(k,x), \tag{1.1}$$

where the potential V is real valued and belongs to $L_1^1(\mathbf{R})$, the class of measurable functions for which $\int_{-\infty}^{\infty} dx(1+|x|)|V(x)|$ is finite. The prime denotes the derivative with respect to the spatial coordinate x. Let us partition the real axis as $\mathbf{R}=\cup_{j=1}^{p}(x_{j-1},x_j)$, with $x_{j-1}<x_j$ for $j=1,...,p$. Here we use the convention $x_0=-\infty$ and $x_p=+\infty$. We obtain a fragmentation of the potential by setting $V(x)=\Sigma_{j=1}^{p} V_j(x)$, where

$$V_j(x)=\begin{cases} V(x), & x\in(x_{j-1},x_j) \\ 0, & \text{elsewhere.} \end{cases} \tag{1.2}$$

In this paper we analyze the relationship between the number of bound states of V and the number of bound states of its fragments. In Sec. II we prove a pair of sharp inequalities relating these numbers (Theorem 2.1); we also study the case $p=2$ in more detail, and in Theorems 2.2 and 2.3 we present criteria that tell us when $N=N_1+N_2$ or $N=N_1+N_2-1$. In Sec. III we give another proof of Theorem 2.1 by using a factorization formula for the scattering matrix and exploiting its small-k asymptotics. We also briefly discuss what happens if we increase the separation distance between two fragments (Theorem 3.1). In Sec. IV we give an example which illustrates various aspects of our results.

The inequality (2.5) in Theorem 2.1 has been proved before by different methods and under stronger assumptions on the potential. In Ref. 1, (2.5) was proved when $p=2$ and the fragments have compact support. In Refs. 2 and 3, some special cases of (2.5) were proved for parity invariant, compactly supported fragments, but, as already mentioned in those references, the parity invariance is not an essential restriction. The method used in Ref. 1 was based on the nodal properties of the zero-energy solutions of the Schrödinger equation but was fairly contrived, while

[a)]Electronic mail: aktosun@plains.nodak.edu
[b)]Electronic mail: klaus@math.vt.edu
[c)]Electronic mail: cornelis@krein.unica.it

 © 1998 American Institute of Physics

the method used in Refs. 2 and 3 relied on a factorization formula[4] for the scattering matrix and the small-k behavior of the scattering coefficients. In the proofs of Theorems 2.1 and 2.2 we use certain properties of the Jost solutions, especially the interlacing property of zeros, in a very straightforward way. As a result, we are able to establish the connection with the factorization method used in Sec. III. Furthermore, no additional technical restrictions are imposed on the potential besides $V \in L_1^1(\mathbf{R})$.

At various places in this paper we need to distinguish between ''generic'' and ''exceptional'' potentials. Recall that a potential is called generic if the corresponding transmission coefficient T vanishes at $k=0$, and that a potential is called exceptional if $T(0)\neq 0$. Equivalently, a potential is generic (exceptional) if for $k=0$ the two Jost solutions are linearly independent (dependent)[5–7].

II. INEQUALITY FOR THE NUMBER OF BOUND STATES

In preparation of this section we first establish some notation and collect some results about the Jost solutions and their nodal properties. Let $f_{l;j}(k,x)$ and $f_{r;j}(k,x)$ denote the Jost solutions from the left and right, respectively, for the fragment V_j. Recall that $f_{l;j}(k,x)=e^{ikx}[1+o(1)]$ as $x\to+\infty$ and $f_{r;j}(k,x)=e^{-ikx}[1+o(1)]$ as $x\to-\infty$. Furthermore, let n_j denote the number of zeros of $f_{r;j}(0,x)$ lying in $(-\infty,x_j)$, m_j the number of zeros of $f_{l;j}(0,x)$ lying in $(x_{j-1},+\infty)$, and N_j the number of bound states of the fragment V_j. Since N_j is equal[8,9] to the number of zeros of either $f_{l;j}(0,x)$ or $f_{r;j}(0,x)$, we conclude that

$$N_j=\begin{cases} n_j, & \text{if } f_{r;j}(0,x_j)f'_{r;j}(0,x_j)\geqslant 0 \quad \text{and} \quad f_{r;j}(0,x_j)\neq 0 \\ n_j+1, & \text{if } f_{r;j}(0,x_j)f'_{r;j}(0,x_j)\leqslant 0 \quad \text{and} \quad f'_{r;j}(0,x_j)\neq 0, \end{cases} \tag{2.1}$$

$$N_j=\begin{cases} m_j, & \text{if } f_{l;j}(0,x_{j-1})f'_{l;j}(0,x_{j-1})\leqslant 0 \quad \text{and} \quad f_{l;j}(0,x_{j-1})\neq 0 \\ m_j+1, & \text{if } f_{l;j}(0,x_{j-1})f'_{l;j}(0,x_{j-1})\geqslant 0 \quad \text{and} \quad f'_{l;j}(0,x_{j-1})\neq 0. \end{cases} \tag{2.2}$$

Note that on $(x_j,+\infty)$ the function $f_{r;j}(0,x)$ is equal to $f'_{r;j}(0,x_j)(x-x_j)+f_{r;j}(0,x_j)$ and that this linear function has the root $x=x_j-f_{r;j}(0,x_j)/f'_{r;j}(0,x_j)$ which lies in $[x_j,+\infty)$ precisely if $f_{r;j}(0,x_j)=0$ or $f_{r;j}(0,x_j)f'_{r;j}(0,x_j)<0$; in this case we have $N_j=n_j+1$. On the other hand, if $f'_{r;j}(0,x_j)=0$ or $f_{r;j}(0,x_j)f'_{r;j}(0,x_j)>0$, then $f_{r;j}(0,x)$ has no zeros in $[x_j,+\infty)$, i.e., all its zeros are in $(-\infty,x_j)$; thus $N_j=n_j$. This proves (2.1). We obtain (2.2) by applying a similar argument to $f_{l;j}(0,x)$. We will also need the Jost solutions for the potential V, which we denote by $f_l(k,x)$ and $f_r(k,x)$, respectively. In the generic case when $k=0$ the following asymptotic relations hold[10] as $x\to+\infty$:

$$f_l(0,x)=1+o(1), \quad f'_l(0,x)=o(1/x), \tag{2.3}$$

$$f_r(0,x)=c_r x+o(x), \quad f'_r(0,x)=c_r+o(1), \tag{2.4}$$

with some constant $c_r\neq 0$.

Theorem 2.1: Suppose that $V\in L_1^1(\mathbf{R})$. Let N denote the number of bound states of V. Then

$$1-p+\sum_{j=1}^{p} N_j \leqslant N \leqslant \sum_{j=1}^{p} N_j, \quad p=1,2,..., \tag{2.5}$$

where both inequalities are sharp.

Proof: It suffices to prove (2.5) for $p=2$ because the general case follows by induction. Let $u(x)$ denote the solution of (1.1) for $k=0$ satisfying the initial conditions $u(x_1)=1$ and $u'(x_1)=0$. Then $u(x)=f_{r;2}(0,x)$ on $x\geqslant x_1$ and $u(x)=f_{l;1}(0,x)$ on $x\leqslant x_1$. Hence $u(x)$ has N_1 zeros on $(-\infty,x_1)$ and N_2 zeros on $(x_1,+\infty)$, i.e., N_1+N_2 zeros in all. Hence, by the interlacing property of zeros, $f_l(0,x)$ has either N_1+N_2 or N_1+N_2-1 zeros. This proves (2.5). To see that the inequalities are sharp, note that a square-well potential of depth $-H^2$ and width w has exactly N bound states, where N is the positive integer satisfying $(N-1)\pi<wH\leqslant N\pi$. Choosing V to be a square-well potential of depth $-\pi^2$ with support $(0,1)$, we obtain $N=1$. Let us partition the interval $(0,1)$ into p nonempty subintervals and hence obtain a fragmentation of V; each fragment still contains exactly one bound state and hence the lower bound in (2.5) becomes equal to N. On

the other hand, consider the square-well potential of depth $-\pi^2$ with support $(0,p)$, and partition $(0,p)$ into the p subintervals $(j-1,j)$ for $j=1,...,p$. Then $N_j=1$, $N=p$, and hence the upper bound in (2.5) becomes equal to N. ■

We remark that the short proof of (2.5) given here was suggested by the referee. Inequality (2.5) also follows from the next theorem that gives us, in case of two fragments, the precise information on whether $N=N_1+N_2$ or $N=N_1+N_2-1$. Let

$$Z(x_1)=\frac{f_{l;2}(0,x_1)}{f'_{l;2}(0,x_1)}-\frac{f_{r;1}(0,x_1)}{f'_{r;1}(0,x_1)}. \tag{2.6}$$

Theorem 2.2: Assume that V is partitioned into two fragments. Then:
(a) If $f'_{r;1}(0,x_1)\neq 0$, $f'_{l;2}(0,x_1)\neq 0$, and $Z(x_1)\geqslant 0$, then $N=N_1+N_2-1$; if $Z(x_1)<0$, then $N=N_1+N_2$.
(b) If $f'_{r;1}(0,x_1)=0$ or $f'_{l;2}(0,x_1)=0$, then $N=N_1+N_2$.

Proof: (a) In order to determine N we will count the zeros of $f_r(0,x)$ that lie in $[x_1,+\infty)$. We do this by using the interlacing property of the zeros of $f_r(0,x)$ and $f_l(0,x)$, noting that $f_l(0,x)=f_{l;2}(0,x)$ on $[x_1,+\infty)$ and $f_r(0,x)=f_{r;1}(0,x)$ on $(-\infty,x_1]$. We already know that n_1 zeros of $f_r(0,x)$ lie in $(-\infty,x_1)$, where n_1 is related to N_1 by (2.1). Upon multiplying $f_{l;2}(0,x)$ and $f_{r;1}(0,x)$ by suitable constants α and β, we can achieve that $\varphi_{l;2}(0,x)=\alpha f_{l;2}(0,x)$ and $\varphi_{r;1}(0,x)=\beta f_{r;1}(0,x)$ satisfy $\varphi'_{l;2}(0,x_1)=\varphi'_{r;1}(0,x_1)=1>0$. Then $Z(x_1)$ in (2.6) becomes

$$Z(x_1)=\varphi_{l;2}(0,x_1)-\varphi_{r;1}(0,x_1).$$

First suppose that $Z(x_1)>0$, which is equivalent to assuming $W[\varphi_l,\varphi_r](x_1)>0$, where $W[g,h](x)=g(x)h'(x)-g'(x)h(x)$ denotes the Wronskian. We first consider the case when $\varphi_{l;2}(0,x)$ has at least one zero on $(x_1,+\infty)$. Suppose that $\varphi_{l;2}(0,x)$ has its zeros at z_j for $j=1,...,m_2$, where $x_1<z_1<z_2<\cdots<z_{m_2}$. If $\varphi_{l;2}(0,x_1)>\varphi_{r;1}(0,x_1)>0$, then $\varphi_{l;2}(0,x)$ has m_2 zeros in $(x_1,+\infty)$ because, by a Wronskian argument, there are no zeros in (x_1,z_1) and there is exactly one zero in each of the intervals (z_1,z_2), (z_2,z_3),...,$(z_{m_2},+\infty)$. To see that there is a zero in $(z_{m_2},+\infty)$, note that by (2.3) and (2.4), $W[\varphi_l,\varphi_r](x)=\alpha\beta c_r>0$. Hence α and βc_r have the same sign. Moreover, if $\alpha>0$, then $\varphi'_{l;2}(0,z_{m_2})>0$ and hence $\varphi_r(0,z_{m_2})<0$. Similarly, if $\alpha<0$, then $\varphi'_{l;2}(0,z_{m_2})<0$ and hence $\varphi_r(0,z_{m_2})>0$. Because $\varphi_r(0,x)=\beta c_r x+o(x)$ as $x\to+\infty$, it follows that $\varphi_r(0,x)$ must have a zero in $(x_1,+\infty)$ and, by the interlacing property, this is the only zero on this interval. Hence, using (2.1) and (2.2), we have $n_1=N_1$, $m_2=N_2-1$, and $N=n_1+m_2=N_1+N_2-1$. The same result holds when $\varphi_{l;2}(0,x)$ has no zeros on $(x_1,+\infty)$. Then $\varphi_r(0,x)$ has no zeros on $(x_1,+\infty)$ either and we have $m_2=0$. If $\varphi_{l;2}(0,x_1)>\varphi_{r;1}(0,x_1)=0$, then the previous argument goes through with only a minor change in counting the zeros because now $\varphi_r(0,x)$ also has a zero at $x=x_1$. We have $n_1=N_1-1$, $m_2=N_2-1$, and $N=n_1+m_2+1=N_1+N_2-1$. If $\varphi_{l;2}(0,x_1)\geqslant 0>\varphi_{r;1}(0,x_1)$, then $\varphi_r(0,x)$ has m_2+1 zeros on $(x_1,+\infty)$ because now there is also a zero in (x_1,z_1). Thus $n_1=N_1-1$, $m_2=N_2-1$, and $N=n_1+m_2+1=N_1+N_2-1$. If $0>\varphi_{l;2}(0,x_1)>\varphi_{r;1}(0,x_1)$, then $n_1=N_1-1$, $m_2=N_2$, and $N=N_1+N_2-1$ because $\varphi_r(0,x)$ has no zeros in (x_1,z_1). All the possibilities with $Z(x_1)>0$ have now been exhausted. If $Z(x_1)<0$, we can apply similar arguments and find that $N=N_1+N_2$. Finally, if $Z(x_1)=0$ because $\varphi_{l;2}(0,x_1)=\varphi_{r;1}(0,x_1)>0$, then $n_1=N_1$, $m_2=N_2-1$, and $N=n_1+m_2=N_1+N_2-1$. If $Z(x_1)=0$ because $\varphi_{l;2}(0,x_1)=\varphi_{r;1}(0,x_1)=0$, then $n_1=N_1-1$, $m_2=N_2-1$, and $N=n_1+m_2+1=N_1+N_2-1$. If $Z(x_1)=0$ because $\varphi_{l;2}(0,x_1)=\varphi_{r;1}(0,x_1)<0$, then $n_1=N_1-1$, $m_2=N_2$, and $N=n_1+m_2=N_1+N_2-1$. This concludes the proof of part (a). The proof of (b) is similar, using (2.1), (2.2), and the Wronskian. The details are omitted. ■

Theorem 2.3: Assume that V is partitioned into two fragments, and let $W(x_1)$ denote the Wronskian $W[f_{l;2}(0,\cdot);f_{r;1}(0,\cdot)](x_1)$. Then:
(i) Suppose N_1 and N_2 are either both even or both odd. If $W(x_1)>0$ $(W(x_1)<0)$, then $N=N_1+N_2$ $(N=N_1+N_2-1)$.
(ii) Suppose N_1 is even and N_2 is odd or vice versa. If $W(x_1)>0$ $(W(x_1)<0)$, then $N=N_1+N_2-1$ $(N=N_1+N_2)$.

Proof: The proof is a consequence of the following observation. If $W(x_1)>0$, then the constant c_r in (2.4) is positive and so N is even, while if $W(x_1)<0$, then c_r is negative and N is odd. ■

If $W(x_1)=0$, then Theorem 2.3 gives no information as to which possibility is realized. However, Theorem 2.2 (a) says that if $W(x_1)=0$ because $Z(x_1)=0$, then $N=N_1+N_2-1$. The only other possibility is that $W(x_1)=0$ because $f'_{r;1}(0,x_1)=f_{l;2}(0,x_1)=0$, in which case $N=N_1+N_2$ by Theorem 2.2 (b).

III. FURTHER OBSERVATIONS

In this section we analyze the result of Theorem 2.1 in conjunction with the scattering matrices corresponding to the fragments of this potential. For simplicity let us consider the fragmentation of V as $V=V_1+V_2$, where V_1 has support in $(-\infty,x_1]$ and V_2 has support in $[x_1,+\infty)$. The analysis for three or more fragments can be carried out by using induction. Let $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}$ be the scattering matrices corresponding to the potentials V_1, V_2, and V, respectively. The scattering coefficients appear in the scattering matrix as follows:

$$\mathbf{S}(k)=\begin{bmatrix} T(k) & R(k) \\ L(k) & T(k) \end{bmatrix}, \tag{3.1}$$

where T is the transmission coefficient, and L and R are the reflection coefficients from the left and from the right, respectively. Similarly, T_j, R_j, and L_j denote the corresponding entries of $\mathbf{S}_j$ for $j=1,2$. Let us define the so-called transition matrix associated with $\mathbf{S}$ as follows:

$$\Lambda(k)=\begin{bmatrix} \dfrac{1}{T(k)} & -\dfrac{R(k)}{T(k)} \\ \dfrac{L(k)}{T(k)} & \dfrac{1}{T(k)^*} \end{bmatrix}=\begin{bmatrix} \dfrac{1}{T(k)} & \dfrac{L(k)^*}{T(k)^*} \\ \dfrac{L(k)}{T(k)} & \dfrac{1}{T(k)^*} \end{bmatrix}, \tag{3.2}$$

where the asterisk denotes complex conjugation. Similarly, let Λ_1 and Λ_2 be the transition matrices corresponding to $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. It is known[4] that

$$\Lambda(k)=\Lambda_1(k)\Lambda_2(k). \tag{3.3}$$

From the (1,1) entry of the matrix product in (3.3) we get

$$\frac{1}{T(k)}=\frac{1-R_1(k)L_2(k)}{T_1(k)T_2(k)}. \tag{3.4}$$

Let $\mathbf{R}^+=(0,+\infty)$. For $k\in\mathbf{R}^+$, let us define the phases $\phi(k)$, $\phi_1(k)$, and $\phi_2(k)$ of the transmission coefficients as follows:

$$T(k)=|T(k)|e^{i\phi(k)}, \quad T_1(k)=|T_1(k)|e^{i\phi_1(k)}, \quad T_2(k)=|T_2(k)|e^{i\phi_2(k)}, \tag{3.5}$$

where it is understood that ϕ, ϕ_1, and ϕ_2 are continuous in $k\in\mathbf{R}^+$ and normalized such that

$$\phi(+\infty)=\phi_1(+\infty)=\phi_2(+\infty)=0. \tag{3.6}$$

Similarly, let

$$1-R_1(k)L_2(k)=|1-R_1(k)L_2(k)|e^{i\omega(k)}, \tag{3.7}$$

where ω is assumed continuous in $k\in\mathbf{R}^+$ and to satisfy $\omega(+\infty)=0$. From (3.4) we obtain

$$\phi(k)=\phi_1(k)+\phi_2(k)-\omega(k), \quad k\in\mathbf{R}^+. \tag{3.8}$$

From Levinson's theorem[11] we have

$$\phi(0+)=\left[N-\frac{d}{2}\right]\pi, \quad \phi_1(0+)=\left[N_1-\frac{d_1}{2}\right]\pi, \quad \phi_2(0+)=\left[N_2-\frac{d_2}{2}\right]\pi, \tag{3.9}$$

where N, N_1, and N_2 denote the number of bound states corresponding to the potentials V, V_1, and V_2, respectively; $d=1$ if V is a generic potential and $d=0$ if V is exceptional; in a similar manner, d_1 and d_2 take values 1 or 0 depending on whether V_1 and V_2 are generic or exceptional. Using (3.9) in (3.8) we obtain

$$N=N_1+N_2+\frac{1}{2}\,[d-d_1-d_2]-\frac{1}{\pi}\,\omega(0+). \tag{3.10}$$

Now let us analyze ω further. Note that R_1 and L_2 are continuous and nonzero and strictly less than one in absolute value for $k\in\mathbf{R}^+$ and that, as $k\to+\infty$, both R_1 and L_2 vanish.

In the following we need to distinguish between the generic case and the exceptional case. When V_1 and V_2 are both generic we have

$$R_1(k)=-1-2ika_{r;1}+\mathrm{o}(k),\quad L_2(k)=-1-2ika_{l;2}+\mathrm{o}(k),\quad k\to 0, \tag{3.11}$$

where

$$a_{r;1}=\left(1-\int_{-\infty}^{x_1}dx\;xV_1(x)f_{r;1}(0,x)\right)\Big/\int_{-\infty}^{x_1}dx\;V_1(x)f_{r;1}(0,x), \tag{3.12}$$

$$a_{l;2}=\left(1+\int_{x_1}^{\infty}dx\;xV_2(x)f_{l;2}(0,x)\right)\Big/\int_{x_1}^{\infty}dx\;V_2(x)f_{l;2}(0,x). \tag{3.13}$$

In the exceptional case we define

$$\gamma_1=\frac{f_{l;1}(0,x)}{f_{r;1}(0,x)}=\frac{1}{f_{r;1}(0,x_1)},\quad \gamma_2=\frac{f_{l;2}(0,x)}{f_{r;2}(0,x)}=f_{l;2}(0,x_1), \tag{3.14}$$

and note that, if V_1, resp. V_2, is exceptional, then

$$R_1(k)=-b_1+\mathrm{o}(1),\quad \text{resp.}\quad L_2(k)=b_2+\mathrm{o}(1),\quad k\to 0, \tag{3.15}$$

where

$$b_j=\frac{\gamma_j^2-1}{\gamma_j^2+1},\quad j=1,2. \tag{3.16}$$

The relations (3.11)–(3.13) follow from p. 146 of Ref. 6; (3.15) was proved in Ref. 12. We remark that the validity of (3.11) depends on the property that V_1 and V_2 are each supported on a semi-infinite interval; this guarantees the convergence of the integrals in the numerators in (3.12) and (3.13). In general, for potentials in $L_1^1(\mathbf{R})$ one can only conclude[12] that the reflection coefficients behave like $-1+\mathrm{o}(1)$ as $k\to 0$ in the generic case.

When both V_1 and V_2 are generic we have

$$1-R_1(k)L_2(k)=-2ik[a_{r;1}+a_{l;2}]+\mathrm{o}(k),\quad k\to 0. \tag{3.17}$$

When both V_1 and V_2 are exceptional we get

$$1-R_1(k)L_2(k)=1+b_1b_2+\mathrm{o}(1),\quad k\to 0. \tag{3.18}$$

When V_1 is generic and V_2 is exceptional we have

$$1-R_1(k)L_2(k)=1+b_2+\mathrm{o}(1),\quad k\to 0, \tag{3.19}$$

and finally, when V_1 is exceptional and V_2 is generic, we have

$$1-R_1(k)L_2(k)=1-b_1+\mathrm{o}(1),\quad k\to 0. \tag{3.20}$$

From (3.15) and (3.18)–(3.20) we see that if at least one of V_1 and V_2 is exceptional, then the quantity $[1-R_1(0)L_2(0)]$ is strictly positive, and hence $\omega(0+)=0$.

If both V_1 and V_2 are generic, the analysis is slightly more complicated: If $a_{r;1}<-a_{l;2}$, then $\omega(0+)=\pi/2$; if $a_{r;1}>-a_{l;2}$, then $\omega(0+)=-\pi/2$. If $a_{r;1}=-a_{l;2}$, then, as $k\to 0$, we get $1-R_1(k)L_2(k)=o(k)$, where we have used (3.17). As a result, (3.4) implies that $k/T(k)=o(1)$ as $k\to 0$, and this, in turn, implies that V is exceptional. Therefore the left-hand side of (3.4) has a limit as $k\to 0$, which means that in fact $1-R_1(k)L_2(k)=O(k^2)$, from which we obtain $\omega(0+)=0$.

It is known[13] that when V_1 and V_2 are both exceptional, then V is exceptional. If exactly one of V_1 and V_2 is exceptional, then V is generic. If both V_1 and V_2 are generic, then V can be exceptional or generic. By using these facts along with the value of $\omega(0+)$ and (3.10), we arrive at the following conclusions:

(i) If both V_1 and V_2 are exceptional, then $N=N_1+N_2$.
(ii) If exactly one of V_1 and V_2 is exceptional and the other is generic, then $N=N_1+N_2$.
(iii) If both V_1 and V_2 are generic and V is also generic, then $\omega(0+)=\pm\pi/2$. In this case, we have $N=N_1+N_2-1$ if $\omega(0+)=\pi/2$, and this happens if $a_{r;1}<-a_{l;2}$ in (3.17); or we have $N=N_1+N_2$ if $\omega(0+)=-\pi/2$, and this happens if $a_{r;1}>-a_{l;2}$.
(iv) If both V_1 and V_2 are generic and V is exceptional, then we must have $\omega(0+)=0$ and $N=N_1+N_2-1$. This happens if $a_{r;1}=-a_{l;2}$ in (3.17).

Summarizing, if $a_{r;1}\leqslant -a_{l;2}$ in (3.17) and both $a_{r;1}$ and $a_{l;2}$ are finite, then we have $N=N_1+N_2-1$; if at least one of $a_{r;1}$ and $a_{l;2}$ is infinite or if $a_{r;1}>-a_{l;2}$, then we have $N=N_1+N_2$.

There is a direct connection between cases (i)–(iv) above and cases (a) and (b) of Theorem 2.2 because the coefficients $a_{r;1}$ and $a_{l;2}$ are related to the quantity $Z(x_1)$ defined in (2.6). To see this recall that $f_{r;1}(0,x)$ and $f_{l;2}(0,x)$ obey the integral equations

$$f_{r;1}(0,x)=1+\int_{-\infty}^{x}dy(x-y)V_1(y)f_{r;1}(0,y), \tag{3.21}$$

$$f_{l;2}(0,x)=1+\int_{x}^{\infty}dy(y-x)V_2(y)f_{l;2}(0,y). \tag{3.22}$$

Hence from (3.21) and (3.22) we obtain

$$f_{r;1}(0,x)=c_{r;1}x+d_{r;1},\quad x>x_1,$$

$$f_{l;2}(0,x)=-c_{l;2}x+d_{l;2},\quad x<x_1,$$

with

$$c_{r;1}=\int_{-\infty}^{x_1}dy\ V_1(y)f_{r;1}(0,y),\quad d_{r;1}=1-\int_{-\infty}^{x_1}dy\ yV_1(y)f_{r;1}(0,y), \tag{3.23}$$

$$c_{l;2}=\int_{x_1}^{\infty}dy\ V_2(y)f_{l;2}(0,y),\quad d_{l;2}=1+\int_{x_1}^{\infty}dy\ yV_2(y)f_{l;2}(0,y). \tag{3.24}$$

Thus from (3.12), (3.13), (3.23), and (3.24) we conclude that

$$a_{r;1}=\frac{d_{r;1}}{c_{r;1}},\quad a_{l;2}=\frac{d_{l;2}}{c_{l;2}}. \tag{3.25}$$

Moreover,

$$\frac{f_{r;1}(0,x_1)}{f'_{r;1}(0,x_1)}=x_1+a_{r;1},\quad \frac{f_{l;2}(0,x_1)}{f'_{l;2}(0,x_1)}=x_1-a_{l;2},$$

and hence

$$Z(x_1) = -a_{r,1} - a_{l;2}.$$

Thus (i) and (ii) above correspond to the possibilities of Theorem 2.2 (b); (i) is the case when $f'_{r;1}(0,x_1) = f'_{l;2}(0,x_1) = 0$, and (ii) is the case when exactly one of $f'_{r;1}(0,x_1)$ and $f'_{l;2}(0,x_1)$ is zero. Case (iii) corresponds to (a) of Theorem 2.2 with $Z(x_1) \neq 0$ and case (iv) corresponds to (a) with $Z(x_1) = 0$.

We conclude this section with a brief look at families of potentials of the form

$$V_\xi(x) = V_1(x) + V_2(x-\xi), \tag{3.26}$$

where ξ is a non-negative parameter and V_1 and V_2 are the two fragments of V. In other words, the parameter ξ controls the separation distance between the two fragments. The next result shows that the number of bound states can only increase if ξ is increased. By virtue of (2.5) it can only increase by one. Since the proof is short we present two versions, one using the method of Sec. II and the other using the method of this section. In the case of compactly supported fragments the result is already known from Refs. 1 and 3.

Theorem 3.1: Let N_ξ denote the number of bound states of V_ξ given in (3.26). Then either $N_\xi = N_1 + N_2$ for all $\xi \geqslant 0$ or there is a unique $\xi_0 \geqslant 0$ such that $N_\xi = N_1 + N_2 - 1$ for $0 \leqslant \xi \leqslant \xi_0$ and $N_\xi = N_1 + N_2$ for $\xi > \xi_0$.

Proof: (a) First, if one of the fragments is exceptional, then we have $N_\xi = N_1 + N_2$ for all $\xi \geqslant 0$. If both fragments are generic, then we let $f_{l;2;\xi}(k,x)$ denote the Jost solution from the left for the potential $V_2(x-\xi)$. Then $f_{l;2;\xi}(0,x) = -c_{l;2}(x-\xi) + d_{l;2}$ for $x < x_1 + \xi$, and thus, by using (2.6) and (3.26), we obtain $Z_\xi(x_1) = -\xi - a_{l;2} - a_{r;1}$. Thus if $Z_0(x_1) < 0$, then, for all $\xi \geqslant 0$, $Z_\xi(x_1) < 0$ and hence $N_\xi = N_1 + N_2$. If $Z_0(x_1) \geqslant 0$, then $Z_{\xi_0}(x_1) = 0$ when $\xi_0 = Z_0(x_1) = -a_{l;2} - a_{r;1}$ and the assertion follows.

(b) Replacing $L_2(k)$ by $e^{2ik\xi}L_2(k)$ in (3.18) we obtain

$$1 - R_1(k)L_{2;\xi}(k) = -2ik[a_{r;1} + a_{l;2} + \xi] + o(k), \quad k \to 0.$$

Now the conclusion follows using (iii) and (iv) above. ■

IV. AN EXAMPLE

The following example illustrates Theorems 2.1 and 3.1. Let

$$V(x) = \begin{cases} A^2, & x \in (0,1) \\ -B^2, & x \in (1,2) \\ 0, & \text{elsewhere,} \end{cases} \tag{4.1}$$

where A and B are some positive constants. We can fragment V as $V = V_1 + V_2$, where V_1 is a square potential barrier of height A^2 with support (0,1) and V_2 is a square well of depth $-B^2$ with support (1,2). Then, a straightforward computation using (3.23)–(3.25) yields $c_{r;1} = A \sinh A$, $d_{r;1} = \cosh A - A \sinh A$, $c_{l;2} = -B \sin B$, $d_{l;2} = \cos B - B \sin B$, and thus

$$a_{r;1} = \frac{1}{A}\coth A - 1, \quad a_{l;2} = -\frac{1}{B}\cot B + 1.$$

Let us demonstrate that by choosing A and B suitably, we can have $N_1 = 0$, $N_2 = 1$, $N = 0$. In other words, the positive fragment V_1 may cancel the bound state caused by the negative fragment V_2, resulting in no bound states for V. Unless B is a multiple of π, both V_1 and V_2 are generic. If we let, for example, $B = \pi/4$, then from (2.6) we get $Z(x_1) = -a_{r;1} - a_{l;2} \geqslant 0$ whenever $A \geqslant A_0$, where A_0 satisfies $A_0 \tanh A_0 = \pi/4$, i.e., $A_0 = 1.0201\ldots$. For $A = A_0$ the potential V is exceptional with no bound state and for $A > A_0$ it is generic with no bound state. Now let us consider the family V_ξ defined in (3.26). If $A < A_0$ and $B = \pi/4$, then $Z_0(x_1) < 0$ and we have $N_\xi = N_1 + N_2 = 1$ for all ξ. If $A = A_0$, then we have $N_0 = 0$ but $N_\xi = 1$ for $\xi > 0$, i.e., $\xi_0 = 0$. If $A > A_0$, then ξ_0 is given by

$$\xi_0 = \frac{1}{B \tan B} - \frac{1}{A \tanh A},$$

and we have $N_\xi = 0$ for $\xi \leqslant \xi_0$ and $N_\xi = 1$ for $\xi > \xi_0$.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DMS-9501053 and by CNR, MURST, and a University of Cagliari Coordinated Research Grant.

[1] M. Klaus, ''Some remarks on double-wells in one and three dimensions,'' Ann. Inst. Henri Poincaré, Sect. A **34**, 405–417 (1981).
[2] M. Sassoli de Bianchi, ''Levinson's theorem, zero-energy resonances, and time delay in one-dimensional scattering systems,'' J. Math. Phys. **35**, 2719–2733 (1994).
[3] M. Sassoli de Bianchi and M. Di Ventra, ''On the number of bound states by one-dimensional finite periodic potentials,'' J. Math. Phys. **36**, 1753–1764 (1995).
[4] T. Aktosun, ''A factorization of the scattering matrix for the Schrödinger equation and for the wave equation in one dimension,'' J. Math. Phys. **33**, 3865–3869 (1992).
[5] L. D. Faddeev, ''Properties of the *S*-matrix of the one-dimensional Schrödinger equation,'' Am. Math. Soc. Trans. **2**, 139–166 (1964) [Tr. Mat. Inst. Steklova **73**, 314–336 (1964) (Russian)].
[6] P. Deift and E. Trubowitz, ''Inverse scattering on the line,'' Commun. Pure Appl. Math. **32**, 121–251 (1979).
[7] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. (Springer, New York, 1989).
[8] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, IV: Analysis of Operators* (Academic, New York, 1978).
[9] T. Aktosun, M. Klaus, and C. van der Mee, ''Wave scattering in one dimension with absorption,'' J. Math. Phys. **39**, 1957–1992 (1998).
[10] T. Aktosun, M. Klaus, and C. van der Mee, ''On the Riemann-Hilbert problem for the one-dimensional Schrödinger equation,'' J. Math. Phys. **34**, 2651–2690 (1993).
[11] R. G. Newton, ''Inverse scattering. I. One dimension,'' J. Math. Phys. **21**, 493–505 (1980).
[12] M. Klaus, ''Low-energy behaviour of the scattering matrix for the Schrödinger equation on the line,'' Inverse Probl. **4**, 505–512 (1988).
[13] T. Aktosun, M. Klaus, and C. van der Mee, ''Factorization of scattering matrices due to partitioning of potentials in one-dimensional Schrödinger-type equations,'' J. Math. Phys. **37**, 5897–5915 (1996).

Symplectic structure for Gaussian diffusions

Augusto Brandão[a)]
Av Prof. Gama Pinto 2, 1669 Lisboa Cedex, Portugal

(Received 24 April 1997; accepted for publication 13 January 1998)

Using information exclusively about classical Hamiltonian boundary value problems (with quadratic time-dependent potentials) we construct probability measures providing a symplectic structure associated with some inhomogeneous Gaussian diffusion processes introduced earlier [T. Kolsrod and J. C. Zambrini, J. Math. Phys. **33**, 1301 (1992)]. © *1998 American Institute of Physics.*
[S0022-2488(98)00309-0]

I. INTRODUCTION

At the end of the 1940s, Feynman introduced a new interpretation of nonrelativistic quantum mechanics.[1] Let us briefly discuss it, in order to introduce some notations, by considering a one-particle system in $\mathbb{R}^3$ (of unit mass) described by the quantum Hamiltonian

$$H=-\frac{\hbar^2}{2}\Delta+V, \tag{1}$$

where $\hbar$ is the Planck constant and V is a regular, bounded below, scalar potential. Quantum dynamics is given (in any time interval, e.g., $[0, T]$) by the solution of the Cauchy problem in $L^2(\mathbb{R}^3)$, for χ in the domain of H, by

$$i\hbar\,\partial_t\psi=H\psi, \quad \psi(0,x)=\chi(x). \tag{2}$$

According to Born's interpretation of $\psi(t,x)$, the probability of finding the particle on a certain region B (B being a Borelian in $\mathbb{R}^3$) at time t is given by

$$\int_B \psi(x,t)\,\overline{\psi}(x,t)\;dx,$$

where $\overline{\psi}$ is the complex conjugate of ψ. Introducing the integral kernel or ''propagator,'' written symbolically as

$$K(x_1,t_1,x_2,t_2)=\int_{\Omega_{x_1}^{x_2}} e^{(i/\hbar)S(\omega)}\mathscr{D}\omega, \quad t_1<t_2,$$

where S is the action of the underlying classical system, i.e., here,

$$S(\omega)=\int_{x_1}^{x_2}\left(\frac{1}{2}\,\dot{\omega}(s)^2-V(\omega(s),s)\right)\,ds,$$

$\omega\in\Omega_{x_1}^{x_2}$ is a continuous path joining x_1 and x_2, and $D\omega$ represents a formal ''measure'' $D\omega=\Pi_{t_1<s<t_2}\,d\omega(s)$ on $\Omega_{x_1}^{x_2}$, the solution to the Schrödinger initial value problem (2) can be represented by

$$\psi(t,y)=\int_{\mathbb{R}^3} K(0,x,t,y)\chi(x)\;dx.$$

a)Electronic mail: abrandao@cc.fc.ul.pt

 © 1998 American Institute of Physics

So the probability that at time t the particle is in B should be

$$\begin{aligned} P(t) &= \int_B \psi(t,y)\overline{\psi}(t,y)\ dy \\ &= \int_B \int_{\mathbb{R}^3} K(0,x,t,y)\chi(x)\overline{\psi}(t,y)\ dx\ dy \\ &= \int_B \int_{\mathbb{R}^3} \chi(x)\overline{\chi}(x)K(0,x,t,y)\ \frac{\overline{\psi}(t,y)}{\overline{\chi}(x)}\ dx\ dy \\ &= \int_B \int_{\mathbb{R}^3} \chi(x)\overline{\chi}(x)\mathbf{K}(0,x,t,y)\ dx\ dy, \end{aligned} \tag{3}$$

where if $\overline{\chi}(x_1)\neq 0$,

$$\mathbf{K}(t_1,x_1,t_2,x_2) = K(t_1,x_1,t_2,x_2)\ \frac{\overline{\psi}(t_2,x_2)}{\overline{\chi}(x_1)}.$$

This suggests that $\mathbf{K}$ is formally the density of a transition probability wrt dx_2, in the sense of the classical theory of Markov processes.[2,3] Also observe that Feynman's path integral formula provides an informal way to describe quantum mechanics exclusively in terms of classical data, the action functional of the system.

Unfortunately, as it is well known today,

(i) $\mathscr{D}_\omega$ does not exist as a measure on $\Omega_{x_1}^{x_2}$.

(ii) $\mathbf{K}$ is complex valued, so it cannot be the transition probability of a real-valued stochastic process. There is no probability space where one can find such a stochastic process compatible with the Born interpretation and the predictions of regular quantum mechanics.

On the other hand, Schrödinger[4] has considered the following associated problem of classical statistical mechanics (Refs. 4–8), when the scalar potential V is zero:

Suppose that the position of a classical Brownian particle in a domain $X\in\mathbb{R}^3$ is observed at times t_1 and t_2 $(t_1<t_2)$ and found to have strictly positive probability densities with respect to Lebesgue measure:

$$\rho_{t_1}(x)\ dx, \quad \rho_{t_2}(z)\ dz.$$

What then, is the most probable evolution of a probability density compatible with those two arbitrary measurements?

Schrödinger shows that the probability density should be of the form

$$\rho(y,t) = \eta(t,y)\,\hat{\eta}(t,y), \quad t_1\leqslant t\leqslant t_2,$$

where the pair $\{\eta,\hat{\eta}\}$ are regular positive solutions of the two adjoint heat equations:

$$\partial_t \eta(t,y) = \tfrac{1}{2}\Delta\,\eta(t,y), \quad t\geqslant t_1, \quad y\in B, \tag{4}$$

$$-\partial_t \hat{\eta}(t,y) = \tfrac{1}{2}\Delta\,\hat{\eta}(t,y), \quad t\leqslant t_2, \quad y\in B, \tag{5}$$

So the problem amounts to proving the existence and uniqueness of positive solutions of the following nonlinear system of equations for $\eta(x,t_1)=\chi(x)$ and $\hat{\eta}(z,t_2)=\hat{\chi}(z)$,

$$\eta_{t_1}(x)\int_{\mathbb{R}^3} h(x,t_1,z,t_2)\,\hat{\eta}_{t_2}(z)\ dz = \rho_{t_1}(x), \tag{6}$$

$$\hat{\eta}_{t_2}(z)\int_{\mathbb{R}^3} \eta_{t_1}(x)h(x,t_1,z,t_2)\ dx = \rho_{t_2}(z), \tag{7}$$

where h denotes the (positive) integral kernel associated with the heat equations (4) and (5).

This problem was not solved by Schrödinger but by Beurling, almost 30 years later, in a much more general setting.[5,9]

In particular, Beurling's result provides us with the existence and uniqueness of the pair $\{\eta, \hat{\eta}\}$ of solutions of the adjoint heat equations for any potential V in the Kato class.[7] Using this, Zambrini[6] established the existence of a class of $\mathbb{R}^3$ Markovian (''Bernstein'') diffusion processes Z_t, for any V in the Kato class, such that

$$P(Z_t \in B) = \int_B \eta(t,y)\, \hat{\eta}(t,y)\, dy, \tag{8}$$

$$\frac{d^2 E(Z_t)}{dt^2} = E[\nabla V(Z_t, t)], \tag{9}$$

where E denotes the (absolute) expectation of the process. Notice the Euclidean change of sign of (9). For such processes Z_t we can define two transition probabilities:

$$P_f(s,x,t,B) = P(Z_t \in B \,|\, Z_s = x) \to Z_{s,x},$$

$$P_b(t,B,u,z) = P(Z_t \in B \,|\, Z_u = z) \to Z^{u,z},$$

where $t \in I = [s, u]$, a fixed time interval, and $Z_{s,x}$, $Z^{u,z}$ denote the associated conditioned processes. The Euclidean analog of Feynman's formula ''transition amplitude''[1,10] can be expressed as

$$P(Z_t \in B) = \int_B \int_{\mathbb{R}^3} \eta(s,x)\, \hat{\eta}(s,x) \rho_{s,x}(t,y)\, dx\, dy, \tag{10}$$

but also as

$$P(Z_t \in B) = \int_B \int_{\mathbb{R}^3} \eta(u,z)\, \hat{\eta}(u,z) \rho^{u,z}(t,y)\, dz\, dy, \tag{11}$$

where $\rho_{s,x}$, $\rho^{u,z}$ are respectively the probability densities of the conditioned processes $Z_{s,x}$, $Z^{u,z}$.

The essential element of Feynman's path integral representation is its relation with classical Lagrangian mechanics, and, in this respect, the quadratic Lagrangians play a very special role. In this case, Feynman shows that the sum over the paths of his formal stochastic process is entirely characterized by classical mechanics. This aspect has also been essential in his use of path integrals in quantum electrodynamics. The main objective of the present paper is to prove an analogous result in the Euclidean counterpart suggested by Schrödinger, i.e., given two Gaussian densities at time s and u, $\rho_s\, dx$, $\rho_u\, dx$, and a time-dependent quadratic potential V, construct (using only classical information) a diffusion process Z_t compatible with the above data and verify

$$\frac{d^2}{dt^2} E(Z_t) = \nabla V(E(Z_t t)).$$

In other words, we want to construct the relevant Gaussian probability measures from purely classical information.

Although we only treat quadratic potentials in this paper, recent developments show that the same treatment holds for some nonlinear equation of motion.[11]

In Sec. II we will describe the underlying classical model. Here the (real time) quantum Hamiltonian given by (1) will be replaced by the Euclidean one:

$$H = \frac{p^2}{2} - V(q,t).$$

The boundary conditions we will impose on the system will take the form of a pair of Lagrangian manifolds at the end points of the time interval. Then, we will show that the resulting classical dynamics can be completely described in terms of the flow of the pair of Lagrangian manifolds (in both senses of time), in particular the solution of the resulting boundary value problem is obtained by the intersection at each time of a pair of Lagrangian manifolds which are the time evolution of the original ones. In Sec. III, using the above-mentioned natural symplectic structure of phase space, we construct the probability measures which are completely characterized at each time by the pair of Lagrangian manifolds at the same instant. The resulting Gaussian density on the time interval will satisfy by construction the required boundary conditions of the quantum problem.

The general framework underlying our study is the one of ''Euclidean Quantum Mechanics.''[10,12] Its relations with regular quantum theory are deeper than the ones of the traditional Euclidian counterpart inspired by the Feynman–Kak formula.[13] For example, it has been proved recently that this framework provides better information about the quantum system than the familiar theorems of quantum mechanics.[14]

II. THE CLASSICAL MODEL

In this paper, our basic dynamical equation will be the classical n-dimensional (Euclidean) Newtonian equation,

$$\ddot{q}(t)=K(t)q(t), \tag{12}$$

on a time interval $I\equiv[s,u]$, where $K(t)$ is a symmetric $n\times n$ matrix on I with a continuous dependence in t. Moreover, we will assume that (12) is disconjugate[15,16] on the time interval I, i.e., Eq. (12) has no trivial solution $q(t)$ such that $q(t_1)=q(t_2)=0$, where t_1 and t_2 are distinct points of I. Equation (12) can be regarded as the classical Jacobi equation for the Euler–Lagrange equations with (Euclidean) Lagrangian of the form

$$L(q,\dot{q},t)=\tfrac{1}{2}|\dot{q}|^2+V(q,t), \tag{13}$$

where $|.|$ denotes the Euclidean norm. Then

$$K(t)=\left(\left.\frac{\partial^2 V}{\partial_{q_i}\,\partial_{q_j}}\right|_{\overline{q}(t)}\right), \quad i,j=1,2,3,...,n,$$

for $\overline{q}(t)$ an extremal of the action functional with Lagrangian L. This is the point of view of the semiclassical [or Wentzel–Kramers–Brillouin (WKB)] approximation in Feynman's approach to quantum mechanics.[1] Equation (12) can be transformed into an Hamiltonian form on $\mathbb{R}^{2n}$ for the Hamiltonian function given by

$$H(q,p,t)=\tfrac{1}{2}|p|^2-\tfrac{1}{2}\langle K(t)q,q\rangle. \tag{14}$$

The correspondent Hamilton's equation of motion will be given by the system:

$$\dot{q}_i=p_i, \quad \dot{p}_i=\sum_{1\leqslant j\leqslant n} K_{ij}(t)q_j, \quad i=1,...,n. \tag{15}$$

We also consider the matrix form of (15), i.e.,

$$\dot{C}_{ij}=D_{ij}, \quad \dot{D}_{ij}=\sum_{1\leqslant l\leqslant n} K_{il}(t)C_{lj}, \tag{16}$$

with the correspondent matrix second-order differential system:

$$\ddot{C}_{ij}(t)=\sum_{1\leqslant l\leqslant n} K_{il}(t)C_{lj}(t). \tag{17}$$

We will denote by $\mathbf{C}_t$ the matrix solution $(C_t,D_t)=(C_t,\dot{C}_t)$ of (16). The disconjugacy assumption for (12) on the time interval I implies that for any solution (C_t,D_t) of (16), C_t can only be noninvertible at one point of I. In particular, for $t_0\in I$, if one defines $E_{t_0}(t)$ as the solution on I of (17) that satisfies the Cauchy data,

$$E_{t_0}(t_0)=0, \quad \dot{E}_{t_0}(t_0)=I_{n\times n}, \tag{18}$$

then $E_{t_0}(t)$ is a nonsingular matrix for $t\neq t_0$.

The linearity of the Hamiltonian system ensures that the phase space has a natural vector space structure that we can identify with $\mathbb{R}^{2n}$. Hence the canonical symplectic structure provided by the two-form[17]

$$\omega=dq\wedge dp=\sum_{i=1}^{n} dq_i\wedge dp_i$$

can be used to define an antisymmetric inner product between elements of the phase space. Given two phase space vectors $X_1\equiv(q_1^1,\ldots,q_n^1,p_1^1,\ldots p_n^1)$ and $X_2\equiv(q_1^2,\ldots,q_n^2,p_1^2,\ldots p_n^2)$, their *symplectic inner product* is defined as

$$\omega(X_1,X_2)=\sum_{i=1}^{2n} q_i^1p_i^2-q_i^2p_i^1. \tag{19}$$

It is easy to prove that if $(X_1(t),X_2(t))$ are solutions of (15) on a time interval I, then $\omega(X_1(t),X_2(t))$ is constant.

We will need a generalization of the symplectic inner product to matrix functions. In order to simplify the notation if $A=(A_{\alpha\beta})$ and $B=(B_{\alpha\beta})$ are $n\times r$ matrices, then the $2n\times r$ matrix $C=(C_{\alpha,\beta})(\alpha=1,2,...,n;\beta=1,2,...,r)$ with $C_{\alpha\beta}=A_{\alpha\beta}$, $C_{n+\alpha,\beta}=B_{\alpha\beta}$ will be denoted as $(A;B)$. Hence if $C=(C_1;C_2)$ and $D=(D_1;D_2)$ are $2n\times n$ matrices and ^{T} denotes the transpose, we will call the matrix function

$$\omega(C,D)=C_1^TD_2-C_2^TD_1 \tag{20}$$

the symplectic inner product of C and D. If $X\in\mathbb{R}^{2n}$ and $C=(C_1;C_2)$, then $\omega(X,C)$ or $\omega(C,X)$ can be defined as $\omega(I_X,C)$ or $\omega(C,I_X)$ where I_X is the diagonal matrix with $(I_X)_{ii}=X_i$ for $i=1,...,2n$.

Although it is true that if $\mathbf{C}_1$ and $\mathbf{C}_2$ are two solutions of (16) on a time interval I, then $\omega(\mathbf{C}_t^1,\mathbf{C}_t^2)$ is a constant matrix on I, but, in contrast with (20), ω is not antisymmetric, hence in general we do not have $\omega(\mathbf{C},\mathbf{C})=0$. The solutions $\mathbf{C}_t$ of (16) [or the solutions C_t of (17)] such that $\omega(\mathbf{C},\mathbf{C})=0$ are sometime called *isotropic* solutions of (17) on I.

Definition 1: A Lagrangian manifold M is defined as an n manifold in a $2n$-dimensional phase space such that for all points (x,p) on M and for all vectors X,Y tangent to M at (x,p) we have

$$\omega(X,Y)=0. \tag{21}$$

Some obvious consequences of this definition are the following:

(1) In $\mathbb{R}^2$ it is clear that a Lagrangian manifold must be a one-dimensional surface, i.e., a curve.

(2) Any manifold of the form $q=\text{const}$ or $p=\text{const}$ on a phase space is a Lagrangian manifold. We will call a *singular Lagrangian manifold associated with a configuration* $x\in\mathbb{R}^n$ a Lagrangian manifold of the form

$$M^x=\{(q,p)|q=x\}. \tag{22}$$

(3) Any graph of a curl-free momentum field $p=p(q)$ is a Lagrangian manifold. If the Lagrangian manifold M is of the form

$$M=\left\{(q,p)\in\mathbb{R}^{2n}\middle| p=\frac{\partial S(q)}{\partial q}\right\} \tag{23}$$

for some regular function $S(q)$, this function is called the *generating function* of the Lagrangian manifold. Given an arbitrary Lagrangian manifold, its generating function may not be defined. A typical example of this is the singular Lagrangian manifold associated with a configuration $x \in \mathbb{R}^n$.

In this paper we only consider linear Lagrangian manifolds, i.e., n-planes on $\mathbb{R}^{2n}$.

A. Evolution of Lagrangian manifolds

Consider the Lagrangian manifolds M_s and M_s^x defined by

$$M_s=\{(q,p)\in\mathbb{R}^{2n}|p=\alpha q-\beta\}, \tag{24}$$

$$M_s^x=\{(q,p)\in\mathbb{R}^{2n}|q=x\}, \tag{25}$$

where α is a non-negative definite invertible $n\times n$ matrix and $\beta,x\in\mathbb{R}^n$.

Our first aim is to explicitly compute the images under the Hamiltonian flow associated with (14) of these two Lagrangian manifolds in a time interval $I=[s,\ u]$.

Definition 2: Given an initial Lagrangian manifold, we define its associated (matrix) function F_t on an interval $I=[s,\ u]$ as an isotropic matrix solution of (17) such that the following hold.

(1) If the Lagrangian manifold is defined by (24), F_t is a nonsingular matrix for $t\in I$ such that

$$\dot{F}_sF_s^{-1}=\alpha. \tag{26}$$

(2) If the Lagrangian manifold is singular [i.e., defined by (25)], then F_t is a nonsingular matrix for $t\in(s,u]$ and satisfies $F(s)=0$. An example of such a function is $E_s(t)$, the solution of (17) on I that satisfies the Cauchy data $E_s(s)=0$, $\dot{E}_s(s)=I_{n\times n}$.

Lemma 1: If F_t and H_t are two associated functions to a Lagrangian manifold M_s or M_s^x on the time interval I, then $\omega(\mathbf{F}_t,\mathbf{H}_t)=0$.

Proof: If the Lagrangian manifold is defined by (24), then both F_t and H_t satisfy (26). Hence,

$$0=\dot{F}_sF_s^{-1}-\dot{H}_sH_s^{-1}=F_s^{-T}(\dot{F}_s^TH_s-F_s^T\dot{H}_s)H_s^{-1}=-F_s^{-T}\omega(\mathbf{F}_t,\mathbf{H}_t)H_s^{-1},$$

which implies that

$$\omega(\mathbf{F}_t,\mathbf{H}_t)=0. \tag{27}$$

If the Lagrangian manifold is defined by (25), then F_t and H_t satisfy $F(s)=H(s)=0$. Since both matrix functions are solutions of (16), their symplectic product is time independent which implies that

$$\omega(\mathbf{F}_t,\mathbf{H}_t)=F_s^T\dot{H}_s-\dot{F}_sH_s=0.$$

□

Proposition 1: The image of the Lagrangian manifolds defined by (24) or (25) under the Hamiltonian flow defined by (14), for $t\in I$, is given by

$$M_t=\{(q,p)|\dot{F}^T(t)q-F^Tp=\delta\}, \tag{28}$$

where $F(t)$ is the associated function to M_s (resp. M_s^x) and $\delta=F_s^T\beta$ (resp. $\delta=x$). Moreover, the expression (28) is independent of the choice of the associated function to M_s, i.e., if H_t is another associated function to M_s on I, then

$$\{(q,p)|\dot{F}^T(t)q-F^Tp=\delta\}=\{(q,p)|\dot{H}^T(t)q-H^Tp=\delta^H\},$$

where $\delta^H=H_s^T\beta$.

For the proof we will need the following Lemma:

Lemma 2: On the time interval $I=[s,u]$ *and for any two configurations* $x,y\in\mathbb{R}^n$, *consider Hamilton's principal function* $R(s,x,t,y)$ *defined*[17] *as the line integral*

$$R(s,x,t,y)=\int_\gamma L\ dt, \tag{29}$$

where L *denotes the Lagrangian defined by (13) along the solution of (12),* $\gamma_\tau=(q(\tau),\dot{q}(\tau))$ *connecting the points (s,x) and (t,y). We have*

$$R(s,x,t,y)=\tfrac{1}{2}\langle\dot{E}_sE_s^{-1}(t)y,y\rangle-\tfrac{1}{2}\langle\dot{E}_t(s)E_t^{-1}(s)x,x\rangle-\langle E_s^{-1}(t)y,x\rangle, \tag{30}$$

where $E_t(\tau)$ *and* $E_s(\tau)$ *are solutions of (16) on the interval I satisfying the Cauchy data*

$$E_t(t)=0,\quad \dot{E}_t(t)=I_{n\times n}; \tag{31}$$

$$E_s(s)=0,\quad \dot{E}_s(s)=I_{n\times n}. \tag{32}$$

Proof: By (29),

$$R(s,x,t,y)=\int_s^t\left[\frac{|\dot{q}_\lambda|^2}{2}+\frac{1}{2}\langle K(\lambda)q_\lambda,q_\lambda\rangle\right]d\lambda,$$

where q_λ is the solution of (12) satisfying $q(s)=x$ and $q(t)=y$ (we will assume for the moment that this solution is well defined on I and is unique). Hence

$$\begin{aligned}R(s,x,t,y)&=\int_s^t\left[\frac{|q_\lambda|^2}{2}+\frac{1}{2}\langle K(\lambda)q_\lambda,q_\lambda\rangle\right]d\lambda\\&=\int_s^t\frac{|\dot{q}_\lambda|^2}{2}\,d\lambda+\frac{1}{2}\int_s^t\langle\ddot{q}_\lambda,q_\lambda\rangle\,d\lambda=\frac{1}{2}\langle\dot{q}_t,y\rangle-\frac{1}{2}\langle\dot{q}_s,x\rangle.\end{aligned} \tag{33}$$

Using the ansatz

$$q_\tau=E_s(\tau)\lambda_1+E_t(\tau)\lambda_2,$$

where $\lambda_i\in\mathbb{R}^n$, $i=1,2$, and taking into account the boundary conditions $[q(s)=x,q(t)=y]$ we will obtain

$$q_\tau=E_s(\tau)E_s^{-1}(t)y+E_t(\tau)E_t^{-1}(s)x. \tag{34}$$

Substituting in (33),

$$R(s,x,t,y)=\tfrac{1}{2}\langle\dot{E}_sE_s^{-1}(t)y,y\rangle-\tfrac{1}{2}\langle\dot{E}_t(s)E_t^{-1}(s)x,x\rangle+\tfrac{1}{2}\langle E_s^{-1}(t)y,x\rangle-\tfrac{1}{2}\langle E_s^{-1}(t)x,y\rangle. \tag{35}$$

As $E_s(\tau)$ and $E_t(\tau)$ are both solutions of (16), the symplectic inner product $\omega(\mathbf{E}_s,\mathbf{E}_t)$ will be constant on the interval I. Hence, using (31) and (32),

$$\omega(\mathbf{E}_s(s),\mathbf{E}_t(s))=\omega(\mathbf{E}_s(t),\mathbf{E}_t(t)),$$

$$E_s^T(s)\dot{E}_t(s)-\dot{E}_s^T(s)E_t(s)=E_s^T(t)\dot{E}_t(t)-\dot{E}_s^T(t)E_t(t),$$

$$-E_t(s)=E_s(t).$$

Using this equality in (35), we obtain the desired result. Observe also that as we are assuming the disconjugacy of (15), $E_s(\tau)$ and $E_t(\tau)$ will be invertible respectively on the time interval $(s,u]$ and $[s,t)$; in particular $E_s(t)$ is invertible, which proves that (34) is the unique solution of (12) that satisfies the required boundary conditions. □

Proof of Proposition 1: Let us assume that the initial Lagrangian manifold is given by (24), i.e.,

$$M_s=\{(q,p)=\alpha q-\beta\},$$

and let us take the associated function to the Lagrangian manifold, F_t, which, by definition, satisfies the following relation at time s:

$$F_sF_s^{-1}=\alpha.$$

Then we can write the generating function for the initial Lagrangian manifold as

$$S(x)=\tfrac{1}{2}\langle F_sF_s^{-1}x,x\rangle-\beta y+\gamma, \tag{36}$$

where γ is an arbitrary real constant.

The idea of the proof is to explicitly compute $\nabla_y S(t,y)$, where $S(t,y)$ will be the generating function for the Lagrangian manifold M_t, since, then we will have[18]

$$M_t=\left\{(q,p)|p=\frac{\partial S}{\partial q}\,(t,q)\right\}, \tag{37}$$

for $S(t,y)$ the solution of the Hamilton–Jacobi equation on I,

$$\frac{\partial S}{\partial y}+H\left(t,y,\frac{\partial S}{\partial y}\right)=0, \tag{38}$$

subject to the Cauchy data

$$S(s.x)=S(x). \tag{39}$$

The solution of the initial value problem defined by (38) and (39) is given by (cf. Ref. 18)

$$S(t,y)=S_s(x)+R(s,x,t,y), \tag{40}$$

where $R(s,x,t,y)$ denotes Hamilton's principal function defined as the line integral of Poincaré–Cartan form $pdq-Hdt$ evaluated along the solution of Hamilton's equation that connects x at time s to y at time t. The starting configuration x may itself be regarded as a function of y and t, i.e., as the starting point of a trajectory that at time t is in the configuration y. A simple computation will give us (cf. Ref. 19) the following expression for x:

$$x=F^{-1}(t)y+F^{-1}(t)E_s(t)\beta, \tag{41}$$

where $F(t)$ is an associated function to M_s. Observing that (40) implies (cf. Ref. 18)

$$\nabla_y S(t,y)=\nabla_y R(s,x,t,y), \tag{42}$$

we obtain, using Lemma 2,

$$\nabla_y R=\tfrac{1}{2}(\dot{E}_sE_s^{-1}(t)+E_s^{-T}\dot{E}_s^T(t))y-[E_s^{-1}(t)]^Tx. \tag{43}$$

As $E_s(t)$ is an isotropic solution of (17), $\dot{E}_s(t)E_s^{-1}(t)$, in particular, is symmetric on the interval $(s,u]$. Hence,

$$\nabla_y R=\dot{E}_sE_s^{-1}(t)y-[E_s^{-1}(t)]^Tx.$$

Using (41) we obtain,

$$\nabla_y S(t,y)=f(t)y-\beta(t), \tag{44}$$

where $f(t)$ and $\beta(t)$ are defined by

$$f(t)=\dot{E}_s(t)E_s^{-1}(t)-[E_s^{-1}(t)]^T F(s)F^{-1}(t),$$

$$\beta(t)=[E_s^{-1}(t)]^T F^{-1}(t)E_s(t)\beta.$$

As $-F(s)=\omega(\mathbf{E}_s(\tau),\mathbf{F}(\tau))$, we can simplify the expression for $f(t)$:

$$\begin{aligned} f(t)&=\dot{E}_s(t)E_s^{-1}(t)-[E_s^{-1}(t)]^T F(s)F^{-1}(t)=[E_s^{-1}(t)]^T[\dot{E}_s^T(t)-F(s)F^{-1}(t)]\\ &=[E_s^{-1}(t)]^T[\dot{E}_s^T(t)F(t)-F(s)]F^{-1}(t)=[E_s^{-1}(t)]^T E_s^T(t)\dot{F}(t)F^{-1}(t)=\dot{F}F^{-1}(t). \end{aligned}$$

As F_t is the associated function for M_s on I, F_t is an isotropic solution of (17) which implies that $\dot{F}F^{-1}(t)$ is a symmetric matrix in I. Using this property we can simplify as well the expression for $\beta(t)$:

$$\begin{aligned} \beta(t)&=[E_s^{-1}(t)]^T F(s)F^{-1}(t)E_s(t)\beta=[\dot{E}_s(t)-f(t)E_s(t)]\beta=[\dot{E}_s(t)-\dot{F}(t)F^{-1}(t)E_s(t)]\beta\\ &=[\dot{E}_s(t)-[F^{-1}(t)]^T\dot{F}^T(t)E_s(t)]\beta=[F^{-1}(t)]^T[F^T\dot{E}_s(t)-\dot{F}^T(t)E_s(t)]\beta\\ &=[F^{-1}(t)]^T\omega(F,E_s)\beta=[F^{-1}(t)]^T F^T(s)\beta=[F^{-1}(t)]^T\delta. \end{aligned}$$

Hence, using (37) and (44) we obtain

$$M_t=\{(q,p)|p=\dot{F}_t F_t^{-1}(t)q-[F^{-1}(t)]^T\delta\}. \tag{45}$$

Using again the fact that F_t is an associated function for M_s on I, we will obtain for $(q,p)\in M_t$

$$F_t^T p-\dot{F}_t^T q=-\delta. \tag{46}$$

If the initial manifold is a singular one associated with a configuration $x\in\mathbb{R}^n$, i.e.,

$$M_s=\{(q,p)|q=x\},$$

then we do not have a generating function for this manifold. However, for $t>s$ the corresponding momentum field is given by

$$M_t=\{(q,p)|p=\nabla R(s,x,t,q)\}.$$

Using Lemma 2 directly we have, for $t>s$,

$$\nabla R(s,x,t,y)=\dot{E}_s(t)E_s^{-1}(t)q-E_s^{-T}(t)x.$$

Hence, as E_s is invertible and isotropic on the interval $(s,u]$, we obtain (28) for the singular Lagrangian manifold case with $\delta=x$:

$$M_t=\{(q,p)|E_s^T(t)p-\dot{E}_s^T(t)q=-x\}.$$

Suppose now that we choose another associated function H_t to M_s on I. In particular H_t will be another solution of (17) on I, hence we will have the following relation between F_t and H_t:

$$H_t=F_t[F_s^{-1}H_s+G(t)\omega(\mathbf{F}_t,\mathbf{H}_t)], \tag{47}$$

where

$$G(t)=\int_s^l F_\tau^{-1}F_\tau^{-T}\,d\tau.$$

Applying Lemma 2, $\omega(\mathbf{F}_t,H_t)=0$ which implies

$$H_t = F_t F_s^{-1} H_s . \tag{48}$$

Consequently, for $(q,p) \in M_t$

$$\dot{H}_t^T p - H_t^T q = [F_s^{-1} H_s]^T [\dot{F}_t^T p - F_t^T q] = [F_s^{-1} H_s]^T F_s^T \beta = H_s^T \beta,$$

which proves that $M_t \subset M_t^H$, where

$$M_t^H = \{(q,p) | \omega(\mathbf{H}_t, \mathbf{I}_{q,p}) = -H_s^T \beta\}.$$

In a similar way we could prove the reverse statement. Hence

$$M_t = M_t^H .$$

□

Now instead of giving an initial Lagrangian manifold or an initial momenta field we are given a final one. Consider the following Lagrangian manifolds,

$$R_u = \{(q,p) | -p = \hat{\alpha} q - \hat{\beta} x\}, \tag{49}$$

$$R_u^z = \{(q,p) | q = z\}, \tag{50}$$

where $\hat{\alpha}$ is a symmetric invertible $n \times n$ matrix and $\hat{\beta}, z \in \mathbb{R}^n$. The trajectories consistent with these final manifolds will be solutions $\gamma(t) = (q(t), p(t))$ of Hamilton's equation of motion on I such that $\gamma(u) \in R_u$ [or $q(u) = z$]. We want to compute for $t \in I$ the time-reversed image R_t of these manifolds.

Definition 3: Given a final Lagrangian manifold we define its *associated function* on an interval $I = [s, u]$, $\hat{F}_t$, as an isotropic matrix solution of (17) such that we have the following.

(1) If the Lagrangian manifold is defined by (49), $\hat{F}_t$ is a nonsingular matrix for $t \in I$ such that

$$-\dot{\hat{F}}_u \hat{F}_u^{-1} = \hat{\alpha}. \tag{51}$$

(2) If the Lagrangian manifold is a singular, i.e., defined by (50), then $\hat{F}_t$ is a nonsingular matrix for $t \in [s,u)$ and satisfies $F(u) = 0$.

As in Lemma 1, we have the following.

Lemma 3: If $\hat{F}_t$ and $\hat{H}_t$ are two associated functions on the time interval I to a Lagrangian manifold R_u, then $\omega(\hat{\mathbf{F}}_t, \hat{\mathbf{H}}_t) = 0$.

Proposition 2: The image of the Lagrangian manifolds defined by (49) [or (50)] under the time reversed Hamiltonian flow defined by (14) is, for $t \in I$, given by

$$R_t = \{(q,p) | \dot{\hat{F}}^T(t) q - \hat{F}^T p = \hat{\delta}\}, \tag{52}$$

where $\hat{F}(t)$ is the associated function to R_u (or R_z^u) and $\hat{\delta} = \hat{F}_u^T \hat{\beta}$ (or $\hat{\delta} = z$). Moreover, the expression (52) is independent of the choice of the associated function to M_s, i.e., if $\hat{H}_t$ is another associated function to R_u on I, then

$$\{(q,p) | \dot{\hat{F}}^T(t) q - \hat{F}^T p = \hat{\delta}\} = \{(q,p) | \dot{\hat{H}}(t) q - \hat{H}^T p = \hat{\delta}^H\},$$

where $\hat{\delta}^H = H_u^T \hat{\beta}$.

Proof: The proof follows the same lines as the proof of the last proposition, i.e., R_t will now be given by

$$R_t = \left\{ (q,p) | -p = \frac{\partial \hat{S}(t,y)}{\partial y} \right\},$$

where $\hat{S}(t,y)$ is the solution for the time-reversed Hamilton–Jacobi equation

$$-\frac{\partial \hat{S}(t,y)}{\partial t}+H\left(t,y,-\frac{\partial \hat{S}(t,y)}{\partial y}\right)=0, \quad y\in\mathbb{R}^n, \quad t\in I, \tag{53}$$

satisfying

$$\hat{S}(u,z)=\hat{S}(z) \tag{54}$$

with $\hat{S}(z)$ denoting the generating function of $\hat{R}_u=\mathscr{T}(R_u)$, where $\mathscr{T}:\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ is defined as $\mathscr{T}(q,p)=(q,-p)$. □

Remark 1: It follows from Propositions 1 and 2 that the images of the Lagrangian manifolds M_s and R_u can be written in terms of the symplectic product, i.e.,

$$M_t=\{(q,p)\,|\,\omega(\mathbf{F}_t,\mathbf{I}_{q,p})=F_s^T\beta\}, \tag{55}$$

$$R_t=\{(q,p)\,|\,\omega(\hat{\mathbf{F}}_t,\mathbf{I}_{q,p})=\hat{\beta}\}, \tag{56}$$

where $\mathbf{I}_{q,p}=(\delta_{ij}q_i;\delta_{ij}p_i)_{t\leqslant i,j\leqslant n}$, $\mathbf{F}_t=(F_t,\dot{F}_t)$, and $\hat{\mathbf{F}}_t=(\hat{F}_t,\dot{\hat{F}}_t)$, with F_t and $\hat{F}_t$ denoting the solutions associated with M_s and R_u.

Remark 2: From the proof of Proposition 1 we can obtain explicit expressions for the solution of the Hamilton–Jacobi equation initial value problem given by (38) and (39). Using (45) we obtain

$$S(t,y)=\langle \dot{F}_t F_t^{-1},y\rangle - F_t^{-T}\beta y+\gamma_t. \tag{57}$$

As the above expression is a solution of the boundary value problem (38) and (39) we can easily show that γ_t is a solution on I of the first-order differential equation:

$$\dot{\gamma}_t+|F_t^{-T}\beta|^2=0, \quad \gamma(s)=\gamma_s. \tag{58}$$

Using Proposition 2 we obtain in the same way a solution of the boundary value problem (53) and (54):

$$\hat{S}(t,y)=-\langle \dot{\hat{F}}_t\hat{F}_t^{-1},y\rangle - \hat{F}_t^{-T}\hat{\beta}y+\hat{\gamma}_t. \tag{59}$$

In this case, $\hat{\gamma}_t$ solves the time-reversed equation on the time interval I, (58), i.e.,

$$\dot{\hat{\gamma}}_t+|\hat{F}_t^{-T}\hat{\beta}|^2=0, \quad \hat{\gamma}(u)=\hat{\gamma}_u. \tag{60}$$

B. Consistent solutions

Consider the pair of Lagrangian manifolds M_s and R_u defined by the associated functions F_t and $\hat{F}_t$ on the interval $I=[s,\ u]$ and the two $\mathbb{R}^n$ vectors $\delta,\hat{\delta}$.

Consider the problem of finding regular trajectories $\mathbf{m}_t\equiv(m(t),\dot{m}(t))\in\mathbb{R}^{2n}$ such that $\mathbf{m}(t)$ is a solution of Hamilton's equations of motion (15) on I satisfying the following boundary conditions:

$$\mathbf{m}(s)\in M_s, \tag{61}$$

$$\mathbf{m}(u)\in R_u. \tag{62}$$

It turns out that this boundary value problem is equivalent to finding, for each $t\in I$, the intersection $L_t\cap R_t$:

Proposition 3: Consider a pair of Lagrangian manifolds M_s and R_u, defined by the associated functions $F_t,\hat{F}_t$ on the time interval I, satisfying $\omega(\mathbf{F}_t,\hat{\mathbf{F}}_t)\neq 0$. Then, for each $t\in I$, the set $M_t\cap R_t$ is a single-element set given by

$$L_t\cap R_t=\{(\mathbf{m}_t)\} \tag{63}$$

with $\mathbf{m}_t=(m_t,\dot{m}_t)$, *where* $\mathbf{m}_t$ *is the solution of Hamilton's equations of motion (15) on I satisfying the boundary conditions given by (61) and (62). Moreover, we can write* $\mathbf{m}_t$ *as*

$$\mathbf{m}_t=\mathbf{F}_t[\omega(\hat{\mathbf{F}},\mathbf{F})]^{-1}\hat{\delta}+\hat{\mathbf{F}}_t[[\omega(\hat{\mathbf{F}},\mathbf{F})]^T]^{-1}\delta, \tag{64}$$

where $\hat{\mathbf{F}}\equiv(\hat{F}_t,\dot{\hat{F}}_t)$ *and* $\mathbf{F}_t\equiv(F_t,\dot{F}_t)$. *This solution* $\mathbf{m}_t$ *will be referred to as the "*M_sR_u *consistent solution."*

Proof: Consider (q,p) in the intersection set. Then, by Propositions 1 and 2,

$$\omega(\mathbf{F}_t,\mathbf{I}_{q,p})=-\delta,\quad \omega(\hat{\mathbf{F}}_t,\mathbf{I}_{q,p})=-\hat{\delta}.$$

Hence using the fact that $\omega(\mathbf{F}_t,\hat{\mathbf{F}}_t)\neq 0$ and solving the system we will obtain the unique solution

$$\mathbf{m}_t=\mathbf{F}_t[\omega(\hat{\mathbf{F}},\mathbf{F})]^{-1}\hat{\delta}+\hat{\mathbf{F}}_t[[\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)]^T]^{-1}\delta,$$

which is obviously a solution of the Hamilton equations of motion that satisfies the required boundary conditions. □

Remark 3: In the conditions of the above proposition we can also consider for any $t\in I$ and $y\in\mathbb{R}^n$ the pairs of Lagrangian manifolds M_s,R_t^y and M_t^y,R_u, where the Lagrangian manifolds R_t^y and M_t^y both denote the singular Lagrangian manifold at time t associated with the configuration y (we adopt different notation since R_t^y is viewed as a final condition with respect to the Lagrangian manifold M_s on the time interval $[s,\ t]$ whereas M_t^y is an initial manifold with respect with the manifold R_u on the time interval $[t,\ u]$). We can take $E_t(\tau)$ as the associated function to R_t^y on the interval $[s,\ u]$ and $E_t(\tau)$ as the associated function to the manifold M_t^y. Denoting by $\gamma_{M_s}^{R_t^y}(\tau)$ and $\gamma_{M_t^y}^{R_y^l}(\tau)$ the consistent solutions to the pairs of Lagrangian manifolds M_t^y,R_u and M_s,R_t^y, respectively, we will have, applying Proposition 3. the following expressions:

$$\gamma_{M_s,R_t^y}(\tau)=\mathbf{F}_\tau[\omega(\hat{\mathbf{E}}_t(\tau),\mathbf{F})]^{-1}y+\hat{\mathbf{E}}_t(\tau)[[\omega(\hat{\mathbf{E}}_t(\tau),F)]^T]^{-1}\delta,$$

$$\gamma^{M_t^y,R_u}(\tau)=\mathbf{E}_t(\tau)[\omega(\hat{\mathbf{F}}(\tau),\mathbf{E}_t)]^{-1}\hat{\delta}+\hat{\mathbf{F}}(\tau)[[\omega(\hat{\mathbf{F}}(\tau),\mathbf{E}_t)]^T]^{-1}y,$$

but as

$$\omega(\hat{\mathbf{E}}_t(\tau),\mathbf{F}(\tau))=\hat{E}_t(t)^T\dot{F}(t)-\dot{\hat{E}}_t(t)^TF(t)=F(t),$$

$$\omega(\hat{\mathbf{F}}(\tau),\mathbf{E_t})=\hat{F}(t)^T\dot{E}_t(t)-\dot{\hat{F}}(t)^TE_t(t)=\hat{F}(t)^T,$$

we obtain

$$\gamma_{M_s,R_t^y}(\tau)=\mathbf{F}_\tau F(t)^{-1}y+\hat{\mathbf{E}}_t(\tau)[F(t)^T]^{-1}\delta, \tag{65}$$

$$\gamma^{M_t^y,R_u}(\tau)=\mathbf{E}_t(\tau)[\hat{F}^T(t)]^{-1}\hat{\delta}+\hat{\mathbf{F}}(\tau)[\hat{F}(t)]^{-1}y. \tag{66}$$

III. EUCLIDEAN (OR PROBABILISTIC) QUANTIZATION

We will show that there is a natural way to associate to a pair of Lagrangian manifolds a Gaussian probability density function defining a $\mathbb{R}^n$-valued diffusion process.

A. Associating a Gaussian probability density to a pair of Lagrangian manifolds

Proposition 3 allows us to consider a regular solution $m(t)$ of $\ddot{q}(t)=K(t)q(t)$ on $I=[s,\ u]$ with boundary values given by the pair of Lagrangians manifolds L_s and R_u. Any configuration $y\neq m(t)$ will determine two points, respectively, $(y,p_1)\in M_t$ and $(y,p_2)\in R_t$, to which a pair of

trajectories $\gamma_{M_s}^{R_t^y}(\tau)=(q^{t,y}(\tau),\dot{q}^{t,y}(\tau))$ and $\gamma_{M_t^y}^{R_u}(\tau)=(q_{t,y}(\tau),\dot{q}_{t,y}(\tau))$ can be associated where $q^{t,y}(\tau)$ and $q_{t,y}(\tau)\tau\in[s,\ u]=I$ denote, respectively, solutions of Hamilton's equation of motion satisfying the relations

$$\gamma_{M_s}^{R_t^y}(s)\in M_s,\quad \gamma_{M_t^y}^{R_u}(u)\in R_u,\quad \gamma_{M_s}^{R_t^y}(t)=(y,p_1),\quad \gamma_{M_t^y}^{R_u}(t)=(y,p_2).$$

Definition 4: Given a pair of Lagrangian manifolds M_s, R_u with associated functions defined on a time interval $I=[s,\ u]$, we will call the area associated to the pair of manifolds the symplectic inner product defined by

$$A_{M_s}^{R_u}(t,y)=\omega(\mathbf{X}_{t,y}(\tau),\mathbf{Y}^{t,y}(\tau)), \tag{67}$$

where $\mathbf{X}_{t,y}(\tau)$ and $\mathbf{Y}^{t,y}(\tau)$ are respectively the solutions of Hamilton's equations (15). defined by

$$\mathbf{X}_{t,y}(\tau)\equiv\gamma_{M_s}^{R_t^y}(\tau)-\mathbf{m}(\tau),\quad \mathbf{Y}^{t,y}(\tau)\equiv\gamma_{M_t^y}^{R_u}(\tau)-\mathbf{m}(\tau).$$

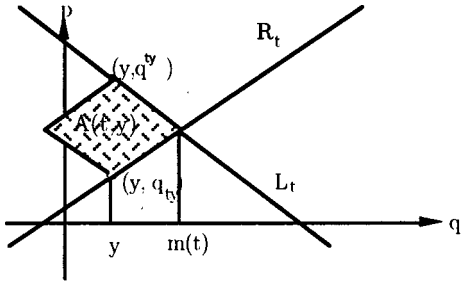

Lemma 4: Consider a pair of Lagrangian manifolds M_s, R_u defined by its associated functions F, $\hat{F}$ on $I=[s,\ u]$ and by the two vectors $\delta,\hat{\delta}\in\mathbb{R}^n$. Then we can write the associated area $A_{M_s}^{R_u}(t,y)$ as

$$A_{M_s}^{R_u}(t,y)=(y-m(t))^TC^{-1}(t)(y-m(t)), \tag{68}$$

where $m(t)$ and $C(t)$ are respectively the projection onto the configuration space of $\mathbf{m}_t$ and the consistent trajectory to the pair of Lagrangian manifold, a symmetric matrix defined by

$$m_t=F_t[\omega(\hat{\mathbf{F}},\mathbf{F})]^{-1}\hat{\delta}+\hat{F}_t[[\omega(\hat{\mathbf{F}},\mathbf{F})]^T]^{-1}\delta, \tag{69}$$

$$C(t)=F(t)[\omega(\hat{\mathbf{F}},\mathbf{F})]^{-1}\hat{F}^T(t). \tag{70}$$

Moreover, $C(t)$ is independent of the choice of the associated functions $F_t,\hat{F}_t$ to the Lagrangian manifolds M_s,R_u.

Proof: As associated functions to the pair of Lagrangian manifolds on the time interval I, they are isotropic, which implies that $\dot{F}_tF_t^{-1}$ and $\dot{\hat{F}}_t\hat{F}_t^{-1}$ are symmetric matrices on the interval I. Hence $\dot{F}_tF_t^{-1}-\dot{\hat{F}}_t\hat{F}_t^{-1}$ is also a symmetric matrix on I. However,

$$\dot{F}_tF_t^{-1}-\dot{\hat{F}}_t\hat{F}_t^{-1}=\hat{F}_t^{-T}\hat{F}_t^T\dot{F}_tF_t^{-1}-\hat{F}_t^{-T}\dot{\hat{F}}_t^TF_tF_t^{-1}=\hat{F}_t^{-T}[\hat{F}_t^T\dot{F}_t-\dot{\hat{F}}_t^TF_t]F_t^{-1}=C^{-1}(t),$$

which proves that, for all $t\in I$, $C(t)$ is a symmetric matrix.

Now we prove (68). By (67) we can write

$$A_{M_s}^{R_u}(t,y)=\omega(\gamma_{M_s}^{R_t^y}(\tau),\gamma_{M_t^y}^{R_u}(\tau))-\omega(\gamma_{M_s}^{R_t^y}(\tau),\mathbf{m}(\tau))-\omega(\mathbf{m}(\tau),\gamma_{M_t^y}^{R_u}(\tau)). \tag{71}$$

Using (65) and (66), we compute the first term of (71):

$$\omega(\gamma_{M_s}^{R_t^y}(\tau),\gamma_{M_t^y}^{R_u}(\tau))=y^T\hat{F}^{-T}\hat{\delta}+y^TF^{-T}\omega(\mathbf{F},\hat{\mathbf{F}})\hat{F}^{-1}y+\delta^TF^{-1}y, \tag{72}$$

but

$$\begin{aligned} y^T\hat{F}^{-T}\hat{\delta}&=y^T\hat{F}^{-T}[F\omega^{-1}(\hat{\mathbf{F}},\mathbf{F})]^{-1}[F\omega^{-1}(\hat{\mathbf{F}},\mathbf{F})]\hat{\delta}\\ &=y^T\hat{F}^{-T}\omega(\hat{\mathbf{F}},\mathbf{F})F^{-1}[F\omega^{-1}(\hat{\mathbf{F}},\mathbf{F})]\hat{\delta}=y^Tc^{-1}(t)F\omega^{-1}(\hat{\mathbf{F}},\mathbf{F})\hat{\delta} \end{aligned} \tag{73}$$

and

$$\delta^TF^{-T}y=\delta^TF^{-T}[\omega^T(\hat{\mathbf{F}},\mathbf{F})\hat{F}^{-1}][\omega^T(\hat{\mathbf{F}},\mathbf{F})\hat{F}^{-1}]^{-1}y=\delta^Tc^{-T}\hat{F}\omega^{-T}(\hat{\mathbf{F}},\mathbf{F})y=y^Tc^{-1}\hat{F}\omega^{-T}(\hat{\mathbf{F}},\mathbf{F})\delta. \tag{74}$$

Hence

$$\begin{aligned} \omega(\gamma_{M_s}^{R_t^y}(\tau),\gamma_{M_t^y}^{R_u}(\tau))&=y^Tc^{-1}[F\omega^{-1}(\hat{\mathbf{F}},\mathbf{F})\hat{\delta}+\hat{F}\omega^{-T}(\hat{\mathbf{F}},\mathbf{F})\delta]+y^TF^{-T}\omega(\mathbf{F},\hat{\mathbf{F}})\hat{F}^{-1}y\\ &=y^Tc^{-1}m(t)+y^Tc^{-1}(t)y=y^Tc^{-1}(t)(m(t)-y). \end{aligned} \tag{75}$$

By similar computations, we also verify that

$$\omega(\gamma_{M_s}^{R_t^y}(\tau),\mathbf{m}(\tau))=\delta^TF^{-1}(m(t)-y), \tag{76}$$

$$\omega(\gamma_{M_t^y}^{R_u}(\tau),\mathbf{m}(\tau))=(y-m(t))^T\hat{F}^{-T}\hat{\delta}. \tag{77}$$

Using (71), (75), (76), and (77) we will obtain

$$\begin{aligned} A_{M_s}^{R_u}(t,y)&=\omega(\gamma_{M_s}^{R_t^y}(\tau),\gamma_{M_t^y}^{R_u}(\tau))-\omega(\gamma_{M_s}^{R_t^y}(\tau),\mathbf{m}(t))-\omega(\mathbf{m}(t),\gamma_{M_t^y}^{R_u}(\tau))\\ &=y^Tc^{-1}(t)(m(t)-y)-\delta^TF^{-1}(m(t)-y)-(y-m(t))^T\hat{F}^{-T}\hat{\delta}\\ &=(y-m_t)C^{-1}(t)(y-m_t). \end{aligned}$$

Now we prove that $C(t)$ is independent of the choice of the associated functions to the Lagrangian manifolds. Let us denote by H_t another function associated to the Lagrangian manifold L_s. Using the argument of the proof of Proposition 2, we have the following relationship between F_t and H_t for $t\in I$:

$$H_t=F_tF_s^{-1}H_s.$$

Therefore,

$$\begin{aligned} H_t[\omega(\hat{\mathbf{F}}_t,\mathbf{H}_t)]^{-1}\hat{F}_t^T&=F_tF_s^{-1}H_s[\omega(\hat{\mathbf{F}}_t,\mathbf{F}_tF_s^{-1}H_s)]^{-1}\hat{F}_t^T\\ &=F_tF_s^{-1}H_s[\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)F_s^{-1}H_s]^{-1}\hat{F}_t^T\\ &=F_t[\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)]^{-1}\hat{F}_t^T=C(t). \end{aligned}$$

A similar argument is valid if one replaces $\hat{F}_t$ by another associated function to the Lagrangian manifold R_u. □

Definition 5: A pair of Lagrangian manifolds M_s, R_u is called a compatible pair on the time interval $I=[s,u]$ if

(i) their associated functions $F_t,\hat{F}_t$ are defined on I, and

(ii) $C(t)$, defined by (70), is positive definite on the interior of I and non-negative definite for $t=s$ and $t=u$.

Proposition 4: Given an arbitrary final (resp. initial) Lagrangian manifold with an associated function $\hat{F}_t$ (resp. F_t) defined on the interval $I=[s,u]$, there exists a family of compatible initial (resp. final) Lagrangian manifolds.

Proof: Consider a final Lagrangian manifold R_u and $\hat{F}_t$, its associated function defined on the interval I. For any $n\times n$ nonsingular matrix σ, let us denote by P the positive definite matrix

$$P\equiv\sigma\sigma^T. \tag{78}$$

We now define a family F_t^P of solutions of the matrix differential equation on I,

$$\ddot{F}_t=K(t)F_t$$

[where, as usual, $K(t)$ denotes a symmetric nonsingular matrix on I], by

$$F_t^P=\hat{F}_tP+\hat{F}_t\int_s^t\hat{F}_\lambda^{-1}F_\lambda^{-T}\,d\lambda, \tag{79}$$

which, in turn, will allow us to define a family of initial Lagrangian manifolds, $M_s^{P,\delta}$:

$$M_s^{P,\delta}=\{(q,p):p=\dot{F}_s^P[F_s^P]^{-1}q-[F_s^P]^{-T}\delta\},$$

compatible with R_u. Observe that for any P fulfilling (78) we have

$$\omega(\hat{F}_t,F_t^P)=I_{n\times n}. \tag{80}$$

An analogous argument is valid if one starts from a fixed initial Lagrangian manifold M_s. □

Remark 4: In fact we can always choose, for two compatible Lagrangian manifolds on I, associated functions such that (80) holds.

Indeed consider F_t and $\hat{F}_t$ any associated functions for the pair of Lagrangian manifolds on the time interval I. If (80) is not verified, we can replace $\hat{F}_t$ by $H\equiv\hat{F}_t\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)^{-T}$ (which is also a function on I associated to the Lagrangian manifold R_u). Then

$$\begin{aligned}\omega(\mathbf{H}_t,\mathbf{F}_t)&=\omega(\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)^{-T}\hat{\mathbf{F}},\mathbf{F}_t)\\&=(\hat{F}_t\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)^{-T})^T\dot{F}_t-(\dot{\hat{F}}\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)^{-T})^TF_t\\&=\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)^{-1}\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)=I_{n\times n}.\end{aligned}$$

Proposition 5: If R_u (resp. M_s) is a final (resp. initial) Lagrangian manifold with associated function $\hat{F}_t$ (resp. F_t) on I, then the pair $M_{t_0}^y,R_u$ (resp. $M_s,R_{t_0}^y$) is compatible in the interval $[t_0,\,u]$ (resp. $[s,\,t_0]$) and for all configurations $y\in\mathbb{R}^n$.

Proof: A solution associated to the singular manifold $M_{t_0}^x$ on the time interval $[t_0,\,u]$ is given by $E_{t_0}(t)$ [i.e., the solution of the matrix equation (17) satisfying the Cauchy data $E_{t_0}(t_0)=0$, $\dot{E}_t0(t_0)=I_{n\times n}$]. The corresponding matrix (70) relative to this case is

$$C(t)=E_{t_0}(t)[\omega(\hat{\mathbf{F}}_t,E_{t_0}(t))]^{-1}\hat{F}(t).$$

To prove that the pair of Lagrangian manifolds are compatible we need to prove that the above expression is non-negative definite. For this we write $E_{t_0}(t)$ on $[t_0,\,u]$ as

$$E_{t_0}(t)=\hat{F}(t)\left[\int_{t_0}^t\hat{F}^{-1}(\lambda)\hat{F}^{-T}(\lambda)\,d\lambda\right]F^T(t_0).$$

Then we obtain for $C(t)$

$$E_{t_0}(t)[\omega(\hat{\mathbf{F}}_t,E_{t_0}(t))]^{-1}\hat{F}^T(t)=E_{t_0}(t)[\hat{F}^T(t_0)]^{-1}\hat{F}^T(t)$$
$$=\hat{F}(t)\left[\int_{t_0}^{t}\hat{F}^{-1}(\lambda)\hat{F}^{-T}(\lambda)\,d\lambda\right]F^T(t_0)[\hat{F}^T(t_0)]^{-1}\hat{F}^T(t)$$
$$=\hat{F}(t)\left[\int_{t_0}^{t}\hat{F}^{-1}(\lambda)\hat{F}^{-T}(\lambda)\,d\lambda\right]\hat{F}^T(t),$$

which is clearly a non-negative definite matrix. In the same way we can prove that, given an initial Lagrangian manifold L_s such that the associated solution $F(t)$ is invertible on the time interval $[s,\,u]$, then, for all configurations $z\in\mathbb{R}^n$ and all $t_0\in I-\{s\}$, the pair of Lagrangian manifolds M_s, $R^z_{t_0}$ are compatible.

Finally, for $s\leqslant t_1<t_2\leqslant u$ and for all configurations $x,\,z\in\mathbb{R}^n$, the pairs of singular Lagrangian manifolds $M^x_{t_1}$, $R^z_{t_2}$ are compatible. In this case the associated functions are $E_{t_1}(t)$ and $\hat{E}_{t_2}(t)\equiv -E_{t_2}(t)$, hence the corresponding $C(t)$ is given by

$$E_{t_1}(t)[\omega(\hat{\mathbf{E}}_{t_2}(t),\mathbf{E}_{t_1}(t))]^{-1}\hat{E}_{t_2}(t),$$

which is clearly a positive definite matrix. □

In a two-dimensional phase space there is no restriction on the pairs of boundary Lagrangian manifolds, i.e., all pairs of Lagrangian manifolds with associated functions on an interval I are compatible.

Let us consider the following pairs of Lagrangian manifolds in $\mathbb{R}^2$:

$$M_s=\{(q,p)\in\mathbb{R}^2|p=\alpha q-\beta\},$$

$$R_u=\{(q,p)\in\mathbb{R}^2|-p=\hat{\alpha}q-\hat{\beta}\},$$

where $\alpha,\,\hat{\alpha},\,\beta,\,\hat{\beta}\in\mathbb{R}$. Consider associated functions F_t, $\hat{F}_t$ on the time interval I. By definition of associated function, both functions have a definite sign on the interior of the interval I. Without loss of generality we can assume also that $\omega(\hat{\mathbf{F}}_t,\mathbf{F}_t)=1$. Hence two cases are possible:

(1) $F_t\hat{F}_t>0$,

(2) $F_t\hat{F}_t<0$.

In the first case, we have $C(t)>0$ on I so the Lagrangian manifolds are compatible. For the second case, we consider the following two Lagrangians manifolds:

$$M'_s=R_u\,, \tag{81}$$

$$R'_u=M_u\,. \tag{82}$$

As it can easily be verified, the correspondent functions for the pair M'_s, R'_u will be given by H_t, $\hat{H}_t$ defined as

$$H_t=\hat{F}_t\,, \tag{83}$$

$$\hat{H}_t=F_t\,. \tag{84}$$

Therefore we will have $H_t\hat{H}_t<0$ and

$$\omega(\hat{\mathbf{H}}_t,\mathbf{H}_t)=\omega(\mathbf{F}_t;\hat{\mathbf{F}}_t)=-1,$$

which implies that $C(t)>0$.

Returning to the general case, we state the main result of this section, i.e., we can associate to any compatible pair of Lagrangian manifolds a Gaussian probability density:

Theorem 1: *A pair of compatible Lagrangian manifolds M_s, R_u define on the time interval I a Gaussian probability density $\rho(t,y)$ whose mean is the consistent trajectory associated with M_s, R_u and variance $C(t)$ given by*

$$C(t)=F_t[\omega[\hat{\mathbf{F}}_t,\mathbf{F}_t)]^{-1}\hat{F}_t^T, \tag{85}$$

where F_t and $\hat{F}_t$ are the functions associated with the Lagrangian manifolds M_s and R_u.

Proof: By Lemma 4, a pair of Lagrangian manifolds defines a symplectic area $A(t,y)$ defined by

$$A(t,y)=(y-m(t))^T C^{-1}(t)(y-m(t)).$$

As we are assuming that the pair of Lagrangian manifolds are compatible, $C(t)$ in the above expression is a non-negative definite symmetric matrix which implies that

$$\rho(y,t)\ dy=\frac{1}{\sqrt{2\pi c(t)}}\exp\ (-A(t,y))\ dy$$

is a Gaussian density function having as mean the consistent trajectories and variance (85). □

B. The associated probability density and a system of adjoint heat equations

Given a pair of compatible Lagrangian manifolds L_s, R_u we know, using the results of the last section, that we have a Gaussian probability density function associated to the manifolds. We now show that this density can be expressed as a product of positive solutions of the adjoint heat equations:

$$\frac{\partial\eta}{\partial t}=H\eta, \tag{86}$$

$$-\frac{\partial\hat{\eta}}{\partial t}=H\hat{\eta}, \tag{87}$$

where $y\in\mathbb{R}^n$ and $t\in I$ and

$$H=-\tfrac{1}{2}\Delta+\langle K(t)y,y\rangle$$

with $K(t)$ denoting a symmetric $n\times n$ positive matrix with a continuous dependence in t.

Proposition 6: Consider a pair of compatible Lagrangian manifolds M_s, R_u on a time interval I. Then (up to a normalizing factor) we can write the associated density as a product of two positive solutions $\eta(t,y)$, $\hat{\eta}(t,y)$ of the adjoint heat equations (86) and (87), satisfying boundary conditions on the interval $I=[s,u]$ which are specified by the pair of Lagrangian manifolds.

Proof: Let us first consider the compatible Lagrangian manifolds on a time interval $I\equiv[s,\ u]$:

$$M_s=\{(q,p)|p=\alpha q-\beta\}, \tag{88}$$

$$R_u=\{(q,p)|-p=\hat{\alpha}q-\hat{\beta}\}, \tag{89}$$

where, as usual, α, $\hat{\alpha}$ are invertible symmetric matrix and β, $\hat{\beta}\in\mathbb{R}^n$. This pair will determine a pair of generating functions (up to additive constants γ_s and γ_u) $S(s,x)$ and $-\hat{S}(u,z)$, namely,

$$S(s,x)=\tfrac{1}{2}\langle\alpha x,x\rangle-\beta x+\gamma_s,\quad -\hat{S}(u,z)=-\tfrac{1}{2}\langle\hat{\alpha}z,z\rangle+\hat{\beta}z+\hat{\gamma}_u.$$

For any $t\in I$, the solutions $S(t,y)$ and $\hat{S}(t,y)$ of the Hamilton Jacobi equations (38) and (53) with boundary conditions given, respectively, by $S(s,x)$ and $S(u,z)$ are given, according to Propositions 1 and 2, by

$$S(t,y)=\tfrac{1}{2}\langle \dot{F}_t F_t^{-1} y,y\rangle - F_t^{-T}\beta y+\gamma_t, \tag{57$'$}$$

$$\hat{S}(t,y)=\tfrac{1}{2}\langle \dot{\hat{F}}_t \hat{F}_t^{-1} y,y\rangle - \hat{F}_t^{-T}\hat{\beta} y+\hat{\gamma}_t, \tag{59$'$}$$

where γ_t and $\hat{\gamma}_t$ are, respectively, solutions on I of the first-order differential equations (58) and (60). Adding (57) with (59) we will obtain

$$S(t,y)+\hat{S}(t,y)=\frac{A(t,y)}{2}+g, \tag{90}$$

where $A(t,y)$ is the symplectic area associated to the pair of manifolds and g denotes a reminder, to be determined. Let us take m_t, the consistent solution to the manifolds M_s, R_u. Obviously $A(t,m_t)=0$, which implies

$$g=S(t,m_t)+\hat{S}(t,m_t).$$

Using (40),

$$g=S_s(m_s)+R(s,m_s,t,m_t)+R(t,m_t,u,m_u)+\hat{S}(u,m_u)=S_s(m_s)+R(s,m_s,u,m_u)+\hat{S}(u,m_u).$$

Hence g does not depend on the variables (t,y), $t\in I$. Using Theorem 1, we can express the associated probability density, $\rho(y,t)$, as

$$\begin{aligned}\rho(t,y)&=\frac{1}{\sqrt{(2\pi)^n \det C(t)}}\exp\left(-\frac{A(t,y)}{2}\right)=\frac{1}{\sqrt{(2\pi)^n \det C(t)}}\exp(-S(t,y)-\hat{S}(t,y)+g)\\&=\frac{1}{\sqrt{(2\pi)^n \det[F_t\omega(\hat{\mathbf{F}},\mathbf{F})\hat{F}_t]}}\exp(-S(t,y)-\hat{S}(t,y)+g)\\&=\frac{\exp(-S(t,y))}{\sqrt{(2\pi)^n|\det[F_t]|}}\frac{\exp(-\hat{S}(t,y))}{\sqrt{(2\pi)^n|\det[\hat{F}_t]|}}\sqrt{\frac{(2\pi)^n}{|\omega(\hat{\mathbf{F}},\mathbf{F})|}}\exp(g),\end{aligned} \tag{91}$$

which implies that this density can be written as

$$\rho(y,t)dy=\mathbf{N}^{-1}\eta(t,y)\,\hat{\eta}(t,y)dy, \tag{92}$$

where $\mathbf{N}$ is a normalizing constant defined by

$$\mathbf{N}=\sqrt{\frac{(2\pi)^n}{|\omega(\hat{\mathbf{F}},\mathbf{F})|}}\exp(g), \tag{93}$$

and $\eta(t,y)$ and $\hat{\eta}(t,y)$ are defined by

$$\eta(t,y)=\frac{\exp(-S(t,y))}{\sqrt{(2\pi)^n|\det[F_t]|}},$$

$$\hat{\eta}(t,y)=\frac{\exp(-\hat{S}(t,y))}{\sqrt{(2\pi)^n|\det[\hat{F}_t]|}},$$

which, as we can easily verify, are positive solutions of the adjoint heat equations (86) and (87). Also, the boundary conditions satisfied by η and $\hat{\eta}$ depend on the pair of Lagrangian manifolds M_s, R_u.

Although (92) was obtained assuming that both Lagrangian manifolds were nonsingular, a similar expression follows in the case where one or both of them are singular. If one considers the

pair M_s^x, R_u (where, as before, M_s^x denotes the singular Lagrangian manifold associated with an initial configuration $x \in \mathbb{R}^n$), then, using (30) and changing $S(t,y)$ for Hamilton's principal function $R(s,x,t,y)$, we obtain

$$R(s,x,t,y)+\hat{S}(t,y)=\tfrac{1}{2}A_{s,x}(t,y)+g_1, \tag{94}$$

where $A_{s,x}(t,y)$ is the symplectic area defined by (67) and associated with the pair of Lagrangian manifolds M_s^x, R_u. Substituting, in (94), y by $m_{s,x}(t)$ (the solution consistent to the Lagrangian manifolds $M_{s,x}$ and R_u), we find that

$$g_1=R(s,x,u,m_{s,x}(u))+\hat{S}(u,m_u)=\hat{S}(s,x).$$

Adapting to this case the calculations used in (91),

$$\begin{aligned}\rho_{s,x}(t,y)&=\frac{\exp(-R_{s,x}(t,y))}{\sqrt{(2\pi)^n \det[E_s(t)]}}\frac{\exp(-\hat{S}(t,y))}{\sqrt{(2\pi)^n|\det[\hat{F}_t]|}}\sqrt{\frac{(2\pi)^n}{|\det[\omega(\hat{\mathbf{F}},\mathbf{E}_s(t))]|}}\exp(g_1)\\&=h(s,x,t,y)\,\hat{\eta}(t,y)\sqrt{\frac{(2\pi)^n}{|\det[\hat{F}(s)]|}}\exp(\hat{S}(s,x))\\&=\frac{h(s,x,t,y)\,\hat{\eta}(t,y)}{\exp\sqrt{(2\pi)^n/|\det[\hat{F}(s)]|}(-\hat{S}(s,x))}\\&=\frac{h(s,x,t,y)\,\hat{\eta}(t,y)}{\hat{\eta}(s,x)}.\end{aligned} \tag{95}$$

Here $h(s,x,t,y)$, $s \leqslant t$, is the fundamental solution of (59) which as can be easily verified, can also be written as

$$\frac{\exp(-R_{s,x}(t,y))}{\sqrt{(2\pi)^n \det[E_s(t)]}}.$$

The cases where one considers the pair of Lagrangian manifolds M_s, R_u^z or M_s^x, R_u^z are treated similarly. We would obtain, respectively,

$$\rho^{u,z}(t,y)=\frac{\eta(t,y)h(t,y,u,z)}{\eta(u,z)}, \tag{96}$$

$$\rho_{s,x}^{u,z}(t,y)=h(s,x,t,y)h(t,y,u,z). \tag{97}$$

□

C. Diffusion processes

In Sec. III A we saw how a pair of compatible Lagrangian manifolds determines a probability density and how this density is written in terms of a pair of solutions of adjoint heat equations with boundary values specified by the pair of Lagrangian manifolds. In this section, we will show that a pair of compatible Lagrangian manifolds determines a Gaussian diffusion process. Our setting will be given by the pair of compatible manifolds M_s, R_u. We will begin by the definition of an $\mathbb{R}^n$-valued diffusion process Z_t on the time interval $[s, u]$ by means of the value of the density associated with the pair M_s, R_u at time s, $\rho(s,x)$ and a transition function. This process will verify $E[Z_t]=m_t$, where E denotes the mathematical expectation and m_t denotes the consistent solution associated with the pair of Lagrangian manifolds. In this sense, we call Z_t the Euclidean quantization of m_t. It is crucial to stress that, as suggested by Feynman's path integral approach,[1] the probabilistic framework is extracted exclusively from the classical model.

Let us first define a transition probability function: Given any configuration $x \in \mathbb{R}^n$ and a Borelian $B \in \mathscr{B}$ we define for t_1, t such that $s \leqslant t_1 < t \leqslant u$ the function:

$$Q^{R_u}_{L^x_{t_1}}(t,B)=\int_B \rho^{R_u}_{L^x_{t_1}}(t,y)\ dy, \tag{98}$$

where $\rho^{R_u}_{L^x_{t_1}}(t,y)$ is the density associated to the pair of Lagrangian manifolds $L^x_{t_1}$, R_u which was given (cf. Theorem 6) by

$$\rho^{R_u}_{L^x_{t_1}}(t,y)=\frac{h(t_1,x,t,y)\,\hat{\eta}(t,y)}{\hat{\eta}(t_1,x)}.$$

We can easily verify that $Q^{R_u}_{L^x_{t_1}}(t,y)$ is a transition probability using the semigroup properties of the solution of the heat equation or using the additivity of the solutions of the Hamilton–Jacobi equation. Together with the initial probability density provided by the density associated with the pair of manifolds M_s, R_u at time $t=s$, i.e., $\rho^{R_u}_{M_s}(s,x)$ and the transition probability defined by (98), we satisfy the conditions of existence of a Gaussian diffusion process Z_t (cf. Ref. 2) called the Euclidean quantization of the classical solution consistent with M_s, R_u. This process will verify, by construction,

$$P(Z_t\in B)=\int_B \rho(t,y)\ dy. \tag{99}$$

However, we could as well start from the ''final'' information, i.e., the value of the probability density associated with the pair of Lagrangian manifolds M_s, R_u at time u, $\rho(u,z)$ and use a time-reversed version of the probability function $\dot{Q}^{R_u}_{L^x_{t_1}}(t,y)$ constructed as follows:

Consider for $t<t_2\leqslant u$ and $y_1\in\mathbb{R}^n$ the density $\rho^{R^z_{t_2}}_{M_s}(t,y)$ associated with the pair $M_sR^z_{t_2}$,

$$\rho^{R^z_{t_2}}_{M_s}(t,y)=\frac{\eta(t,y)h(t,y,t_2,z)}{\eta(t_2,z)},$$

and define for $x\in\mathbb{R}^n$. $B\in\mathscr{B}$, and t_1,t_2 such that $s\leqslant t_1<t_2\leqslant u$ the function

$$Q^{R^z_{t_2}}_{M_s}(t_1,B)=\int_B \rho^{R^z_{t_2}}_{M_s}(t_1,y)\ dy, \tag{100}$$

which is, as is easily verified, a (backward) transition probability function. This will enable us to construct an $\mathbb{R}^n$-valued Gaussian diffusion process Y_t with a backward transition function given by (100) and final probability given by the probability density associated with the pair of manifolds M_s, R_u at time $t=u$, $\rho^{R_u}_{M_s}(u,z)$. This process satisfies (cf. Ref. 20) as well

$$P(Y_t\in B)=\int_B \rho(t,y)\ dy, \tag{101}$$

Y_t and Z_t, as diffusion processes having the same probability density for all t in $[s,\ u]$. We will call this process the Euclidean quantization of the M_sR_u consistent trajectory. Summarizing, we have the following.

Theorem 2: *Given any pair of compatible manifolds on an interval $I=[s,u]$, there exists a Gaussian diffusion process Z_t such that* (i) $E[Z_t]=m_t$, *where m_t is the solution consistent with the pair of Lagrangian manifolds. Hence in particular we have*

$$\frac{d^2}{dt^2}E[Z_t]=K(t)E[Z_t],$$

where $K(t)$ is a symmetric $n\times n$ matrix nonsingular on I. (ii) *The variance of this process is completely described by the functions F_t and $\hat{F}_t$ associated to both Lagrangian manifolds, i.e.,*

$$C(t)=F_t[\omega(\hat{F}_t,F_t)]^{-1}\hat{F}_t^T.$$

D. Uniqueness

One can see that the Euclidean quantization Z_t associated with a pair of compatible Lagrangian manifolds M_s and R_u on a time interval I solves the following Itô standard differential equations (sde's):

$$dZ_t=(A(t)Z_t+a(t))\,dt+dW_t, \tag{102}$$

$$d_*Z_t=(A^*(t)Z_t-a^*(t))\,dt+d_*W_{*t}, \tag{103}$$

with respect to the past and future filtration $(\mathscr{P}_t)$ and $(\mathscr{F}_t)$, respectively. Here d and d_* denote forward and backward differentials and W and W_* are $\mathbb{R}^n$-valued Wiener processes wrt the filtration.[20] The drift terms are given in terms of the associated functions of the Lagrangians manifolds i.e.,

$$f(t)=\dot{\hat{F}}_t\hat{F}_t^{-1},\quad a(t)=\hat{F}_t^{-T}\hat{F}_u^T\hat{\beta},$$

$$f^*(t)=\dot{F}_tF_t^{-1},\quad a^*(t)=F_t^{-T}F_s^T\beta.$$

Now we will prove the reverse statement. Consider an $\mathbb{R}^n$-valued Gaussian diffusion process Z_t with a fixed identity diffusion matrix and satisfying the following hypothesis:

(H1) The expectation of the process $m_t=E(Z_t)$ is a solution of the ordinary differential equation on *I:*

$$\ddot{m}_t=K(t)m_t. \tag{104}$$

This assumption establishes the link between the diffusion process Z_t and the classical Hamiltonian system.

(H2) The variance $C(t)$ of the process is nonzero on the interior of the time interval. Also, let us denote by C_s and C_u the non-negative values of the variance at the endpoints of the interval I, i.e., $C_s=C(s)$ and $C_u=C(u)$.

(H3) Z_t satisfies in I the following forward and backward Itô sde's:

$$dZ_t=(f(t)Z_t+\beta(t))dt+dW_t. \tag{105}$$

$$d_*Z_t=(f^*(t)Z_t-\beta(t)^*)dt+d_*W_{*t}, \tag{106}$$

where $f(t)$, $f^*(t)$ are two nonsingular symmetric matrices on the time interval I.

Lemma 5: Consider a $\mathbb{R}^n$-valued Gaussian diffusion process Z_t defined on the interval I. satisfying the assumptions (H2) and (H3). Then its variance is given by

$$C(t)=M_t^*\omega^{-1}(\mathbf{M}_t,\mathbf{M}_t^*)M_t^T, \tag{107}$$

where M_t and M_t^* satisfy the (deterministic) matrix equations on I.

$$\dot{M}_t=f(t)M_t,\quad M_s=I_{n\times n}, \tag{108}$$

$$\dot{M}_t^*=f^*(t)M_t^*,\quad M_u^*=I_{n\times n}. \tag{109}$$

Moreover, if assumption (H1) is satisfied, then the expectation m_t of the process can be written as

$$m_t=M_t^*\omega^{-1}(\mathbf{M}_t,\mathbf{M}_t^*)\,\delta+M_t\omega^{-T}(\mathbf{M}_t,\mathbf{M}_t^*)\,\delta^*. \tag{110}$$

Proof: The general solutions of (105) and (106) are given respectively by

$$Z_t = M_t\left(Z_s + \int_s^t M_\lambda^{-1} f(\lambda)\, d\lambda + \int_s^t M_\lambda^{-1}\, dW_\lambda \right), \tag{111}$$

$$Z_t = M_t^*\left(Z_u + \int_t^u [M_\lambda^*]^{-1}\beta^*(\lambda)\, d\lambda + \int_t^u [M_\lambda^*]^{-1} dW_{*,\lambda} \right), \tag{112}$$

where M_t and M_t^* satisfy (108) and the last integrals are Itô's forward and backward integrals.[16] Using (111), we can conclude (cf. Ref. 3) that the variance $C(t)$ of the process is the only positive definite matrix solution of the following ordinary differential equation:

$$\dot{C}(t) = f(t)C(t) + C(t)f(t)^T + I, \quad C(s) = C_s.$$

However, $C(t)$ is also the variance of the backward process. Hence the solution of the above ordinary differential equation must also be the solution of

$$\dot{C}(t) = f^*(t)C(t) + C(t)f^*(t)^T - I, \quad C(u) = C_u.$$

As f and f^* are both symmetric matrices, we conclude that $C(t)$ verifies

$$2f(t)C(t) + I = 2f^*(t)C(t) - I.$$

Hence

$$C(t) = (f^*(t) - f(t))^{-1}. \tag{113}$$

Using (108) and (109) we obtain

$$C(t) = (\dot{M}_t^*[M_t^*]^{-1} - \dot{M}_t M_t)^{-1} = (M^{-T}\omega(\mathbf{M}_t, \mathbf{M}_t^*)M^*)^{-1} = M_t^*\omega^{-1}(\mathbf{M}_t, \mathbf{M}_t^*)M_t^T.$$

Taking expectations on both sides of (111) and (112) and differentiating with respect to t we will obtain

$$\dot{m}_t = f(t)m_t + \beta(t), \tag{114}$$

$$\dot{m}_t = f^*(t)m_t - \beta^*(t), \tag{115}$$

obtaining

$$f(t)m_t + \beta(t) = f^*(t)m_t - \beta^*(t),$$

which implies that

$$m_t = (f^*(t) - f(t))^{-1}.$$

By (113) we obtain

$$m_t = C(t)[\beta(t) + \beta^*(t)]. \tag{116}$$

Since m_t is, by assumption, a solution of Hamilton's version of equations of motion (12), we obtain that $\beta(t)$, $\beta^*(t)$ are solutions of the differential equations

$$\dot{\beta}(t) + f(t)\beta(t) = 0,$$

$$\dot{\beta}^*(t) - f^*(t)\beta^*(t) = 0.$$

Using (108) and (109) the solutions are

$$\beta(t)=M_t^{-T}\delta, \tag{117}$$

$$\beta^*(t)=[M_t^*]^{-T}\delta^*, \tag{118}$$

where δ and δ^* are two arbitrary vectors of $\mathbb{R}^n$. Substituting in (116), the following formula holds for the expectation:

$$\begin{aligned} m_t=C(t)[\beta(t)+\beta^*(t)]&=M_t^*\omega^{-1}(\mathbf{M}_t,\mathbf{M}_t^*)M_t^T[M_t^{-T}\delta+[M_t^*]^{-T}\delta^*\\ &=M_t^*\omega^{-1}(\mathbf{M}_t,\mathbf{M}_t^*)\,\delta+M_t\omega^{-T}(\mathbf{M}_t,\mathbf{M}_t^*)\,\delta^*. \end{aligned}$$

□

Theorem 3: *Consider an $\mathbb{R}^n$-valued Gaussian diffusion Z_t on an interval I, satisfying the assumptions (H1), (H2), and (H3). Then Z_t is the Euclidean quantization of the pair of Lagrangian manifolds M_s, R_u given by*

$$M_s=\{(q,p)|p=f_s^*q-\beta^*(s)\}, \tag{119}$$

$$R_u=\{(q,p)|-p=f_uq+\beta_u\}. \tag{120}$$

Proof: Consider F_t, $\hat{F}_t$ defined as

$$F_t=M_t^*, \tag{121}$$

$$\hat{F}_t=M_t, \tag{122}$$

where M_t, M_t^* are defined by (108) and (109). As $f(t)$ and $f^*(t)$ are nonsingular symmetric matrices on the interval I, F_t and $\hat{F}_t$ are disconjugate solutions of (104). Using these functions together with δ and $\hat{\delta}$ defined by (117) and (118) we can introduce a pair of Lagrangian manifolds:

$$M_s=\{(q,p)|p=\dot{F}_sF_s^{-1}q-F_s^{-T}\delta\}, \tag{123}$$

$$R_u=\{(q,p)|-p=\dot{\hat{F}}_u\hat{F}_u^{-1}q-\hat{F}_u^{-T}\hat{\delta}\}. \tag{124}$$

They define a Gaussian process whose mean is given by

$$m_t=F_t[\omega(\hat{\mathbf{F}},\mathbf{F})]^{-1}\delta+\hat{F}_t[[\omega(\hat{\mathbf{F}},\mathbf{F})]^T]^{-1}\delta^* \tag{125}$$

$$=M_t^*[\omega(\mathbf{M}_t,\mathbf{M}_t^*)]^{-1}\delta+M_t[[\omega(\mathbf{M}_t,\mathbf{M}_t^*)]^T]^{-1}\delta^*. \tag{126}$$

and whose variance is given by

$$C(t)=M_t^*\omega^{-1}(\mathbf{M}_t,\mathbf{M}_t^*)M_t^T.$$

□

Remark 5: In Ref. 21 the authors consider diffusions satisfying the sde's (102) and (103) as ''Gaussian Bernstein processes'' associated with the quantum Hamiltonian

$$H=-\tfrac{1}{2}\Delta+\tfrac{1}{2}\langle K(t)x,x\rangle$$

in the following way: The forward and backward drifts are related to a pair of positive solutions η, η^* of the adjoint heat equations (86) and (87):

$$\frac{\nabla\eta_t}{\eta_t}=f(t)x+\beta(t), \quad -\frac{\nabla\eta_t^*}{\eta_t^*}=f^*(t)x-\beta_t^*.$$

From these two relations we can easily obtain (104). In the present paper, however, we insist on starting exclusively from the classical Hamiltonian system. Now we see that these Bernstein

processes are, in fact, associated to a pair of Lagrangian manifolds. It should be noted that in Ref. 21 the necessary assumption of symmetry on the matrices f, f^* is missing.

IV. EXAMPLES

A. The Bernstein bridge

Let H be the time-dependent quadratic Hamiltonian on $\mathbb{R}^{2n}$,

$$H=\tfrac{1}{2}|p|^2-\langle K(t)q,q\rangle,$$

where $K(t)$ denotes a symmetric nonsingular matrix on the interval $[s,\ u]$. The Bernstein bridge associated with H is characterized as a Bernstein diffusion Z_t with boundary data given by $Z_s=x$, $Z_u=z$, where x and z are two configurations in $\mathbb{R}^n$. This implies that the variance of Z_t satisfies $C(s)=C(u)=0$. By Lemma 5, using in particular (108) and (109), the matrices M_t and M_t^* are determined by the boundary conditions

$$M_s=I_{n\times n},\quad M_s^*=0,$$

$$M_u=0,\quad M_u^*=I_{n\times n},$$

which implies that

$$M_t=E_u(t)E_u^{-1}(s),\quad M_t^*=E_s(t)E_s^{-1}(u).$$

Using (109) we can write the variance of the process as

$$C(t)=E_s(t)\omega^{-1}(\hat{\mathbf{E}}_u,\mathbf{E}_s)\hat{E}_u^T(t),\quad t\in[s,u],$$

where $\hat{E}_u(t)=-E_u(t)$. Using now (109), the mean of the process is

$$m_t=E_s(t)\omega^{-1}(\hat{\mathbf{E}}_u,\mathbf{E}_s)z+\hat{E}_u(t)\omega^{-T}(\hat{\mathbf{E}}_u,\mathbf{E}_s)x.$$

The formulas for the mean and variance suggest that we can rescale (M_t and M_t^*) so that

$$M_t=\hat{E}_u(t),\quad M_t^*=E_s(t).$$

This shows that the Bernstein bridge is the Euclidean quantization associated with the pair of Lagrangian manifolds:

$$M_s^x=\{(q,p)|q=x\},\tag{127}$$

$$R_u^z=\{(q,p)|q=z\}.\tag{128}$$

The stochastic differential equations satisfied by the process Z_t are, respectively,

$$dZ_t=\left[\frac{\dot{\hat{E}}_u(t)}{\hat{E}_u(t)}\,Z_t+\frac{z}{\hat{E}_u^{-1}(t)}\right]dt+dW_t,\tag{129}$$

$$dZ_t=\left[\frac{\dot{E}_s(t)}{E_s(t)}\,Z_t-\frac{x}{E_u^{-1}(t)}\right]dt+dW_t,\tag{130}$$

with respect to the past and future filtrations.

When $K=0$ this is the rigorous version of Feynman's example (cf. pp. 179–180 of Ref. 1).

B. Brownian motion starting at $x_0\in\mathbb{R}$

The Brownian motion can be characterized as a Bernstein process associated with the free Hamiltonian

$$H=\frac{p^2}{2}$$

on $I=[0, T]$ that satisfies the forward sde

$$dZ_t=dW_t, \quad Z_0=x.$$

As the forward drift is zero, we must have $\dot{M}_t\equiv 0$ on the time interval and, as $M_T=1\forall t \in I$, we get $M_t\equiv 1$ and $\delta=0$. Using the same reasoning as in the previous example [since $Z_0=x$ we must have $C(0)=0$] we obtain $M_t^*=E_0(t)=t$. With these values, the expectation will be

$$m_t=\frac{\delta^* M_t+\delta M_t^*\delta}{\omega(\mathbf{M}_t,\mathbf{M}_t^*)}=\delta^*,$$

but as $m_0=x$ this will imply that $\delta^*=x$ and $m_t\equiv x\forall t\in I$.

As for the variance of Z_t,

$$C(t)=\frac{M_t M_t^*}{\omega(\mathbf{M}_t,\mathbf{M}_t^*)}=t.$$

Hence the associated Lagrangian manifolds will be, in this case

$$L_0=\{(q,p)|q=x\}, \tag{131}$$

$$R_T=\{(q,p)|p=0\}. \tag{132}$$

The backward stochastic differential equation satisfied by this process is

$$d^*Z_t=\left[\frac{1}{t}Z_t-\frac{x}{T-t}\right]dt+d_*W_t.$$

C. The Ornstein–Uhlenbeck process on the time interval [0,1]

The process is defined as a solution of

$$dZ_t=-\lambda Z_t+dW_t, \quad \lambda>0, \tag{133}$$

$$Z_0=x. \tag{134}$$

Equation (133) shows that we should choose $f(t)\equiv\lambda$ and $\beta(t)\equiv 0$ on the whole interval. This is equivalent to choosing a final Lagrangian manifold (i.e., at time $t=1$) of the form

$$R_1=\{(q,p)|p=\lambda q\}.$$

The missing Lagrangian manifold will be determined by the initial condition (134). Hence we should choose

$$L_0=\{(q,p)|q=x\}.$$

This will correspond to choosing $M^*(t)=e_0(t)$ and $\delta^*=x$, where $e_0(t)$ is the solution of

$$\ddot{q}(t)=\lambda^2 q(t), \quad q(0)=0, \quad \dot{q}(0)=1,$$

i.e.,

$$e_0(t)=\frac{1}{a}\sinh(at), \quad a=\text{sqrt }\lambda.$$

Hence we can conclude that this Ornstein–Uhlenbeck on the interval $[0, 1]$ can be considered as the EQ under the Hamiltonian

$$H=\frac{1}{2}\,p^2-\lambda^2\,\frac{q^2}{2}$$

of the pair of Lagrangians manifolds L_0, R_1. Moreover, the expectation of the process will be

$$m(t)=\frac{\delta M_t^*+\delta^* M_t}{\omega(M_t,M_t^*)}=\frac{xe^{-\lambda t}}{M(0)}=xe^{-\lambda t}$$

and its variance

$$c(t)=\frac{M_t M_t^*}{\omega(M_t,M_t^*)}=\frac{1}{2a}\,(1-e^{-2at}).$$

The backward differential equation will be

$$d_*Z_t=\left[\lambda\,\coth\,(at)Z_t-\frac{x}{\sinh(\lambda t)}\right]dt+dW_t.$$

If Z_0 is $N(x,c_0)$, then one can easily find that

$$M_t=\frac{1}{\lambda}\,\sinh(\lambda t)+c_0e^{-at}$$

and

$$\delta^*=\frac{x}{c_0},$$

which will give us mean and variance

$$m_t=\left[\frac{x}{c_0}\,\frac{1}{\lambda}\,\sinh(\lambda t)+c_0e^{-at}\right], \tag{135}$$

$$c_t=\frac{1}{\lambda}\,e^{-\lambda t}\,\sinh(\lambda t)+c_0e^{-2\lambda t}. \tag{136}$$

The backward differential equation will be

$$d_*Z_t=\left[\lambda\,\frac{1-2\lambda c_0+e^{2\lambda t}}{-1+2\lambda c_0+e^{2\lambda t}}\,Z_t-\frac{2\lambda e^{\lambda t}x}{c_0(-1+\lambda c_0+e^{2\lambda t})}\right]dt+dW_t.$$

[1] R. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965).
[2] N. Ikeda and S. Watanabe, *Stochastic differential equations and diffusion processes* (North Holland, Amsterdam, 1981).
[3] L. Arnold, *Stochastic Differential Equations* (Krieger, Malabar, FL, 1991).
[4] E. Schrödinger, Ann. Inst. Henri Poincaré **2**, 269 (1932).
[5] A. Beurling, Ann. Math. **72**, 189 (1960)
[6] J. C. Zambrini, J. Math. Phys. **27**, 2307 (1986).
[7] S. Albeverio, K. Yasue, and J. C. Zambrini, Ann. Inst. Henri Poincaré Phys. Theor. **49(3)**, 259 (1989)
[8] T. Kolsrud and J. C. Zambrini, ''The general mathematical framework of Euclidean Quatum Mechanics,'' in *Stochastic Analysis and Applications*, edited by A. B. Cruzeiro and J. C. Zambrini, Progress in Probability, Vol. 26 (Birkhäuser Verlag, Boston, 1991).
[9] B. Jamison, Z. Wahrscheinlichkeitstheor. Verwandte Geb. **30**, 65 (1974).
[10] J. C. Zambrini, ''From Quantum Physics to Probability theory and back,'' in *Chaos, the interplay between stochastic and deterministic behaviour*, edited P. Garbaczewski, M. Wolf, and A. Weron, Springer Lect. Notes in Physics, Vol. 475, p. 393 (Springer Verlag, Berlin, 1995).
[11] A. Brandão and T. Kolsrud, ''Phase Space Transformations of Gaussian Diffusions,'' Preprint, TRITA-MAT-1996-MA-21, Stockholm (1996) to appear in Pot. Analysis.

[12] S. Albeverio, K. Yasue, and J. C. Zambrini, Ann. Inst. Henri Poincaré: Phys. Theor. **49(3)**, 259 (1989).
[13] M. Kak, ''On the connections between probability theory and differential and integral equations,'' in Proceedings of the 2nd Berkeley Symposium on Probability and Statistics, edited by J. Neyman (Univ. of California, Berkeley, 1951).
[14] M. Thieullen and J. C. Zambrini, Ann. Inst. Henri Poincaré: Phys. Theor. **67**, 31 (1977).
[15] W. A. Coppel, Lect. Notes Math. **220**, (1974).
[16] W. T. Reid, *Sturmian Theory for Ordinary Differential Equations* (Springer-Verlag, New York, 1980).
[17] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, 1980).
[18] R. G. Littlejohn, J. Stat. Phys. **68**, 7 (1992).
[19] A. Brandão, ''Estruturas Simpleticas e estruturas Gaussianas associadas,'' PhD. thesis, Faculdade de Ciências da Universidade de Lisboa, 1995.
[20] T. Kolsrud and J. C. Zambrini, J. Math. Phys. **33**, 1301 (1992).
[21] J. C. Zambrini and A. B. Cruzeiro, ''Ornestein-Uhlenbeck processes as Bernstein processes,'' in *Progress in Probability 32*, edited by M. Sanz Solé and D. Nualart, Barcelona Seminar on Stochastic Analysis, St. Feliu de Guixolls (Birkhäuser Verlag, Basel, 1993), pp. 40–61.

The effect of configuration mixing on the first-order moments of polarized hydrogen lines

Roberto Casini
High Altitude Observatory, National Center for Atmospheric Research,[a] P.O. Box 3000, Boulder, Colorado 80307-3000

(Received 30 March 1998; accepted for publication 1 May 1998)

We extend previous work on the algebraic first-order moments of polarization profiles of hydrogen lines in the presence of external fields, to include the effect of configuration mixing induced by the applied fields. An algebraic formula for hydrogen line shifts due to the quadratic Stark effect is derived, under the assumption that the applied electric field is weak enough that line quenching induced by field ionization can be neglected. The derivation of such a formula is made possible by an aimed use of standard techniques of projection-operator algebra and of the fundamental implications of the spectral theorem. The results of this paper should help in decreasing the amount of numerical work needed in the calculation of line shifts of hydrogen lines induced by the presence of electric fields, as the preliminary numerical calculation of the Hamiltonian eigenvalues and eigenvectors is avoided, thus making calculations tractable for lines of higher hydrogen series than considered so far, e.g., in applications of pressure-broadening theory. © *1998 American Institute of Physics.* [S0022-2488(98)02609-7]

I. INTRODUCTION

In previous work[1,2] we gave the definition of the Nth-order moments of the Stokes profiles of hydrogen lines, and explicitly worked out the algebraic calculation of the first- and second-order moments as functions of the external fields and of the principal quantum numbers of the transition levels. Those calculations required two fundamental assumptions. The first assumption was that atomic polarization, i.e., the presence of anisotropy in the magnetic level populations and of quantum coherences among state vectors representing those levels, were negligible. The second assumption was that the intensity of the external fields were such that configuration mixing, i.e., the interaction of Hamiltonian eigenstates from different atomic configurations, which in the absence of external fields would be distinguished by different principal quantum numbers, could also be neglected. This last assumption, in particular, implied an upper bound to the intensity of the external fields that were tractable, depending on the principal quantum numbers involved (see Ref. 3, Fig. 1).

In the present work we relax this last constraint, thus allowing for the interaction of distinct atomic configurations. In particular, we will determine how centers of gravity of polarized hydrogen lines are affected by configuration mixing induced by the presence of electric fields (quadratic Stark effect), a phenomenon of relevant importance, for instance, in applications of the theory of pressure broadening of hydrogen lines (see, e.g., Ref. 4).

A limiting assumption that we will consider is that the applied electric fields must be weak enough that line quenching due to field ionization of the atom (see, e.g., Refs. 5–7) can be neglected. Using simple classical arguments, it is possible to provide, for any transition, a qualitative (yet safe) estimate of the maximum electric-field strength allowed by the above assumption as a function of the principal quantum number of the upper level of the transition, n. Following Ref. 6, such an estimate gives

$$E_{\max}(n) \approx 3.2 \times 10^{8} n^{-4} \text{ V cm}^{-1}. \tag{1}$$

[a]The National Center for Atmospheric Research is sponsored by the National Science Foundation.

 © 1998 American Institute of Physics

In spite of the above constraint, the applied electric field can still be strong enough to produce observable shifts of the line center. For instance, Ref. 5 reports a 5% relative shift of the outermost intense components of $H\gamma\,(n=5)$ for a field of 4×10^5 V cm^{-1}, whereas the critical value of the field for line quenching of the same line is estimated around 7×10^5 V cm^{-1}. Equation (1) gives in this case $E_{\max}(5)\approx5.1\times10^5$ V cm^{-1}, showing that a classical treatment of field ionization leads to a rather pessimistic estimate of the critical field strength for line quenching.

II. PERTURBATION EXPANSION OF THE *N*TH-ORDER MOMENTS OF POLARIZED HYDROGEN LINES

In this section we will derive a rigorous formulation of the *N*th-order moments of polarized hydrogen lines in terms of suitable operators, $\mathscr{O}_{\prime}^{(k)}$, $0\leqslant k\leqslant N$, which are known functions of the eigenvalues and eigenvectors of the Hamiltonian $\mathscr{H}$ for the hydrogen atom in the presence of external fields [cf. Eq. (6) below]. We will then consider a truncated second-order perturbation expansion of the Hamiltonian eigenvalues and eigenvectors, which will enable us to derive an approximate (to second order of perturbation) yet workable expression of the operators $\mathscr{O}_{\prime}^{(k)}$. The approximations involved by this second-order perturbation approach will be extensively discussed later in this section.

We start from the definition of the *N*th-order moments of Stokes profiles for the transition between two hydrogen levels of principal quantum numbers n and m $(n>m)$, assuming for the moment that no line broadening mechanism is present [cf. Ref. 2, Eq. (3)],

$$\langle\omega_{\mathbf{k}}^N(i)\rangle=\frac{\Sigma_{\nu\mu}s_{\nu\mu}(i,\mathbf{k})(\omega_{\nu\mu}-\overline{\omega})^N}{\Sigma_{\nu\mu}s_{\nu\mu}(0,\mathbf{k})},\quad i=0,1,2,3. \tag{2}$$

In the above equation, $s_{\nu\mu}(i,\mathbf{k})$ are the strengths of the polarized components of the line at the frequencies $\omega_{\nu\mu}$, where ν and μ, respectively, span the dimensions of the subspaces $\mathscr{E}_n$ and $\mathscr{E}_m$ of the levels n and m. (In the following, a latin index will always represent a principal quantum number of the discrete spectrum, or the energy parameter for the continuum, whereas a greek index will stand for a corresponding degeneracy parameter or set of parameters.) The index i enumerates the four Stokes parameters (I,Q,U,V), and $\mathbf{k}$ is the propagation vector [specifying, in particular, the direction of the line of sight (LOS)]. For the frequency $\overline{\omega}$, about which the moments are calculated, we choose the center of gravity of the line *in the absence of external fields*. With this choice, the first-order moments, as given by Eq. (2), are vanishing when no external fields are present.[1] The normalization of the *N*th-order moments to the integrated intensity, $\Sigma_{\nu\mu}\,s_{\nu\mu}(0,\mathbf{k})$, is chosen because, in the absence of atomic polarization and neglecting configuration mixing, the integrated polarization profiles, $\Sigma_{\nu\mu}\,s_{\nu\mu}(i,\mathbf{k})$, are vanishing for $i=1,2,3$ (see Ref. 1, Appendix A). In this paper, in fact, we will show that electric-field induced configuration mixing is responsible for a nonvanishing (second-order) contribution to such integrated profiles. On the other hand, we will see that such contribution can be safely neglected in our derivation of the algebraic first-order moments.

The strengths $s_{\nu\mu}(i,\mathbf{k})$ are quadratic expressions of the matrix elements of the radial vector operator, $\mathbf{r}$, between Hamiltonian states of the two levels of the transition $n-m$. These expressions assume a particularly convenient form when use is made of the (complex) spherical-vector basis

$$\begin{cases}\mathbf{e}_{\pm1}=\mp\dfrac{1}{\sqrt{2}}\,(\mathbf{e}_{x_1}\pm i\mathbf{e}_{x_2}),\\ \mathbf{e}_0=e_{x_3},\quad x_3\equiv\text{quantization axis},\end{cases}$$

in terms of which a vector operator, $\mathbf{v}$, has expansion $\mathbf{v}=\Sigma_q(-1)^q v_q\mathbf{e}_{-q}$, $q=0,\pm1$. We then have [see, e.g., Ref. 1, Eq. (4)],

$$s_{\nu\mu}(i,\mathbf{k})=\sum_{qq'}(-1)^{q+q'}\mathscr{T}_{-q-q'}(i,\mathbf{k})(r_q)_{n\nu,m\mu}(r_{q'})^*_{n\nu,m\mu},\quad q,q'=0,\pm1, \tag{3}$$

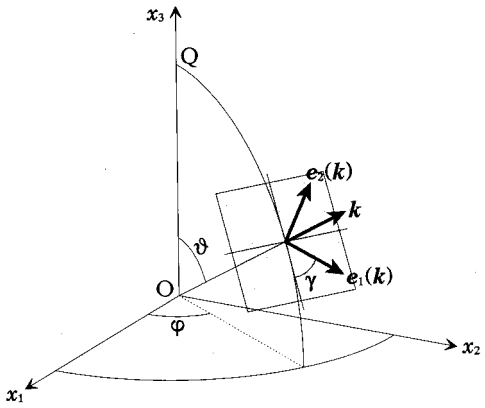

FIG. 1. Frame of reference of the quantization axis (x_3 axis). The plane of polarization, generated by the basis vectors for linear polarization, $\mathbf{e}_1(\mathbf{k})$ and $\mathbf{e}_2(\mathbf{k})$, is normal to the propagation vector, $\mathbf{k}$. The position angle γ fixes the reference direction for linear polarization, $\mathbf{e}_1(\mathbf{k})$, along which the observable $\frac{1}{2}(I+Q)$ is measured (adapted from Ref. 1).

where the (reducible) spherical tensor $\mathscr{T}_{qq'}(i,\mathbf{k})$ depends only on the geometry of the observer (i.e., LOS and position angle of the reference axis for Q-polarization measurement; see Fig. 1) relative to the frame of reference of the quantization axis. The definition and explicit expressions of the components $\mathscr{T}_{qq'}(i,\mathbf{k})$ are also given in Ref. 1 [cf. Eqs. (2) and (3) in that paper]; however, as we will not need them for the following development, they are not reported here.

If we expand the binomial power in Eq. (2) and recall Eq. (3), we see that we ultimately have to evaluate quantities of the form

$$\sigma^{(j)}_{nm}(i,\mathbf{k}) \equiv \sum_{\nu\mu} s_{\nu\mu}(i,\mathbf{k})\omega^j_{\nu\mu} = \sum_{qq'} (-1)^{q+q'} \mathscr{T}_{-q-q'}(i,\mathbf{k}) S^{(j)}_{nm}(q,q'), \quad 0 \leqslant j \leqslant N, \tag{4}$$

where, assuming for convenience of notation $\hbar = 1$ (so that frequencies have the dimensions of energy), and indicating with $\omega_{\ell\lambda}$ and $|\psi_{\ell\lambda}\rangle$ the eigenvalues and *orthonormal* eigenvectors of $\mathscr{H}$ for the generic energy level of principal quantum number ℓ,

$$\begin{aligned} S^{(j)}_{nm}(q,q') &\equiv \sum_{\nu\mu} (r_q)_{n\nu,m\mu} (r_{q'})^*_{n\nu,m\mu} \omega^j_{\nu\mu} \\ &= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \sum_{\nu\mu} \omega^{j-k}_{n\nu} \omega^k_{m\mu} \langle \psi_{n\nu}|r_q|\psi_{m\mu}\rangle \langle \psi_{m\mu}|r^\dagger_{q'}|\psi_{n\nu}\rangle, \end{aligned} \tag{5}$$

having introduced in the last equivalence the bra-ket notation for the matrix elements of the operator r_q, and also considered that $\omega_{\nu\mu} = \omega_{n\nu} - \omega_{m\mu}$.

It is convenient at this point to introduce a suitable orthonormal basis on the Hilbert space of the Hamiltonian of the hydrogen atom. Consistent with Ref. 1, we choose to adopt a basis of degenerate eigenvectors of the Schrödinger Hamiltonian,

$$\mathscr{H}_0 = \frac{p^2}{2m_e} - \frac{e_0^2}{r}$$

(m_e and e_0 are, respectively, the mass and the charge of the electron). According to our notation, such a basis can be chosen in the form $\{|a\,\alpha\rangle\}$, where a is the energy parameter (which reduces to the principal quantum number for a state of the discrete spectrum) and $\alpha \equiv \{(ls)(ji)fm_f\}$. The quantum numbers l, $s=1/2$, j, $i=1/2$, f, and m_f, which account for the high degeneracy of the energy states of the Schrödinger Hamiltonian (for the discrete state of principal quantum number ℓ, the order of degeneracy is $4\ell^2$), are associated, respectively, to the commuting (with $\mathscr{H}_0$ and among themselves) operators L^2, $S^2=3/4$, J^2, $I^2=3/4$, F^2, and F_z, where $\mathbf{L}$ is the electron orbital angular momentum (a.m.), $\mathbf{S}$ is the electron spin a.m., $\mathbf{J}=\mathbf{L}+\mathbf{S}$ is the electron total a.m., $\mathbf{I}$ is the nuclear (proton) spin a.m., $\mathbf{F}=\mathbf{J}+\mathbf{I}$ is the total (electron+proton) a.m., and F_z is the projection of $\mathbf{F}$ on the quantization axis.

Because of the form of Eq. (5), it proves useful to introduce the operator

$$\mathscr{O}_{\ell}^{(k)} \equiv \sum_{\lambda} \omega_{\ell\lambda}^{k} |\psi_{\ell\lambda}\rangle\langle\psi_{\ell\lambda}|, \tag{6}$$

and consider the resolution of the identity, $\mathbf{S}_{a\alpha}|a\alpha\rangle\langle a\alpha| = 1$ (where $\mathbf{S}$ is the symbol of generalized summation over the states of the discrete and continuous spectra of the Schrödinger Hamiltonian). Equation (5) then becomes

$$\begin{aligned}
S_{nm}^{(j)}(q,q') &= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \sum_{\mu} \omega_{m\mu}^{k} \langle\psi_{m\mu}| r_{q'}^{\dagger} \mathscr{O}_{n}^{(j-k)} r_q |\psi_{m\mu}\rangle \\
&= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \mathbf{S}_{a\alpha,a'\alpha'} \langle a\alpha|\mathscr{O}_{n}^{(j-k)}|a'\alpha'\rangle \sum_{\mu} \omega_{m\mu}^{k} \langle a'\alpha'|r_q|\psi_{m\mu}\rangle\langle\psi_{m\mu}|r_{q'}^{\dagger}|a\alpha\rangle \\
&= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \mathbf{S}_{a\alpha,a'\alpha'} \langle a\alpha|\mathscr{O}_{n}^{(j-k)}|a'\alpha'\rangle\langle a'\alpha'|r_q \mathscr{O}_{m}^{(k)} r_{q'}^{\dagger}|a\alpha\rangle \\
&= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \mathbf{S}_{\substack{a\alpha,a'\alpha' \\ b\beta,b'\beta'}} \langle a\alpha|\mathscr{O}_{n}^{(j-k)}|a'\alpha'\rangle\langle a'\alpha'|r_q|b\beta\rangle\langle b\beta|\mathscr{O}_{m}^{(k)}|b'\beta'\rangle \\
&\quad \times \langle b'\beta'|r_{q'}^{\dagger}|a\alpha\rangle.
\end{aligned} \tag{7}$$

If configuration mixing is at all negligible, the operator $\mathscr{O}_{\ell}^{(k)}$ projects onto the subspace $\mathscr{E}_{\ell}$ of the level ℓ, since $|\psi_{\ell\lambda}\rangle \in \mathscr{E}_{\ell}$ in that case. In particular, since the eigenvectors $|\psi_{\ell\lambda}\rangle$ are normalized to unity, for $k=0$ we have $\mathscr{O}_{\ell}^{(0)} = \mathscr{P}_{\ell}$, while for $k \neq 0$ we have $\mathscr{O}_{\ell}^{(k)} = (\mathscr{P}_{\ell}\mathscr{H}\mathscr{P}_{\ell})^k$, $\mathscr{P}_{\ell}$ being the projection operator onto $\mathscr{E}_{\ell}$ (i.e., $\mathscr{P}_{\ell} = \Sigma_{\lambda}|\psi_{\ell\lambda}\rangle\langle\psi_{\ell\lambda}|$, or $\mathscr{P}_{\ell} = \Sigma_{\lambda}|\ell\lambda\rangle\langle\ell\lambda|$ using the representation basis $\{|a\alpha\rangle\}$). This also means that, in the absence of configuration mixing, n and m are ''good'' quantum numbers.

If instead configuration mixing cannot be neglected, then $\mathscr{O}_{\ell}^{(k)}$ operates into the whole Hilbert space of the Hamiltonian $\mathscr{H}$, since the eigenvectors $|\psi_{\ell\lambda}\rangle$ in general have components in other subspaces besides $\mathscr{E}_{\ell}$. This is easily understood if we consider orders of approximation of $|\psi_{\ell\lambda}\rangle$ higher than zeroth, within the framework of the Rayleigh–Schrödinger theory of stationary perturbations for the degenerate case (see, e.g., Ref. 8). Assuming that the eigenvectors in the absence of configuration mixing, $|\psi_{\ell\lambda}^{(0)}\rangle$, are orthogonal and normalized to unity, while choosing for the perturbed eigenvectors the normalization $\langle\psi_{\ell\lambda'}^{(0)}|\psi_{\ell\lambda}\rangle = \delta_{\lambda\lambda'}$, we have, to second order of perturbation,

$$|\psi_{\ell\lambda}\rangle = |\psi_{\ell\lambda}^{(0)}\rangle + \epsilon|\psi_{\ell\lambda}^{(1)}\rangle + \epsilon^2|\psi_{\ell\lambda}^{(2)}\rangle \tag{8a}$$

(as is customary, the dimensionless parameter ϵ is introduced only for convenience, in order to keep track of the perturbation order, and must be put equal to unity at the end of the calculations), where $|\psi_{\ell\lambda}^{(0)}\rangle \in \mathscr{E}_{\ell}$, and

$$|\psi_{\ell\lambda}^{(1)}\rangle = \mathbf{S}'_{s\sigma} \frac{|s\sigma\rangle\langle s\sigma|W|\psi_{\ell\lambda}^{(0)}\rangle}{\omega_{\ell}^{(0)} - \omega_{s}^{(0)}}, \tag{8b}$$

$$|\psi_{\ell\lambda}^{(2)}\rangle = \mathbf{S}'_{s\sigma,t\tau} \frac{|s\sigma\rangle\langle s\sigma|W|t\tau\rangle\langle t\tau|W|\psi_{\ell\lambda}^{(0)}\rangle}{(\omega_{\ell}^{(0)} - \omega_{s}^{(0)})(\omega_{\ell}^{(0)} - \omega_{t}^{(0)})} - \omega_{\ell\lambda}^{(1)} \mathbf{S}'_{s\sigma} \frac{|s\sigma\rangle\langle s\sigma|W|\psi_{\ell\lambda}^{(0)}\rangle}{(\omega_{\ell}^{(0)} - \omega_{s}^{(0)})^2}. \tag{8c}$$

In general, the operator W in the above equations represents the whole perturbation Hamiltonian, $\mathscr{H}_1 \equiv \mathscr{H} - \mathscr{H}_0$. However, in this formulation, we explicitly distinguish between the operators $\mathscr{H}_1$ and W, assuming the latter to be the entire contribution in $\mathscr{H}_1$ that is responsible for configuration mixing [i.e., that has nonvanishing matrix elements in Eqs. (8b) and (8c)]. The prime on the summations in Eqs. (8b) and (8c) means that the subspace $\mathscr{E}_{\ell}$ must be excluded from the sum-

mation range. In fact, as a consequence of our choice for the normalization of the perturbed eigenvectors, $|\psi_{\ell\lambda}\rangle$, the corrections $|\psi^{(k)}_{\ell\lambda}\rangle$ are all orthogonal to $\mathscr{E}_\ell$, for $k>0$. In Eq. (8c), $\omega^{(1)}_{\ell\lambda}$ is the first-order correction to the energy of the unperturbed level, $\omega^{(0)}_\ell$ [see Eq. (9a) below].

The reason for maintaining the notation $|\psi_{\ell\lambda}\rangle$ even when the eigenvectors have components outside $\mathscr{E}_\ell$ is that the corrections $|\psi^{(k)}_{\ell\lambda}\rangle$ are supposed to affect ''continuously'' the eigenvectors in the zeroth-order approximation, $|\psi^{(0)}_\ell\rangle$, so that $|\psi_{\ell\lambda}\rangle$ tends continuously to $|\psi^{(0)}_{\ell\lambda}\rangle$ (and its norm to unity) if the perturbation to $\mathscr{H}_0$ becomes negligible (i.e., for $\epsilon\to 0$). This assumption, however, becomes problematic when the perturbation radically modifies the character of the energy spectrum, as in the presence of a perturbing electric field, which introduces a completely continuous spectrum in contrast with the discrete+continuous spectrum of $\mathscr{H}_0$. In fact, the so-called bound states of the Stark effect actually are ''resonances'' in a continuous spectrum with a typical lifetime that decreases rapidly as the electric-field strength and/or the principal quantum numbers of the levels involved increase, because of the increased probability of tunneling through the ionization potential barrier (see Refs. 5–7). Therefore, in our treatment we will limit ourselves to considering electric-field intensities below the critical value given by Eq. (1), so that the resonant states of the Stark effect can be safely considered as bound states of the Hamiltonian $\mathscr{H}$.

Analogous to Eqs. (8), for the eigenvalue $\omega_{\ell\lambda}$ we can write

$$\omega_{\ell\lambda}=\omega^{(0)}_\ell+\epsilon\omega^{(1)}_{\ell\lambda}+\epsilon^2\omega^{(2)}_{\ell\lambda}, \tag{9a}$$

$$\omega^{(2)}_{\ell\lambda}=\sum_{s\sigma}{}' \frac{\langle\omega^{(0)}_{\ell\lambda}|W|s\sigma\rangle\langle s\sigma|W|\psi^{(0)}_{\ell\lambda}\rangle}{\omega^{(0)}_\ell-\omega^{(0)}_s}. \tag{9b}$$

As is well known, the solution of the perturbation problem to first order yields the eigenvalues $\omega^{(0)}_\ell+\epsilon\omega^{(1)}_{\ell\lambda}$ by diagonalization of the operator $\mathscr{P}_\ell\mathscr{H}\mathscr{P}_\ell$ (with $\mathscr{H}=\mathscr{H}_0+\epsilon\mathscr{H}_1$), along with the zeroth-order eigenvectors, $|\psi^{(0)}_{\ell\lambda}\rangle$. Analogously, the perturbation solution of our problem to second order would yield second-order eigenvalues, along with first-order eigenvectors. The second-order correction to the eigenvectors, Eq. (8c), is also needed if we want to account for configuration mixing in a consistent way, since we must keep at least every second-order contribution to the operators $\mathscr{O}^{(k)}_\ell$.

On the other hand, a rather important consistency problem is originated by our choice for the normalization of the perturbed eigenvectors, $|\psi_{\ell\lambda}\rangle$. In fact, whereas the eigenvectors in the expression of $S^{(j)}_{nm}(q,q')$, Eq. (5), or in the expression of the operator $\mathscr{O}^{(k)}_\ell$, Eq. (6), are supposed to be normalized to unity, the norm of the perturbed eigenvectors of Eqs. (8) is instead, to second order,

$$\langle\psi_{\ell\lambda}|\psi_{\ell\lambda}\rangle=\langle\psi^{(0)}_{\ell\lambda}|\psi^{(0)}_{\ell\lambda}\rangle+\epsilon^2\langle\psi^{(1)}_{\ell\lambda}|\psi^{(1)}_{\ell\lambda}\rangle=1+\epsilon^2\sum_{s\sigma}{}' \frac{\langle\psi^{(0)}_{\ell\lambda}|W|s\sigma\rangle\langle s\sigma|W|\psi^{(0)}_{\ell\lambda}\rangle}{(\omega^{(0)}_\ell-\omega^{(0)}_s)^2}.$$

If we then use the perturbed eigenvectors in Eq. (5) [or Eq. (6)] without correcting for their nonunitary norm, we introduce an error in the expressions of $S^{(j)}_{nm}(q,q')$ that is at least of the same order of perturbation as the corrections that we need to consider for a consistent treatment of configuration mixing. Further in this paper we will be able to show that the second-order corrections introduced by the nonunitary norm of the perturbed eigenvectors either are compensated by the normalization of Eq. (1) or can otherwise be safely neglected in our approximation. For this reason, in the following we will simply ignore these corrections.

III. CALCULATION OF THE FIRST-ORDER MOMENTS IN THE PRESENCE OF CONFIGURATION MIXING

In Sec. II we considered a second-order approximation of the eigenvalues $\omega_{\ell\lambda}$ and eigenvectors $|\psi_{\ell\lambda}\rangle$ that enter the expression of the operators $\mathscr{O}^{(k)}_\ell$ defined in Eq. (6). Through Eqs. (4) and (7), we are then in a position to determine a second-order approximation of the Nth-order moments of polarized hydrogen lines, where configuration mixing is accounted for in a natural way (at least to the lowest order of approximation). We should also note that, because of Eqs. (8) and

(9), the only unknowns in this second-order approximation of the Nth-order moments of hydrogen lines are $\omega^{(1)}_{\ell\lambda}$ and $|\psi^{(0)}_{\ell\lambda}\rangle$. The key point in our algebraic derivation of the first-order moments of hydrogen lines in the presence of configuration mixing is therefore the elimination from the expression of $S^{(1)}_{nm}(q,q')$ [cf. Eq. (7)] of any reference to $\omega^{(1)}_{\ell\lambda}$ and $|\psi^{(0)}_{\ell\lambda}\rangle$ (which otherwise had to be calculated numerically via diagonalization of the perturbation Hamiltonian, $\mathscr{H}_1$), making use of standard techniques of projection-operator algebra and of the fundamental results of the spectral theorem.

A. Calculation of the operators $\mathscr{O}^{(k)}_{\ell}$ for $k=0,1$

For the evaluation of $S^{(1)}_{nm}(q,q')$ we only need to consider the operators $\mathscr{O}^{(k)}_{\ell}$ for $k=0,1$. In the following we will simply write ω_a instead of $\omega^{(0)}_a$ for the degenerate, zeroth-order eigenvalues (the absence of the degeneracy parameter, α, indicating that ω_a is necessarily degenerate).

From Eqs. (6) and (8a), we then have, for $k=0$,

$$\mathscr{O}^{(0)}_{\ell}=\sum_{\lambda}|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|+\epsilon\sum_{\lambda}\left(|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(1)}_{\ell\lambda}|+|\psi^{(1)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|\right)$$

$$+\epsilon^2\sum_{\lambda}\left(|\psi^{(1)}_{\ell\lambda}\rangle\langle\psi^{(1)}_{\ell\lambda}|+|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(2)}_{\ell\lambda}|+|\psi^{(2)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|\right),$$

where explicitly

$$|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(1)}_{\ell\lambda}|=\sum_{s\sigma}{}'\,\frac{|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|W|s\sigma\rangle\langle s\sigma|}{\omega_\ell-\omega_s}=\left(|\psi^{(1)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|\right)^{\dagger},$$

$$|\psi^{(1)}_{\ell\lambda}\rangle\langle\psi^{(1)}_{\ell\lambda}|=\sum_{s\sigma,t\tau}{}'\,\frac{|s\sigma\rangle\langle s\sigma|W|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|W|t\tau\rangle\langle t\tau|}{(\omega_\ell-\omega_s)(\omega_\ell-\omega_t)},$$

$$|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(2)}_{\ell\lambda}|=\sum_{s\sigma,t\tau}{}'\,\frac{|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|W|t\tau\rangle\langle t\tau|W|s\sigma\rangle\langle s\sigma|}{(\omega_\ell-\omega_s)(\omega_\ell-\omega_t)}$$

$$-\omega^{(1)}_{\ell\lambda}\sum_{s\sigma}{}'\,\frac{|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|W|s\sigma\rangle\langle s\sigma|}{(\omega_\ell-\omega_s)^2}=\left(|\psi^{(2)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|\right)^{\dagger}.$$

Therefore

$$\mathscr{O}^{(0)}_{\ell}=\mathscr{P}_\ell+\epsilon\sum_{s\sigma}{}'\,\frac{\mathscr{P}_\ell W|s\sigma\rangle\langle s\sigma|+|s\sigma\rangle\langle s\sigma|W\mathscr{P}_\ell}{\omega_\ell-\omega_s}$$

$$+\epsilon^2\Bigg[\sum_{s\sigma,t\tau}{}'\,\frac{|s\sigma\rangle\langle s\sigma|W\mathscr{P}_\ell W|t\tau\rangle\langle t\tau|}{(\omega_\ell-\omega_s)(\omega_\ell-\omega_t)}$$

$$+\sum_{s\sigma,t\tau}{}'\,\frac{\mathscr{P}_\ell W|t\tau\rangle\langle t\tau|W|s\sigma\rangle\langle s\sigma|+|s\sigma\rangle\langle s\sigma|W|t\tau\rangle\langle t\tau|W\mathscr{P}_\ell}{(\omega_\ell-\omega_s)(\omega_\ell-\omega_t)}$$

$$-\sum_{s\sigma}{}'\,\frac{\mathscr{P}_\ell\mathscr{H}_1\mathscr{P}_\ell W|s\sigma\rangle\langle s\sigma|+|s\sigma\rangle\langle s\sigma|W\mathscr{P}_\ell\mathscr{H}_1\mathscr{P}_\ell}{(\omega_\ell-\omega_s)^2}\Bigg],\tag{10}$$

where for the last term we considered that, to first order of perturbation,

$$\sum_{\lambda}\omega^{(1)}_{\ell\lambda}|\psi^{(0)}_{\ell\lambda}\rangle\langle\psi^{(0)}_{\ell\lambda}|=\mathscr{P}_\ell\mathscr{H}_1\mathscr{P}_\ell.\tag{11}$$

Analogously, to second order,

$$\begin{aligned}\mathscr{O}_\ell^{(1)} &= \omega_\ell \sum_\lambda |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(0)}| + \epsilon \sum_\lambda [\omega_{\ell\lambda}^{(1)} |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(0)}| + \omega_\ell (|\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(1)}| + |\psi_{\ell\lambda}^{(1)}\rangle\langle\psi_{\ell\lambda}^{(0)}|)] \\ &\quad + \epsilon^2 \sum_\lambda [\omega_{\ell\lambda}^{(2)} |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(0)}| + \omega_{\ell\lambda}^{(1)} (|\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(1)}| + |\psi_{\ell\lambda}^{(1)}\rangle\langle\psi_{\ell\lambda}^{(0)}|) \\ &\quad + \omega_\ell (|\psi_{\ell\lambda}^{(1)}\rangle\langle\psi_{\ell\lambda}^{(1)}| + |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(2)}| + |\psi_{\ell\lambda}^{(2)}\rangle\langle\psi_{\ell\lambda}^{(0)}|)] \\ &\equiv \omega_\ell \mathscr{O}_\ell^{(0)} + \epsilon \sum_\lambda (\omega_{\ell\lambda}^{(1)} + \epsilon\omega_{\ell\lambda}^{(2)}) |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(0)}| + \epsilon^2 \sum_\lambda \omega_{\ell\lambda}^{(1)} (|\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(1)}| + |\psi_{\ell\lambda}^{(1)}\rangle\langle\psi_{\ell\lambda}^{(0)}|).\end{aligned} \tag{12}$$

The second addendum in the last equivalence can be expressed through the projection operator $\mathscr{P}_\ell$, in a fashion similar to Eq. (11), since it represents the spectral resolution of a suitable operator on the subspace $\mathscr{E}_\ell$. In fact, it can be proved (see, e.g., Ref. 9) that the zeroth-order eigenvectors, $|\psi_{\ell\lambda}^{(0)}\rangle$, and the energy corrections to any order r of perturbation, $\Delta_{\ell\lambda}^{(r)}(\epsilon) \equiv \sum_{k=1}^r \epsilon^k \omega_{\ell\lambda}^{(k)}$, are the solution of the eigenvalue problem for some suitable operator on the subspace $\mathscr{E}_\ell$ including terms of order $k=1,\ldots,r$ in the perturbation Hamiltonian, $\mathscr{H}_1$. Equation (11) illustrates this property in the particular case $r=1$. To second order we find instead (considering that, by definition, only W contributes to configuration mixing)

$$\sum_\lambda (\omega_{\ell\lambda}^{(1)} + \epsilon\omega_{\ell\lambda}^{(2)}) |\psi_{\ell\lambda}^{(0)}\rangle\langle\psi_{\ell\lambda}^{(0)}| = \mathscr{P}_\ell \mathscr{H}_1 \mathscr{P}_\ell + \epsilon {\sum_{s\sigma}}' \frac{\mathscr{P}_\ell W |s\sigma\rangle\langle s\sigma| W \mathscr{P}_\ell}{\omega_\ell - \omega_s}. \tag{11'}$$

Of course, the expressions of the zeroth-order eigenvectors which are obtained by solving the eigenvalue problem for the operators of Eqs. (11) and (11′) differ for quantities of the second order in the perturbation W. However, for our purposes, these differences are unessential, as they are only responsible for corrections to the operators $\mathscr{O}_\ell^{(0)}$ and $\mathscr{O}_\ell^{(1)}$ of order higher than second. Equation (12) then becomes, to second order,

$$\begin{aligned}\mathscr{O}_\ell^{(1)} &= \omega_\ell \mathscr{O}_\ell^{(0)} + \epsilon \mathscr{P}_\ell \mathscr{H}_1 \mathscr{P}_\ell + \epsilon^2 \Bigg\{ {\sum_{s\sigma}}' \frac{\mathscr{P}_\ell W |s\sigma\rangle\langle s\sigma| W \mathscr{P}_\ell}{\omega_\ell - \omega_s} \\ &\quad + {\sum_{s\sigma}}' \frac{\mathscr{P}_\ell \mathscr{H}_1 \mathscr{P}_\ell W |s\sigma\rangle\langle s\sigma| + |s\sigma\rangle\langle s\sigma| W \mathscr{P}_\ell \mathscr{H}_1 \mathscr{P}_\ell}{\omega_\ell - \omega_s} \Bigg\},\end{aligned} \tag{12'}$$

where we used both Eqs. (11) and (11′).

We must notice how any explicit reference to the unknown quantities $\omega_{\ell\lambda}^{(1)}$ and $|\psi_{\ell\lambda}^{(0)}\rangle$ has been removed from Eqs. (10) and (12′).

B. Calculation of the contribution of the quadratic Stark effect to $\sigma_{nm}^{(1)}(i,\mathbf{k})$

As we stated in the beginning of Sec. III, for the algebraic calculation of the first-order moments of polarized hydrogen lines we need to evaluate the quantity $S_{nm}^{(1)}(q,q')$, which is expressed through the operators $\mathscr{O}_\ell^{(0)}$ and $\mathscr{O}_\ell^{(1)}$ [cf. Eq. (7)]. However, it proves useful to rather consider the operator $\tilde{\mathscr{O}}_\ell^{(1)} \equiv \mathscr{O}_\ell^{(1)} - \omega_\ell \mathscr{O}_\ell^{(0)}$, and correspondingly introduce the new quantity

$$\begin{aligned}\tilde{S}_{nm}^{(1)}(q,q') &= {\sum_{\substack{a\alpha,a'\alpha' \\ b\beta,b'\beta'}}}' \langle a\alpha|\tilde{\mathscr{O}}_n^{(1)}|a'\alpha'\rangle\langle b\beta|\mathscr{O}_m^{(0)}|b'\beta'\rangle\langle a'\alpha'|r_q|b\beta\rangle\langle a\alpha|r_{q'}|b'\beta'\rangle \\ &\quad - {\sum_{\substack{a\alpha,a'\alpha' \\ b\beta,b'\beta'}}}' \langle a\alpha|\mathscr{O}_n^{(0)}|a'\alpha'\rangle\langle b\beta|\tilde{\mathscr{O}}_m^{(1)}|b'\beta'\rangle\langle a'\alpha'|r_q|b\beta\rangle\langle a\alpha|r_{q'}|b'\beta'\rangle,\end{aligned} \tag{13}$$

where we took into account that $\langle a\alpha|r_q|b\beta\rangle=\langle b\beta|r_q^\dagger|a\alpha\rangle$, since $\langle a\alpha|r_q|b\beta\rangle$ is a real quantity because of our choice of the representation basis [cf. Ref. 10, Eq. (23)]. The quantity $\tilde{S}^{(1)}_{nm}(q,q')$ is then related, through Eq. (4), to the quantity

$$\tilde{\sigma}^{(1)}_{nm}(i,\mathbf{k})\equiv\sum_{\nu\mu} s_{\nu\mu}(i,\mathbf{k})(\omega_{\nu\mu}-\omega_{nm}), \tag{14}$$

where $\omega_{nm}\equiv\omega_n-\omega_m$. Further in this section we will find it particularly convenient to express the first-order moments, $\langle\omega_{\mathbf{k}}(i)\rangle$, in terms of $\tilde{\sigma}^{(1)}_{nm}(i,\mathbf{k})$ rather than $\sigma^{(1)}_{nm}(i,\mathbf{k})$.

We then have, considering Eq. (10) and letting $\epsilon\to 1$,

$$\begin{aligned}\langle k\kappa|\mathscr{O}^{(0)}_{\ell}|k'\kappa'\rangle=&\,\delta_{k\ell}\delta_{k'\ell}\delta_{\kappa\kappa'}+\delta_{k\ell}(1-\delta_{k'\ell})\frac{\langle\ell\kappa|W|k'\kappa'\rangle}{\omega_\ell-\omega_{k'}}+\delta_{k'\ell}(1-\delta_{k\ell})\frac{\langle k\kappa|W|\ell\kappa'\rangle}{\omega_\ell-\omega_k}\\&+(1-\delta_{k\ell})(1-\delta_{k'\ell})\sum_\lambda\frac{\langle k\kappa|W|\ell\lambda\rangle\langle\ell\lambda|W|k'\kappa'\rangle}{(\omega_\ell-\omega_k)(\omega_\ell-\omega_{k'})}\\&+\delta_{k\ell}(1-\delta_{k'\ell})\sum_{t\tau}{}'\frac{\langle\ell\kappa|W|t\tau\rangle\langle t\tau|W|k'\kappa'\rangle}{(\omega_\ell-\omega_{k'})(\omega_\ell-\omega_t)}\\&+\delta_{k'\ell}(1-\delta_{k\ell})\sum_{t\tau}{}'\frac{\langle k\kappa|W|t\tau\rangle\langle t\tau|W|\ell\kappa'\rangle}{(\omega_\ell-\omega_k)(\omega_\ell-\omega_t)}\\&-\delta_{k\ell}(1-\delta_{k'\ell})\sum_\lambda\frac{\langle\ell\kappa|\mathscr{H}_1|\ell\lambda\rangle\langle\ell\lambda|W|k'\kappa'\rangle}{(\omega_\ell-\omega_{k'})^2}\\&-\delta_{k'\ell}(1-\delta_{k\ell})\sum_\lambda\frac{\langle k\kappa|W|\ell\lambda\rangle\langle\ell\lambda|\mathscr{H}_1|\ell\kappa'\rangle}{(\omega_\ell-\omega_k)^2}.\end{aligned} \tag{15}$$

Analogously, considering Eq. (12′) and letting $\epsilon\to 1$,

$$\begin{aligned}\langle k\kappa|\tilde{\mathscr{O}}^{(1)}_{\ell}|k'\kappa'\rangle=&\,\delta_{k\ell}\delta_{k'\ell}\langle\ell\kappa|\mathscr{H}_1|\ell\kappa'\rangle+\delta_{k\ell}\delta_{k'\ell}\sum_{s\sigma}{}'\frac{\langle\ell\kappa|W|s\sigma\rangle\langle s\sigma|W|\ell\kappa'\rangle}{\omega_\ell-\omega_s}\\&+\delta_{k\ell}(1-\delta_{k'\ell})\sum_\lambda\frac{\langle\ell\kappa|\mathscr{H}_1|\ell\lambda\rangle\langle\ell\lambda|W|k'\kappa'\rangle}{\omega_\ell-\omega_{k'}}\\&+\delta_{k'\ell}(1-\delta_{k\ell})\sum_\lambda\frac{\langle k\kappa|W|\ell\lambda\rangle\langle\ell\lambda|\mathscr{H}_1|\ell\kappa'\rangle}{\omega_\ell-\omega_k}.\end{aligned} \tag{16}$$

As we anticipated in the previous section, the operator W in general coincides with the perturbation Hamiltonian, $\mathscr{H}_1$. From now on, we will consider the following explicit form of $\mathscr{H}_1$ [cf. Ref. 10, Eq. (2′)]:

$$\mathscr{H}_1=\mathscr{H}_{\rm fs}+\mathscr{H}_{\rm hfs}+\mathscr{H}_E+\mathscr{H}_B^{(1)}+\mathscr{H}_B^{(2)},$$

where $\mathscr{H}_{\rm fs}$ is the fine-structure Hamiltonian [comprehensive of the relevant radiative corrections; see Ref. 10, Eq. (6)], $\mathscr{H}_{\rm hfs}$ is the hyperfine-structure Hamiltonian, $\mathscr{H}_E$ is the electric Hamiltonian (associated with the external, uniform electric field), and $\mathscr{H}_B$ is the magnetic Hamiltonian, responsible for both the linear Zeeman effect (electronic and nuclear; $\mathscr{H}^{(1)}_B$) and the quadratic Zeeman effect (diamagnetic term; $\mathscr{H}^{(2)}_B$). The expressions for all these operators have been given in Ref. 10; however, we will not need them explicitly in the present paper.

If the external magnetic field is not too strong, relative to the principal quantum numbers involved ($B\ll 10^{10}n^{-4}$ G), so that $\|\mathscr{H}^{(2)}_B\|\ll\|\mathscr{H}^{(1)}_B\|$, only the electric Hamiltonian, $\mathscr{H}_E$, contributes sensibly to configuration mixing. In fact, the configuration mixing due to the fine structure terms, $\mathscr{H}_{\rm fs}$

and $\mathscr{H}_{\rm hfs}$, is always negligible, whereas $\mathscr{H}_B^{(1)}$ does not mix different configurations (as can easily be proved through the Wigner–Eckart theorem). So we can just replace W with $\mathscr{H}_E$ in Eqs. (15) and (16). This will enable us to account for the quadratic Stark effect in the algebraic formalism of the first-order moments of polarized hydrogen lines.

To first order of perturbation, the only contribution to $\tilde{S}^{(1)}_{nm}(q,q')$ comes from the matrix element $\langle \ell\kappa|\mathscr{H}_1|\ell\kappa'\rangle$ in Eq. (16). To second order, instead, the only nonvanishing contributions to $\tilde{S}^{(1)}_{nm}(q,q')$ come from products of matrix elements of $\mathscr{H}_E$ only. In fact, all terms in $\mathscr{H}_1$ other than $\mathscr{H}_E$ determine the selection rules $|\Delta l|=0$ and/or $|\Delta l|=2$ for the azimuthal quantum number, l, whereas $\mathscr{H}_E$ yields $|\Delta l|=1$. Then we notice the presence into Eq. (13) of $3j$-symbol products of the form

$$\begin{pmatrix} l & 1 & l' \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} l' & 1 & l'' \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} l'' & 1 & l''' \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} l''' & C & l \\ 0 & 0 & 0 \end{pmatrix},$$

which are nonvanishing only for odd values of $C=|\Delta l|$ because of the electric–dipole selection rule (Laporte's rule), $|\Delta l|=1$.

After substitution of Eqs. (15) and (16) into Eq. (13), $\tilde{S}^{(1)}_{nm}(q,q')$ can be written as a finite sum of contributions originated by all possible products between the terms in Eqs. (15) and (16). Formally, there are 32 such contributions for each of the two summations in Eq. (13), many of which correspond to corrections of order higher than second and so are negligible in our approximation. Only six non-negligible terms remain for each of the summations in Eq. (13), so we can write

$$\tilde{S}^{(1)}_{nm}(q,q')=\sum_{r=1}^{6}(K_r^+-K_r^-). \tag{17}$$

The coefficients $K_1^{\pm}$ of first order in the perturbation Hamiltonian, $\mathscr{H}_1$, are easily found to be

$$K_1^+=\sum_{\alpha\alpha'\beta}\langle n\alpha|.\mathscr{H}_1|n\alpha'\rangle\langle n\alpha'|r_q|m\beta\rangle\langle n\alpha|r'_q|m\beta\rangle,$$

$$K_1^-=\sum_{\alpha\beta\beta'}\langle m\beta|.\mathscr{H}_1|m\beta'\rangle\langle n\alpha|r_q|m\beta\rangle\langle n\alpha|r'_q|m\beta'\rangle.$$

If we compare them with Eq. (9) in Ref. 1, we see that the contribution $K_1^+-K_1^-$ in $\tilde{S}^{(1)}_{nm}(q,q')$ corresponds to the contribution of the perturbation $\mathscr{H}_1$ to the first-order moments of polarized hydrogen lines in the absence of configuration mixing. The remaining five terms of the form $(K_r^+-K_r^-)$ in Eq. (17) then yield second-order corrections, due to electric-field induced configuration mixing, to those first-order moments. The corresponding coefficients $K_r^{\pm}$ are the following:

$$K_2^+=\sum_{b'}{}'(\omega_m-\omega_{b'})^{-1}\sum_{\alpha\alpha'\beta\beta'}\langle n\alpha|.\mathscr{H}_E|n\alpha'\rangle\langle m\beta|.\mathscr{H}_E|b'\beta'\rangle\langle n\alpha'|r_q|m\beta\rangle\langle n\alpha|r_{q'}|b'\beta'\rangle, \tag{18a}$$

$$K_2^-=\sum_{a'}{}'(\omega_n-\omega_{a'})^{-1}\sum_{\alpha\alpha'\beta\beta'}\langle n\alpha|.\mathscr{H}_E|a'\alpha'\rangle\langle m\beta|.\mathscr{H}_E|m\beta'\rangle\langle a'\alpha'|r_q|m\beta\rangle\langle n\alpha|r_{q'}|m\beta'\rangle, \tag{19a}$$

$$K_3^+=\sum_{b}{}'(\omega_m-\omega_b)^{-1}\sum_{\alpha\alpha'\beta\beta'}\langle n\alpha|.\mathscr{H}_E|n\alpha'\rangle\langle b\beta|.\mathscr{H}_E|m\beta'\rangle\langle n\alpha'|r_q|b\beta\rangle\langle n\alpha|r_{q'}|m\beta'\rangle, \tag{18b}$$

$$K_3^- = \sum_a{}' (\omega_n - \omega_a)^{-1} \sum_{\alpha\alpha'\beta\beta'} \langle a\alpha|\mathscr{H}_E|n\alpha'\rangle\langle m\beta|\mathscr{H}_E|m\beta'\rangle\langle n\alpha'|r_q|m\beta\rangle\langle a\alpha|r_{q'}|m\beta'\rangle, \tag{19b}$$

$$K_4^+ = \sum_s{}' (\omega_n - \omega_s)^{-1} \sum_{\alpha\alpha'\sigma\beta} \langle n\alpha|\mathscr{H}_E|s\sigma\rangle\langle s\sigma|\mathscr{H}_E|n\alpha'\rangle\langle n\alpha'|r_q|m\beta\rangle\langle n\alpha|r_{q'}|m\beta\rangle, \tag{18c}$$

$$K_4^- = \sum_s{}' (\omega_m - \omega_s)^{-1} \sum_{\alpha\sigma\beta\beta'} \langle m\beta|\mathscr{H}_E|s\sigma\rangle\langle s\sigma|\mathscr{H}_E|m\beta'\rangle\langle n\alpha|r_q|m\beta\rangle\langle n\alpha|r_{q'}|m\beta'\rangle, \tag{19c}$$

$$K_5^+ = \sum_{a'}{}' (\omega_n - \omega_{a'})^{-1} \sum_{\alpha\alpha'\beta\lambda} \langle n\alpha|\mathscr{H}_E|n\lambda\rangle\langle n\lambda|\mathscr{H}_E|a'\alpha'\rangle\langle a'\alpha'|r_q|m\beta\rangle\langle n\alpha|r_{q'}|m\beta\rangle, \tag{18d}$$

$$K_5^- = \sum_{b'}{}' (\omega_m - \omega_{b'})^{-1} \sum_{\alpha\beta\beta'\lambda} \langle m\beta|\mathscr{H}_E|m\lambda\rangle\langle m\lambda|\mathscr{H}_E|b'\beta'\rangle\langle n\alpha|r_q|m\beta\rangle\langle n\alpha|r_{q'}|b'\beta'\rangle, \tag{19d}$$

$$K_6^+ = \sum_a{}' (\omega_n - \omega_a)^{-1} \sum_{\alpha\alpha'\beta\lambda} \langle a\alpha|\mathscr{H}_E|n\lambda\rangle\langle n\lambda|\mathscr{H}_E|n\alpha'\rangle\langle n\alpha'|r_q|m\beta\rangle\langle a\alpha|r_{q'}|m\beta\rangle, \tag{18e}$$

$$K_6^- = \sum_b{}' (\omega_m - \omega_b)^{-1} \sum_{\alpha\beta\beta'\lambda} \langle b\beta|\mathscr{H}_E|m\lambda\rangle\langle m\lambda|\mathscr{H}_E|m\beta'\rangle\langle n\alpha|r_q|b\beta\rangle\langle n\alpha|r_{q'}|m\beta'\rangle. \tag{19e}$$

It is worth noting that all of the summations of matrix element products in Eqs. (18) and (19) have already been evaluated in Ref. 2. In fact, if we compare Eqs. (6) and (14) in Ref. 2 [see also Eq. (13) in that paper], we realize that sums of the form

$$\begin{aligned} S^{(1)} &= \sum_{\alpha\alpha'\alpha''\alpha'''} \langle a\alpha|\mathscr{H}_E|a'\alpha'\rangle\langle a'\alpha'|\mathscr{H}_E|a''\alpha''\rangle\langle a''\alpha''|r_q|a'''\alpha'''\rangle\langle a\alpha|r_{q'}|a'''\alpha'''\rangle \\ &\equiv \sum_{\alpha\alpha'\alpha''\alpha'''} \langle a\alpha|\mathscr{H}_E|a'\alpha'\rangle\langle a'\alpha'|\mathscr{H}_E|a''\alpha''\rangle\langle a''\alpha''|r_q|a'''\alpha'''\rangle\langle a'''\alpha'''|r_{q'}^\dagger|a\alpha\rangle, \end{aligned}$$

or

$$\hat{S}^{(1)} = S^{(1)}\{r_q \leftrightarrow r_{q'}^\dagger\},$$

yield the following ''electric'' contribution to the quantity $\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})$,

$$[\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})]_E^{(1)} = 4e_0^2 a_0^2 E^2 \sum_{K=0,2} T_K(i) A_K^{(1)} \tag{20a}$$

(a_0 is the Bohr radius), where, using the shorthand notation $\Pi_{\alpha\beta\ldots} = [(2\alpha+1)(2\beta+1)\cdots]^{1/2}$, we have [cf. Ref. 2, Eq. (14)]

$$\begin{aligned} A_K^{(1)} = a_0^{-2} \sum_{ll'l''l'''} \Pi^2_{ll'l''l'''} &\langle al|r|a'l'\rangle\langle a'l'|r|a''l''\rangle\langle a''l''|r|a'''l'''\rangle\langle a'''l'''|r|al\rangle \begin{pmatrix} l & 1 & l' \\ 0 & 0 & 0 \end{pmatrix} \\ &\times \begin{pmatrix} l' & 1 & l'' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l'' & 1 & l''' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l''' & 1 & l \\ 0 & 0 & 0 \end{pmatrix} \begin{Bmatrix} l & l'' & K \\ 1 & 1 & l' \end{Bmatrix} \begin{Bmatrix} l & l'' & K \\ 1 & 1 & l''' \end{Bmatrix}. \end{aligned} \tag{20b}$$

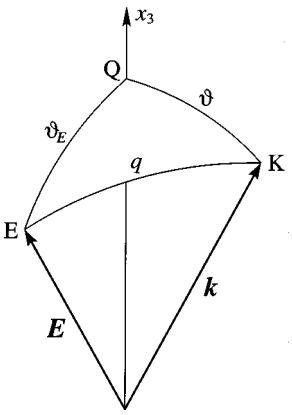

FIG. 2. Geometry of the electric field and the LOS relative to the quantization axis. The angle $\hat{Q}$ in the spherical triangle $\Delta(EQK)$ is equivalent to $(\varphi-\varphi_E)$, where φ and φ_E are respectively the azimuthal angles of the propagation vector, **k**, and of the electric field, **E**, in the frame of reference of the quantization axis.

Analogously, sums of the form

$$S^{(2)}=\sum_{\alpha\alpha'\alpha''\alpha'''}\langle a\alpha|\mathscr{H}_E|a'\alpha'\rangle\langle a''\alpha''|\mathscr{H}_E|a'''\alpha'''\rangle\langle a'\alpha'|r_q|a''\alpha''\rangle\langle a\alpha|r_{q'}|a'''\alpha'''\rangle$$

$$\equiv\sum_{\alpha\alpha'\alpha''\alpha'''}\langle a\alpha|\mathscr{H}_E|a'\alpha'\rangle\langle a'\alpha'|r_q|a''\alpha''\rangle\langle a''\alpha''|\mathscr{H}_E|a'''\alpha'''\rangle\langle a'''\alpha'''|r^{\dagger}_{q'}|a\alpha\rangle,$$

or

$$\hat{S}^{(2)}=S^{(2)}\{r_q\leftrightarrow r^{\dagger}_{q'}\},$$

yield

$$[\tilde{\sigma}^{(1)}_{nm}(i,\mathbf{k})]^{(2)}_E=4e_0^2a_0^2E^2\sum_{K=0,2}T_K(i)A^{(2)}_K, \tag{21a}$$

where [cf. Ref. 2, Eq. (14)]

$$A^{(2)}_K=a_0^{-2}\sum_{ll'l''l'''}\Pi^2_{ll'l''l'''}\langle al|r|a'l'\rangle\langle a'l'|r|a''l''\rangle\langle a''l''|r|a'''l'''\rangle\langle a'''l'''|r|al\rangle\begin{pmatrix}l & 1 & l'\\ 0 & 0 & 0\end{pmatrix}$$

$$\times\begin{pmatrix}l' & 1 & l''\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}l'' & 1 & l'''\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}l''' & 1 & l\\ 0 & 0 & 0\end{pmatrix}\begin{Bmatrix}1 & 1 & K\\ l' & l & 1\\ l'' & l''' & 1\end{Bmatrix}. \tag{21b}$$

The quantities $T_K(i)$ in Eqs. (20a) and (21a), which depend on the field geometry, for $K=0,2$ are given by

$$T_0(i)=\tfrac{1}{3}\delta_{i0}, \tag{22}$$

and

$$T_2(0)=\tfrac{1}{6}(1-3\cos^2 q), \tag{23a}$$

$$T_2(1)=\tfrac{1}{2}\{\sin^2\vartheta_E[\cos 2(\varphi-\varphi_E)-1]-\sin^2 q\}\cos 2\gamma+\sin\vartheta_E\sin(\varphi-\varphi_E)\sin q\cos\hat{K}\sin 2\gamma, \tag{23b}$$

$$T_2(2)=T_2(1)\begin{Bmatrix}\cos 2\gamma\rightarrow-\sin 2\gamma\\ \sin 2\gamma\rightarrow\cos 2\gamma\end{Bmatrix}, \tag{23c}$$

$$T_2(3)=0, \tag{23d}$$

where, for the fundamental formulas of spherical triangles (see Fig. 2),

$$\begin{cases} \cos q = \cos\vartheta\,\cos\vartheta_E + \sin\vartheta\,\sin\vartheta_E\,\cos(\varphi-\varphi_E), \\ \sin q\,\cos\hat{K} = \sin\vartheta\,\cos\vartheta_E - \cos\vartheta\,\sin\vartheta_E\,\cos(\varphi-\varphi_E). \end{cases}$$

In the particular case in which the LOS coincides with the quantization axis, so that $\vartheta=\varphi=0$ and $q=\vartheta_E$, Eqs. (23) reduces to Eqs. (18) in Ref. 2.

To recover Eqs. (20) and (21) from comparison of Eqs. (6) and (14) in Ref. 2, we also had to consider the following result, valid in the absence of configuration mixing [see Ref. 1, Eq. (A3)],

$$\sum_{\nu\mu} s_{\nu\mu}(i,\mathbf{k}) = \delta_{i0}\,\frac{4}{3}\sum_{ll'}\Pi^2_{ll'}\langle nl|r|ml'\rangle^2\begin{pmatrix} l & 1 & l' \\ 0 & 0 & 0\end{pmatrix}^2 \equiv \delta_{i0}\mathscr{N}_{nm}. \tag{24}$$

[The factor 4 in Eqs. (20a), (21a), and (24) is related to the electronic- and nuclear-spin quantum numbers, $s=i=1/2$, and comes from $\Pi^2_{si}=(2s+1)(2i+1)=4$.]

C. Estimate of the contribution of the quadratic Stark effect to $\sigma^{(0)}_{nm}(i,\mathbf{k})$

To completely define the first-order moments of polarized hydrogen lines, according to Eq. (2) we still have to evaluate the integrated profiles, $\Sigma_{\nu\mu}\, s_{\nu\mu}(i,\mathbf{k})$, in the presence of configuration mixing, which in general will differ from the result stated by Eq. (24).

According to Eq. (7), we must evaluate the quantity

$$S^{(0)}_{nm}(q,q') = \sum_{\substack{a\alpha,a'\alpha' \\ b\beta,b'\beta'}} \langle a\alpha|\mathscr{O}^{(0)}_n|a'\alpha'\rangle\langle b\beta|\mathscr{O}^{(0)}_m|b'\beta'\rangle\langle a'\alpha'|r_q|b\beta\rangle\langle a\alpha|r_{q'}|b'\beta'\rangle.$$

From Eq. (15), and applying the same arguments as in the previous calculation of $\tilde{S}^{(1)}_{nm}(q,q')$, we see that there are 15 nonvanishing contributions to $S^{(0)}_{nm}(q,q')$, originating from products between terms of Eq. (15), so we can write

$$S^{(0)}_{nm}(q,q') = \sum_{r=0}^{14} L_r.$$

For $r=0$ we have trivially [cf. Ref. 1, Appendix A, and Eq. (24) above]

$$L_0 = \sum_{\alpha\beta}\langle n\alpha|r_q|m\beta\rangle\langle n\alpha|r_{q'}|m\beta\rangle \equiv \delta_{qq'}\,\frac{4}{3}\sum_{ll'}\Pi^2_{ll'}\langle nl|r|ml'\rangle^2\begin{pmatrix} l & 1 & l' \\ 0 & 0 & 0\end{pmatrix}^2 \equiv \delta_{qq'}\mathscr{N}_{nm},$$

which is the zeroth-order term that would give Eq. (24) if configuration mixing were absent. The remaining 14 coefficients L_r are all of second order in $W\equiv\mathscr{H}_E$, and they also can be expressed through the set of coefficients $A^{(i)}_K(n,m)$ $(i=1,2;\ K=0,2)$ given in Eqs. (20b) and (21b).

We then see that, because of configuration mixing, to second order of perturbation we no longer have $\Sigma_{\nu\mu}\, s_{\nu\mu}(i,\mathbf{k})=0$ for $i\neq 0$. This is a fact that needs to be taken into account in the evaluation of the numerator of Eq. (2). The denominator of Eq. (2), on the contrary, can be safely approximated by $\mathscr{N}_{nm}$ [cf. Eq. (24)], just like in the absence of configuration mixing. To see this in the case of the first-order moments we rewrite

$$\langle\omega_{\mathbf{k}}(i)\rangle = \frac{\Sigma_{\nu\mu}s_{\nu\mu}(i,\mathbf{k})(\omega_{\nu\mu}-\omega_{nm})}{\Sigma_{\nu\mu}s_{\nu\mu}(0,\mathbf{k})} + (\omega_{nm}-\overline{\omega})\,\frac{\Sigma_{\nu\mu}s_{\nu\mu}(i,\mathbf{k})}{\Sigma_{\nu\mu}s_{\nu\mu}(0,\mathbf{k})}.$$

The first term on the right-hand side (rhs) is a quantity of first order in the perturbation $\mathscr{H}_1$ (plus second-order corrections due to configuration mixing), so the second-order corrections in $\Sigma_{\nu\mu}\, s_{\nu\mu}(0,\mathbf{k})$ can be neglected, as they would yield at least third (or higher) order corrections to $\langle\omega_{\mathbf{k}}(i)\rangle$. On the other hand, the second term on the rhs is exactly equal to $(\omega_{nm}-\overline{\omega})$ if $i=0$, whereas it is a quantity of second order in the perturbation $W\equiv\mathscr{H}_E$ if $i\neq 0$, so the second-order corrections to the denominator can be ignored in both cases. Finally, we observe that the difference $(\omega_{nm}-\overline{\omega})$ is essentially determined by the fine structure of the hydrogen atom, so it represents a quantity of the first order in the perturbation $\mathscr{H}_{\rm fs}$. Therefore, the second term on the rhs of

the above equation, for $i\neq 0$, represents a correction to $\langle\omega_{\mathbf{k}}(i)\rangle$ that actually is comparable with terms of order higher than second in the perturbation $\mathscr{H}_1$. For all these reasons, to second order we can safely approximate [cf. Eq. (14)]

$$\langle\omega_{\mathbf{k}}(i)\rangle\approx\mathscr{N}_{nm}^{-1}\sum_{\nu\mu}s_{\nu\mu}(i,\mathbf{k})(\omega_{\nu\mu}-\omega_{nm})+\delta_{i0}(\omega_{nm}-\overline{\omega})=\mathscr{N}_{nm}^{-1}\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})+\delta_{i0}(\omega_{nm}-\overline{\omega}), \tag{25}$$

which manifestly supports our earlier choice of evaluating $\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})$ rather than $\sigma_{nm}^{(1)}(i,\mathbf{k})$.

We note that the above arguments also justify our decision of neglecting the second-order corrections to the first-order moments due to the nonunitary norm of the perturbed eigenvectors (see Sec. II), as these corrections are unessential, to our order of approximation, in the evaluation of the summation in Eq. (25).

D. Algebraic formulation of the first-order moments

Because of the approximation of the denominator $\Sigma_{\nu\mu}\,s_{\nu\mu}(0,\mathbf{k})$ by Eq. (24), the corrections to the first-order moments due to electric-field induced configuration mixing are simply added to the expressions of the first-order moments in the absence of configuration mixing, so we can rewrite Eq. (25) in the form

$$\begin{aligned}\langle\omega_{\mathbf{k}}(i)\rangle&\approx\{\delta_{i0}(\omega_{nm}-\overline{\omega})+\mathscr{N}_{nm}^{-1}[\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})]_{\text{no cm}}\}+\mathscr{N}_{nm}^{-1}[\tilde{\sigma}_{nm}^{(1)}(i,\mathbf{k})]_{\text{cm}}\\&\equiv\langle\omega_{\mathbf{k}}(i)\rangle_{\text{no cm}}+\langle\omega_{\mathbf{k}}(i)\rangle_{E},\end{aligned} \tag{25$'$}$$

where $\langle\omega_{\mathbf{k}}(i)\rangle_{\text{no cm}}$ represents the first-order moments in the absence of configuration mixing [inclusive of the first-order term $\delta_{i0}(\omega_{nm}-\overline{\omega})$], which were evaluated in Ref. 1, whereas $\langle\omega_{\mathbf{k}}(i)\rangle_E$ represents the correction to those first-order moments due to the configuration mixing induced by the presence of the electric field. This correction, as suggested by Eqs. (20a) and (21a), can be written

$$\langle\omega_{\mathbf{k}}(i)\rangle_E=4e_0^2a_0^2E^2\sum_{K=0,2}T_K(i)\tilde{A}_K(n,m), \tag{26}$$

with a suitable choice of the coefficients $\tilde{A}_K(n,m)$.

To provide a compact formulation of these coefficients, it is useful to introduce the following symbols,

$$\begin{aligned}R(al,a'l',a''l'',a'''l''')\equiv a_0^{-2}\Pi_{ll'l''l'''}^2\langle al|r|a'l'\rangle\langle a'l'|r|a''l''\rangle\langle a''l''|r|a'''l'''\rangle\langle a'''l'''|r|al\rangle\\\times\begin{pmatrix}l&1&l'\\0&0&0\end{pmatrix}\begin{pmatrix}l'&1&l''\\0&0&0\end{pmatrix}\begin{pmatrix}l''&1&l'''\\0&0&0\end{pmatrix}\begin{pmatrix}l'''&1&l\\0&0&0\end{pmatrix},\end{aligned} \tag{27}$$

$$f_K^{(1)}(ll'l''l''')\equiv\begin{Bmatrix}l&l''&K\\1&1&l'\end{Bmatrix}\begin{Bmatrix}l&l''&K\\1&1&l'''\end{Bmatrix},\quad f_K^{(2)}(ll'l''l''')\equiv\begin{Bmatrix}1&1&K\\l'&l&1\\l''&l'''&1\end{Bmatrix}, \tag{28}$$

and the ''contracted'' coefficients

$$\mathscr{R}_K^{(i)}(aa'a''a''')\equiv\sum_{ll'l''l'''}R(al,a'l',a''l'',a'''l''')f_K^{(i)}(ll'l''l'''),\quad i=1,2,\quad K=0,2. \tag{29}$$

If we consider the symmetry properties of the $6j$ and $9j$ symbols (see, e.g., Ref. 11) we can easily demonstrate the following symmetries of the coefficients $\mathscr{R}_K^{(i)}$ for even values of K,

$$\mathscr{R}_K^{(1)}(aa'a''a''')=\mathscr{R}_K^{(1)}(aa'''a''a')=\mathscr{R}_K^{(1)}(a''a'aa''')=\mathscr{R}_K^{(1)}(a''a'''aa'),$$

$$\mathscr{R}_K^{(2)}(aa'a''a''')=\mathscr{R}_K^{(2)}(a'aa'''a'')=\mathscr{R}_K^{(2)}(a''a'''aa')=\mathscr{R}_K^{(2)}(a'''a''a'a).$$

The coefficients $\tilde{A}_K(n,m)$ of Eq. (26) then assume a particularly symmetric form. In fact, from Eqs. (18) and (19), after some renaming of summation indexes and collecting of similar terms, we find

$$\mathscr{N}_{nm}\tilde{A}_K(n,m)=\sum_a{}'(\omega_m-\omega_a)^{-1}[2.\mathscr{R}_K^{(2)}(nnma)-2.\mathscr{R}_K^{(1)}(anmm)-\mathscr{R}_K^{(1)}(mnma)]$$

$$-\sum_a{}'(\omega_n-\omega_a)^{-1}[2.\mathscr{R}_K^{(2)}(mmna)-2.\mathscr{R}_K^{(1)}(amnn)-\mathscr{R}_k^{(1)}(nmna)]. \quad (30)$$

Equations (26) and (30) finally give an algebraic expression for the quadratic Stark effect on the centers of gravity of polarized hydrogen lines.

Whereas the numerical calculation of the discrete part of Eq. (30) [or of Eqs. (18) and (19)] is straightforward and does not present any conceptual difficulty, things are much more complicated when integration over the continuum must be considered. A prescription given by Ruffa,[12] on the use of the integral representation of the hydrogenic wave functions for the continuum (see, e.g., Ref. 13), which reduces the calculation of summations of matrix elements over the continuum to integrations in the complex domain, might probably be the best approach to this problem. However, we will not investigate the continuum problem any further in this paper.

E. Generalization of the first-order moments in the presence of line broadening

A final remark must be made about the generalization of the result expressed by Eq. (30) when typical line-broadening mechanisms (e.g., pressure broadening, thermal-Doppler broadening) are also present. In the beginning of Sec. II, in fact, we explicitly noted that Eq. (2) holds in the absence of any broadening mechanism of the line profile other than the external fields.

This problem has already been dealt with in Ref. 2 (see Sec. 5 in that paper). However those previous results must be somewhat modified to account for the no-longer vanishing (because of configuration mixing) integrated polarization profiles, $\Sigma_{\nu\mu}\, s_{\nu\mu}(i,\mathbf{k})$, for $i\neq 0$. Following the notation of Ref. 2, where ϕ_a and ϕ_b are, respectively, the unbroadened Stokes profile and the line profile resulting from the intervening broadening mechanisms, we indicate with $\mathscr{I}_a$ and $\mathscr{I}_b$ the corresponding integrated profiles, and with $\mathscr{N}_a$ and $\mathscr{N}_b$ the corresponding normalization constants. Operating as in Ref. 2 we then easily find for the first-order moment of the broadened Stokes profile (obtained as the convolution product of ϕ_a and ϕ_b),

$$\langle\omega\rangle=\frac{\mathscr{I}_b}{\mathscr{N}_b}\langle\omega_a\rangle+\frac{\mathscr{I}_a}{\mathscr{N}_a}\langle\omega_b\rangle+\frac{\mathscr{I}_a\mathscr{I}_b}{\mathscr{N}_a\mathscr{N}_b}\,\overline{\omega}. \quad (31)$$

As is customary, the line-broadening profile is normalized to unit integrated area, so that $\mathscr{I}_b=\mathscr{N}_b=1$. For the Stokes profile, instead, we must distinguish between the intensity profile ($i=0$) and the other polarization profiles ($i\neq 0$). For the intensity profile we can choose the standard normalization, $\mathscr{I}_a=\mathscr{N}_a=1$, so we recover Eq. (25a) in Ref. 2. For the linear- and circular-polarization profiles, instead, Eq. (31) becomes, recalling Eq. (24),

$$\langle\omega(i)\rangle=\langle\omega_a(i)\rangle+[\langle\omega_b\rangle+\overline{\omega}]\mathscr{N}_{nm}^{-1}\sum_{\nu\mu}\, s_{\nu\mu}(i,\mathbf{k}), \quad i=1,2,3. \quad (32)$$

The second term then determines a second-order correction to the first-order moments of broadened, polarized hydrogen lines, which is induced by configuration mixing. [Compare the above equation with Eq. (25a′) in Ref. 2, where configuration mixing was neglected.] However, if the line-broadening profile, ϕ_b, is *symmetric*, as is almost always the case, we easily see that $\langle\omega_b\rangle+\overline{\omega}=0$, so Eq. (31) reduces to $\langle\omega\rangle=\langle\omega_a\rangle$ for *all* four Stokes profiles. In this case, Eq. (25′) gives directly the algebraic first-order moments of polarized hydrogen lines also in the presence of broadening mechanisms other than the external fields.

IV. CONCLUSIONS

Equation (26), along with Eqs. (30) and (24), provides an algebraic expression for the quadratic Stark effect on the centers of gravity of polarized hydrogen lines.

We remark that the success of the algebraic approach followed in this paper essentially relies on the possibility of expressing the operators $\mathscr{O}_\ell^{(0)}$ and $\mathscr{O}_\ell^{(1)}$, given, respectively, by Eqs. (10) and (12′), in terms of the projection operator $\mathscr{P}_\ell$, so that it becomes possible, making use of standard techniques of projection-operator algebra and of the fundamental results of the spectral theorem, to remove from $\mathscr{O}_\ell^{(0)}$ and $\mathscr{O}_\ell^{(1)}$ any explicit reference to the unknown first-order eigenvalues, $\omega_{\ell\lambda}^{(1)}$, and zeroth-order eigenvectors, $|\psi_{\ell\lambda}^{(0)}\rangle$, which otherwise had to be determined numerically via diagonalization of the perturbation Hamiltonian, $\mathscr{H}_1$.

We think that the availability of an algebraic formula for the centers of gravity of hydrogen lines, accounting for the quadratic Stark effect, might be useful for future estimates of hydrogen line shifts—for instance, in applications of pressure-broadening theory to solar and fusion plasma research—as it opens to the investigation of (relatively) high hydrogen series at the cost of a comparatively modest computational effort, at least for the contribution from the discrete spectrum.

ACKNOWLEDGMENT

The author wishes to thank Philip Judge, of the High Altitude Observatory (Boulder, CO), for reviewing a former version of the paper and for many helpful suggestions.

[1] R. Casini and E. Landi Degl'Innocenti, Astron. Astrophys. **291**, 668 (1994).
[2] R. Casini and E. Landi Degl'Innocenti, Astron. Astrophys. **300**, 309 (1995).
[3] R. Casini and E. Landi Degl'Innocenti, Astron. Astrophys. **308**, 335 (1996).
[4] H. R. Griem, *Spectral Line Broadening by Plasmas* (Academic, New York, 1974).
[5] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Springer, Berlin, 1957).
[6] E. Luc-Koenig and A. Bachelier, J. Phys. B **13**, 1743 (1980).
[7] H. Friedrich, *Theoretical Atomic Physics* (Springer, Berlin, 1990).
[8] J. J. Sakurai, *Modern Quantum Mechanics* (Benjamin, Menlo Park, 1985).
[9] C. E. Soliverez, Am. J. Phys. **35**, 624 (1967).
[10] R. Casini and E. Landi Degl'Innocenti, Astron. Astrophys. **276**, 289 (1993).
[11] D. M. Brink and G. R. Satchler, *Angular Momentum* (Clarendon, Oxford, 1968).
[12] A. R. Ruffa, Am. J. Phys. **41**, 234 (1973).
[13] E. C. Kemble, *Quantum Mechanics* (Dover, New York, 1958).

A geometrical angle on Feynman integrals

A. I. Davydychev[a)]
Physics Department, University of Tasmania,
GPO Box 252-21, Hobart, Tasmania, 7001 Australia
and Institute for Nuclear Physics, Moscow State University, 119899, Moscow, Russia

R. Delbourgo[b)]
Physics Department, University of Tasmania,
GPO Box 252-21, Hobart, Tasmania, 7001 Australia

(Received 22 December 1997; accepted for publication 29 April 1998)

A direct link between a one-loop N-point Feynman diagram and a geometrical representation based on the N-dimensional simplex is established by relating the Feynman parametric representations to the integrals over contents of $(N-1)$-dimensional simplices in non-Euclidean geometry of constant curvature. In particular, the four-point function in four dimensions is proportional to the volume of a three-dimensional spherical (or hyperbolic) tetrahedron which can be calculated by splitting into birectangular ones. It is also shown that the known formula of reduction of the N-point function in $(N-1)$ dimensions corresponds to splitting the related N-dimensional simplex into N rectangular ones. © *1998 American Institute of Physics.* [S0022-2488(98)00709-9]

I. INTRODUCTION

The development of techniques for an efficient calculation of one-loop N-point Feynman diagrams is very important for studying the leading and next-to-leading corrections to elementary particle processes within and beyond the Standard Model. Since any one-loop diagram can be reduced to a combination of scalar integrals,[1] in what follows we shall mainly deal with scalar Feynman integrals.

It is well known (see e.g., in Ref. 2) that results for the three- and four-point functions in four dimensions can be expressed in terms of dilogarithms (or related functions). The first explicit calculation of the general one-loop four-point function was given in Ref. 3 and later on, more compact results were presented in Refs. 2 and 4 (see also in Refs. 5 and 6). Moreover, it is known that (in four dimensions) the five-point function, the ''pentagon,'' can be reduced to a linear combination of four-point functions[7–9] (see also in Refs. 10–13). Using linear dependence of the external momenta, a similar reduction procedure can be applied to the N-point functions with $N \geqslant 6$ (see in Refs. 14 and 8).

As a rule, explicit results for diagrams with several external legs possess a rather complicated analytical structure. For example, separate terms in a sum of dilogarithms may have singularities (cuts) which cancel in the whole sum. As the result, there are certain difficulties in describing analytic continuation to all regions (in external momenta and masses) of interest.

Another approach to the evaluation of one-loop integrals[15] makes it possible to represent them in terms of multiple hypergeometric functions. In Ref. 16, such results for the integrals with an arbitrary number of external legs were presented. As a rule, the corresponding hypergeometric series have a rather restricted region of convergence. In general, the problem of analytic continuation of occurring functions to all other regions of interest is very complicated.

The analytical structure of the results can be better understood if one employs a geometrical interpretation of kinematic invariants and other quantities. For example, the singularities of the general three-point function can be described pictorially through a tetrahedron constructed out of the external and internal momenta. This method can be used to derive Landau equations defining the positions of possible singularities[17] (see also in Refs. 18–20) and a similar approach can be applied to the four-point function[21,3] too.

[a)]Electronic mail: davyd@theory.npi.msu.su
[b)]Electronic mail: Bob.Delbourgo@utas.edu.au

 © 1998 American Institute of Physics

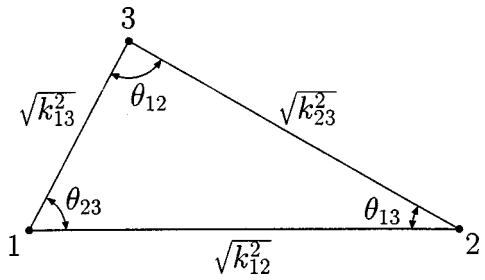

FIG. 1. The triangle associated with the massless three-point function.

Another well-known example of using geometrical ideas is the massless three-point function with arbitrary (off-shell) external momenta. Let us denote them as k_{12}, k_{23}, and k_{31}, so that $k_{12}+k_{23}+k_{31}=0$, and assume that all these momenta are time-like ($k_{jl}^2>0$). Using the standard notation for the ''triangle'' Källen function,

$$\lambda(x,y,z)=x^2+y^2+z^2-2xy-2yz-2zx, \tag{1.1}$$

and assuming that we are in the region $\lambda(k_{12}^2,k_{23}^2,k_{31}^2)\leqslant 0$, the result for the corresponding one-loop scalar Feynman integral (in four dimensions) can be neatly expressed as

$$\frac{2i\pi^2}{\sqrt{-\lambda(k_{12}^2,k_{23}^2,k_{31}^2)}}\{\mathrm{Cl}_2(2\theta_{12})+\mathrm{Cl}_2(2\theta_{23})+\mathrm{Cl}_2(2\theta_{31})\}, \tag{1.2}$$

where the angles θ_{12}, θ_{23}, and θ_{31} are nothing but those of the triangle with sides $\sqrt{k_{12}^2}$, $\sqrt{k_{23}^2}$, and $\sqrt{k_{31}^2}$, as shown in Fig. 1. Moreover, the denominator $\sqrt{-\lambda}$ is nothing but four times the area of this triangle. [Here we use the same normalization as in (2.1). The results for the massless three-point functions can be found, e.g., in Refs. 22, 2, 23. They are closely connected with the results for two-loop massive vacuum diagrams.[24–26] The functions related to the dilogarithm (including the Clausen function $\mathrm{Cl}_2(\theta)$) are defined in Appendix A (see also Ref. 27).]

In this paper, we discuss the geometrical interpretation of the kinematic variables related to the one-loop N-point functions. The generalization of the tetrahedron representation (which was used in the three-point case) is an N-dimensional simplex. Furthermore, we show that there is a direct transition from the Feynman parametric representation to the geometrical description connected with an N-dimensional simplex. We thereby arrive at a ''geometrical'' way of evaluating Feynman integrals.

The paper is organized as follows. In Sec. II we describe the connection of the N-point variables and the corresponding N-dimensional simplex. In Sec. III we show how one can ''geometrize'' the ordinary Feynman parametric representation. In Secs. IV, V, and VI we consider application of these geometrical ideas to the two-, three- and four-point functions, respectively. In Sec. VII we discuss how to use geometrical ideas for the reduction of N-point integrals in $(N-1)$ dimensions, and also for the calculation of the integrals with higher powers of denominators. Finally, in Sec. VIII, we summarize the main ideas and results.

II. BASIC SIMPLICES IN *N* DIMENSIONS

A. *N*-point function and the related simplex

Consider a one-loop scalar N-point ''sun-type'' diagram presented in Fig. 2. The corresponding Feynman integral is

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)\equiv\int\frac{d^nq}{\Pi_{i=1}^N[(p_i+q)^2-m_i^2]^{\nu_i}}, \tag{2.1}$$

where n is the space–time dimension and ν_i are the powers of the propagators. (Here and below, the usual causal prescription for the propagators is understood, i.e., $1/[q^2-m^2]^\nu\leftrightarrow 1/[q^2-m^2+i0]^\nu$.) In general, it depends on $\frac{1}{2}N(N-1)$ momenta invariants k_{jl}^2 ($j<l$), where

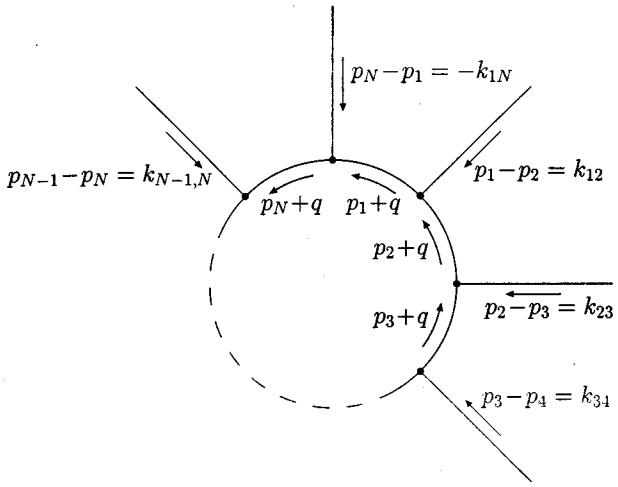

FIG. 2. The one-loop N-point diagram.

$$k_{jl}\equiv p_j-p_l\,, \tag{2.2}$$

and N masses m_i corresponding to the internal propagators. The momenta p_i are auxiliary in the sense that they all can be shifted by a constant vector without changing the momenta k_{jl} (2.2). We shall discuss this freedom below [see Eqs. (2.49)–(2.53)].

It is well known that, using the condition $\Sigma\alpha_i=1$, the standard quadratic form occurring in the denominator of the integrand of the Feynman parametric representation for these integrals [cf. Eqs. (3.1)–(3.2) below] can be rewritten in a homogeneous form,

$$\left[\sum_{j<l}\sum\ \alpha_j\alpha_l k_{jl}^2-\sum\ \alpha_i m_i^2\right]\Rightarrow-\left[\sum\ \alpha_i^2 m_i^2+2\sum_{j<l}\sum\ \alpha_j\alpha_l m_j m_l c_{jl}\right], \tag{2.3}$$

where (in some papers, the notation y_{jl} is used for $\pm c_{jl}$)

$$c_{jl}\equiv\frac{m_j^2+m_l^2-k_{jl}^2}{2m_j m_l}. \tag{2.4}$$

In the region between the corresponding two-particle pseudo-threshold, $k_{jl}^2=(m_j-m_l)^2$, and the threshold, $k_{jl}^2=(m_j+m_l)^2$, we have $|c_{jl}|<1$, and therefore in this region they can be understood as cosines of some angles τ_{jl},

$$c_{jl}=\cos\ \tau_{jl}=\begin{cases}1, & k_{jl}^2=(m_j-m_l)^2\\ -1, & k_{jl}^2=(m_j+m_l)^2.\end{cases} \tag{2.5}$$

The corresponding angles τ_{jl} are

$$\tau_{jl}=\arccos(c_{jl})=\arccos\left(\frac{m_j^2+m_l^2-k_{jl}^2}{2m_j m_l}\right)=\begin{cases}0, & k_{jl}^2=(m_j-m_l)^2\\ \pi, & k_{jl}^2=(m_j+m_l)^2.\end{cases} \tag{2.6}$$

The expressions in other regions should be understood in the sense of analytic continuation, using (when necessary) the causal prescription for the propagators. For example, when $k_{jl}^2<(m_j-m_l)^2(c_{jl}>1)$, i.e., below the pseudothreshold, we should interpret

$$\tau_{jl}=-i\ \mathrm{Arch}(c_{jl})=-\frac{i}{2}\ln\left(\frac{m_j^2+m_l^2-k_{jl}^2+\sqrt{\lambda(m_j^2,m_l^2,k_{jl}^2)}}{m_j^2+m_l^2-k_{jl}^2-\sqrt{\lambda(m_j^2,m_l^2,k_{jl}^2)}}\right), \tag{2.7}$$

where $\lambda(x,y,z)$ is the Källen function defined by Eq. (1.1). When $k_{jl}^2>(m_j+m_l)^2(c_{jl}<-1)$, i.e., above the threshold, we get

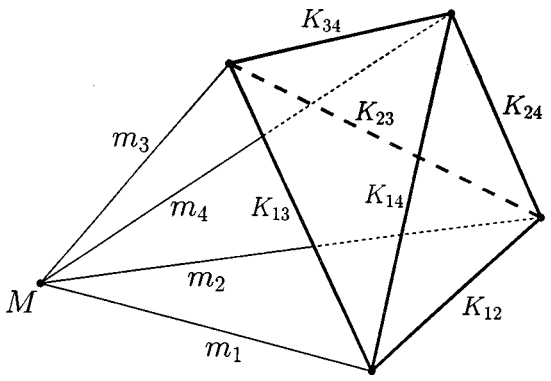

FIG. 3. The basic simplex for $N=4$.

$$\tau_{jl}=\pi+i\ \mathrm{Arch}(-c_{jl})=\pi+\frac{i}{2}\ \ln\left(\frac{k_{jl}^2-m_j^2-m_l^2+\sqrt{\lambda(m_j^2,m_l^2,k_{jl}^2)}}{k_{jl}^2-m_j^2-m_l^2-\sqrt{\lambda(m_j^2,m_l^2,k_{jl}^2)}}\right). \tag{2.8}$$

Note that $\lambda(m_j^2,m_l^2,k_{jl}^2)$ is positive when $k_{jl}^2<(m_j-m_l)^2$ or $k_{jl}^2>(m_j+m_l)^2$.

When the angles τ_{jl} [defined by Eq. (2.6)] are real, the quadratic form on the right-hand side in Eq. (2.3) has a rather simple geometrical interpretation. Let us consider a set of Euclidean ''mass'' vectors whose lengths are m_i. Let them be directed so that the angle between the jth and the lth vectors is τ_{jl}. If we denote the corresponding unit vectors as a_i (so that the ''mass'' vectors are $m_i a_i$), we get

$$(a_j\cdot a_l)=\cos\ \tau_{jl}=c_{jl}. \tag{2.9}$$

This is also valid for the case $j=l$, since $\tau_{jj}=0$ and $c_{jj}=1$, cf. Eqs. (2.5)–(2.6). Now, we can represent Eq. (2.3) as

$$\sum\ \alpha_i^2 m_i^2+2\sum_{j<l}\sum\ \alpha_j\alpha_l m_j m_l c_{jl}=\left(\sum\ \alpha_i m_i a_i\right)^2. \tag{2.10}$$

In N dimensions, if we put all ''mass'' vectors together as emanating from a common origin, they, together with the sides connecting their ends, will define a *simplex* (or hypertetrahedron) which is the *basic* one for a given Feynman diagram. An example of such a simplex for $N=4$ is illustrated by Fig. 3. [Since the four-dimensional Euclidean space is understood, one should consider Fig. 3 just as an illustration (rather than precise picture).] It is easy to see that the length of the side connecting the ends of the jth and the lth mass vectors is $(m_j^2+m_l^2-2m_j m_l c_{jl})^{1/2}=(k_{jl}^2)^{1/2}$, so we shall call it a ''momentum'' side. In total, the *basic* N-dimensional simplex has $\frac{1}{2}N(N+1)$ sides, among them N mass sides (corresponding to the masses $m_1,\ldots,m_N$) and $\frac{1}{2}N(N-1)$ momentum sides (corresponding to the momenta k_{jl}, $j<l$), which meet at $(N+1)$ vertices. Each vertex is a ''meeting point'' for N sides. There is one vertex where all mass sides meet, which we shall call the *mass meeting point*, M. All other vertices are meeting points for $(N-1)$ momentum sides and one mass side. Furthermore, the number of $(N-1)$-dimensional hyperfaces is $(N+1)$. N of them (the *reduced* hyperfaces) can be obtained by excluding one mass side in turn, together with the corresponding vertex and the momentum sides meeting at that vertex. The reduced hyperface without the jth mass side is nothing but the $(N-1)$-dimensional simplex corresponding to an $(N-1)$-point function obtained from the N-point function by shrinking the jth propagator; this is equivalent to the power of propagator vanishing, $\nu_j=0$. The last hyperface, namely the *momentum* hyperface involves all momentum sides but does not include any mass sides. It is associated with the massless N-point function possessing the same external momenta as the basic one.

In general, in the N-point function only $(N-1)$ external momenta are independent, because they are related via the conservation law. We can choose, for example, the vectors $k_{1N},k_{2N},\ldots,k_{N-1,N}$. Considering all scalar products of these momenta, including their squares,

we get $\frac{1}{2}N(N-1)$ invariants, as stated above. However, when n is integer and less than $(N-1)$, only n of the $(N-1)$ momenta are independent, e.g., $k_{1N},k_{2N},\ldots,k_{nN}$. Each of the remaining $(N-n-1)$ vectors k_{jN} $(j=n+1,\ldots,N-1)$ can be fixed by n scalar products $(k_{jN}\cdot k_{lN})$ $(l=1,\ldots,N-1)$. Hence the total number of independent momentum invariants for integer $n\leqslant N-1$ is

$$\tfrac{1}{2}n(n+1)+n(N-n-1)=\tfrac{1}{2}n(2N-n-1). \tag{2.11}$$

while for $n\geqslant N-1$ it remains equal to $\frac{1}{2}N(N-1)$; both expressions coincide for $n=N-1$. The linear dependence between the external momenta reveals itself in the degeneracy of the results associated with the vanishing of certain hypervolumes (see below).

We note that there is a difference between the ''real'' momenta of the N-point diagram, k_{jl}, and the vectors

$$K_{jl}\equiv m_j a_j - m_l a_l, \tag{2.12}$$

corresponding to the momentum sides of the simplex (the same applies to the momenta p_i and P_i discussed in Sec. II D). The momenta k_{jl} are defined in *pseudo-Euclidean* (Minkowski) space, and they are time-like $(k_{jl}^2>0)$ in the region between the threshold and pseudo-threshold. The momentum sides of the simplex, K_{jl}, are vectors in *Euclidean* space, the (Euclidean) squares of their lengths (K_{jl}^2) being equal to k_{jl}^2. In what follows, we shall consider the K_{jl} as Euclidean analogs of the k_{jl} and, *vice versa*, the k_{jl} are pseudo-Euclidean analogs of the K_{jl}. Since the k_{jl}^2 $(j<l)$ form the complete basis of external invariants, all invariant scalars in Minkowski and Euclidean spaces are the same.

The matrix with the components (2.4),

$$\|c\|\equiv\|c_{jl}\|\equiv\begin{pmatrix} 1 & c_{12} & c_{13} & \ldots & c_{1N} \\ c_{12} & 1 & c_{23} & \ldots & c_{2N} \\ c_{13} & c_{23} & 1 & \ldots & c_{3N} \\ \multicolumn{5}{c}{\ldots\ldots\ldots\ldots\ldots\ldots} \\ c_{1N} & c_{2N} & c_{3N} & \ldots & 1 \end{pmatrix}, \tag{2.13}$$

is associated with many geometrical properties of the basic simplex. [The matrix (2.13) is nothing but the Gram matrix of the vectors $a_1,\ldots,a_N$.] In particular, we shall need its determinant,

$$D^{(N)}\equiv\det\|c_{jl}\|, \tag{2.14}$$

and the minors

$$D_{jl}^{(N-1)}\equiv\{\text{minor of (2.13) obtained by eliminating the } j\text{th row and the } l\text{th column}\}. \tag{2.15}$$

In particular, when $j=l$ the minor $D_{jj}^{(N-1)}$ is a *principal* one for the matrix (2.13).

The formulas for the *content* or hypervolume of the basic N-dimensional simplex and its hyperfaces, in terms of the determinant (2.14) and the minors (2.15), are well known in the N-dimensional Euclidean geometry (see for instance Refs. 28 and 29). Thus the content of the N-dimensional simplex is given by

$$V^{(N)}=\frac{1}{N!}\left(\prod_{i=1}^{N} m_i\right)\sqrt{D^{(N)}}. \tag{2.16}$$

The content of the jth $(N-1)$-dimensional reduced hyperface (which does not contain the jth mass side) is

$$\overline{V}_j^{(N-1)}=\frac{1}{(N-1)!}\left(\prod_{i\neq j} m_i\right)\sqrt{D_{jj}^{(N-1)}}, \tag{2.17}$$

whereas the content of the $(N-1)$-dimensional momentum hyperface is

$$\overline{V}_0^{(N-1)}=\frac{1}{(N-1)!}\sqrt{\Lambda^{(N)}}, \tag{2.18}$$

where $\Lambda^{(N)}$ is symmetric combination of the momenta defined in the same way as in Ref. 11,

$$\Lambda^{(N)}\equiv\det\|(k_{jN}\cdot k_{lN})\|. \tag{2.19}$$

Note that the content of the basic simplex can also be presented as

$$V^{(N)}=\frac{1}{N}\,\overline{V}_0^{(N-1)}m_0, \tag{2.20}$$

where m_0 is the distance between the mass meeting point and the momentum hyperface, i.e., the length of the vector of the height of the simplex, H_0. Therefore,

$$m_0\equiv|H_0|=N\,\frac{V^{(N)}}{\overline{V}_0^{(N-1)}}=\left(\prod_{i=1}^{N}m_i\right)\sqrt{\frac{D^{(N)}}{\Lambda^{(N)}}}. \tag{2.21}$$

B. Normals, dihedral angles, and the dual matrix

The dihedral angles between the hyperfaces can be defined via the angles between their normals. Thus, we get $\frac{1}{2}N(N-1)$ dihedral angles between pairs of the reduced hyperfaces and N angles between the momentum hyperface and the reduced ones.

Again, let a_i be unit vectors directed along the mass sides of the basic N-dimensional simplex. The normals to the $(N-1)$-dimensional reduced hyperfaces can be defined as

$$n_{j\lambda}=\frac{\partial}{\partial a_j^{\lambda}}\,(\epsilon_{\nu_1\ldots\nu_N}a_1^{\nu_1}\ldots a_N^{\nu_N}), \tag{2.22}$$

where $\epsilon_{\nu_1\ldots\nu_N}$ is the completely antisymmetric tensor in N dimensions.

Using the well-known expression for the contraction of the product of two ϵ tensors in terms of the determinant with Kronecker δ symbols, it is easy to show that

$$(n_j\cdot n_l)=(-1)^{j+l}D_{jl}^{(N-1)}, \tag{2.23}$$

where $D_{jl}^{(N-1)}$ is defined in (2.15). In particular, n_j^2 is nothing but the principal minor $D_{jj}^{(N-1)}$ related to the $(N-1)$-point function associated with the corresponding reduced hyperface. The cosine of the angle $\tilde{\tau}_{jl}$ between the normals to the jth and the lth reduced hyperfaces is

$$\tilde{c}_{jl}\equiv\cos\,\tilde{\tau}_{jl}=\frac{(n_j\cdot n_l)}{\sqrt{n_j^2n_l^2}}=\frac{(-1)^{j+l}D_{jl}^{(N-1)}}{\sqrt{D_{jj}^{(N-1)}D_{ll}^{(N-1)}}}. \tag{2.24}$$

In particular, this means that the matrix with the elements $\sqrt{D_{jj}^{(N-1)}D_{ll}^{(N-1)}}\tilde{c}_{jl}/D^{(N)}$ is the inverse of the matrix (2.13). Namely,

$$\|c_{jl}\|^{-1}=\frac{1}{D^{(N)}}\,\mathrm{diag}(\sqrt{D_{11}^{(N-1)}},\ldots,\sqrt{D_{NN}^{(N-1)}})\|\tilde{c}_{jl}\|\mathrm{diag}(\sqrt{D_{11}^{(N-1)}},\ldots,\sqrt{D_{NN}^{(N-1)}}), \tag{2.25}$$

where $D^{(N)}$ is defined in (2.14). We shall call the matrix $\|\tilde{c}_{jl}\|$ the *dual* matrix (with respect to $\|c_{jl}\|$). [The matrix $\|\tilde{c}_{jl}\|$ is nothing but the Gram matrix of unit vectors directed along the normals n_i $(i=1,\ldots,N)$.] A trivial corollary of (2.25) is that

$$\tilde{D}^{(N)}\equiv\det\|\tilde{c}_{jl}\|=\frac{(D^{(N)})^{N-1}}{\Pi_{i=1}^{N}D_{ii}^{(N-1)}}. \tag{2.26}$$

Note that the dihedral angles ψ_{jl} are related to the angles $\tilde{\tau}_{jl}$ as

$$\psi_{jl}=\pi-\tilde{\tau}_{jl},\ \ \cos\,\psi_{jl}=-\cos\,\tilde{\tau}_{jl}. \tag{2.27}$$

The normal to the momentum hyperface can be defined as

$$n_{0\lambda}=-\epsilon_{\mu_1\ldots\mu_{N-1}\lambda}(K_{1N})^{\mu_1}\ldots(K_{N-1,N})^{\mu_{N-1}}. \tag{2.28}$$

If we represent $K_{jN}=(m_j a_j-m_N a_N)$ and use the fact that all terms involving two or more $m_N a_N$ disappear in (2.28) (due to the antisymmetry of the ϵ tensor), we get the following representation

$$n_{0\lambda}=-\left(\prod_{i=1}^{N} m_i\right)\sum_{j=1}^{N}\frac{n_{j\lambda}}{m_j}. \tag{2.29}$$

On one hand, the original definition (2.28) leads to

$$n_0^2=\det\|(k_{jN}\cdot k_{lN})\|\equiv\Lambda^{(N)}, \tag{2.30}$$

where $\Lambda^{(N)}$ is defined in (2.19). On the other hand, using (2.29) we obtain

$$n_0^2=\left(\prod_{i=1}^{N} m_i^2\right)\sum_{l=1}^{N}\frac{1}{m_l^2}F_l^{(N)}, \tag{2.31}$$

where [cf. Eq. (4) of Ref. 11]

$$F_l^{(N)}=\sum_{j=0}^{N}(-1)^{j+l}D_{jl}^{(N-1)}\frac{m_l}{m_j}=\frac{\partial}{\partial m_l^2}\,(m_l^2 D^{(N)}), \tag{2.32}$$

i.e., $F_l^{(N)}$ is the determinant obtained from $\det\|c_{jl}\|$ by substituting, instead of the lth column, the mass ratios m_l/m_j (where j is the line number). When taking the partial derivative in (2.32), it is implied that the set of independent variables involves m_i^2 and k_{jl}^2, rather than c_{jl}. In particular, the representation of $F_l^{(N)}$ in terms of determinants (2.32) (see also in Ref. 11) shows that they obey the following set of linear equations:

$$\sum_{l=1}^{N} c_{jl}F_l^{(N)}\frac{1}{m_l}=D^{(N)}\frac{1}{m_j}. \tag{2.33}$$

Another useful representation of n_0 is

$$n_0=-\frac{\Pi m_l}{\sqrt{D^{(N)}}}\sum_{i=1}^{N}\frac{1}{m_i}F_i^{(N)}a_i. \tag{2.34}$$

where, as usually, a_i are the unit vectors directed along the mass sides of the basic simplex. Using Eq. (2.33) (and remembering that $(a_j\cdot a_l)=c_{jl}$), one can easily see that the representation (2.34) is equivalent to Eqs. (2.28) and (2.29). Namely, n_0 (defined by (2.34)) is orthogonal to all $K_{jl}=m_j a_j-m_l a_l$ and n_0^2 is the same as in Eq. (2.31).

C. Splitting the basic simplex

The geometrical meaning of $F_l^{(N)}$ can be understood as follows. Using Eqs. (2.29) and (2.23), we get

$$(n_0\cdot n_l)=-\left(\prod_{i\neq l} m_i\right)F_l^{(N)}. \tag{2.35}$$

Therefore, the cosine of the angle $\tilde{\tau}_{0l}$ between the normals to the momentum hyperface and the lth reduced hyperface is

$$\tilde{c}_{0l} \equiv \cos\ \tilde{\tau}_{0l} = \frac{(n_0 \cdot n_l)}{\sqrt{n_0^2 n_l^2}} = -\frac{(\Pi_{i \neq l} m_i) F_l^{(N)}}{\sqrt{\Lambda^{(N)} D_{ll}^{(N-1)}}}, \tag{2.36}$$

whereas the dihedral angle between these hyperfaces is given by

$$\psi_{0l} = \pi - \tilde{\tau}_{0l}, \quad \cos\ \psi_{0l} = -\cos\ \tilde{\tau}_{0l}. \tag{2.37}$$

The orthogonal projection of the ith reduced hyperface onto the momentum hyperface is

$$\overline{V}_i^{(N-1)} \cos\ \psi_{0i}. \tag{2.38}$$

The sum of all such projections should cover the whole momentum hyperface,

$$\sum_{i=1}^{N} \overline{V}_i^{(N-1)} \cos\ \psi_{0i} = \overline{V}_0^{(N-1)}. \tag{2.39}$$

Using Eqs. (2.36), (2.17), and (2.18), we see that condition (2.39) is equivalent to

$$\left(\prod_{j=1}^{N} m_j^2 \right) \sum_{l=1}^{N} \frac{F_l^{(N)}}{m_l^2} = \Lambda^{(N)}, \tag{2.40}$$

which is the case, cf. Eqs. (2.30) and (2.31). The condition (2.40) is equivalent to Eq. (38) of Ref. 11 (see also in Ref. 3). Equation (2.39) illustrates the geometrical meaning of (2.40). Moreover, we can see that, if we use the ''height'' H_0 for splitting the basic N-dimensional simplex into N rectangular ones whose ''bases'' correspond to the projections (2.38), the content of the ith rectangular simplex $V_i^{(N)}$ is proportional to $F_i^{(N)}$, namely,

$$V_i^{(N)} = \frac{1}{N} \overline{V}_i^{(N-1)} m_0 \cos\ \psi_{0i} = \frac{V^{(N)}}{\Lambda^{(N)}} \left(\prod_{l \neq i} m_l^2 \right) F_i^{(N)}. \tag{2.41}$$

In the same manner we can consider a projection onto one of the reduced hyperfaces. This gives

$$\sum_{l \neq i} \overline{V}_l^{(N-1)} \cos\ \psi_{il} + \overline{V}_0^{(N-1)} \cos\ \psi_{0i} = \overline{V}_i^{(N-1)} \tag{2.42}$$

or

$$\sum_{l=1}^{N} \overline{V}_l^{(N-1)} \tilde{c}_{il} + \overline{V}_0^{(N-1)} \tilde{c}_{0i} = 0. \tag{2.43}$$

In terms of $F_l^{(N)}$, Eq. (2.43) corresponds to definition (2.32).

The cosine of the angle τ_{0j} between the jth mass side and the height H_0 is

$$c_{0j} \equiv \cos\ \tau_{0j} = \frac{m_0}{m_j}. \tag{2.44}$$

Consider the matrix of cosines (similar to (2.13)) associated with the ith rectangular simplex. Its elements are

$$\begin{cases} c_{jl}, & \text{if } j \neq i \text{ and } l \neq i \\ c_{0l}, & \text{if } j = i \text{ and } l \neq i \\ c_{0j}, & \text{if } l = i \text{ and } j \neq i \\ 1, & \text{if } j = l = i. \end{cases} \tag{2.45}$$

Denoting

$$D_i^{(N)} \equiv \{\text{determinant of the matrix } (2.45)\}, \tag{2.46}$$

and using Eqs. (2.16) and (2.21), we get

$$V_i^{(N)} = \left(\prod_{l \neq i} m_l \right) \sqrt{\frac{D_i^{(N)}}{\Lambda^{(N)}}}. \tag{2.47}$$

Comparing Eqs. (2.41) and (2.47), we obtain the following result for the determinant (2.46):

$$D_i^{(N)} = \left(\prod_{l \neq i} m_l \right) \frac{(F_i^{(N)})^2}{\Lambda^{(N)}}. \tag{2.48}$$

D. Choice of the momenta p_i

Let us discuss possibilities for a choice of the momenta p_i in Eq. (2.1) and their Euclidean analogs P_i, such that $K_{jl} = P_j - P_l$. Due to Eq. (2.2), the p_i vectors emanating from an origin, together with the vectors k_{jl}, form an N-dimensional simplex in pseudo-Euclidean space. (For integer $n < N$, this simplex is degenerate, since one cannot use more than n dimensions for external vectors.) Analogously, in Euclidean space the vectors P_i together with the K_{jl} momentum hyperface (taken as a base), form an N-dimensional Euclidean simplex similar to the basic one. Using translational invariance of p_i, we can shift their origin as we like. In particular, we can make one of the p_i vectors zero, whereupon the origin of P_i coincides with one of the vertices belonging to the momentum hyperface. However, this way would break the symmetry of the problem. There are at least two reasonable, *symmetric* ways to fix the origin of p_i (and P_i):

(1) Choose the mass meeting point as the origin of P_i vectors, so that $P_i^2 = p_i^2 = m_i^2$ and the P_i vectors coincide with the mass sides, $P_i = m_i a_i$. In this case, the basic simplex becomes even ''more basic,'' and the integral (2.1) can be expressed as

$$J^{(N)}(n; \nu_1, \ldots, \nu_N) = \int \frac{d^n q}{\prod_{i=1}^N [q^2 + 2(p_i \cdot q)]^{\nu_i}} \Bigg|_{(p_j \cdot p_l) = m_j m_l c_{jl}}. \tag{2.49}$$

Such a choice of p_i simplifies study of singularities via Landau equations.[17,20](In this case, the main Landau singularity is associated with $q \to 0$.) Note that, according to Eq. (2.34), in this case the vector of the height H_0 can be represented as

$$H_0 = \frac{\Pi m_l^2}{\Lambda^{(N)}} \sum_{i=1}^N \frac{1}{m_i^2} F_i^{(N)} P_i. \tag{2.50}$$

However, this choice of p_i cannot be used for integer $n < N$.

(2) Pick, as the origin of P_i vectors, the point of intercept of the height H_0 and the momentum hyperface. This is nothing but a projection of the previous option onto the momentum hyperface. In this case (cf. Eq. (2.44)),

$$p_i^2 = P_i^2 = m_i^2 - m_0^2 = m_i^2 (1 - c_{0i}^2), \tag{2.51}$$

$$(p_j \cdot p_l) = (P_j \cdot P_l) = m_j m_l c_{jl} - m_0^2 = m_j m_l (c_{jl} - c_{0j} c_{0l}), \tag{2.52}$$

and the denominators in the integral (2.1) become

$$[(p_i + q)^2 - m_i^2] \Rightarrow [q^2 + 2(p_i \cdot q) - m_0^2]. \tag{2.53}$$

The projection of Eq. (2.34) onto the momentum hyperface shows that there exists the following relation which is valid for *this* choice of P_i,

$$\sum_{i=1}^N \frac{1}{m_i^2} F_i^{(N)} P_i = 0. \tag{2.54}$$

III. FROM FEYNMAN PARAMETERS TO NON-EUCLIDEAN GEOMETRY

A. Moving from linear to quadratic hypersurface

The standard Feynman parametric representation of a one-loop N-point integral in n dimensions (2.1), corresponding to the diagram presented in Fig. 2, reads

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)=i^{1-n}\pi^{n/2}\,\frac{\Gamma\left(\Sigma\nu_i-\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\int_0^1\cdots\int_0^1\frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i\,\delta(\Sigma\alpha_i-1)}{[\Sigma\sum\limits_{j<l}\alpha_j\alpha_l k_{jl}^2-\Sigma\alpha_i m_i^2]^{\Sigma\nu_i-n/2}}. \tag{3.1}$$

As we have already mentioned (2.3), the integrand can be written in a homogeneous form,

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)=i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\Sigma\nu_i-\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\int_0^1\cdots\int_0^1\frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i\,\delta(\Sigma\alpha_i-1)}{[\Sigma\alpha_i^2m_i^2+2\Sigma\sum\limits_{j<l}\alpha_j\alpha_l m_j m_l c_{jl}]^{\Sigma\nu_i-n/2}}, \tag{3.2}$$

where the ''cosines'' c_{jl} are defined in (2.4). Note that the limits of integration in Eqs. (3.1)–(3.2) can be extended from (0,1) to $(0,\infty)$, since the actual region of integration is defined by the δ function. It corresponds to a part of an $(N-1)$-dimensional hyperplane $\Sigma\alpha_i=1$ which is cut out by the conditions $\alpha_i\geqslant 0$ $(i=1,\ldots,N)$.

Now, let us consider the integral (3.2), and let us use a rescaling which is similar to one used in Refs. 21, 2, 4 (see also in Ref. 25, where the transformations of a general form have been discussed). Let us rescale $\alpha_i=m_i^{-1}\alpha_i'$, so that the δ function becomes

$$\delta\left(\sum\ \frac{\alpha_i'}{m_i}-1\right). \tag{3.3}$$

To restore the argument of the δ function in its original form, let us substitute

$$\alpha_i'=\mathscr{F}(\alpha_1'',\ldots,\alpha_N'')\,\alpha_i'',\ \text{ with }\ \mathscr{F}(\alpha_1,\ldots,\alpha_N)=\frac{\Sigma\alpha_i}{\Sigma\,\dfrac{\alpha_i}{m_i}}. \tag{3.4}$$

Note that the Jacobian of this substitution is $\mathscr{F}^N$ (cf. Appendix B).

Suppressing the primes, we arrive at the following representation:

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)=i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\Sigma\nu_i-\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\,\frac{1}{\Pi m_i^{\nu_i}}\int_0^1\cdots\int_0^1\frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i\,\delta(\Sigma\alpha_i-1)}{\left(\Sigma\,\dfrac{\alpha_i}{m_i}\right)^{n-\Sigma\nu_i}(\alpha^T\|c\|\alpha)^{\Sigma\nu_i-n/2}}, \tag{3.5}$$

where we use matrix notation (2.13),

$$\alpha^T\|c\|\alpha\equiv\sum_{j=1}^N\sum_{l=1}^N c_{jl}\alpha_j\alpha_l=\sum\ \alpha_i^2+2\sum_{j<l}\sum\ \alpha_j\alpha_l c_{jl}. \tag{3.6}$$

Next, let us change the variables from α to α' via

$$\alpha_i=\mathscr{G}(\alpha_1',\ldots,\alpha_N')\,\alpha_i',\ \text{ with }\ \mathscr{G}(\alpha_1',\ldots,\alpha_N')=\frac{\Sigma(\alpha_i')^2+2\Sigma\sum\limits_{j<l}\alpha_j'\alpha_l'c_{jl}}{\Sigma\alpha_i'}. \tag{3.7}$$

Here, the Jacobian of the substitution is $2\mathscr{F}^N$ (see Appendix B). Effectively, we put the quadratic form into the argument of the δ function. However, the linear denominator (coming from the $\mathscr{F}$ function) survives and we arrive at

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)=2i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\Sigma\nu_i-\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\,\frac{1}{\Pi m_i^{\nu_i}}\int_0^\infty\cdots\int_0^\infty\frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i}{\left(\Sigma\,\frac{\alpha_i}{m_i}\right)^{n-\Sigma\nu_i}}\,\delta(\alpha^T\|c\|\alpha-1), \tag{3.8}$$

where we continue to use the matrix notation (3.6).

B. The case $\Sigma\nu_i=n$

In the particular case $\Sigma\nu_i=n$ the linear denominator disappears and Eq. (3.5) gives

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)\big|_{\Sigma\nu_i=n}=i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\,\frac{1}{\Pi m_i^{\nu_i}}\int_0^1\cdots\int_0^1\frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i\,\delta(\Sigma\alpha_i-1)}{(\alpha^T\|c\|\alpha)^{\Sigma\nu_i-n/2}}. \tag{3.9}$$

Here all dependence on the external momenta and masses in the integral is through c_{jl}, Eq. (2.4).

Since the final integral (3.9) is the same as in the equal-mass case, we can formulate the following statement: Let $\Sigma\nu_i=n$, and let the result for the Feynman integral (3.1) with equal masses ($m_i=m$, $i=1,\ldots,N$) be represented by a dimensionless function Ψ depending on the quantities $\bar{\bar{c}}_{jl}\equiv 1-k_{jl}^2/(2m^2)$ (corresponding to Eq. (2.4) in the equal-mass case) as

$$J^{(n)}(n;\nu_1,\ldots,\nu_N)\big|_{\Sigma\nu_i=n,\quad m_i=m}=\frac{1}{m^{\Sigma\nu_i}}\,\Psi(\{\bar{\bar{c}}_{jl}\}). \tag{3.10}$$

Then the result for the corresponding integral with different masses m_i can be expressed in terms of *the same* function Ψ as

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)\big|_{\Sigma\nu_i=n}=\frac{1}{\Pi m_i^{\nu_i}}\,\Psi(\{c_{jl}\}), \tag{3.11}$$

where c_{jl} are defined in Eq. (2.4).

Simplification of the integrals $J^{(N)}(n;\nu_1,\ldots,\nu_N)$ when $\Sigma\nu_i=n$ is not surprising; to some extent, this may be considered a generalization of the so-called ''uniqueness'' formula for massless triangle diagrams[30] to the case of massive N-point integrals.

Whenever $\Sigma\nu_i=n$ the denominator of the integrand of (3.8) disappears, and the result is

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)\big|_{\Sigma\nu_i=n}=2i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\,\frac{1}{\Pi m_i^{\nu_i}}\int_0^\infty\cdots\int_0^\infty\prod\,\alpha_i^{\nu_i-1}d\alpha_i\,\delta(\alpha^T\|c\|\alpha-1). \tag{3.12}$$

In particular, when all $\nu_i=1$ ($n=N$), our integrand is just a δ function,

$$J^{(N)}(N;1,\ldots,1)=2i^{1-2N}\pi^{N/2}\,\frac{\Gamma\left(\frac{N}{2}\right)}{\Pi m_i}\int_0^\infty\cdots\int_0^\infty\prod\,d\alpha_i\,\delta(\alpha^T\|c\|\alpha-1). \tag{3.13}$$

Consider also a very special situation, when all nondiagonal c_{jl} vanish (i.e., $c_{jl}=\delta_{jl}$). Physically, this corresponds to the case where all external momenta squared take the mean values

between the threshold and the pseudo-threshold, $k_{jl}^2=m_j^2+m_l^2$ (for $j<l$). In this case, all the angles τ_{jl} ($j\neq l$) are equal to $\frac{1}{2}\pi$. Employing the representation (3.12) we arrive at a very simple result for this special case,

$$J^{(N)}(n;\nu_1,\ldots,\nu_N)\Big|_{\substack{\Sigma\nu_i=n\\ c_{jl}=\delta_{jl}}}=i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Pi\Gamma\left(\frac{1}{2}\,\nu_i\right)}{2^{N-1}\Pi\Gamma(\nu_i)}\,\frac{1}{\Pi m_i^{\nu_i}}. \tag{3.14}$$

Imagine the N-dimensional Euclidean space of the parameters α_i. The quadratic argument of the δ function in (3.13) defines a $(N-1)$-dimensional hypersurface. Provided that all eigenvalues of this quadratic form are positive, this is an N-dimensional ellipsoid. Otherwise, this is an N-dimensional hyperboloid. So, the integral (3.13) is nothing but the measure (i.e., the content) of the part of this hypersurface which belongs to the region where all α's are non-negative. This region is a 2^Nth part of the whole N-dimensional space.

To calculate the content of this part of the hypersurface, it is reasonable to diagonalize the quadratic form. Suppose there is an N-dimensional rotation transforming the old coordinates α_i into the new ones, β_i, so that

$$\alpha^T\|c\|\alpha=\sum\ \alpha_i^2+2\sum_{j<l}\sum\ \alpha_j\alpha_l c_{jl}\Rightarrow\sum\ \lambda_i\beta_i^2. \tag{3.15}$$

Obviously, the product of λ's must be equal to the determinant $D^{(N)}$ of the matrix (2.13) corresponding to the quadratic form (3.6),

$$\lambda_1\ldots\lambda_N=D^{(N)}. \tag{3.16}$$

Note that this determinant $D^{(N)}$ is related to the content $V^{(N)}$ of the basic N-dimensional simplex via Eq. (2.16).

Now, let us assume that all eigenvalues λ_i are real and positive, i.e., the hypersurface defined by the quadratic form is an N-dimensional ellipsoid. After rotation from α's to β's, we get into the coordinate system corresponding to the principal axes of the ellipsoid. Then, we can rescale

$$\beta_i=\frac{\gamma_i}{\sqrt{\lambda_i}}, \tag{3.17}$$

the Jacobian of this transformation being $(\lambda_1\ldots\lambda_N)^{-1/2}=(D^{(N)})^{-1/2}$, according to (3.16). In this way, our N-dimensional ellipsoid is transformed into a hypersphere. Now all we need to calculate is the content of a part of this hypersphere which is cut out (in the space of γ_i) by the images of the hyperfaces restricting the region where all α_i are positive (in the space of α_i). This content, which we shall denote by $\Omega^{(N)}$, can be understood as the N-dimensional solid angle subtended by the above-mentioned hyperfaces. In terms of $\Omega^{(N)}$, the relevant integral (3.13) can be written as

$$J^{(N)}(N;1,\ldots,1)=i^{1-2N}\pi^{N/2}\,\frac{\Gamma\left(\frac{N}{2}\right)}{\Pi m_i}\,\frac{\Omega^{(N)}}{\sqrt{D^{(N)}}}. \tag{3.18}$$

An interesting fact is that $\Omega^{(N)}$ is also related to the basic simplex. To see this, let us note that effectively (in terms of α's), the transformation $\alpha_i\to\beta_i\to\gamma_i$ is equivalent to using the matrix $\|c\|$ as a new metric. Say, for a vector α the *new* length squared is

$$|\alpha|_c^2\equiv\alpha^T\|c\|\alpha. \tag{3.19}$$

For unit vectors $e_i^{(\alpha)}$, the *new* scalar product is

$$(e_j^{(\alpha)}\cdot e_l^{(\alpha)})_c=c_{jl}\equiv\cos\ \tau_{jl}\,. \tag{3.20}$$

Taking into account that $|e_i^{(\alpha)}|_c^2=1$, we see that in the space with the c-metric (i.e., in the space of γ_i) the angle between $e_j^{(\alpha)}$ and $e_l^{(\alpha)}$ is τ_{jl}. Therefore, we get nothing but the N-dimensional solid angle at the mass meeting point of the basic simplex. It defines the region of integration on the hypersurface of the unit hypersphere in the γ_i coordinate system. We also note that the cosines c_{jl} are invariant under simple rescaling of α's.

It is easy to see that the image of $e_i^{(\alpha)}$ in the γ_i-space is directed along the vector a_i. Moreover, according to Eq. (2.34), the image of the vector with the components $F_i^{(N)}/m_i$ is directed along the height H_0.

To summarize, the following statement is valid: The content of the N-dimensional solid angle $\Omega^{(N)}$ in the space of γ_i is equal to that at the vertex of the basic N-dimensional simplex where all mass sides meet. Moreover, the angles between the corresponding hyperfaces in the space of γ_i and those in the basic simplex are the same. Therefore, the result for the integral (3.13) can be expressed in terms of the content of the basic simplex and the content of its N-dimensional solid angle at the vertex which is common for all mass sides as

$$J^{(N)}(N;1,...,1)=i^{1-2N}\pi^{N/2}\,\frac{\Gamma\left(\frac{N}{2}\right)}{N!}\,\frac{\Omega^{(N)}}{V^{(N)}}. \tag{3.21}$$

Looking at Eq. (3.21), we see that $\Omega^{(N)}$ is indeed the only thing which is to be calculated, since $V^{(N)}$ is known through Eq. (2.16). In the following sections, we shall consider the lowest examples to understand how to calculate $\Omega^{(N)}$.

Nevertheless, the above discussion leads us to the following significant statement: The content $\Omega^{(N)}$ is nothing but the content of a *non-Euclidean* $(N-1)$-dimensional simplex calculated in the spherical (or hyperbolic, depending on the signature of the eigenvalues λ_i) space of constant curvature. The sides of this non-Euclidean simplex are equal to the angles τ_{jl}. Therefore, the problem of calculating Feynman integrals is intimately connected with the problem of calculating the content of a simplex in non-Euclidean geometry.

C. The general case

To understand how to deal with the general case, when $\Sigma\nu_i\neq n$, we need some modification of the above transformations. First of all, we note that Eqs. (2.33) can be rephrased as

$$\sum_{l=1}^{N}\,(\sqrt{F_j^{(N)}}c_{jl}\sqrt{F_l^{(N)}})\,\frac{\sqrt{F_l^{(N)}}}{m_l}=D^{(N)}\,\frac{\sqrt{F_j^{(N)}}}{m_j}. \tag{3.22}$$

In other words, the vector with components

$$f_i=\frac{\sqrt{F_i^{(N)}}}{m_i} \tag{3.23}$$

is an eigenvector of the matrix $\|C\|\equiv\|C_{jl}\|$ with the elements

$$C_{jl}=(\sqrt{F_j^{(N)}}c_{jl}\sqrt{F_l^{(N)}}), \tag{3.24}$$

and the corresponding eigenvalue is equal to $D^{(N)}$. Note that

$$f^2\equiv(f\cdot f\,)=\sum_{i=1}^{N}\,\frac{F_i^{(N)}}{m_i^2}=\frac{\Lambda^{(N)}}{\Pi_{i=1}^{N}m_i^2}, \tag{3.25}$$

$$\det\|C\|=D^{(N)}\prod_{j=1}^{N}\,F_j^{(N)}. \tag{3.26}$$

The idea is that it may be more convenient to use the quadratic form defined by the matrix $\|C_{jl}\|$, rather than $\|c_{jl}\|$, since we know one of its eigenvectors with the corresponding eigenvalue.

After using the masses m_i to rescale the Feynman parameters α_i and ''hiding'' the quadratic form into the argument of the delta function, we obtained the representation (3.8). Analogously, we can use $1/f_i = m_i/\sqrt{F_i^{(N)}}$, rather than m_i, to rescale the original α's. The result is

$$J^{(N)}(n;\nu_1,\ldots,\nu_N) = 2i^{1-2\Sigma\nu_i}\pi^{n/2}\,\frac{\Gamma\left(\Sigma\nu_i - \frac{n}{2}\right)}{\Pi\Gamma(\nu_i)}\left(\prod f_i^{\nu_i}\right)\int_0^\infty\cdots\int_0^\infty \frac{\Pi\alpha_i^{\nu_i-1}d\alpha_i}{(\Sigma\alpha_i f_i)^{n-\Sigma\nu_i}}\,\delta(\alpha^T\|C\|\alpha-1), \tag{3.27}$$

where notation (3.6) has been used again. A nice feature of this representation is that in the denominator we have the contraction (scalar product) of α and the known eigenvector of the matrix $\|C\|$ defining the quadratic form (see Eqs. (3.22)–(3.24)).

If $\nu_1 = \cdots = \nu_N = 1$, we get

$$J^{(N)}(n;1,\ldots,1) = 2i^{1-2N}\pi^{n/2}\Gamma\left(N-\frac{n}{2}\right)\left(\prod f_i\right)\int_0^\infty\cdots\int_0^\infty \frac{\Pi d\alpha_i}{(\Sigma\alpha_i f_i)^{n-N}}\,\delta(\alpha^T\|C\|\alpha-1). \tag{3.28}$$

Now, let us use the same transformations of variables as earlier. Namely, let us first rotate from α_i to the variables β_i such that the quadratic form becomes diagonal,

$$\alpha^T\|C\|\alpha = \sum_{i=1}^N \lambda_i\beta_i^2. \tag{3.29}$$

One of these β's, say β_N, is directed along the eigenvector f. Therefore, $\lambda_N = D^{(N)}$ and $\Pi_{i=1}^{N-1}\lambda_i = \Pi_{j=1}^N F_j^{(N)}$. Moreover, the denominator is also proportional to β_N. Finally, we rescale $\beta_i = \gamma_i/\sqrt{\lambda_i}$, and the denominator becomes

$$\left(\sum \alpha_i f_i\right) \Rightarrow \frac{1}{\Pi m_i}\,\sqrt{\frac{\Lambda^{(N)}}{D^{(N)}}}\,\gamma_N = \frac{1}{m_0}\,\gamma_N, \tag{3.30}$$

where m_0, defined in Eq. (2.21), is the distance between the mass meeting point and the momentum hyperface (i.e., the length of the height H_0).

Now, the transformation $\alpha_i \to \beta_i \to \gamma_i$ is equivalent to using the matrix $\|C\|$ as a new metric. For a vector α the *new* length squared is

$$|\alpha|_C^2 \equiv \alpha^T\|C\|\alpha. \tag{3.31}$$

For unit vectors $e_i^{(\alpha)}$, the *new* scalar product is

$$(e_j^{(\alpha)}\cdot e_l^{(\alpha)})_C = C_{jl}, \tag{3.32}$$

and the corresponding cosine of the angle is

$$\frac{(e_j^{(\alpha)}\cdot e_l^{(\alpha)})_C}{|e_j^{(\alpha)}|_C|e_l^{(\alpha)}|_C} = \frac{C_{jl}}{\sqrt{C_{jj}C_{ll}}} = c_{jl} \equiv \cos\,\tau_{jl}. \tag{3.33}$$

[Note that in the γ_i-space with the C-metric the $e_i^{(\alpha)}$ vectors are not unit, $|e_i^{(\alpha)}|_C^2 = C_{ii} = F_i^{(N)}$.] Again, as in the case of the c-metric, we get the same N-dimensional solid angle as that occurring at the mass meeting point of the basic simplex.

But we can go further. In particular, we can comprehend what the image of the vector f is. Consider the scalar products

$$(e_i^{(\alpha)}\cdot f)_C = (e_i^{(\alpha)}\cdot\|C\|f) = D^{(N)}(e_i^{(\alpha)}\cdot f) = D^{(N)}f_i, \tag{3.34}$$

$$(f\cdot f\ )_C=(f\cdot\|C\|f\ )=D^{(N)}(f\cdot f\ )=\frac{D^{(N)}\Lambda^{(N)}}{\Pi_{j=1}^{N}m_j^2}. \tag{3.35}$$

Hence, the cosine of the angle between the images of f and $e_i^{(\alpha)}$ is [cf. Eq. (2.44)]

$$\frac{(e_i^{(\alpha)}\cdot f\ )_C}{\sqrt{(e_i^{(\alpha)}\cdot e_i^{(\alpha)})_C(f\cdot f\ )_C}}=\sqrt{\frac{D^{(N)}}{\Lambda^{(N)}}}\prod_{j\neq i}\ m_j=\frac{m_0}{m_i}=\cos\ \tau_{0i}\equiv c_{0i}\,. \tag{3.36}$$

Therefore, in the basic simplex, with the mass sides m_i, etc., the image of the vector f is directed along the height H_0!

In terms of γ's, the integral becomes

$$J^{(N)}(n;1,...,1)=2i^{1-2N}\pi^{n/2}\Gamma\left(N-\frac{n}{2}\right)\ \frac{m_0^{n-N-1}}{\sqrt{\Lambda^{(N)}}}\int_{\Omega^{(N)}}\cdots\int\ \frac{\Pi d\gamma_i}{\gamma_N^{n-N}}\ \delta\left(\sum\ \gamma_i^2-1\right), \tag{3.37}$$

where the integration goes over the interior of the N-dimensional solid angle $\Omega^{(N)}$ of the basic simplex. If we define the angle between the ''running'' unit vector and the Nth axis as θ, the above formula means that we should integrate over the hypersurface of the unit hypersphere with the ''weight'' $(\cos\theta)^{N-n}$, within the limits set by $\Omega^{(N)}$,

$$J^{(N)}(n;1,...,1)=i^{1-2N}\pi^{n/2}\Gamma\left(N-\frac{n}{2}\right)\ \frac{m_0^{n-N}}{\sqrt{D^{(N)}}\Pi m_i}\ \Omega^{(N;n)}, \tag{3.38}$$

with

$$\Omega^{(N;n)}\equiv\int_{\Omega^{(N)}}\cdots\int\ \frac{d\Omega_N}{\cos^{n-N}\ \theta}. \tag{3.39}$$

Obviously, $\Omega^{(N;N)}=\Omega^{(N)}$. Using Eq. (2.16), the formula (3.38) can be exhibited in the following form:

$$J^{(N)}(n;1,...,1)=i^{1-2N}\pi^{n/2}\Gamma\left(N-\frac{n}{2}\right)\ \frac{m_0^{n-N}\Omega^{(N;n)}}{N!\,V^{(N)}}. \tag{3.40}$$

D. Splitting the *N*-dimensional solid angle

Now, let us again use the height H_0 to split the basic N-dimensional simplex into N rectangular ones, each time replacing one of the mass sides, m_i, by $m_0=|H_0|$. The content of the ith rectangular simplex $V_i^{(N)}$ is given by Eq. (2.41). Let us denote the corresponding N-dimensional solid angle as $\Omega_i^{(N)}$, so that

$$\Omega_i^{(N;n)}=\int_{\Omega_i^{(N)}}\cdots\int\ \frac{d\Omega_N}{\cos^{n-N}\ \theta},\quad \Omega^{(N;n)}=\sum_{i=1}^{N}\ \Omega_i^{(N;n)}\,. \tag{3.41}$$

Substituting (3.41) into Eq. (3.40), we get

$$J^{(N)}(n;1,...,1)=\sum_{i=1}^{N}\ \frac{V_i^{(N)}}{V^{(N)}}\,J_i^{(N)}(n;1,...,1), \tag{3.42}$$

where $J_i^{(N)}$ are the integrals corresponding to the rectangular simplices. Specifically, in the integral $J_i^{(N)}$ the internal masses are

$$m_1,\ldots,m_{i-1},m_0,m_{i+1},\ldots,m_N, \tag{3.43}$$

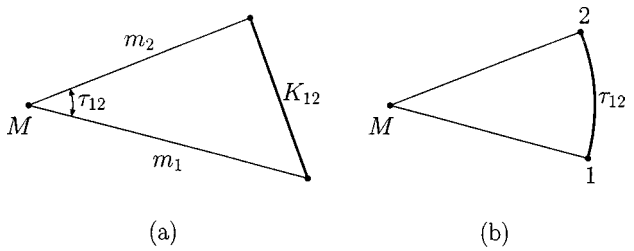

(a) (b)

FIG. 4. Two-point case: (a) the basic triangle and (b) the arc τ_{12}.

while the momenta invariants are

$$\begin{cases} k_{jl}^2, & \text{if } j\neq i \text{ and } l\neq i \\ m_l^2-m_0^2, & \text{if } j=i \\ m_j^2-m_0^2, & \text{if } l=i \end{cases} . \tag{3.44}$$

Using (2.41), we arrive at

$$J^{(N)}(n;1,...,1)=\frac{1}{\Lambda^{(N)}}\left(\prod\ m_i^2\right)\sum_{i=1}^{N}\ \frac{1}{m_i^2}\,F_i^{(N)}J_i^{(N)}(n;1,...,1). \tag{3.45}$$

We are now ready to tackle some specific examples using the geometrical approach.

IV. TWO-POINT FUNCTION

Here, the two-dimensional basic simplex is a triangle and the angle τ_{12} lies between the two mass sides (see Fig. 4). We obtain

$$D^{(2)}=1-c_{12}^2=\sin^2\ \tau_{12},\ \ V^{(2)}=\tfrac{1}{2}m_1m_2\ \sin\ \tau_{12},\ \ \Omega^{(2)}=\tau_{12}, \tag{4.1}$$

$$m_0=m_1m_2\sqrt{\frac{D^{(2)}}{\Lambda^{(2)}}},\ \ \cos\ \tau_{0i}=\frac{m_0}{m_i},\ \ \tau_{01}+\tau_{02}=\tau_{12},\ \ \Lambda^{(2)}=k_{12}^2. \tag{4.2}$$

In two dimensions, from (3.21) we obtain the well-known result

$$J^{(2)}(2;1,1)=\frac{i\pi}{m_1m_2}\,\frac{\tau_{12}}{\sin\ \tau_{12}}, \tag{4.3}$$

while in three dimensions, Eqs. (3.40)–(3.41) yield

$$J^{(2)}(3;1,1)=i\pi^2\,\frac{1}{\sqrt{\Lambda^{(2)}}}\,\{\Omega_1^{(2;3)}+\Omega_2^{(2;3)}\}, \tag{4.4}$$

with

$$\Omega_i^{(2;3)}=\int_0^{\tau_{0i}}\frac{d\theta}{\cos\ \theta}=\ln\left(\frac{1+\sin\ \tau_{0i}}{1-\sin\ \tau_{0i}}\right). \tag{4.5}$$

Combining the logarithms, we get

$$J^{(2)}(3;1,1)=\frac{i\pi^2}{\sqrt{k_{12}^2}}\ln\left(\frac{m_1+m_2+\sqrt{k_{12}^2}}{m_1+m_2-\sqrt{k_{12}^2}}\right). \tag{4.6}$$

In the Euclidean region ($k_{12}^2<0$) the logarithm gives $\arctan(\sqrt{-k_{12}^2}/(m_1+m_2))$, and the result coincides with those presented in Refs. 11 and 31.

In four dimensions, the Γ function in front of the integral (3.38) becomes singular. Introducing dimensional regularization[32] to circumvent that difficulty, we get

$$J^{(2)}(4-2\varepsilon;1,1)=i\pi^{2-\varepsilon}\Gamma(\varepsilon)\,\frac{m_0^{1-2\varepsilon}}{\sqrt{\Lambda^{(2)}}}\,\{\Omega_1^{(2;4-2\varepsilon)}+\Omega_2^{(2;4-2\varepsilon)}\}, \tag{4.7}$$

with

$$\Omega_i^{(2;4-2\varepsilon)}=\int_0^{\tau_{0i}}\frac{d\theta}{\cos^{2-2\varepsilon}\,\theta}. \tag{4.8}$$

Expanding the integrand in ε and taking into account that

$$\int_0^{\tau}\frac{d\theta}{\cos^2\,\theta}\ln(\cos\,\theta)=\tan\,\tau\,\ln(\cos\,\tau)+\tan\,\tau-\tau, \tag{4.9}$$

we reach the well-known result (cf., e.g., in Refs. 2 and 33),

$$J^{(2)}(4-2\varepsilon;1,1)=i\pi^{2-\varepsilon}\Gamma(1+\varepsilon)\times\left\{\frac{1}{\varepsilon}+2-\ln\,m_1-\ln\,m_2+\frac{m_1^2-m_2^2}{k_{12}^2}\ln\frac{m_2}{m_1}-\frac{2m_1m_2}{k_{12}^2}\,\tau_{12}\sin\,\tau_{12}\right\}+\mathcal{O}(\varepsilon). \tag{4.10}$$

Furthermore, the representations (4.7)–(4.8) make it possible to construct next terms of the expansion in ε. For example, the ε term requires just

$$\int_0^{\tau}\frac{d\theta}{\cos^2\,\theta}\ln^2(\cos\,\theta)=\tan\,\tau(\ln^2(\cos\,\tau)+2\ln(\cos\,\tau)+2)-2\tau(1-\ln 2)-\mathrm{Cl}_2(\pi-2\tau) \tag{4.11}$$

(the corresponding result was obtained in Ref. 34; the Clausen function $\mathrm{Cl}_2(\theta)$ is defined in Appendix A). Moreover, the result for an arbitrary space–time dimension or ε can be obtained in terms of the Gauss hypergeometric function (cf. in Refs. 35 and 33), using

$$\Omega_i^{(2;4-2\varepsilon)}=\int_0^{\tau_{0i}}\frac{d\theta}{\cos^{2-2\varepsilon}\,\theta}=2\tan\,\tau_{0i}\;{}_2F_1\left(\left.\begin{matrix}1/2,\varepsilon\\3/2\end{matrix}\right|-\tan^2\,\tau_{0i}\right). \tag{4.12}$$

Using formulas of analytic continuation of the ${}_2F_1$ function, one can establish a connection with the result presented in Eq. (A7) of Ref. 33.

V. THREE-POINT FUNCTION

A. Geometrical picture and the three-dimensional case

Here the three-dimensional basic simplex is a tetrahedron with three mass sides (the angles between these mass sides are τ_{12}, τ_{13}, and τ_{23}) and three momentum sides. It is shown in Fig. 5(a). The volume of this tetrahedron is defined as

$$V^{(3)}=\tfrac{1}{6}m_1m_2m_3\sqrt{D^{(3)}}, \tag{5.1}$$

where (cf. Eqs. (2.13)–(2.14))

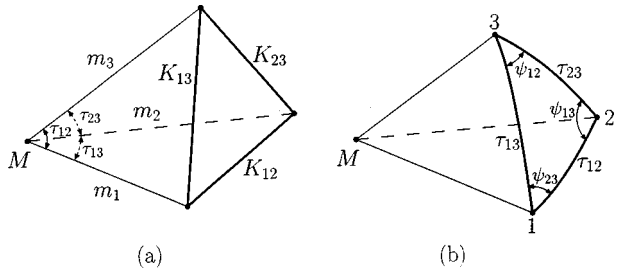

FIG. 5. Three-point case: (a) the basic tetrahedron and (b) the solid angle.

$$D^{(3)} \equiv \begin{vmatrix} 1 & c_{12} & c_{13} \\ c_{12} & 1 & c_{23} \\ c_{13} & c_{23} & 1 \end{vmatrix} = 1 - c_{12}^2 - c_{13}^2 - c_{23}^2 + 2c_{12}c_{13}c_{23}. \tag{5.2}$$

Furthermore, $\Omega^{(3)}$ is the usual solid angle at the vertex derived by the mass sides [cf. Fig. 5(b)]. Its value can be defined as the area of a part of the unit sphere cut out by the three planar faces adjacent to the vertex; in other words, this is the area of a spherical triangle corresponding to this section. The sides of this spherical triangle are obviously equal to the angles τ_{12}, τ_{13}, and τ_{23} while its angles, ψ_{12}, ψ_{13}, and ψ_{23}, are equal to those between the plane faces (we define ψ_{12} as the angle between the sides τ_{13} and τ_{23}, etc.). This definition is in agreement with Eq. (2.27). Using Eq. (2.24) or the well-known formulas of the spherical trigonometry, it is easy to show that

$$\cos\,\psi_{12} = \frac{\cos\,\tau_{12} - \cos\,\tau_{13}\cos\,\tau_{23}}{\sin\,\tau_{13}\sin\,\tau_{23}}, \quad \sin\,\psi_{12} = \frac{\sqrt{D^{(3)}}}{\sin\,\tau_{13}\sin\,\tau_{23}}, \tag{5.3}$$

and analogous expressions for ψ_{13} and ψ_{23}. The solid angle corresponding to spherical triangle can be obtained via

$$\Omega^{(3)} = \psi_{12} + \psi_{13} + \psi_{23} - \pi. \tag{5.4}$$

Using Eqs. (5.3) and (5.4), it is straightforward to show that

$$\begin{aligned} \Omega^{(3)} &= 2\,\arccos\left(\frac{1 + c_{12} + c_{13} + c_{23}}{\sqrt{2(1+c_{12})(1+c_{13})(1+c_{23})}} \right) \\ &= 2\,\arcsin\left(\sqrt{\frac{D^{(3)}}{2(1+c_{12})(1+c_{13})(1+c_{23})}} \right) \\ &= 2\,\arctan\left(\frac{\sqrt{D^{(3)}}}{1 + c_{12} + c_{13} + c_{23}} \right). \end{aligned} \tag{5.5}$$

Finally, the result

$$J^{(3)}(3;1,1,1) = -\frac{i\pi^2}{2m_1m_2m_3} \frac{\Omega^{(3)}}{\sqrt{D^{(3)}}}, \tag{5.6}$$

with $\Omega^{(3)}$ defined by (5.5), corresponds to one obtained in Ref. 11 in a different way. Note that here we have derived it purely geometrically.

However, Eq. (5.4) cannot be easily generalized to the four-dimensional case. This is why it is instructive to reproduce the three-dimensional result (5.5) in a different manner, splitting the basic tetrahedron into rectangular ones (cf. Eq. (3.42)).

Consider a spherical triangle 123 on the unit sphere, cut out by the solid angle of the basic tetrahedron in the γ space, as illustrated in Fig. 6. The points 1,2,3 correspond to the intersections of the mass sides of the basic tetrahedron and the unit sphere. The sides of the triangle are equal

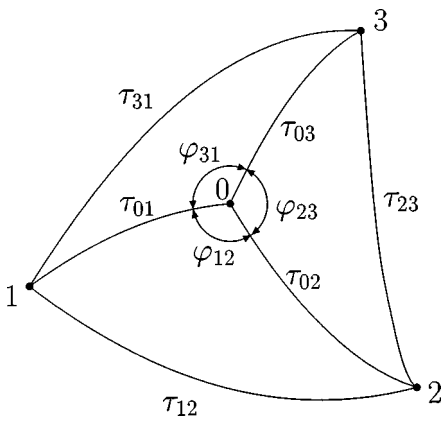

FIG. 6. The spherical triangle 123.

to the angles τ_{12}, τ_{23}, and τ_{31}. There is another point 0, also on the unit sphere, corresponding to the intersection of the height H_0 of the tetrahedron and the unit sphere. The point 0 is connected with each of the points 1,2,3. The corresponding sides 01, 02, and 03 are equal to τ_{01}, τ_{02}, and τ_{03}, respectively, so that

$$\cos\,\tau_{0i}=\frac{m_0}{m_i},\quad m_0=m_1m_2m_3\sqrt{\frac{D^{(3)}}{\Lambda^{(3)}}}, \tag{5.7}$$

with (cf. Eq. (2.19))

$$\Lambda^{(3)}=-\tfrac{1}{4}[(k_{12}^2)^2+(k_{13}^2)^2+(k_{23}^2)^2-2k_{12}^2k_{13}^2-2k_{12}^2k_{23}^2-2k_{13}^2k_{23}^2]=-\tfrac{1}{4}\lambda(k_{12}^2,k_{13}^2,k_{23}^2), \tag{5.8}$$

where $\lambda(x,y,z)$ is defined in Eq. (1.1) (cf. Fig. 1). The triangle 123 is thereby split into three triangles 012, 023, and 031. The angle between 01 and 02 is denoted as φ_{12}, etc. Obviously,

$$\varphi_{12}+\varphi_{23}+\varphi_{31}=2\pi. \tag{5.9}$$

For the calculation of three-dimensional triangle integral, it does not really matter where the point 0 lies, though this becomes essential for the four-dimensional case. As we know, the $N=n$ integrals can be reduced to the integrals with equal masses. So, let us start with the equal mass case. In Fig. 6, this means that $\tau_{01}=\tau_{02}=\tau_{03}\equiv\tau_0$. Therefore, the spherical triangles 012, 023, and 031 are isosceles. Consider one of these triangles, say 012, as illustrated in Fig. 7. The sides 01 and 02 are equal to τ_0. The third side, 12, is equal to τ_{12}. The length of the perpendicular dropped from the point 0 to the side 12 is denoted as η_{12}. This perpendicular intersects the side 12 in a point T_{12} which in the equal-mass case divides the side 12 into two equal parts (each of them equals $\tau_{12}/2$). It also divides the angle φ_{12} into two equal parts. The angles at the vertices 1 and 2 (which are also equal in the equal-mass case) are denoted as κ_{12}. The lengths of two of the sides of a smaller isosceles triangle $01_\xi2_\xi$, 01_ξ and 02_ξ, are $\tau_0\xi$ $(0\leqslant\xi\leqslant1)$. The third side, $1_\xi2_\xi$, is parallel (in the spherical sense) to the side 12 and its length is denoted as $\tau(\xi)$; obviously, $\tau(0)=0$ and $\tau(1)=\tau_{12}$. The perpendicular $0T_{12}$ also splits it into two equal parts of the length $\tau(\xi)/2$. The length of the part of the perpendicular within the triangle $01_\xi2_\xi$ is denoted as $\eta(\xi)$, whereas the angles at the vertices 1_ξ and 2_ξ are $\kappa(\xi)$; obviously, $\eta(0)=0$, $\eta(1)=\eta_{12}$, $\kappa(0)=(\pi-\varphi_{12})/2$, $\kappa(1)=\kappa_{12}$.

The main idea of introducing an auxiliary variable ξ is to simplify the limits of integration. We define the usual spherical integration angle θ in a way that it is the angle between the direction of H_0 and the ''running'' point. For the triangle 012 presented in Fig. 7, let us define that $\varphi=0$ corresponds to the direction of the perpendicular $0T_{12}$. In this way, the limits of the φ integration are $\pm\varphi_{12}/2$. The lower limit of the θ integration is zero, whereas the upper limit depends on φ and varies from τ_0 (at $\varphi=\pm\varphi_{12}/2$) to η_{12} (at $\varphi=0$). If we use, instead of θ, the variable ξ, its limits will be $0\leqslant\xi\leqslant1$, independently of φ. The result of an infinitesimal variation of ξ will be a thin strip based on the side $1_\xi2_\xi$ (cf. Fig. 7). Moreover, all such strips from the triangles 012, 023 and

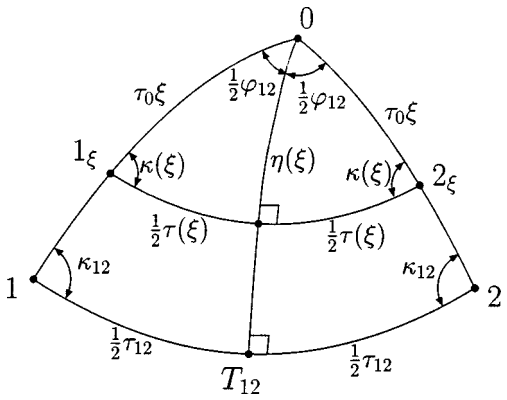

FIG. 7. The isosceles spherical triangle 012.

031 (cf. Fig. 6) can be ''glued'' together. So, the whole result of a variation in ξ is a thin closed spherical triangle. Alternatively, one can perform the azimuthal φ-integration first, and then evaluate the remaining ξ integral. As we shall see, both ways work in the three-point case.

Using formulas of spherical trigonometry (cf. Fig. 7), we get the following relations:

$$\sin(\tfrac{1}{2}\tau(\xi)) = \sin(\tfrac{1}{2}\varphi_{12})\sin(\tau_0\xi),$$

$$\sin\ \eta(\xi) = \sin\ \kappa(\xi)\sin(\tau_0\xi),$$

$$\cos\ \kappa(\xi) = \sin(\tfrac{1}{2}\varphi_{12})\cos\ \eta(\xi),$$

$$\tan\ \kappa(\xi) = (\cos(\tau_0\xi)\tan(\tfrac{1}{2}\varphi_{12}))^{-1},$$

$$\tan\ \eta(\xi) = \cos(\tfrac{1}{2}\varphi_{12})\tan(\tau_0\xi).$$

The situation for arbitrary φ is drawn in Fig. 8. This is a rectangular spherical triangle with an upper vertex 0 and one of the outgoing sides being the perpendicular of the length $\eta(\xi)$, as in Fig. 7. This side corresponds to $\varphi = 0$. The angle at the vertex 0 is equal to φ and a running value of φ is understood. The side opposite to 0 is perpendicular to the one mentioned before, $\eta(\xi)$. This is a part of the corresponding side directed to the point 1_ξ in Fig. 7, i.e. the line of constant ξ. Its length is denoted as $\rho(\xi,\varphi)$. The third side or ''hypotenuse'' has length $\theta(\xi,\varphi)$. The vertex opposite to the $\eta(\xi)$ side is the running point with the coordinates θ, φ. The angle at this vertex is denoted as $\kappa(\xi,\varphi)$.

We want to express θ in terms of ξ and φ. Using the formulas of spherical trigonometry for the triangle in Fig. 8, analogous to ones written above, it is straightforward to show that

$$\tan\ \theta(\xi,\varphi) = \frac{\tan\ \eta(\xi)}{\cos\ \varphi} = \frac{\cos(\tfrac{1}{2}\varphi_{12})}{\cos\ \varphi}\tan(\tau_0\xi). \tag{5.10}$$

The area is obtained by integrating

$$\sin\ \theta\ \ d\theta\ \ d\varphi = -d(\cos\ \theta)d\varphi \Rightarrow -\frac{\partial(\cos\ \theta(\xi,\varphi))}{\partial\xi}\,d\xi\ \ d\varphi, \tag{5.11}$$

where, according to Eq. (5.10),

$$\cos\ \theta(\xi,\varphi) = \left(1 + \frac{\cos^2(\tfrac{1}{2}\varphi_{12})}{\cos^2\ \varphi}\tan^2(\tau_0\xi)\right)^{-1/2}. \tag{5.12}$$

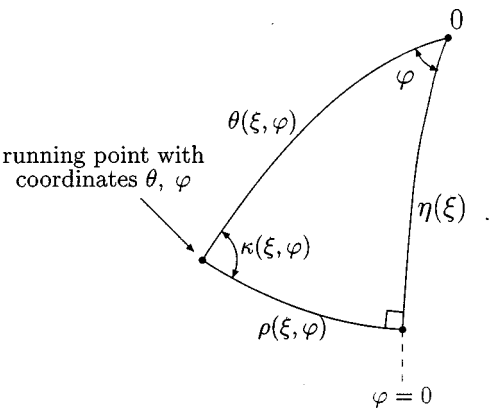

FIG. 8. Integration variables on the sphere.

If we first perform the φ-integral, this effectively corresponds to the area of the infinitesimal strip between two sides of constant ξ,

$$-2d\{\arctan(\cos(\tau_0\xi)\tan(\tfrac{1}{2}\varphi_{12}))\}. \tag{5.13}$$

Then, the remaining ξ integral gives just this arctan. Collecting the results for all triangles (cf. Fig. 6), we get the same as in Eq. (5.5). This can be seen if we observe that the expression (5.13) is nothing but $2d\kappa(\xi)$ and the spherical excess for the triangle $01_\xi 2_\xi$ is $(\varphi_{12}+2\kappa(\xi)-\pi)$. A nice thing about Eq. (5.13) is that in the ξ-integral we can glue all arctan functions into a symmetric answer at the *infinitesimal* level. The same final result can also be obtained when we first integrate over ξ, it being trivial since the integrand is a derivative with respect to ξ.

B. The four-dimensional case

If we consider the four-dimensional three-point function, the only (but very essential!) difference is that we should divide the integrand by $\cos\theta(\xi,\varphi)$, cf. Eq. (3.37). Therefore, the function in the integrand can be written as

$$-\frac{\partial}{\partial\xi}\ln(\cos\,\theta(\xi,\varphi)). \tag{5.14}$$

As in the previous case, one can integrate first over either φ or ξ; both ways lead to equivalent results. Let us integrate first over ξ, this gives

$$\ln\left(\frac{\cos\,\theta(0,\varphi)}{\cos\,\theta(1,\varphi)}\right)=\tfrac{1}{2}\ln\left(1+\frac{\tan^2\,\eta_{12}}{\cos^2\,\varphi}\right). \tag{5.15}$$

In the equal-mass case, the remaining φ-integral becomes

$$\int_0^{\varphi_{12}/2}d\varphi\,\ln\left(1+\frac{\tan^2\,\eta_{12}}{\cos^2\,\varphi}\right)=\frac{1}{2}\int_0^{\varphi_{12}}d\varphi'\left\{\ln(1+2\tan^2\,\eta_{12})+\ln\left(1+\frac{\cos\,\varphi'}{1+2\tan^2\,\eta_{12}}\right)\right.$$
$$\left.-\ln(1+\cos\,\varphi')\right\}, \tag{5.16}$$

where $\varphi'=2\varphi$. Using a result from Ref. 27 (p. 308, Eq. (38)), this integral can be expressed in terms of the Clausen function (see in Appendix A) as

$$\tfrac{1}{2}\tau_{12}\ln\left(\frac{\sin(\frac{1}{2}(\varphi_{12}+\tau_{12}))}{\sin(\frac{1}{2}(\varphi_{12}-\tau_{12}))}\right)+\tfrac{1}{2}\mathrm{Cl}_2(\varphi_{12}+\tau_{12})+\tfrac{1}{2}\mathrm{Cl}_2(\varphi_{12}-\tau_{12})-\mathrm{Cl}_2(\varphi_{12}). \tag{5.17}$$

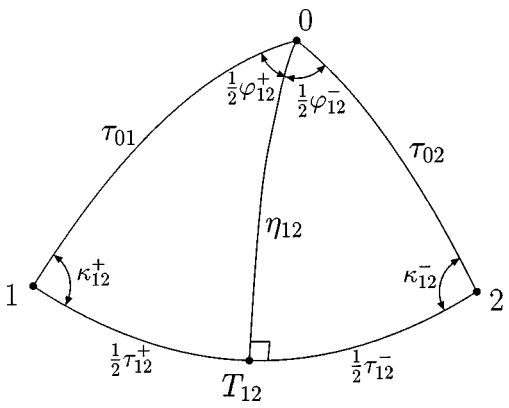

FIG. 9. An asymmetric spherical triangle 012.

The arguments of the Clausen functions have a very transparent geometric interpretation. We note that the logarithmic term can also be presented as

$$\tfrac{1}{2}\tau_{12}\,\ln\!\left(\frac{1+\sin\,\eta_{12}}{1-\sin\,\eta_{12}}\right). \tag{5.18}$$

The function (5.17) can also be expressed in terms of the generalized inverse tangent integral (A8) as

$$\mathrm{Ti}_2(\tan(\tfrac{1}{2}\tau_{12}),\tan(\tfrac{1}{2}\varphi_{12}))-\mathrm{Ti}_2(\tan(\tfrac{1}{2}\tau_{12}),-\tan(\tfrac{1}{2}\varphi_{12})). \tag{5.19}$$

The generalization to the case of unequal masses is easy. To understand why, let us consider Fig. 9. This is again a close-up of the triangle 012 from Fig. 6, but now it is asymmetric, because τ_{01} and τ_{02} are different. We denote the quantities in one of the triangles as $\varphi_{12}^{+}/2$, $\tau_{12}^{+}/2$, $\kappa_{12}^{+}/2$, and in another triangle as $\varphi_{12}^{-}/2$, $\tau_{12}^{-}/2$, $\kappa_{12}^{-}/2$. The height η_{12} is the same for both triangles. Obviously,

$$\tfrac{1}{2}(\varphi_{12}^{+}+\varphi_{12}^{-})=\varphi_{12}\ \text{ and }\ \tfrac{1}{2}(\tau_{12}^{+}+\tau_{12}^{-})=\tau_{12}\,. \tag{5.20}$$

One can immediately see that the result for the integral over this asymmetric triangle is nothing but a half of the sum of the functions (5.17) labeled with plus and with minus. Namely,

$$\begin{aligned}\tfrac{1}{2}\tau_{12}\,\ln\!\left(\frac{1+\sin\,\eta_{12}}{1-\sin\,\eta_{12}}\right)&+\tfrac{1}{4}\mathrm{Cl}_2(\varphi_{12}^{+}+\tau_{12}^{+})+\tfrac{1}{4}\mathrm{Cl}_2(\varphi_{12}^{+}-\tau_{12}^{+})-\tfrac{1}{2}\mathrm{Cl}_2(\varphi_{12}^{+})\\&+\tfrac{1}{4}\mathrm{Cl}_2(\varphi_{12}^{-}+\tau_{12}^{-})+\tfrac{1}{4}\mathrm{Cl}_2(\varphi_{12}^{-}-\tau_{12}^{-})-\tfrac{1}{2}\mathrm{Cl}_2(\varphi_{12}^{-}).\end{aligned} \tag{5.21}$$

Again, the geometric interpretation of the arguments of the Clausen functions is very transparent, thanks to Fig. 9.

Some useful relations are

$$\cos(\tfrac{1}{2}\tau_{12}^{+})=\frac{\cos\,\tau_{01}}{\cos\,\eta_{12}},\ \ \cos(\tfrac{1}{2}\tau_{12}^{-})=\frac{\cos\,\tau_{02}}{\cos\,\eta_{12}},$$

$$\sin(\tfrac{1}{2}\tau_{12}^{+})=\sin\,\tau_{01}\,\sin(\tfrac{1}{2}\varphi_{12}^{+}),\ \ \sin(\tfrac{1}{2}\tau_{12}^{-})=\sin\,\tau_{02}\,\sin(\tfrac{1}{2}\varphi_{12}^{-}),$$

$$\tan(\tfrac{1}{2}\tau_{12}^{+})=\sin\,\eta_{12}\,\tan(\tfrac{1}{2}\varphi_{12}^{+}),\ \ \tan(\tfrac{1}{2}\tau_{12}^{-})=\sin\,\eta_{12}\,\tan(\tfrac{1}{2}\varphi_{12}^{-}).$$

Worth noting is

$$\cos\ \eta_{12}=\frac{m_0\sqrt{k_{12}^2}}{m_1 m_2\ \sin\ \tau_{12}}. \tag{5.22}$$

To get the complete result for the three-point integral in four dimensions, we should sum up the functions (5.21) for all three triangles 012, 023, and 031, and multiply by the overall factor (see Eq. (3.37)) which is

$$-i\pi^2\ \frac{1}{\sqrt{\Lambda^{(3)}}}. \tag{5.23}$$

Note that in this case we get $\sqrt{\Lambda^{(3)}}$ rather than $\sqrt{D^{(3)}}$ in the denominator.

VI. FOUR-POINT FUNCTION

A. The general case

In this case, the four-dimensional simplex has four mass sides and six momentum sides (see Fig. 3). It has five vertices and five three-dimensional hyperfaces. Four of these hyperfaces are the *reduced* ones, corresponding to three-point functions, whereas the fifth one is the momentum hyperface. In fact, this four-dimensional simplex is completely defined by its mass sides m_1, m_2, m_3, m_4 and six ''planar'' angles between them, τ_{12}, τ_{13}, τ_{14}, τ_{23}, τ_{24}, and τ_{34}. According to Eq. (2.16), the content (hypervolume) of this simplex equals

$$V^{(4)}=\frac{1}{24}\ m_1 m_2 m_3 m_4\sqrt{D^{(4)}}, \tag{6.1}$$

with $D^{(4)}$ given by (2.13)–(2.14) at $N=4$,

$$D^{(4)}=\det\|c_{jl}\|,\ \ \|c_{jl}\|=\begin{pmatrix} 1 & c_{12} & c_{13} & c_{14}\\ c_{12} & 1 & c_{23} & c_{24}\\ c_{13} & c_{23} & 1 & c_{34}\\ c_{14} & c_{24} & c_{34} & 1 \end{pmatrix}. \tag{6.2}$$

The four-dimensional four-point function can be exhibited as [cf. Eqs. (3.18) and (3.21)]

$$J^{(4)}(4;1,1,1,1)=\frac{1}{12}\ i\pi^2\ \frac{\Omega^{(4)}}{V^{(4)}}=\frac{2i\pi^2}{m_1 m_2 m_3 m_4}\ \frac{\Omega^{(4)}}{\sqrt{D^{(4)}}}. \tag{6.3}$$

So, the main problem is how to calculate $\Omega^{(4)}$.

In four dimensions, $\Omega^{(4)}$ is the value of the four-dimensional generalization of the solid angle at the vertex of the simplex where all four mass sides meet. In the spherical case, it can be defined as the volume of a part of the unit hypersphere which is cut out from it by the four three-dimensional reduced hyperfaces, each hyperface involving three mass sides of the simplex. This hypersection is a three-dimensional spherical tetrahedron whose six sides (edges) are equal to the angles τ_{jl}. [In the hyperbolic case, this is a hyperbolic tetrahedron whose volume can be obtained by analytic continuation (see below).]

It is illustrated in Fig. 10 where the ith vertex ($i=1,2,3,4$) corresponds to the intersection of the ith mass side of the basic simplex and the unit hypersphere. [It should be stressed that Fig. 10 is just an illustration (rather than a precise picture) since a realistic non-Euclidean tetrahedron should be understood as embedded into four-dimensional Euclidean space.] Furthermore, the dihedral angles of this *three*-dimensional spherical tetrahedron coincide with those of the *four*-dimensional basic simplex, ψ_{jl} (cf. Eqs. (2.24), (2.27)). The dual matrix $\|\tilde{c}_{jl}\|$ as well as its determinant $\tilde{D}^{(4)}$, cf. Eq. (2.26),

$$\|\tilde{c}_{jl}\| = \begin{pmatrix} 1 & \tilde{c}_{12} & \tilde{c}_{13} & \tilde{c}_{14} \\ \tilde{c}_{12} & 1 & \tilde{c}_{23} & \tilde{c}_{24} \\ \tilde{c}_{13} & \tilde{c}_{23} & 1 & \tilde{c}_{34} \\ \tilde{c}_{14} & \tilde{c}_{24} & \tilde{c}_{34} & 1 \end{pmatrix}, \quad \tilde{c}_{jl} = -\cos\,\psi_{jl}, \quad \tilde{D}^{(4)} = \det\|\tilde{c}_{jl}\|, \tag{6.4}$$

are usually referred to as the Gram matrix and determinant of the spherical or hyperbolic tetrahedron. For example, the dihedral angle at the edge 12 of the spherical tetrahedron in Fig. 10 is ψ_{34}, etc. (The digits labeling the edge and the dihedral angle at this edge should cover the whole set 1234.) We note that the dihedral angle is well defined in spherical and hyperbolic spaces, i.e. it remains unchanged along the given edge.

Unfortunately, there are no *simple* relations like (5.4) which might make it possible to express the volume of a spherical (or hyperbolic) tetrahedron in terms of its sides or dihedral angles. In fact, calculation of this volume in an elliptic or hyperbolic space is a well-known problem of non-Euclidean geometry (see e.g. Refs. 36–38). A standard way to solve this problem, say in spherical space, is to split an arbitrary tetrahedron into a set of birectangular ones. For example, the tetrahedron 1234 (shown in Fig. 10) is called *birectangular* (or *double-rectangular*) if (i) the edge 12 is perpendicular to the face 234 and (ii) the opposite edge, 34, is perpendicular to the face 123. It is easy to check that in this case three dihedral angles, ψ_{13}, ψ_{14}, and ψ_{24}, are right angles. The other three are usually denoted[39,40] as

$$\psi_{12} = \alpha, \quad \psi_{23} = \beta, \quad \psi_{34} = \gamma. \tag{6.5}$$

The volume of a birectangular tetrahedron is known (see below) and can be expressed in terms of Lobachevsky or Schläfli functions which can be related to dilogarithms (see Refs. 39 and 40).

The volume of the general tetrahedron must be a symmetrical function of the six edges τ_{jl}, or equivalently of the six dihedral angles ψ_{jl}. When we break it up into a sum of birectangular tetrahedra, this explicit symmetry may be lost, although of course it must be hidden in the properties of the sums of dilogarithmic functions.

The general way to split an arbitrary tetrahedron into a sum of birectangular ones is to fix a point 0, say inside the tetrahedron. [If this point (or the feet of some of the perpendiculars) happen to be outside the tetrahedron, this would mean that some of the volumes of the resulting birectangular tetrahedra should be taken with an opposite sign.] Then, connecting this point 0 with each of the vertices of the tetrahedron 1234, we split it into four smaller tetrahedra: 0123, 0124, 0134, and 0234. Consider one of these tetrahedra, say 0123. Dropping the perpendicular from 0 onto the face 123 and connecting the foot of this perpendicular F_{123} with the vertices 1, 2, and 3, we split 0123 into three *rectangular* tetrahedra, $012F_{123}$, $013F_{123}$, and $023F_{123}$. Then, each of these

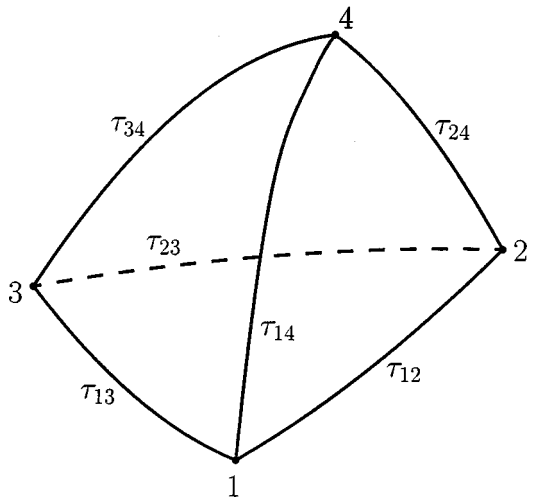

FIG. 10. The spherical tetrahedron 1234.

rectangular tetrahedra (say, $012F_{123}$) can be split into a sum of two *birectangular* tetrahedra by dropping a perpendicular from F_{123} onto the only side belonging to the original tetrahedron 1234 (the edge 12). In this way, we get $4\times3\times2=24$ birectangular tetrahedra.

Since we are free to choose the position of the point 0, there are (at least) two ways to reduce the number of terms involved in this volume splitting in the general case:

(i) The point 0 may coincide with one of the vertices of the original tetrahedron, say with the vertex 4 [cf. Fig. 11(a)]. In this case, we do not need to split into four smaller tetrahedra and can start just by dropping a perpendicular from the point 4 onto the face 123. In this way, we reduce the number of birectangular tetrahedra involved to six. However, the price we pay for this is the loss of explicit symmetry.

(ii) A symmetric choice of the point 0 can be related to the structure of the basic four-dimensional Euclidean simplex. In particular, we can take as the point 0 the intersection of the height H_0 and the unit hypersphere. Moreover, since the general four-point function can be reduced to the equal-mass case (see Eqs. (3.10)–(3.11)), it is enough to consider the case when the point 0 is equidistant from the vertices 1, 2, 3, and 4. In this case, at the first stage of splitting we get four ''*isosceles*'' tetrahedra, signifying for each of them the three sides meeting at the point 0 are equal. Continuing the splitting, we see that due to ''isoscelesness'' the pairs of the tetrahedra produced at the last stage have equal values. Effectively, we get the sum of $4\times3=12$ volumes of birectangular tetrahedra. Although this is twice as much as in the case (i), the advantage is that we are able to preserve the explicit symmetry at all stages.

We note that the possibility to rescale the Feynman parameters in the case $n=N$ is closely connected with the freedom to choose the point 0 as we like. [See Eqs. (3.27)–(3.28). For $n=N$, the denominator disappears. In general, the f_i can take arbitrary positive values.] As we have seen, this rescaling preserves the angles τ_{jl} between the mass vectors, whereas the direction of the height H_0 varies. In particular, we can make it coincide with one of the mass vectors, which corresponds to the splitting (i) considered above. From this point of view, we lose the explicit symmetry since the corresponding rescaling is not symmetric.

Another possibility to reduce the number of birectangular volumes involved is to try to choose the point 0 in such a way that the volumes of the four tetrahedra (produced at the first stage of splitting) are equal (or proportional) to each other. In this case, the further splitting would give just $3\times2=6$ birectangular terms, i.e., the same number as in the case (i); but here we may preserve the explicit symmetry. However, it seems to be difficult (if possible at all) to define the position of the point 0 in this case, because one needs to solve a set of transcendental equations.

In some special examples described below we show that an additional symmetry of the diagram can also be employed.

B. Volume of birectangular tetrahedra in curved space

As we have seen (cf. Eq. (6.5)), the birectangular tetrahedron may be specified by three dihedral angles $\psi_{12}=\alpha$, $\psi_{23}=\beta$, $\psi_{34}=\gamma$, so that the corresponding Gram matrix $\|\tilde{c}_{jl}\|$ and its determinant (cf. Eq. (20) of Ref. 40) are

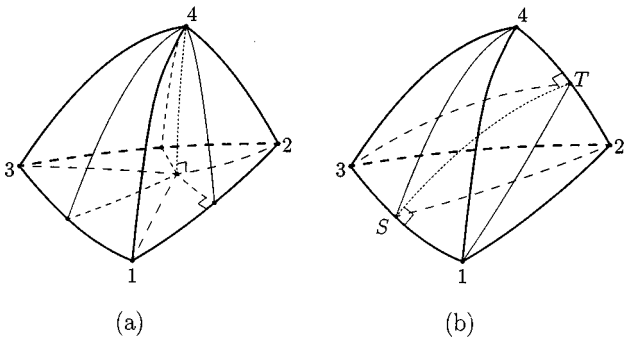

FIG. 11. Different ways of splitting the non-Euclidean tetrahedron.

$$\|\tilde{c}_{jl}\|=\begin{pmatrix} 1 & -\cos\alpha & 0 & 0 \\ -\cos\alpha & 1 & -\cos\beta & 0 \\ 0 & -\cos\beta & 1 & -\cos\gamma \\ 0 & 0 & -\cos\gamma & 1 \end{pmatrix}, \quad \tilde{D}^{(4)}=\sin^2\alpha\,\sin^2\gamma-\cos^2\beta. \tag{6.6}$$

The volume of this birectangular tetrahedron (i.e., its contribution to $\Omega^{(4)}$) can be presented as

$$\Omega^{(4)}=V(\alpha,\beta,\gamma)=\tfrac{1}{4}S(\pi/2-\alpha,\beta,\pi/2-\gamma). \tag{6.7}$$

The notation $S(\alpha,\beta,\gamma)$ corresponds to the Schläfli function as defined in Ref. 39, whereas the notation $V(\alpha,\beta,\gamma)$ was used in Ref. 40 and is proportional to the function originally used by Schläfli.[37] Note that function (6.7) is symmetric with respect to α and γ.

If we define (cf. Eq. (3.21) of Ref. 39)

$$X\equiv\frac{\cos\alpha\,\cos\gamma-\sqrt{\tilde{D}^{(4)}}}{\cos\alpha\,\cos\gamma+\sqrt{\tilde{D}^{(4)}}}, \tag{6.8}$$

the result for the volume can be expressed in terms of dilogarithms,[39]

$$\begin{aligned} 4V(\alpha,\beta,\gamma)=&\tfrac{1}{2}[\mathrm{Li}_2(Xe^{2i\alpha})+\mathrm{Li}_2(Xe^{-2i\alpha})]+\tfrac{1}{2}[\mathrm{Li}_2(Xe^{2i\gamma})+\mathrm{Li}_2(Xe^{-2i\gamma})] \\ &-\tfrac{1}{2}[\mathrm{Li}_2(-Xe^{2i\beta})+\mathrm{Li}_2(-Xe^{-2i\beta})]-\mathrm{Li}_2(-X)-\left(\frac{\pi}{2}-\alpha\right)^2-\left(\frac{\pi}{2}-\gamma\right)^2+\beta^2. \end{aligned} \tag{6.9}$$

When the condition $\tilde{D}^{(4)}\geqslant 0$ ($\sin\alpha\sin\gamma\leqslant\cos\beta$) is obeyed, the tetrahedron exists in the usual geometrical sense, and its volume is real. In this case, one can express $V(\alpha,\beta,\gamma)$ in terms of the real function $\mathrm{Li}_2(r,\theta)\equiv\mathrm{Re}[\mathrm{Li}_2(re^{i\theta})]$ (see Appendix A). However, one should distinguish between the cases $X\geqslant 0$ (or $\cos^2\alpha+\cos^2\beta+\cos^2\gamma\geqslant 1$) and $X\leqslant 0$ (or $\cos^2\alpha+\cos^2\beta+\cos^2\gamma\leqslant 1$). The corresponding expressions are

$$\begin{aligned} 4V(\alpha,\beta,\gamma)\big|_{X\geqslant 0}=&\,\mathrm{Li}_2(X,2\alpha)+\mathrm{Li}_2(X,2\gamma)-\mathrm{Li}_2(X,\pi-2\beta)-\mathrm{Li}_2(-X) \\ &-\left(\frac{\pi}{2}-\alpha\right)^2-\left(\frac{\pi}{2}-\gamma\right)^2+\beta^2, \end{aligned} \tag{6.10}$$

$$\begin{aligned} 4V(\alpha,\beta,\gamma)\big|_{X\leqslant 0}=&\,\mathrm{Li}_2(-X,\pi-2\alpha)+\mathrm{Li}_2(-X,\pi-2\gamma)-\mathrm{Li}_2(-X,2\beta)-\mathrm{Li}_2(-X) \\ &-\left(\frac{\pi}{2}-\alpha\right)^2-\left(\frac{\pi}{2}-\gamma\right)^2+\beta^2. \end{aligned} \tag{6.11}$$

In particular, when $X=-1$ we take into account that $\mathrm{Li}_2(1,\theta)=\frac{1}{4}(\pi-\theta)^2-\frac{1}{12}\pi^2$ and get

$$V(\alpha,\beta,\gamma)\big|_{X=-1}=\tfrac{1}{4}\pi(\alpha+\beta+\gamma-\pi). \tag{6.12}$$

When $\alpha=\beta=\gamma=\frac{1}{2}\pi$, this gives $\frac{1}{8}\pi^2$, which agrees with Eq. (3.14) (at $N=4$, $\nu_i=1$).

When $\tilde{D}^{(4)}<0$ ($\sin\alpha\sin\gamma<\cos\beta$), the tetrahedron does not exist in the usual geometrical sense (in the spherical case), but its volume (more precisely, the analytic continuation of the function corresponding to the volume) can be obtained from Eq. (6.9), and it is imaginary. In this case we define (according to Refs. 39, 40)

$$X=e^{-2i\delta},\ \text{so that}\ \tan\delta=\frac{\sqrt{-\tilde{D}^{(4)}}}{\cos\alpha\,\cos\gamma}, \tag{6.13}$$

and we get

$$V(\alpha,\beta,\gamma)=(1/(8i))[\mathrm{Cl}_2(2\alpha+2\delta)-\mathrm{Cl}_2(2\alpha-2\delta)+\mathrm{Cl}_2(2\gamma+2\delta)$$
$$-\mathrm{Cl}_2(2\gamma-2\delta)-\mathrm{Cl}_2(\pi-2\beta+2\delta)+\mathrm{Cl}_2(\pi-2\beta-2\delta)+2\mathrm{Cl}_2(\pi-2\delta)]. \quad (6.14)$$

Using the connection between the Clausen function and the Lobachevsky function $L(\theta)$ (cf. Eq. (A10) in Appendix A), Eq. (6.14) can be rewritten in terms of the Lobachevsky function. [The function used in Ref. 40 which was denoted by a Cyrillic ''L'' can be expressed in terms of the Clausen function as $\frac{1}{2}\mathrm{Cl}_2(2\theta)$.] The same expression (6.14) gives the volume of a birectangular tetrahedron in the hyperbolic space (see Refs. 39 and 40).

C. Some cases with an additional symmetry

There are a few cases of particular practical interest because of their inherent higher symmetries. In particular, there is the most symmetrical case where all $c_{ij}=c$. Thus all dihedral angles of the full (regular) tetrahedron are equal, to ψ say. In that case the six birectangular tetrahedra corresponding to splitting (i) are all of equal size and $\Omega^{(4)}$ reduces to

$$\Omega^{(4)}=6V(\pi/3,\psi/2,\psi)=\tfrac{3}{2}S(\pi/6,\psi/2,\pi/2-\psi). \quad (6.15)$$

Coxeter gives the identities (Eqs. (4.31)–(4.32) of Ref. 39)

$$S(\pi/6,\psi/2,\pi/2-\psi)=4S(\pi/6,\pi/3,(\pi-\psi)/2)=\tfrac{2}{3}S((\pi-\psi)/2,\psi,(\pi-\psi)/2), \quad (6.16)$$

which he derived algebraically, and which can be used to convert $\Omega^{(4)}$ given by Eq. (6.7) into other forms. We will presently explain the geometrical meaning of these identities.

Another case of interest corresponds to equal mass scattering with equal internal masses; this means $c_{12}=c_{23}=c_{34}=c_{14}$ and in general are not equal to c_{13},c_{24}, which are associated with energy and scattering angle variables. This case has four of the tetrahedral sides equal and different from the other two (which are opposite to one another). Thus there are three distinct c's. We shall see that it is possible here to break up full tetrahedron into four birectangular ones by joining the midpoints of the unequal sides to render the problem more tractable and make the maximum use of what symmetry exists.

We shall carry out our analysis by relaxing the symmetry relations. First we consider the case where all six c_{ij} are equal to c and treat this in three different ways. Then we let $c_{13}\equiv c_s$ and $c_{24}\equiv c_t$ be unequal.

1. Completely symmetrical box diagram

Here all c_{jl} are equal and we are dealing with a completely regular tetrahedron. For example, in the equal-mass case this would mean that all k_{jl}^2 are equal (of course this represents an unphysical point in the Mandelstam diagram), to k^2 say, and all $c=1-k^2/(2m^2)$. We have

$$\|c_{jl}\|=\begin{pmatrix}1&c&c&c\\c&1&c&c\\c&c&1&c\\c&c&c&1\end{pmatrix},\quad \|\tilde{c}_{jl}\|=\begin{pmatrix}1&\tilde{c}&\tilde{c}&\tilde{c}\\\tilde{c}&1&\tilde{c}&\tilde{c}\\\tilde{c}&\tilde{c}&1&\tilde{c}\\\tilde{c}&\tilde{c}&\tilde{c}&1\end{pmatrix},\quad \tilde{c}=-\frac{c}{1+2c}. \quad (6.17)$$

This means that the dihedral angle ψ between each and every pair of hyperplanes equals $\arccos[c/(1+2c)]$. {The generalization of this totally symmetric case to an N-point function is quite easy and leads to $\cos\psi=c/[1+(N-2)c]$.}

We can evaluate the hypervolume in at least three ways (the first two correspond to the ways (i) and (ii) considered in the general case):

(i) By dropping a perpendicular from one corner onto the opposite hyperplane and another perpendicular onto one of the opposite sides, we subdivide the full tetrahedron into six equal birectangular tetrahedra [see Fig. 11(a)]. By inspection, each birectangular tetrahe-

dron has dihedral angles ψ, $\psi/2$, and $\pi/3$. Therefore the total hypervolume equals

$$\Omega^{(4)}=6V(\psi,\psi/2,\pi/3)=\tfrac{3}{2}S(\pi/6,\psi/2,\pi/2-\psi). \tag{6.18}$$

(ii) We can alternatively consider the total volume as four times the volume contained in the tetrahedron generated by one base triangle and the ''center of mass'' point 0: $P_0=\frac{1}{4}(P_1+P_2+P_3+P_4)$ (here, we use the notation $P_i=m_ia_i$, cf. Eq. (2.49)). In this case the matrix associated with the tetrahedron 0123 is found to be

$$\|c_{jl}\|=\begin{pmatrix} 1 & c' & c' & c' \\ c' & 1 & c & c \\ c' & c & 1 & c \\ c' & c & c & 1 \end{pmatrix},\quad c'=\frac{\sqrt{1+3c}}{2}. \tag{6.19}$$

leading to

$$\|\tilde{c}_{jl}\|=\begin{pmatrix} 1 & \tilde{c}' & \tilde{c}' & \tilde{c}' \\ \tilde{c}' & 1 & 1/2 & 1/2 \\ \tilde{c}' & 1/2 & 1 & 1/2 \\ \tilde{c}' & 1/2 & 1/2 & 1 \end{pmatrix},\quad \tilde{c}'=-\sqrt{\frac{1+3c}{2(1+2c)}}. \tag{6.20}$$

This implies that there are two distinct dihedral angles,

$$\tfrac{2}{3}\pi \ \text{ and } \ \arccos\sqrt{\frac{1+3c}{2(1+2c)}}=\tfrac{1}{2}\arccos\left(\frac{c}{1+2c}\right)=\frac{\psi}{2} \tag{6.21}$$

in fact. Actually the result is hardly surprising since it can be deduced by simple inspection of the geometrical figure. Anyhow if we now divide 0123 into six identical birectangular tetrahedra by dropping a perpendicular from 0 onto the center of the triangle 123, these smaller tetrahedra possess dihedral angles $\psi/2$, $\pi/3$, $\pi/3$. Consequently we have an alternative geometrical expression for the full volume,

$$\Omega^{(4)}=24V(\pi/3,\pi/3,\psi/2)=6S(\pi/6,\pi/3,(\pi-\psi)/2). \tag{6.22}$$

Comparing this with the previous answer we have a geometrical explanation of one of the Coxeter identities (6.16). Furthermore, because (see Ref. 39, p. 18) there exists the differential relation

$$dS(\pi/6,\pi/3,\gamma)=-2\arccos\left(\frac{\cos\gamma}{\sqrt{4\cos^2\gamma-1}}\right)d\gamma, \tag{6.23}$$

we get the following integral representation of the four-dimensional solid angle (6.22):

$$\Omega^{(4)}=\pi^2-12\int_0^{(\pi-\psi)/2}d\gamma\,\arccos\left(\frac{\cos\gamma}{\sqrt{4\cos^2\gamma-1}}\right). \tag{6.24}$$

(iii) We take the two midpoints S and T of the opposite sides: $P_S=(P_1+P_3)/2$, $P_T=(P_2+P_4)/2$ [see Fig. 11(b)]. The full volume is then just four times that of the tetrahedron $12ST$, which is already birectangular as can be seen by inspection. In this case the matrix $\|c_{jl}\|$ associated with the tetrahedron $12ST$ has the following nondiagonal elements:

$$c_{12}=c,\quad c_{13}=c_{24}=\sqrt{\frac{1+c}{2}},\quad c_{14}=c_{23}=c\sqrt{\frac{2}{1+c}},\quad c_{34}=\frac{2c}{1+c}. \tag{6.25}$$

The corresponding Gram matrix of the tetrahedron $12ST$ is

$$\|\tilde{c}_{jl}\|=\begin{pmatrix} 1 & 0 & \tilde{c}' & 0 \\ 0 & 1 & 0 & \tilde{c}' \\ \tilde{c}' & 0 & 1 & \tilde{c} \\ 0 & \tilde{c}' & \tilde{c} & 1 \end{pmatrix}, \tag{6.26}$$

with $\tilde{c}$ and $\tilde{c}'$ defined in Eqs. (6.17) and (6.20), respectively. From this we extract the three dihedral angles.

$$\alpha=\gamma=\tfrac{1}{2}\psi \quad \text{and} \quad \beta=\psi, \tag{6.27}$$

where ψ is nothing but the dihedral angle at the side 12, i.e., one of the dihedral angles of the full tetrahedron. This accords completely with the shape of $12ST$. Hence the total volume of the complete tetrahedron 1234 is given by

$$\Omega^{(4)}=4V(\psi/2,\psi,\psi/2)=S((\pi-\psi)/2,\psi,(\pi-\psi)/2). \tag{6.28}$$

Comparing with the earlier expression, this provides a geometric meaning of the second Coxeter identity (6.16).

2. Partially symmetric box diagram

In this case we shall take four sides of the tetrahedron to be equal and two opposite sides to be unequal to them. In the equal-mass case, this would correspond to setting $k_{12}^2=k_{23}^2=k_{34}^2=k_{14}^2=k^2$, $k_{13}^2=s$, $k_{24}^2=t$, and we would get three different c's: $c=1-k^2/2m^2$, $c_s=1-s/2m^2$, $c_t=1-t/2m^2$ (s and t are just the Mandelstam variables). In this case much the simplest way of evaluating the volume of the tetrahedron is to break it up into four equal birectangular tetrahedra by taking the midpoints of the sides 13 and 24, as before; $P_S=(P_1+P_3)/2$, $P_T=(P_2+P_4)/2$. There is enough symmetry in the problem that the full tetrahedron is four times the birectangular tetrahedron $12ST$. The latter is described by the matrix $\|c_{jl}\|$ with the nondiagonal elements,

$$c_{12}=c, \quad c_{13}=\sqrt{\frac{1+c_s}{2}}, \quad c_{24}=\sqrt{\frac{1+c_t}{2}},$$

$$c_{23}=c\sqrt{\frac{2}{1+c_s}}, \quad c_{14}=c\sqrt{\frac{2}{1+c_t}}, \quad c_{34}=\frac{2c}{\sqrt{(1+c_s)(1+c_t)}}. \tag{6.29}$$

The nondiagonal elements of the ''dual'' matrix $\|\tilde{c}_{jl}\|$ are

$$\tilde{c}_{13}=-\sqrt{\frac{(1+c_s)(1+c_t)-4c^2}{2(1+c_t-2c^2)}}, \quad \tilde{c}_{24}=-\sqrt{\frac{(1+c_s)(1+c_t)-4c^2}{2(1+c_s-2c^2)}},$$

$$\tilde{c}_{34}=-c\sqrt{\frac{(1-c_s)(1-c_t)}{(1+c_s-2c^2)(1+c_t-2c^2)}}, \quad \tilde{c}_{12}=\tilde{c}_{14}=\tilde{c}_{23}=0. \tag{6.30}$$

We can now read off the three nontrivial dihedral angles,

$$\alpha=\arccos\left(\sqrt{\frac{(1+c_s)(1+c_t)-4c^2}{2(1+c_t-2c^2)}}\right), \quad \gamma=\arccos\left(\sqrt{\frac{(1+c_s)(1+c_t)-4c^2}{2(1+c_s-2c^2)}}\right),$$

$$\beta=\arccos\left(c\sqrt{\frac{(1-c_s)(1-c_t)}{(1+c_s-2c^2)(1+c_t-2c^2)}}\right). \tag{6.31}$$

Finally then the required volume of the full tetrahedron 1234 is obtained from

$$\Omega^{(4)}=4V(\alpha,\beta,\gamma)=S(\pi/2-\alpha,\beta,\pi/2-\gamma). \tag{6.32}$$

A well-known physical example of partially symmetric box diagram is related to the photon–photon scattering.[41] In this case, $k^2=0$, $c=1$, $\cos\beta=1$, $\beta=0$. Furthermore, $\widetilde{D}^{(4)}=\sin^2\alpha\sin^2\gamma-1$ and we should use formula (6.14) for $V(\alpha,0,\gamma)$.

D. The massless limit

In the case when all masses m_i vanish, the quantities c_{jl} become infinite (cf. Eq. (2.4)) and should be be considered as hyperbolic cosines. The determinant $D^{(4)}$, Eq. (6.2), has the following limit:

$$\begin{aligned}(m_1^2m_2^2m_3^2m_4^2D^{(4)})|_{m_i\to0}\Rightarrow{}&\tfrac{1}{16}[(k_{12}^2k_{34}^2)^2+(k_{13}^2k_{24}^2)^2+(k_{14}^2k_{23}^2)^2\\&-2k_{12}^2k_{34}^2k_{13}^2k_{24}^2-2k_{12}^2k_{34}^2k_{14}^2k_{23}^2-2k_{13}^2k_{24}^2k_{14}^2k_{23}^2]\\&=\tfrac{1}{16}\lambda(k_{12}^2k_{34}^2,k_{13}^2k_{24}^2,k_{14}^2k_{23}^2)\end{aligned} \tag{6.33}$$

with $\lambda(x,y,z)$ defined by Eq. (1.1).

The elements of the dual matrix $\|\widetilde{c}_{jl}\|$ still can be interpreted as cosines of the dihedral angles, namely,

$$\widetilde{c}_{12}=\widetilde{c}_{34}=-\cos\psi_{34}=-\cos\psi_{12}=\frac{k_{13}^2k_{24}^2+k_{14}^2k_{23}^2-k_{12}^2k_{34}^2}{\sqrt{k_{13}^2k_{24}^2k_{14}^2k_{23}^2}}, \tag{6.34}$$

$$\widetilde{c}_{13}=\widetilde{c}_{24}=-\cos\psi_{24}=-\cos\psi_{13}=\frac{k_{14}^2k_{23}^2+k_{12}^2k_{34}^2-k_{13}^2k_{24}^2}{\sqrt{k_{14}^2k_{23}^2k_{12}^2k_{34}^2}}, \tag{6.35}$$

$$\widetilde{c}_{14}=\widetilde{c}_{23}=-\cos\psi_{23}=-\cos\psi_{14}=\frac{k_{12}^2k_{34}^2+k_{13}^2k_{24}^2-k_{14}^2k_{23}^2}{\sqrt{k_{12}^2k_{34}^2k_{13}^2k_{24}^2}}. \tag{6.36}$$

Therefore, in this situation we get nothing but an *ideal* hyperbolic tetrahedron (see pp. 39–40 of Ref. 40), i.e., the tetrahedron whose vertices are all at infinity. The pairs of opposite dihedral angles of the ideal tetrahedron are equal, $\psi_{12}=\psi_{34}$, $\psi_{13}=\psi_{24}$, $\psi_{14}=\psi_{23}$, whereas its volume is given by

$$\Omega^{(4)}=\frac{1}{2i}\,[\mathrm{Cl}_2(2\psi_{12})+\mathrm{Cl}_2(2\psi_{13})+\mathrm{Cl}_2(2\psi_{23})],\quad \psi_{12}+\psi_{13}+\psi_{23}=\pi, \tag{6.37}$$

according to Eq. (41) of Ref. 40 (see also Ref. 42).

First of all we see that Eqs. (6.33)–(6.36) depend only on the products $k_{12}^2k_{34}^2$, $k_{13}^2k_{24}^2$, and $k_{14}^2k_{23}^2$. This corresponds to the ''glueing'' of arguments in the massless case observed in Ref. 43. Moreover, the result (6.37) is of the same form as the massless three-point function in four dimensions (1.2). This accords completely with the reduction of the massless four-point function to the three-point function which has been proven in Ref. 43.

VII. REDUCTION OF INTEGRALS WITH $N>n$ AND RELATED PROBLEMS

A. Integrals with $N>n$

Consider Eq. (3.45) in the case when $n=N-1$,

$$J^{(N)}(N-1;1,\ldots,1)=\frac{1}{\Lambda^{(N)}}\left(\prod\ m_i^2\right)\sum_{i=1}^{N}\ \frac{1}{m_i^2}\,F_i^{(N)}J_i^{(N)}(N-1;1,\ldots,1). \tag{7.1}$$

The integrals on the right-hand side, $J_i^{(N)}(N-1;1,...,1)$, correspond to splitting the basic N-dimensional simplex into N rectangular ones. The ith rectangular simplex is obtained via substituting the ith mass side of the basic simplex by its height H_0 whose length is m_0. Therefore, the corresponding elements of the $\|c\|$ matrix (2.13) should be replaced by (cf. Eqs. (3.43)–(3.44))

$$c_{ji}\to\cos\ \tau_{0j}=\frac{m_0}{m_j},\quad c_{il}\to\cos\ \tau_{0l}=\frac{m_0}{m_l}. \tag{7.2}$$

At this point let us consider what the representation (3.8) gives for $J_i^{(N)}$. The α-integrand becomes

$$\left(\sum_{l\neq i}\ \frac{\alpha_l}{m_l}+\frac{\alpha_i}{m_0}\right)\delta\left((\alpha^T\|c\|\alpha)\Big|_{\alpha_i=0}+2m_0\alpha_i\sum_{l\neq i}\ \frac{\alpha_l}{m_l}+\alpha_i^2-1\right). \tag{7.3}$$

One can see that the first factor in (7.3) is proportional to the derivative of the argument of the δ function with respect to α_i. Therefore, Eq. (7.3) can be expressed as

$$\frac{1}{2m_0}\frac{\partial}{\partial\alpha_i}\ \theta\left((\alpha^T\|c\|\alpha)\Big|_{\alpha_i=0}+2m_0\alpha_i\sum_{l\neq i}\ \frac{\alpha_l}{m_l}+\alpha_i^2-1\right). \tag{7.4}$$

Integrating over α_i, we get

$$J_i^{(N)}(N-1;1,...,1)=\frac{i^{1-2N}\pi^{(N-1)/2}}{m_0^2\Pi_{l\neq i}m_l}\ \Gamma\left(\frac{N+1}{2}\right)\int_0^\infty\cdots\int_0^\infty\prod_{l\neq i}\ d\alpha_l\theta(1-(\alpha^T\|c\|\alpha)|_{\alpha_i=0}). \tag{7.5}$$

Inserting

$$\frac{1}{N-1}\sum_{l\neq i}\ \frac{\partial\alpha_l}{\partial\alpha_l}=1 \tag{7.6}$$

and integrating by parts, the integrand of (7.5) can be transformed as

$$\begin{aligned}\theta(1-(\alpha^T\|c\|\alpha)|_{\alpha_i=0})&\Rightarrow-\frac{1}{N-1}\left(\sum_{l\neq i}\ \alpha_l\frac{\partial}{\partial\alpha_l}\right)\theta(1-(\alpha^T\|c\|\alpha)|_{\alpha_i=0})\\&\Rightarrow\frac{1}{N-1}\ \delta(1-(\alpha^T\|c\|\alpha)|_{\alpha_i=0})\sum_{l\neq i}\ \alpha_l\frac{\partial}{\partial\alpha_l}(\alpha^T\|c\|\alpha)|_{\alpha_i=0}\\&\Rightarrow\frac{2}{N-1}\ \delta((\alpha^T\|c\|\alpha)|_{\alpha_i=0}-1).\end{aligned} \tag{7.7}$$

Using Eq. (3.8), we arrive at

$$J_i^{(N)}(N-1;1,...,1)=-\frac{1}{2m_0^2}\ J^{(N-1)}(N-1;1,...,1)|_{\text{without }i}. \tag{7.8}$$

Therefore, we get

$$J^{(N)}(N-1;1,...,1)=-\frac{1}{2D^{(N)}}\sum_{i=1}^{N}\ \frac{1}{m_i^2}\ F_i^{(N)}J^{(N-1)}(N-1;1,...,1)|_{\text{without }i}. \tag{7.9}$$

This corresponds to Eq. (47) of Ref. 11, when the number of external legs N is equal to the space–time dimension plus one. For $N=3,4,5$, Eq. (7.9) reproduces Eqs. (10), (8), (13) from Ref. 11, respectively. (For $N=3$, the explicit result was presented in Ref. 10.)

Similar reduction formulas have also been considered in Refs. 8–13. In Ref. 14 it was shown how the linear dependence of external momenta in the case $N\geqslant 6$ and for $n=4$ may be employed

to reduce the N-point function to a set of $(N-1)$-point ones. A method for expressing the five-point function in four dimensions in terms of a sum of four-point functions was first shown in Ref. 7, whereas the reduction of the three-point function in two dimensions was studied in Ref. 10 (see also Ref. 18). These results were generalized in Refs. 8 and 9. For example, in Ref. 8 (see also Ref. 10) the representation for the coefficients corresponding to $F_l^{(N)}$ (Eq. (7) of Ref. 8) is given in terms of the propagators whose momenta are solutions to a certain set of equations corresponding to the other propagators. Moreover, it is possible to make a direct as opposed to step-by-step reduction to n-point functions for any $N>n$. More recently, some other approaches to this problem were suggested in Refs. 12 and 13.

It is also instructive to consider another derivation of Eq. (7.9). The representation (3.8) yields

$$J^{(N)}(N-1;1,...,1)=\frac{2i^{1-2N}\pi^{(N-1)/2}}{\Pi m_l}\,\Gamma\left(\frac{N+1}{2}\right)\int_0^\infty\cdots\int_0^\infty\prod\,d\alpha_l\left(\sum_{i=1}^N\,\frac{\alpha_i}{m_i}\right)\delta(\alpha^T\|c\|\alpha-1). \tag{7.10}$$

Using Eq. (2.33) we get

$$\sum_{i=1}^N\,\frac{\alpha_i}{m_i}=\frac{1}{D^{(N)}}\sum_{i=1}^N\,\alpha_i\sum_{l=1}^N\,\frac{c_{il}}{m_l}\,F_l^{(N)}=\frac{1}{D^{(N)}}\sum_{i=1}^N\,\frac{F_i^{(N)}}{m_i}\sum_{l=1}^N\,c_{il}\alpha_l\,. \tag{7.11}$$

Repeating steps (7.3)–(7.7) for the resulting α-integrals, we arrive at the following result:

$$\frac{2i^{1-2N}\pi^{(N-1)/2}}{\Pi m_l}\,\Gamma\left(\frac{N+1}{2}\right)\int_0^\infty\cdots\int_0^\infty\prod\,d\alpha_l\left(\sum_{l=1}^N\,c_{il}\alpha_l\right)\delta(\alpha^T\|c\|\alpha-1)$$

$$=-\frac{1}{2m_i}\,J^{(N-1)}(N-1;1,...,1)\big|_{\text{without }i}\,. \tag{7.12}$$

Substituting (7.11) and (7.12) into (7.10), we obtain once again Eq. (7.9).

A similar approach can also be used to obtain formulas connecting the integrals with different values of the space–time dimension.[13,44,45]

B. Integrals with higher powers of propagators

Equations like (7.12) can also be used for calculating integrals with higher powers of propagators. Such integrals can be associated with derivatives with respect to m_i^2, provided that m_i and k_{jl}^2 are considered as the set of independent variables. Specifically, consider the sum

$$c_{1i}m_1J^{(N)}(N+1;2,1,1,...,1)+c_{2i}m_2J^{(N)}(N+1;1,2,1,...,1)$$

$$+\cdots+c_{iN}m_NJ^{(N)}(N+1;1,...,1,2)\equiv\sum_{l=1}^N\,c_{il}m_lJ^{(N)}(N+1;\{1+\delta_{lj}\}). \tag{7.13}$$

(The notation for the set of arguments $\{\nu_j\}=\{1+\delta_{lj}\}$ means that all powers of denominators are equal to one, except for the ν_l which is equal to two. As usually, it is implied that $c_{ji}=c_{ij}$ and $c_{ii}=1$.) Employing the representation (3.8) for all integrals in (7.13), we obtain the same integral as in Eq. (7.12), multiplied by $(-\pi)$. Therefore, we get a system of equations which can be displayed in the matrix form,

$$\|c_{jl}\|\begin{pmatrix} m_1J^{(N)}(N+1;2,1,1,...,1)\\ m_2J^{(N)}(N+1;1,2,1,...,1)\\ \cdots\\ m_NJ^{(N)}(N+1;1,1,...,1,2)\end{pmatrix}=\frac{\pi}{2}\begin{pmatrix} m_1^{-1}J^{(N-1)}(N-1;1,...,1)\big|_{\text{without }1}\\ m_2^{-1}J^{(N-1)}(N-1;1,...,1)\big|_{\text{without }2}\\ \cdots\\ m_N^{-1}J^{(N\text{-}1)}(N-1;1,...,1)\big|_{\text{without }N}\end{pmatrix}. \tag{7.14}$$

The matrix inverse to $\|c_{jl}\|$ is given by Eq. (2.25). Applying $\|c_{jl}\|^{-1}$ to both sides, we get

$$\begin{pmatrix} (m_1/\sqrt{D_{11}^{(N-1)}})J^{(N)}(N+1;2,1,1,...,1) \\ (m_2/\sqrt{D_{22}^{(N-1)}})J^{(N)}(N+1;1,2,1,...,1) \\ \cdots \\ (m_N/\sqrt{D_{NN}^{(N-1)}})J^{(N)}(N+1;1,1,...,1,2) \end{pmatrix}$$

$$= \frac{\pi}{2D^{(N)}} \|\widetilde{c}_{jl}\| \begin{pmatrix} m_1^{-1}\sqrt{D_{11}^{(N-1)}}J^{(N-1)}(N-1;1,...,1)|_{\text{without } 1} \\ m_2^{-1}\sqrt{D_{22}^{(N-1)}}J^{(N-1)}(N-1;1,...,1)|_{\text{without } 2} \\ \cdots \\ m_N^{-1}\sqrt{D_{NN}^{(N-1)}}J^{(N-1)}(N-1;1,...,1)|_{\text{without } N} \end{pmatrix}. \tag{7.15}$$

(Similar results can also be obtained using the integration-by-parts approach.[46,23])

Using the expression (2.24) for $\widetilde{c}_{jN}$, we arrive at

$$J^{(N)}(N+1;\{1+\delta_{il}\}) = \frac{\pi}{2m_l D^{(N)}} \sum_{j=1}^{N} \frac{(-1)^{j+l}}{m_j} D_{jl}^{(N-1)} J^{(N-1)}(N-1;1,...,1)|_{\text{without } j}. \tag{7.16}$$

In particular, in the three-point four-dimensional case we end up with

$$J^{(3)}(4;1,1,2) = \frac{i\pi^2 \sin \tau_{12}}{2m_1 m_2 m_3^2 D^{(3)}} [\tau_{23} \cos \psi_{13} + \tau_{13} \cos \psi_{23} + \tau_{12}]. \tag{7.17}$$

The results for $J^{(3)}(4;1,2,1)$ and $J^{(3)}(4;2,1,1)$ can be obtained by permutation of indices. The sum of these three integrals is in agreement with the result obtained in Ref. 18.

It is easy to check the consistency of the results (7.16) and (7.9). Using the standard Feynman parametric representation (or the representation (3.8) involving the δ function of the quadratic form) it is straightforward to show that

$$\sum_{j=1}^{N} J^{(N)}(N+1;\{1+\delta_{ij}\}) = -\pi J^{(N)}(N-1;1,...,1). \tag{7.18}$$

[The generalization of this formula to the case of arbitrary n and the powers of the propagators ν_i can be found, e.g., in Ref. 47, Eq. (6).] Considering the sum of the integrals (7.16) and employing the relation (7.18), we arrive at Eq. (7.9).

VIII. CONCLUSION

Let us briefly summarize the main results of the present paper. We have proven that there is a direct link between Feynman parametric representation of a one-loop N-point function and the basic simplex in N-dimensional Euclidean space. To show this, we have changed the surface of integration in the Feynman parametric representation, from a linear one ($\Sigma \alpha_i = 1$) to a quadratic one ($\alpha^T \|c\| \alpha = 1$). The integral representation involving the δ function of this quadratic form is given by Eq. (3.8).

In the case $N = n$, the result for the Feynman integral turns out to be proportional to the ratio of an N-dimensional solid angle at the meeting point of the mass sides to the content of the N-dimensional basic simplex. (Moreover, in the case $N = n$ the result for the diagram with unequal masses can be expressed in terms of the function corresponding to the equal-mass case.) For example, to reproduce the result for the three-dimensional three-point function[11] we just require the expression for the area of a spherical triangle which can easily be calculated as the sum of the angles minus π (in general, an analytic continuation of the area function should be taken).

For the four-dimensional four-point function, the representation (3.8) provides a very interesting connection with the volume of the non-Euclidean (spherical or hyperbolic) tetrahedron. By splitting the non-Euclidean tetrahedron into birectangular ones, the latter can be expressed in terms of dilogarithms or related functions. (A compact integral representation for this volume can be found in Ref. 48.) Different ways of splitting correspond to different representations of the result for the four-point function. This ''freedom'' can be used to construct more compact representations (cf. Refs. 3, 2, 4).

In other cases, such as, for example, the four-dimensional three-point function, we get an extra factor in the integrand corresponding to a power of the cosine of the angle between the height of the basic simplex and the vector of integration. In general, the height of the basic simplex, H_0, plays an essential role in calculation of the integrals with $N \neq n$. It is used to split the basic Euclidean simplex into N rectangular simplices. When $N=n+1$, each integral corresponding to one of the resulting rectangular tetrahedra can be reduced to an $(N-1)$-point function. When $N<n$, this splitting also simplifies the calculation of separate integrals.

The derivations in this paper have been based on a purely geometrical approach to evaluation of N-point Feynman diagrams. In the resulting expressions, all arguments of functions arising possess a straightforward geometrical meaning in terms of the dihedral angles, etc. In particular, this is quite useful for choosing the most convenient kinematic variables to describe the N-point diagrams. We suggest that this approach can serve as a powerful tool for understanding the geometrical structure of loop integrals with several external legs, as well as the structure of the corresponding phase-space integrals (see, e.g., Ref. 49). More optimistically, it may shed light on analytical results for higher loops.

Note added in proof: After our manuscript was accepted for publication Ref. 50 appeared, which deals with nonrelativistic Feynman integrals. It cited Ref. 51 where the relativistic one-loop Feynman amplitudes (in Euclidean quantum field theory) were studied and expressed in terms of definite integrals over simplices. We also note that some representations for the Schläfli function (similar to those discussed in Sec. VI B) can be found in Ref. 52.

ACKNOWLEDGMENTS

One of the authors (A.I.D.) is grateful to the Australian Research Council for a grant which was used to support his visit to Hobart in 1996 (where an essential part of this work has been done), and to the Department of Physics, University of Tasmania for hospitality during the visit. A.I.D.'s research was partly supported by Grant Nos. INTAS–93-0744 and RFBR–96-01-00654.

APPENDIX A: DILOGARITHM AND RELATED FUNCTIONS

Here we present definitions and some representations of the dilogarithm and related functions. More detailed information can be found in Ref. 27.

The Euler's dilogarithm is defined by

$$\mathrm{Li}_2(z) \equiv -z \int_0^1 d\xi \, \frac{\ln \xi}{1-\xi z}. \tag{A1}$$

The Clausen function is related to the imaginary part of the dilogarithm,

$$\mathrm{Cl}_2(\theta) = \mathrm{Im}[\mathrm{Li}_2(e^{i\theta})] = -\int_0^\theta d\theta' \, \ln\left|2\sin\left(\tfrac{1}{2}\theta'\right)\right|, \tag{A2}$$

and it can also be represented as

$$\mathrm{Cl}_2(\theta) = -\sin\theta \int_0^1 \frac{d\xi \, \ln \xi}{1-2\xi\cos\theta+\xi^2}. \tag{A3}$$

Note that $\mathrm{Cl}_2(2\pi-\theta) = -\mathrm{Cl}_2(\theta)$.

If we consider dilogarithm of a general complex argument, $\mathrm{Li}_2(re^{i\theta})$, we get

$$\mathrm{Re}[\mathrm{Li}_2(re^{i\theta})] \equiv \mathrm{Li}_2(r,\theta), \tag{A4}$$

$$\mathrm{Im}[\mathrm{Li}_2(re^{i\theta})] = \omega \ln r + \tfrac{1}{2}[\mathrm{Cl}_2(2\theta)+\mathrm{Cl}_2(2\omega)+\mathrm{Cl}_2(2\chi)], \tag{A5}$$

with $\tan\omega = r\sin\theta/(1-r\cos\theta)$ and $\chi \equiv \pi-\theta-\omega$. We note that

$$\tfrac{1}{2}[\mathrm{Cl}_2(2\theta)+\mathrm{Cl}_2(2\omega)+\mathrm{Cl}_2(2\chi)] = -\sin\theta \int_0^r \frac{d\eta \, \ln \eta}{1-2\eta\cos\theta+\eta^2}. \tag{A6}$$

The imaginary part can also be rewritten as

$$\mathrm{Im}[\mathrm{Li}_2(re^{i\theta})]=\mathrm{Ti}_2(\tan\,\omega)-\mathrm{Ti}_2(\tan\,\omega,\tan\,\theta)=-r\,\sin\,\theta\int_0^1\frac{d\xi\,\ln\,\xi}{1-2r\xi\,\cos\,\theta+r^2\xi^2},\quad\text{(A7)}$$

where $\mathrm{Ti}_2(z)$ and $\mathrm{Ti}_2(z,a)$ are the ordinary and the generalized inverse tangent integrals, respectively,

$$\mathrm{Ti}_2(z)=\int_0^z\frac{dz'}{z'}\arctan\,z',\quad \mathrm{Ti}_2(z,a)=\int_0^z\frac{dz'}{z'+a}\arctan\,z'.\quad\text{(A8)}$$

In particular, $\mathrm{Ti}_2(z)=\mathrm{Ti}_2(z,0)$.

Finally the Lobachevsky function $L(\theta)$ is defined by

$$L(\theta)\equiv-\int_0^\theta d\theta'\,\ln(\cos\,\theta'),\quad(-\tfrac{1}{2}\pi\leqslant\theta\leqslant\tfrac{1}{2}\pi).\quad\text{(A9)}$$

It is related to Clausen's function by

$$\mathrm{Cl}_2(\theta)=-2L\left(\frac{\pi-\theta}{2}\right)+(\pi-\theta)\,\ln 2,\quad L(\theta)=-\tfrac{1}{2}\mathrm{Cl}_2(\pi-2\theta)+\theta\,\ln 2.\quad\text{(A10)}$$

APPENDIX B: ON THE RESCALING OF α'S

Consider the transformation

$$\alpha_i=\mathscr{G}(\alpha'_1,\ldots,\alpha'_N)\alpha'_i,\quad i=1,\ldots,N,\quad\text{(B1)}$$

where the ''scaling'' function $\mathscr{G}$ is the same for all α's. The Jacobian of the transformation is

$$\text{Jacobian}=\left(\prod_{j=1}^N\alpha'_j\,\frac{\partial\mathscr{G}}{\partial\alpha'_j}\right)\begin{vmatrix}1+d_1 & 1 & \cdots & 1\\ 1 & 1+d_2 & \cdots & 1\\ \cdots & \cdots & \cdots & \cdots\\ 1 & 1 & \cdots & 1+d_N\end{vmatrix},\quad d_i\equiv\mathscr{G}\left(\alpha'_i\,\frac{\partial\mathscr{G}}{\partial\alpha'_i}\right)^{-1}.\quad\text{(B2)}$$

Since the determinant in (B2) is equal to

$$\left(\prod_{i=1}^N d_i\right)\left[1+\frac{1}{d_1}+\frac{1}{d_2}+\cdots+\frac{1}{d_N}\right],\quad\text{(B3)}$$

the expression for the Jacobian reduces to

$$\text{Jacobian}=\mathscr{G}^N\left(1+\frac{1}{\mathscr{G}}\sum_{j=1}^N\alpha'_j\,\frac{\partial\mathscr{G}}{\partial\alpha'_j}\right)=\mathscr{G}^N\left(1+\sum_{j=1}^N\alpha'_j\,\frac{\partial(\ln\,\mathscr{G})}{\partial\alpha'_j}\right).\quad\text{(B4)}$$

If $\mathscr{G}$ is a homogeneous function of α's of the order R, i.e.,

$$\mathscr{G}(Z\alpha'_1,\ldots,Z\alpha'_N)=Z^R\mathscr{G}(\alpha'_1,\ldots,\alpha'_N),\quad\text{(B5)}$$

then

$$\sum_{j=1}^N\alpha'_j\,\frac{\partial\mathscr{G}}{\partial\alpha'_j}=R\mathscr{G}\ \text{ and }\ \text{Jacobian}=[1+R]\mathscr{G}^N.\quad\text{(B6)}$$

In particular, if $\mathscr{G}=Q_r/Q'_{r'}$, where Q_r is a polynomial of degree r and $Q'_{r'}$ is a polynomial of degree r' in α', Jacobian$\to[1+(r-r')]\mathscr{G}^N$, i.e., $R=r-r'$. In the text we have encountered the situations when $r=r'=1$ $(R=0)$ and $r=2$, $r'=1$ $(R=1)$.

[1] L. M. Brown and R. P. Feynman, Phys. Rev. **85**, 231 (1952); G. Passarino and M. Veltman, Nucl. Phys. B **160**, 151 (1979).
[2] G. 'tHooft and M. Veltman, Nucl. Phys. B **153**, 365 (1979).
[3] A. C. T. Wu, Mat. Fys. Medd. K. Dan. Vidensk. Selsk. **33**(No. 3), 1 (1961).
[4] A. Denner, U. Nierste, and R. Scharf, Nucl. Phys. B **367**, 637 (1991).
[5] G. J. van Oldenborgh and J. A. M. Vermaseren, Z. Phys. C **46**, 425 (1990); G. J. van Oldenborgh, Phys. Lett. B **282**, 185 (1992).
[6] A. Denner, Fortschr. Phys. **41**, 307 (1993).
[7] F. R. Halpern, Phys. Rev. Lett. **10**, 310 (1963).
[8] B. Petersson, J. Math. Phys. **6**, 1955 (1965).
[9] D. B. Melrose, Nuovo Cimento **40A**, 181 (1965).
[10] G. Källén and J. Toll, J. Math. Phys. **6**, 299 (1965).
[11] B. G. Nickel, J. Math. Phys. **19**, 542 (1978).
[12] W. L. van Neerven and J. A. M. Vermaseren, Phys. Lett. B **137**, 241 (1984).
[13] Z. Bern, L. Dixon, and D. A. Kosower, Phys. Lett. B **302**, 299 (1993); Nucl. Phys. B **412**, 751 (1994).
[14] L. M. Brown, Nuovo Cimento **22**, 178 (1961).
[15] D. S. Kershaw, J. Math. Phys. **15**, 798 (1974); Phys. Rev. D **8**, 2708 (1973); E. E. Boos and A. I. Davydychev, Theor. Math. Phys. **89**, 1052 (1991); D. Kreimer, Z. Phys. C **54**, 667 (1992); Int. J. Mod. Phys. A **8**, 1797 (1993); L. Brücher, J. Franzkowski, and D. Kreimer, Mod. Phys. Lett. A **9**, 2335 (1994).
[16] A. I. Davydychev, J. Math. Phys. **32**, 1052 (1991); **33**, 358 (1992).
[17] L. D. Landau, Nucl. Phys. **13**, 181 (1959).
[18] G. Källen and A. Wightman, Mat. Fys. Skr. K. Dan. Vidensk. Selsk. **1**(No. 6), 1 (1958).
[19] S. Mandelstam, Phys. Rev. **115**, 1742 (1959); R. E. Cutkosky, J. Math. Phys. **1**, 429 (1960); J. C. Taylor, Phys. Rev. **117**, 261 (1960); Yu. A. Simonov, Sov. Phys. JETP **16**, 1599 (1963) [Zh. Eksp. Teor. Fiz. **43**, 2263 (1962)].
[20] R. J. Eden, P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, *The Analytic S-Matrix* (Cambridge University Press, Cambridge, 1966).
[21] R. Karplus, C. M. Sommerfield, and E. H. Wichmann, Phys. Rev. **114**, 376 (1959).
[22] J. S. Ball and T.-W. Chiu, Phys. Rev. D **22**, 2550 (1980); **23**, 3085(E) (1981); H. J. Lu and C. A. Perez, preprint SLAC-PUB-5809, 1992.
[23] A. I. Davydychev, J. Phys. A **25**, 5587 (1992).
[24] J. J. van der Bij and M. Veltman, Nucl. Phys. B **231**, 205 (1984); C. Ford, I. Jack, and D. R. T. Jones, *ibid.* **387**, 373 (1992); A. I. Davydychev and J. B. Tausk, *ibid.* **397**, 123 (1993).
[25] R. Scharf, doctoral thesis, Würzburg, 1994; R. Scharf and J. B. Tausk, Nucl. Phys. B **412**, 523 (1994).
[26] A. I. Davydychev and J. B. Tausk, Phys. Rev. D **53**, 7381 (1996).
[27] L. Lewin, *Polylogarithms and Associated Functions* (North–Holland, Amsterdam, 1981).
[28] D. M. Y. Sommerville, *An Introduction to the Geometry of N Dimensions* (Methuen, London, 1929).
[29] M. G. Kendall, *A Course in the Geometry of n Dimensions* (Charles Griffin, London, 1961).
[30] A. N. Vassiliev, Yu. M. Pis'mak, and Yu. R. Honkonen, Teor. Mat. Fiz. **47**, 291 (1981); N. I. Ussyukina, *ibid.* **54**, 124 (1983); D. I. Kazakov, Phys. Lett. **133B**, 406 (1983).
[31] A. K. Rajantie, Nucl. Phys. B **480**, 729 (1996).
[32] G. 'tHooft and M. Veltman, Nucl. Phys. B **44**, 189 (1972); C. G. Bollini and J. J. Giambiagi, Nuovo Cimento B **12**, 20 (1972).
[33] F. A. Berends, A. I. Davydychev, and V. A. Smirnov, Nucl. Phys. B **478**, 59 (1996).
[34] U. Nierste, D. Müller, and M. Böhm, Z. Phys. C **57**, 605 (1993).
[35] E. E. Boos and A. I. Davydychev, Teor. Mat. Fiz. **89**, 56 (1991).
[36] N. I. Lobatschefsky, *Imaginäre Geometrie, Kasaner Gelehrte Schriften, 1836* (Übersetzung mit Anmerkungen von H. Liebmann, Leipzig, 1904).
[37] L. Schläfli, Q. J. Math. **3**, 54 (1860); **3**, 97 (1860); *Gesammelte Matematische Abhandlungen* (Birkhäuser, Basel, 1953), Band II.
[38] J. L. Coolidge, *The Elements of Non-Euclidean Geometry* (Clarendon, Oxford, 1909); J. Böhm and E. Hertel, *Polyedergeometrie in n-Dimensionalen Räumen Konstanter Krümmung* (Birkhäuser, Basel, 1981).
[39] H. S. M. Coxeter, Q. J. Math. **6**, 13 (1935).
[40] E. B. Vinberg, Usp. Mat. Nauk **48**(No. 2), 17 (1993) [Russian Math. Surveys **48**(No. 2), 15 (1993)].
[41] R. Karplus and M. Neuman, Phys. Rev. **80**, 380 (1950); **83**, 776 (1951); B. De Tollis, Nuovo Cimento **32**, 757 (1964); **35**, 1182 (1965).
[42] J. Milnor, Bull. Am. Math. Soc. **6**, 9 (1982).
[43] N. I. Ussyukina and A. I. Davydychev, Phys. Lett. B **298**, 363 (1993); **305**, 136 (1993); D. J. Broadhurst, *ibid.* **307**, 132 (1993).
[44] Z. Bern and A. G. Morgan, Nucl. Phys. B **467**, 479 (1996).
[45] O. V. Tarasov, Phys. Rev. D **54**, 6479 (1996).
[46] F. V. Tkachov, Phys. Lett. **100B**, 65 (1981); K. G. Chetyrkin and F. V. Tkachov, Nucl. Phys. B **192**, 159 (1981).
[47] A. I. Davydychev, Phys. Lett. B **263**, 107 (1991).
[48] W.-Y. Hsiang, Q. J. Math. **39**, 463 (1988).
[49] B. Almgren, Ark. Fys. **38**(No. 7), 161 (1968); E. Byckling and K. Kajantie, *Particle Kinematics* (Wiley, New York, 1973).
[50] P. Wagner, J. Math. Phys. **39**, 2428 (1998).
[51] N. Ortner and P. Wagner, Ann. Inst. Henri Poincaré: Phys. Théor. **63**, 81 (1995).
[52] R. Kellerhals, Math. Ann. **285**, 541 (1989); in *Structural Properties of Polylogarithms*, edited by L. Lewin (AMS Math. Surveys and Monographs, Providence, RI, 1991), Vol. 37, p. 301.

A proof of the dynamical version of the Bombieri–Taylor conjecture

César R. de Oliveira[a)]
Departamento de Matemática, UFSCar, C.P. 676, São Carlos, SP, 13560-970 Brazil

(Received 3 December 1997; accepted for publication 8 May 1998)

An adaptation of the Bombieri–Taylor conjecture on diffraction has been applied to some time-dependent quantum systems. Here the problem is mathematically stated and a proof is presented. Simple examples give an idea as to what extent the conjecture is valid. © *1998 American Institute of Physics.*
[S0022-2488(98)02709-1]

I. INTRODUCTION

This work is a contribution to the investigations of some quantum systems under general time-dependent Hamiltonians; physical problems with nonperiodic characters have been discussed mainly in the context of discrete Schrödinger operators,[1–4] models in statistical mechanics and dynamical systems.[5–7] Besides the intrinsic physical interest, two other points justify these choices: the discovery of quasicrystals[8,9] and the rather well-developed techniques to treat such problems (mainly in one dimension).

The investigations become even more exciting as soon as one realizes that there are many different kinds of nonperiodicity, ranging from quasiperiodic to random. An effective way to materialize nonperiodicities is by considering sequences of substitution. The autocorrelation of each sequence probes its randomness; there are examples of (nonperiodic) substitution sequences with pure point, pure singular continuous, and even pure absolutely continuous autocorrelation measures.[10,11] Inserting in some way these sequences in physical models and understanding their influences is a problem of major interest currently.

Another attractive area for the investigation of the influence of nonperiodicity is the dynamical behavior of systems with time-dependent Hamiltonians. In this context there is a lack of tools to investigate the asymptotic behavior of such systems as time goes to infinity, at least when compared with statistical mechanics, dynamical systems and the analysis of tight-binding models. Although there is a generalization of the Floquet operator for the case of quasiperiodic forces (finite number of fundamental frequencies),[12] no such generalization seems evident for general time dependence (see, e.g., Ref. 13). An important point here is the quantum autocorrelation measure, a notion that is expected to replace the concept of spectral measures[14] (recall that we have a family of possible different Hamiltonians), but even in the case of periodic Hamiltonians it may be very difficult to compute the autocorrelation functions and the corresponding measures. See Ref. 15 for some rare explicit computations in the case of kicked Hamiltonians driven by the Thue–Morse sequence and Refs. 16–18 for related numerical tricks and comments.

In Ref. 19 (see also Refs. 20 and 21 and references therein for other investigations on nonautonomous quantum systems) another quantity has been proposed to study the dynamical behavior of quantum systems driven by the Fibonacci substitution sequence. It is strongly based on the notion of Fourier transform and on a conjecture of Bombieri and Taylor[22,23] about diffraction by nonperiodic structures, and we shall denote this quantity by $\mathscr{B}$ (see Sec. II for precise statements). It is expected that in some cases $\mathscr{B}$ should be easier to compute than autocorrelation functions and that it gives information about the point part of the autocorrelation measures. Nevertheless, no relation between $\mathscr{B}$ and autocorrelation measures has been explicitly considered yet. Therefore we have a ''dynamical conjecture,'' i.e., it is necessary to clarify the relations between $\mathscr{B}$ and the autocorrelation measures of general time-dependent quantum systems. We refer to this problem as ''the dynamical version of the Bombieri–Taylor conjecture.'' The main

a)Electronic mail: oliveira@dm.ufscar.br

 © 1998 American Institute of Physics

goals of this work are to formulate precisely this problem and present a proof (not its applications).

This work is organized as follows. In Sec. II we briefly recall the original Bombieri–Taylor conjecture and set up a rigorous formulation of its dynamical version. In Sec. III a proof of this ''dynamical conjecture'' is given. In Sec. IV we discuss some particular cases.

II. THE PROBLEM

A. The Bombieri–Taylor prescription

In this subsection we recall the Bombieri–Taylor argument. The general question raised by Bombieri and Taylor[22,23] is ''what nonperiodic distribution of matter would show spots in a diffraction experiment.'' In a diffraction experiment one measures the intensity of the scattered radiation (x-rays, for example) from a material, and this intensity tells us about the microscopic structure of the material. A regularity in the bright spots pattern indicates an organization of the structure and was usually related only to crystals until the experimental discovery of the so-called quasicrystals.[8] See Ref. 24 for a discussion about the definitions of crystals and quasicrystals and Steffen Weber's home page[9] for an updated introduction to the subject.

Let $X=\{x_j\}_{j=0}^{\infty}$ be a countable set in $\mathbb{R}$ with no accumulation point (here it is enough to consider dimension one), which represents the positions of the atoms in a sample of matter. The diffraction pattern is related to the Fourier transform of the density distribution of atoms; in this case the density is represented by the measure

$$\rho(x)=\sum_{x_j\in X}\delta(x-x_j),$$

where δ denotes the Dirac measure. According to Bombieri–Taylor, the set X looks ordered in a diffraction experiment if the Fourier transform of ρ, i.e., $\hat{\rho}$, has the form

$$\hat{\rho}(\xi)=\sum_{\lambda\in Y}a_\lambda\delta(\xi-\lambda)+g(\xi)$$

with Y being a subset of $\mathbb{R}$. The presence of an atomic part (i.e., δ terms) in $\hat{\rho}$ is the indication of long-range correlations in the atomic positions and is an effect of constructive interference. The function g would describe a diffusive background. The Bombieri–Taylor question has been discussed and accurately modified by Hof.[25]

In order to decide if $\hat{\rho}$ has an atomic part at $\xi\in\mathbb{R}$, Bombieri–Taylor argued that this will occur if and only if

$$F_N(\xi)=\sum_{j=0}^{N}\exp(2\pi i\xi x_j)$$

is such that the ratio $F_N(\xi)/(N+1)$ has a nonzero limit as N goes to infinity. Since they did not present a justification of this statement, one calls it a conjecture. Certainly it has its roots in the Wiener–Khinchin theorem.[26] We refer to Refs. 25 and 27 for a discussion about this conjecture, including references of applications to physical models.

B. The dynamical version

In this subsection we discuss the dynamical version of the Bombieri–Taylor conjecture. The physical framework here is composed by quantum systems subjected to kicks at each positive integer instant of time. Let $\{V_j\}$ be a sequence of potentials; the Hamiltonian we consider is

$$H=H_0+\sum_{j=1}^{\infty}V_j\delta(t-j).$$

Notice that a kick does not occur at instant n if V_n is the null function.

We have not specified the Hilbert space $\mathscr{H}$ in which the operators act since, for our results, it is relevant only for the existence of dynamics. In this respect we suppose the operators involved are self-adjoint and, by using a standard construction, we suppose the propagator between $t=(j-1+0)$ and $t=(j+0)$ has been found and is given by a unitary operator U_j acting on $\mathscr{H}$. We shall denote the evolution operator from the initial instant $t=0$ until $t=n+0$ by $\Lambda(n)$, so that

$$\Lambda(n)=U_nU_{n-1}\cdots U_1.$$

The general question to be addressed is about the asymptotic behavior of $\Lambda(n)$ for n large. This kind of question is commonly referred to as the quantum stability problem and for autonomous systems a major role is played by the Hamiltonian spectrum. Continuous spectrum and point spectrum entail different dynamical behaviors. For general nonautonomous systems we do not have a spectral theory at our disposal, so that one has to use other tools to classify the dynamical behavior of these systems. On the basis of the Bombieri–Taylor conjecture and the idea that point spectrum is also a kind of constructive interference, in Ref. 19 a dynamical version of the Bombieri–Taylor conjecture was proposed and applied to two-level systems subjected to kicks modulated along the Fibonacci substitution sequence.

The authors of Ref. 19 argued that a pointlike (''quasiperiodic'') dynamical behavior is present if for some $\omega\in\mathbb{R}$ the quantity

$$G_N(\omega)=\sum_{j=0}^{N} e^{ij\omega}\Lambda(j)$$

is such that $G_N/(N+1)$ has a nonzero limit for $N\to\infty$. The application of this conjecture was also the focus of Refs. 20 and 21.

The admitted generalization of spectrum for nonautonomous systems is given by the support of autocorrelation measures. The autocorrelation function for $\psi\in\mathscr{H}$ (assume $\|\psi\|=1$) is

$$C_\psi(k)=\lim_{N\to\infty}\frac{1}{N+1}\sum_{j=0}^{N}\langle\Lambda(j+k)\psi|\Lambda(j)\psi\rangle$$

for integers numbers $k\geqslant 0$, and we define $C_\psi(-k)=\overline{C_\psi(k)}$ (the limits are supposed to exist). This function is of positive type and so is the Fourier transform of a unique positive measure σ_ψ, the so-called autocorrelation measure associated with ψ. The key observation is that for autonomous systems the autocorrelation measure σ_ψ is the spectral measure μ_ψ of the Hamiltonian operator[14] [this can be easily seen by direct substitution of $\Lambda(n)=\exp(-inH)$ into the limit that defines $C_\psi(k)$].

In analogy with the case of diffraction, the general question now can be put in the following form: which kind of kick mimics a pointlike component in the system time evolution? In other words, which kind of kick generates deltas (atoms) in the autocorrelation measures? Therefore, the dynamical version of the Bombieri–Taylor conjecture is to relate the limit of $G_N/(N+1)$ with the point part of autocorrelation measures.

C. Main results

In this subsection we give a precise formulation of the dynamical version of the conjecture and state the main results of this work. Let $\{U_j\}_{j=1}^{\infty}$ be a sequence of unitary operators acting on the Hilbert space $\mathscr{H}$ and set $\Lambda(n)=U_nU_{n-1}\ldots U_1$, $\Lambda(0)=\mathrm{Id}$ (the identity operator). Suppose that for any $\psi\in\mathscr{H}$ the autocorrelation function exists and denote by σ_ψ the corresponding autocorrelation measure. In general, the proof of the existence of the autocorrelation function is far from trivial and should be analyzed separately in each case.

For $\omega\in[0,2\pi]$ (it is enough to consider ω values in this interval) set

$$\phi_N(\omega)=\frac{1}{N+1}\sum_{j=0}^{N}\exp(ij\omega)\Lambda(j)$$

and

$$\mathscr{B}_\psi(\omega)=\limsup_{N\to\infty}\|\phi_N(\omega)\psi\|^2.$$

Theorem 1: *Suppose the autocorrelation functions of* Λ *exist. Then, for any* $\omega\in[0,2\pi]$, $\psi\in\mathscr{H}$, *we have* $0\leqslant\mathscr{B}_\psi(\omega)\leqslant 2\sigma_\psi(\{\omega\})$.

Corollary 1: If σ_ψ *is a continuous measure, then* $\mathscr{B}_\psi(\omega)=0$ *for any* ω.

Corollary 2: If for some ω *we have* $\mathscr{B}_\psi(\omega)>0$, *then* σ_ψ *has an atom at* ω.

The above Theorem and Corollaries constitute the rigorous formulation of the dynamical version of the Bombieri–Taylor conjecture; see Sec. IV for a nonperiodic example such that $\mathscr{B}_\psi(\omega)<\sigma_\psi(\{\omega\})$. We suspect the factor ''2'' in Theorem 1 is an artifact of our nonoptimal estimates and should be replaced by ''1.'' Another useful concept borrowed from material science is the structure factor.[5,7]

Definition 1: For each ω *set*

$$\Xi_N(\omega)=\frac{1}{N+1}\left[\sum_{j=0}^{N}e^{-ij\omega}\Lambda^*(j)\right]\left[\sum_{j=0}^{N}e^{ij\omega}\Lambda(j)\right].$$

Any weak limit point of the sequence of absolutely continuous measures*

$$d\nu_N(\omega)=\langle\Xi_N(\omega)\psi|\psi\rangle d\omega$$

will be called a structure factor *for* Λ *(associated with* ψ*).*

Theorem 2: *Suppose the autocorrelation functions of* Λ *exist. Then, for each* ψ, Λ *has a unique structure factor and it coincides with the autocorrelation measure* σ_ψ.

In other words, Theorem 2 says that the sequence of measures $d\nu_N(\omega)$ converges in the weak* topology to the autocorrelation measure of ψ (in case this autocorrelation exists). This could be of practical relevance since one can approximate autocorrelation measures by absolutely continuous measures (with respect to Lebesgue measure).

III. PROOFS

We begin with some notation. For any Borelian measure μ on the unit circle S we shall denote its Fourier transform by $\hat{\mu}(k)=\int_0^{2\pi}e^{-ikx}\mu(dx)$, $k\in\mathbb{Z}$, and for any integrable function f we shall also denote $\mu(f)=\int f d\mu$. In many opportunities we shall drop the index ψ.

Proof of Theorem 1: Direct substitution gives

$$\|\phi_N(\omega)\psi\|^2=\frac{1}{(N+1)^2}\sum_{n,j=0}^{N}\exp(i(n-j)\omega)\langle\Lambda(n)\psi|\Lambda(j)\psi\rangle=\frac{1}{N+1}\sum_{k=-N}^{N}\exp(ik\omega)C_N(k),$$

where $C_N(k)$ depends on ψ and, for $-N\leqslant k\leqslant N$, it is given by

$$C_N(k)=\frac{1}{N+1}\sum_{j=0}^{N-k}\langle\Lambda(j+k)\psi|\Lambda(j)\psi\rangle,\quad k\geqslant 0,$$

and $C_N(-k)=\overline{C_N(k)}$. We extend C_N for $|k|>N$ by defining

$$C_N(k)=\frac{1}{N+1}\sum_{j=0}^{N}\langle\Lambda(j+k)\psi|\Lambda(j)\psi\rangle,\quad k>N,$$

and $C_N(-k)=\overline{C_N(k)}$. Therefore, for each N there is a measure σ^N such that $C_N(k)=\hat{\sigma}^N(k)$ for any $k\in\mathbb{Z}$. Notice that, for normalized vectors, σ^N are probability measures.

Since it is supposed that the autocorrelation functions exist, it is readily checked that, for each $k\in\mathbb{Z}$, $C_N(k)\to C_\psi(k)$ for $N\to\infty$; therefore, $\hat{\sigma}^N(k)\to\hat{\sigma}_\psi(k)$ for each integer k and $\sigma^N\to\sigma_\psi$ in the weak* topology.

Now we rewrite

$$\|\phi_N(\omega)\psi\|^2=\frac{1}{N+1}\sum_{k=-N}^{N}\exp(ik\omega)\int_0^{2\pi}\exp(-ikx)\sigma^N(dx)=\sigma^N(f_N),$$

where $f_N(x)=(1/(N+1))\Sigma_{k=-N}^{N}\exp(ik(\omega-x))$ is a continuous function for each N. Notice that $|f_N|\leqslant 2$ and that $f_N\to 2\chi_{\{\omega\}}$ (the characteristic function of the set $\{\omega\}$) for $N\to\infty$.

Let $0<\epsilon<1$ and set $A_\epsilon(\omega)=[\omega-\epsilon,\omega+\epsilon]\cap[0,2\pi]$. For N large we have

$$0\leqslant\sigma^N(f_N)=\sigma^N(\chi_{A_\epsilon}f_N)+\sigma^N((1-\chi_{A_\epsilon})f_N)\leqslant\sigma^N(\chi_{A_\epsilon}|f_N|)+\sigma^N((1-\chi_{A_\epsilon})|f_N|)$$

$$\leqslant 2\sigma^N(\chi_{A_\epsilon})+\sigma^N((1-\chi_{A_\epsilon})|f_N|).$$

Since $\epsilon>0$, it follows that $(1-\chi_{A_\epsilon})|f_N|\to 0$ uniformly and $\sigma^N((1-\chi_{A_\epsilon})|f_N|)\leqslant c\epsilon/N$ for some positive constant c. Therefore,

$$0\leqslant\sigma^N(f_N)\leqslant 2\sigma^N(\chi_{A_\epsilon})+c\epsilon/N.$$

Since A_ϵ is a closed set, it follows that $\limsup_{N\to\infty}\sigma^N(\chi_{A_\epsilon})\leqslant\sigma_\psi(\chi_{A_\epsilon})$.[28] Therefore, taking the lim sup for $N\to\infty$ in the above expression we obtain

$$0\leqslant\mathscr{B}_\psi(\omega)\leqslant 2\sigma_\psi(\chi_{A_\epsilon}).$$

Now consider a sequence $\epsilon_n\downarrow 0$ so that $\cap_{n=1}^{\infty}A_{\epsilon n}=\{\omega\}$; σ_ψ being a finite measure, the theorem follows.

Remark: If $C_N(k)$ converges uniformly to $C_\psi(k)$, then it is possible to show that $\mathscr{B}_\psi(\omega)=\sigma_\psi(\{\omega\})=0$ for all ω.

Proof of Theorem 2: We have $\Xi_N(\omega)=1/(N+1)\Sigma_{j=0}^{N}\Sigma_{k=-j}^{N-j}e^{ik\omega}\Lambda^*(j)\Lambda(j+k)$, so that

$$d\nu_N(\omega)=\left[\frac{1}{N+1}\sum_{j=0}^{N}\sum_{k=-j}^{N-j}e^{ik\omega}\langle\Lambda(j+k)\psi|\Lambda(j)\psi\rangle\right]d\omega;$$

its Fourier transform is given by ($r\in\mathbb{Z}$ and $N>|r|$)

$$\hat{\nu}_N(r)=\frac{1}{N+1}\sum_{j=0}^{N}\sum_{k=-j}^{N-j}\delta_{k,r}\langle\Lambda(j+k)\psi|\Lambda(j)\psi\rangle=\frac{1}{N+1}\sum_{j=0}^{N}\langle\Lambda(j+r)\psi|\Lambda(j)\psi\rangle.$$

Therefore, $\hat{\nu}_N(r)\to\hat{\sigma}_\psi(r)$ for any r and ν_N weak* converges to σ_ψ.

IV. PARTICULAR CASES

A. Example

Here we present an example of a nonperiodic sequence $\{U_j\}$ for which $\mathscr{B}_\psi(\omega)=0$ for any ω but $\sigma_\psi(\{0\})=1$. This shows that the ''if and only if'' part of the dynamical version of the Bombieri–Taylor conjecture cannot hold without further restrictions. This example is a very simple adaptation of an analysis of Allouche and Mendès France[29] and was also used in Ref. 27.

Take $\Lambda(n)=\exp(2\pi in^{1/3})Id$ for each $n\in\mathbb{Z}$ acting on a Hilbert space $\mathscr{H}$. Thus

$$\phi_N(\omega=0)=\frac{1}{N+1}\sum_{j=0}^{N}\exp(2\pi ij^{1/3})\mathrm{Id}.$$

By the Fejer theorem[30] the sequence $\{j^{1/3}\bmod(1)\}_{j=0}^{\infty}$ is uniformly distributed in $[0,1]$. Therefore

$$\phi_N(0)\psi=\frac{1}{N+1}\sum_{j=0}^{N}\exp(2\pi ij^{1/3})\psi\to 0$$

for $N\to\infty$ and $\mathscr{B}_\psi(0)=0$ for any normalized $\psi\in\mathscr{H}$.

On the other hand,

$$d\nu_N(\omega)=\left[\frac{1}{N+1}\sum_{j,n=0}^{N}\exp(i(j-n)\omega)\exp(-i2\pi(j^{1/3}-n^{1/3}))\right]\|\psi\|^2 d\omega$$

so that

$$\hat{\nu}_N(k)=\frac{1}{N+1}\sum_{j=0}^{N}\exp(-i2\pi(j^{1/3}-(j+k)^{1/3})).$$

Since $(j^{1/3}-(j+k)^{1/3})\to 0$ for $j\to\infty$, it follows that $\hat{\nu}_N(k)\to 1$ for any $k\in\mathbb{Z}$, so that ν_N weak* converges to the Dirac measure δ_0. By Theorem 2 we get $\sigma_\psi=\delta_0$; then $\sigma_\psi(\{0\})=1$ for any normalized vector ψ, while $\mathscr{B}_\psi(0)=0$.

This example shows that $\mathscr{B}_\psi$ is too sensitive to phase variations, which sometimes is not so important in quantum mechanics, as its autocorrelation measures attest.

B. Periodic case

Here we discuss the case $\Lambda(n)=U_{n-1}U_{n-1}\cdots U_1U_0$ with $\{U_j\}$ being a periodic sequence (of period T) of unitary operators. We have introduced U_0 just for notational convenience.

Let $U_F=\Lambda(T)$ (the Floquet operator) and $L(\omega)=\Sigma_{j=0}^{T-1}\exp(ij\omega)\Lambda(j)$; for each ω, $L(\omega)$ is a bounded operator on the (in this subsection we suppose separable) Hilbert space $\mathscr{H}$ [$\|L(\omega)\|\leqslant T$ for any ω]. Let $\{\psi_k\}$ be an orthonormal basis (of eigenvectors of U_F, $U_F\psi_k=e^{-iT\alpha_k}\psi_k$, with $0\leqslant T\alpha_k<2\pi$) of the point subspace $\mathscr{H}_p$ of U_F; denote the continuous subspace of U_F by $\mathscr{H}_c$ and define $\mathscr{H}_p^1=\{\psi\in\mathscr{H}_p:\psi=\Sigma_k a_k\psi_k$ with $\Sigma_k|a_k|<\infty\}$. $\mathscr{H}_p^1$ is a dense subspace of $\mathscr{H}_p$.

Theorem 3: *(i) If zero is not an eigenvalue of $L(\omega)$ for all ω (i.e., L is invertible) and $\psi\in\mathscr{H}_p^1$, then $\mathscr{B}_\psi(\omega)>0$ if and only if $\mu_\psi(\{\omega\})>0$, where μ_ψ is the spectral measure of U_F associated with ψ.*

(ii) For $\psi\in\mathscr{H}_c$ we have $\mathscr{B}_\psi(\omega)=0$ for all ω.

Corollary 3: If zero is not an eigenvalue of $L(\omega)$ for all ω, then $\mathscr{B}_\psi(\omega)=0$ for all vectors $\psi\in\mathscr{H}_p^1\oplus\mathscr{H}_c$ if and only if $\omega\neq\alpha_k$ for all k.

Proof: Each N can be written in the form $N=nT+s$, with $0\leqslant s<T$ and n is an integer number. Then

$$\phi_N(\omega)=\frac{1}{N+1}\sum_{j=0}^{N}\exp(ij\omega)\Lambda(j)=\frac{1}{N+1}\left[\sum_{j=0}^{nT}+\sum_{j=nT+1}^{nT+s}\right]\exp(ij\omega)\Lambda(j)$$

$$=L(\omega)\frac{1}{N+1}\sum_{j=0}^{n-1}\exp(ijT\omega)U_F^j+\frac{1}{N+1}\sum_{j=nT+1}^{nT+s}\exp(ij\omega)\Lambda(j).$$

By the inequality $\|[1/(N+1)]\Sigma_{j=nT+1}^{nT+s}\exp(ij\omega)\Lambda(j)\|\leqslant T/(N+1)$ we are justified in discarding the last term in the above expression; it does not contribute to $\mathscr{B}_\psi$. From now on we simply ignore it.

(i) For an eigenvector of U_F, say ψ_k, we have $\phi_N(\omega)\psi_k\approx(1/T)L(\omega)\times(1/n)\Sigma_{j=0}^{n-1}e^{iTj(\omega-\alpha_k)}\psi_k$; for $N\to\infty$ we have $n\to\infty$ and we get $\mathscr{B}_{\psi_k}(\omega)=(1/T^2)\delta_{\omega,\alpha_k}\|L(\omega)\psi_k\|^2$. Since $\|L(\omega)\psi_k\|\neq 0$ we conclude $\mathscr{B}_{\psi_k}(\omega)\neq 0$ if and only if $\omega=\alpha_k(\omega\in[0,2\pi])$.

Now let $\psi=\Sigma_k a_k\psi_k$ with $\Sigma_k|a_k|<\infty$. Then it is possible to justify an interchange of limits and we obtain

$$\mathscr{B}_\psi(\omega)=\frac{1}{T^2}\left\|L(\omega)\sum_k\delta_{\omega,\alpha_k}a_k\psi_k\right\|^2.$$

By the hypothesis that $L(\omega)$ is invertible it follows that $\mathscr{B}_\psi(\omega)\neq 0$ if and only if the vector $\Sigma_k\delta_{\omega,\alpha_k}a_k\psi_k\neq 0$, and this happens if and only if $\omega=\alpha_k$ for some k and $a_k\neq 0$. This proves the first part of Theorem 3.

(ii) Let $\psi\in\mathscr{H}_c$. Taking into account that $(1/T)\|L(\omega)\|\leqslant 1$. We obtain

$$\|\phi_N(\omega)\| \leqslant \left\| \frac{1}{n} \sum_{j=0}^{n-1} \exp(ijT\omega) U_F^j \psi \right\|.$$

(Recall we have discarded a small term.) From this it follows that

$$\|\phi_N(\omega)\|^2 \leqslant \frac{1}{n^2} \sum_{j,k=0}^{n-1} |\langle \psi | U_F^{k-j} \psi \rangle|;$$

the right-hand side of this expression vanishes for $N \to \infty$, as it is shown in Ref. 31, p. 165. This proves the second part of Theorem 3.

Remark: The spectral measures of U_F are closely related to the autocorrelation measures of Λ.[14]

Remark: By using Theorem 3 it is easy to construct examples with $\mathscr{B}_\psi < \sigma_\psi$; choose the periodic sequence $\{U_j\}$ such that $L(\omega) = 0$. For instance, take $U_0 = \mathrm{Id}$. $U_1 = -\mathrm{Id}$, and repeat these values periodically (period 2). In this case $L(0) = 0$ while $\sigma_\psi = \delta_0$ for all normalized vector ψ. Therefore $0 = \mathscr{B}_\psi(0) < \sigma_\psi(\{0\}) = 1$. The example in the previous subsection shows that this can also happen in the nonperiodic case.

C. Period one

Here $T = 1$ and the situation is simpler than the general periodic case. We have a constant sequence of unitary operators, i.e., there is a unitary operator U_F such that $U_j = U_F$ for all $j > 0$ and $\Lambda(n) = U_F^n$ ($U_0 = \mathrm{Id}$). In this case it is readily seen that the autocorrelation measures of Λ are exactly the spectral measures of U_F.

By following the arguments of the previous subsection, with $L(\omega) = \mathrm{Id}$, for $\psi \in \mathscr{H}_p$ ($= \mathscr{H}_p(U_F)$), $\psi = \Sigma_k a_k \psi_k$, one obtains (no additional condition on ψ is needed)

$$\|\phi_N(\omega)\psi\|^2 = \sum_k |a_k|^2 \left| \frac{1}{N+1} \sum_{j=0}^{N} \exp(i(\omega - \alpha_k)) \right|^2.$$

Therefore, for $N \to \infty$ we find $\mathscr{B}_\psi(\omega) = \Sigma_k \delta_{\omega,\alpha_k} |a_k|^2 = \mu_\psi(\{\omega\})$, where μ_ψ is the spectral measure of U_F associated to ψ.

For $\psi \in \mathscr{H}_c$ one also gets $\mathscr{B}_\psi(\omega) = 0$ for all ω, as a consequence of the RAGE theorem[31] (similar to what we argued in the previous section). Summing up,

Theorem 4: *Let U_F be a unitary operator on the separable Hilbert space $\mathscr{H}$ and $\{\mu_\psi\}$ its spectral measures. If $\Lambda(n) = U_F^n$, then $\mathscr{B}_\psi(\omega) = \mu_\psi(\{\omega\}) = \sigma_\psi(\{\omega\})$ for any $\psi \in \mathscr{H}$.*

Remark: Suppose $\{U_j\}$ is a periodic sequence of period T. Defining

$$\tilde{\mathscr{B}}_\psi(\omega) = \limsup_{N \to \infty} \left\| \frac{1}{N+1} \sum_{j=0}^{N} \exp(ijT\omega) U_F^j \psi \right\|^2,$$

one obtains results analogous to Theorem 4 (but now for periodic systems) with $\tilde{\mathscr{B}}_\psi$ instead of $\mathscr{B}_\psi$. What is crucial here is that L becomes the identity operator. This idea of using ''symmetries'' to redefine $\mathscr{B}_\psi$ may also be fruitful in other cases.

ACKNOWLEDGMENTS

I am grateful to A. Hof for sending me some preprints prior to publication and to G. Q. Pellegrino for reading the manuscript. This work was partially supported by CNPq (Brazil).

[1] A. Bovier and J.-M. Ghez, ''Spectral Properties of One-Dimensional Schrödinger Operators with Potentials Generated by Substitutions,'' Commun. Math. Phys. **158**, 45–66 (1993).
[2] A. Bovier and J.-M. Ghez, Erratum to Ref. 1, Commun. Math. Phys. **166**, 431–432 (1994).
[3] A. Sűtö, ''Singular Continuous Spectrum on a Cantor Set of Zero Lebesgue Measure for the Fibonacci Hamiltonian,'' J. Stat. Phys. **56**, 525–531 (1989).
[4] A. Hof, O. Knill, and B. Simon, ''Singular Continuous Spectrum for Palindromic Schrödinger Operators,'' Commun. Math. Phys. **174**, 149–159 (1995).

[5] S. Aubry, C. Godrèche, and J.-M. Luck, ''A Structure Intermediate Between Quasi-Periodic and Random,'' Europhys. Lett. **4**, 639–643 (1987).
[6] S. Aubry, C. Godrèche, and J.-M. Luck, ''Scaling Properties of a Structure Intermediate Between Quasiperiodic and Random,'' J. Stat. Phys. **51**, 1033–1075 (1988).
[7] A. Hof, ''On A Structure Intermediate Between Quasi-Periodic and Random,'' J. Stat. Phys. **84**, 309–320 (1996).
[8] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, ''Metallic Phase with Long-Range Orientational Order and no Translational Symmetry,'' Phys. Rev. Lett. **53**, 1951–1953 (1984).
[9] S. Weber, http://www.nirim.go.jp/~weber/
[10] M. Queffélec, *Substitution Dynamical Systems. Spectral Analysis*, LNM 1294 (Springer-Verlag, Berlin, 1987).
[11] F. Axel and D. Gratias (eds.), *Beyond Quasicrystals* (Les Editions de Physique/Springer-Verlag, Berlin, 1995).
[12] P. M. Blekher, H. R. Jauslin, and J. L. Lebowitz, ''Floquet Spectrum for Two-Level Systems in Quasiperiodic Time-Dependent Fields,'' J. Stat. Phys. **68**, 271–310 (1992).
[13] C. R. de Oliveira, ''Some Remarks Concerning Stability for Nonstationary Quantum Systems,'' J. Stat. Phys. **78**, 1055–1066 (1995).
[14] G. Casati and I. Guarneri, ''Chaos and Special Features of Quantum System Under External Perturbations,'' Phys. Rev. Lett. **50**, 640–643 (1983).
[15] M. Combescure, ''Recurrent Versus Diffusive Dynamics for a Kicked Quantum Oscillator,'' Ann. Inst. Henri Poincaré **57**, 67–87 (1992).
[16] N. F. de Godoy and R. Graham, ''Long-Range Correlations in Quantum Systems with Quasi-Periodic Fibonacci Hamiltonians,'' Europhys. Lett. **16**, 519–524 (1991).
[17] C. R. de Oliveira, ''Numerical Study of the Long-Time Behaviour of Quantum Systems Driven by Thue-Morse Forces. Application to Two-Level Systems,'' Europhys. Lett. **31**, 63–68 (1995).
[18] T. Geisel, ''Quasiperiodicity Versus Mixing Instability in a Kicked Quantum System,'' Phys. Rev. A **41**, 2989–2994 (1990).
[19] J. M. Luck, H. Orland, and U. Smilansky, ''On the Response of a Two-Level Quantum System to a Class of Time-Dependent Quasiperiodic Perturbations,'' J. Stat. Phys. **53**, 551–564 (1988).
[20] R. Graham, ''Nature of the Quasi-Energy Spectrum of Quasi-Periodically Driven N-Level Quantum Systems,'' Europhys. Lett. **8**, 717–721 (1989).
[21] C. R. de Oliveira, ''On Kicked Systems Modulated Along the Thue-Morse Sequence,'' J. Phys. A **27**, L847–L851 (1994).
[22] E. Bombieri and J. E. Taylor, ''Which Distributions of Matter Diffract? An Initial Investigation,'' J. Phys. Colloq. **47**(C3), 19–28 (1986).
[23] E. Bombieri and J. E. Taylor, ''Quasicrystals, Tilings, and Algebraic Number Theory: Some Preliminary Connections,'' in The *Legacy of S. Kovalevskaya*, Contemporary Mathematics Vol. 64 (American Mathematical Society, Providence, RI, 1987).
[24] M. Senechal and J. Taylor, ''Quasicrystals: the View from Les Houches,'' Math. Intell. **12**, 54–64 (1990).
[25] A. Hof, ''On Diffraction by Aperiodic Structures,'' Commun. Math. Phys. **169**, 25–43 (1995).
[26] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1966), Vol. II.
[27] A. Hof, ''Diffraction by Aperiodic Structures,'' in *The Mathematics of Aperiodic Order*, edited by R. V. Moody and J. Padera (Proceedings of NATO-ASI, Waterloo, 1995).
[28] P. Billingsley, *Weak Convergence of Measures: Applications in Probability* (Society for Industrial and Applied Mathematics, Philadelphia, 1971).
[29] J.-P. Allouche and M. Mendès France, ''Automata and Automatic Sequences,'' in Ref. 11.
[30] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences* (Wiley, New York, 1974).
[31] V. Enss and K. Veselić, ''Bound States and Propagating States for Time-Dependent Hamiltonians,'' Ann. Inst. Henri Poincaré **39**, 159–191 (1983).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

The general decomposition theory of SU(2) gauge potential, topological structure and bifurcation of SU(2) Chern density

Yishi Duan[a] and Libin Fu[b]
Institute of Theoretical Physics, Lanzhou University, Lanzhou, Gansu, 730000, People's Republic of China

(Received 7 October 1997; accepted for publication 14 May 1998)

By means of the geometric algebra the general decomposition of SU(2) gauge potential on the sphere bundle of a compact and oriented four-dimensional manifold is given. Using this decomposition theory the SU(2) Chern density has been studied in detail. It shows that the SU(2) Chern density can be expressed in terms of the δ-function $\delta(\phi)$. One can also find that the zero points of the vector fields ϕ are essential to the topological properties of a manifold. It is shown that there exists the crucial case of branch process at the zero points. Based on the implicit function theorem and the Taylor expansion, the bifurcation of the Chern density is detailed in the neighborhoods of the bifurcation points of ϕ. It is pointed out that, since the Chern density is a topological invariant, the sum topological chargers of the branches will remain constant during the bifurcation process. © *1998 American Institute of Physics.* [S0022-2488(98)02809-6]

I. INTRODUCTION

The topological properties of physics systems play important roles in studying some physical problems. It is well known that the gauge potential (connection) and gauge field (curvature) are essential to establish a direct relationship between differential geometry and topological invariants. The decomposition theory of gauge potential provides a powerful method in researching some topological properties. It has been effectively used to study the topological gauge theory of dislocation and disclinations in condensed matter physics,[1] the geometrization of the Planck constant in terms of the space–time defect in general relativity,[2,3] the space–time dislocations in the early universe,[4] and the Gauss–Bonnet–Chern theorem.[3,5] The essential feature of the decomposition shows that the gauge potential has inner structure.[6,1] Generally, the topological characteristics of a manifold are represented by the properties of a smooth vector field on it, or, in other words, the smooth vector fields carry the topological information of a manifold, which inspires us to study the decomposition theory of the gauge potential in terms of the unit vector field on the manifold.

It is well known that the SU(2) gauge theory and the second Chern class has been widely used in discussing many physical problems such as the magnetic monopole,[6–9] and the anomaly in nonlinear σ-models,[10] and especially the problems on the four-manifold,[11,12] the instantons,[13–15] the merons,[16,17] the Donsdson theory,[18] and so on. It urges us to study the decomposition theory of the SU(2) gauge potential and the topological properties on the four-manifold.

In this paper, we will establish a general decomposition theory of the SU(2) gauge potential in terms of the unit vector **n** on the compact and oriented four-dimensional manifold. Also, by means of geometric algebra, we can describe the unit vector **n** as an element of Spin(3).[19] One can show that the general decomposition formula of SU(2) gauge potential has a global property. The Chern density of the principal $P(\pi, M, \mathrm{SU}(2))$ will be studied by using the decomposition formula. One shows that the SU(2) Chern density takes the form of the δ-function $\delta(\boldsymbol{\phi})$. The topological structure of the SU(2) Chern density can be labeled by the Brouwer degrees and the Hopf index. Further research shows that there exist the crucial cases of branch process in the topological

[a]Electronic mail: ysduan@lzu.edu.cn
[b]Corresponding author. Electronic mail: itp2@lzu.edu.cn

 © 1998 American Institute of Physics

density, when the Jacobian $D(\phi/x)=0$. We calculate out the different branches of the Chern density by using the implicit function theorem. It is pointed out that the topological charges will be split at the critical points.

This paper is arranged in five sections. In Sec. II we will study a general decomposition theory of SU(2) gauge potential on a sphere bundle. In Sec. III we will study the topological structure of the SU(2) Chern density by using the decomposition expression given in Sec. II. The bifurcation of the Chern density will be studied in Sec. IV. Last, we present our conclusion.

II. THE DECOMPOSITION THEORY OF SU(2) GAUGE POTENTIAL

In this section we will give the decomposition theory of SU(2) gauge potential in terms of the sphere bundle on a compact and oriented four-dimensional manifold. First we have to give the basic notions which are necessary for our discussions.

Let V be a unit SU(2) Clifford vector

$$V=V^a\sigma_a, \quad a=1,2,3, \tag{1}$$

and

$$V^aV^a=1 \tag{2}$$

in which σ_a are Pauli matrixes and the base of SU(2) Clifford algebra.[20]

The covariant derivative one-form of V is given by

$$DV=dV-[A,V], \tag{3}$$

where A is the SU(2) gauge potential one-form:

$$A=i/2A^a\sigma_a \tag{4}$$

and

$$A^a=A^a_\mu dx^\mu, \quad \mu=0,1,2,3. \tag{5}$$

In gauge theory, the potential one-form undergoes the gauge transformation:

$$A'=SAS^{-1}+dSS^{-1}. \tag{6}$$

In our viewpoint,[6] the gauge potential one-form A can be decomposed and has inner structure. The main feature of the decomposition theory of the gauge potential is that the gauge potential A can be generally decomposed as

$$A=a+b, \tag{7}$$

where a and b are required to satisfy the gauge transformation and vector covariant transformation rules, i.e.,[19]

$$a'=SaS^{-1}+dSS^{-1} \tag{8}$$

and

$$b=SbS^{-1}. \tag{9}$$

From (8) and (9), one can show that the gauge potential A rigorously satisfies the gauge transformation

$$A'=a'+b'=S(a+b)S^{-1}+dSS^{-1}.$$

Let $V_{(i)}$ $(i=1,2,3)$ be an orthonormal basis of SU(2) Clifford vector with the orthogonal relations

$$V_{(i)} \cdot V_{(j)} = \delta_{ij}, \tag{10}$$

i.e.,

$$V^a_{(i)} V^a_{(j)} = \delta_{ij},$$

where $V_{(i)} \cdot V_{(j)}$ is the Clifford scalar product defined by[19]

$$V_{(i)} \cdot V_{(j)} = \tfrac{1}{2}(V_{(i)} V_{(j)} + V_{(j)} V_{(i)}). \tag{11}$$

Since the potential A is also a SU(2) Clifford vector, we have the projection formula

$$A = (A \cdot V_{(i)}) V_{(i)}. \tag{12}$$

Substituting this formula into (3), we obtain

$$DV_{(i)} = dV_{(i)} - (A \cdot V_{(j)})[V_{(j)}, V_{(i)}]. \tag{13}$$

Using (10), (11) and (13), and considering that

$$[V_{(j)}, V_{(i)}] = 2V_{(j)} V_{(i)} - 2V_{(i)} \cdot V_{(j)}, \tag{14}$$

we can rewrite A as

$$A = 1/4 dV_{(i)} V_{(i)} - 1/4 DV_{(i)} V_{(i)}. \tag{15}$$

According to (7), we define that

$$a = \tfrac{1}{4} dV_{(i)} V_{(i)}, \tag{16}$$

and

$$b = \tfrac{1}{4} DV_{(i)} V_{(i)}. \tag{17}$$

It is easy to prove that the decomposition formula above satisfies the requirement in (6), even has global property, and is independent from the local coordinates.

Let a family $\{W, V, U, \ldots\}$ be an open cover of M and S_{uv} be the transition matrix function which satisfy the following condition:[21]

$$S_{uu=1}, \quad S^{-1}_{vu} = S_{uv}, \quad S_{wv} S_{vu} S_{uw} = I; \quad W \cap V \cap U \neq \varnothing. \tag{18}$$

For any two open neighborhoods V and U, if $V \cap U \neq \varnothing$, then

$$V_{(i)v} = S_{vu} V_{(i)} S^{-1}_{vu}, \tag{19}$$

where the subscripts ''$_u$'' and ''$_v$'' represent the open covers U and V correspondingly, and we know that the gauge potential A undergoes the gauge transformation

$$A_v = S_{vu} A_u S^{-1}_{vu} + dS_{vu} S^{-1}_{vu}, \tag{20}$$

which is the fundamental condition for the existence of the gauge potential on the principal $P(\pi, M, \mathrm{SU}(2))$. In the physics terminology S_{vu} is just the gauge transformation (6) in the gauge theory. In the following, for abbreviation, we shall use the notation $S_{vu} = S$. According to (19), we deduce that

$$dV_{(i)v} V_{(i)v} = S dV_{(i)u} V_{(i)u} S^{-1} + dSS^{-1} - V_{(i)v} dSS^{-1} V_{(i)v}. \tag{21}$$

Considering that A is a vector of Clifford algebra, we can see that dSS^{-1} is a vector of the Clifford algebra. Also, then, there exists the following formula:[19]

$$V_{(i)v} dSS^{-1} V_{(i)v} = -3 dSS^{-1}. \tag{22}$$

Noticing that

$$DV_{(i)v}=SDV_{(i)u}S^{-1},$$

from (15), and making use of (19), (21) and (22), one can easily obtain

$$A_v-\tfrac{1}{4}(dV_{(i)v}V_{(i)v}-DV_{(i)v}V_{(i)v})=S[A_u-\tfrac{1}{4}(DV_{(i)u}V_{(i)u}-DV_{(i)u}V_{(i)u})]S^{-1}. \tag{23}$$

The expression above shows that if the decomposition formula on the open neighborhood U,

$$A_v=\tfrac{1}{4}(dV_{(i)v}V_{(i)v}-DV_{(i)v}V_{(i)v}),$$

holds true, the decomposition formula on the open neighborhood U,

$$A_u=\tfrac{1}{4}(dV_{(i)u}V_{(i)u}-DV_{(i)u}V_{(i)u}),$$

must hold true, too. This means that the general decomposition formula (15) has a global property and is independent from the choice of the local coordinates.

We know that the characteristic class is the fundamental topological property, and it is independent of the gauge potential.[22] So, to discuss the Chern class, we can take A as

$$A=\tfrac{1}{4}dU_{(i)}U_{(i)}. \tag{24}$$

One can regard it as a special gauge. By use of this formula, the magnetic monopole can be studied as in Refs. 6–8. In this paper we want to discuss the topological properties of a four-dimensional manifold, a decomposition formula in terms of the sphere bundle over the manifold will be convenient.

Now, let us return to discuss the decomposition theory on the sphere bundle of a compact and oriented four-dimensional manifold. Let M be a compact and oriented four-dimensional manifold and $P(\pi,M,G)$ be a principal bundle with the structure group $G=\mathrm{SU}(2)$. A smooth vector field ϕ^A $(A=0,1,2,3)$ can be found on the base manifold M. We define a unit vector field on M as

$$\begin{gathered} n^A=\phi^A/\|\phi\|, \quad A=0,1,2,3, \\ \|\phi\|=\sqrt{\phi^A\phi^A}, \end{gathered} \tag{25}$$

in which the superscripts ''A'' are the local orthonormal frame index.

In fact, **n** is identified as a section of the sphere bundle over M (or a partial section of the vector bundle over M).[5] We see that the zeros of $\boldsymbol{\phi}$ are just the singular points of **n**. Since the global property of a manifold has close relation with zeros of a smooth vector fields on it, this expression of the unit vector **n** is a very powerful tool in the discussion of the global topology. It naturally guarantees the constraint

$$n^An^A=1. \tag{26}$$

We can express the unit vector **n** in terms of the Clifford algebra[23] as

$$n=n^As_A, \quad A=0,1,2,3, \tag{27}$$

where

$$s=(I,i\boldsymbol{\sigma}), \quad s^\dagger=(I,-i\boldsymbol{\sigma}). \tag{28}$$

We know that n is just an element of Spin(3) in terms of geometric algebra.[19] We can rewrite (26) as

$$nn^\dagger=1. \tag{29}$$

It is easy to see that n has three independent components. Since the orthonormal basis of the SU(2) Clifford one-vectors just have three independent components, we can express $U_{(i)}$ in terms

of an element of Spin(3). From gauge transformations of the orthonormal basis given in Ref. 24, and the spinorial transformation given in Ref. 25, one can expand $U_{(i)}$ in terms of an element of Spin(3) as

$$U^a_{(i)}=2(n^0)^2\delta^{ia}+2n^0n^b\in^{iba}+2n^in^a-\delta^{ia}. \tag{30}$$

It is easy to prove

$$U^a_{(i)}U^a_{(j)}=\delta_{ij}. \tag{31}$$

Using (30) and (24), as a result, the gauge potential A will be expressed as

$$A=dnn^\dagger. \tag{32}$$

Because n is a unit element of the sphere bundle over M, we will discuss the topological property of the principle $P(\pi,M,\mathrm{SU}(2))$ directly by using this formula.

III. THE SU(2) CHERN DENSITY AND ITS INNER STRUCTURE

It is well known that gauge potential (connection) and gauge field (curvature) play essential roles in discussing the topological properties of a manifold. For the principal $P(\pi,M,\mathrm{SU}(2))$ the SU(2) Chern density is an important topological characteristic. In this section we will discuss the SU(2) Chern density by using the decomposition formula (32), which we have just given in the last section.

We know that the second Chern class is the fundamental characteristic of the principal $P(\pi,M,\mathrm{SU}(2))$. It is denoted as

$$C_2(P)=\frac{1}{8\pi^2}\mathrm{Tr}(F\wedge F). \tag{33}$$

The curvature F is defined as

$$F=dA-A\wedge A. \tag{34}$$

One can show that the gauge field F is a generalized function when there exist m singular points z_i $(i=1,2,\ldots,m)$ in the unit vector field n. Let us substitute the formula (32) into (34). Then

$$F\begin{cases}=0 & \text{when } x\neq z_i,\\ \neq 0 & \text{when } x=z_i,\end{cases} \tag{35}$$

where z_i are the singular points of n. This means that the gauge field F vanishes at the region where n has no singular points, but at the singular points of n the gauge field F does not vanish. This feature we will show soon is due to the nontriviality property of the principal $P(\pi,M,\mathrm{SU}(2))$, and is essential to study some physical problem.

The second SU(2) Chern class can be rewrite in terms of Chern Simon:[22,26]

$$C_2(P)=\frac{1}{8\pi^2}d\Omega \tag{36}$$

and

$$\Omega=\frac{1}{8\pi^2}\mathrm{Tr}\left(A\wedge dA-\frac{2}{3}A\wedge A\wedge A\right), \tag{37}$$

which is known as the Chern–Simon form.[27]

Substituting (32) into (36) and considering (29), we obtain

$$C_2(P)=\frac{1}{24\pi^2}\mathrm{Tr}(dn\wedge dn^\dagger\wedge dn\wedge dn^\dagger). \tag{38}$$

In detail, substituting (27) into the formula above,

$$C_2(P)=\frac{1}{24\pi^2}\in^{\mu\nu\lambda\rho}\partial_\mu n^A\partial_\nu n^B\partial_\lambda n^C\partial\rho n^D Tr(s_A s_B^\dagger s_C s_D^\dagger)dx^4$$

$$=\frac{1}{12\pi^2}\in^{\mu\nu\lambda\rho}\in_{ABCD}\partial_\mu n^A\partial_\nu n^B\partial_\lambda n^C\partial_\rho n^D dx^4. \tag{39}$$

By substituting (25) into (39), considering

$$dn^A=\frac{d\phi^A}{\|\phi\|}+\phi^A d\left(\frac{1}{\|\phi\|}\right), \tag{40}$$

we have

$$C_2(P)=-\frac{1}{4\pi^2}\frac{\partial^2}{\partial\phi^A\partial\phi^A}\left(\frac{1}{\|\phi\|^2}\right)D(\phi/x)d^4x, \tag{41}$$

where $D(\phi/x)$ is the Jacobian defined as

$$\epsilon^{ABCD}D(\phi/x)=\epsilon^{\mu\nu\lambda\rho}\partial_\mu\phi^A\partial\nu\phi^B\partial_\lambda\phi^C\partial_\rho\phi^D. \tag{42}$$

By means of the general Green's function formula

$$\frac{\partial^2}{\partial\phi^A\partial\phi^A}\left(\frac{1}{\|\phi\|^2}\right)=-4\pi\delta^4(\boldsymbol{\phi}), \tag{43}$$

we have

$$C_2(P)=\delta^4(\boldsymbol{\phi})D(\phi/x)d^4x. \tag{44}$$

Suppose $\phi^A(x)$ $(A=0,1,2,3)$ have m isolated zeros at $x_\mu=z_\mu^i$ $(i=1,2,\ldots,m)$. According to the δ-function theory,[28] $\delta(\boldsymbol{\phi})$ can be expressed by

$$\delta(\boldsymbol{\phi})=\sum_{i=1}^{m}\frac{\beta_i\delta(\mathbf{x}-\mathbf{z}_i)}{|D(\phi/x)|_{\mathbf{x}=\mathbf{z}_i}}, \tag{45}$$

and one then obtains

$$C_2(P)=\sum_{i=1}^{m}\eta_i\beta_i\delta^4(x-z_i)d^4x, \tag{46}$$

where β_i is a positive integer (the Hopf index of the ith zeros) and η_i is the Brouwer degree:[29]

$$\eta_i=\frac{D(\phi/x)}{|D(\phi/x)|}=\mathrm{sgn}[D(\phi/x)]\big|_{x=z_i}=\pm 1. \tag{47}$$

The meaning of the Hopf index β_i is that the vector field function $\boldsymbol{\phi}$ covers the corresponding region β_i times while $\mathbf{x}$ covers the region of the neighborhood of zero z_i once. From the above discussion, the Chern density $\rho(M)$ is defined as

$$\rho(M)=\sum_{i=1}^{m}\eta_i\beta_i\delta^4(x-z_i), \tag{48}$$

which shows that the topological structure of Chern density ρ is labeled by the Brouwer degrees and the Hopf index. The integration of $\rho(M)$ on M,

$$C_2=\int_M \rho(M)d^4x=\sum_{i=1}^{m}\eta_i\beta_i, \tag{49}$$

is an integer called the Chern number, which is a topological invariant of M.

The result (49) suggests that the zero points of the smooth vector $\boldsymbol{\phi}$ are essential to the topological properties of the base manifold M. On the other hand, the density (48) can be regarded as the density of a system of k classical pointlike particles with topological invariant charges $g_i=\eta_i\beta_i$ on the four-dimensional manifold.

IV. THE BIFURCATION OF CHERN DENSITY

As discussed before, the zeros of the smooth vector $\boldsymbol{\phi}$ play important roles in studying the Chern class of the manifold M. In this section, we will study the properties of the zero points, in other words, the properties of the following equation solutions

$$\begin{aligned}&\phi^0(x^0,x^1,x^2,x^3)=0,\quad \phi^1(x^0,x^1,x^2,x^3)=0,\\ &\phi^2(x^0,x^1,x^2,x^3)=0,\quad \phi^3(x^0,x^1,x^2,x^3)=0.\end{aligned} \tag{50}$$

As we knew before, if the Jacobian determinant

$$D\left(\frac{\phi}{x}\right)=\frac{\partial(\phi^0,\phi^1,\phi^2,\phi^3)}{\partial(x^0,x^1,x^2,x^3)}\neq 0,$$

we will have the isolated solutions of (50). The isolated solutions are called regular points. It is easy to see that the result in Sec. III is based on this condition. However, when this condition fails, the above results will change in some way, and will lead to the branch process of topological density and give rise to the bifurcation.

In order to show this case easily, we want to have some suppose. It is convenient to let $x^0=t$, and suppose $\phi^0=\phi^0(t)$. We denote one of the zero points as $(t^*,\mathbf{z}_i)$. We know that if

$$D\left(\frac{\boldsymbol{\phi}}{\mathbf{x}}\right)\Big|_{(t^*,\mathbf{z}_i)}=\frac{\partial(\phi^1,\phi^2,\phi^3)}{\partial(x^1,x^2,x^3)}\Big|_{(t^*,\mathbf{z}_i)}=0,$$

the Jacobian $D(\phi/x)|_{(t^*,\mathbf{z}_i)}=0$ will be obtained automatically. We will show that this case will lead to the branch process of topological density. In this case, Eqs. (50) will be rewritten as

$$\phi^0(t)=0, \tag{51}$$

and

$$\phi^1(t,x^1,x^2,x^3)=0,\quad \phi^2(t,x^1,x^2,x^3)=0,\quad \phi^3(t,x^1,x^2,x^3)=0. \tag{52}$$

It is well known that when the Jacobian $D(\boldsymbol{\phi}/\mathbf{x})|_{(t^*,\mathbf{z}_i)}=0$, the usual implicit function theorem is of no use. However, if the Jacobian

$$D^1\left(\frac{\boldsymbol{\phi}}{x}\right)\Big|_{(t^*,\mathbf{z}_i)}=\frac{\partial(\phi^1,\phi^2,\phi^3)}{\partial(t,x^2,x^3)}\Big|_{(t^*,\mathbf{z}_i)}\neq 0,$$

we can use the Jacobian $D^1(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)}$ instead of $D(\boldsymbol{\phi}/\mathbf{x})|_{(t^*,\mathbf{z}_i)}=0$, for the purpose of using the implicit function theorem.[30] Then we have a unique solution of Eqs. (52) in the neighborhood of the points $(t^*,\mathbf{z}_i)$:

$$t=t(x^1),\quad x^i=x^i(x^1),\quad i=2,3, \tag{53}$$

with $t^*=t(x^1)$. We call the critical points $(t^*,\mathbf{z}_i)$ the limit points. In the present case, it is easy to know that

$$\left.\frac{dx^1}{dt}\right|_{(t^*,\mathbf{z}_i)} = \frac{D^1(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)}}{D(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)}} = \infty,$$

i.e.,

$$\left.\frac{dt}{dx^1}\right|_{(t^*,\mathbf{z}_i)} = 0.$$

Then we have the Taylor expansion of (53) at the point $(t^*,\mathbf{z}_i)$:

$$t = t^* + \left.\frac{dt}{dx^1}\right|_{(t^*,\mathbf{z}_i)} (x^1 - z_i^1) + \frac{1}{2}\left.\frac{d^2t}{(dx^1)^2}\right|_{(t^*,\mathbf{z}_i)} (x^1 - z_i^1)^2 = t^* + \frac{1}{2}\left.\frac{d^2t}{(dx^1)^2}\right|_{(t^*,\mathbf{z}_i)} (x^1 - z_i^1)^2.$$

Therefore,

$$t - t^* = \frac{1}{2}\left.\frac{d^2t}{(dx^1)^2}\right|_{(t^*,\mathbf{z}_i)} (x^1 - z_i^1)^2, \tag{54}$$

which is a parabola in the x^1-t plane. From (54), we can obtain the two solutions $x_1^1(t)$ and $x_2^1(t)$, which give the branch solutions of the system (50). If $d^2t/(dx^1)^2|_{(t^*,\mathbf{z}_i)} > 0$, we have the branch solutions for $t > t^*$; otherwise, we have the branch solutions for $t > t^*$. These two condition are related to the origin and annihilation of topological charges.[31] Since the Chern number (49) is a topological invariant, the topological numbers of these two must be opposite at the zero point, i.e.,

$$\beta_{i1}\eta_{i1} = -\beta_{i2}\eta_{i2}.$$

For a limit point, it also requires the $D^1(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)} \neq 0$. As for a bifurcation point,[32] it must satisfy a more complete condition. This case will be discussed in the following subsections in detail.

A. The branch process at the bifurcation point

In this subsection, we have the restrictions of the system (50) at the bifurcation point $(t^*,\mathbf{z}_i)$:

$$\left.D\left(\frac{\boldsymbol{\phi}}{\mathbf{x}}\right)\right|_{(t^*,\mathbf{z}_i)} = 0, \quad \left.D^1\left(\frac{\boldsymbol{\phi}}{x}\right)\right|_{(t^*,\mathbf{z}_i)} = 0. \tag{55}$$

These will lead to an important fact that the function relationship between t and x^1 is not unique in the neighborhood of the bifurcation point $(\mathbf{z}_i, t^*)$. It is easy to see from the equation

$$\left.\frac{dx^1}{dt}\right|_{(t^*,\mathbf{z}_i)} = \frac{D^1(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)}}{D(\boldsymbol{\phi}/x)|_{(t^*,\mathbf{z}_i)}}, \tag{56}$$

which, under the restraint (55), directly shows that the direction of the integral curve of Eq. (56) is indefinite at the point $(\mathbf{z}_i, t^*)$. This is why the very point $(\mathbf{z}_i, t^*)$ is called a bifurcation point of the system (50).

Next, we will find a simple way to search for the different directions of all branch curves at the bifurcation point. Assume that the bifurcation point $(\mathbf{z}_i, t^*)$ has been found from (52). We know that, at the bifurcation point $(\mathbf{z}_i, t^*)$, the rank of the Jacobian matrix $[\partial\boldsymbol{\phi}/\partial x]$ is smaller than 3. First, we suppose the rank of the Jacobian matrix $[\partial\boldsymbol{\phi}/\partial x]$ is 2 (the case of a more smaller rank will be discussed later). Suppose that the 2×2 submatrix $J_1(\phi/x)$ is

$$J_1\left(\frac{\phi}{x}\right)=\begin{pmatrix} \frac{\partial\phi^1}{\partial x^2} & \frac{\partial\phi^1}{\partial x^3} \\ \frac{\partial\phi^2}{\partial x^2} & \frac{\partial\phi^2}{\partial x^3} \end{pmatrix}, \tag{57}$$

and its determinant $D_1(\phi/x)$ does not vanish. The implicit function theorem says that there exists one and only one function relation

$$x^i=f^i(x^1,t), \quad i=2,3. \tag{58}$$

We denoted the partial derivatives as

$$f^i_1=\frac{\partial f^i}{\partial x^1}; \quad f^i_t=\frac{\partial f^i}{\partial t}; \quad f^i_{11}=\frac{\partial^2 f^i}{\partial x^1\partial x^1}; \quad f^i_{1t}=\frac{\partial^2 f^i}{\partial x^1\partial x^t}; \quad f^i_{tt}=\frac{\partial f^i}{\partial x^t\partial x^t}.$$

From (50) and (58) we have, for $a=1,2,3$,

$$\phi^a=\phi^a(x^1,f^2(x^1,t),f^3(x^1,t),t)=0, \tag{59}$$

which give

$$\frac{\partial\phi^a}{\partial x^1}=\phi^a_1+\sum_{j=2}^{3}\frac{\partial\phi^a}{\partial f^j}\frac{\partial f^j}{\partial x^1}=0, \tag{60}$$

$$\frac{\partial\phi^a}{\partial t}=\phi^a_t+\sum_{j=2}^{3}\frac{\partial\phi^a}{\partial f^j}\frac{\partial f^j}{\partial t}=0, \tag{61}$$

from which we can get the first-order derivatives of f^i: f^i_1 and f^i_t. Differentiating (60) with respect to x^1 and t, respectively, we get

$$\sum_{j=2}^{3}\phi^a_j f^j_{11}=-\sum_{j=2}^{3}\left[2\phi^a_{j1}f^j_1+\sum_{k=2}^{3}(\phi^a_{jk}f^k_1)f^j_1\right]-\phi^a_{11}, \quad a=1,2,3, \tag{62}$$

$$\sum_{j=2}^{3}\phi^a_j f^j_{1t}=-\sum_{j=2}^{3}\left[\phi^a_{jt}f^j_1+\phi^a_{j1}f^j_t+\sum_{k=2}^{3}(\phi^a_{jk}f^k_t)f^j_1\right]-\phi^a_{1t}, \quad a=1,2,3. \tag{63}$$

The differentiation of (61) with respect to t gives

$$\sum_{j=2}^{3}\phi^a_j f^j_{tt}=-\sum_{j=2}^{3}\left[2\phi^a_{jt}f^j_t+\sum_{k=2}^{3}(\phi^a_{jk}f^k_t)f^j_t\right]-\phi^a_{tt}, \quad a=1,2,3, \tag{64}$$

where

$$\phi^a_{jk}=\frac{\partial^2\phi^a}{\partial x^j\partial x^k}, \quad \phi^a_{jt}=\frac{\partial^2\phi^a}{\partial x^j\partial t}. \tag{65}$$

The differentiation of (61) with respect to x^1 gives the same expression as (63). By making use of the Gaussian elimination method to (62), (63) and (64) we can find the second-order derivatives f^i_{11}, f^i_{1t} and f^i_{tt}. The above discussion does not matter to the last component $\phi^3(\mathbf{x},t)$. In order to find the different values of dx^1/dt at the bifurcation point $(\mathbf{z}_i,t^*)$, let us investigate the Taylor expansion of $\phi^3(\mathbf{x},t)$ in the neighborhood of $(\mathbf{z}_i,t^*)$. Substituting (58) into $\phi^3(\mathbf{x},t)$ we have the function of two variables x^1 and t,

$$F(x^1,t)=\phi^3(x^1,f^2(x^1,t),f^3(x^1,t),t), \tag{66}$$

which, according to (52), must vanish at the bifurcation point

$$F(z_i^1, t^*) = 0. \tag{67}$$

From (66) we have the first-order partial derivatives of $F(x^1,t)$:

$$\frac{\partial F}{\partial x^1} = \phi_1^3 + \sum_{j=2}^{3} \phi_j^3 f_1^j, \quad \frac{\partial F}{\partial t} = \phi_t^3 + \sum_{j=2}^{3} \phi_j^3 f_t^j. \tag{68}$$

Using (60) and (61), the first equation of (55) is expressed as

$$D\left(\frac{\boldsymbol{\phi}}{x}\right)\Big|_{(\mathbf{z}_i, t^*)} = \begin{vmatrix} -\sum_{j=2}^{3} \phi_j^1 f_1^j & \phi_2^1 & \phi_3^1 \\ -\sum_{j=2}^{3} \phi_j^2 f_1^j & \phi_2^2 & \phi_3^2 \\ \phi_1^3 & \phi_2^3 & \phi_3^3 \end{vmatrix}_{(\mathbf{z}_i, t^*)} = 0, \tag{69}$$

which, by Cramer's rule, can be written as

$$D\left(\frac{\boldsymbol{\phi}}{x}\right)\Big|_{(\mathbf{z}_i, t^*)} = \frac{\partial F}{\partial x^1} \det J_1\left(\frac{\phi}{x}\right)\Big|_{(\mathbf{z}_i, t^*)} = 0.$$

Since $\det J_1(\phi/x)|_{(\mathbf{z}_i, t^*)} \neq 0$, the above equation gives

$$\frac{\partial F}{\partial x^1}\Big|_{(\mathbf{z}_i, t^*)} = 0. \tag{70}$$

With the same reasons, we have

$$\frac{\partial F}{\partial t}\Big|_{(\mathbf{z}_i, t^*)} = 0. \tag{71}$$

The second-order partial derivatives of the function F are easily found to be

$$\frac{\partial^2 F}{(\partial x^1)^2} = \phi_{11}^3 + \sum_{j=2}^{3} \left[2\phi_{1j}^3 f_1^j + \phi_j^3 f_{11}^j + \sum_{k=2}^{3} (\phi_{kj}^3 f_1^k) f_1^j \right], \tag{72}$$

$$\frac{\partial^2 F}{\partial x^1 \partial t} = \phi_{1t}^3 + \sum_{j=2}^{3} \left[\phi_{1j}^3 f_t^j + \phi_{tj}^3 f_1^j + \phi_j^3 f_{1t}^j + \sum_{k=2}^{3} (\phi_{jk}^3 f_t^k) f_1^j \right], \tag{73}$$

$$\frac{\partial^2 F}{\partial t^2} = \phi_{tt}^3 + \sum_{j=2}^{3} \left[2\phi_{jt}^3 f_t^j + \phi_j^3 f_{tt}^j + \sum_{k=2}^{3} (\phi_{jk}^3 f_t^k) f_t^j \right], \tag{74}$$

which at $(\mathbf{z}_i, t^*)$ are denoted by

$$A = \frac{\partial^2 F}{(\partial x^1)^2}\Big|_{(\mathbf{z}_i, t^*)}, \quad B = \frac{\partial^2 F}{\partial x^1 \partial t}\Big|_{(\mathbf{z}_i, t^*)}, \quad C = \frac{\partial^2 F}{\partial t^2}\Big|_{(\mathbf{z}_i, t^*)}. \tag{75}$$

Then taking notice of (67), (70), (71) and (75), the Taylor expansion of $F(x^1,t)$ in the neighborhood of the bifurcation point $(\mathbf{z}_i, t^*)$ can be expressed as

$$F(x^1,t) = \tfrac{1}{2}A(x^1 - z_i^1)^2 + B(x^1 - z_i^1)(t - t^*) + \tfrac{1}{2}C(t - t^*)^2, \tag{76}$$

which by (66) is the expression of $\phi^3(\mathbf{x},t)$ in the neighborhood of $(\mathbf{z}_i,t^*)$. The expression (76) is reasonable, which shows that at the bifurcation point $(\mathbf{z}_i,t^*)$ one of the equations (50), $\phi^3(\mathbf{x},t)=0$, is satisfied, i.e.,

$$A(x^1-z_i^1)^2+2B(x^1-z_i^1)(t-t^*)+C(t-t^*)^2=0. \tag{77}$$

Dividing (77) by $(t-t^*)^2$ and taking the limit $t\to t^*$ as well as $x^1\to z_i^1$, we get

$$A\left(\frac{dx^1}{dt}\right)^2+2B\frac{dx^1}{dt}+C=0. \tag{78}$$

In the same way we have

$$C\left(\frac{dt}{dx^1}\right)^2+2B\frac{dt}{dx^1}+A=0. \tag{79}$$

The different directions of the branch curves at the bifurcation point are determined by (78) or (79). It is easy to see that there are at most two different branches.

B. The branch process at a higher degenerated point

In Sec. IV A we have studied the case that the rank of the Jacobian matrix $[\partial\phi/\partial x]$ of Eqs. (50) is $2=3-1$. In this subsection, we consider the case that the rank of the Jacobian matrix is $1=3-2$. Let the $J_2(\boldsymbol{\phi}/x)=\partial\phi^1/\partial x^1$ and suppose that $\det J_2\neq 0$. With the same reasons as when obtaining (58), we can have the function relations

$$x^3=f^3(x^1,x^2,t). \tag{80}$$

Substituting the relations (80) into the last two equations of (50), we have the following two equations with three arguments x^1,x^2,t:

$$\begin{aligned} F_1(x^1,x^2,t)&=\phi^2(x^1,x^2,f^3(x^1,x^2,t),t)=0,\\ F_2(x^1,x^2,t)&=\phi^3(x^1,x^2,f^3(x^1,x^2,t),t)=0. \end{aligned} \tag{81}$$

Calculating the partial derivatives of the function F_1 and F_2 with respect to x^1, x^2 and t, taking notice of (80) and using six similar expressions to (70) and (71), i.e.,

$$\left.\frac{\partial F_j}{\partial x^1}\right|_{(\mathbf{z}_i,t^*)}=0,\quad \left.\frac{\partial F_j}{\partial x^2}\right|_{(\mathbf{z}_i,t^*)}=0,\quad \left.\frac{\partial F_j}{\partial t}\right|_{(\mathbf{z}_i,t^*)}=0,\quad j=1,2, \tag{82}$$

we have the following forms of Taylor expressions of F_1 and F_2 in the neighborhood of $(\mathbf{z}_i,t^*)$:

$$\begin{aligned} F_j(x^1,x^2,t)\approx{}&A_{j1}(x^1-z_i^1)^2+A_{j2}(x^1-z_i^1)(x^2-z_i^2)+A_{j3}(x^1-z_i^1)\\ &\times(t-t^*)+A_{j4}(x^2-z_i^2)^2+A_{j5}(x^2-z_i^2)(t-t^*)+A_{j6}(t-t^*)^2=0,\quad j=1,2. \end{aligned} \tag{83}$$

In the case of $A_{j1}\neq 0, A_{j4}\neq 0$, dividing (83) by $(t-t^*)^2$ and taking the limit $t\to t^*$, we obtain two quadratic equations of dx^1/dt and dx^2/dt:

$$A_{j1}\left(\frac{dx^1}{dt}\right)^2+A_{j2}\frac{dx^1}{dt}\frac{dx^2}{dt}+A_{j3}\frac{dx^1}{dt}+A_{j4}\left(\frac{dx^2}{dt}\right)^2+A_{j5}\frac{dx^2}{dt}+A_{j6}=0,\quad j=1,2. \tag{84}$$

Eliminating the variable dx^1/dt, we obtain an equation of dx^2/dt in the form of a determinant,

$$\begin{vmatrix} A_{11} & A_{12}v+A_{23} & A_{14}v^2+A_{15}v+A_{16} & 0 \\ 0 & A_{11} & A_{12}v+A_{13} & A_{14}v^2+A_{15}v+A_{16} \\ A_{21} & A_{22}v+A_{23} & A_{24}v^2+A_{25}v+A_{26} & 0 \\ 0 & A_{21} & A_{22}v+A_{23} & A_{24}v^2+A_{25}v+A_{26} \end{vmatrix} = 0, \tag{85}$$

where $v=dx^2/dt$, that is a fourth-order equation of dx^2/dt:

$$a_0\left(\frac{dx^2}{dt}\right)^4+a_1\left(\frac{dx^2}{dt}\right)^3+a_2\left(\frac{dx^2}{dt}\right)^2+a_3\left(\frac{dx^2}{dt}\right)+a_4=0. \tag{86}$$

Therefore, we get different directions at the bifurcation point corresponding to different branch curves. The number of different branch curves is at most four. At the end of this section, we conclude that in our theory of topological density there exists the crucial case of branch process. This means that a normal point-like topological charge, which moves through the bifurcation point, may split into several point-like topological charges moving along different branch curves. Since the topological density is invariant, the total charge of the splitting topological particles must be precisely equal to the topological charge of the original particle.

Since the Chern density is identically conserved, the sum of the topological charges of these split topological densities must be equal to that of the original current at the bifurcation point. We suppose that there exist l different branches. Then the topological density of Chern class $\rho(M)$ changes in the following form:

$$\rho(M)=\sum_{i=1}^{k}\rho_i=\sum_{i=1}^{k}\sum_{j=1}^{l}\rho_{ij},$$

where

$$\rho_i=g_i\delta^4(x-z_i), \quad \rho_{ij}=g_{ij}\delta^4(x-z_{ij}), \quad 1\leqslant j\leqslant l. \tag{87}$$

With the same reason, the sum of the topological charges at the bifurcation points will have the following form:

$$g_i=\sum_{i=1}^{l}g_{ij}.$$

for fixed i.

V. CONCLUSION

We have explicitly constructed the gauge potential decomposition theory of SU(2) gauge theory in terms of the sphere bundle on a four-dimensional manifold. An important observation of the application of this theory is that the SU(2) Chern density takes the form of a general function. We also find that the Chern density has the bifurcation process. It is shown that the topological charges are split under this case, and the total charges are conserved. Those features are very important in discussing the topological problems on a four-manifold. Since many problems, not only in theoretical physics but also in differential geometry, are associated with the SU(2) gauge theory, the theory of SU(2) gauge potential decomposition, the inner structure of the SU(2) Chern density will provide important and powerful methods in those fields.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China.

[1] Y. S. Duan and S. L. Zhang, Int. J. Eng. Sci. **28**, 689 (1990); **29**, 153 (1991); **30**, 153 (1992).
[2] Y. S. Duan, S. L. Zhang, and S. S. Feng, J. Math. Phys. **35**, 4463 (1994).
[3] Y. S. Duan and X. H. Meng, J. Math. Phys. **34**, 1149 (1993); Y. S. Duan and X. G. Lee, Helv. Phys. Acta **68**, 513 (1995).
[4] Y. S. Duan, G. H. Yang, and Y. Jing, Helv. Phys. Acta. **70**, 565 (1997).

[5] J. S. Dowker and J. P. Schofield, J. Math. Phys. **31**, 808 (1990).
[6] Y. S. Duan and M. L. Ge, Kexue Tongbao **26**, 282 (1976); Sci. Sin. **11**, 1072 (1979); Y. S. Duan and S. C. Zhao, Commun. Theor. Phys. **2**, 1553 (1983).
[7] Y. S. Duan, SLAC-PUB-3301/84.
[8] Y. S. Duan and J. C. Liu, *Proceedings of Johns Hopkins Workshop 11*, edited by Y. S. Duan, G. Domokos, and S. Kovesi-Domokos (World Scientific, Singapore, 1988).
[9] J. M. F. Labastida and M. Marino, Nucl. Phys. B **456**, 633 (1996).
[10] J. Wess and B. Zumino, Phys. Lett. **37B**, 95 (1971).
[11] E. Witten, *Monopoles and Four-Manifold*, hep-th/9411102, IASSNS-HEP-94-96.
[12] E. Witten, J. Math. Phys. **35**, 5101 (1994).
[13] C. Michael and P. S. Spencer, Phys. Rev. D **52**, 4691 (1996).
[14] A. Galperin and E. Sokatchev, Report BONN-TH-94-27 24 (1994).
[15] P. M. Sotcliffe, Nucl. Phys. B **431**, 97 (1995).
[16] A. Actor, Rev. Mod. Phys. **51**, 461 (1979).
[17] V. de Alfaro, S. Fubini, and G. Furlan, Phys. Lett. B **73**, 463 (1978).
[18] S. Donaldson, Topology **29**, 257 (1990).
[19] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).
[20] S. Okubo, J. Math. Phys. **32**, 1657 (1991).
[21] S. S. Chern, An. da Acad. Brasileira de Ciéncias **35**, 17 (1963).
[22] S. Nash and S. Sen, *Topology and Geometry for Physicists* (Academic, London, 1983); M. W. Hirsch, *Differential Topology* (Springer-Verlag, New York, 1976).
[23] C. Doran, D. Hestenes, F. Sommen, and N. V. Acker, J. Math. Phys. **34**, 3642 (1993); A. Lasenby, C. Doran, and S. Gull, *ibid.* **34**, 3683 (1993).
[24] H. Boerner, *Representation of Groups* (North–Holland, Amsterdam, 1963).
[25] A. Dimakis and F. Müller-Hoisen, Class. Quantum Grav. **8**, 2093 (1991).
[26] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[27] S. S. Chern and J. Simons, Ann. Math. **99**, 48 (1974).
[28] A. S. Schwarz, *Topology For Physicists* (Springer-Verlag, New York, 1994).
[29] H. Hopf, Math. Ann. **96**, 209 (1929).
[30] E. Goursat, *A Course in Mathematical Analysis* Vol. 1 (translated by Earle Raymond Hedrick, 1904).
[31] N. Sakai, Phys. Rev. D **54**, 1548 (1996).
[32] M. Kubicek and M. Marek, *Computational Methods in Bifurcation Theory and Dissipative Structures* (Springer-Verlag, New York, 1983).

The elliptic quantum algebra $A_{q,p}(\widehat{sl_n})$ and its bosonization at level one

Heng Fan and Bo-yu Hou
Institute of Modern Physics, P.O. Box 105, Northwest University, Xian 710069, China[a)]

Kang-jie Shi and Wen-li Yang[b)]
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, China,
and Institute of Modern Physics, P.O. Box 105, Northwest University, Xian 710069, China

(Received 21 November 1997; accepted for publication 5 March 1998)

We extend the work of Foda *et al.* and propose an elliptic quantum algebra $A_{q,p}(\widehat{sl_n})$. Similar to the case of $A_{q,p}(\widehat{sl_2})$, our presentation of the algebra is based on the relation $RLL=LLR^*$, where R and R^* are Z_n symmetric R matrices with the elliptic moduli chosen differently, and a scalar factor is also involved. With the help of the results obtained by Asai *et al.*, we realize type I and type II vertex operators in terms of bosonic free fields for the Z_n symmetric Belavin model. We also give a bosonization for the elliptic quantum algebra $A_{q,p}(\widehat{sl_n})$ at level one.
© *1998 American Institute of Physics.* [S0022-2488(98)03709-8]

I. INTRODUCTION

The investigation of the symmetries in quantum integrable models has been attracting a great deal of interest. Recently, the quantum affine algebra $U_q(\widehat{sl_2})$ has been studied extensively and applied successfully to the XXZ model in the antiferromagnetic regime: see Ref. 1 and the references therein. The R matrix associated with the XXZ model is the six-vertex model which is a trigonometric vertex model. Using the approach of free boson realization of the vertex operators, Jimbo *et al.* obtained integral formulas for the correlation functions and the form factors for the XXZ model.

It is well known that we can obtain the six-vertex model from the Baxter's eight-vertex model[2] by taking a special limit. And we know that the XYZ model, the spin chain equivalent of the eight-vertex model, is a generalization of the XXZ model. Foda *et al.*[3] proposed an elliptic extention of the quantum affine algebra $A_{q,p}(\widehat{sl_2})$ as an algebra of symmetries for the eight-vertex model. The Kyoto group also conjecture that type I and type II vertex operators for the elliptic algebra $A_{q,p}(\widehat{sl_2})$ can be found. So, it is an open problem to give a bosonic free fields realization for vertex operators of the elliptic algebra $A_{q,p}(\widehat{sl_2})$. It is also an interesting problem to extend the elliptic algebra $A_{q,p}(\widehat{sl_2})$ to a more general case $A_{q,p}(\widehat{sl_n})$ which would play the role of the symmetry algebra in the Z_n Belavin model. In this paper, we will study these problems.

It is now believed that the vertex operator approach[1,4] and the method of bosonization are very powerful to study correlation functions of solvable lattice models. It is first formulated for vertex models, and then extended to incorporate face models.[5,6] Lukyanov and Pugai[6] successfully give a bosonic realization of the vertex operators for the ABF model.[7] This result is developed to a more general case by Asai *et al.*[8] They give a bosonization of vertex operators for the $A^{(1)}_{n-1}$ face model,[9] the ABF model being the case $n=2$ with a restricted condition. We know there is a face–vertex correspondence between the $A^{(1)}_{n-1}$ face model and the Z_n Belavin[10] vertex model. When $n=2$, the Z_n symmetric Belavin model reduces to the Baxter's eight-vertex model. In our former work,[11] we used the intertwiners of face–vertex correspondence, and obtained type I vertex operators for the Z_n Belavin vertex model with the help of the results obtained by Asai *et al.*[8] In the present paper, continuing to extend the results in Ref. 8 and our former work in Ref. 11, we give a bosonization of type II vertex operators for the Z_n Belavin model. Using the Miki's

a)Mailing address.
b)Electronic mail: wlyang@nwu.edu.cn; Fax: 0086-029-8303511; Telephone: 0086-029-8302986.

 © 1998 American Institute of Physics

construction,[12] we obtain a bosonic realization for the elliptic algebra $A_{q,p}(\widehat{sl_n})$ (which is first proposed in this paper) at level one. As a special case of $n=2$, this will give the bosonization for $A_{q,p}(\widehat{sl_2})$ algebra at level one.

The paper is organized as follows. In Sec. II we introduce the Z_n symmetric Belavin vertex model and give a definition for the elliptic algebra $A_{q,p}(\widehat{sl_n})$. In Sec. III we define the vertex operators for the algebra $A_{q,p}(\widehat{sl_n})$ at level one. Section IV is devoted to the main results of this paper. Extending the work of Asai *et al.*, we introduce another set of boson oscillators and obtain a bosonization of type II vertex operators for the $A_{n-1}^{(1)}$ face model. Using face–vertex correspondence, we give a free boson realization of type I and type II vertex operators for the Z_n Belavin model. This gives a bosonization of the elliptic algebra $A_{q,p}(\widehat{sl_n})$ at level one. Finally, we give a summary and discussions in Sec. V. The Appendix contains some detailed calculations.

II. THE MODEL AND THE ELLIPTIC ALGEBRA $A_{q,p}(\widehat{sl_n})$

We first introduce some notations. Let $n\in Z^+$ and $n\geqslant 2$, $w\in C$ and $\mathrm{Im}\, w\geqslant 0$, $r\geqslant n+2$ and take real value, $\tau\in C$ with $\mathrm{Im}\,\tau>0$.

Define matrix g,h,I_α with elements taking the values

$$g_{ij}=\omega^i\delta_{ij},\quad \omega=\exp\left\{\frac{2\pi i}{n}\right\},$$

$$h_{jk}=\delta_{j+1,k},\quad I_\alpha=I_{(\alpha_1,\alpha_2)}=g^{\alpha_2}h^{\alpha_1}.$$

The elliptic functions are defined as

$$\theta\begin{bmatrix}a\\ b\end{bmatrix}(z,\tau)=\sum_{m\in Z}\exp\{\pi i[(m+a)^2\tau+2(m+a)(z+b)]\},$$

$$\sigma_\alpha(z,\tau)=\sigma_{(\alpha_1,\alpha_2)}(z,\tau)=\theta\begin{bmatrix}\frac{1}{2}+\frac{\alpha_1}{n}\\ \frac{1}{2}+\frac{\alpha_2}{n}\end{bmatrix}(z,\tau).$$

Redefine $z\equiv vw$, $\tau\equiv rw$. The R matrix of the Z_n symmetric R matrix can be defined as

$$\bar{R}(vw,rw)=\frac{\sigma_0(w,rw)}{\sigma_0(vw+w,rw)}\sum_\alpha W_\alpha(vw,rw)I_\alpha\otimes I_\alpha^{-1}, \tag{1}$$

where

$$W_\alpha(wv,rw)=\frac{\sigma_\alpha\left(vw+\frac{w}{n},rw\right)}{n\sigma_\alpha\left(\frac{w}{n},rw\right)}. \tag{2}$$

It is necessary to introduce other notations,

$$R(v)\equiv R(vw,rw)=x^{2v(1/n-1)}\frac{g_1(v)}{g_1(-v)}\bar{R}(vw,rw), \tag{3}$$

where $x=e^{\pi iw}$ and

$$g_1(v)=\frac{\{x^{2v}x^2\}\{x^{2n+2r-2}x^{2v}\}}{\{x^{2n}x^{2v}\}\{x^{2r}x^{2v}\}}, \tag{4}$$

where $\{z\}=(z;x^{2r},x^{2n})$ and $(z;p_1,\ldots,p_n)\equiv\Pi_{\{n_i\}=0}^{\infty}(1-zp_1^{n_1}\cdots p_m^{n_m})$.

The R matrix $R(v,rw)$ has the following properties:

$$\text{Yang–Baxter equation: } R_{12}(v_1-v_2)R_{13}(v_1-v_3)R_{23}(v_2-v_3)$$
$$=R_{23}(v_2-v_3)R_{13}(v_1-v_3)R_{12}(v_1-v_2);$$
$$\text{unitary: } R_{12}(v)R_{21}(-v)=1; \tag{5}$$
$$\text{cross-unitarity: } \sum_{jl} R_{kj}^{li}(v)R_{i'l}^{jk'}(-v-n)=\delta_{ii'}\delta_{kk'}.$$

The parameters v,w,r in our parametrization for the Z_n symmetric R matrix in Eq. (3) are related to that of Foda *et al.*[8] as follows: $q=e^{i\pi w}=x$, $\xi=x^v$, $p=x^{2r}$.

Define

$$R^*(v)=\Delta_n^2(v)R(vw,(r-c)w),$$
$$\Delta_n(v)=-x[2(n-2)/n]v\,\frac{(x^{2n-2+2v};x^{2n})(x^{2-2v};x^{2n})}{(x^{2+2v};x^{2n})(x^{2n-2-2v};x^{2n})}. \tag{6}$$

Definition: The algebra $A_{q,p}(\widehat{gl_n})$ is generated by $L_{ij}(v)$, satisfying the relation

$$R^+(v_1-v_2)L_1(v_1)L_2(v_2)=L_2(v_2)L_1(v_1)R^{*+}(v_1-v_2), \tag{7}$$

where

$$R^+(v)=R(v)\tau^{-1}(-v+\tfrac{1}{2}),\quad R^{*+}(v)=R^*(v)\tau^{-1}(-v+\tfrac{1}{2}),$$
$$\tau(v)=x^{[2(1-n)/n]v}\,\frac{(x^{1+2v};x^{2n})(x^{2n-1-2v};x^{2n})}{(x^{2n-1+2v};x^{2n})(x^{1-2v};x^{2n})}. \tag{8}$$

By a standard argument based on the antisymmetric fusion for the Z_n symmetry R matrix,[10] we find that the following quantum determinant belongs to the center of $A_{q,p}(\widehat{gl_n})$:

$$q-\det L(v)=\sum_{\sigma\in S_n}\ \operatorname{sign}(\sigma)L_{1,\sigma(1)}(v-n)L_{2,\sigma(2)}(v-n+1)\cdots L_{n,\sigma(n)}(v-1). \tag{9}$$

Therefore, we can impose the further relation $q-\det L(v)=q^{c/2}$ and define the quotient algebra

$$A_{q,p}(\widehat{sl_n})=A_{q,p}(\widehat{gl_n})/(q-\det L(v)-q^{c/2}). \tag{10}$$

Remark: For the case $n=2$, we have $\Delta_n(v)=-1$. The algebra $A_{q,p}(\widehat{sl_n})$ reduces to the original elliptic algebra $A_{q,p}(\widehat{sl_2})$ proposed by Foda et al.[3]

III. BOSONIZATION OF THE ALGEBRA $A_{q,p}(\widehat{sl_n})$ AT LEVEL ONE

In the following, we will mainly restrict our attention to the level one case, and we have $c=1$. The $R^*(v)$ now becomes

$$R^*(v)=\Delta_n^2(v)R(v,(r-1)w). \tag{11}$$

The algebra relation remains as

$$R^+(v_1-v_2)L_1(v_1)L_2(v_2)=L_2(v_2)L_1(v_1)R^{*+}(v_1-v_2). \tag{12}$$

We first introduce a type I vertex operator corresponding to the half-column transfer matrix of the Z_n Belavin model,[5,11,13] and a type II vertex operator of the Z_n Belavin model which is

expected to create the eigenstates of the transfer matrix. We denote the two types of vertex operators as

(i) the vertex operator of type I: $\Phi_i(v)$;

(ii) the vertex operator of type II: $\Psi_i^*(v)$.

These Vertex operators satisfy the Faddeev–Zamolodchikov (ZF) algebra,

$$\Phi_j(v_2)\Phi_i(v_1)=\sum_{lk} R_{lk}^{ij}(v_1-v_2)\Phi_l(v_1)\Phi_k(v_2), \tag{13}$$

$$\Psi_i^*(v_1)\Psi_j^*(v_2)=\sum_{lk} \Psi_k^*(v_2)\Psi_l^*(v_1)R_{ij}^{*lk}(v_1-v_2)\Delta_n^{-1}(v_1-v_2), \tag{14}$$

$$\Phi_i(v_1)\Psi_j^*(v_2)=\tau^{-1}(v_1-v_2)\Psi_j^*(v_2)\Phi_i(v_1). \tag{15}$$

In the next section, we will give a q-boson free field realization of the type I and type II vertex operators listed above.

Introduce Miki's construction,[12]

$$L_{ij}(v)=\Phi_i(v)\Psi_j^*(v-\tfrac{1}{2}). \tag{16}$$

Using relations of the ZF algebra Eq. (13)–Eq. (15), one can prove that the operator matrix L constructed above satisfies the defining relation of the elliptic quantum algebra, Eq. (7). Here the relation $\tau(v+\frac{1}{2})/\tau(-v+\frac{1}{2})=\Delta_n^{-1}(v)$ has been used.

Thus, if the bosonization of the type I and type II vertex operators at level one satisfying Eq. (13)–Eq. (15) can be found, we can find a free boson realization of the elliptic algebra $A_{q,p}(\widehat{sl_n})$ at level one. Moreover, the elliptic algebra $A_{q,p}(\widehat{sl_n})$ and its vertex operators $\Phi_i(v),\Phi_j^*(v)$ at level one have the following relations, which is the generalized relations of Frenkel *et al.* in studying the q-affine algebra:[14]

$$\Phi_2(v_2)L_1(v_1)=R_{12}^{+}(v_1-v_2)L_1(v_1)\Phi_2(v_2),$$

$$L_1(v_1)\Psi_2^*(v_2-\tfrac{1}{2})=\Psi_2^*(v_2-\tfrac{1}{2})L_1(v_1)R_{12}^{*+}(v_1-v_2).$$

IV. BOSONIZATION FOR VERTEX OPERATORS

In this section, we will use a face–vertex correspondence relation to obtain a bosonization of vertex operators for the Z_n Belavin model satisfying Eq. (13)–Eq. (15).

A. Bosonization for the $A_{n-1}^{(1)}$ face model

We first give a brief review of the bosonization of type I vertex operators for the $A_{n-1}^{(1)}$ face model,[8] and then similarly construct the bosonization of type II vertex operators for the $A_{n-1}^{(1)}$ face model.

Let $\epsilon_\mu(1\leqslant\mu\leqslant n)$ be the orthonormal bases in R^n, which are supplied with the inner product $\langle\epsilon_\mu,\epsilon_\nu\rangle=\delta_{\mu\nu}$. Set

$$\overline{\epsilon}_\mu=\epsilon_\mu-\epsilon,\quad \epsilon=\frac{1}{n}\sum_{\mu=1}^{n}\epsilon_\mu.$$

The type $A_{n-1}^{(1)}$ weight Lattice is the linear span of the $\overline{\epsilon}_\mu$: $P=\Sigma_{\mu=1}^{n}Z\overline{\epsilon}_\mu$. Let ω_μ $(1\leqslant\mu\leqslant n-1)$ be the fundamental weights and $\alpha_\mu(1\leqslant\mu\leqslant n-1)$ be the simple roots: $\omega_\mu=\Sigma_{\nu=1}^{\mu}\overline{\epsilon}_\nu$, $\alpha_\mu=\epsilon_\mu-\epsilon_{\mu+1}$.

An ordered pair $(b,a)\in P^2$ is called admissible if and only if there exists μ $(1\leqslant\mu\leqslant n)$ such that $b-a=\overline{\epsilon}_\mu$. An ordered set of four weights $\binom{cd}{ba}\in P^4$ is called an admissible configuration around a face if and only if the pairs (b,a), (c,b), (d,a), and (c,d) are admissible. To each admissible configuration around a face, one can associate the Boltzmann weight for it.[2,7,9]

We introduce the zero mode operators $q_\mu,p_\mu(1\leqslant\mu\leqslant n)$, which satisfy

$$[p_\mu, iq_\nu] = \langle \epsilon_\mu, \epsilon_\nu \rangle = \delta_{\mu\nu}.$$

Set

$$Q_{\overline{\epsilon}_\mu} = q_\mu - \frac{1}{n}\sum_{l=1}^{n} q_l, \quad P_{\overline{\epsilon}_\mu} = p_\mu - \frac{1}{n}\sum_{l=1}^{n} p_l + \frac{1}{\sqrt{r(r-1)}} w_\mu,$$

where $\{w_j\}$ are generic complex numbers, which ensures that the intertwiners $\tilde{\varphi}(v)^{(k)}_{\mu,a}$ and $\overline{\varphi}(v)^{(k)}_{a,\mu}$ (see the following section) could exist. We can also reconstruct the zero mode operators P_α, Q_α (Ref. 8) indexed by $\alpha \in P$, which are Z linear in α and satisfy

$$[P_\alpha, iQ_\beta] = \langle \alpha, \beta \rangle \quad (\alpha, \beta \in P).$$

Now we consider the oscillator part. Define free bosonic oscillators β^j_m $(1 \leqslant j \leqslant n, m \in Z/\{0\})$ satisfying the relations

$$[\beta^j_m, \beta^k_n] = \begin{cases} m \dfrac{[(n-1)m]_x[(r-1)m]_x}{[nm]_x[rm]_x} \delta_{n+m,0}, & j=k, \\ -mx^{sign(j-k)nm} \dfrac{[m]_x[(r-1)m]_x}{[nm]_x[rm]_x} \delta_{n+m,0}, & j \neq k. \end{cases} \tag{17}$$

Here we have used the standard notation $[a]_x = (x^a - x^{-a})/(x - x^{-1})$. The bosonic oscillators should satisfy the constraint $\Sigma_{j=1}^{n} x^{-2jm}\beta^j_m = 0$. One can check that the above constraint is compatible with the commutation relations, Eq. (18). Similarly, we can also introduce another set of bosonic oscillators β'^j_m $(1 \leqslant j \leqslant n,\ m \in Z/\{0\})$ which will be used to construct the type II vertex operators and are related to the original bosons by $\beta'^j_m = \{[rm]_x/[(r-1)m]_x\}\beta^j_m$. Using β^j_m, β'^j_m, we define the operators,

$$S^j_m = (\beta^j_m - \beta^{j+1}_m)x^{-jm}, \quad \Omega^j_m = \sum_{k=1}^{j} x^{(j-2k+1)m}\beta^k_m, \tag{18}$$

$$S'^j_m = (\beta'^j_m - \beta'^{j+1}_m)x^{-jm}, \quad \Omega'^j_m = \sum_{k=1}^{j} x^{(j-2k+1)m}\beta'^k_m. \tag{19}$$

Define the bosonic Fock spaces $\mathscr{F}_{l,k} = c[\{\beta^j_{-1}, \beta^j_{-2}, \ldots\}_{1 \leqslant j \leqslant n}]|l,k\rangle$ and the vacuum vector,

$$|l,k\rangle = e^{i\sqrt{r/(r-1)}Q_l - i\sqrt{(r-1)/r}Q_k}|0,0\rangle. \tag{20}$$

In the following part of this paper, we will construct the bosonic realization of the vertex operators defined in Eq. (13)–Eq. (15) on the space $\Sigma_{l,k\in P} \oplus F_{l,k}$. For the generic r (e.g., an irrational number), we can introduce two types of independent operators $\hat{K}$ and $\hat{L}$ with $\hat{K} = \Sigma_\mu \hat{k}_\mu \overline{\epsilon}_\mu$, $\hat{L} = \Sigma_\mu \hat{l}_\mu \overline{\epsilon}_\mu$, which act on the space $\Sigma_{l,k\in P} \oplus F_{l,k}$ as

$$\hat{k}_\mu F_{l,k} = (w_\mu + k_\mu)F_{l,k}, \quad \text{if } k = \sum_\mu k_\mu \overline{\epsilon}_\mu,$$

$$\hat{l}_\mu F_{l,k} = (w_\mu + l_\mu)F_{l,k}, \quad \text{if } l = \sum_\mu l_\mu \overline{\epsilon}_\mu.$$

Let us introduce some basic operators acting on the space $\Sigma_{l,k\in P} \oplus F_{l,k}$:

$$\eta_j(v) = e^{-i\sqrt{(r-1)/r}(Q_{\omega_j} - i2v \ln x P_{\omega_j})} : e^{-\Sigma_{m\neq 0}(l/m)\Omega^j_m x^{-2vm}} :,$$

$$\xi_j(v) = e^{i\sqrt{(r-1)/r}(Q_{\alpha_j} - i2v \ln x P_{\alpha_j})} : e^{\Sigma_{m\neq 0}(1/m)S^j_m x^{-2vm}} :, \tag{21}$$

$$\eta_j'(v)=e^{i\sqrt{r/(r-1)}(Q_{\omega_j}-i2v\ \ln xP_{w_j})}{:}e^{\Sigma_{m\neq 0}(1/m)\Omega_m'^j x^{-2vm}}{:},$$

$$\xi_j'(v)=e^{-i\sqrt{r/(r-1)}(Q_{\alpha_j}-i2v\ \ln xP_{\alpha_j})}{:}e^{-\Sigma_{m\neq 0}(1/m)S_m'^j x^{-2vm}}{:}.$$

We can obtain the normal order relations which will be presented in the Appendix. From those results, we have the following commutation relations:

$$\xi_j(v_1)\xi_{j+1}(v_2)=-\frac{[v_1-v_2+\frac{1}{2}]}{[v_1-v_2-\frac{1}{2}]}\,\xi_{j+1}(v_2)\xi_j(v_1),$$

$$\xi_j(v_1)\,\eta_j(v_2)=-\frac{[v_1-v_2+\frac{1}{2}]}{[v_1-v_2-\frac{1}{2}]}\,\eta_j(v_2)\xi_j(v_1), \tag{22}$$

$$\xi_j(v_1)\xi_j(v_2)=\frac{[v_1-v_2-1]}{[v_1-v_2+1]}\,\xi_j(v_2)\xi_j(v_1),$$

where

$$[v]=x^{v^2/r-v}(x^{2v};x^{2r})(x^{2r}x^{-2v};x^{2r})(x^{2r};x^{2r})=\sigma\left(\frac{v}{r},-\frac{1}{rw}\right)\times\text{const.} \tag{23}$$

The bosonization of the type I vertex operator in the $A_{n-1}^{(1)}$ face model was given by Asai *et al.*,[8]

$$\phi_\mu(v)=\oint\prod_{j=1}^{\mu-1}\frac{dx^{2v_j}}{2\pi i x^{2v_j}}\,\eta_1(v)\xi_1(v_1)\cdots\xi_{\mu-1}(v_{\mu-1})\prod_{j=1}^{\mu-1}f(v_j-v_{j-1},\pi_{j\mu}), \tag{24}$$

$$\pi_\mu=\sqrt{r(r-1)}P_{\bar{\epsilon}_\mu},\quad \pi_{\mu\nu}=\pi_\mu-\pi_\nu,\quad v_0=v,\quad f(v,w)=\frac{[v+\frac{1}{2}-w]}{[v-\frac{1}{2}]}. \tag{25}$$

The integration contours are simple closed curves around the origin, satisfying

$$x|x^{2v_{j-1}}|<|x^{2v_j}|<x^{-1}|x^{2v_{j-1}}|.$$

We also have another set of relations,

$$\xi_j'(v_1)\xi_{j+1}'(v_2)=-\frac{[v_1-v_2-\frac{1}{2}]'}{[v_1-v_2+\frac{1}{2}]'}\,\xi_{j+1}'(v_2)\xi_j'(v_1),$$

$$\xi_j'(v_1)\,\eta_j'(v_2)=-\frac{[v_1-v_2-\frac{1}{2}]'}{[v_1-v_2+\frac{1}{2}]'}\,\eta_j'(v_2)\xi_j'(v_1), \tag{26}$$

$$\xi_j'(v_1)\xi_j(v_2)=\frac{[v_1-v_2+1]'}{[v_1-v_2-1]'}\,\xi_j'(v_2)\xi_j'(v_1),$$

where

$$[v]'=x^{v^2/(r-1)-v}(x^{2v};x^{2r-2})(x^{2r-2}x^{-2v};x^{2r-2})(x^{2r-2};x^{2r-2})=\sigma\left(\frac{v}{r-1},-\frac{1}{(r-1)w}\right)\times\text{const}'. \tag{27}$$

Construct type II vertex operators in the $A_{n-1}^{(1)}$ face model as

$$\phi'_\mu(v)=\oint\prod_{j=1}^{\mu-1}\frac{dx^{2v_j}}{2\pi i x^{2v_j}}\,\eta'_1(v)\xi'_1(v_1)\cdots\xi'_{\mu-1}(v_{\mu-1})\prod_{j=1}^{\mu-1}f'(v_j-v_{j-1},\pi_{j\mu}),\tag{28}$$

$$f'(v,w)=\frac{[v-\frac{1}{2}+w]'}{[v+\frac{1}{2}]'}.\tag{29}$$

Here the integration contours are simple closed curves around the origin satisfying

$$x|x^{2v_{j-1}}|<|x^{2v_j}|<x^{-1}|x^{2v_{j-1}}|.$$

Define the face Boltzmann weights for admissible configurations around a face,

$$\overline{W}\left(\begin{matrix}a+2\overline{\epsilon}_\mu & a+\overline{\epsilon}_\mu\\ a+\overline{\epsilon}_\mu & a\end{matrix}\middle|v\right)\equiv\overline{W}(a|v)^{\mu\mu}_{\mu\mu}=1,$$

$$\overline{W}\left(\begin{matrix}a+\overline{\epsilon}_\mu+\overline{\epsilon}_\nu & a+\overline{\epsilon}_\nu\\ a+\overline{\epsilon}_\nu & a\end{matrix}\middle|v\right)\equiv\overline{W}(a|v)^{\mu\nu}_{\nu\mu}=\frac{[v+a_{\mu\nu}][1]}{[v+1][a_{\mu\nu}]},\quad \mu\neq\nu,\tag{30}$$

$$\overline{W}\left(\begin{matrix}a+\overline{\epsilon}_\mu+\overline{\epsilon}_\nu & a+\overline{\epsilon}_\mu\\ a+\epsilon_\nu & a\end{matrix}\middle|v\right)\equiv\overline{W}(a|v)^{\mu\nu}_{\mu\nu}=\frac{[v][a_{\mu\nu}-1]}{[v+1][a_{\mu\nu}]},\quad \mu\neq\nu.$$

We also define another Boltzmann weight,

$$\overline{W}'\left(\begin{matrix}a+2\overline{\epsilon}_\mu & a+\overline{\epsilon}_\mu\\ a+\overline{\epsilon}_\mu & a\end{matrix}\middle|v\right)\equiv\overline{W}'(a|v)^{\mu\mu}_{\mu\mu}=1,$$

$$\overline{W}'\left(\begin{matrix}a+\overline{\epsilon}_\mu+\overline{\epsilon}_\nu & a+\overline{\epsilon}_\nu\\ a+\overline{\epsilon}_\nu & a\end{matrix}\middle|v\right)\equiv\overline{W}'(a|v)^{\mu\nu}_{\nu\mu}=\frac{[v+a_{\mu\nu}]'[1]'}{[v+1]'[a_{\mu\nu}]'},\quad \mu\neq\nu,\tag{31}$$

$$\overline{W}'\left(\begin{matrix}a+\overline{\epsilon}_\mu+\overline{\epsilon}_\nu & a+\overline{\epsilon}_\mu\\ a+\epsilon_\nu & a\end{matrix}\middle|v\right)\equiv\overline{W}'(a|v)^{\mu\nu}_{\mu\nu}=\frac{[v]'[a_{\mu\nu}-1]'}{[v+1]'[a_{\mu\nu}]'},\quad \mu\neq\nu.$$

Following the results obtained by Asai *et al.* in Ref. 8, we have

$$\phi_\mu(v_2)\phi_\nu(v_1)=r_1(v_2-v_1)\sum_{\mu',\nu'}\phi_{\mu'}(v_1)\phi_{\nu'}(v_2)\overline{W}(\hat{\pi}/v_1-v_2)^{\mu\nu}_{\nu'\mu'}.\tag{32}$$

Using the method introduced by Asai *et al.* in Ref. 8, and noting that

$$f'(v,\pi_{j\mu}-r)=\frac{[v-\frac{1}{2}+\pi_{j\mu}-r]'}{[v+\frac{1}{2}]'}=-\frac{[v-\frac{1}{2}+\pi_{j\mu}-1]'}{[v+\frac{1}{2}]'}$$

$$=-\frac{[-v+\frac{1}{2}-\pi_{j\mu}+1]'}{[-v-\frac{1}{2}]'}=-f(-v,\pi_{j\mu}-1)|_{r\to r-1},$$

$$f'(v,\pi_{\mu\nu}+r)=-f(-v,\pi_{\mu\nu}+1)|_{r\to r-1},$$

we can derive the following commutation relations:

$$\phi'_\mu(v_1)\phi'_\nu(v_2)=r'_1(v_1-v_2)\sum_{\mu',\nu'}\phi'_{\mu'}(v_2)\phi'_{\nu'}(v_1)\overline{W}'(\hat{\pi}|v_1-v_2)^{\mu\nu}_{\nu'\mu'},\tag{33}$$

$$\phi_\mu(v_1)\phi'_\nu(v_2)=\tau(v_1-v_2)\phi'_\nu(v_2)\phi_\mu(v_1),\tag{34}$$

where

$$r_1(v)=x^{2(r-1)(n-1)v/nr}\frac{g_1(-v)}{g_1(v)},\quad r_1'(v)=x^{2r(n-1)v/n(r-1)}\frac{g_1'(-v)}{g_1'(v)}.$$

One can see that ϕ_μ,ϕ_μ' are the intertwiners between the Fock space,

$$\phi_\mu(v):\ F_{l,k}\to F_{l,k+\bar{\epsilon}_\mu},$$

$$\phi_\mu'(v):\ F_{l,k}\to F_{l+\bar{\epsilon}_\mu,k}.$$

B. Face–vertex correspondence and modular transformation

It is convenient to introduce some notations. Define

$$\bar{R}^{(1)}(v)=\frac{\sigma_0\left(\frac{1}{r},-\frac{1}{rw}\right)}{\sigma_0\left(\frac{v+1}{r},-\frac{1}{rw}\right)}\sum_\alpha\frac{\sigma_\alpha\left(\frac{v}{r}+\frac{1}{nr},-\frac{1}{rw}\right)}{n\sigma_\alpha\left(\frac{1}{nr},-\frac{1}{rw}\right)}I_\alpha\otimes I_\alpha^{-1},$$

$$\bar{R}'^{(1)}(v)=\frac{\sigma_0\left(\frac{1}{r-1},-\frac{1}{(r-1)w}\right)}{\sigma_0\left(\frac{v+1}{r-1},-\frac{1}{(r-1)w}\right)}\sum_\alpha\frac{\sigma_\alpha\left(\frac{v}{r-1}+\frac{1}{n(r-1)},-\frac{1}{(r-1)w}\right)}{n\sigma_\alpha\left(\frac{1}{n(r-1)},-\frac{1}{(r-1)w}\right)}I_\alpha\otimes I_\alpha^{-1},$$

$$\varphi_{\mu,\hat{K}}^{(k)}(v)=\theta^{(k)}\left(\frac{v+n\langle\hat{K},\epsilon_\mu\rangle}{r}+\frac{(n-1)}{r},-\frac{1}{rw}\right),\tag{35}$$

$$\varphi_{\hat{L},\mu}'^{(k)}(v)=\theta^{(k)}\left(\frac{v+n\langle\hat{L},\epsilon_\mu\rangle}{r-1}+P_0,-\frac{1}{(r-1)w}\right),$$

$$\theta^{(k)}(z,\tau)=\theta\begin{bmatrix}-\frac{k}{n}\\ 0\end{bmatrix}(z,n\tau).$$

Although we choose the different function $\theta^{(k)}(v)$ from the ordinary one,[15] the face–vertex correspondence relation is still survived,

$$\sum_{k,l}\bar{R}^{(1)}(v_1-v_2)_{ij}^{kl}\varphi_{\nu,\hat{K}+\bar{\epsilon}_\mu}^{(k)}(v_1)\varphi_{\mu,\hat{K}}^{(l)}(v_2)=\sum_{\mu',\nu'}\bar{W}(\hat{K}|v_1-v_2)_{\nu'\mu'}^{\nu\mu}\varphi_{\nu',\hat{K}}^{(i)}(v_1)\varphi_{\mu',\hat{K}+\bar{\epsilon}_\nu}^{(j)}(v_2),$$

$$\sum_{k,l}\bar{R}'^{(1)}(v_1-v_2)_{ij}^{kl}\varphi_{\hat{L},\mu}'^{(k)}(v_1)\varphi_{\hat{L}-\bar{\epsilon}_\mu,\nu}'^{(l)}(v_2)=\sum_{\mu',\nu'}\bar{W}'(\hat{L}|v_1-v_2)_{\mu'\nu'}^{\mu\nu}\varphi_{\hat{L}-\bar{\epsilon}_{\nu'},\mu'}'^{(i)}(v_1)\varphi_{\hat{L},\nu'}'^{(j)}(v_2).$$

For the generic number $\{w_j\}$, we can introduce intertwiners $\tilde{\varphi}_{\mu,\hat{K}}(v)$ and $\tilde{\varphi}_{\hat{L},\mu}'(v)$, satisfying relations[15]

$$\sum_k\tilde{\varphi}_{\mu\hat{K}}^{(k)}(v)\varphi_{\nu,\hat{K}}^{(k)}(v)=\delta_{\mu\nu},\quad\sum_k\bar{\varphi}_{\hat{L},\nu}'^{(k)}(v)\varphi_{\hat{L},\nu}'^{(k)}(v)=\delta_{\mu\nu}.$$

So it is easy to find the following face–vertex correspondence relations,

$$\sum_{k,l}\bar{R}^{(1)}(v_1-v_2)_{kl}^{ij}\tilde{\varphi}_{\mu,\hat{K}}^{(k)}(v_1)\tilde{\varphi}_{\nu,\hat{K}+\bar{\epsilon}_\mu}^{(l)}(v_2)=\sum_{\mu',\nu'}\bar{W}(\hat{K}|v_1-v_2)_{\mu\nu}^{\mu'\nu'}\tilde{\varphi}_{\mu',\hat{K}+\bar{\epsilon}_{\nu'}}^{(i)}(v_1)\tilde{\varphi}_{\nu',\hat{K}}^{(j)}(v_2),\tag{36}$$

$$\sum_{k,l} \overline{R'}^{(1)}(v_1-v_2)_{kl}^{ij}\overline{\varphi}^{\prime(k)}_{\hat{L}-\overline{\epsilon}_{\nu,\mu}}(v_1)\overline{\varphi}^{\prime(l)}_{\hat{L},\nu}(v_2)=\sum_{\mu',\nu'} \overline{W}'(\hat{L}|v_1-v_2)_{\mu\nu}^{\nu'\mu'}\overline{\varphi}^{(i)}_{\hat{L},\nu'}(v_1)\overline{\varphi}^{(j)}_{\hat{L}-\overline{\epsilon}_{\nu',\mu'}}(v_2). \tag{37}$$

Introduce the modular transformation,

$$\theta\begin{bmatrix}\frac{1}{2}+a\\ \frac{1}{2}+b\end{bmatrix}\left(\frac{z}{\tau},-\frac{1}{\tau}\right)=\theta\begin{bmatrix}\frac{1}{2}+b\\ \frac{1}{2}-a\end{bmatrix}(z,\tau)\exp\ \pi i\left\{\frac{z^2}{\tau}+a-b+2ab\right\}\times\text{const}, \tag{38}$$

where the const only depends on τ. Therefore, we can derive the following relations for Z_n symmetry R matrices $\overline{R}^{(1)}(v)$ and $\overline{R'}^{(1)}(v)$,

$$(M\otimes M)\overline{R}^{(1)}(v)(M^{-1}\otimes M^{-1})=x^{[2v(1-n)/nr]}P\overline{R}(vw,rw)P, \tag{39}$$

$$(M\otimes M)\overline{R'}^{(1)}(v)(M^{-1}\otimes M^{-1})=x^{2v(1-n)/n(r-1)}P\overline{R}(vw,(r-1)w)P, \tag{40}$$

where P is the permutation operator acting on the tensor space $V\otimes V$ as $P(e_i\otimes e_j)=e_j\otimes e_i$, and the matrix $(M)_{lk}=\omega^{-lk}=\exp\{(-2i\pi/n)lk\}$ which have the following properties:

$$MgM^{-1}=h^{-1},\quad MhM^{-1}=g.$$

C. Bosonization for the Z_n Belavin model

Based on the bosonization for the $A_n^{(1)}$ face model and the face–vertex correspondence, we will construct the bosonization of two types of vertex operators for the Z_n Belavin model.

First, define

$$\Phi_j(v)=\sum_{i=1}^{n}\sum_{\mu=1}^{n} M_{ji}\phi_\mu(v)\tilde{\varphi}^{(i)}_{\mu,\hat{K}}(-v), \tag{41}$$

$$\Psi_j(v)=\sum_{i=1}^{n}\sum_{\mu=1}^{n} M_{ji}\phi'_\mu(v)\overline{\varphi}^{(i)}_{-\hat{L},\mu}(-v), \tag{42}$$

which satisfy the commutation relations on the space $\Sigma_{l,k\in P}\oplus F_{l,k}$,

$$\Phi_j(v_2)\Phi_i(v_1)=\sum_{ij} R_{lk}^{ij}(v_1-v_2)\Phi_l(v_1)\Phi_k(v_2), \tag{43}$$

$$\Psi_i(v_1)\Psi_j(v_2)=\sum_{lk} \Psi_k(v_2)\Psi_l(v_1)R_{lk}^{*ij}(v_1-v_2)\Delta_n^{-1}(v_1-v_2), \tag{44}$$

$$\Phi_i(v_1)\Psi_j(v_2)=\tau(v_1-v_2)\Psi_j(v_2)\Phi_i(v_1). \tag{45}$$

Based on the antisymmetric fusion for the Z_n symmetric R matrix,[10,15] we can define

$$\Psi_j^*(v)=\Psi_{(1,\ldots,j-1,j+1,\ldots,n)}(v)$$

$$\equiv\sum_{\sigma\in S_{n-1}} \text{sign}(\sigma)\Psi_{\sigma(1)}(v+n-1)\Psi_{\sigma(2)}(v+n-2)\cdots\Psi_{\sigma(n)}(v+1),$$

and derive the following relations:

$$\Phi_j(v_2)\Phi_i(v_1)=\sum_{lk} R_{lk}^{ij}(v_1-v_2)\Phi_l(v_1)\Phi_k(v_2), \tag{46}$$

$$\Psi_i^*(v_1)\Psi_j^*(v_2)=\sum_{lk}\ \Psi_k^*(v_2)\Psi_l^*(v_1)R_{ij}^{*lk}(v_1-v_2)\Delta_n^{-1}(v_1-v_2), \tag{47}$$

$$\Phi_i(v_1)\Psi_j^*(v_2)=\tau^{-1}(v_1-v_2)\Psi_j^*(v_2)\Phi_i(v_1). \tag{48}$$

Therefore, we obtain the bosonic realization of vertex operators $\Phi_i(v)$ and $\Psi_i^*(v)$ defined in Sec. II. Moreover, we obtain the bosonic realization for $A_{q,p}(\widehat{sl_n})$ algebra at level one through Miki construction, Eq. (16). Thus, this algebra is self-consistent.

Remark: For the case $n=2$ $(\Delta_2=-1)$, the bosonic operators $\Phi_i(v)$ and $\Psi_i^*(v)$ become the type I and type II vertex operators of $A_{q,p}(\widehat{sl_2})$ at level one proposed by Foda *et al.*[3]

Although we have constructed the bosonization of $A_{q,p}(\widehat{sl_n})$ at level one, we have still not found the scaling element which is related to the corner transfer matrix.[1–3] But, for the $A_{n-1}^{(1)}$ face type model, we could construct such a scaling element.[11] Unfortunately, the intertwiners $\tilde{\varphi}_{\mu,\hat{K}}^{(i)}(v)$ and $\tilde{\varphi}_{-\hat{L},\mu}^{(i)}(v)$ break the shift properties of vertex operators $\Phi_i(v)$ and $\Psi_i^*(v)$ in the vertex type model. So, the scaling element D of the elliptic algebra $A_{q,p}(\widehat{sl_n})$ is still an open problem.

V. DISCUSSION

For $n=2$, the algebra $A_{q,p}(\widehat{sl_n})$ reduces to the original $A_{q,p}(\widehat{sl_2})$ algebra which was first proposed by Foda *et al.*[3] The quantum affine algebra $U_q(\widehat{sl_2})$ is a degeneration algebra of $A_{q,p}(\widehat{sl_2})$ with $p=0$.[3] Unfortunately, due to the nontrivial scalar factor $\Delta_n(v)$ when $2<n$, the relation between the elliptic algebra $A_{q,p}(\widehat{sl_n})$ and $U_q(\widehat{sl_n})$ is still an open problem.

Some elliptic algebras were proposed by Felder[16] and its central extension was constructed by Enriquez *et al.*[17] However, our elliptic algebra is quite different from the algebra given by Felder or by Enriquez *et al.* This is a result of the fact that the $A_{q,p}(\widehat{sl_n})$ algebra is based on the relation ''$RLL=LLR^*$'' and $R.R^*$ are chosen with the different elliptic moduli. Moreover, the Drinfeld currents of the dynamically twisted algebra of our elliptic algebra[18] is quite different from that of Felder's algebra and from that of the Enriquez *et al.* algebra.

As discussed in Ref. 8, vertex operators $\phi_\mu(v)$ in the $A_{n-1}^{(1)}$ face model was the q analog of the chiral primary fields of q-deformed W algebras.[16] However, the vertex operators $\Phi_i(v)$ and $\Psi_j^*(v)$ in the Z_n Belavin model are reconstructed by the vertex operators of the $A_{n-1}^{(1)}$ face model through the face–vertex correspondence relations. Moreover, the elliptic algebra $A_{q,p}(\widehat{sl_n})$ at level one can be constructed by Miki construction. So, there exist some relations between the algebra $A_{q,p}(\widehat{sl_n})$ and q-deformed W algebras.[19]

Besides the degenerate algebra $A_{q,0}(\widehat{sl_n})$, there exists another degenerate algebra $A_{\hbar,\eta}(\widehat{sl_n})$,[20] which can be obtained by the scaling limit $(v=u/\hbar,\ q=x,\ p=x^\eta,\ x\to 1)$ of the elliptic algebra $A_{q,p}(\widehat{sl_n})$. The resulting algebra $A_{\hbar,\eta}(\widehat{sl_n})$ is some deformation of the Yangian double $DY(\widehat{sl_n})$.[20,21] In the scaling limit case the R matrix entering the $\Psi^*\Psi^*$ commutation relation would be interpreted as the S matrix for a soliton of the affine Toda theory. (In the case of $n=2$, it is related to sine-Gordon theory.[21]) Moreover, the $\hbar$-deformed W algebra[22] would be related to the degenerate algebra $A_{\hbar,\eta}(\widehat{sl_n})$ in some way.

In our formulation, the algebra $A_{q,p}(\widehat{sl_n})$ is formulated in the framework of the ''RLL'' approach in terms of the L operator. It would be of great importance to find an analogic Drinfeld current for the algebra $A_{q,p}(\widehat{sl_n})$. For a special case $A_{q,p}(\widehat{sl_2})$ with the scaling limit, the Drinfeld currents for the algebra $A_{\hbar,\eta}(\widehat{sl_2})$ were found to be the Gauss coordinates of the L operator.[20] We expect that the same relation would exist for the elliptic algebra $A_{q,p}(\widehat{sl_n})$.

APPENDIX: SOME DETAILED CALCULATIONS

1. The normal order relation for the basic operator

We list the relations as three series: Type I,

$$\eta_1(v_1)\,\eta_1(v_2)=x^{2(r-1)(n-1)v_1/nr}g_1(v_2-v_1):\eta_1(v_1)\,\eta_1(v_2):,$$

$$\begin{aligned}
&\eta_j(v_1)\xi_j(v_2)=x^{[2(1-r)/r]v_1}s(v_2-v_1):\eta_j(v_1)\xi_j(v_2):,\\
&\xi_j(v_2)\,\eta_j(v_1)=x^{[2(1-r)/r]v_2}s(v_1-v_2):\xi_j(v_2)\,\eta_j(v_1):,\\
&\xi_j(v_1)\xi_{j+1}(v_2)=x^{[2(1-r)/r]v_1}s(v_2-v_1):\xi_j(v_1)\xi_{j+1}(v_2):,\\
&\xi_{j+1}(v_2)\xi_j(v_1)=x^{[2(1-r)/r]v_2}s(v_1-v_2):\xi_j(v_1)\xi_{j+1}(v_2):,\\
&\xi_j(v_1)\xi_j(v_2)=x^{[4(r-1)/r]v_1}t(v_2-v_1):\xi_j(v_1)\xi_j(v_2):,\\
&\xi_j(v_1)\xi_l(v_2)=:\xi_j(v_1)\xi_l(v_2): \quad \text{if } |l-j|>1,\\
&\eta_j(v_1)\xi_l(v_2)=:\eta_j(v_1)\xi_l(v_2): \quad \text{if } l\neq j;
\end{aligned} \tag{A1}$$

Type II,

$$\begin{aligned}
&\eta_1'(v_1)\,\eta_1'(v_2)=x^{2r(n-1)v_1/(r-1)n}g_1'(v_2-v_1):\eta_1'(v_1)\,\eta_1'(v_2):,\\
&\xi_j'(v_1)\,\eta_j'(v_2)=x^{[-2r/(r-1)]}s'(v_2-v_1):\xi_j'(v_1)\,\eta_j'(v_2):,\\
&\eta_j'(v_2)\xi_j'(v_1)=x^{[-2r/(r-1)]v_2}s'(v_1-v_2):\eta_j'(v_2)\xi_j'(v_1):,\\
&\xi_j'(v_1)\xi_{j+1}'(v_2)=x^{[-2r/(r-1)]v_1}s'(v_2-v_1):\xi_j'(v_1)\xi_{j+1}'(v_2):,\\
&\xi_{j+1}'(v_2)\xi_j'(v_1)=x^{[-2r/(r-1)]v_2}s'(v_1-v_2):\xi_j'(v_1)\xi_{j+1}'(v_2):,\\
&\xi_j'(v_1)\xi_j'(v_2)=x^{[4r/(r-1)]v_1}t'(v_2-v_1):\xi_j'(v_1)\xi_j'(v_2):,\\
&\xi_j'(v_1)\xi_l'(v_2)=:\xi_j'(v_1)\xi_l'(v_2): \quad \text{if } |l-j|>1,\\
&\eta_j'(v_1)\xi_l'(v_2)=:\eta_j'(v_1)\xi_l'(v_2): \quad \text{if } l\neq j;
\end{aligned} \tag{A2}$$

Type I with type II:

$$\begin{aligned}
&\eta_j(v_1)\xi_j'(v_2)=(x^{2v_1}-x^{2v_2}):\xi_j'(v_2)\,\eta_j(v_1):,\\
&\xi_j'(v_2)\,\eta_j(v_1)=(x^{2v_2}-x^{2v_1}):\eta_j(v_1)\xi_j'(v_2):,\\
&\xi_j(v_1)\,\eta_j'(v_2)=(x^{2v_1}-x^{2v_2}):\xi_j(v_1)\,\eta_j'(v_2):,\\
&\eta_j'(v_2)\xi_j(v_1)=(x^{2v_2}-x^{2v_1}):\eta_j'(v_2)\xi_j(v_1):,\\
&\eta_l(v_1)\xi_j'(v_2)=:\eta_l(v_1)\xi_j'(v_2)=:\xi_j'(v_2)\,\eta_l(v_1), \quad \text{if } l\neq j,\\
&\eta_l'(v_1)\xi_j(v_2)=:\eta_l'(v_1)\xi_j(v_2)=:\xi_j(v_2)\,\eta_l'(v_1), \quad \text{if } l\neq j,\\
&\xi_j(v_1)\xi_{j+1}'(v_2)=(x^{2v_1}-x^{2v_2}):\xi_j(v_1)\xi_{j+1}'(v_2):,\\
&\xi_{j+1}'(v_2)\xi_j(v_1)=(x^{2v_2}-x^{2v_1}):\xi_{j+1}'(v_2)\xi_j(v_1):,
\end{aligned} \tag{A3}$$

$$\xi_j(v_1)\xi_j'(v_2)=x^{-4v_1}\,\frac{1}{(1-xx^{2(v_2-v_1)})(1-x^{-1}x^{2(v_2-v_1)})}:\xi_j(v_1)\xi_j'(v_2):,$$

$$\xi_j'(v_2)\xi_j(v_1)=x^{-4v_2}\,\frac{1}{(1-xx^{2(v_1-v_2)})(1-x^{-1}x^{2(v_1-v_2)})}:\xi_j'(v_2)\xi_j(v_1):,$$

$$\xi_j(v_1)\xi'_l(v_2)=:\xi_j(v_1)\xi'_l(v_2): \quad \text{if } |j-l|>1,$$

$$\xi'_l(v_2)\xi_j(v_1)=:\xi'_l(v_2)\xi_j(v_1): \quad \text{if } |j-l|>1,$$

$$\eta_1(v_1)\eta'_j(v_2)=x^{2v_1[(j-n)/n]}\frac{(x^{2n-j}x^{2(v_2-v_1)};x^{2n})}{(x^jx^{2(v_2-v_1)};x^{2n})}:\eta_1(v_1)\eta'_j(v_2):,$$

$$\eta'_j(v_2)\eta_1(v_1)=x^{2v_2[(j-n)/n]}\frac{(x^{2n-j}x^{2(v_1-v_2)};x^{2n})}{(x^jx^{2(v_1-v_2)};x^{2n})}:\eta'_j(v_2)\eta_1(v_1):,$$

where

$$g_1(v)=\frac{\{x^2x^{2v}\}\{x^{2n+2r-2}x^{2v}\}}{\{x^{2n}x^{2v}\}\{x^{2r}x^{2v}\}}, \quad s(v)=\frac{(x^{2r-1}x^{2v};x^{2r})}{(xx^{2v};x^{2r})},$$

$$t(v)=(1-x^{2v})\frac{(x^2x^{2v};x^{2r})}{(x^{2r-2}x^{2v};x^{2r})},$$

$$g'_1(v)=\frac{\{x^{2v}\}'\{x^{2n+2r-2}x^{2v}\}'}{\{x^{2r}x^{2v}\}'\{x^{2n-2}x^{2v}\}'}, \quad \{z\}'=(z;x^{2r-2},x^{2n}),$$

$$s'(v)=\frac{(x^{2r-1}x^{2v};x^{2r-2})}{(x^{-1}x^{2v};x^{2r-2})}, \quad t'(v)=(1-x^{2v})\frac{(x^{-2}x^{2v};x^{2r-2})}{(x^{2r}x^{2v};x^{2r-2})}. \tag{A4}$$

2. Proof of the commutation relations, Eq. (43)–Eq. (45)

Notice that with the commutation relation for zero mode operators, we have

$$\phi_\nu(v_1)\tilde{\varphi}^{(k)}_{\mu,\hat{K}}(v_2)=\tilde{\varphi}^{(k)}_{\mu,\hat{K}-\bar{\epsilon}_\nu}(v_2)\phi_\nu(v_1), \quad \phi'_\nu(v_1)\tilde{\varphi}^{(k)}_{\mu,\hat{K}}(v_2)=\tilde{\varphi}^{(k)}_{\mu,\hat{K}}(v_2)\phi'_\nu(v_1),$$

$$\phi_\nu(v_1)\overline{\varphi}'^{(k)}_{\hat{L},\mu}(v_2)=\overline{\varphi}'^{(k)}_{\hat{L},\mu}(v_2)\phi_\nu(v_1), \quad \phi'_\nu(v_1)\overline{\varphi}'^{(k)}_{\hat{L},\mu}(v_2)=\overline{\varphi}'^{(k)}_{\hat{L}-\bar{\epsilon}_{\nu,\mu}}(v_2)\phi'_\nu(v_1).$$

Define

$$Z_j(v)=\sum_{\mu=1}^{n}\phi_\mu(v)\tilde{\varphi}^{(i)}_{\mu,\hat{K}}(-v), \quad Z'_j(v)=\sum_{\mu=1}^{n}\phi'_\mu(v)\overline{\varphi}'^{(i)}_{-\hat{L},\mu}(-v).$$

Notice that

$$\overline{W}(\hat{\pi}|v)^{\nu'\mu'}_{\nu\mu}|_{F_{l,k}}=\overline{W}(\hat{K}|v)^{\nu'\mu'}_{\nu\mu}|_{F_{l,k}},$$

using the exchange relation of $\phi_\mu(v)$ in Eq. (32) and the face–vertex correspondence by $\tilde{\varphi}^{(k)}_{\mu,\hat{K}}(v)$ in Eq. (36), we have

$$Z_j(v_2)Z_i(v_1)=\sum_{l,k}r_1(v_2-v_1)(P\overline{R}^{(1)}(v_1-v_2)P)^{ij}_{lk}Z_l(v_1)Z_k(v_2).$$

Notice the properties of the Z_n symmetry R matrix under the modular transformation, Eq. (39), we obtain

$$Z_j(v_2)Z_i(v_1)=\sum_{l,k}(M^{-1}\otimes M^{-1}R(v_1-v_2)M\otimes M)^{ij}_{lk}Z_l(v_1)Z_k(v_2).$$

Due to $\Phi_j(v)=\Sigma_{i=1}^{n}M_{ji}Z_i(v)$, one can obtain Eq. (43). Using the same method and noticing that

$$\overline{W}'(\hat{L}|v)^{\mu\nu}_{\nu'\mu'}=\overline{W}'(-\hat{L}|v)^{\nu\mu}_{\mu'\nu'}, \quad \frac{r_1'(v)}{r_1(-v)|_{r\to r-1}}=\Delta_n(v),$$

$$\overline{W}'(\hat{\pi}|v)^{\nu'\mu'}_{\nu\mu}|_{F_{l,k}}=\overline{W}'(\hat{L}|v)^{\nu'\mu'}_{\nu\mu}|_{F_{l,k}},$$

we can obtain Eq. (44) and Eq. (45).

[1] M. Jimbo and T. Miwa, *Algebraic Analysis of Solvable Lattice Models*, CBMS Regional Conference Series in Mathematics (American Mathematical Society, Providence, RI, 1994), Col. 85.

[2] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, London, 1982).

[3] O. Foda, K. Iohara, M. Jimbo, R. Kedem, T. Miwa, and H. Yan, ''An elliptic quantum algebra for $s\hat{l}_2$,'' Lett. Math. Phys. **32**, 259 (1994); ''Notes on highest weight modules of the elliptic algebra $A_{q,p}(\widehat{sl_2})$,'' Prog. Theor. Phys. Suppl. **118**, 1 (1995).

[4] B. Davies, O. Foda, M. Jimbo, T. Miwa, and A. Nakayashiki, ''Diagonalization of the XXZ Hamiltonian by vertex operators,'' Commun. Math. Phys. **151**, 89 (1993).

[5] O. Foda, M. Jimbo, T. Miwa, K. Miki, and A. Nakayashiki, ''Vertex operators in solvable lattice models,'' J. Math. Phys. **35**, 13 (1994).

[6] S. L. Lukyanov and Ya. Pugai, ''Multi-point local height probabilities in the integrable RSOS model,'' Nucl. Phys. B **473**, 631 (1996).

[7] G. E. Andrews, R. J. Baxter, and P. J. Forrester, ''Eight-vertex SOS model and generalized Rogers–Ramanujan-type identities,'' J. Stat. Phys. **35**, 193 (1984).

[8] Y. Asai, M. Jimbo, T. Miwa, and Ya. Pugai, ''Bosonization of vertex operators for the $A^{(1)}_{n-1}$ face model,'' J. Phys. A **29**, 6595 (1996).

[9] M. Jimbo, T. Miwa, and M. Okado, ''Local state probabilities of solvable lattice models: An $A^{(1)}_{n-1}$ family,'' Nucl. Phys. B **300**, 74 (1988).

[10] A. A. Belavin, ''Dynamicall symmetry of integrable quantum systems,'' Nucl. Phys. B **180**, 189 (1981); M. P. Richey and C. A. Tracy, ''Z_n Baxter model: Symmetries and the Belavin parameterization,'' J. Stat. Phys. **42**, 311 (1986).

[11] H. Fan, B. Y. Hou, K. J. Shi, and W. L. Yang, ''Bosonization of vertex operators for Zn symmetric Belavin model,'' J. Phys. A **30**, 5687 (1997); H. Fan, B. Y. Hou, K. J. Shi, W. L. Yang, and Z. X. Yang, ''Zamolodchikov–Faddeev algebra related to Z_n symmetric elliptic R-matrix,'' *ibid.* **28**, 3157 (1995).

[12] K. Miki, ''Creation/annihilation operators and form factors of the XXZ model,'' Phys. Lett. A **186**, 217 (1994).

[13] Y.-H. Quano, ''Spontaneous polarization of the Z_n-Baxter model,'' Mod. Phys. Lett. A **8**, 3363 (1993).

[14] I. B. Frenkel and N. Reshetikhin, ''Quantum affine algebras and holonomic difference equations,'' Commun. Math. Phys. **146**, 1 (1992).

[15] B. Y. Hou and Y. K. Zhou, ''Fusion procedure and Sklyanin algebra,'' J. Phys. A **23**, 1147 (1990); B. Y. Hou, K. J. Shi, and Z. X. Yang, ''Cyclic representation and function difference representation of Z_n Sklyanin algebra,'' *ibid.* **26**, 4951 (1993).

[16] G. Felder, ''Elliptic quantum groups,'' *Proceedings of the ICMP*, Paris, 1994, hep-th/9412207, 1995, p. 211.

[17] B. Enriquez and G. Felder, ''Elliptic quantum groups $E_{\tau,\eta}(\widehat{sl_2})$ and quasi-Hopf algebra,'' q-alg/9703018.

[18] B. Y. Hou and W. L. Yang, ''Dynamically twisted algebra $A_{q,p;\hat{\pi}}(\widehat{gl_2})$ algebra as current algebra generalizing screening currents of q-deformed Virasoro algebra,'' q-alg/9709024.

[19] H. Awata, H. Kubo, S. Odake, and J. Shiraishi, ''Quantum W_N algebras and Macdonald polynomials,'' Commun. Math. Phys. **179**, 401 (1996); B. Feigin and E. Frenkel, ''Quantum W-algebras and elliptic algebras,'' *ibid.* **178**, 653 (1996).

[20] S. Khoroshkin, D. Lebedev, and S. Pakuliak, ''Elliptic algebra $A_{\hbar,\eta}(\widehat{sl_2})$ in the scaling limit,'' q-alg/9702002; B. Y. Hou, L. Zhao, and X. M. Ding, ''The algebra $A_{\hbar,\eta}(\hat{g})$ and Hopf family of algebra,'' hep-th/9703046.

[21] K. Iohara, ''Bosonic representations of Yangian double $DY_\hbar(g)$ with $g=gl_N, sl_N$,'' J. Phys. A **29**, 4593 (1996); H. Konno, ''Degeneration of the elliptic algebra $A_{q,p}(\widehat{sl_n})$ and form factors in the sine-Gordon theory,'' hep-th/9701034; M. Jimbo, H. Konno, and T. Miwa, ''Massless XXZ model and degeneration of the elliptic algebra $A_{q,p}(\widehat{sl_2})$,'' hep-th/9610079.

[22] B. Y. Hou and W. L. Yang, ''A $\hbar$-deformed Virasoro algebra as a hidden symmetry of the restricted sine-Gordon model,'' hep-th/9612235; ''An $\hbar$-deformation of the W_N algebra and its vertex operators,'' hep-th/9701101.

Coupling constant thresholds of perturbed periodic Hamiltonians

Silvestro Fassari and Martin Klaus[a)]
Department of Mathematics, Virginia Tech, Blacksburg, Virginia 24061

(Received 6 March 1998; accepted for publication 24 March 1998)

We consider Schrödinger operators of the form $H_\lambda = -\Delta + V + \lambda W$ on $L^2(\mathbf{R}^\nu)$ ($\nu = 1$, 2, or 3) with V periodic, W short range, and λ a real non-negative parameter. Then the continuous spectrum of H_λ has the typical band structure consisting of intervals, separated by gaps. In the gaps there may be discrete eigenvalues of H_λ that are functions of the parameter λ. Let (a,b) be a gap and $E(\lambda) \in (a,b)$ an eigenvalue of H_λ. We study the asymptotic behavior of $E(\lambda)$ as λ approaches a critical value λ_0, called a coupling constant threshold, at which the eigenvalue either emerges from or is absorbed into the continuous spectrum. A typical question is the following: Assuming $E(\lambda) \downarrow a$ as $\lambda \downarrow \lambda_0$, is $E(\lambda) - a \sim c(\lambda - \lambda_0)^\alpha$ for some $\alpha > 0$ and $c \neq 0$, or is there an expansion in some other quantity? As one expects from previous work in the case $V = 0$, the answer strongly depends on ν. © *1998 American Institute of Physics.* [S0022-2488(98)03007-2]

I. INTRODUCTION

We consider the family of Schrödinger operators,

$$H_\lambda = H_0 + \lambda W, \tag{1.1}$$

on $L^2(\mathbf{R}^\nu)$, $\nu = 1,2,3$, where

$$H_0 = -\Delta + V, \tag{1.2}$$

V is periodic and W is a short-range perturbation. The parameter λ, called the coupling constant, is assumed to be real and non-negative. Specific conditions will be placed on V and W below; they will guarantee that the spectra of H_0 and H_λ exhibit the qualitative behavior described in this Introduction. One important restriction on V that we should mention at the outset is that in two and three dimensions we assume V to be of a form which allows us to solve the eigenvalue problem for H_0 by separation of variables. As is typical for Schrödinger operators with a periodic potential, the spectrum $\sigma(H_0)$ of H_0 consists of bands separated by gaps, where in one dimension the generic situation is that there are infinitely many gaps, while in two and three dimensions the number of gaps is finite. By a gap in $\sigma(H_0)$ we mean an interval (a,b) such that $a,b \in \sigma(H_0)$ and $(a,b) \cap \sigma(H_0) = \varnothing$. For $\lambda > 0$, the spectrum of H_λ is made up of a continuous part which coincides with the spectrum of H_0 and a discrete part consisting of at most a finite number of eigenvalues (counting multiplicities) in every gap. Since there are no eigenvalues when $\lambda = 0$, by general results from perturbation theory, eigenvalues of H_λ can appear in a gap only by emerging from one of its endpoints as λ is varied. Likewise, eigenvalues can disappear from a gap only by converging to an endpoint. A value λ_0 is called a coupling constant threshold (henceforth abbreviated as c.c.th.) for the family H_λ at the endpoint a, resp. b, of a gap (a,b), if there is an eigenvalue branch $E(\lambda)$ of H_λ such that $E(\lambda) \downarrow a$, resp. $E(\lambda) \uparrow b$, as $\lambda \downarrow \lambda_0$ or $\lambda \uparrow \lambda_0$. The case where $E(\lambda) \downarrow a$ as $\lambda \uparrow \lambda_0$ describes an eigenvalue that approaches the endpoint a from the right and is absorbed into the continuous spectrum at $\lambda = \lambda_0$. The case where $E(\lambda) \downarrow a$ as $\lambda \downarrow \lambda_0$ corresponds to the situation where an eigenvalue appears at a as $\lambda \uparrow \lambda_0 + \epsilon$ ($\epsilon > 0$). The third possibility is that

a)Electronic mail: klaus@math.vt.edu

 © 1998 American Institute of Physics

there is an eigenvalue converging to a both as $\lambda \uparrow \lambda_0$ and as $\lambda \downarrow \lambda_0$. In other words, an eigenvalue comes in from the right and ''turns around'' at a as λ is increased. Of course, analogous possibilities exist at the endpoint b.

The main object of this paper is to study the analytic behavior of $E(\lambda)$ when λ is near a c.c.th. λ_0. Our analysis rests on the Birman–Schwinger principle which allows us to recast the eigenvalue problem for H_λ as an eigenvalue problem for a compact integral operator. Depending on the dimension we will employ slightly different versions of the Birman–Schwinger principle. For any fixed $\lambda \geqslant 0$ we introduce the operators

$$K_{\lambda,E} = W^{1/2}(H_\lambda - E)^{-1}|W|^{1/2}, \quad W^{1/2} = |W|^{1/2}\mathrm{sgn}\ W, \tag{1.3}$$

where sgn denotes the sign function, and call them Birman–Schwinger kernels, since in the applications these operators will be represented by explicit integral kernels. In one dimension the relevant kernel will be $K_{\lambda_0,E}$, where λ_0 is a c.c.th.; in two and three dimensions it will be $K_{0,E}$. The Birman–Schwinger principle says that

$$H_\lambda \psi = E\psi, \tag{1.4}$$

with $E \in (a,b)$ and $\psi \in L^2(\mathbf{R}^\nu)$ if and only if

$$(a)\ \ (\lambda - \lambda_0)K_{\lambda_0,E} f = -f, \quad \nu = 1,$$

$$(b)\ \ \lambda K_{0,E} f = -f, \quad \nu = 2,3,$$

where f and ψ are related by $f = W^{1/2}\psi$ and $\psi = -[\lambda - \lambda_0](H_{\lambda_0} - E)^{-1}|W|^{1/2} f$ in case (a), resp. $\psi = -\lambda(H_0 - E)^{-1}|W|^{1/2} f$, in case (b). Moreover, the multiplicity of E as an eigenvalue of H_λ is the same as the (geometric) multiplicity of -1 as an eigenvalue of $(\lambda - \lambda_0)K_{\lambda_0,E}$, resp. $\lambda K_{0,E}$. Note that the Birman–Schwinger kernel is not self-adjoint unless W has constant sign, so that in general one has to distinguish between the algebraic and geometric multiplicity of an eigenvalue.

The main reason for using different kernels depending on the dimension is that in one dimension it is useful to think of $(\lambda - \lambda_0)W$ as a perturbation of H_{λ_0} because the solutions to $H_{\lambda_0}\psi = E\psi$ can be constructed using familiar techniques from the theory of integral equations and $K_{\lambda_0,E}$ can be expressed in terms of these solutions. One could also use $K_{0,E}$ in one dimension, but we were able to obtain stronger results by using $K_{\lambda_0,E}$. In two and three dimensions it is more convenient to view λW as a perturbation of H_0 because our assumptions on V will allow us to get detailed information about the resolvent $(H_0 - E)^{-1}$ and, therefore, about $K_{0,E}$.

For later use we give some more details of our approach when $\nu = 1$ or 2. We choose the notation appropriate for the one-dimensional case but keep it sufficiently general so that the two-dimensional case is included by setting $\lambda_0 = 0$. Consider the specific situation where $E(\lambda) \uparrow b$ as $\lambda \downarrow \lambda_0$. It turns out that as in the case $V = 0$, the kernel $K_{\lambda_0,E}$ can be decomposed into two parts: a singular part and a regular part. The singular part is of finite rank (rank one if $\nu = 1$) and diverges as $E \uparrow b$, while the regular part has a limit (in Hilbert–Schmidt norm) as $E \uparrow b$. The general form of this decomposition is

$$K_{\lambda_0,E} = Q_{\lambda_0,E} + R_{\lambda_0,E}, \tag{1.5}$$

$$Q_{\lambda_0,E} = \sum_{j=1}^{N} d_{\lambda_0;j}(E) L_{\lambda_0,E;j}, \quad L_{\lambda_0,E;j}\cdot = \omega_{\lambda_0,E;j}(\tilde{\omega}_{\lambda_0 E;j}, \cdot). \tag{1.6}$$

Here and in the sequel $(,)$ denotes the inner product in $L^2(\mathbf{R}^\nu)$ such that $(f,g) = \int_{\mathbf{R}^\nu} \overline{f(x)} g(x) d^\nu x$. The sets of vectors $\{\omega_{\lambda_0,E;j}\}_{j=1}^N$ and $\{\tilde{\omega}_{\lambda_0,E;j}\}_{j=1}^N$ are separately linearly independent and $\lim_{E\uparrow b}\ \omega_{\lambda_0,E;j}$ and $\lim_{E\uparrow b}\ \omega_{\lambda_0,E;j}$ exist in the L^2-norm. The coefficients $d_{\lambda_0;j}(E)$ are real and diverge as $E \uparrow b$ in a specific manner depending on ν. So, in (1.5), $Q_{\lambda_0,E}$ is the singular part and $R_{\lambda_0,E}$ is the regular part. In one dimension $N = 1$, but in two dimensions any $N > 1$ is possible. Note that for $|\lambda - \lambda_0|$ small enough,

$$(\lambda-\lambda_0)K_{\lambda_0,E}f=-f\Leftrightarrow[\lambda-\lambda_0](1+[\lambda-\lambda_0]R_{\lambda_0,E})^{-1}Q_{\lambda_0,E}f=-f.$$

By using (1.6), the right-hand is seen to be equivalent to

$$(\mathbf{I}_N+\mathbf{A})\mathbf{v}=0, \tag{1.7}$$

where $\mathbf{I}_N$ denotes the identity matrix in $\mathbf{C}^N$, $\mathbf{A}$ is the $N\times N$ matrix with entries

$$A_{i,j}=(\lambda-\lambda_0)d_{\lambda_0;j}(E)(\tilde{\omega}_{\lambda_0,E;i},(1+[\lambda-\lambda_0]R_{\lambda_0,E})^{-1}\omega_{\lambda_0,E;j}),$$

and $\mathbf{v}$ is the vector with components $v_j=(\tilde{\omega}_{\lambda_0,E;j},f)$, $j=1,...,N$. If $\mathbf{v}$ is known, then f is given by

$$f=-(\lambda-\lambda_0)\sum_{j=1}^{N} d_{\lambda_0;j}(E)(1+[\lambda-\lambda_0]R_{\lambda_0,E})^{-1}\omega_{\lambda_0,E;j}v_j.$$

From (1.7) we conclude that a nontrivial solution $\mathbf{v}$ exists if and only if $\det[\mathbf{I}_N+\mathbf{A}]=0$, that is,

$$\det[\,\delta_{ij}+(\lambda-\lambda_0)d_{\lambda_0;j}(E)(\tilde{\omega}_{\lambda_0,E;i},(1+[\lambda-\lambda_0]R_{\lambda_0,E})^{-1}\omega_{\lambda_0,E;j})]_{i,j=1}^{N}=0, \tag{1.8}$$

where δ_{ij} is the Kronecker delta. This is the equation which we will use to investigate $E(\lambda)$.

We remark that a similar approach has been used by Simon[1] and Holden[2] when $V=0$ and $\nu=2$. An alternative method to study $E(\lambda)$ is based on eigenvalue perturbation theory. The reader is referred to Refs. 1 and 3–5, and, in the special case when V is a Kronig–Penney potential, to Ref. 6. The eigenvalue perturbation method has the drawback that in some cases, when certain integrals involving W are zero, it becomes rather cumbersome and causes complications because $K_{\lambda_0,E}$ is in general not self-adjoint. At this point we should mention that for most results in this paper we do not require W to have constant sign. It is mainly for these reasons that we have decided not to use eigenvalue perturbation theory in one and two dimensions in this paper.

In three dimensions the situation is different in that the operator $K_{0,E}$ converges to a limit, $K_{0,b}$, as $E\uparrow b$, where $K_{0,b}$ is compact. Then an eigenvalue $E(\lambda)$ is associated with an eigenvalue branch $\tau(E)$ of $K_{0,E}$ such that $\lim_{E\uparrow b}\tau(E)$ exists and is finite, and the eigenvalue $E(\lambda)$ is found by solving $\lambda\tau(E)=-1$ for $E(\lambda)$. Moreover, the c.c.th.'s are given by $-1/\tau_j(b)$, where $\tau_j(b)$ $(j=1,2,...)$ denotes the negative eigenvalues of $K_{0,b}$. Thus, in three dimensions, $\lambda_0=0$ is never a c.c.th. at b (under our assumptions on V and W). Similarly, $\lambda_0=0$ is never a c.c.th. at a and hence, if λ is sufficiently small, H_λ has no eigenvalues in the gap (a,b).

We would like to mention that part of our motivation for studying the c.c.th. behavior of perturbed periodic Hamiltonians came from seminars given by two of our colleagues, Chris Beattie and George Hagedorn, on spectral properties of operators of the type considered here. Both of these individuals have also done numerical calculations of eigenvalue branches in dimensions one and two.[7,8] Further motivation came from earlier work on the subject in the case $V=0$,[1–5] the Kronig–Penney case,[6] and from related work on c.c.th.'s[9,10] and on eigenvalues in spectral gaps.[11–15]

This paper is organized as follows. Section II is devoted to the one-dimensional case. We describe the various threshold behaviors that can occur and find the leading order asymptotics of the eigenvalues. In some cases we also find higher order corrections to the leading term. We will see that generally $E(\lambda)-b\sim -c(\lambda-\lambda_0)^{2m}$ as $\lambda\downarrow\lambda_0$ or $\lambda\uparrow\lambda_0$, with $c>0$ and m a positive integer. Similar results hold at a. However, there are various situations where we can say more. For example, if $b=\inf\sigma(H_0)$ and $\lambda_0=0$, then only $m=1$ and 2 can occur, while if $\lambda_0>0$, then only $m=1$ is possible (Theorems 2.11 and 2.12). If b (or a) is the endpoint of a finite gap, then a similar result holds provided the support of W is suitably restricted (Theorem 2.16). Without such a restriction any m seems to be possible. For an eigenvalue that ''turns around'' we always have $m\geqslant 2$ with m even [Theorem 2.12(ii) and (v)]. Because we have made an effort to keep the

assumptions on W as weak as possible, there are technical details that require a careful discussion and which have caused this section to become larger than originally intended. In order to avoid extensive digressions we have moved the proofs of five technical lemmas to an Appendix.

In Section III we consider some special cases and extensions of the results of Section II, and compare them with similar results in the literature. In particular, we present an example illustrating the case $\lambda_0=0$, $E(\lambda)-a\sim c\lambda^8 (c>0)$, and we briefly discuss the case when $W(x)$ is of the form $\pm\delta(x-x_0)$.

In Section IV we consider the two-dimensional case. In view of the length of Section II we content ourselves with a discussion of the case $\lambda_0=0$. The typical leading behavior at the right endpoint of a gap in two dimensions is $E(\lambda)-b\sim -c_1 e^{-c_2/\lambda^s}$ as $\lambda\downarrow 0$, with c_1, $c_2>0$, and s a positive integer. If $b=\inf\sigma(H_0)$, then generally $s=1$, but $s=2$ is also possible if a certain integral involving W vanishes. At other endpoints any s seems to be possible, but $s=1$ is the generic situation. There is also the possibility that more than one distinct eigenvalue approaches an endpoint as $\lambda\downarrow 0$, or that an eigenvalue is degenerate.

Finally, in Section V we briefly consider the three-dimensional case, restricting ourselves to c.c.th.'s at the bottom of the essential spectrum which we assume to be at zero. The main result, Theorem 5.3, shows that there are two possible behaviors: either $E(\lambda)\sim -c(\lambda-\lambda_0)^2$ or $E(\lambda)\sim -c(\lambda-\lambda_0)$ as $\lambda\downarrow\lambda_0$, with $c>0$, where the second possibility can only occur if zero is an eigenvalue of H_{λ_0}. This type of result was to be expected in light of previous work on the subject.[9]

II. ONE DIMENSION: GENERAL RESULTS

In this section we consider the operator family (1.1) on $L^2(\mathbf{R})$ under the assumptions that V and W are real-valued,

$$V\in L^1_{loc}(\mathbf{R}),\quad V(x+p)=V(x),\quad \text{for some } p>0, \tag{2.1}$$

and that W satisfies one of the following conditions:

$$\int_{-\infty}^{\infty}(1+|x|)|W(x)|dx<\infty, \tag{H1}$$

$$\int_{-\infty}^{\infty}(1+x^2)|W(x)|dx<\infty, \tag{H2}$$

$$\int_{-\infty}^{\infty}e^{c|x|}|W(x)|dx<\infty,\quad c>0. \tag{H3}$$

Condition (2.1) implies that V is $-d^2/dx^2$-form bounded with relative bound zero (see Ref. 16, p. 8). Thus $H_0=-d^2/dx^2+V$ can be defined as a self-adjoint, semi-bounded operator by means of the KLMN theorem (see Ref. 17, Theorem VI.2.1; Ref. 18, Theorem X.17). Alternatively, the differential operator defined by H_0 is limit-point at $\pm\infty$. Hence H_0 can be interpreted as the unique self-adjoint extension of the minimal operator $H_{0,\min}$ with domain $D(H_{0,\min})=\{\psi\in L^2:\psi,\psi'\in AC_{loc}$, supp ψ compact, $H_0\psi\in L^2\}$ (see Ref. 19, p. 41). Here AC_{loc} denotes the set of locally absolutely continuous functions on $\mathbf{R}$. If (H1) holds, then the multiplication operator W is a relatively form compact perturbation of H_0 (cf. Ref. 20, pp. 114 and 369) and this is equivalent to the compactness of the Birman–Schwinger kernel. It is well-known that the spectrum of H_0 is absolutely continuous and of the form

$$\sigma(H_0)=\bigcup_{n=0}^{\infty}[E_{2n},E_{2n+1}],$$

where $-\infty<E_0<E_1\leqslant E_2<E_3\leqslant E_4\cdots$. So the intervals $[E_{2n},E_{2n+1}]$ are the bands and the intervals (E_{2n-1},E_n) (with $E_{-1}=-\infty$) are the gaps. In general a gap may be empty (E_{2n-1}

$=E_n$), but in this paper we will only be concerned with nonempty gaps. The spectral parameter E will always be assumed to lie in a gap or at an endpoint of a gap. We first summarize some well-known facts about the equation

$$H_0\psi = E\psi. \tag{2.2}$$

This equation has the two solutions (also known as Floquet solutions or Bloch waves),

$$\psi_0^{(\pm)}(x,E) = \theta_0(x,E) + m_0^{(\pm)}(E)\,\phi_0(x,E), \tag{2.3}$$

where $\theta_0(x,E)$ and $\phi_0(x,E)$ are solutions of (2.2) satisfying for all E the initial conditions $\theta_0(0,E) = \phi_0'(0,E) = 1$, $\theta_0'(0,E) = \phi_0(0,E) = 0$, and $m_0^{(\pm)}(E)$ are the Titchmarsh–Weyl m-functions associated with the intervals $\mathbf{R}_+ = [0,\infty)$ and $\mathbf{R}_- = (-\infty,0]$, respectively. Explicitly,

$$m_0^{(\pm)}(E) = \frac{\phi_0'(p,E) - \theta_0(p,E)}{2\phi_0(p,E)} \pm i\,\frac{\sin[q(E)]}{\phi_0(p,E)}, \tag{2.4}$$

$$q(E) = \cos^{-1}[\Delta(E)], \quad \Delta(E) = \frac{1}{2}\,[\phi_0'(p,E) + \theta_0(p,E)]. \tag{2.5}$$

Here $\Delta(E)$ is the Floquet determinant which is related to the spectrum of H_0 by $\sigma(H_0) = \{E : |\Delta(E)| \leqslant 1\}$. The branch of $\cos^{-1}$ in (2.5) is chosen such that $q(E)$ is continuous, $\mathrm{Im}[q(E)] > 0$ for $E < E_0$, and $q(E_0) = 0$. The solutions (2.3) are linearly independent for E inside a gap and satisfy

$$[\psi_0^{(+)}(\cdot,E); \psi_0^{(-)}(\cdot,E)] = -\frac{2i\,\sin[q(E)]}{\phi_0(p,E)}, \tag{2.6}$$

where $[f;g]$ denotes the Wronskian of f and g. Moreover, $\psi_0^{(\pm)}(x,E)$ can alternatively be written in the form

$$\psi_0^{(\pm)}(x,E) = e^{\pm ikx}\xi_0^{(\pm)}(x,E), \tag{2.7}$$

where $k = k(E) = q(E)/p$ is called the quasi-momentum, $\xi_0^{(\pm)}(x,E)$ have period p and $\xi_0^{(\pm)}(0,E) = 1$ [since $\psi_0^{(\pm)}(0,E) = 1$]. We further introduce the notation

$$\mathscr{G}_n = (E_{2n-1}, E_{2n}), \quad n = 0,1,2,\ldots,$$

$$\mathscr{B}_n = [E_{2n-2}, E_{2n-1}], \quad n = 1,2,\ldots,$$

so that $\mathscr{G}_n$ denotes the nth gap starting with the zeroth gap $(-\infty, E_0)$, and $\mathscr{B}_n$ denotes the nth band. The function $k(E)$ maps[21,22] $\mathscr{B}_n$ onto the interval $[(n-1)\pi/p, n\pi/p]$ in the k-plane, and it maps $\mathscr{G}_0$ onto the positive imaginary axis, and each $\mathscr{G}_n$ $(n = 1,2,\ldots)$ onto a segment of the form $\mathscr{S}_n = \{k : k = n\pi/p + i\mu, 0 < \mu \leqslant \delta_n\}$ for some $\delta_n > 0$, where $\mu = \mu(E) = \mathrm{Im}[k(E)]$. Here $\delta_n > 0$ is determined by the equation $k(\widetilde{E}_n) = n\pi/p + i\delta_n$, where $\widetilde{E}_n$ is the unique point (cf. Ref. 23, p. 295) in $\mathscr{G}_n$ such that $\Delta'(\widetilde{E}_n) = 0$. The two intervals $(E_{2n-1}, \widetilde{E}_n]$ and $[\widetilde{E}_n, E_{2n})$ are both mapped onto the segment $\mathscr{S}_n$ and hence the function $\mu \mapsto E(n\pi/p + i\mu)$ has two branches, $E^{(1)}(n\pi/p + i\mu)$ and $E^{(2)}(n\pi/p + i\mu)$, such that $E^{(1)}(n\pi/p + i\mu) \to E_{2n-1}$ and $E^{(2)}(n\pi/p + i\mu) \to E_{2n}$ as $\mu \downarrow 0$. Moreover, the functions $\mu \mapsto E^{(j)}(n\pi/p + i\mu)$ $(j = 1,2)$ are analytic, even and satisfy

$$E^{(j)}(n\pi/p + i\mu) - E_{2n-2+j} = \frac{(-1)^n p^2}{2\Delta'(E_{2n-2+j})}\,\mu^2 + O(\mu^4), \quad \mu \downarrow 0. \tag{2.8}$$

This follows from (2.5) by expanding $\Delta(E)$ about the point E_{2n-2+j}, using $\Delta(E_{2n-2+j}) = (-1)^n$ and the fact that $\phi_0(x,E)$ and $\theta_0(x,E)$ are analytic functions of E, hence even functions of μ. From now on we assume that

$$\phi_0(p,E_n) \neq 0, \tag{2.9}$$

at the endpoint under consideration. The case $\phi_0(p,E_n)=0$ will be discussed separately, since it requires special arguments. Note that for $n=0$, (2.9) is automatically satisfied because E_0 is the lowest point of $\sigma(H_0)$. In fact, $\phi_0(p,E_0)=0$ implies that $\phi_0(x,E_0)$ is periodic (with period p) and hence has infinitely many zeros, contradicting that E_0 is the lowest point of $\sigma(H_0)$. Moreover, the solution $\psi_0^{(+)}(x,E_0)$, which is well-defined because $\phi_0(p,E_0)\neq 0$, is periodic and thus has no zeros. Hence, by the Sturm separation theorem, $\phi_0(x,E_0)$ has exactly one zero (at $x=0$). There are two solutions, $F_\lambda^{(\pm)}(x,E)$, of (1.4) which are defined as solutions of the integral equations

$$F_\lambda^{(\pm)}(x,E)=\psi_0^{(\pm)}(x,E)-\lambda\int_x^{\pm\infty}A_0(x,y;E)W(y)F_\lambda^{(\pm)}(y,E)dy, \tag{2.10}$$

$$A_0(x,y;E)=-\frac{i\phi_0(p,E)}{2\sin[q(E)]}\left[\psi_0^{(+)}(x,E)\psi_0^{(-)}(y,E)-\psi_0^{(-)}(x,E)\psi_0^{(+)}(y,E)\right]. \tag{2.11}$$

The solutions $F_\lambda^{(\pm)}(x,E)$ correspond to the standard Jost solutions in the case when $V=0$ and, as usual, (2.10)–(2.11) are obtained by applying the variation of constants formula to (1.4) and using (2.6). For reasons given below, the relevant case for us will be that in which for a fixed $\lambda\geqslant 0$, $F_\lambda^{(+)}(x,E_n)$ and $F_\lambda^{(-)}(x,E_n)$ are linearly dependent. This case will also be called the exceptional case and we will call λ an exceptional value (at a given endpoint E_n) if $[F_\lambda^{(+)}(\cdot,E_n);F_\lambda^{(-)}(\cdot,E_n)]=0$. In the exceptional case we define the constant a_n by

$$F_\lambda^{(+)}(x,E_n)=a_n\,F_\lambda^{(-)}(x,E_n). \tag{2.12}$$

The case when $[F_\lambda^{(+)}(\cdot,E_n);F_\lambda^{(-)}(\cdot,E_n)]\neq 0$ will be called the generic case.

Next we introduce the transmission coefficient $T_\lambda(E)$ associated with (1.4),

$$T_\lambda(E)=-\frac{2i\,\sin[q(E)]}{[F_\lambda^{(+)}(\cdot,E);F_\lambda^{(-)}(\cdot,E)]\phi_0(p,E)}. \tag{2.13}$$

The transmission coefficient will enter naturally into our analysis and bring the advantage that we can draw on recent work on its asymptotic properties.[24] We remark that usually $T_\lambda(E)$ is first defined for $E\in\sigma(H_0)$ and then analytically continued into the gaps. The results of Ref. 24 were shown to be valid when E lies in a gap.

It will sometimes be convenient to number the gaps by using the symbol

$$\tilde{n}=\left[\frac{n+1}{2}\right], \tag{2.14}$$

where $[x]$ denotes the largest integer less than or equal to x. Then E_n is the left (right) endpoint of the gap $\mathscr{G}_{\tilde{n}}$ if n is odd (even). The statement $E\to E_n$ is understood to mean that E approaches E_n from within the gap $\mathscr{G}_{\tilde{n}}$, where E_n may be either endpoint of $\mathscr{G}_{\tilde{n}}$. In terms of μ this can be simply stated as $\mu\downarrow 0$, with the understanding that E is viewed as a function of μ, namely, $E=E^{(j)}(\tilde{n}\pi/p+i\mu)$, with the proper branch labeled by j ($j=n+2-2\tilde{n}$).

The following theorem was proved in Ref. 24 (Theorem 3.3).

Theorem 2.1: Suppose that (H1) and (2.9) hold. Fix $\lambda\geqslant 0$. Then $T_\lambda(E)$ is continuous at every E_n and there are two possible asymptotic behaviors of $T_\lambda(E)$ as $\mu\downarrow 0$ ($E\in\mathscr{G}_{\tilde{n}}$):

(i) In the generic case,

$$T_\lambda(E)=\beta_n\mu+o(\mu), \tag{2.15}$$

with β_n real and nonzero.

(ii) In the exceptional case,

$$T_\lambda(E)=T_\lambda(0)+o(1),\quad T_\lambda(0)=\frac{2a_n}{1+a_n^2}. \tag{2.16}$$

Of course, (2.15) follows immediately from (2.13), along with the fact that $F_\lambda^{(\pm)}(x,E_n)$ are real and

$$\sin[q(E)]=i(-1)^{\tilde{n}}\sinh \mu p, \quad E\in\mathscr{G}_{\tilde{n}}. \tag{2.17}$$

The proof of (2.16) is more involved. Note that if $\lambda=0$, then $F_0^{(\pm)}(x,E)=\psi_0^{(\pm)}(x,E)$, so $T_0(E)=1$ for all E, and we are in case (ii) with $a_n=1$.

Lemma 2.2: Suppose that (H1) and (2.9) hold. Then $[F_\lambda^{(+)}(\cdot,E_n);F_\lambda^{(-)}(\cdot,E_n)]=0$ if and only if

$$\int_{-\infty}^{\infty}\psi_0^{(+)}(x,E_n)W(x)F_\lambda^{(+)}(x,E_n)dx=0. \tag{2.18}$$

Proof: This result was proved in Ref. 24 as a consequence of Lemma 2.1 there. We briefly sketch the argument here, since we will use one formula from the proof later. For $E=E_n$ the integral equation (2.10) for $F_\lambda^{(+)}(x,E_n)$ becomes

$$F_\lambda^{(+)}(x,E_n)=\psi_0^{(+)}(x,E_n)-\lambda\int_x^{\infty}A_0(x,y;E_n)W(y)F_\lambda^{(+)}(y,E_n)dy, \tag{2.19}$$

$$A_0(x,y;E_n)=\lim_{E\to E_n}A_0(x,y;E)=\phi_0(x,E_n)\psi_0^{(+)}(y,E_n)-\psi_0^{(+)}(x,E_n)\phi_0(y,E_n). \tag{2.20}$$

Since $[F_\lambda^{(+)}(\cdot,E_n);F_\lambda^{(-)}(\cdot,E_n)]=0$ if and only if $H_\lambda\psi=E_n\psi$ has a bounded (nontrivial) solution, the assertion (2.18) follows on letting $x\to-\infty$ in (2.19) and using standard estimates (see Ref. 24, or Lemma 2.3 below). ∎

For the following it is convenient to normalize the solutions $F_\lambda^{(\pm)}(x,E)$ by introducing

$$\psi_\lambda^{(\pm)}(x,E)=\frac{F_\lambda^{(\pm)}(x,E)}{F_\lambda^{(\pm)}(0,E)}, \tag{2.21}$$

so that $\psi_\lambda^{(\pm)}(0,E)=1$. Therefore, in addition to (2.9), we make the assumption

$$F_\lambda^{(\pm)}(0,E_n)\neq 0, \tag{2.22}$$

so that by continuity $F_\lambda^{(\pm)}(0,E)\neq 0$ for E near E_n. It turns out that (2.22) is not a serious restriction, since by shifting the origin if necessary, we can always make sure that it holds. This point will be discussed in more detail later when we remove condition (2.9) by the same method.

The integral kernel of the resolvent $(H_\lambda-E)^{-1}$, i.e., the Green's function of H_λ, will be denoted by $G_{\lambda,E}(x,y)$. It is given by

$$G_{\lambda,E}(x,y)=\frac{1}{[\psi_\lambda^{(+)}(\cdot,E);\psi_\lambda^{(-)}(\cdot,E)]}\begin{cases}\psi_\lambda^{(+)}(x,E)\psi_\lambda^{(-)}(y,E), & y<x,\\ \psi_\lambda^{(-)}(x,E)\psi_\lambda^{(+)}(y,E), & y>x.\end{cases} \tag{2.23}$$

For fixed λ and E inside a gap, the Wronskian in the denominator of (2.23) vanishes exactly when E is an eigenvalue of H_λ, and so it is nonzero for E near but different from an endpoint E_n. The Birman–Schwinger kernel (1.3) is given by

$$K_{\lambda,E}(x,y)=W(x)^{1/2}G_{\lambda,E}(x,y)|W(y)|^{1/2}, \quad W(x)^{1/2}=|W(x)|^{1/2}\operatorname{sgn}W(x).$$

In order to split $K_{\lambda,E}$ into a singular and a regular part we write

$$G_{\lambda,E}(x,y)=\frac{1}{G_{\lambda,E}(0,0)}G_{\lambda,E}(x,0)G_{\lambda,E}(0,y)+G^D_{\lambda,E}(x,y). \tag{2.24}$$

where $G^D_{\lambda,E}(x,y)$ is the Green's function of H_λ with an additional Dirichlet condition at the origin. The decomposition (2.24) was used in Ref. 15 to study the number of eigenvalues in $\mathscr{G}_n$ for n large. Then the corresponding decomposition of $K_{\lambda,E}$ is

$$K_{\lambda,E}=d_\lambda(E)L_{\lambda,E}+R_{\lambda,E}, \tag{2.25}$$

where $L_{\lambda,E}$ and $R_{\lambda,E}$ are integral operators with kernels $L_{\lambda,E}(x,y)$ and $R_{\lambda,E}(x,y)$, respectively, given by

$$\begin{aligned} L_{\lambda,E}(x,y)&=\omega_\lambda(x,E)\tilde{\omega}_\lambda(y,E),\\ R_{\lambda,E}(x,y)&=W(x)^{1/2}G^D_{\lambda,E}(x,y)|W(y)|^{1/2}, \end{aligned} \tag{2.26}$$

$$\begin{aligned} \omega_\lambda(x,E)&=\begin{cases} W(x)^{1/2}\psi^{(+)}_\lambda(x,E), & x>0,\\ W(x)^{1/2}\psi^{(-)}_\lambda(x,E), & x<0, \end{cases}\\ \tilde{\omega}_\lambda(y,E)&=\omega_\lambda(y,E)\operatorname{sgn} W(y), \end{aligned} \tag{2.27}$$

$$G^D_{\lambda,E}(x,y)=\begin{cases} \phi_\lambda(x,E)\psi^{(+)}_\lambda(y,E), & 0<x<y,\\ \psi^{(+)}_\lambda(x,E)\phi_\lambda(y,E), & 0<y<x,\\ -\phi_\lambda(x,E)\psi^{(-)}_\lambda(y,E), & y<x<0,\\ -\psi^{(-)}_\lambda(x,E)\phi_\lambda(y,E), & x<y<0,\\ 0, \quad \text{otherwise.} \end{cases} \tag{2.28}$$

Moreover,

$$d_\lambda(E)=G_{\lambda,E}(0,0)=\frac{1}{[\psi^{(+)}_\lambda(\cdot,E);\psi^{(-)}_\lambda(\cdot,E)]}, \tag{2.29}$$

so that, by (2.6) and (2.17),

$$d_0(E)=\frac{(-1)^{\tilde{n}}\phi_0(p,E)}{2\sinh\,\mu p}, \quad E\in\mathscr{G}_{\tilde{n}}. \tag{2.30}$$

Another useful expression for $d_\lambda(E)$ follows from (2.13), (2.17), (2.21), and (2.29), namely,

$$d_\lambda(E)=\frac{(-1)^{\tilde{n}}\phi_0(p,E)T_\lambda(E)F^{(+)}_\lambda(0,E)F^{(-)}_\lambda(0,E)}{2\sinh\,\mu p}, \quad E\in\mathscr{G}_{\tilde{n}}. \tag{2.31}$$

The implicit equation (1.8) for $E(\lambda)$ can be written as

$$[\lambda-\lambda_0](\tilde{\omega}_{\lambda_0,E},(1+[\lambda-\lambda_0]R_{\lambda_0,E})^{-1}\omega_{\lambda_0,E})=-\frac{1}{d_{\lambda_0}(E)}, \tag{2.32}$$

where λ_0 is a c.c.th. and $\omega_{\lambda,E}$ (resp., $\tilde{\omega}_{\lambda,E}$) denotes the function $\omega_\lambda(x,E)$ [resp., $\tilde{\omega}_\lambda(x,E)$] defined in (2.27). Notice that the decomposition (2.25) is of the form (1.5), (1.6) with $N=1$ if we set $\lambda=\lambda_0$, $d_{\lambda_0;1}(E)=d_{\lambda_0}(E)$, $L_{\lambda_0,E;1}=L_{\lambda_0,E}$, $\omega_{\lambda_0,E;1}=\omega_{\lambda_0,E}$, etc.

Our goal now is to find solutions $E=E(\lambda)$ of (2.32) such that $E(\lambda)\to E_n$ as $\lambda\to\lambda_0$. We do this by first finding the solutions in the form $\mu(\lambda)$ and then using (2.8) to obtain $E(\lambda)$. Expanding the left-hand side of (2.32) in powers of $\lambda-\lambda_0$ yields

$$\sum_{j=0}^{\infty}(-1)^j[\lambda-\lambda_0]^{j+1}(\tilde{\omega}_{\lambda_0,E},R^j_{\lambda_0,E}\omega_{\lambda_0,E})=-\frac{1}{d_{\lambda_0}(E)}. \tag{2.33}$$

Now we would like to expand the coefficients $(\tilde{\omega}_{\lambda_0,E}, R^j_{\lambda_0,E}\omega_{\lambda_0,E})$ and the right-hand side of (2.33) in powers of μ and then obtain an expansion of $\mu(\lambda)$ in powers of $\lambda-\lambda_0$. It is not hard to see that a full expansion will generally not be possible under assumptions (H1) or (H2); one needs stronger assumptions like (H3) to do this. Under (H1) or (H2) one can only expect to find a few terms in an expansion of $\mu(\lambda)$. In particular, since the right-hand side of (2.33) vanishes linearly in μ, it will be easy to obtain the leading order term of $\mu(\lambda)$ provided the coefficient with $j=0$, $(\tilde{\omega}_{\lambda_0,E_n}, \omega_{\lambda_0,E_n})$, is nonzero. However, we are also interested in the case when $(\tilde{\omega}_{\lambda_0,E_n}, \omega_{\lambda_0,E_n}) = 0$, and, more generally, in the case when $(\tilde{\omega}_{\lambda_0,E_n}, R^j_{\lambda_0,E_n}\omega_{\lambda_0,E_n}) = 0$ for $j=0,1,\ldots,M-1$ and $(\tilde{\omega}_{\lambda,E_n}, R^M_{\lambda_0,E_n}\omega_{\lambda_0,E_n}) \neq 0$ for some M. Or, what if $(\tilde{\omega}_{\lambda_0,E_n}, R^j_{\lambda_0,E_n}\omega_{\lambda_0,E_n}) = 0$ for all j, if this is possible? In order to answer these questions we need some technical lemmas. Since the proofs of these lemmas are somewhat lengthy but mostly standard and do not contribute much to the understanding of our main results, they have all been consigned to the Appendix. In the following we will always think of E as a function of μ [as explained below (2.14)]. The phrase ''for E near E_n'' shall mean that there is a $\mu_0>0$ sufficiently small, so that the given statement holds for $0 \leqslant \mu \leqslant \mu_0$. The letter C will be used to denote various constants that may depend on μ_0 (but not on μ) and are not necessarily the same at each appearance. Differentiation with respect to μ will be denoted by a dot.

Lemma 2.3: Assume (H1), (2.9), (2.22), and fix $\lambda \geqslant 0$. Then for E near E_n we have the estimates

(i) $|\psi^{(\pm)}_\lambda(x,E)| \leqslant Ce^{\mp\mu x}$, $x \in \mathbf{R}_\pm$,

(ii) $|\phi_\lambda(x,E)| \leqslant C(1+|x|)e^{\mu|x|}$,

(iii) $|\dot{\phi}_\lambda(x,E)| \leqslant C[\mu(1+|x|)^3/(1+\mu|x|)^2]e^{\mu|x|}$.

Furthermore, if $\lambda=0$ and (H1) holds, or if $\lambda>0$ and (H2) holds, then

(iv) $|\dot{\psi}^{(\pm)}_\lambda(x,E)| \leqslant C(1+|x|)e^{\mp\mu x}$, $x \in \mathbf{R}_\pm$.

Proof: See the Appendix. ■

It is true that for E near but different from E_n, the function $\psi^{(\pm)}_\lambda(x,E)$ is differentiable with respect to μ under condition (H1) alone, for any $\lambda \geqslant 0$. However, if $\lambda>0$ and (H2) is not satisfied, then $\dot{\psi}_\lambda(x,E)$ will in general not remain bounded as $\mu \downarrow 0$. In other words, estimate (iv) no longer holds. The evidence for this conclusion comes from a result in Ref. 25 (Theorem 3.1) where, in the case $V=0$, it was proved that for a certain class of potentials that satisfy (H1) but violate (H2), the Jost solutions are in general not differentiable with respect to k $(k=\sqrt{-E})$ at $k=0$.

Lemma 2.4: Assume (H1), (2.9), (2.22) and fix $\lambda \geqslant 0$. Then for E near E_n we have

(i) $G^D_{\lambda,E}(x,y) \leqslant C[1+\min\{|x|,|y|\}]$,

(ii) $R_{\lambda,E} \to R_{\lambda,E_n}$ in Hilbert–Schmidt norm as $E \to E_n$,

(iii) Assume $\lambda=0$ and (H1), or $\lambda>0$ and (H2). Then

$$|\dot{G}^D_{\lambda,E}(x,y)| \leqslant C(1+|x|)(1+|y|). \tag{2.34}$$

Proof: See the Appendix. ■

We note that (i) implies

$$|G^D_{\lambda,E}(x,y)| \leqslant C(1+|x|)^{1/2}(1+|y|)^{1/2}.$$

When $\lambda=0$ a bound of this type was used in Ref. 15. Under condition (H1) all we can say in general is that $\|R_{\lambda,E}-R_{\lambda,E_n}\|_{H.S.}=o(1)$ as $E \to E_n$. However, if (H2) holds, then (2.34) implies that $\dot{R}_{\lambda,E}(x,y)$ is Hilbert–Schmidt and hence $\|R_{\lambda,E}-R_{\lambda,E_n}\|_{H.S.}=O(\mu)$.

Lemma 2.5: In addition to (2.9) and (2.22) assume $\lambda=0$ and (H1), or $\lambda>0$ and (H2). Then, for E near E_n, the functions $\mu \to (\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})$, $s=0,1,2,\ldots$, are continuously differentiable and obey

(i) $|(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})| \leqslant \tilde{C}r^s$,

(ii) $|d(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})/d\mu| \leqslant \tilde{C}r^s$,

for some $\tilde{C}>0$ and $r>0$.

Proof: See the Appendix. ■

The next two lemmas are essential for our study of (2.33). They are variants of the inverse function theorem tailored to our needs.

Lemma 2.6: Suppose $a_j(\mu)$, $j=1,2,...$, and $c(\mu)$ are continuous real-valued functions defined for $0\leqslant\mu\leqslant\mu_0$, satisfying

(a) $|a_j(\mu)|\leqslant Dr^j$, $j=1,2,3,...$, for some $D>0$, $r>0$, and $0\leqslant\mu\leqslant\mu_0$,

(b) $a_j(\mu)=O(\mu)$ as $\mu\to 0$, for $j=1,2,...N-1$, and $a_N(0)\neq 0$ for some $N\geqslant 1$,

(c) $c(\mu)=\dot{c}(0)\mu+o(\mu)$, with $\dot{c}(0)>0$, as $\mu\to 0$,

and consider the equation

$$\sum_{j=1}^{\infty} a_j(\mu)z^j=c(\mu). \tag{2.35}$$

Then the following occur:

(i) If N is odd and $a_N(0)>0$, then there is a unique, real, continuous solution $z(\mu)$ of (2.35) such that $z(\mu)=[\dot{c}(0)/a_N(0)]^{1/N}\mu^{1/N}+o(\mu^{1/N})$ as $\mu\downarrow 0$.

(ii) If N is odd and $a_N(0)<0$, then there is a unique, real, continuous solution $z(\mu)$ such that $z(\mu)=-[-\dot{c}(0)/a_N(0)]^{1/N}\mu^{1/N}+o(\mu^{1/N})$ as $\mu\downarrow 0$.

(iii) If N is even and $a_N(0)>0$, then there are exactly two real, continuous solutions $z_{\pm}(\mu)$ such that $z_{\pm}(\mu)=\pm[\dot{c}(0)/a_N(0)]^{1/N}\mu^{1/N}+o(\mu^{1/N})$ as $\mu\downarrow 0$.

(iv) If N is even and $a_N(0)<0$, then (2.35) has no real solution converging to zero as $\mu\downarrow 0$.

(v) If $a_j(0)=0$ for all j and $|a_j(\mu)|\leqslant C_1 r_1^j\mu$ for some $C_1>0$ and $r_1>0$, then there is no real solution converging to zero as $\mu\downarrow 0$.

Proof: See the Appendix. ■

Lemma 2.7: Suppose that, in addition to (a)–(c) of Lemma 2.6, $a_j(\mu)$ ($j=1,2,...$) and $c(\mu)$ are continuously differentiable, and $|\dot{a}_j(\mu)|\leqslant C_2 r_2^j$, for some $C_2>0$ and $r_2>0$. Then we have the following:

(i) If N is even and $a_N(0)>0$, then (2.35) has a unique solution $\mu(z)$ such that

$$\mu(z)=\frac{a_N(0)}{\dot{c}(0)}\,z^N+o(z^N), \quad z\to 0. \tag{2.36}$$

(ii) If N is odd and $a_N(0)>0$, resp. $a_N(0)<0$, then (2.35) has a solution obeying (2.36) as $z\downarrow 0$, resp. $z\uparrow 0$.

(iii) Suppose that $c(\mu)=\dot{c}(0)\mu+c_2\mu^2+o(\mu^2)$ with $\dot{c}(0)>0$. Then, under the same assumptions on $a_N(0)$ as in (i) and (ii) and with z approaching zero in the same manner as there, we have that

$$\mu(z)=\frac{a_N(0)}{\dot{c}(0)}\,z^N+\left[\frac{a_{N+1}(0)}{\dot{c}(0)}+\frac{a_N(0)\dot{a}_1(0)}{\dot{c}(0)^2}-\frac{c_2 a_1(0)^2}{\dot{c}(0)^3}\,\delta_{N,1}\right]z^{N+1}+o(z^{N+1}). \tag{2.37}$$

In (2.37) the Kronecker symbol $\delta_{N,1}$ has been added only for clarity. It follows from assumption (b) of Lemma 2.6 that $a_1(0)=0$ when $N>1$; thus the coefficient c_2 has no effect on (2.37) when $N>1$. Note that if $a_N(0)=0$, then (2.37) just reduces to (2.36) but with $N+1$ in place of N.

Proof: See the Appendix. ■

Some technical comments on Lemmas 2.6 and 2.7 may be in order. In assumption (c) of Lemma 2.6 we do not intend to imply that $c(\mu)$ is continuously differentiable on $[0,\mu_0]$, only that besides being continuous, it is also differentiable at $\mu=0$. In the application of Lemma 2.6 we will actually need the inverse function $\mu(z)$ of $z(\mu)$ [in case (iii) we need the inverse of each of the two functions $z_{\pm}(\mu)$]. Lemma 2.6 does not guarantee the existence of these inverse functions. For example, the solution $z(\mu)$ of the equation $[1+2\mu\,\sin(1/\mu)]z=\mu(0<\mu\leqslant\mu_0)$ does not have an inverse on $(0,\mu_0]$, no matter how small μ_0. However, in the applications we will have additional information on $\mu(z)$ which will allow us to argue that the inverse function exists. It may be also useful to give a short illustration of what can happen if $a_j(\mu)$ and $c(\mu)$ violate some of the conditions placed on them in Lemmas 2.6 and 2.7. Consider, for example, the equation $z+z^2=\mu+\mu^{1+\epsilon}$ with $1/2<\epsilon<1$, i.e., $a_1(\mu)=a_2(\mu)=1$, $c(\mu)=\mu+\mu^{1+\epsilon}$, and so $\dot{c}(0)=1$. Thus $c(\mu)$ does not fulfill condition (iii) of Lemma 2.7. Then $\mu(z)=z-z^{1+\epsilon}+z^2+o(z^2)$ and we see that a term, $-z^{1+\epsilon}$, has slipped between the linear and quadratic terms so that (2.37) no longer holds. Now replace $a_1(\mu)$ by μ, so that the equation becomes $\mu z+z^2=\mu+\mu^{1+\epsilon}$. Then $\mu(z)$

$=z^2-z^{2(1+\epsilon)}+O(z^3)$ which shows that the ϵ has no effect on the leading term. This is in agreement with Lemma 2.6. Furthermore, if $a_1(\mu)$ does not vanish linearly, uniqueness or existence may break down. For example, the equation $5\sqrt{\mu}z-4z^2=\mu$ has two solutions $\mu(z)=z^2$ and $\mu(z)=16z^2$, and the equation $\mu^{2/3}z-z^2=\mu$ has no real non-negative solution $\mu(z)$ that converges to zero as $z\to 0$.

Theorem 2.8: Assume (2.9) and (2.22). If $\lambda_0>0$ is a c.c.th. at E_n, then λ_0 is an exceptional value. The exceptional values and the c.c.th.'s each form a discrete set with no finite accumulation points.

Proof: Suppose that $[F^{(+)}_{\lambda_0}(\cdot,E_n);F^{(-)}_{\lambda_0}(\cdot,E_n)]\neq 0$. Then, by (2.15) and (2.31), the right-hand side of (2.33) approaches a finite nonzero limit as $E\to E_n$, whereas, by Lemma 2.5, the left-hand side goes to zero as $\lambda\to\lambda_0$, uniformly in E for E near E_n. Therefore, a solution $\mu(\lambda)$ of (2.33) satisfying $\mu(\lambda)\to 0$ as $\lambda\to\lambda_0$ does not exist and hence λ_0 cannot be a c.c.th. This proves the first part. For the second part we allow λ to be complex. Solving (2.19) by iteration, we see that for each x, the functions $\lambda\mapsto F^{(\pm)}_{\lambda}(x,E_n)$ and $\lambda\mapsto F^{(\pm)'}_{\lambda}(x,E_n)$ are entire functions of λ. Hence $\lambda\mapsto[F^{(+)}_{\lambda}(\cdot,E_n);F^{(-)}_{\lambda}(\cdot,E_n)]$ is entire and the conclusion follows. ■

In connection with Theorem 2.8 we remind the reader that $\lambda_0=0$ is always an exceptional value, but, as we will see below, it need not always be a c.c.th. [Theorem 2.11(i)]. Moreover, the restrictions (2.9) and (2.22) can be removed and the first assertion of Theorem 2.8 has a converse if $n=0$ and, under further assumptions, also when $n\geqslant 1$. This will be shown in Theorems 2.14(iii) and 2.16(i).

From now on we will always assume that we are in the exceptional case. In order to determine the threshold behavior of the eigenvalues we have to identify (2.33) with (2.35). Since $\dot{c}(0)$ in Lemma 2.6 must be positive, we need to know the sign of the derivative of $-1/d_{\lambda_0}(E)$ with respect to μ. To this end, we first note that by (2.12), (2.16), and (2.31),

$$-\frac{1}{d_{\lambda_0}(E)}=\frac{(-1)^{\tilde{n}+1}(1+a_n^2)p}{\phi_0(p,E_n)F^{(+)}_{\lambda_0}(0,E_n)^2}\,\mu+o(\mu). \tag{2.38}$$

Now it is known[14] that if (2.9) holds, then

$$\phi_0(p,E_n)>0, \quad n=4j \quad \text{and} \quad n=4j+1,$$

$$\phi_0(p,E_n)<0, \quad n=4j+2 \quad \text{and} \quad n=4j+3, \ j=0,1,\ldots .$$

Therefore,

$$\operatorname{sgn}[(-1)^{\tilde{n}+1}\phi_0(p,E_n)]=(-1)^{n+1},$$

and if we set

$$c_{\lambda_0}(\mu)=\frac{(-1)^n}{d_{\lambda_0}(E)}, \tag{2.39}$$

then

$$\dot{c}_{\lambda_0}(0)=\frac{(-1)^{\tilde{n}+n}(1+a_n^2)p}{\phi_0(p,E_n)F^{(+)}_{\lambda_0}(0,E_n)^2}>0. \tag{2.40}$$

In order to achieve some computational simplications later, we introduce the kernel

$$M_{\lambda,E}(x,y)=d_{\lambda}(E)W(x)^{1/2}|W(y)|^{1/2}\cdot\begin{cases}B_{\lambda}(x,y;E), & y<x,\\ B_{\lambda}(y,x;E), & y>x,\end{cases}$$
$$B_{\lambda}(x,y;E)=\psi^{(+)}_{\lambda}(x,E)\psi^{(-)}_{\lambda}(y,E)-\psi^{(+)}_{\lambda}(x,E_n)\psi^{(+)}_{\lambda}(y,E_n), \tag{2.41}$$

and denote the associated integral operator by $M_{\lambda,E}$. We remark that the integral operator $M_{\lambda,E}$ satisfies

$$K_{\lambda,E}=d_\lambda(E)L_{\lambda,E_n}+M_{\lambda,E}, \tag{2.42}$$

and thus is associated with a decomposition of $K_{\lambda,E}$ slightly different from that in (2.25). When $V=0$ this decomposition corresponds to that used in Refs. 1 and 5. If we assume (H2), then Lemma 2.4(ii) also holds for $M_{\lambda,E}$, namely $M_{\lambda,E}\to M_{\lambda,E_n}$ in the Hilbert–Schmidt norm. Then we could use the implicit equation

$$[\lambda-\lambda_0](\tilde{\omega}_{\lambda_0,E_n},(1+[\lambda-\lambda_0]M_{\lambda_0,E})^{-1}\omega_{\lambda_0,E_n})=-\frac{1}{d_{\lambda_0}(E)},$$

instead of (2.32). In this paper we need (2.41) only when $E=E_0$ and $\lambda=0$. Then we have

$$M_{0,E_0}(x,y)=\frac{-W(x)^{1/2}|W(y)|^{1/2}}{2}\cdot\begin{cases}A_0(x,y;E_0), & y<x,\\ A_0(y,x;E_0), & y>x,\end{cases} \tag{2.43}$$

where $A_0(x,y;E_0)$ is defined in (2.20). If only (H1) is assumed, then the kernel (2.43) will in general not represent a bounded operator. However, here we will only need certain matrix elements like $(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})$. This matrix element exists as an integral over $\mathbf{R}^2$ and is equal to $\lim_{E\to E_0}(\tilde{\omega}_{0,E_0},M_{0,E}\omega_{0,E_0})$, where this limit exists by the estimates in Lemma 2.3.

In preparation of our next theorem we first prove two lemmas. Let λ_0 be an exceptional value and define

$$I_{\lambda_0,j}(E)=(\tilde{\omega}_{\lambda_0,E},R^j_{\lambda_0,E}\omega_{\lambda_0,E}),\quad j=0,1,2,...; \tag{2.44}$$

in particular,

$$I_{\lambda_0,0}(E_n)=(\tilde{\omega}_{\lambda_0,E_n},\omega_{\lambda_0,E_n})=\int_{-\infty}^{\infty}W(x)\psi^{(+)}_{\lambda_0}(x,E_n)^2dx. \tag{2.45}$$

Lemma 2.9: Assume (H1), (2.9), $n=0$, and $\lambda_0=0$. Then

$$(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})=(\tilde{\omega}_{0,E_0},R_{0,E_0}\omega_{0,E_0})+\frac{1}{\dot{c}_0(0)}I_{0,0}(E_0)\dot{I}_{0,0}(E_0). \tag{2.46}$$

Proof: By (2.25), (2.39), (2.42), and (2.44), we have

$$\begin{aligned}(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})-(\tilde{\omega}_{0,E_0},R_{0,E_0}\omega_{0,E_0})&=\lim_{E\to E_0}d_0(E)(\tilde{\omega}_{0,E_0},[L_{0,E}-L_{0,E_0}]\omega_{0,E_0})\\&=2\frac{1}{\dot{c}_0(0)}(\tilde{\omega}_{0,E_0},\omega_{0,E_0})(\tilde{\omega}_{0,E_0},\dot{\omega}_{0,E_0})\\&=\frac{1}{\dot{c}_0(0)}I_{0,0}(E_0)\dot{I}_{0,0}(E_0),\end{aligned}$$

proving (2.46). In the last equation we have used $\dot{I}_{0,0}(E)=2(\tilde{\omega}_{0,E},\dot{\omega}_{0,E})$. Note that, by Lemma 2.3(iv), $\dot{I}_{0,0}(E_0)$ exists under condition (H1). ■

Lemma 2.10: Assume (H1), $W\neq 0$, and $n=0$. Then the following are true:

(i) If $\lambda_0>0$, $[F^{(+)}_{\lambda_0}(\cdot,E_0);F^{(-)}_{\lambda_0}(\cdot,E_0)]=0$, and (2.22) holds, then $I_{\lambda_0,0}(E_0)<0$.

(ii) If $\lambda_0=0$ and $I_{0,0}(E_0)=0$, then $I_{0,1}(E_0)=(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})>0$.

As we shall see later, in (i) it is not really necessary to assume (2.22). The only purpose of this assumption is to ensure that $\psi^{(+)}_{\lambda_0}(x,E_0)$ is defined by (2.21).

Proof: (i) Since $F^{(\pm)}_{\lambda_0}(x,E_0)$ are linearly dependent, (2.18) holds with $n=0$ and $\lambda=\lambda_0$. Multiplying (2.19) for $n=0$ by $W(x)F^{(+)}_{\lambda_0}(x,E_0)$, integrating over $\mathbf{R}$ using an integration by parts and Lemma 2.2, and using (2.21) and (2.45) yields

$$F^{(+)}_{\lambda_0}(0,E_0)^2 I_{\lambda_0,0}(E_0)=\int_{-\infty}^{\infty} W(x)F^{(+)}_{\lambda_0}(x,E_0)^2 dx$$

$$=-2\lambda_0\int_{-\infty}^{\infty} W(x)F^{(+)}_{\lambda_0}(x,E_0)\phi_0(x,E_0)$$

$$\times\left(\int_x^{\infty} W(y)\psi_0^{(+)}(y,E_0)F^{(+)}_{\lambda_0}(y,E_0)dy\right)dx. \tag{2.47}$$

Now the relation $[\phi_0(x,E_0)/\psi_0^{(+)}(x,E_0)]'=1/\psi_0^{(+)}(x,E_0)^2$ [recall that $\psi_0^{(+)}(x,E_0)$ has no zeros] together with an integration by parts yields

$$F^{(+)}_{\lambda_0}(0,E_0)^2 I_{\lambda_0,0}(E_0)=-\lambda_0\int_{-\infty}^{\infty}\frac{1}{\psi_0^{(+)}(x,E_0)^2}$$

$$\times\left(\int_x^{\infty} W(y)\psi_0^{(+)}(y,E_0)F^{(+)}_{\lambda_0}(y,E_0)dy\right)^2 dx<0. \tag{2.48}$$

The boundary terms from the integration by parts vanish on account of Lemma 2.3 and (H1). Hence $I_{\lambda_0,0}(E_0)<0$, proving (i).

(ii) By (2.45) and (2.46), since $I_{0,0}(E_0)=0$, we have that $I_{0,1}(E_0)=(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})$. Then, from (2.43), by an integration by parts and using that $\psi_0^{(+)}(x,E_0)=\psi_0^{(-)}(x,E_0)$, we obtain

$$I_{0,1}(E_0)=2\int_{-\infty}^{\infty} W(x)\psi_0^{(+)}(x,E_0)\phi_0(x,E_0)\left(\int_x^{\infty} W(y)\psi_0^{(+)}(y,E_0)^2 dy\right)dx. \tag{2.49}$$

Proceeding with (2.49) as with (2.47), we obtain

$$I_{0,1}(E_0)=\int_{-\infty}^{\infty}\frac{1}{\psi_0^{(+)}(x,E_0)^2}\left(\int_x^{\infty} W(y)\psi_0^{(+)}(y,E_0)^2 dy\right)^2 dx>0,$$

and (ii) is proved. ∎

In the following the constant

$$\nu_n=\frac{|\phi_0(p,E_n)|F^{(+)}_{\lambda_0}(0,E_n)^2}{(1+a_n^2)\sqrt{2|\Delta'(E_n)|}} \tag{2.50}$$

will frequently appear. We do not explicitly indicate the dependence on λ_0 of this and similar constants.

Theorem 2.11: Assume (H1), $W\neq 0$, and $n=0$. Then we have the following:

(i) $\lambda_0=0$ is a c.c.th. at E_0 if and only if $I_{0,0}(E_0)\leqslant 0$. If this is the case, then there exists a unique eigenvalue $E(\lambda)$ such that

$$\sqrt{E_0-E(\lambda)}=-\nu_0 I_{0,0}(E_0)\lambda+\nu_0(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0})\lambda^2+o(\lambda^2),\quad \lambda\downarrow 0. \tag{2.51}$$

(ii) Suppose that $\lambda_0>0$, $[F^{(+)}_{\lambda_0}(\cdot,E_0);F^{(-)}_{\lambda_0}(\cdot,E_0)]=0$, and that (2.22) holds. Then λ_0 is a c.c.th., $I_{\lambda_0,0}(E_0)<0$, and there is a unique eigenvalue given by

$$\sqrt{E_0-E(\lambda)}=-\nu_0 I_{\lambda_0,0}(E_0)(\lambda-\lambda_0)+o(\lambda-\lambda_0),\quad \lambda\downarrow\lambda_0. \tag{2.52}$$

Proof: Suppose λ_0 is a c.c.th. at E_0. Then we identify (2.33) with (2.35) by means of the substitutions $1/d_{\lambda_0}(E_0)=c_{\lambda_0}(\mu)\to c(\mu)$ [cf. (2.39) with $n=0$], $\lambda-\lambda_0\to z$, and $(-1)^{j+1}(\tilde{\omega}_{\lambda_0,E},R^j_{\lambda_0,E}\omega_{\lambda_0,E})=(-1)^{j+1}I_{\lambda_0,j}(E)\to a_{j+1}(\mu)$ $(j=0,1,...)$. To prove (i), first note that, since $\psi_0^{(+)}(x,E_0)=\psi_0^{(-)}(x,E_0)$, $\lambda_0=0$ is an exceptional value at each E_0. By Lemma 2.6(ii), if $a_1(0)<0$, then (2.35) does not have an acceptable (i.e. positive) solution $z(\mu)$. Hence $a_1(0)=-I_{0,0}(E_0)\geqslant 0$ is necessary for $\lambda_0=0$ to be a c.c.th. To show the sufficiency of the condition $I_{0,0}(E_0)\leqslant 0$ we note that by Lemma 2.10(ii), if $I_{0,0}(E_0)=0$, then $I_{0,1}(E_0)=a_2(0)>0$. So either $a_1(0)>0$ or $a_1(0)=0$ and $a_2(0)>0$. Moreover, Lemma 2.5(ii) guarantees that the coefficients $a_k(\mu)$ are continuously differentiable, and from (2.29) and (2.39) it follows that $c(\mu)$ is continuously differentiable, $\dot{c}(0)>0$ [cf. (2.40)] and $c_2=0$ [because $\phi_0(p,E)$ is an even function of μ]. Hence Lemma 2.7(iii) applies and $\mu(z)$ is given by (2.37). This shows that $\lambda_0=0$ is a c.c.th. It remains to convert (2.37) to (2.51). Considering the bracketed term in (2.37), using (2.46), we have

$$\frac{a_2(0)}{\dot{c}(0)}+\frac{a_1(0)\dot{a}_1(0)}{\dot{c}(0)^2}=\frac{I_{0,1}(E_0)}{\dot{c}(0)}+\frac{I_{0,0}(E_0)\dot{I}_{0,0}(E_0)}{\dot{c}(0)^2}=\frac{1}{\dot{c}(0)}\,(\tilde{\omega}_{0,E_0},M_{0,E_0}\omega_{0,E_0}). \tag{2.53}$$

Now (2.51) follows from (2.37), (2.53), and using (2.8) to relate μ to $\sqrt{E_0-E(\lambda)}$. This proves (i).

(ii) When $\lambda_0>0$, then $a_1(0)=-I_{\lambda_0,0}(E_0)>0$ by Lemma 2.10(i), and hence $z(\mu)=[\dot{c}(0)/a_1(0)]\mu+o(\mu)$ by Lemma 2.6(ii). Note that Lemma 2.7 is not applicable, since $a_k(\mu)$ and $c(\mu)$ are not necessarily continuously differentiable at $\mu=0$. Now we show that the function $z(\mu)$ is invertible for small enough μ. We argue by contradiction. Suppose that $z(\mu_1)=z(\mu_2)$ for $0<\mu_1<\mu_2$. Then $z(\mu)$ assumes an extremum at an interior point of this interval, say at μ^*. If $z(\mu^*)$ is a strict extremum, then there are two sequences $\{\mu_{n,\pm}\}$ such that $\mu_{n,+}\downarrow\mu^*$, $\mu_{n,-}\uparrow\mu^*$, $z(\mu_{n,+})=z(\mu_{n,-})$, and $z(\mu_{n,\pm})\to z(\mu^*)$ as $\mu_{n,\pm}\to\mu^*$. This means that, as $\lambda\to\lambda_0+z(\mu^*)$, two eigenvalue sequences $\{E(i\mu_{n,\pm})\}$ belonging to H_{λ_n} with $\lambda_n=\lambda_0+z(\mu_{n,+})=\lambda_0+z(\mu_{n,-})$ converge to the eigenvalue $E(i\mu^*)$ of $H_{\lambda_0+z(\mu^*)}$. Consequently, by standard results from perturbation theory, $H_{\lambda_0+z(\mu^*)}$ has a discrete eigenvalue of multiplicity two. But this is impossible; hence $z(\mu)$ cannot have a strict minimum at μ^*. If the extremum at μ^* is not strict, then there exists a sequence $\{\mu_n\}$ such that $\mu_n\to\mu^*$ and $z(\mu_n)=z(\mu^*)$. This implies that the values $E(i\mu^*)$ and $E(i\mu_n)$ are all eigenvalues of $H_{\lambda_0+z(\mu^*)}$ and that $E(i\mu^*)$ is an accumulation point of these eigenvalues. Again, this is impossible. Hence $z(\mu)$ is a monotone function. Inverting it we get $\mu(z)=[a_1(0)/\dot{c}(0)]z+o(z)$ and (2.52) follows from (2.8). ■

In the statement of the next theorem, in order to reduce the number of case distinctions, we temporarily lift the restriction $\lambda\geqslant 0$ to the extent that we allow λ to approach 0 from below if the c.c.th. is $\lambda_0=0$.

Theorem 2.12: Suppose that $\lambda_0=0$ or $\lambda_0>0$ and $[F^{(+)}_{\lambda_0}(\cdot,E_n);F^{(-)}_{\lambda_0}(\cdot,E_n)]=0$, $n\geqslant 1$, and that (2.9) and (2.22) hold. Also, suppose that for some $M\geqslant 0$,

$$I_{\lambda_0,j}(E_n)=0,\quad j=0,...,M-1,\quad I_{\lambda_0,M}(E_n)\neq 0. \tag{2.54}$$

Further, assume one of the following:

(a) $\lambda_0=0$ and (H1),

(b) $\lambda_0>0$, $M=0$, and (H1),

(c) $\lambda_0>0$, $M\geqslant 1$, and (H2).

Then the following holds:

(i) If n is odd, M is even, and $I_{\lambda_0,M}(E_n)>0$ [resp., $I_{\lambda_0,M}(E_n)<0$], then there is a unique eigenvalue $E(\lambda)$ of H_λ obeying

$$\sqrt{E(\lambda)-E_n}=\nu_n I_{\lambda_0,M}(E_n)(\lambda-\lambda_0)^{M+1}[1+o(1)],\quad \lambda\downarrow\lambda_0\quad(\text{resp. }\lambda\uparrow\lambda_0). \tag{2.55}$$

(ii) If n is odd, M is odd, and $I_{\lambda_0,M}(E_n)<0$, then there is an eigenvalue obeying

$$\sqrt{E(\lambda)-E_n}=-\nu_n I_{\lambda_0,M}(E_n)(\lambda-\lambda_0)^{M+1}[1+o(1)],\quad \lambda\downarrow\lambda_0\quad\text{and}\quad\lambda\uparrow\lambda_0. \tag{2.56}$$

(iii) If n is odd, M is odd, and $I_{\lambda_0,M}(E_n)>0$, then there is no eigenvalue converging to E_n as either $\lambda\downarrow\lambda_0$ or $\lambda\uparrow\lambda_0$.

(iv) If n is even, M is even, and $I_{\lambda_0,M}(E_n)<0$ [resp. $I_{\lambda_0,M}(E_n)>0$], then there is a unique eigenvalue $E(\lambda)$ of H_λ obeying

$$\sqrt{E_n-E(\lambda)}=-\nu_n I_{\lambda_0,M}(E_n)(\lambda-\lambda_0)^{M+1}[1+o(1)],\quad \lambda\downarrow\lambda_0\quad (\text{resp. } \lambda\uparrow\lambda_0). \tag{2.57}$$

(v) If n is even, M is odd and $I_{\lambda_0,M}(E_n)>0$, then there is an eigenvalue obeying

$$\sqrt{E_n-E(\lambda)}=\nu_n I_{\lambda_0,M}(E_n)(\lambda-\lambda_0)^{M+1}[1+o(1)],\quad \lambda\downarrow\lambda_0\quad \text{and}\quad \lambda\uparrow\lambda_0. \tag{2.58}$$

(vi) If n is even, M is odd, and $I_{\lambda_0,M}(E_n)<0$, then there is no eigenvalue converging to E_n as either $\lambda\downarrow\lambda_0$ or $\lambda\uparrow\lambda_0$.

Proof: Apply Lemma 2.7 if (a) or (c) are assumed and use Lemma 2.6, together with the argument about inverting $z(\mu)$ used in the proof of Theorem 2.11, when (b) is assumed. In identifying (2.33) with (2.35) use (2.39) and the correspondence $(-1)^{j+n+1}I_{\lambda_0,j}(E)\to a_{j+1}(\mu)$; further details are omitted. ■

Some comments about Theorems 2.11 and 2.12 may be in order. According to (2.51) and (2.52) eigenvalues can approach E_0 only as $\lambda\downarrow\lambda_0$. In other words, eigenvalues that appear at E_0 can only move to the left as λ increases. This is well-known and not only true for λ near λ_0 but for all $\lambda>\lambda_0$ (Ref. 20, p. 79; Ref. 26, footnote on p. 89). For an endpoint E_n with n odd, i.e., the left endpoint of a gap, part (i) of Theorem 2.12 describes the two situations where an eigenvalue either appears at E_n as $\lambda\uparrow\lambda_0+\epsilon$ [if $I_{\lambda_0,M}(E_n)>0$] or where an eigenvalue comes in from the right and gets absorbed into the continuous spectrum at E_n [if $I_{\lambda_0,M}(E_n)<0$]. Case (ii) describes the situation where an eigenvalue ''turns around'' at E_n. Note that $M+1$ is even and hence $M+1\geqslant 2$, so that $E(\lambda)-E_n=O(\lambda^4)$ at least. In case (iii), λ_0 is an exceptional value, but there is no eigenvalue converging to E_n as $\lambda\downarrow\lambda_0$ or $\lambda\uparrow\lambda_0$. In other words, λ_0 is an exceptional value but not a c.c.th. For n even, the possibilities are similar as can be seen from (iv)–(vi).

Theorem 2.13: Suppose that W satisfies (H3). Then $\sqrt{E(\lambda)-E_n}$ is given by a convergent power series in $\lambda-\lambda_0$.

Proof: Both sides of (2.32) can be analytically continued, as functions of μ, to a complex neighborhood of $\mu=0$. This follows by generalizing the estimates in Lemmas 2.3 and 2.4 to the case when μ (resp., E) is complex. We omit a detailed discussion and only sketch a few steps. The estimates $|\phi_{\lambda_0}(x,E)|\leqslant C(1+|x|)e^{|\mathrm{Re}\,\mu||x|}$ and $|\psi^{(\pm)}_{\lambda_0}(x,E)|\leqslant Ce^{\mp(\mathrm{Re}\,\mu)x}$ for $x\in\mathbf{R}_\pm$ imply that $|G^D_{\lambda_0,E}(x,y)|\leqslant C[1+\min\{|x|,|y|\}]e^{|\mathrm{Re}\,\mu|(|x|+|y|)}$. The exponential factor appearing here will be absorbed by the decay of W when used in the expressions for $R_{\lambda_0,E}$ and $\omega_{\lambda_0,E}$, provided $2|\mathrm{Re}\,\mu|<c$, where c is the constant in (H3). As a result, the function $h(\mu,z)$ in (A30) (see the Appendix) is jointly analytic in μ and z near $\mu=z=0$, $h(0,0)=0$, and $h_\mu(0,0)\neq 0$. Therefore, by the implicit function theorem for analytic functions, the equation $h(\mu,z)=0$ has, for small z, a unique analytic solution $\mu(z)$. ■

Next we discuss how the above results have to be modified when (2.9) or (2.22) are not satisfied, that is, when $\phi_0(p,E_n)=0$, or $\phi_0(p,E_n)\neq 0$ but $F^{(\pm)}_{\lambda_0}(0,E_n)=0$. These special situations may easily occur for reasons of symmetry. For example, if V and W are even, then the subspaces of odd and even parity, respectively, are reducing subspaces for H_λ. Hence on the subspace of odd parity we have $F^{(\pm)}_{\lambda_0}(0,E_n)=0$ at the threshold. Moreover, either E_1 or E_2 is the first Dirichlet eigenvalue for the eigenvalue problem (2.2) on the interval $[0,p]$ and thus $\phi_0(p,E_j)=0$ for $j=1$ or 2. That our approach breaks down when $\phi_0(p,E_n)=0$ is evident from (2.4) which shows that the Titchmarsh–Weyl coefficient $m_0^{(+)}(E)$ [or $m_0^{(-)}(E)$] becomes singular as $E\to E_n$; the imaginary part of $m_0^{(+)}(E)$ blows up, while the real part approaches a finite limit. Note that $\phi_0(p,E)$ vanishes quadratically in μ as $\mu\downarrow 0$, while $\sin[q(E)]$ only vanishes linearly. This implies that the decomposition (2.25) is no longer useful. One way to deal with this situation would be to choose a different decomposition of the Green's function, and this is also necessary if one wants to give a complete treatment of the case $\phi_0(p,E_n)=0$. However, here we are interested only in

obtaining the leading terms of some of the eigenvalue expansions, and so we choose a different approach. We perform a shift of the origin so that for the shifted problem conditions (2.9) and (2.22) are satisfied. Then we relate the expressions for the shifted problem back to the original problem. We will often refer to Ref. 24, where some details of this method have already been worked out.

Let $b \in \mathbf{R}$, and replace $V(x)$ and $W(x)$ in (1.1) and (1.2) by $V(x+b)$ and $W(x+b)$. Clearly, the spectrum of H_λ is invariant under the shift. We denote the various solutions associated with the shifted problem by means of an additional argument or subscript b, e.g., $\psi_\lambda^{(+)}(x,E;b)$, $\phi_\lambda(x,E;b)$, $\nu_{n;b}$, etc. From (2.45) and (2.50) we have

$$\nu_{n;b} I_{\lambda_0,0}(E_n;b) = \frac{|\phi_0(p,E_n;b)|}{(1+a_n^2)\sqrt{2|\Delta'(E_n)|}} \int_{-\infty}^{\infty} W(x+b) F_{\lambda_0}^{(+)}(x,E_n;b)^2 dx, \tag{2.59}$$

where we used (2.21) for the shifted problem and the fact that the constant a_n [defined in (2.12)] and the discriminant $\Delta(E)$ are invariant under the shift. We now distinguish between two cases.

Case 1: $F_{\lambda_0}^{(\pm)}(0,E_n)=0$ and $\phi_0(p,E_n)\neq 0$: Since

$$\psi_0^{(\pm)}(x,E;b) = \frac{\psi_0^{(\pm)}(x+b,E)}{\psi_0^{(\pm)}(b,E)},$$

$$F_{\lambda_0}^{(\pm)}(x,E;b) = \psi_0^{(\pm)}(x,E;b) + o(1), \quad x \to \pm\infty,$$

it follows that

$$F_{\lambda_0}^{(\pm)}(x,E;b) = \frac{F_{\lambda_0}^{(\pm)}(x+b,E)}{\psi_0^{(\pm)}(b,E)}. \tag{2.60}$$

So we can choose b such that $\psi_0^{(\pm)}(b,E_n)\neq 0$ and $F_{\lambda_0}^{(\pm)}(b,E_n)\neq 0$, and hence

$$F_{\lambda_0}^{(\pm)}(0,E_n;b) = \frac{F_{\lambda_0}^{(\pm)}(b,E_n)}{\psi_0^{(\pm)}(b,E_n)} \neq 0.$$

Furthermore, we have [cf. Ref. 24, Eqs. (3.37)–(3.38)]

$$\phi_0(p,E_n;b) = \psi_0^{(+)}(b,E_n)^2 \phi_0(p,E_n) \neq 0, \tag{2.61}$$

so that (2.9) and (2.22) are both satisfied for the shifted problem. Since $F_{\lambda_0}^{(\pm)}(0,E_n)=0$, we have

$$F_{\lambda_0}^{(\pm)}(x,E_n) = F_{\lambda_0}^{(\pm)'}(0,E_n)\,\phi_{\lambda_0}(x,E_n), \tag{2.62}$$

and so (2.59) can be written as

$$\nu_{n;b} I_{\lambda_0,0}(E_n;b) = \tilde{\nu}_n \tilde{I}_{\lambda_0,0}(E_n), \tag{2.63}$$

$$\tilde{\nu}_n = \frac{|\phi_0(p,E_n)| F_{\lambda_0}^{(+)'}(0,E_n)^2}{(1+a_n^2)\sqrt{2|\Delta'(E_n)|}}, \quad \tilde{I}_{\lambda_0,0}(E_n) = \int_{-\infty}^{\infty} W(x)\,\phi_{\lambda_0}(x,E_n)^2 dx. \tag{2.64}$$

Finally, we find an expression for $I_{\lambda_0,1}(E_n;b) = (\tilde{\omega}_{\lambda_0,E_n;b}, R_{\lambda_0,E_n;b}\omega_{\lambda_0,E_n;b})$ under the assumption that $\tilde{I}_{\lambda_0,0}(E_n)=0$, so that we can also determine the leading order term in the eigenvalue expansion when $\tilde{I}_{\lambda_0,0}(E_n)=0$. By using (2.26) and $\tilde{I}_{\lambda_0,0}(E_n)=0=I_{\lambda_0,0}(E_n;b)$, we obtain

$$I_{\lambda_0,1}(E_n;b)=2\int_{-\infty}^{\infty}W(x+b)\,\phi_{\lambda_0}(x,E_n;b)\,\psi_{\lambda_0}^{(+)}(x,E_n;b)\left(\int_x^{\infty}W(y+b)\,\psi_{\lambda_0}^{(+)}(y,E_n;b)^2dy\right)dx. \tag{2.65}$$

Now using

$$\phi_{\lambda_0}(x,E_n;b)=\theta_{\lambda_0}(b,E_n)\,\phi_{\lambda_0}(x+b,E_n)-\phi_{\lambda_0}(b,E_n)\,\theta_{\lambda_0}(x+b,E_n), \tag{2.66}$$

along with (2.60), (2.61), (2.62) for $x=b$, and $\tilde{I}_{\lambda_0,0}(E_n)=0$, we get

$$\begin{aligned}\nu_{n;b}I_{\lambda_0,1}(E_n;b)&=\tilde{\nu}_n\tilde{I}_{\lambda_0,1}(E),\\ \tilde{I}_{\lambda_0,1}(E_n)&=-2\int_{-\infty}^{\infty}W(x)\,\phi_{\lambda_0}(x,E_n)\,\theta_{\lambda_0}(x,E_n)\left(\int_x^{\infty}W(y)\,\phi_{\lambda_0}(y,E_n)^2dy\right)dx.\end{aligned} \tag{2.67}$$

Case 2: $\phi_0(p,E_n)=0 (n\geqslant 1)$: In place of (2.61) we now have [cf. Ref. 24, Eq. (3.37)]

$$\phi_0(p,E_n;b)=-\phi_0(b,E_n)^2\theta_0'(p,E_n), \tag{2.68}$$

and we see that by choosing b such that $\phi_0(b,E_n)\neq 0$ we can make sure that $\phi_0(p,E_n;b)\neq 0$. The exceptional case is now characterized by the property that $H_{\lambda_0}\psi=E_n\psi$ has a unique (up to constant factors) nontrivial solution. This is the same as saying that the shifted problem is exceptional in the sense defined below Theorem 2.1. Assuming we are in the exceptional case, following Ref. 24 we introduce the solutions

$$\eta_{\lambda_0}^{(\pm)}(x,E)=\frac{F_{\lambda_0}^{(\pm)}(x,E)}{m_0^{(\pm)}(E)}, \tag{2.69}$$

of $H_{\lambda_0}\psi=E\psi$, and note that

$$\eta_{\lambda_0}^{(\pm)}(x,E_n)=\lim_{E\to E_n}\eta_{\lambda_0}^{(\pm)}(x,E) \tag{2.70}$$

exist, and that $\eta_{\lambda_0}^{(\pm)}(x,E_n)$ represent bounded nontrivial solutions of $H_{\lambda_0}\psi=E_n\psi$. Moreover, $\eta_{\lambda_0}^{(+)}(x,E_n)=b_n\eta_{\lambda_0}^{(-)}(x,E_n)$ for some constant b_n, analogous to (2.12), and $\eta_0^{(\pm)}(x,E_n)=\phi_0(x,E_n)$ if $\lambda_0=0$ and $\eta_{\lambda_0}^{(\pm)}(x,E_n)=\phi_0(x,E_n)+o(1)$ as $x\to\pm\infty$ if $\lambda_0>0$. These asymptotics follow from the integral equation that $\eta_{\lambda_0}^{(\pm)}(x,E_n)$ satisfy and which can be obtained from (2.69) and (2.10):

$$\eta_{\lambda_0}^{(\pm)}(x,E_n)=\phi_0(x,E_n)-\lambda_0\int_x^{\pm\infty}\tilde{A}_0(x,y;E_n)W(y)\,\eta_{\lambda_0}^{(\pm)}(y,E_n)dy, \tag{2.71}$$

where $\tilde{A}_0(x,y;E_n)=\phi_0(x,E_n)\,\theta_0(y,E_n)-\theta_0(x,E_n)\,\phi_0(y,E_n)$. Note that $\tilde{A}_0(x,y;E_n)=\lim_{E\to E_n}A_0(x,y;E)$, where $A_0(x,y;E)$ is defined in (2.11), as can be seen by inserting (2.3) and (2.4) in (2.11). $\tilde{A}_0(x,y;E_n)$ replaces the kernel in (2.20) which becomes formally undefined because $\phi_0(p,E_n)=0$. Since $\psi_0^{(\pm)}(b,E)/m_0^{(\pm)}(E)\to\phi_0(b,E_n)$ as $E\to E_n$, by (2.60), (2.69), and (2.70) we obtain

$$F_{\lambda_0}^{(\pm)}(x,E_n;b)=\frac{\eta_{\lambda_0}^{(\pm)}(x+b,E_n)}{\phi_0(b,E_n)},\quad \psi_{\lambda_0}^{(+)}(x,E_n;b)=\frac{\eta_{\lambda_0}^{(+)}(x+b,E_n)}{\eta_{\lambda_0}^{(+)}(b,E_n)}, \tag{2.72}$$

and hence

$$F_{\lambda_0}^{(+)}(x,E_n;b)=b_nF_{\lambda_0}^{(-)}(x,E_n;b). \tag{2.73}$$

Upon inserting (2.68), (2.72), and (2.73) in (2.59), we find that

$$\nu_{n;b} I_{\lambda_0,0}(E_n;b) = \kappa_n J_{\lambda_0,0}(E_n), \tag{2.74}$$

$$\kappa_n = \frac{|\theta_0'(p,E_n)|}{(1+b_n^2)\sqrt{2|\Delta'(E_n)|}}, \quad J_{\lambda_0,0}(E_n) = \int_{-\infty}^{\infty} W(x)\,\eta_{\lambda_0}^{(+)}(x,E_n)^2 dx. \tag{2.75}$$

As in Case 1, by assuming $J_{\lambda_0,0}(E_n)=0$, we can obtain an expression for the product $\nu_{n;b} I_{\lambda_0,1}(E_n;b)$ which we state without a detailed derivation:

$$\nu_{n;b} I_{\lambda_0,1}(E_n;b) = \kappa_n J_{\lambda_0,1}(E_n), \tag{2.76}$$

$$J_{\lambda_0,1}(E_n) = 2\int_{-\infty}^{\infty} W(x) h_{\lambda_0}(x,E_n)\,\eta_{\lambda_0}^{(+)}(x,E_n)\left(\int_x^{\infty} W(y)\,\eta_{\lambda_0}^{(+)}(y,E_n)^2 dy\right) dx. \tag{2.77}$$

Here $h_{\lambda_0}(x,E_n)$ is any solution of $H_{\lambda_0}\psi = E_n\psi$ such that $[\eta_{\lambda_0}(\cdot,E_n); h_{\lambda_0}(\cdot,E_n)]=1$. This concludes our discussion of Case 2.

The expressions on the right-hand sides of (2.63) and (2.74) can be written in a common form that involves an arbitrary bounded, nontrivial solution $\tilde{\psi}_{\lambda_0}(x,E_n)$ of $H_{\lambda_0}\psi = E_n\psi$. If $\phi_0(p,E_n) \neq 0$, we have that $\tilde{\psi}_{\lambda_0}(x,E_n) = \tilde{\psi}_{\pm}\psi_0^{(\pm)}(x,E_n) + o(1)$ as $x \to \pm\infty$ for some nonzero constants $\tilde{\psi}_{\pm}$. Similarly, if $\phi_0(p,E_n)=0$, since $\tilde{\psi}_{\lambda_0}(x,E_n)$ is a multiple of $\eta_{\lambda_0}^{(+)}(x,E_n)$, there are constants $\tilde{\psi}_{\pm}$ such that $\tilde{\psi}_{\lambda_0}(x,E_n) = \tilde{\psi}_{\pm}\phi_0(x,E_n) + o(1)$ as $x \to \pm\infty$. If $\phi_0(p,E_n) \neq 0$, it follows that

$$\nu_n I_{\lambda_0,0}(E_n), \text{ resp. } \tilde{\nu}_n \tilde{I}_{\lambda_0,0}(E_n) = \frac{|\phi_0(p,E_n)|}{\sqrt{2|\Delta'(E_n)|}}\left[\frac{\int_{-\infty}^{\infty} W(x)\tilde{\psi}_{\lambda_0}(x,E_n)^2 dx}{\tilde{\psi}_+^2 + \tilde{\psi}_-^2}\right], \tag{2.78}$$

and, if $\phi_0(p,E_n)=0$, then

$$\kappa_n J_{\lambda_0,0}(E_n) = \frac{|\theta_0'(p,E_n)|}{\sqrt{2|\Delta'(E_n)|}}\left[\frac{\int_{-\infty}^{\infty} W(x)\tilde{\psi}_{\lambda_0}(x,E_n)^2 dx}{\tilde{\psi}_+^2 + \tilde{\psi}_-^2}\right]. \tag{2.79}$$

When $\lambda_0=0$, the right-hand sides of (2.78) and (2.79) can be expressed in yet another form, one that involves the effective mass. Recall [see Ref. 27 and (2.8)] that at the endpoint E_n the effective mass m_n^* is given by

$$m_n^* = \frac{|\Delta'(E_n)|}{p^2}. \tag{2.80}$$

Furthermore, as a consequence of the variation of constants formula [see (2.3.7) and (2.3.9) in Ref. 28] we have

$$\Delta'(E_n) = \begin{cases} -\dfrac{1}{2}\,\phi_0(p,E_n)\displaystyle\int_0^p \psi_0^{(+)}(x,E_n)^2 dx, & \phi_0(p,E_n)\neq 0, \\ \dfrac{1}{2}\,\theta_0'(p,E_n)\displaystyle\int_0^p \phi_0(x,E_n)^2 dx, & \phi_0(p,E_n)=0 \end{cases}. \tag{2.81}$$

If $\lambda_0=0$, then $\tilde{\psi}_0(x,E_n) = \tilde{\psi}_+\psi_0^{(+)}(x,E_n)$ [resp., $\tilde{\psi}_0(x,E_n) = \tilde{\psi}_+\phi_0^{(+)}(x,E_n)$] and $\tilde{\psi}_+ = \tilde{\psi}_-$. Therefore, from (2.78)–(2.81) it follows that

$$\nu_n I_{0,0}(E_n), \ \tilde{\nu}_n \tilde{I}_{0,0}(E_n), \text{ resp. } \kappa_n J_{0,0}(E_n) = \frac{\sqrt{m_n^*}\,p}{\sqrt{2}}\left[\frac{\int_{-\infty}^{\infty} W(x)\tilde{\psi}_0(x,E_n)^2 dx}{\int_0^p \tilde{\psi}_0(x,E_n)^2 dx}\right].$$

In the next theorem we specialize some of the results contained in Lemma 2.10, Theorem 2.11, and Theorem 2.12 to the situations described in Cases 1 and 2. We confine ourselves to the cases $M=0$ and $M=1$ [cf. (2.54)], where, in the present context, the case $M=0$ corresponds to either $\tilde{I}_{\lambda_0,0}(E_n)\neq 0$ or $J_{\lambda_0,0}(E_n)\neq 0$, and the case $M=1$ corresponds to either $\tilde{I}_{\lambda_0,0}(E_n)=0$ and $\tilde{I}_{\lambda_0,1}(E_n)\neq 0$, or $J_{\lambda_0,0}(E_n)=0$ and $J_{\lambda_0,1}(E_n)\neq 0$.

Theorem 2.14: Assume (H1), $W\neq 0$, and $\lambda\geqslant 0$.

(i) If $F^{(\pm)}_{\lambda_0}(0,E_0)=0$, then $\lambda_0>0$, $\tilde{I}_{\lambda_0,0}(E_0)<0$, and there exists a unique eigenvalue obeying

$$\sqrt{E_0-E(\lambda)}=-\tilde{\nu}_0\tilde{I}_{\lambda_0,0}(E_0)(\lambda-\lambda_0)+o(\lambda-\lambda_0),\quad \lambda\downarrow\lambda_0. \tag{2.82}$$

(ii) Suppose that either $\phi_0(p,E_n)\neq 0$, $n\geqslant 1$, $\lambda_0>0$, and $F^{(\pm)}_{\lambda_0}(0,E_n)=0$, or $\phi_0(p,E_n)=0$ and λ_0 is an exceptional value. Also assume (H2) if $M=1$. If $M=0$ or 1, then Theorem 2.12 applies with $I_{\lambda_0,M}(E_n)$ replaced by $\tilde{I}_{\lambda_0,M}(E_n)$ [resp., $J_{\lambda_0,M}(E_n)$], and the constants ν_n replaced by $\tilde{\nu}_n$ (resp., κ_n).

(iii) $\lambda_0>0$ is a c.c.th. at E_0 if and only if λ_0 is an exceptional value. If $\lambda_0>0$ is a c.c.th. at E_n with $n\geqslant 1$, then λ_0 is an exceptional value.

Proof: (i) Since $F^{(\pm)}_0(0,E_0)=1$, it follows that $\lambda_0>0$ and that λ_0 is an exceptional value. The rest follows from Lemma 2.10(i) and Theorem 2.11(ii) by applying a shift and using (2.63). (ii) follows from Theorem 2.12, along with (2.63), (2.67), (2.74), and (2.76). The assertion in (iii) concerning E_0 is a direct consequence of Theorems 2.8 and 2.11, and, in Case 1, of (i) and (ii). The statement concerning E_n with $n\geqslant 1$ has been proved in Theorem 2.8 when $\phi_0(p,E_n)\neq 0$ and $F^{(\pm)}_{\lambda_0}(0,E_n)\neq 0$. If one of these restrictions is dropped, then the result follows by using a shift as above and following the proof of Theorem 2.8. Note that the Wronskian $[F^{(+)}_{\lambda_0}(\cdot,E_n;b);\ F^{(-)}_{\lambda_0}(\cdot,E_n;b)]$ is independent of b by (2.60) and (2.72). ■

We remark that there are connections between some of our results concerning eigenvalue absorption at E_0 and earlier work by Gesztesy and Zhao.[29] In Ref. 29 the concepts of criticality, subcriticality, and supercriticality were defined for general second order linear differential operators. Without going into details, we note that in our context H_λ is critical if and only if $\lambda=0$ or $\lambda=\lambda_{0,\min}$, where $\lambda_{0,\min}=\min\{\lambda_0:\lambda_0$ is a c.c.th. at $E_0\}$. Moreover, assuming $\lambda_{0,\min}>0$, H_λ is subcritical if $0<\lambda<\lambda_{0,\min}$ and supercritical if $\lambda>\lambda_{0,\min}$. This follows (after some arguments) from Lemma 2.10 and Theorems 2.11 and 2.14 here, and Definition 3.2 and Theorem 3.6 in Ref. 29. Also, Lemma 3.18 in Ref. 29 and our Lemma 2.10(i) correspond to each other (ignoring the differences in the assumptions on V).

One may ask whether Lemma 2.10, Theorem 2.11, and Theorem 2.14 [(i) and (iii)] can be extended to endpoints other than E_0. It turns out that for compactly supported W, under some further assumptions on the location of the support of W relative to the zeros of the bounded (nontrivial) solution of $H_0\psi=E_n\psi$, this is possible. The fact that zeros of solutions play a role in threshold problems has been noted earlier (see Ref. 10, Theorem 3.2) and in a related context in Ref. 30 (Theorem 8.2).

Note that when $n\geqslant 1$, then any nontrivial solution of $H_0\psi=E_n\psi$ has infinitely many zeros. If $\phi_0(p,E_n)\neq 0$, let ζ_k, $k\in\mathbf{Z}$, denote the zeros of $\psi^{(+)}_0(x,E_n)$ arranged in ascending order, i.e., $\cdots<\zeta_{-1}<\zeta_0<\zeta_1<\cdots$. Similarly, if $\phi_0(p,E_n)=0$, let η_k denote the zeros of $\phi_0(x,E_n)$ arranged in ascending order.

Lemma 2.15: Assume (H1), $W\neq 0$, and $n\geqslant 1$. Let $\lambda_0\geqslant 0$ be an exceptional value at the endpoint E_n.

(i) Suppose $\phi_0(p,E_n)\neq 0$, $\lambda_0>0$, and supp $W\subset[\zeta_k,\zeta_{k+1}]$ for some fixed $k\in\mathbf{Z}$. Then $I_{\lambda_0,0}(E_n)<0$ [resp., $\tilde{I}_{\lambda_0,0}(E_n)<0$ if $F^{(\pm)}_{\lambda_0}(0,E_n)=0$]. If $\phi_0(p,E_n)=0$ and supp $W\subset[\eta_k,\eta_{k+1}]$ for some fixed $k\in\mathbf{Z}$, then $J_{\lambda_0,0}(E_n)<0$.

(ii) Suppose that $\lambda_0=0$. If $\phi_0(p,E_n)\neq 0$ and

$$\int_{\zeta_k}^{\zeta_{k+1}} W(x)\,\psi^{(+)}_0(x,E_n)^2\,dx=0, \tag{2.83}$$

for all $k \in \mathbf{Z}$, then $I_{0,0}(E_n)=0$ and $I_{0,1}(E_n)>0$.

If $\phi_0(p,E_n)=0$ and

$$\int_{\eta_k}^{\eta_{k+1}} W(x)\,\phi_0(x,E_n)^2 dx=0,$$

for all $k \in \mathbf{Z}$, then $J_{0,0}(E_n)=0$ and $J_{0,1}(E_n)>0$.

Proof: (i) If $\phi_0(p,E_n)\neq 0$, we can mimic the proof of Lemma 2.10(i) except that the relevant integrals now only involve $x \in [\zeta_k,\zeta_{k+1}]$. As in (2.47) and (2.49) we now get

$$\int_{\zeta_k}^{\zeta_{k+1}} W(x) F_{\lambda_0}^{(+)}(x,E_n)^2 dx<0,$$

so that $I_{\lambda_0,0}(E_n)<0$, resp., $\widetilde{I}_{\lambda_0,0}(E_n)<0$ [see (2.62) and (2.64)]. When $\phi_0(p,E_n)=0$, we use (2.71), the relation $[\theta_0(x,E_n)/\phi_0(x,E_n)]'=-1/\phi_0(x,E_n)^2$, and the analog of Lemma 2.2, namely, the fact that

$$\int_{\eta_k}^{\eta_{k+1}} \phi_0(x,E_n)\,W(x)\,\eta_{\lambda_0}^{(+)}(x,E_n)dx=0,$$

to derive

$$J_{\lambda_0,0}(E_n)=-\lambda_0\int_{\eta_k}^{\eta_{k+1}} \frac{1}{\phi_0(x,E_n)^2}\left(\int_x^{\eta_{k+1}} \phi_0(y,E_n)\,W(y)\,\eta_{\lambda_0}^{(+)}(y,E_n)dy\right)^2 dx<0.$$

This proves (i). Considering (ii), it is obvious by summing over all intervals $[\zeta_k,\zeta_{k+1}]$ that $I_{0,0}(E_n)=0$ [resp., $J_{0,0}(E_n)=0$]. If $\phi_0(p,E_n)\neq 0$ and (2.83) holds, then (2.49) becomes

$$I_{0,1}(E_n)=2\sum_{k\in\mathbf{Z}}\int_{\zeta_k}^{\zeta_{k+1}} W(x)\,\psi_0^{(+)}(x,E_n)\,\phi_0(x,E_n)\left(\int_x^{\zeta_{k+1}} W(y)\,\psi_0^{(+)}(y,E_n)^2 dy\right)dx.$$

Using $[\phi_0(x,E_n)/\psi_0^{(+)}(x,E_n)]'=1/\psi_0^{(+)}(x,E_n)^2$ on each interval $[\zeta_k,\zeta_{k+1}]$, along with an integration by parts, we obtain

$$I_{0,1}(E_n)=\sum_{k\in\mathbf{Z}}\int_{\zeta_k}^{\zeta_{k+1}} \frac{1}{\psi_0^{(+)}(x,E_n)^2}\left(\int_x^{\zeta_{k+1}} W(y)\,\psi_0^{(+)}(y,E_n)^2 dy\right)^2 dx>0. \quad \blacksquare$$

If $\phi_0(p,E_n)=0$, then we argue similarly by using (2.77) with $\lambda_0=0$, $\eta_0^{(+)}(x,E_n)=\phi_0(x,E_n)$, and $[h_0(x,E_n)/\phi_0(x,E_n)]'=1/\phi_0(x,E_n)^2$; thus $J_{0,1}(E_n)>0$. $\blacksquare$

Theorem 2.16: Suppose supp $W\subset[\zeta_k,\zeta_{k+1}]$, resp., supp $W\subset[\eta_k,\eta_{k+1}]$ if $\phi_0(p,E_n)=0$, for some $k \in \mathbf{Z}$ and $W\neq 0$. Then the following holds:

(i) $\lambda_0>0$ is a c.c.th. at E_n if and only if λ_0 is an exceptional value. Then $I_{\lambda_0,0}(E_n)<0$ [resp., $\widetilde{I}_{\lambda_0,0}(E_n)<0$, resp., $J_{\lambda_0,0}(E_n)<0$] and the corresponding eigenvalue obeys (2.55) as $\lambda\uparrow\lambda_0$, with $M=0$, if n is odd, and it obeys (2.57) as $\lambda\downarrow\lambda_0$, with $M=0$, if n is even.

(ii) If n is even, then $\lambda_0=0$ is a c.c.th. at E_n if and only if $I_{0,0}(E_n)\leqslant 0$ [resp., $J_{0,0}(E_n)\leqslant 0$]. The corresponding eigenvalue obeys (2.57) as $\lambda\downarrow\lambda_0$, with $M=0$, if $I_{0,0}(E_n)<0$, and it obeys (2.58) as $\lambda\downarrow\lambda_0$ and $\lambda\uparrow\lambda_0$, with $M=1$, if $I_{0,0}(E_n)=0$.

(iii) If n is odd, then $\lambda_0=0$ is a c.c.th. at E_n if and only if $I_{0,0}(E_n)>0$ [resp., $J_{0,0}(E_n)>0$]. The corresponding eigenvalue obeys (2.55) as $\lambda\downarrow\lambda_0$, with $M=0$.

Proof: (i) If $n=0$, then the result has already been established in Theorem 2.14(iii) together with Theorem 2.11(ii) and Theorem 2.14(i). If $n\geqslant 1$ and $\lambda_0>0$ is a c.c.th., then Theorem 2.14(iii) implies that λ_0 is an exceptional value. The converse follows from Lemma 2.15(i), Theorem 2.14(ii), and parts (i) and (iv) of Theorem 2.12. Note that Lemma 2.15(i) implies $I_{\lambda_0,0}(E_n)<0$ [resp., $\widetilde{I}_{\lambda_0,0}(E_n)<0$, resp., $J_{\lambda_0,0}(E_n)<0$], and so $M=0$. This proves (i). To prove (ii), suppose that $\lambda_0=0$ is a c.c.th. and that $I_{0,0}(E_n)>0$ [resp., $J_{0,0}(E_n)>0$]. Then, by Theorem 2.12(iv), an eigen-

value converging to E_n as $\lambda \downarrow 0$ cannot exist. Hence $I_{0,0}(E_n) \leqslant 0$ [resp., $J_{0,0}(E_n) \leqslant 0$]. The converse follows from Theorem 2.12(iv) if $I_{0,0}(E_n) < 0$ [resp., $J_{0,0}(E_n) < 0$], and from Lemma 2.15(ii) combined with Theorem 2.12(v) if $I_{0,0}(E_n) = 0$ [resp., $J_{0,0}(E_n) = 0$]. In (iii), if $I_{0,0}(E_n) > 0$ [resp., $J_{0,0}(E_n) > 0$], then $\lambda_0 = 0$ is a c.c.th. by Theorem 2.12(i). Conversely, if $\lambda_0 = 0$ is a c.c.th., then necessarily $I_{0,0}(E_n) \geqslant 0$ [resp., $J_{0,0}(E_n) \geqslant 0$]. However, if $I_{0,0}(E_n) = 0$ [resp., $J_{0,0}(E_n) = 0$], then Lemma 2.15(ii) implies that we are in case (iii) of Theorem 2.12 with $M = 1$. Hence $\lambda_0 = 0$ is not a c.c.th. This is a contradiction and thus $I_{0,0}(E_n) > 0$ [resp., $J_{0,0}(E_n) > 0$]. ■

Theorem 2.16(i) implies that, under the stated restrictions on the support of W, at the right endpoint of a gap an eigenvalue can only appear as $\lambda \uparrow \lambda_0 + \epsilon$; it cannot get absorbed as $\lambda \uparrow \lambda_0$. Similarly, at the left endpoint an eigenvalue can only get absorbed as $\lambda \uparrow \lambda_0$; there is no eigenvalue near the endpoint for λ slightly larger than λ_0.

Note that the points ζ_k (resp., η_k) are the zeros of the unique (up to constant multiples) bounded, nontrivial solution of $H_0\psi = E_n\psi$. Let $\tilde{\psi}_0(x, E_n)$ denote any such solution. It turns out that if we assume W to be non-negative, then the results of Theorem 2.16 can be strengthened as follows.

Theorem 2.17: Suppose that the support of W lies between two consecutive zeros of $\tilde{\psi}_0(x, E_n)$ and that $W \geqslant 0$ $(W \neq 0)$. If n is even, then there are no c.c.th.'s at E_n. If n is odd, then $\lambda_0 = 0$ is the only c.c.th. at E_n.

Proof: For any c.c.th. $\lambda_0 \geqslant 0$ we have $I_{\lambda_0,0}(E_n) > 0$ [resp., $\tilde{I}_{\lambda_0,0}(E_n) > 0$, resp., $J_{\lambda_0,0}(E_n) > 0$], by (2.45), (2.64), (2.75), and the assumption on W. Now the assertions follow from Theorem 2.16. ■

The final theorem of this section tells us what happens if W satisfies the support condition of Theorem 2.17 at both endpoints of a gap.

Theorem 2.18: Suppose that at each endpoint of the gap (E_n, E_{n+1}) (n odd) the support of W satisfies the condition of Theorem 2.17, and that $W \geqslant 0$ $(W \neq 0)$. Then there is a unique eigenvalue $E(\lambda)$ such that $\lim_{\lambda \downarrow 0} E(\lambda) = E_n$, $E(\lambda)$ is strictly increasing for $\lambda > 0$, $\lim_{\lambda \uparrow \infty} E(\lambda) = E_\infty$ exists, and $E_\infty \in (E_n, E_{n+1}]$.

Proof: By Theorem 2.17, $\lambda_0 = 0$ is the only c.c.th. at E_n. Moreover, by perturbation theory, since W is non-negative (and nontrivial), $E'(\lambda) > 0$ for $\lambda > 0$. Furthermore, the eigenvalue cannot get absorbed at E_{n+1} by Theorem 2.17. Thus $E(\lambda)$ has to converge to a limit as $\lambda \to +\infty$, and this limit must lie in $(E_n, E_{n+1}]$ (it may be the point E_{n+1}). ■

Theorem 2.18 describes the situation where an eigenvalue $E(\lambda)$ is ''trapped.'' The trapping phenomenon has been discussed in detail, and in greater generality, in Ref. 13. There, an interpretation of E_∞ as a Dirichlet eigenvalue of H_0 on $\mathbf{R} \backslash \operatorname{supp} W$ was given. Theorems 2.17 and 2.18 also fit together with more general results obtained in Ref. 11 (see, e.g., Corollary 3.2) which, in the present context, imply that the number of c.c.th.'s at a left endpoint of a gap is always finite, provided W has compact support and is non-negative. Some further results along these lines will appear elsewhere.

III. ONE DIMENSION: FURTHER RESULTS AND SPECIAL CASES

In this section we discuss some special aspects of the results of Section II and we establish the connection with related results that have appeared in the literature.

In Lemma 2.10(ii) and Lemma 2.15(ii), we were able to conclude that $I_{\lambda_0,1}(E_n) > 0$ provided that $I_{\lambda_0,0}(E_n) = 0$. So it is natural to ask whether it is possible that $I_{\lambda_0,0}(E_n) = I_{\lambda_0,1}(E_n) = 0$ ($W \neq 0$, of course). As the following example shows the answer is in the affirmative, at least when $\lambda_0 = 0$. We give the construction for the case when $\phi_0(p, E_n) \neq 0$, but it will be obvious that similar examples can be constructed if $\phi_0(p, E_n) = 0$, showing that $J_{0,0}(E_n) = J_{0,1}(E_n) = 0$ is possible. The construction of an example with $\lambda_0 > 0$ will not be attempted here. Suppose that V is even and W is odd. Then $\psi_0^{(+)}(x, E_n)$ is even, in fact $\psi_0^{(+)}(x, E_n) = \theta_0(x, E_n)$, $\phi_0(x, E_n)$ is odd, and hence $I_{0,0}(E_n) = 0$ by symmetry and (2.45). Furthermore, symmetry implies that $I_{0,j}(E_n) = 0$ for j even. Consider now a specific W of the form

$$W(x)=\begin{cases}1, & x\in[c,c+\epsilon],\\ -1, & x\in[-c-\epsilon,-c],\\ 0, & \text{otherwise},\end{cases}$$

where $c>0$ and $\epsilon>0$ are parameters. The function $h(x)=\int_x^\infty W(y)\psi_0^{(+)}(y,E_n)^2dy$ is even and non-negative. Hence, by symmetry, we can write (2.49) as

$$I_{0,1}(E_n)=4\int_c^{c+\epsilon}W(x)\psi_0^{(+)}(x,E_n)\phi_0(x,E_n)\left(\int_x^{c+\epsilon}W(y)\psi_0^{(+)}(y,E_n)^2dy\right)dx. \tag{3.1}$$

Now let ζ_1 denote the smallest positive zero of $\psi_0^{(+)}(x,E_n)$ and let η_1 denote the smallest positive zero of $\phi_0(x,E_n)$. Then $\zeta_1<\eta_1$ by the interlacing property of zeros. On $(0,\zeta_1)$ we have $\psi_0^{(+)}(x,E_n)>0$ and $\phi_0(x,E_n)>0$, and on (ζ_1,η_1) we have $\psi_0^{(+)}(x,E_n)<0$ and $\phi_0(x,E_n)>0$. Let $\epsilon<\min\{\zeta_1/2,(\eta_1-\zeta_1)/2\}$ and choose $c=\zeta_1/2$. Then $I_{0,1}(E_n)>0$. On the other hand, choosing $c=(\zeta_1+\eta_1)/2$ we get $I_{0,1}(E_n)<0$. So if we vary c from $\zeta_1/2$ to $(\zeta_1+\eta_1)/2$, then there exists a value c for which $I_{0,1}(E_n)=0$. Hence $I_{0,0}(E_n)=I_{0,1}(E_n)=0$ for this value of c. From the above argument we also see that for every sufficiently small $\epsilon>0$ there is a $c=c(\epsilon)$ such that $I_{0,1}(E_n)=0$ and that $c(\epsilon)\to\zeta_1$ as $\epsilon\to0$. The asymptotic behavior of the function $c(\epsilon)$ as $\epsilon\to0$ can be obtained by expanding the right-hand side of (3.1) in powers of ϵ and $c-\zeta_1$ using the approximations $\psi_0^{(+)}(x,E_n)=\psi_0^{(+)'}(\zeta_1,E_n)(x-\zeta_1)+o(x-\zeta_1)$ and $\phi_0(x,E_n)=\phi_0(\zeta_1,E_n)+o(1)$, and setting $I_{0,1}(E_n)=0$. Introducing variables r and φ such that $c-\zeta_1=r\cos\varphi$ and $\epsilon=r\sin\varphi$, it follows that

$$I_{0,1}(E_n)=\frac{[\psi_0^{(+)'}(\zeta_1,E_n)]^3\phi_0(\zeta_1,E_n)}{15}\cdot r^5f(\varphi)\sin^2\varphi+o(r^5),$$

$$f(\varphi)=30\cos\varphi+17\sin\varphi+11\sin3\varphi.$$

The relevant zero of $f(\varphi)$ is $\varphi_0=1.96784...$ which gives $c(\epsilon)=\zeta_1+\epsilon\cot\varphi_0+o(\epsilon)$, where $\cot\varphi_0=-0.41932...$. So $c(\epsilon)<\zeta_1$ and $c(\epsilon)+\epsilon>\zeta_1$ as it should be, since the support of W must contain the point ζ_1 in its interior. We have already mentioned that $I_{0,2}(E_n)=0$ by symmetry. We want to show that $I_{0,3}(E_n)<0$ for ϵ small and $c=c(\epsilon)$. This follows by writing

$$I_{0,3}(E_n)=4\int_c^{c+\epsilon}\phi_0(x,E_n)u(x,E_n)\left(\int_x^{c+\epsilon}\psi_0^{(+)}(y,E_n)u(y,E_n)dy\right)dx,$$

where $u(x,E_n)=[G_0^DW\psi_0^{(+)}](x,E_n)$, and expanding the right-hand side as in the case of $I_{0,1}(E_n)$. The result is

$$I_{0,3}(E_n)=\frac{[\psi_0^{(+)'}(\zeta_1,E_n)]^5\phi_0(\zeta_1,E_n)^3}{11340}\cdot r^9g(\varphi)\sin^4\varphi+o(r^9),$$

$$g(\varphi)=26838\cos\varphi-1386\cos3\varphi-2772\cos5\varphi+16556\sin\varphi+15383\sin3\varphi+1067\sin5\varphi.$$

Since $g(\varphi_0)=9.31172...>0$, $\psi_0^{(+)'}(\zeta_1,E_n)<0$, and $\phi_0(\zeta_1,E_n)>0$, we conclude that $I_{0,3}(E_n)<0$ for small enough r, resp., small enough $\epsilon(r,\epsilon>0)$. By Theorem 2.12, if n is odd, then there is an eigenvalue obeying (2.56) with $\lambda_0=0$ and $M=3$, i.e., $\sqrt{E(\lambda)-E_n}=-\nu_nI_{0,3}(E_n)\lambda^4+o(\lambda^4)$ as $\lambda\downarrow0$. On the other hand, if n is even, then case (vi) of Theorem 2.12 occurs and there is no eigenvalue approaching zero as $\lambda_0\downarrow0$ (or as $\lambda_0\uparrow0$).

Next we specialize Theorems 2.11 and 2.12 and compare them with some known results in the literature. Suppose that $V\neq0$, so that $E_0=0$ and $[0,\infty)$ is the only band. If (H1) holds and $\int_{-\infty}^\infty W(x)dx\leqslant0$, then $\lambda_0=0$ is a c.c.th. and (2.51) reduces to

$$\sqrt{-E(\lambda)}=-\frac{\lambda}{2}\int_{-\infty}^\infty W(x)dx-\frac{\lambda^2}{4}\int_{-\infty}^\infty W(x)|x-y|W(y)dxdy+o(\lambda^2),\quad \lambda\downarrow0, \tag{3.2}$$

which is a well-known result.[1,4,5] If $\lambda_0>0$, we can compare our results with a result in Ref. 10 (Theorem 3.2), although the operators considered there were of a somewhat different form, namely, $-d^2/dx^2+V+\lambda W$ with $V,W\in C_0^\infty(\mathbf{R})$, and the method used there was also different. For simplicity we set $V=0$ and assume that $\lambda_0>0$ is a c.c.th. for $H_\lambda=-d^2/dx^2+\lambda W$ at $E_0=0$. In accordance with the notation established above (2.78), we let $\tilde{\psi}_{\lambda_0}(x,0)$ be any bounded nontrivial solution of $H_{\lambda_0}\psi=0$. Then the limits $\lim_{x\to\pm\infty}\tilde{\psi}_{\lambda_0}(x,0)=\tilde{\psi}_\pm$ exist and are nonzero. Using (2.78) and the appropriate expansion for $E(\lambda)$ [(2.52), (2.82), or Theorem 2.14(ii)] gives

$$E(\lambda)=-\left[\frac{\int_{-\infty}^{\infty}W(x)\tilde{\psi}_{\lambda_0}(x,0)^2dx}{\tilde{\psi}_+^2+\tilde{\psi}_-^2}\right]^2(\lambda-\lambda_0)^2+o([\lambda-\lambda_0]^2),\quad \lambda\downarrow\lambda_0. \tag{3.3}$$

This result agrees with (3.8) in Ref. 10. Furthermore, (3.3) is easily extended to the case $V\neq 0$ with V and W obeying (2.1) and (H1), respectively. The problem where $V(x)=l_\pm/x^2$ for large $\pm x$, with constants $l_\pm$, was studied recently by Engländer and Pinsky.[31]

Our results extend to the case when V and W contain delta functions. To be specific, suppose that $W(x)=\pm\delta(x-x_0)$ and that V still obeys (2.1). We show how this example fits into the context of the present paper. We give details only for the case $W(x)=+\delta(x-x_0)$. Suppose that E is an eigenvalue of H_λ in a certain gap with eigenfunction $\tilde{\psi}_\lambda(x,E)$. Then we can assume that $\tilde{\psi}_\lambda(x,E)=\psi_0^{(+)}(x,E)$ for $x>x_0$ and $\tilde{\psi}_\lambda(x,E)=c\,\psi_0^{(-)}(x,E_n)$ for $x<x_0$ with some constant c. At x_0, the following matching conditions are satisfied:

$$\psi_0^{(+)}(x_0,E)=c\,\psi_0^{(-)}(x_0,E),\quad \psi_0^{(+)'}(x_0,E)-c\,\psi_0^{(-)'}(x_0,E)=\lambda\,\psi_0^{(+)}(x_0,E). \tag{3.4}$$

It is immediately clear that E can be an eigenvalue only when $\psi_0^{(\pm)}(x_0,E)\neq 0$. The equation determining the eigenvalues follows from (3.4) and (2.29):

$$\lambda\,\psi_0^{(+)}(x_0,E)\,\psi_0^{(-)}(x_0,E)=-\frac{1}{d_0(E)}. \tag{3.5}$$

This equation is equivalent to (2.33) with $\lambda_0=0$ as can be seen by inserting W in (2.44) which gives

$$I_{0,j}(E)=\begin{cases}\phi_0(x_0,E)^j\psi_0^{(+)}(x_0,E)^{j+2}, & x_0\geqslant 0,\\ (-1)^j\phi_0(x_0,E)^j\psi_0^{(-)}(x_0,E)^{j+2}, & x_0<0.\end{cases} \tag{3.6}$$

Summing the geometric series in (2.33) yields

$$\frac{\lambda\,\psi_0^{(+)}(x_0,E)^2}{1+\lambda\,\phi_0(x_0,E)\,\psi_0^{(+)}(x_0,E)}=-\frac{1}{d_0(E)},\quad x_0\geqslant 0, \tag{3.7}$$

$$\frac{\lambda\,\psi_0^{(-)}(x_0,E)^2}{1-\lambda\,\phi_0(x_0,E)\,\psi_0^{(-)}(x_0,E)}=-\frac{1}{d_0(E)},\quad x_0<0. \tag{3.8}$$

With the help of (A1) (see the Appendix) it is easy to see that (3.7) and (3.8) are equivalent to (3.5). If $\phi_0(p,E_n)\neq 0$ and λ_0 is a c.c.th., then from (2.3), (2.38), and (3.5) we conclude that $\lambda_0\psi_0^{(+)}(x_0,E_n)^2=0$, so $\lambda_0=0$ or $\psi_0^{(+)}(x_0,E_n)=0$. In the latter case we have $\psi_0^{(\pm)}(x_0,E)=O(\mu)$ as $E\to E_n$. So, in view of (2.38) this means that in the limit $\mu\to 0$ the two sides of (3.5) are incompatible regardless of what c.c.th. λ is converging to. Thus, if $\psi_0^{(+)}(x_0,E_n)=0$, then (3.5) has no solution $E(\lambda)$ such that $E(\lambda)\to E_n$ and so there are no c.c.th.'s in this case. This is in agreement with case (v) of Lemma 2.6, since then, by (3.6), $I_{0,j}(E_n)=0$ for all j. If $\psi_0^{(+)}(x_0,E_n)\neq 0$, then $\lambda_0=0$ is the only c.c.th. Further analysis then shows that n has to be odd, which is in agreement with the fact that W is non-negative, and that (2.55) applies with $M=0$ and $I_{0,0}(E_n)=\psi_0^{(+)}(x_0,E_n)^2$. The case $\phi_0(p,E_n)=0$ and $\psi_0^{(+)}(x_0,E_n)=0$, can be handled similarly, or by a shift as in Section II. The result is that $\lambda_0=0$ is the sole c.c.th. provided $\phi_0(x_0,E_n)\neq 0$ and n is odd. Then Theorem 2.14(ii) applies. In sum we see that at E_n with n odd an eigenvalue

enters the gap as $\lambda \uparrow \lambda + \epsilon$ if and only if x_0 does not coincide with a zero of the bounded solution of $H_0\psi = E_n\psi$. An analogous result holds when $W(x) = -\delta(x - x_0)$ and n is even. Without going into detail we mention that the above results can be further specialized by taking H_0 to be a Kronig-Penney Hamiltonian of the form $V(x) = \Sigma_{m=-\infty}^{\infty} \delta(x - [2m-1]a)$, with $a > 0$, so that $p = 2a$. The reader is referred to Ref. 32 for detailed information about such Hamiltonians. Theorems 2.6.4–2.6.6 (Section III.2.6) of Ref. 32 deal with the existence of eigenvalues when W represents a substitutional or interstitial delta-type impurity and those results are direct consequences of the results presented above.

IV. TWO DIMENSIONS

In two dimensions we assume that V is given by

$$V(x) = V^{(1)}(x_1) + V^{(2)}(x_2), \quad x = (x_1, x_2), \tag{4.1}$$

with $V^{(j)}$ real-valued, $V^{(j)} \in L^1_{loc}(\mathbf{R})$, and

$$V^{(j)}(x_j + p_j) = V^{(j)}(x_j), \quad p_j > 0, \quad j = 1,2. \tag{4.2}$$

Moreover, W is real-valued and satisfies

$$\int_{\mathbf{R}^2} |W(x)|^{1+\gamma} d^2x < \infty, \quad \int_{\mathbf{R}^2} |W(x)|(1+|x|)^{\gamma} d^2x < \infty, \tag{4.3}$$

for some $\gamma > 0$. Here and below the symbol $|\ |$ denotes the Euclidean norm in $\mathbf{R}^{\nu}$. The assumptions (4.1)–(4.2) guarantee that V is relatively $-\Delta$-form bounded with relative bound zero so that $H_0 = -\Delta + V$ can be defined by the method of forms. This follows from the corresponding one-dimensional result. Alternatively, H_0 is the unique self-adjoint extension of the minimal operator $H_{0,\min} = H^{(1)}_{0,\min} \otimes \mathbf{I} + \mathbf{I} \otimes H^{(2)}_{0,\min}$, where $H^{(j)}_{0,\min}$ is the minimal operator associated with the one-dimensional operator,

$$H_0^{(j)} = -\frac{d^2}{dx_j^2} + V^{(j)}(x_j),$$

on $L^2(\mathbf{R})$. Since each $H^{(j)}_{0,\min}$ is essentially self-adjoint, $H_{0,\min}$ is also essentially self-adjoint (Ref. 33, Theorem 8.33), and so $H_0 = \overline{H_{0,\min}}$. The conditions (4.3) imply[1] that $|W|^{1/2}(-\Delta + 1)^{-1}|W|^{1/2}$ is Hilbert–Schmidt. Hence $|W|^{1/2}(-\Delta + 1)^{-1/2} \in \mathscr{T}_4$, where $\mathscr{T}_p$ $(1 \leqslant p \leqslant \infty)$ denotes the usual trace ideals.[34] Let $E < \inf\sigma(H_0)$. Since $(-\Delta + 1)^{1/2}(H_0 - E)^{-1/2}$ is bounded, we conclude that $|W|^{1/2}(H_0 - E)^{-1/2} = [|W|^{1/2}(-\Delta + 1)^{-1/2}][(-\Delta + 1)^{1/2}(H_0 - E)^{-1/2}] \in \mathscr{T}_4$. Thus $|W|^{1/2}(H_0 - E)^{-1}|W|^{1/2}$ is Hilbert–Schmidt, and so W is a relatively form compact perturbation of H_0.

The spectrum of H_0 coincides with the set

$$\bigcup_{n,m=0}^{\infty} \{E_n^{(1)}(\theta_1) + E_m^{(2)}(\theta_2) : \theta_1 \in [0,2\pi], \theta_2 \in [0,2\pi]\}, \tag{4.4}$$

where $E_n^{(j)}(\theta_j)$, $n = 0,1,2,...$, are the eigenvalues, arranged in increasing order, of the spectral problem

$$\begin{gathered} H_0^{(j)}(\theta_j)\phi = \left[-\frac{d^2}{dx_j^2} + V^{(j)}(x_j)\right]\phi = E\phi, \quad j = 1,2, \\ \phi(p_j) = e^{i\theta_j}\phi(0), \quad \phi'(p_j) = e^{i\theta_j}\phi'(0), \quad \theta_j \in [0,2\pi]. \end{gathered} \tag{4.5}$$

The spectrum of $H_0^{(j)}$ consists of the bands $[E_0^{(j)}(0), E_0^{(j)}(\pi)]$, $[E_1^{(j)}(\pi), E_1^{(j)}(0)]$, etc. The sums $E_n^{(1)}(\theta_1) + E_m^{(2)}(\theta_2)$ $(n,m = 0,1,...)$ are the eigenvalues of the closure of the operator $H_0^{(1)}(\theta_1) \otimes \mathbf{I} + \mathbf{I} \otimes H_0^{(2)}(\theta_2)$, and the corresponding normalized eigenfunctions are products of the form

$\psi_n^{(1)}(x_1,\theta_1)\psi_m^{(2)}(x_2,\theta_2)$, where $\psi_n^{(j)}(x_j,\theta_j)$ is the normalized eigenfunction for the eigenvalue problem (4.5). We extend these eigenfunctions to all of $\mathbf{R}$ by using the boundary conditions in (4.5). We also note that

$$\psi_n^{(j)}(x,\theta_j)=\overline{\psi_n^{(j)}(x,2\pi-\theta_j)}, \tag{4.6}$$

$$E_n^{(j)}(2\pi-\theta_j)=E_n^{(j)}(\theta_j), \tag{4.7}$$

and that for n even (resp., n odd) the functions $E_n^{(j)}(\theta_j)$ are strictly monotone increasing (resp., decreasing) on $[0,\pi]$ (Ref. 20, Theorem XIII.89). Moreover, near the edges of a gap in $\sigma(H_0^{(j)})$ we have [cf. (2.8)]

$$E_n^{(j)}(\theta_j)=E_n^{(j)}(0)+c_n^{(j)}\theta_j^2+O(\theta_j^4), \quad \theta_j\to 0, \tag{4.8}$$

$$E_n^{(j)}(\theta_j)=E_n^{(j)}(\pi)+d_n^{(j)}(\theta_j-\pi)^2+O([\theta_j-\pi]^4), \quad \theta_j\to\pi, \tag{4.9}$$

where $(-1)^n c_n^{(j)}>0$ and $(-1)^n d_n^{(j)}<0$. Without loss of generality we may assume that

$$E_0^{(1)}(0)=E_0^{(2)}(0)=0, \tag{4.10}$$

so that the first spectral band is given by $[0, E_0^{(1)}(\pi)+E_0^{(2)}(\pi)]$. We first consider absorption of an eigenvalue at the bottom of the spectrum, that is, at energy zero. For the Birman–Schwinger kernel we use $K_{0,E}$ as defined in (1.3), but to simplify the notation we omit the subscript 0 and just write K_E. For any interval Δ, let P^{Δ} denote the corresponding spectral projection for H_0. We first decompose K_E as follows: Pick $\delta>0$ and write

$$K_E=K_E^{[0,\delta]}+K_E^{(\delta,\infty)}, \tag{4.11}$$

where

$$K_E^{[0,\delta]}=W^{1/2}P^{[0,\delta]}(H_0-E)^{-1}P^{[0,\delta]}|W|^{1/2}, \tag{4.12}$$

$$K_E^{(\delta,\infty)}=W^{1/2}P^{(\delta,\infty)}(H_0-E)^{-1}P^{(\delta,\infty)}|W|^{1/2}. \tag{4.13}$$

Lemma 4.1: If $E<0$ and $\delta>0$, then $K_E^{(\delta,\infty)}$ is Hilbert–Schmidt and has an analytic continuation in the Hilbert–Schmidt norm to a complex neighborhood of $E=0$.

Proof: For $E<0$ write $K_E^{(\delta,\infty)}=ABC$ with $A=W^{1/2}(H_0+1)^{-1/2}$, $B=P^{(\delta,\infty)}(H_0+1)^{1/2}(H_0-E)^{-1}(H_0+1)^{1/2}P^{(\delta,\infty)}$, and $C=(H_0+1)^{-1/2}|W|^{1/2}$. Then A and C both lie in $\mathscr{T}_4$ and B is norm-analytic near $E=0$. Hence the product ABC is analytic in the Hilbert-Schmidt norm for E near 0. ∎

In order to analyze the operator $K_E^{[0,\delta]}$ in (4.12) we use the representation of Ref. 20 (see Theorem XIII.98) which allows us to express the integral kernel of $K_E^{[0,\delta]}$ as

$$K_E^{[0,\delta]}(x,y)=W(x)^{1/2}\int_{\mathscr{D}_\delta}\left[\frac{\psi_0^{(1)}(x_1,\theta_1)\psi_0^{(2)}(x_2,\theta_2)\overline{\psi_0^{(1)}(y_1,\theta_1)\psi_0^{(2)}(y_2,\theta_2)}}{E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E}\,\frac{d^2\theta}{(2\pi)^2}\right]|W(y)|^{1/2}, \tag{4.14}$$

where $\theta=(\theta_1,\theta_2)$ and

$$\mathscr{D}_\delta=\{\theta\in[0,2\pi]^2: E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant\delta\}. \tag{4.15}$$

Let

$$S(x_1,x_2,y_1,y_2;\theta)=\psi_0^{(1)}(x_1,\theta_1)\psi_0^{(2)}(x_2,\theta_2)\overline{\psi_0^{(1)}(y_1,\theta_1)\psi_0^{(2)}(y_2,\theta_2)}, \tag{4.16}$$

$$G(x,y;\theta)=S(x_1,x_2,y_1,y_2;\theta)+S(y_1,x_2,x_1,y_2;\theta)+S(x_1,y_2,y_1,x_2;\theta)+S(y_1,y_2,x_1,x_2;\theta), \tag{4.17}$$

$$H(x,y;\theta)=G(x,y;\theta)-G(x,y;0). \tag{4.18}$$

Note that because $\psi_0^{(j)}(x_j,0)$ is real, we have

$$G(x,y;0)=4\psi_0^{(1)}(x_1,0)\psi_0^{(2)}(x_2,0)\psi_0^{(1)}(y_1,0)\psi_0^{(2)}(y_2,0). \tag{4.19}$$

Moreover, $G(x,y;\theta)$ can be written as

$$\begin{aligned}G(x,y;\theta)=4[&\mathrm{Re}(\psi_0^{(1)}(x_1,\theta_1))\mathrm{Re}(\psi_0^{(1)}(y_1,\theta_1))+\mathrm{Im}(\psi_0^{(1)}(x_1,\theta_1))\mathrm{Im}(\psi_0^{(1)}(y_1,\theta_1))]\\ &\times[\mathrm{Re}(\psi_0^{(2)}(x_2,\theta_2))\mathrm{Re}(\psi_0^{(2)}(y_2,\theta_2))+\mathrm{Im}(\psi_0^{(2)}(x_2,\theta_2))\mathrm{Im}(\psi_0^{(2)}(y_2,\theta_2))].\end{aligned} \tag{4.20}$$

Lemma 4.2: For δ small enough, $K_E^{[0,\delta]}$ can be decomposed as

$$K_E^{[0,\delta]}=b(E;\delta)L+R_E^{[0,\delta]}, \tag{4.21}$$

where

$$b(E;\delta)=\frac{1}{\pi^2}\int_{\mathscr{E}_\delta}\frac{d^2\theta}{E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E},$$

$$\mathscr{E}_\delta=\{\theta\in[0,\pi]^2:E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant\delta\},$$

$$L\cdot=\omega(\tilde{\omega},\cdot),$$

$$\omega(x)=W(x)^{1/2}\psi_0^{(1)}(x_1,0)\psi_0^{(2)}(x_2,0),\quad \tilde{\omega}=\omega\ \mathrm{sgn}\ W, \tag{4.22}$$

and $R_E^{[0,\delta]}$ is the integral operator with kernel

$$R_E^{[0,\delta]}(x,y)=W(x)^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{H(x,y;\theta)}{E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E}\frac{d^2\theta}{(2\pi)^2}\right]|W(y)|^{1/2}. \tag{4.23}$$

Proof: We first notice that for δ sufficiently small, the set $\mathscr{D}_\delta$ in (4.15) consists of four disjoint pieces which are located near the corners of the square $[0,2\pi]^2$. The set $\mathscr{E}_\delta$ is that portion of $\mathscr{D}_\delta$ which contains the origin. By using (4.6), (4.7), (4.16), and (4.17), we can reduce the integral over $\mathscr{D}_\delta$ to one over $\mathscr{E}_\delta$, so that (4.14) becomes

$$K_E^{[0,\delta]}(x,y)=W(x)^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{G(x,y;\theta)}{E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E}\frac{d^2\theta}{(2\pi)^2}\right]|W(y)|^{1/2}. \tag{4.24}$$

Inserting $G(x,y;\theta)=G(x,y;0)+H(x,y;\theta)$ in (4.24) and using (4.19) yields (4.21). Note that (4.3) implies $\omega\in L^2(\mathbf{R}^2)$ so that L is a rank-one operator on $L^2(\mathbf{R}^2)$. ■

Theorem 4.3: For $-\delta<E<0$ with $\delta>0$ sufficiently small we have the convergent expansion

$$\begin{aligned}b(E;\delta)=&-\frac{1}{2\pi^2}\left(\sum_{m=0}^{\infty}a_{2m}E^m\right)\ln(-E)+\frac{1}{2\pi^2}\left(\sum_{m=0}^{\infty}a_{2m}E^m\right)\ln(\delta-E)\\ &+\frac{1}{2\pi^2}\sum_{k=1}^{\infty}\delta^{-k}\left(\sum_{m=k}^{\infty}\frac{a_{2m}\delta^{m+1}}{m-k+1}\right)E^{k-1},\end{aligned} \tag{4.25}$$

with suitable coefficients a_{2m} (to be defined below).

Proof: Define new variables v_1 and v_2 $(v_1,v_2\geqslant 0)$ by

$$v_1=\sqrt{E_0^{(1)}(\theta_1)},\quad v_2=\sqrt{E_0^{(2)}(\theta_2)}. \tag{4.26}$$

For sufficiently small v_1 and v_2 the inverse functions

$$\theta_1 = h_1(v_1), \quad \theta_2 = h_2(v_2), \tag{4.27}$$

exist and, by (4.8) and (4.10), are odd, analytic functions of v_1 and v_2, respectively. By means of the polar coordinates

$$r = \sqrt{v_1^2 + v_2^2}, \quad v_1 = r\cos\varphi, \quad v_2 = r\sin\varphi,$$

we can express $b(E;\delta)$ as

$$b(E;\delta) = \frac{1}{\pi^2}\int_0^{\sqrt{\delta}}\int_0^{\pi/2} \frac{h_1'(r\cos\varphi)h_2'(r\sin\varphi)}{r^2 - E}\, r\,dr\,d\varphi = \frac{1}{\pi^2}\int_0^{\sqrt{\delta}} \frac{u(r)r}{r^2 - E}\, dr, \tag{4.28}$$

$$u(r) = \int_0^{\pi/2} h_1'(r\cos\varphi)h_2'(r\sin\varphi)\,d\varphi. \tag{4.29}$$

The functions $h_1'(r\cos\varphi)$ and $h_2'(r\sin\varphi)$ are even functions of r and hence $u(r)$ can be expanded as

$$u(r) = \sum_{m=0}^{\infty} a_{2m} r^{2m}, \tag{4.30}$$

for $r < r_0$, where $r_0 > 0$ is less than the minimum of the radii of convergence for h_1 and h_2. Inserting (4.30) in (4.28) gives

$$b(E;\delta) = \frac{1}{\pi^2}\sum_{m=0}^{\infty} a_{2m} I_m(E;\delta),$$

$$I_m(E;\delta) = \int_0^{\sqrt{\delta}} \frac{r^{2m+1}}{r^2 - E}\, dr, \quad m = 0,1,2,\ldots . \tag{4.31}$$

Then

$$I_0(E;\delta) = -\tfrac{1}{2}\ln(-E) + \tfrac{1}{2}\ln(\delta - E),$$

and $I_m(E;\delta)$ is given recursively by

$$I_m(E;\delta) - E I_{m-1}(E;\delta) = \frac{\delta^m}{2m}, \quad m = 1,2,\ldots .$$

Thus

$$I_m(E;\delta) = -\frac{1}{2}E^m \ln(-E) + \frac{1}{2}E^m \ln(\delta - E) + \frac{1}{2}\sum_{k=1}^{m} \frac{\delta^{m-k+1}}{m-k+1} E^{k-1}, \tag{4.32}$$

as can be verified by induction on m or, alternately, by differentiation with respect to δ. Inserting (4.32) in (4.31) yields (4.25). It is easy to see that choosing $\delta < r_0^2$ ensures convergence of the various series on the right-hand side of (4.25) for $-\delta < E < 0$. ■

From (4.25) we infer that

$$b(E;\delta) = -\frac{a_0}{2\pi^2}\ln(-E) + c(\delta) - \frac{a_2}{2\pi^2} E \ln(-E) + O(E),$$

where

$$c(\delta) = \frac{1}{2\pi^2}\left[a_0 \ln\delta + \sum_{m=1}^{\infty} \frac{a_{2m}\delta^m}{m}\right].$$

By using (4.8), (4.26), and (4.27), we find $h_j'(0)=1/\sqrt{c_0^{(j)}}$ $(j=1,2)$ and thus, by (4.29) and (4.30),

$$a_0=\frac{\pi}{2\sqrt{c_0^{(1)}c_0^{(2)}}}. \tag{4.33}$$

We let

$$R_{E;\delta}=R_E^{[0,\delta]}+K_E^{(\delta,\infty)}=K_E-b(E;\delta)L, \tag{4.34}$$

and define the operator M_E by

$$K_E=-\frac{a_0}{2\pi^2}\ln(-E)L+M_E, \tag{4.35}$$

so that

$$M_E=\left[c(\delta)-\frac{a_2}{2\pi^2}E\ln(-E)+O(E)\right]L+R_{E;\delta}. \tag{4.36}$$

We now study the behavior of M_E as $E\to 0$. Note that, by (4.35), M_E is independent of δ. Also, note that the representation (4.35) is of the form (1.5) with $N=1$, $\lambda_0=0$, $d_{0;1}(E)=-a_0/(2\pi^2)$, $\omega_{0,E;1}=\omega$ as given by (4.22), and $R_{0,E}=M_E$. We remark that for two-dimensional perturbed periodic Schrödinger operators it has been noted earlier[35] that the Birman-Schwinger kernel diverges as $E\uparrow 0$, and this fact has been discussed in connection with the criticality of such operators.

Lemma 4.4: For every $\alpha\in(0,1]$ and $\theta_j\geqslant 0$ sufficiently small there is a constant C_j such that

$$|\psi_0^{(j)}(x_j,\theta_j)-\psi_0^{(j)}(x_j,0)|\leqslant C_j\,(1+|x_j|)^{\alpha}\theta_j^{\alpha},\quad j=1,2. \tag{4.37}$$

Proof: Let $\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))$ denote the solutions defined in (2.7) associated with the potential $V^{(j)}$ for $E=E_0^{(j)}(\theta_j)$. Thus

$$\psi_0^{(j)}(x_j,\theta_j)=\frac{\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))}{N_0^{(j,+)}(\theta_j)}=\frac{e^{i\theta_j x_j/p_j}\xi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))}{N_0^{(j,+)}(\theta_j)}, \tag{4.38}$$

where

$$N_0^{(j,+)}(\theta_j)=\left(\int_0^{p_j}|\xi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))|^2dx_j\right)^{1/2}. \tag{4.39}$$

Recall that the functions $\psi_0^{(j)}(x_j,\theta_j)$ are normalized over one period. In the following estimates $\tilde{C}_j$ denotes a suitable constant. Similarly to (i) and (iv) of Lemma 2.3 we have

$$|\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))|\leqslant\tilde{C}_j$$

and thus

$$|\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))-\psi_0^{(j,+)}(x_j,0)|\leqslant 2\tilde{C}_j.$$

Moreover,

$$|\dot{\psi}_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))|\leqslant\tilde{C}_j(1+|x_j|),$$

where now the dot denotes differentiation with respect to θ_j. Therefore,

$$|\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))-\psi_0^{(j,+)}(x_j,0)|\leqslant\tilde{C}_j\theta_j(1+|x_j|)$$

and hence, by interpolation,

$$|\psi_0^{(j,+)}(x_j,E_0^{(j)}(\theta_j))-\psi_0^{(j,+)}(x_j,0)|\leqslant 2^{1-\alpha}\tilde{C}_j\theta_j^{\alpha}(1+|x_j|)^{\alpha}$$

for any $\alpha\in(0,1]$. Furthermore,

$$|N_0^{(j,+)}(\theta_j)-N_0^{(j,+)}(0)|=O(\theta_j)$$

by (4.39). Now (4.37) follows from (4.38). ■

Theorem 4.5: There is a Hilbert–Schmidt operator M_0 and $\beta>0$ such that $\|M_E-M_0\|_{H.S.}=O(|E|^{\beta})$ as $E\uparrow 0$.

Proof: In view of Lemma 4.1, (4.34), and (4.36), it suffices to show that $R_E^{[0,\delta]}$ converges in Hilbert–Schmidt norm to a limit $R_0^{[0,\delta]}$ as $E\uparrow 0$, and that $\|R_E^{[0,\delta]}-R_0^{[0,\delta]}\|_{H.S.}=O(|E|^{\beta})$. We write

$$H(x,y;\theta)=[S(x_1,x_2,y_1,y_2;\theta)-\psi_0^{(1)}(x_1,0)\psi_0^{(2)}(x_2,0)\psi_0^{(1)}(y_1,0)\psi_0^{(2)}(y_2,0)]+\cdots, \quad (4.40)$$

where the right-hand side consists of four obvious terms coming from subtracting (4.19) from (4.17). Each of these terms can again be written as a sum of four terms, namely

$$\begin{aligned}&S(x_1,x_2,y_1,y_2;\theta)-\psi_0^{(1)}(x_1,0)\psi_0^{(2)}(x_2,0)\psi_0^{(1)}(y_1,0)\psi_0^{(2)}(y_2,0)\\&\quad=[\psi_0^{(1)}(x_1,\theta_1)-\psi_0^{(1)}(x_1,0)]\psi_0^{(2)}(x_2,\theta_2)\overline{\psi_0^{(1)}(y_1,\theta_1)\psi_0^{(2)}(y_2,\theta_2)}+\cdots, \quad (4.41)\end{aligned}$$

where it is clear how the right-hand side is obtained. Applying Lemma 4.4 to each term on the right-hand side of (4.41) and then using the result in (4.40) gives

$$|H(x,y;\theta)|\leqslant C(1+|x|)^{\alpha}(1+|y|)^{\alpha}|\theta|^{\alpha}, \quad \alpha\in(0,1]. \quad (4.42)$$

Substituting (4.42) in (4.23) leads to θ-integrals that are bounded independently of E, since by using polar coordinates (ignoring irrelevant factors) we have

$$\int_{\mathscr{C}_\delta}\frac{|\theta|^{\alpha}}{E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E}\,d^2\theta\leqslant C\int_0^{\sqrt{\delta}}\frac{r^{\alpha+1}}{r^2-E}\,dr\leqslant C\,\frac{\delta^{\alpha/2}}{\alpha}.$$

Consequently,

$$|R_E^{[0,\delta]}(x,y)|\leqslant C|W(x)|^{1/2}(1+|x|)^{\alpha}(1+|y|)^{\alpha}|W(y)|^{1/2}. \quad (4.43)$$

So if we take α so that $0<\alpha\leqslant\min\{1,\gamma/2\}$, where γ is the constant in (4.3), then (4.43) provides us with the required uniform bound on $|R_E^{[0,\delta]}(x,y)|$ that allows us to conclude, as in the proof of Lemma 2.4(ii), that $R_E^{[0,\delta]}\to R_0^{[0,\delta]}$ in Hilbert–Schmidt norm as $E\uparrow 0$. Taking the derivative of $R_E^{[0,\delta]}$ with respect to E and using the estimate

$$\int_{\mathscr{C}_\delta}\frac{|\theta|^{\alpha}}{[E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E]^2}\,d^2\theta\leqslant C\int_0^{\sqrt{\delta}}\frac{r^{\alpha+1}}{(r^2-E)^2}\,dr\leqslant C|E|^{-1+\alpha/2}\int_0^{\infty}\frac{u^{\alpha+1}}{(u^2+1)^2}\,du,$$

we get

$$\left|\frac{dR_E^{[0,\delta]}}{dE}(x,y)\right|\leqslant C|E|^{-1+\alpha/2}|W(x)|^{1/2}(1+|x|)^{\alpha}(1+|y|)^{\alpha}|W(y)|^{1/2}. \quad (4.44)$$

Thus an integration gives $\|R_E^{[0,\delta]}-R_0^{[0,\delta]}\|_{H.S.}=O(|E|^{\alpha/2})$, and so we can choose any $\beta\leqslant\min\{1/2,\gamma/4\}$ to reach the conclusion of the theorem. ■

According to (1.8) and (4.35) the equation for the bound state reads

$$\lambda(\tilde{\omega},[1+\lambda M_E]^{-1}\omega)=\frac{2\pi^2}{a_0\ln(-E)}, \quad (4.45)$$

and we are looking for solutions obeying $E(\lambda)\uparrow 0$ as $\lambda\downarrow 0$. Since for $|E|$ small, the right-hand side of (4.45) is negative, such solutions can exist only if

$$(\tilde{\omega},\omega)=\int_{\mathbf{R}^2} W(x)\psi_0^{(1)}(x_1,0)^2\psi_0^{(2)}(x_2,0)^2 d^2x\leqslant 0.$$

The next lemma is the analog of Lemma 2.10(ii) in two dimensions.

Lemma 4.6: If $(\tilde{\omega},\omega)=0$ and $W\neq 0$, then $(\tilde{\omega},M_0\omega)>0$.

Proof: $(\tilde{\omega},\omega)=0$ is equivalent to $L\omega=0$, and so, by (4.35) and Theorem 4.5,

$$\lim_{E\uparrow 0}(\tilde{\omega},K_E\omega)=(\tilde{\omega},M_0\omega). \tag{4.46}$$

Let

$$\hat{K}_E=(\operatorname{sgn} W)K_E=|W|^{1/2}(H_0-E)^{-1}|W|^{1/2},\quad E<0.$$

Then $\hat{K}_E$ is an increasing family of positive self-adjoint operators. So

$$(\tilde{\omega},K_E\omega)=(\omega,\hat{K}_E\omega)\geqslant 0,$$

and hence, by (4.46), $(\tilde{\omega},M_0\omega)\geqslant 0$. Furthermore, by (4.46), $(\tilde{\omega},M_0\omega)=0$ implies $(\omega,\hat{K}_E\omega)=0$ for all $E<0$. So $(H_0-E)^{-1/2}|W|^{1/2}\omega=0$ for all $E<0$; in particular,

$$(H_0+1)^{-1/2}|W|^{1/2}\omega=0, \tag{4.47}$$

and thus $|W|^{1/2}\omega=0$. Since $|W|^{1/2}\omega$ is generally not in $L^2(\mathbf{R}^2)$, the last conclusion requires an explanation. If we consider the scale of Hilbert spaces $\mathscr{H}_{+1}\subset L^2(\mathbf{R}^2)\subset\mathscr{H}_{-1}$, where $\mathscr{H}_{+1}=D([H_0+1]^{1/2})$ with the norm $\|f\|_{+1}=\|(H_0+1)^{1/2}f\|$ and $\mathscr{H}_{-1}$ is the completion of $L^2(\mathbf{R})$ in the norm $\|f\|_{-1}=\|(H_0+1)^{-1/2}f\|$, then $|W|^{1/2}\omega\in\mathscr{H}_{-1}$. Moreover, $(H_0+1)^{-1/2}$ maps $\mathscr{H}_{-1}$ to $L^2(\mathbf{R}^2)$ and has a trivial kernel. Since H_0+1 and $-\Delta+1$ have the same form domain, or, equivalently, by multiplying (4.47) from the left by $(-\Delta+1)^{-1/2}(H_0+1)^{1/2}$, we conclude that $(-\Delta+1)^{-1/2}|W|^{1/2}\omega=0$. Then, by using a Fourier transform, we obtain $W(x)=0$, a.e. Since this is not the case, $(\tilde{\omega},M_0\omega)>0$. ■

In the next theorem, $E_0(\lambda)$ denotes the lowest eigenvalue (ground state) of H_λ. Moreover, a_0 is the constant defined in (4.33) and β is as in Theorem 4.5.

Theorem 4.7: Assume (4.1)–(4.3). Then there is an eigenvalue $E_0(\lambda)$ of H_λ satisfying the following:

(i) $E_0(\lambda)$ is simple and the only eigenvalue converging to zero as $\lambda\downarrow 0$.

(ii) If $(\tilde{\omega},\omega)<0$, then

$$E_0(\lambda)=-e^{-h_1(\lambda)}[1+O(e^{-\rho_1/\lambda})],\quad \lambda\downarrow 0, \tag{4.48}$$

where $\rho_1=2\pi^2\gamma_{-1}\beta a_0^{-1}$ and

$$h_1(\lambda)=\frac{2\pi^2}{a_0}\left[\frac{\gamma_{-1}}{\lambda}+\gamma_0+\sum_{n=1}^{\infty}\gamma_n\lambda^n\right],$$

$$\gamma_{-1}=-\frac{1}{(\tilde{\omega},\omega)},\quad \gamma_0=-\frac{(\tilde{\omega},M_0\omega)}{(\tilde{\omega},\omega)^2}, \tag{4.49}$$

with the series in (4.49) being convergent for λ small enough.

(iii) If $(\tilde{\omega},\omega)=0$, then

$$E_0(\lambda)=-e^{-h_2(\lambda)}[1+O(\lambda^{-2}e^{-\rho_2/\lambda^2})],\quad \lambda\downarrow 0, \tag{4.50}$$

where ρ_2 is any constant such that $0<\rho_2<2\pi^2\nu_{-2}\beta a_0^{-1}$, and

$$h_2(\lambda)=\frac{2\pi^2}{a_0}\left[\frac{\nu_{-2}}{\lambda^2}+\frac{\nu_{-1}}{\lambda}+\nu_0+\sum_{n=1}^{\infty}\nu_n\lambda^n\right],$$

$$\nu_{-2}=\frac{1}{(\tilde{\omega},M_0\omega)},\quad \nu_{-1}=\frac{(\tilde{\omega},M_0^2\omega)}{(\tilde{\omega},M_0\omega)^2},\quad \nu_0=\frac{(\tilde{\omega},M_0^2\omega)^2}{(\tilde{\omega},M_0\omega)^3}-\frac{(\tilde{\omega},M_0^3\omega)}{(\tilde{\omega},M_0\omega)^2}. \tag{4.51}$$

(iv) If $(\tilde{\omega},\omega)>0$, then λ_0 is not a c.c.th.

We remark that in the case $V\neq 0$ the coefficients γ_{-1}, γ_0, ν_{-2}, ν_{-1}, and ν_0 reduce to those given by Holden.[2]

Proof: (i) The existence and uniqueness of $E_0(\lambda)$ are obvious if we recall the approach based on eigenvalue perturbation theory.[1,3] However, for later use it is useful to have an independent argument based solely on (4.45). First, existence is obvious from (4.45), Theorem 4.5, and the intermediate value theorem. Uniqueness follows by taking the derivative with respect to E and showing that $[d/dE](\tilde{\omega},[1+\lambda M_E]^{-1}\omega)=O(|E|)^{-1+\alpha/2}$ as $E\uparrow 0$, uniformly in λ, with $0<\alpha\leqslant\min\{1,\gamma/2\}$. This follows from (4.44), which remains valid if we replace $R_E^{[0,\delta]}$ by M_E, and estimates similar to those used in the proof of Lemma 2.5. The restriction on α comes from the fact that the integral $\int_{\mathbf{R}^2}(1+|x|)^{2\alpha}|W(x)|d^2x$ must be finite. Further details are omitted. Since the derivative of the right-hand side of (4.45) is equal to $2\pi^2 a_0^{-1}E^{-1}[\ln(-E)]^{-2}$, and thus dominates for small E, uniqueness follows.

To prove (ii) note that, by (4.45),

$$\ln(-E)=-\frac{2\pi^2}{a_0}\left[\frac{\gamma_{-1}}{\lambda}+\gamma_0+o(1)\right], \tag{4.52}$$

and so, by Theorem 4.5,

$$\|(1+\lambda M_E)^{-1}-(1+\lambda M_0)^{-1}\|_{H.S.}\leqslant C_1\lambda|E|^{\beta}\leqslant C_2\lambda e^{-\rho_1/\lambda}, \tag{4.53}$$

where $\rho_1=2\pi^2\gamma_{-1}\beta a_0^{-1}$. Hence we can write (4.45) as

$$\lambda(\tilde{\omega},(1+\lambda M_0)^{-1}\omega)+O(\lambda^2 e^{-\rho_1/\lambda})=\frac{2\pi^2}{a_0\ln(-E)}. \tag{4.54}$$

Let

$$h_1(\lambda)=-\frac{2\pi^2}{a_0}\frac{1}{\lambda(\tilde{\omega},[1+\lambda M_0]^{-1}\omega)}. \tag{4.55}$$

Then (4.48) and (4.49) follow from (4.54) and (4.55) by expanding $(1+\lambda M_0)^{-1}$.

The proof of (iii) is similar. Instead of (4.52) we now have

$$\ln(-E)=-\frac{2\pi^2}{a_0}\left[\frac{\nu_{-2}}{\lambda^2}+\frac{\nu_{-1}}{\lambda}+\nu_0+o(1)\right],$$

and thus

$$\|(1+\lambda M_E)^{-1}-(1+\lambda M_0)^{-1}\|\leqslant C_1\lambda|E|^{\beta}\leqslant C_2\lambda e^{-\rho_2/\lambda^2},$$

with $0<\rho_2<2\pi^2\nu_{-2}\beta a_0^{-1}$, taking into account the possibility that ν_{-1} may be negative. Since $(\tilde{\omega},\omega)=0$, the function $h_1(\lambda)$ can be rewritten as

$$h_2(\lambda)=\frac{2\pi^2}{a_0}\frac{1}{\lambda^2(\tilde{\omega},M_0[1+\lambda M_0]^{-1}\omega)}.$$

Expanding the denominator leads to (4.51). In place of (4.54) we have from (4.45),

$$-\lambda^2(\tilde{\omega},M_0[1+\lambda M_0]^{-1}\omega)+O(\lambda^2 e^{-\rho_2/\lambda^2})=\frac{2\pi^2}{a_0\ln(-E)}. \tag{4.56}$$

So, by solving (4.56) for E and putting $E=E_0(\lambda)$, we obtain (4.50).

Conclusion (iv) is obvious from (4.45) because the two sides have opposite signs when λ and E are small. ■

If we assume

$$\int_{\mathbf{R}^2} e^{c|x|}|W(x)|d^2x<\infty, \quad c>0, \tag{4.57}$$

then $E_0(\lambda)$ has a convergent expansion in terms of quantities containing factors of the form $e^{-d/\lambda}$ $(d>0)$ and powers of λ. First, note that from (2.3), (2.4), and (2.7), it follows that the normalization constants $N_0^{(j,+)}(\theta_j)$ $(j=1,2)$ in (4.39), originally defined for real θ_j, can be analytically continued to a complex neighborhood of zero and that $N_0^{(j,+)}(\theta_j)$ are even functions of θ_j. Consequently, the real part of $\psi_0^{(j)}(x_j,\theta_j)$ is an even function of θ_j and the imaginary part is an odd function of θ_j. Thus there is a $\tilde{\theta}>0$ such that for $|\theta_j|<\tilde{\theta}$ $(j=1,2)$ we have the convergent expansions

$$\psi_0^{(j)}(x_j,\theta_j)=\alpha_0^{(j)}(x_j)+i\beta_1^{(j)}(x_j)\theta_j+\alpha_2^{(j)}(x_j)\theta_j^2+i\beta_3^{(j)}(x_j)\theta_j^3+\ldots, \tag{4.58}$$

where

$$\alpha_0^{(j)}(x_j)=\frac{\psi_0^{(j,+)}(x_j,0)}{N_0^{(j,+)}(0)}, \quad \beta_1^{(j)}(x_j)=\frac{\phi_0^{(j)}(x_j,0)}{\phi_0^{(j)}(p_j,0)N_0^{(j,+)}(0)},$$

and $\alpha_n^{(j)}(x_j)$ and $\beta_n^{(j)}(x_j)$ are real. Also, recall that $\phi_0^{(j)}(p_j,0)\neq 0$, since $E_0^{(j)}(0)=0=\inf\sigma(H_0^{(j)})$. Equation (4.58) follows from (4.38), (2.3), and (2.4) with $q(E)=q(E_0^{(j)}(\theta_j))=\theta_j$, by expanding $\psi_0^{(j)}(x_j,\theta_j)$ in powers of θ_j. In view of (4.20) and (4.58) the function $H(x,y;\theta)$ in (4.18) is even in θ_1 and θ_2 separately, and hence can be expanded as

$$H(x,y;\theta)=\sum_{n,m=0}^{\infty} A_{nm}(x,y)\theta_1^{2n}\theta_2^{2m}, \tag{4.59}$$

where $A_{00}(x,y)=0$ because $H(x,y;0)=0$. From (4.16), (4.18), and (4.38) we obtain the inequality

$$|H(x,y;\theta)|\leq Ce^{|\theta_1||x_1|/p_1+|\theta_2||x_2|/p_2+|\theta_1||y_1|/p_1+|\theta_2||y_2|/p_2}. \tag{4.60}$$

Let $\tilde{p}=(p_1^{-2}+p_2^{-2})^{1/2}$ and assume $|\theta_j|<\tilde{\theta}(j=1,2)$. Using the Schwarz inequality gives

$$|H(x,y;\theta)|\leq Ce^{\tilde{\theta}\,\tilde{p}(|x|+|y|)}.$$

Thus, by Cauchy's inequality, we obtain

$$|A_{nm}(x,y)|\leq C\,\frac{e^{\tilde{\theta}\,\tilde{p}(|x|+|y|)}}{\tilde{\theta}^{2(n+m)}}. \tag{4.61}$$

Inserting (4.59) in (4.23) and using polar coordinates according to (4.26) and (4.27) gives rise to integrals of the form

$$\rho_{nm}(r)=\int_0^{\pi/2} h_1'(r\,\cos\,\varphi)h_2'(r\,\sin\,\varphi)\cos^{2n}\,\varphi\,\sin^{2m}\,\varphi d\varphi.$$

The functions $\rho_{nm}(r)$ can be expanded as

$$\rho_{nm}(r)=\sum_{s=0}^{\infty} c_{nm;2s}\,r^{2s},$$

where the series converges for $r<r_0$ with r_0 being the same as in (4.30). Again, by Cauchy's inequality, there is a constant $M_{r_0}>0$ (independent of n and m) such that

$$|c_{nm;2s}| \leqslant \frac{M_{r_0}}{r_0^{2s}}. \tag{4.62}$$

Assuming δ to be so small that $\delta < \min\{\tilde{\theta}^2, r_0^2\}$ and $\mathscr{E}_\delta \subset \{\theta \in [0,\pi]^2 : |\theta_j| < \tilde{\theta}, j=1,2\}$, for $-\delta < E < 0$ we have

$$R_E^{[0,\delta]}(x,y) = \frac{W(x)^{1/2}}{4\pi^2}\left[\sum_{n,m,s=0}^{\infty} c_{nm;2s} A_{nm}(x,y) I_{n+m+s}(E;\delta)\right]|W(y)|^{1/2}, \tag{4.63}$$

where $I_{n+m+s}(E;\delta)$ is given by (4.32). Inserting (4.32) in (4.63) gives

$$R_E^{[0,\delta]}(x,y) = \left[\sum_{n=0}^{\infty} D_n^{(1)}(x,y) E^n\right] E \ln(-E) + \sum_{n=0}^{\infty} D_n^{(2)}(x,y) E^n, \tag{4.64}$$

where the coefficients $D_n^{(1)}(x,y)$ and $D_n^{(2)}(x,y)$ obey

$$|D_n^{(j)}(x,y)| \leqslant C|W(x)|^{1/2} e^{\tilde{\theta}\tilde{p}(|x|+|y|)}|W(y)|^{1/2}, \quad j=1,2.$$

Hence the kernels $D_n^{(j)}(x,y)$ are Hilbert–Schmidt if we assume that $\tilde{\theta} \leqslant c/(2\tilde{p})$, where c is the constant in (4.57). By (4.36), and since $K_E^{(\delta,\infty)}$ is analytic at $E=0$, the expansion (4.64) carries over to $R_{E;\delta}$ and to M_E. As a result, (4.45) can be written in the form

$$\ln(-E) = \begin{cases} -\dfrac{2\pi^2}{a_0}\left[\dfrac{\gamma_{-1}}{\lambda} + \gamma_0 + \displaystyle\sum_{\substack{n,m,s=0 \\ n+m+s \geqslant 1}} e_{nms} E^n [E \ln(-E)]^m \lambda^s\right], & (\tilde{\omega},\omega) < 0, \\ -\dfrac{2\pi^2}{a_0}\left[\dfrac{\nu_2}{\lambda^2} + \dfrac{\nu_{-1}}{\lambda} + \nu_0 + \displaystyle\sum_{\substack{n,m=0 \\ s \geqslant -2}} f_{nms} E^n [E \ln(-E)]^m \lambda^s\right], & (\tilde{\omega},\omega) = 0, \end{cases} \tag{4.65}$$

where $f_{00-2} = f_{00-1} = f_{000} = 0$. Let

$$v = \lambda^{-1} e^{-2\pi^2 \gamma_{-1}/(a_0\lambda)}, \quad w = \lambda^{-4} e^{-(2\pi^2/a_0)[\nu_{-2}/\lambda^2 + \nu_{-1}/\lambda]}. \tag{4.66}$$

Theorem 4.8: Assume (4.57). Then the following holds.

(i) If $(\tilde{\omega},\omega) < 0$, then for λ sufficiently small, $E_0(\lambda)$ has the uniformly and absolutely convergent expansion

$$E_0(\lambda) = \sum_{n,m=1}^{\infty} c_{nm} v^n \lambda^m, \tag{4.67}$$

where $c_{11} = -e^{-2\pi^2 \gamma_0/a_0}$.

(ii) If $(\tilde{\omega},\omega) = 0$, then

$$E_0(\lambda) = \sum_{\substack{n=1 \\ m=4}}^{\infty} c_{nm} w^n \lambda^m, \tag{4.68}$$

where $c_{14} = -e^{-2\pi^2 \nu_0/a_0}$.

Proof: (i) Write

$$E_0(\lambda) = e^{-(2\pi^2/a_0)[\gamma_{-1}/\lambda + \gamma_0]}[1+\eta],$$

and substitute it in the first equation in (4.65). After some simplifications we get an equation of the form

$$F(\eta,v,\lambda)=0, \tag{4.69}$$

where F is analytic in all three variables in a neighborhood of zero, $F(0,0,0)=0$, and $[\partial F/\partial\eta](0,0,0)\neq 0$. Thus, by the implicit function theorem for analytic functions, (4.69) has a unique solution $\eta=\eta(v,\lambda)$ given by a convergent power series in the variables v and λ; this implies (4.67). The proof of (ii) is similar: We insert

$$E_0(\lambda)=e^{-(2\pi^2/a_0)[\nu_{-2}/\lambda^2+\nu_{-1}/\lambda+\nu_0]}[1+\eta]$$

in the second equation in (4.65) to obtain an equation of the form $F(\eta,w,\lambda)=0$ whose solution $\eta=\eta(w,\lambda)$ is analytic in both variables for $|w|$ and $|\lambda|$ small. Thus (4.68) follows. ■

There is a slight discrepancy between our expansion for $E_0(\lambda)$ and the corresponding expansion of Ref. 2 in the case $V=0$ in that the quantity w used here contains the factor λ^{-4} in contrast to the factor λ^{-3} in Ref. 2. The reason for having the factors λ^{-1} and λ^{-4} in front of the exponentials in (4.66) is due to the terms $e_{010}\,E\,\ln(-E)$ and $f_{01-2}E\,\ln(-E)\lambda^{-2}$ in (4.65), respectively. These terms arise from expanding $(\tilde{\omega},M_E\omega)$ for E small [cf. (4.45)].

This concludes our discussion of the behavior of the ground state. We now investigate what happens in the first gap, that is in the interval

$$(E_0^{(1)}(\pi)+E_0^{(2)}(\pi),\min\{E_1^{(1)}(\pi),E_1^{(2)}(\pi)\}),$$

provided

$$\min\{E_1^{(1)}(\pi),E_1^{(2)}(\pi)\}>E_0^{(1)}(\pi)+E_0^{(2)}(\pi). \tag{4.70}$$

We only consider the case $\lambda_0=0$ and content ourselves with obtaining the leading terms in the expansion of the eigenvalues. For this purpose we divide our analysis into two parts.

Case 1: Absorption at $E^*=E_0^{(1)}(\pi)+E_0^{(2)}(\pi)$. In this case we decompose K_E as follows:

$$K_E=K_E^{[E^*-\delta,E^*]}+K_E^{\mathbf{R}\setminus[E^*-\delta,E^*]},$$

where the notation [cf. (4.11)–(4.13)] is self-explanatory. The integral associated with $K_E^{[E^*-\delta,E^*]}$ [i.e., the analog of (4.14)] extends over the region $\{\theta\in[0,2\pi]^2: E^*-\delta\leqslant E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant E^*\}$. It contains the point (π,π) and is symmetric with respect to the lines $\theta_1=\pi$ and $\theta_2=\pi$. Using this symmetry and (4.6) and (4.7), the integration can be reduced to one over that portion of the region on which $\theta_1\leqslant\pi$ and $\theta_2\leqslant\pi$. It follows that we can write

$$K_E=\frac{b_0}{2\pi^2}\ln|E-E^*|L+M_E^{(1)}, \tag{4.71}$$

where now $b_0=2^{-1}\pi[d_0^{(1)}d_0^{(2)}]^{-1/2}$, with $d_0^{(1)}$ and $d_0^{(2)}$ being the constants in (4.9), $L\cdot=\omega(\tilde{\omega},\cdot)$, and $\tilde{\omega}=\omega\ \mathrm{sgn}\ W$, where

$$\omega(x)=W(x)^{1/2}\psi_0^{(1)}(x_1,\pi)\psi_0^{(2)}(x_2,\pi).$$

As in Theorem 4.5 one shows that $M_E^{(1)}\to M_{E^*}^{(1)}$ in the Hilbert–Schmidt norm as $E\downarrow E^*$. Note the sign change in the first term in (4.71) as compared to (4.35). Therefore, an eigenvalue can get absorbed at E^* only if

$$(\tilde{\omega},\omega)=\int_{\mathbf{R}^2}W(x)\psi_0^{(1)}(x_1,\pi)^2\psi_0^{(2)}(x_2,\pi)^2d^2x\geqslant 0.$$

Theorem 4.9: Suppose (4.1)–(4.3) and (4.70) hold. Then the following holds:
(i) If $(\tilde{\omega},\omega)>0$, then there is a unique, simple eigenvalue $E(\lambda)$ such that

$$E(\lambda)-E^*=e^{-(2\pi^2/b_0)[\mu_{-1}/\lambda+\mu_0+o(1)]},\quad \lambda\downarrow 0. \tag{4.72}$$

(ii) If $(\tilde{\omega},[M_{E*}^{(1)}]^k\omega)=0$ for $k=0,1,...,n-1$ and $(-1)^n(\tilde{\omega},[M_{E*}^{(1)}]^n\omega)>0$, then there are constants $c_0,\ldots,c_{n+1}$ such that

$$E(\lambda)-E^*=e^{c_0/\lambda^{n+1}+c_1/\lambda^n+\cdots+c_{n+1}+o(1)},\quad \lambda\downarrow 0,$$

where

$$c_0=\frac{2\pi^2(-1)^{n+1}}{b_0(\tilde{\omega},[M_{E*}^{(1)}]^n\omega)}<0.$$

(iii) If $(\tilde{\omega},[M_{E*}^{(1)}]^k\omega)=0$ for $k=0,1,...,n-1$ and $(-1)^n(\tilde{\omega},[M_{E*}^{(1)}]^n\omega)<0$, then λ_0 is not a c.c.th.

Proof: (i) is a special case of (ii). In analogy to (4.45), the equation for $E(\lambda)$ now reads

$$-\lambda(\tilde{\omega},[1+\lambda M_E^{(1)}]^{-1}\omega)=\frac{2\pi^2}{b_0\,\ln|E-E^*|}. \tag{4.73}$$

Then (ii) follows from (4.73) by proceeding as in the proof of Theorem 4.7(ii). The coefficients c_n are obtained by expanding $[1+\lambda M_{E*}^{(1)}]^{-1}$, noting that the difference $[1+\lambda M_E^{(1)}]^{-1}-[1+\lambda M_{E*}^{(1)}]^{-1}$ has no effect on these coefficients because it is controlled by an estimate similar to that in (4.53). Uniqueness and simplicity of the eigenvalue follow as in the proof of Theorem 4.7. ■

Of course, we could be more explicit about the remainder terms in (4.72) and write it in a form like (4.48). If $(\tilde{\omega},\omega)=0$, then in general nothing can be said about the sign of $(\tilde{\omega},M_{E*}^{(1)}\omega)$; that is, there is no analog of Lemma 4.6. The reason is that the proof of Lemma 4.6 uses in an essential way the fact that we are at the bottom of the spectrum.

Case 2: Absorption at $E^*=\min\{E_1^{(1)}(\pi),E_1^{(2)}(\pi)\}$. This case will be divided further into three cases. As we will see, in one of these cases it is possible that two eigenvalues approach E^* as $\lambda\downarrow 0$. We first describe the decompositions of K_E that characterize each case. Then, in Theorem 4.10, we collect the information about the behavior of the eigenvalues that converge to E^* as $\lambda\downarrow 0$.

(a) $E^*=E_1^{(1)}(\pi)$ and $E_1^{(2)}(\pi)>E^*$: In this case, for sufficiently small δ, the relevant region for the θ-integration is given by $\{\theta\in[0,2\pi]^2: E^*\leqslant E_1^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant E^*+\delta\}$. This region has two components containing the points $(\pi,0)$ and $(\pi,2\pi)$, respectively. The singular contribution to K_E again has rank one and we have the decomposition

$$K_E=-\frac{e_0}{2\pi^2}\ln|E-E^*|L+M_E^{(2)},$$

where $e_0=2^{-1}\pi[d_1^{(1)}c_0^{(2)}]^{-1/2}$, $L\cdot=\omega(\tilde{\omega},\cdot)$, $\tilde{\omega}=\omega\ \mathrm{sgn}\ W$, and

$$\omega(x)=W(x)^{1/2}\psi_1^{(1)}(x_1,\pi)\psi_0^{(2)}(x_2,0).$$

Moreover, $M_E^{(2)}\to M_{E*}^{(2)}$ in the Hilbert–Schmidt norm as $E\uparrow E^*$.

(b) $E^*=E_1^{(2)}(\pi)$ and $E_1^{(1)}(\pi)>E^*$: This case is analogous to case (a). We have

$$K_E=-\frac{f_0}{2\pi^2}\ln|E-E^*|L+M_E^{(3)},$$

where $f_0=2^{-1}\pi[d_1^{(2)}c_0^{(1)}]^{-1/2}$, $L\cdot=\omega(\tilde{\omega},\cdot)$, $\tilde{\omega}=\omega\ \mathrm{sgn}\ W$, and

$$\omega(x)=W(x)^{1/2}\psi_0^{(1)}(x_1,0)\psi_1^{(2)}(x_2,\pi).$$

Moreover, $M_E^{(3)}\to M_{E*}^{(3)}$ in the Hilbert–Schmidt norm as $E\uparrow E^*$.

(c) $E^*=E_1^{(1)}(\pi)=E_1^{(2)}(\pi)$: In this case the critical region of integration is $\{\theta\in[0,2\pi]^2: E^* \leqslant E_0^{(1)}(\theta_1)+E_1^{(2)}(\theta_2)\leqslant E^*+\delta\}\cup\{\theta\in[0,2\pi]^2: E^*\leqslant E_1^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant E^*+\delta\}$ and thus has four components located near $(\pi,0)$, $(2\pi,\pi)$, $(\pi,2\pi)$, and $(0,\pi)$. Let

$$\omega_1=W(x)^{1/2}\psi_0^{(1)}(x_1,0)\,\psi_1^{(2)}(x_2,\pi),\quad \omega_2=W(x)^{1/2}\psi_1^{(1)}(x_1,\pi)\,\psi_0^{(2)}(x_2,0). \tag{4.74}$$

Then K_E is of the form

$$K_E=d_1(E)\,\omega_1(\tilde{\omega}_1,\cdot)+d_2(E)\,\omega_2(\tilde{\omega}_2,\cdot)+R_E,$$

where

$$d_1(E)=\frac{1}{\pi^2}\int_{\mathscr{E}_{1,\delta}}\frac{d^2\theta}{E_0^{(1)}(\theta_1)+E_1^{(2)}(\theta_2)-E},$$

$$\mathscr{E}_{1,\delta}=\{\theta\in[0,\pi]^2: E^*\leqslant E_0^{(1)}(\theta_1)+E_1^{(2)}(\theta_2)\leqslant E^*+\delta\},$$

$$d_2(E)=\frac{1}{\pi^2}\int_{\mathscr{E}_{2,\delta}}\frac{d^2\theta}{E_1^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)-E},$$

$$\mathscr{E}_{2,\delta}=\{\theta\in[0,\pi]^2: E^*\leqslant E_1^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)\leqslant E^*+\delta\}.$$

Also, $R_E\to R_{E*}$ in the Hilbert–Schmidt norm as $E\uparrow E^*$. Since ω_1 and ω_2 are linearly independent, the operator $d_1(E)\,\omega_1(\tilde{\omega}_1,\cdot)+d_2(E)\,\omega_2(\tilde{\omega}_2,\cdot)$ has rank two. The coefficients $d_1(E)$ and $d_2(E)$ diverge as $E\uparrow E^*$ according to

$$d_1(E)=-\beta_1\,\ln|E-E^*|+O(1),$$

$$d_2(E)=-\beta_2\,\ln|E-E^*|+O(1),$$

where $\beta_1=(4\pi)^{-1}[d_1^{(2)}c_0^{(1)}]^{-1/2}$, $\beta_2=(4\pi)^{-1}[d_1^{(1)}c_0^{(2)}]^{-1/2}$. This follows from a result analogous to (4.25). Thus (1.8) reads

$$\det\begin{bmatrix}1+\lambda d_1(E)(\tilde{w}_1,[1+\lambda R_E]^{-1}w_1) & \lambda d_2(E)(\tilde{w}_1,[1+\lambda R_E]^{-1}w_2)\\ \lambda d_1(E)(\tilde{w}_2,[1+\lambda R_E]^{-1}w_1) & 1+\lambda d_2(E)(\tilde{w}_2,[1+\lambda R_E]^{-1}w_2)\end{bmatrix}=0. \tag{4.75}$$

The next theorem gives information about the behavior of the eigenvalues for each of the above situations.

Theorem 4.10: Suppose (4.1)–(4.3) and (4.70) hold. Then we have the following:

(a) Suppose $E^*=E_1^{(1)}(\pi)$ and $E_1^{(2)}(\pi)>E^*$: If $(\tilde{\omega},\omega)<0$, then there is a unique, simple eigenvalue obeying

$$E^*-E(\lambda)=e^{-(2\pi^2/e_0)[\sigma_{-1}/\lambda+\sigma_0+o(1)]},\quad \lambda\downarrow 0,$$

where

$$\sigma_{-1}=-\frac{1}{(\tilde{\omega},\omega)},\quad \sigma_0=-\frac{(\tilde{\omega},M_{E^*}^{(2)}\omega)}{(\tilde{\omega},\omega)^2}.$$

If $(\tilde{\omega},\omega)=0$, then a result analogous to Theorem 4.9(ii) holds. If $(\tilde{\omega},\omega)>0$, then $\lambda_0=0$ is not a c.c.th.

(b) Suppose $E^*=E_1^{(2)}(\pi)$ and $E_1^{(1)}(\pi)>E^*$: If $(\tilde{\omega},\omega)<0$, then there is a unique, simple eigenvalue obeying

$$E^*-E(\lambda)=e^{-(2\pi^2/f_0)[\rho_{-1}/\lambda+\rho_0+o(1)]},\quad \lambda\downarrow 0,$$

where

$$\rho_{-1} = -\frac{1}{(\tilde{\omega},\omega)}, \quad \rho_0 = -\frac{(\tilde{\omega},M^{(3)}_{E^*}\omega)}{(\tilde{\omega},\omega)^2}.$$

If $(\tilde{\omega},\omega)=0$, then a result analogous to Theorem 4.9(ii) holds. If $(\tilde{\omega},\omega)>0$, then $\lambda_0=0$ is not a c.c.th.

(c) Suppose $E^*=E_1^{(1)}(\pi)=E_1^{(2)}(\pi)$: Let

$$c_\pm = -\beta_1(\tilde{\omega}_1,\omega_1)-\beta_2(\tilde{\omega}_2,\omega_2)\pm\sqrt{[\beta_1(\tilde{\omega}_1,\omega_1)-\beta_2(\tilde{\omega}_2,\omega_2)]^2+4\beta_1\beta_2(\tilde{\omega}_1,\omega_2)^2}.$$

If $c_\pm>0$, then there are two eigenvalues (counting multiplicities) approaching E^* as $\lambda\downarrow 0$ given by

$$E^*-E(\lambda)=e^{-2/(c_\pm\lambda)+O(1)}, \quad \lambda\downarrow 0. \tag{4.76}$$

If $c_\pm$ are nonzero and of opposite signs, then there is only one eigenvalue, the one corresponding to c_+, that converges to E^* as $\lambda\downarrow 0$, and for it (4.76) holds.

Proof: The proofs in cases (a) and (b) are by now familiar. The result in case (c) follows by expanding the left-hand side of (4.75) for λ small and E near E^*, and using an estimate like (4.53). Further details are omitted. ■

We see from (4.75) that if $W\leqslant 0 (W\neq 0)$, then $c_\pm>0$ by the Schwarz inequality, $(\tilde{\omega}_1,\omega_2)^2 \leqslant (\tilde{\omega}_1,\omega_1)^2(\tilde{\omega}_2,\omega_2)^2$. Thus there are two eigenvalues (counting multiplicities) that converge to E^* as $\lambda\downarrow 0$. As one would expect, if V and W have certain symmetry properties, then the two eigenvalues may degenerate into one eigenvalue of multiplicity two. For example, this happens if $V^{(1)}$ and $V^{(2)}$ in (4.1) are given by identical functions, $V^{(1)}(\xi)=V^{(2)}(\xi)$ $(\xi\in\mathbf{R})$, W is symmetric, $W(x_1,x_2)=W(x_2,x_1)$, and even in x_1 and x_2 separately, and $E_1^{(1)}(\pi)$ is the first Dirichlet eigenvalue of the operator $H_0^{(1)}$ on the interval $[0,p]$. To see this, let $\mathscr{U}$ be the unitary operator $(\mathscr{U}\psi)(x_1,x_2)=\psi(x_2,x_1)$ and let $\mathscr{M}_1=\{\psi\in L^2(\mathbf{R}^2):\psi(x_1,x_2)=\psi(-x_1,x_2)=-\psi(x_1,-x_2)\}$ and $\mathscr{M}_2=\{\psi\in L^2(\mathbf{R}^2):\psi(x_1,x_2)=\psi(x_1,-x_2)=-\psi(-x_1,x_2)\}$. Then $\mathscr{M}_1$ and $\mathscr{M}_2$ are orthogonal, invariant subspaces for H_λ (for any $\lambda\geqslant 0$) such that $\mathscr{U}\mathscr{M}_1=\mathscr{M}_2$ and $\mathscr{U}\mathscr{M}_2=\mathscr{M}_1$. Moreover, $\mathscr{M}_1$ and $\mathscr{M}_2$ are invariant subspaces for K_E. By (4.74) and because of the symmetries of the potentials,

$$\omega_1=W(x)^{1/2}\psi_0^{(1)}(x_1,0)\psi_1^{(1)}(x_2,\pi), \quad \omega_2=W(x)^{1/2}\psi_1^{(1)}(x_1,\pi)\psi_0^{(1)}(x_2,0).$$

Since $\psi_0^{(1)}(\cdot,0)$ is even and $\psi_1^{(1)}(\cdot,\pi)$ is odd, $\omega_1\in\mathscr{M}_1$ and $\omega_2\in\mathscr{M}_2$. It follows that the restrictions of K_E to the invariant subspaces $\mathscr{M}_1$ and $\mathscr{M}_2$ are unitarily equivalent; in fact $K_E\restriction\mathscr{M}_1 = \mathscr{U}(K_E\restriction\mathscr{M}_2)\mathscr{U}$. Thus the eigenvalue in (4.76) is degenerate and of multiplicity two. Furthermore, $(\tilde{\omega}_1,\omega_2)=0$, $(\tilde{\omega}_1,\omega_1)=(\tilde{\omega}_2,\omega_2)$, $\beta_1=\beta_2$, and thus $c_+=c_-=-2\beta_1(\tilde{\omega}_1,\omega_1)$.

If $W\geqslant 0 (W\neq 0)$, then $c_\pm<0$, and there are no eigenvalues that converge to E^* as $\lambda\downarrow 0$. If one or both of $c_\pm$ are zero, then a more detailed investigation is necessary.

We do not discuss here in detail eigenvalues in higher gaps. We only remark that the singular part of K_E can have any rank. The following example shows how rank three might arise. Assume that $E_0^{(1)}=E_0^{(2)}(0)=0$, $E_0^{(1)}(\pi)=E_0^{(2)}(\pi)=1$, $E_1^{(1)}(\pi)=E_1^{(2)}(\pi)=5$, $E_1^{(1)}(0)=E_1^{(2)}(0)=6$, $E_2^{(1)}(0)=E_2^{(2)}(0)=7$, $E_2^{(1)}(\pi)=E_2^{(2)}(\pi)=8$, and $E_3^{(1)}(\pi)=E_3^{(2)}(\pi)=10$. Then the lowest two (open) gaps of the two-dimensional problem are (2,5) and (9,10) (the point {7} represents a closed gap). The singular part of K_E has rank 1 at $E=2$, rank 2 at $E=5$, rank 1 at $E=9$, and rank 3 at $E=10$. In the last instance we have $\omega_1(x)=W(x)^{1/2}\psi_0^{(1)}(x_1,0)\psi_3^{(2)}(x_2,\pi)$ $\omega_2(x)=W(x)^{1/2}\psi_3^{(1)}(x_1,\pi)\psi_0^{(2)}(x_2,0)$, and $\omega_3(x)=W(x)^{1/2}\psi_2^{(1)}(x_1,\pi)\psi_2^{(2)}(x_2,\pi)$ because $E_0^{(1)}(0)+E_3^{(2)}(\pi)=E_3^{(1)}(\pi)+E_0^{(2)}(0)=E_1^{(1)}(\pi)+E_1^{(2)}(\pi)=10$.

V. THREE DIMENSIONS

We briefly discuss the three-dimensional problem under the assumptions that

$$V(x)=V^{(1)}(x_1)+V^{(2)}(x_2)+V^{(3)}(x_3), \quad x=(x_1,x_2,x_3), \tag{5.1}$$

with W, $V^{(j)}$ real-valued, $V^{(j)}\in L^1_{loc}(\mathbf{R})$, and

$$V^{(j)}(x_j+p_j)=V(x_j), \quad p_j>0, \quad j=1,2,3, \tag{5.2}$$

$$W\in L^{3/2}(\mathbf{R}^3), \quad \int_{\mathbf{R}^3}(1+|x|)^{\gamma}|W(x)|d^3x, \tag{5.3}$$

for a suitable $\gamma>0$. As in one and two dimensions, H_0 can be defined by the forms method or as the closure of a suitable minimal operator (cf. Section IV). The condition $W\in L^{3/2}(\mathbf{R}^3)$ implies that $W\in\mathscr{R}$, the class of Rollnick potentials (Ref. 18, p. 170; Ref. 26). Such potentials are relatively form compact with respect to $-\Delta$ and so also with respect to H_0. We only consider eigenvalue absorption at the bottom of the continuous spectrum which we assume to be at $E=0$. As in the two-dimensional case we may suppose that the spectrum of each $H_0^{(j)}$ also begins at zero. The Birman–Schwinger kernel is given by

$$K_E=W^{1/2}(H_0-E)^{-1}|W|^{1/2}, \quad E<0,$$

and K_E is Hilbert–Schmidt because $W\in\mathscr{R}$. To shorten the notation we will write

$$\Psi_0(x,\theta)=\psi_0^{(1)}(x_1,\theta_1)\psi_0^{(2)}(x_2,\theta_2)\psi_0^{(3)}(x_3,\theta_3), \quad \theta=(\theta_1,\theta_2,\theta_3).$$

We also let

$$G(x,y;\theta)=\Psi_0(x,\theta)\overline{\Psi_0(y,\theta)}+\cdots, \tag{5.4}$$

where the dots indicate terms that are obtained from the first one by interchanging any number of elements between the ordered triples $\{x_1,x_2,x_3\}$ and $\{y_1,y_2,y_3\}$, respecting the order of the elements as in (4.17).

Theorem 5.1: Suppose (5.1)–(5.3) hold. Then K_E has the expansion

$$K_E=K_0+(-E)^{1/2}B+o(|E|^{1/2}), \quad E\uparrow 0, \tag{5.5}$$

where the symbol o refers to the Hilbert-Schmidt norm and B is a rank-one operator with kernel

$$B(x,y)=-\frac{1}{4\pi\sqrt{c_0^{(1)}c_0^{(2)}c_0^{(3)}}}W(x)^{1/2}\Psi_0(x,0)\Psi_0(y,0)|W(y)|^{1/2}. \tag{5.6}$$

We remark that if $V=0$, then $E_0^{(k)}(\theta_k)=\theta_k^2/p_k^2$, where the periods p_k are arbitrary positive numbers, so that $c_0^{(k)}=1/p_k^2$, $\Psi_0(x,0)=1/\sqrt{p_1p_2p_3}$. Hence

$$B(x,y)=-\frac{1}{4\pi}W(x)^{1/2}|W(y)|^{1/2},$$

which is a familiar result [see, e.g., Ref. 3, Eq. (2.7)].

Proof: In order to shorten the notation we define

$$E_0(\theta)=E_0^{(1)}(\theta_1)+E_0^{(2)}(\theta_2)+E_0^{(3)}(\theta_3).$$

(i) In place of (4.24) we now have

$$K_E^{[0,\delta]}(x,y)=W(x)^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{G(x,y;\theta)}{E_0(\theta)-E}\frac{d^3\theta}{(2\pi)^3}\right]|W(y)|^{1/2}, \tag{5.7}$$

where $\mathscr{E}_\delta=\{\theta\in[0,\pi]^3:E_0(\theta)\leqslant\delta\}$. We now write (5.7) as

$$K_E^{[0,\delta]}(x,y)=X_E^{[0,\delta]}(x,y)+Y_E^{[0,\delta]}(x,y), \tag{5.8}$$

$$X_E^{[0,\delta]}(x,y)=W(x)^{1/2}G(x,y;0)|W(y)|^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{d^3\theta}{(2\pi)^3[E_0(\theta)-E]}\right], \tag{5.9}$$

$$Y_E^{[0,\delta]}(x,y)=W(x)^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{H(x,y;\theta)}{E_0(\theta)-E}\,\frac{d^3\theta}{(2\pi)^3}\right]|W(y)|^{1/2}, \tag{5.10}$$

$$H(x,y;\theta)=G(x,y;\theta)-G(x,y;0). \tag{5.11}$$

First consider the integral on the right-hand side of (5.9). Following (4.26) and (4.27) we set $v_j=\sqrt{E_0^{(j)}(\theta_j)}$, $\theta_j=h_j(v_j)$, for $j=1,2,3$, and $r=\sqrt{v_1^2+v_2^2+v_3^2}$, $v_1=r\sin\chi\cos\varphi$, $v_2=r\sin\chi\sin\varphi$, $v_3=r\cos\chi$. Note that the domain of integration lies in the first octant. By using these spherical coordinates we obtain

$$\int_{\mathscr{E}_\delta}\frac{d^3\theta}{E_0(\theta)-E}=\int_0^{\sqrt{\delta}}\frac{g(r)r^2}{r^2-E}\,dr, \tag{5.12}$$

where

$$g(r)=\int_0^{\pi/2}\left(\int_0^{\pi/2}h_1'(r\sin\chi\cos\varphi)h_2'(r\sin\chi\sin\varphi)d\varphi\right)h_3'(r\cos\chi)\sin\chi\,d\chi.$$

For $r<r_1$ with $r_1>0$ sufficiently small, we can expand $g(r)$ as

$$g(r)=\sum_{m=0}^{\infty}b_{2m}r^{2m},$$

and insert it in (5.12). Note that [cf. (4.33)]

$$b_0=\frac{\pi}{2\sqrt{c_0^{(1)}c_0^{(2)}c_0^{(3)}}}. \tag{5.13}$$

This gives

$$\int_{\mathscr{E}_\delta}\frac{d^3\theta}{E_0(\theta)-E}=\sum_{m=0}^{\infty}b_{2m}J_m(E;\delta), \tag{5.14}$$

where

$$J_m(E;\delta)=\int_0^{\sqrt{\delta}}\frac{r^{2m+2}}{r^2-E}\,dr=(-1)^m(-E)^{m+1/2}\left[-\frac{\pi}{2}+\arctan\sqrt{-\frac{E}{\delta}}\right]+\sum_{k=1}^{m+1}\frac{\delta^{m-k+3/2}}{2m-2k+3}E^{k-1}. \tag{5.15}$$

Note that in contrast to (4.32) the right-hand side of (5.15) does not contain any logarithmic terms. From (5.14), (5.15), and (5.9), we see that the term $(-E)^{1/2}B$ in (5.5) comes from the first term on the right-hand side of (5.15) for $m=0$ and by using (5.13) and $G(x,y;0)=8\Psi_0(x,0)\Psi_0(y,0)$ which follows from (5.4) and the reality of $\Psi_0(x,0)$. The other terms on the right-hand side of (5.15) only affect either the term $X_0^{[0,\delta]}(x,y)$ or the corrections of order $O(|E|)$. Thus

$$X_E^{[0,\delta]}(x,y)=X_0^{[0,\delta]}(x,y)+(-E)^{1/2}B(x,y)+O(|E|). \tag{5.16}$$

Note that $X_E^{[0,\delta]}(x,y)$ is a Hilbert–Schmidt kernel because $W\in L^1(\mathbf{R}^3)$. Now we turn to the term $Y_E^{[0,\delta]}(x,y)$ in (5.10) which we write as $Y_E^{[0,\delta]}(x,y)=Y_0^{[0,\delta]}(x,y)+[Y_E^{[0,\delta]}(x,y)-Y_0^{[0,\delta]}(x,y)]$, and note that

$$Y_E^{[0,\delta]}(x,y)-Y_0^{[0,\delta]}(x,y)=EW(x)^{1/2}\left[\int_{\mathscr{E}_\delta}\frac{H(x,y;\theta)}{[E_0(\theta)-E]E_0(\theta)}\,\frac{d^3\theta}{(2\pi)^3}\right]|W(y)|^{1/2}. \tag{5.17}$$

As in (4.42) we have the estimate $|H(x,y;\theta)|\leq C(1+|x|)^\alpha(1+|y|)^\alpha|\theta|^\alpha(\alpha\in(0,1])$ which we insert in (5.17). From the θ-integration we get

$$\int_0^{\sqrt{\delta}} \frac{r^\alpha}{r^2-E}\,dr = |E|^{(\alpha-1)/2}\int_0^{\sqrt{\delta/|E|}} \frac{u^\alpha}{u^2+1}\,du.$$

If we pick any $\alpha\in(0,1)$, then the integral on the right-hand side stays bounded as $E\uparrow 0$ and thus the right-hand side is $O(|E|^{(\alpha-1)/2})$. Hence, by (5.17),

$$|Y_E^{[0,\delta]}(x,y)-Y_0^{[0,\delta]}(x,y)|\leqslant C|E|^{(\alpha+1)/2}|W(x)|^{1/2}(1+|x|)^\alpha(1+|y|)^\alpha|W(y)|^{1/2}. \tag{5.18}$$

So, by (5.3), if we choose α such that $0<\alpha<\min\{1,\gamma/2\}$, then $(1+|x|)^\alpha|W(x)|^{1/2}\in L^2(\mathbf{R}^3)$, which says that the right-hand side of (5.18) is Hilbert–Schmidt and of order $o(|E|^{1/2})$. Thus

$$Y_E^{[0,\delta]}(x,y)=Y_0^{[0,\delta]}(x,y)+o(|E|^{1/2}). \tag{5.19}$$

Since $K_E^{(\delta,\infty)}=K_0^{(\delta,\infty)}+O(|E|)$, by adding (5.16) and (5.19) and using $K_0=X_0^{[0,\delta]}(x,y)+Y_0^{[0,\delta]}(x,y)+K_0^{(\delta,\infty)}$, we obtain (5.5). ∎

Recall that λ_0 is a c.c.th. at $E=0$ if and only $\lambda_0 K_0$ has eigenvalue -1, resp., K_0 has eigenvalue $-\lambda_0^{-1}$. Hence $\lambda_0=0$ cannot be a c.c.th. in three dimensions for the class of potentials considered here. The following lemma will help us classify the possible behaviors of the eigenvalues $E(\lambda)$ that converge to zero as $\lambda\downarrow 0$.

Lemma 5.2: Assume (5.1), (5.2), and (5.3) with $\gamma>1$. Suppose that $K_0 f_0=-\lambda_0^{-1}f_0$, $f_0\in L^2(\mathbf{R}^3)$, $f_0\neq 0$, and $Bf_0=0$. Then zero is an eigenvalue of H_{λ_0}. If $-\lambda_0^{-1}$ is an eigenvalue of K_0 of geometric multiplicity k with geometric eigenspace $\mathcal{N}_0$ and $Bf=0$ for all $f\in\mathcal{N}_0$, then zero is an eigenvalue of H_{λ_0} of multiplicity k.

Proof: We will show that

$$\chi_0=H_0^{-1}|W|^{1/2}f_0 := s-\lim_{E\uparrow 0}(H_0-E)^{-1}|W|^{1/2}f_0 \tag{5.20}$$

is an eigenfunction of H_{λ_0} for the eigenvalue zero. To do this we start from the decomposition

$$(H_0-E)^{-1}|W|^{1/2}f_0=(H_0-E)^{-1}P^{[0,\delta]}|W|^{1/2}f_0+(H_0-E)^{-1}P^{(\delta,\infty)}|W|^{1/2}f_0, \tag{5.21}$$

and first establish the strong convergence as $E\uparrow 0$ of each term on the right-hand side. For the second term on the right-hand side of (5.21) this is clear from

$$(H_0-E)^{-1}P^{(\delta,\infty)}|W|^{1/2}f_0=[(H_0-E)^{-1}(H_0+1)^{1/2}P^{(\delta,\infty)}][(H_0+1)^{-1/2}|W|^{1/2}f_0],$$

since $(H_0+1)^{-1/2}|W|^{1/2}$ is a bounded operator and $\lim_{E\uparrow 0}(H_0-E)^{-1}(H_0+1)^{1/2}P^{(\delta,\infty)}$ exists in norm by the spectral theorem. As for the first term on the right-hand side of (5.21), our task is to show that

$$s-\lim_{E\uparrow 0}\int_{\mathbf{R}^3}\left[\int_{\mathscr{E}_\delta}\frac{G(x,y;\theta)}{E_0(\theta)-E}\frac{d^3\theta}{(2\pi)^3}\right]|W(y)|^{1/2}f_0(y)\,d^3y\in L^2(\mathbf{R}^3). \tag{5.22}$$

To do this we split the terms that make up $G(x,y;\theta)$ as follows; we indicate it for the first term on the right-hand side of (5.7):

$$\Psi_0(x,\theta)\overline{\Psi_0(y,\theta)}=\Psi_0(x,\theta)\Psi_0(y,0)+\Psi_0(x,\theta)[\overline{\Psi_0(y,\theta)}-\Psi_0(y,0)]. \tag{5.23}$$

Notice that when (5.23) is substituted in (5.22), the first term in (5.23) gives no contribution because of the assumption $Bf_0=0$ and (5.6). The integral arising from the second term is dealt with as follows. We set $\hat{x}=(x_1/p_1,x_2/p_2,x_3/p_3)$ so that

$$\Psi_0(x,\theta)=e^{i\theta\cdot\hat{x}}\frac{\xi_0^{(1,+)}(x_1,E_0^{(1)}(\theta_1))}{N_0^{(1,+)}(\theta_1)}\frac{\xi_0^{(2,+)}(x_2,E_0^{(2)}(\theta_2))}{N_0^{(2,+)}(\theta_2)}\frac{\xi_0^{(3,+)}(x_3,E_0^{(3)}(\theta_3))}{N_0^{(3,+)}(\theta_3)}, \tag{5.24}$$

from which, by expanding the functions $\xi_0^{(+)}(x_j,E_0^{(j)}(\theta_j))$ in powers of θ_j, we obtain

$$\Psi_0(x,\theta)=\sum_{m,n,s=0}^{\infty} a_{mns}(x)\,\theta_1^m\theta_2^n\theta_3^s, \tag{5.25}$$

where the coefficients $a_{mns}(x)$ obey

$$|a_{mns}(x)|\leqslant M\tilde{\theta}^{-m-n-s}, \tag{5.26}$$

for some $M>0$ and $\tilde{\theta}>0$, and where $|\theta_j|<\tilde{\theta}$. As a result, from (5.22)–(5.25) we obtain

$$\int_{\mathbf{R}^3}\left[\int_{\mathscr{E}_\delta}\frac{G(x,y;\theta)}{E_0(\theta)-E}\frac{d^3\theta}{(2\pi)^3}\right]|W(y)|^{1/2}f_0(y)d^3y=\sum_{m,n,s=0}^{\infty}a_{mns}(x)\int_{\mathscr{E}_\delta}e^{i\theta\cdot\hat{x}}g_{mns}(\theta,E)d^3\theta+\cdots, \tag{5.27}$$

where

$$g_{mns}(\theta,E)=\frac{\theta_1^m\theta_2^n\theta_3^s}{(2\pi)^3[E_0(\theta)-E]}\int_{\mathbf{R}^3}\Delta(y;\theta)|W(y)|^{1/2}f_0(y)d^3y,$$

$$\Delta(y;\theta)=\overline{\Psi_0(y;\theta)}-\Psi_0(y;0),$$

and where the dots in (5.27) indicate further terms similar to the one displayed [cf. (5.4)]. By Lemma 4.4,

$$|\Delta(y;\theta)|\leqslant C(1+|y|)^{\alpha}|\theta|^{\alpha},\quad \alpha\in(0,1],$$

and thus using the Schwarz inequality gives

$$\left|\int_{\mathbf{R}^3}\Delta(y;\theta)|W(y)|^{1/2}f_0(y)d^3y\right|\leqslant C|\theta|^{\alpha}\|W\|_{2\alpha}^{1/2}\ \|f_0\|,$$

where $\|W\|_{2\alpha}=\int_{\mathbf{R}^3}(1+|y|)^{2\alpha}|W(y)|d^3y$. Now, since (5.3) holds with $\gamma>1$, we may choose α such that $1/2<\alpha<\min\{1,\gamma/2\}$. This guarantees that

$$\frac{|\theta|^{m+n+s+\alpha}}{E_0(\theta)-E}\in L^2(\mathscr{E}_\delta),$$

for all $E\leqslant 0$ with the L^2-norm being bounded independently of E. Therefore

$$\|g_{mns}(\theta,E)\|_{L^2(\mathscr{E}_\delta)}\leqslant C\delta^{(m+n+s+\alpha)/2-1/4}\|W\|_{2\alpha}^{1/2}\|f_0\|,$$

for all $E\leqslant 0$. By the Plancherel theorem and (5.26), we conclude that

$$\left\|a_{mns}(x)\int_{\mathscr{E}_\delta}e^{i\theta\cdot\hat{x}}g_{mns}(\theta,E)d^3\theta\right\|_{L^2(\mathbf{R}^s)}\leqslant\tilde{C}\delta^{\alpha/2-1/4}\left(\frac{\sqrt{\delta}}{\tilde{\theta}}\right)^{m+n+s}\|W\|_{2\alpha}^{1/2}\|f_0\|.$$

Thus, if we choose $\sqrt{\delta}<\tilde{\theta}$, then the series on the right-hand side of (5.27) converges absolutely in the L^2-norm. This implies that $\|(H_0-E)^{-1}P^{[0,\delta]}|W|^{1/2}f_0\|$ is bounded independently of E for $E<0$. Moreover, this norm is increasing in E and hence converges as $E\uparrow 0$. In addition, $(H_0-E)^{-1}P^{[0,\delta]}|W|^{1/2}f_0$ converges weakly as $E\uparrow 0$ because $(\varphi,(H_0-E)^{-1}P^{[0,\delta]}|W|^{1/2}f_0)$ converges for any φ from the dense set $\{\varphi\in L^2:\ \varphi\in\mathrm{Ran}\ P^{[\epsilon,\infty)}$ for some $\epsilon>0\}$. These facts imply (5.22), and hence we have shown that $\chi_0=H_0^{-1}|W|^{1/2}f_0\in L^2(\mathbf{R}^3)$. It remains to show that $H_{\lambda_0}\chi_0=0$. We argue as follows. Let $\psi\in\mathscr{D}(H_{\lambda_0})$. Thus $\psi\in\mathscr{D}(H_0^{1/2})$ and so

$$
\begin{aligned}
(\chi_0, H_{\lambda_0}\psi) &= (H_0^{-1}|W|^{1/2}f_0, H_{\lambda_0}\psi) \\
&= \lim_{E\uparrow 0}([H_0-E]^{-1}|W|^{1/2}f_0, H_{\lambda_0}\psi) \\
&= \lim_{E\uparrow 0}([H_0+1]^{-1/2}H_{\lambda_0}[H_0-E]^{-1}|W|^{1/2}f_0, [H_0+1]^{1/2}\psi) \\
&= \lim_{E\uparrow 0}([H_0+1]^{-1/2}|W|^{1/2}(1+\lambda_0 K_E)f_0, [H_0+1]^{1/2}\psi) = 0,
\end{aligned}
$$

because $K_E \to K_0$ in norm and $-\lambda_0^{-1}$ is an eigenvalue of K_0 with eigenvector f_0. Thus $\chi_0 \in \mathscr{D}(H_{\lambda_0}^*) = \mathscr{D}(H_{\lambda_0})$ and $H_{\lambda_0}\chi_0 = 0$. To prove the last assertion, suppose that $f_1, \ldots, f_k$ are linearly independent eigenvectors of K_0 for the eigenvalue $-\lambda_0^{-1}$. Multiplying the equation $c_1 H_0^{-1}|W|^{1/2}f_1 + \cdots + c_k H_0^{-1}|W|^{1/2}f_k = 0$ from the left by $W^{1/2}$ and using that $\lambda_0 W^{1/2}H_0^{-1}|W|^{1/2}f_j = -f_j$ $(j=1,\ldots,k)$ implies $c_1 = \cdots = c_k = 0$. ■

From now on we assume

$$
W(x) \leqslant 0 \quad \text{a.e.}, \tag{5.28}
$$

which implies that $K_E = -|W|^{1/2}(H_0 - E)^{-1}|W|^{1/2}$ is self-adjoint. Assumption (5.28) is made here only for convenience, in order to avoid discussion of certain technical points that arise when we apply perturbation theory to K_E. The case where W changes sign has been discussed in Ref. 3 (Section 8). The next theorem is similar to Theorem 2.3 in Ref. 3 and to Theorem 2.1 in Ref. 9.

Theorem 5.3: Suppose V and W satisfy (5.1), (5.2), (5.3) with $\gamma > 1$, and (5.28). Let λ_0 be a c.c.th. at $E=0$. Suppose that m $(m \geqslant 1)$ eigenvalues approach zero as $\lambda \downarrow \lambda_0$. Then we have the alternatives:

(i) 0 is not an eigenvalue of H_{λ_0}, $m=1$, and there is a unique eigenvalue obeying $E(\lambda) = -c(\lambda-\lambda_0)^2 + o([\lambda-\lambda_0]^2)$ with $c \neq 0$.

(ii) 0 is an eigenvalue of H_{λ_0} in which case at most one eigenvalue is in case (i) while the other eigenvalues obey $E(\lambda) = -c(\lambda-\lambda_0) + o(\lambda-\lambda_0)$ with $c \neq 0$. Moreover, 0 is either an eigenvalue of H_{λ_0} of multiplicity $m-1$ or m, depending on whether or not there is an eigenvalue in case (i).

Proof: By the Birman–Schwinger principle, $-\lambda_0^{-1}$ is an eigenvalue of K_0 of multiplicity m. Let $\mathscr{N}_0$ denote the corresponding eigenspace. For $E<0$ there are m eigenvalues $\tau_j(E)$ $(j=1,\ldots,m)$ of K_E converging to $-\lambda_0^{-1}$ and the $\tau_j(E)$ are strictly decreasing functions of E. We can obtain $\tau_j(E)$ from (5.5) by using eigenvalue perturbation theory and then solve $\lambda\tau_j(E) = -1$ for $E(\lambda)$ to get the eigenvalue branch $E_j(\lambda)$.

Considering (i), since 0 is not an eigenvalue of H_{λ_0}, by Lemma 5.2, we have that $Bf \neq 0$ for all nonzero $f \in \mathscr{N}_0$. However, since B has rank one, this can only happen if $m=1$. Let $f_0 \in \mathscr{N}_0$ with $\|f_0\| = 1$. Then, by (5.6) and (5.28), $(f_0, Bf_0) > 0$. From perturbation theory[17] and (5.5), we infer that

$$
\tau(E) = -\lambda_0^{-1} + (f_0, Bf_0)(-E)^{1/2} + o(|E|^{1/2}),
$$

which leads to $E(\lambda) = -c(\lambda-\lambda_0)^2 + o([\lambda-\lambda_0]^2)$ with $c = \lambda_0^{-4}(f_0, Bf_0)^{-2}$; so (i) is proved.

To prove (ii), notice that B either vanishes on all of $\mathscr{N}_0$ or on a subspace of dimension $m-1$ of $\mathscr{N}_0$. Thus, by Lemma 5.2, 0 is an eigenvalue of H_{λ_0} of multiplicity m or $m-1$, respectively. Now the number of eigenvalues that behave like $-c(\lambda-\lambda_0) + o(\lambda-\lambda_0)$ with $c>0$ as $\lambda \downarrow \lambda_0$ is equal to the multiplicity of 0 as eigenvalue of H_{λ_0}. To see this, let P_0 denote the orthogonal projection onto the kernel of H_{λ_0}. Then

$$
P_0 W P_0 = -\lambda_0^{-1} P_0 H_0 P_0 \leqslant 0,
$$

so that, since H_0 has a trivial kernel, $P_0 W P_0$ also has a trivial kernel on Ran P_0. Thus $P_0 W P_0$, which is self-adjoint and negative, has $p = \dim[\text{Ran } P_0]$ strictly negative eigenvalues, and

$p=m-1$ or m. Then Theorem 2.4(ii) in Ref. 10 implies that the number of eigenvalues that approach 0 as $\lambda\downarrow\lambda_0$ and which, for λ sufficiently close to λ_0, obey an estimate of the form

$$-a(\lambda-\lambda_0)<E(\lambda)<-b(\lambda-\lambda_0), \tag{5.29}$$

with $a>b>0$, is equal to p. However, from the proof of that theorem one also infers that there is a $c>0$ such that for any $\epsilon>0$ we can choose a and b such that $c-\epsilon<b<a<c+\epsilon$ and so that (5.29) is satisfied as $\lambda\downarrow\lambda_0$ ($-c$ is just one of the p nonzero eigenvalues of P_0WP_0). Since $\epsilon>0$ is arbitrary, it follows that we can replace (5.29) by the stronger statement

$$E(\lambda)=-c(\lambda-\lambda_0)+o(\lambda-\lambda_0), \quad \lambda\downarrow\lambda_0. \tag{5.30}$$

If $p=m$, then all m eigenvalues obey (5.30), with the coefficient c depending on the eigenvalue, of course. If $p=m-1$, then there is exactly one eigenvalue that does not obey (5.29). Then B does not vanish on all of $\mathscr{N}_0$ and so, by perturbation theory, there is one eigenvalue that is in case (i), and this must be the eigenvalue that does not obey (5.29). ■

APPENDIX: PROOFS OF LEMMAS 2.3–2.7

1. Proof of Lemma 2.3

We first prove the entire lemma when $\lambda=0$. Obviously, (i) and (iv) immediately follow from (2.7). To prove (ii) and (iii) we note that from (2.3), (2.4), (2.17), and (2.30), we have

$$\phi_0(x,E)=-d_0(E)[\psi_0^{(+)}(x,E)-\psi_0^{(-)}(x,E)]. \tag{A1}$$

In order to simplify the notation we define

$$g(x,\mu)=\xi_0^{(+)}(x,E(\tilde{n}\pi/p+i\mu)), \tag{A2}$$

where we now assume that an endpoint E_n of the gap $\mathscr{G}_{\tilde{n}}$, together with the proper branch of $E(\tilde{n}\pi/p+i\mu)$, have been selected. We temporarily use the shorthand notation $\tilde{\psi}_0^{(\pm)}(x,\mu)=\psi_0^{(\pm)}(x,E(\tilde{n}\pi/p+i\mu))$ and note that, by (2.3), (2.4), and (2.17),

$$\tilde{\psi}_0^{(+)}(x,\mu)=\tilde{\psi}_0^{(-)}(x,-\mu).$$

This implies, by (2.7),

$$\xi_0^{(-)}(x,E(\tilde{n}\pi/p+i\mu))=e^{2i\tilde{n}\pi x/p}g(x,-\mu), \tag{A3}$$

and thus

$$\tilde{\psi}_0^{(\pm)}(x,\mu)=e^{i\tilde{n}\pi x/p}e^{\mp\mu z}g(x,\pm\mu). \tag{A4}$$

Thus, from (A2)–(A4) it follows that [again using $E=E(\tilde{n}\pi/p+i\mu)$]

$$\begin{aligned}\psi_0^{(+)}(x,E)-\psi_0^{(-)}(x,E)=e^{i\tilde{n}\pi x/p}[e^{-\mu x}g(x,\mu)-e^{\mu x}g(x,-\mu)]-e^{i\tilde{n}\pi x/p}\sinh\mu x[g(x,\mu)\\ +g(x,-\mu)]+e^{i\tilde{n}\pi x/p}\cosh\mu x[g(x,\mu)-g(x,-\mu)].\end{aligned} \tag{A5}$$

Next we note that, by (2.30),

$$d_0(E)=\frac{\alpha}{\mu}+\beta\mu+O(\mu^3), \tag{A6}$$

where α and β are constants; in particular,

$$\alpha=\frac{(-1)^{\tilde{n}}\phi_0(p,E_n)}{2p}.$$

Moreover, $d_0(E)$ is an odd function of μ. Furthermore (a dot means $d/d\mu$),

$$g(x,\mu)+g(x,-\mu)=2\xi_0^{(+)}(x,E_n)+O(\mu^2), \tag{A7}$$

$$g(x,\mu)-g(x,-\mu)=2\dot{\xi}_0^{(+)}(x,E_n)\mu+O(\mu^3). \tag{A8}$$

Therefore, on inserting (A5)–(A8) in (A1) we get

$$\phi_0(x,E)=2\alpha\xi_0^{(+)}(x,E_n)e^{i\tilde{n}\pi x/p}\mu^{-1}\sinh\mu x-2\alpha\dot{\xi}_0^{(+)}(x,E_n)e^{i\tilde{n}\pi x/p}\cosh\mu x+O(\mu)\sinh\mu x$$
$$+O(\mu^2)\cosh\mu x. \tag{A9}$$

Now (ii) follows from (A9) and the estimates

$$|\sinh\mu x|\leqslant C\,\frac{\mu|x|}{1+\mu|x|}\,e^{\mu|x|},\quad \cosh\mu x\leqslant Ce^{\mu|x|}. \tag{A10}$$

Similarly, on differentiating (A9) with respect to μ and using (A10), along with the estimate

$$|\sinh\mu x-\mu x\cosh\mu x|\leqslant C\,\frac{\mu^3|x|^3}{(1+\mu|x|)^2}\,e^{\mu|x|}, \tag{A11}$$

and assuming $\mu<1$, we obtain (iii). Thus the lemma has been proved when $\lambda=0$. Now suppose that $\lambda>0$. In view of (2.19) and (2.21) it suffices to prove (i) and (iv) for $F_\lambda^{(\pm)}(x,E)$ instead of $\psi_\lambda^{(\pm)}(x,E)$. By using (2.11), (2.17), (A2), and (A4), we obtain

$$A_0(x,y;E)=2\alpha e^{(i\tilde{n}\pi/p)(x+y)}\xi_0^{(+)}(x,E_n)\xi_0^{(+)}(y,E_n)\mu^{-1}\sinh\mu(x-y)-2\alpha e^{(i\tilde{n}\pi/p)(x+y)}[\dot{\xi}_0^{(+)}$$
$$\times(x,E_n)\xi_0^{(+)}(y,E_n)-\xi_0^{(+)}(x,E_n)\dot{\xi}_0^{(+)}(y,E_n)]\cosh\mu(x-y)+O(\mu)\sinh\mu(x-y)$$
$$+O(\mu^2)\cosh\mu(x-y).$$

Thus, by using (A10), (A11), and obvious variants of them, we obtain the bounds

$$|A_0(x,y;E)|\leqslant C(1+|x-y|)e^{\mu|x-y|}, \tag{A12}$$

$$|\dot{A}_0(x,y;E)|\leqslant C\,\frac{\mu(1+|y-x|)^3}{(1+\mu|y-x|)^2}\,e^{\mu|y-x|}, \tag{A13}$$

To prove (i), use (A12) and (i) with $\lambda=0$ in (2.10). This yields, for $x>0$,

$$|F_\lambda^{(+)}(x,E)|\leqslant Ce^{-\mu x}+C\lambda e^{-\mu x}\int_x^\infty(1+y)|W(y)|e^{\mu y}|F_\lambda^{(+)}(y,E)|dy. \tag{A14}$$

Multiplying (A14) by $e^{\mu x}$ and applying Gronwall's inequality proves (i) for $F_\lambda^{(+)}(x,E)$. The proof for $F_\lambda^{(-)}(x,E)$ is similar. Considering (iv) we start from

$$\dot{F}_\lambda^{(+)}(x,E)=\dot{\psi}_0(x,E)-\lambda\int_x^\infty\dot{A}_0(x,y;E)W(y)F_\lambda^{(+)}(y,E)dy-\lambda\int_x^\infty A_0(x,y;E)W(y)\dot{F}_\lambda^{(+)}(y,E)dy.$$

Thus, by (A12), (A13), and (i) for $\lambda=0$, we conclude that

$$|\dot{F}_\lambda^{(+)}(x,E)|\leqslant C(1+x)e^{-\mu x}+C\lambda e^{-\mu x}\int_x^\infty\frac{\mu(1+y)}{(1+\mu y)^2}\,(1+y)^2|W(y)|dy+C\lambda e^{-\mu x}$$
$$\times\int_x^\infty|W(y)|(1+y)^2\,\frac{e^{\mu y}|\dot{F}_\lambda^{(+)}(y,E)|}{1+y}\,dy. \tag{A15}$$

Dividing both sides by $(1+x)e^{-\mu x}$ and applying Gronwall's inequality yields (iv). Note that condition (H2) enters through the second and third term on the right-hand side of (A15). To prove (ii) we use the fact that $\phi_\lambda(x,E)$ is a solution of the integral equation

$$\phi_\lambda(x,E)=\phi_0(x,E)+\lambda\int_0^x A_0(x,y;E)W(y)\phi_\lambda(y,E)dy. \tag{A16}$$

By using (ii) with $\lambda=0$ and (A12) we obtain (assuming $x>0$)

$$|\phi_\lambda(x,E)|=C(1+x)e^{\mu x}+C(1+x)e^{\mu x}\int_0^x e^{-\mu y}W(y)|\phi_\lambda(y,E)|dy. \tag{A17}$$

Dividing (A17) by $(1+x)e^{\mu x}$ and using Gronwall's inequality yields (ii) for $x>0$. When $x<0$, the argument is similar, so (ii) is proved. The proof of (iii) is similar to the proof of (iv). We differentiate (A16) with respect to μ and use (ii) and (iii) for $\lambda=0$, together with (A12), (A13), and the monotonicity of the function $y\to\mu(1+y)^2/[1+\mu y]^2$ (assuming $\mu<1$). This gives (for $x>0$)

$$\frac{e^{-\mu x}(1+\mu x)^2|\dot\phi_\lambda(x,E)|}{\mu(1+x)^3}\leq C+C\int_0^x(1+y)|W(y)|\,\frac{e^{-\mu y}(1+\mu y)^2|\dot\phi_\lambda(x,E)|}{\mu(1+y)^3}\,dy.$$

An application of Gronwall's inequality yields the desired estimate when $x>0$, and an analogous estimate holds when $x<0$. Note that only condition (H1) is needed here. ■

2. Proof of Lemma 2.4

(i) The bound immediately follows from Lemma 2.3(i), (ii), and (2.28). To prove (ii) use (2.26), (2.28), part (i), and the dominated convergence theorem, to conclude that

$$\|R_{\lambda,E}-R_{\lambda,E_n}\|_{H.S.}^2=\int_{\mathbf{R}^2}|W(x)||G^D_{\lambda,E}(x,y)-G^D_{\lambda,E_n}(x,y)|^2|W(y)|dxdy\to 0,$$

as $E\to E_n$. Finally, (iii) follows by differentiating (2.28) and applying Lemma 2.3. ■

3. Proof of Lemma 2.5

Let $C_1>0$ be a constant such that

$$|\omega_\lambda(x,E)|\leq C_1|W(x)|^{1/2},\quad |\dot\psi_\lambda(x,E)|\leq C_1|W(x)|^{1/2}(1+|x|). \tag{A18}$$

These estimates follow from Lemma 2.3(i), (iv), and (2.27). Let

$$\sigma=\int_{-\infty}^{\infty}|W(x)|(1+|x|)dx.$$

Then using (2.27), Lemma 2.4(i), and

$$\int_{-\infty}^{\infty}(1+\min\{|x|,|z|\})|W(z)|(1+\min\{|z|,|y|\})dz\leq\sigma(1+\min\{|x|,|y|\}),$$

we deduce by induction that

$$|(R^s_{\lambda,E})(x,y)|\leq C^s\sigma^{s-1}|W(x)|^{1/2}[1+\min\{|x|,|y|\}]|W(y)|^{1/2},\quad s=1,2,...,$$

and thus, by using (A18),

$$|(R^s_{\lambda,E}\omega_{\lambda,E})(x)|\leq C_1C^s\sigma^s|W(x)|^{1/2},\quad s=0,1,2,.... \tag{A19}$$

Now (i) follows from (A18) and (A19) which give

$$|(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})| \leq C_1^2 C^s \sigma^{s+1}. \tag{A20}$$

So choosing $\tilde{C} = C_1^2\sigma$, $r = C\sigma$ proves estimate (i). To prove (ii) we formally differentiate $(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})$:

$$\frac{d(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})}{d\mu} = (\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\dot{\omega}_{\lambda,E}) + (\dot{\tilde{\omega}}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E}) + \sum_{j=0}^{s-1} (\tilde{\omega}_{\lambda,E}, R^j_{\lambda,E}\dot{R}_{\lambda,E}R^{s-j-1}_{\lambda,E}\omega_{\lambda,E}). \tag{A21}$$

Then we note that

$$|(R^{*s}_{\lambda,E}\tilde{\omega}_{\lambda,E})(x)| \leq C_1 C^s \sigma^s |W(x)|^{1/2}, \tag{A22}$$

where the * denotes the adjoint, and so

$$|\dot{\tilde{\omega}}(x,E)(R^s_{\lambda,E}\omega_{\lambda,E})(x)| \leq C_1^2 C^s \sigma^s |W(x)|(1+|x|), \tag{A23}$$

$$|\dot{\omega}_\lambda(x,E)(R^{*s}_{\lambda,E}\tilde{\omega}_{\lambda,E})(x)| \leq C_1^2 C^s \sigma^s |W(x)|(1+|x|), \tag{A24}$$

$$|(R^{*j}_{\lambda,E}\tilde{\omega}_{\lambda,E})(x)\dot{R}_{\lambda,E}(x,y)(R^{s-j-1}_{\lambda,E}\omega_{\lambda,E})(y)| \leq C_1^2 C^s \sigma^{s-1} |W(x)|(1+|x|)(1+|y|)|W(y)|. \tag{A25}$$

For the last inequality we used (A20), (A22), and (2.34). Since the various scalar products in (A25) are multiple integrals involving the functions $R_{\lambda,E}(x,y)$, $\omega_\lambda(x,E)$, $\tilde{\omega}_\lambda(x,E)$, and their derivatives with respect to μ, the estimates (A22)–(A25) provide us with the necessary L^1-bounds on the integrands to justify the formal differentiation leading to (A21) (see, e.g., Ref. 36, Theorem 16.8). From (A23)–(A25) we obtain

$$|(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\dot{\omega}_{\lambda,E})| \leq C_1^2 C^s \sigma^{s+1}, \quad |(\dot{\tilde{\omega}}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})| \leq C_1^2 C^s \sigma^{s+1}, \tag{A26}$$

$$|(\tilde{\omega}_{\lambda,E}, R^j_{\lambda,E}\dot{R}_{\lambda,E}R^{s-j-1}_{\lambda,E}\omega_{\lambda,E}| \leq C_1^2 C^s \sigma^{s+1}. \tag{A27}$$

Inserting (A26) and (A27) in (A21) gives

$$\left|\frac{d(\tilde{\omega}_{\lambda,E}, R^s_{\lambda,E}\omega_{\lambda,E})}{d\mu}\right| \leq (2+s) C_1^2 C^s \sigma^{s+1},$$

and so assertion (ii) follows on setting $r = C\sigma$, $\tilde{C} = (2+s)C_1^2\sigma$. ■

4. Proof of Lemma 2.6

In proving (i)–(iii) it suffices to consider the case $a_N(0) > 0$ and, in (iii), to find the solution $z_+(\mu)$; the modifications needed in the other cases are obvious. We always assume that $0 \leq \mu \leq \mu_0$ and $0 \leq z < r^{-1}$ so that the series on the left-hand side of (2.35) converges absolutely, and uniformly in μ. We begin by writing (2.35) in the form

$$a_N(\mu) z^N \left(1 + \sum_{j=1}^{\infty} \frac{a_{N+j}(\mu)}{a_N(\mu)} z^j\right) = \dot{c}(0)\mu \left(1 - \sum_{j=1}^{N-1} \frac{a_j(\mu)}{\mu \dot{c}(0)} z^j + o(1)\right), \tag{A28}$$

where the $o(1)$-term is a function of μ and comes from $c(\mu) = \dot{c}(0)\mu[1+o(1)]$. Note that the infinite sum on the left of (A28) is $O(z)$ uniformly in μ and that the finite sum on the right is also $O(z)$ uniformly in μ because $\mu^{-1}a_j(\mu)$ is bounded by assumption (b). Therefore, for any $\delta \in (0,1)$ there is $z_1 > 0$ and $\mu_1 > 0$ $(z_1 < r^{-1}, \mu_1 \leq \mu_0)$ such that

$$\begin{aligned} a_N(\mu) z^N (1-\delta) &\leq \dot{c}(0)\mu(1+\delta), \\ a_N(\mu) z^N (1+\delta) &\geq \dot{c}(0)\mu(1-\delta), \end{aligned} \tag{A29}$$

for $0 \leqslant z \leqslant z_1$ and $0 \leqslant \mu \leqslant \mu_1$. Define

$$h(\mu,z) = \sum_{j=1}^{\infty} a_j(\mu) z^j - c(\mu), \tag{A30}$$

so that (2.35) becomes

$$h(\mu,z) = 0. \tag{A31}$$

Now choose any two numbers b_1 and b_2 such that $0 < b_1 < [\dot{c}(0)/a_N(0)]^{1/N} < b_2$. Since δ in (A29) is arbitrary, we conclude that for μ_1 and z_1 sufficiently small, every pair $\{\mu,z\}$ satisfying (A31), $0 \leqslant \mu \leqslant \mu_1$, and $0 < z \leqslant z_1$, also satisfies $b_1 \mu^{1/N} < z < b_2 \mu^{1/N}$. Moreover, a simple calculation using (b) and (c) gives

$$\lim_{\mu \to 0} \mu^{-1} h(\mu, b_1 \mu^{1/N}) = a_N(0) b_1^N - \dot{c}(0) < 0,$$

$$\lim_{\mu \to 0} \mu^{-1} h(\mu, b_2 \mu^{1/N}) = a_N(0) b_2^N - \dot{c}(0) > 0.$$

Therefore, if μ is sufficiently small, the function $z \mapsto h(\mu,z)$ has a zero $z(\mu)$ in the interval $[b_1 \mu^{1/N}, b_2 \mu^{1/N}]$. This proves the existence of solutions to (2.35). To prove uniqueness we consider the partial derivative $h_z(\mu,z)$ for fixed μ and $z \in [b_1 \mu^{1/N}, b_2 \mu^{1/N}]$. We have

$$h_z(\mu,z) = \sum_{j=1}^{N-1} j a_j(\mu) z^{j-1} + N a_N(\mu) z^{N-1} + \sum_{j=N+1}^{\infty} j a_j(\mu) z^{j-1}.$$

The first sum is $O(\mu)$ by assumption (b), the term in the middle is bounded from below by $N[a_N(0)/2] b_1^{N-1} \mu^{1-1/N}$ for μ small, and the last sum is $O(\mu)$. Hence the middle term dominates as $\mu \to 0$, and so $h_z(\mu,z) > 0$ on $[b_1 \mu^{1/N}, b_2 \mu^{1/N}]$. Thus the solution $z(\mu)$ is unique. The continuity of $z(\mu)$ follows from the continuity of $h(\mu,z)$ in both variables.

Assertions (iv) and (v) follow directly from (2.35) which becomes inconsistent. First, if N is even and $a_N(0) < 0$, then the two sides of (2.35) have different signs for μ and z small enough. The sign of z does not matter. Secondly, if $a_j(0) = 0$ for all j with the specified bound on $|a_j(\mu)|$, the left-hand side of (2.35) is bounded by $C_1 r_1 (1 - r_1 |z|)^{-1} |z| \mu$ provided $|z| < \min\{r^{-1}, r_1^{-1}\}$. This bound is incompatible with the right-hand side and the assumption that $\dot{c}(0) > 0$. Thus Lemma 2.6 is proved. ■

5. Proof of Lemma 2.7

(i) The existence of a solution $\mu(z)$ follows from (A28) arguing similarly as in the proof of Lemma 2.6. Considering only the case where $a_N(0) > 0$ and assuming $z \geqslant 0$, we see that the solution $\mu(z)$ is located in an interval of the form $[\tilde{b}_1 z^N, \tilde{b}_2 z^N]$, where $0 < \tilde{b}_1 < a_N(0)/\dot{c}(0) < \tilde{b}_2$. The uniqueness is a consequence of the fact that $h(\mu,z)$ is differentiable with respect to μ if $|z| < \min\{r^{-1}, r_2^{-1}\}$ and $h_\mu(\mu,z) = -\dot{c}(0) + o(1)$ on the interval $[\tilde{b}_1 z^N, \tilde{b}_2 z^N]$. The proof of (iii) proceeds by substituting $\mu(z) = [a_N(0)/\dot{c}(0)] z^N [1 + \varphi(z)]$ in (2.35) and using $a_j(\mu) = \dot{a}_j(0)\mu + o(\mu)$ for $j = 1,\ldots,N-1$, and the given expansion for $c(\mu)$, to find the leading contribution of order $O(z)$ to $\varphi(z)$. The result is (2.37). ■

[1] B. Simon, ''The bound states of weakly coupled Schrödinger operators in one and two dimensions,'' Ann. Phys. **97**, 279–299 (1976).
[2] H. Holden, ''On coupling constant thresholds in two dimensions,'' J. Op. Theory **14**, 263–276 (1985).
[3] M. Klaus and B. Simon, ''Coupling constant thresholds in nonrelativistic quantum mechanics. I. Short-range two-body case,'' Ann. Phys. **130**, 251–281 (1980).
[4] R. Blankenbecler, M. L. Goldberger, and B. Simon, ''The bound states of weakly coupled long-range one-dimensional quantum Hamiltonians,'' Ann. Phys. **108**, 69–78 (1977).
[5] M. Klaus, ''On the bound states of Schrödinger operators in one dimension,'' Ann. Phys. **108**, 288–300 (1977).
[6] S. Fassari, ''Spectral properties of the Kronig-Penney Hamiltonian with a localised impurity,'' J. Math. Phys. **30**, 1385–1392 (1989).

[7] C. Beattie (private communication).
[8] G. Hagedorn (private communication).
[9] B. Simon, ''On the absorption of eigenvalues by continuous spectrum in regular perturbation problems,'' J. Funct. Anal. **25**, 338–344 (1977).
[10] M. Klaus, ''Some applications of the Birman-Schwinger principle,'' Helv. Phys. Acta **55**, 49–68 (1982).
[11] P. A. Deift and R. Hempel, ''On the existence of eigenvalues of the Schrödinger operator $H-\lambda W$ in a gap of $\sigma(H)$,'' Commun. Math. Phys. **103**, 461–490 (1986).
[12] S. Alama, P. A. Deift, and R. Hempel, ''Eigenvalue branches of the Schrödinger Operator $H-\lambda W$ in a gap of $\sigma(H)$,'' Commun. Math. Phys. **121**, 291–321 (1989).
[13] F. Gesztesy, D. Gurarie, H. Holden, M. Klaus, L. Sadun, and P. Vogl, ''Trapping and cascading of eigenvalues in the large coupling limit,'' Commun. Math. Phys. **118**, 597–634 (1988).
[14] V. A. Zheludev, ''Perturbation of the spectrum of the one-dimensional Schrödinger operator with a periodic potential,'' in *Topics in Mathematical Physics*, edited by M. Sh. Birman (Consultants Bureau, New York, 1971), Vol 4, pp. 55–75.
[15] F. Gesztesy, and B. Simon, ''A short proof of Zheludev's theorem,'' Trans. Am. Math. Soc. **335**, 329–340 (1993).
[16] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators with Applications to Quantum Mechanics and Global Geometry* (Springer-Verlag, New York, 1987).
[17] T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, New York, 1976).
[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, II: Fourier Analysis, Self-Adjointness* (Academic, New York, 1978).
[19] J. Weidmann, *Spectral Theory of Ordinary Differential Operators, Lecture Notes in Mathematics* 1258 (Springer-Verlag, New York, 1987).
[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, IV: Analysis of Operators* (Academic, New York, 1978).
[21] N. E. Firsova, ''The Riemann surface of a quasimomentum, and scattering theory for a perturbed Hill operator, Mathematical questions in the theory of wave propagation, 7,'' Zap. Naučn. Semin., LOMI **51**, 183–196 (1975) (in Russian).
[22] R. G. Newton, ''Inverse scattering by a local impurity in a periodic potential in one dimension,'' J. Math. Phys. **24**, 2152–2162 (1983).
[23] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-Order Differential Equations, Part II* (Oxford University Press, Oxford, 1958).
[24] D. P. Clemence and M. Klaus, ''Continuity of the *S* matrix for the perturbed Hill's equation,'' J. Math. Phys. **35**, 3285–3300 (1994).
[25] M. Klaus, ''Exact behavior of Jost functions at low energy,'' J. Math. Phys. **29**, 148–154 (1988).
[26] B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms* (Princeton University Press, Princeton NJ, 1971).
[27] C. Kittel, *Introduction to Solid State Physics*, 2nd ed. (Wiley, New York, 1961).
[28] M. S. P. Eastham, *The Spectral Theory of Periodic Differential Equations* (Scottish Academic, Edinburgh-London, 1975).
[29] F. Gesztesy and Z. Zhao, ''On critical and subcritical Sturm-Liouville operators,'' J. Funct. Anal. **98**, 311–345 (1991).
[30] R. Hempel, ''A left-indefinite generalized eigenvalue problem for Schrödinger operators,'' Ludwig-Maximilians-Universität München, Habilitationsschrift, 1987.
[31] J. Engländer and R. G. Pinsky, ''The asymptotic behavior of the principal eigenvalue for small perturbation of critical one-dimensional Schrödinger operators with $V(x)=l_{\pm}/x^2$ for $\pm x \gg 1$,'' J. Funct. Anal. **133**, 501–515 (1995).
[32] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics*, Texts and Monographs in Physics (Springer-Verlag, New York, 1988).
[33] J. Weidmann, *Linear Operators in Hilbert Spaces*, Graduate Texts in Mathematics (Springer-Verlag, New York, 1980).
[34] B. Simon, *Trace Ideal Methods and Their Applications* (Cambridge University Press, Cambridge, 1979).
[35] F. Gesztesy and Z. Zhao, ''On positive solutions of critical Schrödinger operators in two dimensions,'' J. Funct. Anal. **127**, 235–256 (1993).
[36] P. Billingsley, *Probability and Measure* (Wiley, New York, 1986).

JWKB connection-formula problem revisited again via Borel summation

Nanny Fröman and Per Olof Fröman
Department of Theoretical Physics, University of Uppsala, Box 803, S-751 08 Uppsala, Sweden

(Received 12 December 1997; accepted for publication 27 April 1998)

Silverstone [Phys. Rev. Lett. **55**, 2523–2526 (1985)] has expressed the opinion that the traditional version of the JWKB connection formulas associated with a classical turning point is incorrect and that the correct version follows from the Borel-based summability of the asymptotic expansions for the Airy functions. In the present paper we show that this assertion is incorrect. After the Borel summation of the asymptotic expressions for the Airy functions has actually been performed explicitly, there are no connection problems, and thus no connection formulas, but *exact* relations expressing the functions $\mathrm{Ai}(z)$ and $\mathrm{Bi}(z)$ in terms of the Whittaker function, although with two defects: Due to the derivation there appear false restrictions on $\arg z$, and there is an artificial cut that coincides with the Stokes line along the real z-axis. We also show that these *exact* relations can be obtained more appropriately in a direct way with the aid of standard formulas available in handbooks, whereby the defects mentioned do not appear. On the other hand, when one uses truncated asymptotic series, the Stokes phenomenon gives rise to connection problems, and the connection formulas, which one uses for handling these connection problems, are inherently one-directional. This property of the connection formulas is demonstrated and discussed anew, in general terms as well as with special regard to the Airy functions. © *1998 American Institute of Physics.*
[S0022-2488(98)00609-4]

I. INTRODUCTION

The properties of the JWKB connection formulas have been the subject of much discussion during several decades in the past. Misconceptions, apparently due to lack of clarity and precision in the formulation of the connection problem, have given rise to controversies. The discussion between Langer[1] and Jeffreys[2] is well known. Both claimed that the connection formulas are one-directional, but Langer misunderstood Jeffreys and thus objected against Jeffreys who was, however, quite right. In Fröman and Fröman[3] and Olver[4] connection formulas associated with a simple turning point were derived, and rigorous upper bounds for the errors were given. Fröman and Fröman[3] demonstrated anew the one-directional nature of the connection formulas. A still more detailed analysis of the connection formulas, with numerical confirmation of their one-directional nature, was given by N. Fröman.[5] Thus, with the connection problem clearly formulated, the correctness as well as the one-directional nature of the connection formulas has long since been clarified. According to Olver (Ref. 6, p. 517) absences of explicit boundary conditions for the solutions and bounds for the error terms could be regarded as contributory causes of the misunderstandings of the meaning of the connection formulas that have occurred frequently in the literature.

In opposition to the results quoted above, Silverstone[7] claims that the traditional version of the JWKB connection formulas associated with a classical turning point is incorrect and that the correct version follows from the Borel summability of the asymptotic expansions for the Airy functions. To analyze this recipe it is thus sufficient to consider the Airy differential equation,

$$\frac{d^2\psi}{dz^2} - z\psi = 0, \tag{1.1}$$

 © 1998 American Institute of Physics

and to investigate what the result of the Borel summation of the asymptotic expansions for the Airy functions really is.

In Sec. II we write down the asymptotic expansions of the Airy functions, and in Sec. III we perform the Borel summation of these asymptotic expansions. This results in *exact* relations that express the Airy functions in terms of the Whittaker function, although with the defects that there are false restrictions on $\arg z$ and that there is an artificial cut that coincides with the Stokes line along the real z-axis. In Sec. IV we derive the exact formulas expressing the Airy functions in terms of the Whittaker function in a more appropriate, direct way with the aid of standard formulas available in handbooks. This direct procedure yields formulas that correctly reproduce the single-valuedness of the Airy functions, and hence there appear no artificial restrictions on $\arg z$ and no artificial cut. For the differential equation $d^2\psi/dx^2+x\psi=0$ and the first order of the JWKB approximation, we show in Sec. V in a concrete way that Silverstone's[7] Eqs. (24) and (25) are not correct connection formulas. In Sec. VI we reconsider the connection formulas and demonstrate again in an alternative way their one-directional nature. We summarize our conclusions in Sec. VII.

II. ASYMPTOTIC EXPANSIONS OF THE AIRY FUNCTIONS

According to 10.4.59 in Abramowitz and Stegun (Ref. 8, p. 448) we have

$$\mathrm{Ai}(z)\sim\frac{1}{2\sqrt{\pi}}\,z^{-1/4}e^{-\zeta}\sum_{k=0}^{\infty}\frac{c_k}{(-\zeta)^k},\quad z\to\infty,\quad |\arg z|<\pi, \tag{2.1}$$

with

$$\zeta=\tfrac{2}{3}z^{3/2}, \tag{2.2}$$

$$c_k=\frac{\Gamma(3(k+1/6))}{2^k3^{3k}\Gamma(k+1/2)\Gamma(k+1)} \tag{2.3a}$$

$$=\frac{\Gamma(k+1/6)\Gamma(k+5/6)}{2^k\Gamma(k+1)\Gamma(1/6)\Gamma(5/6)}, \tag{2.3b}$$

where the last member has been obtained with the aid of 6.1.19 and 6.1.17 in Abramowitz and Stegun (Ref. 8, p. 256). Using Eq. (5) in Luke (Ref. 9, p. 32), we obtain from (2.3a) the asymptotic formula

$$c_k\sim\frac{(2\pi)^{1/2}[(k-1/2)/e]^{k-1/2}}{2\pi2^k}\sim\frac{\Gamma(k)}{2\pi2^k},\quad k\to\infty. \tag{2.4}$$

Recalling that the Airy function Ai is single-valued, and using 10.4.7 in Abramowitz and Stegun (Ref. 8, p. 446), we obtain

$$\mathrm{Ai}(-z)=e^{+i\pi/3}\mathrm{Ai}(ze^{+i\pi/3})+e^{-i\pi/3}\mathrm{Ai}(ze^{-i\pi/3}). \tag{2.5}$$

With the aid of (2.1) we obtain from (2.5) the formula

$$\mathrm{Ai}(-z)\sim\frac{1}{2\sqrt{\pi}}\,z^{-1/4}\left[e^{+i(\zeta-\pi/4)}\sum_{k=0}^{\infty}\frac{c_k}{(i\zeta)^k}+e^{-i(\zeta-\pi/4)}\sum_{k=0}^{\infty}\frac{c_k}{(-i\zeta)^k}\right],\quad |\arg z|<2\pi/3, \tag{2.6}$$

which we can also obtain by rewriting 10.4.60 in Abramowitz and Stegun (Ref. 8, p. 448).

According to 10.4.9 in Abramowitz and Stegun (Ref. 8, p. 446) we have

$$\mathrm{Bi}(z)=\pm2ie^{\mp i\pi/3}\,\mathrm{Ai}(ze^{\pm2\pi i/3})\mp i\,\mathrm{Ai}(z), \tag{2.7}$$

and with the aid of (2.1) we obtain from (2.7)

$$\mathrm{Bi}(z)\sim\frac{1}{\sqrt{\pi}}\,z^{-1/4}\left[e^{\zeta}\sum_{k=0}^{\infty}\ \frac{c_k}{\zeta^k}\mp\frac{1}{2}\,ie^{-\zeta}\sum_{k=0}^{\infty}\ \frac{c_k}{(-\zeta)^k}\right],$$

$$-\pi<\arg z<\pi/3(\text{upper sign}),\quad -\pi/3<\arg z<\pi(\text{lower sign}). \tag{2.8}$$

From 10.4.9 in Abramowitz and Stegun (Ref. 8, p. 446) we obtain

$$\mathrm{Bi}(-z)=2e^{\pm i\pi/6}\,\mathrm{Ai}(ze^{\mp i\pi/3})\mp i\,\mathrm{Ai}(-z), \tag{2.9}$$

and with the aid of (2.1) we obtain from (2.9) the formula

$$\mathrm{Bi}(-z)\sim\frac{1}{2\sqrt{\pi}}\,z^{-1/4}\left[e^{+i(\zeta+\pi/4)}\sum_{k=0}^{\infty}\ \frac{c_k}{(i\zeta)^k}+e^{-i(\zeta+\pi/4)}\sum_{k=0}^{\infty}\ \frac{c_k}{(-i\zeta)^k}\right],$$

$$|\arg z|<2\pi/3, \tag{2.10}$$

which we can also obtain by rewriting 10.4.64 in Abramowitz and Stegun (Ref. 8, p. 449).

In Sec. III we shall explicitly perform the Borel summation of the asymptotic expansions (2.1), (2.6), (2.8), and (2.10).

III. BOREL SUMMATION OF THE ASYMPTOTIC SERIES

Since

$$\frac{1}{k!}\int_0^{\infty}e^{-t}t^k\,dt=1, \tag{3.1}$$

one can in each term of a divergent, infinite series introduce the left-hand member of (3.1) as a factor. When one then interchanges the summation and the integration, the infinite series in the integrand may be convergent, and the integral may also be convergent. The result of this procedure is called Borel summation of the original divergent series.

Recalling (3.1), one can write (2.1) as

$$\mathrm{Ai}(z)\sim\frac{1}{2\sqrt{\pi}}\,z^{-1/4}e^{-\zeta}\sum_{k=0}^{\infty}\ \frac{c_k}{(-\zeta)^k}\,\frac{1}{k!}\int_0^{\infty}e^{-t}t^k\,dt,\quad |\arg z|<\pi. \tag{3.2}$$

Inserting the expression for c_k according to (2.3b) into (3.2), and interchanging the order of summation and integration, and writing $=$ instead of $\sim$, since at the end we do not get an asymptotic but an *exact* formula, we obtain

$$\mathrm{Ai}(z)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}e^{-\zeta}\int_0^{\infty}dt\ e^{-t}\,\frac{\Gamma(1)}{\Gamma(1/6)\Gamma(5/6)}\sum_{k=0}^{\infty}\ \frac{\Gamma(k+1/6)\Gamma(k+5/6)}{\Gamma(k+1)}\,\frac{1}{k!}\left(-\frac{t}{2\zeta}\right)^k,$$

$$|\arg z|<\pi. \tag{3.3}$$

Recalling (2.3b) and (2.4), we see that the infinite series in (3.3) is convergent when $|-t/(2\zeta)|<1$. Its sum (with the constant factor in front of the summation sign included) is equal to ${}_2F_1(1/6,5/6;1;-t/(2\zeta))$ according to 15.1.1 in Abramowitz and Stegun (Ref. 8, p. 556). Therefore we can write (3.3) as

$$\mathrm{Ai}(z)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}e^{-\zeta}\int_0^{\infty}dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta)),\quad |\arg z|<\pi. \tag{3.4}$$

Using in Buchholz Eqs. (30) and (19) (Ref. 10, pp. 78 and 19, respectively), we obtain

$$\int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta))=e^{\zeta}W_{0,-1/3}(2\zeta)=e^{\zeta}W_{0,1/3}(2\zeta),$$

$$|\arg\ \zeta|<\pi,\ \text{ i.e., }\ |\arg\ z|<2\pi/3. \tag{3.5}$$

Inserting (3.5) into (3.4), we obtain

$$\mathrm{Ai}(z)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}W_{0,-1/3}(2\zeta)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}W_{0,1/3}(2\zeta),\quad |\arg\ z|<2\pi/3. \tag{3.6}$$

Proceeding analogously as above, we shall from (2.6), (2.8), and (2.10) now obtain the functions $\mathrm{Ai}(-z)$, $\mathrm{Bi}(z)$, and $\mathrm{Bi}(-z)$ expressed in terms of the Whittaker function.

We obtain from (2.6),

$$\mathrm{Ai}(-z)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}\Big[e^{i(\zeta-\pi/4)}\int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta e^{-i\pi/2}))$$
$$+e^{-i(\zeta-\pi/4)}\int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta e^{i\pi/2}))\Big],\quad |\arg\ z|<2\pi/3, \tag{3.7}$$

and then from (3.7),

$$\mathrm{Ai}(-z)=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}[e^{-i\pi/4}W_{0,-1/3}(2\zeta e^{-i\pi/2})+e^{+i\pi/4}W_{0,-1/3}(2\zeta e^{i\pi/2})]$$
$$=\frac{1}{2\sqrt{\pi}}\,z^{-1/4}[e^{-i\pi/4}W_{0,1/3}(2\zeta e^{-i\pi/2})+e^{+i\pi/4}W_{0,1/3}(2\zeta e^{i\pi/2})],\quad |\arg\ z|<\pi/3. \tag{3.8}$$

Formula (2.8) yields

$$\mathrm{Bi}(z)=\frac{1}{\sqrt{\pi}}\,z^{-1/4}\Big[e^{\zeta}\int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta e^{\pm i\pi}))$$
$$\mp\frac{1}{2}\,ie^{-\zeta}\int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta))\Big],$$

$$-\pi<\arg\ z<\pi/3(\text{upper sign}),\quad -\pi/3<\arg\ z<\pi(\text{lower sign}), \tag{3.9}$$

where in the first hypergeometric function we have written $-t/(2\zeta e^{\pm i\pi})$ in order to make the formula to be derived below valid in the largest possible region of $\arg z$. From (3.9) we obtain

$$\mathrm{Bi}(z)=\frac{1}{\sqrt{\pi}}\,z^{-1/4}\Big[W_{0,-1/3}(2\zeta e^{\pm i\pi})\mp\frac{1}{2}\,iW_{0,-1/3}(2\zeta)\Big]$$
$$=\frac{1}{\sqrt{\pi}}\,z^{-1/4}\Big[W_{0,1/3}(2\zeta e^{\pm i\pi})\mp\frac{1}{2}\,iW_{0,1/3}(2\zeta)\Big],$$

$$-2\pi/3<\arg\ z<0(\text{upper sign}),\quad 0<\arg\ z<2\pi/3(\text{lower sign}). \tag{3.10}$$

Putting in Eq. (21b) in Buchholz (Ref. 10, p. 20) $\kappa=0$ and $\mu=\pm 2/3$, and replacing z by 2ζ, and using the reflection formula for the Γ-function, we obtain

$$W_{0,\pm 1/3}(2\zeta) = -i[W_{0,\pm 1/3}(2\zeta e^{+i\pi}) - W_{0,\pm 1/3}(2\zeta e^{-i\pi})]. \tag{3.11}$$

With the aid of this formula it is easily seen that from (3.10) one obtains the same formula when one uses the upper signs as when one uses the lower signs, viz.

$$\mathrm{Bi}(z) = \frac{1}{2\sqrt{\pi}}\, z^{-1/4}[W_{0,\pm 1/3}(2\zeta e^{+i\pi}) + W_{0,\pm 1/3}(2\zeta e^{-i\pi})], \quad |\arg z| < 2\pi/3, \tag{3.12}$$

where the region of $\arg z$ has been obtained with the use of both the upper and the lower signs in (3.10).

From (2.10) we obtain

$$\begin{aligned}\mathrm{Bi}(-z) = \frac{1}{2\sqrt{\pi}}\, z^{-1/4}\Bigg[& e^{+i(\zeta+\pi/4)} \int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta e^{-i\pi/2})) \\ & + e^{-i(\zeta+\pi/4)} \int_0^\infty dt\ e^{-t}\ {}_2F_1(1/6,5/6;1;-t/(2\zeta e^{+i\pi/2}))\Bigg], \quad |\arg z| < 2\pi/3,\end{aligned} \tag{3.13}$$

and then from (3.13),

$$\begin{aligned}\mathrm{Bi}(-z) &= \frac{1}{2\sqrt{\pi}}\, z^{-1/4}[e^{+i\pi/4} W_{0,-1/3}(2\zeta e^{-i\pi/2}) + e^{-i\pi/4} W_{0,-1/3}(2\zeta e^{i\pi/2})] \\ &= \frac{1}{2\sqrt{\pi}}\, z^{-1/4}[e^{+i\pi/4} W_{0,1/3}(2\zeta e^{-i\pi/2}) + e^{-i\pi/4} W_{0,1/3}(2\zeta e^{i\pi/2})], \quad |\arg z| < \pi/3.\end{aligned} \tag{3.14}$$

IV. DIRECT DERIVATION OF THE FORMULAS OBTAINED BY BOREL SUMMATION

We have according to 10.4.14 in Abramowitz and Stegun (Ref. 8, p. 447)

$$\mathrm{Ai}(z) = \frac{1}{\pi}\,(z/3)^{1/2} K_{1/3}(\zeta) \tag{4.1}$$

and according to Eq. (8) in Luke (Ref. 9, p. 213)

$$K_{1/3}(\zeta) = \left(\frac{\pi}{2\zeta}\right)^{1/2} W_{0,1/3}(2\zeta). \tag{4.2}$$

Inserting (4.2) into (4.1) and using (2.2), we obtain (3.6), but without the restriction $|\arg z| < 2\pi/3$.

We have according to 10.4.15 in Abramowitz and Stegun (Ref. 8, p. 447)

$$\mathrm{Ai}(-z) = \tfrac{1}{2}(z/3)^{1/2}[e^{+i\pi/6} H^{(1)}_{1/3}(\zeta) + e^{-i\pi/6} H^{(2)}_{1/3}(\zeta)] \tag{4.3}$$

and according to Eqs. (2a) and (2b) in Buchholz (Ref. 10, p. 207)

$$e^{-i\pi/6} H^{(2)}_{1/3}(\zeta) = \left(\frac{2}{\pi\zeta}\right)^{1/2} e^{+i\pi/4} W_{0,1/3}(2\zeta e^{i\pi/2}), \tag{4.4a}$$

$$e^{+i\pi/6} H^{(1)}_{1/3}(\zeta) = \left(\frac{2}{\pi\zeta}\right)^{1/2} e^{-i\pi/4} W_{0,1/3}(2\zeta e^{-i\pi/2}). \tag{4.4b}$$

Inserting (4.4a) and (4.4b) into (4.3) and using (2.2), we obtain (3.8), but without the restriction $|\arg z| < 2\pi/3$.

With the aid of 10.4.14 in Abramowitz and Stegun (Ref. 8, p. 447) and (2.2) we obtain from (2.7)

$$\mathrm{Bi}(z)=\pm\frac{2i}{\pi}\,(z/3)^{1/2}K_{1/3}(\zeta e^{\pm i\pi})\mp\frac{i}{\pi}\,(z/3)^{1/2}K_{1/3}(\zeta). \tag{4.5}$$

When we use Eq. (2c) in Buchholz (Ref. 10, p. 207) we obtain from (4.5) the expressions for $\mathrm{Bi}(z)$ in the last member of (3.10), but without the restrictions on $\arg z$. When we then use Eq. (19) in Buchholz (Ref. 10, p. 19), we obtain the first expressions for $\mathrm{Bi}(z)$ in (3.10), but without the restrictions on $\arg z$.

We have according to 10.4.19 in Abramowitz and Stegun (Ref. 8, p. 447)

$$\mathrm{Bi}(-z)=\tfrac{1}{2}i(z/3)^{1/2}[e^{+i\pi/6}H^{(1)}_{1/3}(\zeta)-e^{-i\pi/6}H^{(2)}_{1/3}(\zeta)]. \tag{4.6}$$

Inserting (4.4a) and (4.4b) into (4.6) and using (2.2), we obtain (3.14), but without the restriction $|\arg z|<\pi/3$.

With the aid of general formulas for special functions we have above in a direct way derived (3.6), (3.8), (3.10), and (3.14), but without obtaining the restrictions on $\arg z$ that appear when these formulas are derived by means of Borel summation. We shall now illuminate the nature of these formulas further by deriving (3.8) from (3.6) and (3.14) from (3.10), still without obtaining any restrictions on $\arg z$. To this purpose it is convenient to use (2.2) to rewrite (3.6), (3.8), (3.10), and (3.14) as

$$\mathrm{Ai}(z)=\frac{(2/3)^{1/6}}{2\sqrt{\pi}}\,\zeta^{-1/6}W_{0,1/3}(2\zeta), \tag{4.7a}$$

$$\mathrm{Ai}(-z)=\frac{(2/3)^{1/6}}{2\sqrt{\pi}}\,\zeta^{-1/6}[e^{-i\pi/4}W_{0,1/3}(2\zeta e^{-i\pi/2})+e^{+i\pi/4}W_{0,1/3}(2\zeta e^{i\pi/2})], \tag{4.7b}$$

$$\mathrm{Bi}(z)=\frac{(2/3)^{1/6}}{\sqrt{\pi}}\,\zeta^{-1/6}\left[W_{0,1/3}(2\zeta e^{\pm i\pi})\mp\frac{1}{2}\,iW_{0,1/3}(2\zeta)\right], \tag{4.8a}$$

$$\mathrm{Bi}(-z)=\frac{(2/3)^{1/6}}{2\sqrt{\pi}}\,\zeta^{-1/6}[e^{+i\pi/4}W_{0,1/3}(2\zeta e^{-i\pi/2})+e^{-i\pi/4}W_{0,1/3}(2\zeta e^{i\pi/2})], \tag{4.8b}$$

respectively. When in (4.7a) we replace z by $-z$ and ζ by $\zeta e^{\pm 3i\pi/2}$ according to (2.2), we obtain (4.7b) by using (3.11) with ζ replaced by $\zeta e^{\pm i\pi/2}$. According to the conclusion in connection with (3.11) and (3.12) in Sec. III, it is clear that one obtains the same function when one uses the upper signs as when one uses the lower signs in (4.8a), and we can therefore use the upper signs when they are most convenient and the lower signs when they are most convenient. When in (4.8a) we replace z by $-z$ and ζ by $\zeta e^{\mp 3i\pi/2}$, where the upper and lower sign belongs together with the upper and lower signs in (4.8a), we obtain (4.8b) with the aid of (3.11) when we there replace ζ by $\zeta e^{\pm i\pi/2}$.

V. DEMONSTRATION OF THE FACT THAT SILVERSTONE'S EQS. (24) AND (25) ARE NOT CORRECT CONNECTION FORMULAS

Silverstone (Ref. 7) claims that the correct connection formulas are expressed by his Eqs. (24) and (25), which, with $\hbar=1$ and the turning point denoted by t, and on noting that $S=|S|$, $dS/dx=|dS/dx|$, $Q=|Q|$, $-dQ/dx=|dQ/dx|$, we can write as

$$\psi\sim 2b|dS/dx|^{-1/2}\cos(|S|+\pi/4)+2a|dS/dx|^{-1/2}\cos(|S|-\pi/4),\quad \arg(x-t)\approx 0, \tag{5.1a}$$

$$\psi\sim 2b|dQ/dx|^{-1/2}\exp(+|Q|)+(a\pm ib)|dQ/dx|^{-1/2}\exp(-|Q|),\quad \arg(t-x)\approx 0,\quad \mp\,\mathrm{Im}\;x>0. \tag{5.1b}$$

In (5.1a) and (5.1b), which are thus equivalent to Silverstone's (Ref. 7) Eqs. (24) and (25), respectively, S and Q are Silverstone's notations for phase-integrals in the classically allowed and the classically forbidden region, respectively. In each one of these phase-integrals the limits of integration are t and x, and the integrand is assigned to an arbitrary order of the JWKB-approximation and is thus a sum in which each term is characterized by the summation index n. In the discussion in the present section we shall, instead of the notations of the previous sections, use the same notations as in (5.1a) and (5.1b), and we shall assume that the differential equation is $d^2\psi/dx^2+x\psi=0$. For the function ψ in (5.1a) and (5.1b) we then have [see Silverstone's (Ref. 7) Eq. (4)]

$$\psi=2\pi^{1/2}[a\ \mathrm{Ai}(-x)+b\ \mathrm{Bi}(-x)]. \tag{5.2}$$

If one retains the whole sum from $n=0$ to $n=\infty$ in the expressions for S and Q in (5.1a) and (5.1b), expands the phase-integral functions to obtain functions corresponding to those in Sec. II, and performs Borel summation, one obtains *exact* expressions for the Airy functions in terms of the Whittaker function, and thus there appear no connection problems and no connection formulas; see Sec. II. However, when the sums in the expressions for S and Q are truncated, the formulas (5.1a) and (5.1b) are *not correct connection formulas*. Below we shall verify this fact for the first order of the JWKB-approximation.

We decompose the function (5.2) into pure phase-integral functions by defining in the classically allowed region complex conjugate, x-dependent coefficients $A(x)$ and $B(x)=A^*(x)$ such that we have *exactly*

$$\psi(x)=A|dS/dx|^{-1/2}\exp(+i|S|)+B|dS/dx|^{-1/2}\exp(-i|S|),\quad x>t, \tag{5.3a}$$

$$\frac{d\psi(x)}{dx}=A\frac{d}{dx}\{|dS/dx|^{-1/2}\exp(+i|S|)\}+B\frac{d}{dx}\{|dS/dx|^{-1/2}\exp(-i|S|)\},\quad x>t, \tag{5.3b}$$

and by defining in the classically forbidden region real, x-dependent coefficients $C(x)$ and $D(x)$ such that we have *exactly*

$$\psi(x)=C|dQ/dx|^{-1/2}\exp(+|Q|)+D|dQ/dx|^{-1/2}\exp(-|Q|),\quad x<t, \tag{5.4a}$$

$$\frac{d\psi(x)}{dx}=C\frac{d}{dx}\{|dQ/dx|^{-1/2}\exp(+|Q|)\}+D\frac{d}{dx}\{|dQ/dx|^{-1/2}\exp(-|Q|)\},\quad x<t. \tag{5.4b}$$

From (5.2), (5.3a), (5.3b), (5.4a), and (5.4b), with the *first-order* JWKB-expressions for S and Q, we obtain with the use of asymptotic formulas for the Airy functions the following approximate expressions for the coefficients in (5.3a) and (5.3b) and (5.4a) and (5.4b):

$$A=a\ \exp(-i\pi/4)+b\ \exp(+i\pi/4), \tag{5.5a}$$

$$B=a\ \exp(+i\pi/4)+b\ \exp(-i\pi/4), \tag{5.5b}$$

$$C=\frac{5a\ \exp(-2|Q|)}{144|Q|^2}+2b, \tag{5.6a}$$

$$D=a+\frac{b\ \exp(+2|Q|)}{27|Q|^2}. \tag{5.6b}$$

Inserting (5.5a) and (5.5b) into (5.3a) and (5.3b) and (5.6a) and (5.6b) into (5.4a) and (5.4b), we obtain approximately

$$\psi(x)=2b|dS/dx|^{-1/2}\cos(|S|+\pi/4)+2a|dS/dx|^{-1/2}\cos(|S|-\pi/4),\quad x>t, \tag{5.7a}$$

$$\psi(x)=\left(\frac{5a\ \exp(-2|Q|)}{144|Q|^2}+2b\right)|dQ/dx|^{-1/2}\ \exp(+|Q|)$$
$$+\left(a+\frac{b\ \exp(+2|Q|)}{27|Q|^2}\right)|dQ/dx|^{-1/2}\ \exp(-|Q|),\quad x<t, \tag{5.7b}$$

and corresponding formulas for $d\psi(x)/dx$; see (5.3b) and (5.4b). It is seen that (5.1a) agrees with (5.7a), and [since $\exp(-2Q)\ll 1$] that the coefficient of the exponentially large phase-integral function in (5.1b) is in satisfactory agreement with (5.7b) unless $|a/b|$ is too large. However, there is an essential disagreement between the coefficient of the exponentially small phase-integral function in (5.1b) and (5.7b) when $b\neq 0$. In the classically forbidden region the function $2\pi^{1/2}\ \mathrm{Bi}(-x)$ does not represent the *pure,* exponentially large phase-integral function but contains a strong admixture of the exponentially small phase-integral function, as displayed in (5.2) and (5.7b). This is the reason why the coefficient of the exponentially small phase-integral function in (5.7b) is very large, and thus differs strongly from the corresponding coefficient in (5.1b), when $b\neq 0$. The above discussion explains why (5.1a) and (5.1b), i.e., Silverstone's (Ref. 7) Eqs. (24) and (25), are not correct connection formulas.

VI. ONE-DIRECTIONAL NATURE OF THE CONNECTION FORMULAS

In every discussion of the connection problem for asymptotic approximations it is of crucial importance to adhere to the rigorous formulation of this problem: Given the asymptotic approximation of a wave function and its derivative at, or in the immediate neighborhood of, a given initial point (without any *further* knowledge of the wave function), what is the asymptotic approximation of the same wave function and its derivative in another region of the complex plane?

Although, as mentioned in Sec. I, the one-directional nature of the connection formulas has already long ago been proved and clarified in differing ways, we shall in this section demonstrate this property again but in an alternative way, whereby we adhere to Silverstone's[7] notations, but with the simplification that we put $\hbar=1$, and with the difference that we denote the turning point by t, while we let x_1 be a *given* point in the classically forbidden region. Thus we consider an *unspecified,* one-dimensional differential equation of the Schrödinger type with one simple classical turning point t on the real x-axis, but with no other transition point close to that axis, and we assume that to the right of t there is a classically allowed region extending to $+\infty$ and to the left of t there is a classically forbidden region extending to $-\infty$.

We shall first discuss the connection formula

$$|dQ/dx|^{-1/2}\ \exp(-|Q|)\rightarrow 2|dS/dx|^{-1/2}\ \cos(|S|-\pi/4), \tag{6.1}$$

and we consider the connection between $-\infty$ and $+\infty$. The *exact* solution of the differential equation that, as $x\rightarrow-\infty$, tends to the function in the left-hand member of (6.1) increases in the classically forbidden region in the direction toward the turning point, and as $x\rightarrow+\infty$ it tends to the function $\Omega|dS/dx|^{-1/2}\cos(|S|-\pi/4+\delta)$, where Ω and δ are constants, Ω being approximately equal to 2, and δ, although in general different from zero, being to its absolute value small compared to unity. When one makes the connection in the direction of the arrow in (6.1), one is thus allowed to put approximately $\Omega=2$ and $\delta=0$, which yields the connection formula (6.1). On the other hand, the *exact* solutions with the phase equal to, or close to, the phase of the cosine in the right-hand member of (6.1) as $x\rightarrow+\infty$, are in the classically forbidden region extremely sensitive to small deviations of the phase from $|S|-\pi/4+\delta$. In fact, the *exact* solution that tends to the function in the right-hand member of (6.1) as $x\rightarrow+\infty$, will tend to infinity as $x\rightarrow-\infty$, unless δ is *exactly* equal to zero, which occurs only in very exceptional situations. When $\delta\neq 0$, although small, the phase of the wave function in the right-hand member of (6.1) is not *exactly* equal to the phase that corresponds to the solution that decreases exponentially when x tends to $-\infty$, and, because of the sensitivity of the solution to small changes of the phase, there will in the classically forbidden region be an admixture of the function $|dQ/dx|^{-1/2}\exp(+|Q|)$, which tends to infinity as x tends to $-\infty$, and hence the connection formula (6.1) is one-directional. This is illustrated in Fig. 6 in N. Fröman (Ref. 5), which is based on numerical results.

For the particular case of the differential equation $d^2\psi/dx^2+x\psi=0$ one has the exceptional situation that $\delta=0$ for the special connection between $-\infty$ and $+\infty$, and therefore the connection formula (6.1) is bidirectional in that special case, but it is one-directional for connection between $-\infty$ and points in the classically allowed region lying at a *finite* distance from the turning point *t*. This can be verified directly by means of the asymptotic formulas for the Airy functions.

We shall now discuss the connection formula

$$|dS/dx|^{-1/2}\cos(|S|+\pi/4)\rightarrow|dQ/dx|^{-1/2}\exp(+|Q|). \tag{6.2}$$

Let $\psi_1(x)$ and $\psi_2(x)$ be the two real, linearly independent *exact* solutions of the differential equation that fulfill the conditions that

$$\psi_1(x) \text{ is represented by } |dQ/dx|^{-1/2}\exp(-|Q|) \text{ when } x\rightarrow-\infty, \tag{6.3}$$

$$\psi_2(x) \text{ is represented by } |dS/dx|^{-1/2}\cos(|S|+\pi/4) \text{ when } x\rightarrow+\infty. \tag{6.4}$$

We define complex conjugate, *x*-dependent coefficients $A_k(x)$ and $B_k(x)=A_k^*(x)$, where $k=1$ or $k=2$, such that in the classically allowed region we have *exactly*

$$\psi_k(x)=A_k|dS/dx|^{-1/2}\exp(+i|S|)+B_k|dS/dx|^{-1/2}\exp(-i|S|), \quad k=1,2, \tag{6.5a}$$

$$\frac{d\psi_k(x)}{dx}=A_k\frac{d}{dx}\{|dS/dx|^{-1/2}\exp(+i|S|)\}+B_k\frac{d}{dx}\{|dS/dx|^{-1/2}\exp(-i|S|)\}, \quad k=1,2, \tag{6.5b}$$

and we define real, *x*-dependent coefficients $C_k(x)$ and $D_k(x)$, where $k=1$ or $k=2$, such that in the classically forbidden region we have *exactly*

$$\psi_k(x)=C_k|dQ/dx|^{-1/2}\exp(+|Q|)+D_k|dQ/dx|^{-1/2}\exp(-|Q|), \quad k=1,2, \tag{6.6a}$$

$$\frac{d\psi_k(x)}{dx}=C_k\frac{d}{dx}\{|dQ/dx|^{-1/2}\exp(+|Q|)\}+D_k\frac{d}{dx}\{|dQ/dx|^{-1/2}\exp(-|Q|)\}, \quad k=1,2. \tag{6.6b}$$

The precise meaning of (6.3) and (6.4) is then that

$$C_1(-\infty)=0, \tag{6.3a$'$}$$

$$D_1(-\infty)=1, \tag{6.3b$'$}$$

$$A_2(+\infty)=\tfrac{1}{2}\exp(+i\pi/4), \tag{6.4a$'$}$$

$$B_2(+\infty)=\tfrac{1}{2}\exp(-i\pi/4). \tag{6.4b$'$}$$

It is known[3] that, apart from the immediate neighborhood of the turning point *t*, we have

$$A_1(x)\approx\exp(-i\pi/4) \quad \text{classically allowed region,} \tag{6.7a}$$

$$B_1(x)\approx\exp(+i\pi/4) \quad \text{classically allowed region,} \tag{6.7b}$$

$$A_2(x)\approx\tfrac{1}{2}\exp(+i\pi/4) \quad \text{classically allowed region,} \tag{6.7c}$$

$$B_2(x)\approx\tfrac{1}{2}\exp(-i\pi/4) \quad \text{classically allowed region,} \tag{6.7d}$$

$$|C_1(x)|\ll\exp[-2|Q(x)|] \quad \text{classically forbidden region,} \tag{6.8a}$$

$$D_1(x)\approx1 \quad \text{classically forbidden region,} \tag{6.8b}$$

$$C_2(x)\approx1 \quad \text{classically forbidden region,} \tag{6.8c}$$

$$|D_2(x)| \ll \exp[2|Q(x)|] \quad \text{classically forbidden region,} \tag{6.8d}$$

and that in general

$$|D_2(x)| \gg 1 \quad \text{classically forbidden region,} \tag{6.8e}$$

although $D_2(x)$ may change sign in the region considered, so that there may exist an *exceptional* point in the classically forbidden region where $D_2(x)$ is equal to zero. In that case the absolute value of $D_2(x)$ is at the most of the order of magnitude unity in a very small interval containing the exceptional point; see Eq. (44b) and Fig. 8 in N. Fröman (Ref. 5). From (6.5a), (6.5b) and (6.7a)–(6.7d) it follows that *approximately*

$$\psi_1(x) = 2|dS/dx|^{-1/2} \cos(|S| - \pi/4) \quad \text{classically allowed region,} \tag{6.9a}$$

$$\psi_2(x) = |dS/dx|^{-1/2} \cos(|S| + \pi/4) \quad \text{classically allowed region.} \tag{6.9b}$$

From (6.6a) and (6.6b) it follows that in the classically forbidden region

$$|dQ(x)/dx|^{-1/2} \exp[|Q(x)|] = \frac{D_2(x)\psi_1(x) - D_1(x)\psi_2(x)}{C_1(x)D_2(x) - C_2(x)D_1(x)}, \tag{6.10a}$$

$$\frac{d}{dx}\{|dQ(x)/dx|^{-1/2} \exp[|Q(x)|]\} = \frac{D_2(x)d\psi_1(x)/dx - D_1(x)d\psi_2(x)/dx}{C_1(x)D_2(x) - C_2(x)D_1(x)}, \tag{6.10b}$$

and with the aid of (6.10a) and (6.10b) we realize that the *exact* solution $\psi(x)$ of the differential equation that at the *fixed* point x_1 in the classically forbidden region fulfills the conditions

$$\psi(x_1) = \{|dQ/dx|^{-1/2} \exp(+|Q|)\}_{x=x_1}, \tag{6.11a}$$

$$\psi'(x_1) = \left\{\frac{d}{dx}[|dQ/dx|^{-1/2} \exp(+|Q|)]\right\}_{x=x_1}, \tag{6.11b}$$

is given by the *exact* formula

$$\psi(x) = \frac{D_2(x_1)\psi_1(x) - D_1(x_1)\psi_2(x)}{C_1(x_1)D_2(x_1) - C_2(x_1)D_1(x_1)}, \quad -\infty < x < +\infty, \tag{6.12}$$

which, with the aid of (6.8a)–(6.8d) can be written *approximately* as

$$\psi(x) = \psi_2(x) - D_2(x_1)\psi_1(x), \quad -\infty < x < +\infty. \tag{6.12'}$$

Using (6.9a) and (6.9b) we can in the classically allowed region write (6.12′) approximately as

$$\psi(x) = |dS(x)/dx|^{-1/2} \cos(|S(x)| + \pi/4) - 2D_2(x_1)|dS(x)/dx|^{-1/2} \cos(|S(x)| - \pi/4). \tag{6.13}$$

If x_1 does not lie very close to a possibly existing *exceptional* point where $D_2(x) = 0$, we have $|D_2(x_1)| \gg 1$ according to (6.8e), and (6.13) can *approximately* be written as

$$\psi(x) = 2|D_2(x_1)||dS(x)/dx|^{-1/2} \cos(|S(x)| + 3\pi/4) \quad \text{when} \quad D_2(x_1) > 0, \tag{6.14a}$$

$$\psi(x) = 2|D_2(x_1)||dS(x)/dx|^{-1/2} \cos(|S(x)| - \pi/4) \quad \text{when} \quad D_2(x_1) < 0. \tag{6.14b}$$

Under the assumption introduced immediately before (6.14a), the amplitude $2|D_2(x_1)|$ in (6.14a) and (6.14b) is large compared to unity, and for an unspecified differential equation we cannot give a general, approximate analytical expression for it. Comparing (6.14a) and (6.14b) with the left-hand member of the connection formula (6.2), we see that this connection formula is one-directional.

The above demonstration of the one-directional nature of the connection formula (6.2) can be summarized in a concrete way as follows. The solution $\psi_2(x)$ that is represented by the left-hand member of (6.2) in the classically allowed region, is in the classically forbidden region approximately given by $|dQ/dx|^{-1/2}\exp(+|Q|)$ with an admixture of $|dQ/dx|^{-1/2}\exp(-|Q|)$ that is insignificant, although the coefficient $D_2(x)$ of the latter phase-integral function in general is much larger than unity to its absolute value. The connection formula (6.2) can thus be used in the direction of the arrow. On the other hand, the solution $\psi(x)$ that at the *fixed* point x_1 in the classically forbidden region is represented by $|dQ/dx|^{-1/2}\exp(+|Q|)$ without any admixture of $|dQ/dx|^{-1/2}\exp(-|Q|)$, is approximately $\psi(x)=\psi_2(x)-D_2(x_1)\psi_1(x)$, and for this solution one obtains in the classically allowed region the approximate formula (6.13), from which one then obtains (6.14a) and (6.14b), unless $|D_2(x)|$ is accidentally at most of the order of magnitude unity. The connection formula (6.2) thus *can not* be used in the direction opposite to the arrow.

The analysis of the connection formula (6.2) performed above is an alternative to the analysis by N. Fröman[5] which was performed entirely within the framework of the *F*-matrix method.[3] The results of these alternative analyses are, of course, equivalent. Figure 8 in N. Fröman (Ref. 5) provides a numerical illustration of a case (involving a Coulomb potential) in which there exists an exceptional point such that, in accordance with (6.14a) and (6.14b), the phase changes abruptly from $3\pi/4$ to $-\pi/4$, and the amplitude goes through a narrow and deep minimum, as x_1 passes through the exceptional point. For another special case, viz. the differential equation $d^2\psi/dx^2+x\psi=0$, one can by straightforward, although somewhat tedious, analytical calculations, in which the asymptotic formulas for the Airy functions are used, show that the wave function $\psi(x)$ that at the *fixed* point x_1 in the classically forbidden region fulfills the initial conditions (6.11a) and (6.11b), is in the classically allowed region given by the approximate formula

$$\psi(x)=\frac{\exp(2|Q(x_1)|)}{27|Q(x_1)|^2}\,|dQ(x)/dx|^{-1/2}\cos[+|Q(x)|+3\pi/4]. \tag{6.15}$$

In this particular case there does not exist an exceptional point in the classically forbidden region, and (6.14a) is valid for every choice of x_1 (not too close to t) in the classically forbidden region.

VII. CONCLUSIONS

The Borel summation of the asymptotic series for the Airy functions, with *infinitely many terms retained*, yields *exact* relations which express the Airy functions in terms of the Whittaker function, and hence there remain no connection problems and no connection formulas after the Borel summation has been performed. Since the Borel summation is not the appropriate method for obtaining these exact relations, there appear false restrictions on $\arg z$ and an artificial cut that coincides with the Stokes line along the real z-axis; see Sec. III. One can derive these exact relations in a more appropriate, direct way with the aid of standard formulas available in handbooks, and then there appear no false restrictions on $\arg z$ and no artificial cut along the Stokes line; see Sec. IV.

On the other hand, when *truncated* asymptotic series are used, the Stokes phenomenon gives rise to connection problems, and the connection formulas for the handling of these problems are inherently one-directional; see Sec. VI.

When a wave function is well defined in the whole complex plane, and its asymptotic series is Borel summable, the accuracy of a truncated Borel sum will increase indefinitely as the number of terms is increased, and on a Stokes line, where the asymptotic series is not Borel summable, it is possible to calculate the wave function via analytic continuation of the result obtained in the region outside the Stokes line. This provides an efficient method for numerically exact calculation on the Stokes line of *a wave function that is well defined in the whole complex plane*,[11] but not of a wave function that is defined only approximately at a point in the classically allowed region, and thus it does not result in a connection formula. In particular, the fact that, by implementing the Borel procedure numerically, one can obtain the sum of an asymptotic series, representing a wave function that is well defined in the whole complex plane, to a much higher accuracy than is possible by partial summation of the original asymptotic series[11] is not relevant for the connection problem, formulated in the beginning of Sec. VI.

Across the Stokes line there is a cut *as regards the Borel summability*. However, the coefficient of the exponentially small term in the asymptotic series does not change discontinuously as z crosses the Stokes line, but is a continuous function of z. This follows in an obvious way from the continuity of the F-matrix in the method for solving connection problems developed by Fröman and Fröman.[3]

In every discussion of the connection formulas it is of primary importance to define clearly what one means by saying that a solution $\psi(x)$ of the differential equation in question at a given point x_0 is represented by a single phase-integral function, $f_1(x)$ or $f_2(x)$, respectively, without any admixture of the other phase-integral function, $f_2(x)$ or $f_1(x)$, respectively, and the only reasonable definition is obviously the requirement that $\psi(x_0)=f_k(x_0)$ and $\psi'(x_0)=f_k'(x_0)$, where $k=1$ or $k=2$, respectively. Adhering to this definition, we have considered the differential equation $d^2\psi/dx^2+x\psi=0$ and the first order of the JWKB-approximation and shown that Silverstone's (Ref. 7) ''correct connection formula,'' given by his Eqs. (24) and (25), is erroneous; see Sec. V.

Silverstone[7] has considered the Borel summation of the asymptotic series for the Airy functions, but not the connection problem for asymptotic approximations, formulated rigorously as in the beginning of Sec. VI. For points in the classically allowed and classically forbidden regions on opposite sides of a simple turning point, the solution of the connection problem is in a correct and satisfactory way given by the one-directional connection formulas (6.1) and (6.2).

As regards transmission through a potential barrier, also discussed by Silverstone,[7] a rigorous treatment by Fröman, Fröman, and Lundborg (Chap. 5 in Ref. 12), in which the barrier is considered as a whole, resulted in a new, very accurate connection formula for a barrier. In particular, this formula accounts in a completely correct way for terms in the transmission coefficient that for a superdense potential barrier are of the order of magnitude $\exp[-4|Q|]$ in Silverstone's[7] notation and with $\hbar=1$. Barrier penetration formulas obtained in treatments, where the connection of the wave function across each one of the two classical turning points of the barrier is treated separately, can sometimes be used, but there are important cases in which they are unsatisfactory, e.g., for energies in the neighborhood of the top of a barrier and for energies in the neighborhood of a resonance.

It is very important to make a clear distinction between the situation when one knows the wave function perfectly well in a certain region of the complex z-plane and the situation when one knows the wave function only approximately in that region. In the former situation one can obtain the wave function in another region be means of *bidirectional* analytic continuation (for instance with the use of Borel summation), while in the latter situation one has only *one-directional* connection formulas. The long dispute concerning the one-directional nature of the connection formulas seems in some cases to be due to an unclear distinction between these two situations.

We can very briefly summarize our conclusions as follows. When one uses Borel summation of infinite asymptotic series, one obtains exact formulas, and thus there remain no connection problems. To use the term connection formulas, when one is referring to exact solutions, is inadequate and gives rise to confusion. When one uses truncated asymptotic series, one is left with the Stokes phenomenon and the inherently one-directional connection formulas.

ACKNOWLEDGMENTS

The present investigation was initiated during the time 6 February–31 March 1995, when the authors were invited to participate in the program *Exponential Asymptotics* at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England. We wish to express our gratitude to the Isaac Newton Institute for a most fruitful time in an extremely stimulating scientific atmosphere. We are also much indebted to Professor Michael Berry for the invitation to participate in the program *Exponential Asymptotics*. We would also like to thank Professor Harris J. Silverstone, who participated in the same program, for interesting and valuable discussions.

[1] R. E. Langer, Bull. Am. Math. Soc. **40**, 545–582 (1934).
[2] H. Jeffreys, Proc. Cambridge Philos. Soc. **52**, 61–66 (1956).
[3] N. Fröman and P. O. Fröman, *JWKB Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965, Russian translation, MIR, Moscow, 1967).
[4] F. W. J. Olver, J. Res. Natl. Bur. Stand. Sect. B **69**, 271–290 (1965).
[5] N. Fröman, Ark. Fys. **31**, 381–408 (1966).
[6] F. W. J. Olver, *Asymptotics and Special Functions* (Academic, New York, London, 1974).

[7] H. J. Silverstone, Phys. Rev. Lett. **55**, 2523–2526 (1985).
[8] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972).
[9] Y. L. Luke, *The Special Functions and their Approximations* (Academic, New York, 1969), Vol. I.
[10] H. Buchholz, *The Confluent Hypergeometric Function*, Springer Tracts in Natural Philosophy, Vol. 15 (Springer, Berlin, 1969).
[11] M. Gailitis and H. J. Silverstone, Theor. Chim. Acta **73**, 105–114 (1988).
[12] N. Fröman and P. O. Fröman, *Phase-Integral Method Allowing Near Lying Transition Points*, Springer Tracts in Natural Philosophy, Vol. 40, edited by C. Truesdell (Springer, New York, 1996), with adjoined papers by A. Dzieciol, N. Fröman, P. O. Fröman, A. Hökback, S. Linnaeus, B. Lundborg, and E. Walles.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Derivative expansion of one-loop effective actions for Yang-Mills fields

T. D. Gargett[a] and I. N. McArthur[b]
Physics Department, The University of Western Australia, Nedlands, Western Australia 6907, Australia

(Received 31 December 1997; accepted for publication 24 April 1998)

Low energy effective actions are computed by integrating out degrees of freedom heavier than some energy scale of physical interest. At the one-loop level, the low energy effective action can be computed in terms of the functional trace of the heat kernel associated with the operator appearing in the part of the action quadratic in the heavy fields. Here, a technique for computing this functional trace is illustrated in the case of heavy scalar fields coupled to a Yang-Mills background. It makes use of the similarity of the functional trace of the heat kernel expressed in momentum space to a Gaussian integral. An explicit computation of the piece of the low energy effective action for the Yang-Mills fields quadratic in derivatives of the field strength is carried out, extending previously known results on the form of the long wavelength limit of the low energy effective action. © *1998 American Institute of Physics.* [S0022-2488(98)02209-9]

I. INTRODUCTION

The study of perturbative low energy effective actions has been very important for phenomenological applications of quantum field theory, where one is often interested in physics at an energy scale small compared with some of the mass scales in an underlying renormalizable quantum field theory. Although particles with masses larger than the energy scale of interest cannot be observed directly, they do contribute to interactions of the ''light'' degrees of freedom as virtual intermediate states. These contributions are summarized in the low energy effective action for the light degrees of freedom, which is obtained by summing Feynman diagrams with heavy degrees of freedom in loops.

In this paper, the case of massless gauge fields interacting with massive matter fields is considered. There has been much work on the computation of the one-loop low energy effective action for the gauge bosons in this case, with the principle focus being on the ''long wavelength limit.'' This is the lowest order term in an expansion of the effective action in derivatives of the field strength, and involves integrating out the massive matter fields in the presence of a *constant* gauge field background. Schwinger[1] used canonical methods to determine the complete form of the effective action in the long wavelength limit for the case of $U(1)$ gauge fields, reproducing earlier results of Heisenberg and Euler.[2] This result has since been rederived by a variety of other methods,[3–5] including extension to non-Abelian gauge groups. Here, the complete form of the first nontrivial correction to Schwinger's result will be computed, namely the term in the effective action involving two derivatives of the field strength.

This calculation is presented as an application of an approach to computing one-loop low energy effective actions developed by the authors in the context of supersymmetric Yang-Mills theories.[6] Using Schwinger proper time regularization, or the closely related zeta function regularization, the one-loop effective action can be computed in terms of the functional trace of the heat kernel for the operator in the piece of the original action quadratic in the quantum fields. The approach of Ref. 6 was to exploit the similarity of the functional trace of the heat kernel expressed in momentum space to a Gaussian integral, which allows Schwinger's result to be derived very efficiently. Here, we show that this method extends naturally to allow computation of the expan-

[a]Electronic mail: tgargett@physics.uwa.edu.au
[b]Electronic mail: mcarthur@physics.uwa.edu.au

 © 1998 American Institute of Physics

sion of the one-loop effective action in terms of covariant derivatives of the field strength of the background gauge fields. This is equivalent to the computation of corrections to the long wavelength limit of the effective action.

The plan of the paper is as follows. In the next section, we review the derivative expansion of the effective action. The technique introduced in Ref. 6 for computing the low energy effective action in the presence of covariantly constant background gauge fields is described in Sec. III, and its extension to nonconstant background gauge fields is discussed in Sec. IV. This is illustrated in Sec. V, where the terms in the effective action linear in the momenta of the gauge bosons are shown to vanish. The main results of the paper are presented in Sec. VI, where the full second order term in the derivative expansion of the effective action is computed. A short discussion follows. Details of computations have been relegated to appendices.

II. DERIVATIVE EXPANSION OF THE EFFECTIVE ACTION

For definiteness, the case of a free massive scalar field coupled to a Yang-Mills background is considered, although the technique readily generalizes to the inclusion of light scalars interacting with the heavy scalars, or to massive fields of different spin. In four-dimensional Euclidean space, the action is

$$S[\phi,A_\mu]=\int d^4x\ \phi^\dagger(-D_\mu D_\mu+m^2)\phi,$$

where D_μ is the covariant derivative in the representation R of the gauge group to which the scalar fields belong. (Since the metric is Euclidean, contracted space–time indices will be written with both indices lowered.) The heat kernel associated with the operator $(-D_\mu D_\mu+m^2)$ is $e^{-tm^2}K(x,x';t)$, where $K(x,x';t)$ is the heat kernel for massless scalars, defined formally as

$$K(x,x';t)=e^{tD_\mu D_\mu}\delta^{(4)}(x,x'). \tag{1}$$

The one-loop effective action for the gauge fields obtained by ''integrating out'' the heavy scalar fields is $\Gamma_{\rm eff}[A_\mu]=-\ln Z[A_\mu]$, with $Z[A_\mu]=\int[d\bar{\phi}][d\phi]e^{-S[\phi,A_\mu]}$. Using the zeta function regularization, $\Gamma_{\rm eff}[A_\mu]=-\zeta'(0)$, where the zeta function is

$$\zeta(s)=\frac{1}{\Gamma(s)}\int_0^\infty dt t^{s-1}e^{-tm^2/\mu^2}\ {\rm Tr}\int d^4x\ \lim_{x\to x'}K(x,x';t/\mu^2). \tag{2}$$

Here, Tr is a trace over gauge indices, and μ is a parameter with dimensions of mass introduced to make t dimensionless.

Thus the problem reduces to that of computing the functional trace

$$\int d^4x\ \lim_{x\to x'}K(x,x';t)\equiv\int d^4x\ K(t)$$

of the massless heat kernel. To do this, it is convenient to introduce a plane wave representation $\delta^{(4)}(x,x')=\int(d^4k/(2\pi)^4)\,e^{ik.(x-x')}$ for the delta function in expression (1) for the kernel. Then

$$K(t)=\lim_{x\to x'}\int\frac{d^4k}{(2\pi)^4}e^{ik.(x-x')}(e^{-ik.(x-x')}e^{tD_\mu D_\mu}e^{ik.(x-x')})=\int\frac{d^4k}{(2\pi)^4}e^{tX_\mu X_\mu}, \tag{3}$$

where

$$X_\mu=D_\mu+ik_\mu. \tag{4}$$

The effective action is a functional of the field strength $F_{\mu\nu}=[D_\mu,D_\nu]$ and its covariant derivatives as a result of gauge invariance (there is no anomaly for a scalar field). In the $U(1)$ case (or background electromagnetic field), the covariant derivatives are ordinary derivatives. The piece of the effective action independent of derivatives of the background field strength is its long

wavelength limit, corresponding to the limit of zero momentum gauge bosons. It is the first term in the expansion of the effective action in derivatives of the field strength, which translates in momentum space to an expansion in powers of momenta of the gauge bosons. Equivalently, this can be viewed as an expansion of the effective action in commutators of covariant derivatives, with the long wavelength limit being a truncation with only single commutators of the covariant derivatives, and corrections to the effective action corresponding to the inclusion of higher order commutators.

As emphasized by Avramidi,[5] the covariant generalization of this in the case of non-Abelian gauge groups is then an expansion of the effective action in commutators of the appropriate covariant derivatives

$$\Gamma_{\text{eff}}[A_\mu]=\int d^4x(\Phi(F)+(DF)\Psi(F)(DF)+\cdots)$$

(where space–time indices have been suppressed). The first term is the long wavelength limit of the effective action, and is computed by integrating out the scalar fields in the presence of a gauge field background with covariantly constant field strength. The aim of this paper is to demonstrate an algorithm for the computation of $K(t)$ in (3) which is well adapted to this expansion of the effective action in covariant derivatives of the background gauge field.

Note that although $K(t)$ as defined by (3) is x-dependent via the covariant derivatives, this will be suppressed for notational convenience. The quantities $X_\mu=D_\mu+ik_\mu\mathbf{1}$ are considered as matrices in the representation R of the Lie algebra of the gauge group.

III. COVARIANTLY CONSTANT BACKGROUND GAUGE FIELDS

We begin by outlining the technique developed in Ref. 6 for computing the functional trace of the heat kernel for the case of covariantly constant background gauge field strength, $D_\mu F_{\nu\rho}=0$. From (3), it follows that $K(t)$ satisfies the linear differential equation

$$\frac{dK(t)}{dt}=K_{\mu\mu}(t), \tag{5}$$

with the notation

$$K_{\mu_1\mu_2\cdots\mu_n}(t)=\int\frac{d^4k}{(2\pi)^4}X_{\mu_1}X_{\mu_2}\cdots X_{\mu_n}e^{tX.X}. \tag{6}$$

Here, $X.X=X_\mu X_\mu$ with X_μ defined by (4). The aim is then to determine $K_{\mu\nu}(t)$ in terms of $K(t)$ to allow this differential equation to be solved for $K(t)$, from which the effective action is computed via (2). The method for doing this introduced in Ref. 6 is to exploit the similarity of expressions of the form (6) to moments of an ordinary Gaussian. To compute

$$\int\frac{d^4k}{(2\pi)^4}k_{\mu_1}k_{\mu_2}\cdots k_{\mu_n}e^{-tk.k},$$

one can use the identity

$$0=\int\frac{d^4k}{(2\pi)^4}\frac{\partial}{\partial k^{\mu_n}}(k_{\mu_1}k_{\mu_2}\cdots k_{\mu_{n-1}}e^{-tk.k})$$

repeatedly to obtain an expression in terms of the ordinary Gaussian $\int[d^4k/(2\pi)^4]e^{-tk.k}$. Similarly, to compute $K_{\mu\nu}(t)$, we use the fact that

$$0=\int\frac{d^4k}{(2\pi)^4}\frac{\partial}{\partial k^{\nu}}(X_\mu e^{tX.X}),$$

as the factor e^{-tk^2} in the integrand ensures that the boundary terms vanish. However, this is now a very nontrivial statement, because the X_μ defined by (4) are not commuting objects; in fact, $[X_\mu, X_\nu] = F_{\mu\nu}$. Doing the derivative with respect to k^ν yields

$$0 = i\delta_{\mu\nu} K(t) + 2it \int \frac{d^4k}{(2\pi)^4} X_\mu \int_0^1 ds (e^{stX.X} X_\nu e^{-stX.X}) e^{tX.X}. \tag{7}$$

The quantity

$$\int_0^1 ds\ e^{stX.X} X_\nu e^{-stX.X} = \sum_{m=0}^{\infty} \frac{t^m}{(m+1)!} ad^{(m)}(X.X)(X_\nu) \tag{8}$$

is readily computed in the case where $F_{\mu\nu} = [X_\mu, X_\nu]$ is covariantly constant; one simply ignores $[X_\mu, [X_\nu, X_\rho]]$ and higher order commutators in the computation of the terms in the series. The result is

$$\int_0^1 ds e^{stX.X} X_\nu e^{-stX.X} = B_{\nu\rho}(t) X_\rho, \tag{9}$$

where

$$B_{\nu\rho}(t) = \sum_{m=0}^{\infty} \frac{(-2t)^m}{(m+1)!} (F^m)_{\nu\rho} = \left[\frac{1 - e^{-2tF}}{2tF} \right]_{\nu\rho}. \tag{10}$$

This expression contains two different matrix products, one over gauge indices and one over space–time indices. The gauge indices on $F_{\mu\nu}$ and on $B_{\mu\nu}(t)$ have been suppressed, but it should not be forgotten that like $K(t)$ itself, they are matrices in the representation R of the gauge group.

Substituting (10) and (9) into (7) and using the fact that the matrix $B_{\nu\rho}(t)$ can be moved through X_μ because the commutators $[X_\mu, F_{\nu\rho}]$ are being set to zero, one finds that

$$0 = \delta_{\mu\nu} K(t) + 2t B_{\nu\rho}(t) K_{\mu\rho}(t).$$

Multiplying both sides by the inverse matrix $B^{-1}_{\lambda\nu}(t)$ gives

$$K_{\mu\nu}(t) = \left[\frac{F}{e^{-2tF} - 1} \right]_{\nu\mu} K(t),$$

so that (5) becomes

$$\frac{dK(t)}{dt} = \mathrm{tr} \left[\frac{F}{e^{-2tF} - 1} \right] K(t), \tag{11}$$

where the trace tr is over space–time indices (but *not* over gauge indices; $K(t)$ is still a matrix with respect to the gauge indices). An essential point here is that because the power series (17) begins with **1**, the inverse matrix $B^{-1}_{\mu\nu}(t)$ does not involve inverse powers of $F_{\mu\nu}$; *so there is never a need to invert the matrix* $F_{\mu\nu}$. Remembering that $K(t)$ is a matrix and using $\mathrm{tr}(F/e^{-2tF} - 1) = \mathrm{tr}(Fe^{2tF}/1 - e^{2tF}) = -\frac{1}{2}\mathrm{tr}[C^{-1}(1 - e^{2tF})^{-1}(d/dt)(1 - e^{2tF})C]$, with C a matrix independent of t, the solution of the differential equation is easily seen to be

$$K(t) = \det[(1 - e^{2tF})C]^{-1/2}.$$

It remains to compute the integration constant C. This is done by requiring that this expression reduces to $\int d^4k/(2\pi)^4 e^{-tk^2} = (16\pi^2 t^2)^{-1}$ in the limit $F \to 0$, from which it follows that $C = 2\pi/F$. Using $\det A = (\det AA^T)^{1/2}$ and $F^T = -F$, this can be symmetrized to give

$$K(t)=\frac{1}{8\pi^2}\det\left[\frac{F^2}{\cosh(2tF)-1}\right]^{1/4}=\frac{1}{16\pi^2t^2}\det\left[\frac{tF}{\sinh\ tF}\right]^{1/2} \tag{12}$$

as obtained by many other authors.[3–5]

For later reference, we make the following point. If expression (12) for $K(t)$ is substituted into (2), it is not possible to perform the t integration in closed form to determine the zeta function and hence the low energy effective action. *Rather, one has to expand $K(t)$ as a power series in t and perform the t integration in (2) order by order in this expansion.* The term of order t^n in the expansion of $K(t)$ gives rise to a contribution to the effective action involving $(n+2)$ factors of the field strength, and so, in the case of a $U(1)$ gauge field, represents an effective vertex for the interaction of $(n+2)$ gauge bosons. So one has to evaluate the low energy effective action as a sum of effective vertices. This is no serious loss, as physically, it is the effective vertices which are of interest. As such, there is really no need to solve the differential equation (11) in closed form; *it suffices to solve for $K(t)$ as a power series expansion in t*. [Indeed, in the paper of Ritz and Delbourgo[7] on the effective n-photon interactions in the low energy effective action for scalar QED, the power series expansion they obtain for $\ln K(t)$ can be obtained by integrating Eq. (11),

$$\frac{dK(t)}{dt}=-\frac{1}{2t}\sum_{m=0}^{\infty}\frac{(-2t)^m}{m!}B_m tr(F^m)K(t),$$

where the definition $z/(e^z-1)=\Sigma_{m=0}^{\infty}(1/m!)\,B_m z^m$ of the Bernoulli numbers B_m is used.] This point will be important later when we examine corrections to the low energy effective action; we will obtain a differential equation for $K(t)$ which is difficult to solve in closed form but which is easily integrated term by term to determine the corrections to $K(t)$ as a power series in t. For later reference, the power series expansion of $K(t)$ in the long wavelength limit, which we denote by $K^{(0)}(t)$, begins

$$K^{(0)}(t)=\frac{1}{16\pi^2t^2}\left[1-\frac{t^2}{12}\text{tr}\ F^2+\frac{t^4}{360}\left(\text{tr}\ F^4+\frac{5}{4}(\text{tr}\ F^2)^2\right)+\cdots\right], \tag{13}$$

where the traces are over space–time indices but not gauge indices. It will be seen later that traces of odd powers of F arise only in corrections to the long wavelength limit, as they can be expressed in terms of covariant derivatives of the field strength.

IV. CORRECTIONS TO THE LONG WAVELENGTH LIMIT

The above procedure for establishing a linear differential equation satisfied by $K(t)$ is easily extended to the case where higher order commutators are included, corresponding to corrections to the long wavelength limit of the effective action, or higher order terms in the (covariant) derivative expansion of the full effective action. The cost is greater complexity. However, all the steps involved are purely combinatoric and algebraic, and as such, have the potential to be implemented using a symbolic computational language. It should be stressed that the procedure outlined below does not allow computation of the kernel itself, but only its functional trace; therefore it can be used to determine the effective action, but not Green's functions in the presence of the background gauge field.

Before explaining the method, it is necessary to introduce some terminology to be adopted, namely the *order* of the approximation to the effective action. The long wavelength limit will be termed *zeroth order* because it involves no covariant derivatives of the field strength $F_{\mu\nu}$. In terms of the computation in the previous section it means including commutators $[X_\mu,X_\nu]$, but setting all higher order commutators to zero. An *nth order* contribution to the effective action will mean one with n covariant derivatives of the field strength; they can all be acting on one field strength, they can all be acting on different field strengths, or anything in between, as all these possibilities are related using integration by parts in x-space in the effective action. In terms of the computation of $K(t)$, the nth order contribution contains n commutators of X_μ with the zeroth order commutators $[X_\nu,X_\rho]=F_{\nu\rho}$, either all acting one $[X_\nu,X_\rho]$, or all on different ones, or anything between.

The equation $dK(t)/dt = K_{\mu\mu}(t)$ is the starting point for an order n computation of $K(t)$, and requires the determination of $K_{\mu\nu}(t)$ to the required order. As above, this is achieved using the identity

$$0 = \int \frac{d^4k}{(2\pi)^4} \frac{\partial}{\partial k^\nu}(X_\mu e^{tX.X}) = i\delta_{\mu\nu}K(t) + 2it\int \frac{d^4k}{(2\pi)^4} X_\mu \left(\int_0^1 ds\ e^{stX.X} X_\nu e^{-stX.X} \right) e^{tX.X}. \tag{14}$$

The computation of the integral over s via the power series expansion (8) cannot be done exactly, and it is at this point that a truncation to a given order n must be made. The result of summing the power series with this truncation can be expressed in the form

$$\int_0^1 ds e^{-stX.X} X_\nu e^{stX.X} = A_\nu(t) + B_{\nu\rho}(t)X_\rho + C_{\nu\rho\sigma}(t)X_\rho X_\sigma + D_{\nu\rho\sigma\lambda}(t)X_\rho X_\sigma X_\lambda + \cdots. \tag{15}$$

The tensors $A_\nu(t), B_{\nu\rho}(t), \ldots$ are independent of X_μ by definition (but not of its commutators). The series on the right hand side will only have a finite number of terms depending on the order of the computation; in fact at order n there are terms with at most $(n+1)$ free X's. This is established by referring back to (8). If $ad^{(m)}(X.X)(X_\nu)$ contains p free X's, then $ad^{(m+1)}(X.X)(X_\nu)$ contains an additional free X only if it also contains an extra higher order commutator. Given that $ad^{(0)}(X.X)(X_\nu) = X_\nu$ and that there are at most n higher order commutators to the order we are considering, the required conclusion easily follows. The computation of the tensors $A_\nu(t), B_{\nu\rho}(t), \ldots$ to the required order is the only nontrivial step in the procedure, as will be seen in Secs. V and VI.

Substitution of the result (15) into (14) yields

$$\begin{aligned} 0 = {} & \frac{1}{2t}\delta_{\mu\nu}K(t) + A_\nu(t)K_\mu(t) + [X_\mu, A_\nu(t)]K(t) + B_{\nu\rho}(t)K_{\mu\rho}(t) \\ & + [X_\mu, B_{\nu\rho}(t)]K_\rho(t) + C_{\nu\rho\sigma}(t)K_{\mu\rho\sigma}(t) + [X_\mu, C_{\nu\rho\sigma}(t)]K_{\rho\sigma}(t) + \cdots. \end{aligned} \tag{16}$$

Now, as was seen in the previous section, $B_{\nu\rho}(t)$ contains a zeroth order contribution; this will be denoted $B^{(0)}_{\nu\rho}(t)$, and

$$B^{(0)}_{\nu\rho}(t) = \sum_{m=0}^{\infty} \frac{(-2t)^m}{(m+1)!}(F^m)_{\nu\rho} = \left[\frac{1 - e^{-2tF}}{2tF} \right]_{\nu\rho}. \tag{17}$$

Recall also that the power series begins with **1**, so that $B^{(0)}_{\nu\rho}(t)$ can be inverted without introducing negative powers of $F_{\mu\nu}$. So, denoting any higher order contribution to $B_{\nu\rho}(t)$ by $\tilde{B}_{\nu\rho}(t)$, one can rearrange Eq. (16) in the form:

$$\begin{aligned} K_{\mu\nu}(t) = {} & -(B^{(0)})^{-1}_{\nu\rho}(t)\{ (1/2t)\, K(t)\, \delta_{\rho\mu} + A_\rho(t)K_\mu(t) + [X_\mu, A_\rho(t)]K(t) + \tilde{B}_{\rho\sigma}(t)K_{\mu\sigma}(t) \\ & + [X_\mu, B_{\rho\sigma}(t)]K_\sigma(t) + C_{\rho\sigma\lambda}(t)K_{\mu\sigma\lambda}(t) + [X_\mu, C_{\rho\sigma\lambda}(t)]K_{\sigma\lambda}(t) + \cdots\}. \end{aligned} \tag{18}$$

We now show by an inductive type of argument that this equation allows the computation of $K_{\mu\nu}(t)$ in terms of $K(t)$ to order n. Since all the tensors $K_{\mu_1\cdots\mu_m}(t)$ are multiplied by objects of at least order 1, it is only necessary to compute these tensors to order $(n-1)$. They are computed in a similar manner to $K_{\mu\nu}(t)$ itself; start this time with the identity

$$0 = \int \frac{d^4k}{(2\pi)^4} \frac{\partial}{\partial k^{\mu_m}}(X_{\mu_1}\cdots X_{\mu_{m-1}} e^{tX.X}) \tag{19}$$

and repeat the steps which led to Eq. (18). To compute $K_{\mu_1\cdots\mu_m}(t)$ to order $(n-1)$ requires a knowledge of the tensors to at most order $(n-2)$ by the same argument as for $K_{\mu\nu}(t)$. Repeating this argument, one finds that it is only necessary to have knowledge of the tensors $K_{\mu_1\cdots\mu_m}(t)$ to order zero.

To complete the argument, we need to determine the tensors $K_{\mu_1\cdots\mu_m}(t)$ to order zero. Beginning with the identity (19), one obtains

$$0 = i\delta_{\mu_m\mu_1}K_{\mu_2\cdots\mu_{m-1}}(t) + \cdots + i\delta_{\mu_m\mu_{m-1}}K_{\mu_1\cdots\mu_{m-2}}(t) + 2it\int\frac{d^4k}{(2\pi)^4}X_{\mu_1}\cdots X_{\mu_{m-1}}\left(\int_0^1 ds e^{stX.X}X_{\mu_m}e^{-stX.X}\right)e^{tX.X}.$$

Substituting the zeroth order result for the integral over s and using the fact that $(B^{(0)})^{-1}_{\nu\rho}(t)$ commutes with X_μ at this order,

$$K_{\mu_1\cdots\mu_m}(t) = -\frac{1}{2t}(B^{(0)})^{-1}_{\mu_m\mu_1}(t)K_{\mu_2\cdots\mu_{m-1}}(t) - \frac{1}{2t}(B^{(0)})^{-1}_{\mu_m\mu_2}(t)K_{\mu_1\mu_3\cdots\mu_{m-1}}(t) - \cdots\frac{1}{2t}(B^{(0)})^{-1}_{\mu_m\mu_{m-1}}(t)K_{\mu_1\cdots\mu_{m-2}}(t).$$

Iterating this using the zeroth order expressions $K^{(0)}_\mu(t)=0$ (as easily established from the identity (19) in the case $m=1$) and

$$K_{\mu\nu}(t) = -\frac{1}{2t}(B^{(0)})^{-1}_{\nu\mu}(t)K(t),$$

it then follows that at zeroth order $K_{\mu_1\cdots\mu_{2m+1}}(t)$ vanishes and

$$K_{\mu_1\cdots\mu_{2m}}(t) = \frac{1}{(-2t)^m}\sum_{perm.}{}' (B^{(0)})^{-1}_{\mu_{i_1}\mu_{i_2}}(t)\cdots(B^{(0)})^{-1}_{\mu_{i_{2m-1}}\mu_{i_{2m}}}K(t), \tag{20}$$

where the prime on the sum indicates the sum over all permutations

$$\{\mu_{i_1},\mu_{i_2},\ldots,\mu_{i_{2m-1}},\mu_{i_{2m}}\}$$

of $\{\mu_1,\mu_2,\ldots,\mu_{2m}\}$ with $i_1>i_2$, $i_3>i_4,\ldots,i_{2m-1}>i_{2m}$. This then completes the argument that $K_{\mu\nu}(t)$ can be consistently computed in terms of $K(t)$ at a given order, thus allowing $K(t)$ to be found via solution of (5).

In the following sections, the above procedure for computing the functional trace of the kernel will be carried out at order 1 and order 2. Before doing so, we make the following comment. Except for the need to order indices because $(B^{(0)})^{-1}_{\mu\nu}(t)$ is not symmetric in μ and ν, the zeroth order expressions for $K_{\mu_1\cdots\mu_m}(t)$ have a ''Wick-type'' appearance, with the role of the free field propagator played by $(B^{(0)})^{-1}_{\mu\nu}(t)$. This analogy is entirely apt, as the Wick expansion for a free field can be derived (see, for example, Ref. 8) from the following identity which is the analog of (19):

$$0 = \int[d\phi]\frac{\delta}{\delta\phi(x_m)}(\phi(x_1)\phi(x_2)\cdots\phi(x_{m-1})e^{-S[\phi]}).$$

However, since the integral in $K(t)$ is an ordinary integral and not a functional integral, the tensors $K_{\mu_1\cdots\mu_m}(t)$ might better be considered as non-Abelian analogs of the usual Gaussian integrals $\int[d^4k/(2\pi)^4]k_{\mu_1}\cdots k_{\mu_m}e^{-tk^2}$, whose values, as pointed out earlier, can also be deduced from the identity

$$0=\int \frac{d^4k}{(2\pi)^4}\frac{\partial}{\partial k^{\mu_m}}(k_{\mu_1}\cdots k_{\mu_{m-1}}e^{-tk^2}).$$

V. THE FIRST ORDER CORRECTION

The first order correction to the effective action vanishes. This is because it is linear in the three index tensor $D_\mu F_{\nu\rho}$, and it is not possible to make a Lorentz invariant expression by contraction with $F_{\mu\nu}$'s. Nevertheless it is instructive to see how this arises in the procedure outlined above before proceeding to calculate the first nontrivial correction.

The only nontrivial step is the computation of

$$\int_0^1 ds e^{stX.X}X_\mu e^{-stX.X}=\sum_{m=0}^{\infty}\frac{t^m}{(m+1)!}ad^{(m)}(X.X)(X_\mu)$$

to first order, i.e., truncated to contain at most one higher order commutator of the form $[X_\mu,[X_\nu,X_\rho]]=(D_\mu F_{\nu\rho})$. The actual computation is described in the next section, and for the current purposes it is only necessary to note that the result is of the form

$$\int_0^1 ds e^{stX.X}X_\mu e^{-stX.X}=A^{(1)}_\mu(t)+B^{(0)}_{\mu\nu}(t)X_\nu+C^{(1)}_{\mu\nu\rho}(t)X_\nu X_\rho,$$

where $B^{(0)}_{\mu\nu}(t)$ is the zeroth order matrix (17), and $A^{(1)}_\mu(t)$ and $C^{(1)}_{\mu\nu\rho}(t)$ are first order, i.e., are linear in $(D_\mu F_{\nu\rho})$. So, substituting into (18), we find that to first order

$$K_{\mu\nu}(t)=-\frac{1}{2t}(B^{(0)})^{-1}_{\nu\mu}(t)K(t)-(B^{(0)})^{-1}_{\nu\rho}(t)A^{(1)}_\rho(t)K_\mu(t)-(B^{(0)})^{-1}_{\nu\rho}(t)C^{(1)}_{\rho\sigma\lambda}(t)K_{\mu\sigma\lambda}(t),$$

where $[X_\mu,A^{(1)}_\rho(t)]$ and $[X_\mu,C^{(1)}_{\rho\sigma\lambda}(t)]$ are omitted because they are of second order. Because their coefficients are first order, $K_\mu(t)$ and $K_{\mu\sigma\lambda}(t)$ are only required to zeroth order, and it was shown in the previous section that the tensors $K_{\mu_1\cdots\mu_{2m+1}}(t)$ all vanish at zeroth order. Thus there is no correction to the zeroth order result $K_{\mu\nu}(t)=-(1/2t)(B^{(0)})^{-1}_{\nu\mu}(t)K(t)$, as expected.

VI. THE SECOND ORDER CORRECTION

The first nontrivial correction to $K(t)$ contains terms which are quadratic in the tensor $(D_\mu F_{\nu\rho})$ and terms which are linear in $(D_\mu D_\nu F_{\rho\sigma})$. Both of these types of terms are required by consistency, as contributions of the second type are related in the effective action to those of the first type by integration by parts in x-space. As mentioned in Sec. IV, the only nontrivial step is the computation of $\int_0^1 ds e^{stX.X}X_\mu e^{-stX.X}$ in the form (15) to second order. This is done in Appendix A, where a diagrammatic technique is developed to aid in the manipulation of the third and fourth rank tensors involved. These diagrammatic techniques allow the determination of the individual terms in the power series (8). In the zeroth order case, this power series could be written compactly in the matrix (with respect to space–time indices) form (17). However, when higher rank tensors are present, this is not always possible. For example, expressions like

$$\sum_{p=0}^{m}\binom{n}{p}(F^p)_{\nu\lambda}(D_\mu F_{\lambda\sigma})(F^{n-p})_{\sigma\rho}$$

are encountered. There is no compact way to write it as a function of the matrix $F_{\mu\nu}$. To allow sums such as this to be written in a compact form, it is necessary to introduce an appropriate notation. First, note that to the order that we are working, the matrices $F_{\mu\nu}$ and $(D_\rho F_{\sigma\lambda})$, considered as matrices in the representation R of the gauge group, commute. This is because $F_{\mu\nu}=[X_\mu,X_\nu]$, so that using the Jacobi identity,

$$[F_{\mu\nu},(D_\rho F_{\sigma\lambda})]=-[[(D_\rho F_{\sigma\lambda}),X_\mu],X_\nu]-[[X_\nu,(D_\rho F_{\sigma\lambda})],X_\mu].$$

Since $(D_\rho F_{\sigma\lambda})$ is itself a commutator, the terms on the right hand side are seen to be third order and so can be dropped. Thus the sum can be rewritten

$$\sum_{p=0}^{m} \binom{n}{p} (-1)^{n-p} (F^p)_{\nu\lambda} (F^{n-p})_{\rho\sigma} (D_\mu F_{\lambda\sigma})$$

using the antisymmetry of $F_{\mu\nu}$ in its space–time indices. We write this as:

$$(F_2 - F_3)^n (D_1 F_{23}).$$

Here, F_2 and F_3 are shorthand for (antisymmetric) matrices in their space–time indices. The ''2'' and ''3'' on $(D_1 F_{23})$ mean that the right-hand indices on the space–time matrices F_2 and F_3 are contracted with space–time indices in the positions represented by ''2'' and ''3'' respectively on the three index tensor $(D_\mu F_{\nu\rho})$. The index ''1'' which is not repeated stands for a free space–time index. The only potential ambiguity is in the ordering of the matrices F_2 and F_3, this time considered as matrices in the representation R of the gauge group, when $(F_2+F_3)^n$ is expanded out. However, the commutator $[F_2, F_3]$ is second order, and since F_2 and F_3 multiply the first order object $(D_1 F_{23})$, their ordering is irrelevant. (In fact, using the Jacobi identity, $[F_{\mu\nu}, F_{\rho\sigma}] = -(D_\nu D_\mu F_{\rho\sigma}) - (D_\mu D_\nu F_{\rho\sigma})$. This means that care must be exercised in using the zeroth order results in Sec. III in the second order computation, as any reordering of the field strengths made in deriving zeroth order results will be invalid if used in the second order calculation. However, no such reorderings were made at zeroth order.) Alternatively, using the antisymmetry of $F_{\mu\nu}$, the expression can also be written as $(D_1 F_{23})(F_3 - F_2)^n$, where it is now the left-hand indices on the space–time matrices F_2 and F_3 which are contracted with $(D_1 F_{23})$. With a little practice, it is easy to become quite adept at translating back and forth between tensor expressions and this notation.

An alternative notation based on differential operators which retains the tensor structure explicitly can also be used. We can write, for example,

$$\sum_{p=0}^{m} \binom{n}{p} (F^p)_{\nu\lambda} (D_\mu F_{\lambda\sigma}) (F^{n-p})_{\sigma\rho} = \frac{d^n}{dz^n} [e^{zF}]_{\nu\lambda} (D_\mu F_{\lambda\sigma}) [e^{zF}]_{\sigma\rho}\big|_{z=0}. \tag{21}$$

The sums in Appendix A can also be done explicitly using this notation.

Returning to the computation of the right hand side of (15) to order 2, the result is

$$\int_0^1 ds e^{stX.X} X_\mu e^{-stX.X} = A_\mu^{(1)}(t) + (B_{\mu\nu}^{(0)}(t) + B_{\mu\nu}^{(2)}(t)) X_\nu + C_{\mu\nu\rho}^{(1)}(t) X_\nu X_\rho + D_{\mu\nu\rho\sigma}^{(2)}(t) X_\nu X_\rho X_\sigma, \tag{22}$$

where the superscript ''(n)'' denotes a contribution of order n to a tensor. The tensors $A_\mu^{(1)}(t)$, etc. are computed in Appendix A using the shorthand notation for tensors introduced above, and the results are given in Eqs. (A5), (A6), (A7) and (A8).

With these results, the rest of the procedure for determining the second order contribution $K^{(2)}(t)$ to $K(t)$ is purely algebraic, as outlined in Sec. IV. The steps are detailed in Appendix B. The result is that

$$K^{(2)}(t) = K^{(0)}(t) M^{(2)}(t), \tag{23}$$

where $M^{(2)}(t)$ satisfies the differential equation

$$\begin{aligned} \frac{dM^{(2)}(t)}{dt} = -(B^{(0)})^{-1}_{\mu\nu}(t) \{ & [X_\mu, A_\nu^{(1)}(t)] K^{(0)}(t) + A_\nu^{(1)}(t) K_\mu^{(1)}(t) \\ & + [X_\mu, B_{\nu\rho}^{(0)}(t)] K_\rho^{(1)}(t) + B_{\nu\rho}^{(2)}(t) K_{\mu\rho}^{(0)}(t) + [X_\mu, C_{\nu\rho\sigma}^{(1)}(t)] K_{\rho\sigma}^{(0)}(t) \\ & + C_{\nu\rho\sigma}^{(1)}(t) K_{\mu\rho\sigma}^{(1)}(t) + D_{\nu\rho\sigma\lambda}^{(2)}(t) K_{\mu\rho\sigma\lambda}^{(0)}(t) \} / K^{(0)}(t). \end{aligned} \tag{24}$$

Although the right hand side of Eq. (24) has been computed in closed form in Appendix B, we have been unable to integrate this equation in closed form. This is *not* because it is not integrable. The power series expansion (see below) of the right hand side begins at order t^2, and involves only positive powers of the field strength $F_{\mu\nu}$. This can be integrated to determine $M^{(2)}(t)$ to any required order in t. [The integration constant clearly vanishes, as $K^{(2)}(t)$ vanishes when DF is set to zero.] However, the complexity of the closed form expression is such that solving the differential Eq. (24) in closed form is a formidable task. As mentioned at the end of Sec. III, the inability to solve in closed form is not seen as a major difficulty in applying these results. In a given physical situation, one is usually interested in a particular effective n-point interaction for the gauge bosons. This requires finding the piece of the effective action with n powers of the field strength, and so, via (2), one is interested in the piece of the functional trace of the heat kernel with n powers of the field strength. A power series expansion of the functional trace of the heat kernel is therefore necessary anyway.

Power series expansion of the right hand side of (24) using the results derived in Appendices A and B is readily achieved order using Mathematica.[9] Although seven different tensor structures appear in the result, some are related by Jacobi identities, and only three of them are independent. The rules for converting the tensor structures into these three forms are given in Appendix C. Applying these, $K^{(2)}(t)$ is given by (23) with $K^{(0)}$ as in (13) and

$$M^{(2)}(t)=\sum_{p,q,r} c_1^{(p,q,r)}(t)(D_\alpha F_{\beta\gamma})(F^p)_{\gamma\rho}(F^q)_{\beta\alpha}(F^r)_{\nu\mu}(D_\mu F_{\nu\rho})+\sum_{p,q,r} c_2^{(p,q,r)}(t)(D_\alpha F_{\beta\gamma})$$
$$\times(F^p)_{\alpha\rho}(F^q)_{\beta\nu}(F^r)_{\gamma\mu}(D_\mu F_{\nu\rho})+\sum_{p,q} c_3^{(p,q)}(t)(D_\alpha D_\beta F_{\gamma\delta})(F^p)_{\gamma\alpha}(F^q)_{\delta\beta}. \qquad (25)$$

We present the expansions to order t^{11} (but stressing the fact that Appendix B contains all the ingredients to compute the right hand side of (24) in closed form and hence allowing the computation of $M^{(2)}(t)$ to any desired order):

$$\sum_{p,q,r} c_1^{(p,q,r)}(t)x^p y^q z^r$$
$$=\frac{t^3}{180}+\frac{t^5}{3780}(-2x^2-4z^2-18xz+51yz)+\frac{t^7}{56700}(3x^4+6z^4-110yz^3+3y^2z^2+26x^3z$$
$$+6x^2z^2-58x^2yz+20xz^3-26xyz^2)+\frac{t^9}{1871100}(-10x^6-20z^6-86x^5z+350yz^5-20y^2z^4$$
$$+131y^3z^3-20x^4z^2+186x^4yz-64x^3z^3+86x^3yz^2-20x^2z^4+284x^2yz^3-10x^2y^2z^2-64xz^5$$
$$+86xyz^4-64xy^2z^3)+\frac{t^{11}}{2554051500}(1382x^8+2764z^8+11864x^7z-47900yz^7+2764y^2z^6$$
$$-34510y^3z^5+1382y^4z^4+2764x^6z^2-25497x^6yz+8770x^5z^3-11864x^5yz^2+2764x^4z^4$$
$$-37890x^4yz^3+1382x^4y^2z^2+8484x^3z^5-11864x^3yz^4+8770x^3y^2z^3+2764x^2z^6$$
$$-37604x^2yz^5+2764x^2y^2z^4-14395x^2y^3z^3+8770xz^7-11864xyz^6+8484xy^2z^5$$
$$-8770xy^3z^4)+O(t^{13}),$$

$$\sum_{p,q,r} c_2^{(p,q,r)}(t)x^p y^q z^r$$

$$= \frac{2\,t^3}{45} + \frac{t^5}{3780}(-23x^2-10y^2-3xz+12yz)$$

$$+ \frac{t^7}{56700}(33z^4+14y^4+25y^2z^2-16y^3z-17yz^3+9xz^3-15xyz^2+12x^2z^2+3xy^2z)$$

$$+ \frac{t^9}{1871100}(-109z^6-46y^6+56yz^5-30xz^5+52y^5z-79y^4z^2+52y^3z^3-81y^2z^4+49xyz^4$$

$$-76x^2z^4-21xy^2z^3+47x^2yz^3-15x^3z^3+44xy^3z^2-34x^2y^2z^2-9xy^4z)$$

$$+ \frac{t^{11}}{2554051500}(15032z^8+6336y^8-7718yz^7+4146xz^7-6745xyz^6+2934xy^2z^5$$

$$-5964xy^3z^4+2631xy^4z^3-5900xy^5z^2+1212xy^6z-7144y^7z+10391x^2z^6-6410x^2yz^5$$

$$+9100x^2y^2z^4-5900x^2y^3z^3+4518x^2y^4z^2+10790y^6z^2+4146x^3z^5-6363x^3yz^4$$

$$+1515x^3y^2z^3-7048y^5z^3+4981x^4z^4+10647y^4z^4-7144y^3z^5+11125y^2z^6)+O(t^{13}),$$

$$\sum_{p,q} c_3^{(p,q)}(t)x^p z^q$$

$$= \frac{t^3}{60}(x-3z)+\frac{t^5}{1260}(-2x^3+4z^3-xz^2+5x^2z)-\frac{t^6}{810}xz^3+\frac{t^7}{56700}(9x^5-22x^4z+5x^3z^2$$

$$-15x^2z^3+4xz^4-17z^5)+\frac{t^8}{5670}xz^5+\frac{t^9}{1871100}(-30x^7+73x^6z-17x^5z^2+49x^4z^3-15x^3z^4$$

$$+47x^2z^5-13xz^6+56z^7)+\frac{t^{10}}{425250}(-9xz^7-2x^3z^5)+\frac{t^{11}}{2554051500}$$

$$\times(4146x^9-10078x^8z+2360x^7z^2-6745x^6z^3+2121x^5z^4-6363x^4z^5+2025x^3z^6$$

$$-6410x^2z^7+1786xz^8-7718z^9)+O(t^{12}).$$

As a check on these results, they can be compared with the known results for the heat kernel $K(t)$. In the limit $t\to 0$, the heat kernel has the asymptotic expansion

$$K(t)=\frac{1}{16\pi^2 t^2}\sum_{n=0}^{\infty} a_n t^n.$$

The coefficients a_n have been computed *exactly* for a small number of values of n by a variety of means: a_3 by Gilkey,[10] a_4 by van de Ven[12] and Avramidi,[11] a_5 by van de Ven[12] and Bel'kov *et al.*,[13] and very recently, a_6 by Fliegner *et al.*[14] and van de Ven.[15] The reduction of a_n to second order (which means neglecting terms with more than two covariant derivatives of the field strength) must coincide with coefficient of t^n in the expansion of $16\pi^2 t^2 K(t)$ to second order via (23) and the power series (13) and (25).

For example, a_3 as computed by Gilkey is

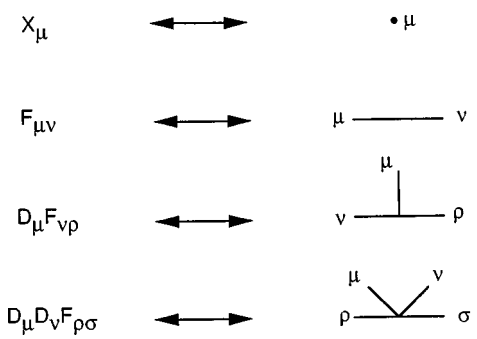

DIAGRAM 1. Representation of the basic tensor structures.

$$a_3=\tfrac{1}{180}(D_\alpha F_{\alpha\gamma})(D_\beta F_{\beta\gamma})+\tfrac{1}{45}(D_\alpha F_{\beta\gamma})(D_\alpha F_{\beta\gamma})-\tfrac{1}{30}F_{\alpha\beta}F_{\beta\gamma}F_{\gamma\alpha}+\tfrac{1}{30}(D_\alpha D_\alpha F_{\beta\gamma})F_{\beta\gamma}. \tag{26}$$

For comparison, the expansions (13) and (25) show that, *to second order*,

$$a_3=\tfrac{1}{180}(D_\alpha F_{\alpha\gamma})(D_\beta F_{\beta\gamma})+\tfrac{2}{45}(D_\alpha F_{\beta\gamma})(D_\gamma F_{\beta\alpha})-\tfrac{1}{20}(D_\alpha D_\beta F_{\alpha\gamma})F_{\gamma\beta}+\tfrac{1}{60}(D_\alpha D_\beta F_{\gamma\beta})F_{\gamma\alpha}. \tag{27}$$

The first terms of (26) and (27) are obviously the same, and the second terms are equivalent after the application of one of the Jacobi identities in Appendix C used to convert tensor structures to one of the three standard forms. The last two terms of (26) are also equivalent to the last two terms in (27); this requires recognition of the fact that despite appearances, $F_{\alpha\beta}F_{\beta\gamma}F_{\gamma\alpha}$ is actually a second order term, as it can be expressed as $\frac{1}{2}[F_{\alpha\beta},F_{\beta\gamma}]F_{\gamma\alpha}$. Application of the Jacobi identity

$$[F_{\alpha\beta},F_{\gamma\delta}]=(D_\alpha D_\beta F_{\gamma\delta})-(D_\beta D_\alpha F_{\gamma\delta})$$

then yields the required equivalence. Note that in the case of a_3, there are no terms higher than second order in the exact expression; this is no longer true for a_n for $n\geqslant 4$.

The coefficients of t^4 and t^5 in the second order expression (23) for $K(t)$ obtained in this paper agree with the reduction to second order of the a_4 and a_5 obtained in Refs. 12, 11, and 14. [These authors actually compute $\int d^4x\ a_n$, so that the expressions obtained for a_n are ambiguous up to total derivatives; this must be borne in mind when making comparison with the results of this paper (and in fact when comparing the exact results for a_n found by different authors).]

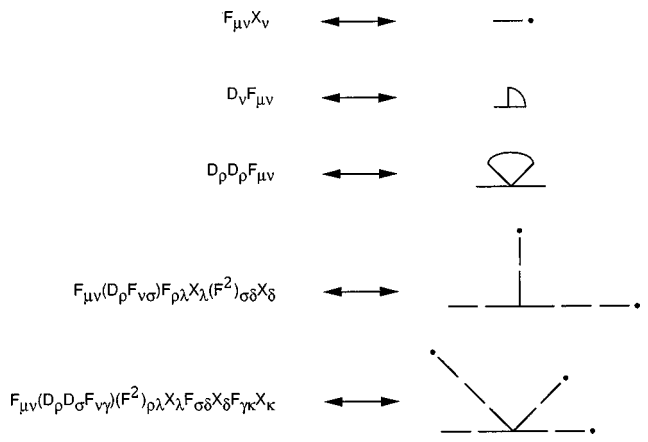

DIAGRAM 2. Examples of more complicated tensor structures.

This is a highly nontrivial check; for example, the expression for a_5 in Ref. 14 contains 36 terms. Comparison with the recently computed a_6 (Refs. 14 and 15) has not been attempted, as the exact expression in Ref. 14 contains 304 terms. Expressions for a_n with $n \geqslant 6$ have not been computed; the results of this paper are equivalent to the computation of a_n to second order for *all* n.

VII. CONCLUSION

The aim of this paper has been to show that the approach to computing the long wavelength limit of one-loop low energy effective actions developed in Ref. 6 is also well adapted to the problem of computing corrections to the long wavelength limit, corresponding to higher order terms in the derivative expansion of the effective action. This has been demonstrated by an explicit computation which yields the complete form for the first nontrivial correction to long wavelength limit of the low energy effective action when massive scalar fields are integrated out in the presence of a Yang-Mills background. The technique readily generalizes to the case where the heavy fields are of different spin, or where there are also light scalars coupled to the heavy fields.

Note added in proof: In a recent preprint (hep-th/9804143), V. P. Gusynin and I. A. Shovkovy present details of computations of higher derivative terms in the effective actions for scalar and spinor QED based on a different method, namely, the worldline (or string-inspired) formalism. This extends their previous work presented in Can. J. Phys. **74**, 282 (1996), where they computed the terms in the effective action for four-dimensional QED up to two derivatives in the field strength.

ACKNOWLEDGMENTS

Both authors are grateful to Professor J. Wess for his hospitality at the University of Munich, where this work was commenced. I.N.M. is also grateful to the Alexander von Humboldt-Stiftung for a Research Fellowship during that period.

APPENDIX A: EVALUATING COMMUTATORS TO SECOND ORDER

In this Appendix, details of the calculation of

$$\int_0^1 ds\; e^{stX.X} X_\mu e^{-stX.X} = \sum_{m=0}^{\infty} \frac{t^m}{(m+1)!} ad^{(m)}(X.X)(X_\mu) \tag{A1}$$

up to order two are presented. It was found convenient to introduce a diagrammatic method for generating the series of commutators $ad^{(m)}(X.X)(X_\mu)$. For a given m, the expression for $ad^{(m)}(X.X)(X_\mu)$ will consist of terms constructed from X_μ, $F_{\mu\nu}$, $(D_\mu F_{\nu\rho})$ and $(D_\mu D_\nu F_{\rho\sigma})$. So, to proceed to the $(m+1)$st commutator with $X.X$, it suffices to know the commutator of $X.X$ with these four quantities. The commutator with $(D_\mu D_\nu F_{\rho\sigma})$ produces terms of order 3 and so need not be considered. To order 2, the other commutators are:

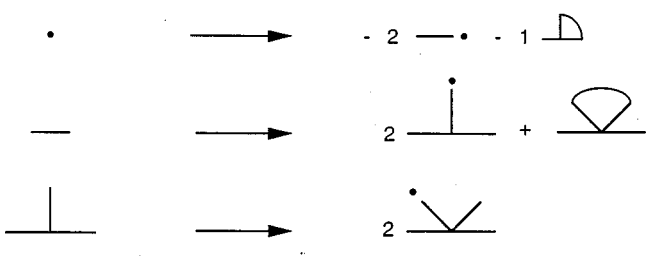

DIAGRAM 3. Rules for constructing commutators.

$$[X.X,X_\mu]=-2F_{\mu\nu}X_\nu-(D_\nu F_{\mu\nu}),$$

$$[X.X,F_{\mu\nu}]=2(D_\rho F_{\mu\nu})X_\rho+(D_\rho D_\rho F_{\mu\nu}), \tag{A2}$$

$$[X.X,D_\mu F_{\nu\rho}]=2(D_\sigma D_\mu F_{\nu\rho})X_\sigma.$$

The diagrammatic method consists of introducing symbols for the tensors X_μ, $F_{\mu\nu}$, $(D_\mu F_{\nu\rho})$ and $(D_\mu D_\nu F_{\rho\sigma})$ which can be assembled in a way which codes the contractions of their space–time indices. In this way, given the diagrams representing $ad^{(m)}(X.X)(X_\mu)$, the diagrams representing the $(m+1)$st commutator are generated by replacing the symbols in the diagrams one at a time by the combinations of symbols dictated by (A2).

The symbols for the tensors X_μ, $F_{\mu\nu}$, $(D_\mu F_{\nu\rho})$ and $(D_\mu D_\nu F_{\rho\sigma})$ are shown in Diagram 1. Contraction of indices is represented by joining the ends of the line segments on the diagrammatic elements appropriately: some examples are illustrated in Diagram 2. Diagram 3 shows the rules (A2) for going from the diagram for $ad^{(m)}(X.X)(X_\mu)$ to that for the $(m+1)$st commutator. The diagrams generated in this manner for $ad^{(m)}(X.X)(X_\mu)$ with $m=0,1,2$ and 3 are shown in Diagram 4. Using this approach, the contributions to $ad^{(m)}(X.X)(X_\mu)$ and their multiplicities are shown in Diagram 5. Although the combinatoric factor associated with a given diagram was initially established by inspection, it is relatively easy to prove by induction that they are correct.

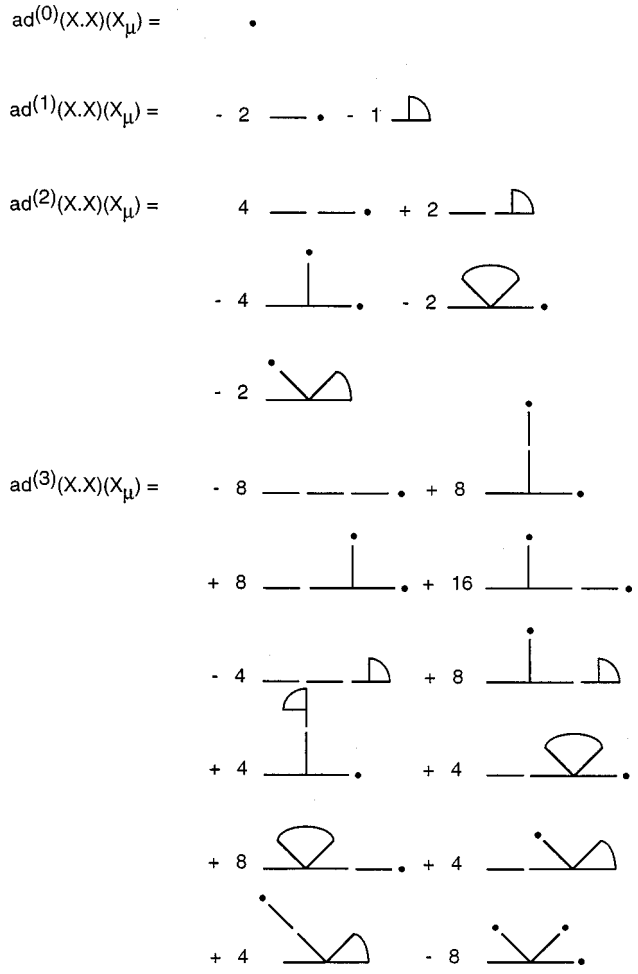

DIAGRAM 4. Representation of $ad^{(m)}(x\cdot x)(x_\mu)$ for $m=0$, 1, 2, 3.

Thus the tensors which appear in (22) corresponding to the expansion of the right hand side of (A1) to order 2 are:

$$A^{(1)}_{\mu}(t) = -\frac{1}{2}\sum_{n=1}^{\infty}\frac{(-2t)^n}{(n+1)!}(F^{n-1})_{\mu\nu}(D_\rho F_{\rho\nu}),$$

$$C^{(1)}_{\mu\nu\rho}(t) = -\sum_{n=2}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-2}\sum_{q=0}^{n-p-2}\binom{p+q+1}{q+1}(F^{n-p-q-2})_{\mu\sigma}(D_\lambda F_{\sigma\gamma})(F^q)_{\lambda\nu}(F^p)_{\gamma\rho},$$

$$\begin{aligned}
B^{(2)}_{\mu\nu}(t) = {} & \frac{1}{2}\sum_{n=3}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-3}\sum_{q=0}^{n-3-p}\left\{\binom{p+q+2}{q+2}(F^{n-3-p-q})_{\mu\rho}(D_\sigma F_{\rho\lambda})(F^q)_{\sigma\gamma}(D_\delta F_{\delta\gamma})(F^p)_{\lambda\nu}\right. \\
& \left.+\binom{p+q+2}{q+1}(F^{n-3-p-q})_{\mu\lambda}(D_\rho F_{\lambda\sigma})(F^q)_{\rho\nu}(F^p)_{\sigma\gamma}(D_\delta F_{\delta\gamma})\right\} \\
& -\frac{1}{2}\sum_{n=2}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-2}(F^{n-2-p})_{\mu\rho}(D_\sigma D_\gamma F_{\rho\gamma})(F^p)_{\sigma\nu}-\frac{1}{2}\sum_{n=2}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-2}(p+1) \\
& \times(F^{n-2-p})_{\mu\rho}(D_\gamma D_\gamma F_{\rho\sigma})(F^p)_{\sigma\nu}-\sum_{n=3}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=1}^{n-2}\sum_{q=0}^{n-p-2}\sum_{r=0}^{p-1}\binom{p+q+1}{q+1} \\
& \times(F^{n-p-q-2})_{\mu\sigma}(D_\lambda F_{\sigma\gamma})(F^q)_{\lambda\rho}(F^r)_{\gamma\alpha}(D_\rho F_{\alpha\beta})(F^{p-r-1})_{\beta\nu}
\end{aligned}$$

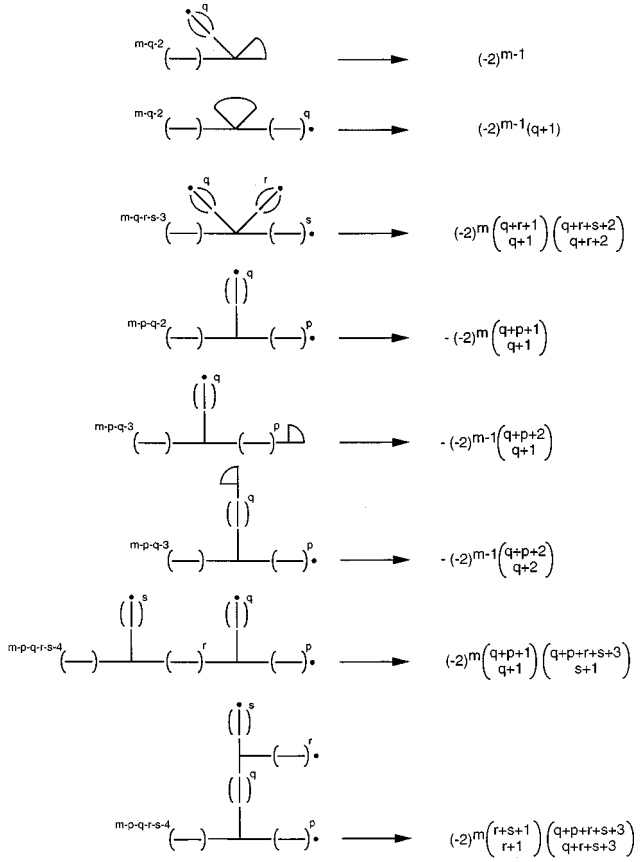

DIAGRAM 5. Contributions to $ad^{(m)}(x\cdot x)(x_\mu)$ and their multiplicities.

$$D^{(2)}_{\mu\nu\rho\sigma}(t)=\sum_{n=4}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-4}\sum_{q=0}^{n-4-p}\sum_{r=0}^{n-4-p-q}\sum_{s=0}^{n-4-p-q-r}\binom{r+s+1}{s+1}\Bigg\{\binom{n-p-1}{q+1}$$
$$\times(F^p)_{\mu\lambda}(D_\gamma F_{\lambda\delta})(F^q)_{\gamma\nu}(F^{n-p-q-r-s-4})_{\delta\alpha}(D_\beta F_{\alpha\epsilon})(F^s)_{\beta\rho}(F^r)_{\epsilon\sigma}+\binom{n-p-1}{q}$$
$$\times(F^p)_{\mu\lambda}(D_\gamma F_{\lambda\delta})(F^{n-4-p-q-r-s})_{\gamma\alpha}(D_\beta F_{\alpha\epsilon})(F^s)_{\beta\nu}(F^r)_{\epsilon\rho}(F^q)_{\delta\sigma}\Bigg\}$$
$$+\sum_{n=3}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-3}\sum_{q=0}^{n-3-p}\sum_{r=0}^{n-3-p-q}\binom{p+q+1}{p+1}\binom{p+q+r+2}{p+q+2}(F^{n-3-p-q-r})_{\mu\gamma}$$
$$\times(D_\delta D_\epsilon F_{\gamma\alpha})(F^p)_{\delta\nu}(F^q)_{\epsilon\rho}(F^r)_{\alpha\sigma}.$$

There is a subtlety associated with the fourth sum in the expression for $B^{(2)}_{\mu\nu}(t)X_\nu$. It actually arises from

$$-\sum_{n=2}^{\infty}\frac{(-2t)^n}{(n+1)!}\sum_{p=0}^{n-2}\sum_{q=0}^{n-p-2}\binom{p+q+1}{q+1}(F^{n-p-q-2})_{\mu\sigma}(D_\lambda F_{\sigma\gamma})(F^q)_{\lambda\nu}X_\nu(F^p)_{\gamma\rho}X_\rho, \quad \text{(A3)}$$

which, upon moving the X_ν to the right through $(F^p)_{\gamma\rho}$, gives $C^{(1)}_{\mu\nu\rho}(t)X_\nu X_\rho$. The commutators induced in this process give rise to a contribution to $B^{(2)}_{\mu\nu}(t)X_\nu$ which is contained in the fourth sum.

The sums in this expression can be carried out explicitly. Using the notation introduced in Sec. VI for tensor contractions, and introducing the quantities

$$E(F)=\frac{e^{2tF}-1}{F},$$

$$H_{p,q}(F_1,F_2,F_3)=\frac{(F_1+F_2)^p}{(F_1+F_2-F_3)}\left[\frac{E(F_1+F_2)}{(F_1+F_2)^q}-\frac{E(F_3)}{F_3^q}\right]-\frac{F_2^p}{(F_2-F_3)}\left[\frac{E(F_2)}{F_2^q}-\frac{E(F_3)}{F_3^q}\right], \quad \text{(A4)}$$

we find

$$A^{(1)}_1(t)=-\frac{1}{4tF_1}\{E(-F_1)-2t\}(D_2F_{21}), \quad \text{(A5)}$$

$$C^{(1)}_{123}(t)=-\frac{1}{2tF_2}H_{1,1}(F_2,F_3,-F_1)(D_2F_{13}), \quad \text{(A6)}$$

$$B^{(2)}_{12}(t)=\frac{1}{4t}(D_3F_{12})\frac{1}{F_3^2}\Bigg\{(F_2+F_3)H_{1,2}(-F_3,-F_2,F_1)+\frac{F_2F_3}{(F_1+F_2)^2}\left[\frac{E(-F_2)}{F_2}+\frac{E(F_1)}{F_1}\right]$$
$$+\frac{2F_3E(-F_2)}{F_2(F_1+F_2)}+\frac{2tF_3E(-F_2)}{(F_1+F_2)}-\frac{2tF_3}{F_2(F_1+F_2)}\Bigg\}(D_4F_{43})+\frac{1}{4t}(D_2F_{13})\frac{1}{F_2F_3}$$
$$\times\left\{\frac{2t}{F_1}-H_{2,2}(-F_3,-F_2,F_1)+\frac{F_3^2}{(F_1+F_3)}\left[\frac{E(F_1)}{F_1^2}-\frac{E(-F_3)}{F_3^2}\right]\right\}(D_4F_{43})$$
$$+\frac{1}{4t}(D_2D_3F_{13})\left\{\frac{1}{(F_1+F_2)}\left[-\frac{E(-F_2)}{F_2}-\frac{E(F_1)}{F_1}\right]+\frac{2t}{F_1F_2}\right\}+\frac{1}{4t}(D_3D_3F_{12})$$
$$\times\left\{\frac{1}{(F_1+F_2)^2}\left[-E(F_1)+\frac{(F_1+2F_2)E(-F_2)}{F_2}\right]+\frac{2tE(-F_2)}{(F_1+F_2)}-\frac{2t}{F_2(F_1+F_2)}\right\}$$
$$-\frac{1}{2t}(D_3F_{14})\frac{1}{F_3(F_2+F_4)}\{H_{1,1}(-F_3,F_2,F_1)-H_{1,1}(-F_3,-F_4,F_1)\}(D_3F_{42}), \quad \text{(A7)}$$

$$D^{(2)}_{1234}(t)=\frac{1}{2t}(D_2F_{15})\,\frac{1}{F_2F_3}\Bigg\{\frac{(F_3+F_4)}{(F_3+F_4+F_5)}\Bigg[\frac{1}{(-F_5)}H_{3,3}(-F_2,-F_5,F_1)-\frac{1}{(F_3+F_4)}H_{3,3}$$
$$(-F_2,F_3+F_4,F_1)\Bigg]+\frac{F_4}{(F_4+F_5)}\Bigg[\frac{1}{F_4}H_{3,3}(-F_2,F_4,F_1)$$
$$-\frac{1}{(-F_5)}H_{3,3}(-F_2,-F_5,F_1)\Bigg]\Bigg\}(D_3F_{54})+\frac{1}{2t}(D_5F_{14})$$
$$\times\Bigg\{\frac{(F_2+F_3)}{F_2(F_2+F_3+F_5)}\Bigg[\frac{1}{(F_2+F_3)^2}H_{3,3}(F_2+F_3,-F_4,F_1)-\frac{1}{F_5^2}H_{3,3}(-F_5,-F_4,F_1)\Bigg]$$
$$+\frac{F_3}{F_2(F_3+F_5)}\Bigg[\frac{1}{F_5^2}H_{3,3}(-F_5,-F_4,F_1)-\frac{1}{F_3^2}H_{3,3}(F_3,-F_4,F_1)\Bigg]\Bigg\}(D_2F_{53})$$
$$-\frac{1}{2t}(D_2D_3F_{14})\frac{1}{F_2}\Bigg\{\frac{1}{(F_2+F_3)}H_{2,2}(-F_2-F_3,-F_4,F_1)$$
$$-\frac{1}{F_3}H_{2,2}(-F_3,-F_4,F_1)\Bigg\}. \tag{A8}$$

It should be noted that in some of the above expressions, several different terms involving a particular $E(F_i)$ appear and could be combined into a single term. However, we have chosen to leave the expression in an unsimplified form involving the combinations $H_{p,q}(F_1,F_2,F_3)$ defined in (A4). This is because such a structure is responsible for many cancellations which occur to ensure that, as power series in t, these expressions involve no negative powers of the field strengths F.

APPENDIX B: *K(t)* TO SECOND ORDER

Here we give details of the determination of the second order contribution $K^{(2)}(t)$ to $K(t)$ using the procedure outlined in Sec. IV. As explained there, it is necessary to determine $K_{\mu\nu}(t)$ to second order via Eq. (18). Using the results $K^{(0)}_{\mu}(t)=0$ and $K^{(0)}_{\mu\nu\rho}(t)=0$ from Sec. IV and $K^{(1)}(t)=0$ from Sec. V, we find

$$K^{(2)}_{\mu\nu}(t)=-(B^{(0)})^{-1}_{\nu\rho}(t)\{(1/2t)\,\delta_{\rho\mu}K^{(2)}(t)+[X_\mu,A^{(1)}_\rho(t)]K^{(0)}(t)$$
$$+A^{(1)}_\rho(t)K^{(1)}_\mu(t)+[X_\mu,B^{(0)}_{\rho\sigma}(t)]K^{(1)}_\sigma(t)+B^{(2)}_{\rho\sigma}(t)K^{(0)}_{\mu\sigma}(t)$$
$$+[X_\mu,C^{(1)}_{\rho\sigma\lambda}(t)]K^{(1)}_{\sigma\lambda}(t)+C^{(1)}_{\rho\sigma\lambda}(t)K^{(1)}_{\mu\sigma\lambda}(t)+D^{(2)}_{\rho\sigma\lambda\delta}(t)K^{(0)}_{\mu\sigma\lambda\delta}(t)\}.$$

This result is to be substituted into the differential equation $dK^{(2)}(t)/dt=K^{(2)}_{\mu\mu}(t)$. By writing

$$K^{(2)}(t)=K^{(0)}(t)M^{(2)}(t),$$

the differential equation for $M^{(2)}(t)$ becomes:

$$\frac{dM^{(2)}(t)}{dt}=-(B^{(0)})^{-1}_{\mu\rho}(t)\{[X_\mu,A^{(1)}_\rho(t)]K^{(0)}(t)+A^{(1)}_\rho(t)K^{(1)}_\mu(t)$$
$$+[X_\mu,B^{(0)}_{\rho\sigma}(t)]K^{(1)}_\sigma(t)+B^{(2)}_{\rho\sigma}(t)K^{(0)}_{\mu\sigma}(t)+[X_\mu,C^{(1)}_{\rho\sigma\lambda}(t)]K^{(0)}_{\sigma\lambda}(t)$$
$$+C^{(1)}_{\rho\sigma\lambda}(t)K^{(1)}_{\mu\sigma\lambda}(t)+D^{(2)}_{\rho\sigma\lambda\delta}(t)K^{(0)}_{\mu\sigma\lambda\delta}(t)\}/K^{(0)}(t). \tag{B1}$$

It remains to determine $K^{(1)}_{\mu}(t)$ and $K^{(1)}_{\mu\nu\rho}(t)$ in terms of $K^{(0)}(t)$, as $K^{(0)}_{\mu\nu}(t)$ and $K^{(0)}_{\mu\nu\rho\sigma}(t)$ have already been expressed in terms of $K^{(0)}(t)$ in Eq. (20). As explained in Sec. IV, this is done via the generalization of the expression (18) to higher order tensors, and the results are:

$$\frac{K^{(1)}_{\mu}(t)}{K^{(0)}(t)} = -(B^{(0)})^{-1}_{\mu\nu}(t)\{A^{(1)}_{\nu}(t) + (1/2t)(B^{(0)})^{-1}_{\rho\sigma}(t)C^{(1)}_{\nu\rho\sigma}(t)\},$$

$$\begin{aligned}\frac{K^{(1)}_{\mu\nu\rho}(t)}{K^{(0)}(t)} = &-(1/2t)(B^{(0)})^{-1}_{\rho\sigma}(t)\{\delta_{\sigma\mu}K^{(1)}_{\nu}(t)/K^{(0)}(t) + \delta_{\sigma\nu}K^{(1)}_{\mu}(t)/K^{(0)}(t)\\ &-\{A^{(1)}_{\sigma}(t)(B^{(0)})^{-1}_{\nu\mu}(t) + [X_{\mu}, B^{(0)}_{\sigma\lambda}(t)](B^{(0)})^{-1}_{\lambda\nu}(t)\\ &+[X_{\nu}, B^{(0)}_{\sigma\lambda}(t)](B^{(0)})^{-1}_{\lambda\mu}(t)\} + 2tC^{(1)}_{\sigma\lambda\delta}(t)K^{(0)}_{\mu\nu\lambda\delta}(t)/K^{(0)}(t)\}.\end{aligned}$$

Assembling the above ingredients, the right hand side of the differential equation (B1) for $M^{(2)}(t)$ can be evaluated by inserting the expressions (A5), (A6), (A7) and (A8) for the tensors $A^{(1)}_{\mu}(t)$, etc. We have done this to obtain a closed form expression for the right hand side of (B1). For the sake of brevity, the expression is not reproduced here, but it can easily be derived using the above information. The integrated power series expansion of this expression is reproduced in Sec. VI to order t^{11}.

APPENDIX C: THE INDEPENDENT TENSOR STRUCTURES

In this Appendix, rules for manipulating the terms in the second order contribution to $K(t)$ into three independent tensor structures are given. The tensor structures which appear in $M^{(2)}(t)$ are of the following forms:

(i) $\lim_{\substack{2\to3\\4\to5}}(D_1F_{23})f(F_1,F_3,F_4)(D_4F_{51})$

(ii) $\lim_{\substack{2\to3\\4\to5}}(D_2F_{31})f(F_1,F_3,F_4)(D_4F_{51})$

(iii) $(D_1F_{23})f(F_1,F_2,F_3)(D_1F_{23})$

(iv) $(D_1F_{23})f(F_1,F_2,F_3)(D_3F_{21})$,

(v) $\lim_{\substack{1\to2\\3\to4}}(D_1D_2F_{34})f(F_2,F_4)$,

(vi) $\lim_{\substack{1\to3\\2\to4}}(D_1D_2F_{34})f(F_3,F_4)$,

where f denotes some functional of the field strengths (which is different for the different tensor structures).

These six tensor structures are not independent. Application of the Jacobi identity allows them all to be expressed in terms of the three tensor structures (ii), (iv) and (vi). The rules for putting the remaining tensor structures in these standard forms by use of the Jacobi identity are:

$$\lim_{\substack{2\to3\\4\to5}}(D_1F_{23})f(F_1,F_3,F_4)(D_4F_{51}) = \lim_{\substack{2\to3\\4\to5}}(D_2F_{31})[f(F_1,-F_3,F_4)-f(F_1,F_3,F_4)](D_4F_{51}),$$

$$(D_1F_{23})f(F_1,F_2,F_3)(D_1F_{23}) = (D_1F_{23})[f(F_1,F_2,F_3)+f(F_1,F_3,F_2)](D_3F_{21}),$$

$$\lim_{\substack{1\to2\\3\to4}}(D_1D_2F_{34})f(F_2,F_4) = \lim_{\substack{1\to3\\2\to4}}(D_1D_2F_{34})[f(F_3,F_4)-f(F_3,-F_4)].$$

The standard tensor structures also satisfy a number of identities as a result of the Jacobi identity and the ability to transpose expressions containing contracted space–time indices. These are

$$0=\lim_{\substack{2\to 3\\4\to 5}}(D_2F_{31})(f(F_1,F_3,F_4)-f(-F_1,F_4,F_3))(D_4F_{51}),$$

$$0=(D_1F_{23})(f(F_1,F_2,F_3)-f(-F_3,-F_2,-F_1))(D_3F_{21}),$$

$$0=(D_1F_{23})(f(F_1,F_2,F_3)-f(F_3,F_2,F_1)$$

$$-f(F_3,F_1,F_2)+f(F_1,F_3,F_2))(D_3F_{21}),$$

$$0=\lim_{\substack{1\to 3\\2\to 4}}(D_1D_2F_{34})(f(F_3,F_4)+f(F_4,F_3)+f(-F_3,-F_4)+f(-F_4,-F_3)).$$

These identities have been applied to reduce the power series (25) for $M^{(2)}(t)$ to its most economical form.

[1] J. Schwinger, Phys. Rev. **82**, 664 (1951).
[2] W. Heisenberg and H. Euler, Z. Phys. **98**, 714 (1936).
[3] M. Brown and M. Duff, Phys. Rev. D **11**, 2124 (1975).
[4] W. Dittrich and M. Reuter, *Effective Lagrangians in Quantum Electrodynamics*, Lecture Notes in Physics 220 (Springer, Berlin, 1985).
[5] I. G. Avramidi, Phys. Lett. B **305**, 27 (1993).
[6] I. N. McArthur and T. D. Gargett, Nucl. Phys. B **497**, 525 (1997).
[7] A. Ritz and R. Delbourgo, Int. J. Mod. Phys. A **11**, 253 (1996).
[8] R. J. Rivers, *Path Integral Methods in Quantum Field Theory* (Cambridge University Press, Cambridge, 1987).
[9] S. Wolfram, *Mathematica* (Cambridge University Press, Cambridge, 1996), 3rd ed.
[10] P. B. Gilkey, J. Diff. Geom. **10**, 601 (1975).
[11] I. G. Avramidi, Nucl. Phys. B **355**, 712 (1991).
[12] A. E. M. van de Ven, Nucl. Phys. B **250**, 593 (1985).
[13] A. A. Bel'kov, A. V. Lanyov, and A. Schaale, Comput. Phys. Commun. **95**, 123 (1996).
[14] D. Fliegner, P. Haberl, M. G. Schmidt, and C. Schubert, *The Higher Derivative Expansion of the Effective Action by the String Inspired Method Part II*, hep-th/9707189.
[15] A. van de Ven, *Index-free Heat Kernel Coefficients*, hep-th/9709152.

Mathematical derivation of chiral anomaly in lattice gauge theory with Wilson's action

Tetsuya Hattori[a)]
Department of Mathematics, Faculty of Science, Rikkyo University, Nishi-Ikebukuro, Tokyo 171, Japan

Hiroshi Watanabe[b)]
Department of Mathematics, Nippon Medical School, 2-297-2, Kosugi, Nakahara, Kawasaki 211, Japan

(Received 7 October 1997; accepted for publication 20 April 1998)

Chiral U(1) anomaly is derived with mathematical rigor for a Euclidean fermion coupled to a smooth external U(1) gauge field on an even dimensional torus as a continuum limit of lattice regularized fermion field theory with the Wilson term in the action. The present work rigorously proves for the first time that the Wilson term correctly reproduces the chiral anomaly. © *1998 American Institute of Physics.* [S0022-2488(98)03609-3]

I. INTRODUCTION

It is widely believed that the continuum limit of a lattice regularized theory of quantized fermion gives the correct chiral anomaly[1–3] if we have Wilson terms in the lattice fermion action.[4] Since Karsten and Smit[5] observed this fact by a perturbative argument, there appeared mathematically more careful analyses,[6,7] in which it was claimed that the Wilson fermion has the expected continuum limit and gives the correct chiral anomaly under some mathematical ansatz on perturbative expansions. However, a further investigation on the validity of the ansatz has not been published so far, and from a mathematical view point, a proof that Wilson's formalism for a lattice fermion is a correct scheme giving the expected anomaly, has not been completed.

On the other hand, in modern physics, the Wilson fermion provides not only a mathematical basis for analytic investigations but also a practical scheme for numerical studies on Euclidean quantum field theories including fermions. In view of this, we cannot help asking for a mathematically rigorous proof that the Wilson term correctly reproduces the chiral anomaly in the continuum limit.

In the present paper, we study Wilson's formalism for a Euclidean lattice fermion coupled to a smooth external U(1) gauge field defined on an even dimensional torus and derive the expected chiral anomaly in the continuum limit with mathematical rigor. The present work rigorously proves for the first time that the Wilson term correctly reproduces the chiral anomaly.

We summarize below three essential mathematical problems in the present study and the strategies to solve the problems.

(1) Chiral oscillation. Let $\mathbb{T}_a^d=(a\mathbb{Z}/L\mathbb{Z})^d$ be a discrete torus with period L and lattice spacing a, where L/a is a positive integer. We denote by $\mathbb{V}_a^d$ the vector space consisting of all functions $u:\mathbb{T}_a^d\to\mathbb{C}^{2^{d/2}}$ defined on $\mathbb{T}_a^d$ with values in the ''fiber'' $\mathbb{C}^{2^{d/2}}$. The vacuum expectation of divergence of chiral current, which is our main quantity of interest, is written in the form of $\mathrm{Tr}(K\gamma_{d+1})$, where K is an operator acting on $\mathbb{V}_a^d$, and γ_{d+1} is the chirality in d dimensions [see (3.1)]. [Note that γ_{d+1} is regarded as an operator acting on $\mathbb{V}_a^d$ instead of the fiber $\mathbb{C}^{2^{d/2}}$ and therefore the multiplicities of the eigenvalues ±1 are not uniformly bounded in the lattice spacing a.]

To take the trace, we choose, as in Refs. 6 and 7, the ''planewave basis,'' i.e., the set of eigenfunctions of the (free) translations on the lattice [see (4.4)]. Though this choice may be standard, it should be underlined because the Wilson term is too weak to make the *trace norm* of

a)Electronic mail: hattori@rkmath.rikkyo.ac.jp
b)Electronic mail: d34335@m-unix.cc.u-tokyo.ac.jp

 © 1998 American Institute of Physics

$K\gamma_{d+1}$ uniformly bounded in the lattice spacing a. In other words, it is essential to ''cancel'' [see, e.g., (6.9)] the highly degenerate positive and negative eigenvalues of γ_{d+1} (the ''chiral oscillation'') before taking the continuum limit. The planewave basis is convenient for this purpose.

Strategy 1: Take the trace with respect to the planewave basis in order to explicitly cancel the chiral oscillation.

It may be illustrative to compare the situation with a previous related work by one of the authors (H.W.).[8–10] There, the chiral oscillation is controlled by introducing an additional heat-kernel regularization, with which the operator in question lies in the trace class, and the continuum limit can be taken without explicit cancellation of chiral oscillation. As a result, the index of the continuum Dirac operator appears, which is equal to the Chern class by the index theorem,[11] hence the proof is completed.

(2) Perturbative expansion. After performing the fermion integration, we have a formula with inverse of operators [see (3.9)]. If we try to expand the inverse operators perturbatively, the convergence of the expansion becomes a problem, as is usual with formal perturbation series. For example, the operator C defined in (3.8) is decomposed as

$$C = C_0 + C_1 \tag{1.1}$$

into a *free part* C_0 (for vanishing gauge potentials) and an *interaction part* C_1, but the Neumann series

$$C^{-1} = \sum_{\ell=0}^{\infty} (-C_0^{-1}C_1)^{\ell} C_0^{-1} \tag{1.2}$$

is not absolutely convergent, unless the gauge field is sufficiently small. The following idea was suggested in Ref. 7 without details:

Strategy 2: Terminate a perturbative expansion of inverse operator up to a finite order, and use ''positivity'' to control the remainder terms.

Namely, we use an identity

$$C^{-1} = C_0^{-1} + \sum_{j=1}^{m-1} (-C_0^{-1}C_1)^j C_0^{-1} + (-C_0^{-1}C_1)^m C^{-1}, \quad m = 1,2,..., \tag{1.3}$$

for a sufficiently large m. What we need is the power-counting property of the operator $(-C_0^{-1}C_1)^j$, and an *a priori bound* arising from positivity of C, to control the remainder terms. See Secs. IV and V for details. It may be suggestive to note that remainder estimates based on *positivity* are used in the convergence proof of cluster expansions in constructive field theory.

(3) Power counting. Strategy 2 should be accompanied by a rigorous power-counting argument. Furthermore, we need asymptotic estimates of operators, when we take the continuum limit. Our last point is to employ the power-counting arguments with mathematical rigor in order to extract the asymptotic behaviors.

Strategy 3: Show and use the fact that the operators in question are ''quasidiagonal'' with respect to the planewave basis.

The class of *quasidiagonal operators* defined in Sec. IV is closed with respect to the summation and the product, and, under a simple assumption, it is also closed with respect to the inverse operation. This calculus, which is a systematic arrangement of the asymptotic analysis of lattice operators in the spirit of power counting, fits our analysis in this paper well. The smoothness assumption on the gauge fields is of relevance here. See Sec. IV for details.

According to the above strategies, we take the continuum limit of $\mathrm{Tr}(K\gamma_{d+1})$ and confirm that the Wilson fermion gives the correct chiral anomaly.

This paper is organized as follows. We describe the model and state the result (Theorem 2.1) in Sec. II. Theorem 2.1 is proved in the subsequent sections. In Sec. III, we derive the expression of chiral anomaly in the form of chiral oscillation. Section IV is devoted to the calculus of quasidiagonal operators. In Sec. V, we apply the framework of Sec. IV to several operators and determine their orders. The proof of Theorem 2.1 is completed in Sec. VI.

The propositions in Secs. III and VI may guide readers who wish to have an overview of our argument.

II. PROBLEM AND RESULT

In this section, we formulate the problem and then state the result in Theorem 2.1.

A. Notation

Let $\mathbb{T}_a^d$ be a discrete torus with lattice spacing a and period L. Throughout this paper, we assume that $0<a<1/2$, $L/a\in\mathbb{Z}$. Let $\mathbb{V}_a^d$ be the set of all mappings

$$u=(u_s)_{s=1,2,\ldots,2^{d/2}}:\quad \mathbb{T}_a^d\to\mathbb{C}^{2^{d/2}}. \tag{2.1}$$

We impose the antiperiodic boundary condition for definiteness, though boundary conditions play no essential roles in this paper. (The antiperiodic boundary condition is required in the proof of the reflection positivity.[12])

Let $(\ ,\ )$ be the inner product on $\mathbb{V}_a^d$ defined by

$$(v,u)=a^d\sum_{x\in\mathbb{T}_a^d}(v(x),u(x))_x, \tag{2.2}$$

where $(\ ,\ )_x$ is the (standard) inner product on $\mathbb{C}^{2^{d/2}}$:

$$(v(x),u(x))_x=\sum_{s=1}^{2^{d/2}}\overline{v_s}(x)u_s(x).$$

Let $A_\mu(x)$, $\mu=1,2,\ldots,d$, be a fixed smooth U(1) gauge field on the continuum torus $\mathbb{T}^d=(\mathbb{R}/L\mathbb{Z})^d$. Using A_μ, we assign the gauge group element

$$U_\mu(x)=\exp\left(-iQ\int_0^a A_\mu(x+se_\mu)ds\right) \tag{2.3}$$

to a directed link $(x,x+e_\mu)$, $x\in\mathbb{T}_a^d$, $\mu=1,2,\ldots,d$, where $Q\in\mathbb{R}$ is a charge and e_μ is a unit vector along the μth axis for $\mu=1,2,\ldots,d$. Denote the free and the covariant translations by $T_{\mu 0}$ and T_μ, respectively:

$$T_{\mu 0}u(x)=u(x+ae_\mu),\quad \mu=1,2,\ldots,d,\quad x\in\mathbb{T}_a^d,\quad u\in\mathbb{V}_a^d, \tag{2.4}$$

$$T_\mu=U_\mu T_{\mu 0},\quad \mu=1,2,\ldots,d, \tag{2.5}$$

where U_μ denotes the multiplication operator defined by

$$U_\mu u(x)=U_\mu(x)u(x),\quad \mu=1,2,\ldots,d,\quad x\in\mathbb{T}_a^d,\quad u\in\mathbb{V}_a^d. \tag{2.6}$$

Note that $T_\mu^*=T_\mu^{-1}$ holds, where T_μ^* is the adjoint with respect to the inner product $(\ ,\ )$. Put

$$D_\mu=\frac{1}{2a}(T_\mu-T_\mu^*),\quad \mu=1,2,\ldots,d, \tag{2.7}$$

and let γ_μ, $\mu=1,2,\ldots,d$, be the d dimensional anti-Hermitian Dirac matrices. Namely,

$$\gamma_\mu\gamma_\nu+\gamma_\nu\gamma_\mu=-2\delta_{\mu\nu}I_{2^{d/2}},\quad \mu,\nu=1,2,\ldots,d, \tag{2.8}$$

where I_ℓ denotes the unit matrix of order ℓ. In what follows, we abbreviate $\gamma_\mu\otimes\dot{I}$ to γ_μ, where $\dot{I}$ denotes the identity operator on the space $\dot{\mathbb{V}}_a^d=\{\varphi:\mathbb{T}_a^d\to\mathbb{C}\}$. Then, the lattice Dirac operator on the discrete torus $\mathbb{T}_a^d$ is by definition

$$\not{D} = \sum_{\mu=1}^{d} \gamma_\mu D_\mu . \tag{2.9}$$

B. Lattice fermion

As is well-known, the operator $\not{D}$ has eigenfunctions called *doublers* which do not approximate any eigenfunctions of the continuum Dirac operator. In order to suppress the contributions of doublers to vacuum expectation values, we introduce the Wilson term:

$$W = -\sum_{\mu=1}^{d} (2I - T_\mu - T_\mu^*), \tag{2.10}$$

where I denotes the identity operator on $\mathbb{V}_a^d$. Then our lattice fermion action is written as

$$S(\overline{\psi}, \psi) = \left(\overline{\psi}, \left(i\not{D} + MI - \frac{r}{2a} W \right) \psi \right), \tag{2.11}$$

where $M > 0$ is the fermion mass and r is a positive constant. (Our analysis allows any $r > 0$, but the proof of reflection positivity[12] requires $0 < r \leqslant 1$.) Using the action S, we define the vacuum expectation by

$$\langle \Phi(\overline{\psi}, \psi) \rangle = \frac{1}{Z(A)} \int \prod_{x \in \mathbb{T}_a^d} d\overline{\psi}(x) d\psi(x) \exp(-S(\overline{\psi}, \psi)) \Phi(\overline{\psi}, \psi), \tag{2.12}$$

where the integrations with respect to $\overline{\psi}$ and ψ are the Grassmann integrations[12,13] and the partition function $Z(A)$ is defined by

$$Z(A) = \int \prod_{x \in \mathbb{T}_a^d} d\overline{\psi}(x) d\psi(x) \exp(-S(\overline{\psi}, \psi)) = \det\left(i\not{D} + MI - \frac{r}{2a} W \right). \tag{2.13}$$

Note that the semipositivity of $-W$,

$$(u, -Wu) = \sum_{\mu=1}^{d} \| (I - T_\mu) u \|^2 \geqslant 0, \tag{2.14}$$

implies

$$\left(u, \left(i\not{D} + MI - \frac{r}{2a} W \right) u \right) \neq 0, \quad u \neq 0, \tag{2.15}$$

hence $Z(A) \neq 0$.

C. Lattice chiral current

We define the lattice chiral current by

$$J_\mu(x) = \tfrac{1}{2}(\overline{\psi}(x), \gamma_{d+1}\gamma_\mu (T_\mu \psi)(x))_x + \tfrac{1}{2}((T_\mu \overline{\psi})(x), \gamma_{d+1}\gamma_\mu \psi(x))_x, \quad \mu = 1,2,...,d, \tag{2.16}$$

where the chirality γ_{d+1} is by definition

$$\gamma_{d+1} = i^{d/2} \gamma_1 \gamma_2 \cdots \gamma_d . \tag{2.17}$$

Put

$$Y(x) = \frac{1}{a} \sum_{\mu=1}^{d} (J_\mu(x) - J_\mu(x - a e_\mu)) - 2M_i(\overline{\psi}(x), \gamma_{d+1}\psi(x))_x, \quad x \in \mathbb{T}_a^d, \tag{2.18}$$

and smear it as

$$Y(\xi)=a^d\sum_{x\in\mathbb{T}_a^d}\xi(x)Y(x) \tag{2.19}$$

by an arbitrary real-valued smooth function ξ defined on the continuum torus. The functional $Y(x)$ is the ''divergence'' of lattice chiral current with a mass correction. Our problem is to calculate $\lim_{a\to 0}\langle Y(\xi)\rangle$.

D. Result

Let us state our result.

Theorem 2.1: For an arbitrary smooth gauge field $A=(A_\mu)$ and for an arbitrary smooth function ξ both defined on the continuum torus $\mathbb{T}^d$, it holds that

$$\lim_{a\to 0}\langle Y(\xi)\rangle=\frac{-2iQ^{d/2}}{(4\pi)^{d/2}(d/2)!}\int_{\mathbb{T}^d}dx\xi(x)\sum_{\mu_1,\mu_2,\ldots,\mu_d=1}^{d}\epsilon_{\mu_1\mu_2\cdots\mu_d}F_{\mu_1\mu_2}(x)F_{\mu_3\mu_4}(x)\cdots F_{\mu_{d-1}\mu_d}(x), \tag{2.20}$$

where $\epsilon_{\mu_1\mu_2\cdots\mu_d}$ is the totally antisymmetric tensor and

$$F_{\mu\nu}(x)=\partial_\mu A_\nu(x)-\partial_\nu A_\mu(x). \tag{2.21}$$

Namely, $\langle Y(x)\rangle$ weakly converges to the Chern class in the continuum limit. In the subsequent sections, we prove this fact with mathematical rigor.

III. CHIRAL OSCILLATION

In this section, we carry out the Grassmann integrations, and write the chiral anomaly as a trace of an operator on $\mathbb{V}_a^d$.

Proposition 3.1: (Refs. 6 and 7) *It holds that*

$$\langle Y(\xi)\rangle=-2i\ \mathrm{Tr}_{\mathbb{V}_a^d}[\Lambda\xi\gamma_{d+1}], \tag{3.1}$$

where

$$\Lambda=\frac{1}{2}\left(i\not{D}+MI-\frac{r}{2a}W\right)^{-1}(i\not{D}+MI)+\frac{1}{2}(i\not{D}+MI)\left(i\not{D}+MI-\frac{r}{2a}W\right)^{-1}. \tag{3.2}$$

Proof: It is easy to see that

$$Y(\xi)=a^d\sum_{x\in\mathbb{T}_a^d}\xi(x)Y(x) \tag{3.3}$$

$$=(\overline{\psi},\xi\gamma_{d+1}i\not{D}\psi)+(i\not{D}\overline{\psi},\xi\gamma_{d+1}\psi)-2Mi(\overline{\psi},\xi\gamma_{d+1}\psi), \tag{3.4}$$

where in (3.4) we used the same symbol ξ for the multiplication operator given by $(\xi u)(x)=\xi(x)u(x)$, $x\in\mathbb{T}_a^d$. Using (3.4) and manipulating the Grassmann integrations, we obtain (3.1). Here, (2.15) ensures the existence of the inverse of $i\not{D}+MI-(r/2a)W$. □

Taking into account the behavior of the Wilson term $-(r/2a)W$ for doublers, one may intuitively see that the Λ works as a ''projection'' to decouple doublers, and according to this picture, one can write the scenario that when the doublers go out of the continuum world, they take away a part of lattice chiral current and produce the chiral unbalance in the continuum world. One may say that this is the origin of chiral anomaly and nonzero index of the Dirac operator.

Put

$$L=MI-\frac{r}{2a}W, \tag{3.5}$$

$$X_{\mu\nu}=D_\mu L^{-1}D_\nu-D_\nu L^{-1}D_\mu, \quad \mu,\nu=1,2,...,d, \tag{3.6}$$

$$B=\frac{1}{2}\sum_{\mu,\nu=1}^{d}\gamma_\mu\gamma_\nu X_{\mu\nu}, \tag{3.7}$$

$$C=L-\sum_{\mu=1}^{d}D_\mu L^{-1}D_\mu. \tag{3.8}$$

Proposition 3.2: (Refs. 6 and 7) *It holds that*

$$\langle Y(\xi)\rangle=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}(\xi W+W\xi)(C+B)^{-1}\gamma_{d+1}\right]. \tag{3.9}$$

Proof: Put

$$H_0=i\not{D}+MI, \tag{3.10}$$

$$H_1=i\not{D}+L, \tag{3.11}$$

and rewrite (3.1) as

$$\langle Y(\xi)\rangle=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}[\gamma_{d+1}(H_0H_1^{-1}+H_1^{-1}H_0)\xi]. \tag{3.12}$$

Since $H_0=H_1+(r/2a)W$ and $\mathrm{Tr}_{\mathbb{V}_a^d}[\xi\gamma_{d+1}]=0$, it holds that

$$\langle Y(\xi)\rangle=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\gamma_{d+1}\frac{r}{2a}(WH_1^{-1}+H_1^{-1}W)\xi\right]. \tag{3.13}$$

Note that

$$H_1L^{-1}H_1^*=L+\not{D}L^{-1}\not{D}=C+B \tag{3.14}$$

implies

$$H_1^{-1}=L^{-1}H_1^*(C+B)^{-1}. \tag{3.15}$$

Furthermore,

$$H_1\gamma_{d+1}=\gamma_{d+1}H_1^* \tag{3.16}$$

holds, while L, $\xi W+W\xi$, and $(C+B)^{-1}$ commute with γ_{d+1}. Then we have

$$\begin{aligned}\langle Y(\xi)\rangle&=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}(\xi W+W\xi)L^{-1}H_1^*(C+B)^{-1}\gamma_{d+1}\right]\\&=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}(\xi W+W\xi)L^{-1}\frac{H_1+H_1^*}{2}(C+B)^{-1}\gamma_{d+1}\right]\\&=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}(\xi W+W\xi)(C+B)^{-1}\gamma_{d+1}\right].\end{aligned} \tag{3.17}$$

□

IV. QUASIDIAGONAL OPERATORS

All the lattice operators appearing in this paper are ''quasidiagonal'' with respect to the plane wave basis. In this section, we define the class of quasidiagonal operators, give a few examples (Lemmas 4.1, 4.2), and prove basic properties (Lemmas 4.3, 4.4).

A. Definition of quasidiagonal operators

We first fix an orthonormal basis of $\mathbb{V}_a^d$. Put

$$\mathbb{T}_a^{d*}=\left\{p=(p_\mu)\in\mathbb{R}^d \middle| p_\mu=\frac{2k_\mu+1}{L}\pi, k_\mu\in\mathbb{Z}, |p_\mu|<\frac{2}{a}\,\pi\right\}. \tag{4.1}$$

We regard $\mathbb{T}_a^{d*}$ as a discrete torus with period $(2/a)\pi$ and spacing $2\pi/L$ and denote by $\langle p\rangle$ the periodic distance on $\mathbb{T}_a^{d*}$:

$$\langle p\rangle=\max_{\mu=1,2,\ldots,d}\langle p_\mu\rangle=\max_{\mu=1,2,\ldots,d}\min_{n\in\mathbb{Z}}\left|p_\mu-\frac{2n}{a}\,\pi\right|. \tag{4.2}$$

For $p\in\mathbb{T}_a^{d*}$, define the antiperiodic plane wave u_p on $\mathbb{T}_a^d$ with a momentum p by

$$u_p(x)=L^{-d/2}\exp(ipx),\quad x\in\mathbb{T}_a^d. \tag{4.3}$$

Furthermore, let χ_α, $\alpha=1,2,\ldots,2^{d/2}$, be the canonical orthonormal basis of $\mathbb{C}^{2^{d/2}}$ with respect to the standard inner product. Then,

$$u_{\alpha p}=\chi_\alpha\otimes u_p\in\mathbb{V}_a^d,\quad \alpha=1,2,\ldots,2^{d/2},\quad p\in\mathbb{T}_a^{d*}, \tag{4.4}$$

constitute an orthonormal basis of $\mathbb{V}_a^d$:

$$(u_{\alpha p},u_{\beta q})=\delta_{\alpha\beta}\delta_{pq},\quad \alpha,\beta=1,2,\ldots,2^{d/2},\quad p,q\in\mathbb{T}_a^{d*}, \tag{4.5}$$

where the inner product $(\ ,\ )$ is the one defined by (2.2).

Let us formulate the class of quasidiagonal operators. Operators such as T_μ or C acting on $\mathbb{V}_a^d$ have dependencies on the lattice spacing a, so that when we define such operators, we in fact define families of operators $T_\mu=T_\mu^{(a)}$, $a>0$, or $C=C^{(a)}$, $a>0$.

Definition: Let $k=k(a,p)$, $p\in\mathbb{T}_a^{d*}$, be a non-negative function satisfying

$$k(a,p)\leqslant c(1+\langle p-q\rangle)^\tau k(a,q),\quad p,q\in\mathbb{T}_a^{d*}, \tag{4.6}$$

for some constants c and τ independent of a, p, and q. Then, for a family of linear operators $K^{(a)}$ on $\mathbb{V}_a^d$, we say that $K^{(a)}$ is *a quasidiagonal operator of the order k* and write $K^{(a)}=\mathscr{O}(k)$, if

$$|(u_{\alpha p},K^{(a)}u_{\beta q})|\leqslant c_\sigma\frac{k(a,p)}{(1+\langle p-q\rangle)^\sigma}, \tag{4.7}$$

holds for all $\sigma\geqslant 0, p,q\in\mathbb{T}_a^{d*}$, and $\alpha,\beta=1,2,\ldots,2^{d/2}$. Here the constant c_σ may depend on σ but not on a, p, q, α, β.

Note that (4.6) and (4.7) imply

$$|(u_{\alpha p},K^{(a)}u_{\beta q})|\leqslant c'_\sigma\frac{k(a,q)}{(1+\langle p-q\rangle)^\sigma} \tag{4.8}$$

for another constant c'_σ. In other words,

$$K^{(a)}=\mathscr{O}(k)\Rightarrow K^{(a)*}=\mathscr{O}(k). \tag{4.9}$$

In the following, we often suppress writing a dependencies of operators explicitly and write K instead of $K^{(a)}$. Then we will simply say that K is quasidiagonal and write $K=\mathscr{O}(k)$.

B. Free parts and multiplication operators

Define the ''free parts'' of D_μ, W, L, and C by

$$D_{\mu 0}=\frac{1}{2a}\,(T_{\mu 0}-T_{\mu 0}^{*}), \tag{4.10}$$

$$W_0=-\sum_{\mu=1}^{d}\,(2-T_{\mu 0}-T_{\mu 0}^{*}), \tag{4.11}$$

$$L_0=MI-\frac{r}{2a}\,W_0, \tag{4.12}$$

$$C_0=L_0-\sum_{\mu=1}^{d}\,L_0^{-1}D_{\mu 0}^{2}, \tag{4.13}$$

respectively. A free part of an operator is equal to the operator when the gauge potentials A_μ, $\mu=1,2,...,d$, vanish. The free part of B vanishes. Since $-W_0$ is semipositive [see (2.14)], L_0 and C_0 are positive definite and therefore invertible.

These operators are *exactly diagonal* with respect to the planewave basis and are the first examples of quasidiagonal operators.

Lemma 4.1: We have the following.

$$T_{\mu 0}=\mathcal{O}(1), \tag{4.14}$$

$$T_{\mu 0}-I=\mathcal{O}(a(1+\langle p\rangle)), \tag{4.15}$$

$$D_{\mu 0}=\mathcal{O}(1+\langle p\rangle), \tag{4.16}$$

$$W_0=\mathcal{O}(a^2(1+\langle p\rangle^2)), \tag{4.17}$$

$$L_0=\mathcal{O}(1+a\langle p\rangle^2), \tag{4.18}$$

$$L_0^{-1}=\mathcal{O}\left(\frac{1}{1+a\langle p\rangle^2}\right), \tag{4.19}$$

$$C_0=\mathcal{O}\left(\frac{1+\langle p\rangle^2}{1+a\langle p\rangle^2}\right), \tag{4.20}$$

$$C_0^{-1}=\mathcal{O}\left(\frac{1+a\langle p\rangle^2}{1+\langle p\rangle^2}\right). \tag{4.21}$$

Proof: (4.14)–(4.17) follow from

$$T_{\mu 0}u_{\alpha p}=e^{iap_\mu}u_{\alpha p}, \tag{4.22}$$

$$D_{\mu 0}u_{\alpha p}=\frac{i}{a}\,\sin(ap_\mu)u_{\alpha p}, \tag{4.23}$$

$$W_{\mu 0}u_{\alpha p}=-2\sum_{\mu=1}^{d}\,(1-\cos(ap_\mu))u_{\alpha p}. \tag{4.24}$$

For example, if $K=T_{\mu 0}-I$, then (4.7) is satisfied by putting $k(a,p)=a\langle p\rangle$. We add an extra ''1'' in (4.15) in order to ensure (4.6): the function $k(a,p)=a(1+\langle p\rangle)$ satisfies (4.6), because

$$1+\langle q+r\rangle\leqslant 1+\langle q\rangle+\langle r\rangle\leqslant(1+\langle r\rangle)(1+\langle q\rangle),\quad q,r\in\mathbb{T}_a^{d*}. \tag{4.25}$$

For the case $K=W_0$, we need

$$1+\langle q+r\rangle^2 \leqslant 1+2(\langle q\rangle^2+\langle r\rangle^2) \leqslant 2(1+\langle r\rangle^2)(1+\langle q\rangle^2), \quad q,r\in\mathbb{T}_a^{d*}. \tag{4.26}$$

The other bounds are shown by using

$$L_0 u_{\alpha p}=\left(M+\frac{r}{2a}\sum_{\mu=1}^{d}(1-\cos ap_\mu)\right)u_{\alpha p}, \tag{4.27}$$

$$C_0 u_{\alpha p}=\left(M+\frac{r}{2a}\sum_{\mu=1}^{d}(1-\cos ap_\mu)+\frac{\Sigma_{\mu=1}^{d}(1/a^2)\sin^2 ap_\mu}{M+(r/2a)\,\Sigma_{\mu=1}^{d}(1-\cos ap_\mu)}\right)u_{\alpha p}. \tag{4.28}$$

□

A multiplication operator determined by a smooth function on $\mathbb{T}^d$ is another example of quasi diagonal operators.

Lemma 4.2: Let ϕ be a multiplication operator defined by

$$\phi u(x)=\varphi(x)u(x); \quad x\in\mathbb{T}_a^d, \quad u\in\mathbb{V}_a^d, \tag{4.29}$$

for a complex valued smooth (periodic) function $\varphi(x)$ on the continuum torus $\mathbb{T}^d$. Then, we have $\phi=\mathcal{O}(1)$.

Furthermore, $\phi=\mathcal{O}(1)$ holds, if ϕ is defined by

$$\phi u(x)=\varphi_a(x)u(x), \quad x\in\mathbb{T}_a^d, \quad u\in\mathbb{V}_a^d, \tag{4.30}$$

for a family of functions $\varphi_a(x)$ on $\mathbb{T}^d$ that are ''uniformly smooth,'' i.e.,

$$\sup_{x\in\mathbb{T}^d}|\partial^\alpha\varphi_a(x)|<c^{(\alpha)}, \quad \alpha=(\alpha_1,\ldots,\alpha_d)\in\{0,1,2,...\}^d, \tag{4.31}$$

where $\{c^{(\alpha)}|\alpha\in\{0,1,...\}^d\}$ is a family of constants independent of a, and ∂^α stands for $\Pi_{k=1}^d\partial_k^{\alpha_k}$.

Proof: We show the first part of the lemma. Since $\varphi(x)$ is smooth and satisfies

$$(u_{\alpha p},\phi u_{\beta q})=\delta_{\alpha\beta}a^d\sum_{x\in\mathbb{T}_a^d}e^{i(q-p)x}\varphi(x), \quad \alpha,\beta=1,2,...,2^{d/2}, \quad p,q\in\mathbb{T}_a^{d*}, \tag{4.32}$$

there exists c_σ such that

$$|(u_{\alpha p},\phi u_{\beta q})|\leqslant\frac{c_\sigma}{(1+\langle p-q\rangle)^\sigma}, \quad \sigma\geqslant 0 \tag{4.33}$$

for $p,q\in\mathbb{T}_a^{d*}$ and for $\alpha,\beta=1,2,...,2^{d/2}$. Then, $K_a=\phi$ and $k(a,p)=1$ satisfy (4.6) and (4.7). The other part of the lemma can be proved in a similar way. □

C. Basic properties

The class of quasidiagonal operators is closed with respect to summation, product, and the inverse operation in the following sense.

Lemma 4.3: Let $K_j=\mathcal{O}(k_j)$, $j=1,2$. Then, we have $K_1+K_2=\mathcal{O}(k_1+k_2)$ and $K_1K_2=\mathcal{O}(k_1k_2)$.

The proof is given in Sec. IV D. Note that, when we say $K=\mathcal{O}(k)$, it is implied that k has the property (4.6) with some constants c, τ.

*Lemma 4.4: Suppose that $K=K_0+K_1$ satisfies the following, where ''*const.*'' stands for some positive constant independent of a and p:*

(1) There exists k_0 satisfying

$$k_0(p)\leqslant\text{const.}(1+\langle p\rangle)^\nu, \quad p\in\mathbb{T}_a^{d*}, \tag{4.34}$$

for some constant ν independent of a and p, such that K_0 is exactly of the order k_0 in the sense that $K_0=\mathcal{O}(k_0)$ and $K_0^{-1}=\mathcal{O}(1/k_0)$ hold.

(2) K_1 is of a lower order than K_0 in the sense that $K_1=\mathcal{O}(k_1)$ and

$$\frac{k_1(p)}{k_0(p)}\leqslant \text{const.}(1+\langle p\rangle)^{-\delta}, \quad p\in \mathbb{T}_a^{d*}, \tag{4.35}$$

for some constant $\delta>0$ independent of a and p.

(3) K^{-1} exists and is uniformly bounded, i.e.,

$$\|K^{-1}\|=\sup_{\substack{u\in \mathbb{V}_a^d \\ u\neq 0}} \frac{\|K^{-1}u\|}{\|u\|}\leqslant c \tag{4.36}$$

for some constant c independent of a.

Then we have $K^{-1}=\mathcal{O}(1/k_0)$.

The proof is given in Sec. IV F.

By means of Lemmas 4.3 and 4.4, we can estimate traces on $\mathbb{V}_a^d$ using the order of operators. As may be seen explicitly from Lemma 4.1, such estimates provide a mathematical justification of the power counting argument.

D. Proof of Lemma 4.3

Let us first show $K_1+K_2=\mathcal{O}(k_1+k_2)$. Assume

$$k_1(a,p)\leqslant c_1(1+\langle p-q\rangle)^{\tau_1}k_1(a,q), \tag{4.37}$$

$$k_2(a,p)\leqslant c_2(1+\langle p-q\rangle)^{\tau_2}k_2(a,q), \tag{4.38}$$

$$|(u_{\alpha p},K_1u_{\beta q})|\leqslant c_{1\sigma}\frac{k_1(a,p)}{(1+\langle p-q\rangle)^{\sigma}}, \tag{4.39}$$

$$|(u_{\alpha p},K_2u_{\beta q})|\leqslant c_{2\sigma}\frac{k_2(a,p)}{(1+\langle p-q\rangle)^{\sigma}}, \tag{4.40}$$

for $\sigma\geqslant 0$, $p,q\in\mathbb{T}_a^{d*}$, and for $\alpha,\beta=1,2,...,2^{d/2}$. Then, putting $K=K_1+K_2$, $k=k_1+k_2$, $\tau=\max(\tau_1,\tau_2)$, $c=\max(c_1,c_2)$, and $c_\sigma=\max(c_{1\sigma},c_{2\sigma})$, we have (4.6) and (4.7), and consequently, $K_1+K_2=\mathcal{O}(k_1+k_2)$.

Let us show $K=K_1K_2=\mathcal{O}(k_1k_2)$. Put $k=k_1k_2$. Then, for $c=c_1c_2$ and for $\tau=\tau_1+\tau_2$, (4.6) holds. In order to show (4.7), we need the following lemma.

Lemma 4.5: If $0\leqslant\rho\leqslant\min\{\sigma,\tau\}$ and $\sigma+\tau-\rho\geqslant d+1$ hold, we have

$$\sum_{r\in\mathbb{T}_a^{d*}}(1+\langle p-r\rangle)^{-\sigma}(1+\langle r-q\rangle)^{-\tau}\leqslant\text{const.}(1+\langle p-q\rangle)^{-\rho}, \tag{4.41}$$

*where ''*const.*'' is independent of p, q, and a.*

Proof: Decompose the summation on the left-hand side of (4.41) as

$$\begin{aligned}&\sum_{r\in\mathbb{T}_a^{d*}}(1+\langle p-r\rangle)^{-\sigma}(1+\langle r-q\rangle)^{-\tau}\\ &\quad=\left(\sum_{\substack{r\in\mathbb{T}_a^{d*}\\ \langle p-r\rangle\geqslant\langle r-q\rangle}}+\sum_{\substack{r\in\mathbb{T}_a^{d*}\\ \langle p-r\rangle<\langle r-q\rangle}}\right)(1+\langle p-r\rangle)^{-\sigma}(1+\langle r-q\rangle)^{-\tau}.\end{aligned} \tag{4.42}$$

Let us estimate the first sum. If $\langle p-r\rangle\geqslant\langle r-q\rangle$, then it holds that

$$2\langle p-r\rangle \geqslant \langle p-r\rangle + \langle r-q\rangle \geqslant \langle p-q\rangle, \tag{4.43}$$

hence

$$\begin{aligned}
&\sum_{\substack{r\in \mathbb{T}_a^{d*}\\ \langle p-r\rangle \geqslant \langle r-q\rangle}} (1+\langle p-r\rangle)^{-\sigma}(1+\langle r-q\rangle)^{-\tau} \\
&\quad \leqslant \sum_{\substack{r\in \mathbb{T}_a^{d*}\\ \langle p-r\rangle \geqslant \langle r-q\rangle}} (1+\langle p-r\rangle)^{-\rho}(1+\langle p-r\rangle)^{-\sigma+\rho}(1+\langle r-q\rangle)^{-\tau} \\
&\quad \leqslant \sum_{r\in \mathbb{T}_a^{d*}} \left(1+\frac{1}{2}\langle p-q\rangle\right)^{-\rho}(1+\langle r-q\rangle)^{-\tau-\sigma+\rho} \leqslant \text{const.}\left(1+\frac{1}{2}\langle p-q\rangle\right)^{-\rho},
\end{aligned} \tag{4.44}$$

where we also used the assumptions $0\leqslant\rho\leqslant\sigma$ and $\tau+\sigma-\rho\geqslant d+1$. The second sum on the right-hand side of (4.42) can be estimated in a similar way. □

Using (4.39), (4.40), (4.38), and (4.41), we have

$$\begin{aligned}
|(u_{\alpha p}, K_1K_2u_{\beta q})| &\leqslant \sum_{\gamma=1}^{2^{d/2}} \sum_{r\in \mathbb{T}_a^{d*}} |(u_{\alpha p}, K_1u_{\gamma r})||(u_{\gamma r}, K_2u_{\beta q})| \\
&\leqslant \sum_{\gamma=1}^{2^{d/2}} \sum_{r\in \mathbb{T}_a^{d*}} \frac{c_{1\sigma}c_{2r}k_1(a,p)k_2(a,r)}{(1+\langle p-r\rangle)^{\sigma}(1+\langle r-q\rangle)^{\tau}} \\
&\leqslant \sum_{\gamma=1}^{2^{d/2}} \sum_{r\in \mathbb{T}_a^{d*}} \frac{c_{1\sigma}c_{2\tau}c_2k_1(a,p)k_2(a,p)}{(1+\langle p-r\rangle)^{\sigma-\tau_2}(1+\langle r-q\rangle)^{\tau}} \leqslant \text{const.}\ \frac{c_{1\sigma}c_{2\tau}c_2k_1(a,p)k_2(a,p)}{(1+\langle p-q\rangle)^{\rho}}
\end{aligned} \tag{4.45}$$

by choosing sufficiently large σ and τ for each ρ. This implies (4.7) for $K=K_1K_2$, hence a proof of Lemma 4.3 is completed.

E. Preparation for a proof of Lemma 4.4

In this section, we prepare Lemmas 4.6, 4.7, and 4.8 for the proof of Lemma 4.4.

The idea of the proof of Lemma 4.4 is to estimate the right-hand side of the identity

$$K^{-1} = \sum_{j=0}^{n-1} J^jK_0^{-1} + J^nK^{-1}, \quad n=1,2,..., \tag{4.46}$$

where

$$J = -K_0^{-1}K_1. \tag{4.47}$$

Lemma 4.3 implies that the first term on the right-hand side of (4.46) is quasidiagonal. For the second term, as we shall see later, J^nK^{-1} for a sufficiently large n has a good power-counting property in spite of the poor information (4.36) on K^{-1}. The key point is that the order of J is strictly lower than 1:

Lemma 4.6: It holds that

$$J = \mathcal{O}((1+\langle p\rangle)^{-\delta}). \tag{4.48}$$

Furthermore, for $n\geqslant \nu/\delta$,

$$J^n = \mathcal{O}(1/k_0) \tag{4.49}$$

holds.

Proof: The assumption (4.35) and Lemma 4.3 imply (4.48). Then we have

$$J^n = \mathcal{O}((1+\langle p\rangle)^{-n\delta}), \quad n = 1,2,\ldots . \tag{4.50}$$

In view of (4.34), we can bound the order of J^n as follows:

$$(1+\langle p\rangle)^{-n\delta} \leqslant \text{const.}\ \frac{(1+\langle p\rangle)^{\nu-n\delta}}{k_0(p)}. \tag{4.51}$$

Choosing $n \geqslant \nu/\delta$, we obtain (4.49). □

The above lemma implies that multiplying by J improves the power-counting properties of operators.

In order to determine the order of the second term on the right-hand side of (4.46), we have to bound

$$\langle p_\mu - q_\mu\rangle^\rho |(u_{\alpha p}, J^n K^{-1} u_{\beta q})| \tag{4.52}$$

for $\rho \geqslant 0, p, q \in \mathbb{T}_a^{d*}$, and for $\alpha, \beta = 1,2,\ldots,2^{d/2}$. To this end, we use commutators with $(1/a)T_{\mu 0}$. Let us write

$$\mathrm{ad}_\mu(R) = \left[\frac{1}{a}\, T_{\mu 0}, R\right] \tag{4.53}$$

for an operator R on $\mathbb{V}_a^d$.

Lemma 4.7: For an operator R on $\mathbb{V}_a^d$, it holds that

$$|(u_{\alpha p}, \mathrm{ad}_\mu(R) u_{\beta q})| \begin{cases} \leqslant \langle p_\mu - q_\mu\rangle |(u_{\alpha p}, R u_{\beta q})| \\ \geqslant \dfrac{1}{\pi}\, \langle p_\mu - q_\mu\rangle |(u_{\alpha p}, R u_{\beta q})|, \end{cases}$$

$$p, q \in \mathbb{T}_a^{d*}, \quad \alpha, \beta = 1,2,\ldots,2^{d/2}, \quad \mu = 1,2,\ldots,d. \tag{4.54}$$

Furthermore, if $R = \mathcal{O}(r)$, then $\mathrm{ad}_\mu(R) = \mathcal{O}(r)$.

Proof: The bounds (4.54) follow from

$$\begin{aligned} |(u_{\alpha p}, \mathrm{ad}_\mu(R) u_{\beta q})| &= \frac{1}{a}\, |(\exp(i p_\mu a) - \exp(i q_\mu a))| |(u_{\alpha p}, R u_{\beta q})| \\ &= \frac{2}{a} \left| \sin\left(\frac{1}{2}\,(p_\mu - q_\mu)a\right)\right| |(u_{\alpha p}, R u_{\beta q})|. \end{aligned} \tag{4.55}$$

The estimate $\mathrm{ad}_\mu(R) = \mathcal{O}(r)$ follows from (4.54) and (4.7). □

Lemma 4.8. (1) For $R = \mathcal{O}(r)$, we have

$$|(u_{\alpha p}, R K^{-1} u_{\beta q})| \leqslant \text{const.}\ r(p), \quad p, q \in \mathbb{T}_a^{d*}, \quad \alpha, \beta = 1,2,\ldots,2^{d/2}, \tag{4.56}$$

where ''const.'' *is a constant independent of a, p, and q.*

(2) Assume that $R = \mathcal{O}(r)$ is at most of the same order as K, i.e.,

$$r(p) \leqslant k_0(p), \quad p \in \mathbb{T}_a^{d*}. \tag{4.57}$$

Then, we have

$$\sum_{s' \in \mathbb{T}_a^{d*}} |(u_{\alpha s'}, RK^{-1}u_{\beta s})| \leqslant \text{const.}, \quad s \in \mathbb{T}_a^{d*}, \quad \alpha, \beta = 1,2,...,2^{d/2}, \tag{4.58}$$

where ''const.'' *is a constant independent of a and s.*

Proof: (1) (4.7) and (4.36) imply

$$|(u_{\alpha p}, Ru_{\gamma s})| \leqslant \frac{\text{const. } r(p)}{(1+\langle p-s\rangle)^{d+1}}, \tag{4.59}$$

$$|(u_{\gamma s}, K^{-1}u_{\beta p})| \leqslant c, \tag{4.60}$$

for $p, s \in \mathbb{T}_a^{d*}$ and for $\alpha, \beta, \gamma = 1,2,...,2^{d/2}$. Then, we have

$$\begin{aligned} |(u_{\alpha p}, RK^{-1}u_{\beta q})| &\leqslant \sum_{s \in \mathbb{T}_a^{d*}} |(u_{\alpha p}, Ru_{\gamma s})||(u_{\gamma s}, K^{-1}u_{\beta q})| \\ &\leqslant \sum_{s \in \mathbb{T}_a^{d*}} \frac{\text{const. } r(p)}{(1+\langle p-s\rangle)^{d+1}} \leqslant \text{const. } r(p), \end{aligned} \tag{4.61}$$

where we used

$$\sum_{s \in \mathbb{T}_a^{d*}} (1+\langle s\rangle)^{-d-1} \leqslant \text{const.} \tag{4.62}$$

(2) Using (4.46), we have

$$\begin{aligned} \sum_{s' \in \mathbb{T}_a^{d*}} |(u_{\alpha s'}, RK^{-1}u_{\beta s})| &\leqslant \sum_{s' \in \mathbb{T}_a^{d*}} \sum_{j=0}^{n-1} |(u_{\alpha s'}, RJ^j K_0^{-1} u_{\beta s})| \\ &\quad + \sum_{s' \in \mathbb{T}_a^{d*}} |(u_{\alpha, s'}, RJ^n K^{-1} u_{\beta s})|. \end{aligned} \tag{4.63}$$

Let us bound the right-hand side of the above inequality. For the first term, (4.48) implies

$$J = \mathcal{O}(1), \tag{4.64}$$

hence

$$RJ^j K_0^{-1} = \mathcal{O}\left(\frac{r}{k_0}\right) \tag{4.65}$$

holds. Then, from (4.65), (4.8), and (4.57), we obtain

$$\begin{aligned} \sum_{s' \in \mathbb{T}_a^{d*}} \sum_{j=0}^{n-1} |(u_{\alpha s'}, RJ^j K_0^{-1} u_{\beta s})| &\leqslant \text{const. } \frac{r(s)}{k_0(s)} \sum_{s' \in \mathbb{T}_a^{d*}} (1+\langle s'-s\rangle)^{-d-1} \\ &\leqslant \text{const.}, \quad s \in \mathbb{T}_a^{d*}, \end{aligned} \tag{4.66}$$

for a ''const.'' independent of a and s. For the second term, note that (4.34) and $r(p) \leqslant k_0(p)$ imply that by choosing a sufficiently large n, we can make the order $r(p)(1+\langle p\rangle)^{-n\delta}$ of RJ^n lower than $(1+\langle p\rangle)^{-d-1}$. Then, using (4.36) for K^{-1} and (4.6) for RJ^n with $\sigma = d+1$, we have

$$\sum_{s' \in \mathbb{T}_a^{d*}} |(u_{\alpha s'}, RJ^n K^{-1} u_{\beta s})| \leqslant \sum_{s', s'' \in \mathbb{T}_a^{d*}} \sum_{\gamma=1}^{2^{d/2}} |(u_{\alpha s'}, RJ^n u_{\gamma s''})||(u_{\alpha s''}, K^{-1} u_{\beta s})|$$

$$\leqslant \sum_{s', s'' \in \mathbb{T}_a^{d*}} \sum_{\gamma=1}^{2^{d/2}} \text{const.} (1+\langle s' \rangle)^{-d-1} (1+\langle s'-s'' \rangle)^{-d-1} \leqslant \text{const.} \tag{4.67}$$

□

F. Proof of Lemma 4.4

In order to show Lemma 4.4, it suffices to estimate the right-hand side of (4.46). For the first term, using (4.64), we have the following bound:

$$\sum_{j=0}^{n-1} J^j K_0^{-1} = \mathcal{O}(1/k_0). \tag{4.68}$$

For the second term, we have to estimate

$$\langle p_\mu - q_\mu \rangle^\rho |(u_{\alpha p}, J^n K^{-1} u_{\beta q})| \tag{4.69}$$

for $\rho \geqslant 0$, $p,q \in \mathbb{T}_a^{d*}$, and for $\alpha, \beta = 1,2,...,2^{d/2}$.

If $\rho = 0$, (4.49) and (4.56) yield

$$|(u_{\alpha p}, J^n K^{-1} u_{\beta p})| \leqslant \frac{\text{const.}}{k_0(p)}, \quad p,q \in \mathbb{T}_a^{d*}, \quad \alpha, \beta = 1,2,...,2^{d/2}. \tag{4.70}$$

Now assume $\rho > 0$. Then Lemma 4.7 implies

$$\langle p_\mu - q_\mu \rangle^\rho |(u_{\alpha p}, J^n K^{-1} u_{\beta q})| \leqslant \pi^\rho |(u_{\alpha p}, \text{ad}_\mu^\rho (J^n K^{-1}) u_{\beta q})|, \quad p,q \in \mathbb{T}_q^{d*}, \quad \alpha, \beta = 1,2,...,2^{d/2}. \tag{4.71}$$

Here we note the equalities

$$\text{ad}_\mu^\rho (J^n K^{-1}) = \sum_{m=0}^{\rho} \binom{\rho}{m} \text{ad}_\mu^{\rho-m}(J^n) \text{ad}_\mu^m (K^{-1}) \tag{4.72}$$

and

$$\text{ad}_\mu^m (K^{-1}) = \sum_{\ell=1}^{m} \sum_{(m_1, m_2, ..., m_\ell)} \frac{(-1)^\ell m!}{m_1! m_2 \cdots m_\ell!} K^{-1} \text{ad}_\mu^{m_1}(K) K^{-1} \text{ad}_\mu^{m_2}(K) K^{-1} \cdots \text{ad}_\mu^{m_\ell}(K) K^{-1},$$

$$m = 1,2,..., \tag{4.73}$$

where the summation $\Sigma_{(m_1, m_2, ..., m_\ell)}$ is taken over all $(m_1, m_2, ..., m_\ell)$'s such that

$$m_1 + m_2 + \cdots + m_\ell = m, \tag{4.74}$$

$$m_1, m_2, ..., m_\ell \geqslant 1. \tag{4.75}$$

Combining them, we obtain

$$\mathrm{ad}_\mu^\rho(J^nK^{-1})=\mathrm{ad}_\mu^\rho(J^n)K^{-1}+\sum_{m=1}^{\rho}\binom{\rho}{m}\sum_{\ell=1}^{m}\sum_{(m_1,m_2,\ldots,m_\ell)}\frac{(-1)^\ell m!}{m_1!m_2!\cdots m_\ell!}$$
$$\times\mathrm{ad}_\mu^{\rho-m}(J^n)K^{-1}\mathrm{ad}_\mu^{m_1}(K)K^{-1}\mathrm{ad}_\mu^{m_2}(K)K^{-1}\cdots\mathrm{ad}_\mu^{m_\ell}(K)K^{-1}. \tag{4.76}$$

Then, it suffices to bound $\mathrm{ad}_\mu^{m'}(J^n)K^{-1}$ and $\mathrm{ad}_\mu^m(K)K^{-1}$.

Note that (4.49) and the latter part of Lemma 4.7 imply $\mathrm{ad}_\mu^{m'}(J^n)=\mathcal{O}(1/k_0)$ for sufficiently large n. Then, using (4.56), we have

$$|(u_{\alpha p},\mathrm{ad}_\mu^{m'}(J^n)K^{-1}u_{\beta s})|\leqslant\frac{\text{const.}}{k_0(p)},\quad p,s\in\mathbb{T}_a^{d*},\quad \alpha,\beta=1,2,\ldots,2^{d/2}. \tag{4.77}$$

Furthermore, the assumptions on the orders of K_0 and K_1 imply $K=\mathcal{O}(k_0)$. This fact and Lemma 4.7 and (4.58) imply

$$\sum_{s\in\mathbb{T}_a^{d*}}\sum_{\beta=1}^{2^{d/2}}|(u_{\beta s},\mathrm{ad}_\mu^m(K)K^{-1}u_{\gamma s'})|\leqslant\text{const.},\quad s'\in\mathbb{T}_a^{d*},\quad \gamma=1,2,\ldots,2^{d/2}. \tag{4.78}$$

Then, (4.71), (4.76), (4.77), and (4.78) imply

$$\langle p_\mu-q_\mu\rangle^\rho|(u_{\alpha p},J^nK^{-1}u_{\beta q})|\leqslant\frac{\text{const.}}{k_0(p)},$$
$$\rho>0,\quad p,q\in\mathbb{T}_a^{d*},\quad \alpha,\beta=1,2,\ldots,2^{d/2}. \tag{4.79}$$

This together with (4.70) yields $J^nK^{-1}=\mathcal{O}(1/k_0)$, which concludes the analysis on the second term on the right-hand side of (4.46), hence the proof of Lemma 4.4.

V. ORDERS OF OPERATORS

Using the framework of quasidiagonal operators, we determine the orders of lattice operators.

A. Orders of interaction parts

Interaction parts of the operators T_μ, D_μ, W, L, and C are by definition

$$T_{\mu 1}=T_\mu-T_{\mu 0}, \tag{5.1}$$
$$D_{\mu 1}=D_\mu-D_{\mu 0}, \tag{5.2}$$
$$W_1=W-W_0, \tag{5.3}$$
$$L_1=L-L_0, \tag{5.4}$$
$$C_1=C-C_0. \tag{5.5}$$

Namely,

$$T_{\mu 1}=(U_\mu-I)T_{\mu 0}, \tag{5.6}$$
$$D_{\mu 1}=\frac{1}{2a}(T_{\mu 1}-T_{\mu 1}^*), \tag{5.7}$$
$$W_1=\sum_{\mu=1}^{d}((U_\mu-I)(T_{\mu 0}-I)+(T_{\mu 0}^*-I)(U_\mu^*-I)+(U_\mu+U_\mu^*-2I)), \tag{5.8}$$

$$L_1 = -\frac{r}{2a} W_1, \tag{5.9}$$

$$C_1 = L_1 - \sum_{\mu=1}^{d} (D_\mu (L^{-1} - L_0^{-1}) D_\mu + D_{\mu 1} L_0^{-1} D_\mu + D_{\mu 0} L_0^{-1} D_{\mu 1}). \tag{5.10}$$

Lemma 5.1: We have the following order estimates.
(1) The orders of the interaction parts:

$$T_{\mu 1} = \mathcal{O}(a), \tag{5.11}$$

$$D_{\mu 1} = \mathcal{O}(1), \tag{5.12}$$

$$W_1 = \mathcal{O}(a^2(1 + \langle p \rangle)), \tag{5.13}$$

$$L_1 = \mathcal{O}(a(1 + \langle p \rangle)). \tag{5.14}$$

(2) The orders of the full operators:

$$T_\mu = \mathcal{O}(1), \tag{5.15}$$

$$D_\mu = \mathcal{O}(1 + \langle p \rangle), \tag{5.16}$$

$$W = \mathcal{O}(a^2(1 + \langle p \rangle^2)), \tag{5.17}$$

$$L = \mathcal{O}(1 + a\langle p \rangle^2). \tag{5.18}$$

Proof: (1) follows from (5.6) to (5.9) and Lemma 4.3 together with Lemma 4.1 and

$$U_\mu - I, U_\mu^* - I = \mathcal{O}(a), \tag{5.19}$$

$$U_\mu + U_\mu^* - 2I = \mathcal{O}(a^2). \tag{5.20}$$

(2) follows from (1) and Lemma 4.1. □

B. Orders of L^{-1} and C^{-1}

Lemma 5.2: The operator L is positive definite and satisfies

$$(u, Lu) \geqslant M \|u\|^2, \tag{5.21}$$

$$\|L^{-1}\| \leqslant M^{-1}. \tag{5.22}$$

Furthermore, it holds that

$$L^{-1} = \mathcal{O}\left(\frac{1}{1 + a\langle p \rangle^2}\right), \tag{5.23}$$

$$L^{-1} - L_0^{-1} = \mathcal{O}\left(\frac{a(1 + \langle p \rangle)}{(1 + a\langle p \rangle^2)^2}\right). \tag{5.24}$$

Proof: The semipositivity of $-W$ [see (2.14)] yields (5.21), from which (5.22) follows. Furthermore, we have (5.23) from (4.18), (4.19), (5.14), and Lemma 4.4, because

$$\frac{a(1 + \langle p \rangle)}{1 + a\langle p \rangle^2} \leqslant \frac{2}{1 + \langle p \rangle}. \tag{5.25}$$

Finally, estimating the right-hand side of

$$L^{-1}-L_0^{-1}=-L^{-1}L_1L_0^{-1}, \tag{5.26}$$

we obtain (5.24). □

Lemma 5.3: The operator C is positive definite and satisfies

$$(u,Cu)\geqslant M\|u\|^2, \tag{5.27}$$

$$\|C^{-1}\|\leqslant M^{-1}. \tag{5.28}$$

Furthermore, it holds that

$$C_1=\mathcal{O}\left(\frac{1+\langle p\rangle}{1+a\langle p\rangle^2}\right), \tag{5.29}$$

$$C=\mathcal{O}\left(\frac{1+\langle p\rangle^2}{1+a\langle p\rangle^2}\right), \tag{5.30}$$

$$C^{-1}=\mathcal{O}\left(\frac{1+a\langle p\rangle^2}{1+\langle p\rangle^2}\right). \tag{5.31}$$

Proof: In the right-hand side of (3.8), L is positive definite (Lemma 5.2) and

$$D_\mu^*=-D_\mu, \quad \mu=1,2,...,d. \tag{5.32}$$

Then, (5.27) and (5.28) hold. Applying Lemmas 4.1, 5.1, and 5.2 to operators on the right-hand side of (5.10), we have (5.29), which, with (4.20), implies (5.30). Finally, we obtain (5.31) from (4.20), (4.21), and (5.28). □

C. Order of $X_{\mu\nu}$

We begin with the following.

Lemma 5.4: If ϕ is a multiplication operator which satisfies (4.30) and (4.31), we have, for $\mu=1,2,...,d$,

$$[T_{\mu 0},\phi]=\mathcal{O}(a), \tag{5.33}$$

$$[T_{\mu 0}^*,\phi]=\mathcal{O}(a), \tag{5.34}$$

$$[T_{\mu 0}+T_{\mu 0}^*,\phi]=\mathcal{O}(a^2(1+\langle p\rangle)). \tag{5.35}$$

Proof: We bound the right-hand side of

$$[T_{\mu 0},\phi]=(T_{\mu 0}\phi T_{\mu 0}^*-\phi)T_{\mu 0}. \tag{5.36}$$

Since $T_{\mu 0}\phi T_{\mu 0}^*$ is a multiplication operator determined by the translation of φ, it holds that

$$T_{\mu 0}\phi T_{\mu 0}^*-\phi=\mathcal{O}(a). \tag{5.37}$$

This proves (5.33). The estimate (5.34) is obtained in a similar way. The estimate (5.35) follows from

$$[T_{\mu 0}+T_{\mu 0}^*,\phi]T_{\mu 0}=2a(T_{\mu 0}\phi T_{\mu 0}^*-\phi)D_{\mu 0}+(T_{\mu 0}\phi T_{\mu 0}^*+T_{\mu 0}^*\phi T_{\mu 0}-2\phi)T_{\mu 0}^* \tag{5.38}$$

with the help of

$$T_{\mu 0}\phi T_{\mu 0}^*+T_{\mu 0}^*\phi T_{\mu 0}-2\phi=\mathcal{O}(a^2). \tag{5.39}$$

□

The refined bound (5.35) due to the cancellation between $T_{\mu 0}$ and $T^*_{\mu 0}$ is essential in the proof of Proposition 6.1.

Lemma 5.5: We have the following order estimates:

$$[D_{\mu 1}, D_{\nu 0}] = \mathcal{O}(1), \tag{5.40}$$

$$[D_{\mu 1}, D_{\nu 1}] = \mathcal{O}(a), \tag{5.41}$$

$$[D_{\mu}, D_{\nu}] = \mathcal{O}(1), \tag{5.42}$$

$$[L_0, D_{\nu 1}] = \mathcal{O}(a(1+\langle p\rangle)), \tag{5.43}$$

$$[L_1, D_{\nu 0}] = \mathcal{O}(a(1+\langle p\rangle)), \tag{5.44}$$

$$[L_1, D_{\nu 1}] = \mathcal{O}(a), \tag{5.45}$$

$$[L, D_{\nu}] = \mathcal{O}(a(1+\langle p\rangle)), \tag{5.46}$$

$$X_{\mu\nu} = \mathcal{O}\left(\frac{a(1+\langle p\rangle^2)}{(1+a\langle p\rangle^2)^2}\right), \tag{5.47}$$

where $\mu, \nu = 1,2,...,d$.

Proof: Put $\eta_\mu = U_\mu - I$. Then, $\eta_\mu = \mathcal{O}(a)$ and

$$[T_{\mu 1}, T_{\nu 0}] = [\eta_\mu, T_{\nu 0}] T_{\mu 0} \tag{5.48}$$

hold. Using (5.33) for $\phi = (1/a)\eta_\mu$ [the functions $U_\mu(x)$ and $\eta_\mu(x)$ can be extended on the whole $\mathbb{T}^d$], we obtain

$$[T_{\mu 1}, T_{\nu 0}] = \mathcal{O}(a^2). \tag{5.49}$$

Similarly, we have

$$[T^*_{\mu 1}, T_{\nu 0}], [T_{\mu 1}, T^*_{\nu 0}], [T^*_{\mu 1}, T^*_{\nu 0}] = \mathcal{O}(a^2). \tag{5.50}$$

Then, (5.40) holds. The estimates

$$[T_{\mu 1}, T_{\nu 1}] = -\eta_\mu [\eta_\nu, T_{\mu 0}] T_{\nu 0} + \eta_\nu [\eta_\mu, T_{\nu 0}] T_{\mu 0} = \mathcal{O}(a^3), \tag{5.51}$$

$$[T^*_{\mu 1}, T_{\nu 1}], [T_{\mu 1}, T^*_{\nu 1}], [T^*_{\mu 1}, T^*_{\nu 1}] = \mathcal{O}(a^3), \tag{5.52}$$

are obtained similarly, from which we have (5.41). The estimate (5.42) follows from (5.40) and (5.41), because $D_{\mu 0}$ and $D_{\nu 0}$ commutes.

Next, using

$$[T_{\mu 0} + T^*_{\mu 0}, T_{\nu 1}] = [T_{\mu 0} + T^*_{\mu 0}, \eta_\nu] T_{\nu 0}, \tag{5.53}$$

$$[T_{\mu 0}, T_{\nu 1} + T^*_{\nu 1}] = T_{\nu 0} [T_{\mu 0}, \eta_\nu + \eta^*_\nu] - 2aD_{\nu 0} [T_{\mu 0}, \eta^*_\nu], \tag{5.54}$$

and Lemma 5.4, and the fact that $\eta_\nu + \eta^*_\nu = \mathcal{O}(a^2)$, we have

$$[T_{\mu 0} + T^*_{\mu 0}, T_{\nu 1}] = \mathcal{O}(a^3(1+\langle p\rangle)), \tag{5.55}$$

$$[T_{\mu 0}, T_{\nu 1} + T^*_{\nu 1}] = \mathcal{O}(a^3(1+\langle p\rangle)). \tag{5.56}$$

Similarly,

$$[T_{\mu 0} + T^*_{\mu 0}, T^*_{\nu 1}] = \mathcal{O}(a^3(1+\langle p\rangle)), \tag{5.57}$$

$$[T^*_{\mu 0}, T_{\nu 1} + T^*_{\nu 1}] = \mathcal{O}(a^3(1+\langle p\rangle)) \tag{5.58}$$

hold and we obtain (5.43) and (5.44). Furthermore, (5.45) is shown by estimating the right-hand sides of

$$[T_{\mu 1} + T^*_{\mu 1}, T_{\nu 1}] = \eta_\mu [T_{\mu 0}, \eta_\nu] T_{\nu 0} + \eta_\nu [\eta_\mu, T_{\nu 0}] T_{\mu 0} + [T^*_{\mu 0}, \eta_\nu] \eta^*_\mu T_{\nu 0} + \eta_\nu T^*_{\mu 0} [\eta^*_\mu, T_{\nu 0}] \tag{5.59}$$

and its adjoint. The estimate (5.46) follows from (5.43) to (5.45). Finally, note that

$$X_{\mu\nu} = D_\mu L^{-1} [D_\nu, L] L^{-1} - D_\nu L^{-1} [D_\mu, L] L^{-1} + [D_\mu, D_\nu] L^{-1}. \tag{5.60}$$

Estimating the right-hand side of (5.60) by means of (5.16), (5.23), (5.42), and (5.46), we obtain (5.47). □

Lemma 5.6: The operator $C+B$ is positive definite and satisfies

$$(u, (C+B)u) \geqslant M \|u\|^2, \tag{5.61}$$

$$\|(C+B)^{-1}\| \leqslant M^{-1}. \tag{5.62}$$

Furthermore, it holds that

$$B = \mathcal{O}\left(\frac{a(1+\langle p\rangle^2)}{(1+a\langle p\rangle^2)^2} \right), \tag{5.63}$$

$$BC^{-1} = \mathcal{O}\left(\frac{a}{1+a\langle p\rangle^2} \right), \tag{5.64}$$

$$(C+B)^{-1} = \mathcal{O}\left(\frac{1+a\langle p\rangle^2}{1+\langle p\rangle^2} \right). \tag{5.65}$$

Proof: (5.61) and (5.62) are obtained from

$$(u, (C+B)u) = (u, Lu) + \sum_{\mu=1}^{d} (D_\mu u, L^{-1} D_\mu u) \tag{5.66}$$

and (5.21). (5.63) is a consequence of (3.7) and (5.47). (5.64) follows from (5.63) and (5.31). Lemma 4.4 then yields (5.65). □

D. Leading term of $X_{\mu\nu}$

We extract the leading term of $X_{\mu\nu}$ and bound the remainder. Put

$$E_{\mu 0} = \tfrac{1}{2}(T_{\mu 0} + T^*_{\mu 0}), \quad \mu = 1,2,...,d. \tag{5.67}$$

Lemma 5.7: For $\mu, \nu = 1,2,...,d$, we have

$$[D_\mu, D_\nu] + iQF_{\mu\nu} E_{\mu 0} E_{\nu 0} = \mathcal{O}(a), \tag{5.68}$$

$$[L, D_\mu] + iarQ \sum_{\rho=1}^{d} F_{\mu\rho} D_{\rho 0} E_{\mu 0} = \mathcal{O}(a), \tag{5.69}$$

where Q is the charge and $F_{\mu\nu}$ is the field strength defined by (2.21).

Proof: Put $\eta_\mu = U_\mu - I$. Recalling (2.3), we have

$$\eta_\mu - T_{\nu 0} \eta_\mu T^*_{\nu 0} = -a\partial_\nu \eta_\mu + \mathcal{O}(a^3) = -ia^2 Q \partial_\nu A_\mu + \mathcal{O}(a^3). \tag{5.70}$$

Then, (5.48) and (5.36) yield

$$[T_{\mu 1}, T_{\nu 0}] = ia^2 Q(_\nu A_\mu) T_{\mu 0} T_{\nu 0} + \mathcal{O}(a^3). \tag{5.71}$$

Similarly, we have

$$[T_{\mu 1}, T^*_{\nu 0}] = -ia^2 Q(\partial_\nu A_\mu) T_{\mu 0} T^*_{\nu 0} + \mathcal{O}(a^3), \tag{5.72}$$

$$[T^*_{\mu 1}, T_{\nu 0}] = -ia^2 Q(\partial_\nu A_\mu) T^*_{\mu 0} T_{\nu 0} + \mathcal{O}(a^3), \tag{5.73}$$

$$[T^*_{\mu 1}, T^*_{\nu 0}] = ia^2 Q(\partial_\nu A_\mu) T^*_{\mu 0} T^*_{\nu 0} + \mathcal{O}(a^3), \tag{5.74}$$

hence

$$[D_{\mu 1}, D_{\nu 0}] = iQ(\partial_\nu A_\mu) E_{\mu 0} E_{\nu 0} + \mathcal{O}(a). \tag{5.75}$$

This together with (5.41) yields (5.68). Let us show (5.69). (5.38) and (5.53) yield

$$[T_{\rho 0} + T^*_{\rho 0}, T_{\mu 1}] = -2ia^3 Q(\partial_\rho A_\mu) D_{\rho 0} T_{\mu 0} + \mathcal{O}(a^3). \tag{5.76}$$

Similarly, we have

$$[T_{\rho 0} + T^*_{\rho 0}, T^*_{\mu 1}] = 2ia^3 Q(\partial_\rho A_\mu) D_{\rho 0} T^*_{\mu 0} + \mathcal{O}(a^3). \tag{5.77}$$

Then,

$$[L_0, D_{\mu 1}] = iarQ \sum_{\rho=1}^{d} (\partial_\rho A_\mu) D_{\rho 0} E_{\mu 0} + \mathcal{O}(a) \tag{5.78}$$

follows. Furthermore, (5.54) implies

$$[L_1, D_{\mu 0}] = -iarQ \sum_{\rho=1}^{d} (\partial_\mu A_\rho) D_{\rho 0} E_{\mu 0} + \mathcal{O}(a). \tag{5.79}$$

Then, (5.69) follows from (5.78), (5.79), and (5.45). □

Lemma 5.8: It holds that

$$X_{\mu\nu} + iQ\left(F_{\mu\nu} E_{\mu 0} E_{\nu 0} L_0^{-1} + ar \sum_{\rho-1}^{d} F_{\rho\nu} L_0^{-2} D_{\rho 0} D_{\mu 0} E_{\nu 0} - ar \sum_{\rho=1}^{d} F_{\rho\mu} L_0^{-2} D_{\rho 0} D_{\nu 0} E_{\mu 0} \right)$$
$$= \mathcal{O}\left(\frac{a(1+\langle p \rangle)}{(1+a\langle p \rangle^2)^2} \right) \tag{5.80}$$

for $\mu, \nu = 1,2,...,d$.

Proof: We extract the leading term from the right-hand side of (5.60). First, (5.68) and (5.24) imply

$$[D_\mu, D_\nu] L^{-1} = -iQF_{\mu\nu} E_{\mu 0} E_{\nu 0} L_0^{-1} + \mathcal{O}\left(\frac{a(1+\langle p \rangle)}{(1+a\langle p \rangle^2)^2} \right). \tag{5.81}$$

Similarly, from (5.69), we have

$$D_\mu L^{-1} [D_\nu, L] L^{-1} = -iarQ \sum_{\rho} D_{\mu 0} L_0^{-1} F_{\rho\nu} D_{\rho 0} D_{\rho 0} E_{\nu 0} L_0^{-1} + \mathcal{O}\left(\frac{a(1+\langle p \rangle)}{(1+a\langle p \rangle^2)^2} \right). \tag{5.82}$$

In the right-hand side of the above equality, we may move $F_{\rho\nu}$ to the front of the term by using the following lemma. [The difference is absorbed in the second term on the right-hand side of (5.82).] Thus we have the desired result. □

Lemma 5.9: If ϕ *is a multiplication operator which satisfies (4.30) and (4.31), we have, for* $\mu = 1,2,...,d$,

$$[D_{\mu 0},\phi]=\mathcal{O}(1), \tag{5.83}$$

$$[E_{\mu 0},\phi]=\mathcal{O}(a), \tag{5.84}$$

$$[W_0,\phi]=\mathcal{O}(a^2(1+\langle p\rangle)), \tag{5.85}$$

$$[L_0^{-1},\phi]=\mathcal{O}\left(\frac{1}{(1+a\langle p\rangle^2)(1+\langle p\rangle)}\right), \tag{5.86}$$

$$[C_0^{-1},\phi]=\mathcal{O}\left(\frac{1+a\langle p\rangle^2}{(1+\langle p\rangle^3)}\right). \tag{5.87}$$

Proof: The estimates (5.83)–(5.85) follow from (5.33) to (5.35). The commutators with L_0^{-1} and with C_0^{-1} are estimated using commutators with L_0 and with C_0, respectively. □

VI. PROOF OF THEOREM 2.1

A. Irrelevant terms

If we formally expand $(C+B)^{-1}$ on the right-hand side of (3.9), we have

$$\langle Y(\xi)\rangle \overset{\text{formally}}{=} \sum_{j=0}^{\infty} z^{(j)}, \tag{6.1}$$

where

$$z^{(j)}=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}\,(\xi W+W\xi)C^{-1}(-BC^{-1})^j\gamma_{d+1}\right],\quad j=0,1,2,\ldots. \tag{6.2}$$

The following proposition is a consequence of the order estimates in Sec. V.

Proposition 6.1: For $j>d/2$, $z^{(j)}$ is irrelevant, i.e.,

$$\lim_{a\to 0}\left(\langle Y(\xi)\rangle-\sum_{j=0}^{d/2} z^{(j)}\right)=0. \tag{6.3}$$

Proof: It suffices to bound the right-hand side of

$$\langle Y(\xi)\rangle-\sum_{j=0}^{d/2} z^{(j)}=-i\ \mathrm{Tr}_{\mathbb{V}_a^d}\left[\frac{r}{2a}\,(\xi W+W\xi)(C+B)^{-1}(-BC^{-1})^{d/2+1}\gamma_{d+1}\right]. \tag{6.4}$$

The order estimates (5.17), (5.65), and (5.64) imply

$$\frac{r}{2a}\,(\xi W+W\xi)(C+B)^{-1}(-BC^{-1})^{d/2+1}\gamma_{d+1}=\mathcal{O}\left(\frac{a^{d/2+2}}{(1+a\langle p\rangle^2)^{d/2}}\right), \tag{6.5}$$

with the help of Lemmas 4.2 and 4.3. Let us take the trace on $\mathbb{V}_a^d$ by means of the planewave basis (4.4). Then, we have from (4.7) with $q=p$,

$$\left|\langle Y(\xi)\rangle-\sum_{j=0}^{d/2} z^{(j)}\right|\leqslant \text{const.}\ \sum p\in\mathbb{T}_a^{d*}\ \frac{a^{d/2+2}}{(1+a\langle p\rangle^2)^{d/2}}\leqslant \text{const.}\ a^2\ \log(1/a)\to 0, a\to 0, \tag{6.6}$$

where we used

$$\sum_{p\in\mathbb{T}_a^{d*}}\ \frac{a^{\ell}}{(1+a\langle p\rangle^2)^{\ell}}\leqslant \text{const.}\begin{cases} a^{\ell-d/2}, & l>d/2\\ \log(1/a), & \ell=d/2.\end{cases} \tag{6.7}$$

□

B. Spin properties

Next we take into account the ''spin properties'' of operators. We classify operators on $\mathbb{V}_a^d$ with respect to the homogeneous degrees in γ matrices. For example, C and B have 0 and 2 homogeneous degrees, respectively. An operator with homogeneous degree 0 is regarded as an operator on $\dot{\mathbb{V}}_a^d=\{\varphi:\mathbb{T}_a^d\to\mathbb{C}\}$. Since $\mathbb{V}_a^d=\mathbb{C}^{2^{d/2}}\otimes\dot{\mathbb{V}}_a^d$, the trace is factorized as:

$$\mathrm{Tr}_{\mathbb{V}_a^d}=\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\,\mathrm{Tr}_{\mathbb{C}^{2^{d/2}}}. \tag{6.8}$$

Proposition 6.2: It holds that

$$z^{(j)}=0, \quad j=0,2,...,d/2-1, \tag{6.9}$$

$$z^{(d/2)}=-i^{d/2+1}\sum_{\mu_1,\mu_2,\ldots,\mu_d=1}^{d}\epsilon_{\mu_1\mu_2\cdots\mu_d}\,\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\left[\frac{r}{2a}\,(\xi W+W\xi)\right.$$
$$\left.\times C^{-1}X_{\mu_1\mu_2}C^{-1}X_{\mu_3\mu_4}C^{-1}\cdots X_{\mu_{d-1}\mu_d}C^{-1}\right], \tag{6.10}$$

where $\epsilon_{\mu_1\mu_2\cdots\mu_d}$ denotes the totally antisymmetric tensor, and all the operators, having homogeneous degree 0, are regarded as defined on the space $\dot{\mathbb{V}}_a^d$.

Proof: The operator $(\xi W+W\xi)C^{-1}(-BC^{-1})^j$ has homogeneous degree $2j$ in γ's. Then, (6.9) is obvious from

$$\mathrm{Tr}_{\mathbb{C}^{2^{d/2}}}(\gamma_{\mu_1}\gamma_{\mu_2}\cdots\gamma_{\mu_\ell}\gamma_{d+1})=0, \quad 0\leqslant\ell<d, \quad \mu_1,\mu_2,\ldots,\mu_\ell=1,2,...,d. \tag{6.11}$$

On the other hand, (6.10) follows from (3.7) and

$$\mathrm{Tr}_{\mathbb{C}^{2^{d/2}}}(\gamma_{\mu_1}\gamma_{\mu_2}\cdots\gamma_{\mu_d}\gamma_{d+1})=(-2i)^{d/2}\epsilon_{\mu_1\mu_2\cdots\mu_d}, \quad \mu_1,\mu_2,\ldots,\mu_d=1,2,...,d. \tag{6.12}$$

□

C. Leading terms

We extract the leading terms from (6.10). Let us introduce the ''normal product'' $:\ :$ that moves multiplication operators to front, e.g.,

$$:\xi W_0+W_0\xi:=2\xi W_0, \tag{6.13}$$

$$:F_{\mu\nu}E_{\mu0}E_{\nu0}L_0^{-1}F_{\mu'\nu'}E_{\mu'0}E_{\nu'0}L_0^{-1}:=F_{\mu\nu}F_{\mu'\nu'}E_{\mu0}E_{\nu0}L_0^{-1}E_{\mu'0}E_{\nu'0}L_0^{-1}. \tag{6.14}$$

Proposition 6.3: Put, for $\mu,\nu=1,2,...,d$,

$$\overline{X}_{\mu\nu}=-iQ\left(F_{\mu\nu}E_{\mu0}E_{\nu0}L_0^{-1}+2ar\sum_{\rho=1}^{d}F_{\rho\nu}L_0^{-2}D_{\rho0}D_{\mu0}E_{\nu0}\right) \tag{6.15}$$

and

$$z_0=-i^{d/2+1}\sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d}\ \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}}^{d}\epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}}$$
$$\times\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\left[:\frac{r}{a}\,\xi W_0C_0^{-1}\overline{X}_{\mu_1\nu_1}C_0^{-1}\overline{X}_{\mu_2\nu_2}C_0^{-1}\cdots\overline{X}_{\mu_{d/2}\nu_{d/2}}C_0^{-1}:\right], \tag{6.16}$$

Then, we have

$$\lim_{a\to0}(z^{(d/2)}-z_0)=0. \tag{6.17}$$

Proof: Insert

$$C^{-1}=C_0^{-1}-C_0^{-1}C_1C^{-1} \tag{6.18}$$

into the right-hand side of (6.10) and expand it with respect to C_1. Lemma 4.1 and Lemma 5.3 imply that C_0^{-1} has the same order as C^{-1} and

$$C_0^{-1}C_1C^{-1}=\mathcal{O}\left(\frac{1}{1+\langle p\rangle}\,\frac{1+a\langle p\rangle^2}{1+\langle p\rangle^2}\right) \tag{6.19}$$

is of lower order than C^{-1} by the factor $1/(1+\langle p\rangle)$. Then, terms containing C_1 vanish in the continuum limit in a similar way as in the proof of Proposition 6.1. By the same reason, we can replace W by W_0 [see (5.13), (5.17), and (4.17)]. Furthermore, we can replace $X_{\mu\nu}$ by

$$-iQ\left(F_{\mu\nu}E_{\mu 0}E_{\nu 0}L_0^{-1}+ar\sum_{\rho=1}^{d}F_{\rho\nu}L_0^{-2}D_{\rho 0}D_{\mu 0}E_{\nu 0}-ar\sum_{\rho=1}^{d}F_{\rho\mu}L_0^{-2}D_{\rho 0}D_{\nu 0}E_{\mu 0}\right) \tag{6.20}$$

[see (5.80)]. Here, the second and third terms of (6.20) give the same contributions to $z^{(d/2)}$ when (5.80) is inserted into the right-hand side of (6.10). Then, we can replace $X_{\mu\nu}$ by $\overline{X}_{\mu\nu}$. Finally in order to replace $\xi W_0+W_0\xi$ by $2\xi W_0$ and to take the normal product, it suffices to bound commutators $[W_0,\xi]$, $[C_0^{-1},F_{\mu\nu}]$, $[E_{\mu' 0}E_{\nu' 0}L_0^{-1},F_{\mu\nu}]$, etc. Those commutators are estimated using Lemma 5.9: By taking a commutator with a multiplication operator, the order of the free part of an operator (with the suffix 0 like $D_{\mu 0}$) is reduced by the factor $1/(1+\langle p\rangle)$, hence the trace containing at least one commutators vanishes in the continuum limit. □

D. Symmetry arguments

In view of (6.15) and (6.16), we see that z_0 is a polynomial in r of order $d/2+1$ if we ignore the r dependencies of L_0^{-1} and C_0^{-1}. The following proposition claims that the cubic and higher order terms vanish.

Proposition 6.4: Put

$$z_1=-iQ^{d/2}\sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d}\ \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d}\epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}}\frac{r}{a}$$
$$\times\mathrm{Tr}_a^{\wedge d}\left[:\xi W_0C_0^{-(d/2+1)}L_0^{-d/2}\prod_{j=1}^{d/2}(F_{\mu_j\nu_j}E_{\mu_j 0}E_{\nu_j 0}):\right], \tag{6.21}$$

$$z_2=-iQ^{d/2}\sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d}\ \sum_{\nu_1,\nu_2,..,\nu_{d/2}}^{d}\epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}}2r^2$$
$$\times\mathrm{Tr}_a^{\wedge d}\left[:\xi W_0C_0^{-(d/2+1)}L_0^{-(d/2+1)}\sum_{\ell=1}^{d/2}\sum_{\rho=1}^{d}F_{\rho\nu_\ell}D_{\rho 0}D_{\mu_\ell 0}E_{\nu_\ell 0}\prod_{\substack{j=1\\ j\neq\ell}}^{d/2}(F_{\mu_j\nu_j}E_{\mu_j 0}E_{\nu_j 0}):\right]. \tag{6.22}$$

Then, we have

$$z_0=z_1+z_2. \tag{6.23}$$

Moreover, in the right-hand side of (6.22), we can fix the value of ρ to μ_ℓ, i.e.,

$$z_2 = -iQ^{d/2} \sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d} \epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}} 2r^2$$
$$\times \mathrm{Tr}_a^{\dot{\vee}d}\left[:\xi W_0 C_0^{-(d/2+1)} L_0^{-(d/2+1)} \sum_{\ell=1}^{d/2} F_{\mu_\ell\nu_\ell} D_{\mu_\ell 0}^2 E_{\nu_\ell 0} \prod_{\substack{j=1 \\ j\neq\ell}}^{d/2} (F_{\mu_j\nu_j} E_{\mu_j 0} E_{\nu_j 0}) : \right]. \quad (6.24)$$

Proof: Insert (6.15) into the summand on the right-hand side of (6.16) and expand it. Let us look at a term generated by the expansion that contains at least two factors of the form $2ar\Sigma_{\rho=1}^{d} F_{\rho\nu} L_0^{-2} D_{\rho 0} D_{\mu 0} E_{\nu 0}$ [the second term of the right-hand side of (6.15)]. Such a term has the form

$$:\cdots 2ar \sum_{\rho_\ell=1}^{d} F_{\rho_\ell\nu_\ell} L_0^{-2} D_{\rho_\ell 0} D_{\mu_\ell 0} E_{\nu_\ell 0} \cdots 2ar \sum_{\rho_{\ell'}=1}^{d} F_{\rho_{\ell'}\nu_{\ell'}} L_0^{-2} D_{\rho_{\ell'} 0} D_{\mu_{\ell'} 0} E_{\nu_{\ell'} 0} \cdots :. \quad (6.25)$$

Then, it is symmetric with respect to μ_ℓ and $\mu_{\ell'}$ hence it vanishes when it is summed up with respect to μ_ℓ and $\mu_{\ell'}$ under the presence of the totally antisymmetric tensor. This gives (6.23).

Next, note that $\mu_1,\nu_1,\ldots,\mu_{d/2},\nu_{d/2}$ are mutually distinct because of the totally antisymmetric tensor. Then, the sum (6.22) is decomposed into partial sums in which ρ is fixed to μ_j or ν_j for $j=1,2,\ldots,d/2$. If $\rho=\nu_\ell$, $F_{\rho\nu_\ell}$ vanishes. It therefore suffices to consider the case $\rho=\mu_j,\nu_j$ for some $j\neq\ell$. Assume $\rho=\mu_j$. In this case, the summand has the form

$$:\cdots F_{\mu_j\nu_\ell} D_{\mu_j 0} D_{\mu_\ell 0} E_{\nu_\ell 0} \cdots F_{\mu_j\nu_j} E_{\mu_j 0} E_{\nu_j 0} \cdots : \quad (6.26)$$

Then, it is symmetric with respect to ν_ℓ and ν_j, hence it vanishes when it is summed up with respect to ν_ℓ and ν_j. The term for $\rho=\nu_j$ similarly vanishes. Thus we have (6.24). □

Combining Propositions 6.1, 6.2, 6.3, and 6.4, we obtain:

Proposition 6.5: Put

$$G = \frac{r}{a} W_0 C_0^{-(d/2+1)} L_0^{-(d/2+1)} \left(L_0 + ar \sum_{\ell=1}^{d} \frac{D_{\ell 0}^2}{E_{\ell 0}} \right) \prod_{j=1}^{d} E_{j0}, \quad (6.27)$$

where we write (with a slight abuse of notation)

$$\frac{D_{\ell 0}^2}{E_{\ell 0}} \prod_{j=1}^{d} E_{j0} = D_{\ell 0}^2 \prod_{\substack{j=1 \\ j\neq\ell}}^{d} E_{j0}. \quad (6.28)$$

Then we have

$$\lim_{a\to 0} (\langle Y(\xi)\rangle + iQ^{d/2} \mathrm{Tr}_a^{\dot{\vee}d}(\xi \epsilon F^{d/2} G)) = 0, \quad (6.29)$$

where

$$\epsilon F^{d/2} = \sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d} \epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}} \prod_{j=1}^{d/2} F_{\mu_j\nu_j}. \quad (6.30)$$

Proof: Rewrite the right-hand sides of (6.21) and (6.24) as:

$$z_1 = -iQ^{d/2} \mathrm{Tr}_a^{\dot{\vee}d}\left(\xi \epsilon F^{d/2} \frac{r}{a} W_0 C_0^{-(d/2+1)} L_0^{-d/2} \prod_{j=1}^{d} E_{j0} \right), \quad (6.31)$$

$$z_2 = -iQ^{d/2}\,\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\Bigg(\xi \sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d} \epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}} \prod_{j=1}^{d/2} F_{\mu_j\nu_j} \times 2r^2 W_0 C_0^{-(d/2+1)} L_0^{-(d/2+1)} \sum_{\ell=1}^{d/2} \frac{D_{\mu_\ell 0}^2}{E_{\mu_\ell 0}} \prod_{j=1}^{d} E_{j0}\Bigg)$$

$$= -iQ^{d/2}\,\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\Bigg(\xi \sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d} \epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}} \prod_{j=1}^{d/2} F_{\mu_j\nu_j} \times 2r^2 W_0 C_0^{-(d/2+1)} L_0^{-(d/2+1)} \frac{1}{2} \sum_{\ell=1}^{d/2} \left(\frac{D_{\mu_\ell 0}^2}{E_{\mu_\ell 0}} + \frac{D_{\nu_\ell 0}^2}{E_{\nu_\ell 0}}\right) \prod_{j=1}^{d} E_{j0}\Bigg)$$

$$= -iQ^{d/2}\,\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}\Bigg(\xi \epsilon F^{d/2} 2r^2 W_0 C_0^{-(d/2+1)} L_0^{-(d/2+1)} \frac{1}{2} \sum_{\ell=1}^{d/2} \frac{D_{\ell 0}^2}{E_{\ell 0}} \prod_{j=1}^{d} E_{j0}\Bigg). \tag{6.32}$$

Then, the proposition follows from (6.3), (6.9), (6.17), (6.23), (6.31), and (6.32). □

E. Coefficients

We take the trace of the left-hand side of (6.29) with respect to the (spinless) planewave basis $u_p, p \in \mathbb{T}_a^{d*}$, defined by (4.3). In the following proposition, $G(p)$ is the eigenvalue of G, i.e., $Gu_p = G(p)u_p$ (see the proof of Lemma 4.1).

Proposition 6.6: Put

$$G(p) = -2r^2 a^d \sum_{\mu=1}^{d} (1-\cos ap_\mu) \prod_{\lambda=1}^{d} \cos ap_\lambda \times \frac{\dfrac{Ma}{r} + \sum_{\nu=1}^{d} (1-\cos ap_\nu) - \sum_{\nu=1}^{d} \dfrac{\sin^2 ap_\nu}{\cos ap_\nu}}{[(Ma + r\Sigma_{\nu=1}^{d}(1-\cos ap_\nu))^2 + \Sigma_{\nu=1}^{d} \sin^2 ap_\nu]^{d/2+1}}, \tag{6.33}$$

$$\kappa_a = \frac{1}{L^d} \sum_{p \in \mathbb{T}_a^{d*}} G(p). \tag{6.34}$$

Then, we have

$$\lim_{a\to 0}\left(\langle Y(\xi)\rangle + i\kappa_a Q^{d/2} a^d \sum_{x \in \mathbb{T}_a^d} \xi(x) \epsilon F^{d/2}(x)\right) = 0, \tag{6.35}$$

where

$$\epsilon F^{d/2}(x) = \sum_{\mu_1,\mu_2,\ldots,\mu_{d/2}=1}^{d} \sum_{\nu_1,\nu_2,\ldots,\nu_{d/2}=1}^{d} \epsilon_{\mu_1\nu_1\cdots\mu_{d/2}\nu_{d/2}} \prod_{j=1}^{d/2} F_{\mu_j\nu_j}(x), \quad x \in \mathbb{T}_a^d. \tag{6.36}$$

Proof: Let ϕ be a multiplication operator determined by a function φ which satisfies (4.30) and (4.31), and let K_0 be a free operator on $\dot{\mathbb{V}}_a^d$, i.e., an operator with u_p, $p \in \mathbb{T}_a^{d*}$ as the set of eigenfunctions. We denote the eigenvalue by $K_0(p)$, i.e., $K_0 u_p = K_0(p)u_p$. Then we see that

$$\mathrm{Tr}_{\dot{\mathbb{V}}_a^d}(\phi K_0) = \sum_{p \in \mathbb{T}_a^{d*}} (u_p, \phi K_0 u_p) = \sum_{p \in \mathbb{T}_a^{d*}} K_0(p)(u_p, \phi u_p) = \sum_{p \in \mathbb{T}_a^{d*}} K_0(p) \frac{a^d}{L^d} \sum_{x \in \mathbb{T}_a^d} \varphi(x). \tag{6.37}$$

The proposition is a consequence of (6.29) and (6.37). □

Now, the proof of Theorem 2.1 is completed if we confirm the following.

Proposition 6.7: (Ref. 7) *It holds that*

$$\lim_{a\to 0} \kappa_a = \frac{2}{(4\pi)^{d/2}(d/2)!}. \tag{6.38}$$

Proof: Put

$$g(a,q) = a^{-d} G(q/a) = -2r^2 \sum_{\mu=1}^{d} (1-\cos q_\mu) \prod_{\lambda=1}^{d} \cos q_\lambda \times \frac{\dfrac{Ma}{r} + \sum_{\nu=1}^{d} (1-\cos q_\nu) - \sum_{\nu=1}^{d} \dfrac{\sin^2 q_\nu}{\cos q_\nu}}{[(Ma + r\Sigma_{\nu=1}^{d}(1-\cos q_\nu))^2 + \Sigma_{\nu=1}^{d} \sin^2 q_\nu]^{d/2+1}} \tag{6.39}$$

for $a>0$, and

$$g(0,q) = \lim_{a\downarrow 0} g(a,q). \tag{6.40}$$

Using the function g, we write

$$\kappa_a = \left(\frac{a}{L}\right)^d \sum_{q\in a\mathbb{T}_a^{d*}} g(a,q). \tag{6.41}$$

For $q\in\mathbb{T}^{d*}=[-\pi,\pi]^d$, let $[q]$ denote the point in $a\mathbb{T}_a^{d*}$ nearest to q. $[q]$ is defined almost everywhere in $\mathbb{T}^{d*}$. Then, it holds that

$$\kappa_a = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^{d*}} dq g(a,[q]). \tag{6.42}$$

Note that, for an arbitrarily small $\epsilon>0$, we can choose $\delta>0$ such that

$$\frac{1}{(2\pi)^d} \int_{B_\delta} dq |g(a,[q])| < \epsilon, \tag{6.43}$$

hold for any $a\geqslant 0$, where $B_\delta=\{q\in\mathbb{R}^d \,|\, |q|<\delta\}$. Removing the small ball B_δ from the hypercube $\mathbb{T}^{d*}$, we can take the continuum limit:

$$\lim_{a\to 0} \frac{1}{(2\pi)^d} \int_{\mathbb{T}^{d*}\setminus B_\delta} dq g(a,[q]) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^{d*}\setminus B_\delta} dq g(0,q). \tag{6.44}$$

Then, it holds that

$$\lim_{a\to 0} \kappa_a = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^{d*}} dq g(0,q). \tag{6.45}$$

It therefore suffices to show

$$\int_{\mathbb{T}^{d*}} dq g(0,q) = \frac{2}{d}\,\Omega_d, \tag{6.46}$$

where Ω_d stands for the area of the d-dimensional unit sphere, i.e., $\Omega_d=(2\pi)^{d/2}/(d-2)!!$. This fact is proved for $d=4$ in Ref. 5 and for all d and all r in Ref. 7. □

(To be strict, our conclusion is consistent with that of Ref. 7, if we include $\Pi_\mu \epsilon_\mu$ and replace s^d-1 by s^{d-1} in the integrand of (A20) in Ref. 7.)

APPENDIX: OPERATORS

We list up the operators that are (globally) used in this paper.

$$T_{\mu 0}u(x)=u(x+ae_{\mu}), \tag{2.4}$$

$$T_{\mu}=U_{\mu}T_{\mu 0}, \tag{2.5}$$

$$D_{\mu}=\frac{1}{2a}\,(T_{\mu}-T_{\mu}^{*}), \tag{2.7}$$

$$\not{D}=\sum_{\mu=1}^{d}\gamma_{\mu}D_{\mu}, \tag{2.9}$$

$$W=-\sum_{\mu=1}^{d}\,(2I-T_{\mu}-T_{\mu}^{*}), \tag{2.10}$$

$$L=MI-\frac{r}{2a}\,W, \tag{3.5}$$

$$X_{\mu\nu}=D_{\mu}L^{-1}D_{\nu}-D_{\nu}L^{-1}D_{\mu}, \tag{3.6}$$

$$B=\frac{1}{2}\sum_{\mu,\nu=1}^{d}\,\gamma_{\mu}\gamma_{\nu}X_{\mu\nu}, \tag{3.7}$$

$$C=L-\sum_{\mu=1}^{d}\,D_{\mu}L^{-1}D_{\mu}, \tag{3.8}$$

$$E_{\mu 0}=\tfrac{1}{2}(T_{\mu 0}+T_{\mu 0}^{*}). \tag{5.67}$$

[1] S. Adler, ''Axial-vector vertex in spinor electrodynamics,'' Phys. Rev. **117**, 2426–2438 (1969).
[2] W. A. Bardeen, ''Anomolous Ward identities in spinor field theories,'' Phys. Rev. **184**, 1848–1859 (1969).
[3] J. Bell and R. Jackiw, ''A PCAC puzzle: $\pi^0\to\gamma\gamma$ in the σ model,'' Nuovo Cimento A **60**, 47–61 (1969).
[4] K. Wilson, in ''New phenomena in subnuclear physics,'' *Proceedings of the International School of Subnuclear Physics, Erice, 1975*, edited by A. Zichichi (Plenum, New York, 1979).
[5] L. K. Karsten and J. Smit, ''Lattice fermions: Species doubling, chiral invariance and the triangle anomaly,'' Nucl. Phys. B **183**, 103–140 (1981).
[6] W. Kerler, ''Axial-vector anomaly in lattice gauge theory,'' Phys. Rev. D **23**, 2384–2387 (1981).
[7] E. Seiler and I. O. Stamatescu, ''Lattice fermions and θ vacuums,'' Phys. Rev. D **25**, 2177–2184 (1982).
[8] H. Watanabe, ''Mathematical construction of chiral anomaly,'' J. Funct. Anal. **97**, 327–350 (1991).
[9] H. Watanabe and A. Yoshioka, ''Chiral anomaly on a spin manifold,'' Proc. Jpn. Acad. Ser. A: Math. Sci. **67**, 101–103 (1991).
[10] H. Watanabe and A. Yoshioka, ''A rigorous construction of the chiral anomaly on a spin manifold,'' Lett. Math. Phys. **30**, 1–22 (1994).
[11] M. Atiyah, R. Bott, and V. Patodi, ''On the heat equation and the index theorem,'' Invent. Math. **19**, 279–330 (1973).
[12] E. Seiler, ''Constructive quantum field theory: Fermions,'' in *Gauge Theories: Fundamental Interactions and Rigorous Results*, edited by P. Dita, V. Georgescu, and R. Purice (Birkhäuser, Boston, 1982), pp. 263–310.
[13] F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966).

Feynman integrals for a class of exponentially growing potentials

Tobias Kuna
Universität Bielefeld, D 33615 Bielefeld, Germany

Ludwig Streit
Universität Bielefeld, D 33615 Bielefeld, Germany and CCM, Universidade da Madeira, P 9000 Funchal, Portugal

Werner Westerkamp
Universität Bielefeld, D 33615 Bielefeld, Germany

(Received 19 May 1997; accepted for publication 16 March 1998)

We construct the Feynman integrands for a class of exponentially growing time-dependent potentials as white noise functionals. We show that they solve the Schrödinger equation. The Morse potential is considered as a special case. © *1998 American Institute of Physics.* [S0022-2488(98)03909-7]

I. INTRODUCTION

As an alternative approach to quantum mechanics Feynman introduced the concept of path integrals,[1] which developed into an extremely useful tool in many branches of theoretical physics.

Unfortunately, Feynman's intuitive idea of averaging over some set of paths is mathematically meaningful only for the heat equation where the underlying structure for the free motion is based on the Brownian paths with Wiener measure. This is stated by the famous Feynman–Kac formula. To write down solutions of the Schrödinger equation as path integrals is much more involved and often less direct. A measure does not exist for the Feynman integral as in the Euclidean case.[2] We do not give a full list of references, but we would like to refer to Ref. 3, their method using infinite dimensional Fresnel integrals and their extensive list of references to further approaches to Feynman integrals. Furthermore, we want to mention one of the most widely used methods, analytic continuation. The calculation itself takes place on the Euclidean side, where we have the whole machinery of probabilistic theory at our disposal. The final step is to perform an analytic continuation in some parameter such as time or mass to get solutions of the Schrödinger equation. In Sec. IV D we will prove that continuation in the mass is incompatible with perturbation theory in the case we are considering. Additionally, we want to mention the method of Doss,[4] using the Feynman–Kac formula and complex scaling. The method of Doss can also be used in the framework of white noise analysis.[5,6]

White noise analysis,[5,7,8] is a framework which offers various generalizations of concepts known from finite dimensional analysis to the infinite dimensional case, among them are differential operators and Fourier transform.

The underlying random variable is not Brownian motion but rather its velocity, white noise. Being independent at each time, white noise provides a suitable infinite dimensional coordinate system. The ''integral'' is understood as the dual pairing of a distribution with a test function, so that the Feynman integrand itself has meaning as a distribution. This allows us to calculate not only the propagator but, more generally, time ordered expectation values. Important for the usefulness of any approach to Feynman integrals is the class of potentials we are able to handle. In white noise analysis Feynman integrals have been constructed for different classes of potentials. The first were proposed by Ref. 9 and by Khandekar and Streit.[10] The latter construction was generalized in Refs. 11 and 12 to a wider class, also allowing time-dependent interactions. Potentials there were given as superpositions of δ functions. Unfortunately, this is restricted to one space dimension. In Ref. 6 another set was considered, the so-called Albeverio–Høegh–Krohn class[3] of potentials that are Fourier transforms of measures. Here the space dimension is arbitrary, on the other hand, the potentials are smooth and bounded. We shall instead consider Laplace transforms of measures, again for arbitrary finite space dimension. In Sec. IV we construct the

 © 1998 American Institute of Physics

Feynman integrand as a white noise distribution and show that the corresponding propagator solves the Schrödinger equation. The potentials are smooth but they grow, in general, exponentially at $\pm\infty$. They are too singular to be handled by Kato–Rellich perturbation theory. Nevertheless we show that the propagator is analytic in the coupling constant and we write it as a Dyson series. In Sec. V we consider the special case of Morse potentials $V(x)=g(e^{-2ax}-2\gamma e^{ax})$ for illustration and for more explicit calculations. This problem is solvable in closed form.[13–16] There the authors derive the Green function, the spectrum, and the eigenfunctions. These quantities are in general not analytic in g. If we change from positive to negative g we also lose the essential self-adjointness of the corresponding Hamilton operator. This dramatic change, however, does not destroy the analyticity of the propagator.

II. WHITE NOISE ANALYSIS

In this section we give a brief overview of concepts and theorems of white noise analysis which we use.[5,8,7]

The starting point of d-dimensional white noise analysis is the real separable Hilbert space

$$L_d^2 := L^2(\mathbb{R})\otimes\mathbb{R}^d, \quad d\in\mathbb{N}, \tag{1}$$

which is unitary isomorphic to a direct sum of d identical copies of $L^2(\mathbb{R})$—the space of real valued square-integrable functions with respect to Lebesgue measure. The norm in L_d^2 is given by

$$|f|_0 := \sum_{j=1}^{d} \int_{\mathbb{R}} f_j^2(s)ds, \quad f\in L_d^2. \tag{2}$$

In this space we choose the densely imbedded nuclear space

$$S_d := S(\mathbb{R})\otimes\mathbb{R}^d. \tag{3}$$

A typical element $\xi\in S_d$ is a d-dimensional vector where each component is a Schwartz test function. By $\{|\cdot|_p\}_{p\in\mathbb{N}}$ we denote a family of Hilbert norms topologizing S_d. Together with the dual space

$$S_d' := S'(\mathbb{R})\otimes\mathbb{R}^d, \tag{4}$$

we obtain the basic nuclear triple

$$S_d \subset L_d^2 \subset S_d'. \tag{5}$$

Let $\mathscr{B}$ be the σ-algebra generated by cylinder sets on S_d'. Using Minlos's theorem we construct a measure space on $(S_d',\mathscr{B})$ by fixing the characteristic function,

$$\int_{S_d'} \exp(i\langle\omega,\xi\rangle)d\mu(\omega) = \exp\left(-\frac{1}{2}|\xi|_0^2\right), \quad \xi\in S_d, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ denotes the dual pairing between S_d' and S_d. The space $(S_d',\mathscr{B},\mu)$ is called the vector valued white noise space. Within this formalism a version of d-dimensional Wiener's Brownian motion is given by

$$B(\tau) := \langle\omega,\mathbb{1}_{[0,\tau)}\rangle := (\langle\omega,\mathbb{1}_{[0,\tau)}\otimes e_1\rangle,\ldots,\langle\omega,\mathbb{1}_{[0,\tau)}\otimes e_d\rangle), \quad \omega\in S_d', \tag{7}$$

where $\{e_1,\ldots,e_d\}$ denotes the canonical basis of $\mathbb{R}^d$. We shall construct a Gel'fand triple with smooth and generalized functions of white noise around the complex Hilbert space

$$L^2(\mu) := L^2(S_d',\mathscr{B},\mu) \tag{8}$$

and we denote the scalar product in this space by

$$((f,g))=\int_{S_d'}\overline{f}(\omega)g(\omega)d\mu(\omega),\quad f,g\in L^2(\mu). \tag{9}$$

We proceed by choosing first a special subspace $(S_d)^1$ of test functionals. Then we construct the Gel'fand triple

$$(S_d)^1\subset L^2(\mu)\subset(S_d)^{-1}. \tag{10}$$

Elements of the space $(S_d)^{-1}$ are called Kondratiev distributions, the well-known Hida distributions form a subspace. Instead of reproducing the explicit construction here,[17,18] we shall characterize the distributions by their T-transforms in Theorem 2 below. Let $\Phi\in(S_d)^{-1}$, then there exist $p,q\in\mathbb{N}_0$ such that we can define for every

$$\xi\in U_{p,q}:=\{\xi\in S_d\,|\,|\xi|_p^2<2^{-q}\} \tag{11}$$

the T-transform by

$$T\Phi(\xi):=\langle\langle\Phi,\exp(i\langle\cdot,\xi\rangle)\rangle\rangle, \tag{12}$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denotes the bilinear extension of the scalar product of $L^2(\mu)$. The definition of the T-transform can be extended via analytic continuation to the complexification of S_d, which we denote by $S_{d,\mathbb{C}}$. Further, we need the definition of holomorphy in a nuclear space.[19]

Definition 1: A function $F:U\to\mathbb{C}$ is holomorphic on an open set $U\subseteq S_{d,\mathbb{C}}$ iff for all $\theta_0\in U$

1. *for any $\theta\in S_{d,\mathbb{C}}$ the mapping $\lambda\mapsto F(\theta_0+\lambda\theta)$ is holomorphic in some neighborhood of 0 in $\mathbb{C}$,*
2. *there exists an open neighborhood U' of θ_0 such that F is bounded on U'.*

F is holomorphic at 0 iff F is holomorphic in a neighborhood of 0.

Now we can give the above-mentioned characterization theorem, which is in the case of Hida distributions due to Refs. 20 and 21, and for Kondratiev distributions to Ref. 18.

Theorem 2: *Let $U\subseteq S_{d,\mathbb{C}}$ be open and $F:U\to\mathbb{C}$ be holomorphic at zero, then there exists a unique $\Phi\in(S_d)^{-1}$ such that $T\Phi=F$. Conversely, let $\Phi\in(S_d)^{-1}$ then $T\Phi$ is holomorphic at zero. The correspondence between F and Φ is a bijection if we identify holomorphic functions which coincide on an open neighborhood of zero.*

As a consequence of the characterization we also have a criterion for sequences and integrals with respect to an additional parameter.

Theorem 3: *Let $(\Phi_n)_{n\in\mathbb{N}}$ be a sequence in $(S_d)^{-1}$, such that there exists $U_{p,q}$, $p,q\in\mathbb{N}_0$, so that*

1. *all $T\Phi_n$ are holomorphic on $U_{p,q}$,*
2. *there exists $C>0$ such that $|T\Phi_n(\theta)|\leq C$ for all $\theta\in U_{p,q}$ and all $n\in\mathbb{N}$,*
3. *$(T\Phi_n(\theta))_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}$ for all $\theta\in U_{p,q}$.*

Then $(\Phi_n)_{n\in\mathbb{N}}$ converges strongly in $(S_d)^{-1}$.

Theorem 4: *Let $(\Lambda,\mathscr{A},\nu)$ be a measure space and $\lambda\mapsto\Phi_\lambda$ a mapping from Λ to $(S_d)^{-1}$. We assume that there exists $U_{p,q}$, $p,q\in\mathbb{N}_0$, such that*

1. *$T\Phi_\lambda$ is holomorphic on $U_{p,q}$ for every $\lambda\in\Lambda$,*
2. *the mapping $\lambda\mapsto T\Phi_\lambda(\theta)$ is measurable for every $\theta\in U_{p,q}$,*
3. *there exists $C\in L^1(\Lambda,\nu)$ such that*

$$|T\Phi_\lambda(\theta)|\leq C(\lambda) \tag{13}$$

for all $\theta\in U_{p,q}$ and for ν-almost all $\lambda\in\Lambda$.

Then there exist $p',q'\in\mathbb{N}_0$, which only depend on p,q, such that Φ_λ is Bochner integrable. In particular,

$$\int_\Lambda\Phi_\lambda d\nu(\lambda)\in(S_d)^{-1} \tag{14}$$

and $T[\int_\Lambda \Phi_\lambda d\nu(\lambda)]$ *is holomorphic on* $U_{p',q'}$. *We may interchange dual pairing and integration*

$$\left\langle\left\langle \int_\Lambda \Phi_\lambda d\nu(\lambda), \varphi \right\rangle\right\rangle = \int_\Lambda \langle\langle \Phi_\lambda, \varphi\rangle\rangle d\nu(\lambda), \quad \varphi \in (S_d)^1. \tag{15}$$

At the end of this section we want to give some examples of distributions.[5]

Example 5: To define the kinetic energy factor in the path integrals one would like to give a meaning to the formal expression

$$\exp\left(c \int_{\mathbb{R}} \omega(t)^2 dt \right),$$

where c is a complex constant. We define the normalized exponential

$$N\exp\left(c \int_{\mathbb{R}} \omega(t)^2 dt \right)$$

as a distribution via the following T-transform:

$$T\left[N\exp\left(c \int_{\mathbb{R}} \omega(t)^2 dt \right)\right](\xi) = \exp\left(\frac{1}{4c-2} \int_{\mathbb{R}} \xi^2(\tau) d\tau \right), \quad c \neq \frac{1}{2}.$$

Example 6: Donsker's delta function. In order to ''pin'' Brownian motion at a point $a \in \mathbb{R}^d$ we want to consider the formal composition of the Dirac delta distribution with Brownian motion: $\delta(B(t)-a)$. This can be given a precise meaning as a Hida distribution.[22] Its T-transform is given by

$$T[\delta(B(t)-a)](\xi) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{1}{2t} \left(\int_0^t i\xi(s) ds + x_0 - a \right)^2 - \frac{1}{2} |\xi|_0^2 \right).$$

III. THE FREE FEYNMAN INTEGRAND

We follow Refs. 9 and 23 in viewing the Feynman integral as a weighted average over Brownian paths. We use a slight change in the definition of the paths, which are here modeled by

$$x(\tau) = x - \sqrt{\frac{\hbar}{m}} \int_\tau^t \omega(\sigma) d\sigma =: x - \sqrt{\frac{\hbar}{m}} \langle \omega, \mathbb{1}_{(\tau,t]} \rangle. \tag{16}$$

In the sequel we set $\hbar = m = 1$ unless otherwise stated. Correspondingly, we define the Feynman integrand for the free motion by[11]

$$I_0(x,t|x_0,t_0) = N\exp\left(\frac{i+1}{2} \int_{\mathbb{R}} \omega^2(\tau) d\tau \right) \delta(x(t_0) - x_0). \tag{17}$$

We recall that the delta distribution $\delta(x(t_0)-x_0)$ is used to fix the starting point and plays the role of an initial distribution. In the sequel instead of $I_0(x,t|x_0,t_0)$ we will often use the shorthand I_0. Thus we get for the T-transform,

$$TI_0(\xi) = \frac{1}{(2\pi i|t-t_0|)^{d/2}} \exp\left[-\frac{i}{2} \int_{\mathbb{R}} \xi^2(\tau) d\tau - \frac{1}{2i|t-t_0|} \left(\int_{t_0}^t \xi(\tau) d\tau + x - x_0 \right)^2 \right]. \tag{18}$$

Not only the expectation ($\xi=0$) but also the T-transform has a physical meaning. Integrating formally by parts we find

$$TI_0(\xi)=\int_{\mathscr{S}_d'} I_0(\omega)\exp\left(-i\int_{t_0}^{t} x(\tau)\cdot\dot{\xi}(\tau)d\tau\right)d\mu(\omega)\exp\left(-\frac{i}{2}\int_{[t_0,t]^c}\xi^2(\tau)d\tau\right)$$
$$\times\exp[ix\cdot\xi(t)-ix_0\cdot\xi(t_0)]. \tag{19}$$

The multiplication denoted by a dot is just the scalar product in $\mathbb{R}^d$. Indeed it is straightforward to verify that

$$K_0^{(\xi)}(x,t|x_0,t_0)=TI_0(\xi)\exp\left(\frac{i}{2}\int_{[t_0,t]^c}\xi^2(\tau)d\tau\right)\exp[ix_0\cdot\xi(t_0)-ix\cdot\xi(t)] \tag{20}$$

is the propagator corresponding to the potential $W=\dot{\xi}(t)\cdot x$, i.e., it obeys the Schrödinger equation

$$(i\partial_t+\tfrac{1}{2}\Delta_d-\dot{\xi}(t)\cdot x)K_0^{(\xi)}(x,t|x_0,t_0)=0 \tag{21}$$

with the initial condition

$$\lim_{t\searrow t_0} K_0^{(\xi)}(x,t|x_0,t_0)=\delta(x-x_0).$$

IV. THE FEYNMAN INTEGRAND FOR A NEW CLASS OF UNBOUNDED POTENTIALS

Now we construct the Feynman integrand for a new class of potentials and calculate the propagators. In Sec. IV C we show that the propagators solve the corresponding Schrödinger equation and in Sec. IV E we generalize to time-dependent potentials.

A. The interactions

Definition 7: Let m be a complex measure on the Borel sets on $\mathbb{R}^d, d\geqslant 1$ *fulfilling the following condition:*

$$\int_{\mathbb{R}^d} e^{C|\alpha|}d|m|(\alpha)<\infty,\quad \forall C>0. \tag{22}$$

We define a potential V on $\mathbb{R}^d$ *by*

$$V(x)=\int_{\mathbb{R}^d} e^{\alpha\cdot x}dm(\alpha). \tag{23}$$

Remark 8: A consequence of the above condition (22) is that the measure *m* is finite. By Lebesgue's dominated convergence theorem we obtain that the potentials are restrictions to the real line of entire functions. In particular, they are locally bounded and without singularities. However, they are in general unbounded at $\pm\infty$.

Remark 9: The time-dependent case will be considered in Sec. IV E.

Example 10: Every finite measure with compact support fulfills the above condition (22).

Example 11: The simplest example is the Dirac measure in one dimension $m:=g\delta_a$ for $a>0$ and $g\in\mathbb{R}$. The associated potential is $V(x)=ge^{ax}$. Obviously all polynomials of exponential functions of the above kind are also in our class, too, e.g., $\sinh(ax)$, $\cosh(ax)$.

Example 12: In particular, the well known Morse potential $V(x):=g(e^{-2ax}-2\gamma e^{-ax})$ with $g,a,x\in\mathbb{R}$ and $\gamma>0$ is included in the above class. We will discuss this potential in Sec. V in more detail.

Example 13: If we choose a Gaussian density, we get potentials of the form $V(x)=ge^{bx^2}$ with $b,x\in\mathbb{R}$.

Example 14: Further entire functions of arbitrary high order of growth are inside the aforementioned class. More explicitly, the measures $d/m(\alpha):=\Theta(\alpha)\exp(-k\alpha^{1+b})d\alpha$ with $b,k>0$ and $x\in\mathbb{R}$ fulfill the condition (22). The corresponding potentials are entire functions of order $1+1/b$, see Ref. 24, Lemma 7.2.1.

B. The Feynman integrand as a generalized white noise functional

In order to handle potentials of the form given above

$$V(x)=\int_{\mathbb{R}^d} e^{\alpha\cdot x}dm(\alpha)$$

within our approach we must give a meaning to the following pointwise multiplication:

$$I=I_0\cdot\exp\left(-i\int_{t_0}^{t}V(x(\tau))d\tau\right), \tag{24}$$

where

$$x(\tau)=x-\int_{\tau}^{t}\omega(s)ds$$

is a path, as in Sec. III. For convenience we assume $t_0<t$ in the sequel. As a first step we formally expand the exponential

$$I=\sum_{n=0}^{\infty}\frac{(-i)^n}{n!}\int_{[t_0,t^n]}\int_{\mathbb{R}^{dn}} I_0\cdot\prod_{j=1}^{n}e^{\alpha_j\cdot x(\tau_j)}\prod_{j=1}^{n}dm(\alpha_j)d^n\tau \tag{25}$$

into a perturbation series. In Theorem 16 we will show the existence of the integrals and the series. However, first we have to give a definition for the product

$$I_0\cdot\prod_{j=1}^{n}e^{\alpha_j\cdot x(\tau_j)}. \tag{26}$$

Products of this type have already been considered, see, e.g., Example 6. In view of the Characterization Theorem 2 it is enough to define the product via its T-transform. Arguing formally we obtain

$$\begin{aligned}T\left(I_0\cdot\prod_{j=1}^{n}e^{\alpha_j\cdot x(\tau_j)}\right)(\xi)&=\int_{S_d'}I_0\cdot\prod_{j=1}^{n}e^{\alpha_j\cdot x(\tau_j)}\exp(i\langle\omega,\xi\rangle)d\mu(\omega)\\&=TI_0\left(\xi+i\sum_{j=1}^{n}\alpha_j\mathbb{1}_{[\tau_j,t)}\right)\exp\left(\sum_{j=1}^{n}\alpha_j\cdot x\right).\end{aligned} \tag{27}$$

We only need to verify that TI_0 is extendable to $\xi+i\Sigma_{j=1}^{n}\alpha_j\mathbb{1}_{(\tau_j,t]}$. This is clearly fulfilled, since the explicit formula (18) extends continuously to all $\xi\in L_d^2$.

Hence we may define the product in this way:

Proposition 15: Let $\tau_j\in[t_0,t]$ for $j=1,\ldots,n$; $t_0<t$ and $\alpha_j\in\mathbb{R}^d$. Then the pointwise product

$$\Phi_n=I_0\cdot\prod_{j=1}^{n}e^{\alpha_j\cdot x(\tau_j)} \tag{28}$$

defined by

$$\begin{aligned}T\Phi_n(\xi)&=TI_0\left(\xi+i\sum_{j=1}^{n}\alpha_j\mathbb{1}_{(\tau_j,t]}\right)\exp\left(\sum_{j=1}^{n}\alpha_j\cdot x\right)\\&=(2\pi i(t-t_0))^{-d/2}\exp\left[-\frac{i}{2}\int_{\mathbb{R}}\left(\xi(s)+i\sum_{j=1}^{n}\alpha_j\mathbb{1}_{(\tau_j,t]}(s)\right)^2ds\right]\\&\quad\times\exp\left\{-\frac{1}{2i(t-t_0)}\left[\int_{t_0}^{t}\xi(s)ds+i\sum_{j=1}^{n}\alpha_j(t-\tau_j)+(x-x_0)\right]^2\right\}\exp\left(\sum_{j=1}^{n}\alpha_j\cdot x\right)\end{aligned}$$

is a Kondratiev distribution.

Proof: Obviously this has an extension in $\xi \in \mathcal{S}_d(\mathbb{R})$ to all $\theta \in \mathcal{S}_{d,\mathbb{C}}(\mathbb{R})$ and fulfills the first part of Definition 1. In order to prove that $\Phi_n \in (\mathcal{S}_d)^{-1}$ by applying Theorem 2, we need a bound

$$
\begin{aligned}
|T\Phi_n(\theta)| &\leqslant (2\pi(t-t_0))^{-d/2} \exp\left\{ \frac{1}{2}|\theta|_0^2 + \int_{\mathbb{R}} \left| \theta(s) \sum_{j=1}^n \alpha_j \mathbb{1}_{(\tau_j,t]}(s) \right| ds \right\} \\
&\quad \times \exp\left\{ \frac{1}{2(t-t_0)} \left[2 \left| \left((x-x_0) + i \sum_{j=1}^n \alpha_j (t-\tau_j) \right) \int_{t_0}^t \theta(s) ds \right| + (t-t_0)|\theta|_0^2 \right. \right. \\
&\quad \left. \left. + 2 \left| (x-x_0) \cdot \sum_{j=1}^n \alpha_j (t-\tau_j) \right| \right] \right\} \exp\left(\sum_{j=1}^n \alpha_j \cdot x \right) \\
&\leqslant (2\pi(t-t_0))^{-d/2} \exp\left(\sum_{j=1}^n |\alpha_j||x-x_0| + \sum_{j=1}^n |\alpha_j||x_0| \right) \\
&\quad \times \exp\left[|\theta|_0^2 + \left(2\sqrt{t-t_0} \sum_{j=1}^n |\alpha_j| + \frac{|x-x_0|}{\sqrt{t-t_0}} \right) |\theta|_0 \right] \\
&=: C_n(\alpha_1, \ldots, \alpha_j, \theta).
\end{aligned} \tag{29}
$$

Thus Φ_n is a Kondratiev distribution, in fact by the above bound it is also a Hida distribution.[21] ■

Now we are able to prove the existence of the integrand.

Theorem 16: *Let V be as in Definition 7. Then*

$$
I := \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_{[t_0,t]^n} \int_{\mathbb{R}^{dn}} I_0 \cdot \prod_{j=1}^n e^{\alpha_j \cdot x(\tau_j)} \prod_{j=1}^n dm(\alpha_j) d^n\tau \tag{30}
$$

exists as a generalized white noise functional. The series converges in the strong topology of $(\mathcal{S}_d)^{-1}$. *The integrals exist in the sense of Bochner integrals. Therefore we can express the T-transform by*

$$
TI(\theta) = \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_{[t_0,t]^n} \int_{\mathbb{R}^{dn}} T\left(I_0 \cdot \prod_{j=1}^n e^{\alpha_j \cdot x(\tau_j)} \right)(\theta) \prod_{j=1}^n dm(\alpha_j) d^n\tau \tag{31}
$$

for all θ *in a neighborhood of zero*

$$
U_{p,q} := \{ \theta \in S_{d,\mathbb{C}} | 2^q |\theta|_p < 1 \} \tag{32}
$$

for some $p,q \in \mathbb{N}_0$.

Proof: We have already shown in Proposition 15 that the product

$$
\Phi_n := I_0 \cdot \prod_{j=1}^n e^{\alpha_j \cdot x(\tau_j)}
$$

is a Kondratiev distribution, moreover, we derived the estimate (29). In order to see that the integrals exist in the sense of Bochner we want to apply Theorem 4. As the *T*-transform of Φ_n is entire in $\xi \in \mathcal{S}_{d,\mathbb{C}}(\mathbb{R})$ and measurable, it remains only to derive a suitable bound

$$\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}} C_n(\alpha_1,\ldots,\alpha_j,\theta)\prod_{j=1}^{n} d|m|(\alpha_j)d^n\tau$$

$$\leqslant (2\pi(t-t_0))^{-d/2}(t-t_0)^n \exp\left(|\theta|_0^2+\frac{|x-x_0|}{\sqrt{t-t_0}}|\theta|_0\right)$$

$$\times\left\{\int_{\mathbb{R}^d}\exp([|x-x_0|+|x_0|+2\sqrt{t-t_0}|\theta|_0]|\alpha|)d|m|(\alpha)\right\}^n, \tag{33}$$

which is finite since the measure satisfies condition (22). Due to Theorem 4 there exists an open neighborhood U independent of n and

$$I_n := \int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}}\Phi_n\prod_{j=1}^{n} dm(\alpha_j)d^n\tau \in (\mathscr{S}_d)^{-1}, \quad \forall n\in\mathbb{N}$$

with TI_n holomorphic on U. To finish the proof we must show that the series converges in $(\mathscr{S}_d)^{-1}$ in the strong sense. For that we apply Theorem 3. We know that TI_n is holomorphic on U and we can bound it by

$$|TI(\theta)|\leqslant\sum_{n=0}^{\infty}\frac{1}{n!}|TI_n(\theta)|$$

$$\leqslant(2\pi(t-t_0))^{-d/2}\exp\left(|\theta|_0^2+\frac{|x-x_0|}{\sqrt{t-t_0}}|\theta|_0\right)$$

$$\times\exp\left\{(t-t_0)\int_{\mathbb{R}^d}\exp[(|x|+2|x_0|+2\sqrt{t-t_0}|\theta|_0)|\alpha|]d|m|(\alpha)\right\}<\infty \tag{34}$$

for $\theta\in U$, so that we prove $I\in(\mathscr{S}_d)^{-1}$. ■

Remark 17: The bound established in the proof above has a trivial, but rather surprising consequence. For the forthcoming discussion it is convenient to show the dependence on the coupling constant explicitly, so that we get

$$TI(\theta)=\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}} T\left(I_0\cdot\prod_{j=1}^{n} e^{\alpha_j\cdot x(\tau_j)}\right)(\theta)\prod_{j=1}^{n} dm(\alpha_j)d^n\tau, \tag{35}$$

which is a perturbation series in the coupling constant. Recall the bound we have already calculated; we obtain that

$$|TI(\theta)|\leqslant\sum_{n=0}^{\infty}\frac{|g|^n}{n!}|TI_n(\theta)|$$

$$\leqslant(2\pi(t-t_0))^{-d/2}\exp\left(|\theta|_0^2+\frac{|x-x_0|}{\sqrt{t-t_0}}|\theta|_0\right)$$

$$\times\exp\left\{|g|(t-t_0)\int_{\mathbb{R}^d}\exp([|x-x_0|+|x_0|+2\sqrt{t-t_0}|\theta|_0]|\alpha|)d|m|(\alpha)\right\} \tag{36}$$

and hence $TI(\theta)$ is entire in the coupling constant g for all fixed x, x_0, $t_0<t$ and $\theta\in U_{p,q}$. This is surprising, since the corresponding Hamilton operators, even if they are essentially self-adjoint for $g>0$, lose this property for $g<0$ in general. Quantities such as eigenvalues and eigenvectors will not be analytic in the coupling constant. [On the other hand under a stronger condition than (22), Albeverio *et al.*[25] have shown that the solution of the Schrödinger equation $\Psi_t(x)$ is analytic in the coupling constant if the initial wave function Ψ_0 as a function of x is from a certain class of analytic functions.]

C. Schrödinger equation

Our aim is to prove that the propagator constructed above indeed solves the Schrödinger equation. We are able to show that for $t_0<t$ the propagator does not only solve it in the sense of distributions, but also in the sense of ordinary functions. Similar to the free case, see Eq. (20), we can also give the test function in the T-transform a physical meaning corresponding to a time-dependent homogeneous external small force in the sense of (32). To compensate the extra factors appearing in the formal integration by parts, see again (20), we consider

$$TI(\theta)\cdot\exp\left[\frac{i}{2}\int_{[t_0,t]^c}\theta^2(s)ds+ix_0\cdot\theta(t_0)-ix\cdot\theta(t)\right]. \tag{37}$$

This then produces the Schrödinger propagator as follows:

Theorem 18: *Let V be as in Definition* 7. *Then*

$$\begin{aligned}K^{(\theta)}(x,t|x_0,t_0)=&\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}(2\pi i(t-t_0))^{-d/2}\\&\times\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}}\exp\left\{-\frac{i}{2}\int_{t_0}^{t}\left(\theta(s)+i\sum_{j=1}^{n}\alpha_j\mathbb{1}_{(\tau_j,t]}(s)\right)^2ds\right\}\\&\times\exp\left\{-\frac{1}{2i(t-t_0)}\left[\int_{t_0}^{t}\theta(s)ds+i\sum_{j=1}^{n}\alpha_j(t-\tau_j)+(x-x_0)\right]^2\right\}\\&\times\exp\left(\sum_{j=1}^{n}\alpha_j\cdot x\right)\exp(ix_0\cdot\theta(t_0)-ix\cdot\theta(t))\prod_{j=1}^{n}dm(\alpha_j)d^n\tau\end{aligned} \tag{38}$$

solves the Schrödinger equation for all $x,x_0,t_0<t$,

$$\left(i\frac{\partial}{\partial t}+\frac{1}{2}\Delta_d-gV(x)-x\cdot\dot{\theta}(t)\right)K^{(\theta)}(x,t|x_0,t_0)=0 \tag{39}$$

with initial condition

$$\lim_{t\searrow t_0}K^{(\theta)}(x,t|x_0,t_0)=\delta(x-x_0). \tag{40}$$

Remark 19: We may also write the propagator as a product of the free propagator with a perturbation series

$$\begin{aligned}K^{(\theta)}(x,t|x_0,t_0)=&K_0^{(\theta)}(x,t|x_0,t_0)\cdot\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}(t-t_0)^n\int_{[0,1]^n}\int_{\mathbb{R}^{dn}}\prod_{j=1}^{n}dm(\alpha_j)d^n\sigma\\&\times\exp\left\{-\frac{i}{2}(t-t_0)\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\cdot\alpha_k(\sigma_j\sigma_k-\sigma_j\wedge\sigma_k)\right]\right\}\exp\left\{\sum_{j=1}^{n}\alpha_j\cdot\left[\sigma_jx\right.\right.\\&\left.\left.+(1-\sigma_j)x_0+\sigma_j\int_{(t-t_0)\sigma_j+t_0}^{t}\theta(s)ds-(1-\sigma_j)\int_{t_0}^{(t-t_0)\sigma_j+t_0}\theta(s)ds\right]\right\}.\end{aligned} \tag{41}$$

Remark 20: The bounds in the following proof also yield that for fixed $\theta\in U_{p,q}$ the above series as a function of x, t, x_0, t_0 is C^∞ and that we are allowed to interchange integration and summation with differentiation.

Proof: It is easy to see that $T\Phi_n(\theta)$ is C^∞ in the variables x, x_0, t_0, t if $t<t_0$. For simplification we introduce the following abbreviation:

$$K^{(\theta)}(x,t|x_0,t_0)=\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}K_n^{(\theta)}(x,t|x_0,t_0), \tag{42}$$

where

$$K_n^{(\theta)}(x,t|x_0,t_0)=\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}} T\Phi_n(\theta)\exp\left(\frac{i}{2}\int_{[t_0,t]^c}\theta^2(s)ds+ix_0\cdot\theta(t_0)-ix\cdot\theta(t)\right)\times\prod_{j=1}^{n} dm(\alpha_j)d^n\tau. \tag{43}$$

By direct computation we obtain

$$\left(i\frac{\partial}{\partial t}+\frac{1}{2}\Delta_d\right)\left[T\Phi_n(\theta)\exp\left(\frac{i}{2}\int_{[t_0,t]^c}\theta^2(s)ds+ix_0\cdot\theta(t_0)-ix\cdot\theta(t)\right)\right]=x\cdot\dot{\theta}(t)\left[T\Phi_n(\theta)\exp\left(\frac{i}{2}\int_{[t_0,t]^c}\theta^2(s)ds+ix_0\cdot\theta(t_0)-ix\cdot\theta(t)\right)\right]. \tag{44}$$

Thus the above product alone solves already the free Schrödinger equation with an extra small external force. The next step is to calculate the derivatives of $K_n^{(\theta)}$. We interchange them with the integrals. Since the domain of the τ-integrals depends on t, we get an extra term compared to (44) if we differentiate with respect to t. Using the theorem of Fubini we recognize that this term corresponds to the potential $V(x)$,

$$\left(i\frac{\partial}{\partial t}+\frac{1}{2}\Delta_d\right)K_n^{(\theta)}(x,t|x_0,t_0)=x\cdot\dot{\theta}(t)K_n^{(\theta)}(x,t|x_0,t_0)+\frac{in}{g}V(x)K_{n-1}^{(\theta)}(x,t|x_0,t_0), \tag{45}$$

which is a typical recursion formula driven by the potential. By summing up we see that $K^{(\theta)}(x,t|x_0,t_0)$ solves formally the Schrödinger equation. It remains to justify the operations above. This can be done similarly to the proof of Theorem 16 if we bound the derivatives of $T\Phi_n(\theta)$ as follows. The derivatives of $T\Phi_n(\theta)$ have the form

$$\left\{a_0(\tau_1,\ldots,\tau_n,\theta)+a_1(\tau_1,\ldots,\tau_n,\theta)\cdot\left(\sum_{j=1}^{n}\alpha_j\right)+a_2(\tau_1,\ldots,\tau_n,\theta)\cdot\left(\sum_{j=1}^{n}\alpha_j\right)^2\right\}\cdot T\Phi_n(\theta), \tag{46}$$

where a_i are continuous in the τ_j. As the τ_j varies only in a compact domain we can bound the derivatives by

$$\{b_0(\theta)+b_1(\theta)+2b_2(\theta)\}\cdot\exp\left(\sum_{j=1}^{n}|\alpha_j|\right)\cdot C_n(\alpha_1,\ldots,\alpha_j,\theta), \tag{47}$$

where C_n is the bound in the proof of Proposition 15 and

$$b_i(\theta):=\sup_{\tau_j\in[t_0,t]}|a_i(\tau_1,\ldots,\tau_n,\theta)|. \tag{48}$$

The rest of the proof can be done as before. ■

D. Continuation to imaginary mass

Using continuation to imaginary mass we obtain a formal perturbation series for the propagator of the heat equation. If the measure in Definition 7 is positive we show that this perturbation series diverges. In the following discussion we need the explicit dependence of the propagator on $\hbar$ and m. For simplification we put $\theta=0$. $K^{(0)}$ is abbreviated as K. Using (41) we obtain

$$K(x,t|x_0,t_0)=K_0(x,t|x_0,t_0)\cdot\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{-ig(t-t_0)}{\hbar}\right)^n\int_{[0,1]^n}\int_{\mathbb{R}^{dn}}\prod_{j=1}^{n}dm(\alpha_j)d^n\sigma$$
$$\times\exp\left\{-\frac{i\hbar}{2m}(t-t_0)\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\cdot\alpha_k(\sigma_j\sigma_k-\sigma_j\wedge\sigma_k)\right]\right\}$$
$$\times\exp\left\{\sum_{j=1}^{n}\alpha_j\cdot[\sigma_j x+(1-\sigma_j)x_0]\right\}. \tag{49}$$

If we perform a formal analytic continuation in the mass from m to im we get

$$K^H(x,t|x_0,t_0)=K_0^H(x,t|x_0,t_0)\cdot\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{-ig(t-t_0)}{\hbar}\right)^n\int_{[0,1]^n}\int_{\mathbb{R}^{dn}}\prod_{j=1}^{n}dm(\alpha_j)d^n\sigma$$
$$\times\exp\left\{-\frac{\hbar}{2m}(t-t_0)\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\cdot\alpha_k(\sigma_j\sigma_k-\sigma_j\wedge\sigma_k)\right]\right\}$$
$$\times\exp\left\{\sum_{j=1}^{n}\alpha_j\cdot[\sigma_j x+(1-\sigma_j)x_0]\right\}. \tag{50}$$

This solves formally the heat equation with potential $-igV(x)$. Usually, convergence properties are easier to handle for the heat equation than for the Schrödinger equation. However, in our case it is the other way around.

Theorem 21: *Let $d=1$ and let m be a positive measure on the Borel sets of $\mathbb{R}$ fulfilling condition (22). If $m(\mathbb{R}\setminus\{0\})>0$ then the power series in g*

$$\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{-ig(t-t_0)}{\hbar}\right)^n\int_{[0,1]^n}\int_{\mathbb{R}^n}\prod_{j=1}^{n}dm(\alpha_j)d^n\sigma\,\exp\left\{-\frac{\hbar}{2m}(t-t_0)\right.$$
$$\left.\times\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\cdot\alpha_k(\sigma_j\sigma_k-\sigma_j\wedge\sigma_k)\right]\right\}\exp\left\{\sum_{j=1}^{n}\alpha_j\cdot[\sigma_j x+(1-\sigma_j)x_0]\right\} \tag{51}$$

diverges for every $g\neq 0$ for any fixed $x_0,x,t_0<t$.

Remark: Analogously we can handle the continuation in $\hbar$ and time.

Proof: Either there exists $a_0>0$ with $m([a_0,\infty))>0$ or $a_0<0$ with $m((-\infty,a_0])>0$. Without loss of generality we assume $a_0>0$. We use the shorthand

$$F(\alpha_1,\ldots,\alpha_n,\sigma_1,\ldots,\sigma_n):=\exp\left\{-\frac{\hbar}{2m}(t-t_0)\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\cdot\alpha_k(\sigma_j\sigma_k-\sigma_j\wedge\sigma_k)\right]\right\}$$
$$\times\exp\left\{\sum_{j=1}^{n}\alpha_j\cdot[\sigma_j x+(1-\sigma_j)x_0]\right\}. \tag{52}$$

For $\frac{3}{16}\leqslant\sigma_j,\sigma_k\leqslant\frac{4}{16}$ we have

$$\sigma_j\sigma_k-\sigma_j\wedge\sigma_k\leqslant-\tfrac{1}{16}. \tag{53}$$

Since the integrand is positive we get

$$\left|\int_{[0,1]^n}\int_{\mathbb{R}^n} F(\alpha_1,\ldots,\alpha_n,\sigma_1,\ldots,\sigma_n)\prod_{j=1}^n dm(\alpha_j)d^n\sigma\right|$$

$$\geqslant \int_{[3/16,4/16]^n}\int_{[a_0,\infty)^n} F(\alpha_1,\ldots,\alpha_n,\sigma_1,\ldots,\sigma_n)\prod_{j=1}^n dm(\alpha_j)d^n\sigma$$

$$\geqslant \int_{[3/16,4/16]^n}\int_{[a_0,\infty)^n} \exp\left\{\frac{\hbar}{2m}(t-t_0)\frac{1}{16}n^2a_0^2\right\}$$

$$\times\exp\left\{-\sum_{j=1}^n |\alpha_j|(|x_0|+|x|)\right\}\prod_{j=1}^n dm(\alpha_j)d^n\sigma$$

$$\geqslant\left(\frac{1}{16}\right)^n \exp\left\{\frac{\hbar}{32m}(t-t_0)a_0^2n^2\right\}\left(\int_{a_0}^\infty \exp\{-|\alpha|(|x_0|+|x|)\}dm(\alpha)\right)^n. \tag{54}$$

By the above assumption for a_0 the last factor does not vanish. Thus we get for the series

$$\sum_{n=0}^\infty \left|\frac{1}{n!}\left(\frac{-ig(t-t_0)}{\hbar}\right)^n\int_{[0,1]^n}\int_{\mathbb{R}^n} F(\alpha_1,\ldots,\alpha_n,\sigma_1,\ldots,\sigma_n)\prod_{j=1}^n dm(\alpha_j)d^n\sigma\right|$$

$$\geqslant\sum_{n=0}^\infty \frac{1}{n!}\left(\frac{|g|(t-t_0)}{16\hbar}\int_{a_0}^\infty \exp\{-|\alpha|(|x_0|+|x|)\}dm(\alpha)\right)^n \exp\left(\frac{\hbar}{32m}(t-t_0)a_0^2n^2\right)=\infty.$$

■

E. Time-dependent potentials

One of the advantages of the Feynman integral is that it can be easily extended to time-dependent potentials.

Theorem 22: *Let m denote a complex measure on the Borel sets of* $\mathbb{R}^d\times[t_0',t']$; $d\geqslant 1$, *such that*

$$\int_{\mathbb{R}^d}\int_{t_0'}^{t'} e^{C|\alpha|}d|m|(\alpha,\tau)<\infty,\quad \forall C>0. \tag{55}$$

Then for $t_0'\leqslant t_0<t\leqslant t'$,

$$I:=\sum_{n=0}^\infty \frac{(-i)^n}{n!}\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}} I_0\cdot\prod_{j=1}^n e^{\alpha_j\cdot x(\tau_j)}\prod_{j=1}^n dm(\alpha_j,\tau_j) \tag{56}$$

exists as a generalized white noise functional in $(\mathcal{S}_d)^{-1}$. *The T-transform fulfills the following equation:*

$$TI(\theta)=\sum_{n=0}^\infty \frac{(-i)^n}{n!}\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}} T\left(I_0\cdot\prod_{j=1}^n e^{\alpha_j\cdot x(\tau_j)}\right)(\theta)\prod_{j=1}^n dm(\alpha_j,\tau_j) \tag{57}$$

for all θ *in a neighborhood of zero.*

Remark 23: If we consider the Schrödinger equation for the whole class defined in the above theorem, the potential becomes a distribution in the time variable. This causes technical difficulties; we only use the special forms given below.

Theorem 24: *Let m be as in Theorem 22. If additionally m has either the special form*

$$dm(\alpha,\tau)=\sum_{j=1}^k dm_j(\alpha)\rho_j(\tau)d\tau \tag{58}$$

with $k\in\mathbb{N}$, m_j complex measures on the Borel sets of $\mathbb{R}^d$ and $\rho_j\in C^0(\mathbb{R},\mathbb{C})$ for all $j=1,...,k$; or the special form

$$dm(\alpha,\tau)=\rho(\alpha,\tau)\ \ d^d\alpha\ \ d\tau \tag{59}$$

where $\rho:\mathbb{R}^d\times[t_0',t']\to\mathbb{C}$ with $\rho(\alpha,\cdot)$ continuous on $[t_0',t']$ for all $\alpha\in\mathbb{R}^d$ and $\sup_{\tau\in[t_0',t']}|\rho(\alpha,\tau)|e^{C|\alpha|}$ in $L^1(\mathbb{R},d^d\alpha)$ for all $C>0$.
Then the propagator

$$\begin{aligned}K^{(\theta)}(x,t|x_0,t_0)=&\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}(2\pi i(t-t_0))^{-d/2}\exp(ix_0\cdot\theta(t_0)-ix\cdot\theta(t))\\&\times\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}}\exp\left\{-\frac{i}{2}\int_{t_0}^{t}\left(\theta(s)+i\sum_{j=1}^{n}\alpha_j\mathbb{1}_{(\tau_j,t]}(s)\right)^2ds\right\}\exp\left\{-\frac{1}{2i(t-t_0)}\right.\\&\times\left.\left[\int_{t_0}^{t}\theta(s)ds+i\sum_{j=1}^{n}\alpha_j(t-\tau_j)+(x-x_0)\right]^2\right\}\exp\left(\sum_{j=1}^{n}\alpha_j\cdot x\right)\prod_{j=1}^{n}dm(\alpha_j,\tau_j)\end{aligned} \tag{60}$$

solves the Schrödinger equation with potential $x\cdot\dot{\theta}(t)+V(x,t)$ for all $t_0'<t_0<t<t'$, where

$$V(x,t):=\sum_{j=1}^{k}\int_{\mathbb{R}^d}e^{\alpha\cdot x}dm_j(\alpha)\rho_j(t) \tag{61}$$

or respectively

$$V(x,t):=\int_{\mathbb{R}^d}e^{\alpha\cdot x}\rho(\alpha,t)d^d\alpha. \tag{62}$$

Proof of Theorems 22 and 24: The proof can be done in a similar way as in the previous subsection. Only the measure $dm(\alpha_j)d\tau_j$ has to be replaced by $dm(\alpha_j,\tau_j)$. More explicitly, we then get instead of bound (33) the following

$$\begin{aligned}&\int_{[t_0,t]^n}\int_{\mathbb{R}^{dn}}C_n(\alpha_1,\ldots,\alpha_j,\theta)\prod_{j=1}^{n}d|m|(\alpha_j,\tau_j)\\&\leqslant(2\pi(t-t_0))^{-d/2}\exp\left(|\theta|_0^2+\frac{|x-x_0|}{\sqrt{t-t_0}}|\theta|_0\right)\\&\quad\times\left\{\int_{\mathbb{R}^d}\int_{t_0}^{t}\exp([|x-x_0|+|x_0|+2\sqrt{t-t_0}|\theta|_0]|\alpha|)d|m|(\alpha,\tau)\right\}^n.\end{aligned} \tag{63}$$

To complete the proof of Theorem 24 we have to justify the interchange of the time derivative with the α-integrals. For this we use an extra condition as, e.g., in the second special form the integrability of $\sup_{\tau\in[t_0',t']}|\rho(\alpha,\tau)|\exp(|\alpha|\,|x|)$. ■

V. A SPECIAL CASE: THE MORSE POTENTIAL

In order to illustrate the remarkable fact that the propagator is analytic in the coupling constant we discuss the Morse potential as a special case of the class of potentials we studied in Sec. IV. This potential has been very useful in molecular and nuclear physics, see, e.g., Ref. 26.

Definition 25: In $L^2(\mathbb{R},\mathbb{C})$ we consider the Hamilton operator

$$H:=-\tfrac{1}{2}\Delta+g(e^{-2ax}-2\gamma e^{-ax}) \tag{64}$$

with $\gamma,g,a\in\mathbb{R}$. As domain we choose

$$D(H) := C_0^\infty(\mathbb{R},\mathbb{C}), \tag{65}$$

the set of infinite differentiable functions with compact support. For $g>0$ *and* $\gamma>0$ *this is called the Morse potential.*

We now collect some well-known results from operator theory.[27,28]

Proposition 26: H is symmetric and

$$D(H^*)=\left\{ f\in L^2(\mathbb{R},\mathbb{C})\cap C^1(\mathbb{R},\mathbb{C}) \,\middle|\, \begin{array}{c} f' \text{ is absolutely continuous} \\ \text{on each compact interval of } \mathbb{R} \text{ and } Hf\in L^2(\mathbb{R},\mathbb{C}) \end{array} \right\}.$$

Since we have a real potential the deficiency indices are equal and thus there exists a self-adjoint extension, which is not necessarily unique.

From Theorems X.8 and X.9 in Ref. 28 we can derive the following proposition:

Proposition 27: H is essentially self-adjoint for $g\geqslant 0$ *and it is not essentially self-adjoint for* $g<0$.

For the Morse potential ($g>0$) there exist treatments by algebraic,[29] and by operator methods.[30] Furthermore, path integral techniques have been applied.[13–15] One uses path-dependent space-time transformation to convert the path integral for the Morse potential into the radial path integral for the harmonic oscillator in three dimensions which is well known. Further refinements were done in Refs. 16 and 31. Using the latter technique one calculates the Green function, the kernel of the resolvent; moreover, one derives the spectrum and the eigenfunctions. Originally the method of path-dependent space-time transformation was applied to calculate the Feynman integral of the hydrogen atom.[32,33] The following formulas are taken from Ref. 16. First we will have a look at the Green function.

$$\begin{aligned} G(x',x;E) &:= \langle x'|(E-H)^{-1}|x\rangle \\ &= -\frac{\Gamma((1+\nu-\gamma\omega)/2)}{\omega|a|/2\Gamma(\nu+1)}\exp\left(\frac{a}{2}(x+x')\right) \\ &\quad\times\{\Theta(a(x-x'))W_{\gamma\omega/2,\nu/2}(\omega e^{-ax'})M_{\gamma\omega/2,\nu/2}(\omega e^{-ax}) \\ &\quad+\Theta(a(x'-x))M_{\gamma\omega/2,\nu/2}(\omega e^{-ax'})W_{\gamma\omega/2,\nu/2}(\omega e^{-ax})\} \end{aligned} \tag{66}$$

with $x',x\in\mathbb{R}$. Θ denotes the Heaviside function with the convention $\Theta(0)=1/2$, Γ is Euler's gamma function, $M_{\varkappa,\mu/2}$ and $W_{\varkappa,\mu/2}$ Whittaker's functions,[34] $\omega := 2\sqrt{2g}/|a|$ and $\nu := 2\sqrt{-2E}/|a|$. The $\sqrt{\cdot}$ denotes the principal branch of the square root with the cut along the negative real half line. Then we get for the spectrum

$$\sigma(H)=[0,\infty)\cup\left\{ -\frac{1}{8}a^2\left(\frac{2\gamma}{|a|}\sqrt{2g}-2n-1\right)^2 \,\middle|\, n\in\mathbb{N}_0 \text{ with } \frac{2\gamma}{|a|}\sqrt{2g}-2n-1>0 \right\} \tag{67}$$

and for the eigenvectors of the discrete eigenvalues

$$\begin{aligned} \Psi_n(x) &= \sqrt{\frac{|a|(\gamma\omega-2n-1)\Gamma(n+1)}{\Gamma(\gamma\omega-n)}}\,\omega^{\gamma\omega/2-(2n+1)/2}\exp\left(-\frac{a}{2}(\gamma\omega-(2n+1))x\right) \\ &\quad\times\exp\left(-\frac{\omega}{2}e^{-ax}\right)L_n^{(\gamma\omega-2n-1)}(\omega e^{-ax}) \end{aligned} \tag{68}$$

where $L_n^{(\mu)}$ is a generalized Laguerre polynomial.[34]

Proposition 28: The Green function (66), the eigenvectors, and the discrete eigenvalues are not analytic in g.

Proof: For the discrete eigenvalues the above statement is obvious. The Green function is not even analytic in $\omega=2\sqrt{2g}/|a|$. Using Ref. 35 we can rewrite (66) as

$$
\begin{aligned}
G(x',x;E) = &-\frac{2\pi}{|a|\sin(\nu\pi)}\exp\left(-\frac{\omega}{2}(e^{-ax'}+e^{-ax})\right)\\
&\times\Bigg\{\Theta(a(x-x'))\frac{{}_1F_1((1+\nu-\gamma\omega)/2;1+\nu,\omega e^{-ax})}{\Gamma(1+\nu)}\\
&\times\Bigg[-\omega^{\nu}e^{-\nu a(x+x')/2}\frac{\Gamma((1+\nu-\gamma\omega)/2)}{\Gamma((1-\nu-\gamma\omega)/2)}\frac{{}_1F_1((1+\nu-\gamma\omega)/2;1+\nu,\omega e^{-ax'})}{\Gamma(1+\nu)}\\
&+e^{\nu a(x'-x)/2}\frac{{}_1F_1((1-\nu-\gamma\omega)/2;1-\nu,\omega e^{-ax'})}{\Gamma(1-\nu)}\Bigg]\\
&+\Theta(a(x'-x))\frac{{}_1F_1((1+\nu-\gamma\omega)/2;1+\nu,\omega e^{-ax'})}{\Gamma(1+\nu)}\\
&\times\Bigg[-\omega^{\nu}e^{-\nu a(x+x')/2}\frac{\Gamma((1+\nu-\gamma\omega)/2)}{\Gamma((1-\nu-\gamma\omega)/2)}\frac{{}_1F_1((1+\nu-\gamma\omega)/2;1+\nu,\omega e^{-ax})}{\Gamma(1+\nu)}\\
&+e^{\nu a(x-x')/2}\frac{{}_1F_1((1-\nu-\gamma\omega)/2;1-\nu,\omega e^{-ax})}{\Gamma(1-\nu)}\Bigg]\Bigg\}
\end{aligned}
$$

with $\nu=2\sqrt{-2E}/|a|$ and ${}_1F_1$ denotes a generalized hypergeometric function.[34] The function

$$
\frac{{}_1F_1(a;b,x)}{\Gamma(b)} \tag{69}
$$

is entire in a, b, x. Thus we only have to investigate $\Gamma((1+\nu-\gamma\omega)/2)\omega^{\nu}$. This is obviously not analytic in ω near $\omega=0$. For the eigenvectors we can proceed along the same line. ■

Although the Green function, the discrete eigenvalues, and the eigenfunctions are not analytic, the propagator has a perturbation series which is uniformly absolutely convergent in the coupling constant for every compact set in the variables x, t, x_0, t_0. For the Morse potential we obtain an expansion for the propagator from (41). Putting $\theta=0$ and doing the $dm(\alpha_j)$ integrations we get

$$
\begin{aligned}
K(x,t|x_0,t_0) = K_0(x,t|x_0,t_0)\cdot\sum_{n=0}^{\infty}\frac{(-ig)^n}{n!}(t-t_0)^n\sum_{j_1,\ldots,j_n=1}^{2}(-2\gamma)^{2n-\sum_{k=1}^{n}j_k}\\
\times\int_{[0,1]^n}\exp\left\{-a\sum_{l=1}^{n}j_l(\sigma_l x+(1-\sigma_l)x_0)\right\}\\
\times\exp\left\{-\frac{i}{2}(t-t_0)a^2\sum_{l=1}^{n}\sum_{k=1}^{n}j_k j_l[\sigma_j\sigma_k-\sigma_j\wedge\sigma_k]\right\}d^n\sigma.
\end{aligned} \tag{70}
$$

In the sum over n the coefficient of g^n is bounded by

$$
\frac{1}{n!}|t-t_0|^n(1+2|\gamma|)^n e^{2n|a|(|x-x_0|+|x_0|)}.
$$

ACKNOWLEDGMENTS

We are grateful to Professor Ch. Bernido, Professor V. Bernido, and Professor Yu. G. Kondratiev for fruitful discussions. We thank also our colleagues M. Grothaus and J. L. da Silva for helpful comments. This work was supported in part by Financiamento Plurianual, JNICT, No. 219/94.

[1] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965).
[2] R. H. Cameron, J. Math. and Phys. **39**, 126 (1960).
[3] S. Albeverio and R. Høegh-Krohn, *Mathematical Theory of Feynman Path Integrals*, LNM Vol. 523 (Springer, Berlin, 1976).
[4] H. Doss, Commun. Math. Phys. **73**, 247 (1980).
[5] T. Hida, H. H. Kuo, J. Potthoff, and L. Streit, *White Noise. An Infinite Dimensional Calculus* (Kluwer, Dordrecht, 1993).
[6] W. Westerkamp, Ph.D. thesis, University of Bielefeld, 1995.
[7] (a) N. Obata, *White Noise Calculus and Fock Space*, LNM Vol. 1577 (Springer, Berlin, 1994); (b) Yu. M. Beresansky and Yu. G. Kondratiev, *Spectral Methods in Infinite-Dimensional Analysis* (Kluwer, Dordrecht, 1995).
[8] H. H. Kuo, *White Noise Distribution Theory* (CRC, Boca Raton, FL, 1996).
[9] M. de Faria, J. Potthoff, and L. Streit, J. Math. Phys. **32**, 2123 (1991).
[10] D. C. Khandekar and L. Streit, Ann. Phys. (Leipzig) **1**, 46 (1992).
[11] A. Lascheck, P. Leukert, L. Streit, and W. Westerkamp, Rep. Math. Phys. **33**, 221 (1993).
[12] M. Grothaus, D. C. Khandekar, J. L. Silva, and L. Streit, J. Math. Phys. **38**, 3278 (1997).
[13] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics and Polymer Physics* (World Scientific, Singapore, 1990).
[14] N. K. Pak and I. Sokmen, Phys. Rev. A **30**, 1629 (1984).
[15] P. Y. Cai, A. Inomata, and R. Wilson, Phys. Lett. **96A**, 117 (1983).
[16] W. Fischer, H. Leschke, and P. Müller, J. Phys. A **25**, 3835 (1992).
[17] Y. G. Kondratiev, Ph.D. thesis, Kiev University, 1978.
[18] Y. G. Kondratiev, P. Leukert, and L. Streit, Acta Appl. Math. **44**, 269 (1996).
[19] J. A. Barroso, *Introduction to Holomorphy*, Mathematical Studies Vol. 106 (North-Holland, Amsterdam, 1985).
[20] J. Potthoff and L. Streit, J. Funct. Anal. **101**, 212 (1991).
[21] Y. G. Kondratiev *et al.*, J. Funct. Anal. **141**, 301 (1996).
[22] H. H. Kuo, Soochow J. Math. **18**, 229 (1992).
[23] T. Hida and L. Streit, Stoch. Proc. Appl. **16**, 55 (1983).
[24] E. Lukacs, *Characteristic Functions*, 2nd ed. (Griffin, London, 1970).
[25] S. Albeverio, Z. Brzeźniak, and Z. Haba (to appear in Potential Anal.).
[26] G. Herzberg, *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules* (Van Nostrand, New York, 1950).
[27] N. Dunford and J. T. Schwartz, *Linear Operators* (Interscience, New York, 1963), Vol. II.
[28] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vols. I and II.
[29] L. Chetouani, L. Guechi, and T. F. Hammann, Helv. Phys. Acta **65**, 1069 (1992).
[30] M. M. Nieto and L. M. Simmons, Jr., Phys. Rev. D **20**, 1342 (1979).
[31] A. Pelster, Ph.D. thesis, Universität Stuttgart, 1996.
[32] I. H. Duru and H. Kleinert, Phys. Lett. **84B**, 185 (1979).
[33] P. Blanchard and M. Sirugue, J. Math. Phys. **22**, 1372 (1981).
[34] *Higher Transcendental Functions*, The Bateman Manuscript Project Vols. I and II, edited by A. Erdélyi (McGraw–Hill, New York, 1953).
[35] H. Buchholz, *The Confluent Hypergeometric Function* (Springer, Berlin, 1969).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

On simulating Liouvillian flow from quantum mechanics via Wigner functions

A. N. Mitra[a)]
244 Tagore Park, Delhi-110009, India

R. Ramanathan
Department of Physics, University of Delhi, Delhi-110007, India

(Received 18 December 1997; accepted for publication 21 February 1998)

The interconnection between quantum mechanics and probabilistic classical mechanics for a free relativistic particle is derived in terms of Wigner functions (WF) for both Dirac and Klein-Gordon (K-G) equations. Construction of WF is achieved by first defining a bilocal 4-current and then taking its Fourier transform w.r.t. the relative 4-coordinate. The K-G and Proca cases also lend themselves to a closely parallel treatment *provided* the Kemmer-Duffin β-matrix formalism is employed for the former. Calculation of WF is carried out in a Lorentz-covariant fashion by standard ''trace'' techniques. The results are compared with a recent derivation due to Bosanac. © *1998 American Institute of Physics.* [S0022-2488(98)02907-7]

I. INTRODUCTION: WIGNER FUNCTION AS LINK BETWEEN QM AND CM

Since the advent of Quantum Mechanics (QM), there have been several investigations aimed at understanding Schrödinger equations and their relativistic counterparts from stochastic/statistical premises in classical mechanics, epitomized by Focker-Planck[1,2] and Liouville equations.[3–6] In the process most such investigations have had to resort at some stage to some sort of ansatz designed to introduce the ''square-root'' of a probability function,[1–4] or make a more direct use of the wave function from the outset.[5,6] This of course does not address the core issue of the foundations of quantum physics, viz., the existence of complex probability amplitudes obeying the ''superposition principle,'' and the need to invoke the ''collapse hypothesis'' to explain the outcome of a quantum measurement. Even without any commitment to resolve this central mystery of conventional quantum physics, it is both interesting and illuminating to explore the classical roots of quantum mechanics by comparing and relating their respective mathematical structures as in several contemporary investigations.[3–6]

Among the several approaches to an alternative ''derivation'' of QM, we would like to pick out for explicit scrutiny, a recent formulation due to Bosanac[4] in which the author makes essential use of the probabilistic (instead of ''trajectory-oriented'') approach to Classical Mechanics (CM) via the Liouville equation, wherein the Uncertainty Principle is sought to be incorporated at the outset, as a means of capturing a good chunk of QM. Our purpose in this regard is more in the nature of an ''inverse'' (to Ref. 4) approach, which stems from a desire to give a more direct role to the QM wave function borne out of its ''being already there,'' thereby according it a more open role in the construction of the Wigner Function,[7] as in other orthodox approaches.[5,6] As we shall see, this approach has the advantage of presenting the treatment of different types of relativistic equations (Dirac, Klein-Gordon, Proca, etc.) within a unified framework.

The central quantity that satisfies the Liouville equation is the Wigner-function[7] which in itself was conceived within the QM framework, as a bilinear in the ''wave function.'' Granting the existence of the wave function in turn implies the existence of its Fourier transform which represents its momentum space counterpart, and the two together guarantee a built-in effect of the Uncertainty Principle, as clearly recognized in Ref. 4. Using such a (hybrid) approach in which the prior existence of the wave function is implicit in the structure of the Wigner function,[7] Bosanac[4] claims to derive the Schrödinger equation from the Liouville equation, consistently with current conservation, apparently as an inverse (?) problem. An extension to the corresponding relativistic

[a)]Electronic mail: insa@giasdl01.vsnl.net.in(subj:a.n.mitra); anmitra@csec.ernet.in

 © 1998 American Institute of Physics

problem, rather surprisingly, seemed to yield only the Dirac equation (spinor), but not the Klein-Gordon (K-G) equation (scalar), a result whose mystery we seek to clarify in this paper and (hopefully) pave the way to a generalization (see below).

Specifically, we have been actuated by a desire to re-examine the issue of Dirac versus K-G,[4] but with a more conscious recognition of an active role of the ''wave function'' at the outset, and in the process we find that the Liouville and current conservation equations for both the Dirac and K-G cases follow from identical premises, *provided* one has no qualms about the free use of *multi-component* wave functions in both Dirac and K-G cases, and not (selectively) for the Dirac case only.[4] Indeed it is easily seen from a detailed perusal of the steps in Ref. 4 that the author's success in deriving the Dirac equation is crucial to his *ansatz* of a 4-component spinor, without which the necessary consistency checks would not be forth-coming. On the other hand he did not seem to consider a similar possibility for the K-G case, viz., the existence of the less familiar Kemmer-Duffin equations for integral spins,[8] wherein the multicomponent structure of the corresponding wave function satisfying a first order equation as in the Dirac case seems to be the right form for putting the K-G and Proca cases on parallel footing with Dirac. Of course the question of whether, in spite of a multicomponent ansatz for the K-G wave function, a bosanac-like treatment[4] of working backwards from the Liouville-cum-current-conservation equations to the K-G equation would have succeeded, is still open, since the Kemmer-Duffin β-matrices are singular, unlike the Dirac matrices which are not (see Sec. IV).

In carrying out this exercise we have been guided solely by the orthodox (Dyson[9]) philosophy that a relativistic quantization of a *single* particle has inherent problems of consistency, whose resolution hinges on the simultaneous existence of an infinite number of particles, i.e., on a *field* concept. Indeed, Dirac's postulation of the infinite sea (of negative energy particles) in the vacuum was enough ground for Pauli-Weisskopf[10] to reinterpret the single particle density from the K-G equation as an average charge density (which could ''locally'' have either sign).[9] Of course from a practical point of view, a single particle interpretation has considerable appeal, but it requires the ''protective cover'' of Feynman's positron theory[11] which mandates $\pm$ time directions to be associated with $\pm$ energy propagations, respectively.

In what follows, we shall keep in the forefront, the results of Ref. 4 for pointwise comparison with the derivation to be presented here under the orthodox ''field'' paradigm[9] which is routinely available in any standard text.[12] In Sec. II, we give a ''direct'' derivation of the Wigner function for the Dirac equation and thence the corresponding Liouville plus current conservation equations, a procedure whose steps are roughly in ''inverse'' order to those of Ref. 4. In Sec. III, we do the corresponding exercise for the K-G and Proca cases via the Kemmer-Duffin[8] formulation,[13] and in so doing, bring out the strong similarity of the two systems. Section IV concludes with a short discussion, including a comparison with Ref. 4.

II. WIGNER FUNCTION FOR THE DIRAC EQUATION

We begin by writing the Dirac equation for a free field $\psi(x)$ in the ''Euclidean'' covariant notation as[13]

$$(\gamma_\mu \partial_\mu + m)\psi(x) = 0; \quad \overline{\psi}(x)(m - \partial_\mu \gamma_\mu) = 0. \tag{2.1}$$

Now define a bilocal current $j_\mu(x_1, x_2)$ as

$$j_\mu(x_1, x_2) = \overline{\psi}(x_1) i \gamma_\mu \psi(x_2). \tag{2.2}$$

Next put $x_1 = x - q$ and $x_2 = x + q$ as in Ref. 4, and define a 4-vector $V_\mu(x,p)$ as

$$V_\mu(x,p) = \pi^{-4} \int d^4q e^{2ip.q} \overline{\psi}(x+q) i \gamma_\mu \psi(x-q). \tag{2.3}$$

The fourth component of V_μ may be identified with the standard WF ρ. As a check on this quantity, its integral w.r.t. d^4p yields the usual 4-current $J_\mu(x)$ as

$$J_\mu(x)=\int d^4p\,V_\mu(x,p)=\overline{\psi}(x)i\gamma_\mu\psi(x), \tag{2.4}$$

from which the standard current conservation equation is recovered as

$$\partial_\mu J_\mu(x)=0. \tag{2.5}$$

Next we make a Gordon reduction of V_μ [always possible for a free field $\psi(x)$] so as to bring out the separate contributions V^c_μ and V^s_μ arising from the ''convective'' and ''spin'' currents, respectively:

$$V^c_\mu(x,p)=\pi^{-4}\int d^4q e^{2ip.q}\overline{\psi}(x+q)\frac{p_\mu}{m}\psi(x-q); \tag{2.6}$$

$$V^s_\mu(x,p)=\pi^{-4}\partial^{(x)}_\nu\int d^4q e^{2ip.q}\overline{\psi}(x+q)\frac{\sigma_{\mu\nu}}{2m}\psi(x-q). \tag{2.7}$$

We check from these that *two* separate conservation equations follow:

$$\partial_\mu V^c_\mu(x,p)=\partial_\mu V^s_\mu(x,p)=0, \tag{2.8}$$

the former being a consequence of the K-G equation for each component of $\psi(x)$ and $\overline{\psi}(x)$, and the latter being identically satisfied.

We now seek to use the Gordon reduction as a key element for deriving the Liouville equation á lá Ref. 4, and in so doing bring out its role in putting our procedure in closer correspondence to Ref. 4. For the central issue is one of demonstrating that the density function $\rho(x,p)$ and the current function $V_i(x,p)$ are related by the ''classical'' velocity factor v_i, so as to ensure the ''correct'' structure of the Liouville equation.[4] To that end we note from (2.3), (2.6), and (2.7) that the effect of Gordon reduction in (3) may be expressed in operator form sandwiched between ψ and $\overline{\psi}$:

$$i\gamma_\mu=\frac{p_\mu}{m}+\frac{1}{2m}\partial^{(x)}_\nu\sigma_{\mu\nu}. \tag{2.9}$$

The spatial part of this equation is seen as the precise counterpart of Eq. (27) of Ref. 4 (after integration w.r.t. d^4pd^4q), with the following identifications: The convective term p_μ in (2.9) has a space part p_i which corresponds to the j_i term of Ref. 4, since it bears the ratio v_i to the time component p_0, as the correct factor connecting the current function $V_i(x,p)$ to the density function $V_0(x,p)$. The space part of the full γ_μ term collectively represents the two terms a_id and $a_i\epsilon$ in Eq. (27) of Ref. 4. Finally the last (derivative) term in (2.9) corresponds to the $\nabla\times\Phi$ term in Eq. (27) of Ref. 4 since it has a zero divergence by itself, as seen from the second part of Eq. (2.8), and therefore has no separate influence on the equation of continuity. Therefore, following the logic of Ref. 4, this term has no effect on the anatomy of the Liouville equation which is thus entirely contained in the first (convective) part of Eq. (2.8). To reveal this structure more explicitly, the $\rho(x,p)$ function of Ref. 4 reads, in our 4×4 Dirac matrix notation, as follows:

$$\rho(x,p)\equiv V_0(x,p)=\pi^{-4}\int d^4q e^{2ip.q}\overline{\psi}(x+q)\gamma_4\psi(x-q), \tag{2.10}$$

which satisfies the Liouville equation

$$\partial_0\rho(x,p)+v_i\partial_i\rho(x,p)=0, \tag{2.11}$$

that is merely a paraphrase of the first part of Eq. (2.8) representing the conservation of the *convective* part of the V_μ function.

To end this section, we outline a construction for the ρ function analogously to Eqs. (51)–(53) of Ref. 4, using a fully covariant procedure. To that end, we write down Fourier representations for the Dirac spinors on the lines of Ref. 12, but adapted to the Euclidean phase convention:[13]

$$\psi(x_1)=(2\pi)^{-3/2}\int d^3p_1(m/E_1)^{1/2}[b_{p_1r_1}u^{r_1}(p_1)e^{ix_1\cdot p_1}+d^*_{p_1r_1}v^{r_1}(p_1)e^{-ix_1\cdot p_1}]; \quad (2.12)$$

$$\overline{\psi}(x_2)=(2\pi)^{-3/2}\int d^3p_2(m/E_2)^{1/2}[b^*_{p_2r_2}\overline{u^{r_2}}(p_2)e^{-ix_2\cdot p_2}+d_{p_2r_2}\overline{v^{r_2}}(p_2)e^{ix_2\cdot p_2}]; \quad (2.13)$$

where b and d are (c-number) electron and positron amplitudes, respectively, and u and v are the corresponding positive and negative energy spinors. The latter are more conveniently expressed in terms of Lorentz boost operators as follows:[12]

$$u^{r_1}(p_1)=[(m-i\gamma\cdot p_1)/(2\sqrt{m(m+E_1)})]u^{r_1}(0); \quad (2.14)$$

$$v^{r_1}(p_1)=[(m+i\gamma\cdot p_1)/(2\sqrt{m(m+E_1)})]v^{r_1}(0). \quad (2.15)$$

These Lorentz boost factors are proportional to energy projection operators for the corresponding spinors. To specify their spin states, we can further multiply these quantities by the spin projection operators P_n,[12]

$$P_n=[1+i\gamma\cdot\hat{n}\gamma_5]/2, \quad (2.16)$$

where $\hat{n}$ is a unit 4-vector representing the direction of spin.[12] After these manipulations on the spinors through the sequence of Eqs. (2.10)–(2.16), and their substitution in the defining Eq. (2.3) for V_μ, the spinor dependence can be totally removed by ''tracing.'' To simplify these expressions note that only the b^*b and d^*d terms will survive, while the cross terms will drop out of the traces, since the $\pm$ energy spinors at a given momentum (zero) are orthogonal to each other. The rest of the simplification is routine and the result can be compactly expressed in terms of two quantities $Tr^{\pm}_{\mu}$ defined as follows:

$$Tr^{\pm}_{\mu}=\frac{\pm m(p_1+p_2)_\mu+\epsilon_{\mu\lambda\rho\sigma}p_{1\lambda}p_{2\rho}\hat{n}_\sigma}{2\sqrt{E_1(m+E_1)E_2(m+E_2)}}. \quad (2.17)$$

Before substituting these quantities in Eq. (2.3) for $V_\mu(x,p)$, note that of the two terms on the right, the first (convective) term has the desired structure for the ''Liouville form,'' Eq. (2.11), but the second (spin) term has not. However, in accordance with the discussion immediately following Eq. (2.9), in line with Ref. 4, one need not insist on the ''Liouville connection'' separately for both the terms since the ''spin'' term (proportional to $\hat{n}$) can be traced back to the ''derivative'' term of (2.9) which has no separate effect on the equation of continuity. Therefore we need only the $\mu=4$ component in Eq. (2.17) above to extract the ρ-function fully, whence the corresponding ''current'' function V_i can be generated by multiplying with $v_i=p_i/p_0$ only.

To proceed further, integration over d^3p_1 and d^3p_2 may be simplified via the transformations $p_{1,2}=p'\pm k/2$, remembering that $x_{2,1}=x\pm q$. Further, the 3D integrals may be converted to 4D integrations over p' and k via the identity

$$d^3p_1d^3p_2=2E_12E_2d^4p'd^4k\delta(p'^2+m^2+k^2/4)\delta(2p'\cdot k), \quad (2.18)$$

making use of the result $\delta(a)\delta(b)=\delta((a+b)/2)\delta(a-b)$. Integration over d^4q now gives $\delta^4(p'\mp p)$ for the terms b^*b and d^*d, respectively; subsequent integration over d^4p' gives $p'=\pm p$. Collecting all these results and substituting in (2.10) gives

$$\rho(x,p)=\pi^{-3}\int d^4k\,\delta(p^2+m^2+k^2/4)\,\delta(2p\cdot k)[b^*(\mathbf{p}+\mathbf{k}/2)b(\mathbf{p}-\mathbf{k}/2)e^{ik.x}$$
$$+d^*(\mathbf{p}+\mathbf{k}/2)d(\mathbf{p}-\mathbf{k}/2)e^{-ik\cdot x}](2mp_0+2i\hat{n}.(\mathbf{p}\times\mathbf{k}))\sqrt{\frac{E_1E_2}{(m+E_1)(m+E_2)}}. \tag{2.19}$$

This expression corresponds to Eqs. (51)–(53) of Ref. 4, after taking due account of the difference in the choice of the respective basis functions (Dirac versus Pauli types), as well as the covariant versus non-covariant notations. Note that the integration in (2.19) is basically angular $d\Omega_k$ in content, like in Ref. 4, since the four-dimensional (4D) integration d^4k is also multiplied by the two δ-functions which take care of the magnitudes of both $\mathbf{k}$ and k_0. Our 3-vector $\mathbf{k}$ is $2\mathbf{k}$ of Ref. 4, and $\hat{n}$ is $\hat{\omega}$ of Ref. 4. Finally, our expression for ρ compactly covers both $\pm$ energy components.

We next turn to the K-G and Proca cases via Kemmer-Duffin equations.

III. WIGNER FUNCTION FOR INTEGRAL SPIN

Because of the relative lack of exposure of the Kemmer-Duffin equations[8] in the contemporary literature, it will first be useful to recall some basic results which are summarized in Roman[13] in the Euclidean phase convention, from which we shall freely draw on several results without much comment. The first order differential equation for integral spin is expressible in β-matrix notation as[13]

$$(\beta_\mu\partial_\mu+m)\psi(x)=0;\overline{\psi}(m-\partial_\mu\beta_\mu)=0, \tag{3.1}$$

where $\psi(x)$ is a multi-component field, each of whose components ψ_i satisfies the K-G equation, and $\overline{\psi}=\psi^\dagger\eta_4$ and $\eta_4=2\beta_4^2-1$. The spin-0 and spin-1 cases correspond to 5- and 10-component irreducible representations, respectively. These components are

$$\begin{aligned}&\text{Spin}-0:\psi_i(i=1-4)=-\partial_\mu\phi(x)/m;\quad \psi_5=\phi,\\ &\text{Spin}-1:\psi_i(i=1-6)=-F_{\mu\nu}/m;\quad \psi_i(i=1-4)=\phi_\mu.\end{aligned} \tag{3.2}$$

The various quantities like bilocal 4-current $j_\mu(x_1,x_2)$, phase space 4-current $V_\mu(x,p)$, and local 4-current $J_\mu(x)$ follow very closely the corresponding constructions (1.2)–(1.5) for the Dirac case, with the replacements: γ_μ by β_μ; γ_4 by η_4. Thus the phase space 4-current is given by

$$V_\mu(x,p)=\pi^{-4}\int d^4q e^{2ip.q}\overline{\psi}(x+q)i\beta_\mu\psi(x-q), \tag{3.3}$$

and its break-up into convective and spin components is achieved via ''Gordon reduction'' analogously to the Dirac case. However the algebra of the β matrices is a bit more involved than that of the γ matrices because the former are singular. Thus[8,13]

$$\beta_\lambda\beta_\mu\beta_\nu+\beta_\nu\beta_\mu\beta_\lambda=\beta_\lambda\delta_{\mu\nu}+\beta_\nu\delta_{\mu\lambda};\quad \eta_4^2=1. \tag{3.4}$$

The ''spin'' operator $S_{\mu\nu}$ is given by

$$iS_{\mu\nu}=\beta_\mu\beta_\nu-\beta_\nu\beta_\mu, \tag{3.5}$$

and, except for a factor of $\frac{1}{2}$, has exactly the same role as that of $\sigma_{\mu\nu}$, including its antisymmetry property. For writing down the counterpart of Gordon reduction, the following equations are needed:

$$(\partial_\mu-\beta_\nu\beta_\mu\partial_\nu)\psi(x)=0;\quad \overline{\psi}(x)(\partial_\mu-\beta_\nu\beta_\mu\partial_\nu)=0. \tag{3.6}$$

With the help of (3.1) and (3.6), and making use of the definition (3.5) for the ''spin,'' the matrix β_μ in (3.3) may be written in a form analogous to (2.9) as

$$i\beta_\mu = \frac{p_\mu}{m} + \partial_\nu^{(x)} \frac{S_{\mu\nu}}{m}, \tag{3.7}$$

which shows a clear division between the convective and spin parts, with the latter having a ''derivative'' structure whose divergence vanishes by itself, as in the second part of Eq. (2.8). Therefore this term has no direct role in the Liouville equation like in the Dirac case, exactly as in Ref. 4.

So far this β-formalism for integral spin covers both spin-0 (K-G) and spin-1 (Proca) cases, and one may wonder why for the former there should be a spin term $S_{\mu\nu}$ at all, in the Gordon reduction (3.7). Actually this term is effectively *zero* for the spin-0 case since, as has been explained in Roman,[13] for the 5×5 representation the eigenvalues of $S_3 = S_{12}$ are nonmeasurable. Indeed the probability density,[13]

$$\rho = -i(\psi_4^*\psi_5 + \psi_5^*\psi_4) = -i(\partial_0\phi^*\phi - \phi^*\partial_0\phi)/m, \tag{3.8}$$

vanishes in the eigenstates that correspond to $S_3 = \pm 1$, so that the presence of the $S_{\mu\nu}$-term in the spin-0 case is purely symbolic. Substituting this information in (3.3), and noting from (3.8) that only the components $\psi_{4,5}$ survive for the phase space density function $\rho(x,p)$ for the K-G case, the final result works out as

$$\rho(x,p) = \frac{1}{m}\pi^{-4} i \int d^4q e^{2ip.q}[\partial_0\phi^*(x+q)\phi(x-q) - \phi^*(x+q)\partial_0\phi(x-q)], \tag{3.9}$$

while Eq. (3.7) ensures that the phase space current function $\mathbf{V}(x,p)$ is related to $\rho(x,p)$ by the factor $\mathbf{p}/p_0 = \mathbf{v}$, so that the structure of the Liouville equation is preserved. As a further check, the integration of (3.9) over d^4p reproduces the K-G density function $\rho(x)$ of (3.8) exactly. The last quantity is of course not positive definite, as expected[13] for the K-G case. Again as in the Dirac case, it is possible to write down an expression for $\rho(x,p)$ closely analogous to (2.19), except for the absence of the ''spin'' term, in the third line. We omit this routine step for brevity.

For the Proca case, the ''spin'' term in the Gordon reduction (3.7) will now contribute to the $\rho(x,p)$ function, while the Liouville connection is again maintained via the convective p_μ term only. From the structure of $\rho(x,p)$, which corresponds to $\mu = 4$, it is seen quite clearly that the ''spin'' contribution arising from the second term of Eq. (3.7) in the Proca case is proportional to $\mathbf{S}\cdot\mathbf{E}\times\mathbf{H}$, where $\mathbf{E}$ and $\mathbf{H}$ are the ''electric'' and ''magnetic'' components of $F_{\mu\nu}$, respectively. Again we skip the Fourier representation (2.19) for the full $\rho(x,p)$ function, as this step will not illuminate the structure any further beyond the above demonstration of a close parallelism between the spin-1/2 and integral spin (0,1) cases.

IV. DISCUSSION AND CONCLUSION

This study was motivated by a desire to understand some results due to Bosanac[4] in which he claims to derive the Dirac equation, starting from the (classical) premises of the Liouville equation. What intrigued us was the assertion[4] that the ''derivation'' leads to the Dirac equation, instead of to the Klein-Gordon equation. To examine this question more closely, we were led to start from the *opposite* direction to Ref. 4, viz., a ''direct'' derivation of the Liouville equation through an explicit construction of the Wigner Functions[7] themselves, as in more orthodox approaches,[5,6] with a view to throwing some light on this apparent asymmetry between the Dirac and the K-G cases. Rather surprisingly, our result indicates a close parallelism between both spin-1/2 and spin-(0,1) cases, brought about by the Kemmer-Duffin[8] formalism[13] for integral spin as a natural counterpart to the Dirac equation for spin-1/2. In both cases, we have made essential use of the Gordon reduction into ''convective'' and ''spin'' contributions to the 4-current, wherein the convective part gives the desired ''Liouville ratio'' p_0:$\mathbf{p}$ between the density and current functions in phase space, but the spin part does not. However the latter has a derivative structure by virtue of which its 4-divergence vanishes identically, so as to yield a ''spin'' current conser-

vation by itself [Eq. (2.8)], and hence has no direct effect on the dynamics of the Liouville equation as such, which is entirely governed by the convective term only. This result seems to accord rather well with the corresponding result of Bosanac,[4] viz., his Eq. (27), despite a vast difference in the respective procedures. In particular, since the Gordon reduction in our derivation has proved crucial for a neat identification of the origin of these two terms as convective and spin contributions, respectively, it may be safely assumed that the same mechanism underlies the derivation in Ref. 4 as well, irrespective of the manner of the derivation.

Without going into the philosophical aspects of the classical foundations of Quantum Mechanics, it is pertinent to ask why only the Dirac equation has been amenable to the Bosanac derivation.[4] For it will be rather naive to argue that no substantial quantum mechanical input has gone into the latter approach, since the very structure of the Wigner Function as a bilinear in the (quantum mechanical) wave function is a tacit acknowledgement of the existence of the latter, with the Uncertainty Principle as a built-in corollary of the same. Further, the structure of the Dirac equation is implicit in the ansatz,[4] Eq. (16), in terms of Pauli spinors F_1, F_2, with anticommutation relations (23)–(25) characteristic of Dirac matrices in the Pauli 2-component form, all of which ingredients have been crucial to the (indirect) derivation of the main Eq. (27). Indeed the entire exercise[4] is a fine example of how to work backwards from the (bilinear) Liouville equation to the (linear) Dirac equation, albeit in the Pauli 2-component form. Yet it goes without saying that the ansatz of a 4-component spinor [satisfying a Dirac-type equation (16)[4]] has been very much of an input in the derivation.

A comparison with our ''direct'' derivation would suggest that a possible reason why the K-G equation was not amenable to the Bosanac derivation was the missing element of a multi-component wave function in his formulation of the K-G problem, something we have sought to rectify through the Kemmer-Duffin β-matrix formalism[8] which restores the formal analogy to the Dirac case. However, and this may be important, even if the Bosanac formalism had included a multi-component wave function for the integral (including zero) spin field, it is doubtful if working backwards from the Liouville to the K-G equation would have proved successful in this case, since the β-matrices are *singular*, unlike the Dirac matrices, so that a routine inversion of steps on the lines of 4 might well not have been possible (we did not succeed in this effort). But there seems to be no problem in the ''forward'' steps (from QM equation to Liouville) going through in the K-G and Proca cases with equal ease, as the above derivation indicates.

The foregoing comparison would suggest that the structure of Quantum Mechanics is probably richer and more varied than can be fully captured in one go, through alternative formulations. And yet the basically indeterministic nature of QM poses a continuous challenge to innovative ideas designed to capture its full flavor from ''objective-realist'' premises, so as to provide a more ''classical'' approach to its uncanny predictions. Reference 4 may be regarded as one more effort in this direction, as part of an ongoing movement[14] for resolving the foundational paradoxes of orthodox QM which has acquired fresh momentum with heavyweights for and against the Copenhagen Interpretation arrayed in comparable strengths on both sides of the Great Divide.[15]

[1] E. Nelson, *Quantum Fluctuations* (Princeton University Press, Princeton, 1987).
[2] A. Kyprianidis, Phys. Rep. **115C**, 1 (1987).
[3] R. K. Varma, in *Microphysical Reality and Quantum Formalism*, edited by A. van der Merwe *et al.* (Kluwer Academic, New York, 1988), pp. 175–194.
[4] S. D. Bosanac, J. Math. Phys. **38**, 3895 (1997).
[5] P. Carruthers and F. Zachariasen, Rev. Mod. Phys. **55**, 245 (1983).
[6] G. R. Shin and J. Rafelski, Ann. Phys. (N.Y.) **243**, 65 (1995).
[7] E. P. Wigner, Phys. Rev. **40**, 749 (1932).
[8] N. Kemmer, Proc. R. Soc. London, Ser. A **175A**, 91 (1939).
[9] F. J. Dyson, *Lectures on Quantum Mechanics* (Laboratory of Nuclear Studies, Cornell University, 1952).
[10] W. Pauli and V. F. Weisskopf, Helv. Phys. Actra **7**, 709 (1934).
[11] R. P. Feynman, Phys. Rev. **76**, 749 (1949).
[12] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), pp. 51–63.
[13] See, for a summary, P. Roman, *Theory of Elementary Particles* (North Holland, Amsterdam, 1964), pp. 146–155.
[14] See, e.g., G. C. Girardi *et al.*, Phys. Today **46**, 13 (1996).
[15] See, e.g., S. W. Hawking and R. Penrose, *Nature of Space and Time* (Princeton University Press, Princeton, 1996).

Two-component formulation of the Wheeler–DeWitt equation

Ali Mostafazadeh[a)]
Theoretical Physics Institute, University of Alberta, Edmonton, Alberta T6G 2J1, Canada and Department of Mathematics, Koç University, Istinye 80860, Istanbul, Turkey

(Received 30 January 1997; accepted for publication 27 April 1998)

The Wheeler–DeWitt equation for the minimally coupled Friedman–Robertson–Walker-massive-scalar-field minisuperspace is written as a two-component Schrödinger equation with an explicitly ''time''-dependent Hamiltonian. This reduces the solution of the Wheeler–DeWitt equation to the eigenvalue problem for a nonrelativistic one-dimensional harmonic oscillator and an infinite series of trivial algebraic equations whose iterative solution is easily found. The solution of these equations yields a mode expansion of the solution of the original Wheeler–DeWitt equation. Further analysis of the mode expansion shows that in general the solutions of the Wheeler–DeWitt equation for this model are doubly graded, i.e., every solution is a superposition of two definite-parity solutions. Moreover, it is shown that the mode expansion of both even- and odd-parity solutions is always infinite. It may be terminated artificially to construct approximate solutions. This is demonstrated by working out an explicit example which turns out to satisfy DeWitt's boundary condition at initial singularity. © *1998 American Institute of Physics.* [S0022-2488(98)01209-2]

I. INTRODUCTION

It is well-known that the Wheeler–DeWitt equation for the minimally coupled Friedman–Robertson–Walker (FRW)-massive-scalar-field (the scalar field is assumed to be homogeneous) minisuperspace can be written in the form of the Klein–Gordon equation (this equation corresponds to simplest choice of factor ordering) in $(1+1)$ dimensions,[1]

$$\left[-\frac{\partial^2}{\partial\alpha^2}+\frac{\partial^2}{\partial\phi^2}+\kappa e^{4\alpha}-m^2e^{6\alpha}\phi^2\right]\psi(\alpha,\phi)=0, \tag{1}$$

where $\alpha=\ln a$, a is the scale factor of the FRW metric

$$ds^2=-dt^2+a^2(t)[dr^2+S_\kappa(r)(d\theta^2+\sin^2\theta d\varphi^2)],$$

$$S_\kappa:=\begin{cases}\sinh r & \text{for } \kappa=-1: \text{ open universe}\\ r & \text{for } \kappa=0: \text{ flat universe}\\ \sin r & \text{for } \kappa=1: \text{ closed universe,}\end{cases}$$

and ϕ is a real scalar field of mass m. It is also well-known that unlike the massless case[2] exact solutions of this equation are not known. The approximate solutions are usually obtained using the semiclassical Wentzel–Kramers–Brillouin (WKB) methods.[3]

Recently, I have used a two-component formulation of the Klein–Gordon equation to study the geometrical phases induced on a Klein–Gordon field due to a dynamical background spacetime geometry and electromagnetic or scalar interactions.[4] In the present article, I shall try to apply the same approach for the treatment of the Wheeler–DeWitt equation (1).

a)Electronic mail: amostafazadeh@ku.edu.tr

 © 1998 American Institute of Physics

II. TWO-COMPONENT FORM OF THE WHEELER–DEWITT EQUATION

The idea of the two-component formalism[5] is to transform the Klein–Gordon equation which is of second order in time-derivative, into a system of two coupled differential equations which are first order in the time-derivative. The latter may be viewed as a Schrödinger equation. In this way one can apply the well-developed techniques of solving the Schrödinger equation to obtain solutions of the Klein–Gordon equation. The application of the two-component formalism for the Klein–Gordon equation is described in detail in Ref. 4. Here I only include the results which are relevant for the solution of Eq. (1).

Equation (1) can be expressed in the form:

$$\left[\eta^{\mu\nu}\frac{\partial}{\partial X^{\mu}}\frac{\partial}{\partial X^{\nu}}+V(X)\right]\psi=0, \tag{2}$$

where $\mu,\ \nu=0,\ 1,\ (\eta^{\mu\nu}):=\mathrm{diag}(-1,1),\ X^0:=\alpha,\ X^1:=\phi,$ and

$$V(X):=\kappa e^{4X^0}-m^2e^{6X^0}(X^1)^2=\kappa e^{4\alpha}-m^2e^{6\alpha}\phi^2=\kappa a^4-m^2a^6\phi^2. \tag{3}$$

Hence, α plays the role of ''time.'' Furthermore, defining

$$\Psi:=\frac{1}{\sqrt{2}}\begin{pmatrix}\psi+i\dot{\psi}\\ \psi-i\dot{\psi}\end{pmatrix}, \tag{4}$$

one can easily check that Eq. (2) is equivalent to the Schrödinger equation

$$i\dot{\Psi}=H\Psi, \tag{5}$$

with the explicitly time-dependent Hamiltonian:

$$H=\frac{1}{2}\begin{pmatrix}1+D & -1+D\\ 1-D & -1-D\end{pmatrix}, \tag{6}$$

where a dot represents differentiation with respect to $X^0=\alpha$ and

$$D:=-\frac{\partial^2}{\partial(X^1)^2}-V=-\frac{\partial^2}{\partial\phi^2}-V. \tag{7}$$

The two-component Hamiltonian (6) has the following remarkable properties.

(1) It belongs to the (dynamical) Lie algebra $\mathrm{su}_s(1,1)\otimes\mathrm{su}_u(1,1)$. Here $\mathrm{su}_s(1,1)$ is the standard representation of su(1,1) with the basis

$$K_3=J_3=\frac{1}{2}\sigma_3,\quad K_a=iJ_{\alpha}=\frac{i}{2}\sigma_a\quad (a=1,2),$$

where σ_i are Pauli matrices, and $\mathrm{su}_u(1,1)$ denotes the unitary representation of su(1,1) constructed out of the position x and momentum p operators of the one-dimensional quantum mechanics, with the basis $\{(x^2-p^2)/4,(x^2+p^2)/4,(xp+px)/4\}$.[6] This is easily seen, by writing H in the form $H=K_-+K_+D$, where $K_{\pm}:=K_3\pm K_2$, and observing that the operator D belongs to $\mathrm{su}_u(1,1)$.

(2) It is the square root of the operator D, in the sense that it satisfies

$$H^2=\begin{pmatrix}D & 0\\ 0 & D\end{pmatrix}=D. \tag{8}$$

This is reminiscent of Dirac's treatment of the noninteracting Klein–Gordon equation in $(3+1)$-dimensional Minkowski space, where the Dirac operator plays the role of the square root of the Klein–Gordon operator.

Next let us examine the eigenvalues E_n and eigenvectors Ψ_n of the Hamiltonian H. In order to have a well-posed eigenvalue problem for H, one should discuss the properties of the space of

two-component wave functions $\mathscr{H}$. This will be done in Sec. III. Here we suffice to note that as a vector space $\mathscr{H}$ is the inner sum of two vector subspaces each of which is isomorphic (as Hilbert spaces) to $L^2(\mathbb{R})$. In view of this observation and the particular form of the Hamiltonian H, the formal solution of the eigenvalue equation

$$H\Psi_n=E_n\Psi_n, \tag{9}$$

for H is given by

$$\Psi_n=\frac{1}{\sqrt{2}}\begin{pmatrix}1+E_n\\1-E_n\end{pmatrix}\Phi_n, \quad D\Phi_n=E_n^2\Phi_n. \tag{10}$$

This shows that the eigenvalue equation (9) for $H:\mathscr{H}\to\mathscr{H}$ is in fact equivalent to the eigenvalue equation (10) for $D:L^2(\mathbb{R})\to L^2(\mathbb{R})$. Since the latter is a a well-posed eigenvalue equation in the Hilbert space $L^2(\mathbb{R})$, so is the former. In particular, the eigenvectors Ψ_n do belong to $\mathscr{H}$.

Next let use substitute Eq. (7) in the second equation in (10). This yields

$$\left[-\frac{1}{2}\frac{\partial^2}{\partial\phi^2}+\frac{1}{2}\omega^2\phi^2\right]\Phi_n=\mathscr{E}_n\Phi_n, \tag{11}$$

where $\omega:=me^{3\alpha}=ma^3$ and $\mathscr{E}_n:=(E_n^2+\kappa e^{4\alpha})/2=(E_n^2+\kappa a^4)/2$. Equation (11) is the eigenvalue equation for a simple harmonic oscillator with frequency ω and unit mass. Hence:

$$\mathscr{E}_n=\omega(n+\tfrac{1}{2})=ma^3(n+\tfrac{1}{2}), \tag{12}$$

$$\Phi_n=N_nH_n(\sqrt{\omega}\phi)e^{-\frac{1}{2}\omega\phi^2}=N_nH_m(\sqrt{ma^3}\phi)e^{-\frac{1}{2}ma^3\phi^2}, \tag{13}$$

where H_n are Hermite polynomials and N_n are normalization constants. The latter are determined by the requirement that Φ_n be normalized with the usual $L^2(\mathbb{R})$ norm. This leads to

$$N_n=\left[\frac{\omega}{\pi(2^nn!)^2}\right]^{1/4}=\left[\frac{ma^3}{\pi(2^nn!)^2}\right]^{1/4}.$$

Note that Φ_n and consequently Ψ_n tend to zero as $a\to0$ or $a\to\infty$.

Substituting Eq. (12) in the expression for $\mathscr{E}_n$, one has

$$E_{\pm n}=\pm\sqrt{\omega(2n+1)-\kappa e^{4\alpha}}=\pm\sqrt{me^{3\alpha}(2n+1)-\kappa e^{4\alpha}}=\pm a\sqrt{ma(2n+1)-\kappa a^2}. \tag{14}$$

Thus, for $\kappa=-1$ or 0, $E_{\pm n}$ are real, whereas for $\kappa=1$ and $a>m$, there are eigenvectors $\Psi_{\pm n}$ with imaginary eigenvalue. They correspond to the integers n satisfying $m(2n+1)<a$.

So far, I have given the solution for the eigenvalue problem for the Hamiltonian H. The eigenvectors of H, however, do not solve the Schrödinger equation. This is mainly because H is explicitly time dependent. The general solution of the Schrödinger equation for an arbitrary explicitly time-dependent Hamiltonian is not known. However, in the cases where the eigenvalue problem for the Hamiltonian is solvable, there are techniques based on the adiabatic approximation[7–9] which may be used to obtain approximate and occasionally exact solutions. These techniques can be employed if the Hamiltonian is self-adjoint. In order to examine the self-adjointness of the Hamiltonian H, one first needs a notion of an inner product on the space $\mathscr{H}$ of two-component wave functions Ψ.

III. HILBERT SPACE STRUCTURE OF $\mathscr{H}$

As a vector space, the space $\mathscr{H}$ of two-component wave functions is isomorphic to $L^2(\mathbb{R})\oplus L^2(\mathbb{R})$. Here by the vector space $L^2(\mathbb{R})\oplus L^2(\mathbb{R})$, I mean the vector space inner sum of two copies of $L^2(\mathbb{R})$, not the Hilbert space inner sum of these spaces where the sum is also endowed with a canonical inner product. Furthermore, in what follows I allow for negative norms and the terms ''inner product'' and ''Hilbert space'' refer to a Hermitian bilinear (sesquilinear) form and

a vector space endowed with such a form, respectively. [A Hermitian bilinear or sesquilinear form on a complex vector space V is a function $\mu: V\times V\to\mathbb{C}$ which is linear in its second variable and satisfies $\mu(v_1,v_2)=\mu(v_2,v_1)^*$ where* denotes complex conjugation.]

The relation $\mathscr{H}=L^2(\mathbb{R})\oplus L^2(\mathbb{R})$ may be viewed as a restriction on the solutions of the Wheeler–DeWitt equation, since for finite α the elements of $L^2(\mathbb{R})\oplus L^2(\mathbb{R})$ vanish at ''spatial'' infinity, i.e., $\phi=\pm\infty$. This choice seems, however, to be a natural and at the same time a rather general one. It is only restrictive as far as the boundary condition at spatial infinity is concerned. It is still quite general, for it does not correspond to any fixed boundary conditions at $\alpha=\pm\infty$. Furthermore, although $\mathscr{H}$ has two vector subspaces with L^2 inner products, an inner product on $\mathscr{H}$ is not *a priori* fixed.

An inner product (sesquilinear form) on $\mathscr{H}$ may be given by the inner product on $L^2(\mathbb{R})$ and a Hermitian two-by-two matrix h. A particularly advantageous choice for h is $h=\sigma^3=\mathrm{diag}(1,-1)$, i.e.,

$$(\Psi,\tilde{\Psi}):=\langle\Psi^1|\tilde{\Psi}^1\rangle-\langle\Psi^2|\tilde{\Psi}^2\rangle, \tag{15}$$

where $\langle\ |\ \rangle$ denotes the inner product on $L^2(\mathbb{R})$. This choice, on the one hand, renders H ''self-adjoint'' in the sense that for all Ψ, $\tilde{\Psi}\in\mathscr{H}$,

$$(\Psi,H\tilde{\Psi})=(H\Psi,\tilde{\Psi}). \tag{16}$$

[Note that the adjoint of a linear operator is well-defined if one has an inner product which defines a positive definite norm. This is because in the definition of the adjoint of a linear operator one uses the Riesz lemma whose proof makes use of the positive definiteness of the norm.[10] Here the term ''self-adjoint'' refers to the validity of Eq. (16).] On the other hand, it reduces to the Klein–Gordon inner product (for more details see Ref. 4) for the solutions (4) of the Schrödinger equation (5).

Adopting this inner product on $\mathscr{H}$, one can compute

$$(\Psi_{\pm m},\Psi_{\pm n})=(E^*_{\pm m}+E_{\pm n})\,\delta_{mn}\,. \tag{17}$$

Here I have used (10) and the fact that Φ_n are orthonormal. In particular, $(\Psi_{\pm n},\Psi_{\pm n})=2\,\mathrm{Re}(E_{\pm n})=\pm 2|\mathrm{Re}(E_n)|$. Hence, the eigenvectors $\Psi_{\pm n}$ with imaginary eigenvalues (if they exist) have zero norm; they are null vectors. This is of course to be expected since the Hamiltonian satisfies Eq. (16). (Note that it is the nonpositive definiteness of the corresponding norm that allows for imaginary eigenvalues and null eigenvectors to exist.) Furthermore, note that the null eigenvectors are only present for the $\kappa=1$ case.

Another interesting consequence of Eq. (17) is that

$$(\Psi_{\pm n},\Psi_{\mp n})=\begin{cases}0 & \text{for } E_n \text{ real}\\ E_{\mp n} & \text{for } E_n \text{ imaginary.}\end{cases} \tag{18}$$

Hence, if we view $\mathscr{H}$ as the linear span of $\Psi_{\pm n}$, it can be orthogonally decomposed as

$$\mathscr{H}=\mathscr{H}^{(-)}\oplus\mathscr{H}^{(0)}\oplus\mathscr{H}^{(+)}, \tag{19}$$

where $\mathscr{H}^{(-)}$, $\mathscr{H}^{(0)}$, and $\mathscr{H}^{(+)}$ are the linear span of the negative-norm, null, and positive-norm eigenvectors. Note that this decomposition is time dependent (it depends on $a=e^\alpha$), and not every null vector belongs to $\mathscr{H}^{(0)}$. Furthermore, Eq. (18) implies that the basis $\{\Psi_n\}$ (respectively, $\{\Psi_{-n}\}$) of $\mathscr{H}^{(+)}$ (respectively, $\mathscr{H}^{(-)}$) is orthogonal, whereas $\mathscr{H}_0$ can be orthogonally decomposed into the sum of two-dimensional subspaces $\mathscr{H}^{(0)}_n=\mathrm{Span}(\Psi_n\oplus\Psi_{-n})$ with $n\leqslant(a/m-1)/2$, namely

$$\mathscr{H}^{(0)}=\bigoplus_{n=1}^{N}\mathscr{H}^{(0)}_n, \tag{20}$$

where $N:=[(a/m-1)/2]$ is the largest integer part of $(a/m-1)/2$.

Although the inner product (15) has a number of interesting properties, in the remainder of this article I shall not assume that $\mathscr{H}$ is endowed with this inner product. In fact, the results presented in the following sections are independent of the choice of the 2×2 matrix h which determines the inner product on $\mathscr{H}$. I shall only use the well-known Hilbert space properties of the space $L^2(\mathbb{R})$.

IV. ADIABATIC APPROXIMATION

Consider the ansatz

$$\Psi = e^{i\zeta_{\mathscr{N}}}\Psi_{\mathscr{N}}, \tag{21}$$

for the solution of the Schrödinger equation (5), where $\zeta_{\mathscr{N}}$ is a complex-valued function of α, and $\mathscr{N}=\pm n\in\mathbb{Z}$. Plugging (21) in (5) and making use of the eigenvalue equation (9) and the form (10) of the eigenvectors $\Psi_{\mathscr{N}}$, one has:

$$[-(\dot{\zeta}_{\mathscr{N}}+E_{\mathscr{N}})(1+E_{\mathscr{N}})+i\dot{E}_{\mathscr{N}}]\Phi_n+i(1+E_{\mathscr{N}})\dot{\Phi}_n=0,$$

$$[-(\dot{\zeta}_{\mathscr{N}}+E_{\mathscr{N}})(1-E_{\mathscr{N}})-i\dot{E}_{\mathscr{N}}]\Phi_n+i(1-E_{\mathscr{N}})\dot{\Phi}_n=0.$$

Adding and subtracting both sides of these equations, one finds:

$$(\dot{\zeta}_{\mathscr{N}}+E_{\mathscr{N}})\Phi_n-i\dot{\Phi}_n=0,\quad \dot{E}_{\mathscr{N}}=0. \tag{22}$$

Here the first equation implies that for $\tilde{n}\neq n$, $A_{\tilde{n}n}:=i\langle\Phi_{\tilde{n}}|\dot{\Phi}_n\rangle=0$. This requirement together with the second relation in (22) signify the conditions under which the adiabatic approximation would yield an exact solution of the Schrödinger equation. Clearly, $\dot{E}_{\mathscr{N}}\neq 0$ for the case under consideration. Thus the adiabatic approximation cannot be exact. However, approximate solutions may be sought if $\dot{E}_{\mathscr{N}}$ and $A_{\tilde{n}n}$ are small.

Performing the necessary calculations, one can show that for arbitrary non-negative integers $\tilde{n}$ and n:

$$A_{\tilde{n}n}:=i\langle\Phi_{\tilde{n}}|\dot{\Phi}_n\rangle=\frac{3i}{4}\,[\sqrt{n(n-1)}\,\delta_{\tilde{n},n-2}-\sqrt{(\tilde{n})(\tilde{n}-1)}\,\delta_{\tilde{n},n+2}], \tag{23}$$

$$\dot{E}_{\mathscr{N}}=\pm\frac{3\left[n+\frac{1}{2}-\frac{2\kappa a}{3m}\right]\sqrt{m}a^{3/2}}{\sqrt{2n+1-\frac{\kappa a}{m}}}\quad\text{where } \mathscr{N}=\pm n. \tag{24}$$

Equation (23) indicates that the adiabatic approximation is not valid, i.e., the solutions of the form (21) do not exist.

Another consequence of Eq. (23) is that in a mode expansion of a solution Ψ of Eq. (5), the even modes $\Psi_{2\mathscr{N}}$ and the odd modes $\Psi_{2\mathscr{N}+1}$ do not mix in the course of evolution. This suggests the existence of a double grading of the set of solutions. In particular one has the *even* and *odd* solutions:

$$\Psi_{\text{even}}=\sum_{\mathscr{N}=-\infty}^{\infty}Z_{2\mathscr{N}}\Psi_{2\mathscr{N}},\quad \Psi_{\text{odd}}=\sum_{\mathscr{N}=-\infty}^{\infty}Z_{2\mathscr{N}+1}\Psi_{2\mathscr{N}+1}. \tag{25}$$

V. MODE EXPANSION OF THE SOLUTIONS

Consider expressing the solution Ψ of the two-component Wheeler–DeWitt equation (5) as a linear combination of the eigenfunctions $\Psi_{\mathscr{N}}$ of the Hamiltonian:

$$\Psi=\sum_{\mathcal{N}=-\infty}^{\infty} Z_{\mathcal{N}}\Psi_{\mathcal{N}},\quad Z_{\mathcal{N}}=Z_{\mathcal{N}}(\alpha). \tag{26}$$

Substituting this equation in (5) and using Eqs. (9) and (10) to simplify the result, one finds

$$\sum_{n=0}^{\infty}\left[(\dot{\sigma}_n+i\sigma_{-n}E_n)\Phi_n+\sigma_n\dot{\Phi}_n\right]=0, \tag{27}$$

$$\sum_{n=0}^{\infty}\left[(\dot{\sigma}_{-n}E_n+i\sigma_nE_n^2+\sigma_{-n}\dot{E}_n)\Phi_n+\sigma_{-n}E_n\dot{\Phi}_n\right]=0, \tag{28}$$

where

$$\sigma_{\pm n}=\sigma_{\pm n}(\alpha):=Z_n\pm Z_{-n},\quad n=0,1,2,\ldots .$$

Note that in general $Z_{-0}\neq Z_0$ and $\sigma_{-0}\neq\sigma_0$.

Next take the inner product of both sides of Eqs. (27) and (28) with $\Phi_{\tilde{n}}$. Making use of Eq. (23) and orthonormality of Φ_n, one has:

$$\dot{\sigma}_n+iE_n\sigma_{-n}+\tfrac{3}{4}\sqrt{(n+1)(n+2)}\sigma_{n+2}-\tfrac{3}{4}\sqrt{n(n-1)}\sigma_{n-2}=0, \tag{29}$$

$$E_n\dot{\sigma}_{-n}+iE_n^2\sigma_n+\dot{E}_n\sigma_{-n}+\tfrac{3}{4}\sqrt{(n+1)(n+2)}E_{n+2}\sigma_{-(n+2)}-\tfrac{3}{4}\sqrt{n(n-1)}E_{n-2}\sigma_{-(n-2)}=0, \tag{30}$$

where $\sigma_{-(0)}=\sigma_{-0}$.

In order to analyze these equations further, let us solve for σ_{-n} from (29) and substitute the result in (30). Doing the same for $\sigma_{-(n\pm2)}$, one obtains:

$$\sigma_{-n}=\frac{i}{E_n}\left[\dot{\sigma}_n+\frac{3}{4}\sqrt{(n+1)(n+2)}\sigma_{n+2}-\frac{3}{4}\sqrt{n(n-1)}\sigma_{n-2}\right], \tag{31}$$

$$\begin{aligned}\ddot{\sigma}_n+[E_n^2-\tfrac{3}{2}(n^2+n+1)]\sigma_n+\tfrac{3}{2}\sqrt{(n+1)(n+2)}\dot{\sigma}_{n+2}-\tfrac{3}{2}\sqrt{n(n-1)}\dot{\sigma}_{n-2}\\ +(\tfrac{3}{4})^2\sqrt{(n+1)(n+2)(n+3)(n+4)}\sigma_{n+4}+(\tfrac{3}{4})^2\sqrt{n(n-1)(n-2)(n-3)}\sigma_{n-4}=0.\end{aligned} \tag{32}$$

Note that Eq. (32) links only σ_{2n}'s, with $n=0,1,\ldots$, thus making the analysis of the coefficients of non-negative modes independent of those of the negative modes.

Equations (31) and (32) clearly display the splitting of the solutions of the Wheeler–DeWitt equation into even and odd solutions. Since these equations are linear and there is no correlation between the even and odd solutions, every solution is a linear combination of an even and an odd solution. Hence without loss of generality, one can restrict to the analysis of Ψ_{even} and Ψ_{odd}.

Before engaging in this analysis, however, I would like to note that the mode expansion (26) for the two-component wave function Ψ, translates into

$$\psi=\sum_{n=0}^{\infty}\sigma_n\Phi_n, \tag{33}$$

for the one-component wave function. Thus, it is the coefficients of the zero and positive modes that determine the solution of the Wheeler–DeWitt equation (1). The negative modes are related to the derivative of ψ with respect to α. Hence, they are also uniquely determined once the zero and positive modes are given.

In the following sections, I shall show that the mode expansion of the solution of the Wheeler–DeWitt equation (1) is always infinite. This is another way of demonstrating the fact that the adiabatic approximation, which corresponds to including only a single mode in this expansion, always fails. Nevertheless, if one tries to blindly apply the adiabatic approximation or a slightly

improved version of it in which one sets all $\sigma_{\pm n}$ to zero, except $\sigma_{\pm n_*}$ for some n_*, then Eqs. (29) and (30) yield $\ddot{\sigma}_{n*}+E_n^2\sigma_{n*}=0$. This corresponds to the eigenvalue equation for a one-dimensional quantum particle of mass 1/2 and potential $\mathscr{V}(\alpha):=-E_n^2=\kappa e^{4\alpha}-(2n+1)me^{3\alpha}$. Specifically, σ_{n*} is the zero-energy eigenfunction. For $\kappa=1$, $\mathscr{V}$ vanishes at $\alpha=-\infty$. It decreases to attain a negative minimum as α increases, and then monotonically increases to $+\infty$ as $\alpha\to+\infty$. Hence, if one requires the zero-energy eigenfunction to vanish at $\alpha=+\infty$ where $\mathscr{V}=+\infty$, then the mass m must satisfy a peculiar quantization condition. This is analogous to the mass quantization advocated in Ref. 11. However, note that the basic hypothesis of this argument, namely the validity of the adiabatic approximation or alternatively the presence of only a single mode in the mode expansion (33), is never satisfied. Hence, the argument leading to the quantization of mass does not apply.

VI. EVEN SOLUTIONS

Let us first prove:

Lemma 1: An even solution $\Psi_{\rm even}=\Sigma_{\mathscr{N}}Z_{2\mathscr{N}}\Psi_{2\mathscr{N}}$ *is nontrivial, i.e.,* $\Psi_{\rm even}\neq 0$, *if and only if* $\sigma_{\pm 0}\neq 0$.

In order to prove this statement, one sets $n=0$ in Eqs. (29) and (30). This leads to

$$\dot{\sigma}_0+iE_0\sigma_{-0}+\frac{3\sqrt{2}}{4}\,\sigma_2=0, \tag{34}$$

$$E_0\dot{\sigma}_{-0}+\dot{E}_0\sigma_{-0}+iE_0^2\sigma_0+\frac{3\sqrt{2}}{4}\,E_2\sigma_{-2}=0. \tag{35}$$

Therefore, if $\sigma_{\pm 0}=0$, both σ_2 and σ_{-2} must vanish. However, by virtue of Eqs. (32) and (31), this is sufficient to conclude that $\sigma_{2\mathscr{N}}=0$, for all $\mathscr{N}\in\mathbb{Z}$. This completes the proof of Lemma 1. Furthermore, one can prove

Lemma 2: Every nontrivial even solution has an infinite mode expansion.

The even solutions may nevertheless be constructed inductively in terms of $\sigma_{\pm 0}$. This is done by noting that Eqs. (31) and (32) yield trivial algebraic equations for $\sigma_{\pm 2(n+1)}$ in terms of $\sigma_{\pm 2n}$. Performing an induction on n, one can express the coefficients of all the modes as linear combinations of $\sigma_{\pm 0}$ and its derivatives. For example, one has

$$\sigma_2=-\frac{1}{\sqrt{2}}\left(\frac{4}{3}\right)(\dot{\sigma}_0+iE_0\sigma_{-0}), \tag{36}$$

$$\sigma_4=\frac{1}{\sqrt{4!}}\left(\frac{4}{3}\right)^2\left[\ddot{\sigma}_0+2iE_0\dot{\sigma}_{-0}+\left(\frac{3}{2}-E_0^2\right)\sigma_0+2i\dot{E}_0\sigma_{-0}\right]. \tag{37}$$

It is easily seen that as a result of this construction σ_{2n} involves the first n derivatives of σ_0, first $n-1$ derivatives of σ_{-0}, and the first $n-1-j$ derivatives of E_{2j}, for $j=0,1,\ldots,n-1$.

Since according to Eq. (24), $\dot{E}_n$ diverges at the zero of E_n, demanding regularity of σ_n would impose certain conditions on $\sigma_{\pm 0}$. In fact, in view of the appearance of E_n^2 in Eq. (32), which is analytic for every α, one can easily show that the only source of possible singularity is the term proportional to E_0 in the expression for σ_2. Thus, in order to ensure the regularity of σ_{2n}, one must require that at the zero of E_0, i.e., $\alpha=\ln(m/\kappa)$, $\sigma_{-0}E_0$ and all its derivatives must vanish. In other words there must be an open neighborhood $\mathscr{O}$ of $\alpha=\ln(m/\kappa)$, in which $\sigma_{-0}(\alpha)$ is of the form $E_0(\alpha)f(\alpha)$ for some smooth (C^∞) function f. In particular $\sigma_{-0}(\ln(m/\kappa))$ must vanish. Note that for the flat ($\kappa=0$) and open ($\kappa=-1$) cases, this condition is void and σ_{2n} are regular if and only if $\sigma_{\pm 0}$ are regular.

Finally note that termination of the mode expansion for a nontrivial even solution is equivalent to vanishing of at least three consecutive σ_{2n}'s. This would impose three functional conditions on $\sigma_{\pm 0}$ which are in general incompatible. This proves Lemma 2.

VII. ODD SOLUTIONS

The structure of the odd solutions is similar to that of the even solutions. The role of $\sigma_{\pm 0}$ is played by $\sigma_{\pm 1}$. In particular, one has

Lemma 3: An odd solution $\Psi_{\rm odd}=\Sigma_{\mathcal{N}} Z_{2\mathcal{N}+1}\Psi_{2\mathcal{N}+1}$ *is nontrivial, if and only if* $\sigma_{\pm 1}:=Z_1 \pm Z_{-1}\neq 0$.

This is a direct consequence of Eqs. (29) and (32) which also provide an inductive construction of the mode coefficients $\sigma_{\pm(2n+1)}$ in terms of $\sigma_{\pm 1}$. For example one has:

$$\sigma_3=-\frac{1}{\sqrt{3!}}\left(\frac{4}{3}\right)(\dot{\sigma}_1+iE_1\sigma_{-1}), \tag{38}$$

$$\sigma_5=\frac{1}{\sqrt{5!}}\left(\frac{4}{3}\right)^2\left[\ddot{\sigma}_1+2iE_1\dot{\sigma}_{-1}+\left(\frac{9}{2}-E_1^2\right)\sigma_1+2i\dot{E}_1\sigma_{-1}\right]. \tag{39}$$

A natural consequence of this inductive construction is

Lemma 4: Every nontrivial odd solution has an infinite mode expansion.

The regularity condition on the odd solutions is analogous to the even case. One can again show that the only source of possible singularity is the term $E_1\sigma_{-1}$ in the expression for σ_3. Hence the regularity of σ_{2n+1} and therefore ψ is ensured by requiring that in the vicinity of the zero of E_1, namely $\alpha=\ln(3m/\kappa)$, σ_{-1} must be of the form $E_1(\alpha)g(\alpha)$, for some smooth function g.

The occurrence of two free functions of α in the construction of even and odd solutions is in complete agreement with the general structure of the Wheeler–DeWitt/Klein–Gordon equation (1) whose solutions depend on two boundary or two initial conditions. The latter are given by two functions of α which in the above procedure can be used to determine $\sigma_{\pm 0}$ or $\sigma_{\pm 1}$ depending on whether the solution is even or odd. In general the solution will be the sum of an even and an odd part. In this case the boundary (initial) conditions would determine both $\sigma_{\pm 0}$ and $\sigma_{\pm 1}$. Once these functions are determined the above iterative construction yields the mode expansion of the general solution.

VIII. REAL SOLUTIONS

Probably the best-known proposal for the solution of the Wheeler–DeWitt equation is that of Hartle and Hawking[12] who suggested the path integral expression $\psi\propto\int \exp(-I)$ for the ground state wave function. Here I stands for the Euclidean action. The Hartle–Hawking wave function is by construction real.

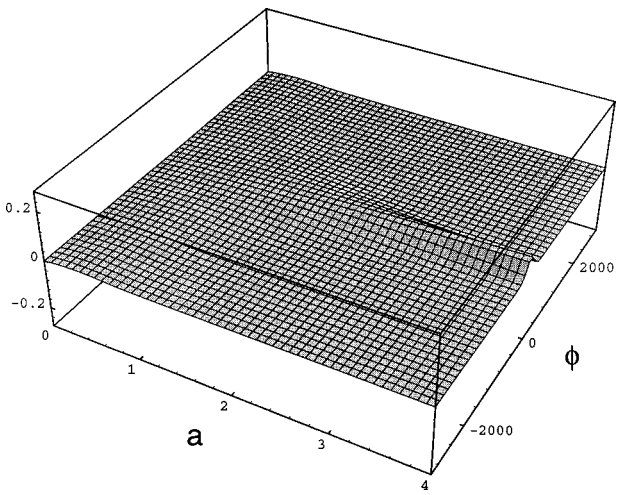

FIG. 1. Plot of $\psi_0=\psi_2=\Sigma_{n=0}^{2}\,\sigma_n\Phi_n$, for $\sigma_{\pm 1}=\sigma_{-0}=0$, $\sigma_0=1$, and $m=10^{-6}$.

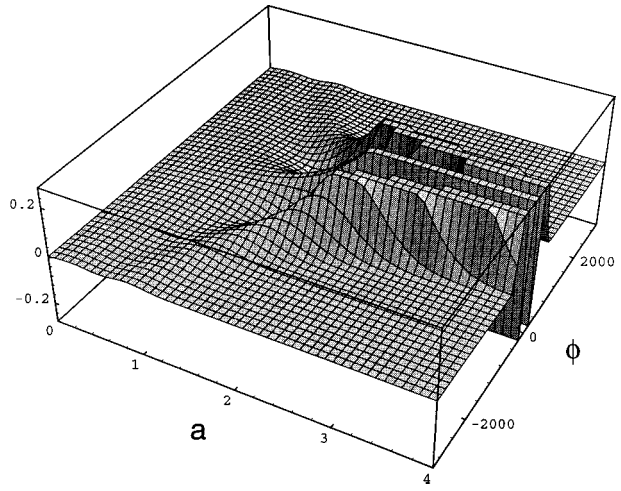

FIG. 2. Plot of $\psi_4:=\Sigma_{n=0}^{4}\ \sigma_n\Phi_n$, for $\sigma_{\pm 1}=\sigma_{-0}=0$, $\sigma_0=1$, and $m=10^{-6}$.

It is not difficult to see that the reality of the one-component solution of the Wheeler–DeWitt equation implies the reality of the coefficient functions σ_n in its mode expansion (33). This is because Φ_n are already real. In view of this observation and Eq. (29), $E_n\sigma_{-n}$ must be imaginary. However, as discussed in Sec. II for the open and flat universes ($\kappa=-1,0$), E_n is always real so that σ_{-n} must be imaginary. For the close universe ($\kappa=1$) E_n is real for $a\leqslant m(2n+1)$ and imaginary otherwise. Hence in this case ($\kappa=1$) the opposite holds for σ_{-n}, i.e., it is imaginary for $a<m(2n+1)$ and real for $a>m(2n+1)$.

These conditions can also be stated as the reality condition on σ_0, σ_1, σ_2, and σ_3 which also generate the solutions. Therefore, they are rather trivial.

IX. APPROXIMATE SOLUTIONS

As shown in the preceding sections, the mode expansion of the solutions never terminates. Therefore, one cannot use it to construct an exact solution unless one can actually find the generic expression for the mode coefficients and sum the series. This is an almost impossible task. However, one can terminate the mode expansion artificially and obtain an infinite class of approximate solutions. $\psi\approx\psi_{n_*}:=\Sigma_{n=0}^{n_*}\ \sigma_n\Phi_n$. The domain of reliability of this solution will in general depend on the boundary (initial) conditions, alternatively the functions $\sigma_{\pm 0}$ and $\sigma_{\pm 1}$.

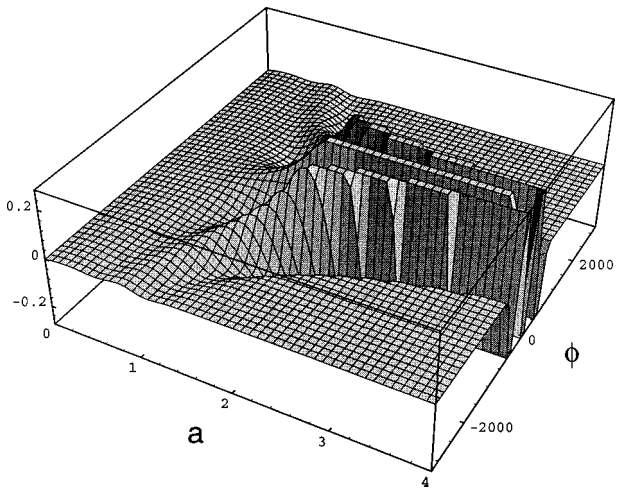

FIG. 3. Plot of $\psi_6:=\Sigma_{n=0}^{6}\ \sigma_n\Phi_n$, for $\sigma_{\pm 1}=\sigma_{-0}=0$, $\sigma_0=1$, and $m=10^{-6}$.

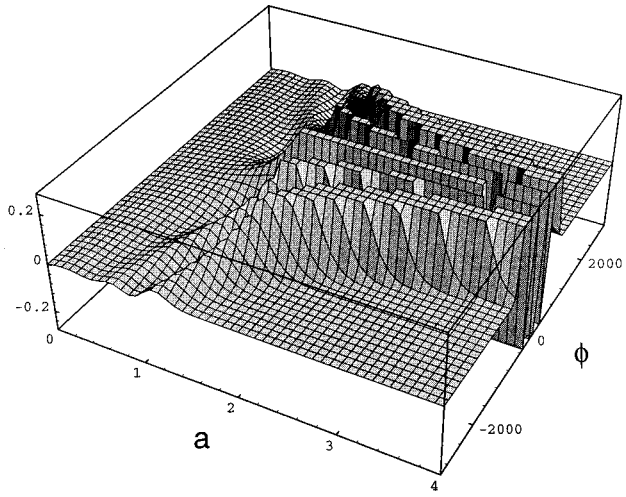

FIG. 4. Plot of $\psi_8 := \Sigma_{n=0}^{8}\ \sigma_n \Phi_n$, for $\sigma_{\pm 1} = \sigma_{-0} = 0$, $\sigma_0 = 1$, and $m = 10^{-6}$.

In this section, I shall examine the simple choice $\sigma_{-0} = \sigma_{\pm 1} = 0$ and $\sigma_0 = 1$. Clearly, this corresponds to a real even solution whose mode coefficients can be obtained using Eq. (32). For example for the closed FRW case where $\kappa = 1$, one has

$$\sigma_2 = 0, \quad \sigma_4 = (\tfrac{2}{3})^{5/2}(\tfrac{3}{2} + a^4 - ma^3), \quad \sigma_6 = \left(\frac{32a^3}{81\sqrt{5}}\right)(3m - 4a),$$

$$\sigma_8 = \frac{4}{243}\sqrt{\frac{2}{35}}\,(4a^8 - 40ma^7 + 36m^2a^6 + 324a^4 - 288ma^3 + 189),$$

$$\sigma_{10} = -\frac{16}{10935\sqrt{7}}\,(96a^5 - 1000ma^4 + 744m^2a^3 + 3604a - 2655m)a^3.$$

Substituting these relations in the expression for ψ_n and making use of the identity[13]

$$H_n(x) = 2^n x^n + \sum_{i=1}^{[n/2]} (-1)^i 2^{n-i}(2i-1)!!\binom{n}{2i} x^{n-2i},$$

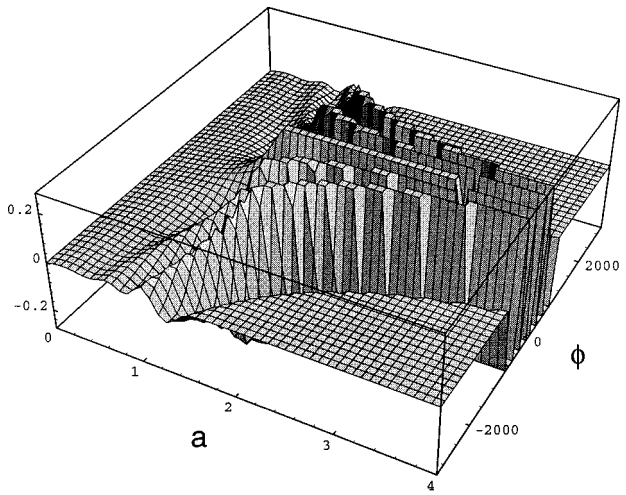

FIG. 5. Plot of $\psi_{10} := \Sigma_{n=0}^{10}\ \sigma_n \Phi_n$, for $\sigma_{\pm 1} = \sigma_{-0} = 0$, $\sigma_0 = 1$, and $m = 10^{-6}$.

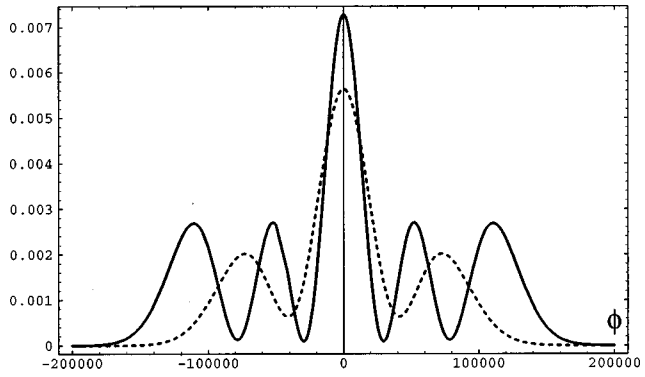

FIG. 6. Plot of the restriction of ψ_6 (dotted line), ψ_8 (dashed line), and ψ_{10} (full line) to the $a=0.1$ line. Note that the graphs of ψ_8 and ψ_{10} coincide for this value of a. This shows the reliability of the finite mode approximation. Note also the range of the values of ϕ over which ψ_n are nonzero.

to calculate the mode functions Φ_n, one can compute $\psi_0=\psi_2$, ψ_4, ψ_6, ψ_8 and ψ_{10}. Figures 1–5 illustrate three-dimensional plots of these functions. Figures 6–9 are plots of the restriction of ψ_6, ψ_8, and ψ_{10} to the lines $a=0.1,1,10,100$. These figures show that for small a the finite mode approximation is quite reliable. In fact, as depicted in Fig. 6, for $a=0.1$, the graphs of ψ_8 and ψ_{10} coincide. For larger a one may improve the approximation by including the contribution of higher modes.

Note also that initially ψ_n are nonzero for a wide range of values of ϕ. However, as a increases the width of this region, which is symmetric with respect to the line $\phi=0$, shrinks. This is easily seen from Figs. 6–9. Note also that for large a, the solutions picks about $\phi=0$ and the approximate solutions become less reliable. As $a\to\infty$, all the approximate solutions seem to vanish except at $\phi=0$.

X. DISCUSSION AND CONCLUSION

Probably the most important aspect of the two-component formulation of the Wheeler–DeWitt equation is the fact that the construction of the solutions in this formulation does not, *a priori*, require fixing a particular inner product or boundary conditions. The freedom in the choice of the inner product is reflected as the independence of the mode expansion from the choice of the Hermitian matrix h which together with the L^2 inner product on $L^2(\mathbb{R})$ determines the inner product on $\mathscr{H}=L^2(\mathbb{R})\oplus L^2(\mathbb{R})$. As far as the boundary conditions are concerned, besides the implicit restriction of vanishing boundary condition at ''spatial'' infinity, the analysis presented here applies generally.

An interesting consequence of the application of the two-component formalism is a specific double grading of the solutions of the Wheeler–DeWitt equation. This allows one to write down

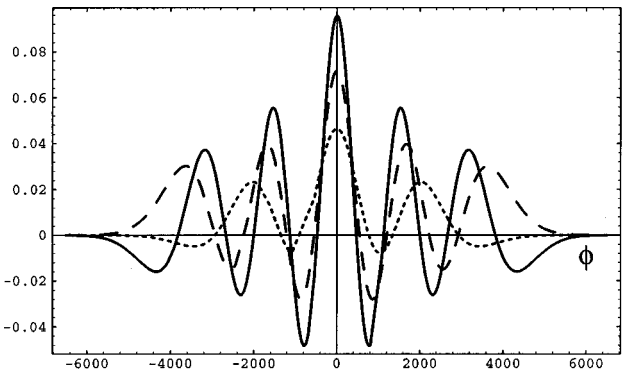

FIG. 7. Plot of the restriction of ψ_6 (dotted line), ψ_8 (dashed line), and ψ_{10} (full line) to the $a=1$ line. Note the range of the values of ϕ over which ψ_n are nonzero.

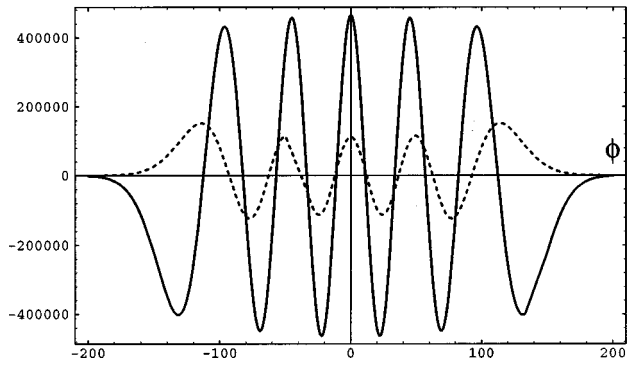

FIG. 8. Plot of the restriction of ψ_8 (dotted line), and ψ_{10} (full line) to the $a=10$ line. Note the range of the values of ϕ over which ψ_n are nonzero.

the most general solution of this equation as a sum of definite-parity (even and odd) solutions whose structure can be easily understood. The two definite-parity parts of the solution do not mix during the evolution. In particular, if the initial wave function has definite parity, its parity remains intact throughout the evolution.

Unfortunately, the mode expansion of the solutions never terminates. Thus, a closed expression for an exact solution cannot be found in this way. However, approximate solutions can be constructed by artificially terminating the mode expansion. The domain of applicability of such approximations depend generally on the boundary (initial) conditions.

An alternative approximation is the adiabatic approximation which is, however, not valid. The ansatz of the adiabatic approximation, i.e., $\psi\approx\sigma_n(\alpha)\Phi_n(\alpha,\phi)$, is reminiscent of the application of the Born–Oppenheimer approximation in quantum cosmology originally advocated by Banks.[14] In this approach one views the gravitational degrees of freedom as the ''slow'' variables and the matter degrees of freedom as the ''fast'' variables and considers solutions of the form $\psi=\psi_{\text{gravity}}(\alpha)\psi_{\text{matter}}(\alpha,\phi)$. The results of this article indicate that for $\psi_{\text{matter}}=\Phi_n$ such an approximation cannot be valid.

Another notable implication of the results of this article is regarding the implicit boundary condition at the initial singularity $a=0$. If one assumes that the coefficient functions σ_n are regular at $a=0$, then as a consequence of the appearance of $a^{3/4}$ in the expression for Φ_n, the one-component wave-function ψ vanishes at $a=0$. This is much in the spirit of an argument of DeWitt.[15] Here it follows as a direct consequence of a regularity condition (which may or may not hold). As seen from Figs. 1 to 5, the choice $\sigma_{-0}=\sigma_{\pm1}=0$ and $\sigma_0=1$ leads to approximate solutions which are regular at $a=0$. They clearly satisfy DeWitt's boundary conditions.

A problem which is not addressed in this article is that of the convergence of the mode expansion in $\mathscr{H}$. This is a difficult problem and I have not been able to find the necessary and sufficient conditions under which the mode expansion converges. In fact, a rigorous convergence

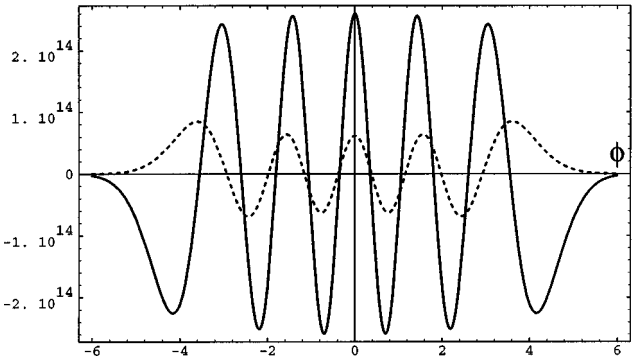

FIG. 9. Plot of the restriction of ψ_8 (dotted line), and ψ_{10} (full line) to the $a=100$ line. Note the range of the values of ϕ over which ψ_n are nonzero.

analysis requires a choice of a positive-definite inner product (norm) on $\mathscr{H}$. However, as demonstrated in Sec. IX, there are regions of the minisuperspace in which the contribution from the higher order modes is small. This may be viewed as an indication of possible pointwise asymptotic convergence of the mode expansion in this regions. This argument is however far from being satisfactory and further investigation of this point is necessary.

Finally, I wish to emphasize the following points.

(1) The fact that the two-component Hamiltonian (6) belongs to the dynamical Lie algebra $\mathrm{su}_s(1,1)\otimes\mathrm{su}_u(1,1)$ suggests a further investigation of the solution of the corresponding Schrödinger equation using the group theoretical methods such as those employed in solving the Schrödinger equation for time-dependent harmonic oscillators.[16] Recall that the Hamiltonian for a time-dependent harmonic oscillator belongs to $\mathrm{su}_u(1,1)$.[6]

(2) The two-component approach to Wheeler–DeWitt equation has a wider domain of applicability than the simple minisuperspace model studied in this article. For example, the addition of an interaction potential for the scalar field would only complicate the defining equation for Φ_n and the inductive construction of the mode expansion. The general structure of the method would however remain intact. In particular, the ϕ-dependence of the solutions would still be given by Φ_n. More generally, for any other minisuperspace, in the gauge in which the shift vector corresponding to DeWitt's supermetric[15] vanishes, the two-component Hamiltonian is of the form (6), where D is some second-order elliptic operator. Therefore, even in this case the developments reported in this article are essentially applicable.

Note added in proof: After the completion of this work, Refs. 17, 18 were brought to my attention where the authors use two different two-component formulations of the Wheeler–DeWitt equation. The methods described in these articles differ substantially from the present article's. I have also noticed a recent paper by Embacher[19] who defines the operator D of Eq. (7) as the (matter) energy operator and obtains the mode expansion (33) in the context of the one-component Wheeler–DeWitt equation. He does not, however, obtain the equations satisfied by the mode coefficients. Instead, he focuses on the semiclassical and adiabatic approximations. The two-component approach developed in this paper justifies the choice of the mode functions made by Embacher in a very natural way. It also makes it possible to obtain the structure of the mode expansion by iteratively solving for the mode coefficients.

ACKNOWLEDGMENTS

I would like to thank B. Darian, M. Razavi, and especially D. Page for invaluable comments and suggestions. I would also like to acknowledge the financial support of the Killam Foundation of Canada.

[1] D. N. Page, in *Gravitation: A Banff Summer Institute*, edited by R. Mann and P. Wesson (World Scientific, Singapore, 1991); D. L. Wiltshire, in *Cosmology: The Physics of the Universe* (World Scientific, Singapore, to be published).
[2] D. J. Kaup and A. P. Vitello, Phys. Rev. D **9**, 1648 (1974).
[3] S. W. Hawking and D. N. Page, Phys. Rev. D **42**, 2655 (1990); see also A. Mostafazadeh, Nucl. Phys. B **509**, 529 (1998), and references therein.
[4] A. Mostafazadeh (unpublished).
[5] H. Feshbach and F. Villars, Rev. Mod. Phys. **30**, 24 (1958); B. R. Holstein, *Topics in Advanced Quantum Mechanics* (Addison–Wesley, Redwood City, CA, 1992).
[6] R. Jackiw, Int. J. Mod. Phys. A **3**, 285 (1988).
[7] L. I. Schiff, *Quantum Mechanics* (McGraw–Hill, New York, 1955).
[8] A. Mostafazadeh, Phys. Rev. A **55**, 1653 (1997).
[9] A. Mostafazadeh, Phys. Lett. A **228**, 7 (1997).
[10] M. Read and B. Simon, *Functional Analysis* (Academic, San Diego, 1980), Vol. I.
[11] C. Kiefer, Phys. Rev. D **38**, 1761 (1988).
[12] J. B. Hartle and S. W. Hawking, Phys. Rev. D **28**, 2960 (1983).
[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980).
[14] T. Banks, Nucl. Phys. B **249**, 332 (1985); R. Brout, Found. Phys. **17**, 603 (1987); R. Brout and G. Venturi, Phys. Rev. D **39**, 2436 (1989); G. Venturi, in *Differential Geometric Methods in Theoretical Physics*, edited by L.-L. Chau and W. Nahm (Plenum, New York, 1990); D. P. Datta, Phys. Rev. D **48**, 5746 (1993).
[15] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967).
[16] H. R. Lewis, Jr. and W. B. Riesenfeld, J. Math. Phys. **10**, 1458 (1969); D. M. Greenberger, *ibid.* **15**, 395 (1974); R. W. Hasse, *ibid.* **16**, 2005 (1975); V. V. Dodonov and V. I. Man'ko, Phys. Rev. A **20**, 550 (1979); L. F. Landovitz, A. M. Levine, and W. M. Schreiber, *ibid.* **20**, 1162 (1979); M. S. Abdalla and R. K. Colegrave, *ibid.* **32**, 1958 (1985); A. B. Nassar and R. T. Berg, *ibid.* **34**, 2462 (1986); C. M. Cheng and P. C. W. Fung, J. Phys. A **21**, 4115 (1988); F. Wolfs and H. J. Korsch, Phys. Rev. A **37**, 1934 (1988); G. S. Agarwal and S. Arun Kumar, Phys. Rev. Lett. **67**, 3665 (1991); A. de

Souza Dutra, C. F. de Souza, and L. C. de Albuquerque, Phys. Lett. A **156**, 371 (1991); S. P. Kim, J. Phys. A **27**, 3927 (1994); J.-Y. Ji and J. K. Kim, Phys. Rev. A **51**, 4268 (1995); A. Mostafazadeh, *ibid.* **55**, 4084 (1997).
[17] S. P. Kim, Phys. Rev. D **46**, 3403 (1992).
[18] T. Dereli, M. Önder, and R. W. Tucker, Phys. Lett. B **324**, 134 (1994).
[19] Franz Embacher, Nucl. Phys. B **479**, 461 (1996).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Quantum Anosov flows: A new family of examples

Ingo J. Peter
Department of Mathematics, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark

Gérard G. Emch
Department of Mathematics, University of Florida, Gainesville, Florida 32601-8105

(Received 13 March 1998; accepted for publication 23 April 1998)

A quantum version is presented for the Anosov system defined by the time evolution implemented by the geodesic coflow on the cotangent bundle of any compact quotient manifold obtained from the Poincaré half-plane. While the canonical Weyl algebra does not close under time evolution, the symplectic structure of these classical systems can be exploited to produce objects akin to the CCR algebras encountered in quantum field theory. This construction allows one to lift both the geodesic and the horocyclic flows to a Weyl algebra describing the quantum dynamics corresponding to the systems under consideration. The Anosov relations as proposed in Ref. 1 are found to be valid for these models. A quantum version of the classical ergodicity of these systems is discussed in the last section. © *1998 American Institute of Physics.* [S0022-2488(98)02109-4]

I. INTRODUCTION: THE CLASSICAL ARCHETYPE

The classical theory of Anosov systems describes a class of unstable dynamical systems. Recently a general axiomatic framework for a possible noncommutative extension of the Anosov theory to the quantum realm has been proposed and developed in this Journal;[1] see also Ref. 2. However, up to now, the only existing examples of quantum Anosov systems are the quantum version of the ''Arnold Cat'' and the modular action of the Lorentz group on a Rindler wedge. Moreover, it is argued in Ref. 1 that the usual, or canonical, quantization of the archetype of classical Anosov systems—namely free motion of a particle moving on a compact Riemannian manifold of constant negative curvature—cannot carry over the Anosov structure: The spectrum of the Laplace–Beltrami operator is discrete, so that the resulting quantum system is not even ergodic. The purpose of the present paper is to propose an alternative quantization scheme that carries over the full Anosov structure.

The classical situation can be summarized as follows. The exponential sensitivity to initial conditions of the geodesic flow in Lobatchevski hyperbolic geometry was noticed already by the end of the last century,[3] and it can be regarded as the prototype for the study of Anosov structures.[4]

Specifically, three flows are defined, first on the cotangent bundle T^*M of the Poincaré half-plane, i.e., the two-dimensional, simply connected Riemannian manifold (M,g) of constant negative curvature $\kappa=-\rho^2$. These flows are the geodesic flows $\mathrm{Fl}_t^{X_H}$ and the two horocyclic flows $\mathrm{Fl}_s^{E_\pm}$. They satisfy the Anosov relations

$$\mathrm{Fl}_{-t}^{X_H}\circ\mathrm{Fl}_s^{E_\pm}\circ\mathrm{Fl}_t^{X_H}=\mathrm{Fl}_{\mathrm{e}^{\pm\rho t}s}^{E_\pm}, \tag{1.1}$$

where the corresponding vector fields X_H and $E_\pm$—for details, see Lemma A4 and Remark A5—will be subsequently normalized by $r^2=g(G_\pm,G_\pm)=g(H_\pm,H_\pm)$; for transparency, we simply chose in this section the usual normalization $r=1$. These complete vector fields then satisfy the Anosov commutation relations (ACRs)

$$[X_\pm,E_\pm]=\pm\rho\,E_\pm\,. \tag{1.2}$$

Suffice it to say here that: (i) the vector field X_H is the Hamiltonian vector field for the Hamiltonian

 © 1998 American Institute of Physics

$$H(x^i,p_i):=\tfrac{1}{2}g^{\sharp}((x^i,p_i),(x^i,p_i))=\tfrac{1}{2}(p_1^2+p_2^2); \tag{1.3}$$

(ii) the vector fields $E_{\pm}$ are the expanding (respectively, contracting) *horocyclic vector fields*; (iii) the constants $\pm\rho$ appearing in these ACRs are to be interpreted as the *Lyapunov coefficients* for this dynamical system; and (iv) while the geodesic flow is a symplectic action, this is *emphatically not* the case for the horocyclic flows (see Proposition A9).

In order to obtain a probability measure invariant under the dynamics, one finally restricts the above flows to the unit cotangent bundle of certain compact quotients Γ/M of the Poincaré half-plane M by co-compact, discrete subgroups Γ of the symmetry group $\mathrm{SL}(2,R)$ of M. It is the resulting dynamical system that is the classical archetype for the quantum structures we want to construct: it is indeed a C-flow, and thus satisfies in particular the following properties (listed in descending strength): it is a K-flow, it is mixing with uniform Lebesgue spectrum, and it is ergodic.[4]

Appendix A presents a fuller, self-contained treatment of this classical model, together with further bibliographical references. The presentation is made in a notation that covers a two-parameter family of models. This is done with two purposes in mind. The first is to facilitate quantization; the second is to allow, at some ulterior stage, a smooth passage to the flat limit where the ergodic behavior disappears. This notation is therefore systematically used in the main body of the paper.

In *Appendix B*, we discuss a no-go theorem showing why the canonical quantum realization of the Weyl algebra does not allow one to implement the time evolution as a geometric action by automorphisms.

In *Sec. II*, we show how to enlarge the quantum algebra to obtain a dynamics implemented by a group of automorphisms; this new Weyl C*-algebra is described in detail.

In *Sec. III*, we construct quantum versions of the horocyclic actions on this algebra implementing the Anosov relations.

Section IV describes how to restrict the quantum hyperbolic systems, when the Poincaré half-plane is replaced by any compact, complete, connected two-dimensional manifolds of constant negative curvature, so that the full quantum Anosov structure can be displayed.

Section V specifically addresses the ergodicity of the quantum systems so obtained, in terms of the existence of states that are invariant under geodesic and horocyclic actions.

II. QUANTUM DYNAMICS

When naturally extended[5,6] to curved configuration spaces, the usual construction of the quantum Weyl group is known to allow a proper description of the kinematics. However, it may fail to sustain a self-contained description of the dynamics. We show in Appendix B that, indeed, this is already the case for the geodesic flow on the Poincaré half-plane.

Consequently, we propose in this section an alternative Weyl algebra, built directly on a class of smooth functions on the *phase space*. The idea in the background is that, classically, the algebra of (exponentiated) observables is given by essentially bounded functions $L^{\infty}(T^*M,\mathbb{C})$ on the phase space. Our definition has the advantage that time evolution can be defined. The first step in this direction is to find a reasonable symplectic structure on the space of smooth functions with compact support $C_c^{\infty}(T^*M,\mathbb{R})$ on the phase space T^*M of our classical models.

Definition 2.1: With the notations introduced in Appendix A, let $C_c^{\infty}(T^*M_{\mu,\lambda},\mathbb{R})$ denote the space of smooth functions of compact support on the cotangent bundle $T^*M_{\mu,\lambda}$. Let H denote the Hamiltonian, and $\{\,.\,,\,.\,\}$ the Poisson bracket. We define a degenerate skew-symmetric bilinear form on $C_c^{\infty}(T^*M_{\mu,\lambda},\mathbb{R})$ by

$$\sigma(f,g):=\int_{T^*M_{\mu,\lambda}} H\{f,g\}\quad \omega\wedge\omega \qquad \forall f,g\in C_c^{\infty}(T^*M_{\mu,\lambda},\mathbb{R}). \tag{2.1}$$

We denote by (V,σ) the degenerate symplectic vector space $(C_c^{\infty}(T^*M_{\mu,\lambda},\mathbb{R}),\sigma)$.

Remark 2.2: In Proposition 2.5 below, we will compute the ''kernel'' of the symplectic form σ and obtain a nondegenerate symplectic vector space. The construction of Weyl algebras with degenerate symplectic vector spaces has been examined in Ref. 7.

Lemma 2.3: The skew-symmetric form σ may also be written equivalently as

$$(a)\quad \sigma(f,g)=\int_{T^*M_{\mu,\lambda}}(fdg-gdf)\wedge dH\wedge\omega,$$

$$(b)\quad \sigma(f,g)=\frac{1}{2}\int_{T^*M_{\mu,\lambda}}(f\{g,H\}-g\{f,H\})\ \omega\wedge\omega,$$

where $f,g\in V$ and H is the Hamiltonian.

Proof: We compute

$$\begin{aligned}(fdg-gdf)\wedge dH\wedge\omega&=-d(H(fdg-gdf)\wedge\omega)+2Hdf\wedge dg\wedge\omega\\&=-d(H(fdg-gdf)\wedge\omega)+H\ i_{X_g}df\ \omega\wedge\omega\\&=-d(H(fdg-gdf)\wedge\omega)+H\{f,g\}\ \omega\wedge\omega.\end{aligned}\tag{2.2}$$

Finally using Stoke's theorem yields (a). In order to show the equality (b), we compute

$$\begin{aligned}(fdg-gdf)\wedge dH\wedge\omega&=(fdg-gdf)\wedge i_{X_H}\omega\wedge\omega\\&=\tfrac{1}{2}(fdg-gdf)\wedge i_{X_H}(\omega\wedge\omega)\\&=\tfrac{1}{2}(i_{X_H}(fdg-gdf))\ \omega\wedge\omega\\&=\tfrac{1}{2}(f\ i_{X_H}dg-g\ i_{X_H}df)\ \omega\wedge\omega\\&=\tfrac{1}{2}(f\{g,H\}-g\{f,H\})\ \omega\wedge\omega.\end{aligned}\tag{2.3}$$

□

Remark 2.4: The naive candidate $\int\{f,g\}\ \omega\wedge\omega$ does not work as a symplectic form, since the integrand is exact. Indeed, for any $f,g\in C_c^\infty(T^*M_{\mu,\lambda},\mathsf{R})$, we have:

$$\begin{aligned}\{f,g\}\ \omega\wedge\omega&=(i_{X_g}df)\ \omega\wedge\omega\\&=df\wedge i_{X_g}(\omega\wedge\omega)\\&=2df\wedge dg\wedge\omega=d((fdg-gdf)\wedge\omega).\end{aligned}\tag{2.4}$$

Proposition 2.5: Let V_0 be the ''kernel'' V_0 of σ defined as

$$V_0:=\cap_{f\in V}\{g\in V|\sigma(f,g)=0\}.$$

Then

$$V_0=\{g\in V|\{g,H\}=0\},$$

i.e., V_0 coincides with the constants of motion. $V_0\subset V$ is an algebra. The symplectic form σ factors to a nondegenerate symplectic form $\tilde{\sigma}:\tilde{V}\times\tilde{V}\to\mathsf{R}$ on the vector space $\tilde{V}:=V/V_0$: $\tilde{\sigma}([f],[g]):=\sigma(f,g)$ for $f,g\in V$, where $[.]:V\to\tilde{V}$ denotes the quotient mapping.

Proof: From Lemma 2.3, we know that $g\in V_0$ iff we have for all $f\in V$:

$$\int_{T^*M_{\mu,\lambda}}f\{g,H\}\ \omega\wedge\omega=\int_{T^*M_{\mu,\lambda}}g\{f,H\}\ \omega\wedge\omega\tag{2.5}$$

and therefore

$$2\int_{T^*M_{\mu,\lambda}} f\{g,H\}\ \omega\wedge\omega=\int_{T^*M_{\mu,\lambda}} (f\{g,H\}+g\{f,H\})\ \omega\wedge\omega$$
$$=\int_{T^*M_{\mu,\lambda}} \{fg,H\}\ \omega\wedge\omega. \tag{2.6}$$

By Remark 2.4 the second line in the above expression vanishes. Hence

$$\int_{T^*M_{\mu,\lambda}} f\{g,H\}\ \omega\wedge\omega=0 \quad \forall f\in V \tag{2.7}$$

and thus $\{g,H\}=0$. □

Remark 2.6: Let us choose a coordinate system on the cotangent bundle $T^*M_{\mu,\lambda}$ where $r=|p|$ is singled out; the Hamiltonian then takes the form $H=H(r)=\frac{1}{2}r^2$, and the corresponding Hamiltonian vector field reads:

$$X_H(x^1,x^2,r,\psi)=r\ \cos\ \psi\cdot s^1+r\ \sin\ \psi\cdot s^2+\frac{2\lambda r}{\mu}\ \sin\ \psi\ \partial_\psi. \tag{2.8}$$

Whence we get the following explicit condition for g to belong to the kernel V_0:

$$0=\{g,H\}=X_H(g)$$
$$=\frac{1}{\mu}r\ \cos\ \psi\ (\partial_{x^1}g)+re^{-2\lambda x^1}\ \sin\ \psi\ (\partial_{x_2}g)+\frac{2\lambda r}{\mu}\ \sin\ \psi(\partial_\psi g). \tag{2.9}$$

Any function $g=g(r)$ depending only on the variable r, which is essentially energy, satisfies this relation. The absence of other constants of the motion is the statement that the classical system is ergodic as long as long as the curvature does not vanish. As the flat situation is reached in the limit where $(\mu,\lambda)=(1,0)$, (2.9) reduces to

$$0=\cos\ \psi\ (\partial_{x^1}g)+\sin\ \psi\ (\partial_{x^2}g) \tag{2.10}$$

from which we read that new constants of the motion do appear in this limit: functions depending not only on r, but also on ψ (i.e., not only the energy, but the angular momentum as well). Thus, the vector space $\tilde{V}:=V/V_0$ can be considered as a space of functions ''living'' on the configuration space, *only* in the flat limit: This fact prompted our formulation—one of the main objects of this paper—of the solution to the quantization problem for the curved case.

Lemma 2.7: Let $\Phi:T^*M_{\mu,\lambda}\to T^*M_{\mu,\lambda}$ be a diffeomorphism, that is ω-symplectic and leaves the Hamiltonian H invariant. Then σ is invariant under

$$\Phi^*: f\in V\mapsto f\circ\Phi\in V.$$

Proof:

$$\sigma(\Phi^*f,\Phi^*g)=\int_{T^*M_{\mu,\lambda}} H\{\Phi^*f,\Phi^*g\}\ \omega\wedge\omega$$
$$=\int_{T^*M_{\mu,\lambda}} H\Phi^*\{f,g\}\ \omega\wedge\omega$$
$$=\int_{T^*M_{\mu,\lambda}} \Phi^*(H\{f,g\}\ \omega\wedge\omega)$$
$$=\int_{T^*M_{\mu,\lambda}} H\{f,g\}\ \omega\wedge\omega=\sigma(f,g), \tag{2.11}$$

where we used that symplectomorphisms may be passed through Poisson bracket, H and ω are

invariant under Φ by hypothesis. □

Definition 2.8: For $[f]\in\widetilde{V}$ consider the *-algebra $\Delta(\widetilde{V},\widetilde{\sigma})$ C-generated by elements $W([f])$ with multiplication law

$$W([f])W([g]):=e^{-i\hbar\sigma([f],[g])}W([f+g]) \quad \forall [f],[g]\in\widetilde{V}$$

and an involution defined by $W([f])^*:=W([-f])$. Note that $\Delta(\widetilde{V},\widetilde{\sigma})$ is linearly generated by the $W([f])$. The elements $W([f])$ become unitaries, the unit of the algebra is $W([0])=:\mathsf{I}$. An arbitrary element $R\in\Delta(\widetilde{V},\widetilde{\sigma})$ may thus be written as a finite sum $R=\Sigma_{i=1}^{N}\beta_i W([f_i])$ for $f_i\in\widetilde{V}$ and $\beta_i\in\mathsf{C}$. One knows[7] that there exists a unique C*-norm on this *-algebra, such that its completion with respect to this norm is a C*-algebra: we denote this C*-algebra by $\mathcal{W}:=\overline{\Delta(\widetilde{V},\widetilde{\sigma})}$, and call it the Weyl C*-algebra of our system.[8,9]

Remarks 2.9: (1) At this level, $\hbar$ is merely a parameter which, if put to zero, makes the Weyl algebra Abelian. In fact, if $\hbar=0$ the map

$$\mathcal{W}\ni W([f])\mapsto \mathrm{e}^{if}\in L^{\infty}(T^*M_{\mu,\lambda},\mathsf{C})$$

provides an isomorphism from the Weyl algebra $\mathcal{W}$ to the classical Abelian algebra of observables $L^{\infty}(T^*M_{\mu,\lambda},\mathsf{C})$. (2) The *-algebra $\Delta(\widetilde{V},\widetilde{\sigma})$ has the following locality property: if $f,g\in V$ such that $\mathrm{supp}(f)\cap\mathrm{supp}(g)=\varnothing$, then $W([f])$ commutes with $W([g])$.

Proposition 2.10: Let Fl^{X_H} be the classical geodesic flow. Then

$$\tau_t W([f]):=W([(\mathrm{Fl}_t^{X_H})^*f])\equiv W([f_t])$$

defines a one-parameter group $\tau:\mathsf{R}\to\mathrm{Aut}(\mathcal{W})$ of *-automorphisms of the Weyl algebra $\mathcal{W}$.

In the sequel, the τ_t will be referred to as *the geodesic automorphisms*.

Proof: Let $\Phi:T^*M_{\mu,\lambda}\to T^*M_{\mu,\lambda}$ be a symplectomorphism on the cotangent bundle, that leaves the Hamiltonian invariant, let $\tau_\Phi W([f]):=W([\Phi^*f])$; this is well-defined. Indeed, if $\{H,f\}=0$, then

$$\{H,\Phi^*f\}=\{H,f\circ\Phi\}=\{H\circ\Phi,f\circ\Phi\}=\{H,f\}\circ\Phi=0. \tag{2.12}$$

Further

$$\begin{aligned}\tau_\Phi W(f)\tau_\Phi W(g)&=W(\Phi^*f)W(\Phi^*g)\\&=W(\Phi^*(f+g))\exp(-i\hbar\sigma(\Phi^*f,\Phi^*g))\\&=W(\Phi^*(f+g))\exp(-i\hbar\sigma(f,g)\\&=\tau_\Phi W(f+g)\exp(-i\hbar\sigma(f,g))\\&=\tau_\Phi(W(f)W(g))\end{aligned} \tag{2.13}$$

shows that τ_Φ defines indeed an automorphism of the Weyl algebra. The one-parameter group property is trivial. Finally observe that the lifted geodesic flow $\mathrm{geo}(.;t):T^*M_{\mu,\lambda}\to T^*M_{\mu,\lambda}$ satisfies all demanded requirements: It is symplectic due to Proposition A.9.1 and leaves the Hamiltonian invariant, as it is contained in the r-sphere bundle $T_r^*M_{\mu,\lambda}$. □

III. EXPANDING AND CONTRACTING AUTOMORPHISMS

We know from Proposition A.9 that, while the geodesic flow is implemented by symplectomorphisms of the cotangent bundle, this is *not* the case for the horocyclic flows, thus making more difficult the quantization of the latter. To achieve this is the task of the present section. The following consideration is motivated by Ref. 10.

Lemma 3.1: Let $[f]\in\widetilde{V}$, $Q:\widetilde{V}\to\mathsf{R}$ be a linear functional (without any specified regularity conditions) on the vector space $\widetilde{V}$ and $S_s:\widetilde{V}\to\widetilde{V}$ be a one-parameter group of linear operators, $s\in\mathsf{R}$. Let

$$\beta_s W([f]) := e^{is\langle Q,[f]\rangle} W(S_s[f]).$$

(1) Then $(\beta_s)_{s\in\mathsf{R}}$ defines a one-parameter group of *-automorphisms of the Weyl algebra $\mathcal{W}$ if and only if $(S_s)_{s\in\mathsf{R}}$ is a one-parameter group of symplectic transformations of the vector space $\tilde{V}$ which satisfies $\langle Q,S_s[f]\rangle = \langle Q,[f]\rangle$.

(2) Moreover, one gains the Anosov commutation relations

$$\tau_{-t}\circ\beta_s\circ\tau_t = \beta_{e^{\rho t}s}$$

if and only if $\langle Q,[f_t]\rangle = e^{\rho t}\langle Q,[f]\rangle$ and $(\mathrm{Fl}_{-t}^{X_H})^* S_s (\mathrm{Fl}_t^{X_H})^* = S_{e^{\rho t}s}$.

Proof: Ad (1): Upon imposing the conditions $\langle Q,[f]\rangle = \langle Q,S_s[f]\rangle$ for all $[f]\in\tilde{V}$ and $s\in\mathsf{R}$ we compute

$$\begin{aligned}
\beta_{s_2}\beta_{s_1}(W([f]) &= \beta_{s_2}(\exp(is_1\langle Q,[f]\rangle)W(S_{s_1}[f])) \\
&= \exp(is_2\langle Q,S_{s_1}[f]\rangle)\exp(is_1\langle Q,[f]\rangle)W(S_{s_1+s_2}[f]) \\
&= \exp(is_1\langle Q,[f]\rangle + is_2\langle Q,S_{s_1}[f]\rangle)W(S_{s_1+s_2}[f]) \\
&= \exp(i(s_1+s_2)\langle Q,[f]\rangle)W(S_{s_1+s_2}[f]) \\
&= \beta_{s_1+s_2}W([f]), \qquad (3.1)
\end{aligned}$$

which shows sufficiency and also necessity of $\langle Q,S_s[f]\rangle = \langle Q,[f]\rangle$. The requirement for $S_s:\tilde{V}\to\tilde{V}$ to be symplectic is necessary and sufficient for $\beta_s:\mathcal{W}\to\mathcal{W}$ to be an automorphism; see, in particular (2.13).

Ad (2): We assume the conditions in part (2) of the statement are valid and get

$$\begin{aligned}
\tau_{-t}\circ\beta_s\circ\tau_t(W([f]) &= \tau_{-t}\circ\beta_s(W([f_t])) \\
&= \tau_{-t}(\exp(is\langle Q,[f_t]\rangle)W(S_s[f_t])) \\
&= \exp(ise^{\rho t}\langle Q,[f]\rangle)W((\mathrm{Fl}_{-t}^{X_H})^* S_s(\mathrm{Fl}_t^{X_H})^*[f]) \\
&= \exp(ise^{\rho t}\langle Q,[f]\rangle)W(S_{e^{\rho t}s}[f]) \\
&= \beta_{e^{\rho t}s}W([f]). \qquad (3.2)
\end{aligned}$$

To see the necessity of the condition follows also from these lines with a spectral argument: It is not possible, if $[f]\neq[g]$, that $W([f]) = e^{i\varphi}W([g])$ for some $\varphi\in[0;2\pi]$ because all $W([f])$ with $[f]\neq 0$ have spectrum the unit circle S^1 whereas the spectrum of $e^{i\varphi}$ clearly consists of a single point. □

Remarks 3.2: (1) Since by Proposition A.9 the horocyclic flows $h_\pm$ are not symplectic we must stress that the above lemma makes it definitely impossible to define q-horocyclic flows as $\beta_s^\pm(W([f])) = W([h_\pm(s)^* f])$ and therefore, after having clarified the nature of Q, we will set $S_s = \mathsf{I}$ in Theorem 3.7 below.

(2) A second idea would be to make use of the vector fields $E_\pm$ defined in Corollary A.8 and the relation $E_\pm(f_t)\circ\mathrm{Fl}_{-t}^{X_H} = \exp(\mp 2\lambda r\mu t)E_\pm(f)$. A naive candidate for the linear functional Q would then be an expression like

$$\langle Q^{(\pm,r)},[f]\rangle = \int_{T_r^* M_{\mu,\lambda}} E_\mp(f)\ \Omega_r .$$

However, this does not work either because $L_{E_\pm}\Omega_r = 0$ by Proposition A.9.2, and therefore

$$\int_{T_r^* M_{\mu,\lambda}} E_\mp(f)\ \Omega_r = -\int_{T_r^* M_{\mu,\lambda}} f\ L_{E_\mp}\Omega_r = 0.$$

The following lemma will show that the linear functional $Q:\tilde{V}\to\mathsf{R}$ may indeed be realized as

a measure on some suitably adapted function subspace of $C^\infty(T^*M_{\mu,\lambda},\mathsf{R})$.

Lemma 3.3: There exist nonzero functions $q_\pm \in C^\infty(U,\mathsf{R})$ such that

$$q_\pm \circ \mathrm{Fl}_t^{X_H} = \exp(\mp(2\lambda r/\mu)\, t) q_\pm .$$

Here $U \subset T^*M_{\mu,\lambda}$ is an open subset of the phase space.

Proof: The functions $q_\pm$ must satisfy the differential equation

$$L_{X_H} q_\pm = \mp \frac{2\lambda r}{\mu} \cdot q_\pm , \tag{3.3}$$

where X_H is the Hamiltonian vector field that generates the geodesic flow. The convenient choice of coordinates used in Remark 2.6 suggests the ansatz $q_\pm = q_\pm(\psi)$ (although it is *not* the most general one). In Lemma 4.2 we will get a whole family of further solutions. The differential equations (3.3) reduce then to

$$\sin\psi \cdot \partial_\psi q_\pm = \mp q_\pm . \tag{3.4}$$

The solutions of these equations are

$$q_-(\psi) = \tan\left(\frac{\psi}{2}\right), \quad q_+(\psi) = \cot\left(\frac{\psi}{2}\right). \tag{3.5}$$

□

Since these solutions do not depend on r, they can be used without modification in the following discussion restricted to $T_r^*M_{\mu,\lambda}$.

Proposition 3.4: Let

$$V^{(r)} := \left\{ f \in C_c^\infty(T^*M_{\mu,\lambda},\mathsf{R}) \,\middle|\, \int_{T_r^*M_{\mu,\lambda}} f q_\pm \, \Omega_r < \infty \right\}. \tag{3.6}$$

Then, for $f \in V^{(r)}$, the functional

$$\langle Q^{(\pm,r)}, [f] \rangle := \int_{T_r^*M_{\mu,\lambda}} f q_\pm \, \Omega_r \tag{3.7}$$

satisfies the conditions of Lemma 3.1.2. Moreover, the space $V^{(r)}$ of functions is invariant under the geodesic flow, namely $(\mathrm{Fl}_{\mathsf{R}}^{X_H})^* V^{(r)} = V^{(r)}$. Equivalence classes of functions modulo constants of motion are written as in Proposition 2.5: $\tilde{V}^{(r)}$. Both spaces $V^{(r)}$ and $\tilde{V}^{(r)}$ are symplectic vector spaces.

Proof: Assume $\{H,f\} = 0$, then

$$\begin{aligned} 0 &= \int_{T_r^*M_{\mu,\lambda}} \{f,H\} q_\pm \, \Omega_r \\ &= \int_{T_r^*M_{\mu,\lambda}} L_{X_H}(f) q_\pm \, \Omega_r \\ &= -\int_{T_r^*M_{\mu,\lambda}} f L_{X_H}(q_\pm) \, \Omega_r = \pm \frac{2\lambda r}{\mu} \int_{T_r^*M_{\mu,\lambda}} f q_\pm \, \Omega_r , \end{aligned}$$

which shows that (in case $\lambda > 0$) the expression $\langle Q^{(\pm,r)}, [f] \rangle$ and also $\tilde{V}^{(r)}$ are well-defined. The two assertions on invariance under geodesic flow and Anosov relations are proved in one step: For $f \in V^{(r)}$ we compute

$$\begin{aligned}\langle Q^{(\pm,r)},[f\circ \mathrm{Fl}_t^{X_H}]\rangle &= \int_{T_r^* M_{\mu,\lambda}} (f\circ \mathrm{Fl}_t^{X_H}) q_\pm \ \Omega_r \\ &= \int_{T_r^* M_{\mu,\lambda}} f(q_\pm \circ \mathrm{Fl}_{-t}^{X_H}) \ \Omega_r \\ &= \exp\left[\pm\left(\frac{2\lambda r}{\mu}\right) t\right] \int_{T_r^* M_{\mu,\lambda}} f q_\pm \ \Omega_r \\ &= \exp\left[\pm\left(\frac{2\lambda r}{\mu}\right) t\right] \langle Q^{(\pm,r)},[f]\rangle.\end{aligned}$$

This shows that with $\langle Q^{(\pm,r)},[f]\rangle<\infty$ one has also $\langle Q^{(\pm,r)},[f_t]\rangle<\infty$ for all $t\in \mathsf{R}$. □

Remarks 3.5: (1) The reader will have noticed that in contrast to Lemma 3.1 the linear functional $Q^{(\pm,r)}$ is not defined on the whole of $\tilde{V}$ but we had to restrict to $\tilde{V}^{(r)}$ in order for (3.7) to make sense.

(2) That $\tilde{V}^{(r)}$ is invariant under the geodesic flow may be understood geometrically, because it can be shown that the loci where $q_\pm$ are not defined, namely $\psi=0$ or $=\pi$, are the trajectories of geodesic flow lines on the cotangent bundle. More precisely, they correspond to geodesics perpendicular to the real axis in the $\mathsf{C}^\uparrow$-picture of the Poincaré half-plane described in A2.

Definition 3.6: The Weyl algebra generated by the Weyl operators $W([f])$ for $f\in V^{(r)}$ will be referred to as *restricted Weyl algebra* and be denoted by $\mathcal{W}^{(r)}$.

Summarizing the above discussions we arrive at the main theorem of this section:

Theorem 3.7: Let $f\in V^{(r)}$, then

$$\chi_s^{(\pm,r)}(W([f])=\exp\left(i\ s\int_{T_r^* M_{\mu,\lambda}} f q_\pm \ \Omega_r\right)\cdot W([f]) \tag{3.8}$$

defines a one-parameter family of *-automorphisms of the restricted Weyl algebra $\mathcal{W}^{(r)}$ such that for all $t,s\in \mathsf{R}$ the Anosov relations hold:

$$\tau_{-t}\circ\chi_s^{(\pm,r)}\circ\tau_t=\chi^{(\pm,r)}_{\exp(\pm(2\lambda/\mu)\,rt)_s}. \tag{3.9}$$

In the sequel, the automorphisms $\chi^{(\pm,r)}:\mathsf{R}\to \mathrm{Aut}(\mathcal{W}^{(r)})$ will be referred as the *expanding*, respectively, *contracting automorphisms* or, simply, the *horocyclic automorphisms*. □

Proposition 3.8: The derivation generating the horocyclic automorphisms is given by

$$\delta^{(\pm,r)}(W([f]))=i\int_{T_r^* M_{\mu,\lambda}} f q_\pm \ \Omega_r\cdot W([f]) \tag{3.10}$$

for $f\in V^{(r)}$. The Weyl operators $W([f])$ such that $f\in V^{(r)}$ provide a dense domain (in the restricted Weyl algebra $\mathcal{W}^{(r)}$) of definition for $\delta^{(\pm,r)}$. We have the commutation relations

$$\tau_{-t}\circ\delta^{(\pm,r)}\circ\tau_t=\exp\left(\pm\frac{2\lambda r}{\mu}t\right)\delta^{(\pm,r)} \tag{3.11}$$

and, upon denoting by δ_τ the derivation generating the one-parameter group of geodesic automorphisms:

$$\delta_\tau\circ\delta^{(\pm,r)}-\delta^{(\pm,r)}\circ\delta_\tau=\frac{2\lambda r}{\mu}\delta^{(\pm,r)}. \tag{3.12}$$

Proof: Everything here follows by differentiation of the corresponding assertions in Theorem 3.7. □

IV. QUANTUM ANOSOV SYSTEMS ON COMPACT QUOTIENT MANIFOLDS

This section describes how to translate geodesic and horocyclic automorphisms defined in the previous sections on $M_{\mu,\lambda}$ to a compact quotient $\Gamma\backslash M_{\mu,\lambda}$. This will be needed in the next section where we will construct states.

Definition 4.1: We choose a fundamental domain F_Γ as described in A.14 with associated cotangent fundamental domain $T^*\check{F}_\Gamma$. $\check{F}_\Gamma$ shall denote the open interior of the (compact) fundamental domain F_Γ. We restrict the space of smooth functions on the cotangent bundle to the space of smooth Γ-periodic functions, i.e.,

$$V_\Gamma := \cap_{\gamma\in\Gamma}\{f\in C_c^\infty(T^*M_{\mu,\lambda},\mathsf{R})\,|\,f\circ U_*(\gamma)=f\}, \tag{4.1}$$

where U_* is defined in A.12. Proposition A.15 states that both the geodesic and horocyclic flows may be restricted to T^*F_Γ in a consistent way. Finally we will need the counterpart of the space of functions defined in Proposition 3.4 which shall be

$$V_{F_\Gamma}^{(r)} := \left\{ f\in V_\Gamma \,\middle|\, \int_{T_r^*F_\Gamma} f q_\pm\ \Omega_r < \infty \right\}. \tag{4.2}$$

These function spaces are endowed with a (degenerate) symplectic form $\sigma_{F_\Gamma}: V_{F_\Gamma}^{(r)}\times V_{F_\Gamma}^{(r)}\to\mathsf{R}$ by setting

$$\sigma_{F_\Gamma}(f,g) := \int_{T^*F_\Gamma} H\{f,g\}\ \omega\wedge\omega \tag{4.3}$$

for $f,g\in V_{F_\Gamma}^{(r)}$. Again, there are the nondegenerate versions $\tilde{V}_{F_\Gamma}^{(r)}$ and $\tilde{\sigma}_{F_\Gamma}$ when we factor out the constants of motion, as we did in Proposition 2.5. The Weyl algebra generated by $W([f])$ for $[f]\in\tilde{V}_{F_\Gamma}^{(r)}$ will be denoted by $\mathcal{W}_{F_\Gamma}^{(r)}$. Since, given a discrete co-compact subgroup Γ of SL(2,R), there are many choices for associated fundamental domains F_Γ, at first sight this could lead to some annoying ambiguity. The following results show that the Weyl algebra does in fact not depend on the choice of F_Γ.

Lemma 4.2: Let $g\in\mathrm{SL}(2,\mathsf{R})$ and $q_\pm$ be the functions computed in Lemma 3.3. Then $U_*(g)^*q_\pm$ also solves the equation $L_{X_H}q_\pm = \mp(2\lambda r/\mu)\,q_\pm$.

Proof: We compute

$$\begin{aligned}
L_{X_H}(U_*(g)^*q_\pm) &= X_H(q_\pm\circ U_*(g))\\
&= \{q_\pm\circ U_*(g), H\}\\
&= \{q_\pm\circ U_*(g), H\circ U_*(g)\}\\
&= \{q_\pm, H\}\circ U_*(g)\\
&= L_{X_H}q_\pm\circ U_*(g)\\
&= \mp\frac{2\lambda r}{\mu}q_\pm\circ U_*(g) = \mp\frac{2\lambda r}{\mu}U_*(g)^*q_\pm .
\end{aligned} \tag{4.4}$$

We used the fact that $U_*(g)$ acts symplectically and leaves the r-sphere bundle $T_r^*M_{\mu,\lambda}$ invariant. □

In order to accomplish a clean passage to the quotient $\Gamma\backslash M_{\mu,\lambda}$ we introduce the following notion:

Definition 4.3: Let F_Γ be an arbitrary (compact) fundamental domain with open interior $\check{F}_\Gamma$. We define the periodic extension $\mathrm{Per}_{F_\Gamma}q_\pm$ of $q_\pm$ with respect to the fundamental domain F_Γ by

$$\mathrm{Per}_{F_\Gamma}q_\pm(U_*(\gamma)\overline{x}) := q_\pm(\overline{x}) \tag{4.5}$$

for all $\gamma\in\Gamma$ and $\overline{x}\in T_r^*\check{F}_\Gamma$.

Proposition 4.4: The F_Γ-periodic extension $\mathrm{Per}_{F_\Gamma}q_\pm$ of $q_\pm$ is well-defined up to a set of measure zero in $T_r^*M_{\mu,\lambda}$. $\mathrm{Per}_{F_\Gamma}q_\pm$ is Γ-periodic and may thus be factorized to the quotient $T_r^*(\Gamma\backslash M_{\mu,\lambda})$ which shall again be denoted by $\mathrm{Per}_{F_\Gamma}q_\pm$. Let $\gamma\in\Gamma$, then there exists a diffeomorphism $R(\gamma):T_r^*(\Gamma\backslash M_{\mu,\lambda})\to T_r^*(\Gamma\backslash M_{\mu,\lambda})$ such that

$$\mathrm{Per}_{U(\gamma)F_\Gamma}q_\pm = R(\gamma)^*\mathrm{Per}_{F_\Gamma}q_\pm \tag{4.6}$$

and

$$R(\gamma)^*\mathrm{Per}_{F_\Gamma}q_\pm\circ\mathrm{Fl}_t^{X_H} = \exp\left(\mp\frac{2\lambda r}{\mu}t\right)R(\gamma)^*\mathrm{Per}_{F_\Gamma}q_\pm \tag{4.7}$$

for all $t\in\mathbb{R}$ and $\gamma\in\Gamma$. The diffeomorphism $R(\gamma):T_r^*(\Gamma\backslash M_{\mu,\lambda})\to T_r^*(\Gamma\backslash M_{\mu,\lambda})$ has the following properties:

(1) $\Pi\circ R(\gamma)=\mathrm{Id}$ where $\Pi:T_r^*(\Gamma\backslash M_{\mu,\lambda})\to\Gamma\backslash M_{\mu,\lambda}$ is the footpoint projection.
(2) $R(\gamma\gamma')=R(\gamma)R(\gamma')$ for all $\gamma,\gamma'\in\mathrm{SL}(2,\mathbb{R})$.
(3) $R(\gamma)$ keeps the volume Ω_r invariant, but is not σ_{F_Γ}-symplectic.

Proof: We may assume that F_Γ is a compact geodesic polygon and therefore $T_r^*\check{F}_\Gamma\cong\check{F}_\Gamma\times S^1$. One has

$$T_r^*M_{\mu,\lambda}\cong\cup_{\gamma\in\Gamma}(U(\gamma)F_\Gamma\times S^1)=\cup_{\gamma\in\Gamma}U(\gamma)F_\Gamma\times S^1.$$

Then the F_Γ-periodic extension $\mathrm{Per}_{F_\Gamma}q_\pm$ is well defined apart from the set

$$N_\pm:=(\cup_{\gamma\in\Gamma}U(\gamma)\partial F_\Gamma\times S^1)\cup(\cup_{\gamma\in\Gamma}U(\gamma)F_\Gamma\times U_*(\gamma).(F_\Gamma,\psi_\pm)),$$

where $\psi_+=0$ and $\psi_-=\pi$ are, in the chart of Lemma 3.3, the angles where $|q_\pm|=\infty$. ∂F_Γ denotes the boundary of the fundamental domain F_Γ. However, the set $N_\pm\subset T_r^*M_{\mu,\lambda}$ has measure zero. For $\gamma,\gamma'\in\Gamma$ and $\overline{x_0}\in T_r^*\check{F}_\Gamma$, $\overline{x}:=U_*(\gamma')\overline{x_0}\in U_*(\gamma')T_r^*\check{F}_\Gamma$ we conclude from

$$\begin{aligned}(\mathrm{Per}_{F_\Gamma}q_\pm\circ U_*(\gamma))(\overline{x}) &= \mathrm{Per}_{F_\Gamma}q_\pm(U_*(\gamma)U_*(\gamma')\overline{x_0})\\ &= \mathrm{Per}_{F_\Gamma}q_\pm(U_*(\gamma\gamma')\overline{x_0})\\ &= q_\pm(\overline{x_0})\\ &= \mathrm{Per}_{F_\Gamma}q_\pm(U_*(\gamma')\overline{x_0})\\ &= \mathrm{Per}_{F_\Gamma}q_\pm(\overline{x})\end{aligned}$$

that $\mathrm{Per}_{F_\Gamma}q_\pm\circ U_*(\gamma)=\mathrm{Per}_{F_\Gamma}q_\pm$, whence $\mathrm{Per}_{F_\Gamma}q_\pm$ is Γ-periodic. Let $F_\Gamma{}'=U(\gamma)F_\Gamma$ be another fundamental domain in the tiling defined by F_Γ. Our aim is to compare $\mathrm{Per}_{F_\Gamma}q_\pm$ and $\mathrm{Per}_{F_\Gamma{}'}q_\pm$. Since it has already been shown that both $\mathrm{Per}_{F_\Gamma}q_\pm$ and $\mathrm{Per}_{F_\Gamma{}'}q_\pm$ are Γ-periodic, we restrict to $T_r^*\check{F}_\Gamma$. With $\overline{x}\in T_r^*\check{F}_\Gamma$ we see from

$$\mathrm{Per}_{F_\Gamma{}'}q_\pm(\overline{x})=q_\pm(U_*(\gamma)\overline{x})=U_*(\gamma)^*q_\pm(\overline{x}) \tag{4.8}$$

that, restricted to $T_r^*\check{F}_\Gamma$, $\mathrm{Per}_{F_\Gamma{}'}q_\pm$ equals $U_*(\gamma)^*q_\pm$. Upon using the chart with $\overline{x}=(x,\psi)$ and

$$
\begin{aligned}
(U_*(\gamma)^* q_\pm)(x,\psi) &= q_\pm(U_*(\gamma).(x,\psi)) \\
&= q_\pm(U(\gamma).x, T_x^* U(\gamma).\psi) \\
&= (\cot,\tan)(\tfrac{1}{2}T_x^*)U(\gamma)\cdot\psi) \\
&= q_\pm(x, T_x^* U(\gamma).\psi)
\end{aligned}
$$

we conclude, again if restricted to $T_r^*\check{F}_\Gamma$, that $U_*(\gamma)^* q_\pm = R(\gamma)^* q_\pm$ with $R(\gamma).(x,\psi) := (x, T_x^* U(\gamma).\psi)$ and further

$$
\mathrm{Per}_{F_{\Gamma'}} q_\pm = \mathrm{Per}_{U(\gamma)F_\Gamma} q_\pm = R(\gamma)^* \mathrm{Per}_{F_\Gamma} q_\pm .
$$

These lines, with the definition of $R(\gamma)$, also imply that $R(\gamma\gamma') = R(\gamma)R(\gamma')$ and with Lemma 4.2 that

$$
R(\gamma)^* \mathrm{Per}_{F_\Gamma} q_\pm \circ \mathrm{Fl}_t^{X_H} = \exp\left(\mp \frac{2\lambda r}{\mu} t \right) R(\gamma)^* \mathrm{Per}_{F_\Gamma} q_\pm .
$$

We used that $U_*(\gamma)$ leaves the r-sphere bundles invariant; the explicit shape of $R(\gamma)$ shows that $\Pi \circ R(\gamma) = \mathrm{Id}$ and also that $R(\gamma)$ factors to a nontrivial diffeomorphism of the quotient $T_r^*(\Gamma\backslash M_{\mu,\lambda})$. $R(\gamma): T_r^*(\Gamma\backslash M_{\mu,\lambda}) \to T_r^*(\Gamma\backslash M_{\mu,\lambda})$ may be interpreted as a base point dependent fiber rotation of the r-sphere bundle $T_r^*(\Gamma\backslash M_{\mu,\lambda})$ over the identity on $\Gamma\backslash M_{\mu,\lambda}$ whence $R(\gamma)^* \Omega_r = \Omega_r$, i.e., $R(\gamma)$ leaves the volume Ω_r invariant, but it is not σ_{F_Γ}-symplectic. □

Remarks 4.5: (1) One has to actually consider the functions $q_\pm$ in the sense of distributions: As they occur in our computations only below the integral their definition only up to a set of measure zero does not harm either.

(2) Our strategy was to compare the function spaces $V_{F_\Gamma}^{(r)}$ over different fundamental domains F_Γ in terms of the functions $q_\pm$ occurring in their definition. To reduce the study of these functions, which are not Γ-periodic, to a fundamental domain $T_r^*\check{F}_\Gamma$ without losing information, is made easier by the very special shape of $q_\pm$: they do not depend on F_Γ but only on the angle ψ which is part of the information already contained in $\check{F}_\Gamma \times S^1 \cong T_r^*\check{F}_\Gamma$.

Proposition 4.6: Let $\gamma \in \Gamma$ and F_Γ, $F_{\Gamma'} = U(\gamma)F_\Gamma$ be two fundamental domains. Then the symplectic vector spaces of functions $V_{F_\Gamma}^{(r)}$ and $V_{F_{\Gamma'}}^{(r)}$ are symplectomorphic.

Proof: We compute

$$
\begin{aligned}
\int_{T_r^* F_{\Gamma'}} f q_\pm \, \Omega_r &= \int_{T_r^*(\Gamma\backslash M_{\mu,\lambda})} f \cdot \mathrm{Per}_{F_{\Gamma'}} q_\pm \, \Omega_r \\
&= \int_{T_r^*(\Gamma\backslash M_{\mu,\lambda})} f \cdot R(\gamma)^* \mathrm{Per}_{F_\Gamma} q_\pm \, \Omega_r \\
&= \int_{T_r^*(\Gamma\backslash M_{\mu,\lambda})} R(\gamma^{-1})^* f \cdot \mathrm{Per}_{F_\Gamma} q_\pm \, \Omega_r = \int_{T_r^* F_\Gamma} R(\gamma^{-1})^* f \cdot q_\pm \, \Omega_r
\end{aligned}
$$

from which we see that $f \in V_{F_\Gamma}^{(r)}$ if and only if $R(\gamma)^* f \in V_{F_{\Gamma'}}$. We rewrite the definition of $V_{F_\Gamma}^{(r)}$ as it was given in Definition 4.1 in terms of an arbitrary but fixed fundamental domain F_Γ:

$$
V_{U(\gamma)F_\Gamma}^{(r)} = \left\{ f \in V_\Gamma \,\middle|\, \int_{T_r^* F_\Gamma} R(\gamma^{-1})^* f \cdot q_\pm \, \Omega_r < \infty \right\}. \tag{4.9}
$$

We note that the symplectic form $\sigma_{F_{\Gamma'}}$ on $V_{U(\gamma)F_\Gamma}^{(r)}$—expressed in terms of the fundamental domain F_Γ—has to be arranged via

$$
\sigma_{U(\gamma)F_\Gamma}(f,g) = \int_{T^* F_\Gamma} H\{f,g\}_\gamma \, \omega \wedge \omega \tag{4.10}
$$

where $\{\,.\,,.\,\}_\gamma$ shall indicate that this Poisson bracket has to be formed with respect to the symplectic form $R(\gamma)^*\omega$; i.e., if $f,g\in V^{(r)}_{U(\gamma)F_\Gamma}$, then $\{f,g\}_\gamma:=(R(\gamma)^*\omega)(X_f,X_g)$ and $i_{X_f}R(\gamma)^*\omega:=df$. Finally we have

$$\{R(\gamma)^*f,R(\gamma)^*g\}_\gamma=\{f,g\}\circ R(\gamma).$$

Then one observes immediately, that the map

$$R(\gamma)^*:V^{(r)}_{F_\Gamma}\to V^{(r)}_{U(\gamma)F_\Gamma},f\mapsto R(\gamma)^*f$$

is a symplectomorphism, i.e.,

$$\begin{aligned}\sigma_{U(\gamma)F_\Gamma}(R(\gamma)f,R(\gamma)g)&=\int_{T^*F_\Gamma}H\{R(\gamma)^*f,R(\gamma)^*g\}_\gamma\ \omega\wedge\omega\\&=\int_{T^*F_\Gamma}(H\{f,g\}\circ R(\gamma))\ \omega\wedge\omega\\&=\int_{T^*F_\Gamma}H\{f,g\}\ \omega\wedge\omega=\sigma_{F_\Gamma}(f,g)\end{aligned}\tag{4.11}$$

for all $\gamma\in\Gamma$. Thereby we had to use the fact that the Hamiltonian H is $R(\gamma)$-invariant, namely $H\circ R(\gamma)=H$, and that $R(\gamma)^*(\omega\wedge\omega)=\omega\wedge\omega$, Proposition 4.4.3. □

Theorem 4.7: Let F_Γ and F'_Γ be two fundamental domains associated to a discrete cocompact subgroup Γ of SL(2,R). Then the Weyl algebras $\mathcal{W}^{(r)}_{F_\Gamma}$ and $\mathcal{W}^{(r)}_{F'_\Gamma}$ are isomorphic and will from now on be written $\mathcal{W}^{(r)}_\Gamma$. Hence the Weyl algebra $\mathcal{W}^{(r)}_\Gamma$ is uniquely determined up to an isomorphism by the subgroup $\Gamma\subset\mathrm{SL}(2,\mathsf{R})$.

Proof: We know from the discussion above that the function spaces $V^{(r)}_{F_\Gamma}$ and $V^{(r)}_{F'_\Gamma}$ are symplectomorphic. Then we give the argument that $R(\gamma)$ makes also sense on the quotient $\tilde{V}^{(r)}_{F_\Gamma}$. Suppose that $\{f,H\}=0$. Then

$$\{R(\gamma)^*f,H\}_\gamma=\{R(\gamma)^*f,R(\gamma)^*H\}_\gamma=\{f,H\}\circ R(\gamma)=0.$$

Therefore the extension of $W([f])\mapsto W([R(\gamma)^*f])$ to $\mathcal{W}^{(r)}_{V_{F_\Gamma}}$ makes sense and defines a *-isomorphism $\mathcal{F}:\mathcal{W}^{(r)}_{F_\Gamma}\to\mathcal{W}^{(r)}_{F_\Gamma'}$ from the same argument as in Proposition 2.10. □

Remark 4.8: In the sequel and in particular in Sec.V we will sometimes refer to the symplectic function spaces $(V^{(r)}_\Gamma,\sigma_\Gamma)$, respectively $(\tilde{V}^{(r)}_\Gamma,\tilde{\sigma}_\Gamma)$. This shall mean that we choose a specific fundamental domain F_Γ with function spaces $(V^{(r)}_{F_\Gamma},\sigma_{F_\Gamma})$ respectively $(\tilde{V}^{(r)}_{F_\Gamma},\tilde{\sigma}_\Gamma)$ as given in Definition 4.1, and we work with the isomorphic copy $\mathcal{W}^{(r)}_{F_\Gamma}$ of the Weyl algebra $\mathcal{W}^{(r)}_\Gamma$.

Finally we come to the main theorem in this section stating the analog of Theorem 3.7 for compact quotients $\Gamma\backslash M_{\mu,\lambda}$.

Theorem 4.9: Let F_Γ and F'_Γ be two fundamental domains with associated function spaces $V^{(r)}_{F_\Gamma}$ and $V^{(r)}_{F'_\Gamma}$ such that $F'_\Gamma=U(\gamma)F_\Gamma$ for some $\gamma\in\Gamma$. We set $\mathcal{F}:\mathcal{W}^{(r)}_{F_\Gamma}\to\mathcal{W}^{(r)}_{F'_\Gamma}$ for the induced isomorphism of Weyl algebras as described in the proof of Theorem 4.7. The geodesic automorphisms on the Weyl algebra $\mathcal{W}^{(r)}_\Gamma$ are in any case well-defined by $\tau_t W([f]):=W([f_t])$. For any $\gamma\in\Gamma$ we define the horocyclic automorphisms for $W([f])\in\mathcal{W}^{(r)}_{U(\gamma)F_\Gamma}$ as

$$\chi^{(\pm,r)}_s W([f]):=\exp\left(is\int_{T^*_rF_\Gamma}R(\gamma^{-1})^*f\cdot q_\pm\ \Omega_r\right)\cdot W([f]).\tag{4.12}$$

Then the following diagram commutes:

$$\begin{array}{ccc} \mathcal{W}^{(r)}_{F_\Gamma} & \chi_s^{(\pm,r)}_{\rightarrow} & \mathcal{W}^{(r)}_{F_\Gamma} \\ \mathcal{F}\downarrow & & \downarrow\mathcal{F} \\ \mathcal{W}^{(r)}_{F'_\Gamma} & \chi_s^{(\pm,r)}_{\rightarrow} & \mathcal{W}^{(r)}_{F'_\Gamma} \end{array}$$

Consequently the horocyclic automorphisms can unambiguously be defined on the Weyl algebra $\mathcal{W}^{(r)}_\Gamma$. Moreover, the Anosov commutation relations

$$\tau_{-t}\circ\chi_s^{(\pm,r)}\circ\tau_t=\chi^{(\pm,r)}_{\exp(\pm(2\lambda r/\mu)\,t)s}$$

also hold on $\mathcal{W}^{(r)}_\Gamma$.

Proof: Let $F'_\Gamma=U(\gamma)F_\Gamma$ and $V^{(r)}_{F'_\Gamma}=R(\gamma)^*V^{(r)}_{F_\Gamma}$. Then

$$\begin{aligned}\mathcal{F}\chi_s^{(\pm,r)}W([f])&=\exp\left(is\int_{T^*_rF_\Gamma}fq_\pm\,\Omega_r\right)\cdot W([R(\gamma)^*f])\\&=\exp\left(is\int_{T^*_rF_\Gamma}R(\gamma)^*f\cdot R(\gamma)^*q_\pm\,\Omega_r\right)\cdot W([R(\gamma)^*f])\\&=\chi_s^{(\pm,r)}W([R(\gamma)^*f])=\chi_s^{(\pm,r)}\mathcal{F}W([f]).\end{aligned}\tag{4.13}$$

In order to see the statement concerning the ACR we observe that the crucial relation occurring in the proof of Proposition 3.4 is

$$\begin{aligned}\int_{T^*_rF_\Gamma}f_t\cdot R(\gamma)^*q_\pm\,\Omega_r&=\int_{T^*_rF_\Gamma}f(R(\gamma)^*q_\pm\circ\mathrm{Fl}^{X_H}_{-t})\,\Omega_r\\&=\exp\left(\pm\frac{2\lambda r}{\mu}t\right)\int_{T^*_rF_\Gamma}f\cdot R(\gamma)^*q_\pm\,\Omega_r.\end{aligned}\tag{4.14}$$

This still holds after passing to the quotient because of Proposition 4.4. □

Remark 4.10: In the spirit of Proposition 3.4 we abbreviate the expression occurring in (4.12) by

$$\int_{T^*_rF_\Gamma}R(\gamma^{-1})^*f\cdot q_\pm\,\Omega_r=\int_{T^*_rF_\Gamma}f\cdot R(\gamma)^*q_\pm\,\Omega_r=:\langle Q^{(\pm,r)}_{U(\gamma)F_\Gamma},[f]\rangle\tag{4.15}$$

where $[f]\in\widetilde{V}^{(r)}_{U(\gamma)F_\Gamma}$. Since this is obviously independent from the choice of the fundamental domain F_Γ, namely

$$\langle Q^{(\pm,r)}_{F_\Gamma},[f]\rangle=\langle Q^{(\pm,r)}_{U(\gamma)F_\Gamma},[R(\gamma)^*f]\rangle$$

we write from now on in the same vein as in Remark 4.8 simply $Q^{(\pm,r)}_\Gamma$. Equation (4.14) then shows that

$$\langle Q^{(\pm,r)}_\Gamma,[f_t]\rangle=\exp\left(\pm\frac{2\lambda r}{\mu}t\right)\langle Q^{(\pm,r)}_\Gamma,[f]\rangle.\tag{4.16}$$

The one-parameter group of horocyclic automorphisms on the Weyl algebra $\mathcal{W}^{(r)}_\Gamma$ may now be rewritten as

$$\chi_s^{(\pm,r)}W([f])=e^{is\langle Q^{(\pm,r)}_\Gamma,[f]\rangle}\cdot W([f]).\tag{4.17}$$

Remark 4.11: We want to mention that the analysis presented above can be adapted to a slightly more general case: While we considered explicitly only fundamental domains sitting in

the same tiling as a given fundamental domain, it is actually possible to allow $F_{\Gamma'}=U(g)F_\Gamma$ for any $g\in \mathrm{SL}(2,\mathsf{R})$ instead of $g\in\Gamma$; and also to take into account ambiguities arising from Lemma 4.2, namely to consider distinct solutions of Lemma 3.3. A straightforward check shows how the arguments, essentially those in Proposition 4.4, go through in the general situation; this mostly involves tracking the effects of the conjugations $\Gamma\to g^{-1}\Gamma g$ with $g\in \mathrm{SL}(2,\mathsf{R})$.

V. INVARIANT STATES AND ERGODICITY

The aim of this section is to construct a state on the Weyl C*-algebra $\mathcal{W}$ which is invariant under the geodesic and the horocyclic automorphisms. We also investigate whether this state can be chosen to be regular in order not to lose field operators; recall that a state Φ on the Weyl algebra $\mathcal{W}$ is regular, if $\mathsf{R}\ni t\mapsto\Phi(W(t[f]))$ is continuous. To illustrate that this requirement is genuine, consider the state $\Phi_0(W([f])):=\delta_{f0}$ where δ_{f0} is the Kronecker symbol; this state is invariant under all possible flows, but it is not regular. The following proposition shows that in fact one cannot expect a state to be both regular and invariant under the horocyclic automorphism group.

Proposition 5.1: Let $\Phi:\mathcal{W}_\Gamma^{(r)}\to\mathsf{C}$ be a state on the Weyl algebra $\mathcal{W}_\Gamma^{(r)}$, which is invariant under the horocyclic automorphism group $\chi^{(\pm,r)}$. Then—if $\chi^{(\pm,r)}$ acts nontrivially—Φ cannot be a regular state.

Proof: Invariance of the state Φ under the horocyclic automorphisms implies

$$\Phi(\chi_s^{(\pm,r)}W([f]))=e^{is\langle Q_\Gamma^{(\pm,r)},[f]\rangle}\Phi(W([f]))=\Phi(W([f])) \tag{5.1}$$

for all $f\in V_\Gamma^{(r)}$. We see that either $\Phi(W([f]))=0$ or $\langle Q_\Gamma^{(\pm,r)},[f]\rangle=0$. Since we assume that the horocyclic automorphisms are nontrivial, there exists an $f\in V_\Gamma^{(r)}$ such that $\langle Q_\Gamma^{(\pm,r)},[f]\rangle\neq 0$, but then $\Phi(W([f]))=0$ and therefore $\lim_{\epsilon\to 0}\Phi(W(\epsilon[f]))=0$ whereas $\Phi(W([0]))=1$. We conclude that Φ cannot be regular. □

Definition 5.2: Let $\tilde{V}_\Gamma^{(r)}$ be the real symplectic vector space as it was described in Remark 4.8. Then for any $[f],[g]\in\tilde{V}_\Gamma^{(r)}$ with H the Hamiltonian

$$\langle[f],[g]\rangle:=\int_{T^*(\Gamma\backslash M_{\mu,\lambda})}\{f,H\}\{g,H\}\ \omega\wedge\omega \tag{5.2}$$

defines a positive definite inner product on $\tilde{V}_\Gamma^{(r)}$. The completion with respect to the norm $\|[f]\|:=\sqrt{\langle[f],[f]\rangle}$ is a Hilbert space and will be denoted by $\mathcal{V}$.

Lemma 5.3: The symplectic form $\tilde{\sigma}_\Gamma:\tilde{V}_\Gamma^{(r)}\times\tilde{V}_\Gamma^{(r)}\to\mathsf{R}$ may be separately extended to the Hilbert space $\mathcal{V}$. More precisely, for any $[f]\in\tilde{V}_\Gamma^{(r)}$ and $h\in\mathcal{V}$ the expression $\tilde{\sigma}_\Gamma([f],h)$ can be uniquely defined.

Proof: For $[f],[g]\in\tilde{V}_\Gamma^{(r)}$ we compute

$$\begin{aligned}\tilde{\sigma}_\Gamma([f],[g])&=\int_{T^*(\Gamma\backslash M_{\mu,\lambda})}H\{f,g\}\ \omega\wedge\omega\\&=\int_{T^*(\Gamma\backslash M_{\mu,\lambda})}(H\{f,g\}-\{f,gH\})\ \omega\wedge\omega\\&=-\int_{T^*(\Gamma\backslash M_{\mu,\lambda})}g\{f,H\}\ \omega\wedge\omega\end{aligned} \tag{5.3}$$

with Remark 2.4. From this we get

$$
\begin{aligned}
|\tilde{\sigma}_\Gamma([f],[g])|^2 &= \Big| \int_{T^*(\Gamma\setminus M_{\mu,\lambda})} H\{f,g\}\ \omega\wedge\omega \Big|^2 \\
&= \left| \int_{T^*(\Gamma\setminus M_{\mu,\lambda})} g\{f,H\}\ \omega\wedge\omega \right|^2 \\
&\leqslant \left(\int_{T^*(\Gamma\setminus M_{\mu,\lambda})} |g\{f,H\}|\ \omega\wedge\omega \right)^2 \\
&\leqslant \int_{T^*(\Gamma\setminus M_{\mu,\lambda})} |g|^2\ \omega\wedge\omega \cdot \int_{T^*(\Gamma\setminus M_{\mu,\lambda})} |\{f,H\}|^2\ \omega\wedge\omega \\
&= \int_{T^*(\Gamma\setminus M_{\mu,\lambda})} |g|^2\ \omega\wedge\omega \cdot \langle [f],[f]\rangle
\end{aligned} \tag{5.4}
$$

which shows, that for fixed $[g]\in\tilde{V}_\Gamma^{(r)}$ the expression $\tilde{\sigma}_\Gamma([f],[g])$ is $\|.\|$-continuous in $[f]$, hence can be extended to $\mathcal{V}$ in one entry. □

Proposition 5.4: There exists a complex structure $J:\mathcal{V}\to\mathcal{V}$ and a real positive inner product $\alpha:\mathcal{V}\times\mathcal{V}\to\mathsf{R}$ such that

$$
\alpha([f],[g]) = -\tilde{\sigma}_\Gamma(J[f],[g]) \Leftrightarrow \alpha(J[f],[g]) = \tilde{\sigma}_\Gamma([f],[g])
$$

for all $[f],[g]\in\tilde{V}_\Gamma^{(r)}$. The completion of $\tilde{V}_\Gamma^{(r)}$ with respect to α will be denoted by $\mathcal{V}^\alpha$. The inner product α is invariant under the geodesic evolution.

Proof: From Lemma 5.3 $\tilde{\sigma}([f],.)$ is a bounded linear functional on the Hilbert space $\mathcal{V}$ for all $[f]\in\tilde{V}_\Gamma^{(r)}$, therefore we may set

$$
\tilde{\sigma}_\Gamma([f],[g]) = \langle A[f],[g]\rangle \tag{5.5}
$$

for a possibly unbounded linear operator A on the Hilbert space $\mathcal{V}$ by Riesz representation theorem. A is densely defined, namely on $\tilde{V}_\Gamma^{(r)}\subset\mathcal{V}$. Skew symmetry of $\tilde{\sigma}_\Gamma$ yields $-A\subseteq A^*$, whence A is closable. W.l.o.g., let A be closed. Closed densely defined operators allow polar decomposition $A=JP$ for a partial isometry J and a positive (unbounded) operator P which is given by $P=\sqrt{A^*A}=\sqrt{-A^2}$ and therefore $A=JP=PJ$. Since the symplectic form $\tilde{\sigma}_\Gamma$ is nondegenerate, the linear operator A is injective, and the same holds for J and P. Uniqueness of polar decomposition gives finally $J^2=-\mathsf{I}$ and $J=-J^*=J^{-1}$. It is also easily checked that $\tilde{\sigma}_\Gamma$ is J-orthogonal:

$$
\begin{aligned}
\tilde{\sigma}_\Gamma(J[f],J[g]) &= \langle AJ[f],J[g]\rangle \\
&= \langle PJ^2[f],J[g]\rangle = -\langle P[f],J[g]\rangle = \langle JP[f],[g]\rangle \\
&= \langle A[f],[g]\rangle = \tilde{\sigma}_\Gamma([f],[g]).
\end{aligned} \tag{5.6}
$$

The next step defines

$$
\alpha([f],[g]) := -\tilde{\sigma}_\Gamma(J[f],[g]) \tag{5.7}
$$

and one checks that α is a positive inner product on the vector space $\tilde{V}_\Gamma^{(r)}$ which is also J-orthogonal:

$$
\alpha(J[f],J[g]) = \tilde{\sigma}_\Gamma([f],J[g]) = -\tilde{\sigma}_\Gamma(J[f],[g]) = \alpha([f],[g]). \tag{5.8}
$$

To see the invariance of α under the automorphism group generated by the geodesic flow we compute

$$
\langle A[f],[g]\rangle = \tilde{\sigma}_\Gamma([f],[g]) = \tilde{\sigma}_\Gamma([f_t],[g_t]) = \langle A[f_t],[g_t]\rangle = \langle (A[f_t])_{-t},[g]\rangle \tag{5.9}
$$

for all $g \in \mathcal{V}$, hence $(\mathrm{Fl}_t^{X_H})^* \circ A = A \circ (\mathrm{Fl}_t^{X_H})^*$ for all $t \in \mathbb{R}$, the linear operator A commutes with time evolution on the dense subspace $\tilde{V}_\Gamma^{(r)}$. Upon setting $J = A\sqrt{-A^2}$ it remains to argue that the square root $\sqrt{-A^2}$ commutes with the time flow, which is done by writing the spectral resolution of the positive operator $-A^2 = \int_0^\infty u \, dE_u$ and observation that $\sqrt{-A^2} = \int_0^\infty \sqrt{u} \, dE_u$. □

Corollary 5.5: The real inner product α and the symplectic form $\tilde{\sigma}_\Gamma$ satisfy the following relation:

$$|\tilde{\sigma}_\Gamma([f],[g])|^2 \leqslant \alpha([f],[f]) \cdot \alpha([g],[g]) \quad \forall [f],[g] \in \tilde{V}_\Gamma^{(r)}.$$

Therefore $\tilde{\sigma}_\Gamma$ may continuously be extended to the Hilbert space $\mathcal{V}^\alpha$.

Proof: For any $[f],[g] \in \tilde{V}_\Gamma^{(r)}$ we have

$$|\tilde{\sigma}_\Gamma([f],[g])|^2 = |\alpha(J[f],[g])|^2 \leqslant \alpha(J[f],J[f]) \cdot \alpha([g],[g]) = \alpha([f],[f]) \cdot \alpha([g],[g]). \tag{5.10}$$

This shows continuity of $\tilde{\sigma}_\Gamma$ with respect to the norm induced by the inner product α on the Hilbert space $\mathcal{V}^\alpha$. □

Corollary 5.6: Let $[f] \in \tilde{V}_\Gamma^{(r)}$, then for any $W([f]) \in \Delta(\tilde{V}_\Gamma^{(r)}, \tilde{\sigma}_\Gamma)$

$$W([f]) \mapsto \Phi(W([f])) := \exp(-\tfrac{1}{2}\alpha([f],[f]))$$

defines a pure quasifree, regular state on the Weyl C*-algebra $\mathcal{W}_\Gamma^{(r)}$. Further the state Φ is invariant with respect to the geodesic automorphism group: $\Phi \circ \tau_t = \Phi$ for $t \in \mathbb{R}$.

Proof: Corollary 5.5 ensures that Φ is a positive linear functional on $\Delta(\tilde{V}_\Gamma^{(r)}, \tilde{\sigma}_\Gamma)$ such that $\Phi(\mathbb{1}) = 1$; thus it can be extended to a state on $\mathcal{W}_\Gamma^{(r)}$. The general theory[8,9] implies that Φ is in fact a pure and regular state. Invariance under τ follows from Proposition 5.4, since α is invariant under the geodesic flow $\mathrm{Fl}_t^{X_H} : \tilde{V}_\Gamma^{(r)} \to \tilde{V}_\Gamma^{(r)}$. □

Remark 5.7: From the proof of Proposition 5.4 we observe that

$$\alpha([f],[g]) = -\tilde{\sigma}_\Gamma(J[f],[g]) = \langle AJ[f],[g] \rangle = \langle P[f],[g] \rangle \tag{5.11}$$

for some positive (unbounded) operator P. Corollary 5.5 states the crucial relation in order that Corollary 5.6 defines a state, in fact Corollary 5.5 ensures positivity. This leads to the conclusion that for any other positive linear operator P' such that $P \leqslant P'$ the expression

$$W([f]) \mapsto \exp\left(-\frac{\hbar}{2} \langle P'[f],[f] \rangle \right) \tag{5.12}$$

will also define a quasifree state on the Weyl algebra $\mathcal{W}_\Gamma^{(r)}$. It is known that these states are pure if and only if $P = P'$, in which case the corresponding GNS representation is equivalent to the Fock representation.[9]

Up to now we succeeded in the construction of a state on the Weyl C*-algebra $\mathcal{W}_\Gamma^{(r)}$ which is invariant under the geodesic automorphisms. The aim of the final step is to produce a state which is also invariant under the group of horocyclic automorphisms. From Lemma 5.1 we know that the following averaging procedure will destroy the regularity of the state constructed in Corollary 5.6.

Proposition 5.8: Let $\Phi : \mathcal{W}_\Gamma^{(r)} \to \mathbb{C}$ be the state constructed in Corollary 5.6; let further $Q_\Gamma^{(\pm,r)}$ be the linear functional on the vector space $\tilde{V}_\Gamma^{(r)}$ as it was explained in Remark 4.10. Then

$$\overline{\Phi}_r(W([f]) := \begin{cases} 0 & \text{if } \langle Q_\Gamma^{(\pm,r)}, [f] \rangle \neq 0 \\ \Phi(W([f]) & \text{otherwise} \end{cases}$$

is a state invariant both under the automorphism groups corresponding to the geodesic flow τ and the horocyclic flows $\chi^{(\pm,r)}$. However $\overline{\Phi}_r$ is not regular.

Proof: We denote by $\mathcal{S}(\mathcal{W}_\Gamma^{(r)})$ the set of states of $\mathcal{W}_\Gamma^{(r)}$. From (4.16) and Corollary 5.6 it follows that $\overline{\Phi}_r \in \mathcal{S}(\mathcal{W}_\Gamma^{(r)})$ is invariant under the geodesic automorphism group τ_t. Further from Theorem 4.9, one sees that, for any $\Phi \in \mathcal{S}(\mathcal{W}_\Gamma^{(r)})$:

$$\Phi(\chi_s^{(\pm,r)} W([f])) = \Phi(e^{is\langle Q_\Gamma^{(\pm,r)},[f]\rangle} W([f])) = e^{is\langle Q_\Gamma^{(\pm,r)},[f]\rangle} \Phi(W([f])). \tag{5.13}$$

One possible averaging procedure is given by

$$\begin{aligned} \overline{\Phi}_r(W([f])) &:= \lim_{s\to\infty} \frac{1}{2s} \int_{-s}^{s} \Phi(\chi_u^{(\pm,r)} W([f]))\, du \\ &= \lim_{s\to\infty} \frac{1}{2s} \int_{-s}^{s} e^{iu\langle Q_\Gamma^{(\pm,r)},[f]\rangle}\, du \cdot \Phi(W([f])) \\ &= \begin{cases} 0 & \text{if } \langle Q_\Gamma^{(\pm,r)},[f]\rangle \neq 0 \\ \Phi(W([f])) & \text{otherwise}. \end{cases} \end{aligned} \tag{5.14}$$

By construction $\overline{\Phi}_r \in \mathcal{S}(\mathcal{W}_\Gamma^{(r)})$ is an invariant state under the horocyclic automorphisms $\chi_s^{(\pm,r)}$. □

Remark 5.9: We want to stress the fact, that for any $r>0$ we constructed a Weyl algebra $\mathcal{W}_\Gamma^{(r)}$ and a state $\overline{\Phi}_r$ which is invariant with respect to *both* the horocyclic automorphisms $\chi_s^{(\pm,r)}$ *and* the dynamics τ_t. Hence, the models described in this paper do satisfy all the defining conditions that lead to the very strong ergodic properties obtained in Ref. 1: they are quantum Anosov flows.

VI. CONCLUSIONS AND OUTLOOK

While the usual irreducible representations—of the system of imprimitivity based on configuration space—led to the impasse already noted in Ref. 1, and discussed more specifically in Appendix B here, our extension to the larger Weyl algebras $\mathcal{W}^{(r)}$ and $\mathcal{W}_\Gamma^{(r)}$ (see Definitions 3.6 and 4.1; and Theorem 4.7) now makes it possible to implement quantum versions of *both* the geodesic *and* the horocyclic flows (see Propositions 2.10 and 3.4; and Theorems 3.7 and 4.9); in particular, these groups of automorphisms satisfy the Anosov commutation relation proposed in Ref. 1, [see here Eq. (3.9) of Theorem 3.7].

As for future extensions of the present work, we think it could be interesting to investigate in more details the properties of the Weyl algebras we constructed here, explore their full physical interpretation, and in the flat case ($\lambda = 0$), to relate them to one of the known Weyl algebras.[11] In view of our Proposition 3.8 and of the general classical structures,[4] it might also be reasonable to try and extend the work of Ref. 1 in terms of relations between *derivations*: compare to the relations (3.11) and (3.12) obtained here for the derivations generating our geodesic and horocyclic automorphisms.

ACKNOWLEDGMENTS

I.P. wishes to thank the Mathematics Department at the University of Florida, Gainesville, for the hospitality he enjoyed during the academic year 1996–1997. Both authors are grateful to Georg Peter, Director of the Sonderpädagogisches Zentrum, Rankweil-Vorderland, Austria for the Zentrum's hospitality in July 1997. I.P. also wants to acknowledge here the partial support of the Land Government of Vorarlberg, Austria.

APPENDIX A: CLASSICAL ANOSOV STRUCTURES FOR THE HYPERBOLIC GEODESIC FLOW

The main purpose of this Appendix is to fix the notation for the classical objects we want to quantize. It may also serve two ancillary functions: to provide an essentially self-contained presentation of the classical structures; and to have a notation flexible enough to show readily how these structures collapse in the limit where the curvature of the configuration space vanishes.

1. Flows on the configuration spaces $M_{\mu,\lambda}$

Definitions A.1: $(M_{\mu,\lambda},g)$ (or simply $M_{\mu,\lambda}$) denotes the Riemannian manifold $\mathbb{R}^2$, with Cartesian coordinates (x^1,x^2), and metric

$$g(x^1,x^2)=\mu^2\ dx^1\otimes dx^1+e^{-4\lambda x^1}dx^2\otimes dx^2 \tag{A1}$$

where $\mu>\lambda\geqslant 0$ are left arbitrary. Upon passing to an orthonormal coframe $\{\sigma^1,\sigma^2\}$ given by

$$\sigma^1(x^1,x^2)=\mu\ dx^1,\sigma^2(x^1,x^2)=e^{-2\lambda x^1}dx^2 \tag{A2}$$

and its dual orthonormal frame field $\{s_1,s_2\}$ defined via $\langle\sigma^i,s_j\rangle:=\delta^i_j$ we write $g^\sharp=s_1\otimes s_1+s_2\otimes s_2$ for the cometric. The cotangent bundle $T^*M_{\mu,\lambda}$ (respectively, the tangent bundle $TM_{\mu,\lambda}$) shall be parametrized as

$$\begin{aligned} T^*M_{\mu,\lambda}\ni(x^i,p_i)\equiv(x^1,x^2,p_1,p_2)\equiv p_1\sigma^1(x^1,x^2)+p_2\sigma^2(x^1,x^2),\\ TM_{\mu,\lambda}\ni(x^i,v^i)\equiv(x^1,x^2,v^1,v^2)\equiv v^1s_1(x^1,x^2)+v^2s_2(x^1,x^2). \end{aligned} \tag{A3}$$

Thus p_i (respectively, v^i) appear as coordinates with respect to the chosen orthonormal coframe $\{\sigma^1,\sigma^2\}$ (respectively, the frame $\{s_1,s_2\}$). The identification of $TM_{\mu,\lambda}$ and $T^*M_{\mu,\lambda}$ via the metric g is now written as $(\Sigma p_i\sigma^i)^\sharp=\Sigma p_is_i$ and conversely $(\Sigma v^is_i)^\flat=\Sigma v^i\sigma^i$. For the ''length'' of covectors (respectively, vectors), we introduce the notations:

$$\begin{aligned} |(x^1,x^2,p_1,p_2)|:=\sqrt{g^\sharp((x^i,p_i),(x^i,p_i))}(=\sqrt{{p_1}^2+{p_2}^2}),\\ |(x^1,x^2,v^1,v^2)|:=\sqrt{g((x^i,v^i),(x^i,v^i))}(=\sqrt{(v^1)^2+(v^2)^2}). \end{aligned} \tag{A4}$$

Proposition A.2: $(M_{\mu,\lambda},g)$ is a two-dimensional, simply connected Riemannian manifold of constant negative (sectional) curvature $\kappa=-4\lambda^2/\mu^2$. For $\mu=2\lambda$

$$\Phi:M_{\mu,\lambda}\to\mathbb{C}^\uparrow,(x^1,x^2)\mapsto x^2+ie^{2\lambda x^1}$$

establishes an isometry from $M_{\mu,\lambda}$ onto Poincaré half-plane $\mathbb{C}^\uparrow:=\{z\in\mathbb{C}\mid \mathrm{Im}\ z>0\}$. In the case $(\mu,\lambda)=(1,0)$, the model becomes the flat $\mathbb{R}^2$.

Proof: Upon using Cartan's structure equations

$$d\sigma^i=-\sum_k\ \omega^i{}_k\wedge\sigma^k,\quad \Omega^i{}_k=d\omega^i{}_k+\sum_j\ \omega^i{}_j\wedge\omega^j{}_k \tag{A5}$$

one readily computes the connection one-form $\omega^1{}_2=(2\lambda/\mu)\ \sigma^2$ and the curvature two-form $\Omega^1{}_2=-(4\lambda^2/\mu^2)\ \sigma^1\wedge\sigma^2$. From these expressions one gets for the (sectional) curvature κ,

$$\kappa(s_1,s_2)=-g(\Omega^2{}_1(s_1,s_2)s_2,s_2)=-\frac{4\lambda^2}{\mu^2}(\sigma^1\wedge\sigma^2)(s_1,s_2)g(s_2,s_2)=-\frac{4\lambda^2}{\mu^2}. \tag{A6}$$

In order to see that Φ is the required isometry one observes that

$$\Phi^*\left(-4\frac{dz\otimes d\bar{z}}{(z-\bar{z})^2}\right)=g \tag{A7}$$

where the expression in the bracket on the left-hand side is the metric of Poincaré half-plane. □

Remark A.3: The advantage of this model for Lobatchevski hyperbolic geometry over Poincaré half-plane or the open-unit-disk model in $\mathbb{C}$ is that it allows one to follow immediately the effect of the ''flattening'' $(\mu,\lambda)\to(1,0)$ since the underlying manifold is already $\mathbb{R}^2$.

Lemma A.4: Let $(x^i,v^i)\in TM_{\mu,\lambda}$ with $0<|(x^i,v^i)|=r=\mathrm{const}$, then:

(1) There exist two vector fields $G_\pm \in \mathcal{X}(M_{\mu,\lambda})$ such that $\nabla_{G_\pm} G_\pm = 0$ and $G_\pm(x^1,x^2) = v^1 s_1(x^1,x^2) + v^2 s_2(x^1,x^2)$: $G_\pm$ are geodesic vector fields having the geodesic through (x^1,x^2) with initial tangent vector (x^i,v^i) as a flow line. The vector fields $G_\pm$ may be chosen such that $g(G_\pm,G_\pm) = r^2 = \text{const}$, and the $G_\pm$ are complete.
(2) For these geodesic vector fields $G_\pm \in \mathcal{X}(M_{\mu,\lambda})$, there exist vector fields $H_\pm \in \mathcal{X}(M_{\mu,\lambda})$ such that $g(G_\pm,H_\pm) = 0$ and $g(H_\pm,H_\pm) = r^2$. The vector fields $H_\pm$ are complete.
(3) The above vector fields satisfy the relations $[G_\pm,H_\pm] = \pm(2\lambda r/\mu) H_\pm$. The integral version of this relation in terms of the respective flows reads

$$\mathrm{Fl}_{-t}^{G_\pm} \circ \mathrm{Fl}_s^{H_\pm} \circ \mathrm{Fl}_t^{G_\pm} = \mathrm{Fl}^{H_\pm}_{\exp(\pm(2\lambda/\mu) rt)_s}. \quad (*)$$

Proof: Choose $(x^1,x^2,v^1,v^2) \in TM_{\mu,\lambda}$ and set $v^1 = r \cos\psi$, $v^2 = r \sin\psi$. We first treat the case $\cos\psi \neq \pm 1$. For $(y^1,y^2) \equiv y \in \mathbb{R}^2$ consider the functions

$$z^\pm(y^1,y^2) = \pm \frac{2\lambda}{\mu} \mathrm{e}^{-2\lambda y^1}(K_\pm - y^2), \quad \text{(A8)}$$

where

$$K_\pm := x^2 \pm \frac{\mu}{2\lambda} e^{2\lambda x^1} \frac{\sin\psi}{1 \mp \cos\psi} \quad \text{(A9)}$$

are determined by the chosen (x^i,v^i). Upon setting

$$G_\pm(y) = \mp r \frac{1 - z_\pm(y)^2}{1 + z_\pm(y)^2} s_1 + r \frac{2 z_\pm(y)}{1 + z_\pm(y)^2} s_2 \quad \text{(A10)}$$

one checks that the vector fields $G_\pm$ are geodesic vector fields indeed. For the vector fields $H_\pm$ we set

$$H_\pm(y) = -r \frac{2 z_\pm(y)}{1 + z_\pm(y)^2} s_1 \mp r \frac{1 - z_\pm(y)^2}{1 + z_\pm(y)^2} s_2. \quad \text{(A11)}$$

From this construction we observe that in case $\cos\psi = 1$ the vector fields G_- and H_- can also be defined: we set $G_+ := rs^1$ and $H_+ := rs^2$. The same applies in the case $\cos\psi = -1$ for G_+ and H_+, where we define $G_- := -rs^1$ and $H_- := -rs^2$. Then $g(G_\pm,H_\pm) = 0$ by construction. A straightforward computation shows that these vector fields also satisfy $[G_\pm,H_\pm] = \pm(2\lambda r/\mu) H_\pm$. As $M_{\mu,\lambda}$ is a complete Riemannian manifold, we conclude from the Hopf–Riemann theorem that maximal geodesics are defined on $\mathbb{R}$ and hence the vector fields $G_\pm$ and also $H_\pm$ are complete. The commutation relation for the flows can be shown generally: assume the flows of two vector fields $X,Y \in \mathcal{X}(M)$ on a manifold M satisfy

$$\mathrm{Fl}_{-t}^X \circ \mathrm{Fl}_s^Y \circ \mathrm{Fl}_t^X = \mathrm{Fl}^Y_{e^{\rho t} s} \quad \text{(A12)}$$

for some $\rho \in \mathbb{R}$. Taking the derivative with respect to s at $s = 0$ of the left-hand side ($x \in M$) yields

$$\frac{d}{ds}\Big|_{s=0} (\mathrm{Fl}_{-t}^X \circ \mathrm{Fl}_s^Y \circ \mathrm{Fl}_t^X)(x) = T(\mathrm{Fl}_{-t}^X).Y(\mathrm{Fl}_t^X(x)) = (\mathrm{Fl}_t^X)^* Y(x) \quad \text{(A13)}$$

and for the right-hand side

$$\frac{d}{ds}\Big|_{s=0} \mathrm{Fl}^Y_{e^{\rho t} s}(x) = e^{\rho t} Y(x) \Rightarrow (\mathrm{Fl}_t^X)^* Y = e^{\rho t} Y. \quad \text{(A14)}$$

Finally upon taking the derivative with respect to t at $t = 0$ we obtain $L_X Y = [X,Y] = \rho Y$ which completes the proof. □

Remark A.5: (1) The vector fields $G_\pm, H_\pm \in \mathcal{X}(M_{\mu,\lambda})$ described in the above lemma (or more precisely their liftings to the vector fields $X_H, E_\pm \in \mathcal{X}(T^*M_{\mu,\lambda})$ defined in Corollary A.8 below) are a special case of a *(classical) Anosov system*. The commutation relations (*) of Lemma A.4 shall be referred to as *Anosov commutation relations* (ACR). The constants $\pm(2\lambda/\mu)\,r$ appearing in the ACR are the *Lyapunov coefficients* for this system. The vector fields $H_\pm$ are the expanding, respectively, contracting *horocyclic vector field*.

(2) The geodesic through (x^1,x^2) with initial tangent vector (x^i,v^i) is an integral curve of both vector fields $G_\pm$.

(3) In two dimensions the vector fields $H_\pm$ are determined by $G_\pm$ uniquely (up to a nonzero function) because the metric is nondegenerate. The vector fields $G_\pm$ and $H_\pm$ depend on the chosen point (x^i,v^i).

(4) Concerning property (3) in Lemma A.4, one can generally conclude that

$$[G,H]=fH \tag{A15}$$

where f is a function. To be more precise, assume

$$\nabla_G G=0, \qquad g(G,H)=0, \quad g(G,G)=\text{const} \tag{A16}$$

then

$$\begin{gathered} g(\nabla_G H,G)=\nabla_G(g(H,G))-g(H,\nabla_G G)=0, \\ g(\nabla_H G,G)=\nabla_H(g(G,G))-g(G,\nabla_H G), \\ g(\nabla_H G,G)=0, \end{gathered} \tag{A17}$$

and therefore, since $g([H,G],G)=g(\nabla_H G,G)-g(\nabla_G H,G)$,

$$g([H,G],G)=0 \tag{A18}$$

and, since the manifold is two dimensional, one concludes that $[H,G]=fH$ for some function f. Note that the conditions on G and H listed at the beginning of this remark would be too weak to conclude that the ''Lyapunov coefficient'' is a constant indeed: Pick a point in $M_{\mu,\lambda}$ and consider all geodesics emanating from this point. After removing the point, this family of geodesics determines a geodesic vector field as well as a vector field that is perpendicular to the geodesic vector field; this is in fact an example where the Lyapunov coefficient is not a constant but a function. However, in this case, the vector fields are not complete either.

(5) A similar situation occurs in deSitter space–time, where one may also implement an Anosov system, however here the ''Lyapunov coefficients'' are only constant on the horizons.[12]

2. Flows on the phase spaces $T^*M_{\mu,\lambda}$

Proposition A.6: The flows $\mathrm{Fl}_t^{G_\pm}$, respectively, $\mathrm{Fl}_s^{H_\pm}$ of the geodesic, respectively, horocyclic vector fields are lifted to the cotangent bundle in the following manner:

$$\mathrm{geo}((x^i,p_i);t):=G_\pm^\flat(\mathrm{Fl}_t^{G_\pm}(x^i)), \qquad h_\pm((x^i,p_i);s):=G_\pm^\flat(\mathrm{Fl}_s^{H_\pm}(x^i)),$$

where the geodesic vector fields $G_\pm$ depend on the point (x^i,p_i) and have to be chosen such that $G_\pm^\flat(x^i)=(x^i,p_i)$. The actions geo and $h_\pm$ mapping $T^*M_{\mu,\lambda}\times\mathsf{R}$ to $T^*M_{\mu,\lambda}$ will be referred to as geodesic coflow and (cotangent) horocyclic flows. The flows geo and $h_\pm$ leave the r-sphere bundle $T_r^*M_{\mu,\lambda}$ invariant, where $T_r^*M_{\mu,\lambda}:=\{(x^i,p_i)\in T^*M_{\mu,\lambda} \mid |(x^i,p_i)|=r\}$. They again satisfy the ACR with the same Lyapunov coefficients:

$$\mathrm{geo}(h_\pm(\mathrm{geo}(x^i,p_i;t);s);-t)=h_\pm\left(x^i,p_i;\exp\left(\pm\frac{2\lambda}{\mu}rt\right)s\right).$$

Proof: $\mathrm{geo}: T^*M_{\mu,\lambda}\times \mathbb{R} \to T^*M_{\mu,\lambda}$ coincides with the common geodesic coflow through (x^i,p_i) and does not depend on which of the two $G_\pm$'s we choose. That $h_\pm : \mathbb{R}\times T^*M_{\mu,\lambda}\times \mathbb{R} \to T^*M_{\mu,\lambda}$ is actually a flow may be seen from

$$h_\pm((x^i,p_i);0) = G_\pm^\flat(\mathrm{Fl}_0^{H_\pm}(x^i)) = G_\pm^\flat(x^i) = (x^i,p_i) \tag{A19}$$

and

$$\begin{aligned} h_\pm(h_\pm((x^i,p_i);s_1);s_2) &= h_\pm(G_\pm^\flat(\mathrm{Fl}_{s_1}^{H_\pm}(x^i));s_2) \\ &= G_\pm^\flat(\mathrm{Fl}_{s_2}^{H_\pm}(\mathrm{Fl}_{s_1}^{H_\pm}(x^i))) \\ &= G_\pm^\flat(\mathrm{Fl}_{s_1+s_2}^{H_\pm}(x^i)) = h_\pm((x^i,p_i);s_1+s_2). \end{aligned} \tag{A20}$$

From the same lines one concludes that the ACR are fulfilled:

$$\begin{aligned} \mathrm{geo}(h_\pm(\mathrm{geo}(x^i,p_i;t);s);-t) &= \mathrm{geo}(h_\pm(G_\pm^\flat(\mathrm{Fl}^{G_\pm}(x^i));s);-t) \\ &= \mathrm{geo}(G_\pm^\flat(\mathrm{Fl}_s^{H_\pm}\circ \mathrm{Fl}_t^{G_\pm}(x^i));-t) \\ &= G_\pm^\flat(\mathrm{Fl}_{-t}^{G_\pm}\circ \mathrm{Fl}_s^{H_\pm}\circ \mathrm{Fl}_t^{G_\pm}(x^i)) \\ &= G_\pm^\flat\left(\mathrm{Fl}^{H_\pm}_{\exp\left(\pm\frac{2\lambda}{\mu}rt\right)s}(x^i)\right) = h_\pm\left(x^i,p_i;\exp\left(\pm\frac{2\lambda}{\mu}rt\right)s\right). \end{aligned} \tag{A21}$$

Since $g^\sharp(G_\pm^\flat,G_\pm^\flat)=r^2=\mathrm{const}$, the flows geo and $h_\pm$ are contained in $T_r^*M_{\mu,\lambda}$ by construction and Lemma A.4.1 and 2. □

3. Symplectic structures

So far we have given an explicit description of an Anosov system on the r-sphere bundle $T_r^*M_{\mu,\lambda}$ of Poincaré half-plane. We still need to state some properties of the geodesic and horocyclic flows which we will use in the quantization carried out in the body of the paper.

Any cotangent bundle carries the structure of a symplectic manifold in a natural manner; in our case the (nondegenerate) symplectic two-form $\omega\in\Omega^2(T^*M_{\mu,\lambda})$ is given at the point $(x^i,p_i)\in T^*M_{\mu,\lambda}$ by

$$\omega_{(x^i,p_i)} = \sum_{j=1}^{2}(\sigma^j\wedge dp_j - p_j d\sigma^j) = \sigma^1\wedge dp_1 + \sigma^2\wedge dp_2 + \frac{2\lambda}{\mu}p_2\ \sigma^1\wedge\sigma^2. \tag{A22}$$

A smooth function $f\in C^\infty(T^*M_{\mu,\lambda},\mathbb{R})$ determines a Hamiltonian vector field $X_f\in\mathcal{X}(T^*M_{\mu,\lambda})$ via $i_{X_f}\omega = df$. Moreover in the sequel we will need a Poisson bracket which is defined as $\{f,g\} := \omega(X_f,X_g)$ for $f,g\in C^\infty(T^*M_{\mu,\lambda},\mathbb{R})$. $\omega\wedge\omega$ will serve as the volume on phase space $T^*M_{\mu,\lambda}$.

Remark A.7: The Hamiltonian flow on the cotangent bundle $T^*M_{\mu,\lambda}$ determined by the Hamiltonian

$$H(x^i,p_i) := \tfrac{1}{2}g^\sharp((x^i,p_i),(x^i,p_i)) = \tfrac{1}{2}(p_1^2+p_2^2) \tag{A23}$$

coincides with the geodesic coflow as given in Proposition A.6.[13] Therefore the corresponding Hamiltonian vector field X_H is in turn the geodesic cospray on the cotangent bundle. The r-sphere bundle may now also be defined as $H^{-1}(r^2/2)$ for $r>0$. Moreover there exists a unique volume Ω_r on $T_r^*(M_{\mu,\lambda})$ such that $\omega\wedge\omega = \Omega_r\wedge dH$.

Corollary A.8: Upon using the notation of Proposition A.6 and Remark A.7, we may define the vector fields $X_H, E_\pm\in\mathcal{X}(T^*M_{\mu,\lambda})$

$$\mathrm{geo}((x^i,p_i);t) = \mathrm{Fl}_t^{X_H}(x^i,p_i) \quad \text{and} \quad h_\pm((x^i,p_i);s) =: \mathrm{Fl}_s^{E_\pm}(x^i,p_i),$$

where $X_H, E_\pm \in \mathcal{X}(T^*M_{\mu,\lambda})$ denote the vector fields generating the geodesic flow and the horocyclic flows on $T^*M_{\mu,\lambda}$. Once restricted to $T_r^*M_{\mu,\lambda}$, they satisfy

$$[X_H, E_\pm] = \pm \frac{2\lambda r}{\mu} E_\pm .$$

Proof: This follows immediately from Lemma A.4.3 and the conclusion of Proposition A.6, with the ACR rewritten as:

$$\mathrm{Fl}_{-t}^{X_H} \circ \mathrm{Fl}_s^{E_\pm} \circ \mathrm{Fl}_t^{X_H} = \mathrm{Fl}_{\exp(\pm(2\lambda/\mu) rt)s}^{E_\pm}. \tag{A24}$$

□

Proposition A.9:

(1) The geodesic co-flow geo: $T_r^*M_{\mu,\lambda} \times \mathbb{R} \to T_r^*M_{\mu,\lambda}$ is a one-parameter group of symplectomorphisms.
(2) The horocyclic flows $h_\pm : T_r^*M_{\mu,\lambda} \times \mathbb{R} \to T_r^*M_{\mu,\lambda}$ do not consist of symplectomorphisms; nevertheless $h_\pm$ leave the volume Ω_r on $T_r^*(M_{\mu,\lambda})$ invariant.

Proof: Ad (1): From the definition of geo, we see that

$$\mathrm{geo}((x^i, p_i); t) := G_\pm^\flat(\mathrm{Fl}_t^{G_\pm}(x^i)) = T_{x^i}^* \mathrm{Fl}_t^{G_\pm} . G_\pm^\flat(x^i)$$

whence the geodesic coflow arises as lift of the diffeomorphisms $\{\mathrm{Fl}_t^{G_\pm} | t \in \mathbb{R}\} \subset \mathrm{Diff}(M_{\mu,\lambda})$ and is therefore symplectic.[13] (An alternate argument would be that the geodesic coflow arises as a Hamiltonian flow). *Ad* (2): Since $h_\pm(T_r^*M_{\mu,\lambda}; s) \subset T_r^*M_{\mu,\lambda}$ $\forall s \in \mathbb{R}$ we have $L_{E_\pm} H = 0$. Assume $\mathrm{Fl}_s^{E_\pm}$ are symplectic $\Leftrightarrow L_{E_\pm}\omega = 0$, then

$$0 = dL_{E_\pm} H = L_{E_\pm} dH = L_{E_\pm} i_{X_H}\omega = i_{[E_\pm, X_H]}\omega + i_{X_H} L_{E_\pm}\omega = \mp \frac{2\lambda r}{\mu} \cdot i_{E_\pm}\omega$$

which yields $E_\pm = 0$, a contradiction if $\lambda > 0$. If $\lambda = 0$ we arrive in the flat case, where the statement may be checked by explicit computation. In order to see that $h_\pm(s)^*\Omega_r = \Omega_r$, it suffices to check that $h_\pm(s)^*(\omega \wedge \omega) = \omega \wedge \omega$; this again is done by explicit computation. □

4. Flows on the unit cotangent bundle of compact quotients $\Gamma \backslash M_{\mu,\lambda}$

Finally, we review in this subsection a family of classical hyperbolic flows that admit a natural invariant probability measure. We recall first[14] that a Riemannian manifold M of dimension $n \geqslant 2$ is complete, connected, and of constant negative curvature $\kappa < 0$, if and only if it is isometric to a quotient $\Gamma \backslash H_\kappa^{(n)}$, where $H_\kappa^{(n)}$ is *the* simply connected n-dimensional hyperbolic space of curvature κ ; and Γ is a group of isometries of $H_\kappa^{(n)}$ acting: (i) freely (i.e., for every $x \in H_\kappa^{(n)}$: $\Gamma_x \equiv \{\gamma \in \Gamma | \gamma[x] = x\}$ is trivial); and (ii) properly discontinuously (i.e., for every $x \in H_\kappa^{(n)}$: there exists a neighborhood U of x such that $\Gamma_U \equiv \{\gamma \in \Gamma | \gamma[U] \cap U \neq \varnothing\}$ is finite; the latter condition implying that the orbit space is Hausdorff. In order to exploit the detailed information gathered on the geodesic flow on the universal covering space $H_\kappa^{(n)}$, we restrict our attention to the case of two-dimensional manifolds. This, however, still allows for spectral pathologies of the type exemplified by the Selberg trace formula[15] where $\Gamma (= \mathrm{SL}(2,\mathbb{Z}))$ acts freely and even properly discontinuously on the Poincaré half-plane, with nevertheless $\Gamma \backslash H_\kappa^{(2)}$ *not* compact, although it is of finite volume. To obtain a good ergodic behavior, based on a natural invariant probability measure, we will henceforth assume that Γ is co-compact. The literature on these manifolds is extensive: see, e.g., Refs. 16–19; for a concrete example of a compact (i.e., not just finite-volume) fundamental domain, see e.g., Ref. 20.

Proposition A.10: If $\lambda > 0$ the symmetry group of $M_{\mu,\lambda}$ is $\mathrm{SL}(2,\mathbb{R})$. Since $\dim M_{\mu,\lambda} = 2$ and $\dim \mathrm{SL}(2,\mathbb{R}) = 3$, the Riemannian manifold $M_{\mu,\lambda}$ is a maximally symmetric space. The isotropy groups are all isomorphic to $\mathrm{SO}(2,\mathbb{R})$. The group H_2 occurring in the Iwasawa decomposition $\mathrm{SL}(2,\mathbb{R}) = H_2 \cdot \mathrm{SO}(2,\mathbb{R})$ acts freely and transitively on $M_{\mu,\lambda}$.

Proof: Solving Killing equation $L_K g = 0$ provides us with three linear independent Killing vector fields $K_i \in \mathcal{X}(M_{\mu,\lambda}), i \in \{1,2,3\}$,

$$K_1(x^1,x^2) = \mu s^1 + 2\lambda x^2 e^{-2\lambda x^1} s^2,$$

$$K_2(x^1,x^2) = e^{-2\lambda x^1} s^2, \tag{A25}$$

$$K_3(x^1,x^2) = -\mu x^2 s^1 + \left(\frac{\mu^2}{2\lambda} \sinh(2\lambda x^1) - \lambda (x^2)^2 e^{-2\lambda x^1} \right) s^2.$$

We remark that for $(\mu,\lambda) = (1,0)$ these Killing vector fields generate the Euclidean group $E_2 = \mathrm{SO}(2,\mathbb{R}) \cdot \mathbb{R}^2$. In general, the Lie brackets are given by

$$[K_1,K_2] = -2\lambda K_2, \quad [K_1,K_3] = \mu^2 K_2 + 2\lambda K_3, \quad [K_2,K_3] = -K_1. \tag{A26}$$

If $\lambda > 0$ the applications

$$\hat{K}_1 := \frac{1}{\lambda} K_1 \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\hat{K}_2 := \frac{1}{\sqrt{\lambda}} K_2 \mapsto \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{A27}$$

$$\hat{K}_3 := \frac{1}{\sqrt{\lambda}} \left(\frac{\mu^2}{4\lambda} K_2 + K_3 \right) \mapsto \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

provide a Lie algebra (anti-) isomorphism of the Lie algebra of Killing vector fields onto the Lie algebra $\mathsf{sl}(2,\mathbb{R})$ of the special linear group $\mathrm{SL}(2,\mathbb{R})$. $K_3(0,0) = 0$ and therefore generates the isotropy group at the point $(0,0) \in M_{\mu,\lambda}$, which is one-dimensional and has to be compact, whence is $\mathrm{SO}(2,\mathbb{R})$. Since $M_{\mu,\lambda}$ is connected, the isotropy groups at different points are all isomorphic via conjugation. The vector fields K_1 and K_2 have no zeros on $M_{\mu,\lambda}$, hence the group H_2 generated by their flows acts effectively; a dimension argument shows transitivity. □

Lemma A.11: Let $U: \mathrm{SL}(2,\mathbb{R}) \to \mathrm{Isom}(M_{\mu,\lambda}), \gamma \mapsto U(\gamma)$ denote the representation of $\mathrm{SL}(2,\mathbb{R})$ as isometries on $M_{\mu,\lambda}$. Then we have:

(1) $\Phi \circ \mathrm{Fl}_t^X = \mathrm{Fl}_t^{\Phi_* X} \circ \Phi$ for $\Phi \in \mathrm{Diff}(M_{\mu,\lambda})$ and $X \in \mathcal{X}(M_{\mu,\lambda})$.
(2) $U(\gamma)_* G_\pm$ are again geodesic vector fields.
(3) $g(U(\gamma)_* G_\pm, U(\gamma)_* H_\pm) = 0$ for all $\gamma \in \mathrm{SL}(2,\mathbb{R})$.

Proof: *Ad* (1) $\Phi_* X := T\Phi \circ X \circ \Phi^{-1}$, consider

$$\frac{d}{dt}(\Phi \circ \mathrm{Fl}_t^X) = T\Phi . \frac{d}{dt} \mathrm{Fl}_t^X = T\Phi \circ X \circ \mathrm{Fl}_t^X = \Phi_* X (\Phi \circ \mathrm{Fl}_t^X) \tag{A28}$$

and

$$\Phi(\mathrm{Fl}_0^X(x)) = \Phi(x). \tag{A29}$$

This shows (1). *Ad* (2): Remember that isometries map geodesics to geodesics, then use (1). *Ad* (3): This is clear. □

Definition A.12: We define the induced action $U_*: \mathrm{SL}(2,\mathbb{R}) \to \mathrm{Diff}(T_r^* M_{\mu,\lambda})$ of the group $\mathrm{SL}(2,\mathbb{R})$ on the r-sphere bundle $T_r^* M_{\mu,\lambda}$ by

$$U_*: \mathrm{SL}(2,\mathbb{R}) \to \mathrm{Diff}(T_r^* M_{\mu,\lambda}), \gamma \mapsto U_*(\gamma) := ((x^i,p_i) \mapsto (U(\gamma).x^i, T^*_{x^i} U(\gamma).p_i)) \tag{A30}$$

where $U: \mathrm{SL}(2,\mathbb{R}) \to \mathrm{Diff}(M_{\mu,\lambda})$ is the action on the configuration space $M_{\mu,\lambda}$.

Theorem A.13: Let $U_*:\mathrm{SL}(2,\mathbb{R}) \to \mathrm{Diff}(T_r^* M_{\mu,\lambda})$ be the representation of $\mathrm{SL}(2,\mathbb{R})$ on the r-sphere bundle as defined in A.12. Then $U_*(\gamma)$ commutes with both the geodesic and horocyclic flows:

$$U_*(\gamma)\mathrm{geo}(.,t) = \mathrm{geo}(.,t)U_*(\gamma) \quad \text{and} \quad U_*(\gamma)h_\pm(.,s) = h_\pm(.,s)U_*(\gamma) \tag{A31}$$

for all $t,s \in \mathbb{R}$ and $\gamma \in \mathrm{SL}(2,\mathbb{R})$. Moreover, $U_*(\gamma)$ are symplectomorphisms for all $\gamma \in \Gamma$.

Proof: Taking into account $U_*(\gamma).X^\flat = (U(\gamma)_* X)^\flat$ and Lemma A.11, we consider the ''raised'' flows on $TM_{\mu,\lambda}$:

$$\begin{aligned} TU(\gamma).G_\pm(\mathrm{Fl}_t^{G_\pm}(x^i)) &= TU(\gamma).G_\pm(U(\gamma^{-1})U(\gamma).\mathrm{Fl}_t^{G_\pm}(x^i)) \\ &= (U_*(\gamma)G_\pm)(U(\gamma).\mathrm{Fl}_t^{G_\pm}(x^i)) \\ &= (U(\gamma)_* G_\pm)(\mathrm{Fl}_t^{U(\gamma)^* G_\pm}(U(\gamma).x^i)) \end{aligned} \tag{A32}$$

which shows that U_* commutes with geo; the proof for $H_\pm$ follows the same lines. Finally $U_*(\gamma)$ are symplectomorphisms as lifts of diffeomorphisms of the configuration space $M_{\mu,\lambda}$. □

We are now in position to describe the dynamics on *smooth quotients manifolds* $\Gamma\backslash M_{\mu,\lambda}$ where $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$ is a discrete *co-compact* subgroup of the symmetry group $\mathrm{SL}(2,\mathbb{R})$ of the model $M_{\mu,\lambda}$.

Definition A.14: Let $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$ be a discrete co-compact subgroup of $\mathrm{SL}(2,\mathbb{R})$ acting on $M_{\mu,\lambda}$ via a representation U. There exists a compact subset $F_\Gamma \subset M_{\mu,\lambda}$ such that $\cup_{\gamma\in\Gamma} U(\gamma).F_\Gamma = M_{\mu,\lambda}$, and there are no two Γ-equivalent points in the interior of F_Γ. The boundary of F_Γ consists of geodesic arcs. F_Γ is called *fundamental domain* for the discrete co-compact group $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$. We recall (see the beginning of Appendix A 4) that the compact quotients considered in this paper are smooth manifolds.

Proposition A.15: $\mathrm{Fl}_t^{X_H}$, $\mathrm{Fl}_s^{E_\pm}: T^*M_{\mu,\lambda} \to T^*M_{\mu,\lambda}$, the geodesic and horocyclic flows on the cotangent bundle $T^*M_{\mu,\lambda}$, factor to the corresponding flows—again denoted by $\mathrm{Fl}_t^{X_H}$, respectively, $\mathrm{Fl}_s^{E_\pm}$—on the cotangent bundle $T^*(\Gamma\backslash M_{\mu,\lambda})$ of the quotient $\Gamma\backslash M_{\mu,\lambda}$. Moreover, the flows on the quotient cotangent bundle $T^*(\Gamma\backslash M_{\mu,\lambda})$ again satisfy the ACR. Furthermore, the quotient sphere bundles $T_r^*(\Gamma\backslash M_{\mu,\lambda})$ admit a probability measure ν that is invariant under both of these flows.

Proof: It follows from Theorem A.13 that the factorization is well-defined and satisfies the ACR. The projections of geodesics are indeed the geodesics on $\Gamma\backslash M_{\mu,\lambda}$ since $\pi: M_{\mu,\lambda} \to \Gamma\backslash M_{\mu,\lambda}$ is a Riemannian covering. Finally, the images of the horocycles are the corresponding horocycles on $\Gamma\backslash M_{\mu,\lambda}$, because they remain perpendicular to the geodesics. The probability measure ν is obtained by normalizing the measure Ω_r on the quotient determined by $\omega \wedge \omega = \Omega_r \wedge dH$; thereby $\omega \in \Omega^2(T^*(\Gamma\backslash M_{\mu,\lambda}))$ is the canonical symplectic form on the cotangent bundle of the quotient $\Gamma\backslash M_{\mu,\lambda}$, see Remark A.7 (and also Remark A.17 below). Proposition A.9 implies that this measure is invariant under both the geodesic and the horocyclic flows. □

This concludes our review of the main ingredients of the classical system, the quantization of which is the purpose of the body of this paper. While our review emphasized the structures we shall actually rely on, the singular importance of this system is that it provides a model for some of the strongest ergodic property one can expect from a classical Hamiltonian dynamical system. These properties are summarized in the following result.

Theorem A.16: The geodesic flow on the unit cotangent bundle $T_1^* M_{\mu,\lambda}$ is a C-flow. It thus satisfies in particular the following properties in the ergodic hierarchy: it is a K-flow, it is mixing with uniform Lebesgue spectrum, and it is ergodic.

This result is textbook staple; it is referred to in Ref. 4 as the Lobatchevski–Hadamard Theorem; see their Theorem 14.3, and its proof in their Appendix 21; see also Refs. 21 and 22. The metric transitivity itself had been known already for more than sixty years,[23] while the group theoretical approach which we shall exploit in our quantization has been pioneered, for the classical version of the model, in Ref. 24.

Remark A.17: In the body of the paper, we deal with the Anosov system on some compact quotient $\Gamma\backslash M_{\mu,\lambda}$ of Lobatchevski hyperbolic geometry $M_{\mu,\lambda}$. There, we use on the quotient only

the system of flows as given in (A.15), and from these we produce all the further objects we need, e.g., the corresponding vector fields. Neither do we intend to factor the symplectic form of $T^*M_{\mu,\lambda}$ to $T^*(\Gamma\backslash M_{\mu,\lambda})$; indeed, $T^*(\Gamma\backslash M_{\mu,\lambda})$ is a cotangent bundle on its own right, and therefore carries its own symplectic structure.

APPENDIX B: A NO-GO DYNAMICAL THEOREM

1. The Weyl group

To find the appropriate Weyl group for a physical system described by the cotangent bundle of some configuration space (i.e., systems with finitely many degrees of freedom) one usually looks for a group acting transitively, effectively and symplectically on the phase space under consideration, Ref. 5. In our case, the phase space is the cotangent bundle $T^*M_{\mu,\lambda}$ of Poincaré half-plane, and the group may be chosen, Ref. 6, to be $G^4_\lambda := H^2_\lambda \oslash \mathbb{R}^2$ with group multiplication law

$$(a^1,a^2,b_1,b_2)(c^1,c^2,d_1,d_2)=(a^1+c^1,a^2+\mathrm{e}^{2\lambda a^1}c^2,b_1+d_1,b_2+\mathrm{e}^{-2\lambda a^1}d_2) \tag{B1}$$

which leads to the Lie algebra $\mathfrak{g}^4_\lambda$:

$$[K_1,K_3]=[K_2,K_3]=[K_2,K_4]=[K_3,K_4]=0, \tag{B2}$$

$$[K_1,K_2]=2\lambda K_2,[K_1,K_4]=-2\lambda K_4. \tag{B2}$$

Moreover one has to make a transition to a central extension W^5_λ

$$0\to S^1\to W^5_\lambda\to G^4_\lambda\to 0 \tag{B3}$$

determined by the cocycle $c: G^4_\lambda\times G^4_\lambda\to S^1$:

$$c((a^1,a^2,b_1,b_2),(c^1,c^2,d_1,d_2))=\exp(i(a_1d_1+a_2d_2e^{-2\lambda a^1})). \tag{B4}$$

The Lie algebra $\mathfrak{w}^5_\lambda$ of W^5_λ is then given by

$$\begin{gathered}[K_2,K_3]=[K_3,K_4]=0,\\ [K_1,K_2]=2\lambda K_2,[K_1,K_4]=-2\lambda K_4,\\ [K_1,K_3]=[K_2,K_4]=K_5,\\ [K_i,K_5]=0 \quad \text{for } i=1,\ldots,4.\end{gathered} \tag{B5}$$

2. Action of G^4_λ on $T^*M_{\mu,\lambda}$.

Take $(x^1,x^2,p_1,p_2)\in T^*M_{\mu,\lambda}$, then

$$(a^1,a^2,b_1,b_2).(x^1,x^2,p_1,p_2)=\left(a^1+x^1,a^2+e^{2\lambda a^1}x^2,p_1+\frac{b_1}{\mu},p_2+e^{2\lambda(a^1+x^1)}b_2\right) \tag{B6}$$

One also defines an action of the central extension W^5_λ on $T^*M_{\mu,\lambda}$ by letting S^1 act trivially. This action determines corresponding fundamental vector fields $\zeta_{K_i}\in\mathfrak{X}(T^*M_{\mu,\lambda})$ which are given by:

$$\begin{gathered}\zeta_{K_1}(x^i,p_i)=\partial_{x^1}+2\lambda x^2\partial_{x^2},\\ \zeta_{K_2}(x^i,p_i)=\partial_{x^2},\\ \zeta_{K_3}(x^i,p_i)=\frac{1}{\mu}\partial_{p_1},\end{gathered} \tag{B7}$$

$$\zeta_{K_4}(x^i,p_i)=\mathrm{e}^{2\lambda x^1}\partial_{p_2},$$

$$\zeta_{K_5}(x^i,p_i)=0.$$

One checks that this is a symplectic action on the phase space $T^*M_{\mu,\lambda}$.

3. Behavior under the geodesic flow

B.1. The No-go theorem: The Weyl algebra is not closed under the action of the geodesic flow, unless the configuration space is flat.

Proof: The theorem asserts that, in case $\lambda\neq 0$, the vector space generated by the functions f_j, which in turn are determined up to a constant by $i_{\zeta_{K_j}}\omega=df_j$, is not closed under time evolution. This is equivalent to the statement that the vector space generated by the fundamental vector fields $\{\zeta_{K_j}|1\leqslant j\leqslant 4\}$ is not closed under the geodesic flow, i.e., that $[X_H,\zeta_{K_j}]$ is not a linear combination of the ζ_{K_j}'s with *constant* coefficients. Upon recalling that the action of the geodesic flow is generated by the vector field

$$X_H(x^i,p_i)=\frac{p_1}{\mu}\partial_{x^1}+p_2e^{2\lambda x^1}\partial_{x^2}-\frac{2\lambda}{\mu}p_2^2\partial_{p_1}+\frac{2\lambda}{\mu}p_1p_2\partial_{p_2} \tag{B8}$$

we find indeed:

$$[X_H,\zeta_{K_j}](x^i,p_i)=\begin{pmatrix} 0 & 2\lambda p_2e^{2\lambda x^1} & 0 & 0\\ 0 & 0 & 0 & 0\\ -\dfrac{1}{\mu^2} & \dfrac{2\lambda}{\mu^2}x^2 & 0 & -\dfrac{2\lambda}{\mu^2}p_2e^{-2\lambda x^1}\\ 0 & -e^{4\lambda x^1} & \lambda e^{2\lambda x^1}p_2 & 0 \end{pmatrix}\cdot(\zeta_{K_j}(x^i,p_i)). \tag{B9}$$

For $(\mu,\lambda)=(1,0)$ this matrix of coefficients becomes independent from (x^i,p_i). This is exactly the reduction to the situation as one knows it from ''flat'' quantum mechanics, Ref. 5. □

Note that the above no-go theorem is lurking behind the negative result of Sec. V in Ref. 1. It is thus to circumvent this theorem that we devised the quantization scheme presented in Sec. II of the present paper.

[1] G. G. Emch, H. Narnhofer, W. Thirring, and G. L. Sewell, ''Anosov Actions on Non-Commutative Algebras,'' J. Math. Phys. **35**, 5582–5599 (1994); G. G. Emch, ''Chaotic dynamics in non-commutative geometry,'' in *Quantizations, Deformations and Coherent States* (Polish Scientific, Warsaw, 1997), pp. 33–42. G. G. Emch, ''Beyond Irreducibility and Back,'' Rep. Math. Phys. **40**, 187–193 (1997).

[2] W. Thirring, ''What are the quantum mechanical Lyapunov exponents?'' in *Proceedings of the 34. Internationale Universitätswochen für Kern– und Teilchenphysik, Schladming, Austria*, 4–11 March 1995 (Springer, Berlin, 1996), pp. 223–237.

[3] J. Hadamard, ''Les surfaces à courbures opposées et leurs lignes géodésiques,'' J. Math. Pures et Appliquées **4**, 27–74 (1898).

[4] V. I. Arnol'd and A. Avez, *Ergodic Problems of Classical Mechanics* (Benjamin, New York, 1968).

[5] C. J. Isham, ''Topological and Global Aspects of Quantum Theory,'' *in Relativity, Groups, Topology* II, Les Houches, Session XL, edited by B. S. DeWitt and R. Stora (Elsevier, New York, 1984), pp. 1059–1290.

[6] J. Bertrand, G. G. Emch, and G. Rideau, ''The Cohomology of the Classical and Quantum Weyl CCR in Curved Spaces,'' Lett. Math. Phys. **34**, 149–158 (1995).

[7] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, ''The Smallest C*-algebra for Canonical Commutation Relations,'' Commun. Math. Phys. **32**, 231–243 (1973).

[8] G. G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley, New York, 1972).

[9] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics*, 2nd ed. (Springer, New York, 1996), Vol. II.

[10] H. Narnhofer and W. Thirring, ''KMS States for the Weyl Algebra,'' Lett. Math. Phys. **27**, 133–142 (1993).

[11] G. G. Emch, ''Quantum and classical mechanics on homogeneous Riemannian manifolds,'' J. Math. Phys. **23**, 1785–1791 (1982).

[12] H. Narnhofer, I. Peter, and W. Thirring, ''How hot is the de Sitter Space?,'' Int. J. Mod. Phys. B **10**, 1507–1520 (1996).

[13] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin, New York, 1978).

[14] J. A. Wolf, *Spaces of Constant Negative Curvature* (McGraw–Hill, New York, 1967).
[15] D. A. Heijhal, *The Selberg trace formula for PSL(2,R)*, Lect. Notes Math. Vol. 548 (Springer, New York, 1976).
[16] H. Poincaré, ''Théorie des groupes Fuchsiens,'' Acta Math. **1**, 1–62 (1882); English translation by J. Stillwell, *H. Poincaré, Papers on Fuchsian functions* (Springer, New York, 1985).
[17] I. M. Gel'fand, M. I. Graev, and I. I. Pyatetstkiĭ-Sapiro, *Representation Theory and Automorphic Functions* (Saunders, Philadelphia, 1969).
[18] A. F. Beardon, *The Geometry of Discrete Groups* (Springer, New York, 1983).
[19] B. Iversen, *Hyperbolic Geometry* (Cambridge University Press, London, 1992).
[20] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer, New York, 1990).
[21] E. Hopf, *Statistik der Geodätischen Linien in Manigfaltigkeiten Negativer Krümmung* (Ber. Verh. Sächs. Akad. Wiss., Leipzig, 1939), 91, pp. 261–304; see also: ''Ergodic theory and the geodesic flow on surfaces of constant negative curvature,'' Bull. Am. Math. Soc. **77**, 863–877 (1971).
[22] D. V. Anosov, ''Geodesic flows on compact Riemannian manifolds of negative curvature,'' Proceedings of the Steklov Institute, Moscow, 1967, p. 90.
[23] G. Hedlund, ''Metric transistivity of the geodesics on closed surfaces of constant negative curvature,'' Ann. Math. **35**, 787 (1935); ''The dynamics of geodesic flows,'' Bull. Am. Math. Soc. **45**, 241–246 (1939).
[24] I. M. Gel'fand and S V. Fomim, ''Geodesic flows on manifolds of constant negative curvature,'' Trans. Am. Math. Soc. **1**, 49–67 (1955).

Thermodynamic fermion loop in a constant magnetic field

Haru-Tada Sato[a)]
The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, DK-2100 Copenhagen, Denmark

(Received 25 August 1997; accepted for publication 30 April 1998)

The one-loop effective potential of a thermodynamic fermion loop in a constant magnetic field is studied. As expected, it can be interpreted literally as a discretized sum of $(D-2)$-dimensional energy density above the Dirac sea. Large/small mass expansions of the potential are also examined. © *1998 American Institute of Physics.* [S0022-2488(98)04109-7]

I. INTRODUCTION

Recently, effects of the magnetic field on the chiral symmetry breaking have been intensively studied.[1–3] In the paper,[4] we obtained a simple analytic form of the thermodynamic effective potential for a fermion loop coupled to constant magnetic field in D dimensions. On the contrary, the corresponding (D-dimensional) effective potential for the electric field case[5–9] cannot easily be organized into a concise form like the magnetic case due to a number of singularities; in other words, possessing an imaginary part.[10] Nevertheless we derived a few analytic expressions for the real parts of the effective potential and its gap equation in the simplest (zero temperature) case through the electromagnetic dual transformation; i.e., rotating a magnetic parameter $\xi=eB$ into a pure imaginary number $\xi=i\rho=ieE$.[4] This rotation procedure cannot always be applied to an intermediate process of calculation, depending on whether or not the potential can be regarded as a complex analytic function of ξ. It is expected to be valid, when the potential is written in terms of a special analytic function. In this way, a compact form of D-dimensional thermodynamic effective potential for the electric case might be obtained.

However, in the nonzero temperature case, there is no useful analytic expression suitable for that purpose; in other words, the electromagnetic dual symmetry becomes unclear. Thus, it is certainly important to know some behaviors of the potential in a constant magnetic field for finding or examining a behavior of thermodynamic potential in a constant electric field. Associated with this strategy, here we present some more analytic representations (in terms of special functions) of the thermodynamic effective potential for a Fermion loop coupled to a constant magnetic field.

The paper is organized as follows. In Sec. II, we discuss a physical interpretation of the effective potential. As is well known, effective potentials possess the meaning of energy density, and can actually be derived from the traces over energy spectra.[11,12] However, the derivation of our potential is *not* based on such a trace computation, and we would like to clarify that a mathematical expression of the (finite temperature) potential indicates the physical meaning exactly. There is also a dimensional reduction from D to $D-2$ in the presence of a magnetic field, and we hence expect that the structure should be $D-2$-dimensional energy density averaged by the Fermi–Dirac state density.

We then present, in Secs. III and IV, large/small mass expansions of the thermodynamic effective potential with keeping β *finite* (One should note that this is certainly a different argument from Ref. 4, which mainly concentrated on an infinite β limit). In order to see a validity for the present results, we shall reproduce a zero temperature potential (with $\mu\neq 0$) from those expansions. Namely, applying a low-temperature expansion to each mass expansion, we shall see that the leading terms of the low-temperature expansions coincide with the zero-temperature potential obtained in Ref. 4, where none of these (mass and temperature) expansions were used and its T

a)Electronic mail: h.sato@thphys.uni-heidelberg.de
Present address: Institut für Theoretische Physik, Univ. Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany.

 © 1998 American Institute of Physics

$=0$ situation was rather direct. In this sense, in this paper another proof of the zero-temperature potential of finite density is provided.

II. STRUCTURE OF THE EFFECTIVE POTENTIAL

The basic parts of the model Lagrangian are

$$\mathscr{L}=\bar{\psi}i\gamma^{\mu}(\partial_{\mu}-ieA_{\mu})\psi-m\bar{\psi}\psi, \tag{1}$$

and the D-dimensional thermodynamic effective potential is derived at one loop,[4]

$$V_{\beta,\mu}(m;\xi)=-\frac{\mathrm{tr}[1]}{2\beta}\int_0^{\infty}\frac{ds}{s}\frac{\Theta_2\left(s\,\frac{2\mu}{\beta},is\,\frac{4\pi}{\beta^2}\right)}{(4\pi s)^{(D-1)/2}}\,s\xi\,\coth(s\xi)e^{-s(m^2-\mu^2)}, \tag{2}$$

where Θ_2 is the elliptic theta function of the second kind and tr[1] stands for the trace of the gamma matrix unit. This can also be rewritten in the form

$$V_{\beta,\mu}(m;\xi)=V_R(m;\xi)+\tilde{V}_{\beta,\mu}(m;\xi), \tag{3}$$

where the first term V_R is a renormalized effective potential for the original Lagrangian (1), or one may add a four-Fermi interaction to the Lagrangian (1) if necessary. However, such a Fermion interaction does not contribute to the second term $\tilde{V}_{\beta,\mu}$, which is the pure thermodynamic part of the effective potential (2) and is renormalization free. Since our present interest is the thermodynamic part, which possesses a model-independent structure, we focus on the latter potential, and its explicit form is given by[4]

$$\tilde{V}_{\beta,\mu}(m;\xi)=\frac{\mathrm{tr}[1]\xi}{(4\pi)^{D/2}}\left[\frac{1}{2}\,\mathscr{O}_{\beta}(m)+\sum_{l=1}^{\infty}\mathscr{O}_{\beta}(\sqrt{m^2+2l\xi})\right], \tag{4}$$

where

$$\mathscr{O}_{\beta}(\sigma)=4\sum_{n=1}^{\infty}(-1)^n\cosh(n\beta\mu)\left(\frac{\beta n}{2\sigma}\right)^{1-D/2}K_{D/2-1}(n\beta\sigma). \tag{5}$$

Note that Eq. (4) is composed of the sum over a discrete energy spectrum $\epsilon_n=\pm\sqrt{m^2+2n\xi}$ due to the presence of a magnetic field. Since ξ possesses the mass dimension 2, the fundamental object $\mathscr{O}_{\beta}(\sigma)$ in the pure thermal effective potential possesses mass dimension $D-2$. Hence Eq. (5) can be interpreted as an effective potential in $D-2$ dimensions. We can regard this fact as an example of the dimensional reduction from D to $D-2$. Furthermore, here we suggest one more interesting feature: From this specific potential form, we can read clearly the common fact that the effective potential is understood as an energy density. Although this statement is very natural, however, the above potential forms, Eqs. (2) and (5), do not explain this structure explicitly.

Let us reveal here the desired structure, introducing another expression for Eq. (5),

$$\mathscr{O}_{\beta}(\sigma)=\frac{-2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\left[\int_0^{\infty}\frac{(t^2+2\sigma t)^{(D-3)/2}}{1+e^{\beta(t+\sigma+\mu)}}\,dt+(\mu\to-\mu)\right], \tag{6}$$

where we have used an integral representation for K_{ν} and then performed the sum on n. Changing the integration variable with $\epsilon=t+\sigma$, we obtain

$$\mathscr{O}_{\beta}(\sigma)=\frac{-2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\left[\int_{\sigma}^{\infty}\frac{(\epsilon^2-\sigma^2)^{(D-3)/2}}{1+e^{\beta(\epsilon-\mu)}}\,d\epsilon+(\mu\to-\mu)\right]. \tag{7}$$

The RHS of this equation can be interpreted as follows. The denominator of the integrand is related to the state density $N(\epsilon)$ averaged by the Fermi–Dirac statistics,

$$N(\epsilon)=\frac{1}{1+e^{\beta(\epsilon-\mu)}}, \tag{8}$$

and the numerator with integration measure carrying the mass dimension $D-2$ (the same as $\mathcal{O}_\beta$),

$$f(\epsilon)d\epsilon=(\epsilon^2-\sigma^2)^{(D-3)/2}\,d\epsilon, \tag{9}$$

is understood as an energy density for the infinitesimal interval $d\epsilon$. The integration means nothing but a total sum of this energy density over the whole state density $N(\epsilon)$, and the lower bound of the integration is restricted by the Dirac sea of the fermion mass σ. In addition, in the case of zero temperature, $N(\epsilon)$ takes 1 or 0, and the Fermi energy is given at the point $\epsilon=\mu$. Thus we have

$$\mathcal{O}_\infty(\sigma)\equiv\lim_{\beta\to\infty}\mathcal{O}_\beta(\sigma)=\int_\sigma^\mu f(\epsilon)d\epsilon, \tag{10}$$

and this is exactly the result given in Ref. 4,

$$\mathcal{O}_\infty(\sigma)=-\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D+1}{2}\right)}(2\sigma)^{(D-3)/2}(\mu-\sigma)^{(D-1)/2}F\left(\frac{3-D}{2},\frac{D-1}{2};\frac{D+1}{2};\frac{\sigma-\mu}{2\sigma}\right)\theta(\mu-\sigma). \tag{11}$$

Next, we consider the further details of $\mathcal{O}_\beta(\sigma)$ (for finite β), which we did not get through in the previous paper.[4] We present below its large/small mass expansions at finite β. Obviously, in Eq. (6), there are two ways of performing a binomial expansion for the quantity $f(\epsilon)$; i.e., for $(t^2+2\sigma t)^\nu$, where $\nu=(D-3)/2$.

III. THE LARGE MASS EXPANSION

First, let us consider the large σ case (σ^{-1} expansion). After changing $\beta t\to t$ and expanding around $\sigma=\infty$, we can perform the integration seen in Eq. (6) for each expansion mode,

$$\mathcal{O}_\beta(\sigma)=\frac{-2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\beta^{2-D}\sum_{k=0}^{\infty}\binom{\frac{D-3}{2}}{k}(2\beta\sigma)^{(D-3)/2-k}\Gamma\left(\frac{D-1}{2}+k\right)$$
$$\times\exp[-\beta(\sigma+\mu)]\Phi\left(\frac{D-1}{2}+k;\frac{1}{2}+\frac{i\beta}{2\pi}(\sigma+\mu),1\right)+(\mu\to-\mu), \tag{12}$$

where $\Phi(z;s,a)$ is the Lerch transcendental function, defined by

$$\Phi(z;s,a)=\sum_{n=0}^{\infty}\frac{e^{2\pi ins}}{(n+a)^z}=\frac{1}{\Gamma(z)}\int_0^\infty\frac{e^{-at}t^{z-1}}{1-e^{-t+2\pi is}}\,dt\quad(\mathrm{Re}\,z>1,\ 0<a\leqslant 1). \tag{13}$$

Since the Lerch function is transformed into the generalized zeta function by the relation

$$e^{2\pi is}\Phi(1-z;s,1)=(2\pi)^{-z}\Gamma(z)[e^{(\pi i/2)z}\zeta(z,s)+e^{-(\pi i/2)z}\zeta(z,1-s)], \tag{14}$$

we can transform Eq. (12) into another representation,

$$\mathcal{O}_\beta(\sigma)=\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\sum_{k=0}^{\infty}\binom{\frac{D-3}{2}}{k}(2\sigma)^{[(D-3)/2]-k}(2\pi)^{[(D-3)/2]+k}\Gamma\left(\frac{D-1}{2}+k\right)\Gamma\left(\frac{3-D}{2}-k\right)$$
$$\times 2\beta^{[(1-D)/2]-k}\,\mathrm{Re}\left[e^{(\pi i/2)[(3-D)/2-k]}\zeta\left(\frac{3-D}{2}-k,\ \frac{1}{2}+\frac{i\beta}{2\pi}(\sigma+\mu)\right)+(\mu\to-\mu)\right]. \tag{15}$$

When $D=3$, the above summations in Eqs. (12) and (15) consist of only the $k=0$ term, and the latter expression (15) contains superficial divergence (in the gamma function). In this sense, the former (12) may rather be a fundamental expression. Anyway, both expressions coincide with the following in $D=3$ (see also the first literature of Ref. 1),

$$\mathcal{O}_\beta(\sigma)=-2\frac{\sqrt{\pi}}{\beta}\ln[(1+e^{-\beta(\sigma+\mu)})(1+e^{-\beta(\sigma-\mu)})]. \tag{16}$$

Nevertheless, the latter representation (15) is convenient to consider the low-temperature expansion (β^{-1} expansion). Let us find that the leading term of the expansion reproduces the previous result (11). The β^{-1} expansion can be performed through the asymptotic formula for the zeta function,

$$\zeta(z,a)\sim\frac{1}{z-1}\,a^{1-z},\quad |a|\to\infty. \tag{17}$$

For example, the highest power of β is calculated as follows:

$$\mathrm{Re}\left[e^{(\pi i/2)[(3-D)/2-k]}\zeta\left(\frac{3-D}{2}-k,\ \frac{1}{2}+\frac{i\beta}{2\pi}(\sigma\pm\mu)\right)\right]$$
$$=\frac{1}{\frac{1-D}{2}-k}\left(\frac{\beta}{2\pi}|\sigma\pm\mu|\right)^{[(D-1)/2]+k}\mathrm{Re}[e^{(\pi i/2)[(3-D)/2-k]}e^{(\pi i/2)[k+(D-1)/2]\mathrm{sign}(\sigma\pm\mu)}]+\cdots$$
$$=\frac{(\beta/2\pi)^{[(D-1)/2]+k}}{\frac{1-D}{2}-k}\times\begin{cases}0, & \text{for } +,\\ \cos\left[\left(1-k-\frac{D}{2}\right)\pi\right]|\mu-\sigma|^{(D-1)/2+k}\theta(\mu-\sigma), & \text{for } -,\end{cases}$$
$$+\cdots. \tag{18}$$

Substituting this result onto the RHS of Eq. (15), we derive

$$\mathcal{O}_\beta(\sigma)=\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\sum_{k=0}^{\infty}\binom{\frac{D-3}{2}}{k}\frac{(2\sigma)^{[(D-3)/2]-k}}{\frac{1-D}{2}-k}(\mu-\sigma)^{[(D-1)/2]+k}\theta(\mu-\sigma)+\mathcal{O}(\beta^{-1}), \tag{19}$$

where we have applied the following relation to eliminate the cosine factor seen in Eq. (18):

$$\Gamma\left(\frac{1}{2}+z\right)\Gamma\left(\frac{1}{2}-z\right)=\frac{\pi}{\cos\,\pi z},\quad \text{for } z=1-k-\frac{D}{2}. \tag{20}$$

Furthermore, using

$$\frac{1}{\frac{1-D}{2}-k}=\frac{\Gamma\left(\frac{1-D}{2}-k\right)}{\Gamma\left(\frac{3-D}{2}-k\right)}=\frac{(-1)^k\pi}{\cos\left(D\,\frac{\pi}{2}\right)\Gamma\left(\frac{3-D}{2}-k\right)\Gamma\left(\frac{1+D}{2}+k\right)}, \tag{21}$$

and

$$F(\alpha,\beta;\gamma;z)=\sum_{n=0}^{\infty}\frac{z^n}{n!}\,\frac{\Gamma(1-\alpha)\Gamma(1-\beta)\Gamma(\gamma)}{\Gamma(1-\alpha-n)\Gamma(1-\beta-n)\Gamma(\gamma+n)}, \tag{22}$$

we find

$$\mathcal{O}_\beta(\sigma)=\frac{2\pi^{3/2}(2\sigma)^{(D-3)/2}(\mu-\sigma)^{(D-1)/2}}{\cos\left(D\,\frac{\pi}{2}\right)\Gamma\left(\frac{3-D}{2}\right)\Gamma\left(\frac{D-1}{2}\right)\Gamma\left(\frac{D+1}{2}\right)}$$
$$\times F\left(\frac{3-D}{2},\,\frac{D-1}{2};\,\frac{D+1}{2};\,\frac{\sigma-\mu}{2\sigma}\right)\theta(\mu-\sigma)+\mathcal{O}(\beta^{-1}). \tag{23}$$

The leading term is now identified with the quantity (11) through the formula (20) for $z=(D-2)/2$.

IV. THE SMALL MASS EXPANSION

Second, let us compute the small mass expansion. The calculation is parallel to the above σ^{-1} expansion. Expanding $(t^2+2\sigma t)^\nu$ for small σ and integrating each mode, we obtain the following two representations:

$$\mathcal{O}_\beta(\sigma)=\frac{-2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\,\beta^{2-D}\sum_{k=0}^{\infty}\binom{\frac{D-3}{2}}{k}(2\beta\sigma)^k\Gamma(D-2-k)$$
$$\times\exp[-\beta(\sigma+\mu)]\Phi\left(D-2-k;\,\frac{1}{2}+\frac{i\beta}{2\pi}\,(\sigma+\mu),1\right)+(\mu\to-\mu), \tag{24}$$

and

$$\mathcal{O}_\beta(\sigma)=\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\sum_{k=0}^{\infty}\binom{\frac{D-3}{2}}{k}(2\sigma)^k(2\pi)^{D-3-k}\,\frac{\Gamma\left(\frac{D-1-k}{2}\right)\Gamma\left(\frac{3-D+k}{2}\right)}{\cos\left[\frac{\pi}{2}\,(3-D+k)\right]}$$
$$\times\beta^{2-D+k}\,\mathrm{Re}\left[e^{(\pi i/2)(3-D+k)}\zeta\left(3-D+k,\,\frac{1}{2}+\frac{i\beta}{2\pi}\,(\sigma+\mu)\right)+(\mu\to-\mu)\right], \tag{25}$$

where the Lerch transformation (14) and the following formula are applied:

$$\Gamma(D-2-k)\Gamma(3-D+k)=\frac{\Gamma\left(\frac{D-1-k}{2}\right)\Gamma\left(\frac{3-D+k}{2}\right)}{\cos\left[\frac{\pi}{2}\,(3-D+k)\right]}. \tag{26}$$

Again, let us perform the β^{-1} expansion in Eq. (25) and sum up the leading contribution similarly to the former case. The resulting equation is

$$\mathcal{O}_\beta(\sigma)=\frac{-2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\frac{(\mu-\sigma)^{D-2}}{D-2}F\left(\frac{D-1}{2},D-1;3-D;\frac{-2\sigma}{\mu-\sigma}\right)\theta(\mu-\sigma)+\mathcal{O}(\beta^{-1}). \quad (27)$$

The value of $\mathcal{O}_\beta$ at the origin $\sigma=0$ for finite β can be obtained by putting $\sigma=0$ into (25). The $k=0$ term is only nonzero, and we get $\mathcal{O}_\beta(0)$;

$$\mathcal{O}_\beta(0)=\frac{2}{\sqrt{\pi}}\left(\frac{2\pi}{\beta}\right)^{D-2}\Gamma\left(\frac{3-D}{2}\right)\mathrm{Re}\ \zeta\left(3-D,\frac{1}{2}+i\frac{\beta\mu}{2\pi}\right). \quad (28)$$

For infinite β, we just put $\sigma=0$ in (27), and only the leading term survives;

$$\mathcal{O}_\infty(0)=-\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D-1}{2}\right)}\frac{\mu^{D-2}}{D-2}. \quad (29)$$

V. CONCLUDING REMARKS

In closing this paper, we make two interesting remarks on Eq. (29). The first one is that we can see that Eq. (29) is exactly a dimensionally reduced potential by $D\to D-2$ up to a normalization constant factor. It can be seen by comparison with the potential of the continuum theory (i.e., $\xi=0$ case):[13]

$$\lim_{\beta\to\infty}\tilde{V}_{\beta,\mu}(0)=-2\frac{\mathrm{tr}[1]}{(4\pi)^{D/2}}\frac{2\sqrt{\pi}}{\Gamma\left(\frac{D+1}{2}\right)}\frac{\mu^D}{D}. \quad (30)$$

Second, Eq. (29) is divergent if we take the limit $D\to 2$. This *is* an interesting point, although a magnetic analysis in $D=2$ *per se* is meaningless. The reason is as follows: As pointed out in the Introduction as well as in Ref. 4, we can obtain an effective potential for the electric case at $T=0$ from the present case through rotating $\xi\to i\rho$. This rotation makes the quantity (29) purely imaginary, and it certainly means that the imaginary part of the (electric case) potential is divergent in $D=2$.[6]

[1] V. P. Gusynin, V. A. Miransky, and I. A. Shovkovy, Phys. Rev. D **52**, 4718 (1995); Nucl. Phys. B **462**, 249 (1996); Phys. Rev. Lett. **73**, 3499 (1994); V. P. Gusynin and I. A. Shovkovy, Phys. Rev. D **56**, 5251 (1997).
[2] C. N. Leung, Y. J. Ng, and A. W. Ackley, Phys. Rev. D **54**, 4181 (1996); D. S. Lee, C. N. Leung, and Y. J. Ng, *ibid.* **55**, 6504 (1997).
[3] D. M. Gitman, S. D. Odintsov, and Yu. I. Shilnov, Phys. Rev. D **54**, 2968 (1996); B. Geyer, L. N. Granda, and S. D. Odintsov, Mod. Phys. Lett. A **11**, 2053 (1996).
[4] S. Kanemura, H-T. Sato, and H. Tochimura, Nucl. Phys. B **517**, 567 (1998).
[5] K. G. Klimenko, Theor. Math. Phys. **89**, 1287 (1992); Z. Phys. C **54**, 323 (1992).
[6] S. Kawati, A. Konisi, and H. Miyata, Phys. Rev. D **28**, 1537 (1983).
[7] S. P. Klevansky and R. H. Lemmer, Phys. Rev. D **39**, 3478 (1989).
[8] S. P. Klevansky, Rev. Mod. Phys. **64**, 649 (1992).
[9] A. Yu. Babansky, E. V. Gorbar, and G. V. Shchepanyuk, Phys. Lett. B **419**, 272 (1998).
[10] M. Stone, Phys. Rev. D **14**, 3568 (1976); J. Schwinger, Phys. Rev. **82**, 664 (1951).
[11] D. Cangemi and G. Dunne, Ann. Phys. (N.Y.) **249**, 582 (1996).
[12] J. O. Andersen and T. Haugset, cond-mat/9410084.
[13] T. Inagaki, T. Kouno, and T. Muta, Int. J. Mod. Phys. A **10**, 2241 (1995).

Amplitude equations for electrostatic waves: Multiple species

John David Crawford[a] and Anandhan Jayaraman[b]
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 16 June 1997; accepted for publication 20 April 1998)

The amplitude equation for an unstable electrostatic wave is analyzed using an expansion in the mode amplitude $A(t)$. In the limit of weak instability, i.e., $\gamma \to 0^+$ where γ is the linear growth rate, the nonlinear coefficients are singular and their singularities predict the dependence of $A(t)$ on γ. Generically the scaling $|A(t)| = \gamma^{5/2} r(\gamma t)$ as $\gamma \to 0^+$ is required to cancel the coefficient singularities to all orders. This result predicts the electric field scaling $|E_k| \sim \gamma^{5/2}$ will hold universally for these instabilities (including beam-plasma and two-stream configurations) throughout the dynamical evolution and in the time-asymptotic state. In exceptional cases, such as infinitely massive ions, the coefficients are less singular and the more familiar trapping scaling $|E_k| \sim \gamma^2$ is recovered. © *1998 American Institute of Physics.* [S0022-2488(98)03009-6]

I. INTRODUCTION

A. Overview and background

We recently presented a detailed analysis of the amplitude equation for an unstable electrostatic mode in an unmagnetized Vlasov plasma [henceforth (I)].[1] In that work, the amplitude equation was studied using an expansion in the wave amplitude, and only the electron motion was considered with the ions treated as a fixed neutralizing background. In this paper we study the effect of including mobile ions and investigate the mode amplitude equation for an electrostatic wave in a multispecies unmagnetized plasma. As in (I), we focus on the regime of weak instability and carefully analyze the structure of the amplitude equation in the limit $\gamma \to 0^+$ where γ is the linear growth rate for the unstable wave.

Incorporating the effects of finite ion inertia is a significant step for several reasons. With finite mass, the ion motion occurs on a finite time scale $T_i < \infty$ and in the limit of very weak growth rates this will be a fast time scale, i.e., $T_i \ll \gamma^{-1}$ as $\gamma \to 0^+$. Over the duration of the instability, the ions can readily respond to the electric field of the wave especially resonant ions near the phase velocity. This ion response is always present even for high frequency modes such as plasma waves since at the phase velocity the wave frequency is Doppler-shifted to zero. In addition, there are modes such as ion-acoustic waves that depend on ion density oscillations and are suppressed when the ions are treated as fixed. Allowing for mobile ions is a prerequisite, if we seek the amplitude equation for an unstable ion-acoustic mode.

This paper can be read independently, but the reader is referred to (I) for a discussion of the history of the problem and additional introductory material. An essential difference between our approach and previous work lies in the choice of unperturbed state. Earlier theories assumed an equilibrium with a neutrally stable mode and obtained ill-defined expansion coefficients.[2–6] This can be avoided by taking the weakly unstable equilibrium as the unperturbed state; a choice that naturally leads one to work with the unstable manifold. A brief summary of our results on the multispecies case has appeared elsewhere.[7]

In the limit $\gamma \to 0^+$, the amplitude equation defines a kind of singular perturbation problem whose detailed features reveal asymptotic scaling behavior of the nonlinear wave. This is a key idea behind our approach and it can be formulated more precisely as follows. The mode eigen-

[a]Electronic mail: jdc@minerva.phyast.pitt.edu
[b]Electronic mail: anand@merlin.phyast.pitt.edu

 © 1998 American Institute of Physics

value $\lambda=\gamma-i\omega$ can be complex (beam-plasma) or real (two-stream), but in either case the equations for the mode amplitude $A(t)=\rho(t)e^{-i\theta(t)}$ have the form

$$\dot{\rho}=\gamma\rho+a_1\rho^3+a_2\rho^5+\mathcal{O}(\rho^7), \tag{1}$$

$$\dot{\theta}=\omega+a_1'\rho^2+a_2'\rho^4+\mathcal{O}(\rho^6). \tag{2}$$

Since the Vlasov equilibrium is assumed to be spatially homogeneous, the evolution of $\rho(t)$ decouples from the phase $\theta(t)$. Suppose for simplicity that the coefficients a_j are constant (ignoring their dependence on γ), then we may view (1) as a singular perturbation problem with the linear term $\gamma\rho$ representing the perturbation. If $\gamma>0$, then there is always a neighborhood of the equilibrium $\rho=0$ where the perturbation dominates the unperturbed system and can completely change the dynamics.

As in other such singular problems, a possible strategy is to seek a (singular) change of variables which transforms (1) into a regular perturbation problem. Thus it is natural to define a new amplitude $r(\tau)$ by

$$\rho(t)=\gamma^{\beta}r(\gamma^{\delta}t) \tag{3}$$

and rewrite (1) in terms of $r(\tau)$

$$\frac{dr}{d\tau}=\gamma^{1-\delta}r(\tau)+\gamma^{2\beta-\delta}a_1(\gamma)r^3+\gamma^{4\beta-\delta}a_2(\gamma)r^5+\cdots. \tag{4}$$

If possible, the choice of β and δ should be made so that (i) each term in (4) is well behaved as $\gamma\to 0^+$ and (ii) the effect of $\gamma>0$ is a regular perturbation of the system at $\gamma=0$. The latter condition requires $\delta=1$; otherwise the linear term will continue to be a singular perturbation. The exponent β then needs to be chosen (if possible) to achieve a balance between the nonlinear terms and the linear term. In the simplest situation, when the coefficients $a_1,a_2,\ldots$ have well-defined *finite* limits as $\gamma\to 0$, then one has the exponent $\beta=1/2$ and this yields

$$\frac{dr}{d\tau}=r(\tau)+a_1(\gamma)r^3+\gamma a_2(\gamma)r^5+\cdots. \tag{5}$$

The unperturbed system $\dot{r}=r+a_1(0)r^3$ now includes the linear term and near $r=0$ the effect of small γ is unimportant. The change of variables (3) is singular at $\gamma=0$ as expected, but nevertheless we have determined the overall scaling $\beta=1/2$ for nonlinear solutions near the equilibrium in the regime of weak instability. In addition, since terms in (4) at fifth order or higher are at least of order $\mathcal{O}(\gamma)$, the unperturbed problem truncates to a simple balance between the linear instability and the dominant nonlinear term.

For an unstable electrostatic wave, the situation is quite different because the nonlinear coefficients a_j blow up as $\gamma\to 0$ and the exponent $\beta=1/2$ does not yield a well-behaved equation for $r(\tau)$. We attack the problem of finding a satisfactory value for β in two steps. First, we determine what choice for β will control the singularity of the cubic coefficient a_1 in (1), and then we investigate whether this choice will also remove the singularities of the higher order nonlinear terms. Typically, the cubic coefficient diverges asymptotically like γ^{-4} in which case we must choose $\beta\geqslant 5/2$ in order to obtain a finite cubic term in the rescaled amplitude equation (4).

Our analysis shows that setting $\beta=5/2$ yields a theory that is finite to all orders as $\gamma\to 0$ at $\gamma=0$ retains (rescaled) nonlinear terms that balance the linear instability. Choosing a larger exponent $\beta=5/2+\epsilon$ would also yield a finite theory to all orders, but now the cubic term is of order $\mathcal{O}(\gamma^{2\epsilon})$ and all the higher order terms ($j>1$) are of order $\mathcal{O}(\gamma^{2\epsilon j})$; there is no balance between the linear and nonlinear terms at $\gamma=0$. For $\epsilon>0$, the larger exponent suppresses nonlinear effects too strongly. In this sense, the scaling

$$\rho(t)=\gamma^{5/2}r(\gamma t) \tag{6}$$

is uniquely determined. Of the two steps, the second is by far the most difficult; it is relatively straightforward to compute a_1 and determine its divergence (Sec. III). Verifying that a given choice of β works to all orders requires a detailed study of the recursion relations of the theory and the final theorem is proved by induction (Sec. IV).

In special cases, the cubic coefficient is less singular, diverging like γ^{-3} instead of γ^{-4}. The limit of infinitely massive ions is the best known example of this situation, and in (I) we found that $\beta=2$ was the scaling needed to make the nonlinear amplitude equation finite to all orders. In this paper, we identify two additional examples where the cubic coefficient has a γ^{-3} singularity: if the effects of the resonant ions are negligible (in a sense we make precise in Sec. III D) or if the electron and ion masses are equal, i.e., for an electron-positron plasma. In the former case we prove that the exponent $\beta=2$ is correct, but the precise value of β for the electron-positron system is not unambiguously determined due to special complications in the singularity structure of the amplitude expansion.

B. Notation

Our model is a one-dimensional, multispecies Vlasov plasma defined by

$$\frac{\partial F^{(s)}}{\partial t}+v\frac{\partial F^{(s)}}{\partial x}+\kappa^{(s)}E\,\frac{\partial F^{(s)}}{\partial v}=0, \tag{7}$$

$$\frac{\partial E}{\partial x}=\sum_s\int_{-\infty}^{\infty}dv F^{(s)}(x,v,t). \tag{8}$$

Here x, t and v are measured in units of u/ω_e, ω_e^{-1} and u, respectively, where u is a chosen velocity scale and $\omega_e^2=4\pi e^2 n_e/m_e$. The plasma length is L with periodic boundary conditions and we adopt the normalization

$$\int_{-L/2}^{L/2}dx\int_{-\infty}^{\infty}dv F^{(s)}(x,v,t)=\left(\frac{z_s n_s}{n_e}\right)L, \tag{9}$$

where $q_s=ez_s$ is the charge of species s and $\kappa^{(s)}\equiv q_s m_e/em_s$. Note that $\kappa^{(e)}=-1$ for electrons and that the normalization (9) for negative species makes the distribution function negative.

Let $F_0(v)$ and $f(x,v,t)$ denote the multicomponent fields for the equilibrium and perturbation, respectively, and κ the matrix of mass ratios,

$$f\equiv\begin{pmatrix} f^{(s_1)}\\ f^{(s_2)}\\ \vdots\end{pmatrix},\quad F_0\equiv\begin{pmatrix} F_0^{(s_1)}\\ F_0^{(s_2)}\\ \vdots\end{pmatrix},\quad \kappa\equiv\begin{pmatrix} \kappa^{(s_1)} & 0 & 0 & \cdots\\ 0 & \kappa^{(s_2)} & 0 & \cdots\\ \vdots & \vdots & \vdots & \end{pmatrix}, \tag{10}$$

then the system (7) and (8) can be concisely expressed as

$$\frac{\partial f}{\partial t}=\mathscr{L}f+\mathscr{N}(f), \tag{11}$$

where the linear operator is defined by

$$\mathscr{L}f=\sum_{l=-\infty}^{\infty}e^{ilx}(L_l f_l)(v), \tag{12}$$

$$(L_l f_l)(v)=\begin{cases}0 & l=0\\ -il[v f_l(v)+\kappa\cdot\eta_l(v)\Sigma_{s'}\int_{-\infty}^{\infty}dv' f_l^{(s')}(v')] & l\neq 0,\end{cases} \tag{13}$$

with $\eta_l(v)\equiv-\partial_v F_0/l^2$, and the nonlinear operator $\mathscr{N}$ is

$$\mathcal{N}(f)=\sum_{m=-\infty}^{\infty} e^{imx}\sum_{l=-\infty}^{\infty}{}' \frac{i}{l}\left(\kappa\cdot\frac{\partial f_{m-l}}{\partial v}\right)\sum_{s'}\int_{-\infty}^{\infty}dv' f_l^{(s')}(v'). \tag{14}$$

In the spatial Fourier expansion (12), l denotes an integer multiple of the basic wave number $2\pi/L$, and a primed summation as in (14) omits the $l=0$ term. The notation $\kappa\cdot\eta_l(v)$ or $\kappa\cdot\partial_v f_{m-l}$ denotes matrix multiplication.

An equilibrium $F_0(v)$ depends on parameters such as the densities or temperatures of the various species, but this dependence will be suppressed unless it is explicitly needed.

Symmetries of the model (11) and the equilibrium $F_0(v)$ are important qualitative features of the problem, and play a significant role in Sec. II when we formulate the amplitude expansions. Spatial translation, $\mathcal{T}_a:(x,v)\to(x+a,v)$, and reflection, $\mathcal{R}:(x,v)\to(-x,-v)$, act as operators on $f(x,v,t)$ in the usual way: if α denotes an arbitrary transformation then $(\alpha f)(x,v)\equiv f(\alpha^{-1}(x,v))$ where $(\alpha f)(x,v)$ denotes the transformed distribution function. The operators $\mathcal{L}$ and $\mathcal{N}$ commute with $\mathcal{T}_a$ due to the spatial homogeneity of F_0, and if $F_0(v)=F_0(-v)$, then $\mathcal{L}$ and $\mathcal{N}$ also commute with the reflection operator $\mathcal{R}$. With periodic boundary conditions, x is a periodic coordinate so $\mathcal{T}_a$ and $\mathcal{R}$ generate O(2), the symmetry group of the circle. If only the translation symmetry is present, then the group is SO(2).

An inner product is needed in Sec. II to derive the amplitude equation. For two multicomponent fields of (x,v), e.g., $B=(B^{(s_1)},B^{(s_2)},B^{(s_3)},\ldots)$ and $D=(D^{(s_1)},D^{(s_2)},D^{(s_3)},\ldots)$, we define their inner product by

$$(B,D)\equiv\sum_s\int_{-L/2}^{L/2}dx\int_{-\infty}^{\infty}dv B^{(s)}(x,v)^*D^{(s)}(x,v)=\int_{-L/2}^{L/2}dx\langle B,D\rangle, \tag{15}$$

where

$$\langle B,D\rangle\equiv\sum_s\int_{-\infty}^{\infty}dv B^{(s)}(x,v)^*D^{(s)}(x,v). \tag{16}$$

C. Summary of linear theory

The spectral theory for $\mathcal{L}$ is well established, and the needed results are simply recalled to establish our notation.[8–10] The eigenvalues $\lambda=-ilz$ of $\mathcal{L}$ are determined by the roots $\Lambda_l(z)=0$ of the ''spectral function,''

$$\Lambda_l(z)\equiv 1+\int_{-\infty}^{\infty}dv\,\frac{\Sigma_s\kappa^{(s)}\eta_l^{(s)}(v)}{v-z}. \tag{17}$$

If the contour in (17) is replaced by the Landau contour for $\mathrm{Im}(z)<0$ then we have the linear dielectric $\epsilon_l(z)$; for $\mathrm{Im}(z)>0$, $\Lambda_l(z)$ and $\epsilon_l(z)$ are the same function. For SO(2) symmetry, the eigenvalues are generically complex; for O(2) symmetry, when the equilibrium is also reflection-symmetric, then real eigenvalues can occur as in a two-stream instability; see, for example, Ref. 11.

Associated with an eigenvalue $\lambda=-ilz$ is the multicomponent eigenfunction $\Psi(x,v)=e^{ilx}\psi(v)$ where

$$\psi(v)=-\frac{\kappa\cdot\eta_l}{v-z}. \tag{18}$$

There is also an associated adjoint eigenfunction $\tilde{\Psi}(x,v)=e^{ilx}\tilde{\psi}(v)/L$ satisfying $(\tilde{\Psi},\Psi)=1$ with

$$\tilde{\psi}(v)=-\frac{1}{\Lambda_l'(z)^*(v-z^*)}. \tag{19}$$

Note that all components of $\tilde{\psi}(v)$ are the same. The normalization in (19) assumes that the root of $\Lambda_l(z)$ is simple and is chosen so that $\langle \tilde{\psi}, \psi \rangle = 1$. The adjoint determines the projection of $f(x,v,t)$ onto the eigenvector, and this projection defines the time-dependent amplitude of Ψ, i.e., $A(t) \equiv (\tilde{\Psi}, f)$.

II. AMPLITUDE EQUATION ON THE UNSTABLE MANIFOLD

A. Unstable linear modes

The equilibrium $F_0(v)$ is assumed to support a ''single'' unstable mode in the following sense. We shall assume that E^u, the unstable subspace for $\mathscr{L}$, is two dimensional. With translation symmetry and periodic boundary conditions, this is the simplest instability problem that can be posed.

Henceforth, let k denote the wave number of this unstable mode that is associated with the root $\Lambda_k(z_0) = 0$ which we assume to be simple, i.e., $\Lambda_k'(z_0) \neq 0$. The corresponding eigenvector is

$$\Psi_1(x,v) = e^{ikx}\psi(v) = e^{ikx}\left(- \frac{\kappa \cdot \eta_k}{v - z_0} \right). \tag{20}$$

The root $z_0 = v_p + i\gamma/k$ determines the phase velocity $v_p = \omega/k$ and the growth rate γ of the linear mode as the real and imaginary parts of the eigenvalue $\lambda = -ikz_0 = \gamma - i\omega$. From (17) and $\Lambda_k(z_0) = 0$, we obtain

$$\Lambda_l(z_0) = (l^2 - k^2)/l^2; \tag{21}$$

this identity will be useful shortly.

When $F_0(v)$ lacks reflection symmetry, then Ψ_1 typically has a nonzero phase velocity and λ is complex. In this case, the identities $\Lambda_k(z) = \Lambda_{-k}(z)$ and $\Lambda_k(z)^* = \Lambda_k(z^*)$ imply three additional modes: Ψ_1^*, Ψ_2, and Ψ_2^* where

$$\Psi_2(x,v) = e^{ikx}\psi(v)^*. \tag{22}$$

These eigenfunctions correspond to eigenvalues λ^*, $-\lambda^*$, and $-\lambda$, respectively, and fill out the eigenvalue quartet characteristic of Hamiltonian systems. In the absence of reflection symmetry, the eigenvalues are typically simple and the unstable subspace is two dimensional.

When $F_0(v)$ is reflection-symmetric, the eigenvalues may be either real or complex. In either case Ψ_j and $\mathscr{R} \cdot \Psi_j$ are linearly independent so the eigenvalues have multiplicity two. Now E^u is two dimensional only in the case of a real eigenvalue since a multiplicity-two complex conjugate pair implies a four-dimensional unstable subspace; this latter possibility is not considered further.

The components of the distribution function along the unstable eigenvectors Ψ_1 and Ψ_1^* are identified by writing

$$f(x,v,t) = [A(t)\Psi(x,v) + \text{c.c.}] + S(x,v,t), \tag{23}$$

where $A(t) = (\tilde{\Psi}, f)$ is the mode amplitude for Ψ and $(\tilde{\Psi}, S) = 0$. In (23), the subscript on Ψ_1 has been dropped, and $\tilde{\Psi} = \exp(ikx)\tilde{\psi}/L$ is the adjoint function for z_0 from (19).

The action of translation $\mathscr{T}_a$ and reflection $\mathscr{R}$ on $f(x,v,t)$ implies an action by these operators on the variables (A,S). From (23) we note that $\mathscr{T}_a f(x,v,t) = f(x-a,v,t)$ is equivalent to

$$\mathscr{T}_a(A, S(x,v)) = (e^{-ika}A, S(x-a,v)), \tag{24}$$

and reflection $\mathscr{R} f(x,v,t) = f(-x,-v,t)$ yields

$$\mathscr{R}(A, S(x,v)) = (A^*, S(-x,-v)). \tag{25}$$

When these transformations are symmetries, these relations are important for organizing the amplitude expansions introduced in Sec. III.

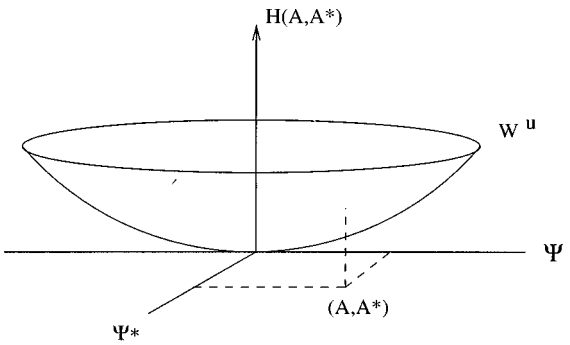

FIG. 1. Schematic representation of the unstable manifold.

B. Derivation of the amplitude equation

In the (A,S) variables, the Vlasov equation (11) becomes:

$$\dot{A}=\lambda A+(\tilde{\Psi},\mathcal{N}(f)), \tag{26}$$

$$\frac{\partial S}{\partial t}=\mathcal{L}S+\mathcal{N}(f)-[(\tilde{\Psi},\mathcal{N}(f))\Psi+\text{c.c.}], \tag{27}$$

where

$$(\tilde{\Psi},\mathcal{N}(f))=-i\sum_{l=-\infty}^{\infty}{}' \frac{1}{l}\langle\partial_v\tilde{\psi},\kappa\cdot f_{k-l}\rangle\sum_{s'}\int_{-\infty}^{\infty}dv' f_l^{(s')}(v'). \tag{28}$$

In writing (26) we have used the adjoint relationship $(\tilde{\Psi},\mathcal{L}S)=(\mathcal{L}^{\dagger}\tilde{\Psi},S)=\lambda(\tilde{\Psi},S)=0$ and in (28) an integration by parts shifts the velocity derivative onto $\tilde{\psi}$. These coupled equations are equivalent to (11); however, by restricting them to the unstable manifold we obtain an autonomous equation for $A(t)$. This procedure is briefly summarized; a more detailed description is provided in (I).

The unstable subspace E^u is invariant under the linear dynamics $\partial_t f=\mathcal{L}f$, but this no longer holds once the nonlinear terms $\mathcal{N}(f)$ are included. Instead we assume there is a two-dimensional invariant surface, the unstable manifold W^u, that represents the nonlinear deformation of E^u. The unstable manifold is tangent to E^u at the equilibrium, and near F_0 it can be described by a function $H(x,v,A,A^*)$ which locates the manifold ''above'' the point $[A\Psi+A^*\Psi^*]$ in E^u; the geometry is illustrated in Fig. 1.

Thus

$$f^u(x,v,t)=[A(t)\Psi(x,v)+\text{c.c.}]+H(x,v,A(t),A^*(t)) \tag{29}$$

represents a distribution function on W^u. In this expression, the evolution of S can be determined from H, i.e.,

$$S(x,v,t)=H(x,v,A(t),A^*(t))=\begin{pmatrix} H^{(s_1)}(x,v,A(t),A^*(t)) \\ H^{(s_2)}(x,v,A(t),A^*(t)) \\ \vdots \end{pmatrix}. \tag{30}$$

When this representation is substituted into (26) and (27) we obtain

$$\dot{A}=\lambda A+(\tilde{\Psi},\mathcal{N}(f^u)), \tag{31}$$

$$\left.\frac{\partial S}{\partial t}\right|_{f^u}=\mathcal{L}H+\mathcal{N}(f^u)-[(\tilde{\Psi},\mathcal{N}(f^u))\Psi+\text{c.c.}]. \tag{32}$$

Note that (31) defines an *autonomous* two-dimensional flow describing the self-consistent nonlinear evolution of the unstable mode; this is the amplitude equation we shall study.

Certain general features of the amplitude equation follow from the transformations (24) and (25) when these are symmetries of the Vlasov equation (11). For our problem, the amplitude equation always has translation symmetry (24) and we can apply standard results on the form of such symmetric equations. In particular, a two-dimensional vector field $\dot{A}=V(A,A^*)$ that is symmetric with respect to $A\to Ae^{-ika}$ can be written as $V(A,A^*)=Ap(\sigma)$ where $\sigma=|A|^2$ and $p(\sigma)$ is a smooth function determined from V.[12] Hence we know the right hand side of (31) takes the general form,

$$Ap(\sigma)=\lambda A+(\tilde{\Psi},\mathcal{N}(f^u)), \tag{33}$$

where the function $p(\sigma)$ must be determined from the Vlasov equation. Typically $p(\sigma)$ is complex-valued, however, when F_0 is reflection-symmetric then $p(\sigma)$ is forced to be real.

C. Analysis of $H(x,v,A,A^*)$

An equation for H follows by requiring consistency between (30) and (32). Equating the time derivative of (30) with the right hand side of (32) gives

$$\frac{\partial H}{\partial A}\dot{A}+\frac{\partial H}{\partial A^*}\dot{A}^*=\mathcal{L}H+\mathcal{N}(f^u)-[(\tilde{\Psi},\mathcal{N}(f^u))\Psi+\text{c.c.}], \tag{34}$$

which is to be solved for H subject to (35). Since the manifold is tangent to E^u at the equilibrium, H must satisfy

$$0=H(x,v,0,0)=\frac{\partial H}{\partial A}(x,v,0,0)=\frac{\partial H}{\partial A^*}(x,v,0,0). \tag{35}$$

The symmetries of the problem impose general restrictions on H because the unstable manifold is mapped onto itself by a symmetry transformation. For a translation this invariance means that $f^u(x-a,v,t)$ corresponds to a point on the manifold and thus can be written as,

$$(\mathcal{T}_a f^u)(x,v)=[A\Psi(x-a,v)+\text{c.c.}]+H(x-a,v,A,A^*) \tag{36}$$

$$=[A'\Psi(x,v)+\text{c.c.}]+H(x,v,A',A'^*). \tag{37}$$

Consistency between these two expressions requires that the transformed amplitude A' is related to the initial amplitude A by $A'=e^{-ika}A$ [cf. (24)], and this implies that H satisfies

$$H(x-a,v,A,A^*)=H(x,v,e^{ika}A,e^{-ika}A^*). \tag{38}$$

If, in addition, the system (26) and (27) has reflection symmetry, then similar reasoning yields

$$H(-x,-v,A,A^*)=H(x,v,A^*,A). \tag{39}$$

These general symmetry properties of H usefully constrain the form of the Fourier expansion

$$H(x,v,A,A^*)=\sum_{l=-\infty}^{\infty}e^{ilx}H_l(v,A,A^*)=\sum_{l=-\infty}^{\infty}e^{ilx}\begin{pmatrix}H_l^{(s_1)}(v,A,A^*)\\ H_l^{(s_2)}(v,A,A^*)\\ \vdots\end{pmatrix}. \tag{40}$$

Applying (38) shows that the components H_l must vanish unless l is an integer multiple of k and that the nonzero components have the general form

$$H_0(v,A,A^*)=\sigma h_0(v,\sigma),\quad H_k(v,A,A^*)=A\sigma h_1(v,\sigma),$$

$$H_{mk}(v,A,A^*)=A^m h_m(v,\sigma)\quad \text{for}\quad m\geqslant 2, \tag{41}$$

where $H_{-l}=H_l^*$. The functions h_m are not determined by symmetry; however, if reflection symmetry also holds, then

$$h_m(-v,\sigma)=h_m(v,\sigma)^*. \tag{42}$$

The reasoning leading to these symmetry results can be briefly outlined. The translation symmetry (38) requires

$$e^{-ila}H_l(v,A,A^*)=H_l(v,Ae^{-ika},A^*e^{ika}). \tag{43}$$

By choosing $a=2\pi/k$, this implies we find $H_l=0$ unless $\exp(2\pi i(l/k))=1$, hence the nonzero components satisfy $l=mk$ for integer m. Now recall that a function $G(v,A,A^*)$ which is invariant under $A\to Ae^{-ika}$ has the form $G(v,A,A^*)=g(v,\sigma)$ where $\sigma=|A|^2$ and $g(v,\sigma)$ is a smooth function determined by G. The relation in (43) implies that $G(v,A,A^*)=(A^*)^mH_{mk}(v,A,A^*)$ is such an invariant function. The factor $(A^*)^m$ in G requires that the corresponding $g(v,\sigma)$ contain an overall factor of σ^m; hence we set $g(v,\sigma)=\sigma^m h(v,\sigma)$. The resulting expression, $(A^*)^mH_{mk}=\sigma^m h(v,\sigma)$, then implies $H_{mk}=A^m h(v,\sigma)$. This general form, together with the requirement (35), leads to general expressions for the Fourier components of H in (41).

D. Equations for $p(\sigma)$ and $h_m(v,\sigma)$

The calculation of the amplitude equation (31) reduces to a determination of the function $p(\sigma)$ in (33); this requires the evaluation of $(\tilde{\Psi},\mathcal{N}(f^u))$ from (28) and involves the Fourier components of f^u

$$f_l^u=[A\psi(v)\delta_{l,k}+A^*\psi(v)^*\delta_{l,-k}]+H_l(v,A,A^*). \tag{44}$$

Combining (28) with (41) and (44) yields

$$\begin{aligned}(\tilde{\Psi},\mathcal{N}(f^u))=\frac{-iA\sigma}{k}\Bigg\{&\langle\partial_v\tilde{\psi},\kappa\cdot(h_0-h_2)\rangle+\frac{\Gamma_2}{2}\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle\\&+\sigma\Big[\Gamma_1\langle\partial_v\tilde{\psi},\kappa\cdot h_0\rangle-\Gamma_1^*\langle\partial_v\tilde{\psi},\kappa\cdot h_2\rangle+\frac{\Gamma_2}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_1^*\rangle-\frac{\Gamma_2^*}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_3\rangle\Big]\\&+\sum_{l=3}^{\infty}\frac{\sigma^{l-2}}{l}[\Gamma_l\langle\partial_v\tilde{\psi},\kappa\cdot h_{l-1}^*\rangle-\sigma\Gamma_l^*\langle\partial_v\tilde{\psi},\kappa\cdot h_{l+1}\rangle]\Bigg\}.\end{aligned} \tag{45}$$

In this expression, sum over the velocity integrals of $h_m^{(s)}$ is denoted by

$$\Gamma_m(\sigma)\equiv\sum_s\int_{-\infty}^{\infty}dv\,h_m^{(s)}(v,\sigma). \tag{46}$$

Now comparing (33) and (45) provides the desired expression for p

$$\begin{aligned}p(\sigma)=\lambda-\frac{i\sigma}{k}\Bigg\{&\langle\partial_v\tilde{\psi},\kappa\cdot(h_0-h_2)\rangle+\frac{\Gamma_2}{2}\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle+\sigma\Big[\Gamma_1\langle\partial_v\tilde{\psi},\kappa\cdot h_0\rangle-\Gamma_1^*\langle\partial_v\tilde{\psi},\kappa\cdot h_2\rangle\\&+\frac{\Gamma_2}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_1^*\rangle-\frac{\Gamma_2^*}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_3\rangle\Big]+\sum_{l=3}^{\infty}\frac{\sigma^{l-2}}{l}[\Gamma_l\langle\partial_v\tilde{\psi},\kappa\cdot h_{l-1}^*\rangle\\&-\sigma\Gamma_l^*\langle\partial_v\tilde{\psi},\kappa\cdot h_{l+1}\rangle]\Bigg\}.\end{aligned} \tag{47}$$

This expression for p involves the functions $h_m(v,\sigma)$ which define H. The equations that determine $h_m(v,\sigma)$ follow from the general equation for H in (34). In presenting these relations we shall use $\mathscr{P}_\perp$, the orthogonal projection defined by the unstable mode. Let $\mathscr{P}$ denote the projection operator onto $\psi(v)$ of a multicomponent function $g(v)$,

$$(\mathscr{P}g)(v) \equiv \langle \tilde{\psi}, g \rangle \psi(v), \tag{48}$$

with the orthogonal projection $\mathscr{P}_\perp \equiv I - \mathscr{P}$ given by

$$(\mathscr{P}_\perp g)(v) = g(v) - \psi(v)\langle \tilde{\psi}, g \rangle. \tag{49}$$

With the previous expression for $(\tilde{\Psi}, \mathscr{N}(f^u))$ in (45) and the notation in (41) for the Fourier components of H, the components of (34) take the form:

$$(p+p^*)\left[h_0 + \sigma \frac{\partial h_0}{\partial \sigma}\right] = \frac{i}{k}\frac{\partial}{\partial v}\kappa \cdot \left\{\left[\psi^* + \sigma(h_1^* - \psi\Gamma_1^*) + \sigma^2 h_1^*\Gamma_1 + \sum_{l=2}^{\infty} \frac{\sigma^{l-1}}{l} h_l^*\Gamma_l\right] - \text{c.c.}\right\}, \tag{50}$$

$$\begin{aligned}\left[(2p+p^*)h_1 - (L_k h_1) + (p+p^*)\sigma\frac{\partial h_1}{\partial \sigma}\right] &= \frac{i}{k}\mathscr{P}_\perp \frac{\partial}{\partial v}\kappa \cdot \left\{h_0 - h_2 + \frac{1}{2}\psi^*\Gamma_2 + \sigma\left[h_0\Gamma_1 - h_2\Gamma_1^* f + \frac{1}{2}h_1^*\Gamma_2 - \frac{1}{2}h_3\Gamma_2^*\right]\right. \\ &\left. + \sum_{l=3}^{\infty} \frac{\sigma^{l-2}}{l}[h_{l-1}^*\Gamma_l - \sigma h_{l+1}\Gamma_l^*]\right\},\end{aligned} \tag{51}$$

$$\begin{aligned}\left[2ph_2 - L_{2k}h_2 + (p+p^*)\sigma\frac{\partial h_2}{\partial \sigma}\right] &= \frac{i}{k}\frac{\partial}{\partial v}\kappa \cdot \left\{\psi + \sigma\left[h_1 + \psi\Gamma_1 - h_3 + \frac{1}{2}h_0\Gamma_2 + \frac{1}{3}\psi^*\Gamma_3\right]\right. \\ &\left. \times\sigma^2\left[h_1\Gamma_1 - h_3\Gamma_1^* - \frac{1}{2}h_4\Gamma_2^* + \frac{1}{3}h_1^*\Gamma_3\right] - \frac{\sigma^3}{3}h_5\Gamma_3^* + \sum_{l=4}^{\infty}\frac{\sigma^{l-2}}{l}[h_{l-2}^*\Gamma_l - \sigma^2 h_{l+2}\Gamma_l^*]\right\},\end{aligned} \tag{52}$$

and

$$\begin{aligned}\left[mph_m - L_{mk}h_m + (p+p^*)\sigma\frac{\partial h_m}{\partial \sigma}\right] &= \frac{i}{k}\frac{\partial}{\partial v}\kappa \cdot \left\{h_{m-1} + \frac{\psi}{m-1}\Gamma_{m-1} + \sum_{l=2}^{m-2}\frac{h_{m-l}}{l}\Gamma_l + \sigma\left[h_{m-1}\Gamma_1\right.\right. \\ &\left. - h_{m+1} + \frac{h_1}{m-1}\Gamma_{m-1} + \frac{h_0}{m}\,amma_m + \frac{\psi^*}{m+1}\Gamma_{m+1}\right] + \sigma^2\left[-h_{m+1}\Gamma_1^* + \frac{h_1^*}{m+1}\Gamma_{m+1}\right] \\ &\left. + \sum_{l=m+2}^{\infty}\frac{\sigma^{l-m}}{l}h_{l-m}^*\Gamma_l - \sum_{l=2}^{\infty}\frac{\sigma^l}{l}h_{m+l}\Gamma_l^*\right\}\end{aligned} \tag{53}$$

for $m=0,1,2$, and $m>2$, respectively.

Together with (47) these component equations determine the functions $p(\sigma)$ and $\{h_m(\sigma)\}_{m=0}^{\infty}$; however, they cannot be solved except using the expansions introduced in the next section. From a practical standpoint, we have achieved a reduction of the problem to the analysis of functions of a *single* real variable, i.e., σ. In the study of the amplitude equation (31) this reduction represents a very useful simplification.

III. EXPANSIONS, RECURSION RELATIONS, AND PINCHING SINGULARITIES

The amplitude equation on the unstable manifold,

$$\dot{A}=Ap(\sigma), \tag{54}$$

is analyzed by expressing $p(\sigma)$, $h_m(v,\sigma)$, and $\Gamma_m(\sigma)$ as power series in σ:

$$p(\sigma)=\sum_{j=0}^{\infty} p_j\sigma^j, \quad h_m(v,\sigma)=\sum_{j=0}^{\infty} h_{m,j}(v)\sigma^j, \quad \Gamma_m(\sigma)=\sum_{j=0}^{\infty} \Gamma_{m,j}\sigma^j, \tag{55}$$

where $\Gamma_{m,j}=\int dv\,\Sigma_s h_{m,j}^{(s)}(v)$.

The coefficients p_j and $h_{m,j}(v)$ are calculated by recursively solving (47) and (50)–(53). A recursion relation for p_j follows from (47) by inserting the expansions above and solving at each order in σ; this calculation yields

$$p_0=\lambda, \tag{56}$$

$$p_j=\frac{i}{k}[\mathscr{A}_j+\mathscr{B}_j] \quad \text{for} \quad j\geqslant 1, \tag{57}$$

where

$$\mathscr{A}_j=-\langle\partial_v\tilde{\psi},\kappa\cdot(h_{0,j-1}-h_{2,j-1})\rangle-\frac{1}{2}\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle\Gamma_{2,j-1}-\sum_{l=0}^{j-2}[\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,j-l-2}\rangle\Gamma_{1,l}$$
$$-\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,j-l-2}\rangle\Gamma_{1,l}^*], \tag{58}$$

$$\mathscr{B}_j=-\sum_{l=0}^{j-2}\left\{\frac{1}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_{1,j-l-2}^*\rangle\Gamma_{2,l}+\sum_{m=0}^{l}\left[\frac{\Gamma_{j-l+1,m}}{j-l+1}\langle\partial_v\tilde{\psi},\kappa\cdot h_{j-l,l-m}^*\rangle\right.\right.$$
$$\left.\left.-\frac{\Gamma_{j-l,m}^*}{j-l}\langle\partial_v\tilde{\psi},\kappa\cdot h_{j-l+1,l-m}\rangle\right]\right\}. \tag{59}$$

Here and below, a summation is understood to be omitted if the lower limit exceeds the upper limit. This organization of the terms in (58) and (59) will turn out to distinguish to different singular behaviors: $\mathscr{A}_j\sim\gamma^{-(5j-1)}$ and $\mathscr{B}_j\sim\gamma^{-(5j-2)}$.

The corresponding relations for $h_{m,j}(v)$ are determined similarly by applying the expansions (55) to the general equations (50)–(53); the resulting equations are solved for $h_{m,j}(v)$. Our results are conveniently stated in terms of the resolvent operator $R_l(w)\equiv(w-L_l)^{-1}$ and certain auxiliary functions $I_{m,j}(v)$ whose detailed expressions are provided below. For $m=0$, $h_{0,j}$ is simply

$$h_{0,j}(v)=\frac{I_{0,j}(v)}{(1+j)(\lambda+\lambda^*)}, \tag{60}$$

and the coefficients $h_{m,j}$ for $m>0$ require the resolvent

$$h_{m,j}(v)=R_{mk}(w_{m,j})I_{m,j} \tag{61}$$

with the complex numbers $w_{m,j}$ defined by

$$w_{m,j} \equiv (j+\delta_{m,1})(\lambda+\lambda^*)+m\lambda = 2(j+\delta_{m,1})\gamma + m\lambda. \tag{62}$$

A general expression for $R_l(w)$ follows from (13) by solving $(w-L_l)f=g$ for f.[10,13] Here both $g(v)=(g^{(s_1)}(v),g^{(s_2)}(v),\ldots)$ and f are multicomponent fields, and $R_l(w)$ acts by

$$R_l(w)g = \begin{pmatrix} (R_l(w)g)^{(s_1)}(v) \\ (R_l(w)g)^{(s_2)}(v) \\ \vdots \end{pmatrix}, \tag{63}$$

where

$$(R_l(w)g)^{(s)}(v) = \frac{1}{il(v-iw/l)}\left[g^{(s)}(v) - \frac{\kappa^{(s)}\eta_l^{(s)}}{\Lambda_l(iw/l)} \sum_{s'} \int_{-\infty}^{\infty} dv' \, \frac{g^{(s')}(v')}{v'-iw/l} \right]. \tag{64}$$

Note the following significant feature: the arguments $w_{m,j}$ of the resolvent in (61) determine *poles* of $h_{m,j}(v)$ located at $v=z_{m,j}$ where $z_{m,j} \equiv iw_{m,j}/mk$. For $m \geqslant 1$, these poles always fall in the upper half-plane above the phase velocity v_p:

$$z_{m,j} = z_0 + \frac{i\gamma d_{m,j}}{k} = v_p + \frac{i\gamma(1+d_{m,j})}{k}, \tag{65}$$

where $d_{m,j} \equiv 2(j+\delta_{m,1})/m$.

A. Recursion relations for $I_{m,j}$

The auxiliary functions $I_{0,j}(v)$ are defined by

$$I_{0,0}(v) = \frac{i}{k}\frac{\partial}{\partial v}\kappa\cdot(\psi^*-\psi) \tag{66}$$

for $j=0$, and for $j \geqslant 1$

$$I_{0,j}(v) = -\sum_{n=0}^{j-1}(1+n)(p_{j-n}+p_{j-n}^*)h_{0,n}(v) + \frac{i}{k}\frac{\partial}{\partial v}\kappa\cdot\left\{ \left[h_{1,j-1}^* - \psi\Gamma_{1,j-1}^* + \sum_{n=0}^{j-2} h_{1,n}^*\Gamma_{1,j-n-2} + \sum_{n=0}^{j-1}\sum_{l=0}^{n} \frac{h_{j-n+1,l}^*}{j-n+1}\Gamma_{j-n+1,n-l} \right] - \text{c.c.} \right\}. \tag{67}$$

For $m=1$, the functions $I_{1,j}(v)$ are given by

$$\begin{aligned} I_{1,j}(v) = &-\sum_{n=0}^{j-1}[(2+n)p_{j-n}+(1+n)p_{j-n}^*]h_{1,n} + \frac{i}{k}\mathscr{P}_\perp\frac{\partial}{\partial v}\kappa\cdot\Bigg\{ h_{0,j} - h_{2,j} + \frac{1}{2}\psi^*\Gamma_{2,j} \\ &+ \sum_{n=0}^{j-1}\Bigg[h_{0,n}\Gamma_{1,j-n-1} - h_{2,n}\Gamma_{1,j-n-1}^* + \frac{1}{2}h_{1,n}^*\Gamma_{2,j-n-1} - \frac{1}{2}h_{3,n}\Gamma_{2,j-n-1}^* \\ &+ \sum_{m=0}^{n}\left(\frac{h_{j-n+1,m}^*}{j-n+2}\Gamma_{j-n+2,n-m} \right)\Bigg] - \sum_{n=0}^{j-2}\sum_{m=0}^{n}\left[\frac{h_{j-n+2,m}}{j-n+1}\Gamma_{j-n+1,n-m}^* \right]\Bigg\}. \end{aligned} \tag{68}$$

The notation $\mathscr{P}_\perp g(v)$ was defined in (49): $\mathscr{P}_\perp g(v) = g(v) - \psi(v)\langle\tilde{\psi},g\rangle$.

For $m=2$ the functions $I_{2,j}(v)$ are given by

$$I_{2,0}(v) = \frac{i}{k}\frac{\partial}{\partial v}\kappa\cdot\psi \tag{69}$$

for $j=0$, and for $j \geqslant 1$

$$I_{2,j}(v)=-\sum_{n=0}^{j-1}[(2+n)p_{j-n}+np_{j-n}^{*}]h_{2,n}+\frac{i}{k}\frac{\partial}{\partial v}\kappa\cdot\Bigg\{h_{1,j-1}+\psi\Gamma_{1,j-1}-h_{3,j-1}+\frac{1}{3}\psi^{*}\Gamma_{3,j-1}$$
$$+\frac{1}{2}\sum_{n=0}^{j-1}h_{0,n}\Gamma_{2,j-n-1}+\sum_{n=0}^{j-2}\Bigg[h_{1,n}\Gamma_{1,j-n-2}-h_{3,n}\Gamma_{1,j-n-2}^{*}-\frac{1}{2}h_{4,n}\Gamma_{2,j-n-2}^{*}$$
$$+\frac{1}{3}h_{1,n}^{*}\Gamma_{3,j-n-2}+\sum_{m=0}^{n}\frac{h_{j-n,m}^{*}}{j-n+2}\Gamma_{j-n+2,n-m}\Bigg]-\sum_{n=0}^{j-3}\frac{h_{5,n}}{3}\Gamma_{3,j-n-3}^{*}$$
$$-\sum_{n=0}^{j-4}\sum_{m=0}^{n}\frac{h_{j-n+2,m}}{j-n}\Gamma_{j-n,n-m}^{*}\Bigg\}. \tag{70}$$

For $m>2$, the functions $I_{m,j}(v)$ are given by

$$I_{m,j}(v)=-\sum_{n=0}^{j-1}[(m+n)p_{j-n}+np_{j-n}^{*}]h_{m,n}(v)+\frac{i}{k}\frac{\partial}{\partial v}\kappa\cdot\Bigg\{h_{m-1,j}+\frac{\psi}{m-1}\Gamma_{m-1,j}$$
$$+\sum_{l=2}^{m-2}\sum_{n=0}^{j}\frac{h_{m-l,n}}{l}\Gamma_{l,j-n}-h_{m+1,j-1}+\frac{\psi^{*}}{m+1}\Gamma_{m+1,j-1}+\sum_{n=0}^{j-1}\Bigg[h_{m-1,n}\Gamma_{1,j-n-1}$$
$$+\frac{h_{1,n}}{m-1}\Gamma_{m-1,j-n-1}+\frac{h_{0,n}}{m}\Gamma_{m,j-n-1}\Bigg]+\sum_{n=0}^{j-2}\Bigg[-h_{m+1,n}\Gamma_{1,j-n-2}^{*}+\frac{h_{1,n}^{*}}{m+1}\Gamma_{m+1,j-n-2}$$
$$+\sum_{l=0}^{n}\left(\frac{h_{j-n,l}^{*}}{j+m-n}\Gamma_{j+m-n,n-l}-\frac{h_{m+j-n,l}}{j-n}\Gamma_{j-n,n-l}^{*}\right)\Bigg]\Bigg\}; \tag{71}$$

in this last expression, if a subscript is negative the term is understood to be omitted, e.g., for $j=0$, $h_{m+1,j-1}$ is omitted.

B. Useful identities

The relations (57), (60) and (61), and (66)–(71) can be applied systematically to calculate p_j and $h_{m,j}$ to any order. The leading coefficient p_0 is determined by linear theory (56) and from the linear eigenfunction ψ one can also calculate $h_{0,0}$ and $h_{2,0}$, c.f. (66) and (69), respectively. These two coefficients then suffice to calculate p_1 from (57). From $\{p_1,h_{0,0},h_{2,0}\}$, the coefficients $h_{1,0}$ and $h_{3,0}$ can be determined and then $h_{0,1}$ and $h_{2,1}$. This provides the input to calculate p_2, and from this point on the structure of the calculation to all orders falls into the simple pattern summarized in Table I.

For the remainder of the paper, our analysis of the expansion (55) is facilitated by several identities which we summarize here. These relations allow $\Gamma_{m,j},h_{m,j}$, and certain important integrals to be obtained very simply from the auxiliary functions $I_{m,j}$.

If $m=0$ we note that (60) and (66) and (67) imply $\Gamma_{0,j}=0$. For $m>0$, by substituting (61) into $\Gamma_{m,j}=\int dv\,\Sigma_s h_{m,j}^{(s)}$ and rearranging, we obtain

$$\Gamma_{m,j}=\frac{-i/mk}{\Lambda_{mk}(z_{m,j})}\sum_s\int_{-\infty}^{\infty}dv\,\frac{I_{m,j}^{(s)}(v)}{v-z_{m,j}}\quad(m>0). \tag{72}$$

With (72) for $\Gamma_{m,j}$, the general form for $h_{m,j}$ in (61) can be reexpressed as

$$h_{m,j}(v)=\left(\frac{-i}{mk}\right)\frac{I_{m,j}(v)}{v-z_{m,j}}-\left(\frac{\kappa\cdot\eta_{mk}}{v-z_{m,j}}\right)\Gamma_{m,j}\quad(m>0). \tag{73}$$

In this way $\Gamma_{m,j}$ and $h_{m,j}$ are written explicitly in terms of $I_{m,j}$.

The coefficients p_j depend on integrals of the form $\langle\partial_v\tilde{\psi},\kappa\cdot h_{m,j}\rangle$, and it is helpful to express these integrals directly in terms of $I_{m,j}$ also. From (60) we obtain the identity

TABLE I. Order of calculation of $h_{m,j}(v)$ and p_j from $\psi(v)$. The flow of calculation of the $h_{m,j}(v)$ is indicated by moving downward. From $\psi(v)$, $h_{0,0}$ and $h_{2,0}$ can be calculated and then p_1 determined; $h_{1,0}$ and $h_{3,0}$ are calculated next from $\{p_1, h_{0,0}, h_{2,0}\}$ and then $h_{0,1}$ and $h_{2,1}$ can be evaluated. This then determines p_2, and so forth. For $j_1 \geqslant 2$, p_{j_1} requires prior calculation of $h_{m,j}$ for $0 \leqslant m \leqslant j_1+1$ and $0 \leqslant j \leqslant j_1 - m + 1 - 2(\delta_{m,0}+\delta_{m,1})$.

	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$\cdots$
p_0		$\psi(v)$						
p_1	$h_{0,0}$	–	$h_{2,0}$					
		$h_{1,0}$		$h_{3,0}$				
p_2	$h_{0,1}$		$h_{2,1}$					
		$h_{1,1}$			$h_{4,0}$			
p_3	$h_{0,2}$		$h_{2,2}$	$h_{3,1}$				
		$h_{1,2}$				$h_{5,0}$		
p_4	$h_{0,3}$		$h_{2,3}$	$h_{3,2}$	$h_{4,1}$			
		$h_{1,3}$					$h_{6,0}$	
p_5	$h_{0,4}$		$h_{2,4}$	$h_{3,3}$	$h_{4,2}$	$h_{5,1}$		
$\vdots$								

$$\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa}\cdot h_{0,j}\rangle = \frac{1}{2\gamma(1+j)\Lambda_k'(z_0)} \int_{-\infty}^{\infty} dv\ \frac{\Sigma_s \boldsymbol{\kappa}^{(s)} I_{0,j}^{(s)}}{(v-z_0)^2}; \tag{74}$$

similarly, by integrating (73) we find

$$\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa}\cdot h_{m,j}\rangle = \left(\frac{-i}{mk\Lambda_k'(z_0)}\right)\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s \boldsymbol{\kappa}^{(s)} I_{m,j}^{(s)}}{(v-z_0)^2(v-z_{m,j})} - \left(\frac{\Gamma_{m,j}}{\Lambda_k'(z_0)}\right)\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s (\boldsymbol{\kappa}^{(s)})^2 \boldsymbol{\eta}_{mk}^{(s)}}{(v-z_0)^2(v-z_{m,j})}. \tag{75}$$

C. Analysis of the cubic coefficient

It is instructive at this point to evaluate and examine the cubic coefficient p_1 in detail. This coefficient illustrates the occurrence of pinching singularities and the procedures needed to evaluate such singular integrals.

From (57) we have

$$p_1 = -\frac{i}{k}\left[\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa}\cdot(h_{0,0}-h_{2,0})\rangle + \frac{\Gamma_{2,0}}{2}\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa}\cdot\psi^*\rangle\right] \tag{76}$$

so $h_{0,0}(v)$, $h_{2,0}(v)$ and $\Gamma_{2,0}$ are needed to calculate p_1. From (60) and (66) we have

$$h_{0,0}(v) = -\frac{1}{k^2}\frac{\partial}{\partial v}\left[\frac{\boldsymbol{\kappa}^2\cdot\boldsymbol{\eta}_k}{(v-z_0)(v-z_0^*)}\right], \tag{77}$$

where $\boldsymbol{\kappa}^2$ denotes the square of the mass matrix. From (21), (69) and (72) we obtain

$$\Gamma_{2,0} = -\frac{2}{3k^2}\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\boldsymbol{\kappa}^2\cdot\boldsymbol{\eta}_k)^{(s)}}{(v-z_0)^3}, \tag{78}$$

and inserting (69) into (73) yields

$$h_{2,0}(v) = \frac{1}{2k^2}\left(\frac{\boldsymbol{\kappa}\cdot\partial_v\psi}{v-z_0}\right) - \frac{\Gamma_{2,0}}{4}\left(\frac{\boldsymbol{\kappa}\cdot\boldsymbol{\eta}_k}{v-z_0}\right). \tag{79}$$

Here we have made the substitution $\boldsymbol{\eta}_{2k} = \boldsymbol{\eta}_k/4$.

We are concerned with the form of the various integrals in p_1 in the asymptotic regime of small growth rate $\gamma \to 0^+$. The behavior of these integrals is determined by any complex singularities of the integrand that approach the real velocity axis as $\gamma \to 0^+$. We shall always assume that whatever complex analytic structure characterizes F_0 there are never any singularities that move to the real axis in this limit and therefore the singularities in η_k can be regarded as irrelevant in analyzing a given integral.

The singularities of interest are associated with the explicit poles in the denominator of a given integrand. For example, $\Gamma_{2,0}$ has a third-order pole at $v = z_0$ which approaches the real axis from above as $\gamma \to 0^+$. In this case, there are no corresponding poles in the lower half-plane, and consequently no pinching singularity develops; therefore $\Gamma_{2,0}$ has a finite limit as the growth rate goes to zero. Similarly, we see by inspection that $h_{2,0}$ has a third-order pole at $v = z_0$ and consequently the integrand of $\langle \partial_v \tilde{\psi}, \kappa \cdot h_{2,0} \rangle$ has a fifth-order pole at $v = z_0$. However, again there is no pinching singularity and this integral also has a finite limit as $\gamma \to 0^+$.

The remaining integrals, $\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle$ and $\langle \partial_v \tilde{\psi}, \kappa \cdot h_{0,0} \rangle$, are more interesting. The simplest example is $\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle$ whose integrand has a double pole at $v = z_0$ and a simple pole at $v = z_0^*$:

$$\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle = -\frac{1}{\Lambda_k'(z_0)} \sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{(v - z_0)^2 (v - z_0^*)}. \tag{80}$$

As $\gamma \to 0^+$, these poles approach the real axis from above and below; this creates a pinching singularity and the integral has a singular limit. We analyze this integral in two steps that can be applied to all such singular integrals: a partial fraction expansion extracts the singularity and the dispersion relation is applied to eliminate integrals over the electron distribution function. The motivation for the second step is to incorporate the one relation between the different distribution functions and to expose the effect of treating the ions as mobile.

The partial fraction expansion of the integrand yields

$$\frac{1}{(v - z_0)^2 (v - z_0^*)} = \left(\frac{ik}{2\gamma} \right)^2 \left[\frac{1}{v - z_0^*} - \frac{1}{v - z_0} \right] - \left(\frac{ik}{2\gamma} \right) \frac{1}{(v - z_0)^2} \tag{81}$$

so that our integral becomes

$$\sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{(v - z_0)^2 (v - z_0^*)} = \left(\frac{ik}{2\gamma} \right)^2 \left[\sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{v - z_0^*} - \sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{v - z_0} \right] - \left(\frac{ik}{2\gamma} \right) \sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{(v - z_0)^2}. \tag{82}$$

On the right hand side of this expression all the integrals are free of pinching singularities and therefore have finite limits as $\gamma \to 0^+$. The singularities of the original integral are now explicitly shown in the factors γ^{-2} and γ^{-1}.

The dispersion relation for the unstable mode $\Lambda_k(z_0) = 0$ implies the identities

$$\int_{-\infty}^{\infty} dv \, \frac{\eta_k^{(e)}(v)}{v - z_0} = 1 + \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s' \kappa^{(s)} \eta_k^{(s)}(v)}{v - z_0}, \tag{83}$$

$$\int_{-\infty}^{\infty} dv \, \frac{\eta_k^{(e)}(v)}{(v - z_0)^2} = -\Lambda_k'(z_0) - \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s^* \kappa^{(s)} \eta_k^{(s)}(v)}{(v - z_0)^2}, \tag{84}$$

where we have used $\kappa^{(e)} = -1$ and the primed sum indicates that the electron term is omitted. Substitution of (83) and (84) for the electron integrals in (82) yields

$$\sum_s \int_{-\infty}^{\infty} \frac{dv (\kappa^2 \cdot \eta_k)^{(s)}}{(v - z_0)^2 (v - z_0^*)} = \left(\frac{ik}{2\gamma} \right)^2 \left[-2i \sum_s{}' \kappa^{(s)} (1 + \kappa^{(s)}) \mathrm{Im} \int_{-\infty}^{\infty} \frac{dv \, \eta_k^{(s)}}{v - z_0} \right]$$

$$+\left(\frac{ik}{2\gamma}\right)\left[\Lambda_k'(z_0)+\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)})\int_{-\infty}^{\infty}\frac{dv\,\eta_k^{(s)}}{(v-z_0)^2}\right]. \quad (85)$$

Combining this with (80) we conclude that $\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle$ typically diverges like γ^{-2} in the small growth rate limit. There are exceptions, however; for example, in the limit of fixed ions, $\kappa^{(s)}\to 0$ for $s\neq e$, then we are left with

$$\sum_s\int_{-\infty}^{\infty}\frac{dv(\kappa^2\cdot\eta_k)^{(s)}}{(v-z_0)^2(v-z_0^*)}\to\left(\frac{ik}{2\gamma}\right)\Lambda_k'(z_0), \quad (86)$$

and the divergence is merely γ^{-1}.

We treat the remaining integral $\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle$ in the same fashion. This integrand now has a fourth-order pole at $v=z_0$ and a simple pole at $v=z_0^*$:

$$\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle=-\frac{2}{k^2\Lambda_k'(z_0)}\sum_s\int_{-\infty}^{\infty}\frac{dv(\kappa^3\cdot\eta_k)^{(s)}}{(v-z_0)^4(v-z_0^*)}. \quad (87)$$

Again we make a partial fraction expansion

$$\frac{1}{(v-z_0)^4(v-z_0^*)}=\left(\frac{ik}{2\gamma}\right)^4\left[\frac{1}{v-z_0^*}-\frac{1}{v-z_0}\right]-\left(\frac{ik}{2\gamma}\right)^3\frac{1}{(v-z_0)^2}-\left(\frac{ik}{2\gamma}\right)^2\frac{1}{(v-z_0)^3}$$
$$-\left(\frac{ik}{2\gamma}\right)\frac{1}{(v-z_0)^4}, \quad (88)$$

and eliminate the electron integrals to find

$$\sum_s\int_{-\infty}^{\infty}\frac{dv(\kappa^3\cdot\eta_k)^{(s)}}{(v-z_0)^4(v-z_0^*)}=\left(\frac{ik}{2\gamma}\right)^4\left[+2i\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\,\mathrm{Im}\int_{-\infty}^{\infty}\frac{dv\,\eta_k^{(s)}}{v-z_0}\right]$$
$$+\left(\frac{ik}{2\gamma}\right)^3\left[-\Lambda_k'(z_0)+\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\int_{-\infty}^{\infty}\frac{dv\,\eta_k^{(s)}}{(v-z_0)^2}\right]$$
$$+\left(\frac{ik}{2\gamma}\right)^2\left[-\frac{\Lambda_k''(z_0)}{2}+\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\int_{-\infty}^{\infty}\frac{dv\,\eta_k^{(s)}}{(v-z_0)^3}\right]$$
$$+\left(\frac{ik}{2\gamma}\right)\left[-\frac{\Lambda_k'''(z_0)}{6}+\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\int_{-\infty}^{\infty}\frac{dv\,\eta_k^{(s)}}{(v-z_0)^4}\right]. \quad (89)$$

Thus for $\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle$ the typical divergence is γ^{-4}.

Combining these results with p_1 in (76), we obtain the asymptotic form of the cubic coefficient as $\gamma\to 0^+$

$$p_1=\frac{1}{\gamma^4}[c_1(\gamma)-\gamma d_1(\gamma)+\mathcal{O}(\gamma^2)], \quad (90)$$

where c_1 and d_1 are nonsingular functions of γ defined by

$$c_1(\gamma)=-\frac{k}{4\Lambda_k'(z_0)}\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\,\mathrm{Im}\left(\int_{-\infty}^{\infty}dv\,\frac{\eta_k^{(s)}}{v-z_0}\right), \quad (91)$$

$$d_1(\gamma)=\frac{1}{4}-\frac{1}{4\Lambda_k'(z_0)}\sum_s{}'\kappa^{(s)}(1-\kappa^{(s)2})\int_{-\infty}^{\infty}dv\,\frac{\eta_k^{(s)}}{(v-z_0)^2}; \quad (92)$$

this result was briefly presented without derivation in an earlier paper.[7] When the leading term is nonzero at $\gamma=0$ [i.e., $c_1(0)\neq 0$], then p_1 diverges like γ^{-4} as a result of the divergence in $\langle \partial_v \tilde{\psi}, \kappa\cdot h_{0,0}\rangle$.

The significance of this result is best understood in light of our earlier discussion of scaling the amplitude $A(t)$ to obtain an amplitude equation free of singular features at small growth rates. With our expression for p_1, the amplitude equation for the unstable mode is

$$\dot{A}=A\left[\lambda+\frac{[c_1(0)+\mathcal{O}(\gamma)]}{\gamma^4}\left|A\right|^2+p_2(\gamma)\left|A\right|^4+\cdots\right], \tag{93}$$

which we rewrite using polar variables $A=\rho e^{-i\theta}$ as

$$\dot{\rho}=\gamma\rho+\frac{1}{\gamma^4}[\mathrm{Re}(c_1(0))+\mathcal{O}(\gamma)]\rho^3+\mathrm{Re}(p_2)\rho^5+\cdots, \tag{94}$$

$$\dot{\theta}=\omega-\frac{1}{\gamma^4}[\mathrm{Im}(c_1(0))+\mathcal{O}(\gamma)]\rho^2-\mathrm{Im}(p_2)\rho^4+\cdots. \tag{95}$$

Now we introduce a scaled amplitude $\rho(t)=\gamma^\beta r(\gamma t)$ following (3) and anticipate that $\delta=1$ is required to remove the singular effect of the linear term:

$$\frac{dr}{d\tau}=r+\frac{\gamma^{2\beta-1}}{\gamma^4}[\mathrm{Re}(c_1(0))+\mathcal{O}(\gamma)]r^3+\gamma^{4\beta-1}\ \mathrm{Re}(p_2)r^5+\cdots, \tag{96}$$

$$\dot{\theta}=\omega-\frac{\gamma^{2\beta}}{\gamma^4}[\mathrm{Im}(c_1(0))+\mathcal{O}(\gamma)]r^2-\gamma^{4\beta}\ \mathrm{Im}(p_2)r^4+\cdots. \tag{97}$$

The exponent β should be chosen so that the nonlinear coefficients have finite limits as $\gamma\to 0^+$. The most stringent requirements on β arise from the coefficients in the $dr/d\tau$ equation, and the cubic term requires $2\beta\geqslant 5$ to absorb the γ^{-4} singularity. Setting $\beta=5/2$ suffices to cancel the singularities in p_1 and allow the cubic nonlinearity to formally balance with the linear term at small γ:

$$\frac{dr}{d\tau}=r+[\mathrm{Re}(c_1(0))+\mathcal{O}(\gamma)]r^3+\gamma^9\ \mathrm{Re}(p_2)r^5+\cdots, \tag{98}$$

$$\dot{\theta}=\omega-\gamma[\mathrm{Im}(c_1(0))+\mathcal{O}(\gamma)]r^2-\gamma^{10}\ \mathrm{Im}(p_2)r^4+\cdots. \tag{99}$$

This choice for β assumes that $\mathrm{Re}(c_1(0))\neq 0$ and that the singularities at fifth (and higher) order are also removed by the same scaling that absorbs the cubic singularity. Typically the condition $\mathrm{Re}(c_1(0))\neq 0$ is satisfied, but there are interesting exceptions and they are discussed next. The adequacy of the $\beta=5/2$ scaling in controlling higher-order singularities is studied in Sec. IV.

D. Special cases: $c_1(0)=0$

Evaluating $c_1(0)$ as $\gamma\to 0^+$ yields

$$c_1(0)=-\frac{\pi k}{4\Lambda_k'(v_p+i0)}\sum_s{}'\,\kappa^{(s)}(1-\kappa^{(s)2})\,\eta_k^{(s)}(v_p(0)), \tag{100}$$

where $v_p(0)$ denotes the linear phase velocity at zero growth rate. Since the first factor $(\pi k/4\Lambda_k')$ cannot vanish, the special case $c_1(0)=0$ will only arise if the species sum (which excludes the electrons) vanishes. By carefully adjusting the parameters of a given model, one can presumably arrange for exact cancellations between different terms in this sum, but there is little motivation to study such artificial situations.

However, there are three natural circumstances in which $c_1(0)=0$ because every term in the sum vanishes independently, and the cubic coefficient is clearly less singular:

(a) Infinitely massive ions: $\kappa^{(s)}=0$ for all $s\neq e$.
(b) Zero slope for the resonant ions: $\eta_k^{(s)}(v_p(0))=0$ for all $s\neq e$.
(c) An electron-positron plasma: $\kappa^{(p)2}=1$ for positrons.
In all three cases, the cubic coefficient has the asymptotic form

$$p_1=-\frac{1}{\gamma^3}[d_1(0)+\mathcal{O}(\gamma)] \tag{101}$$

with $d_1(0)$ given by the full formula (92) for the second case and $d_1(0)=1/4$ in the first and third cases.

In the rescaled equation (96), this weakened singularity requires only that $2\beta\geqslant 4$, and setting $\beta=2$ yields

$$\frac{dr}{d\tau}=r+[\mathrm{Re}(d_1(0))+\mathcal{O}(\gamma)]r^3+\gamma^7\ \mathrm{Re}(p_2)r^5+\cdots, \tag{102}$$

$$\dot{\theta}=\omega-[\mathrm{Im}(d_1(0))+\mathcal{O}(\gamma)]r^2-\gamma^8\ \mathrm{Im}(p_2)r^4+\cdots. \tag{103}$$

In (I), we have shown that the exponent $\beta=2$ suffices to absorb the singularities in (102) and (103) to *all* orders for a plasma with fixed ions; this exponent corresponds to the relatively familiar trapping scaling for the electric field of an unstable mode. In Sec. V, we provide a new proof of this result and also establish the same conclusion for the case (b) where each ion distribution is flat at the phase velocity: $\eta_k^{(s)}(v_p)=0$ for all $s\neq e$. This will occur, for example, if the ions are sufficiently cold and do not populate the resonant region in velocity space. On the other hand, when there is an exact reflection symmetry $F_0(v)=F_0(-v)$, we can have an unstable mode with $v_p=0$ and $\eta_k^{(s)}(0)=0$ holds even though the ion distribution function is nonzero at $v=0$. This latter situation arises for instabilities of reflection-symmetric equilibria due to real eigenvalues, e.g., a reflection-symmetric two-stream instability.

Our understanding of the third example, an electron-positron plasma, is less complete. In Sec. V B we obtain by explicit calculation $\mathrm{Re}(p_2)\sim\gamma^{-8}$ so the rescaled fifth-order term in (102) remains singular unless $\beta\geqslant 9/4$. However, the singularity structure of the full expansion is more complicated and we have not determined the effects of singularities at higher order.

IV. SINGULARITY STRUCTURE OF THE EXPANSION

Our goal is a systematic analysis of the singularities of p_j to all orders in the amplitude equation (54). The detailed calculation of p_j rapidly becomes prohibitively laborious, but the recursion relations (57)–(59) determining the higher order coefficients in terms of lower order quantities (cf. Table I) can be analyzed to determine the properties of p_j. This study requires an accurate estimate of the pinching singularities in the integrals found in p_j. For this purpose we introduce an ''index'' which allows the divergence of a given integral to be assessed by a simple counting procedure. As in our discussion of the cubic coefficient, we assume the analytic singularities of F_0 do not affect the divergence behavior; the precise assumption on F_0 is given in (108) below.

A. Definition of the index

At every order in the expansion, the pinching singularities have a common structure. For $n>0$, define

$$D_n(\alpha,v)\equiv\frac{1}{(v-\alpha_1)(v-\alpha_2)\cdots(v-\alpha_n)}, \tag{104}$$

where $\alpha\equiv(\alpha_1,\ldots,\alpha_n)$ and also define $D_0(\alpha,v)\equiv 1$. Evaluating p_j for $j\geqslant 1$ involves integrands of the form

$$\mathcal{I}(v)=D_m(\beta,v)*D_n(\alpha,v)\sum_s{}'\ (\kappa^{(s)})^{m'}\frac{\partial^q\eta_l^{(s)}}{\partial v^q} \tag{105}$$

with $m+n\geqslant 1$, $m'\geqslant 1$ and $q\geqslant 0$. The poles in $\mathcal{G}(v)$ may be written as

$$\alpha_j = z_0 + i\gamma\nu_j/k, \quad j=1,\ldots,n, \tag{106}$$

$$\beta_j^* = z_0^* - i\gamma\zeta_j/k, \quad j=1,\ldots,m; \tag{107}$$

since they always lie along the vertical line $\mathrm{Re}(v)=v_p(\gamma)$. The specific pole locations are given by the numbers $\nu_j\geqslant 0$ and $\zeta_j\geqslant 0$ which vary depending on the integral. In all cases, these numbers are independent of F_0; in particular ν_j and ζ_j are independent of γ.

The complex-analytic singularities of F_0 are assumed to be unimportant for the amplitude expansions in the following sense. For any $n\geqslant 1$ and $q\geqslant 0$, we assume that

$$\lim_{\gamma\to 0^+}\left|\int_{-\infty}^{\infty} dv\, D_n(\alpha,v)\,\frac{\partial^q \eta_l^{(s)}}{\partial v^q}\right| < \infty \tag{108}$$

holds for each species and arbitrary choices of ν_j in (106). For example, it is sufficient for each distribution function $F_0^{(s)}(v)$ to be analytic on a neighborhood of $v_p(0)$, the linear phase velocity at zero growth rate; in this case the limit in (108) can be evaluated by the Plemej formulas for Cauchy integrals.[14]

The *index* of $\mathcal{G}(v)$ in (105) is defined by

$$\mathrm{Ind}[\mathcal{G}] \equiv m+n+q-1, \tag{109}$$

and $\mathrm{Ind}[\mathcal{G}]\geqslant 0$ since we assume $m+n\geqslant 1$. For example, for ψ in (20), the integrand of $\Gamma_{2,0}$ in (78) and the integrand of $\langle \partial_v\tilde{\psi}, \kappa\cdot\psi^*\rangle$ in (80), we obtain the indices

$$\mathrm{Ind}\left[\sum_s \kappa^{(s)}\psi^{(s)}\right]=0, \quad \mathrm{Ind}\left[\frac{\Sigma_s(\kappa^2\cdot\eta_k)^{(s)}}{(v-z_0)^3}\right]=2, \quad \mathrm{Ind}\left[\frac{\Sigma_s(\kappa^2\cdot\eta_k)^{(s)}}{(v-z_0)^2(v-z_0^*)}\right]=2. \tag{110}$$

If $mn=0$, our assumption (108) ensures that $\int dv\,\mathcal{G}(v)$ is nonsingular, but when $mn\neq 0$, then as $\gamma\to 0^+$, the integral of $\mathcal{G}$ develops a pinching singularity at the phase velocity $v_p(0)$. The index of $\mathcal{G}$ estimates the worst possible divergence of $\int dv\,\mathcal{G}$ due to this singularity.

Lemma IV.1: For $\mathcal{G}(v)$ *in* (105) *with* $mn\neq 0$, *the integral of* $\mathcal{G}$ *satisfies*

$$\lim_{\gamma\to 0^+}\gamma^J\left|\int_{-\infty}^{\infty} dv\,\mathcal{G}(v)\right| < \infty \tag{111}$$

with $J=\mathrm{Ind}[\mathcal{G}]$. *However, if*

$$\sum_s{}' \kappa^{(s)}[(-1)^{m'}+(\kappa^{(s)})^{m'-1}]\,\eta_k^{(s)}(v_p(0))=0, \tag{112}$$

then the integral of $\mathcal{G}$ *is less singular and* (111) *holds with* $J=\mathrm{Ind}[\mathcal{G}]-1$. *In* (112) *the prime on the summation indicates that the sum omits the electron species.*

Proof: By integrating by parts q times we can reduce to the case with $q=0$. With $q=0$, consider the simplest possibility $m+n=2$, and make a partial fraction expansion of the integrand

$$\int_{-\infty}^{\infty} dv\,\frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)(v-\beta_1^*)} = \frac{(-ik/\gamma)}{(2+\nu_1+\zeta_1)}\left[\int_{-\infty}^{\infty} dv\,\frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)} - \int_{-\infty}^{\infty} dv\,\frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\beta_1^*)}\right]. \tag{113}$$

The integrals on the right are nonsingular with limits at $\gamma=0$ given by the Plemej formula

$$\lim_{\gamma\to 0^+}\left[\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)}-\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\beta_1^*)}\right]=2\pi i\sum_s\ (\kappa^{(s)})^{m'}\eta_l^{(s)}(v_p(0)). \tag{114}$$

From the limit $\gamma\to 0^+$ of the dispersion relation $\Lambda_k(z_0)=0$, we obtain $\Sigma_s(\kappa^{(s)})\,\eta_k^{(s)}(v_p(0))=0$ for the imaginary part. Since $\kappa^{(e)}=-1$, this can be rewritten as $\eta_k^{(e)}=\Sigma_s'(\kappa^{(s)})\,\eta_k^{(s)}(v_p(0))$, and used to eliminate $\eta_k^{(e)}$ from (114):

$$\lim_{\gamma\to 0^+}\left[\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)}-\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\beta_1^*)}\right]$$
$$=2\pi i\left(\frac{k^2}{l^2}\right)\sum_s{}'\,\kappa^{(s)}[(-1)^{m'}+(\kappa^{(s)})^{m'-1}]\,\eta_k^{(s)}(v_p(0)). \tag{115}$$

This proves the lemma for $m+n=2$. It is important that the limit in (115) is independent of the parameters ν_1 and ζ_1 that locate the poles α_1 and β_1 for $\gamma>0$.

For integrals with $m+n>2$, by expanding the integrand in partial fractions, they can be reexpressed in terms of integrals with $m+n-1$ multiplied by a factor of γ^{-1}. A simple induction argument then establishes (111) for general $mn\neq 0$. (Recall that if $mn=0$ then the integrals are nonsingular.)

The general significance of the condition in (112) becomes clear from the form of the partial fraction expansion. For example, consider the expansion for $(m,n)=(2,1)$:

$$\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)(v-\alpha_2)(v-\beta_1^*)}=\frac{(-ik/\gamma)}{(2+\nu_1+\zeta_1)}\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_1)(v-\alpha_2)}$$
$$-\frac{(-ik/\gamma)^2}{(2+\nu_1+\zeta_1)(2+\nu_2+\zeta_1)}\left[\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_2)}\right.$$
$$\left.-\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\beta_1^*)}\right]. \tag{116}$$

The second term is the dominant singularity γ^{-2} and at $\gamma=0$ the bracketed integrals yield the same factor obtained previously in (115). When this factor is zero, the singularity of the $(m,n)=(2,1)$ integral is reduced from 2 to 1.

In the partial fraction expansion of the general case, the most singular terms that diverge according to (111) will always be proportional to the difference of two nonsingular integrals:

$$\left[\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\alpha_j)}-\int_{-\infty}^{\infty} dv\ \frac{\Sigma_s(\kappa^{(s)})^{m'}\eta_l^{(s)}}{(v-\beta_{j'}^*)}\right] \tag{117}$$

whose limit at $\gamma=0$ is given by (115). Thus when this limit is zero, the singularity of the term must drop by 1. This proves the second part of the lemma. □

The exceptional situation defined by the condition in (112) is a generalization of the feature noted previously in our analysis of the cubic coefficient. Setting $m'=3$ in (112) yields $c_1(0)=0$ from (100); the special circumstance that corresponded to a less singular integral at third order.

Our application of Lemma IV.1 to the recursion relations for $I_{m,j}$ requires generalization of the index to allow for sums of functions with well-defined indices and products of $\mathcal{G}$ with singular functions of γ. In each case, the generalized index is defined so that Eq. (111) remains true, i.e., the index for the composite function gives the *maximal* possible divergence of its integral.

First, if $\mathcal{G}_1(v)$ and $\mathcal{G}_2(v)$ have indices satisfying $\mathrm{Ind}[\mathcal{G}_1]\geqslant\mathrm{Ind}[\mathcal{G}_2]$ then we define the index of the sum to be the larger index:

$$\mathrm{Ind}[\mathscr{G}_1+\mathscr{G}_2]\equiv\mathrm{Ind}[\mathscr{G}_1]. \tag{118}$$

Clearly, $\mathrm{Ind}[\mathscr{G}_1]$ gives the maximal possible divergence of $\int dv(\mathscr{G}_1+\mathscr{G}_2)$. Second, if $q(\gamma)$ is a function of γ with the asymptotic behavior $q(\gamma)\sim\gamma^{-\nu}$ as $\gamma\to 0^+$, then we define the index of $q(\gamma)\mathscr{G}(v)$ to be

$$\mathrm{Ind}[q\mathscr{G}]\equiv\mathrm{Ind}[\mathscr{G}]+\nu. \tag{119}$$

Estimates of the form (111) still hold for $q\mathscr{G}$ with $J=\mathrm{Ind}[q\mathscr{G}]$. This completes the definition of the index.

By applying (118) and (119), the indices of $I_{m,j}$ and $h_{m,j}$ may be determined. For example, from (66) and (69) we have the indices

$$\mathrm{Ind}\left[\sum_s \kappa^{(s)}I_{0,0}^{(s)}\right]=1,\quad \mathrm{Ind}\left[\sum_s \kappa^{(s)}I_{2,0}^{(s)}\right]=1, \tag{120}$$

and from (77) and (79)

$$\mathrm{Ind}\left[\sum_s \kappa^{(s)}h_{0,0}^{(s)}\right]=2,\quad \mathrm{Ind}\left[\sum_s \kappa^{(s)}h_{2,0}^{(s)}\right]=2. \tag{121}$$

The index for $h_{2,0}^{(s)}$ is correct provided $\Gamma_{2,0}$ does not diverge more strongly than γ^{-2}. We can confirm this immediately from the index in (110); in fact, $\Gamma_{2,0}$ is nonsingular, as noted already in Sec. III C. In Sec. IV B, these indices are applied to determine the singularity of p_1. We stress again that when applied to a composite function $\mathscr{G}(v)$ the estimate in (111) does not necessarily determine the true singularity, but only an upper bound on the possible divergence of the integral.

There are several immediate consequences of the index definition worth stating. First, complex conjugation does not alter the index, $\mathrm{Ind}[\mathscr{G}]=\mathrm{Ind}[\mathscr{G}^*]$; second, if $G(v)$ has a well-defined index, then dividing G by $(v-\alpha)$ or $(v-\beta^*)$ simply increases the index of G by one:

$$\mathrm{Ind}[\mathscr{G}/(v-\alpha)]=\mathrm{Ind}[\mathscr{G}/(v-\beta^*)]=\mathrm{Ind}[\mathscr{G}]+1. \tag{122}$$

Here α,β are defined as in (106) and (107) and thereby increase the index of every term in G by one; this implies (122). Finally, from (105), differentiating $\mathscr{G}$ also raises the index by one, $\mathrm{Ind}[\partial_v\mathscr{G}]=1+\mathrm{Ind}[\mathscr{G}]$.

B. Analysis of p_1 using the index

The above relations along with recursion relations determine the indices of $I_{m,j}$ and $h_{m,j}$ to all orders. This information permits an estimate of the singularities of the coefficients p_j for $j\geqslant 1$. In practice, this must be done recursively following the organization of the amplitude expansions in Table I. The procedure is illustrated here for the first level of Table I; more general results on the index to all orders are provided in Secs. IV C and V.

Table I begins with ψ, $h_{0,0}$ and $h_{2,0}$ whose indices were evaluated in (110) and (121). From this information, we can estimate the integrals required to evaluate p_1 in (76). The integrand of $\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle$ has index 2, and the integrands of $\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle$ and $\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle$ have index 4. In each case, the factor $\partial_v\tilde{\psi}^{(s)*}$ increases the index by two from the values obtained in (110) and (121). Now Lemma IV.1 tells us that $\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle$ cannot diverge more strongly than γ^{-2}, and $\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle$ and $\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle$ cannot diverge more strongly than γ^{-4}.

These conclusions imply that the singularity of p_1 cannot be worse than γ^{-4}; this was found in Sec. III C by an explicit calculation of p_1. The explicit evaluation, of course, provided the important additional result that the γ^{-4} singularity actually occurs unless one of the exceptional cases corresponding to $c_1(0)=0$ is considered. The foregoing index analysis is easier but yields less complete information.

C. The main result

The main result on the singularity structure of the amplitude expansions can now be proved.

Theorem IV.1: *For $j \geqslant 1$, the coefficients in the expansion of the amplitude equation* (54) *satisfy*

$$\lim_{\gamma \to 0^+} \gamma^{5j-1} |p_j| < \infty. \tag{123}$$

Let $J_{m,j} \equiv (2m+5j-3)+4\delta_{m,0}+5\delta_{m,1}$, then for $j \geqslant 0$, $m \geqslant 0$ and $m' \geqslant 0$, the indices of $I^{(s)}_{m,j}$ and $h^{(s)}_{m,j}$ obey

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} I^{(s)}_{m,j}\right] \leqslant J_{m,j}. \tag{124}$$

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{m,j}\right] \leqslant J_{m,j}+1, \tag{125}$$

and the integrals in (72) *and* (75) *satisfy*

$$\lim_{\gamma \to 0^+} \gamma^{J_{m,j}+1} |\Gamma_{m,j}| < \infty, \tag{126}$$

$$\lim_{\gamma \to 0^+} \gamma^{J_{m,j}+3-\delta_{m,1}} |\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa} \cdot h_{m,j} \rangle| < \infty. \tag{127}$$

In addition, integrals over $I^{(s)}_{1,j}(v)$ satisfy

$$\lim_{\gamma \to 0^+} \gamma^{J_{1,j}+n-1} \left| \int_{-\infty}^{\infty} dv D_n(\alpha, v) \sum_s (\kappa^{(s)})^{m'} I^{(s)}_{1,j}(v) \right| < \infty \tag{128}$$

for $n \geqslant 1$ where $D_n(\alpha, v)$ is defined in (104) *with upper half-plane poles $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ of the form* (106) *but otherwise arbitrary.*

The relations in (124)–(128) are of interest chiefly for their utility in proving (123). We are able to prove (124) and (125) as upper bounds, but we expect them to hold as equalities in most cases. For $m=1$, the bounds in (126) and (127) give stronger estimates on the integrals than one would obtain simply from the index of the integrand calculated using (124) and (125). Similarly, the bounds in (128) are needed because they provide sharper estimates for the $m=1$ case than are available using the index information (124) and (125) alone.

Even though $h_{1,j}$ and $h^*_{1,j}$ have the same index, it is very important to recognize that the conclusion in (127) may not hold for $\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa} \cdot h^*_{1,j} \rangle$. The most singular terms in $\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa} \cdot h_{1,j} \rangle$ are not in general the most singular terms in $\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa} \cdot h^*_{1,j} \rangle$; for $\langle \partial_v \tilde{\psi}, \boldsymbol{\kappa} \cdot h^*_{1,j} \rangle$, if we wish to avoid explicitly evaluating the integral, our only means of estimating the divergence is by calculating the index of the integrand from (125).

Proof: (1) The proof is by induction using the organization of Table I. In Secs. IV A and IV B, the relations in (123)–(127) have been explicitly verified for $I_{0,0}$, $I_{2,0}$, $h_{0,0}$, $h_{2,0}$, $\Gamma_{0,0}$, $\Gamma_{2,0}$, and p_1. This proves the theorem for the first level of Table I, but does not establish the validity of (128) for the initial case $j=0$ which enters in the second level of Table I. By substituting the recursion relation (68) for $I_{1,0}$ into (128) we obtain

$$\gamma^{3+n}\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} I_{1,0}^{(s)}(v)$$

$$= \gamma^{3+n}\left(\frac{i}{k}\right)\int_{-\infty}^{\infty} dv D_n(\alpha,v)\partial_v\left(\kappa\cdot\left[h_{0,0}-h_{2,0}+\frac{1}{2}\psi^*\Gamma_{2,0}\right]\right)$$

$$+ \gamma^{3+n}\left(\frac{i}{k}\right)\left\langle \partial_v\tilde{\psi},\kappa\cdot\left[h_{0,0}-h_{2,0}+\frac{1}{2}\psi^*\Gamma_{2,0}\right]\right\rangle\int_{-\infty}^{\infty} dv D_n(\alpha,v)\psi(v). \quad (129)$$

On the right hand side, the index of the first integrand is $(n+3)$ so the factor of γ^{n+3} ensures a finite limit. In the second term, the integral over $D_n\psi$ is nonsingular since all poles are in the upper half-plane, but the coefficient $\langle\partial_v\tilde{\psi},\kappa\cdot[h_{0,0}-h_{2,0}+\frac{1}{2}\psi^*\Gamma_{2,0}]\rangle$ diverges like γ^{-4} (cf. Sec. III C). Thus for $n\geqslant 1$ the second term will also have a finite limit; this verifies (128) for $I_{1,0}$.

(2) The relations (123)–(128) are extended by induction to all coefficients $I_{m,j}$, $h_{m,j}$, and p_j using the recursion relations (57)–(59) and (66)–(71). Assume that (123)–(128) are valid down to some arbitrary level of Table I, and consider what the recursion relations imply for the coefficients evaluated at the next level. We use (m_1,j_1) to denote subscripts of coefficients such as I_{m_1,j_1}, h_{m_1,j_1} and p_{j_1} that are to be evaluated from lower order quantities assumed to satisfy (123)–(128).

The induction argument has three parts. First, we show that the properties (125)–(127) must hold for h_{m_1,j_1} and Γ_{m_1,j_1} if I_{m_1,j_1} satisfies two relations, namely the index identity (124)

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} I_{m_1,j_1}^{(s)}\right]\leqslant J_{m_1,j_1}, \quad (130)$$

and the $m_1=1$ estimate (128)

$$\lim_{\gamma\to 0^+}\gamma^{J_{1,j_1}+n-1}\left|\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} I_{1,j_1}^{(s)}(v)\right|<\infty. \quad (131)$$

In the second part of the proof, we verify that (130) and (131) do hold for I_{m_1,j_1}. Finally, in the third part of the proof, we verify that (125)–(127) imply (123) for p_{j_1}. It is important to note from Table I that the quantities h_{m_1,j_1} depend on p_j for $j<j_1$, but are independent of p_{j_1} so the reasoning is not circular.

(3) **(Part 1):** Suppose that (130) and (131) hold, then the three relations (125)–(127) are easily obtained as follows. First, consider (126) which holds trivially for $m_1=0$ since $\Gamma_{0,j_1}=0$. For $m_1\geqslant 2$, (126) follows from (130) and the identity (72). For $m_1=1$, we have to allow for the fact that $\Lambda_k(z_{1,j_1})\sim\gamma$ as $\gamma\to 0^+$, and apply (131) to estimate the integral in (72). Next (125) follows by applying (130) and (126) to the identity (73). Each term in (73) separately has a maximum index of $(J_{m_1,j_1}+1)$ so this must bound the index of the sum.

Finally consider (127). For $m_1\neq 1$, this relation follows from (125), and for $m_1=1$ we evaluate (127) using the identity in (73):

$$\gamma^{J_{1,j_1}+2}\langle\partial_v\tilde{\psi},\kappa\cdot h_{1,j_1}\rangle = -\gamma^{5j_1+6}\left(\frac{i}{k}\right)\langle\partial_v\tilde{\psi},\kappa\cdot I_{1,j_1}/(v-z_{1,j_1})\rangle$$

$$-\gamma^{5j_1+6}\Gamma_{1,j_1}\langle\partial_v\tilde{\psi},\kappa^2\cdot\eta_k/(v-z_{1,j_1})\rangle. \quad (132)$$

The first term on the right is nonsingular by virtue of (131), and the second term is also nonsingular by virtue of (126).

(4) **(Part 2):** Thus the crux of the matter is to verify (130) and (131) from the recursion relations for $I_{m_1,j_1}^{(s)}$, i.e., from (66)–(71). In this step, we assume (123)–(128) apply to the quantities appearing on the right hand side of these recursion relations, then the properties in (130) and (131) are verified term by term.

Verification of (130)*—the index of* I_{m_1,j_1}: The index of each term on the right hand side of the recursion relations for $I^{(s)}_{m_1,j_1}(v)$ can be evaluated by applying (123)–(127); this exercise shows that all of these terms have an index less than or equal to J_{m_1,j_1} which establishes (130) for $I^{(s)}_{m_1,j_1}$. A few examples from (68), the recursion relation for $(m_1,j_1)=(1,j_1)$, illustrate these index calculations; note that in these examples $J_{1,j_1}=5j_1+4$.

(a) The terms in (68) that depend on p_j have the form $[(2+n)p_{j_1-n}+(1+n)p^*_{j_1-n}]h^{(s)}_{1,n}$; from (125) we have

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{1,n}\right] \leqslant J_{1,n}+1=5n+5, \tag{133}$$

and from (123) the singularity of p_{j_1-n} is determined, hence

$$\mathrm{Ind}\left[p_{j_1-n}\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{1,n}\right] \leqslant 5(j_1-n)-1+(5n+5)=5j_1+4 \tag{134}$$

which is consistent with (130) for $(m_1,j_1)=(1,j_1)$.

(b) Next consider the term $\mathscr{P}_\perp \kappa\cdot\partial_v h_{0,j_1}$ in (68); from the definition (49) we have that the index of $\mathscr{P}_\perp \kappa\cdot\partial_v h_{0,j_1}$ is equal to the largest index obtained from the two terms $\kappa\cdot\partial_v h_{0,j_1}$ and $\psi(v)\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,j_1}\rangle$. These terms have individual indices

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} \partial_v h^{(s)}_{0,j_1}\right] \leqslant 5j_1+3, \tag{135}$$

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} \psi^{(s)}(v)\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,j_1}\rangle\right] \leqslant 5j_1+4, \tag{136}$$

where the first index (135) comes from (125) and the second follows from (127). Hence

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} (\mathscr{P}_\perp \kappa\cdot\partial_v h_{0,j_1})^{(s)}\right] \leqslant 5j_1+4 \tag{137}$$

which is consistent with (130) for $(m_1,j_1)=(1,j_1)$.

(c) A more subtle example is the term $\mathscr{P}_\perp \kappa\cdot\partial_v h^*_{1,n}\Gamma_{2,j_1-n-1}$ in (68) which expands to

$$\mathscr{P}_\perp \kappa\cdot\partial_v h^*_{1,n}\Gamma_{2,j_1-n-1}=\kappa\cdot\partial_v h^*_{1,n}\Gamma_{2,j_1-n-1}+\Gamma_{2,j_1-n-1}\psi(v)\langle\partial_v\tilde{\psi},\kappa\cdot h^*_{1,n}\rangle. \tag{138}$$

Now the index of each term can be evaluated to give

$$\mathrm{Ind}\left[\Gamma_{2,j_1-n-1}\sum_s (\kappa^{(s)})^{m'} \partial_v h^{(s)*}_{1,n}\right] \leqslant 5j_1+2, \tag{139}$$

$$\mathrm{Ind}\left[\Gamma_{2,j_1-n-1}\sum_s (\kappa^{(s)})^{m'} \psi^{(s)}\langle\partial_v\tilde{\psi},\kappa\cdot h^*_{1,n}\rangle\right] \leqslant 5j_1+4; \tag{140}$$

thus

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} (\mathscr{P}_\perp \kappa\cdot\partial_v h^*_{1,n})^{(s)}\Gamma_{2,j_1-n-1}\right] \leqslant 5j_1+4 \tag{141}$$

which is consistent with (130) for $(m_1,j_1)=(1,j_1)$. In this last example, (139) is obtained from (125) and (126); the second result (140) differs from (136) since (127) cannot be applied and the

singularity of $\langle \partial_v \tilde{\psi}, \kappa \cdot h^*_{1,n} \rangle$ must be estimated from the index in (125). In a similar way the index of each term appearing on the right in the recursion relations (66)–(71) for $I_{m_1,j_1}(v)$ can be estimated with the final result in (130).

(5) *Verification of* (131): We integrate the recursion relation for $I_{1,j_1}(v)$ and examine the resulting expression to establish (131). When (68), the recursion for I_{1,j_1}, is inserted into (131) we obtain

$$\begin{aligned}
&\gamma^{J_{1,j_1}+n-1}\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} I^{(s)}_{1,j_1}(v) \\
&\quad = -\gamma^{5j_1+n+3}\sum_{l=0}^{j_1-1}[(2+l)p_{j_1-l}+(1+l)p^*_{j_1-l}]\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{1,l}(v) \\
&\qquad + \gamma^{5j_1+n+3}\left(\frac{i}{k}\right)\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'+1} G^{(s)}(v) \\
&\qquad + \gamma^{5j_1+n+3}\left(\frac{i}{k}\right)\langle \partial_v \tilde{\psi}, \kappa \cdot G\rangle \int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} \psi^{(s)}(v),
\end{aligned} \tag{142}$$

where $G(v)=h_{0,j_1}-h_{2,j_1}+\cdots$ denotes the collection of terms enclosed by brackets in (68). The relations (124)–(127) hold for these terms by assumption, and it is straightforward to verify that $\mathrm{Ind}[G]\leqslant 5j_1+2$; this immediately implies that the terms in (142) involving G are finite as $\gamma\to 0^+$. For the remaining terms in (142) we have $p_{j_1-l}\sim\gamma^{-5(j_1-l)+1}$ from (123), hence we need to prove

$$\lim_{\gamma\to 0^+}\gamma^{5l+4+n}\left|\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{1,l}(v)\right| < \infty \quad (n\geqslant 1) \tag{143}$$

to establish (131). The formula in (125) bounds the index of the integrand in (143) by only $(5l+n+5)$ which is not sufficient. Instead, we evaluate (143) using the identity (73)

$$\begin{aligned}
\int_{-\infty}^{\infty} dv D_n(\alpha,v)\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{1,l}(v) &= -\left(\frac{i}{k}\right)\int_{-\infty}^{\infty} dv D_n(\alpha,v)\,\frac{\Sigma_s(\kappa^{(s)})^{m'} I^{(s)}_{1,l}(v)}{(v-z_{1,l})} \\
&\quad -\Gamma_{1,l}\int_{-\infty}^{\infty} dv D_n(\alpha,v)\,\frac{\Sigma_s(\kappa^{(s)})^{m'+1}\eta^{(s)}_k}{(v-z_{1,l})}.
\end{aligned} \tag{144}$$

The integral in the second term is nonsingular and $\Gamma_{1,l}\sim\gamma^{-5l-5}$ from (126), thus the desired estimate in (143) holds for the second term. It remains to verify

$$\lim_{\gamma\to 0^+}\gamma^{5l+4+n}\left|\int_{-\infty}^{\infty} dv D_n(\alpha,v)\,\frac{\Sigma_s(\kappa^{(s)})^{m'} I^{(s)}_{1,l}(v)}{(v-z_{1,l})}\right| < \infty \quad (n\geqslant 1) \tag{145}$$

for the first term in (144); this follows from (128) with $j\to l$ and $D_n\to D_n/(v-z_{1,l})$.
The proof of (143) establishes (131) for I_{1,j_1}.

(6) **(Part 3):** *Verification of* (123) *for* p_{j_1}: It is straightforward to check that application of (125)–(127) to the right hand side of the recursion relations (57)–(59) yields the estimate in (123). A few examples from (58) and (59) suffice to illustrate how this is carried out.

(a) Consider the term $\langle \partial_v \tilde{\psi}, \kappa \cdot h_{0,j_1-1}\rangle$ in (58); from (127) the singularity of this integral is at most

$$|\langle \partial_v \tilde{\psi}, \kappa \cdot h_{0,j_1-1}\rangle| \sim \left(\frac{1}{\gamma}\right)^{3+J_{0,j_1-1}} = \left(\frac{1}{\gamma}\right)^{5j_1-1} \tag{146}$$

which is consistent with (123).

(b) Other terms will involve products of integrals; for example, $\Gamma_{2,j_1-1}\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle$ in (58). Using (82) and (126) we determine that the singularity of this product is at most

$$|\Gamma_{2,j_1-1}\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle| \sim \left(\frac{1}{\gamma}\right)^{1+J_{2,j_1-1}} \left(\frac{1}{\gamma}\right)^2 = \left(\frac{1}{\gamma}\right)^{5j_1-1} \tag{147}$$

which again is consistent with (123).

(c) A final example is the product $\Gamma_{2,l}\langle \partial_v \tilde{\psi}, \kappa \cdot h^*_{1,j_1-l-2} \rangle$ with $0 \leqslant l \leqslant j_1 - 2$ from (59); applying (125) and (126) to this shows a maximum singularity of

$$|\Gamma_{2,l}\langle \partial_v \tilde{\psi}, \kappa \cdot h^*_{1,j_1-l-2} \rangle| \sim \left(\frac{1}{\gamma}\right)^{1+J_{2,l}} \left(\frac{1}{\gamma}\right)^{3+J_{1,j_1-l-2}} = \left(\frac{1}{\gamma}\right)^{5j_1-1} \tag{148}$$

which is consistent with (123).

This completes the proof. □

D. Implications of the main result

Theorem IV.1 allows us to evaluate the $\gamma \to 0^+$ limit of the theory. Rewriting the amplitude equation (54) using polar variables $A(t) = \rho(t)\exp(-i\theta(t))$, and the scaling

$$\rho(t) = \gamma^{5/2} r(\gamma t) \tag{149}$$

from Sec. III C, yields

$$\frac{dr}{d\tau} = r + \sum_{j=1}^{\infty} [\gamma^{5j-1} \operatorname{Re}(p_j)] r^{2j+1}, \tag{150}$$

$$\frac{d\theta}{dt} = \omega - \gamma \sum_{j=1}^{\infty} [\gamma^{5j-1} \operatorname{Im}(p_j)] r^{2j}, \tag{151}$$

where $\tau = \gamma t$. The result (123) on the divergence of the coefficients p_j ensures that the terms in (150) and (151) are now finite as $\gamma \to 0^+$. Introducing the asymptotic coefficients

$$c_j(0) = \lim_{\gamma \to 0^+} [\gamma^{5j-1} p_j] \quad (j \geqslant 1) \tag{152}$$

the amplitude equation for a weakly unstable wave becomes

$$\frac{dr}{d\tau} = r + \sum_{j=1}^{\infty} [\operatorname{Re}(c_j(0)) + \mathcal{O}(\gamma)] r^{2j+1}. \tag{153}$$

Unless $\operatorname{Re}(c_j(0)) = 0$ in each term, the scaling in (149) is required to obtain a finite theory. The circumstances that could force $c_1(0) = 0$ were discussed in Sec. III C, and will be analyzed more completely in Sec. V.

The electric field is obtained from Poisson's equation $(imk)E_{mk}(t) = \Sigma_s \int_{-\infty}^{\infty} dv\, f^{u\,(s)}_{mk}$ using the form of f^u_{mk} in (44)

$$E_{mk}(t) = \begin{cases} -iA(1 + \sigma\Gamma_1(\sigma))/k, & m = 1 \\ -iA^m \Gamma_m(\sigma)/mk, & m > 1. \end{cases} \tag{154}$$

The small growth rate behavior of $\Gamma_m(\sigma)$ can be inferred from (126) and the expansion $\Gamma_m(\sigma) = \Sigma_j \Gamma_{m,j} \sigma^j$. Using $\sigma^j = \gamma^{5j} r^{2j}$, we find an expansion of Γ_m with a nonsingular limit

$$\gamma^{2m-2+5\delta_{m,1}} \Gamma_m = \sum_{j=0}^{\infty} \gamma^{J_{m,j}+1} \Gamma_{m,j} r^{2j}, \tag{155}$$

and this motivates our definition of Γ^c_m

$$\Gamma_m^c = \lim_{\gamma\to 0^+} \gamma^{2m-2+5\delta_{m,1}}\Gamma_m = \sum_{j=0}^{\infty} b_{m,j}(0) r^{2j}, \tag{156}$$

where

$$b_{m,j}(0) \equiv \lim_{\gamma\to 0^+} [\gamma^{J_{m,j}+1}\Gamma_{m,j}]. \tag{157}$$

With this notation, the asymptotic form of the Fourier expansion $E(x,t)=\Sigma_m E_{mk}(t)\exp(imkx)$ of the field can be evaluated

$$E(x,t) = -\frac{i\gamma^{5/2}}{k}\Bigg[r(\tau)(1+\Gamma_1^c r(\tau)^2 + \mathcal{O}(\gamma)) e^{i(kx-\theta(t))} + \sum_{m=2}^{\infty} \gamma^{(m-1)/2} \frac{(\Gamma_m^c r(\tau)^m + \mathcal{O}(\gamma))}{m} e^{im(kx-\theta(t))} \Bigg] + \text{c.c.} \tag{158}$$

The asymptotic form may be rewritten more simply as

$$\frac{iE(x,t)}{\gamma^{5/2}} = \left[\frac{r(\tau)}{k}(1+\Gamma_1^c r(\tau)^2) e^{i(kx-\theta(t))} + \text{c.c.} \right] + \mathcal{O}(\sqrt{\gamma}); \tag{159}$$

this describes an electric field comprised of a single Fourier component whose coefficient scales like $\gamma^{5/2}$ and evolves nonlinearly according to the asymptotic amplitude equation (153). The $\gamma^{5/2}$ scaling yields a dramatically smaller field than would be expected from the trapping scaling γ^2 characteristic of a model with infinitely massive ions.

An analogous discussion of the asymptotic distribution function $f(x,v,t)$ is substantially more complicated because of singularities in v that emerge as $\gamma\to 0^+$. This issue is briefly treated in Sec. VI, but we defer a full analysis to a companion paper.[15]

V. SPECIAL CASES

Section III D identified three settings for which $c_1(0)=0$ and the γ^{-4} singularity of p_1 is replaced by a γ^{-3} divergence:

(a) Infinitely massive ions: $\kappa^{(s)}=0$ for all $s\neq e$,
(b) zero slope for the resonant ions: $\eta_k^{(s)}(v_p(0))=0$ for all $s\neq e$,
(c) an electron-positron plasma: $\kappa^{(p)^2}=1$ for positrons $(s=p)$.

From the partial fraction expansion in Lemma IV.1, we found the general condition required to reduce the singularity of a standard integral,

$$\int_{-\infty}^{\infty} dv\, \mathcal{G}(v) = \int_{-\infty}^{\infty} dv\, D_m(\beta,v)^* D_n(\alpha,v) \sum_s (\kappa^{(s)})^{m'} \frac{\partial^q \eta_l^{(s)}}{\partial v^q} \quad (mn>0) \tag{160}$$

from $\gamma^{-m-n-q+1}$ to $\gamma^{-m-n-q+2}$; namely,

$$\sum_s{}' \kappa^{(s)}[(-1)^{m'} + (\kappa^{(s)})^{m'-1}] \eta_k^{(s)}(v_p(0)) = 0, \tag{161}$$

where the primed sum omits the electrons. The specific instance of (161) that yields $c_1(0)=0$ corresponds to $m'=3$.

The validity of this general condition varies between the three settings listed above. For both (a) and (b), (161) holds quite generally for all integers $m'\geqslant 0$; however, for the third setting (c), the condition reduces to

$$[(-1)^{m'}+1]\eta_k^{(p)}(v_p(0)) = 0, \tag{162}$$

and this will be generally satisfied only for m' odd.

A. Cases (a) and (b)

It is relatively simple to adapt Theorem IV.1 to accommodate the first two settings where (161) holds uniformly. The following result restates Theorem IV.1 of (I) under more general assumptions [note that the definition of the index in this paper differs from (I)].[1]

Theorem V.1: *Assume that* (161) *holds for all* $m' \geqslant 0$; *then for* $j \geqslant 1$, *the coefficients in the expansion of the amplitude equation* (54) *satisfy*

$$\lim_{\gamma \to 0^+} \gamma^{4j-1} |p_j| < \infty. \tag{163}$$

Let $J_{m,j} \equiv (2m+5j-3)+4\delta_{m,0}+5\delta_{m,1}$, *then for* $j \geqslant 0$, $m \geqslant 0$ *and* $m' \geqslant 0$, *the indices of* $I^{(s)}_{m,j}$ *and* $h^{(s)}_{m,j}$ *obey*

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} I^{(s)}_{m,j}\right] \leqslant J_{m,j} - j - \delta_{m,1}. \tag{164}$$

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} h^{(s)}_{m,j}\right] \leqslant J_{m,j} + 1 - j - \delta_{m,1}, \tag{165}$$

and the integrals in (72) *and* (75) *satisfy*

$$\lim_{\gamma \to 0^+} \gamma^{J_{m,j}-j} |\Gamma_{m,j}| < \infty, \tag{166}$$

$$\lim_{\gamma \to 0^+} \gamma^{J_{m,j}+2-j-\delta_{m,1}} |\langle \partial_v \tilde{\psi}, \kappa \cdot h_{m,j} \rangle| < \infty. \tag{167}$$

The proof is simpler since (165)–(167) follow directly from (164) and there is no need for special estimates such as (128) in the $m=1$ case.

Proof: (1) The induction proof follows that for Theorem IV.1, and we only highlight the new features brought in by (161). The recursion relations for $I_{m,j}$ are finite sums; each term of which has the form $q(\gamma)\mathcal{G}(v)$ where $\mathrm{Ind}[\mathcal{G}(v)] \geqslant 1$ and the coefficient $q(\gamma)$ may be singular as $\gamma \to 0^+$. According to Lemma IV.1, with (161) in force, if $J = \mathrm{Ind}[q(\gamma)\mathcal{G}(v)]$, then $\gamma^{J-1} q(\gamma) \int dv\, \mathcal{G}(v)$ remains finite as $\gamma \to 0^+$.

(2) We have already shown that (161) implies (163) for $j=1$ so the theorem has been proved for the first level of Table I. As before, we assume that (163)–(167) are valid down to some arbitrary level of Table I, and consider what the recursion relations imply for the coefficients evaluated at the next level. Again, we use (m_1, j_1) to denote subscripts of coefficients such as I_{m_1,j_1}, h_{m_1,j_1} and p_{j_1} that are to be evaluated from lower order quantities assumed to satisfy (163)–(167). Following the organization of the previous proof, we first check that (164) for $I^{(s)}_{m_1,j_1}$ implies (165)–(167) for $h^{(s)}_{m_1,j_1}$, Γ_{m_1,j_1}, and $\langle \partial_v \tilde{\psi}, \kappa \cdot h_{m_1,j_1} \rangle$, respectively. Then (164) is verified from the recursion relation for $I^{(s)}_{m_1,j_1}$, and finally (163) is shown to follow from (165)–(167).

(3) **(Part 1):** Assume that (164) holds for $I^{(s)}_{m_1,j_1}$:

$$\mathrm{Ind}\left[\sum_s (\kappa^{(s)})^{m'} I^{(s)}_{m_1,j_1}\right] \leqslant J_{m_1,j_1} - j_1 - \delta_{m_1,1}, \tag{168}$$

then the properties (165)–(167) are easily obtained as follows. First, the integrand of Γ_{m_1,j_1} in (72) has index

$$\mathrm{Ind}\left[\sum_s I^{(s)}_{m_1,j_1} \Big/ (v - z_{m,j})\right] \leqslant J_{m_1,j_1} - j_1 - \delta_{m_1,1} + 1 \tag{169}$$

from (164) so the expression

$$\gamma^{J_{m_1,j_1}-j_1-\delta_{m_1,1}}\int_{-\infty}^{\infty} dv\,\frac{\Sigma_s I_{m,j}^{(s)}(v)}{v-z_{m,j}} \tag{170}$$

remains finite as $\gamma\to 0^+$; this proves (166) since the $\delta_{m_1,1}$ in the exponent compensates for the added singularity of $\Lambda_k(z_{1,j_1})$ in (72). Next, (165) follows by applying (168) and (166) to the identity (73). Finally, the integrand in (167) has index

$$\mathrm{Ind}\left[\sum_s \partial_v\tilde{\psi}^{(s)}\kappa^{(s)}h_{m,j}^{(s)}\right]\leqslant J_{m_1,j_1}-j_1-\delta_{m_1,1}+3 \tag{171}$$

from (165) so the bound in (167) follows from (161) and Lemma IV.1.

(4) **(Part 2):** *Verification of* (168)*—the index of* I_{m_1,j_1}: The consequences of (161) for the index calculation are incorporated in (163)–(166). The index of each term on the right hand side of the recursion relations for $I_{m_1,j_1}^{(s)}(v)$ can be evaluated by applying these conclusions regarding the lower order coefficients; this is done exactly as in the proof of (130) in Theorem IV.1. The bound in (168) follows as before.

(5) **(Part 3):** *Verification of* (163) *for* p_{j_1}: It is straightforward to check that application of (165)–(167) to the right hand side of the recursion relations (57)–(59) yields the estimate in (163) just as in the proof of Theorem IV.1. □

In light of the discussion in Sec. IV D, the theorem tells us that the scaling in (149) is replaced by

$$\rho(t)=\gamma^2 r(\gamma t) \tag{172}$$

for the special cases (a) and (b). This modification in turn implies an electric field characterized by the ''trapping scaling'' $E_k\sim\gamma^2$; the discussion in (I) provides detailed expressions for the asymptotic form of $E(x,t)$ and $f^{(e)}(x,v,t)$ in the limit $\gamma\to 0^+$ when the ions are fixed.[1]

B. An electron-positron plasma

The result $p_j\sim\gamma^{-4j+1}$ for the two special cases just considered is not correct for the electron-positron system with $\kappa^{(p)}=-\kappa^{(e)}=1$. Although $p_1\sim\gamma^{-3}$ does hold, we find at fifth order $p_2\sim\gamma^{-8}$ instead of $p_2\sim\gamma^{-7}$. Thus the trapping scaling of the amplitude in (172) is not sufficient to obtain a finite expansion. We summarize the evaluation of p_2 from (57)–(59), but leave the comprehensive analysis of this case as an open problem.

The exact recursion relation for p_2 is

$$\begin{aligned}-ikp_2=&-\langle\partial_v\tilde{\psi},\kappa\cdot(h_{0,1}-h_{2,1})\rangle-\tfrac{1}{2}\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle\Gamma_{2,1}-\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle\Gamma_{1,0}+\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle\Gamma_{1,0}^*\\&-\left\{\frac{1}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_{1,0}^*\rangle\Gamma_{2,0}+\frac{\Gamma_{3,0}}{3}\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}^*\rangle-\frac{\Gamma_{2,0}^*}{2}\langle\partial_v\tilde{\psi},\kappa\cdot h_{3,0}\rangle\right\},\end{aligned} \tag{173}$$

and, in the bracketed terms, $\Gamma_{2,0}$ and $\Gamma_{3,0}$ are nonsingular so Theorem IV.1 ensures the divergence of these terms cannot exceed γ^{-4}; hence they are subdominant:

$$\begin{aligned}-ikp_2=&-\langle\partial_v\tilde{\psi},\kappa\cdot(h_{0,1}-h_{2,1})\rangle-\tfrac{1}{2}\langle\partial_v\tilde{\psi},\kappa\cdot\psi^*\rangle\Gamma_{2,1}-\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle\Gamma_{1,0}\\&+\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle\Gamma_{1,0}^*+\mathcal{O}(\gamma^{-4}).\end{aligned} \tag{174}$$

Here consider the two terms involving $\Gamma_{1,0}$. In the general case, Theorem IV.1 gives $\Gamma_{1,0}\sim\gamma^{-5}$, but this drops to γ^{-4} when we take (161) into account for m' odd. The integral $\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle$ is nonsingular so $\langle\partial_v\tilde{\psi},\kappa\cdot h_{2,0}\rangle\Gamma_{1,0}^*\sim\gamma^{-4}$ and we neglect this term also. Next $\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle$ is written out to find

$$\langle\partial_v\tilde{\psi},\kappa\cdot h_{0,0}\rangle=\frac{i}{2k\gamma}\langle\partial_v\tilde{\psi},\kappa^2\cdot\partial_v(\psi^*-\psi)\rangle; \tag{175}$$

this integrand has index equal to 3, but contains an odd power of κ so (161) holds and the divergence cannot exceed γ^{-2}. Overall, $\langle \partial_v \tilde{\psi}, \kappa \cdot h_{0,0} \rangle \Gamma_{1,0}$ has a maximum singularity of γ^{-7}. A careful examination of the first term in (174) yields the same estimate $\langle \partial_v \tilde{\psi}, \kappa \cdot (h_{0,1} - h_{2,1}) \rangle \sim \gamma^{-7}$.

The dominant singularity in p_2 arises from the remaining term

$$p_2 = -\frac{i\Gamma_{2,1}}{2k} \langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle + \mathcal{O}(\gamma^{-7}), \tag{176}$$

and from (85) and (115), we obtain

$$\langle \partial_v \tilde{\psi}, \kappa \cdot \psi^* \rangle = -\left(\frac{k}{\gamma}\right)^2 \left[\frac{i\pi \eta_k^{(p)}(v_p(0))}{\Lambda_k'(v_p + i0)} + \mathcal{O}(\gamma) \right]. \tag{177}$$

The evaluation of $\Gamma_{2,1}$ starts from (72)

$$\Gamma_{2,1} = \frac{-i/2k}{\Lambda_{2k}(z_{2,1})} \sum_s \int_{-\infty}^{\infty} dv \, \frac{I_{2,1}^{(s)}(v)}{v - z_{2,1}}, \tag{178}$$

and the recursion relation (70) for $I_{2,1}$

$$I_{2,1}(v) = -2p_1 h_{2,0} + \frac{i}{k} \frac{\partial}{\partial v} \kappa \cdot \left\{ h_{1,0} + \psi \Gamma_{1,0} - h_{3,0} + \frac{1}{3} \psi^* \Gamma_{3,0} + \frac{1}{2} h_{0,0} \Gamma_{2,0} \right\}. \tag{179}$$

Note that $\Lambda_{2k}(z_{2,1}) = 3/4 + \mathcal{O}(\gamma)$ from (21) so any singularities arise from the integral

$$\int_{-\infty}^{\infty} dv \, \frac{\Sigma_s I_{2,1}^{(s)}(v)}{v - z_{2,1}} = -2p_1 \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s h_{2,0}^{(s)}(v)}{v - z_{2,1}} + \frac{i}{k} \int_{-\infty}^{\infty} \frac{dv}{(v - z_{2,1})^2} \sum_s \kappa^{(s)} \left[h_{1,0}^{(s)} + \psi^{(s)} \Gamma_{1,0} - h_{3,0}^{(s)} + \frac{1}{3} \psi^{(s)*} \Gamma_{3,0} + \frac{1}{2} h_{0,0}^{(s)} \Gamma_{2,0} \right]. \tag{180}$$

The strongest singularity here occurs in the integral over $h_{1,0}^{(s)}$, and all other terms in (180) are subdominant; thus we can focus attention on the expression

$$\int_{-\infty}^{\infty} dv \, \frac{\Sigma_s I_{2,1}^{(s)}(v)}{v - z_{2,1}} = \frac{i}{k} \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s \kappa^{(s)} h_{1,0}^{(s)}}{(v - z_{2,1})^2} + \mathcal{O}(\gamma^{-4}). \tag{181}$$

When $h_{1,0}^{(s)}$ is evaluated using (73) we obtain

$$\int_{-\infty}^{\infty} dv \, \frac{\Sigma_s I_{2,1}^{(s)}(v)}{v - z_{2,1}} = \frac{1}{k^2} \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s \kappa^{(s)} I_{1,0}^{(s)}}{(v - z_{2,1})^2 (v - z_{1,0})} + \mathcal{O}(\gamma^{-4}). \tag{182}$$

Now the recursion relation (68) for $I_{1,0}^{(s)}$ is applied, and examination of the various terms shows that the dominant singularity arises from the occurrence of ψ^* in $h_{0,0}$ so that (178) can be reduced to

$$\Gamma_{2,1} = \left(\frac{i[1 + \mathcal{O}(\gamma)]}{3\gamma k^5} \right) \int_{-\infty}^{\infty} dv \, \frac{\Sigma_s (\kappa^{(s)})^3 \partial_v^2 \psi^{(s)}(v)^*}{(v - z_{2,1})^2 (v - z_{1,0})} + \mathcal{O}(\gamma^{-5}). \tag{183}$$

Finally a partial fraction expansion of the remaining integral yields a γ^{-5} singularity

$$\int_{-\infty}^{\infty} dv \, \frac{\Sigma_s (\kappa^{(s)})^3 \partial_v^2 \psi^{(s)}(v)^*}{(v - z_{2,1})^2 (v - z_{1,0})} = \frac{1}{\gamma^5} \left[-\left(\frac{\pi}{8}\right) k^5 \eta_k^{(p)}(v_p(0)) + \mathcal{O}(\gamma) \right], \tag{184}$$

and leads to the asymptotic form of p_2 from (176)

TABLE II. Summary of the exponent β obtained from the amplitude equation. The cases (a), (b) and (c) are the exceptional cases discussed in Sec. V.

System	β	Comment
$c_1(0)\neq 0$	$\frac{5}{2}$	generic result for multiple mobile species
(a) $c_1(0)=0$	2	infinitely massive ions
(b) $c_1(0)=0$	2	$\eta_k^{(s)}(v_p)=0$ for each ion species
(c) $c_1(0)=0$	$\frac{9}{4}\leqslant\beta\leqslant\frac{5}{2}$	an electron-positron plasma

$$p_2=\frac{1}{\gamma^8}\left[\left(\frac{i\pi^2 k}{48}\right)\frac{\eta_k^{(p)}(v_p)^2}{\Lambda_k'(v_p+i0)}+\mathcal{O}(\gamma)\right]+\mathcal{O}(\gamma^{-7}). \tag{185}$$

Since $\mathrm{Im}\,\Lambda_k'(v_p+i0)$ is typically nonzero, this shows that $\mathrm{Re}(p_2)\sim\gamma^{-8}$ and implies that the trapping scaling (172) leaves a residual singularity at fifth order. With $\rho(t)=\gamma^\beta r(\gamma t)$ the rescaled equation becomes

$$\frac{dr}{d\tau}=r+\gamma^{2\beta-1}\,\mathrm{Re}(p_1)r^3+\gamma^{4\beta-1}\,\mathrm{Re}(p_2)r^5+\cdots \tag{186}$$

so $\beta\geqslant 9/4$ is required to overcome the γ^{-8} divergence in p_2.

VI. DISCUSSION

Our analysis of an unstable electrostatic wave expresses the Vlasov distribution function $f(x,v,t)$ in terms of the wave amplitude, i.e.,

$$f(x,v,t)=[A(t)e^{ikx}\psi(v)+\mathrm{c.c.}]+\sum_{m=-\infty}^{\infty}e^{imkx}H_{mk}(v,A,A^*), \tag{187}$$

where $A(t)$ evolves according to an amplitude equation,

$$\dot{A}(t)=Ap(|A^2|). \tag{188}$$

In the small growth rate limit, the eigenvalue of the unstable mode approaches the imaginary axis and, at $\gamma=0$, merges with a neutrally stable continuous spectrum on the axis.[1,10] This ''interaction'' between the unstable mode and the neutrally stable continuum leads to singular features in the asymptotic form of the amplitude equation (188).[16]

These singularities are physically significant since they reflect the very strong nonlinear interaction between the wave and the resonant particles at the linear phase velocity. Remarkably, the quantitative form of the singularities can be consistently interpreted as fixing overall scalings of the nonlinear solution with the linear growth rate in the asymptotic limit $\gamma\to 0^+$. This interpretation is achieved through a singular transformation $|A(t)|=\gamma^\beta r(\gamma t)$ of the mode amplitude in which the exponent β is fixed by the criterion that the rescaled dynamics,

$$\frac{dr}{d\tau}=r(\tau)\frac{\mathrm{Re}(p(\gamma^{2\beta}r(\tau)^2))}{\gamma}, \tag{189}$$

is free of singularities as $\gamma\to 0^+$. Our conclusions are summarized in Table II; with the notable exception of the electron-positron system, we can rigorously establish an exact value of $\beta=5/2$ in the typical instability and $\beta=2$ for certain exceptional models. In the electron-positron case, we have fixed only a range of possible β values from the explicit calculation of $p(|A|^2)$ through fourth order in $|A|$; the complexity of the recursion relations in this case prevents the iterative determination of the dominant singularities at each order.

The shift from $\beta=2$ to $\beta=5/2$ as the true asymptotic scaling due to finite mass ions at the phase velocity of the wave is our most significant conclusion regarding the physical character of the nonlinear wave. Since previous studies of the asymptotic scaling of single wave instabilities

have focused primarily on plasma waves and assumed fixed ions, this shift was not anticipated and was discovered purely as a consequence of the singularities in the amplitude equation (188). Numerical simulations of unstable ion-acoustic waves or unstable plasma waves in systems with finite mass ratios would be of considerable interest in this connection. In particular, for plasma waves one expects a crossover in the asymptotic scaling from the trapping scaling ($\beta=2$) to $\beta=5/2$ at sufficiently small growth rates. Predictions of this crossover have been made elsewhere based on the competition between the γ^{-4} and γ^{-3} singularities in the cubic coefficient.[7]

In this paper we have not analyzed the asymptotic form of $f(x,v,t)$, but it is clear from the expansions for $H(x,v,A,A^*)$ that this is an interesting question. For example, consider the leading coefficient in $h_0(v)$ from (77),

$$h_{0,0}(v)=-\frac{1}{k^2}\frac{\partial}{\partial v}\left[\frac{\kappa^2\cdot\eta_k}{(v-z_0)(v-z_0^*)}\right]; \tag{190}$$

there are poles in the complex-velocity plane at z_0 and z_0^* that approach the real velocity axis at v_p in the limit $\gamma\to 0^+$. Thus the coefficient becomes a singular function at the phase velocity when $\gamma=0$; on the other hand, for $v\neq v_p$, the coefficient is well-behaved at small growth rates. In the distribution function (187) this particular coefficient is multiplied by $|A|^2$ so that after rescaling the amplitude we have a term of the form $r^2\gamma^{2\beta}h_{0,0}(v)$. Now the emerging singularity in velocity space is weighted by a factor of $\gamma^{2\beta}$, and one would like to understand whether the resulting product is finite or singular at $\gamma=0$. A detailed analysis of this problem will be presented in a forthcoming paper, but the primary conclusion can be simply summarized.[15] Inside a neighborhood of width $\mathcal{O}(\gamma)$, and centered on $v=v_p$, the singularities in the coefficients of H are balanced by the factors of γ introduced by the rescaling of $|A|$ and one can obtain finite, nonzero limits for the weighted coefficients. Outside this small neighborhood of the phase velocity, the factors of γ force the weighted coefficients to vanish at $\gamma=0$.

The connection between coefficient singularities and nonlinear scalings that is studied in this paper is relevant in other settings. As mentioned above, a key feature of our instabilities is the emergence of the critical eigenvalues from a neutrally stable continuous spectrum. It has been appreciated recently that this characteristic is shared by a diverse set of additional examples which include the onset of linear instability in fluid shear flows,[17–20] instabilities of solitary waves,[21–23] synchronization of large systems of coupled oscillators,[24] and bifurcations in ''mean field'' descriptions of the dynamics of bubble clouds in fluids.[25] In the phase-dynamical models of coupled oscillators, the amplitude equations of the weakly unstable modes have been analyzed in the $\gamma\to 0^+$ limit.[26,27] In these studies, the coefficients can be singular or nonsingular depending on how the oscillators are coupled, but when singularities occur in the limit of small growth they can be absorbed by singular rescalings of the amplitude in a procedure analogous to the method used in this paper. The specific values of the scaling exponents are different, but yield predictions for the emergence of the nonlinear synchronized state.

It seems likely that the analysis of singular amplitude equations provides a general means to determine scaling exponents for weakly unstable systems. The exponents in turn provide a quantitative measure of the strong nonlinear interaction between the neutrally stable modes and the weakly unstable modes.

ACKNOWLEDGMENT

This work was supported by NSF Grant No. PHY-9423583.

[1] J. D. Crawford, ''Amplitude equations for electrostatic waves: universal singular behavior in the limit of weak instability,'' Phys. Plasmas **2**, 97–128 (1995).
[2] E. Frieman, S. Bodner, and P. Rutherford, ''Some new results on the quasi-linear theory of plasma instabilities,'' Phys. Fluids **6**, 1298–1305 (1963).
[3] A. Simon and M. Rosenbluth, ''Single-mode saturation of the bump-on-tail instability: immobile ions,'' Phys. Fluids **19**, 1567–1580 (1976).
[4] A. Simon, ''Single mode saturation of the bump-on-tail instability: II. Mobile ions,'' Phys. Fluids **20**, 79–82 (1977).
[5] P. Janssen and J. Rasmussen, ''Limit cycle behavior of the bump-on-tail instability,'' Phys. Fluids **24**, 268–273 (1981).
[6] C. Burnap, M. Miklavcic, B. Willis, and P. Zweifel, ''Single-mode saturation of a linearly unstable plasma,'' Phys. Fluids **28**, 110–115 (1985).

[7] J. D. Crawford and A. Jayaraman, ''Nonlinear saturation of an electrostatic wave: mobile ions modify trapping scaling,'' Phys. Rev. Lett. **77**, 3549–3552 (1996).
[8] N. G. van Kampen, ''On the theory of stationary waves in plasmas,'' Physica (Amsterdam) **21**, 949–963 (1955); see also N. G. van Kampen and B. U. Felderhof, *Theoretical Methods in Plasma Physics* (North-Holland, Amsterdam, 1967).
[9] K. Case, ''Plasma oscillations,'' Ann. Phys. (N.Y.) **7**, 349–364 (1959); also ''Plasma oscillations,'' Phys. Fluids **21**, 249–257 (1978).
[10] J. D. Crawford and P. Hislop, ''Application of the method of spectral deformation to the Vlasov-Poisson model,'' Ann. Phys. (N.Y.) **189**, 265–317 (1989).
[11] S. W. H. Cowley, ''Growing plasma oscillations for symmetrical douple-humped velocity distributions,'' J. Plasma Phys. **4**, 297–300 (1970).
[12] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory: Vol. II*, Appl. Math. Sci. 69 (Springer-Verlag, New York, 1988).
[13] P. Hislop and J. D. Crawford, ''Application of spectral deformation to the Vlasov-Poisson system II: mathematical results,'' J. Math. Phys. **30**, 2819–2837 (1989).
[14] N. I. Muskhelishvili, *Singular Integral Equations* (Dover, New York, 1992), pp. 42 and 43.
[15] J. D. Crawford and A. Jayaraman, ''First principles justification of a 'single wave model' for electrostatic instabilities,'' Phys. Plasmas (submitted), available at the LANL preprint archive as physics/9804014.
[16] J. D. Crawford, in ''Modern Mathematical Methods in Transport Theory'' in *Operator Theory: Advances and Applications*, edited by W. Greenberg and J. Polewczak (Birkhauser Verlag, Basel, 1991), Vol. 51, pp. 97–108.
[17] K. Case, ''Stability of inviscid plane Couette flow,'' Phys. Fluids **3**, 143–148 (1960).
[18] R. J. Briggs, J. D. Daugherty, and R. H. Levy, ''Role of Landau damping in crossed-field electron beams and inviscid shear flow,'' Phys. Fluids **13**, 421–432 (1970).
[19] S. A. Maslowe, ''Critical layers in shear flows,'' Annu. Rev. Fluid Mech. **18**, 405–432 (1986).
[20] N. J. Balmforth, D. Del-Castillo-Negrete, and W. R. Young, ''Dynamics of vorticity defects in shear,'' J. Fluid Mech. **333**, 197–230 (1997).
[21] R. Pego and M. I. Weinstein, ''Eigenvalues and instabilities of solitary waves,'' Philos. Trans. R. Soc. London, Ser. A **340**, 47–94 (1992).
[22] R. Pego and M. I. Weinstein, ''Evans function, Melnikov's integral, and solitary wave instabilities,'' in *Differential Equations with Applications to Mathematical Physics*, edited by W. F. Ames, E. M. Harrel, and J. V. Herod (Academic, Orlando, 1993).
[23] R. Pego, P. Smereka, and M. I. Weinstein, ''Oscillatory instability of solitary waves in a continuum model of lattice vibrations,'' Nonlinearity **8**, 921–941 (1995).
[24] S. Strogatz and R. Mirollo, ''Stability of incoherence in a population of coupled oscillators,'' J. Stat. Phys. **63**, 613–635 (1991).
[25] G. Russo and P. Smereka, ''Kinetic theory for bubbly flow I: collisionless case,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **56**, 327–357 (1996).
[26] J. D. Crawford, ''Amplitude expansions for instabilities in populations of globally-coupled oscillators,'' J. Stat. Phys. **74**, 1047–1084 (1994).
[27] J. D. Crawford, ''Scaling and singularities in the entrainment of globally coupled oscillators,'' Phys. Rev. Lett. **74**, 4341–4344 (1995).

Multivector fields and connections: Setting Lagrangian equations in field theories

Arturo Echeverría-Enríquez, Miguel C. Muñoz-Lecanda,[a)]
and Narciso Román-Roy[b)]
Departamento de Matemática Aplicada y Telemática, Edificio C-3, Campus Norte UPC, C/Jordi Girona 1, E-08034, Barcelona, Spain

(Received 8 July 1997; accepted for publication 14 May 1998)

The integrability of multivector fields in a differentiable manifold is studied. Then, given a jet bundle $J^1E \to E \to M$, it is shown that integrable multivector fields in E are equivalent to integrable connections in the bundle $E \to M$ (that is, integrable jet fields in J^1E). This result is applied to the particular case of multivector fields in the manifold J^1E and connections in the bundle $J^1E \to M$ (that is, jet fields in the repeated jet bundle J^1J^1E), in order to characterize integrable multivector fields and connections whose integral manifolds are canonical lifting of sections. These results allow us to set the Lagrangian evolution equations for first-order classical field theories in three equivalent geometrical ways (in a form similar to that in which the Lagrangian dynamical equations of nonautonomous mechanical systems are usually given). Then, using multivector fields, we discuss several aspects of these evolution equations (both for the regular and singular cases); namely, the existence and nonuniqueness of solutions, the integrability problem and Noether's theorem, giving insights into the differences between mechanics and field theories.
© *1998 American Institute of Physics.* [S0022-2488(98)04009-2]

I. INTRODUCTION

The theory of *multisymplectic manifolds* (that is, differentiable manifolds endowed with a closed nondegenerate k-form, with $k \geqslant 2$) has been revealed as a powerful tool for geometrically describing some problems in physics. In particular, jet bundles and their duals, equipped with specific multisymplectic forms, are the suitable geometrical frameworks for describing the Lagrangian and Hamiltonian formalisms of first-order classical field theories (a nonexhaustive list of references is Refs. 1–9).

As a consequence of this fact, the study of multisymplectic manifolds and their properties has increased lately;[10–12] in particular, those concerning the behavior of *multisymplectic Hamiltonian systems*, which are the generalization of the corresponding symplectic case. This generalization requires the use of *multivector fields* as a fundamental tool, and their contraction with differential forms, which is the intrinsic formulation of the systems of partial differential equations locally describing the field. Hence, the study of the integrability of such equations (that is, of the corresponding multivector fields), is of considerable interest and constitutes the first aim in this paper.

A particular situation of special relevance arises when multivector fields in fiber bundles are considered. In these cases, a result on the equivalence of integrable multivector fields and jet fields, or connections in the corresponding first-order jet bundles, can be stated.

Furthermore, in the jet bundle description of classical field theories, the evolution equations are usually obtained using the multisymplectic form in order to characterize the critical sections which are solutions of the problem. Nevertheless, in mechanics we can write the evolution equations in a more geometric–algebraic manner using vector fields, and then obtain the critical sections as integral curves of these vector fields. Different attempts have been made to achieve the same goal with the evolution equations of field theories. So, for the Lagrangian formalism, this is done in two different ways: by using *Ehresmann connections* in a jet bundle[13,14] or, what is

[a)]Electronic mail: MATMCML@MAT.UPC.ES
[b)]Electronic mail: MATNRR@MAT.UPC.ES

 © 1998 American Institute of Physics

equivalent, their associated *jet fields*.[4] Moreover, in Refs. 15–18, the evolution equations in the Hamiltonian formalism and generalized Poisson brackets are stated (mainly in local coordinate terms) using multivector fields (although, as far as we know, the first use of multivector fields in the realm of field theories can be found in Ref. 6). A further goal of this work is to carry out a deeper and pure geometrical analysis of these procedures for the Lagrangian formalism, showing that all these ways are, in fact, equivalent.

Thus, the aims of this paper are the following.

(1) To study the integrability of multivector fields in differentiable manifolds in general.
(2) To set the equivalence between jet fields in the jet bundle J^1E and a certain kind of multivector fields in E. This equivalence will in fact be used for jet fields in the ''repeated'' jet bundle J^1J^1E and the corresponding multivector fields in the jet bundle J^1E.
(3) To use these results to state the Lagrangian evolution equations for first-order classical field theories using multivector fields; and to prove this formalism is equivalent to the above mentioned ways of setting these equations.
(4) To use, in particular, this *multivector field formalism* to study some features of Lagrangian field theories; namely, the existence and nonuniqueness of solutions, the problem of the integrability and Noether's theorem. In this way the differences between mechanics and field theories will be made evident.

The structure of the paper is the following.

The first part is devoted to an analysis of the first two items. In particular, the following.

In Sec. II A we introduce the notion and characterization of *integrable multivector fields* in a manifold, as well as several properties for them.

Section II B is devoted to the presentation of an algorithm for finding the submanifold (if it exists) where a nonintegrable multivector field is integrable. This algorithm is applied to solve a particular system of partial differential equations in Sec. II C.

In Secs. II D and II E, we study the integrability of jet fields, connections and special kinds of multivector fields in jet bundles; and their equivalence. This equivalence is considered for the particular case of jet fields in J^1J^1E, and the suitable multivector fields in J^1E.

In Sec. II F, after giving different geometrical characterizations for *semi-holonomic* and *holonomic* multivector fields, they are finally related with semi-holonomic and holonomic jet fields.

In the second part, the applications to Lagrangian first-order field theories are considered. Thus we have the following.

Section III A is a short background on Lagrangian first-order classical field theory.

Section III B is devoted to establishing the evolution equations for classical field theories (the *Euler–Lagrange equations*) using multivector fields in J^1E, showing that this is equivalent to using jet fields in J^1J^1E or their associated connections in $J^1E\to M$.

In Secs. III C and III D, the analysis of the evolution equations is made for regular and singular Lagrangian field theories, using multivector fields. In particular, we show that the existence of integrable multivector fields solution of the equations is not guaranteed (even in the regular case). The nonuniqueness of solutions is also discussed, as well as the integrability of these solutions. Furthermore, algorithmic procedures for finding multivector fields solutions in the singular case is outlined in Sec. III D.

In Sec. III E we deal with the *generalized symmetries* of a Lagrangian system, and a version of *Noether's theorem* using the multivector field formalism is proved.

An example, which is a quite general version of many typical models in field theories, is analyzed in Sec. III F.

Finally, a summary with the main results in the work is given.

An appendix is included as a remainder of the definitions and main properties of the canonical structures in a jet bundle, which are used in some parts of the work.

All manifolds are real, paracompact, connected and C^∞. All maps are C^∞. The sum over crossed repeated indices is understood.

II. MULTIVECTOR FIELDS AND CONNECTIONS

We begin this work by studying the integrability of multivector fields in differentiable manifolds in general; and setting the equivalence between jet fields and multivector fields in jet bundles.

A. Multivector fields in differentiable manifolds

Let E be an n-dimensional differentiable manifold. Sections of $\Lambda^m(TE)$ are called *m-multivector fields* in E. We will denote by $\mathcal{X}^m(E)$ the set of m-multivector fields in E. In general, given a m-multivector field $Y\in\mathcal{X}^m(E)$, for every $p\in E$, there exists an open neighborhood $U_p\subset E$ and $Y_1,\ldots,Y_r\in\mathcal{X}(U_p)$ such that

$$Y=\sum_{U_p 1\leqslant i_1<\cdots<i_m\leqslant r} f^{i_1\cdots i_m}Y_{i_1}\wedge\cdots\wedge Y_{i_m}\,, \tag{1}$$

with $f^{i_1\cdots i_m}\in C^\infty(U_p)$ and $m\leqslant r\leqslant \dim E$. The situation we are interested in is when, for every $p\in E$, we can take $r=m$; therefore[19,20] we have the following.

Definition 1: An m-multivector field $Y\in\mathcal{X}^m(E)$ *is said to be* decomposable *iff there are* $Y_1,\ldots,Y_m\in\mathcal{X}(E)$ *such that* $Y=Y_1\wedge\cdots\wedge Y_m$.

The multivector field $Y\in\mathcal{X}^m(E)$ *is said to be* locally decomposable *iff, for every* $p\in E$, *there exists an open neighborhood* $U_p\subset E$ *and* $Y_1,\ldots,Y_m\in\mathcal{X}(U_p)$ *such that* $Y=_{U_p} Y_1\wedge\cdots\wedge Y_m$.

Every multivector field $Y\in\mathcal{X}^m(E)$ defines an operation $i(Y)$ of degree $-m$ in the algebra of differential forms $\Omega(E)$. If $\Omega\in\Omega^k(E)$ is a differentiable k-form in E; using (1), this is

$$i(Y)\Omega=\sum_{U_p 1\leqslant i_1<\cdots<i_m\leqslant r} f^{i_1\cdots i_m}i(Y_{i_1}\wedge\cdots\wedge Y_{i_m})\Omega=\sum_{1\leqslant i_1<\cdots<i_m\leqslant r} f^{i_1\cdots i_m}i(Y_{i_1})\cdots i(Y_{i_m})\Omega,$$

if $k\geqslant m$, and it is obviously equal to zero if $k<m$. The k-form Ω is said to be *j-nondegenerate* (for $1\leqslant j\leqslant k-1$) iff, for every $p\in E$ and $Y\in\mathcal{X}^j(E)$, $i(Y_p)\Omega_p=0\Leftrightarrow Y_p=0$ (see Refs. 10 and 11 for more information on these topics).

Let D be an m-dimensional distribution in E; that is, an m-dimensional subbundle of TE. Obviously sections of $\Lambda^m D$ are m-multivector fields in E. The existence of a nonvanishing global section of $\Lambda^m D$ is equivalent to the orientability of D. We are interested in the relation between nonvanishing m-multivector fields in E and m-dimensional distributions in TE. So we set the following.

Definition 2: A nonvanishing m-multivector field $Y\in\mathcal{X}^m(E)$ *and an m-dimensional distribution* $D\subset TE$ *are* locally associated *iff there exists a connected open set* $U\subseteq E$ *such that* $Y|_U$ *is a section of* $\Lambda^m D|_U$.

As a consequence of this definition, if $Y,Y'\in\mathcal{X}^m(E)$ are nonvanishing multivector fields locally associated, on the same connected open set U, with the same distribution D, then there exists a nonvanishing function $f\in C^\infty(U)$ such that $Y'=_U fY$. This fact defines an equivalence relation in the set of nonvanishing m-multivector fields in E, whose equivalence classes will be denoted by $\{Y\}_U$. Therefore we can state the following.

Theorem 1: *There is a bijective correspondence between the set of m-dimensional orientable distributions D in TE and the set of the equivalence classes* $\{Y\}_E$ *of nonvanishing, locally decomposable m-multivector fields in E.*

Proof: Let $\omega\in\Omega^m(E)$ be an orientation form for D. If $p\in E$ there exists an open neighborhood $U_p\subset E$ and $Y_1,\ldots,Y_m\in\mathcal{X}(U_p)$, with $i(Y_1\wedge\cdots\wedge Y_m)\omega>0$, such that $D|_{U_p}=\mathrm{span}\{Y_1,\ldots,Y_m\}$. Then $Y_1\wedge\cdots\wedge Y_m$ is a representative of a class of m-multivector fields associated with D in U_p. But the family $\{U_p;p\in E\}$ is a covering of E; let $\{U_\alpha;\alpha\in A\}$ be a locally finite refinement and $\{\rho_\alpha;\alpha\in A\}$ a subordinate partition of unity. If $Y_1^\alpha,\ldots,Y_m^\alpha$ is a local basis of D in U_α, with $i(Y_1^\alpha\wedge\cdots\wedge Y_m^\alpha)\omega>0$, then $Y=\Sigma_\alpha\rho_\alpha Y_1^\alpha\wedge\cdots\wedge Y_m^\alpha$ is a global representative of the class of nonvanishing m-multivector fields associated with D in E.

The converse is trivial because if $Y|_U=Y_1^1\wedge\cdots\wedge Y_m^1=Y_1^2\wedge\cdots\wedge Y_m^2$, for different sets $\{Y_1^1,\ldots,Y_m^1\}$, $\{Y_1^2,\ldots,Y_m^2\}$, then $span\ \{Y_1^1,\ldots,Y_m^1\}=span\ \{Y_1^2,\ldots,Y_m^2\}$. ■

Comments:

- If $Y \in \mathcal{X}^m(E)$ is a nonvanishing m-multivector field and $U \subseteq E$ is a connected open set, the distribution associated with the class $\{Y\}_U$ will be denoted by $\mathcal{D}_U(Y)$. If $U=E$ we will write simply $\mathcal{D}(Y)$.
- If D is a nonorientable m-dimensional distribution in TE, for every $p \in E$, there exists an open neighborhood $U_p \subset E$ and a nonvanishing multivector field $Y \in \mathcal{X}^m(U)$ such that $D|_{U_p} = \mathcal{D}_U(Y)$.

Definition 3: Let $Y \in \mathcal{X}^m(E)$ be a multivector field.

1. *A submanifold $S \hookrightarrow E$, with $dim\ S = m$, is said to be an* integral manifold *of Y iff, for every $p \in S$, Y_p spans $\Lambda^m T_p S$.*
2. *Y is said to be an* integrable multivector field *on an open set $U \subseteq E$ iff, for every $p \in U$, there exists an integral manifold $S \hookrightarrow U$ of Y, with $p \in S$.*
3. *Y is said to be* integrable *iff it is integrable in E.*

Obviously, every integrable multivector field is nonvanishing.

Now, bearing in mind the statement in theorem II A, we can define the following.

Definition 4: Let $Y \in \mathcal{X}^m(E)$ be a multivector field.

1. *Y is said to be* involutive *on a connected open set $U \subseteq E$ iff it is locally decomposable in U and its associated distribution $\mathcal{D}_U(Y)$ is involutive.*
2. *Y is said to be* involutive *iff it is involutive on E.*
3. *Y is said to be* locally involutive *around $p \in E$ iff there is a connected open neighborhood $U_p \ni p$ such that Y is involutive on U_p.*

Note that if Y is locally involutive around every $p \in E$, then it is involutive.

Now, the classical *Frobenius' theorem* can be reformulated in this context as follows.

Proposition 1: A nonvanishing and locally decomposable multivector field $Y \in \mathcal{X}^m(E)$ is integrable on a connected open set $U \subseteq E$ if, and only if, it is involutive on U.

Note that if a multivector field Y is integrable, then so is every other in its equivalence class $\{Y\}$, and all of them have the same integral manifolds.

As is well known, an m-dimensional distribution D is integrable if, and only if, it is locally spanned by a set of vector fields $(Y_1, \ldots, Y_m)$ such that $[Y_\mu, Y_\nu] = 0$, for every pair Y_μ, Y_ν. Thus a multivector field $Y \in \mathcal{X}^m(E)$ is integrable if, and only if, for every $p \in E$, there exists an open neighborhood $U_p \subset E$ and $Y_1, \ldots, Y_m \in \mathcal{X}(U_p)$ such that

1. $Y_1, \ldots, Y_m$ span $\mathcal{D}_{U_p}(Y)$.
2. $[Y_\mu, Y_\nu] = 0$, for every pair Y_μ, Y_ν.

Then there is a nonvanishing function $f \in C^\infty(U_p)$ such that $Y|_{U_p} = f Y_1 \wedge \cdots \wedge Y_m$.

Now, let $\{\tau^\mu\}$ be the local one-parameter groups of diffeomorphisms of Y_μ around p (for $\mu = 1, \ldots, m$). The second condition above implies that $\tau^\mu \circ \tau^\nu = \tau^\nu \circ \tau^\mu$, for every μ, ν. Then, if S_p is the integral manifold of Y through p, there exist open neighborhoods $V(0) \subset \mathbf{R}^m$ and $W_p \subset U_p \subset E$, and a map

$$\tau : V(0) \times W_p \to U_p$$

$$(t,q) \mapsto \tau_t(q) := (\tau^1_{t_1} \circ \ldots \circ \tau^m_{t_m})(q) \quad (t = (t_1, \ldots, t_m)),$$

verifying that

1. τ is injective.
2. $\tau_{t+s} = \tau_t \circ \tau_s$, for $t, s, t+s \in V(0)$.
3. For every $q \in W_p$, the set $\{\tau_t(q); t \in V(0)\}$ is an open neighborhood of q in S_q.

The map τ is called the m-flow associated with the multivector field Y (see Ref. 21 for the terminology and notation). Therefore we can define the following.

Definition 5: A multivector field $Y \in \mathcal{X}^m(E)$ is said to be a dynamical multivector field *iff*

1. *Y is integrable.*
2. *For every* $p \in E$, *there exists an open neighborhood* $U_p \subset E$ *and* $Y_1, \ldots, Y_m \in \mathcal{X}(U_p)$ *such that* $[Y_\mu, Y_\nu] = 0$, *for every pair* Y_μ, Y_ν, *and* $Y|_{U_p} = Y_1 \wedge \cdots \wedge Y_m$.

Thus, as a consequence of all these comments we have the following.

Proposition 2: Let $\{Y\}$ *be a class of integrable m-multivector fields. Then there is a representative Y of the class which is a dynamical multivector field.*

B. Integrability algorithm for locally decomposable multivector fields

Every locally decomposable multivector field $Y \in \mathcal{X}^m(E)$ defines a system of (first-order) partial differential equations, whose solutions (if they exist) give locally the integral manifolds of Y.

Nevertheless, in many applications, we have a locally decomposable multivector field which is not integrable in E, but integrable in a submanifold of E. What this means is that the corresponding system of partial differential equations has no solution everywhere in E, but only on the points of that submanifold. For instance, this problem arises when we look for solutions of evolution equations in field theories (as we will see later).

Next we present an algorithm which allows us to find this submanifold. Although we will maintain the global notation (E for the manifold), this is a local algorithm; that is, we are in fact working on suitable open sets in E. Hence, let $Y \equiv \wedge_{\mu=1}^{m} Y_\mu$.

• *Integrability condition*:

The condition for Y to be integrable is equivalent to demanding that the distribution spanned by $Y_1, \ldots, Y_m$ is involutive. Then, let $Z_1, \ldots, Z_{n-m} \in \mathcal{X}(E)$, such that $Y_1, \ldots, Y_m$, $Z_1, \ldots, Z_{n-m}$ are a local basis of $\mathcal{X}(E)$. Therefore, for every couple Y_μ, Y_ν $(1 \leqslant \mu, \nu \leqslant m)$ we have

$$[Y_\mu, Y_\nu] = \xi^\rho_{\mu\nu} Y_\rho + \zeta^l_{\mu\nu} Z_l,$$

for some functions $\xi^\rho_{\mu\nu}, \zeta^l_{\mu\nu}$. Consider the system $\zeta^l_{\mu\nu} = 0$ and let

$$E_1 = \{p \in E;\ \zeta^l_{\mu\nu}(p) = 0,\ \forall \mu, \nu, l\};$$

we have three options.

1. $E_1 = E$.
   Then the distribution spanned by $Y_1, \ldots, Y_m$ is involutive, and the multivector field Y is integrable in E. If, in addition, all the functions $\xi^\rho_{\mu\nu}$ are zero everywhere, then Y is also a dynamical multivector field.
2. $E_1 = \varnothing$.
   Then the distribution spanned by $Y_1, \ldots, Y_m$ is not involutive at any point in E, and hence the multivector field Y is not integrable.
3. E_1 is a proper subset of E.
   In this case we assume that E_1 is a closed submanifold of E and the functions $\zeta^l_{\mu\nu}$ are the constraints locally defining E_1. The distribution spanned by $Y_1, \ldots, Y_m$ is involutive on E_1; that is, the multivector field Y is involutive on E_1. Moreover, if all the functions $\xi^\rho_{\mu\nu}$ are zero everywhere in E_1, then Y is also a dynamical multivector field in E_1.
   Obviously, $Y : E_1 \to \Lambda^m TE|_{E_1}$. If, in addition, $Y : E_1 \to \Lambda^m TE_1$, then Y is an integrable multivector field tangent to E_1. Nevertheless, this is not the case in general, so we need the following.

• *Tangency condition*:

Consider the set

$$E_2 := \{p \in E_1;\ Y(p) \in \Lambda^m T_p E_1\}.$$

For E_2 we have the same problem, so we define inductively, for $i > 1$,

$$E_i := \{p \in E_{i-1};\ Y(p) \in \Lambda^m T_p E_{i-1}\},$$

and assume that we obtain a sequence of nonempty (closed) submanifolds $\cdots \subset E_i \subset \cdots \subset E_1 \subset E$.

Observe that the necessary and sufficient condition for a locally decomposable multivector field $Y=Y_1\wedge\cdots\wedge Y_m$ to be tangent to E_i is that Y_μ is tangent to E_i, for every μ. Hence, if $E_f:=\cap_{1\geqslant 1}E_i$, and assuming that E_f is a nonempty (closed) submanifold of E with $\dim E_f\geqslant m$, we have that $Y:E_f\to\Lambda^m TE_f$; this is involutive and therefore integrable on the manifold E_f.

Using the constraints, we have that, if $\{\zeta_{\alpha_i}^{(i)}\}$ is a basis of constraints defining locally E_i in E_{i-1}, the tangency condition is

$$0\underset{E_i}{=}Y_\mu(\zeta_{\alpha_i}^{(i)})\ ,\quad \forall\mu,\alpha_i\,,$$

that is, we have

$$E_{i+1}:=\{p\in E_i\,;Y_\mu(\zeta_{\alpha_i}^{(i)})(p)=0,\forall\mu,\alpha_i\},$$

and again we have the same situation as in the case of $E_1\subset E$.

This algorithm ends in one of the following two options.

1. We obtain a submanifold $E_f\hookrightarrow E$, with $\dim E_f\geqslant m$, where Y is an integrable multivector field (and, maybe, dynamical).
2. We obtain a submanifold E_f with $\dim E_f<m$, or the empty set. Then the problem has no solution in E.

We will call this procedure the *integrability algorithm* for multivector fields.

Remark: As is clear, although we have developed this algorithm for locally decomposable multivector fields, the procedure is obviously applicable for distributions in general.

C. An example

In order to show how the above algorithm of integrability is applied we take the following example of a system of partial differential equations which does not verify the integrability condition (see Ref. 22),

$$\frac{\partial y_1}{\partial x_1}=y_1-x_1-x_2\,,\quad \frac{\partial y_1}{\partial x_2}=y_1{}^2-x_1{}^2-x_2{}^2-2x_1-2x_2-2x_1x_2\,,$$

$$\frac{\partial y_2}{\partial x_1}=-y_2+x_1+x_2\,,\quad \frac{\partial y_2}{\partial x_2}=1.$$

For this system we have $E\equiv\mathbf{R}^4$, with local coordinates (x_1,x_2,y_1,y_2). The associated distribution is locally spanned by the vector fields

$$Y_1=\frac{\partial}{\partial x_1}+(y_1-x_1-x_2)\frac{\partial}{\partial y_1}+(-y_2+x_1+x_2)\frac{\partial}{\partial y_2}\,,$$

$$Y_2=\frac{\partial}{\partial x_2}+(y_1{}^2-x_1{}^2-x_2{}^2-2x_1-2x_2-2x_1x_2)\frac{\partial}{\partial y_1}+\frac{\partial}{\partial y_2}\,,$$

and a representative of the corresponding class of locally decomposable multivector fields $\{Y\}\subset\mathcal{X}^2(E)$ is $Y=Y_1\wedge Y_2$.

• *Integrability condition*:

As local basis of $\mathcal{X}(E)$ we take the set $(Y_1,Y_2,\partial/\partial y_1,\partial/\partial y_2)$. Then the condition of involutivity of the above distribution leads to

$$[Y_1,Y_2]=((x_1+x_2-y_1)^2-1)\frac{\partial}{\partial y_1}=0,$$

which gives the constraint $\zeta^{(1)}\equiv(x_1+x_2-y_1)^2-1=(x_1+x_2-y_1-1)(x_1+x_2-y_1+1)\equiv\zeta_a^{(1)}\zeta_b^{(1)}$; and we have the two disjoint submanifolds,

$$E_1^a := \{(x_1, x_2, y_1, y_2) | x_1 + x_2 - y_1 - 1 = 0\},$$

$$E_1^b := \{(x_1, x_2, y_1, y_2) | x_1 + x_2 - y_1 + 1 = 0\},$$

and the tangency condition must hold separately for each one of them.

• *Tangency condition*:

1. *Tangency condition for* E_1^a: We obtain that $Y_1(\zeta_a^{(1)}) \underset{E_1^a}{=} 2$; then Y_1 is not tangent to E_1^a at any point, and the system has no solution on E_1^a.
2. *Tangency condition for* E_1^b: In this case we have that $Y_1(\zeta_a^{(1)}) \underset{E_1^b}{=} 0$ and $Y_2(\zeta_a^{(1)}) \underset{E_1^b}{=} 0$. Then $Y|_{E_1^b} \in \mathcal{X}^2(E_1^b)$ and it is integrable on E_1^b.

• *Integration of the system on* E_1^b:

The general solution of the system on E_1^b is

$$y_1 = x_1 + x_2 + 1, \quad y_2 = x_1 + x_2 - 1 + ce^{-x_1} \quad (c \in \mathbf{R}),$$

and there exists an integral manifold through every point on E_1^b. (Compare with Ref. 22, pp. 47, 53.)

D. Multivector fields in fiber bundles

The particular situation in which we are interested is the study of multivector fields in fiber bundles and, in particular, in jet bundles. Bearing this in mind, first we state some results concerning multivector fields in fiber bundles in general.

Let $\tau: E \to M$ be a fiber bundle with $\dim M = m$. We denote by $\mathrm{V}(\tau)$ the vertical bundle associated with τ [that is, $\mathrm{V}(\tau) = \mathrm{Ker}\, T\tau$], and by $\mathcal{X}^{\mathrm{V}(\tau)}(E)$ the corresponding sections or vertical vector fields.

We are interested in the case that the integral manifolds of multivector fields are sections of the projection τ. In order to characterize this kind of multivector fields we set the following.

Definition 6: A multivector field $Y \in \mathcal{X}^m(E)$ *is said to be* transverse to the projection τ *(or* τ-transverse*) iff, at every point* $y \in E$, $(i(Y)(\tau^* \omega))_y \neq 0$, *for every* $\omega \in \Omega^m(M)$ *with* $\omega(\tau(y)) = 0$.

Comments:

• This condition is equivalent to $(\Lambda^m T\tau \circ Y)(y) \neq 0$, for every $y \in E$.

• If $Y \in \mathcal{X}^m(E)$ is locally decomposable, then Y is τ-transverse if, and only if, $T_y\tau(\mathcal{D}(Y)) = T_{\tau(y)}M$, for every $y \in E$. [Remember that $\mathcal{D}(Y)$ is the m-distribution associated to Y.]

Then we can state the following.

Theorem 2: *Let* $Y \in \mathcal{X}^m(E)$ *be integrable. Then* Y *is* τ*-transverse if, and only if, its integral manifolds are local sections of* $\tau: E \to M$.

Proof: Consider $y \in E$, with $\tau(y) = x$. In a neighborhood of y there exists $Y_1, \ldots, Y_m \in \mathcal{X}(E)$ such that $Y_1, \ldots, Y_m$ span $\mathcal{D}(Y)$ and $Y = Y_1 \wedge \cdots \wedge Y_m$. But, as Y is τ-transverse, $(i(Y) \times (\tau^* \omega))_y \neq 0$, for every $\omega \in \Omega^m(M)$ with $\omega(x) \neq 0$. Thus, taking into account the second comment above, $\mathcal{D}(Y)$ is a τ-transverse distribution and $Y_\mu \notin \mathcal{X}^{\mathrm{V}(\tau)}(E)$ at any point, for every Y_μ. Now, let $S \hookrightarrow E$ be the integral manifold of $\mathcal{D}(Y)$ passing through y then $T_y S = \mathrm{span}\{(Y_1)_y, \ldots, (Y_m)_y\}$. As a consequence of all of this, and again taking into account the second comment above, for every point $y \in S$, $T_y\tau(\mathcal{D}(Y)) = T_{\tau(y)}M$, then $\tau|_S$ is a local diffeomorphism and S is a local section of τ.

The converse is obvious. ■

Observe that, in this case, if $\phi: U \subset M \to E$ is a local section with $\phi(x) = y$ and $\phi(U)$ is the integral manifold of Y through y, then $T_y(\mathrm{Im}\ \phi)$ is $\mathcal{D}_y(Y)$.

Now, considering the diagram

$$\begin{array}{ccc} \Lambda^m\mathrm{T}(\pi^{-1}(U)) & \overset{\Lambda^m\mathrm{T}\pi}{\underset{\Lambda^m\mathrm{T}\phi}{\rightleftarrows}} & \Lambda^m\mathrm{T}U \\ \Lambda^m\tau_E \big\downarrow\big\uparrow & Y & \big\downarrow \Lambda^m\tau_U \\ \pi^{-1}(U) & \underset{\pi}{\overset{\phi}{\rightleftarrows}} & U\subset M \end{array}$$

we have the following.

Proposition 3: $Y\in\mathcal{X}^m(E)$ is integrable and π-transverse if, and only if, for every point $y\in E$, there exists a local section $\phi:U\subset M\to E$ such that $\phi(\pi(y))=y$; and a nonvanishing function $f\in C^\infty(E)$ such that $\Lambda^m T\phi = fY\circ\phi\circ\Lambda^m\tau_U$.

Proof: If Y is integrable and π-transverse, then by theorem 2, for every $y\in E$, with $\pi(y)=x$, there is an integral local section $\phi:U\subset M\to E$ of Y at y such that $\phi(x)=y$. Then, by definition 3, Y_y spans $\Lambda^m T_y\phi$, and hence the relation in the statement holds and the above diagram becomes commutative on the open set U.

Conversely, if the relation holds, then Im ϕ is an integral manifold of Y at y; then Y is integrable and, as ϕ is a section of π, Y is necessarily π-transverse. ■

E. Multivector fields and connections in jet bundles

Next we are going to establish the relation between multivector fields and connections in jet bundles. Let $\pi:E\to M$ be a fiber bundle, $\pi^1:J^1E\to E$ the corresponding first-order jet bundle of local sections of π, and $\overline{\pi^1}=\pi\circ\pi^1:J^1E\to M$. If (x^μ,y^A) is a local system of coordinates adapted to the projection π ($\mu=1,\ldots,m$, $A=1,\ldots,N$), let (x^μ,y^A,v^A_μ) be the induced local system of coordinates in J^1E.

With the aim of relating the elements in $\mathcal{X}^m(E)$ to connections in $\pi:E\to M$, we briefly recall several ways of giving a connection (see Ref. 14).

Definition 7: A connection in $\pi:E\to M$ *is one of the following equivalent elements.*

1. *A global section of* $\pi^1:J^1E\to E$ *(that is, a mapping* $\Psi:E\to J^1E$ *such that* $\pi^1\circ\Psi=\mathrm{Id}_E$*). It is called a* jet field.
2. *A subbundle* $H(E)$ *of* TE *such that* $TE=V(\pi)\oplus H(E)$*. It is called a* horizontal subbundle.
3. *A* π*-semibasic* 1*-form* $\boldsymbol{\nabla}$ *on* E *with values in* TE*, such that* $\boldsymbol{\nabla}^*\alpha=\alpha$*, for every* π*-semibasic form* $\alpha\in\Omega^1(E)$*. It is called the* connection form *or* Ehresmann connection.

The equivalence among these elements can be seen in Ref. 14, or it suffices to observe their local expressions. In a natural chart (x^μ,y^A,v^A_μ) we have

$$\Psi=(x^\mu,y^A,\Gamma^A_\rho(x^\mu,y^A)),$$

$$\boldsymbol{\nabla}=dx^\mu\otimes\left(\frac{\partial}{\partial x^\mu}+\Gamma^A_\mu(x^\mu,y^A)\frac{\partial}{\partial y^A}\right), \tag{2}$$

$$H(E)=\mathrm{span}\left\{\frac{\partial}{\partial x^\mu}+\Gamma^A_\mu(x^\mu,y^A)\frac{\partial}{\partial y^A}\right\}.$$

Comments:

- The horizontal subbundle $H(E)$ is also denoted $\mathcal{D}(\Psi)$ when it is considered as the distribution associated with Ψ.
- A jet field $\Psi:E\to J^1E$ (resp., an Ehresmann connection $\boldsymbol{\nabla}$) is said to be orientable iff $\mathcal{D}(\Psi)$ is an orientable distribution on E.
- If M is orientable, then every connection is also orientable.

Let $\Psi:E\to J^1E$ be a jet field. A section $\phi:M\to E$ is said to be an *integral section* of Ψ iff $\Psi\circ\phi=j^1\phi$ (where $j^1\phi:M\to J^1E$ denotes the canonical lifting of ϕ). Ψ is said to be an *integrable*

jet field iff it admits integral sections through every point of E or, what is equivalent, if, and only if, $\mathcal{D}(\Psi)$ is an involutive distribution [that is, $\mathcal{D}(\Psi)$ is integrable]. Locally, if $\phi=(x^\mu,f^A(x^\nu))$, then ϕ is an integral section of Ψ if, and only if, ϕ is a solution of the following system of partial differential equations:

$$\frac{\partial f^A}{\partial x^\mu}=\Gamma^A_\mu\circ\phi. \tag{3}$$

As is well known, a jet field Ψ is integrable if, and only if, the curvature of the connection form $\mathbf{\nabla}$ associated with Ψ vanishes identically; where the *curvature* of a connection $\mathbf{\nabla}$ is the (2,1)-tensor field defined by

$$\mathcal{R}(Z_1,Z_2):=(\mathrm{Id}-\mathbf{\nabla})([\mathbf{\nabla}(Z_1),\mathbf{\nabla}(Z_2)])=i([\mathbf{\nabla}(Z_1),\mathbf{\nabla}(Z_2)])(\mathrm{Id}-\mathbf{\nabla}),$$

for every $Z_1,Z_2\in\mathcal{X}(E)$. Using the coordinate expressions of the connection form $\mathbf{\nabla}$ and the 1-jet field Ψ, a simple calculation leads to

$$\mathcal{R}=\frac{1}{2}\left(\frac{\partial\Gamma^B_\eta}{\partial x^\mu}-\frac{\partial\Gamma^B_\mu}{\partial x^\eta}+\Gamma^A_\mu\frac{\partial\Gamma^B_\eta}{\partial y^A}-\Gamma^A_\eta\frac{\partial\Gamma^B_\mu}{\partial y^A}\right)(dx^\mu\wedge dx^\eta)\otimes\frac{\partial}{\partial y^B}\,.$$

Hence, from this local expression of $\mathcal{R}$ we obtain the local integrability conditions of the equations (3).

Now, the relation between multivector fields and connections (jet fields) is given as a particular case of theorem 1.

Theorem 3: *There is a bijective correspondence between the set of orientable jet fields* $\Psi:E\to J^1E$ *(that is, the set of orientable connections* $\mathbf{\nabla}$ *in* $\pi:E\to M$*) and the set of the equivalence classes of locally decomposable and* π*-transverse multivector fields* $\{Y\}\subset\mathcal{X}^m(E)$*. They are characterized by the fact that* $\mathcal{D}(\Psi)=\mathcal{D}(Y)$.

In addition, the orientable jet field Ψ *is integrable if, and only if, so is Y, for every* $Y\in\{Y\}$.

Proof: If Ψ is an orientable jet field in J^1E, let $\mathcal{D}(\Psi)$ be its horizontal distribution. Then, taking $D\equiv\mathcal{D}(\Psi)$, we construct $\{Y\}$ by applying theorem 1 and, since the distribution $\mathcal{D}(\Psi)$ is π-transverse, the result follows immediately. The proof of the converse statement is similar.

Moreover, Ψ is integrable if, and only if, $\mathcal{D}(\Psi)=\mathcal{D}(Y)$ also is. Therefore it follows that Y is also integrable, for $Y\in\{Y\}$, and conversely. ■

As is obvious, recalling the local expression (2), we obtain the following local expression for a particular representative multivector field Y of the class $\{Y\}$ associated with the jet field Ψ:

$$Y\equiv\bigwedge_{\mu=1}^{m}Y_\mu=\bigwedge_{\mu=1}^{m}\left(\frac{\partial}{\partial x^\mu}+\Gamma^A_\mu\frac{\partial}{\partial y^A}\right),$$

and $\phi=(x^\mu,f^A(x^\nu))$ is an integral section of Y if, and only if, ϕ is a solution of the system of partial differential equations (3).

Remarks:

- If M is an orientable manifold and $\omega\in\Omega^m(M)$ is the volume form in M, then the above representative can be obtained from the relation $i(Y)(\pi^*\omega)=i(\partial/\partial x^1\wedge\cdots\wedge\partial/\partial x^m)\times(\pi^*\omega)$.
- It is interesting to point out that, if we consider the integrability algorithm described in Sec. II B for this kind of multivector fields, then the integrability condition (given by the involutivity condition for the associated distribution) is equivalent now to the condition $\mathcal{R}=0$ ($\mathcal{R}$ being the curvature of the associated jet field). Furthermore, in this case, at the end of the algorithm, we will obtain directly a *dynamical multivector field* as the representative of the class of integrable multivector fields.

In Lagrangian classical field theory we are interested in multivector fields in the bundle $\overline{\pi}^1:J^1E\to M$. Hence, all the above considerations must be adapted to the bundle projections $J^1J^1E\xrightarrow{\pi^1_1}J^1E\xrightarrow{\overline{\pi}^1}M$. In this way we have the following.

1. Let $X \in \mathcal{X}^m(J^1E)$ be integrable. Then X is $\overline{\pi}^1$-transverse if, and only if, its integral manifolds are local sections of $\overline{\pi}^1 : J^1E \to M$ (theorem 2).
2. $X \in \mathcal{X}^m(J^1E)$ is integrable and $\overline{\pi}^1$-transverse if, and only if, for every point $\overline{y} \in J^1E$, there exists a local section $\psi : U \subset M \to J^1E$ such that $\psi(\overline{\pi}^1(\overline{y})) = \overline{y}$; and a nonvanishing function $f \in \mathrm{C}^\infty(J^1E)$ such that $\Lambda^m T\psi = fX \circ \psi \circ \Lambda^m \tau_U$ (proposition 3).
3. In order to relate multivector fields in J^1E with connections in $\overline{\pi}^1 : J^1E \to M$, let $\mathcal{Y} : J^1E \to J^1J^1E$ be a jet field, and $\boldsymbol{\nabla}$ and $H(J^1E)$ its associated connection form and horizontal subbundle, respectively. If (x^μ, y^A, v^A_μ) is a natural local chart in J^1E, then the induced chart in J^1J^1E is denoted $(x^\mu, y^A, v^A_\mu, a^A_\rho, b^A_{\rho\mu})$. We have the following local expressions for these elements:

$$\mathcal{Y} = (x^\mu, y^A, v^A_\mu, F^A_\rho(x,y,v), G^A_{\mu\rho}(x,y,v)),$$

$$\boldsymbol{\nabla} = dx^\mu \otimes \left(\frac{\partial}{\partial x^\mu} + F^A_\mu \frac{\partial}{\partial y^A} + G^A_{\mu\rho} \frac{\partial}{\partial v^A_\rho} \right),$$

$$H(J^1E) = \mathrm{span} \left\{ \frac{\partial}{\partial x^\mu} + F^A_\mu \frac{\partial}{\partial y^A} + G^A_{\mu\rho} \frac{\partial}{\partial v^A_\rho} \right\}.$$

4. Let $\psi = (x^\mu, f^A(x), g^A_\mu(x))$ be a local section of $\overline{\pi}^1$. It is an integral section of $\mathcal{Y}$ if, and only if, ψ is a solution of the following system of partial differential equations:

$$\frac{\partial f^A}{\partial x^\mu} = F^A_\mu \circ \psi\,, \quad \frac{\partial g^A_\rho}{\partial x^\mu} = G^A_{\rho\mu} \circ \psi, \tag{4}$$

and the necessary and sufficient condition for this system to be integrable is that the curvature $\mathcal{R}$ of $\mathcal{Y}$ vanishes everywhere; that is, in coordinates,

$$\begin{aligned} 0 = & \left(\frac{\partial F^B_\eta}{\partial x^\mu} + F^A_\mu \frac{\partial F^B_\eta}{\partial y^A} + G^A_{\gamma\mu} \frac{\partial F^B_\eta}{\partial v^A_\gamma} - \frac{\partial F^B_\mu}{\partial x^\eta} - F^A_\eta \frac{\partial F^B_\mu}{\partial y^A} - G^A_{\rho\eta} \frac{\partial F^B_\mu}{\partial v^A_\rho} \right) (dx^\mu \wedge dx^\eta) \otimes \frac{\partial}{\partial y^B} \\ & + \left(\frac{\partial G^B_{\rho\eta}}{\partial x^\mu} + F^A_\mu \frac{\partial G^B_{\rho\eta}}{\partial y^A} + G^A_{\gamma\mu} \frac{\partial G^B_{\rho\eta}}{\partial v^A_\gamma} - \frac{\partial G^B_{\rho\mu}}{\partial x^\eta} - F^A_\eta \frac{\partial G^B_{\rho\mu}}{\partial y^A} - G^A_{\gamma\eta} \frac{\partial G^B_{\rho\mu}}{\partial v^A_\gamma} \right) (dx^\mu \wedge dx^\eta) \otimes \frac{\partial}{\partial v^B_\rho}\,. \end{aligned} \tag{5}$$

5. From the above local expressions we obtain for a representative multivector field X of the class $\{X\}$ associated with the jet field $\mathcal{Y}$, the expression

$$X \equiv \bigwedge_{\mu=1}^{m} X_\mu = \bigwedge_{\mu=1}^{m} \left(\frac{\partial}{\partial x^\mu} + F^A_\mu \frac{\partial}{\partial y^A} + G^A_{\mu\rho} \frac{\partial}{\partial v^A_\rho} \right). \tag{6}$$

Moreover, as in the case of the bundle E, if M is an orientable manifold and $\omega \in \Omega^m(M)$ is the volume form in M, then the above representative can be obtained from the relation

$$i(Y)(\overline{\pi}^{1*}\omega) = i\left(\frac{\partial}{\partial x^1} \wedge \cdots \wedge \frac{\partial}{\partial x^m} \right)(\overline{\pi}^{1*}\omega).$$

F. Holonomic multivector fields and holonomic jet fields

We wish to characterize the integrable multivector fields in J^1E whose integral manifolds are canonical prolongations of sections of π. In order to achieve this we first recall the same situation for jet fields.[4,14]

Definition 8: A jet field $\mathcal{Y}$ is said to be *holonomic* iff it is integrable and its integral sections $\psi: M \to J^1E$ are holonomic; that is, they are canonical liftings of sections $\phi: M \to E$.

Let $\mathbf{y} \in J^1J^1E$ with $\mathbf{y} \overset{\pi_1^1}{\mapsto} y \overset{\overline{\pi}^1}{\mapsto} y \overset{\pi}{\mapsto} x$, and $\psi: M \to J^1E$ a representative of $\mathbf{y}$; that is, $\mathbf{y} = T_x\psi$. Consider the section $\phi = \pi^1 \circ \psi: M \to E$, and let $j^1\phi$ be its canonical prolongation. Then we are able to define another natural projection,

$$\begin{array}{rccc} j^1\pi^1 & : & J^1J^1E & \mapsto & J^1E \\ & & \mathbf{y} & \mapsto & j^1(\pi^1 \circ \psi)(\overline{\pi}_1^1(\mathbf{y})). \\ & & (x^\mu, y^A, v^A_\mu, a^A_\rho, b^A_{\mu\rho}) & \mapsto & (x^\mu, y^A, a^A_\rho) \end{array}$$

Definition 9: A jet field $\mathcal{Y}: J^1E \to J^1J^1E$ *is said to be a* Second Order Partial Differential Equation (SOPDE) *or a* semi-holonomic *jet field (or also that it verifies the* SOPDE condition), *iff it is also a section of the projection* $j^1\pi^1$*; that is,* $j^1\pi^1 \circ \mathcal{Y} = \mathrm{Id}_{J^1E}$.

This is equivalent to saying that $\mathcal{Y}$ *is a section of the projection* $\hat{J}^2E \to J^1E$*, where* $\hat{J}^2E$ *denotes the* semi-holonomic 2-jet manifold *(see Ref. 14, p. 173).*

The relation among integrable, holonomic and SOPDE jet fields (connections) is the following.

Theorem 4: A jet field $\mathcal{Y}: J^1E \to J^1J^1E$ is holonomic if, and only if, it is integrable and SOPDE.

The condition $j^1\pi^1 \circ \mathcal{Y} = \mathrm{Id}_{J^1E}$ is locally expressed as follows: $\mathcal{Y} = (x^\mu, y^A, v^A_\mu, F^A_\rho, G^A_{\nu\rho})$ is a SOPDE if, and only if, $F^A_\rho = v^A_\rho$.

For multivector fields in J^1E, first we define the following.

Definition 10: A multivector field $X \in \mathcal{X}^m(J^1E)$ is said to be holonomic iff

1. X is integrable.
2. X is $\overline{\pi}^1$-transverse.
3. The integral sections $\psi: M \to J^1E$ of X are holonomic.

In order to study the SOPDE question for multivector fields, we adopt two different approaches: The first one, which is similar to the characterization of *second order* (that is, holonomic) vector fields in time-dependent mechanics, is based on the study of the two different bundle structures of TJ^1E over TE. The second one uses the *structure canonical form* $\theta \in \Omega^1(J^1E, \pi^{1*}V(\pi))$ (see the Appendix).

In the first approach, we have a natural vector bundle projection $T\pi^1: TJ^1E \to TE$ and another one $\kappa: TJ^1E \to TE$ defined by

$$\kappa(\overline{y,u}) := T_{\overline{\pi}^1(\overline{y})}\phi(T_{\overline{y}}\overline{\pi}^1(\overline{u})),$$

where $(\overline{y,u}) \in TJ^1E$ and $\phi \in \overline{y}$. If $(W; x^\mu, y^A, v^A_\mu)$ is a local natural chart in J^1E and $\overline{y} \in J^1E$ with $\overline{y} \overset{\overline{\pi}^1}{\to} y \overset{\pi}{\to} x$, then

$$\kappa\left(\alpha^\mu \frac{\partial}{\partial x^\mu}\bigg|_{\overline{y}} + \beta^A \frac{\partial}{\partial y^A}\bigg|_{\overline{y}} + \lambda^A_\mu \frac{\partial}{\partial v^A_\mu}\bigg|_{\overline{y}}\right) = \alpha^\mu \frac{\partial}{\partial x^\mu}\bigg|_y + v^A_\mu(\overline{y})\alpha^\mu \frac{\partial}{\partial y^A}\bigg|_y.$$

This projection is extended in a natural way to $\Lambda^m TJ^1E$, and so we have the following diagram:

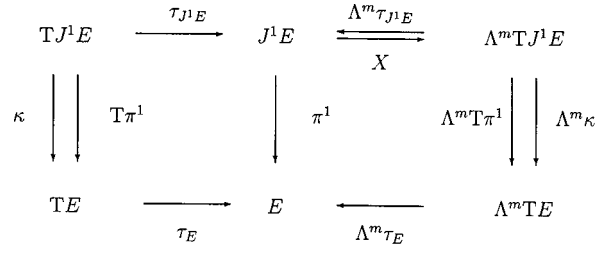

and we can define the following.

Definition 11: A $\overline{\pi^1}$*-transverse multivector field* $X \in \mathcal{X}^m(J^1E)$ *verifies the* SOPDE condition *(we also say that it is a* SOPDE *or a* semi-holonomic multivector field*) iff* $\Lambda^m\kappa \circ X = \Lambda^m T\pi^1 \circ X$.

The condition for X to be $\overline{\pi^1}$-transverse is necessary in order to relate this definition to the fact that, if X is integrable, then its integral sections are canonical prolongations of sections of π (as will be seen later).

With regard to the second approach, observe that for every $X \in \mathcal{X}(J^1E)$, we obtain that $\theta(X) \equiv i(\theta)X$ is a section of the bundle $\pi^{1*}V(\pi) \to J^1E$ [that is, an element of $\Gamma(\pi^{1*}V(\pi))$]. Then we have a natural extension of the action of θ to every $X \in \mathcal{X}^m(J^1E)$. Furthermore, denoting $V^m(\pi) := \ker \Lambda^m T\pi$ and recalling the definition of θ (see definition A in the Appendix) we can define the following.

Definition 12: The structure canonical *m*-form *of* J^1E *is the m-form* θ^m *in* J^1E *with values on* $\pi^{1*}V^m(\pi)$, *which is defined by*

$$(\theta^m(X))(\overline{y}) \equiv (i(\theta^m)X)(\overline{y}) := (\Lambda^m T_{\overline{y}}\overline{\pi}^1 - \Lambda^m T_{\overline{y}}(\phi \circ \overline{\pi^1}))(X_{\overline{y}}),$$

where $X \in \mathcal{X}^m(J^1E)$, $\overline{y} \in J^1E$ and ϕ is a representative of $\overline{y}$.

Observe that $\theta^m \in \Omega^m(J^1E, \pi^{1*}V^m(\pi))$ and that $\theta^m \neq \overbrace{\theta \wedge \cdots \wedge \theta}^{m}$.

Now we can state the following.

Proposition 4: Let $X \in \mathcal{X}^m(J^1E)$ *be* $\overline{\pi^1}$*-transverse and locally decomposable. Then the following conditions are equivalent.*

1. *X is a SOPDE.*
2. $i(\theta^m)X = 0$.
3. $i(\theta)X = 0$.
4. $(W; x^\mu, y^A, v^A_\mu)$ *is a natural chart in* J^1E, *then the local expression of X* is

$$X \equiv \bigwedge_{\mu=1}^{m} X_\mu = \bigwedge_{\mu=1}^{m} f_\mu \left(\frac{\partial}{\partial x^\mu} + v^A_\mu \frac{\partial}{\partial y^A} + G^A_{\mu\rho} \frac{\partial}{\partial v^A_\rho} \right), \tag{7}$$

where f_μ *are nonvanishing functions.*

5. $i(\alpha)X = 0$, *for every* $\alpha \in \mathcal{M}_c$; *where* $\mathcal{M}_c$ *denotes the contact module* in J^1E *(see the Appendix).*

Proof: Let $(W; x^\mu, y^A, v^A_\mu)$ be a natural chart in J^1E. Suppose that $X \in \mathcal{X}^m(J^1E)$ is $\overline{\pi^1}$-transverse and locally decomposable; then in this chart we have

$$X = \bigwedge_{\mu=1}^{m} f_\mu \left(\frac{\partial}{\partial x^\mu} + F^A_\mu \frac{\partial}{\partial y^A} + G^A_{\mu\rho} \frac{\partial}{\partial v^A_\rho} \right),$$

with f_μ nonvanishing functions:

$(1 \Leftrightarrow 2)$ It is obvious.

$(2 \Leftrightarrow 4)$ If $\overline{y} \in U \subset J^1E$ and $\phi \in \overline{y}$, then

$$\begin{aligned}(\theta^m(X))(\overline{y}) &= (\Lambda^m T_{\overline{y}}\overline{\pi}^1 - \Lambda^m T_{\overline{y}}(\phi \circ \overline{\pi^1}))(X_{\overline{y}}) \\ &= \bigwedge_{\mu=1}^{m} f_\mu(\overline{y}) \left(\frac{\partial}{\partial x^\mu} + F^A_\mu(\overline{y}) \frac{\partial}{\partial y^A} \right)_y - \bigwedge_{\mu=1}^{m} f_\mu(\overline{y}) \left(\frac{\partial}{\partial x^\mu} + v^A_\mu(\overline{y}) \frac{\partial}{\partial y^A} \right)_y,\end{aligned}$$

therefore $\theta^m(X) = 0 \Leftrightarrow F^A_\mu = v^A_\mu$, for every A, μ,

$(3 \Leftrightarrow 4)$ Since $\theta = (dy^B - v^B_\rho\, dx^\rho) \otimes \dfrac{\partial}{\partial y^B}$, we have that

$$i(\theta)X=\Bigg[f^1(F_1^B-v_1^B)\wedge\bigwedge_{\mu\neq 1} f_\mu\Bigg(\frac{\partial}{\partial x^\mu}+F_\mu^A\frac{\partial}{\partial y^A}+G_{\mu\rho}^A\frac{\partial}{\partial v_\rho^A}\Bigg)$$
$$-f^2(F_2^B-v_2^B)\wedge\bigwedge_{\mu\neq 2} f_\mu\Bigg(\frac{\partial}{\partial x^\mu}+F_\mu^A\frac{\partial}{\partial y^A}+G_{\mu\rho}^A\frac{\partial}{\partial v_\rho^A}\Bigg)+\ \dots\ \Bigg]\otimes\frac{\partial}{\partial y^B};$$

therefore $i(X)\theta=0 \Leftrightarrow F_\mu^A=v_\mu^A$, for every A,μ.

($4 \Leftrightarrow 5$) It is immediate from item 4, and bearing in mind the local expressions given in the Appendix. ■

The relation between integrable, holonomic and SOPDE multivector fields in J^1E is the following.

Theorem 5: *A multivector field $X\in\mathcal{X}^m(J^1E)$ is holonomic if, and only if, it is integrable and SOPDE.*

Proof: Let $\overline{y}\in J^1E$ be and $\psi:M\to J^1E$ an integral section of X at $\overline{y}$, with $\psi(x)=\overline{y}$. For $Z\in\Lambda^m T_xM$ with $Z_x\neq 0$, as ψ is an integral section, there exists $\lambda\in\mathbf{R}$ with $\lambda\neq 0$ such that $\Lambda^m T_x\psi(Z_x)=\lambda X_{\overline{y}}$ and hence

$$(i(\theta)X)(\overline{y})=\frac{1}{\lambda}i(\theta)(\Lambda^m T_x\psi(Z_x))=\frac{1}{\lambda}i(\psi^*\theta)Z_x=0;$$

therefore $i(X)\theta=0 \Leftrightarrow \psi^*\theta=0$, which is the necessary and sufficient condition for ψ to be holonomic (see proposition A in the Appendix). ■

Finally, as a consequence of theorems 3, 4 and 5 we have the following theorem.

Theorem 6: There is a bijective correspondence between the set of orientable jet fields $\mathcal{Y}:J^1E\to J^1J^1E$ (i.e., the set of orientable connections $\boldsymbol{\nabla}$ in $\overline{\pi}^1:J^1E\to M$) and the set of the equivalence classes of locally decomposable and $\overline{\pi}^1$-transverse multivector fields $\{X\}\subset\mathcal{X}^m(J^1E)$. They are characterized by the fact that $\mathcal{D}(\mathcal{Y})=\mathcal{D}(X)$. In addition we have the following.

1. The jet field $\mathcal{Y}$ is integrable if, and only if, so is X, for every $X\in\{X\}$.
2. The jet field $\mathcal{Y}$ is a SOPDE if, and only if, so is X, for every $X\in\{X\}$.
3. The jet field $\mathcal{Y}$ is holonomic if, and only if, so is X, for every $X\in\{X\}$.

If $\mathcal{Y}$ is a SOPDE, then we can choose [see the local expressions (6) and (7)]

$$X\equiv\bigwedge_{\mu=1}^m\Bigg(\frac{\partial}{\partial x^\mu}+v_\mu^A\frac{\partial}{\partial y^A}+G_{\mu\rho}^A\frac{\partial}{\partial v_\rho^A}\Bigg), \tag{8}$$

as a representative multivector field of $\{X\}$. For a SOPDE multivector field or jet field, on the other hand, if a section $j^1\phi=(x^\mu,f^A,\partial f^A/\partial x^\mu)$ has to be an integral section of such a field, the necessary and sufficient condition is that ϕ is the solution of the following system of (second-order) partial differential equations:

$$G_{\nu\rho}^A\Bigg(x^\mu,f^A,\frac{\partial f^A}{\partial x^\mu}\Bigg)=\frac{\partial^2 f^A}{\partial x^\rho\ \partial x^\nu}\,, \tag{9}$$

which justifies the nomenclature. It is important to remark that, since the integrability of a class of multivector fields is equivalent to demanding that the curvature $\mathcal{R}$ of the connection associated with this class vanishes everywhere; the system (9) has solution if, and only if, the following additional system of equations [which arise from the condition (5) for a SOPDE] holds (for every $B,\ \mu,\ \rho,\ \eta$):

$$\begin{aligned}0&=G_{\eta\mu}^B-G_{\mu\eta}^B,\\ 0&=\frac{\partial G_{\eta\rho}^B}{\partial x^\mu}+v_\mu^A\frac{\partial G_{\eta\rho}^B}{\partial y^A}+G_{\mu\gamma}^A\frac{\partial G_{\eta\rho}^B}{\partial v_\gamma^A}-\frac{\partial G_{\mu\rho}^B}{\partial x^\eta}-v_\eta^A\frac{\partial G_{\mu\rho}^B}{\partial y^A}-G_{\eta\gamma}^A\frac{\partial G_{\mu\rho}^B}{\partial v_\gamma^A}\,.\end{aligned} \tag{10}$$

Observe that this is a system of $\frac{1}{2}Nm(m-1)$ linear relations (the first group of equations) and $\frac{1}{2}Nm^2(m-1)$ partial differential equations (the second group).

III. APPLICATION TO LAGRANGIAN CLASSICAL FIELD THEORIES

At this point, our goal is to show that the Lagrangian formalism for field theories can also be established using jet fields in J^1J^1E, their associated connections in J^1E or, equivalently, multivector fields in J^1E. Then, we use multivector fields for discussing the main features of the evolution equations.

A. Background on Lagrangian field theories

Henceforth we assume that M is an m-dimensional oriented manifold and $\omega \in \Omega^m(M)$ is the volume m-form on M.

A *classical field theory* is described by its *configuration bundle* $\pi: E \to M$ (where M is an oriented manifold with volume form ω), and a *Lagrangian density* which is a $\overline{\pi^1}$-semibasic m-form on J^1E. A Lagrangian density is usually written as $\mathcal{L} = \pounds(\overline{\pi^1}^* \omega)$, where $\pounds \in C^\infty(J^1E)$ is the *Lagrangian function* associated with $\mathcal{L}$ and ω. In a natural system of coordinates this expression is $\mathcal{L} = \pounds(x^\mu, y^A, v^A_\mu) dx^1 \wedge \cdots \wedge dx^m$. Then a *Lagrangian system* is a couple $((E,M;\pi),\mathcal{L})$. The states of the field are the sections of π which are critical for the functional,

$$\begin{aligned} \mathbf{L} \quad : \quad & \Gamma_c(M,E) \to \mathbf{R}, \\ & \phi \quad \to \int_M (j^1\phi)^* \mathcal{L}, \end{aligned}$$

where $\Gamma_c(M,E)$ is the set of compact supported sections of π.

In order to characterize these critical sections, and using the *vertical endomorphism* $\mathcal{V}$ of the bundle J^1E (see the Appendix), we construct the differentiable forms

$$\Theta_{\mathcal{L}} := i(\mathcal{V})\mathcal{L} + \mathcal{L} \in \Omega^m(J^1E) \; ; \quad \Omega_{\mathcal{L}} := -d\Theta_{\mathcal{L}} \in \Omega^{m+1}(J^1E),$$

which are called the *Poincaré–Cartan m* and $(m+1)$-forms associated with the Lagrangian density $\mathcal{L}$. The Lagrangian system is said to be *regular* iff $\Omega_{\mathcal{L}}$ is 1-nondegenerate and, as a consequence, $(J^1E, \Omega_{\mathcal{L}})$ is a multisymplectic manifold.[10]

In a natural chart (x^μ, y^A, v^A_μ) in J^1E, the expressions of the above forms are

$$\begin{aligned} \Theta_{\mathcal{L}} &= \frac{\partial \pounds}{\partial v^A_\mu} dy^A \wedge d^{m-1}x_\mu - \left(\frac{\partial \pounds}{\partial v^A_\mu} v^A_\mu - \pounds \right) d^m x, \\ \Omega_{\mathcal{L}} &= -\frac{\partial^2 \pounds}{\partial v^B_\nu \, \partial v^A_\mu} dv^B_\nu \wedge dy^A \wedge d^{m-1}x_\mu - \frac{\partial^2 \pounds}{\partial y^B \, \partial v^A_\mu} dy^B \wedge dy^A \wedge d^{m-1}x_\mu \\ &\quad + \frac{\partial^2 \pounds}{\partial v^B_\nu \, \partial v^A_\mu} v^A_\mu dv^B_\nu \wedge d^m x + \left(\frac{\partial^2 \pounds}{\partial y^B \, \partial v^A_\mu} v^A_\mu - \frac{\partial \pounds}{\partial y^B} + \frac{\partial^2 \pounds}{\partial x^\mu \, \partial v^B_\mu} \right) dy^B \wedge d^m x \end{aligned} \tag{11}$$

[where $d^{m-1}x_\mu \equiv i(\partial/\partial x^\mu) d^m x$] and the condition of regularity is equivalent to demanding that $\det(\partial^2 \pounds / \partial v^A_\mu \partial v^B_\nu)(\overline{y}) \neq 0$, for every $\overline{y} \in J^1E$.

Then, the (compact-supported) critical sections $\phi: M \to E$ of the variational problem can be characterized as follows.

Proposition 5: Let $((E,M;\pi),\mathcal{L})$ *be a Lagrangian system. The critical sections of the variational problem posed by* $\mathcal{L}$ *are those which satisfy the following equivalent conditions.*

1. $d/dt|_{t=0} \int_M (j^1\phi_t)^* \mathcal{L} = 0$, being $\phi_t = \tau_t \circ \phi$, *where* $\{\tau_t\}$ *is a local one-parameter group of any* π*-vertical and compact supported vector field* $Z \in \mathcal{X}(E)$.
2. $(j^1\phi)^* i(X)\Omega_{\mathcal{L}} = 0$, *for every* $X \in \mathcal{X}(J^1E)$.
3. *The local components of* ϕ *satisfy the Euler–Lagrange equations*:

$$\left(\frac{\partial \pounds}{\partial y^A}\right)_{j^1\phi} - \frac{\partial}{\partial x^\mu}\left(\frac{\partial \pounds}{\partial v^A_\mu}\right)_{j^1\phi} = 0.$$

(See, for instance, Ref. 4 for the proof, and also Refs. 1, 5, 8, 9 as complementary references on all these topics.)

B. The evolution equations in terms of connections and multivector fields

As is well known, the evolution equations in analytical mechanics can be written in a geometric–algebraic manner using vector fields, and then the solutions of the variational problem can be obtained as integral curves of these vector fields. Now we wish to do the same with the evolution equations of field theories, using jet fields (connections) and their equivalent multivector fields, and then obtaining the critical sections as integral sections of these elements.

As a previous step, we must define the action of jet fields on differential forms (see Ref. 4).

Let $\mathcal{Y}: J^1E \to J^1J^1E$ be a jet field. A map $\overline{\mathcal{Y}}: \mathcal{X}(M) \to \mathcal{X}(J^1E)$ can be defined in the following way: let $Z \in \mathcal{X}(M)$, then $\overline{\mathcal{Y}}(Z) \in \mathcal{X}(J^1E)$ is the vector field defined as

$$\overline{\mathcal{Y}}(Z)(\overline{y}) := (T_{\overline{\pi}^1(\overline{y})}\psi)(Z_{\overline{\pi}^1(\overline{y})}),$$

for every $\overline{y} \in J^1E$ and $\psi \in \mathcal{Y}(\overline{y})$. This map is an element of $\Omega^1(M) \otimes_M \mathcal{X}(J^1E)$ and its local expression is

$$\overline{\mathcal{Y}}\left(f^\mu \frac{\partial}{\partial x^\mu}\right) = f^\mu\left(\frac{\partial}{\partial x^\mu} + F^A_\mu \frac{\partial}{\partial y^A} + G^A_{\rho\mu}\frac{\partial}{\partial v^A_\rho}\right).$$

This map induces an action of $\mathcal{Y}$ on $\Omega(J^1E)$. In fact, let $\xi \in \Omega^{m+j}(J^1E)$, and with $j \geqslant 0$, we define $i(\mathcal{Y})\xi: \mathcal{X}(M) \times \overset{(m)}{\ldots} \times \mathcal{X}(M) \to \Omega^j(J^1E)$ given by

$$((i(\mathcal{Y})\xi)(Z_1, \ldots, Z_m))(\overline{y}; X_1, \ldots, X_j) := \xi(\overline{y}; \overline{\mathcal{Y}}(Z_1), \ldots, \overline{\mathcal{Y}}(Z_m), X_1, \ldots, X_j),$$

for $Z_1, \ldots, Z_m \in \mathcal{X}(M)$ and $X_1, \ldots, X_j \in \mathcal{X}(J^1E)$. This is a $C^\infty(M)$-linear and alternate map on the vector fields $Z_1, \ldots, Z_m$. This map $i(\mathcal{Y})\xi$ so-defined, extended by zero to forms of degree $p < m$, is called the *inner contraction* of the jet field $\mathcal{Y}$ and the differential form ξ.

The following result characterizes the action of integrable jet fields on forms (see Ref. 4 for the proof).

Proposition 6: Let $\mathcal{Y}$ be an integrable jet field and $\xi \in \Omega^p(J^1E)$ with $p \geqslant m$. The necessary and sufficient condition for the integral sections of $\mathcal{Y}$, $\psi: M \to J^1E$, to verify the relation $\psi^* i(X)\xi = 0$, for every $X \in \mathcal{X}(J^1E)$, is $i(\mathcal{Y})\xi = 0$.

At this point the following theorem gives the characterization of the critical sections of a Lagrangian field theory in terms of connections (jet fields) and multivector fields.

Theorem 7: *Let $((E, M; \pi), \mathcal{L})$ be a Lagrangian system. The critical sections of the variational problem posed by $\mathcal{L}$ satisfy the following equivalent conditions.*

1. *They are the integral sections of a holonomic jet field $\mathcal{Y}_\mathcal{L}: J^1E \to J^1J^1E$, such that*

$$i(\mathcal{Y}_\mathcal{L})\Omega_\mathcal{L} = 0.$$

2. *They are the integral sections of a holonomic jet field $\mathcal{Y}_\mathcal{L}: J^1E \to J^1J^1E$, such that*

$$i(\mathbf{\nabla}_\mathcal{L})\Omega_\mathcal{L} = (m-1)\Omega_\mathcal{L},$$

   where $\mathbf{\nabla}_\mathcal{L}$ is the associated connection form.

3. *They are the integral sections of a class of holonomic multivector fields $\{X_\mathcal{L}\} \subset \mathcal{X}^m(J^1E)$, such that*

$$i(X_\mathcal{L})\Omega_\mathcal{L} = 0.$$

Proof: These results are consequences of all the above results. In particular, the following holds.

1. It follows from the second item in proposition 5 and proposition 6.
2. See Refs. 13 and 9 for this proof.
3. It is a consequence of the first item and theorem 6.

■

The conditions in this theorem are the version of the Euler–Lagrange equations in terms of jet fields, connection forms and multivector fields, respectively.

C. Analysis of the evolution equations for regular Lagrangians

Using these formulations and, in particular, the multivector fields formalism, it is possible to explore some important properties of the evolution equations in field theories, specifically the existence and multiplicity of solutions. As a first situation, we can consider the case of regular Lagrangian systems.

First, let us recall that the equivalent problem in (time-dependent) mechanics consists in finding vector fields $X_{\mathcal{L}} \in \mathcal{X}(TQ \times \mathbf{R})$ [where TQ is the velocity phase space of the system), such that they verify the dynamical equation $i(X_{\mathcal{L}})\Omega_{\mathcal{L}} = 0$, and they are holonomic vector fields [their integral curves are canonical lifting of curves in Q; that is, sections of the projection $Q \times \mathbf{R} \to \mathbf{R}$]. In this case, the regularity of the Lagrangian function assures the existence and uniqueness of a holonomic vector field which is the solution of the intrinsic equation of motion, with the additional condition $i(X_{\mathcal{L}})dt = 1$.

In field theories we search for (classes of) nonvanishing and locally decomposable multivector fields $X_{\mathcal{L}} \in \mathcal{X}^m(J^1E)$ such that the following occurs.

$$\begin{aligned} &1.\ \text{The equation } i(X_{\mathcal{L}})\Omega_{\mathcal{L}} = 0 \text{ holds.} \\ &2.\ X_{\mathcal{L}} \text{ are SOPDE.} \\ &3.\ X_{\mathcal{L}} \text{ are integrable.} \end{aligned} \tag{12}$$

Remarks:

- If a multivector field satisfies condition 2 and also verifies that $i(X_{\mathcal{L}})(\overline{\pi}^1{}^*\omega) = 1$, then its local expression is (6).
- Let us recall that, if $\{X_{\mathcal{L}}\} \subset \mathcal{X}^m(J^1E)$ is a class of locally decomposable and $\overline{\pi}^1$-transverse multivector fields, then condition 3 is equivalent to demanding that the curvature $\mathcal{R}$ of the connection associated with this class vanishes everywhere.

Following the terminology introduced in Refs. 13 and 14 for connections, we define the following.

Definition 13: $X_{\mathcal{L}} \in \mathcal{X}^m(J^1E)$ is said to be an Euler–Lagrange multivector field *for $\mathcal{L}$ iff it is locally decomposable and verifies conditions 1 and 2 of definition 12.*

Regularity of Lagrangians assures the existence of Euler–Lagrange multivector fields (or connections). In fact we have the following.

Theorem 8: (Existence and local multiplicity of Euler–Lagrange solutions). *Let $((E,M;\pi),\mathcal{L})$ be a regular Lagrangian system.*

1. *There exist classes of locally decomposable and $\overline{\pi}^1$-transverse multivector fields $\{X_{\mathcal{L}}\} \subset \mathcal{X}^m(J^1E)$ (and hence equivalent jet fields $\mathcal{Y}_{\mathcal{L}} : J^1E \to J^1J^1E$ with associated connection forms $\boldsymbol{\nabla}_{\mathcal{L}}$), such that*
2. *(a) the following equivalent conditions hold:*

$$i(X_{\mathcal{L}})\Omega_{\mathcal{L}} = 0\,, \quad i(\mathcal{Y}_{\mathcal{L}})\Omega_{\mathcal{L}} = 0\,, \quad i(\boldsymbol{\nabla}_{\mathcal{L}})\Omega_{\mathcal{L}} = (m-1)\Omega_{\mathcal{L}}\,. \tag{13}$$

(b) $X_{\mathcal{L}} \in \{X_{\mathcal{L}}\}$ are SOPDE (and therefore so is the corresponding $\mathcal{Y}_{\mathcal{L}}$).

In a local system the above solutions depend on $N(m^2-1)$ arbitrary functions.

Proof: In order to prove this, we use the multivector field formalism (for another proof of these statements using the Ehresmann connection formalism see Refs. 13 and 14).

1. First we analyze the local existence of solutions and then their global extension.

In a chart of natural coordinates in J^1E, the expression of $\Omega_{\mathcal{L}}$ is (11), and taking the multi-

vector field given in (6) as the representative of the class $\{X_{\mathcal{L}}\}$, from the relation $i(X_{\mathcal{L}})\Omega_{\mathcal{L}}=0$ we obtain the following conditions.

• The coefficients on dv^A_μ must vanish:

$$0=(F^B_\mu-v^B_\mu)\frac{\partial^2 \pounds}{\partial v^A_\nu\ \partial v^B_\mu} \quad (\text{for every } B,\nu). \tag{14}$$

But, if $\mathcal{L}$ is regular, the matrix $(\partial^2\pounds/\partial v^A_\nu\ \partial v^B_\mu)$ is regular. Therefore $F^B_\mu=v^B_\mu$ (for every B,μ); which proves that if $X_{\mathcal{L}}$ exists it is a SOPDE.

• The coefficients on dy^A must vanish:

$$0=\frac{\partial \pounds}{\partial y^A}-\frac{\partial^2 \pounds}{\partial x^\mu\ \partial v^A_\mu}-\frac{\partial^2 \pounds}{\partial y^B\ \partial v^A_\mu}F^B_\mu-\frac{\partial^2 \pounds}{\partial v^B_\nu\ \partial v^A_\mu}G^B_{\nu\mu}+\frac{\partial^2 \pounds}{\partial y^A\ \partial v^B_\mu}(F^B_\mu-v^B_\mu) \quad (A,B=1,\dots,N), \tag{15}$$

and taking into account that we have obtained $F^B_\mu=v^B_\mu$, these relations lead to

$$\frac{\partial^2 \pounds}{\partial v^B_\nu\ \partial v^A_\mu}G^B_{\nu\mu}=\frac{\partial \pounds}{\partial y^A}-\frac{\partial^2 \pounds}{\partial x^\mu\ \partial v^A_\mu}-\frac{\partial^2 \pounds}{\partial y^B\ \partial v^A_\mu}v^B_\mu \quad (A,B=1,\dots,N), \tag{16}$$

which is a system of N linear equations on the functions $G^B_{\nu\mu}$. This is a compatible system as a consequence of the regularity of $\mathcal{L}$, since the matrix of the coefficients has (constant) rank equal to N (observe that the matrix of this system is obtained as a rearrangement of rows of the Hessian matrix).

• Finally, from the above conditions, we obtain that the coefficients on dx^μ vanish identically.

These results allow us to assure the local existence of (classes of) multivector fields satisfying the desired conditions. The corresponding global solutions are then obtained using a partition of unity subordinated to a covering of J^1E made of local natural charts.

[Observe that, if $j^1\phi=(x^\mu,f^A,\partial f^A/\partial x^\mu)$ is an integral section of $X_{\mathcal{L}}$, then $v^A_\mu=\partial f^A/\partial x^\mu$ and $G^A_{\nu\mu}=\partial^2 f^A/\partial x^\nu\ \partial x^\mu$, and therefore Eqs. (16) are the Euler–Lagrange equations for the section ϕ.]

2. In a chart of natural coordinates in J^1E, the expression of a SOPDE multivector field $X_{\mathcal{L}}\in\{X_{\mathcal{L}}\}$ is given by (8). So, it is determined by the Nm^2 coefficients $G^B_{\nu\mu}$, which are related by the N independent equations (16). Therefore, there are $N(m^2-1)$ arbitrary functions.

■

Comments:

- If, in addition, $\{X_{\mathcal{L}}\}$ is integrable; then every $X_{\mathcal{L}}\in\{X_{\mathcal{L}}\}$ is a holonomic multivector field (and, therefore, $\mathcal{Y}_{\mathcal{L}}$ is also a holonomic jet field).
- As additional conditions on the functions $G^A_{\rho\mu}$ must be imposed in order to assure that $X_{\mathcal{L}}$ is integrable; hence the number of arbitrary functions will be in general less than $N(m^2-1)$.

The last problem consists in studying if it is possible to find a class of integrable Euler–Lagrange multivector fields. This question is posed in the following terms: taking the suitable representatives in each class, the set of Euler–Lagrange multivector fields $X_{\mathcal{L}}$ can be locally written in the form (8), where the coefficients $G^A_{\mu\nu}$ verify Eqs. (16). This is a linear system and, then, if $\mathcal{G}^A_{\mu\nu}$ is a particular solution of this system, we can write

$$X_{\mathcal{L}}=\bigwedge_{\mu=1}^{m}\left(\frac{\partial}{\partial x^\mu}+v^A_\mu\frac{\partial}{\partial y^A}+\mathcal{G}^B_{\mu\gamma}\frac{\partial}{\partial v^B_\gamma}+g^C_{\mu\eta}\frac{\partial}{\partial v^C_\eta}\right),$$

where $g^C_{\eta\mu}$ satisfy the homogeneous linear system,

$$\frac{\partial^2 \pounds}{\partial v^C_\eta\ \partial v^A_\mu}g^C_{\eta\mu}=0 \quad (A=1,\dots,N).$$

As the problem now is to select the integrable Euler–Lagrange multivector fields, from the solutions of this system we can choose those that make the corresponding multivector field $X_{\mathcal{L}}$ verify the integrability condition; that is, the curvature of the associated connection $\nabla_{\mathcal{L}}$ vanishes [Eqs. (10)].

As far as we know, since we have a system of partial differential equations with linear restrictions, there is no way for assuring the existence of an integrable solution, or how to select it; observe that, considering the Euler–Lagrange equations for the coefficients $G^A_{\mu\nu}$ [Eqs. (16)], together with the integrability conditions [Eqs. (10)], we have $N+\frac{1}{2}Nm(m-1)$ linear equations and $\frac{1}{2}Nm^2(m-1)$ partial differential equations. Then, if the set of linear restrictions [made of the Eqs. (16) and the first group of Eqs. (10)] allow us to isolate $N+\frac{1}{2}Nm(m-1)$ coefficients $G^A_{\mu\nu}$ as functions on the remaining ones; and the set of $\frac{1}{2}Nm^2(m-1)$ partial differential equations [the second group of equations (10)] on these remaining coefficients satisfies the conditions on the *Cauchy–Kowalewska theorem*,[23] then the existence of integrable Euler–Lagrange multivector fields is assured (see the example in Sec. III F).

On the other hand, if we can eventually select some particular Euler–Lagrange multivector field solution, then the integrability algorithm developed in Sec. II B can be applied in order to find a submanifold where this multivector field is integrable (if it exists).

D. Discussion on the evolution equations for singular Lagrangian field theories

Next we discuss the case of singular Lagrangian systems ($\Omega_{\mathcal{L}}$ is 1-degenerate). As in the regular case, the existence of multivector fields (or connections) verifying conditions (12) is not assured. The difference with the regular case lies in the following facts.

1. A locally decomposable and $\overline{\pi^1}$-transverse multivector field $X\in\mathcal{X}^m(J^1E)$ verifying condition 1 on J^1E (the evolution equation) does not necessarily exist.
2. If such a multivector field exists, it does not necessarily hold condition 2 (the SOPDE condition).

As in the regular case, the existence of Euler–Lagrange multivector fields does not imply their integrability. Nevertheless, it is possible for these integrable multivector fields to exist on a submanifold of J^1E. So we can state the following problem: to look for a submanifold $S\hookrightarrow J^1E$ where integrable Euler–Lagrange multivector fields $X_{\mathcal{L}}\in\mathcal{X}^m(J^1E)$ exist; and then their integral sections are contained in S. Observe that these conditions imply that $\overline{\pi^1}|_S:S\to M$ is onto on M, because $X_{\mathcal{L}}$ is SOPDE.

As a first step, we do not consider the integrability condition. [In this way, the situation is quite analogous to that in (time-dependent) mechanics.[24,25]]

The procedure is algorithmic (from now on we suppose that all the multivector fields are locally decomposable).

- First, let S_1 be the set of points of J^1E where Euler–Lagrange multivector fields do exist; that is,

$$S_1:=\left\{\overline{y}\in J^1E;\ \exists X_{\mathcal{L}}\in\mathcal{X}^m(J^1E)\ \text{ such that }\left\{\begin{array}{c}(i(X_{\mathcal{L}})\Omega_{\mathcal{L}})(\overline{y})=0\\(i(X_{\mathcal{L}})(\overline{\pi^1}{}^*\omega))(\overline{y})=1\\X_{\mathcal{L}}\ \text{is a SOPDE at}\ \overline{y}\end{array}\right\}\right\}.$$

We assume that S_1 is a nonempty (closed) submanifold of J^1E.

This is the *compatibility condition*.

- Now, denote by $\mathcal{X}^m_{\mathcal{L}}(J^1E,S_1)$ the set of multivector fields in J^1E which are Euler–Lagrange on S_1. Let $X_{\mathcal{L}}:S_1\to\Lambda^m TJ^1E|_{S_1}$ be in $\mathcal{X}^m_{\mathcal{L}}(J^1E,S_1)$. If, in addition, $X_{\mathcal{L}}:S_1\to\Lambda^m TS_1$; that is, $X_{\mathcal{L}}\in\mathcal{X}^m(S_1)$, then we say that $X_{\mathcal{L}}$ is a solution on S_1. Nevertheless, this last condition is not assured except perhaps in a set of points $S_2\subset S_1\subset J^1E$, which we will assume to be a (closed) submanifold, and which is defined by

$$S_2:=\{\overline{y}\in S_1;\ \exists X_{\mathcal{L}}\in\mathcal{X}^m_{\mathcal{L}}(J^1E,S_1)\ \text{ such that }\ X_{\mathcal{L}}(\overline{y})\in\Lambda^m T_{\overline{y}}S_1\}.$$

This is the so-called *consistency* or *tangency condition*.

- In this way, a sequence of (closed) submanifolds, $\ldots \subset S_i \subset \cdots \subset S_1 \subset J^1E$, is assumed to be obtained, each one of them being defined as

$$S_i := \{\overline{y} \in S_{i-1};\ \exists X_{\mathcal{L}} \in \mathcal{X}^m_{\mathcal{L}}(J^1E, S_{i-1}) \ \text{ such that } \ X_{\mathcal{L}}(\overline{y}) \in \Lambda^m T_{\overline{y}} S_{i-1}\}.$$

It is important to point out that, for every submanifold S_i of this sequence, $\overline{\pi}^1|_{S_i}: S_i \to M$ must be on M.

- There are two possible options for the final step of this algorithm, namely the following.

1. The algorithm ends by giving a submanifold $S_f \hookrightarrow J^1E$, with $\dim S_f \geqslant m$ (where $S_f = \cap_{i \geqslant 1} S_i$) and Euler–Lagrange multivector fields $X_{\mathcal{L}} \in \mathcal{X}^m(S_f)$. S_f is then called the *final constraint submanifold*.
2. The algorithm ends by giving a submanifold S_f with $\dim S_f < m$, or the empty set. Then there is no Euler–Lagrange multivector fields $X_{\mathcal{L}} \in \mathcal{X}^m(S_f)$. In this case the Lagrangian system has no solution.

Comment:

The problem considered here can also be treated in an alternative manner by splitting the procedure into two steps (see Ref. 13 for a more detailed discussion on this method, using Ehresmann connections).

1. First, searching for a submanifold P_f of J^1E and locally decomposable multivector fields $X_{\mathcal{L}} \in \mathcal{X}^m(J^1E)$, such that

(a) $(i(X_{\mathcal{L}})\Omega_{\mathcal{L}})|_{P_f} = 0$.

(b) $(i(X_{\mathcal{L}})(\overline{\pi}^{1*}\omega))|_{P_f} = 1$.

(c) $X_{\mathcal{L}}: P_f \to \Lambda^m T P_f$.

This procedure is called the *constraint algorithm*.

2. Second, searching for a submanifold S_f of P_f and locally decomposable multivector fields $X_{\mathcal{L}} \in \mathcal{X}^m(J^1E)$, such that

(a) $(i(X_{\mathcal{L}})\Omega_{\mathcal{L}})|_{S_f} = 0$.

(b) $X_{\mathcal{L}}$ is a SOPDE on S_f.

(c) $X_{\mathcal{L}}: S_f \to \Lambda^m T S_f$.

It is then clear that, in some particular cases, locally decomposable and $\overline{\pi}^1$-transverse multivector fields, solutions of the evolution equations on some submanifold P_f, can exist, but none of them being SOPDE at any point belonging to P_f.

The local treatment of the singular case shows significant differences to the regular one. In fact, although the starting equations are the same in both cases [Eqs. (14) and (15)]; since the matrix $(\partial^2 \pounds / \partial v^A_\nu \, \partial v^B_\mu)$ is not regular, the first group does not imply the SOPDE condition. Nevertheless, we can impose this condition by making $F^A_\mu = v^A_\mu$, for every A, μ. Therefore, we have the system of equations for the coefficients $G^A_{\mu\nu}$ which is (16) again. As we have stated, this system is not compatible in general, and S_1 is the closed submanifold where the system is compatible. Then, there are Euler–Lagrange multivector fields on S_1, but the number of arbitrary functions on which they depend is not the same as in the regular case, since it depends on the dimension of S_1 and the rank of the Hessian matrix of $\pounds$. Now the tangency condition must be analyzed in the same way as in the algorithm of Sec. II B.

Finally, the question of integrability must be considered. We make the same remarks as at the end of the above section, but for the submanifold S_f instead of J^1E.

E. Noether's theorem for multivector fields

Next we recover the idea of *conserved quantity* and state Noether's theorem in terms of multivector fields. First we need a previous lemma.

Lemma 1: Let $Y \in \mathcal{X}^m(J^1E)$ be a holonomic multivector field. If $\zeta \in \Omega^m(J^1E)$ belongs to the ideal generated by the contact module [this ideal is denoted $\mathcal{I}(\mathcal{M}_c)$], then $i(Y)\zeta = 0$.

Proof: Consider $\overline{y}\in J^1E$, with $\overline{\pi}^1(\overline{y})=x$, and a section $\phi:M\to E$ such that $j^1\phi$ is an integral section of Y passing through $\overline{y}$; that is $j^1\phi(x)=\overline{y}$. Then, as Y is SOPDE, there exists $f\in C^\infty(J^1E)$ with

$$(\Lambda^m T_x j^1\phi)\left(\frac{\partial}{\partial x^1}\wedge\cdots\wedge\frac{\partial}{\partial x^m}\right)=(fY)\overline{_{y}},$$

and therefore

$$\begin{aligned}(i(Y)\zeta)(\overline{y})&=(i(Y)\zeta)(j^1\phi(x))\\&=i(Y_{j^1\phi(x)})(\zeta(j^1\phi(x)))\\&=\frac{1}{f(\overline{y})}i\left((\Lambda^m T_x j^1\phi)\left(\frac{\partial}{\partial x^1}\wedge\cdots\wedge\frac{\partial}{\partial x^m}\right)\right)(\zeta(j^1\phi(x)))\\&=\frac{1}{f(\overline{y})}i\left(\frac{\partial}{\partial x^1}\wedge\cdots\wedge\frac{\partial}{\partial x^m}\right)((j^1\phi)^*\zeta)(x))=0,\end{aligned}$$

because ζ belongs to the ideal generated by the contact module and then $(j^1\phi)^*\zeta=0$. ■

Now we are able to prove the following.

Theorem 9 (Noether): *Let $((E,M;\pi),\mathcal{L})$ be a Lagrangian system and $X_{\mathcal{L}}\in\mathcal{X}^m(J^1E)$ an integrable Euler–Lagrange multivector field. Let $X\in\mathcal{X}(J^1E)$ be a vector field satisfying the following conditions.*

1. *X* preserves the contact module: $L(X)\mathcal{M}_c\subset\mathcal{M}_c$.
2. There exist $\xi\in\Omega^{m-1}(J^1E)$, and $\alpha\in\mathcal{I}(\mathcal{M}_c)\cap\Omega^m(J^1E)$ such that $L(X)\mathcal{L}=d\xi+\alpha$.

Then $i(X_{\mathcal{L}})d(\xi-i(X)\Theta_{\mathcal{L}})=0$.

Proof: First we have that

$$L(X)\Theta_{\mathcal{L}}=i(X)d\Theta_{\mathcal{L}}+di(X)\Theta_{\mathcal{L}},$$

however, as $\Theta_{\mathcal{L}}:=i(\mathcal{V})\mathcal{L}+\mathcal{L}\equiv\vartheta_{\mathcal{L}}+\mathcal{L}$,

$$L(X)\Theta_{\mathcal{L}}=L(X)(\vartheta_{\mathcal{L}}+\mathcal{L})=L(X)\vartheta_{\mathcal{L}}+d\xi+\alpha\ ,$$

and $L(X)\vartheta_{\mathcal{L}}\in\mathcal{I}(\mathcal{M}_c)$, because $\vartheta_{\mathcal{L}}\in\mathcal{I}(\mathcal{M}_c)$. So we obtain

$$i(X)d\Theta_{\mathcal{L}}+di(X)\Theta_{\mathcal{L}}=d\xi+\zeta,$$

with $\zeta=\alpha+L(X)\vartheta_{\mathcal{L}}\in\mathcal{I}(\mathcal{M}_c)$. Then

$$d(\xi-i(X)\Theta_{\mathcal{L}})=i(X)d\Theta_{\mathcal{L}}-\zeta,$$

and therefore

$$i(X_{\mathcal{L}})d(\xi-i(X)\Theta_{\mathcal{L}})=i(X_{\mathcal{L}})i(X)d\Theta_{\mathcal{L}}+i(X_{\mathcal{L}})\zeta=0,$$

since $i(X_{\mathcal{L}})i(X)\Omega_{\mathcal{L}}=0$, for every $X\in\mathcal{X}(J^1E)$, because $X_{\mathcal{L}}$ is a solution of the Euler–Lagrange equations, and $i(X_{\mathcal{L}})\zeta=0$, by the previous lemma. ■

*Definition 14: Let $((E,M;\pi),\mathcal{L})$ be a Lagrangian system. An (*infinitesimal) generalized symmetry *of the Lagrangian system is a vector field $X\in\mathcal{X}(J^1E)$ such that it verifies conditions 1 and 2 of the above theorem.*

In particular, X is said to be an infinitesimal natural symmetry *if it is the canonical prolongation of a vector field $Z\in\mathcal{X}(E)$.*

Remark: As a special case, if $L(X)\mathcal{L}=0$, then $i(X_{\mathcal{L}})di(X)\Theta_{\mathcal{L}}=0$. These are the so-called *symmetries of the Lagrangian*.

An immediate consequence of the above theorem is the following fact: if $\alpha=\xi-i(X)\Theta_{\mathcal{L}} \in \Omega^{m-1}(J^1E)$, then $i(X_{\mathcal{L}})\alpha=0$, and hence, if $J:N\hookrightarrow J^1E$ is an integral submanifold of $X_{\mathcal{L}}$, we have $J^*d\alpha=0$ and then $dJ^*\alpha=0$. But the integral submanifolds of $X_{\mathcal{L}}$ are images of canonical lifting of sections, $j^1\phi:M\to J^1E$; therefore $0=(j^1\phi)^*d\alpha=d(j^1\phi)^*\alpha$; that is the following.

Theorem 10 (Noether): *Let X be an infinitesimal generalized symmetry of the system $((E,M;\pi),\mathcal{L})$. Then, the $(m-1)$-form $\xi-i(X)\Theta_{\mathcal{L}}$ is closed on the critical sections of the variational problem posed by $\mathcal{L}$.*

F. Example

(See also Ref. 26.)

Most of the (quadratic) Lagrangian systems in field theories can be modeled as follows: $\pi:E\to M$ is a trivial bundle (usually $E=M\times\mathbf{R}^N$) and then $\pi^1:J^1E\to E$ is a vector bundle. g is a metric in this vector bundle, Γ is a connection of the projection π^1 and $f\in C^\infty(E)$ is a potential function. Then the Lagrangian function is

$$\pounds(\overline{y})=\tfrac{1}{2}g(\overline{y}-\Gamma(\pi^1(\overline{y})),\overline{y}-\Gamma(\pi^1(\overline{y})))+\pi^{1*}f(\overline{y}),$$

and in natural coordinates takes the form[26]

$$\pounds=\tfrac{1}{2}a^{\mu\nu}_{AB}(y)(v^A_\mu-\Gamma^A_\mu(x))(v^B_\nu-\Gamma^B_\nu(x))+f(y).$$

In order to apply the formalism which has been developed in this work, we consider a model where the matrix of the coefficients $a^{\mu\nu}_{AB}$ is regular and symmetric at every point [that is, $a^{\mu\nu}_{AB}(y)=a^{\nu\mu}_{BA}(y)$]. This fact is equivalent to the nondegeneracy of the metric g.

The associated Poincaré–Cartan $(m+1)$-form is then

$$\begin{aligned}\Omega_{\mathcal{L}}=&-a^{\nu\mu}_{BA}\,dv^B_\nu\wedge dy^A\wedge d^{m-1}x_\mu\\&-\frac{\partial a^{\nu\mu}_{CA}}{\partial y^B}(v^C_\nu-\Gamma^C_\nu)dy^B\wedge dy^A\wedge d^{m-1}x_\mu+a^{\nu\mu}_{BA}v^A_\mu\,dv^B_\nu\wedge d^mx\\&+\left(\frac{\partial a^{\nu\mu}_{CA}}{\partial y^B}(v^C_\nu-\Gamma^C_\nu)v^A_\mu-\frac{\partial a^{\nu\mu}_{CA}}{\partial y^B}(v^C_\nu-\Gamma^C_\nu)-\frac{\partial f}{\partial y^B}-a^{\mu\nu}_{AB}\frac{\partial\Gamma^A_\nu}{\partial x^\mu}\right)dy^B\wedge d^mx,\end{aligned}$$

and the system is obviously regular. Then, taking (8) as the local expression for representatives of the corresponding classes of Euler–Lagrange multivector fields $\{X_{\mathcal{L}}\}\subset\mathcal{X}^m(J^1E)$, their components $G_{\mu\nu}$ are related by Eqs. (16), which in this case are

$$a^{\nu\mu}_{BA}G^B_{\nu\mu}=\frac{1}{2}\frac{\partial a^{\nu\mu}_{CB}}{\partial y^A}(v^C_\nu-\Gamma^C_\nu)(v^B_\mu-\Gamma^B_\mu)+\frac{\partial f}{\partial y^A}+a^{\mu\nu}_{BA}\frac{\partial\Gamma^B_\nu}{\partial x^\mu}-\frac{\partial a^{\nu\mu}_{CA}}{\partial y^B}v^B_\mu(v^C_\nu-\Gamma^C_\nu).\tag{17}$$

This system allows us to isolate N of these components as functions of the remaining $N(m^2-1)$; and then it determines a family of (classes of) Euler–Lagrange multivector fields. In order to obtain an integrable class, the condition of integrability $\mathcal{R}=0$ (where $\mathcal{R}$ is the curvature of the associated connection) must hold; that is, Eqs. (10) must be added to the last system.

A simpler case is obtained when Γ is an integrable connection. Then there exist local natural charts in J^1E such that $\Gamma^A_\mu=0$;[3] and then $\pounds=\frac{1}{2}a^{\mu\nu}_{AB}(y)v^A_\mu v^B_\nu+f(y)$. If, in addition, we consider that the matrix of coefficients is $a^{\mu\nu}_{AB}(y)=\delta_{AB}\delta^{\mu\nu}$ (that is, we take an orthonormal frame for the metric g), then we have that

$$\pounds=\tfrac{1}{2}\,\delta_{AB}\delta^{\mu\nu}v^A_\mu v^B_\nu+f(y).$$

In this case Eqs. (17) reduce to

$$\delta_{BA}\delta^{\nu\mu}G^B_{\nu\mu}=\frac{\partial f}{\partial y^A}\,.$$

From this system we can isolate N of the coefficients $G^A_{\mu\nu}$; for instance, if $\mu,\nu=0,1,\ldots,m-1$, those for which $\mu=\nu=0$: Thus

$$\delta_{AB}G^B_{00}=\frac{\partial f}{\partial y^A}-\sum_{\mu=1}^{m-1}\delta_{AB}G^B_{\mu\mu}.$$

Therefore the Euler–Lagrange multivector fields are

$$X_{\mathcal{L}}=\bigwedge_{\mu=0}^{m-1}\left(\frac{\partial}{\partial x^\mu}+v^A_\mu\frac{\partial}{\partial y^A}+\delta^0_\mu\left[\frac{\partial f}{\partial y^A}-\sum_{\nu=1}^{m-1}\delta_{AB}G^B_{\nu\nu}\right]\frac{\partial}{\partial v^A_0}+\sum_{\mu=\gamma\neq 0}G^B_{\mu\gamma}\frac{\partial}{\partial v^B_\gamma}+\sum_{\mu\neq\gamma}G^C_{\mu\gamma}\frac{\partial}{\partial v^C_\gamma}\right).$$

Now, if we look for integrable Euler–Lagrange multivector fields, the integrability conditions (10) must be imposed.

As a particular case, we can consider this last model with dim $M=2$. Equations (16) are

$$\delta^{BA}\frac{\partial f}{\partial y^A}-G^B_{00}-G^B_{11}=0,$$

and the expression of the Euler–Lagrange multivector fields will then be

$$X_{\mathcal{L}}=\left(\frac{\partial}{\partial x^0}+v^A_0\frac{\partial}{\partial y^A}+G^A_{00}\frac{\partial}{\partial v^A_0}+G^A_{01}\frac{\partial}{\partial v^A_1}\right)$$
$$\wedge\left(\frac{\partial}{\partial x^1}+v^A_1\frac{\partial}{\partial y^A}+G^A_{10}\frac{\partial}{\partial v^A_0}+\left[G^A_{00}-\delta^{AB}\frac{\partial f}{\partial y^B}\right]\frac{\partial}{\partial v^A_1}\right).$$

The integrability conditions (10) are

$$0=G^B_{01}-G^B_{10},$$

$$0=\frac{\partial G^B_{10}}{\partial x^0}+v^A_0\frac{\partial G^B_{10}}{\partial y^A}+G^A_{00}\frac{\partial G^B_{10}}{\partial v^A_0}+G^A_{01}\frac{\partial G^B_{10}}{\partial v^A_1}$$
$$-\frac{\partial G^B_{00}}{\partial x^1}-v^A_1\frac{\partial G^B_{00}}{\partial y^A}-G^A_{10}\frac{\partial G^B_{00}}{\partial v^A_0}-\left(G^A_{00}-\delta^{AC}\frac{\partial f}{\partial y^C}\right)\frac{\partial G^B_{00}}{\partial v^A_1},$$

$$0=\frac{\partial G^B_{00}}{\partial x^0}+v^A_0\frac{\partial G^B_{00}}{\partial y^A}-v^A_0\frac{\partial^2 f}{\partial y^{A2}}+G^A_{00}\frac{\partial G^B_{00}}{\partial v^A_0}+G^A_{01}\frac{\partial G^B_{00}}{\partial v^A_1}$$
$$-\frac{\partial G^B_{01}}{\partial x^1}-v^A_1\frac{\partial G^B_{01}}{\partial y^A}-G^A_{10}\frac{\partial G^B_{01}}{\partial v^A_0}-\left(G^A_{00}-\delta^{AC}\frac{\partial f}{\partial y^C}\right)\frac{\partial G^B_{01}}{\partial v^A_1}.$$

The first relation imposes $G^B_{01}=G^B_{10}$, and then the others are a system of $2N$ partial differential equations on the functions G^B_{00} and G^B_{01}. For this system the hypothesis of the Cauchy–Kowalewska theorem holds,[23] and hence there exist integrable multivector fields solutions of the Euler–Lagrange equations for $\mathcal{L}$.

IV. CONCLUSIONS AND OUTLOOK

We have studied the integrability of multivector fields in a differentiable manifold and the relation between integrable jet fields and multivector fields in jet bundles, using this to give alternative geometric formulations of the Lagrangian formalism of classical field theories (of first order). In particular we have the following.

• We have introduced the notion of *integrable multivector fields* in a manifold E, first proving that every orientable distribution in E is equivalent to a class of *nonvanishing, locally decomposable* multivector fields, and then demanding that this distribution be integrable. The idea of *multiflow* for a certain kind of integrable multivector fields (which we call *dynamical*) is studied, showing in addition that from every class of integrable multivector fields we can select a representative which is dynamical.

In addition, an algorithmic procedure for finding the (maximal) submanifold of E where a locally decomposable multivector field is integrable (if it exists) is also developed.

The results obtained can be summarized in the following table:

$$\begin{array}{ccccc} \text{Integrable} & & \text{Integrable m.v.f.} & & \left\{\begin{array}{l}\left\{\begin{array}{l}\text{Non}-\text{van. m.v.f.}\\ \text{Loc. dec. m.v.f.}\end{array}\right\}(class)\\ \text{Involutive m.v.f. } (class)\end{array}\right. \\ \text{Orientable} & \Leftrightarrow & (class) & \Leftrightarrow & \text{Orientable} \\ \text{distribution} & & \downarrow & & \text{distribution} \\ & & \text{Dynamical m.v.f.} & & \\ & & (representative) & & \end{array}$$

• We have established the relation between integrable jet fields in J^1E and a certain kind of integrable m-multivector fields in E, by proving that every orientable jet field Ψ is equivalent to a class of nonvanishing m-multivector fields $\{Y\}$, and conversely, in such a way that their associated distributions $\mathcal{D}(\Psi)$ and $\mathcal{D}(Y)$ are the same. Of course, in order to make this equivalence consistent, it is necessary for the distribution $\mathcal{D}(Y)$ to be transverse to the projection $\pi: E \to M$. So we have

$$\left.\begin{array}{c}\text{Transverse m.v.f.}\\ \text{Locally decomp. m.v.f.}\end{array}\right\}(class) \Leftrightarrow \text{Orientable j.f.}$$

• We have applied these ideas to prove that every orientable jet field $\mathcal{Y}$ in J^1J^1E is equivalent to a class of locally decomposable and nonvanishing m-multivector fields $\{X\}$ in J^1E, and conversely. In this case, the consistency of this equivalence requires the distribution $\mathcal{D}(X)$ to be transverse to the projection $\overline{\pi}^1: J^1E \to M$.

The integrability and the SOPDE condition for multivector fields in this framework are then discussed, giving several equivalent characterizations for SOPDE multivector fields, which lead finally to establishing the equivalence between classes of holonomic multivector fields and orientable, holonomic jet fields. So we have the following summarizing scheme:

$$\begin{array}{ccccc} \text{Holonomic m.v.f} & & \left\{\begin{array}{l}\left\{\begin{array}{c}\overline{\pi}^1-\text{Transverse m.v.f.}\\ \text{Involutive m.v.f.}\end{array}\right\}(class) \quad \Leftrightarrow \quad \left\{\begin{array}{l}\text{Orientable j.f.}\\ \text{Integrable j.f.}\end{array}\right.\\ \left\{\begin{array}{c}\text{Loc. decom. m.v.f.}\\ \text{SOPDE m.v.f.}\end{array}\right\}(class) \quad \Leftrightarrow \quad \left\{\begin{array}{l}\text{SOPDE j.f.}\\ \text{Orientable j.f.}\end{array}\right.\end{array}\right\} \text{Holonomic} \\ (class) & & & & \text{j.f.} \end{array}$$

• We prove that the evolution equations for first-order classical field theories in the Lagrangian formalism (*Euler–Lagrange equations*) can be written in three equivalent geometric ways: using multivector fields in J^1E, jet fields in J^1J^1E or their associated Ehresmann connections in J^1E. These descriptions allow us to write the evolution equations for field theories in an analogous way to the dynamical equations for (time-dependent) mechanical systems.

• Using the formalism with multivector fields, we show that the evolution equations can have no integrable solutions in J^1E, for neither regular nor singular Lagrangian systems.

In the regular case, Euler–Lagrange multivector fields (that is, SOPDE and the solution of the Lagrangian evolution equations) always exist; but they are not necessarily integrable. In the singular case, not even the existence of such an Euler–Lagrange multivector field is assured. In both cases, the multivector field solution (if it exists) is not unique. These features are significant differences in relation to the analogous situation in mechanics.

• With regard to these facts, for the singular case we outline an algorithmic procedure in order to obtain the submanifold of J^1E where Euler–Lagrange multivector fields exist.

Further research on these latter topics will be carried out in the future.

• A version of Noether's theorem using multivector fields is also proved.

In a forthcoming paper devoted to the Hamiltonian formalism of field theories, we will extend these results to the dual bundle of J^1E.

ACKNOWLEDGMENTS

We wish to thank Mr. Jeff Palmer for his assistance in preparing the English version of the manuscript. We thank Prof. I. V. Kanatchikov for his helpful comments. We are also grateful for the financial support of the CICYT TAP97-0969-C03-01.

APPENDIX: CANONICAL STRUCTURES OF A JET BUNDLE

(Following Ref. 5. See also Ref. 4.)

Consider a first-order jet bundle $J^1E \overset{\pi^1}{\to} E \overset{\pi}{\to} M$. Let $\phi: M \to E$ be a section of π, $x \in M$ and $y = \phi(x)$. The *vertical differential* of the section ϕ at the point $y \in E$ is the map

$$\begin{aligned} d^v_y\phi \quad &: T_yE \to \quad V_y(\pi), \\ &\quad u \ \to u - T_y(\phi \circ \pi)u. \end{aligned}$$

If (x^μ, y^A) is a natural local system of E and $\phi = (x^\mu, \phi^A(x^\mu))$, then

$$d^v_y\phi\left(\frac{\partial}{\partial x^\mu}\right)_y = -\left(\frac{\partial \phi^A}{\partial x^\mu}\right)_y \left(\frac{\partial}{\partial y^A}\right)_y, \quad d^v_y\phi\left(\frac{\partial}{\partial y^A}\right)_y = \left(\frac{\partial}{\partial y^A}\right)_y.$$

Observe that the vertical differential splits T_yE into a vertical component and another which is tangent to the image of ϕ at the point y. In addition, $d^v_y\phi(u)$ is the projection of u on the vertical part of this splitting.

Definition 15: Consider $\overline{y} \in J^1E$ *with* $\overline{y} \overset{\overline{\pi^1}}{\mapsto} y \mapsto {}^{\pi}x$ *and* $\overline{u} \in T_{\overline{y}}J^1E$. *The* structure canonical form *of* J^1E *is a* 1*-form* θ *in* J^1E *with values on* ${\pi^1}^*V(\pi)$ *which is defined by*

$$\theta(\overline{y};\overline{u}) := (d^v_y\phi)(T_{\overline{y}}\pi^1(\overline{u})) = (T_{\overline{y}}\pi^1 - T_{\overline{y}}(\phi \circ \overline{\pi^1}))(\overline{u}),$$

where the section ϕ *is a representative of* $\overline{y}$.

Definition 16: Let θ *be the structure canonical form of* J^1E. *It can be considered as a* $C^\infty(J^1E)$*-linear map* $\theta: \Gamma(J^1E, {\pi^1}^*V(\pi))^* \to \Omega^1(J^1E)$. *The image of this map is called the* contact module *or* Cartan distribution *of* J^1E. *It is denoted by* $\mathcal{M}_c$.

The expression of θ in a natural local system of J^1E is $\theta = (dy^A - v^A_\mu\, dx^\mu) \otimes \partial/\partial y^A$. Then $\mathcal{M}_c$ is generated by the forms $\theta^A = dy^A - v^A_\mu\, dx^\mu$ in this open set.

If $\phi: M \to E$ is a section of π, we denote by $j^1\phi: M \to J^1E$ its canonical prolongation. Then, a section $\psi: M \to J^1E$ is said to be *holonomic* iff $\psi = j^1\phi$, for some ϕ. Holonomic sections can be characterized using the structure canonical form or, equivalently, the contact module, as follows.

Proposition 7: Let $\psi: M \to J^1E$ *be a section of* $\overline{\pi^1}$. *The following assertions are equivalent.*

1. ψ *is a holonomic section.*
2. $\psi^*\theta = 0$.
3. $\psi^*\alpha = 0$, *for every* $\alpha \in \mathcal{M}_c$.

Definition 17: 1. For every $\overline{y} \in J^1E$, *consider the canonical isomorphism,*

$$\mathcal{S}_{\overline{y}}: T^*_{\overline{\pi^1}(\overline{y})}M \otimes V_{\pi^1(\overline{y})}(\pi) \to V_{\overline{y}}(\pi^1),$$

which consists of associating with an element $\alpha \otimes v \in T^*_{\overline{\pi^1}(\overline{y})}M \otimes V_{\pi^1(\overline{y})}(\pi)$ *the directional derivative in* $\overline{y}$ *with respect to* $\alpha \otimes v$. *Taking into account that* $\alpha \otimes v$ *acts in* $J^1_{\overline{y}}E$ *by translation,*

$$\mathcal{S}_{\overline{y}}(\alpha\otimes v):=D_{\alpha\otimes v}(\overline{y}):f\mapsto\lim_{t\to 0}\frac{f(\overline{y}+t(\alpha\otimes v))-f(\overline{y})}{t},$$

for $f\in \mathrm{C}^\infty(J^1_yE)$. *Then we have the following isomorphism of* $C^\infty(J^1E)$*-modules:*

$$\mathcal{S}:\Gamma(J^1E,\overline{\pi^1}^*T^*M)\otimes\Gamma(J^1E,\pi^{1*}V(\pi))\to\Gamma(J^1E,V(\pi^1)),$$

which is called the vertical endomorphism $\mathcal{S}$.

2. *As* $\mathcal{S}\in\Gamma(J^1E,(\pi^{1*}V(\pi))^*)\otimes\Gamma(J^1E,V(\pi^1))\otimes\Gamma(J^1E,\overline{\pi^1}^*TM)$ [*where all the tensor products are on* $\mathrm{C}^\infty(J^1E)$]*; the* vertical endomorphism $\mathcal{V}$ *arises from the natural contraction between the factors* $\Gamma(J^1E,(\pi^{1*}V(\pi))^*)$ *of* $\mathcal{S}$ *and* $\Gamma(J^1E,\pi^{1*}V(\pi))$ *of* θ*; that is*

$$\mathcal{V}=i(\mathcal{S})\theta\in\Omega^1(J^1E)\otimes\Gamma(J^1E,V(\pi^1))\otimes\Gamma(J^1E,\overline{\pi^1}^*TM).$$

If (x^μ,y^A,v^A_μ) is a natural system of coordinates, the local expression of $\mathcal{S}$ and $\mathcal{V}$ are given in this local system by

$$\mathcal{S}=\xi^A\otimes\frac{\partial}{\partial v^A_\mu}\otimes\frac{\partial}{\partial x^\mu};\quad \mathcal{V}=(dy^A-v^A_\mu\,dx^\mu)\otimes\frac{\partial}{\partial v^A_\nu}\otimes\frac{\partial}{\partial x^\nu},$$

where $\{\xi^A\}$ is the local basis of $\Gamma(J^1E,\pi^{1*}V(\pi))^*$ which is dual of $\{\partial/\partial y^A\}$.

[1] E. Binz, J. Sniatycki, and H. Fisher, *The Geometry of Classical Fields* (North-Holland, Amsterdam, 1988).
[2] J. F. Cariñena, M. Crampin, and L. A. Ibort, ''On the multisymplectic formalism for first order field theories,'' Diff. Geom. Appl. **1**, 345–374 (1991).
[3] A. Echeverría-Enríquez and M. C. Muñoz-Lecanda, ''Variational calculus in several variables: A Hamiltonian approach,'' Ann. Inst. Henri Poincaré **56**, 27–47 (1992).
[4] A. Echeverría-Enríquez, M. C. Muñoz-Lecanda, and N. Roman-Roy, ''Geometry of Lagrangian first-order classical field theories,'' Fortschr. Phys. **44**, 235–280 (1996).
[5] P. L. García, ''The Poincaré–Cartan invariant in the calculus of variations,'' in Symposia Mathematica Vol. XIV, Convegro di Geometria Simplettica e Fisica Matematica, INDAM, Rome, 1973 (Academic Press, London, 1974), p. 203–267.
[6] H. Goldschmidt and S. Sternberg, ''The Hamilton–Cartan formalism in the calculus of variations,'' Ann. Inst. Fourier Grenoble **23**, 203–267 (1973).
[7] M. J. Gotay, J. Isenberg, J. E. Marsden, R. Montgomery, J. Śniatycki, and P. B. Yasskin, ''Momentum maps and classical relativistic fields,'' GIMMSY, 1990.
[8] G. Sardanashvily, ''Five lectures on jet manifold methods in field theory,'' preprint hep-th/9411089, 1994.
[9] D. J. Saunders, ''The Cartan form in Lagrangian field theories,'' J. Phys. A **20**, 333–349 (1987).
[10] F. Cantrijn, L. A. Ibort, and M. de León, ''On the geometry of multisymplectic manifolds,'' to appear in Austr. J. Math.
[11] F. Cantrijn, L. A. Ibort, and M. De León, ''Hamiltonian structures on multisymplectic manifolds,'' Rnd. Sem. Math. Univ. Pol. Torino **54**, 225–236 (1996).
[12] L. A. Ibort, A. Echeverría-Enríquez, M. C. Muñoz-Lecanda, and N. Román-Roy, ''Invariant forms and automorphisms of multisymplectic manifolds,'' preprint, 1997.
[13] M. de León, J. Marín-Solano, and J. C. Marrero, ''Ehresmann connections in classical field theories,'' *Proceedings of the III Fall Workshop: Differential Geometry and its Applications*, Anales de Física, Monografías 1995, Vol. 2, pp. 73–89.
[14] D. J. Saunders, *The Geometry of Jet Bundles*, London Mathematica Society Lecture Notes, Ser. **142** (Cambridge, University Press, Cambridge, 1989).
[15] I. Kanatchikov, ''On the canonical structure of the De Donder-Weyl covariant Hamiltonian formulation of field theory I,'' Preprint PITHA 93/41, hep-th/9312162 (1993).
[16] I. Kanatchikov, ''From the Poincaré-Cartan form to a Gerstenhaber algebra of the Poisson brackets in field theory,'' *Proceedings of the XIII-th International Workshop on Geometric Methods in Physics* (Plenum, New York, 1995).
[17] I. Kanatchikov, ''Novel algebraic structures from the polysymplectic form in field theory,'' *GROUP21, Physical Applications and Mathematical Aspects of Geometry, Groups and Algebras*, edited by H. A. Doebner, W. Scherer, and C. Schulte (World Scientific, Singapore, 1997), Vol. 2, p. 894.
[18] I. V. Kanatchikov, ''Canonical structure of Classical Field Theory in the polymomentum phase space,'' Rep. Math. Phys. **41**(1), 49–90 (1998).
[19] Y. Choquet-Bruhat, C. De Witt-Morette, and A. Dillard-Bleik, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1977).
[20] R. L. Bryant, S. S. Chern, R. B. Gardner, H. L. Goldschmidt, and P. A. Griffiths, *Exterior Differential Systems* (Springer-Verlag, New York, 1991).
[21] L. Conlon, *Differentiable Manifold. A First Course* (Birkhauser, Boston, MA, 1993).
[22] C. B. Allendoerfer, *Calculus of Several Variables and Differentiable Manifolds* (Macmillan, New York, 1974).
[23] J. Dieudonne, *Elements d'Analyse* (Gauthier-Villars, Paris, 1974), Vol. IV.

[24] D. Chinea, M. De León, and J. C. Marrero, ''The constraint algorithm for time-dependent Lagrangrians, '' J. Math. Phys. **7**, 3410–3447 (1994).
[25] M. De León, J. Marín, and J. C. Marrero, ''The constraint algorithm in the jet formalism,'' Diff. Geom. Appl. **6**, 275–300 (1996).
[26] G. Sardanashvily, *Generalized Hamiltonian Formalism for Field Theory. Constraint Systems* (World Scientific, Singapore, 1995).

Limit analysis of the diffraction of a plane wave by a one-dimensional periodic medium

D. Felbacq[a] and B. Guizal
LASMEA UMR 6602, Complexe Universitaire des Cézeaux 63177 Aubiére Cedex, France

F. Zolla[b]
LOE ESA Faculté des Sciences de St. Jérôme 13397 Marseille Cedex 13, France

(Received 1 December 1997; accepted for publication 5 May 1998)

In this paper we address the problem of the reflection coefficient limit in a periodic stratified structure when the number of periods tends to infinity. We show that the graph of the reflected energy admits a superior envelope and we exhibit an explicit formula. © *1998 American Institute of Physics.* [S0022-2488(98)00809-3]

I. INTRODUCTION

The study of multilayered structures is an old one,[1] and has generated a tremendous amount of literature. Nevertheless it continues to be studied since it is a good laboratory for the general study of periodic structures, e.g., photonic crystals. Besides, modern technology allows its realization and there is a need for theoretical works, for instance, in the optimization of Bragg mirrors. In this paper we address the problem of reflection coefficient behavior when the number of periods tends to infinity. In the very simple case of a symmetric slab, it is known that when the thickness of the slab tends to infinity, the reflection coefficient oscillates and does not converge towards that of a diopter. As could be expected, the same happens for general periodic structure, provided that the concerned wavelength does not belong to a forbidden band. Nevertheless, it is possible to exhibit a weaker result: We show in this paper that when the number of periods tends to infinity, the graph of the reflection coefficient admits a superior envelope, whose explicit expression is derived.

II. SETTING OF THE PROBLEM

We consider in $\mathbb{R}^3$, endowed with a triaxial Cartesian coordinate system $(0,x,y,z)$, a periodic obstacle C, with period h in the x direction (cf. Fig. 1). C is enlightened by a monochromatic plane wave, polarized parallel to (Oz), under the incidence θ

$$\mathbf{U}^i=\exp(ik(x\ \cos\ \theta-y\ \sin(\theta))\exp(-i\omega t))\mathbf{z}, \tag{1}$$

with $k=\omega\sqrt{\varepsilon_0\mu_0}$. C is characterized by its relative permittivity ε, which is assumed to be real and piecewise continuous. The discontinuity points of ε over one period are labeled by $\{x_l\}, l \in\{0,\ldots,p-1\}$. We assume that C is made of N periods. Besides, we set

$$\beta:]0,Nh[\to\mathbb{R}^+ x\mapsto k\sqrt{\varepsilon(x)-\sin^2(\theta)} \tag{2}$$

and $\beta_0=k\ \cos\ \theta$.

Under these conditions, the electromagnetic field is described by a function $U(x)$ and $U(x)\exp(iky\ \sin\ \theta)$ represents the z-component of the electric field (E// polarization), resp. magnetic field (H// polarization). Maxwell equations lead to the following problem, where

$$F=\begin{pmatrix} U \\ p^{-1}\dot{U}\end{pmatrix}\quad 0\leqslant x\leqslant Nh:\frac{dF}{dx}=\begin{pmatrix} 0 & p \\ -p^{-1}\beta^2 & 0\end{pmatrix}F, \tag{3}$$

$$x\leqslant 0:U(x)=\exp(i\beta_0 x)+r_N(k)\exp(-i\beta_0 x), \tag{4}$$

[a]Electronic mail: felbacq@lasmea.univ-bpclermont.fr
[b]Electronic mail: zolla@loe.u-3mrs.fr

 © 1998 American Institute of Physics

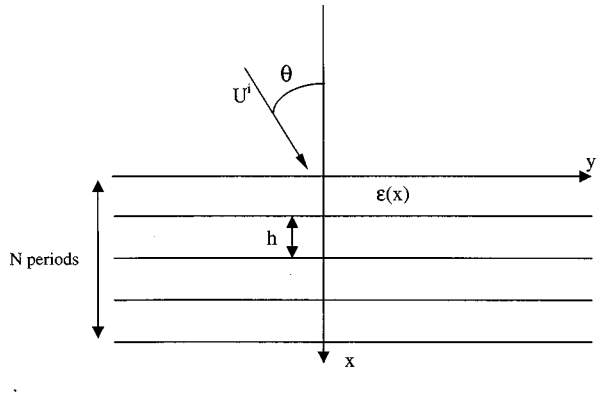

FIG. 1. Scheme of the diffraction problem.

$$x \geqslant Nh: U(x) = t_N(k)\exp(i\beta_0 x),$$

with $p \equiv 1$ for the E// polarization, $p \equiv \varepsilon$ for the H// polarization and $U \in C(\mathbb{R}), p^{-1}\dot{U} \in C(\mathbb{R})$. In order to get a useful representation of the field, we will need the following result.

Lemma 1: There exists $\mathbf{T} = (\mathrm{T}_{ij}) \in C(\mathbb{R}^+, S\mathbb{L}_2(\mathbb{R}))$, *such that* $F((n+1)h) = \mathbf{T}F(nh)$. $\mathbf{T}$ *is called the monodromy matrix of (3).*

Proof: The regularity of ε implies that

$$F = \begin{pmatrix} U \\ p^{-1}\dot{U} \end{pmatrix}$$

verifies an equality of the following type: $F(x_{l+1}) = T_l F(x_l)$ where $T_l \in C(\mathbb{R}^+, S\mathbb{L}_2(\mathbb{R}))$ (cf. Ref. 2). Consequently, the same kind of relation holds for one period; $\mathbf{T}$ is defined by setting

$$\mathbf{T}: \mathbb{R}^+ \to \mathrm{S}\mathbb{L}_2(\mathbb{R}), \quad k \mapsto \prod_{l=0}^{p-1} T_l(k). \tag{5}$$

For the entire structure, we deduce a global relation:

$$\mathbf{T}^N \begin{pmatrix} 1 + r_N(k) \\ i\beta_0(1 - r_N(k)) \end{pmatrix} = t_N(k) \begin{pmatrix} 1 \\ i\beta_0 \end{pmatrix}. \tag{6}$$

We are interested in the behavior of the reflected field when the number of periods tends to infinity. In that case, the relevant quantity seems to be the reflected energy, which is the square modulus of the reflection coefficient. The aim of this paper is, therefore, to study the limit of $R_N := k \mapsto |r_N(k)|^2$ when $N \to +\infty$.

III. STATEMENT OF THE RESULTS

As we are interested in the behavior of the graph of the reflected energy, we use the concept of Γ-convergence, whose meaning is recalled:

Definition 2: *A sequence of functions* (f_N), *defined on a compact subset* Ω *of* $\mathbb{R}$, Γ*-converges towards* f *if, by definition, the two following propositions hold:*

($\star$) $\forall (x_N) \to x$, $liminf\ f_N(x_N) \geqslant f(x)$

($\star\star$) *For all* $x \in \Omega$, *there exists a sequence* (x_N) *verifying* $(x_N) \to k$ *and limsup* $f_N(x_N) \leqslant f(x)$.

Our main result is:

Theorem 3: *(1) We assume that* $\frac{1}{2}|Tr(\mathbf{T})| > 1$, *and we denote by* μ *the eigenvalue of* $\mathbf{T}$ *of smallest modulus.*

When $N \to \infty$ (R_N) *converges pointwise towards 1,* (F_N) *converges pointwise towards* F_∞ *verifying* $\mathbf{T}F_\infty = \mu F_\infty$.

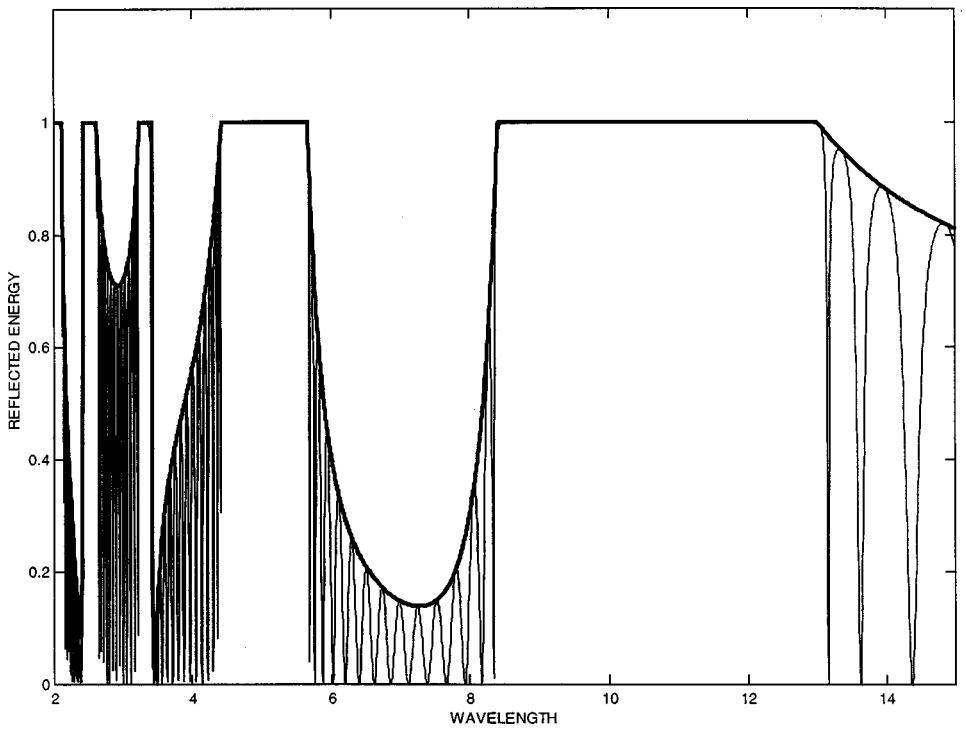

FIG. 2. Reflected energy and gamma-limit for a periodic structure with two layers: $h_1=1$, $h_2=1$,$\varepsilon_1=12.25$, $\varepsilon_2=2.25$. The thick line corresponds to the graph of (8).

(2) We assume that $\frac{1}{2}|Tr(\mathbf{T})|<1$.
When $N\to\infty$, $(1-R_N)$ Γ*-converges towards*

$$k\mapsto(1-\tfrac{1}{4}Tr(\mathbf{T})^2)(\mathrm{T}_{12}\beta_0-\mathrm{T}_{21}\beta_0^{-1})^{-2}. \tag{7}$$

The limit of the superior envelope of the graph of R_N *is, therefore, the graph of:*

$$R_\Gamma : k\mapsto 1-(1-\tfrac{1}{4}Tr(\mathbf{T})^2)(\mathrm{T}_{12}\beta_0-\mathrm{T}_{21}\beta_0^{-1})^{-2}. \tag{8}$$

Numerical Example: We give in Fig. 2 the square modulus of the reflection coefficient of a Bragg mirror (i.e., two layers in one period), with permittivities 12.25 and 2.25 and thicknesses 1 enlightens under normal incidence by a plane wave. The thick line corresponds to the superior limit.

Remark: The case $\frac{1}{2}|Tr(\mathbf{T})|=1$ is a peculiar one, as it may correspond to a transparent medium, i.e., $\mathbf{T}$ is the identity, or to a nondiagonalizable matrix $\mathbf{T}$. Indeed, 1 is then the only eigenvalue. In the first case, (8) admits a discontinuity point, while in the second, R_Γ is equal to 1 and (8) is continuous.

IV. PROOFS

(1) We assume that $\frac{1}{2}|Tr(\mathbf{T})|>1$. The eigenvalues of $\mathbf{T}$ are distinct and real, we denote by μ the smallest one ($|\mu|<1$). $\mathbf{T}$ can be put in diagonal form

$$\mathbf{T}=\mathbf{P}\begin{pmatrix}\mu & 0\\ 0 & \mu^{-1}\end{pmatrix}\mathbf{P}^{-1},\mathbf{P}=\begin{pmatrix}a & c\\ b & d\end{pmatrix}\in S\mathbb{L}_2(\mathbb{R}).$$

Let us denote

$$F_N(0)=\begin{pmatrix}\zeta_N\\ \xi_N\end{pmatrix}.$$

The condition (6) gives

$$\frac{\zeta_N}{\xi_N}=\frac{iac\beta_0\mu^{2N}-iac\beta_0-bc\mu^{2N}+ad}{iad\beta_0\mu^{2N}-ibc\beta_0-bd\mu^{2N}+bd}. \tag{9}$$

Going to the limit, we get: $\lim_{N\to+\infty}\zeta_N/\xi_N=a/b$, which proves that $F_N(0)$ tends towards an eigenvector associated to μ. Moreover, the equalities $\zeta_N=1+r_N, \xi_N=i\beta_0(1-r_N)$ imply that

$$r=\lim_{M\to\infty} r_N=\frac{i\beta_0 a-b}{i\beta_0 a+b}. \tag{10}$$

Finally: $|r|=1$.

(2) We now assume that $\frac{1}{2}|Tr(\mathbf{T})|<1$. We set $\Phi: k\mapsto \arccos(\frac{1}{2}Tr(\mathbf{T}))$.

It can easily be shown (cf. Ref. 3) that

$$\mathbf{T}^N=\begin{pmatrix} T_{11}S_N-S_{N-1} & T_{12}S_N \\ T_{21}S_N & T_{22}S_N-S_{N-1}\end{pmatrix},$$

with $S_N=\sin(N\Phi)\sin(\Phi)^{-1}$. Equation (6) allows us to write: $R_N=1-4f_N$ where

$$\begin{aligned} f_N&: k\mapsto\{4+\sin^2(N\Phi)[g(k)-4]\}^{-1},\\ g&: k\mapsto(T_{12}\beta_0-T_{21}\beta_0^{-1})^2\sin(\Phi)^{-2}.\end{aligned} \tag{11}$$

Our aim is to show the Γ-convergence of (f_N) towards $1/g$, that is, the two following propositions:

($\star$) $\forall(k_N)\to k$, *liminf* $f_N(k_N)\geqslant 1/g(k)$

($\star\star$) *For all* k *such that* $\frac{1}{2}|Tr(\mathbf{T})|<1$, *there exists a sequence* (k_N) *verifying* $(k_N)\to k$ *and limsup* $f_N(k_N)\leqslant 1/g(k)$.

First we note that $R_N\geqslant 0$ implying $g(k)-4\geqslant 0$. As a consequence, we deduce $f_N(k)\geqslant 1/g(k)$, for all k such that $\frac{1}{2}|Tr(\mathbf{T})|<1$. Which proves ($\star$).

For ($\star\star$), we first prove:

Lemma 4: *For* $\Phi\in[0,\pi]$, *there exists* $(\Phi_N)\to\Phi$, *such that* $\sin(N\Phi_N)\to 1$. This lemma is obvious by choosing $\Phi_N=\Phi+(\pi/2-u_N)/(N)$, where $u_N\equiv N\Phi[2\pi]$.

Let us now consider an open interval U verifying: $\forall k\in U, \frac{1}{2}|Tr(\mathbf{T})|<1$ and choose $k_0\in U$. We set $\Phi_0=\Phi(k_0)$ and $V=\Phi(U)$. From the above lemma, there exists $(\Phi_N)\to\Phi_0$ such that $\sin(N\Phi_N)\to 1$. As Φ is continuous, we deduce: $W:=\Phi^{-1}(\{\overline{\Phi_N}\})\cap U\neq\varnothing$ ($\overline{A}$ denotes the closure of the set A). Noticing that $k_0\in\overline{W}$, we conclude that there exists $k_N\to k_0$ such that $f_N(k_N)\to 1/g(k_0)$.

V. CONCLUSION

For a general multilayered periodic structure, we have shown that, when the number of periods tends to infinity, the superior envelope of the reflectivity admits a limit. This result gives a regularization of the reflection coefficient and could be of use for the optimization of such structures.

[1] L. Brillouin and M. Parodi, *Propagation des ondes dans les milieux périodiques* (Masson, Paris, 1956).
[2] V. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer-Verlag, Berlin, 1983).
[3] J. Lekner, ''Light in periodically stratified media,'' J. Opt. Soc. Am. A **11**, 2892–2899 (1994).

Black hole thermodynamics and spectral analysis

F. Belgiorno[a)]
Dipartimento di Fisica, Università di Milano, 20133 Milano, Italy

(Received 5 January 1998; accepted for publication 6 April 1998)

We deal with some problems concerning quantum matter fields thermodynamics on static spherosymmetric black hole backgrounds. By means of variable separation, we study the self-adjointness of the Euclidean wave operators relevant in the calculation of the partition function and of the thermal Green functions for the case of scalar quantum fields in thermal equilibrium with a Schwarzschild black hole. In this respect, the role of the horizon surface is analyzed. An approach to this topic, related to completeness of the Euclidean manifold, is also discussed. It is found that, if the horizon surface is not included in the Euclidean manifold (or if it cannot be included, as in the case of a temperature different from the black hole one), essential self-adjointness is missing and boundary conditions on the horizon are required. We also analyze some qualitative spectral properties of the given wave operators after introducing an infrared regularization by means of a spherical box. We find that in the standard Euclidean path integral approach the spectrum is purely discrete. The same analysis is carried out for the noncovariant choice of the functional measure in the Euclidean path integral. An analogous study is carried for an extremal Reissner–Nordström black hole and it is sketched in the nonextremal one. The latter is completely analogous to the Schwarzschild case, whereas in the extremal case it is found that the corresponding Riemannian manifold is complete for every choice of the temperature and that at finite four volume there is a nonvoid continuous spectrum. Physical implications of our results are discussed. © *1998 American Institute of Physics.* [S0022-2488(98)03509-9]

I. INTRODUCTION

In the framework of the Euclidean approach to black hole thermodynamics it is postulated that the partition function for scalar quantum matter fields in thermal equilibrium with a black hole is given by[1]

$$Z=\int_{P(\beta)}[D\phi]\exp(-S[\phi]), \tag{1}$$

where $S[\phi]$ is the classical action for the free field ϕ,

$$S[\phi]=\frac{1}{2}\int d^4x\sqrt{g}\phi(-\Delta+M^2+\xi R)\phi,$$

and Δ is the Laplacian on functions; M is the mass of the field and ξ is an arbitrary coupling constant of the field with the Ricci scalar R. $P(\beta)$ indicates that periodic boundary conditions on the Euclidean time are imposed and $[D\phi]$ is generally chosen to be the covariant functional measure,

$$[D\phi]=\prod_x g(x)^{1/4}\,d\phi(x),$$

where $g(x)\equiv\det(g_{ab}(x))$. In the configuration space, a possible choice for the measure is the noncovariant one:

a)Electronic mail: belgiorno@mi.infn.it

 © 1998 American Institute of Physics

$$[D\phi]_{NC}=\prod_x \sqrt{g^{00}(x)}g(x)^{1/4}\, d\phi(x).$$

This measure is the one, e.g., derived by choosing a Liouville measure in phase space.[2] For the sake of completeness, we will also treat this case, although it is not the standard choice. A criticism about the validity of the noncovariant choice is found in Ref. 3 and a further discussion is given in Ref. 2. In any case, results relative to the noncovariant measure are equivalent to the ones that can be obtained by working on the so called optical manifold, as it will result from the following section.

In (1) one has to specify the background metric and the manifold; a necessary condition in order to give a physical meaning to a partition function is that the metric is stationary or static, unless some adiabatic expansion in the time dependence of the manifold is considered. In the following, we will deal with static spherosymmetric asymptotically flat black hole manifolds. Their metric for the external Euclidean region can be written as

$$ds^2=g_{00}(r)d\tau^2+\frac{1}{g_{00}(r)}dr^2+r^2\, d\Omega^2. \tag{2}$$

The function $g_{00}(r)$ is characterized by one or more zeros, the greatest of which, say r_{bh}, identifies the event horizon. The corresponding manifold (in the above coordinates) is diffeomorphic to $\mathscr{M}=S^1\times(r_{bh},\infty)\times S^2$. Actually, as it is generally accepted, at the Hawking temperature the thermodynamical manifold in the case of the Schwarzschild nondegenerate Killing horizon and for the covariant measure, also encloses the point $r=r_{bh}$, so that the above metric corresponds only to an incomplete chart partially covering the actual thermodynamical manifold $\mathscr{M}_E$. A detailed discussion on this topic is found in Sec. III. A spherical boundary at $r=R>r_{bh}$ is also introduced as a natural infrared regulator necessary in order to obtain finiteness for global quantities like the partition function, free energy, entropy and internal energy. In every case, such a regularization corresponds to a finite four-volume.

In the case of free scalar fields, for the covariant functional measure, equation (1) becomes

$$Z=\frac{1}{\sqrt{\mathrm{Det}(-\Delta+m^2+\xi R)}}, \tag{3}$$

so in the free field case the calculation of the partition function reduces to the calculation of a functional determinant of the wave operator,

$$A\equiv-\Delta+M^2+\xi R.$$

If the noncovariant choice is made, one gets

$$Z=\frac{1}{\sqrt{\mathrm{Det}[g_{00}(-\Delta+m^2+\xi R)]}}; \tag{4}$$

the operator involved in the functional determinant is then[4]

$$A_{NC}\equiv g_{00}A.$$

The operators A and A_{NC} also determine the thermal Green functions.[4]

The zeta function approach is the most useful tool in order to perform partition function calculations, but we will not pursue it in this paper. Our work consists mainly in the analysis of essential self-adjointness and of some qualitative spectral properties of the above wave operators on black hole manifold $\mathscr{M}$ and its relations, if it is possible, with the analysis on $\mathscr{M}_E$. In particular, the role of the horizon is analyzed in relation to the deficiency indices of the wave operators; essential self-adjointness (in the absence of the box) is regarded as a property ensuring that physics, i.e., thermodynamical features of the system, is well defined without any necessity of introducing boundary conditions on the horizon.

We will also discuss the case in which a conical defect γ is present in the manifold (see also Sec. III), i.e., when $\gamma \equiv \beta/\beta_h \neq 1$, where β is the inverse temperature and β_h is the inverse of the Hawking temperature. As usual, the case $\gamma=1$ will be called ''on shell,'' whereas the case $\gamma \neq 1$ will be called ''off shell.'' We recall that, although thermal equilibrium with the black hole can be obtained only at the Hawking temperature, going off shell appears as an intermediate computational tool in some approaches to black hole thermodynamics. For a general discussion on these topics see, e.g., Refs. 5, 6 and references therein.

The results we will obtain will be inferred by the analysis of the partial wave operators (p.w.o.) arising by variable separation on the black hole manifold. Other tools, particularly in the case of qualitative spectral analysis, could be used, but we prefer variable separation as it allows us to treat straightforwardly all cases. Finally, some results about qualitative spectral properties of the p.w.o. we will obtain overlap partially with the ones contained in Ref. 4 as far as the Schwarzschild on shell case is considered.

II. VARIABLE SEPARATION

The minimal domain for the wave operator A is $C_0^\infty(\mathscr{M})$,[7,8] which is dense in $L^2(\mathscr{M}, d^4x\sqrt{g})$. We can write

$$L^2(\mathscr{M}, d^4x\sqrt{g}) = L^2(S^1, d\tau) \otimes L^2(I, dr') \otimes L^2(S^2, d\Omega),$$

where I is the ''radial'' interval and dr' is the suitable measure on it. We denote by K_n the eigenspace corresponding to the eigenvalue n^2, $n \in \mathbb{Z}$ of the operator $-\partial_\tau^2$ with a periodic boundary condition on τ and by S_l the eigenspace corresponding to the eigenvalue $l \in \mathbb{N}$ of the Laplace–Beltrami operator $-\Delta_\Omega$ on S^2; then one gets

$$L^2(S^1, d\tau) \otimes L^2(I, dr') \otimes L^2(S^2, d\Omega) = \sum_{n,l \in \mathbb{N}}^{\oplus} V_{nl},$$

where

$$V_{nl} = L^2(I, dr') \otimes K_n \otimes S_l;$$

cf. also Ref. 9, chapter X, Example 4. The subspaces V_{nl} are reducing subspaces for A. We so obtain the orthogonal decomposition of the operator A in the direct sum of partial wave operators A_{nl} defined by restricting A to $D(A) \cap V_{nl}$; such restrictions will be called the V_{nl}-parts of A. We get

$$A = \sum_{n,l \in \mathbb{N}}^{\oplus} A_{nl} \otimes \mathrm{Id}_n \otimes \mathrm{Id}_l, \tag{5}$$

where $\mathrm{Id}_n, \mathrm{Id}_l$ are the identity operators on K_n, S_l, respectively. The direct sum in (5) is defined on the linear hull $L\{C_0^\infty(I) \otimes K_n \otimes S_l : n,l \in \mathbb{N}\}$ that is also a core for its closure. Moreover, the closures of A restricted to $L\{C_0^\infty(I) \otimes K_n \otimes S_l : n,l \in \mathbb{N}\}$ and of A defined on $C_0^\infty(\mathscr{M})$ are identical.

Indeed, let us consider the closure $\overline{A}$ of A on $C_0^\infty(\mathscr{M})$. We will show that $L\{C_0^\infty(I) \otimes K_n \otimes S_l : n,l \in \mathbb{N}\}$ is a core for $\overline{A}$. It is known[10] that a subspace reduces a closable operator if and only if it reduces its closure; in our case we have an orthogonal decomposition of the whole Hilbert space into the orthogonal sum of reducing subspaces for the operator A and for its closure $\overline{A}$. Then in the latter case, it holds that

$$\overline{A} = \sum_{n,l \in \mathbb{N}}^{\oplus} \overline{A}_{nl} \otimes \mathrm{Id}_n \otimes \mathrm{Id}_l. \tag{6}$$

We recall that, if X is a reducing subspace for a closable (closed) operator A, the X-part of A is closable (closed) and the closure of the X-part of A is the X-part of the closure of A.[10] This means that the closure of the V_{nl}-parts of A coincide with the V_{nl}-parts of $\overline{A}$. Moreover, from the

corollary to theorem 4.4.5 of Ref. 10 one can conclude that the closure of the direct sum (5) of closable operators A_{nl} is the direct sum of the closures $\overline{A_{nl}}$. The thesis then follows from the equality between the closure of the direct sum in (5) and the orthogonal decomposition of $\overline{A}$ in (6).

This means that problems of self-adjointness of the operator A on $C_0^\infty(\mathscr{M})$ are reduced to the analogous problems for the p.w.o., A_{nl}. Particularly, the following relation between the deficiency indices $(N_+(\overline{A}),N_-(\overline{A}))$ of $\overline{A}$ and the ones of the p.w.o. (cf. also Ref. 9, chapter X, Problem 1a) holds:

$$N_+(\overline{A})=\sum_{n,l\in\mathbb{N}} N_+(\overline{A_{nl}})\cdot \dim K_n\cdot \dim S_l,$$
$$N_-(\overline{A})=\sum_{n,l\in\mathbb{N}} N_-(\overline{A_{nl}})\cdot \dim K_n\cdot \dim S_l. \tag{7}$$

We will assume the commutativity of the self-adjoint extension of the wave operator A with the operators $-\partial_\tau^2$ and $-\Delta_\Omega$ in any case in which this property is not automatically given. This means that we will consider only those self-adjoint extensions of A having the same reducing subspaces V_{nl} as A; these extensions are obtained by extending the partial wave operators A_{nl} in $L^2(I,dr')$. Then the relevant spectral problem,

$$A\psi_N=\lambda_N\psi_N,$$

by means of the factorization of the eigenfunctions,

$$\psi(\tau,r,\theta,\phi)=\zeta(r)f(\tau)Y_{lm}(\theta,\phi) \tag{8}$$

[$Y_{lm}(\theta,\phi)$ are the spherical harmonics], becomes a spectral problem for a family of one-dimensional Schrödinger-like operators defined on an interval (the radial one). Indeed, given the suitable domain for A_{nl} to be self-adjoint, the above eigenvalue problem reduces to

$$A_{nl}\zeta_{\alpha nl}=\lambda_{\alpha nl}\zeta_{\alpha nl}. \tag{9}$$

We also recall the following formula, relating the spectrum of the above selected self-adjoint extension of the operator A with the spectrum of the self-adjoint extensions of the p.w.o.:

$$\sigma(A)=\overline{\bigcup_{n,l\in\mathbb{N}}\sigma(A_{nl})}. \tag{10}$$

The noncovariant operator A_{NC} is treated in an analogous way. The only difference is in the Hilbert space measure. Indeed, A_{NC} becomes formally self-adjoint in the Hilbert space $L^2(\mathscr{M},d^4x(\sqrt{g}/g_{00}))$.

We recall also that the following (unitary) transformation[11] on the field $\phi=\overline{\phi}/\sqrt{g_{00}}$ from $L^2(\mathscr{M},d^4x(\sqrt{g}/g_{00}))$ to $L^2(\mathscr{M},d^4x(\sqrt{g}/g_{00}^2))$ carries the eigenvalue problem for A_{NC} into the eigenvalue problem for

$$A_{opt}\equiv-\Delta_h+q(x), \tag{11}$$

where Δ_h is the Laplacian on scalar functions for the so called optical metric $h_{\mu\nu}$

$$ds_{opt}^2=d\tau^2+\frac{1}{g_{00}^2(r)}dr^2+\frac{r^2}{g_{00}(r)}d\Omega^2, \tag{12}$$

and the potential $q(x)$ is

$$q(x)=\frac{1}{\sqrt{g_{00}}}(\Delta_h\sqrt{g_{00}})+g_{00}(M^2+\xi R).$$

About this equivalence, cf. also Ref. 7, chapter 3. The p.w.o. analysis is sufficient in order to establish self-adjointness properties of A_{NC}. Nevertheless, a ''direct'' approach is possible and it is found in Appendix C.

III. SCHWARZSCHILD BLACK HOLE

The function $g_{00}(r)$ in (2) is given by

$$g_{00}(r)=1-\frac{r_{bh}}{r},$$

where r_{bh} is the black hole radius, associated with the black hole mass m by $r_{bh}=2m$. (We use natural units $\hbar=c=G=k_B=1$.) We will start by treating the case of generic γ on the manifold $\mathscr{M}$. It is useful to pass to rescaled coordinates:

$$\tau\to\frac{\tau}{2r_{bh}\gamma}\in[0,2\pi);\quad r\to\frac{r}{2r_{bh}}\in\left(\frac{1}{2},\infty\right).$$

We will call these rescaled coordinates again τ, r, unless explicitly remarked. So one gets

$$ds^2=(2r_{bh})^2\left(\left(1-\frac{1}{2r}\right)\gamma^2d\tau^2+\frac{1}{\left(1-\frac{1}{2r}\right)}dr^2+r^2d\Omega^2\right).$$

By means of the new variable

$$y=\sqrt{2}\sqrt{r-\tfrac{1}{2}}\in(0,\infty),$$

the metric becomes

$$ds^2=(2r_{bh})^2\left(\frac{y^2}{(1+y^2)}\,\gamma^2d\tau^2+(1+y^2)dy^2+\frac{1}{4}(1+y^2)^2d\Omega^2\right). \tag{13}$$

This choice of the ''radial'' variable is particularly suitable in order to explore the behavior of the metric near the horizon, that is, for $y\to0^+$; indeed, as it is well known, near the horizon the manifold has a conical structure:

$$ds^2\sim(2r_{bh})^2\left(y^2\gamma^2d\tau^2+dy^2+\frac{1}{4}d\Omega^2\right), \tag{14}$$

that is, it is equivalent to $C_\gamma\times S^2$, where C_γ is a two-dimensional cone that degenerates to a punctured plane if $\gamma=1$. In the presence of a conical defect $\gamma\neq1$, a careful analysis has to be done in order to take into account curvature distributional contributions[12] at the tip of the cone.

If $\gamma=1$, then the Riemannian manifold can be extended smoothly up to include the surface $y=0$. In fact, in the framework of the so called Euclidean path integral approach,[1] the thermodynamical manifold is required to be the one representing the so called Euclidean section of the maximal extension of the Schwarzschild spacetime, characterized by $\gamma=1$ and $R=0$ and including $y=0$, that is $r=r_{bh}$. Note that the metric in (2) is related with a coordinate chart not covering the value $r=r_{bh}$. This value corresponds to a set of fixed points for the Killing vector ∂_τ and can be covered, e.g., by means of the Kruskal Euclidean coordinates T, X,[1]

$$ds^2=\frac{4r_{bh}^3}{r}\exp\left(-\frac{r}{r_{bh}}\right)(dT^2+dX^2)+r^2d\Omega^2.$$

The relation with the incomplete chart is

$$T^2+X^2=\left(\frac{r}{r_{bh}}-1\right)\exp\left(\frac{r}{r_{bh}}\right),\quad \frac{T}{X}=\tan\left(\frac{\tau}{2r_{bh}}\right).$$

The manifold covered by the given Kruskal chart corresponds to the Euclidean section, that is $\mathscr{M}_E=\mathbb{R}^2\times S^2$. Note that its fundamental group is trivial, i.e., $\mathscr{M}_E$ is simply connected, whereas $\Pi_1(\mathscr{M})=\mathbb{Z}$. Note also the relation between T, X and ''conical'' coordinates:

$$T=\exp\left(\frac{y^2+1}{2}\right)y\ \sin(\tau),\quad X=\exp\left(\frac{y^2+1}{2}\right)y\ \cos(\tau). \tag{15}$$

It is evident that the manifold $\mathscr{M}_E$ is not compatible with the choice of the noncovariant functional measure and with the introduction of a conical defect. Both these cases require us to consider a manifold $\mathscr{M}=S^1\times(0,\infty)\times S^2$, that is, the manifold cannot include the bifurcation surface $y=0$. So a modification in topology of the manifold is required w.r.t. the Euclidean path integral approach (cf. also Ref. 4).

We will start by discussing the on shell case for the covariant measure choice; a comparison between the case in which the manifold is $\mathscr{M}$ and the case in which the manifold is $\mathscr{M}_E$ will be carried out. (As it is known,[4] this choice for the manifold cannot allow a tree level gravitational entropy contribution.) Then an analysis of the off shell case and of the noncovariant one, both necessarily defined only on $\mathscr{M}$, will be made.

As a final remark of this section, we stress that in the following we will deal with operators defined on metrics of the form

$$ds^2=c^2(G(y)d\tau^2+F(y)dy^2+H^2(y)d\Omega^2),$$

where c is a constant with the dimensions of length. We will carry our analysis about p.w.o. on the rescaled metric,

$$ds^2_{\text{rescaled}}=G(y)d\tau^2+F(y)dy^2+H^2(y)d\Omega^2.$$

It is evident that

$$A=\frac{1}{c^2}A_{\text{rescaled}}.$$

A. p.w.o. analysis on shell

The wave operator A in the metric (13) with $\gamma=1$ can be written as

$$A=\frac{1}{yf(y)^2}(-\partial_y(yf(y)\partial_y))-\frac{f(y)}{y^2}\partial_\tau^2-\frac{4}{f(y)^2}\Delta_\Omega+M^2+\xi R,$$

where

$$f(y)=1+y^2.$$

We recall that the Euclidean section is Ricci-flat, so $R=0$. The (rescaled) p.w.o. operators in the metric (13) are

$$A_{nl}=\frac{1}{yf(y)^2}(-\partial_y(yf(y)\partial_y)+V_{nl}(y)), \tag{16}$$

where

$$V_{nl}(y)=yf(y)^2R_{nl}(y),$$

$$R_{nl}(y)=\frac{f(y)}{y^2}n^2+\frac{4l(l+1)}{f(y)^2}+M^2,$$

$$y \in (0,\infty).$$

The p.w.o. obtained has a Sturm–Liouville form and the powerful theorems concerning Sturmian operators can be used.

The p.w.o. A_{nl} in (16) defines a singular Sturm–Liouville operator on an interval $I \subset \mathbb{R}$. Indeed, it is an operator of the form

$$L\varphi(y) = \frac{1}{k(y)}\left(-\frac{d}{dy}\left(p(y)\frac{d\varphi(y)}{dy}\right) + q(y)\varphi(y)\right).$$

Conditions to be satisfied by $k(y)$, $p(y)$, $q(y)$ are, e.g., found in Refs. 13, 14. In our case we get $k(y) = yf(y)^2 > 0$, $p(y) = yf(y)$, $q(y) = V_{nl}(y)$ that are all smooth functions for $y \in (0,\infty)$. Note that $y = 0$ is a singular point for L. One can define the maximal and the minimal operators[13] induced by our p.w.o. In our case the minimal operator $T_0(A_{nl})$ can be defined as follows:

$$D(T_0(A_{nl})) = C_0^\infty(I) \cap L^2(I, k(y)dy),$$

$$T_0(A_{nl})\varphi = A_{nl}\varphi \quad \text{for} \quad \varphi \in D(T_0(A_{nl}));$$

the maximal operator $T(A_{nl})$ can be defined as (cf. Ref. 7)

$$D(T(A_{nl})) = \{\varphi \in C^\infty(I) \,|\, \varphi, A_{nl}\varphi \in L^2(I, k(y)dy)\} = D(T_0^*(A_{nl})) \cap C^\infty(I),$$

$$T(A_{nl})\varphi = A_{nl}\varphi \quad \text{for} \quad \varphi \in D(T(A_{nl})).$$

$L \equiv A_{nl}$ is a formal elliptic differential operator formally self-adjoint. Ellipticity implies that a weak solution of the equation $(L - \lambda)\varphi = 0$ for $\lambda \in \mathbb{C}$ is also a strong solution and a smooth solution of the same equation.[7] So deficiency subspaces of the minimal operator are necessarily subspaces of $C^\infty(I) \cap L^2(I, k(y)dy)$, and are spanned by ordinary (i.e., nondistributional) solutions of the given differential equation.

1. Deficiency indices of A

We now study the behavior of the solution of the differential equation

$$A_{nl}\varphi = \lambda\varphi, \tag{17}$$

near $y = 0$. We want to determine if, according to Weyl's alternative, in a right neighbor of 0, all the solutions of (17) are square integrable for every $\lambda \in \mathbb{C}$, that is, $\varphi \in L^2((0,c], k(y)dy)$, $0 < c < \infty$ for every $\lambda \in \mathbb{C}$; or if, for every $\lambda \in \mathbb{C}$ it exists a solution not square integrable. In the former case, one gets the so called limit circle case (LCC); in the second one the so called limit point case (LPC) is verified.

Equation (17) is equivalent to the following one:

$$\left(-\partial_y^2 - \frac{1+3y^2}{y(1+y^2)}\partial_y + (1+y^2)(R_{nl}(y) - \lambda)\right)\varphi(y) = 0,$$

and it is a second order differential equation with rational coefficients. Near $y = 0$ one gets

$$\left(-\partial_y^2 - \left(\frac{1}{y} + O(y)\right)\partial_y + \left(\frac{n^2}{y^2} + 2n^2 + 4l(l+1) + M^2 - \lambda + O(y^2)\right)\right)\varphi(y) = 0,$$

and the following root for the indicial equation:

$$\alpha_1 = |n|, \quad \alpha_2 = -|n|.$$

A fundamental system for $y \to 0^+$ is given by

$$\zeta_+(y)=y^{|n|}\sum_{k=0}^{\infty}a_k y^k,\quad \zeta_-(y)=y^{-|n|}\sum_{k=0}^{\infty}b_k y^k+\zeta_+(y)\log(y).$$

For $n\neq 0$, we get the LPC: there are for every choice of λ solutions not belonging to $L^2((0,c],y(1+y^2)^2dy)$, where $(0,c]$ represents a right neighbor of 0. For $n=0$ one has the LCC: both the independent solutions of our equation are square integrable in a right neighbor of 0. Moreover, it is easy to show that the LPC is verified at infinity for every p.w.o.

For a real formal differential operator of order 2 one can find straightforwardly the deficiency indices of T_0 from the knowledge of the behavior of τ according to Weyl's alternative. Indeed, according to theorem 5.7 of Ref. 14, the deficiency indices are (0,0) if one gets the LPC at both the extremes; (1,1) if one finds the LPC at one extreme and the LCC at the other; (2,2) if LCC is verified at both the extremes.

Summarizing our results, we have that the partial wave operators with $n\neq 0$ are characterized by the limit point case at 0 and at infinity, so they have deficiency indices (0,0); the partial wave operators with $n=0$ are characterized by the limit circle case at 0 and have deficiency indices (1,1).

This means that for $n\neq 0$ in this case the p.w.o. are essentially self-adjoint, whereas for $n=0$ there is a one parameter family of self-adjoint extensions for each p.w.o.

The deficiency indices of A are given by the formula (7) and they are evidently equal and infinite. Indeed, even if there is only a value of n contributing nonzero deficiency indices, namely $n=0$, we have

$$N_+(\overline{A})=\sum_{l\in\mathbb{N}}(2l+1)N_+(\overline{A}_{0l})=\infty,\quad N_-(\overline{A})=\sum_{l\in\mathbb{N}}(2l+1)N_-(\overline{A}_{0l})=\infty$$

due to the fact that one has to sum over $l\in\mathbb{N}$. This means that the wave operator A is not essentially self-adjoint on $\mathscr{M}$ and that boundary conditions on the horizon have to be defined. We postpone a discussion about suitable boundary conditions to subsection III B.

2. Spectral analysis

We limit ourselves at first at the case in which a spherical boundary surrounds the black hole. This means $y\in(0,y_b]$. Note that $y=y_b$ is a regular point for the p.w.o. and the LCC is verified there. Deficiency indices then become (2,2) in the $n=0$ case and (1,1) in the $n\neq 0$ one. Moreover, and a Dirichlet boundary condition is assumed at $y=y_b$. Deficiency indices of A are (∞,∞).

Let us analyze some qualitative properties of the spectrum. In order to study the spectral properties of the p.w.o. it is useful to introduce the variable z defined by

$$\frac{dz}{dy}=\sqrt{1+y^2};\tag{18}$$

the transformation (18) is smooth and one obtains

$$z(y)=\tfrac{1}{2}(y\sqrt{1+y^2}+\log(y+\sqrt{1+y^2})),\quad z\in(0,z_b].\tag{19}$$

The metric becomes

$$ds^2=(2r_{bh})^2\left(\frac{y(z)^2}{f(y(z))}d\tau^2+dz^2+\frac{1}{4}f(y(z))^2d\Omega^2\right),\tag{20}$$

where $f(y(z))=1+y(z)^2$.

The wave operator on shell on the basis (8) gives rise to the following p.w.o.:

$$A_{nl}=-\frac{1}{P(z)}\partial_z(P(z)\partial_z)+V_{nl}(z),$$

$$P(z)=y(z)(1+y(z)^2)^{3/2},$$

$$V_{nl}(z)=n^2\frac{(1+y(z)^2)}{y(z)^2}+4l(l+1)\frac{1}{(1+y(z)^2)^2}+M^2.$$

The following Liouville unitary transformation:

$$U:L^2((0,z_b],P(z)dz)\to L^2((0,z_b],dz),$$

$$(U\varphi)(z)=\sqrt{P(z)}\varphi(z)\equiv u(z),$$

carries (17) into $B_{nl}u(z)=\lambda u(z)$, where

$$B_{nl}=-\frac{d^2}{dz^2}+Q_{nl}(z),$$

$$Q_{nl}(z)=V_{nl}(z)+\frac{1}{2}\left(\frac{dG(z)}{dz}+\frac{G(z)^2}{2}\right),$$

$$G(z)=\frac{d\ \log(P(z))}{dz};$$

it holds that $A_{nl}=U^{-1}B_{nl}U$ and also $\varphi\in L^2((0,z_b],P(z)dz)\Leftrightarrow u\in L^2((0,z_b],dz)$. The potential term is given by

$$Q_{nl}(z)=\left(n^2-\frac{1}{4}\right)\frac{1}{z^2}\frac{z^2}{y(z)^2}+W_{nl}(z),$$

$$W_{nl}(z)=n^2+M^2+\frac{4l(l+1)}{(1+y(z)^2)^2}+\frac{1}{4(1+y(z)^2)^3}(11+3y(z)^2+y(z)^4);$$

$W_{nl}(z)$ is a bounded smooth function in $(0,z_b]$. For $z\sim 0^+$, that is for $y\sim 0^+$, one gets a singular behavior of the operators B_{nl}. Note that our B_{nl}, for $I=(0,\infty)$, represents a particular case of the Schrödinger-like operators studied in Ref. 15.

The $n\neq 0$ case does not require the definition of a boundary condition at $z=0$. The choice of the Dirichlet boundary condition at $z=z_b$ defines a self-adjoint extension of the p.w.o. Then we can apply criterion (25) of Ref. 16, p. 1598, in order to determine if the essential spectrum σ_e of our self-adjoint extension is void. We recall that all the self-adjoint extensions of our minimal operators have the same essential spectrum, because they have finite deficiency indices. The above criterion requires that, given an operator of the form

$$B=-\frac{d^2}{dz^2}+Q(z),\quad z\in(0,z_b],$$

where $Q(z)$ is real and continuous function, if

$$\liminf_{z\to 0} z^2Q(z)>-\tfrac{1}{4},$$

then $\sigma_e(B)=\varnothing$ and B is bounded from below. Our $Q_{nl}(z)$ satisfies the request to be a real continuous function and it holds that

$$\lim_{z\to 0^+} z^2Q_{nl}(z)=n^2-\tfrac{1}{4}>-\tfrac{1}{4},$$

so the essential spectrum is void. [We used $\lim_{z\to 0^+}[z^2/y(z)^2]=1$.] This means that for $n\neq 0$ there is only a discrete spectrum for our operators.

In the $n=0$ case, from Ref. 16 (corollary 12, p. 1402), one infers that the essential spectrum is again void: indeed each A_{0l} is a formally self-adjoint differential operator of order 2 having

deficiency indices (2,2). Note that this is true for every choice of the boundary condition at $z=0$ as far as the resulting operator is self-adjoint. Moreover, the p.w.o. A_{0l} are bounded from below, as it can be inferred from Sturmian oscillation theory.[16,14]

Summarizing, we found that the p.w.o. have a purely discrete spectrum for all n,l and for every value of the mass M (due to the fact that the Schwarzschild solution is Ricci-flat; the other coupling ξ is irrelevant).

As expected, the presence of a box discretizes the spectrum of the wave operator A, and this is expected to be on physical grounds a necessary condition in order to get a finite partition function.

3. No box case

We will adopt the so called ''decomposition method.'' See also Ref. 14, chapter 11. Our analysis is based on Ref. 16, theorem 4, p. 1438.

Let L be a formal differential operator defined on an interval I with end points a,b and let $I=I_1\cup I_2$, where I_1,I_2 are subintervals of I. Let L_1 and L_2 be the restrictions of L to I_1 and I_2, respectively. Then

$$\sigma_e(L)=\sigma_e(L_1)\cup\sigma_e(L_2).$$

In our case we choose

$$I_1=(0,z_b], \quad I_2=[z_b,\infty).$$

Let us analyze the essential spectrum of the operators $B_{nl}|_{[z_b,\infty)}$. Given

$$\lim_{z\to\infty} Q_{nl}(z)=M^2+n^2,$$

according to theorem 16, p. 1448 of Ref. 16, it results in

$$\sigma_e(B_{nl}|_{[z_b,\infty)})=\{\lambda|\lambda\geqslant M^2+n^2\}. \tag{21}$$

Summarizing, in the absence of the box, the essential spectrum is nonvoid for every value of n and for every value of M.

B. Boundary conditions at $y=0$ and analysis on $\mathscr{M}_E$

We have to choose a boundary condition for our operators at the singular point $y=0$. In the case $y\in(0,\infty)$ there is no boundary condition to choose at infinity (LPC); we have a one parameter family of self-adjoint extensions of the p.w.o., A_{0l}. In the case $y\in(0,y_b]$, *a priori* one could choose mixed boundary conditions, i.e., periodic boundary conditions. *A priori*, the given family of p.w.o. with $n=0$ admits a four-parameter family of self-adjoint extensions. We are not interested in the general case and we limit ourselves to selecting a Dirichlet boundary condition at $y=y_b$ and to choosing separated boundary conditions for $y=0$ and $y=y_b$. In fact, the presence of a spherical box should not modify the boundary condition one has at $y=0$ when there is no box. This amounts necessarily to choosing a separated boundary condition. With a fixed boundary condition at $y=y_b$ and separated boundary conditions, we have *a priori* also in the case with a spherical box a one-parameter family of self-adjoint extensions for the p.w.o., A_{0l}. Note that, A_{nl} being bounded from below, we can choose the Friedrichs extension that would require a separated set of boundary conditions too, as it results from Ref. 16, remark $(D5)$, p. 1545. At the regular point $y=y_b$, this amounts to selecting just a Dirichlet boundary condition $\zeta(y_b)=0$.

When the point $y=0$ is characterized by the LCC, the boundary condition, i.e., the self-adjoint extension, is selected according to the following physical criterion: If the thermal manifold is chosen to be $\mathscr{M}$, then the singularity appearing at $y=0$ is due to incompleteness (but not to a singularity; no singularity arises in the curvature invariants as $y=0$, as far as the on shell case is considered) in the manifold. A natural choice for the boundary condition in $y=0$ can be the one not introducing ''extra-sources'' at the singularity itself (cf. the appendix of Ref. 17 and the discussion of the off shell case).

In a certain sense, the singularity is a removable one. In our case, the natural choice is the boundary condition corresponding to Friedrichs' ''ausgezeichnete Randbedingung,'' [18] which in the case in which $y \in (0,\infty)$ corresponds to the choice of the Friedrichs extension for A_{0l}. This boundary condition at $y=0$ and the Dirichlet one at $y=y_b$ define the Friedrichs extension for A_{0l} in the $y \in (0,y_b]$ case.

We note that our choice corresponds in fact to the choice of the Friedrichs extension for A. Indeed, in general the Friedrichs extension of a direct sum $C=\Sigma_k^{\oplus} C_k$, $k \in \mathbb{N}$ of operators C_k bounded from below is equal to the direct sum of the Friedrichs extensions of the operators C_k (cf. Ref. 13, exercise 5.43). This fact implies in our case that the Friedrichs extension of C has the same reducing subspaces as C.

It is interesting to see what happens if the thermal manifold is the whole Euclidean section $\mathscr{M}_E$. We get that *the wave operator A is essentially self-adjoint* on $C_0^{\infty}(\mathscr{M}_E)$. Indeed, $(\mathscr{M}_E, g_{\mu\nu})$, where $g_{\mu\nu}$ is the metric given in Kruskal coordinates, is a complete Riemannian manifold and, given a complete Riemannian manifold $\mathscr{M}_R$, the wave operator $A=-\Delta+M^2$ defined on $C_0^{\infty}(\mathscr{M}_R)$ is essentially self-adjoint.[7,8] For the completeness of the Euclidean section, see, e.g., Ref. 19, example (a), p. 271 and see also Appendix B herein. So the wave operator A on $\mathscr{M}_E$ gets $(N_+(\overline{A}),N_-(\overline{A}))=(0,0)$. Moreover, A is also bounded from below,[8] and both these conditions imply that its unique self-adjoint extension is the Friedrichs extension.[9] The introduction of a spherical boundary in $\mathscr{M}_E$ gives rise to a manifold like $D\times S^2$, where D is a 2D disk. A becomes not essentially self-adjoint, and a self-adjoint extension is obtained if, e.g., a Dirichlet boundary condition is imposed. Then it is known that the spectrum of A is purely discrete and that Friedrichs extension of the minimal operator corresponds to a Dirichlet boundary condition.[7]

A question that could be raised concerns which relation exists (for simplicity we consider the case of no spherical box) between the Friedrichs extension $\overline{A}$ of the essentially self-adjoint wave operator A on $C_0^{\infty}(\mathscr{M}_E)$, and the Friedrichs extension A^F of the direct sum of p.w.o. It is easy to show that they actually coincide. For a demonstration see Appendix D. This means that, according to the above reasoning, the only possible choice for the boundary condition at $y=0$ for A_{0l} is the one implementing the Friedrichs extension in the absence of the box, which is the ''ausgezeichnete Randbedingung.'' See also the conclusions. [In the case of $\mathscr{M}_E$, one has that $y=0$ is a coordinate singularity due to a bad choice of the chart. In the given chart for the Euclidean section, the problem is analogous to the one related to the use of a polar coordinate chart in the plane, where the origin cannot be covered by the given chart and the wave equation becomes singular there. This singularity is only a coordinate singularity. This analogy can be easily realized by means of the relations (15).]

An explicit implementation of the ''ausgezeichnete Randbedingung'' is given in Ref. 18, sec. 3. One has to deal with an operator A_{nl} semibounded at $y=0$. A real fundamental system of solutions ψ,ω of the equation $A_{nl}u=\mu u$ exists such that they do not vanish in $0<y\leqslant s<y_b$ (s generic) and that

$$\lim_{y\to 0} \frac{\omega}{\psi}=0.$$

Then the ''ausgezeichnete Randbedingung'' is

$$\lim_{y\to 0} \frac{u}{\psi}=0,$$

and it is not dependent on μ, ψ, ω. Such a real fundamental system in our case $n=0,l$ near $y=0$ can be chosen as

$$\omega=1+a_1 y+O(y^2),$$

$$\psi=b_1 y+(1+a_1 y)\log(y)+O(y^2),$$

so Friedrichs' boundary condition is

$$\lim_{y\to 0}\frac{\zeta(y)}{\log(y)}=0.$$

C. Off shell case

If one chooses an inverse temperature β not equal to β_h, then distributional curvature singularities develop near $y=0$ in (13).[12] One can show that the Ricci scalar is given by

$$R=2(1-\gamma)\,\frac{\delta(y)}{y}.$$

We will start by treating simply the case of minimal coupling $\xi=0$. The $\xi\neq 0$ case is shortly discussed at the end of this subsection. The minimal coupling case is somehow special, because no curvature singularity appears in the wave equation, the only term having distributional nature being the coupling to the Ricci scalar curvature. It is evident that the only change in the on shell case formulas consists in the following substitution:

$$n^2\to\frac{n^2}{\gamma^2}.$$

1. Deficiency indices of A

The eigenvalue equation is

$$\left(-\partial_y^2-\frac{1+3y^2}{y(1+y^2)}\partial_y+(1+y^2)\left(\frac{(1+y^2)}{y^2}\,\frac{n^2}{\gamma^2}+\frac{4l(l+1)}{(1+y^2)^2}+M^2-\lambda\right)\right)\varphi(y)=0,$$

and near $y=0$ one gets

$$\left(-\partial_y^2-\left(\frac{1}{y}+O(y)\right)\partial_y+\left(\frac{n^2}{\gamma^2}\frac{1}{y^2}+2\,\frac{n^2}{\gamma^2}+4l(l+1)+M^2-\lambda+O(y^2)\right)\right)\varphi(y)=0. \qquad (22)$$

The corresponding indicial equation has the following roots:

$$\alpha_1=\frac{|n|}{\gamma},\quad \alpha_2=-\frac{|n|}{\gamma}.$$

The presence of the conical defect affects the characterization of the Weyl's alternative. Both the solutions belong to $L^2((0,c],y(1+y^2)^2dy)$, where $(0,c]$ represents a right neighbor of 0, if

$$\frac{|n|}{\gamma}<1.$$

This means that, if we fix a value of $\gamma<1$, for $n\neq 0$, we get the LPC; for $n=0$ one has the LCC. The choice $\gamma<1$ gives the same qualitative properties as the on shell case, and it is natural if one considers a high temperature off shell limit ($\beta\to 0^+\Leftrightarrow\gamma\to 0^+$).

If we fix $\gamma>1$, i.e., if we fix a temperature lower that the black hole one, then the LCC holds for every n such that $|n|<\gamma$. LPC is verified definitively. Indeed, by posing $\gamma=[\gamma]+\epsilon$, $\epsilon\in[0,1)$, one gets

$$\epsilon>0\Rightarrow \text{LPC},\forall n,\quad \text{such that } |n|\geqslant\bar{n}\equiv[\gamma]+1;$$

$$\epsilon=0\Rightarrow \text{LPC},\forall n,\quad \text{such that } |n|\geqslant\bar{n}\equiv[\gamma].$$

The LPC holds at infinity. Then all the p.w.o. with $|n|<\bar{n}$ have deficiency indices (1,1), whereas all the p.w.o. with $|n|\geqslant\bar{n}$ have deficiency indices (0,0).

The deficiency indices of the wave operator A result to be (∞,∞) also in the off shell case, for every choice of γ. Indeed it holds that

$$N_+(\bar{A}) = \sum_{l\in\mathbb{N}} (2l+1) N_+(\bar{A}_{0l}) + 2\sum_{n=1}^{\bar{n}-1} \sum_{l\in\mathbb{N}} (2l+1) N_+(\bar{A}_{nl}) = \infty,$$

$$N_-(\bar{A}) = \sum_{l\in\mathbb{N}} (2l+1) N_-(\bar{A}_{0l}) + 2\sum_{n=1}^{\bar{n}-1} \sum_{l\in\mathbb{N}} (2l+1) N_-(\bar{A}_{nl}) = \infty.$$

From this point of view, there is a remarkable difference w.r.t. Laplacian operators defined on metric cones like

$$ds^2 = dy^2 + y^2 g_{M_k},$$

where $y \in (0,\infty)$ and g_{M_k} is the (y independent) metric on a k-dimensional compact manifold.[20] Indeed, in the latter case, with self-evident notation, we get

$$\Delta = \frac{1}{y^k}\partial_y(y^k\partial_y) + \frac{1}{y^2}\Delta_k, \tag{23}$$

and only a finite number of eigenvalues of the elliptic operator Δ_k, defined on a compact manifold, can contribute to the deficiency indices of the laplacian Δ, that, as a consequence, gets finite deficiency indices. Moreover, as shown in Ref. 20, the wave operator in (23) is not essentially self-adjoint only for $k=1,2$. The case $k=1$ is extensively discussed in Ref. 21.

Note that for $\gamma>1$ one gets, w.r.t. the $\gamma\leqslant 1$ case, a number of further self-adjoint extensions to define: For each value of l, there are $2[\gamma]$ further self-adjoint extensions to define for $\epsilon>0$ and $2([\gamma]-1)$ for $\epsilon=0$. An analogous problem exists for quantum mechanics on cones with conical defect $\gamma>1$.[21]

The lack of essential self-adjointness of the wave operators can be interpreted in terms of incompleteness of the conical metric. See also Appendix B. The corresponding Riemannian manifold is also nonextendible (cf. Ref. 20 for the metric cone case), so one cannot ''remove'' the singularity in a different chart. The choice of the boundary condition at $y=0$ is then less straightforward in the off shell case, because it lacks the possibility to interpretate the singularity at $y=0$ as a coordinate one. The ''ausgezeichnete Randbedingung'' could be naively privileged as the most directly connected to the on shell case (no conical defect).

Actually, a more careful analysis should be done. We limit ourselves to propose the following conjecture. It is known[22] that a delta-like model potential in a Schrödinger-like equation can be implemented by means of a suitable boundary condition on the Schrödinger operator (bounded from below) without the delta-like potential on a suitable domain. Choices of self-adjoint extensions different from Friedrichs' one correspond to such delta-like potentials. The same could be expected to be true also in this case, so boundary conditions other than the ''ausgezeichnete'' one could be reasonably related to distributional potentials on the two-dimensional surface $y=0$. Then one could also expect that the nonminimal case, characterized by the presence of a ''delta'' potential on $y=0$, can be dealt with by means of the same self-adjoint extensions techniques as for analogous problems in ordinary flat space quantum mechanics.[22,15] If the above hypothesis is right, that is, if it is possible to define the model potential via the self-adjoint extension method, then boundary conditions at $y=0$ different from Friedrichs' one could be sufficient for implementing the nonminimal coupling case. (If this is true, it is reasonable to conjecture that suitable self-adjoint extensions of the $n=0$ case are enough also for the case $\gamma>1$. The other cases could define other distributional interactions we are not interested in. For $n\neq 0$ such that $|n|<\bar{n}$ it would seem natural in this framework to impose Friedrichs boundary condition at $y=0$.)

In the given purely conjectural framework, the coupling ξ, that on shell is irrelevant, becomes a bare coupling related via the renormalization group[22] to the self-adjoint extension parameters and the naive approach to privilege the Friedrichs' boundary condition also in the case $\xi\neq 0$ should be better justified.

Before a discussion about qualitative spectral properties, we characterize the self-adjoint extensions of the p.w.o., A_{nl}, in the presence of a spherical boundary at $y=y_b$. Obviously, defi-

ciency indices become (2,2) for $|n|<\overline{n}$ and (1,1) for $|n|\geqslant\overline{n}$. Taking into account that we will limit ourselves to consider a Dirichlet separated boundary condition at $y=y_b$, it holds for the essential self-adjointness domain on $\mathscr{U}_{nl}$ for the self-adjoint extensions of A_{nl},

$$\mathscr{U}_{nl}=\{\zeta|\zeta\in C^{\infty}(0,y_b]\cap L^2((0,y_b],y(1+y^2)^2\ dy),$$

$$A_{nl}\zeta\in L^2((0,y_b],y(1+y^2)^2\ dy),[\zeta,\zeta_1](y=0)=0;\quad \zeta(y=y_b)=0\},$$

where $[u,v]\equiv p(y)(u(dv/dy)-(du/dy)v)$; $\zeta_1\not\equiv 0$ satisfies Eq. (17) for Im $\lambda\neq 0$ and is such that $[\zeta_1,\zeta_1](0)=0$.[23]

Given the general structure of the equation near the regular singular point $y=0$ [cf. equation (22)], we can express the condition at that point by following Ref. 23, chapter 8 or, equivalently, Ref. 24. As a preliminary step, we give some definitions. Let $\alpha_1\geqslant\alpha_2$ be the roots of the indicial equation. Moreover, define the following fundamental system:

$$\zeta_+(y)=y^{\alpha_1}\sum_{k=0}^{\infty}\ \omega_k y^k,$$

$$\zeta_-(y)=y^{\alpha_2}\sum_{k=0}^{\infty}\ \psi_k y^k+c\zeta_+(y)\log(y);$$

the coefficients ω_k, ψ_k, c are uniquely determined. (Particularly, it holds that $\omega_0=1$ and $\psi_0=1$ and $c=0$ for $\alpha_1-\alpha_2\neq 0,1,2,\ldots$; $\psi_0=0$ and $c=1$ for $\alpha_1-\alpha_2=0$; $\psi_0=1$ and $\psi_n=0$ for $\alpha_1-\alpha_2=n$, $n=1,2,\ldots$.) Then let us suitably define

$$\omega_2(y)=\omega_0+\omega_1 y,$$

$$\psi_2(y)=\psi_0+\psi_1 y,\quad \zeta_{1,2}(y)=y^{\alpha_1}\omega_2(y),$$

$$\zeta_{2,2}(y)=y^{\alpha_2}\psi_2(y)+c\zeta_{1,2}(y)\log(y).$$

(It is easy to show that the parameter σ appearing in Ref. 23 is equal to 2 in our case. For further details, see Ref. 23.) Then, define for $\zeta(y)\in\mathscr{U}_{nl}$,

$$a_0=\lim_{y\to 0}\frac{\zeta(y)}{y^{\alpha_2}},\quad a_1=\lim_{y\to 0}\frac{\zeta(y)-a_0\zeta_{2,2}(y)}{y^{\alpha_1}},$$

for real $\alpha_1\neq\alpha_2$ and

$$a_0=\lim_{y\to 0}\frac{\zeta(y)}{y^{\alpha_1}\log(y)},\quad a_1=\lim_{y\to 0}\frac{(\zeta(y)-a_0\zeta_{2,2}(y)}{y^{\alpha_1}},$$

for real $\alpha_1=\alpha_2$. The above numbers a_0,a_1 satisfy the conditions to be the so called *Rellich initial numbers*[23] and eventually this means that $\mathscr{U}_{nl}$ can be written as

$$\mathscr{U}_{nl}=\{\zeta|\zeta\in C^{\infty}(0,y_b]\cap L^2((0,y_b],y(1+y^2)^2dy),$$

$$A_{nl}\zeta\in L^2((0,y_b],y(1+y^2)^2dy),\ \zeta(y=y_b)=0;\quad a_0\cos(s)+a_1\sin(s)=0,\ s\in[0,\pi)\}.$$

The equivalence between the condition $a_0\cos(s)+a_1\sin(s)=0$ and a condition like

$$\nu a_0=a_1;\quad \nu\in(-\infty,+\infty]$$

is trivial. The ''ausgezeichnete Randbedingung'' at $y=0$ is equivalent to

$$a_0=0,$$

that is, $s=0$ or $\nu=\infty$. We recall that in our case $\alpha_1=|n|/\gamma>\alpha_2=-|n|/\gamma$ if $n\neq 0$ and $\alpha_1=\alpha_2=0$ for $n=0$.

The above characterization is general and it holds for every choice of γ.

2. Spectral analysis

All results about qualitative properties of the spectrum remain unaltered w.r.t. the on shell case. All the p.w.o. with deficiency indices (2,2) have, from Ref. 16 (corollary 12, p. 1402), a void essential spectrum. The same is true in the case of deficiency indices (1,1): the same choice of the variable z and of the Liouville transformation as in the on shell case carries (17) into $B_{nl}u(z)=\lambda u(z)$, where

$$B_{nl}=-\frac{d^2}{dz^2}+Q_{nl}(z),\quad Q_{nl}(z)=\left(\frac{n^2}{\gamma^2}-\frac{1}{4}\right)\frac{1}{z^2}\frac{z^2}{y(z)^2}+W_{nl}(z),$$

$$W_{nl}(z)=\frac{n^2}{\gamma^2}+M^2+\frac{4l(l+1)}{(1+y(z)^2)^2}+\frac{1}{4(1+y(z)^2)^3}(11+3y(z)^2+y(z)^4).$$

From

$$\lim_{z\to 0^+} z^2Q_{nl}(z)=\frac{n^2}{\gamma^2}-\frac{1}{4}>-\frac{1}{4},$$

and from criterion (25) of Ref. 16, p. 1598, it is proven that the essential spectrum of any self-adjoint extension is void. If $y\in(0,\infty)$, then every p.w.o. gets an essential spectrum contribution, and a completely analogous approach to the one for the $\gamma=1$ case gives

$$\sigma_e(A_{nl})=\left\{\lambda\,\middle|\,\lambda\geqslant\frac{n^2}{\gamma^2}+M^2\right\}.$$

D. Noncovariant measure

In the case when one chooses the noncovariant measure, essential self-adjointness properties and qualitative properties of the spectrum of the wave operator change remarkably. The following discussion will hold both for the on shell case $\gamma=1$ and for the off shell case with minimal coupling $\xi=0$.

The partial wave operators are

$$A_{nl}=\frac{y}{f(y)^3}(-\partial_y(yf(y)\partial_y)+V_{nl}(y)),\quad V_{nl}(y)=\frac{f(y)^3}{y}v_{nl}(y),$$

$$v_{nl}(y)=\frac{4y^2}{f(y)^3}l(l+1)+M^2\frac{y^2}{f(y)}+\frac{n^2}{\gamma^2},\quad f(y)=1+y^2.$$

The Hilbert space containing the domain is $L^2(I,[f(y)^3/y]dy)$, where $I=(0,\infty)$ if no spherical box is introduced. For $y\to 0^+$ the asymptotic form of the differential equation (17),

$$\left(-\partial_y^2-\frac{1}{y}\partial_y+\frac{1}{y^2}\left(\frac{n^2}{\gamma^2}-\lambda\right)\right)\zeta(y)=0,$$

gives the following secular roots:

$$\alpha_+=\sqrt{\frac{n^2}{\gamma^2}-\lambda},\quad \alpha_-=-\sqrt{\frac{n^2}{\gamma^2}-\lambda}.$$

By choosing, e.g., $\lambda=0$ it is straightforward to realize that at $y=0$ the LPC is verified as well as at infinity (cf. also Ref. 23, theorem 1, p. 251). The deficiency indices are then (0,0) for each A_{nl}. So for A_{NC} one obtains $(N_+,N_-)=(0,0)$.

The introduction of a spherical box restricts the y-interval to $(0,y_b]$. The point $y=y_b$ is a regular point for each p.w.o., so that the LCC is verified there.

Qualitative spectral properties can be inferred by means of the following variable transformation (tortoise-like),

$$z=-\log(y)-\frac{y^2}{2}, \quad z\in[-z_b,\infty),$$

such that the rescaled metric becomes

$$ds^2=\frac{y(z)^2}{(1+y(z)^2)}(d\tau^2+dz^2)+\frac{1}{4}(1+y(z)^2)^2\, d\Omega^2; \tag{24}$$

the Liouville transformation,

$$\varphi=\frac{\zeta(y(z))}{f(y(z))},$$

then carries Eq. (17) into an eigenvalue equation for the operator,

$$B_{nl}=-\partial_z^2+Q_{nl}(z),$$

$$Q_{nl}(z)=\frac{4y(z)^2}{f(y(z))^3}l(l+1)+M^2\frac{y(z)^2}{f(y(z))}+\frac{n^2}{\gamma^2}+\frac{4y(z)^2}{f(y(z))^4}.$$

It holds

$$\lim_{z\to\infty} Q_{nl}(y(z))=\lim_{y\to 0^+} Q_{nl}(y)=\frac{n^2}{\gamma^2},$$

so, according to theorem 3, p. 1448, of Ref. 16, one has

$$\sigma_e(A_{nl})=\left\{\lambda\,\middle|\,\lambda\geqslant\frac{n^2}{\gamma^2}\right\}.$$

A noncovariant measure, that is, the choice of a noncovariant scalar product, gives rise to a nonvoid essential spectrum for every n. This means that spectral properties of the p.w.o. are deeply modified by the choice of the measure.

By means of the decomposition method, we can also see what happens if there is no spherical box, i.e., for $y\in(0,+\infty)=(0,y_b]\cup[y_b,+\infty)$. The analysis on the latter interval is completely analogous to the one on the former and one gets

$$\sigma_e(A_{nl}|_{[y_b,+\infty)})=\left\{\lambda\,\middle|\,\lambda\geqslant\frac{n^2}{\gamma^2}+M^2\right\}.$$

[The variable to choose is $z=\log(y)+y^2/2\in[c,\infty)$; the Liouville transformation and the potential $Q_{nl}(y)$ are the same as in the first interval.] Then the essential spectrum of the p.w.o. in the absence of the box proves to be the same as if there is the box.

Regarding the nonminimally coupled off shell case, we limit ourselves to note that there is a distributional contribute to the potential given by

$$\xi R=\xi 2(1-\gamma)\frac{y\,\delta(y)}{f(y)},$$

and that, for noncovariant measure, the problem of studying the nonminimal case appears to be even more difficult than for a covariant choice of the measure. Indeed, as it is clear for $y\in(0,\infty)$, there is no possibility to define distributional contributions by means of the self-adjoint

extension technique: The $\xi=0$ wave operator A is essentially self-adjoint, so that there is no freedom to define self-adjoint extensions.

IV. EXTREMAL BLACK HOLES

Reissner–Nordström electrically charged black holes are the most direct black hole spherosymmetrical extensions of the Schwarzschild solution. They are characterized by two geometrical hair, the mass m and the charge e and by a Ricci scalar $R=0$ (but they are not Ricci flat). We will consider only uncharged scalar matter fields on these charged black hole backgrounds. Then a discussion of the nonextreme case is entirely analogous to the one for the Schwarzschild case, and for sake of completeness it is sketched in Appendix A. Here we will focus our study on the degenerate black hole case: when $e^2=m^2$ the event horizon is not a bifurcate Killing horizon and near the horizon there is no conical singularity. In the Schwarzschild gauge the metric in (2) has

$$g_{00}(r)=\frac{(r-r_{bh})^2}{r^2},$$

and we can choose an arbitrary value for the period β of the imaginary time τ due to the fact that there is no conical structure near the horizon. Rescaling by r_{bh} the variables τ,r (we will still call them τ,r) and introducing the variable $y=r-1\in(0,y_b]$ one obtains

$$ds^2=(r_{bh})^2\left(\frac{y^2}{(y+1)^2}d\tau^2+\frac{(y+1)^2}{y^2}dy^2+(y+1)^2d\Omega^2\right).$$

We stress that in this case no thermodynamical manifold is defined as different from $\mathscr{M}=S^1\times(0,\infty)\times S^2$. So we will work exclusively on $\mathscr{M}$. The Riemannian manifold $(\mathscr{M},g_{\mu\nu})$ results to be complete, as it is shown in Appendix B. So the wave operator A has to be essentially self-adjoint in the absence of a spherical box. This fact will result also from the p.w.o. analysis. Note that completeness of the given Riemannian manifold implies that it is nonextendible.

The partial wave operators are

$$A_{nl}=\frac{1}{(y+1)^2}(-\partial_y(y^2\partial_y)+Q_{nl}(y)),$$

$$Q_{nl}(y)=(y+1)^2q_{nl}(y),$$

$$q_{nl}(y)=\frac{(y+1)^2}{y^2}n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2+\frac{l(l+1)}{(y+1)^2}+M^2,$$

and the Hilbert space containing the domain is $L^2(I,(1+y)^2dy)$, where $I=(0,\infty)$ in the absence of the box.

A. Deficiency indices of *A*

Equation (17) is

$$\left(-\partial_y^2-\frac{2}{y}\partial_y+\frac{(1+y)^2}{y^2}(q_{nl}(y)-\lambda)\right)\varphi(y)=0. \tag{25}$$

Then for $y\to0^+$ one gets

$$\left(-\partial_y^2-\frac{2}{y}\partial_y+\left(\left(\frac{n^2}{y^4}+\frac{4n^2}{y^3}\right)\left(\frac{2\pi r_{bh}}{\beta}\right)^2+\frac{6n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2+l(l+1)+M^2-\lambda}{y^2}+O\left(\frac{1}{y}\right)\right)\right)\varphi(y)=0. \tag{26}$$

For $n\neq0$ there is an irregular singularity for $y\to0$ and at the leading order the solutions are (cf. Ref. 16, p. 1432)

$$\varphi_1 = \exp\left(\frac{|n|\left(\frac{2\pi r_{bh}}{\beta}\right)}{y} \right) y^{-2|n|(2\pi r_{bh}/\beta)}(1+O(y)),$$

$$\varphi_2 = \exp\left(\frac{-|n|\left(\frac{2\pi r_{bh}}{\beta}\right)}{y} \right) y^{2|n|(2\pi r_{bh}/\beta)}(1+O(y)).$$

They are characterized by an essential singularity at $y=0$ and, existing for every λ a solution not square integrable near the origin, one gets the limit point case.

If $n=0$, then the singularity in (26) becomes a regular one and the corresponding indicial equation admits the following roots:

$$\alpha_1 = \tfrac{1}{2}(-1+c_\lambda), \quad \alpha_2 = \tfrac{1}{2}(-1-c_\lambda),$$

$$c_\lambda = \sqrt{1+4l(l+1)+4M^2-4\lambda}.$$

It is easy to see that, for every choice of λ, l, M one finds the LPC. (If $M>0$ one can choose, e.g., $\lambda=0$ and verify for every l the LPC according to Weyl's alternative (cf. Ref. 13). The same value $\lambda=0$ is suitable for $M=0$, $l>0$. If $M=0$, $l=0$ then a suitable choice in order to see that one has LPC could be $\lambda=1/4$. See also theorem 1, p. 251, of Ref. 23.)

The LPC is verified also at infinity. Then the p.w.o. have deficiency indices (0,0) and there is no boundary condition to choose at $y=0$ in the extremal case.

B. Spectral properties

In the following, $I=(0,y_b]$. We define the new variable z,

$$\frac{dz}{dy} = -\left(1+\frac{1}{y}\right), \quad z(y) = -y-\log(y), \quad z\in[z_b,\infty),$$

in such a way as to get the following rescaled metric:

$$ds^2 = \frac{y(z)^2}{(y(z)+1)^2} d\tau^2 + dz^2 + (y(z)+1)^2\, d\Omega^2;$$

by posing

$$\varphi = \frac{g(y)}{s(y)}, \quad s(y) = \sqrt{y(y+1)},$$

Eq. (25) becomes

$$[\partial_z^2 + (\lambda - R_{nl}(z))]g(z) = 0,$$

where

$$R_{nl}(z) = \frac{\partial_z^2 s(y(z))}{s(y(z))} + \frac{Q_{nl}(y(z))}{(y(z)+1)^2}$$
$$= \frac{1}{4}\frac{(4y(z)+1)}{(y(z)+1)^4} + \frac{(y(z)+1)^2}{y(z)^2} n^2 \left(\frac{2\pi r_{bh}}{\beta}\right)^2 + \frac{l(l+1)}{(y(z)+1)^2} + M^2.$$

[We note that these coordinates allow us to determine very easily if the LPC is verified at $y=0 \Leftrightarrow z=\infty$. Indeed, being $R_{nl}(z)$ is a positive function, conditions for LPC at $z=\infty$ according to

the corollary to theorem 6.3 of Ref. 14 are easily verified.] We can determine if the essential spectrum is void or not according to Ref. 16, theorem 16, p. 1448. We have to find $\lim_{z\to+\infty} R_{nl}(z)$, which is equivalent to $\lim_{y\to 0^+} R_{nl}(y)$. We get

$$n\neq 0 \Rightarrow \lim_{z\to+\infty} R_{nl}(z)=+\infty \Leftrightarrow \sigma_e(A_{nl})=\varnothing,$$

$$n=0 \Rightarrow \lim_{z\to+\infty} R_{0l}(z)=\tfrac{1}{4}+l(l+1)+M^2 \Leftrightarrow \sigma_e(A_{0l})=\{\lambda\,|\,\lambda\geqslant \tfrac{1}{4}+l(l+1)+M^2\}.$$

Due to the nonvoid essential spectrum contribution of the partial wave operators with $n=0$, one finds that $\sigma_e(A)\neq\varnothing$. Physical implications of this very interesting result are briefly discussed in the conclusions.

It is simple to show that, in the absence of the box, i.e., when $y\in(0,+\infty)\Leftrightarrow z\in(-\infty,+\infty)$, the essential spectrum is nonvoid for every value of n, and $\sigma_e(A_{nl})=\{\lambda\,|\,\lambda\geqslant M^2+n^2(2\pi r_{bh}/\beta)^2\}$.

C. Noncovariant measure

The noncovariant choice of the measure leads to the following partial wave operators:

$$A_{nl}=\frac{y^2}{(y+1)^4}(-\partial_y(y^2\partial_y)+S_{nl}(y)),$$

$$S_{nl}(y)=\frac{(y+1)^4}{y^2}n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2+l(l+1)+M^2(1+y)^2.$$

The domain of the p.w.o. is contained in $L^2((0,\infty),[(y+1)^4/y^2]dy)$. The study of Weyl's alternative amounts to determining if near $y=0$ either the LPC or the LCC holds for the solutions of the equation

$$\left(-\partial_y^2-\frac{2}{y}\partial_y+\frac{\left(n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2-\lambda\right)(1+y)^4}{y^4}+\frac{l(l+1)}{y^2}+\frac{M^2(1+y)^2}{y^2}\right)\zeta=0.$$

For all the possible values of the parameters, the LPC holds and so the deficiency indices of the p.w.o. are (0,0) (cf. also corollary of theorem 6.3 of Ref. 14 in the following coordinate choice). Then the wave operator A is essentially self-adjoint.

Qualitative spectral properties in the spherical box are easily found by means of the following variable transformation:

$$z(y)=\frac{1}{y}-2\log(y)-y;\quad z\in[z_b,\infty), \tag{27}$$

such that the rescaled metric becomes

$$ds^2=\frac{y^2}{(y+1)^2}(d\tau^2+dz^2)+(y+1)^2\,d\Omega^2.$$

Then the following Liouville transformation:

$$\zeta(y(z))=\frac{\varphi(y(z))}{1+y(z)},$$

allows us to get the eigenvalue equation

$$(-\partial_z^2+Q_{nl}(z))\varphi=\lambda\varphi,$$

where

$$Q_{nl}(z)=n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2+M^2\,\frac{y(z)^2}{(1+y(z))^2}+\frac{y(z)^2}{(1+y(z))^4}l(l+1)+\frac{2y(z)^3}{(1+y(z))^6}.$$

According to theorem 16, p. 1448 of Ref. 16, given

$$\lim_{z\to\infty} Q_{nl}(y(z))=\lim_{y\to 0} Q_{nl}(y)=n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2,$$

we find that the essential spectrum is the set

$$\sigma_e(A_{nl})=\left\{\lambda\,\middle|\,\lambda\geqslant n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2\right\}, \tag{28}$$

and it is nonvoid for all the p.w.o.

The essential spectrum contribution arising from the p.w.o. restricted to the interval $y\in[y_b,+\infty)$ is given by

$$\sigma_e(A_{nl}|_{[y_b,\infty)})=\left\{\lambda\,\middle|\,\lambda\geqslant n^2\left(\frac{2\pi r_{bh}}{\beta}\right)^2+M^2\right\},$$

so that the essential spectrum of the p.w.o. is the same as in (28).

V. CONCLUSIONS

The analysis of scalar field wave operators on the Euclidean section of a Schwarzschild black hole has been carried in the so called on shell case and in the off shell one for minimal coupling to the curvature. In the case of the covariant choice of the functional measure, the properties of the physical self-adjoint extension of the wave operators to be chosen show that on shell the natural choice for the thermal manifold is $\mathscr{M}_E$ because of the nature of the ''singularity'' at the horizon to be only a coordinate one. Indeed, even if one starts with $\mathscr{M}$, the boundary condition to be chosen is such that the corresponding self-adjoint extension coincides with the one calculated on $\mathscr{M}_E$. The off shell case is different from the on shell one, because going off shell means being forced to work on an incomplete Riemannian manifold and to impose a boundary condition at $r=r_{bh}$ that cannot be selected on firm physical grounds (the shift from the equilibrium temperature being unphysical) and whose choice is also nontrivial from a mathematical point of view.

By posing a spherical box around the black hole, corresponding in all cases to a finite Euclidean four-volume V_4, we have found that the essential spectrum is void only if the covariant measure for the path integral is chosen.

In the case of the noncovariant choice, one finds that A_{NC} is essentially self-adjoint on $C_0^\infty(\mathscr{M})$ and has a nonvoid essential spectrum in a finite volume (cf. also Ref. 4) unless another infrared regularization is introduced near the horizon. The equivalence of the spectral problem of A_{NC} with the spectral problem of A_{opt} on the optical manifold gives a qualitative insight about spectral properties; indeed, the corresponding four-volume $V_{4\,optical}$ is divergent unless a further infrared regularization is introduced near the ''radial'' coordinate $r=r_{bh}$.

An analogous study has been generalized to the case of electrically neutral scalar fields on nondegenerate Reissner–Nordström black hole manifolds and analogous results have been obtained.

The extremal case has revealed some very peculiar features. First of all, the corresponding Riemannian manifold is complete for any β-periodicity of the Euclidean time τ, and so it lacks the possibility to define in this way a preferred geometric temperature for the thermal equilibrium of the quantum field on the extremal black hole background. Any choice of β ensures essential self-adjointness of the wave operator on $C_0^\infty(\mathscr{M})$. From this point of view thermal quantum field theory on the extremal background is not different from, e.g., thermal quantum field theory in flat space and there is no apparent contradiction in interpreting this fact as a lack of thermal effects associated with the given black hole background, at least as far as the Euclidean path integral

approach is pursued. A second relevant feature is the presence of a nonvoid essential spectrum, whichever choice of the measure one makes. So in the extremal case it is not possible to get a discretization of physical states by posing a spherical box with reflecting boundary conditions at a fixed distance from the event horizon. Only by also introducing a further boundary near the degenerate horizon can one get a discrete spectrum. Anyway, it has to be noted that for the covariant choice the only contribution to the essential spectrum of the wave operator comes from the p.w.o. with $n=0$ and that the subspace $n=0$ in the Euclidean approach to ''standard'' finite temperature quantum field theory corresponds to a zero temperature contribution (vacuum contribution);[25] so it does not seem to affect the thermal properties of the system.

Further investigations on these topics will be the subject of future works.

ACKNOWLEDGMENTS

The author wishes to thank especially F. Valz-Gris for his very helpful remarks and his careful bibliographic suggestions. Thanks are also addressed to A. S. Cattaneo, S. Desantis, A. Santambrogio, and M. Spreafico for some discussions about differential geometry.

APPENDIX A: NONEXTREMAL REISSNER–NORDSTRÖM BLACK HOLES

The nondegenerate Reissner–Nordström black hole is characterized by a mass m and by an electric charge e, with $m^2<e^2$. A conical structure near the event horizon is again present and this gives rise qualitatively to the same analysis as in the Schwarzschild case. The metric in the external region has the same form as in (2) with

$$g_{00}(r)=\frac{(r-r_{bh})(r-r_c)}{r^2},$$

where r_{bh} is the event horizon radius and r_c is the Cauchy horizon radius, with $r_c<r_{bh}$. The conical defect is avoided if the period β is fixed at the following value:

$$\beta_h=\frac{2\pi}{k}=4\pi\frac{r_{bh}^2}{r_{bh}-r_c}, \quad k=\frac{r_{bh}-r_c}{2r_{bh}^2}.$$

We define the rescaled variables $\tau\to k\tau\in[0,2\pi)$; $r\to kr\in(kr_{bh},\infty)$ (if a spherical box at $r=R>r_{bh}$ is introduced then $r\to kr\in(kr_{bh},kR]$). By means of the change of variable,

$$z=\sqrt{2}\sqrt{r-kr_{bh}}\in(0,\infty),$$

one finds

$$ds^2=\frac{1}{k^2}\left(z^2\frac{h(z)}{f(z)}d\tau^2+\frac{f(z)}{h(z)}dz^2+\frac{1}{4}f(z)d\Omega^2\right),$$

$$f(z)=(\delta+z^2)^2, \quad h(z)=\delta^2+z^2,$$

$$\delta=2kr_{bh}=1-\frac{r_c}{r_{bh}}\in(0,1).$$

Note that the value $\delta=0$ corresponds to the extremal case, whereas $\delta=1$ is the Schwarzschild one. We will consider only noncharged scalar matter, so there is no need to introduce a chemical potential and a grand canonical partition function.

1. Deficiency indices of *A*

Also in this case it is possible to show that the Euclidean manifold $\mathcal{M}=S^1\times(0,\infty)\times S^2$ can be extended to a complete Riemannian manifold $\mathcal{M}_E$. The principle is the same as in the Schwarzschild case. See also Appendix B. Then the wave operator A is essentially self-adjoint on $C_0^\infty(\mathcal{M}_E)$.

The partial wave operators become in this case

$$A_{nl}=-\frac{1}{zf(z)}\partial_z(zh(z)\partial_z)+V_{nl}(z),$$

$$V_{nl}(z)=\frac{n^2}{z^2}\frac{f(z)}{h(z)}+\frac{4l(l+1)}{f(z)}+M^2.$$

The differential equation (17) is equivalent to

$$\left(-\partial_z^2-\frac{3z^2+\delta^2}{z(z^2+\delta^2)}\partial_z+\frac{(z^2+\delta)^2}{z^2+\delta^2}(V_{nl}(z)-\lambda)\right)\zeta=0,$$

and its asymptotic expansion for $z\to 0^+$ is

$$\left(-\partial_z^2-\frac{1}{z}\partial_z+\frac{n^2}{z^2}\right)\zeta=0.$$

The indicial equation is the same as in the Schwarzschild case, and one finds the LCC for all the p.w.o. with $n=0$ and the LPC for all the p.w.o. with $n\neq 0$.

2. Spectral analysis

In the case of the p.w.o. with $n=0$, one has the LCC at both the extremes of the interval $(0,z_b]$ and the deficiency indices are (2,2), so one can conclude that the spectrum is purely discrete. In order to study the case with $n\neq 0$, it is useful to introduce the following variable:

$$\frac{dx}{dz}=\frac{z^2+\delta}{\sqrt{z^2+\delta^2}},$$

$$x(z)=\frac{1}{2}z\sqrt{z^2+\delta^2}+\delta\left(1-\frac{\delta}{2}\right)\log|z+\sqrt{z^2+\delta^2}|+C\in(0,x_b],$$

$$C=-\delta\left(1-\frac{\delta}{2}\right)\log|\delta|,$$

in such a way that the metric becomes

$$ds^2=\frac{1}{k^2}\left(z(x)^2\frac{\delta^2+z(x)^2}{(\delta+z(x)^2)^2}d\tau^2+dx^2+\frac{1}{4}(\delta+z(x)^2)^2d\Omega^2\right). \tag{29}$$

The Liouville transformation,

$$\zeta(x)=\frac{1}{\sqrt{z(x)(\delta+z(x)^2)(\delta^2+z(x)^2)^{1/2}}}\phi(x),$$

carries the p.w.o.,

$$A_{nl}=-\partial_x^2-(\partial_x\log(z(x)(\delta+z(x)^2)(\delta^2+z(x)^2)^{1/2})\partial_x+V_{nl}(z(x)),$$

into a form suitable for applying theorems and criteria of Ref. 16,

$$B_{nl}=-\partial_x^2+V_{nl}(z(x))+G(z(x)),$$

$$G=\tfrac{1}{2}\left(\partial_x H+\tfrac{1}{2}H^2\right),$$

$$H=\partial_x\log(z(x)(\delta+z(x)^2)(\delta^2+z(x)^2)^{1/2}),$$

where

$$G=-\frac{\delta^6-2\delta^4 z(x)^2(2+\delta)-\delta^2 z(x)^4(4+20\delta-9\delta^2)-8z(x)^6\delta(2-\delta)}{4z(x)^2(\delta+z(x)^2)^4(\delta^2+z(x)^2)}.$$

We have then

$$B_{nl}=-\partial_x^2+v_{nl}(z(x)),$$

$$v_{nl}(z(x))=\left(n^2-\frac{1}{4}\right)\frac{1}{z(x)^2}+w_{nl}(z(x)),$$

$$w_{nl}=\text{smooth bounded function},$$

and according to criterion (25), p. 1598 of Ref. 16, given

$$\lim_{x\to 0^+} x^2 v_{nl}(z(x))=n^2-\tfrac{1}{4}>-\tfrac{1}{4},$$

one can conclude that the essential spectrum of all the p.w.o. with $n\neq 0$ is empty. Summarizing, one gets a purely discrete spectrum for all the p.w.o.

The off shell case is entirely analogous to the one discussed in the Schwarzschild case; also analogous is the study of the noncovariant measure spectrum.

APPENDIX B: COMPLETENESS OF THE EUCLIDEAN MANIFOLDS

In order to show that the Euclidean manifold in the on shell case $(\mathscr{M}_E, g_{\mu\nu})$ is a complete Riemannian manifold one can proceed as follows. $(\mathscr{M}_E, g_{\mu\nu})$ can be seen as the warped product[26] of the 2D Riemannian manifold $(\mathbb{R}^2, g_{ab})$, where the metric g_{ab} is diagonal (with entries, e.g., in the Schwarzschild case both equal to $(4r_{bh}^3/r)\exp(-r/r_{bh})$), with the 2D sphere S^2 (canonical metric). The warping function[26] is $r^2>0$. In general, a warped product $(A, g_{\mu\nu}(A)) \times_h (B, g_{\mu\nu}(B))$ of two Riemannian manifolds with warping function $h>0$ is complete if and only if $(A, g_{\mu\nu}(A))$ and $(B, g_{\mu\nu}(B))$ are complete Riemannian manifolds.[26] The 2D sphere is compact and then is complete. In order to show that $(\mathbb{R}^2, g_{ab})$ is complete we note that in the ''conical'' chart (29) $\tau\in[0,2\pi); x\in(0,\infty)$ the metric is written in polar coordinates originating at 0, so one can refer to sec. 4 of Ref. 27.

One can equivalently use the following trick, which allows us to give a unified treatment also for studying completeness of the Riemannian manifold in the off shell case and in the extremal one. Define[28] a divergent curve on a manifold $\mathscr{M}$ as a differentiable mapping $\alpha:[0,\infty)\to\mathscr{M}$ such that for any compact set $K\subset\mathscr{M}$ there exists $t_0\in(0,\infty)$ with $\alpha(t)\notin K$, $\forall t>t_0$. The length of the divergent curve is defined by

$$\lim_{t\to\infty}\int_0^t dt|\alpha'(t)|.$$

$(\mathscr{M}, g_{\mu\nu})$ is complete if and only if the length of any divergent curve is unbounded.[28] In all cases the metric gets the form

$$ds_2^2=\eta^2(x)d\tau^2+dx^2,$$

with $\tau\in[0,2\pi)$, $\eta(x)$ a suitable function and $x\in I$. The interval I is equal to $(0,\infty)$ in the nondegenerate case and if $\gamma=1$ can become $[0,\infty)$; it is $(-\infty,+\infty)$ in the extremal case. For such a metric all lines $\tau=\text{const}$ are geodesics. If $I=(0,\infty)$, then the ''radial'' geodesics $\tau=\text{const}$ starting from a given K and ''going towards'' the point $x=0$ are divergent curves of finite length, so the manifold is necessarily incomplete. The same is not true if $I=[0,\infty)$ because in this case all divergent curves have to ''point towards'' $x=\infty$ and so they have to be unbounded.

If $I=\mathbb{R}$ one can use the fact that the divergent curves given by the ''radial'' geodesics have to diverge (in length) at $x=+\infty$ and/or at $x=-\infty$. Indeed, one can, e.g., consider (without loss of generality) compact sets of the type $K=S^1\times[x_0,x_1]$ with $-\infty<x_0<x_1<\infty$. Given a divergent curve ending, say, at $x_d>x_1$ one has that

$$\int_0^{t_d} dt|\alpha'(t)| = \int_0^{t_d} dt \sqrt{\left(\frac{dx}{dt}\right)^2 + \eta^2(x)\left(\frac{d\tau}{dt}\right)^2} \geqslant \int_0^{t_d} dt \left|\frac{dx}{dt}\right|;$$

the last term above can be assumed to represent the length of a purely radial geodesic (suitably reparametrized) having the same starting point and with an end point having the same radial coordinate as the given divergent curve. It is evident that such a radial geodesic is a divergent curve too and that its length is unbounded. Then also the extremal manifold is complete.

Concerning the optical Riemannian manifold $(\mathscr{M},h_{\mu\nu})$, the optical metric is a warped product of a 2D metric,

$$ds_2^2 = d\tau^2 + \frac{1}{g_{00}^2(r)} dr^2,$$

with the canonical metric on S^2. The warping function is $r^2/g_{00}(r)$. Also in this case one can limit showing the completeness of the above 2D metric, but this proves to be an easy task. Indeed, the given metric is a metric on a product manifold $S^1\times I$, where $I=(r_{bh},\infty)$, and it is complete if each factor is complete.[29] A tortoise-like change of radial coordinate carries the above metric into

$$ds_2^2 = d\tau^2 + dx^2,$$

where $x\in\mathbb{R}$. Then each factor is complete.

APPENDIX C: ESSENTIAL SELFADJOINTNESS OF A_{NC}

In this section we show that the noncovariant wave operator (for minimal coupling $\xi=0$ in the conical case) is essentially self-adjoint on $C_0^\infty(\mathscr{M})$ without involving a p.w.o. analysis. The starting point is the well known equivalence between A_{NC} and $A_{opt}\equiv-\Delta_h+q(x)$ given in (11). We can then translate the problem of self-adjointness properties of A_{NC} into the analogous one for A_{opt} on the optical manifold. We know that the optical manifold is a complete Riemannian manifold, so surely the optical Laplacian $-\Delta_h$ is essentially self-adjoint on $C_0^\infty(\mathscr{M})\cap L^2(\mathscr{M},d^4x\sqrt{h})$. We can show that the sum $-\Delta_h+q(x)$ is essentially self-adjoint on the same domain. Indeed, we have, e.g. using the tortoise ''radial'' coordinate x, that

$$q(y(x)) = M^2 \frac{y(x)^2}{(1+y(x)^2)} - \frac{1}{(1+y(x)^2)^4}$$

in the Schwarzschild case (if $\gamma\neq1$ then $\xi=0$); in the extremal case it holds that

$$q(y(x)) = M^2 \frac{y(x)^2}{(1+y(x))^2} + 2\frac{y(x)^3}{(1+y(x))^6}.$$

The above potentials are smooth bounded functions on $\mathscr{M}$. As operators in $L^2(\mathscr{M},d^4x\sqrt{h})$; they are self-adjoint bounded multiplication operators Q_q having domain $D(Q_q)=\{\phi|q\phi\in L^2(\mathscr{M},d^4x\sqrt{h})\}=L^2(\mathscr{M},d^4x\sqrt{h})$. In particular, we get trivially $C_0^\infty(\mathscr{M})\subset D(Q_q)$. Then we can use the Rellich–Kato theorem[13] because Q_q, being bounded, is also $(-\Delta_h)$-bounded with $(-\Delta_h)$-bound 0.

We can conclude that A_{NC} is essentially self-adjoint on $C_0^\infty(\mathscr{M})\cap L^2(\mathscr{M},d^4x(\sqrt{g}/g_{00}))$.

APPENDIX D: FRIEDRICHS EXTENSION FOR THE ON SHELL CASE

We want to show that the Friedrichs extension A^F of the direct sum of the p.w.o. coincides with the closure $\overline{A}$ of the essentially self-adjoint operator A defined on $C_0^\infty(\mathscr{M}_E)$.

The first step is to show that a core $D(A^F)$ of A^F is contained in the core $C_0^\infty(\mathscr{M}_E)$ of $\overline{A}$. We have that $D(A^F)$ is given by

$$D(A^F)=L\{D_{nl}\otimes K_n\otimes S_l : n,l\in\mathbb{N}\},$$

$$D_{nl}=C_0^\infty(0,\infty) \quad \forall n\neq 0,$$

$$D_{0l}=C_0^\infty(0,\infty)+L\{w_F\},$$

where w_F, as it will result from the following, is a smooth function such that the Friedrichs boundary condition is implemented. This core is contained in $C_0^\infty(\mathscr{M}_E)$ and the given boundary condition is the only one allowing such an inclusion. For analogous discussions in the case of Schrödinger operators in quantum mechanics, see Ref. 14, sec. 17.F.

In fact, one can obtain a core for a general self-adjoint extension of A_{0l} as follows:

$$D_{0l}=C_0^\infty(0,\infty)+L\{\cos(s)w_1+\sin(s)w_2\},$$

where w_1, w_2 are smooth functions such that[14]

$$w_i=\begin{cases} u_i(y), & \text{for } y \text{ close to } 0,\\ 0, & \text{for } y \text{ close to } \infty,\end{cases}$$

and u_1, u_2 are a fundamental system for the equation $(A_{0l}-\lambda)u=0$. In our case a fundamental system has the following leading order behavior near $y=0$:

$$u_1(y)\sim 1, \quad u_2(y)\sim \log(y).$$

We want to see for which value of s the function

$$\varphi_s(y)\equiv\cos(s)w_1(y)+\sin(s)w_2(y)$$

is in $C_0^\infty(\mathscr{M}_E)$. By taking the limit $\lim_{y\to 0}\varphi_s(y)$ it results that the only value of s allowing us to get a finite limit in 0 and so to include D_{0l} in $C_0^\infty(\mathscr{M}_E)$ is $s=0$ and it corresponds to the Friedrichs extension.

We have obtained that the restriction A^F_{core} of A^F to its core is such that $A^F_{core}\subset A$. Note that A^F_{core} is obviously essentially self-adjoint. This implies that $(A^F_{core})^*=A^F\supset A^*=\overline{A}$. But both the operators $(A^F_{core})^*$ and $\overline{A}$ are self-adjoint. The equality $A^F=\overline{A}$ follows.

[1] G. W. Gibbons and S. W. Hawking, ''Action integrals and partition functions in quantum gravity,'' Phys. Rev. D **15**, 2752 (1977).
[2] M. Hatsuda, P. van Nieuwenhuizen, W. Troost, and A. Van Proyen, ''The regularized phase space path integral measure for a scalar field coupled to gravity,'' Nucl. Phys. B **335**, 166 (1990).
[3] D. J. Toms, ''Functional measure for quantum field theory on curved spacetime,'' Phys. Rev. D **35**, 3796 (1987).
[4] J. L. R. Barbon and R. Emparan, ''Quantum black hole entropy and Newton constant renormalization,'' Phys. Rev. D **52**, 4527 (1995).
[5] V. P. Frolov, D. V. Fursaev, and A. I. Zelnikov, ''Black hole entropy: Off shell versus on shell,'' Phys. Rev. D **54**, 2711 (1996).
[6] F. Belgiorno and M. Martellini, ''Hawking radiation entropy and horizon divergences,'' Phys. Rev. D **53**, 7073 (1996).
[7] H. O. Cordes, *Spectral Theory of Linear Differential Operators and Comparison Algebras*, London Mathematical Society Lecture Note Series 76 (Cambridge University Press, Cambridge, 1987).
[8] R. M. Wald, ''On the Euclidean approach to quantum field theory in curved spacetime,'' Commun. Math. Phys. **70**, 221 (1979).
[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II. Fourier Analysis and Self-Adjointness* (Academic, New York, 1975).
[10] W. Mlak, *Hilbert Spaces and Operator Theory* (Kluwer Academic, Dordrecht, 1991).
[11] S. A. Fulling, *Aspects of Quantum Field Theory in Curved Spacetime*, London Mathematical Society Student Texts 17 (Cambridge University Press, Cambridge, 1989); S. A. Fulling and S. N. M. Ruijsenaars, ''Temperature, periodicity and horizons,'' Phys. Rep. **152**, 135 (1987).
[12] D. V. Fursaev and S. N. Solodukhin, ''Description of the Riemannian geometry in presence of conical defects,'' Phys. Rev. D **52**, 2133 (1995).
[13] J. Weidmann, *Linear Operators in Hilbert Spaces* (Springer-Verlag, New York, 1980).

[14] J. Weidmann, *Spectral Theory of Ordinary Differential Operators*, Lecture Notes in Mathematics 1258 (Springer-Verlag, Berlin, 1987).
[15] W. Bulla and F. Gesztesy, ''Deficiency indices and singular boundary conditions in quantum mechanics,'' J. Math. Phys. **26**, 2520 (1985).
[16] N. Dunford and J. T. Schwartz, *Linear Operators. Part II: Spectral Theory* (Interscience, New York, 1963), Chap. XIII.
[17] H. Rumpf and H. K. Urbantke, ''Covariant 'in–out' formalism for creation by external fields,'' Ann. Phys. (N.Y.) **114**, 332 (1978).
[18] F. Rellich, ''Halbbeschränkte gewohnliche Differentialoperatoren zweiter Ordnung,'' Math. Ann. **122**, 343 (1951).
[19] A. L. Besse, *Einstein Manifolds* (Springer-Verlag, Berlin, 1987).
[20] C. Emmerich and H. Römer, ''Orbifolds as configuration spaces of systems with gauge symmetries,'' Commun. Math. Phys. **129**, 69 (1990).
[21] B. S. Kay and U. M. Studer, ''Boundary conditions for quantum mechanics on cones and fields around cosmic string,'' Commun. Math. Phys. **139**, 103 (1991).
[22] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics* (Springer-Verlag, Berlin, 1988).
[23] G. Hellwig, *Differential Operators of Mathematical Physics* (Addison-Wesley, Reading, MA, 1964).
[24] F. Rellich, ''Die zulässigen randbedingungen bei den singulären eigenwertproblemen der matematischen physik,'' Math. Z. **49**, 702 (1943/44).
[25] J. S. Dowker and G. Kennedy, ''Finite temperature and boundary effects in static space-times,'' J. Phys. A **11**, 895 (1978).
[26] R. L. Bishop and B. O'Neill, ''Manifolds of negative curvature,'' Trans. Am. Math. Soc. **145**, 1 (1969).
[27] J. L. Kazdan and F. W. Warner, ''Curvature functions for open 2-manifolds,'' Ann. Math. **99**, 203 (1974).
[28] M. P. do Carmo, *Riemannian Geometry* (Birkhäuser, Boston, 1992).
[29] R. L. Bishop and R. L. Crittenden, *Geometry of Manifolds* (Academic, New York, 1964).

Macroscopic models for semiconductor heterostructures

P. Degond
Mathématiques pour l'Industrie et la Physique, UMR CNRS 5640, URF MIG, Université Paul Sabatier, 118, route de Narbonne 31062, Toulouse Cedex, France

C. Schmeiser
Institut für Angewandte und Numerische Mathematik, TU Wien, Wiedner Hauptstrasse 8-10, 1040 Wien, Austria

(Received 8 October 1997; accepted for publication 14 May 1998)

The semiconductor Boltzmann equation is considered within a material with a space-dependent band characteristic. Its fluid limit under strong elastic collisions leads to the so-called spherical harmonics expansion (SHE) model. For this model, the influence of smoothly space-dependent or abruptly discontinuous band characteristics is analyzed. For the latter case which occurs at heterojunction interfaces, kinetic boundary layers are analyzed and higher order transmission conditions are derived. In a further step, transmission conditions through heterojunctions for the energy-transport and drift-diffusion models (which are obtained as moments of the SHE model) are constructed. © *1998 American Institute of Physics.* [S0022-2488(98)02409-8]

I. INTRODUCTION

The drift-diffusion (DD) model of semiconductors is, by far, the most simple and widely used macroscopic semiconductor model.[1,2] Its derivation from the more fundamental kinetic model, the semiconductor Boltzmann equation, has been made mathematically rigorous in Refs. 3 and 4.

However, the search for more accurate descriptions of carrier transport in semiconductors has been motivated by the recent developments in microelectronics technology. Several models have been proposed to improve the DD model, including the hydrodynamic (HD) model,[5] the energy-transport (ET) model,[6] and the spherical harmonic expansion (SHE) model.[7,8] Among these, the SHE and ET models have been proved to derive from a diffusion approximation of the Boltzmann transport equation,[9,10,7,11] while this question still remains open for the HD model. The diffusion approximation naturally provides consistent closure relations to the SHE and ET models while phenomenological arguments must be invoked to close the HD model. For this reason, in the present paper, we shall concentrate on the SHE and ET models, and on the DD model which stands as a submodel of the ET model.

The available derivations of the SHE and ET models concern materials with homogeneous crystalline structure.[7,10,9] The goal of the present paper is to extend them to materials with smoothly space-dependent band diagrams and to abrupt heterojunctions. Smoothly space-dependent band diagrams appear in materials with graded stoechiometry (like $Al_xGa_xAs_{1-x}$) which are used in certain microelectronics components. Heterojunctions are now widely used in various applications. The DD model has been extended to smoothly space-dependent band diagrams in Ref. 12 and to heterojunctions in Ref. 13.

In Sec. II, we show that the SHE model for smoothly space-dependent band diagrams is structurally identical as for space independent band diagrams, provided that an adequate set of variables and unknowns are used. Similar conclusions are drawn for the ET and DD models in Sec. VI.

For abrupt heterojunctions, the problem is more complicated. We start with the interface problem at the level of the Boltzmann equation. Interface conditions at the band discontinuity are set up in terms of reflection-transmission matrices which take into account carrier scattering by the irregularities of the potential at the interface. Various properties of these interface conditions are reviewed in Sec. III, together with a few practical examples.

In Sec. IV, we proceed to the diffusion approximation of the Boltzmann equation complemented with the interface conditions at the heterojunction. Transmission conditions for the SHE

 © 1998 American Institute of Physics

model are naturally derived from this procedure. As usual, the order of accuracy of the diffusion approximation is measured in terms of the dimensionless mean free path, ratio of the mean free path (associated with elastic collisions) to the typical size of the device. If the first order of accuracy is retained, two natural continuity relations are derived, one relating the values of the SHE-distribution function on each side, the other one relating its fluxes on each side. At the second order of accuracy, the continuity condition on the SHE distribution function is replaced by a condition which links its jump at the interface to the common value of its fluxes on each side of the interface. The coefficient in this combination is similar to the ''extrapolation length'' originating from the presence of kinetic boundary layers[14,15] and introduced in the semiconductor context in Refs. 16 and 17. It is obtained as the equilibrium limit of two half space stationary kinetic problems connected with the interface conditions, which will later be referred to as a ''two-sided Milne problem,'' in reference to the Milne problem of kinetic theory.[14] In Sec. V, it is shown that the two-sided Milne problem is solvable under reasonable assumptions on the solution of one-sided Milne problems with inflow boundary conditions. For special cases these assumptions can be verified. An approximation procedure in the spirit of, e.g., Ref. 15, is carried out leading (for rotationally invariant models) to explicit formulas for the above-mentioned coefficient in the second-order transmission condition.

Transmission conditions at heterojunction interfaces are then derived for the ET and the DD models by a moment expansion of the corresponding transmission conditions for the SHE model (Sec. VI). This procedure supposes that there is no boundary layer in the ET or DD limit of the SHE model, or at least, that the influence of this boundary layer is negligible. This point will be discussed in detail in future work. The obtained first-order transmission condition for the DD model reduces to the classical one.[13] For the DD model, transmission conditions of the form of the second-order conditions have been derived in Ref. 18. They correspond to a very rough approximation of the solution of the two-sided Milne problem.

We believe that the use of these second-order transmission conditions will significantly improve the accuracy of the numerical treatment of heterojunction interfaces. Future work will be devoted to the numerical implementation of these conditions and their validation against experimental and numerical data.

II. THE SPHERICAL HARMONIC EXPANSION (SHE) MODEL

The SHE model is derived from the semiconductor Boltzmann equation through a diffusion approximation procedure using elastic collisions as the dominant scattering mechanism. We refer to Ref. 9 for a discussion of the validity of the elastic approximation of the phonon collision operator. Here, our starting point will be the following Boltzmann equation:

$$\frac{\partial f^\alpha}{\partial t}+\frac{1}{\alpha}\,(\nabla_k\epsilon\cdot\nabla_x f^\alpha-\nabla_x\epsilon\cdot\nabla_k f^\alpha)=\frac{1}{\alpha^2}\,Q(f^\alpha)+S(f^\alpha), \tag{II.1}$$

where $f^\alpha=f^\alpha(x,k,t)$ is the electron distribution function depending on the position $x\in\mathbf{R}^3$, the momentum $k\in B$, and the time $t>0$. The Brillouin zone B is the elementary cell of the dual lattice L^* and is identified with the torus $B=\mathbf{R}^3/L^*$. We suppose that B does not depend on x, which means that the interatomic distances remain the same as the material changes. This hypothesis is reasonable because abrupt heterojunctions or materials with graded stoechiometry are based on materials which share almost the same lattice characteristics.

The function $\epsilon=\epsilon(x,k,t)$ is the total particle energy of an electron in the lattice. It can be decomposed in the following way:

$$\epsilon(x,k,t)=e(x,k)-V(x,t)=e(x,k)-\chi(x)-\phi(x,t), \tag{II.2}$$

with $e(x,0)=0$. The function $e(x,k)$ is the kinetic energy measured from the bottom of the conduction band, while $-V(x,t)$ is the energy of the bottom of the conduction band measured from an arbitrary fixed origin.[19] The potential $V(x,t)$ can itself be decomposed into the Coulomb potential $\phi(x,t)$ arising from the electron space-charge, the doping and the applied bias, which is smooth and may depend upon t, and the electron affinity $\chi(x)$, which only depends on the material and is therefore independent of t. In the present section, we shall first suppose that $\chi(x)$

and $e(x,k)$ are smooth. The space-dependent kinetic energy $e(x,k)$ takes into account the variations of the band diagram as the material smoothly changes. A typical form of $e(x,k)$ is that of a parabolic band diagram with a space-dependent effective mass:

$$e(x,k)=\frac{k^2}{2m_*(x)}. \tag{II.3}$$

However, in the remainder of this section, any smooth C^1 function e: $(x,k)\in\mathbf{R}^3\times B\to\mathbf{R}$ will be admissible. We denote by $v(x,k)=\nabla_k\epsilon(x,k,t)=\nabla_k e(x,k)$ the particle velocity.

The collision operator $Q(f)$ takes into account both impurity scattering and phonon scattering in its elastic limit.[9,7] It is written:

$$Q(f)(k)=\int_B \Phi(x,k,k')\,\delta(\epsilon(k')-\epsilon(k))(f(k')-f(k))dk', \tag{II.4}$$

where the (x,t) dependence of ϵ, f, and $Q(f)$ has been omitted for brevity. The δ function in (II.4) expresses the elastic character of the interaction. $\Phi(x,k,k')$ is the transition matrix associated with the considered interactions. We shall not specify it (see Ref. 20), assuming only its positivity and invariance under exchange of k and k'.

The collision operator $S(f)$ takes into account all the other interactions, these being considered as weaker than the elastic interaction: The inelastic correction to the phonon collisions, the electron–electron collisions, and possibly, other interaction mechanisms. Their exact expression will not be needed here and we refer to Ref. 9 for more details.

Equation (II.1) is already written in scaled form. The scaling is detailed in Ref. 9. Let us just recall that the energy scale is that of the applied bias, and not that of the thermal energy associated with the lattice temperature, and consequently, to some extent, high field effects are included. The parameter α represents the scaled mean free path, i.e., the ratio of the mean free path for elastic interactions to the typical size of the phenomena under consideration. It is supposed small ($\alpha\ll 1$).

The SHE model is obtained as the formal limit of (II.1) when α tends to zero,[9,7,21] by means of the Hilbert expansion:

$$f^\alpha=f^0+\alpha f^1+\alpha^2 f^2+\cdots. \tag{II.5}$$

After inserting the ansatz (II.5) in (II.1), terms of equal powers of α at the left- and right-hand sides of (II.1) are identified. This formal expansion procedure will be carried out in the following, and a rigorous justification is left as the subject for future work. This means in particular that the occurring functions will be assumed sufficiently smooth such that the computations are justified.

Before proceeding, we recall the coarea formula[22]

$$\int_B \psi(k)dk=\int_{-\infty}^{\infty}\int_{\mathcal{S}_{(x,\epsilon,t)}} \psi(k)\,\frac{dS_{(x,\epsilon,t)}(k)}{|\nabla_k\epsilon(x,k,t)|}\,d\epsilon, \tag{II.6}$$

where $\mathcal{S}_{(x,\epsilon,t)}$ is the ((x,t) dependent) manifold

$$\mathcal{S}_{(x,\epsilon,t)}=\{k\in B,\epsilon(x,k,t)=\epsilon\},$$

in k-space and $dS_{(x,\epsilon,t)}(k)$ is the Euclidean surface element on $\mathcal{S}_{(x,\epsilon,t)}$. We denote by $dN_{(x,\epsilon,t)}(k)$ the coarea and by $N(x,\epsilon,t)$ the energy density of states:

$$dN_{(x,\epsilon,t)}(k)=\frac{dS_{(x,\epsilon,t)}(k)}{|\nabla_k\epsilon(x,k,t)|},\quad N(x,\epsilon,t)=\int_{\mathcal{S}_{(x,\epsilon,t)}} dN_{(x,\epsilon,t)}(k). \tag{II.7}$$

The coarea formula (II.6) is valid for every $\psi\in C(\overline{B})$, if

$$N(x,\epsilon,t)<\infty\quad \forall(x,\epsilon,t)\in\mathbf{R}^3\times\mathbf{R}\times\mathbf{R}^+ \tag{II.8}$$

holds. This is an assumption on the set where $\nabla_k\epsilon$ vanishes and on its behavior close to this set.

The collisions modeled by the operator Q do not change the energy of a particle. Therefore, we can consider Q as an operator acting on $L^2(\mathscr{S}_{(x,\epsilon,t)}, dN_{(x,\epsilon,t)})$ for every fixed (x,ϵ,t). From Ref. 9, we have

Lemma II.1: (i) The null space of Q consists of the constants only.
(ii) The equation $Q(f)=g$ is solvable if and only if

$$\int_{\mathscr{S}_{(x,\epsilon,t)}} g(k)\,dN_{(x,\epsilon,t)}(k)=0, \tag{II.9}$$

and its solution f is unique provided that

$$\int_{\mathscr{S}_{(x,\epsilon,t)}} f(k)\,dN_{(x,\epsilon,t)}(k)=0. \tag{II.10}$$

(iii) Q is a nonpositive self-adjoint operator, i.e.,

$$\int_{\mathscr{S}_{(x,\epsilon,t)}} Q(f)g\,dN_{(x,\epsilon,t)}=\int_{\mathscr{S}_{(x,\epsilon,t)}} fQ(g)\,dN_{(x,\epsilon,t)},$$

$$\int_{\mathscr{S}_{(x,\epsilon,t)}} Q(f)f\,dN_{(x,\epsilon,t)}\leqslant 0,$$

with equality implying that f is in the null space of Q.

A more precise functional setting and additional properties of Q can be found in Ref. 9 or Ref. 21. A consequence of point (i) is that if $Q(f)=0$ holds for a function $f(x,k,t)$, then $f(x,k,t)=F(x,\epsilon(x,k,t),t)$ for a function F of position, energy, and time.

The Hilbert procedure leads to the following sequence of equations for f^i, $i=0,1,2$:

$$Q(f^0)=0, \tag{II.11}$$

$$Q(f^1)=\nabla_k\epsilon\cdot\nabla_x f^0-\nabla_x\epsilon\cdot\nabla_k f^0, \tag{II.12}$$

$$Q(f^2)=\nabla_k\epsilon\cdot\nabla_x f^1-\nabla_x\epsilon\cdot\nabla_k f^1+\frac{\partial f^0}{\partial t}-S(f^0). \tag{II.13}$$

From (II.11) and Lemma II.1 (i), it follows that there exists $F=F(x,\epsilon,t)$, such that

$$f^0(x,k,t)=F(x,\epsilon(x,k,t),t). \tag{II.14}$$

The function $F(x,\epsilon,t)$ cannot be specified further by using Eq. (II.11) alone. To derive the equation for F that we are looking for, we must solve the next order equations (II.12) and (II.13). Equation (II.12) reduces to

$$Q(f^1)=\nabla_k\epsilon\cdot\nabla_x F.$$

Since $\nabla_x F$ only depends on the energy, f^1 can be written as

$$f^1=-\lambda(x,k,t)\cdot\nabla_x F+F^1(x,\epsilon(x,k,t),t), \tag{II.15}$$

where F^1 is arbitrary and λ is the solution of

$$-Q(\lambda)=\nabla_k\epsilon, \quad \int_{\mathscr{S}_{(x,\epsilon,t)}} \lambda\,dN_{(x,\epsilon,t)}=0. \tag{II.16}$$

Existence and uniqueness of solutions of (II.16) are guaranteed by Lemma II.1 (ii), since $\int_{\mathscr{S}}\nabla_k\epsilon\,dN=0$. Note that (II.16) must be understood componentwise. Then we have the

Theorem II.2: Let f^0 and f^1 be given by (II.14) and (II.15). Then, a solution f^2 of (II.13) exists if and only if $F(x,\epsilon,t)$ satisfies

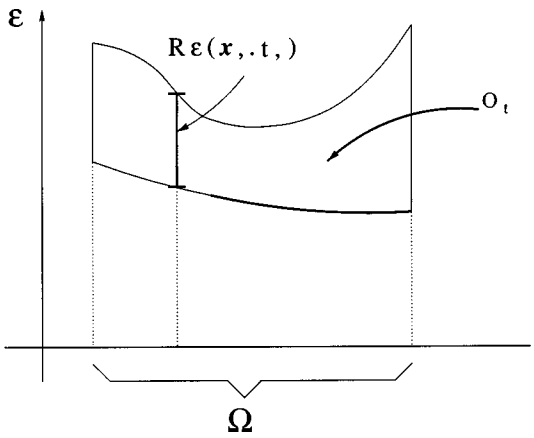

FIG. 1. Domain of definition of F: Case of a smooth band diagram.

$$N(x,\epsilon,t)\left(\frac{\partial F}{\partial t}-\frac{\partial \phi}{\partial t}\frac{\partial F}{\partial \epsilon}\right)-\nabla_x\cdot(D(x,\epsilon,t)\nabla_x F)=\overline{S}(F), \tag{II.17}$$

with the diffusion matrix $D(x,\epsilon,t)$ given by

$$D(x,\epsilon,t)=\int_{\mathscr{S}_{(x,\epsilon,t)}}\nabla_k\epsilon\otimes\lambda\, dN_{(x,\epsilon,t)}, \tag{II.18}$$

where $\otimes$ denotes the tensor product, and

$$\overline{S}(F)=\int_{\mathscr{S}_{(x,\epsilon,t)}}S(F)dN_{(x,\epsilon,t)}.$$

Proof: With (II.15), (II.13) is written as

$$Q(f^2)=\nabla_k\epsilon\cdot\nabla_x F^1-\nabla_x\cdot(\nabla_k\epsilon(\lambda\cdot\nabla_x F))+\nabla_k\cdot(\nabla_x\epsilon(\lambda\cdot\nabla_x F))+\frac{\partial F}{\partial t}+\frac{\partial F}{\partial\epsilon}\frac{\partial\epsilon}{\partial t}-S(F).$$

Then, the solvability condition (II.9), together with the identities [valid for any smooth function $u(x,k,t)$, denoting for short, $\mathscr{S}_\epsilon:=\mathscr{S}_{(x,\epsilon,t)}$, $dN_\epsilon:=dN_{(x,\epsilon,t)}$]:

$$\nabla_x\left(\int_{\mathscr{S}_\epsilon}u\ dN_\epsilon\right)=\int_{\mathscr{S}_\epsilon}\nabla_x u\ dN_\epsilon-\frac{\partial}{\partial\epsilon}\left(\int_{\mathscr{S}_\epsilon}u\nabla_x\epsilon\ dN_\epsilon\right),$$

$$\int_{\mathscr{S}_\epsilon}\nabla_k u\ dN_\epsilon=\frac{\partial}{\partial\epsilon}\int_{\mathscr{S}_\epsilon}\nabla_k\epsilon\ u\ dN_\epsilon,$$

leads to the result. Here, we have used (II.2) to ensure that $\partial\epsilon/\partial t=-\partial\phi/\partial t$ is independent of k.■

Remark II.3: The SHE model (11.17) with an x-dependent band diagram is identical to the standard SHE model expressed in terms of the total energy variable (compare to Ref. 9, Eqs. (III.23), (III.24)). The only difference lies in the position-dependent density of states N and diffusivity D. We also note that, following Ref. 9, the diffusivity is a symmetric non negative matrix, which satisfies

$$D\geqslant\frac{C}{N}\int_{\mathscr{S}_\epsilon}\nabla_k\epsilon\otimes\nabla_k\epsilon\, dN_\epsilon,$$

where $C>0$ is a constant related to the coercivity constant of Q and $(\nabla_k\epsilon\nabla_k\epsilon)$ is the tensor square of $\nabla_k\epsilon$. We shall not detail this point here. ■

Remark II.4: $F(x,\epsilon,t)$ is defined on the (time-dependent) domain (*see Fig. 1*)

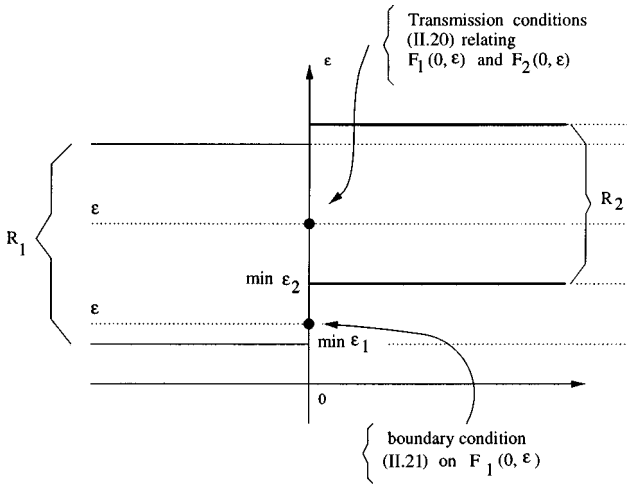

FIG. 2. Domain of definition of F: Case of an abrupt heterojunction at Σ.

$$\mathcal{O}_t=\{(x,\epsilon)\in\Omega\times\mathbf{R}|N(x,\epsilon,t)>0\}=\{(x,\epsilon)\in\Omega\times\mathbf{R}|\,\epsilon\in\mathcal{R}\epsilon(x,\cdot,t)\},\tag{II.19}$$

where Ω is the device domain in x-space and $\mathcal{R}\epsilon(x,\cdot,t)$ is the numerical range of the energy function $k\in B\to\epsilon(x,k,t)$. Therefore, boundary conditions must be prescribed for $x\in\partial\Omega$ or $\epsilon\in\partial\mathcal{R}\epsilon(x,\cdot,t)$. For $x\in\partial\Omega$, we refer to Ref. 16 where this question is addressed in detail. For $\epsilon\in\partial\mathcal{R}\epsilon(x,\cdot,t)$, it is enough to note that the diffusion constant D (see (II.18)) vanishes, because the density of states vanishes. Thus, the parabolic problem (II.17) degenerates (see Ref. 9 for details) and no boundary condition needs to be prescribed. ■

Now, instead of a smoothly x-dependent band diagram, we consider a heterojunction, modeled as an abrupt interface Σ separating two media Ω_1 and Ω_2 which have different band diagram characteristics. Let us denote by $e_i(k)$ and χ_i, $i=1,2$, the kinetic energy and the affinity of each material. For simplicity, we now assume that e_i and χ_i do not vary within Ω_i. Therefore

$$\epsilon(x,k,t)=e_i(k)-\chi_i-\phi(x,t),\quad x\in\Omega_i,\quad k\in B,\quad t>0.$$

For $x\in\Sigma$, let

$$\epsilon_i(x,k,t)=\lim_{y\to x,y\in\Omega_i}\epsilon(y,k,t).$$

The jump discontinuity of the band diagram characteristics across Σ results in similar jump discontinuities of the coefficients $N(x,\epsilon,t)$ and $D(x,\epsilon,t)$ [see formulas (II.7) and (II.18)]. In a weak formulation of (II.17) this does not cause any problem, and the resulting interface conditions are continuity of F and of the normal flux $\nu^T D\nabla_x F$ across Σ:

$$F_1=F_2,\quad \nu^T D_1\nabla_x F_1=\nu^T D_2\nabla_x F_2,\quad x\in\Sigma.\tag{II.20}$$

Here, $\nu(x)$ denotes the normal vector to Σ at a given point x oriented say from 1 to 2, ν^T, the transpose of ν, and

$$\nu^T D\nabla_x F=\sum_{i,j}\nu_i D_{ij}\frac{\partial F}{\partial x_j}.$$

Remark II.5: For $x\in\Sigma$, conditions (II.20) are only valid for the values of ϵ such that $(x,\epsilon)\notin\partial\mathcal{O}_t$, where $\mathcal{O}_t$ is the domain of definition of F (II.19). When $(x,\epsilon)\in\partial\mathcal{O}_t$, for instance, if $\min\epsilon_1(x,\cdot,t)<\epsilon<\min\epsilon_2(x,\cdot,t)$, a boundary condition must be assigned to F_1 instead (see Fig. 2). But since electrons of energy ϵ are totally reflected by the potential barrier, there is no net flux across the heterojunction, and the boundary condition is naturally the zero flux condition:

$$\nu^T D_1 \nabla_x F_1(x,\epsilon,t) = 0. \tag{II.21}$$

At variance with Remark 2.3, D_1 is not necessarily degenerate because ϵ is not an extremal value of $\epsilon_1(x,\cdot,t)$. A similar condition must of course be assigned to F_2 when (x,ϵ) is an end point of the domain of definition of F_2. ■

Transmission conditions (II.20) and (II.21) may be justified by successively performing a diffusion limit of the Boltzmann equation for a smoothed band diagram and then letting the smoothing parameter go to zero. A natural question is whether these two procedures commute, or equivalently, does the diffusion limit of a kinetic model with discontinuous band diagram lead to transmission conditions (II.20), (II.21)? In Sec. III, we set up the heterojunction transmission problem at the kinetic level. Then, in Sec. IV, we show that conditions (II.20) and (II.21) are deduced from a first-order diffusion approximation of this problem. We then derive more accurate conditions from a second-order diffusion approximation.

III. THE KINETIC TRANSMISSION PROBLEM AT A HETEROJUNCTION

We focus on a transmission problem which describes a one-dimensional heterojunction. We let $\Omega_1=(-\infty,0)$, $\Omega_2=(0,\infty)$ and we denote by $f_i^\alpha(x,k,t)$, $x\in\Omega_i$, $k\in B$, $t\geqslant 0$, the distribution function in Ω_i, which solves (with obvious notations)

$$\frac{\partial f_i^\alpha}{\partial t}+\frac{1}{\alpha}\left(v_{i,x}\frac{\partial f_i^\alpha}{\partial x}+\frac{\partial \phi}{\partial x}\frac{\partial f_i^\alpha}{\partial k_x}\right)=\frac{1}{\alpha^2}Q_i(f_i^\alpha)+S_i(f_i^\alpha), \tag{III.1}$$

where $e_i(k)$, $i=1,2$, is the kinetic energy in each domain Ω_i and $v_{i,x}(k)=(\partial e_i/\partial k_x)(k)$ is the associated velocity in the x-direction. We recall that the affinity χ_i is constant over Ω_i and that the total energy is given by

$$\epsilon(x,k,t)=e_i(k)-\chi_i-\phi(x,t), \quad x\in\Omega_i.$$

For simplicity, we set $\phi(0,t)=0$ and we denote by $\epsilon_i(k)$ the right and left limits of ϵ at 0:

$$\epsilon_i(k)=e_i(k)-\chi_i. \tag{III.2}$$

The numerical range of ϵ_i is denoted by $\mathscr{R}_i$.

In the following, we shall assume (which is satisfied in practice[20]) a reflection symmetry of the model. As a consequence, the functions $e_i(k)$, $i=1,2$ are even and the collision operators are even in the sense that for any arbitrary smooth function $f(k)$, we have:

$$Q_i(f(-k))=Q_i(f)(-k), \quad S_i(f(-k))=S_i(f)(-k). \tag{III.3}$$

We now set up boundary conditions for f_i at $x=0$. Particle scattering by the interface is due to the localized potential which sits at the heterojunction because of the change in crystalline structure. It may be described at the kinetic level by scattering operators mutually relating the incoming and outgoing parts (relative to Ω_i) of the two distribution functions f_i, $i=1,2$ at the interface. The spirit is similar to gas–surface interaction operators.[23] However, here we assume the scattering to be of purely elastic nature and the theory of Ref. 23 must be slightly adapted.

We respectively denote by f_i^+ and f_i^- the outgoing and incoming parts of the distribution function f_i at $x=0$:

$$f_i^\pm = f_i|_{\{k\in B,\pm\sigma_i v_{i,x}(k)>0\}} \quad \text{with} \quad \sigma_1=+1 \quad \text{and} \quad \sigma_2=-1. \tag{III.4}$$

The reflection-transmission problem for f_i at $x=0$ is written as

$$f_i^- = \mathscr{T}_i(f_1^+, f_2^+), \quad i=1,2, \tag{III.5}$$

with, for $k\in B$ such that $\sigma_i v_{i,x}(k)<0$:

$$\mathcal{T}_i(f_1^+, f_2^+)(k) = \int_{\{k' \in B, \sigma_i v_{i,x}(k') > 0\}} R_i(k' \to k)\, \delta(\epsilon_i(k) - \epsilon_i(k'))\, f_i(k')\, dk' + \int_{\{k' \in B, \sigma_j v_{j,x}(k') > 0\}} T_{j \to i}(k' \to k)\, \delta(\epsilon_i(k) - \epsilon_j(k'))\, f_j(k')\, dk'. \quad \text{(III.6)}$$

In (III.6), the first integral accounts for the reflection of particles of Ω_i by the interface, while the second one describes their transmission from Ω_j into Ω_i. The delta functions express that reflection as well as transmission occur with conservation of total particle energy. The coefficients $R_i(k' \to k)$ and $T_{i \to j}(k' \to k)$ are non-negative measures derived from the scattering operator of the underlying microscopic dynamics.[24]

For $\epsilon \in \mathbf{R}$, we introduce

$$\mathcal{S}_\epsilon^i = \{k \in B, \epsilon_i(k) = \epsilon\}, \quad dN_\epsilon^i(k) = \frac{dS_\epsilon^i(k)}{|\nabla_k \epsilon_i(k)|}, \quad \text{(III.7)}$$

$dS_\epsilon^i(k)$ being the Euclidean surface element on $\mathcal{S}_\epsilon^i$. Also, we introduce

$$\mathcal{S}_\epsilon^{i,\pm} = \{k \in \mathcal{S}_\epsilon^i, \quad \pm \sigma_i v_{i,x}(k) > 0\}$$

and σ_i given by (III.4). Then, using the coarea formula (II.6), relation (III.6) can be rewritten for $k \in B$ such that $\sigma_i v_{i,x}(k) < 0$:

$$\mathcal{T}_i(f_1^+, f_2^+)(k) = \int_{\mathcal{S}^{i,+}_{\epsilon_i(k)}} R_i(k' \to k) f_i(k')\, dN^i_{\epsilon_i(k)}(k') + \int_{\mathcal{S}^{j,+}_{\epsilon_i(k)}} T_{j \to i}(k' \to k) f_j(k')\, dN^j_{\epsilon_i(k)}(k'), \quad \text{(III.8)}$$

with the convention that $j \neq i$ in (III.8). In particular, we note that, if $\epsilon_i(k) \notin \mathcal{R}_j$ [where we recall that $\mathcal{R}_i$ denotes the numerical range of $\epsilon_i(k)$], the second integral in (III.8) is identically zero (because $\mathcal{S}^j_{\epsilon_i(k)} = \emptyset$). Thus, for such particle energies, only reflection occurs, as expressed by the first integral in (III.8).

We now set up conditions for the reflection-transmission operators $\mathcal{T}_i$ to provide a physically sound description of electron scattering by the interface.

Hypothesis 1: Current conservation.

$$|v_{i,x}(k)| = \int_{\mathcal{S}^{i,-}_{\epsilon_i(k)}} |v_{i,x}(k')| R_i(k \to k') dN^i_{\epsilon_i(k)}(k') + \int_{\mathcal{S}^{j,-}_{\epsilon_i(k)}} |v_{j,x}(k')| T_{i \to j}(k \to k') dN^j_{\epsilon_i(k)}(k'), \quad \text{(III.9)}$$

holds for all $k \in B$ such that $\sigma_i v_{i,x}(k) > 0$, with again $j \neq i$ in (III.9).

With property (III.9), the flux of particles of given energy ϵ is continuous through the interface, as asserted by the following Lemma, the proof of which is a simple computation:

Lemma III.1: Let $f_i(k)$, $i = 1,2$, be connected through relation (III.5) with R_i and $T_{i \to j}$ satisfying Hypothesis 1. Then, we have,

$$\int_{\mathcal{S}^1_\epsilon} v_{1,x} f_1 dN^1_\epsilon = \int_{\mathcal{S}^2_\epsilon} v_{2,x} f_2 dN^2_\epsilon, \quad \forall \epsilon \in \mathcal{R}_1 \cap \mathcal{R}_2, \quad \text{(III.10)}$$

$$\int_{\mathcal{S}^i_\epsilon} v_{i,x} f_i dN^i_\epsilon = 0, \quad \forall \epsilon \in \mathcal{R}_i \backslash \mathcal{R}_j. \quad \text{(III.11)}$$

Conversely, for (III.10) and (III.11) to be satisfied for any pair (f_1, f_2) connected by (III.5), R_i and $T_{i\to j}$ must satisfy (III.9).

Hypothesis 2: Reciprocity: For all k, k' in B such that $\epsilon_i(k)=\epsilon_i(k')$ and $\sigma_i v_{i,x}(k)>0$, $\sigma_i v_{i,x}(k')<0$, we assume

$$|v_{i,x}(k')|R_i(k\to k')=|v_{i,x}(-k)|R_i(-k'\to -k). \tag{III.12}$$

Similarly, for all k, k' in B such that $\epsilon_i(k)=\epsilon_j(k')$ and $\sigma_i v_{i,x}(k)>0, \sigma_j v_{j,x}(k')<0$, we assume

$$|v_{j,x}(k')|T_{i\to j}(k\to k')=|v_{i,x}(-k)|T_{j\to i}(-k'\to -k), \tag{III.13}$$

where σ_i is given by (III.4).

Relations (III.12) and (III.13) result from the time reversibility (and thus, the unitarity of the scattering matrix) of the microscopic particle dynamics at the interface (see, e.g., Ref. 24). The validity of the reciprocity relation is discussed in Ref. 23 for gas–surface interaction.

Combining (III.9) with (III.12) and (III.13), and using the evenness of $e_i(k)$, we obtain:

$$1=\int_{\mathscr{S}^{i,+}_{\epsilon_i(k)}} R_i(k'\to k)\,dN^i_{\epsilon_i(k)}(k')+\int_{\mathscr{S}^{j,+}_{\epsilon_i(k)}} T_{j\to i}(k'\to k)\,dN^j_{\epsilon_i(k)}(k'), \tag{III.14}$$

for all k such that $\sigma_i v_{i,x}(k)<0$. It follows that the constant function equal to 1 is a particular solution of the interface condition (III.5).

Hypothesis 3: Transmittivity. Some transmission occurs whenever it is energetically possible:

$$\overline{T}_{j\to i}(k,\epsilon):=\int_{\mathscr{S}^{j,+}_{\epsilon}} T_{j\to i}(k'\to k)\,dN^j_{\epsilon}(k')>0 \tag{III.15}$$

for every $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$ and for k in a subset of $\mathscr{S}^{i,-}_{\epsilon}$ with positive measure.

Note that by the reciprocity relation (III.13), (III.15) holds for both $T_{1\to 2}$ and $T_{2\to 1}$, even if it is only required for one of them. As mentioned above, only reflection occurs for particle energies not in the common range of ϵ_1 and ϵ_2:

$$\overline{T}_{j\to i}(\cdot,\epsilon)=0, \quad \epsilon\in\mathscr{R}_i\backslash\mathscr{R}_j. \tag{III.16}$$

Lemma III.2: Let R_i and $T_{i\to j}$ satisfy Hypotheses 1–3. Assume, in addition, that the functions $e_i(k)$ are even. Then, two functions $F_1(\epsilon_1(k))$ and $F_2(\epsilon_2(k))$ of the energy satisfy the transmission condition (III.5) if and only if

$$F_1(\epsilon)=F_2(\epsilon) \quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2. \tag{III.17}$$

Proof: Using (III.14), the transmission conditions for F_1 and F_2 can be written as

$$\overline{T}_{j\to i}(k,\epsilon)(F_1(\epsilon)-F_2(\epsilon))=0. \tag{III.18}$$

The result follows immediately from (III.15) and (III.16). ■

With (III.14), relation (III.8) expresses that f_i^- is a convex combination of f_1^+ and f_2^+. We deduce the following version of the Darrozes–Guiraud inequality[23] for our particular transmission process:

Lemma III.3: Let f_i satisfy the transmission condition (III.5), with reflection-transmission matrices satisfying Hypotheses 1 and 2 and with even functions $e_i(k)$. Then, we have [with σ_i given by (III.4)]:

$$\int_{\mathscr{S}^2_{\epsilon}} v_{2,x}|f_2|^2\,dN^2_{\epsilon}\leqslant\int_{\mathscr{S}^1_{\epsilon}} v_{1,x}|f_1|^2\,dN^1_{\epsilon}, \quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{III.19}$$

$$\sigma_i\int_{\mathscr{S}^i_{\epsilon}} v_{i,x}|f_i|^2\,dN^i_{\epsilon}\geqslant 0, \quad \forall\,\epsilon\in\mathscr{R}_i\backslash\mathscr{R}_j. \tag{III.20}$$

Proof: Using Jensen's inequality on (III.8), assuming $\epsilon \in \mathscr{R}_1 \cap \mathscr{R}_2$, and $\epsilon_i(k) = \epsilon$, we get

$$|f_i^-(k)|^2 \leqslant \int_{\mathscr{S}_\epsilon^{i,+}} R_i(k' \to k) |f_i(k')|^2 dN_\epsilon^i(k') + \int_{\mathscr{S}_\epsilon^{j,+}} T_{j\to i}(k' \to k) |f_j(k')|^2 dN_\epsilon^j(k'). \tag{III.21}$$

Then, multiplying (III.21) by $v_{i,x}(k)$, integrating over $\mathscr{S}_\epsilon^{i,-}$, summing up the contributions of f_1 and f_2 and using (III.9) we obtain (III.19). The inequality (III.20) can be proven in a similar way. ■

We now give two examples of transmission operators $\mathscr{T}_i$ which satisfy the above Hypotheses 1–3.

Example 1: Semiclassical transmission of particles between media with spherically symmetric band diagrams across a perfectly plane interface. From Hamiltonian mechanics, we know that the transmission of a particle across a perfectly plane interface occurs with conservation of parallel momentum and total energy. We assume that transmission of a particle with momentum k from side i to side j produces a particle with momentum k'. Denoting by y and z the coordinates parallel to the heterojunction, the transmission process can be written as:

$$k_y' = k_y, \quad k_z' = k_z, \quad \epsilon_j(k') = \epsilon_i(k) \tag{III.22}$$

with ϵ_i given by (III.2). We suppose that e_i is spherically symmetric: $e_i(k) = \overline{e_i}(|k|)$, with a strictly monotone $\overline{e_i}$ [parabolic band diagrams (II.3) being obviously covered by this assumption]. Relation $\epsilon = \overline{e_i}(|k|) - \chi_i$ can be inverted and we denote by $|k| = \kappa_i(\epsilon)$ the inverse relation. From the first two relations in (III.22), we deduce that

$$|k'|^2 - k_x'^2 = |k|^2 - k_x^2,$$

and, with the last relation (III.22) and the definition of κ_i, we obtain that the transverse momentum is transmitted according to

$$k_x' = \sigma_i \psi_i(k)^{1/2}, \quad \psi_i(k) = k_x^2 - |k|^2 + \kappa_j(\epsilon_i(k))^2, \tag{III.23}$$

if $\psi_i(k) > 0$, the particle being reflected (with $k_x' = -k_x$) otherwise.

Now, we define the reflection-transmission operator $\mathscr{T}_i$ according to:

$$\mathscr{T}_i(f_1^+, f_2^+)(k) = (1 - \overline{T}_{j\to i}(k)) f_i^+(rk) + \overline{T}_{j\to i}(k) f_j^+(t_{j\to i}k), \tag{III.24}$$

with $i = 1,2$, $j \neq i$, $k \in B$, such that $\sigma_i k_x \leqslant 0$ and with $\overline{T}_{j\to i}(k)$ being the transmission probability of the semiclassical transmission process, satisfying:

$$0 \leqslant \overline{T}_{j\to i}(k) \leqslant 1, \quad \overline{T}_{j\to i}(k) = 0 \quad \text{if } \psi_i(k) < 0. \tag{III.25}$$

The case of a purely classical transmission is covered with $\overline{T}_{j\to i}(k) = 1$ when $\psi_i(k) > 0$ and $\overline{T}_{j\to i}(k) = 0$ otherwise.

In (III.24), r and $t_{j\to i}$ are particle reflection and transmission operators defined by inverting (III.23):

$$t_{j\to i}k = (-\sigma_i \psi_i(k)^{1/2}, k_y, k_z) \quad \text{if } \psi_i(k) > 0, \tag{III.26}$$

$$rk = (-k_x, k_y, k_z), \tag{III.27}$$

for $k \in B$, such that $\sigma_i k_x \leqslant 0$.

The reflection and transmission matrices associated with (III.24) can be determined from:

$$T_{j\to i}(k' \to k)\, \delta(\epsilon_i(k) - \epsilon_j(k')) = \overline{T}_{j\to i}(k)\, \delta(k' - t_{j\to i}k), \tag{III.28}$$

$$R_i(k' \to k)\, \delta(\epsilon_i(k) - \epsilon_i(k')) = (1 - \overline{T}_{j\to i}(k))\, \delta(k' - rk). \tag{III.29}$$

We easily deduce that Hypotheses 1 and 2 hold if and only if $\overline{T}_{j\to i}$ satisfy the following properties:

$$\overline{T}_{j\to i}(k_x,-k_y,-k_z)=\overline{T}_{j\to i}(k_x,k_y,k_z),\quad \overline{T}_{j\to i}(k)=\overline{T}_{i\to j}(rt_{j\to i}k). \tag{III.30}$$

The last property is the usual quantum-mechanical reciprocity relation.[25] A simple computation using (III.28) shows that the notation $\overline{T}_{j\to i}$ for the transmission probabilities is compatible with its definition in (III.15). Thus, the assumption

$$\overline{T}_{j\to i}(k)>0 \quad \text{for } \psi_i(k)>0,$$

is sufficient for Hypothesis 3 to be satisfied.

Semiclassical transmission probabilities $\overline{T}_{j\to i}$ can be deduced from a scattering analysis of a one-dimensional Schrödinger model of the heterojunction. A simple example is the formula[18]

$$\overline{T}_{j\to i}(k)=\frac{4k_x(t_{j\to i}k)_x m_i m_j}{(k_x m_j+(t_{j\to i}k)_x m_i)^2}$$

for the case of a step barrier and parabolic bands $e_i(k)=k^2/(2m_i)$ on both sides. ■

Example 2: Purely isotropic reflection-transmission operators: Here we assume that the reflection-transmission coefficients are such that the incoming distributions f_i^- only depend on $\epsilon_i(k)$. This implies that $R_i(k\to k')$ and $T_{i\to j}(k\to k')$ only depend on k. Then, Hypothesis 2 implies that R_i and $T_{i\to j}$ are of the form

$$R_i(k\to k')=\hat{R}_i(\epsilon_i(k))|v_{i,x}(k)|,\quad T_{i\to j}(k\to k')=\hat{T}(\epsilon_i(k))|v_{i,x}(k)|.$$

Moreover, by Hypothesis 1, we deduce that $\hat{R}_i$ and $\hat{T}$ must satisfy the relations

$$1=\alpha_i(\epsilon)\hat{R}_i(\epsilon)+\alpha_j(\epsilon)\hat{T}(\epsilon),\quad i\neq j,$$

for $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$ with

$$\alpha_i(\epsilon)=\int_{\mathscr{S}_\epsilon^{i,-}}|v_{i,x}|dN_\epsilon^i. \tag{III.31}$$

Therefore, R_i and $T_{i\to j}$ are given in terms of the single arbitrary function $\hat{T}(\epsilon)$:

$$R_i(k\to k')=|v_{i,x}(k)|\,\frac{1-\alpha_j\hat{T}}{\alpha_i}\,(\epsilon_i(k)),\quad T_{i\to j}(k\to k')=|v_{i,x}(k)|\hat{T}(\epsilon_i(k)), \tag{III.32}$$

with $0\leqslant\hat{T}(\epsilon)\leqslant\min(1/\alpha_1(\epsilon),1/\alpha_2(\epsilon))$ for $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$ and $\hat{T}(\epsilon)=0$ otherwise. By $\overline{T}_{i\to j}=\alpha_i\hat{T}$, the assumption $\hat{T}(\epsilon)>0$ for $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$ is sufficient for Hypothesis 3 to be satisfied. ■

IV. DIFFUSION APPROXIMATION OF THE HETEROJUNCTION MODEL

In this section we assume that the transmission operator is given by (III.6) and satisfies Hypotheses 1–3. We start with the definition of an order k approximate solution:

Definition IV.1: f_i^α is an order k approximate solution of the Boltzmann transmission problem (III.1) and (III.5) if

(i) f_i^α solves the Boltzmann equation (III.1) up to order α^k terms in each domain Ω_i:

$$\frac{\partial f_i^\alpha}{\partial t}+\frac{1}{\alpha}\left(v_{i,x}\frac{\partial f_i^\alpha}{\partial x}+\frac{\partial\phi}{\partial x}\frac{\partial f_i^\alpha}{\partial k_x}\right)-\frac{1}{\alpha^2}\,Q_i(f_i^\alpha)-S_i(f_i^\alpha)=\mathscr{O}(\alpha^k), \tag{IV.1}$$

(ii) f_i^α satisfies the transmission condition (III.5) up to order α^k terms:

$$f_i^\alpha(0,\cdot,t)^-=\mathscr{T}_i(f_1^\alpha(0,\cdot,t)^+,\ f_2^\alpha(0,\cdot,t)^+)+\mathscr{O}(\alpha^k). \tag{IV.2}$$

(iii) The flux of particles of given energy ϵ given by

$$J_i^\alpha(x,\epsilon,t)=\frac{1}{\alpha}\int_{\mathscr{S}_{(x,\epsilon,t)}} v_{i,x}(k) f_i^\alpha(x,k,t)\,dN_{(x,\epsilon,t)}(k),$$

is continuous through the interface up to order α^k terms, i.e.,

$$J_1^\alpha(0,\epsilon,t)-J_2^\alpha(0,\epsilon,t)=\mathscr{O}(\alpha^k),\quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{IV.3}$$

$$J_i^\alpha(0,\epsilon,t)=\mathscr{O}(\alpha^k),\quad \forall\,\epsilon\in\mathscr{R}_i\backslash\mathscr{R}_j. \tag{IV.4}$$

Condition (iii) needs some explanation: On first glance it seems superfluous, since flux conservation is already included in the transmission conditions. However, the analysis below shows that (iii) is needed both to completely specify an order k approximate solution and to enable the construction of an order $k+1$ approximation.

In this section, we are concerned with the derivation of interface conditions for the SHE distribution function $F(x,\epsilon,t)$ across the heterojunction at $x=0$. We shall successively derive first- and second-order transmission conditions, by respectively constructing order 1 and order 2 approximate solutions of the Boltzmann transmission problem (III.1), (III.5). We start with the first-order transmission conditions. Following Sec. II, the Hilbert expansion (II.5) of f_i^α in each domain Ω_i leads to an approximate solution of the form

$$f_i^\alpha(x,k,t)=F_i^0(x,\epsilon(x,k,t),t)+\alpha\left(F_i^1(x,\epsilon(x,k,t),t)-\lambda_{i,x}(x,k,t)\,\frac{\partial F_i^0}{\partial x}\,(x,\epsilon(x,k,t),t)\right)$$
$$+\alpha^2 f_i^2(x,k,t), \tag{IV.5}$$

where F_i^0 is a solution of the SHE equation in Ω_i:

$$N_i(x,\epsilon,t)\left(\frac{\partial F_i^0}{\partial t}-\frac{\partial\phi}{\partial t}\,\frac{\partial F_i^0}{\partial\epsilon}\right)-\frac{\partial}{\partial x}\left(D_i(x,\epsilon,t)\,\frac{\partial F_i^0}{\partial x}\right)=\overline{S}_i(F_i^0). \tag{IV.6}$$

Here, $N_i(x,\epsilon,t)$ is the density of states (II.7) in each domain Ω_i, and $D_i(x,\epsilon,t)$ is the scalar which corresponds to the $x-x$ entry of the diffusion matrix (II.18) in Ω_i:

$$D_i(x,\epsilon,t)=\int_{\mathscr{S}_{(x,\epsilon,t)}}\lambda_{i,x}(x,k,t)v_{i,x}(k)\,dN_{(x,\epsilon,t)}(k). \tag{IV.7}$$

The following theorem states that the first-order interface conditions coincide with those established in Sec. II.

Theorem IV.2: Let f_i^α be defined by (IV.5) with F_i^0 solving (IV.6). Then, f_i^α is an order 1 approximate solution of the Boltzmann transmission problem (III.1), (III.5) if and only if F_i^0 verifies the following continuity relations:

$$F_1^0(0,\epsilon,t)=F_2^0(0,\epsilon,t),\quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{IV.8}$$

$$D_1\,\frac{\partial F_1^0}{\partial x}\,(0,\epsilon,t)=D_2\,\frac{\partial F_2^0}{\partial x}\,(0,\epsilon,t),\quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{IV.9}$$

$$D_i\,\frac{\partial F_i^0}{\partial x}\,(0,\epsilon,t)=0,\quad \forall\,\epsilon\in\mathscr{R}_i\backslash\mathscr{R}_j. \tag{IV.10}$$

Proof: By construction, f_i^α is an order 1 approximate solution of the Boltzmann equation (III.1) inside each Ω_i. Now, inserting ansatz (IV.5) into the transmission condition (III.5), we get

$$(f_i^\alpha)^- - \mathscr{T}_i((f_1^\alpha)^+,(f_2^\alpha)^+)=[F_i^{0-}-\mathscr{T}_i(F_1^{0+},F_2^{0+})]+\mathscr{O}(\alpha).$$

By Lemma (III.2), imposing (IV.8) is necessary and sufficient to guarantee that

$$F_i^{0-} - \mathcal{T}_i(F_1^{0+}, F_2^{0+}) = 0,$$

and thus to ensure (IV.2) with $k=1$. Finally, the flux of particles of energy ϵ associated with (IV.5) is given at $x=0$ by

$$J_i^\alpha(0,\epsilon,t) = -\left(D_i \frac{\partial F_i^0}{\partial x} \right)(0,\epsilon,t) + \mathcal{O}(\alpha).$$

Therefore condition (IV.3) requires (IV.9) while (IV.4) requires (IV.10). ■

We now proceed to the derivation of second-order transmission conditions. The forthcoming procedure is inspired from the now classical method of Ref. 14, already applied to semiconductors in Refs. 16 and 26, to derive higher order boundary conditions and kinetic-fluid coupling algorithms. We start again with the Hilbert expansion (away from the heterojunction), like in (II.5), but retaining an additional term:

$$f_i^\alpha = F_i^0 + \alpha\left(F_i^1 - \lambda_{i,x} \frac{\partial F_i^0}{\partial x} \right) + \alpha^2 f_i^2 + \alpha^3 f_i^3. \tag{IV.11}$$

Inserting ansatz (IV.11) in (IV.1) and identifying terms of equal powers of α yields the equations (II.13) for the determination of f_i^2 and

$$Q_i(f_i^3) = \frac{\partial \epsilon}{\partial k_x}\frac{\partial f_i^2}{\partial x} - \frac{\partial \epsilon}{\partial x}\frac{\partial f_i^2}{\partial k_x} + \frac{\partial f_i^1}{\partial t} - S_i(f_i^1). \tag{IV.12}$$

Here, we suppose that S_i is linear (a nonlinear S_i would just result in slightly more complicated calculations). The first step of the procedure is to look for conditions which ensure the solvability of system (II.11)–(II.13), (IV.12), and thus, the existence of the terms involved in expansion (IV.11). This guarantees that ansatz (IV.11) satisfies (IV.1) with $k=2$.

Recall that the solution of (II.11) and (II.12) is given by:

$$f_i^0(x,k,t) = F_i^0(x,\epsilon_i(k),t), \tag{IV.13}$$

$$f_i^1(x,k,t) = -\lambda_{i,x}(k) \frac{\partial F_i^0}{\partial x}(x,\epsilon_i(k),t) + F_i^1(x,\epsilon_i(k),t). \tag{IV.14}$$

Proposition IV.3: Let f_i^0 and f_i^1 be respectively given by (IV.13) and (IV.14). Then, there exist f_i^2 and f_i^3, respectively solutions of (II.13), (IV.12), if and only if F_i^0 and F_i^1 both satisfy the SHE equation (IV.6).

Proof: That the SHE-equation for F_i^0 is necessary for the existence of f_i^2, has already been verified in Sec. II.

To find the solvability condition for (IV.12), we need an explicit expression of f_i^2. For that purpose, let h_i^2 be the unique solution of (II.13) in which F_i^1 is set to zero, and satisfying (II.10). With hypothesis (III.3) and the uniqueness statement of Lemma II.1 (ii), we deduce that h_i^2 is an even function with respect to k. Then, the most general solution f_i^2 of (II.13) is

$$f_i^2 = h_i^2 - \lambda_{i,x}\frac{\partial F_i^1}{\partial x} + F_i^2, \tag{IV.15}$$

where $F_i^2 = F_i^2(x,\epsilon,t)$ is arbitrary. Inserting (IV.15) in (IV.12) yields

$$\begin{aligned} Q_i^0(f_i^3) = &\left(\frac{\partial \epsilon}{\partial k_x}\frac{\partial}{\partial x} - \frac{\partial \epsilon}{\partial x}\frac{\partial}{\partial k_x} \right)\left(-\lambda_{i,x}\frac{\partial (F_i^1)}{\partial x} \right) + \frac{\partial (F_i^1)}{\partial t} - S_i(F_i^1) \\ &+ \left(\frac{\partial \epsilon}{\partial k_x}\frac{\partial}{\partial x} - \frac{\partial \epsilon}{\partial x}\frac{\partial}{\partial k_x} \right)(h_i^2 + (F_i^2)) + \frac{\partial}{\partial t}\left(-\lambda_{ix}\frac{\partial (F_i^0)}{\partial x} \right) - S_i\left(-\lambda_{i,x}\frac{\partial (F_i^0)}{\partial x} \right). \end{aligned}$$

Here the notation (F_i^m), $m=0,1$ is a shorthand for $F_i^m(x,\epsilon_i(x,k,t),t)$.

The last line consists of odd terms in k which vanish after integration over $\mathscr{S}_{(x,\epsilon,t)}$. Therefore, the solvability condition (II.9) reduces to the integration of the first line, which obviously yields the SHE equation (IV.6) for F_i^1. ■

With Proposition IV.3, ansatz (IV.11) satisfies condition (IV.1) with $k=2$. We must now find transmission conditions on F_i^0 and F_i^1 which guarantee that (IV.11) is also of order 2 at the interface, i.e., satisfies (IV.2), (IV.3), (IV.4) with $k=2$. Conditions (IV.8)–(IV.10) must be imposed on F_i^0 because they are necessary to guarantee the order 1 approximation (Theorem IV.2). However, whatever transmission conditions are prescribed to F_i^1, it is impossible to make ansatz (IV.11) an order 2 approximate solution at the interface. Indeed, if this were the case, the first-order term f_i^1 given by (IV.14) would have to satisfy the transmission condition (III.5). But f_i^1 involves $\lambda_{i,x}(k)$, which is not a function of the energy only and which cannot be canceled at the interface by just a suitable choice of F_i^1. Therefore, we need to modify ansatz (IV.11) by introducing a transmission layer term, which is exponentially small inside each domain Ω_i, but which cancels the first-order term f_i^1 at the interface. The definition of this transmission layer term is the second step of the procedure.

To derive the transmission layer term, we start from the kinetic problem (III.1) and we stretch the position variable in the direction normal to the boundary. Letting

$$E(x,t)=\frac{\partial \phi}{\partial x}(x,t), \quad \theta_i^\alpha(\xi,k,t)=f_i^\alpha(\alpha\xi,k,t),$$

we get:

$$\frac{\partial \theta_i^\alpha}{\partial t}+\frac{1}{\alpha^2}\,v_{i,x}\,\frac{\partial \theta_i^\alpha}{\partial \xi}-\frac{1}{\alpha}\,E(\alpha x,t)\,\frac{\partial \theta_i^\alpha}{\partial k_x}=\frac{1}{\alpha^2}\,Q_i(\theta_i^\alpha)+S_i(\theta_i^\alpha). \tag{IV.16}$$

We suppose that $E(\alpha x,t)=\mathscr{O}(1)$ as $\alpha\to 0$, which implies that there exists no sharp transition layer in the electrostatic potential. Keeping only the leading order terms of (IV.16), we obtain the following kinetic layer equation with unknown $\theta_i(\xi,k)$, $\sigma_i\xi<0$, $k\in B$, $i=1,2$:

$$v_{i,x}(k)\,\frac{\partial \theta_i}{\partial \xi}=Q_i(\theta_i). \tag{IV.17}$$

Since the transmission layer term θ_i must cancel the first-order term f_i^1 at the interface, we need to consider inhomogeneous transmission conditions at the interface. Namely, given two functions $\ell_i(k)$, $i=1,2$, for $k\in B$, we impose

$$(\theta_i(0,\cdot)-\ell_i)^-=\mathscr{T}_i((\theta_1(0,\cdot)-\ell_1)^+,(\theta_2(0,\cdot)-\ell_2)^+). \tag{IV.18}$$

Layer terms need to have limits as $|\xi|\to\infty$. At least, we look for bounded solutions of problem (IV.17), (IV.18). Integrating (IV.17) over $\mathscr{S}_\epsilon^i$ [see (III.7) for the definition of $\mathscr{S}_\epsilon^i$], we get

$$\frac{\partial}{\partial \xi}J_i(\xi,\epsilon)=0, \quad J_i(\xi,\epsilon)=\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,\theta_i(\xi,k)\,dN_\epsilon^i(k).$$

Therefore, $J_i(\xi,\epsilon)=J_i(\epsilon)$ is constant with respect to ξ. Then, we multiply (IV.17) by $\lambda_{i,x}(k)$, integrate over $\mathscr{S}_\epsilon^i$, use the self-adjointness of Q_i [property (iii) of Lemma II.1] and get:

$$\frac{\partial}{\partial \xi}K_i(\xi,\epsilon)+J_i(\epsilon)=0, \quad K_i(\xi,\epsilon)=\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\lambda_{i,x}(k)\,\theta_i(\xi,k)\,dN_\epsilon^i(k), \tag{IV.19}$$

or

$$K_i(\xi,\epsilon)=-J_i(\epsilon)\xi+K_i(0,\epsilon). \tag{IV.20}$$

Bounded solutions θ_i (leading to bounded K_i) therefore satisfy the zero flux condition:

$$J_i(\epsilon)=\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,\theta_i(\xi,k)\,dN_\epsilon^i(k)=0,\quad \epsilon\in\mathscr{R}_i. \tag{IV.21}$$

The transmission condition (IV.18) and Lemma III.1 imply that the flux produced by $\theta_1(0,\cdot)-\ell_1$ is equal to the flux produced by $\theta_2(0,\cdot)-\ell_2$. Since by (IV.21) the fluxes produced by the $\theta_i(0,\cdot)$ vanish, a necessary condition for the existence of a bounded solution of (IV.17), (IV.18) is given by the flux continuity condition

$$\int_{\mathscr{S}_\epsilon^1} v_{1,x}\ell_1\,dN_\epsilon^1=\int_{\mathscr{S}_\epsilon^2} v_{2,x}\ell_2\,dN_\epsilon^2,\quad \forall\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{IV.22}$$

$$\int_{\mathscr{S}_\epsilon^i} v_{i,x}\ell_i\,dN_\epsilon^i=0,\quad \forall\epsilon\in\mathscr{R}_i\backslash\mathscr{R}_j. \tag{IV.23}$$

For the homogeneous version of (IV.17) and (IV.18) ($\ell_1=\ell_2=0$) an easy computation shows that bounded solutions are given by

$$\theta_i(\xi,k)=\Theta(\epsilon_i(k)),\quad i=1,2, \tag{IV.24}$$

where Θ is an arbitrary bounded function of ϵ.

Problem (IV.17), (IV.18) belongs to the class of Milne problems, which are widely used in many areas of kinetic theory.[14,27,16,15,17] In the present paper is the first occurrence (to our knowledge) of a ''two-sided'' Milne problem. In Sec. V an existence theorem will be proved under strong assumptions on the solutions of the corresponding one-sided Milne problems for $\xi>0$ and $\xi<0$ with inflow boundary conditions. For rotationally symmetric band diagrams $e_i(k)$ and collision operators Q_i these assumptions can be verified (see Sec. V). For more general band diagrams, they are conjectures which will be dealt with in future work.

In the rest of this section, we shall assume the results of the existence theorem: Under conditions (IV.22), (IV.23), problem (IV.17), (IV.18) has a bounded solution, which is unique up to the addition of a function of the type (IV.24) with arbitrary Θ. For $|\xi|\to\infty$, $\theta_i(\xi,k)$ converges exponentially to an element $\theta_i^\infty(\epsilon_i(k))$ of the kernel of Q_i.

Finally, a simple computation shows that when θ_i is a bounded solution corresponding to the inhomogeneities ℓ_i, then $A(\epsilon_i(k))\,\theta_i(\xi,k)$ is a bounded solution corresponding to the inhomogeneities $A(\epsilon_i(k))\ell_i(k)$ for arbitrary functions $A(\epsilon)$.

Once the transmission layer problem is set up, the third step of the procedure is to find a modified ansatz (IV.11) which provides a true order 2 approximate solution of the Boltzmann transmission problems (III.1) and (III.5). We introduce the following modified Hilbert expansion:

$$f_i^\alpha=F_i^0(x,\epsilon,t)+\alpha\Bigg[-\lambda_{i,x}(k)\,\frac{\partial F_i^0}{\partial x}(x,\epsilon,t)+F_i^1(x,\epsilon,t)+\left(\theta_i\left(\frac{x}{\alpha},k\right)-\theta_i^\infty(\epsilon_i(k))\right)J_{\mathrm{int}}^0(\epsilon_i(k),t)\Bigg]+\alpha^2 f_i^2+\alpha^3 f_i^3. \tag{IV.25}$$

We recall that F_i^0 and F_i^1 satisfy the SHE-equation and that F_i^0 satisfies the first-order transmission conditions (IV.8), (IV.9), and (IV.10). We denote by $J_{\mathrm{int}}^0(\epsilon,t)$, the common value in expression (IV.9), i.e.,

$$J_{\mathrm{int}}^0(\epsilon,t)=\left(D_i\,\frac{\partial F_i^0}{\partial x}\right)(0,\epsilon,t),\quad \epsilon\in\mathscr{R}_1\cap\mathscr{R}_2. \tag{IV.26}$$

J_{int}^0 is the (leading order) flux of particles of energy ϵ at the interface. The transmission layer corrector term θ_i is a solution of the two-sided Milne problem (IV.17), (IV.18) with $\ell_i(k)$ given by

$$\ell_i(k)=D_i^{-1}(0,\epsilon_i(k))\lambda_{i,x}(k),$$

if $\epsilon_i(k)$ belongs to $\mathscr{R}_1 \cap \mathscr{R}_2$ and $\ell_i(k) = 0$ otherwise. From the definition (IV.7) of D_i, we have for $\epsilon \in \mathscr{R}_1 \cap \mathscr{R}_2$:

$$\int_{\mathscr{S}^i_\epsilon} v_{i,x} \ell_i dN^i_\epsilon = D_i(0,\epsilon,t)^{-1} \int_{\mathscr{S}^i_\epsilon} v_{i,x}(k) \lambda_{i,x}(k) dN^i_\epsilon(k) = 1,$$

so that the flux continuity conditions (IV.22), (IV.23) are satisfied. With our assumptions on the two-sided Milne problem, the layer term θ_i exists and the difference

$$\theta(\epsilon) = \theta_2^\infty(\epsilon) - \theta_1^\infty(\epsilon) \tag{IV.27}$$

of its asymptotic values when $|\xi| \to \infty$ is determined uniquely. The next theorem provides interface conditions on F_i^1, necessary for an ansatz of the form (IV.25) to be an order 2 approximate solution of the Boltzmann transmission problem.

Theorem IV.4: Let $F_i^0(x,\epsilon,t)$ and $F_i^1(x,\epsilon,t)$ satisfy the SHE equation (IV.6), and let F_i^0 satisfy the transmission conditions (IV.8), (IV.9), (IV.10). Then, for f_i^α given by ansatz (IV.25) to be an order 2 approximate solution of the Boltzmann transmission problem (III.1), (III.5), it is necessary that F_i^1 satisfies:

$$F_2^1(0,\epsilon,t) - F_1^1(0,\epsilon,t) = \theta(\epsilon) J^0_{\mathrm{int}}(\epsilon,t), \quad \forall \epsilon \in \mathscr{R}_1 \cap \mathscr{R}_2, \tag{IV.28}$$

together with the current relations (IV.9) and (IV.10).

Proof: Inserting ansatz (IV.25) into the transmission condition (III.5), we get, at $x = 0$:

$$\begin{aligned}
&(f_i^\alpha)^- - \mathscr{T}_i((f_1^\alpha)^+, (f_2^\alpha)^+) \\
&\quad = (F_i^0(\epsilon_i)^- - \mathscr{T}_i(F_1^0(\epsilon_1)^+, F_2^0(\epsilon_2)^+) + \alpha \Bigg[\left(\theta_i J^0_{\mathrm{int}} - \lambda_{i,x} \frac{\partial F_i^0}{\partial x} \right)^- \\
&\qquad - \mathscr{T}_i \left(\left(\theta_1 J^0_{\mathrm{int}} - \lambda_{1,x} \frac{\partial F_1^0}{\partial x} \right)^+, \left(\theta_2 J^0_{\mathrm{int}} - \lambda_{2,x} \frac{\partial F_2^0}{\partial x} \right)^+ \right) \\
&\qquad + \alpha [(F_i^1 - \theta_i^\infty J^0_{\mathrm{int}})^- - \mathscr{T}_i((F_1^1 - \theta_1^\infty J^0_{\mathrm{int}})^+, (F_2^1 - \theta_2^\infty J^0_{\mathrm{int}})^+)] + \mathscr{O}(\alpha^2)
\end{aligned} \tag{IV.29}$$

The order α^0 term at the right-hand side of (IV.29) is zero because F_i^0 satisfies (IV.8) and because of Lemma III.2. The second one of the order α terms is also zero by a similar application of Lemma III.2 as soon as transmission condition (IV.28) is imposed to F_i^1. For the first one of the order α terms, we use the fact that $\theta_i J^0_{\mathrm{int}}(\epsilon_i)$ is a solution of the two-sided Milne problem (IV.17), (IV.18), and (IV.21) associated with

$$\ell_i(k) = D_i(\epsilon_i)^{-1} \lambda_{i,x}(k) J^0_{\mathrm{int}}(\epsilon_i) = \lambda_{i,x}(k) \frac{\partial F_i^0}{\partial x}(\epsilon_i).$$

Therefore $\theta_i J^0_{\mathrm{int}}(\epsilon_i) - \lambda_{i,x} \partial F_i^0 / \partial x(\epsilon_i)$ satisfies the interface condition (III.5), which shows that the first one of the order α terms in (IV.29) also vanishes. We deduce that ansatz (IV.25) satisfies the transmission condition (III.5) up to order α^2 terms.

Finally, we turn to the flux condition (IV.3). Inserting ansatz (IV.25) into (IV.3) and recalling that fluxes of functions of ϵ only vanish, that the flux of θ_i is zero by definition [condition (IV.21)], and that condition (IV.9) holds for F_i^0, we get for $\epsilon \in \mathscr{R}_1 \cap \mathscr{R}_2$:

$$J_1^\alpha(0,\epsilon,t) - J_2^\alpha(0,\epsilon,t) = \alpha \left[\int_{\mathscr{S}^1_\epsilon} v_{1,x}(k) f_1^2(0,k,t) dN^1_\epsilon(k) - \int_{\mathscr{S}^2_\epsilon} v_{2,x}(k) f_2^2(0,k,t) dN^2_\epsilon(k) \right] + \mathscr{O}(\alpha^2). \tag{IV.30}$$

But, with (IV.15), given that F_i^2 is a function of ϵ only and that h_i^2 is an even function of k, the order α term in (IV.30) reduces to

$$-\left(\int_{\mathscr{S}_\epsilon^1} v_{1,x}(k)\lambda_{1,x}(k)dN_\epsilon^1(k)\right)\frac{\partial F_1^1}{\partial x}(\epsilon)+\left(\int_{\mathscr{S}_\epsilon^2} v_{2,x}(k)\lambda_{2,x}(k)dN_\epsilon^2(k)\right)\frac{\partial F_2^1}{\partial x}(\epsilon).$$

The integral terms in the above expressions are nothing else than $D_i(\epsilon)(=D_i(0,\epsilon,t))$ according to (IV.7). Thus, condition (IV.9) for F_i^1 ensures that the order α term in (IV.30) is identically zero and that relation (IV.3) holds with $k=2$. The same arguments can be used to derive relation (IV.4) with $k=2$, which completes the proof of Theorem IV.4. ■

The construction of a complete $O(\alpha^2)$-approximation would require $O(\alpha^2)$ and $O(\alpha^3)$ layer terms. Let instead f_i^2 and f_i^3 be outer approximations given by Proposition IV.3. Inserting ansatz (IV.25) into the Boltzmann equation (III.1) and using the proof of Proposition IV.3, we obtain:

$$\begin{aligned}
&\partial_t f_i^\alpha+\frac{1}{\alpha}\left(v_{i,x}\frac{\partial f_i^\alpha}{\partial x}+\frac{\partial\phi}{\partial x}\frac{\partial f_i^\alpha}{\partial k_x}\right)-\frac{1}{\alpha^2}Q_i(f_i^\alpha)-S_i(f_i^\alpha)\\
&\quad=\alpha^2\left[\partial_t f_i^2+\alpha\partial_t f_i^3-S_i(f_i^2)-\alpha S_i(f_i^3)+\left(v_{i,x}\frac{\partial}{\partial x}+\frac{\partial\phi}{\partial x}\frac{\partial}{\partial k_x}\right)f_i^3\right]\\
&\qquad+\left(\alpha\partial_t+\frac{\partial\phi}{\partial x}\frac{\partial}{\partial k_x}-\alpha S_i\right)\left(\left(\theta_i\left(\frac{x}{\alpha},k\right)-\theta_i^\infty(\epsilon_i(k))\right)J_0^{\mathrm{int}}(\epsilon_i(k),t)\right).
\end{aligned}\tag{IV.31}$$

The first line of the right-hand side of (IV.31) is $O(\alpha^2)$ while the last line of (IV.31) is exponentially small away from $x=0$. Therefore, f_i^α is a formal approximate solution of order 2 to the Boltzmann equation (III.1), in domains of the form $\{|x|\geqslant x_0\}$.

The last step of the procedure is to combine F_i^0 and F_i^1 into a single SHE distribution function:

$$F_i^\alpha=F_i^0+\alpha F_i^1,$$

which satisfies the SHE equation (IV.6) in each Ω_i, the flux conditions (IV.9) and (IV.10) and the jump relation [deduced from (IV.8) for F_i^0 and from (IV.28) for F_i^1]:

$$F_2^\alpha(0,\epsilon,t)-F_1^\alpha(0,\epsilon,t)=\alpha\theta(\epsilon)J_{\mathrm{int}}^0(\epsilon,t).\tag{IV.32}$$

However, in (IV.32), J_{int}^0 is the flux of F_i^0 and not that of F_i^α. It would thus be desirable to replace it by J_{int}^α [with J_{int}^α being defined by (IV.26) with F_i^0 replaced by F_i^α], in order to obtain a self-contained transmission problem for F_i^α. But first, we must check if this problem would be well-posed.

Lemma IV.5: We have

$$\theta(\epsilon)\geqslant 0,\quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2.\tag{IV.33}$$

Proof: We apply the Darrozes–Guiraud inequality (III.19) to the transmission condition

$$(\theta_i-D_i^{-1}\lambda_{i,x})^-=\mathscr{T}_i((\theta_1-D_1^{-1}\lambda_{1,x})^+,(\theta_2-D_2^-\lambda_{2,x})^+),$$

and obtain for $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$,

$$\begin{aligned}
&\int_{\mathscr{S}_\epsilon^2} v_{2,x}(k)\left|\theta_2(0,k)-\lambda_{2,x}(k)D_2(\epsilon)^{-1}\right|^2dN_\epsilon^2(k)\\
&\qquad\leqslant\int_{\mathscr{S}_\epsilon^1} v_{1,x}(k)\left|\theta_1(0,k)-\lambda_{1,x}(k)D_1(\epsilon)^{-1}\right|^2dN_\epsilon^1(k).
\end{aligned}\tag{IV.34}$$

We develop:

$$\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,|\theta_i(0,k)-\lambda_{i,x}(k)D_i(\epsilon)^{-1}|^2 dN_\epsilon^i(k)$$

$$= \int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,|\theta_i(0,k)|^2 dN_\epsilon^i(k) + \left(\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)|\lambda_{i,x}(k)|^2 dN_\epsilon^i(k)\right) D_i(\epsilon)^{-2}$$

$$-2\left(\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\lambda_{i,x}(k)\,\theta_i(0,k) dN_\epsilon^i(k)\right) D_i(\epsilon)^{-1}. \tag{IV.35}$$

Multiplying Eq. (IV.17) by $\theta_i(\xi,k)$, integrating over k, and using property (iii) of Lemma II.1, we get

$$\frac{1}{2}\frac{\partial}{\partial\xi}\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,|\theta_i(\xi,k)|^2 dN_\epsilon^i(k) = \int_{\mathscr{S}_\epsilon^i} Q_i(\theta_i)\,\theta_i dN_\epsilon^i(k) \leqslant 0.$$

Moreover, since $\lim_{|\xi|\to\infty}\theta_i(\xi,k)=\theta_i^\infty(\epsilon_i(k))$ is an even function of k, we have

$$\lim_{|\xi|\to\infty}\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,|\theta_i(\xi,k)|^2 dN_\epsilon^i(k)=0.$$

We deduce

$$\sigma_i\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\,|\theta_i(0,k)|^2 dN_\epsilon^i(k)\leqslant 0, \quad \sigma_1=1,\sigma_2=-1. \tag{IV.36}$$

Then, using that $\lambda_{i,x}(k)$ is an odd function of k, we obtain:

$$\int_{\mathscr{S}_\epsilon^i} v_{i,x}|\lambda_{i,x}|^2 dN_\epsilon^i=0. \tag{IV.37}$$

Finally, we have

$$\int_{\mathscr{S}_\epsilon^i} v_{i,x}(k)\lambda_{i,x}(k)\,\theta_i(0,k) dN_\epsilon^i(k)=K_i(0,\epsilon),$$

where the K-integral $K_i(\xi,\epsilon)$ has been defined at (IV.19) and is constant with respect to ξ [because of (IV.20) and because $J_i=0$]. Thus, using (IV.7):

$$K_i(0,\epsilon)=\lim_{|\xi|\to\infty} K_i(\xi,\epsilon)=D_i(\epsilon)\,\theta_i^\infty(\epsilon). \tag{IV.38}$$

Inserting (IV.35), (IV.36), (IV.37) and (IV.38) into (IV.34) leads to the result. ■

We may now introduce $\tilde{F}_i^\alpha$, a solution of the SHE equation (IV.6) in each domain Ω_i, with the flux relations (IV.9) and (IV.10) at $x=0$ and the jump relation

$$\tilde{F}_2^\alpha-\tilde{F}_1^\alpha-\alpha\theta\tilde{J}_\alpha^{\text{int}}=0, \quad \forall\,\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2, \tag{IV.39}$$

where $\tilde{J}_\alpha^{\text{int}}$ is the common value of the current at the interface [formula (IV.26) with F_i^0 replaced by $\tilde{F}_i^\alpha$]. Relation (IV.39) is the final form of the order 2 transmission condition for the SHE distribution, that we are looking for. We note that $\theta\geqslant 0$ is a sufficient condition for the SHE equation with transmission conditions (IV.9), (IV.10), and (IV.39) to be well-posed. As a matter of example, we look at the following model problem.

Lemma IV.6: The stationary one-dimensional transmission problem for $F(x)$, $x\in(-1,1)$ given by

$$-(DF')'+F=S, \quad x\in(-1,1), \tag{IV.40}$$

$$F(-1)=F(1)=0, \tag{IV.41}$$

$$D_1F_1'=D_2F_2':=J^{\text{int}}, \quad F_1'=\lim_{x\to 0^-}F'(x), \quad F_2'=\lim_{x\to 0^+}F(x), \tag{IV.42}$$

$$F_2-F_1-\theta J^{\text{int}}=0 \quad F_1=\lim_{x\to 0^-}F(x), \quad F_2=\lim_{x\to 0^+}F(x), \tag{IV.43}$$

with $D(x)=D_1$ for $x\in(-1,0)$, $D(x)=D_2$ for $x\in(0,1)$, is well-posed for $\theta\geqslant 0$.

Proof: Multiplying (IV.40) by F, integrating by parts and using (IV.41), (IV.42), and (IV.43), leads to the following variational formulation:

$$\int_{-1}^{1} D|F'|^2\,dx+\int_{-1}^{1}|F|^2\,dx+\theta^{-1}|F_2-F_1|^2=\int_{-1}^{1}FS\,dx. \tag{IV.44}$$

If $\theta=0$, the term $\theta^{-1}|F_2-F_1|^2$ must be replaced by 0. From (IV.44), it is easy to prove existence of solution by Lax–Milgram theory. ■

The difference $F_i^\alpha-\tilde{F}_i^\alpha$ satisfies the same SHE problem (IV.6), (IV.9), (IV.10), but with an inhomogeneous transmission condition (IV.39) where the right-hand side is of order α^2 [use (IV.32)]. A stability theory derived from the existence theory of Lemma IV.6 then shows that $\tilde{F}_i^\alpha-F_i^\alpha$ is of order α^2. By virtue of (IV.25) and of evenness considerations, we have, away from $x=0$:

$$\frac{1}{N_i(x,\epsilon,t)}\int_{\mathscr{S}_{(x,\epsilon,t)}} f_i^\alpha(k)\,dN_\epsilon^i(k)=F_i^\alpha(x,\epsilon,t)+\mathscr{O}(\alpha^2).$$

Therefore, as a final theorem we can state:

Theorem IV.7: Let $\tilde{F}_i^\alpha$ be the solution of the transmission problem for the SHE equation (IV.6) with transmission conditions (IV.9), (IV.10), (IV.39). Then $\tilde{F}_i^\alpha(\epsilon)$ is formally an order 2 approximation of the energy-shell average of f_i^α, away from the interface.

V. THE KINETIC TRANSMISSION LAYER PROBLEM

In this section, we are concerned with the existence and uniqueness theory of the transmission layer problem (IV.17), (IV.18). In particular, rotationally symmetric models will be treated in detail. We shall start, however, dealing with the general case.

Our analysis will be based on assumptions on the one-sided Milne problems

$$v_{i,x}\frac{\partial\theta_i}{\partial\xi}=Q_i(\theta_i), \quad \xi\in I_i=\sigma_i(-\infty,0], \quad k\in\mathscr{S}_\epsilon^i, \tag{V.1}$$

$$\theta_i^- \text{ given on } \mathscr{S}_\epsilon^{i,-}, \tag{V.2}$$

for $\epsilon\in\mathscr{R}_i$, where the abbreviation $\theta_i^\pm=\theta_i(0,\cdot)^\pm$ is used. Note that we use the fact that the Milne problem is decoupled in the energy direction.

Hypothesis 4: For $i=1$ or 2, $\epsilon\in\mathscr{R}_i$ and arbitrary $\theta_i^-\in L^2(\mathscr{S}_\epsilon^{i,-},dN_\epsilon^i)$ the problem (V.1), (V.2) has a unique solution in $L^\infty(I_i,L^2(\mathscr{S}_\epsilon^i,dN_\epsilon^i))$ such that the Albedo operator $\mathscr{A}_i$, determined from $\theta_i^+=\mathscr{A}_i\theta_i^-$ is compact from $L^2(\mathscr{S}_\epsilon^{i,-},dN_\epsilon^i)$ to $L^2(\mathscr{S}_\epsilon^{i,+},dN_\epsilon^i)$. Furthermore, as $|\xi|\to\infty$, θ_i converges exponentially to a constant.

It will be shown below that this assumption can be verified for rotationally symmetric models. Its verification for more general cases remains the subject of future work.

Theorem V.1: Assume the hypotheses 1–4. Then the two-sided Milne problem (IV.17), (IV.18) has a bounded solution iff $\mathscr{C}_1$ and $\mathscr{C}_2$ satisfy the solvability conditions (IV.22), (IV.23). In this case, the solution is unique up to the addition of a solution of the form (IV.24) of the homogeneous problem.

Proof: With the help of the Albedo operators, the transmission conditions (IV.18) can be written as

$$\theta_i^- - \mathcal{T}_i(\mathcal{A}_1\theta_1^-, \mathcal{A}_2\theta_2^-) = \ell_i^- - \mathcal{T}_i(\ell_1^+, \ell_2^+),$$

reducing the problem to a fixed point problem for the θ_i^-:

$$\begin{pmatrix} \theta_1^- \\ \theta_2^- \end{pmatrix} - K\begin{pmatrix} \theta_1^- \\ \theta_2^- \end{pmatrix} = \begin{pmatrix} \ell_1^- \\ \ell_2^- \end{pmatrix} - \begin{pmatrix} \mathcal{T}_1(\ell_1^+, \ell_2^+) \\ \mathcal{T}_2(\ell_1^+, \ell_2^+) \end{pmatrix}, \tag{V.3}$$

with

$$K\begin{pmatrix} \theta_1^- \\ \theta_2^- \end{pmatrix} = \begin{pmatrix} \mathcal{T}_1(\mathcal{A}_1\theta_1^-, \mathcal{A}_2\theta_2^-) \\ \mathcal{T}_2(\mathcal{A}_1\theta_1^-, \mathcal{A}_2\theta_2^-) \end{pmatrix}. \tag{V.4}$$

An application of Jensen's inequality as in the proof of Lemma III.3 and an integration with respect to k shows that $\mathcal{T}_i$ is a continuous operator from $L^2(\mathcal{S}_\epsilon^{1,+}, dN_\epsilon^1) \times L^2(\mathcal{S}_\epsilon^{2,+}, dN_\epsilon^2)$ to $L^2(\mathcal{S}_\epsilon^{i,-}, dN_\epsilon^i)$. Therefore, by Hypothesis 4, K is a compact operator on $L^2(\mathcal{S}_\epsilon^{1,-}, dN_\epsilon^1) \times L^2(\mathcal{S}_\epsilon^{2,-}, dN_\epsilon^2)$.

Let now $\ell_i = 0$, $i = 1,2$. Multiplying (IV.17) by θ_i, integrating it with respect to k, and using Lemma II.1 (iii) yields

$$\frac{1}{2}\frac{\partial}{\partial \xi}\int_{\mathcal{S}_\epsilon^i} v_{i,x}\theta_i^2 dN_\epsilon^i = \int_{\mathcal{S}_\epsilon^i} Q_i(\theta_i)\theta_i dN_\epsilon^i \leqslant 0. \tag{V.5}$$

By Hypothesis 4, θ_i tends to a constant independent of $k \in \mathcal{S}_\epsilon^i$ as $|\xi|$ tends to infinity. Therefore, integrating (V.5) with respect to ξ over I_i and using the oddness of $v_{i,x}$ leads to

$$\int_{\mathcal{S}_\epsilon^1} v_{1,x}(k)\theta_1(0,k)^2 dN_\epsilon^1(k) \leqslant 0 \leqslant \int_{\mathcal{S}_\epsilon^2} v_{2,x}(k)\theta_2(0,k)^2 dN_\epsilon^2(k).$$

Combined with the Darrozes–Guiraud inequality (Lemma III.3), this implies that

$$\int_{\mathcal{S}_\epsilon^i} v_{i,x}(k)\theta_i(0,k)^2 dN_\epsilon^i(k) = 0, \quad i = 1,2.$$

Therefore, we have:

$$\int_{I_i}\int_{\mathcal{S}_\epsilon^i} Q_i(\theta_i)\theta_i dN_\epsilon^i d\xi = 0,$$

which implies, with Lemma II.1 (iii) that θ_i is constant on $\mathcal{S}_\epsilon^i$: $\theta_i = \theta_i(\xi, \epsilon)$. Then using Eq. (IV.17), we see that θ_i is also independent of ξ: $\theta_i = \theta_i(\epsilon)$. Finally, Lemma III.2 implies

$$\theta_1(\epsilon) = \theta_2(\epsilon),$$

completing the proof for the homogeneous case.

We now turn to the inhomogeneous case. It has already been observed in Sec. IV that (IV.22) and (IV.23) are necessary solvability conditions. Since $I - K$ (where I is the identity operator) is a Fredholm operator on $L^2(\mathcal{S}_\epsilon^{1,-}, dN_\epsilon^1) \times L^2(\mathcal{S}_\epsilon^{2,-}, dN_\epsilon^2)$ whose kernel is the one-dimensional vector space spanned by the constants, its range is closed and of codimension 1, and (IV.22) and (IV.23) are thus also sufficient for the existence of a solution. ■

We now specialize to a spherically symmetric model and, thus, consider a spherically symmetric band diagram

$$e_i(k) = \overline{e_i}(|k|) \tag{V.6}$$

with a strictly monotone $\overline{e_i}$, and rotationally symmetric collision and transmission operators. The parabolic band diagram (II.3) is a particular case of (V.6).

Because of the monotonicity hypothesis on $\overline{e_i}$, there exists a one-to-one correspondence between $|k|$ and $\epsilon=\epsilon_i(k)=\overline{e_i}(|k|)-\chi_i$. We denote the inverse relation by $|k|=\kappa_i(\epsilon)$. It will be convenient to introduce a new set of variables $\epsilon\in\mathscr{R}_i$, $\mu\in[-1,1]$, $\varphi\in[0,2\pi)$ such that

$$k=\kappa_i(\epsilon)(\mu,\sqrt{1-\mu^2}\cos\varphi,\sqrt{1-\mu^2}\sin\varphi). \tag{V.7}$$

The measure $dN^i_\epsilon(k)$ is given in these coordinates by

$$dN^i_\epsilon(k)=\overline{v_i}(\epsilon)^{-1}\kappa_i^2(\epsilon)d\mu d\varphi \quad \text{with } \overline{v_i}(\epsilon)=\frac{\partial\overline{e_i}}{\partial|k|}(\kappa_i(\epsilon)). \tag{V.8}$$

The hypothesis of rotational symmetry for Q_i implies that the matrix element $\Phi_i(k,k')$ [see (II.4)] is invariant under all orthogonal transformations of $\mathbf{R}^3$, i.e., $\Phi_i(Rk,Rk')=\Phi(k,k')$, for all R in the orthogonal group $O_3(\mathbf{R})$. Therefore, Φ_i is of the form

$$\Phi_i(k,k')=\tilde{\Phi}_i\left(\epsilon,\frac{k}{|k|}\cdot\frac{k'}{|k'|}\right) \tag{V.9}$$

[since Φ_i needs to be defined only when $\epsilon_i(k)=\epsilon_i(k')=\epsilon$]. A special case is the relaxation time model

$$\Phi_i=\Phi_i(\epsilon)=\frac{\overline{v_i}(\epsilon)}{4\pi\kappa_i(\epsilon)^2\tau_i(\epsilon)}, \tag{V.10}$$

where the collision operator takes the form

$$Q_i(f)=\frac{1}{\tau_i(\epsilon)}\left(\frac{1}{4\pi}\int_{-1}^{1}\int_0^{2\pi}f(\epsilon,\mu',\varphi')d\varphi'd\mu'-f(\epsilon,\mu,\varphi)\right). \tag{V.11}$$

The hypothesis of rotational symmetry for the transmission operator $\mathscr{T}_i$ is that R_i and $T_{i\to j}$ are invariant under all the orthogonal transformations that leave the interface Σ globally invariant. Therefore, $R_i(k\to k')$ for $\epsilon_i(k)=\epsilon_i(k')=\epsilon\in\mathscr{R}_i$ and $T_{i\to j}(k\to k')$ for $\epsilon_i(k)=\epsilon_j(k')=\epsilon$ are of the form:

$$\begin{aligned}
R_i(k\to k')&=\tilde{R}_i\left(\epsilon,\frac{k_yk_y'+k_zk_z'}{\sqrt{k_y^2+k_z^2}\sqrt{k_y'^2+k_z'^2}};\frac{k_x}{|k_x|}\to\frac{k_x'}{|k_x'|}\right),\\
T_{i\to j}(k\to k')&=\tilde{T}_{i\to j}\left(\epsilon,\frac{k_yk_y'+k_zk_z'}{\sqrt{k_y^2+k_z^2}\sqrt{k_y'^2+k_z'^2}};\frac{k_x}{|k_x|}\to\frac{k_x'}{|k_x'|}\right).
\end{aligned} \tag{V.12}$$

With the spherical symmetry assumption and the relaxation time model for the collision operator, Hypothesis 4 can be verified:

Lemma V.2: Assume (V.6) and (V.10). Then Hypothesis 4 holds.

Proof: We have to analyze the one-sided Milne problem

$$\overline{v_i}\mu\frac{\partial\theta_i}{\partial\xi}=Q_i(\theta_i),\quad \xi\in I_i,\quad \mu\in[-1,1],\quad \varphi\in[0,2\pi), \tag{V.13}$$

$$\theta_i(0,\mu,\varphi)=\theta_B(\mu,\varphi),\quad \sigma_i\mu<0, \tag{V.14}$$

where Q_i is given by (V.11). Here and in the following we skip the dependencies on ϵ since it is only a parameter. For the solution we make the ansatz

$$\theta_i(\xi,\mu,\varphi)=\overline{\theta_i}(\xi,\mu)+\begin{cases}e^{-\xi/|\mu|}(\theta_B(\mu,\varphi)-\overline{\theta_B}(\mu)) & \sigma_i\mu<0\\ 0 & \sigma_i\mu>0,\end{cases} \tag{V.15}$$

where $\overline{\theta_B}$ is the average with respect to φ of θ_B:

$$\overline{\theta_B}(\mu)=\frac{1}{2\pi}\int_0^{2\pi}\theta_B(\mu,\varphi)d\varphi.$$

It is easily checked that (V.15) is a solution of (V.13), (V.14) if $\overline{\theta_i}$ satisfies the φ-averaged problem

$$\overline{v}_i\mu\,\frac{\partial\overline{\theta_i}}{\partial\xi}=\frac{1}{\tau_i}\left(\frac{1}{2}\int_{-1}^{1}\overline{\theta_i}(\xi,\mu')d\mu'-\overline{\theta_i}(\xi,\mu)\right),\quad \xi\in I_i,\quad \mu\in[-1,1] \tag{V.16}$$

$$\overline{\theta_i}(0,\mu)=\overline{\theta_B}(\mu),\quad \sigma_i\mu<0. \tag{V.17}$$

Now the proof can be completed by references to the literature. The φ-averaged problem is a version of the Milne problem of neutron transport. Results on existence, uniqueness, and exponential decay to an equilibrium state can be found in Ref. 14. This provides existence for the full problem (V.13), (V.14). Uniqueness can be shown analogously to the corresponding result for (V.16) and (V.17) in Ref. 14. By (V.15), the Albedo operator for the full problem is the composition of the φ-average and the Albedo operator for the averaged problem. The compactness of the latter follows from its representation in terms of the Chandrasekhar H-function.[27] Compactness of the full Albedo operator is then obvious. ■

We now return to the general problem without symmetry assumptions and investigate a scheme which, in the spirit of those reviewed in Ref. 15, approximately solves the two-sided Milne problem and provides approximate values of the parameter θ (IV.27). In the following, a fixed value of $\epsilon\in\mathscr{R}_1\cap\mathscr{R}_2$ is considered. We start by applying the Albedo operators to the rows of (V.3) and obtaining a fixed point problem for the θ_i^+:

$$\theta_i^+=\mathscr{A}_i(\mathscr{T}_i(\theta_1^+,\theta_2^+)+\varphi_i^-),\quad i=1,2. \tag{V.18}$$

Now the Albedo operators are approximated by $\tilde{\mathscr{A}}_i$, $i=1,2$, defined by the requirements that a constant function is returned and that the vanishing flux property of the full problem is preserved:

$$f^+=\tilde{\mathscr{A}}_i f^-\Leftrightarrow f^+=f^+(\epsilon)\quad\text{and}\quad\int_{\mathscr{S}_\epsilon^i}v_{i,x}f\,dN_\epsilon^i=0.$$

The approximating Albedo operators are explicitly given by

$$\tilde{\mathscr{A}}_i f^-=\frac{1}{\alpha_i}\int_{\mathscr{S}_\epsilon^{i,-}}|v_{i,x}|f^-dN_\epsilon^i,$$

with α_i defined in (III.31). When (V.18) is approximated by

$$\theta_i^+=\tilde{\mathscr{A}}_i(\mathscr{T}_i(\theta_1^+,\theta_2^+)+\varphi_i^-),\quad i=1,2 \tag{V.19}$$

the θ_i^+ are constant on $\mathscr{S}_\epsilon^{i,+}$. Application of the transmission operator to constant data gives

$$\mathscr{T}_i(\theta_1^+,\theta_2^+)=\theta_i^++(\theta_j^+-\theta_i^+)\overline{T}_{j\to i}, \tag{V.20}$$

where $\overline{T}_{j\to i}$ has been defined in (III.15) and (III.14) has been used. Substitution in (V.19) gives

$$\theta_i^+-\theta_j^+=\frac{\tilde{\mathscr{A}}_i\varphi_i^-}{\tilde{\mathscr{A}}_i\overline{T}_{j\to i}}, \tag{V.21}$$

where we have used that constant factors can be taken out of $\tilde{\mathscr{A}}_i$ and that $\tilde{\mathscr{A}}_i$ reduces to the identity when applied to constants. The positivity of the denominator in the right-hand side of (V.21) is a consequence of the transmittivity Hypothesis 3.

Of course, the question arises if the result for $\theta_1^+ - \theta_2^+$ is the same for the two formulas ($i=1$, $j=2$ and $i=2$, $j=1$) contained in (V.21). With the help of the reciprocity relation (III.13) it is easy to show that

$$\alpha_1 \tilde{\mathscr{A}}_1 \overline{T}_{2\to 1} = \alpha_2 \tilde{\mathscr{A}}_2 \overline{T}_{1\to 2}$$

holds. On the other hand, a straightforward computation using the flux conservation property of the transmission operator (Lemma III.1) shows that

$$\alpha_1 \tilde{\mathscr{A}}_1 \varphi_1^- = -\alpha_2 \tilde{\mathscr{A}}_2 \varphi_2^-$$

holds iff the ℓ_i satisfy the solvability condition (IV.22). The approximating problem requires the same solvability condition as the original one.

The *K*-integral (IV.19)

$$K_i = \int_{\mathscr{S}_\epsilon^i} v_{i,x} \lambda_{i,x} \theta_i \, dN_\epsilon^i$$

is constant in each Ω_i. Evaluation at $|\xi| = \infty$ gives

$$K_i = D_i \theta_i^\infty .$$

With the transmission conditions (IV.18) and with (V.20), $\theta_i(\xi = 0)$ can be expressed in terms of the θ_i^+:

$$K_i = \int_{\mathscr{S}_\epsilon^{i,-}} v_{i,x} \lambda_{i,x} \varphi_i^- \, dN_\epsilon^i + D_i(\theta_i^+ + \beta_i(\theta_j^+ - \theta_i^+)),$$

with

$$\beta_i = \frac{1}{D_i} \int_{\mathscr{S}_\epsilon^{i,-}} v_{i,x} \lambda_{i,x} \overline{T}_{j\to i} \, dN_\epsilon^i .$$

Note that $\overline{T}_{j\to i} \leqslant 1$ (III.14) and the oddness of $v_{i,x}$ and $\lambda_{i,x}$ imply

$$\beta_i \leqslant \tfrac{1}{2}.$$

By elimination of K_i, θ_i^∞ can be computed. Taking the difference gives

$$\theta = \theta_2^\infty - \theta_1^\infty = (1-\beta_1-\beta_2)(\theta_2^+ - \theta_1^+) + \frac{1}{D_2} \int_{\mathscr{S}_\epsilon^{2,-}} v_{2,x} \lambda_{2,x} \varphi_2^- \, dN_\epsilon^2 - \frac{1}{D_1} \int_{\mathscr{S}_\epsilon^{1,-}} v_{1,x} \lambda_{1,x} \varphi_1^- \, dN_\epsilon^1 . \tag{V.22}$$

The formulation of the second-order transmission conditions for the SHE-model requires the value of θ for the choice $\ell_i = \lambda_{i,x}/D_i$ of the inhomogeneities. In this case, our result can be simplified.

Lemma V.3: Assume $\ell_i = \lambda_{i,x}/D_i$. Then

$$\alpha_i \tilde{\mathscr{A}}_i \varphi_i^- = \sigma_i (1-\beta_1-\beta_2).$$

Proof: With

$$\varphi_i^- = \frac{\lambda_{i,x}}{D_i} - \mathscr{T}_i\left(\frac{\lambda_{1,x}^+}{D_1}, \frac{\lambda_{2,x}^+}{D_2}\right)$$

we obtain $\alpha_i \tilde{\mathscr{A}}_i \varphi_i^- = A + B + C$ with

$$A=\frac{1}{D_i}\int_{\mathscr{S}_\epsilon^{i,-}}|v_{i,x}|\lambda_{i,x}dN_\epsilon^i=\frac{\sigma_i}{2},$$

as well as

$$B=-\frac{1}{D_i}\int_{\mathscr{S}_\epsilon^{i,-}}\int_{\mathscr{S}_\epsilon^{i,+}}R_i(k'\to k)|v_{i,x}(k)|\lambda_{i,x}(k')dN_\epsilon^i(k')dN_\epsilon^i(k)=\sigma_i(\tfrac{1}{2}-\beta_i) \qquad \text{(V.23)}$$

by the reciprocity relation (III.12) and by (III.14). Similarly we show

$$C=-\frac{1}{D_j}\int_{\mathscr{S}_\epsilon^{i,-}}\int_{\mathscr{S}_\epsilon^{j,+}}T_{j\to i}(k'\to k)|v_{i,x}(k)|\lambda_{j,x}(k')dN_\epsilon^j(k')dN_\epsilon^i(k)=\sigma_j\beta_j, \qquad \text{(V.24)}$$

completing the proof. ■

After substitution of (V.21), (V.22) becomes

$$\theta=\frac{(1-\beta_1-\beta_2)^2}{\alpha_i\tilde{\mathscr{K}}_i\overline{T}_{j\to i}}-\sum_{i=1}^{2}\frac{\sigma_i}{D_i}\int_{\mathscr{S}_\epsilon^{i,-}}v_{i,x}\lambda_{i,x}\varphi_i^-dN_\epsilon^i, \qquad \text{(V.25)}$$

which is explicit up to the computation of $\lambda_{i,x}$.

A simplification can be achieved by returning to the assumptions (V.6), (V.10), and (V.12) of spherical symmetry and the relaxation time model for the collision operator. In this case, $\lambda_{i,x}$ can be computed explicitly:

$$\lambda_{i,x}=\tau_i v_{i,x}=\tau_i\overline{v}_i\mu.$$

The quantities needed in (V.25) are

$$\beta_i=\frac{3}{2}\int_{\sigma_i\mu<0}\mu^2\overline{T}_{j\to i}(\epsilon,\mu)d\mu,$$

$$\alpha_i\tilde{\mathscr{K}}_i\overline{T}_{j\to i}=2\pi\kappa_i^2\int_{\sigma_i\mu<0}|\mu|\overline{T}_{j\to i}(\epsilon,\mu)d\mu,$$

with

$$\overline{T}_{j\to i}(\epsilon,\mu)=\frac{\kappa_j^2}{\overline{v}_j}\int_{\sigma_j\mu'>0}\int_0^{2\pi}\tilde{T}_{j\to i}(\epsilon,\cos\varphi;\mu'\to\mu)d\varphi d\mu',$$

as well as

$$\frac{1}{D_i}\int_{\mathscr{S}_\epsilon^{i,-}}v_{i,x}\lambda_{i,x}\varphi_i^-dN_\epsilon^i=\frac{3}{2}\int_{\sigma_i\mu<0}\mu^2\varphi_i^-(\epsilon,\mu)d\mu,$$

with

$$\varphi_i^-(\epsilon,\mu)=\frac{3\mu}{4\pi\kappa_i^2}-\frac{3}{4\pi\overline{v}_i}\int_{\sigma_i\mu'>0}\int_0^{2\pi}\mu'\tilde{R}_i(\epsilon,\cos\varphi;\mu'\to\mu)d\varphi d\mu'$$
$$-\frac{3}{4\pi\overline{v}_j}\int_{\sigma_j\mu'>0}\int_0^{2\pi}\mu'\tilde{T}_{j\to i}(\epsilon,\cos\varphi;\mu'\to\mu)d\varphi d\mu'.$$

We conclude this section by applying our results to the examples of transmission operators from Sec. III.

Example 1: Semiclassical transmission conditions (III.24). Let k_1, k_2 be such that $k_2 = t_{2\to 1} k_1$ or $k_1 = t_{1\to 2} k_2$ [see (III.26)]. Relation (III.23) translates to the (ϵ,μ) geometry into $\epsilon_1 = \epsilon_2 = \epsilon$ and:

$$\kappa_1^2(\epsilon)(1-\mu_1^2) = \kappa_2^2(\epsilon)(1-\mu_2^2). \tag{V.26}$$

We now suppose, without loss of generality, that $\kappa_1^2(\epsilon)/\kappa_2^2(\epsilon) > 1$ and define $\mu_0 \in (0,1)$ by $\kappa_2^2(\epsilon) = (1-\mu_0^2)\kappa_1^2(\epsilon)$. Equation (V.26) is then written

$$1-\mu_1^2 = (1-\mu_0^2)(1-\mu_2^2). \tag{V.27}$$

It allows us to define the particle transmission process $t_{j\to i}$ in the (ϵ,μ) system of coordinates by $\mu_j = t_{j\to i}\mu_i$ with, for all $\mu_2 > 0$ and $\mu_1 < -\mu_0$:

$$t_{1\to 2}\mu_2 = \sqrt{1-(1-\mu_0^2)(1-\mu_2^2)}, \quad t_{2\to 1}\mu_1 = -\sqrt{1-\frac{1-\mu_1^2}{1-\mu_0^2}}. \tag{V.28}$$

If $-\mu_0 < \mu_1 < 0$, $t_{2\to 1}\mu_1$ is not defined. The reflection operator r is simply defined by $r\mu_i = -\mu_i$. Rotational symmetry for R_i and $T_{i\to j}$ respectively, given by (III.28) and (III.29) simply means that the transmission probability $T_{j\to i}$ satisfies: $T_{j\to i} = T_{j\to i}(\epsilon_i, \mu_i)$ for all $\sigma_i\mu_i < 0$, $i = 1,2$, and with $0 \leqslant \overline{T}_{j\to i} \leqslant 1$. The reciprocity relation (III.30) gives:

$$\overline{T}_{2\to 1}(\epsilon,\mu) = \begin{cases} \overline{T}_{1\to 2}(\epsilon, r t_{2\to 1}\mu), & \mu < -\mu_0 \\ 0 \quad \text{else.} \end{cases} \tag{V.29}$$

The transmission and reflection rates (V.12) are now given by

$$\tilde{R}_i(\epsilon;\mu\to\mu') = \frac{\overline{v}_i}{2\pi\kappa_i^2}\,(1-\overline{T}_{j\to i}(\epsilon,\mu'))\,\delta(\mu+\mu'),$$

$$\tilde{T}_{i\to j}(\epsilon;\mu\to\mu') = \frac{\overline{v}_i}{2\pi\kappa_i^2}\,\overline{T}_{j\to i}(\epsilon,\mu')\,\delta(\mu - t_{i\to j}\mu').$$

Of course, purely classical transmission is described within this framework with $\overline{T}_{1\to 2}(\mu) = 1$, for $\mu \in [0,1]$. Now we substitute into the above formulas for the spherically symmetric case and compute

$$\beta_1 + \beta_2 = \frac{3\kappa_2^2}{2}\int_0^1 \mu\left(\frac{t_{1\to 2}\mu}{\kappa_1^2} + \frac{\mu}{\kappa_2^2}\right)\overline{T}_{1\to 2}(\epsilon,\mu)\,d\mu,$$

$$\alpha_i \tilde{\mathscr{B}}_i \overline{T}_{j\to i} = 2\pi\kappa_2^2 \int_0^1 \mu\,\overline{T}_{1\to 2}(\epsilon,\mu)\,d\mu,$$

$$-\sum_{i=1}^{2}\frac{\sigma_i}{D_i}\int_{\mathscr{S}_\epsilon^{i,-}} v_{i,x}\lambda_{i,x}\varphi_i^- dN_\epsilon^i = \frac{9}{16\pi}\left[\frac{1}{\kappa_1^2} + \frac{1}{\kappa_2^2} - 2\kappa_2^2\int_0^1 \mu\left(\frac{t_{1\to 2}\mu}{\kappa_1^2} + \frac{\mu}{\kappa_2^2}\right)^2 \overline{T}_{1\to 2}(\epsilon,\mu)\,d\mu\right].$$

Note that these results apply to the case $\kappa_2^2/\kappa_1^2 < 1$. In the converse situation $\kappa_2^2/\kappa_1^2 > 1$, the roles of 1 and 2 must be exchanged. The above formulas cannot be simplified further without specifying the values of the transmission coefficients. In the case of purely classical transmission ($\overline{T}_{1\to 2} = 1$) we substitute in (V.25) and obtain

$$\theta = \frac{9}{32\pi\kappa_1^2}\,\mu_0^4\left[1 + \frac{8}{9}\frac{\mu_0^2}{1-\mu_0^2} + \frac{1}{2(1-\mu_0^2)^{3/2}}\ln\left(\frac{1+\sqrt{1-\mu_0^2}}{1-\sqrt{1-\mu_0^2}}\right)\right]. \tag{V.30}$$

We note that $\theta=0$ when $\kappa_1=\kappa_2$ ($\mu_0^2=1-\kappa_2^2/\kappa_1^2=0$), i.e., when there is no heterojunction at all. On the other hand, $\theta\to\infty$ when $\kappa_2\to 0$, i.e., when ϵ tends to the boundary of the region $\mathscr{R}_1\cap\mathscr{R}_2$. Thanks to that, when ϵ crosses the boundary between $\mathscr{R}_1\cap\mathscr{R}_2$ and $\mathscr{R}_1\setminus\mathscr{R}_2$, there is a continuous shift from the transmission conditions (IV.39), (IV.9), to the Neumann condition (IV.10). ■

Example 2: Isotropic reflection-transmission conditions. We consider the reflection-transmission rates (III.32). Clearly, the corresponding transmission operator is rotationally symmetric and the reflection and transmission rates (V.12) are

$$\tilde{R}_i(\epsilon;\mu\to\mu')=\frac{\overline{v_i}(\epsilon)}{\pi\kappa_i(\epsilon)^2}\,(1-\pi\kappa_j(\epsilon)^2\hat{T}(\epsilon))|\mu|,$$

$$\tilde{T}_{i\to j}(\epsilon;\mu\to\mu')=\overline{v_i}(\epsilon)\hat{T}(\epsilon)|\mu|$$

with $\hat{T}$ being the transmission coefficient from (III.32), satisfying $0\leqslant\hat{T}\pi\kappa_i^2\leqslant 1$, $i=1,2$. If, in addition, relaxation time collision operators are considered, the above formulas yield, after some easy computations:

$$\theta=\frac{1}{\pi^2\kappa_1^2\kappa_2^2}\left(\frac{1}{\hat{T}}-\frac{15\pi}{32}\,(\kappa_1^2+\kappa_2^2)\right)\geqslant 0.$$

■

VI. APPLICATION TO THE ENERGY-TRANSPORT AND DRIFT-DIFFUSION MODELS

The energy-transport (ET) and drift-diffusion (DD) models are usually directly derived from the semiconductor Boltzmann equation (II.1) by a diffusion approximation using appropriate assumptions on the dominant scattering (see Refs. 10 and 13). However, using the knowledge that the equilibrium distribution of electrons is a Fermi–Dirac distribution, they can also be derived from the SHE model by the moment method, using a Fermi–Dirac ansatz to compute the fluxes. We shall use this method to derive the ET and DD models for an x-dependent band diagram, and their associated transmission conditions at a heterojunction interface.

We first consider a band diagram which depends smoothly on x and define the macroscopic density n, total energy $\mathscr{E}_{\rm tot}$, particle current j_n and total energy current $j_{\mathscr{E}_{\rm tot}}$ according to

$$\begin{pmatrix} n \\ n\mathscr{E}_{\rm tot}\end{pmatrix}=\int_{\mathbf{R}} F\begin{pmatrix}1\\ \epsilon\end{pmatrix}N d\epsilon,\quad \begin{pmatrix} j_n \\ j_{\mathscr{E}_{\rm tot}}\end{pmatrix}=-\int_{\mathbf{R}} D\nabla_x F\begin{pmatrix}1\\ \epsilon\end{pmatrix}d\epsilon, \tag{VI.1}$$

where F is a solution of the SHE equation (II.17). We also let q_n and $q_{\mathscr{E}_{\rm tot}}$ the particle and total energy gain rates due to the collisions modeled by $S(F)$:

$$\begin{pmatrix} q_n \\ q_{\mathscr{E}_{\rm tot}}\end{pmatrix}=\int_{\mathbf{R}} S(F)\begin{pmatrix}1\\ \epsilon\end{pmatrix}d\epsilon.$$

Then, multiplying (II.17) by 1 and ϵ and integrating over ϵ leads to the particle and total energy conservation equations:

$$\frac{\partial n}{\partial t}+\nabla_x\cdot j_n=q_n, \tag{VI.2}$$

$$\frac{\partial}{\partial t}\,(n\mathscr{E}_{\rm tot})+n\,\frac{\partial\phi}{\partial t}+\nabla_x\cdot j_{\mathscr{E}_{\rm tot}}=q_{\mathscr{E}_{\rm tot}}. \tag{VI.3}$$

The ET model is obtained by closing these equations assuming that F is a Fermi–Dirac distribution:

$$F(x,\epsilon,t)=\mathscr{F}_{\epsilon_F,T}(\epsilon):=\frac{1}{\exp\left(\frac{\epsilon-\epsilon_F}{T}\right)+1}, \tag{VI.4}$$

where $\epsilon_F=\epsilon_F(x,t)$ and $T=T(x,t)$ are, respectively, the scaled Fermi energy and temperature. We define the functions $n(\epsilon_F,T;x,t)$ and $n\mathscr{E}_{\rm tot}(\epsilon_F,T;x,t)$ by:

$$\begin{pmatrix} n \\ n\mathscr{E}_{\rm tot}\end{pmatrix}(\epsilon_F,T;x,t)=\int_{\mathbf{R}}\mathscr{F}_{\epsilon_F,T}(\epsilon)\begin{pmatrix}1\\ \epsilon\end{pmatrix}N(x,\epsilon,t)\,d\epsilon, \tag{VI.5}$$

where the (x,t) dependence of n and $n\mathscr{E}_{\rm tot}$ comes from the dependence of the band diagram characteristics upon (x,t). We also define the diffusion matrices $D_{ij}=D_{ij}(\epsilon_F,T;x,t)$, $i=1,2$, $j=1,2$, by

$$D_{ij}(\epsilon_F,T;x,t)=\int_{\mathbf{R}}\mathscr{F}_{\epsilon_F,T}(\epsilon)(1-\mathscr{F}_{\epsilon_F,T}(\epsilon))D(x,\epsilon,t)\,\epsilon^{i+j-2}d\epsilon. \tag{VI.6}$$

We note that each D_{ij} is a non-negative symmetric 3×3 matrix, that $D_{12}=D_{21}$, and that under weak geometric assumptions on $e(x,k)$, the 6×6 matrix

$$D=\begin{pmatrix}D_{11} & D_{12}\\ D_{21} & D_{22}\end{pmatrix}$$

is symmetric positive definite (see Ref. 9 for details). With these definitions, inserting (VI.4) in (VI.1) leads to

$$j_n=-D_{11}\nabla_x\left(\frac{\epsilon_F}{T}\right)-D_{12}\nabla_x\left(-\frac{1}{T}\right), \tag{VI.7}$$

$$j_{\mathscr{E}_{\rm tot}}=-D_{21}\nabla_x\left(\frac{\epsilon_F}{T}\right)-D_{22}\nabla_x\left(-\frac{1}{T}\right). \tag{VI.8}$$

The ET model written in terms of the Fermi energy-temperature variables consists in equations (VI.2), (VI.3), (VI.7), (VI.8) together with the closure relation (VI.5). The DD model is just the ET model in which the temperature is kept constant equal to the lattice temperature. If the lattice temperature is uniform, its formulation in terms of the Fermi-energy variable consists of (VI.2) and (VI.7) in which $\nabla_x(-1/T)=0$. The ET and DD models are sometimes more conveniently expressed in terms of the chemical potential ϵ_F+V rather than in terms of the Fermi energy ϵ_F.[10]

Now, we turn toward discontinuous band diagrams. We shall use the results of Secs. IV and V to derive appropriate transmission conditions for the ET and DD models. We begin with the first-order transmission conditions (IV.8), (IV.9), (IV.10). First, inserting the Fermi–Dirac ansatz (VI.4) into relation (IV.8) leads to the continuity of the temperatures and of the Fermi levels, i.e., with obvious notations:

$$\epsilon_{F_1}=\epsilon_{F_2},\quad T_1=T_2. \tag{VI.9}$$

Then, inserting the Fermi–Dirac ansatz (VI.4) and taking the first two moments (in ϵ) of relations (IV.9), (IV.10) leads to:

$$\left[\int_{\mathscr{R}_1\cap\mathscr{R}_2}D_i(\epsilon)(\mathscr{F}_i(1-\mathscr{F}_i))(\epsilon)\left(\nabla\left(\frac{\epsilon_{F_i}}{T_i}\right)+\epsilon\nabla\left(-\frac{1}{T_i}\right)\right)\begin{pmatrix}1\\ \epsilon\end{pmatrix}d\epsilon\right]=0, \tag{VI.10}$$

$$\int_{\mathscr{R}_i\setminus\mathscr{R}_j}D_i(\epsilon)(\mathscr{F}_i(1-\mathscr{F}_i))(\epsilon)\left(\nabla\left(\frac{\epsilon_{F_i}}{T_i}\right)+\epsilon\nabla\left(-\frac{1}{T_i}\right)\right)\begin{pmatrix}1\\ \epsilon\end{pmatrix}d\epsilon=0,\quad i=1,2, \tag{VI.11}$$

where $\mathscr{F}_i := \mathscr{F}_{\epsilon_{F_i},T_i}$ and where, for a quantity A, we denote by $[A]$ its jump across the heterojunction: $[A_i] = A_2 - A_1$. But (VI.11) can be added to the corresponding term of (VI.10) to give the flux continuity condition [using (VI.6)–(VI.8)]:

$$j_{n_1} = j_{n_2} := j_n^{\text{int}}, \quad (j_{\mathscr{E}_{\text{tot}}})_1 = (j_{\mathscr{E}_{\text{tot}}})_2 := j_{\mathscr{E}_{\text{tot}}}^{\text{int}}. \tag{VI.12}$$

The first-order transmission conditions for the ET model at a heterojunction interface thus consist in (VI.9), (VI.12). Only the continuity of ϵ_F and j_n needs to be retained for the DD model. These conditions are classical for the DD model, and are derived or used by many authors (see, e.g., Ref. 13).

We now turn toward the second-order transmission condition (IV.39). Of course, the continuity of the fluxes (VI.12) remains unchanged in the second-order conditions. Now, taking the moments of (IV.39), we get:

$$\int_{\mathscr{R}_1 \cap \mathscr{R}_2} \frac{1}{\theta(\epsilon)} (F_2^\alpha - F_1^\alpha) \begin{pmatrix} 1 \\ \epsilon \end{pmatrix} d\epsilon = \alpha \int_{\mathscr{R}_1 \cap \mathscr{R}_2} J_\alpha^{\text{int}} \begin{pmatrix} 1 \\ \epsilon \end{pmatrix} d\epsilon. \tag{VI.13}$$

Inserting the Fermi–Dirac ansatz (VI.4) in the right-hand side of (VI.13) and using (VI.10)–(VI.12), we obtain:

$$\int_{\mathscr{R}_1 \cap \mathscr{R}_2} J_\alpha^{\text{int}} \begin{pmatrix} 1 \\ \epsilon \end{pmatrix} d\epsilon = \begin{pmatrix} j_n^{\text{int}} \\ j_{\mathscr{E}_{\text{tot}}}^{\text{int}} \end{pmatrix}. \tag{VI.14}$$

Now, the Fermi–Dirac ansatz (VI.4) and Taylor's formula lead to:

$$F_2^\alpha - F_1^\alpha = \mathscr{F}_{\epsilon_{F_2},T_2} - \mathscr{F}_{\epsilon_{F_1},T_1} = \mathscr{G}(\epsilon) \left(\left[\frac{\epsilon_F}{T} \right] + \epsilon \left[-\frac{1}{T} \right] \right), \tag{VI.15}$$

with

$$\mathscr{G}(\epsilon) = \int_0^1 \mathscr{F}_s(\epsilon)(1 - \mathscr{F}_s(\epsilon)) ds = \mathscr{G}(\epsilon; \epsilon_{F_1}, T_1, \epsilon_{F_2}, T_2),$$

$$\mathscr{F}_s(\epsilon) = \left(\exp\left((1-s) \frac{\epsilon - \epsilon_{F_1}}{T_1} + s \frac{\epsilon - \epsilon_{F_2}}{T_2} \right) + 1 \right)^{-1},$$

$$\left[\frac{\epsilon_F}{T} \right] = \frac{\epsilon_{F_2}}{T_2} - \frac{\epsilon_{F_1}}{T_1}, \quad \left[-\frac{1}{T} \right] = -\left(\frac{1}{T_2} - \frac{1}{T_1} \right).$$

Inserting (VI.15) into the left-hand side of (VI.13), we are led to

$$\int_{\mathscr{R}_1 \cap \mathscr{R}_2} \frac{1}{\theta(\epsilon)} (F_2^\alpha - F_1^\alpha) \begin{pmatrix} 1 \\ \epsilon \end{pmatrix} d\epsilon = \Lambda \begin{pmatrix} \left[\frac{\epsilon_F}{T} \right] \\ \left[-\frac{1}{T} \right] \end{pmatrix}, \tag{VI.16}$$

where $\Lambda = (\Lambda_{ij})_{i=1,2;j=1,2} = \Lambda(\epsilon_{F_1}, T_1, \epsilon_{F_2}, T_2)$ is the 2×2 matrix given by:

$$\Lambda_{k\ell} = \int_{\mathscr{R}_1 \cap \mathscr{R}_2} \frac{1}{\theta(\epsilon)} \mathscr{G}(\epsilon) \epsilon^{k+\ell-2} d\epsilon, \quad k = 1,2; \ell = 1,2. \tag{VI.17}$$

Following (VI.13), equating (VI.16) and (VI.14) leads to the following transmission condition for the ET model:

$$\Lambda(\epsilon_{F_1}, T_1, \epsilon_{F_2}, T_2)\begin{pmatrix} \left[\dfrac{\epsilon_F}{T}\right] \\ \left[-\dfrac{1}{T}\right] \end{pmatrix} = \alpha \begin{pmatrix} j_n^{\text{int}} \\ j_{\mathscr{E}_{\text{tot}}}^{\text{int}} \end{pmatrix}. \tag{VI.18}$$

For Λ to be finite, we need $1/\theta(\epsilon)$ to be integrable. This is an additional requirement to be imposed on the reflection-transmission operators R_i, $T_{i\to j}$. This requirement is satisfied for classical reflection-transmission across a plane interface (V.30), see the last discussion of example 1, Sec. V. When θ^{-1} is integrable, it is an easy matter to check that Λ is a symmetric positive-definite matrix.

Therefore, second-order transmission conditions for the ET model consist of the flux continuity conditions (VI.12), and the jump conditions (VI.18). The symmetry and positive definiteness of Λ is a sufficient condition for formal well-posedness of the problem, as can be viewed from an easy extension of Lemma IV.6 to vector valued functions (left to the reader). Second-order transmission conditions for the DD model are simply obtained by taking the first condition (VI.12) and the first component of (VI.18), (with $[-1/T]=0$). At the level of the DD model, interface conditions of this form have been considered in Ref. 18. There the computation of Λ_{11} corresponds to using the crude approximation $\theta_2^+ - \theta_1^+$ from (V.21) for the value of θ.

Clearly, at equilibrium, when the currents j_n^{int} and $j_{\mathscr{E}_{\text{tot}}}^{\text{int}}$ are zero, condition (VI.18) reduces to the continuity of the Fermi energy and temperature (VI.9). In spite of their apparent complexity, the second-order transmission conditions can easily be implemented in the framework of finite element codes or box method codes by means of weak formulations. This point will be developed in future work.

ACKNOWLEDGMENTS

The work of the second author (C.S.) has been supported by the Austrian Science Foundation under Grant No. P11308-MAT. It has been carried out during a visit of C.S. at Univ. Paul Sabatier, Toulouse. The first author (P.D.) acknowledges the support of the EC through the HCM program ''Nonlinear Spatio Temporal Structures in Semiconductors, Fluids and Oscillator Ensembles.''

[1] S. Selberherr, *Analysis and Simulation of Semiconductor Devices* (Springer, Wien, 1984).
[2] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations* (Springer, Wien, 1990).
[3] F. Golse and F. Poupaud, ''Limite fluide des équations de Boltzmann des semiconducteurs pour une statistique de Fermi-Dirac,'' Asymptotic Analysis **6**, 135–160 (1992).
[4] F. Poupaud, ''Diffusion approximation of the linear semiconductor equation: Analysis of boundary layers,'' Asymptotic Analysis **4**, 293–317 (1991).
[5] G. Baccarani and M. R. Wordeman, ''An investigation of steady-states velocity overshoot in silicon,'' Solid-State Electron. **28**, 407–416 (1985).
[6] K. Souissi, F. Odeh, H. H. K. Tang, and A. Gnudi, ''Comparative studies of hydrodynamic and energy-transport model,'' COMPEL (The International Journal for Computations in Mathematics in Electrical and Electronic Engineering) **13**, 439–453 (1994).
[7] P. Dmitruk, A. Saul, and L. Reyna, ''High electric field approximation in semiconductor devices,'' Appl. Math. Lett. **5**, 99–102 (1992).
[8] A. Gnudi, D. Ventura, G. Baccarani, and F. Odeh, ''Two-dimensional MOSFET simulation by means of a multidimensional spherical harmonic expansion of the Boltzmann transport equation,'' Solid-State Electron. **36**, 575–581 (1993).
[9] N. Ben Abdallah and P. Degond, ''On a hierarchy of macroscopic models for semiconductors,'' J. Math. Phys. **37**, 3306–3333 (1996).
[10] N. Ben Abdallah, P. Degond, and S. Genieys, ''An energy-transport model for semiconductors derived from the Boltzmann equation,'' J. Stat. Phys. **84**, 205–231 (1996).
[11] J. Mohring, ''The Spherical Harmonics Expansion Method,'' Dissertation, Univ. Kaiserslautern, 1996.
[12] A. H. Marshack and K. M. Van-Vliet, ''Electrical current in solids with position dependent band structure,'' Solid-State Electron. **21**, 417–427 (1977).
[13] F. Poupaud and A. Yamnahakki, ''Drift-diffusion models for heterostructures in semiconductors,'' in *Pitman Research Notes* (Longman, London, 1996).
[14] C. Bardos, R. Santos, and R. Sentis, ''Diffusion approximation and computation of the critical size,'' Trans. Am. Math. Soc. **284**, 617–649 (1984).
[15] F. Golse, ''Knudsen layers from a computational viewpoint,'' Transp. Theory Stat. Phys. **21**, 211–236 (1992).
[16] P. Degond and C. Schmeiser, ''Kinetic boundary layers and fluid-kinetic coupling in semiconductors'' (unpublished).
[17] A. Yamnahakki, ''Second order boundary conditions for the drift-diffusion equations of semiconductors,'' Math. Methods Appl. Sci. **5**, 429–456 (1995).

[18] C. M. Wu and E. S. Yang, ''Carrier transport across heterojunction interfaces,'' Solid-State Electron. **22**, 241–248 (1979).
[19] S. M. Sze, *Physics of Semiconductor Devices* (Wiley, New York, 1969).
[20] *Hot Electron Transport in Semiconductors*, edited by L. Reggiani (Springer, Berlin, 1985).
[21] C. Schmeiser and A. Zwirchmayr, ''Elastic and drift-diffusion limits of electron-phonon interaction in semiconductors,'' Math. Methods Appl. Sci. **8**, 37–53 (1998).
[22] H. Federer, *Geometric Measure Theory* (Springer, Berlin, 1969).
[23] C. Cercignani, *The Boltzmann Equation and its Applications* (Springer, Berlin, 1988).
[24] F. Nier, ''A semi-classical picture of quantum scattering,'' Ann. Scient. Ec. Norm. Sup. 4^e série **29**, 149–183 (1996).
[25] A. Messiah, *Mécanique Quantique, Tome 1* (Dunod, Paris, 1995).
[26] A. Klar, ''Asymptotic induced domain decomposition methods for kinetic and drift diffusion semiconductor equation'' (unpublished).
[27] S. Chandrasekhar, *Radiative Transfer* (Dover, New York, 1960).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Propagation of waves through non-Gaussian longitudinally correlated random media

Joseph Gozani[a)]
Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado at Boulder, Boulder, Colorado 80309-0449

(Received 8 May 1998; accepted for publication 18 May 1998)

Using successive approximations, we derive a hierarchical set of integro-differential equations for the evolution of the statistical moments of a wave propagating through non-Gaussian longitudinally correlated media. A closure procedure in the spirit of the Ornstein–Uhlenbeck process produces the most general set of equations currently available. The procedure can be applied to a large class of Cauchy problems in mathematical physics. © *1998 American Institute of Physics.* [S0022-2488(98)02509-2]

I. INTRODUCTION

This paper extends previous formulations[1–4] of classical wave propagation through random media to account for the evolution of statistical moments of a wave propagating through longitudinally correlated non-Gaussian medium. Mathematically, this problem is described by a stochastic Schrödinger equation whose potential is the random refractivity. Therefore, we look for equations that govern the evolution of the statistical moments of the wave fields. Our formulation addresses the closure problem of an infinite hierarchy of coupled equations. Thus, it shares common features with other branches of mathematical physics, e.g., the Schrödinger wave equation with random potentials,[5] the Bogoliobov–Born–Green–Kirkwood–Yvon (BBGKY) hierarchy in statistical physics,[6] and the hierarchical set of equations due to Keller and Friedmann in statistical turbulence theory.[7]

The motivation for this problem emerged from laser-propagation experiments through the atmosphere, where the random refractivity due to the turbulent temperature and humidity induce fluctuations on the light waves. When dealing with turbulence, the notion of intermittency is necessarily introduced. The intermittency is a type of nonstationarity of the turbulence due to bursts of thermal plumes. These waveforms often have a ramplike shape on which the small-scale random process is superimposed (e.g., Refs. 8 and 9). Actually, except for the scales, there is no distinction between the small and the large processes; they are part of the complete turbulent process. Thus, the turbulent process is admittedly non-Gaussian, and additionally large-scale boundary effects, e.g., orography, introduce a large-scale correlation along the propagation path. To obtain stable results, averaging over the intermittency is a must. However, one may speculate that it is possible to measure the intermittency of turbulence using laser scintillometry, and, for this, excessive averaging is detrimental.

Although its motivation came from atmospheric optics and from turbulence theory, nonstationary statistics is of great concern to all theories that intend to interpret an experiment. One remedy that is often used is applying a conditional sampling. Here, the experiment is used to examine a theory, which is partial, but otherwise well defined. A second remedy is to randomize the locally well-established partial theory. This can lead to flexibility in interpreting experiments by introducing more parameters. We caution, however, that unless the physical model of the large-scale variability is well understood, this procedure can lead to superfluous flexibility and incorrect interpretations.

In the case of wave propagation, the nonstationarity along the path introduces a longitudinal correlation that prevents the closing of the differential equations, thus path integrals must be used.[10] Additionally, the departure from the Gaussian model means that the tail of the distribution

[a)]Electronic mail: gozani@colorado.edu

 © 1998 American Institute of Physics

has a non-negligible effect, and a second-order description of the medium does not apply.[11] In both cases, the current theory is too limited to be practically useful.[12] The present paper addresses both deviations from the regular random process.

We note that whereas, conceptually, our procedure is simple, the complexity due to dimensionality and hierarchy is quite large. We could handle the excessive bookkeeping only by using multi-index notation, where we treated arrays symbolically as if they were scalars. Eventually, we can close the hierarchy at any level we like and choose a manageable difficulty. Completing the formulation would require a proof of convergence that currently we do not have. We content ourselves by noting that many elements of this procedure were already proven correct in previous studies. Finally, we introduce preparatory steps that seem to be required for a proof of convergence.

The paper is based on four prior journal articles that handled only one aspect of the problem we solve. The formulation is rather complicated; therefore, we summarize their contributions and relationship to the present formulation. Reference 1 obtained the hierarchical statistical moment equations for a Gaussian longitudinally-correlated medium. Reference 2 proved the convergence of this hierarchy as the order of the hierarchy is increased. Reference 3 obtained the statistical closure for a non-Gaussian δ-correlated medium; in this case, no hierarchy was involved and the results can be derived by other methods. Reference 4 suggested an asymptotic expansion based on the hierarchy for a non-Gaussian medium with the inner scale of turbulence as a small parameter. This is its closure assumption, but note that the asymptotic expansion was applied to the entire hierarchy of integro-differential equations.

The present paper extends Refs. 1–3 to account for the evolution of statistical moments of wave propagating through longitudinally correlated non-Gaussian medium. It is different from Ref. 4 in two major respects: (a) the formulation in nonasymptotic, and (b) it uses the δ-correlated model as a closure assumption only for the highest-order equation. The latter model is given a meaning by an asymptotic expansion of the integral term of the integro-differential equation. One may note that our equations are in the configuration domain, whereas Ref. 4 provides a partial Laplace transform on the timelike variable. This is, of course, only a minor difference that depended on the goal of the analysis.

Let us portray the plan of the paper. Section II is the main thrust of the paper. It defines the mathematical problem and provides all the necessary steps of the derivation toward a statistical closure. It also gives a simple example that helps to clarify how to use the method. In Sec. III we suggest an approximation of the δ-correlated process that is meaningful in the context of this closure. In Sec. IV we discuss extensions of the present closure for the sake of proving convergence to the genuine solution. Finally, in Sec. V we summarize the paper and suggest future directions.

II. MATHEMATICAL FORMULATION

A. Definitions

We begin with the well-established Schrödinger-type equation (e.g., Ref. 13):

$$[2ik\partial_z+\boldsymbol{\Delta}_{\mathbf{x}}-\boldsymbol{\Delta}_{\mathbf{y}}+k^2\nu(z,\mathbf{x},\mathbf{y})]\,\gamma(z,\mathbf{x},\mathbf{y})=0, \tag{1}$$

$$\gamma(z=0,\mathbf{x},\mathbf{y})=\gamma_0(\mathbf{x},\mathbf{y})\in L_2(\mathbb{R}^{2(N+M)}),\quad \mathbf{x}\in\mathbb{R}^{2N},\quad \mathbf{y}\in\mathbb{R}^{2M}. \tag{2}$$

This equation describes the stochastic evolution of the product of multiple scalar fields:

$$\gamma(z,\mathbf{x},\mathbf{y})=\prod_{n=1}^{N}\psi(z,x_n)\prod_{m=1}^{M}\psi^*(z,y_n) \tag{3}$$

scattered forward under the narrow-angle approximation. Here $\psi(z,x_n)$ is the complex field envelope, the symbol* denotes complex conjugation, z is the propagation range, $x_n\in\mathbb{R}^2$ and $y_m\in\mathbb{R}^2$ are two-dimensional position vectors transverse to the propagation direction, and k is the reference wave number. In (1) and (2), $\mathbf{x}\equiv(x_1,\ldots,x_N)$ and $\mathbf{y}\equiv(y_1,\ldots,y_M)$ are multi-vectors and $\boldsymbol{\Delta}_{\mathbf{v}}\equiv(\Delta_1+\cdots+\Delta_s)$ is a multi-Laplacian operator with $\mathbf{v}=\mathbf{x}$ or $\mathbf{v}=\mathbf{y}$ and $s=N$ or $s=M$, respectively.

Equations (1) and (2) comprise a Cauchy problem. The timelike variable $z \in [0, Z]$ denotes the propagation range. The spacelike variables are $\mathbf{x}$ and $\mathbf{y}$. Solutions of this equation are well posed in $L([0, Z], H)$, where $H = L_2[\mathbb{R}^{2(N+M)}]$, a separable complex Hilbert space.

The last term in (1), $\nu(z,\mathbf{x},\mathbf{y})$, is the *random scattering potential* that accounts for parametric fluctuations defined over an appropriate probability space. It is defined as

$$\nu(z,\mathbf{x},\mathbf{y}) = \sum_{n=1}^{N} \mu(z,x_n) - \sum_{m=1}^{M} \mu(z,y_m) = \int d^2x' \ \alpha(x')\mu(z,x'), \tag{4}$$

where

$$\alpha(x') = \sum_{n=1}^{N} \delta(x'-x_n) - \sum_{m=1}^{M} \delta(x'-y_m), \tag{5}$$

and $\mu(z,x)$ is a random electric susceptibility. A stochastic process is fully characterized by the *characteristic functional*

$$\Phi_\mu[\omega] \equiv \left\langle \exp\left[i \int dz' \int dx' \ \omega(z',x')\mu(z',x') \right] \right\rangle, \tag{6}$$

and the *cumulant-generating functional* $\Theta_\mu[\omega] \equiv \ln \Phi_\mu[\omega]$. In (6), $\langle \cdot \rangle$ denotes an ensemble average. Here, the subscript μ identifies the process. However, whereas all our derivations could be worked out with $\mu(z,x)$ leading to lengthy expressions, we prefer to formulate all in terms of $\nu(z,\mathbf{x},\mathbf{y})$ that produces shorter, albeit equivalent, expressions. Therefore, we will subsequently omit the subscript of the characteristic functionals that identify the random process, in an understanding it means Φ_ν and Θ_ν.

The characteristic functional of $\nu(z,\mathbf{x},\mathbf{y})$ reads

$$\begin{aligned}\Phi_z[w(z,\mathbf{x},\mathbf{y})] &\equiv \left\langle \exp\left[i \int_0^z dz' \int d\mathbf{x}' \ d\mathbf{y}' \ w(z',\mathbf{x}',\mathbf{y}')\nu(z',\mathbf{x}',\mathbf{y}') \right] \right\rangle \\ &\equiv \exp\{\Theta_z[w(z,\mathbf{x},\mathbf{y})]\} = \exp\left[\sum_{m=1}^{\infty} \frac{1}{m!} \int_0^z dz_1' \int d\mathbf{x}_1' \ d\mathbf{y}_1' \cdots \int_0^z dz_m' \int d\mathbf{x}_m' \ d\mathbf{y}_m' \right. \\ &\quad \left. \times K_\nu^{(m)}(z_1',\ldots,z_m';\mathbf{x}_m',\mathbf{y}_m',\ldots,\mathbf{x}_m',\mathbf{y}_m') w(z_1',\mathbf{x}_1',\mathbf{y}_1') \cdots w(z_m',\mathbf{x}_m',\mathbf{y}_m') \right],\end{aligned} \tag{7}$$

where

$$K_\nu^{(m)}(z_1,\ldots,z_m;\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_m,\mathbf{y}_m) = \langle \nu(z_1,\mathbf{x}_1,\mathbf{y}_1) \cdots \nu(z_m,\mathbf{x}_m,\mathbf{y}_m) \rangle_c \tag{8}$$

are the cumulants (semi-invariants) of order m, with $\langle \cdot \rangle_c$ denoting the cumulant average. In (7), the subscript z of Φ_z and Θ_z signifies dependency on z through the upper limit of the integral as the process is limited to $[0, z]$. From (7),

$$K_\nu^{(m)}(z_1,\ldots,z_m;\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_m,\mathbf{y}_m) \equiv \left. \frac{\delta^m \Theta_z[w(z',\mathbf{x}',\mathbf{y}')]}{\delta w(z_1',\mathbf{x}_1',\mathbf{y}_1') \cdots \delta w(z_m',\mathbf{x}_m',\mathbf{x}_m',\mathbf{y}_m')} \right|_{[w]=0}. \tag{9}$$

For quasi-homogeneity in z, the cumulants of $\nu(z,\mathbf{x},\mathbf{y})$ assume the form

$$\langle \nu(z_1,\mathbf{x}_1,\mathbf{y}_1) \cdots \nu(z_m,\mathbf{x}_m,\mathbf{y}_m) \rangle_c = K_\nu^{(m)}(z_1, z_2 - z_1, \ldots, z_n - z_1; \mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_m,\mathbf{y}_m), \tag{10}$$

with a peak dependency of the differences. In this case, the dependency on z_1 expresses the large-scale variability along the path.

Subsequently, we denote $\mathbf{r} = (\mathbf{x},\mathbf{y}) \in \mathbb{R}^{2(N+M)}$, where $\mathbf{r}$ is an array of $(N+M)$ two-dimensional transverse position vectors. Similarly, $\gamma(z,\mathbf{r}) \equiv \gamma(z,\mathbf{x},\mathbf{y})$, $\nu(z,\mathbf{r}) \equiv \nu(z,\mathbf{x},\mathbf{y})$, $w(z,\mathbf{r}) \equiv w(z,\mathbf{x},\mathbf{y})$, and $\mathscr{L}_\mathbf{r} \equiv \boldsymbol{\Delta}_\mathbf{x} - \boldsymbol{\Delta}_\mathbf{y}$ is an ultrahyperbolic operator. This notation simplifies the descrip-

tion because most of the analysis is done with respect to the timelike variable, z, and not with respect to the spacelike variables $\mathbf{x}$ and $\mathbf{y}$. Finally we note that our derivation has a broader context than atmospheric optics, because it applies as well to a Cauchy problem with a more general linear, although deterministic operator $\mathscr{L}_{\mathbf{r}}$.

For the sake of brevity, we define additional multi-vectors. Let ${}^m\boldsymbol{\rho}$ and ${}^m\mathbf{r}$ be $1\times m$ arrays of spacelike multi-vectors describing the cluster of points on the plane transverse to the propagation direction:

$$\begin{gathered} {}^m\boldsymbol{\rho}\equiv(\mathbf{r}_1,\ldots,\mathbf{r}_m), \quad \mathbf{r}_s=(\mathbf{x}_s,\mathbf{y}_s), \quad s=1,\ldots,m, \\ {}^m\mathbf{r}\equiv(\mathbf{r}_1,\ldots,\mathbf{r}_m), \quad \mathbf{r}_1=\mathbf{r}_2=\cdots=\mathbf{r}_m=\mathbf{r}. \end{gathered} \tag{11}$$

Similarly, the timelike $1\times m$ vectors are

$$\begin{gathered} {}^m\boldsymbol{\tau}\equiv(z_1,\ldots,z_m), \\ {}^m\mathbf{z}\equiv(z_1,\ldots,z_m), \quad z_1=z_2=\cdots=z_m=z, \\ {}^m\mathbf{0}\equiv(z_1,\ldots,z_m), \quad z_1=z_2=\cdots=z_m=0, \\ \pm{}^m\boldsymbol{\infty}\equiv(z_1,\ldots,z_m), \quad z_1\to\pm\infty, \quad z_2\to\pm\infty,\ldots, \quad z_m\to\pm\infty. \end{gathered} \tag{12}$$

The left-superscript m is required constantly because multi-vectors of different lengths are used in the same expressions. Finally, we define a multi-index variational derivative

$$\mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})=\frac{\delta^n}{\delta\nu(z_1,\mathbf{r}_1)\cdots\delta\nu(z_n,\mathbf{r}_n)}. \tag{13}$$

B. Derivation

First we average (1) and (2):

$$[2ik\partial_z+\mathscr{L}_{\mathbf{r}}]\langle\gamma(z,\mathbf{r})\rangle=-k^2\langle\nu(z,\mathbf{r})\gamma(z,\mathbf{r})\rangle, \tag{14}$$

$$\langle\gamma(z=0,\mathbf{r})\rangle=\gamma_0(\mathbf{r}). \tag{15}$$

The term on the right-hand side of (14) is yet statistically unclosed. Contrary to the tractable situation when a δ-correlated process or a Gaussian process is assumed, the term $\langle\nu(z,\mathbf{r})\gamma[\nu(\cdot)]\rangle$ cannot be closed in an exact form, hence a purely partial differential equation is unavailable. On the other hand, it is possible to obtain a set of integro-differential equations by successive approximations. To close this average, we use the expansion derived by Klyatskin[14] using operators with variational derivatives,

$$\begin{aligned} \langle\nu(z,\mathbf{r})\gamma[\nu(\cdot)]\rangle &\equiv\left\langle\frac{1}{i(\delta/i\,\delta\nu(z',\mathbf{r}'))}\,\frac{\partial}{\partial z}\,\Theta_z\left[\frac{\delta}{i\,\delta\nu(z',\mathbf{r}')}\right]\gamma[\nu(\cdot)]\right\rangle \\ &=\sum_{m=0}^{\infty}\frac{1}{m!}\int_{{}^m\mathbf{0}}^{{}^m\mathbf{z}}d^m\boldsymbol{\tau}'\int d^m\boldsymbol{\rho}'\,K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')\left\langle\frac{\delta^m\gamma(z,\mathbf{r})}{\delta\nu(z_1',\mathbf{r}_1')\cdots\delta\nu(z_m',\mathbf{r}_m')}\right\rangle, \end{aligned} \tag{16}$$

where

$$K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau},{}^m\boldsymbol{\rho})=\langle\nu(z,\mathbf{r})\nu(z_1,\mathbf{r}_1)\cdots\nu(z_m,\mathbf{r}_m)\rangle_c. \tag{17}$$

In (16), $\delta/\delta\nu(z_m,\mathbf{r}_m)$, $m\in\mathbb{N}$, denotes a variational derivative, where $\delta\nu(z_n,\mathbf{r}_n)$ is a surface differing from zero only in the vicinity of $(z_m,\mathbf{r}_m)\in([0,z],\mathbb{R}^{2(N+M)})$. The index $m=0$ means that no integration is involved on the right-hand side of (16), and, therefore, there in no need for multi-vectors for this term.

Let us take the nth variational derivative of (1) and average the result. Upon applying the variational derivatives as indicated in (16) for $z_1,\ldots,z_n<z$, we find that $\langle \mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$ follows an integro-differential equation similar to (14), i.e.,

$$[2ik\partial_z+\mathscr{L}_{\mathbf{r}}]\langle \mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle=-k^2\sum_{m=0}^{\infty}\frac{1}{m!}\int_{m_0}^{m\mathbf{z}}d^m\boldsymbol{\tau}'\int d^m\boldsymbol{\rho}'\,K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')\times\langle \mathbf{D}^{m+n}({}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle, \tag{18}$$

where

$$\begin{aligned}\mathbf{D}^{m+n}({}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})&\equiv\mathbf{D}^m({}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')\mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\\&=\mathbf{D}^{m+n}(({}^n\boldsymbol{\tau},{}^m\boldsymbol{\tau}'),({}^n\boldsymbol{\rho},{}^m\boldsymbol{\rho}'))\\&=\mathbf{D}^{m+n}({}^{m+n}\boldsymbol{\tau},{}^{m+n}\boldsymbol{\rho}).\end{aligned} \tag{19}$$

Note that in (18), $\mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})(\nu(z,\mathbf{r})\,\gamma(z,\mathbf{r}))$ produces terms that contain

$$\frac{\delta\nu(z,\mathbf{r})}{\delta\nu(z_n,\mathbf{r}_n)}=\delta(\mathbf{r}-\mathbf{r}_n)\,\delta(z-z_n)=0,\quad z_n<z, \tag{20}$$

hence are dropped.

We seek the related initial conditions for this hierarchical set. The initial condition for $\langle\gamma\rangle$ at $z=0$ is γ_0, which is deterministic. The initial condition for $\langle \mathbf{D}^n({}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$ is found using the integro-differential equation:

$$2ik\gamma(z,\mathbf{r})=2ik\gamma_0(\mathbf{r})-\int_0^z dz'\;\mathscr{L}_{\mathbf{r}}\gamma(z',\mathbf{r})-k^2\int_0^z dz'\;\nu(z',\mathbf{r})\,\gamma(z',\mathbf{r}). \tag{21}$$

Applying $\delta/\delta\nu(z_1,\mathbf{r}_1)$ for $z_1<z$ we obtain

$$\begin{aligned}2ik\,\frac{\delta\gamma(z,\mathbf{r})}{\delta\nu(z_1,\mathbf{r}_1)}&=-\int_{z_1}^z dz'\;\mathscr{L}_{\mathbf{r}}\,\frac{\gamma(z',\mathbf{r})}{\delta\nu(z_1,\mathbf{r}_1)}\\&\quad-k^2\int_{z_1}^z dz'\,\nu(z',\mathbf{r})\,\frac{\gamma(z',\mathbf{r})}{\delta\nu(z_1,\mathbf{r}_1)}-k^2\,\delta(\mathbf{r}-\mathbf{r}_1)\,\delta(z-z_1)\,\gamma(z_1,\mathbf{r}_1).\end{aligned} \tag{22}$$

Because causality implies that $\delta/\delta\nu(z_1,\mathbf{r}_1)\equiv 0$ for $z_1>z$, we seek a boundary condition for $z_1\uparrow z$. In this limit we obtain

$$\frac{\delta\gamma(z,\mathbf{r})}{\delta\nu(z_1,\mathbf{r}_1)}=i\,\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}_1)\,\gamma(z_1,\mathbf{r}_1). \tag{23}$$

Averaging (23) we obtain the initial condition for the first derivative

$$\left\langle\frac{\delta\gamma(z,\mathbf{r})}{\delta\nu(z_1,\mathbf{r}_1)}\right\rangle=i\,\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}_1)\langle\gamma(z_1,\mathbf{r}_1)\rangle, \tag{24}$$

which is shown to depend on a lower element of the hierarchy. We use (23) to get the initial condition for higher elements by applying the variational derivative successively and averaging:

$$\begin{aligned}\langle \mathbf{D}^m({}^m\boldsymbol{\tau},{}^m\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle&\equiv\langle \mathbf{D}^m(z_1,\mathbf{r}_1;\ldots;z_m,\mathbf{r}_m)\,\gamma(z,\mathbf{r})\rangle\\&=i\,\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}_m)\langle \mathbf{D}^{m-1}({}^{m-1}\boldsymbol{\tau},{}^{m-1}\boldsymbol{\rho})\,\gamma(z,\mathbf{r}_m)\rangle,\end{aligned} \tag{25}$$

where $0\leqslant z_m\leqslant\cdots\leqslant z_2\leqslant z_1\leqslant z$.

C. δ-correlation closure

A δ-correlated process with respect only to z at the hierarchy-order $m=M$ immediately closes the hierarchical set. In this case, all the cumulants are δ-correlated in z, i.e.,

$$K_{\nu}^{(m)}({}^{m}\boldsymbol{\tau},{}^{m}\boldsymbol{\rho})=K_{\mathrm{eff}}^{(m)}(z,{}^{m}\boldsymbol{\rho})\,\delta(z_1-z_2)\cdots\delta(z_{m-1}-z_m). \tag{26}$$

The cumulant-generating functional in this case assumes the form

$$\begin{aligned}\Theta_z^{\mathrm{eff}}[w(z,\mathbf{r})]&=\sum_{m=1}^{\infty}\frac{i^m}{m!}\int_{-{}^{m}\boldsymbol{\infty}}^{{}^{m}\boldsymbol{\infty}}d^m\boldsymbol{\tau}'\int d^m\boldsymbol{\rho}'K_{\nu}^{(m)}({}^{m}\boldsymbol{\tau}',{}^{m}\boldsymbol{\rho})w(z_1',\mathbf{r}_1')\cdots w(z_m',\mathbf{r}_m')\\&\quad\times\theta(z_1)\cdots\theta(z_m)\,\theta(z-z_1)\cdots\theta(z-z_m)\\&=\sum_{m=1}^{\infty}\frac{i^m}{m!}\int_0^z dz'\int d^m\boldsymbol{\rho}'K_{\mathrm{eff}}^{(m)}(z',{}^{m}\boldsymbol{\rho}')w(z',\mathbf{r}_1')\cdots w(z',\mathbf{r}_m'),\end{aligned} \tag{27}$$

where $\pm{}^{m}\boldsymbol{\infty}$ is defined in (12), and we used the step function defined as $\theta(z<0)\equiv 0;\ \theta(z\geqslant 0)\equiv 1$.

In the case of the δ-correlated process, the closure formula admits a jump discontinuity as $\langle\nu(z'\to z,\mathbf{r})\,\gamma[\,\cdot\,]\rangle$. The result is[14]

$$\begin{aligned}\langle\nu(z,\mathbf{r})\,\gamma[\,\nu(\cdot)\,]\rangle&\equiv\left\langle\frac{1}{i(\delta/i\,\delta\nu_z(\mathbf{r}))}\,\frac{\partial}{\partial z}\,\Theta_z^{\mathrm{eff}}\left[\frac{\delta}{i\,\delta\nu_z(\mathbf{r})}\right]\gamma[\,\nu(\cdot)\,]\right\rangle\\&=\sum_{m=0}^{\infty}\frac{1}{(m+1)!}\int d^m\boldsymbol{\rho}'K_{\mathrm{eff}}^{(m+1)}(z,\mathbf{r};{}^{m}\boldsymbol{\rho}')\langle\mathbf{D}^m({}^{m}\mathbf{z},{}^{m}\boldsymbol{\rho}')\,\gamma(z,\mathbf{r})\rangle,\end{aligned} \tag{28}$$

where

$$\mathbf{D}^m({}^{m}\mathbf{z},{}^{m}\boldsymbol{\rho}')\,\gamma(z,\mathbf{r})\equiv\frac{\delta^m\gamma(z,\mathbf{r})}{\delta\nu_z(\mathbf{r}_1')\cdots\delta\nu_z(\mathbf{r}_m')}, \tag{29}$$

and we note that in (29), $\delta/\delta\nu_z(\mathbf{r}_n')$, $n\in\mathbb{N}$, means a *partial variational derivative* with a fixed z plane, i.e., $\delta\nu_z(\mathbf{r}_n')$ is a surface differing from zero only in the vicinity of $\mathbf{r}_n'\in\mathbb{R}^{2(N+M)}$.

Thus, the equation of the Mth-order successive approximation reads

$$\begin{aligned}[2ik\partial_z+\mathscr{L}_{\mathbf{r}}]\langle\mathbf{D}^M({}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle&=-k^2\sum_{m=0}^{\infty}\frac{1}{(m+1)!}\left(i\,\frac{k}{2}\right)^m\int d^m\boldsymbol{\rho}'\;K_{\mathrm{eff}}^{(m+1)}(z,\mathbf{r};{}^{m}\boldsymbol{\rho}')\\&\quad\times\langle\mathbf{D}^{m+M}({}^{m}\mathbf{z},{}^{m}\boldsymbol{\rho}',{}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle.\end{aligned} \tag{30}$$

Now we have to find $\langle\mathbf{D}^{m+M}({}^{m}\mathbf{z},{}^{m}\boldsymbol{\rho}',{}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$. Using (23) and letting ${}^{m}\boldsymbol{\tau}\uparrow{}^{m}\mathbf{z}$, we obtain

$$\langle\mathbf{D}^{m+M}({}^{m}\mathbf{z},{}^{m}\boldsymbol{\rho}',{}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle=\left(i\,\frac{k}{2}\right)^m\delta({}^{m}\mathbf{r}-{}^{m}\boldsymbol{\rho}')\langle\mathbf{D}^M({}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle. \tag{31}$$

Because $\langle\mathbf{D}^M({}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$ is known from the initial condition, we obtain

$$\begin{aligned}[2ik\partial_z+\mathscr{L}_{\mathbf{r}}]\langle\mathbf{D}^M({}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle&=-k^2\sum_{m=0}^{\infty}\frac{1}{(m+1)!}\left(i\,\frac{k}{2}\right)^m K_{\mathrm{eff}}^{(m+1)}(z,\mathbf{r};{}^{m}\mathbf{r})\\&\quad\times\langle\mathbf{D}^M({}^{M}\boldsymbol{\tau},{}^{M}\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle,\end{aligned} \tag{32}$$

where the integral over ${}^{m}\boldsymbol{\rho}'$ collapsed because of the $\delta({}^{m}\mathbf{r}-{}^{m}\boldsymbol{\rho}')$ in (31). In (32), the functional

$$\sum_{m=0}^{\infty} \frac{i^m}{(m+1)!} K_{\rm eff}^{(m+1)}(z,\mathbf{r};{}^m\mathbf{r})\beta^m(z)\Big|_{\beta(z)=k/2} = \frac{1}{i\beta(z)}\,\partial_z \ln \Phi_z^{\rm eff}[\beta(z')\delta(\mathbf{r}-\mathbf{r}')]\Big|_{\beta(z')=k/2}$$

$$\equiv \frac{1}{i\beta(z)}\,\partial_z \Theta_z^{\rm eff}[\beta(z')\delta(\mathbf{r}-\mathbf{r}')]\Big|_{\beta(z')=k/2}. \quad (33)$$

Here we remind the reader that for $w(z',\mathbf{r}')=\beta(z')\delta(\mathbf{r}-\mathbf{r}')$, the characteristic functionals of a process that is δ-correlated in z read

$$\Phi_z^{\rm eff}[\beta(z')\delta(\mathbf{r}-\mathbf{r}')] \equiv \left\langle \exp\left[i\int_0^z dz'\ \nu(z',\mathbf{r})\beta(z')\right]\right\rangle$$

$$\Theta_z^{\rm eff}[\beta(z')\delta(\mathbf{r}-\mathbf{r}')] \equiv \sum_{m=1}^{\infty}\ \frac{i^m}{m!}\ \int_0^z dz'\ K_{\rm eff}^{(m)}(z';{}^m\mathbf{r})\beta^m(z'). \quad (34)$$

Note that because $\beta(z)=k/2$ is a constant,

$$\frac{\delta}{i\,\delta\beta(z)} = \frac{1}{i\beta(z)}\,\partial_z = -i\,\frac{2}{k}\,\partial_z. \quad (35)$$

Accordingly, the effective cumulants are obtained from

$$K_{\rm eff}^{(m)}(z;{}^m\mathbf{r}) = \frac{\delta^m \Theta_z^{\rm eff}[\beta(z')\delta(\mathbf{r}-\mathbf{r}')]}{\delta\beta(z_1)\cdots\delta\beta(z_m)}\Big|_{\beta(z)=k/2} = \left(-i\,\frac{2}{k}\right)^m \partial_z^m \Theta_z^{\rm eff}\left[\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}')\right], \quad (36)$$

where ∂_z^m operates only on the limits of the integral in (34).

Thus, we finally obtain a set of M equations, $M-1$ of which are in (18) with a boundary condition in (25), and the Mth equation is

$$[2ik\partial_z + \mathscr{L}_{\mathbf{r}}]\langle \mathbf{D}^M({}^M\boldsymbol{\tau},{}^M\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle = \partial_z \Theta_z^{\rm eff}\left[\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}')\right]\langle \mathbf{D}^M({}^M\boldsymbol{\tau},{}^M\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle, \quad (37)$$

with the boundary condition

$$\langle \mathbf{D}^M({}^M\boldsymbol{\tau},{}^M\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle = i\,\frac{k}{2}\,\delta(\mathbf{r}-\mathbf{r}_1)\langle \mathbf{D}^{M-1}({}^{M-1}\boldsymbol{\tau},{}^{M-1}\boldsymbol{\rho})\,\gamma(z,\mathbf{r}_1)\rangle. \quad (38)$$

By truncating the order of the hierarchy at $m=M$, we also determine the number of cumulants involved in the approximation. Thus, although we do not need to approximate $\Theta_z^{\rm eff}$ in (37), we can decide upon its truncation order, i.e., $m=M$ according to the accuracy in comparison with $\Theta_z^{\rm eff}$. A schematic array, where the cumulants involved are truncated at $M=4$, is

$$\begin{bmatrix} K^{(1)} & K^{(2)} & K^{(3)} & K^{(4)} & \searrow_0 \\ K^{(1)} & K^{(2)} & K^{(3)} & \searrow_0 & \searrow_0 \\ K^{(1)} & K^{(2)} & \searrow_0 & \searrow_0 & \searrow_0 \\ K_{\rm eff}^{(1)} & K_{\rm eff}^{(2)} & K_{\rm eff}^{(3)} & K_{\rm eff}^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (39)$$

In (39), the rows represent the order of the hierarchy, the $\{\searrow_0\}$ indicates discarded cumulants and the ellipsis a continuation of a possibly truncated $\Theta_z^{\rm eff}$ of the δ-correlated process indicated by the effective cumulants defined in (36). The vertical dots beyond $M=4$ indicate that the governing equations are the same and, hence, can be ignored.

A truncation of $M=1$ leads readily to the moment equations for a non-Gaussian δ-correlated medium determined already by Klyatskin.[3,14]

D. Example

To exemplify the procedure and bring down to earth the multi-dimensional multi-vector notations let us analyze a simple example. We discard the operator $\mathscr{S}_{\mathbf{r}}$ because it does not alter the principle steps of the procedure and simplifies the explanation considerably. Consider the ordinary differential equation

$$\begin{aligned} &\frac{d\psi(z)}{dz} = i\,\frac{k}{2}\,\mu(z)\psi(z), \\ &\psi(0) = \psi_0, \quad 0 \leqslant z \leqslant Z. \end{aligned} \tag{40}$$

In (40), the medium is represented by $\mu(z)$, which is a zero-mean stationary non-Gaussian process. We would like to obtain an approximate closed set of integro-differential equations for its average $\langle\psi\rangle$, taking into account a finite correlation length. A zero-mean process has a vanishing first cumulant, i.e., $K^{(1)}(z)\equiv 0$. Additionally stationarity means that the cumulants $K^{(m+1)}(z,\tau_1',\ldots,\tau_m') = K^{(m+1)}(z-\tau_1',\ldots,z-\tau_m')$, and that $\partial_z K_{\text{eff}}^{(m)}(z)=0$, $\forall m$. Denoting the effective cumulants by the constants $\kappa_m \equiv K_{\text{eff}}^{(m)}$ we present the effective characteristic functional as

$$\Phi_z^{\text{eff}}\left[\frac{k}{2}\right] = \exp\left[z\sum_{m=1}^{\infty}\frac{1}{m!}\left(\frac{ik}{2}\right)^m \kappa_m\right] \equiv \exp\left(z\,\frac{d}{dz}\,\Theta_z^{\text{eff}}\left[\frac{k}{2}\right]\right). \tag{41}$$

Let us choose a truncation order of $M=3$ that leads to a hierarchy of three equations. The first equation reads

$$\begin{aligned} \frac{d\langle\psi(z)\rangle}{dz} &= i\,\frac{k}{2}\int_0^z d\tau_1\, K^{(2)}(z-\tau_1)\left\langle\frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle \\ &\quad + i\,\frac{k}{2}\,\frac{1}{2!}\int_0^z d\tau_1\int_0^z d\tau_2\, K^{(3)}(z-\tau_1,z-\tau_2)\left\langle\frac{\delta^2\psi(z)}{\delta\mu(\tau_1)\delta\mu(\tau_2)}\right\rangle, \end{aligned} \tag{42}$$

subject to the initial condition

$$\langle\psi(0)\rangle = \psi_0. \tag{43}$$

The second equation reads

$$\frac{d}{dz}\left\langle\frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle = i\,\frac{k}{2}\int_0^z d\tau_2\, K^{(2)}(z-\tau_2)\left\langle\frac{\delta^2\psi(z)}{\delta\mu(\tau_2)\delta\mu(\tau_1)}\right\rangle, \quad 0\leqslant\tau_1\leqslant z, \tag{44}$$

subject to the initial condition

$$\left\langle\frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle\Bigg|_{z=\tau_1} = i\,\frac{k}{2}\,\langle\psi(\tau_1)\rangle. \tag{45}$$

The third equation, on which we apply the δ-correlation ansatz, reads

$$\frac{d}{dz}\left\langle\frac{\delta^2\psi(z)}{\delta\mu(\tau_2)\delta\mu(\tau_1)}\right\rangle = \frac{d}{dz}\,\Theta_z^{\text{eff}}\left[\frac{k}{2}\right]\left\langle\frac{\delta^2\psi(z)}{\delta\mu(\tau_2)\delta\mu(\tau_1)}\right\rangle, \quad 0\leqslant\tau_2,\ \ \tau_1\leqslant z, \tag{46}$$

subject to the initial condition

$$\left\langle\frac{\delta^2\psi(z)}{\delta\mu(\tau_2)\delta\mu(\tau_1)}\right\rangle\Bigg|_{z=\tau_1} = i\,\frac{k}{2}\left\langle\frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle. \tag{47}$$

A Gaussian $\mu(z)$ in (42) implies $K^{(n)}(\cdot)\equiv 0$ for $n>2$, hence the obvious distinction between ours and the Gaussian case. Choosing $M=2$ leads to a *practically Gaussian* case, thus to show any

improvement of the proposed procedure, we need at least $M=3$. The distinction between our formulation and the non-Gaussian δ-correlation version becomes clear if we note that all the equations in the hierarchy of the δ-correlation formulation follow the same partial differential equation (37) and, in particular, (46), whereas our formulation follows this partial differential equation only from $M=3$ onward.

In general, the set of three equations must be solved numerically. Sometimes the last equation can be solved analytically. Then, this solution is inserted in the lower-order equations of the hierarchy. Often, the resulting integro-differential equations cannot be solved analytically.

In the present case we can solve the last equation, whose solution is

$$\left\langle \frac{\delta^2\psi(z)}{\delta\mu(\tau_1)\delta\mu(\tau_2)}\right\rangle = i\,\frac{k}{2}\exp\left[(z-\tau_1)\,\frac{d}{dz}\,\Theta_z^{\mathrm{eff}}\left[\frac{k}{2}\right]\right]\left\langle \frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle. \tag{48}$$

Inserting (48) into the second-order equation (44) gives

$$\frac{d}{dz}\left\langle \frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle = \left(i\,\frac{k}{2}\right)^2\,\frac{1}{2!}\int_0^z d\tau_2\,K^{(2)}(z-\tau_2)\,\exp\left[(z-\tau_1)\,\frac{d}{dz}\,\Theta_z^{\mathrm{eff}}\left[\frac{k}{2}\right]\right]\left\langle \frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle, \tag{49}$$

and inserting (49) into the first equation (42) gives

$$\begin{aligned}\frac{d\langle\psi(z)\rangle}{dz} &= i\,\frac{k}{2}\int_0^z d\tau_1\,K^{(2)}(z-\tau_1)\left\langle \frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle + i\left(\frac{k}{2}\right)^2\,\frac{1}{2!}\int_0^z d\tau_1\int_0^z d\tau_2\,K^{(3)}(z-\tau_1,z-\tau_2)\\ &\quad\times\exp\left[(z-\tau_1)\,\frac{d}{dz}\,\Theta_z^{\mathrm{eff}}\left[\frac{k}{2}\right]\right]\left\langle \frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle.\end{aligned} \tag{50}$$

Integrating both (49) and (50) with respect to z and incorporating the initial condition (43) and (45) the set of two coupled equations to be solved is

$$\begin{aligned}\langle\psi(z)\rangle &= \psi_0 + i\,\frac{k}{2}\int_0^z dz'\int_0^{z'} d\tau_1\,K^{(2)}(z'-\tau_1)\left\langle \frac{\delta\psi(z')}{\delta\mu(\tau_1)}\right\rangle\\ &\quad + i\left(\frac{k}{2}\right)^2\,\frac{1}{2!}\int_0^z dz'\int_0^{z'} d\tau_1\int_0^{z'} d\tau_2\,K^{(3)}(z'-\tau_1,z'-\tau_2)\\ &\quad\times\exp\left[(z-\tau_1)\,\frac{d}{dz}\,\Theta_z^{\mathrm{eff}}\left[\frac{k}{2}\right]\right]\left\langle \frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle,\end{aligned} \tag{51}$$

and

$$\begin{aligned}\left\langle \frac{\delta\psi(z)}{\delta\mu(\tau_1)}\right\rangle &= i\,\frac{k}{2}\,\langle\psi(\tau_1)\rangle + \left(i\,\frac{k}{2}\right)^3\,\frac{1}{2!}\int_0^z d\tau_1\int_0^z d\tau_2\,K^{(2)}(z-\tau_2)\\ &\quad\times\exp\left[(z-\tau_1)\,\frac{d}{dz}\,\Theta_z^{\mathrm{eff}}\left[\frac{k}{2}\right]\right]\left\langle \frac{\delta\psi(\tau_1)}{\delta\mu(\tau_2)}\right\rangle.\end{aligned} \tag{52}$$

The Gaussian case is obtained by setting $K^{(3)}(\cdot)=0$, and $d\Theta_z^{\mathrm{eff}}/dz=(ik/2)^2\kappa_2/2!$.

III. THE δ-CORRELATION LIMIT

The δ-correlated process does not occur in nature, thus it should be given a physical meaning. We invoked a δ-correlated process only in the last equation, whereas in the lower-order equations, the full cumulant is used. Other closure techniques involve the assumption that either a high-order cumulant (moment) depends on the lower order cumulants (moments), or at the $(n-1)$th level the nth cumulant (moment) vanishes or is asymptotically small. To assign a physical meaning to the effective cumulant defined in (26) and (27), we note that a δ-correlated process can be considered

the limit of the Ornstein–Uhlenbeck process,[5] i.e., white Gaussian noise. This process and its δ-correlated limit are both phenomenologically convenient forms to write the partial integral on the right-hand side of (18) as

$$\mathscr{I} \equiv \int_{{}^m\mathbf{0}}^{{}^m\mathbf{z}} d\,{}^m\boldsymbol{\tau}'\; K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')\langle \mathbf{D}^{m+n}({}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$$

$$\simeq \langle \mathbf{D}^{m+n}({}^m\mathbf{z},{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle \int_{{}^m\mathbf{0}}^{{}^m\mathbf{z}} d\,{}^m\boldsymbol{\tau}'\, K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}'), \tag{53}$$

where ${}^m\mathbf{0}$ and ${}^m\mathbf{z}$ are defined in (12), and the final step that follows (30) is immediately applicable. This means that $\langle \mathbf{D}^{m+n}(\cdot)\,\gamma(z,\mathbf{r})\rangle$ is supposed to have longer timelike scales than $K_\nu^{(m+1)}(\cdot)$. In this case the integro-differential formulation leads to a single partial differential equation. We term this case the *stochastic limit*. In fact, we expect in many cases that the timelike correlation length of $\langle \mathbf{D}^{m+n}(\cdot)(z,\mathbf{r})\rangle$ is of the order of the propagation length, whereas the scale of the $K_\nu^{(m+1)}(\cdot)$ is of the order of $L<(m+1)l$, where l is the largest correlation scale of the medium. Thus, $z\gg L$ is an intuitive definition of the δ-correlated approximation, and $l/z\to 0$ must hold uniformly for an arbitrary $m\leqslant M$.

We stress that in the *coherent limit*, which is opposite to the stochastic limit mentioned above, a partial differential equation cannot be obtained. In this case, and in the intermediate coherent-stochastic case, we have to use the proposed successive approximations to obtain a set of statistically closed, albeit integro-differential equations.

The closure procedure can be filled with content, if we define the characteristic scale to be

$$l_j(z,{}^m\mathbf{z},{}^m\boldsymbol{\rho}') \equiv \lim_{{}^m\boldsymbol{\tau}'\uparrow{}^m\mathbf{z}} \left| \frac{K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')}{\partial_{z_j'} K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')} \right|, \quad j=1,\ldots,m. \tag{54}$$

Then, we assume that $K_\nu^{(m+1)}(\cdot)$ attains its maximum as ${}^m\boldsymbol{\tau}\uparrow{}^m\mathbf{z}$, but this point is not a stationary point. We evaluate the integral $\mathscr{I}$ using a Laplace-type expansion in z/l_j and integrate the exponential integrand by parts;

$$\mathscr{I} \equiv \int_{{}^m\mathbf{0}}^{{}^m\mathbf{z}} d\,{}^m\boldsymbol{\tau}'\; K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}')\langle \mathbf{D}^{m+n}({}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle$$

$$\sim \langle \mathbf{D}^{m+n}({}^m\mathbf{z},{}^m\boldsymbol{\rho}',{}^n\boldsymbol{\tau},{}^n\boldsymbol{\rho})\,\gamma(z,\mathbf{r})\rangle K_\nu^{(m+1)}(z,\mathbf{r};{}^m\mathbf{z},{}^m\boldsymbol{\rho}') \prod_{j=1}^{m} l_j(z,{}^m\mathbf{z},{}^m\boldsymbol{\rho}')(1-e^{-z/l_j}), \tag{55}$$

where $z\gg l_j$ leads to the stochastic limit, i.e., the δ-correlated case. Comparing (55) and (26) and (27), the effective cumulant can be found if we note that $K_\nu^{(m+1)}(z,\mathbf{r};{}^m\mathbf{z},{}^m\boldsymbol{\rho}')$ is a function of z/l_j for $j=1,\ldots,m$. Thus,

$$K_{\mathrm{eff}}^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\rho}') = \int_{{}^m\mathbf{0}}^{{}^m\infty} d\,{}^m\boldsymbol{\tau}'\; K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}) \sim K_\nu^{(m+1)}(z,\mathbf{r};{}^m\mathbf{z},{}^m\boldsymbol{\rho}') \prod_{j=1}^{m} l_j(z,{}^m\mathbf{z},{}^m\boldsymbol{\rho}'). \tag{56}$$

Finally, if ${}^m\boldsymbol{\tau}={}^m\mathbf{z}$ happened to be a stationary point, then using a scale related to the square root of the normalized determinant of the Hessian of $K_\nu^{(m+1)}(\cdot)$ leads to a similar result with a Gaussian integral.

IV. DISCUSSION: A CONSISTENT δ-CORRELATION CLOSURE

Although outside the scope of this paper, predicting how a convergence proof might evolve, we note that the cumulant discard does not represent systematically the cumulants of our closure device, i.e., the δ-correlation model. Thus, we propose another closure procedure that is best summarized by the schematic array, where, for example, the δ-correlation assumption is applied at the hierarchy order $M=3$. In this case the array is

$$\begin{bmatrix} K^{(1)} & K^{(2)} & K^{(3)} & K_3^{(4)} & \cdots \\ K^{(1)} & K^{(2)} & K_3^{(3)} & K_3^{(4)} & \cdots \\ K_{\text{eff}}^{(1)} & K_{\text{eff}}^{(2)} & K_{\text{eff}}^{(3)} & K_{\text{eff}}^{(4)} & \cdots \\ K_{\text{eff}}^{(1)} & K_{\text{eff}}^{(2)} & K_{\text{eff}}^{(3)} & K_{\text{eff}}^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{57}$$

where effective cumulants $\{K_{\text{eff}}^{(m)}\}$ can be readily obtained by using (26) or (56), and the $\{K_M^{(m)}\}$ may be the effective cumulants or hybrids defined as the regular cumulants up to the order M and δ-correlated onward:

$$K_M^{(m+1)}(z,\mathbf{r};{}^M\boldsymbol{\tau},{}^m\boldsymbol{\rho}) \simeq \int_0^\infty dz'_{M+1} \cdots \int_0^\infty dz'_m \; K_\nu^{(m+1)}(z,\mathbf{r};{}^m\boldsymbol{\tau}',{}^m\boldsymbol{\rho}). \tag{58}$$

Regarding this infinite system, we note that once we applied a δ-correlation closure, hierarchy orders higher than M follow the same integro-differential equation. In the example above with $M=3$, this is expressed in (57) by identical rows three and four, etc. Finally, the hybrid cumulant $K_M^{(m)}$ seems to be the best option although its use must be justified. The complete representation of all the cumulants, albeit partly in an approximated effective form, seems to be required for a proof of convergence to a genuine solution, which we intend to explore in the future.

V. CONCLUSIONS

We presented here a hierarchy of integro-differential equations that can be used to obtain the governing equations for a general Cauchy problem with random parametric fluctuations. To the best of our knowledge, this is the most general formulation that has been presented in this context. As we mentioned previously, currently there is no proof of convergence of this procedure, although limiting cases are known to be correct. To this end, we suggested also an alternative closure procedure that seems to be required for a proof of convergence.

The formulation is rather complicated, but in a few cases, e.g., the second-statistical moment [$N=M=1$ in (2)], or when $\mathscr{L}_{\mathbf{r}}$ is a first-order differential operator, it is possible to obtain tractable equations. In the future, we intend to show these equations for simple physical cases that clarify the difference between the Gaussian and the non-Gaussian cases and to prove the convergence of the procedure to the genuine solution.

ACKNOWLEDGMENTS

This research was supported by the Air Force Materiel Command, USAF, Contract No. F19628-94-C-0137. Partial support was also provided by the Environmental Technology Laboratory/ERL/NOAA.

[1] V. I. Klyatskin and V. I. Tatarskii, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **14**, 1400 (1971).
[2] R. V. Bobrik, Teor. Mat. Fiz. **68**, 301 (1986).
[3] V. I. Klyatskin, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **12**, 1506 (1974).
[4] R. V. Bobryk, Waves Random Media **3**, 267 (1993).
[5] N. G. Van Kampen, Phys. Rep. **24C**, 171 (1976).
[6] N. N. Bogolyubov, in *Studies in Statistical Mechanics*, edited by J. De Boer and G. E. Uhlenbeck (North-Holland, Amsterdam, 1962), Vol. 1, p. 5.
[7] A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics: Mechanics of Turbulence* (MIT, Cambridge, MA, 1971), Vol. 1, p. 375.
[8] J. M. Wilczak, J. Atmos. Sci. **41**, 3537 (1984).
[9] T. Kikuchi and O. Chiba, Boundary-Layer Meteorol. **31**, 51 (1985).
[10] J. Gozani, Opt. Lett. **17**, 559 (1992).
[11] J. Gozani, J. Electromagn. Waves Appl. **11**, 409 (1997).
[12] J. Gozani, Phys. Rev. E **53**, 6486 (1996).
[13] M. I. Charnotskii, J. Gozani, V. I. Tatarskii, and V. U. Zavorotny, in *Progress in Optics*, edited by E. Wolf (Elsevier Science, Amsterdam, 1993), Vol. XXXII, p. 203.
[14] V. I. Klyatskin, *Stochastic Equations and Waves in Randomly-Inhomogeneous Media* (Moscow, Nauka, 1980) (in Russian).

Maximally even sets and the devil's-staircase phase diagram for the one-dimensional Ising antiferromagnet with arbitrary-range interaction

Richard J. Krantz
Department of Physics, Metropolitan State College of Denver, Denver, Colorado 80217-3362

Jack Douthett
Department of Arts and Science, ATVI Community College, Albuquerque, New Mexico 87106

Steven D. Doty
Department of Physics, Metropolitan State College of Denver, Denver, Colorado 80217-3362

(Received 9 June 1997; accepted for publication 1 May 1998)

By extending the formalism of maximally even sets to the one-dimensional antiferromagnetic Ising model with arbitrary-range interaction in an applied magnetic field it is shown that the system exhibits a devil's-staircase magnetic phase diagram. A ''generalized'' similarity dimension (a generalization for nonuniform fractals) of the phase diagram is calculated and compared to previous results. It is shown that previous estimates of the dimension of the phase diagram for this system differs from this ''generalized'' similarity dimension. © *1998 American Institute of Physics.* [S0022-2488(98)00509-X]

I. INTRODUCTION

The role the one-dimensional Ising model plays in the analysis of various physical systems has been discussed previously by Bak and Bruinsma.[1,2] They show that, in the infinite limit, the magnetic phase diagram for a one-dimensional array of Ising spins with a $1/x^{\alpha}$ $(\alpha>1)$ antiferromagnetic interaction is a complete devil's staircase with an estimated ''average'' fractal dimension of $2/(1+\alpha)$.

Recently, Douthett and Krantz[3] investigated the one-dimensional antiferromagnetic spin-1/2 Ising model at zero temperature and zero applied field using the formalism of maximally/minimally even sets.[4,5] Assuming periodic boundary conditions, they show that the configurational energy, for a given density of down (up) spins, $\rho=N_-/N(\rho=N_+/N)$ where N is the number of lattice sites and $N_-(N_+)$ are the numbers of down (up) spins, interacting via a strictly convex interaction is minimized by a maximally even distribution of up and down spins where the maximally even distribution is given by the following algorithm.[4]

> Let i be a fixed integer such that $0\leqslant i\leqslant N-1$, and assign down-spins (up-spins) to sites $\lfloor (Nk+i)/N_{\pm}\rfloor$ where $k=0,1,...,N_{\pm}-1$. Assign up-spins (down-spins) to the remaining sites. Then each down- and up-spin spectrum consists of one or two consecutive integers. Conversely, if each down-spin (up-spin) spectrum consists of one or two consecutive integers, then there exists an i, $0\leqslant i\leqslant N-1$, such that the down-spins (up-spins) occupy the sites $\lfloor (Nk+i)/N_{\pm}\rfloor$, $k=0,1,...,N_{\pm}-1$, and the up-spins (down-spins) occupy the rest.

A restricted form of this algorithm has been discussed by Clough and Myerson.[6,7] The kth spectrum of down (up) spins is defined to be the set of distinct, clockwise distances between down (up) spin pairs that have $k-1$ down (up) spins between them. Previously, Hubbard[8] observed that, for a given down-spin density, the minimum configurational energy occurs when each spectrum consists of one or two consecutive integers. Douthett and Krantz[3] proceeded to apply their formalism to a discrete system of one-dimensional spin-1/2 Ising spins with an exponentially de-

 © 1998 American Institute of Physics

creasing antiferromagnetic pairwise interaction. Then, they extended the results to describe the configurational energy in the limit of large N. In this paper the theory is generalized to include the effects of an applied magnetic field.

II. FIELD DEPENDENCE

We start with the antiferromagnetic Hamiltonian:

$$H_{\text{anti}}=\sum_{\substack{i,j=0\\ i\neq j}}^{N-1} J(|i-j|)\sigma_i\sigma_j-\sum_{i=0}^{N-1}\sigma_i h, \tag{1}$$

where $J(|i-j|)$ is the absolute value of the ''reduced'' pairwise interaction (where we have absorbed the usual constant factors) between the spins at lattice sites i and j, the σ's represent the orientations of the spins $(+1$ or $-1)$ on the ith and jth sites relative to the axis defined by the reduced applied field, h, and N is the number of lattice sites. As shown in Ref. 3 the configurational energy for a given fraction of down spins $(\rho=N_-/N)$ is minimized by a maximally even distribution of up and down spins where the ρ's are given by ratios in the Farey series. The minimum energy per spin is given by

$$U_{\min}(\rho)=\begin{cases}\displaystyle\sum_{k=1}^{\infty} J(k) & \text{if } \rho=0\\ \displaystyle\sum_{k=0}^{\infty} f_k(\rho) & \text{otherwise,}\end{cases} \tag{2a}$$

where

$$f_k(\rho)=\begin{cases}(1-4\rho)\displaystyle\sum_{j=1}^{\infty} J(j) & \text{if } k=0\\ 4\rho[(1-\{k/\rho\})J(\lfloor k/\rho\rfloor)+\{k/\rho\}J(\lceil k/\rho\rceil)] & \text{otherwise}\end{cases} \tag{2b}$$

and $\{x\}$ represents the fractional part of x, $\lfloor x\rfloor$ is the floor function of x, and $\lceil x\rceil$ is the ceiling function of x. For a given n, the sum $\Sigma_{k=0}^{n} f_k(\rho)$ calculates the energy for the $\lceil n/\rho\rceil$th near neighbors $(\rho\leqslant 1/2)$. This function is strictly convex, differentiable almost everywhere (ae), and is symmetric about $\rho=1/2$ on the interval $[0,1]$. Note that $\rho=0$ corresponds to all up spins (no down spins) and $\rho=1$ corresponds to all down spins. In an applied field, h, the average energy per spin as a function of the down-spin density, ρ, is

$$U_h(\rho)=U_{\min}(\rho)-(1-2\rho)h. \tag{3}$$

In order to describe the phase diagram, the values of ρ that minimize $U_h(\rho)$ are plotted versus h. The result is the so-called devil's-staircase phase diagram which will be described in more detail below. Similar results for this model have been obtained by other means (see Aubry[9,10] and Ishimura[11]).

It is possible to find the left and right end points (h_l and h_r) of the field intervals for a given rational fraction, ρ, by taking the right and left derivatives of $U_h(\rho)$, setting each to zero, and solving for the fields. Then for $0<\rho<1$,

$$h_l=-(1/2)D^+U_{\min}(\rho)=2\left[\sum_{k=1}^{\infty} J(k)-\sum_{k=1}^{\infty} g(\lceil k/\rho\rceil-1)\right], \tag{4a}$$

$$h_r=-(1/2)D^-U_{\min}(\rho)=2\left[\sum_{k=1}^{\infty} J(k)-\sum_{k=1}^{\infty} g(\lfloor k/\rho\rfloor)\right], \tag{4b}$$

where $g(m)=(m+1)J(m)-mJ(m+1)$ and the symbols D^- and D^+ represent the left- and right-hand derivatives, respectively.

Letting $\rho = p/q$, where p and q are coprime, the second sums in each of the field terms may be rearranged to express the field end points for a given p and q as

$$h_l(p/q) = 2\left[\sum_{k=1}^{\infty} J(k) - \sum_{j=1}^{p}\sum_{k=1}^{\infty} g(q(k-1)+\lceil qj/p\rceil - 1)\right] \tag{5a}$$

and

$$h_r(p/q) = 2\left[\sum_{k=1}^{\infty} J(k) - \sum_{j=1}^{p}\sum_{k=1}^{\infty} g(q(k-1)+\lfloor qj/p\rfloor)\right]. \tag{5b}$$

Taking the difference between $h_l(p/q)$ and $h_r(p/q)$ gives the length of the field step in the devil's-staircase phase diagram:

$$\Delta h(q) = h_r(p/q) - h_l(p/q) = 2q\sum_{k=1}^{\infty} [k(J(qk-1) - 2J(qk) + J(qk+1))]. \tag{6}$$

Note that Δh depends only on the denominator q. This expression is equivalent to Eq. (5) of Ref. 1 which was arrived at by calculating the configurational energy cost to flip one down-spin up, rearranging the spins in a new minimum configuration, and solving for one field end point when the energy cost was equal to zero, indicating that this configuration was a stable state. Similarly, the other field end point was determined by calculating the energy cost to flip one up-spin down, rearranging the spins, setting the energy change to zero, and solving for the field.

III. EXPONENTIAL INTERACTION

We have chosen an antiferromagnetic interaction of the form:

$$\mathscr{J}(k) = -J_0 \exp(-\alpha k), \tag{7}$$

where α and J_0 are positive constants. Then, $J(k) = J_0 e^{-\alpha k}$ where, $k = 1,2,3,...$ denotes near-neighbor, next-near-neighbor, next-next-near-neighbor pairs, etc., respectively.[3] Shown in Fig. 1 is the reduced configurational energy per spin ($U_h/J_0 h_T$) versus the down-spin density, ρ, for several (reduced) applied fields ($h/J_0 h_T$) assuming an exponentially decreasing antiferromagnetic interaction with $\alpha = 0.5$. In the denominator of the reduced quantities, h_T is the field at which a down-spin density of zero occurs. As the applied field is increased the down-spin density that minimizes the configurational energy decreases as more spins align along the field. The reduced applied field ranges from zero, where the minimum configurational energy occurs when $\rho = 0.5$ and the up and down spins alternate on the lattice, to unity, where the minimum configurational energy occurs when $\rho = 0.0$ and all the spins are up.

For convenience, the configurational energy, in Fig. 1, has been calculated for fractional densities that occur in the Farey series $F_8 = \{0,1/8,1/7,1/6,1/5,1/4,2/7,1/3,3/8,2/5,3/7,1/2\}$. The origin of the devil's-staircase phase diagram becomes clearer when one examines the details of Fig. 1 (see the insert). For each rational ρ, $U_h(\rho)$ has a corner with different right and left derivatives. For a given ρ, the right derivative is zero when the field is h_l. As the field is increased that ρ minimizes U_h until the field reaches h_r at which point the left derivative is zero. This occurs at every rational density. For example, for a reduced field near 0.2 a down-spin density of 2/5 minimizes the configurational energy (see Fig. 1). As the field is increased, a down-spin density of 2/5 still minimizes the energy until the configurational energy is decreased enough such that the next (smallest) down-spin density minimizes the configurational energy. For the applied fields in between, the down-spin density of 2/5 minimizes the configurational energy and a step in the phase diagram occurs.

A similar behavior is seen in the insert of Fig. 1 for a value of the down-spin density near 1/3. This type of behavior occurs in the limit of large N as well, yielding a devil's-staircase phase diagram with the field width of each step given by Eq. (6). The explicit form of the field width at each step for our example interaction is given below.

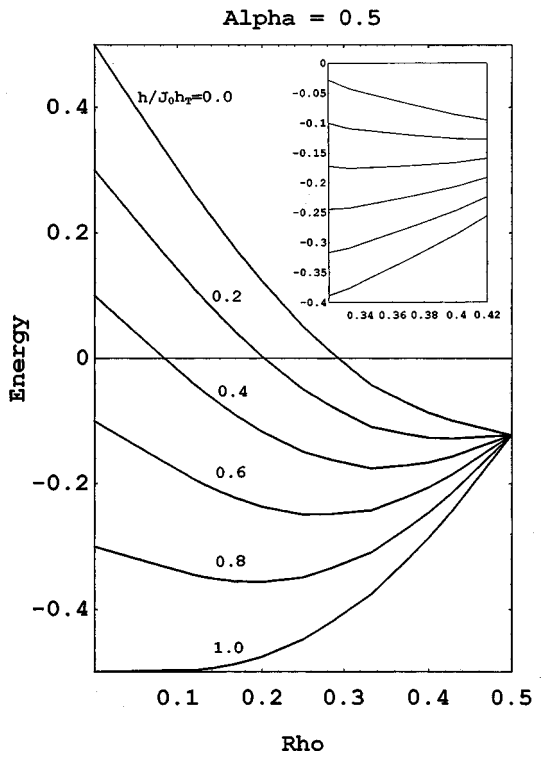

FIG. 1. Reduced configurational energy (U_h/J_0h_T) vs down-spin density (ρ) for various reduced applied fields (h/J_0h_T). The insert shows a portion of the plot for down-spin densities between 0.32 and 0.42. See the text for details.

For the exponentially decreasing antiferromagnetic interaction the equations that describe the applied field end points [Eqs. (5a) and (5b)] may be simplified:

$$h_l=\begin{cases} 2J_0 r/(1-r) & \text{if } \rho=0 \\ 2J_0\left[r/(1-r)-\sum_{k=1}^{p} G(q,\lceil qk/p\rceil-1)\right] & \text{otherwise} \end{cases} \tag{8a}$$

and

$$h_r=\begin{cases} -2J_0 r/(1-r) & \text{if } \rho=0 \\ 2J_0\left[r/(1-r)-\sum_{k=1}^{p} G(q,\lfloor qk/p\rfloor)\right] & \text{otherwise,} \end{cases} \tag{8b}$$

where $r=e^{-\alpha}$ and $G(q,m)=[g(m)-r^{2q}g(m-q)]/(1-r^q)^2$.

The applied field step width as a function of q, only, becomes

$$\Delta h(q)=2qr^{q-1}(1-r)^2/(1-r^q)^2, \quad q\geqslant 2. \tag{9}$$

Thus, for exponential interactions, the infinite sums in Eqs. (5a), (5b), and (6) can be expressed in closed form.

IV. THE DEVIL'S STAIRCASE

Shown in Fig. 2, is the resulting devil's-staircase magnetic phase diagram for the exponentially decreasing antiferromagnetic interaction with $\alpha=0.5$. As described in Eqs. (6) and (9), the applied field step widths depend only on the denominator of the rational fraction of the appropriate down-spin density. As expected, the phase diagram shows symmetry about zero applied field and a down-spin density of one-half. Equivalently, the system can be described either by down-spin or up-spin density versus applied field.

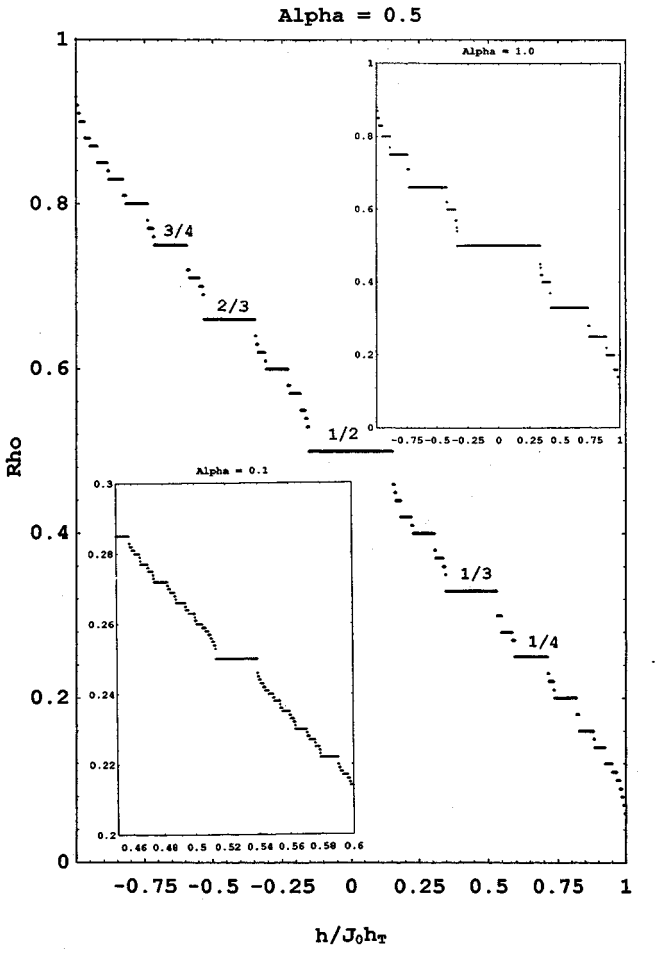

FIG. 2. Devil's-staircase phase diagram, Down-spin density (ρ) vs reduced applied field (h/J_0h_T) for the exponentially decreasing antiferromagnetic interaction with $\alpha=0.5$. [Upper Insert: Phase diagram with $\alpha=1.0$. Lower Insert: Phase diagram with $\alpha=0.1$.]

For comparison, the upper insert of Fig. 2 shows the resulting devil's staircase for a shorter range interaction, $\alpha=1.0$. The reduced field step widths for a given q are wider, relatively, as h_T, the field at which a down-spin density of zero occurs $[h_T=2e^{-\alpha}/(1-e^{-\alpha})]$, decreases with increasing α. As α increases further, i.e., the interaction becomes of shorter and shorter range and the spins become more and more independent, the $\rho=1/2$ reduced field step width approaches 2 (from -1 at the left-hand side to $+1$ at the right-hand side). In absolute terms the field interval from $-h_T$ to h_T is getting smaller. At these end points the density switches discontinuously to 0 or 1 depending on the direction of the applied field. For further comparison, the devil's-staircase phase diagram for longer ranged interactions, $\alpha=0.1$ is shown in the lower insert of Fig. 2, where for convenience, we have shown only a limited portion of the phase diagram.

V. DIMENSIONS

Previously, Bruinsma and Bak[2] estimated the ''average'' fractal dimension of a devil's-staircase magnetic phase diagram for a one-dimensional antiferromagnetic Ising model with a $1/x^{\alpha}$ interaction. They find a dimension by counting the number of phase transitions, $N(q)$, that one could measure with a precision of $\Delta h(q)$ and use

$$D_F=\lim_{q\to\infty}\ln(N(q))/\ln(H_T/\Delta h(q)),$$

TABLE I. ''Generalized'' similarity dimension of the devil's-staircase phase diagram for an exponentially decreasing antiferromagnetic interaction with different ranges, α, in the large q limit.

$J=J_0\exp(-\alpha k)$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.50$
''Generalized'' similarity dimension	0.3983	0.3255	0.2849	0.1851

where H_T is the range of the applied field. Because the allowed densities are rational fractions contained in the Farey series, the number of phases that will occur, $N(q)$, is just equal to the sum of Euler phi functions up to q, $\Phi(q)$. In the limit of large q this sum behaves like $3q^2/\pi^2$.[12] For a $1/x^\alpha$ interaction the ratio $H_T/\Delta h(q)$ can be shown to approach $q^{1+\alpha}$ in the limit of large q. Therefore, Bruinsma and Bak[2] estimate the fractal dimension as $2/(1+\alpha)$ in the limit of large q. For our exponential interaction D_F is zero for all $\alpha>0$.

The simple form of this dimension is appealing, however, this dimension is nonstandard and as such cannot be compared to fractal dimensions in general. In particular, it is easy to show that D_F differs from the capacity/box dimension for two-scale Cantor sets. Let Δ_1 and Δ_2 be the ratios to scale down the lengths to construct the two-scale Cantor set. If for the two scale ratios, Δ_1 and Δ_2, $\Delta_1<\Delta_2$ then D_F is dominated by the smaller scale factor:

$$D_F=-\log 2/\log \Delta_1. \tag{10}$$

On the other hand, the capacity dimension for this two-scale Cantor set is the solution to

$$\Delta_1^{D_c}+\Delta_2^{D_c}=1. \tag{11}$$

These two dimensions are equal if and only if $\Delta_1=\Delta_2$.

Jackson[13] generalizes Eq. (11) for an n-scale Cantor set. In the case of a Farey series Cantor set, there are an infinite number of scales; so we calculate the similarity dimension for nonuniform fractals (a generalization of the capacity dimension for nonuniform cantor sets) as described by Hilborn.[14] This generalization accounts for the changing scale factor at each level. For the purposes of our calculation this reduces to numerically finding the similarity dimension for nonuniform fractals, D_s, thus:

$$\sum_{k=1}^{\Phi(q)}(\Delta L_q(k))^{D_s}=\left(4\sum_{k=1}^{\infty}J(k)\right)^{D_s}, \tag{12}$$

where $\Delta L_q(k)$ denote the difference between the left and right field end points of subsequent stable phases, i.e., the ΔL_q's represent the ''gaps'' in the phase diagram. The fact that the ΔL_q's are different at each level accounts for the changing scale at each level mentioned above for nonuniform fractals—to which this ''generalized'' similarity dimension applies, as pointed out by Hillborn.[14] For exponential interactions the right-hand side of Eq. (12) becomes $(4J_0r/(1-r))^{D_s}$. The right-hand side of Eq. (12) is just $2h_T$, the length of the horizontal axis—from the field at which the spins are all up to the field at which the spins are all down.

Shown in Table I are the ''generalized'' similarity dimensions for the phase diagrams for exponentially decreasing antiferromagnetic interactions with varying ranges. As seen in this table the ''generalized'' fractal dimension depends on the range of the interaction. The longer the range, the larger the ''generalized'' similarity dimension. The approach of the value of the generalized similarity dimension in the large denominator limit is shown in Fig. 3 for the four values of the range, α, given in Table I. The horizontal axis is the largest denominator, q, occurring in the Farey series which generates the rational fractions representing the down-spin densities, ρ. For large q the number of allowable phases is $3\,q^2/\pi^2$ as mentioned above. This means that for a denominator of 3700 the distinct number of terms in Eq. (12) required to calculate the dimension is some 2.1 million. (If the phase diagram were not symmetric about $\rho=0$ such a calculation would require some 4.2 million terms.)

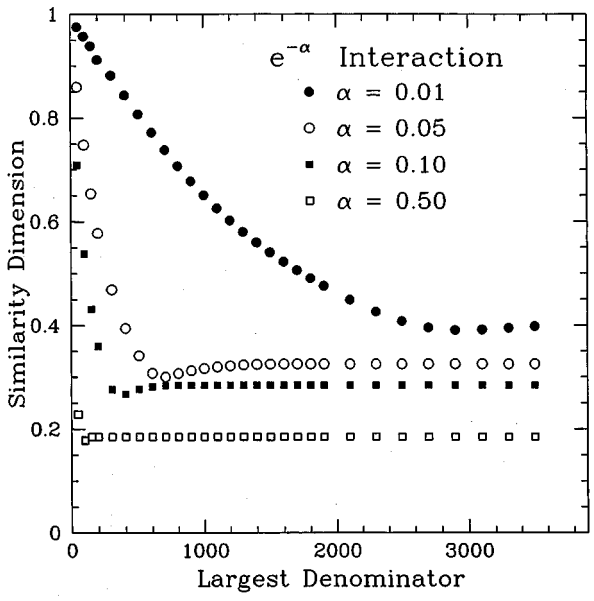

FIG. 3. ''Generalized'' similarity dimension as a function of the largest denominator in the Farey series.

Shown in Fig. 3 is the dependence of the generalized similarity dimension on q, which indicates that the fractal phase diagram is not self-similar at all levels, which is to be expected for nonuniform fractals.

Similar observations have been made by Bruinsma and Bak[2] with respect to their ''average'' dimension and the $1/x^2$ interaction. The ''average'' fractal dimension calculated by these researchers approaches $2/(1+\alpha)$, or 2/3 in this case, in the limit of large q, and approaches this limit from below as self similar subsets each with dimension $1/(1+\alpha)$, or 1/3 are combined.[2] In Table II we show the generalized similarity dimension calculated for the $1/x^2$ interaction. The generalized similarity dimension for this interaction approaches a limit of 0.645 which is different than the 2/3 ''average'' dimension calculated by Bruinsma and Bak.[2] As there is no formal connection between the average dimension of Bruinsma and Bak[2] and the generalized similarity dimension, we can only note that such a difference occurs.

VI. SUMMARY

The analysis of the one-dimensional antiferromagnetic spin- 1/2 Ising model with an arbitrary-range exponential interaction using the formalism of maximally even sets[3] has been extended to

TABLE II. ''Generalized'' similarity dimension of the devil's-staircase phase diagram for a $1/x^2$ antiferromagnetic interaction as q, the largest denominator in the Farey series (which determines the allowable down-spin densities) increases.

$J \propto 1/x^{\alpha}$	$\alpha = 2.0$
q	D_s
5	0.603
10	0.619
15	0.627
20	0.631
25	0.633
30	0.635
100	0.643
174	0.645
200	0.645

include the effect of an applied magnetic field. It is shown that the magnetic phase diagram is a complete devil's staircase in which the field step width can be expressed in closed form.

The resulting fractal is shown to be a nonuniform Cantor set. The generalized similarity dimension[14] for this case is calculated numerically and compared to a previous closed form estimate of the fractal dimension.[2] Our results suggest that the generalized similarity dimension for this system differs from the closed form estimate.

ACKNOWLEDGMENTS

Thanks to Colin O'Donnell for his help in generating Figs. 1 and 2. Thanks to the reviewer for suggesting earlier references of which we were unaware.

[1] P. Bak and R. Bruinsma, Phys. Rev. Lett. **49**, 249 (1982).
[2] R. Bruinsma and P. Bak, Phys. Rev. B **27**, 5824 (1983).
[3] J. Douthett and R. Krantz, J. Math. Phys. **37**, 3334 (1996).
[4] J. Clough and J. Douthett, J. Music Theor. **35**, 93 (1991).
[5] S. Block and J. Douthett, J. Music Theor. **38**, 21 (1994).
[6] J. Clough and G. Myerson, J. Music Theor. **29**, 249 (1985).
[7] J. Clough and G. Myerson, Am. Math. Monthly **93**, 695 (1986).
[8] J. Hubbard, Phys. Rev. B **17**, 494 (1978).
[9] S. Aubry, in *Solitons and Condensed Matter*, edited by A. R. Bishop and T. Schneider (Springer, Berlin, 1978), p. 264.
[10] S. Aubry, J. Phys. C **16**, 2497 (1983).
[11] N. Ishimura, J. Phys. Soc. Jpn. **54**, 4752 (1985).
[12] See, for example, R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation For Computer Science*, 2nd ed. (Addison–Wesley, New York, 1994), pp. 133–139.
[13] E. A. Jackson, *Perspectives of Nonlinear Dynamics* (Cambridge University Press, Cambridge, 1994), Vol. 1, pp. 72 and 187–190.
[14] R. Hilborn, *Chaos and Nonlinear Dynamics; An Introduction for Scientists and Engineers* (Oxford University Press, New York, 1994), pp. 394 and 401–405.

Analytic-bilinear approach to integrable hierarchies. I. Generalized KP hierarchy

L. V. Bogdanov[a] and B. G. Konopelchenko
Consortium EINSTEIN,[b] Dipartimento di Fisica dell'Università di Lecce and Sezione INFN, 73100 Lecce, Italy

(Received 9 June 1997; accepted for publication 18 May 1998)

An analytic-bilinear approach for construction and study of integrable hierarchies, in particular, the KP hierarchy is proposed. It starts with the generalized Hirota identity for the Cauchy–Baker–Akhiezer (CBA) function and leads to a generalized KP hierarchy in the form of compact functional equations containing a special shift operator. A generalized KP hierarchy incorporates the basic KP hierarchy, modified KP hierarchy, KP singularity manifold equation hierarchy, and corresponding hierarchies of linear problems. Different ''vertical'' levels of the generalized KP hierarchy are connected via invariants of the Combescure symmetry group. The resolution of functional equations also gives rise to the τ function and the addition formulas for it. © *1998 American Institute of Physics.* [S0022-2488(98)03209-5]

I. INTRODUCTION

The Kadomtsev–Petviashvili (KP) equation and the whole KP hierarchy of equations are significant parts of the theory of integrable equations. They arise in various fields of physics from hydrodynamics to string theory. They are also the tools to solve several problems in mathematics from the differential geometry of surfaces to an algebraic geometry.

The KP hierarchy has been described and studied within the framework of different approaches. The Sato approach[1] (see also Refs. 2–6) and the $\bar{\partial}$-dressing method[7–10] are, perhaps, the most beautiful and powerful among them. The infinite-dimensional Grassmannian, pseudodifferential operators, Hirota bilinear identity, and τ function are the basic ingredients of the Sato approach, which is, in essence, algebraic. In contrast, the $\bar{\partial}$ method, based on the nonlocal $\bar{\partial}$ problem for the wave function, mostly uses the analytic properties of the wave function. These two approaches look completely different. On the other hand, each of them has its own advantages. So one could expect that their marriage could be rather profitable.

A bridge between the Sato approach and the $\bar{\partial}$-dressing method has been established by the observation that the Hirota bilinear identity can be derived from the $\bar{\partial}$-dressing method.[10,11] Elements of the approach which combine the characteristic features of both methods, namely, the Hirota bilinear identity from the Sato approach and the analytic properties of solutions from the $\bar{\partial}$-dressing method, have been considered in Refs. 12, 13.

In this paper we are devoted to the analytic-bilinear approach to integrable hierarchies. It is based on the generalized Hirota bilinear identity for the wave function with simple analytic properties [the Cauchy–Baker–Akhiezer (CBA) function]. This approach allows us to derive generalized hierarchies of integrable equations in terms of the CBA function $\psi(\lambda,\mu,\mathbf{x})$ in a compact, finite form. Such generalized equations contain integrable equations in their usual form, their modified partners, and corresponding linear problems.

The generalized Hirota bilinear identity provides various functional equations for the CBA function. The resolution of some of them leads to the introduction of the τ function, while others give rise to the addition theorems for the τ function. The properties of the τ function, such as the associated closed 1-form and its global definition, also arise in a simple manner within the ap-

[a]Permanent address: L. D. Landau Institute for Theoretical Physics, Kosygin str. 2, Moscow 117940, GSP1, Russia; electronic mail: leonid@landau.ac.ru
[b]European Institute for Nonlinear Studies via Transnationally Extended Interchanges.

 © 1998 American Institute of Physics

proach under consideration. Generic discrete transformations of the CBA function and of the τ function are presented in the determinant form. These transformations are the generalized forms of the usual iterated Darboux transformations.

The generalized KP hierarchy is presented also in the ''moving'' frame, depending on the parameter. This generalized KP hierarchy written in the different ''moving'' frames contains the Darboux system of equations.

In addition to the usual infinite-dimensional symmetries, the generalized KP hierarchy possesses the symmetries given by the Combescure transformations. The invariants of these symmetry transformations have the compact forms in terms of the CBA function. The generalized KP hierarchy written in terms of these invariants coincide with the usual KP hierarchy, mKP hierarchy, and SM-KP hierarchy.

In this paper we make an emphasis of the concepts of generalized hierarchy and Combescure symmetry transformation invariants. We tried to derive and present a generalized KP hierarchy in a rather elementary manner.

The second part of the paper is devoted to the generalized multicomponent KP and two-dimensional Toda lattice hierarchies. In that paper we present more general and sophisticated forms of generalized integrable hierarchies.

The paper is organized as follows. The generalized Hirota identity is introduced in Sec. II. The generalized KP hierarchy is derived in Sec. III. Combescure symmetry transformations are discussed in Sec. IV. The generalized KP hierarchy in the ''moving'' frame is considered in Sec. V. The τ function is introduced in Sec. VI. The addition formulas for the τ function are also obtained here. Transformations of the CBA function and the τ function given by the determinant formulas and the Darboux transformations are discussed in Sec. VII. Closed one-form variational formulas for the τ function are presented in Sec. VIII.

II. GENERALIZED HIROTA IDENTITY

The famous Hirota bilinear identity provides us with a condensed and compact form of the integrable hierarchies. Here we will derive the generalized Hirota bilinear identity in the frame of the $\overline{\partial}$-dressing method.

The $\overline{\partial}$-dressing method (see Refs. 7–10) is based on the nonlocal $\overline{\partial}$ problem of the form

$$\overline{\partial}_\lambda(\chi(\lambda,\mathbf{x})-\eta(\lambda,\mathbf{x}))=\int\int_{\mathbb{C}} d\mu\wedge d\overline{\mu}\ \chi(\mu,\mathbf{x})g^{-1}(\mu,\mathbf{x})R(\mu,\lambda)g(\lambda,\mathbf{x}),$$
$$(\chi(\lambda,\mathbf{x})-\eta(\lambda,\mathbf{x}))_{|\lambda|\to\infty}\to 0, \tag{1}$$

where $\lambda\in\mathbb{C}$, $\mathbf{x}$ is some set of dynamical variables, $\overline{\partial}_\lambda=\partial/\partial\overline{\lambda}$, $\eta(\lambda,\mathbf{x})$ is a rational function of λ (normalization). In a general case, the function $\chi(\lambda,\mathbf{x})$ and the kernel $R(\lambda,\mu)$ are matrix-valued functions.

The dependence of the solution $\chi(\lambda,\mathbf{x})$ of the problem (1) on the dynamical variables is hidden in the function $g(\lambda,\mathbf{x})$. For the case of continuous variables, g is a product of the functions $g_i=\exp(K_i x_i)$, where $K_i(\lambda)$ are, in general, commuting matrix meromorphic functions.

It is assumed that the problem (1) is uniquely solvable. The $\overline{\partial}$-dressing method allows us to construct and solve wide classes of nonlinear partial differential equations (PDEs) which correspond to the different choice of the functions $K_i(\lambda)$.

Here we will assume that the kernel $R(\lambda,\mu)$ is equal to zero in some open subset G of the complex plane with respect to λ and to μ. This subset should typically include all zeros and poles of the considered class of functions $g(\lambda)$ and a neighborhood of infinity.

In this case the solution of the problem (1) normalized by η is the function

$$\chi(\mathbf{x},\lambda)=\eta(\mathbf{x},\lambda)+\varphi(\mathbf{x},\lambda),$$

where $\eta(\lambda)$ is a rational function of λ (normalization), all poles of $\eta(\lambda)$ belong to G, and $\varphi(\lambda)$ decreases as $\lambda\to\infty$ and is *analytic* in G.

The special class of solutions of the $\bar{\partial}$ problem (1) normalized by $(\lambda-\mu)^{-1}$ [$\eta(\lambda)=(\lambda-\mu)^{-1}$, $\mu\in G$] is of particular importance for the whole $\bar{\partial}$-dressing method.[8] Let us consider the $\bar{\partial}$ problem (1) for such solutions and the corresponding dual problem for the dual function $\chi^*(\lambda,\mu,\mathbf{x}')$:

$$\bar{\partial}_\lambda\chi(\lambda,\mu;g_1)=2\pi i\delta(\lambda-\mu)+\int\int_{\mathbb{C}} d\nu\wedge d\bar{\nu}\,\chi(\nu,\mu;g_1)g_1(\nu)R(\nu,\lambda)g_1(\lambda)^{-1},$$
$$\bar{\partial}_\lambda\chi^*(\lambda,\mu;g_2)=-2\pi i\delta(\lambda-\mu)-\int\int_{\mathbb{C}} d\nu\wedge d\bar{\nu}\,g_2(\nu)R(\lambda,\nu)g_2(\lambda)^{-1}\chi^*(\nu,\mu;g_2), \tag{2}$$

where $g_1=g(\lambda,\mathbf{x})$, $g_2=g(\lambda,\mathbf{x}')$, functions $\chi(\lambda,\mu;g_1)$ and $\chi^*(\lambda,\mu;g_2)$ are treated as functionals of the functions g_1, g_2, or, if the parametrization of g_1 and g_2 in terms of the variables $\mathbf{x}$, $\mathbf{x}'$ is given, just as functions of the variables $\mathbf{x}$, $\mathbf{x}'$.

After simple calculations, one gets

$$\int\int_G d\nu\wedge d\bar{\nu}\,\frac{\partial}{\partial\nu}(\chi(\nu,\lambda;g_1)g_1(\nu)g_2(\nu)^{-1}\chi^*(\nu,\mu;g_2))$$
$$=\int_{\partial G}\chi(\nu,\lambda;g_1)g_1(\nu)g_2(\nu)^{-1}\chi^*(\nu,\mu;g_2)d\nu=0. \tag{3}$$

In the particular case $g_1=g_2$ from (3), it follows that in $\bar{G}$ the function $\chi(\lambda,\mu)$ is equal to $-\chi^*(\mu,\lambda)$ (see also Ref. 14). So, finally, we have

$$\int_{\partial G}\chi(\nu,\mu;g_1)g_1(\nu)g_2^{-1}(\nu)\chi(\lambda,\nu;g_2)d\nu=0. \tag{4}$$

Here the function $\chi(\lambda,\mu)$ possesses the following analytical properties:

$$\bar{\partial}_\lambda\chi(\lambda,\mu)=2\pi\mathrm{i}\delta(\lambda-\mu),\quad -\bar{\partial}_\mu\chi(\lambda,\mu)=2\pi\mathrm{i}\delta(\lambda-\mu),$$

where $\delta(\lambda-\mu)$ is a δ function, or, in other words, $\chi\to(\lambda-\mu)^{-1}$ as $\lambda\to\mu$ and $\chi(\lambda,\mu)$ is an analytic function of both variables λ,μ for $\lambda\neq\mu$.

In the particular case $\lambda=\mu=0$, the relation (4) is nothing but the usual Hirota identity for the KP wave function $\chi(\nu,0,\mathbf{x})$ and the dual KP wave function $\chi^*(\nu,0,\mathbf{x}')$ (see, e.g., Refs. 1–4). Just in this case (for $\lambda=\mu=\infty$), the Hirota bilinear identity has been derived from the $\bar{\partial}$ problem in Ref. 11.

The identity (4) represents itself as the generalization of the Hirota identity which is bilocal with respect to the dynamical variables $\mathbf{x}$ and the spectral variables (λ,μ). It can be considered as the point of departure without any reference to the $\bar{\partial}$-dressing method. Namely, the generalized Hirota bilinear relation (4) is the starting point of the analytic-bilinear approach which we will consider in this paper. The double bilocality (with respect to $\mathbf{x}$ and (λ,μ)) provides us an additional freedom which will allow us to represent integrable hierarchies in a unified and condensed form.

Introducing the function $\psi(\lambda,\mu,\mathbf{x})$ via

$$\psi(\lambda,\mu;g)=g^{-1}(\mu)\chi(\lambda,\mu;g)g(\lambda), \tag{5}$$

one gets another form of the generalized Hirota equation,

$$\int_{\partial G}\psi(\nu,\mu;g_1)\psi(\lambda,\nu;g_2)d\nu=0. \tag{6}$$

Note that in the framework of the algebro-geometric technique, the function $\psi(\lambda,\mu)$ corresponds to the Cauchy–Baker–Akhiezer kernel on the Riemann surface (see Ref. 15). We will refer to the function $\psi(\lambda,\mu)$ as the Cauchy–Baker–Akhiezer (CBA) function.

We will assume in what follows that we are able to find solutions for the relation (4) somehow. In particular, it can be done by the $\bar{\partial}$-dressing method.

The general setting of the problem of solving (4) requires some modification of the Segal–Wilson Grassmannian approach.[4] Let us consider two linear spaces, $W(g)$ and $\tilde{W}(g)$, defined by the function $\chi(\lambda,\mu)$ [satisfying (4)] via equations connected with Eq. (4),

$$\int_{\partial G} f(\nu;g)\chi(\lambda,\nu;g)d\nu=0, \tag{7}$$

$$\int_{\partial G} \chi(\nu,\mu;g)h(\nu;g)d\nu=0; \tag{8}$$

here $f(\lambda)\in W$, $h(\lambda)\in\tilde{W}$; $f(\lambda)$, $h(\lambda)$ are defined in $\bar{G}$.

It follows from the definition of linear spaces $W,\tilde{W}$ that

$$\begin{aligned} f(\lambda)&=2\pi\mathrm{i}\int\int_G \eta(\nu)\chi(\lambda,\nu)d\nu\wedge d\bar{\nu}, \quad \eta(\nu)=\left(\frac{\partial}{\partial\bar{\nu}}f(\nu)\right),\\ h(\mu)&=-2\pi\mathrm{i}\int\int_G \chi(\nu,\mu)\tilde{\eta}(\nu)d\nu\wedge d\bar{\nu}, \quad \tilde{\eta}(\nu)=\left(\frac{\partial}{\partial\bar{\nu}}h(\nu)\right). \end{aligned} \tag{9}$$

These formulas, in some sense, provide an expansion of the functions f,h in terms of the basic function $\chi(\lambda,\mu)$. The formulas (9) readily imply that linear spaces $W,\tilde{W}$ are transversal to the space of holomorphic functions in G (the transversality property).

From the other point of view, these formulas define a map of the space of functions (distributions) on $\bar{G}$ $\eta,\tilde{\eta}$ to the spaces $W,\tilde{W}$. We will call $\eta(\tilde{\eta})$ a *normalization* of the corresponding function belonging to $W(\tilde{W})$.

The dynamics of linear spaces $W,\tilde{W}$ looks very simple:

$$W(g)=W_0g^{-1}; \quad \tilde{W}(g)=g\tilde{W}_0; \tag{10}$$

here $W_0=W(g=1)$, $\tilde{W}_0=\tilde{W}(g=1)$ [the formulas (10) follow from identity (4) and the formulas (9)].

To introduce a dependence on several variables (maybe of a different type), one should consider a product of corresponding functions $g(\lambda)$ (all of them commute).

The formula (4) is a basic tool for our construction. Analytic properties of the CBA kernel accompanied with the different choices of the functions g_1 and g_2 will provide us with various compact and useful relations.

III. GENERALIZED KP HIERARCHY

In the present paper we will consider the scalar KP hierarchy. It is generated by the generalized Hirota formula (4) where G is a unit disk and

$$g(\lambda,\mathbf{x})=\exp\left(\sum_{i=1}^{\infty} x_i\lambda^{-i}\right). \tag{11}$$

Let us consider the formula (4) with

$$g_1g_2^{-1}=g(\nu,\mathbf{x})g^{-1}(\nu,\mathbf{x}')=\exp\left(\sum_{i=1}^{\infty}(x_i-x_i')\nu^{-i}\right)=\frac{\nu-a}{\nu-b}, \tag{12}$$

where a and b are arbitrary complex parameters, $a,b\in G$. Since $\ln(1-\epsilon)=-\sum_{i=1}^{\infty}\epsilon^i/i$, one has

$$x_i' - x_i = \frac{1}{i}a^i - \frac{1}{i}b^i.$$

Substituting the expression (12) into (4), one gets

$$\left(\frac{\mu-a}{\mu-b}\right)\chi(\lambda,\mu,\mathbf{x}+[a]) - \left(\frac{\lambda-a}{\lambda-b}\right)\chi(\lambda,\mu,\mathbf{x}+[b])$$
$$+(b-a)\chi(\lambda,b,\mathbf{x}+[a])\chi(b,\mu,\mathbf{x}+[b]) = 0, \quad \lambda \neq \mu, \tag{13}$$

where $\mathbf{x}+[a]=x_i+[a]_i$, $0 \leqslant i < \infty$, $[a]_i=(1/i)a^i$. Equation (13) is the simplest functional equation for $\chi(\lambda,\mu,\mathbf{x})$ which follows from the generalized Hirota equation (4).

Residues of the LHS of (13) at the poles $\mu=b$ and $\lambda=b$ vanish identically. Evaluating the LHS of (13) at $\mu=a$ and $\lambda=a$, one gets the equations

$$\left(\frac{\lambda-a}{\lambda-b}\right)\chi(\lambda,a,\mathbf{x}+[b]) = (b-a)\chi(\lambda,b,\mathbf{x}+[a])\chi(b,a,\mathbf{x}+[b]), \tag{14}$$

$$\left(\frac{\mu-a}{\mu-b}\right)\chi(a,\mu,\mathbf{x}+[a]) = (a-b)\chi(a,b,\mathbf{x}+[a])\chi(b,\mu,\mathbf{x}+[b]). \tag{15}$$
$$a \neq b.$$

Equations (14) and (15) imply that

$$(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x}) = \frac{(\lambda-a)\chi(\lambda,a,\mathbf{x}-[a]+[\mu])}{(\mu-a)\chi(\mu,a,\mathbf{x}-[a]+[\mu])} = \frac{(a-\mu)\chi(a,\mu,\mathbf{x}+[a]-[\lambda])}{(a-\lambda)\chi(a,\lambda,\mathbf{x}+[a]-[\lambda])}. \tag{16}$$

Since $(\mu-a)\chi(\mu,a)\to 1$ as $\mu-a\to 0$, one gets from (16),

$$(\lambda-\mu)^2\chi(\lambda,\mu,\mathbf{x}+[\lambda])\chi(\mu,\lambda,\mathbf{x}+[\mu]) = -1. \tag{17}$$

We will solve the equations (14)–(17) in the next section. Now, let us consider the particular form of Eq. (13) for $b=0$. In terms of the CBA function, it reads as

$$\psi(\lambda,\mu,\mathbf{x}+[a]) - \psi(\lambda,\mu,\mathbf{x}) = a\psi(\lambda,\mathbf{x}+[a])\tilde{\psi}(\mu,\mathbf{x}); \quad x_i' - x_i = \frac{1}{i}a^i, \tag{18}$$

where we have introduced the Baker–Akhiezer and dual Baker–Akhiezer functions $\psi(\lambda,\mathbf{x}) = g(\lambda)\chi(\lambda,0,\mathbf{x})$, $\tilde{\psi}(\mu,\mathbf{x})=g(\mu)^{-1}\chi(0,\mu,\mathbf{x})$. This equation is a condensed finite form of the whole KP-mKP hierarchy. The expansion of this relation over a generates the KP-mKP hierarchies (and dual hierarchies) and linear problems for them. Indeed, Eq. (18) is equivalent to the infinite system of equations,

$$p_{i+1}(\tilde{\partial})\psi(\lambda,\mu,\mathbf{x}) = \tilde{\psi}(\mu,\mathbf{x})p_i(\tilde{\partial})\psi(\lambda,\mathbf{x}), \tag{19}$$

where p_i are Schur polynomials and $\tilde{\partial}=(\partial/\partial x_1,(1/2)(\partial/\partial x_2),\ldots,(1/n)(\partial/\partial x_n),\ldots)$. The first three equations from (19) are (we are using the equivalent symmetric form)

$$\psi(\lambda,\mu,\mathbf{x})_x = \psi(\lambda,\mathbf{x})\tilde{\psi}(\mu,\mathbf{x}), \tag{20}$$

$$\psi(\lambda,\mu,\mathbf{x})_y = \psi(\lambda,\mathbf{x})_x\tilde{\psi}(\mu,\mathbf{x}) - \psi(\lambda,\mathbf{x})\tilde{\psi}(\mu,\mathbf{x})_x, \tag{21}$$

$$\psi(\lambda,\mu,\mathbf{x})_t = \tfrac{1}{4}\psi(\lambda,\mu,\mathbf{x})_{xxx} - \tfrac{3}{4}\psi(\lambda,\mathbf{x})_x\tilde{\psi}(\mu,\mathbf{x})_x$$
$$+\tfrac{3}{4}(\psi(\lambda,\mathbf{x})_y\tilde{\psi}(\mu,\mathbf{x}) - \psi(\lambda,\mathbf{x})\tilde{\psi}(\mu,\mathbf{x})_y), \tag{22}$$

$$x=x_1; \quad y=x_2; \quad t=x_3.$$

A combination of the first two equations gives equivalently the equations

$$\psi(\lambda,\mu,\mathbf{x})_y-\psi(\lambda,\mu,\mathbf{x})_{xx}=-2\psi(\lambda,\mathbf{x})\tilde{\psi}(\mu,\mathbf{x})_x, \tag{23}$$

$$\psi(\lambda,\mu,\mathbf{x})_y+\psi(\lambda,\mu,\mathbf{x})_{xx}=2\psi(\lambda,\mathbf{x})_x\tilde{\psi}(\mu,\mathbf{x}). \tag{24}$$

Evaluating the first equation at $\mu=0$, the second at $\lambda=0$ (it is easier to do it rewriting the equation in terms of the function χ), one gets

$$f(\mathbf{x})_y-f(\mathbf{x})_{xx}=u(\mathbf{x})f(\mathbf{x}), \tag{25}$$

$$\tilde{f}(\mathbf{x})_y+\tilde{f}(\mathbf{x})_{xx}=-u(\mathbf{x})\tilde{f}(\mathbf{x}), \tag{26}$$

where $u(\mathbf{x})=-2\chi(0,0)_x$ and

$$f=\int \psi(\lambda,0)\rho(\lambda)d\lambda,$$

$$\tilde{f}=\int \tilde{\rho}(\mu)\psi(0,\mu)d\mu;$$

$\rho(\lambda)$ and $\tilde{\rho}(\mu)$ are some arbitrary functions.

In a similar manner, one obtains from (20)–(22) the equations

$$f_t-f_{xxx}=\tfrac{3}{2}uf_x+\tfrac{3}{4}(u_x+\partial_x^{-1}u_y)f, \tag{27}$$

$$\tilde{f}_t-\tilde{f}_{xxx}=\tfrac{3}{2}u\tilde{f}_x+\tfrac{3}{4}(u_x-\partial_x^{-1}u_y)\tilde{f}. \tag{28}$$

Both the linear system (25), (27) for the wave function f and the linear system (26), (28) for the wave function $\tilde{f}$ give rise to the same KP equation,

$$u_t=\tfrac{1}{4}u_{xxx}+\tfrac{3}{2}uu_x+\tfrac{3}{4}\partial_x^{-1}u_{yy}. \tag{29}$$

To derive linear problems for the mKP and dual mKP equations, we integrate Eqs. (20), (23), (24), and (22) with the two arbitrary functions $\rho(\lambda)$, $\tilde{\rho}(\mu)$,

$$\Phi(\mathbf{x})_x=f(\mathbf{x})\tilde{f}(\mathbf{x}), \tag{30}$$

$$\Phi(\mathbf{x})_y-\Phi(\mathbf{x})_{xx}=-2f(\mathbf{x})\tilde{f}(\mathbf{x})_x, \tag{31}$$

$$\Phi(\mathbf{x})_y+\Phi(\mathbf{x})_{xx}=2f(\mathbf{x})_x\tilde{f}(\mathbf{x}), \tag{32}$$

$$\Phi(\mathbf{x})_t-\Phi(\mathbf{x})_{xxx}=-\tfrac{3}{2}f(\mathbf{x})_x\tilde{f}(\mathbf{x})_x-\tfrac{3}{4}(f(\mathbf{x})\tilde{f}(\mathbf{x})_y-f(\mathbf{x})_y\tilde{f}(\mathbf{x})), \tag{33}$$

where

$$\Phi=\int\int \tilde{\rho}(\mu)\psi(\lambda,\mu)\rho(\lambda)d\lambda\, d\mu.$$

Using the first equation to exclude f from the second (and $\tilde{f}$ from the third), we obtain

$$\Phi_y-\Phi_{xx}=v(\mathbf{x})\Phi_x, \tag{34}$$

$$\Phi_y+\Phi_{xx}=-\tilde{v}(\mathbf{x})\Phi_x, \tag{35}$$

where $v=-2[\tilde{f}(\mathbf{x})_x/\tilde{f}(\mathbf{x})]$, $\tilde{v}=2[f(\mathbf{x})_x/f(\mathbf{x})]$.

Similarly, one gets from (22),

$$\Phi_t-\Phi_{xxx}=\tfrac{3}{2}v(\mathbf{x})\Phi_{xx}+\tfrac{3}{4}(v_x+\tfrac{1}{2}v^2+\partial_x^{-1}v_y)\Phi_x, \tag{36}$$

$$\Phi_t-\Phi_{xxx}=\tfrac{3}{2}\tilde{v}(\mathbf{x})\Phi_{xx}+\tfrac{3}{4}(\tilde{v}_x+\tfrac{1}{2}\tilde{v}^2-\partial_x^{-1}\tilde{v}_y)\Phi_x. \tag{37}$$

The system (34), (36) gives rise to the mKP equation,

$$v_t=\tfrac{1}{4}v_{xxx}-\tfrac{3}{8}v^2v_x+\tfrac{3}{4}v_x\partial_x^{-1}v_y+\tfrac{3}{4}\partial_x^{-1}v_{yy}, \tag{38}$$

while the system (35), (37) leads to the dual mKP equation, which is obtained from (38) by the substitution $v\to\tilde{v}$, $t\to -t$, $y\to -y$, $x\to -x$.

So the function Φ is simultaneously a wave function for the mKP and dual mKP linear problems with different potentials, defined by the dual KP (KP) wave functions.

Using the equation (22) and relations (34) and (35), one also obtains an equation for the function Φ,

$$\Phi_t=\tfrac{1}{4}\Phi_{xxx}+\tfrac{3}{8}\frac{\Phi_y^2-\Phi_{xx}^2}{\Phi_x}+\tfrac{3}{4}\Phi_xW_y, \quad W_x=\frac{\Phi_y}{\Phi_x}. \tag{39}$$

This equation first arose in the Painlève analysis of the KP equation as a singularity manifold equation.[16]

It is a tedious but absolutely straightforward check that the expansion of (18) in higher orders of a generates (**1**) the whole hierarchy of KP singularity manifold equations for $\psi(\lambda,\mu)$ [or $\Phi(\mathbf{x})$]; (**2**) the hierarchy of linear problems for the mKP and dual mKP equations, where $\psi(\lambda,\mu)$ [or $\Phi(\mathbf{x})$] is the common wave function and $v=-2(\ln\tilde{\psi}(\mu,\mathbf{x}))_x$, $\tilde{v}=-2(\ln\psi(\lambda,\mathbf{x}))_x$ are the potentials; (**3**) the mKP hierarchy for v and the dual mKP hierarchy for $\tilde{v}$; (**4**) the hierarchies of the KP linear problems for $\psi(\lambda,\mathbf{x})$ and the dual KP linear problems for $\tilde{\psi}(\mu,\mathbf{x})$ and; finally, (**5**) the KP hierarchy of equations for $u=-2\chi(0,0)_x$.

Note also one interesting consequence of the formula (17),

$$\chi(0,\lambda,\mathbf{x})=-\frac{1}{\lambda^2}\chi^{-1}(\lambda,0,\mathbf{x}+[\lambda]). \tag{40}$$

IV. KP HIERARCHY IN THE "MOVING FRAME." DARBOUX EQUATIONS AS THE HORIZONTAL SUBHIERARCHY

Now let us consider the expansion of the LHS of (13) over $\epsilon=a-b$, where $\epsilon\to 0$. In the first order in ϵ, one gets

$$\Delta_1(b)\psi(\lambda,\mu,\mathbf{x})=\psi(b,\mu,\mathbf{x})\psi(\lambda,b,\mathbf{x}), \tag{41}$$

where

$$\Delta_1(b)=\sum_{n=1}^{\infty}b^{n-1}\frac{\partial}{\partial x_n}.$$

In the higher orders in ϵ one obtains the hierarchy of equations of the form (20)–(22) and their higher analogs with the substitution $\psi(\lambda,0,\mathbf{x})\to\psi(\lambda,b,\mathbf{x})$, $\psi(0,\mu,\mathbf{x})\to\psi(b,\mu,\mathbf{x})$ and

$$\frac{\partial}{\partial x_i}\to\Delta_i(b)=\sum_{n=i}^{\infty}\frac{n!}{n(n-i)!i!}b^{n-i}\frac{\partial}{\partial x_i}.$$

Such a substitution is, in fact, nothing but the change of dynamical variables (or the coordinates on the group of functions g). Indeed, it is not difficult to show that $\Delta_i(b)=\partial/\partial x_i(b)$, where the dynamical variables $x_i(b)$ are defined by the relation

$$\sum_{i=1}^{\infty} \frac{x_i(b)}{(\lambda-b)^i}=\sum_{i=1}^{\infty} \frac{x_i}{(\lambda)^i}. \tag{42}$$

It is clear that

$$\left[\frac{\partial}{\partial x_i(\lambda')},\frac{\partial}{\partial x_i(\lambda)}\right]=0. \tag{43}$$

Note one interesting property of the derivatives $\partial/\partial x_i(b)$, namely,

$$\left[\frac{\partial}{\partial \lambda},\frac{\partial}{\partial x_i(\lambda)}\right]=(i+1)\frac{\partial}{\partial x_{i+1}(\lambda)}. \tag{44}$$

So the operator $\partial/\partial\lambda$ is a ''master symmetry'' for all vector fields $\partial/\partial x_i(\lambda)$.

The expansion of Eq. (13) up to the third order in ϵ gives the equations

$$\frac{\partial}{\partial x_1(b)}\psi(\lambda,\mu,\mathbf{x}(b))=\psi(b,\mu,\mathbf{x}(b))\psi(\lambda,b,\mathbf{x}(b)), \tag{45}$$

$$\frac{\partial}{\partial x_2(b)}\psi(\lambda,\mu,\mathbf{x}(b))=\frac{\partial}{\partial x_1(b)}\psi(\lambda,b,\mathbf{x}(b))\cdot\psi(b,\mu,\mathbf{x}(b)) - \psi(\lambda,b,\mathbf{x}(b))\cdot\frac{\partial}{\partial x_1(b)}\psi(b,\mu,\mathbf{x}(b)), \tag{46}$$

$$\begin{aligned}&\frac{\partial}{\partial x_3(b)}\psi(\lambda,\mu,x(b))\\ &=\frac{1}{4}\frac{\partial^3}{\partial x_1(b)^3}\psi(\lambda,\mu,\mathbf{x}(b))-\frac{3}{4}\frac{\partial}{\partial x_1(b)}\psi(\lambda,b,\mathbf{x}(b))\cdot\frac{\partial}{\partial x_1(b)}\psi(b,\mu,\mathbf{x}(b))\\ &\quad+\frac{3}{4}\left(\frac{\partial}{\partial x_2(b)}\psi(\lambda,b,\mathbf{x}(b))\cdot\psi(b,\mu)-\psi(\lambda,b,\mathbf{x}(b))\cdot\frac{\partial}{\partial x_2(b)}\psi(b,\mu,\mathbf{x}(b))\right).\end{aligned} \tag{47}$$

The analogs of Eqs. (23) and (24) have the form

$$\frac{\partial}{\partial x_2(b)}\psi(\lambda,\mu,\mathbf{x}(b))-\frac{\partial^2}{\partial x_1(b)^2}\psi(\lambda,\mu,\mathbf{x}(b))+2\psi(\lambda,b)\frac{\partial}{\partial x_1(b)}\psi(b,\mu,\mathbf{x}(b))=0, \tag{48}$$

$$\frac{\partial}{\partial x_2(b)}\psi(\lambda,\mu,\mathbf{x}(b))+\frac{\partial^2}{\partial x_1(b)^2}\psi(\lambda,\mu,\mathbf{x}(b))-2\frac{\partial}{\partial x_1(b)}\psi(\lambda,b)\cdot\psi(b,\mu,\mathbf{x}(b))=0. \tag{49}$$

Equations (45)–(47) and higher equations again give rise to the generalized KP hierarchy, but now in coordinates $x_i(b), i=1,2,3,\ldots$. For such a KP hierarchy written in the ''moving'' frame, the parameter b is an arbitrary one, but fixed.

Let us consider now equations of the type (45) written for several values of b. We denote $x_1(b_\alpha)=\xi_\alpha$, $\alpha=1,2,\ldots,n$. Equations (45), taken for $b=b_\alpha$, $\lambda=b_\beta$, $\mu=b_\gamma(\alpha\neq\beta\neq\gamma)$, look like

$$\frac{\partial}{\partial \xi_\alpha}\psi_{\beta\gamma}=\psi_{\beta\alpha}\psi_{\alpha\gamma},\quad \alpha\neq\beta\neq\gamma, \tag{50}$$

where $\psi_{\alpha\beta}=\psi(b_\alpha,b_\beta,\mathbf{x})$. The system (50) is just a well-known system of N^2-N resonantly interacting waves.

Integrating Eqs. (45) over μ with the function $\rho(\mu)$ and evaluating the result at $b=b_\alpha$, $\gamma=b_\beta$, one gets

$$\frac{\partial f_\beta}{\partial \xi_\alpha} = \psi_{\beta\alpha} f_\alpha \quad (\alpha \neq \beta), \tag{51}$$

where $f_\beta = \int d\mu \psi(b_\beta, \mu)\rho(\mu)$. Analogously, one gets

$$\frac{\partial f^*_\beta}{\partial \xi_\alpha} = \psi_{\alpha\beta} f^*_\alpha \quad (\alpha \neq \beta), \tag{52}$$

where $f^*_\beta = \int d\lambda \psi(\lambda, b_\beta)\rho^*(\lambda)$ and $\rho^*(\lambda)$ is an arbitrary function. The systems (51) and (52) are the linear problem and dual linear problem for Eqs. (50), respectively.

Expressing $\psi_{\alpha\beta}$ via f_α and f^*_α, one gets from (51) and (52) [using (50)] the same system for f_α and f^*_α,

$$\frac{\partial^2 H_\alpha}{\partial \xi_\beta \partial \xi_\gamma} = \frac{1}{H_\beta}\frac{\partial H_\beta}{\partial \xi_\gamma}\frac{\partial H_\alpha}{\partial \xi_\beta} + \frac{1}{H_\gamma}\frac{\partial H_\gamma}{\partial \xi_\beta}\frac{\partial H_\alpha}{\partial \xi_\gamma} \quad (\alpha \neq \beta \neq \gamma \neq \alpha). \tag{53}$$

The system (53) is the Darboux system which was introduced for the first time in the differential geometry of surfaces[14] and then was rediscovered in the matrix form within the ∂-dressing method in the paper.[5] Note that the Darboux equations in the variables of the type $x_1(b_\alpha)$ have also appeared in the paper[17] within completely different approach.

One can treat the Darboux equations (53) with different n as the horizontal subhierarchy of the whole generalized KP hierarchy. Note that Eqs. (45)–(47) and their higher analogs give rise to the higher resonantly interacting waves equations.

V. COMBESCURE SYMMETRY TRANSFORMATIONS FOR THE GENERALIZED KP HIERARCHY

Let us consider now the symmetries of the equations derived above. All the higher equations of the hierarchy are, as usual, the symmetries of each member of the hierarchy. Here we will discuss another type of symmetry.

Since $\rho(\lambda)$ and $\tilde{\rho}(\mu)$ are arbitrary functions, Eqs. (39) and the hierarchy (18) possess the symmetry transformation

$$\Phi(\rho(\lambda), \tilde{\rho}(\mu)) \to \Phi' = \Phi(\rho'(\lambda), \tilde{\rho}'(\mu)).$$

This transformation is, in fact, the transformation which changes the normalization of the wave functions. The fact that such transformations are connected with the so-called Combescure transformations, known for a long time in differential geometry, was pointed out in Ref. 12.

The Combescure transformation was introduced last century within the study of the transformation properties of surfaces (see, e.g., Refs. 18, 19). It is a transformation of surface such that the tangent vector at a given point of the surface remains parallel. The Combescure transformation is essentially different from the well-known Bäcklund and Darboux transformations. The Combescure transformation plays an important role in the theory of the systems of a hydrodynamical type.[20] It is also of great interest for the theory of $(2+1)$-dimensional integrable systems.[21] Combescure symmetry transformations are an essential part of the analytic-bilinear approach.

The Combescure transformation can be characterized in terms of the corresponding invariants. The simplest of these invariants for the mKP equation is just the potential of the KP equation L operator expressed through the wave function

$$u = \frac{f(\mathbf{x})_y - f(\mathbf{x})_{xx}}{f(\mathbf{x})}, \tag{54}$$

$$u = \frac{\tilde{f}(\mathbf{x})_y - \tilde{f}(\mathbf{x})_{xx}}{\tilde{f}(\mathbf{x})}, \tag{55}$$

or, in terms of the solution for the mKP (dual mKP) equation,

$$v'_y+v'_{xx}-\tfrac{1}{2}((v')^2)_x=v_y+v_{xx}-\tfrac{1}{2}(v^2)_x, \tag{56}$$

$$\tilde{v}'_y-\tilde{v}'_{xx}-\tfrac{1}{2}((\tilde{v}')^2)_x=\tilde{v}_y-\tilde{v}_{xx}-\tfrac{1}{2}(\tilde{v}^2)_x. \tag{57}$$

The solutions of the mKP equations are transformed only by a subgroup of the Combescure symmetry group corresponding to the change of the weight function $\tilde{\rho}(\mu)$ (left subgroup), and they are invariant under the action of the subgroup corresponding to $\rho(\lambda)$ (vice versa for the dual mKP).

All the hierarchy of the Combescure transformation invariants is given by the expansion over ϵ near the point $\mathbf{x}$ of the relation (18), rewritten in the form

$$\frac{\partial}{\partial\epsilon}\left(\frac{\tilde{f}(\mathbf{x}')-\tilde{f}(\mathbf{x})}{\epsilon\tilde{f}(\mathbf{x})}\right)=-\frac{1}{2}\frac{\partial}{\partial\epsilon}\partial_{x'}^{-1}u(\mathbf{x}'),\quad x'_i-x_i=\frac{1}{i}\epsilon^i; \tag{58}$$

$$\frac{\partial}{\partial\epsilon}\left(\frac{f(\mathbf{x})-f(\mathbf{x}')}{\epsilon f(\mathbf{x})}\right)=\frac{1}{2}\frac{\partial}{\partial\epsilon}\partial_{x'}^{-1}u(\mathbf{x}'),\quad x'_i-x_i=-\frac{1}{i}\epsilon^i. \tag{59}$$

The expansion of the left part of these relations gives the Combescure transformation invariants in terms of the wave functions $\tilde{f}$, f. To express them in terms of the mKP equation (the dual mKP equation) solution, one should use the formulas

$$v=-2\frac{\tilde{f}_x}{\tilde{f}},\quad \tilde{f}=\exp\left(-\frac{1}{2}\partial_x^{-1}v\right); \tag{60}$$

$$\tilde{v}=2\frac{f_x}{f},\quad f=\exp\left(\frac{1}{2}\partial_x^{-1}\tilde{v}\right). \tag{61}$$

It is also possible to consider special Combescure transformations, keeping invariant the KP equation (the dual KP equation) wave functions [i.e., solutions for the dual mKP (mKP) equations]. The first invariants of this type are

$$\frac{\Phi'_x(\mathbf{x})}{\tilde{f}'(\mathbf{x})}=\frac{\Phi_x(\mathbf{x})}{\tilde{f}(\mathbf{x})}, \tag{62}$$

$$\frac{\Phi'_x(\mathbf{x})}{f'(\mathbf{x})}=\frac{\Phi_x(\mathbf{x})}{f(\mathbf{x})}. \tag{63}$$

All the hierarchies of the invariants of this type are generated by the expansion of the left part of the following relations over ϵ:

$$\left(\frac{\Phi(\mathbf{x}')-\Phi(\mathbf{x})}{\tilde{f}(\mathbf{x})}\right)=\epsilon f(\mathbf{x}'),\quad x'_i-x_i=\frac{1}{i}\epsilon^i; \tag{64}$$

$$\left(\frac{\Phi(\mathbf{x})-\Phi(\mathbf{x}')}{f(\mathbf{x})}\right)=\epsilon\tilde{f}(\mathbf{x}'),\quad x'_i-x_i=-\frac{1}{i}\epsilon^i. \tag{65}$$

Now let us consider the equation (39) and all the hierarchies given by the relation (18). This equation admits the Combescure group of symmetry transformations $\Phi(\rho(\lambda),\tilde{\rho}(\mu))\to\Phi'=\Phi(\rho'(\lambda),\tilde{\rho}'(\mu))$ consisting of two subgroups (the right and left Combescure transformations). These subgroups have the following invariants:

$$v=\frac{\Phi_y-\Phi_{xx}}{\Phi_x} \tag{66}$$

and

$$\tilde{v}=\frac{\Phi_y+\Phi_{xx}}{\Phi_x}. \tag{67}$$

From (34), (35) it follows that they just obey the mKP and dual mKP equation, respectively. The invariant for the full Combescure transformation can be obtained by the substitution of the expression for v via Φ (66) to the formula (56). It reads as

$$u=\partial_x^{-1}\left(\frac{\Phi_y}{\Phi_x}\right)_y-\frac{\Phi_{xxx}}{\Phi_x}+\frac{\Phi_{xx}^2-\Phi_y^2}{2\Phi_x^2}. \tag{68}$$

From (25), (26), (54), (55), (56), and (57) it follows that u solves the KP equation.

So there is an interesting connection between Eq. (39), mKP–dual mKP equations, and the KP equation. Equation (39) is the unifying equation. It possesses a Combescure symmetry transformations group. After the factorization of Eq. (39) with respect to one of the subgroups (right or left), one gets the mKP or dual mKP equation in terms of the invariants for the subgroup (66), (67). The factorization of Eq. (39) with respect to the full Combescure transformations group gives rise to the KP equation in terms of the invariant of group (68).

In other words, the invariant of Eq. (39) under the full Combescure group is described by the KP equation, while the invariants under the action of its right and left subgroups are described by the mKP or dual mKP equations.

VI. τ FUNCTION AND ADDITION FORMULAS

Now we will analyze the functional equations (13)–(17). Evaluating (16) at $a=0$, one gets

$$\chi(\lambda,\mu,\mathbf{x})=\frac{1}{\lambda-\mu}\,\frac{\lambda\chi(\lambda,0,\mathbf{x}+[\mu])}{\mu\chi(\mu,0,\mathbf{x}+[\mu])}, \tag{69}$$

and the relation

$$\frac{\lambda\chi(\lambda,0,\mathbf{x}+[\mu])}{\mu\chi(\mu,0,\mathbf{x}+[\mu])}=\frac{\mu\chi(0,\mu,\mathbf{x}-[\lambda])}{\lambda\chi(0,\lambda,\mathbf{x}-[\lambda])}. \tag{70}$$

Using the relation (40), one obtains from the above equation the relation which contains the function $\chi(\lambda,0,\mathbf{x})$ only. It is of the form

$$\chi(\lambda,0,\mathbf{x})\chi(\mu,0,\mathbf{x}-[\lambda])=\chi(\mu,0,\mathbf{x})\chi(\lambda,0,\mathbf{x}-[\mu]), \tag{71}$$

or, equivalently,

$$\frac{\chi(\lambda,0,\mathbf{x}-[\mu])}{\chi(\lambda,0,\mathbf{x})}=\frac{\chi(\mu,0,\mathbf{x}-[\lambda])}{\chi(\mu,0,\mathbf{x})}. \tag{72}$$

It is straightforward to check that Eqs. (14)–(17) are satisfied identically due to the formulas (69) and (71) or (72). So the problem is reduced to resolution of the functional equation (71) or (72).

Introducing the function $\phi(\lambda,\mu,\mathbf{x})=(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})$ analytic in the unit disk we get

$$\chi(\lambda,\mu,\mathbf{x})=\frac{1}{\lambda-\mu}\,\frac{\phi(\lambda,0,\mathbf{x}+[\mu])}{\phi(\mu,0,\mathbf{x}+[\mu])}, \tag{73}$$

and the equation

$$\frac{T_\mu\phi(\lambda,0,\mathbf{x})}{\phi(\lambda,0,\mathbf{x})}=\frac{T_\lambda\phi(\mu,0,\mathbf{x})}{\phi(\mu,0,\mathbf{x})}, \tag{74}$$

where $T_\lambda f(\mathbf{x})=f(\mathbf{x}-[\lambda])$.

Remark: Note that for the function $R(a,b,\mathbf{x})=\phi(a,0,\mathbf{x}+[a]+[b])$, Eq. (15) implies

$$R(a,\lambda)R(\lambda,b)R(b,a)=R(a,b)R(b,\lambda)R(\lambda,a).$$

So the problem of resolution of Eqs. (14) and (15) is connected also with the functional equation which is of the form of the triangle (Yang–Baxter) equation, well known in the theory of quantum solvable models (see, e.g., Ref. 22).

A symmetric structure of Eq. (74) clearly suggests a substitution resolving it,

$$\phi(\lambda,0,\mathbf{x})=\frac{T_\lambda \tau(\mathbf{x})}{\tau(\mathbf{x})}=\frac{\tau(\mathbf{x}-[\lambda])}{\tau(\mathbf{x})}, \tag{75}$$

where $\tau(\mathbf{x})$ is an arbitrary function.

Using (69), we get

$$\chi(\lambda,\mu,\mathbf{x})=\frac{1}{(\lambda-\mu)}\frac{\tau(\mathbf{x}-[\lambda]+[\mu])}{\tau(\mathbf{x})}. \tag{76}$$

We would like to emphasize here that the functions $\phi(\lambda,0,\mathbf{x})$, $\phi(0,\mu,\mathbf{x})$ define, in fact, Baker–Akhiezer and dual (adjoint) Baker–Akhiezer functions which satisfy a standard Hirota bilinear identity [it follows directly from (4); the only minor difference with the common approach is that we work in the neighborhood of zero, not in the neighborhood of infinity]. The formula (75) for the Baker–Akhiezer function is well known (see Refs. 1–4). We have shown that expression (75) for the BA function implies expression (76) for the CBA function. So this point gives a connection to the standard technique. Expression (76) could serve as a definition to introduce the CBA function to the context of the Grassmannian technique (see also Refs. 15, 13). And we could refer to the known results, considering the question of the existence of the τ function providing a substitution (75) for the arbitrary BA function. But we prefer to give an independent proof here, because the use of functional equations for the CBA function makes this proof quite simple and elegant, and it also provides a new expression for the closed one-form defining the τ function.

To prove the existence of the τ function, we use Eq. (41) taken at $\lambda\to\mu$. Equation (41), in terms of the function χ, reads as

$$\Delta_1(b)\chi(\lambda,\mu,\mathbf{x})+\left(\frac{1}{\lambda-b}-\frac{1}{\mu-b}\right)\chi(\lambda,\mu,\mathbf{x})=\chi(\lambda,b,\mathbf{x})\chi(b,\mu,\mathbf{x}). \tag{77}$$

Taking $\lambda\to\mu$, we get

$$\Delta_1(\mu)\chi_{\mathrm{r}}(\lambda,\lambda,\mathbf{x})=\Delta_1(\lambda)\chi_{\mathrm{r}}(\mu,\mu,\mathbf{x})=\chi(\lambda,\mu,\mathbf{x})\chi(\mu,\lambda,\mathbf{x})+(\lambda-\mu)^{-2}, \tag{78}$$

where $\chi_{\mathrm{r}}(\lambda,\mu,\mathbf{x})=\chi(\lambda,\mu,\mathbf{x})-(\lambda-\mu)^{-1}$. This equation implies that there exists a function $\tau(\mathbf{x})$, such that

$$\Delta_1(\lambda)\ln\ \tau(\mathbf{x})=-\chi_{\mathrm{r}}(\lambda,\lambda,\mathbf{x}), \tag{79}$$

or, using the expansion in λ,

$$d\ \ln\ \tau(\mathbf{x})=-\sum_{n=1}^{\infty}\ \frac{\partial^{n-1}}{\partial\lambda^{n-1}}\chi_{\mathrm{r}}(\lambda,\lambda,\mathbf{x})\Big|_{\lambda=0}dx_n. \tag{80}$$

[(78) guarantees that the cross-derivatives are equal, or, in other words, that the form (80), defining the τ function, is closed].

Moreover, using (41), it is also possible to express the function $\chi(\lambda,\mu,\mathbf{x})$ through the τ function. Taking the limit $b\to\lambda$, one obtains

$$\Delta_1(\lambda)\chi(\lambda,\mu,\mathbf{x})+\left(\frac{\partial}{\partial\lambda}-\frac{1}{\mu-\lambda}\right)\chi(\lambda,\mu,\mathbf{x})=\chi_{\mathrm{r}}(\lambda,\lambda,\mathbf{x})\chi(\lambda,\mu,\mathbf{x}), \tag{81}$$

or, for the function $\phi(\lambda,\mu,\mathbf{x})=(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})$, using the definition of the τ function (80),

$$\left(\Delta_1(\lambda)+\frac{\partial}{\partial\lambda}\right)\ln\ \phi(\lambda,\mu,\mathbf{x})=-\Delta_1(\lambda)\ln\ \tau(\mathbf{x}). \tag{82}$$

Equation (82) is a nonhomogeneous infinite-dimensional wave equation. A general solution for it is a sum of the general solution of a homogeneous equation and the special solution of a nonhomogeneous equation, it can be written as

$$\ln\ \phi(\lambda,\mu,\mathbf{x})=F(\mu,\mathbf{x}-[\lambda])-\ln\ \tau(\mathbf{x}). \tag{83}$$

Then, using the condition $\phi(0,0,\mathbf{x})=1$, one obtains

$$\phi(\lambda,0,\mathbf{x})=\frac{\tau(\mathbf{x}-[\lambda])}{\tau(\mathbf{x})}, \tag{84}$$

Q.E.D.

A similar argument leads to the formula

$$\phi(0,\mu,\mathbf{x})=\frac{\tau(\mathbf{x}+[\mu])}{\tau(\mathbf{x})}. \tag{85}$$

So the τ function that we have introduced by the definition (80) is a usual τ function of the KP hierarchy. The next step is to find the formula expressing the CBA function through the τ function; given the formula (84), we already have done it in the beginning of the section. So we have proved that there exists a τ function such that the CBA function is expressed through it by the formula (76).

Now we will use the general equation (13). Substituting (76) into (13), one gets

$$\begin{aligned}&(a-\mu)(\lambda-b)\tau(\mathbf{x}+[a]+[\mu])\tau(\mathbf{x}+[\lambda]+[b])+(\lambda-a)(\mu-b)\tau(\mathbf{x}+[\lambda]+[a])\tau(\mathbf{x}+[b]+[\mu])\\&\quad+(b-a)(\lambda-\mu)\tau(\mathbf{x}+[b]+[a])\tau(\mathbf{x}+[\lambda]+[\mu])=0.\end{aligned} \tag{86}$$

It is nothing but the simplest addition formula for the τ function derived in Ref. 1, which is closely connected with the Fay's trisecant formula.[23]

The generalized Hirota identity gives rise also to other addition formulas from Ref. 1. Indeed, let us choose in (4),

$$g(\mathbf{x})g^{-1}(\mathbf{x}')=\prod_{\alpha=1}^{n}\ \frac{\nu-a_\alpha}{\nu-b_\alpha}, \tag{87}$$

where n is an arbitrary integer and $\mathbf{x}'-\mathbf{x}=\Sigma_{\alpha=1}^{n}[a_\alpha]-[b_\alpha]$. In this case, Eq. (4) gives

$$\begin{aligned}&\prod_{\alpha=1}^{n}\ \frac{\mu-a_\alpha}{\mu-b_\alpha}\chi\left(\lambda,\mu,\mathbf{x}+\sum_{\alpha=1}^{n}\ [a_\alpha]\right)-\prod_{\alpha=1}^{n}\ \frac{\lambda-a_\alpha}{\lambda-b_\alpha}\chi\left(\lambda,\mu,\mathbf{x}+\sum_{\alpha=1}^{n}\ [b_\alpha]\right)+\sum_{\alpha=1}^{n}\ (b_\alpha-a_\alpha)\\&\times\prod_{\gamma,\gamma\neq\alpha}\ \frac{b_\alpha-a_\gamma}{b_\alpha-b_\gamma}\chi\left(\lambda,b_\alpha,\mathbf{x}+\sum_{\delta=1}^{n}\ [a_\delta]\right)\chi\left(b_\alpha,\mu,\mathbf{x}+\sum_{\delta=1}^{n}\ [b_\delta]\right)=0.\end{aligned} \tag{88}$$

Substituting the expression (76) into (88) and shifting $\mathbf{x}\to\mathbf{x}+[\lambda]$, one gets

$$\prod_{\alpha=1}^{n} \frac{(\mu-a_\alpha)}{(\mu-b_\alpha)(\mu-\lambda)} \tau\left(\mathbf{x}+[\mu]+\sum_{\gamma=1}^{n}[a_\gamma]\right)\tau\left(\mathbf{x}+[\lambda]+\sum_{\gamma=1}^{n}[b_\gamma]\right)$$

$$-\prod_{\alpha=1}^{n} \frac{(\lambda-a_\alpha)}{(\lambda-b_\alpha)(\lambda-\mu)} \tau\left(\mathbf{x}+[\mu]+\sum_{\gamma=1}^{n}[b_\gamma]\right)\tau\left(\mathbf{x}+[\lambda]+\sum_{\gamma=1}^{n}[a_\gamma]\right)$$

$$+\sum_{\alpha=1}^{n} \frac{(b_\alpha-a_\alpha)}{(b_\alpha-\mu)(\lambda-b_\alpha)} \prod_{\gamma,\gamma\neq\alpha} \frac{b_\alpha-a_\gamma}{b_\alpha-b_\gamma} \tau\left(\mathbf{x}+[b_\alpha]+\sum_{\beta=1}^{n}[a_\beta]\right)$$

$$\times\tau\left(\mathbf{x}+[\lambda]+[\mu]+\sum_{\beta,\beta\neq\alpha}[b_\beta]\right)=0. \tag{89}$$

It is not difficult to check (after some renotations) that Eq. (89) coincides with the Plücker's relations for a universal Grassmannian manifold (see Theorem 3 of Ref. 1). That means, according to Theorems 1–3 of the paper,[1] that τ is the τ function of the KP hierarchy. Note that the formulas (76), (89) provide solutions simultaneously for the KP, mKP (dual mKP), and SM-KP hierarchies.

VII. DETERMINANT FORMULAS FOR TRANSFORMATIONS

The analytic-bilinear approach allows us to represent in a compact form, not only the integrable hierarchies but also rather general transformations acting in the space of solutions.

We will consider here the transformations which are equivalent to the action of an arbitrary meromorphic function $g(\lambda)$ on the CBA function $\chi(\lambda,\mu,x)$. Let $g(\lambda)$ be a meromorphic function in G which has simple poles at the points a_i $(i=1,2,\ldots,n)$ and simple zeros at the points b_i $(i=1,2,\ldots,n)$, i.e., $g(\lambda)=\Pi_{i=1}^{n}(\lambda-a_i)/(\lambda-b_i)$. To construct the transformed CBA function $\chi^\dagger(\lambda,\mu)$, it is sufficient to find a solution of the equation

$$\int_{\partial G}\chi(\nu,\mu)g(\nu)\chi^\dagger(\lambda,\nu)d\nu=0,$$

where $\chi^\dagger$ has the same normalization $((\lambda-\mu)^{-1})$ as χ.

The simplest way to find $\chi^\dagger$ consists in using the following consequence of Eq. (4),

$$\int_{\partial G}\int_{\partial G} d\nu\, d\rho\ \chi(\nu,\mu)g(\nu)\chi^\dagger(\rho,\nu,g)g^{-1}(\rho)\chi(\lambda,\rho)=0.$$

Using this formula, one finds

$$\chi^\dagger(\lambda,\mu)=g^{-1}(\lambda)g(\mu)\frac{\det\Delta_{n+1}}{\det\Delta_n}, \tag{90}$$

where

$$\Delta_{m,\alpha\beta}=\chi(a_\alpha,b_\beta),\quad \alpha,\beta=1,2,\ldots,m,$$

and $a_{n+1}=\lambda$, $b_{n+1}=\mu$.

The formula (90) defines the generic discrete transformation of χ. In terms of the τ function, this transformation has the form

$$\frac{\tau^\dagger(\mathbf{x}-[\lambda]+[\mu])}{\tau^\dagger(\mathbf{x})}=(\lambda-\mu)g^{-1}(\lambda)g(\mu)\frac{\det F_{n+1}}{\tau(\mathbf{x})\det F_n},$$

where

$$F_{n,\alpha\beta}=\frac{\tau(\mathbf{x}-[a_\alpha]+[b_\beta])}{a_\alpha-b_\beta}.$$

In the simplest case $n=1$, we have ($b_1=b$, $a_1=a$)

$$\chi^{\dagger}(\lambda,\mu)=\frac{(\lambda-b)(\mu-a)}{(\lambda-a)(\mu-b)}\frac{\begin{vmatrix}\chi(a,b) & \chi(a,\mu)\\ \chi(\lambda,b) & \chi(\lambda,\mu)\end{vmatrix}}{\chi(a,b)} \tag{91}$$

and

$$\frac{\tau^{\dagger}(\mathbf{x}-[\lambda]+[\mu])}{\tau^{\dagger}(\mathbf{x})}=\frac{(\lambda-\mu)(a-b)(\lambda-b)(\mu-a)}{(\lambda-a)(\mu-b)}\frac{\begin{vmatrix}\dfrac{\tau(\mathbf{x}-[a]+[b])}{a-b} & \dfrac{\tau(\mathbf{x}-[a]+[\mu])}{a-\mu}\\ \dfrac{\tau(\mathbf{x}-[\lambda]+[b])}{\lambda-b} & \dfrac{\tau(\mathbf{x}-[\lambda]+[\mu])}{\lambda-\mu}\end{vmatrix}}{\tau(\mathbf{x})\,\tau(\mathbf{x}-[a]+[b])},$$

where $|A|=\det A$. One can represent the transformations (90) also in terms of the function $\psi(\lambda,\mu)$. In that form the determinant formulas (90) taken at $\lambda=0$ or $\mu=0$ are very similar to the determinant formulas for the iterated Darboux transformations (see, e.g., Ref. 24). The formulas (90) give us the Darboux transformations for all subhierarchies (KP, mKP, dual mKP, SM-KP hierarchies) of the generalized KP hierarchy.

The determinant formulas (90) provide us also the multilinear relations for the τ function. Since $\tau^{\dagger}(\mathbf{x})=\tau(\mathbf{x}+\Sigma_{i=1}^{n}(-[a_i]+[b_i]))$, the formula (90) is the n-linear relation for the τ function. It is easy to check that in the simplest case $n=1$ this formula gives the simplest addition formula (86). At $n=2$ the formula looks like

$$\tau(\mathbf{x})\,\tau(\mathbf{x}-[\lambda]+[\mu]+[b_1]-[a_1]+[b_2]-[a_2])\times\begin{vmatrix}\dfrac{\tau(\mathbf{x}-[a_1]+[b_1])}{a_1-b_1} & \dfrac{\tau(\mathbf{x}-[a_1]+[b_2])}{a_1-b_2}\\ \dfrac{\tau(\mathbf{x}-[a_2]+[b_1])}{a_2-b_1} & \dfrac{\tau(\mathbf{x}-[a_2]+[b_2])}{a_2-b_2}\end{vmatrix}$$

$$=\frac{(\lambda-\mu)(\mu-a_1)(\mu-a_2)(\lambda-b_1)(\lambda-b_2)}{(\mu-b_1)(\mu-b_2)(\lambda-a_1)(\lambda-a_2)}\,\tau(\mathbf{x}+[b_1]-[a_1]+[b_2]-[a_2])$$

$$\times\begin{vmatrix}\dfrac{\tau(\mathbf{x}-[\lambda]+[\mu])}{\lambda-\mu} & \dfrac{\tau(\mathbf{x}-[\lambda]+[b_1])}{\lambda-b_1} & \dfrac{\tau(\mathbf{x}-[\lambda]+[b_2])}{\lambda-b_2}\\ \dfrac{\tau(\mathbf{x}-[a_1]+[\mu])}{a_1-\mu} & \dfrac{\tau(\mathbf{x}-[a_1]+[b_1])}{a_1-b_1} & \dfrac{\tau(\mathbf{x}-[a_1]+[b_2])}{a_1-b_2}\\ \dfrac{\tau(\mathbf{x}-[a_2]+[\mu])}{a_2-\mu} & \dfrac{\tau(\mathbf{x}-[a_2]+[b_1])}{a_2-b_1} & \dfrac{\tau(\mathbf{x}-[a_2]+[b_2])}{a_2-b_2}\end{vmatrix}. \tag{92}$$

It is cumbersome but straightforward to check that the relation (92) is satisfied due to the higher addition formulas (89) with $n=2$.

VIII. τ FUNCTION AS A FUNCTIONAL

The formula (76) can be written both for the variables $\mathbf{x}$ and in terms of the function g (g is defined on the unit circle); in the second case the τ function is treated as a functional of the function g,

$$\chi(\lambda,\mu;g)=\frac{\tau(g\times g_{(\lambda,\mu)})}{\tau(g)(\lambda-\mu)} \tag{93}$$

(see Ref. 13), where $g_{(\lambda,\mu)}(\nu)=(\nu-\lambda)/(\nu-\mu)$. The expression in terms of g will be important for the second part of our paper to transfer the notion of the τ function to the multicomponent case. So in this section we will give independent proof of the unique existence of the τ function, providing an expression (93) for an arbitrary CBA function (in the class of g analytic and having

no zeros outside the unit circle, normalized by 1 at infinity). A definition of the τ function in terms of a closed variational one-form arises in this context. The second part of the proof (existence) could be conducted similar to the proof in Sec. VI, but here we prefer to give a straightforward calculation to demonstrate some peculiar features of the direct derivation (especially the use of a determinant formula).

Differentiating (93) with respect to λ, one obtains

$$-\frac{1}{\tau(g)} \oint \frac{\delta\tau(g\times g_{(\lambda,\mu)})}{\delta \ln g(\nu)} \frac{1}{\nu-\lambda} d\nu = \frac{\partial}{\partial\lambda}(\lambda-\mu)\chi(\lambda,\mu;g), \tag{94}$$

or for $\lambda=\mu$

$$-\frac{1}{\tau(g)} \oint \frac{\delta\tau(g)}{\delta \ln g(\nu)} \frac{1}{\nu-\lambda} d\nu = \chi_{\mathrm{r}}(\lambda,\lambda;g), \tag{95}$$

where $\chi_{\mathrm{r}}(\lambda,\mu;g)=\chi(\lambda,\mu;g)-(\lambda-\mu)^{-1}$. This equation can be rewritten in the form

$$-\frac{1}{\tau(g)} \oint \frac{\delta\tau(g)}{\delta \ln g(\nu)} \frac{1}{\nu-\lambda} d\nu = \oint \chi_{\mathrm{r}}(\nu,\nu;g) \frac{1}{\nu-\lambda} d\nu. \tag{96}$$

Relation (96) implies that the functionals $-[1/\tau(g)][\delta\tau(g)/\delta \ln g(\nu)]$ and $\chi_{\mathrm{r}}(\nu,\nu;g)$ are identical for the class of functions analytic outside the unit circle and decreasing at infinity. Thus, for $\delta g/g$ belonging to this class, one has

$$-\delta \ln \tau(g) = \oint \chi_{\mathrm{r}}(\nu,\nu;g) \frac{\delta g(\nu)}{g(\nu)} d\nu. \tag{97}$$

This expression gives a variational 1-form defining the τ function. It is easy to prove using the identity (4) that this form is *closed*. Indeed, according to (4),

$$\delta\chi(\lambda,\mu;g) = \oint \chi(\nu,\mu;g) \frac{\delta g(\nu)}{g(\nu)} \chi(\lambda,\nu;g) d\nu,$$

$$\delta\chi_{\mathrm{r}}(\lambda,\lambda;g) = \oint \chi_{\mathrm{r}}(\nu,\lambda;g) \frac{\delta g(\nu)}{g(\nu)} \chi_{\mathrm{r}}(\lambda,\nu;g) d\nu. \tag{98}$$

So the variation of (97) gives

$$-\delta^2 \ln \tau(g) = \oint \oint \chi_{\mathrm{r}}(\nu,\lambda;g)\chi_{\mathrm{r}}(\lambda,\nu;g) \frac{\delta' g(\nu)}{g(\nu)} \frac{\delta g(\lambda)}{g(\lambda)} d\nu\, d\lambda. \tag{99}$$

The symmetry of the kernel of second variation with respect to λ, ν implies that the form (97) is closed.

So the formula (97) gives the definition of the τ function in terms of the closed 1-form (and also proves the uniqueness of the τ function satisfying the formula (93)].

For the standard KP coordinates,

$$\frac{\delta g(\lambda)}{g(\lambda)} = \sum_{n=1}^{\infty} \frac{dx_n}{\lambda^n}.$$

This formula allows us to obtain a closed 1-form in terms of dx_n,

$$-\delta \ln \tau(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{\partial^{n-1}}{\partial\nu^{n-1}} \chi_{\mathrm{r}}(\nu,\nu;\mathbf{x})\Big|_{\nu=0} dx_n, \tag{100}$$

which coincides with the expression (80). For $x=x_1$ this formula immediately gives the standard formula

$$\frac{\partial^2}{\partial x^2}\ln\,\tau=\frac{1}{2}u,$$

where u is a solution for the KP equation.

It is possible to prove that the function τ defined as the solution of the relation (97) satisfies the global formula (93) (existence). To do this, we will show, using the formula (97), that the derivatives of the difference of logarithms of the LHS and the RHS of the expression (93) with respect to $\lambda,\mu,\bar{\lambda},\bar{\mu},$ and the variation with respect to $g,$ are equal to zero for arbitrary λ,μ,g. That means that the LHS and RHS of (76) could differ only by the factor, and the normalization of the function χ implies that this factor is equal to 1.

First we will calculate the derivative with respect to λ,

$$\begin{aligned}&\frac{\partial}{\partial\lambda}(\ln\,\chi(\lambda,\mu)(\lambda-\mu)-\ln\,\tau(g\times g_{(\lambda,\mu)})+\ln\,\tau(g))\\&\quad=\frac{\partial}{\partial\lambda}\ln\,\chi(\lambda,\mu)(\lambda-\mu)+\oint\chi_{\mathrm{r}}(\nu,\nu;g\times g_{(\lambda,\mu)})\,\frac{1}{\lambda-\nu}d\nu\\&\quad=\frac{\partial}{\partial\lambda}\ln\,\chi(\lambda,\mu)(\lambda-\mu)+\chi_{\mathrm{r}}(\lambda,\lambda;g\times g_{(\lambda,\mu)}).\end{aligned}\tag{101}$$

The second term can be found in terms of $\chi(\lambda,\mu;g)$ using the determinant formula (91). The formula (91) gives

$$\chi(\lambda,\mu;g\times g_{(\lambda,\mu')})=\frac{(\lambda-\mu')(\mu-\lambda)}{\mu-\mu'}\,\frac{\det\begin{pmatrix}\chi(\lambda,\mu;g) & \chi(\lambda,\mu';g)\\ \frac{\partial}{\partial\lambda}\chi(\lambda,\mu;g) & \frac{\partial}{\partial\lambda}\chi(\lambda,\mu';g)\end{pmatrix}}{\chi(\lambda,\mu';g)}.\tag{102}$$

Taking the regular part of this formula at $\mu=\lambda$, one immediately obtains that the expression (101) is equal to zero.

The case with the derivative over μ is analogous; derivatives over $\bar{\lambda}$ and $\bar{\mu}$ immediately give zero. So now we proceed to the calculation of variation,

$$\begin{aligned}&\delta(\ln\,\chi(\lambda,\mu)(\lambda-\mu)-\ln\,\tau(g\times g_{(\lambda,\mu)})+\ln\,\tau(g))\\&\quad=\frac{1}{\chi(\lambda,\mu)}\oint\chi(\nu,\mu;g)\frac{\delta g(\nu)}{g(\nu)}\chi(\lambda,\nu;g)d\nu\\&\qquad-\oint\chi_{\mathrm{r}}(\nu,\nu;g\times g_{(\lambda,\mu)})\,\frac{\delta g(\nu)}{g(\nu)}d\nu+\oint\chi(\nu,\nu;g)\frac{\delta g(\nu)}{g(\nu)}d\nu\end{aligned}\tag{103}$$

[we have used (98) and (97)]. Using the determinant formula (90) to transform the second term, one concludes that the variation is equal to zero.

Below we will give a brief sketch of the derivation of a 1-form in terms of the Baker–Akhiezer function. Differentiation of (76) (with shifted arguments) with respect to λ and μ gives

$$\left(\frac{\partial}{\partial\lambda}+\frac{\partial}{\partial x_1(\lambda)}\right)\ln[(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})]=-\frac{\partial}{\partial x_1(\lambda)}\ln\,\tau(\mathbf{x}),\tag{104}$$

$$\left(\frac{\partial}{\partial\mu}-\frac{\partial}{\partial x_1(\mu)}\right)\ln[(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})]=\frac{\partial}{\partial x_1(\mu)}\ln\,\tau(\mathbf{x}),\tag{105}$$

where $\partial/\partial x_1(\lambda)=\Sigma_{n=1}^{\infty}\lambda^{n-1}(\partial/\partial x_n)$. Equation (104) implies that

$$\frac{\partial}{\partial x_n}\ln\ \tau(\mathbf{x}) = -\frac{1}{(n-1)!}\frac{\partial^{n-1}}{\partial\lambda^{n-1}}\left(\left(\frac{\partial}{\partial\lambda}+\frac{\partial}{\partial x_1(\lambda)}\right)\ln[(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})]\right)\Bigg|_{\lambda=0}.$$

Hence, we have the closed 1-form ω,

$$\omega(\mathbf{x},d\mathbf{x}) = -\sum_{n=1}^{\infty}\ \frac{1}{(n-1)!}\frac{\partial^{n-1}}{\partial\lambda^{n-1}}\left(\left(\frac{\partial}{\partial\lambda}+\frac{\partial}{\partial x_1(\lambda)}\right)\ln[(\lambda-\mu)\chi(\lambda,\mu,\mathbf{x})]\right)\Bigg|_{\lambda=0} dx_n, \quad (106)$$

and $\omega = d\ \ln\ \tau$. A similar expression for ω can be obtained using (105). At $\mu = 0$ the formula (106) is equivalent to the formula found in Ref. 3.

ACKNOWLEDGMENTS

The first author (L.B.) is grateful to the Dipartimento di Fisica dell'Università and Sezione INFN, Lecce, for hospitality and support; L.B. also acknowledges partial support from the Russian Foundation for Basic Research under Grant No. 96-01-00841. We are also grateful to the referee for helping to improve the paper.

[1] M. Sato, RIMS, Kokyuroku, Kyoto Univ. **439**, 30 (1981) (in Japanese); M. Sato and Y. Sato, in *Nonlinear Partial Differential Equations in Applied Science*, Proceedings of the U.S.–Japan Seminar, Tokyo, 1982 (North-Holland, Amsterdam, 1983).
[2] M. Jimbo and T. Miwa, Publ. RIMS, Kyoto Univ. **19**, 943 (1983).
[3] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, ''Transformation groups for soliton equations,'' in *Nonlinear Integrable Equations—Classical and Quantum Theory*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983).
[4] G. Segal and G. Wilson, Publ. Math. I.H.E.S. **61**, 1 (1985).
[5] P. van Moerbeke, in *Lectures on Integrable Systems*, edited by O. Babelon *et al.* (World Scientific, Singapore, 1994).
[6] M. Adler and P. van Moerbeke, Adv. Math. **108**, 140 (1994).
[7] V. E. Zakharov and S. V. Manakov, Zap. Nauchn. Semin. LOMI **133**, 77 (1984); V. E. Zakharov and S. V. Manakov, Funkt. Anal. Ego Pril. **19**, 11 (1985)(in Russian).
[8] L. V. Bogdanov and S. V. Manakov, J. Phys. A **21**, L537 (1988).
[9] V. E. Zakharov, in *Inverse Methods in Action*, edited by P. C. Sabatier (Springer-Verlag, Berlin, 1990).
[10] B. G. Konopelchenko, *Solitons in Multidimensions* (World Scientific, Singapore, 1993).
[11] R. Carroll and B. Konopelchenko, Lett. Math. Phys. **28**, 307 (1993).
[12] L. V. Bogdanov and B. G. Konopelchenko, J. Phys. A **28**, L173 (1995).
[13] L. V. Bogdanov, Physica D **87**, 58 (1995).
[14] A. I. Zenchuk and S. V. Manakov, Teor. Mat. Fiz. **105**, 371 (1995).
[15] P. G. Grinevich and A. Yu. Orlov, in *Modern Problems of Quantum Field Theory* (Springer-Verlag, Berlin, 1989).
[16] J. Weiss, J. Math. Phys. **24**, 1405 (1983).
[17] F. Nijhoff and J.-M. Maillet, in *Nonlinear Evolution*, Proceedings of IV NEEDS, edited by J. J. P. Leon (World Scientific, Singapore, 1988).
[18] G. Darboux, *Lecons sur les Systémes Orthogonaux et les Coordonneés Curvilignes* (Hermann, Paris, 1910).
[19] L. P. Eisenhart, *Transformation of Surfaces* (Princeton University Press, Princeton, NJ, 1923).
[20] S. P. Tsarev, Math. USSR Izv. **37**, 347 (1991); ''Classical differential geometry and integrability of systems of hydrodynamical type,'' preprint hep-th 9303092, Exeter Workshop, 1992.
[21] B. G. Konopelchenko and W. K. Schief, ''Lamé and Zakharov–Manakov systems: Combescure, Darboux and Bäcklund transformations,'' preprint AM 93/9, UNSW, Sydney, 1993.
[22] M. Jimbo, in *Yang–Baxter Equation in Integrable Systems* (World Scientific, Singapore, 1990).
[23] T. Shiota, Invent. Math. **83**, 333 (1986).
[24] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer-Verlag, Berlin, 1991).

Analytic-bilinear approach to integrable hierarchies. II. Multicomponent KP and 2D Toda lattice hierarchies

L. V. Bogdanov[a] and B. G. Konopelchenko
Consortium EINSTEIN,[b] Dipartimento di Fisica dell'Università di Lecce and Sezione INFN, 73100 Lecce, Italy

(Received 9 June 1997; accepted for publication 4 September 1997)

An analytic-bilinear approach for the construction and study of integrable hierarchies is discussed. Generalized multicomponent KP and 2D Toda lattice hierarchies are considered. This approach allows us to represent generalized hierarchies of integrable equations in a condensed form of finite functional equations. A generalized hierarchy incorporates basic hierarchy, modified hierarchy, singularity manifold equation hierarchy, and corresponding linear problems. Different levels of generalized hierarchy are connected via invariants of Combescure symmetry transformation. The resolution of functional equations also leads to the τ function and addition formulas to it. © *1998 American Institute of Physics.*
[S0022-2488(98)03309-X]

I. INTRODUCTION

The Kadomtsev–Petviashvili (KP) equation and the whole KP hierarchy as well as the multicomponent KP and two-dimensional Toda lattice hierarchies are among the most important ingredients of the theory of integrable partial differential equations. They arise in various fields of physics from hydrodynamics to string theory.

Integrable hierarchies have been described and studied within the framework of different approaches. The Sato approach[1–6] and the $\bar{\partial}$-dressing method (see Refs. 7–9 and a review[10]) are among the most beautiful and powerful among them. These two approaches look completely different. The Sato approach is in essence algebraic with its infinite-dimensional Grassmannian, pseudodifferential operators, the Hirota bilinear identity, and the τ function. In contrast, the $\bar{\partial}$-dressing method, based on the nonlocal $\bar{\partial}$ problem, mostly uses the technique of the complex variable theory.

Recently we proposed (Refs. 11, 12) an approach which combines the characteristic features of both these methods, namely, the Hirota bilinear identity from the Sato approach and analytic properties of the wave functions from the $\bar{\partial}$-dressing method. This analytic-bilinear approach is based on the generalized Hirota bilinear identity for the wave function $\psi(\lambda,\mu)$ with simple analytic properties [the Cauchy–Baker–Akhiezer (CBA) function]. The CBA function [or, to be more precise, its ''nonsingular part'' $\chi(\lambda,\mu)$] is defined in some region G of the complex plane and it is characterized by the generalized Hirota bilinear identity on the boundary of this region.[13] The CBA function arises both in the $\bar{\partial}$-dressing method[8] and in the finite gap integration method,[14] and it can be introduced also in the context of the general Grassmannian approach.[1,6]

According to the generalized Hirota bilinear identity, the CBA function is a functional of the function g defined on the boundary of G (in fact, g is a map of ∂G to $\mathbb{C}^{\times}$ (i.e., $\mathbb{C}-0$), or, in other words, it is an element of the loop group). One can choose some special subgroups of the generic loop group parametrized by a finite number of variables (''times'') to get particular examples of integrable equations. This is a standard approach in the $\bar{\partial}$-dressing method. One may also try to use some generic loop group and extract a full infinite set of equations. So the notion of the hierarchy arises. The dependence of the CBA function on g encodes the dynamics of the system. The hierarchy can be written down in variational form, i.e., as equations describing variations of

[a]Permanent address: L. D. Landau Institute for Theoretical Physics, Kosygin str. 2, Moscow 117940, GSP1, Russia; electronic mail: Leonid@landau.ac.ru
[b]European Institute for Nonlinear Studies via Transnationally Extended Interchanges.

 © 1998 American Institute of Physics

the objects of theory under the arbitrary variation of group element g. To get a hierarchy in terms of times, one should introduce a parametrization of the loop group.

The distinctive feature of an analytic-bilinear approach is that our primary object is the CBA function and hierarchies are written in the form of equations for the CBA function (we call these equations *generalized* hierarchies). The analytic-bilinear approach allows us to describe generalized hierarchies in a compact finite form. Such a generalized hierarchy contains a basic integrable hierarchy, a hierarchy of equations for wave functions (modified hierarchy), and a hierarchy of equations for CBA-type wave functions (we call it the singularity manifold equations (SM) hierarchy, because in the scalar KP case it coincides with the singularity manifold equations obtained from the Painleve analysis of KP; however, our context is completely different), and corresponding linear problems. The relation between three hierarchies of equations can be described in terms of invariants of the Combescure symmetry group (Refs. 11, 12).

The generalized Hirota bilinear identity provides various functional equations for the CBA function. The resolution of some of them leads to the introduction of the τ function, while others give rise to the addition theorems for the τ function. The properties of the τ function, such as the associated closed 1-form and its global definition also arise in a simple manner within the analytic-bilinear approach.

The first part of this paper[12] was devoted to the scalar KP hierarchy. In this case G is just a domain of the complex plane. Without the loss of generality, we choose G as a unit disc. The group element g is defined on the unit circle and can be represented in a form

$$g(\lambda)=\lambda^n g_{\text{in}}\times g_{\text{out}},$$

where n is the index of the function, g_{in} and g_{out} are functions analytic and having no zeros, respectively, inside (outside) the unit circle on the complex plane (i.e., elements of the loop groups Γ_- and Γ_+). In the scalar KP hierarchy case, we have no opportunity to treat the index variable without some additional construction, so we put the index to zero. The functions g_{out} and g_{in} are parametrized in a standard way by infinite sets of ''positive'' and ''negative'' times. But it appears that in the scalar KP hierarchy case the dependence of the objects of theory on g_{in} (i.e., on negative times in our notations) is trivial, so the scalar KP hierarchy is defined effectively by the subgroup Γ_+ of the loop group Γ.

In the second part of this paper we consider the generalized multicomponent KP hierarchy and the two-dimensional Toda lattice hierarchy (for other approaches to the usual multicomponent KP hierarchy and 2D Toda lattice hierarchy see Refs. 1, 2, 4, 5, 15, 16).

There are two equivalent approaches to investigating a multicomponent hierarchy. One could start with the generalized Hirota identity for the matrix-valued CBA function. In this case G is again a unit disc, and g is a diagonal matrix function on the unit circle (i.e., a set of N functions g_n). Another way is to start with the scalar CBA function defined in the region G consisting of N disconnected domains (unit discs for simplicity). In this case the dynamics is also defined by the set of N functions g_n, and the CBA function consists, in fact, of $N\times N$ disconnected components. These two representations are exactly equivalent and are associated with the multicomponent loop group Γ^N. In our paper we will use both of them.

As in the scalar case, the objects of multicomponent KP theory depend trivially on the subgroup Γ_-^N. The difference with the scalar case consists of the presence of discrete variables (indices), though only the lattice vectors with a zero general index are admissible.

Since the generalized Hirota bilinear identity is our primary object, it is quite natural to proceed from the hierarchy corresponding to the simplest choice of G—the unit disk—to the hierarchies corresponding to G consisting of several unit discs, i.e., from the scalar KP hierarchy to the multicomponent KP hierarchy. But on the other hand, the loop group is also a fundamental object of the theory, so one would like to find a hierarchy with a full-scale action of the generic loop group on the objects of theory. This hierarchy is known as the Toda lattice hierarchy. In our approach it can be introduced as a two-component KP hierarchy with the special choice of G consisting of two discs with the common boundary (the unit circle): one disc around the origin and one disc around infinity. In this case, up to some trivial dynamics, one can glue two special loops defining the two-component KP hierarchy to obtain a generic loop. In the case of a Toda lattice

hierarchy the discrete index variable, positive and negative time variables, get their natural interpretation.

We first consider the multicomponent KP hierarchy. We start with the generalized Hirota identity for the matrix-valued CBA function and derive various finite functional equations for this function. Part of these functional equations present the compact form of the generalized multicomponent KP hierarchy which includes equations for potentials (the standard multicomponent KP hierarchy), equations for the wave functions (the modified multicomponent KP hierarchy), equations for CBA-type wave functions (singularity manifold equation hierarchy), and corresponding linear problems.

Resolution of functional equations introduces the set of τ functions. We discuss also the associated closed 1-form and its global definition. The simplest integrable systems are given by the resonantly interacting wave equations and by the scalar Darboux system which is its modified partner.

We present a compact form of *equations* of multicomponent KP hierarchy, the modified KP hierarchy, and the singularity manifold equation hierarchy, using shift operators. Similar formulas for the first two cases can be found in Ref. 17.

The two-component case which corresponds to the generalized Davey–Stewartson (DS) hierarchy is considered in detail.

The description of the generalized Toda lattice hierarchy within the analytic-bilinear approach requires minor modifications of the multicomponent KP case. For simplicity we consider here only the Abelian Toda lattice hierarchy. It appears that a modified Toda lattice equation (i.e., the equation for the Toda lattice wave functions) coincides with the 2D Volterra system,[18] which was recently discussed in the literature in different contexts.[19,20] We derive a 2D Toda lattice singularity manifold equation, which was also discussed in Ref. 21. The invariants of the Combescure transformations group (Refs. 11, 12) are found. We also present the hierarchies of equations using shift operators.

The paper is organized as follows. The generalized Hirota bilinear identity and some basic formulas are discussed in Sec. II. A generalized multicomponent KP hierarchy is considered in Sec. III. Compact forms of equations of generalized multicomponent KP hierarchy are presented in Sec. IV. Section V is devoted to the generalized DS hierarchy. The τ function for the DS hierarchy is introduced in Sec. VI. The τ function and corresponding addition formulas for the multicomponent KP hierarchy are considered in Sec. VII. A closed variational 1-form and the global definition for the τ function are given in Sec. VIII. The basic equations for the generalized 2D Toda lattice hierarchy are derived in Sec. IX. Concrete equations associated with the generalized 2DTL hierarchy are presented in Sec. X. Addition formulas for the 2DTL hierarchy are given in Sec. XI.

II. GENERALIZED HIROTA IDENTITY, FUNCTIONAL EQUATIONS, AND GENERALIZED MULTICOMPONENT KP HIERARCHY

The analytic-bilinear approach starts with the generalized Hirota bilinear identity,

$$\int_{\partial G}\chi(\nu,\mu;g_1)g_1(\nu)g_2^{-1}(\nu)\chi(\lambda,\nu;g_2)d\nu=0, \tag{1}$$

where $\chi(\lambda,\mu;g)$ is the CBA function normalized by the condition $\chi\to(\lambda-\mu)^{-1}$ as $\lambda\to\mu$ and $\chi(\lambda,\mu)$ is an analytic function of both variables λ,μ for $\lambda\neq\mu$ in the domain G with the boundary ∂G. The quantity $g=g(\lambda,\boldsymbol{x})$ specifies the dependence on the infinite set of dynamical variables $\boldsymbol{x}=(x_1,x_2,\dots)$. In terms of the wave function $\psi(\lambda,\mu;g)$ defined via

$$\psi(\lambda,\mu;g)=g^{-1}(\mu)\chi(\lambda,\mu;g)g(\lambda), \tag{2}$$

the generalized Hirota bilinear identity looks like

$$\int_{\partial G}\psi(\nu,\mu;g_1)\psi(\lambda,\nu;g_2)d\nu=0. \tag{3}$$

Note that the generalized Hirota bilinear identity is bilocal both in dynamical and spectral variables.

In this paper we consider the case for which χ and ψ take their value in the full linear group $\mathrm{GL}(N,\mathbb{C})$, G is a unit disc with a center $\lambda=0$. Elements $g(\lambda,\boldsymbol{x})$ belong to maximal Abelian subgroup of $\mathrm{GL}(N,\mathbb{C})$.

It is important to note that there is an alternative way to treat the matrix problem (1), starting from the *scalar* identity but with the region G consisting of several disconnected domains. In this case one has to choose domain G consisting of N unit discs D_α with the center at $\lambda=0$. Now the scalar CBA function effectively consists of $N\times N$ disconnected components and can be interpreted as a matrix function defined on the unit disc:

$$\chi_{\alpha\beta}(\lambda,\mu)=\chi(\lambda,\mu),\quad \lambda\in D_\beta,\quad \mu\in D_\alpha. \tag{4}$$

The function g in this framework is a set of N functions defined on the unit circle. Thus, rewriting the scalar identity in matrix notations, we get *exactly* the same matrix generalized Hirota identity, where g is a diagonal matrix.

So the matrix structure in the scalar case is hidden in the multicomponent domain of the definition of the CBA function. The advantage of the scalar treatment is the opportunity to preserve most of the structure of the one-component KP hierarchy case almost intact and to use geometric intuition in the matrix case. We will use both representations of the generalized Hirota identity in the paper.

The dependence on dynamical variables is hidden in the set of N functions g defined on the unit circle. The most natural choice would be to treat sets of arbitrary functions on the circle having no zeros, forming the group by multiplication. Arbitrary g can be represented in the form

$$g(\lambda)=\lambda^n g_{\mathrm{in}}\times g_{\mathrm{out}},$$

where discrete variable n is the index of the function, g_{in} and g_{out} are functions analytic and having no zeros, respectively, inside (outside) the unit circle on the complex plane. We will use the following parametrization for g_{in} and g_{out}:

$$g_{\alpha\ \mathrm{out}}=\exp\left(\sum_{n=1}^{\infty} x_{(\alpha)n}\lambda^{-n}\right), \tag{5}$$

$$g_{\alpha\ \mathrm{in}}=\exp\left(\sum_{-\infty}^{n=-1} x_{(\alpha)n}\lambda^{-n}\right). \tag{6}$$

In these formulas we have omitted ''the zero time'' x_0, that defines simply some constant factor in front of g and, according to (1), it does not change the function χ.

So, in the general case the set of functions g_α can be parametrized by N discrete times n_α and N infinite sets of continuous variables $x_{(\alpha)n}$, $n\in\mathbb{Z}$. But in the case of the KP hierarchy the dependence of the function $\chi(\lambda,\mu;g)$ on g_{in} (in other words, on the ''negative times'') is trivial; according to (1) (the ''scalar'' form), we get

$$\chi(\lambda,\mu;g\times g_{\mathrm{in}})=g_{\mathrm{in}}^{-1}(\mu)\chi(\lambda,\mu;g)g_{\mathrm{in}}(\lambda). \tag{7}$$

It occurs that other important objects (KP solution, τ function) also depend on negative times trivially. Thus, for simplicity, from the very beginning we omit the dependence on the negative times.

As for discrete times, we will see that the case of multicomponent KP hierarchy allows only sets of discrete times with the zero general index ($\sum_{\alpha=1}^{N} n_\alpha=0$).

So the case of multicomponent KP hierarchy seems a bit asymmetric with respect to positive and negative times, and it also does not allow us to consider arbitrary lattice vectors of discrete times. The most natural way to resolve these problems is to consider the multicomponent Toda-lattice hierarchy (see below). The objects of the Toda lattice hierarchy have a natural dependence on positive, negative, and discrete times.

To find special solutions for (1), one may use some additional construction (e.g., a $\bar{\partial}$-dressing method or algebro-geometric technique). But it is possible to develop a procedure of solving (1) in terms of (1) itself. In this case the function $\chi(\lambda,\mu,g=1)=\chi_0(\lambda,\mu)$ should be treated as initial data, and one can start from arbitrary $\chi_0(\lambda,\mu)$ possessing the necessary analytical properties. Then the dynamics of the function $\chi(\lambda,\mu)$ can be defined via the variational form of identity (1),

$$\delta\chi(\lambda,\mu;g)=\oint \chi(\nu,\mu;g)\frac{\delta g(\nu)}{g(\nu)}\chi(\lambda,\nu;g)d\nu. \tag{8}$$

So the problem of solving (1) is reduced to (infinite number in terms of variables x_n) differential equations with an initial condition. Taking the second variation in $\delta(\ln g)$, it is easy to prove that this system is self-consistent (i.e., the cross-derivatives are equal), and it is a correctly defined variational equation for the function $\chi(\lambda,\mu,g)$. Going further, one could take higher variational derivatives (all of them have a very simple form) and represent a formal solution in terms of a Taylor-type series.

Another approach to solving (1) is to extract from it some linear integral equations. Indeed, let us choose some initial condition $\chi_0(\lambda,\mu)$. Then for the function $\chi(\lambda,\mu,g)$ from (1) we have a pair of equations,

$$\int_{\partial G}\chi(\nu,\mu;g)g(\nu)\chi_0(\lambda,\nu)d\nu=0, \tag{9}$$

$$\int_{\partial G}\chi_0(\nu,\mu)g^{-1}(\nu)\chi(\lambda,\nu;g)d\nu=0. \tag{10}$$

These relations can be interpreted as linear integral equations for the function $\chi(\lambda,\mu;g)$; integral equations of such a type arise in the context of a nonlocal Riemann problem. Using the technique developed in Ref. 13, one may show that if both of these equations are solvable [in the class of functions $\chi(\lambda,\mu)$ with the analytical properties specified above, here they play a role of *normalization*], then the solution for both equations is the same and it is unique.

There is a very important class of functions g corresponding to some special discrete transformations—meromorphic functions—for which the dynamics can be found explicitly. Namely, the solution for both equations (9), (10) exists (and it is unique), if and only if the meromorphic function g has an equal number of zeros and poles in G (taking into account the multiplicity), and it is given by the determinant formula.[13] This statement means that the dependence of χ on discrete (index) variables is correctly defined for the *zero general index case* and, moreover, it can be represented in explicit form.[13]

III. GENERALIZED MULTICOMPONENT KP HIERARCHY

Proceeding with the multicomponent KP case, we choose the following parametrization for g_{out} (matrix representation):

$$g(\lambda,\boldsymbol{x})=\exp\left[\sum_{\alpha=1}^{N}\sum_{n=1}^{\infty}\frac{E_\alpha}{\lambda^n}x_{(\alpha)n}\right], \tag{11}$$

where the matrices E_α form a basis of the commutative subalgebra of diagonal matrices: $(E_\alpha)_{\beta\gamma}=\delta_{\alpha\beta}\delta_{\beta\gamma}$ $(\alpha,\beta,\gamma=1,\ldots,N)$. So we have an N infinite series of dynamical variables $x_{(\alpha)n}$.

Equations (1) or (3) with g of the form (11) generically define the multicomponent KP hierarchy. To extract from them finite functional equations similar to the scalar case we consider the Hirota identity (1) for the dynamical variables $x'_{(\alpha)n}$ and $x_{(\alpha)n}$ related in a special way, namely,

$$x'_{(\alpha)n}-x_{(\alpha)n}=\frac{1}{n}a_\alpha^n-\frac{1}{n}b_\alpha^n, \tag{12}$$

where a_α, b_α $(0\geqslant\alpha\geqslant N)$ are arbitrary parameters. We will also use vector notations $\boldsymbol{x}=(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(N)})$, $\boldsymbol{x}_{(\alpha)}=(x_{(\alpha)1},x_{(\alpha)2},\ldots)$. In these terms the preceding formula looks like

$$\boldsymbol{x}'_{(\alpha)}-\boldsymbol{x}_{(\alpha)}=[a_\alpha]-[b_\alpha], \tag{13}$$

or

$$\boldsymbol{x}'-\boldsymbol{x}=[\boldsymbol{a}]-[\mathbf{b}], \tag{14}$$

where $[\boldsymbol{a}]_{(\alpha)i}=(1/i)\,a_\alpha^i$. For such x and x' due to the formula $\log(1-\epsilon)=-\Sigma_{n=1}^{\infty}(\epsilon^n/n)$ one has

$$(g_1(\nu)g_2^{-1}(\nu))_{\alpha\beta}=\delta_{\alpha\beta}\frac{\nu-a_\alpha}{\nu-b_\alpha},\quad 0\geqslant\alpha,\quad \beta\geqslant N. \tag{15}$$

Substitution of the expression (15) into (1) gives

$$\left(\frac{\mu-a_\alpha}{\mu-b_\alpha}\right)\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(\frac{\lambda-a_\beta}{\lambda-b_\beta}\right)\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])$$
$$+\sum_{\gamma=1}^{N}(b_\gamma-a_\gamma)\chi_{\gamma\beta}(\lambda,b_\gamma,\boldsymbol{x}+[\boldsymbol{a}])\chi_{\alpha\gamma}(b_\gamma,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0. \tag{16}$$

Equation (16) represents the final functional form of the generalized multicomponent KP hierarchy. It encodes the usual multicomponent KP hierarchy and its modified version together with their linear problems. To demonstrate this fact we, similar to the scalar case,[12] consider the particular form of equation (16) with $b_\alpha=0$, $0\geqslant\alpha\geqslant N$,

$$\left(1-\frac{a_\alpha}{\mu}\right)\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(1-\frac{a_\beta}{\lambda}\right)\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x})=\sum_{\gamma=1}^{N}a_\gamma\chi_{\gamma\beta}(\lambda,0,\boldsymbol{x}+[\boldsymbol{a}])\chi_{\alpha\gamma}(0,\mu,\boldsymbol{x})=0, \tag{17}$$

or, in terms of the CBA and Baker–Akhiezer functions,

$$\psi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\psi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x})=\sum_{\gamma=1}^{N}a_\gamma\psi_{\gamma\beta}(\lambda,\boldsymbol{x}+[\boldsymbol{a}])\tilde{\psi}_{\alpha\gamma}(\mu,\boldsymbol{x}), \tag{18}$$

where $\psi_{\gamma\beta}(\lambda,\boldsymbol{x})=\chi(\lambda,0,\boldsymbol{x})g_\beta(\lambda)$, $\tilde{\psi}_{\gamma\beta}(\mu,\boldsymbol{x})=\chi(0,\mu,\boldsymbol{x})g_\gamma^{-1}(\mu)$. The expansion of (18) in the Taylor series in $a_{(\gamma)}$ gives rise in the first order to the following equation:

$$\frac{\partial\psi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x})}{\partial x_{(\gamma)1}}=\psi_{\gamma\beta}(\lambda,\boldsymbol{x})\tilde{\psi}_{\alpha\gamma}(\mu,\boldsymbol{x}). \tag{19}$$

Evaluating the equation (17) with $\alpha\neq\beta\neq\gamma\neq\alpha$ at $\lambda=\mu=0$, one obtains the system

$$\frac{\partial P_{\alpha\beta}}{\partial\xi_\gamma}=P_{\alpha\gamma}P_{\gamma\beta}, \tag{20}$$

where $P_{\alpha\beta}=\chi_{\alpha\beta}(0,0,\boldsymbol{x})$, $\xi_\gamma=x_{(\gamma)1}$, which is nothing but the Darboux system in terms of rotation coefficients[22] or resonantly interacting waves equation.[7] Equation (17) with $\alpha\neq\gamma$ and $\mu=0$ gives the system

$$\frac{\partial\Psi_\alpha}{\partial\xi_\gamma}=P_{\alpha\gamma}\Psi_\gamma, \tag{21}$$

where

$$\Psi_\alpha = \sum_{\beta=1}^{N} \oint d\lambda \; f_\beta(\lambda)\psi_{\alpha\beta}(\lambda, \boldsymbol{x}), \tag{22}$$

and $f_\beta(\lambda)$ are arbitrary functions. The system (21) is a well-known linear problem for equations (20).

Equation (17) with $\beta \neq \gamma$ and $\lambda = 0$ gives the adjoint linear problem for the system (20),

$$\frac{\partial \tilde{\Psi}_\beta}{\partial \xi_\gamma} = P_{\gamma\beta}\tilde{\Psi}_\gamma, \tag{23}$$

where $\tilde{\Psi}_\beta = \Sigma_{\alpha=1}^{N} \oint d\mu \tilde{f}_\alpha(\mu)\tilde{\psi}_{\alpha\beta}(\mu, \boldsymbol{x})$ and $\tilde{f}_\alpha(\mu)$ are arbitrary functions.

Further, (21) implies that $P_{\alpha\gamma} = (1/\Psi_\gamma)(\partial\Psi_\alpha/\partial\xi_\gamma)$. Substituting this expression into (20), one gets

$$\partial_\beta\partial_\gamma\Psi_\alpha = \Psi_\beta^{-1}(\partial_\gamma\Psi_\beta)(\partial_\beta\Psi_\alpha) + \Psi_\gamma^{-1}(\partial_\beta\Psi_\gamma)(\partial_\gamma\Psi_\alpha), \tag{24}$$

where $\partial_\beta = \partial/\partial\xi_\beta$. This is a well-known Darboux–Zakharov–Manakov system which has been discovered in the theory of systems of conjugate surfaces first[22] and then has been rediscovered within the $\bar{\partial}$-dressing method in Ref. 7. This system is a modified partner of the system (20). To get the linear problem for it, it is sufficient to substitute the expression $P_{\alpha\gamma} = (1/\Psi_\gamma)(\partial\Psi_\alpha/\partial\xi_\gamma)$ into (23).

Thus the system (17) contains the equation (20), its modified partner (24), and their linear problems.

The higher order terms in the expansion of (17) in a_γ give rise to the higher analogs of equation (19). It is a straightforward check that these equations contain the higher order equations (20), their modified analogs, and corresponding linear problems.

The generalized multicomponent KP hierarchy described by functional equation (16) has a rather special structure. In particular, it contains N independent generalized scalar KP hierarchies. They are obviously given by equations (16) with $\alpha = \beta$ and $b_\gamma = a_\gamma$ for all γ except $\gamma = \alpha$. It is also quite clear that the number of effective independent dynamical variables is equal to three. For instance, the systems (20) and (24) are overdetermined and represent a collection of commuting closed systems defined for every triad of dynamical variables.

Note also that one can get the multicomponent KP hierarchy starting with the scalar generalized Hirota identity (1). In this case one has to choose a disconnected domain G consisting of N unit discs D_α with the center at $\lambda = 0$. In this case the scalar CBA function effectively consists of $N \times N$ disconnected components and can be interpreted as a matrix function. So the matrix structure in the scalar case is hidden in the multicomponent domain of definition of the CBA function. The function $g(\lambda)$ is defined on the set of N unit circles S_α and consists in fact of N components g_α defined on a unit circle.

The advantage of the scalar treatment is the opportunity to preserve most of the structure of the one-component KP hierarchy case almost without modification.

Using the parametrization (5), we represent the function $g(\lambda)$ (defined on the set of N unit circles S_α) as

$$g(\boldsymbol{x}, \lambda) = \prod_{\alpha=1}^{N} \exp\left(\sum_{i=1}^{\infty} x_{(\alpha)i} K_\alpha^i \right), \tag{25}$$

where

$$K_\alpha(\lambda) = \frac{1}{\lambda}, \quad \lambda \in D_\alpha;$$

$$K_\alpha(\lambda) = 0, \quad \lambda \notin D_\alpha.$$

In this case the transformation group element g_0 providing just the same shift of the $\boldsymbol{x}$ variables takes the form

$$g_0(\nu)=(g_1(\nu)g_2^{-1}(\nu))=\frac{\nu-a_\alpha}{\nu-b_\alpha},\quad \nu\in D_\alpha. \tag{26}$$

The analog of equation (16) reads as

$$g_0(\mu)\chi(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-g_0(\lambda)\chi(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])$$
$$+\sum_{\gamma=1}^{N}(b_\gamma-a_\gamma)\chi(\lambda,b_{(\gamma)},\boldsymbol{x}+[\boldsymbol{a}])\chi(b_\gamma,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0. \tag{27}$$

It seems to be a scalar equation, but if we take into account that the function χ consists of a set of disconnected components and denote $\chi_{\alpha\beta}(\lambda,\mu)=\chi(\lambda,\mu)$, $\lambda\in D_\beta$, $\mu\in D_\alpha$, then this equation gives rise exactly to the system (16).

The analog of equation (17) is, respectively,

$$\prod_{\alpha=1}^{N}(1-a_\alpha K_\alpha(\mu))\chi(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\prod_{\alpha=1}^{N}(1-a_\alpha K_\alpha(\lambda))\chi(\lambda,\mu,\boldsymbol{x})$$
$$=\sum_{\gamma=1}^{N}a_{(\gamma)}\chi_\gamma(\lambda,\boldsymbol{x}+[\boldsymbol{a}])\tilde{\chi}_\gamma(\mu,\boldsymbol{x})=0, \tag{28}$$

and in terms of CBA and BA functions one gets

$$\psi(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\psi(\lambda,\mu,\boldsymbol{x})=\sum_{\gamma=1}^{N}a_\gamma\psi_\gamma(\lambda,\boldsymbol{x}+[\boldsymbol{a}])\tilde{\psi}_\gamma(\mu,\boldsymbol{x}), \tag{29}$$

where we have introduced the notations

$$\begin{aligned}
\chi_\gamma(\lambda,\boldsymbol{x})&=\chi(\lambda,0_\gamma,\boldsymbol{x}),\\
\tilde{\chi}_\gamma(\mu,\boldsymbol{x})&=\chi(0_\gamma,\mu,\boldsymbol{x}),\\
\psi_\alpha(\lambda,\boldsymbol{x})&=g(\lambda)\chi_\alpha(\lambda,\boldsymbol{x}),\\
\tilde{\psi}_\alpha(\mu,\boldsymbol{x})&=g^{-1}(\mu)\tilde{\chi}_\alpha(\mu,\boldsymbol{x});
\end{aligned} \tag{30}$$

0_α denotes the center point of the disc D_α.

We will also use the following notations, most of them exactly corresponding to notations introduced in the frame of matrix representation: for potentials,

$$\chi_{\alpha\beta}(\boldsymbol{x})=\chi_{\mathrm{r}}(0_\beta,0_\alpha,\boldsymbol{x}),$$

$$\chi_{\mathrm{r}}(\lambda,\mu,\boldsymbol{x})=\chi(\lambda,\mu,\boldsymbol{x})-\delta_{\alpha\beta}\frac{1}{\lambda-\mu},\quad \lambda\in D_\alpha,\quad \mu\in D_\beta;$$

for wave functions,

$$\begin{aligned}
\Phi(\boldsymbol{x})&=\int\int_{\partial G}d\lambda\ d\mu\ f(\lambda)\tilde{f}(\mu)\psi(\lambda,\mu,\boldsymbol{x}),\\
\Psi_\alpha(\boldsymbol{x})&=\int_{\partial G}d\lambda\ f(\lambda)\psi_\alpha(\lambda,\boldsymbol{x}),\\
\tilde{\Psi}_\alpha(\boldsymbol{x})&=\int_{\partial G}d\mu\ \tilde{f}(\mu)\tilde{\psi}_\alpha(\mu,\boldsymbol{x}),
\end{aligned} \tag{31}$$

where $f(\lambda),\tilde{f}(\mu)$ are arbitrary functions defined on the boundary of G (i.e., on N unit circles).

IV. EQUATIONS OF GENERALIZED KP HIERARCHY IN COMPACT FORM

The basic equation of the multicomponent KP hierarchy (17) looks very simple and contains all the information about generalized hierarchy—the hierarchy of linear problems and the hierarchy of equations. However, sometimes this information could be excessive and one would prefer to have a compact form of the *equations* of the hierarchy. It is possible to extract all the hierarchies of equations from (17) in terms of shift operators. To derive the equations of the KP hierarchy, we will use generic shifts of times. For each triad of different shifts (i.e., shifts with different parameters) it is possible to derive a closed system of equations, representing, in fact, the multicomponent KP hierarchy.

As we will see later, the multicomponent KP hierarchy in a general case incorporates also discrete times not reducible to the shifts of continuous times. So equations of a multicomponent KP hierarchy can also contain these essentially discrete times. It is possible to include them to equations of this section, but now we will derive the equations of KP hierarchy containing only continuous times. Discrete times will be discussed later in the context of the Toda lattice hierarchy.

Thus, we will derive a compact form of the equation of the multicomponent KP hierarchy, the modified KP hierarchy, and the singularity manifold equation hierarchy. It is interesting to note that these equations are obtained in fact in the form of lattice equations, and all the hierarchies are given by the expansion of these lattice equations with respect to the parameters of the lattice.

In this section we will use the matrix form of the equation (17), i.e.,

$$\Delta\chi(\lambda,\mu)+\frac{1}{\lambda}\chi(\lambda,\mu)A-\frac{1}{\mu}A\cdot T\chi(\lambda,\mu)=\chi(0,\mu)AT\chi(\lambda,0). \tag{32}$$

Here the common argument $\boldsymbol{x}$ is omitted, the function $\chi(\lambda,\mu,\boldsymbol{x})$ is $N\times N$ matrix-valued function, A is a diagonal matrix, $A_{pq}=\delta_{pq}a_q$, and we have introduced the shift operator and difference operator,

$$T\chi(\lambda,\mu,\boldsymbol{x})=\chi(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}]),\quad \boldsymbol{a}=(a_1,\ldots,a_N),$$

$$\Delta\chi(\lambda,\mu,\boldsymbol{x})=(T-1)\chi(\lambda,\mu,\boldsymbol{x}).$$

If one takes the first term of expansion of the relation (32) at $\lambda=0$, $\mu=0$, one obtains an identity containing the potential $u=\chi_{\mathrm{r}}(0,0,\boldsymbol{x})$ and two higher terms of expansion of $\chi_{\mathrm{r}}(\lambda,\mu,\boldsymbol{x})$. It is possible to cancel these terms by combining three equations (32) with different matrices A_k, $k\in\{1,2,3\}$ and get an equation for the potential u.

First, it is easy to cancel the terms containing A_i/λ combining pairs of equations,

$$\Delta_i\chi(\lambda,\mu)A_j-\Delta_j\chi(\lambda,\mu)A_i+\frac{1}{\mu}\left(A_iT_i\chi(\lambda,\mu)A_j-A_jT_j\chi(\lambda,\mu)A_i\right)$$
$$=\chi(0,\mu)A_iT_i\chi(\lambda,0)A_j-\chi(0,\mu)A_jT_j\chi(\lambda,0)A_i. \tag{33}$$

And then, multiplying this equation by the operator A_kT_k from the left, taking the sum by different permutations of indices i,j,k, and using the first term of expansion by λ,μ at $\lambda=0$, $\mu=0$, one obtains the equation

$$\sum_{(ijk)}\epsilon_{ijk}A_kT_k(\Delta_iu-uA_iT_iu)A_j=0, \tag{34}$$

where $u=\chi_{\mathrm{r}}(0,0,\boldsymbol{x})$, and the summation goes over different permutations of indices. This extremely simple equation generates all equations of multicomponent KP hierarchy (not containing essentially discrete variables). A similar equation can be found in Ref. 17.

The scalar case of (34) represents a scalar KP hierarchy in the form of the superposition formula for three Bäcklund transformations. To simplify the derivation of equations of the KP

hierarchy from this equation, it is possible to take the first order of expansion in terms of the parameter a_1, performing, in fact, a transition $\Delta_1 \to \partial/\partial x_1$, $T_1 \to 1$ and thus distinguishing the continuous variable $x := x_1$ (obtaining, in fact, the superposition formula for *two* Bäcklund transformations).

It is easy to check that the choice of three elementary shifts $T_i : \boldsymbol{x}_{(i)} \to \boldsymbol{x}_{(i)} + [a_i]$ (only the variables $\boldsymbol{x}_{(i)}$ are shifted) in the formula (34) in the first nontrivial order of expansion $(a_i a_j a_k)$ leads to the standard form of the three waves interaction equation.

Using cross-differences of equations (32), it is possible to obtain equations for the wave functions (''modified equations''),

$$\sum_{(ijk)} \epsilon_{ijk} A_k T_k(\tilde{\Psi}^{-1} \cdot (T_i \tilde{\Psi})) A_j = 0. \tag{35}$$

Equations of this type for the DS case and N-wave equations case were derived in Ref. 23; see also the lattice form in Ref. 17.

And finally, it is easy to check using (32) that Φ satisfies the equation

$$(T_1\Delta_3\Phi)\cdot(T_3\Delta_1\Phi)^{-1}\cdot(T_3\Delta_2\Phi)\cdot(T_2\Delta_3\Phi)^{-1}\cdot(T_2\Delta_1\Phi)\cdot(T_1\Delta_2\Phi)^{-1} = 1. \tag{36}$$

This symmetric equation represents a multicomponent singularity manifold equation hierarchy. In the scalar case it reads as

$$(T_2\Delta_1\Phi)(T_3\Delta_2\Phi)(T_1\Delta_3\Phi) = (T_2\Delta_3\Phi)(T_3\Delta_1\Phi)(T_1\Delta_2\Phi). \tag{37}$$

We would like to remind the reader here of the main points of the concept of the Combescure symmetry transformation. The term Combescure symmetry transformation originates from geometry, where it is used for the transformation of surface, keeping all tangent vectors parallel.[22] For the integrable Darboux system describing triply conjugate systems of surfaces this transformation has a very simple interpretation in the contexts of theory of integrable systems, and in fact a general class of integrable equations (''modified'' equations) possesses this type of transformation (though the geometrical interpretation in this case may be missing). Our concept of the Combescure transformation is inspired mainly by integrability, not geometry (see Ref. 11).

Integrable equations for the wave functions (i.e., modified equations and singularity manifold type equations) possess the symmetry transformations corresponding to the freedom of choice of the wave function for the same CBA function. In spectral terms this symmetry is described by the change of weight functions $\rho(\lambda)$, $\eta(\mu)$. In terms of equations themselves this symmetry is represented using the Combescure invariants. The solutions are connected by the Combescure symmetry transformation, if Combescure invariants for them are the same. The Combescure symmetry group consists of two subgroups (''left'' and ''right''), corresponding to the change of one weight function [$f(\lambda)$ or $\tilde{f}(\mu)$; see notations (31); in our case $f(\lambda)$ and $\tilde{f}(\mu)$ are matrix-valued functions, multiplication by $f(\lambda)$ is from the right, by $\tilde{f}(\mu)$ is from the left, and here we will use the terms ''left'' and ''right'' Combescure invariants according to the rules of matrix multiplication]. Singularity manifold-type equations possess three types of invariants: left, right, and complete invariants. Modified equations possess only one type of invariant (left or right) and are invariant under an opposite Combescure subgroup. In this sense the basic hierarchy is invariant under the action of the full Combescure group.

We will write down some of the Combescure invariants of equations (35), (36). Combescure invariants for the equation (35) read as

$$C_{ij} = (\tilde{\Psi})^{-1}(\Delta_i \tilde{\Psi} A_j - \Delta_j \tilde{\Psi} A_i). \tag{38}$$

The left Combescure invariants for the equation (36) are given by

$$C'_{ij} = (\Delta_i \Phi)(\Delta_j \Phi)^{-1}. \tag{39}$$

V. GENERALIZED DS HIERARCHY

Now we will consider the 2-component case of the multicomponent KP hierarchy in more detail. In this section we derive the DS hierarchy and the hierarchy of corresponding linear problems, the mDS hierarchy and linear problems, and in the end we present the compact form of the hierarchy of Combescure transformation invariants for the mDS hierarchy. To perform all the derivations in terms of hierarchy, we use the formulas containing two continuous times ($x_1 := x_{(1)1}$ and $x_2 := x_{(2)1}$ and discrete shifts $[a_1]$ and $[a_2]$, the first acting on $\boldsymbol{x}_{(1)}$, the second on $\boldsymbol{x}_{(2)}$). These shifts change all the times of the hierarchy, and the expansion in a_1, a_2 will generate the hierarchy in terms of partial differential equations.

To derive DS hierarchy and linear problems, we will use the following special cases of the formula (28):

$$\frac{1}{a_1}\Delta_1\chi(\lambda,\mu,\boldsymbol{x})+K_1(\lambda)\chi(\lambda,\mu,\boldsymbol{x})-K_1(\mu)T_1\chi(\lambda,\mu,\boldsymbol{x})=\tilde{\chi}_1(\mu,\boldsymbol{x})T_1\chi_1(\lambda,\boldsymbol{x}), \tag{40}$$

$$\frac{1}{a_2}\Delta_2\chi(\lambda,\mu,\boldsymbol{x})+K_2(\lambda)\chi(\lambda,\mu,\boldsymbol{x})-K_2(\mu)T_2\chi(\lambda,\mu,\boldsymbol{x})=\tilde{\chi}_2(\mu,\boldsymbol{x})T_2\chi_2(\lambda,\boldsymbol{x}), \tag{41}$$

$$\partial_1\chi(\lambda,\mu,\boldsymbol{x})+(K_1(\lambda)-K_1(\mu))\chi(\lambda,\mu,\boldsymbol{x})=\tilde{\chi}_1(\mu,\boldsymbol{x})\chi_1(\lambda,\boldsymbol{x}), \tag{42}$$

$$\partial_2\chi(\lambda,\mu,\boldsymbol{x})+(K_2(\lambda)-K_2(\mu))\chi(\lambda,\mu,\boldsymbol{x})=\tilde{\chi}_2(\mu,\boldsymbol{x})\chi_2(\lambda,\boldsymbol{x}), \tag{43}$$

where we have introduced shift operators and difference operators,

$$T_1\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})=\chi(\lambda,\mu,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}),$$

$$T_2\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})=\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}+[a_2]),$$

$$\Delta_1\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})=(T_1-1)\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}),$$

$$\Delta_2\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})=(T_2-1)\chi(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}).$$

These operators can also be represented as differential operators,

$$T_1=\exp\left(\sum_{n=1}^{\infty}\frac{a_1^n}{n}\partial_{(1)n}\right),$$

$$T_2=\exp\left(\sum_{n=1}^{\infty}\frac{a_2^n}{n}\partial_{(2)n}\right).$$

The coefficient of expansion of T_1 in a_1, T_2 in a_2 are given explicitly in terms of Schur's polynomials $p_N(\partial_{(1)1}/1,\ldots,\partial_{(1)N}/N)$, $p_N(\partial_{(2)1}/1,\ldots,\partial_{(2)N}/N)$.

The L operator for the DS hierarchy is obtained by taking (42) at $\lambda=0_2$, (43) at $\lambda=0_1$ [and then passing on to the wave functions (31)],

$$\begin{aligned}&\partial_1\Psi_2=u\Psi_1,\quad u=\chi_{21}(\boldsymbol{x}),\quad \partial_1=\frac{\partial}{\partial x_1},\quad x_1=x_{(1)1};\\ &\partial_2\Psi_1=v\Psi_2,\quad v=\chi_{12}(\boldsymbol{x}),\quad \partial_2=\frac{\partial}{\partial x_2},\quad x_2=x_{(2)1};\end{aligned} \tag{44}$$

while the hierarchy of A operators is obtained by taking, respectively, (40) at $\mu=0_2$, (41) at $\mu=0_1$,

$$\frac{\Delta_1}{a_1}\Psi_2 = uT_1\Psi_1,$$
$$\frac{\Delta_2}{a_2}\Psi_1 = vT_2\Psi_2. \tag{45}$$

One can also obtain dual L and A operators, taking (40), (42) at $\mu = 0_2$, (41), (43) at $\lambda = 0_1$.

It is possible to get the DS hierarchy through compatibility conditions, but we prefer to extract it directly from the basic formulas (40), (41), (42), (43).

Taking the difference of equations (40), (42) at $\lambda = 0_1$, $\mu = 0_2$, and also using similar combinations, one obtains the system of equations

$$\begin{gathered}
(1/a_1(T_1-1)-\partial_1)u = (T_1-1)\chi_{11}u,\\
(1/a_1(1-T_1^{-1})-\partial_1)v = (T_1^{-1}-1)\chi_{11}v,\\
(1/a_2(T_2-1)-\partial_2)v = (T_2-1)\chi_{22}v,\\
(1/a_2(1-T_2^{-1})-\partial_2)u = (T_2^{-1}-1)\chi_{22}u,\\
\partial_1\chi_{22} = uv,\\
\partial_2\chi_{11} = uv.
\end{gathered} \tag{46}$$

This system of equations represents a compact form of the DS hierarchy. The expansion of this system in a_1, a_2 generates the DS equation and higher equations of the hierarchy. It is useful to note that this expansion has the structure of recursion relations, expressing the derivatives in higher times through derivatives in lower times.

It is possible to obtain a DS hierarchy from the formula (34). One could start with the three shifts $T_1: \boldsymbol{x}_{(1)} \to \boldsymbol{x}_{(1)} + [a_1]$, $T_2: \boldsymbol{x}_{(2)} \to \boldsymbol{x}_{(2)} + [a_2]$, $T_3: \boldsymbol{x}_{(1)} \to \boldsymbol{x}_{(1)} + [a_3]$ (or $T_3: \boldsymbol{x}_{(2)} \to \boldsymbol{x}_{(2)} + [a_3]$ for the part of the hierarchy with higher times connected with the disc D_2). Then, taking the first order of expansion in a_1, a_2 and thus performing a transition from two difference operators to derivatives $\Delta_1 \to \partial/\partial x_1$, $T_1 \to 1$, $\Delta_2 \to \partial/\partial x_2$, $T_2 \to 1$, and finally writing down matrix components of equation (34), one obtains the DS hierarchy exactly in the form (46).

The zeroth order of expansion just gives trivial identities. The first order of expansion of the first pair of equations in a_1 and the second pair in a_2 gives the DS system

$$\begin{gathered}
\tfrac{1}{2}(\partial_{(1)2} + \partial_1^2)u = u\partial_1\chi_{11},\\
\tfrac{1}{2}(\partial_{(1)2} - \partial_1^2)v = -v\partial_1\chi_{11},\\
\tfrac{1}{2}(\partial_{(2)2} + \partial_2^2)v = v\partial_2\chi_{22},\\
\tfrac{1}{2}(\partial_{(2)2} - \partial_2^2)u = -u\partial_2\chi_{22},\\
\partial_1\chi_{22} = uv,\\
\partial_2\chi_{11} = uv,
\end{gathered} \tag{47}$$

where $\partial_{(1)2} = \partial/\partial x_{(1)2}$, $\partial_{(2)2} = \partial/\partial x_{(2)2}$. The usual form of the DS system (see, e.g., Ref. 24) is a closed subsystem of (47) written in terms of the time t defined as $\partial_t = \partial_{(1)2} + \partial_{(2)2}$.

The second order of expansion, with the use of the DS equations (47), gives

$$\begin{gathered}
\tfrac{1}{3}(\partial_{(1)3} - \partial_1^3)u = u\partial_2^{-1}\partial_1(u\partial_1 v) - \partial_1(u\partial_2^{-1}\partial_1 uv),\\
\tfrac{1}{3}(\partial_{(1)3} - \partial_1^3)v = v\partial_2^{-1}\partial_1(v\partial_1 u) - \partial_1(v\partial_2^{-1}\partial_1 uv),
\end{gathered} \tag{48}$$

$$\begin{aligned}
&\tfrac{1}{3}(\partial_{(2)3}-\partial_2^3)v = v\partial_1^{-1}\partial_2(v\partial_2 u)-\partial_2(v\partial_1^{-1}\partial_2 uv),\\
&\tfrac{1}{3}(\partial_{(2)3}-\partial_2^3)u = u\partial_1^{-1}\partial_2(u\partial_2 v)-\partial_2(u\partial_1^{-1}\partial_2 uv),
\end{aligned} \tag{49}$$

which is the higher DS system (Ref. 25). The reduction $u=1$ gives rise to the split Nizhnik–Veselov–Novikov equation for v,[26,27] while the reduction $u=v$ leads to the modified Nizhnik–Veselov–Novikov equation.[25]

To proceed to the modified DS hierarchy case, we integrate equations (40), (41), (42), (43) over λ,μ with some weight functions and obtain equations for wave functions,

$$\begin{aligned}
&\frac{\Delta_1}{a_1}\Phi = \tilde{\Psi}_1 T_1 \Psi_1,\\
&\frac{\Delta_2}{a_2}\Phi = \tilde{\Psi}_2 T_2 \Psi_2,\\
&\partial_1\Phi = \tilde{\Psi}_1\Psi_1,\\
&\partial_2\Phi = \tilde{\Psi}_2\Psi_2.
\end{aligned} \tag{50}$$

The L operator for a modified DS hierarchy can be obtained from the last pair of equations, using equations (44) and a dual DS L operator,

$$\partial_1\partial_2\Phi = (\partial_2 U)\partial_1\Phi + (\partial_1 V)\partial_2\Phi, \tag{51}$$

where $U=\ln\Psi_1$, $V=\ln\Psi_2$. The hierarchy of A operators arises if one uses the second pair of equations (50) to exclude $\tilde{\Psi}_1,\tilde{\Psi}_2$ from the first pair,

$$\begin{aligned}
&\frac{\Delta_1}{a_1}\Phi = W_1\partial_1\Phi, \quad W_1 = \frac{T_1\Psi_1}{\Psi_1};\\
&\frac{\Delta_2}{a_2}\Phi = W_2\partial_2\Phi, \quad W_2 = \frac{T_2\Psi_2}{\Psi_2}.
\end{aligned} \tag{52}$$

The compatibility condition for (51), (52) gives the modified DS hierarchy,

$$\frac{1}{a_1}\Delta_1\partial_2 U + (T_1\partial_2 U)(\partial_1 W_1) + (T_1\partial_1 V)(\partial_2 W_1) + W_1(\partial_2 U)\Delta_1\partial_1 V = \partial_1\partial_2 W_1 + \partial_1(W_1\partial_2 U), \tag{53}$$

$$\frac{1}{a_1}\Delta_1\partial_1 V + W_1(\Delta_1\partial_1 V)(\partial_1 V) - \partial_1(W_1\partial_1 V) = 0,$$

$$W_1 = \exp(\Delta_1 U),$$

and for times with the subscript (2),

$$\frac{1}{a_2}\Delta_2\partial_1 V + (T_2\partial_1 V)(\partial_2 W_2) + (T_2\partial_2 U)(\partial_1 W_2) + W_2(\partial_1 V)\Delta_2\partial_2 U = \partial_2\partial_1 W_2 + \partial_2(W_2\partial_1 V), \tag{54}$$

$$\frac{1}{a_2}\Delta_2\partial_2 U + W_2(\Delta_2\partial_2 U)(\partial_2 U) - \partial_2(W_2\partial_2 U) = 0,$$

$$W_2 = \exp(\Delta_2 V).$$

The zeroth order of expansion of these equations in a_1,a_2 gives a modified DS equation,[23]

$$\tfrac{1}{2}\partial_2(\partial_{(1)2}U-\partial_1^2U-(\partial_1U)^2)+\partial_1(\partial_2U\cdot\partial_1V)=0,$$
$$\partial_{(1)2}V+\partial_1^2V+(\partial_1V)=\partial_1V\cdot\partial_1U, \tag{55}$$

$$\tfrac{1}{2}\partial_1(\partial_{(2)2}V-\partial_2^2V-(\partial_2V)^2)+\partial_2(\partial_2U\cdot\partial_1V)=0,$$
$$\partial_{(2)2}U+\partial_2^2U+(\partial_2U)^2=\partial_2V\cdot\partial_2U. \tag{56}$$

The hierarchy of Combescure invariants (Ref. 11) for the modified DS hierarchy is obtained by a representation of the DS solutions u,v through the mDS solutions U,V,

$$u=\exp(-T_1U)\frac{\Delta_1}{a_1}\exp(V),$$
$$v=\exp(-T_2V)\frac{\Delta_2}{a_2}\exp(U). \tag{57}$$

VI. τ FUNCTION FOR THE GENERALIZED DS HIERARCHY

Now we will analyze the functional equations (16) for the case $N=2$. For convenience we present them here explicitly:

$$\left(\frac{\mu-a_1}{\mu-b_1}\right)\chi_{11}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(\frac{\lambda-a_1}{\lambda-b_1}\right)\chi_{11}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_1-a_1)\chi_{11}(\lambda,b_1,\boldsymbol{x}+[\boldsymbol{a}])$$
$$\times\chi_{11}(b_1,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_2-a_2)\chi_{21}(\lambda,b_2,\boldsymbol{x}+[\boldsymbol{a}])\chi_{12}(b_2,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0, \tag{58}$$

$$\left(\frac{\mu-a_2}{\mu-b_2}\right)\chi_{22}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(\frac{\lambda-a_2}{\lambda-b_2}\right)\chi_{22}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_1-a_1)\chi_{12}(\lambda,b_1,\boldsymbol{x}+[\boldsymbol{a}])$$
$$\times\chi_{21}(b_1,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_2-a_2)\chi_{22}(\lambda,b_2,\boldsymbol{x}+[\boldsymbol{a}])\chi_{22}(b_2,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0, \tag{59}$$

$$\left(\frac{\mu-a_1}{\mu-b_1}\right)\chi_{12}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(\frac{\lambda-a_2}{\lambda-b_2}\right)\chi_{12}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_1-a_1)\chi_{12}(\lambda,b_1,\boldsymbol{x}+[\boldsymbol{a}])$$
$$\times\chi_{11}(b_1,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_2-a_2)\chi_{22}(\lambda,b_2,\boldsymbol{x}+[\boldsymbol{a}])\chi_{12}(b_2,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0, \tag{60}$$

$$\left(\frac{\mu-a_2}{\mu-b_2}\right)\chi_{21}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{a}])-\left(\frac{\lambda-a_1}{\lambda-b_1}\right)\chi_{21}(\lambda,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_1-a_1)\chi_{11}(\lambda,b_1,\boldsymbol{x}+[\boldsymbol{a}])$$
$$\times\chi_{21}(b_1,\mu,\boldsymbol{x}+[\boldsymbol{b}])+(b_2-a_2)\chi_{21}(\lambda,b_2,\boldsymbol{x}+[\boldsymbol{a}])\chi_{22}(b_2,\mu,\boldsymbol{x}+[\boldsymbol{b}])=0. \tag{61}$$

First it is clear that equations (58) with $a_2=b_2$ and (59) with $a_1=b_1$ which coincide with those for scalar KP hierarchy give rise to the following expressions for χ_{11} and χ_{22} (see Ref. 12),

$$\chi_{11}(\lambda,\mu,\boldsymbol{x})=\frac{1}{(\lambda-\mu)}\frac{\tau_1(\boldsymbol{x}_{(1)}-[\lambda]+[\mu],\boldsymbol{x}_{(2)})}{\tau_1(\boldsymbol{x})}, \tag{62}$$

$$\chi_{22}(\lambda,\mu,\boldsymbol{x})=\frac{1}{(\lambda-\mu)}\frac{\tau_2(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}-[\lambda]+[\mu])}{\tau_2(\boldsymbol{x})}, \tag{63}$$

where τ_1 and τ_2 are the components of the τ function which depend on the variables $\boldsymbol{x}=(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})$.

Let us consider now equation (59) at $\mu=a_1$ and $a_2=b_2$. Substituting the expression (62) into this equation, we obtain

$$\tau_1(\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)}+[a_2])\chi_{12}(\lambda,a_1,\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)}+[a_2])$$

$$= \tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2])\chi_{12}(\lambda,a_1,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2]). \tag{64}$$

For $b_1=0$ after the change $\boldsymbol{x}_{(2)}+[a_2]\to\boldsymbol{x}_{(2)}$, one gets

$$\tau_1(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})\chi_{12}(\lambda,a_1,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})\chi_{12}(\lambda,0,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}). \tag{65}$$

Denoting the RHS of this equation as $\tilde{\tau}_{12}(\lambda,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})$, we rewrite it as

$$\chi_{12}(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \frac{\tilde{\tau}_{12}(\lambda,\boldsymbol{x}_{(1)}+[\mu],\boldsymbol{x}_{(2)})}{\tau_1(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})}. \tag{66}$$

Then the substitution of (66) into equation (64) with $\mu=a_1$ and $\lambda=a_2$ gives rise to

$$\tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[b_2])\ \frac{\tilde{\tau}_{12}(a_2,\boldsymbol{x}_{(1)}+[a_1]+[b_1],\boldsymbol{x}_{(2)}+[a_2])}{\tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2])}, \tag{67}$$

$$= \tau_2(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[b_2])\ \frac{\tilde{\tau}_{12}(b_2,\boldsymbol{x}_{(1)}+[a_1]+[b_1],\boldsymbol{x}_{(2)}+[b_2])}{\tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2])}. \tag{68}$$

This equation implies that

$$\tilde{\tau}_{12}(a,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \frac{\tau_{12}(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}-[a])}{\rho(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})} \tag{69}$$

and

$$\rho(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \frac{\tau_2(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})}{\tau_1(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})}, \tag{70}$$

where τ_{12} is an arbitrary function. Substitution of (69) into (66) gives

$$\chi_{12}(\lambda,\mu,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \frac{\tau_{12}(\lambda,\boldsymbol{x}_{(1)}+[\mu],\boldsymbol{x}_{(2)}-[\lambda])}{\tau_2(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})}. \tag{71}$$

Now we proceed with equation (61). Substitution of the expression (63) into this equation with $\mu=a_2$, $a_1=b_1$ gives

$$\chi_{21}(\lambda,a_2,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[b_2]) = \frac{\tau_2(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2])}{\tau_2(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[b_2])}\chi_{21}(\lambda,b_2,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}+[a_2]). \tag{72}$$

After the change $\boldsymbol{x}_{(1)}+[a_1]\to\boldsymbol{x}_{(1)}$ and the choice $b_2=0$ one gets

$$\tau_2(\boldsymbol{x})\chi_{21}(\lambda,a_2,\boldsymbol{x}) = \tau_2(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}+[a_2])\chi_{21}(\lambda,0,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}+[a_2]). \tag{73}$$

From this equation one concludes that

$$\chi_{21}(\lambda,\mu,\boldsymbol{x}) = \frac{\tilde{\tau}_{21}(\lambda,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}+[\mu])}{\tau_2(\boldsymbol{x})}, \tag{74}$$

where $\tilde{\tau}_{21}(\lambda,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) = \tau_2(\boldsymbol{x})\chi_{21}(\lambda,0,\boldsymbol{x})$.

Substituting now with $\mu=a_2$, $\lambda=a_1$, $b_2=0$ we obtain

$$\frac{\tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})}{\tau_2(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})}\ \tilde{\tau}_{21}(a_1,\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)}) = \frac{\tau_1(\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)})}{\tau_2(\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)})}\ \tilde{\tau}_{21}(b_1,\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)}). \tag{75}$$

Hence

$$\tilde{\tau}_{21}(a_1,\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)})=\frac{\tau_{21}(\boldsymbol{x}_{(1)}-[a_1],\boldsymbol{x}_{(2)})}{\tilde{\rho}(\boldsymbol{x})} \tag{76}$$

and

$$\tilde{\rho}(\boldsymbol{x})=\frac{\tau_1(\boldsymbol{x})}{\tau_2(\boldsymbol{x})}. \tag{77}$$

Substitution of (76) and (77) into (74) gives

$$\chi_{21}(\lambda,\mu,\boldsymbol{x})=\frac{\tau_{21}(\boldsymbol{x}_{(1)}-[\lambda],\boldsymbol{x}_{(2)}+[\mu])}{\tau_1(\boldsymbol{x})}. \tag{78}$$

To establish the relation between τ_1 and τ_2 we consider again equation (75). Due to (76) one gets

$$\frac{\tau_1(\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)})}{\tau_2(\boldsymbol{x}_{(1)}+[b_1],\boldsymbol{x}_{(2)})}=\frac{\tau_1(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})}{\tau_2(\boldsymbol{x}_{(1)}+[a_1],\boldsymbol{x}_{(2)})}. \tag{79}$$

Since a_1 and b_1 are arbitrary parameters, one concludes that

$$\frac{\tau_1(\boldsymbol{x})}{\tau_2(\boldsymbol{x})}=f(\boldsymbol{x}_{(2)}). \tag{80}$$

In the same way, it follows from (60) that

$$\frac{\tau_1(\boldsymbol{x})}{\tau_2(\boldsymbol{x})}=f(\boldsymbol{x}_{(1)}). \tag{81}$$

Thus

$$\frac{\tau_1(\boldsymbol{x})}{\tau_2(\boldsymbol{x})}=\text{const}. \tag{82}$$

Having in mind the formulas (62) and (63), one can put $\tau_1=\tau_2$ without the loss of generality.

So for the 2-component KP hierarchy one has

$$\chi_{11}(\lambda,\mu,\boldsymbol{x})=\frac{1}{(\lambda-\mu)}\,\frac{\tau(\boldsymbol{x}_{(1)}-[\lambda]+[\mu],\boldsymbol{x}_{(2)})}{\tau(\boldsymbol{x})}, \tag{83}$$

$$\chi_{22}(\lambda,\mu,\boldsymbol{x})=\frac{1}{(\lambda-\mu)}\,\frac{\tau(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}-[\lambda]+[\mu])}{\tau(\boldsymbol{x})}, \tag{84}$$

$$\chi_{12}(\lambda,\mu,\boldsymbol{x})=\frac{\tau_{12}(\boldsymbol{x}_{(1)}+[\mu],\boldsymbol{x}_{(2)}-[\lambda])}{\tau(\boldsymbol{x})}, \tag{85}$$

$$\chi_{21}(\lambda,\mu,\boldsymbol{x})=\frac{\tau_{21}(\boldsymbol{x}_{(1)}-[\lambda],\boldsymbol{x}_{(2)}+[\mu])}{\tau(\boldsymbol{x})}. \tag{86}$$

At this step τ, τ_{12}, and τ_{21} are arbitrary functions. Substitution of the expressions (83)–(86) into the general functional equations (58)–(61) gives rise to the four bilinear functional equations for these functions which represent themselves with the addition formulas. We will write them down explicitly for the multicomponent case later.

VII. τ FUNCTION AND ADDITION FORMULAS FOR THE *N*-COMPONENT KP HIERARCHY

It will be shown in the next section that the objects of multicomponent KP hierarchy depend not only on continuous times, but also on discrete times (indices). The addition formulas arising from the discrete dynamics will be considered in the context of the Toda lattice hierarchy. Here we derive only the addition formulas following from the basic formula (16) defining the dynamics for continuous variables.

For the general multicomponent KP hierarchy the system (16) looks rather complicated. But its analysis is simplified considerably due to the special structure of equations (16). It is not difficult to see that for two fixed values of α and β, taking $a_\gamma = b_\gamma$, $\gamma \neq \alpha$, $\gamma \neq \beta$, the system reduces to four equations of the DS hierarchy (58)–(61) for the variables $\boldsymbol{x}_{(\alpha)}$ and $\boldsymbol{x}_{(\beta)}$. Hence, in these variables we have the formulas

$$\chi_{\alpha\alpha}(\lambda,\mu,\boldsymbol{x}) = \frac{1}{(\lambda-\mu)} \frac{\tau(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(\alpha)}-[\lambda]+[\mu],\ldots,\boldsymbol{x}_{(N)})}{\tau(\boldsymbol{x})},$$
$$\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}) = \frac{\tau_{\alpha\beta}(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(\alpha)}+[\mu],\ldots,\boldsymbol{x}_{(\beta)}-[\lambda],\ldots,\boldsymbol{x}_{(N)})}{\tau(\boldsymbol{x})} \quad (\alpha \neq \beta). \tag{87}$$

Here indices α and β are arbitrary. Therefore the formulas (87) take place for all α and β.

To compactify these formulas we will introduce the shift operator $T_\alpha(\lambda)$ $(\alpha = 1,\ldots,N)$,

$$T_\alpha(\lambda) f(\boldsymbol{x}) = f(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(\alpha)}+[\lambda],\ldots,\boldsymbol{x}_{(N)}). \tag{88}$$

Using such operators, one can present the formulas (87) in the form

$$\chi_{\alpha\alpha}(\lambda,\mu,\boldsymbol{x}) = \frac{1}{\lambda-\mu} \frac{T_\alpha(\mu)T_\alpha^{-1}(\lambda)\tau(\boldsymbol{x})}{\tau(\boldsymbol{x})}, \quad \alpha = 1,\ldots,N,$$
$$\chi_{\alpha\beta}(\lambda,\mu,\boldsymbol{x}) = \frac{T_\alpha(\mu)T_\beta^{-1}(\lambda)\tau_{\alpha\beta}(\boldsymbol{x})}{\tau(\boldsymbol{x})}, \quad \alpha \neq \beta. \tag{89}$$

In the special cases $\lambda = 0$ or $\mu = 0$ these formulas reduce to the standard formulas for the Baker–Akhiezer function.[3,14,16] Separating the diagonal case $\alpha = \beta$ and the off diagonal one, one gets

$$\begin{aligned}
&(a_\alpha-\mu)(\lambda-b_\alpha)\tau(\boldsymbol{x}+[\boldsymbol{b}])T_\alpha(\mu)T_\alpha^{-1}(\lambda)\tau(\boldsymbol{x}+[\boldsymbol{a}]) + (b_\alpha-\mu)(\lambda-a_\alpha)\tau(\boldsymbol{x}+[\boldsymbol{a}])\\
&\quad\times T_\alpha(\mu)T_\alpha^{-1}(\lambda)\tau(\boldsymbol{x}+[\boldsymbol{b}]) + (b_\alpha-a_\alpha)(\lambda-\mu)T_\alpha(\mu)T_\alpha^{-1}(b_\alpha)\tau(\boldsymbol{x}+[\boldsymbol{b}])T_\alpha(b_\alpha)\\
&\quad\times T_\alpha^{-1}(\lambda)\tau(\boldsymbol{x}+[\boldsymbol{a}]) + (\mu-\lambda)(\mu-b_\alpha)(\lambda-b_\alpha)\sum_{\gamma\neq\alpha}(b_\gamma-a_\gamma)T_\alpha(\mu)\\
&\quad\times T_\gamma^{-1}(b_\gamma)\tau_{\alpha\gamma}(\boldsymbol{x}+[\boldsymbol{b}])T_\gamma(b_\gamma)T_\beta^{-1}(\lambda)\tau_{\gamma\beta}(\boldsymbol{x}+[\boldsymbol{a}]) = 0
\end{aligned} \tag{90}$$

and

$$\begin{aligned}
&(\mu-a_\alpha)(\lambda-b_\beta)\tau(\boldsymbol{x}+[\boldsymbol{b}])T_\alpha(\mu)T_\beta^{-1}(\lambda)\tau_{\alpha\beta}(\boldsymbol{x}+[\boldsymbol{a}]) + (\mu-b_\alpha)(\lambda-a_\beta)\tau(\boldsymbol{x}+[\boldsymbol{a}])\\
&\quad\times T_\alpha(\mu)T_\beta^{-1}(\lambda)\tau_{\alpha\beta}(\boldsymbol{x}+[\boldsymbol{b}]) + (b_\alpha-a_\alpha)(\lambda-b_\beta)T_\alpha(\mu)T_\alpha^{-1}(b_\alpha)\tau(\boldsymbol{x}+[\boldsymbol{b}])T_\alpha(b_\alpha)\\
&\quad\times T_\beta^{-1}(\lambda)\tau_{\alpha\beta}(\boldsymbol{x}+[\boldsymbol{a}]) + (\mu-b_\alpha)(\lambda-b_\beta)\sum_{\gamma\neq\alpha,\beta}(b_\gamma-a_\gamma)T_\alpha(\mu)T_\gamma^{-1}(b_\gamma)\tau_{\alpha\gamma}(\boldsymbol{x}\\
&\quad+[\boldsymbol{b}])T_\gamma(b_\gamma)T_\beta^{-1}(\lambda)\tau_{\gamma\beta}(\boldsymbol{x}+[\boldsymbol{a}]) = 0.
\end{aligned} \tag{91}$$

Equations (90) and (91) represent the extension of the simplest addition formula for the τ function in the scalar case (see Ref. 12) to the multicomponent case. They are related to the generalization

of the bilinear relations for the Plücker coordinates to the multicomponent case. On the other hand they could be connected with the possible multidimensional extensions of the Fay's trisecant formula.[28]

More general addition formulas for the τ function follow from the general functional equation for the function $\chi_{\alpha\beta}$ which corresponds to the choice

$$(g_1(\boldsymbol{x})g_2^{-1}(\boldsymbol{x}))_{\alpha\beta}=\delta_{\alpha\beta}\prod_{k=1}^{N}\frac{\nu-a_{\alpha k}}{\nu-b_{\alpha k}}. \tag{92}$$

VIII. τ-FUNCTIONAL SUBSTITUTION AND CLOSED 1-FORM

The formulas (83), (84), (85), (86) indicate that it is possible to preserve in the multicomponent case the general formula connecting the τ function and the CBA function in the scalar case. It would be more consistent first to define the τ function through the CBA function in terms of the closed one-form, and then to prove that this definition implies the general formula. On the other hand, one may treat the general formula as the definition and then prove the correctness of this definition reducing it to closed 1-form and showing the equivalence of ''local'' and ''global'' formulas (that implies the unique existence of the τ function, at least in some neighborhood of $g=1$). In this paper we would rather follow the second way.

So we define the τ function by the relation

$$\chi(\lambda,\mu,g)=\chi_0(\lambda,\mu)\frac{\tau(g_{(\lambda,\mu)}g)}{\tau(g)}. \tag{93}$$

Here $\chi_0(\lambda,\mu)$ is just the function $(\lambda-\mu)^{-1}$ extended as 1 for λ, μ in different D_i, and the function $g_{(\lambda,\mu)}$ is an extension of the function $(\nu-\lambda)/(\nu-\mu)$ to the case of disconnected G,

$$\chi_0(\lambda,\mu)=\begin{cases}\dfrac{1}{\lambda-\mu}, & \text{for } \lambda\in D_i, \quad \mu\in D_i,\\ 1, & \text{for } \lambda\in D_i, \quad \mu\in D_j, \quad i\neq j.\end{cases} \tag{94}$$

The function $g_{(\lambda,\mu)}$ [in fact, it is a set of N loop group elements $g_i(\nu)$ depending on the parameters λ,μ] is the ratio of two terms,

$$\begin{gathered}g_{(\lambda,\mu)}=\frac{g_{(\lambda)}}{g_{(\mu)}},\\ g_{(\lambda)i}(\nu)=\begin{cases}\nu-\lambda, & \text{for } \lambda\in D_i,\\ 1, & \text{for } \lambda\notin D_i.\end{cases}\end{gathered} \tag{95}$$

To extract the formulas (83), (84), (85), (86) from the formula (93), one should recall that the loop group defining the multicomponent KP hierarchy can be decomposed in the following way:

$$g_i(\nu)=g_{i\mathrm{out}}(\nu)\times\nu^{n_i}, \quad \sum_{i=1}^{N}n_i=0, \tag{96}$$

where the first term defines the dependence on continuous times and the second—on the discrete times (indices). If λ,μ belong to the same D_i, using explicit parametrization of the loop group elements, we obtain exactly the formulas (83), (84). For $\lambda\in D_k$, $\mu\in D_p$, $k\neq p$, $g_{(\lambda,\mu)}$ has a factorization,

$$\begin{gathered}g_{(\lambda,\mu)k}(\nu)=\frac{\nu-\lambda}{\nu}\times\nu,\\ g_{(\lambda,\mu)p}(\nu)=\frac{\nu}{\nu-\mu}\times\nu^{-1},\end{gathered} \tag{97}$$

$$g_{(\lambda,\mu)i}(\nu)=1, \quad i\neq k, \quad i\neq p.$$

Using these formulas, we obtain the formulas (85), (86), with the components of the τ function interpreted as a τ function with shifted index variables (the general index zero is preserved). For example, in the two-component case the function τ_{12} corresponds to $\tau(n_1-1,n_2+1)$. To express the CBA function with zero indices, we need to know the set of functions τ, τ_{ij}, $i\neq j$. Then for other admissible index vectors the CBA function and the τ function can be found explicitly through the determinant formulas.[13] Thus, the discrete dynamics is exactly solvable in terms of initial data (CBA function with zero indices or the set of τ functions τ, τ_{ij}, $i\neq j$).

Returning to our original purpose, we will extract the definition of the τ function in terms of the closed 1-form from (2) and then prove that both definitions are equivalent. Our derivation will mostly repeat the derivation in the scalar case given in Ref. 12, but we would like to mention it due to the importance for the following consideration.

Differentiating (93) with respect to λ, one obtains

$$-\frac{1}{\tau(g)}\oint \frac{\delta\tau(g\times g_{(\lambda,\mu)})}{\delta \ln g_\alpha(\nu)}\frac{1}{\nu-\lambda}d\nu=\frac{\partial}{\partial\lambda}(\lambda-\mu)\chi_{\alpha\alpha}(\lambda,\mu;g), \tag{98}$$

or for $\lambda=\mu$,

$$-\frac{1}{\tau(g)}\oint \frac{\delta\tau(g)}{\delta \ln g_\alpha(\nu)}\frac{1}{\nu-\lambda}d\nu=\chi_{r\alpha\alpha}(\lambda,\lambda;g), \tag{99}$$

where $\chi_{r\alpha\beta}(\lambda,\mu;g)=\chi_{\alpha\beta}(\lambda,\mu;g)-\delta_{\alpha\beta}(\lambda-\mu)^{-1}$, ''$r$'' is for ''regularized.'' This equation can be rewritten in the form

$$-\frac{1}{\tau(g)}\oint \frac{\delta\tau(g)}{\delta \ln g_\alpha(\nu)}\frac{1}{\nu-\lambda}d\nu=\oint \chi_{r\alpha\alpha}(\nu,\nu;g)\frac{1}{\nu-\lambda}d\nu. \tag{100}$$

Relation (100) implies that the linear functionals $-[1/\tau(g)][\delta\tau(g(\nu))/\delta \ln g]$ and $\chi_r(\nu,\nu;g)$ are identical for arbitrary variations $\delta g/g$ analytic outside each D_i and decreasing at infinity. Thus for $\delta g/g$ belonging to this class of functions the variation of the τ function is given by

$$-\delta \ln \tau(g)=\sum_{\alpha=1}^{N}\oint \chi_{r\alpha\alpha}(\nu,\nu;g)\frac{\delta g_\alpha(\nu)}{g_\alpha(\nu)}d\nu. \tag{101}$$

This expression defines a variational 1-form defining the τ function. It is easy to prove using the identity (1) that this form is *closed*. Indeed, according to (1) [here we are using scalar notations, it is not difficult to extract equations in terms of components $\chi_{\alpha\beta}(\lambda,\mu;g)$],

$$\begin{aligned}\delta\chi(\lambda,\mu;g)&=\int_{\partial G}\chi(\nu,\mu;g)\frac{\delta g(\nu)}{g(\nu)}\chi(\lambda,\nu;g)d\nu,\\ \delta\chi_r(\lambda,\lambda;g)&=\int_{\partial G}\chi_r(\nu,\lambda;g)\frac{\delta g(\nu)}{g(\nu)}\chi_r(\lambda,\nu;g)d\nu.\end{aligned} \tag{102}$$

So the variation of (101) gives

$$-\delta^2 \ln \tau(g)=\int_{\partial G}\int_{\partial G}\chi_r(\nu,\lambda;g)\chi_r(\lambda,\nu;g)\frac{\delta' g(\nu)}{g(\nu)}\frac{\delta g(\lambda)}{g(\lambda)}d\nu d\lambda. \tag{103}$$

The symmetry of the kernel of second variation with respect to λ,ν implies that the form (101) is closed.

So the formula (101) gives the definition of the τ function in terms of the closed 1-form. For the standard KP coordinates,

$$\frac{\delta g_\alpha(\lambda)}{g_\alpha(\lambda)} = \sum_{n=1}^{\infty} \frac{dx_{(\alpha)n}}{\lambda^n}.$$

This formula allows us to obtain a closed 1-form in terms of $dx_{\alpha n}$,

$$-d\ \ln\ \tau(\boldsymbol{k},\boldsymbol{x}) = \sum_{\alpha=1}^{N}\sum_{n=0}^{\infty} \frac{\partial^n}{\partial \nu^n}\,\chi_{r\alpha\alpha}(\nu,\nu,\boldsymbol{k},\boldsymbol{x})\Big|_{\nu=0} dx_{(\alpha)n}. \tag{104}$$

To prove that the ''local'' definition in terms of the closed 1-form implies the global definition (93), it is sufficient to show that logarithmic derivatives of the LHS and the RHS of (93) with respect to λ,μ and variational logarithmic derivatives with respect to $\delta g_\alpha / g_\alpha$ are equal in virtue to (101) and generalized Hirota bilinear identity (1). Then the lhs and the rhs of (93) could differ only by the factor, and the normalization of the function χ implies that this factor is equal to 1.

Here we will prove only the equality of logarithmic variations; the proof for derivatives is analogous (see Ref. 12).

The logarithmic variation of (93) gives

$$\begin{aligned}
&\delta(\ln\ \chi(\lambda,\mu;g)(\lambda-\mu) - \ln\ \tau(g\times g_{(\lambda,\mu)}) + \ln\ \tau(g)) \\
&\quad = \frac{1}{\chi(\lambda,\mu)} \int_{\partial G} \chi(\nu,\mu;g)\,\frac{\delta g(\nu)}{g(\nu)}\,\chi(\lambda,\nu;g)\,d\nu \\
&\qquad - \int_{\partial G} (\chi_r(\nu,\nu;g\times g_{(\lambda,\mu)}) - \chi_r(\nu,\nu;g))\,\frac{\delta g(\nu)}{g(\nu)}\,d\nu
\end{aligned} \tag{105}$$

[we have used (102) and (101)]. After the use of the determinant formula,[13]

$$\chi(\lambda,\mu;g\times g_{(\lambda',\mu')}) = g^{-1}_{(\lambda',\mu')}(\lambda)\,g_{(\lambda',\mu')}(\mu)\,\frac{\begin{vmatrix}\chi(\lambda',\mu') & \chi(\lambda',\mu)\\ \chi(\lambda,\mu') & \chi(\lambda,\mu)\end{vmatrix}}{\chi(\lambda',\mu')}, \tag{106}$$

to transform the second term, one concludes that the variation is equal to zero.

IX. GENERALIZED 2D TODA LATTICE HIERARCHY. BASIC EQUATIONS

We will consider here for simplicity only scalar 2D Toda lattice hierarchy. It is generated by the generalized Hirota identity with the domain G which consists of two disconnected regions, i.e., in some sense it is a two-point KP hierarchy. But it also possess some fundamental special features and can be treated as a primary object with respect to the scalar KP hierarchy.

Without loss of generality one can choose G as

$$G = D_0 \cup D_\infty,$$

where D_0 and D_∞ are unit discs around the origin and infinity, respectively. The two-component KP hierarchy is generated by the pair of functions $g_{\rm in}$, $g_{\rm out}$ defined on the boundary of D_0, D_∞, respectively, having a zero general index (let their individual indices be $n,-n$), $g_{\rm out}$ is analytic and has no zeros outside the unit circle, and $g_{\rm in}$ inside the unit circle. It is easy to see, that dynamics defined by the pair of these functions coincides with the dynamics defined by one function for both boundaries,

$$g = \lambda^n g_{\rm in} g_{\rm out}, \tag{107}$$

up to some trivial dynamics, defined by a function analytic in $G = D_0 \cup D_\infty$ ($g = g_{\rm in}$, $\lambda \in D_0$; $g = g_{\rm out}$, $\lambda \in D_\infty$).

As we know, an *arbitrary* function on the circle having no zeros (arbitrary loop) can be represented in the form (107). So in the case of the Toda lattice hierarchy the objects of the theory nontrivially depend on the whole loop group (i.e., on discrete time, positive, and negative con-

tinuous times); the τ function in this case is a functional of arbitrary g on the circle. Discrete dynamics is correctly defined, because in this case λ^n is a function having in G an equal number of zeros and poles.

The explicit parametrization of the loop group element g in terms of times is given by

$$g(n,\boldsymbol{x},\nu)=g(n,\boldsymbol{x}_+,\boldsymbol{x}_-,\nu)=\nu^n\exp\left(\sum_{k=1}^{\infty}x_{(+)k}\nu^{-k}+x_{(-)k}\nu^k\right). \tag{108}$$

The compact form of the 2D Toda lattice hierarchy in terms of the function $\chi(\lambda,\mu,k,\boldsymbol{x})$ can be obtained from (1) using special combinations of group elements $g_1g_2^{-1}$ with simple analytical properties. To define dynamics with respect to continuous times, let us take

$$g_0(\nu):=g_1(\nu)g_2^{-1}(\nu)=\frac{\nu-a_+}{\nu-b_+}\frac{\nu-a_-}{\nu-b_-},$$

$$a_+,b_+\in D_0,\quad a_-,b_-\in D_\infty, \tag{109}$$

$$g_0(\nu)=g_+\times(g_-)^{-1},\quad g_+=\frac{\nu-a_+}{\nu-b_+},\quad g_-=\frac{\nu-b_-}{\nu-a_-}.$$

Substituting the expression for $g_1g_2^{-1}$ into the generalized Hirota bilinear identity (1), one obtains

$$\begin{aligned}&g_0(\mu)\chi(\lambda,\mu,g\times g_+)-g_0(\lambda)\chi(\lambda,\mu,g\times g_-)+\mathrm{Res}(g_0)|_{b_+}\chi(\lambda,b_+,g\times g_+)\chi(b_+,\mu,g\times g_-)\\&\quad+\mathrm{Res}(g_0)|_{b_-}\chi(\lambda,b_-,g\times g_+)\chi(b_-,\mu,g\times g_-)=0.\end{aligned} \tag{110}$$

The equations of the Toda lattice hierarchy are generated by equation (110) with $b_+=0$, $b_-=\infty$, or, in terms of g_0,

$$\begin{aligned}g_0(\nu)=g_1(\nu)g_2^{-1}(\nu)=\left(1-\frac{\nu}{a_-}\right)\left(1-\frac{a_+}{\nu}\right)=\exp\left(\sum_{i=1}^{\infty}\frac{1}{i}\left(\left(\frac{\nu}{a_-}\right)^i+\left(\frac{a_+}{\nu}\right)^i\right)\right),\\|a_+|<1,\quad |a_-|>1.\end{aligned} \tag{111}$$

Substituting into equation (110) parametrization of loop group elements in terms of times, we obtain

$$\begin{aligned}&\left(1-\frac{\mu}{a_-}\right)\left(1-\frac{a_+}{\mu}\right)\chi(\lambda,\mu,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)-\left(1-\frac{\lambda}{a_-}\right)\left(1-\frac{a_+}{\lambda}\right)\\&\times\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-])+-a_+\chi(\lambda,0,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)\chi(0,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-])\\&-\frac{1}{a_-}\chi(\lambda,\infty,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)\chi(\infty,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-])=0,\end{aligned} \tag{112}$$

where $\chi(\lambda,\infty)=\chi(\lambda,\mu)\times\mu|_{\mu=\infty}$. In terms of CBA and BA functions $\psi(\lambda,\mu)$ and $\psi_0(\lambda)$, $\tilde{\psi}_0(\mu)$ $\psi_\infty(\lambda)$, $\tilde{\psi}_\infty(\mu)$ we have

$$\begin{aligned}&\psi(\lambda,\mu,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)-\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-])\\&\quad=a_+\psi_0(\lambda,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)\tilde{\psi}_0(\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-])\\&\qquad+\frac{1}{a_-}\psi_\infty(\lambda,n,\boldsymbol{x}_++[\boldsymbol{a}_+],\boldsymbol{x}_-)\tilde{\psi}_\infty(\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_--[\boldsymbol{a}_-]).\end{aligned} \tag{113}$$

To get equations defining the discrete dynamics, we take

$$g_0(\nu)=g_1(\nu)g_2^{-1}(\nu)=\frac{1}{\lambda} \tag{114}$$

or

$$g_0(\nu)=g_1(\nu)g_2^{-1}(\nu)=\lambda. \tag{115}$$

In the first case the equation for the function χ reads as

$$\frac{1}{\mu}\chi(\lambda,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)-\frac{1}{\lambda}\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,0,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)\chi(0,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-), \tag{116}$$

and in the second case,

$$\mu\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)-\lambda\chi(\lambda,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,\infty,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\chi(\infty,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-). \tag{117}$$

In terms of CBA and BA functions one gets

$$\psi(\lambda,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)-\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\psi_0(\lambda,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)\tilde{\psi}_0(\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$

$$\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)-\psi(\lambda,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)=\psi_\infty(\lambda,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\tilde{\psi}(\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-). \tag{118}$$

Equations (113), (118) define the generalized 2D Toda lattice hierarchy. The discrete dynamics can be resolved in both directions explicitly [i.e., $\chi(n+1)$, $\chi(n-1)$ can be expressed through $\chi(n)$] using the determinant formula.[13]

X. TODA LATTICE HIERARCHY: 2DTL, 2D VOLTERRA CHAIN, 2DTL SINGULARITY MANIFOLD EQUATION

At this point we would like to derive the simplest equations of the generalized 2D Toda lattice hierarchy. They contain three variables: one discrete variable n and two continuous variables $x_{(+)1}$, $x_{(-)1}$ (we will denote them x_+ and x_-). We will derive three different equations corresponding to different levels of the tower of equations of the generalized hierarchy: the main equation (the standard two-dimensional Toda lattice equation), the modified equation (i.e., the equation for the Baker–Akhiezer function), and the singularity manifold-type equation (arising here as an equation for the CBA function). It is interesting to note that the second and the third equations (not to mention the first) are already known in literature in different contexts (see Refs. 18–21).

The basic equations for our derivation are equations (116), (117) and the zeroth order terms of expansion of (112) in a_+, $1/a_-$, i.e.,

$$\left(\frac{\partial}{\partial x_+}+\frac{1}{\lambda}-\frac{1}{\mu}\right)\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,0,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\chi(0,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$

$$\left(\frac{\partial}{\partial x_-}+\lambda-\mu\right)\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,\infty,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\chi(\infty,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-). \tag{119}$$

In terms of CBA and BA functions instead of (119) we have

$$\frac{\partial}{\partial x_+}\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\psi_0(\lambda,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\tilde{\psi}_0(\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$

$$\frac{\partial}{\partial x_-}\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)=\psi_\infty(\lambda,n,\boldsymbol{x}_+,\boldsymbol{x}_-)\tilde{\psi}_\infty(\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-). \tag{120}$$

Combining (116), (117), and (119), one obtains

$$\left(\frac{\partial}{\partial x_+}+\frac{1}{\lambda}\right)\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)-\frac{1}{\lambda}\,\chi(\lambda,\mu,n-1,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,0,n,\boldsymbol{x}_+,\boldsymbol{x}_-)(\chi(0,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)$$
$$-\chi(0,\mu,n-1,\boldsymbol{x}_+,\boldsymbol{x}_-)),$$
$$\left(\frac{\partial}{\partial x_-}+\lambda\right)\chi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)-\lambda\chi(\lambda,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)=\chi(\lambda,\infty,n,\boldsymbol{x}_+,\boldsymbol{x}_-)(\chi(\infty,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)$$
$$-\chi(\infty,\mu,n+1,\boldsymbol{x}_+,\boldsymbol{x}_-)). \tag{121}$$

Taking the first equation at $\lambda=\infty$, $\mu=0$, the second at $\lambda=0$, $\mu=\infty$, one obtains the following equations:

$$\partial_+u(n)=(U(n)-U(n-1))u(n),$$
$$\partial_-v(n)=(V(n)-V(n+1))v(n). \tag{122}$$

Here we have introduced the notations

$$u(n)=\chi(\infty,0,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$
$$v(n)=\chi(0,\infty,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$
$$U(n)=\chi_r(0,0,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$
$$V(n)=\chi_r(\infty,\infty,n,\boldsymbol{x}_+,\boldsymbol{x}_-),$$

and we omit arguments $\boldsymbol{x}_+,\boldsymbol{x}_-$ common for all functions.

Equations (119) taken, respectively, at $\lambda=\infty$, $\mu=\infty$; $\lambda=0$, $\mu=0$ give rise to the following relations:

$$\partial_+V(n)=u(n)v(n)-1,$$
$$\partial_-U(n)=u(n)v(n)-1, \tag{123}$$

which together with (122) form a closed system of equations.

There is also an additional relation implied by (116), (117):

$$u(n+1)v(n)=1. \tag{124}$$

This formula shows that the functions $u(n)$ and $v(n)$ are not independent.

The system of equations (122), (123), (124) represent some special form of classical 2D Toda lattice equations. To show it, let us derive more standard forms of the Toda lattice from this system. First, it is easy to show that

$$\partial_+\partial_-\ln u(n)=u(n)v(n)-u(n-1)v(n-1),$$
$$\partial_+\partial_-\ln v(n)=u(n)v(n)-u(n+1)v(n+1). \tag{125}$$

Thus, for the function $h(n)=u(n)v(n)$ we have an equation,

$$\partial_+\partial_-\ln h(n)=2h(n)-h(n-1)-h(n+1), \tag{126}$$

which is the 2D Toda lattice equation in the Darboux form.

Since, according to (124), the functions $u(n)$ and $v(n)$ are not independent, we can write an equation for one function,

$$\partial_+\partial_-\ln u(n)=\frac{u(n)}{u(n+1)}-\frac{u(n-1)}{u(n)}. \tag{127}$$

In terms of $\theta(n)=\ln u(n)$ we have

$$\partial_+\partial_-\theta(n)=\exp(\theta(n)-\theta(n+1))-\exp(\theta(n-1)-\theta(n)). \tag{128}$$

This is just a standard form of the 2D Toda lattice equation. An original bilinear form, introduced by Toda, follows from addition formulas for the τ function, which will be given in the next section.

Now we proceed to the modified Toda lattice equation. In this case we will use the equations for generic wave functions, which are given by (118) and (120) integrated over λ and μ with some weight functions $\rho(\lambda)$ and $\eta(\mu)$ [in fact, each of them in the matrix notations is a pair of functions defined as an ''in'' unit circle and an ''out'' unit circle; see also (31)]:

$$\begin{aligned}\Phi(n)-\Phi(n+1)&=\Psi_0(n+1)\tilde{\Psi}_0(n),\\ \Phi(n)-\Phi(n-1)&=\Psi_\infty(n-1)\tilde{\Psi}_\infty(n)\end{aligned} \tag{129}$$

(we have omitted common arguments $\boldsymbol{x}_+,\boldsymbol{x}_-$).

Combining these equations, one obtains a linear problem for the modified Toda lattice equation,

$$\begin{aligned}\partial_+\Phi&=W_+(1-T)\Phi,\quad W_+=\frac{\psi_0}{T\psi_0},\\ \partial_-\Phi&=W_-(1-T^{-1})\Phi,\quad W_-=\frac{\psi_\infty}{T^{-1}\psi_\infty}.\end{aligned} \tag{130}$$

The compatibility condition gives the modified Toda lattice equations,

$$\begin{aligned}\partial_-W_+&=W_+(1-T)W_-,\\ \partial_+W_-&=W_-(1-T^{-1})W_+.\end{aligned} \tag{131}$$

These equations appeared in the literature[18] under the name of a two-dimensional Volterra chain; they were discussed quite recently in Refs. 19, 20. It is interesting to note that in Ref. 19 they arose in the geometrical context.

Proceeding further, one can obtain an equation for the function Φ, using linear operators of a modified Toda lattice to express the functions W_+, W_- via the function Φ and then substituting the final result into equations (131). Both equations give rise to the same equation,

$$\partial_+\partial_-\Phi+\partial_+\Phi\cdot\partial_-\Phi\left(\frac{1}{T\Phi-\Phi}+\frac{1}{T^{-1}\Phi-\Phi}\right)=0. \tag{132}$$

This equation was also introduced in different contexts in Refs. 19, 20. Another interesting feature of equation (132) could be its connection with the Painleve analysis of the 2DTL. Making a comparison with the one-dimensional Toda lattice case,[21] and recalling the analogous equation for the scalar KP case, it is natural to suggest that the equation (132) represents a singularity manifold equation for the 2DTL.

Equation (132) generates 2DTL and a two-dimensional Volterra chain in different ways: first, in terms of Combescure invariants, and second, in terms of expansion in λ,μ of the function $\psi(\lambda,\mu,n,\boldsymbol{x}_+,\boldsymbol{x}_-)$ which satisfies equation (132).

The Combescure invariants for (132) are given by the formulas

$$\begin{aligned}W_+&=-\frac{\partial_+\Phi}{\Delta\Phi},\\ W_-&=\frac{\partial_-T\Phi}{\Delta\Phi},\end{aligned} \tag{133}$$

(we have written down only ''left'' invariants). For a two-dimensional Volterra chain the Combescure invariants look like

$$u = W_- \exp(-\Delta^{-1}(W_- + TW_+)),$$
$$v = W_+ \exp(\Delta^{-1}(W_- + TW_+)). \tag{134}$$

Substituting (133) in (134), one obtains expressions of the full Combescure group invariants for singularity manifold equation (132).

At last, we present the formulas for the hierarchies of the two-dimensional Toda, Volterra, and 2D Toda singularity manifold equations. The derivation is very similar to the case of DS hierarchy, so we will omit it.

The 2DTL hierarchy can be represented as an equation,

$$T_+ T_-^{-1} u + \frac{a_+}{a_-} Tu = T_+ u T_-^{-1} u \left(\frac{1}{u} + \frac{a_+}{a_-} \frac{1}{T_+ T_-^{-1} T^{-1} u} \right), \tag{135}$$

where a convention for shift and difference operators is standard,

$$T_+ : \boldsymbol{x}_+ \to \boldsymbol{x}_+ + [a_+],$$
$$T_- : \boldsymbol{x}_- \to \boldsymbol{x}_+ + [a_-^{-1}],$$
$$\Delta_+ = T_+ - 1, \tag{136}$$
$$\Delta_- = (T_- - 1).$$

The 2D Volterra hierarchy reads as

$$a_- \Delta_- W_+ + (T_- W_+)(TW_-) - (T_+ W_-) W_+ = 0,$$
$$1/a_+ \Delta_+ W_- + (T_+ W_-)(T^{-1} W_+) - (T_- W_+) W_- = 0, \tag{137}$$

where $W_+ = T_+ \Psi_0 / T\Psi_0$, $W_- = T_- \Psi_\infty / T^{-1} \Psi_\infty$.

Finally, the 2DTL singularity manifold equation hierarchy looks like

$$\{T_-(T_+ - 1)\Phi\}\{T(T_- - T^{-1})\Phi\}\{T_+(1 - T^{-1})\Phi\} = \{(T_+ - 1)\Phi\}\{T_+(T_- - T^{-1})\Phi\} \times \{T_-(T - 1)\Phi\}. \tag{138}$$

This equation possesses a symmetry $T_+ \to T_-$, $T_- \to T_+$, $T \to T^{-1}$. Though this symmetry is not obvious it can be easily verified.

XI. ADDITION FORMULAS FOR TODA-LATTICE HIERARCHY

The τ-functional substitution,

$$\chi(\lambda, \mu, g) = \frac{1}{\lambda - \mu} \frac{\tau(g_{(\lambda,\mu)} g)}{\tau(g)},$$
$$g_{(\lambda,\mu)} = \frac{\lambda - \nu}{\mu - \nu}, \tag{139}$$

to (110), (118) leads to addition formulas for the τ function, the expansion of which gives the equations of the Toda lattice hierarchy in terms of the τ function.

The simplest addition formulas follow from (118),

$$\frac{1}{\mu(\lambda-\mu)}\,\tau(g\times g_d g_{(\lambda,\mu)})\tau(g)-\frac{1}{\lambda(\lambda-\mu)}\tau(g\times g_{(\lambda,\mu)})\tau(g\times g_d)$$
$$+\frac{1}{\lambda\mu}\tau(g\times g_d g_{(\lambda,0)})\tau(g\times g_{(0,\mu)})=0,\quad g_d(\nu)=\nu. \tag{140}$$

This is a functional form of addition formulas, containing four components characterized by the position of λ,μ with respect to the unit circle. Using parametrization of loop group elements in terms of times, we obtain a standard form of addition formulas. It is convenient to write down first the factorization formulas for the function $g_{(\lambda,\mu)}$, representing it as a product of functions analytic inside and outside the unit circle for different positions of λ and μ. For λ,μ both belonging to D_0 (D_∞) the factorization is trivial, because $g_{(\lambda,\mu)}$ is analytic outside (inside) the unit circle. For λ,μ in different domains it is given by

$$g_{(\lambda_+,\mu_-)}=g_d g_{(\lambda_+,0)}g_{(\infty,\mu_-)},$$
$$g_{(\lambda_-,\mu_+)}=g_d^{-1}g_{(\lambda_-,\infty)}g_{(0,\mu_+)}; \tag{141}$$

here the subscript $+,-$ just indicates the position inside or outside the unit circle.

In terms of time variables, the action of loop group elements represented in the formula (140) (taking into account the factorization formulas) looks like

$$g_{(\lambda_+,\mu_+)}\to \boldsymbol{x}_+ +[\mu_+]-[\lambda_+],$$
$$g_{(\lambda_-,\mu_-)}\to \boldsymbol{x}_- +[\mu_-^{-1}]-[\lambda_-^{-1}],$$
$$g_{(\lambda_+,\mu_-)}\to n+1,\boldsymbol{x}_+ -[\lambda_+],\boldsymbol{x}_- +[\mu_-^{-1}], \tag{142}$$
$$g_{(\lambda_-,\mu_+)}\to n-1,\boldsymbol{x}_+ +[\mu_+],\boldsymbol{x}_- -[\lambda_-^{-1}].$$

Substituting these formulas in (140), we obtain a standard form of addition formulas for the Toda lattice hierarchy. It occurs that off-diagonal components of (140) give the same formula with interchanged λ,μ, so we have three addition formulas,

$$\lambda\,\tau_n^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\lambda^{-1}])}-\mu\,\tau_{n-1}^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_-)}\tau_{n+1}^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\lambda^{-1}])}=(\lambda-\mu)\,\tau_n^{(\boldsymbol{x}_+,\boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_- +[\lambda^{-1}])}, \tag{143}$$
$$|\lambda|>1,\quad |\mu|<1,$$

$$\lambda\,\tau_{n+1}^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+ +[\lambda],\boldsymbol{x}_-)}-\mu\,\tau_n^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_-)}\tau_{n+1}^{(\boldsymbol{x}_+ +[\lambda],\boldsymbol{x}_-)}=(\lambda-\mu)\,\tau_{n+1}^{(\boldsymbol{x}_+,\boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+ +[\mu]+[\lambda],\boldsymbol{x}_-)}, \tag{144}$$
$$|\lambda|<1,\quad |\mu|<1,$$

$$\lambda\,\tau_n^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\mu^{-1}])}\tau_{n-1}^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\lambda^{-1}])}-\mu\,\tau_{n-1}^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\mu^{-1}])}\tau_n^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\lambda^{-1}])}$$
$$=(\lambda-\mu)\,\tau_{n-1}^{(\boldsymbol{x}_+,\boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+,\boldsymbol{x}_- +[\mu^{-1}]+[\lambda^{-1}])},\quad |\lambda|>1,\quad |\mu|>1. \tag{145}$$

In fact, we have used only the first equation of the pair of equations (118), but the second equation gives exactly the same set of addition formulas.

The formula (143) is a well-known simplest addition formula for the Toda lattice hierarchy. The formulas (144), (145) do not shift negative (positive) times and it is connected with the Bäcklund transformations for the KP equation.

To get more generic addition formulas, one could use the formula (110). To get more symmetric expressions, we will use a special version of this formula corresponding to the choice

$$g_0(\nu) := g_1(\nu) g_2^{-1}(\nu) = \frac{\nu-\xi}{\nu-\eta}, \tag{146}$$

where we do not specify from the beginning the position of the points λ_1, μ_1. This choice leads to the following functional form of the addition formula:

$$\begin{aligned}(\mu-\eta)(\lambda-\xi)\tau(gg_{(\xi,\eta)}g_{(\lambda,\mu)})\tau(g) - (\lambda-\eta)(\mu-\xi)\tau(gg_{(\lambda,\mu)})\tau(gg_{(\xi,\eta)}) \\ +(\lambda-\mu)(\eta-\xi)\tau(gg_{(\xi,\mu)})\tau(gg_{(\lambda,\eta)}) = 0.\end{aligned} \tag{147}$$

The formula (147) contains 16 components depending on the position of four parameters with respect to the unit circle. It is convenient to introduce the function showing the position of each parameter,

$$i_\lambda = \begin{cases} \frac{1}{2}, & |\lambda|<1, \\ -\frac{1}{2}, & |\lambda|>1. \end{cases} \tag{148}$$

Then the addition formulas take the form

$$\begin{aligned}(\mu-\eta)(\lambda-\xi)\tau_{n-i_\eta-i_\mu}^{(\boldsymbol{x}+[[\eta]]+[[\mu]])}\tau_{n-i_\lambda-i\xi}^{(\boldsymbol{x}+[[\lambda]]+[[\xi]])} - (\lambda-\eta)(\mu-\xi)\tau_{n-i_\mu-i_\xi}^{(\boldsymbol{x}+[[\xi]]+[[\mu]])}\tau_{n-i_\lambda-i_\eta}^{(\boldsymbol{x}+[[\lambda]]+[[\eta]])} \\ +(\lambda-\mu)(\eta-\xi)\tau_{n-i_\lambda-i_\mu}^{(\boldsymbol{x}+[[\mu]]+[[\lambda]])}\tau_{n-i_\eta-i_\xi}^{(\boldsymbol{x}+[[\eta]]+[[\xi]])} = 0,\end{aligned} \tag{149}$$

where

$$(\boldsymbol{x}+[[\lambda]]) = \begin{cases} (\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_-) & |\lambda|<1, \\ (\boldsymbol{x}_+, \boldsymbol{x}_- + [\lambda^{-1}]), & |\lambda|>1. \end{cases}$$

Due to its symmetric form, the formula (149) contains only five different addition formulas corresponding to all four parameters inside the unit circle: one parameter outside, two parameters outside, three parameters outside, all four parameters outside. The first and the last case just give the KP addition formulas, so we will consider only the remaining three cases. First we write down the addition formulas for the case when two variables (say, λ and μ) are inside the unit circle:

$$\begin{aligned}(\mu-\eta)(\lambda-\xi)\tau_n^{(\boldsymbol{x}_+ + [\mu], \boldsymbol{x}_- + [\eta^{-1}])}\tau_n^{(\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_- + [\xi^{-1}])} \\ -(\lambda-\eta)(\mu-\xi)\tau_n^{(\boldsymbol{x}_+ + [\mu], \boldsymbol{x}_- + [\xi^{-1}])}\tau_n^{(\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_- + [\eta^{-1}])} \\ +(\lambda-\mu)(\eta-\xi)\tau_{n-1}^{(\boldsymbol{x}_+ + [\mu]+[\lambda], \boldsymbol{x}_-)}\tau_{n+1}^{(\boldsymbol{x}_+, \boldsymbol{x}_- + [\eta^{-1}]+[\xi^{-1}])} = 0.\end{aligned} \tag{150}$$

The structure of this addition formula is similar to the formula (143), and in fact (150) is a generalized version of (143).

When one parameter (say, λ) is inside the unit circle, we obtain the following addition formula:

$$\begin{aligned}(\mu-\eta)(\lambda-\xi)\tau_{n+1}^{(\boldsymbol{x}_+, \boldsymbol{x}_- + [\eta^{-1}]+[\mu^{-1}])}\tau_n^{(\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_- + [\xi^{-1}])} \\ \times(\lambda-\eta)(\mu-\xi)\tau_{n+1}^{(\boldsymbol{x}_+, \boldsymbol{x}_- + [\xi^{-1}]+[\mu^{-1}])}\tau_n^{(\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_- + [\eta^{-1}])} \\ +(\lambda-\mu)(\eta-\xi)\tau_n^{(\boldsymbol{x}_+ + [\lambda], \boldsymbol{x}_- + [\mu^{-1}])}\tau_{n+1}^{(\boldsymbol{x}_+, \boldsymbol{x}_- + [\eta^{-1}]+[\xi^{-1}])} = 0.\end{aligned} \tag{151}$$

And the last case (λ outside the unit circle) gives

$$\begin{aligned}(\mu-\eta)(\lambda-\xi)\tau_{n-1}^{(\boldsymbol{x}_+ + [\eta]+[\mu], \boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+ + [\xi], \boldsymbol{x}_- + [\lambda^{-1}])} - (\lambda-\eta)(\mu \\ -\xi)\tau_{n-1}^{(\boldsymbol{x}_+ + [\xi]+[\mu], \boldsymbol{x}_-)}\tau_n^{(\boldsymbol{x}_+ + [\eta], \boldsymbol{x}_- + [\lambda^{-1}])}\end{aligned}$$

$$+(\lambda-\mu)(\eta-\xi)\tau_n^{(\boldsymbol{x}_+ +[\mu],\boldsymbol{x}_- +[\lambda^{-1}])}\tau_{n-1}^{(\boldsymbol{x}_+ +[\eta]+[\xi],\boldsymbol{x}_-)}=0. \tag{152}$$

The formulas (151), (152) generalize the formulas (144), (145).

ACKNOWLEDGMENTS

The first author (L.B.) is grateful to the Dipartimento di Fisica dell'Università and Sezione INFN, Lecce, for hospitality and support; L.B. also acknowledges partial support from the Russian Foundation for Basic Research under Grant No 96-01-00841 and by INTAS under Grant No. 93-166.

[1] M. Sato, RIMS Kokyuroku, Kyoto Univ. **439**, 30 (1981) (in Japanese); M. Sato and Y. Sato, in *Nonlinear Partial Differential Equations in Applied Science*, Proceedings of the U.S.–Japan Seminar, Tokyo, 1982 (North-Holland, Amsterdam, 1983).
[2] M. Jimbo and T. Miwa, Publ. RIMS Kyoto Univ. **19**, 943 (1983).
[3] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, ''Transformation groups for soliton equations,'' in *Non-Linear Integrable Equations—Classical and Quantum Theory*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983).
[4] K. Ueno and K. Takasaki, Adv. Stud. Pure Math. **1**, 4 (1984).
[5] K. Takasaki, Rev. Math. Phys. **1**, 1 (1989).
[6] G. Segal and G. Wilson, Publ. Math. I.H.E.S. **61**, 1 (1985).
[7] V. E. Zakharov and S. V. Manakov, Zap. Nauchn. Semin. LOMI **133**, 77 (1984); Funkt. Anal. Ego Pril. **19**, 11 (1985) (in Russian).
[8] L. V. Bogdanov and S. V. Manakov, J. Phys. A **21**, L537 (1988).
[9] V. E. Zakharov, in *Inverse Methods in Action*, edited by P. C. Sabatier (Springer-Verlag, Berlin, 1990).
[10] B. G. Konopelchenko, *Solitons in Multidimensions* (World Scientific, Singapore, 1993).
[11] L. V. Bogdanov and B. G. Konopelchenko, J. Phys. A **30**, 1591 (1997).
[12] L. V. Bogdanov and B. G. Konopelchenko, ''Analytic-bilinear approach to integrable hierarchies. I. Generalized KP hierarchy,'' preprint solv-int/9609009, 1996.
[13] L. V. Bogdanov, Physica D **87**, 58 (1995).
[14] P. G. Grinevich and A. Yu. Orlov, in *Modern Problems of Quantum Field Theory* (Springer-Verlag, Berlin, 1989).
[15] V. G. Kac and J. W. Van de Leur, in *Important Developments in Soliton Theory*, edited by A. S. Fokas and V. E. Zakharov (Springer-Verlag, Berlin, 1993), p. 302.
[16] L. A. Dickey, *Soliton Equations and Hamiltonian Systems* (World Scientific, Singapore, 1991).
[17] F. Nijhoff and H. W. Capel, Inverse Probl. **6**, 567 (1990).
[18] A. N. Leznov, M. V. Saveliev, and V. G. Smirnov, Lett. Math. Phys. **4**, 445 (1980).
[19] E. Ferapontov, Teor. Mat. Fiz. **110**, 86 (1997).
[20] A. B. Shabat and R. T. Yamilov, Phys. Lett. A **227**, 15 (1997).
[21] J. D. Gibbon and M. Tabor, J. Math. Phys. **26**, 1956 (1985).
[22] G. Darboux, *Lecons Sur les Systémes Orthogonaux et les Coordonneés Curvilignes* (Hermann, Paris, 1910).
[23] B. G. Konopelchenko, Rev. Math. Phys. **2**, 399 (1990).
[24] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevsky, *Theory of Solitons* (Consultants Bureau, New York, 1984).
[25] L. V. Bogdanov, Teor. Mat. Fiz. **70**, 309 (1987).
[26] L. P. Nizhnik, DAN SSSR **254**, 332 (1980) (in Russian).
[27] A. P. Veselov and S. P. Novikov, DAN SSSR **279**, 20 (1984) (in Russian).
[28] T. Shiota, Invent. Math. **83**, 333 (1986).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Supersymmetric Drinfeld–Sokolov reduction

F. Delduc and L. Gallot
Laboratoire de Physique, ENS Lyon, 46 Allée d'Italie, 69364 Lyon, France

(Received 25 February 1998; accepted for publication 2 April 1998)

The Drinfeld–Sokolov construction of integrable hierarchies, as well as its generalizations, may be extended to the case of loop superalgebras. A sufficient condition on the algebraic data for the resulting hierarchy to be invariant under supersymmetry transformation is given. The method used is a construction of the hierarchies in superspace, where supersymmetry is manifest. Several examples are discussed. © *1998 American Institute of Physics.* [S0022-2488(98)01009-3]

I. INTRODUCTION

One can find in the literature many examples of supersymmetric hierarchies (see, e.g., Refs. 1–7) but a general discussion of these hierarchies in the Drinfeld–Sokolov (DS) reduction framework is still lacking. The construction of a hierarchy in the Drinfeld–Sokolov approach relies on some algebraic data.[8–11] In particular, one needs to introduce a loop algebra, together with an integer grading of this algebra, and a semisimple positively graded element Λ in the loop algebra. One may simply replace everywhere the words ''loop algebra'' by ''loop superalgebra,'' and the construction carries through without major difficulties. The hierarchy which will be obtained in this way contains both commuting and anticommuting (Grassmann odd) fields. However, in general it will not be supersymmetric, that is to say it will not contain an odd conserved charge whose Poisson bracket with itself is the charge associated with translation invariance. Another way to put it is that it will not admit a symmetry transformation relating commuting to anticommuting fields in such a way that the commutator of two such transformations is a space translation. The simplest example of such a hierarchy will be given in Sec. III.[12,13]

Suppose now that we wish to construct supersymmetric hierarchies. It turns out that a very simple additional requirement on the input data is sufficient to ensure that the hierarchy will possess supersymmetry. Namely, the semisimple element Λ, necessarily an even element of the superalgebra, should be the square of an odd graded element Ψ of the superalgebra. This will be shown in Sec. II by constructing the hierarchy from the input data directly in superspace, where supersymmetry is manifest at all steps of the procedure.

Several examples of the general construction are gathered in Sec. III. In particular, we shall give the Drinfeld–Sokolov formulations for the supersymmetric hierarchies studied in the Gelfand–Dickey approach in Ref. 14. Some examples of supersymmetric extensions of the homogeneous hierarchies, which require the use of a twisted loop superalgebra, are also studied.

II. DS REDUCTION

In this section, we shall recapitulate the Drinfeld–Sokolov construction for the supersymmetric Korteweg–de Vries (KdV) type hierarchies in $N=1$ superspace. As a generalization of the construction in Ref. 4, we shall give the supersymmetric version of the construction given in Ref. 11. Some details, common to both the bosonic and the supersymmetric cases will be omitted and we refer to the literature[8–11] for details.

A. Superspace

We shall consider an $N=1$ superspace with one space coordinate x and one Grassmann coordinate θ. The pair of coordinates (x,θ) will be denoted by $\tilde{x}$. The supersymmetric covariant derivative is defined by

 © 1998 American Institute of Physics

TABLE I. The symmetry properties.

$[A,B\}=$	$A\in{}_+\mathcal{G}$	$A\in{}_-\mathcal{G}$
$B\in{}_+\mathcal{G}$	$-[B,A\}$	$-[\hat{B},A\}$
$B\in{}_-\mathcal{G}$	$-[B,\hat{A}\}$	$[\hat{B},\hat{A}\}$

$$D=\frac{\partial}{\partial\theta}+\theta\partial,\quad D^2=\partial\equiv\frac{\partial}{\partial x}. \tag{1}$$

We define the integration of a superfield $H(\tilde{x})$ over the $N=1$ superspace to be

$$\int d\tilde{x}\ H(\tilde{x})=\int dx\ DH|_{\theta=0}. \tag{2}$$

A superalgebra $\mathcal{G}$[15] is a Z_2 graded algebra $\mathcal{G}=\mathcal{G}_{\bar{0}}\oplus\mathcal{G}_{\bar{1}}$ with the product law $[\ ,\}$ satisfying for any $a\in\mathcal{G}_\alpha$ and $b\in\mathcal{G}_\beta$,

$$[a,b\}=-(-)^{\alpha\beta}[b,a\},\quad [a,[b,c\}\}=[[a,b\},c\}+(-)^{\alpha\beta}[b,[a,c\}\}. \tag{3}$$

We shall denote by a caret the automorphism of $\mathcal{G}$ that reverses the sign of odd elements,

$$M\in\mathcal{G},\quad M=M_{\bar{0}}+M_{\bar{1}}\to\hat{M}=M_{\bar{0}}-M_{\bar{1}}. \tag{4}$$

We shall only consider classical superalgebras[15] equipped with a bilinear invariant form denoted by $\langle\ ,\rangle$.

In the following, we shall use superfields taking values in the tensor product $\underline{\mathcal{G}}$ of a superalgebra $\mathcal{G}$ with some Grassmann algebra $\mathcal{G}r=\mathcal{G}r_{\bar{0}}\oplus\mathcal{G}r_{\bar{1}}$. An element in this space is called ''even'' if it belongs to ${}_+\underline{\mathcal{G}}=\mathcal{G}_{\bar{0}}\otimes\mathcal{G}r_{\bar{0}}\oplus\mathcal{G}_{\bar{1}}\otimes\mathcal{G}r_{\bar{1}}$ and ''odd'' if it belongs to ${}_-\underline{\mathcal{G}}=\mathcal{G}_{\bar{0}}\otimes\mathcal{G}r_{\bar{1}}\oplus\mathcal{G}_{\bar{1}}\otimes\mathcal{G}r_{\bar{0}}$. If we denote by T^a a basis vector of $\mathcal{G}$, an element A in $\underline{\mathcal{G}}$ may be expanded as

$$A=\sum_a A_aT^a,\quad A_a\in\mathcal{G}r. \tag{5}$$

The supercommutator of two elements A and B in $\underline{\mathcal{G}}$ will be defined as

$$[A,B\}=\sum_{a,b}A_aB_b[T^a,T^b\}. \tag{6}$$

With this definition, the supercommutator satisfies the symmetry properties listed in Table I.
If $A(\tilde{x})$ is a superfield taking values in ${}_+\underline{\mathcal{G}}$ or ${}_-\underline{\mathcal{G}}$, we extend the definition of the supercommutator in order to include the derivatives

$$[D,A(\tilde{x})\}=\Sigma_a(DA_a(\tilde{x}))T^a,$$

$$[D,A(\tilde{x})\}=\mp[\hat{A}(\tilde{x}),D\}\quad\text{if } A\in{}_\pm\underline{\mathcal{G}}.$$

We denote by $\mathcal{F}[J]\in\mathcal{G}r$ a functional of the superfield $J(\tilde{x})$ which takes values in ${}_+\underline{\mathcal{G}}$ or ${}_-\underline{\mathcal{G}}$. The functional derivative of $\mathcal{F}[J]$ with respect to $J(\tilde{x})$, denoted by $\delta\mathcal{F}/[\delta J(\tilde{x})]$, is defined by

$$\frac{d}{d\epsilon}\mathcal{F}[J+\epsilon r]|_{\epsilon=0}=\int d\tilde{x}\left\langle\frac{\delta\mathcal{F}}{\delta J(\tilde{x})},r(\tilde{x})\right\rangle, \tag{7}$$

where $r(\tilde{x})$ is any superfield with the same parity as $J(\tilde{x})$, and ϵ is a real parameter. Notice that the functional derivative of $\mathcal{F}$ has the same parity as $\mathcal{F}$ if J is odd, and the opposite one if J is even.

Following the lines of Ref. 11, a supersymmetric KdV system and its modified system will be associated with the choice of a sextuplet $(\mathscr{A}, d_1, d_0, \Psi; \alpha^0, \beta^0)$. $\mathscr{A}$ is an affine superalgebra (we shall not restrict ourselves to affine superalgebras possessing a purely odd root system) with vanishing center, that is a twisted loop superalgebra

$$\mathscr{A} = \mathscr{L}(\mathscr{G}, \tau) \subset \mathscr{G} \otimes C[\lambda, \lambda^{-1}] \tag{8}$$

attached to a finite dimensional classical superalgebra $\mathscr{G}$ with an automorphism τ of finite order. d_1 and d_0 are two compatible integer gradations of $\mathscr{A}$ ($[d_1, d_0] = 0$). The subsets $\mathscr{A}_m^n = \mathscr{A}^n \cap \mathscr{A}_m$ with $\mathscr{A}^n = \{X \in \mathscr{A} | d_1(X) = nX\}$, $\mathscr{A}_m = \{X \in \mathscr{A} | d_0(X) = mX\}$ define a bigrading of $\mathscr{A}$. We also assume that

$$\mathscr{A}^{>0} \subset \mathscr{A}_{\geqslant 0}, \quad \mathscr{A}^{<0} \subset \mathscr{A}_{\leqslant 0}. \tag{9}$$

In the sextuplet, Ψ is supposed to be an odd element of $\mathscr{A}$ with positive d_1-grade, $d_1(\Psi) = k\Psi$ with $k > 0$, whose square $\Lambda = \frac{1}{2}[\Psi, \Psi]$ is semisimple,

$$\mathscr{A} = \mathrm{Ker}(\mathrm{ad}\ \Lambda) \oplus \mathrm{Im}(\mathrm{ad}\ \Lambda). \tag{10}$$

We denote by $\mathscr{Z}$ the center of the kernel $\mathscr{K} = \mathrm{Ker}(\mathrm{ad}\ \Lambda)$. $\mathscr{Z}$ contains only even elements of $\mathscr{A}$.

We finally assume that $\mathscr{A}^0$ can be decomposed into the direct sum of two of its subalgebras as

$$\mathscr{A}^0 = \alpha^0 \oplus \beta^0 \quad \text{with} \quad \alpha^0 = \mathscr{A}^0_{\geqslant 0}, \quad \beta^0 \subset \mathscr{Z}^0. \tag{11}$$

The superalgebra $\mathscr{A}$ is then decomposed into $\mathscr{A} = \alpha \oplus \beta$, where $\alpha = \mathscr{A}^{>0} \oplus \alpha^0$ and $\beta = \beta^0 \oplus \mathscr{A}^{<0}$ are subalgebras. The usual decomposition[4] is recovered if one requires the supplementary condition

$$\mathscr{A}^0 \subset \mathscr{A}_0, \tag{12}$$

which implies $\alpha^0 = \mathscr{A}^0$, $\beta^0 = \{0\}$.

We introduce the odd Lax operator

$$\mathscr{L} = D + q(\tilde{x}) + \Psi, \tag{13}$$

where the odd superfield $q(\tilde{x})$ takes values in ${}_-\mathscr{A}$. We shall write evolution equations on $\mathscr{L}$ of the zero curvature type

$$\frac{\partial}{\partial t}\mathscr{L} = [A, \mathscr{L}], \tag{14}$$

where the even superfield $A(\tilde{x})$ takes values in ${}_+\mathscr{A}$. From the odd Lax operator $\mathscr{L}$ we may obtain an even operator by

$$\mathscr{L}_x = \tfrac{1}{2}[\hat{\mathscr{L}}, \mathscr{L}] = \partial + Q - \Lambda, \tag{15}$$

where the even superfield $Q(\tilde{x}) = Dq + \frac{1}{2}[\hat{q}, q\} - [\psi, q\}$ takes values in ${}_+\mathscr{A}$. It will evolve as

$$\frac{\partial}{\partial t}\mathscr{L}_x = [\hat{A}, \mathscr{L}_x]. \tag{16}$$

The phase space

$$\Theta = \{\mathscr{L} = D + q + \Psi | q \in {}_-\mathscr{A}^{<k} \cap {}_-\mathscr{A}^{\geqslant 0} \cap {}_-\mathscr{A}_{\geqslant 0}\} \tag{17}$$

will define a modified KdV type system, while a KdV type system will be obtained from

$$\mathscr{Q} = \{\mathscr{L} = D + q + \Psi | q \in {}_-\mathscr{A}^{\leqslant k} \cap {}_-\mathscr{A}_{\geqslant 0}\}. \tag{18}$$

In both cases, the construction relies on the following version of the formal dressing procedure. For any $q \in {}_-\mathscr{A}^{<k}$, there exists a unique $F(q(\tilde{x})) \in {}_+(\mathrm{Im}(\mathrm{ad}\ \Lambda))^{<0}$ such that

$$\mathscr{L}_0 = e^{\hat{F}} \mathscr{L} e^{-F} = D + H + \Psi, \tag{19}$$

where $H \in {}_-(\mathrm{Ker}(\mathrm{ad}\ \Lambda))^{<k}$. At any finite grade, the components of both F and H are differential polynomials in terms of those of q. By this we mean that the graded components of F and H are polynomials in the components of q and of a finite number of its derivatives Dq, ∂q, etc. Suppose that this is true up to grade $-n+k$ for H and $-n$ for F, then, at the next grade, Eq. (19) reduces to

$$H^{k-n-1} = P(H^{k-1},\ldots,H^{k-n},F^{-1},\ldots,F^{-n}) + [\hat{F}^{-n-1},\Psi\}. \tag{20}$$

It is not hard to show that the transformation ad Ψ may be restricted to the image Im (ad Λ), and that this restriction is an isomorphism. In other words, there is a unique element G^{-n-1-k} in ${}_+\mathrm{Im}(\mathrm{ad}\ \Lambda)$ such that

$$F^{-n-1} = [\hat{G}^{-n-1-k},\Psi\}.$$

Then Eq. (20) may be written as

$$H^{k-n-1} = P(H^{k-1},\ldots,H^{k-n},F^{-1},\ldots,F^{-n}) + [\Lambda,G^{-n-1-k}]. \tag{21}$$

From the fact that Λ is semisimple, see Eq. (10), we conclude that there exists one and only one choice for $G^{-n-1-k} \in \mathrm{Im}(\mathrm{ad}\ \Lambda)$ such that this equation holds at this grade. For later purposes, we define for any constant $b \in \mathscr{Z}^{\geqslant 0}$ the functional

$$\mathscr{H}_b(q) = \int d\tilde{x} \langle b, H(q) \rangle. \tag{22}$$

and the even superfield

$$B_b(q) = e^{-\hat{F}} b e^{\hat{F}}. \tag{23}$$

The graded components of $B_b(q)$ are differential polynomials in the components of q. Moreover, we have the property

$$[b,\mathscr{L}_0] = 0 \Rightarrow [B_b(q),\mathscr{L}] = 0. \tag{24}$$

According to a standard calculation, the functional derivative of $\mathscr{H}_b(q)$ is

$$\frac{\delta}{\delta q} \mathscr{H}_b(q) = B_b(q). \tag{25}$$

B. The modified system

The formal dressing procedure may in particular be applied to the phase space Θ defined in Eq. (17). We shall construct on this phase space an integrable system of the modified KdV type. The splitting of $\mathscr{A}$,

$$\mathscr{A} = \alpha + \beta, \tag{26}$$

yields an r-matrix,

$$\mathscr{R} = \tfrac{1}{2}(\mathscr{P}_\alpha - \mathscr{P}_\beta), \tag{27}$$

where $\mathscr{P}_\alpha$, $\mathscr{P}_\beta$ are the projectors on the respective subalgebras α, β along the subalgebras β, α. $\mathscr{R}$ satisfies the (super) classical modified Yang–Baxter equation

$$\mathscr{R}([X,\mathscr{R}(Y)\}+[\mathscr{R}(X),Y\}) = [\mathscr{R}(X),\mathscr{R}(Y)\}+\tfrac{1}{4}[X,Y\}, \tag{28}$$

which is a sufficient condition for the r-bracket $[\ ,\ \}_{\mathscr{R}}$ defined by

$$[X,Y\}_{\mathscr{R}} = [X,\mathscr{R}(Y)\}+[\mathscr{R}(X),Y\}, \tag{29}$$

to satisfy the super Jacobi identity. The evolution equation for the modified system, associated with the constant element $b \in \mathscr{Z}^{\geqslant 0}$, is given by

$$\frac{\partial}{\partial t_b} q = [\mathscr{R}(B_b(q)),\mathscr{L}\}. \tag{30}$$

Using Eq. (24), this evolution equation may be written as

$$\frac{\partial}{\partial t_b} q = [\mathscr{P}_\alpha(B_b(q)),\mathscr{L}\} = -[\mathscr{P}_\beta(B_b(q)),\mathscr{L}\}, \tag{31}$$

from which one easily checks that (30) is a consistent evolution equation. Using that $[B_a(\theta),B_b(\theta)]=0$ for all $a,b \in \mathscr{Z}^{\geqslant 0}$ and that, as a consequence of (30), $(\partial/\partial t_a)B_b(\theta) = [\mathscr{R}(B_a(\theta)),B_b(\theta)\}$, together with the modified classical Yang–Baxter equation (28) for $\mathscr{R}$, it can be shown that the flows associated with different elements of $\mathscr{Z}^{\geqslant 0}$ commute:

$$\left[\frac{\partial}{\partial t_a}, \frac{\partial}{\partial t_b}\right] q = 0, \quad \forall a,b \in \mathscr{Z}^{\geqslant 0}. \tag{32}$$

Finally, the functionals $\mathscr{H}_a(q)$ are conserved quantities for this evolution equation

$$\frac{\partial}{\partial t_b} \mathscr{H}_a(q) = 0. \tag{33}$$

We shall now describe the Hamiltonian formalism for this hierarchy of evolution equations. For this, it will be convenient to introduce a Poisson bracket on the local functionals of the fields in the space

$$\mathscr{M} = \{\mathscr{L} = D + q + \Psi \,|\, q \in _\mathscr{A}^{<k}\}, \tag{34}$$

which contains as subsets both Θ and $\mathscr{Q}$. The Poisson bracket is inferred from the existence of the r-bracket $[\ ,\ \}_{\mathscr{R}}$ and reads

$$\{f,g\}_{\mathscr{R}} = \int d\tilde{x} \left\langle \left[\frac{\delta f}{\delta q}, \frac{\delta g}{\delta q}\right]_{\mathscr{R}}, q + \Psi \right\rangle + (-1)^{[g]+1} \left\langle \mathscr{R}\,\frac{\delta f}{\delta q}, D\,\frac{\hat{\delta} g}{\delta q} \right\rangle + (-1)^{[g]+1} \left\langle \frac{\delta f}{\delta q}, D\mathscr{R}\,\frac{\hat{\delta} q}{\delta q} \right\rangle, \tag{35}$$

where $[g]=0$ if g is an even functional, and $[g]=1$ if g is odd. The important fact here is that $\Theta \subset \mathscr{M}$ is a Poisson submanifold, so that we may restrict the Poisson bracket to Θ. The evolution equation (30) may be written in the following Hamiltonian form:

$$\frac{\partial}{\partial t_b} f = \{f,\mathscr{H}_b(q)\}_{\mathscr{R}}, \tag{36}$$

and the Hamiltonians are in involution with respect to the Poisson bracket

$$\{\mathscr{H}_a(q),\mathscr{H}_b(q)\}_{\mathscr{R}} = 0. \tag{37}$$

C. The KdV type system

We shall describe here the construction of a KdV type system and its relation with the modified system. We shall require the nondegeneracy condition

$$\mathrm{Ker}(\mathrm{ad}(\Lambda)) \cap \mathscr{A}_0^{<0} = \{0\}. \tag{38}$$

The phase space of the KdV type system is the factor space $\mathscr{Q}/\mathscr{N}$ where $\mathscr{N}$ is the group of gauge transformations e^{γ} acting on $\mathscr{Q}$ as

$$e^{\gamma}: \mathscr{L} \mapsto e^{\gamma} \mathscr{L} e^{-\gamma} \quad \gamma(\tilde{x}) \in {}_+\mathscr{A}_0^{<0}. \tag{39}$$

Let us define a subspace $\mathscr{Q}_V$ of the phase space $\mathscr{Q}$ by

$$\mathscr{Q}_V = \{\mathscr{L} = D + q_V + \Psi \,|\, q_V(\tilde{x}) \in {}_-V\}, \tag{40}$$

where V is a vector space such that

$$\mathscr{A}^{<k} \cap \mathscr{A}_{\geqslant 0} = [\Psi, \mathscr{A}_0^{<0}] \oplus V. \tag{41}$$

Due to the nondegeneracy condition (38), for any field configuration q in $\mathscr{Q}$, there is a unique field-dependent gauge transformation such that the image is $\mathscr{Q}_V$. In other words, $\mathscr{Q}_V$ is a model for $\mathscr{Q}/\mathscr{N}$. When regarded as functions on $\mathscr{Q}$, the components of $q_V(q)$ are differential polynomials which freely generate the set of gauge invariant differential polynomials on $\mathscr{Q}$. We shall refer to $\mathscr{Q}_V$ as a Drinfeld–Sokolov gauge.

To construct an integrable hierarchy on $\mathscr{Q}/\mathscr{N}$, we shall first exhibit commuting flows on $\mathscr{Q}$ by means of the formal dressing procedure. Similar to the modified case, we define commuting flows on $\mathscr{Q}$ by

$$\frac{\partial}{\partial t_b} q = [\mathscr{R}(B_b(q)), \mathscr{L}], \tag{42}$$

for any $b \in \mathbb{Z}^{\geqslant 0}$. It is not a difficult task to verify that the conditions on the sextuplet $(\mathscr{A}, d_1, d_0, \Psi; \alpha^0, \beta^0)$ ensure that (42) gives a consistent evolution equation on $\mathscr{Q}$. This equation has a gauge invariant meaning, which is to say that the evolution of gauge invariant quantities is given by gauge invariant differential polynomials. This is a consequence of the uniqueness of the formal dressing procedure.[11] If we use $\mathscr{Q}_V$ as a model for $\mathscr{Q}/\mathscr{N}$, then the evolution equation (42) on $\mathscr{Q}_V$ takes the form

$$\frac{\partial}{\partial t_b} q_V = [\mathscr{R}(B_b(q_V)) + \eta_b(q_V), D + q_V + \Psi], \tag{43}$$

where $\eta_b(q_V) \in {}_+\mathscr{A}_0^{<0}$ is a uniquely determined differential polynomial in q_V. The flows thus defined on $\mathscr{Q}_V$ commute.

The Hamiltonian interpretation of these flows is as follows. Notice first that $\mathscr{Q} \subset \mathscr{M}$ is a Poisson submanifold with respect to the Poisson bracket $\{\ ,\ \}_{\mathscr{R}}$, hence the evolution equation (42) can be put in the Hamiltonian form

$$\frac{\partial}{\partial t_b} f = \{f, \mathscr{H}_b\}_{\mathscr{R}}, \tag{44}$$

where $\mathscr{H}_b$ is obtained from the formal dressing procedure. We know that if f is gauge invariant then the right-hand side of (44) is gauge invariant. Since $\mathscr{H}_b$ is a gauge invariant functional of $\mathscr{Q}$, by the uniqueness property, we are then led to suspect that the Poisson bracket $\{f, g\}_{\mathscr{R}}$ of any two gauge invariant functionals on $\mathscr{Q}$ is again gauge invariant. Close inspection of the Poisson bracket shows that this is indeed the case. As a consequence, a nonlinear Poisson bracket is defined on $\mathscr{Q}/\mathscr{N}$ by identifying the local functionals on $\mathscr{Q}/\mathscr{N}$ with gauge invariant functionals on $\mathscr{Q}$. The KdV type hierarchy on $\mathscr{Q}/\mathscr{N}$ is generated by the Hamiltonians $\mathscr{H}_b$ and this induced bracket. The evolution equations of this hierarchy (43) hence take the form (44) where f is a gauge invariant functional.

The modified type system is related to the KdV type system by a Miura map. One can first check that $\Theta \in \mathscr{Q}$ and $\mathscr{Q}/\mathscr{N}$ have the same dimension

$$\dim(\mathscr{A}^{<k}\cap\mathscr{A}^{\geqslant 0}\cap\mathscr{A}_{\geqslant 0})=\dim(\mathscr{A}^{<k}\cap\mathscr{A}_{\geqslant 0})-\dim(\mathscr{A}^{<0}\cap\mathscr{A}_{0}), \tag{45}$$

which means that the two systems have the same number of independent fields. The map

$$\mu:\Theta\to\mathscr{Q}/\mathscr{N} \tag{46}$$

induced by the projection $\mathscr{Q}\to\mathscr{Q}/\mathscr{N}$ is the Miura map. μ maps the flows of the modified system onto those of the KdV type system. From the Hamiltonian point of view, μ is a Poisson map from the linear Poisson bracket $\{\ ,\ \}_{\mathscr{R}}$ on Θ to the nonlinear Poisson bracket on $\mathscr{Q}/\mathscr{N}$ induced by the Poisson bracket $\{\ ,\ \}_{\mathscr{R}}$ on $\mathscr{Q}$.

Remark: In this construction, the Hamiltonians are associated with the elements in $\mathscr{Z}$, the center of Ker (ad Λ). Inami and Kanno[4] associated the Hamiltonians with the elements in a vector space $\tilde{\mathscr{K}}$ such that the following decomposition of Ker (ad Λ) holds

$$\mathscr{K}=[\mathscr{K},\mathscr{K}\}\oplus\tilde{\mathscr{K}}. \tag{47}$$

$[\mathscr{K},\mathscr{K}\}$ is the commutant of $\mathscr{K}$. Indeed, an evolution equation of the hierarchy may be written on the dressed operator $\mathscr{L}_0$ and reads

$$\partial_{t_b}\mathscr{L}_0=[A_b,\mathscr{L}_0\}, \tag{48}$$

where $A_b\in\mathscr{K}$. If one decomposes the dressed operator into $\mathscr{L}_0=D+H_{[\mathscr{K},\mathscr{K}\}}+H_{\tilde{\mathscr{K}}}+\Psi$, then one obtains the conservation law

$$\partial_{t_b}H_{\tilde{\mathscr{K}}}=-D(\hat{A}_b)_{\tilde{\mathscr{K}}}\Rightarrow\partial_{t_b}\int d\tilde{x}H_{\tilde{\mathscr{K}}}=0. \tag{49}$$

In the examples that we shall look at hereafter, the set of conserved quantities associated with $\tilde{\mathscr{K}}$ is identified with the one associated with $\mathscr{Z}$. This is a consequence of the fact that $\tilde{\mathscr{K}}$ may be chosen in a way such that $\mathscr{Z}$ and $\tilde{\mathscr{K}}$ are isotropic and conjugated with respect to the invariant bilinear form $\langle\ ,\ \rangle$ on $\mathscr{A}$.

III. EXAMPLES

A. A reminder of the scalar formalism

In this section, we wish to study the matrix Lax formulation for the two series of $N=2$ KdV hierarchies described in Ref. 14. Before starting, we recall some basic facts about the $N=1$ scalar Lax formulation of these hierarchies.

This formalism, which is an analog of the Gelfand–Dickey formalism, is based on the use of the algebra $\mathscr{C}$ of even $N=1$ pseudo-differential operators (Ψ DOs) of the form

$$L=\sum_{i<M}u_iD^i. \tag{50}$$

The coefficient functions u_i are $N=1$ superfields. We define the residue of the pseudodifferential operator L by res $L=u_{-1}$. The trace of L is the integral of the residue

$$\mathrm{Tr}\ L=\int d\tilde{x}\ \mathrm{res}\ L,\quad \mathrm{Tr}[L,L']=0. \tag{51}$$

The algebra $\mathscr{C}$ can be decomposed into the direct sum of two associative subalgebras $\mathscr{C}=\mathscr{C}_{\geqslant 1}\oplus\mathscr{C}_{\leqslant 0}$ where $L=L_{\geqslant 1}$ is in $\mathscr{C}_{\geqslant 1}$ if it is a differential operator containing only strictly positive powers of D, and $L=L_{\leqslant 0}$ is in $\mathscr{C}_{\leqslant 0}$ if it is a pseudodifferential operator containing only negative or zero powers of D. We denote by $\mathscr{R}$ the classical r-matrix associated with this splitting:

$$\mathscr{R}(L)=\tfrac{1}{2}(L_{\geqslant 1}-L_{\leqslant 0}). \tag{52}$$

Then the commuting flows of an $N=2$ KP hierarchy, for an operator L of the form

$$L=D^{2n}+\sum_{i=1}^{\infty} U_i D^{2n-i-1}, \tag{53}$$

are defined by

$$\partial_{t_k} L=[.\mathscr{R}(L^{k/n}),L]. \tag{54}$$

This $N=2$ KP hierarchy is the so-called nonstandard supersymmetric KP hierarchy.[16,17] Due to the fact that the r-matrix $\mathscr{R}$ is not skew symmetric, these flows are Hamiltonian with respect to a linear and two quadratic Poisson brackets. We introduce these Poisson structures on the linear functionals of L of the type $l_X(L)=\mathrm{Tr}(LX)$, X being some element in $\mathscr{C}$ which does not depend on the phase space fields $\{U_i\}$. The linear structure is

$$\{l_X,l_Y\}_{(1)}(L)=\mathrm{Tr}(L[X,Y]_{\mathscr{R}}), \tag{55}$$

and the two quadratic ones[18,14] are

$$\{l_X,l_Y\}^a_{(2)}(L)=\mathrm{Tr}(LX(LY)_{\geqslant 0}-XL(YL)_{\geqslant 0})+\int d^2\tilde{x}(-\psi_Y\ \mathrm{res}[L,X] + \mathrm{res}[L,Y]\mathrm{res}(XLD^{-1})-\mathrm{res}[L,X]\mathrm{res}(YLD^{-1})), \tag{56}$$

$$\{l_X,l_Y\}^b_{(2)}(L)=\mathrm{Tr}(LX(LY)_{\geqslant 0}-XL(YL)_{\geqslant 0})+\int d^2\tilde{x}(\psi_Y\ \mathrm{res}[L,X] + \mathrm{res}[L,Y]\mathrm{res}(LXD^{-1})-\mathrm{res}[L,X]\mathrm{res}(LYD^{-1})), \tag{57}$$

where the quantity ψ_X is defined up to a constant by $D\psi_X=\mathrm{res}[L,X]$. The two quadratic brackets are of the $abcd$ type.[19] Introducing the Hamiltonians $\mathscr{H}_k=(n/k)\mathrm{Tr}(L^{k/n})$, the $N=2$ KP evolution equations (54) may be written in the Hamiltonian form

$$\partial_{t_k} l_X(L)=\{l_X,\mathscr{H}_{k+n}\}_{(1)}(L)=\{l_X,\mathscr{H}_k\}^{a,b}_{(2)}(L). \tag{58}$$

The Hamiltonians form a family of quantities in involution with respect to any of these three brackets. The two quadratic brackets are related to each other by an invertible Poisson map p of $\mathscr{C}$ defined by

$$p(L)=\partial^{-1}L^t, \tag{59}$$

where L^t is the operator adjoint to L. One has the relation

$$\{l_X\circ p,l_Y\circ p\}^a_{(2)}=-\{l_X,l_Y\}^b_{(2)}\circ p, \tag{60}$$

which gives an equivalence between the two quadratic structures. However, there is no relation between the Hamiltonians $\mathrm{Tr}(L^{k/n})$ and $\mathrm{Tr}(p(L)^{k/(n-1)})$.

As in the bosonic case, $N=2$ KdV hierarchies are consistent reductions of the KP hierarchy. Indeed, a first series of KdV hierarchies[3,4] is obtained by considering the restriction $L=L_{\geqslant 1}$, that is to say

$$L=D^{2n}+\sum_{i=1}^{2n-2} w_i D^{2n-i-1}, \tag{61}$$

which defines a Poisson submanifold of the quadratic bracket $\{\ ,\ \}^a_{(2)}$. The Poisson algebra on the fields w_i is the $N=2\mathscr{W}_n$ algebra. The restriction

$$L=D^{2n-2}+\sum_{i=1}^{2n-3} w_i D^{2n-i-3}+D^{-1}w_{2n-2} \tag{62}$$

defines a Poisson submanifold of both the quadratic bracket $\{\ ,\ \}^b_{(2)}$ and the linear one $\{\ ,\ \}_{(1)}$ and then gives a second series of KdV hierarchies. The two series of KdV hierarchies are related to each other by the Poisson map p.

The two series of $N=2$ KdV hierarchies correspond to two series of $N=2$ modified KdV hierarchies as we shall now explain.

Let us write first the operator (61) in the factorized form

$$\begin{aligned} L=&(D+\theta_{2n-2})(D+\theta_{2n-2}+\theta_{n-1})(D+\theta_{2n-3}+\theta_{n-1})\cdots \\ &\times(D+\theta_n+\theta_2)(D+\theta_n+\theta_1)(D+\theta_1)D, \end{aligned} \tag{63}$$

where the Miura fields θ_k satisfy the following Poisson algebra of the Gardner type:

$$\{f,g\}=-\int d\tilde{x}\sum_{i,j=1}^{2n-2} C_{ij}D\left(\frac{\delta f}{\delta\theta_i}\right)\frac{\delta g}{\delta\theta_j}. \tag{64}$$

The symmetric invertible matrix C is

$$C=\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}, \tag{65}$$

where the $(n-1)\times(n-1)$ upper triangular matrix A has ones on the diagonal and above the diagonal: $A_{ij}=0$ if $j>i$ and $A_{ij}=1$ if $j\leqslant i$. The comparison between Eqs. (61) and (63) yields the Miura transformation which gives the KdV fields in terms of the Miura fields. A standard calculation[20] shows that the Miura transformation is a Poisson map from the Gardner type bracket (64) to the first quadratic bracket $\{\ ,\ \}^a_{(2)}$.

A similar statement for the second series of KdV hierarchies is the following. Let us write the operator (62) in the factorized form

$$\begin{aligned} L=&D^{-1}(D+\theta_1)(D+\theta_n+\theta_1)(D+\theta_n+\theta_2)\cdots \\ &\times(D+\theta_{2n-3}+\theta_{n-1})(D+\theta_{2n-2}+\theta_{n-1})(D+\theta_{2n-2}), \end{aligned} \tag{66}$$

which yields the Miura transformation by comparison with (62). The Miura fields satisfy the Gardner Poisson algebra (64). Then, the Miura transformation is an (anti) Poisson map from the Gardner type bracket (64) to the second quadratic bracket $\{\ ,\ \}^b_{(2)}$.

B. A matrix formulation for the first series

The first series of $N=2$ KdV hierarchies has been studied in Refs. 3 and 4. It can be described within the usual Drinfeld–Sokolov formalism with an appropriate choice of the sextuplet $(\mathscr{A},d_1,d_0,\Psi;\alpha^0,\beta^0)$ such that $\beta^0=\{0\}$.

We consider $\mathscr{A}=sl(n|n)\otimes\mathbf{C}[\lambda,\lambda^{-1}]$. A $2n\times 2n$ matrix in $sl(n|n)$ has vanishing supertrace

$$\mathrm{str}\ M=\sum_{i=1}^{2n}(-1)^{i+1}M_{ii}.$$

The center of the superalgebra $\mathscr{A}$ is spanned by the matrices $\lambda^k I$ where I is the $2n\times 2n$ identity matrix. $\mathscr{A}$ is endowed with the two compatible gradations

$$d_1=2n\lambda\partial_\lambda+\mathrm{ad}\ \mathscr{K}_I,\quad \mathscr{K}_I=\Sigma_{k=1}^{2n}\frac{2n+1-2k}{2}e_{k,k}, \tag{67}$$

$$d_0=2\lambda\partial_\lambda+\mathrm{ad}\ \mathscr{K},\quad \mathscr{K}=\Sigma_{k=1}^{2n-1}e_{k,k}. \tag{68}$$

Here d_1 is the principal gradation but d_0 is not the homogeneous gradation. Notice that this choice for d_0 is the only point where our formalism differs from that of Refs. 3 and 4. We have made this

choice in order to be able to define a Drinfeld–Sokolov gauge. The end result, though, is the same as in Refs. 3 and 4. The assumptions in (9) are satisfied, as well as the supplementary condition $\mathscr{A}^0 \subset \mathscr{A}_0$. We choose for Ψ the odd element of d_1-grade one

$$\Psi = \sum_{k=1}^{2n-1} e_{k,k+1} + \lambda e_{2n,1} \tag{69}$$

whose square Λ is semisimple.

1. The modified system

We shall first identify the modified system defined by the above choice of algebraic data with the system characterized by the factorized Lax operator (63). We parametrize the phase space Θ in (17) by

$$q(\tilde{x}) = \sum_{k=1}^{n-1} \theta_k (e_{2k-1,2k-1} + e_{2k,2k}) + \sum_{k=n}^{2n-2} \theta_k (e_{2k-2n+2,2k-2n+2} + e_{2k-2n+3,2k-2n+3}). \tag{70}$$

We have fixed the gauge freedom induced by the center of $\mathscr{A}$ by setting the coefficient of $e_{2n,2n}$ in q to zero. The explicit evaluation of the Poisson bracket (35) on Θ yields exactly the bracket given in (64). The linear problem $\mathscr{L}\Psi = 0$, where $\Psi = (\psi_1, \psi_2, \ldots, \psi_{2n})^t$, leads to the eigenvalue equation

$$\begin{aligned} L\psi_{2n} = &(D + \theta_{2n-2})(D + \theta_{2n-2} + \theta_{n-1})(D + \theta_{2n-3} + \theta_{n-1}) \cdots \\ &\times (D + \theta_n + \theta_2)(D + \theta_n + \theta_1)(D + \theta_1) D \psi_{2n} = \lambda \psi_{2n}. \end{aligned} \tag{71}$$

This allows us to identify the Hamiltonians of both systems and finally the two modified systems themselves.

2. The KdV type system

It is easily checked that the nondegeneracy condition (38) is satisfied in this case at hand. In order to identify the KdV type system associated with the above algebraic data as the nth $N = 2$ KdV hierarchy of the first series, we parametrize the phase space $\mathscr{Q}/\mathscr{N}$ by a convenient DS gauge $\mathscr{Q}_V$ whose general element is

$$q_V(\tilde{x}) = \sum_{k=1}^{2n-2} (-)^k U_k e_{2n-1,k}. \tag{72}$$

From the linear problem $\mathscr{L}\Psi = 0$, we obtain the eigenvalue equation on ψ_{2n},

$$L\psi_{2n} = (D^{2n} + U_1 D^{2n-2} + U_2 D^{2n-3} + \cdots + U_{2n-2} D)\psi_{2n} = \lambda \psi_{2n}. \tag{73}$$

The correspondence between the two parametrizations of L in (71) and (73) provides the Miura map μ: $\Theta \to \mathscr{Q}/\mathscr{N} = \mathscr{Q}_V$ which is a Poisson map when Θ is equipped with the linear bracket (35) and $\mathscr{Q}/\mathscr{N}$ is equipped with the nonlinear bracket obtained as the reduction of the linear bracket on $\mathscr{Q}$. Using the identification between the linear bracket on Θ with the Gardner type bracket in (64), we conclude that the nonlinear bracket on $\mathscr{Q}/\mathscr{N}$ coincides with the bracket $\{\ ,\ \}^a_{(2)}$ on the operator L. The identification with the nth $N = 2$ KdV hierarchy is then established.

C. A matrix formulation for the second series

The second series of $N = 2$ KdV hierarchies can be recovered within the generalized Drinfeld–Sokolov formalism with an appropriate choice of the sextuplet $(\mathscr{A}, d_1, d_0, \Psi; \alpha^0, \beta^0)$.

We consider $\mathscr{A} = sl(n|n-1) \otimes \mathbf{C}[\lambda, \lambda^{-1}]$. A $2n-1 \times 2n-1$ matrix in $sl(n|n-1)$ has vanishing supertrace

$$\mathrm{str}\, M = \sum_{i=1}^{2n-1} (-1)^{i+1} M_{ii}\,.$$

$\mathscr{A}$ is endowed with the two compatible gradations

$$d_1 = (2n-2)\lambda\partial_\lambda + \mathrm{ad}\, \mathscr{K}_{\mathrm{II}}, \quad \mathscr{K}_{\mathrm{II}} = \Sigma_{k=1}^{2n-1}(n-k)e_{k,k}\,, \tag{74}$$

$$d_0 = \lambda\partial_\lambda\,, \tag{75}$$

satisfying the assumptions in (9), but not the supplementary condition (12). Here d_0 is the homogeneous gradation but d_1 is not the principal gradation. We choose for Ψ the odd element of d_1-grade one

$$\Psi = \sum_{k=1}^{2n-2} e_{k,k+1} + \lambda(e_{2n-1,1} + e_{2n,2}) \tag{76}$$

whose square Λ is semisimple. The d_1-grade zero part of the center $\mathscr{Z}$ of the kernel is spanned by the matrix

$$\Lambda_0 = \lambda e_{2n-1,1} + e_{1,1} + 2\sum_{k=2}^{2n-2} e_{k,k} + e_{2n-1,2n-1} + \lambda^{-1} e_{1,2n-1}\,, \tag{77}$$

and we take $\beta^0 = \mathrm{span}\{\Lambda_0\}$, so that the decomposition (11) holds.

1. The modified system

We wish to identify the modified system defined by the above choice of algebraic data with the system belonging to the factorized Lax operator (66). For this we now parametrize the phase space Θ in (17) by

$$\begin{aligned} q(\tilde{x}) = {} & \xi H + \Sigma_{k=1}^{n-1}\theta_k(e_{2n-2k,2n-2k} + e_{2n-2k+1,2n-2k+1}) \\ & + \Sigma_{k=n}^{2n-2}\theta_k(e_{4n-3-2k,4n-3-2k} + e_{4n-2-2k,4n-2-2k}) + \lambda(\theta_1 + \theta_{2n-2})e_{2n-1,1}\,, \end{aligned} \tag{78}$$

where $H = (n-1)(\Sigma_{k=0}^{n-1} e_{2k+1,2k+1}) + n(\Sigma_{k=1}^{n-1} e_{2k,2k})$ is an element in $\mathscr{K}^0$ which has the property that it cannot be written as the commutator of two elements in $\mathscr{K}$. H is also dual to Λ_0 since $\langle H, \Lambda_0\rangle = 2(1-n) \neq 0$. Let us define the linear functional of q,

$$\mathscr{I}[q] \equiv \int d\tilde{x} f(\tilde{x})\langle \Lambda_0, q\rangle = 2(1-n)\int d\tilde{x} f(\tilde{x})\xi. \tag{79}$$

Using the property that Λ_0 is an element of β^0 and the expression of the Poisson structure (35), it is not difficult to show that the functional $\mathscr{I}[q]$ has vanishing Poisson bracket with any other functional $\mathscr{G}[q]$. Hence ξ can be set to zero. The explicit evaluation of the Poisson bracket (35) on Θ then yields exactly the bracket given in (64). The linear problem $\mathscr{L}\Psi = 0$ leads to the eigenvalue equation

$$\begin{aligned} L\psi_1 = {} & D^{-1}(D+\theta_1)(D+\theta_1+\theta_n)(D+\theta_2+\theta_n)\cdots \\ & \times(D+\theta_{2n-3}+\theta_{n-1})(D+\theta_{2n-2}+\theta_{n-1})(D+\theta_{2n-2})\psi_1 = 2\lambda\psi_1\,. \end{aligned} \tag{80}$$

This allows us to identify the Hamiltonians of both systems and finally the two modified systems themselves.

2. The KdV type system

In order to identify the KdV type system associated with the above algebraic data as the nth $N=2$ KdV hierarchy of the second series, we parametrize the phase space $\mathscr{Q}/\mathscr{N}$ by a convenient DS gauge $\mathscr{Q}_V$ whose general element is

$$q_V(\tilde{x}) = \sum_{k=1}^{2n-2} (-)^k U_k e_{2n-1,2n-1-k} + \xi H. \tag{81}$$

The field ξ can be set to zero by identical considerations as in the modified case. From the linear problem $\mathscr{L}\Psi = 0$, we obtain the eigenvalue equation on ψ_1,

$$L\psi_1 = D^{-1}(D^{2n-1} + U_1 D^{2n-3} + \cdots U_{2n-3} D + U_{2n-2})\psi_1 = 2\lambda\psi_1. \tag{82}$$

The correspondence between the two parametrizations of L in (80) and (82) provides the Miura map $\mu: \Theta \to \mathscr{Q}/\mathscr{N} = \mathscr{Q}_V$. The identification with the nth $N=2$ KdV hierarchy is then established by the same arguments as in the case of the first series.

IV. OTHER EXAMPLES

In this section we shall give the Drinfeld–Sokolov reduction procedure for some interesting, supersymmetric or not, integrable systems.

A. Fermionic extension of KdV

The fermionic extension of KdV[12,13] is a nonsupersymmetric bi-Hamiltonian integrable hierarchy. Its second Hamiltonian structure is the $N=1$ superconformal algebra. The evolution equations of the hierarchy are

$$\partial_{t_k} L = [L_{\geqslant 0}^{(2k+1)/n}, L], \tag{83}$$

where the Lax operator L is

$$L = \partial^2 - u - \xi\partial^{-1}\xi, \tag{84}$$

where $\xi(x)$ is a Grassmann odd field. This hierarchy can be recovered within the usual (nonsupersymmetric) Drinfeld–Sokolov formalism based on a superalgebra rather than a Lie algebra. We consider the loop superalgebra $\mathscr{A} = \mathrm{osp}(1|2) \otimes C[\lambda, \lambda^{-1}]$. The supertrace of the 3×3 matrix M is defined by

$$\mathrm{str}(M) = M_{1,1} - M_{2,2} + M_{3,3} \tag{85}$$

and a matrix in osp$(1|2)$ may be parametrized as

$$M = \begin{pmatrix} a & \alpha & b \\ \beta & 0 & \alpha \\ c & -\beta & -a \end{pmatrix}. \tag{86}$$

$\mathscr{A}$ is equipped with the two compatible gradations

$$d_1 = 4\lambda\partial_\lambda + \mathrm{ad}(e_{11} - e_{33}), \tag{87}$$

$$d_0 = \lambda\partial_\lambda. \tag{88}$$

d_1 is the principal gradation and d_0 is the homogeneous one. They satisfy the conditions (9) as well as (12). Finally, we choose for Λ the even d_1-grade two element

$$\Lambda = e_{1,3} + \lambda e_{3,1}. \tag{89}$$

Λ is semisimple. Notice that Λ cannot be written as the square of an odd element in $\mathscr{A}$. This is the reason why the system associated with this choice of Λ is not supersymmetric. The lax operator is

$$\mathscr{L} = \partial + Q + \Lambda \tag{90}$$

with $Q \in \mathscr{A}^{\leqslant 1} \cap \mathscr{A}_{\geqslant 0}$. The gauge algebra is the set of λ independent lower triangular matrices and a convenient DS gauge is

$$Q_V=\begin{pmatrix} 0 & 0 & 0 \\ \xi & 0 & 0 \\ u & -\xi & 0 \end{pmatrix}. \tag{91}$$

The linear problem $\mathscr{L}\Psi=0$ where $\Psi=(\psi_1,\psi_2,\psi_3)^t$ leads to the eigenvalue equation

$$L\psi_1=(\partial^2-u-\xi\partial^{-1}\xi)\psi_1=\lambda\psi_1. \tag{92}$$

The relation with the fermionic extension of KdV is then obtained by using standard arguments.

B. $N=4$ KdV, extensions of $N=2$ KdV

1. $N=4$ KdV

The $N=4$ KdV hierarchy[21,22,14] can be viewed as a reduction of the $N=2$ KP hierarchy[14] for which the scalar Lax operator in $N=1$ superspace is

$$L=\partial-W_1+D^{-1}W_2+D^{-1}\bar{\phi}D^{-1}\phi, \tag{93}$$

where W_1, ϕ, and $\bar{\phi}$ are even superfields, while W_2 is an odd superfield. This hierarchy can be recovered within the generalized DS formalism. We consider the algebra $\mathscr{A}=sl(2|2)\otimes C[\lambda,\lambda^{-1}]$. The supertrace of a 4×4 matrix M is

$$\mathrm{str}\ M=M_{11}-M_{22}-M_{33}+M_{44}.$$

This superalgebra is equipped with the two compatible gradations

$$d_1=2\lambda\partial_\lambda+\mathrm{ad\ Diag}(1,0,0,-1), \tag{94}$$

$$d_0=\lambda\partial_\lambda. \tag{95}$$

We choose the d_1-grade one odd element Ψ to be

$$\Psi=e_{1,2}+e_{2,4}+\lambda e_{2,1}+\lambda e_{4,2}. \tag{96}$$

$\Lambda=\Psi^2$ is semisimple. We then define

$$\beta^0=\mathscr{Z}^0=\mathrm{span}(\Lambda_0), \tag{97}$$

where $\Lambda_0=\lambda e_{4,1}+e_{1,1}+2e_{2,2}+e_{4,4}+\lambda^{-1}e_{1,4}$. The decomposition (11) holds. A convenient DS gauge for this system is then

$$\mathscr{L}=D+\begin{pmatrix} u+\xi & 1 & 0 & 0 \\ \lambda+W_1 & u-\xi & 0 & 1 \\ \phi & 0 & u-\xi & 0 \\ W_2 & \lambda & \bar{\phi} & u+\xi \end{pmatrix}. \tag{98}$$

The superfield u is the coefficient of the center of $sl(2|2)$, hence it can be set consistently to zero. The superfield ξ is the coefficient of an element H which belongs to $\mathscr{K}^0$, dual to Λ_0. Hence ξ is central in the Poisson algebra and can be set consistently to zero. The linear problem $\mathscr{L}\Psi=0$ then leads to the eigenvalue equation

$$L\psi_1=(\partial-W_1+D^{-1}W_2+D^{-1}\bar{\phi}D^{-1}\phi)\psi_1=2\lambda\psi_1, \tag{99}$$

leading to the identification of this matrix hierarchy with the $N=4$ KdV hierarchy. One can find in Ref. 14 an infinite series of hierarchies where the $N=1$ scalar Lax operator has the same pseudo-differential part as in Eq. (93), but where the differential part is more general. A matrix Lax formulation for these hierarchies is found by generalizing the above scheme to the superalgebra $sl(n|n)$.

2. Fermionic extension of the $N=2$ $a=4$ KdV hierarchy

The fermionic extension of the $N=2$ $a=4$ KdV hierarchy is very similar to the $N=4$ KdV hierarchy just discussed. It is a reduction of the $N=2$ KP hierarchy and its scalar lax operator[14] is

$$L=\partial-W_1+D^{-1}W_2+D^{-1}\bar{\phi}D^{-1}\phi, \tag{100}$$

where W_1 is an even superfield while W_2, ϕ, and $\bar{\phi}$ are odd superfields.

Its matrix interpretation is also very similar to the one of the $N=4$ KdV hierarchy as we shall see now. We choose the loop superalgebra $\mathscr{A}=sl(3|1)\otimes C[\lambda,\lambda^{-1}]$. The supertrace of a 4×4 matrix M is

$$\text{str } M=M_{11}-M_{22}+M_{33}+M_{44}.$$

$\mathscr{A}$ is equipped with the two compatible gradations

$$d_1=2\lambda\partial_\lambda+\text{ad Diag}(1,0,0,-1), \tag{101}$$

$$d_0=\lambda\partial_\lambda. \tag{102}$$

We choose the d_1-grade one odd element Ψ to be

$$\Psi=e_{1,2}+e_{2,4}+\lambda e_{2,1}+\lambda e_{4,2}. \tag{103}$$

The even element $\Lambda=\Psi^2$ is semisimple. We then take

$$\beta^0=\mathscr{Z}^0=\text{span}(\Lambda_0), \tag{104}$$

where $\Lambda_0=\lambda e_{4,1}+e_{1,1}+2e_{2,2}+e_{4,4}+\lambda^{-1}e_{1,4}$. The decomposition (11) holds.

A convenient DS gauge for this system is then

$$\mathscr{L}=D+\begin{pmatrix} u+\xi & 1 & 0 & 0 \\ \lambda+W_1 & u+2\xi & 0 & 1 \\ \phi & 0 & -u & 0 \\ W_2 & \lambda & \bar{\phi} & u+\xi \end{pmatrix}. \tag{105}$$

The element $H=e_{1,1}+2e_{2,2}+e_{4,4}$ belongs to $\mathscr{K}^0$ and is dual to Λ_0. Hence ξ is central in the Poisson algebra and can be set consistently to zero. The Abelian gauge invariance generated by the grade zero element $K=e_{1,1}+e_{2,2}-e_{3,3}+e_{4,4}$ which commutes with Ψ can be used to set u to zero. The linear problem $\mathscr{L}\Psi=0$ then leads to the eigenvalue equation

$$L\psi_1=(\partial-W_1+D^{-1}W_2+D^{-1}\bar{\phi}D^{-1}\phi)\psi_1=2\lambda\psi_1. \tag{106}$$

This scheme can be generalized to $sl(n+1|n-1)$ without difficulty, giving a matrix Lax formulation for some hierarchies described in Ref. 14.

C. $N=2$ homogeneous hierarchies

Homogeneous bosonic hierarchies[8–10] are associated with a loop algebra $\mathscr{A}=\mathscr{G}\otimes C[\lambda,\lambda^{-1}]$ for which one takes $d_1=d_0$ to be the homogeneous gradation. As a consequence, in this case there exists only an analog of the modified KdV type equations. The simplest such hierarchies have a Lax operator of the type

$$\mathscr{L}=\partial+q+\Lambda, \tag{107}$$

where the semisimple grade one element $\Lambda=\lambda H$ is such that H belongs to the Cartan subalgebra of $\mathscr{G}$. In this case q takes its values in $\mathscr{A}_0=\mathscr{G}$. The flows of the hierarchy take the particular form

$$\partial_{t_b}q=[(B_b[q])_{-1},\lambda H], \tag{108}$$

which means that only the components of q in $(\mathrm{Im}(\mathrm{ad}\,\Lambda))_0$ have nontrivial evolution equations. The components of q in $(\mathrm{Ker}(\mathrm{ad}\,\Lambda)_0)$ can be consistently set to zero at the level of evolution equations. If H is regular, $(\mathrm{Ker}(\mathrm{ad}\,\Lambda)_0)$ is the Cartan subalgebra of $\mathcal{G}$ but if H is not regular, $(\mathrm{Ker}(\mathrm{ad}\,\Lambda)_0)$ is a bigger subalgebra of $\mathcal{G}$.

The simplest examples are associated with $\mathcal{G}=sl(2)$ or $sl(3)$. When $\mathcal{G}=sl(2)$, the above construction leads to the nonlinear Schrödinger (NLS) hierarchy with the Lax operator

$$\mathcal{L}=\partial+\begin{pmatrix}0 & \phi\\ \bar{\phi} & 0\end{pmatrix}+\lambda\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{109}$$

When $\mathcal{G}$ is the rank two Lie algebra $sl(3)$, there exists a one parameter family of such integrable hierarchies. Each hierarchy is associated with a Cartan element

$$H_t=\mathrm{Diag}(1,-2t,2t-1). \tag{110}$$

In the generic case, H_t is regular. The phase space contains six fields associated with the roots of $sl(3)$. In the particular cases when $t=-\frac{1}{2}$, $\frac{1}{4}$ or 1, H_t is not regular and the phase space contains only four fields.

In the following, we shall study two examples of $N=2$ supersymmetric extensions of the homogeneous hierarchies. We consider the twisted loop superalgebra

$$\mathcal{A}=\bigoplus_{k\in Z}\lambda^{2k}\mathcal{G}_0\oplus\lambda^{2k+1}\mathcal{G}_1, \tag{111}$$

where $\mathcal{G}$ is a classical superalgebra, and $d_1=d_0$ is the homogeneous gradation. The choice $\mathcal{G}=sl(2|1)$ will lead to the $N=2$ NLS hierarchy. In the case when $\mathcal{G}=sl(3|1)$, there are only three possible choices of the odd element Ψ. They lead to generalizations of the $sl(3)$ hierarchies associated with the particular values $t=-\frac{1}{2},\frac{1}{4},1$.

1. The N=2 NLS hierarchy

The $N=2$ NLS hierarchy[23] can be viewed as a reduction of the $N=2$ KP hierarchy[14] associated with the $N=1$ scalar Lax operator

$$L=\partial-\phi D^{-1}\bar{\phi}D, \tag{112}$$

where ϕ and $\bar{\phi}$ are fermionic fields. It can be obtained in the DS formalism as a homogeneous hierarchy for which $\mathcal{G}=sl(2|1)$. We choose the supertrace of a 3×3 matrix M to be

$$\mathrm{str}\,M=M_{11}-M_{22}+M_{33}.$$

We choose the grade one odd element Ψ to be

$$\Psi=\lambda(e_{1,2}+e_{2,1}). \tag{113}$$

$\Lambda=\Psi^2=\lambda^2(e_{1,1}+e_{2,2})$ is a semisimple element of $\mathcal{A}$. Notice that $e_{1,1}+e_{2,2}$ is not the Cartan element appearing in the bosonic operator (109). It differs from it by $\frac{1}{2}(e_{1,1}+2e_{2,2}+e_{3,3})$, which is an element of the center of $\mathcal{G}_0$. Since Ψ has grade one, the phase space belongs to the grade zero part of the superalgebra, so that the matrix Lax operator $\mathcal{L}$ contains only odd superfields

$$\mathcal{L}=D+\begin{pmatrix}u_0 & \lambda & \phi\\ \lambda & u_0+u_1 & 0\\ \bar{\phi} & 0 & u_1\end{pmatrix}. \tag{114}$$

The grade zero element $H=e_{1,1}+e_{2,2}$ which belongs to the center of the kernel $\mathcal{Z}$ generates an Abelian gauge invariance. A consequence of this gauge invariance is that the field u_1 has trivial equations of evolution

$$\partial_{t_b}u_1=-\partial_{t_b}\,\mathrm{str}(Hq)=-\mathrm{str}(H[(B_b(q))_{-1},\Psi\})=0 \tag{115}$$

for any b in $\mathcal{Z}(\mathcal{K})^{>0}$. Hence u_1 can be set consistently to zero in the evolution equations. The gauge invariance can also be used to set $u_0=0$. The linear problem $\mathcal{L}\Psi=0$ then leads to the eigenvalue equation

$$L\psi_2=(\partial-\phi D^{-1}\bar{\phi}D)\psi_2=\lambda^2\psi_2 \tag{116}$$

and the hierarchy thus obtained is the $N=2$ NLS hierarchy.

2. The $sl(3|1)$ example

We choose here $\mathcal{G}=sl(3|1)$. We choose the supertrace of a 4×4 matrix M to be

$$\mathrm{str}\ M=M_{11}+M_{22}+M_{33}-M_{44}.$$

The only three possible choices of the element Ψ are

$$\Psi_i=\lambda(e_{i,4}+e_{4,i}),\quad i=1,2,3. \tag{117}$$

The semisimple element $\Lambda=\Psi^2$ is then

$$\Lambda_i=\frac{\lambda^2}{3}\,(U+(3\delta_{i,1}-1)H_{t_i}),\quad i=1,2,3. \tag{118}$$

where $U=\mathrm{Diag}(1,1,1,3)$ belongs to the center of $\mathcal{G}_0^-$ and $t_1=\frac{1}{4}$, $t_2=1$, $t_3=-\frac{1}{2}$. The $N=2$ hierarchies thus considered are generalizations of the particular bosonic $sl(3)$ hierarchies. With similar arguments as used previously, it can be shown that the phase space may be reduced to four odd superfields. For example, in the case $i=1$, the Lax operator $\mathcal{L}$ takes the form

$$\mathcal{L}=D+\begin{pmatrix} 0 & \phi_1 & \phi_2 & \lambda \\ \bar{\phi}_1 & 0 & 0 & 0 \\ \bar{\phi}_2 & 0 & 0 & 0 \\ \lambda & 0 & 0 & 0 \end{pmatrix}. \tag{119}$$

The linear problem $\mathcal{L}\Psi=0$ then leads to the eigenvalue equation

$$L\psi_4=\left(\partial-\sum_{k=1}^{2}\phi_i D^{-1}\bar{\phi}_i D\right)\psi_4=\lambda^2\psi_4 \tag{120}$$

where one recognizes the $N=1$ scalar Lax operator of the two components $N=2$ KP hierarchy. The generalization to $sl(n|1)$ is *a priori* straightforward.

[1] C. A. Laberge and P. Mathieu, Phys. Lett. B **215**, 718 (1988).
[2] P. Labelle and P. Mathieu, J. Math. Phys. **32**, 923 (1991).
[3] T. Inami and H. Kanno, J. Phys. A **25**, 3729 (1992).
[4] T. Inami and H. Kanno, Int. J. Mod. Phys. A **7**, 419 (1992).
[5] C. Morosi and L. Pizzochero, Commun. Math. Phys. **158**, 267 (1993).
[6] S. Krivonos and A. Sorin, Phys. Lett. B **357**, 94 (1995).
[7] S. Krivonos, A. Sorin, and F. Toppan, Phys. Lett. A **206**, 146 (1995).
[8] V. G. Drinfeld and V. V. Sokolov, Sov. Math. Dokl. **23**, 457 (1981); J. Sov. Math. **30**, 1975 (1985).
[9] M. F. de Groot, T. J. Hollowood, and J. L. Miramontes, Commun. Math. Phys. **145**, 57 (1992).
[10] N. J. Burroughs, M. F. de Groot, T. J. Hollowood, and J. L. Miramontes, Commun. Math. Phys. **153**, 1879 (1993).
[11] F. Delduc, L. Fehér and L. Gallot (unpublished).
[12] B. A. Kupershmidt, *Elements of Superintegrable Systems* (Reidel, Dordrecht, 1987).
[13] M. Gurses and O. Oguz, Phys. Lett. **108A**, 437 (1985).
[14] F. Delduc and L. Gallot, Commun. Math. Phys. (to be published).
[15] V. G. Kac, Commun. Math. Phys. **53**, 31 (1977).
[16] J. C. Brunelli and A. Das, Rev. Math. Phys. **7**, 1181 (1995).
[17] S. Ghosh and S. Paul, Phys. Lett. B **341**, 293 (1995).
[18] A. Das and S. Panda, Mod. Phys. Lett. A **11**, 723 (1996).
[19] L. Freidel and J. M. Maillet, Phys. Lett. B **262**, 278 (1991).

[20] L. A. Dickey, *Soliton Equations and Hamiltonian Systems*, Advanced Series in Mathematical Physics Vol. 12 (World Scientific, Singapore, 1991).
[21] F. Delduc and E. Ivanov, Phys. Lett. B **309**, 312 (1993).
[22] E. Ivanov and S. Krivonos, Phys. Lett. A. (to be published).
[23] G. Roelofs and P. Kersten, J. Math. Phys. **33**, 2185 (1992).

A one-dimensional many-body integrable model from Z_n Belavin model with open boundary conditions

Heng Fan
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Institute of Modern Physics, Northwest University, P.O. Box 105, Xi'an, 710069, People's Republic of China

Bo-Yu Hou and Guang-Liang Li
Institute of Modern Physics, Northwest University, P.O. Box 105, Xi'an, 710069, People's Republic of China

Kang-Jie Shi
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Institute of Modern Physics, Northwest University, P.O. Box 105, Xi'an, 710069, People's Republic of China

Yan-Shen Wang
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, People's Republic of China and Zhejiang Institute of Modern Physics, Zhejiang University, Hangzhou, 310027, People's Republic of China

(Received 2 December 1997; accepted for publication 25 March 1998)

We use a factorized L operator to construct an integrable model with open boundary conditions. By taking the trigonometric ($\tau \to \sqrt{-1}\infty$) and scaling limits ($w \to 0$), we get a Hamiltonian of a classical integrable system. It shows that this integrable system is similar to those found by Calogero *et al.* [J. Math. Phys. **10**, 2191 (1969); **12**, 419 (1971)]. © *1998 American Institute of Physics.* [S0022-2488(98)01709-5]

I. INTRODUCTION

In recent decades, a series of one-dimensional integrable many-body systems has been found.[1–6] One effective method to prove the integrability of the many-body systems is the Lax representation, which means that we can construct the complete set of integrals of the motion. The Lax representation for the elliptic Calogero–Moser model was found by Krichever,[7] and the Ruijsenaars–Macdonald commuting difference operators can also prove the integrability of the many-body systems.[5,6,8,9]

It is known that there are many two-dimensional exactly solved models in statistical mechanics.[10–15] The integrability is proved by the commuting transfer matrix. Recently, some relations between the one-dimensional many-body systems and the two-dimensional solvable models were found (see Refs. 16–19 and the references therein). Hasegawa[16] found that the L operator[20–22] for the two-dimensional Z_n Belavin model can be related to the Krichever's Lax matrix[7] and the commuting difference operators given by the L operators are similar to the Ruijsenaars–Macdonald difference operators. Hasegawa studied the Z_n Belavin model by imposing the periodic boundary conditions. We will study the Z_n Belavin model by imposing the open boundary conditions for the case of open boundary conditions (see Refs. 23 and 24 and the references therein). Using the factorized L operators and one solution of the reflection equation, we can construct commuting difference operators which should be equivalent to the Ruijsenaars–Macdonald difference operators. By taking a special limit, we find a trigonometric integrable model which is similar to the trigonometric model found by Calogero *et al.*[2–4] Principally, we can obtain a classical integrable system if we can find a solution to the reflection equation.

The paper is organized as follows: In Sec. II, we will introduce the Z_n symmetric Belavin model. The factorized L operator will be given in Sec. III. The commuting difference operators connected with the transfer matrix with open boundary conditions will be given in Sec. IV. The special limit is taken in Sec. V and the integrable model is found in this section. Section VI containes a summary and discussions.

 © 1998 American Institute of Physics

II. THE Z_n SYMMETRIC BELAVIN R MATRIX

The Z_n symmetric Belavin R matrix[14,15] is given as

$$R_{12}=\frac{1}{n}\sum_{\alpha\in Z_n^2} W_\alpha(z) I_\alpha\otimes I_\alpha^{-1}, \tag{1}$$

with

$$\begin{gathered} W_\alpha(z)=\frac{\sigma_\alpha(z+\eta)}{\sigma_\alpha(\eta)}, \quad I_\alpha=g^{\alpha_2}h^{\alpha_1}, \quad h_{ij}=\delta_{i+1,j}, \\ g_{ij}=\omega^i\delta_{i,j}, \quad \omega=e^{2\pi\sqrt{-1}/n}, \quad i,j\in Z_n, \\ \sigma_\alpha(z)\equiv\theta\begin{bmatrix}\frac{1}{2}+\frac{\alpha_1}{n}\\ \frac{1}{2}+\frac{\alpha_2}{n}\end{bmatrix}\quad(z,\tau), \\ \theta\begin{bmatrix}a\\ b\end{bmatrix}(z,\tau)\equiv\sum_{m\in Z} e^{\pi\sqrt{-1}(m+a)^2\tau+2\pi\sqrt{-1}(m+a)(z+b)}. \end{gathered} \tag{2}$$

This R matrix satisfies the Yang–Baxter equation (YBE),[10,11]

$$R_{12}(z_1-z_2)R_{13}(z_1-z_3)R_{23}(z_2-z_3)=R_{23}(z_2-z_3)R_{13}(z_1-z_3)R_{12}(z_1-z_2), \tag{3}$$

where $R_{12}(z)$, $R_{13}(z)$, and $R_{23}(z)$ act in $C^n\otimes C^n\otimes C^n$ with $R_{12}(z)=R(z)\otimes 1$, $R_{23}(z)=1\otimes R(z)$, etc. One can find that the R matrix also satisfies the following unitary and cross-unitary properties:

$$R_{12}(z_1-z_2)R_{21}(z_2-z_1)=\rho(z_1-z_2)\cdot\mathrm{id}, \tag{4}$$

$$R_{21}^{t_2}(z_2-z_1-nw)R_{12}^{t_2}(z_1-z_2)=\tilde{\rho}(z_1-z_2)\cdot\mathrm{id}, \tag{5}$$

where

$$\rho(z)=\frac{h(z+w)h(-z+w)}{h^2(w)}, \tag{6}$$

$$\tilde{\rho}(z)=\frac{h(z)h(-z-nw)}{h^2(w)}, \tag{7}$$

w is defined by $w\equiv n\eta$, $h(z)\equiv\sigma_0(z)$, and t_i means the transposition in the ith space.

Assume an operator matrix $L(z)$ satisfies the Yang–Baxter relation (YBR)

$$R_{12}(z_1-z_2)L_1(z_1)L_2(z_2)=L_2(z_2)L_1(z_1)R_{12}(z_1-z_2), \tag{8}$$

with $L_1(z_1)=L(z_1)\otimes 1$ and $L_2(z_2)=1\otimes L(Z_2)$. For the periodic boundary conditions, its transfer matrix is defined by $t(z)=\mathrm{tr}\,L(z)$. We can prove that these transfer matrices with different spectra commute with each other $\{[t(z_1),t(z_2)]=0\}$. For the open boundary conditions, Sklyanin[23] proposed a systematic approach to handle the problems which involve the reflection equation (RE)

$$R_{12}(z_1-z_2)K_1(z_1)R_{21}(z_1+z_2)K_2(z_2)=K_2(z_2)R_{12}(z_1+z_2)K_1(z_1)R_{21}(z_2-z_1), \tag{9}$$

where the reflecting K matrix is a solution of the RE. In order to construct the integrable models, we also need a dual reflection equation (DRE) which is associated with the cross-unitary relation of the R matrix. For Z_n Belavin model, the dual RE takes the form[24]

$$R_{12}(z_1-z_2)\tilde{K}_1(z_1)R_{21}(-z_1-z_2-nw)\tilde{K}_2(z_2)=\tilde{K}_2(z_2)R_{12}(-z_1-z_2-nw)\tilde{K}_1(z_1)R_{21}(z_2-z_1). \tag{10}$$

If we define the transfer matrix as $t(z)=\mathrm{tr}\,[\tilde{K}(z)L(z)K(z)L^{-1}(-z)]$, with the help of unitary and cross-unitary relations, one can prove that $[t(z_1),t(z_2)]=0$, which means the model under consideration is integrable.

Additionally, there is an isomorphism between $K(z)$ and $\tilde{K}(z)$:

$$\Phi: K(z)\to\tilde{K}(z)=K\left(-z-\frac{nw}{2}\right). \tag{11}$$

Given a solution $K(z)$ of the RE, we can find a solution $\tilde{K}(z)$ of the DRE. For Z_n Belavin model, a solution $K(z)$ to the RE is as follows:[24]

$$K(z)=K_0(z)K_0(0), \tag{12}$$

where

$$K_0(z)=\sum_{\alpha\in Z_n^2} U_{2\alpha}(z)\omega^{2\alpha_1\alpha_2}I_{2\alpha}, \tag{13}$$

$$U_{2\alpha}(z)=\frac{\sigma_{2\alpha}(z+c)}{\sigma_{2\alpha}(c)}, \tag{14}$$

where c is an arbitrary constant.

III. THE FACTORIZED *L* MATRIX

Jimbo *et al.*[25] defined the intertwiner of the Z_n model as an n-element column vector $\phi_{a,a+\hat{\mu}}(z)$ whose jth element is

$$\phi^{(j)}_{a,a+\hat{\mu}}(z)=\theta^{(j)}(z-nw\overline{a}_\mu,n\tau), \tag{15}$$

$$\theta^{(j)}(z-nw\overline{a}_\mu,n\tau)=\theta\begin{bmatrix}\frac{1}{2}-\frac{j}{n}\\ \frac{1}{2}\end{bmatrix}(z-nw\overline{a}_\mu,n\tau), \tag{16}$$

where $\overline{a}_\mu=a_\mu-(1/n)\Sigma_\nu a_\nu+\delta_\mu$, $a\in Z^n$, and δ_μ are some generic numbers. Using the intertwiner, the face–vertex correspondence can be written as

$$R_{12}(z_1-z_2)^{i'j'}_{ij}\phi^{(i')}_{a-\hat{\mu}-\hat{\nu},\alpha-\hat{\mu}}(z_1)\phi^{(j')}_{a-\hat{\mu},a}(z_2)=\sum_\kappa W\begin{bmatrix}a-\hat{\mu}-\hat{\nu} & a-\hat{\mu}\\ a-\hat{\kappa} & a\end{bmatrix}(z_1-z_2)\times\phi^{(j)}_{a-\hat{\mu}-\hat{\nu},a-\hat{\kappa}}(z_2)\phi^{(i)}_{a-\hat{\kappa},a}(z_1), \tag{17}$$

where

$$W\begin{bmatrix}a-\hat{\mu}-\hat{\nu} & a-\hat{\mu}\\ a-\hat{\kappa} & a\end{bmatrix}(z)$$

is the face Boltzmann weight of $A^{(1)}_{n-1}$ interaction-round-a-face (IRF) model.[25] It is defined as

$$W\begin{bmatrix} a-2\hat{\mu} & a-\hat{\mu} \\ a-\hat{\mu} & a \end{bmatrix}(z)=\frac{h(z+w)}{h(w)},$$

$$W\begin{bmatrix} a-\hat{\mu}-\hat{\nu} & a-\hat{\mu} \\ a-\hat{\mu} & a \end{bmatrix}(z)=\frac{h(z-a_{\mu\nu}w)}{h(-a_{\mu\nu}w)}, \tag{18}$$

$$W\begin{bmatrix} a-\hat{\mu}-\hat{\nu} & a-\hat{\nu} \\ a-\hat{\mu} & a \end{bmatrix}(z)=\frac{h(z)h((a_{\mu\nu}+1)w)}{h(w)h(a_{\mu\nu}w)}. \tag{19}$$

The other face Boltzmann weights are defined as zeroes where $a_{\mu\nu}=\overline{a}_{\mu}-\overline{a}_{\nu}$. At the same time, we can also find the n-element row vectors $\tilde{\phi}$ and $\overline{\phi}$ which satisfy the following relations[20–22,25]

$$\tilde{\phi}^{(k)}_{a-\hat{\mu},a}(z)\,\phi^{(k)}_{a-\hat{\nu},a}(z)=\delta_{\mu\nu}, \tag{20}$$

$$\overline{\phi}^{(k)}_{a,a+\hat{\mu}}(z)\,\phi^{(k)}_{a,a+\hat{\nu}}(z)=\delta_{\mu\nu}. \tag{21}$$

The above equation can also be written as

$$\sum_{\mu}\ \phi_{a-\hat{\mu},a}(z)\,\tilde{\phi}_{a-\hat{\mu},a}(z)=I, \tag{22}$$

$$\sum_{\mu}\ \phi_{a,a+\hat{\mu}}(z)\,\overline{\phi}_{a,a+\hat{\mu}}(z)=I. \tag{23}$$

Using those results, the face–vertex correspondence can be written in other forms:

$$\tilde{\phi}^{(i)}_{a-\hat{\mu},a}(z_1)R_{12}(z_1-z_2)^{i'j'}_{ij}\,\phi^{(j')}_{a-\hat{\nu},a}(z_2)=\sum_{\kappa}\ W\begin{bmatrix} a-\hat{\mu}-\hat{\kappa} & a-\hat{\nu} \\ a-\hat{\mu} & a \end{bmatrix}(z_1-z_2)$$
$$\times\,\phi^{(j)}_{a-\hat{\mu}-\hat{\kappa},a-\hat{\mu}}(z_2)\,\tilde{\phi}^{(i')}_{a-\hat{\mu}-\hat{\kappa},a-\hat{\nu}}(z_1), \tag{24}$$

$$\overline{\phi}^{(j)}_{a,a+\hat{\mu}}(z_2)R_{12}(z_1-z_2)^{i'j'}_{ij}\,\phi^{(i')}_{a-\hat{\nu}-\hat{\kappa},a-\hat{\mu}}(z_1)=\sum_{\kappa}\ W\begin{bmatrix} a & a+\hat{\nu} \\ a+\hat{\mu} & a+\hat{\mu}+\hat{\kappa} \end{bmatrix}(z_1-z_2)$$
$$\times\,\phi^{(i)}_{a+\hat{\mu},a+\hat{\mu}+\hat{\kappa}}(z_1)\,\overline{\phi}^{(j')}_{a+\hat{\nu},a+\hat{\mu}+\hat{\kappa}}(z_2), \tag{25}$$

$$\tilde{\phi}^{(i)}_{a-\hat{\mu},a}(z_1)\,\tilde{\phi}^{(j)}_{a-\hat{\mu}-\hat{\nu},a-\hat{\mu}}(z_2)R_{12}(z_1-z_2)^{i'j'}_{ij}=\sum_{\kappa}\ W\begin{bmatrix} a-\hat{\mu}-\hat{\nu} & a-\hat{\kappa} \\ a-\hat{\mu} & a \end{bmatrix}(z_1-z_2)$$
$$\times\,\tilde{\phi}^{(j')}_{a-\hat{\kappa},a}(z_2)\,\tilde{\phi}^{(i')}_{a-\hat{\mu}-\hat{\nu},a-\hat{\kappa}}(z_1), \tag{26}$$

$$\overline{\phi}^{(i)}_{a+\hat{\mu},a+\hat{\mu}+\hat{\nu}}(z_1)\,\overline{\phi}^{(j)}_{a,a+\hat{\mu}}(z_2)R_{12}(z_1-z_2)^{i'j'}_{ij}=\sum_{\kappa}\ W\begin{bmatrix} a & a+\hat{\kappa} \\ a+\hat{\mu} & a+\hat{\mu}+\hat{\nu} \end{bmatrix}(z_1-z_2)$$
$$\times\,\overline{\phi}^{(j')}_{a+\hat{\kappa},a+\hat{\mu}+\hat{\nu}}(z_2)\,\overline{\phi}^{(i')}_{a,a+\hat{\kappa}}(z_1). \tag{27}$$

With the help of those face–vertex correspondence relations, we then can construct the $L(z)$ matrix[20–22] which meet the YBR. Let

$$f(a,\mu,z)^{j}_{i}=\phi^{(i)}_{a-\hat{\mu},a}(z+\xi_1)\,\tilde{\phi}^{(j)}_{a-\hat{\mu},a}(z+\xi_2), \tag{28}$$

where ξ_1 and ξ_2 are arbitrary complex numbers. Then from the face–vertex correspondence relations (17) and (26), we have

$$R(z_1-z_2)_{ij}^{i'j'}f(a-\hat{\mu},\nu,z_1)_{i'}^{i''}f(a,\nu,z_2)_{j'}^{j''}+\mu\rightleftarrows\nu$$
$$=f(a-\hat{\mu},\nu,z_2)_{j}^{j'}f(a,\nu,z_1)_{i}^{i'}R(z_1-z_2)_{i'j'}^{i''j''}+\mu\rightleftarrows\nu, \tag{29}$$

where $\mu\leftrightarrow\nu$ means the same form as the term before them, while the μ and ν exchange with each other. We introduce the difference operator Γ_μ,

$$\Gamma_\mu f(a)=f(a+\hat{\mu})\Gamma_\mu, \tag{30}$$

and we have

$$R(z_1-z_2)_{ij}^{i'j'}\Gamma_\nu f(a,\nu,z_1)_{i'}^{i''}\Gamma_\mu f(a,\nu,z_2)_{j'}^{j''}+\mu\leftrightarrow\nu$$
$$=\Gamma_\nu f(a,\nu,z_2)_{j}^{j'}\Gamma_\mu f(a,\nu,z_1)_{i}^{i'}R(z_1-z_2)_{i'j'}^{i''j''}+\mu\leftrightarrow\nu. \tag{31}$$

We can find that the above equation is sufficient to prove the following YBR:

$$R(z_1-z_2)_{ij}^{i'j'}L(a,z_1)_{i'}^{i''}L(a,z_2)_{j'}^{j''}=L(a,z_2)_{j}^{j'}L(a,z_1)_{i}^{i'}R(z_1-z_2)_{i'j'}^{i''j''}, \tag{32}$$

where $L(a,z_k)=\Sigma_\mu\Gamma_\mu f(a,\mu,z_k)$. Let

$$g(a,\mu,z)_i^j=\phi_{a,a+\hat{\mu}}^{(i)}(z+\xi_2)\overline{\phi}_{a,a+\hat{\mu}}^{(j)}(z+\xi_1). \tag{33}$$

There is

$$\sum_\nu\Gamma_\nu f(a,\nu,z)_i^j\sum_\mu\Gamma_{-\mu}g(a,\mu,z)_j^k=\delta_{ik}. \tag{34}$$

It shows that $L^{-1}(a,z)=\Sigma_\mu\Gamma_{-\mu}g(a,\mu,z)$. Then the transfer matrix with open boundary conditions is

$$t(z)=\mathrm{tr}\,\tilde{K}(z)L(a,z)K(z)L^{-1}(a,-z)$$
$$=\mathrm{tr}\,\tilde{K}(z)\sum_\nu\Gamma_\nu f(a,\nu,z)K(z)\sum_\mu\Gamma_{-\mu}g(a,\mu,-z)$$
$$=\sum_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu F_{\mu\nu}^{(1)}(a,z)F_{\mu\nu}^{(2)}(a,z)=\sum_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu G_{\mu\nu}(a,z), \tag{35}$$

with

$$F_{\mu\nu}^{(1)}(a,z)=\tilde{\phi}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}^{(i)}(z+\xi_2)K(z)_i^j\phi_{a,a+\hat{\mu}}^{(j)}(-z+\xi_2), \tag{36}$$

$$F_{\mu\nu}^{(2)}(a,z)=\overline{\phi}_{a,a+\hat{\mu}}^{(k)}(-z+\xi_1)\tilde{K}(z)_k^l\phi_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}^{(l)}(z+\xi_1), \tag{37}$$

$$G_{\mu\nu}(a,z)=F_{\mu\nu}^{(1)}(a,z)F_{\mu\nu}^{(2)}(a,z). \tag{38}$$

IV. THE CALCULATION OF TRANSFER MATRIX $t(z)$

Substituting the $K(z)$ (12) into Eq. (36), we obtain

$$F_{\mu\nu}^{(1)}(a,z)=\tilde{\phi}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}^{(i)}(z)\sum_{\gamma\in Z_n^2}U_{2\gamma}(z)\omega^{2\gamma_1\gamma_2}[g^{2\gamma_2}h^{2\gamma_1}K_0(0)]_i^j\phi_{a,a+\hat{\mu}}^{(j)}(-z). \tag{39}$$

Here we let $\xi_1=\xi_2=0$. We can prove that

$$g^{\beta-1}h^{\alpha}K_0(0)\phi_{a,a+\hat{\mu}}(z)=(-1)^{\beta-1}e^{2\pi\sqrt{-1}(\alpha/n)(\alpha\tau/2+z-nw\overline{a}_\mu+1/2)}\phi_{a,a+\hat{\mu}}(-z+2nw\overline{a}_\mu-\alpha\tau-\beta) \tag{40}$$

(see Appendix). The above equation (39) can be written as

$$F^1_{\mu\nu}(a,z)=\sum_{\gamma\in Z_n^2}U_{2\gamma}(z)e^{2\pi\sqrt{-1}2\gamma_1\gamma_2/n}e^{2\pi\sqrt{-1}(2\gamma_1/n)(2\gamma_1\tau/2-z-nw\overline{a}_\mu+1/2)}\tilde{\phi}^{(i)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(z)$$
$$\times\phi^{(i)}_{a,a+\hat{\mu}}(z+2nw\overline{a}_\mu-2\gamma_1\tau-2\gamma_2-1). \tag{41}$$

From Eq. (23) we know $\tilde{\phi}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(z)$ can be obtained from the inverse of the matrix $\tilde{M}$ whose elements are

$$\tilde{M}_{i\lambda}(z)=\phi^{(i)}_{a+\hat{\mu}-\hat{\lambda},a+\hat{\mu}}(z)=\theta^{(i)}(z-nw(\overline{a}_\lambda+\delta_{\mu\lambda}-1)). \tag{42}$$

So we get

$$\tilde{\phi}^{(i)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(z)\phi^{(i)}_{a,a+\hat{\mu}}(z+2nw\overline{a}_\mu-2\gamma_1\tau-(2\gamma_2+1))=\frac{\det\tilde{M}}{\det\tilde{M}}. \tag{43}$$

Substituting the ν column elements of matrix $\tilde{M}$ with the corresponding elements of column vector $\phi_{a,a+\hat{\mu}}(z+2nw\overline{a}_\mu-2\gamma_1\tau-(2\gamma_2+1))$ while keeping other matrix elements unchanged, we obtain the matrix $\tilde{M}'$.

Suppose the elements of a matrix A are defined as $A_{ij}=\theta^{(i)}(nz_j)$. One can prove that the determinant of the matrix A has the results[20–22]

$$\det A=C(\tau)h\left(\sum_i z_i-\frac{n-1}{2}\right)\prod_{i<k}h(z_i-z_k). \tag{44}$$

Using this result, we obtain

$$\tilde{\phi}^{(i)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(z)\phi^{(i)}_{a,a+\hat{\mu}}(z+2nw\overline{a}_\mu-2\gamma_1\tau-(2\gamma_2+1))$$
$$=\frac{h\left(-z+w\delta+w(1-n)+\frac{n-1}{2}-w(\overline{a}_\nu+\delta_{\mu\nu}-1+\overline{a}_\mu)+\frac{1}{n}(2\gamma_1\tau+2\gamma_2+1)\right)}{h\left(-z+w\delta+w(1-n)+\frac{n-1}{2}\right)}$$
$$\times\prod_{j\neq\nu}\frac{h\left(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu)+\frac{1}{n}(2\gamma_1\tau+2\gamma_2+1)\right)}{h(-w(\overline{a}_j+\delta_{\mu j}-\overline{a}_\nu-\delta_{\mu\nu}))} \tag{45}$$

($\delta=\Sigma_i\delta_i$). With the help of the formula

$$\sigma_0\left(z+\frac{1}{n}(\alpha\tau+\beta)\right)=e^{-2\pi\sqrt{-1}(\alpha/n)(\alpha\tau/2n+z+1/2+\beta/n)}\sigma_{\alpha,\beta}(z,\tau), \tag{46}$$

finally we obtain

$$F^{(1)}_{\mu\nu}(a,z)=\sum_{\gamma\in Z_n^2}U_{2\gamma}(z)e^{-2\pi\sqrt{-1}(2\gamma_1\gamma_2/n)}e^{2\pi\sqrt{-1}(2\gamma_1/n)}$$
$$\times\frac{\sigma_{2\gamma_1,2\gamma_2+1}(-z+w\delta+w(1-n)+(n-1)/2-w(\overline{a}_\nu+\delta_{\mu\nu}-1+\overline{a}_\mu))}{\sigma_0(-z+w\delta+w(1-n)+(n-1)/2)}$$

$$\prod_{j\neq\nu}\frac{\sigma_{2\gamma_1,2\gamma_2+1}(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu))}{\sigma_0(-w(\overline{a}_j+\delta_{\mu j}-\overline{a}_\nu-\delta_{\mu\nu}))}. \tag{47}$$

Substituting the $\tilde{K}(z)$ into Eq. (37), considering the isomorphism relation between $\tilde{K}(z)$ and $K(z)$(11), we obtain

$$F^{(2)}_{\mu\nu}(a,z)=\overline{\phi}^{(k)}_{a,a+\hat{\mu}}(-z)\sum_{\gamma\in Z_n^2}U_{2\gamma}\left(-z-\frac{nw}{2}\right)\omega^{2\gamma_1\gamma_2}[g^{2\gamma_2}h^{2\gamma_1}K_0(0)]^l_k\phi^{(l)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(z). \tag{48}$$

Using Eq. (40), we have

$$\begin{aligned}F^{(2)}_{\mu\nu}(a,z)=&\sum_{\gamma\in Z_n^2}U_{2\gamma}\left(-z-\frac{nw}{2}\right)e^{2\pi\sqrt{-1}(2\gamma_1\gamma_2/n)}e^{2\pi\sqrt{-1}(2\gamma_1/n)(2\gamma_1\tau/2+z-nw(\overline{a}_\nu+\delta_{\mu\nu}-1)+1/2)}\\&\times\overline{\phi}^{(k)}_{a,a+\hat{\mu}}(-z)\phi^{(k)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(-z+2nw(\overline{a}_\nu+\delta_{\mu\nu}-1)-2\gamma_1\tau-2\gamma_2-1).\end{aligned} \tag{49}$$

From Eq. (23), we also know $\overline{\phi}_{a,a+\hat{\mu}}(-z)$ can be obtained from the inverse of the matrix $\overline{M}$ whose elements are

$$\overline{M}_{i\lambda}(-z)=\phi^{(i)}_{a,a+\hat{\lambda}}(-z)=\theta^{(i)}(-z-nw\overline{a}_\lambda). \tag{50}$$

So, we obtain

$$\overline{\phi}^{(k)}_{a,a+\hat{\mu}}(-z)\phi^{(k)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(-z+2nw(\overline{a}_\nu+\delta_{\mu\nu}-1)-2\gamma_1\tau-2\gamma_2-1)=\frac{\det\overline{M}'}{\det\overline{M}}. \tag{51}$$

Substituting the μ column elements of matrix $\overline{M}$ with the corresponding elements of column vector $\phi_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(-z+2nw(\overline{a}_\nu+\delta_{\mu\nu}-1)-2\gamma_1\tau-2\gamma_2-1)$, holding other matrix elements unchanged, we get the matrix M'. Using this result of (44), we obtain

$$\begin{aligned}&\overline{\phi}^{(k)}_{a,a+\hat{\mu}}(-z)\phi^{(k)}_{a+\hat{\mu}-\hat{\nu},a+\hat{\mu}}(-z+2nw(\overline{a}_\nu+\delta_{\mu\nu}-1)-2\gamma_1\tau-2\gamma_2-1)\\&\quad=\frac{h(z+w\delta+(n-1)/2-w(\overline{a}_\mu+\delta_{\mu\nu}-1+\overline{a}_\nu)+(1/n)(2\gamma_1\tau+2\gamma_2+1))}{h(z+w\delta+(n-1)/2)}\\&\qquad\times\prod_{j\neq\mu}\frac{h(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu)+(1/n)(2\gamma_1\tau+2\gamma_2+1))}{h(-w(\overline{a}_j-\overline{a}_\mu))}.\end{aligned} \tag{52}$$

Similarly, with the help of Eq. (46), we get

$$\begin{aligned}F^{(2)}_{\mu\nu}(a,z)=&\sum_{\gamma\in Z_n^2}U_{2\gamma}\left(-z-\frac{nw}{2}\right)e^{-2\pi\sqrt{-1}(2\gamma_1\gamma_2/n)}e^{2\pi\sqrt{-1}(2\gamma_1/n)}\\&\times\frac{\sigma_{2\gamma_1,2\gamma_2+1}(z+w\delta+(n-1)/2-w(\overline{a}_\mu+\delta_{\mu\nu}-1+\overline{a}_\nu))}{\sigma_0(z+w\delta+(n-1)/2)}\\&\times\prod_{j\neq\mu}\frac{\sigma_{2\gamma_1,2\gamma_2+1}(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu))}{\sigma_0(-w(\overline{a}_j-\overline{a}_\mu))}.\end{aligned} \tag{53}$$

V. TAKING THE TRIGONOMETRIC AND SCALING LIMITS

When $\tau\to\sqrt{-1}\infty$, considering $2\gamma_1/n\neq m$ (m is an integer number), we can find that

$$\theta\begin{bmatrix}\frac{1}{2}+\frac{2\gamma_1}{n}\\ \frac{1}{2}+\frac{2\gamma_2+1}{n}\end{bmatrix}(z,\tau)\to f(\tau)e^{2\pi\sqrt{-1}(2\gamma_1/n+1/2+m)(z+1/2+(2\gamma_2+1)/n)}. \tag{54}$$

Then,

$$U_{2\gamma}(z)\to e^{2\pi\sqrt{-1}(2\gamma_1/n+1/2+m)z} \tag{55}$$

$$\begin{aligned}F^{(1)}_{\mu\nu}(a,z)\to&\sum_{\gamma_1}\frac{e^{2\pi\sqrt{-1}(2\gamma_1/n+1/2+m)z}e^{2\pi\sqrt{-1}(2\gamma_1/n)}}{\sin\pi(-z+w\delta+w(1-n)+(n-1)/2)}\prod_{j\neq\nu}\frac{1}{\sin\pi(\bar{a}_j+\delta_{\mu j}-\bar{a}_\nu-\delta_{\mu\nu})}\\&\times\sum_{\gamma_2}e^{-2\pi\sqrt{-1}(2\gamma_1\gamma_2/n)}f^n(\tau)e^{2\pi\sqrt{-1}(2\gamma_1/n+1/2+m)(-z+n-1/2-nw\bar{a}_\mu+2\gamma_2)}\\&\times\sum_{\gamma_2}e^{2\pi\sqrt{-1}(2\gamma_1\gamma_2/n)}=\sum_{\gamma_2}(e^{2\pi\sqrt{-1}(2\gamma_1/n)})^{\gamma_2}=0.\end{aligned} \tag{56}$$

We can see that from $F^{(2)}_{\mu\nu}(a,z)$, we can obtain the same result. So we only pay attention to the case where $2\gamma_1/n=m$. The next step is to consider this case.

We know there is an arbitrary parameter c in $U_{2\gamma}(z)$. Let $c=\epsilon\tau+c'$ with $\epsilon<1/n$. We have

$$\theta\begin{bmatrix}\frac{1}{2}+\frac{2\gamma_1}{n}\\ \frac{1}{2}+\frac{2\gamma_2}{n}\end{bmatrix}(z+c,\tau)=e^{-2\pi\sqrt{-1}\epsilon(2\epsilon\tau/n+z+c'+1/2+2\gamma_2/n)}\theta\begin{bmatrix}\frac{1}{2}+\frac{2\gamma_1}{n}+\epsilon\\ \frac{1}{2}+\frac{2\gamma_2}{n}\end{bmatrix}(z+c',\tau). \tag{57}$$

Then

$$\begin{aligned}U_{2\gamma}(z)=&e^{-2\pi\sqrt{-1}\epsilon z}\frac{\theta\begin{bmatrix}\frac{1}{2}+\frac{2\gamma_1}{n}+\epsilon\\ \frac{1}{2}+\frac{2\gamma_2}{n}\end{bmatrix}(z+c',\tau)}{\theta\begin{bmatrix}\frac{1}{2}+\frac{2\gamma_1}{n}+\epsilon\\ \frac{1}{2}+\frac{2\gamma_2}{n}\end{bmatrix}(c',\tau)},\\&\to e^{-2\pi\sqrt{-1}\epsilon z}e^{2\pi\sqrt{-1}(1/2+2\gamma_1/n+\epsilon+m)z},\end{aligned} \tag{58}$$

which have no relation to γ_2 and do not include any dynamics variables like $\bar{a}_\mu$, so we handle it as an irrelevant constant when taking the trigonometric limit. The conclusion (56) still holds when $c=\epsilon\tau+c'$. Now, we can obtain ($2\gamma_1/n=m$)

$$\begin{aligned}F^{(1)}_{\mu\nu}(a,z)=&\sum_{\gamma_2}\frac{\sin\pi(-z+w\delta+w(1-n)+(n-1)/2-w(\bar{a}_\nu+\delta_{\mu\nu}-1+\bar{a}_\mu)+(2\gamma_2+1)/n)}{\sin\pi(-z+w\delta+w(1-n)+(n-1)/2)}\\&\times\prod_{j\neq\nu}\frac{\sin\pi(-w(\bar{a}_j+\delta_{\mu j}-1+\bar{a}_\mu)+(2\gamma_2+1)/n)}{\sin\pi(-w(\bar{a}_j+\delta_{\mu j}-\bar{a}_\nu-\delta_{\mu\nu}))},\end{aligned} \tag{59}$$

$$F^{(2)}_{\mu\nu}(a,z)=\sum_{\gamma_2}\ \frac{\sin\ \pi(z+w\delta+(n-1)/2-w(\overline{a}_\nu+\delta_{\mu\nu}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}{\sin\ \pi\left(z+w\delta+\dfrac{n-1}{2}\right)}$$

$$\times\prod_{j\neq\mu}\ \frac{\sin\ \pi(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu)+(2\gamma_2+1)/n)}{\sin\ \pi(-w(\overline{a}_j-\overline{a}_\mu))}. \tag{60}$$

For the n conservations that do not depend on the spectrum parameter z, when $z\to-\sqrt{-1}\infty$, we can get the integrable Hamiltonian of the system. So there are

$$F^{(1)}_{\mu\nu}(a,z)=\sum_{\gamma_2}\ e^{\pi\sqrt{-1}(-w(\overline{a}_\nu+\delta_{\mu\nu}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}$$

$$\times\prod_{j\neq\nu}\ \frac{\sin\ \pi(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}{\sin\ \pi(-w(\overline{a}_j+\delta_{\mu j}-\overline{a}_\nu-\delta_{\mu\nu}))}, \tag{61}$$

$$F^{(2)}_{\mu\nu}(a,z)=\sum_{\gamma_2}\ e^{-\pi\sqrt{-1}(-w(\overline{a}_\nu+\delta_{\mu\nu}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}$$

$$\times\prod_{j\neq\mu}\ \frac{\sin\ \pi(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}{\sin\ \pi(-w(\overline{a}_j+\overline{a}_\mu))}. \tag{62}$$

Expanding $\sin\ \pi(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)$ in $F^{(1)}_{\mu\nu}(a,z)$ and $\sin\ \pi(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu)+(2\gamma_2'+1)/n)$ in $F^{(2)}_{\mu\nu}(a,z)$ respectively as follows,

$$\frac{1}{2\sqrt{-1}}\left[e^{\pi\sqrt{-1}(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}-e^{-\pi\sqrt{-1}(-w(\overline{a}_j+\delta_{\mu j}-1+\overline{a}_\mu)+(2\gamma_2+1)/n)}\right], \tag{63}$$

$$\frac{1}{2\sqrt{-1}}\left[e^{\pi\sqrt{-1}(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu)+(2\gamma_2+1)/n)}-e^{-\pi\sqrt{-1}(-w(\overline{a}_j+\delta_{\mu\nu}-1+\overline{a}_\nu)+(2\gamma_2+1)/n)}\right], \tag{64}$$

and summing over γ_2 and γ_2' we can find that all the terms of $F^{(1)}_{\mu\nu}(a,z)$ and $F^{(2)}_{\mu\nu}(a,z)$ have the forms

$$\sum_{\gamma_2}\ e^{(2\pi\sqrt{-1}/n)m\gamma_2},\quad m=\underbrace{1\pm1\pm\cdots\pm1}_{n-1}=n,n-2,n-4,\ldots,-n+2, \tag{65}$$

and

$$\sum_{\gamma_2'}\ e^{(2\pi\sqrt{-1}/n)m\gamma_2'},\quad m=\underbrace{1\pm1\pm\cdots\pm1}_{n-1}=n,n-2,n-4,\ldots,-n+2. \tag{66}$$

Only for the case $m=\pm n,0$ are the above summations not equal to zeroes. When $n=$even, m can take n or 0, but it is rather complex to calculate the case $m=0$. For simplicity, we only consider $n=$odd, so m can only take n. We have

$$F^{(1)}_{\mu\nu}(a,z)\rightsquigarrow e^{\pi\sqrt{-1}\Sigma_i(-w(\overline{a}_i+\delta_{\mu i}-1+\overline{a}_\mu)+1/n)}\prod_{j\neq\nu}\ \frac{1}{\sin\ \pi(-w(\overline{a}_j+\delta_{\mu j}-\overline{a}_\nu-\delta_{\mu\nu}))}, \tag{67}$$

$$F^{(2)}_{\mu\nu}(a,z)\rightsquigarrow e^{-\pi\sqrt{-1}\Sigma_i(-w(\overline{a}_i+\delta_{\mu\nu}-1+\overline{a}_\nu)+1/n)}\prod_{j\neq\mu}\ \frac{1}{\sin\ \pi(-w(\overline{a}_j-\overline{a}_\mu))}, \tag{68}$$

$$G_{\mu\nu}(a,z)\rightsquigarrow e^{\pi\sqrt{-1}(nw(\bar{a}_\nu-\bar{a}_\mu)+nw\delta_{\mu\nu}-w)}\prod_{j\neq\nu}\frac{1}{\sin\pi(-w(\bar{a}_j+\delta_{\mu j}-\bar{a}_\nu-\delta_{\mu\nu}))}$$

$$\times\prod_{k\neq\mu}\frac{1}{\sin\pi(-w(\bar{a}_k-\bar{a}_\mu))}. \tag{69}$$

For $\bar{a}_j-\bar{a}_k=a_j-a_k+\delta_j-\delta_k$, and $H=\Sigma_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu G_{\mu\nu}(a,z)$, in which $\Gamma_\mu f(a)=f(a+\hat{\mu})\Gamma_\mu$, we obtain,

$$H=\sum_{\mu\nu}\Gamma_\nu\left[e^{\pi\sqrt{-1}(nw(a_\nu+\delta\nu))}\prod_{j\neq\nu}\frac{1}{\sin\pi(-w(a_j-a_\nu+\delta_j-\delta_\nu))}\right]\Gamma_{-\mu}\left[e^{-\pi\sqrt{-1}(nw(a_\mu+\delta_\mu))}\right.$$

$$\left.\times\prod_{k\neq\mu}\frac{1}{\sin\pi(-w(a_k-a_\mu+\delta_k-\delta_\mu))}\right]. \tag{70}$$

In quantum theory, one can set $\hat{p}_\mu=(\hbar/\sqrt{-1})(\partial/\partial x_\mu)$,

$$e^{(\hbar/\sqrt{-1})(\partial/\partial x_\mu)}f(x)=f\left(x+\frac{\hbar}{\sqrt{-1}}\hat{\mu}\right)e^{(\hbar/\sqrt{-1})(\partial/\partial x_\mu)}. \tag{71}$$

Compared with $\Gamma_\mu f(a)=f(a+\hat{\mu})\Gamma_\mu$, let $\Gamma_\mu\rightsquigarrow e(\hbar/\sqrt{-1})(\partial/\partial x_\mu)$, $-\sqrt{-1}w(a_\mu+\delta_\mu)\rightsquigarrow x_\mu$, and $w\rightsquigarrow\hbar$. There is

$$H\rightsquigarrow\sum_{\mu\nu}e^{\hat{p}_\nu}\left[e^{-n\pi x_\nu}\prod_{j\neq\nu}\frac{1}{\sinh\pi(x_j-x_\nu)}\right]e^{-\hat{p}_\mu}\left[e^{n\pi x_\mu}\prod_{k\neq\mu}\frac{1}{\sinh\pi(x_k-x_\mu)}\right]. \tag{72}$$

When $\hat{p}_\lambda\rightsquigarrow p_\lambda$, it becomes a Hamiltonian of a classical integrable system.

We also can begin directly from Eqs. (59) and (60). According to the expanding method used for getting the results (67) and (68), we have

$$F^{(1)}_{\mu\nu}(a,z)\rightsquigarrow[e^{\pi\sqrt{-1}(-z+(n-1)/2+1-nw\bar{a}_\mu)}-e^{-\pi\sqrt{-1}(-z+(n-1)/2+1-nw\bar{a}_\mu)}]$$

$$\times\prod_{j\neq\nu}\frac{1}{\sin\pi(-w(\bar{a}_j-\bar{a}_\nu+\delta_{\mu j}-\delta_{\mu\nu}))}, \tag{73}$$

$$F^{(2)}_{\mu\nu}(a,z)\rightsquigarrow[e^{\pi\sqrt{-1}(z+(n-1)/2+1-nw(\bar{a}_\nu+\delta_{\mu\nu}-1))}-e^{-\pi\sqrt{-1}(z+(n-1)/2+1-nw(\bar{a}_\nu+\delta_{\mu\nu}-1))}]$$

$$\times\prod_{j\neq\mu}\frac{1}{\sin\pi(-w(\bar{a}_j-\bar{a}_\mu))}, \tag{74}$$

while ignoring those sine functions which have no any dynamics variables. Let

$$g_{\mu\nu}=\prod_{j\neq\nu}\frac{1}{\sin\pi(-w(\bar{a}_j+\delta_{\mu j}-\bar{a}_\nu-\delta_{\mu\nu}))}\prod_{k\neq\mu}\frac{1}{\sin\pi(-w(\bar{a}_k-\bar{a}_\mu))}. \tag{75}$$

Then we have

$$G_{\mu\nu}(a,z)=\{e^{\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}+e^{-\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}$$

$$-e^{\pi\sqrt{-1}(-2z-nw(\bar{a}_\nu+1-\bar{a}_\nu-\delta_{\mu\nu}))}-e^{-\pi\sqrt{-1}(-2z-nw(\bar{a}_\mu+1-\bar{a}_\nu-\delta_{\mu\nu}))}\}g_{\mu\nu}. \tag{76}$$

For different z, $H(z)$ commute with each other, so we can obtain three commutable Hamiltonians H^i from H

$$H^i=\sum_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu G^i_{\mu\nu},\quad i=1,2,3, \tag{77}$$

with

$$G^1_{\mu\nu}=\{e^{\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}+e^{-\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}\}g_{\mu\nu}, \tag{78}$$

$$G^2_{\mu\nu}=e^{\pi\sqrt{-1}(-2z-nw(\bar{a}_\mu+1-\bar{a}_\nu-\delta_{\mu\nu}))}g_{\mu\nu}, \tag{79}$$

$$G^3_{\mu\nu}=e^{-\pi\sqrt{-1}(-2z-nw(\bar{a}_\mu+1-\bar{a}_\nu-\delta_{\mu\nu}))}g_{\mu\nu}. \tag{80}$$

In $G^1_{\mu\nu}$, if we have $\delta_\mu\to\delta_\mu+\rho$, where ρ is commutable, we may still get two integrable Hamiltonians

$$H^{11}=\sum_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu G^{11}_{\mu\nu}, \tag{81}$$

$$H^{12}=\sum_{\mu\nu}\Gamma_{-\mu}\Gamma_\nu G^{12}_{\mu\nu}, \tag{82}$$

with

$$G^{11}_{\mu\nu}=e^{\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}g_{\mu\nu}, \tag{83}$$

$$G^{12}_{\mu\nu}=e^{-\pi\sqrt{-1}(n+1-nw(\bar{a}_\mu+\bar{a}_\nu+\delta_{\mu\nu}-1))}g_{\mu\nu}. \tag{84}$$

Here H^{11} and H^{12} all can commute with H^2 and H^3, but they may not commute with each other.

An integrable system requires its Poisson bracket $H=f_1,f_2,\ldots,f_n$ satisfies

$$\{f_i,f_j\}=\sum_{lm}\left(\frac{\partial f_i}{\partial p_l}\frac{\partial f_j}{\partial q_m}-\frac{\partial f_i}{\partial q_l}\frac{\partial f_j}{\partial p_m}\right)=0. \tag{85}$$

When we take variables as changed, i.e., $q_i\to q_i'=\chi q_i+\rho_i$, p_i remain unchanged. In the new variable set $\{q_i',p_i\}$, the above equation (85) still holds, so the system is still integrable. Taking some kind of limit after the variable substitution, we may achieve a new integrable system, i.e., in Eq. (72), let $x_j\to x_j'=\chi x_j$. Then there is

$$H'=\sum_{\mu\nu}e^{p_\nu}\left[e^{-n\pi(1/\chi)x'_\nu}\prod_{j\neq\nu}\frac{1}{\sinh(\pi/\chi)(x_j'-x_\nu')}\right]e^{-p_\mu}\left[e^{n\pi(1/\chi)x'_\mu}\prod_{k\neq\mu}\frac{1}{\sinh(\pi/\chi)(x_k'-x_\mu')}\right]. \tag{86}$$

When $\chi\to\infty$, take the lowest order term. Then we obtain

$$H''=\sum_{\mu\nu}e^{p_\nu-p_\mu}\prod_{j\neq\nu}\frac{1}{x_j-x_\nu}\prod_{k\neq\mu}\frac{1}{x_k-x_\mu}. \tag{87}$$

This may be an integrable Hamiltonian.

Additionally, we can let $x_0\to x_0+\rho_0$ with others unchanged in H, and let $\sinh(\pi(x_0-x_i))$ become $\frac{1}{2}e^{\pi(x_0-x_i)}$ to obtain a new Hamiltonian. We also can change two variables or more to get a new Hamiltonian, i.e., let $\delta_0\gg\delta_1\gg\cdots\gg\delta_{n-1}\gg1\to\infty$, $x_i\to x_i+\delta_i$. There are $\sinh\pi(x_i-x_j)\to\frac{1}{2}e^{\pi(x_i-x_j+\delta_i-\delta_j)}$ $(i<j)$. Then

$$H\to\sum_{\mu\nu}e^{p_\nu-p_\mu}e^{-n\pi(x_\mu-x_\nu)}\prod_{j<\nu}e^{-\pi(x_j-x_\nu)}\prod_{j>\nu}e^{-\pi(x_\nu-x_j)}\prod_{k<\mu}e^{-\pi(x_k-x_\nu)}\prod_{k>\mu}e^{-\pi(x_\nu-x_k)}. \tag{88}$$

Like this, we can get the number of integrable Hamiltonians.

VI. SUMMARY AND DISCUSSIONS

We study the Z_n Belavin model with the open boundary conditions. By the factorized L operators, we constructed the commuting difference operators which are similar to the Ruijsenaars–Macdonald difference operators. We obtained a one-dimensional many-body integrable model from these commuting difference operators. By taking the special limit, this model is similar to the Calogero model.

The one-dimensional integrable models presented in this paper are obtained from the transfer matrix operators with open boundary conditions. Generally, for different solutions of the reflection equations, we can define different transfer matrices. Thus, to some extent, we give a systematic approach to obtain one-dimensional many-body integrable models. Besides the Z_n Belavin model, there are many two-dimensional exactly solvable models in statistical mechanics. It is interesting to find relations between those models with some one-dimensional many-body systems.

ACKNOWLEDGMENTS

This work is supported in part by the Natural Science Foundation of China. H. F. is supported by the NWU Fund.

APPENDIX

$$
\begin{aligned}
[g^{\beta}h^{\alpha}K_0(0)\phi_{a,a+\hat{\mu}}(z)]^{(i)} &= \omega^{\beta i}\delta_{ij}\delta_{j+\alpha,k}\overline{\delta}_{k+l,0}\theta^{(l)}(z-nw\overline{a}_{\mu},n\tau)\\
&= \omega^{\beta i}\theta\begin{bmatrix}\dfrac{1}{2}+\dfrac{i+\alpha}{n}\\ \dfrac{1}{2}\end{bmatrix}(z-nw\overline{a}_{\mu},n\tau)\\
&= \omega^{\beta i}\theta\begin{bmatrix}-\dfrac{1}{2}-\dfrac{i+\alpha}{n}\\ -\dfrac{1}{2}\end{bmatrix}(-z+nw\overline{a}_{\mu},n\tau)\\
&= (-1)\omega^{\beta i}e^{2\pi\sqrt{-1}(1/2-(i-\alpha)/n\beta}\theta\begin{bmatrix}\dfrac{1}{2}-\dfrac{i+\alpha}{n}\\ \dfrac{1}{2}\end{bmatrix}(-z+nw\overline{a}_{\mu},n\tau).
\end{aligned}
$$

Using the formula

$$
\theta^{(i)}(z+\alpha\tau+\beta,n\tau)=e^{-2\pi\sqrt{-1}(\alpha/2)(\alpha\tau/2+z+1/2+\beta)}e^{2\pi\sqrt{-1}(1/2-i(i-2)/n)\beta}\theta\begin{bmatrix}\dfrac{1}{2}-\dfrac{i-\alpha}{n}\\ \dfrac{1}{2}\end{bmatrix}(z,n\tau), \tag{A1}
$$

there is

$$
\begin{aligned}
\cdots &= (-1)^{(\beta-1)}e^{2\pi\sqrt{-1}(\alpha/n)(\alpha\tau/2+z-nw\overline{a}_{\mu}+1/2)}\omega^{i}\theta\begin{bmatrix}\dfrac{1}{2}-\dfrac{i}{n}\\ \dfrac{1}{2}\end{bmatrix}(-z+nw\overline{a}_{\mu}-\alpha\tau-\beta,n\tau)\\
&= (-1)^{(\beta-1)}e^{2\pi\sqrt{-1}(\alpha/n)(\alpha\tau/2+z-nw\overline{a}_{\mu}+1/2)}\omega^{i}\phi^{(i)}_{a,a+\hat{\mu}}(-z+2nw\overline{a}_{\mu}-\alpha\tau-\beta,n\tau)\\
&= (-1)^{(\beta-1)}e^{2\pi\sqrt{-1}(\alpha/n)(\alpha\tau/2+z-nw\overline{a}_{\mu}+1/2)}[g\phi_{a,a+\hat{\mu}}(-z+2nw\overline{a}_{\mu}-\alpha\tau-\beta,n\tau)]^{(i)}.
\end{aligned} \tag{A2}
$$

By moving the g from the right-hand side of the equation to the left-hand side, we obtain the conclusion (40).

[1] M. A. Olshanetski and A. M. Perelomov, Phys. Rep. **94**, 313 (1983).
[2] F. Calogero *et al.*, J. Math. Phys. **10**, 2191 (1969); **12**, 419 (1971).
[3] B. Sutherland, J. Math. Phys. **12**, 246 (1971); Phys. Rev. A **4**, 2019 (1971).
[4] J. Moser, Adv. Math. **16**, 197 (1975).
[5] S. N. M. Ruijsenaars, Commun. Math. Phys. **110**, 191 (1987).
[6] S. N. M. Ruijsenaars and H. Schneider, Ann. Phys. **170**, 370 (1986).
[7] I. M. Krichever, Funct. Anal. Appl. **14**, 282 (1980).
[8] I. G. Macdonald, ''A new class of symmetric functions,'' in *Actes Seminaire Lotharringen*, Publ. Inst. Rech. Math. Adv. Strasbourg (1988), p. 131.
[9] I. G. Macdonald, ''Orthogonal polynomials associated with root systems,'' in *Orthogonal Polynomials*, edited by P. Nevai (Kluwer, Dordrecht, 1990), pp. 311–318.
[10] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic, New York, 1982).
[11] C. N. Yang, Phys. Rev. Lett. **19**, 1312 (1967).
[12] V. V. Bazhanov, Phys. Lett. B **159**, 321 (1985); Commun. Math. Phys. **113**, 471 (1987).
[13] M. Jimbo, Commun. Math. Phys. **102**, 537 (1986).
[14] A. A. Belavin, Nucl. Phys. B: Field Theory Stat. Syst. **180**, [FS2], 189 (1981).
[15] M. P. Richey and C. A. Tracy, J. Stat. Phys. **42**, 311 (1986).
[16] K. Hasegawa, ''Ruijsenaars' commuting difference operators as commuting transfer matrices,'' preprint.
[17] G. Felder and A. Varchenko, ''Elliptic quantum groups and Ruijsenaars model,'' preprint.
[18] J. Avan, O. Babelon, and E. Billey, Commun. Math. Phys. **178**, 50 (1996).
[19] B. Y. Hou, K. J. Shi, Y. S. Wang, and L. Zhao, ''Ruijsenaars-Macdonald Type Difference Operators From Z_n Belavin Model With Open Boundary Condition,'' preprint (1997).
[20] Y. H. Quano and A. Fujii, Mod. Phys. Lett. A **6**, 3635 (1991).
[21] Y. H. Quano, Mod. Phys. Lett. A **8**, 3363 (1993).
[22] B. Y. Hou, K. J. Shi, W. L. Yang, and Z. X. Yang, Phys. Lett. A **178**, 73 (1993).
[23] E. K. Sklyanin, J. Phys. A **21**, 2375 (1988).
[24] H. Fan, B. Y. Hou, K. J. Shi, and Z. X. Yang, Phys. Lett. A **200**, 109 (1996); Nucl. Phys. B **478**, 723 (1996).
[25] M. Jimbo, T. Miwa, and M. Okado, Nucl. Phys. B: Field Theory Stat. Syst. **300**, [FS22], 74 (1988).

Direct proof of integrability of Calogero–Moser model

Cezary Gonera[a)]
Department of Field Theory, University of Łódź, Pomorska 149/153, 90-236 Łódź, Poland

(Received 13 August 1997; accepted for publication 3 April 1998)

The straightforward proof of involutiveness of the set of integrals of motion for the general Calogero–Moser model is given for the classical as well as the quantum case. © *1998 American Institute of Physics.* [S0022-2488(98)01307-3]

I. INTRODUCTION

The Calogero–Moser–Sutherland (CMS) model,[1–3] perhaps a paradigm of solvable finite-dimensional Hamiltonian systems,[4] although discovered more than a quarter century ago, still attracts much attention. It has been observed to have interesting connections and physical applications in various fields like fractional statistics,[5] the quantum Hall effect,[6] spin chains,[7] two-dimensional gravity,[8] and QCD[9] or higher spin theories.[10] Apart from its relevance in these or other branches of theoretical physics, it provides a theoretical laboratory for studying diverse techniques commonly used in mathematical physics; the R-matrix method,[11] the inverse scattering method,[12] the collective field methods,[13] and the W algebras.[14] Paraphrasing the famous Lenin statement about an electron, one can say that the CMS model is inexhaustible.

The classical and quantum integrability of the CMS model has been proven by different methods in numerous papers. The most tedious step is a demonstration that the set of functionally independent conserved quantities is in involution. This may be achieved by arguments based on asymptotics,[15] Lax matrix techniques,[16] a recursion procedure,[17] or by more recent and rather advanced R-matrix procedures.[11]

Our aim in this article is to give direct formal proof of classical and quantum integrability of the CMS model (without a harmonic oscillator term) using only elementary methods, mainly the normal ordering procedure. The proof is valid for most general potential given by the Weierstrass function (which, in special cases, reduces to well-known rational, trigonometric, and hyperbolic potentials). The starting point is the expression for integrals of motion given by Sawada and Kotera[18] and Wojciechowski.[17] The involutivity appears then to be guaranteed by the set of identities obeyed by Weierstrass functions which generalize the well-known addition formula; a similar phenomenon appears in the relativistic case.[19] Given this formula and the notion of normal ordering the proof becomes quite easy.

II. PROOF OF INTEGRABILITY

The Calogero–Moser model[1–3] is defined by the following choice of Hamiltonian:

$$H=\frac{1}{2}\sum_{i=1}^{N} p_i^2+\frac{g}{2}\sum_{i,j=1}^{N}{}' V(x_i-x_j), \tag{1}$$

where Σ' denotes summation over $i\neq j$ and V is the Weierstrass function (which, in particular cases, reduces to rational, trigonometric, or hyperbolic functions).

The classical integrals of motion can be conveniently defined as follows:[18] one starts with the quantity

$$J_N=\exp\left(-\frac{g}{2}\sum_{i,j=1}^{N}{}' V(x_i-x_j)\,\frac{\partial^2}{\partial p_i \partial p_j}\right)\prod_{k=1}^{N} p_k\,. \tag{2}$$

[a)]Electronic mail: cgonera@krysia.uni.odz.pl

 © 1998 American Institute of Physics

Obviously, J_N is a completely symmetric polynomial of the Nth order in p's. In fact, it is simply the determinant of the Lax matrix.[4] Other integrals of motion are defined recursively:

$$J_{N-1}=\left\{\sum_{k=1}^{N} x_k, J_N\right\}, \tag{3}$$

where $\{,\}$ is a standard Poisson bracket. These integrals can be compactly written as follows. Define

$$J(\alpha)=\exp\left(\frac{g}{2}\sum_{i,j=1}^{N}{}' V(x_i-x_j)\,\frac{\partial^2}{\partial p_i \partial p_j}\right)\prod_{k=1}^{N}(p_k+\alpha). \tag{4}$$

Then, Eq. (3) implies

$$J(\alpha)=\sum_{n=0}^{N}\frac{\alpha^n}{n!}J_{N-n}. \tag{5}$$

It follows from Eq. (4) that $J(\alpha)$ is the determinant of the Lax matrix with diagonal elements p_k replaced by $p_k+\alpha$. Therefore, J_n are, up to a constant, the coefficients of the characteristic equation for the Lax matrix. It can be shown that the Poisson brackets $\{J_n, J_{n'}\}$ all vanish. Moreover, the Hamiltonian (1) is expressible in terms of J_n's. Therefore, they are involutive integrals of motion for the classical Calogero–Moser model.

As was noticed by Calogero, Marchioro, and Ragnisco,[20] there are no ordering ambiguities in defining $\hat{J}_n$'s in quantum theory. Therefore we expect the quantum Calogero–Moser model to also be integrable with $\hat{J}_n$'s providing the complete set of commuting integrals of motion. Our aim here is to give a simple formal proof of this statement.

Before entering into details let us stress that although it is true that there is no ordering problem with the definition of quantum counterparts of classical integrals of motion J_n's (actually, this is exactly why these integrals were chosen as a starting point; it would be much more troublesome if one would like to start with traces of powers of the Lax matrix, where ordering problems are significant), it does not imply that there are no ordering problems when one considers the commutators of such integrals. On the contrary, these ordering ambiguities are very relevant because even the product of ''harmless'' operators contains many terms for which the ordering problem is nontrivial. In fact, the cancellations among higher-order terms in $\hbar$ are nontrivial and result from identities (15) below those valid for elliptic functions; the point is that in the reordering of x's and p's in such a way that the classical result for the Poisson bracket can be used, one gets extra terms which cancel only because of Eq. (15). This is just the main point of the proof.

Let us first note that it is sufficient to prove that

$$[\hat{J}_N, \hat{J}(\alpha)]=0, \tag{6}$$

for any α. To see this the following reasoning is sufficient. The involutiveness of the $\hat{J}_n$'s is equivalent to the equation

$$[\hat{J}(\alpha), \hat{J}(\beta)]=0, \tag{7}$$

for all $\alpha,\beta\in\mathbb{R}$. However, the transformation $\hat{x}_i\to\hat{x}_i$, $\hat{p}_i\to\hat{p}_i+\text{const}$ is a unitary transformation, so due to $\hat{J}_N=\hat{J}(0)$, Eq. (7) is implied by

$$[\hat{J}_N, \hat{J}(\beta-\alpha)]=0, \tag{8}$$

i.e., by Eq. (6).

The product of $\hat{x}$ and $\hat{p}_i$ operators will be called normally ordered if all $\hat{p}_i$ factors stand rightmost. With such a definition to any polynomial function $f(x,p)$ defined on classical phase space (which can be explicitly $\hbar$ dependent), there corresponds a unique operator,

$$\hat{f}=:f(\hat{x},\hat{p}):, \tag{9}$$

where ''$:\cdots:$'' means normal ordering defined above. In order to simplify the notation, Eq. (9) will be written henceforth as

$$\hat{f}=:f(x,p):, \tag{10}$$

and it is understood that $:\cdots:$ means the operation of replacing the classical x, p variables by corresponding operators followed by normal ordering. It is obvious that our definition can be extended to include a nonpolynomial function (using, for example, the coordinate representation or formal power expansions).

The following important formulas hold:

$$:f(x,p)::g(x,p):=:f\left(x,p-i\hbar\,\frac{\partial}{\partial y}\right)g(y,p)|_{x=y}:, \tag{11a}$$

$$:f(x,p)::g(x,p):=:g\left(x-i\hbar\,\frac{\partial}{\partial q},p\right)f(x,q)|_{p=q}:. \tag{11b}$$

One can prove the above relations by starting with polynomials in x and p and generalizing accordingly. We shall use the notion of normal ordering to prove Eq. (6).

Due to the fact that there are no ordering ambiguities in quantizing $\hat{J}_n$'s, the classical expressions can be taken as symbols of their quantum counterparts. Using the formulas (11) it is straightforward to write the symbol of the commutator $[\hat{J}(\alpha),\hat{J}_N]$:

$$[\hat{J}(\alpha),\hat{J}_N]=:\exp\left(\frac{-g}{2}\sum_{i,j=1}{}'\left(V(x_i-x_j-i\hbar)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\right)\frac{\partial^2}{\partial p_i\partial p_j}-\frac{g}{2}\sum_{i,j}{}'\,V(x_i-x_j)\frac{\partial^2}{\partial q_i\partial q_j}\right)$$
$$\cdot\left(\prod_k\,(p_k+\alpha)q_k-\prod_k\,p_k(q_k+\alpha)\right):|_{p=q}. \tag{12}$$

Therefore it is sufficient to prove that

$$\exp\left(\frac{-g}{2}\sum_{i,j=1}{}'\left(V(x_i-x_j-i\hbar)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\right)\frac{\partial^2}{\partial p_i\partial p_j}-\frac{g}{2}\sum_{i,j}{}'\,V(x_i-x_j)\frac{\partial^2}{\partial q_i\partial q_j}\right)$$
$$\times\left(\prod_k\,(p_k+\alpha)q_k-\prod_k\,p_k(q_k+\alpha)\right):|_{p=q}=0. \tag{13}$$

For further convenience let us define the objects:

$$X_A(p,q,\alpha)\equiv\prod_{i\notin A}\,(p_i+\alpha)q_i-\prod_{i\notin A}\,p_i(q_i+\alpha)\equiv:Y_A(p,q,\alpha)-Y_A(q,p,\alpha), \tag{14}$$

for any subset $A\subset(1,...,N)$. The following useful properties are immediate consequences of the above definitions:

(i) If $F(q,\partial/\partial q;p,\partial/\partial p)=F(p,\partial/\partial p;q,\partial/\partial q)$, then $F(p,\partial/\partial p;q,\partial/\partial q)X_A(p,q,\alpha)|_{p=q}=0$;

(ii) $\partial^2X_A/\partial p_i^2=\partial^2X_A/\partial q_i^2=0$;

(iii)

$$\prod_{j\in\beta}\,\frac{\partial^2}{\partial p_j\partial q_j}\,X_A=\begin{cases}0, & A\cap B\neq\varnothing,\\ X_{A\cup B}, & A\cap B=\varnothing;\end{cases}$$

(iv)

$$\frac{\partial X_A(p,q,\alpha)}{\partial p_i}=\begin{cases}0, & i\in A,\\ q_i X_{A\cup\{i\}}-\alpha Y_{A\cup\{i\}}(q,p,\alpha), & i\in A,\end{cases}$$

$$\frac{\partial X_A(p,q,\alpha)}{\partial q_i}=\begin{cases}0, & i\in A;\\ p_i X_{A\cup\{i\}}+\alpha Y_{A\cup\{i\}}(p,q,\alpha), & i\notin A;\end{cases}$$

(v)

$$\frac{\partial Y_A}{\partial q_i}=\begin{cases}0, & i\in A\\ (p_i+\alpha)Y_{A\cup\{i\}} & i\notin A,\end{cases}$$

$$\frac{\partial Y_A}{\partial p_i}=\begin{cases}0, & i\in A\\ q_i Y_{A\cup\{i\}}, & i\notin A,\end{cases}$$

Note that

$$X_\varnothing=\prod_k (p_k+\alpha)q_k-\prod_k p_k(q_k+\alpha).$$

Our proof is based on the following identity proven in the Appendix:

$$\sum_{i_1,\ldots,i_N} P(x_{i_1}-x_{i_2})P(x_{i_2}-x_{i_3})\cdots P(x_{i_{N-2}}-x_{i_{N-1}})P'(x_{i_{N-1}}-x_{i_N})=0, \tag{15}$$

where $P(x)$ is the Weierstrass function, $P'(x)$ its derivative, and the sum extends over all permutations of numbers $(1,\ldots,N)$.

In order to prove Eq. (13), we expand $V(x_i-x_j-i\hbar(\partial/\partial q_i-\partial/\partial q_j))$ in powers of $i\hbar(\partial/\partial q_i-\partial/\partial q_j)$. Let us note that due to property (ii) above,

$$\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)^n X_A=0, \tag{16}$$

for any A and $n>2$. Hence, we have only to show that

$$\exp\left(\frac{-g}{2}\sum_{i,j}{}' V(x_i-x_j)\left(\frac{\partial^2}{\partial p_i\partial p_j}+\frac{\partial^2}{\partial q_i\partial q_j}\right)+\frac{i\hbar g}{2}\sum_{i,j}{}' V(x_i-x_j)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\frac{\partial^2}{\partial p_i\partial p_j}\right)$$
$$\cdot\exp\left(\frac{g\hbar^2}{4}\sum_{i,j}{}' V''(x_i-x_j)\frac{\partial^2}{\partial p_i\partial p_j}\frac{\partial^2}{\partial q_i\partial q_j}\right)X_\varnothing|_{p=q}=0. \tag{17}$$

Now, (iii) above implies that

$$\exp\left(\frac{g\hbar^2}{4}\sum_{i,j}{}' V''(x_i-x_j)\frac{\partial^2}{\partial p_i\partial p_j}\frac{\partial^2}{\partial q_i\partial q_j}\right)X_\varnothing$$

can be written as a linear combination (with coefficients depending only on x's) of the quantities X_A, $A=\varnothing$, $\{i,j\}$, etc. Moreover, it is easy to see that

$$\exp\left(\frac{-g}{2}\sum_{i,j}{}' V(x_i-x_j)\frac{\partial^2}{\partial p_i\partial p_j}+\frac{\partial^2}{\partial q_i\partial q_j}\right)\exp\left(\frac{i\hbar g}{2}\sum_{i,j}{}' V'(x_i-x_j)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\frac{\partial^2}{\partial p_i\partial p_j}\right)X_A$$
$$=\exp\left(\frac{-g}{2}\sum_{i,j\notin A}{}' V(x_i-x_j)\left(\frac{\partial^2}{\partial p_i\partial p_j}+\frac{\partial^2}{\partial q_i\partial q_j}\right)\right)\exp\left(\frac{i\hbar g}{2}\sum_{i,j\notin A}{}' V'(x_i-x_j)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\frac{\partial^2}{\partial p_i\partial p_j}\right)X_A,$$

which implies that it is sufficient to show

$$\exp\left(\frac{-g}{2}\sum_{i,j}{}'\ V(x_i-x_j)\left(\frac{\partial^2}{\partial p_i\partial p_j}+\frac{\partial^2}{\partial q_i\partial q_j}\right)\right)$$
$$\times\exp\left(\frac{i\hbar g}{2}\sum_{i,j}{}'\ V'(x_i-x_j)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\frac{\partial^2}{\partial p_i\partial p_j}\right)X_A|_{p=q}=0, \tag{18}$$

for any number of degrees of freedom. Let us denote

$$D\equiv-\frac{g}{2}\sum_{i,j}{}'\ V(x_i-x_j)\left(\frac{\partial^2}{\partial p_i\partial p_j}+\frac{\partial^2}{\partial q_i\partial q_j}\right),$$
$$C\equiv\frac{i\hbar g}{2}\sum_{i,j}{}'\ V(x_i-x_j)\left(\frac{\partial}{\partial q_i}-\frac{\partial}{\partial q_j}\right)\frac{\partial^2}{\partial p_i\partial p_j}. \tag{19}$$

Equation (18) reads now as

$$e^De^CX_\varnothing|_{p=q}=0, \tag{20}$$

and the expansion in powers of C is equivalent to the one in powers of $\hbar$. Obviously, this expansion has a finite number of terms.

In order to prove Eq. (20) we use the induction with respect to the powers of C. In zeroth order in $\hbar$ Eq. (20) follows immediately from the property (i) and expresses the trivial fact that classical dynamical variables commute.

Now, by the induction hypothesis we assume that

$$e^DC^nX_\varnothing|_{p=q}=0, \tag{21a}$$

for any number of particles, i.e. (as explained above) that

$$e^DC^nX_A|_{p=q}=0, \tag{21b}$$

for any subset A. Due to the properties (iii)–(v) we have

$$e^DC^{n+1}X_\varnothing=-\frac{g}{2}\,e^DC^n\sum_{i,j}{}'\ V'(x_i-x_j)(q_j-q_i)X_{\{i,j\}}$$
$$=-\frac{g}{2}\,e^D\sum_{i,j}{}'\ V'(x_i-x_j)(q_j-q_i)C^nX_{\{i,j\}}. \tag{22}$$

The last equality follows from the observation that the commutator $[C,q_j-q_i]$ acting on $X_{\{i,j\}}$ gives zero, by property (iv). Therefore we have to show that

$$e^D\sum_{i,j}{}'\ V'(x_i-x_j)(q_j-q_i)C^nX_{\{i,j\}}|_{p=q}=0. \tag{23}$$

Writing

$$e^D(q_j-q_i)=((q_j-q_i)+[D,q_j-q_i])e^D, \tag{24}$$

one obtains

$$e^D\sum_{ij}{}'\ V'(x_i-x_j)(q_j-q_i)C^nX_{\{ij\}}|_{p=q}=\sum_{ij}{}'\ V'(x_i-x_j)(q_j-q_i)e^DC^nX_{\{ij\}}|_{p=q}$$
$$-2g\sum_{i_1,i_2,i_3}{}'\ V'(x_{i_1}-x_{i_2})V(x_{i_2}-x_{i_3})e^DC^n\left.\frac{\partial X_{\{ij\}}}{\partial q_{i_3}}\right|_{p=q}. \tag{25}$$

The first term on the right-hand side of Eq. (25) vanishes by the induction hypothesis. As for the second one, we shall show that

$$\sum_{i_1,i_2,i_3}{}' V'(x_{i_1}-x_{i_2})V(x_{i_2}-x_{i_3})e^D C^n \frac{\partial X_{\{ij\}}}{\partial q_{i_3}}$$

$$\approx (-g)^{m-3}\sum_{i_1,\ldots,i_m}{}' V'(x_{i_1}-x_{i_2})V(x_{i_2}-x_{i_3})\cdots V(x_{m-1}-x_{i_m})e^D\cdot C^n \frac{\partial X_{\{i_1\ldots i_{m-1}\}}}{\partial \eta_{i_m}}, \tag{26}$$

where $\eta=p$ or q and ''$\approx$'' stands for equality up to the terms which vanish when $p=q$.

Let us assume that $\eta=p$ (for $\eta=q$ the proof goes along the same lines). It follows from the property (iv) that

$$\frac{\partial X_{\{i_1\ldots i_{m-1}\}}}{\partial p_{i_m}} = q_{i_m} X_{\{i_1\ldots i_m\}}(p,q,\alpha) - \alpha Y_{\{i_1\ldots i_m\}}(q,p,\alpha). \tag{27}$$

Inserting (27) onto the right-hand side of Eq. (26) we get two terms. The one including the function $Y_{\{i_1\cdots i_m\}}$ vanishes due to the identity (15) because $C^n Y_{\{i_1\cdots i_m\}}$ does not depend on the permutation of indices $i_1,\ldots,i_m$. As for the remaining term, we write it as

$$e^D q_{i_m} C^n X_{\{i_1\cdots i_m\}} = (q_{i_m}+[D,q_{i_m}])e^D C^n X_{\{i_1\ldots i_m\}}$$

$$= q_{i_m} e^D C^n X_{\{i_1\cdots i_m\}} - g\sum_{i_{m+1}}{}' V(x_{i_m}-x_{i_{m+1}})e^D C^n \frac{\partial X\{i_1\cdots i_m\}}{\partial q_{i_{m+1}}}. \tag{28}$$

By the induction hypothesis (with respect to n) the first term on the right-hand side gives no contribution for $p=q$ while the second one yields Eq. (26) for $m+1$; induction with respect to m (with n fixed) therefore yields Eq. (26).

Putting $m=N$ in Eq. (26), using (iv), and taking into account our basic identity (15), we get

$$\sum_{i_1,i_2,i_3}{}' V'(x_{i_1}-x_{i_2})v(x_{i_2}-x_{i_3})e^D C^n \left.\frac{\partial X\{i,j\}}{\partial q_{i_3}}\right|_{p=q} = 0,$$

which concludes the proof of Eq. (21) for $n\to n+1$.

So the proof of integrability is completed.

APPENDIX: PROOF OF THE IDENTITY (15)

Below we sketch the proof of the basic identity (15).

Let, for any $n\geqslant 3$,

$$K_n(x_1,\ldots,x_n) \equiv \sum_{(i_1,\ldots,i_n)\subset(1,\ldots,n)} P(x_{i_1}-x_{i_2})\cdots P(x_{i_{n-2}}-x_{i_{n-1}})P'(x_{i_{n-1}}-x_{i_n}), \tag{A1}$$

where $P(x)$ is the Weierstrass function, $P'(x)$ is its derivative, and the sum on the right-hand side extends over all permutations of numbers 1,...,n.

We claim that

$$K_n(x_1,\ldots,x_n)=0, \tag{A2}$$

for any $n\geqslant 3$.

The proof is by induction. For $n=3$, (A2) reduces to well-known addition theorem for the Weierstrass function.[21]

Let us assume that (A2) holds for some n. Define

$$\Delta \equiv \{n\omega_1 + m\omega_2 \in \mathbb{C} | n,m \in \mathbb{Z}\},$$

where ω_1, ω_2 are periods of P.

Let us single out one of the x variables, say x_1, and consider K_{n+1} as a function of this variable. Assume that $x_i - x_j \notin \Delta$ for any $2 \leqslant i \neq j \leqslant n+1$. Then K_{n+1}, viewed as a function of x_1, is double periodic and analytic except possible poles of, at most, third order coming from single factors on the right-hand side of Eq. (A1) and located at the points $x_1 - x_k \in \Delta$, $k = 2,...,n+1$. We claim that K_{n+1} is analytic on the whole x_1-complex plane. Obviously, it is sufficient to consider the poles appearing due to $x_1 - x_2 \in \Delta$; moreover, due to double periodicity in x_1 we can restrict ourselves to the case $x_1 \to x_2$. Let us call $z \equiv x_1 - x_2$ and select the singular terms of K_{n+1}. A careful but easy analysis gives

$$K_{n+1}(x_1,\ldots,x_{n+1}) = \frac{2}{z^2} K_n(x_2,\ldots,x_n) - \frac{1}{z} \frac{\partial K_n(x_2,\ldots,x_{n+1})}{\partial x_2} + O(1). \tag{A3}$$

By the induction assumption the singular term vanishes. Therefore $K_{n+1}(x_1,x_2,\ldots,x_{n+1})$, considered as a function of x_1, is doubly periodic and analytic provided $x_i - x_j \notin \Delta$, $2 \leqslant i \neq j \leqslant n+1$. Consequently, it is a constant. This reasoning can be, of course, repeated for any variable x_i. We conclude that K_{n+1} is constant on the open connected set,

$$\Lambda = \{(x_1,\ldots,x_{n+1}) \in \mathbb{C}^{n+1} | x_i - x_j \notin \Delta, 1 \leqslant 1, j \leqslant n+1\} \tag{A4}$$

But K_{n+1} is meromorphic on $\mathbb{C}^{n+1}$ and $\overline{\Lambda} = \mathbb{C}^{n+1}$; hence K_{n+1} is constant. Now, K_{n+1} is odd under the inversion $x_i \to -x_i$, $i = 1,...,n+1$, so it vanishes.

ACKNOWLEDGMENTS

Numerous discussions with S. Giller, P. Kosiński, P. Maślanka, and B. Nowak are acknowledged. An unknown referee is kindly acknowledged for useful remarks which allowed us to make the paper more clear. This work was supported by KBN Grant No. 2 P03B 076 10.

[1] F. Calogero, J. Math. Phys. **10**, 2191, 2197 (1969); **12**, 429 (1971); **15**, 1425 (1974).
[2] B. Sutherland, J. Math. Phys. **12**, 246 (1971); Phys. Rev. A **4**, 2019 (1971); *ibid.* **5**, 1372 (1972).
[3] J. Moser, Adv. Math. **16**, 1 (1975).
[4] M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **94**, 313 (1983); **71**, 313 (1981).
[5] A. P. Polychronakos, Nucl. Phys. B **324**, 597 (1989); S. B. Jsakov and J. M. Leinaas, *ibid.* **463**, 194 (1996); J. M. Leinacaas, ''Generalized statistics and the algebra of observables,'' hep-th/9611167.
[6] N. Kawakami, Phys. Rev. Lett. **71**, 275 (1993); O. Azuma and S. Iso, Phys. Lett. B **331**, 107 (1994).
[7] F. D. M. Haldane, Phys. Rev. Lett. **60**, 635 (1988); B. S. Shastry, *ibid.* **60**, 639 (1988).
[8] I. Andric, A. Jevicki, and M. Levine, Nucl. Phys. B **215**, 307 (1983); A. Jevicki, Nucl. Phys. B **376**, 75 (1992).
[9] J. A. Minahan and A. P. Polychronakos, Phys. Lett. B **312**, 155 (1993); Nucl. Phys. B **336**, 288 (1994).
[10] M. S. Fradkin and M. A. Vasiliev, Ann. Phys. **177**, 63 (1987); M. A. Vasiliev, Int. J. Mod. Phys. A **6**, 115 (1991).
[11] O. Babelon and C-M. Viallet, Phys. Lett. B **237**, 411 (1990); J. Avan and M. Talon, *ibid.* **303**, 33 (1993); J. Avan, O. Babelon, and M. Talon, Alg. Anal. **6**, 67 (1994); E. K. Sklyanin, *ibid.* **6**, 227 (1994); H. W. Barden and T. Suzuki, Lett. Math. Phys. **30**, 147 (1994).
[12] H. Ujino, K. Hikami, and M. Wadati, J. Phys. Soc. Jpn. **61**, 3425 (1992); H. Ujino, K. Hikami, and M. Wadati, *ibid.* **62**, 3035 (1993); H. Ujino and M. Wadati, *ibid.* **64**, 4121 (1995); M. Wadati, K. Hikami, and H. Ujino, Chaos Solitons Fractals **3**, 627 (1993).
[13] J. Avan and A. Jevicki, Phys. Lett. B **272**, 17 (1991); Commun. Math. Phys. **150**, 149 (1992); R. Caracciolo, A. Lerda, and G. R. Zemba, Phys. Lett. B **352**, 304 (1995); D. Sen and R. K. Bhaduri, ''Applications of the collective field theory for the Calogero-Sutherland model,'' cond-mat/9702152.
[14] H. Ujino and M. Wadati, J. Phys. Soc. Jpn. **63**, 3585 (1994); H. Ujino and M. Wadati, *ibid.* **62**, 4203 (1993); Phys. Rev. Lett. **73**, 1191 (1994); V. Narayana and M. Sivakumar, Mod. Phys. Lett. A **11**, 763 (1996).
[15] J. Moser, Adv. Math. **16**, 1 (1975).
[16] M. A. Olshanetsky and A. M. Perelomov, Invent. Math. **37**, 93 (1976).
[17] S. Wojciechowski, Lett. Nuovo Cimento **18**, 103 (1977).
[18] K. Sawada and T. Kotera, J. Phys. Soc. Jpn. **39**, 1614 (1975).
[19] S. N. M. Ruijsenaars, Commun. Math. Phys. **110**, 191 (1987).
[20] F. Calogero, C. Marchioro, and O. Ragnisco, Lett. Nuovo Cimento **13**, 383 (1975).
[21] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, Cambridge, 1973).

Some new results on the Blaszak–Marciniak lattice: Bäcklund transformation and nonlinear superposition formula

Xing-Biao Hu
CCAST (World Laboratory), P. O. Box 8730, Beijing 100080, People's Republic of China and State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific Engineering Computing, Academia Sinica, P. O. Box 2719, Beijing 100080, People's Republic of China

Zuo-Nong Zhu
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China

(Received 21 January 1998; accepted for publication 8 April 1998)

The so-called Blaszak–Marciniak lattice is considered. By the dependent variable transformation, the Blaszak–Marciniak lattice is transformed into a quart-linear form. By introducing an auxiliary variable, we further transform it into the bilinear form. A corresponding Bäcklund transformation is obtained. Furthermore, a nonlinear superposition formula is proved rigorously. As an application of the obtained results, soliton solutions are derived. © *1998 American Institute of Physics.* [S0022-2488(98)01409-1]

I. INTRODUCTION

The so-called Blaszak–Marciniak (BM) lattice is[1,2]

$$\begin{aligned}\frac{d}{dt}a(n)&=c(n+1)-c(n-1),\\ \frac{d}{dt}b(n)&=a(n-1)c(n-1)-a(n)c(n),\\ \frac{d}{dt}c(n)&=c(n)(b(n)-b(n+1)),\end{aligned}\tag{1}$$

which was first derived as an application of *r*-matrix formalism to the algebra of shifts operators. Some research on the BM lattice has been conducted. For examples, Blaszak and Marciniak established the bi-Hamiltonian structure of this system.[1] An integrable symplectic map connected with the BM lattice and its hierarchy was obtained by Wu and Geng.[2] The purpose of this paper is to derive a Bäcklund transformation for the BM lattice and establish a nonlinear superposition formula for it. As a result, multisoliton solutions are derived step by step.

We now consider the system (1). By the dependent variable transformation

$$a(n)=\frac{D_t^2 f(n+1)\cdot f(n+1)}{2f(n)f(n+2)},\quad b(n)=\left(\ln\frac{f(n)}{f(n+1)}\right)_t,\quad c(n)=\frac{f(n)f(n+2)}{f^2(n+1)},$$

(1) can be transformed into the following form:

$$\frac{1}{2}\left[\frac{D_t^2 f(n+1)\cdot f(n+1)}{f(n)f(n+2)}\right]_t=\frac{f(n+1)f(n+3)}{f^2(n+2)}-\frac{f(n+1)f(n-1)}{f^2(n)},\tag{2}$$

or equivalently

$$\tfrac{1}{2}D_t(D_t^2 f(n)\cdot f(n))\cdot(e^{D_n}f(n)\cdot f(n))=f^2(n-1)f(n)f(n+2)-f(n)f(n-2)f^2(n+1),\tag{3}$$

 © 1998 American Institute of Physics

where the bilinear operators are defined as follows:[3–5]

$$D_x^m D_t^n a\cdot b\equiv\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial x'}\right)^m\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial t'}\right)^n a(x,t)b(x',t')|_{x'=x,t'=t}.$$

$$\exp(\delta D_n)a(n)\cdot b(n)\equiv\exp\left[\delta\left(\frac{\partial}{\partial n}-\frac{\partial}{\partial n'}\right)\right]a(n)b(n')|_{n'=n}=a(n+\delta)b(n-\delta).$$

By introducing an auxiliary variable z and using (A1), we can decouple (3) into the following bilinear form:

$$(D_t^2-2D_z e^{D_n})f(n)\cdot f(n)=0, \tag{4}$$

$$(D_z D_t-4\sinh^2(\tfrac{1}{2}D_n))f(n)\cdot f(n)=0. \tag{5}$$

The paper is organized as follows. In Sec. II we give a Bäcklund transformation for Eqs. (4) and (5). Then in Sec. III, we give a brief proof of a nonlinear superposition formula. Some particular solutions of Eqs. (4) and (5) are then found through this formula. Finally, Sec. IV summarizes the obtained results. Appendix A lists some bilinear operator identities made use of in this paper. In Appendix B, a detailed deduction of relation (17) is given.

II. A BÄCKLUND TRANSFORMATION FOR THE BM LATTICE

In this section, we derive a Bäcklund transformation for Eqs. (4) and (5). The result obtained is

Proposition 1: A Bäcklund transformation for (4) and (5) is

$$(D_t+\lambda^{-1}e^{-D_n}+\mu)f(n)\cdot g(n)=0, \tag{6}$$

$$(D_z e^{-(1/2)D_n}-\lambda e^{(1/2)D_n}+\gamma e^{-(1/2)D_n})f(n)\cdot g(n)=0, \tag{7}$$

$$[\lambda^{-1}D_t e^{-(1/2)D_n}-D_z e^{(1/2)D_n}-\gamma e^{(1/2)D_n}+\omega e^{-(1/2)D_n}]f(n)\cdot g(n)=0, \tag{8}$$

where λ, μ, γ, and ω are arbitrary constants.

Proof: Let $f(n)$ be a solution of Eqs. (4) and (5). If we can show that Eqs. (6)–(8) guarantee that the following two relations:

$$P_1\equiv(D_t^2-2D_z e^{D_n})g(n)\cdot g(n)=0,$$

$$P_2\equiv(D_z D_t-4\sinh^2(\tfrac{1}{2}D_n))g(n)\cdot g(n)=0,$$

hold, then Eqs. (6)–(8) form a Bäcklund transformation.

Note that (5) is just the bilinear form for the two-dimensional Toda lattice. From Ref. 6 we know that (6) and (7) guarantee $P_2=0$. Therefore, we only need to prove $P_1=0$. By use of (A2)–(A6) and (6)–(8), we have

$$\begin{aligned}-f^2(n)P_1=&\,2D_t(D_t f(n)\cdot g(n))\cdot f(n)g(n)-4\sinh(\tfrac{1}{2}D_n)\\&\times\{[D_z e^{(1/2)D_n}f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n}f(n)\cdot g(n)]\\&-[e^{(1/2)D_n}f(n)\cdot g(n)]\cdot[D_z e^{-(1/2)D_n}f(n)\cdot g(n)]\}\\=&-2\lambda^{-1}D_t[e^{-D_n}f(n)\cdot g(n)]\cdot f(n)g(n)-4\sinh(\tfrac{1}{2}D_n)\\&\times[D_z e^{(1/2)D_n}f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n}f(n)\cdot g(n)]-4\gamma\sinh(\tfrac{1}{2}D_n)\\&\times[e^{(1/2)D_n}f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n}f(n)\cdot g(n)]\end{aligned}$$

$$
\begin{aligned}
&= 4\lambda^{-1}\sinh(\tfrac{1}{2}D_n)[D_t e^{-(1/2)D_n} f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n} f(n)\cdot g(n)] - 4\sinh(\tfrac{1}{2}D_n)\\
&\times[D_z e^{(1/2)D_n} f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n} f(n)\cdot g(n)] - 4\gamma\sinh(\tfrac{1}{2}D_n)\\
&\times[e^{(1/2)D_n} f(n)\cdot g(n)]\cdot[e^{-(1/2)D_n} f(n)\cdot g(n)] = 0.
\end{aligned}
$$

Thus we have completed the proof of proposition 1.

By use of (6)–(8), we can easily obtain the following solutions from the trivial solution $f(n)=1$:

$$g(n) = 1 + \exp(\eta)$$

with

$$\lambda = \left(\frac{e^{2p}}{1+e^{p}}\right)^{(1/3)}, \quad \mu = -\lambda^{-1}, \quad \gamma = \lambda, \omega = \lambda,$$

where $\eta = pn + qz + rt + \eta^0$, $q = \lambda(1-e^{-p})$, $r = \lambda^{-1}(e^{p}-1)$.

III. A NONLINEAR SUPERPOSITION FORMULA

In the following, we shall simply denote, without confusion, $f(n,t)=f(n)$ or f. The result reached is

Proposition 2: Let f_0 be a solution of Eqs. (4) and (5) and suppose that $f_i (i=1,2)$ are solutions of (4) and (5) which are related to f_0 under the BT Eqs. (6)–(8) with parameters $(\lambda_i,\mu_i,\gamma_i,\omega_i)$, i.e., $f_0 \overset{(\lambda_i,\mu_i,\gamma_i,\omega_i)}{\longrightarrow} f_i (i=1,2)$, $\lambda_1\lambda_2 \neq 0$, $f_j \neq 0$ $(j=0,1,2)$. Then f_{12} defined by

$$\exp(-\tfrac{1}{2}D_n) f_0\cdot f_{12} = c[\lambda_1 \exp(-\tfrac{1}{2}D_n) - \lambda_2 \exp(\tfrac{1}{2}D_n)] f_1\cdot f_2 \quad (c \text{ is a nonzero constant}), \tag{9}$$

is a new solution which is related to f_1 and f_2 under the BT (6)–(8) with parameters $(\lambda_2,\mu_2,\gamma_2,\omega_2)$, $(\lambda_1,\mu_1,\gamma_1,\omega_1)$, respectively.

Proof: Same as the deduction in Ref. 6, we can show that

$$(D_t + \lambda_2^{-1} e^{-D_n} + \mu_2) f_1\cdot f_{12} = 0, \tag{10}$$

$$(D_t + \lambda_1^{-1} e^{-D_n} + \mu_1) f_2\cdot f_{12} = 0, \tag{11}$$

$$(D_z e^{-(1/2)D_n} - \lambda_2 e^{(1/2)D_n} + \gamma_2 e^{-(1/2)D_n}) f_1\cdot f_{12} = 0, \tag{12}$$

$$(D_z e^{-(1/2)D_n} - \lambda_1 e^{(1/2)D_n} + \gamma_1 e^{-(1/2)D_n}) f_2\cdot f_{12} = 0, \tag{13}$$

and

$$-D_t f_1\cdot f_2 + (\mu_1-\mu_2) f_1 f_2 - \frac{1}{c\lambda_1\lambda_2} e^{-D_n} f_0\cdot f_{12} = 0. \tag{14}$$

Therefore, in order to prove proposition 2, it suffices to show that

$$[\lambda_2^{-1} D_t e^{-(1/2)D_n} - D_z e^{(1/2)D_n} - \gamma_2 e^{(1/2)D_n} + \omega_2 e^{-(1/2)D_n}] f_1\cdot f_{12} = 0, \tag{15}$$

$$[\lambda_1^{-1} D_t e^{-(1/2)D_n} - D_z e^{(1/2)D_n} - \gamma_1 e^{(1/2)D_n} + \omega_1 e^{-(1/2)D_n}] f_2\cdot f_{12} = 0. \tag{16}$$

From

$$\frac{1}{\lambda_1}[(D_z e^{-(1/2)D_n}-\lambda_1 e^{(1/2)D_n}+\gamma_1 e^{-(1/2)D_n})f_0(n)\cdot f_1(n)]f_2\left(n-\tfrac{1}{2}\right),$$

$$-\frac{1}{\lambda_2}[(D_z e^{-(1/2)D_n}-\lambda_2 e^{(1/2)D_n}+\gamma_2 e^{-(1/2)D_n})f_0(n)\cdot f_2(n)]f_1(n-\tfrac{1}{2})=0,$$

we can deduce that (for detail, see Appendix B)

$$[\lambda_2 D_z e^{(1/2)D_n}+\lambda_1 D_z e^{-(1/2)D_n}-2\lambda_2\gamma_1 e^{(1/2)D_n}+2\lambda_1\gamma_2 e^{-(1/2)D_n}]f_1(n)\cdot f_2(n)$$
$$+\frac{1}{c}D_z e^{-(1/2)D_n}f_0(n)\cdot f_{12}(n)=0. \tag{17}$$

Since f_1 and f_2 are two solutions of Eq. (4), we have, by use of (A2)–(A9), (9), (10), (12), (17), and $f_0 \underset{(\lambda_2,\mu_2,\gamma_2,\omega_2)}{\rightarrow} f_2$, that

$$0=[(D_t^2-2D_z e^{D_n})f_1\cdot f_1]f_2^2-f_1^2[(D_t^2-2D_z e^{D_n})f_2\cdot f_2],$$

$$=2D_t(D_t f_1\cdot f_2)\cdot f_1 f_2-4\sinh(\tfrac{1}{2}D_n)\{[D_z e^{(1/2)D_n}f_1\cdot f_2]\cdot[e^{-(1/2)D_n}f_1\cdot f_2]$$
$$-[e^{(1/2)D_n}f_1\cdot f_2]\cdot[D_z e^{-(1/2)D_n}f_1\cdot f_2]\}$$

$$=-2D_t\left[\frac{1}{c\lambda_1\lambda_2}e^{-D_n}f_0\cdot f_{12}\right]\cdot f_1 f_2-\frac{4}{\lambda_1}\sinh\left(\frac{1}{2}D_n\right)\{[(\lambda_1 D_z e^{-(1/2)D_n}$$
$$+\lambda_2 D_z e^{(1/2)D_n})f_1\cdot f_2]\cdot[e^{(1/2)D_n}f_1\cdot f_2]-[(\lambda_1 e^{-(1/2)D_n}-\lambda_2 e^{(1/2)D_n})f_1\cdot f_2]\cdot[D_z e^{(1/2)D_n}f_1\cdot f_2]\}$$

$$=-\frac{2}{c\lambda_1\lambda_2}e^{-(1/2)D_n}\{[D_t e^{-(1/2)D_n}f_0\cdot f_2]\cdot[e^{-(1/2)D_n}f_1\cdot f_{12}]$$
$$-[e^{-(1/2)D_n}f_0\cdot f_2]\cdot[D_t e^{-(1/2)D_n}f_1\cdot f_{12}]\}-\frac{4}{\lambda_1}\sinh\left(\frac{1}{2}D_n\right)\left\{\left[-\frac{1}{c}D_z e^{-(1/2)D_n}f_0\cdot f_{12}\right.\right.$$
$$\left.\left.-2\gamma_2\lambda_1 e^{-(1/2)D_n}f_1\cdot f_2\right]\cdot[e^{(1/2)D_n}f_1\cdot f_2]-\frac{1}{c}[e^{-(1/2)D_n}f_0\cdot f_{12}]\cdot[D_z e^{(1/2)D_n}f_1\cdot f_2]\right\}$$

$$=-\frac{2}{c\lambda_1\lambda_2}e^{-(1/2)D_n}\{[D_t e^{-(1/2)D_n}f_0\cdot f_2]\cdot[e^{-(1/2)D_n}f_1\cdot f_{12}]$$
$$-[e^{-(1/2)D_n}f_0\cdot f_2]\cdot[D_t e^{-(1/2)D_n}f_1\cdot f_{12}]\}+\frac{2}{c\lambda_1}e^{-(1/2)D_n}\{[D_z e^{(1/2)D_n}f_0\cdot f_2]\cdot[e^{-(1/2)D_n}f_1$$
$$\cdot f_{12}]-[e^{-(1/2)D_n}f_0\cdot f_2]\cdot[D_z e^{(1/2)D_n}f_1\cdot f_{12}]+[e^{(1/2)D_n}f_0\cdot f_2]\cdot[D_z e^{-(1/2)D_n}f_1\cdot f_{12}]$$
$$-[D_z e^{-(1/2)D_n}f_0\cdot f_2]\cdot[e^{(1/2)D_n}f_1\cdot f_{12}]\}+\frac{8\gamma_2}{\lambda_1}\sinh\left(\frac{1}{2}D_n\right)[(\lambda_1 e^{-(1/2)D_n}$$
$$-\lambda_2 e^{(1/2)D_n})f_1\cdot f_2]\cdot[e^{(1/2)D_n}f_1\cdot f_2]$$

$$=-\frac{2}{c\lambda_1\lambda_2}e^{-(1/2)D_n}\{[D_t e^{-(1/2)D_n}f_0\cdot f_2]\cdot[e^{-(1/2)D_n}f_1\cdot f_{12}]$$
$$-[e^{-(1/2)D_n}f_0\cdot f_2]\cdot[D_t e^{-(1/2)D_n}f_1\cdot f_{12}]\}+\frac{2}{c\lambda_1}e^{-(1/2)D_n}\{[D_z e^{(1/2)D_n}f_0\cdot f_2]\cdot[e^{-(1/2)D_n}f_1\cdot 12]$$
$$-[e^{-(1/2)D_n}f_0\cdot f_2]\cdot[D_z e^{1/2D_n}f_1\cdot f_{12}]-[e^{(1/2)D_n}f_0\cdot f_2]\cdot[\gamma_2 e^{-(1/2)D_n}f_1\cdot f_{12}]$$

$$+[\gamma_2 e^{-(1/2)D_n} f_0 \cdot f_2]\cdot[e^{(1/2)D_n} f_1 \cdot f_{12}]\} + \frac{8\gamma_2}{\lambda_1 c}\sinh\left(\frac{1}{2}D_n\right)[e^{-(1/2)D_n} f_0 \cdot f_{12}]\cdot[e^{(1/2)D_n} f_1 \cdot f_2]$$

$$= -\frac{2}{c\lambda_1} e^{-(1/2)D_n}\{[(\lambda_2^{-1} D_t e^{-(1/2)D_n} - D_z e^{(1/2)D_n} + \gamma_2 e^{(1/2)D_n}) f_0 \cdot f_2]\cdot[e^{-(1/2)D_n} f_1 \cdot f_{12}]$$

$$-[e^{-(1/2)D_n} f_0 \cdot f_2]\cdot[(\lambda_2^{-1} D_t e^{-(1/2)D_n} - D_z e^{(1/2)D_n} + \gamma_2 e^{(1/2)D_n}) f_1 \cdot f_{12}]\}$$

$$+\frac{4\gamma_2}{\lambda_1 c} e^{-(1/2)D_n}\{[e^{(1/2)D_n} f_0 \cdot f_2]\cdot[e^{-(1/2)D_n} f_1 \cdot f_{12}] - [e^{-(1/2)D_n} f_0 \cdot f_2]\cdot[e^{(1/2)D_n} f_1 \cdot f_{12}]\}$$

$$= \frac{2}{c\lambda_1} e^{-(1/2)D_n}[e^{-(1/2)D_n} f_0 \cdot f_2]\cdot[(\lambda_2^{-1} D_t e^{-(1/2)D_n} - D_z e^{(1/2)D_n} - \gamma_2 e^{(1/2)D_n}$$

$$+\omega_2 e^{-(1/2)D_n}) f_1 \cdot f_{12}]\},$$

which implies that (15) holds. Similarly we can prove (16) also holds. Therefore we have completed the proof of the proposition 2.

As an application of the result, we can construct soliton solutions of the BM lattice. Choose for example $f_0 = 1$, $c = 1/(\lambda_1 - \lambda_2)$. It is easily verified that

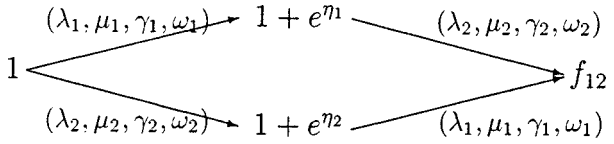

where

$$f_{12} = 1 + \frac{\lambda_1 e^{-p_1/2} - \lambda_2 e^{p_1/2}}{\lambda_1 - \lambda_2} e^{\eta_1} + \frac{\lambda_1 e^{p_2/2} - \lambda_2 e^{-p_2/2}}{\lambda_1 - \lambda_2} e^{\eta_2} + \frac{\lambda_1 e^{(p_2 - p_1)/2} - \lambda_2 e^{(p_1 - p_2)/2}}{\lambda_1 - \lambda_2} e^{\eta_1 + \eta_2}$$

with

$$\eta_i = p_i n + q_i z + r_i t + \eta_i^0, \quad q_i = \lambda_i(1 - e^{-p_i}), \quad r_i = \lambda_i^{-1}(e^{p_i} - 1)$$

and

$$\lambda_i = \left(\frac{e^{2p_i}}{1 + e^{p_i}}\right)^{1/3}, \quad \mu_i = -\lambda_i^{-1}, \quad \gamma_i = \lambda_i, \quad \omega_i = \lambda_i.$$

In general, along this line, we can obtain multisoliton solutions for the BM lattice (4) and (5) step by step.

IV. SUMMARY

The so-called Blaszak–Marciniak lattice has been considered. By the dependent variable transformation, the BM lattice is transformed into a quart-linear form. By introducing an auxiliary variable, we further transform it into the bilinear form. We have given a Bäcklund transformation for the BM lattice as well as a nonlinear superposition formula for it. As an application of the obtained results, soliton solutions are derived.

ACKNOWLEDGMENTS

The authors would like to express their sincere thanks to Professor Hong-Ci Huang for valuable discussions. X.B.H. gratefully acknowledges support by the Croucher Foundation, the Chinese Academy of Sciences and the National Natural Science Foundation of China. Z.N.Z. is supported by Faculty Research Grant of HKBU.

APPENDIX A: HIROTA BILINEAR OPERATOR IDENTITIES

The following bilinear operator identities hold for arbitrary functions a, b, c, and d.

$$2D_t[D_z e^{D_n} a(n)\cdot(n)]\cdot[e^{D_n} a(n)\cdot a(n)] = a^2(n-1)D_t D_z a(n+1)\cdot a(n+1) - a^2(n+1)D_t D_z a(n-1)\cdot a(n-1), \tag{A1}$$

$$(D_t^2 a\cdot a)b^2 - a^2 D_t^2 b\cdot b = 2D_t(D_t a\cdot b)\cdot ab, \tag{A2}$$

$$[D_z e^{D_n} a\cdot a]b^2 - a^2[D_z e^{D_n} b\cdot b] = 2\sinh(\tfrac{1}{2}D_n)\{[D_z e^{(1/2)D_n} a\cdot b]\cdot[e^{-(1/2)D_n} a\cdot b] - [e^{(1/2)D_n} a\cdot b]\cdot[D_z e^{-(1/2)D_n} a\cdot b]\}, \tag{A3}$$

$$D_t(e^{-D_n} a\cdot b)\cdot ab = -2\sinh(\tfrac{1}{2}D_n)[D_t e^{-(1/2)D_n} a\cdot b]\cdot[e^{-(1/2)D_n} a\cdot b], \tag{A4}$$

$$D_t a\cdot a = 0, \tag{A5}$$

$$\sinh(\tfrac{1}{2}D_n)a\cdot a = 0, \tag{A6}$$

$$D_t(e^{-D_n} a\cdot b)\cdot cd = e^{-(1/2)D_n}\{[D_t e^{-(1/2)D_n} a\cdot d]\cdot[e^{-(1/2)D_n} c\cdot b] - [e^{-(1/2)D_n} a\cdot d]\cdot[D_t e^{-(1/2)D_n} c\cdot b]\}, \tag{A7}$$

$$\begin{aligned} &2\sinh(\tfrac{1}{2}D_n)\{[D_t e^{-(1/2)D_n} a\cdot b]\cdot[e^{(1/2)D_n} c\cdot d] + [e^{-(1/2)D_n} a\cdot b]\cdot[D_t e^{(1/2)D_n} c\cdot d]\} \\ &= e^{-(1/2)D_n}\{[D_t e^{(1/2)D_n} a\cdot d]\cdot[e^{-(1/2)D_n} c\cdot b] + [e^{(1/2)D_n} a\cdot d]\cdot[D_t e^{-(1/2)D_n} c\cdot b] \\ &\quad - [D_t e^{-(1/2)D_n} a\cdot d]\cdot[e^{(1/2)D_n} c\cdot b] - [e^{-(1/2)D_n} a\cdot d]\cdot[D_t e^{(1/2)D_n} c\cdot b]\}, \end{aligned} \tag{A8}$$

$$\begin{aligned} &2\sinh(\tfrac{1}{2}D_n)[e^{-(1/2)D_n} a\cdot b]\cdot[e^{(1/2)D_n} c\cdot d] \\ &= e^{-(1/2)D_n}\{[e^{(1/2)D_n} a\cdot d]\cdot[e^{-(1/2)D_n} c\cdot b] - [e^{-(1/2)D_n} a\cdot d]\cdot[e^{(1/2)D_n} c\cdot b]\}. \end{aligned} \tag{A9}$$

APPENDIX B: DEDUCTION OF (17)

By use of (9), we have

$$\begin{aligned} 0 &= \frac{1}{\lambda_1}\left[D_z f_0\left(n-\frac{1}{2}\right)\cdot f_1\left(n+\frac{1}{2}\right) - \lambda_1 f_0\left(n+\frac{1}{2}\right) f_1\left(n-\frac{1}{2}\right) + \gamma_1 f_0\left(n-\frac{1}{2}\right) f_1\left(n+\frac{1}{2}\right)\right] f_2\left(n-\frac{1}{2}\right) \\ &\quad - \frac{1}{\lambda_2}\left[D_z f_0\left(n-\frac{1}{2}\right)\cdot f_2\left(n+\frac{1}{2}\right) - \lambda_2 f_0\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)\right. \\ &\quad \left. + \gamma_2 f_0\left(n-\frac{1}{2}\right) f_2\left(n+\frac{1}{2}\right)\right] f_1\left(n-\frac{1}{2}\right) \\ &= -\frac{1}{c\lambda_1\lambda_2} f_{0z}\left(n-\frac{1}{2}\right)[c(\lambda_1 e^{-(1/2)D_n} - \lambda_2 e^{(1/2)D_n}) f_1(n)\cdot f_2(n)] \\ &\quad + f_0\left(n-\frac{1}{2}\right)\left[-\lambda_1^{-1} f_{1z}\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right) + \lambda_2^{-1} f_1\left(n-\frac{1}{2}\right) f_{2z}\left(n+\frac{1}{2}\right)\right. \\ &\quad \left. + \frac{\gamma_1}{\lambda_1} f_1\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right) - \frac{\gamma_2}{\lambda_2} f_1\left(n-\frac{1}{2}\right) f_2\left(n+\frac{1}{2}\right)\right] \\ &= f_0\left(n-\frac{1}{2}\right)\left[-\frac{1}{c\lambda_1\lambda_2} f_{0z}\left(n-\frac{1}{2}\right) f_{12}\left(n+\frac{1}{2}\right) - \lambda_1^{-1} f_{1z}\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)\right. \end{aligned}$$

$$+\lambda_2^{-1} f_1\left(n-\frac{1}{2}\right) f_{2z}\left(n+\frac{1}{2}\right)+\frac{\gamma_1}{\lambda_1} f_1\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)-\frac{\gamma_2}{\lambda_2} f_1\left(n-\frac{1}{2}\right) f_2\left(n+\frac{1}{2}\right)\Bigg]$$

$$=f_0\left(n-\frac{1}{2}\right)\Bigg[-\frac{1}{2c\lambda_1\lambda_2} D_z f_0\left(n-\frac{1}{2}\right)\cdot f_{12}\left(n+\frac{1}{2}\right)-\frac{1}{2c\lambda_1\lambda_2}[e^{-(1/2)D_n} f_0(n)\cdot f_{12}(n)]_z$$

$$-\lambda_1^{-1} f_{1z}\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)+\lambda_2^{-1} f_1\left(n-\frac{1}{2}\right) f_{2z}\left(n+\frac{1}{2}\right)+\frac{\gamma_1}{\lambda_1} f_1\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)$$

$$-\frac{\gamma_2}{\lambda_2} f_1\left(n-\frac{1}{2}\right) f_2\left(n+\frac{1}{2}\right)\Bigg]$$

$$=f_0\left(n-\frac{1}{2}\right)\Bigg\{-\frac{1}{2c\lambda_1\lambda_2} D_z e^{-(1/2)D_n} f_0(n)\cdot f_{12}(n)-\frac{1}{2\lambda_1\lambda_2}\Bigg[\lambda_1 f_1\left(n-\frac{1}{2}\right) f_2\left(n+\frac{1}{2}\right)$$

$$-\lambda_2 f_1\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)\Bigg]_z-\lambda_1^{-1} f_{1z}\left(n+\frac{1}{2}\right) f_2\left(n-\frac{1}{2}\right)+\lambda_2^{-1} f_1\left(n-\frac{1}{2}\right) f_{2z}\left(n+\frac{1}{2}\right)$$

$$+\frac{\gamma_1}{\lambda_1} e^{(1/2)D_n} f_1(n)\cdot f_2(n)-\frac{\gamma_2}{\lambda_2} e^{-(1/2)D_n} f_1(n)\cdot f_2(n)\Bigg\}$$

$$=f_0\left(n-\frac{1}{2}\right)\Bigg\{-\frac{1}{2c\lambda_1\lambda_2} D_z e^{-(1/2)D_n} f_0(n)\cdot f_{12}(n)-\frac{1}{2\lambda_1} D_z e^{(1/2)D_n} f_1(n)\cdot f_2(n)$$

$$-\frac{1}{2\lambda_2} D_z e^{-(1/2)D_n} f_1(n)\cdot f_2(n)+\frac{\gamma_1}{\lambda_1} e^{(1/2)D_n} f_1(n)\cdot f_2(n)-\frac{\gamma_2}{\lambda_2} e^{-(1/2)D_n} f_1(n)\cdot f_2(n)\Bigg\}$$

$$=0,$$

which implies that (17) holds.

[1] M. Blaszak and K. Marciniak, J. Math. Phys. **35**, 4661 (1994).
[2] Y. Wu and X. Geng, J. Math. Phys. **37**, 2338 (1996).
[3] R. Hirota, *Direct Methods in Soliton Theory, in Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980).
[4] R. Hirota and J. Satsuma, Prog. Theor. Phys. Suppl. **59**, 64 (1976).
[5] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[6] X. B. Hu, J. Phys. A: Math. Gen. **27**, 201 (1994).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Binary Darboux–Bäcklund transformations for the Manin–Radul super KdV hierarchy

Jiin-Chang Shaw[a)]
Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

Ming-Hsien Tu[b)]
Department of Physics, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

(Received 27 October 1997; accepted for publication 17 April 1998)

We construct the supersymmetric extensions of the Darboux–Bäcklund transformations (DBTs) for the Manin–Radul super Korteweg–de Vries (KdV) hierarchy using the super-pseudo-differential operators. The elementary DBTs are triggered by the gauge operators constructed from the wave functions and adjoint wave functions of the hierarchy. Iterating these elementary DBTs, we obtain not only Wronskian-type but also binary-type superdeterminant representations of the solutions. © *1998 American Institute of Physics.* [S0022-2488(98)00109-1]

I. INTRODUCTION

In the past few decades, the Darboux–Bäcklund transformations (DBTs) have been shown to be an efficient method to obtain the soliton solutions of the nonlinear integrable systems including Korteweg–de Vries (KdV), Kadomtsev–Petviashili (KP), sine–Gordon, and nonlinear Schrödinger equations (see, for example, Ref. 1), etc. In contrast to these results, the application of the DBTs to superintegrable systems is quite limited. The reasons are partly due to the complexity of the supersymmetric formulation of superintegrable systems. However, as far as the supersoliton solutions are concerned, the supersymmetric generalization of the DBTs is urgently needed.

Recently, Liu[2] proposed a DBT for the Manin–Radul super KdV (MR sKdV) equation.[3] The DBT is triggered by a gauge operator which is a first-order superdifferential operator (SDO) parametrized by a wave function of the associated linear system. By iterating such a DBT, Liu and Mañas[4] obtained the supersymmetric version of the Crum transformation for the MR sKdV equation and showed that the solutions can be expressed in terms of Wronskian superdeterminants.

In this paper, we will investigate the DBTs for the MR sKdV hierarchy from a super-pseudo-differential operator (SΨDO) point of view. Motivated by the non-supersymmetric cases[1,5–12] and the works described above, we introduce the adjoint DBT which is triggered by an adjoint wave function of the system. This adjoint DBT in combination with the previous one forms a binary DBT which is constructed from a wave function and an adjoint wave function. Iterating these elementary DBTs, we obtain not only Wronskian-type but also binary-type superdeterminant representations for the solutions of the MR sKdV hierarchy which have not been obtained previously.

Our paper is organized as follows: In Sec. II, some basic facts about SΨDO are recalled and the MR sKdV hierarchy is defined. In Sec. III, we introduce the elementary DBTs which are triggered by the gauge operators constructed from the wave functions and the adjoint wave functions of the MR sKdV system. In Sec. IV, we iterate these DBTs to obtain the solutions of the MR sKdV hierarchy. Concluding remarks are presented in Sec. V.

II. PRELIMINARIES

Before considering a specific superintegrable system, let us recall some basic facts of the SΨDO which is defined by

[a)]Electronic mail: shaw@math.nctu.edu.tw
[b)]Electronic mail: mhtu@phys.nthu.edu.tw

 © 1998 American Institute of Physics

$$\Lambda = D^N + \sum_{i=-\infty}^{N-1} U_i D^i, \quad N \in \mathbf{Z}, \tag{2.1}$$

where the supercovariant derivative $D \equiv \partial_\theta + \theta\partial$ satisfies $D^2 = \partial$ and θ is the Grassmann variable ($\theta^2 = 0$) which, together with the even variable $x \equiv t_1$, defines the $(1|1)$ superspace with coordinate (x, θ). The formal inverse of D is introduced by $D^{-1} = \theta + \partial_\theta \partial^{-1}$, which satisfies $DD^{-1} = D^{-1}D = 1$. The coefficients U_i are superfields that depend on the variables x, θ, and t_i and can be represented by $U_i = u_i(t) + \theta v_i(t)$. Since the SΨDO is assumed to be homogeneous under Z_2 grading, the grading of the superfield U_i is $|U_i| = N + i \pmod 2$. Here we refer the parity of U_i to be even if $|U_i| = 0$ and odd if $|U_i| = 1$. The supercovariant derivative D satisfies the supersymmetric version of the Leibniz rule:[3]

$$D^i U = \sum_{k=0}^{\infty} (-1)^{|U|(i-k)} \begin{bmatrix} i \\ k \end{bmatrix} (D^k U) D^{i-k}, \tag{2.2}$$

where the superbinomial coefficients $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]$ are defined by

$$\begin{bmatrix} i \\ k \end{bmatrix} = \begin{cases} \begin{pmatrix} [i/2] \\ [k/2] \end{pmatrix} & \text{for } 0 \leqslant k \leqslant i \quad \text{and } (i,k) \neq (0,1) \mod 2 \\ (-1)^{[k/2]} \begin{bmatrix} -i+k-1 \\ k \end{bmatrix} & \text{for } i < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

It is convenient to separate Λ into the positive and negative parts as follows:

$$\Lambda_+ = \sum_{i \geqslant 0} U_i D^i, \quad \Lambda_- = \sum_{i \leqslant -1} U_i D^i. \tag{2.4}$$

The superresidue (sres) of Λ is defined by

$$\text{sres}\,(\Lambda) = U_{-1}, \tag{2.5}$$

and the conjugate operation ''*'' for SΨDOs can be defined as follows

$$(PQ)^* = (-1)^{|P||Q|} Q^* P^*, \tag{2.6}$$

which implies that $U^* = U$ for arbitrary superfield U and

$$(\partial^i)^* = (-1)^i \partial^i, \quad (D^i)^* = (-1)^{i(i+1)/2} D^i. \tag{2.7}$$

A simple calculation shows that

$$\Lambda^* = \sum_i (-1)^N (-1)^{i(i-1)/2} D^i U_i. \tag{2.8}$$

With these definitions in hand, let us provide some useful identities which will simplify the computations involving compositions of SΨDOs.

Lemma 1:

$$(\Lambda^*)_+ = (\Lambda_+)^*, \quad (\Lambda^*)_- = (\Lambda_-)^*, \tag{2.9}$$

$$(D^{-1}\Lambda)_- = D^{-1}(\Lambda^*)_0 + D^{-1}(\Lambda)_-, \tag{2.10}$$

$$(\Lambda D^{-1})_- = (\Lambda)_0 D^{-1} + \Lambda_- D^{-1}, \tag{2.11}$$

$$\text{sres}\,(\Lambda) = \text{sres}\,(\Lambda^*), \quad (D\ \text{sres}\,(\Lambda)) = \text{sres}\,(D\Lambda - (-1)^{|\Lambda|}\Lambda D), \tag{2.12}$$

$$\text{sres}\,(\Lambda D^{-1})=(\Lambda)_0, \quad \text{sres}\,(D^{-1}\Lambda)=(-1)^{|\Lambda|}(\Lambda^*)_0, \tag{2.13}$$

$$\text{sres}\,(D^{-1}\Lambda_1\Lambda_2 D^{-1})=\text{sres}\,(D^{-1}(\Lambda_1^*)_0\Lambda_2 D^{-1})+\text{sres}\,(D^{-1}\Lambda_1(\Lambda_2)_0 D^{-1}), \tag{2.14}$$

where $\Lambda_1=(\Lambda_1)_+$ and $\Lambda_2=(\Lambda_2)_+$.

Proof: The proofs for these identities are straightforward. Here we only give the proof for the second identity. From the left-hand side, we have

$$(D^{-1}\Lambda)_-=(D^{-1}\Lambda_+)_-+D^{-1}\Lambda_-. \tag{2.15}$$

Then using (2.2), we obtain

$$\begin{aligned}(D^{-1}\Lambda_+)_- &= \left(\sum_{k=0}^{\infty}\sum_{i=0}^{N}(-1)^{(N+i)(1+k)}\begin{bmatrix}-1\\ k\end{bmatrix}(D^k U_i)D^{-1-k+i}\right)_- \\ &= \sum_{l=0}^{\infty}\sum_{i=0}^{N}(-1)^{(N+i)(1+i+l)}(-1)^{[(i+l)/2]}(D^{i+l}U_i)D^{-1-l} \\ &= \sum_{l=0}^{\infty}(-1)^{N(1+l)}(-1)^{[l/2]}(D^l(\Lambda^*)_0)D^{-1-l}=D^{-1}(\Lambda^*)_0. \end{aligned}\tag{2.16}$$

Here the relation $(-1)^{[i/2]}=(-1)^{i(i-1)/2}$ has been used to reach the third line of (2.16). □

The MR sKdV hierarchy is defined, in Lax form, as

$$\partial_{t_n}L=[P_n,\,L], \quad n=1,3,5,..., \tag{2.17}$$

with

$$L=\partial^2+v_1D+v_0, \tag{2.18}$$

$$P_n=L_+^{n/2}, \tag{2.19}$$

where the coefficients v_1 and v_0 are superfields depending on the variables $(\theta,x,t_3,t_5,\ldots)$ with grading $|v_i|=i \pmod 2$. We can rewrite the hierarchy equations (2.17) as follows:

$$\partial_{t_m}P_n-\partial_{t_n}P_m+[P_n,\,P_m]=0, \tag{2.20}$$

which is called the zero-curvature condition and is equivalent to the whole set of equations of (2.17). If we can find a set of superfields $\{v_1,v_0\}$ and hence a corresponding set of superdifferential operators (SDO) $\{P_n\}$ satisfying (2.20), then we have a solution to the MR sKdV hierarchy. In fact, the MR sKdV hierarchy was constructed originally from the MR sKP hierarchy[3] associated with the odd SΨDO: $\Lambda_{\text{MR}}=D+\Sigma_{i=0}^{\infty}U_{-i}D^{-i}$ by the reduction $L=(\Lambda_{\text{MR}}^4)_+$ and has been shown to be bi-Hamiltonian.[13,14] Substituting (2.18) and (2.19) into the Lax equation (2.17) for $n=3$, one obtains

$$\partial_{t_3}v_0=\tfrac{1}{4}\partial_x(v_{0xx}+3v_0^2+3v_1(Dv_0)), \tag{2.21}$$

$$\partial_{t_3}v_1=\tfrac{1}{4}\partial_x(v_{1xx}+3v_1(Dv_1)+6v_1v_0), \tag{2.22}$$

which is the MR sKdV equation. By setting $v_1=0$, Eqs. (2.21) and (2.22) reduce to the KdV equation. The Lax equation (2.17) can be viewed as the compatibility condition of the linear system

$$L\phi=\lambda\phi, \quad \partial_{t_n}\phi=(P_n\phi)_0, \tag{2.23}$$

where ϕ and λ are called wave function and spectral parameter of the hierarchy, respectively. On the other hand, we can also introduce adjoint wave function ψ which satisfies the linear system

$$L^*\psi=\eta\psi, \quad \partial_{t_n}\psi=-((P_n)^*\psi)_0. \tag{2.24}$$

For convenience, throughout this paper, ϕ and ψ will stand for wave function and adjoint wave function, respectively. Of course, it should be realized that both ϕ and ψ are superfields in this formalism. In next section, we will use these (adjoint) wave functions to construct the DBTs for the MR sKdV hierarchy.

III. THE ELEMENTARY DB TRANSFORMATIONS

We consider the following transformation:

$$L\to\hat{L}=TLT^{-1}, \tag{3.1}$$

where $T=T(\theta,x,t_3,t_5,\ldots)$ is any reasonable SΨDO. To guarantee that such transformations can generate new solutions of the MR sKdV hierarchy, the transformed Lax operator $\hat{L}$ should preserve the form of (2.18) and satisfy the Lax equation (2.17). It is easy to show that under the transformation (3.1), the operator P_n then is transformed as

$$P_n\to\hat{P}_n=TP_nT^{-1}+\partial_{t_n}TT^{-1}, \tag{3.2}$$

which preserves the zero curvature condition (2.20). Note that although P_n is a SDO, the right-hand side of (3.2) will, in general, not be a purely SDO. However, if we suitably choose the gauge operator T such that $\hat{P}_n$, as defined by (3.2), is also a purely SDO, then $\{\hat{L},\hat{P}_n\}$ represents a valid new solution to the MR sKdV hierarchy. To formulate the DBTs of the MR sKdV hierarchy, let us introduce a superfield Ω called bilinear potential[6] (or squared eigenfunction potential) which is constructed from a wave function and an adjoint wave function and will be useful later on.

Lemma 2: For any pair of ϕ and ψ, there exists a bilinear potential $\Omega(\psi,\phi)$ satisfying

$$(D\Omega(\psi,\phi))=\psi\phi, \tag{3.3}$$

$$\partial_{t_n}\Omega(\psi,\phi)=\operatorname{sres}\,(D^{-1}\psi P_n\phi D^{-1}). \tag{3.4}$$

Proof: To prove the existence of the bilinear potential Ω, we have to show that Eqs. (3.3) and (3.4) are compatible. First, we note that (3.3) is consistent with (3.4) for $n=1$. Since

$$\Omega_x=\operatorname{sres}\,(D^{-1}\psi P_1\phi D^{-1})=\operatorname{sres}\,(D^{-1}\psi D^2\phi D^{-1})=(D\psi)\phi+(-1)^{|\psi|}\psi(D\phi)=(D(D\Omega)), \tag{3.5}$$

this implies that

$$(D\Omega)_x=(D(D(D\Omega)))=(D\Omega_x). \tag{3.6}$$

In general, from (3.3) and (3.4) we have

$$\begin{aligned}(D\partial_{t_n}\Omega)&=(D\,\operatorname{sres}\,(D^{-1}\psi P_n\phi D^{-1}))\\&=\operatorname{sres}\,(\psi P_n\phi D^{-1}-(-1)^{|\psi||\phi|}D^{-1}\psi P_n\phi)\\&=\psi(P_n\phi)-(-1)^{|\psi||\phi|}\phi(P_n^*\psi)\\&=\psi\phi_{t_n}+\psi_{t_n}\phi\\&=\partial_{t_n}(D\Omega).\end{aligned} \tag{3.7}$$

Second, from (3.4) and (2.14), we have

$$\begin{aligned}\partial_{t_m}(\partial_{t_n}\Omega)-\partial_{t_n}(\partial_{t_m}\Omega)&=\operatorname{sres}\,\partial_{t_m}(D^{-1}\psi P_n\phi D^{-1})-\operatorname{sres}\,\partial_{t_n}(D^{-1}\psi P_m\phi D^{-1})\\&=\operatorname{sres}\,(D^{-1}\psi(\partial_{t_m}P_n-\partial_{t_n}P_m)\phi D^{-1})\end{aligned}$$

$$
\begin{aligned}
&-\operatorname{sres}\,(D^{-1}\psi(P_m^*\psi)P_n\phi D^{-1})+\operatorname{sres}\,(D^{-1}\psi P_n(P_m\phi)D^{-1})\\
&+\operatorname{sres}\,(D^{-1}\psi(P_n^*\psi)P_m\phi D^{-1})-\operatorname{sres}\,(D^{-1}\psi P_m(P_n\phi)D^{-1})\\
&=\operatorname{sres}\,(D^{-1}\psi(\partial_{t_m}P_n-\partial_{t_n}P_m+[P_n,\,P_m])\phi D^{-1})=0. \qquad (3.8)
\end{aligned}
$$

Here the zero-curvature condition of the MR sKdV hierarchy has been used in the last line. □

Remarks: $\Omega(\psi,\phi)$ is a superfield with parity $(-1)^{|\psi||\phi|+1}$ and can be represented by $\Omega(\psi,\phi)=\int^x \psi_2\phi_1+(-1)^{|\psi|}\int^x \psi_1\phi_2+\theta\psi_1\phi_1$, where $\phi=\phi_1+\theta\phi_2$ and $\psi=\psi_1+\theta\psi_2$.

Proposition 1:[2] Let χ be an even wave function of the linear system. Then the gauge operator

$$T=\chi D\chi^{-1}=D+\alpha,\quad \alpha\equiv-\frac{(D\chi)}{\chi}, \qquad (3.9)$$

triggers the following DBT:

$$\hat{L}=TLT^{-1}=\partial^2+\hat{v}_1D+\hat{v}_0, \qquad (3.10)$$

$$\hat{\phi}=(T\phi)=\chi(D\chi^{-1}\phi), \qquad (3.11)$$

$$\hat{\psi}=((T^{-1})^*\psi)=\chi^{-1}\Omega(\psi,\chi), \qquad (3.12)$$

where the transformed coefficients are given by

$$\hat{v}_1=-v_1-2\alpha_x, \qquad (3.13)$$

$$\hat{v}_0=v_0+(Dv_1)+2\alpha(v_1+\alpha_x). \qquad (3.14)$$

Proof: We first show that $\hat{L}$ has the same form as L. From (3.10) we have

$$\hat{L}_-=(\chi D\chi^{-1}L\chi D^{-1}\chi^{-1})_-=\chi(D\chi^{-1}(L\chi))D^{-1}\chi^{-1}=0, \qquad (3.15)$$

which implies that $\hat{L}$ is a SDO. A simple calculation shows that the transformed coefficients $\hat{v}_1$ and $\hat{v}_0$ are given by (3.13) and (3.14), respectively. Furthermore, the hierarchy equation for $\hat{L}$ is given by

$$\partial_{t_n}\hat{L}=[TP_nT^{-1}+\partial_{t_n}TT^{-1},\,\hat{L}]. \qquad (3.16)$$

Using the identity

$$\chi D\chi^{-1}\Lambda_+\chi D^{-1}\chi^{-1}=(\chi D\chi^{-1}\Lambda\chi D^{-1}\chi^{-1})_+ +\chi(D\chi^{-1}(\Lambda_+\chi)_0)D^{-1}\chi^{-1} \qquad (3.17)$$

and then substituting $\Lambda=L^{n/2}$ (n odd), we obtain

$$\hat{P}_n=TP_nT^{-1}+\partial_{t_n}TT^{-1}=(TL^{n/2}T^{-1})_+ +\chi(D\chi^{-1}\partial_{t_n}\chi)D^{-1}\chi^{-1}+\partial_{t_n}TT^{-1}=\hat{L}_+^{n/2}. \qquad (3.18)$$

Hence

$$\partial_{t_n}\hat{L}=[\hat{L}_+^{n/2},\,\hat{L}]. \qquad (3.19)$$

The evolution equations for $\hat{\phi}$ and $\hat{\psi}$ can be verified in a similar way. From (3.11), we have

$$\partial_{t_n}\hat{\phi}=\partial_{t_n}T\phi+T\partial_{t_n}\phi=\partial_{t_n}TT^{-1}\hat{\phi}+TP_nT^{-1}\hat{\phi}=(\hat{P}_n\hat{\phi})_0. \qquad (3.20)$$

Similarly,

$$\partial_{t_n}\hat{\psi}=\partial_{t_n}(T^{-1})^*\psi+(T^{-1})^*\partial_{t_n}\psi=((T^{-1})^*(\partial_{t_n}T)^*+(T^{-1})^*P_n^*T^*)\hat{\psi}=-(\hat{P}_n^*\hat{\psi})_0. \qquad (3.21)$$

□

Having described the first construction of the DBT, we now turn to another construction using the adjoint wave function.

Proposition 2: Let μ be an even adjoint wave function of the linear system (2.24). Then the gauge operator

$$S=\mu^{-1}D^{-1}\mu=(D-\beta)^{-1},\quad \beta\equiv-\frac{(D\mu)}{\mu}, \tag{3.22}$$

triggers the following adjoint DB transformation:

$$\hat{L}=SLS^{-1}=\partial^2+\hat{v}_1D+\hat{v}_0, \tag{3.23}$$

$$\hat{\phi}=(S\phi)=\mu^{-1}\Omega(\mu,\phi), \tag{3.24}$$

$$\hat{\psi}=((S^{-1})^*\psi)=-\mu(D\mu^{-1}\psi), \tag{3.25}$$

where

$$\hat{v}_1=-v_1+2\beta_x, \tag{3.26}$$

$$\hat{v}_0=v_0+(Dv_1)-2(D\beta)_x-2\beta(v_1-\beta_x). \tag{3.27}$$

Proof: The proof is similar to the Proposition 1. Here we only mention that Eq. (3.19) can be verified easily by using the identity

$$(\mu^{-1}D^{-1}\mu\Lambda\mu^{-1}D\mu)_-=\mu^{-1}D^{-1}\mu\Lambda_-\mu^{-1}D\mu-(-1)^{|\Lambda|}\mu^{-1}D^{-1}(D\mu^{-1}(\Lambda_+^*\mu)_0)\mu \tag{3.28}$$

instead of (3.17). □

Proposition 3: For any pair of even wave function χ and odd adjoint wave function μ, the gauge operator

$$R=1-\chi\Omega(\mu,\chi)^{-1}D^{-1}\mu \tag{3.29}$$

triggers the following binary DB transformation:

$$\hat{L}=RLR^{-1}=\partial^2+\hat{v}_1D+\hat{v}_0, \tag{3.30}$$

$$\hat{\phi}=(R\phi)=\phi-\chi\Omega(\mu,\chi)^{-1}\Omega(\mu,\phi), \tag{3.31}$$

$$\hat{\psi}=((R^{-1})^*\psi)=\psi-\mu\Omega(\mu,\chi)^{-1}\Omega(\psi,\chi), \tag{3.32}$$

where

$$\hat{v}_1=v_1-2\Gamma_x,\quad \Gamma\equiv\frac{(D\Omega)}{\Omega}, \tag{3.33}$$

$$\hat{v}_0=v_0+2(D\Gamma)_x+2v_1\Gamma-2\Gamma\alpha_x+2\alpha\Gamma_x+2\Gamma\Gamma_x. \tag{3.34}$$

Proof: The gauge operator R is just a composition of the DB transformation T and the adjoint DB transformation S. To see this, let us first perform a DB transformation triggered by the wave function χ. The odd adjoint wave function μ is thus transformed to $\hat{\mu}=\chi^{-1}\Omega(\mu,\chi)$ which is an even one. Then a subsequent adjoint DB transformation S triggered by $\hat{\mu}$ is performed and the composition of these two transformations is given by $R=\hat{\mu}^{-1}D^{-1}\hat{\mu}\chi D\chi^{-1}$ $=\chi\Omega^{-1}D^{-1}\Omega D\chi^{-1}=1-\chi\Omega^{-1}D^{-1}\mu$. Therefore, the remaining part of the proof is just a corollary of the Propositions 1 and 2. □

Remarks: The binary DB transformation can also be constructed from a pair of χ and μ with $|\chi|=1$ and $|\mu|=0$. A direct calculation shows that $R=\hat{\chi}D\hat{\chi}^{-1}\mu^{-1}D^{-1}\mu=1-\chi\Omega^{-1}D^{-1}\mu$, which has the same form as (3.29). We also note that the transformed coefficients (3.33) and (3.34) are just the ones in Ref. 2 where a different approach was presented.

IV. ITERATION OF THE DB TRANSFORMATIONS

We have introduced the elementary DBTs which contain the DBT, the adjoint DBT, and the binary DBT triggered by the gauge operators *T*, *S*, and *R*, respectively. Using these elementary transformations as the building blocks, the more complicated transformations can be constructed from the compositions of these gauge operators. In Ref. 4, by iterating the DB transformation *T*, the so-called Crum transformation for the MR sKdV equation was constructed and the Wronskian superdeterminant representations for the solutions were obtained. This construction starts with *n* wave functions χ_i, $i=0,...,n-1$ of the linear system (2.23) with parity $(-1)^i$. Using χ_0, say, to perform the first DBT of Proposition 1, then χ_i are transformed to $\hat{\chi}_i$. It is obvious that $\hat{\chi}_0=0$. The next step is to perform a subsequent DBT triggered by $\hat{\chi}_1$, say, which leads to the new wave functions $\hat{\hat{\chi}}_i$ with $\hat{\hat{\chi}}_1=0$. Iterating this process such that all the wave functions are used up, then an *n*-step DBT with gauge operator T_n is obtained.

Proposition 4:[4] Let χ_i, $i=0,...,n-1$ be wave functions of the linear system (2.23) with parity $(-1)^i$. Then after *n* iterations of the DBT of Proposition 1, the transformed Lax operator becomes

$$\hat{L}=T_nLT_n^{-1},\quad T_n=D^n+\sum_{i=0}^{n-1}a_iD^i, \tag{4.1}$$

$$=\partial^2+\hat{v}_1D+\hat{v}_0, \tag{4.2}$$

where the coefficients a_i are defined by

$$T_n\chi_j=0,\quad j=0,...,n-1, \tag{4.3}$$

and

$$\hat{v}_1=(-1)^nv_1-2(a_{n-1})_x, \tag{4.4}$$

$$\hat{v}_0=v_0-2(a_{n-2})_x-a_{n-1}((-1)^nv_1+\hat{v}_1)+\frac{1-(-1)^n}{2}(Dv_1). \tag{4.5}$$

Remarks: Due to the fact that the parity of the gauge operator T_n depends on *n*, the cases for *n* being even or odd are rather different. In Ref. 4, the coefficients a_i for both cases have been obtained separately by solving the linear equation (4.3) and the transformed fields $\hat{v}_1$ and $\hat{v}_0$ can be expressed in a compact form in terms of the superdeterminant. We will present such a calculation in a more general case involving the binary DBTs.

Now let us turn to a similar construction using the adjoint wave functions.

Proposition 5: Let μ_i, $i=0,...,n-1$ be adjoint wave functions of the linear system (2.24) with parity $(-1)^i$. Then after *n* iterations of the adjoint DBT of Proposition 2, the transformed Lax operator becomes

$$\hat{L}=S_nLS_n^{-1},\quad (S_n^{-1})^*=(-1)^{[n/2]}\left(D^n+\sum_{i=0}^{n-1}b_iD^i\right) \tag{4.6}$$

$$=\partial^2+\hat{v}_1D+\hat{v}_0, \tag{4.7}$$

where the coefficients b_i are defined by

$$(S_n^{-1})^*\mu_j=0,\quad j=0,...,n-1, \tag{4.8}$$

and

$$\hat{v}_1=(-1)^n v_1+2(b_{n-1})_x, \tag{4.9}$$

$$\hat{v}_0=v_0-2(b_{n-2})_x+b_{n-1}((-1)^n v_1+\hat{v}_1)+\frac{1+(-1)^n}{2}(Dv_1)-(D\hat{v}_1). \tag{4.10}$$

Proof: Essentially, the proof is exactly the same as Proposition 4. □

Proposition 6: Let $\chi_0,\ldots,\chi_{n-1}$ be even wave functions of the linear system (2.23) and $\mu_0,\ldots,\mu_{n-1}$ be odd adjoint wave functions of the linear system (2.24). Then after n iterations of the binary DBT of Proposition 3, the transformed Lax operator becomes

$$\hat{L}=R_n L R_n^{-1}, \quad R_n=1-\sum_{i=0}^{n-1} c_i D^{-1}\mu_i, \tag{4.11}$$

$$=\partial^2+\hat{v}_1 D+\hat{v}_0, \tag{4.12}$$

where the coefficient c_i are defined by

$$R_n\chi_j=0, \quad j=0,\ldots,n-1, \tag{4.13}$$

and

$$\hat{v}_1=v_1-2\sum_{i=0}^{n-1}(c_i\mu_i)_x, \tag{4.14}$$

$$\hat{v}_0=v_0+(v_1+\hat{v}_1)\sum_{i=0}^{n-1} c_i\mu_i+2\sum_{i=0}^{n-1}(c_i(D\mu_i))_x. \tag{4.15}$$

Proof: A straightforward calculation. We only remark that the form of the gauge operator R_n can be verified by induction. □

Proposition 7: The transformed coefficients, $\hat{v}_1$ and $\hat{v}_0$, in Eqs. (4.14) and (4.15) can be expressed as

$$\hat{v}_1=v_1-2(D^3 \ln\det\Omega), \tag{4.16}$$

$$\hat{v}_0=v_0+(v_1+\hat{v}_1)(D\ln\det\Omega)+2\sum_{i=0}^{n-1}\left(D^2\left(\frac{\det\Omega^{(i)}}{\det\Omega}(D\mu_i)\right)\right), \tag{4.17}$$

where $\Omega_{ij}\equiv\Omega(\mu_i,\chi_j)$ and $\Omega^{(i)}$ is constructed from Ω with its ith row replaced by $(\chi_0,\ldots,\chi_{n-1})$.

Proof: The proof follows easily from the Cramer's formula. □

Remarks: The expressions of the transformed fields $\hat{v}_1$ and $\hat{v}_0$ show that they are unchanged under the interchange of any two wave functions χ_i and χ_j with $i\neq j$. Hence the permutability of DBTs is still maintained in this supersymmetric formalism. We also note that the case for $|\chi_i|=1$ and $|\mu_i|=0$ gives the same result.

Finally let us discuss a more general DBT in which the numbers of wave functions and adjoint wave functions are unequal.

Proposition 8: Suppose there are n adjoint wave functions $\mu_0,\ldots,\mu_{n-1}$ and $n+m$ wave functions $\chi_0,\ldots,\chi_{n+m-1}$ of the MR sKdV system. Assume the parities of μ_i are all odd, whereas among χ_i, $\chi_0,\ldots,\chi_{n+[m+1/2]-1}$ are even and $\chi_{n+[m+1/2]},\ldots,\chi_{n+m-1}$ are odd. Then after performing the DB transformations triggered by these (adjoint) wave functions, the transformed Lax operator reads

$$\hat{L}=Q_{(n,m)}LQ_{(n,m)}^{-1}, \quad Q_{(n,m)}=D^m+\sum_{i=0}^{m-1} d_i D^i+\sum_{i=0}^{n-1} e_i D^{-1}\mu_i, \tag{4.18}$$

$$=\partial^2+\hat{v}_1 D+\hat{v}_0, \tag{4.19}$$

where the coefficients d_i and e_i are defined by

$$Q_{(n,m)}\chi_i=0, \quad i=0,...,n+m-1, \tag{4.20}$$

and

$$\hat{v}_1=(-1)^m v_1-2(d_{m-1})_x, \tag{4.21}$$

$$\hat{v}_0=v_0-2(d_{m-2})_x-d_{m-1}((-1)^m v_1+\hat{v}_1)+\frac{1-(-1)^m}{2}(Dv_1). \tag{4.22}$$

Proof: The gauge operator $Q_{(n,m)}$ can be realized easily as follows: since the number of wave functions is larger than that of the adjoint wave functions, thus part of the even wave functions, say, $\chi_0,\ldots,\chi_{n-1}$, would be paired with odd adjoint wave functions $\mu_0,\ldots,\mu_{n-1}$ to form the binary DB transformation R_n of Proposition 7 and the residual wave functions $\chi_n,\ldots,\chi_{n+m-1}$ will form the DB transformation T_m of Proposition 4. Hence $Q_{(n,m)}=T_m R_n$ and its expression in (4.18) can be verified by induction. □

Notice that the parity of the gauge operator $Q_{(n,m)}$ depends on m, therefore the case for m being even or odd should be discussed separately.

For $m=2k$, it is convenient to define the following row vectors:

$$\boldsymbol{\chi}^{[0]}=(\chi_0,\chi_1,\ldots,\chi_{n+k-1}), \quad \boldsymbol{\chi}^{[1]}=(\chi_{n+k},\ldots,\chi_{n+2k-1}), \tag{4.23}$$

$$\mathbf{p}^{[0]}=(e_0,e_1,\ldots,e_{n-1},d_0,d_2,\ldots,d_{2k-2}), \quad \mathbf{p}^{[1]}=(d_1,d_3,\ldots,d_{2k-1}), \tag{4.24}$$

$$\mathbf{q}^{[i]}=\partial_x^k\boldsymbol{\chi}^{[i]}, \quad i=0,1, \tag{4.25}$$

and matrices

$$\Omega_{ij}^{[0]}=\Omega(\mu_i,\chi_j), \quad i=0,...,n-1; j=0,...,n+k-1, \tag{4.26}$$

$$\Omega_{ij}^{[1]}=\Omega(\mu_i,\chi_{n+k+j}), \quad i=0,...,n-1; j=0,...,k-1, \tag{4.27}$$

$$\Xi^{[i]}=\begin{pmatrix} \boldsymbol{\chi}^{[i]} \\ \partial_x\boldsymbol{\chi}^{[i]} \\ \vdots \\ \partial_x^{k-1}\boldsymbol{\chi}^{[i]} \end{pmatrix}, \quad i=0,1, \tag{4.28}$$

where the superscript in brackets stands for the parity of the corresponding vector or matrix.

Proposition 9: After solving the linear equation (4.20) for $m=2k$, the transformed coefficients, $\hat{v}_1$ and $\hat{v}_0$ in Eqs. (4.21) and (4.22) can be expressed as

$$\hat{v}_1=v_1+2\left(\frac{\det(\hat{D}-\hat{C}A^{-1}B)}{\det(D-CA^{-1}B)}\right)_x, \tag{4.29}$$

$$\hat{v}_0=v_0+2\left(\frac{\operatorname{sdet}\hat{\mathbf{M}}}{\operatorname{sdet}\mathbf{M}}\right)_x-(v_1+\hat{v}_1)\left(\frac{\det(\hat{D}-\hat{C}A^{-1}B)}{\det(D-CA^{-1}B)}\right), \tag{4.30}$$

where the supermatrix

$$\mathbf{M}=\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with

$$A=\begin{pmatrix} \Omega^{[0]} \\ \Xi^{[0]} \end{pmatrix}, \quad B=\begin{pmatrix} \Omega^{[1]} \\ \Xi^{[1]} \end{pmatrix}, \tag{4.31}$$

$$C=(D\Xi^{[0]}),\quad D=(D\Xi^{[1]}),$$

and its superdeterminant is defined by[15]

$$\text{sdet }\mathbf{M}=\frac{\det(A-BD^{-1}C)}{\det D}=\frac{\det A}{\det(D-CA^{-1}B)}. \tag{4.32}$$

The matrix $\hat{\mathbf{M}}$ is constructed from $\mathbf{M}$ with its $(n+k)$th row replaced by its ∂_x derivation, whereas $\hat{D}$ and $\hat{C}$ are constructed from D and C with their last rows replaced by their D derivation.

Proof: For $m=2k$, the transformed coefficients $\hat{v}_1$ and $\hat{v}_0$ are expressed as

$$\hat{v}_1=v_1-2(d_{2k-1})_x, \tag{4.33}$$

$$\hat{v}_0=v_0-2(d_{2k-2})_x+(v_1+\hat{v}_1)d_{2k-1}. \tag{4.34}$$

Since the parity of the operator $Q_{(n,2k)}$ is even, we can rewrite the linear equation (4.20) as follows:

$$(\mathbf{p}^{[0]},\mathbf{p}^{[1]})\mathbf{M}=-(\mathbf{q}^{[0]},\mathbf{q}^{[1]}). \tag{4.35}$$

Now multiplying $\mathbf{M}^{-1}$ to the right of (4.35), we have

$$(\mathbf{p}^{[0]},\mathbf{p}^{[1]})=-(\mathbf{q}^{[0]},\mathbf{q}^{[1]})\begin{pmatrix}(A-BD^{-1}C)^{-1} & -AB^{-1}(D-CA^{-1}B)\\ -D^{-1}C(A-BD^{-1}C)^{-1} & (D-CA^{-1}B)^{-1}\end{pmatrix}, \tag{4.36}$$

which implies

$$\mathbf{p}^{[0]}(A-BD^{-1}C)=-(\mathbf{q}^{[0]}-\mathbf{q}^{[1]}D^{-1}C), \tag{4.37}$$

$$\mathbf{p}^{[1]}(D-CA^{-1}B)=-(\mathbf{q}^{[1]}-\mathbf{q}^{[0]}A^{-1}B). \tag{4.38}$$

Using Cramer's rule, we obtain

$$d_{2k-2}=\mathbf{P}^{[0]}_{n+k}=-\frac{\text{sdet }\hat{\mathbf{M}}}{\text{sdet }\mathbf{M}}, \tag{4.39}$$

$$d_{2k-1}=\mathbf{P}^{[1]}_{k}=-\frac{\det(\hat{D}-\hat{C}A^{-1}B)}{\det(D-CA^{-1}B)}, \tag{4.40}$$

which lead to the results (4.29) and (4.30). □

For $m=2k+1$, some notations in Eqs. (4.23)–(4.28) should be modified to

$$\boldsymbol{\chi}^{[0]}=(\chi_0,\chi_1,\ldots,\chi_{n+k}),\quad \boldsymbol{\chi}^{[1]}=(\chi_{n+k+1},\ldots,\chi_{n+2k}), \tag{4.41}$$

$$\mathbf{p}^{[0]}=(d_1,d_3,\ldots,d_{2k-1}),\quad \mathbf{p}^{[1]}=(e_0,e_1,\ldots,e_{n-1},d_0,d_2,\ldots,d_{2k}), \tag{4.42}$$

$$\mathbf{q}^{[i]}=(D^{2k+1}\boldsymbol{\chi}^{[i]}),\quad i=0,1, \tag{4.43}$$

$$\Omega^{[0]}_{ij}=\Omega(\mu_i,\chi_j),\quad i=0,\ldots,n-1,\quad j=0,\ldots,n+k, \tag{4.44}$$

$$\Omega^{[1]}_{ij}=\Omega(\mu_i,\chi_{n+k+j+1}),\quad i=0,\ldots,n-1,\quad j=0,\ldots,k-1, \tag{4.45}$$

$$\Xi^{[i]}=\begin{pmatrix}\boldsymbol{\chi}^{[0]}\\ \partial_x\boldsymbol{\chi}^{[0]}\\ \vdots\\ \partial_x^k\boldsymbol{\chi}^{[0]}\end{pmatrix},\qquad i=0,1. \tag{4.46}$$

Then we can define another supermatrix

$$\mathbf{N}=\begin{pmatrix} \breve{A} & \breve{B} \\ \breve{C} & D \end{pmatrix}$$

where the matrices $\breve{A}$ and $\breve{B}$ are constructed from the matrices A and B by taking into account the replacements (4.41)–(4.46), whereas $\breve{C}$ does the same thing in addition to removing the last row.

Proposition 10: Solving the linear equation (4.20) for $m=2k+1$, the transformed coefficients, $\hat{v}_1$ and $\hat{v}_0$, in Eqs. (4.21) and (4.22) can be expressed as

$$\hat{v}_1=-v_1+2\left(\frac{\det(\tilde{A}-\tilde{B}D^{-1}\breve{C})}{\det(\breve{A}-\breve{B}D^{-1}\breve{C})}\right)_x, \tag{4.47}$$

$$\hat{v}_0=v_0+(Dv_1)+2\left(\frac{\operatorname{sdet}\mathbf{N}}{\operatorname{sdet}\tilde{\mathbf{N}}}\right)_x+(v_1-\hat{v}_1)\left(\frac{\det(\tilde{A}-\tilde{B}D^{-1}\breve{C})}{\det(\breve{A}-\breve{B}D^{-1}\breve{C})}\right), \tag{4.48}$$

where $\tilde{\mathbf{N}}$ is constructed from $\mathbf{N}$ with its last row replaced by its ∂_x derivation, whereas $\tilde{A}$ and $\tilde{B}$ are constructed from $\breve{A}$ and $\breve{B}$ with their last rows replaced by their D derivation.

Proof: For $m=2k+1$, the linear equation (4.20) now becomes

$$(\mathbf{p}^{[1]},\mathbf{p}^{[0]})\mathbf{N}=-(\mathbf{q}^{[1]},\mathbf{q}^{[0]}). \tag{4.49}$$

Following the procedures for the case of $m=2k$, it is easy to show that

$$d_{2k}=\mathbf{p}^{[1]}_{n+k+1}=\frac{\det(\tilde{A}-\tilde{B}D^{-1}\breve{C})}{\det(\breve{A}-\breve{B}D^{-1}\breve{C})}, \tag{4.50}$$

$$d_{2k-1}=\mathbf{p}^{[0]}_{k}=-\frac{\operatorname{sdet}\mathbf{N}}{\operatorname{sdet}\tilde{\mathbf{N}}}, \tag{4.51}$$

which imply that the transformed coefficients $\hat{v}_1$ and $\hat{v}_0$ are given by (4.47) and (4.48), respectively. □

Remarks: (1) In both cases ($m=2k$ or $2k+1$), the expressions for $\hat{v}_1$ and $\hat{v}_0$ are, indeed, independent of the order of the even wave functions as well as the odd ones. (2) For the case that the number of the adjoint wave functions is larger than that of the wave functions in a DBT, we can get a similar result just by exchanging the roles played by them. (3) When we set $n=0$, the binary type expressions for $\hat{v}_1$ and $\hat{v}_0$ then reduce to the Wronskian-type ones obtained in Ref. 4. Especially in the $m=2k$ case, it is possible to write d_{2k-1} in terms of superdeterminant.

V. CONCLUDING REMARKS

We have studied the solutions of the MR sKdV hierarchy by using DBTs. In addition to the previously known DBT,[2,4] we provide the adjoint DBT which can be combined with the former one to form the binary DBT. Using these elementary DBTs, we obtain not only Wronskian-type but also binary-type solutions for the MR sKdV hierarchy. The super soliton solutions then can be constructed from the trivial one which corresponds to $v_1=v_0=0$ by using the formulas derived in Sec. IV. Finally, we would like to remark that our approach involves only the algebra of SΨDOs, hence the formulation is general enough to extend to the other cases such as super KP hierarchies and their reductions. We will leave these discussions to another publication.

Note added in proof: After submission of this manuscript for publication we were informed by Q. P. Liu that Propositions 6 and 7 were also obtained in Ref. 16 by using the vectorial Darboux transformation.

ACKNOWLEDGMENT

This work is supported by the National Science Council of the Republic of China under Grant No. NSC-87-2811-M-007-0025.

[1] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer, Heidelberg, 1991).
[2] Q. P. Liu, Lett. Math. Phys. **35**, 115 (1995).
[3] Y. I. Manin and A. O. Radul, Commun. Math. Phys. **98**, 65 (1997).
[4] Q. P. Liu and M. Mañas, Phys. Lett. B **396**, 133 (1997).
[5] L. L. Chau, J. C. Shaw, and H. C. Yen, Commun. Math. Phys. **149**, 263 (1992).
[6] W. Oevel, Physica A **195**, 533 (1993); W. Oevel and W. Schief, in *Applications of Analytic and Geometric Methods in Nonlinear Differential Equations*, edited by P. Clarkson (Kluwer, Dordrecht, 1993).
[7] J. J. C. Nimmo, in *Nonlinear Evolution Equations and Dynamical Systems*, edited by V. G. Makhankov, A. R. Bishop, and D. D. Holm (World Scientific, Singapore, 1995).
[8] H. Aratyn, E. Nissimov, and S. Pacheva, Phys. Lett. A **210**, 293 (1995); **228**, 164 (1997).
[9] H. Aratyn, E. Nissimov, and S. Pacheva, Commun. Math. Phys. **193**, 493 (1998).
[10] L. L. Chau, J. C. Shaw, and M. H. Tu, J. Math. Phys. **38**, 4128 (1997).
[11] L. Loris and R. Willox, Inverse Probl. **13**, 849 (1997).
[12] J. C. Shaw and M. H. Tu, J. Math. Phys. **38**, 5756 (1997).
[13] W. Oevel and Z. Popowicz, Commun. Math. Phys. **139**, 441 (1991).
[14] J. M. Figueroa-O'Farrill, J. Mas, and E. Ramos, Rev. Math. Phys. **3**, 479 (1991).
[15] B. Dewitt, *Supermanifolds* (Cambridge U.P., Cambridge, 1992).
[16] Q. P. Liu and M. Mañas, Phys. Lett. B **394**, 337 (1997).

Integrable semi-discretization of the coupled modified KdV equations

Takayuki Tsuchida,[a] Hideaki Ujino, and Miki Wadati
Department of Physics, Graduate School of Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku,
Tokyo 113-0033, Japan

(Received 3 February 1998; accepted for publication 13 May 1998)

The discrete version of the inverse scattering method proposed by Ablowitz and Ladik is extended to solve multi-component systems. The extension enables one to solve the initial value problem, which proves directly the complete integrability of a semi-discrete version of the coupled modified Korteweg–de Vries (KdV) equations and their hierarchy. It also provides a procedure to obtain conservation laws and multi-soliton solutions of the hierarchy. © *1998 American Institute of Physics.* [S0022-2488(98)01909-4]

I. INTRODUCTION

In the theory of solitons, the inverse scattering method (ISM) is known to be one of the most useful tools to analyze various integrable models.[1] The ISM has been widely applied to models with continuous coordinates such as the KdV equation,[2] the modified KdV (mKdV) equation,[3–5] and the nonlinear Schrödinger (NLS) equation.[6] The class of Lax pairs to perform the ISM has been extended from the 2×2 AKNS formulation[7] to 3×3 (Refs. 8–10) and, more generally, $N\times N$ matrix formulations[11,12] so as to solve multi-component integrable models including the three-wave interaction equations[8] and the coupled NLS equations (Manakov system).[9] Recently the authors[13] proposed a new extension of the ISM to solve the coupled modified KdV (cmKdV) equations,

$$\frac{\partial u_i}{\partial t}+6\left(\sum_{j,k=0}^{M-1} C_{jk}u_j u_k\right)\frac{\partial u_i}{\partial x}+\frac{\partial^3 u_i}{\partial x^3}=0, \quad i=0,1,...,M-1, \tag{1.1}$$

which was shown to have an infinite number of local conservation laws by Svinolupov.[14] Iwao and Hirota obtained soliton solutions by the direct method (Hirota's method).[15]

On the other hand, Flaschka,[16] Manakov,[17] and Ablowitz and Ladik[18,19] showed that the discrete version of the ISM formulation is also applicable to discrete integrable models such as the Toda lattice, the Volterra model, the discrete mKdV equation, and the discrete NLS equation. However, the class of discrete models which are integrable by the ISM is not large enough compared with that of continuous models.

In the present paper, we propose a matrix generalization of the Ablowitz–Ladik hierarchy and we solve the semi-discrete coupled modified KdV (sd-cmKdV) equations, or the coupled modified Volterra equations,[20]

$$\frac{\partial u_n^{(i)}}{\partial t}=\left(1+\sum_{j,k=0}^{M-1} C_{jk}u_n^{(j)}u_n^{(k)}\right)(u_{n+1}^{(i)}-u_{n-1}^{(i)}), \quad i=0,1,...,M-1, \tag{1.2}$$

and the higher flows in the hierarchy. This model reduces to the cmKdV equations (1.1) in the continuum limit by some variable transformations. Hisakado[20] revealed the connection between the sd-cmKdV equations and the 2-D Toda lattice. Ohta[21] obtained an N-soliton solution for the model which is equivalent to the sd-cmKdV equations by the Hirota's method. Hirota[22] studied a similar model,

[a]Electronic mail: tsuchida@monet.phys.s.u-tokyo.ac.jp

 © 1998 American Institute of Physics

$$\frac{\partial u_n^{(i)}}{\partial t}=\sum_{j,k=0}^{M-1} C_{jk}u_n^{(j)}u_n^{(k)}\cdot(u_{n+1}^{(i)}-u_{n-1}^{(i)}),\quad i=0,1,...,M-1, \tag{1.3}$$

in the open-end case and obtained the so-called molecule solutions. However, for the first time, in the present paper the initial value problem of the sd-cmKdV equations is solved and the most general soliton solutions are given.

The paper is organized as follows. In Sec. II, we introduce a Lax formulation for the sd-cmKdV equations and prove that the model has an infinite number of conservation laws. In Sec. III, we perform the ISM for the sd-cmKdV equations. As a result, the initial value problem is solved and the *N*-soliton solution is obtained. The last section, Sec. IV, is devoted to the summary and discussions.

II. GENERAL FORMULATION

In this section, we propose a Lax formulation for the sd-cmKdV equations, which is a new extension of the formulation by Ablowitz and Ladik.[1,18,19] The Lax formulation gives us an infinite number of conservation laws.

A. Normalization

We introduce an M-component vector $\mathbf{u}_n(t)$ and a constant $M\times M$ matrix G,

$$\mathbf{u}_n=(u_n^{(0)},u_n^{(1)},\ldots,u_n^{(M-1)})^T,\quad G=(-C_{ij}),$$

where the superscript T means the transposition. In this notation, the sd-cmKdV equations (1.2) are rewritten as

$$\mathbf{u}_{n,t}=(1-\mathbf{u}_n^TG\mathbf{u}_n)(\mathbf{u}_{n+1}-\mathbf{u}_{n-1}). \tag{2.1}$$

In the following, we assume that G is a real symmetric regular matrix and thus may set

$$G=P^T\Lambda P,$$

$$\Lambda=\mathrm{diag}(\lambda_0,\ldots,\lambda_{M-1}),\quad \lambda_j\neq 0\quad \text{for all } j.$$

Here P is a real orthogonal matrix,

$$P^TP=PP^T=I.$$

In terms of a new set of dependent variables $\mathbf{v}_n=(v_n^{(0)},v_n^{(1)},\ldots,v_n^{(M-1)})^T$ defined by

$$\mathbf{v}_n\equiv P\mathbf{u}_n,$$

Eq. (2.1) is cast into

$$\mathbf{v}_{n,t}=(1-\mathbf{v}_n^T\Lambda\mathbf{v}_n)(\mathbf{v}_{n+1}-\mathbf{v}_{n-1}), \tag{2.2}$$

or more explicitly

$$\frac{\partial v_n^{(i)}}{\partial t}=\left(1-\sum_{j=0}^{M-1}\lambda_jv_n^{(j)2}\right)(v_{n+1}^{(i)}-v_{n-1}^{(i)}),\quad \lambda_j\neq 0,\quad i=0,1,...,M-1. \tag{2.3}$$

Changing a scale of v_i by $\sqrt{|\lambda_i|}\cdot v_n^{(i)}\to v_n^{(i)}$, we obtain normalized sd-cmKdV equations,

$$\frac{\partial v_n^{(i)}}{\partial t}=\left(1-\sum_{j=0}^{M-1}\varepsilon_jv_n^{(j)2}\right)(v_{n+1}^{(i)}-v_{n-1}^{(i)}),\quad \varepsilon_j=\mathrm{sgn}(\lambda_j)=\pm 1,\quad i=0,1,...,M-1. \tag{2.4}$$

The above expression of the sd-cmKdV equations is more convenient than Eq. (1.2) to perform the ISM. Thus, we mainly deal with Eq. (2.4) as the sd-cmKdV equations in what follows. Further, the dependent variables $\{v_n^{(i)}\}$ are assumed to be real.

B. Lax pair

We introduce a set of auxiliary linear equations

$$\Psi_{n+1}=L_n\Psi_n, \quad \Psi_{n,t}=M_n\Psi_n. \tag{2.5}$$

Here Ψ_n is a $(p+q)$-component vector and L_n, M_n are $(p+q)\times(p+q)$ matrices. The compatibility condition of Eq. (2.5) is given by

$$L_{n,t}+L_nM_n-M_{n+1}L_n=O. \tag{2.6}$$

We call L_n, M_n the Lax pair and Eq. (2.6) (a semi-discrete version of) the zero-curvature condition or, simply, the Lax equation. We employ the following form of the Lax pair,

$$L_n=z\begin{bmatrix} I_1 & O \\ O & O \end{bmatrix}+\begin{bmatrix} O & Q_n \\ R_n & O \end{bmatrix}+\frac{1}{z}\begin{bmatrix} O & O \\ O & I_2 \end{bmatrix}=\begin{bmatrix} zI_1 & Q_n \\ R_n & \dfrac{1}{z}I_2 \end{bmatrix}, \tag{2.7}$$

$$\begin{aligned} M_n&=z^2\begin{bmatrix} aI_1 & O \\ O & O \end{bmatrix}+z\begin{bmatrix} O & aQ_n \\ aR_{n-1} & O \end{bmatrix}+\begin{bmatrix} -aQ_nR_{n-1}+c_1I_1 & O \\ O & bR_nQ_{n-1}+c_2I_2 \end{bmatrix} \\ &\quad+\frac{1}{z}\begin{bmatrix} O & -bQ_{n-1} \\ -bR_n & O \end{bmatrix}+\frac{1}{z^2}\begin{bmatrix} O & O \\ O & -bI_2 \end{bmatrix} \\ &=\begin{bmatrix} z^2aI_1-aQ_nR_{n-1}+c_1I_1 & zaQ_n-\dfrac{1}{z}bQ_{n-1} \\ zaR_{n-1}-\dfrac{1}{z}bR_n & -\dfrac{1}{z^2}bI_2+bR_nQ_{n-1}+c_2I_2 \end{bmatrix}, \end{aligned} \tag{2.8}$$

where z is the spectral parameter which satisfies $z_t=0$. Here I_1 and I_2 are respectively the $p\times p$ and $q\times q$ unit matrices, Q_n is a $p\times q$ matrix, and R_n is a $q\times p$ matrix.

Substituting Eqs. (2.7) and (2.8) into Eq. (2.6), we obtain a set of matrix equations

$$Q_{n,t}-aQ_{n+1}-bQ_{n-1}+(c_2-c_1)Q_n+aQ_{n+1}R_nQ_n+bQ_nR_nQ_{n-1}=O, \tag{2.9a}$$

$$R_{n,t}+bR_{n+1}+aR_{n-1}+(c_1-c_2)R_n-bR_{n+1}Q_nR_n-aR_nQ_nR_{n-1}=O. \tag{2.9b}$$

Choosing the constants as

$$a=-b=1, \quad c_1=c_2(=0),$$

we obtain

$$Q_{n,t}-(Q_{n+1}-Q_{n-1})+(Q_{n+1}R_nQ_n-Q_nR_nQ_{n-1})=O, \tag{2.10a}$$

$$R_{n,t}-(R_{n+1}-R_{n-1})+(R_{n+1}Q_nR_n-R_nQ_nR_{n-1})=O. \tag{2.10b}$$

We call this model (2.10) the semi-discrete (sd-) matrix mKdV equation.[13] If we set

$$R_n=-I, \quad Q_n=W_n-I, \tag{2.11}$$

we obtain a matrix version of the Lotka–Volterra equation,

$$W_{n,t}=W_{n+1}W_n-W_nW_{n-1}, \tag{2.12}$$

where W_n is a square matrix.

When we take an alternative choice

$$a=b=i, \quad c_2-c_1=2i,$$

we obtain

$$iQ_{n,t}+(Q_{n+1}+Q_{n-1}-2Q_n)-(Q_{n+1}R_nQ_n+Q_nR_nQ_{n-1})=O, \tag{2.13a}$$

$$iR_{n,t}-(R_{n+1}+R_{n-1}-2R_n)+(R_{n+1}Q_nR_n+R_nQ_nR_{n-1})=O. \tag{2.13b}$$

We call this model the semi-discrete (sd-) matrix nonlinear Schrödinger (NLS) equation. The integrable model (2.13) or, more generally, (2.9) was found by D. Trubatch and M. J. Ablowitz, as far as the authors know.

C. Conservation laws

In this subsection, we present a systematic method to construct local conservation laws for the system (2.9) which includes the sd-matrix mKdV equation and the sd-matrix NLS equation as its reductions. We start from a special class, $p=q=l$, of Eq. (2.5),

$$\begin{bmatrix}\Psi_{1n+1}\\ \Psi_{2n+1}\end{bmatrix}=\begin{bmatrix}F_{1n} & Q_n\\ R_n & F_{2n}\end{bmatrix}\begin{bmatrix}\Psi_{1n}\\ \Psi_{2n}\end{bmatrix}, \tag{2.14}$$

$$\begin{bmatrix}\Psi_{1n}\\ \Psi_{2n}\end{bmatrix}_t=\begin{bmatrix}A_n & B_n\\ C_n & D_n\end{bmatrix}\begin{bmatrix}\Psi_{1n}\\ \Psi_{2n}\end{bmatrix}, \tag{2.15}$$

where all the entries in Eqs. (2.14) and (2.15) are assumed to be $l\times l$ square matrices. The following is a discrete version of the method for continuous systems.[13,23] The zero-curvature condition (2.6) gives

$$F_{1n,t}+F_{1n}A_n+Q_nC_n-A_{n+1}F_{1n}-B_{n+1}R_n=O, \tag{2.16a}$$

$$F_{2n,t}+F_{2n}D_n+R_nB_n-D_{n+1}F_{2n}-C_{n+1}Q_n=O, \tag{2.16b}$$

$$Q_{n,t}+F_{1n}B_n+Q_nD_n-A_{n+1}Q_n-B_{n+1}F_{2n}=O, \tag{2.16c}$$

$$R_{n,t}+F_{2n}C_n+R_nA_n-D_{n+1}R_n-C_{n+1}F_{1n}=O. \tag{2.16d}$$

Defining an $l\times l$ square matrix Γ_n by

$$\Gamma_n\equiv\Psi_{2n}\Psi_{1n}^{-1}, \tag{2.17}$$

we can prove the following relations from Eqs. (2.14) and (2.15):

$$\Gamma_{n+1}=(R_n+F_{2n}\Gamma_n)(F_{1n}+Q_n\Gamma_n)^{-1}, \tag{2.18}$$

$$\Gamma_{n,t}=C_n+D_n\Gamma_n-\Gamma_nA_n-\Gamma_nB_n\Gamma_n. \tag{2.19}$$

Equation (2.19) can be interpreted as a matrix version of the Ricatti equation. Using Eqs. (2.16) and (2.18), we can rewrite Eq. (2.19) as

$$\begin{aligned}(Q_n\Gamma_n+F_{1n})_t(Q_n\Gamma_n+F_{1n})^{-1}=&A_{n+1}-(Q_n\Gamma_n+F_{1n})A_n(Q_n\Gamma_n+F_{1n})^{-1}\\ &+B_{n+1}\Gamma_{n+1}-(Q_n\Gamma_n+F_{1n})B_n\Gamma_n(Q_n\Gamma_n+F_{1n})^{-1}.\end{aligned} \tag{2.20}$$

Taking the trace of Eq. (2.20), we get

$$\mathrm{tr}\{\log(Q_n\Gamma_n+F_{1n})\}_t=\mathrm{tr}(A_{n+1}+B_{n+1}\Gamma_{n+1})-\mathrm{tr}(A_n+B_n\Gamma_n). \tag{2.21}$$

Assuming that L_n is expressed as Eq. (2.7), we have

$$F_{1n}=zI, \quad F_{2n}=\frac{1}{z}\,I. \tag{2.22}$$

The matrices Q_n and R_n are square matrices in this case. Then Eqs. (2.21) and (2.18) are cast into

$$\left\{\mathrm{tr}\,\log\left(I+\frac{1}{z}\,Q_n\Gamma_n\right)\right\}_t=\mathrm{tr}(A_{n+1}+B_{n+1}\Gamma_{n+1})-\mathrm{tr}(A_n+B_n\Gamma_n), \tag{2.23}$$

$$zQ_n\Gamma_n=Q_nR_{n-1}+\frac{1}{z}\,Q_nQ_{n-1}^{-1}(Q_{n-1}\Gamma_{n-1})-(Q_n\Gamma_n)(Q_{n-1}\Gamma_{n-1}). \tag{2.24}$$

Equation (2.23) has the form of the local conservation law. It shows that $\mathrm{tr}\{\log(I+Q_n\Gamma_n/z)\}$ is a generating function of the conserved densities for Eq. (2.9). We expand $Q_n\Gamma_n$ with respect to the inverse of the spectral parameter z in the following form:

$$Q_n\Gamma_n=\sum_{j=1}^{\infty}\frac{1}{z^{2j-1}}\,f_n^{(j)}. \tag{2.25}$$

Substituting Eq. (2.25) into Eq. (2.24), we obtain a recursion formula for $f_n^{(j)}$:

$$f_n^{(j)}=Q_nR_{n-1}\delta_{j,1}+Q_nQ_{n-1}^{-1}f_{n-1}^{(j-1)}-\sum_{k=1}^{j-1}f_n^{(k)}f_{n-1}^{(j-k)}, \quad j=1,2,\ldots. \tag{2.26}$$

The formula (2.26) yields $f_n^{(j)}$; for instance,

$$f_n^{(1)}=Q_nR_{n-1},$$

$$f_n^{(2)}=Q_nR_{n-2}-Q_nR_{n-1}Q_{n-1}R_{n-2},$$

$$\begin{aligned}f_n^{(3)}=&Q_nR_{n-3}-Q_nR_{n-2}Q_{n-2}R_{n-3}-Q_nR_{n-1}Q_{n-1}R_{n-3}-Q_nR_{n-2}Q_{n-1}R_{n-2}\\&+Q_nR_{n-1}Q_{n-1}R_{n-2}Q_{n-2}R_{n-3}+Q_nR_{n-1}Q_{n-1}R_{n-2}Q_{n-1}R_{n-2}.\end{aligned}$$

We put Eq. (2.25) into $\mathrm{tr}\{\log(I+Q_n\Gamma_n/z)\}$ and expand it with respect to $1/z$:

$$\mathrm{tr}\left\{\log\left(I+\frac{1}{z^2}\,f_n^{(1)}+\frac{1}{z^4}\,f_n^{(2)}+\frac{1}{z^6}\,f_n^{(3)}+\cdots\right)\right\}=\mathrm{tr}\left\{\frac{1}{z^2}\,f_n^{(1)}+\frac{1}{z^4}\left(f_n^{(2)}-\frac{1}{2}\,f_n^{(1)2}\right)+\cdots\right\}.$$

Thus, the first two conserved densities given by this expansion are

$$J_n^{(-1)}=f_n^{(1)}=Q_nR_{n-1}, \tag{2.27a}$$

$$J_n^{(-2)}=\mathrm{tr}\{f_n^{(2)}-\tfrac{1}{2}\,(f_n^{(1)})^2\}=\mathrm{tr}\{Q_nR_{n-2}-Q_nR_{n-1}Q_{n-1}R_{n-2}-\tfrac{1}{2}\,(Q_nR_{n-1})^2\}. \tag{2.27b}$$

Next we expand $Q_n\Gamma_n$ with respect to z. For this purpose, we rewrite Eq. (2.24) as

$$Q_n\Gamma_n=-zQ_nR_n+zQ_nQ_{n+1}^{-1}(Q_{n+1}\Gamma_{n+1})(Q_n\Gamma_n+zI). \tag{2.28}$$

We substitute the following expansion,

$$Q_n\Gamma_n=\sum_{j=1}^{\infty}z^{2j-1}g_n^{(j)}, \tag{2.29}$$

into Eq. (2.28) and obtain a recursion formula,

$$g_n^{(j)} = -Q_n R_n \delta_{j,1} + Q_n Q_{n+1}^{-1} g_{n+1}^{(j-1)} + Q_n Q_{n+1}^{-1} \sum_{k=1}^{j-1} g_{n+1}^{(k)} g_n^{(j-k)}. \tag{2.30}$$

By use of the formula (2.30), the first three of the coefficients $g_n^{(j)}$ are given by

$$g_n^{(1)} = -Q_n R_n,$$

$$g_n^{(2)} = -Q_n R_{n+1}(I - Q_n R_n),$$

$$g_n^{(3)} = -Q_n R_{n+2}(I - Q_n R_n) + Q_n R_{n+1} Q_n R_{n+1}(I - Q_n R_n) + Q_n R_{n+2} Q_{n+1} R_{n+1}(I - Q_n R_n).$$

We put Eq. (2.29) into $\mathrm{tr}\{\log(I + Q_n \Gamma_n / z)\}$ and expand it with respect to z. Then we obtain

$$\begin{aligned}\mathrm{tr}\,\{\log(I + g_n^{(1)} + z^2 g_n^{(2)} + z^4 g_n^{(3)} + \cdots)\} = \mathrm{tr}\{&\log(I + g_n^{(1)}) + z^2 g_n^{(2)}(I + g_n^{(1)})^{-1} \\ &+ z^4[g_n^{(3)}(I + g_n^{(1)})^{-1} - \tfrac{1}{2}\{g_n^{(2)}(I + g_n^{(1)})^{-1}\}^2] + \cdots\}.\end{aligned}$$

Thus, the first three conserved densities given by this expansion are

$$J_n^{(0)} = \mathrm{tr}\{\log(I + g_n^{(1)})\} = \mathrm{tr}\{\log(I - Q_n R_n)\}, \tag{2.31a}$$

$$J_n^{(1)} = Q_n^{-1} g_n^{(2)}(I + g_n^{(1)})^{-1} Q_n = -R_{n+1} Q_n, \tag{2.31b}$$

$$\begin{aligned}J_n^{(2)} &= \mathrm{tr}[g_n^{(3)}(I + g_n^{(1)})^{-1} - \tfrac{1}{2}\{g_n^{(2)}(I + g_n^{(1)})^{-1}\}^2] \\ &= \mathrm{tr}\{-Q_n R_{n+2} + Q_n R_{n+2} Q_{n+1} R_{n+1} + \tfrac{1}{2}(Q_n R_{n+1})^2\}.\end{aligned} \tag{2.31c}$$

We should pay attention to the fact that all the entries of matrices $J_n^{(-1)}$ and $J_n^{(1)}$ are conserved densities for Eq. (2.9). We can prove it by a straightforward calculation. That is the reason why we dropped tr in Eqs. (2.27a) and (2.31b).

D. Reduction of the Lax pair for the semi-discrete coupled modified KdV equations

In this subsection, we show a reduction of the sd-matrix mKdV equation to the sd-cmKdV equations. We recursively define $2^{m-1} \times 2^{m-1}$ matrices $Q_n^{(m)}$ and $R_n^{(m)}$ by

$$Q_n^{(1)} = \mu_0 v_n^{(0)} + i v_n^{(1)}, \quad R_n^{(1)} = \varepsilon_1(\mu_0 v_n^{(0)} - i v_n^{(1)}), \tag{2.32}$$

$$Q_n^{(m+1)} = \begin{bmatrix} Q_n^{(m)} & -\varepsilon_{2m+1}(\mu_{2m} v_n^{(2m)} + i v_n^{(2m+1)}) I_{2^{m-1}} \\ -(\mu_{2m} v_n^{(2m)} - i v_n^{(2m+1)}) I_{2^{m-1}} & -R_n^{(m)} \end{bmatrix}, \tag{2.33}$$

$$R_n^{(m+1)} = \begin{bmatrix} R_n^{(m)} & -\varepsilon_{2m+1}(\mu_{2m} v_n^{(2m)} + i v_n^{(2m+1)}) I_{2^{m-1}} \\ -(\mu_{2m} v_n^{(2m)} - i v_n^{(2m+1)}) I_{2^{m-1}} & -Q_n^{(m)} \end{bmatrix}. \tag{2.34}$$

Here $I_{2^{m-1}}$ is the $2^{m-1} \times 2^{m-1}$ unit matrix, each ε_i is a constant which is either 1 or -1, and μ_{2m} is a constant that satisfies

$$\mu_{2m}^2 = \frac{\varepsilon_{2m}}{\varepsilon_{2m+1}} = \varepsilon_{2m} \varepsilon_{2m+1}.$$

For $Q_n^{(m)}$ and $R_n^{(m)}$ defined by Eqs. (2.32)–(2.34), we can prove a simple relation,

$$Q_n^{(m)} R_n^{(m)} = R_n^{(m)} Q_n^{(m)} = \sum_{j=0}^{2m-1} \varepsilon_j v_n^{(j)2} \cdot I_{2^{m-1}}, \tag{2.35}$$

by induction. Then substituting $Q_n^{(m)}$ and $R_n^{(m)}$ for Q_n and R_n in the sd-matrix mKdV equation (2.10), we obtain

$$v_{n,t}^{(i)}-(v_{n+1}^{(i)}-v_{n-1}^{(i)})+\left(\sum_{j=0}^{M-1}\varepsilon_j v_n^{(j)2}\right)(v_{n+1}^{(i)}-v_{n-1}^{(i)})=0,\quad \varepsilon_j=\pm 1,\quad i=0,1,...,M-1, \tag{2.36}$$

where we set $M=2m$. Setting

$$\mu_{2j}=1,\quad j=0,1,...,m-1,$$

and

$$q_n^{(j)}=v_n^{(2j)}+iv_n^{(2j+1)},$$

we get an alternative representation of the sd-cmKdV equations in terms of complex variables,

$$q_{n,t}^{(i)}-(q_{n+1}^{(i)}-q_{n-1}^{(i)})+\left(\sum_{j=0}^{m-1}\varepsilon_{2j}|q_n^{(j)}|^2\right)(q_{n+1}^{(i)}-q_{n-1}^{(i)})=0,\quad i=0,1,...,m-1. \tag{2.37}$$

In the following, we choose

$$\varepsilon_i=-1,\quad i=0,1,...,2m-1,$$

$$\mu_{2j}=1,\quad j=0,1,...,m-1,$$

and consider a self-focusing type of the sd-cmKdV equations (2.4), that is,

$$\frac{\partial v_n^{(i)}}{\partial t}=\left(1+\sum_{j=0}^{M-1}v_n^{(j)2}\right)(v_{n+1}^{(i)}-v_{n-1}^{(i)}),\quad i=0,1,...,M-1. \tag{2.38}$$

In this case, the recursion relations of $Q_n^{(m)}$ and $R_n^{(m)}$ are given by

$$Q_n^{(1)}=v_n^{(0)}+iv_n^{(1)},\quad R_n^{(1)}=-v_n^{(0)}+iv_n^{(1)}, \tag{2.39}$$

$$Q_n^{(m+1)}=\begin{bmatrix} Q_n^{(m)} & (v_n^{(2m)}+iv_n^{(2m+1)})I_{2^{m-1}} \\ -(v_n^{(2m)}-iv_n^{(2m+1)})I_{2^{m-1}} & -R_n^{(m)} \end{bmatrix}, \tag{2.40}$$

$$R_n^{(m+1)}=\begin{bmatrix} R_n^{(m)} & (v_n^{(2m)}+iv_n^{(2m+1)})I_{2^{m-1}} \\ -(v_n^{(2m)}-iv_n^{(2m+1)})I_{2^{m-1}} & -Q_n^{(m)} \end{bmatrix}. \tag{2.41}$$

For example, the Lax matrix L_n for the four-component sd-cmKdV equations (2.38) is given by

$$L_n=\begin{bmatrix} z & 0 & v_n^{(0)}+iv_n^{(1)} & v_n^{(2)}+iv_n^{(3)} \\ 0 & z & -v_n^{(2)}+iv_n^{(3)} & v_n^{(0)}-iv_n^{(1)} \\ -v_n^{(0)}+iv_n^{(1)} & v_n^{(2)}+iv_n^{(3)} & \frac{1}{z} & 0 \\ -v_n^{(2)}+iv_n^{(3)} & -v_n^{(0)}-iv_n^{(1)} & 0 & \frac{1}{z} \end{bmatrix}. \tag{2.42}$$

The relation (2.35) for the above $Q_n^{(m)}$ and $R_n^{(m)}$ is

$$Q_n^{(m)}R_n^{(m)}=R_n^{(m)}Q_n^{(m)}=-\sum_{j=0}^{2m-1}v_n^{(j)2}\cdot I_{2^{m-1}}. \tag{2.43}$$

This relation plays a crucial role in performing the ISM. In addition, a simple relation between $Q_n^{(m)}$ and $R_n^{(m)}$ holds,

$$R_n = -Q_n^{\dagger}, \tag{2.44}$$

where the symbol † denotes the Hermitian conjugate.

Because the sd-cmKdV equations are given as a reduction of the sd-matrix mKdV equation, the results in Sec. II C assure that the sd-cmKdV equations have an infinite number of local conservation laws. Explicit forms of the first four conserved densities for the original sd-cmKdV equations (1.2) are given by

$$\log\left(1+\sum_{j,k} C_{jk}u_n^{(j)}u_n^{(k)}\right), \tag{2.45}$$

$$\begin{gathered} u_{n+1}^{(j)}u_n^{(k)}-u_n^{(j)}u_{n+1}^{(k)} \quad \text{for all } j,k(j\neq k), \\ \sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}, \end{gathered} \tag{2.46}$$

$$\begin{aligned} &\left(1+\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}\right)\cdot\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_n^{(k)}+\left(\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}\right)^2 \\ &\quad -\frac{1}{2}\sum_{j,k} C_{jk}u_n^{(j)}u_n^{(k)}\cdot\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}, \end{aligned} \tag{2.47}$$

$$\begin{aligned} &\left(1+\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_{n+2}^{(k)}\right)\left(1+\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}\right)\cdot\sum_{j,k} C_{jk}u_{n+3}^{(j)}u_n^{(k)} \\ &\quad +2\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_n^{(k)}\cdot\left(1+\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}\right)\left(\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}+\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_{n+1}^{(k)}\right) \\ &\quad -\left(\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}\cdot\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_{n+2}^{(k)}+\sum_{j,k} C_{jk}u_{n+2}^{(j)}u_{n+1}^{(k)}\cdot\sum_{j,k} C_{jk}u_n^{(j)}u_n^{(k)}\right) \\ &\quad \cdot\left(1+\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}\right)+\frac{4}{3}\left(\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}\right)^3 \\ &\quad -\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_n^{(k)}\cdot\sum_{j,k} C_{jk}u_n^{(j)}u_n^{(k)}\cdot\sum_{j,k} C_{jk}u_{n+1}^{(j)}u_{n+1}^{(k)}. \end{aligned} \tag{2.48}$$

III. INVERSE SCATTERING METHOD

In this section we consider the scattering and inverse scattering problems associated with the $2l\times 2l$ ($l=2^{m-1}$) matrix (2.7),

$$\Psi_{n+1}=L_n\Psi_n, \quad L_n=\begin{bmatrix} zI & Q_n \\ R_n & \dfrac{1}{z}I \end{bmatrix}, \tag{3.1}$$

under the constraints (2.43) and (2.44),

$$R_n=-Q_n^{\dagger}, \quad Q_nR_n=-\sum_{j=0}^{2m-1} v_n^{(j)2}\cdot I\equiv-\sigma_nI. \tag{3.2}$$

The rapidly decreasing boundary conditions are assumed,

$$Q_n, R_n\to O \quad \text{as} \quad n\to\pm\infty. \tag{3.3}$$

Considering time dependence of the scattering data, we can solve the initial value problem of the sd-cmKdV equations (2.38). Some of the main ideas in what follows are an extension and a modification of the analyses in Refs. 13, 24, and 25 for the matrix KdV equation and the matrix mKdV equation in the continuous theory. The notation used in this section follows the one in Ref. 1.

A. Scattering problem

Let $\Psi_n(z)$ and $\Phi_n(z)$ be solutions of Eq. (3.1) composed by $2l(=2^m)$ rows and $l(=2^{m-1})$ columns. We can show that

$$\Psi_{n+1}^{\dagger}\left(\frac{1}{z^*}\right)\Phi_{n+1}(z)=(I-Q_nR_n)\Psi_n^{\dagger}\left(\frac{1}{z^*}\right)\Phi_n(z)=\rho_n\Psi_n^{\dagger}\left(\frac{1}{z^*}\right)\Phi_n(z).$$

Here $*$ denotes the complex conjugate and ρ_n is defined by

$$\rho_n\equiv 1+\sigma_n=1+\sum_{j=0}^{2m-1} v_n^{(j)2}. \tag{3.4}$$

We employ the following definition of the matrix Wronskian $W_n[\Psi,\ \Phi]$,

$$W_n[\Psi,\ \Phi]\equiv\Psi_n^{\dagger}\left(\frac{1}{z^*}\right)\Phi_n(z), \tag{3.5}$$

which satisfies

$$W_{n+1}[\Psi,\ \Phi]=(I-Q_nR_n)W_n[\Psi,\ \Phi]=\rho_nW_n[\Psi,\ \Phi]. \tag{3.6}$$

Using Eq. (3.6) repeatedly, we get

$$W_{\infty}[\Psi,\ \Phi]=\tau W_{-\infty}[\Psi,\ \Phi], \tag{3.7}$$

where τ is defined by

$$\tau\equiv\prod_{n=-\infty}^{\infty}\rho_n=\prod_{n=-\infty}^{\infty}\left(1+\sum_{j=0}^{2m-1} v_n^{(j)2}\right), \tag{3.8}$$

and assumed to be finite. It should be noticed that τ is a conserved quantity [see Eq. (2.45)].

We introduce Jost functions $\phi_n,\bar{\phi}_n$ and $\psi_n,\bar{\psi}_n$ which satisfy the boundary conditions,

$$\phi_n\sim\begin{bmatrix} I \\ O \end{bmatrix}z^n \qquad \text{as} \quad n\to-\infty, \tag{3.9a}$$

$$\bar{\phi}_n\sim\begin{bmatrix} O \\ -I \end{bmatrix}z^{-n} \quad \text{as} \quad n\to-\infty, \tag{3.9b}$$

and

$$\psi_n\sim\begin{bmatrix} O \\ I \end{bmatrix}z^{-n} \qquad \text{as} \quad n\to+\infty, \tag{3.9c}$$

$$\bar{\psi}_n\sim\begin{bmatrix} I \\ O \end{bmatrix}z^{n} \qquad \text{as} \quad n\to+\infty. \tag{3.9d}$$

Here O and I respectively stand for the $l\times l$ zero matrix and the $l\times l$ unit matrix. It can be shown that ϕ_nz^{-n},ψ_nz^n are analytic outside the unit circle ($|z|>1$) on the complex z plane, and that

$\overline{\phi}_n z^n, \overline{\psi}_n z^{-n}$ are analytic inside the unit circle ($|z|<1$) on the z plane, when Q_n, R_n go to O sufficiently rapidly as $n \to \pm\infty$. We assume the following summation representation of the Jost functions ψ_n and $\overline{\psi}_n$:

$$\psi_n = \sum_{n'=n}^{\infty} z^{-n'} K(n,n'), \tag{3.10a}$$

$$\overline{\psi}_n = \sum_{n'=n}^{\infty} z^{n'} \overline{K}(n,n'), \tag{3.10b}$$

where $K(n,n')$ and $\overline{K}(n,n')$ are column vectors which consist of two $l \times l$ square matrices,

$$K(n,n') = \begin{bmatrix} K_1(n,n') \\ K_2(n,n') \end{bmatrix}, \quad \overline{K}(n,n') = \begin{bmatrix} \overline{K_1}(n,n') \\ \overline{K_2}(n,n') \end{bmatrix}.$$

We substitute Eq. (3.10) into Eq. (3.1). Equating the terms with the same power of z, we get the relations for K_1 and K_2,

$$K_1(n,n) = O, \tag{3.11a}$$

$$Q_n K_2(n,n) = -K_1(n,n+1), \tag{3.11b}$$

$$K_1(n,n+j+1) = K_1(n+1,n+j) - Q_n K_2(n,n+j) \quad (j \geqslant 1), \tag{3.11c}$$

$$R_n K_1(n,n+j) = K_2(n+1,n+j) - K_2(n,n+j-1) \quad (j \geqslant 1), \tag{3.11d}$$

and for $\overline{K}_1$ and $\overline{K}_2$,

$$\overline{K}_2(n,n) = O, \tag{3.12a}$$

$$R_n \overline{K}_1(n,n) = -\overline{K}_2(n,n+1), \tag{3.12b}$$

$$Q_n \overline{K}_2(n,n+j) = \overline{K}_1(n+1,n+j) - \overline{K}_1(n,n+j-1) \quad (j \geqslant 1), \tag{3.12c}$$

$$\overline{K}_2(n,n+j+1) = \overline{K}_2(n+1,n+j) - R_n \overline{K}_1(n,n+j) \quad (j \geqslant 1). \tag{3.12d}$$

Because a pair of the Jost functions ϕ_n and $\overline{\phi}_n$, or ψ_n and $\overline{\psi}_n$, forms a fundamental system of the solutions of the scattering problem (3.1), we can set

$$\phi_n(z) = \overline{\psi}_n(z) A(z) + \psi_n(z) B(z), \tag{3.13a}$$

$$\overline{\phi}_n(z) = \overline{\psi}_n(z) \overline{B}(z) - \psi_n(z) \overline{A}(z). \tag{3.13b}$$

Here the coefficients $\{A(z), \overline{A}(z), B(z), \overline{B}(z)\}$ are n-independent $l \times l$ matrices which are called scattering data.

To derive the formula of the ISM rigorously and concisely, we assume that Q_n and R_n are on compact support. The result is, however, applicable to larger classes of potentials Q_n and R_n. According to the asymptotic behaviors of the Jost functions (3.9a)–(3.9d), we obtain

$$W_{-\infty}[\phi,\ \phi] = W_{-\infty}[\overline{\phi},\ \overline{\phi}] = W_{\infty}[\psi,\ \psi] = W_{\infty}[\overline{\psi},\ \overline{\psi}] = I, \tag{3.14a}$$

$$W_n[\phi,\ \overline{\phi}] = W_n[\psi,\ \overline{\psi}] = O, \tag{3.14b}$$

$$A(z) = W_{\infty}[\overline{\psi},\ \phi], \tag{3.14c}$$

$$\overline{A}(z) = -W_{\infty}[\psi,\ \overline{\phi}], \tag{3.14d}$$

$$B(z)=W_\infty[\psi,\ \phi], \tag{3.14e}$$

$$\overline{B}(z)=W_\infty[\overline{\psi},\ \overline{\phi}]. \tag{3.14f}$$

The expressions (3.14c) and (3.14d) show that $A(z)$ and $\overline{A}(z)$ are respectively analytic outside the unit circle ($|z|>1$) and inside the unit circle ($|z|<1$). Using the relations (3.14), we obtain the following relations among $A(z)$, $\overline{A}(z)$, $B(z)$, and $\overline{B}(z)$:

$$A^\dagger\left(\frac{1}{z^*}\right)A(z)+B^\dagger\left(\frac{1}{z^*}\right)B(z)=\tau\, I, \tag{3.15a}$$

$$\overline{A}^\dagger\left(\frac{1}{z^*}\right)\overline{A}(z)+\overline{B}^\dagger\left(\frac{1}{z^*}\right)\overline{B}(z)=\tau\, I, \tag{3.15b}$$

$$A^\dagger\left(\frac{1}{z^*}\right)\overline{B}(z)-B^\dagger\left(\frac{1}{z^*}\right)\overline{A}(z)=O. \tag{3.15c}$$

These relations are written in a matrix form as

$$\begin{bmatrix} A^\dagger(1/z^*) & B^\dagger(1/z^*) \\ \overline{B}^\dagger(1/z^*) & -\overline{A}^\dagger(1/z^*) \end{bmatrix}\begin{bmatrix} A(z) & \overline{B}(z) \\ B(z) & -\overline{A}(z) \end{bmatrix}=\tau\begin{bmatrix} I & O \\ O & I \end{bmatrix},$$

which leads to the inversion of Eq. (3.13),

$$\tau\overline{\psi}_n(z)=\phi_n(z)A^\dagger\left(\frac{1}{z^*}\right)+\overline{\phi}_n(z)\overline{B}^\dagger\left(\frac{1}{z^*}\right), \tag{3.16a}$$

$$\tau\psi_n(z)=\phi_n(z)B^\dagger\left(\frac{1}{z^*}\right)-\overline{\phi}_n(z)\overline{A}^\dagger\left(\frac{1}{z^*}\right). \tag{3.16b}$$

Equation (3.16a) is used in Appendix A.

B. Gel'fand–Levitan–Marchenko equations

Multiplying $A(z)^{-1}$ and $\overline{A}(z)^{-1}$ from the right to Eqs. (3.13a) and (3.13b), respectively, we obtain

$$\phi_n(z)A(z)^{-1}=\overline{\psi}_n(z)+\psi_n(z)B(z)A(z)^{-1}, \tag{3.17a}$$

$$\overline{\phi}_n(z)\overline{A}(z)^{-1}=-\psi_n(z)+\overline{\psi}_n(z)\overline{B}(z)\overline{A}(z)^{-1}. \tag{3.17b}$$

We substitute Eq. (3.10) into the right side of Eq. (3.17a) and operate

$$\frac{1}{2\pi i}\oint_C dz\ z^{-m-1}\quad (m\geqslant n),$$

where C denotes a contour along the unit circle $|z|=1$. Then we obtain

$$J=\overline{K}(n,m)+\sum_{n'=n}^{\infty}K(n,n')F_C(n'+m),$$

where F_C and J are defined by

$$F_C(n'+m)\equiv\frac{1}{2\pi i}\oint_C B(z)A(z)^{-1}z^{-(n'+m)-1}\,dz,$$

$$J \equiv \frac{1}{2\pi i} \oint_C \phi_n z^{-n} A(z)^{-1} z^{n-m-1}\, dz. \tag{3.18}$$

We notice that $\phi_n z^{-n}$ and $A(z)$ are analytic outside the unit circle C, $|z|>1$. The inverse of $A(z)$, i.e., $A(z)^{-1}$, is given by

$$A(z)^{-1} = \frac{1}{\det A(z)} \tilde{A}(z),$$

where $\tilde{A}$ is the cofactor matrix of A. Thus the singularities of the integrand of Eq. (3.18) in $|z|>1$ come from the zeros of $\det A(z)$. We assume that $1/\det A(z)$ has $2N$ isolated simple poles $\{z_1, z_2, \ldots, z_{2N}\}$ and is regular on the unit circle C [see Eq. (3.35) for the reason why we set the number of poles $2N$]. We set

$$J_{\infty,n} = \lim_{z\to\infty} \phi_n z^{-n} A(z)^{-1},$$

and use the residue theorem. Then the integral J is computed as

$$J = -\sum_{j=1}^{2N} \psi_n(z_j) C_j z_j^{-m-1} + J_{\infty,n}\delta_{n,m} = -\sum_{j=1}^{2N} \sum_{n'=n}^{\infty} K(n,n') C_j z_j^{-(n'+m)-1} + J_{\infty,n}\delta_{n,m},$$

where C_j is the residue matrix of $B(z)A(z)^{-1}$ at $z=z_j$. Defining

$$F_D(n'+m) \equiv \sum_{j=1}^{2N} C_j z_j^{-(n'+m)-1},$$

we arrive at the discrete version of the Gel'fand–Levitan–Marchenko equation,

$$\overline{K}(n,m) + \sum_{n'=n}^{\infty} K(n,n') F(n'+m) = J_{\infty,n}\delta_{n,m} \quad (m \geqslant n). \tag{3.19}$$

Here $F(n'+m)$ is defined by

$$\begin{aligned} F(n'+m) &\equiv F_C(n'+m) + F_D(n'+m) \\ &= \frac{1}{2\pi i} \oint_C B(z)A(z)^{-1} z^{-(n'+m)-1} dz + \sum_{j=1}^{2N} C_j z_j^{-(n'+m)-1}. \end{aligned} \tag{3.20}$$

Similarly, we operate

$$\frac{1}{2\pi i} \oint_C dz\, z^{m-1} \quad (m \geqslant n)$$

on Eq. (3.17b). Then we obtain

$$\overline{J} = -K(n,m) + \sum_{n'=n}^{\infty} \overline{K}(n,n') \overline{F}_C(n'+m),$$

where

$$\overline{F}_C(n'+m) = \frac{1}{2\pi i} \oint_C \overline{B}(z)\overline{A}(z)^{-1} z^{n'+m-1}\, dz,$$

$$\overline{J} = \frac{1}{2\pi i} \oint_C \overline{\phi}_n z^n \overline{A}(z)^{-1} z^{-n+m-1}\, dz.$$

We notice that $\overline{\phi}_n z^n$ and $\overline{A}(z)$ are analytic inside the unit circle C, $|z|<1$. We assume that $1/\det \overline{A}(z)$ has $2N$ isolated simple poles $\{\overline{z}_1,\overline{z}_2,\ldots,\overline{z}_{2N}\}$ and is regular on the unit circle C. We set

$$\overline{J}_{0,n}=\lim_{z\to 0}\ \overline{\phi}_n z^n \overline{A}(z)^{-1},$$

and use the residue theorem. The integral $\overline{J}$ is given by

$$\overline{J}=\sum_{k=1}^{2\overline{N}} z^{-(n'+m)-1}\ \overline{\psi}_n(\overline{z}_k)\overline{C}_k\overline{z}_k^{m-1}+\overline{J}_{0,n}\delta_{n,m}=\sum_{k=1}^{2\overline{N}}\sum_{n'=n}^{\infty}\overline{K}(n,n')\overline{C}_k\overline{z}_k^{n'+m-1}+\overline{J}_{0,n}\delta_{n,m},$$

where $\overline{C}_k$ is the residue matrix of $\overline{B}(z)\overline{A}(z)^{-1}$ at $z=\overline{z}_k$. Defining

$$\overline{F}_D(n'+m)\equiv-\sum_{k=1}^{2\overline{N}}\ \overline{C}_k\overline{z}_k^{n'+m-1},$$

we obtain the counterpart of the discrete Gel'fand–Levitan–Marchenko equation,

$$K(n,m)-\sum_{n'=n}^{\infty}\ \overline{K}(n,n')\overline{F}(n'+m)=-\overline{J}_{0,n}\delta_{n,m}\quad(m\geqslant n).\tag{3.21}$$

Here $\overline{F}(n'+m)$ is defined by

$$\begin{aligned}\overline{F}(n'+m)&\equiv\overline{F}_C(n'+m)+\overline{F}_D(n'+m)\\&=\frac{1}{2\pi i}\ \oint_C\overline{B}(z)\overline{A}(z)^{-1}z^{n'+m-1}\,dz-\sum_{k=1}^{2\overline{N}}\ \overline{C}_k\overline{z}_k^{n'+m-1}.\end{aligned}\tag{3.22}$$

From Eqs. (3.11b) and (3.11d), we obtain

$$\begin{aligned}K_2(n,n)&=(I-R_nQ_n)^{-1}K_2(n+1,n+1)\\&=(I-R_nQ_n)^{-1}(I-R_{n+1}Q_{n+1})^{-1}K_2(n+2,\ n+2)\\&\vdots\\&=\prod_{i=n}^{\infty}\ (I-R_iQ_i)^{-1}=\prod_{i=n}^{\infty}\ \rho_i^{-1}\cdot I,\end{aligned}$$

where $\lim_{n\to\infty} K_2(n,n)=I$ is used. Similarly, from Eqs. (3.12b) and (3.12c) with $\lim_{n\to\infty}\overline{K}_1(n,n)=I$, we obtain

$$\begin{aligned}\overline{K}_1(n,n)&=(I-Q_nR_n)^{-1}\overline{K}_1(n+1,\ n+1)\\&=(I-Q_nR_n)^{-1}(I-Q_{n+1}R_{n+1})^{-1}\overline{K}_1(n+2,\ n+2)\\&\vdots\\&=\prod_{i=n}^{\infty}\ (I-Q_iR_i)^{-1}=\prod_{i=n}^{\infty}\ \rho_i^{-1}\cdot I.\end{aligned}$$

Thus, it is natural to set

$$K(n,m)=\kappa(n,m)\prod_{i=n}^{\infty}\ (I-R_iQ_i)^{-1}=\kappa(n,m)\prod_{i=n}^{\infty}\ \rho_i^{-1}\quad(m\geqslant n),\tag{3.23}$$

$$\overline{K}(n,m)=\overline{\kappa}(n,m)\prod_{i=n}^{\infty}\ (I-Q_iR_i)^{-1}=\overline{\kappa}(n,m)\prod_{i=n}^{\infty}\ \rho_i^{-1}\quad(m\geqslant n).\tag{3.24}$$

Here $\kappa(n,m)$ and $\overline{\kappa}(n,m)$ are column vectors whose elements are $l\times l$ square matrices:

$$\kappa(n,m)=\begin{bmatrix}\kappa_1(n,m)\\ \kappa_2(n,m)\end{bmatrix},\quad \overline{\kappa}(n,m)=\begin{bmatrix}\overline{\kappa}_1(n,m)\\ \overline{\kappa}_2(n,m)\end{bmatrix}.$$

In particular, $\kappa(n,n)$ and $\overline{\kappa}(n,n)$ are given by

$$\kappa(n,n)=\begin{bmatrix} O \\ I \end{bmatrix}, \tag{3.25}$$

$$\overline{\kappa}(n,n)=\begin{bmatrix} I \\ O \end{bmatrix}. \tag{3.26}$$

Putting Eq. (3.23) with (3.25) and Eq. (3.24) with (3.26) into Eq. (3.11) and Eq. (3.12), respectively, we obtain the relations for κ_1, κ_2, $\overline{\kappa}_1$, and $\overline{\kappa}_2$,

$$-\kappa_1(n,n+1)=Q_n, \tag{3.27}$$

$$-\overline{\kappa}_2(n,n+1)=R_n, \tag{3.28}$$

$$\kappa_1(n,n+j+1)=\rho_n\kappa_1(n+1,n+j)-Q_n\kappa_2(n,n+j) \quad (j\geqslant 1), \tag{3.29}$$

$$R_n\kappa_1(n,n+j)=\rho_n\kappa_2(n+1,n+j)-\kappa_2(n,n+j-1) \quad (j\geqslant 1), \tag{3.30}$$

$$Q_n\overline{\kappa}_2(n,n+j)=\rho_n\overline{\kappa}_1(n+1,n+j)-\overline{\kappa}_1(n,n+j-1) \quad (j\geqslant 1), \tag{3.31}$$

$$\overline{\kappa}_2(n,n+j+1)=\rho_n\overline{\kappa}_2(n+1,n+j)-R_n\overline{\kappa}_1(n,n+j) \quad (j\geqslant 1). \tag{3.32}$$

In terms of κ and $\overline{\kappa}$, the Gel'fand–Levitan–Marchenko equations (3.19) and (3.21) for $m>n$ are expressed as

$$\overline{\kappa}(n,m)+\begin{bmatrix} O \\ I \end{bmatrix}F(n+m)+\sum_{n'=n+1}^{\infty}\kappa(n,n')F(n'+m)=\begin{bmatrix} O \\ O \end{bmatrix} \quad (m>n), \tag{3.33}$$

$$\kappa(n,m)-\begin{bmatrix} I \\ O \end{bmatrix}\overline{F}(n+m)-\sum_{n'=n+1}^{\infty}\overline{\kappa}(n,n')\overline{F}(n'+m)=\begin{bmatrix} O \\ O \end{bmatrix} \quad (m>n). \tag{3.34}$$

It should be noted that the scattering problem (3.1) gives the symmetry properties of the scattering data. Iterating Eq. (3.1), we can prove that $A(z),\overline{A}(z)$ are polynomials in z of even degree and $B(z),\overline{B}(z)$ are polynomials in z of odd degree. This fact leads to

$$\det A(z)=\det A(-z), \tag{3.35}$$

$$\det \overline{A}(z)=\det \overline{A}(-z), \tag{3.36}$$

which mean the eigenvalues $z_j,\overline{z_k}$ should appear as positive–negative pairs. Further, we have

$$B(z)A(z)^{-1}=-B(-z)A(-z)^{-1}, \tag{3.37}$$

$$\overline{B}(z)\overline{A}(z)^{-1}=-\overline{B}(-z)\overline{A}(-z)^{-1}. \tag{3.38}$$

Therefore, we can simplify the forms of F and $\overline{F}$ as

$$F(n+m)=\begin{cases} 2F_R(n+m), & m=n+2j-1, \\ O, & m=n+2j, \end{cases} \quad j\geqslant 1, \tag{3.39a}$$

$$F_R(n+m)=\frac{1}{2\pi i}\int_{C_R}B(z)A(z)^{-1}z^{-(n+m)-1}\,dz+\sum_{j=1}^{N}C_jz_j^{-(n+m)-1}, \tag{3.39b}$$

and

$$\overline{F}(n+m)=\begin{cases} 2\overline{F}_R(n+m), & m=n+2j-1, \\ 0, & m=n+2j, \end{cases} \quad j\geqslant 1, \tag{3.40a}$$

$$\overline{F}_R(n+m)=\frac{1}{2\pi i}\int_{C_R}\overline{B}(z)\overline{A}(z)^{-1}z^{n+m-1}\,dz-\sum_{k=1}^{\overline{N}}\overline{C}_k\overline{z}_k^{n+m-1}. \tag{3.40b}$$

Here C_R denotes a contour along the right-half portion of the unit circle C.

The symmetry properties of F and $\overline{F}$ give rise to those of κ and $\overline{\kappa}$. From Eqs. (3.33) and (3.34), we obtain

$$\kappa_1(n,m)=\begin{cases} \kappa_{1R}(n,m), & m=n+2j-1, \\ 0, & m=n+2j, \end{cases} \quad j\geqslant 1, \tag{3.41a}$$

$$\kappa_2(n,m)=\begin{cases} \kappa_{2R}(n,m), & m=n+2j, \\ 0, & m=n+2j-1, \end{cases} \quad j\geqslant 1, \tag{3.41b}$$

$$\overline{\kappa}_1(n,m)=\begin{cases} \overline{\kappa}_{1R}(n,m), & m=n+2j, \\ 0, & m=n+2j-1, \end{cases} \quad j\geqslant 1, \tag{3.41c}$$

$$\overline{\kappa}_2(n,m)=\begin{cases} \overline{\kappa}_{2R}(n,m), & m=n+2j-1, \\ 0, & m=n+2j, \end{cases} \quad j\geqslant 1. \tag{3.41d}$$

Considering the above symmetry properties, we obtain the simplified Gel'fand–Levitan–Marchenko equations for κ_{1R} and $\overline{\kappa}_{2R}$:

$\kappa_{1R}(n,m)$ $(m>n, m-n=\text{odd})$

$$=2\overline{F}_R(n+m)-4\sum_{\substack{n'=n+2 \\ n'-n=\text{even}}}^{\infty}\sum_{\substack{n''=n+1 \\ n''-n=\text{odd}}}^{\infty}\kappa_{1R}(n,n'')F_R(n''+n')\overline{F}_R(n'+m), \tag{3.42}$$

$\overline{\kappa}_{2R}(n,m)$ $(m>n, m-n=\text{odd})$

$$=-2F_R(n+m)-4\sum_{\substack{n'=n+2 \\ n'-n=\text{even}}}^{\infty}\sum_{\substack{n''=n+1 \\ n''-n=\text{odd}}}^{\infty}\overline{\kappa}_{2R}(n,n'')\overline{F}_R(n''+n')F_R(n'+m). \tag{3.43}$$

C. Time dependence of the scattering data

Under the rapidly decreasing boundary conditions (3.3), the asymptotic form of the Lax matrix M_n for the sd-matrix mKdV equation (2.10) is given by

$$M_n\to\begin{bmatrix} z^2 I & O \\ O & \dfrac{1}{z^2}I \end{bmatrix} \quad \text{as} \quad n\to\pm\infty. \tag{3.44}$$

We define time-dependent Jost functions by

$$\phi_n^{(t)} \equiv \phi_n e^{z^2 t} \sim \begin{bmatrix} I \\ O \end{bmatrix} z^n e^{z^2 t} \qquad \text{as} \quad n \to -\infty,$$

$$\overline{\phi}_n^{(t)} \equiv \overline{\phi}_n e^{(1/z^2)t} \sim \begin{bmatrix} O \\ -I \end{bmatrix} z^{-n} e^{(1/z^2)t} \quad \text{as} \quad n \to -\infty,$$

$$\psi_n^{(t)} \equiv \psi_n e^{(1/z^2)t} \sim \begin{bmatrix} O \\ I \end{bmatrix} z^{-n} e^{(1/z^2)t} \quad \text{as} \quad n \to +\infty,$$

$$\overline{\psi}_n^{(t)} \equiv \overline{\psi}_n e^{z^2 t} \sim \begin{bmatrix} I \\ O \end{bmatrix} z^n e^{z^2 t} \qquad \text{as} \quad n \to +\infty.$$

From the relations

$$\frac{\partial \phi_n^{(t)}}{\partial t} = M_n \phi_n^{(t)}, \quad \frac{\partial \overline{\phi}_n^{(t)}}{\partial t} = M_n \overline{\phi}_n^{(t)},$$

we obtain

$$\frac{\partial \phi_n}{\partial t} = (M_n - z^2 I)\phi_n, \quad \frac{\partial \overline{\phi}_n}{\partial t} = \left(M_n - \frac{1}{z^2} I \right) \overline{\phi}_n. \tag{3.45}$$

We put the definitions of the scattering data,

$$\phi_n(z) = \overline{\psi}_n(z) A(z,t) + \psi_n(z) B(z,t), \tag{3.46a}$$

$$\overline{\phi}_n(z) = \overline{\psi}_n(z) \overline{B}(z,t) - \psi_n(z) \overline{A}(z,t), \tag{3.46b}$$

into Eq. (3.45). Then taking the limit $n \to +\infty$, we obtain

$$A_t(z,t) = O, \tag{3.47a}$$

$$B_t(z,t) = -\left(z^2 - \frac{1}{z^2} \right) B(z,t), \tag{3.47b}$$

and

$$\overline{A}_t(z,t) = O, \tag{3.47c}$$

$$\overline{B}_t(z,t) = \left(z^2 - \frac{1}{z^2} \right) \overline{B}(z,t). \tag{3.47d}$$

From Eq. (3.47), the time dependences of BA^{-1}, C_j and $\overline{BA}^{-1}, \overline{C}_k$, are respectively concluded as

$$B(z,t)A(z,t)^{-1} = B(z,0)A(z,0)^{-1} e^{-(z^2 - 1/z^2)t}, \tag{3.48}$$

$$C_j(t) = C_j(0) e^{-(z_j^2 - 1/z_j^2)t}, \tag{3.49}$$

and

$$\overline{B}(z,t)\overline{A}(z,t)^{-1} = \overline{B}(z,0)\overline{A}(z,0)^{-1} e^{(z^2 - 1/z^2)t}, \tag{3.50}$$

$$\overline{C}_k(t) = \overline{C}_k(0) e^{(\overline{z}_k^2 - 1/\overline{z}_k^2)t}. \tag{3.51}$$

The above result gives explicitly time-dependent forms of $F_R(n,t)$ and $\overline{F}_R(n,t)$ for odd n,

$$F_R(n,t)=\frac{1}{2\pi i}\int_{C_R} B(z,0)A(z,0)^{-1}z^{-n-1}e^{-(z^2-1/z^2)t}\,dz+\sum_{j=1}^{N} C_j(0)z_j^{-n-1}e^{-(z_j^2-1/z_j^2)t}, \tag{3.52}$$

$$\overline{F}_R(n,t)=\frac{1}{2\pi i}\int_{C_R} \overline{B}(z,0)\overline{A}(z,0)^{-1}z^{n-1}e^{(z^2-1/z^2)t}dz-\sum_{k=1}^{\overline{N}} \overline{C}_k(0)\overline{z}_k^{n-1}e^{(\overline{z}_k^2-1/\overline{z}_k^2)t}. \tag{3.53}$$

D. Initial value problem

According to the constraints $R_n=-Q_n^{\dagger}$ and $Q_nR_n=-\sigma_nI$, we have some additional relations besides Eqs. (3.35)–(3.38), which make the further analysis simple.

The first additional relation is

$$\det \overline{A}(z)=\left\{\det A\left(\frac{1}{z^*}\right)\right\}^*, \tag{3.54}$$

which is proved in Appendix A. This relation gives us useful information about the numbers and the positions of the poles of $A(z)^{-1}$ and $\overline{A}(z)^{-1}$. That is,

$$\overline{N}=N, \quad \overline{z_k}=\frac{1}{z_k^*}. \tag{3.55}$$

Due to Eq. (3.15c), we have the second additional relation,

$$\overline{B}(z)\overline{A}(z)^{-1}=\left\{B\left(\frac{1}{z^*}\right)A\left(\frac{1}{z^*}\right)^{-1}\right\}^{\dagger}, \tag{3.56}$$

which leads to

$$\overline{B}(z)\overline{A}(z)^{-1}=\{B(z)A(z)^{-1}\}^{\dagger} \quad (\text{on } |z|=1), \tag{3.57a}$$

$$\overline{C_k}=-\frac{1}{z_k^{*2}}C_k^{\dagger}. \tag{3.57b}$$

The relations (3.55) and (3.57) give a relation between $\overline{F_R}(n,t)$ and $F_R(n,t)$,

$$\overline{F_R}(n,t)=F_R(n,t)^{\dagger}. \tag{3.58}$$

So as to make the ISM applicable to the sd-cmKdV equations, we have to take account of the internal symmetries of Q_n and R_n defined by Eqs. (2.39)–(2.41).

If we set $Q_n^{(m)}$ and $R_n^{(m)}$ for $m\geqslant 2$ as

$$Q_n^{(m)}=v_n^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1} v_n^{(k)}e_k, \quad R_n^{(m)}=-v_n^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1} v_n^{(k)}e_k, \tag{3.59}$$

the following important relations for $2^{m-1}\times 2^{m-1}$ matrices $\{e_i\}$ hold:

$$\{e_i,e_j\}_+=-2\delta_{ij}\mathbb{1}, \tag{3.60}$$

$$e_k^{\dagger}=-e_k, \tag{3.61}$$

$$\text{tr } e_k=0. \tag{3.62}$$

Here $\{\cdot,\cdot\}_+$ means the anticommutator and $\mathbb{1}$ is the $2^{m-1}\times 2^{m-1}$ unit matrix, which has been denoted as $I_{2^{m-1}}$ so far. From Eq. (3.60), we have

$$\text{tr}(e_ie_j)=-2^{m-1}\delta_{ij}. \tag{3.63}$$

Now, we discuss the initial value problem of the sd-cmKdV equations. Considering the scattering problem (3.1) with the potentials $Q_n^{(m)}$ and $R_n^{(m)}$ for $m \geqslant 2$, we can show the following conditions on the scattering data.

Proposition III.D.1: (1) The determinant of $A(z)$ satisfies

$$\det A(z) = \{\det A(z^*)\}^*, \tag{3.64}$$

as a function of z. Thus the poles of $1/\det A(z)$ outside the unit circle should appear as pairs which are situated symmetric with respect to the real axis. Therefore, we replace N in Sec. III B with $2N$ and can set the values of $2N$ poles in the right portion outside the unit circle C as

$$z_{2j-1} = \xi_j + i\eta_j = a_j e^{i\theta_j}, \quad j = 1,2,...,N, \tag{3.65a}$$

$$z_{2j} = z_{2j-1}^* = \xi_j - i\eta_j = a_j e^{-i\theta_j}, \quad j = 1,2,...,N, \tag{3.65b}$$

where

$$a_j > 1, \quad 0 < \theta_j \leqslant \frac{\pi}{2},$$

for $\theta_j \neq 0$. The condition (3.65) should be interpreted as follows. If $\theta_j = 0$, it does not need its counterpart. The values of the remaining $2N$ poles in the left portion outside the unit circle are given by

$$z_{2N+k} = -z_k, \quad k = 1,2,...,2N.$$

(2) The reflection coefficient $B(z)A(z)^{-1}$ on $|z| = 1$ should be expressed as

$$B(z)A(z)^{-1} = r^{(0)}\mathbb{1} + \sum_{k=1}^{2m-1} r^{(k)} e_k. \tag{3.66}$$

Here $r^{(0)}$ and $r^{(k)}$ are complex functions of z and t which satisfy

$$r^{(0)}(z^*) = r^{(0)}\left(\frac{1}{z}\right) = r^{(0)}(z)^*, \quad r^{(k)}(z^*) = r^{(k)}\left(\frac{1}{z}\right) = r^{(k)}(z)^*, \tag{3.67}$$

for $|z| = 1$.

(3) The residue matrices $\{C_1, C_2, \ldots, C_{2N-1}, C_{2N}\}$ should be expressed as

$$C_{2j-1} = c_j^{(0)}\mathbb{1} + \sum_{k=1}^{2m-1} c_j^{(k)} e_k, \quad j = 1,2,...,N, \tag{3.68a}$$

$$C_{2j} = c_j^{(0)*}\mathbb{1} + \sum_{k=1}^{2m-1} c_j^{(k)*} e_k, \quad j = 1,2,...,N, \tag{3.68b}$$

where $c_j^{(0)}$ and $c_j^{(k)}$ are complex functions of t. For example, $\{C_1, C_2, \ldots, C_{2N}\}$ for the four-component $(m = 2)$ sd-cmKdV equations are given by

$$C_{2j-1} = \begin{bmatrix} \alpha_j & \beta_j \\ -\gamma_j & \delta_j \end{bmatrix}, \quad j = 1,2,...,N,$$

$$C_{2j} = \begin{bmatrix} \delta_j^* & \gamma_j^* \\ -\beta_j^* & \alpha_j^* \end{bmatrix}, \quad j = 1,2,...,N,$$

in a different notation.

A proof of the statements is given in Appendix B.

Considering the above conditions, we get explicit expressions of $F_R(n,t)$ and $\overline{F_R}(n,t)$ for odd n:

$$
\begin{aligned}
F_R(n,t) &= \frac{1}{2\pi i}\int_{C_R} B(z)A(z)^{-1}z^{-n-1}\,dz + \sum_{j=1}^{2N} C_j z_j^{-n-1} \\
&= \frac{1}{2\pi i}\int_{C_R}\left(r^{(0)}\mathbb{1} + \sum_{k=1}^{2m-1} r^{(k)}e_k\right)z^{-n-1}\,dz + \sum_{j=1}^{2N} C_j z_j^{-n-1} \\
&= \frac{1}{2\pi i}\int_{C_{UR}}\left\{(r^{(0)}z^{-n-1} + r^{(0)*}z^{*-n+1})\mathbb{1} + \sum_{k=1}^{2m-1}(r^{(k)}z^{-n-1} + r^{(k)*}z^{*-n+1})e_k\right\}dz \\
&\quad + \sum_{j=1}^{N}\left\{(c_j^{(0)}z_j^{-n-1} + c_j^{(0)*}z_j^{*-n-1})\mathbb{1} + \sum_{k=1}^{2m-1}(c_j^{(k)}z_j^{-n-1} + c_j^{(k)*}z_j^{*-n-1})e_k\right\}, \\
&= \frac{1}{2\pi i}\int_{C_{UR}}\left\{(r^{(0)}z^{-n-1} + r^{(0)*}z^{n-1})\mathbb{1} + \sum_{k=1}^{2m-1}(r^{(k)}z^{-n-1} + r^{(k)*}z^{n-1})e_k\right\}dz \\
&\quad + \sum_{j=1}^{N}\left\{(c_j^{(0)}z_j^{-n-1} + c_j^{(0)*}z_j^{*-n-1})\mathbb{1} + \sum_{k=1}^{2m-1}(c_j^{(k)}z_j^{-n-1} + c_j^{(k)*}z_j^{*-n-1})e_k\right\}, \qquad (3.69)
\end{aligned}
$$

$$
\begin{aligned}
\overline{F_R}(n,t) &= F_R(n,t)^\dagger \\
&= \frac{1}{2\pi i}\int_{C_R}\left(r^{(0)*}\mathbb{1} - \sum_{k=1}^{2m-1} r^{(k)*}e_k\right)z^{n-1}\,dz + \sum_{j=1}^{2N} C_j^\dagger z_j^{*-n-1} \\
&= \frac{1}{2\pi i}\int_{C_{UR}}\left\{(r^{(0)}z^{-n-1} + r^{(0)*}z^{n-1})\mathbb{1} - \sum_{k=1}^{2m-1}(r^{(k)}z^{-n-1} + r^{(k)*}z^{n-1})e_k\right\}dz \\
&\quad + \sum_{j=1}^{N}\left\{(c_j^{(0)}z_j^{-n-1} + c_j^{(0)*}z_j^{*-n-1})\mathbb{1} - \sum_{k=1}^{2m-1}(c_j^{(k)}z_j^{-n-1} + c_j^{(k)*}z_j^{*-n-1})e_k\right\}, \qquad (3.70)
\end{aligned}
$$

where C_{UR} denotes the quadrant (upper-right portion) of the unit circle contour C. Because $B(z)A(z)^{-1}$ and C_j depend on t as Eqs. (3.48) and (3.49), the time dependences of $r^{(0)}, r^{(k)}$ and $c_j^{(0)}, c_j^{(k)}$ are given by

$$
r^{(0)}(z,t) = r^{(0)}(z,0)e^{-(z^2-1/z^2)t}, \quad r^{(k)}(z,t) = r^{(k)}(z,0)e^{-(z^2-1/z^2)t}, \qquad (3.71)
$$

$$
c_j^{(0)}(t) = c_j^{(0)}(0)e^{-(z_j^2-1/z_j^2)t}, \quad c_j^{(k)}(t) = c_j^{(k)}(0)e^{-(z_j^2-1/z_j^2)t}. \qquad (3.72)
$$

It should be noted that $F_R(n,t)$ and $\overline{F_R}(n,t)$ for odd n are expressed as

$$
F_R(n,t) = f^{(0)}(n,t)\mathbb{1} + \sum_{k=1}^{2m-1} f^{(k)}(n,t)e_k,
$$

$$
\overline{F_R}(n,t) = f^{(0)}(n,t)\mathbb{1} - \sum_{k=1}^{2m-1} f^{(k)}(n,t)e_k,
$$

where the real functions $f^{(0)}(n,t)$ and $f^{(k)}(n,t)$ satisfy

$$
\partial_t f^{(0)}(2n+1,t) - f^{(0)}(2n+3,t) + f^{(0)}(2n-1,t) = 0,
$$

$$
\partial_t f^{(k)}(2n+1,t) - f^{(k)}(2n+3,t) + f^{(k)}(2n-1,t) = 0.
$$

Considering the relation (3.58), we get the time-dependent Gel'fand–Levitan–Marchenko equations,

$$\kappa_{1R}(n,m;t) \quad (m>n, m-n=\text{odd})$$

$$=2F_R(n+m,t)^\dagger-4\sum_{\substack{n_1=n+2\\ n_1-n=\text{even}}}^{\infty}\sum_{\substack{n_2=n+1\\ n_2-n=\text{odd}}}^{\infty}\kappa_{1R}(n,n_2;t)F_R(n_2+n_1,t)F_R(n_1+m,t)^\dagger, \tag{3.73}$$

$$\overline{\kappa}_{2R}(n,m;t) \quad (m>n, m-n=\text{odd})$$

$$=-2F_R(n+m,t)-4\sum_{\substack{n_1=n+2\\ n_1-n=\text{even}}}^{\infty}\sum_{\substack{n_2=n+1\\ n_2-n=\text{odd}}}^{\infty}\overline{\kappa}_{2R}(n,n_2;t)F_R(n_2+n_1,t)^\dagger F_R(n_1+m,t), \tag{3.74}$$

where $F_R(n,t)$ for odd n is given by Eq. (3.69).

We can solve the initial value problem of the sd-cmKdV equations as follows.

(1) For given potentials at $t=0$, $Q_n(0)$ and $R_n(0)$ which are expressed as Eq. (3.59), we solve the scattering problem (3.1), and obtain scattering data $\{B(z)A(z)^{-1}, z_j, C_j\}$ or, more concretely, $\{r^{(0)}, r^{(k)}, z_j, c_j^{(0)}, c_j^{(k)}\}$.
(2) The time dependence of the scattering data is given by Eqs. (3.48) and (3.49) [or (3.71) and (3.72)].
(3) We substitute the time-dependent scattering data into the Gel'fand–Levitan–Marchenko equations (3.73) and (3.74). Solving the equations, we reconstruct the time-dependent potentials

$$Q_n(t)=-\kappa_{1R}(n,n+1;t), \tag{3.75}$$

$$R_n(t)=-\overline{\kappa}_{2R}(n,n+1;t). \tag{3.76}$$

This directly proves the complete integrability of the sd-cmKdV equations (2.38). The same procedure is applicable to solve the initial value problem of the higher flows of the sd-cmKdV hierarchy. For instance, we can consider the Lax equation (2.6) with the L_n-matrix (3.1) whose time derivative is replaced with the second flow of the sd-cmKdV hierarchy. As a solution of such a Lax equation, we get an M_n-matrix which contains from z^4 to $1/z^4$ terms. Correspondingly, the time dependences of the scattering data (3.48)–(3.51) should be replaced by, for instance,

$$B(z,t)A(z,t)^{-1}=B(z,0)A(z,0)^{-1}e^{-(z^4-1/z^4)t}, \tag{3.77}$$

$$C_j(t)=C_j(0)e^{-(z_j^4-1/z_j^4)t}, \tag{3.78}$$

and

$$\overline{B}(z,t)\overline{A}(z,t)^{-1}=\overline{B}(z,0)\overline{A}(z,0)^{-1}e^{(z^4-1/z^4)t}, \tag{3.79}$$

$$\overline{C}_k(t)=\overline{C}_k(0)e^{(\overline{z}_k^4-1/\overline{z}_k^4)t}. \tag{3.80}$$

Employing the above time dependences, we can solve the initial value problem of the second flow of the sd-cmKdV hierarchy:

$$\frac{\partial v_n^{(i)}}{\partial t}=\left(1+\sum_{j=0}^{M-1}v_n^{(j)2}\right)\left\{\left(1+\sum_{k=0}^{M-1}v_{n+1}^{(k)2}\right)v_{n+2}^{(i)}-\left(1+\sum_{k=0}^{M-1}v_{n-1}^{(k)2}\right)v_{n-2}^{(i)}\right.$$
$$\left.-\left(\sum_{k=0}^{M-1}v_{n+1}^{(k)2}-\sum_{k=0}^{M-1}v_{n-1}^{(k)2}\right)v_n^{(i)}+2\left(\sum_{k=0}^{M-1}v_{n+1}^{(k)}v_n^{(k)}+\sum_{k=0}^{M-1}v_n^{(k)}v_{n-1}^{(k)}\right)(v_{n+1}^{(i)}-v_{n-1}^{(i)})\right\},$$
$$i=0,1,\ldots,M-1. \tag{3.81}$$

E. Soliton solutions

Assuming the reflection-free potentials which satisfy for z on $|z|=1$,

$$B(z)=\overline{B}(z)=O,$$

we can construct soliton solutions of the sd-cmKdV equations. In this case, $F_R(n,t)$ and $\overline{F}_R(n,t)$ for odd n are given by

$$F_R(n,t)=\sum_{j=1}^{2N} C_j(t)z_j^{-n-1}, \quad C_j(t)=C_j(0)e^{-(z_j^2-1/z_j^2)t}, \tag{3.82}$$

$$\overline{F}_R(n,t)=-\sum_{k=1}^{2N} \overline{C}_k(t)\overline{z}_k^{n-1}, \quad \overline{C}_k(t)=\overline{C}_k(0)e^{(\overline{z}_k^2-1/\overline{z}_k^2)t}. \tag{3.83}$$

To solve Eq. (3.42) [or Eq. (3.73)] with Eqs. (3.82) and (3.83), we set

$$\kappa_{1R}(n,m)=\sum_{k=1}^{2N} P_k(n)\overline{C}_k(t)\overline{z}_k^{n+m-1} \quad (m-n=\text{odd}). \tag{3.84}$$

Substituting Eq. (3.84) into Eq. (3.42) or Eq. (3.73), we have

$$P_k(n)-4\sum_{l=1}^{2N}\sum_{j=1}^{2N}\left(\frac{\overline{z}_l}{z_j}\right)^{2n}\frac{\overline{z}_k^2}{(z_j^2-\overline{z}_k^2)(z_j^2-\overline{z}_l^2)}P_l(n)\overline{C}_l(t)C_j(t)=-2I. \tag{3.85}$$

We introduce a matrix S defined by

$$\begin{aligned} S_{lk}&\equiv\delta_{lk}I-4\sum_{j=1}^{2N}\left(\frac{\overline{z}_l}{z_j}\right)^{2n}\frac{\overline{z}_k^2}{(z_j^2-\overline{z}_k^2)(z_j^2-\overline{z}_l^2)}\overline{C}_l(t)C_j(t)\\ &=\delta_{lk}I+4\sum_{j=1}^{2N}\frac{1}{z_j^{2n}z_l^{*2n}(z_j^2z_k^{*2}-1)(z_j^2z_l^{*2}-1)}C_l(t)^{\dagger}C_j(t), \quad 1\leqslant l,k\leqslant 2N. \end{aligned} \tag{3.86}$$

Then Eq. (3.85) is expressed by

$$(\,P_1\,P_2\cdots P_{2N}\,)\begin{pmatrix} S_{11} & \cdots & S_{1\,2N}\\ \vdots & \ddots & \vdots\\ S_{2N1} & \cdots & S_{2N\,2N}\end{pmatrix}=-2(\underbrace{I\,I\cdots I}_{2N}).$$

Similarly, to solve Eq. (3.43) [or Eq. (3.74)] with Eqs. (3.82) and (3.83), we set

$$\overline{\kappa}_{2R}(n,m)=\sum_{j=1}^{2N}\overline{P}_j(n)C_j(t)z_j^{-(n+m)-1} \quad (m-n=\text{odd}). \tag{3.87}$$

Substituting Eq. (3.87) into Eq. (3.43) or Eq. (3.74), we have

$$\overline{P}_j(n)-4\sum_{l=1}^{2N}\sum_{k=1}^{2N}\frac{\overline{z}_k^{2n+2}}{z_l^{2n}}\frac{1}{(z_l^2-\overline{z}_k^2)(z_j^2-\overline{z}_k^2)}\overline{P}_l(n)C_l(t)\overline{C}_k(t)=-2I. \tag{3.88}$$

In terms of a matrix $\overline{S}$,

$$\bar{S}_{lk} \equiv \delta_{lk} I - 4 \sum_{j=1}^{2N} \frac{\bar{z}_j^{2n+2}}{z_l^{2n}} \frac{1}{(z_l^2 - \bar{z}_j^2)(z_k^2 - \bar{z}_j^2)} C_l(t) \bar{C}_j(t)$$

$$= \delta_{lk} I + 4 \sum_{j=1}^{2N} \frac{1}{z_l^{2n} z_j^{*2n} (z_l^2 z_j^{*2} - 1)(z_k^2 z_j^{*2} - 1)} C_l(t) C_j(t)^\dagger, \quad 1 \leqslant l,k \leqslant 2N, \tag{3.89}$$

we rewrite Eq. (3.88) as

$$(\bar{P}_1 \bar{P}_2 \cdots \bar{P}_{2N}) \begin{pmatrix} \bar{S}_{11} & \cdots & \bar{S}_{1\,2N} \\ \vdots & \ddots & \vdots \\ \bar{S}_{2N1} & \cdots & \bar{S}_{2N\,2N} \end{pmatrix} = -2(\underbrace{I\, I \cdots I}_{2N}).$$

Thus the N-soliton solution of the sd-cmKdV equations (2.38) is given by

$$Q_n^{(m)}(t) = -\kappa_{1R}(n, n+1; t)$$

$$= -\sum_{k=1}^{2N} P_k(n) \bar{C}_k(t) \frac{1}{z_k^{*2n}} = -2(\underbrace{I\, I \cdots I}_{2N}) S^{-1} \begin{pmatrix} C_1(t)^\dagger \dfrac{1}{z_1^{*2n+2}} \\ C_2(t)^\dagger \dfrac{1}{z_2^{*2n+2}} \\ \vdots \\ C_{2N}(t)^\dagger \dfrac{1}{z_{2N}^{*2n+2}} \end{pmatrix}, \tag{3.90a}$$

$$R_n^{(m)}(t) = -\bar{\kappa}_{2R}(n, n+1; t)$$

$$= -\sum_{j=1}^{2N} \bar{P}_j(n) C_j(t) \frac{1}{z_j^{2n+2}} = 2(\underbrace{I\, I \cdots I}_{2N}) \bar{S}^{-1} \begin{pmatrix} C_1(t) \dfrac{1}{z_1^{2n+2}} \\ C_2(t) \dfrac{1}{z_2^{2n+2}} \\ \vdots \\ C_{2N}(t) \dfrac{1}{z_{2N}^{2n+2}} \end{pmatrix}. \tag{3.90b}$$

Rigorously, Eq. (3.90) includes solitons, breathers, and their coexistence. To extract soliton solutions, we should impose appropriate conditions on the residue matrices $\{C_j\}, \{\bar{C}_k\}$. As a criterion of soliton, we employ the condition that each solitary wave observed in the first conserved density,

$$\log\left\{1 + \sum_{j=0}^{2m-1} (v_n^{(j)}(t))^2\right\}, \tag{3.91}$$

has a time-independent shape. Calculating an asymptotic form of (3.91) at $n \to +\infty$, we find that the criterion gives necessary conditions

$$C_{2j-1}\bar{C}_{2j} = \bar{C}_{2j} C_{2j-1} = C_{2j}\bar{C}_{2j-1} = \bar{C}_{2j-1} C_{2j} = O, \quad j = 1,2,...,N. \tag{3.92}$$

Equation (3.92) is simplified to

$$\sum_{i=0}^{2m-1} (c_j^{(i)})^2 = 0, \quad j = 1,2,...,N. \tag{3.93}$$

To check the sufficiency of the conditions (3.92) or (3.93), we consider the $N=1$ case. For this purpose, we set

$$\begin{aligned}\overline{C_1} &= -\frac{1}{z_1^{*2}}\, C_1^\dagger \\ &= -\frac{1}{z_1^{*2}}\left(c_1^{(0)} * \mathbb{1} - \sum_{k=1}^{2m-1} c_1^{(k)} * e_k \right) \\ &\equiv \overline{c}_1^{(0)} \mathbb{1} + \sum_{k=1}^{2m-1} \overline{c}_1^{(k)} e_k , \end{aligned} \tag{3.94}$$

$$\overline{C_2} = \overline{c}_1^{(0)} * \mathbb{1} + \sum_{k=1}^{2m-1} \overline{c}_1^{(k)} * e_k , \tag{3.95}$$

$$\overline{z_1} = \frac{1}{z_1^*} = e^{-W+i\theta}, \quad W>0, \tag{3.96}$$

$$e^{\phi_0} = \frac{\sinh 2W}{\sqrt{2 \sum_{j=0}^{2m-1} |\overline{c}_1^{(j)}(0)|^2}}. \tag{3.97}$$

Here Eq. (3.93) for $N=1$ is rewritten as

$$\sum_{j=0}^{2m-1} (\overline{c}_1^{(j)}(0))^2 = 0. \tag{3.98}$$

With the help of the conditions (3.92), Eq. (3.90) for $N=1$ yields

$$\begin{aligned} Q_n^{(m)}(t) &= \operatorname{sech}\left\{ 2nW + \frac{1}{2}\left(z_1^2 + z_1^{*2} - \frac{1}{z_1^2} - \frac{1}{z_1^{*2}} \right) t + \phi_0 \right\} \frac{\sinh 2W}{\sqrt{2 \sum_{j=0}^{2m-1} |\overline{c}_1^{(j)}(0)|^2}} \\ &\quad \cdot \left\{ \left(\frac{z_1}{z_1^*} \right)^n \overline{C}_1(0) e^{(1/2)(z_1^2 - z_1^{*2} - 1/z_1^2 + 1/z_1^{*2})t} + \left(\frac{z_1^*}{z_1} \right)^n \overline{C}_2(0) e^{-(1/2)(z_1^2 - z_1^{*2} - 1/z_1^2 + 1/z_1^{*2})t} \right\} \\ &= \operatorname{sech}\{2nW + 2(\sinh 2W \cos 2\theta)t + \phi_0\} \frac{\sinh 2W}{\sqrt{2 \sum_{j=0}^{2m-1} |\overline{c}_1^{(j)}(0)|^2}} \\ &\quad \cdot \{ \overline{C}_1(0) e^{2i\{n\theta + (\cosh 2W \sin 2\theta)t\}} + \overline{C}_2(0) e^{-2i\{n\theta + (\cosh 2W \sin 2\theta)t\}} \}, \end{aligned} \tag{3.99a}$$

$$R_n^{(m)}(t) = -Q_n^{(m)}(t)^\dagger. \tag{3.99b}$$

Equation (3.99) can be expressed explicitly for each component as

$$\begin{aligned} v_n^{(i)}(t) &= \operatorname{sech}\{2nW + 2(\sinh 2W \cos 2\theta)t + \phi_0\} \frac{\sinh 2W}{\sqrt{2 \sum_{j=0}^{2m-1} |\overline{c}_1^{(j)}(0)|^2}} \\ &\quad \cdot \{ \overline{c}_1^{(i)}(0) e^{2i\{n\theta + (\cosh 2W \sin 2\theta)t\}} + \overline{c}_1^{(i)} * (0) e^{-2i\{n\theta + (\cosh 2W \sin 2\theta)t\}} \}, \quad i = 0,1,\ldots,2m-1. \end{aligned} \tag{3.100}$$

For the above solution, we have

$$\sum_{i=0}^{2m-1} (v_n^{(i)}(t))^2 = \sinh^2 2W \operatorname{sech}^2\{2nW + 2(\sinh 2W \cos 2\theta)t + \phi_0\}, \tag{3.101}$$

which satisfies the criterion given before. Thus, Eq. (3.100) with (3.98) is interpreted as a one-soliton solution for the sd-cmKdV equations (2.38) with $M=2m$. The structure of the one-soliton solution is essentially the same as that for the continuous cmKdV equations (1.1).[13]

This result shows that the conditions (3.92) are necessary and sufficient for Eq. (3.90) to be the pure N-soliton solution. It is noted that the number of the arbitrary parameters is larger than that of the continuous coupled NLS equations.

A comment is in order. Ohta[21] obtained an N-soliton solution for semi-discrete coupled NLS equations, which can be cast into the sd-cmKdV equations by a simple transformation. The above N-soliton solution (3.90) with the constraints (3.92) contains more parameters for general $M=2m$ than that of Ohta's. We conjecture that Ohta's result coincides with ours by a particular choice of those parameters.

It is not evident whether Eqs. (3.90a) and (3.90b) can be expressed as

$$Q_n^{(m)}(t)=v_n^{(0)}(t)\mathbb{1}+\sum_{k=1}^{2m-1} v_n^{(k)}(t)e_k, \tag{3.102a}$$

$$R_n^{(m)}(t)=-v_n^{(0)}(t)\mathbb{1}+\sum_{k=1}^{2m-1} v_n^{(k)}(t)e_k, \tag{3.102b}$$

without using the products of e_i such as $e_ie_j, e_ie_je_k$. Further, $v_n^{(0)}, v_n^{(k)}$ should be real in Eqs. (3.102a) and (3.102b). Noting the fact that summations and products of real quaternions are real quaternions, either Eq. (3.102a) or (3.102b) can be proved for $m=2$ (four-component sd-cmKdV equations) by using the Neumann–Liouville expansion (see Appendix C). It is left for a future problem to prove both Eqs. (3.102a) and (3.102b) for $M=2m\geqslant 4$ rigorously.

IV. SUMMARY AND DISCUSSIONS

In this paper, we have studied the semi-discrete version of the coupled mKdV equations (sd-cmKdV equations). To solve the model, we have presented a new extension of the discrete version of the ISM proposed by Ablowitz and Ladik.[18,19] By this extension, we have shown for the first time that the initial value problem of the sd-cmKdV equations is solvable by the ISM. We remark that both of the constraints (3.2) play an important role in the process of the ISM, while only the former constraint is necessary for the continuous theory. Further, the extension enables us to solve semi-discrete coupled NLS equations or, more generally, semi-discrete coupled Hirota equations[26] by the ISM. The details will be reported in separate papers.

The sd-cmKdV equations can be transformed into other completely integrable models. If we take new dependent variables as

$$v_{2n}^{(i)}=V_n^{(i)}, \quad v_{2n-1}^{(i)}=-I_n^{(i)},$$

the sd-cmKdV equations (2.4) are cast into a new coupled version of the self-dual network equations,[18,27]

$$\frac{1}{1-\sum_{j=0}^{M-1}\varepsilon_j V_n^{(j)2}}\frac{\partial V_n^{(i)}}{\partial t}=I_n^{(i)}-I_{n+1}^{(i)}, \tag{4.1a}$$

$$\frac{1}{1-\sum_{j=0}^{M-1}\varepsilon_j I_n^{(j)2}}\frac{\partial I_n^{(i)}}{\partial t}=V_{n-1}^{(i)}-V_n^{(i)}, \tag{4.1b}$$

with $\varepsilon_j=\pm 1$.

If we set one variable to be a constant in Eq. (2.4), for instance,

$$\varepsilon_0 v_n^{(0)2}=1 \quad \text{for all } n,$$

the sd-cmKdV equations are cast into

$$\frac{\partial v_n^{(i)}}{\partial t} = -\left(\sum_{j=1}^{M-1} \varepsilon_j v_n^{(j)2} \right)(v_{n+1}^{(i)} - v_{n-1}^{(i)}), \quad i = 1,2,...,M-1. \tag{4.2}$$

This system was studied by Hirota, who gave its molecule solutions.[22]

Shifting the variables $v_n^{(i)}$ by $v_n^{(i)} \to a_i v_n^{(i)} + b_i$, we obtain a coupled version of the generalized Volterra model:[27]

$$\frac{\partial v_n^{(i)}}{\partial t} = \left(\alpha + \sum_j \beta_j v_n^{(j)} + \sum_j \gamma_j v_n^{(j)2} \right)(v_{n+1}^{(i)} - v_{n-1}^{(i)}). \tag{4.3}$$

We believe that our extensions of the Ablowitz–Ladik hierarchy will be useful in the study of the discrete multi-component integrable models.

ACKNOWLEDGMENTS

One of the authors (TT) would like to express his hearty thanks to Professor K. Sogo, Dr. N. Sasa, Dr. M. Hisakado, and Dr. M. Shiroishi for many useful comments and fruitful discussions. While writing the paper, the authors were informed by Dr. Hisakado that the similar Lax formulation in terms of the Clifford algebra was used in the work of Eichenherr and Pohlmeyer.[28] The authors are grateful to Dr. Y. Ohta, D. Trubatch, and Professor M. J. Ablowitz for informing their results prior to publication.

In order to check the conservation laws of the semi-discrete coupled mKdV equations, the authors partly utilized profitable *Mathematica* algorithm coded by Ü. Göktaş, W. Hereman, and G. Erdmann.[29]

APPENDIX A: PROOF OF EQ. (3.54)

We start from a counterpart of the linear equations (2.5),

$$\Psi_{n+1} = L_n \Psi_n, \tag{A1}$$

where Ψ_n is assumed to be a square matrix. Then, we get a chain of identities,

$$\Psi_{n+1}\Psi_n^{-1} = L_n, \quad \frac{\det \Psi_{n+1}}{\det \Psi_n} = \det L_n,$$

$$\det \Psi_\infty = \det \Psi_{-\infty} \cdot \prod_{n=-\infty}^{\infty} \det L_n. \tag{A2}$$

We consider the case that L_n is given by

$$L_n = \begin{bmatrix} zI & Q_n \\ R_n & \dfrac{1}{z} I \end{bmatrix}, \tag{A3}$$

with $l \times l$ square matrices I, Q_n, and R_n, which satisfy the constraints (3.2). Equation (A2) gives

$$\det \Psi_\infty = \det \Psi_{-\infty} \cdot \left(\prod_{n=-\infty}^{\infty} \rho_n \right)^l = \tau^l \cdot \det \Psi_{-\infty},$$

where ρ_n and τ are defined by Eqs. (3.4) and (3.8),

$$\rho_n=(\det L_n)^{1/l}=1+\sigma_n=1+\sum_{j=0}^{2m-1} v_n^{(j)2},$$

$$\tau=\prod_{n=-\infty}^{\infty}\rho_n.$$

It should be noted that τ is a conserved quantity, which is proved by the zero-curvature condition (2.6) [see also Eq. (2.45)].

If we take $[\phi_n\ \psi_n]$ as Ψ_n, we get

$$\lim_{n\to-\infty}\det[\overline{\phi}_n\ \overline{\psi}_n]=\frac{1}{\tau^l}\cdot\det A^{\dagger}\left(\frac{1}{z^*}\right),$$

$$\lim_{n\to+\infty}\det[\overline{\phi}_n\ \overline{\psi}_n]=\det\overline{A}(z),$$

which lead to the formula to be proved,

$$\det\overline{A}(z)=\left\{\det A\left(\frac{1}{z^*}\right)\right\}^*. \tag{A4}$$

APPENDIX B: PROOF OF PROPOSITION III.D.1

Let us recall $\Gamma_n=\Psi_{2n}\Psi_{1n}^{-1}$ defined in Sec. II C. We assume

$$\lim_{|z|\to\infty}\Gamma_n=O. \tag{B1}$$

Then the relation (2.18) with Eq. (2.22) becomes

$$\begin{aligned}\Gamma_{n+1}&=\left(R_n+\frac{1}{z}\Gamma_n\right)(zI+Q_n\Gamma_n)^{-1}\\&=\left(R_n+\frac{1}{z}\Gamma_n\right)z^{-1}\sum_{j=0}^{\infty}(-1)^j\left(\frac{1}{z}Q_n\Gamma_n\right)^j\\&=\sum_{j=0}^{\infty}\frac{(-1)^j}{z^{j+1}}R_nQ_n\cdot\Gamma_n(Q_n\Gamma_n)^{j-1}+\sum_{j=0}^{\infty}\frac{(-1)^j}{z^{j+2}}\Gamma_n(Q_n\Gamma_n)^j.\end{aligned} \tag{B2}$$

Here Q_n and R_n for the $2m$-component sd-cmKdV equations (2.38) ($m\geqslant 2$) are given by

$$Q_n^{(m)}=v_n^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}v_n^{(k)}e_k,\quad R_n^{(m)}=-v_n^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}v_n^{(k)}e_k, \tag{B3}$$

which lead to

$$Q_n^{(m)}R_n^{(m)}=-\sum_{j=0}^{2m-1}v_n^{(j)2}\cdot\mathbb{1}, \tag{B4}$$

as is mentioned in (2.43).

We first show the following theorem and corollary.

Theorem B.1: Let X,Y,Z be $2^{m-1}\times 2^{m-1}$ matrices given by

$$X=x^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}x^{(k)}e_k,\quad Y=y^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}y^{(k)}e_k,\quad Z=z^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}z^{(k)}e_k,$$

where the coefficients $x^{(0)},x^{(k)}$; $y^{(0)},y^{(k)}$; $z^{(0)},z^{(k)}$ are real. Then there exist real numbers $w^{(0)},w^{(k)}$ that satisfy

$$W\equiv XYZ+ZYX=w^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}w^{(k)}e_k. \tag{B5}$$

A proof is given in Ref. 13. □

Corollary B.2: Let X,Y be $2^{m-1}\times 2^{m-1}$ matrices given by

$$X=x^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}x^{(k)}e_k,\quad Y=y^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}y^{(k)}e_k,$$

where the coefficients $x^{(0)},x^{(k)}$; $y^{(0)},y^{(k)}$ are real. Then there exist real numbers $s_j^{(0)},s_j^{(k)}$ that satisfy

$$\begin{aligned}S_j&\equiv(XY)^{j-1}X=X(YX)^{j-1}\\&=s_j^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}s_j^{(k)}e_k.\end{aligned} \tag{B6}$$

The statement can be proved from Theorem B.1, by induction. □

Using Corollary B.2, we can prove by induction that Γ_n is expressed by

$$\Gamma_n=\gamma_n^{(0)}\mathbb{1}+\sum_{k=1}^{2m-1}\gamma_n^{(k)}e_k, \tag{B7}$$

where the coefficients $\gamma_n^{(0)}$ and $\gamma_n^{(k)}$ satisfy

$$\gamma_n^{(0)}(z)=\{\gamma_n^{(0)}(z^*)\}^*,\quad \gamma_n^{(k)}(z)=\{\gamma_n^{(k)}(z^*)\}^*. \tag{B8}$$

In other words, $\gamma_n^{(0)}(z)$ and $\gamma_n^{(k)}(z)$ are polynomials in $1/z$ with real coefficients. This fact is particularly easy to understand when the potentials Q_n and R_n are on compact support.

We recall the asymptotic behavior of the Jost function ϕ_n at $n\to\pm\infty$,

$$\phi_n\equiv\begin{bmatrix}\phi_{1n}\\ \phi_{2n}\end{bmatrix}\sim\begin{bmatrix}I\\ O\end{bmatrix}z^n \qquad \text{as } n\to-\infty,$$

$$\sim\begin{bmatrix}A(z)z^n\\ B(z)z^{-n}\end{bmatrix} \quad \text{as } n\to+\infty.$$

These relations yield

$$\lim_{n\to-\infty}\phi_{1n}z^{-n}=I,\quad \lim_{n\to+\infty}\phi_{1n}z^{-n}=A(z). \tag{B9}$$

We see that

$$\lim_{|z|\to\infty}\phi_{2n}\phi_{1n}^{-1}=O.$$

Then we can replace Γ_n in Eq. (B7) with $\phi_{2n}\phi_{1n}^{-1}$. Thus $\det A(z)$ is expressed as

$$\begin{aligned}
\det A(z) &= \exp\{\mathrm{tr}\ \log A(z)\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} [\log(\phi_{1n+1}z^{-n-1}) - \log(\phi_{1n}z^{-n})] \right\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} \log[\phi_{1n+1}\phi_{1n}^{-1}z^{-1}] \right\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} \log\left[I + \frac{1}{z} Q_n^{(m)}\phi_{2n}\phi_{1n}^{-1} \right] \right\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j} \left(\frac{1}{z} Q_n^{(m)}\Gamma_n \right)^j \right\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j\cdot z^j} Q_n^{(m)} \cdot (\Gamma_n Q_n^{(m)})^{j-1}\Gamma_n \right\} \\
&= \exp\left\{ \mathrm{tr} \sum_{n=-\infty}^{\infty} \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j\cdot z^j} Q_n^{(m)} \left(s_{j,n}^{(0)}\mathbb{1} + \sum_{k=1}^{2m-1} s_{j,n}^{(k)} e_k \right) \right\} \\
&= \exp\left\{ 2^{m-1} \sum_{n=-\infty}^{\infty} \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j\cdot z^j} \left(v_n^{(0)} s_{j,n}^{(0)} - \sum_{k=1}^{2m-1} v_n^{(k)} s_{j,n}^{(k)} \right) \right\}. \qquad \text{(B10)}
\end{aligned}$$

At the last equality, we have used Eqs. (3.62) and (3.63). Due to Eq. (B8), $\det A(z)$, as a function of z, satisfies

$$\det A(z) = \{\det A(z^*)\}^*. \qquad \text{(B11)}$$

This is a proof of Proposition III.D.1 (1).

Further, we obtain

$$\begin{aligned}
B(z)A(z)^{-1} &= \lim_{n\to+\infty} \phi_{2n}\phi_{1n}^{-1}\cdot z^{2n} \\
&= \lim_{n\to+\infty} \left[\gamma_n^{(0)}(z) z^{2n}\mathbb{1} + \sum_{k=1}^{2m-1} \gamma_n^{(k)}(z) z^{2n} e_k \right] \\
&= r^{(0)}(z)\mathbb{1} + \sum_{k=1}^{2m-1} r^{(k)}(z) e_k, \qquad \text{(B12)}
\end{aligned}$$

with conditions

$$r^{(0)}(z) = \{r^{(0)}(z^*)\}^*, \quad r^{(k)}(z) = \{r^{(k)}(z^*)\}^*. \qquad \text{(B13)}$$

Using Eqs. (B12) and (B13), we can straightforwardly prove Proposition III.D.1 (2) and (3).

APPENDIX C: NEUMANN–LIOUVILLE EXPANSION

We begin with the Gel'fand–Levitan–Marchenko equations (3.42) and (3.43),

$$\begin{aligned}
&\kappa_{1R}(n,m) \quad (m>n, m-n = \text{odd}) \\
&= 2\overline{F}_R(n+m) - 4\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty} \kappa_{1R}(n, n+2l_2-1) F_R(2n+2l_2+2l_1-1)\overline{F}_R(n+2l_1+m),
\end{aligned} \qquad \text{(C1)}$$

$$\overline{\kappa}_{2R}(n,m) \quad (m>n, m-n=\text{odd})$$
$$= -2F_R(n+m) - 4\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty} \overline{\kappa}_{2R}(n,n+2l_2-1)\overline{F}_R(2n+2l_2+2l_1-1)F_R(n+2l_1+m). \tag{C2}$$

By successive approximations, we obtain the Neumann–Liouville expansion for κ_{1R} and $\overline{\kappa}_{2R}$,

$$\kappa_{1R}(n,n+1)$$
$$= 2\overline{F}_R(2n+1) - 8\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty} \overline{F}_R(2n+2l_2-1)F_R(2n+2l_2+2l_1-1)\overline{F}_R(2n+2l_1+1)+\cdots$$
$$+2\cdot(-4)^j\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\cdots\sum_{l_{2j}=1}^{\infty} \overline{F}_R(2n+2l_{2j}-1)F_R(2n+2l_{2j}+2l_{2j-1}-1)$$
$$\cdot\overline{F}_R(2n+2l_{2j-1}+2l_{2j-2}-1)\cdots\overline{F}_R(2n+2l_3+2l_2-1)F_R(2n+2l_2+2l_1-1)$$
$$\cdot\overline{F}_R(2n+2l_1+1)+\cdots, \tag{C3}$$

$$\overline{\kappa}_{2R}(n,n+1)$$
$$= -2F_R(2n+1) + 8\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty} F_R(2n+2l_2-1)\overline{F}_R(2n+2l_2+2l_1-1)F_R(2n+2l_1+1)+\cdots$$
$$-2\cdot(-4)^j\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\cdots\sum_{l_{2j}=1}^{\infty} F_R(2n+2l_{2j}-1)\overline{F}_R(2n+2l_{2j}+2l_{2j-1}-1)$$
$$\cdot F_R(2n+2l_{2j-1}+2l_{2j-2}-1)\cdots F_R(2n+2l_3+2l_2-1)\overline{F}_R(2n+2l_2+2l_1-1)$$
$$\cdot F_R(2n+2l_1+1)+\cdots. \tag{C4}$$

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[2] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).
[3] M. Wadati, J. Phys. Soc. Jpn. **32**, 1681 (1972).
[4] M. Wadati, J. Phys. Soc. Jpn. **34**, 1289 (1973).
[5] R. Hirota, J. Phys. Soc. Jpn. **33**, 1456 (1972).
[6] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[7] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **31**, 125 (1973).
[8] D. J. Kaup, Stud. Appl. Math. **55**, 9 (1976).
[9] S. V. Manakov, Sov. Phys. JETP **38**, 248 (1974).
[10] N. Sasa and J. Satsuma, J. Phys. Soc. Jpn. **60**, 409 (1991).
[11] P. J. Caudrey, Physica D **6**, 51 (1982).
[12] R. Beals and R. R. Coifman, Inverse Probl. **5**, 87 (1989).
[13] T. Tsuchida and M. Wadati, J. Phys. Soc. Jpn. **67**, 1175 (1998).
[14] S. I. Svinolupov, Commun. Math. Phys. **143**, 559 (1992).
[15] M. Iwao and R. Hirota, J. Phys. Soc. Jpn. **66**, 577 (1997).
[16] H. Flaschka, Prog. Theor. Phys. **51**, 703 (1974).
[17] S. V. Manakov, Sov. Phys. JETP **40**, 269 (1975).
[18] M. J. Ablowitz and J. F. Ladik, J. Math. Phys. **16**, 598 (1975).
[19] M. J. Ablowitz and J. F. Ladik, J. Math. Phys. **17**, 1011 (1976).
[20] M. Hisakado, J. Phys. Soc. Jpn. **66**, 1939 (1997), solv-int/9701022.
[21] Y. Ohta, ''Pfaffian solution for coupled discrete nonlinear Schrödinger equation,'' preprint; in Proceedings of the Brussels Meeting II on *Integrability and Chaos in Discrete Systems, 1997* (in press).
[22] R. Hirota, J. Phys. Soc. Jpn. **66**, 2530 (1997).
[23] M. Wadati, H. Sanuki, and K. Konno, Prog. Theor. Phys. **53**, 419 (1975).
[24] M. Wadati and T. Kamijo, Prog. Theor. Phys. **52**, 397 (1974).
[25] M. Wadati, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980), pp. 287–299.
[26] R. Hirota, J. Math. Phys. **14**, 805 (1973).
[27] M. Wadati, Prog. Theor. Phys. Suppl. **59**, 36 (1976).
[28] H. Eichenherr and K. Pohlmeyer, Phys. Lett. B **89**, 76 (1979).
[29] Ü. Göktaş, W. Hereman, and G. Erdmann, Phys. Lett. A **236**, 30 (1997).

Possibly degenerate tetrad gravity and Maxwell–Dirac fields

Daniel Canarutto[a)]
Dipartimento di Matematica Applicata ''G. Sansone,'' Via S. Marta 3, 50139 Firenze, Italia

(Received 20 November 1997; accepted for publication 19 May 1998)

I present an integrated approach to tetrad gravity and Maxwell–Dirac fields, which is not singular in the case when the tetrad is degenerate. All field equations are linear. The geometric data consist only of a complex vector bundle $\mathbf{S}\to\mathbf{M}$ with two-dimensional fibers, where $\mathbf{M}$ is a real 4-manifold. All needed structures are functorially derived from $\mathbf{S}$. Any considered object which is not a functorial construction is assumed to be a dynamical field. © *1998 American Institute of Physics.* [S0022-2488(98)03809-2]

I. INTRODUCTION

In the classical two-spinor formalism[1,2] the correspondence $v^a \leftrightarrow v^{AA\cdot}$ between spacetime and spinor indices is usually considered as ''adapted'' to the spacetime metric, which represents the gravitational field, according to $g_{ab}\leftrightarrow \boldsymbol{\epsilon}_{AB}\boldsymbol{\epsilon}_{A\cdot B\cdot}$. A somewhat different attitude consists in viewing that correspondence as the gravitational field itself. Namely, if $\mathbf{S}$ is a complex vector bundle with two-dimensional fibers, one can formulate a ''tetrad-gravity'' approach in which the tetrad is valued in the Hermitian subspace $\mathbf{H}$ of $\mathbf{S}\otimes\overline{\mathbf{S}}$. We obtain an integrated formulation of Einstein–Cartan–Maxwell–Dirac fields by assuming *minimal geometric data*: All the needed geometric structures are derived functorially from $\mathbf{S}$, without fixing any extra background field; if we consider an object which is not functorially determined from $\mathbf{S}$, then it is a dynamical field of the theory. So, the tetrad is a map $\Theta: T\mathbf{M}\to\mathbf{H}$; a linear connection $\mathrm{\Gamma}$ on $\mathbf{S}$ naturally splits into the gravitational and electromagnetic parts; the Dirac field ψ is a section of $\mathbf{W}:=\mathbf{S}\oplus\overline{\mathbf{S}}^{\bigstar}$ (there is a natural Clifford map $\mathbf{H}\to\mathrm{End}\mathbf{W}$). Even coupling factors naturally arise as covariantly constant sections of (tensor) powers of a real line bundle $\mathbf{L}$, where $\mathbf{L}^2:=\mathbf{L}\otimes\mathbf{L}$ is defined to be the Hermitian subbundle of $\wedge^2\mathbf{S}\otimes\wedge^2\overline{\mathbf{S}}$. This $\mathbf{L}$ can be seen as the ''space of length units'' (a spacetime object tensorialized by some power of $\mathbf{L}$ might be called a ''conformal density'' [1]). In general I will not assume a given $\boldsymbol{\epsilon}\in\wedge^2\mathbf{S}^{\bigstar}$, or a given Hermitian metric on the fibers of $\mathbf{S}$ (a global fixed $\boldsymbol{\epsilon}$ implies that $\wedge^2\mathbf{S}$ is a trivializable bundle, thus excluding the possibility of describing topological effects).

If we translate *directly* the standard Einstein–Cartan–Maxwell–Dirac theory in terms of the above said fields Θ, $\mathrm{\Gamma}$, and ψ, then we are forced to use Θ^{-1} (in the standard formulation, the inverse $g^{\#}$ of the spacetime metric is essential). However, it has been suggested[3] that a theory which remains nonsingular also when Θ is degenerate may be interesting in view of quantization: it would allow metric signature change, and the degenerate case could be seen as a geometrical basis for a stochastical approach to quantum theory. From the mathematical point of view, the nonsingularity requirement constitutes a selection criterion which leads to a particularly simple formulation.

The novelties of this paper, with respect to the previous one,[4] are essentially the following: now, also the treatment of the electromagnetic field satisfies the nonsingularity criterion and is well integrated with the rest—as a consequence, all Lagrangians and equations are first order; the study of the electromagnetic gauge in this context (Sec. VII) is new; the description of two-spinor connections is improved, leading to simpler expressions of the field equations, and in particular of the generalized Dirac equation; some conventions have been changed in order to have a description more consistent and readily comparable with standard approaches.

a)Electronic mail: canarutto@dma.unifi.it

 © 1998 American Institute of Physics

II. TWO-SPINORS REVISITED

A. Complex spaces

If X is any set and $f:X\to\mathbb{C}$ then we denote by $\overline{f}$ the conjugate map, $\overline{f}(x):=\overline{f(x)}$.

If $\mathbf{V}$ is a finite-dimensional complex vector space then we shall denote by $\mathbf{V}^\star$ and $\mathbf{V}^*$ its complex and real dual spaces, respectively. Moreover we shall denote by $\overline{\mathbf{V}}^\star$ the antidual space (i.e., the space of all antilinear maps $\mathbf{V}\to\mathbb{C}$), and by $\overline{\mathbf{V}}:=(\mathbf{V}^\star)^\star$ the conjugate space (then $\overline{\mathbf{V}}^\star\cong\mathbf{V}^\star$). We have the conjugation anti-isomorphisms $\mathbf{V}^\star\to\overline{\mathbf{V}}^\star$ and $\mathbf{V}\to\overline{\mathbf{V}}$.

Let (ζ_a) be a basis of $\mathbf{V}$, $a=1,\ldots,n$, and (z^a) the dual basis of $\mathbf{V}^\star$. Then we have the conjugate bases $(\overline{\zeta}_{a\cdot}):=(\overline{\zeta_a})$ of $\overline{\mathbf{V}}$ and $(\overline{z}^{\overline{a}}):=(\overline{z^a})$ of $\overline{\mathbf{V}}^\star$. If $v=v^a\zeta_a$ and $\lambda=\lambda_a z^a$ then $\overline{v}=\overline{v}^{\overline{a}}\overline{\zeta}_{a\cdot}$ and $\overline{\lambda}=\overline{\lambda}_{a\cdot}\overline{z}^{\overline{a}}$ with $\overline{v}^{\overline{a}}=\overline{v^a}$, $\overline{\lambda}_{a\cdot}=\overline{\lambda_a}$.

Conjugation can be naturally extended to tensor products of the above spaces with any number of factors. If τ is a tensor then $\overline{\tau}$ has dotted indices in the place of nondotted indices of τ, and vice-versa. A tensor $w\in\mathbf{V}\otimes\overline{\mathbf{V}}$ is said to be Hermitian if $\overline{w}=w^T$, where the superscript T denotes transposition. In coordinates this means $\overline{w}^{b\cdot a}=w^{ab\cdot}$. We have the *real* decomposition $\mathbf{V}\otimes\overline{\mathbf{V}}=H\mathbf{V}\oplus iH\mathbf{V}$ into Hermitian and anti-Hermitian subspaces.

If $\mathbf{V}$ is one-dimensional then $\mathbf{V}\cong\mathbf{V}^{1/2}\otimes\mathbf{V}^{1/2}$ where $\mathbf{V}^{1/2}$ is its *square root* space, which is unique up to an isomorphism. Similarly one defines $\mathbf{V}^{1/p}$ for $p\in\mathbb{N}$, hence also $\mathbf{V}^{q/p}$ for $p,q\in\mathbb{N}$. If we set $\mathbf{V}^{-1}:=\mathbf{V}^\star$, the power of a one-dimensional complex vector space is naturally defined for all rational exponents.

B. Two-spinor space

Let $\mathbf{S}$ be a two-dimensional complex vector space. We set $\mathbf{\Lambda}^2:=\wedge^2\mathbf{S}$ and identify $\mathbf{\Lambda}^{-2}:=\mathbf{\Lambda}^{2\star}$ with $\wedge^2\mathbf{S}^\star$ through the rule[5] $\omega(s\wedge s'):=\frac{1}{2}\omega(s,s')$, $\forall\,\omega\in\wedge^2\mathbf{S}^\star$, $s,\,s'\in\mathbf{S}$, where $s\wedge s':=\frac{1}{2}(s\otimes s'-s'\otimes s)$. If $\omega\neq 0$ then it has a unique ''inverse'' or ''dual'' element $\omega^{-1}\equiv\omega^\star$ such that $\omega(\omega^{-1})=1$.

The Hermitian subspace of $\mathbf{\Lambda}^2\otimes\overline{\mathbf{\Lambda}^2}$ is a real vector space with a distinguished orientation; its positive semispace will be indicated by $\mathbf{L}^2$, thus $\mathbf{\Lambda}^2\otimes\overline{\mathbf{\Lambda}^2}=\mathbb{C}\otimes\mathbf{L}^2$. Moreover we have the square root semispace $\mathbf{L}$, characterized by $\mathbf{L}^2\cong\mathbf{L}\otimes\mathbf{L}$. If r is any rational number then we define $\mathbf{\Lambda}^r:=(\mathbf{\Lambda}^2)^{r/2}$, hence $\mathbf{\Lambda}^r\otimes\overline{\mathbf{\Lambda}^r}=\mathbb{C}\otimes\mathbf{L}^r$. In particular we obtain $(r=1)$ the square root $\mathbf{\Lambda}$ of $\mathbf{\Lambda}^2$.

Next, consider the complex two-dimensional space

$$\mathbf{U}:=\mathbf{L}^{-1/2}\otimes\mathbf{S},$$

and observe that the one-dimensional space $\wedge^2\mathbf{U}=\mathbf{L}^*\otimes\wedge^2\mathbf{S}$ has a natural Hermitian metric; hence there is the distinguished set of normalized symplectic forms on $\mathbf{U}$, differing by a phase factor. This is our 2-*spinor space*. We say that elements of $\mathbf{U}$ and of its tensor algebra are ''conformally invariant,'' while tensorializing by $\mathbf{L}^r$ we obtain ''conformal densities'' of weight r.

Any basis (ξ_A) of $\mathbf{S}$ determines the basis $w:=\boldsymbol{\epsilon}^{AB}\xi_A\wedge\xi_B$ of $\mathbf{\Lambda}^2$, where $\boldsymbol{\epsilon}^{AB}$ is the antisymmetric Ricci matrix, and the basis $l:=\sqrt{w\otimes\overline{w}}$ of $\mathbf{L}$. Then we also have the induced *normalized* basis $(\zeta_A):=(l^{-1/2}\otimes\xi_A)$ of $\mathbf{U}$ and its dual basis (z^A); the induced (mutually dual) bases $\boldsymbol{\epsilon}:=\boldsymbol{\epsilon}_{AB}z^A\wedge z^B\in\wedge^2\mathbf{U}^\star$ and $\boldsymbol{\epsilon}^*\equiv\boldsymbol{\epsilon}^{-1}=\boldsymbol{\epsilon}^{AB}\zeta_A\wedge\zeta_B\in\wedge^2\mathbf{U}$ are then normalized too.

Note that, in contrast to the usual 2-spinor formalism, no ''symplectic'' 2-form $\boldsymbol{\epsilon}$ is fixed in general. Two-spinor indices may be lowered and raised by $\boldsymbol{\epsilon}_{AB}$ and its inverse $\boldsymbol{\epsilon}^{AB}$, but this operation is not intrinsic. However, $\boldsymbol{\epsilon}\otimes\overline{\boldsymbol{\epsilon}}$ is an intrinsic object (Sec. II C). Also, one has a distinguished symplectic 2-form on the space $\mathbf{U}':=(\wedge^2\mathbf{U})^{-1/2}\otimes\mathbf{U}=\mathbf{\Lambda}^{-1}\otimes\mathbf{S}$.

C. Lorentz structure

There is a natural bilinear form $\tilde{g}$ on $\mathbf{U}\otimes\overline{\mathbf{U}}$, given for decomposable elements by

$$\tilde{g}(p\otimes\overline{q},r\otimes\overline{s})=\boldsymbol{\epsilon}(p,r)\overline{\boldsymbol{\epsilon}}(\overline{q},\overline{s}),$$

where $\boldsymbol{\epsilon}$ is any normalized symplectic form. If we consider the decomposition of $\mathbf{U}\otimes\overline{\mathbf{U}}$ into its Hermitian and anti-Hermitian real subspaces, namely

$$\mathbf{U}\otimes\overline{\mathbf{U}}=\mathbf{H}\oplus i\mathbf{H},$$

then it turns out that the restriction of $\tilde{g}$ to $\mathbf{H}$ is a Lorentz metric with signature $(+,-,-,-)$; in fact a normalized basis (ζ_A) induces the basis $(\tau_\lambda := \tau_\lambda^{AB'}\zeta_A\otimes\overline{\zeta_{B'}})$ of $\mathbf{H}$ (where the $(\tau_\lambda^{AB'})$'s are the Pauli matrices divided by $\sqrt{2}$) which turns out to be a $\tilde{g}$-orthonormal basis, i.e., $\tilde{g}(\tau_\lambda,\tau_\mu)=\eta_{\lambda\mu}$. Moreover $\tilde{g}$ determines, up to sign, a normalized volume form $\tilde{\eta}$ on $\mathbf{H}$.

Next we observe that an element of $\mathbf{U}\otimes\overline{\mathbf{U}}$ can be seen as a linear map $\mathbf{U}^\star\to\mathbf{U}$, while an element of $\mathbf{U}^\star\otimes\mathbf{U}^\star$ can be seen as a linear map $\mathbf{U}\to\overline{\mathbf{U}}^\star$. Then we define the map

$$\tilde{\gamma}:\mathbf{U}\otimes\overline{\mathbf{U}}\to\mathrm{End}(\mathbf{U}\oplus\overline{\mathbf{U}}^\star):y\mapsto\tilde{\gamma}(y):=\sqrt{2}(y+\overline{y^b}),$$

where y^b is the element of $\mathbf{U}^\star\otimes\overline{\mathbf{U}}^\star$ corresponding to y via the ''index lowering'' isomorphism induced by $\tilde{g}$.

The restriction of $\tilde{\gamma}$ to $\mathbf{H}$ turns out to be a Clifford map. Thus we are led to regard $\mathbf{W}:=\mathbf{U}\oplus\overline{\mathbf{U}}^\star$ as the space of Dirac spinors, decomposed into its Weyl subspaces. Given $\psi=u+\chi\in\mathbf{W}$, its ''Dirac adjoint'' is $\overline{\psi}=\overline{\chi}+\overline{u}\in\mathbf{U}^\star\oplus\overline{\mathbf{U}}\equiv\mathbf{W}^\star$.

III. TWO-SPINOR CONNECTIONS

We consider any manifold $\mathbf{M}$ and a complex vector bundle $\mathbf{S}\to\mathbf{M}$ with two-dimensional fibers. Base coordinates are indicated by (x^a). According to the constructions of the Sec. II, we now have the bundles $\mathbf{\Lambda}^2$, $\mathbf{L}$, $\mathbf{U}$, $\mathbf{H}$ over $\mathbf{M}$, with smooth natural structures. Moreover we have the semivector bundles $\mathbf{L}^r$, with r rational. But observe that, for given r, according to the bundle topology we may have no $\mathbf{\Lambda}^r=(\mathbf{\Lambda}^2)^{r/2}$, or we may have one or more (not isomorphic) such bundles. We always obtain a unique $\mathbf{\Lambda}^r$ by a suitable restriction of the base manifold; in particular, locally we obtain a bundle $\mathbf{\Lambda}=(\wedge^2\mathbf{S})^{1/2}$.

A complex-linear connection $\mathrm{Б}$ on $\mathbf{S}\to\mathbf{M}$ induces connections on each of the above bundles. Consider a frame (ξ_A) of $\mathbf{S}$ and the induced frames l of $\mathbf{L}$ and (ξ_A) of $\mathbf{U}$ (Sec. II B). Denote by $\mathrm{Б}_a{}^A{}_B$ the coefficients of $\mathrm{Б}$ in the frame (ξ_A). Then one finds

$$\nabla l=-G_a\mathrm{d}x^a\otimes l,\quad \nabla\epsilon=2iA_a\mathrm{d}x_a\otimes\epsilon,$$

where

$$A_a=\frac{1}{4i}(\mathrm{Б}_a{}^A{}_A-\overline{\mathrm{Б}}_a{}^{A'}{}_{A'}),\quad G_a=\tfrac{1}{4}(\mathrm{Б}_a{}^A{}_A+\overline{\mathrm{Б}}_a{}^{A'}{}_{A'}).$$

Since G_a and A_a are real, we see that $\mathrm{Б}$ actually yields a linear connection on $\mathbf{L}$, and that the induced linear connection on $\wedge^2\mathbf{U}$ is Hermitian.

Next consider the linear connection $\tilde{\Gamma}$ induced by $\mathrm{Б}$ on $\mathbf{U}\otimes\overline{\mathbf{U}}$; a calculation shows that its coefficients $\tilde{\Gamma}^\mu_{a\lambda}$ in the Hermitian frame (τ_λ) are real, hence $\tilde{\Gamma}$ is reducible to a real linear connection on the Hermitian subbundle $\mathbf{H}$. Moreover $\tilde{\Gamma}$ is metric, i.e., $\nabla\tilde{g}=0$.

A basic result (see also Ref. 4) is now that, conversely, $\mathrm{Б}$ can be unically expressed in terms of the induced connections $\tilde{\Gamma}$, G, and A by

$$\mathrm{Б}_a{}^A{}_B=(G_a+iA_a)\delta^A{}_B+\tfrac{1}{2}\tilde{\Gamma}_a{}^{AA'}{}_{BA'},$$

where $\tilde{\Gamma}_a{}^{AA'}{}_{BB'}:=\tilde{\Gamma}^\lambda_{a\mu}\tau_\lambda{}^{AA'}\tau^\mu{}_{BB'}$. A similar relation holds among the curvature tensors of the considered connections, namely

$$R_{ab}{}^A{}_B=-(dG+i\mathrm{d}A)_{ab}\delta^A{}_B+\tfrac{1}{2}\tilde{R}_{ab}{}^{AA'}{}_{BA'}.$$

IV. GAUGE TRANSFORMATIONS

The principal bundle of all frames of $\mathbf{S}$ has the structure group $\mathrm{Gl}(2,\mathbb{C})$. For $K\in\mathrm{Gl}(2,\mathbb{C})$, consider the transformations induced on $\wedge^2\mathbf{S}$, $\mathbf{L}$, $\mathbf{U}$, $\wedge^2\mathbf{U}$, and $\mathbf{H}$:

$$K_{\wedge^2\mathbf{S}} = \det K, \qquad K_{\mathbf{L}} = |\det K|, \qquad K_{\mathbf{U}} = |\det K|^{-1/2} K,$$

$$K_{\wedge^2\mathbf{U}} = |\det K|^{-1} \det K = \exp(i \arg\det K),$$

$$\widetilde{K}^{\mu}_{\lambda} := (K_{\mathbf{H}})^{\mu}_{\lambda} = |\det K|^{-1} \tau_{\lambda}{}^{AA\cdot} K^{B}_{A} \overline{K}^{B\cdot}_{A\cdot} \tau^{\mu}{}_{BB\cdot}.$$

Since $\mathrm{Gl}(2,\mathbb{C})$ is connected we see that $K_{\mathbf{H}}$ preserves orientation, namely we have a way of selecting ''positive'' orientations on **H**: those orientations for which the ''Pauli basis'' (τ_λ) induced by any basis of **S** is positively oriented.

In this paper I will assume that the curvature of the connection G on **L** vanishes, $dG = 0$. This amounts to gauging away the conformal (''dilaton'') symmetry, so that the fundamental bundle is $\mathbf{U} \to \mathbf{M}$. The gauge group now reduces to transformations of the form $|\det K|^{-1/2} K$, with $K \in \mathrm{Gl}(2,\mathbb{C})$. Namely the gauge group is constituted by all complex 2×2 matrices having the module of the determinant equal to 1. This group is

$$\widetilde{\mathrm{Sl}}(2,\mathbb{C}) := (\mathrm{U}(1) \times \mathrm{Sl}(2,\mathbb{C}))/\mathbb{Z}_2,$$

where $\mathbb{Z}_2$ is identified with the normal subgroup of $\mathrm{U}(1)\times\mathrm{Sl}(2,\mathbb{C})$ generated by $(-1, -\mathbf{1})$.

V. TWO-SPINOR TETRAD

A. Tetrad and related fields

From now on I assume that **M** is a real four-dimensional manifold. Consider a linear morphism

$$\Theta : T\mathbf{M} \to \mathbf{S} \otimes \overline{\mathbf{S}} = \mathbb{C} \otimes \mathbf{L} \otimes \mathbf{H},$$

i.e., a section

$$\Theta : \mathbf{M} \to \mathbb{C} \otimes \mathbf{L} \otimes \mathbf{H} \otimes T^*\mathbf{M}$$

(all tensor products are over **M**). Its coordinate expression is

$$\Theta = \Theta^{\lambda}_{a}\, l \otimes \tau_{\lambda} \otimes \mathrm{d}x^a = \Theta^{AA\cdot}_{a}\, l \otimes \zeta_A \otimes \overline{\zeta}_{A\cdot} \otimes \mathrm{d}x^a,$$

with Θ^{λ}_{a}, $\Theta^{AA\cdot}_{a} : \mathbf{M} \to \mathbb{C}$. If we restrict our attention to the case when Θ is nondegenerate and valued in the Hermitian subspace $\mathbf{L}\otimes\mathbf{H} \subset \mathbf{S}\otimes\overline{\mathbf{S}}$, then we can view Θ as a ''scaled'' *tetrad* (or *soldering form*, or *vierbein*). In that case the coefficients Θ^{λ}_{a} are real while $\Theta^{AA\cdot}_{a}$ are Hermitian, $\overline{\Theta}^{A\cdot A}_{a} = \Theta^{AA\cdot}_{a}$. By abuse of language we shall call any Θ a (possibly degenerate) tetrad.

From Θ we obtain the following objects (in general, an object ''living'' in **H** will be distinguished by a tilde from its corresponding—via Θ—spacetime object):

$$g := \Theta^* \tilde{g} : \mathbf{M} \to \mathbb{C} \otimes \mathbf{L}^2 \otimes T^*\mathbf{M} \otimes T^*\mathbf{M},$$

$$\eta := \Theta^* \tilde{\eta} : \mathbf{M} \to \mathbb{C} \otimes \mathbf{L}^4 \otimes \wedge^4 T^*\mathbf{M},$$

$$\gamma := \tilde{\gamma} \circ \Theta : T\mathbf{M} \to \mathbb{C} \otimes \mathbf{L} \otimes \mathrm{End}(\mathbf{W}).$$

We have the coordinate expressions:

$$g = \eta_{\lambda\mu} \Theta^{\lambda}_{a} \Theta^{\mu}_{b}\, l^2 \otimes \mathrm{d}x^a \otimes \mathrm{d}x^b = \epsilon_{AB} \epsilon_{A\cdot B\cdot} \Theta^{AA\cdot}_{a} \Theta^{BB\cdot}_{b}\, l^2 \otimes \mathrm{d}x^a \otimes \mathrm{d}x^b,$$

$$\eta = \det(\Theta)\, l^4 \otimes \mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3,$$

$$\gamma = \sqrt{2}\, \Theta^{AA\cdot}_{a}\, l \otimes (\zeta_A \otimes \overline{\zeta}_{A\cdot} + \epsilon_{AB}\epsilon_{A\cdot B\cdot} \overline{z}^{B\cdot} \otimes z^B) \otimes \mathrm{d}x^a.$$

If and only if Θ is real and nondegenerate, the above objects turn out to be a conformal Lorentz metric, the corresponding (up to sign, unique) conformal volume form and a conformal Clifford map.

B. Cotetrad

The following construction holds for any (possibly degenerate) Θ. For $p=0,1,2,3,4$ consider the map

$$\check{\Theta}^{(p)}:\wedge^p(\mathbf{S}\otimes\overline{\mathbf{S}})\otimes(\wedge^p T^*\mathbf{M})\to\mathbb{C}\otimes\mathbf{L}^4\otimes\wedge^4 T^*\mathbf{M}$$

defined by

$$\check{\Theta}^{(p)}(\xi):=\frac{1}{(4-p)!}\circ\tilde{*}[\xi\wedge(\wedge^{4-p}\Theta)],$$

where $\tilde{*}$ is the Hodge operator, in the exterior algebra $\wedge(\mathbf{H})$, determined by $\tilde{\eta}$. In particular, we shall be concerned with $\check{\Theta}^{(2)}$ and $\check{\Theta}:=\check{\Theta}^{(1)}$, which I call the *co-tetrad* (it may be seen as a kind of ''pseudoinverse'' of Θ, defined even if Θ is degenerate). We have the coordinate expression

$$\check{\Theta}(\xi)=\check{\Theta}^a_\lambda\xi^\lambda_a l^4\otimes\mathrm{d}^4x:=\frac{1}{3!}\,\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\Theta^\mu_b\Theta^\nu_c\Theta^\rho_d\xi^\lambda_a l^4\otimes\mathrm{d}^4x.$$

Clearly $\check{\Theta}$ can also be seen as a linear map $\mathbf{S}\otimes\overline{\mathbf{S}}\to\mathbb{C}\otimes\mathbf{L}^4\otimes T\mathbf{M}\otimes\wedge^4 T^*\mathbf{M}$ over $\mathbf{M}$. Using this point of view, in the nondegenerate case we can write $\check{\Theta}=\Theta^{-1}\otimes\eta$.

C. Tetrad and connections

A further point about the nondegenerate case deals with connections, and will turn out to be essential for recovering the standard interpretation: if $\mathbb{B}$ is a complex-linear connection on $\mathbf{S}$, and G and $\tilde{\Gamma}$ are the induced connections on $\mathbf{L}$ and $\mathbf{H}$, then a nondegenerate tetrad $\Theta:T\mathbf{M}\to\mathbf{L}\otimes\mathbf{H}$ yields a unique connection Γ on $T\mathbf{M}$, characterized by the condition

$$\nabla[\Gamma\otimes\tilde{\Gamma}]\Theta=0.$$

Moreover Γ is metric, i.e., $\nabla[\Gamma]g=0$. Denoting by ${\Gamma_a}^\lambda{}_\mu$ the coefficients of Γ in the frame $\Theta_\lambda:=\Theta^{-1}(l\otimes\tau_\lambda)$ we obtain

$${\Gamma_a}^\lambda{}_\mu={\tilde{\Gamma}_a}^\lambda{}_\mu+2G_a\delta^\lambda{}_\mu.$$

In general, the connection Γ will have nonvanishing torsion, with coordinate expression

$$\Theta^\lambda_c T^c{}_{ab}=\partial_{[a}\Theta^\lambda_{b]}+\Theta^\mu_{[a}{\tilde{\Gamma}_{b]}}^\lambda{}_\mu+2\Theta^\lambda_{[a}G_{b]}.$$

The curvature tensors of Γ and $\tilde{\Gamma}$ are related by ${R_{ab}}^\lambda{}_\mu={\tilde{R}_{ab}}^\lambda{}_\mu$, or

$${R_{ab}}^c{}_d={\tilde{R}_{ab}}^\lambda{}_\mu\Theta^c_\lambda\Theta^\mu_d.$$

Remark: In the degenerate case there is no induced spacetime connection Γ, but one still has the two-spinor connection $\mathbb{B}$ and its decomposition.

VI. FIELDS AND FIELD EQUATIONS

I am going to present an approach to Einstein–Cartan–Maxwell–Dirac fields, in which the fundamental fields are Θ, $\mathbb{B}$, F, and ψ; in Sec. VI E. I will add to these the neutrino field. All objects and equations are nonsingular even if Θ is degenerate. For simplicity, henceforth I will take Θ to be Hermitian-valued, but extension to the more general complexified case is straightforward. I will make the further assumption $\mathrm{d}G=0$ (vanishing of the curvature tensor of the connection induced on $\mathbf{L}$), so that the G-part of $\mathbb{B}$ can be gauged away.

I'll define Lagrangian 4-forms by means of the co-tetrad $\breve{\Theta}$; these can be translated into the usual Lagrangians in the nondegenerate case. Coupling constants arise naturally, as covariantly constant sections of **L**, from the requirement that the Lagrangian be conformally invariant. I will get first-order Lagrangians and equations. In Sec. VII I will examine some extensions of the present formulation, related to the electromagnetic gauge.

A. Gravitational Lagrangian

The gravitational field is represented by Θ (which can be viewed as a ''square root'' of the metric) and the traceless part of B, namely $\widetilde{\Gamma}$. We set

$$\mathscr{L}_g := \tfrac{1}{4}\boldsymbol{\kappa}^{-1}\breve{\Theta}^{(2)}(\widetilde{R}^{\#}) = \tfrac{1}{8}\boldsymbol{\kappa}^{-1}\widetilde{*}\circ(\widetilde{R}^{\#}\wedge\Theta\wedge\Theta):\mathbf{M}\to\wedge^4 T^*\mathbf{M},$$

where $\widetilde{R}^{\#}:\mathbf{M}\to\wedge^2 T^*\mathbf{M}\otimes\wedge^2\mathbf{H}$ is the curvature tensor of $\widetilde{\Gamma}$ with one index raised via $\tilde{g}$, and $\boldsymbol{\kappa}:\mathbf{M}\to\mathbf{L}^2$ is a covariantly constant section called the *gravitational coupling factor*. Note how this is necessary in order to obtain a true 4-form on **M**. Setting $\boldsymbol{\kappa}=\kappa l^2$ one has the coordinate expression $\mathscr{L}_g=\ell_g \mathrm{d}^4x$ with

$$\ell_g = \tfrac{1}{8}\kappa^{-1}\boldsymbol{\epsilon}_{\lambda\mu\nu\rho}\boldsymbol{\epsilon}^{abcd}\widetilde{R}_{ab}{}^{\lambda\mu}\Theta_c{}^{\nu}\Theta_d{}^{\rho},$$

which in the nondegenerate case equals $-(1/2\kappa)R\ \det\Theta$.

B. Electromagnetic Lagrangian

The electromagnetic potential and the electromagnetic field are considered independent fields. These are the connection A induced by B on $\wedge^2 U$ and a spinor field

$$\widetilde{F}:\mathbf{M}\to\mathbf{L}^{-2}\otimes\Lambda^2\mathbf{H}^*.$$

The Lagrangian density is the sum of two terms[6]

$$\mathscr{L}_{\mathrm{em}} = -\tfrac{1}{2}\Theta^{(2)}(\mathrm{d}A\otimes\widetilde{F}) + \tfrac{1}{4}(\widetilde{F}\cdot\widetilde{F})\,\eta;$$

its coordinate expression is $\mathscr{L}_{\mathrm{em}}=\ell_{\mathrm{em}}\mathrm{d}^4x$ with

$$\ell_{\mathrm{em}} = -\tfrac{1}{4}\boldsymbol{\epsilon}^{abcd}\boldsymbol{\epsilon}_{\lambda\mu\nu\rho}\partial_a A_b\widetilde{F}^{\lambda\mu}\Theta_c{}^{\nu}\Theta_d{}^{\rho} + \tfrac{1}{4}\widetilde{F}^{\alpha\beta}\widetilde{F}_{\alpha\beta}\ \det\Theta.$$

Since $\widetilde{F}$ does not appear in the other terms of the total Lagrangian, the $\widetilde{F}$ component of the field equations is immediately seen to yield

$$-\tfrac{1}{2}\boldsymbol{\epsilon}^{abcd}\boldsymbol{\epsilon}_{\lambda\mu\nu\rho}\partial_a A_b\Theta_c{}^{\nu}\Theta_d{}^{\rho} + \widetilde{F}_{\lambda\mu}\ \det\Theta = 0,$$

which in the nondegenerate case gives $F := \Theta^*\widetilde{F} = 2\,\mathrm{d}A \Rightarrow \mathscr{L}_{\mathrm{em}} = -\tfrac{1}{4}F^2\,\eta$.

C. Dirac Lagrangian

Recall that the Dirac spinor field and its ''adjoint'' are the fields

$$\psi = u+\chi:\mathbf{M}\to\mathbf{L}^{-3/2}\otimes\mathbf{W} = \mathbf{L}^{-3/2}\otimes(\mathbf{U}\oplus\overline{\mathbf{U}}^{\bigstar}),$$

$$\overline{\psi} = \overline{\chi}+\overline{u}:\mathbf{M}\to\mathbf{L}^{-3/2}\otimes(\mathbf{U}^{\bigstar}\oplus\overline{\mathbf{U}}) = \mathbf{L}^{-3/2}\otimes\mathbf{W}^{\bigstar},$$

so that if $\psi'=u'+\chi'$ then $\langle\overline{\psi},\psi'\rangle = \langle\overline{\chi},u'\rangle + \langle\chi',\overline{u}\rangle$ is valued in $\mathbb{C}\otimes\mathbf{L}^{-3}$. We set

$$\mathscr{L}_D = \frac{i}{\sqrt{2}}\,\breve{\Theta}(\nabla u\otimes\overline{u} - u\otimes\nabla\overline{u} + \tilde{g}^{\#}(\overline{\chi}\otimes\nabla\chi - \nabla\overline{\chi}\otimes\chi)) - m(\langle\chi,\overline{u}\rangle + \langle\overline{\chi},u\rangle)\,\eta,$$

where m: $\mathbf{M}\to\mathbf{L}^{-1}$ is a covariantly constant section. Note how we replaced the combination $mc/\hbar$ with m only, which is essentially the usual setting $\hbar=c=1$. Actually, in the present context, the particle's mass and the Planck constant appear only in this combination. We have the coordinate expression $\mathscr{L}_D=\ell_D d^4x$ with

$$\ell_D=\frac{i}{\sqrt{2}}\,\breve{\Theta}^a_{AA\cdot}(\nabla_a u^A\overline{u}^{A\cdot}-u^A\nabla_a\overline{u}^{A\cdot}+\epsilon^{AB}\overline{\epsilon}^{A\cdot B\cdot}(\overline{\chi}_B\nabla_a\chi_{B\cdot}-\nabla_a\overline{\chi}_B\chi_{B\cdot}))$$
$$-m(\overline{\chi}_A u^A+\chi_{A\cdot}\overline{u}^{A\cdot})\det\Theta.$$

D. Field equations

The Θ component of the field equations reads

$$\tfrac{1}{2}\,\kappa^{-1}\epsilon_{\lambda\mu\nu\rho}\epsilon^{abcd}R_{ab}{}^{\lambda\mu}\theta^\rho_d=\tfrac{1}{2}\,\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\partial_a A_b\widetilde{F}^{\lambda\mu}\Theta^\rho_d-\tfrac{1}{4}\,\widetilde{F}^2\breve{\Theta}^c_\nu$$
$$-\frac{i}{4}\,\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\Theta^\mu_b\Theta^\rho_d(\overline{\psi}\widetilde{\gamma}^\lambda\nabla_a\psi-\overline{\widetilde{\gamma}^\lambda\nabla_a\psi}\psi)+m\overline{\psi}\psi\breve{\Theta}^c_\nu.$$

In the nondegenerate case, the left-hand side amounts essentially to the Einstein tensor, and the right-hand side can be viewed as the sum of the energy-momentum tensors of the electromagnetic field and of the Dirac field. The $\widetilde{\Gamma}$ component gives the equation for torsion

$$\tfrac{1}{2}\kappa^{-1}\epsilon_{\lambda\mu\nu\rho}\epsilon^{abcd}(\partial_b\Theta^\nu_c+\Theta^\sigma_b\widetilde{\Gamma}_c{}^\nu{}_\sigma)\Theta^\rho_d=\frac{i}{4\sqrt{2}}\,[(\breve{\Theta}^a_{AC\cdot}\tau_{[\lambda}{}^{DC\cdot}\tau_{\mu]DA\cdot}-\breve{\Theta}^a_{CA\cdot}\tau_{[\lambda}{}^{CD\cdot}\tau_{\mu]AD\cdot})u^A\overline{u}^{A\cdot}$$
$$+(\breve{\Theta}^{aBC\cdot}\tau_{[\lambda}{}^{DB\cdot}\tau_{\mu]DC\cdot}-\breve{\Theta}^{aCB\cdot}\tau_{[\lambda}{}^{BD\cdot}\tau_{\mu]CD\cdot})\overline{\chi}_B\chi_{B\cdot}].$$

From this we see that the spinor field is a source for torsion, and that in this context we cannot formulate a torsion-free theory.

The A-component is

$$\epsilon^{abcd}(\mathrm{d}*F)_{bcd}+\breve{\Theta}^a_\lambda(\overline{\psi}\widetilde{\gamma}^\lambda\psi)=0.$$

In the nondegenerate case this gives the (second) Maxwell equation $\frac{1}{2}*\mathrm{d}*F=j$, where $j^a:=\Theta^a_\lambda(\overline{\psi}\widetilde{\gamma}^\lambda\psi)$ is the Dirac current.

The ψ component, namely the $\overline{u}$ and $\overline{\chi}$ components, give the following modified form of the Dirac equation

$$\begin{cases}\sqrt{2}i\breve{\Theta}^a_{AA\cdot}\nabla_a u^A-m\chi_{A\cdot}\det\Theta+\mathscr{B}_{AA\cdot}u^A=0,\\ \sqrt{2}i\breve{\Theta}^{aAA\cdot}\nabla_a\chi_{A\cdot}-mu^A\det\Theta+\mathscr{B}^{AA\cdot}\chi_{A\cdot}=0,\end{cases}$$

where

$$\mathscr{B}_{AA\cdot}=\frac{i}{\sqrt{2}}\left[\partial_a\breve{\Theta}^a_{AA\cdot}+\frac{1}{2}\,(\breve{\Theta}^a_{BA\cdot}\widetilde{\Gamma}_a{}^{BC\cdot}{}_{AC\cdot}+\breve{\Theta}^a_{AB\cdot}\widetilde{\Gamma}_a{}^{CB\cdot}{}_{CA\cdot})\right]$$

(in the nondegenerate case just divide by $\det\Theta$ and take into account $\breve{\Theta}=\Theta^{-1}\det\Theta$; dropping the $\mathscr{B}$ terms, this is essentially the usual 2-spinor form of the Dirac equation[1]). The ψ-component gives the conjugate equation.

Finally note that we can express $\partial_a\Theta^\mu_b$ algebraically in terms of $\widetilde{\Gamma}$ and ψ via the torsion equation, and replace for this derivative in $\mathscr{B}_{AA\cdot}$ and $\mathscr{B}^{AA\cdot}$. These kind of torsion-related nonlinearities are known to arise in the context of Einstein–Cartan–Dirac fields.[7–11] Their possible physical meaning is an argument under discussion.

E. Neutrino field

Dropping the mass term in the Dirac Lagrangian poses no mathematical problem, and one obtains two decoupled Weyl equations (with an extra term related to the torsion). How do we ''turn off'' the electromagnetic interaction? Since A, being a connection, is not a section of a vector bundle, apparently it makes no sense to let it be zero. However, if $\mathbf{\Lambda}^2 := \wedge^2\mathbf{S}$ has a root $\mathbf{\Lambda}$ (Secs. II B, III), then we can take the neutrino bundle to be

$$\mathbf{U}' := (\wedge^2\mathbf{U})^{-1/2} \otimes \mathbf{U} = \mathbf{\Lambda}^{-1} \otimes \mathbf{S}.$$

I said that $\mathbf{U}'$ has a natural symplectic structure; moreover any two-spinor connection on $\mathbf{S}$ is easily seen to yield a flat connection on $\wedge^2\mathbf{U}'$. By the way, tensorializing $\mathbf{U}$ by powers of $\mathbf{\Lambda}$ is a natural way of introducing particles with different electric charges interacting with the same electromagnetic potential.

VII. ELECTROMAGNETIC GAUGE

The standard 2-spinor formalism considers certain decompositions of *world-tensors*—we may say elements in the tensor algebra $\otimes\mathbf{H}$—which are important for various reasons. In particular, the reduced forms of the fields are useful in the study of the Cauchy problem.

The decomposition procedure is essentially based on fixing a normalized 2-form $\boldsymbol{\epsilon}$: $\mathbf{M} \to \wedge^2\mathbf{U}^\star$. However, a choice of $\boldsymbol{\epsilon}$ is equivalent to that of an electromagnetic gauge, because the connection A is made into a true 1-form $A_a = -i/4\langle \nabla_a \boldsymbol{\epsilon}, \boldsymbol{\epsilon}^{-1} \rangle$. In the usual setting $\boldsymbol{\epsilon}$ is assumed to be a covariantly constant, fixed object, and the electromagnetic gauge is associated with the choice of a *different* normalized 2-form α (Ref. 1, Sec. V A).

In order to recover that setting, without assuming extra background structure, one might take the neutrino bundle $\mathbf{U}'$ as the fundamental 2-spinor bundle. Then one might reintroduce A as a *new* connection on $\wedge^2\mathbf{U}'$ (possibly flat in the neutrino case). I'll consider the above said decompositions only in relation to the choice of a gauge, allowing $\nabla\boldsymbol{\epsilon} \neq 0$ in general. Locally, we may view the condition $\nabla\boldsymbol{\epsilon} = 0$ as referring to the flat connection determined by the gauge on $\wedge^2\mathbf{U}^\star$; then A becomes a true 1-form as the difference between two connections.[12]

A. Fixing a gauge

In this section and in the next one I assume a fixed gauge. Then, locally, one can find normalized frames of $\mathbf{U}$ such that $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_{AB} z^A \wedge z^B$. Consider now a new field Φ: $\mathbf{M} \to \mathbf{L}^{-2} \otimes \vee^2\mathbf{U}$ (symmetrized tensor product) defined as the contraction between $\overline{F}$ and $\boldsymbol{\epsilon}$, in coordinates $\Phi_{AB} = \frac{1}{2}\overline{\boldsymbol{\epsilon}}^{A\cdot B\cdot}\widetilde{F}_{AA\cdot BB\cdot}$. We can take Φ instead of $\widetilde{F}$ as dynamic object of the theory; as far as the calculation of the field equations is concerned, we find no essential differences. An interesting fact, however, is that—by expressing $d{*}F$ in terms of $\nabla{*}F$ antisymmetrized—the two (groups of) Maxwell equations can now be written as the single equation

$$j^{AA\cdot} \det \Theta = 2\check{\Theta}^{aBA\cdot}(\nabla_a \Phi_{BC} + 2iA_a \Phi_{BC})\boldsymbol{\epsilon}^{AC} - i\boldsymbol{\epsilon}^{abcd}\boldsymbol{\epsilon}_{\lambda\mu\nu\rho}\Theta_a^{AA\cdot}\Theta_b^{BB\cdot}T^h{}_{cd}\Theta_h^{CC\cdot}\Phi_{BC}\overline{\boldsymbol{\epsilon}}_{B\cdot C\cdot}.$$

If the torsion vanishes and $\nabla\boldsymbol{\epsilon} = 0$, then we recover the usual two-spinor form of the Maxwell equations (Ref. 1, formula 5.1.52).

Consider the relation between Φ and A derived as one as the field equations. In the nondegenerate case it can be written

$$\Phi_{AB} = \tfrac{1}{2}\nabla_a A_b \overline{\boldsymbol{\epsilon}}^{A\cdot B\cdot}(\Theta^a_{AA\cdot}\Theta^b_{BB\cdot} + \Theta^a_{BA\cdot}\Theta^b_{AB\cdot}) + \tfrac{1}{2}T^c{}_{ab}A_c\overline{\boldsymbol{\epsilon}}^{A\cdot B\cdot}\Theta^a_{AA\cdot}\Theta^b_{BB\cdot}.$$

Assuming the Lorenz gauge

$$\nabla_a A_b \overline{\boldsymbol{\epsilon}}^{A\cdot B\cdot}\boldsymbol{\epsilon}^{AB}\Theta^a_{AA\cdot}\Theta^b_{BB\cdot} = 0$$

this becomes

$$\Phi_{AB} = \overline{\boldsymbol{\epsilon}}^{A\cdot B\cdot}\Theta^a_{AA\cdot}\Theta^b_{BB\cdot}(\nabla_a A_b + \tfrac{1}{2}T^c{}_{ab}A_c).$$

B. Schwinger Lagrangian

When a gauge is fixed, and A is made into a true 1-form, one can modify $\mathscr{L}_{\text{em}}$ into a general relativistic ''Schwinger'' Lagrangian:

$$\mathscr{L}_S = -\tfrac{1}{4}\Theta^{(2)}(A\wedge\nabla\widetilde{F} + dA\otimes\widetilde{F}) + \tfrac{1}{4}\widetilde{F}\cdot\widetilde{F}\,\eta,$$

or

$$\mathscr{L}_S = \tfrac{1}{8}\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\Theta^{\nu}_{c}\Theta^{\rho}_{d}(\nabla_a\widetilde{F}^{\lambda\mu}A_b - \partial_a A_b\widetilde{F}^{\lambda\mu}) + \tfrac{1}{4}\widetilde{F}^{\alpha\beta}\widetilde{F}_{\alpha\beta}\ \det\Theta.$$

As before, the $\widetilde{F}$ component of the field equations gives the equation relating $\widetilde{F}$ to A,

$$2\widetilde{F}_{\lambda\mu}\ \det\Theta - (\partial_a A_b + \tfrac{1}{2}A_h T^h{}_{ab})\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\Theta^{\nu}_{c}\Theta^{\rho}_{d} = 0.$$

So, in the nondegenerate case, $F = 2dA + A\lrcorner T$. The A component, coupling to the Dirac current, is

$$\tfrac{1}{8}\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}[\nabla_b\widetilde{F}^{\lambda\mu}\Theta^{\nu}_{c}\Theta^{\rho}_{d} + \partial_a(\widetilde{F}^{\lambda\mu}\Theta^{\nu}_{c}\Theta^{\rho}_{d})].$$

In the nondegenerate case this reads

$$-\tfrac{1}{2}(\nabla_h F^{ah} + \partial_b F_{cd}\epsilon^{abcd})\det\Theta,$$

which is different from the usual divergence of F because of the torsion. Also, we have the $\widetilde{\Gamma}$ component

$$(\mathscr{E}_S)^{a}_{\lambda\mu} = \tfrac{1}{8}\epsilon^{abcd}A_b\Theta^{\nu}_{c}\Theta^{\rho}_{d}\widetilde{F}^{\alpha}{}_{[\lambda}\epsilon_{\mu]\alpha\nu\rho},$$

which enters the torsion equation, and the Θ component

$$(\mathscr{E}_S)^{c}_{\nu} = \tfrac{1}{4}\epsilon^{abcd}\epsilon_{\lambda\mu\nu\rho}\Theta^{\rho}_{d}(\nabla_a\widetilde{F}^{\lambda\mu}A_b - \partial_a A_b\widetilde{F}^{\lambda\mu}) + \tfrac{1}{4}\widetilde{F}^{\alpha\beta}\widetilde{F}_{\alpha\beta}\breve{\Theta}^{c}_{\nu},$$

which enters the gravitational equation.

C. The gauge as a dynamical object

Rather than fixing a gauge, we might assume that the gauge itself is a field of the theory. We denote it by ω and write its local expression as $\omega = e^{-2it}\epsilon$; locally we may regard the function t: $\mathbf{M}\to\mathbb{R}$ as the dynamical gauge field. Moreover we consider a further field

$$X:\mathbf{M}\to\mathbf{L}^{-3}\otimes\mathbf{H},$$

whose meaning will result from the field equations. Write the Lagrangian in terms of $w := \omega^{-1}$ as

$$\mathscr{L}_{\text{gauge}} = -\frac{i}{2}\,\breve{\Theta}[X\lrcorner(\nabla w\otimes\overline{w} - \nabla\overline{w}\otimes w)] + qX^2\eta,$$

where the (covariantly) constant $q\in\mathbf{L}^2$ is needed in order to have a conformally invariant form. The coordinate expression is

$$\mathscr{L}_{\text{gauge}} = 2(\partial_a t - A_a)X^{\lambda}\breve{\Theta}^{a}_{\lambda} + q\,\eta_{\lambda\mu}X^{\lambda}X^{\mu}\ \det\Theta.$$

The X and t components of the field equations are, respectively,

$$qX_{\lambda}\ \det\Theta + (\partial_a t - A_a)\breve{\Theta}^{a}_{\lambda} = 0, \qquad \partial_a(X^{\lambda}\breve{\Theta}^{a}_{\lambda}) = 0.$$

In the nondegenerate case these become

$$qX_{\lambda} = A'_a\Theta^{a}_{\lambda}, \quad A' := A - \mathrm{d}t,$$

$$(\nabla_a X^\lambda + T^b{}_{ab} X^\lambda)\,\breve{\Theta}^a_\lambda = 0.$$

The first equation then determines qX as the spinor representative of A', the 1-form determined by the gauge ω, while the second is the Lorenz condition modified by the torsion.

The Lagrangian $\mathscr{L}_{\text{gauge}}$ contributes also to the A and Θ components of the total field equations, so we have also a ''gauge current'' and a ''gauge stress-energy tensor,'' whose possible physical meanings are to be discussed.

ACKNOWLEDGMENTS

This work has been supported by Italian MURST (national and local funds) and by GNFM of the Consiglio Nazionale delle Ricerche.

[1] R. Penrose and W. Rindler, *Spinors and Space-Time. I: Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge, 1984).
[2] R. Penrose and W. Rindler, *Spinors and Space-Time. II: Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge, 1988).
[3] A. Jadczyk, ''Symmetry of Einstein–Yang–Mills system and dimensional reduction,'' JGP **1**, 97–126 (1984).
[4] C. Canarutto and A. Jadczyk, ''Two-spinors and Einstein–Cartan–Maxwell–Dirac fields,'' Nuovo Cimento B **113**, 49–67 (1997).
[5] This contraction, defined in such a way to respect the usual conventions in two-spinor literature, corresponds to half standard tensor-algebra contraction and 1/4 standard exterior-algebra contraction.
[6] In the nondegenerate case, this turns out to be essentially the Lagrangian used in the ADM formalism. Compared to the Schwinger Lagrangian, it lacks a term of the kind $A\cdot\delta F$, which however is not gauge-invariant. We discuss this point in Sec. VII.
[7] F. Gronwald and F. W. Hehl, ''On the Gauge Aspects of Gravity,'' *Proceedings of the 14th Course of the School of Cosmology and Gravitation on Quantum Gravity, Erice, Italy, May 1995*, edited by P. G. Bergamann, V. de Sabbata, and H.-J. Treder (World Scientific, Singapore, 1996).
[8] F. W. Hehl, J. D. McCrea, E. W. Mielke, and Y. Ne'eman, ''Metric-affine gauge theory of gravity: Field equations, Noether identities, world spinors, and breaking of dilaton invariance,'' Phys. Rep. **258**, 1–171 (1995).
[9] A. Jakubiec and J. Kijowski, ''On Interaction of the Unified Maxwell–Einstein Field with Spinorial Matter,'' Lett. Math. Phys. **9**, 1–11 (1985).
[10] T. Regge, ''The group manifold approach to unified gravity,'' in *Relativity, groups and topology II*, Les Houches 1983, edited by B. S. DeWitt and R. Stora (North-Holland, Amsterdam, 1984), pp. 933–1005.
[11] M. Ferraris and J. Kijowski, ''Unified Geometric Theory of Electromagnetic and Gravitational Interactions,'' Gen. Relativ. Gravit. **14**, 37–47 (1982).
[12] From a global point of view, the two approaches are not equivalent.

Group-theoretical quantization of 2+1 gravity in the metric-torus sector

P. Hájíček
Institute for Theoretical Physics, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

(Received 17 June 1997; accepted for publication 5 August 1997)

A symmetry based quantization method of reparametrization invariant systems is described; it will work for all systems that possess complete sets of perennials whose Lie algebras close and those that generate sufficiently large symmetry groups. The construction leads to a quantum theory including a Hilbert space, a complete system of operator observables, and a unitary time evolution. The method is applied to the $2+1$ gravity. The paper is restricted to the metric-torus sector, zero cosmological constant Λ and it makes strong use of the so-called homogeneous gauge; the chosen algebra of perennials is that of Martin. Two frequent problems are tackled. First, the Lie algebra of perennials does not generate a group of symmetries. The notion of group completion of a reparametrization invariant system is introduced so that the group does act; the group completion of the physical phase space of our model is shown to add only some limit points to it so that the ranges of observables are not unduly changed. Second, a relatively large number of relations between observables exists; they are transferred to the quantum theory by the well-known methods of Kostant and Kirillov. In this way, a uniqueness of the physical representation of some extension of Martin's algebra is shown. The Hamiltonian is defined by a systematic procedure due to Dirac; for the torus sector, the result coincides with that by Moncrief. The construction may be extended to higher genera and nonzero Λ of the $2+1$ gravity, because some complete sets of perennials are well-known and there are no obstructions to the closure of the algebra. © *1998 American Institute of Physics.* [S0022-2488(98)00409-5]

I. INTRODUCTION

The $(2+1)$-dimensional gravity is a popular model. It has been utilized as a laboratory for studies of specific problems in quantum gravity (for reviews, see Refs. 1 and 2). Our aim is to illustrate a symmetry based method of quantization of the so-called *reparametrization invariant systems* (RIS); a RIS is a constrained system whose dynamics is completely determined by its constraints (the value of the Hamiltonian is zero). For this purpose, $2+1$ gravity seems particularly interesting, because it is a generally covariant RIS and the structure of its constraints is very similar to that of the general relativity.

The (technical) starting point of the method is a choice of a set of functions on the phase space that satisfy three conditions. First, they should have vanishing Poisson brackets with all constraints. Such quantities have been called ''first-class'' by Dirac[3] and they are often called ''observables'' today. The name ''observables'' is, however, not justified in the case of RIS's. This difficulty, which is connected to the problem of time, has been first noticed by Kuchař,[4] who has introduced a special name, *perennials*, for the first-class quantities of RIS's. An attempt at a mathematical formulation of Kuchař's ideas can be found in Ref. 5, where a construction of observables from perennials is described that necessarily makes use of some additional (time) structures (cf. Sec. II). The second condition on the set is that it separates the classical solutions. This means that for two different solutions there is at least one function from the set that takes on different values at these solutions. This property is called *completeness*: it was first introduced by Bergmann[6] and it was later used by Ashtekar[7] as one of the basic properties within the algebraic quantization method. The third condition on the set is that it forms a closed algebra with respect to linear combinations and Poisson brackets. Thus, it will be a Lie algebra **g**.

 © 1998 American Institute of Physics

However, availability of perennials in the general relativity, or even their existence, has been questioned.[4] Indeed, one *can* invent RIS's that possess no perennials. However, these are mathematical constructs whose physics is strange; they cannot be reduced, even locally, and their physical phase spaces are not manifolds. It turns out that physically reasonable systems do possess complete systems of perennials (for proof, see Ref. 5). True, perennials of a special kind, for example those that are linear in momenta or are local in certain sense are already shown to be absent in the general relativity.[8,9] We shall show, however, how these proofs are transcended for the $2+1$ gravity model (to which the proof methods of Kuchař, Anderson, and Torre may be applicable). The third condition, that the perennials form a Lie algebra, might also be a source of problems; there are finite dimensional symplectic manifolds that do not admit a complete finite set of perennials whose algebra closes. However, it seems that such systems are again rather artificial with strange physics and that they also would be very difficult to quantize by any existing method.

The complete algebra of perennials plays a twofold role in the theory. First, as already mentioned, observables can be constructed from them. We assume that these observables comprise the most important and directly measurable properties of the system—this condition, although a little vague from the mathematical point of view, should influence the choice of the algebra **g**. Second, a group G of symmetries is generated by **g**, if some conditions are satisfied. Such a group, if it exists, can strongly simplify the task of finding suitable representation of the algebra **g** (that is, quantizing the system). The corresponding methods are those of the so-called *group quantization* (see Ref. 10 for a review); a modification of the methods suitable for the RIS's has been suggested by Rovelli.[11] The existence of the group G can however also help to solve the problem of time that afflicts the quantization of RIS's. The idea of this use for G stems from an old paper by Dirac;[12] in this paper, a time evolution has been constructed for a system of relativistic particles on the Minkowski spacetime. Dirac's ideas have been extended to general RIS's and developed to a coherent theory in Refs. 5 and 13. Although already published, this theory is far from being well known. We give, therefore, a short pedagogic exposition in Sec. II to make the paper self-contained.

The group G of symmetries is obtained from the algebra **g** in two steps. First, each Lie algebra determines a unique (simply connected) abstract Lie group G. Second, the Hamiltonian vector fields of the elements of the algebra determine an action of G on the phase space, if the vector fields are complete—we shall meet a prominent example of incomplete Hamiltonian vector fields in the $2+1$ gravity model. The action is then unique; let us call it *Hamiltonian action* of G. In some cases, the center of G will contain a nontrivial subgroup G_c, whose elements act trivially. Two cases must be distinguished: isolated elements of G in G_c and Lie subgroups of G in G_c. One can simplify G by taking the quotient with respect to the isolated elements, but not with respect to the Lie part of G_c.[10]

The $2+1$ gravity model can be considered as a dynamics of an ISO(1,2) affine connection of a three-dimensional manifold M.[14] If some conditions are satisfied, then the affine connection defines a Lorentzian metric on M and this property is preserved by the dynamics (we shall cut out the singularities!). We assume that Cauchy surface Σ is the torus $(S^1\times S^1)$ and that the metric is well-defined; in such a case, we speak of the *metric-torus sector*. This sector alone gives a solvable but unexpectedly interesting model so that all the general points above can be illustrated in a rather nontrivial way. We shall limit ourselves to this sector and we will work in a particular gauge, the so-called homogeneous gauge, in which the three-metric depends only on the time coordinate. Let us remark that there is no reason why our method should not work for higher genera. One can try to use the loop variables described in Refs. 15 or 16; the topology of the physical phase spaces for all higher genera is contractible and there are global Darboux coordinates; thus, there are no obstructions to the existence of a complete system of functions with a closed algebra.

In the present paper, we will choose the complete set of perennials that has been published by Martin.[15] The perennials are directly related to the topological degrees of freedom of the system and can be expressed as a kind of Wilson loop variable. For the metric-torus sector, Moncrief observed that Martin's perennials form the six-dimensional algebra **g** isomorph to iso(1,2). We shall introduce new canonical coordinates that are adapted to Martin's perennials; the constraint becomes formally the mass-shell condition for a rest-mass-zero relativistic particle. Thus, further four perennials can be immediately written down; together with the old ones, they form a ten-

dimensional algebra so(2,3). The corresponding group G is the conformal group of three-dimensional flat spacetime; G is isomorph to SO(2,3).

The problem of existence of the Hamiltonian action for G is nontrivial. Our calculation will reveal that only a subgroup $G_0 \subset G$ with the structure of $\mathrm{SO}(1,2)\times\mathbf{R}$ acts on the phase space of the system. It is easy to observe, however, that the phase space can be extended so that the whole group G has the Hamiltonian action on the extended space. We call such extensions *group completions*. A minimal group completion is unique under some quite general conditions. It turns out that the minimal group completion of the physical phase space consists, in our case, of adding ''relatively few points'' in such a way that the ranges of observables are not changed except for adding some boundary points to them.

The metric-torus sector has two degrees of freedom. Thus, a ten-dimensional algebra like $\mathbf{g}$ will exhibit six independent relations. Using the Kostant-Kirillov method, we find that the group G does not possess any physical representation. This has to do with Van Hove theorem: the physical phase space is too small and the algebra $\mathbf{g}$ of functions is too large to be represented without deformation. We find, however, that one of its maximal subgroups, G_1, which is seven-dimensional, and isomorph to $(\mathrm{SO}(1,2)\times\mathbf{R})\otimes_s\mathbf{R}^3$, possesses a unique physical representation and we calculate the form of the operators representing $\mathbf{g}_1$. Here, ''$\otimes_s$'' denotes the semidirect product of groups. Three independent classical relations for G_1 can be written down in terms of (generalized) Casimir operators.

For the construction of time evolution à la Dirac, we can use only the four-dimensional subgroup G_0; then the Hamiltonian action of G_0 on the constraint surface (and so on the classical spacetimes) provides the interpretation of the corresponding unitary transformations in the Hilbert space. It turns out that this action ''goes in time direction'' and so a time evolution can be constructed. The candidates for the Hamiltonian that generates everywhere time evolution towards future form a three-dimensional family. Most of these operators are unbounded from below. Thus, the condition that an operator generates time evolution towards the future does not necessarily guarantee that the operator has a non-negative spectrum. However, there is exactly one Hamiltonian that *is* bounded from below (and it is even nonnegative). The positive Hamiltonian coincides with the Hamiltonian written down by Moncrief[17] and the time coincides with the constant mean external curvature.

Some surprise is that the quantum mechanics we have constructed are not equivalent to the ''ordinary'' quantum mechanics of the rest-mass-zero free particle in three-dimensional Minkowski spacetime, in spite of the fact that the algebra of observables is so(2,3)-like for the particle, and that we managed to introduce new variables in which the constraint coincides with the ordinary mass-shell condition for such a particle. The explanation is that the global structure of the torus configuration space is very different from that of the particle (which is the three-dimensional Minkowski spacetime): the former is only a subset of the latter, namely, the inside of the light cone of the origin. The points outside of the light cone correspond to timelike two-surfaces evolving in a spacelike direction (the signature of the spacetime remains $+1$).

The plan of the paper is as follows. In Sec. II, we briefly describe the mathematical apparatus that has turned out advantageous for the study of RIS's, their observable properties, and their time evolution. No detail and proofs are given, because these can be found in already published papers.[5,13] In Sec. III, we list our starting assumptions and equations concerning the $2+1$ gravity model. They are mostly taken over from Refs. 17 and 18, where more detail can be found. One nontrivial but plausible assumption is that the so-called homogeneous gauge can be chosen in which the model becomes finite-dimensional.[18,19]

In Sec. IV, we study the problem of action of the group, define the group completion of a RIS and the weak action of the group. We derive the group completion of the model and prove that the group completed physical phase space contains the original one as an open dense subset. We define the action of G on the completed physical phase space, give the form of all observables obtained from the algebra $\mathbf{g}$ on the physical phase space and list all independent relations.

Section V describes an application of Dirac's time evolution idea to our model. We find that the dynamics is much more unique than in the case of free relativistic particle studied by Dirac (who found three inequivalent ''forms of relativistic dynamics''); we find only one ''form.'' Finally, in Sec. VI, after a brief description of the Kostant-Kirillov method, we derive the physical representation.

II. EXAMPLE: THE RELATIVISTIC PARTICLE

We consider a free relativistic particle of mass $m>0$ on the four-dimensional Minkowski space–time with coordinates x^μ and the metric $\eta_{\mu\nu}=\mathrm{diag}(-1,1,1,1)$. This is one of the systems studied originally by Dirac.[12] Let the conjugate momenta be p_μ and the constraint be $\mathscr{H}=(1/2)\times(p\cdot p+m^2)$. We shall use the abbreviation $A\cdot B:=\eta_{\mu\nu}A^\mu B^\nu$ throughout the paper.

The manifold $\mathbf{R}^8$ with the coordinates x^μ and p_μ, and with the symplectic form $\tilde{\Omega}=dp_\mu\wedge dx^\mu$ will be called extended phase space and denoted by $\tilde{\Gamma}$. The submanifold Γ of $\tilde{\Gamma}$ defined by $\mathscr{H}=0$ and $p_0<0$ is the *constraint surface*. The function $\mathscr{H}$ is the so-called *super-Hamiltonian*. It defines Γ, generates reparametrizations along the particle trajectories, and generates the dynamics,

$$\dot{x}^\mu=\mathscr{N}\{x^\mu,\mathscr{H}\},\quad \dot{p}_\mu=\mathscr{N}\{p_\mu,\mathscr{H}\}, \tag{1}$$

where $\mathscr{N}$ is an arbitrary function, the so-called *lapse*. The arcs that are determined by maximal solutions of Eqs. (1) will be called *c-orbits*. A *c*-orbit will be typically denoted by γ. *c*-orbits represent classical solutions. The quotient space $\bar{\Gamma}=\Gamma/\gamma$ will be called *physical phase space*; we will assume that $\bar{\Gamma}$ is a manifold. In this case, there is a symplectic form $\bar{\Omega}$ on $\bar{\Gamma}$ that is uniquely determined by $\tilde{\Omega}$. In our example, Γ is seven-dimensional, *c*-orbits are one-dimensional, so $\bar{\Gamma}$ is six-dimensional. The dimension of the physical space is the double of the physical degrees of freedom.

An important notion is that of *transversal surface*. This is any submanifold Γ_i of Γ such that each *c*-orbit intersects Γ_i exactly once and in a transversal direction (any vector at an intersection point that is simultaneously tangential to the *c*-orbit and to the transversal surface is necessarily the zero vector). The importance of transversal surfaces for the description of time evolution in relativistic theories has been recognized by Dirac.[12] Each transversal surface Γ_i carries a unique symplectic form Ω_i, the pull back of $\tilde{\Omega}$ to Γ_i. The symplectic space (Γ_i,Ω_i) can be identified with $(\bar{\Gamma},\bar{\Omega})$ by the map that sends each *c*-orbit γ from $\bar{\Gamma}$ to the intersection point of γ with Γ_i. An example of transversal surface, which we denote by Γ_0, is given by the equations $x^0=0$ and $\mathscr{H}=0$. As coordinates on Γ_0, the functions x^1, x^2, x^3, p_1, p_2, and p_3 can be chosen; in these coordinates, $\Omega_0=dp_k\wedge dx^k$. As we can see, Γ_0 defines a particular *time instant*, namely $x^0=0$. This is a general property of transversal surfaces; for example, in the general relativity, a transversal surface defines a unique Cauchy surface in any generic space–time solution (i.e., the solution that admits no symmetry).

The Poincaré group (in fact, only the componet of identity thereof) ISO (1,3) acts on $\tilde{\Gamma}$ in the usual way, leaving both $\tilde{\Omega}$ and Γ invariant. Such transformations in $\tilde{\Gamma}$ are called *symmetries*. The action is generated by ten functions p_μ, $J_k:=\epsilon_{klm}x_l p_m$ and $K_k:=x_k p_0-x_0 p_k$, $k=1,2,3$ (the indices are lowered and raised by the metric). The functions have vanishing Poisson brackets with $\mathscr{H}$ (as $\mathscr{H}$ is an invariant), so they are perennials. They form a Lie algebra that is isomorphic to iso(1,3) with respect to linear combinations and Poisson brackets, and the set separates *c*-orbits. Thus, we have an example of a complete Lie algebra $\mathbf{g}$ of perennials.

In this case, the group that is determined by $\mathbf{g}$ is $\mathrm{Sl}(2,\mathbf{C})\otimes_S\mathbf{R}^4$. The Hamiltonian action of the group is just the above action of ISO(1,3): The center of Sl(2,**C**) contains only one element that is different from identity, which acts trivially, so we can restrict ourselves to the group ISO(1,3).

Important general properties of symmetries are (see Refs. 11 and 13) (a) a symmetry sends *c*-orbits onto *c*-orbits and (b) it sends transversal surfaces onto transversal surfaces. The property (b) is crucial to the construction of dynamics á la Dirac;[12] by the transformations of the group G, a time instant is sent into another time instant. Dirac also proposed to choose maximally symmetric (with respect to G) transversal surfaces. Such a choice not only minimizes the number of ''different time instants,'' but also simplifies the operator observables.[12] There are three inequivalent maximally symmetric transversal surfaces for the relativistic particle; (a) the spacelike plane $x^0=0$, (b) the pair of hyperboloids $x\cdot x=1$, and (c) the null plane $x^0-x^1=0$. This leads to Dirac's ''three forms of dynamics.''

At first sight, it seems that perennials *are* observables. Thus, p_k are the momenta and J_k are the angular momenta of the particle. Such an association is, however, more difficult for p_0 and K_k. It is clearly not reasonable to put $-p_0$ equal to the energy, because only $\sqrt{\mathbf{p}^2+m^2}$ will lead

to the observed spectrum of the particle. Of course, $\sqrt{\mathbf{p}^2+m^2}$ is the value of $-p_0$ at the constraint surface Γ. More problems are encountered, if we try to interpret K_k as an observable. We can utilize p_μ together with K_k to form the functions

$$X_t^k := \frac{K_k}{p_0} + \frac{p_k}{p_0}\, t$$

for any constant t: as functions of perennials, X_t^k are themselves perennials. They can be interpreted as coordinates of the particle at the time $x^0=t$. It turns out, that this is a quite general way of forming observables from perennials; for it to work, some ''time structure'' is necessary. Here, we have used the (standard) family of transversal surfaces Γ_t defined by the equations $x^0=t$, as well as the family of maps generated by p_0 and sending Γ_t to $\Gamma_{t'}$ for each pair (t,t'). One easily verifies that

$$\frac{\partial X_t^k}{\partial t} + \{X_t^k, p_0\} = 0.$$

The class $\{X_t^k | t \in \mathbf{R}\}$ can, therefore, be considered as a Heisenberg observable, and so an observable turns out to be a class of perennials. The apparently trivial relation between perennials and observables for the particular perennials p_k and J_k in this case (the functions in the class are t-independent and coincide in form with the original perennials) is due to the fact that these perennials are connected in a special way to the standard time structure (represented by Γ_t and p_0) that is determined in the phase space by an inertial frame; p_k and J_k have vanishing Poisson brackets with p_0 and they leave (via Poisson bracket) each time instant Γ_t invariant. A systematic theory is given in Ref. 5.

An important technical tool for the present paper will be *projection to a transversal surface*. One can project perennials and symmetries. The projection o_1 of a perennial o to the transversal surface Γ_1 is a function on Γ_1 defined as the pullback of o to Γ_1: $o_1 := o|_{\Gamma_1}$. This projection preserves three operations: linear combinations, multiplications of functions, and Poisson brackets (of perennials). In particular, for the Poisson brackets, we have $\{o_1, o_1'\}_1 = \{o, o'\}$, where $\{\cdot,\cdot\}_1$ is the Poisson bracket in (Γ_1, Ω_1), $\{\cdot,\cdot\}$ that of $(\tilde{\Gamma}, \tilde{\Omega})$, o_1 the projection of o and o_1' that of o' to Γ_1 (for the proof, see Ref. 13). Projections of perennials to a given transversal surface can (sometimes!) be considered as Schrödinger observables (for more detail, see Ref. 5).

Similar to perennials, symmetries can also be projected. Let ϕ be a symmetry and $p \in \Gamma_1$. Then, the projection $\phi_1 : \Gamma_1 \to \Gamma_1$ of ϕ to Γ_1 is defined by $\{\phi_1(p)\} := \gamma_{\phi(p)} \cap \Gamma_1$, where $\{a\}$ is the set with the element a and γ_q is the c-orbit through the point q. One can show that ϕ_1 preserves Ω_1 as well as the group multiplication of symmetries; thus, the projection of a group of symmetries is a group of symplectic maps.[5] Moreover, if o generates a one-dimensional group of symmetries ϕ_t, then o_1 generates a one-dimensional group of symplectic maps ψ_t and it holds that $\psi_t = \phi_{t1}$ for all t, where ϕ_{t1} is the projection of ϕ_t to Γ_1.[13]

Finally, let Γ_1 and Γ_2 be two transversal surfaces, $p \in \Gamma_1$ and ρ: $\Gamma_1 \to \Gamma_2$ defined by $\{\rho(p)\} := \gamma_p \cap \Gamma_2$. Then, the pull back of Ω_2 by ρ is Ω_1 and the pull back of o_2 is o_1 for any perennial o with projection o_1 to Γ_1 and o_2 to Γ_2. Thus, ρ realizes the equivalence of all Γ_i's.[13]

Another important notion that we can illustrate with the relativistic particle model is that of *relation*. The projections of the ten perennials p_μ, J_k, and K_k to the transversal surface Γ_0 are linearly independent. However, there must be four functional relations between them, because Γ_0 is only six-dimensional. There cannot be more independent relations, as the functions form a complete system. The relations can be written as

$$\mathbf{p}^2 = -m^2, \quad \epsilon^{\mu\nu\rho\sigma} p_\nu J_{\rho\sigma} = 0,$$

where $J_{\rho\sigma} := \epsilon_{\rho\sigma\mu\nu} x^\mu p^\nu$. These relations have a close connection to the values of Casimir elements for this particular representation of the Poincaré group. Pohlmayer[20] has discussed the general case.

In the quantum version, we shall try to preserve these relations so that they become similar relations between operators (that may be deformed: some additional terms proportional to $\hbar$ may

appear). There also can be spectral conditions like $p_0<0$, that should be satisfied in the quantum theory. A unitary representation of the group G that preserves the relations between its generators and that satisfies the spectral condition is called *physical representation*.

III. THE HOMOGENEOUS GAUGE

In this section, we return to the $2+1$ gravity and briefly summarize our starting assumptions and equations. More detail can be found in Refs. 17 and 18. We shall consider only the metric-torus sector of the $2+1$ gravity system. In arbitrary coordinates x^1 and x^2 on arbitrary spacelike surfaces $t=\text{const}$ with the manifold structure Σ, the metric of the spacetime M has the form

$$ds^2=-N^2dt^2+g_{ab}(dx^a-N^adt)(dx^b-N^bdt)$$

and the ADM action for the model reads

$$S=\int_{\mathbf{R}}dt\int_{\Sigma}d^2x\left(\pi^{ab}\frac{\partial g_{ab}}{\partial t}-\mathcal{N}\mathcal{H}-\mathcal{N}^a\mathcal{H}_a\right),$$

where

$$\mathcal{H}=-\frac{1}{\sqrt{g}}[\pi^{ab}\pi_{ab}-(\pi^a_a)^2]+\sqrt{g}\mathcal{R},$$

$$\mathcal{H}_a=-2\nabla_b\pi^b_a.$$

∇_a is the covariant derivative associated with the metric g_{ab} and $\mathcal{R}$ is the curvature scalar of the two-surfaces $t=\text{const}$.

The analysis of the model simplifies enormously, if one can choose the so-called ''homogeneous gauge.'' It is the choice of coordinates such that the fields g_{ab} and π^{ab} are independent of x^1 and x^2.[17,19] A rigorous proof that such coordinates exist for each classical solution of the model has not yet been published. If each solution, however, admits at least one Cauchy surface of constant mean curvature (CMC), which seems to be a very plausible conjecture, then the existence can be shown.[18] The existence of the CMC surface has been proven by Andersson for genera higher than 1 and there are some ideas even for genus 1.[21] We will assume, that there is such a Cauchy surface; in any case, one can consider this as a part of the definition of the system, and if the conjecture is invalid, then this system will not be completely equivalent to the $2+1$ gravity.

In the homogeneous gauge, the metric can be taken in the form[17]

$$ds^2=-N(t)^2dt^2+e^{2\mu(t)}(dx^1)^2+e^{2\nu(t)}(dx^2+\beta(t)dx^1)^2. \tag{2}$$

A straightforward calculation then leads to the action

$$S=\int dt(p_\mu\dot{\mu}+p_\nu\dot{\nu}+p_\beta\dot{\beta}-\mathcal{N}\mathcal{H}),$$

where

$$\mathcal{H}=\tfrac{1}{2}(e^{-\mu-\nu}p_\mu p_\nu-e^{\mu-3\nu}p_\beta^2).$$

Moncrief then performs two canonical transformations; the first is

$$q^1=\nu-\mu,\quad q^2=\beta,\quad q^3=\nu+\mu,$$

$$p_\mu=-p_1+p_3,\quad p_\nu=p_1+p_3,\quad p_\beta=p_2,$$

so that the super-Hamiltonian becomes

$$\mathcal{H}=\tfrac{1}{2}e^{-q^3}(p_3^2-p_1^2-e^{-2q^1}p_2^2).$$

These coordinates are advantageous for visualization of the geometry of the system. The extended phase space is $\tilde{\Gamma}=T^*\mathbf{R}^3$ with the canonical coordinates q^1, q^2, q^3, p_1, p_2, and p_3; the meaning of q^3 and p_3 is

$$q^3=\log\sqrt{g},\quad p_3=\frac{\sqrt{g}}{\mathscr{N}}\frac{\partial q^3}{\partial t}.$$

The constraint surface Γ is the light cone in the momentum space

$$p_3^2-p_1^2-e^{-2q^1}p_2^2=0.$$

The $p_3>0$ half of the cone represents expanding, the $p_3<0$ half contracting, and the cusp $p_3=0$ represents the static tori—solutions with higher symmetry. We shall adhere to the convention that the coordinate t on the tori spacetimes is future oriented for $\mathscr{N}>0$.

The structure of the constraint surface Γ and of the physical phase space $\overline{\Gamma}=\Gamma/\gamma$ is spoilt by the points of higher symmetry.[22] The c-orbits with $p_3\neq 0$ are curves, those with $p_3=0$ are just points. As the static solutions form a set of measure zero, we can cut them away. Thus, we consider only that part of the system that satisfies the condition

$$p_3\neq 0.$$

The new system will have a disconnected phase space $\tilde{\Gamma}'=\tilde{\Gamma}'_+\cup\tilde{\Gamma}'_-$, a constraint surface $\Gamma'=\Gamma'_+\cup\Gamma'_-$ and a physical phase space $\overline{\Gamma}'=\overline{\Gamma}'_+\cup\overline{\Gamma}'_-$, where $\tilde{\Gamma}'_+:=\{x\in\tilde{\Gamma}|p_3>0\}$, $\Gamma'_+:=\Gamma\cap\tilde{\Gamma}'_+$, and $\overline{\Gamma}'_+:=\Gamma'_+/\gamma$ (similarly for $p_3<0$).

After this truncation, Moncrief performs the second transformation,

$$T=\ln|p_3|-q^3,\quad p_T=-p_3,$$

the other variables remaining the same; the super-Hamiltonian then reads

$$\mathscr{H}=-\frac{e^T}{2p_T}(p_T^2-p_1^2-e^{-2q^1}p_2^2). \tag{3}$$

The meaning of the variable T is given by

$$g^{ab}K_{ab}=\frac{1}{2\mathscr{N}}g^{ab}\frac{\partial g_{ab}}{\partial t}=\epsilon e^T,$$

where $\epsilon=\pm 1$ and the sign is determined by $\epsilon=\text{sign } p_3=-\text{sign } p_T$. Thus, ϵ is just the sign of the CMC of the surface $t=\text{const}$.

Martin's[15] constants of motion (perennials) are in these coordinates given by[17]

$$C_1=-(\epsilon/2)e^T\{[e^{-q^1}+(q^2)^2e^{q^1}](p_T+p_1)-2e^{-q^1}p_1+2q^2e^{-q^1}p_2\}, \tag{4}$$

$$C_2=-(\epsilon/2)e^T[e^{q^1}(p_T+p_1)], \tag{5}$$

$$C_3=-(\epsilon/2)e^T[q^2e^{q^1}(p_T+p_1)+e^{-q^1}p_2], \tag{6}$$

$$C_4=\tfrac{1}{2}\{[e^{-2q^1}-(q^2)^2]p_2+2q^2p_1\}, \tag{7}$$

$$C_5=\tfrac{1}{2}p_2, \tag{8}$$

$$C_6=p_1-q^2p_2. \tag{9}$$

They can be expressed as a kind of loop integrals by means of the original fields $g_{ab}(x)$ and $\pi^{ab}(x)$.[15]

Moncrief observed that the Poisson brackets of the variables P_μ and J^μ defined by

$$P_0 := \tfrac{1}{2}(C_1 + C_2), \quad P_1 := \tfrac{1}{2}(C_1 - C_2), \quad P_2 := C_3, \tag{10}$$

and

$$J^0 := -C_4 + C_5, \quad J^1 := -C_4 - C_5, \quad J_2 = -C_6,$$

form a Lie algebra isomorphic to iso(1,2). If we introduce the abbreviations

$$A := A^\mu P_\mu, \quad C := C_\mu J^\mu,$$

then

$$\{A, A'\} = 0, \quad \{A, C\} = (\varepsilon^{\rho\mu\nu} A_\mu C_\nu) P_\rho, \quad \{C, C'\} = (\varepsilon_{\rho\mu\nu} C^\mu C^\nu) J^\rho. \tag{11}$$

Here, we raise and lower the indices of X^μ and P_μ by the Minkowskian three-metric diag$(-1,1,1)$, $\varepsilon^{\rho\mu\nu}$ and $\varepsilon_{\rho\mu\nu}$ are the usual antisymmetric symbols ($\varepsilon_{\rho\mu\nu}$ is not $\varepsilon^{\rho\mu\nu}$ with lowered indices).

Formulas (4)–(9) imply the following four equations:

$$C_1 C_2 - C_3^2 = \tfrac{1}{4} e^{2T} (p_T^2 - p_1^2 - e^{-2q^1} p_2^2), \tag{12}$$

$$\{C_4, \mathscr{H}'\} = \{C_5, \mathscr{H}'\} = \{C_6, \mathscr{H}'\} = 0,$$

where

$$\mathscr{H}' = p_T^2 - p_1^2 - e^{-2q^1} p_2^2.$$

Thus, all C's are perennials.

There also are some discrete symmetries that originate from the nonuniqueness of the metric (2) for a given torus geometry (class group transformations, see e.g., Ref. 23). We shall not discuss the question whether these transformations are to be considered as symmetries or as gauge transformations. If $\mathbf{X}_1$ and $\mathbf{X}_2$ form a basis of a lattice in $\mathbf{E}^2$ defining the torus, then the metric has the form

$$g_{11} = \mathbf{X}_1 \cdot \mathbf{X}_1, \quad g_{12} = \mathbf{X}_1 \cdot \mathbf{X}_2, \quad g_{22} = \mathbf{X}_2 \cdot \mathbf{X}_2,$$

where $\mathbf{a} \cdot \mathbf{b}$ denotes the scalar product of the vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbf{E}^2$. The following two transformations of the basis generate the whole group of the discrete transformations (''large diffeomorphisms''):

$$\mathbf{X}_1' = \mathbf{X}_2, \quad \mathbf{X}_2' = \mathbf{X}_1,$$

so that

$$g_{11}' = g_{22}, \quad g_{12}' = g_{12}, \quad g_{22}' = g_{11}, \tag{13}$$

and

$$\mathbf{X}_1' = \mathbf{X}_1, \quad \mathbf{X}_2' = \mathbf{X}_1 + \mathbf{X}_2,$$

so that

$$g_{11}' = g_{11}, \quad g_{12}' = g_{11} + g_{12}, \quad g_{22}' = g_{11} + 2g_{12} + g_{22}. \tag{14}$$

Using Eq. (2), we obtain from Eq. (13) for μ, ν, and β,

$$e^{2\mu'} = \frac{e^{2(\mu+\nu)}}{e^{2\mu} + \beta^2 e^{2\nu}}, \quad e^{2\nu'} = e^{2\mu} + \beta^2 e^{2\nu},$$

$$\beta' = \beta \frac{e^{2\nu}}{e^{2\mu} + \beta^2 e^{2\nu}}.$$

The corresponding transformation of q^1, q^2, and q^3 is

$$e^{q'^1} = e^{q^1}[(q^2)^2 + e^{-2q^1}], \tag{15}$$

$$q'^2 = \frac{q^2}{(q^2)^2 + e^{-2q^1}}, \tag{16}$$

$$q'^3 = q^3. \tag{17}$$

This gives for the momenta

$$p_1' = \frac{(q^2)^2 - e^{-2q^1}}{(q^2)^2 + e^{-2q^1}} p_1 + \frac{2q^2 e^{-2q^1}}{(q^2)^2 + e^{-2q^1}} p_2,$$

$$p_2' = 2q^2 p_1 - [(q^2)^2 - e^{-2q^1}] p_2.$$

One can then easily verify that

$$p_1'^2 + e^{-2q'^1} p_2'^2 = p_1^2 + e^{-2q^1} p_2^2,$$

so that the super-Hamiltonian (3) is invariant.

The transformation (14) gives

$$\mu' = \mu, \quad \nu' = \nu, \quad \beta' = 1 + \beta.$$

Thus,

$$q'^1 = q^1, \quad q'^2 = 1 + q^2, \quad q'^3 = q^3, \tag{18}$$

and

$$p_1' = p_1, \quad p_2' = p_2, \quad p_3' = p_3.$$

Again, $\mathscr{H}$ is conserved.

The transformations (15)–(17) imply

$$C_1' = C_2, \quad C_2' = C_1, \quad C_3' = C_3,$$

and (18) implies

$$C_1' = C_1 + C_2 + 2C_3, \quad C_2' = C_2, \quad C_3' = C_2 + C_3.$$

These are both integral transformations with determinant 1 as one expects for loop variables, if the loops are just permuted or linearly combined.

IV. GROUP COMPLETION OF THE PHASE SPACE

A. Completion by ISO(1,2)

In this section, we will investigate which transformations are generated by the perennials P_μ and J^μ in the phase space of the system. This task will be simplified, if we use coordinates that are adapted to the perennials in the following sense. C_1, C_2, and C_3 action via Poisson brackets in the phase space can be projected to the configuration space spanned by T, q^1, and q^2, and the projections are the vector fields $\hat{C}_1$, $\hat{C}_2$, and $\hat{C}_3$ given by replacing p_T, p_1, and p_2 by $\partial/\partial T$, $\partial/\partial q^1$, and $\partial/\partial q^2$ in expressions (4)–(6). The perennials C_1, C_2, and C_3 have vanishing Poisson

brackets with each other, so $\hat{C}_1$, $\hat{C}_2$, and $\hat{C}_3$ will define a holonomous frame; the corresponding coordinates are the desired ones. Let us first simplify these vectors by the transformation

$$u=T, \quad v=q^1-T, \quad y=q^2,$$

so that

$$\hat{C}_1=-(\epsilon/2)e^{-v}[(1+y^2e^{2u+2v})\partial_u+2y\partial_y-2\partial_v],$$

$$\hat{C}_2=-(\epsilon/2)e^{2u+v}\partial_u,$$

$$\hat{C}_3=-(\epsilon/2)e^{2u+v}(y\partial_u+e^{-2u-2v}\partial_y).$$

Now, we look for pairs of independent functions that are annihilated by each of these differential operators. The method of characteristics suggests that we study integral curves of the vector fields. The integral curve of $\hat{C}_1$ is defined by

$$\dot{u}=A(1+y^2e^{2u+2v}), \quad \dot{y}=A(2y), \quad \dot{v}=A(-2),$$

where $A=-\epsilon e^{-v}/2$. Thus,

$$(1/y)\dot{y}+\dot{v}=0,$$

$$-2e^{-2u-v}\dot{u}+2ye^{v}\dot{y}+(-e^{-2u-v}+y^2e^{v})\dot{v}=0,$$

and we have

$$\hat{C}_1(ye^{v})=0, \quad \hat{C}_1(e^{-2u-v}+y^2e^{v})=0.$$

The integral curve of $\hat{C}_2$ satisfies

$$\dot{u}=-\tfrac{1}{2}e^{2u+v}, \quad \dot{y}=0, \quad \dot{v}=0,$$

hence,

$$\hat{C}_2v=0, \quad \hat{C}_2y=0.$$

Similarly for $\hat{C}_3$,

$$\dot{u}=By, \quad \dot{y}=Be^{-2u-2v}, \quad \dot{v}=0,$$

where $B=-(\epsilon/2)e^{2u+v}$. Thus,

$$-e^{-2u-2v}\dot{u}+y\dot{y}=0,$$

and

$$\hat{C}_3v=0, \quad \hat{C}_3(e^{-2u-2v}+y^2)=0.$$

The pair of independent functions we have found for each vector field determines all functions that are annihilated by the field. We can easily find three independent functions such that each vector field annihilates exactly two of them. The results can be summarized in the following table:

$\hat{C}_1$		ye^{v},	$e^{-2u-v}+y^2e^{v}$,
$\hat{C}_2$	v,	ye^{v},	
$\hat{C}_3$	v,		$e^{-2u-v}+y^2e^{v}$.

Hence, the following transformation will simplify the vector fields:

$$\xi = e^{-2u-v} + y^2 e^v, \quad \eta = v, \quad \zeta = y e^v.$$

Indeed, we obtain that

$$\hat{C}_1 = \epsilon e^{-\eta} \partial_\eta, \quad \hat{C}_2 = \epsilon \partial_\xi, \quad \hat{C}_3 = -(\epsilon/2) \partial_\zeta.$$

Finally, the transformation

$$X^0 = \epsilon(e^\eta + \xi), \quad X^1 = \epsilon(e^\eta - \xi), \quad X^2 = -2\epsilon\zeta,$$

gives

$$\hat{P}_\rho = \frac{\partial}{\partial X^\rho}, \quad \rho = 0,1,2.$$

Composing all the transformations, we express X^ρ by means of the original variables T, q^1, and q^2,

$$X^0 = \epsilon e^{-T}[e^{q^1} + e^{-q^1} + (q^2)^2 e^{q^1}], \tag{19}$$

$$X^1 = \epsilon e^{-T}[e^{q^1} - e^{-q^1} - (q^2)^2 e^{q^1}], \tag{20}$$

$$X^2 = \epsilon e^{-T}[-2q^2 e^{q^1}]. \tag{21}$$

The class group transformations in terms of the coordinates X read

$$X'^0 = (3/2)X^0 + (1/2)X^1 - X^2,$$

$$X'^1 = -(1/2)X^0 + (1/2)X^1 + X^2,$$

$$X'^2 = -X^0 - X^1 + X^2,$$

which is an element of SO(1,2), and

$$X'^0 = X^0, \quad X'^1 = -X^1, \quad X'^2 = X^2.$$

The relations (19)–(21) together with (4)–(6) and (10) define a symplectic embedding ι: $\tilde{\Gamma}' \mapsto T^*M^3$ of the phase space $\tilde{\Gamma}'$ of our system, each component of which is spanned by the coordinates T, q^1, q^2, p_t, p_1, and p_2, into the cotangent bundle $T^*\mathbf{M}^3$ of the three-dimensional Minkowski space $\mathbf{M}^3$ with the coordinates (X^μ, P_μ). A very important point is that the image $\iota(\tilde{\Gamma}')$ is only a proper subset of $T^*\mathbf{M}^3$, namely the cotangent bundle of the inside of the light cone of the origin. Indeed, calculating $X \cdot X$ from (19)–(21), we obtain the identity

$$X \cdot X = -4e^{-2T} = -(4/\tau^2).$$

At the points of the light cone, the CMC is infinite; this surface represents the singularity of the torus dynamics. For the CMC τ, we obtain

$$\tau = \frac{2\epsilon}{\sqrt{-X \cdot X}}. \tag{22}$$

From Eq. (19), it follows that $\epsilon X^0 > 0$. Thus $\tilde{\Gamma}'_+ (\tilde{\Gamma}'_-)$ is mapped on the inside of the future (past) light cone. Moreover, Eqs. (4)–(6) and (10) yield

$$P_0 = (\epsilon e^T/4)(V_T p_T - V_1 p_1 - e^{-2q^1} V_2 p_2),$$

where

$$V_T=e^{q^1}+e^{-q^1}+(q^2)^2e^{q^1},\quad V_1=e^{q^1}-e^{-q^1}+(q^2)^2e^{q^1},\quad V_2=2q^2e^{q^1}.$$

We easily verify the identity

$$-V_T^2+V_1^2+e^{-2q^1}V_2^2=-4.$$

Thus, (V_T,V_1,V_2) is a ''timelike vector'' oriented towards the future ($V_T>0$) and (p_T,p_1,p_2) is a ''null vector'' (at the constraint surface, see (3)). Their ''scalar product'' $-V_Tp_T+V_1p_1+e^{-2q^1}V_2p_2$ is, therefore, negative (positive) if $p_T>0$ ($p_T<0$). As the sign of p_T and of ϵ are correlated, it follows that

$$P_0<0$$

everywhere at the constraint surface. This, together with Eqs. (10) and (12) imply that the points of the constraint surface satisfy the conditions

$$P\cdot P=0,\quad P_0<0 \tag{23}$$

with respect to the new variables (X^μ,P_μ). Let us denote by $\mathscr{P}$ the set of points in the momentum space with the coordinates P_μ that satisfy Eq. (23).

The transformation inverse to (19)–(21) is well-defined only inside the light cone, and can be written with respect to the coordinates as follows:

$$e^T=\frac{2}{\sqrt{-X\cdot X}},\quad e^{q^1}=\epsilon\,\frac{X^0+X^1}{\sqrt{-X\cdot X}},\ q^2=-\frac{X^2}{X^0+X^1}.$$

As X^0+X^1 is positive (negative) inside the future (past) half cone, e^{q^1} will be always positive. If we try to extend the transformation to the whole of $\mathbf{M}^3$, then we discover the meaning of the points lying outside the light cone. A simple calculation leads to the following complex transformation of the original variables,

$$t=it',\quad \nu=\nu'+(i\pi/2),$$

μ and β remaining unchanged, where t' and ν' are real. Then, all new momenta are real, but

$$e^{2T}=-p_3^2e^{-2\mu-2\nu'},\quad e^{2q^1}=-e^{-2\mu+2\nu'},$$

so that e^{2T} and e^{2q^1} become negative as necessary. Thus, we obtain space–times with the metric

$$ds^2=\mathscr{N}^2(t')dt'^2+e^2\mu(t')(dx^1)^2-e^{2\nu'(t')}[dx^2+\beta(t')dx^1]^2,$$

which have the Lorentz signature, but are acausal. A complete null geodesic crossing from the inside to the outside of the light cone represents an analytic three-dimensional space–time analogous to the Taub-NUT solution (see e.g., Ref. 24) with a Cauchy horizon and the Taub-NUT-like incompleteness at the cross point, if the singularity is viewed as a lightlike torus.

Martin's perennials are push-forwarded by ι just into the usual generators of the Poincaré group in the three-dimensional Minkowski spacetime $\mathbf{M}^3$: P_ρ and $J^\rho=\epsilon^{\rho\mu\nu}X_\mu P_\nu$. The corresponding group action is not transitive on $T^*\mathbf{M}^3$; the orbits are classified by the well-known invariants $P\cdot P$ and sign P_0, if $P\cdot P\leqslant 0$. However, each orbit of the group ISO(1,2) intersects $\iota(\tilde{\Gamma}')$, so no superfluous orbit has been added. Inside of the light cone, only the subgroup SO(1,2) acts; P's do not define any group acion on $\iota(\tilde{\Gamma}')$, because the corresponding vector fields are badly incomplete there. We can interpret our construction as follows.

The system of six Martin's functions form a complete algebra perennials; only three of them, the generators of SO(1,2), can be integrated to give a group action on the phase space $\tilde{\Gamma}'$; the three Hamiltonian vector fields corresponding to the perennials P_ρ are incomplete in $\tilde{\Gamma}'$. However, the Lie algebra generated by the six perennials defines a group, ISO(1,2). With the standard symplectic form, $\Omega_{\mathrm{ISO}}=dP_\mu\wedge dX^\mu$ of cotangent bundles, $T^*\mathbf{M}^3$ is a phase space, on which this group

does act. There is a map, ι, that sends $\tilde{\Gamma}'$ in $T^*\mathbf{M}^3$; ι is a symplectic imbedding and it pushes forward Martin's perennials into the generators of the action of ISO(1,2) on $T^*\mathbf{M}^3$. Thus, the map ι is equivariant for the two actions of the subgroup SO(1,2) of ISO(1,2). Such a space $T^*\mathbf{M}^3$ together with such a map ι can be called *minimal group completion of the phase space* $\tilde{\Gamma}'$ *corresponding to the complete Lie algebra of Martin's perennials*. The completion is minimal in the sense that there is no smaller completion (subspace of $T^*\mathbf{M}^3$), because each orbit of ISO(1,2) in $T^*\mathbf{M}^3$ intersects $\iota(\tilde{\Gamma}')$.

The constraint surface Γ' will be mapped to the subset of $T^*\mathbf{M}^3$ given by $X\cdot X<0$ and Eq. (23). The group ISO(1,2) does not act on $\iota(\Gamma')$ even if the generators of the group are tangential to $\iota(\Gamma')$: again, the translations are incomplete within this surface. However, there is a unique completion of $\iota(\Gamma')$ in $T^*\mathbf{M}^3$ on which the group acts, which we call $\Gamma_{\rm ISO}$. This surface is determined just by Eq. (23). Clearly, $(T^*\mathbf{M}^3, \Omega_{\rm ISO}, \Gamma_{\rm ISO})$ is a reparametrization invariant system that defines the corresponding c-orbits; they coincide with the maximal null geodesics in $\mathbf{M}^3$. Moreover, the image $\iota(\gamma)$ of each c-orbit γ in Γ' lies completely within some of the c-orbits of $(T^*\mathbf{M}^3, \Omega_{\rm ISO}, \Gamma_{\rm ISO})$—namely, in that null geodesic that extends $\iota(\gamma)$. Thus, we also have a well-defined map, $\bar{\iota}$: $\Gamma'\mapsto\Gamma_{\rm ISO}$, where $\Gamma_{\rm ISO}$ is the physical phase space of $(T^*\mathbf{M}^3, \Omega_{\rm ISO}, \Gamma_{\rm ISO})$.

In this sense, we can speak about a *group completion of the reparametrization invariant system*.

The completion constructed above has an important property which makes them interesting for physics. Let (M,Ω) be a symplectic manifold and let a set $\mathbf{g}$ of functions form a Lie algebra with respect to linear combinations and Poisson brackets. Let G be the (abstract) simply connected group that is determined by $\mathbf{g}$. Let (M_G,Ω_G,ι) be a minimal group completion of (M,Ω) by G. Thus, G acts on (M_G,Ω_G) as a group of symplectic diffeomorphisms. Let the image $\iota(M)$ be an open dense subset of M_G. Then, we say that the group G has a *weak action* on (M,Ω). For the above construction, we shall show the theorem:

Theorem 1: *The group ISO(1,2) has a weak action on the physical phase space* $\tilde{\Gamma}'$.

Proof: Consider the two surfaces $\Sigma_\pm$ defined by $X\cdot X=1$ and $\pm X^0>0$, respectively, inside the light cone of the origin in $\mathbf{M}^3$; the manifolds $\Sigma_\pm\times\mathscr{P}$ are global transversal surfaces in $\iota(\Gamma'_+)$ and $\iota(\Gamma'_-)$, as they are intersected by all null geodesics inside of the light cone. Thus, $\bar{\Gamma}'$ can be identified with $(\Sigma_+\times\mathscr{P})\cup(\Sigma_-\times\mathscr{P})$. Next consider the surface Σ in $\mathbf{M}^3$ defined by $X^0=0$. Clearly, the manifold $\Sigma\times\mathscr{P}$ is a global transversal surface for the group completed system, because it is intersected by any null geodesic at exactly one point; we can identify $\Gamma_{\rm ISO}$ with $\Sigma\times\mathscr{P}$.

Every point of $\Sigma_\pm\times\mathscr{P}$ defines a unique null geodesic; this geodesic intersects $\Sigma\times\mathscr{P}$ at precisely one point. Thus we have a well-defined map ρ: $(\Sigma_+\times\mathscr{P})\cup(\Sigma_-\times\mathscr{P})\mapsto\Sigma\times\mathscr{P}$ (the map ρ and its properties have been described in Sec. II). It is easy to see that $\bar{\iota}$ can be identified with ρ and so we have to show that $\rho(\Sigma_+\times\mathscr{P})\cup(\Sigma_-\times\mathscr{P})$ is open and dense in $\Sigma\times\mathscr{P}$.

Let us introduce the coordinates u and v at the surfaces Σ_ϵ by

$$X^0=\epsilon\sqrt{1+u_1^2+u_2^2},\quad X^1=u_1,\quad X^2=u_2,$$

and the coordinates x_1 and x_2 at Σ by $X^1=x_1$ and $X^2=x_2$. Let us consider null geodesics with a definite three-momenta of the form

$$P_\mu=p(1,\cos\alpha,\sin\alpha),\tag{24}$$

where p is a (negative) number. The null geodesic with the momenta (24) starting at the point (x_1,x_2) of Σ will intersect Σ_ϵ at the point

$$t=\epsilon\sqrt{1+u_1^2+u_2^2},\quad u_1=x_1+\epsilon\cos\alpha\sqrt{1+u_1^2+u_2^2},$$

$$u_2=x_2+\epsilon\sin\alpha\sqrt{1+u_1^2+u_2^2}.$$

Solving for (x_1,x_2), we obtain a description of ρ in terms of the coordinates u_1, u_2, ϵ, x_1, and x_2,

$$x_1 = u_1 - \epsilon \cos \alpha \sqrt{1+u_1^2+u_2^2}, \tag{25}$$

$$x_2 = u_2 - \epsilon \sin \alpha \sqrt{1+u_1^2+u_2^2}. \tag{26}$$

To see which part of Σ is hit, we introduce new variables,

$$x_1' = x_1 \cos \alpha + x_2 \sin \alpha, \quad x_2' = -x_1 \sin \alpha + x_2 \cos \alpha,$$

$$u_1' = u_1 \cos \alpha + u_2 \sin \alpha, \quad u_2' = -u_1 \sin \alpha + u_2 \cos \alpha,$$

and observe that

$$u_1^2 + u_2^2 = u_1'^2 + u_2'^2.$$

Then, Eqs. (25) and (26) are equivalent to

$$x_1' = u_1' - \epsilon \sqrt{1+u_1'^2+u_2'^2}, \quad x_2' = u_2'. \tag{27}$$

Thus, as u_2' runs through $\mathbf{R}$, so does x_2'. For a fixed x_2',

$$x_1' = u_1' - \epsilon \sqrt{1+u_1'^2+x_2'^2},$$

hence

$$\frac{dx_1'}{du_1'} = 1 - \epsilon \frac{u_1'}{\sqrt{1+u_1'^2+x_2'^2}} > 0, \quad \epsilon = \pm 1.$$

For $\epsilon = +1$, $x_1' \to -\infty$ as $u_1' \to -\infty$. On the other side, we obtain

$$\lim_{u_1'=\infty} x_1' = \lim_{u_1'=\infty} \frac{-1-x_2'^2}{u_1' + \sqrt{1+u_1'^2+x_2'^2}} = 0.$$

For $\epsilon = -1$, $x_1' \to \infty$ as $u_1' \to \infty$. On the other side, we obtain

$$\lim_{u_1'=-\infty} x_1' = \lim_{u_1'=-\infty} \frac{1+x_2'^2}{|u_1'| + \sqrt{1+u_1'^2+x_2'^2}} = 0.$$

Thus, Σ_+ is mapped to $x_1 \cos \alpha + x_2 \sin \alpha < 0$ and Σ_- is mapped to $x_1 \cos \alpha + x_2 \sin \alpha > 0$. Only the straight line $x_1 \cos \alpha + x_2 \sin \alpha = 0$ is missing from each surface $P_1 = \text{const}$, $P_2 = \text{const}$. It follows that $\rho(\tilde{\Gamma}')$ is open and dense. Q.E.D.

There is a standard way of construction of group completions, if the Lie algebra of observables is complete. Let (M, Ω) be a symplectic manifold and let $\mathbf{g}$ be a Lie algebra generated by a complete system of functions on M. Let G be the unique simply connected group whose Lie algebra coincides with $\mathbf{g}$. Let $\mathbf{g}^*$ be the dual linear space to $\mathbf{g}$ and let ad* be the coadjoint representation of G on $\mathbf{g}^*$. The orbits of the action ad* of G in $\mathbf{g}^*$ are homogeneous symplectic spaces of the group G according to a beautiful result of Kirillov.[25] Moreover, from a basis of $\mathbf{g}$, a (basis-independent) map Π: $M \mapsto \omega$ of M into an orbit ω can be constructed; Π is the so-called momentum map. Then, the manifold ω with Kirillov's symplectic structure and with Π as ι is the desired (minimal) completion.

As an example consider the manifold M with the coordinates q and p given by $q^2 + p^2 < 1$, equipped with the symplectic form $\Omega = dp \wedge dq$. Let the algebra $\mathbf{g}$ be generated by the functions q, p, and 1 (constant function). The corresponding group is the three-dimensional Heisenberg group defined on $\mathbf{R}^3$ with group law

$$(a_1,b_1,r_1)\cdot(a_2,b_2,r_2)=\left(a_1+a_2,b_1+b_2,r_1+r_2+\frac{b_1a_2-b_2a_1}{2}\right).$$

The space dual to the algebra is $\mathbf{R}^3$ with the coordinates A, B, and R, the orbits of the group are the planes $R=\text{const}$ and the momentum map is

$$A=q, \quad B=p, \quad R=1.$$

Thus, the orbit ω on which M is mapped is given by the equation $R=1$. The image of M is the disk $A^2+B^2<1$, and the group does not act even weakly.

B. SO(2,3) completion

There is a motivation to look for further symmetries; the subgroup SO(1,2) that acts on $\tilde{\Gamma}'$ is too small to define a time evolution à la Dirac. The simple form of the reparametrization invariant system $(T^*\mathbf{M}^3, \Omega_{\mathrm{ISO}}, \Gamma_{\mathrm{ISO}})$ allows us to see immediately that there are more perennials than just the generators of the three-dimensional Poincaré group; we have also the conformal isometries. The so-called dilatation is generated by

$$D:=X^\mu P_\mu, \tag{28}$$

and the so-called conformal accelerations are generated by

$$Q_0:=(X\cdot X)P_0+2X^0(X\cdot P), \tag{29}$$

$$Q_1:=(X\cdot X)P_1-2X^1(X\cdot P), \tag{30}$$

$$Q_2:=(X\cdot X)P_2-2X^2(X\cdot P). \tag{31}$$

It is easy to verify that the Poisson brackets of these variables with $P\cdot P$ weakly vanish. Let us denote $B^\mu Q_\mu$ by B. Then, the Lie algebra of A, B, C, and D is given by Eqs. (11) and

$$\{A,B\}=2(A\cdot B)D-2(\varepsilon_{\rho\mu\nu}A^\mu B^\nu)J^\rho, \quad \{B,C\}=(\varepsilon^{\rho\mu\nu}B_\mu C_\nu)Q_\rho, \tag{32}$$

$$\{A,D\}=-A, \quad \{B,B'\}=0, \quad \{B,D\}=B, \quad \{C,D\}=0. \tag{33}$$

This is the Lie algebra of the group SO(2,3). Indeed, let Z^0, Z^1, Z^3, W^0, and W^1 be coordinates in $\mathbf{R}^5$ with the metric

$$dS^2=-(dZ^0)^2-(dW^0)^2+(dZ^1)^2+(dZ^2)^2+(dW^1)^2; \tag{34}$$

to obtain the algebra (11), (32), and (33), we have to just identify

$$J^\rho\mapsto\varepsilon^{\rho\mu\nu}Z_\mu\frac{\partial}{\partial Z^\nu}, \quad D\mapsto W_0\frac{\partial}{\partial W^1}-W_1\frac{\partial}{\partial W^0},$$

$$P_\rho\mapsto\left(Z_\rho\frac{\partial}{\partial W^0}-W_0\frac{\partial}{\partial Z^\rho}\right)-\left(Z_\rho\frac{\partial}{\partial W^1}-W_1\frac{\partial}{\partial Z^\rho}\right),$$

$$Q_\rho\mapsto\left(Z_\rho\frac{\partial}{\partial W^0}-W_0\frac{\partial}{\partial Z^\rho}\right)+\left(Z_\rho\frac{\partial}{\partial W^1}-W_1\frac{\partial}{\partial Z^\rho}\right),$$

where the indices are lowered by the metric (34).

The Hamiltonian vector field corresponding to the dilatation is complete not only within $T^*\mathbf{M}^3$, but even within $\iota(\tilde{\Gamma}')$. The Hamiltonian vector fields corresponding to Q's are incomplete within $T^*\mathbf{M}^3$; however, there is still a chance to construct additional observables from Q's, so we

have to construct the next completion. This completion is well-known; it is the cotangent bundle $T^*\overline{\mathbf{M}}^3$ of the compactified Minkowski space–time $\overline{\mathbf{M}}^3$.[26] Let us briefly describe the construction, because we shall need some details of it.

Consider the three-dimensional Einstein cosmology space–time $\mathbf{M}_E^3$ with the coordinates τ, ϑ, and φ and the metric

$$ds^2 = -d\tau^2 + d\vartheta^2 + \sin^2\vartheta d\varphi^2.$$

The null geogesics that start at the point $\tau = \tau_0$, $\vartheta = 0$, are given by

$$\tau = \tau_0 + \lambda, \quad \vartheta = \lambda, \quad \varphi = \text{const}.$$

These geodesic form a null cone that refocuses at the point $\tau = \tau_0 + \pi$, $\vartheta = \pi$. Similarly, all null geodesics through the point $\tau = \tau_0$, $\vartheta = \pi$, have the form

$$\tau = \tau_0 + \lambda, \quad \vartheta = \pi - \lambda, \quad \varphi = \text{const},$$

and they refocus at $\tau = \tau_0 + \pi$, $\vartheta = 0$. As it is well-known, (e.g., Ref. 26), the compactified Minkowski space–time $\overline{\mathbf{M}}^3$ is obtained from $\mathbf{M}_E^3$ by the identification of each point $(\tau, \vartheta, \varphi)$ with the point $(\tau + \pi, \pi - \vartheta, \varphi + \pi)$. There is a conformal isometry ϕ that sends $\mathbf{M}^3$ into $\overline{\mathbf{M}}^3$ such that the set $\phi(\mathbf{M}^3)$ lies in the future of the null cone between the points $(\tau = -\pi, \vartheta = 0)$ and $(\tau = 0, \vartheta = \pi)$, and in the past of the null cone between $(\tau = 0, \vartheta = \pi)$ and $(\tau = \pi, \vartheta = 0)$.

Another copy of Minkowski space–time lies between the points $(\tau = -\pi, \vartheta = \pi)$, $(\tau = 0, \vartheta = 0)$, and $(\tau = \pi, \vartheta = \pi)$. By the above identification and the two conformal isometries, the point X^μ of the first Minkowski space–time will be mapped to the point Y^μ of the second one given by

$$Y^\mu = \frac{X^\mu}{X \cdot X}, \tag{35}$$

if the inertial coordinates X^μ and Y^μ are chosen properly. We will make some use of these two patches of $\overline{\mathbf{M}}^3$; the fact that they do not cover $\overline{\mathbf{M}}^3$ will not be important. Let us call them U and V.

The map ϕ is a diffeomorphism of three-dimensional manifolds and it can be extended to an isomorphism ϕ_{cot} of the cotangent bundles of these manifolds. As the conformal group SO(2,3) acts transitively in $T^*\overline{\mathbf{M}}^3$, the triad $(T^*\overline{\mathbf{M}}^3, \Omega_{\text{SO}}, \phi_{\text{cot}})$ is the desired minimal SO(2,3) completion; the form Ω_{SO} is the standard symplectic form of cotangent bundles; observe that it is exact. The set $\phi_{\text{cot}}(T^*\mathbf{M}^3)$ is equal to T^*U, and so it is open and dense in $T^*\overline{\mathbf{M}}^3$. Thus, SO(2,3) *acts weakly* on $T^*\overline{\mathbf{M}}^3$.

Let the canonical coordinates in the cotangent bundles T^*U and T^*V be X^μ, P_μ, and Y^μ, Q_μ, respectively. Then the transformation (35) between X^μ and Y^μ leads to the following transformation between P_μ and Q_μ:

$$Q_\mu = (X \cdot X) P_\mu - 2(X \cdot P) X_\mu. \tag{36}$$

The inverse map is

$$X^\mu = \frac{Y^\mu}{Y \cdot Y}, \quad P_\mu = (X \cdot X) Q_\mu - 2(X \cdot Q) Y_\mu. \tag{37}$$

Comparison with Eqs. (29)–(31) shows that the use of the letter Q for this coordinate will not lead to any confusion with the notation for the generators of conformal accelerations. From the transformation formulas (36) and (37), we easily verify that the symplectic form Ω_{SO} at the points where the patches overlap satisfies

$$dP_\mu \wedge dX^\mu = dQ_\mu \wedge dY^\mu.$$

The SO(2,3) completion of the RIS $(T^*\mathbf{M}^3, \Omega_{\text{ISO}}, \Gamma_{\text{ISO}})$ is $(T^*\overline{\mathbf{M}}^3, \Omega_{\text{SO}}, \Gamma_{\text{SO}})$, where Γ_{SO} is given by the equations

$$g_i^{\mu\nu}P_\mu^i P_\nu^i=0, \quad P_0^i<0,$$

within each chart T^*U_i, where the metric $g_i^{\mu\nu}$ is the metric of the conformal chart U_i, P_μ^i the canonical coordinate in the fibers of T^*U_i, and we allow only charts that have the same time orientation.

It follows that the projection of the *c*-orbits to the configuration space $T^*\overline{\mathbf{M}}^3$ are complete null geodesics; they are closed curves (topologically S^1). The ϕ-images of the *c*-orbits of the system $(T^*\underline{\mathbf{M}}^3,\Omega_{\mathrm{ISO}},\Gamma_{\mathrm{ISO}})$ are the null geodesics in the chart U; each of them is completed by one point in $\mathbf{M}^3$. The null geodesics in $\mathbf{M}^3$ that do not contain any ϕ-images form the light cone of the origin of the chart V $(Y^\mu=0)$. Hence, the surface Γ_0 given by the equations $X^0=0$, $P_0=-\sqrt{P_1^2+P_2^2}$ within the chart T^*U, and by $Y^0=0$, $Q_0=-\sqrt{Q_1^2+Q_2^2}$ within the chart T^*V is a global transversal surface. We can introduce coordinates $(x_1,x_2)\in\mathbf{R}^2$, $(p_1,p_2)\in\mathbf{R}^2\backslash\{0\}$ in $\Gamma_0\cap T^*U$ and $(y_1,y_2)\in\mathbf{R}^2$, $(q_1,q_2)\in\mathbf{R}^2\backslash\{0\}$ in $\Gamma_0\cap T^*V$ such that the imbedding equations are

$$\left.\begin{array}{lll} X^0=0, & X^1=x_1, & X^2=x_2, \\ P_0=-\sqrt{p_1^2+p_2^2}, & P_1=p_1, & P_2=p_2, \end{array}\right\} \tag{38}$$

and

$$Y^0=0, \quad Y^1=y_1, \quad Y^2=y_2,$$

$$Q_0=-\sqrt{q_1^2+q_2^2}, Q_1=q_1, \quad Q_2=q_2.$$

One easily verifies that Γ_0 defined in this way is a smooth surface and that the transformation formulas (35), (36), and (37) imply the relations

$$x_k=\frac{y_k}{y_1^2+y_2^2}, \quad p_k=(y_1^2+y_2^2)q_k-2(y_1q_1+y_2q_2)y_k. \tag{39}$$

In particular, Eq. (36) implies that $Q_0<0$ if $P_0<0$. The pullback Ω_0 of the symplectic form Ω_{SO} to Γ_0 is given in these coordinates by

$$\Omega_0=dp_k\wedge dx_k=dq_k\wedge dy_k,$$

the last inequality following from Eqs. (39). Ω_0 is exact.

The manifold Γ_0 is a bundle with the fiber $\mathscr{P}\cong S^1\times\mathbf{R}$ given by $x_k=\mathrm{const}$ or $y_k=\mathrm{const}$. The base space is S^2, and the coordinates (x_1,x_2) and (y_1,y_2) are nothing but the two stereographic projection charts of S^2. The symplectic space (Γ_0,Ω_0) is a homogeneous symplectic space of the group SO(2,3), which acts on Γ_0 by the projection of symmetries (see Sec. II); each element of the conformal group maps null geodesics in null geodesics. $(\underline{\Gamma}_0,\Omega_0)$ can be identified with the group completed physical phase space $\overline{\Gamma}_{\mathrm{SO}}$. The composition $\phi\circ\overline{\iota}$ of the ISO(1,2)-completion and the SO(2,3)-completion gives the image $\phi(\overline{\iota}(\overline{\Gamma}'))$ as an open dense subspace of Γ_0; thus, the conformal group SO(2,3) acts weakly (and transitively) on $\overline{\Gamma}'$.

The generators of the action of SO(2,3) on Γ_0 are Hamiltonian vector fields of the projections of the perennials P_μ, Q_μ, J^μ, and D from $T^*\mathbf{M}^3$ to Γ_0 (see Sec. II). In the patch (x_k,p_k), the projections coincide with the functions

$$\left.\begin{array}{lll} \overline{P}_0:=-\sqrt{\mathbf{p}\cdot\mathbf{p}}, & \overline{Q}_0:=-(\mathbf{x}\cdot\mathbf{x})\sqrt{\mathbf{p}\cdot\mathbf{p}}, & \overline{J}^0:=x_1p_2-x_2p_1, \\ \overline{P}_1:=p_1, & \overline{Q}_1:=(\mathbf{x}\cdot\mathbf{x})p_1-2(\mathbf{x}\cdot\mathbf{p})x_1, & \overline{J}^1:=-x_2\sqrt{\mathbf{p}\cdot\mathbf{p}}, \\ \overline{P}_2:=p_2, & \overline{Q}_2:=(\mathbf{x}\cdot\mathbf{x})p_2-2(\mathbf{x}\cdot\mathbf{p})x_2, & \overline{J}^2:=x_1\sqrt{\mathbf{p}\cdot\mathbf{p}}, \\ & & \overline{D}:=\mathbf{x}\cdot\mathbf{p}. \end{array}\right\} \tag{40}$$

These functions will be considered as observables. Here, we have used the abbreviation $\mathbf{u}\cdot\mathbf{v}:=u_1v_1+u_2v_2$. The expressions within the other patch, (y_k,q_k), are analogous, one just has to exchange P's and Q's, write y for x and q for p. Via Poisson brackets, the functions generate the

Lie algebra of SO(2,3). They are ten functions of four variables; thus, there will be six relations. These relations can be systematically written down, if we solve the definitions of $\overline{P_1}$, $\overline{P^2}$, $\overline{J^1}$, and $\overline{J^2}$ for p_1, p_2, x_1, and x_2 and substitute the results into the other definitions,

$$\overline{P_0^2}=\overline{P_1^2}+\overline{P_2^2}, \quad \overline{J^0}\,\overline{P_0}+\overline{J^1}\,\overline{P_1}+\overline{J^2}\,\overline{P_2}=0, \tag{41}$$

$$\overline{Q_0^2}=\overline{Q_1^2}+\overline{Q_2^2}, \quad \overline{D}=\frac{\overline{P_1}\,\overline{J^2}-\overline{P_2}\,\overline{J^1}}{\sqrt{\overline{\mathbf{P}}\cdot\overline{\mathbf{P}}}}, \tag{42}$$

$$\overline{Q_1}=\frac{\overline{P_1}}{\overline{\mathbf{P}}\cdot\overline{\mathbf{P}}}[(\overline{J^1})^2-(\overline{J^2})^2]+\frac{\overline{P_2}}{\overline{\mathbf{P}}\cdot\overline{\mathbf{P}}}[2\overline{J^1}\,\overline{J^2}], \tag{43}$$

$$\overline{Q_2}=\frac{\overline{P_1}}{\overline{\mathbf{P}}\cdot\overline{\mathbf{P}}}[2\overline{J^1}\,\overline{J^2}]-\frac{\overline{P_2}}{\overline{\mathbf{P}}\cdot\overline{\mathbf{P}}}[(\overline{J^1})^2-(\overline{J^2})^2]. \tag{44}$$

The two Eqs. (41) are relations concerning also ISO(1,2) alone and the four relations (42)–(44) can be used to calculate the remaining generators of SO(2,3). A quadratic relation follows:

$$-(\overline{J^0})^2+(\overline{J^1})^2+(\overline{J^2})^2=\overline{D^2}. \tag{45}$$

The Poisson brackets of the four quadratic expressions $P\cdot P$, $Q\cdot Q$, $J\cdot P$, and $-D^2+J\cdot J$ with the generators P_μ, Q_μ, J^μ, and D are mostly vanishing or proportional to $P\cdot P$; they are (generalized) Casimirs of some subgroups.

As it was explained in Sec. II, the action of SO(2,3) on Γ_0 is generated by the Hamiltonian vector fields of the observables (40). This defines a linear map from the Lie algebra so(2,3) into vector fields on Γ_0. We can describe the map, if we choose a basis of so(2,3) and list the images of the elements of the basis. Let the basis be

$$(\mathscr{P}_0,\mathscr{P}_1,\mathscr{P}_2,\mathscr{Q}_0,\mathscr{Q}_1,\mathscr{Q}_2,\mathscr{T}^0,\mathscr{T}^1,\mathscr{T}^2,\mathscr{D}). \tag{46}$$

Here, we denote the abstract elements of the Lie algebra by upper case calligraphic letters to distinguish them from the corresponding perennials or vector fields. The association with the vector fields is

$$\mathscr{P}_0\mapsto-\frac{p_k}{\sqrt{\mathbf{p}\cdot\mathbf{p}}}\frac{\partial}{\partial x_k}, \quad \mathscr{P}_k\mapsto\frac{\partial}{\partial x_k}, \tag{47}$$

$$\mathscr{Q}_0\mapsto-\frac{\mathbf{x}\cdot\mathbf{x}}{\sqrt{\mathbf{p}\cdot\mathbf{p}}}p_k\frac{\partial}{\partial x_k}+2\sqrt{\mathbf{p}\cdot\mathbf{p}}x_k\frac{\partial}{\partial p_k}, \tag{48}$$

$$\mathscr{Q}_k\mapsto[(\mathbf{x}\cdot\mathbf{x})\delta_{kl}-2x_kx_l]\frac{\partial}{\partial x_l}+[(\mathbf{x}\cdot\mathbf{p})\delta_{kl}+x_kp_l-x_lp_k]\frac{\partial}{\partial p_l}, \tag{49}$$

$$\mathscr{T}^0\mapsto-x_2\frac{\partial}{\partial x_1}+x_1\frac{\partial}{\partial x_2}-p_2\frac{\partial}{\partial p_1}+p_1\frac{\partial}{\partial p_2}, \tag{50}$$

$$\mathscr{T}^1\mapsto-\frac{x_2p_k}{\sqrt{\mathbf{p}\cdot\mathbf{p}}}\frac{\partial}{\partial x_k}+\sqrt{\mathbf{p}\cdot\mathbf{p}}\frac{\partial}{\partial p_2}, \tag{51}$$

$$\mathscr{T}^2\mapsto\frac{x_1p_k}{\sqrt{\mathbf{p}\cdot\mathbf{p}}}\frac{\partial}{\partial x_k}-\sqrt{\mathbf{p}\cdot\mathbf{p}}\frac{\partial}{\partial p_1}, \tag{52}$$

$$\mathscr{D} \mapsto x_1 \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_2} - p_1 \frac{\partial}{\partial p_1} - p_2 \frac{\partial}{\partial p_2}. \tag{53}$$

We shall need the form of these vector fields for the construction of the physical representation of the group in Sec. VI.

V. THE HAMILTONIAN

In this section, we study the time evolution of the $2+1$ gravity model. We are going to apply Dirac's idea; a choice of transversal surfaces of maximal symmetry, and a comparison of different time levels using symmetry operations.

There are two problems that prevent a straightforward application. First, the group SO(2,3) is too large to have a representation that satisfies all conditions on *physical* representation (see the next section). Second, only a four-dimensional subgroup of SO(2,3) has the *action* on the phase space of the system that is associated with the corresponding Lie algebra of perennials.

The largest subgroups that have physical representations are G_1 and G_2 with the structure $(\mathrm{SO}(1,2)\times\mathbf{R})\otimes_S\mathbf{R}^3$. G_1 is generated by $\mathscr{T}^0$, $\mathscr{T}^1$, $\mathscr{T}^2$, $\mathscr{D}$, $\mathscr{P}_0$, $\mathscr{P}_1$, and $\mathscr{P}_2$ and G_2 by $\mathscr{T}^0$, $\mathscr{T}^1$, $\mathscr{T}^2$, $\mathscr{D}$, $\mathscr{Q}_0$, $\mathscr{Q}_1$, and $\mathscr{Q}_2$. Consider G_1. The corresponding group completion of the $2+1$ gravity coincides with the ISO(1,2)-completion that was constructed in the previous section, because the only additional element, the dilatation $\mathscr{D}$, acts on $\tilde{\Gamma}'$. Let us, therefore, restrict ourselves to G_1.

The group G_1 acts weakly on the physical phase space Γ'. This is important for the quantum generators of the group to have suitable spectra. On the other hand, the weak action is not sufficient for the construction of a time evolution according to Dirac. Let us explain these two points.

Consider the two-dimensional disk example described at the end of the previous section. The Heisenberg group did not act even weakly. On the disk, the values of the classical observables q and p are bounded; $q^2+p^2<1$. On the group completion, which is a whole plain, the values of q fill up the interval $(-\infty,\infty)$ and similarly for p. This follows from the structure of the Lie algebra (see, e.g., Ref. 10). If we represent q and p by self-adjoint operators satisfying the canonical commutation relations, then this structure forces the spectra again to fill up the whole real axis. The corresponding quantum mechanics contains semiclassical wave packets with average values of $\hat{q}$ and $\hat{p}$ that lie far away from possible classical values. Let us call this the *problem of ranges*. On the other hand, if a group acts weakly, then the only change of the range of classical values that is motivated is an addition of some limit points. This would happen in any case in the quantum mechanics, because the spectrum of any self-adjoint operator is a closed subset of **R**. In this respect, the weak action does not differ from an ordinary action.

However, for Dirac's idea to work, we have to find a family of maximally symmetric transversal surfaces and a sufficiently large subset of the symmetry group so that all such surfaces can be obtained from one by the action of the subset. We emphasize that this has to work within the constraint surface of the classical theory so that we can interpret the transformations. Clearly, for these purposes, the weak action is not adequate. First, the maximally symmetric surfaces will lie in the group completed constraint surface, but not, in general, inside that of the system; the intersection of such surfaces with the constraint surface of the system will not be, in general, globally transversal. Second, an image of a globally transversal surface by an element of the group that has only a weak action will not, in general, lie inside the constraint surface. It is easy to construct examples of this kind for the action of G_1. This means that only the subgroup G_0 is at our disposal for Dirac's construction.

The group $G_0=\mathrm{SO}(1,2)\times\mathbf{R}$ generated by $\mathscr{T}^0$, $\mathscr{T}^1$, $\mathscr{T}^2$, and $\mathscr{D}$ is a Cartesian product of two simple groups. Its three-dimensional subgroups are of two types:

(1) SO (1,2), which is the *unique* subgroup of this type, because it is a normal subgroup;
(2) Subgroups that leave a null plane invariant; an example is the subgroup of the plane $X^0+X^1=0$, which is generated by $\mathscr{T}^0+\mathscr{T}^1$, $\mathscr{T}^2$ and $\mathscr{D}$. All other subgroups of this type are (group-) similar to this one.

The surfaces in Γ' symmetric with respect to SO (1,2) satisfy $-(X^0)^2+(X^1)^2+(X^2)^2=\mathrm{const}(\mathrm{const}<0)$, have two components (with $X^0>0$ and $X^0<0$), and form a one-dimensional

family. The components coincide with the CMC surfaces, and the union of both components is globally transversal. On the other hand, the projection of the surfaces to $\mathbf{M}^3$ that are invariant with respect to the null plane groups are of course these null planes; the surfaces do not lie inside the null cone $\iota(\tilde{\Gamma}')$ and their intersections with $\iota(\tilde{\Gamma}')$ are not transversal. We summarize the results:

Theorem 2: *The* $2+1$ *gravity model possesses a unique one-dimensional family of maximally symmetric globally transversal surfaces. Each such surface has two components that are CMC surfaces with opposite values of the CMC.*

The second step of the construction is to find a subgroup that would carry us along the family of the CMC surfaces. Thus, it must be a subgroup whose elements are representatives of all classes of $G_0/\mathrm{SO}(1,2)$. However, $G_0/\mathrm{SO}(1,2)=\mathbf{R}$, so the desired subgroup is generated by just one element, which must have the form $\mathscr{D}+a\mathscr{T}^0+b\mathscr{T}^1+c\mathscr{T}^2$, where a, b, and c are three arbitrary reals. We can normalize the generator in this way, as the overall factor does not change the subgroup, and the factor in front of $\mathscr{D}$ must be nonzero; a nontrivial motion of the CMC surface is generated just by the $\mathscr{D}$ term. Let us study how it acts on the CMC τ. Using Eqs. (22) and (28), we find that

$$\{\tau,D\}=-\frac{2\epsilon}{\sqrt{-X\cdot X}}.$$

Thus, for $\epsilon>0$, the action of D diminishes τ, and for $\epsilon<0$, it enlarges τ. $\epsilon>0$ ($\epsilon<0$) means that we are in the future (past) light cone of the origin. In the future light cone, we have expanding tori ($\tau>0$) and they expand from the ''big bang'' $\tau=\infty$ to the maximal expansion state $\tau=0$. Thus, D generates evolution towards future here. In the past light cone ($\tau<0$), we have contracting tori that start at the maximal expansion state $\tau=0$ and finish at the ''big crunch'' $\tau=-\infty$. Thus D generates an evolution towards past. Luckily enough, there is a smooth Hamiltonian H that evolves everything towards the future; $H=D$ inside the future light cone and $H=-D$ inside the past light cone. Let us calculate the corresponding perennial $\overline{H}$ on the physical phase space $\overline{\Gamma}_{\mathrm{ISO}}$. From the proof of theorem 1, it follows that the portion $\overline{\Gamma}'_+$ of the physical phase space that corresponds to the future half of the light cone is given by $x_1\cos\alpha+x_2\sin\alpha<0$ and that $\overline{\Gamma}'_-$ corresponding to the past one by $x_1\cos\alpha+x_2\sin\alpha>0$. However, Eq. (24) implies that

$$p_1=p\cos\alpha,\quad p_2=p\sin\alpha,$$

where $p<0$. Here x_1, x_2, p_1, and p_2 are the coordinates that we have chosen in the physical phase space $\overline{\Gamma}$ (cf. Eq. (38)). Then, from the last of Eq. (40) it follows that $\overline{D}$ is positive for the expanding tori and negative for the contracting ones, or

$$\overline{H}=|\overline{D}|. \tag{54}$$

In the general case, we start from the function

$$\overline{D}+a\overline{J^0}+b\overline{J^1}+c\overline{J^2}.$$

As it is only the D-part which leads to changes in τ, the Hamiltonian that evolves towards the future corresponding to the above function is

$$\overline{H}=\mathrm{sign}(\overline{D})(\overline{D}+a\overline{J^0}+b\overline{J^1}+c\overline{J^2}).$$

Equations (40) lead to

$$\overline{H}=\mathrm{sign}(\mathbf{x}\cdot\mathbf{p})[\mathbf{x}\cdot\mathbf{p}+a(x_1p_2-x_2p_1)+bx_2\sqrt{\mathbf{p}\cdot\mathbf{p}}-cx_1\sqrt{\mathbf{p}\cdot\mathbf{p}}].$$

Let us change the variables x_1, x_2, p_1, and p_2 to $\overline{D}$, $\overline{J^0}$, p, α; we obtain

$$\overline{H}=|\overline{D}|(1+b\sin\alpha-c\cos\alpha)+\mathrm{sign}(\overline{D})\overline{J^0}(a+b\cos\alpha+c\sin\alpha). \tag{55}$$

We can see that $\overline{H}$ is unbounded from below except for the case that $a=b=c=0$, because $\overline{J^0}$ can take on arbitrary values independent of $\overline{D}$ and α. Hence, (54) is the only one from the three-dimensional family of possible candidates for a Hamiltonian that is bounded from below (and even positive). This is, of course, nothing but an intriguing observation; there is no *a priori* reason for the generator of the time evolution, even if it evolves towards the future, to be bounded from below or positive, unless it plays simultaneously another role, for example that of the total energy of the system. We also observe that the dynamics simplifies strongly if we choose (54) in comparison with all other candidates; $\overline{J^\mu}$ becomes time independent, and P_μ just scales with time. The next comment is that the choice (54) leads to the dynamics that has been obtained by Moncrief.[17] Finally, it is easy to see that there will be no problem to define the quantum mechanical operator $\hat{H}$ from H, if the operators $\hat{D}$ and $\hat{J}^\mu$ are given, because $\hat{D}$ will commute with all $\hat{J}$'s. The corresponding problem of ranges will be automatically solved, if we define $|\hat{H}|$ by the spectral theorem.

VI. THE PHYSICAL REPRESENTATION

The physical representation of the algebra so(2,3) would map each element of the algebra to a linear operator on a common invariant dense domain $\mathscr{K}_0$ in a Hilbert space $\mathscr{K}$; the map R must satisfy the following conditions:

(1) R is linear, $R(1)=id$, and
$$(i/\hbar)R(\{X,Y\})=R(X)R(Y)-R(Y)R(X)$$
for all X, $Y\in$ so(2,3) on $\mathscr{K}_0$;
(2) the operators $R(X)$ for all $X\in$ so(2,3) are essentially self-adjoint on $\mathscr{K}_0$;
(3) the problem of ranges is satisfactorily solved;
(4) the operators $R(X)$ for all $X\in$ so(2,3) satisfy algebraic relations that go over to (41)–(44) in the classical limit.

In general, the group method of quantization of an algebra $\mathbf{g}$ of observables on a symplectic manifold (M,Ω) is to find a unitary representation on a Hilbert space $\mathscr{K}$ of the group G corresponding to the algebra. Then, the generators of the group action on $\mathscr{K}$ satisfy automatically the conditions (1) and (2), but a part of condition (3) $(P_0<0)$ and condition (4) can pose problems.

In this section, we are going to use an old idea of finding the physical representation by the group way; the Kostant–Kirillov method of orbits. This method works quite generally for finite systems. Let us briefly describe the steps of the method (for more detail see Refs. 25 and 27).

The method of orbits is based on the momentum map Π determined by the algebra of observables $\mathbf{g}$ (in this way, the relations and ranges are encoded). $\Pi(M)$ is a particular orbit ω of G in the linear space $\mathbf{g}^*$ dual to the Lie algebra $\mathbf{g}$, where the group acts via the coadjoint representation.

The method starts with a choice of a point $F\in\omega$ and with calculating the stabilizer $G_F\subset G$ of F. Then, the subalgebra $\mathbf{n}_F$ called subordinate to F must be found satisfying the conditions

(1) $\langle F,[X,Y]\rangle=0 \quad \forall X,Y\in\mathbf{n}_F$,
(2) $\mathrm{codim}_{\mathbf{g}}\mathbf{n}_F=(1/2)\dim\ \omega$,
(3) Pukanszky's condition; let $\mathbf{n}_F^\perp$ be the subspace of $\mathbf{g}^*$ that annihilates $\mathbf{n}_F$, then, $F+\mathbf{n}_F^\perp\subset\omega$.

One can show that $\mathbf{g}_F\subset\mathbf{n}_F$. The subalgebra $\mathbf{n}_F$ generates a subgroup N_F of G and one must find a one-dimensional unitary representation $R_{\mathbf{n}}$ of N_F such that

$$R_{\mathbf{n}}(\exp\ X)=\exp(\langle F,X\rangle)$$

in a neighborhood of the identity of N_F. Such a representation will exist, if Kirillov's symplectic form of ω is integral (its integral over any two-cycle is an integer). The physical representation is then just the unitary representation of G induced by N_F (see Refs. 25 and 28).

In our case, $M=\Gamma_0$ and as the group we first take SO(2,3). The momentum map Π will be described in terms of a coordinate system in $\mathbf{g}^*$; the coordinate system will be associated with the basis that is dual to (46); let the corresponding coordinates be $(\xi_\mu,\zeta_\mu,\theta^\mu,\delta)$. Then, the momentum map is given by

$$\left.\begin{array}{lll} \xi_0=-\sqrt{\mathbf{p}\cdot\mathbf{p}}, & \zeta_0=-(\mathbf{x}\cdot\mathbf{x})\sqrt{\mathbf{p}\cdot\mathbf{p}}, & \theta^0=x_1p_2-x_2p_1, \\ \xi_1=p_1, & \zeta_1=(\mathbf{x}\cdot\mathbf{x})p_1-2(\mathbf{x}\cdot\mathbf{p})x_1, & \theta^1=-x_2\sqrt{\mathbf{p}\cdot\mathbf{p}}, \\ \xi_2=p_2, & \zeta_2=(\mathbf{x}\cdot\mathbf{x})p_2-2(\mathbf{x}\cdot\mathbf{p})x_2, & \theta^2=x_1\sqrt{\mathbf{p}\cdot\mathbf{p}}, \\ & & \delta=\mathbf{x}\cdot\mathbf{p}. \end{array}\right\} \tag{56}$$

As the group SO(2,3) has a trivial center, the momentum map is a symplectic isomorphism and the homogeneous symplectic space (Γ_0,Ω_0) of G can be identified with the orbit ω. Then, the Kirillov symplectic form Ω_0 is exact and so it is trivially integral.

Let us choose the point F corresponding to the point $u\in\Gamma_0$ that is given by the values of coordinates $x_1=x_2=0$, $p_1=1$, and $p_2=0$. From Eqs. (56), we calculate the coordinates of F to be $(-1,1,0,0,0,0,0,0,0,0)$. The map of $\mathbf{g}$ into $T_F\omega$ is given by the values of the vector fields (47)–(53) at the point u,

$$\mathscr{P}_0\mapsto-\frac{\partial}{\partial x_1},\quad \mathscr{P}_k\mapsto\frac{\partial}{\partial x_k}, \tag{57}$$

$$\mathscr{Q}_0\mapsto 0,\quad \mathscr{Q}_k\mapsto 0, \tag{58}$$

$$\mathscr{T}^0\mapsto\frac{\partial}{\partial p_2},\quad \mathscr{T}^1\mapsto\frac{\partial}{\partial p_2},\quad \mathscr{T}^2\mapsto-\frac{\partial}{\partial p_1}, \tag{59}$$

$$\mathscr{D}\mapsto-\frac{\partial}{\partial p_1}. \tag{60}$$

The kernel of the map is, therefore, the subalgebra $\mathbf{g}_F$ generated by $\mathscr{P}_0+\mathscr{P}_1$, $\mathscr{T}^0-\mathscr{T}^1$, $\mathscr{D}-\mathscr{T}^2$, and $\mathscr{Q}_\mu$, $\mu=0,1,2$. The algebra $\mathbf{g}_F$ has to be extended to $\mathbf{n}_F$. Thus, we have to find a two-dimensional subspace of $\mathbf{g}/\mathbf{g}_F$ which is invariant with respect to $\mathbf{g}_F$. $\mathbf{g}/\mathbf{g}_F$ is four-dimensional; we choose $[\mathscr{P}_1]$, $[\mathscr{P}_2]$, $[\mathscr{T}^1]$, and $[\mathscr{T}^2]$ as its basis. The action of $Y\in\mathbf{g}_F$ on $\mathbf{g}/\mathbf{g}_F$ is given by $[X]\mapsto\pi([X,Y])$, where X is a representation of an element of the basis of $\mathbf{g}/\mathbf{g}_F$, $[X]$ is the corresponding class, and π is the projector from $\mathbf{g}$ to $\mathbf{g}/\mathbf{g}_F$. A straightforward calculation gives

	$\mathscr{P}_0+\mathscr{P}_1$	$\mathscr{T}^0-\mathscr{T}^1$	$\mathscr{D}-\mathscr{T}^2$	$\mathscr{Q}_0$	$\mathscr{Q}_1$	$\mathscr{Q}_2$
$[\mathscr{P}_1]$	0	$[-\mathscr{P}_2]$	0	$[2.\mathscr{T}^2]$	$[2.\mathscr{T}^2]$	$[-2.\mathscr{T}^1]$
$[\mathscr{P}_2]$	0	0	$[-\mathscr{P}_2]$	$[-2.\mathscr{T}^1]$	$[2.\mathscr{T}^1]$	$[2.\mathscr{T}^2]$
$[\mathscr{T}^1]$	$[\mathscr{P}_2]$	$[\mathscr{T}^2]$	$[-\mathscr{T}^1]$	0	0	0
$[\mathscr{T}^2]$	0	0	0	0	0	0

Our task is to find a two-dimensional (in general complex) common invariant subspace of all six transformations. The Abelian subalgebra generated by $\mathscr{Q}_0$, $\mathscr{Q}_1$, and $\mathscr{Q}_2$ is represented by triangular matrices. They have a common invariant subspace T spanned by $[\mathscr{T}^1]$ and $[\mathscr{T}^2]$; there is no complex linear combination $[a\mathscr{P}_1+b\mathscr{P}_2]$ that would be mapped by all $\mathscr{Q}_\mu$, $\mu=0,1,2$, to a one-dimensional subspace of T. Thus, T is the only two-dimensional invariant subspace of all $\mathscr{Q}_\mu$, $\mu=0,1,2$. However, this subspace is not invariant with respect to $\mathscr{P}_0+\mathscr{P}_1$. Hence, no subordinate algebra $\mathbf{n}_F$ exists for the whole group SO(2,3).

In fact, if we just want to have a unitary representation of SO(2,3) by complex functions on a two-dimensional manifold M (this reflects the fact that we have two physical degrees of freedom), then such a representation will determine a definite action of SO(2,3) on M that will be transitive, or else the representation will not be irreducible. Then, $M=\mathrm{SO}(2,3)/G_M$, where G_M is a stabilizer of a point of M. Thus, SO(2,3) had to admit an eight-dimensional subgroup. However, there is no such subgroup.[28]

There is, however, $\mathbf{n}_F$, if we restrict ourselves to some subgroup of SO(2,3); the largest are G_1, generated by $(\mathscr{P}_0,\mathscr{P}_1,\mathscr{P}_2,\mathscr{T}^0,\mathscr{T}^1,\mathscr{T}^2,\mathscr{D})$ and G_2 generated by $(\mathscr{Q}_0,\mathscr{Q}_1,\mathscr{Q}_2,\mathscr{T}^0,\mathscr{T}^1,\mathscr{T}^2,\mathscr{D})$. Γ_0 is still a space where G_1 or G_2 act; they do not act transitively, however: the points with $x_1=x_2=0$ are invariant with respect to G_2 and those with $y_1=y_2=0$

with respect to G_1. These points have to be cut out in respective cases. Thus, the coordinate patch (x_k,p_k) is a homogeneous space of G_1 and (y_k,q_k) that of G_2. The action of the group G_1 on T^*U is the same as that of G_2 on T^*V in the respective coordinates. Let us consider G_1 and $T^*\mathbf{M}^3$ only. In fact, only this case is a group extension of our original system, namely a G_1-extension.

From the corresponding part of the table, we can see immediately that there are two different invariant subspaces; T_1 spanned by $[\mathscr{P}_1]$ and $[\mathscr{P}_2]$ and T_2 spanned by $[\mathscr{P}_2]$ and $[\mathscr{T}^2]$. It is easy to see that there are no others. From T_2, we obtain the subalgebra $\mathbf{n}_{2F}$ generated by $\mathscr{P}_0+\mathscr{P}_1$, $\mathscr{P}_2$, $\mathscr{T}^0-\mathscr{T}^1$, $\mathscr{T}^2$ and $\mathscr{D}$; $\mathbf{n}_{2F}$ satisfies the conditions (1) and (2), but it does not satisfy Pukanszky's condition. Indeed, the subspace $\mathbf{n}_{2F}^{\perp}$ that annihilates it has the form $(a,-a,0,b,b,0,0)$ where $(a,b)\in\mathbf{R}^2$. The subset $F+\mathbf{n}_{2F}^{\perp}$ is given by $(-1+a,1-a,0,b,b,0,0)$. This will lie in ω if the equations

$$-\sqrt{\mathbf{p}\cdot\mathbf{p}}=-1+a,\quad p_1=1-a,\quad p_2=0,$$

$$x_1p_2-x_2p_1=b,\quad -x_2\sqrt{\mathbf{p}\cdot\mathbf{p}}=b,\quad x_1\sqrt{\mathbf{p}\cdot\mathbf{p}}=0,$$

$$x\cdot\mathbf{p}=0,$$

have solutions for any a and b. However, the first equation implies that $a<1$. Thus, this algebra is not admissible. Similar calculation for T_1 shows that the corresponding algebra $\mathbf{n}_{1F}$ satisfies all three conditions and so it is the only possibility. Let us concentrate on $\mathbf{n}_{1F}$, which is generated by $\mathscr{P}_0$, $\mathscr{P}_1$, $\mathscr{P}_2$, $\mathscr{T}^0-\mathscr{T}^1$, and $\mathscr{D}-\mathscr{T}^2$.

The map of the Lie algebra $\mathbf{g}_1$ of the group G_1 into $T_F\omega$ given by Eqs. (47) and (50)–(53) sends $\mathbf{n}_{1F}$ on the subspace $E_F\in T_F\omega$ spanned by the vectors

$$\frac{\partial}{\partial x_1},\quad \frac{\partial}{\partial x_2}. \tag{61}$$

We can use the action ad* of the group G_1 to bring the subspace from the point F to any other point of ω; this is a well-defined procedure, because $\mathbf{n}_{1F}$ is invariant with respect to the stabilizer G_{1F} of F. The result is the subspace spanned by (61) at any point (x_1,x_2,p_1,p_2) of ω. Indeed, we can use the four one-dimensional subgroups of G_1 generated by $\mathscr{P}_1$, $\mathscr{P}_2$, $\mathscr{T}^0$, and $\mathscr{T}^2$. The corresponding vector fields given by Eqs. (47), (50), and (52) describe the action of these generators on ω; their projections to the submanifold $x_1=x_1^0$, $x_2=x_2^0$ for any constant x_1^0 and x_2^0 is independent of x_1^0 and x_2^0. Thus, the curve defined by

$$x_1=\gamma_1(t),\quad x_2=\gamma_2(t),\quad p_1=p_1^0,\quad p_2=p_2^0, \tag{62}$$

with the real constants p_1^0 and p_2^0 will be mapped onto a curve of the form

$$x_1=\gamma_1'(t),\quad x_2=\gamma_2'(t),\quad p_1=p_1'^0,\quad p_2=p_2'^0$$

by any element of the group. Hence, the subspace E_F will be mapped to the subspace $E_{F'}$ spanned by the vectors (61) at any point F' of ω with $x_1=x_2=0$ and p_1 and p_2 arbitrary. The vector fields (61) (which are now used in their role of the action of $\mathscr{P}_1$ and $\mathscr{P}_2$) can easily be integrated; they generate the maps

$$(x_1,x_2,p_1,p_2)\mapsto(x_1+a,x_2+b,p_1,p_2)$$

with arbitrary a and b. Thus, curves of the form (62) are mapped to

$$x_1=\gamma_1(t)+a,\quad x_2=\gamma_2(t)+b,\quad p_1=p_1^0,\quad p_2=p_2^0.$$

Hence, $E_{F'}$ goes over to $E_{F''}$ spanned again by (61) in an arbitrary point F'' of ω. The resulting subbundle of the tangent bundle is called *polarization*. It is an integrable subbundle, its integral manifolds $\mathscr{E}$ being given by $p_1=\text{const}$, $p_2=\text{const}$.

At this stage, it is much quicker to guess the form of the operators representing the Lie algebra than to calculate the representation according to the general procedure by Kostant–Kirillov. The unitary representation of G_1 that we are going to construct is induced by the representation $R_{\mathbf{n}}$ of the subgroup N_{1F} that is generated by the subalgebra $\mathbf{n}_{1F}$. Thus, the Hilbert space will be built from complex functions on the homogeneous space G_1/N_{1F}. This may be identified with the manifold $\Gamma_{\mathrm{ISO}}/\mathscr{E}$ that is just $\mathbf{R}^2\backslash\{0\}$ with the coordinates p_1 and p_2 and which we have denoted by $\mathscr{P}$ in Sec. IV A. If we look at the formula for the induced representation [see e.g., Ref. 28, p. 479, formula (15)], we can see that there will be three kinds of terms in the operators representing the Lie algebra of G_1. From the representation of N_{1F}, multiplicative terms will come; they must clearly be multiplications by $-\sqrt{\mathbf{p}\cdot\mathbf{p}}$, p_1, and p_2 for the operators $\hat{P}_0$, $\hat{P}_1$, and $\hat{P}_2$. From the action of G_1 on the classes G_1/N_{1F}, differential operators come; they must be projections of the vector fields (47), (50)–(53) to the space $\mathscr{P}$ multiplied by $-i$,

$$J^0_{\mathrm{diff}}=ip_2\frac{\partial}{\partial p_1}-ip_1\frac{\partial}{\partial p_2},$$

$$J^1_{\mathrm{diff}}=-i\sqrt{\mathbf{p}\cdot\mathbf{p}}\,\frac{\partial}{\partial p_2},\quad J^2_{\mathrm{diff}}=i\sqrt{\mathbf{p}\cdot\mathbf{p}}\,\frac{\partial}{\partial p_1},$$

$$D_{\mathrm{diff}}=ip_1\frac{\partial}{\partial p_1}+ip_2\frac{\partial}{\partial p_2}.$$

Finally, there will be terms coming from the Radon–Nikodym derivative that will correct the differential operators. Such terms have the general form

$$\frac{i}{2\sigma}\,(\xi_{\mathrm{diff}}\sigma),$$

where σ is a quasi-invariant measure on G_1/N_{1F}. Different but equivalent measures will lead to unitarily equivalent representations. A choice that strongly simplifies the correction terms is

$$\sigma=\frac{1}{\sqrt{\mathbf{p}\cdot\mathbf{p}}}.$$

Then, finally, the operators must have the form

$$\hat{P}_0\psi(p)=-\sqrt{\mathbf{p}\cdot\mathbf{p}}\psi(p),\quad \hat{P}_k\psi(p)=p_k\psi(p), \tag{63}$$

$$\hat{J}^0\psi(p)=ip_2\frac{\partial\psi(p)}{\partial p_1}-ip_1\frac{\partial\psi(p)}{\partial p_2}, \tag{64}$$

$$\hat{J}^1\psi(p)=-i\sqrt{\mathbf{p}\cdot\mathbf{p}}\,\frac{\partial\psi(p)}{\partial p_2},\quad \hat{J}^2=i\sqrt{\mathbf{p}\cdot\mathbf{p}}\,\frac{\partial\psi(p)}{\partial p_1}, \tag{65}$$

$$\hat{D}=ip_1\frac{\partial\psi}{\partial p_1}+ip_2\frac{\partial\psi}{\partial p_2}+\frac{i}{2}\,\psi. \tag{66}$$

It is straightforward but tedious to verify that this guessed operators coincide with those that would follow from the full general construction of the representation.

An interesting question is: What happened with the relations? For the group G_1, we have only three relations, and we can take Eqs. (41) and (45). Composing the corresponding operators on a common invariant domain [say, $C^\infty_0(\mathscr{P})$], we obtain easily,

$$-\hat{P}_0^2+\hat{P}_1^2+\hat{P}_2^2=0, \tag{67}$$

$$\hat{J}^0\hat{P}_0+\hat{J}^1\hat{P}_1+\hat{J}^2\hat{P}_2=0, \tag{68}$$

$$-\hat{D}^2-(\hat{J}^0)^2+(\hat{J}^1)^2+(\hat{J}^2)^2=\tfrac{1}{4}. \tag{69}$$

Thus only the last relation has been deformed (of course, there will be $\hbar^2/4$ at the right-hand side, if $\hbar$, which has been set equal 1, is restored). The left-hand sides of Eqs. (67)–(69) are generalized Casimirs of the group G_1 in the following sense. Let $E(\mathbf{g}_1)$ be the envelopping algebra of $\mathbf{g}_1$, let $\mathscr{H}=-\mathscr{P}_0^2+\mathscr{P}_1^2+\mathscr{P}_2^2$ represent the constraint and let $\mathscr{I}(\mathscr{H})$ be the ideal in $E(\mathbf{g}_1)$ generated by $\mathscr{H}$. Then, the classes of the left-hand sides of Eqs. (67)–(69) in $E(\mathbf{g}_1)/\mathscr{I}(\mathscr{H})$ commute with all elements of $E(\mathbf{g}_1)/\mathscr{I}(\mathscr{H})$.

It is interesting to observe that each state ψ of the Hilbert space $\mathscr{K}$ of the physical representation must satisfy the equation

$$(-\hat{P}_0^2+\hat{P}_1^2+\hat{P}_2^2)\psi=0, \tag{70}$$

which follows from Eq. (67). This is a point where the group method touches the operator constraint method of quantization, because Eq. (70) has the form of the operator constraint equation of our system (for more detail, see Ref. 20).

ACKNOWLEDGMENTS

Helpful discussions with J. Bičák, K. Kuchař, R. Loll, V. Moncrief, and L. Ziewer are gratefully acknowledged. The author is thankful to the National Science Foundation Grant No. PHY9507719, to the Tomalla Foundation, Zurich and to the Swiss Nationalfonds for support. Thanks go to the Max-Planck-Institut für Gravitationsphysik, Potsdam for hospitality and support.

[1] S. Carlip, ''Six ways to quantize $(2+1)$-dimensional gravity'' in *General Relativity and Relativistic Astrophysics, Proceedings*, edited by R. B. Mann and R. G. McLenaghan (World Scientific, Singapore, 1994), pp. 215–234.
[2] R. Loll, J. Math. Phys. **36**, 6494 (1995).
[3] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Yeshiva University Press, New York, 1964).
[4] K. V. Kuchař, ''Canonical quantum gravity,'' in *General Relativity and Gravitation 1992*, edited by R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi (IOP, Bristol, 1993), pp. 119–150.
[5] P. Hájíček, Class. Quantum Grav. **13**, 1353 (1996).
[6] P. G. Bergmann, Rev. Mod. Phys. **33**, 510 (1961).
[7] A. Ashtekar, *Lectures on Nonperturbative Canonical Gravity* (World Scientific, Singapore, 1991).
[8] K. V. Kuchař, J. Math. Phys. **22**, 2640 (1981).
[9] Anderson and C. Torre, Commun. Math. Phys. **176**, 479 (1996).
[10] C. J. Isham, ''Topological and global aspects of quantum theory,'' in *Relativity, Groups and Topology II*, edited by B. C. DeWitt and R. Stora (Elsevier, New York, 1984).
[11] C. Rovelli, Nuovo Cimento B **100**, 343 (1987).
[12] P. A. M. Dirac, Rev. Mod. Phys. **21**, 392 (1949).
[13] P. Hájíček, J. Math. Phys. **36**, 4612 (1995).
[14] E. Witten, Nucl. Phys. B **311**, 46 (1988).
[15] S. P. Martin, Nucl. Phys. B **327**, 178 (1989).
[16] R. Loll, Class. Quantum Grav. **12**, 1655 (1995).
[17] V. Moncrief, J. Math. Phys. **31**, 2978 (1990).
[18] V. Moncrief, J. Math. Phys. **30**, 2907 (1989).
[19] N. Manojlovic and A. Mikovic, Nucl. Phys. B **385**, 571 (1992).
[20] K. Pohlmayer, Commun. Math. Phys. **114**, 351 (1988).
[21] L. Andersson (private communication).
[22] J. Louko and D. Marolf, Class. Quantum Grav. **11**, 311 (1994).
[23] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications. I,II* (Springer, Berlin, 1985/88).
[24] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, Cambridge, 1973).
[25] A. A. Kirillov, *Elements of the Theory of Representations* (Springer, Berlin, 1976).
[26] R. Penrose and W. Rindler, *Spinors and Space–Time* (Cambridge University Press, Cambridge, 1986), Vol. II, p. 297.
[27] B. Kostant, in *Lectures in Modern Analysis and Applications. III*, edited by C. J. Taam (Springer, Berlin, 1970).
[28] A. O. Barut and R. Rączka, *Theory of Group Representations and Applications* (Polish Scientific Publishers, Warsaw, 1980).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Existence theorems for hairy black holes in $\mathfrak{su}(N)$ Einstein–Yang–Mills theories

N. E. Mavromatos and E. Winstanley[a)]
University of Oxford, Theoretical Physics, 1 Keble Road OX1 3NP, United Kingdom

(Received 15 December 1997; accepted for publication 13 April 1998)

We establish the existence of hairy black holes in $\mathfrak{su}(N)$ Einstein–Yang–Mills theories, described by $N-1$ parameters, corresponding to the nodes of the gauge field functions. © *1998 American Institute of Physics.* [S0022-2488(98)01807-6]

I. INTRODUCTION

Interest in non-Abelian Einstein–Yang–Mills theories was sparked by the discovery of particlelike[1] and non-Reissner–Nordström (''hairy'') black hole[2] solutions when the gauge group is $\mathfrak{su}(2)$. Since then a plethora of hairy black holes, possessing nontrivial geometry and a field structure outside the event horizon, have been found (see, for example, Ref. 3), including colored black holes in $\mathfrak{su}(N)$ Einstein–Yang–Mills theories.[4,5] Many of these objects have as a fundamental requirement for the existence of hair, a non-Abelian gauge field, often coupled to other fields, such as a Higgs field.[6] As in Ref. 6 the existence of gauge hair is not surprising in itself, since the gauge field force is long range. However, the non-Abelian nature of the field is important for evading the no-hair theorem, for example, for scalar fields coupled to the gauge field. The vast majority of these solutions have been found only numerically, with analytic work in this area at present being limited to an extensive study of the $\mathfrak{su}(2)$ case,[7–9] and some analysis of the field equations for general N.[10,11]

In this paper we continue this analysis of the coupled Einstein–Yang–Mills equations for an $\mathfrak{su}(N)$ gauge field, and prove analytically the existence of ''genuine'' hairy black hole solutions for every N. By a ''genuine'' $\mathfrak{su}(N)$ black hole, we mean a solution which is not simply the result of embedding a smaller gauge group in $\mathfrak{su}(N)$. As in the $\mathfrak{su}(2)$ case, the solutions are labeled by the number of nodes of each of the $N-1$ nonzero functions required to describe the gauge field. We shall prove that for each integer n_1, there are an infinite number of sequences of integers $n_{N-1} \geqslant n_{N-2} \geqslant \cdots \geqslant n_1$ corresponding to black hole solutions. It is to be expected that, in fact, *every* such sequence corresponds to a black hole solution, but unfortunately we are unable to prove this analytically, although we shall present a numerically based argument for $\mathfrak{su}(3)$ and numerical investigations [such as that done for $\mathfrak{su}(5)$ in Ref. 4] for higher dimension groups which indicate that this is, in fact, the case. The method used to prove the main theorem of this paper is remarkably simple, drawing only on elementary topological ideas.

The structure of the paper is as follows. In Sec. II we review briefly the $\mathfrak{su}(N)$ Einstein–Yang–Mills field equations and the ansatz and notations we shall employ in the rest of the paper. Next, we state some elementary properties of these equations, including results from Ref. 11. The remainder of the paper has a similar progression of ideas as in Ref. 7, and we continue by first analyzing the behavior of solutions to the field equations in the two asymptotic regimes, at infinity and close to the event horizon. Although these forms were discussed in Ref. 11, we present here a shorter proof. Section V is devoted to a discussion of the flat space solutions, which will be important for later propositions. The results here are somewhat weaker than in Ref. 7 due to the fact that for $\mathfrak{su}(N)$, we have $N-1$ variables and therefore have a $2(N-1)$-dimensional phase space rather than a phase plane as in the $\mathfrak{su}(2)$ case, so the powerful Poincaré–Bendixson theory no longer applies. We now consider integrating the field equations outward from the event horizon and consider the various possible behaviors of the resulting solutions. As in Ref. 7 there are three types of solutions: the regular black holes we are seeking, singular solutions (in which the lapse function vanishes outside the event horizon), and oscillating solutions in which the geometry is not

a)Electronic mail: winstan@thphys.ox.ac.uk

 © 1998 American Institute of Physics

asymptotically flat. The main results of this paper are in Secs. VII and VIII. An inductive argument is used to prove the existence of solutions for $\mathfrak{su}(N)$, assuming existence for $\mathfrak{su}(N-1)$ [since we have rigorous theorems for $\mathfrak{su}(2)$[7,8]]. The argument is presented in detail for $\mathfrak{su}(3)$ in Sec. VII and a brief outline of the extension to general N in Sec. VIII. Finally, a summary and our conclusions are presented in Sec. IX.

II. ANSATZ AND FIELD EQUATIONS

In this section we first describe the ansatz we are using and outline the field equations.

The field equations for an $\mathfrak{su}(N)$ Yang–Mills gauge field coupled to gravity have been derived in Ref. 10 for a spherically symmetric geometry. We take the line element, in the usual Schwarzschild coordinates, to be

$$ds^2=-S^2\mu dt^2+\mu^{-1}dr^2+r^2d\theta^2+r^2\sin^2\theta d\phi^2, \tag{1}$$

where the metric functions μ and S are functions of r alone and

$$\mu(r)=1-\frac{2m(r)}{r}. \tag{2}$$

A spherically symmetric $\mathfrak{su}(N)$ gauge potential may be written in the form[10]

$$\mathcal{A}=\mathcal{A}dt+Bdr+\frac{1}{2}(C-C^H)d\theta-\frac{i}{2}[(C+C^{\mathrm{H}})\sin\theta+D\cos\theta]d\phi, \tag{3}$$

where $D=\mathrm{Diag}\{k_1,k_2,\ldots,k_N\}$ with $k_1\geqslant k_2\geqslant\cdots\geqslant k_N$ integers whose sum is zero. In addition, C is a strictly upper triangular complex matrix such that $C_{ij}\neq 0$ only if $k_i=k_j+2$ and C^{H} its Hermitian conjugate, and A, B are anti-Hermitian matrices that commute with D. An irreducible representation of $\mathfrak{su}(N)$ can be constructed by taking[10]

$$D=\mathrm{Diag}\{N-1,N-3,\ldots,-N+3,-N+1\}. \tag{4}$$

In this case A, B have trace zero and can be written as

$$A_{jj}=i\left\{-\frac{1}{N}\sum_{k=1}^{j-1}k\mathcal{A}_k+\sum_{k=j}^{N-1}\left(1-\frac{k}{N}\right)\mathcal{A}_k\right\}, \tag{5}$$

for real functions $\mathcal{A}_k$, and similarly for B. The only nonvanishing entries of C are

$$C_{j,j+1}=\omega_j e^{i\gamma}, \tag{6}$$

where ω_j and γ_j are real functions. All the functions in this ansatz depend only on r. Now we make the following simplifying assumption:[10]

$$\mathcal{A}_j=0,\quad \mathcal{B}_j+\gamma_j'=0,\quad \forall_j. \tag{7}$$

The remaining gauge freedom can then be used to set $B=0$,[12] which implies that the γ_j are constants, which we choose to be zero for simplicity.

The gauge field equations for the ω_j then take the form[10,11] (where we have set $\kappa=2$ in Ref. 10)

$$r^2\mu\omega_j''+(2m-2r^3p_\theta)\omega_j'+[1-\omega_j^2+\tfrac{1}{2}(\omega_{j-1}^2+\omega_{j+1}^2)]\omega_j=0, \tag{8}$$

where

$$p_\theta=\frac{1}{4r^4}\sum_{j=1}^{N}[\omega_j^2-\omega_{j-1}^2-N-1+2j]^2, \tag{9}$$

and the Einstein equations can be simplified to

$$m' = \mu G + r^2 p_\theta, \quad \frac{S'}{S} = \frac{2G}{r}, \tag{10}$$

where

$$G = \sum_{j=1}^{N-1} \omega_j'^2. \tag{11}$$

Note that $\omega_0 \equiv 0 \equiv \omega_N$, so that these are the usual equations[2] in the $\mathfrak{su}(2)$ case, and have a very similar, but slightly coupled, structure for general N. The equations (8)–(11) possess two symmetries.[11] First, the substitution $\omega_j \to -\omega_j$ for any fixed j leaves the equations invariant, exactly as in the $\mathfrak{su}(2)$ case. Second, there is an additional symmetry under the transformation $j \to N-j$ for all j. We will not assume that this symmetry is respected by the solutions of the field equations, so that the $\mathfrak{su}(3)$ case is not necessarily trivial.

We are concerned in this paper with black holes, which will have a regular event horizon at $r = r_h$, where $\mu = 0$. The equations in their present form are singular at r_h, so in order to produce a set of equations which are regular as $\mu \to 0$, we define a new independent variable τ by[7]

$$\frac{dr}{d\tau} = r\sqrt{\mu}, \tag{12}$$

denote $d/d\tau$ by; and define new dependent variables κ, U_j, and Ψ as follows:[7]

$$\Psi = \sqrt{\mu}, \quad U_j = \Psi \omega_j', \quad \kappa = \frac{1}{2\Psi}\,(1 + \Psi^2 + 2\mu G - 2r^2 p_\theta). \tag{13}$$

Then the field equations take the form[7]

$$\dot{r} = r\Psi, \tag{14}$$

$$\dot{\omega}_j = r U_j, \tag{15}$$

$$\dot{\Psi} = (\kappa - \Psi)\Psi - 2\mu G, \tag{16}$$

$$(S\Psi)^{\cdot} = S(\kappa - \Psi)\Psi, \tag{17}$$

$$\dot{U}_j = -(\kappa - \Psi)U_j - \frac{1}{r}\left[1 - \omega_j^2 + \frac{1}{2}\,(\omega_{j+1}^2 + \omega_{j-1}^2)\right]\omega_j, \tag{18}$$

$$\dot{\kappa} = -\kappa^2 + 1 + 2\mu G. \tag{19}$$

The main thrust of this article is to prove the existence of regular, black hole solutions of the field equations for general N with $N-1$ degrees of freedom. In other words, we begin with a regular event horizon at $r = r_h$, where $\mu = 0$, and integrate outward with increasing r. Later we shall classify the possible behavior of the solutions as r increases. We are interested primarily in those solutions which possess the physical properties of a black hole geometry, namely, for which $\mu > 0$ for all $r > r_h$, the ω_j and their derivatives are finite for all $r > r_h$ and the spacetime becomes flat in the limit $r \to \infty$. We shall refer to such solutions as *regular black hole solutions*.

III. ELEMENTARY RESULTS

In this section we state a few elementary results, which will prove useful in a later analysis. First, two lemmas which are proved in the $\mathfrak{su}(N)$ case exactly as in the $\mathfrak{su}(2)$ case; see Ref. 7.

Lemma III.1: If $\mu(r_0) < 1$ *for some* r_0 *then* $\mu(r) < 1$ *for all* $r \geqslant r_0$.

Lemma III.2: As long as $0 < \mu < 1$ *all field variables are regular functions of* r.

These two lemmas show that for a regular event horizon at $r=r_h$, where $\mu=0$, then $\mu<1$ for all $r>r_h$ and the field variables are regular functions as long as $\mu>0$. The black hole solutions we seek approach the Schwarzschild geometry as $r\to\infty$, so we may assume that all ω_j are bounded for all r (from Lemma III.2 each ω_j is bounded on every closed interval $[r_h, r_1]$, so we are simply assuming that ω_j remains bounded as $r\to\infty$). It will be proved in Sec. VI that this assumption is, in fact, valid for solutions having $\mu>0$ for all r. Define a quantity M_j for each j to be the lowest upper bound on ω_j^2, i.e.,

$$\omega_j^2\leqslant M_j, \quad \forall_j=1,2,...,N-1. \tag{20}$$

It has been shown in Ref. 8 that the following result is true for $N=2$, and proved, in general, with the assumption that all ω_j are bounded via an elegant method in Ref. 11.

Theorem III.3 $M_j\leqslant j(N-j)$:

Part of the proof of this theorem in Ref. 11 involves a knowledge of where ω_j may have maxima or minima. The result following is less powerful than in the $\mathfrak{su}(2)$ case due to the coupling in the gauge field equation (8).

Proposition III.4: As long as $\mu(r)>0$, the function $\omega_j(r)$ cannot have maxima in the regions

$$\omega_j>\sqrt{1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)} \quad \textit{and} \quad 0>\omega_j>-\sqrt{1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)}, \tag{21}$$

or minima in the regions

$$\omega_j<-\sqrt{1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)} \quad \textit{and} \quad 0<\omega_j<\sqrt{1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)}. \tag{22}$$

Proof: When $\omega_j'=0$, from (8),

$$\mu\omega_j''=-\frac{1}{r^2}\left[1-\omega_j^2+\frac{1}{2}\,(\omega_{j+1}^2+\omega_{j-1}^2)\right]\omega_j, \tag{23}$$

so that ω_j will have a maximum if

$$\omega_j>0 \quad \text{and} \quad \omega_j^2<1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2), \tag{24}$$

or

$$\omega_j<0 \quad \text{and} \quad \omega_j^2>1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2). \tag{25}$$

Similarly, ω_j will have a minimum if

$$\omega_j>0 \quad \text{and} \quad \omega_j^2>1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2), \tag{26}$$

or

$$\omega_j<0 \quad \text{and} \quad \omega_j^2<1+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2). \tag{27}$$

□

IV. ASYMPTOTIC BEHAVIOR OF SOLUTIONS

We seek solutions to the above field equations representing black holes with a regular event horizon at $r=r_h$ and finite total energy density. In order to prove the local existence of solutions of the field equations with the desired asymptotic behavior, we shall apply the following theorem.[7]

Theorem IV.1: *Consider a system of differential equations for $n+m$ functions* $\mathbf{a}=(a_1,a_2,\ldots,a_n)$ *and* $\mathbf{b}=(b_1,b_2,\ldots,b_m)$ *of the form*

$$x\,\frac{da_i}{dx}=x^{p_i}f_i(x,\mathbf{a},\mathbf{b}), \quad x\,\frac{db_i}{dx}=-\lambda_i b_i+x^{q_i}g_i(x,\mathbf{a},\mathbf{b}), \tag{28}$$

with constants $\lambda_i > 0$ *and integers* p_i, $q_i \geqslant 1$ *and let* $\mathscr{C}$ *be an open subset of* $\mathbb{R}^n$ *such that the functions* f_i *and* g_i *are analytic in a neighborhood of* $x=0$, $\mathbf{a}=\mathbf{c}$, $\mathbf{b}=\mathbf{0}$, *for all* $\mathbf{c} \in \mathscr{C}$. *Then there exists an n-parameter family of solutions of the system (28) such that*

$$a_i(x) = c_i + O(x^{p_i}), \quad b_i(x) = O(x^{q_i}), \tag{29}$$

where $a_i(x)$ *and* $b_i(x)$ *are defined for* $\mathbf{c} \in \mathscr{C}$, *and* $|x| < x_0(\mathbf{c})$ *and are analytic in* x *and* $\mathbf{c}$.

In this section we consider only the equations for μ and the ω_j. The equation for S (10) is independent of the others and hence can be integrated immediately once we have the other field variables, with an additional parameter. This procedure gives a finite answer on any closed, bounded interval on which the remaining variables are finite. Results similar to Propositions IV.2 and IV.3 were proved in Ref. 11, but here we are using a more straightforward approach.

A. Behavior at infinity

As $r \to \infty$, in order to have finite total energy density, the geometry must approach the Schwarzschild solution, that is,

$$m \to M = \text{const}, \quad G \to 0, \quad p_\theta \to 0. \tag{30}$$

Hence each ω_j must approach a constant value, which is fixed by (30) to be

$$\omega_j^2 \to j(N-j). \tag{31}$$

The equations for the ω_j are then automatically satisfied in the limit as $r \to \infty$.

Proposition IV.2: There exists an N-parameter family of local solutions of (8)–(11) near $r = \infty$ *analytic in* c_j, M, *and* r^{-1} *such that*

$$\mu(r) = 1 - \frac{2M}{r} + O\left(\frac{1}{r^4}\right), \quad \omega_j(r) = \sqrt{j(N-j)} - \frac{c_j}{r} + O\left(\frac{1}{r^2}\right). \tag{32}$$

Proof: Introduce new variables,

$$x = \frac{1}{4}, \quad \psi_j = r(\sqrt{j(N-j)} - \omega_j), \quad \xi_j = r^2 \mu \omega_j', \quad \lambda = r(1-\mu). \tag{33}$$

Then

$$x \frac{d\lambda}{dx} = -\frac{2}{x}(\mu G + r^2 p_\theta), \tag{34}$$

where

$$G = x^4 \sum_{j=1}^{N-1} \frac{\xi_j^2}{\mu^2} \tag{35}$$

and

$$p_\theta = \frac{x^4}{4} \sum_{j=1}^{N} [x^2 \psi_j^2 - x^2 \psi_{j-1}^2 - 2x\psi_j \sqrt{j(N-j)} + 2x\psi_{j-1}\sqrt{(j-1)(N-j+1)}]^2, \tag{36}$$

that is,

$$x \frac{d\lambda}{dx} = -x^3 f_\lambda, \tag{37}$$

where f_λ is a polynomial in x, μ^{-1}, ξ's, and ψ's. Similarly,

$$x\frac{d\psi_j}{dx}=-\psi_j+\frac{1}{\mu}\xi_j=-\psi_j+\xi_j+xf_{\psi_j}, \tag{38}$$

where f_{ψ_j} is analytic in x, λ, ψ's, and ξ's in a neighborhood of $x=0$. Using the field equations, we also have

$$\begin{aligned}x\frac{d\xi_j}{dx}=&-2\xi_j+2\psi_j[j(N-j)]-\psi_{j-1}\sqrt{j(j-1)(N-j)(N-j+1)}\\&-\psi_{j+1}\sqrt{j(j+1)(N-j)(N-j-1)}+xf_{\xi_j},\end{aligned} \tag{39}$$

where f_{ξ_j} is a polynomial in x, μ^{-1}, ψ's, and ξ's. The algebra simplifies if we define new functions α_j, β_j such that

$$\psi_j=\alpha_j+\beta_j, \quad \xi_j=\alpha_j+\Lambda\beta_j, \tag{40}$$

where Λ is a real constant not equal to unity. The equations for α_j and β_j then read as

$$\begin{aligned}(\Lambda-1)x\frac{d\alpha_j}{dx}=&2\alpha_j-2[j(N-j)]\alpha_j+\sqrt{j(j-1)(N-j)(N-j+1)}\alpha_{j-1}\\&+\sqrt{j(j+1)(N-j)(N-j-1)}\alpha_{j+1}+(\Lambda^2+\Lambda)\beta_j-2[j(N-j)]\beta_j\\&+\sqrt{j(j-1)(N-j)(N-j+1)}\beta_{j-1}+\sqrt{j(j+1)(N-j)(N-j-1)}\beta_{j+1}+xf_{\alpha_j},\end{aligned} \tag{41}$$

$$\begin{aligned}(\Lambda-1)x\frac{d\beta_j}{dx}=&-2\alpha_j+2[j(N-j)]\alpha_j-\sqrt{j(j-1)(N-j)(N-j+1)}\alpha_{j-1}\\&-\sqrt{j(j+1)(N-j)(N-j-1)}\alpha_{j+1}+(1-3\Lambda)\beta_j+2[j(N-j)]\beta_j\\&-\sqrt{j(j-1)(N-j)(N-j+1)}\beta_{j-1}-\sqrt{j(j+1)(N-j)(N-j-1)}\beta_{j+1}+xf_{\beta_j},\end{aligned}$$

where the f_{α_j} and f_{β_j} are analytic functions of x, μ^{-1}, α's, and β's. Consider the matrix $\mathscr{M}_{N-1}$ whose entries are

$$\begin{pmatrix}2(N-1) & -\sqrt{(N-1)2(N-2)} & \cdots\\ -\sqrt{(N-1)2(N-2)} & 2.2(N-2) & \cdots\\ 0 & -\sqrt{2(N-2)3(N-3)} & \cdots\\ \vdots & \vdots & \vdots\end{pmatrix}. \tag{42}$$

For a vector $\mathbf{q}=(q_1,\ldots,q_{N-1})$, let $p_j=\sqrt{j(N-j)}q_j$ for $j=1,\ldots,N-1$. Then

$$\mathbf{q}^T\mathscr{M}_{N-1}\mathbf{q}=p_1^2+p_{N-1}^2+(p_1-p^2)^2+\cdots(p_{N-2}-p_{N-1})^2. \tag{43}$$

Hence, the matrix $\mathscr{M}_{N-1}$ is real, symmetric, and positive definite, and will have positive real eigenvalues $\mathscr{E}_i$, and corresponding eigenvectors $\mathbf{v}_i$. We may expand our variable vectors in terms of eigenvectors of $\mathscr{M}_{N-1}$ as follows:

$$\boldsymbol{\alpha}=\sum_{i=1}^{N-1}A_i(x)\mathbf{v}_i, \quad \boldsymbol{\beta}=\sum_{i=1}^{N-1}B_i(x)\mathbf{v}_i, \tag{44}$$

$$\mathbf{f}_\alpha=\sum_{i=1}^{N-1}F_i(x)\mathbf{v}_i, \quad \mathbf{f}_\beta=\sum_{i=1}^{N-1}G_i(x)\mathbf{v}_i, \tag{45}$$

in terms of which the equations (41) now become

$$(\Lambda-1)x\,\frac{dA_i}{dx}=(2-\mathscr{E}_i)A_i+(\Lambda^2+\Lambda-\mathscr{E}_i)B_i+xF_i\,,$$
$$(\Lambda-1)x\,\frac{dB_i}{dx}=(\mathscr{E}_i-2)A_i+(1-3\Lambda+\mathscr{E}_i)B_i+xG_i\,. \tag{46}$$

This transformation has removed the coupling between the j's, but the A's and B's are still coupled. Repeating the procedure with the matrix $\mathscr{P}_i$ will decouple these equations, where

$$\mathscr{P}_i=\begin{pmatrix} 2-\mathscr{E}_i & \Lambda^2+\Lambda-\mathscr{E}_i \\ -2+\mathscr{E}_i & 1-3\Lambda+\mathscr{E}_i \end{pmatrix}. \tag{47}$$

This matrix has eigenvalues

$$\mu_\pm=\tfrac{1}{2}(1-\Lambda)(3\pm\sqrt{1+4\mathscr{E}_i}), \tag{48}$$

and corresponding eigenvectors $\mathbf{V}_\pm$. Writing

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix}=C_+(x)\mathbf{V}_++C_-(x)\mathbf{V}_-\,,\quad \begin{pmatrix} F_i \\ G_i \end{pmatrix}=D_+(x)\mathbf{V}_++D_-(x)\mathbf{V}_-\,, \tag{49}$$

the equations are now

$$(\Lambda-1)x\,\frac{dC_\pm}{dx}=\mu_\pm C_\pm+xD_\pm\,. \tag{50}$$

In the case where $\mu_-<0$, Theorem IV.1 applies directly, and we have

$$C_-=O(x); \tag{51}$$

when $\mu_-=0$ (corresponding to $\mathscr{E}_i=2$),

$$C_-=K_-+O(x), \tag{52}$$

for some constant K_-. The other eigenvalue μ_+ is always strictly positive, and in this situation Theorem IV.1 does not apply. However, Eq. (50) can be integrated using the standard Picard method (see, for example, Ref. 13) to give an analytic solution,

$$C_+=K_++O(x). \tag{53}$$

Both $C_\pm$ will be analytic in $K_\pm$ and x in some neighborhood of $x=0$ and substituting back through Eqs. (44) and (49) gives solutions of the form (32), as required. □

B. Behavior at the event horizon

We assume that at $r=r_h$ there is a nondegenerate event horizon, namely, $\mu(r_h)=0$, but $\mu'(r_h)>0$ is finite.

Proposition IV.3: There exists an N-parameter family of local solutions of (8)–(11) near $r=r_h$, *analytic in* r_h, $\omega_{j,h}$ *and* r *such that*

$$\mu(r_h+\rho)=\mu'(r_h)+O(\rho),$$
$$\omega_j(r_h+\rho)=\omega_{j,h}+w_j'(r_h)+O(\rho^2), \tag{54}$$

where $\mu'(r_h)$ *and* $\omega_j'(r_h)$ *are functions of the* $\omega_{j,h}$.

Proof: Let $\rho=r-r_h$ be the new independent variable, and define

$$x=r,\quad \lambda=\frac{\mu}{\rho},\quad \psi_j=\omega_j\,,\quad \xi_j=\frac{\mu\omega_j'}{\rho}. \tag{55}$$

Then the field equations take the form

$$\rho \frac{dx}{d\rho} = \rho, \tag{56}$$

$$\rho \frac{d\lambda}{d\rho} = -\lambda + \left[\frac{1}{x} - 2xp_\theta\right] + \rho \frac{\lambda}{x}[1-2G] = -\lambda + \rho H_\lambda + F_\Lambda, \tag{57}$$

$$\rho \frac{d\psi_j}{d\rho} = \frac{\rho \xi_j}{\lambda}, \tag{58}$$

$$\rho \frac{d\xi_j}{d\rho} = -\xi_j + \rho H_j + F_j, \tag{59}$$

where the F's and H's are polynomials in x^{-1}, λ^{-1}, and the other variables, and the F's depend only on x and ψ_j's. Next, define

$$\tilde{\xi}_j = \xi_j - F_j, \quad \tilde{\lambda} = \lambda - F_\lambda, \tag{60}$$

whose derivatives are given by

$$\rho \frac{d\tilde{\xi}_j}{d\rho} = -\tilde{\xi}_j + \rho G_j, \tag{61}$$

$$\rho \frac{d\tilde{\lambda}}{d\rho} = -\tilde{\lambda} + \rho G_\lambda. \tag{62}$$

Here the G's are analytic in $x^{-1}, \lambda^{-1}, \lambda, x, \tilde{\lambda}, \psi_j, \tilde{\xi}_j$. Applying Theorem IV.1, there exist solutions of the form

$$x = r_h + \rho, \quad \psi_j = \omega_{j,h} + O(\rho), \quad \tilde{\lambda}, \tilde{\xi}_y = O(\rho), \tag{63}$$

which gives the behavior (54), together with the required analyticity. From the field equations (8)–(11), setting $\mu(r_h) = 0$, gives the following relations:

$$\mu'(r_h) = \tfrac{1}{2} r_h^2 p_\theta(r_h), \tag{64}$$

$$\omega_j'(r_h) = -\frac{[1 - \omega_{j,h}^2 + \frac{1}{2}(\omega_{j-1,h}^2 + \omega_{j+1,h}^2)\omega_{j,h}]}{r_h - r_h^3 \rho_\theta(r_h)}, \tag{65}$$

where

$$p_\theta(r_h) = \frac{1}{4r_h^4} \sum_{j=1}^{N} [\omega_{j,h}^2 - \omega_{j-1,h}^2 - N - 1 + 2j]^2. \tag{66}$$

□

V. FLAT SPACE SOLUTIONS

The behavior of the gauge field equations in the flat space limit will be useful later in Sec. VI, when we come to examine the properties of regular solutions. We are not concerned here with the global existence of flat space solutions, but only the local properties pertinent to the curved space problem.

The analysis here is similar to that in Ref. 7, but note that for general N there will be $N-1$ coupled gauge field equations. The powerful Poincaré–Bendixson theory of autonomous systems is not applicable when $N > 2$, and so we will be able to derive only correspondingly weaker results,

and cannot draw a phase portrait. Fortunately, the theorems later require only a knowledge of the nature of the critical points and the local behavior close to these points, which is the subject of this section.

In flat space the gauge field equations reduce to

$$r^2\omega_j''+[1-\omega_j^2+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)]\omega_j=0. \tag{67}$$

These equations can be made autonomous by changing variables to $\tau=\log r$, and denoting $d/d\tau$ by

$$\ddot{\omega}_j-\dot{\omega}_j+[1-\omega_j^2+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)]\omega_j=0. \tag{68}$$

This system of $N-1$ coupled equations has critical points when

$$[1-\omega_j^2+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)]\omega_j=0, \quad \forall j=1,2,...,N-1. \tag{69}$$

If $\omega_j\neq 0$ for all j, then

$$\omega_j=\sqrt{j(N-j)}. \tag{70}$$

There is also a critical point when $\omega_j=0$ for all j.

The other critical points can be described as follows. Suppose that $\omega_i=0$ and $\omega_{i+k}=0$, but that $\omega_{i+m}\neq 0$ for $m=1,...,k-1$ (where $k\geqslant 2$). The critical point is then described by the equations

$$\begin{aligned} 0&=\omega_i\,,\\ 0&=1-\omega_{i+1}^2+\tfrac{1}{2}(\omega_i^2+\omega_{i+2}^2),\\ &\vdots\\ 0&=1-\omega_{i+k-1}^2+\tfrac{1}{2}(\omega_{i+k-2}^2+\omega_{i+k}^2),\\ 0&=\omega_{i+k}\,, \end{aligned} \tag{71}$$

which have the solution

$$\omega_{i+m}=\sqrt{m(k-m)}, \quad \text{for } m=1,...,k-1. \tag{72}$$

In other words, if we have a run of k nonzero ω's with zero ω_i at each end, then the solution for the nonzero ω's is the same as the $N=k$ case with no zero ω's. We can also put together a run of as many zero ω's as we like in the critical point.

In order to classify the critical points, we linearize the field equations. Let

$$\omega_j(\tau)=\omega_j^{(0)}+\epsilon_j(\tau), \tag{73}$$

where $\omega_j^{(0)}$ is the value of ω_j at the critical point and ϵ_j is a small perturbation. The equation for ϵ_j is, to first order.

$$0=\ddot{\epsilon}_j-\dot{\epsilon}_j+\epsilon_j[1-\omega_j^{(0)2}+\tfrac{1}{2}(\omega_{j+1}^{(0)2}+\omega_{j-1}^{(0)2})]+\omega_j^{(0)}[-2\epsilon_j\omega_j^{(0)}+\epsilon_{j+1}\omega_{j+1}^{(0)}+\epsilon_{j-1}\omega_{j-1}^{(0)}]. \tag{74}$$

We shall take the two cases we need to consider in turn.

First, the case where $\omega_j^{(0)}=0$, when the equation for ϵ_j reduces to

$$\ddot{\epsilon}_j-\dot{\epsilon}_j+\epsilon_j[1+\tfrac{1}{2}(\omega_{j+1}^{(0)2}+\omega_{j-1}^{(0)2})]=0. \tag{75}$$

In order to find the nature of the critical point, let

$$\epsilon_j=e^{\lambda\tau}; \tag{76}$$

then λ satisfies the equation

$$\lambda^2-\lambda+[1+\tfrac{1}{2}(\omega_{j+1}^{(0)2}+\omega_{j-1}^{(0)2})]=0. \tag{77}$$

This implies that

$$\lambda=\tfrac{1}{2}\pm\tfrac{1}{2}i\alpha, \tag{78}$$

where α is the positive real number given by

$$\alpha^2=3+2\omega_{j+l}^{(0)2}+2\omega_{j-1}^{(0)2}. \tag{79}$$

We conclude that, in the $(\epsilon_j,\dot{\epsilon}_j)$ plane, we have an unstable focal point, as found in the $\mathfrak{su}(2)$ case in Ref. 7.

Second, we consider the situation in which there is at least one $\omega_j^{(0)}$ which is nonzero. Suppose that $\omega_i^{(0)}=0$, and $\omega_{i+k}^{(0)}=0$, but $\omega_{i+m}^{(0)}\neq 0$ for $m=1,...,k-1$, where we include the case that $i=0$ and $k=N$. Then we have a series of coupled perturbation equations:

$$\begin{aligned}
0&=\ddot{\epsilon}_{i+1}-\dot{\epsilon}_{i+1}-2\epsilon_{i+1}\omega_{i+1}^{(0)2}+\epsilon_{i+2}\omega_{i+2}^{(0)}\omega_{i+1}^{(0)},\\
0&=\ddot{\epsilon}_{i+2}-\dot{\epsilon}_{i+2}-2\epsilon_{i+2}\omega_{i+2}^{(0)2}+\epsilon_{i+3}\omega_{i+3}^{(0)}\omega_{i+2}^{(0)}+\epsilon_{i+1}\omega_{i+1}^{(0)}\omega_{i+2}^{(0)},\\
&\vdots\\
0&=\ddot{\epsilon}_{i+k-1}-\dot{\epsilon}_{i+k-1}-2\epsilon_{i+k-1}\omega_{i+k-1}^{(0)2}+\epsilon_{i+k-2}\omega_{i+k-2}^{(0)}\omega_{i+k-1}^{(0)}.
\end{aligned} \tag{80}$$

Define a vector $\boldsymbol{\epsilon}$ by

$$\boldsymbol{\epsilon}=(\epsilon_{i+1},\ldots,\epsilon_{i+k-1})^T, \tag{81}$$

and consider solutions of the form

$$\boldsymbol{\epsilon}=\boldsymbol{\epsilon}^{\lambda r}\mathbf{q}, \tag{82}$$

where $\mathbf{q}$ is a constant vector. Then $\lambda^2-\lambda$ are eigenvalues of the matrix $\mathscr{M}_{k-1}$ given by (42). As discussed in Sec. IV, the matrix $\mathscr{M}_{k-1}$ has positive real eigenvalues $\mathscr{E}_j$. Then

$$\lambda=\frac{1\pm\sqrt{1+4\mathscr{E}_j}}{2}=j+1,-j \tag{83}$$

will have positive and negative real values, and there is a saddle point. Again this is in direct analogy with the $\mathfrak{su}(2)$ case.[7]

VI. GLOBAL BEHAVIOR OF THE SOLUTIONS

In this section we investigate the behavior of solutions of the regular field equations (14)–(19) as functions of τ. From Sec. IV, we know that, given any starting values $\omega_{j,h}$ for the gauge field functions, then there is a local solution of the field equations in a neighborhood of the regular event horizon at $r=r_h$, $\tau=0$, which is analytic in τ and the initial parameters. Furthermore, from Lemma III.2, as long as $0<\Psi<1$ the solutions remain regular. Therefore, as we integrate out in τ from the event horizon there are only three possibilities.

(1) There is a $\tau_0>0$ such that $\Psi(\tau_0)=0$.
(2) For all $\tau>0$, we have $\Psi(\tau)>0$ and $r(\tau)$ remains bounded as $\tau\to\infty$.
(3) For all $\tau>0$, we have $\Psi(\tau)>0$ and $r(\tau)\to\infty$ as $\tau\to\infty$.

We refer to solutions of the first type as *singular* solutions. Those of type (2) are known as *oscillating* solutions in Ref. 7, and we retain their terminology. It will be shown below (Proposition VI.4), that there is a maximum value $r_{\max}$ of r for oscillating solutions. Finally, we denote by S_∞ solutions of the first type for which $r(\tau_0)<r_{\max}$. Our first task is to show that solutions of type

(3) are precisely the black hole solutions we are seeking. We begin by proving the assumption made at the end of Sec. II, namely, that for solutions of type (3), each ω_j is bounded for all r.

Lemma VI.1: If $\Psi(\tau)>0$ *for all* τ *and* $\lim_{\tau\to\infty} r(\tau)=\infty$ *then all the* ω_j *remain bounded as* $r\to\infty$.

Proof: From the local existence Propositions IV.3 and IV.2 and Lemma III.2, each ω_j is an analytic function of r as long as $\mu>0$. Introduce a new variable $x=r^{-1}$; then each ω_j is an analytic function of x in a neighborhood of $x=0$, except possibly at $x=0$, and can be written as a Laurent series:

$$\omega_j(x)=\sum_{n=-\infty}^{\infty} a_n^j x^n. \tag{84}$$

In order for $\mu>0$ for all r, it must be the case that $2mr^{-1}=2mx<1$, although m itself need not necessarily remain bounded. Hence $2mx$ is bounded in a neighborhood of $x=0$, and, in particular, is analytic at $x=0$. Therefore we can write

$$2mx=\sum_{n=0}^{\infty} b_n x^n. \tag{85}$$

Thus

$$m=\frac{1}{2}\sum_{n=0}^{\infty} b_n x^{n-1}, \quad \frac{dm}{dr}=-x^2\frac{dm}{dx}=-\frac{1}{2}\sum_{n=0}^{\infty} b_n(n-1)x^n, \tag{86}$$

so that dm/dr is analytic in a neighborhood of $x=0$. From the field equations,

$$\frac{dm}{dr}=\mu G+r^2 p_\theta, \tag{87}$$

where

$$G=\sum_{j=1}^{N-1}\left(\frac{d\omega_j}{dr}\right)^2, \quad r^2 p_\theta=\frac{x^2}{4}\sum_{j=1}^{N}[\omega_j^2-\omega_{j-1}^2-N-1+2j]^2. \tag{88}$$

Since both G and $r^2 p_\theta$ are positive, they must each be analytic in a neighborhood of $x=0$. Consider G first, using

$$\frac{d\omega_j}{dr}=-x^2\frac{d\omega_j}{dx}=-\sum_{n=-\infty}^{\infty} a_n^j n x^{n+1}, \tag{89}$$

and since $(d\omega_j/dr)^2$ must be analytic near $x=0$, it must be the case that

$$a_n^j=0, \quad \forall n<-1. \tag{90}$$

Then

$$\omega_j^2=\frac{a_{-1}^{j2}}{x^2}+\sum_{n=-1}^{\infty} c_n^j x^n, \tag{91}$$

for constants c_n^j and hence

$$\omega_j^2-\omega_{j-1}^2-N-1+2j=\frac{1}{x^2}[a_{-1}^{j2}-a_{-1}^{(j-1)2}]+\sum_{n=-1}^{\infty} d_n^j x^n. \tag{92}$$

Since $r^2 p_\theta$ is analytic,

$$a^{j}_{-1} = \pm a^{j-1}_{-1}, \quad \forall j. \tag{93}$$

But $\omega_0 \equiv 0$, so $a^{0}_{-1} = 0$, and thus $a^{j}_{-1} = 0$ for all j. We conclude that ω_j is analytic at $x = 0$, and therefore finite as $x = 0$. Therefore ω_j is bounded for all $r \in [r_h, \infty)$. □

Proposition VI.2: If $\Psi(\tau) > 0$ *for all* τ *and* $\lim_{\tau\to\infty} r(\tau) = \infty$, *then the solution tends to one of the flat space critical points, with the exception of the origin.*

The proof of this proposition closely follows Proposition 14 of Ref. 7, the main thrust of which is the following lemma.

Lemma VI.3: If $\Psi(\tau) > 0$ *for all* τ *and* $\lim_{\tau\to\infty} r(\tau) = \infty$ *then*

$$\lim_{\tau\to\infty} \Psi(\tau) = 1. \tag{94}$$

Proof: There are two cases to consider.

(1) There is at least one ω_j which has zeros for arbitrarily large r.
(2) All ω_j have only a finite number of zeros.

Case 1: The proof in this situation follows exactly that of Proposition 14 of Ref. 7 [Eqs. (74)–(76)] since all the ω_j's are bounded (Lemma VI.1).

Case 2: In this case there is a $T_1 > 0$ such that for all $\tau > T_1$ no ω_j has a zero. However, since the equations governing the ω_j are coupled, the ω_j do not necessarily have to be monotonic, provided Proposition III.4 is satisfied.

Suppose, initially, that there is a $T_2 > T_1$ such that for all $\tau > T_2$ every ω_j is monotonic. In this situation each ω_j has a limit and so $\lim_{\tau\to\infty} U_j = 0$ for all j. Then the proof of Proposition 14 of Ref. 7 carries over directly to show that $m(\tau)$ is bounded and $\lim_{\tau\to\infty} \Psi(\tau) = 1$. The proof proceeds as follows. First, in analogy to Eq. (74) of Ref. 7, integrating the field equations (14)–(19) gives, for each j,

$$|\Psi U_j(\tau_2) - \Psi U_j(\tau_1)| \leqslant \frac{c_j}{r(\tau_1)}, \tag{95}$$

for all $\tau_2 > \tau_1 > T_2$, and some constant c_j, since all the ω_j are bounded. Fix τ_1 for the moment; then for all $\tau_3 > \tau_1$,

$$\int_{\tau_1}^{\tau_3} r \Psi U_j^2 d\tau' \leqslant c_j' \int_{\tau_1}^{\tau_3} \dot{\omega}_j d\tau' \leqslant c_j'', \tag{96}$$

for constants c_j' and c_j'', since ω_j has a limit. Hence

$$m(\tau_3) - m(\tau_1) \leqslant \frac{1}{2} \sum_{j=1}^{N-1} c_j'', \tag{97}$$

which is bounded.

Next, turn to the other extreme situation, where ω_j^2 has minima for arbitrarily large r, for every j. If ω_j^2 has a minimum at $r = r_0$, then by Proposition III.4,

$$\omega_j^2(r_0) \geqslant 1 + \tfrac{1}{2}(\omega_{j+1}^2(r_0) + \omega_{j-1}^2(r_0)). \tag{98}$$

Define N_j to be the greatest lower bound of the set of minimum values of ω_j^2 for $\tau > T_1$, where N_j can be zero (although ω_j^2 cannot). Then we have

$$N_j \geqslant 1 + \tfrac{1}{2}(N_{j+1} + N_{j-1}). \tag{99}$$

These inequalities may be solved to give

$$N_j \geqslant j(N - j), \quad \forall j. \tag{100}$$

However, since $\omega_j^2 \leqslant j(N-j)$ for all j from Theorem III.3, it follows that $\omega_j^2(\tau) = j(N-j)$ for sufficiently large τ and all j. Therefore $U_j(\tau)$ is zero for sufficiently large τ; hence $m(\tau)$ is bounded as $\tau \to \infty$ and $\lim_{\tau\to\infty} \Psi(\tau) = 1$.

The remaining intermediate case is where ω_j and ω_{j+l}, where $l > 1$, are monotonic for all sufficiently large τ, while ω_{j+i}^2, $i = 1,...,l-1$ have minima for arbitrarily large τ. We include the possibility that $j = 0$, to cover the case where there is only one ω_i which has a limit. Define L_1 and L_2 by

$$\omega_j^2 \to L_1, \quad \omega_{j+l}^2 \to L_2, \tag{101}$$

as $\tau \to \infty$. Given $\epsilon > 0$, there is a T_1 such that

$$|\omega_j^2 - L_1| < \epsilon, \quad |\omega_{j+l}^2 - L_2| < \epsilon, \quad \forall\, \tau > T_1. \tag{102}$$

If ω_{j+1}^2 has a minimum at $\tau_0 > T_1$, then

$$\omega_{j+1}^2(\tau_0) \geqslant 1 + \tfrac{1}{2}(\omega_{j+2}^2(\tau_0) + \omega_j^2(\tau_0)) \geqslant 1 + \tfrac{1}{2}(\omega_{j+2}^2(\tau_0) + L_1 - \epsilon). \tag{103}$$

With N_{j+1} as before, then

$$N_{j+1} \geqslant 1 + \tfrac{1}{2} N_{j+2} + K_1, \tag{104}$$

where $K_1 = \frac{1}{2}(L_1 - \epsilon)$. Similarly,

$$N_{j+l-1} \geqslant 1 + \tfrac{1}{2} N_{j+l-2} + K_2, \tag{105}$$

where $K_2 = \frac{1}{2}(L_2 - \epsilon)$. The remaining inequalities for $N_{j+2},\ldots, N_{j+l-2}$ are exactly as before (99). The previous method can be repeated to give

$$N_{j+i} \geqslant i(l-i) + K_1 + K_2, \quad i = i,\ldots,l-1. \tag{106}$$

Now define $\check{N}_i$ to be the lowest upper bound of the set of maximum values of ω_i^2 (which exists since ω_i^2 is bounded). If ω_{j+1}^2 has a maximum at $\tau_0 > T_1$, then

$$\omega_{j+1}^2(\tau_0) \geqslant 1 + \tfrac{1}{2}(\omega_{j+2}^2(\tau_0) + \omega_j^2(\tau_0)) \geqslant 1 + \tfrac{1}{2}(\omega_{j+2}^2 + L_1 + \epsilon). \tag{107}$$

The analysis now proceeds exactly as in the previous paragraph and yields

$$\tilde{N}_{j+i} \leqslant i(l-i) + \tilde{K}_1 + \tilde{K}_2, \quad i = 1,...,l-1, \tag{108}$$

where $\tilde{K}_i = \frac{1}{2}(L_i + \epsilon)$ for $i = 1,2$. In addition, by definition it must be the case that

$$\tilde{N}_{j+i} \geqslant N_{j+i}, \tag{109}$$

which means that

$$i(l-i) + \tfrac{1}{2}(L_1 + L_2) - \epsilon \leqslant N_{j+i} \leqslant \tilde{N}_{j+i} \leqslant i(l-i) + \tfrac{1}{2}(L_1 + L_2) + \epsilon, \tag{110}$$

for all $\epsilon > 0$, whence $\tilde{N}_{j+i} = N_{j+i}$ for all i and

$$\omega_{j+i}^2 = i(l-i) + \tfrac{1}{2}(L_1 + L_2), \tag{111}$$

for sufficiently large τ.

In conclusion, then, we have shown that the intermediate case has some ω_j's which are monotonic for sufficiently large τ, with the remaining ω_j's being constant for sufficiently large τ. Hence, in this case also m is bounded as $\tau \to \infty$ and $\lim_{\tau\to\infty} \Psi(\tau) = 1$. □

Proof of Proposition VI.2: In order to show that the geometry becomes flat as $\tau \to \infty$, it remains to show that $S \to 1$ as $\tau \to \infty$. From the proof of Lemma VI.3, for ω_j having a limit as $\tau \to \infty$, we have

$$rU_i = \dot{\omega}_i \to 0, \tag{112}$$

as $\tau \to \infty$. In case 1 of the proof of Lemma VI.3, the proof of Ref. 7 carries straight over to show that $U_i \to 0$ as $\tau \to \infty$ also in this situation. Now

$$\frac{\dot{S}}{S} = \Psi G = \frac{1}{\Psi} \sum_{i=1}^{N-1} U_i^2 \leqslant \frac{C}{\Psi r^2}, \tag{113}$$

for some constant C and sufficiently large τ. Therefore

$$\frac{S'}{S} = \frac{\dot{S}}{S} \frac{1}{r\Psi} \leqslant \frac{C}{\mu r^3} \leqslant \frac{\tilde{C}}{r^3}, \tag{114}$$

for some constant $\tilde{C}$ and sufficiently large τ, as $\mu \to 1$ as $\tau \to \infty$. This means that S has a finite limit as $\tau \to \infty$.

The field equations only involve S'/S rather than just S itself. Therefore S is defined only up to a multiplicative constant, and without loss of generality we may therefore take the finite limit of S as $\tau \to \infty$ to be 1, so that the space–time is asymptotically flat.

Let $\delta(\tau) = 2\Psi - \kappa - 1$ be a small perturbation; then the equation for ω_j reads as

$$\ddot{\omega}_j + [1 - \omega_j^2 + \tfrac{1}{2}(\omega_{j+1}^2 + \omega_{j-1}^2)]\omega_j = (1 + \delta(\tau))\dot{\omega}_j, \tag{115}$$

as $\tau \to \infty$. Since δ is small. it does not alter the position or nature of the critical points as compared with the exactly flat space case. Lemma VI.3 showed that each $\dot{\omega}_j$, $\ddot{\omega}_j \to 0$ as $\tau \to \infty$. Hence, ω_j must approach one of the critical points whose nature was elucidated in Sec. V. The flat space analysis showed that, if $\omega_j = 0$, then there is an unstable focal point in the $(\omega_j, \dot{\omega}_j)$ plane. Hence our solution cannot approach this value, unless $\omega_j \equiv 0$. The solution has to tend to one of the saddle points along a stable direction, and, in direct analogy with the $\mathfrak{su}(2)$ case,[14] there are no solutions where $\omega_j \to 0$ as $\tau \to \infty$. The solution must therefore be a member of the family found in the local existence theorem IV.2. □

The power of proposition VI.2 lies in that, if we can prove the existence of solutions for which $\Psi > 0$ for all $\tau > 0$ and $r \to \infty$, then these solutions are automatically the regular black holes (and, if $r_h \to 0$, soliton solutions[1]) we seek. We close this section by determining the asymptotic behavior of solutions of type (2).

Proposition VI.4: If $\Psi(\tau) > 0$ for all $\tau > 0$ and $r(\tau)$ remains bounded, then $\Psi \to 0$ and $\kappa \to 1$ as $\tau \to \infty$. In addition, there is at least one j for which $\omega_j \to 0$ as $\tau \to \infty$. and this ω_j has infinitely many zeros.

Proof: Since r is monotonic increasing and bounded, it has a limit r_0. In addition, m is monotonic increasing and bounded since the positivity of μ implies that

$$m < \frac{r}{2} \leqslant \frac{r_0}{2}. \tag{116}$$

As $\tau \to \infty$, $\dot{r} \to 0$ and hence $\Psi \to 0$ from (14). Consider the quantity

$$E = -\frac{r^2}{4}(1 + \Psi^2 - 2\kappa\Psi) = -\frac{r^2}{2}(-\mu G + r^2 p_\theta). \tag{117}$$

Then

$$\dot{E} = 2r^2\mu G\Psi - r^2\kappa\mu G < r^2\mu G(1 - \kappa) < 0, \tag{118}$$

for sufficiently large τ for which $\Psi < 1/2$, and since $\kappa > 1$.[7] Hence

$$E \to -\frac{r_0^2}{4}, \tag{119}$$

as $\tau\to\infty$, since E is monotonically decreasing for sufficiently large τ. Then, following Ref. 7 we have $\kappa\to 1$ as $\tau\to\infty$.

Let $\delta(\tau)=\kappa-2\Psi-1$ be small: then the equation for ω_j reads as

$$\ddot{\omega}_j=\dot{\omega}_j(-1-\delta)-\omega_j[1-\omega_j^2+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)]. \tag{120}$$

Replacing τ by $-\tau$ yields the equation

$$\ddot{\omega}_j=\dot{\omega}_j(1+\delta)-\omega_j[1-\omega_j^2+\tfrac{1}{2}(\omega_{j+1}^2+\omega_{j-1}^2)], \tag{121}$$

which is the equation for flat space solutions, so that ω_j must approach one of the critical points. The fact that $E\neq 0$ means that p_θ cannot tend to zero as $\tau\to\infty$ because μG does vanish in the limit. Hence, at least one of the ω_j's must be zero as $\tau\to\infty$. Since δ is small it alters neither the position nor the characteristics of the critical points, and hence this ω_j will go into the focus at zero. Therefore it oscillates infinitely many times before it hits zero. □

We may determine the value of r_0 as follows. Since κ is finite and we have (119), it follows that $r^2 p_\theta\to 1/2$ as $\tau\to\infty$. Hence

$$r_0^2=\frac{1}{2}\lim_{\tau\to\infty}\sum_{j=1}^{N}(\omega_j^2-\omega_{j-1}^2-N-1+2j)^2. \tag{122}$$

The maximum possible value of r_0^2 is when all $\omega_j\to 0$ as $\tau\to\infty$. In this case,

$$r_{0,\max}^2=\frac{1}{2}\sum_{j=1}^{N}(-N-1+2j)^2=\frac{1}{6}N(N-1)(N+1). \tag{123}$$

The minimum possible value of r_0^2 is when only one $\omega_j\to 0$ as $\tau\to\infty$, when

$$r_0^2=j^2(N-j)^2, \tag{124}$$

which has a minimum when $j=1$ or $N-1$, and

$$r_0^2=(N-1)^2. \tag{125}$$

With this value of r_0, we follow Ref. 7 and denote by S_∞ singular solutions for which μ vanishes outside the event horizon when $r<r_{0,\max}$.

VII. $\mathfrak{su}(3)$ BLACK HOLES

We are now in a position to prove the existence of regular black holes. We begin in this section with $\mathfrak{su}(3)$, which we shall study in some detail, before proceeding to the general case in the next section. The method in general is exactly the same as in this section, but it is hoped that by doing the simplest specific case, where we have just two gauge field degrees of freedom, explicitly, the proof will be more transparent.

Genuine $\mathfrak{su}(3)$ solutions [as opposed to embedded $\mathfrak{su}(2)$ solutions which are discussed below] have been found numerically in Ref. 11. There it was conjectured that, analogous to the $\mathfrak{su}(2)$ case, black hole solutions exist for an infinite, but discrete, set of points in the parameter space given by the values at the event horizon of the functions describing the gauge degrees of freedom.

First, we review briefly the way in which $\mathfrak{su}(2)$ solutions may be embedded in $\mathfrak{su}(3)$. For neutral black holes, numerical solutions are discussed in Ref. 5, where both $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$ embeddings are discussed. The $\mathfrak{so}(3)$ embedding contains scaled $\mathfrak{su}(2)$ black holes as well as genuine $\mathfrak{so}(3)$ solutions possessing two degrees of freedom, which are indexed by the number of nodes of each of the gauge field functions. The method for embedding charged black holes can be found in Ref. 4, where the various possible embeddings are described in detail and numerical solutions presented for the various embeddings in $\mathfrak{su}(5)$. Such solutions possess Coulomb charge arising from one or more of the gauge degrees of freedom. The remaining gauge field functions do not contribute to the charge and have structures similar to the neutral solutions. The black holes are once again indexed by the number of nodes of the gauge functions.

Neutral black holes. Let $\omega_1(r)=\sqrt{2}\omega(r)$ and $\omega_2(r)=\sqrt{2}\omega(r)$. Then the field equations (8)–(11) become

$$r^2\mu\omega''=-(2m-2r^3p_\theta)\omega'-(1-\omega^2)\omega, \tag{126}$$

$$m'=(\mu G+r^2p_\theta), \tag{127}$$

$$\frac{S'}{S}=\frac{2G}{r}, \tag{128}$$

where

$$\mu=1-\frac{2m}{r}, \quad G=4\omega'^2, \quad p_\theta=\frac{2}{r^4}(\omega^2-1)^2. \tag{129}$$

In order to extract the precise form of the $\mathfrak{su}(2)$ equations, define a new independent variable R and a new dependent variable M by

$$R=\tfrac{1}{2}r, \quad M=\tfrac{1}{2}m, \tag{130}$$

with the other field variables remaining unchanged. Then the equations are

$$R^2\mu\frac{d^2\omega}{dR^2}=-\left[2M-\frac{(\omega^2-1)^2}{2R}\right]\frac{d\omega}{dR}-(1-\omega^2)\omega, \tag{131}$$

$$\frac{dM}{dR}=\left[\mu\left(\frac{d\omega}{dR}\right)^2+\frac{1}{2R^2}(\omega^2-1)^2\right], \tag{132}$$

$$\frac{1}{S}\frac{dS}{dR}=\frac{2}{R}\left(\frac{d\omega}{dR}\right)^2. \tag{133}$$

These are now exactly the $\mathfrak{su}(2)$ field equations, and the existence of regular black hole solutions has been proved in Refs. 7 and 9.

Charged black holes. Suppose $\omega_2(r)\equiv 0$ [the case $\omega_1(r)\equiv 0$ is identical]. Then

$$r^2\mu\omega_1''=-(2m-2r^3p_\theta)\omega_1'-(1-\omega_1^2)\omega_1, \tag{134}$$

$$m'=\left[\omega_1'^2+\frac{1}{2r^2}(\omega_1^2-1)^2+\frac{3}{2r^2}\right], \tag{135}$$

$$\frac{S'}{S}=\frac{2\omega_1'^2}{r}. \tag{136}$$

While these equations are *not* the same as those for $\mathfrak{su}(2)$ neutral black holes, the equations for κ, Ψ, U, ω_1, r, and S as functions of τ are identical. Hence the analysis of Ref. 7 carries over immediately to prove the existence of black hole solutions of this form.

Our first existence proposition states the existence of genuine $\mathfrak{su}(3)$ black holes, that is, black holes which do not belong to one of the two families outlined above. The proof is remarkably simple once we have set up our notation suitably.

Proposition VII.1: There exist regular black hole solutions of the $\mathfrak{su}(3)$ *Einstein–Yang–Mills equations which do not correspond to embedded* $\mathfrak{su}(2)$ *charged or neutral black holes.*

The strategy in proving this proposition is very similar to that used in Ref. 7. Generic values of the starting parameters $(\omega_{1,h},\omega_{2,h})$ lead to a singular solution, but we will prove that there must be some values of the starting parameters which do not lead to a singular solution. First, we rule out the possibility of solutions of type (2). From Sec. VI, any solution for which $\mu(\tau)>0$ for all τ and $r(\tau)$ remains bounded as $\tau\to\infty$, has $r\to r_0$ as $\tau\to\infty$, where

$$(N-1)^2 \leqslant r_0^2 \leqslant \tfrac{1}{6}N(N+1)(N-1). \tag{137}$$

For $N=3$, therefore, $r_0=2$. Fixing $r_h>2$ will therefore rule out the possibility of such solutions. This is analogous to the value $r_h>1$ which rules out such solutions in the $\mathfrak{su}(2)$ case. At the end of this section we shall return to the existence proof for $r_h \leqslant 2$.

For every pair of starting values $(\omega_{1,h},\omega_{2,h})$ there are then just two possibilities.

(1) $\mu(\tau)>0$ for all τ and we have a regular black hole solution.
(2) There is a $\tau_0>0$ such that $\mu(\tau_0)=0$, and we have a singular solution.

Define a new variable R by

$$R=\frac{r-r_h}{r}, \tag{138}$$

and R_m by the maximum value of R for each solution, that is, in case one $R_m=1$ (corresponding to $r\to\infty$), and in case two $R_m=R(\tau_0)$. For each solution we define n_i to be the number of zeros of the function ω_i between $\tau=0$ and $\tau=\infty$ in case 1 and $\tau=\tau_0$ in case (2), where n_i can be infinite. Defining new variables N_i by

$$N_i=\frac{n_i}{1+n_i}, \tag{139}$$

we allow the possibility that $N_i=1$.

Each pair of nonzero starting values $(\omega_{1,h},\omega_{2,h})$ can then be mapped to the three quantities (R_m,N_1,N_2) for the corresponding solution, giving rise to a map from $\mathbb{R}^2$ [$(\omega_{1,h},\omega_{2,h})$ space] to $\mathbb{B}^2\times[0,1]$ [(R_m,N_1,N_2) space], where $\mathbb{B}$ is the discrete set $\{0,\frac{1}{2},\frac{2}{3},\frac{3}{4},\ldots,1\}$. Call this map f. Figures 1–3 sketch the nature of $(\omega_{1,h},\omega_{2,h})$ space, $\mathbb{B}\times[0,1]$, and $\mathbb{B}^2$, respectively. Note that this map will not be 1–1 for $N=3$, although it was conjectured in Ref. 7 that for $N=2$ the map is 1–1.

Note that since we know the nature of the solution space when one of the $\omega_{i,h}$ vanishes (since in this case $\omega_i\equiv 0$), we do not need to extend the map f to the coordinate axes in $(\omega_{1,h},\omega_{2,h})$ space. With this notation, we are ready to prove Proposition VII.1, once we have the following lemma.

Lemma VII.2: Suppose that for starting values $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *there is a singular solution with* $\mu(\overline{\tau}_0)=0$ *and the gauge field functions having node structure* $(\overline{n}_1,\overline{n}_2)$. *Then the map* f *is continuous at* $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$.

Proof: From Proposition IV.3, the field variables are continuous in τ and the starting parameters. Thus, all starting values $(\omega_{1,h},\omega_{2,h})$ in a sufficiently small neighborhood of $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ will give rise to a singular solution with $\mu(\tau_0)=0$ where τ_0 is close to $\overline{\tau}_0$, the values of the field variables at τ_0 will be close to those at $\overline{\tau}_0$ in the original solution, and the node structure will be $(\overline{n}_1,\overline{n}_2)$ since the gauge field functions cannot have double zeros (Proposition III.4). In other words, f is continuous at $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$. □

Proof of Proposition VII.1: Consider the open subset of the $(\omega_{1,h},\omega_{2,h})$ plane given by

$$D=\{(\omega_{1,h},\omega_{2,h}):0<\omega_{1,h}<\omega_{2,h}\}. \tag{140}$$

The subset $D'=\{(\omega_{1,h},\omega_{2,h}):0<\omega_{2,h}<\omega_{1,h}\}$ can be treated similarly. The symmetries of the field equations (8)–(11) mean that it is sufficient to consider only the first quadrant of the $(\omega_{1,h},\omega_{2,h})$ plane. From Ref. 7 we know that along the line $\omega_{1,h}=\omega_{2,h}$ there are singular solutions with node structure (n_1,n_1) for all $n_1=1,2,\ldots$. Therefore there are neighborhoods of D corresponding to singular solutions with node structure (n_1,n_1) for all $n_1=1,2,\ldots$. Hence $f(D)$ contains points corresponding to (R_m,N_1,N_1) for all $N_1\in\mathbb{B}$. Therefore $f(D)$ is not a connected set. Therefore f cannot be continuous everywhere on D because D is connected. Hence, we conclude that there exists at least one $(\omega_{1,h},\omega_{2,h})\in D$ corresponding to a black hole solution because f is continuous for all $(\omega_{1,h},\omega_{2,h})$ corresponding to singular solutions. □

This approach allows us to see quite simply how the transversality property conjectured in Ref. 7 arises in the $\mathfrak{su}(2)$ case. Along the lien of values of ω_h, singular solutions having different

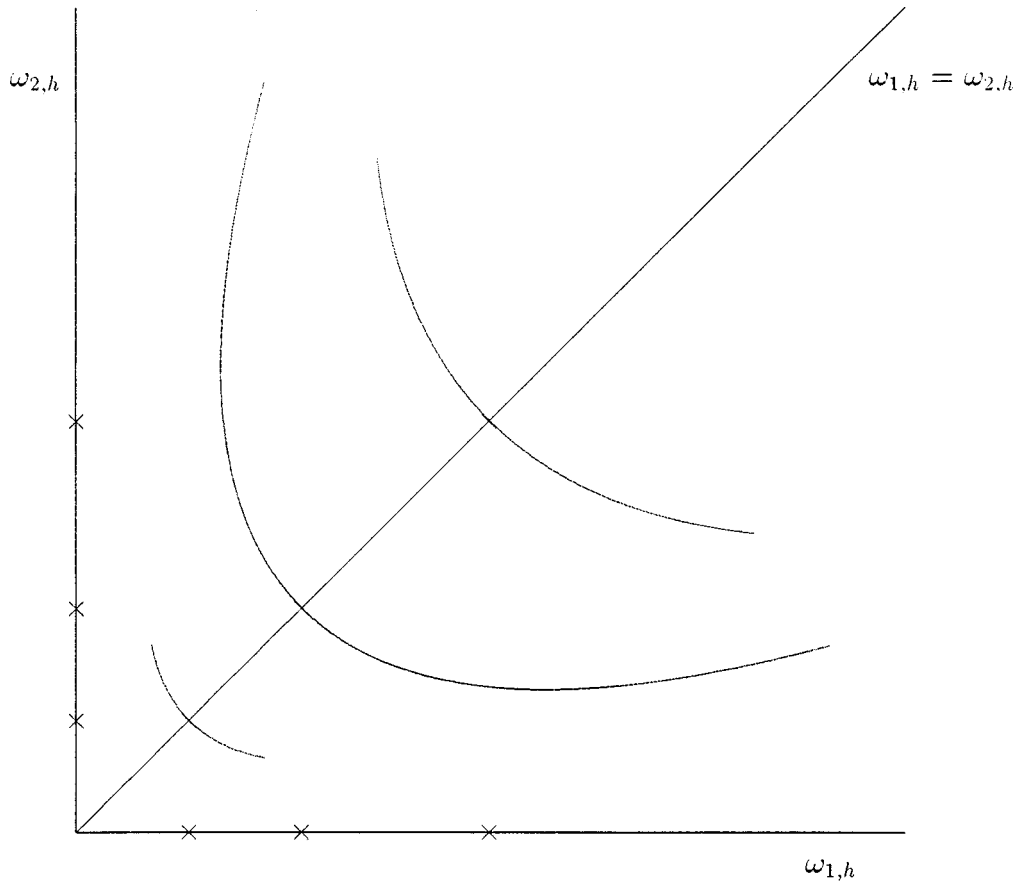

FIG. 1. The $(\omega_{1,h},\omega_{2,h})$ plane, with a schematic representation of the lines representing regular black hole solutions, which are also denoted by crosses.

numbers of nodes must be separated by at least one regular solution. For $\mathfrak{su}(2)$, it is known that values of ω_h sufficiently close to 1 correspond to singular solutions with ω having one node, while sufficiently small ω_h correspond to solutions having as many nodes as we like. In addition, if we have a regular solution with n nodes, then all singular solutions with ω_h sufficiently close have either n or $n+1$ nodes. Starting with the singular solutions with one node, and decreasing ω_h, we first hit a regular solution with one node (there being only the trivial solution with no nodes). Then there may be more singular solutions with one node or two nodes. In the former case there must be another regular solution with one node; in the latter case there will next be a regular solution with two nodes. Either way, there must be a regular solution with two nodes before we can move on to singular solutions with $n>2$ nodes. The process continues for all n.

Having shown that there exist genuine $\mathfrak{su}(3)$ black holes, the main result of this section is the following theorem, which we prove at the moment for the case $r_h>2$, returning to the case $r_h \leqslant 2$ at the end of the section.

Theorem VII.3: *Given* $\overline{n_1}=0,1,\ldots$ *. then there exist regular black hole solutions of the* $\mathfrak{su}(3)$ *Einstein–Yang–Mills equations with* $\omega_1(r)$ *having* $\overline{n_1}$ *nodes and* $\omega_2(r)$ *having* n_2 *nodes, for infinitely many* $n_2 \geqslant \overline{n_1}$.

A similar result holds with the roles of ω_1 and ω_2 reversed. Note that this result is slightly weaker than the corresponding theorem for $\mathfrak{su}(2)$, since we cannot guarantee that every combination of (n_1,n_2) is the node structure of the gauge fields for some black hole solution, only that an infinite number of such combinations does occur for each n_1. At the end of this section we shall give an argument, based on a numerical analysis, that, in fact, black holes exist for all (n_1,n_2), although we are not able to prove this analytically. The proof of Theorem VII.3 will proceed via a series of lemmas.

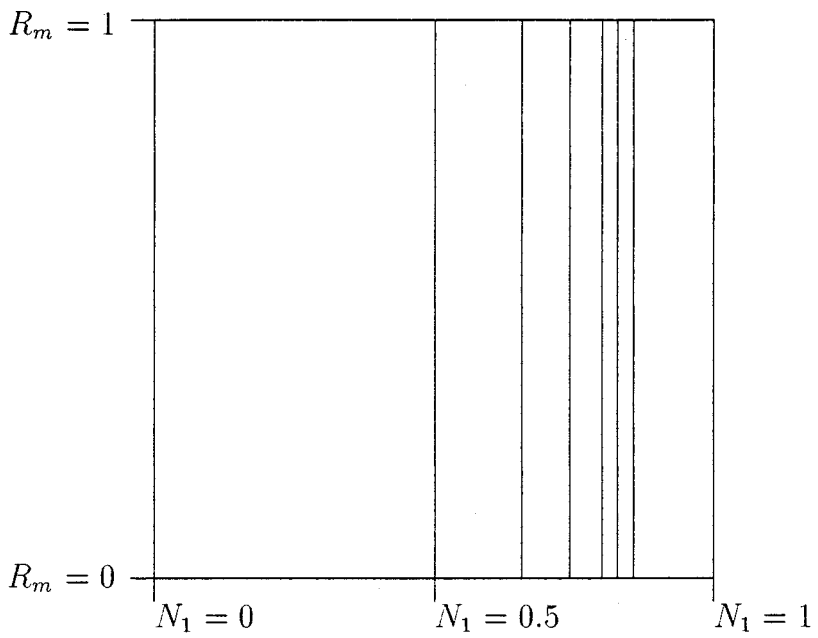

FIG. 2. Schematic representation of the set $\mathbb{B}\times[0,1]$, corresponding to the space of parameters of the $\mathfrak{su}(3)$ solutions in Sec. VII. The horizontal axis represents N_1, the number of nodes of the gauge function ω_1, and the vertical axis R_m, which characterizes the maximum value of the radial coordinate r for each solution.

Lemma VII.4: Suppose that the starting parameters $(0,\overline{\omega}_{2,h})$ *correspond to a charged regular black hole solution in which* $\omega_2(r)$ *has* $\overline{n}_2$ *zeros. Then, given* n_0, *for all sufficiently small* $\omega_{1,h}$ *and* $\omega_{2,h}$ *sufficiently close to* $\overline{\omega}_{2,h}$, *the solutions are regular or singular solutions with* $n_1 \geqslant n_0$ *and* $n_2 \geqslant \overline{n}_2$.

Proof: By continuity, all field variables will remain close to the original charged solution until $r \gg 1$ and the geometry is approximately flat. At this point ω_1 will be very small and ω_2 will be

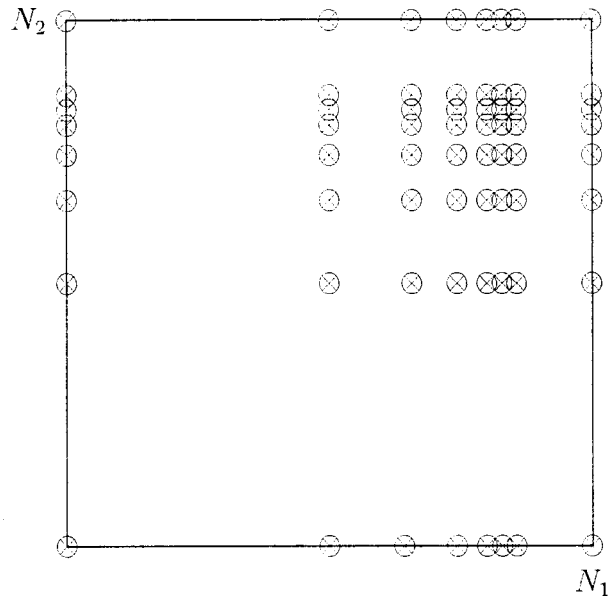

FIG. 3. The projection of the parameter space of solutions onto the set $\mathbb{B}\times\mathbb{B}$. The axes are N_1 and N_2, the number of nodes of the respective gauge functions ω_1 and ω_2. The crosses indicate schematically the points in this set.

close to 1. The equations for $\epsilon_1=\omega_1$ and $\epsilon_2=\omega_2-1$ as functions of τ are, in this regime, to first order,

$$0=\ddot{\epsilon}_1-\dot{\epsilon}_1+\tfrac{3}{2}\epsilon_1, \quad 0=\ddot{\epsilon}_2-\dot{\epsilon}_2-2\epsilon_2. \tag{141}$$

This corresponds to a focus in the $(\omega_1,\dot{\omega}_1)$ plane, and hence with $\omega_{1,h}$ sufficiently small, ϵ_1 will have at least n_0 zeros. □

Lemma VII.5: If $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *leads to a regular black hole solution with* ω_1 *having* $\overline{n}_1$ *nodes and* ω_2 *having* $\overline{n}_2$ *nodes, then all* $(\omega_{1,h},\omega_{2,h})$ *sufficiently close to* $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *lead to regular or singular solutions with* ω_1 *having at least* $\overline{n}_1$ *zeros and* ω_2 *having at least* $\overline{n}_2$ *zeros.*

Proof: Since the solutions are continuous in the starting parameters and r, for $(\omega_{1,h},\omega_{2,h})$ sufficiently close to $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$, the gauge function ω_1 will have $\overline{n}_1$ zeros and ω_2 will have $\overline{n}_2$ zeros for $r<r_1$ for some r_1 and $\omega_i(r_1)$ will be close to their asymptotic values. □

From the proof of Proposition VII.1, it is expected that the $(\omega_{1,h},\omega_{2,h})$ plane will be partitioned into open sets containing singular solutions with (n_1,n_2) nodes by lines of regular solutions. These lines of regular solutions will occur as the horizon radius, r_h, is varied. Numerical investigations[4,5,11] indicate that for fixed r_h regular solutions exist at discrete points in the $(\omega_{1,h},\omega_{2,h})$ plane, the positions of these discrete points varying as r_h is varied. Lemma VII.5 shows that lines corresponding to regular solutions with different numbers of nodes must be disjoint, and, furthermore, that the region corresponding to singular solutions with $(\overline{n}_1,\overline{n}_2)$ nodes will be bounded by lines of regular solutions having (n_1,n_2) nodes, where $n_1\leqslant\overline{n}_1$ and $n_2\leqslant\overline{n}_2$. Lemma VII.5 is somewhat weaker than the corresponding result for $\mathfrak{su}(2)$,[7] since the coupling between the ω's means that we cannot bound from above the number of zeros of the gauge field functions by analytic arguments. Below we shall address this question further by numerical investigations. This leads directly to our existence theorem VII.3 being slightly weaker than for $N=2$.

Lemma VII.6: Suppose $(0,\overline{\omega}_{2,h})$ *corresponds to a charged regular solution, where* ω_2 *has* $\overline{n}_2$ *nodes. Given* n_0, *let* D_0 *be the largest neighborhood of* $(0,\overline{\omega}_{2,h})$ *for which* $(\omega_{1,h},\omega_{2,h})$ *correspond to regular or singular solutions with* $n_1\geqslant n_0$ *and* $n_2\geqslant\overline{n}_2$. *Suppose, in addition, that there is* $(\hat{\omega}_{1,h},\hat{\omega}_{2,h})\in D_0$ *corresponding to a singular solution with* $n_1=n_0$ *and* $n_2=\overline{n}_2$. *Then there is* $(\omega^R_{1,h},\omega^R_{2,h})\in D_0$ *corresponding to a regular solution with* $n_1=n_0$ *and* $n_2=\overline{n}_2$.

Proof: Let $\check{n}_0>n_0$. Then there is a neighborhood $\tilde{D}_0\subset D_0$, $\tilde{D}_0\neq D_0$ for which all solutions are regular or singular solutions with $n_1\geqslant\tilde{n}_0$ and $n_2\geqslant\overline{n}_2$. In $\tilde{D}_0$ there must be at least one singular solution having $n_1\geqslant\tilde{n}_0$ and $n_2\geqslant\overline{n}_2$, by Lemma VII.5, corresponding to starting values $(\tilde{\omega}_{1,h},\tilde{\omega}_{2,h})$. Consider a curve joining $(\tilde{\omega}_{1,h},\tilde{\omega}_{2,h})$ and $(\hat{\omega}_{1,h},\hat{\omega}_{2,h})$ and lying in D_0. Then there must be at least one regular solution along this curve. Let $(\omega^R_{1,h},\omega^R_{2,h})$ be the regular solution closest to $(\hat{\omega}_{1,h},\hat{\omega}_{2,h})$, having $n_1=n^R_1$ and $n_2=n^R_2$. Since $(\omega^R_{1,h},\omega^R_{2,h})\in D_0$, it follows that $n^R_1\geqslant n_0$ and $n^R_2\geqslant\overline{n}_2$. Also, from Lemma VII.5, $n^R_1\leqslant n_0$ and $n^R_2\leqslant\overline{n}_2$, since sufficiently close to $(\omega^R_{1,h},\omega^R_{2,h})$ there are singular solutions having $n_1=n_0$ and $n_2=\overline{n}_2$. Therefore $n^R_1=n_0$ and $n^R_2=\overline{n}_2$. □

Lemma VII.7: Given $\overline{n}_2$ *and* $n_0>\overline{n}_2$, *then there exist regular and singular solutions with* $n_2=\overline{n}_2$ *and* $n_1\geqslant n_0$.

Proof: Consider the point $(0,\overline{\omega}_{2,h})$ which corresponds to the charged regular solution with $n_2=\overline{n}_2$. Then all $(\omega_{1,h},\omega_{2,h})$ sufficiently close to $(0,\overline{\omega}_{2,h})$ are regular or singular solutions with $n_1\geqslant n_0$ and $n_2\geqslant\overline{n}_2$. Let all such $(\omega_{1,h},\omega_{2,h})$ form a neighborhood D_0 of $(0,\overline{\omega}_{2,h})$. From Ref. 7, there exist $(0,\hat{\omega}_{2,h})\in D_0$ corresponding to singular solutions having $\omega_1\equiv 0$ and $n_2=\overline{n}_2$. By continuity, all $(\omega_{1,h},\omega_{2,h})$ sufficiently close to $(0,\hat{\omega}_{2,h})$ correspond to singular solutions having $n_1\geqslant n_0$ and $n_2=\overline{n}_2$. Hence, by Lemma VII.6, there are also in this neighborhood regular black hole solutions having $n_1\geqslant n_0$ and $n_2=\overline{n}_2$. □

Proof of Theorem VII.3: Fix $\overline{n}_2$. Then we have regular solutions with node structure $(\overline{n}_2,\overline{n}_2)$. Now let $n_0=\overline{n}_2+1$, then from Lemma VII.7 there are regular solutions having $n_2=\overline{n}_2$ and $n_1\geqslant n_0$. Let n_0' be the smallest such n_1. Now set $n_0=n_0'+1$ and repeat the process. □

In order to guarantee the existence of black hole solutions having node structure $(\overline{n}_1,\overline{n}_2)$ for every pair of integers $(\overline{n}_1,\overline{n}_2)$, we would require the following lemma in addition to Lemma VII.7.

Lemma VII.8: Suppose $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *corresponds to a regular black hole solution in which* ω_1 *has* $\overline{n_1}$ *nodes and* ω_2 *has* $\overline{n_2}$ *nodes. Then, for* $(\omega_{1,h},\omega_{2,h})$ *sufficiently close to* $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$*, solutions for which* ω_2 *still has* $\overline{n_2}$ *nodes are such that* ω_1 *has either* $\overline{n_1}$ *or* $\overline{n_1}+1$ *nodes.*

The argument we now give for Lemma VII.8 is not an analytic proof because we require a numerical analysis.

Proof: By continuity, we know that for $(\omega_{1,h},\omega_{2,h})$ sufficiently close to $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$, the gauge function ω_1 has $\overline{n_1}$ zeros and ω_2 has $\overline{n_2}$ zeros for $r<r_1$ for some r_1, and $\omega_i(r_1)$ will be close to their asymptotic values. Since we are considering only solutions for which ω_2 has $\overline{n_2}$ nodes, for $r>r_1$, then ω_2 will be of one sign. Therefore we may consider the new dependent variable

$$\psi=\frac{\omega_1}{\omega_2}, \tag{142}$$

which will have the same number of zeros as $\omega_1(r)$ for $r>r_1$. Then the equation satisfied by ψ is

$$r^2\mu\psi''+(2m-2r^3p_\theta)\psi'+2r^2\mu\,\frac{\omega_2'}{\omega_2}\,\psi'+\tfrac{3}{2}\omega_2^2\psi(1-\psi^2)=0. \tag{143}$$

On some interval, $r_1<r<r_2$, the geometry will be very nearly flat, and it will remain flat until one of ω_1, ω_2 or their derivatives approach $O(\sqrt{r})$. Next, consider the flat space equations for ψ and ω_2, which have the autonomous form (where $\tau=\log r$)

$$\begin{aligned} 0&=\ddot{\omega}_2-\dot{\omega}_2+(1-\omega_2^2+\tfrac{1}{2}\psi^2\omega_2^2)\omega_2,\\ 0&=\ddot{\psi}-\dot{\psi}+\frac{2\dot{\omega}_2}{\omega_2}\,\dot{\psi}+\frac{3}{2}\,\omega_2^2\psi(1-\psi^2). \end{aligned} \tag{144}$$

Without loss of generality, we may assume that we are close to the critical point where $\psi=1$ and $\omega_2=\sqrt{2}$. Using the substitution

$$\psi=1+\epsilon_1,\quad \omega_2=\sqrt{2}(1+\epsilon_2), \tag{145}$$

the linearized equations close to this critical point are

$$0=\ddot{\epsilon}_1-\dot{\epsilon}_1-6\epsilon_1,\quad 0=\ddot{\epsilon}_2-\dot{\epsilon}_2+2\epsilon_1-2\epsilon_2. \tag{146}$$

The analysis of Sec. V has already shown that we have a saddle point here; this notation is convenient merely because the perturbation in ψ decouples from that of ω_2 in the linear approximation. From the form of the equation (143), it is clear that $|\psi|$ will be monotonic, increasing for $|\psi|>1$. In other words, for $\epsilon_1>0$ initially, $\epsilon_1>0$ always and ω_1 will have no additional zeros for $r>r_1$. Hence, we need only consider the case where $\epsilon_1<0$ initially. Unfortunately, the nonlinear perturbation equations cannot be integrated analytically, and so we present a numerical argument. For initial ϵ_1, ϵ_2 sufficiently small, the values of ϵ_1 and ϵ_2 will remain close to those on the unstable manifold for all $\tau<\tau_1$ for some τ_1, where τ_1 can be taken to be as large as we like by making the initial perturbations sufficiently small. Numerical integration of the nonlinear equations with the initial point on the unstable manifold and $\epsilon_1<0$ shows that ϵ_2 increases monotonically and ϵ_1 decreases monotonically with just one zero, and cuts through $\epsilon_1=-2$ (corresponding to $\psi=-1$), from whence ψ must be monotonically decreasing (see Figs. 4 and 5). Hence, in this case ω_1 has $\overline{n_1}+1$ zeros. □

Lemma VII.8 allows us to prove that every combination of integers (n_1,n_2) must correspond to a black hole solution. In the proof of Theorem VII.3, we begin with regular solutions with node structure $(\overline{n_2},\overline{n_2})$, and then from Lemma VII.8 there are either regular or singular solutions with node structure $(\overline{n_2}+1,\overline{n_2})$. Using Lemma VII.6, there are then regular solutions having this node structure, and the proof of Theorem VII.3 follows as before.

So far we have proved the existence of infinitely many $\mathfrak{su}(3)$ black holes only for $r_h>2$. The remainder of this section will be spent proving the result for $r_h\leqslant 2$. In this case, the points at

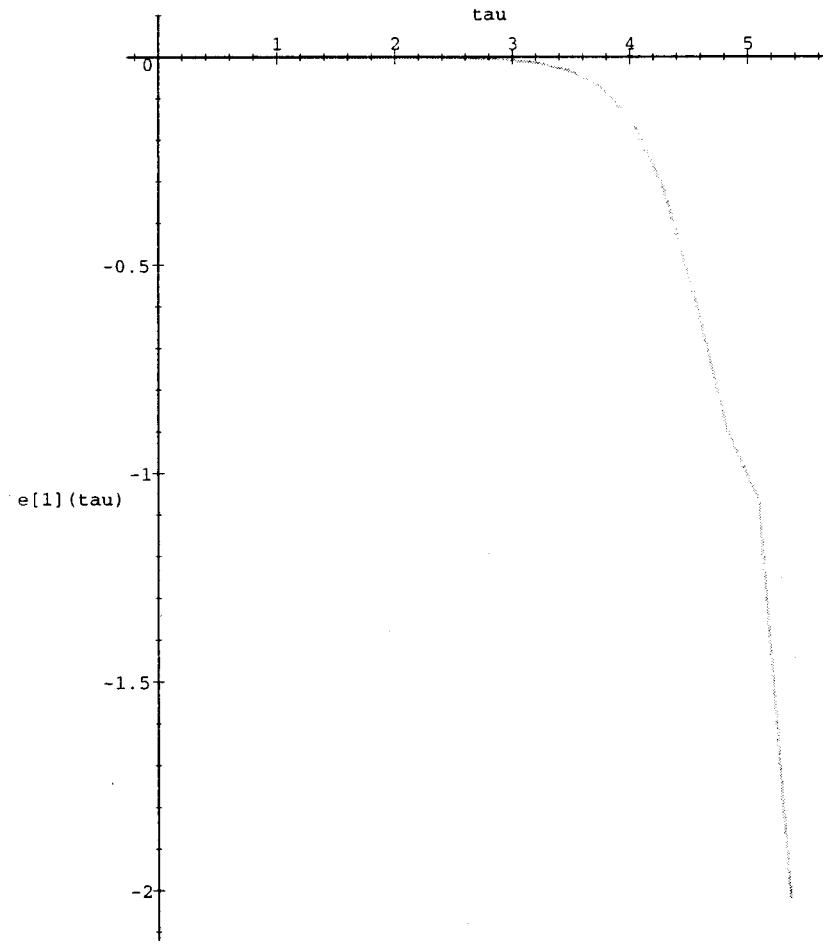

FIG. 4. The su(3) perturbation $\epsilon_1(\tau)$ along the unstable manifold. The value $\tau=0$ corresponds to the critical point. Note that $\psi=1+\epsilon_1$ monotonically decreases through 0 and -1, from which point it will continue to monotonically decrease.

which the map f is not continuous (which must still exist by the argument used in proving Proposition VII.1) do not necessarily correspond to regular black hole solutions: they could also be oscillating solutions. A couple of lemmas concerning oscillating solutions are required before the proofs of Proposition VII.1 and Theorem VII.3 can be extended.

Lemma VII.9: If $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *leads to an oscillating solution with* $\omega_1(\tau)\to 0$ *and* $\omega_2(\tau)\to 1$ *as* $\tau\to\infty$, *and* ω_2 *has* $\overline{n}_2$ *zeros, then all* $(\omega_{1,h},\omega_{2,h})$ *sufficiently close to* $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ *lead to one of the following types of solution.*

(1) *An oscillating solution in which* $\omega_1(\tau)\to 0$ *and* $\omega_2(\tau)\to 1$ *as* $\tau\to\infty$, *with* ω_2 *having at least* $\overline{n}_2$ *zeros.*

(2) *Given* n_0, *either a regular or a singular solution with* $n_1\geqslant n_0$ *and* $n_2\geqslant\overline{n}_2$.

(3) *An* S_∞ *solution.*

Proof: Since in the original solution ω_1 has infinitely many zeros and the solutions are analytic in τ and the starting values, there is some τ_1 such that all solutions with $(\omega_{1,h},\omega_{2,h})$ sufficiently close to $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})$ must have at least n_0 zeros of ω_1 and $\overline{n}_2$ zeros of ω_2 for $\tau<\tau_1$, and all variables will be close to their asymptotic values. □

Lemma VII.10: There exists $(\overline{\omega}_{1,h},\overline{\omega}_{2,h})\in D$ *corresponding to an oscillating solution.*

Proof: From Ref. 7 we know that there are points on the line $\omega_{1,h}=\omega_{2,h}$ corresponding to S_∞ solutions. By continuity, there is a neighborhood D_0 of the line $\omega_{1,h}=\omega_{2,h}$ corresponding to S_∞ solutions. However, Lemma VII.5 still applies in this case, so that the regular solutions on the line $\omega_{1,h}=\omega_{2,h}$ will have singular solutions in D sufficiently close to them for which $r(\tau_0)>2$. Therefore, by Lemma VII.9, there must be points in D which correspond to oscillating solutions. □

The proofs of Proposition VII.1 and Theorem VII.3 are now exactly the same as for $r_h>2$.

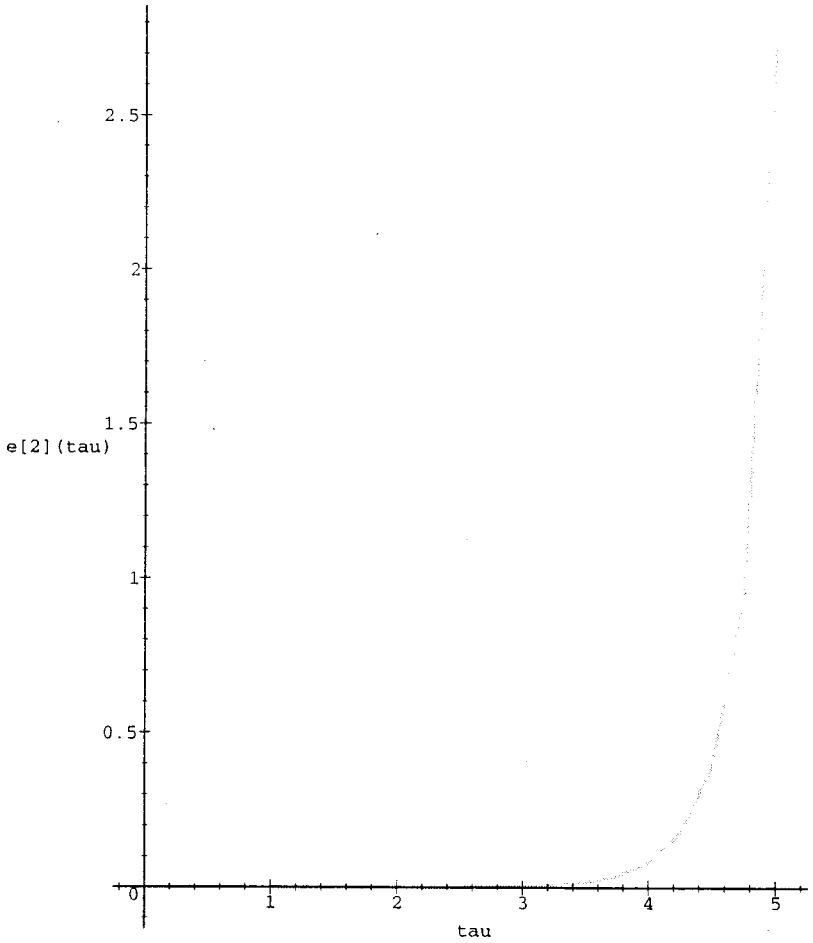

FIG. 5. The $\mathfrak{su}(3)$ perturbation ϵ_2 along the unstable manifold. Note that this is a monotonically increasing function, which is in accordance with the gauge function ω_2 having no further zeros.

Since oscillating solutions have an infinite number of zeros for at least one of the gauge field functions, and using Lemma VII.9, regular solutions are still needed to separate singular solutions having different node structures.

VIII. $\mathfrak{su}(N)$ BLACK HOLES

The detailed discussion of the previous section now enables us to proceed quite swiftly to the analogs of Proposition VII.1 and Theorem VII.3 for general N. The method of proof will be by induction, which was the basic idea used in the previous section, where we proved existence for $\mathfrak{su}(3)$ exploiting known results about $\mathfrak{su}(2)$.

The first step is to illustrate how solutions which are from $\mathfrak{su}(n)$, where $n<N$ may be embedded in the $\mathfrak{su}(N)$ framework. First, exactly as in the $N=3$ case, we may embed neutral $\mathfrak{su}(2)$ solutions.[11] We set $\omega_j(r)=\sqrt{j(N-j)}\omega(r)$ and introduce a scaled variable $R=\lambda_N r$, where

$$\lambda_N=(\tfrac{1}{6}N(N+1)(N-1))^{1/2}. \tag{147}$$

Then the field equations are exactly the same as the $\mathfrak{su}(2)$ ones, and the existence of solutions follows directly.[11] Second, charged, effectively $\mathfrak{su}(N-1)$ solutions can be generated by setting $\omega_1\equiv 0$ (or $\omega_{N-1}\equiv 0$).[4] Although the equation for m' in this situation is not the same as for neutral $\mathfrak{su}(N-1)$ black holes, due to the charge, the regular equations discussed in Sec. II are unchanged, so the proof of existence of neutral $\mathfrak{su}(N-1)$ black holes extends naturally to this case.

For $N>3$, there are additional embeddings which arise from setting at least one of $\omega_2,\ldots,\omega_{N-2}\equiv 0$, so that the gauge field equations decouple into two or more sets of coupled

components, the sets being coupled to each other only through the metric. Solutions of this form have been found numerically in Ref. 4. We illustrate how the existence of black holes of this type may be proved by considering the simplest case, which arises when $N=4$. In this case, we have three nonzero gauge field functions, ω_1, ω_2, ω_3. If we set $\omega_2\equiv 0$, then the field equations take the form

$$r^2\mu\omega_1''=-(2m-2r^3p_\theta)\omega_1'-(1-\omega_1^2)\omega_1, \tag{148}$$

$$r^2\mu\omega_3''=-(2m-2r^3p_\theta)\omega_3'-(1-\omega_3^2)\omega_3, \tag{149}$$

$$m'=(\mu G+r^2p_\theta), \tag{150}$$

$$\frac{S'}{S}=\frac{2G}{r}, \tag{151}$$

where

$$G=\omega_1'^2+\omega_3'^2, \quad p_\theta=\frac{1}{2r^4}([\omega_1^2-1]^2+[\omega_3^2-1]^2+8). \tag{152}$$

These equations look very much like two uncoupled $\mathfrak{su}(2)$ degrees of freedom, however, the two ω's are (albeit weakly) coupled. The fact that we have two gauge field functions here means that we can use the methods of Sec. VII to prove the existence of regular black hole solutions. In fact, all the results of that section carry directly over to this situation on replacing ω_2 there by ω_3 here. There is one exception. The equations (148), (149) are slightly different from those in Sec. VII, and this enables us to strengthen Lemma VII.5.

Lemma VIII.1: If $(\overline{\omega}_{1,h},\overline{\omega}_{3,h})$ *leads to a regular black hole solution with* ω_1 *having* $\overline{n}_1$ *nodes and* ω_3 *having* $\overline{n}_3$ *nodes, then all* $(\omega_{1,h},\omega_{3,h})$ *sufficiently close to* $(\overline{\omega}_{1,h},\overline{\omega}_{3,h})$ *lead to regular or singular solutions with* ω_1 *having either* $\overline{n}_1$ *or* $\overline{n}_1+1$ *zeros and* ω_3 *having either* $\overline{n}_3$ *or* $\overline{n}_3+1$ *zeros.*

Proof: Since the solutions are continuous in the starting parameters and r, for $(\omega_{1,h},\omega_{3,h})$ sufficiently close to $(\overline{\omega}_{1,h},\overline{\omega}_{3,h})$, there is some r_1 such that ω_1 will have $\overline{n}_1$ zeros and ω_3 will have $\overline{n}_3$ zeros for $r<r_1$. There will also be an $r_2>r_1$ such that all field variables are close to their asymptotic values for $r\in[r_1,r_2]$. The flat space field equations (see Sec. V) decouple completely since $\omega_2\equiv 0$, and so the phase plane analysis of Ref. 7 is valid in this case. The phase plane analysis enables us to be more precise about the number of zeros of the perturbed gauge field functions, unlike the more general case where the phase space had more than two coupled dimensions. Therefore the conclusions of Ref. 7 can be applied directly here and each gauge field function can have, at most, one zero for $r>r_2$. □

With this more powerful lemma, the analysis of Sec. VII reaches the conclusion that there are regular black hole solutions of the required form for each r_h and node structure (n_1,n_3).

For general N, similar embeddings of the form of two or more $\mathfrak{su}(n)$ $(n<N)$ type solutions separated by at least one $\omega_j\equiv 0$ are possible, and the existence proof follows the lines outlined above for $\mathfrak{su}(4)$, using the existence of solutions for the various $\mathfrak{su}(n)$. The allowed node structures will be exactly the same as those for the constituent $\mathfrak{su}(n)$ components. We refer the reader to Ref. 4 for further details of the construction of solutions of this form.

The remainder of this section will be spent outlining briefly the proof of the following theorem. By ''types'' in the statement of the theorem we mean the node structures of the gauge field functions.

Theorem VIII.2: *There exist infinitely many types of regular black hole solutions of the* $\mathfrak{su}(N)$ *Einstein–Yang–Mills equations including genuine* $\mathfrak{su}(N)$ *solutions and embedded solutions.*

All that remains to prove this theorem is to show the existence of genuine $\mathfrak{su}(N)$ solutions in which the node structures of the gauge field functions are not all the same. We assume for an inductive hypothesis that the theorem holds for all $n<N$. The situation is slightly complicated by the fact that our parameter space consisting of the $\omega_{j,h}$'s is now $N-1$ dimensional. We consider the $\omega_{1,h}=0$ hyperplane, on which we know the solution space by the inductive hypothesis. The

map $\int$ (see sec. VII) now takes $\mathbb{R}^{N-1}$ to $\mathbb{B}^{N-1}\times[0,1]$. Again, we need only consider the action of f on those parts of $\mathbb{R}^{N-1}$ in which no $\omega_{j,h}$ vanishes. We restrict attention to one of the disjoint spaces so generated without loss of generality. Lemma VII.2 and the proof of the $\mathfrak{su}(N)$ equivalent of proposition VII.1 now carry over to prove the existence of genuine $\mathfrak{su}(N)$ black holes. Lemmas VII.4, VII.6, and VII.7 now hold for a point $(0,\overline{\omega}_{2,h},\ldots,\overline{\omega}_{N-1,h})$ corresponding to a charged solution, and Lemmas VII.5 and VII.9 carry straight over to a point $(\overline{\omega}_{1,h},\overline{\omega}_{2,h},\ldots,\overline{\omega}_{N-1,h})$ which leads to a regular black hole solution or oscillating solution, respectively. These are the ingredients necessary for the simple argument which leads to the proof of Theorem VII.3 and hence Theorem VIII.2 for both $r_h^2>\frac{1}{6}N(N+1)(N-1)$ (in which case no oscillating solutions exist), and $r_h^2\leqslant\frac{1}{6}N(N+1)(N-1)$. Note that, as in the $\mathfrak{su}(3)$ case, we are not able to prove analytically that every sequence of integers $(n_1,n_2,\ldots,n_{N-1})$ corresponds to a black hole solution whose gauge functions have this node structure, although it might reasonably be expected that this is indeed the case.

IX. RESULTS AND CONCLUSIONS

In this paper we have proved the existence of a vast number of hairy black holes in $\mathfrak{su}(N)$ Einstein–Yang–Mills theories. The solutions are described by $N-1$ parameters, corresponding to the number of nodes of the gauge field functions. The result of Ref. 12 tells us that each of these solutions will have a topological instability, similar to the flat-space ''sphaleron,'' as was the case for $\mathfrak{su}(2)$ black holes. This instability does not necessarily diminish the physical importance of these objects, as they may have an important role in processes such as cosmological particle creation.[15]

We now briefly mention another important reason for studying $\mathfrak{su}(N)$ black holes, namely the behavior of the solutions as $N\to\infty$. Black holes in $\mathfrak{su}(\infty)$ Einstein–Yang–Mills theory would possess an infinite amount of hair, requiring an infinite number of parameters to describe their geometry. This might be analogous to the ''W-hair'' found in noncritical string theory,[16] and have drastic consequences for the Hawking radiation, information loss, and quantum decoherence processes associated with black holes. The existence of infinite amounts of hair might also render such objects stable. It is already known that the $\mathfrak{su}(\infty)$ Lie algebra is simply that of the diffeomorphisms of the sphere.[17] This means that the limit as $N\to\infty$ cannot be taken smoothly, in particular, the results of the present paper are only valid when N is finite. We hope to return to this matter in a subsequent publication.

ACKNOWLEDGMENTS

The work of N.E.M. is supported by a P.P.A.R.C. advanced fellowship and that of E.W. is supported by a fellowship at Oriel College, Oxford.

[1] R. Bartnik and J. McKinnon, Phys. Rev. Lett. **61**, 141 (1988).
[2] P. Bizon, Phys. Rev. Lett. **64**, 2844 (1990).
[3] B. R. Greene, S. D. Mathur, and C. M. O'Neill, Phys. Rev. D **47**, 2242 (1993); P. Kanti, N. E. Mavromatos, J. Rizos, K. Tamvakis, and E. Winstanley, *ibid.* **54**, 5049 (1996); T. Torii, K-I. Maeda, and T. Tachizawa, *ibid.* **52**, 4272 (1995); S. Droz, M. Heusler, and N. Straumann, Phys. Lett. B **268**, 371 (1991); G. Lavrelashvili and D. Maison, Nucl. Phys. B **410**, 407 (1993).
[4] B. Kleihaus, J. Kunz, and A. Sood, preprint hep-th/9705179.
[5] B. Kleihaus, J. Kunz, and A. Sood, Phys. Lett. B **354**, 240 (1995).
[6] N. E. Mavromatos and E. Winstanley, Phys. Rev. D **53**, 3190 (1996).
[7] P. Breitenlohner, P. Forgács, and D. Maison, Commun. Math. Phys. **163**, 141 (1994).
[8] J. A. Smoller and A. G. Wasserman, J. Math. Phys. **36**, 4301 (1995).
[9] J. Smoller, A. Wasserman, and S.-T. Yau, Commun. Math. Phys. **154**, 377 (1993).
[10] H. P. Künzle, Class. Quantum Grav. **8**, 2283 (1991).
[11] H. P. Künzle, Commun. Math. Phys. **162**, 371 (1994).
[12] O. Brodbeck and N. Straumann, Zürich University preprint, ZU-TH 38/94, gr-qc/9411058.
[13] E. A. Coddington and N. Levinson, *The Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).
[14] A. A. Ershov and D. V. Gal'tsov, Phys. Lett. A **150**, 159 (1990).
[15] G. W. Gibbons and A. R. Steif, Phys. Lett. B **320**, 245 (1994).
[16] See, for example, J. Ellis, N. E. Mavromatos, and D. V. Nanopoulos, ''A non-critical string approach to black holes, time and quantum dynamics,'' Lectures presented at the International School of Sub-Nuclear Physics, Erice, published in the *Proceedings, Sub-Nuclear Series* (World Scientific, Singapore 1994), Vol. 31, p. 1.
[17] E. G. Floratos, J. Iliopoulos, and G. Tiktopoulos, Phys. Lett. B **217**, 285 (1989).

Covariant spin structure

Gennadi A. Sardanashvily
Department of Theoretical Physics, Moscow State University, 117234 Moscow, Russia

(Received 5 December 1997; accepted for publication 29 April 1998)

Dirac fermion fields associated with different tetrad gravitational fields and under general covariant transformations are described by sections of the composite bundle $S\to\Sigma\to X^4$, which is both the Dirac spinor bundle over the tetrad bundle Σ and the natural one over X^4. As a natural bundle, $S\to X^4$ admits general covariant transformations which are those of Dirac spin structures. A different way is to consider a background spin structure. We find gauge transformations which preserve this spin structure, but act on effective tetrad fields as general covariant transformations.
© *1998 American Institute of Physics.* [S0022-2488(98)01109-8]

I. INTRODUCTION

Metric and metric-affine theories of gravity in the absence of fermion fields are formulated on the natural bundles over a world manifold X^4 such that there exist canonical lifts of diffeomorphisms of X^4 onto these bundles. These lifts are general covariant transformations. The invariance of gravitational Lagrangians under general covariant transformations leads to the energy-momentum conservation laws in these gravitation theories.[1–5] A problem arises because of Dirac fermion fields.

Remark: Manifolds throughout are real, finite-dimensional, Hausdorff, second-countable and connected. By a world manifold X^4 is meant a four-dimensional noncompact oriented manifold which is parallelizable. Such a manifold admits a Dirac spin structure which module isomorphisms is unique.[6,7] This property remains true for all spin structures on X^4 which are generated by the twofold universal covering groups.[8]

Recall that a Dirac spin structure on a world manifold X^4 is said to be a pair (P^h,z_h) of a principal spin bundle $P^h\to X^4$ with the structure spin group $L_s=SL(2,\mathbf{C})$ and a principal bundle morphism

$$z_h: P^h\to LX, \tag{1}$$

over X^4 from P^h to the principal bundle $LX\to X^4$ of oriented linear frames in the tangent bundle TX of X^4.[9–11] The structure group of LX is $GL_4=GL^+(4,\mathbf{R})$. Every Dirac spin structure factorizes through the morphism

$$z_h: P^h\to L^hX\subset LX,$$

where L^hX, called a reduced Lorentz structure, is a principal Lorentz sub-bundle of the frame bundle LX whose structure group is the proper Lorentz group $L=SO^0(1,3)$. Note that a reduced Lorentz structure is not preserved under general covariant transformations of the frame bundle LX. From the physical viewpoint, it means that a Dirac spin structure provides spontaneous breaking of world symmetries.[12,13]

By the well-known theorem,[14] there is one-to-one correspondence between the reduced Lorentz sub-bundles L^hX of the frame bundle LX and the global sections of the quotient bundle

$$\Sigma=LX/L\to X^4, \tag{2}$$

called the tetrad bundle. Its elements are oriented frames in TX module Lorentz transformations. The bundle Σ is the twofold covering of the bundle Σ_g of pseudo-Euclidean bilinear forms in TX, whose global sections are pseudo-Riemannian metrics on X^4. Global sections h of $\Sigma\to X^4$ are tetrad fields on X^4. Accordingly, a Dirac spin structure P^h which factorizes through L^hX is said to be associated with the tetrad field h.

 © 1998 American Institute of Physics

Following the standard terminology of gauge theory,[15–18] one may say that tetrad fields, associated with reduced Lorentz structures, are Higgs fields corresponding to spontaneous breaking of world symmetries.[12,13,15]

Dirac fermion fields in the presence of a tetrad field h are described by sections of the P^h-associated spinor bundle

$$S^h=(P^h\times V)/L_s\to X^4, \tag{3}$$

whose typical fiber V carries the spinor representation of the spin group L_s. To describe Dirac fermion fields and, in particular, to be provided with the Dirac operator, the spinor bundle S^h (3) must be represented as a sub-bundle of the bundle of Clifford algebras, that is, as a spinor structure on the cotangent bundle T^*X in the terminology of Lawson and Michelson.[9] The crucial point is that, for different S^h and $S^{h'}$, these representations are not equivalent though the spin structures P^h and $P^{h'}$ are isomorphic (see Sec. IV). Roughly speaking, for different tetrad fields h, the Clifford representations

$$\gamma_h(dx^\lambda)=h^\lambda_a\gamma^a,$$

of co-frames dx^λ by Dirac's matrices are not equivalent.[12,13]

It follows that every Dirac fermion field must be described in a pair (s_h,h) with a certain tetrad field h, and Dirac fermion fields in the presence of different tetrad fields fail to be given by sections of the same spinor bundle. This fact exhibits the physical nature of gravity as a Higgs field.

There are two ways to describe Dirac fermion fields in the presence of different gravitational fields and under general covariant transformations.

(i) Gravitational fields are identified with different tetrad fields on a world manifold X^4, and the totality of fermion-gravitation pairs (s_h,h) is examined. The goal is to construct a bundle over X^4 whose sections exhaust all these pairs.

(ii) A background spin structure (defined, e.g., by the whole fermion matter of the Universe) and the associated tetrad field on a world manifold X^4 are considered, while different gravitational fields lead to different effective tetrad (or metric) fields, which do not change the background spin structure. The key point is to find gauge transformations over diffeomorphisms of X^4 which both keep the background geometry and act on effective tetrad fields as general covariant transformations.

It should be emphasized that, in both variants, the equations of motion are equivalent at least locally, and so are the equations for a gravitational field if a Lagrangian is independent of a background field.

Following the first variant, let us consider the universal twofold covering group $\widetilde{GL}_4$ of the group GL_4 and the corresponding twofold covering bundle $\widetilde{LX}$ of the frame bundle LX.[9,19–21] The bundle $\widetilde{LX}\to X^4$ inherits the general covariant transformations of the frame bundle LX. However, the spinor representation of the group $\widetilde{GL}_4$ is infinite dimensional. Therefore, the $\widetilde{LX}$-associated spinor bundle describes infinite-dimensional ''world'' spinor fields, but not the Dirac ones. The theory of world spinors has been developed.[22]

In contrast with this world spinor model, our purpose here is to describe the totality of familiar Dirac fermion fields on a world manifold, without appealing to the spinor representation of the general linear group.

We use the fact that the frame bundle LX is the principal bundle $LX\to\Sigma$ over the tetrad bundle Σ (2) with the structure Lorentz group L. Since the diagram

$$\begin{array}{ccc} \widetilde{GL}_4 & \longrightarrow & GL_4 \\ \uparrow & & \uparrow \\ L_s & \xrightarrow{z_L} & L \end{array} \tag{4}$$

commutes (see Sec. V), the bundle $\widetilde{LX}$ is an L_s-principal bundle over the same tetrad bundle $\Sigma = \widetilde{LX}/L_s$. Then the commutative diagram

$$\begin{array}{ccc} \widetilde{LX} & \xrightarrow{\tilde{z}} & LX \\ & \searrow\swarrow & \\ & \Sigma & \end{array} \tag{5}$$

provides a Dirac spin structure on the tetrad bundle Σ. We will show that, for any tetrad field h, the restriction $h^*\widetilde{LX}$ of the L_s-principal bundle $\widetilde{LX}\to\Sigma$ to $h(X^4)\subset\Sigma$ is isomorphic to the L_s-principal sub-bundle P^h of the bundle $\widetilde{LX}\to X^4$ such that the diagram

$$\begin{array}{ccc} \widetilde{LX} & \xrightarrow{\tilde{z}} & LX \\ \uparrow & & \uparrow \\ P^h & \xrightarrow{z_h} & L^hX \end{array} \tag{6}$$

commutes. Then general covariant transformations of the bundle $\widetilde{LX}\to X^4$ take the form of automorphisms of the principal spin bundle $\widetilde{LX}\to\Sigma$ over general covariant transformations of the tetrad bundle Σ. These are desired general covariant transformations of Dirac spin structures (1) as the restrictions to $h(X)\subset\Sigma$ of the diagram (5), called the covariant spin structure.

Now let us consider the spinor bundle

$$S=(\widetilde{LX}\times V)/L_s\to\Sigma, \tag{7}$$

associated with the principal spin bundle $\widetilde{LX}\to\Sigma$. Its typical fiber is the Dirac spinor space V. Given a tetrad field h, the restriction h^*S of $S\to\Sigma$ to $h(X)\subset\Sigma$ is a sub-bundle of the composite bundle

$$S\to\Sigma\to X^4. \tag{8}$$

This sub-bundle is isomorphic to the P^h-associated spinor bundle S^h (3) whose sections s_h describe Dirac fermion fields in the presence of the tetrad field h (see Sec. II). It follows that sections of the composite bundle (8) projected onto different tetrad fields $h:X^4\to\Sigma$ exhaust the totality of pairs (s_h,h) of Dirac fermion fields and tetrad fields. The configuration space of this totality is the first order jet manifold J^1S of the composite bundle $S\to X^4$.[5,23] In this model, tetrad fields are dynamic. They are treated as gravitational fields. We will construct the total Dirac operator and the total Lagrangian on the configuration space J^1S whose restrictions to $h(X)\subset\Sigma$, for any tetrad field h, recover the familiar Dirac operator and the familiar Dirac's Lagrangian of fermion fields in the presence of the tetrad field h and a general linear connection K on X^4.

Note that the bundle $S\to X^4$ is not a spinor bundle. It is provided with the structure of the $\widetilde{LX}$-associated bundle with the structure group $\widetilde{GL}_4$ which acts on the typical fiber $(\widetilde{GL}_4\times V)/L_s$ by the induced representation (see Sec. II). Therefore, general covariant transformations of $\widetilde{LX}$ yield the corresponding automorphisms of the bundle $S\to X^4$, which takes the form of automorphisms of the spinor bundle $S\to\Sigma$ over the general covariant transformations of the tetrad bundle $\Sigma\to X^4$. We will construct the canonical lift onto S of vector fields on X^4 which is the generator of these transformations. Then the energy-momentum conservation law can be derived.[5,23]

Following the variant (ii), we consider a background spin structure P^h associated with a background tetrad field h. In this model, gravitational fields are identified with the sections of the LX-associated group bundle $Q\to X$. The canonical morphism $Q\times\Sigma\to\Sigma$ restricted to $h(X)\subset\Sigma$

defines effective tetrad fields.[24] We will construct automorphisms of LX over diffeomorphisms of X which both preserve the background spin structure $P^h \to LX$ and act on effective tetrad fields as general covariant transformations.

II. REDUCED STRUCTURE

The reduced structure language provides the adequate mathematical formulation of gauge models with broken symmetries.[15–18]

Let $\pi_{PX}: P \to X$ be a principal bundle with a structure group G, which acts freely and transitively

$$R_g: p \mapsto pg, \quad p \in P, \quad g \in G, \tag{9}$$

on P on the right. Let

$$Y = (P \times V)/G, \tag{10}$$

be a P-associated bundle with a typical fiber V on which the structure group G acts on the left. Recall that the quotient (10) is defined by identification of elements (p,v) and $(pg, g^{-1}v)$ for all $g \in G$. Its elements will be denoted by $(p,v)\cdot G$. To say more exactly, the bundle Y (10) is canonically associated with P. In particular, every automorphism Φ of a principal bundle P [which, by definition, is equivariant $R_g \circ \Phi = \Phi \circ R_g$, $\forall g \in G$, under the canonical action (9)] yields the corresponding automorphism

$$\Phi_Y: (P \times V)/G \to (\Phi(P) \times V)/G,$$

of the P-associated bundle Y (10). Recall that every vertical automorphism Φ of P takes the form $p \to p\phi(p)$ where ϕ is a G-valued equivariant function on P, i.e., $\phi(pg) = g^{-1}\phi(p)g$, $\forall g \in G$.

Let H be a Lie subgroup of G. We have the composite bundle

$$P \to P/H \to X,$$

where $\pi_{\Sigma X}: P/H \to X$ is a P-associated bundle, denoted by Σ, with the typical fiber G/H, and $\pi_{P\Sigma}: P \to P/H$ is a principal bundle with the structure group H. A H-principal sub-bundle P^h of P is called a reduced structure.[25,26] By the well-known theorem,[14] there is one-to-one correspondence between the global sections h of the quotient bundle $\Sigma \to X$ and the H-principal sub-bundles P^h of P. Such a sub-bundle P^h is isomorphic to the pull-back $h^*P = \pi_{P\Sigma}^{-1}(h(X))$ over X of the bundle $P \to \Sigma$ by h. The following assertion takes place.[5,26]

Proposition 1: Every vertical automorphism Φ of the principal bundle $P \to X$ sends a reduced sub-bundle P^h onto a reduced sub-bundle $P^{h'}$ which is isomorphic to P^h as a H-principal bundle. Conversely, let two reduced sub-bundles P^h and $P^{h'}$ of a principal bundle P be isomorphic to each other as H-principal bundles. Then every isomorphism $\Phi: P^h \to P^{h'}$ over X can be extended to a vertical automorphism of P.

If the quotient G/H is homeomorphic to an Euclidean space, all H-principal sub-bundles of P are isomorphic to each other as H-principal bundles.[27] This also takes place if P is a trivial bundle.

Given a reduced sub-bundle P^h of a principal bundle P, let

$$Y^h = (P^h \times V)/H, \tag{11}$$

be the P^h-associated bundle with a typical fiber V. If $P^{h'}$ is another reduced sub-bundle of P which is isomorphic to P^h, the bundles Y^h and $Y^{h'}$ are isomorphic, but not canonically isomorphic in general.

Proposition 2: Let P^h be a H-principal sub-bundle of a G-principal bundle P. Let Y^h be the P^h-associated bundle (11) with a typical fiber V. If V carries a representation of the whole group G, the fiber bundle Y^h is canonically isomorphic to the P-associated fiber bundle (10).

Proof: Every element of Y can be represented as $(p,v)\cdot G$, $p \in P^h$. Then the desired isomorphism is

$$Y^h \ni (p,v)\cdot H \mapsto (p,v)\cdot G \in Y.$$

It follows that, given a H-principal sub-bundle P^h of P, any P-associated bundle Y with the structure group G is canonically equipped with a structure of the P^h-associated fiber bundle Y^h with the structure group H. Briefly, we will write

$$Y=(P\times V)/G\cong(P^h\times V)/H=Y^h.$$

However, P^h- and $P^{h'}$-associated bundle structures on Y are not equivalent because, given bundle atlases Ψ^h of P^h and $\Psi^{h'}$ of $P^{h'}$, the union of the associated atlases of Y has necessarily G-valued transition functions between the charts from Ψ^h and $\Psi^{h'}$.

In gauge theory on the principal bundle P, sections h of the quotient bundle Σ are treated as Higgs fields, while sections s_h of the P^h-associated bundle Y^h (11) describe matter fields in the presence of the Higgs fields h. From the physical viewpoint, the structure group G of P is said to be the group of broken symmetries because matter fields carries only a representation of its subgroup H, and a reduced structure $P^h\subset P$ which is not preserved under automorphisms of P.

In general, Y^h is not associated or canonically associated with other H-principal sub-bundles of P. It follows that matter fields can be represented only by pairs with Higgs fields.

To describe the totality of these pairs (s_h,h) for all Higgs fields, let us consider the composite bundle

$$Y \xrightarrow{\pi_{Y\Sigma}} \Sigma \xrightarrow{\pi_{\Sigma X}} X, \tag{12}$$

where $Y\to\Sigma$ is the bundle

$$Y_\Sigma=(P\times V)/H,$$

associated with the H-principal bundle $P\to\Sigma$.[5,28] There is the canonical isomorphism $i_h: Y^h \to h^*Y$ of the P^h-associated bundle Y^h to the sub-bundle of $Y\to X$ which is the restriction

$$h^*Y=(h^*P\times V)/H\cong(P^h\times V)/H=Y^h,$$

of the bundle Y_Σ to $h(X)\subset\Sigma$, i.e.,

$$i_h(Y^h)=\pi_{Y\Sigma}^{-1}(h(X)). \tag{13}$$

Then every global section s_h of Y^h corresponds to the global section $i_h\circ s_h$ of the composite bundle (12). Conversely, every global section s of the composite bundle (12) which projects onto a section $h=\pi_{Y\Sigma}\circ s$ of the bundle $\Sigma\to X$ takes its values into the sub-bundle $i_h(Y^h)\subset Y$ in accordance with the relation (13). Hence, there is one-to-one correspondence between the sections of the bundle Y^h and the sections of the composite bundle (12) which cover h.

Remark: The total space of the composite bundle $Y\to X$ (12) has the structure of the P-associated bundle

$$Y=(P\times(G\times V)/H)/G,$$

where the elements (p,g,v) and $(pab,b^{-1}g,a^{-1}v)$ for all $a\in H$ and $b\in G$ are identified. Its typical fiber is the quotient $(G\times V)/H$ of the product $G\times V$ by identification of the elements (g,v) and $(ag,a^{-1}v)$ for all $a\in H$. The group G act on this typical fiber by the rule

$$G\ni b: (g,v)\cdot H\to(bg,v)\cdot H,$$

which is the induced representation of G by the identic representation of H. In particular, if the typical fiber V of the composite bundle $Y\to X$ admits the action of the group G, these two bundle structures on Y are equivalent.

The feature of the dynamics of field systems on composite bundles consists in the following.[5,28] Let the composite bundle Y (12) be coordinated by $(x^\lambda, \sigma^m, y^i)$, where (x^λ, σ^m) are bundle coordinates on $\Sigma \to X$. Its first order jet manifold is provided with the adapted coordinates $(x^\lambda, \sigma^m, y^i, \sigma^m_\lambda, y^i_\lambda)$. Let

$$A = dx^\lambda \otimes (\partial_\lambda + A^i_\lambda \partial_i) + d\sigma^m \otimes (\partial_m + A^i_m \partial_i), \tag{14}$$

be a principal connection on the bundle $Y \to \Sigma$. This connection defines the splitting

$$VY = VY_\Sigma \underset{Y}{\oplus} (Y \underset{\Sigma}{\times} V\Sigma),$$

$$\dot{y}^i \partial_i + \dot{\sigma}^m \partial_m = (\dot{y}^i - A^i_m \dot{\sigma}^m)\partial_i + \dot{\sigma}^m (\partial_m + A^i_m \partial_i),$$

of vertical tangent bundles. Using this splitting, one can construct the first-order differential operator

$$\begin{aligned} &\widetilde{D}: J^1 Y \to T^*X \underset{Y}{\otimes} VY_\Sigma, \\ &\widetilde{D} = dx^\lambda \otimes (y^i_\lambda - A^i_\lambda - A^i_m \sigma^m_\lambda)\partial_i, \end{aligned} \tag{15}$$

on the composite bundle Y. The operator (15) possesses the following important property. Given a global section h of Σ, its restriction

$$\begin{aligned} &\widetilde{D}_h = \widetilde{D} \circ J^1 i_h : J^1 Y^h \to T^*X \otimes VY^h, \\ &\widetilde{D}_h = dx^\lambda \otimes (y^i_\lambda - A^i_\lambda - A^i_m \partial_\lambda h^m)\partial_i, \end{aligned} \tag{16}$$

to Y^h is exactly the familiar covariant differential relative to the principal connection

$$A_h = dx^\lambda \otimes [\partial_\lambda + (A^i_m \partial_\lambda h^m + A^i_\lambda)\partial_i],$$

on the bundle $Y^h \to X$, which is induced by the principal connection (14) on the bundle $Y \to \Sigma$ by the imbedding i_h.

III. LORENTZ STRUCTURE

An example of a reduced structure is a Lorentz reduced structure in gravitation theories which is deduced from the equivalence principle,[15] and accompanies a Dirac fermion matter.[12,13]

Let $\pi_{LX}: LX \to X^4$ be the frame bundle. Given the holonomic frames $\{\partial_\mu\}$ in the tangent bundle TX, every element $\{H_a\}$ of LX takes the form $H_a = H^\mu_a \partial_\mu$, where H^μ_a is a matrix element of the group GL_4. The frame bundle LX is provided with the bundle coordinates (x^λ, H^μ_a). In these coordinates, the canonical action of the structure group GL_4 on LX reads

$$R_g : H^\mu_a \mapsto H^\mu_b g^b_a, \quad g \in GL_4.$$

As is well-known, the frame bundle LX is equipped with the canonical $\mathbf{R}^4$-valued one-form, which is given by the coordinate expression

$$\theta_{LX} = H^a_\mu dx^\mu \otimes t_a, \tag{17}$$

where $\{t_a\}$ is a fixed basis for $\mathbf{R}^4$ and H^b_μ is the inverse matrix of H^μ_a.

Since a world manifold is parallelizable, the structure group GL_4 of the frame bundle LX is reducible to the Lorentz group L. The corresponding L-principal sub-bundle L^hX is a reduced Lorentz structure. Since LX is trivial, any two Lorentz sub-bundles L^hX and $L^{h'}X$ are isomorphic to each other. By virtue of Proposition 1, there exists a vertical bundle automorphism Φ of LX which sends L^hX onto $L^{h'}X$.

As was mentioned above, there is one-to-one correspondence between the Lorentz sub-bundles L^hX of LX and the global sections h of the tetrad bundle Σ (2) with the typical fiber GL_4/L.

Every tetrad field h defines an associated Lorentz atlas $\Psi^h=\{U_\zeta, z^h_\zeta\}$ of the frame bundle LX where the corresponding local sections z^h_ζ of LX take their values into the Lorentz sub-bundle L^hX. Given a Lorentz atlas Ψ^h, the pull-back

$$z^{h*}_\zeta \theta_{LX}=h^a\otimes t_a=h^a_\lambda dx^\lambda\otimes t_a, \tag{18}$$

of the canonical form θ_{LX} (17) by a local section z^h_ζ is said to be a (local) tetrad form. It determines the tetrad coframes $\{h^a\}$ in the cotangent bundle $T^*X\to X^4$. Their coefficients h^a_μ and the inverse matrix elements $h^\mu_a=H^\mu_a\circ z^h_\zeta$ are called tetrad functions.

Given a Lorentz sub-bundle L^hX, let us consider the associated bundle of Minkowski spaces

$$M^hX=(L^hX\times M)/L, \tag{19}$$

where M is provided with the Minkowski metric η. By virtue of Proposition 2, this bundle is isomorphic to the cotangent bundle T^*X. However, the Minkowski structures M^hX and $M^{h'}X$ on T^*X for different tetrad fields h and h' are not equivalent.

IV. DIRAC SPIN STRUCTURE

Every bundle of Minkowski spaces M^hX (19) over a world manifold is extended to the bundle of Clifford algebras C^hX with the fibers generated by the fibers of M^hX.[10] This bundle C^hX has the structure group $\mathrm{Aut}(\mathbf{C}_{1,3})$ of inner automorphisms of the Clifford algebra $\mathbf{C}_{1,3}$. In general, C^hX does not contain a spinor sub-bundle because a spinor subspace V (a minimal left ideal) of $\mathbf{C}_{1,3}$ is not stable under inner automorphisms of $\mathbf{C}_{1,3}$. As was shown,[29,30] a spinor sub-bundle of C^hX exists if the transition functions of C^hX can be lifted from $\mathrm{Aut}(\mathbf{C}_{1,3})$ to the Clifford group $G_{1,3}$. This agrees with the usual condition of existence of a spin structure which holds for X^4. Such a spinor sub-bundle is the bundle S^h (3) associated with the universal twofold covering

$$z_h: P^h\to L^hX, \quad z_h\circ R_g=R_{z_{L(g)}}, \quad \forall g\in L_s,$$

of L^hX. This is the h-associated Dirac spin structure on a world manifold.

There exists the bundle morphism

$$\gamma_h: T^*X\otimes S^h=(P^h\times(M\otimes V))/L_s\to(P^h\times\gamma(M\otimes V))/L_s=S^h, \tag{20}$$

whereby γ is meant the left action of $M\subset\mathbf{C}_{1,3}$ on $V\subset\mathbf{C}_{1,3}$.[12,13] One can think of (20) as being the representation of covectors to X^4 by the Dirac γ-matrices on elements of the spinor bundle S^h. Relative to an atlas $\{z_\zeta\}$ of P^h and to the associated Lorentz atlas $\{z_h\circ z_\zeta\}$ of LX, the representation (20) reads

$$y^A(\gamma_h(h^a(x)\otimes v))=\gamma^{aA}{}_B y^B(v), \quad v\in S^h_x,$$

where y^A are the corresponding bundle coordinates of S^h, and h^a are the tetrad coframes (18). For brevity, we will write

$$\hat h^a=\gamma_h(h^a)=\gamma^a, \quad \hat d x^\lambda=\gamma_h(dx^\lambda)=h^\lambda_a(x)\gamma^a.$$

Let A_h be a principal connection on S^h, and let

$$D: J^1S^h\to T^*X\underset{S^h}{\otimes}S^h,$$

$$D=(y^A_\lambda-A^{ab}{}_\lambda L_{ab}{}^A{}_B y^B)dx^\lambda\otimes\partial_A$$

where

$$L_{ab}=\tfrac{1}{4}[\gamma_a,\gamma_b], \tag{21}$$

be the corresponding covariant differential. The first-order differential Dirac operator is defined on S^h by the composition

$$\begin{aligned}&\Delta_h=\gamma_h\circ D: J^1S^h\to T^*X\otimes S^h\to S^h,\\ &y^A\circ\Delta_h=h^\lambda_a\gamma^{aA}{}_B(y^B_\lambda-\tfrac{1}{2}A^{ab}{}_\lambda L_{ab}{}^A{}_B y^B).\end{aligned} \tag{22}$$

The h-associated spinor bundle S^h is equipped with the fiber spinor metric

$$a_h(v,v')=\tfrac{1}{2}(v^+\gamma^0 v'+v'^+\gamma^0 v),\quad v,v'\in S^h.$$

Using this metric and the Dirac operator (22), one can define Dirac's Lagrangian

$$\begin{aligned}L_h=\Bigg\{\frac{i}{2}h^\lambda_q\Bigg[y^+_A(\gamma^0\gamma^q)^A{}_B\left(y^B_\lambda-\frac{1}{2}A_\lambda{}^{ab}L_{ab}{}^B{}_C y^C\right)\\ -\left(y^+_{\lambda A}-\frac{1}{2}A_\lambda{}^{ab}y^+_C L^+_{ab}\right)(\gamma^0\gamma^q)^A{}_B y^B\Bigg]-m y^+_A(\gamma^0)^A{}_B y^B\Bigg\}\det(h^a_\mu),\end{aligned} \tag{23}$$

on J^1S^h which describes Dirac fermion fields in the presence of a tetrad field h and a principal connection A_h on S_h.

We consider the general case of a principal connection A_h on S_h generated by a general linear connection on a world manifold as follows. Let ω_K be a connection form on LX of a general linear connection

$$K=dx^\lambda\otimes\left(\partial_\lambda+K_\lambda{}^\mu{}_\nu\dot{x}^\nu\frac{\partial}{\partial\dot{x}^\mu}\right), \tag{24}$$

on X^4, called a world connection. By virtue of the well-known theorem,[14] the pull-back $z^*_h\omega_L$ over P^h of the Lorentz part ω_L of ω_K is the connection form of the spin connection

$$K_h=dx^\lambda\otimes[\partial_\lambda+\tfrac{1}{4}(\eta^{kb}h^a_\mu-\eta^{ka}h^b_\mu)(\partial_\lambda h^\mu_k-h^\nu_k K_\lambda{}^\mu{}_\nu)L_{ab}{}^A{}_B y^B\partial_A], \tag{25}$$

on S^h, where L_{ab} are the generators (21).[23,31–33]

Remark: There is one-to-one correspondence between the world connections and the sections of the quotient bundle

$$C_K=J^1LX/GL_4, \tag{26}$$

whereby J^1LX implies the first-order jet manifold of the frame bundle $LX\to X^4$. With respect to the holonomic frames in TX, the bundle C_K is coordinated by $(x^\lambda,k_\lambda{}^\nu{}_\alpha)$ so that, for any section K of $C_K\to X^4$,

$$k_\lambda{}^\nu{}_\alpha\circ K=K_\lambda{}^\nu{}_\alpha,$$

are the coefficients of the world connection K (24).

Motivated by the connection (25), one can obtain the canonical lift

$$\widetilde{\tau}=\tau^\lambda\partial_\lambda+\tfrac{1}{4}(\eta^{kb}h^a_\mu-\eta^{ka}h^b_\mu)(\tau^\lambda\partial_\lambda h^\mu_k-h^\nu_k\partial_\nu\tau^\mu)L_{ab}{}^A{}_B y^B\partial_A, \tag{27}$$

of vector fields τ on X onto the spinor bundle S^h.[5,23] The lift (27) is brought into the form

$$\widetilde{\tau}=\tau_{\{\}}-\tfrac{1}{4}(\eta^{kb}h^a_\mu-\eta^{ka}h^b_\mu)h^\nu_k\nabla_\nu\tau^\mu L_{ab}{}^A{}_B y^B\partial_A,$$

where $\tau_{\{\}}$ is the horizontal lift of τ by means of the spin Levi-Civita connection for the tetrad field h, and $\nabla_\nu\tau^\mu$ are the covariant derivatives of τ relative to the Levi-Civita connection.[34,35]

The canonical lift (27) fails to be a generator of general covariant transformations because it does not involve transformations of tetrad fields. To define general covariant transformations, one must consider spinor structures associated with different tetrad fields. The difficulty arises because, though the principal spinor bundles P^h and $P^{h'}$ are isomorphic, the associated structures of the bundles of Minkowski spaces M^hX and $M^{h'}X$ (19) on the cotangent bundle T^*X are not equivalent, and so are the representations γ_h and $\gamma_{h'}$ (20).[12,13] Indeed, let

$$t^* = t_\mu dx^\mu = t_a h^a = t'_a h'^a,$$

be an element of T^*X. Its representations γ_h and $\gamma_{h'}$ (20) read

$$\gamma_h(t^*) = t_a \gamma^a = t_\mu h^\mu_a \gamma^a, \quad \gamma_{h'}(t^*) = t'_a \gamma^a = t_\mu h'^\mu_a \gamma^a.$$

There is no isomorphism Φ_s of S^h onto $S^{h'}$ which can obey the condition

$$\gamma_{h'}(t^*) = \Phi_s \gamma_h(t^*) \Phi_s^{-1}, \quad \forall t^* \in T^*X.$$

We thus observe the phenomenon of symmetry breaking in gravitation theory which exhibits the physical nature of gravity as a Higgs field.

V. COVARIANT SPIN STRUCTURE

We start from the following two facts.

Remark: The L-principal bundle

$$P_L := GL_4 \to GL_4/L, \tag{28}$$

is trivial. In accordance with the classification theorem,[27] a G-principal bundle over an n-dimensional sphere S^n is trivial if the homotopy group $\pi_{n-1}(G)$ is trivial. The base GL_4/L is homeomorphic to $S^3 \times \mathbf{R}^7$. Let us consider the morphism f_1 of S^3 into GL_4/L, $f_1(p) = (p,0)$, and the pull-back L-principal bundle $f_1^* P_L \to S^3$. Since L is homeomorphic to $RP^3 \times \mathbf{R}^3$ and $\pi_2(L) = 0$, this bundle is trivial. Let f_2 be the projection of GL_4/L onto S^3. Then the pull-back L-principal bundle $f_2^*(f_1^* P_L) \to GL_4/L$ is also trivial. Since the composition $f_1 \circ f_2$ of GL_4/L into GL_4/L is homotopic to the identity morphism of GL_4/L, the bundle $f_2^*(f_1^* P_L) \to GL_4/L$ is equivalent to the bundle P_L.[27] It follows that the bundle (28) is also trivial.

Remark: The diagram (4) commutes. The restriction of the universal covering group $\widetilde{GL}_4 \to GL_4$ to the Lorentz group L is obviously a covering space of L. Let us show that this is the universal covering space. Indeed, any noncontractible cycle in GL_4 belongs to some subgroup $SO(3) \subset GL_4$ and the restriction of the covering bundle $\widetilde{GL}_4 \to GL_4$ to $SO(3)$ is the universal covering of $SO(3)$. Since the L is homotopic to its maximal compact subgroup $SO(3)$, its universal covering space belongs to $\widetilde{GL}_4$.

As a consequence, we have the commutative diagram (6)[36] and the covariant spin structure (5).

The covariant spin structure is unique. Since the bundle $\Sigma \to X^4$ is trivial, the set of nonequivalent spin structures on Σ is in bijective correspondence with the cohomology group $H^1(S^3 \times \mathbf{R}^7 \times X^4; \mathbf{Z}_2)$. Since the cohomology group $H^1(S^3; \mathbf{Z}_2)$ is trivial and a spin structure on S^3 is unique,[37] one can show that nonequivalent spin structures on Σ are in bijective correspondence with those on X^4.

Let us consider the composite spinor bundle S (8) where $S \to \Sigma$ is the spinor bundle (7) associated with the L_s-principal bundle $\widetilde{LX} \to \Sigma$. Given a tetrad field h, there is the canonical isomorphism

$$i_h : S^h = (P^h \times V)/L_s \to (h^* \widetilde{LX} \times V)/L_s,$$

of the h-associated spinor bundle S^h (3) onto the restriction h^*S of the spinor bundle $S \to \Sigma$ to $h(X) \subset \Sigma$. Thence, every global section s_h of the spinor bundle S^h corresponds to the global

section $i_h \circ s_h$ of the composite spinor bundle (3). Conversely, every global section s of the composite spinor bundle (8), which projects onto a tetrad field h, takes its values into the sub-bundle $i_h(S^h) \subset S$.

Let the frame bundle $LX \to X^4$ be provided with a holonomic atlas $\{U_\zeta, T\phi_\zeta\}$, and let the principal bundles $\widetilde{LX} \to \Sigma$ and $LX \to \Sigma$ have the associated atlases $\{U_\epsilon, z^s_\epsilon\}$ and $\{U_\epsilon, z_\epsilon = \widetilde{z} \circ z^s_\epsilon\}$, respectively. With these atlases, the composite spinor bundle S is equipped with the bundle coordinates $(x^\lambda, \sigma^\mu_a, y^A)$, where $(x^\lambda, \sigma^\mu_a)$ are coordinates on Σ such that σ^μ_a are the matrix components of the group element $(T\phi_\zeta \circ z_\epsilon)(\sigma)$, $\sigma \in U_\epsilon$, $\pi_{\Sigma X}(\sigma) \in U_\zeta$. For any tetrad field h, we have $(\sigma^\lambda_a \circ h)(x) = h^\lambda_a(x)$ where $h^\lambda_a(x) = H^\lambda_a \circ z_\epsilon \circ h$ are the tetrad functions with respect to the Lorentz atlas $\{z_\epsilon \circ h\}$ of $L^h X$.

The spinor bundle $S \to \Sigma$ is the sub-bundle of the bundle of Clifford algebras which is generated by the bundle of Minkowski spaces

$$E_M = (LX \times M)/L \to \Sigma,$$

associated with the L-principal bundle $LX \to \Sigma$. Since the bundles LX and P_L (28) are trivial, so is the bundle $E_M \to \Sigma$. Hence, it is isomorphic to the product $\Sigma \times_X T^*X$. Then there exists the representation

$$\gamma_\Sigma : T^*X \otimes_\Sigma S = (\widetilde{LX} \times (M \otimes V))/L_s \to (\widetilde{LX} \times \gamma(M \otimes V))/L_s = S, \tag{29}$$

given by the coordinate expression

$$\hat{d}x^\lambda = \gamma_\Sigma(dx^\lambda) = \sigma^\lambda_a \gamma^a.$$

Restricted to $h(X) \subset \Sigma_T$, this representation recovers the morphism γ_h (20).

Using this representation, one can construct the total Dirac operator on the composite spinor bundle S as follows. Since the bundles $\widetilde{LX} \to \Sigma$ and $\Sigma \to X^4$ are trivial, let us consider a principal connection A (14) on the L_s-principal bundle $\widetilde{LX} \to \Sigma$ given by the local connection form

$$A = (A_\lambda{}^{ab} dx^\lambda + A^{kab}_\mu d\sigma^\mu_k) \otimes L_{ab}, \tag{30}$$

$$\begin{aligned} A_\lambda{}^{ab} &= -\tfrac{1}{2}(\eta^{kb}\sigma^a_\mu - \eta^{ka}\sigma^b_\mu)\sigma^\nu_k K_\lambda{}^\mu{}_\nu, \\ A^{kab}_\mu &= \tfrac{1}{2}(\eta^{kb}\sigma^a_\mu - \eta^{ka}\sigma^b_\mu), \end{aligned} \tag{31}$$

where K is a world connection on X^4. This connection defines the associated spin connection

$$A_S = dx^\lambda \otimes (\partial_\lambda + \tfrac{1}{2} A_\lambda{}^{ab} L_{ab}{}^A{}_B y^B \partial_A) + d\sigma^\mu_k \otimes (\partial^k_\mu + \tfrac{1}{2} A^{kab}_\mu L_{ab}{}^A{}_B y^B \partial_A), \tag{32}$$

on the spinor bundle $S \to \Sigma$. The choice of the connection (30) is motivated by the fact that, given a tetrad field h, the restriction of the spin connection (32) to S^h is exactly the spin connection (25).

The connection (32) yields the first order differential operator $\widetilde{D}$ (15) on the composite spinor bundle $S \to X^4$ which reads

$$\widetilde{D} : J^1 S \to T^*X \otimes_\Sigma S,$$

$$\begin{aligned} \widetilde{D} &= dx^\lambda \otimes [y^A_\lambda - \tfrac{1}{2}(A_\lambda{}^{ab} + A^{kab}_\mu \sigma^\mu_{\lambda k}) L_{ab}{}^A{}_B y^B] \partial_A \\ &= dx^\lambda \otimes [y^A_\lambda - \tfrac{1}{4}(\eta^{kb}\sigma^a_\mu - \eta^{ka}\sigma^b_\mu)(\sigma^\mu_{\lambda k} - \sigma^\nu_k K_\lambda{}^\mu{}_\nu) L_{ab}{}^A{}_B y^B] \partial_A. \end{aligned} \tag{33}$$

The corresponding restriction $\widetilde{D}_h$ (16) of the operator $\widetilde{D}$ (33) to $J^1 S^h \subset J^1 S$ recovers the familiar covariant differential on the h-associated spinor bundle $S^h \to X^4$ relative to the spin connection (27).

Combining (29) and (33), we obtain the first-order differential operator

$$\Delta=\gamma_\Sigma\circ\widetilde{D}: J^1S\to T^*X\underset{\Sigma}{\otimes}S\to S,$$

$$y^B\circ\Delta=\sigma_a^\lambda\gamma^{aB}{}_A[y^A_\lambda-\tfrac{1}{4}(\eta^{kb}\sigma^a_\mu-\eta^{ka}\sigma^b_\mu)(\sigma^\mu_{\lambda k}-\sigma^\nu_k K_\lambda{}^\mu{}_\nu)L_{ab}{}^A{}_B y^B], \tag{34}$$

on the composite spinor bundle $S\to X^4$. One can think of Δ as being the total Dirac operator on S since, for every tetrad field h, the restriction of Δ to $J^1S^h\subset J^1S$ is exactly the Dirac operator Δ_h (22) on the spinor bundle S^h in the presence of the tetrad field h and the spin connection (25).

Thus, we come to the model of metric-affine gravity and Dirac fermion fields. The total configuration space of this model is the jet manifold J^1Y of the bundle product

$$Y=C_K\underset{\Sigma}{\times}S, \tag{35}$$

where C_K is the bundle of world connections (26). It is coordinated by $(x^\mu,\sigma^\mu_a,k_\mu{}^\alpha{}_\beta,y^A)$, and J^1Y is provided with the adapted coordinates

$$(x^\mu,\sigma^\mu_a,k_\mu{}^\alpha{}_\beta,y^A,\sigma^\mu_{\lambda a},k_{\lambda\mu}{}^\alpha{}_\beta,y^A_\lambda).$$

The bundle (35) can be endowed with the spin connection

$$A_Y=dx^\lambda\otimes(\partial_\lambda+\widetilde{A}_\lambda{}^{ab}L_{ab}{}^A{}_B y^B\partial_A)+d\sigma^\mu_k\otimes(\partial^k_\mu+A^{kab}_\mu L_{ab}{}^A{}_B y^B\partial_A), \tag{36}$$

where A^{kab}_μ is given by the expression (31), and

$$\widetilde{A}_\lambda{}^{ab}=-\tfrac{1}{2}(\eta^{kb}\sigma^a_\mu-\eta^{ka}\sigma^b_\mu)\sigma^\nu_k k_\lambda{}^\mu{}_\nu.$$

Using the connection (36), we obtain the first-order differential operator

$$\widetilde{D}_Y: J^1Y\to T^*X\underset{\Sigma}{\otimes}S,$$

$$\widetilde{D}_Y=dx^\lambda\otimes[y^A_\lambda-\tfrac{1}{4}(\eta^{kb}\sigma^a_\mu-\eta^{ka}\sigma^b_\mu)(\sigma^\mu_{\lambda k}-\sigma^\nu_k k_\lambda{}^\mu{}_\nu)L_{ab}{}^A{}_B y^B]\partial_A, \tag{37}$$

and the total Dirac operator

$$\Delta_Y=\gamma_\Sigma\circ\widetilde{D}: J^1Y\to T^*X\underset{\Sigma}{\otimes}S\to S,$$

$$y^B\circ\Delta_Y=\sigma_a^\lambda\gamma^{aB}{}_A[y^A_\lambda-\tfrac{1}{4}(\eta^{kb}\sigma^a_\mu-\eta^{ka}\sigma^b_\mu)(\sigma^\mu_{\lambda k}-\sigma^\nu_k k_\lambda{}^\mu{}_\nu)L_{ab}{}^A{}_B y^B], \tag{38}$$

on the bundle $Y\to X^4$. Given a section $K: X\to C_K$, the restrictions of the spin connection A_Y (36), the operator $\widetilde{D}_Y$ (37) and the Dirac operator Δ_Y (38) to K^*Y are exactly the spin connection (32) and the operators (33) and (34), respectively.

The total Lagrangian on the configuration space J^1Y of metric-affine gravity and fermion fields is the sum

$$L=L_{MA}+L_D. \tag{39}$$

A metric-affine Lagrangian L_{MA} depends on the metric coordinates

$$\sigma_{\mu\nu}=\sigma^a_\mu\sigma^b_\nu\eta_{ab} \tag{40}$$

and the curvature

$$R_{\lambda\mu}{}^\alpha{}_\beta=k_{\lambda\mu}{}^\alpha{}_\beta-k_{\mu\lambda}{}^\alpha{}_\beta+k_\mu{}^\alpha{}_\epsilon k_\lambda{}^\epsilon{}_\beta-k_\lambda{}^\alpha{}_\epsilon k_\mu{}^\epsilon{}_\beta.$$

Dirac's Lagrangian is

$$L_D=\left\{\frac{i}{2}\,\sigma_q^{\lambda}\left[y_A^+(\gamma^0\gamma^q)^A{}_B\left(y_\lambda^B-\frac{1}{4}\,(\eta^{kb}\sigma_\mu^a-\eta^{ka}\sigma_\mu^b)(\sigma_{\lambda k}^{\mu}-\sigma_k^{\nu}k_\lambda{}^\mu{}_\nu)L_{ab}{}^B{}_C y^C\right)\right.\right.$$
$$\left.-\left(y_{\lambda A}^+-\frac{1}{4}\,(\eta^{kb}\sigma_\mu^a-\eta^{ka}\sigma_\mu^b)(\sigma_{\lambda k}^{\mu}-\sigma_k^{\nu}k_\lambda{}^\mu{}_\nu)y_C^+L_{ab}^+{}^C{}_A\right)(\gamma^0\gamma^q)^A{}_B y^B\right]$$
$$\left.-my_A^+(\gamma^0)^A{}_B y^B\right\}\sqrt{|\sigma|},\quad \sigma=\det(\sigma_{\mu\nu}). \tag{41}$$

By construction, it is Hermitian. Note that, in fact, the Lagrangian L_D depends only on the torsion $k_\lambda{}^\mu{}_\nu-k_\nu{}^\mu_\lambda$ of a world connection, while the pseudo-Riemannian part is given by the derivative coordinates $\sigma_{\lambda k}^{\mu}$.

VI. GENERAL COVARIANT TRANSFORMATIONS

The frame bundle $LX\to X^4$ belongs to the category of natural bundles. Every diffeomorphism f of X^4 gives rise canonically to the automorphism

$$\widetilde{f}:(x^\lambda,H^\lambda{}_a)\mapsto(f^\lambda(x),\partial_\mu f^\lambda H^\mu{}_a), \tag{42}$$

of LX and to the corresponding automorphisms (general covariant transformations)

$$\widetilde{f}:T=(LX\times W)/GL_4\to(\widetilde{f}(LX)\times W)/GL_4,$$

of any LX-associated bundle T. In particular, if $T=TX$, the lift $\widetilde{f}=Tf$ is the familiar tangent morphism to f.

The lift (42) yields the canonical horizontal lift $\widetilde{\tau}$ of every vector field τ on X^4 onto LX-associated bundles. For instance, such a lift onto a tensor bundle $(\overset{m}{\otimes}TX)\otimes(\overset{k}{\otimes}T^*X)$ reads

$$\widetilde{\tau}=\tau^\mu\partial_\mu+[\partial_\nu\tau^{\alpha_1}\dot{x}_{\beta_1\cdots\beta_k}^{\nu\alpha_2\cdots\alpha_m}+\cdots-\partial_{\beta_1}\tau^\nu\dot{x}_{\nu\beta_2\cdots\beta_k}^{\alpha_1\cdots\alpha_m}-\cdots]\,\frac{\partial}{\partial\dot{x}_{\beta_1\cdots\beta_k}^{\alpha_1\cdots\alpha_m}}.$$

There also exists the canonical lift

$$\widetilde{\tau}_K=\tau^\mu\partial_\mu+[\partial_\nu\tau^\alpha k_\mu{}^\nu{}_\beta-\partial_\beta\tau^\nu k_\mu{}^\alpha{}_\nu-\partial_\mu\tau^\nu k_\nu{}^\alpha{}_\beta+\partial_{\mu\beta}\tau^\alpha]\,\frac{2}{\partial k_\mu{}^\alpha{}_\beta}, \tag{43}$$

of τ onto the bundle of world connections C_K (26), though this is not a LX-associated bundle.

Since the covariant spin structure is unique, the $\widetilde{GL}_4$-principal bundle $\widetilde{LX}\to X^4$ as well as the frame bundle LX admits the lift of any diffeomorphism f of the base X^4.[19]

This lift is defined by the commutative diagram

$$\begin{array}{ccc}\widetilde{LX} & \xrightarrow{\widetilde{f}} & \widetilde{LX}\\ {\scriptstyle\widetilde{z}}\downarrow & & \downarrow{\scriptstyle\widetilde{z}}\\ LX & \xrightarrow{\widetilde{f}} & LX\\ \downarrow & & \downarrow\\ X & \xrightarrow{f} & X\end{array}$$

The associated morphism of the spinor bundle $S\to\Sigma$ is given by the relation

$$\widetilde{f}_S:(p,v)\cdot L_s\to(\widetilde{f}(p),v)\cdot L_s,\quad p\in\widetilde{LX},\quad v\in S.$$

Because $\widetilde{f}$ is equivarient, this is a fiber-to-fiber automorphism of the bundle $S\to\Sigma$ over the automorphism $\widetilde{f}_\Sigma$ of the tetrad bundle $\Sigma\to X^4$ induced by the diffeomorphism f of X^4. Thus, we have the commutative diagram

$$\begin{array}{ccc} S & \xrightarrow{\widetilde{f}_S} & S \\ \downarrow & & \downarrow \\ \Sigma & \xrightarrow{\widetilde{f}_\Sigma} & \Sigma \\ \downarrow & & \downarrow \\ X^4 & \xrightarrow{f} & X^4 \end{array}$$

of general covariant transformations of the spinor bundle S.

Accordingly, there exists a canonical lift $\widetilde{\tau}_S$ onto S of every vector field τ on X^4. The goal is to find its coordinate expression. Difficulties arise because the tetrad coordinates σ_a^μ of Σ depend on the choice of an atlas of the bundle $LX\to\Sigma$. Therefore, noncanonical vertical components appear in the coordinate expression of $\widetilde{\tau}_S$.

The lift τ_Σ of a vector field τ on X^4 onto the tetrad bundle Σ can be derived from the relation (40) and the canonical lift

$$\widetilde{\tau}=\tau^\lambda\partial_\lambda+(\partial_\nu\tau^\alpha\sigma^{\nu\beta}+\partial_\nu\tau^\beta\sigma^{\nu\alpha})\frac{\partial}{\partial\sigma^{\alpha\beta}}, \tag{44}$$

of τ onto the bundle of pseudo-Riemannian metric Σ_g identified with an open sub-bundle of the tensor bundle $\overset{2}{\vee}TX$, and provided with coordinates $(x^\lambda,\sigma^{\mu\nu})$. We have

$$\tau_\Sigma=\tau^\lambda\partial_\lambda+\partial_\nu\tau^\mu\sigma_c^\nu\frac{\partial}{\partial\sigma_c^\mu}+Q_c^\mu\frac{\partial}{\partial\sigma_c^\mu},$$

where the terms Q_c^μ obey the condition

$$(Q_a^\mu\sigma_b^\nu+Q_a^\nu\sigma_b^\mu)\eta^{ab}=0.$$

Let us consider a horizontal lift of the vector field τ_Σ onto the spinor bundle $S\to\Sigma$ by means of the spin connection (32). It reads

$$\begin{aligned} A_S\widetilde{\tau}_\Sigma=\tau^\lambda\partial_\lambda+\partial_\nu\tau^\mu\sigma_c^\nu\frac{\partial}{\partial\sigma_c^\mu}+\frac{1}{4}(\eta^{kb}\sigma_\mu^a-\eta^{ka}\sigma_\mu^b)\sigma_k^\nu(\partial_\nu\tau^\mu-K_\lambda{}^\mu{}_\nu\tau^\nu)(L_{ab}{}^A{}_By^B\partial_A+L_{ab}^+{}^A{}_By_A^+\partial^B)\\ +Q_c^\mu\frac{\partial}{\partial\sigma_c^\mu}+\frac{1}{4}Q_k^\mu(\eta^{kb}\sigma_\mu^a-\eta^{ka}\sigma_\mu^b)(L_{ab}{}^A{}_By^B\partial_A+L_{ab}^+{}^A{}_By_A^+\partial^B). \end{aligned}$$

This leads us to the desired canonical lift of τ onto S:

$$\widetilde{\tau}_S=\tau^\lambda\partial_\lambda+\partial_\nu\tau^\mu\sigma_c^\nu\frac{\partial}{\partial\sigma_c^\mu}+Q_c^\mu\frac{\partial}{\partial\sigma_c^\mu}+\frac{1}{4}Q_k^\mu(\eta^{kb}\sigma_\mu^a-\eta^{ka}\sigma_\mu^b)(L_{ab}{}^A{}_By^B\partial_A+L_{ab}^+{}^A{}_By_A^+\partial^B),$$

which can be brought into the form

$$\widetilde{\tau}_S=\tau^\lambda\partial_\lambda+\partial_\nu\tau^\mu\sigma_c^\nu\frac{\partial}{\partial\sigma_c^\mu}+\vartheta,$$

$$\vartheta = \tfrac{1}{4} Q_k^\mu (\eta^{kb}\sigma_\mu^a - \eta^{ka}\sigma_\mu^b)\left[-L_{ab}{}^d{}_c \sigma_d^\nu \frac{\partial}{\partial \sigma_c^\nu} + L_{ab}{}^A{}_B y^B \partial_A + L^+_{ab}{}^A{}_B y^+_A \partial^B \right],$$

where $L_{ab}{}^d{}_c$ are generators of the Lorentz group in the Minkowski space. The term ϑ is the above-mentioned noncanonical part of this lift. The corresponding total vector field on the fibred product Y (35) reads

$$\widetilde{\tau}_Y = \widetilde{\tau} + \vartheta,$$

$$\widetilde{\tau} = \tau^\lambda \partial_\lambda + \partial_\nu \tau^\mu \sigma_c^\nu \frac{\partial}{\partial \sigma_c^\mu} + \widetilde{\tau}_K, \tag{45}$$

where $\overline{\tau}_K$ is the canonical lift (43) onto C_K. Its canonical part $\overline{\tau}$ (45) is the generator of a local one-parameter group of general covariant transformations of the bundle Y, whereas the vertical vector field ϑ is the generator of a local one-parameter group of vertical Lorentz automorphisms of the bundle $S \to \Sigma$. By construction, the total Lagrangian L (39) obeys the relations

$$\mathbf{L}_{J^1\vartheta} L_D = 0, \tag{46}$$

$$\mathbf{L}_{J^1\widetilde{\tau}} L_{MA} = 0, \quad \mathbf{L}_{J^1\widetilde{\tau}} L_D = 0. \tag{47}$$

The relation (46) leads to the Nöther conservation law, while the equalities (47) lead to the energy-momentum one.[23]

VII. BACKGROUND SPIN STRUCTURE

Let us consider the case of a background spin structure and the corresponding background tetrad field h.

Given a tetrad field h, any general covariant transformation of the frame bundle LX can be written as the composition $\widetilde{f} = \Phi \circ \widetilde{f}_h$ of its automorphism $\widetilde{f}_h$ over f which preserves L^hX and some vertical automorphism

$$\Phi : p \mapsto p\,\phi(p), \quad p \in LX, \tag{48}$$

where ϕ is a GL_4-valued equivariant function on LX. Since LX is trivial, the automorphism $\widetilde{f}_h$ exists. Indeed, let z^h be a global section of L^hX. Put

$$\widetilde{f}_h : L_xX \ni p = z^h(x)g \mapsto z^h(f(x))g \in L_{f(x)}X.$$

The automorphism $\widetilde{f}_h$ restricted to L^hX induces an automorphism of the principal bundle P^h and the corresponding automorphism $\widetilde{f}_s$ of the spinor bundle S^h, which preserves the representation (20).

Turn now to the vertical automorphism Φ. Let us consider the group bundle $Q \to X$ associated with LX. Its typical fiber is the group GL_4 which acts on itself by the adjoint representation. Let $(x^\lambda, q^\lambda_\mu)$ be coordinates on Q. There exist the left and right canonical actions of Q on any LX-associated bundle T:

$$\rho_{l,r} : Q \underset{X}{\times} T \to T,$$

$$\rho_l : ((p,g)\cdot GL_4, (p,w)\cdot GL_4) \mapsto (p, gw)\cdot GL_4,$$

$$\rho_r : ((p,g)\cdot GL_4, (p,w)\cdot GL_4) \mapsto (p, g^{-1}w)\cdot GL_4.$$

Given a vertical automorphism Φ (48) of LX, the corresponding vertical automorphisms of an associated bundle T and the group bundle Q read

$$\Phi : (p,w)\cdot GL_4 \mapsto (p, \phi(p)w)\cdot GL_4,$$

$$\Phi:(p,g)\cdot GL_4\mapsto(p,\phi(p)g\phi^{-1}(p))\cdot GL_4.$$

For any Φ (48), there exists the fiber-to-fiber morphism

$$\overline{\Phi}:(p,q)\cdot GL_4\mapsto(p,\phi(p)q)\cdot GL_4,$$

of the group bundle Q such that

$$\rho_l(\overline{\Phi}(Q)\times T)=\Phi(\rho_l(Q\times T)), \tag{49}$$

$$\rho_r(\overline{\Phi}(Q)\times\Phi(T))=\rho_r(Q\times T). \tag{50}$$

For instance, if $T=T^*X$, the expressions (49) and (50) take the coordinate form

$$\rho_r:(x^\lambda,q^\lambda_\mu,\dot{x}_\mu)\mapsto(x^\lambda,\dot{x}_\lambda q^\lambda_\mu),$$

$$\overline{\Phi}:(x^\lambda,q^\lambda_\mu)\mapsto(x^\lambda,S^\lambda_\nu q^\nu_\mu),$$

$$\rho_r(x^\lambda,S^\lambda_\nu q^\nu_\mu,\dot{x}_\alpha(S^{-1})^\alpha_\lambda)=(x^\lambda,\dot{x}_\lambda q^\lambda_\mu).$$

Hence, we obtain the representation

$$\gamma_Q:(Q\times T^*X)\underset{Q}{\otimes}(Q\times S^h)\to(Q\times S^h),$$

$$\gamma_Q=\gamma_h\circ\rho_r:(q,t^*)\mapsto\dot{x}_\lambda q^\lambda_\mu\hat{d}x^\mu=\dot{x}_\lambda q^\lambda_\mu h^\mu_a\gamma^a, \tag{51}$$

on elements of the spinor bundle S^h. Let q_e be the canonical global section of the group bundle $Q\to X^4$ whose values are the unit elements of the fibers of Q. Then the representation γ_Q (51) restricted to $q_e(X^4)$ comes to the representation γ_h (20).

Let h be a background tetrad field, while sections $q(x)$ of the group bundle Q are dynamic variables treated as gravitational fields. There is the canonical morphism

$$\rho_l:Q\underset{X}{\times}\Sigma\to\Sigma,$$

$$\rho_l:((p,g)\cdot GL_4,(p,\sigma)\cdot GL_4)\mapsto(p,g\sigma)\cdot GL_4,\quad p\in LX,$$

$$\rho_l:(x^\lambda,q^\lambda_\mu,\sigma^\mu_a)\mapsto(x^\lambda,q^\lambda_\mu\sigma^\mu_a).$$

This morphism restricted to $h(X^4)\subset\Sigma$ takes the form

$$\rho_h:Q\to\Sigma,$$

$$\rho_h:((p,g)\cdot L,(p,\sigma_0)\cdot L)\to(p,g\sigma_0)\cdot L,\quad p\in L^hX,$$

$$\rho_h:(x^\lambda,q^\lambda_\mu)\mapsto(x^\lambda,q^\lambda_\mu h^\mu_a), \tag{52}$$

where σ_0 is the center of the quotient GL_4/L.

Let Σ_h, coordinated by $\overline{\sigma}^\mu_a$, be the quotient of the bundle Q by the kernel $\mathrm{Ker}_h\rho_h$ of the morphism (52) with respect to the section h. This is isomorphic to the tetrad bundle Σ provided with the Lorentz structure of a L^hX-associated bundle. Then the representation (51), which is constant on $\mathrm{Ker}_h\rho_h$, reduces to the representation

$$(\Sigma_h\times T^*X)\underset{\Sigma_h}{\otimes}(\Sigma_h\times S^h)\to(\Sigma_h\times S^h),$$

$$(\tilde{\sigma},t^*)\mapsto\dot{x}_\lambda\tilde{\sigma}^\lambda_a\gamma^a.$$

Thence, one can think of a section $\widetilde{h}\neq h$ of the bundle Σ_h as being an effective tetrad field, and can treat $\widetilde{g}^{\mu\nu}=\widetilde{h}^\mu_a\widetilde{h}^\nu_b\eta_{ab}$ as an effective metric. It should be emphasized that the Greek indices go down and go up by means of the background metric $g^{\mu\nu}=h^\mu_a h^\nu_b\eta_{ab}$.

Given a general covariant transformation $\widetilde{f}=\Phi\circ\widetilde{f}_h$ of the frame bundle LX, let us consider the morphism

$$\widetilde{f}_Q:\quad Q\to\overline{\Phi}\circ\widetilde{f}_h(Q),\quad S^h\to\widetilde{f}_s(S^h),\quad T^*X\to\widetilde{f}(T^*X). \tag{53}$$

This preserves the representation (51), i.e., $\gamma_Q\circ\widetilde{f}_Q=\widetilde{f}_s\circ\gamma_Q$, and yields the general covariant transformation $\widetilde{\sigma}^\lambda_a\mapsto\partial_\mu f^\lambda\widetilde{\sigma}^\mu_a$ of the bundle Σ_h.

Using a spin connection K_h (25) and the representation γ_Q (51), one can construct the Dirac operator

$$\Delta_Q=q^\lambda{}_\mu h^\mu_a\gamma^a D_\lambda.$$

Restricted to $q_e(X^4)$, this operator recovers the Dirac operator Δ_h (22) on the spinor bundle S^h in the presence of the background tetrad field h and the world connection K.

In particular, we come to the metric-affine generalization of Logunov's model[38] whose configuration space is the jet manifold J^1Y of the product

$$Y=Q\underset{X}{\times}C_K\underset{X}{\times}S^h. \tag{54}$$

It is coordinated by $(x^\mu,q^\mu_\nu,k_\alpha{}^\mu{}_\nu,y^A)$. A total Lagrangian on this configuration space is the sum

$$L=L_{MA}+L_D+L_q(q,g), \tag{55}$$

where L_{MA} is a metric-affine Lagrangian expressed into the effective metric $\widetilde{\sigma}^{\mu\nu}=\widetilde{\sigma}^\mu_a\widetilde{\sigma}^\nu_b\eta^{ab}$, L_D is Dirac's Lagrangian (41), expressed into the effective tetrad coordinates $\widetilde{\sigma}^\lambda_a=q^\lambda_\mu h^\mu_a$, while the Lagrangian L_q depends on gravitational fields q and the background metric g.

For the sake of simplicity, let us replace the bundle Q in the product (54) with the bundle Σ_h. There exists the canonical lift $\widetilde{\tau}_{Y'}$ onto $\Sigma_h\times C_K\times S^h$ of every vector field τ on X^4. It coincides with the vector field $\widetilde{\tau}_Y$ (45) where σ^λ_a are replaced with $\widetilde{\sigma}^\lambda_a$. This lift is the generator of gauge transformations induced by morphisms $\widetilde{f}_Q$ (53). Then using the standard procedure,[5,23] obtain the energy-momentum conservation law

$$\partial_\lambda(\tau^\lambda L_q)+(\partial_\alpha\tau^\mu\widetilde{g}^{\alpha\nu}+\partial_\alpha\tau^\nu\widetilde{g}^{\alpha\mu})\,\frac{\partial L_q}{\partial\widetilde{g}^{\mu\nu}}\approx d_\lambda\left(2\tau^\mu\widetilde{g}^{\lambda\alpha}\,\frac{\partial L_q}{\partial\widetilde{g}^{\alpha\mu}}+\tau^\lambda L_q-d_\mu U^{\mu\lambda}\right), \tag{56}$$

where

$$U=2\,\frac{\partial L_{AM}}{\partial R_{\mu\lambda}{}^\alpha{}_\nu}\,(\partial_\nu\tau^\alpha-k_\sigma{}^\alpha{}_\nu\tau^\sigma),$$

is the generalized Komar superpotential.[3,23] If $L_q=0$, we obtain exactly the energy-momentum conservation law of the metric-affine gravitation theory, but with an effective metric. In the case of the Logunov's Lagrangian

$$L_q=\lambda_3 g_{\mu\nu}\widetilde{\sigma}^{\mu\nu}\sqrt{|\widetilde{\sigma}|},$$

the equality (56) comes to the well-known condition

$$\nabla_\alpha(\widetilde{g}^{\alpha\mu}\sqrt{|\widetilde{g}|})\approx 0,$$

where ∇_α are covariant derivatives relative to the Levi-Civita connection of the background metric g.

[1] J. Novotný, in *Geometrical Methods in Physics. Proceeding of the Conference on Differential Geometry and its Applications (Czechoslovakia 1983)*, edited by D. Krupka (University of J. E. Purkyně, Brno, 1984), p. 207.
[2] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, Gen. Relativ. Gravit. **26**, 637 (1994).
[3] G. Giachetta and G. Sardanashvily, Class. Quantum Grav. **13**, L67 (1996).
[4] G. Sardanashvily, J. Math. Phys. **38**, 847 (1997).
[5] G. Giachetta, L. Mangiarotti, and G. Sardanashvily, *New Lagrangian and Hamiltonian Methods in Field Theory* (World Scientific, Singapore, 1997).
[6] R. Geroch, J. Math. Phys. **9**, 1739 (1968).
[7] S. Avis and C. Isham, Commun. Math. Phys. **72**, 103 (1980).
[8] W. Greub and H.-R. Petry, in *Differential Geometric Methods in Mathematical Physics II*, Lecture Notes in Mathematics (Springer-Verlag, Berlin, 1978), Vol. 676, p. 217.
[9] H. Lawson and M.-L. Michelson, *Spin Geometry* (Princeton University Press, Princeton, 1989).
[10] I. Benn and R. Tucker, *An Introduction to Spinors and Geometry with Applications in Physics* (Adam Hilger, Bristol, 1987).
[11] B. Van der Heuvel, J. Math. Phys. **35**, 1668 (1994).
[12] G. Sardanashvily, Int. J. Theor. Phys. **30**, 721 (1991).
[13] G. Sardanashvily and O. Zakharov, *Gauge Gravitation Theory* (World Scientific, Singapore, 1992).
[14] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Vol. 1* (Wiley, New York, 1963).
[15] D. Ivanenko and G. Sardanashvily, Phys. Rep. **94**, 1 (1983).
[16] A. Trautman, *Differential Geometry for Physicists* (Bibliopolis, Naples, 1984).
[17] L. Nikolova and V. Rizov, Rep. Math. Phys. **20**, 287 (1984).
[18] M. Keyl, J. Math. Phys. **32**, 1065 (1991).
[19] L. Dabrowski and R. Percacci, Commun. Math. Phys. **106**, 691 (1986).
[20] R. Percacci, *Geometry on Nonlinear Field Theories* (World Scientific, Singapore, 1986).
[21] S. Switt, J. Math. Phys. **34**, 3825 (1993).
[22] F. Hehl, J. McCrea, E. Mielke, and Y. Ne'eman, Phys. Rep. **258**, 1 (1995).
[23] G. Sardanashvily, Class. Quantum Grav. **14**, 1371 (1997).
[24] G. Sardanashvily, Teor. Mat. Fiz. **112**, N2 (1998) (in Russian).
[25] S. Kobayashi, *Transformation Groups in Differential Geometry* (Springer-Verlag, Berlin, 1972).
[26] F. Gordejuela and J. Masqué, J. Phys. A **28**, 497 (1995).
[27] N. Steenrod, *The Topology of Fibre Bundles* (Princeton University Press, Princeton, 1972).
[28] G. Sardanashvily, J. Math. Phys. **33**, 1546 (1992).
[29] W. Rodrigues and J. Vaz, in *Gravity, Particles and Space-Time*, edited by P. Pronin and G. Sardanashvily (World Scientific, Singapore, 1996), p. 307.
[30] I. Benn and R. Tucker, in *Spinors in Physics and Geometry*, edited by A. Trautman and G. Furlan (World Scientific, Singapore, 1988).
[31] V. Ponomarev and Yu. Obukhov, Gen. Relativ. Gravit. **14**, 309 (1982).
[32] A. Aringazin and A. Mikhailov, Class. Quantum Grav. **8**, 1685 (1991).
[33] R. Tucker and C. Wang, Class. Quantum Grav. **12**, 2587 (1995).
[34] Y. Kosmann, Ann. di Matematica Pura et Applicata **91**, 317 (1972).
[35] F. Fatibene, M. Ferraris, M. Francaviglia, and M. Godina, in *Differential Geometry and its Applications (Proceedings of the 6th International Conference, Brno, August 28-September 1, 1995)*, edited by J. Janyška, I. Kolář, and J. Slovák (Masaryk University Press, Brno, 1996), p. 549.
[36] R. Fulp, J. Lawson, and L. Norris, Int. J. Theor. Phys. **33**, 1011 (1994).
[37] L. Dabrowski and A. Trautman, J. Math. Phys. **27**, 2022 (1986).
[38] A. Logunov, Yu. Loskutov, and M. Mestvirishvili, Int. J. Mod. Phys. A **3**, 2067 (1988).

New varieties of Gowdy space–times

Masayuki Tanimoto[a)]
Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto 606-8502, Japan

(Received 2 February 1998; accepted for publication 1 May 1998)

Gowdy space–times are generalized to admit two commuting spatial local Killing vectors, and some new varieties of them are presented, which are all closely related to Thurston's geometries. Explicit spatial compactifications, as well as the boundary conditions for the metrics, are given in a systematic way. A short comment on an implication of their dynamics toward the initial singularity is made. © *1998 American Institute of Physics.* [S0022-2488(98)00209-6]

I. INTRODUCTION

Spatially compact inhomogeneous space–times admitting two commuting spatial Killing vectors are known as Gowdy space–times,[1] which have recently been paid a great deal of attention as favorable models for studies of asymptotic behavior toward the cosmological initial singularity.[2–6] Since Gowdy space–times provide the simplest inhomogeneous cosmologies, it seems natural to use them in the kind of studies first developed based on Bianchi homogeneous cosmologies[7] by Belinski *et al.*[8] This article's aim is not, however, to elucidate the dynamics of Gowdy space–times, but to point out some new features pertaining to their varieties, which may open new windows for looking at the initial singularity.

As is well known, Bianchi homogeneous cosmologies are classified by the local structures of simply transitive three-dimensional groups, Bianchi I–IX groups. Spatially compact Bianchi cosmologies[9–11] are more diverse than the open ones, since possible compact topologies are in general very diverse for each Bianchi (universal covering) cosmology. Nevertheless, Gowdy space–times having lower symmetry than these homogeneous ones are known to have very limited varieties, i.e., only T^3, $S^2\times S^1$, S^3, and lens spaces as their possible spatial topologies. Why is their diversity so poor? In fact, all the Bianchi cosmologies except types VIII and IX admit two commuting Killing vectors, and, moreover, compactifications are possible except for types IV and $\mathrm{VI}_{a\neq 0}$.[12,9] Since these compact Bianchi models can be thought of as (if they exist) homogeneous limits of compact inhomogeneous models which admit two commuting Killing vectors, it seems that Gowdy models should be more diverse. The solution to this paradox lies in whether the Killing vectors are local or global. That is, the restriction for the possible topologies of Gowdy space–times is a consequence of the definition that the two commuting Killing vectors must be globally defined, while spatially compact Bianchi cosmologies admit in general only local Killing vectors.

If we consider Killing vectors for the simplification of Einstein's equation, which is local in nature, the imposition of the globality of the Killing vectors is evidently unnecessary. We therefore generalize Gowdy space–times in this article to admit two commuting local Killing vectors. (This type of generalization was also considered by Rendall[13] in a different approach from ours.) We will actually find that there exist rich (topologically infinitely many) varieties of Gowdy space–times.

As is well known in the theory of three-dimensional topology, Thurston[14] enumerated eight types of homogeneous three-manifolds, called model geometries, S^3, E^3, H^3, $S^2\times E^1$, $H^2\times E^1$, $\widetilde{SL}(2,\mathbf{R})$, nilgeometry (Nil), and solvegeometry (Sol), and proved in essence that any compact three-dimensional manifold which admits a locally homogeneous Riemannian metric is a compact quotient of one of these eight types of homogeneous manifold (see Sec. 3.8 of Ref. 14). We effectively utilize these model geometries as in the compact locally homogeneous cases.[9,10]

In the next section, we show how we can apply Thurston's theorem to find (generalized)

[a)]JSPS Research Fellow. Electronic mail: tanimoto@yukawa.kyoto-u.ac.jp

 © 1998 American Institute of Physics

Gowdy space–times, and in the subsequent two sections present three new types of Gowdy space–times as examples. The final section is devoted to conclusions, including a comment on the dynamics of the three Gowdy space–times.

II. POSSIBLE TOPOLOGIES

To find possible topologies of Gowdy space–times, we consider a ''homogenization'' of a compact Riemannian three-manifold (M,h_{ab}) which admits two commuting local Killing vectors. Namely, suppose we can smoothly deform the metric h_{ab}, preserving the two local Killing vectors, and can take a locally homogeneous limit $(M,h^{\rm lh}_{ab})$ of the Gowdy space (M,h_{ab}). Since the universal cover $(\tilde{M},\tilde{h}^{\rm lh}_{ab})$ of $(M,h^{\rm lh}_{ab})$ is homogeneous, it must be one of the BKSN types, i.e., Bianchi homogeneous three-manifolds and the Kantowski-Sachs-Nariai (KSN) homogeneous three-manifold.[15,16] On the other hand, since the homogenization $(M,h^{\rm lh}_{ab})$ is compact, $(M,h^{\rm lh}_{ab})$ must be homeomorphic to a compact quotient of one of the eight Thurston model geometries.

A (model) *geometry* is the pair $(\tilde{M},G)$ of a manifold $\tilde{M}$ and a group G of diffeomorphisms on $\tilde{M}$, such that G acts transitively on $\tilde{M}$ with compact isotropy subgroups. Since G acts transitively on $\tilde{M}$, $(\tilde{M},G)$ can be thought of as an equivalence class of the homogeneous manifolds whose isometry group is isomorphic to (or includes a subgroup isomorphic to) G.[10] For convenience we call $(\tilde{M},G)$ a *subgeometry* of $(\tilde{M},G')$ if G is a subgroup of another transitive group G' with compact isotropy subgroups. Moreover, if geometry $(\tilde{M},G)$ is not a subgeometry of any geometry, then we call $(\tilde{M},G)$ a *maximal* geometry, and if $(\tilde{M},G)$ does not have any subgeometry, then we call $(\tilde{M},G)$ a *minimal* geometry. While the eight Thurston geometries are maximal geometries, all the BKSN types except Bianchi IV and $\mathrm{VI}_{a\neq 0}$ can be thought of as the minimal geometries of Thurston geometries.[9,12] This correspondence helps apply Thurston geometries to space–time models.

The group G of the geometry $(\tilde{M},G)$ which corresponds to $(\tilde{M},\tilde{h}^{\rm lh}_{ab})$ should include a subgroup isomorphic to one of $\mathbf{R}^2$, $\mathbf{R}\times\mathrm{U}(1)$, and $\mathrm{U}(1)\times\mathrm{U}(1)$, corresponding to the commuting two Killing vectors. All Thurston geometries satisfy this condition. More precisely, what include $\mathbf{R}^2$ are geometries E^3, H^3, $H^2\times E^1$, $\widetilde{\mathrm{SL}}(2,\mathbf{R})$, Nil, and Sol. Geometry $S^2\times E^1$ includes $\mathbf{R}\times\mathrm{U}(1)$, and S^3 includes $\mathrm{U}(1)\times\mathrm{U}(1)$. Moreover, all the minimal geometries of these Thurston geometries except Bianchi VIII ($\widetilde{\mathrm{SL}}(2,\mathbf{R})$ as the maximal one) and IX (S^3, similarly) also satisfy this condition. Any compact quotient for all the homogeneous three-manifolds except Bianchi VIII and IX can therefore be a homogenization of a Gowdy space. Even for Bianchi VIII and IX, there can be a homogenization when imposing a fourth Killing vector on them.

This observation tells what homogeneous manifold can be the universal cover of a homogeneous limit of a Gowdy space. Once such a homogenization with isometry group G is fixed, the universal cover of the corresponding Gowdy space can be determined as follows. That is, we may find the subgroup H of G such that the actions are smooth along the two commuting Killing vectors, but discrete along the third one. The H-invariant metric is the universal cover metric of the Gowdy space (or space–time), if the fundamental group can be represented in H. In fact, the H-invariance is a necessary condition for the universal cover to admit a (spatially) compact quotient, so we need to check that. As we will see, it is easy to write such an H-invariant metric if using the invariant one-forms for the corresponding Bianchi type. The KSN type has already been discussed in Ref. 1, so we do not consider it in this article.

In the following, we show how these ideas work by presenting three examples, which are all new.

III. GOWDY ON *NIL*×R

As a first example let us consider Nil, which possesses Bianchi II as its minimal geometry. Note that the Bianchi II homogeneous spaces are characterized by the following commutation relations for the three Killing vectors

$$[\xi_1,\xi_2]=-\xi_3,\quad [\xi_2,\xi_3]=0,\quad [\xi_3,\xi_1]=0. \tag{1}$$

In terms of coordinate basis they can be represented by

$$\xi_1=\frac{\partial}{\partial x}+y\,\frac{\partial}{\partial z},\quad \xi_2=\frac{\partial}{\partial y},\quad \xi_3=\frac{\partial}{\partial z}. \tag{2}$$

For future use, we take this opportunity to write the finite actions generated by these ξ_i's:

$$\begin{pmatrix} a\\ b\\ c\end{pmatrix}\begin{pmatrix} x\\ y\\ z\end{pmatrix}=\begin{pmatrix} a+x\\ b+y\\ c+z+ay\end{pmatrix}, \tag{3}$$

where the component vectors

$$\left\{\begin{pmatrix} a\\ b\\ c\end{pmatrix}\equiv e^{c\xi_3}e^{b\xi_2}e^{a\xi_1}\,|\,a,b,c\in\mathbf{R}\right\} \tag{4}$$

form the Bianchi II group G_{II}. ($e^{a\xi_i}$ denotes the one-parameter group of diffeomorphisms generated by ξ_i.) Note that the ''spatial point'' (x,y,z) is the image of the origin (0,0,0) by the ''group element'' (x,y,z), so that Eq. (3) itself gives the multiplication rule in the Bianchi II group.

From the commutation relations (1), we find two possible choices of obtaining a Gowdy space–time model, i.e., one (Type 1) is to keep ξ_2 and ξ_3 as the two commuting Killing vectors and consider inhomogeneity along ξ_1, and the other (Type 2) is to keep ξ_1 and ξ_3 as Killing vectors and consider inhomogeneity along ξ_2. We consider Type 1 first.

Note that the invariant one-forms of Bianchi II, given by

$$\sigma^1=dx,\quad \sigma^2=dy,\quad \sigma^3=dz-xdy, \tag{5}$$

are globally defined on $\tilde{M}=\mathbf{R}^3$, so that we can expand any metric on $\tilde{M}$ by these one-forms,

$$dl^2=h_{ij}\sigma^i\sigma^j, \tag{6}$$

with globally defined metric functions h_{ij}. Note also that the subgroup H_{II} of G_{II} of which action is discrete along ξ_1 is formed by

$$\left\{\begin{pmatrix} 2m\pi\\ b\\ c\end{pmatrix}\,\middle|\, m\in\mathbf{Z},b,c\in\mathbf{R}\right\}. \tag{7}$$

(The choice of the interval for the first component is arbitrary. Our choice of 2π is just for definiteness.) Since the invariant one-forms (5) are invariant under H_{II} (and G_{II} by definition), the metric (6) is invariant under H_{II} iff the metric functions h_{ij} are, too. This requirement is equivalent to h_{ij}'s depending only on x and being periodic with period 2π, i.e.,

$$h_{ij}=h_{ij}(x)=h_{ij}(x+2\pi). \tag{8}$$

The homogeneous limit can be achieved when $h_{ij}=$constants. So, we have found the inhomogeneous metric of (the universal cover of) a Gowdy space, given by Eq. (6) with the boundary condition (8).

Now, we can write down the ''appropriate'' space–time metric by[1] imposing the two-surface orthogonality and choosing the isothermal coordinates for the reference surface. After all, we obtain the space–time metric

$$ds^2=e^{-\lambda/2}t^{-1/2}(-dt^2+(\sigma^1)^2)+R[e^P(\sigma^2)^2+2e^PQ\sigma^2\sigma^3+(e^PQ^2+e^{-P})(\sigma^3)^2], \tag{9}$$

where P, Q, R, and λ are functions of t and x and are periodic in x with period 2π. We have followed the parametrization of a recent paper[6] to make comparisons easier. The isometry group for the space–time is unchanged, given by H_{II} [(7)].

This metric should be considered as the universal cover metric $\tilde{g}_{ab}$ of a Gowdy space–time $(M\times\mathbf{R},g_{ab})$. Any topology of M is possible if the fundamental group $\pi_1(M)$ can be represented

in H_{II}. One can choose such a $\pi_1(M)$ in the list of compact quotients modeled on Nil, presented in Ref. 9. For example, choose the manifold ''$b/1(n)$,'' characterized by

$$\pi_1(M_n)=\langle g_1,g_2,g_3;[g_1,g_2]=g_3^n,[g_1,g_3]=1,[g_2,g_3]=1\rangle, \tag{10}$$

where n is a positive integer parametrizing the family $b/1$ of topologies on Nil. Then, using the multiplication rule (3), we can find the representation up to conjugations

$$\Gamma_n=\{g_1,g_2,g_3\}=\left\{\begin{pmatrix}2p\pi\\ {g_1}^2\\ 0\end{pmatrix},\begin{pmatrix}2q\pi\\ {g_2}^2\\ {g_2}^3\end{pmatrix},\begin{pmatrix}0\\ 0\\ \frac{2\pi}{n}(p{g_2}^2-q{g_1}^2)\end{pmatrix}\right\}, \tag{11}$$

where $p,q\in Z$, and $g_1^2,g_2^2,g_2^3\in\mathbf{R}$. Thus, we have obtained the Gowdy space–time $(M_n\times\mathbf{R}, g_{ab})=(\tilde{M}_n\times\mathbf{R},\tilde{g}_{ab})/\Gamma_n$, where the action of Γ_n is defined by Eq. (3). We may think of the real parameters (g_1^2,g_2^2,g_2^3) as the moduli parameters for the space–time.

Type 2 can be obtained by interchanging the roles of σ^1 and σ^2. For example, the metric can be obtained from Eq. (9) by transforming $\sigma^1\to\sigma^2$, $\sigma^2\to\sigma^1$, $P(t,x)\to P(t,y)$, etc. The isometry group thereof is formed by the actions

$$\left\{\left.\begin{pmatrix}a\\ 2m\pi\\ c\end{pmatrix}\right| m\in Z,a,c\in\mathbf{R}\right\}. \tag{12}$$

As can be easily checked, we can represent the fundamental group (10) into this isometry group in a form similar to Eq. (11).

Let us consider the vacuum Einstein equations. To be specific, we consider the Type 1. (As for the final results, we will present for both Types 1 and 2.) Note that, if neglecting the boundary conditions, our metric (9) is essentially the same as the metric given in Ref. 1,

$$ds^2=e^{-\lambda/2}t^{-1/2}(-dt^2+dx^2)+R[e^P dy^2+2e^P Q dy dz+(e^P Q^2+e^{-P})dz^2], \tag{13}$$

where P, Q, R, and λ are functions of t and x. In fact, our metric (9) is obtained from this one by transforming

$$P\to P+\ln[(1-xQ)^2+x^2e^{-2P}],\quad Q\to\frac{Q(1-xQ)^2-xe^{-2P}}{(1-xQ)^2+x^2e^{-2P}} \tag{14}$$

(though this does not preserve the periodicity of P and Q). Moreover, the role of R as the area function of the group orbits, consisting of flat T^2's,[17] is the same for both metrics. This can be checked by noticing that the natural volume element of the second term in the metric (9) is given by $R dy\wedge dz$. As a result, the function R of our metric satisfies the same key equation as that of metric (13), i.e.,[1]

$$\partial_{tt}R-\partial_{xx}R=0. \tag{15}$$

This equation can also be checked by a direct substitution into the vacuum Einstein equation. Then, since the group orbits for our spatial manifold do not degenerate everywhere (except at the initial singularity), R can be taken as $t(\equiv e^{-\tau})$, as in the Gowdy model on $T^3\times\mathbf{R}$.[1] Here $t=0(\tau=+\infty)$ corresponds to the initial singularity. With this choice of R, the remaining independent Einstein's equations for our metric (9) are found by a direct calculation to be

$$\begin{aligned}&\ddot{P}-e^{-2\tau}P''-e^{2P}\dot{Q}^2+e^{-2\tau}[e^{2P}(Q^2\pm Q')^2\pm 2Q'-e^{-2P}]=0,\\ &\ddot{Q}-e^{-2\tau}Q''+2\dot{P}\dot{Q}-2e^{-2\tau}[P'Q'\pm(P'\mp Q)(Q^2+e^{-2P})]=0,\end{aligned} \tag{16}$$

and

$$\begin{aligned}&\lambda' - 2(P' \mp 2Q)\dot{P} \mp 2[e^{2P}(Q^2 \pm Q') - 1]\dot{Q} = 0,\\ &\dot{\lambda} - \dot{P}^2 - e^{2P}\dot{Q}^2 - e^{-2\tau}[e^{2P}(Q^2 \pm Q')^2 + P'^2 + 2Q^2 \mp 2(Q' + 2P'Q) + e^{-2P}] = 0,\end{aligned} \tag{17}$$

where dot and dash denote τ and x (or y for the Type 2) derivatives, respectively. The upper and lower signs are for Types 1 and 2, respectively. Note that P and Q are not constrained, since λ does not appear in Eqs. (16). This is one of the advantages of vacuum Gowdy models. The integrability condition for the constraint equations (17) for λ is automatically satisfied with Eqs. (16). The boundary conditions for P, Q, and λ are the spatially periodic ones. The Hamiltonian for the dynamical equations (16) can be guessed from straightforwardly reducing the Einstein–Hilbert action with the metric (9). It is given by

$$H = \frac{1}{2}\int_0^{2\pi} d\theta[\pi_P^2 + e^{-2P}\pi_Q^2] + e^{-2\tau}[(P' \mp 2Q)^2 - 2(Q^2 \pm Q') + e^{2P}(Q^2 \pm Q')^2 + e^{-2P}], \tag{18}$$

where π_P and π_Q are the conjugate momenta of P and Q, respectively. The integration measure θ is x for Type 1, or y for Type 2.

IV. GOWDY ON SOL×R AND VII$_0$×R

The next examples correspond to Bianchi VI$_0$ and VII$_0$ as minimal geometries. Their maximal geometries are Sol and E^3, respectively. Since the local structures for Bianchi VI$_0$ and VII$_0$ are similar, we treat them in parallel in this section. The basic procedure is the same as that of the previous section.

We first observe the commutation relations of the sets of the three Killing vectors of the Bianchi groups

$$[\xi_1, \xi_2] = 0, \quad [\xi_2, \xi_3] = -\xi_1, \quad [\xi_3, \xi_1] = \xi_2 : \ \mathrm{VI}_0, \tag{19}$$

$$[\xi_1, \xi_2] = 0, \quad [\xi_2, \xi_3] = -\xi_1, \quad [\xi_3, \xi_1] = -\xi_2 : \ \mathrm{VII}_0. \tag{20}$$

(We write ''tags'' to distinguish the two types as above.) In terms of coordinate basis,

$$\xi_1 = \frac{\partial}{\partial x} + \frac{\partial}{\partial y}, \quad \xi_2 = \frac{\partial}{\partial x} - \frac{\partial}{\partial y}, \quad \xi_3 = \frac{\partial}{\partial z} - x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} : \ \mathrm{VI}_0, \tag{21}$$

$$\xi_1 = \frac{\partial}{\partial x}, \quad \xi_2 = \frac{\partial}{\partial y}, \quad \xi_3 = \frac{\partial}{\partial z} - y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y} : \ \mathrm{VII}_0. \tag{22}$$

The Bianchi groups are formed by the finite actions of ξ_i's, which we take as

$$G_{\mathrm{VI}_0} = \left\{ \begin{pmatrix} a \\ b \\ c \end{pmatrix} \equiv e^{(a/2)(\xi_1+\xi_2)} e^{(b/2)(\xi_1-\xi_2)} e^{c\xi_3} \middle| a, b, c \in \mathbf{R} \right\}, \tag{23}$$

$$G_{\mathrm{VII}_0} = \left\{ \begin{pmatrix} a \\ b \\ c \end{pmatrix} \equiv e^{a\xi_1} e^{b\xi_2} e^{c\xi_3} \middle| a, b, c \in \mathbf{R} \right\}. \tag{24}$$

The actions or multiplications are given by

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a + e^{-c}x \\ b + e^{c}y \\ c + z \end{pmatrix} : \ \mathrm{VI}_0, \quad \begin{pmatrix} a \\ b \\ c \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \mathscr{R}_c \begin{pmatrix} x \\ y \end{pmatrix} \\ c + z \end{pmatrix} : \ \mathrm{VII}_0, \tag{25}$$

where $\mathscr{R}_c$ is the rotation matrix by angle c.

From the commutation relations (19) and (20), we find sole possibility of obtaining the Gowdy spaces, the ones inhomogeneous along ξ_3. Using the invariant one-forms

$$\sigma^1=\frac{1}{\sqrt{2}}(e^z dx+e^{-z}dy),\quad \sigma^2=\frac{1}{\sqrt{2}}(-e^z dx+e^{-z}dy),\quad \sigma^3=dz:\ \mathrm{VI}_0, \tag{26}$$

$$\sigma^1=\cos z dx+\sin z dy,\quad \sigma^2=-\sin z dx+\cos z dy,\quad \sigma^3=dz:\ \mathrm{VII}_0, \tag{27}$$

and the periodic functions

$$P=P(\tau,z)=P(\tau,z+z_0),\quad Q=Q(\tau,z)=Q(\tau,z+z_0),$$
$$\lambda=\lambda(\tau,z)=\lambda(\tau,z+z_0), \tag{28}$$

we can immediately write the same space–time metric for the two types as

$$ds^2=e^{-\lambda/2}e^{\tau/2}(-e^{-2\tau}d\tau^2+(\sigma^3)^2)+e^{-\tau}[e^P(\sigma^1)^2+2e^P Q\sigma^1\sigma^2+(e^P Q^2+e^{-P})(\sigma^2)^2]. \tag{29}$$

As we will see later, the period z_0 is specified depending upon the topology. The isometry groups H_{VI_0} and H_{VII_0} for the space–time metric (29) is formed by the discrete actions along ξ_3 and can be written in the same form

$$\left\{\begin{pmatrix} a \\ b \\ mz_0 \end{pmatrix}\middle| a,b\in\mathbf{R},m\in\mathbf{Z}\right\}. \tag{30}$$

In the following, we consider the vacuum Einstein equations first, since the compactifications must be discussed separately for each universal cover.

In the metric (29), we wrote the area function of the group orbit as $e^{-\tau}=t$, since, as in the Nil case, the group orbits do not degenerate except at the initial singularity for both types. The remaining Einstein equations are then found to be

$$\ddot{P}-e^{-2\tau}P''-e^{2P}\dot{Q}^2+e^{-2\tau}[e^{2P}(Q'+Q^2\mp 1)^2+2Q'-e^{-2P}]=0,$$
$$\ddot{Q}-e^{-2\tau}Q''+2\dot{P}\dot{Q}-2e^{-2\tau}[P'Q'+(P'-Q)(e^{-2P}+Q^2\mp 1)]=0, \tag{31}$$

and

$$\lambda'-2(P'-2Q)\dot{P}-2[e^{2P}(Q'+Q^2\mp 1)-1]\dot{Q}=0,$$
$$\dot{\lambda}-\dot{P}^2-e^{2P}\dot{Q}^2-e^{-2\tau}[(P'-2Q)^2-2(Q'+Q^2\mp 1)+e^{2P}(Q'+Q^2\mp 1)^2+e^{-2P}]=0, \tag{32}$$

where dot and dash denote, respectively, τ and z derivatives. The upper and lower signs are for $\mathrm{Sol}(\mathrm{VI}_0)$ and VII_0, respectively. The integrability condition for the constraint equations (32) for λ is automatically satisfied with Eqs. (31). The Hamiltonian for the dynamical equations (31) is given by

$$H=\frac{1}{2}\int_0^{z_0}dz[\pi_P^2+e^{-2P}\pi_Q^2]+e^{-2\tau}[(P'-2Q)^2-2(Q'+Q^2)+e^{2P}(Q'+Q^2\mp 1)^2+e^{-2P}]. \tag{33}$$

A. Compactification for Gowdy on Sol×R

We can take any topology of M of the Gowdy space–time $(M\times\mathbf{R},g_{ab})$ on $\mathrm{Sol}\times\mathbf{R}$ if we can represent the fundamental group $\pi_1(M)$ into H_{VI_0} (30) together with, if needed, the disconnected components defined with the discrete isometry

$$h:(x,y,z)\to(-x,-y,z). \tag{34}$$

As an explicit example, we consider the major sequence of compact quotients presented in Ref. 9. The fundamental groups are parametrized by an integer n such that $|n|>2$, and given by

$$\pi_1(M_n)=\langle g_1,g_2,g_3;[g_1,g_2]=1,\quad g_3g_1g_3^{-1}=g_2,\quad g_3g_2g_3^{-1}=g_1^{-1}g_2^n\rangle. \tag{35}$$

(This does not, however, exhaust all the compact quotients.[18]) We find that the representations are, up to conjugations, given by

$$\Gamma_n=\left\{\begin{pmatrix}\alpha u_1\\ \beta u_2\\ 0\end{pmatrix},\begin{pmatrix}\alpha v_1\\ \beta v_2\\ 0\end{pmatrix},\begin{pmatrix}0\\ 0\\ z_0\end{pmatrix}\right\}, \tag{36}$$

for $n>2$, and

$$\Gamma_n=\left\{\begin{pmatrix}\alpha u_1\\ \beta u_2\\ 0\end{pmatrix},\begin{pmatrix}\alpha v_1\\ \beta v_2\\ 0\end{pmatrix},h\circ\begin{pmatrix}0\\ 0\\ z_0\end{pmatrix}\right\}, \tag{37}$$

for $n<-2$, with $\alpha,\beta\in\mathbf{R}$. In these representations, (u_1,v_1), (u_2,v_2), and z_0 are determined in such a way that $\mathrm{sign}(n)e^{-z_0}$ and $\mathrm{sign}(n)e^{z_0}$ are the eigenvalues of matrix $\left(\begin{smallmatrix}0 & 1\\ -1 & n\end{smallmatrix}\right)$, and the corresponding normalized eigenvectors are (u_1,v_1) and (u_2,v_2), respectively. In particular, $e^{z_0}=|n+\sqrt{n^2-4}|/2$. We thus have obtained the Gowdy space–time $(M_n\times\mathbf{R},g_{ab})=(\tilde{M}_n\times\mathbf{R},\tilde{g}_{ab})/\Gamma_n$, where the universal cover metric $\tilde{g}_{ab}$ is given by Eq. (29) with one-forms (26).

B. Compactification for Gowdy on $\mathrm{VII}_0\times\mathbf{R}$

The Gowdy model that is most frequently picked up in the literature is the one on $T^3\times\mathbf{R}$, which would have been obtained by our procedure starting from Bianchi I. However, another $T^3\times\mathbf{R}$ model can be obtained from Bianchi VII_0, which we pick up here.

The fundamental group of T^3 is the infinite group with three commuting generators:

$$\pi_1(T^3)=\langle g_1,g_2,g_3;[g_1,g_2]=1,\quad [g_2,g_3]=1,\quad [g_3,g_1]=1\rangle. \tag{38}$$

The general solution of the representation into ''G_{VII_0}'' has already given in Eq. (97) of Ref. 9:

$$\Gamma=\left\{\begin{pmatrix}{g_1}^1\\ {g_1}^2\\ 2l\pi\end{pmatrix},\begin{pmatrix}{g_2}^1\\ {g_2}^2\\ 2m\pi\end{pmatrix},\begin{pmatrix}{g_3}^1\\ {g_3}^2\\ 2n\pi\end{pmatrix}\right\}, \tag{39}$$

where $g_1^1,g_1^2,g_2^1,g_2^2,g_3^1,g_3^2\in\mathbf{R}$, and $l,m,n\in Z$. We immediately notice that this representation is effective even in ''H_{VII_0}'' if we choose the period z_0 to 2π. So, this representation with $z_0=2\pi$ is the general solution. The only difference from the locally homogeneous case is that there are no effective conjugations. The representation (39) therefore gives the final form of the covering group. Finally, we have obtained the Gowdy space–time $(T^3\times\mathbf{R},g_{ab})=(\mathbf{R}^4,\tilde{g}_{ab})/\Gamma$, where the universal cover metric $\tilde{g}_{ab}$ is given by Eq. (29) with one-forms (27).

V. CONCLUSIONS

We have generalized Gowdy space–times to admit two commuting local Killing vectors. By this generalization, we gained rich varieties of new Gowdy space–times, but any advantages the original has, e.g., the simplicity of the vacuum Einstein equations, have not been lost. In this sense, the original definition demanding global existence of the two commuting Killing vectors was too much restricted. Our generalization is natural and useful for physical applications.

We have presented three new Gowdy space–times, on $\mathrm{Nil}\times\mathbf{R}$, on $\mathrm{Sol}\times\mathbf{R}$, and on $\mathrm{VII}_0\times\mathbf{R}$, which are closely related to Thurston geometries, Nil, Sol, and E^3, respectively. We have given not only the vacuum Einstein equations with boundary conditions but also an explicit representation of the covering group for each case. These three new models and the one called $T^3\times\mathbf{R}$ model in the literature have common features like (1) the group orbits do not degenerate

everywhere except at the initial singularity, and (2) there are two dynamical variables (i.e., P and Q). These features make the four models very similar. In fact, the only essential difference is the boundary condition for the metric. Since we wrote the space–time metrics in a suitable way for each case, their Einstein equations look different from each other, but the boundary conditions for the metric functions are the same, simply periodic. We could have written the space–time metric in a common form, e.g., like Eq. (13), but in that case, the boundary conditions for the metric functions would have taken inconvenient forms, as we have seen in the case of Nil.

It is worth noting that we can interpret the difference of the boundary conditions for the metrics as the difference of the ''background metrics.'' For example, the spatial metric of the (conventional) $T^3\times\mathbf{R}$ model can be smoothly deformed locally flat, so that we can think that the background is flat, while, say, the spatial metric of the Nil$\times\mathbf{R}$ model cannot be deformed locally flat but locally Bianchi II, so that the background is the Bianchi II locally homogeneous curved space in this case.

Here, we comment on the dynamics near the initial singularity of our three models. The Gowdy space–time on $T^3\times\mathbf{R}$ is conjectured[3] to be asymptotically velocity term dominated (AVTD).[2] Recent investigation by Berger and Garfinkle[6] supports this, *except for* measure-zero nongeneric spatial points. They succeeded in explaining the phenomenon by a potential picture, showing the nongeneric points correspond to the points where $Q'=0$. In our new models, such points correspond to points such that $Q'\pm Q^2=0$ (for Types 1 and 2 of Nil) and $Q'+Q^2\mp 1=0$ (for Sol and VII$_0$). Note that the points where $Q'=0$ are inevitable, since Q is periodic, but our four conditions are not necessarily satisfied in any spatial point. One may therefore expect that the Gowdy Nil$\times\mathbf{R}$, Sol$\times\mathbf{R}$, and VII$_0\times\mathbf{R}$ models are AVTD *everywhere*, so the ''curved backgrounds'' improve the AVTD behavior.

As a final remark, an extension of our method to the so-called $U(1)$ models[19] and other similar ones would also be possible, which is now under development, as well as a complete classification and further study of Gowdy models.

Note after the completion of this work: Recently, Weaver, Isenberg, and Berger[20] applied a Gowdy model on Sol$\times\mathbf{R}$ with magnetic field to examine the Mixmaster behavior toward the initial singularity. The boundary conditions imposed there are consistent with ours, though they look different since they used a metric similar to Eq. (13).

ACKNOWLEDGMENTS

The author thanks Professor Rendall for drawing his attention to Ref. 13. He also acknowledges financial support from the Japan Society for the Promotion of Science and the Ministry of Education, Science and Culture.

[1] R. H. Gowdy, Phys. Rev. Lett. **27**, 826 (1971); , 1102 (1971) (Erratum); Ann. Phys. **83**, 203 (1974).
[2] J. Isenberg and V. Moncrief, Ann. Phys. **199**, 84 (1990).
[3] B. Grubišić and V. Moncrief, Phys. Rev. D **47**, 2371 (1993).
[4] B. K. Berger and V. Moncrief, Phys. Rev. D **48**, 4676 (1993).
[5] S. D. Hern and J. M. Stewart, preprint, DAMTP R-97/41 (gr-qc/9708038, 1997).
[6] B. K. Berger and D. Garfinkle, preprint (gr-qc/9710102, 1997).
[7] See, e.g., L. D. Landau and E. M. Lifshitz, *Classical Theory of Fields* (MIT, Reading, MA, 1971).
[8] V. A. Belinski, I. M. Khalatnikov, and E. M. Lifshitz, Adv. Phys. **19**, 525 (1970); **31**, 639 (1982).
[9] T. Koike, M. Tanimoto, and A. Hosoya, J. Math. Phys. **35**, 4855 (1994).
[10] M. Tanimoto, T. Koike, and A. Hosoya, J. Math. Phys. **38**, 350 (1997).
[11] M. Tanimoto, T. Koike, and A. Hosoya, J. Math. Phys. **38**, 6560 (1997).
[12] H. V. Fagundes, Phys. Rev. Lett. **54**, 1200 (1985).
[13] A. D. Rendall, Commun. Math. Phys. **189**, 145 (1997).
[14] W. P. Thurston, *Three-Dimensional Geometry and Topology, Vol. 1* (Princeton U.P., Princeton, 1997).
[15] H. Nariai, Sci. Rep. Tohoku Univ., Ser. I **34**, 160 (1950); **35**, 62 (1951).
[16] R. Kantowski and R. K. Sachs, J. Math. Phys. **7**, 443 (1967).
[17] The group actions are well defined on $(\tilde{M},\tilde{h}_{ab})$, and the orbits form Euclid planes E^2. The compactification of $(\tilde{M},\tilde{h}_{ab})$ compactifies these E^2 to two-tori T^2, which we call the group orbits as well, though the group actions are defined only locally on the quotient.
[18] H. Kodama, Prog. Theor. Phys. **99**, 173 (1998).
[19] V. Moncrief, Ann. Phys. **167**, 118 (1986).
[20] M. Weaver, J. Isenberg, and B. K. Berger, preprint (gr-qc/9712055, 1997).

On the classification of subalgebras of the conformal algebra with respect to inner automorphisms

L. F. Barannyk
Institute of Mathematics, Higher Pedagogical School, ul. Arciszewskiego 22 b, 76200 Słupsk, Poland

P. Basarab-Horwath
Mathematics Department, University of Linköping, 581 83 Linköping, Sweden

W. I. Fushchych
Institute of Mathematics, Ukrainian Academy of Sciences, Tereshchenkivska Street 3, 252004 Kyïv, Ukraine

(Received 3 January 1997; accepted for publication 12 January 1998)

We give a complete justification of the classification of inequivalent subalgebras of the conformal algebra with respect to the inner automorphisms of the conformal group, and we perform the classification of the subalgebras of the conformal algebra $AC(1,3)$. © *1998 American Institute of Physics.* [S0022-2488(98)02309-3]

I. INTRODUCTION

The necessity of classifying the subalgebras of the conformal algebra is motivated by many problems in mathematics and mathematical physics.[1,2] The conformal algebra $AC(1,n)$ of Minkowski space $\mathbf{R}_{1,n}$ contains the extended Poincaré algebra $A\tilde{P}(1,n)$ and the full Galilei algebra $AG_4(n-1)$ (also known as the optical algebra). The classification of the subalgebras of the conformal algebra $AC(1,n)$ is almost reducible to the classification of the subalgebras of the algebras $A\tilde{P}(1,n)$ and $AG_4(n-1)$.

Patera, Winternitz and Zassenhaus[1] have given a general method for the classification of the subalgebras of inhomogeneous transformations. Using this method, the classification of the subalgebras $AP(1,n)$, $A\tilde{P}(1,n)$, and $AG_4(n-1)$ was carried out in Refs. 1–9 for $n=2,3,4$. In Refs. 7–11, this general method was supplemented by many structural results which made possible the algorithmization of the classification of the subalgebras of the Euclidean, Galilean, and Poincaré algebras for spaces of arbitrary dimensions. Indeed, this was done in Refs. 9 and 10, where the subalgebras of $AC(1,n)$ were classified up to conjugation under the conformal group $C(1,n)$ for $n=2,3,4$.

In order to perform the symmetry reduction of differential equations, it is necessary to identify the subalgebras of the symmetry algebra (of the equation) which give the same systems of basic invariants. This observation has led to the introduction in Ref. 12 of the concept of *I*-maximal subalgebras: a subalgebra F is said to be *I*-maximal if it contains every subalgebra of the symmetry algebra with the same invariants as F. In Ref. 13, all *I*-maximal subalgebras of $AC(1,4)$, classified upto $C(1,4)$-conjugation, were found in the representation defined on the solutions of the eikonal equation. Using these subalgebras, reductions of the eikonal and Hamilton–Jacobi equations to differential equations of lower order were obtained in Refs. 9 and 12. We note that the list of *I*-maximal subalgebras for a given algebra can differ according to the equation being investigated.

In the above works, the question of the connection between conjugation of the subalgebras of the algebra $A\tilde{P}(1,n)$ under the group $\tilde{P}(1,n)$ (or the group $AdA\tilde{P}(1,n)$ of inner automorphisms of the algebra $A\tilde{P}(1,n)$) and the conjugacy of these subalgebras under the group $C(1,n)$ was not dealt with. This, and the same problem for subalgebras of the Galilei algebra $AG_4(n-1)$, is the problem we address in the present article.

Since the group analysis of differential equations is of a local nature, we concentrate on conjugacy of the subalgebras under the group of inner automorphisms of the algebra $AC(1,n)$. Going over to conjugacy under $C(1,n)$ is not complicated, and requires only a further identification of the subalgebras under the action of at most three discrete symmetries. The results of this

 © 1998 American Institute of Physics

paper allow us to obtain a full classification of the subalgebras of $AC(1,n)$ for low values of n. On the basis of these results, we give at the end of this paper a classification of the algebra $AC(1,3)$ with respect to its group of inner automorphisms. The list of subalgebras obtained in this way can be used for the symmetry reduction of any system of differential equations which are invariant under $AC(1,3)$.

II. MAXIMAL SUBALGEBRAS OF THE CONFORMAL ALGEBRA

We denote by AdL the group of inner automorphisms of the Lie algebra L. Unless otherwise stated, conjugacy of subalgebras of L means conjugacy with respect to the group AdL. We consider AdL_1 as a subgroup of AdL_2 whenever L_1 is a subalgebra of L_2. The connected identity component of a Lie group H is denoted by H_1.

Let $\mathbf{R}_{1,n}$, $(n\geqslant 2)$, be Minkowski space with metric $g_{\alpha\beta}$, where $(g_{\alpha\beta})=\mathrm{diag}[1,-1,...,-1]$ and $\alpha,\beta=0,1,...,n$. The transformation defined by the equations

$$x_\alpha=x_\alpha(y_0,y_1,\ldots,y_n),\quad \alpha=0,1,...,n$$

of a domain $U\subset\mathbf{R}_{1,n}$ into $\mathbf{R}_{1,n}$, is said to be conformal if

$$\frac{\partial x_\mu}{\partial y^\alpha}\frac{\partial x_\nu}{\partial y^\beta}\,g^{\mu\nu}=\lambda(x)g_{\alpha\beta},$$

where $\lambda(x)\neq 0$ and $x=(x_0,x_1,\ldots,n)$. The conformal transformations of $\mathbf{R}_{1,n}$ form a Lie group, the conformal group $C(1,n)$. The Lie algebra $AC(1,n)$ of the group $C(1,n)$ has as its basis the generators of pseudorotations $J_{\alpha\beta}$, the translations P_α, the nonlinear conformal translations K_α, and the dilatations D, where $\alpha,\beta=0,1,...,n$. These generators satisfy the following commutation relations:

$$\begin{aligned}
&[J_{\alpha\beta},J_{\gamma\delta}]=g_{\alpha\delta}J_{\beta\gamma}+g_{\beta\gamma}J_{ga\delta}-g_{\alpha\gamma}J_{\beta\delta}-g_{\beta\delta}J_{\alpha\gamma},\\
&[P_\alpha,J_{\beta\gamma}]=g_{\alpha\beta}P_\gamma-g_{\alpha\gamma}P_\beta,\quad [P_\alpha,P_\beta]=0,\\
&[K_\alpha,J_{\beta\gamma}]=g_{\alpha\beta}K_\gamma-g_{\alpha\gamma}K_\beta,\ [K_\alpha,K_\beta]=0,\\
&[D,P_\alpha]=P_\alpha,\ [D,K_\alpha]=-K_\alpha,\ [D,J_{\alpha\beta}]=0,\\
&[K_\alpha,P_\beta]=2(g_{\alpha\beta}D-J_{\alpha\beta}).
\end{aligned}\tag{1}$$

The pseudo-orthogonal group $O(2,n+1)$ is the multiplicative group of all $(n+3)\times(n+3)$ real matrices C satisfying $C^tE_{2,n+1}C=E_{2,n+1}$, where $E_{2,n+1}=\mathrm{diag}[1,1,-1,...,-1]$. We denote by I_{ab} the $(n+3)\times(n+3)$ matrix whose entries are zero except for 1 in the (a,b) position, with $a,b=1,2,...,n+3$. The Lie algebra $AO(2,n+1)$ of $O(2,n+1)$ has as its basis the following operators:

$$\Omega_{12}=I_{12}-I_{21},\ \Omega_{ab}=-I_{ab}+I_{ba}\ \ (a<b;a,b=3,...,n+3),$$

$$\Omega_{ia}=-I_{ia}-I_{ai}\quad (i=1,2;a=3,...,n+3),$$

which satisfy the commutation relations

$$[\Omega_{ab},\Omega_{cd}]=\rho_{ad}\Omega_{bc}+\rho_{bc}\Omega_{ad}-\rho_{ac}\Omega_{bd}-\rho_{bd}\Omega_{ac}\quad (a,b,c,d=1,2,...,n+3),$$

where $(\rho_{ab})=E_{2,n+1}$. Let us denote by $\mathbf{R}_{2,n+1}$ the pseudo-Euclidean space of $n+3$ dimensions with metric ρ_{ab}. The matrices of the group $O(2,n+1)$ and the algebra $AO(2,n+1)$ will be identified with operators acting on the left on $\mathbf{R}_{2,n+1}$. Then, with this convention, $O(2,n+1)$ is the group of isometries of $\mathbf{R}_{2,n+1}$.

It is known (see for instance Ref. 9) that there is a homomorphism Ψ; $O(2,n+1)\to C(1,n)$ with kernel $\{\pm E_{n+3}\}$, where $\{E_{n+3}\}$ is the unit $(n+3)\times(n+3)$ matrix. Thus we are able to

identify $O(2,n+1)$ with $C(1,n)$. This homomorphism of groups induces an isomorphism f of the corresponding Lie algebras, f: $AO(2,n+1)\to AC(1,n)$, which is given by

$$f(\Omega_{\alpha+2,\beta+2})=J_{\alpha\beta},\quad f(\Omega_{1,\alpha+2}-\Omega_{a+2,n+3})=P_{\alpha},$$

$$f(\Omega_{1,\alpha+2}+\Omega_{\alpha+2,n+3})=K_{\alpha},\quad f(\Omega_{1,n+3})=-D\quad(\alpha,\beta=0,1,...,n).$$

We shall in this article identify the two algebras, using this isomorphism, so that we can write the previous equations as

$$\Omega_{\alpha+2,\beta+2}=J_{\alpha\beta},\quad \Omega_{1,\alpha+2}-\Omega_{\alpha+2,n+3}=P_{\alpha},$$

$$\Omega_{1,\alpha+2}+\Omega_{\alpha+2,n+3}=K_{\alpha},\quad \Omega_{1,n+3}=-D\quad(\alpha<\beta;\alpha,\beta=0,1,...,n).$$

We shall use the matrix realization of the conformal algebra.

Each matrix C which belongs to the identity component $O_1(2,n+1)$ of the group $O(2,n+1)$ is a product of matrices which are rotations in the x_1x_2 and x_ax_b planes ($a<b$; $a,b=3,...,n+3$) and hyperbolic rotations in the x_ix_a planes ($i=1,2$; $a=3,...,n+3$). Thus each such matrix C can be given as a finite product of matrices of the form $\exp X$, where $X\in AO(2,n+1)$. From this, it follows that each inner automorphism of the algebra $AO(2,n+1)$ is a mapping

$$\varphi_C: Y\to CYC^{-1}, \tag{2}$$

where $Y\in AO(2,n+1)$ and $C\in O_1(2,n+1)$, and conversely each mapping of this type is an inner automorphism of the algebra $AO(2,n+1)$.

In the process of our investigation mappings of the above type (2) will occur for certain matrices $C\in O(2,n+1)$, so we call these types of mappings $O(2,n+1)$-automorphisms of the algebra $AO(2,n+1)$ corresponding to the matrix C.

If G is the group of $O(2,n+1)$-automorphisms of the algebra $AO(2,n+1)$, and H is the subgroup of G consisting of its inner automorphisms, then H is normal in G and $[G:H]\leqslant 4$. Representatives of the cosets of G/H different from the identity will be

$$C_1=\mathrm{diag}[-1,1,...,1,-1],\quad C_2=\mathrm{diag}[1,1,-1,1,...,1],$$
$$C_3=\mathrm{diag}[-1,1-1,1,...,1,-1], \tag{3}$$

or

$$C_1=\mathrm{diag}[1,-1,1,...,1,-1,1],\quad C_2=\mathrm{diag}[1,1,-1,1,...,1],$$
$$C_3=\mathrm{diag}[1,-1,-1,1...,1,-1,1]. \tag{4}$$

Given a subspace V of $\mathbf{R}_{2,n+1}$, there is a maximal subalgebra of $AO(2,n+1)$ which leaves V invariant. We call this algebra the normalizer in $AO(2,n+1)$ of the subspace V.

Let $Q_1,\ldots,Q_{n+3}$ be a system of unit vectors in $\mathbf{R}_{2,n+1}$. Then the normalizer in $AO(2,n+1)$ of the isotropic subspace $\langle Q_1+Q_{n+3}\rangle$ is the extended Poincaré algebra

$$\tilde{AP}(1,n)=\langle P_0,P_1,\ldots,P_n\rangle\uplus(AO(1,n)\oplus\langle D\rangle),$$

where $\uplus$ denotes semidirect sum, and $\oplus$ denotes direct sum of algebras; $AO(1,n)=\langle J_{\alpha\beta}:\alpha,\beta=0,1,...,n\rangle$. The normalizer in $AO(2,n+1)$ of the completely isotropic subspace $\langle Q_1+Q_{n+3},Q_2+Q_{n+2}\rangle$ is the full Galilei algebra

$$AG_4(n-1)=\langle M,P_1,\ldots,P_{n-1},G_1,\ldots,G_{n-1}\rangle\uplus(AO(n-1)\oplus\langle R,S,T\rangle\oplus\langle Z\rangle),$$

where

$$M=P_0+P_n,\quad G_a=J_{0a}-J_{an}\quad(a=1,...,n-1),\quad R=-(J_{0n}+D),$$

$$S=\tfrac{1}{2}(K_0+K_n),\quad T=\tfrac{1}{2}(P_0-P_n),\quad Z=J_{0n}-D.$$

The generators of the algebra $AG_4(n-1)$ satisfy the following commutation relations:

$$[J_{ab},J_{cd}]=g_{ad}J_{bc}+g_{bc}J_{ad}-g_{ac}J_{bd}-g_{bd}J_{ac},$$

$$[G_a,J_{bc}]=g_{ab}G_c-g_{ac}G_b;\ [P_a,J_{bc}]=g_{ab}P_c-g_{ac}P_b,$$

$$[G_a,G_b]=0;\ [P_a,G_b]=\delta_{ab}M;\ [G_a,M]=0;\ [P_a,M]=0,$$

$$[J_{ab},M]=0;\ [R,S]=2S;\ [R,T]=-2T;\ [T,S]=R,$$

$$[Z,R]=[Z,S]=[Z,T]=[Z,J_{ab}]=0,$$

$$[R,G_a]=G_a;\ [R,P_a]=-P_a;\ [R,M]=0;\ [R,J_{ab}]=0,$$

$$[S,G_a]=0;\ [S,P_a]=-G_a;\ [S,M]=0;\ [S,J_{ab}]=0,$$

$$[T,G_a]=P_a;\ [T,P_a]=0;\ [T,M]=0;\ [T,J_{ab}]=0,$$

$$[Z,G_a]=-G_a;\ [Z,P_a]=-P_a;\ [Z,M]=-2M,$$

with $a,b,c,d=1,...,n-1$.

From these commutation relations we find that

$$\langle R,S,T\rangle=ASL(2,\mathbf{R});\ \langle R,S,T\rangle\oplus\langle Z\rangle=AGL(2,\mathrm{R}),$$

where $\mathbf{R}$ denotes the field of real numbers.

Let F be a reducible subalgebra of $AO(2,n+1)$. That is, there exists in $\mathbf{R}_{2,n+1}$ a nontrivial subspace W which is invariant under F. If W is isotropic, then there exists a totally isotropic subspace $W_0\subset W$ which is invariant under F. Since dim W_0 is 1 or 2, then, by Witt's theorem,[14] there exists an isometry $C\in O(2,n+1)$ such that CW_0 is either $\langle Q_1+Q_{n+3}\rangle$ or $\langle Q_1+Q_{n+3},Q_2+Q_{n+2}\rangle$. Taking into account that the matrices (3) do not change these subspaces and represent all the components of the group $O(2,n+1)$ different from the identity component $O_1(2,n+1)$, then we may assume that the above C lies in $O_1(2,n+1)$, the identity component. Thus there exists an inner automorphism φ of the algebra $AO(2,n+1)$ such that either $\varphi(F)\subset A\tilde{P}(1,n)$ or $\varphi(F)\subset AG_4(n-1)$.

If W is a nondegenerate subspace, then, by Witt's theorem, it is isometric with one of the following subspaces: $\mathbf{R}_{1,k}(k\geqslant 2)$, $\mathbf{R}_{2,k}(k\geqslant 1)$, $\mathbf{R}_k(k\geqslant 1)$. Each of the isometries (3) leaves invariant each of these subspaces, so that we may assume that the isometry which maps W onto one of these subspaces belongs to $O_1(2,n+1)$. From this, it follows that a subalgebra F is conjugate under the group of inner automorphisms of the algebra $AO(2,n+1)$ to a subalgebra of one of the following algebras:

(1) $AO'(1,k)\oplus AO''(1,n-k+1)$, where $AO'(1,k)=\langle\Omega_{ab}:a,b=1,3,...,k+2\rangle$ and $AO''(1,n-k+1)=\langle\Omega_{ab}:a,b=2,k+3,...,n+3\rangle$ with $n\geqslant 3$ and $k=2,...,[(n+1)/2]$;
(2) $AO(2,k)\oplus AO(n-k+1)$, where $AO(n-k+1)=\langle\Omega_{ab}:a,b=k+3,...,n+3\rangle$ with $k=0,1,...,n$.

In order to classify the subalgebras of these direct sums it is necessary to know the irreducible subalgebras of algebras of the type $AO(1,m)$ $(m\geqslant 2)$ and $AO(2,m)$, $(m\geqslant 3)$. It has been shown in Ref. 15 that $AO(1,m)$ has no irreducible subalgebras different from $AO(1,m)$. In Refs. 16 and 17 it has been shown that every semisimple irreducible subalgebra of $AO(2,m)$ $(m\geqslant 3)$ can be mapped by an automorphism of this algebra onto one of the following algebras:

(1) $AO(2,m)$;
(2) $ASU[1,(m/2)]$ when m is even;
(3) $\langle\Omega_{14}+\sqrt{3}\Omega_{13}+\Omega_{25},-\Omega_{15}+\Omega_{24}-\sqrt{3}\Omega_{23},\Omega_{12}-2\Omega_{45}\rangle$ when $m=3$.

It follows then that when $m>3$ is odd, the algebra $AO(2,m)$ has no irreducible subalgebras other than $AO(2,m)$. If $m=2k$ and $k\geqslant 2$, then, up to inner automorphisms, $AO(2,m)$ has two nontrivial maximal irreducible subalgebras: $ASU(1,k)\oplus\langle Y\rangle$, and $ASU(1,k)'\oplus\langle Y'\rangle$, where

$$Y=\mathrm{diag}[J,\ldots,J],\quad Y'=\mathrm{diag}[J,-J,J,\ldots,J]$$

with

$$J=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}.$$

We note that a subalgebra L of $AG_4(n-1)$ is conjugate under $Ad\ AO(2,n+1)$ with a subalgebra the algebra $A\tilde{P}(1,n)$ if and only if the projection of L onto $AGL(2,\mathbf{R})=\langle R,S,T\rangle\oplus\langle Z\rangle$ is conjugate under $Ad\ AGL(2,\mathbf{R})$ with a subalgebra of the algebra $\langle R,T,Z\rangle$.

III. CONJUGACY UNDER *Ad AP*(1,*n*) OF SUBALGEBRAS OF THE POINCARÉ ALGEBRA *AP*(1,*n*)

The Poincaré group $P(1,n)$ is the multiplicative group of matrices

$$\begin{pmatrix}\Delta & Y\\ 0 & 1\end{pmatrix},$$

where $\Delta\in O(1,n)$ and $Y\in\mathbf{R}^{n+1}$. Let I'_{ab}, $a,b=0,1,...,n+1$ be the $(n+2)\times(n+2)$ matrix whose entries are all zero except for the ab-entry, which is unity. Then a basis for $AP(1,n)$ is given by the matrices

$$J_{0a}=-I'_{0a}-I'_{a0},\quad J_{ab}=-I'_{ab}+I'_{ba},$$

$$P_0=I'_{0,n+1},\quad P_a=I'_{a,n+1},$$

with $a<b;a,b=1,...,n$. These basis elements obey the commutation relations (1). It is sometimes useful in calculations to identify elements of $AO(1,n)$ with matrices of the form

$$X=\begin{pmatrix}0 & \beta_{01} & \beta_{02} & \ldots & \beta_{0n}\\ \beta_{01} & 0 & \beta_{12} & \ldots & \beta_{1n}\\ \beta_{02} & -\beta_{12} & 0 & \ldots & \beta_{2n}\\ . & . & . & . & .\\ \beta_{0n} & -\beta_{1n} & -\beta_{2n} & \ldots & 0\end{pmatrix}$$

and elements of the space $U=\langle P_0,\ldots,P_n\rangle$ are represented by $n+1$-dimensional columns Y. In this case, we take

$$P_0=\begin{pmatrix}1\\0\\.\\.\\.\\0\end{pmatrix},\quad P_1=\begin{pmatrix}0\\1\\.\\.\\.\\0\end{pmatrix},\quad \ldots,\quad P_n=\begin{pmatrix}0\\0\\.\\.\\.\\1\end{pmatrix}$$

and with this notation it is easy to see that $[X,Y]=XY$. We endow the space U with the metric of the pseudo-Euclidean space $\mathbf{R}_{1,n}$, so that the inner product of two vectors

$$\begin{pmatrix} x_0 \\ x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

is $x_0y_0-x_1y_1-\cdots-x_ny_n$. The projection of $AP(1,n)$ onto $AO(1,n)$ is denoted by $\hat{\varepsilon}$. We also note that $AO(n)$, contained in $AO(1,n)$, is generated by J_{ab} $(a<b;\ a,b=1,...,n)$.

Let B be a Lie subalgebra of the algebra $AO(1,n)$ which has no invariant isotropic subspaces in $\mathbf{R}_{1,n}$. Then B is conjugate under $Ad\ AO(1,n)$ to a subalgebra of $AO(n)$ or to $AO(1,k)\oplus C$ where $k\geqslant 2$ and C is a subalgebra of the orthogonal algebra $AO'(n-k)$ generated by the matrices J_{ab} $(a,b=k+1,...,n)$. In the first case, B is not conjugate to any subalgebra of $AO(n-1)$.

Proposition 1: Let B be a subalgebra of $AO(n)$ which is not conjugate to a subalgebra of $AO(n-1)$. If L is a subalgebra of $AP(1,n)$ and $\hat{\varepsilon}(L)=B$, then L is conjugate to an algebra $W\uplus C$, where W is a subalgebra of $\langle P_1,...,P_n\rangle$, and C is a subalgebra of $B\oplus\langle P_0\rangle$. Two subalgebras $W_1\uplus C_1$ and $W_2\uplus C_2$ of this type are conjugate to each other under $Ad\ AP(1,n)$ if and only if they are conjugate under $Ad\ AO(n)$.

Proof: The algebra B is a completely reducible algebra of linear transformations of the space U and annuls only the subspace $\langle P_0\rangle$ (other than the null subspace itself). Thus, by Theorem 1.5.3,[9] the algebra L is conjugate to an algebra of the form $W\uplus C$ where $W\subset\langle P_1,...,P_n\rangle$ and $C\subset B\oplus\langle P_0\rangle$. Now let $W_1\uplus C_1$ and $W_2\uplus C_2$ be of this form, conjugate under $Ad\ AP(1,n)$. Then there exists a matrix $\Gamma\in P_1(1,n)$ such that $\varphi_\Gamma(W_1\uplus C_1)=W_2\uplus C_2$, and from this it follows that $\varphi_\Lambda(B_1)=B_2$ for some $\Lambda\in O_1(1,n)$. Let $V=\langle P_1,...,P_n\rangle$. Since $[B_1,V]=V$, then $[B_2,\varphi_\Lambda(V)]=\varphi_\Lambda(V)$ and $\varphi_\Lambda(V)=V$. Thus we can assume that $\Lambda=\mathrm{diag}[1,\Lambda_1]$ where $\Lambda_1\in SO(n)$, so that the given algebras are conjugate under $Ad\ AO(n)$. The converse is obvious.

Proposition 2: Let $B=AO(1,k)\oplus C$, where $k\geqslant 2$ and $C\subset AO'(n-k)$. If L is a subalgebra of $AP(1,n)$ and $\hat{\varepsilon}(L)=B$ then L is conjugate to $L_1\oplus L_2$ where $L_1=AO(1,k)$ or $L_1=AP(1,k)$, and L_2 is a subalgebra of the Euclidean algebra $AE'(n-k)$ with basis P_a,J_{ab}, $(a,b=k+1,...,n)$. Two subalgebras of this form, $L_1\oplus L_2$ and $L_1'\oplus L_2'$ are conjugate under $Ad\ AP(1,n)$ if and only if $L_1=L_1'$ and L_2 is conjugate to L_2' under the group of $E'(n-k)$-automorphisms.

Proof: The proof is as in the proof of Proposition 1.

Lemma 1: If $C\in O(1,n)$ and $C(P_0+P_n)=\lambda(P_0+P_n)$ then $\lambda\neq 0$ and

$$C=\begin{pmatrix} \dfrac{1+\lambda^2(1+\mathbf{v}^2)}{2\lambda} & \lambda\mathbf{v}^tB & \dfrac{-1+\lambda^2(1-\mathbf{v}^2)}{2\lambda} \\ \mathbf{v} & B & -\mathbf{v} \\ \dfrac{-1+\lambda^2(1+\mathbf{v}^2)}{2\lambda} & \lambda\mathbf{v}^tB & \dfrac{1+\lambda^2(1-\mathbf{v}^2)}{2\lambda} \end{pmatrix}, \tag{5}$$

where $B\in O(n-1)$, $\mathbf{v}$ is an $(n-1)$-dimensional column vector, $\mathbf{v}^2$ is the scalar square of $\mathbf{v}$ and $\mathbf{v}^t$ is the transpose of $\mathbf{v}$. Conversely, every matrix C of this form satisfies $C(P_0+P_n)=\lambda(P_0+P_n)$.

Proof: Proof is by direct calculation.

Lemma 2: Let $C\in O(1,n)$ have the form (5), with $\lambda>0$. Then

$$C=\mathrm{diag}[1,B,1]\exp[(-\ln\lambda)J_{0n}]\exp(-\beta_1G_1-\cdots-\beta_{n-1}G_{n-1}),$$

where $G_a=J_{0a}-J_{an}$ and

$$\begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-1} \end{pmatrix}=B^{-1}\mathbf{v}.$$

Proof: Direct calculation gives us

$$\exp(-\theta J_{0n})=\begin{pmatrix}\cosh\theta & 0 & \sinh\theta\\ 0 & E_{n-1} & 0\\ \sinh\theta & 0 & \cosh\theta\end{pmatrix}$$

and

$$\exp(-\beta_1 G_1-\cdots-\beta_{n-1}G_{n-1})=\begin{pmatrix}1+\frac{\mathbf{b}^2}{2} & \mathbf{b}^t & \frac{\mathbf{b}^2}{2}\\ \mathbf{b} & E_{n-1} & -\mathbf{b}\\ \frac{\mathbf{b}^2}{2} & \mathbf{b}^t & 1-\frac{\mathbf{b}^2}{2}\end{pmatrix},$$

where $\mathbf{b}=(\beta_1,\ldots,\beta_{n-1})^t$. On putting $\lambda=\exp\theta$ we have

$$\cosh\theta=\frac{\lambda^2+1}{2\lambda},\quad \sinh\theta=\frac{\lambda^2-1}{2\lambda}.$$

Since we have

$$\begin{pmatrix}\frac{\lambda^2+1}{2\lambda} & 0 & \frac{\lambda^2-1}{2\lambda}\\ 0 & E_{n-1} & 0\\ \frac{\lambda^2-1}{2\lambda} & 0 & \frac{\lambda^2+1}{2\lambda}\end{pmatrix}\begin{pmatrix}1+\frac{\mathbf{b}^2}{2} & \mathbf{b}^t & -\frac{\mathbf{b}^2}{2}\\ \mathbf{b} & E_{n-1} & -\mathbf{b}\\ \frac{\mathbf{b}^2}{2} & \mathbf{b}^t & 1-\frac{\mathbf{b}^2}{2}\end{pmatrix}$$

$$=\begin{pmatrix}\frac{1+\lambda^2(1+\mathbf{b}^2)}{2\lambda} & \lambda\mathbf{b}^t & \frac{-1+\lambda^2(1-\mathbf{b}^2)}{2\lambda}\\ \mathbf{b} & E_{n-1} & -\mathbf{b}\\ \frac{-1+\lambda^2(1+\mathbf{b}^2)}{2\lambda} & \lambda\mathbf{b}^t & \frac{1+\lambda^2(1-\mathbf{b}^2)}{2\lambda}\end{pmatrix},$$

then

$$\exp(-\theta J_{0n})\exp(-\beta_1 G_1-\cdots-\beta_{n-1}G_{n-1})=\mathrm{diag}[1,\beta^{-1},1]C$$

from which it follows directly that

$$C=\mathrm{diag}[1,B,1]\exp[(-\ln\lambda)J_{0n}]\exp(-\beta_1 G_1-\cdots-\beta_{n-1}G_{n-1})$$

and the lemma is proved.

The set of F of matrices of the form (5) with $\lambda>0$ is a group under multiplication. The mapping

$$C\to\begin{pmatrix}\lambda B & \lambda\mathbf{v}\\ 0 & 1\end{pmatrix}$$

is an isomorphism of the group F onto the extended Euclidean group $\tilde{E}(n-1)$. Thus we shall mean the group F when talking of the extended Euclidean group, and the connected identity component $\tilde{E}_1(n-1)$ will be identified with the group of matrices of the form (5) with $\lambda>0$ and $B\in SO(n-1)$. From Lemma 2 it follows that the Lie algebra AF of the group F is generated by the basis elements J_{ab},G_a,J_{0n} $(a<b;\ a,b=1,\ldots,n-1)$.

Lemma 3: If $C\in O_1(1,n)$ and $C(P_0+P_n)=\lambda(P_0+P_n)$ then $\lambda>0$ and $B\in SO(n-1)$ in (5).

Proof: Since

$$\frac{1+\lambda^2(1+\mathbf{v}^2)}{2\lambda}>0,$$

then we have $\lambda>0$. From Lemma 2, $\mathrm{diag}[1,B,1]\in O_1(1,n)$, so that $\det B>0$. Thus $B\in SO(n-1)$ and the lemma is proved.

Lemma 4: If $C\in O(1,n)$ and $\pm C\notin\widetilde{E}(n-1)$ then $C=\pm A_1C'A_2$ where $A_1,A_2\in\widetilde{E}(n-1)$ and $C'=\mathrm{diag}[1,...,1,-1]$.

Proof: We can choose a matrix $\Lambda\in O(n-1)$ so that $\Lambda C(P_0+P_n)=\alpha P_0+\beta P_1+\gamma P_n$ where $\alpha^2-\beta^2-\gamma^2=0$. If $\beta\neq 0$ then $\alpha-\gamma\neq 0$. Let $\theta=\beta/(\alpha-\gamma)$. Then,

$$\exp(\theta G_1)(\alpha P_0+\beta P_1+\gamma P_n)=\frac{\alpha-\gamma}{2}\,(P_0-P_n)$$

and so there exists a matrix $\Gamma\in\widetilde{E}(n-1)$ such that $\Gamma C(P_0+P_n)=\lambda(P_0+P_n)$ or $\Gamma C(P_0+P_n)=\lambda(P_0-P_n)$. In the first case, $\pm\Gamma C\in\widetilde{E}(n-1)$ so that then we have $\pm C\in\widetilde{E}(n-1)$, which is impossible. In the second case, $C'\Gamma C(P_0+P_n)=\lambda(P_0+P_n)$. For $\lambda>0$ we find $C'\Gamma C\in\widetilde{E}(n-1)$. Put $C'\Gamma C=A_2$, $\Gamma=A_1^{-1}$. Then $C=A_1C'A_2$. If $\lambda<0$ then we put $-C'\Gamma C=A_2$, in which case $C=-A_1C'A_2$, and the lemma is proved.

Lemma 5: If $C\in O_1(1,n)$ and $C\notin\widetilde{E}_1(n-1)$, then $C=D_1QD_2$, where $D_1,D_2\in\widetilde{E}_1(n-1)$, and $Q=\mathrm{diag}[1,-1,1,...,1,-1]$.

Proof: If $\pm C\in\widetilde{E}(n-1)$, then $C(P_0+P_n)=\gamma(P_0+P_n)$. By Lemma 3, $\gamma>0$ and $C\in\widetilde{E}_1(n-1)$, which contradicts the assumption. Thus, $\pm C\notin\widetilde{E}(n-1)$. By Lemma 4, $C=\pm A_1C'A_2$. From this it follows that $C=D_1\Gamma D_2$, where $D_1,D_2\in\widetilde{E}_1(n-1)$, and Γ is one of the matrices $\pm C',\pm Q$. However, $\Gamma\in O_1(1,n)$, since $\Gamma=D_1^{-1}CD_2^{-1}$, find from this it follows that $\Gamma=Q$. The Lemma is proved.

Direct calculation shows that the normalizer of the space $\langle P_0+P_n\rangle$ in $AO(1,n)$ is generated by the matrices G_a,J_{ab},J_{0n} $(a,b=1,...,n-1)$, which satisfy the commutation relations

$$[G_a,J_{bc}]=g_{ab}G_c-g_{ac}G_b,\ [G_a,G_b]=0,\ [G_a,J_{0n}]=G_a.$$

This means that the normalizer of the space $\langle P_0+P_n\rangle$ in the algebra $AO(1,n)$ is the extended Euclidean algebra

$$A\widetilde{E}(n-1)=\langle G_1,\ldots,G_{n-1}\rangle\uplus(AO(n-1)\oplus\langle J_{0n}\rangle)$$

in an $(n-1)$-dimensional space, where the generators of translations are $G_1,\ldots,G_{n-1}$ and the generator of dilatations is the matrix J_{0n}.

Let K be a subalgebra of $AP(1,n)$ such that its projection onto $AO(1,n)$ has an invariant isotropic subspace in Minkowski space $\mathbf{R}_{1,n}$. The subalgebra K is conjugate under $AdAP(1,n)$ with a subalgebra of the algebra $\mathcal{A}=AG_1(n-1)\uplus\langle J_{0n}\rangle$ where $AG_1(n-1)$ is the usual Galilei algebra with basis M,T,P_a,G_a,J_{ab} $(a,b=1,...,n-1)$, and $M=P_0+P_n,T=\frac{1}{2}(P_0-P_n)$.

Proposition 3: Let L_1 and L_2 be subalgebras of $\mathcal{A}$, with L_1 not conjugate under $Ad\ \mathcal{A}$ to any subalgebra having zero projection onto $\langle G_1,\ldots,G_{n-1}\rangle$. If $\varphi(L_1)=L_2$ for some $\varphi\in AdAP(1,n)$, then there exists an inner automorphism ψ of the algebra $\mathcal{A}$ with $\psi(L_1)=L_2$.

Proof: Since $Ad\ \mathcal{A}$ contains automorphisms which correspond to matrices of the form

$$\exp\left(\sum_{\gamma=0}^{n}a_\gamma P_\gamma\right) \tag{6}$$

and since $P(1,n)$ is a semidirect product of the group of matrices of the form (6) and the group $O(1,n)$ of matrices of the form $\mathrm{diag}[\Delta,1]$, then we may assume that $\varphi=\varphi_C$ with $C\in O_1(1,n)$. If $C\notin\widetilde{E}_1(n-1)$, then by Lemma 5, $C=D_1QD_2$. In that case we find that

$$(D_1QD_2)\hat{\varepsilon}(L_1)(D_2^{-1}QD_1^{-1})=\hat{\varepsilon}(L_2),$$

whence

$$Q(D_2\hat{\varepsilon}(L_1)D_2^{-1})Q=D_1^{-1}\hat{\varepsilon}(L_2)D_1. \tag{7}$$

However,

$$QG_aQ=Q(J_{0a}-J_{an})Q=\begin{cases} J_{0a}+J_{an} & \text{when } a\neq 1 \\ -(J_{01}+J_{1n}) & \text{when } a=1 \end{cases}.$$

This means that $QG_aQ\notin\mathcal{A}$. Because of this, the left-hand side of (7) does not belong to $\mathcal{A}$, whereas the right-hand side of (7) is a subalgebra of $\mathcal{A}$. This then implies that we must have $C\in\tilde{E}_1(n-1)$ and thus we have $\psi(L_1)=L_2$ for some $\psi\in Ad\mathcal{A}$.

Proposition 4: Let $\tilde{\mathcal{A}}$ be a Lie algebra with basis $P_0,P_a,P_n,J_{ab},J_{0n}$ $(a,b=1,...,n-1)$ and let L_1,L_2 be subalgebras of $\tilde{\mathcal{A}}$ such that at least one of them has a nonzero projection onto $\langle J_{0n}\rangle$. If $\varphi(L_1)=L_2$ for some $\varphi\in AdAP(1,n)$, then there exists an inner automorphism $\psi\in\tilde{\mathcal{A}}$ so that either $\psi(L_1)=L_2$ or $\psi(L_1)=\varphi_Q(L_2)$ where $Q=\mathrm{diag}[1,-1,1,...,1,-1]$.

Proof: As in the proof of Proposition 3, we may assume that $\varphi=\varphi_C$ where $C\in O_1(1,n)$. We shall also assume that the projection of L_1 onto $\langle J_{0n}\rangle$ is nonzero. If $C\in\tilde{E}_1(n-1)$ and $C\notin\tilde{O}_1(n-1)$ then the projection of the algebra $\varphi(L_1)$ onto $\langle G_1,\ldots,G_{n-1}\rangle$ is nonzero, and hence the projection of L_2 onto $\langle G_1,\ldots,G_{n-1}\rangle$ is nonzero, which contradicts the assumptions of the proposition. Thus, if $C\in\tilde{E}_1(n-1)$ then $\varphi\in Ad\tilde{\mathcal{A}}$.

Let $C\notin\tilde{E}_1(n-1)$. By Lemma 5, $C=D_1QD_2$ where $D_1,D_2\in\tilde{E}_1(n-1)$. Then $\varphi(L_1)=L_2$ can be written as

$$\varphi_Q[\sigma_{D_2}(L_1)]=\varphi_{D_1^{-1}}(L_2).$$

If $D_2\notin\tilde{O}_1(n-1)$ then the projection of $\varphi_{D_2}(L_1)$ onto $\langle G_1,\ldots,G_{n-1}\rangle$ is nonzero and hence $\varphi_Q[\varphi_{D_2}(L_1)]$ does not belong to $\mathcal{A}$. But then $\varphi_{D_1^{-1}}(L_2)$ is also not in $\mathcal{A}$. This is a contradiction. Thus $D_1,D_2\in\tilde{O}_1(n-1)$. From this it follows that $\varphi_Q(\psi(L_1))=L_2$ where $\psi=\varphi_D$ is an inner automorphism of the algebra $\tilde{\mathcal{A}}$. This proves the proposition.

Proposition 5: Suppose $2\leqslant m\leqslant n-1$. Let F be a subalgebra of the algebra $AO(m)$ which is not conjugate under $AdAO(m)$ to a subalgebra of $AO(m-1)$, and let L be a subalgebra of $\langle P_0,P_1,\ldots,P_n\rangle\uplus F$ such that $\hat{\varepsilon}(L)=F$. Then L is conjugate to an algebra $W\uplus K$, where W is a subalgebra of $\langle P_1,\ldots,P_m\rangle$ and K is a subalgebra of $F\oplus\langle P_0,P_{m+1},\ldots,P_n\rangle$. Two subalgebras $W_1\uplus K_1$ and $W_2\uplus K_2$ of this type are conjugate under $AdAP(1,n)$ if and only if there exists an automorphism $\psi\in AdAO(m)\times AdAO(1,n-m)$ such that $\psi(W_1\uplus K_1)=W_2\uplus K_2$ or $\psi(W_1\uplus K_1)=Q(W_2\uplus K_2)Q$ where

$$AO(1,n-m)=\langle J_{\alpha\beta}:\alpha,\beta=0,m+1,...,n\rangle$$

and $Q=\mathrm{diag}[1,-1,1,...,1,-1]$.

IV. CONJUGACY OF SUBALGEBRAS OF THE EXTENDED POINCARÉ ALGEBRA $A\tilde{P}(1,n)$ UNDER $AdAC(1,n)$

Lemma 6: If $C\in O(2,n+1)$ and $C(Q_1+Q_{n+3})=\lambda(Q_1+Q_{n+3})$ then $\lambda\neq 0$ and

$$C=\begin{pmatrix} \dfrac{1+\lambda^2(1-\mathbf{v}^2)}{2\lambda} & -\lambda\mathbf{v}^tE_{1,n}B & \dfrac{-1+\lambda^2(1+\mathbf{v}^2)}{2\lambda} \\ \mathbf{v} & B & -\mathbf{v} \\ \dfrac{-1+\lambda^2(1-\mathbf{v}^2)}{2\lambda} & -\lambda\mathbf{v}^tE_{1,n}B & \dfrac{1+\lambda^2(1+\mathbf{v}^2)}{2\lambda} \end{pmatrix}, \tag{8}$$

where $B\in O(1,n)$, $E_{1,n}=\mathrm{diag}[1,-1,...,-1]$, $\mathbf{v}$ is an $(n+1)\times 1$ matrix and $\mathbf{v}^2$ is its scalar square in $\mathbf{R}_{1,n}$. Conversely, every matrix C of the form (8) satisfies the condition $C(Q_1+Q_{n+3})=\lambda(Q_1+Q_{n+3})$.

Proof: Direct calculation.

Lemma 7: Let $C\in O(2,n+1)$ have the form (8), with $\lambda>0$. Then

$$C=\mathrm{diag}[1,B,1]\exp[(\ln\lambda)D]\exp(-\beta_0P_0-\beta_1P_1-\cdots-\beta_nP_n),$$

where

$$\begin{pmatrix}\beta_0\\ \beta_1\\ \cdot\\ \cdot\\ \cdot\\ \beta_n\end{pmatrix}=B^{-1}\mathbf{v}.$$

Proof: The proof of Lemma 7 is similar to that of Lemma 2.

The mapping

$$f:C\to\begin{pmatrix}\lambda B & \lambda\mathbf{v}\\ 0 & 1\end{pmatrix}$$

is a homomorphism of the group of matrices (8) onto the extended Poincaré group $\tilde{P}(1,n)$. The kernel of this homomorphism is the group of order two, $\{-E_{n+3},E_{n+3}\}$. Let us denote by H the set of matrices of the form (8) with $\lambda>0$. Then f is an isomorphism of H onto $\tilde{P}(1,n)$. For this reason we shall, in the remainder of this article, mean the group H when referring to $\tilde{P}(1,n)$. Its Lie algebra is the extended Poincaré algebra $A\tilde{P}(1,n)$ given in Sec. II.

Lemma 8: Let $C\in O_1(2,n+1)$ and let it be of the form (8) with $\lambda>0$. Then $B\in O_1(1,n)$.

Remark 1: Note that when $\lambda<0$ it is possible that B does not belong to $O_1(2,n+1)$.

Lemma 9: If $C\in O(2,n+1)$ and $\pm C\notin\tilde{P}(1,n)$ then either $C=\pm A_1QA_2$ or $C=A_1F(\theta)A_2$, where $A_1,A_2\in\tilde{P}(1,n)$, $Q=\mathrm{diag}[1,\ldots,1,-1]$ and $F(\theta)=\exp[(\theta/2)(K_0+P_0+K_n-P_n)]$.

Proof: There exists a matrix $\Lambda\in\tilde{P}(1,n)$ such that

$$\Lambda C(Q_1+Q_{n+3})=\alpha_1Q_1+\alpha_2Q_2+\alpha_3Q_{n+2}+\alpha_4Q_{n+3},$$

where $\alpha_1^2+\alpha_2^2-\alpha_3^2-\alpha_4^2=0$ and $\alpha_2\alpha_3\geqslant 0$. If $\alpha_1\neq\alpha_4$ then, as in the proof of Lemma 4, we obtain that

$$\exp(\beta_0P_0+\beta_nP_n)\Lambda C(Q_1+Q_{n+3})=\gamma(Q_1\pm Q_{n+3})$$

for some real numbers β_0,β_n,γ. From this it follows that

$$\Gamma\ \exp(\beta_0P_0+\beta_nP_n)\Lambda C(Q_1+Q_{n+3})=\lambda(Q_1+Q_{n+3}),$$

where $\lambda>0$ and $\Gamma=\pm E_{n+3}$ or $\Gamma=\pm Q$. By Lemma 6 and Lemma 7, we obtain

$$\Gamma\ \exp(\beta_0P_0+\beta_nP_n)\Lambda C=\tilde{\Lambda},\ \ \tilde{\Lambda}\in\tilde{P}(1,n).$$

Since $\pm C\notin\tilde{P}(1,n)$, then $\Gamma=\pm Q$, and so $C=\pm A_1QA_2$, where $A_1=\Lambda^{-1}\exp(-\beta_0P_0-\beta_nP_n)$, $A_2=\tilde{\Lambda}$.

If $\alpha_1=\alpha_4$, then also $\alpha_2=\alpha_3$. It is easy to verify that

$$F(\theta)\Lambda C(Q_1+Q_{n+3})=(\alpha_1\cos\theta+\alpha_2\sin\theta)(Q_1+Q_{n+3})+(\alpha_2\cos\theta-\alpha_1\sin\theta)(Q_2+Q_{n+2}).$$

If $\alpha_1=0$ then we put $\theta=(\pi/2)$, when $\alpha_2>0$ and $\theta=-(\pi/2)$, when $\alpha_2<0$. If $\alpha_1\neq 0$ then we let $\alpha_2\cos\theta-\alpha_1\sin\theta=0$. In that case,

$$\tan\theta=\frac{\alpha_2}{\alpha_1},\ \ \alpha_1\cos\theta+\alpha_2\sin\theta=\alpha_1\cos\theta(1+\tan^2\theta).$$

We choose the value of θ so that $\alpha_1\cos\theta>0$. With this choice of θ we have

$$F(\theta)\Lambda C(Q_1+Q_{n+3})=\lambda(Q_1+Q_{n+3}),$$

where $\lambda>0$. But then, as a result of Lemma 6 and Lemma 7, $F(\theta)\Lambda C=\tilde{\Lambda}$, $\tilde{\Lambda}\in\tilde{P}(1,n)$, and so $C=A_1F(-\theta)A_2$, where $A_1=\Lambda^{-1}$, $A_2=\tilde{\Lambda}$. The result is proved.

Lemma 10: Let L_1 and L_2 be subalgebras of $A\tilde{P}(1,n)$ which are not conjugate under $A\tilde{P}(1,n)$ to subalgebras of $A\tilde{O}(1,n)=AO(1,n)\oplus\langle D\rangle$. Then L_1,L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under $AdA\tilde{P}(1,n)$ or if one of the following conditions holds:

(1) n is an odd number and there exists an automorphism $\psi\in AdA\tilde{P}(1,n)$ with $\psi(L_1)=C_2L_2C_2^{-1}$ (see Eq. (3) for notation);
(2) there exist automorphisms $\psi_1,\psi_2\in A\tilde{P}(1,n)$ with

$$\psi_1(L_1)=F(\theta)[\psi_2(L_2)]F(-\theta).$$

Proof: Let $CL_1C^{-1}=L_2$ for some $C\in O_1(2,n+1)$. By Lemma 9, we may assume that $\pm C\in\tilde{P}(1,n)$ or that C is one of the matrices $\pm A_1QA_2$, $A_1F(\theta)A_2$ (we use the notation of Lemma 9). If $C\in\tilde{P}(1,n)$ then, by Lemma 8, C belongs to the identity component of the group $\tilde{P}(1,n)$ and thus φ_C is an inner automorphism of the algebra $A\tilde{P}(1,n)$. Now suppose $-C\in\tilde{P}(1,n)$. Then by Lemma 7, $C=-\mathrm{diag}[1,B,1]\Delta$, where $B\in O(1,n)$ and $\Delta\in\tilde{P}_1(1,n)$. Thus we may assume that $C=-\mathrm{diag}[1,B,1]$. From this it follows that $B\in O_1(1,n)$ for odd n and we have

$$\mathrm{diag}[1,1,-1,1,...,1,1]B\in O_1(1,n).$$

For even n this means that the algebras L_1,L_2 are conjugate to each other under $AdA\tilde{P}(1,n)$ or that there exists an automorphism $\psi\in AdA\tilde{P}(1,n)$ such that $\psi(L_1)=C_2L_2C_2^{-1}$.

Let $C=\pm A_1QA_2$. Then $C=\Gamma_1\Delta\Gamma_2$ with $\Gamma_1,\Gamma_2\in\tilde{P}(1,n)$ and $\Delta=\pm\mathrm{diag}[1,\epsilon_1,1,...,1,\epsilon_2,-1]$ with ϵ_1, $\epsilon_2\in\{-1,1\}$. Clearly, $\Delta\in O_1(2,n+1)$. When $C=A_1QA_2$ we have $\epsilon_1=1$, $\epsilon_2=-1$ and when $C=-A_1QA_2$, $\epsilon_1=1$, $\epsilon_2=(-1)^n$. Since

$$\Delta P_n\Delta^{-1}=\pm K_n,\quad \Delta P_\alpha\Delta^{-1}=\pm K_\alpha$$

with $\alpha<n$, then from $\Gamma_1^{-1}L_2\Gamma_1=\Delta(\Gamma_2L_1\Gamma_2^{-1})\Delta^{-1}$ it follows that the algebra $\Gamma_1^{-1}L_2\Gamma_1$ has a nonzero projection onto $\langle K_0,K_1,\ldots,K_n\rangle$, which is impossible. Thus the matrix C is different from $\pm A_1QA_2$.

Now let $C=A_1F(\theta)A_2$. If Γ is one of the matrices (4), then $\Gamma F(\theta)\Gamma^{-1}=F(\pm\theta)$, so that

$$C=A_1'F(\theta)A_2'\Delta,$$

where A_1', $A_2'\in\tilde{P}_1(1,n)$ and $\Delta=E$ or Δ is one of the matrices (4). Since Δ can be represented as a product of matrices in $O_1(2,n)$, then the last case is impossible, and we have proved the Lemma.

Theorem 1: Let L_1 and L_2 be subalgebras of $A\tilde{P}(1,n)$ which are not conjugate under $A\tilde{P}(1,n)$ to subalgebras of $A\tilde{O}(1,n)$ and such that their projections onto $AO(1,n)$ have no invariant isotropic subspace in $\mathbf{R}_{1,n}$. The subalgebras L_1 and L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under $AdA\tilde{P}(1,n)$ or when there exists an automorphism $\psi\in AdA\tilde{P}(1,n)$ such that $\psi(L_1)=C_2L_2C_2^{-1}$, where $C_2=\mathrm{diag}[1,1,-1,1,...,1]$.

Proof: By Lemma 10 we may assume that $\psi_1(L_1)=F(\theta)[\psi_2(L_2)]F(-\theta)$ for some $\psi_1,\psi_2\in A\tilde{P}(1,n)$. Under the given assumptions, the projection of $\psi_2(L_2)$ onto $AO(1,n)$ contains an element of the form

$$X=\sum_{b=1}^{n-1}(\alpha_bJ_{0b}+\gamma_bJ_{bn})+\sum_{b,c=1}^{n-1}\sigma_{bc}J_{bc},$$

where $\alpha_q\neq-\gamma_q$ for some $q(1\leqslant q\leqslant n-1)$. Since

$$F(\theta)J_{0q}F(-\theta)=J_{0q}\cos\theta+\tfrac{1}{2}(K_q+P_q)\sin\theta$$

and

$$F(\theta)J_{qn}F(-\theta)=J_{qn}\cos\theta+\tfrac{1}{2}(K_q-P_q)\sin\theta$$

we have that $F(\theta)XF(-\theta)$ contains the term

$$F(\theta)[\alpha_q J_{0q}+\gamma_q J_{qn}]F(-\theta)=(\alpha_q J_{0q}+\gamma_q J_{qn})\cos\theta+\tfrac{1}{2}[\alpha_q(K_q+P_q)+\gamma_q(K_q-P_q)]\sin\theta$$

and from this it follows that $(\alpha_q+\gamma_q)\sin\theta=0$ so that $\sin\theta=0$. But then $\theta=m\pi$. When $m=2d$ we have $F(\theta)=E_{n+3}$. When $m=2d+1$ then $F(\theta)=\mathrm{diag}[-1,-1,E_{n-1},-1,-1]$. However,

$$F(\theta)[\psi_2(L_2)]F(-\theta)=(-F(\theta))[\psi_2(L_2)](-F(-\theta))$$

from which it follows that we may assume that $\psi_1(L_1)=C[\psi_2(L_2)]C^{-1}$ where $C=\mathrm{diag}[1,1,-E_{n-1},1,1]$. If n is odd, then φ_C is an inner automorphism of $A\tilde{P}(1,n)$. If n is even, then $\varphi_{C_2}\varphi_C$ is an inner automorphism of the algebra $A\tilde{P}(1,n)$. In the first case, $\psi_3(L_1)=L_2$ where $\psi_3=\psi_2^{-1}\varphi_C^{-1}\psi_1$ is an inner automorphism of the algebra $A\tilde{P}(1,n)$. In the second case, $\psi(L_1)=\varphi_{C_2}(L_2)$ for some $\psi\in AdA\tilde{P}(1,n)$. The theorem is proved.

Theorem 2: Let L_1 and L_2 be subalgebras of $A\tilde{O}(1,n)$ having no invariant isotropic subspaces in $\mathbf{R}_{1,n}$. The subalgebras L_1,L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under $AdA\tilde{O}(1,n)$ or when there exists an automorphism $\psi\in AdA\tilde{O}(1,n)$ such that $\psi(L_1)=CL_2C^{-1}$ where C is one of the $(n+3)\times(n+3)$ matrices

$$\mathrm{diag}[1,1,-1,1,...,1],\ \ \mathrm{diag}[1,...,1,-1],\ \ \mathrm{diag}[1,...,1,-1,-1].$$

We note that $A\tilde{O}(1,n)\subset AO(2,n+1)$ and that the matrix C is $(n+3)\times(n+3)$.

V. SUBALGEBRAS OF THE FULL GALILEI ALGEBRA

Lemma 11: Let $C\in O(2,n+1)$ and $W=\langle Q_1+Q_{n+3},Q_2+Q_{n+2}\rangle$. If $CW=W$, then

$$C=\exp[\theta(S+T)]\mathrm{diag}[1,\varepsilon,K,\varepsilon,1]\exp(\alpha R+\beta Z)\exp\left(\sum_{i=1}^{n-1}\gamma_i G_i\right)\exp\left(\delta M+\lambda T+\sum_{i=1}^{n-1}\mu_i P_i\right), \quad (9)$$

where $\varepsilon=\pm1$, $K\in O(n-1)$.

Proof: We have

$$C(Q_1+Q_{n+3})=\alpha_1(Q_1+Q_{n+3})+\alpha_2(Q_2+Q_{n+2})$$

and so

$$F(-\theta)C(Q_1+Q_{n+3})=(\alpha_1\cos\theta-\alpha_2\sin\theta)(Q_1+Q_{n+3})+(\alpha_2\cos\theta+\alpha_1\sin\theta)(Q_2+Q_{n+2}).$$

If $\alpha_1=0$ then we put $\theta=(3\pi/2)$ when $\alpha_2>0$ and $\theta=(\pi/2)$ when $\alpha_2<0$. If $\alpha_1\neq0$ then we put $\alpha_1\sin\theta+\alpha_2\cos\theta=0$ and then $\tan\theta=-\alpha_2/\alpha_1$ and $\alpha_1\cos\theta-\alpha_2\sin\theta=\alpha_1\cos\theta(1+\tan^2\theta)$. We choose θ so that $\alpha_1\cos\theta>0$. For this choice of θ we have $F(-\theta)C(Q_1+Q_{n+3})=\xi(Q_1+Q_{n+3})$, where $\xi>0$. Using Lemma 7, we obtain

$$F(-\theta)C=A=\mathrm{diag}[1,B,1]\exp([\ln\xi]D)\exp\left(-\sum_{i=0}^{n}\beta_i P_i\right)\in\tilde{P}(1,n),$$

where $B\in O(1,n)$. Then $C=F(\theta)A$. The matrix A has the form (8). Direct calculation gives

$$A(Q_2+Q_{n+2})=\alpha(Q_1+Q_{n+3})+\beta Q_2+\gamma Q_{n+2}+\sum_{i=3}^{n+1}\delta_i Q_i.$$

From this it follows that

$$F(\theta)A(Q_2+Q_{n+2})=(\alpha\cos\theta+\beta\sin\theta)Q_1+(-\alpha\sin\theta+\beta\cos\theta)Q_2+(\gamma\cos\theta-\alpha\sin\theta)Q_{n+2}+(\gamma\sin\theta+\alpha\cos\theta)Q_{n+3}+\sum_{i=3}^{n+1}\delta_i Q_i.$$

Now we have $F(\theta)A(Q_2+Q_{n+2})\in W$, from which we have

$$\alpha\cos\theta+\beta\sin\theta=\gamma\sin\theta+\alpha\cos\theta,\quad -\alpha\sin\theta+\beta\cos\theta=\gamma\cos\theta-\alpha\sin\theta$$

and so we conclude that $\beta=\gamma$ and $\delta_j=0$, $j=3,...,n+1$. But in that case we have

$$\mathrm{diag}[1,B,1](Q_2+Q_{n+2})=\beta(Q_2+Q_{n+2}).$$

By Lemma 2, we have

$$\pm B=\mathrm{diag}[1,K,1]\exp[(-\ln|\beta|)J_{0n}]\exp\left(\sum_{i=1}^{n-1}\gamma_i G_i\right),$$

where $K\in O(n-1)$. We note that

$$K_0+P_0-K_n-P_n=2(S+T),\quad J_{0n}=\tfrac{1}{2}(Z-R),\quad D=-\tfrac{1}{2}(Z+R),$$

$$P_0=\tfrac{1}{2}(M+2T),\quad P_n=\tfrac{1}{2}(M-2T),\quad [D,G_a]=0,\quad [D,J_{0n}]=0.$$

The lemma is proved.

Lemma 12: Let $C\in O_1(2,n+1)$ and $W=\langle Q_1+Q_{n+3},Q_2+Q_{n+2}\rangle$. If $CW=W$ then the matrix C has the form (9) with $\varepsilon=1$ and $K\in SO(n-1)$.

Proof: From the conditions of Lemma 11 and the fact that we ask for $C\in O_1(2,n+1)$, it follows that $\mathrm{diag}[1,\varepsilon,K,\varepsilon,1]\in O_1(2,n+1)$. It follows now that $\varepsilon>0$ and that

$$\begin{vmatrix} K & 0\\ 0 & \varepsilon\end{vmatrix}>0$$

and thus we have $\varepsilon=1$ and $|K|>0$, whence $K\in SO(n-1)$. This proves the lemma.

The matrices of the form (9) with $\varepsilon=1$ and $K\in SO(n-1)$ form a group under multiplication, which we denote by $G_4(n-1)$ since its Lie algebra is the full Galilei algebra $AG_4(n-1)$. It is easy to see that $G_4(n-1)\subset O_1(2,n+1)$.

Lemma 13: If $C\in O_1(2,n+1)$ but $C\notin G_4(n-1)$, then $C=A_1\Gamma A_2$, where $A_1,A_2\in G_4(n-1)$ and Γ is one of the matrices

$$\Gamma_1=\mathrm{diag}[1,...,1,-1],\quad \Gamma_2=\mathrm{diag}[1,1,-1,1...,1,-1,1]. \tag{10}$$

Proof: Let

$$C(Q_1+Q_{n+3})=\sum_{i=1}^{n+3}\alpha_i Q_i,\quad \alpha_1^2+\alpha_2^2-\alpha_3^2-\cdots-\alpha_{n+3}^2=0.$$

There exists a matrix $\Lambda=\mathrm{diag}[1,1,\Delta,1,1]$ with $\Delta\in SO(n-1)$ such that $\Lambda C(Q_1+Q_{n+3})$ does not contain $Q_4,\ldots,Q_{n+1}$. Hence we may assume $\alpha_1^2+\alpha_2^2-\alpha_3^2-\alpha_{n+2}^2-\alpha_{n+3}^2=0$.

Since

$$S+T=\tfrac{1}{2}(K_0+P_0+K_n-P_n)=\Omega_{12}+\Omega_{n+2,n+3},$$

then, up to a factor $\exp[\theta(S+T)]$, we may suppose that $\alpha_1\neq 0$, $\alpha_2=0$. If $\alpha_1^2=\alpha_{n+3}^2$ then $\alpha_3=0$, $\alpha_{n+2}=0$. Assume $\alpha_1\neq\alpha_{n+3}$. As in the proof of Lemma 4, we find that

$$\exp(\beta_1P_1+\beta_2P_2)(\alpha_1Q_1+\alpha_3Q_3+\alpha_{n+2}Q_{n+2}+\alpha_{n+3}Q_{n+3})=\alpha_1'Q_1+\alpha_{n+3}'Q_{n+3},$$

where $\alpha_1'^2-\alpha_{n+3}'^2=0$. Thus there exists a matrix $A_1\in G_4(n-1)$ such that

$$A_1^{-1}C(Q_1+Q_{n+3})=\gamma(Q_1\pm Q_{n+3}),$$
$$A_1^{-1}C(Q_2+Q_{n+2})=\delta_1Q_1+\delta_2Q_2+\delta_3Q_3+\delta_4Q_{n+2}+\delta_5Q_{n+3}. \tag{11}$$

Since the pseudo-orthogonal transformations preserve the scalar product, it follows that the right-hand sides in (11) are also orthogonal, which implies that $\gamma(\delta_1\mp\delta_5)=0$ so that $\delta_5=\pm\delta_1$. If $\delta_2\neq\delta_4$ then multiplying the left- and right-hand sides in (11) by $\exp(\theta G_1)$ does not change the right-hand side of the first equality, and allows us to eliminate δ_3 by transforming it into 0. If $\delta_2=\delta_4$, then one easily deduces that $\delta_3=0$. Thus we may assume that $\delta_3=0$. But then we have $\delta_4=\pm\delta_2$ because $\delta_5=\pm\delta_1$ and $\delta_1^2+\delta_2^2-\delta_4^2-\delta_5^2=0$.

Let $W=\langle Q_1+Q_{n+3},Q_2+Q_{n+2}\rangle$. The above reasoning implies that for some matrix $A_1\in G_4(n-1)$ we have $\Gamma A_1^{-1}CW=W$ where Γ is one of the matrices (10). The fact that $\Gamma A_1^{-1}C\in O_1(2,n+1)$ implies, using Lemma 12, $\Gamma A_1^{-1}C=A_2\in G_4(n-1)$. Thus $C=A_1\Gamma A_2$ and the lemma is proved.

Lemma 14: The subalgebras L_1 and L_2 of $AG_4(n-1)$ are conjugate under $AdAC(1,n)$ if and only if they are conjugate under $AdAG(n-1)$ or if there exist automorphisms ψ_1,ψ_2 in $AdAG_4(n-1)$ with $\psi_1(L_1)=\Gamma[\psi_2(L_2)]\Gamma^{-1}$, where Γ is one of the matrices (10).

Proof: The result follows immediately from Lemma 13.

In the following table we give the action on the full Galilei algebra $AG_4(n-1)$ of the automorphisms $\varphi_{\Gamma_1},\varphi_{\Gamma_2},\varphi_{C_1},\varphi_{C_4},\varphi_{C_5}$ where

$$C_4=\exp\left(\frac{\pi}{2}(S+T)\right),\quad C_5=\exp(\pi(S+T))$$

(see (3) and (10) for the notation).

Theorem 3: Let L_1 and L_2 be subalgebras of $AG_4(n-1)$ which are not conjugate under $AdAG_4(n-1)$ with subalgebras of

$$\langle M,T,P_1,\ldots,P_{n-1}\rangle\uplus(AO(n-1)\otimes\langle D,J_{0n}\rangle)$$

and

$$AO(n-1)\oplus\langle S+T,Z\rangle.$$

Then the subalgebras L_1 and L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under $AdAG_4(n-1)$.

Proof: If the subalgebras L_1 and L_2 are conjugate under $AdAG_4(n-1)$ then they are conjugate under $AdAC(1,n)$. Now suppose that they are conjugate under $AdAC(1,n)$. In order to prove their conjugacy under $AdAG_4(n-1)$ it is sufficient (by Lemma 14) to show that for an arbitrary $\psi\in AdAG_4(n-1)$ and for each matrix Γ of the form (10), the subalgebra $\Gamma\psi(L_1)\Gamma^{-1}$ either equals $\psi(L_1)$ or is not contained in $AG_4(n-1)$, for then the only possibility is that they are conjugate under $AdAG_4(n-1)$.

If the projection of $\psi(L_1)$ onto $\langle G_1,\ldots,G_{n-1}\rangle$ is nonzero, then, using Table I, the subalgebra $\Gamma\psi(L_1)\Gamma^{-1}$ contains an element Y whose projection for some a, $1\leqslant a\leqslant n-1$ onto $\langle J_{0a},J_{an}\rangle$ is of the form $\lambda(J_{0a}+J_{an})$ with $\lambda\neq0$. If $\Gamma\psi(L_1)\Gamma^{-1}\subset AG_4(n-1)$, then the projection of Y onto $\langle J_{0a},J_{an}\rangle$ would have the form $\mu(J_{0a}-J_{an})$, which would imply $\lambda=\mu=-\mu=0$, an obvious contradiction.

Now let the projection of $\psi(L_1)$ onto $\langle G_1,\ldots,G_{n-1}\rangle$ be zero. Denote by $\tau\psi(L_1)$ the projection of $\psi(L_1)$ onto $\langle R,S,T\rangle$ If $\tau\psi(L_1)=\langle R,S,T\rangle$, then $\langle R,S,T\rangle\subset\psi(L_1)$. From this it follows that $\Gamma_2\psi(L_1)\Gamma_2^{-1}$ is not a subset of $AG_4(n-1)$. If we assume that $\Gamma_1\psi(L_1)\Gamma_1^{-1}\subset AG_4(n-1)$, we obtain, from Table I, that the projection of $\psi(L_1)$ onto $\langle P_1,\ldots,P_n,M\rangle$ is zero, and consequently we have either $\psi(L_1)=\langle R,S,T\rangle$ or $\psi(L_1)=\langle R,S,T\rangle\oplus\langle Z\rangle$. In this case, $\Gamma_1\psi(L_1)\Gamma_1^{-1}=\psi(L_1)$. If $\tau\psi(L_1)=\langle R+\alpha S,T+\beta S\rangle$, with $\alpha\neq0$, then $\Gamma_2\psi(L_1)\Gamma_2^{-1}$ is not contained in $AG_4(n-1)$. If we had $\Gamma_1\psi(L_1)\Gamma_1^{-1}\subset AG_4(n-1)$, then the projection of $\psi(L_1)$ onto $\langle P_1,\ldots,P_n,M\rangle$ would be zero. But then $\psi(L_1)$ would be conjugate under $AdAG_4(n-1)$ with a subalgebra of $AO(n-1)\oplus\langle R,T,Z\rangle$, which contradicts the assumptions of the theorem. The theorem is proved.

Theorem 4: Let L_1 and L_2 be subalgebras of the algebra

TABLE I. Action of automorphisms on elements of $AG_4(n-1)$ for $n \geqslant 2$.

Element of $AG_4(n-1)$	φ_{Γ_1}	φ_{Γ_2}	φ_{C_1}	φ_{C_4}	φ_{C_5}	Restrictions
P_1	K_1	$-P_1$	$-P_1$	$-G_1$	$-P_1$	
P_a	K_a	P_a	$-P_a$	$-G_a$	$-P_a$	$a=2,...,n-1$
M	K_0-K_n	$2T$	$-M$	M	M	
G_1	$J_{01}+J_{1n}$	$-(J_{01}+J_{1n})$	G_1	P_1	$-G_1$	
G_a	$J_{0a}+J_{an}$	$J_{0a}+J_{an}$	G_a	P_a	$-G_a$	$a=2,...,n-1$
J_{1a}	J_{1a}	$-J_{1a}$	J_{1a}	J_{1a}	J_{1a}	$a=2,...,n-1$
J_{ab}	J_{ab}	J_{ab}	J_{ab}	J_{ab}	J_{ab}	$a,b=2,...,n-1$
R	$-R$	Z	R	$-R$	R	
S	T	$\frac{1}{2}(K_0-K_n)$	$-S$	T	S	
T	S	$\frac{1}{2}M$	$-T$	S	T	
Z	$-Z$	R	Z	Z	Z	

$$L=\langle M,T,P_1,\ldots,P_{n-1}\rangle \uplus (AO(n-1)\oplus\langle D,J_{0n}\rangle)$$

having nonzero projection on $\langle J_{0n}\rangle$ and $\langle D\rangle$ and are not conjugate under AdL with subalgebras of the algebra $\langle M,T\rangle \uplus (AO(n-1)\oplus\langle D,J_{0n}\rangle)$. Then L_1 and L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under AdL or if there exists an automorphism $\psi \in AdL$ such that $\psi(L_1)=\Lambda L_2 \Lambda^{-1}$ where Λ is one of the matrices $\Gamma_2, C_5, \Gamma_2 C_5$ (see Table I).

Proof: If $\psi \in AdAG_4(n-1)$, then $\psi=\varphi_C$ where C is a matrix of the form (9). By theorem IV.3.4 of Ref. 9, the subalgebra L_1 is, up to an automorphism of $AdAG_4(n-1)$, one of the following algebras:

(1) $(U_1+U_2+U_3)\uplus F$, where $U_1 \subset \langle M\rangle$, $U_2 \subset \langle T\rangle$, $U_3 \subset \langle P_1,\ldots,P_{n-1}\rangle$ and $F \subset AO(n-1)\oplus\langle D,J_{0n}\rangle$;
(2) $(U_1+U_2)\uplus F$, where $U_1 \subset \langle T\rangle$, $U_2 \subset \langle P_1,\ldots,P_{n-1}\rangle$ and F is a subalgebra of $AO(n-1)\oplus\langle R,M\rangle$;
(3) $(U_1+U_2)\uplus F$, where $U_1 \subset \langle M\rangle$, $U_2 \subset \langle P_1,\ldots,P_{n-1}\rangle$ and F is a subalgebra of $AO(n-1)\oplus\langle Z,T\rangle$.

By assumption, the projection of L_1 onto $\langle P_1,\ldots,P_{n-1}\rangle$ is nonzero.

If $\psi(L_1)=L_2$, then in formula (9) $\theta=0$ or $\theta=\pi$ because for other values of θ the projection of $\psi(L_1)$ onto $\langle G_1,\ldots,G_{n-1}\rangle$ is nonzero. For this reason, $\gamma_1=\cdots=\gamma_{n-1}=0$ and so $\psi \in AdL$ or $\varphi_{C_5}\psi \in AdL$. Let there be automorphisms $\psi_1,\psi_2 \in AdAG_4(n-1)$ with $\Gamma\psi_1(L_1)\Gamma=\psi_2(L_2)$ where Γ is one of the matrices (10). If AdL did not contain ψ_1 and $\varphi_{C_5}\psi_1$, then the projection of $\psi_1(L_1)$ on $\langle G_1,\ldots,G_{n-1}\rangle$ would be nonzero, and so, by Table I, $\psi_2(L_2)$ would not be in $AG_4(n-1)$. Thus ψ_j or $\varphi_{C_5}\psi_j$ belongs to AdL for each $j=1,2$. For $\Gamma=\Gamma_1$ the projection of $\Gamma\psi_1(L_1)\Gamma$ onto $\langle K_1,\ldots,K_{n-1}\rangle$ is nonzero, so we have $\Gamma=\Gamma_2$. In this case $\Gamma\psi_2(L_2)\Gamma=\psi_2'(\Gamma L_2\Gamma)$. Using Lemma 14, the theorem is proved.

In a similar way, one proves the following results.

Theorem 5: Let B be a subalgebra of the algebra

$$N=\langle M,P_1,\ldots,P_{n-1}\rangle \uplus (AO(n-1)\oplus\langle D,T\rangle$$

and let B have nonzero projection onto $\langle D\rangle$. Then B is conjugate under $AdAC(1,n)$ to the algebra

$$F=(W_1\oplus W_2)\uplus E, \tag{12}$$

where E is a subalgebra of the algebra $AO(n-1)\oplus\langle D\rangle$, $W_1 \subset \langle P_1,\ldots,P_{n-1}\rangle$ and W_2 is one of the algebras $0, \langle P_0\rangle, \langle P_n\rangle, \langle M\rangle, \langle P_0,P_n\rangle$. If $W_2=\langle P_n\rangle$ or $W_2=\langle P_0,P_n\rangle$ then the subalgebra $W_1\uplus E$ is not conjugate under $AdAO(n-1)$ with any subalgebra of $\langle P_1,\ldots,P_{n-2}\rangle \uplus (AO(n-2)\oplus\langle D\rangle)$. Subalgebras F_1, F_2 of the type (12) of the algebra N with nonzero projection onto $\langle D\rangle$, which are not conjugate under AdN to subalgebras of $\langle M,T\rangle \uplus (AO(n-1)\oplus\langle D\rangle)$, will be conjugate under

$AC(1,n)$ if and only if they are conjugate under AdL or when there exists an automorphism $\psi \in AdL$ with $\psi(F_1)=\Gamma_2 F_2 \Gamma_2^{-1}$ (see (10)), where $L=AO(n-1)$ (we consider $AdAO(n-1)$ to be a subgroup of $AdAC(1,n)$).

Theorem 6: Let B be a subalgebra of the algebra

$$N=\langle M,P_1,\ldots,P_{n-1}\rangle \uplus (AO(n-1)\oplus\langle J_{0n},T\rangle$$

and let B have nonzero projection onto $\langle J_{0n}\rangle$. Then B is conjugate under $AdAC(1,n)$ with the algebra

$$F=W\uplus E, \tag{13}$$

where E is a subalgebra of the algebra $\langle P_1,\ldots,P_{n-1}\rangle \uplus (AO(n-1)\oplus\langle J_{0n}\rangle)$ and W is one of the algebras 0, $\langle M\rangle$, $\langle P_0,P_n\rangle$. Let $L=N\uplus\langle D\rangle$. Subalgebras F_1,F_2 of the type (13) of the algebra N which are not conjugate under AdN with subalgebras of the algebra $\langle M\rangle\uplus(AO(n-1)\oplus\langle J_{0n},T\rangle$, will be conjugate under $AdAC(1,n)$ if and only if they are conjugate under AdL or if there exists an automorphism $\psi\in AdL$ with $\psi(F_1)=\Lambda F_2\Lambda^{-1}$ where Λ is one of the matrices $\Gamma_2, C_5, \Gamma_2 C_5$ (see Table I).

Theorem 7: Let L_1,L_2 be subalgebras of the algebra $L=\langle M,S+T,Z\rangle\oplus AO(n-1)$ which have nonzero projection onto $\langle S+T\rangle$. The algebras L_1 and L_2 are conjugate under $AdAC(1,n)$ if and only if they are conjugate under AdL or if there exists an automorphism $\psi\in AdL$ such that $\psi(L_1)=\Gamma_1 L_2 \Gamma_1^{-1}$ (see Table I).

VI. SUBALGEBRAS OF $AC(1,3)$

We recall that in this article the conformal algebra $AC(1,3)$ is realized as the pseudo-orthogonal algebra $AO(2,4)$. It turns out that it is convenient to divide the subalgebras of $AO(2,4)$ into seven classes:

(1) subalgebras not having invariant isotropic subspaces in $\mathbf{R}_{2,4}$;
(2) subalgebras conjugate to subalgebras of $AG_1(2)$;
(3) subalgebras conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03}\rangle$ and having nonzero projection onto $\langle J_{03}\rangle$;
(4) subalgebras conjugate to subalgebras of $AP(1,3)$ but not conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03}\rangle$;
(5) subalgebras conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03},D\rangle$ but not conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03}\rangle$;
(6) subalgebras conjugate to subalgebras of $A\tilde{P}(1,3)$ but not conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03},D\rangle$;
(7) subalgebras conjugate to subalgebras of $AG_4(2)$ but not conjugate to subalgebras of $A\tilde{P}(1,3)$.

Since subalgebras conjugate under $AdAC(1,3)$ are identified, we omit mentioning conjugacy when referring to classes. So, for instance, we shall consider the second class as consisting of subalgebras of $AG_1(2)$. In order to have a better survey of subalgebras it is convenient to split the classes into subclasses corresponding to certain properties of the projections of the subalgebras of a class onto the homogeneous part of the algebra.

The division of the set of subalgebras of $AC(1,3)$ into the classes (1)–(7) allows us easily to construct the set of subalgebras of each of the algebras $AG_1(2), AP(1,3), A\tilde{P}(1,3), AG_4(2)$. Up to conjugacy under $AdAC(1,3)$ we have

(a) the set of subalgebras of $AG_1(2)$ coincides with class (2);
(b) the set of subalgebras of $AP(1,3)$ is the union of classes (2), (3) and (4);
(c) the set of subalgebras of $A\tilde{P}(1,3)$ coincides with the union of classes (2)–(6);
(d) the set of subalgebras of $AG_4(2)$ is the union of classes (2), (3), (5), and (7).

We use the notation $F:U_1,\ldots,U_m$ for $U_1\uplus F,\ldots,U_m\uplus F$.

A. Subalgebras not possessing invariant isotropic subspaces in $\mathbf{R}_{2,4}$

This class is divided into subclasses by the existence for the subalgebras of invariant irreducible subspaces of a particular kind in the space $\mathbf{R}_{2,4}$.

1. Irreducible subalgebras of $AO(2,4)$

$$AC(1,3);$$

$$ASU(1,2)=\langle P_0+K_0+2J_{12},P_0+K_0+K_3-P_3,P_1+K_1+2J_{02},P_3+K_3+K_0-P_0,K_2-P_2+2J_{13},P_2+K_2-2J_{01},D+J_{03},K_1-P_1-2J_{23}\rangle;$$

$$ASU'(1,2)=\langle P_0+K_0-2J_{12},P_0+K_0+K_3-P_3,P_1+K_1-2J_{02},P_3+K_3+K_0-P_0,K_2-P_2-2J_{13},P_2+K_2+2J_{01},D+J_{03},K_1-P_1+2J_{23}\rangle;$$

$$ASU(1,2)\oplus\langle P_0+K_0-2J_{12}-K_3+P_3\rangle;$$

$$ASU'(1,2)\oplus\langle P_0+K_0+2J_{12}-K_3+P_3\rangle;$$

$$\langle P_0+K_0-2J_{12}-K_3+P_3\rangle\oplus\langle P_1+K_1+2J_{02},P_3+K_3+K_0-P_0,K_2-P_2+2J_{13}\rangle;$$

$$\langle P_0+K_0+2J_{12}-K_3+P_3\rangle\oplus\langle P_1+K_1-2J_{02},P_3+K_3+K_0-P_0,K_2-P_2-2J_{13}\rangle.$$

2. Irreducible subalgebras of $AO(1,4):AC(3)$

3. Irreducible subalgebras of $AO(2,3)$

$$AC(1,2);$$

$$\langle P_2+K_2+\sqrt{3}(P_1+K_1)+K_0-P_0,D+J_{02}-\sqrt{3}J_{01},P_0+K_0-2(K_2-P_2)\rangle;$$

$$\langle P_2+K_2-\sqrt{3}(P_1+K_1)+K_0-P_0,D+J_{02}+\sqrt{3}J_{01},P_0+K_0-2(K_2-P_2)\rangle.$$

4. Subalgebras of $AO(2,2)\oplus AO(2)$ with irreducible projection onto $AO(2,2)$

$$\langle J_{01}-D,K_0-P_0-P_1-K_1,P_0+K_0-K_1+P_1\rangle\oplus\langle P_0+K_0+K_1-P_1\rangle\oplus F,$$

where $F=0$ or $F=\langle J_{23}\rangle$;

$$\langle J_{01}+D,K_0-P_0+P_1+K_1,P_0+K_0+K_1-P_1\rangle\oplus\langle P_0+K_0-K_1+P_1\rangle\oplus F,$$

where $F=0$ or $F=\langle J_{23}\rangle$;

$$AC(1,1),AC(1,1)\oplus\langle J_{23}\rangle,\quad \text{where } AC(1,1)=\langle P_0,P_1,K_0,K_1,J_{01},D\rangle;$$

$$\langle J_{01}-D,K_0-P_0-P_1-K_1,P_0+K_0-K_1+P_1\rangle\oplus\langle P_0+K_0+K_1-P_1+\alpha J_{23}\rangle(\alpha\neq 0);$$

$$\langle J_{01}+D,K_0-P_0+P_1+K_1,P_0+K_0+K_1-P_1\rangle\oplus\langle P_0+K_0-K_1+P_1+\alpha J_{23}\rangle(a\neq 0).$$

5. Subalgebras of the type $AO(2,1)\oplus F$ with $F\subset AO(3)$

$AC(1)\oplus L$, where $AC(1)=\langle D,P_0,K_0\rangle$, and L is one of the algebras: $0,\langle J_{12}\rangle,\langle J_{12},J_{13},J_{23}\rangle$.

6. Subalgebras of $AO(2)\oplus AO(4)$ having an irreducible projection

$$\langle P_0+K_0\rangle;\ \langle P_0+K_0\rangle\oplus\langle 2J_{12}+\alpha(K_3-P_3)\rangle(|\alpha|\leqslant 1);$$

$$\langle P_0+K_0\rangle\oplus\langle J_{12},K_3-P_3\rangle;\ \langle P_0+K_0\rangle\oplus\langle J_{12},J_{13},J_{23};$$

$$\langle P_0+K_0\rangle\oplus\langle 2J_{12}+\varepsilon(K_3-P_3),2J_{13}-\varepsilon(K_2-P_2),2J_{23}+\varepsilon(K_1-P_1)(\varepsilon=\pm 1)\rangle;$$

$$\langle P_0+K_0\rangle\oplus\langle 2J_{12}+\varepsilon(K_3-P_3),2J_{13}-\varepsilon(K_2-P_2),2J_{23}+\varepsilon(K_1-P_1)\rangle$$

$$\oplus\langle 2J_{12}-\varepsilon(K_3-P_3)\rangle(\varepsilon=\pm 1);$$

$$\langle P_0+K_0\rangle\oplus\langle K_1-P_1,K_2-P_2,K_3-P_3,J_{12},J_{13},J_{23}\rangle;$$

$$\langle P_0+K_0+2\alpha J_{12}\rangle(\alpha\neq 0,|\alpha|\neq 1);$$

$$\langle P_0+K_0+2\alpha J_{12}+\beta(K_3-P_3)\rangle(\alpha\neq 0,|\alpha|\neq 1,\beta\geqslant\alpha,\beta\neq 1);$$

$$\langle 2J_{12}+\alpha(P_0+K_0),K_3-P_3+\beta(P_0+K_0)\rangle(\alpha\neq 0,\beta\geqslant 0,\ \text{with}\ |\alpha|\neq 1\ \text{when}\ \beta=0);$$

$$\langle\alpha(P_0+K_0)+2\varepsilon J_{12}-K_3+P_3\rangle\oplus\langle 2\varepsilon J_{12}+K_3-P_3,2\varepsilon J_{13}-K_2+P_2,2\varepsilon J_{23}+K_1-P_1\rangle(\alpha\geqslant 0);$$

$$\langle 2\varepsilon J_{12}+K_3-P_3,2\varepsilon J_{13}-K_2+P_2,2\varepsilon J_{23}+K_1-P_1\rangle(\varepsilon=\pm 1);$$

$$\langle 2\varepsilon J_{12}+K_3-P_3,2\varepsilon J_{13}-K_2+P_2,2\varepsilon J_{23}+K_1-P_1\rangle\oplus\langle 2\varepsilon J_{12}-K_3+P_3\rangle(\varepsilon=\pm 1);$$

$$\langle K_1-P_1,K_2-P_2,K_3-P_3,J_{12},J_{13},J_{23}\rangle.$$

7. Subalgebras of $AO(1,2)\oplus AO(1,2)$

$$\langle P_1+K_1,P_2+K_2,J_{12}\rangle\oplus\langle K_0-P_0,K_3-P_3,J_{03}\rangle;$$

$$\langle P_1+K_1+2\varepsilon J_{03},P_2+K_2+K_0-P_0,2\varepsilon J_{12}+K_3-P_3\rangle(\varepsilon=\pm 1);$$

$$\langle P_1+K_1,P_2+K_2,J_{12}\rangle\oplus\langle K_3-P_3\rangle.$$

B. Subalgebras of $AG_1(2)$

The classical Galilei algebra $AG_1(2)$ is the semidirect sum of a solvable ideal, generated by $\langle P_1,P_2,M,T\rangle$, and the Euclidean algebra $AE(2)=\langle G_1,G_2,J_{12}\rangle$. The projection of $AG_1(2)$ onto $AO(1,3)$ coincides with $AE(2)$, which has, up to inner automorphisms, the subalgebras $0,\langle J_{12}\rangle,\langle G_1\rangle,\langle G_1,G_2\rangle,\langle G_1,G_2,J_{12}\rangle$. The first two subalgebras are completely reducible algebras of linear transformations of Minkowski space $\mathbf{R}_{1,3}$, whereas the others are not of this type. Thus we divide this class into two subclasses *A* and *B*.

1. Subalgebras with completely reducible projection onto $AO(1,3)$

$$0,\langle P_0\rangle,\langle P_1\rangle,\langle M\rangle,\langle P_0,P_3\rangle,\langle M,P_1\rangle,\langle P_1,P_2\rangle,\langle M,P_1,P_2\rangle,\langle P_0,P_1,P_2\rangle,$$

$$\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$\langle J_{12}\rangle:0,\langle P_0\rangle,\langle P_3\rangle,\langle M\rangle,\langle P_0,P_3\rangle,\langle P_1,P_2\rangle,\langle P_0,P_1,P_2\rangle,\langle M,P_1,P_2\rangle,$$

$$\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$\langle J_{12}+P_0\rangle:0,\langle P_3\rangle,\langle P_1,P_2\rangle,\langle P_1,P_2,P_3\rangle;$$

$$\langle J_{12}\pm P_3\rangle:0,\langle P_0\rangle,\langle P_1,P_2\rangle,\langle P_0,P_1,P_2\rangle;$$

$$\langle J_{12}\pm 2T\rangle:0,\langle M\rangle,\langle P_1,P_2\rangle,\langle M,P_1,P_2\rangle.$$

2. Subalgebras whose projection onto $AO(1,3)$ is not completely reducible

$$\langle G_1\rangle:\langle P_2\rangle,\langle M,P_1\rangle,\langle M,P_2\rangle,\langle M,P_1+\alpha P_2\rangle,\langle M,P_1,P_2\rangle,\langle P_0,P_1,P_3\rangle,$$

$$\langle P_0,P_1,P_2,P_3\rangle(\alpha\neq 0);$$

$$\langle G_1 \pm P_2\rangle : 0, \langle M\rangle, \langle M, P_1\rangle, \langle P_0, P_1, P_3\rangle;$$

$$\langle G_1 + 2T\rangle : 0, \langle P_2\rangle, \langle M\rangle, \langle M, P_1\rangle, \langle M, P_2\rangle, \langle M, P_1 + \alpha P_2\rangle (\alpha \neq 0), \langle M, P_1, P_2\rangle;$$

$$\langle G_1, G_2\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1 + \varepsilon P_2, G_2 - \varepsilon P_1, M\rangle, \langle G_1 + \varepsilon P_2, G_2 - \varepsilon P_1 + \alpha P_2, M\rangle (\varepsilon = \pm 1, \alpha \neq 0);$$

$$\langle G_1 + \alpha P_2, G_2 + 2T, M, P_1\rangle (\alpha \in \mathbf{R});$$

$$\langle G_1 \pm P_2, G_2, M, P_1\rangle, \langle G_1, G_2 + 2T, M, P_1, P_2\rangle;$$

$$\langle G_1, G_2, J_{12}\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, G_2, J_{12} \pm 2T, M, P_1, P_2\rangle, \langle G_1 + \varepsilon P_2, G_2 - \varepsilon P_1, J_{12}, M\rangle (\varepsilon = \pm 1).$$

C. Subalgebras of $AG_1(2) \uplus \langle J_{03}\rangle$ with nonzero projection onto $\langle J_{03}\rangle$

We divide also this class into two subclasses which are distinguished by whether or not they have a completely reducible projection onto $AO(1,3)$.

1. Subalgebras with completely reducible projection onto $AO(1,3)$

$$\langle J_{03}\rangle : 0, \langle P_1\rangle, \langle M\rangle, \langle P_0, P_3\rangle, \langle M, P_1\rangle,$$
$$\langle P_1, P_2\rangle, \langle P_0, P_1, P_3\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle J_{03} + P_1\rangle : 0, \langle P_2\rangle, \langle M\rangle, \langle P_0, P_3\rangle, \langle M, P_2\rangle, \langle P_0, P_2, P_3\rangle;$$

$$\langle J_{12} + \alpha J_{03}\rangle : 0, \langle M\rangle, \langle P_0, P_3\rangle, \langle P_1, P_2\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle (\alpha \neq 0);$$

$$\langle J_{12}, J_{03}\rangle : 0, \langle M\rangle, \langle P_0, P_3\rangle, \langle P_1, P_2\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle.$$

2. Subalgebras with projections onto $AO(1,3)$ which are not completely reducible

$$\langle G_1, J_{03}\rangle : 0, \langle M\rangle, \langle P_2\rangle, \langle M, P_1\rangle, \langle M, P_2\rangle, \langle M, P_1 + \alpha P_2\rangle (\alpha \neq 0),$$
$$\langle M, P_1, P_2\rangle, \langle P_0, P_1, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, J_{03} + P_2\rangle : 0, \langle M\rangle, \langle M, P_1\rangle, \langle M, P_1 + \alpha P_2\rangle (\alpha \neq 0), \langle P_0, P_1, P_3\rangle;$$

$$\langle G_1, J_{03} + P_1\rangle : \langle M\rangle, \langle M, P_2\rangle;$$

$$\langle G_1, J_{03} + P_1 + \alpha P_2, M\rangle (\alpha \neq 0);$$

$$\langle G_1, G_2, J_{03}\rangle : 0, \langle M\rangle, \langle M, P_1\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, G_2, J_{03} + P_1, M\rangle, \langle G_1, G_2, J_{03} + P_2, M, P_1\rangle;$$

$$\langle G_1, G_2, J_{12} + \alpha J_{03}\rangle : 0, \langle M\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle (\alpha \neq 0);$$

$$\langle G_1, G_2, J_{12}, J_{03}\rangle : 0, \langle M\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle.$$

D. Subalgebras of $AP(1,3)$ which are not conjugate to subalgebras of $AG_1(2) \uplus \langle J_{03}\rangle$

This class consists of those subalgebras of the Poincaré algebra $AP(1,3)$ whose projection onto $AO(1,3)$ do not possess isotropic invariant subspaces in $\mathbf{R}_{1,3}$. Since the projections are simple algebras, then each subalgebra of the fourth class splits. The full list of such algebras is

$$AO(1,2):0,\langle P_3\rangle,\langle P_0,P_1,P_2\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$AO(3):0,\langle P_0\rangle,\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$AO(1,3):0,\langle P_0,P_1,P_2,P_3\rangle.$$

E. Subalgebras of $AG_1(2)\uplus\langle J_{03},D\rangle$ which are not conjugate to subalgebras of $AG_1(2)\uplus\langle J_{03}\rangle$

Let K be a subalgebra of $AG_1(2)\uplus\langle J_{03},D\rangle$ with nonzero projection onto $\langle D\rangle$, and let $\hat{\theta}$ be the projection of K onto $\langle J_{03},D\rangle$. By Propositions IV.2.3 and IV.2.5 in Ref. 9, the algebra K, as a subalgebra of $A\tilde{P}(1,3)$, is split whenever $\hat{\theta}(K)$ is one of the subalgebras 1) $\langle D\rangle$; 2) $\langle \gamma D - J_{03}\rangle$ $(\gamma\neq\pm 1,0,2)$; 3) $\langle D,J_{03}\rangle$. This leads us to dividing this class of subalgebras into two subclasses of nonsplittable subalgebras K of $A\tilde{P}(1,3)$, denoted by D and E, for which the projection onto $\langle G_1,G_2\rangle$ is non-zero, and for which $\hat{\theta}(K)$ is $\langle J_{03}\mp D\rangle$ and $\langle J_{03}-2D\rangle$ respectively. It is also useful to distinguish the subclass A of subalgebras having zero projection onto $\langle G_1,G_2\rangle$. The subalgebras in this subclass differ from the other subalgebras in that their projections onto $AO(1,3)$ are completely reducible algebras of linear transformations of Minkowski space $\mathbf{R}_{1,3}$. All the other subalgebras are split, and we divide them formally into subclasses B and C, depending on the dimension of their projection onto $\langle D,J_{03}\rangle$.

1. Subalgebras with zero projection on $\langle G_1,G_2\rangle$

$$\langle D\rangle:\langle P_0\rangle,\langle P_0,P_3\rangle,\langle P_0,P_1,P_2\rangle,\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$\langle J_{12}+\alpha D\rangle:\langle P_0\rangle,\langle P_3\rangle,\langle P_0,P_3\rangle,\langle P_0,P_1,P_2\rangle,\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle(\alpha>0);$$

$$\langle J_{12},D\rangle:\langle P_0\rangle,\langle P_3\rangle,\langle P_0,P_3\rangle,\langle P_0,P_1,P_2\rangle,\langle P_1,P_2,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$\langle J_{03}+\alpha D\rangle(0<\alpha\leqslant 1);$$

$$\langle J_{03}+\alpha D,M\rangle(0<|\alpha|\leqslant 1);$$

$$\langle J_{03}+\alpha D\rangle:\langle P_1\rangle,\langle P_0,P_3\rangle,\langle P_1,P_2\rangle,\langle P_0,P_1,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle(\alpha>0);$$

$$\langle J_{03}+\alpha D\rangle:\langle M,P_1\rangle,\langle M,P_1,P_2\rangle(\alpha\neq 0);$$

$$\langle J_{03}-D\pm 2T\rangle:0,\langle P_1\rangle,\langle M\rangle,\langle P_1,P_2\rangle,\langle M,P_1\rangle,\langle M,P_1,P_2\rangle;$$

$$\langle J_{03},D\rangle:0,\langle P_1\rangle,\langle M\rangle,\langle P_0,P_3\rangle,\langle P_1,P_2\rangle,\langle M,P_1\rangle,\langle M,P_1,P_2\rangle,\langle P_0,P_1,P_3\rangle,\langle P_0,P_1,P_2,P_3\rangle;$$

$$\langle \epsilon J_{12}+\alpha J_{03}+\beta D\rangle(0<\alpha\leqslant\beta,\epsilon=\pm 1);$$

$$\langle J_{12}+\alpha J_{03}+\beta D,M\rangle(0<|\alpha|\leqslant|\beta|;$$

$$\langle \epsilon J_{12}+\alpha J_{03}+\beta D\rangle:\langle P_0,P_3\rangle,\langle P_1,P_2\rangle,\langle P_0,p_1,P_2,P_3\rangle(\epsilon=\pm 1,\alpha,\beta>0);$$

$$\langle J_{12}+\alpha J_{03}+\beta D,M,P_1,P_2\rangle(\alpha\neq 0,\beta\neq 0);$$

$$\langle J_{12}+\alpha(J_{03}-D\pm 2T)\rangle:0,\langle M\rangle,\langle P_1,P_2\rangle,\langle M,P_1,P_2\rangle(\alpha\neq 0);$$

$$\langle J_{12}+\alpha J_{03},D\rangle:0,\langle M\rangle,\langle P_1,P_2\rangle,\langle P_0,P_3\rangle,\langle M,P_1,P_2\rangle,\langle P_0,P_1,P_2,P_3\rangle(\alpha\neq 0);$$

$$\langle J_{03}+\alpha D,J_{12}+\beta D\rangle:\langle P_0,P_3\rangle,\langle P_1,P_2\rangle,\langle P_0,P_1,P_2,P_3\rangle(\alpha^2+\beta^2\neq 0);$$

$$\langle J_{03}+\alpha D,J_{12}+\beta D\rangle(|\alpha|\leqslant 1,\beta\geqslant 0,|\alpha|+\beta\neq 0);$$

$$\langle J_{03}+\alpha D,J_{12}+\beta D,M\rangle(|\alpha|\leqslant 1,\beta\geqslant 0,|\alpha|+\beta\neq 0);$$

$$\langle J_{03}+\alpha D, J_{12}+\beta D, M, P_1, P_2\rangle (\alpha,\beta\in\mathbf{R}, \alpha^2+\beta^2\neq 0);$$

$$\langle J_{03}-D\pm 2T, J_{12}+2\alpha T\rangle : 0, \langle M\rangle, \langle P_1, P_2\rangle, \langle M, P_1, P_2\rangle;$$

$$\langle J_{03}-D, J_{12}\pm T\rangle : 0, \langle M\rangle, \langle P_1, P_2\rangle, \langle M, P_1, P_2\rangle;$$

$$\langle J_{03}, J_{12}, D\rangle : 0, \langle M\rangle, \langle P_0, P_3\rangle, \langle P_1, P_2\rangle, \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle.$$

2. Subalgebras with two-dimensional projection onto $\langle J_{03}, D\rangle$ and nonzero projection onto $\langle G_1, G_2\rangle$

$$\langle G_1, J_{03}, D\rangle : \langle P_2\rangle, \langle M, P_1\rangle, \langle M, P_2\rangle, \langle M, P_1+\alpha P_2\rangle, \langle M, P_1, P_2\rangle,$$

$$\langle P_0, P_1, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle (\alpha\neq 0);$$

$$\langle G_1, G_2, J_{03}, D\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, G_2, J_{12}+\alpha J_{03}, D\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle (\alpha\neq 0);$$

$$\langle G_1, G_2, J_{03}+\alpha D, J_{12}+\beta D, M, P_1, P_2\rangle (|\alpha|\leqslant 1, \beta\geqslant 0, |\alpha|+\beta\neq 0);$$

$$\langle G_1, G_2, J_{03}+\alpha D, J_{12}+\beta D, P_0, P_1, P_2, P_3\rangle (\alpha^2+\beta^2\neq 0);$$

$$\langle G_1, G_2, J_{03}, J_{12}, D\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle.$$

3. Split subalgebras with one-dimensional projection onto $\langle J_{03}, D\rangle$ and nonzero projection onto $\langle G_1, G_2\rangle$

$$\langle G_1+D\rangle : \langle P_0, P_1, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, D\rangle : \langle P_0, P_1, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1+D, G_2, P_0, P_1, P_2, P_3\rangle \langle G_1, G_2, D, P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, J_{03}+\alpha D\rangle : \langle P_2\rangle, \langle M, P_1\rangle, \langle M, P_2\rangle, \langle M, P_1+\beta P_2\rangle (|\alpha|\leqslant 1, \alpha\neq 0, \beta\neq 0);$$

$$\langle G_1, J_{03}+\alpha D\rangle : \langle M, P_1, P_2\rangle, \langle P_0, P_1, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle (\alpha\neq 0);$$

$$\langle G_1, G_2, J_{03}+\alpha D, M, P_1, P_2\rangle (0<|\alpha|\leqslant 1);$$

$$\langle G_1, G_2, J_{03}+\alpha D, P_0, P_1, P_2, P_3\rangle (\alpha\neq 0);$$

$$\langle G_1, G_2, J_{12}+\alpha D, P_0, P_1, P_2, P_3\rangle (\alpha\neq 0);$$

$$\langle G_1, G_2, J_{12}, D, P_0, P_1, P_2, P_3\rangle;$$

$$\langle G_1, G_2, J_{12}+\alpha J_{03}+\beta D, M, P_1, P_2\rangle (0<|\alpha|\leqslant|\beta|);$$

$$\langle G_1, G_2, J_{12}+\alpha J_{03}+\beta D, P_0, P_1, P_2, P_3\rangle (\beta\neq 0).$$

4. Nonsplit subalgebras of $AG_1(2)\uplus\langle J_{03}\mp D\rangle$ with nonzero projection onto $\langle G_1, G_2\rangle$ and $\langle J_{03}\mp D\rangle$

$$\langle J_{03}-D, G_1\pm P_2\rangle : 0, \langle M\rangle, \langle M, P_1\rangle, \langle P_0, P_1, P_3\rangle;$$

$$\langle J_{03}-D\pm 2T, G_1+\alpha P_2, M, P_1\rangle;$$

$$\langle J_{03}-D\pm 2T, G_1, M, P_1, P_2\rangle, \langle J_{03}+D+M, G_1, P_2\rangle;$$

$$\langle J_{03}-D, G_1+\epsilon P_2, G_2-\epsilon P_1+\alpha P_2, M\rangle (\epsilon=\pm 1, \alpha\in\mathbf{R});$$

$$\langle J_{03}-D, G_1\pm P_2, G_2, M, P_1\rangle \langle J_{03}-D\pm 2T, G_1, G_2, P_1, P_2, M\rangle;$$

$$\langle J_{12}+\alpha(J_{03}-D), G_1+\epsilon P_2, G_2-\epsilon P_1, M\rangle (\epsilon=\pm 1, \alpha\neq 0);$$

$$\langle J_{12}+\alpha(J_{03}-D\pm 2T), G_1, G_2, M, P_1, P_2\rangle (\alpha\neq 0);$$

$$\langle J_{12}\pm 2T, J_{03}-D, G_1, G_2, M, P_1, P_2\rangle;$$

$$\langle J_{12}+2\alpha T, J_{03}-D\pm 2T, G_1, G_2, M, P_1, P_2\rangle (\alpha\in\mathbf{R});$$

$$\langle J_{12}, J_{03}-D, G_1+\epsilon P_2, G_2-\epsilon P_1, M\rangle (\epsilon=\pm 1).$$

5. Nonsplit subalgebras of $AG_1(2)\uplus\langle J_{03}-2D\rangle$ with nonzero projection onto $\langle G_1, G_2\rangle$ and $\langle J_{03}-2D\rangle$

$$\langle J_{03}-2D, G_1+2T\rangle : 0, \langle M\rangle, \langle P_2\rangle, \langle M, P_1\rangle, \langle M, P_2\rangle, \langle M, P_1+\alpha P_2\rangle, \langle M, P_1, P_2\rangle (\alpha\neq 0);$$

$$\langle J_{03}-2D, G_1, G_2+2T\rangle : \langle M, P_1\rangle, \langle M, P_1, P_2\rangle.$$

F. Subalgebras of $A\tilde{P}(1,3)$ not conjugate to subalgebras of $AP(1,3)$ and of $AG_1(2)\uplus\langle J_{03}, D\rangle$

This class consists of those subalgebras of $AP(1,3)$ whose projection onto $AO(1,3)$ do not have invariant isotropic subspaces in $\mathbf{R}_{1,3,}$ and with a nonzero projection onto $\langle D\rangle$. We have

$$AO(1,2)\oplus\langle D\rangle : 0, \langle P_3\rangle, \langle P_0, P_1, P_2\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$AO(3)\oplus\langle D\rangle : 0, \langle P_0\rangle, \langle P_1, P_2, P_3\rangle, \langle P_0, P_1, P_2, P_3\rangle;$$

$$AO(1,3)\oplus\langle D\rangle : 0, \langle P_0, P_1, P_2, P_3\rangle.$$

G. Subalgebras of $AG_4(2)$ which are not conjugate to subalgebras of $A\tilde{P}(1,3)$

Let K be a subalgebra of $AG_4(2)$ and $\tau(K)$ its projection onto $AGL(2,\mathbf{R})$. By Propositions V.2.1 and V.2.2 of Ref. 9, the algebra K belongs to this class if and only if $\tau(K)$ is conjugate to one of the following algebras: $\langle S+T\rangle$, $\langle S+T\rangle+\langle Z\rangle$ (subdirect sum), $ASL(2,\mathbf{R})=\langle R,S,T\rangle$, $AGL(2,\mathbf{R})=\langle R,S,T,Z\rangle$. Because of this, we divide this seventh class into three subclasses, each of which consists of subalgebras having a corresponding projection onto $AGL(2,\mathbf{R})$; those subalgebras whose projections are either $ASL(2,\mathbf{R})$ or $AGL(2,\mathbf{R})$ are put into the same subclass.

1. Subalgebras whose projection onto $AGL(2,R)$ is $\langle S+T\rangle$

$$\langle S+T\rangle : 0, \langle M\rangle, \langle G_1, P_1, M\rangle, \langle G_1-\alpha^{-1}P_2, G_2+\alpha P_1, M\rangle, \langle G_1, G_2, P_1, P_2, M\rangle (0<|\alpha|\leqslant 1);$$

$$\langle S+T\pm M\rangle, \langle S+T+\alpha J_{12}\pm M\rangle (\alpha\neq 0);$$

$$\langle S+T+\alpha J_{12}\rangle : 0, \langle M\rangle, \langle G_1+\epsilon P_2, G_2-\epsilon P_1, M\rangle, \langle G_1, G_2, P_1, P_2, M\rangle (\epsilon=\pm 1, \alpha\neq 0);$$

$$\langle S+T+\epsilon J_{12}\rangle : \langle G_1+\epsilon P_2\rangle, \langle G_1+\epsilon P_2, M\rangle, \langle G_1+\epsilon P_2, G_1-\epsilon P_2, G_2+\epsilon P_1, M\rangle (\epsilon=\pm 1);$$

$$\langle S+T+\epsilon J_{12}\pm M, G_1+\epsilon P_2\rangle (\epsilon=\pm 1);$$

$$\langle S+T+\epsilon J_{12}+\epsilon G_1+P_2\rangle : 0, \langle M\rangle, \langle G_2-\epsilon P_1, M\rangle,$$

$$\langle G_1-\epsilon P_2, G_2+\epsilon P_1, M\rangle \langle G_2-\epsilon P_1, G_1-\epsilon P_2, G_2+\epsilon P_1, M\rangle (\epsilon=\pm 1);$$

$$\langle J_{12}, S+T\rangle : 0, \langle M\rangle, \langle G_1+\epsilon P_2, G_2-\epsilon P_1, M\rangle, \langle G_1, G_2, P_1, P_2, M\rangle (\epsilon=\pm 1);$$

$$\langle J_{12}\pm M, S+T+\alpha M\rangle(\alpha\in\mathbf{R});$$

$$\langle J_{12}, S+T\pm M\rangle.$$

2. Subalgebras whose projection onto $AGL(2,\mathbf{R})$ is the subdirect sum $\langle S+T\rangle+\langle Z\rangle$

$$\langle S+T+\alpha Z\rangle:0,\langle M\rangle,\langle G_1,P_1,M\rangle,\langle G_1-\beta^{-1}P_2,G_2+\beta P_1,M\rangle,$$

$$\langle G_1,G_2,P_1,P_2,M\rangle(0<|\beta|\leqslant 1,\alpha\neq 0);$$

$$\langle S+T,Z\rangle:0,\langle M\rangle,\langle G_1,P_1,M\rangle,\langle G_1-\alpha^{-1}P_2,G_2+\alpha P_1,M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle(0<|\alpha|\leqslant 1);$$

$$\langle S+T+\alpha J_{12}+\beta Z\rangle:0,\langle M\rangle,\langle G_1+\epsilon P_2,G_2-\epsilon P_1,M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle(\epsilon=\pm 1,\alpha\neq 0,\beta>0);$$

$$\langle S+T+\alpha J_{12},Z\rangle:0,\langle M\rangle,\langle G_1+\epsilon P_2,G_2-\epsilon P_1,M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle(\epsilon=\pm 1,\alpha\neq 0);$$

$$\langle S+T+\epsilon J_{12}+\alpha Z\rangle:\langle G_1+\epsilon P_2\rangle,\langle G_1+\epsilon P_2,M\rangle,$$

$$\langle G_1+\epsilon P_2,G_1-\epsilon P_2,G_2+\epsilon P_1,M\rangle(\epsilon=\pm 1,\alpha\neq 0);$$

$$\langle S+T+\epsilon J_{12},Z\rangle:\langle G_1+\epsilon P_2\rangle,\langle G_1+\epsilon P_2,M\rangle,\langle G_1+\epsilon P_2,G_1-\epsilon P_2,G_2+\epsilon P_1,M\rangle(\epsilon=\pm 1);$$

$$\langle J_{12}+\alpha Z,S+T+\beta Z\rangle:0,\langle M\rangle,\langle G_1+\epsilon P_2,G_2-\epsilon P_1,M\rangle,$$

$$\langle G_1,G_2,P_1,P_2,M\rangle(\epsilon=\pm 1,|\alpha|+|\beta|\neq 0);$$

$$\langle J_{12},S+T,Z\rangle:0,\langle M\rangle,\langle G_1+\epsilon P_2,G_2-\epsilon P_1,M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle(\epsilon=\pm 1).$$

3. Subalgebras whose projection onto $AGL(2,\mathbf{R})$ contains $ASL(2,\mathbf{R})$

$$\langle R,S,T\rangle:0,\langle M\rangle,\langle G_1,P_1,M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle;$$

$$\langle J_{12}\rangle\oplus\langle R,S,T\rangle:0,\langle M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle;$$

$$\langle J_{12}\pm M\rangle\oplus\langle R,S,T\rangle;$$

$$\langle R,S,T,Z\rangle:0,\langle M\rangle,\langle G_1,P_1,M\rangle,\langle G_1,G_2,P_1,P_2M\rangle;$$

$$\langle R,S,T\rangle\oplus\langle J_{12}+\alpha Z\rangle:0,\langle M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle(\alpha\neq 0);$$

$$\langle R,S,T\rangle\oplus\langle J_{12},Z\rangle:0,\langle M\rangle,\langle G_1,G_2,P_1,P_2,M\rangle.$$

ACKNOWLEDGMENTS

This work was supported in part by INTAS and DKNT of Ukraine. P. Basarab-Horwath thanks the Wallenberg fund of Linköping University and the Magnusson Fund of the Swedish Academy of Sciences for travel grants, and the Mathematics Institute of the Ukrainian Academy of Sciences in Kyïv for its hospitality.

Professor Wilhelm Fuschych died on April 7, 1997, after a short illness. This is a great loss for his family, his many students, and for the scientific community. His many and deep contributions to the field of symmetry analysis of differential equations have made the Kyïv school of symmetries known throughout the world. We take this opportunity to express our deep sense of loss as well as our gratitude for all the encouragement in research that Wilhelm Fushchych gave during the years we knew him.

[1] J. Patera, P. Winternitz, and H. Zassenhaus, ''Continuous subgroups of the fundamental groups of physics. I. General method and the Poincaré group,'' J. Math. Phys. **16**, 1597–1614 (1975).

[2] G. Burdet, J. Patera, M. Perrin, and P. Winternitz, ''The optical group and its subgroups,'' J. Math. Phys. **19**, 1758–1780 (1978).
[3] J. Patera, P. Winternitz, and H. Zassenhaus, ''Continuous subgroups of the fundamental groups of physics. II. The similitude group,'' J. Math. Phys. **16**, 1615–1624 (1975).
[4] J. Patera, P. Winternitz, R. T. Sharp, and H. Zassenhaus, ''Subgroups of the similitude group of three-dimensional Minkowski space,'' Can. J. Phys. **54**, 950–961 (1976).
[5] G. Burdet, J. Patera, M. Perrin, and P. Winternitz, ''Sous-algèbres de Lie de l'algèbre de Schrödinger,'' Ann. Sc. Math. Quebec **2**, 81–108 (1978).
[6] W. I. Fushchich, A. F. Barannik, L. F. Barannik, and V. M. Fedorchuk, ''Continuous subgroups of the Poincaré group $P(1,4)$,'' J. Phys. A **18**, 2893–2899 (1985).
[7] L. F. Barannik and W. I. Fushchich, ''On subalgebras of the Lie algebra of the extended Poincaré group $P(1,n)$,'' J. Math. Phys. **28**, 1445–1458 (1987).
[8] L. F. Barannik and W. I. Fushchich, ''On continuous subgroups of the generalized Schrödinger groups,'' J. Math. Phys. **30**, 280–290 (1989).
[9] W. I. Fushchich, L. F. Barannyk, and A. F. Barannyk, ''Subgroup analysis of the Galilei and Poincaré groups and reduction of nonlinear equations'' (in Russian) Kyiv. Naukova Dumka (1991).
[10] L. F. Barannik and W. I. Fushchich, ''On continuous subgroups of the conformal group of Minkowski space $\mathbf{R}_{1,n}$'' (in Russian) (preprint) Ukrainian Academy of Sciences, Kyiv, 88.34 (1988).
[11] W. I. Fushchich, L. F. Barannyk, and A. F. Barannyk, ''Continuous subgroups of the generalized Euclidean group'' (in Russian) Ukr. J. Math. **38**, 67–72 (1986).
[12] A. F. Barannyk, L. F. Barannyk, and W. I. Fushchich, ''Reduction and exact solutions of the eikonal equation'' (in Russian) Ukr. J. Math. **43**, 461–474 (1991).
[13] A. F. Barannyk, L. F. Barannyk, and W. I. Fushchich, ''Connected subgroups of the conformal group $C(1,4)$'' (in Russian) Ukr. J. Math. **43**, 870–884 (1991).
[14] S. Lang, *Algebra* (Addison–Wesley, Reading, 1965).
[15] Su-Shing Chen, ''On subgroups of the noncompact real exceptional Lie group F_4^*,'' Math. Ann. **204**, 271–284 (1973).
[16] M. S. Taufik, ''On semisimple subalgebras of some simple real Lie algebras'' (in Russian) Usp. Mat. Nauk **30**, 261–262 (1975).
[17] M. S. Taufik, ''On maximal subalgebras in classical real Lie algebras'' (in Russian) Voprosy teorii grupp i gomologicheskoj algebry. Jaroslavl Univ. 148–168 (1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Quommutator deformations of spl(N,1)

Y. Brihaye
Department of Mathematical Physics, University of Mons, Av. Maistriau, B-7000 Mons, Belgium

(Received 22 August 1997; accepted for publication 23 April 1998)

We analyze the Witten-Woronowicz's type deformations of the Lie superalgebras spl(N,1) and obtain deformations parametrized by $N(N-1)/2$ independent parameters for $N>2$ and by three parameters for $N=2$. For some of these algebras, finite dimensional representations can be constructed in terms of finite difference operators, providing operators that are relevant for the classification of quasiexactly solvable, finite difference equations. © *1998 American Institute of Physics.* [S0022-2488(98)03207-1]

I. INTRODUCTION

The study of deformations of classical Lie algebras has attracted much attention in the last few years. Many types of deformations have been proposed. In some of them, the commutators of the generators are set equal to nonlinear (in fact transcendental) functions of some of the generators; they are the Drinfeld–Jimbo deformations.[1,2] Other types of deformations consist in keeping the right-hand sides of the structure relations linear in the generators while the commutators of the left-hand side are deformed into quommutators, such deformations are called Witten–Woronowicz type.[3,4] In this second approach, all the generators are treated on the same footing and the quommutator relations provide normal ordering rules which allow one to write the elements of the enveloping algebra in a canonical form. This last property is useful for the kind of applications that we have in mind, namely the study of the quasiexactly solvable (QES), finite difference operators. The QES (differential or finite difference) equations refer to a class of spectral equations for which a finite part of the spectrum can be obtained by solving a system of algebraic equations.[5,6] In this respect, the linear operators which preserve a finite dimensional vector space of smooth functions are naturally related to QES equations; such operators are also called quasiexactly solvable. The invariant space is typically a direct sum of modules of polynomials of given degrees.

The observation that QES operators possess hidden algebras[5,6] and that they can be related to some abstract algebras has motivated several tentatives to classify these operators and to study their properties in more detail (see Refs. 7 and 8 for recent reviews). The scalar QES operators of one variable, preserving the vector space of polynomials of degree at most n, are related to the enveloping algebra of the Lie algebra sl(2).[9] Choosing, more generically, V as a direct sum of modules of polynomials, it is likely[10–14] that a finite number of basic QES operators can be put in correspondence with the generators of an abstract algebra, suitably represented in terms of differential operators. The more general QES operators can then be obtained as the polynomials in the basic operators, in other words they are the elements of an enveloping algebra.

The present understanding of QES equations suggests that the scalar QES equations are related to the classical Lie algebras[10,11] while the systems of coupled QES equations are related to superalgebras.[10,12,14] For instance, the superalgebra osp(2,2) [or equivalently spl(2,1)] plays a crucial role[10,12] for the systems of two equations of one variable. In the case of systems of two equations of V variables, the underlying symmetry is related to the superalgebras spl($V+1$,1).[14]

On the other hand it seems that, if we want to describe the algebraic structure underlying the QES finite difference equations, the representations of some deformed algebra emerge in a natural way. For instance, the scalar, finite difference QES operators in one variable[9] are related to the Witten deformation of sl(2). Owing to this example and the results of Ref. 14 it is natural to try to classify the deformations of Witten–Woronowicz type of the Lie superalgebras spl(N,1). This is the object of this paper; Drinfeld–Jimbo type deformations of general Lie superalgebras are discussed in Ref. 15.

 © 1998 American Institute of Physics

II. THE ALGEBRA spl(N,1)

The graded Lie algebra spl(N,1) is generated by N^2 bosonic generators, N fermionic and N antifermionic generators. We will denote them, respectively, by E_a^b, V_a and $\overline{V^a}$, $a,b=1,...,N$. The structure relations read as follows:

$$\{V_a, V_b\}=0, \tag{1}$$

$$\{\overline{V^a}, \overline{V^b}\}=0, \tag{2}$$

$$\{V_a, \overline{V^b}\}=E_a^b, \tag{3}$$

$$[E_a^b, V_c]=\delta_c^b V_a-\delta_a^b V_c, \tag{4}$$

$$[E_a^b, \overline{V^c}]=\delta_a^b \overline{V^c}-\delta_a^c \overline{V^b}, \tag{5}$$

$$[E_a^b, E_c^d]=\delta_c^b E_a^d-\delta_a^d E_c^b. \tag{6}$$

As indicated in (6), the generators E_a^b form a gl(N) subalgebra. Equations (4) [resp. (5)] show that the generators V_a's (resp. $\overline{V^b}$) form an N-dimensional representation of gl(N) under the adjoint action of the E_a^b.

We will consider the deformations of spl(N,1) obtained by replacing the commutators (4)–(6) and the anticommutators (1)–(3) respectively, by quommutators and antiquommutators:

$$[A,B]_q=AB-qBA, \quad \{A,B\}_q=AB+qBA, \tag{7}$$

where q represents a parameter of deformation. More specifically, we classify the deformations obeying the following conditions:

(i) Each commutator (resp. antiquommutator) (1)–(6) defining the classical algebra is replaced by a quommutator (resp. antiquommutator) (7) with its own parameter q.
(ii) The couples of generators which (anti-) commute are imposed to (anti-) quommute.
(iii) The right hand side of (3)–(5) is kept linear in the generators; on the right-hand side of (6) we allow additional terms of the form $V_f \overline{V^g}$.

III. DEFORMATIONS OF spl(N,1)

In order to present the deformations of spl(N,1) it is convenient to introduce a set of $N(N-1)/2$ arbitrary parameters labeled q_{ab} such that

$$q_{ba}=\frac{1}{q_{ab}}, \quad a,b=1,\ldots,N. \tag{8}$$

The deformations of spl(N,1) obeying the conditions above and compatible with the associativity conditions (i.e., the Jacobi identities) read as follows:

$$V_a V_b+q_{ab}V_b V_a=0, \tag{9}$$

$$\overline{V^a}\,\overline{V^b}+q_{ab}\overline{V^b}\,\overline{V^a}=0, \tag{10}$$

$$V_a \overline{V^b}+q_{ba}\overline{V^b}V_a=E_a^b, \tag{11}$$

$$E_a^b V_c-q_{ac}q_{cb}V_c E_a^b=\delta_c^b V_a-\delta_a^b V_c, \tag{12}$$

$$E_a^b \overline{V^c}-q_{bc}q_{ca}\overline{V^c}E_a^b=\delta_a^b \overline{V^c}-q_{ba}\delta_a^c \overline{V_b}, \tag{13}$$

$$E_a^b E_c^d-q_{ac}q_{cb}q_{bd}q_{da}E_c^d E_a^b=\delta_b^c E_a^d-q_{ba}q_{ac}q_{cb}\delta_a^d E_c^b. \tag{14}$$

The normalization of the generators has been chosen in such a way that the right-hand sides of (11) and (12) look as simple as possible. The bosonic generators close under the quommutators (14), leading to a deformation of the Lie algebra gl(N). It can be checked that the products of three and of four q's defining the structure constants in (14) depend effectively on $(N-1)(N-2)/2$ parameters among the whole set of the $N(N-1)/2$ parameters q_{ab}. In fact, the algebra (14) coincides with the deformation of gl(N) constructed in Ref. 16. It is amusing to note that the parameters q_{ab}, used to label the deformations in Ref. 16, appear here as defining the antiquommutators (9) and (10) of the fermionic operators. Note, however, that another quommutator is used in Ref. 16, it can be related to (14) by an appropriate renormalization of the generators.

IV. DEFORMATIONS OF spl(2,1)

The analysis of Ref. 16 reveals that the algebra gl(2) leads to a richer set of deformations than for the generic case gl(N) for $N>2$. A similar result holds for the deformations of spl(N,1). We now discuss the deformations of spl(2,1) which, remember, is equivalent to the (perhaps more popular) graded Lie algebra osp(2,2).

The most general deformation of spl(2,1) obeying the three restrictions above and compatible with associativity depends on three parameters, say p,r,s.

The different relations read as follows.

(1) For the fermionic–fermionic relations,

$$V_1V_1=V_2V_2=0, \quad \overline{V}^1\overline{V}^1=\overline{V}^2\overline{V}^2=0, \tag{15}$$

$$\{V_1,V_2\}_{(p/sr)}=0, \quad \{\overline{V}^1,\overline{V}^2\}_{(s/pr)}=0, \tag{16}$$

$$\{\overline{V}^j,V_j\}=E_j^j, \quad j=1,2, \tag{17}$$

$$\{\overline{V}^1,V_2\}_{psr}=E_2^1, \quad \{\overline{V}^2,V_1\}_{(ps/r)}=E_1^2. \tag{18}$$

(2) For the fermionic–bosonic relations,

$$\begin{gathered}
[E_1^1,V_1]=0, \quad [E_2^2,V_1]_{s^2}=V_1, \\
[E_1^2,V_1]_{ps/r}=0, \quad [E_2^1,V_1]_{sr/p}=-\frac{sr}{p}V_2, \\
[E_1^1,V_2]_{p^2}=V_2, \quad [E_2^2,V_2]=0, \\
[E_1^2,V_2]_{p/sr}=-\frac{p}{sr}V_1, \quad [E_2^1,V_2]_{psr}=0,
\end{gathered} \tag{19}$$

$$\begin{gathered}
[E_1^1,\overline{V}^1]=0, \quad [E_2^2,\overline{V}^1]_{1/s^2}=-\frac{1}{s^2}\overline{V}^1, \\
[E_1^2,\overline{V}^1]_{ps/r}=\overline{V}^2, \quad [E_2^2,\overline{V}^1]_{1/psr}=0, \\
[E_1^1,\overline{V}^2]_{1/p^2}=-\frac{1}{p^2}\overline{V}^2, \quad [E_2^2,\overline{V}^2]=0, \\
[E_1^2,\overline{V}^2]_{r/ps}=0, \quad [E_2^1,\overline{V}^2]_{s/pr}=\overline{V}^1.
\end{gathered} \tag{20}$$

(3) For the bosonic–bosonic operators,

$$[E_1^1,E_2^2]=0, \tag{21}$$

$$[E_1^1,E_1^2]_{1/p^2}=-\frac{1}{p^2}E_1^2,\quad [E_2^2,E_1^2]_{s^2}=E_1^2, \tag{22}$$

$$[E_1^1,E_2^1]_{p^2}=E_2^1,\quad [E_1^1,E_2^1]_{1/s^2}=-\frac{1}{s^2}E_2^1, \tag{23}$$

$$[E_1^2,E_2^1]_{s^2/p^2}=E_1^1-\frac{s^2}{p^2}E_2^2+(s^2-1)V_1\overline{V^1}-\frac{s^2}{p^2}(p^2-1)V_2\overline{V^2}. \tag{24}$$

The three parameters of the deformation, noted p,r,s, are intrinsic and cannot be eliminated by a rescaling of the generators. This is seen easily from the way these parameters enter in the different coefficients defining the quommutator. The undeformed osp(2,2) algebra is recovered by $p=s=r=1$. In the limit $s^2=p^2=1$, the generic case of the previous section is recovered: There is only one deforming parameter, noted r. The gl(2) subalgebra is undeformed.

Keeping r,s,p arbitrary and decoupling the fermionic generators from Eq. (24) leads, together with (21), (22) and (23), to the two parameter deformation of gl(2) obtained in Ref. 17 (a four-parameter deformation of gl(2) was further obtained in Ref. 16).

V. REPRESENTATIONS

Restricting the three parameters p,r,s in the algebra presented in Sec. IV to the case

$$p=s^{-1}=q^{1/2},\quad r=1 \tag{25}$$

leads to a one parameter deformation of osp(2,2) parametrized by q which we will refer to as $\mathrm{osp}(2,2)_q$. The following operator q-commutes with the eight generators of the algebra

$$C=V_1\overline{V^1}-qV_2\overline{V^2}-q^2E_2^1E_1^2+qE_1^1E_2^2-E_1^1. \tag{26}$$

As for the Witten type deformation of sl(2),[9] it is possible to construct some representations of $\mathrm{osp}(2,2)_q$ in terms of a finite difference operator D_q:

$$D_qf(x)=\frac{f(x)-f(qx)}{(1-q)x},\quad D_qx^n=[n]_qx^{n-1},\quad [n]_q\equiv\frac{1-q^n}{1-q}. \tag{27}$$

The representations we want to discuss act on the vector space $P(n-1)\oplus P(n)$ [where $P(n)$ denotes the module of polynomials of degree n in the variable x]. It is therefore of dimension $2n-1$. The fermionic generators are represented by

$$\begin{gathered} V_1=\sigma_-,\quad V_2=x\sigma_-,\\ \overline{V_1}=q^{-n}(xD_q-[n]_q)\sigma_+,\quad \overline{V_2}=q^{-1}D_q\sigma_+, \end{gathered} \tag{28}$$

where $\sigma_\pm=(\sigma_1\pm i\sigma_2)/2$. The operators E_j^i are easily constructed with (16) and (17). It appears that the operator (26) vanishes for these representations.

Since the representations above are irreducible, all 2×2 matrix finite difference operators preserving $P(n-1)\oplus P(n)$ are elements of the enveloping algebra generated by (28).[7] These operators are therefore quasiexactly solvable.

VI. DEFORMATIONS OF osp(1,2)

Recently, a Witten type deformation of the super Lie algebra osp(1,2) was constructed.[18] We showed that this deformation is the only one to fulfill the three requirements of Sec. II, its finite dimensional representations can be expressed in terms of the operator (27). Again, the space of the representation is $P(n-1)\oplus P(n)$ and the two fermionic generators of osp(1,2) are represented by 2×2 matrix operators. Using exactly the same notation as in Ref. 18 we find

$$V_- = \begin{pmatrix} 0 & D_{q^2} \\ 1 & 0 \end{pmatrix}, \quad V_+ = \begin{pmatrix} 0 & q^{-2n}(xD_{q^2} - [n]_{q^2}) \\ x & 0 \end{pmatrix}. \tag{29}$$

The three bosonic operators, H, J_-, J_+ can then be computed through the structure of osp(1,2), namely,

$$H = \{V_-, V_+\}_q, \quad J_- = \{V_-, V_-\}_q, \quad J_+ = \{V_+, V_+\}_q. \tag{30}$$

Correspondingly, the Casimir operator [Eq. (13) in Ref. 18] has a value $C = -\frac{1}{2}[-n-\frac{1}{2}]_{q^2}$.

VII. CONCLUDING REMARKS

The representations of spl(V+1,1) formulated in terms of differential operators provide the building blocks for the construction of the QES operators preserving more general polynomial spaces of V variables.[14] This result, and the fact that the quommutators deformations of super Lie algebras have not been studied systematically (at least to our knowledge), motivate a study of the deformations of the Lie superalgebras spl(N,1) by means of quommutators.

On the other hand, the most interesting examples of QES systems (e.g., the relativistic Coulomb problem,[19] the doubly periodic Lame equation and the stability of the sphaleron in the Abelian Higgs model[12]) are related to the algebra osp(2,2). Generalizations of this algebra, e.g., its deformations, and the corresponding realizations in terms of finite difference operators therefore deserve some attention. The operators (28) could be relevant for the study of some coupled problems arising in discrete quantum mechanics in Ref. 20.

On a more abstract level we mention that the operators (28) can be used to construct finite difference operators preserving the vector space $P(m) \oplus P(n)$. The algebraic structure underlying these operators is probably determined by a series of deformed, nonlinear superalgebras indexed by $|m-n|$. Finally, it would be interesting to relate the two-parameter deformation of osp(2,2) of Ref. 21 with our deformation. The occurrence of some mapping between the two structures would allow us to transport the Hopf structure to our case.

[1] V. Drinfeld, Proceedings of the International Congress of Mathematicians, Berkeley, CA, 1986 (unpublished), p. 798.
[2] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985).
[3] E. Witten, Nucl. Phys. B **330**, 285 (1990).
[4] S. Woronowicz, Commun. Math. Phys. **111**, 613 (1987).
[5] A. V. Turbiner, Commun. Math. Phys. **119**, 467 (1988).
[6] A. G. Ushveridze, *Quasi Exact Solvability in Quantum Mechanics* (IOP, Bristol, 1993).
[7] A. V. Turbiner, Am. Math. Soc. Trans. **160**, 263 (1994).
[8] A. V. Turbiner, in *Lie-algebras and Quasi-exactly-solvable Differential Equations*, CRC Handbook of Lie Group Analysis and Differential Equations, Vol. 3, edited by N. Ibragimov (Chemical Rubber, Cleveland, OH, 1995), Chap. 12, p. 331.
[9] A. V. Turbiner, J. Phys. A **25**, L1087 (1992).
[10] M. A. Shifman and A. V. Turbiner, Commun. Math. Phys. **120**, 347 (1989).
[11] A. Gonzalez-Lopez, N. Kamran, and P. J. Olver, J. Phys. A **24**, 3995 (1991).
[12] Y. Brihaye and P. Kosinski, J. Math. Phys. **35**, 3089 (1994).
[13] Y. Brihaye, S. Giller, C. Gonera, and P. Kosinski, J. Math. Phys. **36**, 4340 (1995).
[14] Y. Brihaye and J. Nuyts, Commun. Math. Phys. (in press).
[15] R. Floreanini, V. Spiridonov, and L. Vinet, Commun. Math. Phys. **137**, 149 (1991).
[16] D. B. Fairlie and J. Nuyts, J. Math. Phys. **35**, 3794 (1994).
[17] D. B. Fairlie and C. K. Zachos, Phys. Lett. B **256**, 43 (1991).
[18] W. Chung (unpublished).
[19] Y. Brihaye, N. Devaux, and P. Kosinski, Int. J. Mod. Phys. A **10**, 4633 (1995).
[20] P. B. Wiegmann and A. V. Zabrodin, Nucl. Phys. B **451**, 699 (1995).
[21] P. Parashar, J. Phys. A **27**, 3803 (1994).

On an easy transition from operator dynamics to generating functionals by Clifford algebras

Bertfried Fauser[a)]
Eberhard-Karls-Universität, Institut für Theoretische Physik, Auf der Morgenstelle 14, 72072 Tübingen, Germany

(Received 22 December 1997; accepted for publication 14 April 1998)

Clifford geometric algebras of multivectors are treated in detail. These algebras are built over a graded space and exhibit a grading or multivector structure. The careful study of the endomorphisms of this space makes it clear that opposite Clifford algebras have to be used for obtaining all endomorphisms. Based on these mathematics, we give a fully Clifford algebraic account on generating functionals, which is thereby geometric. The field operators are shown to be Clifford and opposite Clifford maps. This picture, relying on geometry, does not need positivity in principle. Furthermore, we propose a transition from operator dynamics to generating functionals, which is based on the algebraic techniques. As a calculational benefit, this transition is considerably short compared to standard ones. The transition is not injective (unique) and depends additionally on the choice of an ordering. We obtain a direct and constructive connection between orderings and the explicit form of the functional Hamiltonian. These orderings depend on the propagator of the theory and thus on the ground state. This is invisible in path integral formulations. The method is demonstrated within two examples, a nonlinear spinor field theory and spinor QED. Antisymmetrized and normal-ordered functional equations are derived in both cases. © *1998 American Institute of Physics.* [S0022-2488(98)02207-5]

I. INTRODUCTION

Modern quantum field theory is treated currently by means of path integrals.[1,2] Path integrals provide formal solutions of functional differential equations. Such functional differential equations encode the matrix element hierarchies of Dyson–Schwinger–Freese.[3–6] Generating functionals describe in a condensed form an infinite set of vacuum or transition matrix elements with a single mathematical entity. They prove to be useful in manipulations of the whole hierarchy.

Since path integrals are very compact in their formulation, one does not actually have an access to their meaning. Moreover, one is troubled with convergence problems, which can be treated up to now only by heuristic methods or in trivial cases; see chapter 6 of Ref. 2.

It seems therefore to be a good idea to study generating functionals and functional differential equations in their own right. Furthermore, it is known from ordinary differential equations that one can achieve information about the solutions without being able to integrate the differential equations. We expect to be able to support the usage of path integrals, at least their algebraic properties, in studying operator dynamics with the help of functional differential equations.

The connection between Lie groups and spin groups was established in Ref. 7, where the authors employed the bi-vector space, which is closed under the commutator product. To be able to treat nonlinear Hamiltonians (see the examples) we extend the mechanism to the whole graded space, which turns all algebra elements into Lie group elements.

The paper is organized in three logical parts. First, in section II, we establish the mathematical apparatus necessary to treat generating functionals algebraically correct. This will be done in the case of antisymmetry (fermions), but might be equally well established for symmetric algebras (bosons); see Ref. 8. This part is mathematically rigorous.

In the second part, consisting of sections III and IV, the infinite dimensional case of QFT is treated formally, since the transition to infinite many generators causes serious convergence prob-

[a)]Electronic mail: Bertfried.Fauser@uni-tuebingen.de

 © 1998 American Institute of Physics

lems. We give an algebraic account on generating functionals based on the Clifford algebra of multivectors. Field operators are shown to be Clifford endomorphisms of the graded space underlying the Grassmann algebra built over the Schwinger sources. This opens for the first time an algebraically motivated approach to quantum field theoretic functionals. In section IV we use the Heisenberg equation to give a very simple transition from operator dynamics into the space of generating functionals. Once more, this can be achieved only if one admits the Clifford geometric point of view. A single operator dynamics might result in different functional equations. This peculiar observation is an essential insight to understanding the computational power of functional differential equations and path integrals. The different functional equations obtained are subjected to different orderings and ultimately to different vacua. The correct choice of this ordering by means of normal-ordering with the help of a specific propagator selects a unique vacuum encoded implicitly in the theory, as was shown in Ref. 9. The Clifford geometric approach given in this paper makes it possible to investigate these connections and to have a constructive tool to implement inequivalent vacua in QFT.

In the third part, sections V and VI, examples of the method are given. A nonlinear spinor field theory, e.g., a Hubbard or Nambu–Jona-Lasinio type model, and spinor QED as an example of a boson-fermion coupling theory are treated. Both examples are dealt with in the antisymmetric and normal-ordered case. Explicit calculations of the functional equations shows furthermore the efficiency of our method. The usually necessary vertex-regularization is no longer apparent, since the algebraic treatment accounts for this transition correctly. This was know earlier on the operator level.[10] It is novel that one can achieve the normal-ordered functional equation in one single step.

In the conclusion the results are summarized and the relevance to path integral calculations is discussed.

II. CLIFFORD GEOMETRIC ALGEBRA OF MULTIVECTORS

There are many possibilities to introduce Clifford algebras, each of them emphasize a different point of view. In our case, it is of utmost importance to have the Clifford algebra built over a graded linear space. This grading is obtained from the space underlying a Grassmann algebra. The Clifford algebra is then related to the endomorphism algebra of this space. This construction, the Chevalley deformation,[11] was originally invented to be able to treat Clifford algebras over fields of $char=2$; see the appendix of Ref. 12 by Lounesto and Ref. 13. However, we use this construction in an entirely different context. With the help of the construction of Riesz,[12] one is able to reconstruct the multivector structure and thereby a correspondence between the linear spaces underlying the Clifford algebra and the Grassmann algebra in use. This reconstruction depends on an automorphism J, which is arbitrary; see Ref. 14. In fact this is just the reversed direction of our construction following Chevalley given below.

Let $T(V)$ be the tensor algebra built over the $\mathbb{K}$-linear space V. The field $\mathbb{K}$ will be either $\mathbb{R}$ or $\mathbb{C}$. With $V^0 \simeq \mathbb{K}$ we have

$$T(V)=\mathbb{K}\oplus V\oplus V\otimes_{\mathbb{K}} V\oplus\cdots . \tag{2.1}$$

The tensor algebra is associative and unital. In $T(V)$ one has bilateral or two-sided ideals, which can be used to construct new algebras by factorization. As an example we define the Grassmann algebra in this way.

Definition 1: The Grassmann algebra $\bigwedge(V)$ is the factor algebra of the tensor algebra w.r.t. the bilateral ideal

$$\begin{aligned} I_{Gr}&=\{y\,|\,y=a\otimes x\otimes x\otimes b,\quad a,b\in T(V),x\in V\}\\ \bigwedge(V)&=\pi(T(V))=\frac{T(V)}{I_{Gr}}=\mathbb{K}\oplus V\oplus V\wedge V\oplus\cdots . \end{aligned} \tag{2.2}$$

The canonical projection π: $T(V)\mapsto\bigwedge(V)$ maps the tensor product $\otimes$ onto the exterior or wedge product denoted by $\wedge$. ■

One may note that the factorization preserves the grading naturally inherited by the tensor algebra, since the ideal I_{Gr} is homogeneous. Defining homogeneous parts of $T(V)$ and $\wedge(V)$ by $T^k(V)=V\otimes\cdots\otimes V$ and $\wedge^k(V)=V\wedge\cdots\wedge V$ k-factors, we obtain $\pi(T^k(V))=\wedge^k(V)$.

Proceeding to Clifford algebras requires a further structure, the quadratic form.

Definition 2: The map $Q:V\mapsto\mathbb{K}$, satisfying

$$\text{(i)}\quad Q(\alpha x)=\alpha^2 Q(x),\quad \alpha\in\mathbb{K},x\in V,$$

$$\text{(ii)}\quad B_p(x,y)=\tfrac{1}{2}\,(Q(x+y)-Q(x)-Q(y)), \tag{2.3}$$

where $B_p(x,y)$ is a symmetric bilinear form, is called a quadratic form. ■

It is tempting to introduce an ideal I_{Cl},

$$I_{Cl}=\{y|y=a\otimes(x\otimes x-Q(x)\mathbb{1})\otimes b,\quad a,b\in T(V),x\in V\}, \tag{2.4}$$

to obtain the Clifford algebra by a factorization procedure. However, since we are interested in arbitrary bilinear forms underlying a Clifford algebra, we will take another approach, which is wide enough for such a structure. Furthermore, the Clifford algebra does not have an intrinsic multivector structure, but is only $\mathbb{Z}_2$ graded, since the ideal I_{Cl} is inhomogeneous but $\mathbb{Z}_2$ graded.

Let V^* be the space of linear forms on V, i.e., $V^*\simeq\text{lin}[V,\mathbb{K}]$. Elements $\omega\in V^*$ act on elements $x\in V$; but there is no natural connection between V and V^*. However, we can find a set of x_i which span V and dual elements ω_k acting on the x_i in a canonical way,

$$\omega_k(x_i)=\delta_{ki}\,. \tag{2.5}$$

This allows us to introduce a map $*$: $V\mapsto V^*$, $x_i^*=\omega_i$ which may be called an Euclidean dual isomorphism.[15] The pair (V^*,V) is connected by this duality which constitutes a pairing $\langle\cdot|\cdot\rangle$: $V^*\times V\mapsto\mathbb{K}$. Since V^* is isomorphic to V in finite dimensions, it is natural to build a Grassmann algebra $\wedge(V^*)$ over it. This is the algebra of Grassmann multiforms.

It is further a natural thing to extend the pairing of the grade-one space and its dual to the whole algebras $\wedge(V)$ and $\wedge(V^*)$, as can be seen by its frequent occurrence in literature.[13,14,16–19] This can be done by the following.

Definition 3: Let $\tau,\eta\in\wedge(V^*)$, $\omega\in V^*$, $u,v\in\wedge(V)$ and $x\in V$; then we can define a canonical action of $\wedge(V^*)$ on $\wedge(V)$ requiring

$$\text{(i)}\quad \omega(x)=\langle\omega|x\rangle,$$

$$\text{(ii)}\quad \omega(u\wedge v)=w(u)\wedge v+\hat{u}\wedge\omega(v), \tag{2.6}$$

$$\text{(iii)}\quad (\tau\wedge\eta)(u)=\tau(\eta(u)),$$

where $\hat{u}$ is the main involution $\hat{V}=-V$ extended to $\wedge(V)$. ■

In fact we have given by definition 3 an isomorphism between the Grassmann algebra of multiforms $\wedge(V^*)$ and the dual Grassmann algebra $[\wedge(V)]^*$. This can be made much clearer by writing

$$y\lrcorner x=\omega_y(x)=\langle\omega_y|y\rangle=B(y,x), \tag{2.7}$$

where we have used the canonical identification of V and V^* via the Euclidean dual isomorphism. One should be very careful in the distinction of $\wedge(V^*)$ and $[\wedge(V)]^*$, since they are isomorphic but not equivalent. Furthermore, we emphasize that in writing $y\lrcorner$ we make explicit use of a special dual isomorphism encoded in the contraction,

$$.\lrcorner:V\mapsto V^*,\quad y\to y\lrcorner=\omega_y\,. \tag{2.8}$$

Since there is no natural, either is mathematically motivated or even better, functorial relation between V and V^*, we are called to seek physically motivated reasons to select a dual isomorphism.

Theorem 4: Let (V,Q) be a pair of a $\mathbb{K}$-linear space V and Q a quadratic form as in definition 2. There exists an injection γ called Clifford map from V into the associative unital algebra $Cl(V,Q)$ which satisfies

$$\gamma_x\gamma_x = Q(x)\mathbb{1}. \tag{2.9}$$

■

Definition 5: The (smallest) algebra $Cl(V,Q)$ generated by $\mathbb{1}$ and γ_x is called (the) Clifford algebra of Q over V. ■

By polarization of this relation we get the usual commutation relations; $x,y \in V$,

$$\gamma_x\gamma_y + \gamma_y\gamma_x = 2B_p(x,y), \tag{2.10}$$

where $B_p(x,y)$ is the symmetric polar form of Q as defined in (2.3).

Remarks: (i) We could have obtained this result directly factoring the tensor algebra with the ideal (2.4). (ii) There exists Clifford algebras which are universal, in this case it is convenient to speak from the Clifford algebra over (V,Q). (iii) If $V \simeq \mathbb{K}^n \simeq \mathbb{C}^n$ or $\mathbb{R}^n$, we denote $Cl(V,Q)$ also by $Cl(\mathbb{C}^n) \simeq Cl_n$ or $Cl(\mathbb{R}_{p,q})$ where the pair p,q enumerate the number of positive and negative eigenvalues of Q. We can also give the dimension n and signature $s=p-q$ to classify all quadratic forms over $\mathbb{R}$. In the case of the complex field, one remains with the dimension, as can be seen, e.g., from the Weyl unitary trick, letting $x_i \to ix_i$ which flips the sign. We do not use sesquilinear forms here, which could be included nevertheless. ■

We will now use Chevalley-deformations to construct the Clifford algebra of multivectors. The main idea is that we can decompose the Clifford map as

$$\gamma_x = x\lrcorner + x\wedge. \tag{2.11}$$

We have thus a natural action of γ_x on the space $F\bigwedge(V)$.

Theorem 6: (Chevalley) Let $F\bigwedge(V)$ be the space underlying the Grassmann algebra over V and $\gamma_x: V \mapsto \mathrm{End}(F\bigwedge(V))$, $x \in V$ be defined as in (2.11); then γ_x is a Clifford map. ■

We have shown that Cl is a subalgebra of the endomorphism algebra of $F\bigwedge(V)$,

$$Cl \subseteq \mathrm{End}(F\bigwedge(V)). \tag{2.12}$$

It is possible to interpret $x\lrcorner$ and $x\wedge$ as annihilating and creation operators (on the space underlying the Grassmann algebra).[8]

With help of the relations (2.6) we can then lift this Clifford map to multivector actions. No symmetry requirement has to be made on the contraction. This leads to the following.

Definition 7: (Clifford algebra of multivectors) Let $B: V\times V \mapsto \mathbb{K}$ be an arbitrary bilinear form. The Clifford algebra $Cl(V,B)$ obtained from lifting the Clifford map,

$$\gamma_x = x\lrcorner + x\wedge = \langle x|\cdot\rangle + x\wedge = B(x,\cdot) + x\wedge, \tag{2.13}$$

to $\mathrm{End}(F\bigwedge(V))$ using the relations (2.6) is called the Clifford algebra of multivectors. ■

Note that $B(x,\cdot)=\omega_x$ is a map from $V \to V^*$ and incorporates a dual isomorphism. It is clear from the construction that $Cl(V,B)$ has a multivector structure or say a $\mathbb{Z}_n$-grading inherited from the graded space $F\bigwedge(V)$.

B admits a decomposition into symmetric and antisymmetric parts $B=G+F$. The symmetric part $G=B_p$ corresponds to a quadratic form Q; see (2.3).

Theorem 8: The Clifford algebra $Cl(V,Q) \simeq Cl(V,G)$ is isomorphic as Clifford algebra to $Cl(V,B)$, if B admits a decomposition $B=G+F$, $G^T=G$, $F^T=-F$. ■

A proof can be found for low dimensions in Ref. 19 and in general in Ref. 20. However, this result was implicitly known to physicists; see Refs. 10, 18, 21. In fact, this is the old Wick rule of QFT. We will insist on the $\mathbb{Z}_n$-grading and therefore carefully distinguish Clifford algebras of multivectors with a common quadratic form Q but different contractions B.

Definition 9: The opposite algebra A^{op} of an algebra A with product $m(a,b)=ab$ is defined to be the same linear space FA underlying A endowed with the opposite or transposed product $m^{op}(a,b)=m(b,a)=ba$. ■

Theorem 10: (Chevalley) The opposite Clifford algebra $Cl^{op}(V,Q)$ of $Cl(V,Q)$ is isomorphic to $Cl(V,-Q)$,

$$Cl^{op}(V,Q)\simeq Cl(V,-Q). \tag{2.14}$$

■

We can generalize this theorem to Clifford algebras of multivectors as the following.

Theorem 11: For Clifford algebras of multivectors it holds that

$$Cl^{op}(V,B)\simeq Cl(V,-B^T), \tag{2.15}$$

where T denotes the transposition of the bilinear form $B(x,y)^T=B(y,x)$. ■

The most general linear transformation on $F\wedge(V)$ is achieved by left and right translations, we have

$$\begin{aligned}\mathrm{End}(F\wedge V)\simeq\wedge(V)\otimes[\wedge(V)]^*&\simeq Cl(V,B)\hat{\otimes}Cl^{op}(V,B)\\ &\simeq Cl(V,B)\hat{\otimes}Cl(V,-B^T)\simeq Cl(V\oplus V,B\oplus -B^T);\end{aligned} \tag{2.16}$$

see, e.g., Ref. 22 theorem 5.5 or Ref. 23. This is the algebra which will be used in the subsequent sections.

The fact that in the above construction one has the graded tensor product is reflected by the following formulas. Let

$$\gamma_x=x\lrcorner_B+x\wedge, \quad \gamma_x^{op}=x\lrcorner_{-B^T}+x\wedge, \tag{2.17}$$

be the Clifford maps into $Cl(V,B)$ and $Cl^{op}(V,B)\simeq Cl(V,-B^T)$; then we obtain the following commutation relations:

$$\begin{aligned}\gamma_x\gamma_y+\gamma_y\gamma_x&=+2G(x,y),\\ y_x^{op}\gamma_y^{op}+\gamma_y^{op}\gamma_x^{op}&=-2G(x,y),\\ \gamma_x\gamma_y^{op}+\gamma_y^{op}\gamma_x&=0.\end{aligned} \tag{2.18}$$

Remember the decompositions $B=G+F$ and $B^{op}=-B^T=-G+F$. If one wants to treat ordinary left and right translations,

$$L_ax=ax, \quad R_ax=xa, \tag{2.19}$$

one has to introduce a further involution which accounts for the grading in the above graded tensor product of Cl and Cl^{op}. One defines

$$L_x=\gamma_x, \quad R_x=\gamma_x^{op}(\hat{\cdot}), \tag{2.20}$$

where $(\hat{\cdot})$ indicates the grade involution operator acting to the right. We obtain in this case the commutation relations as

$$\begin{aligned}L_xL_y+L_yL_x&=\gamma_x\gamma_y+\gamma_y\gamma_x=2G(x,y),\\ R_xR_y+R_yR_x&=\gamma_x^{op}\hat{\gamma}_y^{op}(\hat{\cdot})+\gamma_y^{op}\hat{\gamma}_x^{op}(\hat{\cdot})=2G(x,y)(\hat{\cdot}),\\ L_xR_y-R_yL_x&=\gamma_x\gamma_y^{op}(\hat{\cdot})-\gamma_y^{op}\hat{\gamma}_x(\hat{\cdot})=0.\end{aligned} \tag{2.21}$$

We give some further notations. Let $\{j_i\}$ be a set of elements spanning $V\simeq\langle j_1,\ldots,j_n\rangle$ and $\{\partial_k\}$ be a set of dual elements. Building the Grassmann algebras $\wedge(V)$, $\wedge(V^*)$ and defining the action of the forms via (2.6), one obtains the relations

$$\text{(i)}\quad j_i\wedge j_i=0=\partial_i\wedge\partial_i, \tag{2.22}$$

$$\text{(ii)}\quad \partial_i j_k + j_k \partial_i = B_{ik} + B_{ki} = 2G_{ik}\,.$$

The space $\mathbf{V} = V \oplus V^*$ is thus spanned by (note the order of indices)

$$\{e_1,\ldots,e_{2n}\} = \{j_1,\ldots,j_n,\partial_1,\ldots,\partial_n\}. \tag{2.23}$$

Usually, the choice of the contraction is taken to be the canonical one, with $G_{ik} = \delta_{ik}$. This leads to the so-called quantum algebra.[24] The resulting algebra is $Cl(\mathbf{V}, B^{can})$, where

$$[B^{can}(e_i,e_j)] = [B^{can}_{ij}] = \begin{vmatrix} 0_{n\times n} & \frac{1}{2}\,\mathbb{1}_{n\times n} \\ \frac{1}{2}\,\mathbb{1}_{n\times n} & 0_{n\times n} \end{vmatrix}. \tag{2.24}$$

This leads to the far more restricted commutation relations of Sallers quantum algebra (which is in fact a CAR algebra),

$$\begin{aligned} &\text{(i)}\quad j_i j_k + j_k j_i = 0 = \partial_i \partial_k + \partial_k \partial_i\,, \\ &\text{(ii)}\quad \partial_i j_k + j_k \partial_i = \delta_{ik}\,. \end{aligned} \tag{2.25}$$

However, see Ref. 25 for the usefulness of a more general bilinear form B. We will employ the j_i sources and ∂_k duals to construct field operators. A more general bilinear form $B = G + F$ is then introduced explicitly via the quantization G and the propagator F in terms of j sources and ∂ duals.

III. GENERATING FUNCTIONALS

Generating functionals originated from two major considerations. The first arose from studying covariant formulated perturbation theory;[3] see also Refs. 26, 27, to be able to handle free-free transitions of particles via a scattering matrix. The main ingredients are vacuum expectation values of time-ordered products of field operators. Since in- and out-going states are assumed to be free, one can retain the Fock representation, which, however, is not valid in the interacting case. Path integrals originated out of this branch of physics.[28]

A second approach was developed to treat QFT in an axiomatic way.[29] The main tool are vacuum expectation values of ordinary operator products. But, beside some structural results, the most prominent of them, the Wightman reconstruction theorem, has not led to a computationally useful tool,[30] apart from low dimensional models.[1]

Both of the constructions assume the following.

(1) The existence of field operators, and operators built from them, the Hermitian of which constitute the observables of the theory.
(2) The existence of a positive, linear, normalized functional, a state which allows the calculation of vacuum expectation values and a statistical interpretation.

There are further requirements necessary, but even the given ones are questionable. It is well known from C^*-algebra theory, that representations exist in the thermodynamic limit, in which no field operators do exist, but only in some larger closure of the algebra.

We will remain with the first assumption, but it tends to weaken the second one. It is mainly positivity which might be questioned in this context. QFT deals already with indefinite structures in describing interactions, e.g., Faddeev–Popov ghosts. Additionally we know, that on orthogonal Clifford algebras, which are related to fermionic fields, there are no nonsingular forms in dimensions higher than two (generators).[22,31]

It is clear that a statistical interpretation breaks down without positivity. But such an interpretation is not called for in the interaction region. Only in- and out-going states, which should have a particle interpretation, have to bear this feature. This idea was already proposed by Heisenberg.

However, we will go a step further and propose a geometric notation, which is valid in positive, indefinite and even singular cases. This makes it possible to decide afterwards in which situations positivity may be obtained and a statistical interpretation is meaningful; see Ref. 9. The Dirac theory was already treated within this method.[32]

We will treat fermionic quantum fields, but give in section VI also an example for a boson-fermion coupling theory, which can be developed straightforwardly.

Our starting point is the quantization rule imposed on field operators,

$$\{\psi_{I_1},\psi_{I_2}\}_+ = A_{I_1 I_2}, \tag{3.1}$$

where ψ_I is a ''super field'' defined as

$$\psi_I = \psi_{\Lambda K} = \begin{cases} \psi_K & \Lambda = 1, \\ \psi_K^\alpha & \Lambda = 2. \end{cases} \tag{3.2}$$

The anti-involution α is taken, e.g., to be the Hermitian or charge conjugation. $A_{I_1 I_2}$ is a symmetric c-number function of space-time and algebraic freedoms which all together are summarized in the multi-index $I=\{\Lambda,\alpha,\ldots\}$.

The relation (3.1) can be seen as the defining relation of a Clifford algebra $Cl(\langle\psi\rangle,A)$, when we identify the field operators as generators. To be able to give an algebraic basis of this Clifford algebra, we need an ordering. Let $\mathscr{P}$ denote such an ordering, which acts on the indices of the field operators. Clifford monomials in the field operators are then given by

$$e_n := \mathscr{P}(\psi_{I_1},\ldots,\psi_{I_n}), \tag{3.3}$$

where $\mathscr{P}$ might be, e.g., time-ordering, antisymmetry for equal times or normal-ordering w.r.t. a vacuum state. The linear space underlying $Cl(\langle\psi_I\rangle,A)$ is thus the linear span of the e_n Clifford monomials,

$$FCl(\langle\psi_I\rangle,A) \simeq \langle e_n\rangle. \tag{3.4}$$

As we emphasized in section II, it is of the utmost importance to have a graded space underlying those algebras, which leads to the Clifford algebra of multivectors of definition 7. In quantum field theory this space is obtained from a Grassmann algebra by introducing the anticommuting Schwinger sources j.[4] Indeed, due to the relation $j_I^2=0$ these are Grassmann generators. Assuming that the indices are from the same set as in the case of the ψ_I's, we build the Grassmann algebra over V with $V=\langle j_I\rangle$,

$$\wedge(\langle j_I\rangle) = \wedge(V). \tag{3.5}$$

Of course, we can identify the spaces underlying $Cl(\langle\psi_I\rangle,A)$ and $\wedge(\langle j_I\rangle)$ by means of Chevalley's identification,

$$FCl(\langle\psi_I\rangle,A) \simeq F\wedge(\langle j_I\rangle); \tag{3.6}$$

see section II. This is the graded space physicists make actual use of in their calculations. This remains true for path integrals also, as can be seen by their functional derivatives and the explicit occurrence of Schwinger source terms $\psi_I j_I$ in the exponent. It is therefore convenient to relate all entities to this space.

We know from Chevalley deformation, that the Clifford map ψ_I constitutes an endomorphism of $F\wedge(\langle j_I\rangle)$ which does not respect the grading. ψ_I may be decomposed in a grade lowering or annihilation operator ∂_I and a grade rising or creation operator $1/2 A_{I_1 I_2} j_{I_2}\wedge$,

$$\psi_I = \partial_I + \tfrac{1}{2} A_{I I_1} j_{I_1}\wedge, \tag{3.7}$$

where we omitted some superfluous factors. The ∂_I's are the canonical duals w.r.t. the Euclidean dual isomorphism. We have thus the commutation relations

$$\{j_I, j_K\}_+ = 0 = \{\partial_I, \partial_K\}_+, \quad \{J_I, \partial_K\}_+ = \delta_{IK}, \tag{3.8}$$

which results with (3.7) in the quantization condition (3.1),

$$\{\psi_I, \psi_K\}_+ = A_{IK}. \tag{3.9}$$

The next step towards a QFT generating functional is to project the Clifford monomials $e_n = \mathscr{P}(\psi_{I_1}, \ldots, \psi_{I_n})$ onto the graded space $F\wedge(\langle j_I \rangle)$. This is in fact a representation of the Clifford morphism obtained by lifting the ψ_I action to the whole algebra, as was described in section II. We write this projection as π: $FCl(V,A) \mapsto F\wedge(V)$,

$$\begin{aligned}\pi(e_n) &= \pi(\mathscr{P}(\psi_{I'_1}, \ldots, \psi_{I'_n})) \\ &= \langle \partial_{I_n} \wedge \cdots \wedge \partial_{I_1} \mathscr{P}(\psi_{I'_1}, \ldots, \psi_{I'_n}) \rangle_\pi j_{I_1} \wedge \cdots \wedge j_{I_n} \\ &= \langle 0 | \mathscr{P}(\psi_{I_1}, \ldots, \psi_{I_n}) | p \rangle j_{I_1} \wedge \cdots \wedge j_{I_n} = \rho_n(I_1, \ldots, I_n | p) j_{I_1} \wedge \cdots \wedge j_{I_n}.\end{aligned} \tag{3.10}$$

The index π parameterizes the linear normed functionals on $FCl(V,A)$, which is thereafter encoded in the bra $\langle 0|$ and ket $|p\rangle$ of the notation common to physicists. Observe that

$$\langle \partial_{I_n} \wedge \cdots \wedge \partial_{I_1} \mathscr{P}(\psi_{I_1}, \ldots, \psi_{I_n}) \rangle_\pi = \langle 0 | \mathscr{P}(\psi_{I_1}, \ldots, \psi_{I_n}) | p \rangle = \rho_n(I_1, \ldots, I_n | p) \tag{3.11}$$

is a c-number transition matrix element. If $|p\rangle$ is set equal to $|0\rangle$ we obtain vacuum expectation values which are known to be sufficient to describe the theory.

If a general element of $F\wedge(\langle j_I \rangle)$ is written down, this reads as

$$\begin{gathered}|\mathscr{P}(j,p)\rangle_F = \sum_{i=0}^{n} \frac{i^n}{n!} \rho_n(I_1, \ldots, I_n | p) j_{I_1} \wedge \cdots \wedge j_{I_n} |0\rangle_F, \\ \rho_n(I_1, \ldots, I_n | p) = \langle 0 | \mathscr{P}(\psi_{I_1}, \ldots, \psi_{I_n}) | p \rangle,\end{gathered} \tag{3.12}$$

and is thus exactly the definition of a fermionic generating functional in quantum field theory.[6,33] To emphasize the ''vector''-state character, in fact the left linear structure, we have added a functional ket $|0\rangle_F$ which is usually subjected to the relations

$${}_F\langle 0 | j_I = 0 = \partial_I |0\rangle_F. \tag{3.13}$$

These writings mimic in fact a linear form imposed on the algebra. We have thus given a Clifford algebraic account on quantum field theoretic generating functionals. The Clifford algebra involved was shown to be a Clifford algebra of multivectors built over a graded space $F\wedge(\langle j_I \rangle)$.

One may note that due to the explicit introduction of a basis $\{j_I\}$ the functional is also basis dependent, especially the matrix elements $\rho_n(I_1, \ldots, I_n | p)$. However, from our general considerations it is clear that the functional ''state'' $|\mathscr{P}(j,p)\rangle_F \in \wedge(\langle j_I \rangle)$ is an invariant object and is basis independent.

IV. TRANSITION FROM OPERATOR DYNAMICS TO GENERATING FUNCTIONALS

The most common case in quantum field theory is that an operator dynamics is given *a priori*. Furthermore, quantization calls for the Hamiltonian picture which breaks explicitly covariance.[34] This is not a problem as long as exact calculations are performed, since one is able to return to the manifest covariant theory.

Given a Hamiltonian, we can use the Heisenberg equation to write down the dynamics

$$i\dot{\psi}_I = [\psi, H]_-, \quad H = H[\psi]. \tag{4.1}$$

H is assumed to generate a one parameter family of automorphisms by integrating (4.1) to

$$\psi_I(t) = e^{iHt} \psi_I(0) e^{-iHt}. \tag{4.2}$$

It can be shown under the restrictive conditions of C^*-theory, this is always possible.

We know that generating functionals are constructed by lifting the ψ_I-action to $F\wedge(\langle j_I\rangle)$, so it is a natural question to ask for a Clifford morphism extension of (4.1). This will be the desired functional equation.

Since it is not quite clear which $H[\psi]$ integrates to a group action, we assume this explicitly. However, the space underlying a Clifford algebra can easily be made into a Lie-algebra by introducing the commutator product as a Lie-product. This has not to be confused with the Lie-algebras obtained as bi-vector sub-algebras under the commutator product.

A straightforward method to obtain derivatives of generating functionals is the Baker–Campbell–Hausdorff formula. If one writes the generating functional formally as

$$e^{\psi(t)} \simeq |\mathscr{P}(j,p)\rangle_F , \tag{4.3}$$

one has to calculate

$$\begin{aligned} i\partial_0 e^{\psi(t)} &= \sum_{k=0}^{\infty} [\psi,H]_k , \\ [\psi,H]_k &:= [[\psi,H]_{k-1},H]_- , \quad k \geqslant 1, \\ i\partial_0\psi &= [\psi,H]_0 = [\psi,H], \end{aligned} \tag{4.4}$$

which is quite tedious. Furthermore, one has to apply the BCH formula a second time to eliminate the field operators, which prolongates the calculations enormously. The result is then the functional equation

$$i\partial_0 |\mathscr{P}(j,p)\rangle_F = H[j,\partial]^{\mathscr{P}} |\mathscr{P}(j,p)\rangle_F , \tag{4.5}$$

with the functional Hamiltonian $H[j,\partial]^{\mathscr{P}}$. This Hamiltonian depends in its explicit form on the ordering implicit in $|\mathscr{P}(j,p)\rangle_F$.

We will circumvent these lengthy calculations using the opposite algebra $Cl^{op}(\langle\psi_I\rangle,A)$. To account for generating functionals, we reinterpret the field operators as Clifford endomorphisms on the space $F\wedge(\langle j_I\rangle)$ spanned by the Grassmann sources. If we allow only a left action, we can not reach every endomorphism on this space. However, the algebra $\wedge(\langle j_I\rangle)\otimes\wedge(\langle j_I\rangle^*)$ was shown to be isomorphic to $\mathrm{End}(F\wedge(\langle j_I\rangle))$. We have therefore to construct the opposite Clifford algebra to handle right actions. In theorem 2.14 we obtained the isomorphism

$$Cl^{op}(\langle\psi_I\rangle,A) \simeq Cl(\langle\psi_I\rangle,-A). \tag{4.6}$$

The decomposition into the creation and annihilation part reads then for the antisymmetric ordering as

$$\psi_I = \partial_I + \tfrac{1}{2} A_{II'} j_{I'}\wedge , \quad \psi_I^{op} = \partial_I - \tfrac{1}{2} A_{II'} j_{I'}\wedge . \tag{4.7}$$

ψ^{op} acts as an operator of right action from the left on the space $F\wedge(\langle j_I\rangle)$. Indeed, one finds

$$\partial_I = \tfrac{1}{2}(\psi_I + \psi_i^{op}), \quad j_I = A_{II'}(\psi_{I'} - \psi_{I'}^{op}), \tag{4.8}$$

where we assumed $AA = \mathbb{1}$, according to the Euclidean dual isomorphism. It is now an easy task to find the desired lifting of the Clifford morphism.

Let $g_n = \pi(e_n) = \rho_n(I_1,\ldots,I_n|p) j_{I_1}\wedge\cdots\wedge j_{I_n}$ be the projected monomial. The action of $H[\psi]$ on this monomial g_n may be defined as

$$i\dot{g}_n = [g_n, H[\psi]]_- = g_n H[\psi] - H[\psi] g_n , \tag{4.9}$$

$$= H[\psi^{op}(\hat{\cdot})] g_n - H[\psi] g_n . \tag{4.10}$$

A detailed investigation of the grading properties should then follow. However, to reproduce the results in the literature, the grade involution is suppressed. The action of the Hamiltonian on multivectors is thus defined to be

$$i\dot{g}_n=(H[\psi^{op}]-H[\psi])g_n. \tag{4.11}$$

Summing up the monomials yields the functional equation

$$i\partial_0|\mathscr{P}(j,a)\rangle_F=(H[\psi^{op}]-H[\psi])|\mathscr{P}(j,a)\rangle_F=H[j,\partial]^{\mathscr{P}}|\mathscr{P}(j,a)\rangle_F, \tag{4.12}$$

where $H[j,\partial]^{\mathscr{P}}$ is the desired functional Hamiltonian acting on the whole Dyson–Schwinger–Freese hierarchy encoded in the generating functional.

Instead of tedious calculations by means of the BCH formula, we have simply to decompose ψ and ψ^{op} into j-∂-parts and sum up the terms. In the above case of an antisymmetric ordering we have

$$H[j,\partial]^{as}=H[\psi^{op}]-H[\psi]=H[\partial-1/2Aj]-H[\partial+1/2Aj]. \tag{4.13}$$

The antisymmetric ordering was introduced in the second line due to the special form of the decomposition of ψ_I into annihilating and creating parts. In the case of an arbitrary ordering, we have to use the Clifford algebra of multivectors with an arbitrary bilinear form $B=G+F=A+F$ (G and A are synonyms). The resulting decompositions are (without factors)

$$\psi=\partial+Aj+Fj, \quad \psi^{op}=\partial-Aj+Fj, \tag{4.14}$$

due to theorem 2.15. We obtain as F-ordered or say normal-ordered functional Hamiltonian

$$H[j,\partial]^{\mathscr{N}}=H[\psi^{op}]-H[\psi]=H[\partial-Aj+Fj]-H[\partial+Aj+Fj], \tag{4.15}$$

and the normal-ordered functional equation

$$i\partial_0|\mathscr{N}(j,a)\rangle_F=(H[\psi^{op}]-H[\psi])|\mathscr{N}(j,a)\rangle_F=H[j,\partial]^{\mathscr{N}}|\mathscr{N}(j,a)\rangle_F. \tag{4.16}$$

The usefulness of this approach will be shown in the following sections by explicit calculations. Bosons can be treated in an analogous way using symplectic Clifford algebras.[8] We have not accounted for all factors i and 1/2 occurring for historical reasons in the definition of generating functionals; this will be done in the examples.

V. NONLINEAR SPINOR FIELD MODEL

It will be illuminating to treat every model in two versions. First the pure antisymmetric case which is the one-time limit of the time-ordered functionals and afterwards the more general normal-ordered case, where F is either the correct propagator of the theory or taken as a parameter.

We use a compact notation, where spinor operators and charge conjugate spinor operators are united,

$$\psi_I=\psi_{\Lambda K}=\begin{cases}\psi_K, & \Lambda=1,\\ \psi_K^c, & \Lambda=2.\end{cases} \tag{5.1}$$

It is useful to work with charge conjugate field operators, since they allow a compact notation of the field equations. We define our model as (summation and integration over repeated indices implicitly assumed).

$$\begin{aligned} i\partial_0\psi_I&=D_{II_1}\psi_{I_1}+gV_I^{\{I_1I_2I_2\}_{as}}\psi_{I_1}\psi_{I_2}\psi_{I_3},\\ \{\psi_{I_1}\psi_{I_2}\}_+&=A_{I_1I_2}, \end{aligned} \tag{5.2}$$

where D_{II_1} is a kinetic term and $V_I^{\{I_1I_2I_3\}_{as}}$ a constant vertex. The anticommutator is denoted by $A_{I_1I_2}$ and we require the identity $A_{I_1I_2}A_{I_2I_3}=\delta_{I_1I_3}$. For explicit expressions see, e.g., (6.10). But it is our intention not to give explicit expressions for D, V and A, since we want to demonstrate the model independency of our method.

The antisymmetry of $V_I^{\{I_1I_2I_2\}_{as}}$ can be pushed over to the field operator products. In this case one has to be careful since there are relations of the type

$$:\psi_{I_2}\wedge\psi_{I_3}\wedge\psi_{I_4}:=\psi_{I_2}\dot\wedge\psi_{I_3}\dot\wedge\psi_{I_4}=\psi_{I_2}\wedge\psi_{I_3}\wedge\psi_{I_4}+F_{I_2I_3}\psi_{I_4}-F_{I_2I_4}\psi_{I_3}+F_{I_3I_4}\psi_{I_2}, \tag{5.3}$$

which result from the normal-ordering.[10] We have thus carefully to distinguish the different wedge products, which ultimately lead to different vacua.[9] Literally this is done by the $:\cdots:$ notations used by physicists, but there without an algebraic theory behind. From the dynamics (5.2) we obtain the Hamilton operators

$$H[\psi]^\wedge=\frac{1}{2}A_{I_1I_3}D_{I_3I_2}\psi_{I_1}\wedge\psi_{I_2}+\frac{g}{4}A_{I_1I_5}V_{I_1}^{I_2I_3I_4}\psi_{I_1}\wedge\psi_{I_2}\wedge\psi_{I_3}\wedge\psi_{I_4} \tag{5.4}$$

and

$$H[\psi]^{\dot\wedge}=\frac{1}{2}A_{I_1I_3}D_{I_3I_2}\psi_{I_1}\dot\wedge\psi_{I_2}+\frac{g}{4}A_{I_1I_5}V_{I_1}^{I_2I_3I_4}\psi_{I_1}\dot\wedge\psi_{I_2}\dot\wedge\psi_{I_3}\dot\wedge\psi_{I_4}, \tag{5.5}$$

where the superscript wedges indicate the usage of the wedge or dotted wedge products.

A. Antisymmetric case

The antisymmetric case is connected with the normal wedge. Furthermore, the field operators corresponding to this ordering are given as

$$\begin{aligned}\psi&=\frac{1}{i}\partial_i-\frac{i}{2}A_{II_1}j_{I_1}\wedge\simeq\partial-Aj,\\ \psi^{op}&=\frac{1}{i}\partial_i+\frac{i}{2}A_{II_1}j_{I_1}\wedge\simeq\partial+Aj.\end{aligned} \tag{5.6}$$

Using the functional Heisenberg equation (4.12), we have to calculate

$$i\partial_0|\mathscr{A}(j,a)\rangle_F=E_{0a}|\mathscr{A}(j,a)\rangle_F=(H[\partial+Aj]-H[\partial-Aj])|\mathscr{A}(j,a)\rangle_F, \tag{5.7}$$

where $|a\rangle$ is assumed to be an energy eigen-''state'' here and in the sequel. Hence, we have to decompose the Clifford map ψ and ψ^{op} into creation and annihilation parts w.r.t. the space $F\wedge(\langle j_I\rangle)$. The first term of (5.4) is given as

$$T_1[\psi]^\wedge=\tfrac{1}{2}A_{I_1I_3}D_{I_3I_2}\psi_{I_1}\wedge\psi_{I_2}, \tag{5.8}$$

which results in functional form as

$$\begin{aligned}T_1[j,\partial]^\wedge&=T_1[\partial+Aj]-T_1[\partial-Aj]\\ &=\frac{1}{2}A_{I_1I_3}D_{I_3I_2}\left(\frac{1}{i}\partial_{I_2}+\frac{i}{2}A_{I_2I_2'}j_{I_2'}\right)\left(\frac{1}{i}\partial_{I_1}+\frac{i}{2}A_{I_1I_1'}j_{I_1'}\right)\\ &\quad-\left(\frac{1}{i}\partial_{I_2}-\frac{i}{2}A_{I_2I_2'}j_{I_2'}\right)\left(\frac{1}{i}\partial_{I_1}-\frac{i}{2}A_{I_1I_1'}j_{I_1'}\right)\\ &=\frac{1}{2}A_{I_2I_3}D_{I_3I_2}\left(\frac{1}{i^2}\{\partial_{I_2},\partial_{I_1}\}_+ +\frac{i^2}{4}A_{I_2I_2'}A_{I_1I_1'}\{j_{I_2'},j_{I_1'}\}_+ +4\frac{i}{2i}A_{I_2I_2'}j_{I_2'}\partial_{I_1}\right).\end{aligned} \tag{5.9}$$

Using the commutation relations of the j,∂ generators and the antisymmetry of AD, we obtain

$$T_1[j,\partial]^\wedge = -\tfrac{1}{2} A_{I_2 I_3} D_{I_3 I_1} 2 A_{I_2 I_2'} j_{I_2'} \partial_{I_1} = -D_{I_2' I_1} j_{I_2'} \partial_{I_1} = -D_{I_1 I_2} j_{I_1} \partial_{I_2}. \tag{5.10}$$

The interaction term,

$$T_2[\psi]^\wedge = \frac{g}{4} A_{I_1 I_5} V_{I_5}^{I_2 I_3 I_4} \psi_{I_1} \wedge \psi_{I_2} \wedge \psi_{I_3} \wedge \psi_{I_4}, \tag{5.11}$$

is decomposed as

$$\begin{aligned} T_2[j,\partial]^\wedge &= T_2[\partial + Aj] - T_2[\partial - Aj] \\ &= \frac{g}{4} A_{I_1 I_5} V_{I_5}^{I_2 I_3 I_4} \Bigg\{ \left(\frac{1}{i} \partial_{I_4} + \frac{i}{2} A_{I_4 I_4'} j_{I_4'} \right) \left(\frac{1}{i} \partial_{I_3} + \frac{i}{2} A_{I_3 I_3'} j_{I_3'} \right) \\ &\quad \times \left(\frac{1}{i} \partial_{I_2} + \frac{i}{2} A_{I_2 I_2'} j_{I_2'} \right) \left(\frac{1}{i} \partial_{I_1} + \frac{i}{2} A_{I_1 I_1'} j_{I_1'} \right) - \left(\frac{1}{i} \partial_{I_4} - \frac{i}{2} A_{I_4 I_4'} j_{I_4'} \right) \\ &\quad \times \left(\frac{1}{i} \partial_{I_3} - \frac{i}{2} A_{I_3 I_3'} j_{I_3'} \right) \left(\frac{1}{i} \partial_{I_2} - \frac{i}{2} A_{I_2 I_2'} j_{I_2'} \right) \left(\frac{1}{i} \partial_{I_1} - \frac{i}{2} A_{I_1 I_1'} j_{I_1'} \right) \Bigg\}. \end{aligned} \tag{5.12}$$

Since $T_2[j,\partial]^\wedge$ is an even function of the generators, we remain with terms of an odd number of j and ∂ generators. Furthermore we use the binomial theorem. This can be done, since new commutator terms does not arise due to antisymmetry of AV in $I_1,\ldots,I_4$. Accounting finally for a factor 2 from the two commutator terms, the result is

$$\begin{aligned} T_2[j,\partial]^\wedge &= g A_{I_4 I_5} V_{I_5}^{I_1 I_2 I_3} A_{I_4 I_4'} j_{I_4'} \partial_{I_3} \partial_{I_2} \partial_{I_1} + \frac{g}{4} A_{I_4 I_5} V_{I_5}^{I_1 I_2 I_3} A_{I_4 I_4'} j_{I_4'} A_{I_3 I_3'} j_{I_3'} A_{I_2 I_2'} j_{I_2'} \partial_{I_1} \\ &= g j_{I_1} V_{I_1}^{I_2 I_3 I_4} \left(\partial_{I_4} \partial_{I_3} \partial_{I_2} + \frac{1}{4} A_{I_4 I_4'} A_{I_3 I_3'} j_{I_4'} j_{I_3'} \partial_{I_2} \right). \end{aligned} \tag{5.13}$$

The functional Hamiltonian $H[j,\partial]^\wedge$ is the sum of $T_1^\wedge$ and $T_2^\wedge$. Therefore, the functional energy equation reads as

$$E_{a0} | \mathscr{A}(j,a) \rangle_F = \{ D_{I_1 I_2} j_{I_1} \partial_{I_2} + g j_{I_1} V_{I_1}^{I_2 I_3 I_4} (\partial_{I_3} \partial_{I_2} \partial_{I_1} + \tfrac{1}{4} A_{I_4 I_4'} A_{I_3 I_3'} j_{I_4'} j_{I_3'} \partial_{I_2}) \} | \mathscr{A}(j,a) \rangle_F, \tag{5.14}$$

in accordance with formula (3.98) in Ref. 21.

B. Normal-ordered case

The normal-ordered functional equation is related to the dotted wedge. It contains explicitly the propagator. This results in another functional Hamiltonian $H[j,\partial]^\wedge$. The functional energy equation reads as

$$i \partial_0 | \mathscr{N}(j,a) \rangle_F = E_{a0} | \mathscr{N}(j,a) \rangle_F = (H[\psi^{op}] - H[\psi]) | \mathscr{N}(j,a) \rangle_F, \tag{5.15}$$

where

$$\begin{aligned} \psi^{op} &= \frac{1}{i} \partial + \frac{i}{2} A_j \dot{\wedge} = \frac{1}{i} \partial + \frac{i}{2} Aj \wedge + iFj \wedge \simeq \partial + Aj + Fj, \\ \psi &= \frac{1}{i} \partial - \frac{i}{2} Aj \dot{\wedge} = \frac{1}{i} \partial - \frac{i}{2} Aj \wedge + iFj \wedge \simeq \partial - Aj + Fj, \end{aligned} \tag{5.16}$$

TABLE I. Resulting terms from $T_2[j,\partial]$ of the normal-ordered nonlinear spinor field functional equation.

$(\#\partial,\#Aj,\#Fj)$	$P_{k_1\cdots k_n}$	Factor	Resulting term
(3,1,0)	$\frac{4!}{3!}=4$	$4\frac{i}{2i^3}=-2$	$-A_{I_4I_4'}j_{I_4'}\partial_{I_3}\partial_{I_2}\partial_{I_1}$
(2,1,1)	$\frac{4!}{2!}=12$	$12\frac{i^2}{2i^2}=6$	$6A_{I_4I_4'}F_{I_3I_3'}j_{I_4'}j_{I_3'}\partial_{I_2}\partial_{I_1}$
(1,3,0)	$\frac{4!}{3!}=4$	$4\frac{i^3}{2^3i}=-\frac{1}{2}$	$-\frac{1}{2}A_{I_4I_4'}A_{I_3I_3'}A_{I_2I_2'}j_{I_4'}j_{I_3'}j_{I_2'}\partial_{I_1}$
(1,1,2)	$\frac{4!}{2!}=12$	$12\frac{i^3}{2i}=-6$	$-6A_{I_4I_4'}F_{I_3I_3'}F_{I_2I_2'}j_{I_4'}j_{I_3'}j_{I_2'}\partial_{I_1}$
(0,3,1)	$\frac{4!}{3!}=4$	$4\frac{i^4}{2^3}=\frac{1}{2}$	$\frac{1}{2}A_{I_4I_4'}A_{I_3I_3'}A_{I_2I_2'}F_{I_1I_1'}j_{I_4'}j_{I_3'}j_{I_2'}j_{I_1'}$
(0,1,3)	$\frac{4!}{3!}=4$	$4\frac{ii^3}{2}=2$	$2A_{I_4I_4'}F_{I_3I_3'}F_{I_2I_2'}F_{I_1I_1'}j_{I_4'}j_{I_3'}j_{I_2'}j_{I_1'}.$

are the Clifford maps related to the dotted wedge, written in terms of the undotted ones. This allows us to use the same j and ∂ generators. Since the operator dynamics is formally the same, we have once more to calculate T_1 and T_2. Now

$$\begin{aligned}\dot{T_1[j,\partial]\hat{}} &= T_1[\partial+Aj+Fj]-T_1[\partial-Aj+Fj]\\ &= \frac{1}{2}A_{I_1I_3}D_{I_3I_2}\left\{\left(\frac{1}{i}\partial_{I_2}+\frac{i}{2}A_{I_2I_2'}j_{I_2'}+iF_{I_2I_2'}j_{I_2'}\right)\left(\frac{1}{i}\partial_{I_1}+\frac{i}{2}A_{I_1I_1'}j_{I_1'}+iF_{I_1I_1'}j_{I_1'}\right)-\left(\frac{1}{i}\partial_{I_2}\right.\right.\\ &\quad\left.\left.-\frac{i}{2}A_{I_2I_2'}j_{I_2'}+iF_{I_2I_2'}j_{I_2'}\right)\left(\frac{1}{i}\partial_{I_1}-\frac{i}{2}A_{I_1I_1'}j_{I_1'}+iF_{I_1I_1'}j_{I_1'}\right)\right\}\end{aligned}\tag{5.17}$$

is no longer odd or even in j,∂ and we can only use antisymmetry arguments. Inspection of the equations shows that only odd numbers of A's can occur. These are

$$\begin{aligned}Aj\partial:&\ 2\frac{i}{2i}A_{I_2I_2'}j_{I_2'}\partial_{I_1},\\ AFjj:&\ 2\frac{i^2}{2}A_{I_2I_2'}F_{I_1I_1'}j_{I_2'}j_{I_1'}.\end{aligned}\tag{5.18}$$

With the same considerations as above, we obtain

$$\dot{T_1[j,\partial]\hat{}}=-D_{I_1I_2}j_{I_1}\partial_{I_2}+D_{I_1I_3}F_{I_3I_2}j_{I_1}j_{I_2}.\tag{5.19}$$

The interaction term T_2 is treated along the same lines. We have to use the polynomial theorem and the polynomial coefficients,

$$P_{k_1\cdots k_n}:=\frac{k!}{k_1!\cdots k_n!},\quad k=\sum k_n.\tag{5.20}$$

The resulting six terms occurring twice are given in Table I.

After juggling the indices and summarizing we obtain

$$\begin{aligned}\dot{T_2[j,\partial]\hat{}}=gV_{I_1}^{I_2I_3I_4}[&-j_{I_1}\partial_{I_4}\partial_{I_3}\partial_{I_2}+3F_{I_2I_2'}j_{I_1}j_{I_2'}\partial_{I_3}\partial_{I_4}\\ &-(3F_{I_3I_3'}F_{I_2I_2'}+\tfrac{1}{4}A_{I_3I_3'}A_{I_2I_2'})j_{I_1}j_{I_3'}j_{I_2'}\partial_{I_4}\\ &-(3F_{I_3I_3'}F_{I_2I_2'}+\tfrac{1}{4}A_{I_3I_3'}A_{I_2I_2'})F_{I_1I_1'}j_{I_1}j_{I_3'}j_{I_2'}j_{I_1'}].\end{aligned}\tag{5.21}$$

The normal-ordered functional energy equation obtained directly from the operator dynamics in one single step reads then as

$$\begin{aligned}E_{a0}|\mathcal{N}(j,a)\rangle_F &= H[j,\partial]\dot{\wedge}|\mathcal{N}(j,a)\rangle_F \\ &= \{D_{I_1I_2}j_{I_1}\partial_{I_2} - D_{I_1I_3}F_{I_3I_2}j_{I_1}j_{I_2} \\ &\quad + gV_{I_1}^{I_2I_3I_4}[j_{I_1'}\partial_{I_4}\partial_{I_3}\partial_{I_2} - 3F_{I_2I_2'}j_{I_1}j_{I_2'}\partial_{I_3}\partial_{I_4} \\ &\quad + (3F_{I_3I_3'}F_{I_2I_2'} + \tfrac{1}{4}A_{I_3I_3'}A_{I_2I_2'})j_{I_1}j_{I_3'}j_{I_2'}\partial_{I_4} \\ &\quad + (3F_{I_3I_3'}F_{I_2I_2'} + \tfrac{1}{4}A_{I_3I_3'}A_{I_2I_2'})F_{I_1I_1'}j_{I_1}j_{I_3'}j_{I_2'}j_{I_1'}]|\mathcal{N}(j,a)\rangle_F. \end{aligned} \tag{5.22}$$

$H[j,\partial]\dot{\wedge}$ is the normal-ordered functional Hamiltonian. The result agrees with (3.106) in Ref. 21.

Remark: In QFT the normal-ordered functional equation is obtained in two steps. Only after deriving the antisymmetric functional equation normal-ordering is performed explicitly via the nonperturbative Wick transformation,

$$\begin{aligned}|\mathcal{A}(j,a)\rangle_F &= e^{-1/2jFj}|\mathcal{N}(j,a)\rangle_F, \\ H[j,\partial]\dot{\wedge} &= e^{-1/2jFj}H[j,\partial]\wedge e^{+1/2jFj}.\end{aligned} \tag{5.23}$$

Observe that j commutes with $\exp(-1/2jFj)$, but not ∂, so that the transition can be given in terms of generators also,

$$\begin{aligned}\partial e^{-1/2jFj} &= e^{-1/2jFj}(\partial - Fj), \\ d &:= \partial - Fj, \\ H[j,\partial]\dot{\wedge} &= H[j,d]\wedge;\end{aligned} \tag{5.24}$$

compare with Ref. 32. This method requires explicit vertex regularization if the products $\wedge$ and $\dot{\wedge}$ are not explicitly treated as distinct $j_{I_1}\dot{\wedge}j_{I_2} = j_{I_1}\wedge j_{I_2} + F_{I_1I_2}$, as was shown in Ref. 10.

VI. SPINOR QUANTUM ELECTRODYNAMICS

We give spinor QED as an example for a boson-fermion coupling theory. We have to deal with a constrained theory, since spinor QED is a gauge theory. Especially on the quantum level this is a difficult task.[35] Hence in a first step we will eliminate the gauge freedom classically and perform the quantization afterwards. This is only possible in QED, because of the simple Abelian gauge coupling. In non-Abelian Yang–Mills Theories as QCD one has a nonlinear term in the Gauss' law, and a classical elimination leads to difficulties. However, this can be done using functional techniques.[35,36]

We start with the equations of classical spinor ED,

$$\begin{aligned}&\text{(i)}\quad \partial_\nu F^{\mu\nu} + \frac{ie_0}{2}\,\Psi C\gamma^\mu\sigma^2\Psi = 0, \\ &\text{(ii)}\quad (i\gamma^\mu\partial_\mu - m_0)\Psi + e_0 A_\mu\gamma^\mu\sigma^3\Psi = 0, \\ &\text{(iii)}\quad F^{\mu\nu} := \partial^\mu A^\nu - \partial^\nu A^\mu,\end{aligned} \tag{6.1}$$

where we have used the definition

$$\Psi = \Psi_{\alpha\Lambda} = \begin{cases}\Psi_\alpha, & \Lambda = 1, \\ \Psi^c_\alpha = C_{\alpha\beta}\overline{\Psi}_\beta, & \Lambda = 2\end{cases} \tag{6.2}$$

and omit the matrix indices. By definition the sigma matrices act on the super index Λ distinguishing the field and its charge conjugated field, which has to be treated as an independent quantity. The gamma matrices and the charge conjugation matrix C act on the spinor index α of the Ψ field.

We use the Coulomb gauge and set

$$\partial_k A^k = 0. \tag{6.3}$$

Furthermore we split the ''bosonic'' fields A and E into longitudinal and transverse parts,

$$\begin{gathered} A_k \equiv A_k^{tr}, \quad \text{because of} \quad \partial_k A^k = 0, \\ E_k = E_k^{tr} + E_k^l, E_k := -F^{0k} = F_{0k}. \end{gathered} \tag{6.4}$$

Furthermore we decompose the field equations into dynamic ones and the remaining constraint, the Gauss' law. Looking at (i) of (6.1) yields

$$\begin{gathered} \nu = 0: \ i\partial_0 E_k = -\tfrac{1}{2} e_0 \Psi C \gamma^k \sigma^2 \Psi + i(\partial_j \partial^k A^j - \partial_j \partial^j A^k), \\ \nu = k: \ \partial^k E_k = -i \frac{e_0}{2} \Psi C \gamma_0 \sigma^2 \Psi, \end{gathered} \tag{6.5}$$

and from the definition of E^k,

$$i\partial_0 A^{tr,k} = iE^k - i\partial_k A^0. \tag{6.6}$$

With the help of the current conservation the Gauss' law, $\nu = k$ from (6.5), can now be used to yield an expression for the A_0 field in terms of spinors,

$$A_0 = -i \frac{e_0}{2} \Delta^{-1} \Psi C \gamma_0 \sigma^2 \Psi. \tag{6.7}$$

The Δ^{-1} symbol is used for the integral kernel of the inverse of the Laplacian. For a further compactification of the notation we introduce a ''super field'' B_η in the ''bosonic'' fields also,

$$B_k^\eta := \begin{cases} A_k^{tr}, & \eta = 1, \\ E_k^{tr}, & \eta = 2. \end{cases} \tag{6.8}$$

The equations of motion reads then as

$$\begin{gathered} i\partial_0 \Psi_{I_1} = D_{I_1 I_2} \Psi_{I_2} + W^k_{I_1 I_2} B_k \Psi_{I_2} + U^{I_2 I_3 I_4}_{I_1} \Psi_{I_2} \Psi_{I_3} \Psi_{I_4}, \\ i\partial B_{K_1} = L_{K_1 K_2} B_{K_2} + J^{I_1 I_2}_{K_1} \Psi_{I_1} \Psi_{I_2}, \end{gathered} \tag{6.9}$$

with the definitions

$$\begin{gathered} I := \{\alpha, \Lambda, \mathbf{r}\}, \quad K := \{k, \eta, \mathbf{z}\}, \quad P^{tr} := 1 - \Delta^{-1} \boldsymbol{\nabla} \otimes \boldsymbol{\nabla}, \\ D_{I_1 I_2} := -(i\gamma_0 \gamma^k \partial_k - \gamma_0 m)_{\alpha_1 \alpha_2} \delta_{\Lambda_1 \Lambda_2} \delta(\mathbf{r}_1 - \mathbf{r}_2), \\ W^K_{I_1 I_2} := e_0 (\gamma_0 \gamma^k)_{\alpha_1 \alpha_2} \delta(\mathbf{r}_1 - \mathbf{r}_2) \delta(\mathbf{r}_1 - \mathbf{z}) \delta_{1\eta} \sigma^3_{\Lambda_1 \Lambda_2}, \\ U^{I_2 I_3 I_4}_{I_1} := -\frac{i}{8\pi} e_0^2 \left[(C\gamma_0)_{\alpha_2 \alpha_3} \delta_{\alpha_1 \alpha_4} \sigma^2_{\Lambda_2 \Lambda_3} \sigma^2_{\Lambda_1 \Lambda_4} \frac{\delta(\mathbf{r}_2 - \mathbf{r}_3) \delta(\mathbf{r}_1 - \mathbf{r}_4)}{|\mathbf{r}_1 - \mathbf{r}_2|} \right]_{\alpha s\{I_2 I_3 I_4\}}, \\ L_{K_1 K_2} := i\delta(\mathbf{z}_2 - \mathbf{z}_1) \delta_{k_1 k_2} \delta_{\eta_1 1} \delta_{\eta_2 2} + i\Delta(\mathbf{z}_1) \delta(\mathbf{z}_1 - \mathbf{z}_2) \delta_{k_1 k_2} \delta_{\eta_1 2} \delta_{\eta_2 1}, \\ J^{I_1 I_2}_K := -\tfrac{1}{2} e_0 P^{tr}(\mathbf{z} - \mathbf{r}_1) \delta(\mathbf{r}_1 - \mathbf{r}_2) (C\gamma^k)_{\alpha_1 \alpha_2} \delta_{2\eta} \sigma^2_{\Lambda_1 \Lambda_2}. \end{gathered} \tag{6.10}$$

We impose commutation relations on the Ψ and B fields which become thereby operators acting on a suitable state space. These states should belong to a Hilbert space, if the theory would be renormalized. The renormalization was discussed in Ref. 37 in a perturbative manner. As a postulate we introduce the equal time commutation relations,

$$\begin{aligned}
&\text{(i)}\quad \{\Psi_{I_1},\Psi_{I_2}\}_+^t := A_{I_1I_2} = C\gamma_0\sigma^1\delta(\mathbf{r}_1-\mathbf{r}_2),\\
&\text{(ii)}\quad [B_K,\Psi_I]_-^t := 0,\\
&\text{(iii)}\quad [B_{K_1},B_{K_2}]_-^t := C_{K_1K_2}.
\end{aligned} \tag{6.11}$$

Equation (i) of (6.11) is the canonical commutation relation for Ψ_α and $\overline{\Psi}_\alpha$, the Dirac adjoint field. In our notation with charge conjugate spinor operators, the somehow unusual $A_{I_1I_2}$ occurs. Because of the super field notation we have all four commutators belonging to the Ψ_α and $\overline{\Psi}_\alpha$ integrated in this relation.

Equation (ii) of (6.11) states that the boson fields are looked at as elementary fields. B_K is not a function of the Ψ_I fields. Equation (ii) (6.11) is usually postulated in QED.

Relation (iii) of (6.11) specifies the boson commutator. The special form of the c-number function $C_{K_1K_2}$ can be calculated as a consequence of requiring a consistent quantum field theory;[37] see below (6.13).

The set of equations (6.9) and (6.11) are the defining relations of Coulomb gauged quantized spinor ED. All previous steps are classical and only for convenience state a somehow self-consistent QFT. If one would quantize first and then try to eliminate Gauss' law, one would yield another QFT! The functional Hamiltonians (see below) would then have infinitely many terms.[38]

We obtain the Hamiltonian corresponding to the equations (6.9) as

$$\begin{aligned}
H(\Psi,B) = {}& \tfrac{1}{2}A_{I_1I_3}D_{I_3I_2}\Psi_{I_1}\Psi_{I_2} + \tfrac{1}{2}A_{I_1I_3}W^K_{I_3I_2}B_K\Psi_{I_1}\Psi_{I_2}\\
&+ \tfrac{1}{4}A_{I_1I_5}U^{I_2I_3I_4}_{I_5}\Psi_{I_1}\Psi_{I_2}\Psi_{I_3}\Psi_{I_4} + \tfrac{1}{2}C_{K_1K_3}L_{K_3K_2}B_{K_1}B_{K_2}.
\end{aligned} \tag{6.12}$$

There is no term $J_K^{I_1I_2}$ in the Hamiltonian. This term was eliminated with help of the commutator, (ii) of (6.11). Taking the time derivative and using the dynamics, one gets

$$C_{K_1K}W^K_{I_1I_2}\Psi_{I_2} = 2A_{I_1I}J_K^{II_2}\Psi_{I_2}. \tag{6.13}$$

This sort of consistency relation may be motivated by the physical assumption, that there is no asymmetry in the boson-fermion and fermion-boson interaction (action equals reaction, Newton). Furthermore a consistency consideration leads to the above relation.[37]

A. Antisymmetric case

We give the definitions of generating functionals for spinor QED. We introduce b sources and ∂^b duals,

$$\begin{aligned}
[b_{K_1},b_{K_2}]_- = 0 = [\partial^b_{K_1},\partial^b_{K_2}]_-,\\
[\partial^b_{K_1},b_{K_2}]_- = \delta_{K_1K_2},
\end{aligned} \tag{6.14}$$

analogous to the Schwinger sources. Therewith we define

$$\begin{aligned}
|.\mathscr{A}(a,j,b)\rangle_F &= \sum_{m,n=0}^{\infty}\frac{i^n}{n!m!}\,\tau_{nm}(I_1,\ldots,I_n,K_1,\ldots,K_m|a),\\
&\quad j_{I_1}\wedge\cdots\wedge j_{I_n}b_{K_1}\circ\cdots\circ b_{K_m}|0\rangle_F,\\
\tau_{nm}(I_1,\ldots,I_n,K_1,\ldots,K_m|a) &:= \langle 0|.\mathscr{A}(\psi_{I_1\cdots\psi_{I_n}}).\mathscr{S}(B_{K_1}\cdots B_{K_m})|a\rangle
\end{aligned} \tag{6.15}$$

and

$$|\mathscr{N}(a,j,b)\rangle_F = \sum_{m,n=0}^{\infty} \frac{i^n}{n!m!}\, \phi_{nm}(I_1,\ldots,I_n,K_1,\ldots,K_m|a),$$

$$j_{I_1}\dot{\wedge}\cdots\dot{\wedge} j_{I_n} b_{K_1}\circ\cdots\circ b_{K_m}|0\rangle_F,$$

$$\phi_{nm}(I_1,\ldots,I_n,K_1,\ldots,K_m|a) := \langle 0|\mathscr{N}(\psi_{I_1}\cdots\psi_{I_n})\mathscr{S}(B_{K_1}\cdots B_{K_m})|a\rangle, \tag{6.16}$$

where $\circ$ denotes the symmetric product of the b_K's, $\mathscr{A}$ the antisymmetrizer, $\mathscr{S}$ the symmetrizer and $\mathscr{N}$ the normal-ordering and $|a\rangle$ an energy eigen-''state.'' We assume, furthermore, that the j,∂ and b,∂^b generators commute. We do not consider bosonic normal-ordering. The field operators in the antisymmetric case are given as

$$\begin{aligned} \psi_I &= \frac{1}{i}\,\partial_I - \frac{i}{2}\,A_{II'}J_{I'}\wedge, \quad \psi_I^{op} = \frac{1}{i}\,\partial_I + \frac{i}{2}\,A_{II'}j_{I'}\wedge, \\ B_K &= \partial_K^b - \tfrac{1}{2}\,C_{KK'}b_{K'}\circ, \quad B_K^{op}\partial_K^b + \tfrac{1}{2}\,C_{KK'}b_{K'}\circ. \end{aligned} \tag{6.17}$$

Inserting these operators into (4.12) with the Hamiltonian (6.12) yields

$$\begin{aligned} E_{0a}|\mathscr{A}(a,j,b)\rangle = \{ & D_{I_1I_2}j_{I_1}\partial_{I_2} + W^K_{I_1I_2}j_{I_1}\partial_{I_2}\partial^b_K + L_{K_1K_2}b_{K_1}\partial^b_{K_2} + U^{I_2I_3I_4}_{I_1}j_{I_1} \\ & \times(\partial_{I_2}\partial_{I_3}\partial_{I_4} - \tfrac{1}{4}\,A_{I_3I_3'}A_{I_2I_2'}j_{I_2'}j_{I_3'}j_{I_4}) \\ & + J_K^{I_1I_2}b_K(\partial_{I_1}\partial_{I_2} + \tfrac{1}{4}\,A_{I_1I_1'}A_{I_2I_2'}j_{I_1'}j_{I_2'})\}|\mathscr{A}(a,j,b)\rangle, \end{aligned} \tag{6.18}$$

where the $A_{I_nI_m}$ terms stem from the field quantization of the fermions. Boson quantization terms would again occur, if the $J_K^{I_1I_2}$ would be eliminated in the above mentioned way. This result can be obtained from the normal-ordered equation by setting $F\equiv 0$, so we treat that case in more detail.

B. Normal-ordered case

The (fermion) normal-ordered equation is obtained by using the normal-ordered fermions,

$$\begin{aligned} \psi_I &= \frac{1}{i}\,\partial_I - \frac{i}{2}\,A_{II'}j_{I'}\wedge + iF_{II'}j_{I'}\wedge, \\ \psi_I^{op} &= \frac{1}{i}\,\partial_I + \frac{i}{2}\,A_{II'}j_{I'}\wedge + iF_{II'}j_{I'}\wedge; \end{aligned} \tag{6.19}$$

the boson operators remain unchanged. Before we proceed, we give the commutators of the fields, not displayed commutators, vanish. The bosons obey

$$\begin{aligned} [\partial^b_{K_1} \pm \tfrac{1}{2}\,C_{K_1K_1'}b_{K_1'}, \partial^b_{K_2} \pm \tfrac{1}{2}\,C_{K_2K_2'}b_{K_2'}]_- &= \mp C_{K_1K_2}, \\ [\partial^b_{K_1} \pm \tfrac{1}{2}\,C_{K_1K_1'}b_{K_1'}, \partial^b_{K_2} \mp \tfrac{1}{2}\,C_{K_2K_2'}b_{K_2'}]_- &= 0, \end{aligned} \tag{6.20}$$

while we have for the normal-ordered fermions,

$$\begin{aligned} \left\{\frac{1}{i}\,\partial_{I_1} \pm \frac{i}{2}\,A_{I_1I_1'}j_{I_1'} + iF_{I_1I_1'}j_{I_1'},\ \frac{1}{i}\,\partial_{I_2} \pm \frac{i}{2}\,A_{I_2I_2'}j_{I_2'} + iF_{I_2I_2'}j_{I_2'}\right\}_+ &= \pm A_{I_1I_2}, \\ \left\{\frac{1}{i}\,\partial_{I_1} \pm \frac{i}{2}\,A_{I_1I_1'}j_{I_1'} + iF_{I_1I_1'}j_{I_1'},\ \frac{1}{i}\,\partial_{I_2} \mp \frac{i}{2}\,A_{I_2I_2'}j_{I_2'} + iF_{I_2I_2'}j_{I_2'}\right\}_+ &= 0. \end{aligned} \tag{6.21}$$

We calculate directly the (fermion) normal-ordered functional Hamiltonian from $\dot{H[\Psi,B]^{\wedge}}$. Therefore we have to build

$$\dot{H[j,b,\partial,\partial_b]^{\wedge}} = H\left[\frac{1}{i}\,\partial_I + \frac{i}{2}\,A_{II'}j_{I'} + iF_{II'}j_{I'}, \partial^b_K + \frac{1}{2}\,C_{KK'}b_{K'}\right] - H\left[\frac{1}{i}\,\partial_I - \frac{i}{2}\,A_{II'}j_{I'} + iF_{II'}j_{I'}, \partial^b_K - \frac{1}{2}\,C_{KK'}b_{K'}\right]. \tag{6.22}$$

We have according to (6.12) four terms T_i. The T_1 term is of the same form as in the nonlinear spinor field model,

$$\dot{T_1[j,\partial]^{\wedge}} = D_{I_1I_2}j_{I_1}\partial_{I_2} - D_{I_1I_3}F_{I_3I_2}j_{I_1}j_{I_2}, \tag{6.23}$$

where the renormalization was discussed in Ref. 37. The term $\dot{T_2^{\wedge}}$ is given as

$$\begin{aligned}\dot{T_2[j,\partial]^{\wedge}} &= \tfrac{1}{2}\,A_{I_1I_3}W^K_{I_3I_2}[(\partial^b_K + \tfrac{1}{2}\,C_{KK'}b_{K'})(\cdots)(\cdots) - \left(\partial^b_K - \tfrac{1}{2}\,C_{KK'}b_{K'}\right)(\cdots)(\cdots)] \\ &= \tfrac{1}{2}\,A_{I_1I_3}W^K_{I_3I_2}[\partial^b_K\{(\cdots)(\cdots) - (\cdots)(\cdots)\} + \tfrac{1}{2}\,C_{KK'}b_{K'}\{(\cdots)(\cdots) + (\cdots)(\cdots)\}].\end{aligned} \tag{6.24}$$

The first term is just the same as before, while the second yields

$$2.\ \text{brace}\ \{\ \} = -2\partial_{I_1}\partial_{I_2} - 2\,\tfrac{1}{4}\,A_{I_1I_1'}A_{I_2I_2'}j_{I_1'}j_{I_2'} + 4F_{I_1I_1'}j_{I_1'}\partial_{I_2} - 2F_{I_1I_1'}F_{I_2I_2'}j_{I_1'}j_{I_2'}. \tag{6.25}$$

We may use (6.13) to obtain

$$\begin{aligned}\tfrac{1}{4}\,A_{I_1I_3}W^K_{I_3I_2}C_{KK'}b_{K'} &= -\tfrac{1}{4}\,A_{I_1I_3}C_{K'K}b_{K'}W^K_{I_3I_2} \\ &= -\tfrac{1}{2}\,A_{I_1I_3}A_{I_3I_4}J^{I_4I_2}_{K'}b_{K'} \qquad (6.26)\\ &= -\tfrac{1}{2}\,J^{I_1I_2}_K b_K, \qquad (6.27)\end{aligned}$$

and finally the result

$$\begin{aligned}\dot{T_2[j,\partial]^{\wedge}} &= W^K_{I_1I_2}[j_{I_1}\partial_{I_2} - F_{I_2I_2'}j_{I_1}j_{I_2'}]\partial^b_K + J^{I_1I_2}_k b_K \\ &\quad\times[\partial_{I_1}\partial_{I_2} - 2F_{I_1I_1'}j_{I_1'}\partial_{I_2} + (F_{I_1I_1'}F_{I_2I_2'} + \tfrac{1}{4}\,A_{I_1I_1'}A_{I_2I_2'})j_{I_1'}j_{I_2'}].\end{aligned} \tag{6.28}$$

The third term will be omitted, as it has the same structure as T_2 in the nonlinear spinor model. The last term runs along the lines of T_1 and has not changed at all from (6.18). Our result is then

$$\begin{aligned}E_{0a}|\mathcal{N}(a,j,b)\rangle = \{&D_{I_1I_2}j_{I_1}\partial_{I_2} - D_{I_1I_3}F_{I_3I_2}j_{I_1}j_{I_2} + W^K_{I_1I_2}[j_{I_1}\partial_{I_2} - F_{I_2I_2'}j_{I_1}j_{I_2'}]\partial^b_K \\ &+ J^{I_1I_2}_k b_K[\partial_{I_1}\partial_{I_2} - 2F_{I_1I_1'}j_{I_1'}\partial_{I_2} + (F_{I_1I_1'}F_{I_2I_1'} + \tfrac{1}{4}\,A_{I_1I_1'}A_{I_2I_2'})j_{I_1'}j_{I_2'}] \\ &+ U^{I_1I_2I_3}_J j_J[\partial_{I_1}\partial_{I_2}\partial_{I_3} - 3F_{I_3I_4}j_{I_4}\partial_{I_2}\partial_{I_1} + (3F_{I_3I_4}F_{I_2I_5} + \tfrac{1}{4}\,A_{I_3I_4}A_{I_2I_5})j_{I_4}j_{I_5}\partial_{I_1} \\ &- (F_{I_3I_4}F_{I_2I_5}F_{I_1I_6} + \tfrac{1}{4}\,A_{I_3I_4}A_{I_2I_5}A_{I_1I_6})j_{I_4}j_{I_5}j_{I_6}] + L_{K_1K_2}b_{K_1}\partial^b_{K_2}\}|\mathcal{N}(a,j,b)\rangle,\end{aligned} \tag{6.29}$$

which coincides with (8.59) in Ref. 21 and (63) in Ref. 37. We have succeeded in deriving the normal-ordered functional equation in one single step by Clifford algebraic considerations. No normal-ordering problems, as a vertex normal ordering, was necessary. No singular intermediate equation has occurred. And as a benefit, the calculation was considerable short.

VII. CONCLUSION

We have identified Clifford geometric algebras of multivectors as the appropriate mathematical tool in modeling quantum field theoretic generating functions. These algebras were studied in section II. Great emphasis was made on the connection of the spaces underlying the Clifford and Grassmann algebras, which induce the multivector structure in a Clifford algebra. This structure is ubiquitously used in physics, but not inherent in an ordinary Clifford algebra. However, if a Clifford algebra is constructed from generators and relations, one has implicitly imposed an ordering, which results in the explicit multivector structure. Only the Clifford algebra of multivectors can sytematically account for these situations and uncover the role played by the propagator. The propagator or say antisymmetric part of the contraction directly parameterizes the vacua underlying the theory as was shown in Ref. 9.

The third section develops a complete algebraic approach to quantum field theoretic generating functionals by means of Clifford algebras of multivectors. It was shown that the field operators can be seen as Clifford maps acting on the Grassmann algebra built from the Schwinger source terms. The vacuum was shown to be involved in the definition of matrix elements. This result remains true for path integrals, which rely on the same generating functionals. However, in the very compact notation of path integrals no direct access to these structures is possible.

In section IV we gave the transition from operator dynamics to generating functionals. This transition was motivated by the Heisenberg equation. However, there is no unique lifting of such an expression. To be able to reproduce the results in literature, we chose the opposite action, not the right action. Further investigations on that topic seem to be necessary. The last sections, V and VI, gave examples, which showed the easiness and computational advantages of this method. A nonlinear spinor field was treated in the antisymmetric and the normal-ordered case. The same was done for spinor QED as an example for a boson-fermion coupling theory. The most remarkable result is that it was possible to derive normal-ordered generating functionals directly, that is in one single step, from the operator dynamics. Furthermore, this showed that the algebraic theory has full control over the involved orderings and thereby over the involved vacuum states. This was investigated in detail in Ref. 9. The method was already applied to Dirac theory with some success.[32]

ACKNOWLEDGMENTS

Two long e-mail letters from Professor Z. Oziewicz, which improved the semantics and clarity of section II, were a valuable help in preparing this work. The support of Mrs. Ursula Wieland is gratefully acknowledged. The author would like to thank the referee for pointing out the importance of Ref. 7 for this work.

[1] J. Glimm and A. Jaffe, *Quantum physics. A Functional Integral Point of View* (Springer-Verlag, Berlin, 1981).
[2] R. J. Rivers, *Path Integral Methods in Quantum Field Theory* (Cambridge University Press, Cambridge, 1988).
[3] F. J. Dyson, ''The radiation theories of Tomonaga, Schwinger and Feynman,'' Phys. Rev. **75**, 486–502 (1949); ''The *S*-matrix in electrodynamics,'' *ibid.* **75**, 1736–1755 (1949); ''Divergence of perturbation theory in quantum electrodynamics,'' *ibid.* **85**, 631–632 (1952).
[4] J. Schwinger, ''The theory of quantized fields I,'' Phys. Rev. **82**, 914–927 (1951); ''The theory of quantized fields II,'' *ibid.* **91**, 713–728 (1953); ''The theory of quantized fields III,'' *ibid.* **91**, 728–740 (1953); ''The theory of quantized fields IV,'' *ibid.* **92**, 1283–1299 (1953); ''The theory of quantized fields V,'' *ibid.* **93**, 615–628 (1954); ''The theory of quantized fields VI,'' *ibid.* **94**, 1362–1384 (1954).
[5] E. Freese, ''Gebundene Teilchen und Streuprobleme in der Quantenfeldtheorie,'' Z. Naturforsch. **8a**, 776–790 (1951).
[6] D. Lurié, *Particles and Fields* (Interscience, New York, 1968).
[7] C. Doran, D. Hestenes, F. Sommen, and N. Van Acker, ''Lie groups as spin groups,'' J. Math. Phys. **34**, 3642–3669 (1993).
[8] A. Crumeyrolle, ''Orthogonal and symplectic Clifford algebras, spinor structures,'' *Mathematics and its Applications* (Kluwer, Dordrecht, 1990).
[9] B. Fauser, ''Clifford geometric parameterization of inequivalent vacua,'' preprint (hepth/9710047).
[10] B. Fauser, ''Vertex normal ordering as a consequence of nonsymmetric bilinearforms in Clifford algebras,'' J. Math. Phys. **37**, 72–83 (1996).
[11] C. Chevalley, *The Algebraic Theory of Spinors* (Columbia University Press, New York, 1954).
[12] M. Riesz, ''Clifford numbers and spinors,'' *The Institute of Fluid Dynamics and Applied Mathematics*, Lecture Series No. 38, University of Maryland, 1958, edited by E. F. Bolinder and P. Lounesto (Kluwer, Dordrecht, 1993), reprinted as a facsimile.
[13] P. Lounesto, *Clifford Algebras and Spinors* (Cambridge University Press, Cambridge, 1997).

[14] B. Fauser, ''Clifford algebraic remark on the Mandelbrot set of two-component number systems,'' Adv. Appl. Clifford Algebr. **6**, 1–26 (1996).
[15] M. Haft, ''Konjugationen und diskrete symmetrien,'' Thesis, Ludwig-Maximilian Univ., München, 1996.
[16] Z. Oziewicz, ''From Grassmann to Clifford,'' in *Clifford Algebras and Their Application in Mathematical Physics, I*, Workshop held at Canterbury, UK, edited by J. S. R. Chisholm and A. K. Common (Kluwer, Dordrecht, 1986), pp. 245–256.
[17] B. Fauser and H. Stumpf, ''Positronium as an example of algebraic composite calculations,'' in *Proceedings of the International Conference on the Theory of the Electron*, Advances in Applied Clifford Algebras, Cuautitlan, Mexico 1995, edited by J. Keller and Z. Oziewicz, 1997, Vol. 7(S), pp. 399–418.
[18] B. Fauser, ''Clifford-algebraische Formulierung und Regulatität der Quantenfeldtheorie,'' Thesis, Tübingen Univ., 1996.
[19] R. Ablamowicz and P. Lounesto, ''On Clifford algebras of a bilinear form with an anti-symmetric part,'' in *Clifford Algebras with Numeric and Symbolic Computations*, edited by R. Ablamowicz, P. Lounesto, and J. M. Parra (Birkhäuser, Boston, 1996), pp. 167–188.
[20] Z. Oziewicz, ''Clifford algebra of multivectors'' in Ref. 17, pp. 467–486.
[21] H. Stumpf and Th. Borne, *Composite Particle Dynamics in Quantum Field Theory* (Vieweg, Braunschweig, 1994).
[22] A. J. Hahn, *Quadratic Algebras, Clifford Algebras and Arithmetic Witt Groups* (Springer-Verlag, Berlin, 1994).
[23] I. M. Benn and R. W. Tucker, *Introduction to Spinors and Geometry with Applications in Physics* (Adam Hilger, Bristol, 1987).
[24] H. Saller, ''Quantum algebras I,'' Nuovo Cimento B **108**, 603–630 (1993); ''Quantum algebras II,'' *ibid.* **109**, 255–280 (1994).
[25] B. Fauser, ''Hecke algebras as subalgebras of Clifford geometric algebras of multivectors,'' preprint (q-alg/9710020).
[26] H. Umezawa and A. Visconti, ''General theory of propagators,'' Nuovo Cimento **1**, 1079–1103 (1955).
[27] M. K. Polivanov, ''On a new derivation of the equations for the Green's functions in quantum electrodynamics,'' Dokl. Akad. Nauk SSSR **100**, 1061–1063 (1955) (in Russian).
[28] L. D. Faddeev and A. A. Slavnov, *Gauge Fields—Introduction to Quantum Theory* (Benjamin Cummings, New York, 1980).
[29] A. S. Wightman, ''Quantum field theory in terms of vacuum expectation values,'' Phys. Rev. **101**, 860–866 (1956).
[30] R. Haag, *Local Quantum Physics* (Springer-Verlag, Berlin, 1992).
[31] M. A. Knus, *Quadratic and Hermitian Forms Over Rings* (Springer-Verlag, Berlin, 1991).
[32] B. Fauser, ''Dirac theory from a field theoretic point of view,'' in *Clifford Algebras and Their Application in Mathematical Physics, IV*, Workshop held at Aachen, edited by K. Habetha (Kluwer, Dordrecht, in press).
[33] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).
[34] P. A. M. Dirac, ''Lectures on quantum field theory,'' Belfer Graduate School of Science, Yeshiva University (New York, 1966).
[35] H. Stumpf and W. Pfister, ''Resolution of constraints and gauge equivalence in algebraic Schrödinger representation of quantum electrodynamics,'' Z. Naturforsch. **51a**, 1045–1066 (1996).
[36] H. Stumpf and W. Pfister, ''Algebraic Schrödinger representation of quantum chromodynamics in temporal gauge and resolution of constraints,'' Z. Naturforsch. **52a**, 220–240 (1997).
[37] H. Stumpf, B. Fauser, and W. Pfister, ''Composite particle theory in quantum electrodynamics,'' Z. Naturforsch. **48a**, 765–776 (1993).
[38] T. D. Lee and C. N. Yang, ''Theory of charged vector mesons interacting with the electromagnetic field,'' Phys. Rev. **128**, 885–898 (1962).

Maximal sign-invariants of quasilinear parabolic equations with gradient diffusivity

Victor A. Galaktionov[a)]
Department of Mathematics, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom and Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya Sq. 4, 125047 Moscow, Russia

Sergey A. Posashkov[b)]
Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya Sq. 4, 125047 Moscow, Russia

(Received 15 October 1997; accepted for publication 10 April 1998)

We consider quasilinear parabolic equations with gradient diffusivity $u_t = \mathrm{div}(|\nabla u|^\sigma \nabla u) + f(u)$, $x \in \mathbf{R}^N$, $t>0$, where $\sigma \neq -1$ is a fixed constant and $f(u)$ is a given smooth function. We also study quasilinear parabolic equations with a gradient-dependent coefficient $u_t = h(|\nabla u|)\Delta u + f(u)$, with a smooth function $h(p)$. For both classes of equations we derive first-order sign-invariants, i.e., first-order operators preserving their signs on the evolution orbits $\{u(\cdot,t), t>0\}$. We give a complete description of (maximal) sign-invariants of prescribed structures. As a consequence, we construct new exact solutions of some quasilinear equations. © *1998 American Institute of Physics.* [S0022-2488(98)01609-0]

Dedicated to Professor S. Kurdyumov on his 70th birthday.

I. INTRODUCTION

We consider quasilinear parabolic equations with gradientlike diffusivity

$$u_t = \mathrm{div}(|\nabla u|^\sigma \nabla u) + f(u), \quad x \in \mathbf{R}^N, \quad t>0, \tag{1.1}$$

where $\sigma \neq -1$ is a fixed constant and $f(u)$ is a sufficiently smooth function to be prescribed later. It is well-known that such types of diffusionlike operators occur, for instance, in the theory of non-Newtonian liquids and in some turbulence problems, see Ref. 1, and references therein. We also study quasilinear parabolic equations with a general nonlinearity in the diffusion term of the gradient-dependent type

$$u_t = h(|\nabla u|)\Delta u + f(u), \tag{1.2}$$

with a sufficiently smooth function $h(p)>0$. Such nonlinearities are widely used in mean curvature flow equations related to differential geometry and to problems of motion of a closed hypersurface in $\mathbf{R}^N$ by its mean curvature, see Refs. 2 and 3. Such equations are important in phase transition phenomena.[4] The one-dimensional Eq. (1.1) ($N=1$) corresponds to the function $h(p) = (\sigma+1)p^\sigma$ in (1.2). As a typical example we consider the equation

$$u_t = \frac{u_{xx}}{1+(u_x)^2} - \frac{1}{u}, \quad x \in \mathbf{R}, \quad t>0, \tag{1.3}$$

which describes (after a suitable surface parametrization) the evolution of cylindrically symmetric hypersurfaces moving by mean curvature in $\mathbf{R}^3$, see Refs. 5 and 6. On the other hand, a one-dimensional quasilinear parabolic equation with both gradient-dependent coefficients

$$u_t = \frac{u_{xx}}{1+(u_x)^2} + f(u)(1+(u_x)^2)^{1/2},$$

a)Fax: +44 (1225) 826 492, Electronic mail: vag@maths.bath.ac.uk
b)Fax: (7-095) 972-0737, Electronic mail: pos@spp.keldysh.ru

 © 1998 American Institute of Physics

is interpreted[7] as an equation of motion for the graph of u

$$\Gamma_t=\{(x,y):\quad y=u(x,t),\quad x\in\mathbf{R}\},$$

with the equation

$$V=k+f(y),$$

where $V=u_t/(1+(u_x)^2)^{1/2}$ denotes the upward velocity of Γ_t and $k=u_{xx}/(1+(u_x)^2)^{3/2}$ is the upward curvature of Γ_t. This is an example of a curve shortening equation with driving force term $f(y)$. Blow-up of the derivative u_x for a more general equation of such type

$$u_t=a(u_x)u_{xx}+f(u)g(u_x),$$

was studied in Ref. 7.

Let us describe the main techniques of the present paper. We introduce a general first-order Hamilton–Jacobi (HJ) operator of the form

$$H(u)=H(u_t,\nabla u,u),$$

where the function $H(\cdot)$ is assumed to be smooth enough.

We say that the operator (HJ) is a sign-invariant of the given nonlinear parabolic equation if it preserves both signs, $\geqslant 0$ and $\leqslant 0$, on its solutions $u(x,t)$. This means that

$$H(u)\geqslant 0(\leqslant 0)\ \text{ in }\ \mathbf{R}^N\quad \text{for } t=0\ \Rightarrow\ H(u)\geqslant 0(\leqslant 0)\ \text{ in }\ \mathbf{R}^N\quad \text{for } t>0. \tag{1.4}$$

These properties are understood in the sense of a formal application of the Maximum Principle for linear parabolic equations, see details in Ref. 8.

Some ideas of sign invariants were originated from the qualitative theory of nonlinear parabolic equations where similar barrier techniques (one-sided as usual) play a fundamental role in the study of existence, uniqueness, differential, and asymptotic properties of wide classes of solutions. We refer to Refs. 8 and 9 where these questions are discussed in detail for other classes of quasilinear parabolic equations, and necessary references are available.

In this paper we present the results on the structure of the first-order sign-invariants admitted by the equations with gradient diffusivity. We study **maximal** sign-invariants describing all possible sign-invariants of a prescribed structure. We also construct the corresponding sets of exact solutions (i.e., those which can be expressed in terms of finite-dimensional dynamical systems) using the zero-invariance property of sign-invariants. Namely, under assumption (1.4) the operator (HJ) is also a zero-invariant of the evolution equation, i.e., there holds

$$H(u)=0\quad \text{for } t=0\ \Rightarrow\ H(u)=0\quad \text{for } t>0. \tag{1.5}$$

This is a differential constraint, a well-known approach for constructing exact solutions of nonlinear evolution equations, see Ref. 8. We claim that under certain restrictions a consistent differential constraint for a quasilinear second-order parabolic equation can be extended to a sign-invariant which describes properties of a much wider (than purely exact ones) class of solutions $u(x,t)$.

In Sec. II we consider sign-invariants of the form $u_t-\psi(u)$ and the corresponding exact solutions. Sections III and IV are devoted to the sign-invariants of the form $u_r-g(r)F(u)$, $r=|x|$, in the class of spatially symmetric solutions $u=u(r,t)$.

Sign-invariants for quasilinear parabolic equations can give new exact finite-dimensional solutions.[8,9] On the other hand, the corresponding differential inequalities $H(u)\geqslant 0(\leqslant 0)$ can define a generalized property of B-convexity (concavity) of solutions preserved in time,[10] which is treated as convexity (concavity) relative to the corresponding family B of exact solutions. Under certain hypotheses on initial-boundary data and the coefficients of the equation, the differential inequalities can allow us to derive some *a priori* bounds on more general solutions of the parabolic equation under consideration.

II. NONSTATIONARY AUTONOMOUS MAXIMAL SIGN-INVARIANT

We are looking for a sign-invariant for the general Eq. (1.2) of the form

$$H(u)\equiv u_t-\psi(u), \tag{2.1}$$

with a smooth function $\psi(u)$ depending on the coefficients of the equation. The structure of operator (2.1) comes from the conditions of ψ-criticality of solutions to quasilinear parabolic equations.[11] Given the particular structure (2.1) of the sign-invariants depending on a single unknown function $\psi(u)$ we perform our further research in terms of the following.

Definition 2.1: The set

$$\mathscr{H}_{\max}=\{H:\ \exists\psi \text{ such that } (2.1) \text{ is a sign-invariant of } (1.2)\}$$

is said to be a maximal sign-invariant of equation (1.2) of the type (2.1).

We begin with a preliminary description of the maximal sign-invariant.

Proposition 2.2: (2.1) is a sign-invariant to (1.2) if and only if the functions $h(p)$ and $\psi(u)$ satisfy the identity for all $(u,p)\in\mathbf{R}\times\mathbf{R}_+$

$$\mathscr{T}(u,p)=f'\psi-f\psi'+\frac{h'(p)}{h(p)}\,p(\psi-f)\psi'+h(p)p^2\psi''\equiv 0. \tag{2.2}$$

Proof: It is not difficult to calculate that the function $J=u_t-\psi(u)$, assumed to be sufficiently smooth, solves a linear parabolic equation of the form

$$J_t=\mathscr{L}J+\mathscr{T}, \tag{2.3}$$

where $\mathscr{L}$ is a second-order elliptic operator and $\mathscr{T}$ is as given in (2.2). Therefore, by the Maximum Principle, the function $J(\cdot,t)$ preserves its sign (both the positive and negative ones) given on the parabolic boundary of the domain if $\mathscr{T}=0$. See details in Ref. 8.

Let us now prove that

$$\mathscr{H}_{\max}=\{H(u)=u_t-\psi(u):\ \psi \text{ solves Eq. } (2.2)\},$$

is the maximal sign-invariant for Eq. (1.2). Indeed, assume for contradiction that there exist constants $\mu\in\mathbf{R}$, $\nu\in\mathbf{R}_+$ such that $\mathscr{T}(\mu,\nu)\neq 0$ and, say, >0. Then one can construct a smooth initial function $u_0\equiv u(x,0)$ such that $J(x,0)\equiv h(|\nabla u_0|)\Delta u_0+f(u_0)-\psi(u_0)=0$ in a neighborhood of the origin $x=0$, and $u_0(0)=\mu$, $|\nabla u_0|(0)=\nu$. For instance, one can take a symmetric function $u_0(r)$, $r=|x-x_0|$, with an arbitrary fixed $x_0\in\mathbf{R}^N$, $x_0\neq 0$. Then $u_0(|x_0|)=\mu$, $u_0'(|x_0|)=\nu$ and the existence of such a function $u_0(r)$ follows from the standard theory of ordinary differential equations. As usual, for a formal application of the maximum principle we assume that such a u_0 can be continued beyond a neighborhood of the origin to provide us with a local-in-time solution $J(x,t)$ of Eq. (2.3). Then a standard formal application of the maximum principle, based on the interior regularity results for uniformly parabolic equations implies that $J(x,t)>0$ for all small $t>0$ and $x\approx 0$. Therefore, the negative sign, $J\leqslant 0$, is not preserved on all the evolution orbits.□

We now give a detailed complete description of the maximal sign-invariant and the corresponding exact solutions. It follows from (2.2) that the right-hand side, as a function of p only, belongs to the linear span

$$W=\mathscr{L}\left\{1,\frac{h'(p)}{h(p)}\,p,\ h(p)p^2\right\}. \tag{2.4}$$

Therefore we consider three cases.

Case 1: dim $W=3$. Then since all the three terms in (2.2) are linearly independent we arrive at the system

$$f'\psi-f\psi'=0,\quad (\psi-f)\psi'=0,\quad \psi''=0, \tag{2.5}$$

which is easily solved with respect to the function $\psi(u)$. We have

$$\psi(u)=au+b, \quad f(u)=\lambda(au+b), \tag{2.6}$$

where either $\lambda=1$ or $a=0$. In particular, if $\lambda=1$, i.e., $\psi(u)=f(u)=au+b$ we have that the equation

$$u_t=h(|\nabla u|)\Delta u+au+b \quad (a\neq 0),$$

admits the exact solution

$$u(x,t)=\frac{1}{a}\,[g(x)e^{at}-b],$$

where $g(x)$ is an arbitrary harmonic function. If $a=0$ then the equation

$$u_t=h(|\nabla u|)\Delta u+b,$$

has the exact solution

$$u(x,t)=g(x)+dt,$$

where g solves the elliptic equation

$$h(|\nabla g|)\Delta g+b-d=0 \ \text{ in } \mathbf{R}^N.$$

Case 2: dim $W=2$. Without loss of generality, we may assume that there exist constants $\alpha\neq 0$, $\beta\neq 0$ such that

$$\frac{h'}{h}\,p+\alpha hp^2+\beta=0. \tag{2.7}$$

Substituting $h'p/h$ from (2.7) into (2.2) we conclude that the functions $f(u)$ and $\psi(u)$ satisfy the following system of the ODE's (ordinary differential equation)

$$\psi''=\alpha(\psi-f\,)\psi', \quad \beta\psi\psi'=f'\psi+(\beta-1)f\psi'. \tag{2.8}$$

The second one is easily solved with respect to the function f:

$$f=A_0\psi^{1-\beta}+\psi. \tag{2.9}$$

Here and later on A_0, B_0,… denote free constants of integration.

Substituting (2.9) into the first equation and integrating once we get the first-order ODE

$$\psi'=B_0+\frac{\alpha A_0}{\beta-2}\,\psi^{2-\beta}, \quad \beta\neq 2, \tag{2.10}$$

$$\psi'=B_0-\alpha A_0\,\log\,\psi, \quad \beta=2. \tag{2.11}$$

The system (2.7), (2.9)–(2.11) defines a multidimensional set of Eq. (1.2) which admit sign-invariant (2.1). The structure of the corresponding exact solutions is calculated from the identity $u_t=\psi(u)$ so that it takes the form

$$u(x,t)=\Phi^{-1}(t+\rho(x)), \tag{2.12}$$

where Φ^{-1} is the inverse function to

$$\Phi(u)=\int_1^u dz/\psi(z), \tag{2.13}$$

and $\rho(x)$ is a smooth function to be determined later by substituting (2.12) into equation (1.2). We then arrive at the following result.

Theorem 2.3: *(i) If $\beta \neq 2$ in (2.7) then the equation*

$$u_t = \frac{(2-\beta)\Delta u}{C_0|\nabla u|^\beta + \alpha|\nabla u|^2} + A_0\psi^{1-\beta}(u) + \psi(u), \tag{2.14}$$

where ψ solves (2.10), admits sign-invariant (2.1) and the exact solution (2.12) where ρ solves an elliptic equation of the form

$$\Delta\rho + B_0|\nabla\rho|^2 + \frac{A_0C_0}{2-\beta}|\nabla\rho|^\beta = 0. \tag{2.15}$$

(ii) If $\beta = 2$ then the parabolic equation takes the form

$$u_t = \frac{\Delta u}{(C_0 + \alpha \log|\nabla u|)|\nabla u|^2} + \frac{A_0}{\psi(u)} + \psi(u), \tag{2.16}$$

with ψ satisfying (2.11), and ρ in (2.12) solves the semilinear elliptic equation

$$\Delta\rho + (\mu + \nu \log|\nabla\rho|)|\nabla\rho|^2 = 0, \quad \mu = B_0 + A_0C_0, \quad \nu = \alpha A_0. \tag{2.17}$$

We now begin to study properties of particular simpler solutions of such type.

Example 2.1: We consider the case $\alpha = 2$ and $\beta = 0$ in (2.7), where h solves

$$\frac{h'}{h}\, p = -2hp^2. \tag{2.18}$$

Then $h(p) = (1+p^2)^{-1}$ and (2.2) reads $\mathscr{T}(u,p) \equiv f'\psi - f\psi' + h(p)p^2[\psi'' - 2(\psi - f)\psi'] = 0$. This yields the system of the ODE's [cf. (2.8)]

$$f'\psi - f\psi' = 0, \quad \psi'' - 2(\psi - f\,)\psi' = 0.$$

The general solution of this system is $f(u) = \lambda\psi(u)$, $\lambda \in \mathbf{R}$, where ψ solves the ODE

$$\psi'' - 2(1-\lambda)\psi\psi' = 0.$$

This representation includes the equation with $f \equiv 0$, i.e., $\lambda = 0$, see a remark below.

We now consider the solution with $\lambda = 1/2$ so that $\psi(u) = 2f(u)$ where f solves $f'' - 2ff' = 0$ and hence

$$f'(u) - f^2(u) = \gamma, \tag{2.19}$$

γ being a free constant. We consider three cases.

(i) If $\gamma = 0$ in (2.19) then $f(u) = -1/u$. Integrating the ordinary differential equation

$$u_t = \psi(u) \equiv -\frac{2}{u},$$

we obtain the explicit solution [cf. (2.12)]

$$u_*(x,t) = (\tilde{\rho}(x) - 4t)^{1/2}. \tag{2.20}$$

Substituting (2.20) into the corresponding parabolic equation [which coincides with (1.3) in the one-dimensional case]

$$u_t = \frac{\Delta u}{1+|\nabla u|^2} - \frac{1}{u}, \tag{2.21}$$

we arrive at the following linear elliptic equation for $\tilde{\rho}$:

$$\Delta\widetilde{\rho}+2=0, \quad x\in\mathbf{R}^N. \tag{2.22}$$

Let us discuss a few properties of such symmetric explicit solutions, $u_*=u_*(r,t)$, $r=|x|$. Then (2.22) is easily integrated:

$$\begin{aligned}\widetilde{\rho}(r)&=-\frac{1}{N}\,r^2+\frac{c_0}{2-N}\,r^{2-N}+c_1, \quad N\geqslant 3,\\ \widetilde{\rho}(r)&=-\frac{r^2}{2}+c_0\,\log r+c_1, \quad N=2,\end{aligned} \tag{2.23}$$

where c_0 and c_1 are arbitrary constants. In the one-dimensional case $N=1$, (2.20) is obviously the exact self-similar solution ($c_0=c_1=0$ by translations in x and t)

$$u_*(x,t)=2\sqrt{-t}\left(1-\frac{\xi^2}{4}\right)^{1/2}, \quad \xi=\frac{x}{\sqrt{-t}}, \tag{2.24}$$

describing the disappearance (extinction) of the solution at the origin $x=0$ at the finite time $t=0$. One can see that (2.24) is defined in the shrinking domain $\{|x|<2\sqrt{-t},\ t<0\}$ and it is not a weak solution of the free boundary problem to (2.21) for $x\in\mathbf{R}$ in the standard sense of integration by parts.

On the other hand, if $N\neq 1$, then solution (2.20), (2.23) is not self-similar and has the form ($c_1=0$ by translation in t and $c_0=1$)

$$\begin{aligned}u_*(r,t)&=\left(-\frac{r^2}{N}-\frac{r^{2-N}}{N-2}-4t\right)^{1/2}, \quad N\geqslant 3,\\ u_*(r,t)&=\left(-\frac{r^2}{2}+\log r-4t\right)^{1/2}, \quad N=2.\end{aligned} \tag{2.25}$$

It describes extinction of the solution to the initial-boundary value problem (IBVP) on the sphere $\{|x|\equiv r_0=(N/2)^{1/N}\}$ at the moment

$$T=-\frac{1}{4(N-2)}\left(\frac{N}{2}\right)^{2/N}, \quad N\geqslant 3; \quad T=-\frac{1}{8}, \quad N=2. \tag{2.26}$$

The behavior as $t\to T^-$, $r\to r_0$, is asymptotically self-similar.

(ii) If $\gamma=1$ in (2.19) then $f(u)=\tan u$. By integrating the ODE $u_t=2\tan u$ we deduce that

$$u_*(x,t)=\arcsin(\widetilde{\rho}(x)e^{2t}), \quad \widetilde{\rho}=e^{2\rho}, \tag{2.27}$$

where ρ solves the elliptic Eq. (2.15) with $\beta=0$, $A_0'=-1/2$, $B_0=C_0=2$. By substituting into equation

$$u_t=\frac{\Delta u}{1+|\nabla u|^2}+\tan u, \tag{2.28}$$

we obtain that $\widetilde{\rho}(x)$ solves the elliptic equation

$$\Delta\widetilde{\rho}-\widetilde{\rho}=0, \quad x\in\mathbf{R}^N. \tag{2.29}$$

(iii) Consider now the last case $\gamma=-1$. Then, the parabolic equation takes the form

$$u_t=\frac{\Delta u}{1+|\nabla u|^2}+\frac{1+e^{2u}}{1-e^{2u}}, \tag{2.30}$$

which admits the exact solution

$$u(x,t)=\pm\log[\sqrt{e^{-4t}\widetilde{\rho}^2(x)-1}+e^{-2t}\widetilde{\rho}(x)],\quad \widetilde{\rho}=\tfrac{1}{2}e^{-2\rho}, \tag{2.31}$$

where ρ satisfies (2.15) and $\widetilde{\rho}(x)$ solves the elliptic equation

$$\Delta\widetilde{\rho}+\widetilde{\rho}=0. \tag{2.32}$$

Similarly, the equation

$$u_t=\frac{\Delta u}{1+|\nabla u|^2}+\frac{1-e^{2u}}{1+e^{2u}},$$

admits the exact solution

$$u(x,t)=\pm\log[\sqrt{e^{-4t}\widetilde{\rho}^2(x)+1}+e^{-2t}\widetilde{\rho}(x)],\quad \widetilde{\rho}=\tfrac{1}{2}e^{-2\rho},$$

and $\widetilde{\rho}$ satisfies (2.32). In particular, in the one-dimensional case we deduce the following 2π-periodic solution

$$u(x,t)=\pm\log[\sqrt{e^{-4t}\sin^2 x+1}+e^{-2t}\sin x].$$

Remark 1: In the case of Eq. (1.1), $N=1$, where $h'(p)p/h(p)=\sigma$, we obtain the following two independent ODE's

$$\psi''=0,\quad f'\psi-f\psi'+\sigma(\psi-f)\psi'=0.$$

This yields the linear sign invariant $u_t-(au+b)$ in the case $f(u)=c(au+b)^{\sigma+1}+(au+b)$. The corresponding heat Eq. (1.1) is well-known to admit an explicit solution in separate variables describing for $\sigma>0$ standing regionally localized blow-up patterns (the so-called *S*-regime of localization), see Ref. 12.

Remark 2: It follows from (2.8) that in the case of the pure gradient-diffusivity equation with $f\equiv 0$, there exists a nontrivial sign-invariant (2.1) provided that $\beta=0$. Then the system (2.8) reduces to a single ODE of the form $\psi''=\alpha\psi\psi'$. In particular, in the case (2.18), with $\alpha=2$, $\beta=0$ in (2.7), we obtain $\psi''-2\psi\psi'=0$ whence

$$\psi'-\psi^2=\gamma,\quad \gamma\in\mathbf{R}.$$

We then deduce that the equation

$$u_t=\frac{\Delta u}{1+|\nabla u|^2},$$

admits the exact solution (2.12) where $\rho(x)$ solves the elliptic equation [cf. (2.15)]

$$\Delta\rho+\gamma|\nabla\rho|^2-1=0.$$

In particular, there exist all the exact solutions (2.20), (2.24), (2.25), and (2.31) where t has to be replaced by $t/2$.

Case 3: dim $W=1$. It is easy to see that we can restrict out attention to the function $h(p)=p^{-2}$. Then identity (2.2) reduces to a single ODE of the form

$$(\psi'+f\psi-\psi^2)'=0. \tag{2.33}$$

Integrating it once we obtain

$$\psi'+f\psi-\psi^2=\gamma. \tag{2.34}$$

Given a smooth function $f(u)$ we fix an arbitrary solution $\psi(u)$ to (2.34). Then we obtain the structure (2.12) of the explicit solutions. Substituting into the parabolic equation

$$u_t=\frac{\Delta u}{|\nabla u|^2}+f(u), \tag{2.35}$$

and using (2.34) we deduce that $\rho(x)$ satisfies the following semilinear elliptic equation:

$$\Delta\rho+\gamma|\nabla\rho|^2=0, \quad x\in\mathbf{R}^N, \tag{2.36}$$

which reduces to the Laplace equation $\Delta p=0$ in $\mathbf{R}^N$ by setting $p(x)=e^{\gamma\rho(x)}$. Recall that due to the single Eq. (2.33) on the function ψ the above explicit solutions to (2.35) exist for arbitrary lower-order terms $f(u)$, i.e., the set of such quasilinear parabolic equations is *infinite-dimensional*. On the other hand, for $N=1$ via the well-known transformation[13] $x=\psi(y,t)$ where ψ is such that $u(\psi(y,t),t)\equiv y$, the one-dimensional Eq. (2.35) reduces to the linear heat equation with convection $\psi_t=\psi_{yy}-f(y)\psi_y$. In terms of ψ this implies a certain linear superposition property of solutions.

This completes the analysis of identity (2.2) on the subspace (2.4).

III. NONAUTONOMOUS STATIONARY SIGN-INVARIANT

In this section we consider sign-invariants of the stationary structure which is nonautonomous in the spatial variable $x\in\mathbf{R}^N$. We restrict our attention to a class of radial symmetric solutions $u=u(r,t)$, $r=|x|\geqslant 0$. In this class the first-order operator has the form

$$H_*(u)\equiv u_r-g(r)F(u) \quad \text{for } t>0, \tag{3.1}$$

with yet unknown smooth non-negative functions $g(r)\not\equiv$const and $F(u)\neq 0$ a.e. The idea of such sign-invariants goes to the gradient bounds introduced[14] for semilinear parabolic equations. Generalizations to the quasilinear gradient-dependent operators can be found in Ref. 15. We apply operator (3.1) to Eq. (1.1) written in the equivalent radially symmetric form (we suppose for convenience that $u_r>0$)

$$u_t=(u_r)^\sigma\left[(\sigma+1)u_{rr}+\frac{N-1}{r}\,u_r\right]+f(u), \tag{3.2}$$

with $\sigma\neq -1$.

Proposition 3.1: Operator (3.1) is a sign-invariant of (3.2) if and only if the functions $g(r)$ and $F(u)$ satisfy the identity

$$\begin{aligned}\frac{\mathscr{T}(r,u)}{(\sigma+1)F^{\sigma+1}}\equiv{}&(g^\sigma g')'(r)+2(\sigma+1)(g^{\sigma+1}g')(r)F'+g^{\sigma+3}(r)[FF''+\sigma(F')^2]\\&+g(r)\left(\frac{f}{F}\right)'\frac{F^{1-\sigma}}{(\sigma+1)}+g^\sigma(r)\,\frac{N-1}{r}\left[g'(r)+g^2(r)\,\frac{\sigma}{\sigma+1}\,F'-\frac{g(r)}{(\sigma+1)r}\right]\\&=0.\end{aligned} \tag{3.3}$$

Proof: The proof is straightforward, see the previous one in Sec. II. Indeed, direct computations lead to the parabolic Eq. (2.3) for the function $J=H_*(u)$ with an elliptic operator $\mathscr{L}$ and the lower order term $\mathscr{T}$ given in (3.3). □

We now give a description of the maximal sign-invariant $\mathscr{H}_{\max}$, i.e., the whole set of sign-invariants of the form (3.1) with nontrivial (see above) functions g and F.

3.1. Dimension: $N=1$. Here the Greek letters α,β,γ... denote arbitrary constants.

Theorem 3.2: *Let $N=1$. Then $\mathscr{H}_{\max}=\mathscr{H}_1\cup\mathscr{H}_2$ where the sets $\mathscr{H}_{1,2}$ are defined as follows: (i) $H_*\in\mathscr{H}_1$ if g satisfies the ODE*

$$(g^\sigma g')'=\beta g^{\sigma+1}g'+\gamma g^{\sigma+3}+\delta g, \quad \gamma=-\sigma\left[\frac{\beta}{2(\sigma+1)}\right]^2, \tag{3.4}$$

and F is a linear function

$$F(u)=-\frac{\beta u}{2(\sigma+1)}. \tag{3.5}$$

(ii) $H_*\in\mathscr{H}_2$ if g and F satisfy the ODE's

$$g^\sigma g'=\gamma g^{\sigma+2}+\delta, \tag{3.6}$$

$$(F^\sigma F')'+2\gamma(\sigma+1)F^{\sigma-1}F'+\gamma^2(\sigma+2)F^{\sigma-1}=0. \tag{3.7}$$

The corresponding function $f(u)$ is then given by

$$\text{(i)}\quad f(u)=\frac{2\delta(\sigma+1)^2}{\sigma\beta}\left(-\frac{\beta u}{2(\sigma+1)}\right)^{\sigma+1}, \tag{3.8}$$

$$\text{(ii)}\quad f(u)=\frac{\delta(\sigma+1)}{\gamma}F^{\sigma+1}F'+\rho F. \tag{3.9}$$

Proof of Theorem 3.2: For $N=1$, the right-hand side of (3.3), as a function of the space variable r, is a linear combination (with coefficients depending on u) of four terms which should be linearly dependent (otherwise no nontrivial sign-invariant exists). Therefore, we assume that

$$\alpha_1(g^\sigma g')'+\beta_1 g^{\sigma+1}g'+\gamma_1 g^{\sigma+3}+\delta_1 g=0. \tag{3.10}$$

We need to consider two cases

Case 1: $\alpha_1\neq 0$. Let $\alpha_1=-1$, i.e.

$$(g^\sigma g')'=\beta_1 g^{\sigma+1}g'+\gamma_1 g^{\sigma+3}+\delta_1 g. \tag{3.11}$$

Substituting (3.11) into (3.3) we obtain a linear combination of the form

$$I_1(u)g^{\sigma+1}g'+I_2(u)g^{\sigma+3}+I_3(u)g=0 \tag{3.12}$$

on the linear space $\mathscr{W}=\mathscr{L}\{g^{\sigma+1}g',\ g^{\sigma+3},\ g\}$. Then we have: (i) If $\dim\mathscr{W}=3$ then we get the system of ODE's for F and f

$$I_1\equiv\beta_1+2(\sigma+1)F'=0,\quad I_2\equiv\gamma_1+FF''+\sigma(F')^2=0,$$

$$I_3\equiv\delta_1+\left(\frac{f}{F}\right)'\frac{1}{(\sigma+1)F^{\sigma-1}}=0, \tag{3.13}$$

which corresponds to the case $H_*\in\mathscr{H}_1$.

(ii) If $\dim\mathscr{W}=2$ then there exist constants γ and δ such that

$$g^{\sigma+1}g'=\gamma g^{\sigma+3}+\delta g. \tag{3.14}$$

Dividing (3.14) over $g\not\equiv 0$, differentiating once and comparing with (3.11) we conclude that such a function g exists if and only if

$$\gamma_1=\gamma[\gamma(\sigma+2)-\beta_1],\quad \delta_1=\delta[\gamma(\sigma+2)-\beta_1]. \tag{3.15}$$

One can see that this corresponds to the case $H_*\in\mathscr{H}_2$. Any less dimensional reduction of $\mathscr{W}$ is impossible.

Case 2: $\alpha_1=0$. This reduces to (3.14) and to the same analysis. □

The sign-invariants from $\mathscr{H}_1$ and the corresponding exact solutions in separate variables are already known and studied in detail.[12,16] Observe that the ODE (3.7) always admits a linear solution $F(u)=Cu$ with either $C=-\gamma$ or $C=-\gamma(\sigma+2)/\sigma$, which corresponds, after the transformations $u=e^{\rho Ct}v$, to stationary or separate-variable solutions.

We now consider new exact solutions generated by sign-invariants from $\mathscr{H}_2$. These solutions are new even for the semilinear heat Eq. (1.1) with $\sigma=0$.

Theorem 3.3: *The one-dimensional equation*

$$u_t=(\sigma+1)(u_x)^\sigma u_{xx}+f(u), \tag{3.16}$$

where $f(u)$ *is given by (3.9) with F satisfying (3.7), admits the exact solution* $u(x,t)$ *given by*

$$\Phi(u)\equiv\int_1^u dz/F(z)=\int_0^x g(y)dy+C_0(t), \tag{3.17}$$

with $g(y)$ *being a solution to (3.6) with the boundary conditions*

$$g(0)=\alpha>0. \tag{3.18}$$

The function C_0 *solves the ODE*

$$\frac{dC_0}{dt}=(\sigma+1)\left(\alpha^{\sigma+2}+\frac{\delta}{\gamma}\right)F^\sigma(\Phi^{-1}(C_0))(F'(\Phi^{-1}(C_0))+\gamma)+\rho, \quad t>0. \tag{3.19}$$

Proof: Substituting (3.17) into (3.16) and using (3.14) we obtain the identity

$$\frac{dC_0}{dt}=(\sigma+1)\left(g^{\sigma+2}+\frac{\delta}{\gamma}\right)F^\sigma(F'+\gamma)+\rho\equiv G(x,t). \tag{3.20}$$

One can check that under the hypotheses of the theorem $G_x\equiv 0$. In order to prove this identity one can differentiate $G(x,\cdot)$ and substitute $g^{\sigma+1}g'$ from (3.14). Finally, (3.7) yields the result.

Hence (3.20) is equivalent to the equation $C_0'=G(0,t)$ and (3.19) follows. □

3.2 Dimensions: $N>1$. If $N\neq 1$ then, in general, (3.3) with $\sigma\neq 0$ is a linear combination in a five-dimensional space of the form

$$W_\sigma=\mathscr{L}\left\{\varphi,\ g^{\sigma+1}g',\ g,\ g^{\sigma+3},\ \frac{g^{\sigma+2}}{r}\right\}, \quad \varphi=\nabla\cdot(g^\sigma\nabla g)-\frac{N-1}{\sigma+1}\frac{g^{\sigma+1}}{r^2}. \tag{3.21}$$

If $\sigma=0$ then the space is four-dimensional

$$W_0=\mathscr{L}\{\varphi,\ gg',\ g,\ g^3\}, \quad \varphi=\Delta g-(N-1)\frac{g}{r^2}. \tag{3.22}$$

The semilinear case $\sigma=0$. We now prove the main result on completeness for the semilinear equation with $\sigma=0$.

Theorem 3.4: *Let* $\sigma=0$. *A sign-invariant of the form (3.1) exists in the two cases: (i) for the equation with a special logarithmic nonlinearity*

$$u_t=\Delta u+Au\ \log u+Bu, \tag{3.23}$$

and (ii) for a class of equations when H_* *contains the special functions* $g(r)=c/r$ *or* $g(r)=cr$.

Proof: Consider the space W_0. In view of (3.3) with $\sigma=0$ we have to have that [see (3.22)]

$$\varphi=\alpha gg'+\beta g+\gamma g^3. \tag{3.24}$$

If we assume that

$$gg',\ g,\ \text{and}\ g^3\ \text{are linearly independent}, \tag{3.25}$$

then substituting (3.24) into (3.3) we arrive at the system

$$2F'+\alpha=0, \quad FF''+\gamma=0, \quad F(f/F)'+\beta=0,$$

whence $\gamma=0$ and F is a linear function. After translation in u and necessary rescaling in the independent variables we get (3.23).

If (3.25) does not hold then g satisfies

$$g'=\varepsilon g^2+\delta, \quad \text{so that } g''=2\varepsilon^2 g^3+2\varepsilon\delta g. \tag{3.26}$$

Substituting (3.26) and (3.24) we arrive at a cubic algebraic equation for g

$$(2\varepsilon^2-\alpha\varepsilon-\gamma)g^3+\frac{\varepsilon(N-1)}{r}\,g^2+\left(2\varepsilon\delta-\frac{N-1}{r^2}-\alpha\delta-\beta\right)g+\frac{\delta(N-1)}{r}=0.$$

One can see that it matches with (3.26) if and only if $\varepsilon\delta=0$. Indeed, if $\varepsilon\delta\neq 0$ then by (3.26) the function $g(r)$ is given either by an exponential-like function or by the tan. Both functions do not satisfy the cubic equation. Therefore either $g'=\delta$ $(\varepsilon=0)$ or $g'=\varepsilon g^2$ $(\delta=0)$ completing the proof. □

In the case (ii) nontrivial exact solutions can be found in Sec. III in Ref. 8.

Equation (3.23) is well-known to admit separable solutions $u=M(t)\psi(x)$ and others. See p. 93 in Ref. 17. Exact solutions corresponding to the functions $g(r)=r$ or $1/r$ were constructed in Ref. 8. Observe that theorem 3.4 proves a kind of completeness of the results of Sec. III in Ref. 8.

The corresponding maximal sign-invariant is presented in the proof of the theorem.

The quasilinear case: $\sigma\neq 0$. The analysis presented below gives a complete study of the sign-invariants (3.1) in the quasilinear case. We stress our attention to some particular situations, which correspond to new types of exact solutions, and omit details of the study of the known types or trivial solutions.

Consider the space W_σ. First of all it follows from (3.3) that the function φ given in (3.21) must satisfy

$$\varphi=\alpha g^{\sigma+1}g'+\beta g+\gamma g^{\sigma+3}+\delta\,\frac{g^{\sigma+2}}{r}. \tag{3.27}$$

(i) Similar to the case $\sigma=0$ one can show that the assumption [cf. (3.25)]

$$g^{\sigma+1}g', \quad g, \quad g^{\sigma+3} \text{ and } g^{\sigma+2}/r \text{ are linearly independent,} \tag{3.28}$$

leads to the equation

$$u_t=\operatorname{div}(|\nabla u|^\sigma\nabla u)+Au^{\sigma+1}+Bu, \tag{3.29}$$

which admits separable solutions.[12]

(ii) Assume now that (3.28) is false so that

$$g^\sigma g'=\varepsilon g^{\sigma+2}+\nu+\mu\,\frac{g^{\sigma+1}}{r}. \tag{3.30}$$

We study compatibility conditions for the system (3.27), (3.30). Differentiating (3.30) and substituting $(g^\sigma g')'$ into (3.27), as in the semilinear case, we arrive at the algebraic equation

$$Ag^{\sigma+3}+Bg+C\,\frac{g^{\sigma+2}}{r}+D\,\frac{g^{\sigma+1}}{r^2}+E\,\frac{1}{r}=0, \tag{3.31}$$

with the coefficients

$$A=\varepsilon^2(\sigma+2)-\alpha\varepsilon-\gamma,\quad B=\varepsilon\nu(\sigma+2)-\alpha\nu-\beta,$$
$$C=\varepsilon[\mu(2\sigma+3)+N-1]-\alpha\mu-\delta,$$
$$D=(\sigma+1)\left(\mu-\frac{1}{\sigma+1}\right)\left(\mu-\frac{1-N}{\sigma+1}\right),\tag{3.32}$$
$$E=\nu(\sigma+1)\left(\mu-\frac{1-N}{\sigma+1}\right).$$

Setting $w=rg(r)$ we obtain the system

$$r^{\sigma+2}(Bw+E)=-(Aw^{\sigma+3}+Cw^{\sigma+2}+Dw^{\sigma+1}),\tag{3.33}$$

$$rw^{\sigma}w'=\varepsilon w^{\sigma+2}+(\mu+1)w^{\sigma+1}+\nu r^{\sigma+2}.\tag{3.34}$$

It follows from (3.33) that $\sigma=-2$ plays a special role.

A: Let $\sigma\neq-2$. Then we need to consider two possibilities.

A.1: It follows from (3.33), (3.34) that under a suitable choice of the parameters there exists a special solution $w=$const, i.e., the function

$$g(r)=\frac{1}{r},\tag{3.35}$$

which will be considered in detail in Sec. IV. Observe that the function (3.35) gives a consistent type of (3.1) sign-invariants (3.1) for a wide class of quasilinear parabolic equations, see Sec. III in Ref. 8.

A.2: We now look for a nontrivial solution

$$w\not\equiv\text{const}.\tag{3.36}$$

Consider first the special case of (3.33)

$$B=E=0\quad\text{and}\quad A=C=D=0.\tag{3.37}$$

Then from (3.32) we get two cases: either $\mu=1/(\sigma+1)$ or $\mu=(1-N)/(\sigma+1)$.

If $\mu=1/(\sigma+1)$ $(D=0)$ and $\nu=0$ $(E=0)$ then (3.34) is solved explicitly:

$$g(r)=\left[-\frac{\varepsilon(\sigma+1)}{\sigma+2}\,r+kr^{-1/(\sigma+1)}\right]^{-1},\quad k=\text{const}.\tag{3.38}$$

The identity (3.3) reads

$$g^{\sigma+3}[\alpha\varepsilon+\gamma+2(\sigma+1)\varepsilon F'+FF''+\sigma(F')^2]+\frac{g^{\sigma+2}}{r}\left[\alpha\mu+\delta+\left(2\mu(\sigma+1)+\frac{(N-1)\sigma}{\sigma+1}\right)F'\right]$$
$$+g\left[\alpha\nu+\beta+2(\sigma+1)\nu F'+\left(\frac{f}{F}\right)'\frac{F^{1-\sigma}}{\sigma+1}\right]\equiv0.\tag{3.39}$$

Then similarly to the case $N=1$, from (3.39) we obtain the parabolic equation (3.2) with a linear term $f(u=cu)$ (and $F(u)=cu$), which admits a separable solution.

If $\mu=(1-N)/(\sigma+1)$ $(D=E=0)$ then (3.3) leads to $F(u)=-\varepsilon u$ and $f(u)=-(\sigma+1)\nu(-\varepsilon u)^{\sigma+1}$. Moreover, the resulting exact solution is stationary.

Assume now that $BE\neq0$. Then substituting $r^{\sigma+2}$ from (3.33):

$$r^{\sigma+2}=-\frac{w^{\sigma}(Aw^3+Cw^2+Dw)}{Bw+E},$$

into Eq. (3.34) rewritten in the form

$$\frac{dr^{\sigma+2}}{dw}=\frac{(\sigma+2)r^{\sigma+2}w^{\sigma}}{\varepsilon w^{\sigma+2}+(\mu+1)w^{\sigma+1}+\nu r^{\sigma+2}},$$

we arrive at a fifth-order polynomial of the independent variable $w\in\mathbf{R}$ with the coefficients depending on the parameters introduced in (3.27) and in (3.30), which has to vanish identically. This gives extra five equations on the parameters. The sixth coefficient of $w^0=1$ is $DE[E(\mu-1/(\sigma+1))-D\nu]=0$ by (3.32). It can be shown that this case provides no new solutions.

B: Let $\sigma=-2$. Then (3.33) and (3.34) take the form

$$(A+B)w+(C+E)+Dw^{-1}=0, \tag{3.40}$$

$$rw^{-2}w'=\varepsilon+(\mu+1)w^{-1}+\nu. \tag{3.41}$$

B.1: There exists a special solution $w=\text{const}$, to be considered in Sec. IV.

B.2: Assume (3.36) holds. Then (3.40) implies that

$$A+B=0,\quad C+E=0,\quad D=0.$$

Since $D=-(\mu+1)(\mu-N+1)$ by (3.32), we consider two possibilities.

B.2.1: Let $\mu=N-1$. Then $D=E=0$ and hence $C=0$. From (3.30) we then get the function

$$g(r)=\left(-\frac{\rho}{N}\,r+kr^{1-N}\right)^{-1},\quad \rho=\varepsilon+\nu. \tag{3.42}$$

In this case the coefficients of the second term $1/r$ on the right-hand side of (3.39) vanish identically, $g^{\sigma+3}\equiv g$ for $\sigma=-2$, and (3.39) reduces to a single equation for two functions f and F:

$$FF''-2(F')^2-2\rho F'-\left(\frac{f}{F}\right)'F^3=0. \tag{3.43}$$

Integrating it once, we obtain

$$f(u)=\frac{1}{F(u)}\,[F'(u)+\rho]+MF(u),\quad M=\text{const}. \tag{3.44}$$

B.2.2: If $\mu=-1$ then (3.30) takes the form

$$g^{-2}g'=\rho-\frac{1}{gr},\quad \rho=\varepsilon+\nu,$$

and hence

$$g(r)=(kr-\rho r\,\log r)^{-1}\quad(\rho\neq 0). \tag{3.45}$$

Substituting the parameters into (3.39) we obtain a linear combination on the subspace $\mathscr{L}\{g,r^{-1}\}$ and hence two linearly independent differential equations on f and F which give separate-variable solutions.

IV. SIGN-INVARIANTS AND NEW SOLUTIONS FOR $N>1$

Using the results on the maximal sign-invariants in Sec. III we now construct new exact solutions of the quasilinear Eq. (3.2) with $N>1$. It follows from Theorem 3.4 ($\sigma=0$) and further results for $\sigma\neq 0$ that nontrivial solutions exist in two cases only: Either (3.35) or (3.42), $\sigma=-2$.

4.1. The case: $g=1/r$. For convenience and for further applications we first perform calculations of the sign-invariants for a power-like function

$$g(r)=r^l. \tag{4.1}$$

Then from (3.3) we arrive at the four term identity

$$\mathscr{T}\equiv r^{l(\sigma+3)}I_1(u)+r^{l(\sigma+2)-1}I_2(u)+r^{l(\sigma+1)-2}I_3(u)+r^lI_4(u), \tag{4.2}$$

with the coefficients (we suppose that $F>0$)

$$I_1=F^2(F^{\sigma+1})'', \quad I_2=[2l(\sigma+1)^2+\sigma(N-1)]F^{\sigma+1}F', \tag{4.3}$$

$$I_3=[l(\sigma+1)-1][N-1+(\sigma+1)l]F^{\sigma+1}, \quad I_4=f'F-fF'. \tag{4.4}$$

Therefore, identity $\mathscr{T}\equiv 0$ yields four differential equations on the two functions f and F which is inconsistent in general.

From (4.2) it is clear that the case $l=-1$ plays a special role (other critical exponents l will be discussed in the final remarks of this section). Namely, if $l=-1$ and $\sigma\neq -2$ then (4.2) reduces to a two term identity on the span $\mathscr{L}\{r^{-(\sigma+3)},r^{-1}\}$ which determines a system of two ODE's

$$F^2(F^{\sigma+1})''+aF^{\sigma+1}F'+bF^{\sigma+1}=0,$$

$$a=\sigma(N-1)-2(\sigma+1)^2, \quad b=(\sigma+2)(\sigma+2-N), \tag{4.5}$$

$$f'F-fF'=0. \tag{4.6}$$

The last equation yields

$$f(u)=\lambda F(u), \quad \lambda=\text{const}. \tag{4.7}$$

Thus, the quasilinear equation with gradient diffusivity

$$u_t=|u_r|^\sigma\left[(\sigma+1)u_{rr}+\frac{N-1}{r}\,u_r\right]+\lambda F(u), \tag{4.8}$$

where $F(u)$ is an arbitrary solution of the ODE (4.5) admits the sign-invariant

$$H(u)=u_r-\frac{1}{r}\,F(u). \tag{4.9}$$

We now construct the corresponding exact solution.

Theorem 4.1: *The parabolic Eq. (4.8), (4.5) admits the solution*

$$v(r,t)\equiv\int_1^{u(r,t)}\frac{dz}{F(z)}=\log r+C_0(t), \tag{4.10}$$

where

$$C_0(t)=-\frac{1}{\sigma+2}\log(1+C_1e^{(\sigma+2)\lambda t})+\lambda t+C_2, \tag{4.11}$$

and C_1,C_2 are arbitrary constants.

Proof: The function v satisfies the evolution equation

$$v_t=F^\sigma(v_r)^\sigma\left[(\sigma+1)(v_{rr}+(v_r)^2F')+\frac{N-1}{r}\,v_r\right]+\lambda. \tag{4.12}$$

Substituting (4.10), which directly follows from (4.9) ($H(u)=0$), we get the equation

$$C_0'=\frac{1}{r^{\sigma+2}}\,\Phi(v)+\lambda, \tag{4.13}$$

where

$$\Phi(v)=F^{\sigma}(u)[(\sigma+1)(F'(u)-1)+N-1].$$

One can see from the ODE (4.5) that $\Phi'=(\sigma+2)\Phi$ whence $\Phi(v)=de^{(\sigma+2)v}$, with $d=\Phi(0)$. Finally, substituting $\Phi(v)$ into (4.13) and using (4.10) we arrive at the ODE $C_0'=de^{(\sigma+2)C_0}$, $t>0$, which gives (4.11). □

Example 4.1: The exact solution (4.10) depends on the ODE (4.5). It is integrated two times if $\sigma=1$ when it takes the form

$$(F^2)''+aF'+b=0, \quad a=N-9, \quad b=3(3-N).$$

If $N=3$ then $b=0$ and the general solution is

$$F(u)-C\ \log(F(u)+C)=3u, \quad C=\text{const}.$$

If $N\neq 3$ then we get the following algebraic equation for F

$$(F(u)-3u)^3\left(F(u)+\frac{N-3}{2}\,u\right)^{(N-3)/2}=C.$$

4.2. The case: $\sigma=-2$, $\mu=N-1$. Recall that the function (3.42) satisfies equation (3.30)

$$g^{-2}g'=\rho+\frac{N-1}{rg(r)}. \tag{4.14}$$

Exact solutions, which follow from the zero-invariance of the operator (3.1) are given in the following statement, which gives exact solutions to an infinite-dimensional family of quasilinear solutions.

Theorem 4.2: *The quasilinear equation*

$$u_t=\frac{1}{(u_r)^2}\left(-u_{rr}+\frac{N-1}{r}\,u_r\right)+\frac{F'(u)+\rho}{F(u)}+MF(u), \tag{4.15}$$

where $F\not\equiv 0$ *is an arbitrary smooth function and* $M=\text{const}$, *admits the exact solution* $u(r,t)$ *given by*

$$\Phi(u)\equiv\int_1^u\frac{dz}{F(z)}=-\frac{1}{\rho}\log\left(1-\frac{\rho}{kN}\,r^N\right)+Mt. \tag{4.16}$$

Observe that (4.15) is a parabolic equation backward in time.

Proof: Note that $f(u)$ in (4.15) is given by (3.44) and the solution (4.16) takes the form

$$\Phi(u)=\int_0^{\tau}g(y)dy+C_0(t), \tag{4.16a}$$

where g is specified in (3.42). Therefore as in the proof of Theorem 3.3, from (4.16a) we deduce that $u_t=FC_0'$, $u_r=Fg$, $u_{rr}=FF'g^2+Fg'$, and substituting these expressions into (4.15) we obtain the equation

$$C_0'=\frac{1}{F^2}\left(-g^{-2}g'+\rho+\frac{N-1}{rg(r)}\right)+M.$$

In view of (4.14) we get:

$$C_0'=M, \tag{4.17}$$

and we arrive at the solution (4.16). □

Example 4.2: Setting e.g., $F(u)=u^2$ we obtain that the quasilinear equation

$$u_t=\frac{1}{(u_r)^2}\left(-u_{rr}+\frac{N-1}{r}\,u_r\right)+\frac{Mu^4+2u+\rho}{u^2}$$

admits the solution

$$u(r,t)=\left[\frac{1}{\rho}\log\left(1-\frac{\rho}{kN}\,r^N\right)-Mt\right]^{-1}.$$

Final Remarks: 1. As far as we know the sign-invariant (3.1), (3.42), (3.43) for Eq. (4.15) with $N>1$ is a unique example of such a sign-invariant with a nonpower function $g(r)$ given in (3.42) with $\rho k\neq 0$. If we set $k=0$ in (3.42) then $g(r)\sim 1/r$ coincides with (4.1) for $l=-1$. On the other hand, solution (4.16a) then coincides with (4.10) [cf. (4.11), $\sigma=-2$).

2. It follows from identity (4.2) that the case $l=N-1$, $\sigma=-2$ is also a special one for the existence of a sign-invariant. It corresponds to $\rho=0$ in (3.42) and in (4.14).

3. The last special exponent is $l=1/(\sigma+1)$. Note that the corresponding differential inequality with the negative sign, $u_r+r^{1/(\sigma+1)}F(u)\leqslant 0$, plays an important part in the analysis of blow-up of radial solutions of the quasilinear Eq. (3.2) with $f(u)\sim u^p$, $p>1$. See Ref. 15. Substituting $l=1/(\sigma+1)$ into (4.2) we get the following ODE's on the functions f and F

$$(F^{\sigma+1})''=0,\quad f'F-fF'+F^{\sigma+1}F'[2+\sigma(N+1)]=0. \tag{4.18}$$

This yields the functions

$$F(u)=(au+b)^{1/(\sigma+1)},\quad f(u)=-\frac{2+\sigma(N+1)}{\sigma}\,(au+b)+\alpha(au+b)^{1/(\sigma+1)}. \tag{4.19}$$

In this case operator (3.1) with $g(r)=r^{1/(\sigma+1)}$ is also zero-invariant of Eq. (1.1) and exact solutions exist. However, more general solutions can be constructed by another invariant approach.[16] For instance, using the transformation $u=v^{(\sigma+1)/\sigma}$ in the equation

$$u_t=\mathrm{div}(|\nabla u|^{\sigma}\nabla u)+\alpha u^{1/(\sigma+1)}+\beta u,$$

we obtain the equation

$$v_t=\mathscr{B}(v)\equiv\mu^{-\sigma}|v_r|^{\sigma}\left[(\sigma+1)\left(vv_{rr}+\frac{1}{\sigma}\,(v_r)^2\right)+\frac{N-1}{r}\,vv_r\right]+\mu(\alpha+\beta v),$$

where $\mu=\sigma/(\sigma+1)$. The operator $\mathscr{B}$ admits the two-dimensional invariant subspace $W_2=\mathscr{L}\{1,|x|^{\gamma}\}$, $\gamma=(\sigma+2)/(\sigma+1)$, i.e., $\mathscr{B}(W_2)\subseteq W_2$. Thus, there exist exact solutions of the form $v_*(x,t)=C_0(t)+C_1(t)|x|^{\gamma}\in W_2$, where the coefficients $\{C_0(t),C_1(t)\}$ solve a nonlinear dynamical system. See details and generalizations in Ref. 16.

4. The structure of the exact solution (4.16) [as well as (2.12)] shows that it can be treated as a solution on a one-dimensional subspace or on a set invariant under a nonlinear operator of the equation. See different examples in Refs. 8,9,18, and 19.

ACKNOWLEDGMENTS

This paper was written while the second author visited the Department of Mathematical Sciences, the University of Bath, supported by a grant from the London Mathematical Society. He is deeply grateful to both institutions for their hospitality.

[1] G. I. Barenblatt, *Similarity, Self-Similarity and Intermediate Asymptotics* (Consultants Bureau, New York, 1978).
[2] K. A. Brakke, *The Motion of a Surface by Its Mean Curvature* (Princeton University Press, Princeton, New Jersey, 1978).
[3] Y.-G. Chen, Y. Giga, and S. Goto, ''Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations,'' J. Diff. Geom. **33**, 749–786 (1991).
[4] M. Gurtin, ''Towards a nonequilibrium thermodynamics of two-phase material,'' Arch. Ration. Mech. Anal. **100**, 275–312 (1988).

[5] L. C. Evans and J. Spruck, ''Motion of level sets by mean curvature. I,'' J. Diff. Geom. **33**, 635–681 (1991).
[6] H. M. Soner and P. E. Souganidis, ''Singularities and uniqueness of cylindrically symmetric surfaces moving by mean curvature,'' Comm. Partial Differ. Equat. **18**, 859–894 (1993).
[7] Y. Giga, ''Interior derivative blow-up for quasilinear parabolic equations,'' Disc. Cont. Dyn. Systems **1**, 449–461 (1995).
[8] V. A. Galaktionov, ''Quasilinear heat equations with first-order sign invariants and new explicit solutions,'' Nonlinear Anal., Theory, Meth. and Appl. **23**, 1595–1621 (1994).
[9] V. A. Galaktionov and S. A. Posashkov, ''Explicit solutions of quasilinear heat equations with general first-order sign-invariants,'' Physica D **99**, 217–236 (1996).
[10] V. A. Galaktionov and J. L. Vazquez, ''Geometrical properties of the solutions of one-dimensional nonlinear parabolic equations,'' Math. Ann. **303**, 741–769 (1995).
[11] V. A. Galaktionov, ''Two methods of comparison of solutions of parabolic equations,'' Sov. Phys. Dokl. **25**, 250–252 (1980).
[12] V. A. Galaktionov, ''Conditions for global non-existence and localization of solutions of the Cauchy problem for a class of non-linear parabolic equations,'' USSR Comput. Math. Math. Phys. **23**, 36–44 (1983).
[13] M. L. Storm, ''Heat conduction in simple metals,'' J. Appl. Phys. **22**, 940–951 (1951).
[14] A. Friedman and B. McLeod, ''Blow-up of positive solutions of semilinear heat equation,'' Indiana Univ. Math. J. **34**, 425–447 (1985).
[15] V. A. Galaktionov and S. A. Posashkov, ''Single point blow-up for *N*-dimensional quasi-linear equations with gradient diffusion and source,'' Indiana Univ. Math. J. **40**, 1041–1060 (1991).
[16] V. A. Galaktionov and S. A. Posashkov, ''Exact solutions and invariant spaces for non-linear gradient diffusion equations,'' USSR Comput. Math. Math. Phys. **34**, 313–321 (1994).
[17] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov and A. P. Mikhailov, *Blow-up in Quasilinear Parabolic Equations* (Nauka, Moscow, 1987; Walter de Gruyter, Berlin, 1995).
[18] V. A. Galaktionov, ''Invariant subspaces and new explicit solutions to evolution equations with quadratic nonlinearities,'' Proc. R. Soc. Edinburgh, Sect. A: Math. **A125**, 225–246 (1995); Report No. AM-91-11, School of Math., Univ. of Bristol (1991).
[19] V. A. Galaktionov, S. A. Posashkov, and S. R. Svirshchevskii, ''On invariant sets and explicit solutions of nonlinear equations with quadratic nonlinearities,'' Differ. and Integral Equat. **8**, 1997–2024 (1995).

On the representations of a *Q*-oscillator algebra

A. Guichardet
Ecole Polytechnique, Centre de Mathematiques, U.R.A 169 du C.N.R.S., F-91128 Palaiseau Cedex, France

(Received 27 October 1997; accepted for publication 27 April 1998)

We give a classification of all irreducible representations of the Arik–Coon oscillator algebra without the assumption that the number operator has a discrete nondegenerate spectrum. © *1998 American Institute of Physics.* [S0022-2488(98)01509-6]

I. INTRODUCTION

In theoretical physics literature a great number of papers have been devoted to the description of various quantum oscillator algebras (see e.g., Refs. 1–6). Here we choose the one generated by elements a, a^+, N with the relations

$$[N,a^+]=a^+, \quad [N,a]=-a, \quad aa^+-qa^+a=1,$$

where q is a fixed strictly positive real number. This algebra is called Arik–Coon oscillator algebra and denoted by $\tilde{A}_0$ in Ref. 5; its irreducible representations are described in Refs. 4 and 5 under the assumption that the spectrum of N is discrete and nondegenerate.

The aim of this paper is, on the one hand, to prove that this condition is automatically fulfilled and, on the other hand, to relate this question with the construction of C^*-algebras ''à la Woronowicz'' and with the theory of crossed products of C^*-algebras by their automorphism groups.

In order to do this, we shall need definitions more precise than the above ones; we shall replace the (rather vague) notion of a representation of $\tilde{A}_0$ by that of a quadruple $\mathcal{T}=(E,T,V,N)$ where E is a complex Hilbert space; T and N are (generally unbounded) self-adjoint operators in E with domains D_T, D_N, T being moreover positive; V is a partially isometric operator in E satisfying

(1) $V(D_T)\subset D_T$, $V^*(D_T)\subset D_T$,
(2) $VV^*=I$, $V^*V=$orthogonal projection onto $(\mathrm{Ker}\ T)^\perp$,
(3) $VT=(qT+I)V$ on D_T,
(4) T and N commute, i.e., their spectral projections commute,
(5) $V(D_N)\subset D_N$, $V^*(D_N)\subset D_N$,
(6) $NV-VN=-V$ on D_N.

The formal link with the above definition is $T=a^+a$, $a=VT^{1/2}$ (polar decomposition).

We call commutant of such a quadruple $\mathcal{T}$ the von Neumann algebra M of all bounded operators in E commuting with V, V^* and with the spectral projections of T and N; we say that $\mathcal{T}$ is irreducible if M is reduced to the scalars. The equivalence between two quadruples is the unitary equivalence.

The paper is organized as follows. In Sec. II we classify all triples (E,T,V); in Sec. III we construct a C^*-algebra A whose representations are in bijective correspondance with these triples (a similar construction has been done in Ref. 7 for the quantum euclidean group in two dimensions); in Sec. IV we classify all quadruples $\mathcal{T}$ and show that the corresponding C^*-algebra is a crossed product of A by the group $\mathbf{R}$ of real numbers.

Here I would like to thank the referee for his useful suggestions.

 © 1998 American Institute of Physics

II. CLASSIFICATION OF TRIPLES $\mathscr{S}=(E,T,V)$

Theorem 1: *Every irreducible triple $\mathscr{S}=(E,T,V)$ is equivalent to one and only one of the following list:*

(i) If $q\geqslant 1$, only one triple $\mathscr{S}_0$ where $E_0=1^2(N)$ with orthonormal basis (e_n), $n\geqslant 0$, $T_0(e_n)=[n]_q e_n$ where $[n]_q=(q^n-1)/(q-1)$, $V_0 e_n=e_{n-1}$ (equal to 0 if $n=0$).
(ii) If $q<1$.

(a) A one-parameter family $\mathscr{S}_\theta$ ($\theta\in\mathbf{U}$=set of complex numbers of modulus 1) where $E_\theta=\mathbf{C}$, $T_\theta=(1-q)^{-1}$, $V_\theta=\theta$.
(b) A one-parameter family $\mathscr{S}_\mu$ ($\mu\in]q,1]$) where $E_\mu=1^2(\mathbf{Z})$ with orthonormal basis (e_n), $n\in\mathbf{Z}$, $T_\mu e_n=(q^n\mu+(1-q)^{-1})e_n$, $V_\mu e_n=e_{n-1}$.
(c) One triple $\mathscr{S}_0$ with the same formulas as in (i).

Sketch of proof:

(1) We denote by ρ the function of a real variable defined by $\rho(x)=qx+1$; its iterated ρ^n, for $n\in\mathbf{Z}$, are given by $\rho^n(x)=q^n x+[n]_q$. Fix $x\geqslant 0$.
  (a) If $q>1$, the sequence $\rho^n(x)$ is strictly increasing from $(1-q)^{-1}<0$ to $+\infty$
  (b) If $q<1$ and $x>(1-q)^{-1}$, it is strictly decreasing from $+\infty$ to $(1-q)^{-1}>0$
  (c) If $q<1$ and $x=(1-q)^{-1}$, it is constant
  (d) If $q<1$ and $x<(1-q)^{-1}$, it is strictly increasing from $-\infty$ to $(1-q)^{-1}>0$.
  Moreover if $x\neq(1-q)^{-1}$, there exists a neighborhood V of x such that the sets $\rho^n(V)$, for $n\in\mathbf{Z}$, are mutually disjoint.
(2) Let $\mathscr{S}$ be a triple (E,T,V), f a real valued bounded borel function on $\mathbf{R}$; then the bounded operator $f(T)$ satisfies $Vf(T)=f(\rho(T))\,V$ (indeed this is true on the domain D_∞, intersection of the domains of all positive powers of T, for f a polynomial, and remains true for a general f by continuity). In particular, if $f\circ\rho=f$, $f(T)$ belongs to the commutant of $\mathscr{S}$.
(3) If $\mathscr{S}$ is irreducible, the spectrum spT is either reduced to the point $(1-q)^{-1}$, or is a collection of nondegenerate eigenvalues of the form $\rho^n(x)$ with x fixed and n variable. The proof of this fact is very similar to those of Proposition 3.2 in Ref. 8 and Proposition 6.3 in Ref. 3.
(4) At this point, the proof is well known (see e.g., Ref. 5).

Remark: (On domains of self-adjointness for $a+a^+$ and $i(a-a^+$.) We choose the triple $\mathscr{S}_0$ with $q>1$; setting $A_+=a+a^+$, $A_-=i(a-a^+)$, we have

$$A_+ e_n=[n]_q^{1/2}e_{n-1}+[n+1]_q^{1/2}e_{n+1},$$

$$A_- e_n=i[n]_q^{1/2}e_{n-1}-i[n+1]_q^{1/2}e_{n+1},$$

these operators have been studied in Ref. 9 where the following properties are established. Let D denote the set of all vectors in $1^2(\mathbf{N})$ having a finite number of nonzero coordinates; clearly A_+ and A_- are defined and symmetric on D; let $\overline{A_+}$ and $\overline{A_-}$ be their closures. Then

(a) each of $\overline{A_+}$ and $\overline{A_-}$ has a one parameter family of self-adjoint extensions which are obtained by adjoining one dimension to its domain;
(b) every such self-adjoint extension has a purely discrete spectrum;
(c) $\overline{A_+}$ and $\overline{A_-}$ have no common domain on which they are self-adjoint.

Theorem 2: *Every triple $\mathscr{S}$ is a direct integral of irreducible ones.* The proof is the same as in the theory of group representations: One diagonalizes a maximal commutative subalgebra in the commutant of $\mathscr{S}$.

III. CONSTRUCTION OF A C^*-ALGEBRA ASSOCIATED TO THE TRIPLES $\mathscr{S}$

We shall consider only the (most complicated) case where $q<1$; the idea is to introduce the direct integral of all irreducible triples $\mathscr{S}$ of theorem 1 (ii), and replace everywhere T by the bounded operator $S=e^{-T}$. Set

$$\tilde{E}=\int_{\theta\in\mathbf{U}}^{\oplus}E_\theta\,d\theta\oplus\int_{\mu\in]q,1]}^{\oplus}E_\mu\,d\mu\oplus E_0=L^2(\mathbf{U})\oplus L^2(]q,1];1^2(\mathbf{Z}))\oplus 1^2(\mathbf{N}),$$

$$\tilde{T}=(1-q)^{-1}\oplus\int^{\oplus}T_\mu\,d\mu\oplus T_0,$$

$$\tilde{V}=\int^{\oplus}\theta\,d\theta\oplus\int^{\oplus}V_\mu\,d\mu\oplus V_0,$$

$$\tilde{S}=e^{-\tilde{T}},$$

$$B=\text{sub-}C^*\text{-algebra of }\mathscr{L}(\tilde{E})\text{ generated by }\tilde{S}\text{ and }\tilde{V},$$

$$A=\text{closed two-sided ideal of }B\text{ generated by }\tilde{S}.$$

The following lemma is easily proved:

Lemma 1: (i) *For every positive number a, we have* $\tilde{S}^a\in B$ *and*

$$\tilde{V}\tilde{S}^a=e^{-a}\tilde{S}^{-qa}\tilde{V}.$$

(ii) *The linear combinations of monomials* $(\tilde{V}^*)^m\tilde{V}^r\tilde{S}^P$, *where m and r are positive integers and P is a nonzero polynomial with respect to q and* q^{-1} *with positive integral coefficients, form a sub-*-algebra everywhere dense in A, which we denote by* A^0.

Theorem 3: *There is a bijective correspondance between representations* (E,π) *of A and triples* $\mathscr{S}=(E,T,V)$ *given by* $T=\log\pi(\tilde{S})$, $V=\pi(\tilde{V})$.

Sketch of proof: By theorem 2 we can limit ourselves to irreducible representations and triples. If (E,π) is an irreducible representation of *A*, $\pi(\tilde{S})$ is nonzero, its kernel is invariant and distinct from *E*, hence null; $\pi(\tilde{S})$ is positive and injective, $\log\pi(\tilde{S})$ exists, and $(E,\log\pi(\tilde{S}),\pi(\tilde{V}))$ is a triple.

Conversely consider an irreducible triple $\mathscr{S}$, and take for instance $\mathscr{S}$ of the form $\mathscr{S}_{\mu_0}$; we can define a representation π of A^0 by setting, for every

$$X=\sum_{m,r,P}\lambda_{m,r,P}(\tilde{V}^*)^m\tilde{V}^r\tilde{S}^P\in A^0,$$

$$\pi(X)=\sum_{m,r,P}\lambda_{m,r,P}(V_{\mu_0}^*)^mV_{\mu_0}^rS_{\mu_0}^P.$$

Then we have only to prove that π is continuous. But *X* contains as a direct summand the direct integral of all

$$X_\mu=\sum_{m,r,P}\lambda_{m,r,P}(V_\mu^*)^mV_\mu^rS_\mu^P$$

whence

$$\|X\|\geqslant\|\int^{\oplus}X_\mu\,d\mu\|=\sup_\mu\|X_\mu\|\geqslant\|X_{\mu_0}\|=\|\pi(X)\|.$$

IV. CLASSIFICATION OF QUADRUPLES $\mathscr{T}$

Theorem 4: (i) *If* $q\geqslant 1$, *every irreducible quadruple* $\mathscr{T}=(E,T,V,N)$ *is equivalent to one and only one of the one-parameter family* $\mathscr{T}_\gamma(\gamma\in\mathbf{R})$ *where* $E_\gamma=1^2(\mathbf{N})$ *with orthonormal basis* $(e_n)(n\geqslant 0)$, $T_\gamma e_n=[n]_q e_n$, $V_\gamma e_n=e_{n-1}$ *(equal to 0 if* $n=0$*)*, $N_\gamma e_n=(n+\gamma)e_n$.
(ii) *If* $q<1$, *every irreducible quadruple* $\mathscr{T}=(E,T,V,N)$ *is equivalent to one and only one of the*

following list:
(a) *a two-parameter family* $\mathscr{T}_{\mu,\gamma}$ $(\mu\in]q,1]$, $\gamma\in\mathbf{R})$ *where* $E_{\mu,\gamma}=1^2(\mathbf{Z})$ *with orthonormal basis* $(e_n)(n\in\mathbf{Z})$, $T_{\mu,\gamma}e_n=(q^n\mu+(1-q)^{-1})e_n$, $V_{\mu,\gamma}e_n=e_{n-1}$, $N_{\mu,\gamma}e_n=(n+\gamma)e_n$,
(b) *a one-parameter family* $\mathscr{T}_{0,\gamma}(\gamma\in[0,1[$ *as in* (a) *but with* $\mu=0$,
(c) *a one-parameter family* $\mathscr{T}_{\gamma}(\gamma\in\mathbf{R})$ *with the same formulas as in* (i).

N.B.: This result is compatible with Ref. 5, 4.1.2 if one defines the numbers $\widetilde{v_0}$ and c of Ref. 5 by $\widetilde{v_0}=\gamma$, $c=q^{-\gamma}(q^{\gamma}-1)/(q-1)$ in case (i) or (iic), and by $\widetilde{v_0}=\gamma$, $c=(q-1)^{-1}-q^{-\gamma}\mu$ in cases (iia) and (iib).

Theorem 5: *Define an action* α *of the group* $\mathbf{R}$ *by automorphisms of the C*-algebra A by* $\alpha_t(\widetilde{S})=\widetilde{S}$, $\alpha_t(\widetilde{V})=e^{-it}\widetilde{V}$ *and form the corresponding crossed-product* $A\times\mathbf{R}$. *Then there exists a bijective correspondance between representations of* $A\times\mathbf{R}$ *and quadruples* $\mathscr{T}$.

Sketch of proofs of Theorems 4 and 5: (a) We can replace each quadruple $\mathscr{T}=(E,T,V,N)$ by the quadruple (E,T,V,U) where U is the one-parameter unitary group $U_t=e^{itN}$; conditions (4)–(6) then become

$$U_tTU_t^{-1}=T,\quad U_tVU_t^{-1}=e^{-it}V. \tag{7}$$

Theorem 3 implies that quadruples $\mathscr{T}$ are in bijective correspondence with covariant representations of the pair $(A,\mathbf{R})$ (for more on this notion, see for instance Ref. 10); then theorem 5 is a consequence of the very definition of $A\times\mathbf{R}$.
(b) We now have to classify the irreducible covariant representations (π,U) of the pair $(A,\mathbf{R})$. To do this we can use the method explained in Ref. 10, analogous to the Mackey machinery (or ''Bargmann's little group method''), well known to construct representations of group extensions: One chooses an orbit X of the action of $\mathbf{R}$ on the dual space $\hat{A}$ of A, then a point ζ in X, then a covariant representation of the pair $(A,\mathbf{R}_\zeta)$ where $\mathbf{R}_\zeta$ is the stabilizer of ζ in $\mathbf{R}$, and induces it to $(A,\mathbf{R})$.

In case $q\geqslant 1$, $\tilde{A}$ is reduced to one single point, described in theorem 1(i), hence $\mathbf{R}_\zeta=\mathbf{R}$; moreover we have a particular U satisfying (7), namely $U_t^0e_n=e^{itn}e_n$; if U is another solution of (7), $U_t\,(U_t^0)^{-1}$ commutes with T and V, hence is a scalar which we can write as $e^{it\gamma}$, $\gamma\in\mathbf{R}$.

Consider now the case $q<1$; the orbits of $\mathbf{R}$ in $\hat{A}$ are the set X of all $\mathscr{S}_\theta$ $(\theta\in\mathbf{U})$; the various points $\mathscr{S}_\mu$, the point $\mathscr{S}_0$.

In the last two cases the proof is similar to that of the case $q\geqslant 1$, and we get families $\mathscr{T}_{\mu,\gamma}$ and $\mathscr{T}_\gamma$. Take now the orbit X and the point $\zeta=\mathscr{S}_1$; we have $\mathbf{R}_\zeta=2\pi\mathbf{Z}$, so that U is induced by a character of $2\pi\mathbf{Z}$, of the form $2\pi k\to\lambda^k$, $\lambda\in\mathbf{U}$; the space E of U can be realized as the set of complex functions on $\mathbf{R}$ satisfying

$$f(t+2\pi k)=\lambda^k f(t),\quad \int_0^{2\pi}|f(t)|^2dt<+\infty,$$

with the action $(U_tf)(s)=f(t+s)$. Taking the orthonormal basis of E consisting of the functions $e_n(t)=(2\pi)^{-1/2}e^{i(\gamma+n)t}$ with $\gamma\in[0,1]$ and $e^{2\pi i\gamma}=\lambda$, we get the family $\mathscr{T}_{0,\gamma}$.

V. CONCLUDING REMARKS

The same method can be applied to other algebras, for instance to the algebra defined (formally) by the same generators N,a,a^+ as before but with the relations

$$[N,a^+]=a^+,\quad [N,a]=-a, aa^+-qa^+a=0,$$

which is a deformation of the enveloping algebra of the euclidean group $E(2)$. In this case, for $q\neq 1$, besides the trivial one dimensional representation, one finds a two-parameter family of irreducible quadruples $\mathscr{T}_{\mu,\gamma}$ $(\mu\in]1,q]$ or $]q,1]$ according as $q>1$ or $q<1$, $\gamma\in\mathbf{R})$ where $E_{\mu,\gamma}=1^2(\mathbf{Z})$ with orthonormal basis (e_n) $(n\in\mathbf{Z})$, $T_{\mu,\gamma}e_n=q^n\mu e_n$, $V_{\mu,\gamma}e_n=e_{n-1}$, $N_{\mu,\gamma}e_n=(n+\gamma)e_n$.

[1] M. Arik and D. D. Coon, J. Math. Phys. **17**, 524 (1976).
[2] W. Pusz and S. Woronowicz, Rep. Math. Phys. **27**, 231 (1989).

[3] A. Guichardet, Annales Polonici Math. **66**, 123–135 (1997).
[4] P. P. Kulish, Theor. Math. Phys. **86**, 108–110 (1991).
[5] C. Quesne and N. Vansteenkiste, Helv. Phys. Acta **69**, 141–157 (1996).
[6] G. Rideau, Lett. Math. Phys. **24**, 147–153 (1992).
[7] S. L. Woronowicz, Lett. Math. Phys. **23**, 251–263 (1991).
[8] L. L. Vaksman and Ia. S. Soibelman, Funct. Anal. Appl. **22**, 170–181 (1988).
[9] A. Kempf, J. Math. Phys. **35**, 4483 (1994).
[10] M. Takesaki, Acta Math. **119**, 273–303 (1967).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Algebraic characterization of gauge anomalies on a nontrivial bundle

P. John, O. Moritsch, and M. Schweda[a)]
Institut für Theoretische Physik, Technische Universität Wien, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria

S. P. Sorella
UERJ, Universidade do Estado do Rio de Janeiro, Departamento de Física Teórica, Instituto de Física, Rua São Francisco Xavier, 524, 20550-013, Maracanã, Rio de Janeiro, Brazil

(Received 28 April 1997; accepted for publication 5 September 1997)

We discuss the algebraic way of solving the descent equations corresponding to the BRST consistency condition for the gauge anomalies and the Chern–Simons terms on a nontrivial bundle. The method of decomposing the exterior space–time derivative as a BRST commutator is extended to the present case. © *1998 American Institute of Physics.* [S0022-2488(98)00607-0]

I. INTRODUCTION

Recently, a new method for solving the descent equations corresponding to the BRST consistency condition for anomalies and generalized Chern–Simons terms has been successfully proposed.[1] This method, originally developed for trivial bundles and flat space–time, relies on the introduction of an algebraic operator which allows us to decompose the exterior derivative as a BRST commutator. This important property has been shown to give a very simple and elegant way of obtaining solutions of the consistency condition. Moreover, these solutions have been proven[2] to be cohomological equivalent to those provided by the so called *Russian formula*.[3–6]

Our aim in the present work is to extend the previous algebraic setup to nontrivial bundles, allowing thus for curved space-time and global properties. Although the method can be applied to several field models, including the gravitational anomalies[7,8] and the topological field theories,[9,10] we shall consider here, as an explicit example, the case of a non-Abelian gauge theory of the Yang–Mills type. We shall focus, in particular, on the algebraic characterization of the gauge anomalies and of the generalized Chern–Simons terms defined on nontrivial bundles.

However, before going any further, let us make some useful remarks on the functional space which shall be used in order to define the BRST differential. In what follows we shall work in the functional space of the local form-polynomials built up essentially with the one-form gauge connection A, the zero form Faddeev–Popov ghost field c and their exterior derivatives. This means that we shall restrict ourselves to the space of the local polynomials constructed by taking suitable exterior products of (A, c, dA, dc), d denoting the space–time differential. As it is well known, this functional space is not the most general space which can be assumed for the BRST differential, as it does not span the set of solutions of the BRST cohomology which explicitly contains the space–time metric, as for instance terms of the Yang–Mills type $F\wedge *F^2$. (The symbol * denotes the usual Hodge operation.) Nevertheless, the space of form-polynomials is known to be sufficiently large in order to include all nontrivial solutions of the descent equations which turn out to be independent of the metric. In fact, as proven by Refs. 11–13, these solutions are exactly given by the gauge anomalies and the Chern–Simons terms which are known to be written in terms of pure differential forms. In addition, being the latters integrated forms of maximum degree, they are left invariant by diffeomorphism transformations. Let us also underline that the present analysis is particularly adapted to the case of the topological theories of the Schwartz type[9] which, being related to a zero curvature condition,[14] can be naturally formulated in terms of form-field polynomials.

a)Electronic mail: mschweda@tph.tuwien.ac.at

 © 1998 American Institute of Physics

The paper is organized as follows. In Sec. II we briefly recall some well-known properties of the nontrivial bundles. Sections III and IV are devoted to the introduction of the BRST differential and of the decomposition of the exterior derivative as a BRST commutator. Finally, in Secs. V and VI we deal with the solutions of the descent equations on nontrivial bundles.

II. GAUGE CONNECTIONS ON NONTRIVIAL BUNDLES

In order to discuss the global structure of the gauge anomalies let us recall some properties of the gauge connections defined on a principal bundle $P(M,G)$, M being an arbitrary space–time base manifold M_{2k-2} of even dimension $(2k-2)$ and G a compact Lie group. (A general introduction to the theory of bundles can be found in Refs. 15 and 16.)

Connections on $P(M,G)$ are locally represented by Lie algebra valued one-forms,

$$A=A_\mu dx^\mu=A^a T^a=A^a_\mu dx^\mu T^a, \tag{1}$$

where T^a are the anti-Hermitian generators of G,

$$[T^a,T^b]=f^{abc}T^c, \quad \mathrm{Tr}(T^aT^b)=\delta^{ab}, \tag{2}$$

f^{abc} being the totally antisymmetric structure constants. The associated two-form field strength is given by

$$F=\tfrac{1}{2}F_{\mu\nu}dx^\mu dx^\nu=dA+A^2=dA+\tfrac{1}{2}\{A,A\}, \tag{3}$$

where a wedge product has to be understood. It obeys the Bianchi identity

$$DF=dF+[A,F]=0, \tag{4}$$

where $D=dx^\mu D_\mu$ is the covariant exterior derivative with respect to the gauge field A and $d=dx^\mu\partial_\mu$ is the exterior space–time derivative.

Gauge transformations of $P(M,G)$ on M are locally represented by means of a G-valued gauge connection A_μ transforming as

$$\delta_G A_\mu=-D_\mu\Omega=-\partial_\mu\Omega-[A_\mu,\Omega], \tag{5}$$

where Ω is a G-valued infinitesimal gauge parameter.

Infinitesimal diffeomorphisms on M are represented by vector fields v^μ. They form an infinite dimensional Lie algebra on M, denoted by $\mathrm{Diff}(M)$. The gauge field transforms under the infinitesimal diffeomorphisms as a vector, i.e.,

$$\delta_D A_\mu=-L_v A_\mu=-v^\nu\partial_\nu A_\mu-(\partial_\mu v^\nu)A_\nu, \tag{6}$$

where L_v denotes the Lie derivative along the infinitesimal vector field v^μ. The inner derivative i_v with respect to the vector field v^μ is defined as usual by

$$i_v dx^\mu=v^\mu. \tag{7}$$

As it is well-known, the inner derivative i_v, the Lie derivative L_v and the exterior derivative d form an infinite dimensional graded Lie subalgebra

$$\begin{aligned}
&\{i_v,i_{v'}\}=0, \quad [L_v,i_{v'}]=i_{[v,v']},\\
&\{i_v,d\}=L_v, \quad [L_v,L_{v'}]=L_{[v,v']},\\
&[L_v, \quad d]=0,\{d,d\}=0,
\end{aligned} \tag{8}$$

with $(v,v')\in\mathrm{Diff}(M)$.

Let us come now to the global properties of a principle bundle. A bundle is called trivial if it has the form $P=M\times G$. Bundles over contractible base manifolds M are trivial.[15,16] In the case of a trivial bundle the set of fields introduced so far is sufficient in order to describe the algebraic

structure of the gauge anomalies. However, as shown in Ref. 3, for the case of a nontrivial bundle one has to introduce a further gauge field, called the reference gauge connection $\mathring{A}$. Such a new connection is chosen to be a fixed background field, due to the fact that the action of v on A is defined only up to a gauge transformation. With the help of the reference connection $\mathring{A}$, all vector fields v on M can be lifted to G. Again, the infinitesimal gauge transformations form an infinite dimensional Lie algebra, now denoted by $\tilde{g}$. Let U be an open trivializing subset of M, $U \subset M$. Then the elements of $\tilde{g}$ are represented on U by pairs of parameters (Ω, v)

$$\Omega \in \Lambda^0(U,g), \quad v \in \mathrm{Diff}(U), \tag{9}$$

where Λ^0 denotes the space of zero-forms with values in the Lie algebra g of G. The reference connection is locally represented by the one-form $\mathring{A}$ on U with values in g,

$$\mathring{A} = \mathring{A}_\mu dx^\mu \in \Lambda^1(U,g). \tag{10}$$

The pairs of parameters obey the commutation relations

$$\begin{aligned}
&[(\Omega',0),(\Omega,0)] = ([\Omega',\Omega],0),\\
&[(0,v'),(0,v)] = (i_{v'} i_v \mathring{F},[v',v]),\\
&[(0,v),(\Omega,0)] = (-L_v\Omega - [i_v \mathcal{A},\Omega],0),
\end{aligned} \tag{11}$$

where $\mathring{F}$ is the two-form field strength associated to $\mathring{A}$,

$$\mathring{F} = \tfrac{1}{2}\mathring{F}_{\mu\nu} dx^\mu dx^\nu = d\mathring{A} + \mathring{A}^2 = d\mathring{A} + \tfrac{1}{2}\{\mathring{A},\mathring{A}\}, \tag{12}$$

which fulfills the Bianchi identity

$$\mathring{D}\mathring{F} = d\mathring{F} + [\mathring{A},\mathring{F}] = 0. \tag{13}$$

The reference connection $\mathring{A}$ ensures that the commutators can be patched together in a consistent way on the overlap $U \cap U'$ of two open subsets U and U' of M. This is achieved by replacing the exterior derivative d in the Lie derivative L_v by the covariant exterior derivative $\mathring{D}$. In fact, one has

$$[(0,v),(\Omega,0)] = (-\mathring{\mathcal{L}}_v\Omega,0), \tag{14}$$

with $\mathring{\mathcal{L}}_v = i_v\mathring{D} + \mathring{D}i_v$. Note that the ordinary Lie derivative L_v in (6) can be rewritten as $L_v = i_v d + d i_v$.

Different reference connections $\mathring{A}$ on P yield isomorphic Lie algebras $\tilde{g}$. Note that the pure gauge transformations $(\Omega,0)$ form an ideal of $\tilde{g}$, while the transformations with respect to the vector fields $(0,v)$ in general do not form a subalgebra of $\tilde{g}$.

Moreover, if the bundle P is trivial, one can choose $U = M$ and $\mathring{A} = 0$. Then $\tilde{g}$ is the semi-direct product of $\Lambda^0(M,g)$ and $\mathrm{Diff}(M)$, meaning that every nontrivial bundle is locally trivial.

The affine space of all connections on P carries an affine representation $\mathcal{R}$ of $\tilde{g}$ given locally by

$$\mathcal{R}(\Omega,v)A = -D(\Omega + i_v(A - \mathring{A})) - i_v F, \tag{15}$$

where $A,\mathring{A}\in\Lambda^1(U,g)$ are the local expressions on U of the connections on P and D is the covariant exterior derivative with respect to A. [$\Lambda^1(U,g)$ denotes the space of the one-forms with values in the Lie algebra g.] Finally, by definition, the fixed reference connection $\mathring{A}$ does not transform under $\tilde{g}$,

$$\mathcal{R}(\Omega,v)\mathring{A}=0. \tag{16}$$

III. THE BRST TRANSFORMATIONS

In order to introduce the BRST differential s, let us replace, as usual, the infinitesimal parameters (Ω,v) introduced in the previous section with Grassmann ghost fields. We will introduce therefore two ghost fields, c and ξ^μ, associated to the gauge and to the diffeomorphism transformations, respectively. Both ghosts have form degree zero and ghost number one.

Concerning the BRST transformations[17] of the various fields and ghosts, we proceed by making use of the so-called Maurer–Cartan horizontality conditions.[5,6] For this purpose we define the nilpotent differential operator of total degree one,

$$\tilde{d}=d-s, \quad \tilde{d}^2=0, \tag{17}$$

where d is the exterior derivative. (The total degree is defined as the sum of the form degree and of the ghost number.) Furthermore, we define the generalized gauge connection[5,6] according to

$$\tilde{A}=A+c+i_\xi(A-\mathring{A}), \tag{18}$$

with i_ξ as the inner derivative with respect to the diffeomorphism ghost ξ^μ,

$$i_\xi dx^\mu=\xi^\mu. \tag{19}$$

The generalized reference connection is simply given by

$$\mathring{\tilde{A}}=\mathring{A}, \tag{20}$$

expressing the fact that $\mathring{A}$ is a fixed background field which does not transform.

The corresponding generalized field strengths of total degree two are given by

$$\tilde{F}=\tilde{d}\tilde{A}+\tilde{A}^2 \quad \text{and} \quad \mathring{\tilde{F}}=\tilde{d}\mathring{\tilde{A}}+\mathring{\tilde{A}}^2, \tag{21}$$

and obey the generalized Bianchi identities

$$\tilde{D}\tilde{F}=\tilde{d}\tilde{F}+[\tilde{A},\tilde{F}]=0 \quad \text{and} \quad \mathring{\tilde{D}}\mathring{\tilde{F}}=\tilde{d}\mathring{\tilde{F}}+[\mathring{\tilde{A}},\mathring{\tilde{F}}]=0, \tag{22}$$

with $\tilde{D}$ and $\mathring{\tilde{D}}$ the generalized covariant exterior derivatives with respect to $\tilde{A}$ and to $\mathring{\tilde{A}}$.

The Maurer–Cartan horizontality conditions can then be expressed by

$$\tilde{F}=\tfrac{1}{2}F_{\mu\nu}\tilde{d}x^\mu\tilde{d}x^\nu, \quad \mathring{\tilde{F}}=\tfrac{1}{2}\mathring{F}_{\mu\nu}\tilde{d}x^\mu\tilde{d}x^\nu, \tag{23}$$

with $\tilde{d}x^\mu$ the generalized differential of degree one,

$$\tilde{d}x^\mu=dx^\mu+\xi^\mu. \tag{24}$$

Expanding these conditions according to the ghost number and the form degree one gets the following set of nilpotent BRST transformations:

$$sc = c^2 - \mathring{\mathcal{L}}_\xi c + \tfrac{1}{2} i_\xi i_\xi \mathring{F},$$

$$sA = D(c + i_\xi (A - \mathring{A})) - i_\xi F, \quad s\mathring{A} = 0, \tag{25}$$

$$sF = [(c + i_\xi (A - \mathring{A})), F] - \mathcal{L}_\xi F, \quad s\mathring{F} = 0,$$

where $\mathcal{L}_\xi = i_\xi D - D i_\xi$ and $\mathring{\mathcal{L}}_\xi = i_\xi \mathring{D} - \mathring{D} i_\xi$ are the covariant Lie derivatives with respect to A and $\mathring{A}$, respectively. [Notice that the relative sign in the definition of the Lie derivatives has now been changed due to the fact that ξ is an odd variable, carrying ghost number one. This implies that i_ξ is even and that $\mathcal{L}_\xi$, $\mathring{\mathcal{L}}_\xi$ have odd degree.] The above BRST transformations are completed by giving the transformation of the diffeomorphism ghost,

$$s\xi^\mu = -\xi^\nu \partial_\nu \xi^\mu. \tag{26}$$

Instead of the ghost c it turns out to be more useful to use the following combination:

$$\hat{c} = c + i_\xi (A - \mathring{A}), \tag{27}$$

as the basic variable. In terms of the new variable $\hat{c}$, the BRST transformations read now as

$$s\xi^\mu = -\xi^\nu \partial_\nu \xi^\mu, \quad s\hat{c} = \hat{c}^2 - \hat{F},$$

$$sA = D\hat{c} - i_\xi F, \quad s\mathring{A} = 0, \tag{28}$$

$$sF = [\hat{c}, F] - \mathcal{L}_\xi F, \quad s\mathring{F} = 0,$$

with $\hat{F}$ defined as

$$\hat{F} = \tfrac{1}{2} i_\xi i_\xi F. \tag{29}$$

Notice, finally, that $\hat{F}$ has form degree zero, ghost number two, and that it transforms covariantly, i.e.,

$$s\hat{F} = [\hat{c}, \hat{F}]. \tag{30}$$

IV. DECOMPOSITION FORMULA AND ALGEBRAIC RELATIONS

As already remarked in the Introduction and as done in previous works,[1,8] we shall assume that the functional space the BRST operator will act upon is identified with the space of form-polynomials constructed out of the variables $(\xi^\mu,\ c,\ A - \mathring{A})$ and the differentials $(d\xi^\mu,\ dc,\ dA,\ d\mathring{A})$. Of course, on the local space of form-polynomials this set of variables can be always replaced by the equivalent basis given by $(\xi^\mu,\ d\xi^\mu,\ \hat{c},\ \mathring{D}\hat{c},\ (A - \mathring{A}),\ F,\ \mathring{F})$. The presence of the covariant derivative $\mathring{D}$ and of the curvatures $(F,\ \mathring{F})$ yields an easy understanding of the latter choice, which turns out to be more convenient in order to solve the BRST consistency condition. We remark also that the use of the combination $(A - \mathring{A})$ relies on the fact that it emerges rather naturally in the BRST transformations (28) and in the formula (27). Let us recall, finally, that the space of form-polynomials is the most suitable choice in order to discuss the anomalies and the Chern–Simons terms, which, as it is well known, can be written in terms of differential forms.[11–13]

Let us now introduce an operator δ of total degree zero defined in the following way:[1,8]

$$\delta\xi^\mu = -dx^\mu, \quad \delta\varphi = 0, \quad \text{for } \varphi = (c, A, \mathring{A}, F, \mathring{F}). \tag{31}$$

From the above expression one sees that δ increases the form degree and decreases the ghost number by one unit, respectively, and that it acts only on the diffeomorphism ghost. The usefulness of the operator δ relies on the fact that it allows us to decompose the covariant exterior space–time derivative $\mathring{D}$ as a BRST commutator. In fact, we have

$$[s,\delta]=-\mathring{D}. \tag{32}$$

Let us remark that the presence of the covariant derivative $\mathring{D}$ in Eq. (32) makes all the difference with respect to the case of a trivial bundle, where only the ordinary exterior space–time derivative d appears. The operators $(s,\ \delta,\ \mathring{D})$ give rise to the following algebraic relations:

$$\begin{gathered}
[s,\delta]=-\mathring{D},\quad [\mathring{D},\delta]=[d,\delta]=2\mathcal{G},\\
\{s,\mathring{D}\}=\{s,d\}=s^2=d^2=0,\quad [\mathcal{G},\delta]=0,\\
\{\mathcal{G},\mathring{D}\}=\{\mathcal{G},d\}=0,\quad \mathcal{G}\mathcal{G}=0,\quad \{\mathcal{G},s\}=\mathcal{K}=\mathring{D}\mathring{D},\quad [\mathcal{K},s]=0,\\
[\mathcal{K},\mathring{D}]=[\mathcal{K},d]=0,\quad [\mathcal{K},\delta]=0,\quad [\mathcal{K},\mathcal{G}]=0,
\end{gathered} \tag{33}$$

where the operator $\mathcal{G}$ is defined as

$$\mathcal{G}\hat{c}=\mathring{F},\quad \mathcal{G}\varphi=0,\quad \text{for}\ \ \varphi=(\xi^\mu,d\xi^\mu,\mathring{D}\hat{c},(A-\mathring{A}),F,\mathring{F}). \tag{34}$$

It decreases the ghost number by one unit and increases the form degree by two units, hence it is an operator of total degree one.

Let us also note that in the case of a trivial bundle $(\mathring{A}=0)$ the operator $\mathcal{G}$ is absent. This is consistent with the results found in Ref. 8. In other words, its presence is related to the global properties of the bundle.

V. COHOMOLOGY AND DESCENT EQUATIONS

It is well-known that the search of the anomalies and of the generalized Chern–Simons actions corresponding to a given set of field transformations can be done in a purely algebraic way by solving the BRST consistency conditions in the space of the integrated local field polynomials.

This leads to the study of the nontrivial solutions of the following cohomology problem:

$$s\Delta=0,\quad \Delta\neq s\overline{\Delta}, \tag{35}$$

where Δ and $\overline{\Delta}$ are integrated local field polynomials. Setting $\Delta=\int \mathcal{Q}_n^g$, the condition (35) translates at the nonintegrated level as

$$s\mathcal{Q}_n^g+d\mathcal{Q}_{n-1}^{g+1}=0, \tag{36}$$

where $\mathcal{Q}_n^g$ is some local polynomial in the fields with ghost number g and form degree n, n denoting the dimension of the space–time. $\mathcal{Q}_n^g$ is said to be nontrivial if

$$\mathcal{Q}_n^g\neq s\overline{\mathcal{Q}}_n^{g-1}+d\overline{\mathcal{Q}}_{n-1}^g. \tag{37}$$

In this case the integral of $\mathcal{Q}_n^g$ on space–time identifies a cohomology class of the BRST operator s and, according to its ghost number, it corresponds to an invariant Chern–Simons action $(g=0)$ or to an anomaly $(g=1)$.

The nonintegrated equation (36), due to the algebraic Poincaré Lemma,[18,12] is easily seen to generate a tower of descent equations,

$$
\begin{aligned}
&s\mathcal{Q}_n^g + d\mathcal{Q}_{n-1}^{g+1} = 0,\\
&s\mathcal{Q}_{n-1}^{g+1} + d\mathcal{Q}_{n-2}^{g+2} = 0,\\
&\cdots\cdots\\
&\cdots\cdots\\
&s\mathcal{Q}_1^{g+n-1} + d\mathcal{Q}_0^{g+n} = 0,\\
&s\mathcal{Q}_0^{g+n} = 0.
\end{aligned}
\tag{38}
$$

As it has been well-known for several years, these equations can be solved by using a transgression procedure based on the so-called *Russian formula*.[3,5,6,19,11,20–22,4,13] More recently an alternative way of finding nontrivial solutions of the ladder (38) has been proposed and successfully applied in the study of the Yang–Mills gauge anomalies.[1] The method makes use of the decomposition formula (32). One easily verifies that once the decomposition has been found, successive applications of the operator δ on the zero-form $\mathcal{Q}_0^{g+n}$ which solves the last equation of the tower (38) give an explicit nontrivial solution for the higher cocycles.

It is a remarkable fact that solving the last equation of the tower of descent equations (38) is only a problem of local BRST cohomology instead of a modulo-d one. One sees then that, due to the operator δ, the study of the cohomology of s modulo d is essentially reduced to the study of the local cohomology of s. The latter can be, e.g., systematically analyzed by using the powerful technique of the spectral sequences.[23]

VI. CHERN–SIMONS TERMS AND GAUGE ANOMALIES

In order to solve the descent equations (38) we use the following strategy. First, we look at the general nontrivial solution of the last descent equation. Then we solve the tower iteratively by using the operators δ and $\mathcal{G}$, as done in Ref. 1.

Let $n=2k$ be the dimension of the base manifold M of a nontrivial bundle $P(M,G)$ with structure group G. Introducing a covariant BRST operator according to

$$
S = s - \hat{c}; \tag{39}
$$

one obtains a remarkable correspondence between the form sector and the gauge sector of the theory,

$$
DF = dF + [A,F] = 0, \quad S\hat{F} = s\hat{F} - [\hat{c},\hat{F}]. \tag{40}
$$

Let us define now the interpolating shifted gauge ghost,

$$
\hat{c}(t) = t\hat{c}, \quad t \in [0,1], \tag{41}
$$

with $\hat{c}(0)=0$ and $\hat{c}(1)=\hat{c}$, and the associated ghost field strength,

$$
\hat{F}(t) = -s\hat{c}(t) + \hat{c}(t)\hat{c}(t), \tag{42}
$$

with $\hat{F}(0)=0$ and $\hat{F}(1)=\hat{F}$. With the help of the interpolating generalized covariant BRST operator,

$$
S_t = s - \hat{c}(t), \tag{43}
$$

with $S_0=s$ and $S_1=S$, one finds the following identities:

$$\frac{d\hat{F}(t)}{dt} = -S_t\hat{c}, \quad S_t\hat{F}(t) = 0. \tag{44}$$

Therefore, in a space–time with dimension $2k$ one has

$$\mathrm{Tr}(\hat{F}^k) = \mathrm{Tr}(\hat{F}^k(1) - \hat{F}^k(0)) = \mathrm{Tr}\int_0^1 dt\ \frac{d}{dt}\ \hat{F}^k(t) = k\ \mathrm{Tr}\int_0^1 dt\ \frac{d\hat{F}(t)}{dt}\ \hat{F}^{k-1}(t)$$

$$= -k\ \mathrm{Tr}\int_0^1 dt\ (S_t\hat{c})\hat{F}^{k-1}(t) = -s\left(k\ \mathrm{Tr}\int_0^1 dt\ \hat{c}\ \hat{F}^{k-1}(t)\right). \tag{45}$$

Using the nilpotency of the BRST operator and the fact that $\mathrm{Tr}(\hat{F}^k) \neq 0$ in a space–time with dimension $2k$, one may conclude that $k\ \mathrm{Tr}\int_0^1 dt\ \hat{c}\ \hat{F}^{k-1}(t)$ is nontrivial. Since $\mathrm{Tr}(\hat{F}^k)$ contains the product of $2k$ fermionic diffeomorphism ghosts ξ^μ, it follows:

> *In a space–time with dimension $2k-1$ a nontrivial solution of the local equation $s\mathcal{Q}_0^{2k-1} = 0$ on a nontrivial bundle can be represented by the integrated parametric formula,*

$$\mathcal{Q}_0^{2k-1} = -k\ \mathrm{Tr}\int_0^1 dt\ \hat{c}\ \hat{F}^{k-1}(t). \tag{46}$$

Acting with the operator $\mathcal{G}$ on the last of the descent equations (38) and using the algebra (33) one finds

$$\mathcal{G}s\mathcal{Q}_0^{2k-1} = 0 = -s\mathcal{G}\mathcal{Q}_0^{2k-1} + \mathring{D}\mathring{D}\mathcal{Q}_0^{2k-1}, \tag{47}$$

which, due to the vanishing of the last term in (47), implies that $\mathcal{G}\mathcal{Q}_0^{2k-1}$ is BRST-closed,

$$s\mathcal{G}\mathcal{Q}_0^{2k-1} = 0. \tag{48}$$

The general solution of (48) has the form

$$\mathcal{G}\mathcal{Q}_0^{2k-1} = s\overline{\mathcal{Q}}_2^{2k-3} + \hat{\mathcal{Q}}_2^{2k-2}, \tag{49}$$

with a possible nontrivial part $\hat{\mathcal{Q}}_2^{2k-2}$,

$$s\hat{\mathcal{Q}}_2^{2k-2} = 0, \hat{\mathcal{Q}}_2^{2k-2} \neq s\hat{\mathcal{Q}}_2^{2k-3}. \tag{50}$$

Recalling now that the operator $\mathcal{G}$ acting on the variable $\hat{c}$ yields $\mathring{F}$, we infer that

$$\hat{\mathcal{Q}}_2^{2k-2} = \mathrm{Tr}(\mathring{F}\mathcal{P}(\hat{c},\hat{F})) = \mathring{F}^a\mathcal{P}^a(\hat{c},\hat{F}), \tag{51}$$

with $\mathcal{P}$ some polynomial in the fields $\hat{c}$ and $\hat{F}$ having form degree zero and ghost number $2k-2$.

In the following we will show that $\mathcal{G}\mathcal{Q}_0^{2k-1}$ is not only BRST-closed but also BRST-exact,

$$\mathcal{G}\mathcal{Q}_0^{2k-1} = s\overline{\mathcal{Q}}_2^{2k-3}. \tag{52}$$

The exactness of $\mathcal{G}\mathcal{Q}_0^{2k-1}$ will be proven by showing that the polynomial $\mathcal{P}^a$ is BRST trivial, due to the fact that the variable $\mathring{F}$ is BRST invariant. Recalling then that $\mathcal{P}^a$ depends only on $(\hat{c},\hat{F})$ due to the fact that it has form degree zero, let us first express the rigid gauge transformations of $(\hat{c},\hat{F})$ as BRST anticommutators, i.e.,

$$\left\{s, \frac{\partial}{\partial\hat{c}^a}\right\} = \delta^a_{rig.} = f^{abc}\hat{c}^b\frac{\partial}{\partial\hat{c}^c} + f^{abc}\hat{F}^b\frac{\partial}{\partial\hat{F}^c}\ . \tag{53}$$

(Here we have taken into account only the part of the BRST operator relative to the fields $\hat{c}$, $\hat{F}$.) It is easily checked by using the Jacobi identity that the operator $\delta^a_{rig.}$ commutes with the BRST operator,

$$[\delta^a_{rig.}, s] = 0. \tag{54}$$

Furthermore, one has

$$\delta^a_{rig.} \mathring{F}^b = 0. \tag{55}$$

Decomposing now Eq. (51) in terms of the eigenvalues of $\delta^a_{rig.}$, we get

$$\delta^a_{rig.}(\mathring{F}^b \mathcal{P}^b)_{(l)} = l(\mathring{F}^b \mathcal{P}^b)_{(l)}, \quad \mathring{F}^b \mathcal{P}^b = \sum_l (\mathring{F}^b \mathcal{P}^b)_{(l)}. \tag{56}$$

For $l \neq 0$ one obtains

$$(\mathring{F}^b \mathcal{P}^b)_{(l)} = \frac{1}{l} \delta^a_{rig.}(\mathring{F}^b \mathcal{P}^b)_{(l)} = \frac{1}{l} \left\{ s, \frac{\partial}{\partial \hat{c}^a} \right\} (\mathring{F}^b \mathcal{P}^b)_{(l)} = s\left(\frac{1}{l} \frac{\partial}{\partial \hat{c}^a} (\mathring{F}^b \mathcal{P}^b)_{(l)} \right), \tag{57}$$

since from Eq. (50) and Eq. (56) it follows that

$$s(\mathring{F}^b \mathcal{P}^b)_{(l)} = 0. \tag{58}$$

However, contributions of the type (57) can be neglected, since they are BRST trivial. What remains is the term corresponding to the eigenvalue $l = 0$,

$$\delta^a_{rig.}(\mathring{F}^b \mathcal{P}^b)_{(0)} = 0. \tag{59}$$

The above equation states that $(\mathring{F}^b \mathcal{P}^b)_{(0)}$ has to be invariant under rigid gauge transformations. However, since $\mathring{F}$ is left invariant by $\delta^a_{rig.}$, Eq. (59) implies that the polynomial $\mathcal{P}(\hat{c}, \hat{F})$ has to be invariant as well. It follows therefore that $(\mathring{F}^b \mathcal{P}^b)_{(0)}$ vanish. This completes our proof.

Repeating and generalizing these arguments to higher $\mathcal{G}$-cocycles one sees that the identity (52) generates a subtower of descent equations between the operators s and $\mathcal{G}$,

$$\begin{aligned}
&\mathcal{G}\mathcal{Q}_0^{2k-1} = s\overline{\mathcal{Q}}_2^{2k-3}, \\
&\mathcal{G}\overline{\mathcal{Q}}_2^{2k-3} = s\overline{\mathcal{Q}}_4^{2k-5}, \\
&\cdots\cdots \\
&\cdots\cdots \\
&\mathcal{G}\overline{\mathcal{Q}}_{2k-4}^{3} = s\overline{\mathcal{Q}}_{2k-2}^{1}, \\
&\mathcal{G}\overline{\mathcal{Q}}_{2k-2}^{1} = \mathcal{P}^0_{2k}(\mathring{F}),
\end{aligned} \tag{60}$$

which end up with a polynomial $\mathcal{P}^0_{2k}(\mathring{F})$ of ghost number zero and form degree $2k$. [In this case one has to incorporate in the rigid gauge transformations of (53), also the contributions of the remaining fields with a nonvanishing form degree.] Moreover, from Eq. (34) it follows that $\mathcal{P}^0_{2k}(\mathring{F})$ is nothing else than the Chern form of $\mathring{F}$ in $2k$ space–time dimensions.

Equipped with this algebraic set up we are now ready to discuss the solutions of the descent equations (38). Before analyzing the general case of a space–time of dimension $n = 2k-1$ let us first solve, as an explicit example, the three dimensional case $(k = 2)$,

$$s\mathcal{Q}_3^0+d\mathcal{Q}_2^1=0,\quad s\mathcal{Q}_2^1+d\mathcal{Q}_1^2=0,\quad s\mathcal{Q}_1^2+d\mathcal{Q}_0^3=0,\quad s\mathcal{Q}_0^3=0. \tag{61}$$

The nontrivial solution of the last descent equation follows from (46) with $(k=2)$,

$$\mathcal{Q}_0^3=-2\,\mathrm{Tr}\int_0^1 dt\ \hat{c}\ \hat{F}(t)=\mathrm{Tr}\left(\frac{1}{3}\hat{c}^3-\hat{F}\hat{c}\right). \tag{62}$$

Acting with the operator δ on the last equation one gets

$$[\delta,s]\mathcal{Q}_0^3+s\,\delta\mathcal{Q}_0^3=0, \tag{63}$$

which, using the decomposition (32), becomes

$$s(\delta\mathcal{Q}_0^3)+d\mathcal{Q}_0^3=0. \tag{64}$$

(Remark that all polynomials $\mathcal{Q}$ are gauge invariant and therefore the covariant exterior derivative $\mathring{D}$ reduces to the ordinary one.) This equation shows that $\delta\mathcal{Q}_0^3$ provides a solution for the cocycle $\mathcal{Q}_1^2$ in Eqs. (61).

In order to find an expression for the next cocycle, we apply the operator δ on Eq. (64). Using then the algebraic relations (33), one has

$$s\left(\frac{\delta^2}{2}\mathcal{Q}_0^3\right)-\mathcal{G}\mathcal{Q}_0^3+d(\delta\mathcal{Q}_0^3)=0. \tag{65}$$

According to (52), the term $\mathcal{G}\mathcal{Q}_0^3$ in (65) can be rewritten as

$$\mathcal{G}\mathcal{Q}_0^3=s\,\mathrm{Tr}(\mathring{\hat{F}}\hat{c})=s\,\overline{\mathcal{Q}}_2^1, \tag{66}$$

so that Eq. (65) becomes

$$s\left(\frac{\delta^2}{2}\mathcal{Q}_0^3-\overline{\mathcal{Q}}_2^1\right)+d(\delta\mathcal{Q}_0^3)=0. \tag{67}$$

Repeating again the previous steps, for the highest level we obtain

$$s\left(\frac{\delta^3}{3!}\mathcal{Q}_0^3-\delta\overline{\mathcal{Q}}_2^1\right)+d\left(\frac{\delta^2}{2}\mathcal{Q}_0^3-\overline{\mathcal{Q}}_2^1\right)=0, \tag{68}$$

from which it follows that $\mathcal{Q}_3^0$ can be identified with $[(\delta^3/3!)\mathcal{Q}_0^3-\delta\overline{\mathcal{Q}}_2^1]$, showing then the usefulness of the operators δ and $\mathcal{G}$ in solving the tower (61).

Summarizing, the algebraic solution of the three dimensional tower of descent equations (61) in the case of a nontrivial bundle is given by

$$\mathcal{Q}_3^0=(\delta^3/3!)\mathcal{Q}_0^3-\delta\overline{\mathcal{Q}}_2^1, \tag{69}$$

$$\mathcal{Q}_2^1=(\delta^2/2)\mathcal{Q}_0^3-\overline{\mathcal{Q}}_2^1, \tag{70}$$

$$\mathcal{Q}_1^2=\delta\mathcal{Q}_0^3, \tag{71}$$

with

$$\mathcal{Q}_0^3=\mathrm{Tr}(\tfrac{1}{3}\hat{c}^3-\hat{F}\hat{c}), \tag{72}$$

$$\mathcal{Q}_1^2=\mathrm{Tr}(-(A-\mathring{A})\hat{c}^2+(A-\mathring{A})\hat{F}+i_\xi F\hat{c}), \tag{73}$$

$$\mathcal{Q}_2^1=\mathrm{Tr}((A-\mathring{A})^2\hat{c}-(A-\mathring{A})i_\xi F-(F+\mathring{\hat{F}})\hat{c}), \tag{74}$$

$$\mathcal{Q}_3^0=\mathrm{Tr}(-\tfrac{1}{3}(A-\mathring{A})^3+(A-\mathring{A})(F+\mathring{F})). \tag{75}$$

In particular, $\mathcal{Q}_3^0$ is recognized to be the three-dimensional Chern–Simons form defined on nontrivial bundles, while $\mathcal{Q}_2^1$ yields the two-dimensional gauge anomaly which, by using the definition of the shifted gauge ghost (27), can be rewritten as

$$\mathcal{Q}_2^1=-\mathrm{Tr}(c\ \mathring{D}A+c\ d\mathring{A}+\mathring{F}i_\xi(A-\mathring{A})). \tag{76}$$

Moreover, in the case of a trivial bundle ($\mathring{A}=0$) the anomaly (76) reduces to the well-known familiar expression

$$\mathcal{Q}_2^1=-\mathrm{Tr}(c\ dA). \tag{77}$$

Finally, acting with the exterior derivative d on $\mathcal{Q}_3^0$ one obtains the related Chern form

$$d\mathcal{Q}_3^0=\mathrm{Tr}(F^2-\mathring{F}^2). \tag{78}$$

Let us come now to the general case of a space–time with dimension $n=2k-1$,

$$\begin{gathered} s\mathcal{Q}_{2k-1}^0+d\mathcal{Q}_{2k-2}^1=0,\\ s\mathcal{Q}_{2k-2}^1+d\mathcal{Q}_{2k-3}^2=0,\\ \cdots\cdots\\ \cdots\cdots\\ s\mathcal{Q}_1^{2k-2}+d\mathcal{Q}_0^{2k-1}=0,\\ s\mathcal{Q}_0^{2k-1}=0. \end{gathered} \tag{79}$$

It is straightforward to iterate the previous construction to obtain the solution of the descent Eqs. (79). The latter is given by

$$\begin{gathered} \mathcal{Q}_0^{2k-1}=-k\ \mathrm{Tr}\int_0^1 dt\ \hat{c}\ \hat{F}^{k-1}(t),\\ \mathcal{Q}_{2p}^{2k-1-2p}=\frac{\delta^{2p}}{(2p)!}\mathcal{Q}_0^{2k-1}-\sum_{q=0}^{p-1}\frac{\delta^{2q}}{(2q)!}\overline{\mathcal{Q}}_{2p-2q}^{2k-1-2p+2q}, \end{gathered} \tag{80}$$

for the even space–time sector and

$$\begin{gathered} \mathcal{Q}_1^{2k-2}=\delta\mathcal{Q}_0^{2k-1},\\ \mathcal{Q}_{2p+1}^{2k-2-2p}=\frac{\delta^{2p+1}}{(2p+1)!}\mathcal{Q}_0^{2k-1}-\sum_{q=0}^{p-1}\frac{\delta^{2q+1}}{(2q+1)!}\overline{\mathcal{Q}}_{2p-2q}^{2k-1-2p+2q}, \end{gathered} \tag{81}$$

for the odd space–time sector and $p=1,2,\ldots,k$. The solutions (80) and (81) generalize the results of Ref. 1 to the case of a nontrivial bundle.

ACKNOWLEDGMENT

O. Moritsch was supported in part by the ''Österreichische Nationalbank'' under Contract Grant No. 5393.

[1] S. P. Sorella, Commun. Math. Phys. **157**, 231 (1993).
[2] S. P. Sorella and L. Tataru, Phys. Lett. B **324**, 3, 51 (1994).
[3] J. Mañes, R. Stora, and B. Zumino, Commun. Math. Phys. **102**, 157 (1985).
[4] F. Langouche, T. Schücker, and R. Stora, Phys. Lett. B **145**, 342 (1984); R. Stora, ''Algebraic structure of chiral anomalies,'' lecture given at the GIFT seminar, 1–8 June 1985, Jaca, Spain, preprint LAPP-TH-143; T. Schücker, Commun. Math. Phys. **109**, 167 (1987).
[5] J. Thierry-Mieg, J. Math. Phys. **21**, 2834 (1980); Nuovo Cimento A **56**, 396 (1980); Phys. Lett. B **147**, 430 (1984); ''Classical geometrical interpretation of ghost fields and anomalies in Yang–Mills theory and quantum gravity,'' *Symposium on Anomalies, Geometry, Topology*, edited by W. A. Bardeen and A. R. White (World Scientific, Singapore, 1985), p. 239.
[6] L. Baulieu, Nucl. Phys. B **241**, 557 (1984); L. Baulieu and J. Thierry-Mieg, Phys. Lett. B **145**, 53 (1984).
[7] M. Werneck de Oliveira and S. P. Sorella, Int. J. Mod. Phys. A **9**, 2979 (1994); O. Moritsch, M. Schweda, and S. P. Sorella, Class. Quantum Grav. **11**, 1225 (1994).
[8] O. Moritsch, M. Schweda, and T. Sommer, Class. Quantum Grav. **12**, 2059 (1995).
[9] D. Birmingham, M. Blau, M. Rakowski, and G. Thompson, Phys. Rep. **209**, 129 (1990).
[10] E. Guadagnini, N. Maggiore, and S. P. Sorella, Phys. Lett. B **255**, 65 (1991); C. Lucchesi, O. Piguet, and S. P. Sorella, Nucl. Phys. B **395**, 325 (1993); A. Brandhuber, O. Moritsch, M. W. de Oliveira, O. Piguet, and M. Schweda, Nucl. Phys. B **341**, 173 (1994).
[11] M. Dubois-Violette, M. Talon, and C. M. Viallet, Commun. Math. Phys. **102**, 105 (1985); Phys. Lett. B **158**, 231 (1985); Ann. Inst. Henri Poincaré **44**, 103 (1986); M. Dubois-Violette, M. Henneaux, M. Talon, and C. M. Viallet, Phys. Lett. B **267**, 81 (1991); M. Henneaux, Commun. Math. Phys. **140**, 1 (1991); M. Dubois-Violette, M. Henneaux, M. Talon, and C. M. Viallet, Phys. Lett. B **289**, 361 (1992).
[12] F. Brandt, N. Dragon, and M. Kreuzer, Nucl. Phys. B **332**, 224 (1990).
[13] F. Brandt, N. Dragon, and M. Kreuzer, Nucl. Phys. B **340**, 187 (1990).
[14] M. Carvalho, L. C. Q. Vilar, C. A. G. Sasaki, and S. P. Sorella, J. Math. Phys. **37**, 5310 (1996); **37**, 5325 (1996).
[15] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).
[16] M. Göckeler and T. Schücker, *Differential Geometry, Gauge Theories, and Gravity* (Cambridge University Press, Cambridge, 1987).
[17] C. Becchi, A. Rouet, and R. Stora, Ann. Phys. (N.Y.) **98**, 287 (1976); I. W. Tyutin, *Gauge Invariance in Field Theory and Statistical Physics*, Lebedev Institute, preprint FIAN No. 39, 1975.
[18] L. Bonora and P. Cotta-Ramusino, Commun. Math. Phys. **87**, 589 (1983).
[19] L. Alvarez-Gaumé and E. Witten, Nucl. Phys. B **234**, 269 (1983).
[20] G. Bandelloni, Nuovo Cimento A **88**, 1 (1985); **88**, 31 (1985).
[21] L. Alvarez-Gaumé and P. Ginsparg, Ann. Phys. (N.Y.) **161**, 423 (1985).
[22] L. Bonora, P. Pasti, and M. Tonin, J. Math. Phys. **27**, 2259 (1986).
[23] J. A. Dixon, Commun. Math. Phys. **139**, 495 (1991).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Classification of low-dimensional Lie super-bialgebras

C. Juszczak and J. T. Sobczyk
Institute of Theoretical Physics, University of Wrocław, Pl. Maksa Borna 9, 50-204 Wrocław, Poland

(Received 29 December 1997; accepted for publication 19 May 1998)

A thorough analysis of Lie super-bialgebra structures on Lie super-algebras osp(1;2) and super e(2) is presented. A combined technique of computer algebraic computations and subsequent identification of equivalent structures is applied. In all the cases, Poisson–Lie brackets on supergroups are found. The possibility of quantizing them in order to obtain quantum groups is discussed. It turns out to be straightforward for all but one structure for the super-E(2) group. © *1998 American Institute of Physics.* [S0022-2488(98)03409-4]

I. INTRODUCTION

About 10 years of intensive research in quantum groups[1,2] demonstrated the richness of investigated mathematical structures. No wonder that, except for efforts to find their physical applications, many authors tried to classify them from a mathematical point of view. The classifications are based on many different ideas. We do not wish to provide an exhaustive list of them but would like to mention some well-known examples. If the quantum SL(2) group is defined as one that preserves a nondegenerate bilinear form, two inequivalent deformations appear.[3] More complicated analysis of the group GL(3) leads to a conclusion that 26 inequivalent deformations exist in this case.[4] Reference 5 contains a very recent attempt to classify quantum SL(3) groups. A classification of all the 4×4 matrices satisfying the Yang–Baxter (YB) equation was done in Ref. 6. Later this result was used in order to list possible quantum deformations of the supergroup GL(1,1).[7] An interest in deformations of $D=4$ relativistic symmetries led to a classification of deformations of the Lorentz and Poincaré groups.[8]

From the early years of interest in quantum groups it is well known that they are closely related to notions of Lie bialgebras and Lie–Poisson groups. Suppose that a parameter q is introduced such that the value $q=1$ corresponds to an undeformed universal enveloping algebra $U(\mathcal{G})$ or a classical commutative algebra of functions on a Lie group G (with the Lie algebra $\mathcal{G}$). Then the structure of the Hopf algebra of the quantum group gives rise in the classical limit to an extra structure on G or $\mathcal{G}$. Then G becomes a Poisson–Lie group and $\mathcal{G}$ a Lie bialgebra.[9] These structures can be viewed as possible *directions* of quantum deformations. A fundamental fact about them is that all can be quantized, i.e., they arise as the classical limit from a *bona fide* quantum deformation.[10] This suggests that an effective approach to classify quantum groups could be by classifying their classical limits and that seems much easier to pull off. The study of Lie bialgebras becomes, in some cases, even easier, as they might be of the coboundary type, which happens if they can be described by means of a classical r-matrix satisfying the (modified) classical Yang–Baxter equation. Cohomological properties of semisimple Lie groups imply that all related Lie bialgebras are, in fact, coboundaries.[9] More elaborate argument shows that the same is true in the case of inhomogeneous groups of space–time symmetries for any signature of metric for dimensionality of space–time $D>2$.[11] Thus the problem of classification of Lie bialgebra structures contains as a subproblem a classification of solutions of the (modified) classical Yang–Baxter equation which has been studied by many authors. It should be stressed, however, that (in general) Lie algebras admit bialgebra structures that are not coboundaries. Recently, many low-dimensional Lie algebras were investigated for bialgebra structures, revealing a surprising number of possibilities. The two-dimensional Galilei algebra admits nine inequivalent Lie bialgebra structures and only one out of them is a coboundary.[12] The central extension of the Galilei algebra by the mass operator admits 26 inequivalent Lie bialgebra structures out of which eight are coboundaries.[13] Similar analysis aimed mainly at the construction of quantum groups was performed before in the cases of Heisenberg–Weyl and oscillator Lie algebras.[14] The classification of

 © 1998 American Institute of Physics

the possible classical r-matrices (and automatically of Lie bialgebra structures) was done in the case of Lorentz and Poincaré algebras.[15] Classical r-matrices for SL(3) were listed in Ref. 16. Corresponding R-matrices satisfying the Yang–Baxter equation were found in Ref. 17.

In this paper we will study quantum deformations of supergroups and their classical limits—Lie super-bialgebras. Quantum supergroups have been studied by many authors. Knowledge of R-matrices and low-dimensional representations of osp(1,2), su(1,1), gl(2,1), and sl(1,2) has been used to construct integrable models.[18] To our knowledge no systematic investigation of Lie super-bialgebras has yet been undertaken. We decided to study two cases: osp(1,2) and the supersymmetric extension of the e(2) algebra, both treated as complex. The first one is interesting as it plays in the supersymmetric case a role analogous to sl(2). Many papers devoted to quantum supergroups are based on a quantum deformation proposed by Ref. 19. A natural question is: are other deformations possible as is in the case of sl(2)? The supersymmetric extension of e(2) chosen as the e(2) group is one which has a simple structure but admits several inequivalent Lie bialgebras.[20,21] There are four of them, with one being, strictly speaking, a one-parameter family. Two structures are generated by classical r-matrices. All the quantum deformations of the group E(2) are known which could be helpful in finding quantum deformations in the supersymmetric case. Some quantum deformations of super-e(2) algebra have already been discussed in the literature. They can be [in analogy to the e(2) case] obtained by means of a contraction procedure.[22]

The problem of classifying Lie bialgebra structures for a given Lie algebra is mostly a computational one. We found it effective to use a computer at various steps. For a Lie algebra with N generators it is necessary to deal with equations containing the number of parameters which increases as N^3. Some of the equations are linear but the remaining (co-Jacobi identities) are quadratic. The complexity of the problem grows quickly with N. The biggest obstacle in continuing with this program for large N is primarily the fact that not everything can be done by computer. The most delicate part of the problem is to single out orbits under the action of the group of automorphisms of $\mathcal{G}$ in a computer-produced space of solutions.

Our first result is that in the osp(1,2) case all the Lie super-bialgebras are coboundaries. We found three independent classical r-matrices. One solution is quite obvious. As sl(2) is a Lie subalgebra of osp(1,2), any r-matrix for the former algebra satisfying the (not modified!) classical YB equation is automatically the r-matrix for the latter. The third solution corresponds to the quantum deformation of osp(1,2) described in Ref. 19. Relations for quantum OSp(1,2) are given in Ref. 23. The second solution is a new one and requires a detailed study.

In the case of super-e(2) we found six families of independent Lie super-bialgebra structures. In their most general form they are not coboundaries, but four of them contain coboundary members. Some links to the classification of analogous structures on e(2) can be established. Case (i) in our list is clearly analogous with δ_2 of Ref. 21. Case (iv) corresponds to δ_1. It is the infinitesimal form of quantum deformation introduced in Ref. 22. The existence of r-matrix r_2 is again obvious as it consists only of generators of e(2) and satisfies the (not modified!) classical YB equation.

For all the Lie super-bialgebra structures we calculated corresponding Lie–Poisson brackets. Their form is such that in five cases of the super-E(2) group it is easy to go through with the program of constructing quantum supergroups.[24,25] It is sufficient to change Lie–Poisson brackets into (anti) commutators.

The paper is organized as follows. In Sec. II all the basic concepts and notation used in the rest of the paper are introduced. Sections III and IV contain derivations of inequivalent Lie super-bialgebras for osp(1,2) and super-e(2). In Sec. V Poisson–Lie brackets for both cases are presented. In Sec. VI our conclusions and some final remarks are presented.

After completing our study we learned about the paper in which by applying the Drinfeld twisting procedure a new quantum supergroup structure on OSp(1,2) was found.[26] It is clear that it corresponds to the Lie super-bialgebra structure number (1) on our list presented in Sec. IV.

II. BASIC DEFINITIONS AND NOTATION

Super-Lie-algebra $\mathcal{G}$ (Ref. 27) is a graded vector space

$$\mathcal{G}=\mathcal{G}_B\oplus\mathcal{G}_F \tag{1}$$

with the grade function grade($\mathcal{G}_B$)=0, grade($\mathcal{G}_F$)=1. A Lie superalgebra structure is provided by a linear mapping

$$[,]:\mathcal{G}\otimes\mathcal{G}\to\mathcal{G} \tag{2}$$

satisfying the requirements of (graded) antisymmetry and Jacobi identity. In order to express them it is useful to introduce a basis in $\mathcal{G}$ $\{g_i\}\subset\mathcal{G}_B\cup\mathcal{G}_F$ and structure constants

$$[g_i,\ g_j]=c_{ij}{}^k g_k\,. \tag{3}$$

Structure constants have to satisfy

$$c_{ij}{}^k=0 \quad \text{whenever grade}\ (g_i)+\text{grade}\ (g_j)\neq\text{grade}\ (g_k)(\text{mod}\ 2), \tag{4}$$

$$c_{ij}{}^k=-z(i,j)c_{ji}{}^k, \tag{5}$$

and

$$c_{ij}{}^k c_{kl}{}^m z(i,l)+c_{jl}{}^k c_{ki}{}^m z(j,i)+c_{li}{}^k c_{kj}{}^m z(l,j)=0, \tag{6}$$

where

$$z(i,j)=(-1)^{\text{grad}\ (g_i)\text{grad}\ (g_j)}. \tag{7}$$

Lie super-bialgebra structure (see, e.g., Refs. 28 and 29) is a linear mapping

$$\delta:\mathcal{G}\to\mathcal{G}\otimes\mathcal{G}, \tag{8}$$

which in the chosen basis reads

$$\delta(g_i)=f_i{}^{kl}g_k\otimes g_l\,. \tag{9}$$

Here δ has to satisfy several requirements. First of all, it makes the dual linear space $\mathcal{G}^*$ a Lie super-algebra:

$$f_k{}^{ij}=0 \quad \text{whenever grad}\ (g_i)+\text{grad}\ (g_j)\neq\text{grad}\ (g_k)(\text{mod}\ 2), \tag{10}$$

$$f_k{}^{ij}=-z(i,j)f_k{}^{ji}, \tag{11}$$

$$f_i{}^{kj}f_j{}^{lm}z(k,m)+f_i{}^{lj}f_j{}^{mk}z(l,k)+f_i{}^{mj}f_j{}^{kl}z(m,l)=0. \tag{12}$$

Moreover, structure constants c and f have to be related:

$$c_{ij}{}^k f_k{}^{lm}=f_i{}^{lk}c_{kj}{}^m+c_{kj}{}^l f_i{}^{km}z(m,j)+c_{jk}{}^l f_j{}^{km}+f_j{}^{lk}c_{ik}{}^m z(i,l). \tag{13}$$

The coboundary Lie super-bialgebra is a pair $(\mathcal{G},r)$, where $\mathcal{G}$ is a Lie super-bialgebra and $r\in\mathcal{G}_B\wedge\mathcal{G}_B\oplus\mathcal{G}_F\wedge\mathcal{G}_F\subset\mathcal{G}\wedge\mathcal{G}$. such that for every $g_i\in\mathcal{G}$,

$$\delta(g_i)=[r,\ g_i\otimes 1+1\otimes g_i]. \tag{14}$$

The Schouten bracket is defined as

$$[[r,\ r]]\equiv[r_{12},\ r_{13}]+[r_{12},\ r_{23}]+[r_{13},\ r_{23}], \tag{15}$$

where $r_{12}=r\otimes 1$, $r_{23}=1\otimes r$, and $r_{13}=(1\otimes\tau)r_{12}$, where τ is linear and $\tau(g_i\otimes g_j)=z(i,j)g_j\otimes g_i$.

Here r satisfies the classical Yang–Baxter equation (CYBE) if

$$[[r,\ r]]=0 \tag{16}$$

and the modified CYBE if

$$\forall g_i \in \mathcal{G} \quad [[[r,\, r]],\ g_i \otimes 1 \otimes 1 + 1 \otimes g_i \otimes 1 + 1 \otimes 1 \otimes g_i] = 0. \tag{17}$$

The super-e(2) Lie algebra is spanned by the set of generators $\{H, P_+, P_-, D_+, D_-\}$ of grading $\mathrm{grad}\,(H) = \mathrm{grad}\,(P_\pm) = 0$ and $\mathrm{grad}\,(D_\pm) = 1$, which fulfill the following (anti) commutation relations:

$$\begin{gathered}
[H,\, P_\pm] = \pm P_\pm, \quad [H,\, D_\pm] = \pm \tfrac{1}{2} D_\pm, \quad [P_+,\, P_-] = 0, \\
\{D_+,\, D_-\} = 0, \quad \{D_\pm,\, D_\pm\} = P_\pm, \\
[P_\pm,\, D_\pm] = 0, \quad [P_\pm,\, D_\mp] = 0.
\end{gathered} \tag{18}$$

A convenient parametrization of the classical super-E(2) group is obtained by means of exponentiation:

$$g(s,a,b,\xi,\eta) = \exp(sH)\exp(aP_+)\exp(bP_-)\exp(\xi D_+)\exp(\eta D_-). \tag{19}$$

Coproducts are

$$\Delta(s) = s \otimes 1 + 1 \otimes s, \tag{20}$$

$$\Delta(a) = 1 \otimes a + a \otimes \exp(-s) + \frac{1}{2}\, \xi \otimes \xi\, \exp\left(-\frac{s}{2}\right), \tag{21}$$

$$\Delta(b) = 1 \otimes b + b \otimes \exp(s) + \frac{1}{2}\, \eta \otimes \eta\, \exp\left(\frac{s}{2}\right), \tag{22}$$

$$\Delta(\xi) = 1 \otimes \xi + \xi \otimes \exp\left(-\frac{s}{2}\right), \tag{23}$$

$$\Delta(\eta) = 1 \otimes \eta + \eta \otimes \exp\left(\frac{s}{2}\right). \tag{24}$$

The Lie superalgebra osp(1,2) is spanned by the set of generators $\{H, X_+, X_-, V_+, V_-\}$ of grading $\mathrm{grad}(H) = \mathrm{grad}(X_\pm) = 0$ and $\mathrm{grad}(V_\pm) = 1$, which fulfill the following (anti) commutation relations:

$$\begin{gathered}
[H,\, X_\pm] = \pm X_\pm, \quad [H,\, V_\pm] = \pm \tfrac{1}{2} V_\pm, \quad [X_+,\, X_-] = 2H, \\
\{V_+,\, V_-\} = -\tfrac{1}{2} H, \quad \{V_\pm,\, V_\pm\} = \pm \tfrac{1}{2} X_\pm, \\
[X_\pm,\, V_\pm] = 0, \quad [X_\pm,\, V_\mp] = V_\pm.
\end{gathered} \tag{25}$$

In the case of supergroup OSp(1,2) we will use more implicit parametrization by means of 3×3 supermatrices subject to certain constraints:[19]

$$g = \begin{pmatrix} a & \alpha & b \\ \gamma & e & \beta \\ c & \delta & d \end{pmatrix}, \tag{26}$$

where $e = 1 + \alpha\delta$, $\gamma = c\alpha - a\delta$, $\beta = d\alpha - b\delta$, and $ad - bc + \alpha\delta = 1$. Variables denoted by Greek letters are of Grassmanian type. Coproducts follow from matrix multiplication of elements of G.

III. LIE SUPER-BIALGEBRAS FOR osp(1,2)

The problem studied in Secs. III and IV can be formulated in the following way:

Given a set of structure constants $c_{ij}{}^{k}$, find all sets of structure constants $f_{m}{}^{np}$ that give rise to Lie super-bialgebras. Two such sets are considered equivalent if they can be made equal by a change of the basis in $\mathscr{G}$. Verify if obtained structures are of coboundary type.

Initial steps in the analysis can be made using a computer. A most general form of f's is assumed. Constraints imposed by the set of linear equations coming from the cocycle condition (13) are first taken into account. Then the set of quadratic equations coming from the super co-Jacobi conditions (12) is to be solved.

In the case of osp(1,2) we discover that all the possible bialgebra structures are coboundaries and that the classical r-matrix is in one of two possible forms:

$$\begin{aligned} r_a &= x(X_+\wedge X_- + 2V_+\wedge V_-) + y(H\wedge X_+ - V_+\wedge V_+) + z(H\wedge X_- - V_-\wedge V_-), \\ r_b &= \pm\sqrt{uv}X_+\wedge X_- + uH\wedge X_+ + vH\wedge X_-, \end{aligned} \tag{27}$$

where x, y, z, u, v are arbitrary complex numbers.

Identification of automorphisms of superalgebra osp(1,2) is fairly straightforward. They cannot mix fermions with bosons (grading is preserved). As the superalgebra is generated by just the two fermions V_+ and V_-, every automorphism is determined by the following transformation of the fermions:

$$\tilde{V}_+ = aV_+ + bV_-, \quad \tilde{V}_- = cV_+ + dV_-. \tag{28}$$

We easily derive that under the above transformation

$$\begin{aligned} \tilde{H} &= -acX_+ + (ad+bc)H + bdX_-, \\ \tilde{X}_+ &= a^2X_+ - 2abH - b^2X_-, \\ \tilde{X}_- &= -c^2X_+ + 2cdH + d^2X_-. \end{aligned} \tag{29}$$

The operators $\{\tilde{H},\tilde{X}_+,\tilde{X}_-,\tilde{V}_+,\tilde{V}_-\}$ obey the same super-commutation relations as $\{H,X_+,X_-,V_+,V_-\}$ if and only if

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = 1. \tag{30}$$

Under the transformations (28) and (29) the parameters x, y, z of the r-matrix r_a given in (27) transform as follows:

$$\begin{pmatrix} \tilde{y} & -\tilde{x} \\ -\tilde{x} & \tilde{z} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} y & -x \\ -x & z \end{pmatrix}\begin{pmatrix} a & c \\ b & d \end{pmatrix}. \tag{31}$$

We notice that this is exactly the way the symmetric form transforms under the change of basis. If we take into account the Sylvester theorem we see that we can make the matrix

$$\begin{pmatrix} y & -x \\ -x & z \end{pmatrix}$$

equal diag (y,y) or diag $(1,0)$ [without the condition (30) it would be possible to make it equal diag $(1,1)$ or diag $(1,0)$].

Thus r_a, if different from zero, is equivalent to one of the following:

$$r_a = \begin{cases} r_2 \equiv H\wedge X_+ - V_+\wedge V_+, & \text{if } x^2 - yz = 0 \\ r_3 \equiv t(H\wedge X_+ - V_+\wedge V_+ + H\wedge X_- - V_-\wedge V_-), & \text{if } x^2 - yz \neq 0. \end{cases} \tag{32}$$

By means of analogous reasoning we arrive at the conclusion that the r-matrix r_b is always equivalent to

$$r_1 \equiv H \wedge X_+ . \tag{33}$$

Here r_3 can also be written in equivalent form as $\tilde{t}(X_+ \wedge X_- + 2V_+ \wedge V_-)$. This is a way to see that it corresponds to the deformation of osp(1,2) presented in Ref. 19. As it stands it satisfies only modified CYBE. One can, however, supplement it by the symmetric ad-invariant expression

$$\tilde{t}(H \otimes H + X_+ \otimes X_- + X_- \otimes X_+ + 2(V_+ \otimes V_- - V_- \otimes V_+)) \tag{34}$$

and obtain an r-matrix satisfying the (not modified) CYBE. This is a way in which a universal R-matrix with the classical limit r_3 can satisfy the Yang–Baxter equation and the quasitriangular deformation of osp(1,2) is obtained.

The parameter t is present in (32) as it cannot be removed by means of the transformation (31). It means that for different values of t we have inequivalent Lie superbialgebras. However, t becomes irrelevant upon quantization since it can be absorbed into a deformation parameter.

IV. LIE SUPER-BIALGEBRAS FOR SUPER-e(2)

The initial steps of the analysis are made using a computer as in the case of osp(1,2). After solving equations quadratic in structure constants it turns out that most of the bialgebra structures are not coboundaries. This makes the investigation more involved since we cannot use the r-matrix formulation and must explicitly state the co-Lie structure for each solution found. The possibilities found by means of the computer are:

Case A:

$$\begin{aligned}
&\delta(H) = H \wedge (aP_+ + bP_-) + cP_+ \wedge P_- , \\
&\delta(P_+) = P_+ \wedge bP_- , \quad \delta(P_-) = -aP_+ \wedge P_- , \\
&\delta(D_+) = \tfrac{1}{2}(aP_+ - bP_-) \wedge D_+ \pm \sqrt{ab} P_+ \wedge D_- , \\
&\delta(D_-) = \tfrac{1}{2}(aP_+ - bP_-) \wedge D_- \pm \sqrt{ab} P_- \wedge D_+ ;
\end{aligned} \tag{35}$$

Case B:

$$\begin{aligned}
&\delta(H) = a(H \wedge P_+ - \tfrac{1}{2} D_+ \wedge D_+) + b(H \wedge P_- + \tfrac{1}{2} D_- \wedge D_-) + cP_+ \wedge P_- , \\
&\delta(P_+) = P_+ \wedge bP_- + d(2H \wedge P_+ - D_+ \wedge D_+), \\
&\delta(P_-) = -aP_+ \wedge P_- + d(2H \wedge P_- + D_- \wedge D_-), \\
&\delta(D_+) = -\tfrac{1}{2}(aP_+ + bP_-) \wedge D_+ + d(H \wedge D_+), \\
&\delta(D_-) = -\tfrac{1}{2}(aP_+ + bP_-) \wedge D_- + d(H \wedge D_-),
\end{aligned} \tag{36}$$

where a, b, c, d are arbitrary complex numbers such that $cd = 0$.

The set of automorphisms of super-e(2) is generated by three transformations:

(1) $\tilde{H} = H + \alpha P_+ + \beta P_- , \quad \tilde{P}_\pm = P_\pm , \quad \tilde{D}_\pm = D_\pm .$
(2) $\tilde{H} = -H, \quad \tilde{P}_\pm = P_\mp , \quad \tilde{D}_\pm = D_\mp .$
(3) $\tilde{H} = H, \quad \tilde{P}_+ = \alpha^2 P_+ , \quad \tilde{D}_+ = \alpha D_+ , \quad \tilde{P}_- = \beta^2 P_- , \quad \tilde{D}_- = \beta D_- .$

When taken into account they lead to a conclusion that a and b can be scaled out to take the value of 1 or 0. It is also possible to show that when $d \neq 0$, then by taking $\tilde{H} = H + (a/2d)P_+ + (b/2d)P_-$, we can make $a = b = c = 0$.

Finally we arrive at six families of bialgebra structures whose representatives can be described by means of the following substitutions:

(i) Case A where $a = b = 0$,

(ii) Case A where $a=1$, $b=0$,
(iii) Case A where $a=b=1$,
(iv) Case B where $a=b=c=0$,
(v) Case B where $a=1$, $b=d=0$,
(vi) Case B where $a=b=1$, $d=0$.

Classical r-matrices exist if both c and d vanish. Their general forms for cases A and B are

$$r_{\mathrm{A}}=aH\wedge P_{+}-bH\wedge P_{-}+\sqrt{ab}D_{+}\wedge D_{-}+fP_{+}\wedge P_{-},$$
$$r_{\mathrm{B}}=a(H\wedge P_{+}-\tfrac{1}{2}D_{+}\wedge D_{+})-b(H\wedge P_{-}+\tfrac{1}{2}D_{-}\wedge D_{-})+fP_{+}\wedge P_{-}. \tag{37}$$

In both cases the term $fP_{+}\wedge P_{-}$ is irrelevant since its commutators (14) with all the generators of super-e(2) vanish.

After substitutions we obtain

$$r_{(\mathrm{ii})}=H\wedge P_{+},$$
$$r_{(\mathrm{iii})}=H\wedge P_{+}-H\wedge P_{-}+D_{+}\wedge D_{-},$$
$$r_{(\mathrm{v})}=H\wedge P_{+}-\tfrac{1}{2}D_{+}\wedge D_{+}, \tag{38}$$
$$r_{(\mathrm{vi})}=H\wedge P_{+}-\tfrac{1}{2}D_{+}\wedge D_{+}-H\wedge P_{-}-\tfrac{1}{2}D_{-}\wedge D_{-}.$$

Only $r_{(\mathrm{ii})}$ and $r_{(\mathrm{v})}$ satisfy the CYBE.

V. POISSON–LIE BRACKETS

Lie super-bialgebras are in 1:1 correspondence with Poisson–Lie structures on supergroups.[28] Poisson–Lie brackets satisfy the following properties:

$$\{f,g\}=-z(f,g)\{g,f\}, \tag{39}$$
$$\{f,gh\}=\{f,g\}h+z(f,g)g\{f,h\}, \tag{40}$$
$$0=z(f,h)\{f,\{g,h\}\}+z(g,f\,)\{g,\{h,f\}\}+z(h,g)\{h,\{f,g\}\}, \tag{41}$$
$$\Delta\{f,g\}=\{\Delta f,\Delta g\}. \tag{42}$$

Poisson–Lie brackets are most easily introduced by means of left- and right-invariant vector fields on a supergroup. In the super case one should distinguish left- and right-hand side derivatives of superfunctions. In the coboundary case with a classical r-matrix r_{kl}, Poisson–Lie brackets are given by[25]

$$\{\phi,\psi\}=(Y_k^{(r)}\phi)r_{kj}(Y_j^{(l)}\psi)-(X_k^{(r)}\phi)r_{kj}(X_j^{(l)}\psi), \tag{43}$$

where $Y_k^{(r,l)}$ denotes left-invariant right (r) or left (l) derivatives and $X_k^{(r,l)}$ right-invariant right (r) or left (l) derivatives. Here Y_k and X_k can be derived from the coproducts.

We present in the following how they act on generators of both supergroups.

(a) Super-E(2)

	a	b	s	ξ	η
$Y_H^{(r,l)}$	$-a$	b	1	$-\xi/2$	$\eta/2$
$Y_{P_+}^{(r,l)}$	1	0	0	0	0
$Y_{P_-}^{(r,l)}$	0	1	0	0	0
$Y_{D_-}^{(r)}$	0	$\eta/2$	0	0	1
$Y_{D_-}^{(l)}$	0	$-\eta/2$	0	0	1
$Y_{D_+}^{(r)}$	$\xi/2$	0	0	1	0
$Y_{D_+}^{(l)}$	$-\xi/2$	0	0	1	0
$X_H^{(r,l)}$	0	0	1	0	0
$X_{P_+}^{(r,l)}$	$\exp(-s)$	0	0	0	0
$X_{P_-}^{(r,l)}$	0	$\exp(s)$	0	0	0
$X_{D_-}^{(r)}$	0	$-\eta\exp(s/2)/2$	0	0	$\exp(s/2)$
$X_{D_-}^{(l)}$	0	$\eta\exp(s/2)/2$	0	0	$\exp(s/2)$
$X_{D_+}^{(r)}$	$-\xi\exp(-s/2)/2$	0	0	$\exp(-s/2)$	0
$X_{D_+}^{(l)}$	$\xi\exp(-s/2)/2$	0	0	$\exp(-s/2)$	0

(44)

(b) OSp(1,2):

	a	α	b	c	δ	d
$Y_H^{(r,l)}$	$a/2$	0	$-b/2$	$c/2$	0	$-d/2$
$Y_{X_+}^{(r,l)}$	0	0	a	0	0	c
$Y_{X_-}^{(r,l)}$	b	0	0	d	0	0
$Y_{V_+}^{(r)}$	0	$a/2$	$\alpha/2$	0	$c/2$	$\delta/2$
$Y_{V_+}^{(l)}$	0	$a/2$	$-\alpha/2$	0	$c/2$	$-\delta/2$
$Y_{V_-}^{(r)}$	$-\alpha/2$	$b/2$	0	$-\delta/2$	$d/2$	0
$Y_{V_-}^{(l)}$	$\alpha/2$	$b/2$	0	$\delta/2$	$d/2$	0
$X_H^{(r,l)}$	$a/2$	$\alpha/2$	$b/2$	$-c/2$	$-\delta/2$	$-d/2$
$X_{X_+}^{(r,l)}$	c	δ	d	0	0	0
$X_{X_-}^{(r,l)}$	0	0	0	a	α	b
$X_{V_+}^{(r)}$	$-\gamma/2$	$e/2$	$-\beta/2$	0	0	0
$X_{V_+}^{(l)}$	$\gamma/2$	$e/2$	$\beta/2$	0	0	0
$X_{V_-}^{(r)}$	0	0	0	$-\gamma/2$	$e/2$	$-\beta/2$
$X_{V_-}^{(l)}$	0	0	0	$\gamma/2$	$e/2$	$\beta/2$

(45)

In the non-coboundary cases Poisson brackets can be calculated by applying the supersymmetric version of the method described in Ref. 21. It is necessary to solve the cocycle equation for an element $\Phi: G \to \mathcal{G} \wedge \mathcal{G}$,

$$\Phi(g_1 g_2) = \Phi(g_1) + g_1 \Phi(g_2) g_1^{-1}, \tag{46}$$

TABLE I. Poisson Lie structures for the group OSp(1,2).

	1	2	3[a]
$\{a,b\}$	$a^2+\alpha\delta-1$	a^2-1	a^2+b^2-1
$\{a,c\}$	$-c^2$	$-c^2$	$1-c^2-a^2$
$\{a,d\}$	$c(a-d)$	$c(a-d)$	$(c-b)(a-d)$
$\{b,c\}$	$-c(a+d)$	$-c(a+d)$	$-(b+c)(a+d)$
$\{b,d\}$	$1-d^2-\alpha\delta$	$1-d^2$	$1-b^2-d^2$
$\{c,d\}$	c^2	c^2	c^2+d^2-1
$\{a,\alpha\}$	$c\alpha-a\delta$		$b\alpha$
$\{b,\alpha\}$	$d\alpha-b\delta$	$-a\alpha$	$-a\alpha$
$\{c,\alpha\}$	$c\delta$	$c\delta$	$c\delta+a\alpha-b\delta$
$\{d,\alpha\}$	$d\delta$	$(d-a)\delta$	$d\delta+b\alpha-a\delta$
$\{a,\delta\}$	$-c\delta$	$-c\delta$	$-c\delta-a\alpha+d\alpha$
$\{b,\delta\}$	$-d\delta$	$-d\delta-c\alpha$	$-(d\delta+b\alpha+c\alpha)$
$\{c,\delta\}$			$d\delta$
$\{d,\delta\}$		$-c\delta$	$-c\delta$
$\{\alpha,\alpha\}$	$2\delta\alpha$	$1-a^2$	$1-a^2-b^2$
$\{\alpha,\delta\}$		$-ac$	$-ac-bd$
$\{\delta,\delta\}$		$-c^2$	$1-c^2-d^2$

[a]All the entries should be multiplied by the parameter t [see (32)].

satisfying ''initial conditions'' determined by the Lie super-bialgebra structure in consideration.

The whole procedure has to be followed in every detail only in the case of Lie super-bialgebra structures (i) and (iv) from the list presented in Sec. III. One finds

$$\Phi_{(\mathrm{i})}(g)=csP_+\wedge P_-, \tag{47}$$

$$\begin{aligned}\Phi_{(\mathrm{iv})}(g)=&-2ae^{s}P_+\wedge H-ae^{s}D_+\wedge D_+-2be^{-s}P_-\wedge H\\&+2abP_-\wedge P_++be^{-s}D_-\wedge D_-+\xi e^{s/2}H\wedge D_+\\&-a\xi e^{3s/2}P_+\wedge D_++\xi be^{-s/2}P_-\wedge D_++\eta e^{-s/2}H\wedge D_-\\&-a\eta e^{s/2}P_+\wedge D_-+\eta be^{-3s/2}P_-\wedge D_--\tfrac{1}{2}\xi\eta D_+\wedge D_-.\end{aligned} \tag{48}$$

For $\Phi(g)=\Phi^{jk}(g)g_j\wedge g_k$ the Poisson–Lie brackets are

$$\{\phi,\psi\}=(X_j^{(r)}\phi)\Phi^{jk}(X_k^{(l)}\psi). \tag{49}$$

In all the remaining cases one can use the fact that they become coboundaries when $c=0$. The parameter c is present in $\delta(H)$ as $cP_+\wedge P_-$. Clearly it appears in Φ as $csP_+\wedge P_-$. It is thus possible to calculate Poisson–Lie brackets using classical r-matrices given in (38) and to add the

TABLE II. Poisson Lie structures for the group super-E(2).

	(i)	(ii)	(iii)	(iv)[a]	(v)	(vi)
$\{a,b\}$	cs	$-b+cs$	$a-b+cs$	$-2ab$	$-b+cs$	$a-b+cs$
$\{a,e^s\}$		$1-e^s$	$1-e^s$	$-2ae^s$	$1-e^s$	$1-e^s$
$\{b,e^s\}$			e^s-e^{2s}	$-2be^s$		e^s-e^{2s}
$\{a,\xi\}$		$\xi/2$	$\xi/2$		$-\xi e^{-s}/2$	$-\xi e^{-s}/2$
$\{a,\eta\}$		$-\eta/2$	$\xi-(\eta/2)$	$-a\eta$	$-\eta/2$	$-\eta/2$
$\{b,\xi\}$			$\eta-(\xi/2)$	ξb		$-\xi/2$
$\{b,\eta\}$			$\eta/2$			$-\eta e^s/2$
$\{e^s,\xi\}$				ξe^s		
$\{e^s,\eta\}$				ηe^s		
$\{\xi,\xi\}$				$-2a$	$e^{-s}-1$	$e^{-s}-1$
$\{\xi,\eta\}$				$-\xi\eta/2$		
$\{\eta,\eta\}$				$2b$		e^s-1

[a]All the entries should be multiplied by the parameter d [see (36)].

extra term to $\{a,b\}$ at the very end of computations. The complete sets of relations making OSp(1,2) and super-E(2) Poisson–Lie supergroups are given in Tables I and II. In the case of OSp(1,2), in order to obtain shorter formulas all the Poisson brackets have been multiplied by the factor 2. The fact that these Poisson–Lie brackets satisfy conditions (39)–(42) was verified explicitly with the use of a computer.

VI. DISCUSSION AND FINAL REMARKS

The relevance of the knowledge of Poisson–Lie brackets lies in the fact that it gives a deep insight into a structure of the corresponding quantum group, in some cases determining it completely.[24]

It is easy to see that in five cases of brackets on the super-E(2) group [an exception is the structure (iv)] it is straightforward to integrate it and obtain a quantum group. In fact, the rhs's of Poisson brackets contain no products of functions which, upon quantization, could lead to ordering ambiguities. It is therefore sufficient to replace $\{,\}\rightarrow(1/h)[,]$ where h is a quantum deformation parameter. By applying duality techniques it must be then possible to obtain quantum deformations of super-e(2) enveloping algebra. Deformation of the super-e(2) enveloping algebra discussed in Ref. 22 corresponds to the structure (iv) of our classification.

In the case of OSp(1,2) a quantization of the first structure appeared very recently in Ref. 26. The third structure is the standard one introduced in Ref. 19. A quantum version of the second structure is missing.

ACKNOWLEDGMENTS

The authors would like to thank Professor J. Lukierski for valuable discussions. The support of KBN Grant No. 2P 30208706 is also acknowledged.

[1] L. D. Faddeev, N. Yu. Reshetikhin, and L. A. Takhtajan, Leningrad Math. J. **1**, 193 (1990).
[2] V. G. Drinfeld, Quantum Groups, Proc. ICM (Berkeley 1986) (Academic, New York, 1987) Vol. 1, p. 798–820.
[3] M. Dubois-Violette and G. Launer, Phys. Lett. B **245**, 175 (1990).
[4] H. Ewen and O. Ogievetsky, ''Classification of GL(3) quantum matrix groups,'' preprint LMV-TPW 94/23 Munich, 1994.
[5] Ch. Ohn, ''Quantum SL(3) with classical representation theory,'' University of Reins, preprint, 1997.
[6] J. Hietarinta, J. Math. Phys. **34**, 1725 (1993).
[7] L. Frappat, V. Hussin, and G. Rideau, ''Classification of the quantum deformations of the superalgebra $gl(1|1)$,'' preprint CRM 2476, Montreal 1997.
[8] S. Woronowicz and S. Zakrzewski, Compositio Mathematica **90**, 211 (1994); P. Podleś and S. Woronowicz, Commun. Math. Phys. **178**, 61 (1996).
[9] V. G. Drinfeld, Sov. Math. Dokl. **27**, 68 (1983).
[10] N. Reshetikhin, Duke Math. J. **61**, 143 (1992); P. Etingof and D. Kazhdan, Selecta Math. **2** (N1), 1 (1996).
[11] S. Zakrzewski, ''Poisson-Poincaré Groups,'' hep-th/9412099.
[12] E. Kowalczyk, Acta Phys. Pol. B **28**, 1893 (1997).
[13] A. Opanowicz, ''Lie Bialgebra Structures for Centrally Extended Two-Dimensional Galilei Algebra and their Lie-Poisson Counterparts,'' q-alg 9710028.
[14] A. Ballesteros and F. J. Herranz, J. Phys. A **29**, 4307 (1996); A. Ballesteros, F. J. Herranz, and P. Parashar, ''Quantum Heisenberg-Weyl Algebras,'' q-alg 9610007.
[15] S. Zakrzewski, Lett. Math. Phys. **32**, 11 (1994).
[16] A. Stolin, Math. Scand. **69**, 81 (1991).
[17] M. Gerstenhaber and A. Giaquinto, ''Boundary Solutions of the Quantum Yang-Baxter Equation and Solutions in Three Dimensions,'' q-alg 9710033.
[18] H. Saleur, Nucl. Phys. B **336**, 363 (1990); H. Hinrichsen and V. Rittenberg, Phys. Lett. B **275**, 350 (1992); M. D. Gould, K. E. Hibberd, J. R. Links, and Y.-Z. Zhang, Phys. Lett. A **212**, 156 (1996); D. Arnaudon, ''Algebraic approach to q-deformed supersymmetric variants of the Hubbard model with pair hoppings,'' preprint ENSLAPP-A-650/97, Lyon, 1997.
[19] P. P. Kulish, ''Quantum superalgebra osp(2|1),'' preprint RIMS-615, Kyoto, 1988; P. P. Kulish and N. Yu. Reshetikhin, Lett. Math. Phys. **18**, 143 (1989).
[20] S. Zakrzewski, ''Poisson space-time symmetry and corresponding elementary systems,'' in *Proceedings of the II International Wigner Symposium*, Goslar 1991, edited by H. D. Doebner and V. K. Dobrev, pp. 111–123.
[21] J. Sobczyk, J. Phys. A **29**, 2887 (1996).
[22] E. Celeghini, R. Giachetti, P. P. Kulish, E. Sorace, and M. Tarlini, J. Phys. A **24**, 5675 (1991).
[23] M. Chaichian and P. P. Kulish, ''Quantum Group Covariant Systems,'' in *From Field Theory to Quantum Groups*,

birthday volume dedicated to Jerzy Lukierski, edited by B. Jancewicz and J. Sobczyk (World Scientific, Singapore, 1996), pp. 99–111.
[24] S. Zakrzewski, J. Phys. A **27**, 2075 (1994).
[25] P. Kosiński, J. Lukierski, P. Maślanka, and J. Sobczyk, J. Phys. A **27**, 6827 (1994).
[26] E. Celeghini and P. P. Kulish, J. Phys. A **31**, L79 (1998).
[27] V. G. Kac, Commun. Math. Phys. **53**, 31 (1977).
[28] N. Andruskiewitsch, Abn. Math. Sem. Univ. Hamburg **63**, 147 (1993).
[29] L. Pittner and P. Uray, J. Math. Phys. **36**, 944 (1995).

Intertwining operators for a degenerate double affine Hecke algebra and multivariable orthogonal polynomials

Saburo Kakei[a)]
Department of Mathematical Sciences, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153-0041, Japan

(Received 18 February 1998; accepted for publication 6 May 1998)

Several properties of the multivariable Hermite and Laguerre polynomials associated with the quantum Calogero models are investigated by using the operators that intertwine representations of a degenerate version of the double affine Hecke algebra. As applications, raising operators and shift operators for the polynomials are constructed in a unified manner. © *1998 American Institute of Physics.* [S0022-2488(98)00909-8]

I. INTRODUCTION

There are intimate relations between quantum mechanics and special functions. Wave functions for some systems can be explicitly written in terms of suitable special functions. Recent studies on integrable quantum many-particle systems reveal that wave functions of some special cases can be written in terms of multivariable analogs of classical orthogonal polynomials.[1–9] In the present paper, we shall consider orthogonal polynomials associated with the quantum Calogero models confined in harmonic potential:[10–13]

$$\mathscr{H}_A=\frac{1}{2}\sum_{j=1}^{N}\left(-\frac{\partial^2}{\partial x_j^2}+x_j^2\right)+\sum_{j<k}\frac{\beta(\beta-1)}{(x_j-x_k)^2}, \tag{1}$$

$$\mathscr{H}_B=\frac{1}{2}\sum_{j=1}^{N}\left\{-\frac{\partial^2}{\partial z_j^2}+z_j^2+\frac{\gamma(\gamma-1)}{z_j^2}\right\}+\sum_{j<k}\left\{\frac{\beta(\beta-1)}{(z_j-z_k)^2}+\frac{\beta(\beta-1)}{(z_j+z_k)^2}\right\}, \tag{2}$$

where we assume that β is a non-negative integer. The subscripts ''A'' and ''B'' signify that the corresponding Hamiltonian is invariant under the action of the Weyl group of A_{N-1} type or B_N type, respectively. (For the latter convenience, we use the letter ''z'' as the coordinates of the B_N-type model.) In these cases, the polynomial part of the wave functions can be regarded as the multivariable generalization of the Hermite (A_{N-1} case) and Laguerre (B_N case) polynomials and has been studied by several authors.[1–9]

Since special functions are related to representation theory, it may be challenging to investigate the algebraic aspect of the multivariable orthogonal polynomials. In the case of the Macdonald polynomials, it has been revealed that the algebraic structure behind the polynomials is the double affine Hecke algebras.[14–17] Since the Jack polynomials can be regarded as a degenerate case of the Macdonald polynomials, their algebraic structure is a degenerate version of the double affine Hecke algebra.[18,19] In previous works,[5,6] we have shown that the Hamiltonians (1) and (2) share the same algebraic structure, that is, a subalgebra of the degenerate double affine Hecke algebra. From this viewpoint, each of the models (1) and (2) corresponds to an individual representation of the algebra.

In this paper, we will investigate several properties of the multivariable Hermite and Laguerre polynomials by using intertwining operators between representation spaces. As applications of the intertwining operators, we will construct raising operators and shift operators for such polynomials. In the case of the Jack polynomials and the multivariable Hermite polynomials, raising operators have already been constructed.[8,20] Our raising operators constructed below have a different

[a)]Electronic mail: kakei@poisson.ms.u-tokyo.ac.jp

 © 1998 American Institute of Physics

form but share the same properties (see Sec. III). We would like to stress that our algebraic method gives a unified viewpoint in order to understand the common structure of the models.

This article is organized as follows. In Sec. II, we recall the relation between Dunkl-type operators and multivariable orthogonal polynomials. The main tools in this article, the intertwining operators, are defined also in Sec. II. Using the intertwining operators, we construct raising operators in Sec. III and shift operators in Sec. IV. The shift operators give an algebraic proof of the norm formula of the multivariable orthogonal polynomials. The last section is devoted to concluding remarks.

II. DUNKL-TYPE OPERATORS AND MULTIVARIABLE ORTHOGONAL POLYNOMIALS

In this section, we define our notation and review the theory of several kinds of multivariable orthogonal polynomials.

In Ref. 21, Dunkl has introduced differential-exchange operators, now called ''Dunkl operators,'' which are associated with root systems. For the A_{N-1}-type root system, the operators are defined as

$$D_j^A=\frac{\partial}{\partial x_j}+\beta\sum_{k(\neq j)}\frac{1-s_{jk}}{x_j-x_k},\quad j=1,\ldots,N,$$

where s_{ij} are elements of the symmetric group $\mathfrak{S}_N$. An element s_{ij} acts on functions of $x_1,\ldots,x_N$ as an operator which permutes arguments x_i and x_j. We remark that the operators D_j preserve the space of the polynomials of variables $x_1,\ldots,x_N$, which we denote $\mathbb{C}[x]$.

Cherednik introduced another version of the Dunkl operators that mutually commute[18] (see also Refs. 22 and 23):

$$\hat{D}_j^A=x_jD_j^A+\beta\sum_{k(<j)}s_{jk}. \tag{3}$$

The algebra generated by the elements $x_j^{\pm 1}$, $\hat{D}_j^A$, and s_{ij} is isomorphic to the degenerate double affine Hecke algebra $\mathfrak{H}'$ associated with the A_{N-1}-type root system.[18,19] We remark that the elements $x_j^{\pm 1}$, D_j^A, and s_{ij} also generate $\mathfrak{H}'$ since D_j^A and $\hat{D}_j^A$ are related through (3). We denote by $\mathfrak{H}_0'$ the subalgebra of $\mathfrak{H}'$ generated by $\hat{D}_j^A$ and s_{ij}, which is isomorphic to the degenerate affine Hecke algebra. We further denote by $\tilde{\mathfrak{H}}'$ the subalgebra of $\mathfrak{H}'$ generated by x_j, $\hat{D}_j^A$, and s_{ij}. In terms of generators, the defining relations are

$$[\hat{D}_i^A,\hat{D}_j^A]=[x_i,x_j]=0,$$

$$s_j^2=1,\quad s_js_{j+1}s_j=s_{j+1}s_js_{j+1},$$

$$[s_i,s_j]=0,\quad |i-j|\neq 1,$$

$$x_is_{ij}=s_{ij}x_j,\quad x_is_{jk}=s_{jk}x_i,\quad i\neq j,k,$$

$$\hat{D}_{j+1}^As_j-s_j\hat{D}_j^A=\beta,\quad s_j\hat{D}_{j+1}^A-\hat{D}_j^As_j=\beta,$$

$$[s_i,\hat{D}_j^A]=0,\quad j\neq i,i+1,$$

$$[\hat{D}_i^A,x_j]=\begin{cases}-\beta x_js_{ij}, & i>i,\\ x_i+\beta\left(\displaystyle\sum_{k(<i)}x_ks_{ik}+\sum_{k(>i)}x_is_{ik}\right), & i=j,\\ -\beta x_is_{ij}, & i<j,\end{cases}$$

where $s_j=s_{j,j+1}$ $(j=1,\ldots,n-1)$ are the simple transpositions.

For a partition $\lambda=(\lambda_1,\ldots,\lambda_N)$ and $w\in\mathfrak{S}_N$, we introduce the notation $x_w^\lambda=x_{w(1)}^{\lambda_j}\cdots x_{w(N)}^{\lambda_N}$. The action of $\hat{D}_j^A$ on monomials x_w^λ is given explicitly by[22–24]

$$\hat{D}_j^A x_w^\lambda = (w(\lambda+\beta\delta))_j x_w^\lambda + \sum_{(\mu,w')<(\lambda,w)} u_{ww'}^{\lambda\mu} x_{w'}^\mu, \tag{4}$$

with $\delta=(N-1,N-2,\ldots,0)$. Here the ordering $<$ is defined as follows: For two pairs (λ,w), (μ,w'), where λ, μ are partitions and $w,w'\in\mathfrak{S}_N$, we define the ordering $<$ as

$$(\mu,w')<(\lambda,w)\Leftrightarrow\begin{cases}\text{(i)} & \mu<_{\mathrm{D}}\lambda,\\ \text{(ii)} & \text{if } \mu=\lambda, \text{ then } w'<_{\mathrm{B}}w,\end{cases}$$

where $<_{\mathrm{D}}$ is the dominance ordering for partitions,[25] and $<_{\mathrm{B}}$ is the Bruhat ordering for the elements of $\mathfrak{S}_N$ (see, for example, Ref. 26). Since the action (4) is triangular, one can diagonalize the operators $\hat{D}_j^A$ simultaneously by a suitable choice of bases of $\mathbb{C}[x]$.[22–24] Such a basis is called a *nonsymmetric Jack polynomial*.

Definition 2.1:[22–24] *A nonsymmetric Jack polynomial $E_w^\lambda(x)$, labeled with the partition $\lambda=(\lambda_1,\ldots,\lambda_N)$ and the element $w\in\mathfrak{S}_N$, is characterized by the following properties:*

(i) $E_w^\lambda(x)=x_w^\lambda+\sum_{(\mu,w')<(\lambda,w)} u_{ww'}^{\lambda\mu}x_{w'}^\mu$,

(ii) $E_w^\lambda(x)$ *is joint eigenfunction for the operators* $\hat{D}_j^A$.

We note that our definition of the nonsymmetric Jack polynomials is slightly different from the one in the references cited above.

The symmetric Jack polynomials are obtained by symmetrizing E_w^λ, i.e.,

$$J_\lambda(x)=\frac{1}{\#\mathfrak{S}_N^\lambda}\prod_{v\in\mathfrak{S}_N} v(E_w^\lambda),$$

where $\mathfrak{S}_N^\lambda$ is a subgroup of $\mathfrak{S}_N$ that preserves λ. The Jack polynomials $J_\lambda(x)$ are pairwise orthogonal with respect to the following scalar product:

$$\langle f(x),g(x)\rangle_J^{(\beta)}=(-1)^{\beta N(N-1)/2}[f\overline{g}(\phi_s^{(\beta)})^2]_0, \tag{5}$$

where $[\cdot]_0$ stands for the constant term and $\overline{g}=g(x^{-1})$. We note that the operators $\hat{D}_j^A$ are self-adjoint with respect to this scalar product.

Next we define an analog of the creation and annihilation operators associated with the Hamiltonian (1):

$$\tilde{a}_j^\dagger=\frac{1}{\sqrt{2}}\,\tilde{\phi}_A^{-1}\circ(-D_j^A+x_j)\circ\tilde{\phi}_A,\quad \tilde{a}_j=\frac{1}{\sqrt{2}}\,\tilde{\phi}_A^{-1}\circ(D_j^A+x_j)\circ\tilde{\phi}_A,$$

with $\tilde{\phi}_A=\Pi_{k=1}^N\exp(-x_k^2/2)$. In Ref. 5, we have introduced a set of commuting operators:

$$\tilde{h}_j^A=\tilde{a}_j^\dagger\tilde{a}_j+\beta\sum_{k(<j)} s_{jk}.$$

These operators give another representation of $\tilde{\mathfrak{H}}'$ on $\mathbb{C}[x]$:

$$\rho^A(x_j)=\tilde{a}_j^\dagger,\quad \rho^A(\hat{D}_j^A)=\tilde{h}_j^A,\quad \rho^A(s_{ij})=s_{ij}.$$

The representation ρ^A is related to the A_{N-1}-type Calogero model. Indeed, the Hamiltonian (1) can be represented by $\tilde{h}_j^A$:

$$\mathscr{H}_A=\phi_A^{(\beta)}\circ\mathrm{Res}\left(\sum_{j=1}^N\tilde{h}_j^A\right)\circ(\phi_A^{(\beta)})^{-1}+\frac{N}{2},$$

with $\phi_A^{(\beta)}=\Pi_{j<k}|x_j-x_k|^\beta \Pi_{j=1}^N \exp(-x_j^2/2)$ a ground state wave function. The notation $\mathrm{Res}(X)$ means that action of X is restricted to symmetric functions of the variables $x_1,\ldots,x_N$. The scalar product corresponds to this case is the following:

$$\langle f,g\rangle_H^{(\beta)}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x)g(x)(\phi_A^{(\beta)})^2\, dx_1\cdots dx_N. \tag{6}$$

By a direct calculation, we see that the operator $\tilde{a}_j^\dagger$ is adjoint of $\tilde{a}_j$ with respect to the scalar product (6) for all $j=1,...,N$. Multivariable Hermite polynomials are defined by using this scalar product.[1,4] In fact, the definition in Ref. 1 and that in Ref. 4 are slightly different. Here we shall follow Ref. 4.

Definition 2.2:[4] *Multivariable Hermite polynomials* $H_\lambda^{(\beta)}(x)$ *are characterized by the following properties:*

(i) $H_\lambda^{(\beta)}=m_\lambda(x)+\sum_{\mu(<_D\lambda)} u_{\lambda\mu}^A m_\mu(x)$,
(ii) $\langle H_\lambda^{(\beta)}(x),m_\mu(x)\rangle_H^{(\beta)}=0$ *for all* $\mu<_D\lambda$.

We now introduce the intertwining operator σ^A, which is a linear operator on $\mathbb{C}[x]$ such that

$$\sigma^A(f(x_1,\ldots,x_N))=f(\tilde{a}_1^\dagger,\ldots,\tilde{a}_N^\dagger)\cdot 1 \quad \text{for all} \quad f(x_1,\ldots,x_N)\in\mathbb{C}[x].$$

The intertwiner σ^A enjoys the following property.

Theorem 2.3: $\sigma^A(Qf(x))=\rho^A(Q)\sigma^A(f(x))$ *for all* $Q\in\mathfrak{H}_0'$, $f(x)\in\mathbb{C}[x]$.

Proof: Since both Q and $f(x)$ are elements of $\tilde{\mathfrak{H}}'$, it suffices to prove $\sigma^A(P\cdot 1)=\rho^A(P)\cdot 1$ for all $P\in\tilde{\mathfrak{H}}'$. We then note that every element P of $\tilde{\mathfrak{H}}'$ can be represented in the following form:

$$P=\sum_{\mathbf{n}}\sum_{w\in\mathfrak{S}_N} p_{\mathbf{n},w}(x)(\hat{D}^A)^{n_1}\cdots(\hat{D}_N^A)^{n_N}w,$$

where $p_{\mathbf{n},w}(x)$ are some polynomials. Since $\hat{D}_j^A\cdot 1=\tilde{h}_j^A\cdot 1=\beta(j-1)$ for all j, we obtain the desired result. □

Using the intertwiner σ^A, we can construct an operator representation of $H_\lambda^{(\beta)}(x)$.

Proposition 2.4:[5,8,9] *Multivariable Hermite polynomials* $H_\lambda^{(\beta)}(x)$ *are related to the Jack polynomials as follows:*

$$H_\lambda^{(\beta)}(x)=2^{-|\lambda|/2}\sigma^A(J_\lambda^{(\beta)}(x))=2^{-|\lambda|/2}J_\lambda^{(\beta)}(\tilde{a}^\dagger)\cdot 1.$$

To prove this relation, it is enough to show $\sigma^A(J_\lambda^{(\beta)}(x))$ satisfies the second property of Definition 2.2. One can find the proofs in Refs. 5 and 9.

Now we turn to the B_N-type model (2). Dunkl operators associated with the B_N-type root system are defined as follows:[13,21]

$$D_j^B=\frac{\partial}{\partial z_j}+\beta\sum_{k(\neq j)}\left(\frac{1-s_{jk}}{z_j-z_k}+\frac{1-t_jt_ks_{jk}}{z_j+z_k}\right)+\gamma\,\frac{1-t_j}{z_j}, \tag{7}$$

where s_{jk} and t_j are elements of the B_N-type Weyl group. An element s_{ij} acts the same as in the A_{N-1}-case and t_j acts as a sign change, i.e., replaces the coordinate z_j by $-z_j$.

We then define Cherednik-type commuting operators associated with (7):

$$\hat{D}_j^B=z_jD_j^B+\beta\sum_{k(<j)}(s_{jk}+t_jt_ks_{ik}).$$

If we make a change of the variables $x_j=z_j^2/2$, $\mathrm{Res}^{(t)}(\hat{D}_j^B)$ is equivalent to $2\hat{D}_j^A$ where the notation $\mathrm{Res}^{(t)}(X)$ means the action of the operator X is restricted to the functions with the symmetry $t_jf(z)=f(z)$.[6] From this fact, we can define the representation of $\tilde{\mathfrak{H}}'$ on $\mathbb{C}[z^2]$:

$$x_j\mapsto\tfrac{1}{2}z_j^2,\quad \hat{D}_j^A\mapsto\tfrac{1}{2}\hat{D}_j^B,\quad s_{ij}\mapsto s_{ij}.$$

We now introduce creation and annihilation operators for the B_N case:

$$\tilde{b}_j^\dagger=\frac{1}{\sqrt{2}}\,\tilde{\phi}_B^{-1}\circ(-D_j^B+z_j)\circ\tilde{\phi}_B,\quad \tilde{b}_j=\frac{1}{\sqrt{2}}\,\tilde{\phi}_B^{-1}\circ(D_j^B+z_j)\circ\tilde{\phi}_B,$$

with $\tilde{\phi}_B=\Pi_{k=1}^N \exp(-z_k^2/2)$. Since the commutation relations of these operators are the same as those of z_j and D_j^B by construction, we can define commuting operators $\tilde{h}_j^B$ as follows:[6]

$$\tilde{h}_j^B=\kappa(\hat{D}_j^B)=\tilde{b}_j^\dagger\tilde{b}_j+\beta\sum_{k(<j)}(s_{jk}+t_jt_ks_{ik}).$$

These operators give another representation of $\tilde{\mathfrak{H}}'$ on $\mathbb{C}[z^2]$:

$$\rho^B(x_j)=\tfrac{1}{2}(\tilde{b}_j^\dagger)^2,\quad \rho^B(\hat{D}_j^A)=\tfrac{1}{2}\tilde{h}_j^B,\quad \rho^B(s_{ij})=s_{ij}.$$

The representation ρ^B is related to the B_N-type Calogero Hamiltonian (2) as follows:

$$\mathscr{H}_B=\phi_B^{(\beta)}\circ\mathrm{Res}\left(\sum_{j=1}^N\tilde{h}_j^B\right)\circ(\phi_B^{(\beta)})^{-1}+\left(\frac{1}{2}+\gamma\right)N,$$

with $\phi_B^{(\beta)}=\Pi_{j<k}|z_j^2-z_k^2|^\beta\Pi_{j=1}^N|z_j|^\gamma\exp(-z_j^2/2)$ the ground state wave function of (2). The scalar product associated with this model is

$$\langle f(z),g(z)\rangle_L^{(\beta)}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(z)g(z)(\phi_B^{(\beta)})^2\,dz_1\cdots dz_N. \tag{8}$$

By a direct calculation, we can show that the operator $\tilde{b}_j^\dagger$ is adjoint of $\tilde{b}_j$ with respect to the scalar product (8), and hence the operator $\tilde{h}_j^B$ is self-adjoint for all $j=1,...,N$. Now we define multivariable Laguerre polynomials.[4]

Definition 2.5:[4] *Multivariable Laguerre polynomials $L_\lambda^{(\beta)}(z)$ are characterized by the following properties:*

(i) $L_\lambda^{(\beta)}(z)=m_\lambda(z^2)+\sum_{\mu(<_D\lambda)}u_{\lambda\mu}m_\mu(z^2)$,
(ii) $\langle L_\lambda^{(\beta)}(z),m_\mu(z^2)\rangle_L^{(\beta)}=0$ *for all* $\mu<_D\lambda$.

We define a linear map of $\mathbb{C}[x]$ to $\mathbb{C}[z^2]$ by using ρ^B:

$$\sigma^B(f(x_1,\ldots,x_N))=f((\tilde{b}_1^\dagger)^2/2,\ldots,(\tilde{b}_N^\dagger)^2/2)\cdot 1\quad\text{for all}\quad f(x_1,\ldots,x_N)\in\mathbb{C}[x].$$

As in the A_{N-1} case, the intertwiner σ^B enjoys the following property.

Theorem 2.6: $\sigma^B(Qf(x))=\rho^B(Q)\sigma^B(f(x))\cdot 1$ *for all* $Q\in\mathfrak{H}_0'$, $f(x_1,\ldots,x_N)\in\mathbb{C}[x]$. One can prove this statement in the same fashion as Theorem 2.3, so we omit the details.

We can construct an operator representation of $L_\lambda^{(\beta)}(z)$ by using the intertwiner σ^B.

Proposition 2.7:[6] *Multivariable Laguerre polynomials $L_\lambda^{(\beta)}(z)$ are related to the Jack polynomials as follows:*

$$L_\lambda^{(\beta)}(z)=\sigma^B(J_\lambda^{(\beta)}(x))=J_\lambda^{(\beta)}((\tilde{b}^\dagger)^2/2)\cdot 1.$$

III. CONSTRUCTION OF RAISING OPERATORS

As is shown in the previous section, the multivariable Hermite and Laguerre polynomials are expressed in terms of the Jack polynomials whose arguments are Dunkl-type operators. Some properties of the multivariable Hermite and Laguerre polynomials are obtained directly from those of the Jack polynomials simply by applying ρ^A or ρ^B. As an example, we will construct raising operators for the polynomials.

Lapointe and Vinet constructed raising operators for the Jack polynomials.[20] Using their raising operators, they obtained a Rodorigues-type formula for the Jack polynomials. After the work of Lapointe and Vinet, Ujino and Wadati constructed raising operators for the multivariable Hermite polynomials.[8] The raising operators above are constructed by the use of Dunkl operators of Heckman type which are noncommutative, so their relation to the degenerate double affine Hecke algebra was not clear.

On the other hand, Kirillov and Noumi gave another expression of raising operators by using Cherednik operators.[16] In our notation, their raising operators are expressed in the following form:

$$B_m^{\mathrm{J}}=\sum_{k_1<\cdots<k_m} x_{k_1}x_{k_2}\cdots x_{k_m}(\hat{D}_{k_1}^A+\beta(2-k_1))(\hat{D}_{k_2}^A+\beta(3-k_2))\cdots(\hat{D}_{k_m}^A+\beta(m-k_m+1)).$$

We recall an important property of these operators.

Theorem 3.1:[16] *Actions of the operators $B_m^{\mathrm{J}}\in\tilde{\mathfrak{H}}'$ on the Jack polynomials are given by*

$$B_m^{\mathrm{J}}J_\lambda^{(\beta)}(x)=\prod_{j=1}^{m}(\lambda_j+\beta(m-j+1))J_{\lambda+(1^m)}^{(\beta)}(x),$$

where $\lambda+(1^m)=(\lambda_1+1,\ldots,\lambda_N+1)$.

Applying σ^A or σ^B to B_m^{J}, we obtain raising operators for the Hermite case or the Laguerre case, respectively:

$$B_m^{\mathrm{H}}=\sum_{k_1<\cdots<k_m}\tilde{a}_{k_1}^\dagger\tilde{a}_{k_2}^\dagger\cdots\tilde{a}_{k_m}^\dagger(\tilde{h}_{k_1}^A+\beta(2-k_1))(\tilde{h}_{k_2}^A+\beta(3-k_2))\cdots(\tilde{h}_{k_m}^A+\beta(m-k_m+1)),$$

$$B_m^{\mathrm{L}}=\sum_{k_1<\cdots<k_m}\tilde{b}_{k_1}^\dagger\tilde{b}_{k_2}^\dagger\cdots\tilde{b}_{k_m}^\dagger(\tilde{h}_{k_1}^B+\beta(2-k_1))(\tilde{h}_{k_2}^B+\beta(3-k_2))\cdots(\tilde{h}_{k_m}^B+\beta(m-k_m+1)).$$

From Theorem 3.1 and Propositions 2.4 and 2.7, we immediately have the following:

Proposition 3.2: (i) $B_m^{\mathrm{H}}H_\lambda^{(\beta)}(x)=2^{-m/2}\Pi_{j=1}^m(\lambda_j+\beta(m-j+1))H_{\lambda+(1^m)}^{(\beta)}(x)$,
(ii) $B_m^{\mathrm{L}}L_\lambda^{(\beta)}(z)=\Pi_{j=1}^m(\lambda_j+\beta(m-j+1))L_{\lambda+(1^m)}^{(\beta)}(z)$.

Applying the raising operators repeatedly, one can obtain Rodorigues-type formulas for the multivariable Hermite and Laguerre polynomials:

$$H_\lambda^{(\beta)}(x)=2^{|\lambda|/2}\prod_{(i,j)\in\lambda}(\lambda_i-j+\beta(\lambda_j'-i+1))^{-1}(B_N^{\mathrm{H}})^{\lambda_N}(B_{N-1}^{\mathrm{H}})^{\lambda_{N-1}-\lambda_N}\cdots(B_1^{\mathrm{H}})^{\lambda_1-\lambda_2}\cdot 1,$$

$$L_\lambda^{(\beta)}(z)=\prod_{(i,j)\in\lambda}(\lambda_i-j+\beta(\lambda_j'-i+1))^{-1}(B_N^{\mathrm{L}})^{\lambda_N}(B_{N-1}^{\mathrm{L}})^{\lambda_{N-1}-\lambda_N}\cdots(B_1^{\mathrm{L}})^{\lambda_1-\lambda_2}\cdot 1,$$

where $\lambda'=(\lambda_1',\lambda_2',\ldots)$ is the conjugate partition to λ.

This expression of the raising operators for the multivariable Hermite polynomials is different from those of Ref. 8. The operators for the multivariable Laguerre polynomials have not been presented, as far as we know.

IV. CONSTRUCTION OF SHIFT OPERATORS

In this section, we construct shift operators for the multivariable Hermite and Laguerre polynomials. Each of such shift operators are related to one of the scalar products (5), (6), and (8) (see Theorems 4.2, 4.6, and 4.9 to follow). However, properties related to scalar product cannot be obtained simply by applying ρ^A or ρ^B. It requires a little more effort to construct shift operators.

A. Shift operators for the Jack polynomials

In this subsection, we review the method of constructing shift operators for the Jack polynomials by the use of the Cherednik operators. We note that the shift operators constructed below are different from that of Ref. 27, where Heckman-type operators are used. Our method is based on a

lecture note of Kirillov Jr.[15] All results given in this section can be obtained by taking limiting procedures on those of Ref. 15. However, all proofs given here are algebraic and we avoid using a limiting procedure so that we can apply the results to the Hermite and Laguerre cases.

Consider elements of $\mathfrak{H}_0'$ that have the following forms:

$$\mathscr{X}=\prod_{i<j}(x_i-x_j),\quad \mathscr{Y}_J=\prod_{i<j}(\beta-\hat{D}_i^A+\hat{D}_j^A),\quad \hat{\mathscr{Y}}_J=\prod_{i<j}(-\beta-\hat{D}_i^A+\hat{D}_j^A).$$

We note that these operators preserve $\mathbb{C}[x]$. From the defining relations of $\mathfrak{H}_0'$, we know that

$$\begin{aligned}(s_j+1)(-\beta-\hat{D}_j^A+\hat{D}_{j+1}^A)&=(-\beta-\hat{D}_{j+1}^A+\hat{D}_j^A)(s_j-1),\\ (s_j-1)(\beta-\hat{D}_j^A+\hat{D}_{j+1}^A)&=(\beta-\hat{D}_{j+1}^A+\hat{D}_j^A)(s_j+1),\\ s_j(c-\hat{D}_j^A+\hat{D}_k^A)(c-\hat{D}_{j+1}^A+\hat{D}_k^A)&=(c-\hat{D}_j^A+\hat{D}_k^A)(c-\hat{D}_{j+1}^A+\hat{D}_k^A)s_j,\end{aligned}\tag{9}$$

with c is an arbitrary constant and $k\neq j,\ j+1$. Then, if we define $\mathbb{C}[x]^{\mathfrak{S}_N}$ and $\mathbb{C}[x]^{-\mathfrak{S}_N}$ as

$$\mathbb{C}[x]^{\mathfrak{S}_N}=\{f(x)\in\mathbb{C}[x]\,|\,(s_j-1)f(x)=0\},$$

$$\mathbb{C}[x]^{-\mathfrak{S}_N}=\{f(x)\in\mathbb{C}[x]\,|\,(s_j+1)f(x)=0\},$$

we see that

$$\mathscr{X}\in\mathbb{C}[x]^{-\mathfrak{S}_N},\quad \mathscr{Y}_J(\mathbb{C}[x]^{\mathfrak{S}_N})=\mathbb{C}[x]^{-\mathfrak{S}_N},\quad \hat{\mathscr{Y}}_J(\mathbb{C}[x]^{-\mathfrak{S}_N})=\mathbb{C}[x]^{\mathfrak{S}_N}.\tag{10}$$

We now introduce shift operators for the Jack polynomials:

$$G_J=\mathscr{X}^{-1}\mathscr{Y}_J,\quad \hat{G}_J=\hat{\mathscr{Y}}_J\mathscr{X}.$$

The operators G_J and $\hat{G}_J$ enjoy the following properties:

Lemma 4.1.[15] (i) $G_J J_{\lambda+\delta}^{(\beta)},\ \hat{G}_J J_\lambda^{(\beta+1)}\in\mathbb{C}[x]^{\mathfrak{S}_N}$.

(ii) $G_J J_{\lambda+\delta}^{(\beta)}=c_\lambda^{(\beta+1)}m_\lambda+$ ''lower terms'' with respect to $<_D$, with $c_\lambda^{(\beta+1)}=\Pi_{i<j}\{\lambda_{N-j+1}-\lambda_{N-i+1}+j-i+\beta(j-i-1)\}$.

(iii) $\hat{G}_J J_\lambda^{(\beta+1)}=\tilde{c}_\lambda^{(\beta+1)}m_{\lambda+\delta}+$ ''lower terms'' with respect to $<_D$, with $\tilde{c}_\lambda^{(\beta+1)}=\Pi_{i<j}\{\lambda_{N-j+1}-\lambda_{N-i+1}+j-i+\beta(j-i+1)\}$.

Proof: (i) Follows from (10).

(ii) For the longest element w_0 of $\mathfrak{S}_N$, i.e., $w_0(j)=N-j+1$, Eq. (4) reduces to

$$\hat{D}_j^A x_{w_0}^\lambda=(\lambda_{N-j+1}+\beta(j-1))x_{w_0}^\lambda+\sum_{(\mu,w')<(\lambda,w_0)}u_{w_0w'}^{\lambda\mu}x_{w'}^\mu.$$

Using this relation, we can calculate the action of $\mathscr{Y}_J$:

$$\begin{aligned}\mathscr{Y}_J J_{\lambda+\delta}^{(\beta)}&=\prod_{i<j}(\beta-\hat{D}_i^A+\hat{D}_j^A)(x_{w_0}^{\lambda+\delta}+\text{''lower terms'' with respect to }<)\\ &=c_\lambda^{(\beta+1)}x_{w_0}^{\lambda+\delta}+\text{''lower terms'' with respect to }<.\end{aligned}\tag{11}$$

On the other hand, (10) implies that $\mathscr{Y}_J J_{\lambda+\delta}^{(\beta)}$ is divisible by $\mathscr{X}$. Together with (i), this concludes the proof.

(iii) Can also be proved in a similar way.

The following theorem implies that $\hat{G}_J$ is, in a sense, adjoint of G_J:

Theorem 4.2:[15] *For* $f,g\in\mathbb{C}[x]^{\mathfrak{S}_N}$, $\langle G_J f,g\rangle_J^{(\beta+1)}=\langle f,\hat{G}_J g\rangle_J^{(\beta)}$.

To prove this theorem, we introduce symmetrizer $\mathscr{P}_+$ and antisymmetrizer $\mathscr{P}_-$ as

$$\mathscr{P}_+=\frac{1}{\#\mathfrak{S}_N}\sum_{w\in\mathfrak{S}_N} w,\quad \mathscr{P}_-=\frac{1}{\#\mathfrak{S}_N}\sum_{w\in\mathfrak{S}_N}(-1)^{l(w)}w,$$

where $l(w)$ is the length of the element w. We further prepare a lemma.

Lemma 4.3:[15] $\mathscr{P}_-(\mathscr{Y}_J-\hat{\mathscr{Y}}_J)=\Sigma_j\hat{g}_j(\hat{D}_1^A,\ldots,\hat{D}_N^A)(s_j-1)$ *for some* $\hat{g}_j(x_1,\ldots,x_N)\in\mathbb{C}[x]$.

It should be remarked that this lemma is a degenerate version of a lemma given in Ref. 15. We will give a proof in the Appendix for the reader's convenience.

Now we go back to the proof of Theorem 4.2.

Proof of Theorem 4.2: It is clear that $\mathscr{P}_+$ does not affect the constant term of polynomials. Then, for $f,g\in\mathbb{C}[x]^{\mathfrak{S}_N}$, we know that

$$\begin{aligned}\langle G_J,f,g\rangle_J^{(\beta+1)}&=(-1)^{(\beta+1)N(N-1)/2}[(\mathscr{X}^{-1}\mathscr{Y}_Jf)\overline{g}(\phi_S^{(\beta+1)})^2]_0\\&=(-1)^{(\beta+1)N(N-1)/2}[\mathscr{P}_+((\mathscr{Y}_Jf)\overline{g}\mathscr{X}(\phi_S^{(\beta)})^2)]_0\\&=(-1)^{\beta N(N-1)/2}[\mathscr{P}_-(\mathscr{Y}_Jf)\overline{g}\overline{\mathscr{X}}(\phi_S^{(\beta)})^2]_0.\end{aligned}$$

From Lemma 4.3, we see that $\mathscr{P}_-(\mathscr{Y}_J-\hat{\mathscr{Y}}_J)f=0$ for all $f\in\mathbb{C}[x]^{\mathfrak{S}_N}$. Hence we can replace $\mathscr{Y}_J$ by $\hat{\mathscr{Y}}_J$:

$$\langle G_J,f,g\rangle_J^{(\beta+1)}=(-1)^{\beta N(N-1)/2}[\mathscr{P}_-(\hat{\mathscr{Y}}_Jf)\overline{g}\overline{\mathscr{X}}(\phi_S^{(\beta)})^2]_0=\langle\hat{\mathscr{Y}}_Jf,\mathscr{X}g\rangle_J^{(\beta)}=\langle f,\hat{G}_Jg\rangle_J^{(\beta)}.$$

In the last equality, we have used the self-adjointness of $\hat{D}_j^A$. □

Using Theorem 4.2, we can evaluate the action of G_J and $\hat{G}_J$ on the Jack polynomials.

Proposition 4.4:[15] $G_JJ_{\lambda+\delta}^{(\beta)}=c_\lambda^{(\beta+1)}J_\lambda^{(\beta+1)}$, $\hat{G}_JJ_\lambda^{(\beta+1)}=\tilde{c}_\lambda^{(\beta+1)}J_{\lambda+\delta}^{(\beta)}$, *where the constants* $c_\lambda^{(\beta)}$ *and* $\tilde{c}_\lambda^{(\beta)}$ *are defined in Lemma 4.1.*

Proof: Assume $\mu<_D\lambda$. Then we have

$$\langle G_JJ_{\lambda+\delta}^{(\beta)},m_\mu\rangle_J^{(\beta)}=\langle J_{\lambda+\delta}^{(\beta)},\hat{G}_Jm_\mu\rangle_J^{(\beta)}=\langle J_{\lambda+\delta}^{(\beta)},m_{\mu+\delta}+\text{lower terms}\rangle_J^{(\beta)}=0,$$

where we have used Lemma 4.1 and Theorem 4.2. From this fact, along with Lemma 4.1 (ii), we see that $G_JJ_{\lambda+\delta}^{(\beta)}$ coincides with $c_\lambda^{(\beta+1)}J_\lambda^{(\beta+1)}$. The latter can be proved in a similar way. □

With these preliminaries, it is possible to derive the following result.

Proposition 4.5:[25]

$$\langle J_\lambda^{(\beta)},J_\lambda^{(\beta)}\rangle_J^{(\beta)}=N!\prod_{k=1}^{\beta}\prod_{i<j}\frac{\lambda_i-\lambda_j-k+\beta(j-i+1)}{\lambda_i-\lambda_j+k+\beta(j-i-1)}.\tag{12}$$

Proof: From Proposition 4.4, it follows that

$$\begin{aligned}\langle J_\lambda^{(\beta+1)},J_\lambda^{(\beta+1)}\rangle_J^{(\beta+1)}&=\frac{1}{c_\lambda^{(\beta+1)}}\langle GJ_{\lambda+\delta}^{(\beta)},J_\lambda^{(\beta+1)}\rangle_J^{(\beta+1)}\\&=\frac{1}{c_\lambda^{(\beta+1)}}\langle J_{\lambda+\delta}^{(\beta)},\hat{G}_JJ_\lambda^{(\beta+1)}\rangle_J^{(\beta)}\\&=\frac{\tilde{c}_\lambda^{(\beta+1)}}{c_\lambda^{(\beta+1)}}\langle J_{\lambda+\delta}^{(\beta)},J_{\lambda=\delta}^{(\beta)}\rangle_J^{(\beta)}.\end{aligned}$$

Applying this relation repeatedly, we have

$$\langle J_\lambda^{(\beta)},J_\lambda^{(\beta)}\rangle_J^{(\beta)}=\prod_{k=0}^{\beta-1}\frac{\tilde{c}_{\lambda+k\delta}^{(\beta-k)}}{c_{\lambda+k\delta}^{(\beta-k)}}\langle J_{\lambda+\beta\delta}^{(\beta=0)},J_{\lambda+\beta\delta}^{(\beta=0)}\rangle_S^{(\beta=0)},$$

which gives the desired result.

The norm formula (12) can be rewritten into the following form:

$$\langle J_\lambda^{(\beta)}, J_\lambda^{(\beta)} \rangle_{\mathrm{J}}^{(\beta)} = \frac{(N\beta)!}{(\beta!)^N} \prod_{(i,j)\in\lambda} \frac{j-1+\beta(N-i+1)}{j+\beta(N-i)} \cdot \frac{\lambda_i - j + 1 + \beta(\lambda'_j - i)}{\lambda_i - j + \beta(\lambda'_j - i + 1)}. \tag{13}$$

A proof of the equivalence between (12) and (13) is given by induction on $|\lambda|$.

B. Shift operators for the multivariable Hermite polynomials

In this section, we will construct shift operators for the multivariable Hermite polynomials. It should be noted that Heckman has constructed shift operators for the Hamiltonian H_A without harmonic potential.[28] However, for the application to norm formulas, it is necessary to compute actions of the shift operators on polynomials explicitly. Our method gives a unified and straightforward way to compute such actions.

To construct shift operators, we first introduce $\mathscr{Y}_{\mathrm{H}}$ and $\hat{\mathscr{Y}}_{\mathrm{H}}$ as follows:

$$\mathscr{Y}_{\mathrm{H}} = \rho^A(\mathscr{Y}_{\mathrm{J}}) = \prod_{i<j} (\beta - \tilde{h}_i^A + \tilde{h}_j^A), \quad \hat{\mathscr{Y}}_{\mathrm{H}} = \rho^A(\hat{\mathscr{Y}}_{\mathrm{J}}) = \prod_{i<j} (-\beta - \tilde{h}_i^A + \tilde{h}_j^A).$$

Using these operators, we define shift operators for the multivariable Hermite polynomials:

$$G_{\mathrm{H}} = \mathscr{X}^{-1} \mathscr{Y}_{\mathrm{H}}, \quad \hat{G}_{\mathrm{H}} = \hat{\mathscr{Y}}_{\mathrm{H}} \mathscr{X}.$$

We stress that we have used same $\mathscr{X}$ as (9), and therefore $G_{\mathrm{H}} \neq \rho^A(G_{\mathrm{J}})$ and $\hat{G}_{\mathrm{H}} \neq \rho^A(\hat{G}_{\mathrm{J}})$. This reflects the characteristics of the scalar products (5) and (6).

If we apply ρ^A to (9), we have

$$(s_j + 1)(-\beta - \tilde{h}_j^A + \tilde{h}_{j+1}^A) = (-\beta - \tilde{h}_{j+1}^A + \tilde{h}_j^A)(s_j - 1),$$

$$(s_j - 1)(\beta - \tilde{h}_j^A + \tilde{h}_{j+1}^A) = (\beta - \tilde{h}_{j+1}^A + \tilde{h}_j^A)(s_j + 1),$$

$$s_j (c - \tilde{h}_j^A + \tilde{h}_k^A)(c - \tilde{h}_{j+1}^A + \tilde{h}_k^A) = (c - \tilde{h}_j^A + \tilde{h}_k^A)(c - \tilde{h}_{j+1}^A + \tilde{h}_k^A) s_j,$$

with c an arbitrary constant and $k \neq j,\ j+1$. These relations imply that

$$\mathscr{Y}_{\mathrm{H}}(\mathbb{C}[x]^{\mathfrak{S}_N}) = \mathbb{C}[x]^{-\mathfrak{S}_N}, \quad \hat{\mathscr{Y}}_{\mathrm{H}}(\mathbb{C}[x]^{-\mathfrak{S}_N}) = \mathbb{C}[x]^{\mathfrak{S}_N}. \tag{14}$$

Furthermore, if we apply σ^A to (11), we see that

$$\mathscr{Y}_{\mathrm{H}} H_{\lambda+\delta}^{(\beta)} = c_\lambda^{(\beta+1)} x_{w_0}^{\lambda+\delta} + \text{“lower terms” with respect to } \prec. \tag{15}$$

For the proof of the shift relations for the Jack polynomials (Proposition 4.4), Theorem 4.2 played a crucial role. Here we state an analogous result for the Hermite case:

Theorem 4.6: For $f,\ g \in \mathbb{C}[x]^{\mathfrak{S}_N}$, $\langle G_{\mathrm{H}} f, g \rangle_{\mathrm{H}}^{(\beta+1)} = \langle f, \hat{G}_{\mathrm{H}} g \rangle_{\mathrm{H}}^{(\beta)}$.

Proof: The proof is similar to that of Theorem 4.2. For $f,\ g \in \mathbb{C}[x]^{\mathfrak{S}_N}$, we know that

$$\langle G_{\mathrm{H}} f, g \rangle_{\mathrm{H}}^{(\beta+1)} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathscr{P}_-(\mathscr{Y}_{\mathrm{H}} f)\, g\, \mathscr{X}(\phi_A^{(\beta)})^2\, dx_1 \cdots dx_N.$$

On the other hand, applying ρ^A to Lemma 4.3, we obtain

$$\mathscr{P}_-(\mathscr{Y}_{\mathrm{H}} - \hat{\mathscr{Y}}_{\mathrm{H}}) = \sum_j \hat{g}_j(\tilde{a}_1^\dagger, \ldots, \tilde{a}_N^\dagger)(s_j - 1) \quad \text{for some } \hat{g}_j(x_1, \ldots, x_N) \in \mathbb{C}[x].$$

From this relation, we find that $\mathscr{P}_-(\mathscr{Y}_{\mathrm{H}} - \hat{\mathscr{Y}}_{\mathrm{H}}) f = 0$ for all $f \in \mathbb{C}[x]^{\mathfrak{S}_N}$. Hence we can replace $\mathscr{Y}_{\mathrm{H}}$ by $\hat{\mathscr{Y}}_{\mathrm{H}}$:

$$\langle G_{\mathrm{H}} f, g \rangle_{\mathrm{H}}^{(\beta+1)} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathscr{P}_-(\hat{\mathscr{Y}}_{\mathrm{H}} f)\, g\, \mathscr{X}(\phi_A^{(\beta)})^2\, dx_1 \cdots dx_N = \langle \hat{\mathscr{Y}}_{\mathrm{H}} f, \mathscr{X} g \rangle_{\mathrm{H}}^{(\beta)} = \langle f, \hat{G}_{\mathrm{H}} g \rangle_{\mathrm{H}}^{(\beta)}.$$

In the last equality, we have used the self-adjointness of the operator $\tilde{h}_j^A$. □

Now we are in position to state the following.

Proposition 4.7: $G_{\mathrm{H}}H_{\lambda+\delta}^{(\beta)}=c_\lambda^{(\beta+1)}H_\lambda^{(\beta+1)}$, $\hat{G}_{\mathrm{H}}H_\lambda^{(\beta+1)}=\tilde{c}_\lambda^{(\beta+1)}H_{\lambda+\delta}^{(\beta)}$, *where the constants* $c_\lambda^{(\beta)}$ *and* $\tilde{c}_\lambda^{(\beta)}$ *are defined in Lemma 4.1.*

Proof: From (14) and (15), we know that $(c_\lambda^{(\beta+1)})^{-1}G_{\mathrm{H}}H_{\lambda+\delta}^{(\beta)}$ satisfies the first condition of Definition 2.2. So it is sufficient to prove the orthogonality which can be shown in the same way as Proposition 4.4. The second equation can be proved in similar way. □

Using Proposition 4.7 and Theorem 4.6, we can prove the norm formula for $H_\lambda^{(\beta)}$.

Proposition 4.8:[1,4]

$$\langle H_\lambda^{(\beta)},H_\lambda^{(\beta)}\rangle_{\mathrm{H}}^{(\beta)}=\frac{\pi^{N/2}N!}{2^{|\lambda|+\beta N(N-1)/2}}\prod_{j=1}^{N}(\lambda_j+\beta(N-j))!\prod_{k=1}^{\beta}\prod_{i<j}\frac{\lambda_i-\lambda_j-k+\beta(j-i+1)}{\lambda_i-\lambda_j+k+\beta(j-i-1)},\tag{16}$$

where $|\lambda|=\Sigma_j\lambda_j$.

Proof: Using Proposition 4.7 and Theorem 4.6, we see that

$$\langle H_\lambda^{(\beta+1)},H_\lambda^{(\beta+1)}\rangle_{\mathrm{H}}^{(\beta+1)}=\frac{\tilde{c}_\lambda^{(\beta+1)}}{c_\lambda^{(\beta+1)}}\langle H_{\lambda+\delta}^{(\beta)},H_{\lambda+\delta}^{(\beta)}\rangle_{\mathrm{H}}^{(\beta)}.$$

On the other hand, since $H_\lambda^{(\beta=0)}(x)$ is direct product of the (one-variable) Hermite polynomials, one can evaluate the norm easily:

$$\langle H_\lambda^{(\beta=0)},H_\lambda^{(\beta=0)}\rangle_{\mathrm{H}}^{(\beta=0)}=\frac{\pi^{N/2}\cdot\#\mathfrak{G}_N^\lambda}{2^{|\lambda|}}\prod_{j=1}^{N}\lambda_j!.$$

Using these relations, one arrives at the formula above. □

The norm formula (16) can be rewritten into the following form:[1]

$$\langle H_\lambda^{(\beta)},H_\lambda^{(\beta)}\rangle_{\mathrm{H}}^{(\beta)}=\frac{\pi^{N/2}}{2^{|\lambda|+\beta N(N-1)/2}}\cdot\frac{\Pi_{j=1}^{N}(j\beta)!}{(\beta!)^N}\times\prod_{(i,j)\in\lambda}\frac{\{j-1+\beta(N-i+1)\}\{\lambda_i-j+1+\beta(\lambda_j'-i)\}}{\lambda_i-j+\beta(\lambda_j'-i+1)}.$$

It should be remarked that other proofs of these formulas have been given via limiting procedure.[1,4]

C. Shift operators for the multivariable Laguerre polynomials

We first define $\mathscr{X}_{\mathrm{L}}$, $\mathscr{Y}_{\mathrm{L}}$, and $\hat{\mathscr{Y}}_{\mathrm{L}}$ as follows:

$$\mathscr{X}_{\mathrm{L}}=\prod_{i<j}(z_i^2-z_j^2),\quad \mathscr{Y}_{\mathrm{L}}=\rho^B(\mathscr{Y}_{\mathrm{J}})=\prod_{i<j}(\beta-\tilde{h}_i^B/2+\tilde{h}_j^B/2),$$

$$\hat{\mathscr{Y}}_{\mathrm{L}}=\rho^B(\hat{\mathscr{Y}}_{\mathrm{J}})=\prod_{i<j}(-\beta-\tilde{h}_i^B/2+\tilde{h}_j^B/2).$$

After same discussion as in the previous subsection, we see that

$$\mathscr{Y}_{\mathrm{L}}(\mathbb{C}[z^2]^{\mathfrak{G}_N})=\mathbb{C}[z^2]^{-\mathfrak{G}_N},\quad \hat{\mathscr{Y}}_{\mathrm{L}}(\mathbb{C}[z^2]^{-\mathfrak{G}_N})=\mathbb{C}[z^2]^{\mathfrak{G}_N},\tag{17}$$

and

$$\mathscr{Y}_{\mathrm{H}}L_{\lambda+\delta}^{(\beta)}=c_\lambda^{(\beta+1)}z_{w_0}^{2(\lambda+\delta)}+\text{“lower terms” with respect to }\prec.\tag{18}$$

Now we define the shift operators for the Laguerre case:

$$G_{\mathrm{L}}=\mathscr{X}_{\mathrm{L}}^{-1}\mathscr{Y}_{\mathrm{L}},\quad \hat{G}_{\mathrm{L}}=\hat{\mathscr{Y}}_{\mathrm{L}}\mathscr{X}_{\mathrm{L}}.$$

These operators enjoy the following properties:

Theorem 4.9: *For* $f,g\in\mathbb{C}[z^2]^{\mathfrak{G}_N}$, $\langle G_{\mathrm{L}}f,g\rangle_{\mathrm{L}}^{(\beta+1)}=\langle f,\hat{G}_{\mathrm{L}}g\rangle_{\mathrm{L}}^{(\beta)}$.

Proof: The proof for this case is also similar to that of Theorem 4.2. For $f,g\in\mathbb{C}[z^2]^{\mathfrak{G}_N}$, we know that

$$\langle G_{\mathrm{L}},f,g\rangle_{\mathrm{L}}^{(\beta+1)}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathscr{P}_{-}(\mathscr{Y}_{\mathrm{L}}f\ )g.\mathscr{X}_{\mathrm{L}}(\phi_B^{(\beta)})^2\ dz_1\cdots dz_N.$$

On the other hand, applying ρ^B to Lemma 4.3, we find that

$$\mathscr{P}_{-}(\mathscr{Y}_{\mathrm{L}}-\hat{\mathscr{Y}}_{\mathrm{L}})=\sum_j\ \hat{g}_j(\tilde{b}_1^{\dagger},\ldots,\tilde{b}_N^{\dagger})(s_j-1)\quad\text{for some}\ \ \hat{g}_j(x_1,\ldots,x_N)\in\mathbb{C}[x].$$

From this relation, we see that $\mathscr{P}_{-}(\mathscr{Y}_{\mathrm{L}}-\hat{\mathscr{Y}}_{\mathrm{L}})f=0$ for all $f\in\mathbb{C}[z^2]^{\mathfrak{G}_N}$. Hence we can replace $\mathscr{Y}_{\mathrm{L}}$ by $\hat{\mathscr{Y}}_{\mathrm{L}}$:

$$\langle G_{\mathrm{H}}f,g\rangle_{\mathrm{L}}^{(\beta+1)}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathscr{P}_{-}(\hat{\mathscr{Y}}_{\mathrm{L}}f\ )g.\mathscr{X}_{\mathrm{L}}(\phi_A^{(\beta)})^2\ dz_1\cdots dz_N=\langle\hat{\mathscr{Y}}_{\mathrm{L}},f,.\mathscr{X}_{\mathrm{L}}g\rangle_{\mathrm{L}}^{(\beta)}=\langle f,\hat{G}_{\mathrm{L}}g\rangle_{\mathrm{L}}^{(\beta)}.$$

In the last equality, we have used the self-adjointness of the operator $\tilde{h}_j^B$. □

We then state the following results:

Proposition 4.10: $G_{\mathrm{L}}L_{\lambda+\delta}^{(\beta)}=c_{\lambda}^{(\beta+1)}L_{\lambda}^{(\beta+1)}$, $\hat{G}_{\mathrm{L}}L_{\lambda}^{(\beta+1)}=\tilde{c}_{\lambda}^{(\beta+1)}L_{\lambda+\delta}^{(\beta)}$, *where the constants* $c_{\lambda}^{(\beta)}$ *and* $\tilde{c}_{\lambda}^{(\beta)}$ *are defined in Lemma 4.1.*

Proof: From (17) and (18), we know that $(c_{\lambda}^{(\beta+1)})^{-1}G_{\mathrm{L}}L_{\lambda+\delta}^{(\beta)}$ satisfies the first condition of Definition 2.5 up to a constant factor. So it is sufficient to prove the orthogonality which can be shown in the same way as Proposition 4.4. The second equation can be proven in a similar way.□

Using Proposition 4.10 and Theorem 4.9, we can prove the norm formula for $L_{\lambda}^{(\beta)}$.

Proposition 4.11:[1,4]

$$\langle L_{\lambda}^{(\beta)},L_{\lambda}^{(\beta)}\rangle_{\mathrm{L}}^{(\beta)}=N!\prod_{j=1}^{N}\ (\lambda_j+\beta(N-j))!\prod_{j=1}^{N}\ \Gamma(\lambda_j+\beta(N-j)+\gamma+1/2)$$
$$\times\prod_{k=1}^{\beta}\prod_{i<j}\ \frac{\lambda_i-\lambda_j-k+\beta(j-i+1)}{\lambda_i-\lambda_j+k+\beta(j-i-1)},\tag{19}$$

where $\Gamma(\cdot)$ *denotes the gamma function.*

Proof: The proof of this proposition is similar to the Hermite case. We only note the following formula for the case $\beta=0$:

$$\langle L_{\lambda}^{(\beta=0)},L_{\lambda}^{(\beta=0)}\rangle_{\mathrm{H}}^{(\beta=0)}=(\#\mathfrak{G}_N^{\lambda})\prod_{j=1}^{N}\ \{\lambda_j!\cdot\Gamma(\lambda_j+\gamma+1/2)\},$$

which follows from the norm formula of the one-variable Laguerre polynomials. □

The norm formula (19) can be rewritten into the following form:[1]

$$\langle L_{\lambda}^{(\beta)},L_{\lambda}^{(\beta)}\rangle_{\mathrm{H}}^{(\beta)}=\frac{\Pi_{j=1}^{N}(j\beta)!}{(\beta!)^N}\prod_{j=1}^{N}\ \Gamma(\lambda_j+\beta(N-j)+\gamma+1/2)$$
$$\times\prod_{(i,j)\in\lambda}\ \frac{\{j-1+\beta(N-i+1)\}\{\lambda_i-j+1+\beta(\lambda_j'-i)\}}{\lambda_i-j+\beta(\lambda_j'-i+1)}.$$

It should be remarked that other proofs of these formulas have been given via limiting procedure.[1,4]

V. CONCLUDING REMARKS

In this paper, we have constructed the raising and shift operators for the multivariable Hermite and Laguerre polynomials by using the intertwining operators σ^A and σ^B.

We restrict ourselves to symmetric polynomials though the operators σ^A and σ^B are applicable to the nonsymmetric case, i.e., we can obtain the nonsymmetric counterparts of the multivariable Hermite and Laguerre polynomials:

$$E_w^{(\mathrm{H})\lambda}(x)=2^{-|\lambda|/2}E_w^{\lambda}(\tilde{a}^{\dagger})\cdot 1, \quad E_w^{(\mathrm{L})\lambda}(z)=E_w^{\lambda}((\tilde{b}^{\dagger})^2/2)\cdot 1.$$

Baker and Forrester named these polynomials nonsymmetric Hermite and Laguerre polynomials, respectively, and studied their properties.[2,3] We note that some of their results may be obtained directly from the corresponding properties of the Jack polynomials by applying the intertwiners.

Our constructs are based on the degenerate double affine Hecke algebra, so it is expected that the results given here extend to the nondegenerate case. As van Diejen[4] already proposed the difference counterpart of the Hamiltonians $\mathscr{H}_A$ and $\mathscr{H}_B$, it would be nice to clarify the algebraic structure of his models. We hope to report on them in the near future.

APPENDIX: PROOF OF LEMMA 4.3

In this Appendix, we will give a proof of Lemma 4.3. We remark again that the proof given in this section is a limiting case of Ref. 15.

We begin by examining some properties of the antisymmetrizer.

Lemma A1:[15] (i) *The antisymmetrizer $\mathscr{P}_-$ is divisible by $1+(-1)^{l(w_0)}w_0$ both on the left and on the right.*

(ii) *For all $j=1,\ldots,N-1$, the antisymmetrizer $\mathscr{P}_-$ is divisible by $1-s_j$ both on the left and on the right.*

Proof: (i) $\mathfrak{G}_N$ can be divided into pairs (w,ww_0). Then, rewriting into the summation over such pairs, we have

$$\mathscr{P}_-=\sum_{(w,ww_0)}\{(-1)^{l(w)}w+(-1)^{l(ww_0)}ww_0\}=\sum_{(w,ww_0)}(-1)^{l(w)}w\{1+(-1)^{l(w_0)}w_0\}.$$

Divisibility on the left is proved similarly.

(ii) Can also be proved by similar discussion. □

From Lemma A1 (ii), we know that $\mathrm{Ker}\,\mathscr{P}_-\supset\Sigma_j\,\mathrm{Ker}(1-s_j)$. To describe the kernel of the antisymmetrizer, we first investigate kernels of $1-s_j$ and their union.

Lemma A2:[15] (i) *Let V be a representation of $\mathfrak{G}_N$, and denote $V_j=\mathrm{Ker}(1-s_j)$, $V'=\Sigma_j V_j$. Then V' is $\mathfrak{G}_N$-invariant.*

(ii) *Assume V is a finite-dimensional irreducible representation of $\mathfrak{G}_N$. Then we have*

$$V'=\begin{cases}0 & \textit{if } V \textit{ is the sign representation,}\\ V & \textit{otherwise.}\end{cases}$$

Proof: (i) From the definition of V_j, it follows that $s_j(s_iv)=s_iv$ for all $v\in s_iV_j$. If we introduce $v_\pm=(v\pm s_iv)/2$, we see that $s_j(v_+-v_-)=v_+-v_-$, which means $v_+-v_-\in V_j$. Since $v_+\in V_i$ by definition, we obtain $v=v_++v_-\in V_i+V_j$. This leads to $s_iV_j\subset V_i+V_j$, which concludes the proof.

(ii) From (i), it follows that V' is a subrepresentation. Due to the irreducibility, V' can be either 0 or V. If $V'=0$, then we have $V_j=0$ for all j. This means that $1-s_j$ is invertible, i.e., for all $v\in V$, there exists u such that $v=(1-s_j)u$. Then we obtain $s_jv=-v$ for all $v\in V$, i.e., V is the sign representation. □

From Lemma A2(ii), it immediately follows that

$$\mathrm{Ker}\,\mathscr{P}_-=\sum_j\ \mathrm{Ker}(1-s_j)$$

for any finite-dimensional representation of $\mathfrak{G}_N$. Note this identity also holds for the representation of $\mathfrak{G}_N$ in the space of polynomials $\mathbb{C}[x]$, since this representation is a direct sum of finite-dimensional representations.

We now introduce operators $\hat{s}_j$ as

$$\hat{s}_j = s_j + \beta \frac{s_j - 1}{x_j - x_{j+1}}.$$

Using these operators, we can define another representation of the degenerate affine Hecke algebra $\mathfrak{h}_0'$ on $\mathbb{C}[x]$:

$$\rho'(\hat{D}_j^A) = x_j, \quad \rho'(s_j) = \hat{s}_j.$$

Using the isomorphism ρ', we introduce the deformed antisymmetrizer $\mathscr{P}_-^{(\beta)}$ as

$$\mathscr{P}_-^{(\beta)} = \rho'(\mathscr{P}_-) = \frac{1}{\#\mathfrak{G}_N} \sum_{w \in \mathfrak{G}_N} (-1)^{l(w)} \hat{w},$$

where $\hat{w} = \rho'(w)$.

Lemma A3:[15] Ker $\mathscr{P}_-^{(\beta)} =$ Ker $\mathscr{P}$ *for the action of* $\mathscr{P}_-^{(\beta)}$ *in* $\mathbb{C}[x]$.

Proof: By discussion similar to Proposition A1, we know that $\mathscr{P}_-^{(\beta)}$ is divisible by $\hat{s}_j - 1$ both on the left and on the right for every $j = 1,...,N-1$. Hence we have

$$\text{Ker } \mathscr{P}_-^{(\beta)} \supset \sum_j \text{Ker}(1 - \hat{s}_j) = \sum_j \text{Ker}(1 - s_j) = \text{Ker } \mathscr{P}, \tag{A1}$$

and thus $\dim(\text{Ker } \mathscr{P}_-^{(\beta)}) \geqslant \dim(\text{Ker } \mathscr{P}_-)$. On the other hand, if we denote $\mathbb{C}[x]_n$ as the space of polynomials of order n, it is clear that $\mathscr{P}_-^{(\beta)}$ preserves $\mathbb{C}[x]_n$. Since $\dim(\text{Ker } \mathscr{P}_-^{(\beta)})$ cannot decrease under specialization, it follows that $\dim(\text{Ker } \mathscr{P}_-^{(\beta)}) \leqslant \dim(\text{Ker } \mathscr{P}_-^{(\beta=0)}) = \dim(\text{Ker } \mathscr{P}_-)$, and hence we have

$$\dim(\text{Ker } \mathscr{P}_-^{(\beta)}) = \dim(\text{Ker } \mathscr{P}_-). \tag{A2}$$

Thus it follows from (A1) and (A2) that Ker $\mathscr{P}_-^{(\beta)} =$ Ker $\mathscr{P} = \Sigma_j$ Ker$(1 - s_j)$. □

We then define $\mathscr{Y}'$ and $\mathscr{Y}'$ as

$$\mathscr{Y}' = \rho'(\mathscr{Y}_J) = \prod_{i<j} (\beta - x_i + x_j), \quad \hat{\mathscr{Y}}' = \rho'(\hat{\mathscr{Y}}_J) = \prod_{i<j} (-\beta - x_i + x_j).$$

Lemma A4:[15]

$$\mathscr{P}_-(\mathscr{Y}' - \hat{\mathscr{Y}}')f = 0 \quad \text{for all } f \in \mathbb{C}[x]^{\mathfrak{G}_N}.$$

Proof: We can show that

$$(1 + (-1)^{N(N-1)/2} w_0)(\mathscr{Y}' - \hat{\mathscr{Y}}') = (\mathscr{Y}' - \hat{\mathscr{Y}}')(1 - w_0)$$

by direct calculation. Considering the action on $\mathbb{C}[x]^{\mathfrak{G}_N}$, we have

$$(1 + (-1)^{N(N-1)/2} w_0)(\mathscr{Y}' - \hat{\mathscr{Y}}')f = 0,$$

for all $f \in \mathbb{C}[x]^{\mathfrak{G}_N}$. Using this formula and Proposition A1 (i), we obtain the desired result. □

From Lemmas A3 and A4, we know that

$$\mathscr{P}_-^{(\beta)}(\mathscr{Y}' - \hat{\mathscr{Y}}')f = 0 \quad \text{for all } f \in \mathbb{C}[x]^{\mathfrak{G}_N}.$$

On the other hand, the following statement can easily be proven:

Lemma A5: Let $\hat{A}$ *be an operator of the form* $\hat{A}=\Sigma_{w\in\mathfrak{G}_N} g_w \hat{w}$ *with* $g_w\in\mathbb{C}[x]$. *If* $\hat{A}f=0$ *for all* $f\in\mathbb{C}[x]^{\mathfrak{G}_N}$, *then* $\hat{A}$ *can be represented in the following form:*

$$\hat{A}=\sum_j \hat{g}_j(\hat{s}_j-1) \quad \text{for some } \hat{g}_j\in\mathbb{C}[x]^{\mathfrak{G}_N}.$$

Proof: The operator $\hat{A}$ can be rewritten as

$$\hat{A}=\sum_{j_1,\ldots,j_k} \hat{g}'_{j_1,\ldots,j_k}(\hat{s}_{j_1}-1)\cdots(\hat{s}_{j_1}-1)+\hat{g}'_0 \quad \text{for some } \hat{g}_{j_1,\ldots,j_k}\in\mathbb{C}[x].$$

Then the assumption of the proposition means $\hat{g}'_0=0$, which gives the desired result. □

Applying Lemma A5 to Lemma A4, we conclude that

$$\mathscr{S}_-^{(\beta)}(\mathscr{Y}'-\hat{\mathscr{Y}}')=\sum_j \hat{g}_j(x_1,\ldots,x_N)(\hat{s}_j-1) \quad \text{for some } \hat{g}_j(x_1,\ldots,x_N)\in\mathbb{C}[x]. \tag{A3}$$

Applying $(\rho')^{-1}$ completes the proof of Lemma 4.3.

[1] T. H. Baker and P. J. Forrester, Commun. Math. Phys. **188**, 175 (1997).
[2] T. H. Baker and P. J. Forrester, Nucl. Phys. B **492**, 682 (1997).
[3] T. H. Baker and P. J. Forrester, ''Non-symmetric Jack polynomials and integral kernels,'' preprint 1996 (solv-int/9612003).
[4] J. F. van Diejen, Commun. Math. Phys. **188**, 467 (1997).
[5] S. Kakei, J. Phys. A **29**, L619 (1996).
[6] S. Kakei, J. Phys. A **30**, L535 (1997).
[7] K. Sogo, J. Phys. Soc. Jpn. **65**, 3097 (1996).
[8] H. Ujino and M. Wadati, J. Phys. Soc. Jpn. **65**, 2423 (1996).
[9] H. Ujino and M. Wadati, J. Phys. Soc. Jpn. **66**, 345 (1997).
[10] F. Calogero, J. Math. Phys. **10**, 2197 (1969).
[11] F. Calogero, J. Math. Phys. **12**, 419 (1970).
[12] B. Sutherland, J. Math. Phys. **12**, 246 (1970).
[13] T. Yamamoto, Phys. Lett. A **208**, 293 (1995).
[14] I. Cherednik, Ann. Math. **141**, 191 (1995).
[15] A. A. Kirillov Jr., ''Lectures on the affine Hecke algebras and Macdonald conjectures'' (available from anonymous ftp-site at euler.math.ualberta.ca, in the directory /ftp/pub/hecke.tex).
[16] A. N. Kirillov and M. Noumi, ''Affine Hecke algebras and raising operators for Macdonald polynomials,'' preprint 1996 (q-alg/9605004).
[17] I. G. Macdonald, ''Affine Hecke algebras and orthogonal polynomials,'' Séminaire Bourbaki, 47ème année, 1994–95, No. 797.
[18] I. Cherednik, Invent. Math. **106**, 411 (1991).
[19] I. Cherednik, Commun. Math. Phys. **169**, 441 (1995).
[20] L. Lapointe and L. Vinet, Commun. Math. Phys. **178**, 425 (1996).
[21] C. F. Dunkl, Trans. Am. Math. Soc. **311**, 167 (1989).
[22] D. Bernard, M. Gaudin, F. D. M. Haldane, and V. Pasquier, J. Phys. A **26**, 5219 (1993).
[23] F. Knop and S. Sahi, Invent. Math. **128**, 9 (1997).
[24] E. M. Opdam, Acta Math. **175**, 75 (1995).
[25] I. G. Macdonald, *Symmetric functions and Hall polynomials*, 2nd Ed., Oxford Mathematical Monographs (Clarendon, 1995).
[26] J. E. Humphreys, *Reflection groups and Coxeter groups* (Cambridge U.P., Cambridge, 1990).
[27] G. L. Heckman, Invent. Math. **103**, 341 (1991).
[28] G. L. Heckman, in *Harmonic analysis on reductive groups*, edited by W. Barker and P. Sally, Progress in Mathematics 101 (Birkhäuser, Boston, 1991), pp. 181–191.

Lie-algebraic structure from inhomogeneous Hopf algebras

M. Lagraa[a] and N. Touhami[b]
Laboratoire de Physique Théorique, Université d'Oran Es-Sénia, 31100, Algérie

(Received 18 November 1997; accepted for publication 23 April 1998)

We construct the vector space dual to the space of right-invariant differential forms constructed from a first-order differential calculus on inhomogeneous quantum groups. We show that this vector space is equipped with a structure of a Hopf algebra which closes on a noncommutative Lie algebra satisfying a Jacobi identity.
© *1998 American Institute of Physics.* [S0022-2488(98)03107-7]

I. INTRODUCTION

It is well known that the underlying symmetry of the physical systems is governed by the Poincaré group. In fact, the most important observables in field theories such as the energy momentum and the angular momentum are related to the Lie algebra generators of the translations and the rotations. Then, it is especially interesting to study the noncommutative version of the Lie algebras of inhomogeneous Hopf algebras. In addition, due to the presence of the translations, one can hope to obtain new insight on the differential structure of the noncommutative space–time geometry on which it is possible to construct consistently quantum field theories or an Einstein–Cartan gravity.

Different approaches to construct Lie algebras of inhomogeneous quantum groups have been proposed.[1–6] Among them are those which are based on the q-deformed Lie algebra of the Lorentz group augmented consistently by translation generators as q-deformed four-vector derivatives in Minkowski space.[1,2] Other approaches are based on the contraction of q-deformed anti-de Sitter algebra U(O(3,2))[4,5] or on the projective method where a bicovariant differential calculus on inhomogeneous quantum groups such as $\mathrm{IGL}_{q,r}(N)$ or $\mathrm{ISO}_{q,r}(N)$ is obtained as a quotient of those of $\mathrm{GL}_{q,r}(N+1)$ or $\mathrm{SO}_{q,r}(N+1)$ with respect to a suitable Hopf ideal.[6]

In this paper, one presents a quite different approach based on a bicovariant differential calculus on an inhomogeneous Hopf algebra studied in our previous paper.[7] This approach allows us to construct the vector space dual to the right-invariant differential one-forms which is equipped with a Hopf algebra structure which closes on a quantum Lie algebra satisfying a quantum Jacobi identity. Throughout this paper we use the convention of those of Refs. 7 and 8.

II. DIFFERENTIAL CALCULUS ON INHOMOGENEOUS QUANTUM GROUPS

In this section, we start by recalling the definition and some basic properties of the inhomogeneous Hopf algebra $\mathscr{B}$ as well as the bicovariant first-order differential calculus on it.

The quantum groups $\mathscr{G}$ are abstract objects described by corresponding Hopf algebra $\mathscr{B}=\mathrm{poly}(\mathscr{G})$. An inhomogeneous quantum group, $\mathscr{G}$, is generated by a homogeneous quantum group $\mathscr{H}$ whose corresponding unital Hopf algebra is $\mathscr{A}=\mathrm{poly}(\mathscr{H})$ and translations described by elements p^n, $n=1,...,N$ corresponding to an irreducible representation $\Lambda^n{}_m$ of $\mathscr{H}$.[8]

We denote $\Delta:\mathscr{B}\to\mathscr{B}\otimes\mathscr{B}$, the comultiplication (algebra homomorphism), $\epsilon:\mathscr{B}\to C$ the counit (character), and $S:\mathscr{B}\to\mathscr{B}$ the coinverse (algebra antihomomorphism) which satisfy the coassociativity $(\Delta\otimes id)\Delta=(id\otimes\Delta)\Delta$, $(\epsilon\otimes id)\Delta=(id\otimes\epsilon)\Delta=id$, and $m\circ(S\otimes id)\Delta=m\circ(id\otimes S)\Delta=I\epsilon$, where $m:\mathscr{B}\otimes\mathscr{B}\to\mathscr{B}$ is the multiplication map $[m(a\otimes b)=ab]$ and $I_{\mathscr{A}}=I_{\mathscr{B}}=I$ is the unity of $\mathscr{B}$.

These maps act on the generators of $\mathscr{B}$ as:

[a]Electronic mail: lagraa.@elbahia.cerist.dz
[b]Electronic mail: touhami@elbahia.cerist.dz

 © 1998 American Institute of Physics

$$\Delta(\Lambda^n{}_m)=\Lambda^n{}_k\otimes\Lambda^k{}_m,\quad \Delta(p^n)=\Lambda^n{}_k\otimes p^k+p^n\otimes I,$$

$$\epsilon(\Lambda^n{}_m)=\delta^n_m,\quad \epsilon(p^n)=0,$$

$$S(\Lambda^n{}_k)\Lambda^k{}_m=\Lambda^n{}_m S(\Lambda^k{}_m)=\epsilon(\Lambda^k_m),$$

$$S(\Lambda^n{}_k)p^k+S(p^n)=\Lambda^n{}_n S(p^k)+p^n=\epsilon(p^n).$$

To construct a $\mathscr{B}$-bimodule of one-forms Γ, we consider an exterior bicovariant derivative $d:\mathscr{B}\to\Gamma(\Delta_L d=(id\otimes d)\Delta$ and $\Delta_R d=(d\otimes id)\Delta)$ (Ref. 7) from which we construct a right-invariant basis (Maurer–Cartan forms) of the bimodule Γ over $\mathscr{B}$,

$$\Theta^n_m=d\Lambda^n_k S(\Lambda^k_m),\quad \Pi^n=dp^n-\Theta^n_k p^k, \tag{1}$$

which transforms as

$$\Delta_R(\Theta^n_m)=\Theta^n_m\otimes I, \tag{2}$$

$$\Delta_R(\Pi^n)=\Pi^n\otimes I, \tag{3}$$

$$\Delta_L(\Theta^n_m)=\Lambda^n{}_{\ell}S(\Lambda^k{}_m)\otimes\Theta^{\ell}_k, \tag{4}$$

$$\Delta_L(\Pi^n)=\Lambda^n{}_k\otimes\Pi^k+\Lambda^n{}_k S(p^{\ell})\otimes\Theta^k_{\ell}. \tag{5}$$

Following Ref. 7, we can state that there exists linear functionals $f^n{}_m$, $f^{nk}{}_m$, f^n_{km}, and $f^{n\ell}_{mk}\in\mathscr{B}'$ (the dual vector space of $\mathscr{B}$) such that

$$\Pi^n a=(a\star f^n_k)\Pi^k+(a\star f^{n\ell}_k)\Theta^k_{\ell}, \tag{6}$$

$$\Theta^n_m a=(a\star f^n_{mk})\Pi^k+(a\star f^{n\ell}_{mk})\Theta^k_{\ell}, \tag{7}$$

and

$$b\Pi^n=\Pi^k(b\star f^n_k\circ S)+\Theta^k_{\ell}(b\star f^{n\ell}_k\circ S), \tag{8}$$

$$b\Theta^n_m=\Pi^k(b\star f^n_{mk}\circ S)+\Theta^k_{\ell}(b\star f^{n\ell}_{mk}\circ S), \tag{9}$$

where the convolution product of a functional $f\in\mathscr{B}'$ and an element a of $\mathscr{B}$ is defined as $(a\star f)=(f\otimes id)\Delta(a)$ or $(f\star a)=(id\otimes f)\Delta(a)$ and the convolution product of two functionals is defined as $(f_1\star f_2)(a)=(f_1\otimes f_2)\Delta(a)$.

From (6) and (7), we deduce

$$f^n{}_m(I)=\delta^n_m,\quad f^{nm}_{k\ell}(I)=\delta^n_{\ell}\delta^m_k,\quad f^{nm}_k(I)=0,\quad f^n_{km}(I)=0 \tag{10}$$

and

$$f^n{}_m(ab)=f^n{}_k(a)f^k{}_m(b)+f^{nk}_{\ell}(a)f^{\ell}_{km}(b), \tag{11}$$

$$f^{nk}_{m\ell}(ab)=f^{np}_{mq}(a)f^{qk}_{p\ell}(b)+f^n_{mq}(a)f^{qk}_{\ell}(b), \tag{12}$$

$$f^n_{m\ell}(ab)=f^{np}_{mq}(a)f^q_{p\ell}(b)+f^n_{mk}(a)f^k_{\ell}(b), \tag{13}$$

$$f^{nk}_{\ell}(ab)=f^n{}_q(a)f^{qk}_{\ell}(b)+f^{np}_q(a)f^{qk}_{p\ell}(b), \tag{14}$$

for any a and $b\in\mathscr{B}$. From (10) to (14), we can see that the convolution product of different functionals generates an algebra $\mathscr{B}^0\subset\mathscr{B}'$ which can be equipped with a Hopf algebra structure where the comultiplication is given by

$$\Delta'(f^n{}_m)=f^n_k\otimes f^k{}_m+f^{nk}_{\ell}\otimes f^{\ell}_{km},$$

$$\Delta'(f^{nk}_{m\ell})=f^{np}_{mq}\otimes f^{qk}_{p\ell}+f^n_{mq}\otimes f^{qk}_{\ell},$$

$$\Delta'(f^n_{m\ell})=f^{np}_{mq}\otimes f^q_{p\ell}+f^n_{mk}\otimes f^k_{\ell},$$

$$\Delta'(f^{nk}_{\ell})=f^n_q\otimes f^{qk}_{\ell}+f^{np}_q\otimes f^{qk}_{p\ell},$$

the counit is given by $\epsilon'(f^n{}_m)=\delta^n_m\epsilon$, $\epsilon'(f^{nk}_{m\ell})=\delta^n_{\ell}\delta^k_m\epsilon$, and $\epsilon'(f^n_{mk})=\epsilon'(f^{nk}_m)=0$, and the antipode is given by $S'(f)=f\circ S$. One can easily verify that these maps satisfy the Hopf algebra axioms

$$(\Delta'\otimes id)\Delta'=(id\otimes\Delta')\Delta',$$

$$(\epsilon'\otimes id)\Delta'=(id\otimes\epsilon')\Delta'=id,$$

$$m'\circ(S'\otimes id)\Delta'=m'\circ(id\otimes S')\Delta'=I_{\mathscr{B}^0}\epsilon',$$

where m' is the multiplication map defined as $m'\circ(f_1\otimes f_2)=f_1\star f_2$.

In the following one assumes

$$f^{nk}_{m\ell}=\tilde{f}^k{}_m\star f^n_{\ell},\quad f^{nk}_m=\tilde{\eta}^k\star f^n{}_m,\quad f^n_{km}=\tilde{\eta}_k\star f^n{}_m, \tag{15}$$

where $\tilde{\eta}^n$, $\tilde{\eta}_n$, and $\tilde{f}^n{}_m\in\mathscr{B}^0$. One deduces from (10) and (15) that

$$\tilde{\eta}^n(I)=0,\quad \tilde{\eta}_n(I)=0,\quad \tilde{f}^n{}_m(I)=\delta^n_m. \tag{16}$$

The fact that $\mathscr{A}$ is a Hopf subalgebra of $\mathscr{B}$, we have[7]

$$\tilde{\eta}_n(a)=0,\quad a\in\mathscr{A}. \tag{17}$$

From (6) to (7), and (17), one can show[7] that the commutative rules between the elements of $\mathscr{A}$ are given by

$$\Lambda^n{}_k(f^k{}_m\star a)=(a\star f^n{}_k)\Lambda^k{}_m, \tag{18}$$

$$\Lambda^n{}_k(\tilde{f}^n{}_k\circ S\star a)=(a\star\tilde{f}^n{}_k\circ S)\Lambda^k{}_m, \tag{19}$$

where the functionals $f^n{}_m$ and $\tilde{f}^n{}_m$ satisfy

$$f^n{}_m(ab)=f^n{}_k(a)f^k{}_m(b),\quad \tilde{f}^n{}_m(ab)=\tilde{f}^k{}_m(a)\tilde{f}^n{}_k(b) \tag{20}$$

for any a and $b\in\mathscr{A}$. The commutation rules between the elements of $\mathscr{A}$ and the translations are given by

$$p^n a=(a\star\tilde{f}^k{}_n\circ S)p^k+a\star\tilde{\eta}^n\circ S-\Lambda^n{}_k(\tilde{\eta}^k\circ S\star a), \tag{21}$$

and those between the translations are given by

$$p^n p^m=f^n{}_k(\Lambda^m_{\ell})p^{\ell}p^k-(f^n{}_k(\Lambda^m_{\ell})-\delta^n{}_{\ell}\delta^m{}_k)\tilde{\eta}^{\ell}(S(\Lambda^k{}_p))p^p+\tilde{\eta}^n(S(p^m))-\Lambda^n{}_k\Lambda^m_{\ell}\tilde{\eta}^k(S(p^{\ell})). \tag{22}$$

In addition, the generators of $\mathscr{A}$ satisfy the following relations:

$$S(\Lambda^k{}_m)\Lambda^a{}_c\tilde{\eta}_k(p^c)=\tilde{\eta}_m(p^a), \tag{23}$$

$$\Lambda^k{}_m\tilde{\eta}_k(S(p^a))=\tilde{\eta}_m(S(p^b))\Lambda^a{}_b. \tag{24}$$

Finally it is shown in Ref. 7 that this formalism is consistent if the different functionals satisfy the following conditions:

$$\tilde{R}=R+I\otimes Q=R^{-1}+R^{-1}(I\otimes Q), \tag{25}$$

$$Q=\lambda I \quad \lambda\in\mathscr{C}, \quad \lambda\neq -1, \tag{26}$$

$$T=-\tilde{R}T \quad \text{for } \lambda\neq 0 \tag{27}$$

$$\text{or } (\tilde{R}+I\otimes I)T-(\Lambda\perp\Lambda)(\tilde{R}+I\otimes I)T=0 \quad \text{for } \lambda=0, \tag{28}$$

$$(R\otimes I)(I\otimes R)(R\otimes I)=(I\otimes R)(R\otimes I)(I\otimes R), \tag{29}$$

$$(Z\otimes I)R+(\tilde{R}\otimes I)(I\otimes Z)R=(I\otimes R)(Z\otimes I)+(I\otimes R)(\tilde{R}\otimes I)(I\otimes Z), \tag{30}$$

$$(R\otimes I-I\otimes I\otimes I)\ ((I\otimes Z)Z-(Z\otimes I)Z)+T\otimes I-(I\otimes\tilde{R})(\tilde{R}\otimes I)(I\otimes T)=0, \tag{31}$$

and

$$(I\otimes R-I\otimes I\otimes I)((\tilde{Z}\otimes I)T-(I\otimes\tilde{Z})T)-(Z\otimes I)-(\tilde{R}\otimes I)(I\otimes Z)T=0 \tag{32}$$

with

$$\tilde{Z}=-RZ, \tag{33}$$

where $R^{n'}_{km}=f^{n}{}_{m}(\Lambda^{'}{}_{k})$, $\tilde{R}^{n'}_{km}=\tilde{f}^{n}{}_{m}(S(\Lambda^{'}{}_{k}))$, $Q^{n}_{k}=\tilde{\eta}_{k}(S(p^{n}))$, $Z^{nk}_{q}=\tilde{\eta}^{n}(S(\Lambda^{k}{}_{q}))$, $\tilde{Z}^{nk}_{m}=\tilde{f}^{n}{}_{m}(S(p^{k}))$ and $T^{nm}=\tilde{\eta}^{n}(S(p^{m}))$.

Remark (2,1): Equations (25) and (26) imply for $\lambda=0$, $R=\tilde{R}=R^{-1}$, and the R-matrix is subject to a strong constraint $R^{2}=I\otimes I$ used in Ref. 8. In this particular case, (18) and (19) reduce to the same commutation rule of the homogeneous part, which gives for $a=\Lambda^{a}{}_{b}$:

$$\Lambda^{n}{}_{k}\Lambda^{a}{}_{c}R^{kc}_{bm}=R^{na}_{ck}\Lambda^{c}{}_{b}\Lambda^{k}{}_{m},$$

and (21) implies

$$p^{n}\Lambda^{a}{}_{b}=R^{na}_{ck}\Lambda^{c}{}_{b}p^{k}+Z^{na}_{c}\Lambda^{c}{}_{b}-\Lambda^{c}{}_{b}\Lambda^{a}{}_{c}Z^{kc}_{b}.$$

Finally, using (28) and setting $T=-2T'$, (22) implies

$$(R^{nm}_{k'}-\delta^{n}_{k}\delta^{m}_{'})(p^{k}p^{'}-Z^{k'}_{q}p^{q}+T'^{k'}-\Lambda^{k}_{q}\Lambda^{'}_{p}T'^{qp})=0,$$

which are exactly the commutation rules between the elements of the inhomogeneous group $\mathscr{B}$ obtained in Ref. 8.

Remark (2,2): For $\lambda=0$, we also have from (31) the following relation:

$$(R\otimes I-I\otimes I\otimes I)((I\otimes Z)Z-(Z\otimes I)Z)+T\otimes I-(I\otimes R)(R\otimes I)(I\otimes T)=0,$$

which is the condition required by Podles in Ref. 9 for the existence of a four-dimensional covariant differential calculus on a quantum Minkowski space.

III. LIE-ALGEBRAIC STRUCTURE FROM INHOMOGENEOUS HOPF ALGEBRAS

In the following, we shall construct the vector space $\mathscr{T}$ dual to the right-invariant differential one-forms and the quantum commutators of different elements of $\mathscr{T}$. We start by

Proposition (3,1):

The one-form

$$X=\tilde{f}^{a}{}_{m}(\Lambda^{m}{}_{b})\Theta^{b}{}_{a} \tag{34}$$

is left- and right-invariant.

Proof: X is right-invariant. To show the left-invariance, we replace $a \in \mathscr{A}$ by $S^{-1}(\Lambda^a{}_b)$ into (19) then we apply S on both sides to have

$$\Lambda^c{}_b S(\Lambda^n{}_k)\tilde{f}^k{}_m(\Lambda^a{}_c) = \tilde{f}^n{}_k(\Lambda^c{}_b)S(\Lambda^k{}_m)\Lambda^a{}_c .$$

Setting the indices $a = m$, we get

$$\Lambda^c{}_b S(\Lambda^n{}_k)\tilde{f}^k{}_m(\Lambda^m{}_c) = \tilde{f}^n{}_m(\Lambda^m{}_b). \tag{35}$$

Using this relation and (4), we get

$$\Delta_L(X) = \Lambda^c{}_b S(\Lambda^n{}_k)\tilde{f}^k{}_m(\Lambda^m{}_c)\otimes \theta^b{}_n = \tilde{f}^n{}_m(\Lambda^m{}_b)\otimes \Theta^b{}_n = I\otimes X \tag{36}$$

which shows the left-invariance of X. Q.E.D.

Remark (3,1): Instead of (19), we can use (18) to have another left- and right-invariant one-form: $X = f^a{}_m(S^{-1}(\Lambda^m_b))\Theta^b{}_a$.

Remark (3,2): Equations (6)–(9) show that the classical limit of the functionals are, for any $a \in \mathscr{B}$, $f^k{}_m(a) = \delta^k_m \epsilon(a)$ and $f^n{}_k(a) = \delta^n_k \epsilon(a)$ implying $f^n{}_k(S^{-1}(\Lambda^m_b)) = \delta^n_b$ and $\tilde{f}^k{}_m(\Lambda^m{}_c) = \delta^k_c$ and, therefore (34) can be considered as the quantum trace which reduces to the usual trace in the classical limit. In fact (36) shows the invariance of the quantum trace under the left adjoint coaction action.[12] We can now express the action of the derivative d on an element of $a \in \mathscr{B}$ as $da = Xa - aX$ by following the method of differential calculus on the homogeneous quantum groups of Ref. 10. From the left- and the right-invariance of X, it is easy to check the left- and the right-covariance of d. In the following we set $V^a{}_b = \tilde{f}^a{}_k(\Lambda^k{}_b)$. Using the form of X and (9), we get

$$da = Xa - aX = V^a_b(\Theta^b{}_a(a*\epsilon) - \Theta^k_{\not{}}(a*f^{b\not{}}_{ak}\circ S) - \Pi^k(a*f^b_{ak}\circ S)), \tag{37}$$

which can be rewritten under the form

$$da = \Theta^b{}_a(a*L^a{}_b) + \Pi^a(a*L_a), \tag{38}$$

where the linear functionals on $\mathscr{B}$,

$$L^a{}_b = V^a_b\epsilon - V^c{}_d(\tilde{f}^a{}_c * f^d{}_b)\circ S \tag{39}$$

and

$$L_a = -V^c{}_d(\tilde{\eta}_c * f^d{}_a)\circ S, \tag{40}$$

are the generators of the Lie algebra of the inhomogeneous Hopf algebra $\mathscr{B}$. Let us note that by virtue of (17)

$$L_a(a) = 0, \quad a \in \mathscr{A}, \tag{41}$$

and from (37) to (38), we deduce $L^a{}_b(I) = 0$ and $L_a(I) = 0$.

To construct the exterior product of one-forms $\in \Gamma$, we consider, in addition to the right-invariant basis (1), the left-invariant basis $\tilde{\Theta}^a{}_b = S(\Lambda^a{}_c S(\Lambda^d{}_c))\Theta^c{}_d$ and $\tilde{\Pi}^a = S(\Lambda^a{}_b)\Pi^b + S(\Lambda^a{}_b S(p^c))\Theta^b{}_c$ which transforms as

$$\Delta_L(\tilde{\Theta}^a{}_b) = I\otimes\tilde{\Theta}^a{}_b, \quad \Delta_L(\tilde{\Pi}^a) = I\otimes\tilde{\Pi}^a, \tag{42}$$

$$\Delta_R(\tilde{\Theta}^a{}_b) = \tilde{\Theta}^c{}_d\otimes S(\Lambda^a{}_c S(\Lambda^d{}_d)), \tag{43}$$

$$\Delta_R(\tilde{\Pi}^a) = \tilde{\Pi}^b\otimes S(\Lambda^a{}_b) + \tilde{\Theta}^b{}_c\otimes S(\Lambda^a{}_b S(p^c)). \tag{44}$$

Now we consider the bimodule automorphism $\sigma : \Gamma^{\otimes 2} \to \Gamma^{\otimes 2}$ such that $\sigma(\tilde{\omega}\otimes_{\mathscr{B}}\omega) = \omega\otimes_{\mathscr{B}}\tilde{\omega}$ for any left-invariant element $\tilde{\omega}\in\Gamma$ and any right-invariant element $\omega\in\Gamma$, which gives for the one-forms $\Theta^a{}_b$,

$$\sigma(\tilde{\Theta}^e{}_f\otimes_{\mathscr{B}}\Theta^g{}_h)=S(\Lambda^e{}_a S(\Lambda^b{}_f))\sigma(\Theta^a{}_b\otimes_{\mathscr{B}}\Theta^g{}_h)=\Theta^g{}_h\otimes_{\mathscr{B}}S(\Lambda^e{}_d S(\Lambda^a{}_b))\Theta^a{}_b.$$

Using (7) and (17), we get

$$\sigma(\Theta^a{}_b\otimes_{\mathscr{B}}\Theta^c{}_d)=(\tilde{f}^g{}_d * f^c{}_h)(S(\Lambda^a{}_e S(\Lambda^f{}_b)))(\Theta^h{}_g\otimes_{\mathscr{B}}\Theta^e{}_f). \tag{45}$$

Similar considerations give

$$\sigma(\Theta^a{}_b\otimes_{\mathscr{B}}\Pi^c)=f^c{}_f(S(\Lambda^a{}_d S(\Lambda^e_b)))(\Pi^f\otimes_{\mathscr{B}}\Theta^d{}_e)+(\tilde{\eta}^g * f^c{}_f)(S(\Lambda^a{}_d S(\Lambda^e{}_b)))(\Theta^f{}_g\otimes_{\mathscr{B}}\Theta^d{}_e), \tag{46}$$

$$\begin{aligned}\sigma(\Pi^a\otimes_{\mathscr{B}}\theta^b{}_c)=&(\tilde{f}^e{}_c * f^b{}_d)(S(\Lambda^a{}_f))(\Theta^d{}_e\otimes_{\mathscr{B}}\Pi^f)+(\tilde{\eta}_c * f^b{}_d)(S(\Lambda^a{}_f S(p^g)))(\Pi^d\otimes_{\mathscr{B}}\Theta^f{}_g)\\&+(\tilde{f}^e{}_c * f^b{}_d)(S(\Lambda^a{}_f S(p^g)))(\Theta^d{}_e\otimes_{\mathscr{B}}\Theta^f{}_g)\end{aligned} \tag{47}$$

and

$$\begin{aligned}\sigma(\Pi^a\otimes_{\mathscr{B}}\Pi^b)=&f^b{}_c(S(\Lambda^a{}_d))(\Pi^e\otimes_{\mathscr{B}}\Pi^d)+f^b{}_c(S(\Lambda^a{}_d S(p^e)))(\Pi^c\otimes_{\mathscr{B}}\Theta^d{}_e)\\&+(\tilde{\eta}^d * f^b{}_c)(S(\Lambda^a{}_e))(\Theta^c{}_d\otimes_{\mathscr{B}}\Pi^e)+(\tilde{\eta}^d * f^b{}_c)(S(\Lambda^a{}_e S(p^f)))(\Theta^c{}_d\otimes_{\mathscr{B}}\Theta^e{}_f).\end{aligned} \tag{48}$$

In the following we write (38) and (45)–(48) under the compact form

$$da=\Pi^A(a * L_A), \tag{49}$$

$$\sigma(\Pi^A\otimes_{\mathscr{B}}\Pi^B)=\sigma^{AB}_{CD}(\Pi^C\otimes_{\mathscr{B}}\Pi^D), \tag{50}$$

where $\Pi^A=\Theta^a{}_b$ and $L_A=L^a_b$ when the indices A take the values ${}^a{}_b, {}^c{}_d, \ldots$, $\Pi^A=\Pi_a$ and $L_A=L_a$ when the indices A take the values $a,b,\ldots$ and $\sigma^{AB}{}_{CD}$ are given by the coefficients multiplying the tensor products of the one-forms Θ and Π in (45)–(48). Let us note that from (49), we have $d(a_{(1)})S(a_{(2)})=\Pi^A L_A(a)=\Pi(a)$ which tranforms as

$$\begin{aligned}\Delta_L(d(a_{(1)})S(a_{(2)}))=\Delta_L(\Pi(a))&=a_{(1)}S(a_{(4)})\otimes d(a_{(2)})S(a_{(3)})\\&=a_{(1)}S(a_{(3)})\otimes\Pi(a_{(2)})\\&=(id\otimes\Pi)Ad(a),\end{aligned} \tag{51}$$

where the left adjoint coaction action $Ad:\mathscr{B}\to\mathscr{B}\otimes\mathscr{B}$ is defined as $Ad(a)=a_{(1)}S(a_{(3)})\otimes a_{(2)}$, and satisfies

$$(\Delta\otimes id)Ad=(id\otimes Ad)Ad. \tag{52}$$

Here $a_{(1)}$, $a_{(2)}$, and $a_{(3)}$ are elements of $\mathscr{B}$ such that $a_{(1)}\otimes a_{(2)}\otimes a_{(3)}=(id\otimes\Delta)\Delta(a)$.

Theorem (3,1):

1. For any $a,b\in\mathscr{B}$

$$L^a{}_b(ab)=L^a{}_b(a)\epsilon(b)+(\tilde{f}^a{}_c * f^d_b)(S(a))L^c{}_d(b)+(\tilde{\eta}^a * f^c{}_b)(S(a))L_c(b), \tag{53}$$

$$L_a(ab)=L_a(a)\epsilon(b)+f^b{}_a(S(a))L_b(b)+(\tilde{\eta}_c * f^d{}_a)(S(a))L^c{}_d(b). \tag{54}$$

2. For any $a\in\mathscr{B}$, the generators $L^a{}_b$ and L_a close on the following quanium Lie algebra

$$\begin{aligned}[L^b{}_a,L^d{}_c](a)=&(V^b{}_a\delta^g_c\delta^d_h-V^k{}_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c S(\Lambda^d{}_h))))L^h{}_g(a))\\&-V^k_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c S(p^d)))L_g(a)),\end{aligned} \tag{55}$$

$$[L^b{}_a,L_c](a)=(V^b{}_a\delta^g_c-V^k{}_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c))L_g(a)), \tag{56}$$

$$[L_a, L^c{}_d](a) = V^k_n(\tilde{\eta}_k * f^n{}_a)(S(\Lambda^g{}_c S(p^d)))L_g(a)), \tag{57}$$

and

$$[L_a, L_b](a) = 0, \tag{58}$$

satisfying the Jacobi identity

$$[L_A,[L_B,L_C]](a) = [[L_A,L_B],L_C](a) + \sigma^{EF}{}_{AB}[L_E,[L_F,L_C]](a). \tag{59}$$

Proof: Using the Leibniz rule of the derivative d, (38) and (8) to (9), one obtains

$$\begin{aligned} d(ab) &= \Theta^b{}_a(ab * L^a{}_b) + \Pi^a(ab * L_a) \\ &= \Theta^b{}_a(a * L^a{}_b(b + \Pi^a(a * L_a)b + a(\Theta^d{}_c(b * L^c{}_d) + a\Pi^c)(b * L_c)) \\ &= \Theta^b{}_a((a * L^a{}_b)b + (a * (\tilde{f}^a{}_c * f^d{}_b) \circ S)(b * L^c{}_d)) + (a * \tilde{\eta}^a * f^c{}_b) \circ S)(b * L_c) + \Pi^a((a * L_a)b \\ &\quad + (a * f^b{}_a \circ S)(b * L_b) + (a * (\tilde{\eta}_c * f^d{}_a) \circ S)(b * L^c{}_d)) \end{aligned}$$

for any $a, b \in \mathscr{B}$. Identifying the coefficients multiplying $\Theta^b{}_a$ and Π_a, then acting ϵ, one obtains the twisted Leibniz rules for $L^a{}_b$ and L_a as (53) and (54).

Applying the exterior derivative d on both sides of (49), using $d^2 = 0$, and taking into account the grade of d, we get

$$d^2 a = 0 = d(\Pi^A)(a * L_A) - (\Pi^A \wedge \Pi^B)(a * (L_A * L_B), \tag{60}$$

for any $a \in \mathscr{B}$, leading to the Maurer–Cartan formula

$$d\Pi^A L_A(a) = d\Pi(a) = \Pi^A \wedge \Pi^B (L_A * L_B)(a) \tag{61}$$

$$= (\Pi^A \otimes \Pi^B)[L_A, L_B](a), \tag{62}$$

where the exterior product of two one-forms Π^A and Π^B is given by $\Pi^A \wedge \Pi^B = (\Pi^C \otimes \Pi^D) \times (\delta^A_C \delta^B_C - \sigma^{AB}{}_{CD})$ and the commutator $[L_A, L_B](a) = (\delta^C_A \delta^D_B - \sigma^{CD}_{AB})(L_C * L_D)(a)$ for any $a \in \mathscr{B}$.

On the other hand the extension of the derivation (37) to the forms gives

$$d\Pi^A L_A(a) = X \wedge \Pi^A L_A(a) + \Pi^A L_A(a) \wedge X = (I - \sigma)(X \otimes \Pi(a) + \Pi(a) \otimes X).$$

From the left-invariance of X and the property of σ, it follows

$$\begin{aligned} d\Pi(a) &= X \otimes \Pi(a) - \sigma(\Pi(a) \otimes X) \\ &= \Theta^a_b \otimes \Theta^c_d - \sigma(\Theta^c_d \otimes \Theta^a_b)) V^b_a L^d_c(a) + (\Theta^a_b \otimes \Pi^c + \sigma(\Pi^c \otimes \Theta^a_b))) V^b_a L_c(a). \end{aligned}$$

We can now use the different stages of the theorem (5,3) of Ref. 10 to get

$$d\Pi^A L_A(a) = (\Pi^A \otimes \Pi^B)[L_A, L_B](a) = (\Pi^A \otimes \Pi^B)(L_A \otimes L_B) Ad(a). \tag{63}$$

Comparing (61) with (63) we get

$$[L_A, L_B](a) = (L_A * L_B - \sigma^{CD}{}_{AB}(L_C * L_D))(a) = (L_A \otimes L_B) Ad(a)$$

for any $a \in \mathscr{B}$. Using (45)–(48), (63) can be written as

$$\begin{aligned} d\Pi^A L_A(a) &= (\Theta^a{}_b \otimes \Theta^c{}_d)((V^b{}_a \delta^g_c \delta^d_h - V^k{}_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c S(\Lambda^d{}_h)))L^h{}_g(a)) \\ &\quad - V^k{}_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c S(p^d)))L_g(a)) + (\Theta^a{}_b \otimes \Pi^c)(V^b{}_a \delta^g_c \\ &\quad - V^k_n(\tilde{f}^b{}_k * f^n{}_a)(S(\Lambda^g{}_c))L_g(a)) - (\Pi^a \otimes \Theta^c{}_d) V^k_n(\tilde{\eta}_k * f^n{}_a)(S(\Lambda^g{}_c S(p^d)))L_g(a). \end{aligned} \tag{64}$$

Developing now (64) and comparing the coefficients multiplying the different tensor products of Θ_b^a and Π^a with the latter relation, we get (55)–(58).

Finally, we can establish the quantum Jacobi identity (59) by using (52) and (64) as

$$\begin{aligned}[L_A,[L_B,L_C]](a)&=(L_A\otimes L_B\otimes L_C)(id\otimes Ad)Ad(a)\\&=(L_A\otimes L_B\otimes L_C)(\Delta\otimes id)Ad(a)\\&=[(L_A{*}L_B),L_C](a)\\&=[[L_A,L_B],L_C](a)+\sigma^{DE}AB[(L_D{*}L_E),L_C](a)\\&=[[L_A,L_B]](a)+\sigma^{DE}{}_{AB}[L_D,[L_E,L_C]](a).\end{aligned}$$

Q.E.D.

As a consequence of the relations (53) and (54) of this theorem, the universal enveloping algebra $U(\mathscr{T})$ can be equipped with a Hopf algebra structure whose comultiplication is given by

$$\Delta'(L^a{}_b)=L^a{}_b\otimes\epsilon+S'(f^{da}_{cb})\otimes L^c_d+S'(f^{ca}_b\otimes L_c,$$

$$\Delta'(L_a)=L_a\otimes\epsilon+S'(f^b{}_a)\otimes L_b+S'(f^d_{ca})\otimes L^c_d,$$

the antipode S' by

$$S'(L^a{}_b)=-S'^2(f^{da}_{cb}){*}L^c{}_d-S'^2(f^{ca}_b){*}L_c,$$

$$S'(L_a)=-S'^2(f^b{}_a){*}L_b-S'^2(f^d_{ca}){*}L^c{}_d,$$

and the counit by

$$\epsilon'(L^a{}_b)=\epsilon'(L_b)=0.$$

Using the Hopf algebra structure of $\mathscr{B}^0$, one can see that the axioms of the Hopf algebra are satisfied for $U(\mathscr{T})$.

Remark (3,3): By virtue of (41), the relations (55)–(58) reduce to the homogeneous part of (55) which is the Lie bracket verified by the generators of the Lie algebra of the homogeneous quantum groups.[11,12] For any $a\in\mathscr{B}$ of the form $\mathscr{A}p^n+\mathscr{A}$, the convolution product $(L_a{*}L_b)(a)=0$, due to (41), hence (58) is nontrivial only when it is applied on an element of $\mathscr{B}$ which contains products of two translations and more.

ACKNOWLEDGMENTS

We are grateful to M. Dubois-Violette for valuable discussions. One of the authors (M.L.) wishes to thank the members of the L.P.T.H.E. of Orsay University for their hospitality during November 1997 where part of the work on this paper was done.

[1] O. Ogiesvestsky, W. B. Schmidke, J. Wess, and B. Zumino, Commun. Math. Phys. **150**, 459 (1992).
[2] M. Schlicker, W. Weich, and R. Weixler, Lett. Math. Phys. **27**, 217 (1993).
[3] V. K. Dobrev, J. Phys. A **26**, 1317 (1993).
[4] J. Lukierski, A. Nowicki, H. Ruegg, and V. N. Tolstoy, Phys. Lett. B **264**, 331 (1991).
[5] J. Lukierski, and A. Nowicki Phys. Lett. B **279**, 299 (1992).
[6] L. Castellani, Lett. Math. Phys. **30**, 233 (1994); Commun. Math. Phys. **171**, 383 (1995); P. Aschieri and L. Castellani, Int. J. Mod. Phys. A **11**, 4513 (1996).
[7] M. Lagraa and N. Touhami, ''The noncommutative Hopf algebra,'' q-alg9705005.
[8] P. Podlés and S. L. Woronowicz, Commun. Math. Phys. **185**, 325 (1997).
[9] P. Podlés, Commun. Math. Phys. **181**, 569 (1996); J. Math. Phys. **38**, 4475 (1997).
[10] S. L. Woronowicz, Commun. Math. Phys. **122**, 125 (1989).
[11] U. Carow-Watamura, M. Schliecker, S. Watamura, and W. Weich, Commun. Math. Phys. **142**, 605 (1991).
[12] M. Lagraa, Int. J. Mod. Phys. A **11**, 699 (1996); A. B. Hammou and M. Lagraa, J. Math. Phys. **38**, 4462 (1997).

Iwasawa decomposition of affine Kac–Moody algebras using Satake diagrams

K. C. Pati and D. Parashar
Department of Physics and Astrophysics, University of Delhi, Delhi-110007, India

(Received 13 February 1997; accepted for publication 26 March 1998)

We carry out the Iwasawa decomposition of the affine Kac–Moody algebras within the general framework of involutive automorphisms which are determined with the help of the Satake diagrams corresponding to these algebras. The parabolic subalgebras essential for the construction of induced representations are computed from the Langlands decomposition. In particular, we consider the untwisted versions, viz., $A_1^{(1)}$ and $A_2^{(1)}$, as representative examples for the purpose of illustration. © *1998 American Institute of Physics.* [S0022-2488(98)01309-7]

I. INTRODUCTION

Recent times have witnessed tremendous activity in developing formal techniques whereby ordinary Lie algebras in general and Kac–Moody algebras in particular could account for a variety of problems with comparative ease and mathematical elegance. It is now well established that the affine Kac–Moody algebras[1,2] play a crucial role in the context of two-dimensional conformal field theories,[3] string theories,[4] and soluble statistical models,[5] among others. Inspired by this strong motivation, we address the study of the affine Kac–Moody algebras but restrict ourselves only to the case of untwisted versions, mainly for reasons of simplicity. The present work makes a preliminary attempt to obtain induced representations of these algebras. In doing so, we are led to devise the Satake diagrams for the affine Kac–Moody algebras in a manner analogous to that of the ordinary Lie algebras.[6] The Satake diagrams so constructed provide the necessary ingredients for the computation of involutive automorphisms of these algebras, a detailed pedagogical description appears elsewhere.[7,8] These involutive automorphisms in turn are used to facilitate the Iwasawa and Langlands decompositions[9] from which the parabolic subalgebras are readily computed and the corresponding representations obtained through Schmidt construction.[10,11]

The paper is organized as follows. In Sec. II, we briefly outline the essential steps for the construction of Satake diagrams from the Dynkin diagrams of the affine Kac–Moody algebras. All possible Satake diagrams for $A_1^{(1)}$ and $A_2^{(1)}$ along with the root transformations have been computed in detail. It is gratifying to observe that our results are in complete accord with those of Cornwell[8] for the untwisted affine Kac–Moody algebras $A_1^{(1)}$ and $A_2^{(1)}$. Section III is devoted to a discussion of the procedure for carrying out the Iwasawa and Langlands decompositions of these algebras within the present framework to enable us to compute parabolic subalgebras. The explicit construction of the Iwasawa and Langlands decompositions are done in Sec. IV for the untwisted affine algebras $A_1^{(1)}$ and $A_2^{(1)}$. It is shown that this leads to parabolic subalgebras which serve as a basis for obtaining the corresponding induced representations for $A_1^{(1)}$ and $A_2^{(1)}$ through the Schmidt construction. Concluding remarks are appended in Sec. V.

II. AFFINE KAC–MOODY ALGEBRAS AND SATAKE DIAGRAMS

An affine Kac–Moody algebra is generated by a set of generators h_{α_i}, $e_{\pm\alpha_i}$, $i=0,1,2,...,l$, where l is the rank of the associated Lie algebra, with the defining relations

$$[h_{\alpha_i},h_{\alpha_j}]=0, \quad [e_{\alpha_i},e_{-\alpha_j}]=\delta_{ij}h_{\alpha_i},$$

$$[e_{\pm\alpha_i},h]=\pm\alpha_i(h)e_{\pm\alpha_i}, \tag{2.1}$$

to complement the Serre's relations. The linear forms α_0, $\alpha_1,\ldots,\alpha_l$ are its simple roots, where $\alpha_0=\delta-\theta$, θ being the highest root and δ the imaginary root. h_{α_0}, $h_{\alpha_1},\ldots,h_{\alpha_l}$ together with the

 © 1998 American Institute of Physics

TABLE I. The Dynkin diagrams corresponding to the affine untwisted algebras.

Affine Untwisted Algebras	Dynkin Diagrams
$A_1^{(1)}$	
$A_l^{(1)}$	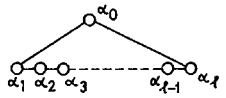
$B_l^{(1)}$	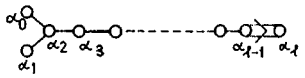
$C_l^{(1)}$	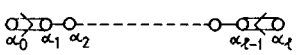
$D_l^{(1)}$	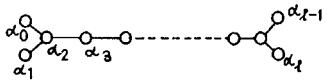
$E_6^{(1)}$	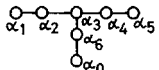
$E_7^{(1)}$	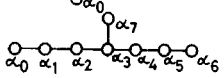
$E_8^{(1)}$	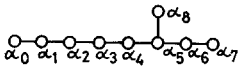
$F_4^{(1)}$	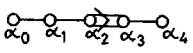
$G_2^{(1)}$	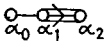

scaling element d, such that $\alpha_1(d)=0$, $i=1,2,...,l$ and $\alpha_0(d)=1$, constitute a basis of the Cartan subalgebra of affine algebra. The Dynkin diagrams of the affine untwisted Kac–Moody algebra are displayed in Table I.

The construction of the Satake diagrams proceeds as follows. Let R be a root systems of affine Kac–Moody algebra. For $\alpha\in R$, let $\overline{\alpha}=\alpha-\sigma(\alpha)$, where σ stands for the involutive automorphism of α. Let us introduce $R_-=\{\overline{\alpha}|\overline{\alpha}\neq 0,\alpha\in R\}$ and $R_0=\{\alpha\in R|\overline{\alpha}=0\}$. If B_- (resp B) denotes the basis of R_- (resp R) and B_0 a basis of R_0 then $B_0=B\cap R_0$. If $B_-=B\backslash B_0=\{\alpha_i\}$ and $B_0=\{\beta_i\}$, then it can be shown that

$$-\sigma(\alpha_i)=\alpha_{\pi(i)}+\sum_l \eta_{il}\beta_l\,. \tag{2.2}$$

where π are the involutive permutations of $(0,1,2,...,l)$ and η_{il} are non-negative integers. We can now associate with B its Satake diagram. Denote the root α_i by $\circ$ as usual (white roots) and the root β_i by $\bullet$ (black roots). If $\pi(i)=k$, it is indicated by

The Satake diagram determines the involution of R uniquely. We note that $\sigma(\beta_i)=\beta_i$ and if $\alpha\in R$ then $\alpha+\sigma(\alpha)\notin R$. As an illustration, consider the following examples.

TABLE II. Satake diagrams of $A_1^{(1)}$ along with their root automorphisms.

Satake Diagrams of $A_1^{(1)}$		Root Automorphisms
(i)	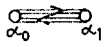	$-\sigma(\alpha_0)=\alpha_0$ $-\sigma(\alpha_1)=\alpha_1$
(ii)		$-\sigma(\alpha_0)=\alpha_1$ $-\sigma(\alpha_1)=\alpha_0$
(iii)	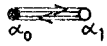	$-\sigma(\alpha_1)=\alpha_1+2\alpha_0$ $\sigma(\alpha_0)=\alpha_0$
(iv)	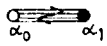	$-\sigma(\alpha_0)=\alpha_0+2\alpha_1$ $\sigma(\alpha_1)=\alpha_1$

A. Satake diagrams of $A_1^{(1)}$

The Cartan matrix of $A_1^{(1)}$ is

$$C=\begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}.$$

The two simple roots of $A_1^{(1)}$ are α_0 and α_1, where $\alpha_0=\delta-\alpha_1$. The possible Satake diagrams of $A_1^{(1)}$ along with their root automorphisms are depicted in Table II.

B. Satake diagrams of $A_2^{(1)}$

The Cartan matrix of $A_2^{(1)}$ is

$$C=\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

The three simple roots of $A_2^{(1)}$ are α_0, α_1, and α_2, where $\alpha_0=\delta-(\alpha_1+\alpha_2)$. The possible Satake diagrams of $A_2^{(1)}$ along with their root automorphisms are given by Table III.

III. IWASAWA AND LANGLANDS DECOMPOSITIONS OF AFFINE KAC–MOODY ALGEBRA

We now extend the notion of direct determination of Iwasawa and Langlands decompositions of Lie algebras. Let $\underset{\sim}{g}$ be a real Kac-Moody algebra generated by its compact real form $\underset{\sim}{g}_K$ by an involutive automorphism defined w.r.t. the Cartan subalgebra $\underset{\sim}{h}$ of $\underset{\sim}{g}_C$, $\underset{\sim}{g}_C$ being the complexification of $\underset{\sim}{g}$. The following commutation relations are satisfied by the elements of $\underset{\sim}{g}_C$.

$$[e_\alpha,h]=\alpha(h)e_\alpha, \quad h\in\underset{\sim}{h}, \quad \alpha\in\Delta,$$

$$[e_\alpha,e_\beta]=\begin{cases} N_{\alpha\beta}e_{\alpha+\beta} & \text{if } \alpha+\beta \text{ is a root} \\ 0 \quad \text{otherwise,} \end{cases} \tag{3.1}$$

$$[e_\alpha,e_{-\alpha}]=h_\alpha, \quad h_\alpha\in\underset{\sim}{h}.$$

Here Δ denotes the set of roots of $\underset{\sim}{g}_C$ w.r.t. $\underset{\sim}{h}$ and the Killing form is defined as $B(e_\alpha,e_{-\alpha})=-1$. Here $\alpha(h)=B(h,h_\alpha)$ is the Killing form of $\underset{\sim}{g}_C$. The compact real form $\underset{\sim}{g}_K$ may be taken

TABLE III. Satake diagrams of $A_2^{(1)}$ along with their root automorphisms.

Satake Diagrams of $A_2^{(1)}$	Root Automorphisms
(i)	$-\sigma(\alpha_0)=\alpha_0$ $-\sigma(\alpha_1)=\alpha_1$ $-\sigma(\alpha_2)=\alpha_2$
(ii) (a) (b) (c) 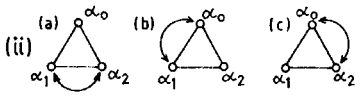	(a) $-\sigma(\alpha_1)=\alpha_2$, $-\sigma(\alpha_2)=\alpha_1$, $-\sigma(\alpha_0)=\alpha_0$ (b) $-\sigma(\alpha_1)=\alpha_0$, $-\sigma(\alpha_0)=\alpha_1$, $-\sigma(\alpha_2)=\alpha_2$ (c) $-\sigma(\alpha_0)=\alpha_2$, $-\sigma(\alpha_2)=\alpha_0$, $-\sigma(\alpha_1)=\alpha_1$
(iii) (a) (b) (c) 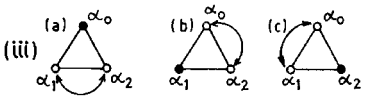	(a) $-\sigma(\alpha_1)=\alpha_2+\alpha_0$, $-\sigma(\alpha_2)=\alpha_1+\alpha_0$, $\sigma(\alpha_0)=\alpha_0$ (b) $-\sigma(\alpha_0)=\alpha_2+\alpha_1$, $-\sigma(\alpha_2)=\alpha_0+\alpha_1$, $\sigma(\alpha_1)=\alpha_1$ (c) $\sigma(\alpha_2)=\alpha_2$, $-\sigma(\alpha_1)=\alpha_0+\alpha_2$, $-\sigma(\alpha_0)=\alpha_1+\alpha_2$
(iv) (a) (b) (c) 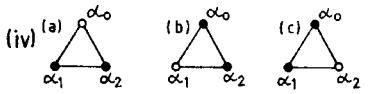	(a) $-\sigma(\alpha_0)=\alpha_0+2\alpha_1+2\alpha_2$, $\sigma(\alpha_1)=\alpha_1$, $\sigma(\alpha_2)=\alpha_2$ (b) $-\sigma(\alpha_1)=\alpha_1+2\alpha_0+2\alpha_2$, $\sigma(\alpha_0)=\alpha_0$, $\sigma(\alpha_2)=\alpha_2$ (c) $-\sigma(\alpha_2)=\alpha_2+2\alpha_1+2\alpha_0$, $\sigma(\alpha_0)=\alpha_0$, $\sigma(\alpha_1)=\alpha_1$

to consist of ih_α, $\alpha=\alpha_0,\alpha_1,\ldots,\alpha_l$ and $(e_\alpha+e_{-\alpha})$, $i(e_\alpha-e_{-\alpha})\forall\alpha$. Let $\underset{\sim}{k}$ be the maximal compact subalgebra of $\underset{\sim}{g}$ defined such that $a\in\underset{\sim}{k}$ iff $a\in\underset{\sim}{g}_C$ and $\sigma a=a$. Let $\underset{\sim}{p}$ be the subspace of $\underset{\sim}{g}$ such that $a\in\underset{\sim}{p}$ iff $\tilde{a}\in\underset{\sim}{g}_C$ and $\sigma a=-a$. Thus $\underset{\sim}{k}$ and $\underset{\sim}{p}$ are given by

$$\underset{\sim}{k}=\{ih_\alpha,\alpha=\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_l \text{ and } (e_\alpha+e_{-\alpha}),\ i(e_\alpha-e_{-\alpha})\forall\alpha|\exp\ \alpha(h)=+1\},\quad (3.2)$$

$$\underset{\sim}{p}=\{i(e_\alpha+e_{-\alpha}),\quad (e_\alpha-e_{-\alpha})\forall\alpha|\exp\ \alpha(h)=-1\}.\quad (3.3)$$

Let $\underset{\sim}{a}$ be the maximal Abelian subalgebra of $\underset{\sim}{p}$ with dimension $|m_1|$ and $\underset{\sim}{m}$ be the centralizer of $\underset{\sim}{a}$ in $\underset{\sim}{k}$. Complexification of $\underset{\sim}{a}\oplus\underset{\sim}{m}$ together with a scaling element d' gives a Cartan subalgebra $\underset{\sim}{h}'$ of $\underset{\sim}{g}_C$ with basis H_0', $H_1',\ldots,H_l'$ and d'. There exists an inner automorphism $V:\underset{\sim}{h}'\to\underset{\sim}{h}$, i.e.,

$$H_j=VH_j',\quad \text{where } V=\prod_\alpha V_\alpha,\quad \alpha\in\Delta.\quad (3.4)$$

Let Δ^+ be the set of positive roots, then

$$h_\alpha=\sum_{j=0}^{l} b_j(\alpha)H_j.\quad (3.5)$$

Thus $\alpha\in\Delta^+$ iff $b_j(\alpha)>0$, where j is the least index such that $b_j(\alpha)\neq 0$. The positive roots can be again divided into the following classes:

$$\text{(i)}\quad \Delta_+^+=\{\alpha|\alpha\in\Delta^+,\quad \alpha(h)\neq\alpha(V\sigma V^{-1}h)\forall h\in\underset{\sim}{h}\},\quad (3.6)$$

$$\text{(ii)}\quad \Delta_-^+=\{\alpha|\alpha\in\Delta^+,\quad \alpha(h)=\alpha(V\sigma V^{-1}h)\forall h\in\underset{\sim}{h}\}.\quad (3.7)$$

Let the subalgebra $\tilde{\underline{\eta}}$ be spanned by the elements $V^{-1}e_\alpha$ for $\alpha \in \Delta_+^+$ and $\underline{n} = \tilde{\underline{\eta}} \cap g$, where $\tilde{\underline{\eta}}$ and $\underline{n}$ are nilpotent subalgebras of $\underline{g}_C$ and $\underline{g}$, respectively. Thus the Iwasawa decomposition of $\underline{g}$ is given by

$$\underline{g} = \underline{k} \oplus \underline{a} \oplus \underline{n}. \tag{3.8}$$

A minimal parabolic subalgebra is defined to be any subalgebra that is conjugate to

$$\underline{p}_1 = \underline{m} \oplus \underline{a} \oplus \underline{n}. \tag{3.9}$$

Any subalgebra of g containing the minimal parabolic subalgebra is a general parabolic subalgebra. There exist $2^{|\tilde{m}_1|}$ classes of parabolic subalgebras of $\underline{g}$ and in each such class there is a standard parabolic subalgebra which can be obtained through the following prescription. Let Σ be the set of roots for $\underline{a}$ and ψ be the set of positive roots in Σ. Let θ denote the subset of ψ and $\langle\theta\rangle$ the set of roots in Σ which arise as a linear combination of roots in θ. Define $\langle\theta\rangle_\pm = \Sigma_\pm \cap \langle\theta\rangle$, where Σ_+ and Σ_- denote the positive and negative roots in Σ. Let $\eta_+(\theta)$, $\eta_-(\theta)$, and $\eta(\theta)$ denote the subspaces of $\underline{a}$ corresponding to $\langle\theta\rangle_+$, $\langle\theta\rangle_-$, and $\Sigma_\pm - \langle\theta\rangle_\pm$, respectively. Now define

$$a_\theta = \{a \in \underline{a} | \lambda(a) = 0 \forall \lambda \in \theta\}, \tag{3.10}$$

and let $a(\theta)$ be the orthogonal component of a_θ in $\underline{a}$ with respect to the Cartan Killing form, then

$$p_\theta = m_\theta \oplus a_\theta \oplus \eta_\theta, \tag{3.11}$$

is a parabolic subalgebra of g, where

$$m_\theta = m \oplus \eta_+(\theta) \oplus \eta_-(\theta) \oplus a(\theta). \tag{3.12}$$

A real Cartan subalgebra $\underline{h}$ is said to be σ invariant if

$$\underline{h} = (\underline{h} \cap \underline{k}) \oplus (\underline{h} \cap \underline{p}). \tag{3.13}$$

A parabolic subalgebra p_θ is said to be cuspidal if there exists a σ-invariant real Cartan subalgebra $\underline{h}$ such that

$$a_\theta = \underline{h} \cap \underline{p}. \tag{3.14}$$

This shows that the minimal parabolic subalgebra is cuspidal.

IV. IWASAWA AND LANGLANDS DECOMPOSITIONS OF $A_1^{(1)}$

Let us consider the involutive automorphism of $A_1^{(1)}$ determined by any one of the Satake diagrams, say (iii), of Table II. The positive roots of $A_1^{(1)}$ are given by

$$\Delta = \{k\alpha_0 + (k-1)\alpha_1, \quad (k-1)\alpha_0 + k\alpha_1, \quad k(\alpha_1 + \alpha_0) \quad \text{where } k \in \mathbb{Z}^+\}. \tag{4.1}$$

These roots are expressed in terms of roots α_1 and imaginary root δ as

$$\Delta = \{\alpha_1, \pm\alpha_1 + n\delta, \text{ where } n \in \mathbb{Z}^+\}. \tag{4.2}$$

From the Satake diagram we see that the basic root automorphisms are given by

$$\sigma(\alpha_0) = \alpha_0, \quad -\sigma(\alpha_1) = \alpha_1 + 2\alpha_0. \tag{4.3}$$

The automorphisms of other roots are determined from the above. For example,

$$\sigma(2\alpha_1 + 3\alpha_0) = -(\alpha_0 + 2\alpha_1), \quad \sigma(3\alpha_1 + 4\alpha_0) = -(2\alpha_0 + 3\alpha_1),$$

and so on. Similarly,

$$\sigma(k(\alpha_1+\alpha_0))=-k(\alpha_1+\alpha_0).$$

From all these relations we see that

$$\exp\ \alpha(h)=+1 \quad \text{for}\ \ \alpha=(k-1)\alpha_1+k\alpha_0 \tag{4.4}$$

and

$$\exp\ \alpha(h)=-1 \quad \text{for}\ \ \alpha=k\alpha_1+(k-1)\alpha_0, \quad k(\alpha_1+\alpha_0). \tag{4.5}$$

So, for $A_1^{(1)}$, k is given by

$$\underset{\sim}{k}=\{ih_\alpha \quad \text{for}\ \ \alpha=\alpha_0,\alpha_1 \quad \text{and}\ \ (e_\alpha+e_{-\alpha}), \quad i(e_\alpha-e_{-\alpha}) \quad \text{for}\ \ \alpha=(k-1)\alpha_1+k\alpha_0\}, \tag{4.6}$$

and similarly $\underset{\sim}{p}$ is given by

$$\underset{\sim}{p}=\{i(e_\alpha+e_{-\alpha}), \quad (e_\alpha-e_{-\alpha}) \quad \text{for}\ \ \alpha=k\alpha_1+(k-1)\alpha_0, \quad k(\alpha_1+\alpha_0)\}. \tag{4.7}$$

We now select a maximal Abelian subalgebra $\underset{\sim}{a}$ in the vector space $\underset{\sim}{p}$. It is clear that $\underset{\sim}{a}$ is one dimensional and may be chosen to have a basis element

$$H_0'=i(e_\alpha+e_{-\alpha}) \quad \text{with}\ \ \alpha=\alpha_1. \tag{4.8}$$

Here $R_A=\{\alpha_1\}$ and R_M is empty. m is one dimensional and its basis element may be taken to be

$$-iH_1'=ih_\beta, \quad \text{where}\ \ \beta=\alpha_1+\alpha_0 \quad \text{as}\ \ \alpha_1(h_\beta)=0, \tag{4.9}$$

so that $H_1'=-h_{\alpha_1+\alpha_0}$.

Note that H_0', H_1' together with a scaling element d' such that $\alpha_0(d')=1$ and $\alpha_1(d')=0$, are the elements of the Cartan subalgebra H'. The inner automorphism of $A_1^{(1)}$ is given by $V_\alpha^{(1)}$ which is obtained from V_{α_1} (inner automorphism of A_1) by the following definition:

$$V_{\alpha_1}^{(1)}(e_{\alpha+n\delta})=t^n\otimes V_{\alpha_1}(e_\alpha) \tag{4.10a}$$

and

$$V_\alpha^{(1)}(d')=d. \tag{4.10b}$$

Applying this to the Cartan subalgebra H' of $A_1^{(1)}$, we obtain

$$H_0=-\left\{\frac{2}{(\alpha_1,\alpha_1)}\right\}^{1/2}h_{\alpha_1}=-2^{1/2}h_{\alpha_1}, \quad H_1=-(h_{\alpha_0}+h_{\alpha_1}), \tag{4.11}$$

along with the scaling element d. With respect to this Cartan subalgebra the set of positive roots is given by

$$\Delta^+=\{-((k-1)\alpha_0+k\alpha_1),(k\alpha_0+(k-1)\alpha_1),-k(\alpha_1+\alpha_0)\}, \tag{4.12}$$

and Δ_+^+ and Δ_-^+ can be written as

$$\Delta_-^+=\{-k(\alpha_1+\alpha_0)\}, \quad \Delta_+^+=\{-((k-1)\alpha_0+k\alpha_1),(k\alpha_0+(k-1)\alpha_1)\}. \tag{4.13}$$

For $A_1^{(1)}$, the basis elements of $\tilde{\underset{\sim}{n}}=\underset{\sim}{n}$ are given by $V_{\alpha_1}^{(1)^{-1}}e_{\alpha'}$ where $\alpha\in\Delta_+^+$ and we see that these are given by the structures

$$
\begin{aligned}
V_{\alpha_1}^{(1)^{-1}} e_{-\alpha_1} &= -\tfrac{1}{2}(e_{\alpha_1} - e_{-\alpha_1}) - \tfrac{1}{2} i 2^{1/2} h_{\alpha_1}, \\
V_{\alpha_1}^{(1)^{-1}} e_{\alpha_0} &= V_{\alpha_1}^{(1)^{-1}} (e_{-\alpha_1+\delta}) = t \otimes \{-\tfrac{1}{2}(e_{\alpha_1} - e_{-\alpha_1}) - \tfrac{1}{2} i 2^{1/2} h_{\alpha_1}\}, \\
V_{\alpha_1}^{(1)^{-1}} e_{-((k-1)\alpha_0 + k\alpha_1)} &= V_{\alpha_1}^{(1)^{-1}} e_{-\alpha_1+(1-k)\delta} = t^{(1-k)} \otimes \{-\tfrac{1}{2}(e_{\alpha_1} - e_{-\alpha_1}) - \tfrac{1}{2} i 2^{1/2} h_{\alpha_1}\}, \\
V_{\alpha_1}^{(1)^{-1}} e_{(k\alpha_0 + (k-1)\alpha_1)} &= V_{\alpha_1}^{(1)^{-1}} e_{-\alpha_1+k\delta} = t^k \otimes \{-\tfrac{1}{2}(e_{\alpha_1} - e_{-\alpha}) - \tfrac{1}{2} i 2^{1/2} h_{\alpha_1}\}.
\end{aligned}
\tag{4.14}
$$

The required Iwasawa decomposition is

$$
A_1^{(1)} = \underset{\sim}{k} \oplus \underset{\sim}{a} \oplus \underset{\sim}{n},
$$

where $\underset{\sim}{k}, \underset{\sim}{a}, \underset{\sim}{n}$ are determined by Eqs. (4.6), (4.8), and (4.14), respectively.

There are $2^{|m_1|}$ classes of parabolic subalgebras of affine Kac–Moody algebra, where $|m_1|$ is the dimension of $\underset{\sim}{a}$. For $A_1^{(1)}$ we see that $|m_1| = 1$ and there will, therefore, be two parabolic subalgebras, one being the minimal parabolic subalgebra and the other the algebra itself. The minimal parabolic subalgebra is given by

$$
\underset{\sim}{p}_1 = \underset{\sim}{m} \oplus \underset{\sim}{a} \oplus \underset{\sim}{n}, \tag{4.15}
$$

where

$$
\underset{\sim}{m} = \{i h_\alpha \quad \text{for} \quad \alpha = \alpha_1 + \alpha_0\}, \quad \underset{\sim}{a} = \{i(e_\alpha + e_{-\alpha}) \quad \text{for} \quad \alpha = \alpha_1\},
$$

and $\underset{\sim}{n}$ is given by Eq. (4.14). The minimal parabolic subalgebra, being the cuspidal parabolic subalgebra, leads to the following definition of a series of representation:

$$
\rho = \mathrm{ind}_{P_1}^{G}(\sigma \otimes \tau), \tag{4.16}
$$

where $\sigma \in \hat{M}$, $\tau \in \hat{N}$, and $P_1 = MAN$, is the cuspidal subgroup. Note that P_1, M, A, N, and G are the groups corresponding to the algebras $\underset{\sim}{p}_1$, $\underset{\sim}{m}$, $\underset{\sim}{a}$, $\underset{\sim}{n}$, and $\underset{\sim}{g}$, respectively. This defines the principal series representation. As P_1 is minimal parabolic and $\tilde{\sigma} = 1_M$, M being Abelian, is the trivial representation, so ρ is irreducible.

A. Iwasawa decomposition of $A_2^{(1)}$

We consider the involutive automorphisms of $A_2^{(1)}$ determined by the Satake diagram iv(b) of Table III. The basic root automorphisms are given by

$$
\sigma(\alpha_0) = \alpha_0, \quad -\sigma(\alpha_1) = \alpha_1 + 2\alpha_0 + 2\alpha_2, \quad \sigma(\alpha_2) = \alpha_2. \tag{4.17}
$$

In terms of roots ''α'' and ''δ'' the basic root automorphisms can be rewritten as

$$
\sigma(\delta - \alpha_1 - \alpha_2) = \delta - \alpha_1 - \alpha_2, \quad -\sigma(\alpha_1) = -\alpha_1 + 2\delta, \quad \sigma(\alpha_2) = \alpha_2. \tag{4.18}
$$

The positive roots of $A_2^{(1)}$ are given by

$$
\Delta = \{\alpha_1, \alpha_2, \quad \pm\alpha_1 + n\delta, \quad \pm\alpha_2 + n\delta, \quad \pm(\alpha_1 + \alpha_2) + n\delta, \quad n\delta, \quad \text{where} \quad n \in \mathbb{Z}^+\}. \tag{4.19}
$$

We can apply the basic root automorphism to find out the automorphisms of other roots and it is easy to see that

$$
\exp\, \alpha(h) = +1 \quad \text{for} \quad \alpha = \alpha_2, \quad -\alpha_1 + n\delta, \quad \alpha_2 + n\delta, \quad -(\alpha_1 + \alpha_2) + n\delta, \tag{4.20}
$$

$$
\exp\, \alpha(h) = -1 \quad \text{for} \quad \alpha = \alpha_1, \quad \alpha_1 + n\delta, \quad -\alpha_2 + n\delta, \quad (\alpha_1 + \alpha_2) + n\delta, \quad n\delta. \tag{4.21}
$$

For $A_2^{(1)}$ the structures for $\underset{\sim}{k}$ and $\underset{\sim}{p}$ are given by

$$\underset{\sim}{k}=\{ih_\alpha \quad \text{for } \alpha=\alpha_0,\alpha_1,\alpha_2, \text{ and } (e_\alpha+e_{-\alpha}), \quad i(e_\alpha-e_{-\alpha})$$

for α given by Eq. (4.20)} and (4.22)

$$\underset{\sim}{p}=\{i(e_\alpha+e_{-\alpha}), \quad (e_\alpha-e_{-\alpha}) \quad \text{for } \alpha \text{ given by Eq. (4.21)}\}. \tag{4.23}$$

Again, in an analogous manner, we select a maximal Abelian subalgebra $\underset{\sim}{a}$; we see that $\underset{\sim}{a}$ is one dimensional and its basis element is given by

$$H_0'=i(e_\alpha+e_{-\alpha}) \quad \text{for } \alpha=\alpha_1. \tag{4.24}$$

Here $R_A=\{\alpha_1\}$, and R_M is empty. m is two dimensional and its basis elements are given by

$$H_1'=-(h_{\alpha_1}+2h_{\alpha_2}), \quad H_2'=-h_\delta. \tag{4.25}$$

The elements H_0', H_1', H_2' together with the scaling element d' form the Cartan subalgebra $\underset{\sim}{h}'$. Defining the inner automorphism of $A_2^{(1)}$ as before, we obtain

$$H_0=-2^{1/2}h_{\alpha_1}, \quad H_1=-(h_{\alpha_1}+2h_{\alpha_2}), \quad H_2=-h_\delta \quad \text{and } d. \tag{4.26}$$

The set of positive roots with respect to this Cartan subalgebra is given by

$$\begin{aligned}\Delta^+=\{&-\alpha_1,\alpha_2,-(\alpha_1+\alpha_2),-(\alpha_1+n\delta),(-\alpha_1+n\delta),\alpha_2+n\delta,\\ &-(-\alpha_2+n\delta),-((\alpha_1+\alpha_2)+n\delta),-(\alpha_1+\alpha_2)+n\delta,-n\delta)\},\end{aligned} \tag{4.27}$$

The Δ_+^+ and Δ_-^+ are then written as

$$\Delta_-^+=\{-n\delta\}, \quad \Delta_+^+=\Delta^+/\Delta_-^+. \tag{4.28}$$

The elements of $\underset{\sim}{\tilde{\eta}}$ are determined by the structure $V_{\alpha_1}^{(1)^{-1}}e_\alpha$, where $\alpha\in\Delta_+^+$. These are given by

$$\begin{aligned}V_{\alpha_1}^{(1)^{-1}}e_{-\alpha_1}&=-\tfrac{1}{2}(e_{\alpha_1}-e_{-\alpha_1})-\tfrac{1}{2}i2^{-1/2}h_{\alpha_1},\\ V_{\alpha_1}^{(1)^{-1}}e_{\alpha_2}&=2^{-1/2}e_{\alpha_2}-i2^{-1/2}(\mathrm{Sgn}\, N_{\alpha_1,\alpha_2})e_{\alpha_1+\alpha_2},\\ V_{\alpha_1}^{(1)^{-1}}e_{-(\alpha_1+\alpha_2)}&=-i2^{-1/2}(\mathrm{Sgn}\, N_{\alpha_1,-(\alpha_1+\alpha_2)})e_{-\alpha_2}+2^{-1/2}e_{-(\alpha_1+\alpha_2)}.\end{aligned} \tag{4.29}$$

The other elements are obtained by using the relations

$$\begin{aligned}&V_{\alpha_1}^{(1)^{-1}}e_{-\alpha_1\pm n\delta}=t^{\pm n}\otimes V_{\alpha_1}^{(1)^{-1}}e_{-\alpha_1}, \quad V_{\alpha_1}^{(1)^{-1}}e_{\alpha_2\pm n\delta}=t^{\pm n}\otimes V_{\alpha_1}^{(1)^{-1}}e_{\alpha_2},\\ &\text{and } V_{\alpha_1}^{(1)^{-1}}e_{-(\alpha_1+\alpha_2)\pm n\delta}=t^{\pm n}\otimes V_{\alpha_1}^{(1)^{-1}}e_{-(\alpha_1+\alpha_2)}.\end{aligned} \tag{4.30}$$

Now $\underset{\sim}{n}$ can be calculated from $\underset{\sim}{\tilde{\eta}}$ and some of its typical elements are given by

$$-\frac{1}{2}(e_{\alpha_1}-e_{-\alpha_1})-\frac{i}{2}2^{-1/2}h_{\alpha_1},$$

$$2^{-1/2}(e_{\alpha_2}-e_{-\alpha_2})-i2^{-1/2}(\mathrm{Sgn}\, N_{\alpha_1,\alpha_2})\{e_{\alpha_1+\alpha_2}-e_{-(\alpha_1+\alpha_2)}\},$$

$$i2^{-1/2}(e_{(\alpha_1+\alpha_2)}-e_{-(\alpha_1+\alpha_2)})+2^{-1/2}(\mathrm{Sgn}\, N_{\alpha_1,-(\alpha_1+\alpha_2)})\{e_{\alpha_2}-e_{-\alpha_2}\}.$$

The remaining elements are similarly obtained by using the above relations (4.30).

For $A_2^{(1)}$ also we see that there will be two parabolic subalgebras, as is the case with $A_1^{(1)}$; one being the minimal parabolic and the other the algebra itself. The minimal parabolic subalgebra is given by

$$\underset{\sim}{p}_1 = \underset{\sim}{m} \oplus \underset{\sim}{a} \oplus \underset{\sim}{n}. \tag{4.31}$$

V. CONCLUSION

Before closing, a few remarks are pertinent. In the present investigation, we have computed the Iwasawa and Langlands decompositions of the affine untwisted Kac–Moody algebras $A_1^{(1)}$ and $A_2^{(1)}$ with the express purpose of determining the parabolic subalgebras which ultimately provide the induced representation. These constructions have been facilitated by the involutive automorphisms derived from the Satake diagrams of these algebras. As mentioned earlier our results are in agreement with those of Cornwell[8] for root automorphisms. We believe that this method is readily amenable to generalization to the case of affine twisted Kac–Moody algebras with suitable redefinition of inner automorphism, supplemented possibly with some extra rule. Such studies are in progress and will be reported in a subsequent communication.

ACKNOWLEDGMENT

We express our deep gratitude to Dr. R. S. Kaushal for fruitful discussions and helpful suggestions.

[1] V. G. Kac, *Infinite Dimensional Lie Algebras*, 2nd Ed. (Cambridge University Press, Cambridge, 1985).
[2] P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).
[3] P. Furlan, G. M. Sotkov, I. T. Todorov, Two Dimensional Conformal Field Theory, SISSA/ISAS, Trieste (Preprint) 60/88/EP.
[4] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory, Vol. I and II* (Cambridge University Press, Cambridge, 1986).
[5] M. Jimbo and T. Miwa, Publ. RIMS, Kyoto Univ. **19**, 943 (1983); D. I. Olive and N. Turok, Nucl. Phys. B **257**, 277 (1985).
[6] O. Loos, *Symmetric Spaces Vol. II* (Benjamin, New York, 1969).
[7] F. Levstein, J. Algebra **114**, 489 (1988).
[8] J. F. Cornwell, J. Phys. A: Math. & Gen. **25**, 2311 (1992); **25**, 2335 (1992); **25**, 3955 (1992).
[9] J. F. Cornwell, J. Math. Phys. **16**, 1992 (1975); **20**, 547 (1979).
[10] R. P. Lipsman, *Lect. Notes in Math., Vol. 338* (Springer, Berlin, 1974).
[11] K. C. Pati and D. Parashar, J. Phys. A: Math. & Gen. **31**, 767 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 9 SEPTEMBER 1998

Cohomology of Lie superalgebras and their generalizations

M. Scheunert[a)]
Physikalisches Institut der Universität Bonn, Nussallee 12, D–53115 Bonn, Germany

R. B. Zhang[b)]
Department of Pure Mathematics, University of Adelaide, Adelaide, Australia

(Received 24 February 1997; accepted for publication 13 April 1998)

The cohomology groups of Lie superalgebras and, more generally, of ε Lie algebras, are introduced and investigated. The main emphasis is on the case where the module of coefficients is nontrivial. Two general propositions are proved, which help to calculate the cohomology groups. Several examples are included to show the peculiarities of the super case. For $L=sl(1|2)$, the cohomology groups $H^1(L,V)$ and $H^2(L,V)$, with V a finite-dimensional simple graded L-module, are determined, and the result is used to show that $H^2(L,U(L))$ [with $U(L)$ the enveloping algebra of L] is trivial. This implies that the superalgebra $U(L)$ does not admit any nontrivial formal deformations (in the sense of Gerstenhaber). Garland's theory of universal central extensions of Lie algebras is generalized to the case of ε Lie algebras. © *1998 American Institute of Physics.* [S0022-2488(98)01607-7]

I. INTRODUCTION

Cohomology is an important tool in mathematics. Its range of applications contains algebra and topology as well as the theory of smooth manifolds or of holomorphic functions. The cohomology theory of Lie algebras (whose generalization is the main objective of the present paper) has its origins in the work of E. Cartan, but the foundation of the theory, as an independent topic of research, is due to Chevalley and Eilenberg,[1] Koszul,[2] and Hochschild and Serre.[3] A unifying treatment of the cohomology theory of groups, associative algebras, and Lie algebras has been given by Cartan and Eilenberg.[4]

As is well-known, several classical results in Lie algebra theory have a cohomological interpretation. Thus, if L is a Lie algebra, the structure of the extensions of L-modules (and hence the question of semi-simplicity of L-modules) is described by the 1-cohomology of L, and the structure of the Lie algebra extensions (and hence the Levi, Malcev theorem) is related to the 2-cohomology.[1,5]

In view of these remarks, it is hardly surprising that shortly after the birth of supersymmetry, a paper appeared[6] in which the most basic constructions and results of the classical theory are generalized to the case of Lie superalgebras.[7,8] Later on, Leites and Fuks calculated the cohomology groups of the classical Lie superalgebras with trivial coefficients[9] (see also Ref. 10), and Leites and Serganova used cohomological methods to determine systems of generators and relations for these algebras.[11] In addition, we should mention a letter by Retakh and Feigin[12] and the papers by Tripathy and coworkers (see Ref. 13 and the papers cited therein). Apart from this, very little seems to be known about the cohomology of Lie superalgebras, in particular, about the cohomology with nontrivial coefficients.

After the present work was posted on the Internet and submitted for publication, Gruson sent us a copy of the galley proofs of an article,[14] in which she proves the finiteness of the homology of certain modules over a Lie superalgebra. She also drew our attention to a paper by Tanaka,[15] in which the author calculates the dimensions of the homology and cohomology groups of $sl(2|1)$ with coefficients in an arbitrary finite-dimensional simple module. This has a direct bearing on part of our work, and we shall comment on it later on.

Our present interest in this topic has been stimulated by the theory of quantum algebras.[16,17] It is known that, for any semi-simple Lie algebra L, the enveloping algebra $U(L)$ does not admit

a)Electronic mail: scheunert@physik.uni-bonn.de
b)Electronic mail: rzhang@maths.adelaide.edu.au

 © 1998 American Institute of Physics

any nontrivial formal deformations in the sense of Gerstenhaber.[18] We would like to know for which of the simple Lie superalgebras the same holds true. Proceeding as in Ref. 19, this amounts to showing that the second cohomology group $H^2(L,U(L))$ is trivial. Unfortunately, we are far from proving a general result of this type. In fact, all we can say at present is that this is true if L is one of the algebras $osp(1|2n)$ (this is easy to see and, certainly, was already known to Kac[7]), or if L is equal to $sl(1|2)$ (this will be proved in the present work). Moreover, in the meantime we have proved that $H^2(L,U(L))$ is also trivial for the algebras $sl(m|1)$ with $m\geqslant 3$. This will be published elsewhere. Accordingly, our paper is going to be explorative in character.

Let us briefly describe the setup of the present article. In Sec. II we collect the basic definitions and constructions, which will be needed in the subsequent sections. Using the graded multilinear algebra as described in Ref. 20 this is just a simple transcription of the classical situation to the graded case. In doing so, we shall follow the approach of Ref. 3 (see also Ref. 1). In addition, we comment on some peculiarities of the graded case, and we prove two propositions, which will be the main tools in our calculations. The first one shows that it is sufficient to investigate cochains and coboundaries with certain invariance properties; the second proposition gives a general criterion for the cohomology with values in a given module to be trivial.

Section III contains several examples. In the first example we consider the cohomology of semi-simple Lie algebras. This example is included mainly in order to show how far we can get, under the most favorable circumstances, by only using the results of Sec. II. A short look at this example reveals that the same arguments can be applied to the Lie superalgebras $osp(1|2n)$. This will be the content of the second example. The third example contains some special information on the cohomology with trivial coefficients (again well-known in the case of semi-simple Lie algebras). In the fourth example we use the second proposition of Sec. II to determine, for the general linear Lie superalgebras $gl(m|n)$, which of the finite-dimensional simple graded $gl(m|n)$-modules might have a nontrivial cohomology. This discussion also covers the case of the special linear Lie superalgebras $sl(m|n)$, provided that $m\neq n$.

In Sec. IV we generalize Garland's theory of universal central coverings (i.e., extensions) of Lie algebras[21] to the general graded setting. As a simple example, we determine the universal covering of the simple Lie superalgebras $sl(n|n)$/center, with $n\geqslant 2$.

In Sec. V we consider the Lie superalgebra $L=sl(1|2)$ and determine the finite-dimensional simple graded L-modules V for which $H^1(L,V)$ or $H^2(L,V)$ is nontrivial. The result is then used to show that $H^2(L,U(L))$ is trivial. Of course, the modules V in question can be read off from Ref. 15. Nevertheless, we think it is worthwhile to present our own calculations. First of all, this makes our paper self-contained in this respect. Secondly, since we only determine $H^1(L,V)$ and $H^2(L,V)$, our calculations are much simpler and the results much more explicit than those of Ref. 15, where $\dim H^n(L,V)$ is determined for arbitrary n. Finally, this part of our paper is meant to explain how one might investigate $H^2(L,U(L))$ for more general Lie superalgebras L.

Section VI contains a brief discussion of our results. The paper is closed by two appendices. In Appendix A we describe the analog of the exterior (i.e., Grassmann) algebra in the general graded setting. Appendix B contains some special information related to Sec. V. This will be used to show that, for $L=sl(1|2)$, the cohomology group $H^1(L,U(L))$ is nontrivial.

We close this introduction by explaining some of our conventions. The base field will always be a field $\mathbb{K}$ of characteristic zero. We use the standard notation $\mathbb{Z}$ for the ring of integers and $\mathbb{N}$ for the set of natural numbers (according to our convention, $\mathbb{N}$ contains 0). The set of strictly positive integers will be denoted by $\mathbb{N}_*$. As explained above, in Sec. II and Sec. IV as well as in Appendix A we are going to consider the so-called ε Lie algebras[22] (also called color Lie algebras[23]). In these sections, Γ is an arbitrary Abelian group and ε is a commutation factor on Γ with values in $\mathbb{K}$. The reader who is only interested in Lie superalgebras may simply set $\Gamma=\mathbb{Z}_2$ or $\Gamma=\mathbb{Z}$ and choose ε to be the standard commutation factor of supersymmetry, defined by $\varepsilon(\alpha,\beta)=(-1)^{\alpha\beta}$ for all $\alpha,\beta\in\Gamma$. The choice $\Gamma=\mathbb{Z}_2$ leads to general Lie superalgebras, for $\Gamma=\mathbb{Z}$, we obtain the consistently $\mathbb{Z}$-graded Lie superalgebras in the sense of Ref. 7. The multiplication in an ε Lie algebra (and hence in a Lie superalgebra) will be denoted by a pointed bracket $\langle,\rangle$.

II. COHOMOLOGY OF ε LIE ALGEBRAS: DEFINITIONS AND SOME BASIC RESULTS

Let us first introduce the notation which will be used throughout the present section. Quite generally, we adopt the conventions of Ref. 20. In the following, Γ denotes an Abelian group,

$\varepsilon:\Gamma\times\Gamma\to\mathbb{K}$ a commutation factor on Γ with values in $\mathbb{K}$, L an ε Lie algebra,[22] and V a graded L-module. All gradations are understood to be Γ-gradations. If an element of a graded vector space is homogeneous of a certain degree, then this degree will be denoted by the lower case ''Greek analog'' of the letter denoting the element itself. For example, if A,b',g,x are homogeneous elements, their degrees will be denoted by $\alpha,\beta',\gamma,\xi\in\Gamma$, respectively. Only occasionally (but in all cases where there could be any doubts) the degrees will be specified explicitly.

Let $n\geqslant 1$ be a natural number and let

$$\mathrm{Lgr}_n(L,V)=\mathrm{Lgr}_n(\underbrace{L,\ldots,L}_{n};V) \tag{2.1}$$

be the graded vector space of all n-linear mappings of $L^n=L\times\cdots\times L$ into V which can be written as a sum of homogeneous n-linear mappings of L^n into V. Recall that $\mathrm{Lgr}_n(L,V)$ is equal to the space of all n-linear mappings of L^n into V if (for example) Γ is finite or if L is finite-dimensional.

According to Ref. 20 $\mathrm{Lgr}_n(L,V)$ is a graded L-module: If $A\in L$ and $g\in\mathrm{Lgr}_n(L,V)$ are homogeneous, the action of A on g is defined by

$$\begin{aligned}(A_L(g))(A_1,\ldots,A_n)&=A_V(g(A_1,\ldots,A_n))\\&\quad-\sum_{r=1}^{n}\varepsilon(\alpha,\gamma+\alpha_1+\cdots+\alpha_{r-1})g(A_1,\ldots,\langle A,A_r\rangle,\ldots,A_n),\end{aligned} \tag{2.2}$$

for all homogeneous elements $A_1,\ldots,A_n\in L$. [Recall that the ε-commutator of two elements $A,B\in L$ is denoted by $\langle A,B\rangle$. Quite generally, if V is a graded L-module, the representative of an element $A\in L$ in V will be denoted by A_V. If x is an element of V, we shall frequently simplify the notation and write $A\cdot x$ in place of $A_V(x)$.]

On the other hand, again according to Ref. 20, there is a natural representation of the symmetric group $\mathcal{S}_n$ in $\mathrm{Lgr}_n(L,V)$. To describe it, we define a map

$$\varepsilon_n:\mathcal{S}_n\times\Gamma^n\to\mathbb{K}, \tag{2.3}$$

by

$$\varepsilon_n(\pi^{-1};\gamma_1,\ldots,\gamma_n)=\prod_{\substack{i<j\\ \pi(i)>\pi(j)}}\varepsilon(\gamma_i,\gamma_j), \tag{2.4}$$

for all $\pi\in\mathcal{S}_n$ and $\gamma_1,\ldots,\gamma_n\in\Gamma$. The basic property of this map is that

$$\varepsilon_n(\pi\tau;\gamma_1,\ldots,\gamma_n)=\varepsilon_n(\pi;\gamma_1,\ldots,\gamma_n)\varepsilon_n(\tau;\gamma_{\pi(1)},\ldots,\gamma_{\pi(n)}), \tag{2.5}$$

for all $\pi,\tau\in\mathcal{S}_n$ and all $\gamma_1,\ldots,\gamma_n\in\Gamma$. Then the representation $\pi\to\check{S}_\pi$ of $\mathcal{S}_n$ in $\mathrm{Lgr}_n(L,V)$ is given by

$$(\check{S}_\pi(g))(A_1,\ldots,A_n)=\varepsilon_n(\pi;\alpha_1,\ldots,\alpha_n)g(A_{\pi(1)},\ldots,A_{\pi(n)}), \tag{2.6}$$

for all $\pi\in\mathcal{S}_n$, $g\in\mathrm{Lgr}_n(L,V)$, and all homogeneous elements $A_1,\ldots,A_n\in L$. It is known that $\check{S}_\pi$ is an automorphism of the graded L-module $\mathrm{Lgr}_n(L,V)$, for all $\pi\in\mathcal{S}_n$.

Now let $C^n(L,V)$ be the set of all ε-skew-symmetric elements of $\mathrm{Lgr}_n(L,V)$, i.e., the set of all elements $g\in\mathrm{Lgr}_n(L,V)$ such that

$$\check{S}_\pi g=\mathrm{sgn}(\pi)g, \tag{2.7}$$

for all $\pi\in\mathcal{S}_n$, where $\mathrm{sgn}(\pi)$ denotes the signum of the permutation π. We extend this definition to the case of integers $n\leqslant 0$ and set

$$C^n(L,V)=\{0\},\quad \text{if } n\leqslant -1, \tag{2.8}$$

$$C^0(L,V)=V. \tag{2.9}$$

Occasionally, it is useful to identify $C^0(L,V)$ with $\mathrm{Lgr}(\mathbb{K},V)$. Then $C^n(L,V)$ can be identified with $\mathrm{Lgr}(\wedge_\varepsilon^n L,V)$, where $\wedge_\varepsilon^n L$ is the subspace of degree n of the ε-exterior algebra $\wedge_\varepsilon L$ of L. A few comments on the ε-exterior algebra of a Γ-graded vector space are contained in Appendix A. Obviously, $C^n(L,V)$ is a graded L-module, for all n. The elements of $C^n(L,V)$ are called the *n-cochains* of L with values in V.

In the subsequent discussion we follow Ref. 3. First we define, for all integers n, a map

$$C^n(L,V)\times L\to C^{n-1}(L,V), \tag{2.10}$$

denoted by

$$(g,A)\to g_A, \tag{2.11}$$

by

$$g_A=0, \quad \text{if } n\leqslant 0, \tag{2.12}$$

$$g_A=g(A), \quad \text{if } n=1, \tag{2.13}$$

$$g_A(A_2,\ldots,A_n)=g(A,A_2,\ldots,A_n), \quad \text{if } n\geqslant 2, \tag{2.14}$$

where $g\in C^n(L,V)$ and $A,A_2,\ldots,A_n\in L$. It is easy to see that this map is bilinear, homogeneous of degree zero, and L-invariant, i.e.,

$$B\cdot g_A=(B\cdot g)_A+\varepsilon(\beta,\gamma)g_{\langle B,A\rangle}, \tag{2.15}$$

for all homogeneous elements $A,B\in L$ and $g\in C^n(L,V)$.

Next we define, again for all integers n, the linear *coboundary operator*

$$\delta^n: C^n(L,V)\to C^{n+1}(L,V), \tag{2.16}$$

as follows: We set

$$\delta^n=0, \quad \text{if } n\leqslant -1, \tag{2.17}$$

and define δ^n for $n\geqslant 0$ inductively by

$$(\delta^n(g))_A=\varepsilon(\gamma,\alpha)A\cdot g-\delta^{n-1}(g_A), \tag{2.18}$$

where $g\in C^n(L,V)$ and $A\in L$ are homogeneous. [Note that Eq. (2.18) is trivially satisfied if $n\leqslant -1$.] As it stands, Eq. (2.18) defines a linear map of $C^n(L,V)$ into $\mathrm{Lgr}_{n+1}(L,V)$ which is homogeneous of degree zero, but it is easy to see by induction on $n\geqslant 0$ that $\delta^n g$ is ε-skew-symmetric and hence an element of $C^{n+1}(L,V)$. One can now show by induction on n that δ^n is a homomorphism of graded L-modules, i.e., that

$$A\cdot(\delta^n g)=\delta^n(A\cdot g), \tag{2.19}$$

for all n and all $A\in L$, $g\in C^n(L,V)$, and then (very easily) that

$$\delta^{n+1}\circ\delta^n=0, \tag{2.20}$$

for all n. Finally, one can verify inductively that the following explicit formula holds for $n\geqslant 1$:

$$
\begin{aligned}
&(\delta^n g)(A_0,A_1,\ldots,A_n)\\
&=\sum_{r=0}^{n}(-1)^r\varepsilon(\gamma+\alpha_0+\cdots+\alpha_{r-1},\alpha_r)A_r\cdot g(A_0,\ldots,\widehat{A_r},\ldots,A_n)\\
&+\sum_{r<s}(-1)^s\varepsilon(\alpha_{r+1}+\cdots+\alpha_{s-1},\alpha_s)g(A_0,\ldots,A_{r-1},\langle A_r,A_s\rangle,A_{r+1},\ldots,\widehat{A_s},\ldots,A_n),
\end{aligned}
\tag{2.21}
$$

where $g\in C^n(L,V)$ and $A_0,A_1,\ldots,A_n\in L$ are homogeneous, and where the sign $\widehat{\ }$ indicates that the element below it must be omitted. Empty sums (like $\alpha_0+\cdots+\alpha_{r-1}$ for $r=0$ and $\alpha_{r+1}+\cdots+\alpha_{s-1}$ for $s=r+1$) are set equal to zero. In particular, for $n=1$ we have

$$
(\delta^1 g)(A_0,A_1)=\varepsilon(\gamma,\alpha_0)A_0\cdot g(A_1)-\varepsilon(\gamma+\alpha_0,\alpha_1)A_1\cdot g(A_0)-g(\langle A_0,A_1\rangle),
\tag{2.22}
$$

and for $n=2$ we obtain

$$
\begin{aligned}
(\delta^2 g)(A_0,A_1,A_2)=&\varepsilon(\gamma,\alpha_0)A_0\cdot g(A_1,A_2)-\varepsilon(\gamma+\alpha_0,\alpha_1)A_1\cdot g(A_0,A_2)\\
&+\varepsilon(\gamma+\alpha_0+\alpha_1,\alpha_2)A_2\cdot g(A_0,A_1)-g(\langle A_0,A_1\rangle,A_2)\\
&+\varepsilon(\alpha_1,\alpha_2)g(\langle A_0,A_2\rangle,A_1)+g(A_0,\langle A_1,A_2\rangle).
\end{aligned}
\tag{2.23}
$$

Occasionally, it is useful to rewrite Eq. (2.21) in the form

$$
\begin{aligned}
&2(\delta^n g)(A_0,A_1,\ldots,A_n)\\
&\quad=\sum_{r=0}^{n}(-1)^r\varepsilon(\gamma+\alpha_0+\cdots+\alpha_{r-1},\alpha_r)(A_r\cdot g)(A_0,\ldots,\widehat{A_r},\ldots,A_n)\\
&\quad+\sum_{r=0}^{n}(-1)^r\varepsilon(\gamma+\alpha_0+\cdots+\alpha_{r-1},\alpha_r)A_r\cdot g(A_0,\ldots,\widehat{A_r},\ldots,A_n).
\end{aligned}
\tag{2.24}
$$

Remark 2.1: The relations (2.18), (2.19), (2.20) are the basic properties of the coboundary operators δ^n. Of course, they can be checked directly [regarding Eq. (2.21) as the definition of δ^n], but it is much easier to proceed as in Ref. 3, i.e., to follow the steps described above.

Remark 2.2: We could use the ε-skew-symmetry of g to shift the argument $\langle A_r,A_s\rangle$ in the second sum on the right hand side of Eq. (2.21) to the first place. Then this equation would take a form which is similar to the one familiar from the nongraded theory. However, this would lead to additional ε factors which we wanted to avoid.

Next let $Z^n(L,V)$ denote the kernel of δ^n and let $B^n(L,V)$ denote the image of δ^{n-1}. Equation (2.19) implies that $Z^n(L,V)$ and $B^n(L,V)$ are graded submodules of $C^n(L,V)$, and according to Eq. (2.20) we have

$$
B^n(L,V)\subset Z^n(L,V).
\tag{2.25}
$$

The elements of $Z^n(L,V)$ are called *n-cocycles*, the elements of $B^n(L,V)$ are the *n-coboundaries.* Because of (2.25) we can construct the so-called *cohomology groups*,

$$
H^n(L,V)=Z^n(L,V)/B^n(L,V).
\tag{2.26}
$$

Two elements of $Z^n(L,V)$ are said to be *cohomologous* if their residue classes modulo $B^n(L,V)$ coincide, i.e., if their difference lies in $B^n(L,V)$.

Of course, the cohomology groups are graded L-modules, too. Actually, their L-module structure is trivial. To prove this we have to show that for any cocycle $g\in Z^n(L,V)$ and any $A\in L$, the cocycle $A\cdot g$ is a coboundary, and this follows directly from Eq. (2.18).

The following consequence of the foregoing result can be used to facilitate the calculation of the cohomology groups.

Proposition 2.1: We use the notation introduced at the beginning of the present section. Let L' *be a graded subalgebra of* L.

(a) If the L'*-module* $Z^n(L,V)$ *is semi-simple [which is true if the* L'*-module* $C^n(L,V)$ *is semi-simple], then for any cocycle* $g \in Z^n(L,V)$ *there exists an* L'*-invariant cocycle* $g' \in Z^n(L,V)$ *which is cohomologous to* g.

(b) If the L'*-module* $C^{n-1}(L,V)$ *is semi-simple, then any* L'*-invariant coboundary* $b \in B^n(L,V)$ *is equal to* $\delta^{n-1}(g)$ *with an* L'*-invariant cochain* $g \in C^{n-1}(L,V)$.

(Semi-simplicity may be understood in the graded sense.)

Proof: (a) If the graded L'-module $Z^n(L,V)$ is semi-simple, there exists a graded L'-submodule $X^n(L,V)$ of $Z^n(L,V)$ which is complementary to $B^n(L,V)$. By definition, the canonical map,

$$Z^n(L,V) \to H^n(L,V), \tag{2.27}$$

induces an isomorphism,

$$X^n(L,V) \to H^n(L,V), \tag{2.28}$$

of graded L'-modules. Since the L'-module structure of $H^n(L,V)$ is trivial, the same is true for the L'-module structure of $X^n(L,V)$, i.e., all elements of $X^n(L,V)$ are L'-invariant. Thus all we have to do is to choose an element $g' \in X^n(L,V)$ such that the canonical images of g and g' in $H^n(L,V)$ coincide.

(b) If the graded L'-module $C^{n-1}(L,V)$ is semi-simple, there exists a graded L'-submodule $Y^{n-1}(L,V)$ of $C^{n-1}(L,V)$ which is complementary to the kernel of δ^{n-1} [i.e., to $Z^{n-1}(L,V)$]. Then δ^{n-1} induces an isomorphism,

$$Y^{n-1}(L,V) \to B^n(L,V), \tag{2.29}$$

of graded L'-modules. This implies (b) and completes the proof of the proposition.

Next we consider some constructions which can be carried out with the cohomology groups. Let L' be a second ε Lie algebra and let

$$\omega : L' \to L \tag{2.30}$$

be a homomorphism of ε Lie algebras. Then the action of L' on V, defined by

$$A' \cdot x = \omega(A') \cdot x, \tag{2.31}$$

for all $A' \in L'$ and $x \in V$, makes V into a graded L'-module, which we denote by V^ω. If $g \in C^n(L,V)$ with some $n \geqslant 1$, then the map $L'^n \to V$, defined by

$$(A'_1, \ldots, A'_n) \to g(\omega(A'_1), \ldots, \omega(A'_n)), \tag{2.32}$$

for all $A'_1, \ldots, A'_n \in L'$, is an element of $C^n(L', V^\omega)$. This assignment defines a linear map,

$$C^n(L,V) \to C^n(L', V^\omega). \tag{2.33}$$

We extend this definition to the cases $n \leqslant 0$ by defining this map to be equal to id_V if $n = 0$, and to zero if $n \leqslant -1$. Obviously, the map (2.33) is homogeneous of degree zero, and it is easy to see that it is compatible with the coboundary operators, in the obvious sense. In particular, it maps cocycles into cocycles and coboundaries into coboundaries, thus inducing a corresponding map for the cohomology groups:

$$H^n(L,V) \to H^n(L', V^\omega). \tag{2.34}$$

Similarly, let

$$f : V \to W \tag{2.35}$$

be a linear map of V into a second graded L-module W and suppose that f is homogeneous of degree φ and L-invariant. Recall[20] that L-invariance means that

$$f(A\cdot x)=\varepsilon(\varphi,\alpha)A\cdot f(x), \tag{2.36}$$

for all $x\in V$, and for all $A\in L$ which are homogeneous of degree α (stated differently, f is L-linear in the graded sense). For any integer n, we define the map

$$f_c^n: C^n(L,V)\to C^n(L,W), \tag{2.37}$$

by

$$f_c^n=0, \quad \text{if } n\leqslant -1, \tag{2.38}$$

$$f_c^0=f, \quad \text{if } n=0, \tag{2.39}$$

$$f_c^n(g)=f\circ g, \quad \text{if } n\geqslant 1. \tag{2.40}$$

Obviously, f_c^n is homogeneous of degree φ, and it is easy to see that

$$\delta_W^n\circ f_c^n=f_c^{n+1}\circ\delta_V^n, \tag{2.41}$$

where, temporarily, we have marked the coboundary operators for cochains with values in V and W by the subscripts V and W, respectively. Equation (2.41) implies that f_c^n maps cocycles into cocycles and coboundaries into coboundaries, and hence induces a linear map,

$$f^n: H^n(L,V)\to H^n(L,W). \tag{2.42}$$

Of course, f^n is also homogeneous of degree φ. If U is a third graded L-module and if $f': U\to V$ is a linear map which is homogeneous and L-invariant, then, obviously,

$$(f\circ f')_c^n=f_c^n\circ f'^n_c \quad \text{and} \quad (f\circ f')^n=f^n\circ f'^n. \tag{2.43}$$

As a first application of this construction, choose $\sigma\in\Gamma$ and consider the graded L-module V^σ which is obtained from V by a shift of the gradation by σ (see Ref. 20). Recall that, considered as a vector space, V and V^σ coincide, and that the actions of L on V and V^σ are also the same, but that the gradations are related by

$$V^\sigma_\gamma=V_{\gamma+\sigma}, \tag{2.44}$$

for all $\gamma\in\Gamma$. Define the linear map

$$f: V^\sigma\to V, \tag{2.45}$$

by

$$f(x)=\varepsilon(\sigma,\xi)x, \tag{2.46}$$

for all $x\in V^\sigma_\xi$, $\xi\in\Gamma$. Then f is bijective and homogeneous of degree σ, and it is easy to see that f is L-invariant. Consequently,

$$f^n: H^n(L,V^\sigma)\to H^n(L,V) \tag{2.47}$$

is a bijective linear map which is homogeneous of degree σ. Stated differently, f^n is an isomorphism of the graded vector space $H^n(L,V^\sigma)$ onto the graded vector space $H^n(L,V)^\sigma$.

For our second application we suppose that V is the direct sum of a family $(V_i)_{i\in I}$ of graded submodules,

$$V=\bigoplus_{i\in I} V_i. \tag{2.48}$$

Considering the canonical injections $V_i \to V$ and projections $V \to V_i$, it is easy to construct a natural linear map,

$$\bigoplus_{i \in I} H^n(L,V_i) \to H^n(L,V), \tag{2.49}$$

which is injective and homogeneous of degree zero. If I is finite or if L is finite-dimensional, this map is even bijective, i.e., an isomorphism of graded vector spaces. The details are obvious and may be left to the reader.

The third application is less trivial. Consider three graded L-modules U, V, W and a short exact sequence,

$$\{0\} \to U \overset{j}{\to} V \overset{p}{\to} W \to \{0\}, \tag{2.50}$$

where the linear mappings j and p are homogeneous (of degrees ι and π, respectively) and L-invariant. Then there exists a long exact sequence,

$$\ldots \to H^n(L,U) \overset{j^n}{\to} H^n(L,V) \overset{p^n}{\to} H^n(L,W) \overset{\delta^n}{\to} H^{n+1}(L,U) \to \ldots, \tag{2.51}$$

where the maps j^n and p^n have been defined above. The definition of the so-called *connecting homomorphisms* δ^n [not to be confounded with the coboundary operators δ^n of Eq. (2.16)] is standard but a little involved (see Refs. 24,25). Since, in the applications, this definition generally is of little use, we do not give it here, but only mention that δ^n is homogeneous of degree $-\iota - \pi$.

Next we want to construct certain product maps between the spaces of cochains. Let V and W be two graded L-modules and let m and n be two integers. We are going to define a bilinear map,

$$C^m(L,V) \times C^n(L,W) \to C^{m+n}(L, V \otimes W), \tag{2.52}$$

denoted by

$$(g,h) \to g \odot h, \tag{2.53}$$

as follows. If $m<0$ or $n<0$, this map is equal to zero. Now assume that $m,n \geqslant 0$, and let $g \in C^m(L,V)$ and $h \in C^n(L,W)$ be homogeneous of degrees γ and η, respectively. If $m=n=0$, we set

$$g \odot h = g \otimes h; \tag{2.54}$$

if $m=0$, $n>0$, we set

$$(g \odot h)(A_1, \ldots, A_n) = g \otimes h(A_1, \ldots, A_n), \tag{2.55}$$

for all $A_1, \ldots, A_n \in L$; if $m>0$, $n=0$, we set

$$(g \odot h)(A_1, \ldots, A_m) = \varepsilon(\eta, \alpha_1 + \cdots + \alpha_m) g(A_1, \ldots, A_m) \otimes h, \tag{2.56}$$

for all homogeneous elements $A_1, \ldots, A_m \in L$. Finally, suppose that $m,n \geqslant 1$. In that case, we define the $(m+n)$-linear map $g \overline{\otimes} h$ (the graded tensor product of g and h) by

$$\begin{aligned}(g \overline{\otimes} h)&(A_1, \ldots, A_m, A_{m+1}, \ldots, A_{m+n}) \\ &= \varepsilon(\eta, \alpha_1 + \cdots + \alpha_m) g(A_1, \ldots, A_m) \otimes h(A_{m+1}, \ldots, A_{m+n}),\end{aligned} \tag{2.57}$$

for all homogeneous elements $A_1, \ldots, A_{m+n} \in L$ and set

$$g \odot h = \frac{1}{m!n!} \sum_{\pi \in \mathcal{S}_{m+n}} \operatorname{sgn}(\pi) \check{S}_\pi (g \overline{\otimes} h), \tag{2.58}$$

where $\check{S}_\pi$ has been defined in Eq. (2.6).

Remark 2.3: If $m,n\geqslant 1$, let $\mathcal{S}(m,n)$ be the set of all permutations $\pi\in\mathcal{S}_{m+n}$ such that π is increasing on $\{1,2,\ldots,m\}$ and on $\{m+1,m+2,\ldots,m+n\}$ (the appropriate set of shuffle permutations). Then Eq. (2.58) can also be written in the form

$$g\odot h=\sum_{\pi\in\mathcal{S}(m,n)}\operatorname{sgn}(\pi)\check{S}_\pi(g\overline{\otimes}h). \tag{2.59}$$

Obviously, in all cases $g\odot h$ is an element of $C^{m+n}(L,V\otimes W)$. Moreover, it follows from the general theory of graded L-modules[20] that the bilinear map (2.52) is homogeneous of degree zero and L-invariant, i.e., that

$$A\cdot(g\odot h)=(A\cdot g)\odot h+\varepsilon(\alpha,\gamma)g\odot(A\cdot h), \tag{2.60}$$

for all homogeneous elements $A\in L$, $g\in C^m(L,V)$, and $h\in C^n(L,W)$.

The product is also compatible with the other operations introduced above. In fact, using (2.59) it is not difficult to show that for all integers m,n and all elements $A\in L_\alpha$, $g\in C^m(L,V)$, and $h\in C^n(L,W)_\eta$,

$$(g\odot h)_A=\varepsilon(\eta,\alpha)g_A\odot h+(-1)^m g\odot h_A. \tag{2.61}$$

Using this relation and induction on $m+n$ it is easy to rederive Eq. (2.60) and then, again by induction on $m+n$, to obtain the following generalized Leibniz rule:

$$\delta^{m+n}(g\odot h)=(\delta^m g)\odot h+(-1)^m g\odot(\delta^n h), \tag{2.62}$$

which holds for all integers m,n and all $g\in C^m(L,V)$ and $h\in C^n(L,W)$. An immediate consequence of this equation is that the product maps (2.52) induce analogous ones for the cohomology groups:

$$H^m(L,V)\times H^n(L,W)\to H^{m+n}(L,V\otimes W). \tag{2.63}$$

As a final result about the product $\odot$ we notice that it is associative in the obvious sense: If U is a third graded L-module, we have for all integers ℓ,m,n,

$$f\odot(g\odot h)=(f\odot g)\odot h, \tag{2.64}$$

where $f\in C^\ell(L,U)$, $g\in C^m(L,V)$, and $h\in C^n(L,W)$ [of course, it is understood that $U\otimes(V\otimes W)$ and $(U\otimes V)\otimes W$ are canonically identified]. The proof can be carried out by induction on $\ell+m+n$.

In the foregoing discussion, the product of two cocycles $g\in C^m(L,V)$ and $h\in C^n(L,W)$ takes its values in $V\otimes W$. This is the generic case, to which the other possibilities can be reduced. If U is a third graded L-module and if

$$b:V\times W\to U \tag{2.65}$$

is a homogeneous L-invariant bilinear map, we can use the corresponding linear map,

$$\tilde{b}:V\otimes W\to U, \tag{2.66}$$

to construct the maps,

$$\tilde{b}_c^n:C^n(L,V\otimes W)\to C^n(L,U), \tag{2.67}$$

which, when composed with the product maps (2.52), yield the maps

$$C^m(L,V)\times C^n(L,W)\to C^{m+n}(L,U). \tag{2.68}$$

Of course, these have properties analogous to those of $\odot$, in particular, they induce bilinear maps,

$$H^m(L,V)\times H^n(L,W)\to H^{m+n}(L,U), \tag{2.69}$$

which are homogeneous of the same degree as b.

We close this section by showing that for certain coefficient modules V the cohomology groups $H^n(L,V)$ must vanish. Let $U(L)$ be the enveloping algebra of L and let $U_+(L)$ denote the ideal of $U(L)$ which is generated by L. We say that an element X of $U(L)$ (and, by abuse of language, also the representative X_V of X in an L-module V) *does not contain a constant term*, in case $X\in U_+(L)$.

Proposition 2.2: Let L be a finite-dimensional ε Lie algebra, let V be a graded L-module, and let C be a homogeneous element of the ε center of $U(L)$ (i.e., a homogeneous Casimir element of L). Suppose that C does not contain a constant term, and that the operator C_V (the representative of C in V) is invertible. Then we have

$$H^n(L,V)=\{0\}, \quad \text{if } n\neq 0. \tag{2.70}$$

Proof: Let $(E_i)_{i\in I}$ be a basis of L, with E_i homogeneous of degree ϵ_i. We define for all $A\in L$ and $j\in I$,

$$\langle A,E_j\rangle=\sum_{i\in I}\mathrm{ad}_{ij}(A)E_i, \tag{2.71}$$

i.e., $A\to(\mathrm{ad}_{ij}(A))$ is the adjoint representation, written in matrix form.

Let L^* be the coadjoint L-module, i.e., the graded dual of L endowed with the representation contragredient to the adjoint one, and let $(E_i')_{i\in I}$ be the basis of L^* dual to $(E_i)_{i\in I}$, defined by

$$E_i'(E_j)=\delta_{ij}, \quad \text{for all } i,j\in I. \tag{2.72}$$

Then E_i' is homogeneous of degree $-\epsilon_i$.

Now consider an L-invariant r-linear form ϕ on L^* which is homogeneous of degree η. Then

$$C(\phi)=\sum_{i_1,\ldots,i_r\in I}\phi(E_{i_1}',\ldots,E_{i_r}')E_{i_r}\ldots E_{i_1} \tag{2.73}$$

is a Casimir element of L [i.e., an element of the ε center of $U(L)$] which is homogeneous of degree η (see Ref. 26). We define the following elements $C_i(\phi)$; $i\in I$, of $U(L)$:

$$C_i(\phi)=\sum_{i_2,\ldots,i_r\in I}\phi(E_i',E_{i_2}',\ldots,E_{i_r}')E_{i_r}\ldots E_{i_2}. \tag{2.74}$$

Obviously, we have

$$C(\phi)=\sum_{i\in I}C_i(\phi)E_i; \tag{2.75}$$

moreover, $C_i(\phi)$ is homogeneous of degree $\eta-\epsilon_i$, and the L-invariance of ϕ implies that

$$\langle A,C_i(\phi)\rangle=-\sum_{j\in I}\varepsilon(\alpha,\eta-\epsilon_j)\mathrm{ad}_{ij}(A)C_j(\phi), \tag{2.76}$$

for all homogeneous elements $A\in L$ and all $i\in I$.

Now let C be the Casimir element described in the proposition, and suppose that C is homogeneous of degree η. Then it is known[26] that C can be written as a sum of Casimir elements of the form $C(\phi)$ considered above. Consequently, there exist elements $C_i\in U(L)$; $i\in I$, which are homogeneous of degree $\eta-\epsilon_i$ and satisfy

$$C=\sum_{i\in I}C_iE_i, \tag{2.77}$$

as well as

$$\langle A, C_i \rangle = -\sum_{j \in I} \varepsilon(\alpha, \eta - \epsilon_j) \mathrm{ad}_{ij}(A) C_j , \tag{2.78}$$

for all homogeneous elements $A \in L$ and all $i \in I$.

After these preparations the proof of the proposition is easy. For any integer $n \geqslant 1$ we define a linear map,

$$d_n : C^n(L,V) \to C^{n-1}(L,V) \tag{2.79}$$

by

$$(d_n(g))(A_2, \ldots, A_n) = \sum_{i \in I} \varepsilon(\epsilon_i, \gamma) C_i \cdot g(E_i, A_2, \ldots, A_n), \tag{2.80}$$

for all homogeneous elements $g \in C^n(L,V)$ and all $A_2, \ldots, A_n \in L$ (with the obvious interpretation if $n=1$). Then one can check that

$$d_{n+1} \circ \delta^n + \delta^{n-1} \circ d_n = (C_V)^n_c , \tag{2.81}$$

where C_V is the representative of C in V (see the analogous but simpler calculation in the proof of Proposition VII.5.6 in Ref. 27). Of course, C_V is homogeneous of degree η and L-invariant. Thus if $g \in C^n(L,V)$ is a cocycle, the Eqs. (2.81) and (2.41) imply that

$$g = \delta^{n-1}((C_V^{-1})^{n-1}_c d_n(g)); \tag{2.82}$$

hence g is a coboundary and the proposition is proved.

The case $n=0$ (not covered by the proposition) can easily be dealt with: Since $B^0(L,V) = \{0\}$, we have

$$H^0(L,V) = Z^0(L,V) = \{x \in V | A \cdot x = 0, \text{ for all } A \in L\}, \tag{2.83}$$

i.e., $H^0(L,V)$ is the space of all L-invariant elements of V.

Remark 2.4: The definitions and constructions of the present section (but not the two propositions) can be generalized by choosing, for the basic domain of scalars, instead of the field $\mathbb{K}$ an arbitrary associative ε-commutative graded algebra S over $\mathbb{K}$. This follows immediately from the multilinear algebra over S as sketched in the appendix of Ref. 28. A generalization of this type could be useful if one considers the deformations of a superalgebra over a Grassmann algebra, and the deformation parameter is allowed to be odd. (The algebra L may also act on S by ε-derivations; for the super case, see Ref. 6.)

III. EXAMPLES

A. Semi-simple Lie algebras

In our first example we consider a (finite-dimensional) semi-simple Lie algebra L and a finite-dimensional L-module V, and we want to describe the cohomology groups $H^n(L,V)$. Needless to say, this case is well-known; we consider it here in order to show how our propositions can be applied under the most favorable circumstances.

The main input is that all finite-dimensional L-modules are semi-simple (i.e., the corresponding representations are completely reducible). Because of the isomorphism (2.49) we may, therefore, assume that the L-module V is simple (of course, in practice it may be difficult to decompose an L-module into simple submodules). If V is nontrivial, it is well-known that there exists a quadratic Casimir element C of L such that $C_V \neq 0$. Since V is simple, it follows that C_V is invertible. Thus Proposition 2.2 implies that $H^n(L,V) = \{0\}$ if $n \neq 0$, and the same is true for $n = 0$ since 0 is the sole L-invariant element of V.

Thus we may now assume that V is the trivial L-module $\mathbb{K}$ and that $n \geqslant 0$. According to (a) of Proposition 2.1 every cohomology class contains an L-invariant element. Conversely, it follows

from Eq. (2.24) that every L-invariant cochain (with values in $\mathbb{K}$) is a cocycle. Moreover, if g and g' are cohomologous elements of $Z^n(L,\mathbb{K})$ and if g and g' are L-invariant, then (b) of Proposition 2.1 tells us that $g-g'=\delta^{n-1}b$ with an L-invariant cochain $b\in C^{n-1}(L,\mathbb{K})$. As we just have seen, this implies that $\delta^{n-1}b=0$, i.e., that $g=g'$.

Summarizing, we have proved that $H^n(L,\mathbb{K})$ *with* $n\geqslant 1$ *can be identified with the space of all* L*-invariant skew symmetric* n*-linear forms on* L.

The cohomology groups $H^n(L,\mathbb{K})$ with $n\in\{0,1,2\}$ can easily be determined. Obviously, we have

$$H^0(L,\mathbb{K})=\mathbb{K} \tag{3.1}$$

[see Eq. (2.83)]. Let us next recall that the commutator algebra $[L,L]$ of L is equal to L. In view of Eq. (2.22), this implies that $Z^1(L,\mathbb{K})=\{0\}$ and hence

$$H^1(L,\mathbb{K})=\{0\}. \tag{3.2}$$

Moreover, it also follows from $[L,L]=L$ that every L-invariant bilinear form on L is symmetric. This shows that

$$H^2(L,\mathbb{K})=\{0\}. \tag{3.3}$$

On the other hand, it is easy to see that

$$H^3(L,\mathbb{K})\neq\{0\}. \tag{3.4}$$

In fact, if ϕ is the Killing form of L, the assignment

$$(A,B,C)\to\phi([A,B],C) \tag{3.5}$$

(with $A,B,C\in L$) defines a nonzero, L-invariant, skew symmetric, trilinear form on L.

B. The Lie superalgebras $osp(1|2n)$

For (finite-dimensional) simple Lie superalgebras the situation is much more complicated, even for the basic classical ones. However, there is one particular class of algebras (already mentioned by Kac in Sec. 5.5.3 of Ref. 7), namely, the $osp(1|2n)$ algebras, for which the discussion of the preceding example goes through almost verbatim. All we have to do is to interpret it in the super sense and to recall the bijection (2.47). The pertinent information on the $osp(1|2n)$ algebras is contained in Ref. 29. As already mentioned, the cohomology groups $H^n(L,\mathbb{K})$ of the nonexceptional classical simple Lie superalgebras have been described in Ref. 9 (see also Chap. 2, § 6 of Ref. 10 and Ref. 14; in particular, the exceptional algebras are considered in the latter article).

C. Trivial coefficients

Let us return to the general assumptions (where L is an arbitrary ε Lie algebra), and let us consider the case where $V=\mathbb{K}$ is the trivial graded L-module. As noted in Example A, Eq. (2.24) implies that every L-invariant cochain is a cocycle. We would like to mention a classical construction of such cocycles (which, in the case of simple Lie algebras, essentially does the whole job; see Ref. 30 for more details).

Let $m\geqslant 0$ be an integer and let $\phi\in \mathrm{Lgr}_{m+1}(L,\mathbb{K})$ be an L-invariant $(m+1)$-linear form. We want to construct an L-invariant cochain out of ϕ. An obvious ansatz is to consider the ε-skew-symmetrization of the $2(m+1)$-linear form on L defined by

$$(A_0,A_1,\ldots,A_{2m+1})\to\phi(\langle A_0,A_1\rangle,\langle A_2,A_3\rangle,\ldots,\langle A_{2m},A_{2m+1}\rangle), \tag{3.6}$$

for all $A_i\in L$. However, the L-invariance of ϕ and the ε Jacobi identity imply, that the ε-skew-symmetrization of the form (3.6) is equal to zero. The well-known way out is to consider for $m\geqslant 1$ the ε-skew-symmetrization of the $(2m+1)$-linear form $\psi\in \mathrm{Lgr}_{2m+1}(L,\mathbb{K})$ defined by

$$\psi(A_0,A_1,\ldots,A_{2m})=\phi(\langle A_0,A_1\rangle,\langle A_2,A_3\rangle,\ldots,\langle A_{2m-2},A_{2m-1}\rangle,A_{2m}), \tag{3.7}$$

for all $A_i\in L$.

The situation simplifies a little if ϕ is ε-symmetric (and L-invariant). In that case it is sufficient to ε-skew-symmetrize ψ in $A_1,A_2,\ldots,A_{2m-1}$, i.e., already the $(2m+1)$-linear form on L defined by

$$(A_0,A_1,\ldots,A_{2m})\to\sum_{\pi\in\mathcal{S}_{2m-1}}\operatorname{sgn}(\pi)\varepsilon_{2m-1}(\pi;\alpha_1,\ldots,\alpha_{2m-1}) \tag{3.8}$$

$$\phi(\langle A_0,A_{\pi(1)}\rangle,\langle A_{\pi(2)},A_{\pi(3)}\rangle,\ldots,\langle A_{\pi(2m-2)},A_{\pi(2m-1)}\rangle,A_{2m})$$

(for homogeneous elements $A_i\in L$) is ε-skew-symmetric and, of course, L-invariant. Note that the trilinear form (3.5) is the simplest example of the foregoing construction.

The possibility to obtain L-invariant cochains from L-invariant ε-symmetric multilinear forms is welcome and interesting, since the latter forms make their appearance in quite a different context. In fact, if L is finite-dimensional and admits a nondegenerate L-invariant bilinear form which is homogeneous of degree zero, then there is a close relationship between the Casimir elements of L and the ε-symmetric L-invariant multilinear forms on L. For more details, we refer the reader to Ref. 26. However, we note that we do not know of any reason to hope that the situation might be as nice as for semi-simple Lie algebras.

D. Simple $gl(m|n)$-modules for which all Casimir operators are equal to zero

In this example we assume that the field $\mathbb{K}$ is *algebraically closed.* We want to determine all those finite-dimensional simple graded $gl(m|n)$-modules, for which all Casimir operators without a constant term are equal to zero. According to Proposition 2.2, among the finite-dimensional simple graded modules, only these might have a nontrivial cohomology. In view of the detailed knowledge of the Casimir elements of $gl(m|n)$ and of their eigenvalues in a highest weight module[31,26,32,33] (see also Refs. 34–36), it is not difficult to solve this problem.

We are going to use the notation and results of the cited references. To begin with, let us recall some of them. Let m and n be any strictly positive integers. The Lie superalgebra $gl(m|n)$ is defined through its elementary representation in $\mathbb{K}^{m+n}$. If $(e_i)_{1\leqslant i\leqslant m+n}$ is the canonical basis of $\mathbb{K}^{m+n}$, we define the gradation of $\mathbb{K}^{m+n}$ by demanding that the elements $e_1,\ldots,e_m$ are even and the elements $e_{m+1},\ldots,e_{m+n}$ are odd. Correspondingly, we define

$$\sigma_i=\begin{cases}1, & \text{if } 1\leqslant i\leqslant m,\\ -1, & \text{if } m+1\leqslant i\leqslant m+n.\end{cases} \tag{3.9}$$

In Ref. 33 certain generators X_{ij} of the Lie superalgebra under consideration have been introduced. For $gl(m|n)$, these are just the usual basic $(m+n)\times(m+n)$ matrices E_{ij}; $1\leqslant i,j\leqslant m+n$.

Now let $s\geqslant 1$ be an integer and let C_s be the element of the universal enveloping algebra $U(gl(m|n))$ defined by

$$C_s=\sum_{i_1,\ldots,i_s}\sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_{s-1}}X_{i_s i_1}X_{i_1 i_2}\ldots X_{i_{s-1}i_s}\,. \tag{3.10}$$

Then C_s is a Casimir element of $gl(m|n)$, and (as shown in Ref. 32) the supercenter $Z(gl(m|n))$ of $U(gl(m|n))$ is generated (as an algebra) by the unit element and these elements C_s. Stated differently, every Casimir element can be written as a polynomial in the C_s; $s\geqslant 1$. [By the way, none of the C_s may be dropped. In particular, the algebra $Z(gl(m|n))$ is not finitely generated.]

The elements X_{ii}; $1\leqslant i\leqslant m+n$ span the standard Cartan subalgebra H of $gl(m|n)$. Let ε_i; $1\leqslant i\leqslant m+n$ be the linear form on H which satisfies

$$\varepsilon_i(X_{jj})=\delta_{ij}\,, \tag{3.11}$$

for $1 \leqslant j \leqslant m+n$. We choose the usual system of positive roots of $gl(m|n)$ with respect to H, namely,

$$\Delta^+ = \{\varepsilon_i - \varepsilon_j | 1 \leqslant i < j \leqslant m+n\}. \tag{3.12}$$

Now let $\Lambda \in H^*$ be any linear form on H and let V be a graded highest weight $gl(m|n)$-module with highest weight Λ. If $C \in Z(gl(m|n))$ is any Casimir element, the corresponding Casimir operator C_V is a scalar multiple of id_V,

$$C_V = \chi_\Lambda(C) id_V, \tag{3.13}$$

where χ_Λ is the so-called central character associated to Λ. To solve our problem, we need the eigenvalues $\chi_\Lambda(C)$. According to the foregoing remarks, it is sufficient to know the $\chi_\Lambda(C_s)$; $s \geqslant 1$. These have been investigated in Ref. 33. To describe the pertinent result, we introduce some more notations.

Let 2ρ be the sum of the even positive roots minus the sum of the odd positive roots, i.e.,

$$\rho = \frac{1}{2} \sum_{i<j} \sigma_i \sigma_j (\varepsilon_i - \varepsilon_j). \tag{3.14}$$

We set

$$\lambda = \Lambda + \rho, \tag{3.15}$$

and define the numbers r_i and ℓ_i through

$$r_i = \sigma_i \rho(X_{ii}), \tag{3.16}$$

$$\ell_i = \sigma_i \lambda(X_{ii}). \tag{3.17}$$

If $C \in Z(gl(m|n))$ is fixed, we have

$$\chi_\Lambda(C) = P_C(\ell_1, \ldots, \ell_{m+n}), \tag{3.18}$$

where P_C is a polynomial in $m+n$ indeterminates $Y_1, \ldots, Y_{m+n}$. If C runs through all of $Z(gl(m|n))$, the polynomials P_C form an algebra $T(m|n)$ (the algebra of supersymmetric polynomials in super dimension $(m|n)$;[31] for a recent reference see Ref. 37), and the map of $Z(gl(m|n))$ onto $T(m|n)$ defined by $C \to P_C$ is an algebra isomorphism (the generalized Harish–Chandra isomorphism[26]). Finally, the algebra $T(m|n)$ is generated by the unit element and the polynomials

$$Q_s = \sum_{i=1}^{m+n} \sigma_i (Y_i^s - r_i^s), \tag{3.19}$$

with $s \geqslant 1$ [see the end of Sec. 2 in Ref. 33, in particular, the lines below Eq. (2.36) and Eq. (2.37)]. Note that the constant term of a Casimir element C is equal to $\chi_0(C)$. Consequently, this term is equal to zero if and only if P_C can be written as a polynomial in the Q_s without a constant term.

Next we recall that every finite-dimensional simple graded $gl(m|n)$-module is a highest weight module (this is the sole place where we use that the base field is algebraically closed). Summarizing, we conclude that our problem can be reformulated as follows.

Determine all linear forms $\Lambda \in H^*$ which correspond to finite-dimensional simple graded $gl(m|n)$-modules and satisfy

$$Q_s(\ell_1, \ldots, \ell_{m+n}) = 0, \tag{3.20}$$

for all $s \geqslant 1$.

Let us set

$$L_i=\Lambda(X_{ii}), \tag{3.21}$$

for $1\leqslant i\leqslant m+n$. [Note that in contrast to Eqs. (3.16), (3.17) we have not included a factor of σ_i on the right hand side.] Then the first condition is satisfied if and only if

$$L_i-L_{i+1}\in\mathbb{N}, \tag{3.22}$$

for $1\leqslant i\leqslant m-1$ and for $m+1\leqslant i\leqslant m+n-1$.

The validity of the relations (3.20) is easily seen to be equivalent to the following condition.

Up to the ordering, the numbers,

$$\ell_1,\ell_2,\ldots,\ell_m,r_{m+1},r_{m+2},\ldots,r_{m+n}, \tag{3.23}$$

coincide with the numbers

$$r_1,r_2,\ldots,r_m,\ell_{m+1},\ell_{m+2},\ldots,\ell_{m+n}. \tag{3.24}$$

The subsequent discussion depends on whether $m=n$, $m>n$, or $m<n$. Using Eq. (3.22) and some special properties of the numbers r_i, we first conclude that the conditions (3.20) are satisfied if and only if the following holds true.

If $m=n$, we have

$$\ell_{m+i}=\ell_{m+1-i}, \quad \text{for } 1\leqslant i\leqslant m; \tag{3.25}$$

if $m>n$, there exists a number $k\in\{0,1,\ldots,n\}$ such that

$$\ell_{m+i}=\ell_{m+1-i}, \quad \text{for } 1\leqslant i\leqslant k, \tag{3.26}$$

$$r_{j-n}=\ell_{j-k}, \quad \text{for } n+1\leqslant j\leqslant m, \tag{3.27}$$

$$\ell_{m+i}=\ell_{n+1-i}, \quad \text{for } k+1\leqslant i\leqslant n; \tag{3.28}$$

and if $m<n$, there exists a number $h\in\{0,1,\ldots,m\}$ such that

$$\ell_{m+i}=\ell_{m+1-i}, \quad \text{for } 1\leqslant i\leqslant h, \tag{3.29}$$

$$\ell_{h+j}=r_{m+j}, \quad \text{for } m+1\leqslant j\leqslant n, \tag{3.30}$$

$$\ell_{n+i}=\ell_{m+1-i}, \quad \text{for } h+1\leqslant i\leqslant m. \tag{3.31}$$

Expressed in terms of the coefficients L_i, our final result reads as follows.

Suppose first that $m=n$. Choose the numbers $L_1,L_2,\ldots,L_m$ in accordance with Eq. (3.22) and define

$$L_{m+i}=-L_{m+1-i}, \quad \text{for } 1\leqslant i\leqslant m. \tag{3.32}$$

Suppose next that $m>n$. Let k be any element of $\{0,1,\ldots,n\}$. Choose the numbers $L_1,L_2,\ldots,L_m$ in accordance with Eq. (3.22) and such that

$$L_{n+1-k}=L_{n+2-k}=\cdots=L_{m-k}=n-k. \tag{3.33}$$

Define

$$L_{m+i}=-L_{m+1-i}, \quad \text{for } 1\leqslant i\leqslant k, \tag{3.34}$$

$$L_{m+i}=-L_{n+1-i}-(m-n), \quad \text{for } k+1\leqslant i\leqslant n \tag{3.35}$$

(for $k=0$, resp., $k=n$, the first, resp., second of these equations should be dropped).

Finally, suppose that $m<n$. Let h be any element of $\{0,1,\ldots,m\}$. Choose the numbers $L_{m+1},L_{m+2},\ldots,L_{m+n}$ in accordance with Eq. (3.22) and such that

$$L_{m+1+h}=L_{m+2+h}=\cdots=L_{n+h}=-(m-h). \tag{3.36}$$

Define

$$L_i=-L_{m+n+1-i}+(n-m), \quad \text{for } 1\leqslant i\leqslant m-h, \tag{3.37}$$

$$L_i=-L_{2m+1-i}, \quad \text{for } m-h+1\leqslant i\leqslant m \tag{3.38}$$

(for $h=m$, resp., $h=0$, the first, resp., second of these equations should be dropped).

Then, in all three cases, our conditions are satisfied. Conversely, every solution to our conditions is of this form.

We note that the relation (3.20) with $s=1$ is equivalent to

$$\sum_{i=1}^{m+n} L_i=0, \tag{3.39}$$

and it is easily checked that the numbers L_i specified above satisfy this condition.

Remark 3.1: Needless to say the modules we have found are atypical.[35] Actually, Eqs. (3.25) – (3.31) show that they are maximally atypical in the sense that Λ satisfies the maximal number [equal to min(m,n)] of atypicality conditions.

The foregoing results can immediately be extended to the special linear Lie superalgebras $sl(m|n)$, provided that $m\neq n$. In fact, all we have to do is interpret them appropriately, as explained in Ref. 33. The elements X_{ij} are now given by

$$X_{ij}=E_{ij}-\frac{1}{d}\sigma_i\delta_{ij}I, \tag{3.40}$$

where we have set $d=m-n$, and where I is the unit matrix. Furthermore, the linear forms ε_i are given by

$$\varepsilon_i(X_{jj})=\delta_{ij}-\frac{1}{d}\sigma_j. \tag{3.41}$$

This implies that

$$\sum_{i=1}^{m+n} X_{ii}=0 \tag{3.42}$$

and

$$\sum_{i=1}^{m+n} \sigma_i\varepsilon_i=0. \tag{3.43}$$

Apart from these modifications, everything goes through, and the highest weights of the modules we are looking for are given by the same formulas as above. [We have already noted that the relation (3.39), which should be satisfied because of Eq. (3.42), indeed holds true.]

IV. CENTRAL EXTENSIONS OF ε LIE ALGEBRAS

We start with the basic definitions. Let L be an ε Lie algebra. A *central extension* of L by an ε Lie algebra H is an exact sequence of ε Lie algebras,

$$\{0\}\rightarrow H\overset{\iota}{\rightarrow}E\overset{\pi}{\rightarrow}L\rightarrow\{0\}, \tag{4.1}$$

such that $\iota(H)$ is in the ε center of E. Thus ι and π are homomorphisms of ε Lie algebras (which implies that they are homogeneous of degree zero), ι is injective, π is surjective, and the kernel of π is equal to the image of ι. In particular, H is Abelian. Thus H is nothing but a graded vector space, endowed with the trivial structure of an ε Lie algebra. If

$$\{0\}\to H'\overset{\iota'}{\to}E'\overset{\pi'}{\to}L\to\{0\} \tag{4.2}$$

is a second central extension of L, then by a *morphism* from the central extension (4.1) to the central extension (4.2), we mean a pair (φ,ψ) of ε Lie algebra homomorphisms,

$$\varphi:E\to E',\quad \psi:H\to H', \tag{4.3}$$

such that

$$\pi=\pi'\circ\varphi,\quad \varphi\circ\iota=\iota'\circ\psi, \tag{4.4}$$

i.e., such that the corresponding diagram is commutative. (Obviously, ψ is uniquely determined by φ.) The morphism (φ,ψ) is said to be an *isomorphism* of the central extensions if φ and ψ are bijective. Actually, if one of the maps φ and ψ is bijective, then so is the other.

Consider next the case where the second central extension has the form

$$\{0\}\to H\overset{\iota'}{\to}E'\overset{\pi'}{\to}L\to\{0\}. \tag{4.5}$$

Then the extensions (4.1) and (4.5) are said to be *equivalent* if there exists a morphism (φ,ψ) from the extension (4.1) to the extension (4.5) such that $\psi=id_H$ (which implies that φ is bijective).

As in the Lie algebra case, the central extensions of L by H can be classified by means of the cohomology groups $H^2(L,H)$, where H is regarded as a graded vector space, endowed with the trivial L-module structure. To show this, we choose a homogeneous *section* σ (in the category of vector spaces) of π, i.e., a homogeneous linear map,

$$\sigma:L\to E \tag{4.6}$$

(necessarily of degree zero) such that

$$\pi\circ\sigma=id_L. \tag{4.7}$$

Then we have

$$\pi(\langle\sigma(A),\sigma(B)\rangle-\sigma(\langle A,B\rangle))=0, \tag{4.8}$$

for all $A,B\in L$. Consequently, there exists a unique map,

$$g:L\times L\to H, \tag{4.9}$$

such that

$$\iota(g(A,B))=\langle\sigma(A),\sigma(B)\rangle-\sigma(\langle A,B\rangle), \tag{4.10}$$

for all $A,B\in L$, and it is easy to see that g is a homogeneous 2-cocycle of degree zero, i.e., $g\in Z^2(L,H)_0$.

Consider in addition the central extension (4.5), choose a homogeneous section σ' of π', and construct the corresponding 2-cocycle g'. Then the central extensions (4.1) and (4.5) are equivalent if and only if g and g' are cohomologous.

Conversely, let g be any element of $Z^2(L,H)_0$. Make the graded vector space $L\times H$ into an ε Lie algebra by defining

$$\langle(A,X),(B,Y)\rangle=(\langle A,B\rangle,g(A,B)), \tag{4.11}$$

for all $A,B\in L$ and all $X,Y\in H$. This ε Lie algebra will be denoted by $L(g)$. Let

$$\sigma:L\to L\times H,\quad \iota:H\to L\times H \tag{4.12}$$

be the canonical injections and let

$$\pi:L\times H\to L \tag{4.13}$$

be the canonical projection. Then

$$\{0\}\to H\overset{\iota}{\to}L(g)\overset{\pi}{\to}L\to\{0\} \tag{4.14}$$

is a central extension of L by H, σ is a homogeneous section of π, and the corresponding 2-cocycle of L with values in H is just the 2-cocycle g we started with.

Summarizing, we conclude that there is a bijection between the set of equivalence classes of central extensions of L by H and $H^2(L,H)_0$, the homogeneous component of degree zero of the cohomology group $H^2(L,H)$. [More generally, an analogous result holds for all those extensions (4.1) of L for which H is only assumed to be Abelian; for more details, see Ref. 5.]

In the Lie algebra case, all this is well-known, and it is not at all surprising that this material can be generalized to the present graded setting. We have included these results in order to fix our notation and as a preparation for the investigations to follow. In fact, we now are ready to generalize Garland's theory of universal central extensions of Lie algebras.[21] As we are going to see, most of Garland's statements and proofs can be adopted almost verbatim. In addition, we comment on the *homological* background of these results, and we show how universal central extensions can be constructed.

Let us agree to call an ε Lie algebra L *perfect* if it is equal to its own commutator algebra, i.e., if $\langle L,L\rangle=L$. Using this notion, we call the central extension (4.1) a *covering* of L if E is perfect. In this case, we also call E or π or the pair (E,π) a covering of L, and we shall say that E [or π or (E,π)] covers L. Obviously, if an ε Lie algebra admits a covering, then it is perfect.

Lemma 4.1: If the central extension (4.1) is a covering of L, then there exists at most one morphism from (4.1) to any second central extension of L.

Proof: Let (φ,ψ) and (φ',ψ') be two morphisms from the central extension (4.1) to the central extension (4.2). Obviously, we have

$$(\varphi-\varphi')(\langle A,B\rangle)=\langle\varphi(A)-\varphi'(A),\varphi(B)\rangle+\langle\varphi'(A),\varphi(B)-\varphi'(B)\rangle, \tag{4.15}$$

for all $A,B\in E$. On the other hand, we have

$$\pi'(\varphi(A)-\varphi'(A))=\pi(A)-\pi(A)=0, \tag{4.16}$$

and hence

$$\varphi(A)-\varphi'(A)\in\iota'(H'), \tag{4.17}$$

for all $A\in E$. Consequently, the right hand side of Eq. (4.15) vanishes, which implies that $\varphi=\varphi'$, and hence also $\psi=\psi'$.

Definition 4.1: We say that a covering of an ε Lie algebra L is universal, *if for every central extension of L there is a unique morphism (in the sense of central extensions) from the covering to the central extension.*

We note that the uniqueness requirement in Definition 4.1 automatically follows from Lemma 4.1. Obviously, the definition and Lemma 4.1 imply the following proposition.

Proposition 4.2: Any two universal coverings of an ε Lie algebra L are isomorphic as central extensions, and the two mutually inverse isomorphisms are unique.

We have already noted that if an ε Lie algebra admits a covering, then it is perfect. Conversely, we have the following.

Proposition 4.3: If the ε Lie algebra L is perfect, then L has a universal covering.

Proof: Let $\wedge_\varepsilon L$ be the ε exterior algebra of the graded vector space L (see Appendix A for more information), let $\wedge_\varepsilon$ denote the multiplication in $\wedge_\varepsilon L$, and let $W'=\wedge_\varepsilon^2 L$ be the subspace of $\wedge_\varepsilon L$ of $\mathbb{Z}$-degree 2. Consider the subspace I of W' which is generated by all elements of the form

$$-\langle A,B\rangle\wedge_\varepsilon C+\varepsilon(\beta,\gamma)\langle A,C\rangle\wedge_\varepsilon B+A\wedge_\varepsilon\langle B,C\rangle, \tag{4.18}$$

with homogeneous elements $A,B,C\in L$. Obviously, I is a graded subspace of W'. We set $W=W'/I$ and denote the canonical image of $A\wedge_\varepsilon B$ in W by $f(A,B)$, for all $A,B\in L$. By definition,

$$f:L\times L\to W \tag{4.19}$$

is a 2-cocycle which is homogeneous of degree zero. Thus we have the central extension

$$\{0\}\to W\xrightarrow{\bar{\iota}}L(f)\xrightarrow{\bar{\pi}}L\to\{0\}. \tag{4.20}$$

Consider an arbitrary central extension of L by an Abelian ε Lie algebra H. We know that this extension is equivalent to the extension (4.14) with a suitable 2-cocycle $g\in Z^2(L,H)_0$. Since g is a 2-cocycle, there exists a unique linear map

$$\psi':W\to H, \tag{4.21}$$

such that

$$\psi'(f(A,B))=g(A,B), \tag{4.22}$$

for all $A,B\in L$, and ψ' is homogeneous of degree zero. We define a map

$$\varphi':L(f)=L\times W\to L\times H=L(g) \tag{4.23}$$

by

$$\varphi'(A,X)=(A,\psi'(X)), \tag{4.24}$$

for all $A\in L$ and $X\in W$. Then (φ',ψ') is a morphism from the central extension (4.20) to the central extension (4.14).

Now let $\hat{L}=\langle L(f),L(f)\rangle$ be the commutator algebra of $L(f)$. Since L is perfect, we have $\hat{L}+W=L(f)$ and hence

$$\bar{\pi}(\hat{L})=L, \tag{4.25}$$

$$\hat{L}=\langle\hat{L}+W,\hat{L}+W\rangle=\langle\hat{L},\hat{L}\rangle. \tag{4.26}$$

Thus $\hat{L}$ is perfect. If $\hat{H}$ is the preimage of $\hat{L}$ under $\bar{\iota}$, the central extension (4.20) induces a central extension,

$$\{0\}\to\hat{H}\xrightarrow{\hat{\iota}}\hat{L}\xrightarrow{\hat{\pi}}L\to\{0\}, \tag{4.27}$$

which is a covering of L. Furthermore, if φ is the restriction of φ' to $\hat{L}$ and if ψ is the restriction of ψ' to $\hat{H}$, the pair (φ,ψ) is a morphism from the central extension (4.27) to the central extension (4.14). This shows that (4.27) is a universal covering of L, and thus the proposition is proved.

In the present work we do not consider the homology of ε Lie algebras (it is fairly obvious how this should be introduced). Nevertheless, it would be rather unsatisfactory not to mention the homological interpretation of the preceding proof.

Consider the linear map,

$$\partial_2 : \overset{2}{\wedge}_\varepsilon L \to L, \tag{4.28}$$

given by

$$\partial_2(A \wedge_\varepsilon B) = -\langle A, B\rangle, \tag{4.29}$$

and the linear map,

$$\partial_3 : \overset{3}{\wedge}_\varepsilon L \to \overset{2}{\wedge}_\varepsilon L, \tag{4.30}$$

specified by

$$\partial_3(A \wedge_\varepsilon B \wedge_\varepsilon C) = -\langle A, B\rangle \wedge_\varepsilon C + \varepsilon(\beta, \gamma)\langle A, C\rangle \wedge_\varepsilon B + A \wedge_\varepsilon \langle B, C\rangle, \tag{4.31}$$

for all homogeneous elements $A, B, C \in L$. The ε Jacobi identity is equivalent to

$$\partial_2 \circ \partial_3 = 0. \tag{4.32}$$

Let $Z_2(L)$ be the kernel of ∂_2 and let $B_2(L)$ be the image of ∂_3. Then Eq. (4.32) implies that $B_2(L) \subset Z_2(L)$, and a closer look at the definition of $\hat{H}$ in the proof of Proposition 4.3 shows that

$$\hat{H} = Z_2(L)/B_2(L) = H_2(L) \tag{4.33}$$

(where the last equation is a definition). By definition, $C^n(L, \mathbb{K})$ is the graded dual of $\wedge_\varepsilon^n L$ [see the remark below Eq. (2.9)], and the coboundary map,

$$\delta^n : C^n(L, \mathbb{K}) \to C^{n+1}(L, \mathbb{K}), \tag{4.34}$$

is the graded transpose of the boundary map,

$$\partial_{n+1} : \overset{n+1}{\wedge}_\varepsilon L \to \overset{n}{\wedge}_\varepsilon L \tag{4.35}$$

(we are considering the cases $n = 1,2$ only, but, in fact, this is true for all n). Consequently, the graded dual $\hat{H}^{*\mathrm{gr}}$ of $\hat{H}$ is canonically isomorphic to $H^2(L, \mathbb{K})$.

More explicitly, the canonical isomorphism $\phi : H^2(L, \mathbb{K}) \to \hat{H}^{*\mathrm{gr}}$ can be described as follows. Let $g \in Z^2(L, \mathbb{K})_\gamma$ be a cocycle which is homogeneous of degree γ. We have to give the image of the cohomology class $g + B^2(L, \mathbb{K})$ in $\hat{H}^{*\mathrm{gr}}$. The cocycle g can be regarded as an element $g_0 \in Z^2(L, \mathbb{K}^\gamma)_0$ [recall Eq. (2.44)]. Let

$$\{0\} \to \mathbb{K}^\gamma \to E \to L \to \{0\} \tag{4.36}$$

be a central extension of L which belongs to the equivalence class corresponding to $g_0 + B^2(L, \mathbb{K}^\gamma)$, and let (φ_g, ψ_g) be the morphism from the universal covering (4.27) to (4.36). Then ψ_g can be regarded as a linear form $\hat{H} \to \mathbb{K}$ which is homogeneous of degree γ. This form depends only on the cohomology class $g + B^2(L, \mathbb{K})$ and is the image we are looking for.

The inverse of the isomorphism ϕ can be described even more simply. Let $\hat{g} \in Z^2(L, \hat{H})_0$ be a cocycle such that the cohomology class $\hat{g} + B^2(L, \hat{H})$ corresponds to the equivalence class of the universal covering (4.27). If λ is any element of $\hat{H}^{*\mathrm{gr}}$, then $\lambda \circ \hat{g}$ is an element of $Z^2(L, \mathbb{K})$ and we have

$$\phi^{-1}(\lambda) = \lambda \circ \hat{g} + B^2(L, \mathbb{K}). \tag{4.37}$$

If L is a perfect ε Lie algebra for which $H^2(L,\mathbb{K})$ is finite-dimensional (of dimension s, say), the preceding discussion leads to the following construction of a universal covering of L. Let $(g_r)_{1\leqslant r\leqslant s}$ be a family of homogeneous 2-cochains $g_r\in Z^2(L,\mathbb{K})$, such that the cohomology classes $g_r+B^2(L,\mathbb{K})$ form a basis of $H^2(L,\mathbb{K})$, and let γ_r denote the degree of g_r. Choose a graded vector space H such that H has a homogeneous basis $(e_r)_{1\leqslant r\leqslant s}$, with e_r homogeneous of degree $-\gamma_r$. Define the bilinear map,

$$g:L\times L\to H, \tag{4.38}$$

by

$$g(A,B)=\sum_{r=1}^{s} g_r(A,B)e_r, \tag{4.39}$$

for all $A,B\in L$. Then $g\in Z^2(L,H)_0$, and the corresponding central extension (4.14) is a universal covering of L.

Remark 4.1: Let

$$\{0\}\to\hat{H}\overset{\hat{\iota}}{\to}\hat{L}\overset{\hat{\pi}}{\to}L\to\{0\} \tag{4.40}$$

be a universal covering of the ε Lie algebra L. Then $\hat{L}$ is a universal covering of every ε Lie algebra E which covers L. In fact, if

$$\{0\}\to H\to E\to L\to\{0\} \tag{4.41}$$

is a covering of L and if (φ,ψ) is a morphism from (4.40) to (4.41), then it is easy to see that

$$\{0\}\to\ker(\varphi)\to\hat{L}\overset{\varphi}{\to}E\to\{0\} \tag{4.42}$$

is a covering of E. One can then show directly that this covering is universal, but the simplest way to prove this is to invoke Theorem 4.1.

Remark 4.2: The universal coverings have a property which, apparently, is more general than the one required in their definition. Let

$$\{0\}\to H\overset{\iota}{\to}E\overset{\pi}{\to}L\to\{0\} \tag{4.43}$$

be a universal covering of an ε Lie algebra L and let

$$\{0\}\to H'\overset{\iota'}{\to}E'\overset{\pi'}{\to}L'\to\{0\} \tag{4.44}$$

be a central extension of an ε Lie algebra L'. If $\varrho:L\to L'$ is a homomorphism (of ε Lie algebras), there exists a unique pair (φ,ψ) of homomorphisms,

$$\varphi:E\to E',\quad \psi:H\to H', \tag{4.45}$$

such that

$$\varrho\circ\pi=\pi'\circ\varphi,\quad \varphi\circ\iota=\iota'\circ\psi, \tag{4.46}$$

i.e., such that the corresponding diagram is commutative.

Remark 4.3: One might think to generalize the definition of a universal covering to all (not necessarily perfect) ε Lie algebras, as follows: A central extension (4.1) of an ε Lie algebra L is said to be universal if, for any central extension (4.2) of L, there exists a unique morphism from (4.1) to (4.2). Actually, it can be shown that if (4.1) is universal in this sense, then L and E are perfect, and nothing beyond the universal coverings has been obtained.

Definition 4.2: (a) An ε Lie algebra L is said to be simply connected, *if for every central extension*

$$\{0\}\to H\overset{\iota}{\to}E\overset{\pi}{\to}L\to\{0\} \tag{4.47}$$

of L, there is a unique homomorphism of ε Lie algebras $\sigma:L\to E$, such that $\pi\circ\sigma=id_L$.

(b) A covering,

$$\{0\}\to H\to E\to L\to\{0\}, \tag{4.48}$$

of an ε Lie algebra L is said to be simply connected, in case E is simply connected.

Every simply connected ε Lie algebra L is perfect. In fact, choose $H=L/\langle L,L\rangle$ and consider the trivial central extension,

$$\{0\}\to H\to L\times H\to L\to\{0\} \tag{4.49}$$

of L, where $H\to L\times H$ is the canonical injection and $L\times H\to L$ is the canonical projection. Of course, the canonical injection $L\to L\times H$ is a homomorphism σ of the type described in (a) of Definition 4.2. But if $\varphi:L\to L/\langle L,L\rangle$ is the canonical homomorphism, the map

$$\sigma':L\to L\times H \tag{4.50}$$

defined by

$$\sigma'(A)=(A,\varphi(A)), \tag{4.51}$$

for all $A\in L$ is likewise. Consequently, the uniqueness of σ implies our claim.

Theorem 4.1: *A covering of an ε Lie algebra L is universal if and only if it is simply connected.*

Proof: Let

$$\{0\}\to\hat{H}\overset{\hat{\iota}}{\to}\hat{L}\overset{\hat{\pi}}{\to}L\to\{0\} \tag{4.52}$$

be a simply connected covering of L. We have to show that there exists a (necessarily unique) morphism from this covering to an arbitrary central extension (4.1) of L.

Define

$$E'=\{(A,B)\in E\times\hat{L}\,|\,\pi(A)=\hat{\pi}(B)\}. \tag{4.53}$$

Obviously, E' is a graded subalgebra of the ε Lie algebra $E\times\hat{L}$. Let $\varrho:E'\to\hat{L}$ be the projection onto the second factor of $E\times\hat{L}$. Then ϱ is a surjective homomorphism of ε Lie algebras, and the kernel of ϱ is equal to

$$\ker(\varrho)=\{(\iota(X),0)\,|\,X\in H\}, \tag{4.54}$$

which is contained in the center of E'. Since $\hat{L}$ is simply connected, there exists a unique homomorphism of ε Lie algebras $\sigma:\hat{L}\to E'$, such that $\varrho\circ\sigma=id_{\hat{L}}$. Let $\lambda:\hat{L}\to E$ denote the composition of σ with the projection of E' onto the first factor of $E\times\hat{L}$. The definitions imply that

$$\sigma(B)=(\lambda(B),B)\in E', \tag{4.55}$$

for all $B\in\hat{L}$ and hence (by definition of E') that $\pi\circ\lambda=\hat{\pi}$. In particular, we have $\pi\circ\lambda\circ\hat{\iota}=0$, which shows that

$$(\lambda\circ\hat{\iota})(\hat{H})\subset\iota(H). \tag{4.56}$$

Consequently, there exists a unique homomorphism of ε Lie algebras $\mu:\hat{H}\to H$ such that $\lambda\circ\hat{\iota}=\iota\circ\mu$. Then (λ,μ) is a morphism from the covering (4.52) to the central extension (4.1).

Conversely, let us assume that (4.52) is a universal covering of L. Consider a central extension,

$$\{0\}\to H\overset{\iota}{\to}E\overset{\pi}{\to}\hat{L}\to\{0\}. \tag{4.57}$$

We have to show that there exists a unique homomorphism of ε Lie algebras $\sigma:\hat{L}\to E$ such that $\pi\circ\sigma=id_{\hat{L}}$.

If σ exists, the pair $(\sigma,0)$ is a morphism from the covering

$$\{0\}\to\{0\}\to\hat{L}\overset{id}{\to}\hat{L}\to\{0\} \tag{4.58}$$

to the central extension (4.57). Since $\hat{L}$ is perfect, Lemma 4.1 implies that σ is uniquely determined. Hence it remains to prove that σ exists.

Let $\tilde{E}=\langle E,E\rangle$ be the commutator algebra of E. Since $\hat{L}$ is perfect, we have $\pi(\tilde{E})=\hat{L}$ and hence $E=\tilde{E}+\iota(H)$. This in turn implies that

$$\tilde{E}=\langle\tilde{E}+\iota(H),\tilde{E}+\iota(H)\rangle=\langle\tilde{E},\tilde{E}\rangle, \tag{4.59}$$

i.e., that $\tilde{E}$ is perfect. Let $\tilde{H}$ be the preimage of $\tilde{E}$ under ι. Then the central extension (4.57) induces a central extension,

$$\{0\}\to\tilde{H}\to\tilde{E}\to\hat{L}\to\{0\}, \tag{4.60}$$

and the homomorphism σ exists, if the analogous homomorphism exists for (4.60). Thus we may assume that E is perfect.

Let $\nu:E\to L$ be the (surjective) homomorphism of ε Lie algebras defined by $\nu=\hat{\pi}\circ\pi$. We are going to show that the kernel of ν is contained in the center of E. Thus suppose that A is a homogeneous element of E such that $\hat{\pi}(\pi(A))=0$. Then $\pi(A)$ is central in $\hat{L}$. Consequently, we have

$$\pi(\langle A,B\rangle)=\langle\pi(A),\pi(B)\rangle=0, \tag{4.61}$$

for all $B\in E$, i.e., $\langle A,B\rangle$ belongs to the kernel of π and hence to the center of E. If B and B' are any two homogeneous elements of E, this implies that

$$\langle A,\langle B,B'\rangle\rangle=\langle\langle A,B\rangle,B'\rangle+\varepsilon(\alpha,\beta)\langle B,\langle A,B'\rangle\rangle=0, \tag{4.62}$$

and since E is perfect it follows that A lies in the center of E.

Thus

$$\{0\}\to\ker(\nu)\to E\overset{\nu}{\to}L\to\{0\} \tag{4.63}$$

is a central extension. Since we are assuming that (4.52) is a universal covering, there exists a morphism (φ,ψ) from (4.52) to (4.63). We are going to show that φ is the homomorphism σ we are looking for. To do so we have to show that $\pi\circ\varphi=id_{\hat{L}}$. Let $\omega:\hat{L}\to\hat{L}$ be the linear map defined by $\omega=\pi\circ\varphi-id_{\hat{L}}$. Since (φ,ψ) is a morphism of central extensions, we have

$$\hat{\pi}\circ\omega=\hat{\pi}\circ(\pi\circ\varphi)-\hat{\pi}=\nu\circ\varphi-\hat{\pi}=0, \tag{4.64}$$

and hence ω takes its values in the center of $\hat{L}$. This implies that

$$\omega(\langle A,B\rangle)=\langle\omega(A),(\pi\circ\varphi)(B)\rangle+\langle A,\omega(B)\rangle=0, \tag{4.65}$$

for all $A,B\in\hat{L}$. Since $\hat{L}$ is perfect, it follows that $\omega=0$ and the theorem is proved.

Example: The universal covering of $sl(n|n)/\mathbb{K}\cdot I_{2n}$. For abbreviation, we write L instead of $sl(n|n)/\mathbb{K}\cdot I_{2n}$ (here and in the following, I_{2n} denotes the $2n\times 2n$ unit matrix). It is well-known[8] that L inherits from $gl(n|n)$ a consistent $\mathbb{Z}$-gradation such that

$$L=L_{-1}\oplus L_0\oplus L_1. \tag{4.66}$$

To begin with, we note that

$$\{0\}\to\mathbb{K}\overset{\iota}{\to}sl(n|n)\overset{\pi}{\to}L\to\{0\} \tag{4.67}$$

is a central extension of L, where ι is defined by

$$\iota(c)=\frac{c}{2n}I_{2n}, \tag{4.68}$$

for all $c\in\mathbb{K}$, and where π is the canonical homomorphism. Obviously, there exists a unique linear map,

$$\sigma:L\to sl(n|n), \tag{4.69}$$

such that

$$\sigma(\pi(A))=A-\frac{1}{2n}\,\mathrm{Tr}(A)I_{2n}, \tag{4.70}$$

for all $A\in sl(n|n)$, and we have $\pi\circ\sigma=id_L$. It follows that the 2-cocycle $g\in Z^2(L,\mathbb{K})$ defined by Eq. (4.10) is given by

$$g(\pi(A),\pi(B))=\mathrm{Tr}(\langle A,B\rangle), \tag{4.71}$$

for all $A,B\in sl(n|n)$. Obviously, g is L_0-invariant (but, of course, not L-invariant).

If $n=1$, the Lie superalgebra L is Abelian and hence the theory of universal coverings does not apply (see Remark 4.3). Thus from now on we shall assume that $n\geqslant 2$. Then L is simple and hence perfect. Our main task is to determine $H^2(L,\mathbb{K})$. Since $L_0\simeq sl(n)\times sl(n)$ is semi-simple, Proposition 2.1 implies that every cohomology class contains an L_0-invariant cocycle. On the other hand, there does not exist a nonzero L_0-invariant linear form on L, i.e., $C^1(L,\mathbb{K})$ does not contain a nontrivial L_0-invariant element. Consequently, Proposition 2.1 shows that $H^2(L,\mathbb{K})$ can be identified with the space of all L_0-invariant 2-cochains. Thus we are going to determine this space.

For $n\geqslant 3$, this is easy. Let us first recall that every L_0-invariant bilinear form on L_0 is symmetric. Using this fact, a short look at the representations of L_0 carried by L_0 and $L_{\pm 1}$ shows that there exists, up to the normalization, a unique nonzero super-skew-symmetric L_0-invariant bilinear form on L. As we already know, g has these properties, moreover, g is a 2-cocycle. Thus we conclude, for $n\geqslant 3$, that $H^2(L,\mathbb{K})$ is one-dimensional and that (4.67) is a universal covering of $L=sl(n|n)/\mathbb{K}\cdot I_{2n}$.

Due to the peculiarities of the algebra $sl(2|2)/\mathbb{K}\cdot I_4$, the case $n=2$ is more interesting. In principle, we could proceed in this case as above. However, we prefer to use a different realization of $sl(2|2)/\mathbb{K}\cdot I_4$ which makes the special properties of this algebra manifest (and which has already been used in the description of the super derivations of this algebra[8]).

Let V be a two-dimensional vector space, let ψ be a nondegenerate skew-symmetric bilinear form on V, and let $sp(\psi)$ [equal to $sl(V)$] be the Lie algebra of all linear maps of V into itself leaving the form ψ invariant. Define the bilinear map,

$$P:V\times V\to sp(\psi), \tag{4.72}$$

by

$$P(x,y)z=\psi(x,z)y+\psi(y,z)x, \tag{4.73}$$

for all $x,y,z\in V$. We set

$$L_{\bar{0}}=sp(\psi)\times sp(\psi),\quad L_{\bar{1}}=V\otimes V\otimes V, \tag{4.74}$$

and regard L_1^- as an L_0^--module, where the first (resp., second) factor of L_0^- acts in the natural way on the first (resp., second) tensorial factor of L_1^-. For any $Q \in sp(\psi)$, let Q_1 and Q_2 be the elements of L_0^- defined by

$$Q_1=(Q,0), \quad Q_2=(0,Q). \tag{4.75}$$

Using these data, the $\mathbb{Z}_2$-graded vector space,

$$L=L_0^- \oplus L_1^-, \tag{4.76}$$

becomes a Lie superalgebra, if the bracket of two elements of L_1^- is chosen such that

$$\langle x\otimes y\otimes z, x'\otimes y'\otimes z'\rangle = cP_1(x,x')\psi(y,y')\psi(z,z') - c\psi(x,x')P_2(y,y')\psi(z,z'), \tag{4.77}$$

for all $x,y,\ldots,z' \in V$, where $c \in \mathbb{K}$ is an arbitrary nonzero constant.

This Lie superalgebra is isomorphic to $sl(2|2)/\mathbb{K}\cdot I_4$. More precisely, one can easily show the following.

Let (v_+, v_-) be any basis of V. If we define

$$L_0=L_0^-, \quad L_{\pm 1}=V\otimes V\otimes \mathbb{K}v_\pm, \tag{4.78}$$

then

$$L=L_{-1}\oplus L_0\oplus L_1 \tag{4.79}$$

is a consistent $\mathbb{Z}$-gradation of L, and L and $sl(2|2)/\mathbb{K}\cdot I_4$ are isomorphic as $\mathbb{Z}$-graded Lie superalgebras.

The arbitrariness in the choice of $v_\pm$ reflects the peculiarity of $sl(2|2)/\mathbb{K}\cdot I_4$ we have been alluding to: This algebra admits a large group of outer automorphisms. In fact, let s be an arbitrary element of the group $Sp(\psi)$ [equal to $SL(V)$] and let $\tilde{s}$ be the linear map of L into itself, which induces the identity map on L_0^-, and on L_1^- is given by

$$\tilde{s}(x\otimes y\otimes z)=x\otimes y\otimes s(z), \tag{4.80}$$

for all $x,y,z \in V$. Then $\tilde{s}$ is an automorphism of the Lie superalgebra L.

It is easy to determine the L_0^--invariant super-skew-symmetric bilinear forms on L. Let g be such a form. For the same reasons as above, $g(A,B)$ has to vanish if at least one of the two homogeneous elements $A,B \in L$ is even. On the other hand, the restriction of g to $L_1^- \times L_1^-$ is given by

$$g(x\otimes y\otimes z, x'\otimes y'\otimes z') = \psi(x,x')\psi(y,y')\varphi(z,z'), \tag{4.81}$$

for all $x,y,\ldots,z' \in V$, where φ is an arbitrary symmetric bilinear form on V. Conversely, it is not difficult to check that every bilinear form g with these properties is, in fact, an L_0^--invariant 2-cocycle.

Obviously, the map assigning to any symmetric bilinear form φ on V the L_0^--invariant 2-cocycle specified above is a vector space isomorphism of the spaces under consideration. Consequently, we have

$$\dim H^2(sl(2|2)/\mathbb{K}\cdot I_4, \mathbb{K}) = 3. \tag{4.82}$$

The universal covering of L can now be constructed along the lines specified around Eq. (4.39). Actually, this is not really necessary, since this algebra is known: It is obtained if one sets out to construct the Serre presentation of $sl(2|2)$ and forgets about the so-called supplementary relations. More precisely, the universal covering of $sl(2|2)/\mathbb{K}\cdot I_4$ is the Lie superalgebra $s(2,2)$ considered in Ref. 38. In fact, there is a natural homomorphism of the Lie superalgebra $s(2,2)$ onto the Lie superalgebra $sl(2|2)/\mathbb{K}\cdot I_4$ whose kernel is equal to the three-dimensional center of $s(2,2)$; moreover, it is known that $s(2,2)$ is perfect. Since the universal covering of $sl(2|2)/\mathbb{K}\cdot I_4$ must cover $s(2,2)$ as well (see Remark 4.1), it follows that this universal covering must be

isomorphic to $s(2,2)$. Actually, it is not difficult to construct this isomorphism explicitly. According to Remark 4.1, $s(2,2)$ is also a universal covering of $sl(2|2)$. [Note that this is at variance with Theorem 2.6.1 of Ref. 10, but it is consistent with Ref. 9. On the other hand, in the latter reference, the definition of the two basic 2-cocycles of $sl(2|2)$ is not correct, and it is easy to see that, contrary to what is stated, we have $H^3(L,\mathbb{K})\neq\{0\}$ for both $L=sl(2|2)$ and $L=sl(2|2)/\mathbb{K}\cdot I_4$. We think that these discrepancies and mistakes are due to some trivial slips or misprints.]

V. COHOMOLOGY OF THE LIE SUPERALGEBRA $sl(1|2)$

Throughout the present section we shall assume that the base field $\mathbb{K}$ is *algebraically closed.* Let $U(sl(1|2))$ denote the enveloping algebra of $sl(1|2)$. Our main goal is to show that

$$H^2(sl(1|2),U(sl(1|2)))=\{0\}. \tag{5.1}$$

In order to prove this, we shall first determine $H^1(sl(1|2),V)$ and $H^2(sl(1|2),V)$, for all finite-dimensional simple graded $sl(1|2)$-modules V. Actually, we only need to know those modules V for which these cohomology groups are nontrivial. As explained in the Introduction, this information could be extracted from Ref. 15, but we think it is worthwhile to present our own calculations.

As an abbreviation, we write L instead of $sl(1|2)$. We are going to use the notation and results of Ref. 39. Let us first recall some of the most relevant facts. The basis of L we are going to use consists of four even elements $Q_\pm$, Q_3, B, and four odd elements $V_\pm$ and $W_\pm$. The elements $Q_\pm$, Q_3 span an $sl(2)$ subalgebra of $L_{\bar 0}$; they are normalized such that

$$\langle Q_+,Q_-\rangle=2Q_3. \tag{5.2}$$

The one-dimensional subspace $\mathbb{K}\cdot B$ is the center of $L_{\bar 0}$, and the two-dimensional subspace,

$$H=\mathbb{K}\cdot B+\mathbb{K}\cdot Q_3, \tag{5.3}$$

is a Cartan subalgebra of L. Thus any linear form λ on H (in particular, any weight of an L-module) can and will be identified with the pair (b,q) of values which λ takes on B and Q_3, respectively. For example, $V_\pm$ and $W_\pm$ are root vectors corresponding to the roots $(\frac{1}{2},\pm\frac{1}{2})$ and $(-\frac{1}{2},\pm\frac{1}{2})$, respectively. They are normalized such that $(V_\pm)$ and $(W_\pm)$ are standard $sl(2)$-doublets and satisfy

$$\langle V_\pm,W_\pm\rangle=\pm Q_\pm, \tag{5.4}$$

$$\langle V_\pm,W_\mp\rangle=-Q_3\pm B. \tag{5.5}$$

The reader who prefers a matrix realization may set

$$Q_+=E_{23},\quad Q_-=E_{32},\quad Q_3=\tfrac{1}{2}(E_{22}-E_{33}),\quad B=-\tfrac{1}{2}(2E_{11}+E_{22}+E_{33}), \tag{5.6}$$

$$V_+=E_{21},\quad V_-=E_{31},\quad W_+=E_{13},\quad W_-=-E_{12}, \tag{5.7}$$

where the E_{ij} are the usual basic 3×3 matrices. [The elements are chosen such that the $\mathbb{Z}$-gradation of $sl(1|2)$ as specified in Ref. 8 and the $\mathbb{Z}$-gradation given by Eq. (5.11) coincide.]

For later use we note that the linear map,

$$\omega:L\to L, \tag{5.8}$$

defined by

$$\omega(Q_\pm)=Q_\pm,\quad \omega(Q_3)=Q_3,\quad \omega(B)=-B, \tag{5.9}$$

$$\omega(V_\pm)=W_\pm,\quad \omega(W_\pm)=V_\pm, \tag{5.10}$$

is an automorphism of the Lie superalgebra L.

The simple roots of L with respect to H are chosen such that V_+ and W_+ are the associated root vectors. Thus both of the simple roots are odd, i.e., we do not use a so-called distinguished basis of the root system. For $sl(1|2)$, this choice has certain advantages, for example, it is compatible with the automorphism ω.

For any weight (b,q); $b,q\in\mathbb{K}$, and any $\gamma\in\mathbb{Z}_2$, there exists (up to isomorphism) a unique simple graded highest weight L-module $V(b,q,\gamma)$ such that the highest weight vectors (which are proportional to each other) have weight (b,q) and $\mathbb{Z}_2$-degree γ. Of course, $V(b,q,\gamma)$ and $V(b,q,\gamma+\overline{1})$ can be transformed into each other by a shift of the $\mathbb{Z}_2$-gradation [see Eq. (2.44)]. The module $V(b,q,\gamma)$ is finite-dimensional if and only if $b=q=0$ or $q\in\frac{1}{2}\mathbb{N}_*$, and any finite-dimensional simple graded L-module is of this type. Moreover, it is atypical[35] if and only if $b=\pm q$, and its dimension is equal to $8q$ (resp., $4q+1$) in the typical (resp., atypical) case.

The Casimir elements of L are known, and the eigenvalues of the Casimir operators in a graded highest weight module have been calculated in terms of a generating function[26,32,33] (see also Example D of Sec. III). In particular, the image of the generalized Harish–Chandra isomorphism is known. It can be regarded as an algebra of polynomials in b,q and, as such, consists of all polynomials of the form $a+(b^2-q^2)P$, where $a\in\mathbb{K}$ is a constant and P is a polynomial in b and (b^2-q^2). This implies that, for a graded highest weight module with highest weight (b,q), all Casimir operators without a constant term are equal to zero, if and only if $b=\pm q$.

Next we recall that L has a consistent $\mathbb{Z}$-gradation, which is fixed by the requirement that the elements $Q_\pm$, Q_3, and B are homogeneous of degree 0, the elements $V_\pm$ are homogeneous of degree 1, and the elements $W_\pm$ are homogeneous of degree -1. Thus we can apply our general theory both in the $\mathbb{Z}_2$-graded and in the $\mathbb{Z}$-graded sense. Actually, this essentially amounts to a matter of formulation (at least in the present article). This is because, for all L-modules V for which a $\mathbb{Z}$-gradation will be of interest, this gradation is simply given by the action of B, as follows:

$$V_r=\{x\in V|2B\cdot x=rx\}, \tag{5.11}$$

for all $r\in\mathbb{Z}$. In particular, this is true for L itself and for the enveloping algebra $U(L)$ (both endowed with the adjoint action of L).

As a final preparatory remark we notice that there is a natural embedding of the Lie superalgebra $osp(1|2)$ into $sl(1|2)$. In fact, if we define

$$U_\pm=\tfrac{1}{2}(V_\pm+W_\pm), \tag{5.12}$$

the elements $Q_\pm$, Q_3, $U_\pm$ satisfy the commutation relations of $osp(1|2)$ as specified in Ref. 39. In particular, we have

$$\langle U_\pm,U_\pm\rangle=\pm\tfrac{1}{2}Q_\pm, \tag{5.13}$$

$$\langle U_\pm,U_\mp\rangle=-\tfrac{1}{2}Q_3. \tag{5.14}$$

The $\mathbb{Z}_2$-graded subalgebra of L spanned by these elements will be denoted by G; by construction, it is isomorphic to $osp(1|2)$.

We recall that the finite-dimensional graded G-modules are semi-simple. The finite-dimensional simple graded G-modules can be labeled by a pair (q,γ), with $q\in\frac{1}{2}\mathbb{N}$ and $\gamma\in\mathbb{Z}_2$: By definition, a highest weight vector v of the corresponding module is homogeneous of degree γ and satisfies $Q_3\cdot v=qv$. We denote this module (which is fixed up to isomorphism) by $U(q,\gamma)$; its dimension is equal to $4q+1$.

Of course, every graded L-module can also be regarded as a graded G-module. For example, choose $b\in\mathbb{K}$, $q\in\frac{1}{2}\mathbb{N}_*$, $\gamma\in\mathbb{Z}_2$, and regard $V(b,q,\gamma)$ as a G-module. If $b\neq\pm q$, this module decomposes into the direct sum of two simple graded G-modules, isomorphic to $U(q,\gamma)$ and $U(q-\frac{1}{2},\gamma+\overline{1})$, respectively. On the other hand, the G-module $V(\pm q,q,\gamma)$ is simple and isomorphic to $U(q,\gamma)$. (Obviously, the latter statement remains true for $q=0$.)

In particular, consider the adjoint L-module L. Regarded as a G-module, it decomposes into two G-submodules: One of them is G itself, the other is spanned by $X_\pm$ and B, with

$$X_{\pm}=\tfrac{1}{2}(V_{\pm}-W_{\pm}). \tag{5.15}$$

The elements $X_{\pm}$ have been chosen such that $X_{\pm}$ and B form a basis of $U(\frac{1}{2},\overline{1})$ as specified in Ref. 39. We remark that $Q_{\pm}$, Q_3, $iX_{\pm}$ also satisfy the commutation relations of $osp(1|2)$.

After these preliminaries we are ready to tackle our main problem: We want to determine the cohomology groups $H^1(L,V)$ and $H^2(L,V)$, with V a finite-dimensional simple graded L-module. In view of the preceding remarks, we deduce from Proposition 2.2 that $H^n(L,V)=\{0\}$ for all n if V is typical. Consequently, we only have to consider the cases where $V=V(\pm q,q,\gamma)$, with $q\in\frac{1}{2}\mathbb{N}$ and $\gamma\in\mathbb{Z}_2$. Actually, if ϱ is the representation afforded by $V(q,q,\gamma)$, then $\varrho\circ\omega$ is (equivalent to) the representation afforded by $V(-q,q,\gamma)$ [where the automorphism ω has been defined by Eqs. (5.8) – (5.10)]. Recalling the isomorphism (2.34) we may assume, therefore, that $V=V(q,q,\gamma)$. On the other hand, the bijection (2.47) shows that it is sufficient to consider only one of the two choices for γ. According to Eq. (5.11), we shall choose $\gamma=\overline{2q}$ (where $\overline{r}=r+2\mathbb{Z}\in\mathbb{Z}_2$ denotes the residue class of the integer r modulo 2), and assume that $V(q,q,\overline{2q})$ is endowed with the consistent $\mathbb{Z}$-gradation defined by Eq. (5.11). To simplify the notation, we write $V(q)$ instead of $V(q,q,\overline{2q})$. Then we have to determine the cohomology groups $H^n(L,V(q))$, with $n\in\{1,2\}$.

The case $n=1$ is simple. We present it in some detail to explain our approach. According to Proposition 2.1, any 1-cocycle of L with values in $V(q)$ is cohomologous to a G-invariant 1-cocycle. Regarded as graded G-modules, L is isomorphic to $U(1,\overline{0})\oplus U(\frac{1}{2},\overline{1})$, and $V(q)$ is isomorphic to $U(q,\overline{2q})$. But a nonzero G-invariant linear map of a G-module $U(p,\overline{2p})$ into a G-module $U(q,\overline{2q})$ (with $p,q\in\frac{1}{2}\mathbb{N}$) exists if and only if $p=q$, and such a map must be even. Consequently, a nonzero G-invariant linear map $g:L\to V(q)$ exists if and only if $q\in\{\frac{1}{2},1\}$. Moreover, if this condition is satisfied, a map of the type in question is even and uniquely determined up to a scalar factor.

On the other hand, let V be any graded L-module, and let $g:L\to V$ be a G-invariant 1-cocycle which is homogeneous of degree $\overline{0}$. Then the G-invariance and the cocycle condition imply that, for all homogeneous elements $X\in G$ and $Y\in L$,

$$X\cdot g(Y)=g(\langle X,Y\rangle)=X\cdot g(Y)-\varepsilon(\xi,\eta)Y\cdot g(X), \tag{5.16}$$

and hence that

$$Y\cdot g(X)=0. \tag{5.17}$$

In particular, we conclude that $g(\langle X,X'\rangle)=0$ for all $X,X'\in G$, which implies that

$$g(X)=0, \tag{5.18}$$

for all $X\in G$.

Thus all we have to do is to determine $H^1(L,V(\frac{1}{2}))$. To construct a nonzero G-module homomorphism of L into $V(\frac{1}{2})$ we introduce a suitable basis in $V(\frac{1}{2})$. According to Ref. 39 there exists a basis (e_+,e_0,e_-) of $V(\frac{1}{2})$ such that $e_{\pm}$ are odd and e_0 is even, such that $e_{\pm}$ and e_0 are weight vectors corresponding to the weights $(\frac{1}{2},\pm\frac{1}{2})$ and $(1,0)$, respectively, and such that

$$V_{\pm}\cdot e_{\pm}=0,\quad V_{\pm}\cdot e_{\mp}=\mp e_0 \tag{5.19}$$

$$W_{\pm}\cdot e_{\pm}=W_{\pm}\cdot e_{\mp}=0, \tag{5.20}$$

$$V_{\pm}\cdot e_0=0,\quad W_{\pm}\cdot e_0=-e_{\pm}. \tag{5.21}$$

(Obviously, this basis is uniquely fixed up to a nonzero overall factor.) The action of $Q_{\pm}$ on the basis vectors can be derived from the commutation relations and the formulas above; it is such that $(e_{\pm})$ is a standard $sl(2)$-doublet and e_0 an $sl(2)$-singlet. It follows that $e_{\pm}$ and e_0 also form a basis of the G-module $V(\frac{1}{2})\simeq U(\frac{1}{2},\overline{1})$ as described in Ref. 39. Consequently, the linear map,

$$g: L \to V(\tfrac{1}{2}), \tag{5.22}$$

defined by

$$g(X) = 0, \quad \text{if } X \in G, \tag{5.23}$$

$$g(B) = e_0, \tag{5.24}$$

$$g(V_\pm - W_\pm) = 2e_\pm, \tag{5.25}$$

is a homomorphism of graded G-modules. Since

$$g(V_\pm + W_\pm) = g(2U_\pm) = 0, \tag{5.26}$$

it follows that

$$g(V_\pm) = -g(W_\pm) = e_\pm. \tag{5.27}$$

It is now easy to check that g is a 1-cocycle. Moreover, since $V(\frac{1}{2})$ does not contain a nonzero G-invariant element, g is not a coboundary. Summarizing, we have shown that

$$H^1(L, V(q)) = \{0\}, \quad \text{if } q \neq \tfrac{1}{2}, \tag{5.28}$$

$$\dim H^1(L, V(\tfrac{1}{2})) = 1. \tag{5.29}$$

The fact that $V(\frac{1}{2})$ is (essentially) the sole finite-dimensional simple graded L-module V for which $H^1(L,V)$ is nontrivial has already been mentioned in the *Additional remarks* at the end of Ref. 7.

The cocycle g can be replaced by a simpler one. Let us regard L as a consistently $\mathbb{Z}$-graded Lie superalgebra and $V(\frac{1}{2})$ as a $\mathbb{Z}$-graded L-module [recall that the $\mathbb{Z}$-gradations are chosen according to Eq. (5.11)]. Then g can be decomposed into its $\mathbb{Z}$-homogeneous components,

$$g = g_0 + g_2, \tag{5.30}$$

where g_r, $r \in \{0,2\}$ is homogeneous of degree r. Explicitly, the nonvanishing values of g_0 and g_2 on the basis vectors $Q_\pm$, Q_3, B, $V_\pm$, $W_\pm$ are given by

$$g_0(V_\pm) = e_\pm, \tag{5.31}$$

$$g_2(B) = e_0, \quad g_2(W_\pm) = -e_\pm. \tag{5.32}$$

Both g_0 and g_2 are 1-cocycles. Hence exactly one of them is not a coboundary. Obviously, g_0 cannot be a coboundary, for $V(\frac{1}{2})$ does not contain a nonzero $\mathbb{Z}$-homogeneous element of degree 0. Consequently, g_2 must be a coboundary. In fact, we have

$$g_2(X) = X \cdot e_0, \tag{5.33}$$

for all $X \in L$.

The 1-cocycle g_0 can be used to construct higher order cocycles, as follows. Choose any integer $n \geqslant 1$. Using the product of cocycles defined in Eq. (2.58) and the associativity of this product [see Eq. (2.64)], we can construct the n-cocycle $g_0^{\odot n} = g_0 \odot \cdots \odot g_0$ (n factors g_0) with values in $V(\frac{1}{2})^{\otimes n}$, which is given by

$$g_0^{\odot n}(X_1, \ldots, X_n) = \sum_{\pi \in \mathcal{S}_n} \operatorname{sgn}(\pi) \varepsilon_n(\pi; \xi_1, \ldots, \xi_n) g_0(X_{\pi(1)}) \otimes \cdots \otimes g_0(X_{\pi(n)}), \tag{5.34}$$

for all homogeneous elements $X_i \in L$. Obviously, $g_0^{\odot n}$ is nonzero and takes its values in the graded L-submodule $W(n)$ of super-skew-symmetric tensors in $V(\frac{1}{2})^{\otimes n}$.

Let $f_n: L^n \to W(n)$ denote the n-cocycle defined by $g_0^{\odot n}$. Of course, f_n is $\mathbb{Z}$-homogeneous of degree zero. Since the degree of a nonzero $\mathbb{Z}$-homogeneous element of $W(n)$ is at least equal to n, there does not exist a nonzero $(n-1)$-linear map of L^{n-1} into $W(n)$, which is $\mathbb{Z}$-homogeneous of degree zero. Consequently, f_n is not a coboundary. On the other hand, it is easy to see that $W(n)$, even when regarded as a $\mathbb{Z}$-graded L-module, is isomorphic to $V(\frac{n}{2})$. Thus we have proved that

$$H^n(L, V(\tfrac{n}{2})) \neq \{0\}, \quad \text{for all } n \geqslant 1. \tag{5.35}$$

Remark 5.1: Our results show that Proposition 2.2 exactly describes the finite-dimensional simple graded L-modules with a nontrivial cohomology, and the same holds for the $osp(1|2n)$ algebras. It would be interesting to know whether this is also true for other simple Lie superalgebras.

Let us next determine the cohomology groups $H^2(L, V(q))$. Since we are going to proceed as for $H^1(L, V(q))$, it should suffice to give the main intermediate steps. For any graded L-module V, the graded L-module $C^2(L, V)$ is canonically isomorphic to $\mathrm{Lgr}(L \wedge_\varepsilon L, V)$ (see Appendix A). Thus we have to determine the structure of $L \wedge_\varepsilon L$. It is known that this module is isomorphic to the graded submodule of super-skew-symmetric tensors in $L \otimes L$. This submodule is semi-simple (whereas $L \otimes L$ is not), and its simple submodules can be determined, with the result that

$$L \wedge_\varepsilon L \simeq V(\tfrac{1}{2}, \tfrac{3}{2}, \overline{1}) \oplus V(-\tfrac{1}{2}, \tfrac{3}{2}, \overline{1}) \oplus V(0,1,\overline{0}). \tag{5.36}$$

This implies that, regarded as a graded G-module, $L \wedge_\varepsilon L$ is isomorphic to

$$2\, U(\tfrac{3}{2}, \overline{1}) \oplus 3\, U(1, \overline{0}) \oplus U(\tfrac{1}{2}, \overline{1}). \tag{5.37}$$

Recalling that the graded G-module $V(q)$ is isomorphic to $U(q, \overline{2q})$, we conclude as before that $H^2(L, V(q))$ is trivial unless $q \in \{\frac{1}{2}, 1, \frac{3}{2}\}$.

Let us consider the case $q = \frac{1}{2}$. Visibly, there exists, up to a scalar factor, a unique nonzero G-invariant linear map $\overline{f}$ of $L \wedge_\varepsilon L$ into $V(\frac{1}{2})$, and this map is even and vanishes on the submodule W isomorphic to

$$W \simeq V(\tfrac{1}{2}, \tfrac{3}{2}, \overline{1}) \oplus V(-\tfrac{1}{2}, \tfrac{3}{2}, \overline{1}). \tag{5.38}$$

Let $f \in C^2(L, V(\frac{1}{2}))$ be the associated bilinear map. We are going to show that f is not a cocycle. According to Proposition 2.1, this will imply that

$$H^2(L, V(\tfrac{1}{2})) = \{0\}. \tag{5.39}$$

In fact, the product map $\langle , \rangle$ of L induces an L-module homomorphism of $L \wedge_\varepsilon L$ onto L, whose kernel is equal to W. Consequently, there exists a unique homomorphism of graded G-modules,

$$g': L \to V(\tfrac{1}{2}), \tag{5.40}$$

such that

$$f(X, Y) = g'(\langle X, Y \rangle), \tag{5.41}$$

for all $X, Y \in L$. But we know that there exists, up to a scalar factor, a unique homomorphism of graded G-modules of L into $V(\frac{1}{2})$, namely, the map g defined by Eqs. (5.22) – (5.25). Thus, by multiplying f by a suitable scalar factor, we may assume that

$$f(X, Y) = g(\langle X, Y \rangle), \tag{5.42}$$

for all $X, Y \in L$. If f would be a 2-cocycle, its $\mathbb{Z}$-homogeneous components f_r, given by

$$f_r(X,Y)=g_r(\langle X,Y\rangle), \quad r\in\{0,2\}, \tag{5.43}$$

would be 2-cocycles as well. However, using Eq. (5.33), it is easy to see that f_2 is not a cocycle.

To proceed, we remark that the condition $q\in\{\frac{1}{2},1,\frac{3}{2}\}$ derived above can be improved by regarding $L\wedge_\varepsilon L$ and $V(q)$ as $L_{\bar 0}$-modules. The same type of argument as before then shows that $H^2(L,V(q))$ is trivial unless $q\in\{0,\frac{1}{2},1\}$; moreover, there exists, up to a scalar factor, a unique $L_{\bar 0}$-invariant linear map of $L\wedge_\varepsilon L$ into $V(1)$. Since (again by Proposition 2.1) any 2-cocycle on L with values in $V(1)$ is cohomologous to an $L_{\bar 0}$-invariant cocycle, we conclude from Eq. (5.35) that $H^2(L,V(1))$ is one-dimensional. Summarizing, we have shown that

$$H^2(L,V(q))=\{0\}, \quad \text{if } q\neq 1, \tag{5.44}$$

$$\dim H^2(L,V(1))=1. \tag{5.45}$$

As an application of the preceding result, let us now prove that $H^2(L,U(L))$ is trivial. The following lemma contains the pertinent information on the L-module $U(L)$.

Lemma 5.1: Any finite-dimensional atypical simple graded subquotient of the adjoint L-module $U(L)$ is isomorphic to one of the three graded L-modules $V(0,0,\overline{0})$ and $V(\pm\frac{1}{2},\frac{1}{2},\overline{1})$.

Proof: Consider two graded submodules $U'\subset U$ of $U(L)$ such that U/U' is a finite-dimensional atypical simple graded L-module. Then U/U' is isomorphic to one of the modules $V(\pm q,q,\gamma)$, with $q\in\mathbb{N}$ and $\gamma\in\mathbb{Z}_2$, and we have to show that $q\in\{0,\frac{1}{2}\}$ and $\gamma=\overline{2q}$.

Let ad denote the adjoint representation of L in $U(L)$. Then ad B is diagonalizable, and the sole eigenvalues of this operator are 0, $\pm\frac{1}{2}$, and ± 1. Since, for $q\geqslant\frac{1}{2}$, the module $V(\pm q,q,\gamma)$ contains nonzero weight vectors with weight $(\pm q\pm\frac{1}{2},q-\frac{1}{2})$, the cases $q\geqslant 1$ are not possible. On the other hand, an eigenvector of ad B is odd if and only if the corresponding eigenvalue is equal to $\pm\frac{1}{2}$, and it is even otherwise. This implies that $\gamma=\overline{2q}$ and proves the lemma.

Remark 5.2: As shown in Appendix B, the adjoint L-module $U(L)$ really has graded subquotients isomorphic to $V(0,0,\overline{0})$ and $V(\pm\frac{1}{2},\frac{1}{2},\overline{1})$.

Remark 5.3: A decomposition of the adjoint L-module $U(L)$ into indecomposable submodules has been constructed in a recent paper by Benamor.[40] (We are grateful to the referee for drawing our attention to this article.) This decomposition is rather complicated, thus we think it would not be adequate to invoke that reference to prove the simple lemma above.

It is now easy to see that $H^2(L,U(L))=\{0\}$. The image of any 2-cochain of L with values in $U(L)$ is contained in a finite-dimensional graded L-submodule of $U(L)$ [use the canonical filtration of $U(L)$]. Thus it is sufficient to show that $H^2(L,V)$ is trivial, for any finite-dimensional graded L-submodule V of $U(L)$. To prove this we choose an increasing sequence of graded L-submodules of $U(L)$,

$$\{0\}=V_0\subset V_1\subset V_2\subset\cdots\subset V_n=V, \tag{5.46}$$

such that V_r/V_{r-1} is a simple graded L-module, for $1\leqslant r\leqslant n$. [Up to the opposite ordering, (5.46) is a Jordan–Hölder sequence of the graded L-module V.] We show by induction on r that

$$H^2(L,V_r)=\{0\}. \tag{5.47}$$

The case $r=0$ is trivial. Now suppose that Eq. (5.47) is true for some $r\in\{0,1,\ldots,n-1\}$. We have a short exact sequence of graded L-modules:

$$\{0\}\to V_r\to V_{r+1}\to V_{r+1}/V_r\to\{0\}. \tag{5.48}$$

Consider the following part of the exact cohomology sequence (2.51) associated to (5.48):

$$\cdots\to H^2(L,V_r)\to H^2(L,V_{r+1})\to H^2(L,V_{r+1}/V_r)\to\cdots\ . \tag{5.49}$$

By assumption, $H^2(L,V_r)$ is trivial, and Lemma 5.1, combined with Eq. (5.44), shows that $H^2(L,V_{r+1}/V_r)$ is trivial as well. Thus, $H^2(L,V_{r+1})$ must be trivial.

Summarizing, we have shown that

$$H^2(L,U(L))=\{0\}, \tag{5.50}$$

i.e., we have proved Eq. (5.1). On the other hand, we shall see in Appendix B that

$$H^1(L,U(L))\neq\{0\}, \tag{5.51}$$

in contrast to what is known for semi-simple Lie algebras.

VI. DISCUSSION

In the present work we have taken some exploratory steps towards a better understanding of the cohomology of Lie superalgebras and their generalizations. It is hardly surprising that the basic definitions and formal techniques known from the cohomology of Lie algebras can immediately be generalized to the general graded setting. In particular, this applies to Garland's theory of universal central extensions of Lie algebras. The picture changes if we try to actually calculate the cohomology groups of Lie superalgebras, especially, when the coefficients are nontrivial. Our main tools in this project were two simple propositions proved in Sec. II.

As for many questions from the theory of Lie superalgebras, the $osp(1|2n)$ algebras behave very much like simple Lie algebras. On the other hand, already for $sl(1|2)$ the situation changes drastically. For any integer $n\geqslant 0$, there exists a finite-dimensional simple graded $sl(1|2)$-module V such that $H^n(sl(1|2),V)$ is nontrivial. Moreover, in Appendix B we are going to show that

$$H^1(sl(1|2),U(sl(1|2)))\neq\{0\}. \tag{6.1}$$

Thus it is remarkable that, nevertheless,

$$H^2(sl(1|2),U(sl(1|2)))=\{0\}. \tag{6.2}$$

Actually, one of the main reasons for our present investigation was to try to prove the analogous result for as many simple Lie superalgebras L as possible, from which it follows that the associative superalgebra $U(L)$ does not admit any nontrivial formal deformations in the sense of Gerstenhaber.[18] To see this, all we have to do is to transcribe the corresponding discussion of Ref. 19 to the present setting. Then Eq. (5.50) implies that the Hochschild 2-cohomology (in the super sense) of the associative superalgebra $U(L)$ with values in $U(L)$ [considered as a graded $U(L)$-bimodule] vanishes. In this connection, we recall that there is an isomorphism relating the cohomology of a Lie algebra L with values in $U(L)$ to the Hochschild cohomology of the associative algebra $U(L)$ (see Theorem 5.1 of chapter XIII in Ref. 4). It would be interesting to know whether an analogous isomorphism exists in the general graded setting.

The methods used in our investigation are completely elementary, and already for the 3-cohomology of $sl(1|2)$ the calculations become quite extensive. On the other hand, in the meantime we have shown that $H^2(L,U(L))=\{0\}$ for $L=sl(m|1)$, $m\geqslant 3$, but up to now we were not able to prove or disprove this for $L=sl(3|2)$. Thus it might turn out that some more profound techniques must be used both in the study of higher order cohomology groups and in the case of more complicated Lie superalgebras.

ACKNOWLEDGMENTS

We should like to thank C. Gruson for sending us a copy of the galley proofs of her article[14] and for drawing our attention to Ref. 15. The present work was initiated during a visit of the first-named author to the Department of Pure Mathematics of the University of Adelaide. The kind invitation by the second-named author and the hospitality extended to the first-named author, both in the Mathematics and in the Physics Department, are gratefully acknowledged.

APPENDIX A: THE ε-EXTERIOR ALGEBRA OF A GRADED L-MODULE

We use the notation introduced at the beginning of Sec. II. Let V be a graded L-module and let

$$T(V)=\bigoplus_{n\in\mathbb{Z}} T_n(V) \tag{A1}$$

be the tensor algebra of V.[20] It is known that $T(V)$ is a $\mathbb{Z}\times\Gamma$-graded associative algebra with a unit element and that

$$T_n(V)=\{0\}, \quad \text{if } n\leqslant -1, \tag{A2}$$

$$T_0(V)=\mathbb{K}, \tag{A3}$$

$$T_n(V)=\underbrace{V\otimes\cdots\otimes V}_{n \text{ factors}}, \quad \text{if } n\geqslant 1. \tag{A4}$$

Moreover, $T(V)$ is a graded L-module, and L acts on $T(V)$ by ε-derivations. Note also that each $T_n(V)$ is a graded L-submodule of $T(V)$.

Now let $J(V,\varepsilon)$ denote the two-sided ideal of $T(V)$, generated by the elements of the form

$$x\otimes y+\varepsilon(\xi,\eta)y\otimes x, \tag{A5}$$

where x and y are homogeneous elements of V. Then the quotient algebra,

$$\wedge_\varepsilon V=T(V)/J(V,\varepsilon), \tag{A6}$$

is called the *ε-exterior (or ε-Grassmann) algebra* of V. The multiplication in $\wedge_\varepsilon V$ will be denoted by $\wedge_\varepsilon$.

Since $J(V,\varepsilon)$ is a $\mathbb{Z}\times\Gamma$-graded ideal, the algebra $\wedge_\varepsilon V$ inherits from $T(V)$ a canonical $\mathbb{Z}\times\Gamma$-gradation. In particular, we have

$$\wedge_\varepsilon V=\bigoplus_{n\in\mathbb{Z}} \overset{n}{\wedge}_\varepsilon V, \tag{A7}$$

where $\wedge_\varepsilon^n V$ is the canonical image of $T_n(V)$ in $\wedge_\varepsilon V$. It follows that

$$\overset{n}{\wedge}_\varepsilon V=\{0\}, \quad \text{if } n\leqslant -1, \tag{A8}$$

$$\overset{0}{\wedge}_\varepsilon V=\mathbb{K}, \tag{A9}$$

$$\overset{1}{\wedge}_\varepsilon V=V. \tag{A10}$$

For $n\geqslant 2$ we also use the notation

$$\overset{n}{\wedge}_\varepsilon V=V\wedge_\varepsilon\cdots\wedge_\varepsilon V \quad (n \text{ factors}). \tag{A11}$$

On the other hand, $J(V,\varepsilon)$ is a Γ-graded L-submodule of $T(V)$, and hence $\wedge_\varepsilon V$ inherits from $T(V)$ the structure of a Γ-graded L-module. Then L acts on $\wedge_\varepsilon V$ by ε-derivations, and each $\wedge_\varepsilon^n V$ is a Γ-graded L-submodule of $\wedge_\varepsilon V$.

The theory of ε-exterior algebras can easily be reduced to the theory of ε'-symmetric algebras,[20] with ε' a commutation factor on a suitable Abelian group Γ'. In fact, set $\Gamma'=\mathbb{Z}\times\Gamma$ and define the commutation factor ε' on Γ' by

$$\varepsilon'((r,\alpha),(s,\beta))=(-1)^{rs}\varepsilon(\alpha,\beta), \tag{A12}$$

for all $r,s\in\mathbb{Z}$ and all $\alpha,\beta\in\Gamma$. Regard V as a Γ'-graded vector space (denoted by V'), such that, for all $(r,\gamma)\in\Gamma'$,

$$V'_{(1,\gamma)}=V_\gamma, \tag{A13}$$

$$V'_{(r,\gamma)}=\{0\}, \quad \text{if } r\neq 1. \tag{A14}$$

On the other hand, regard L as a Γ'-graded ε' Lie algebra (denoted by L'), such that, for all $(r,\gamma)\in\Gamma'$,

$$L'_{(0,\gamma)}=L_\gamma, \tag{A15}$$

$$L'_{(r,\gamma)}=\{0\}, \quad \text{if } r\neq 0, \tag{A16}$$

while leaving the product map unchanged. Obviously, V' and $T(V)$ have a natural structure of a Γ'-graded L'-module.

A priori, $T(V')$ is a $\mathbb{Z}\times\Gamma'=\mathbb{Z}\times\mathbb{Z}\times\Gamma$-graded vector space, where the $\mathbb{Z}$-gradation corresponding to the first factor is the canonical $\mathbb{Z}$-gradation of the tensor algebra $T(V')$, and where the one corresponding to the second factor stems from the $\mathbb{Z}$-gradation of V'. Obviously, these two $\mathbb{Z}$-gradations coincide. Thus we may simply forget about one of them and regard $T(V')$ as a Γ'-graded algebra and a Γ'-graded L'-module. A similar remark applies to the ε'-symmetric algebra $S(V',\varepsilon')$ of V'.

With these conventions, $T(V)$ and $T(V')$ coincide, and $\wedge_\varepsilon V$ is equal to $S(V',\varepsilon')$ (where these statements may be interpreted both in the sense of Γ'-graded algebras and of Γ'-graded L'-modules).

The foregoing observation implies that for any result about ε'-symmetric algebras there is an analogous one for ε-exterior algebras. In particular, from Sec. 12 of Ref. 20 we draw the following conclusions.

Let $n\geqslant 1$ be an integer and let

$$\omega_n: V^n\to\overset{n}{\wedge}_\varepsilon V \tag{A17}$$

be the canonical mapping defined by

$$\omega_n(x_1,\ldots,x_n)=x_1\wedge_\varepsilon\cdots\wedge_\varepsilon x_n, \tag{A18}$$

for all $x_i\in V$. Then ω_n is an ε-skew-symmetric n-linear map, which is homogeneous of degree zero and L-invariant, and the pair $(\wedge^n_\varepsilon V,\omega_n)$ has the following universal property.

For any vector space W and any ε-skew-symmetric n-linear map,

$$g: V^n\to W, \tag{A19}$$

there exists a unique linear map,

$$\hat{g}:\overset{n}{\wedge}_\varepsilon V\to W, \tag{A20}$$

such that

$$g=\hat{g}\circ\omega_n. \tag{A21}$$

Now suppose in addition that W is a graded L-module, and let $\mathrm{Lgr}^a_n(V,W;\varepsilon)$ denote the graded L-submodule of $\mathrm{Lgr}_n(V,\ldots,V;W)$ consisting of all elements which are ε-skew-symmetric. Then the assignment $g\to\hat{g}$ described above induces a graded L-module isomorphism of $\mathrm{Lgr}^a_n(V,W;\varepsilon)$ onto $\mathrm{Lgr}(\wedge^n_\varepsilon V,W)$.

Next, let $T_n^a(V,\varepsilon)$ denote the graded L-submodule of $T_n(V)$ consisting of all ε-skew-symmetric tensors in $T_n(V)$. The ε-skew-symmetrization,

$$t \to \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \operatorname{sgn}(\pi) S_\pi(t), \tag{A22}$$

defines a graded L-module homomorphism of $T_n(V)$ onto $T_n^a(V,\varepsilon)$ (recall that the mappings S_π have been defined in Ref. 20). Using the universal property of $(\wedge_\varepsilon^n V, \omega_n)$, it follows that there exists a unique surjective graded L-module homomorphism,

$$\chi_n : \overset{n}{\wedge}_\varepsilon V \to T_n^a(V,\varepsilon), \tag{A23}$$

such that

$$\chi_n(x_1 \wedge_\varepsilon \cdots \wedge_\varepsilon x_n) = \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \operatorname{sgn}(\pi) S_\pi(x_1 \otimes \cdots \otimes x_n), \tag{A24}$$

for all $x_i \in V$, and it is easy to see that χ_n is injective, and hence an isomorphism of graded L-modules.

On the other hand, from the Poincaré, Birkhoff, Witt theorem[22] we deduce that to any homogeneous basis of V there corresponds, in the obvious way, a homogeneous basis of $\wedge_\varepsilon V$.

Finally, if V and W are two graded L-modules, then $\wedge_\varepsilon(V \oplus W)$ (regarded as a $\mathbb{Z}\times\Gamma$-graded algebra and as a Γ-graded L-module) is canonically isomorphic to $(\wedge_\varepsilon V) \overline{\otimes}_{\varepsilon'} (\wedge_\varepsilon W)$, where $\overline{\otimes}_{\varepsilon'}$ denotes the graded tensor product of $\mathbb{Z}\times\Gamma$-graded algebras with respect to ε'. This is a special case of an analogous result for ε Lie algebras: If G and L are two ε Lie algebras, the (universal) enveloping algebra $U(G\times L)$ of the direct product $G\times L$ of G and L is canonically isomorphic to the graded tensor product $U(G) \overline{\otimes}_\varepsilon U(L)$ of the Γ-graded algebras $U(G)$ and $U(L)$ (with respect to ε). The proof follows directly from the universal property of enveloping algebras.

In particular, let $U = U_{\bar{0}} \oplus U_{\bar{1}}$ be a $\mathbb{Z}_2$-graded vector space. Applying the preceding result to $V = U_{\bar{0}}$ and $W = U_{\bar{1}}$ (regarded as $\mathbb{Z}_2$-graded vector spaces), we conclude that the super-exterior algebra $\wedge_\varepsilon U$ is canonically isomorphic to $(\wedge U_{\bar{0}}) \overline{\otimes} S(U_{\bar{1}})$, where $\wedge U_{\bar{0}}$ [resp., $S(U_{\bar{1}})$] is the *usual* exterior (resp., symmetric) algebra of the vector space $U_{\bar{0}}$ (resp., $U_{\bar{1}}$), and where $\overline{\otimes}$ denotes the graded tensor product of $\mathbb{Z}$-graded algebras with respect to the commutation factor $(r,s) \to (-1)^{rs}$. In particular, we have the vector space isomorphism

$$\overset{n}{\wedge}_\varepsilon U \simeq \bigoplus_{r=0}^{n} ((\wedge^r U_{\bar{0}}) \otimes S_{n-r}(U_{\bar{1}})). \tag{A25}$$

This enables us to make contact with the definition of the cohomology of Lie superalgebras given in Ref. 10. First, we notice that we have chosen a different overall sign for the coboundary operator (which does not matter). Moreover, there is another discrepancy, which (in our language) amounts to dropping γ in the first sum on the right hand side of Eq. (2.21). This is serious, for it spoils the L-invariance of δ^n [but not the ε-skew-symmetry of $\delta^n g$ and not the validity of the fundamental equation (2.20)]. We suspect that this is a misprint.

APPENDIX B: THE TENSORIAL SQUARE OF THE ADJOINT $sl(1|2)$-MODULE

We keep the notation of Sec. V and write L instead of $sl(1|2)$. Our goal is to investigate the graded L-module $L\otimes L$. Of course, $L\otimes L$ decomposes into the direct sum of the graded L-submodule $T_2^s(L,\varepsilon)$ of super-symmetric tensors with the graded L-submodule $T_2^a(L,\varepsilon)$ of super-skew-symmetric tensors,

$$L \otimes L = T_2^s(L,\varepsilon) \oplus T_2^a(L,\varepsilon). \tag{B1}$$

Moreover, it is known that $T_2^s(L,\varepsilon)$ [resp., $T_2^a(L,\varepsilon)$] is isomorphic to $S_2(L,\varepsilon)$ (resp., to $\wedge_\varepsilon^2 L$) (see Ref. 20 and Appendix A).

The decomposition of $T_2^a(L,\varepsilon)$ into the direct sum of simple graded submodules has already been used in Sec. V, we have

$$T_2^a(L,\varepsilon)\simeq V(\tfrac{1}{2},\tfrac{3}{2},\overline{1})\oplus V(-\tfrac{1}{2},\tfrac{3}{2},\overline{1})\oplus V(0,1,\overline{0}). \tag{B2}$$

Thus all we have to do is to investigate $T_2^s(L,\varepsilon)$. This L-module is not semi-simple. In fact, we find that

$$T_2^s(L,\varepsilon)\simeq V(0,2,\overline{0})\oplus V(0,1,\overline{0})\oplus V_8, \tag{B3}$$

where V_8 is an 8-dimensional indecomposable graded L-module, which can be described as follows. It has a basis consisting of four even weight vectors t,s,v,w and four odd weight vectors $v_\pm, w_\pm$, corresponding to the weights (0,0), (0,0), (1,0), $(-1,0)$ and $(\frac{1}{2},\pm\frac{1}{2})$, $(-\frac{1}{2},\pm\frac{1}{2})$, respectively. The action of L on V_8 is fixed by the equations

$$V_\pm\cdot t=v_\pm,\quad W_\pm\cdot t=w_\pm, \tag{B4}$$

$$V_\pm\cdot v_\mp=\pm v,\quad W_\pm\cdot w_\mp=\pm w\ , \tag{B5}$$

$$W_\pm\cdot v=v_\pm,\quad V_\pm\cdot w=w_\pm, \tag{B6}$$

$$W_\pm\cdot v_\mp=V_\pm\cdot w_\mp=\pm s\ , \tag{B7}$$

$$V_\pm\cdot s=W_\pm\cdot s=0, \tag{B8}$$

while the remaining images of the basis vectors under the action of $V_\pm$ or $W_\pm$ are equal to zero. We note that these equations fix the normalization of the basis vectors up to an overall factor. The even vectors t,s,v,w are $sl(2)$-singlets, while $(v_\pm)$ and $(w_\pm)$ are standard $sl(2)$-doublets. Obviously, t generates the L-module V_8 (but, of course, it is not a highest weight vector). Needless to say, t can be replaced by any vector of the form $t+\lambda s$, with $\lambda\in\mathbb{K}$.

We do not want to specify the eight basis vectors inside $T_2^s(L,\varepsilon)$. Suffice it to note that t can be chosen such that

$$t=Q_+\otimes Q_-+Q_-\otimes Q_++2\,Q_3\otimes Q_3+2\,B\otimes B. \tag{B9}$$

Remark B.1: The decomposition (B3) of $T_2^s(L,\varepsilon)$ has been found independently of Ref. 40. As already noted, in this article a decomposition of $U(L)$ into indecomposable L-submodules has been determined, and (B3) is just the first nontrivial part of it. The module V_8 has been known since the early days of the theory of Lie superalgebras. In fact, it was proved in Sec. 3.B of Ref. 39 that, for both $\gamma\in\mathbb{Z}_2$, we have

$$V(0,\tfrac{1}{2},\gamma)\otimes V(0,\tfrac{1}{2},\gamma)\simeq V(0,1,\overline{0})\oplus V_8, \tag{B10}$$

thus showing that the tensor product of two (typical) simple graded L-modules need not be semi-simple.

The graded L-module V_8 has the following nontrivial graded submodules:

$$V_1=\mathbb{K}s, \tag{B11}$$

$$V_4=\mathbb{K}v_+\oplus\mathbb{K}v_-\oplus\mathbb{K}v\oplus\mathbb{K}s, \tag{B12}$$

$$\overline{V_4}=\mathbb{K}w_+\oplus\mathbb{K}w_-\oplus\mathbb{K}w\oplus\mathbb{K}s, \tag{B13}$$

$$V_7=V_4+\overline{V_4}. \tag{B14}$$

Obviously, V_1 is the submodule consisting of the L-invariant elements of V_8. The modules V_4 and $\overline{V_4}$ are indecomposable, and we have

$$V_7/\overline{V_4} \simeq V_4/V_1 \simeq V(\tfrac{1}{2}, \tfrac{1}{2}, \overline{1}), \tag{B15}$$

$$V_7/V_4 \simeq \overline{V_4}/V_1 \simeq V(-\tfrac{1}{2}, \tfrac{1}{2}, \overline{1}). \tag{B16}$$

Finally, V_8/V_7 is a trivial one-dimensional L-module.

Let us next calculate $\dim H^1(L,V_8)$. Using the information obtained in Sec. V, this is easily done by means of the long cohomology sequence (2.51), applied to various short exact sequences of graded L-modules. In fact, considering the sequence

$$\{0\} \to V_1 \to V_4 \to V_4/V_1 \to \{0\}, \tag{B17}$$

and the analogous one with V_4 replaced by $\overline{V_4}$, we obtain

$$\dim H^1(L,V_4) = \dim H^1(L,\overline{V_4}) = 1, \tag{B18}$$

$$H^2(L,V_4) = H^2(L,\overline{V_4}) = \{0\}. \tag{B19}$$

Considering next the sequence

$$\{0\} \to V_4 \to V_7 \to V_7/V_4 \to \{0\}, \tag{B20}$$

we conclude that

$$\dim H^1(L,V_7) = 2. \tag{B21}$$

Finally, from the cohomology sequence corresponding to

$$\{0\} \to V_7 \to V_8 \to V_8/V_7 \to \{0\} \tag{B22}$$

and from Eq. (2.83) we deduce that

$$\dim H^1(L,V_8) = 1. \tag{B23}$$

The corresponding 1-cocycles can easily be determined. In fact, define two linear maps,

$$g: L \to V_4, \quad \overline{g}: L \to \overline{V_4}, \tag{B24}$$

by

$$g(V_\pm) = v_\pm, \quad \overline{g}(W_\pm) = w_\pm, \tag{B25}$$

$$-g(B) = \overline{g}(B) = s, \tag{B26}$$

with $g(X) = \overline{g}(Y) = 0$ for all other elements X, Y of our standard basis of L. Then g and $\overline{g}$ are 1-cocycles with values in V_4 and $\overline{V_4}$, respectively, and they are not coboundaries. If we regard g and $\overline{g}$ as 1-cocycles with values in V_7, their cohomology classes in $H^1(L,V_7)$ are linearly independent. Finally, if we regard g and $\overline{g}$ as 1-cocycles with values in V_8, they are still not coboundaries, but g and $-\overline{g}$ are cohomologous, with

$$g + \overline{g} = \delta^0 t. \tag{B27}$$

The foregoing results have an immediate bearing on the cohomology of $U(L)$. In fact, it is known[26] that $U(L)$, regarded as a graded L-module (and also as a graded coalgebra) is canonically isomorphic to $S(L,\varepsilon)$. Consequently, there exists a graded L-submodule of $U(L)$, which is a

direct summand of $U(L)$ and which is isomorphic to V_8. (Note that the simple graded subquotients of V_8 are exactly those which according to Lemma 5.1 are allowed.) Using the isomorphism (2.49), we conclude from Eq. (B23) that

$$H^1(L,U(L))\neq\{0\}. \tag{B28}$$

[1] C. Chevalley and S. Eilenberg, Trans. Am. Math. Soc. **63**, 85 (1948).
[2] J.-L. Koszul, Bull. Soc. Math. France **78**, 65 (1950).
[3] G. Hochschild and J.-P. Serre, Ann. Math. **57**, 591 (1953).
[4] H. Cartan and S. Eilenberg, *Homological Algebra* (Princeton University Press, Princeton, NJ, 1956).
[5] *Théorie des Algèbres de Lie. Topologie des Groupes de Lie*, Séminaire *Sophus Lie*, 1e année: 1954/55, École Normale Supérieure, Secrétariat mathématique, 11 rue Pierre Curie, Paris 5e, 1955.
[6] D. A. Leites, Funkt. Anal. Pril. **9**(4), 75 (1975).
[7] V. G. Kac, Adv. Math. **26**, 8 (1977).
[8] M. Scheunert, *The Theory of Lie Superalgebras*, Lecture Notes in Mathematics, Vol. 716 (Springer-Verlag, Berlin, 1979).
[9] D. B. Fuks and D. A. Leites, Dokl. Bolg. Akad. Nauk **37**, 1595 (1984).
[10] D. B. Fuks, *Cohomology of Infinite-Dimensional Lie Algebras* (Consultants Bureau, New York, 1986).
[11] D. Leites and V. Serganova, *Topological and Geometrical Methods in Field Theory*, edited by J. Mickelsson and O. Pekonen (World Scientific, Singapore, 1992), p. 194.
[12] V. S. Retakh and B. L. Feigin, Usp. Mat. Nauk **37**, 233 (1982).
[13] K. C. Tripathy and M. K. Patra, J. Math. Phys. **31**, 2822 (1990).
[14] C. Gruson, Ann. Inst. Fourier Grenoble **47**, 531 (1997).
[15] J. Tanaka, J. Math. Kyoto Univ. **35**, 733 (1995).
[16] V. Drinfeld, J. Sov. Math. **41**, 898 (1988) (expanded version of a report to the International Congress of Mathematicians, Berkeley, 1986).
[17] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985).
[18] M. Gerstenhaber, Ann. Math. **79**, 59 (1964).
[19] C. Kassel, *Quantum Groups* (Springer-Verlag, New York, 1995).
[20] M. Scheunert, J. Math. Phys. **24**, 2658 (1983).
[21] H. Garland, Inst. Hautes Études Sci. Publ. Math. **52**, 181 (1980).
[22] M. Scheunert, J. Math. Phys. **20**, 712 (1979).
[23] V. Rittenberg and D. Wyler, Nucl. Phys. B **139**, 189 (1978).
[24] S. MacLane, *Homology* (Springer-Verlag, Berlin, 1963).
[25] R. Bott and L. W. Tu, *Differential Forms in Algebraic Topology* (Springer-Verlag, New York, 1982).
[26] M. Scheunert, J. Math. Phys. **24**, 2671 (1983).
[27] P. J. Hilton and U. Stammbach, *A Course in Homological Algebra* (Springer-Verlag, New York, 1971).
[28] M. Scheunert, J. Math. Phys. **28**, 1180 (1987).
[29] D. Ž. Djoković, J. Pure Appl. Alg. **7**, 217 (1976).
[30] W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature, and Cohomology* (Academic, New York, 1976), Vol. III.
[31] M. Scheunert, *Differential Geometric Methods in Mathematical Physics*, edited by S. Sternberg (Reidel, Dordrecht, 1984), p. 115.
[32] M. Scheunert, *Differential Geometric Methods in Theoretical Physics*, edited by H. D. Doebner and J. D. Hennig (World Scientific, Singapore, 1987), p. 275.
[33] M. Scheunert, J. Math. Phys. **24**, 2681 (1983).
[34] V. G. Kac, Commun. Alg. **5**, 889 (1977).
[35] V. G. Kac, *Differential Geometrical Methods in Mathematical Physics*, Lecture Notes in Mathematics, Vol. 676, edited by K. Bleuler, H. R. Petry, and A. Reetz (Springer-Verlag, Berlin, 1978), p. 597.
[36] V. G. Kac, Proc. Natl. Acad. Sci. USA **81**, 645 (1984).
[37] A. Molev, ''Factorial supersymmetric Schur functions and super Capelli identities,'' Mathematics Research Report No. MRR 001-96, Centre for Mathematics and its Applications, Australian National University, Canberra, q-alg/9606008.
[38] M. Scheunert, J. Math. Phys. **34**, 3780 (1993).
[39] M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977).
[40] H. Benamor, Commun. Alg. **25**, 715 (1997).

Realizations of su(1,1) and $U_q(\mathrm{su}(1,1))$ and generating functions for orthogonal polynomials

J. Van der Jeugt[a] and R. Jagannathan[b]
Department of Applied Mathematics and Computer Science, University of Ghent, Krijgslaan 281-S9, B-9000 Gent, Belgium

(Received 5 March 1998; accepted for publication 3 April 1998)

Positive discrete series representations of the Lie algebra su(1,1) and the quantum algebra $U_q(\mathrm{su}(1,1))$ are considered. The diagonalization of a self-adjoint operator (the Hamiltonian) in these representations and in tensor products of such representations is determined, and the generalized eigenvectors are constructed in terms of orthogonal polynomials. Using simple realizations of su(1,1), $U_q(\mathrm{su}(1,1))$, and their representations, these generalized eigenvectors are shown to coincide with generating functions for orthogonal polynomials. The relations valid in the tensor product representations then give rise to new generating functions for orthogonal polynomials, or to Poisson kernels. In particular, a group theoretical derivation of the Poisson kernel for Meixner–Pollaczek and Al-Salam–Chihara polynomials is obtained. © *1998 American Institute of Physics.* [S0022-2488(98)01809-X]

I. INTRODUCTION

The Lie algebra su(1,1) [or so(2,1)] has been extensively used as a spectrum generating algebra in many simple quantum systems, such as the nonrelativistic Coulomb problem, the isotropic harmonic oscillator, Schrödinger's relativistic equation, and the Dirac–Coulomb problem (Ref. 1, and references therein). In some interacting boson models the Hamiltonian H can be written as a linear combination of the su(1,1) basis elements $J_0, J_\pm$, such that $H^\dagger = H$, see, e.g., the superfluid bose system.[2,3] In such boson models, the representations that play a role are the positive discrete series representations of su(1,1), often denoted by $\mathscr{D}^+(k)$. For a given Hamiltonian $H = aJ_0 + bJ_+ + cJ_- + d$, with $a^* = a$, $d^* = d$ and $b^* = c$, the important questions are: (1) its spectrum (continuous or discrete); and (2) the expansion of its eigenvectors in terms of the boson states $|k,n\rangle$ [corresponding to eigenstates of J_0 in the representations $\mathscr{D}^+(k)$].

In Ref. 4 a particular element, $H = 2J_0 - J_+ - J_-$, was chosen, and it was shown that the expansion coefficients of the eigenstates of H in the basis $|k,n\rangle$ of the representation $\mathscr{D}^+(k)$ are given as generalized Laguerre polynomials (see also Ref. 5). The main purpose of Ref. 4 was then to consider tensor product representations, and deduce in a Lie algebraic way addition or convolution formulas for orthogonal polynomials [in this case for generalized Laguerre and Jacobi polynomials, cf. Ref. 4, (1.1)]. Following this, the most general Hamiltonian with real coefficients (up to an overall constant and scaling factor) $H = \sigma J_0 - J_+ - J_-$ ($\sigma \in \mathbb{R}$) was considered in Ref. 6 and it was shown that the spectrum of H depends upon whether $|\sigma| = 2$, $|\sigma| < 2$ or $|\sigma| > 2$ (see also Ref. 7 for a particular representation). In each of these cases, the expansion coefficients of the Hamiltonian eigenstates in the $|k,n\rangle$-basis are orthogonal polynomials. In Sec. II, the main results of Refs. 4 and 6 are summarized in the present framework of eigenstates (generalized eigenvectors).

Although the physical applications or interpretations are clear from the above observations, this paper deals primarily with a number of mathematical consequences of choosing particular realizations for the positive discrete series representations.

The general framework for the positive discrete series representations is outlined in Sec. II. The three types of operators are given, together with their spectra and expansion coefficients

[a]Research Associate of the Fund for Scientific Research—Flanders (Belgium).
Electronic mail: Joris.VanderJeugt@rug.ac.be
[b]Permanent address: Institute of Mathematical Sciences, CIT Campus, Madras 600113, India.
Electronic mail: jagan@imsc.ernet.in

 © 1998 American Institute of Physics

(orthogonal polynomials). The tensor product of two positive discrete series representations is considered, and, just as there are ''coupled'' and ''uncoupled'' eigenvectors for J_0 in the tensor product, one can also define coupled and uncoupled eigenstates of the Hamiltonian. The expansion coefficients of uncoupled eigenstates into coupled eigenstates (in an irreducible component of the tensor product), cf. Eq. (2.13), are again orthogonal polynomials (Jacobi, continuous Hahn, or Hahn polynomials).

All the relations given in Sec. II (deduced in Refs. 4 and 6) are obtained by Lie algebraic methods only. The purpose of the present paper is to examine the implications of these relations once a realization for the Lie algebra su(1,1) and its representations $\mathscr{D}^+(k)$ are chosen. For su(1,1), there exists a classical realization such that the J_0-diagonal basis states $|k,n\rangle$ of $\mathscr{D}^+(k)$ are simply the monomials z^n. This realization is considered in Sec. III. As a consequence, the eigenstates of the Hamiltonian H simply become generating functions for the orthogonal polynomials (i.e., the expansion coefficients), and also the J_0-diagonal basis states in an irreducible component of the tensor product become simple functions. Then, it remains to consider the relation connecting coupled and uncoupled eigenstates of the Hamiltonian in the tensor product, and investigate its explicit form in this realization. For the case of Laguerre and Jacobi polynomials, this is done in Sec. IV; the outcome is a generalization of classical generating functions for Laguerre polynomials. For the case of Meixner and Hahn polynomials, this analysis is performed in Sec. V; the outcome is the Poisson kernel formula for Meixner (or Meixner–Pollaczek) polynomials, Eq. (5.5). Although the formula thus obtained is known, the method used to reach it is original and interesting: in this Lie algebraic framework it is a simple consequence of a very general expansion [Eq. (2.13)] in a particular realization.

Then we turn our attention to the q-analog of the results obtained so far. Since the method is purely algebraic, it works not only for the Lie algebra su(1,1) but also for the quantum algebra $U_q(\mathrm{su}(1,1))$.[8] In Sec. VI, the quantum algebra $U_q(\mathrm{su}(1,1))$ is defined, and the positive discrete series representations with a standard basis are given. Next, a self-adjoint operator playing the role of Hamiltonian is considered, and its eigenstates are expanded in the standard basis with expansion coefficients proportional to Al-Salam–Chihara polynomials. Just as for the su(1,1) case, coupled and uncoupled eigenstates of this operator in the tensor product are constructed and the expansion of one into the other is given in terms of Askey–Wilson polynomials. The results presented in Sec. VI are mainly taken from Ref. 6. Just as for su(1,1), the purpose is now to investigate the implication of these formulas once a realization is chosen. For $U_q(\mathrm{su}(1,1))$, this realization is given in Sec. VII, and the standard basis for positive discrete series representations consist again of monomials z^n. Finally, the relation connecting coupled and uncoupled eigenstates is shown to give rise to the (symmetric) Poisson kernel formula for Al-Salam–Chihara polynomials in this realization. Although this formula has recently been found by means of classical methods, the present deduction is simple, purely algebraic, and a natural q-analog of the Poisson kernel for Meixner–Pollaczek polynomials in the su(1,1) framework.

In a final section, two different realizations of su(1,1) and its representations are considered. These realizations have been used in physical models. We show again how our general results of Sec. II have interesting implications for special functions. In particular, integrals over products of Laguerre and Hermite polynomials are expressed in terms of a Meixner polynomial. This section is completed by some general comments and further outlook.

II. THE LIE ALGEBRA su(1,1) AND POSITIVE DISCRETE SERIES REPRESENTATIONS

The Lie algebra su(1,1) is generated by $J_0, J_\pm$ subject to the relations

$$[J_0, J_\pm] = \pm J_\pm, \quad [J_+, J_-] = -2J_0, \tag{2.1}$$

with the conditions $J_0^\dagger = J_0$ and $J_\pm^\dagger = J_\mp$. The positive discrete representations[9] $\mathscr{D}^+(k)$ are labeled by a positive real number k. The representation space is $\ell^2(\mathbb{Z}_+)$, with orthonormal basis vectors denoted by $e_n^{(k)}$, with $n = 0,1,2,\ldots$ (and sometimes denoted by $|k,n\rangle$). The explicit action of the generators in this representation (k) is given by

TABLE I. Orthogonal polynomials appearing in the formal eigenvector $v^{(k)}(x)$.

X	$l_n^{(k)}(x)$	Polynomial	Eigenvalue $\lambda(x)$	Spectrum
X_2	$\sqrt{\frac{n!}{(2k)_n}}\,L_n^{(2k-1)}(x)$	generalized Laguerre	x	$0\leqslant x<\infty$
X_ϕ	$\sqrt{\frac{n!}{\Gamma(2k+n)}}\,P_n^{(k)}(x;\phi)$	Meixner–Pollaczek	$2x\,\sin\phi$	$x\in\mathbb{R}$
X_c	$\sqrt{\frac{(2k)_n}{n!}}\,c^n M_n(x;2k;c^2)$	Meixner	$(c-1/c)(k+x)$	$x\in\mathbb{N}$

$$J_0 e_n^{(k)} = (n+k)e_n^{(k)},$$
$$J_+ e_n^{(k)} = \sqrt{(n+1)(2k+n)}\,e_{n+1}^{(k)}, \tag{2.2}$$
$$J_- e_n^{(k)} = \sqrt{n(2k+n-1)}\,e_{n-1}^{(k)}.$$

In two previous papers,[4,6] a recurrence operator in the Lie algebra su(1,1) was related to a Jacobi matrix. Rather than working with the spectral theory of Jacobi matrices, it will be more appropriate for this paper to use the (equivalent) notion of formal or generalized eigenvectors, to be identified with eigenstates of a Hamiltonian. In Ref. 6, the formal eigenvectors

$$v^{(k)}(x) = \sum_{n=0}^{\infty} l_n^{(k)}(x) e_n^{(k)} \tag{2.3}$$

of the self-adjoint operator $X=\sigma J_0 - J_+ - J_-$ ($\sigma\in\mathbb{R}$) in the representation (k) were studied; in a boson model such as referred to in the Introduction, the Hamiltonian H is essentially equal to this operator X. The formal vector $v^{(k)}(x)$ is a generalized eigenvector of X for the eigenvalue $\lambda(x)$ provided $l_n^{(k)}(x)$ satisfies a three-term recurrence relation.[4,6] This leads to an identification of $l_n^{(k)}(x)$ with orthogonal polynomials in x. These polynomials associated with X are of different type for $|\sigma|=2$, $|\sigma|<2$ or $|\sigma|>2$. Labelling the operator in the three distinct cases as follows:

$$X_2 = 2J_0 - J_+ - J_-, \tag{2.4}$$

$$X_\phi = -2\cos\phi J_0 + J_+ + J_-, \quad 0<\phi<\pi, \tag{2.5}$$

$$X_c = -(c+1/c)J_0 + J_+ + J_-, \quad 0<c<1, \tag{2.6}$$

we can give a summary of some results obtained earlier in Table I. The spectrum of X is determined by the support of the measure of the associated orthogonal polynomials being continuous for $|\sigma|=2$ and $|\sigma|<2$ and discrete for $|\sigma|>2$.

In Table I, $(\alpha)_n=\Gamma(\alpha+n)/\Gamma(\alpha)$ is the common notation for the Pochhammer symbol in terms of the classical Γ function. The definition of the orthogonal polynomials in terms of hypergeometric series is as follows:[10]

$$L_n^{(\alpha)}(x) = \frac{(\alpha+1)_n}{n!}\,{}_1F_1\left[\begin{matrix}-n\\ \alpha+1\end{matrix};x\right], \quad (\alpha>-1); \tag{2.7}$$

$$P_n^{(\lambda)}(x;\phi) = \frac{(2\lambda)_n}{n!}e^{in\phi}\,{}_2F_1\left[\begin{matrix}-n,\lambda+ix\\ 2\lambda\end{matrix};1-e^{-2i\phi}\right], \quad (\lambda>0); \tag{2.8}$$

$$M_n(x;\beta;c) = {}_2F_1\left[\begin{matrix}-n,-x\\ \beta\end{matrix};1-1/c\right], \quad (\beta>0). \tag{2.9}$$

The notation for hypergeometric series is the standard one.[11,12]

TABLE II. Orthogonal polynomial appearing in the expansion (2.13).

X	$S_j^{(k_1,k_2)}(x_1,x_2)$	Polynomial
X_2	$C_1(-1)^j(x_1+x_2)^j P_j^{(2k_1-1,2k_2-1)}\left(\frac{x_2-x_1}{x_2+x_1}\right)$	Jacobi
X_ϕ	$C_2(-2\sin\phi)^j p_j(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2))$	continuous Hahn
X_c	$C_3(-c+1/c)^j(-x_1-x_2)_j Q_j(x_1;2k_1-1,2k_2-1,x_1+x_2)$	Hahn

By considering the tensor product decomposition for su(1,1) representations, new summation formulas or convolution theorems were obtained for these polynomials. The tensor product decomposes as follows:[9]

$$(k_1)\otimes(k_2)=\bigoplus_{j=0}^{\infty}(k_1+k_2+j). \tag{2.10}$$

The ''coupled basis vectors'' are written in terms of the uncoupled ones by means of the Clebsch–Gordan coefficients:

$$e_n^{(k_1k_2)k}=\sum_{n_1,n_2} C_{n_1,n_2,n}^{k_1,k_2,k}\; e_{n_1}^{(k_1)}\otimes e_{n_2}^{(k_2)}. \tag{2.11}$$

Herein, $k=k_1+k_2+j$ for some integer $j\geqslant 0$, and the sum is such that $n_1+n_2=j+n$. Explicit expressions for the Clebsch–Gordan coefficients are given, e.g., in Ref. 4.

In the tensor product space, the generalized eigenvectors of $\Delta(X)=X\otimes 1+1\otimes X$ can be considered. Again, there are ''uncoupled'' eigenvectors $v^{(k_1)}(x_1)v^{(k_2)}(x_2)$ with eigenvalue $\lambda(x_1+x_2)$, but also ''coupled'' eigenvectors $v^{(k_1k_2)k}(x_1+x_2)$ defined as follows:

$$v^{(k_1k_2)k}(x_1+x_2)=\sum_{n=0}^{\infty} l_n^{(k)}(x_1+x_2)e_n^{(k_1k_2)k}, \tag{2.12}$$

where $k=k_1+k_2+j$ for some integer $j\geqslant 0$. It was shown in Refs. 4 and 6 that

$$v^{(k_1)}(x_1)v^{(k_2)}(x_2)=\sum_{j=0}^{\infty} S_j^{(k_1,k_2)}(x_1,x_2)v^{(k_1k_2)k_1+k_2+j}(x_1+x_2), \tag{2.13}$$

where S is again an orthogonal polynomial given in Table II, according to which of the three cases is considered. The constants in this table are determined by

$$C_1=(j!/((2k_1)_j(2k_2)_j(2k_1+2k_2+j-1)_j))^{1/2}, \tag{2.14}$$

$$C_2=(j!(2k_1+2k_2+2j-1)\Gamma(2k_1+2k_2+j-1)/(\Gamma(2k_1+j)\Gamma(2k_2+j)))^{1/2}, \tag{2.15}$$

$$C_3=C_2(2k_1)_j/j!. \tag{2.16}$$

The polynomials appearing in Table II are defined as follows:[10]

$$P_n^{(\alpha,\beta)}(x)=\frac{(\alpha+1)_n}{n!}\,{}_2F_1\left[\begin{matrix}-n,n+\alpha+\beta+1\\ \alpha-1\end{matrix};\frac{1-x}{2}\right], \tag{2.17}$$

$$p_n(x;a,b,c,d)=i^n\frac{(a+c)_n(a+d)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n,n+a+b+c+d-1,a+ix\\ a+c,a+d\end{matrix};1\right], \tag{2.18}$$

$$Q_n(x;a,b,N)={}_3F_2\left[\begin{matrix}-n,n+a+b+1,-x\\ a+1,-N\end{matrix};1\right]. \tag{2.19}$$

TABLE III. Explicit form of the vectors $v^{(k)}(x,z)$ in the present realization.

X	Series	Generating function
X_2	$\Sigma_{n=0}^{\infty} L_n^{(2k-1)}(x) z^n$	$(1-z)^{-2k} e^{xz/(z-1)}$
X_ϕ	$\Gamma(2k)^{-1} \Sigma_{n=0}^{\infty} P_n^{(k)}(x;\phi) z^n$	$(1-e^{i\phi}z)^{-k+ix}(1-e^{-i\phi}z)^{-k-ix}/\Gamma(2k)$
X_c	$\Sigma_{n=0}^{\infty} \frac{(2k)_n c^n}{n!} M_n(x;2k,c^2) z^n$	$(1-z/c)^x (1-cz)^{-x-2k}$

A number of interesting new convolution identities for orthogonal polynomials were constructed in Ref. 6 from the relation $(k=k_1+k_2+j)$:

$$\sum_{n_1+n_2=n+j} C^{k_1,k_2,k}_{n_1,n_2,n} \; l_{n_1}^{(k_1)}(x_1) l_{n_2}^{(k_2)}(x_2) = l_n^{(k)}(x_1+x_2) S_j^{(k_1,k_2)}(x_1,x_2). \tag{2.20}$$

In the simplest case, where $l_n^{(k)}(x)$ corresponds to a generalized Laguerre polynomial and $S_j^{(k_1,k_2)}(x_1,x_2)$ to a Jacobi polynomial, this relation reduces to Ref. 4, (1.1). Recently, this equation has been applied to find transformation brackets for $U(N)$ boson models.[13]

III. REALIZATION OF su(1,1)

In this section we give a realization of su(1,1) and of the discrete series representation $\mathscr{D}^+(k)$. Assume that $k>1/2$, and consider the Hilbert space of analytic functions $f(z)$ $(z\in\mathbb{C})$ on the unit disc $|z|<1$, with inner product[9]

$$(f_1,f_2)=\frac{2k-1}{\pi}\int\int_{|z|<1} f_1(z)\overline{f_2(z)}(1-|z|^2)^{2k-2}\,dxdy \quad (z=x+iy). \tag{3.1}$$

The following functions form an orthonormal basis:

$$e_n^{(k)} \equiv e_n^{(k)}(z) = \sqrt{\frac{(2k)_n}{n!}}\, z^n, \tag{3.2}$$

since $(e_n^{(k)}, e_m^{(k)}) = \delta_{m,n}$. The realization of the su(1,1) basis elements reads as follows:

$$J_0 = z\frac{d}{dz} + k, \quad J_- = \frac{d}{dz}, \quad J_+ = z^2\frac{d}{dz} + 2kz. \tag{3.3}$$

It is easy to verify that the action of these operators on the basis (3.2) is indeed the same as in (2.2).

Next, consider the formal vectors (2.3), which are now denoted by $v^{(k)}(x,z)$ since in this realization the basis vectors become functions of z. Since the vectors $e_n^{(k)}$ are proportional to z^n, one can see that the $v^{(k)}(x,z)$ are in fact generating functions for the polynomials $l_n^{(k)}(x)$. The explicit forms of the vectors is given in Table III, for each of the three cases considered here.

For each of the cases, the generating function can be found in Ref. 10. One can also find it by explicitly solving the first order differential equation

$$Xv^{(k)}(x,z) = \lambda(x) v^{(k)}(x,z), \tag{3.4}$$

where X is one of (2.4)–(2.6), in the realization (3.3).

Next, consider the tensor product of two representations $(k_1)\otimes(k_2)$ in the present realization, and the coupled vector

$$e_n^{(k_1k_2)k}(z_1,z_2) = \sum_{n_1,n_2} C^{k_1,k_2,k}_{n_1,n_2,n} \; e_{n_1}^{(k_1)}(z_1) e_{n_2}^{(k_2)}(z_2). \tag{3.5}$$

Using properties of the generating function for the Clebsch–Gordan coefficients, this expression can be written as follows:

$$e_n^{(k_1k_2)k}(z_1,z_2)=\left[\frac{(2k_1)_j(2k_2)_j(2k_1+2k_2+2j)_n}{j!n!(2k_1+2k_2+j-1)_j}\right]^{1/2}(-z_1+z_2)^j z_1^n\,{}_2F_1\left[\begin{matrix}-n,2k_2+j\\2k_1+2k_2+2j\end{matrix};1-z_2/z_1\right], \tag{3.6}$$

where, as before, $k=k_1+k_2+j$.

In the following sections we shall consider the explicit forms of (2.12) and (2.13) in this realization, and its consequences.

IV. THE CASE OF JACOBI POLYNOMIALS

Consider Eq. (2.13) in the case $X=X_2$,

$$v^{(k_1)}(x_1,z_1)v^{(k_2)}(x_2,z_2)=\sum_{j=0}^{\infty} S_j^{(k_1,k_2)}(x_1,x_2)v^{(k_1k_2)k_1+k_2+j}(x_1+x_2,z_1,z_2); \tag{4.1}$$

in this situation the polynomials S_j are Jacobi polynomials. Using the formulas of Table II, and the new variables

$$s=x_2+x_1, \quad x_1=s(1-r)/2,$$

$$r=(x_2-x_1)/(x_2+x_1), \quad x_2=s(1+r)/2, \tag{4.2}$$

this equation becomes

$$v^{(k_1)}(s(1-r)/2,z_1)v^{(k_2)}(s(1+r)/2,z_2)$$
$$=\sum_{j=0}^{\infty} C_1(-1)^j s^j P_j^{(2k_1-1,2k_2-1)}(r)v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2). \tag{4.3}$$

Now we can integrate with respect to the weight function of Jacobi polynomials:

$$\int_{-1}^{1} P_j^{(2k_1-1,2k_2-1)}(r)v^{(k_1)}\left(\frac{s(1-r)}{2},z_1\right)v^{(k_2)}\left(\frac{s(1+r)}{2},z_2\right)(1-r)^{2k_1-1}(1+r)^{2k_2-1}\,dr$$
$$=C_1h_j(-1)^j s^j v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2), \tag{4.4}$$

where h_j is the norm of the Jacobi polynomials of degree j, i.e.,

$$h_j=\frac{2^{2k_1+2k_2-1}}{2k_1+2k_2+2j-1}\,\frac{\Gamma(2k_1+j)\Gamma(2k_2+j)}{j!\Gamma(2k_1+2k_2+j-1)}. \tag{4.5}$$

Using the expressions for the functions v given in Table III, we obtain

$$C_1h_j(-1)^j s^j v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2)$$
$$=(1-z_1)^{-2k_1}(1-z_2)^{-2k_2}e^{(s/2)\{[z_1/(z_1-1)]+[z_2/(z_2-1)]\}}$$
$$\times\int_{-1}^{1} P_j^{(2k_1-1,2k_2-1)}(r)e^{j(z_1-z_2)/2(z_1-1)(z_2-1)}(1-r)^{2k_1-1}(1+r)^{2k_2-1}\,dr. \tag{4.6}$$

Lemma 1:

$$I=\int_{-1}^{1}(1-r)^a(1+r)^b P_j^{(a,b)}(r)e^{cr}\,dr$$
$$=2^{a+b+1}\frac{\Gamma(a+j+1)\Gamma(b+j+1)}{j!\Gamma(a+b+2j+2)}e^{-c}(2c)^j\ {}_1F_1\left[\begin{matrix}b+j+1\\ a+b+2j+2\end{matrix};2c\right]. \tag{4.7}$$

Proof: Writing

$$P_j^{(a,b)}(r)=\frac{1}{2^j}\sum_{m=0}^{j}\binom{j+a}{m}\binom{j+b}{j-m}(r-1)^{j-m}(r+1)^m, \tag{4.8}$$

and using

$$\int_{-1}^{1}(1-r)^\alpha(1+r)^\beta e^{\gamma r}\,dr=2^{\alpha+\beta+1}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}\,e^{-\gamma}\ {}_1F_1\left[\begin{matrix}\beta+1\\ \alpha+\beta+2\end{matrix};2\gamma\right], \tag{4.9}$$

one obtains, after some simplifications,

$$I=(-1)^j 2^{a+b+1}e^{-c}\frac{\Gamma(a+j+1)\Gamma(b+j+1)}{j!\Gamma(a+b+j+2)}\sum_{j=0}^{m}\frac{(-j)_m}{m!}\ {}_1F_1\left[\begin{matrix}b+m+1\\ a+b+j+2\end{matrix};2c\right]. \tag{4.10}$$

In the sum over m, the ${}_1F_1$ is written explicitly as a series, the two summation variables are changed, and then Vandermonde's summation theorem can be applied; putting $\alpha=a+b+j+2$ and $\beta=b+1$, this gives explicitly

$$\sum_{j=0}^{m}\frac{(-j)_m}{m!}\ {}_1F_1\left[\begin{matrix}\beta+m\\ \alpha\end{matrix};\gamma\right]=\sum_{m=0}^{j}\frac{(-j)_m}{m!}\sum_{k=0}^{\infty}\frac{(\beta+m)_k\gamma^k}{(\alpha)_k k!}$$
$$=\sum_{k=0}^{\infty}\frac{(\beta)_k\gamma^k}{(\alpha)_k k!}\sum_{m=0}^{j}\frac{(b+k)_m(-j)_m}{(b)_m m!}$$
$$=\sum_{k=0}^{\infty}\frac{(\beta)_k\gamma^k}{(\alpha)_k k!}\frac{(-k)_j}{(b)_j}. \tag{4.11}$$

In the last sum k goes from j up to ∞, so let $l=k-j$, and rewrite the series; one finds

$$\frac{(-1)^j\gamma^j}{(\alpha)_j}\sum_{l=0}^{\infty}\frac{(\beta+j)_l\gamma^l}{(\alpha+j)_l l!}=\frac{(-1)^j\gamma^j}{(\alpha)_j}\ {}_1F_1\left[\begin{matrix}\beta+j\\ \alpha+j\end{matrix};\gamma\right]. \tag{4.12}$$

Putting this back in (4.10) finally proves the lemma.

As a consequence, we now have the following explicit form for $v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2)$:

$$v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2)=(-1)^j\left[\frac{(2k_1)_j(2k_2)_j}{j!(2k_1+2k_2+j-1)_j}\right]^{1/2}(1-z_1)^{-2k_1-j}(1-z_2)^{-2k_2-j}$$
$$\times(z_1-z_2)^j e^{sz_1/(z_1-1)}\ {}_1F_1\left[\begin{matrix}2k_1+j\\ 2k_1+2k_2+2j\end{matrix};\frac{s(z_1-z_2)}{(z_1-1)(z_2-1)}\right]. \tag{4.13}$$

On the other hand, we know from (2.12) that ($k=k_1+k_2+j$)

$$v^{(k_1k_2)k}(s,z_1,z_2)=\sum_{n=0}^{\infty}l_n^{(k)}(s)e_n^{(k_1k_2)k}(z_1,z_2), \tag{4.14}$$

with $l_n^{(k)}(s)$ given in Table I and $e_n^{(k_1k_2)k}(z_1,z_2)$ given in (3.6). Thus we obtain

$$\sum_{n=0}^{\infty} L_n^{(2k)}(s) z_1^n \, {}_2F_1\left[\begin{matrix} -n, 2k_2+j \\ 2k_1+2k_2+2j \end{matrix}; 1-z_2/z_1\right]$$

$$= (1-z_1)^{-2k_1-j}(1-z_2)^{-2k_2-j} e^{sz_1/(z_1-1)} \, {}_1F_1\left[\begin{matrix} 2k_1+j \\ 2k_1+2k_2+2j \end{matrix}; \frac{s(z_1-z_2)}{(z_1-1)(z_2-1)}\right]. \quad (4.15)$$

This can be rewritten in a more appropriate form:

Proposition 2:

$$\sum_{n=0}^{\infty} L_n^{(b-1)}(s) z_1^n \, {}_2F_1\left[\begin{matrix} -n, a \\ b \end{matrix}; 1-\frac{z_2}{z_1}\right] = (1-z_1)^{a-b}(1-z_2)^{-a} e^{sz_1/(z_1-1)} \, {}_1F_1\left[\begin{matrix} a \\ b \end{matrix}; \frac{s(z_1-z_2)}{(z_1-1)(z_2-1)}\right]. \quad (4.16)$$

This is a generalization of two classical generating functions for the Laguerre polynomials. For $a=0$, or $z_1=z_2$, it reduces to the first classical result

$$\sum_{n=0}^{\infty} L_n^{(b-1)}(s) z_1^n = (1-z_1)^{-b} e^{sz_1/(z_1-1)}; \quad (4.17)$$

for $z_2=0$, it reduces to

$$\sum_{n=0}^{\infty} L_n^{(b-1)}(s) \frac{(b-a)_n}{(b)_n} z_1^n = (1-z_1)^{-b} \, {}_1F_1\left[\begin{matrix} b-a \\ b \end{matrix}; \frac{sz_1}{(z_1-1)}\right]. \quad (4.18)$$

Note that (4.16) can be written in a more general form:

$$\sum_{n=0}^{\infty} {}_1F_1\left[\begin{matrix} -n \\ b \end{matrix}; x\right] \, {}_2F_1\left[\begin{matrix} -n, a \\ b \end{matrix}; y\right] \frac{(b)_n}{n!} z^n$$

$$= (1-z)^{a-b}(1-z+yz)^{-a} e^{xz/(z-1)} \, {}_1F_1\left[\begin{matrix} a \\ b \end{matrix}; \frac{xyz}{(1-z)(1-z+yz)}\right]. \quad (4.19)$$

If one is not interested in the particular form of $v^{(k_1k_2)k_1+k_2+j}(s,z_1,z_2)$, but only in comparing the two expressions of this vector in order to obtain (4.15), many simplifications take place. In particular, note that (4.15) does not explicitly depend upon j but only upon $2k_1+j$ and $2k_2+j$. Thus, one can obtain the same formula (4.16) just by working out only the vector with $j=0$. This will be done for the remaining cases in the following sections.

V. THE CASE OF MEIXNER AND MEIXNER–POLLACZEK POLYNOMIALS

Consider the case $X=X_c$; then (4.1) holds with S_j proportional to a Hahn polynomial. First, assume that $s=x_1+x_2$ is a positive integer (later this condition will disappear by analytic continuation). We can then apply the (discrete) orthogonality relation for Hahn polynomials, and obtain the $j=0$ component

$$v^{(k_1k_2)k_1+k_2}(s,z_1,z_2) \frac{(2k+1+2k_2-1)_{s+1}}{s!(2k_1+2k_2-1)}$$

$$= \sum_{x_1=0}^{s} \frac{(2k_1)_{x_1}(2k_2)_{s-x_1}}{x_1!(s-x_1)!} v^{(k_1)}(x_1,z_1) v^{(k_2)}(s-x_1,z_2). \quad (5.1)$$

With the explicit forms for $v^{(k_1)}$ and $v^{(k_2)}$ from Table III, this gives rise to

$$v^{(k_1k_2)k_1+k_2}(s,z_1,z_2) = (1-cz_1)^{-2k_1}(1-z_2/c)^s(1-cz_2)^{-s-2k_2} \, {}_2F_1\left[\begin{matrix} -s, 2k_1 \\ 2k_1+2k_2 \end{matrix}; 1-\frac{z_2}{z_1}\right]. \quad (5.2)$$

On the other hand, (4.14) now becomes

$$v^{(k_1k_2)k_1+k_2}(s,z_1,z_2)=\sum_{n=0}^{\infty}\frac{(2k_1+2k_2)_n}{n!}M_n(s;2k_1+2k_2,c^2)c^nz_1^n\,{}_2F_1\left[\begin{matrix}-n,2k_2\\2k_1+2k_2\end{matrix};1-\frac{z_2}{z_1}\right]. \tag{5.3}$$

Applying a series transformation on the last ${}_2F_1$ yields

$$\sum_{n=0}^{\infty}\frac{(b)_n}{n!}M_n(s;b,c^2)c^nz_2^n\,{}_2F_1\left[\begin{matrix}-n,a\\b\end{matrix};1-\frac{z_2}{z_1}\right]$$

$$=(1-cz_1)^{-a}(1-cz_2)^{a-b}\left(\frac{1-z_2/c}{1-cz_2}\right)^s\,{}_2F_1\left[\begin{matrix}-s,a\\b\end{matrix};\frac{(c-1/c)(z_2-z_1)}{(1-cz_1)(1-z_2/c)}\right]. \tag{5.4}$$

Given the Meixner polynomials in terms of a ${}_2F_1$, this leads to the following series expression:

Proposition 3:

$$\sum_{n=0}^{\infty}{}_2F_1\left[\begin{matrix}-n,a\\c\end{matrix};x\right]{}_2F_1\left[\begin{matrix}-n,b\\c\end{matrix};y\right]\frac{(c)_nz^n}{n!}$$

$$=(1-z+xz)^{-a}(1-z+yz)^{-b}(1-z)^{a+b-c}\,{}_2F_1\left[\begin{matrix}a,b\\c\end{matrix};\frac{xyz}{(1-z+xz)(1-z+yz)}\right]. \tag{5.5}$$

This formula was already given in Ref. 14, p. 85, eq. (12), and can be interpreted as the Poisson kernel for Meixner or Meixner–Pollaczek polynomials.[15] Observe that (4.19) is a limiting case of (5.5): putting $x=x'/a$ in (5.5) and taking the limit $a\to\infty$ leads to (4.19).

The case of Meixner–Pollaczek polynomials leads essentially to the same formulas.

VI. THE ALGEBRA $U_q(\mathrm{su}(1,1))$

Let $U_q(\mathrm{sl}(2,\mathbb{C}))$ $(0<q<1)$ be the complex unital associative algebra generated by A,B,C,D subject to the relations

$$AD=1=DA,\quad AB=q^{1/2}BA,\quad AC=q^{-1/2}CA,\quad BC-CB=\frac{A^2-D^2}{q^{1/2}-q^{-1/2}}. \tag{6.1}$$

(Compared to Ref. 6, q is replaced by $q^{1/2}$ in order to have q as the basis of the basic hypergeometric series appearing later.) This algebra can be equipped with a comultiplication, counit, and antipode, turning it into a Hopf algebra.[16,8] The Hopf $*$-algebra $U_q(\mathrm{su}(1,1))$ has the following $*$-structure:

$$A^*=A,\quad B^*=-C,\quad C^*=-B,\quad D^*=D. \tag{6.2}$$

The positive discrete representations are labeled by a positive real number k. The representation space is again $\ell^2(\mathbb{Z}_+)$, with orthonormal basis vectors denoted by $e_n^{(k)}$, with $n=0,1,2,\ldots$. The explicit action of the generators in this representation (k) is given by[6]

$$Ae_n^{(k)}=q^{(k+n)/2}e_n^{(k)}, \tag{6.3}$$

$$Ce_n^{(k)}=q^{(1-2k-2n)/4}\frac{\sqrt{(1-q^n)(1-q^{2k+n-1})}}{q^{1/2}-q^{-1/2}}\,e_{n-1}^{(k)}, \tag{6.4}$$

$$Be_n^{(k)}=q^{-(1+2k+2n)/4}\frac{\sqrt{(1-q^{n+1})(1-q^{2k+n})}}{q^{-1/2}-q^{1/2}}\,e_{n+1}^{(k)}. \tag{6.5}$$

Let $s\in\mathbb{R}\backslash\{0\}$, and define

$$Y_s = q^{1/4}B - q^{-1/4}C + \frac{s^{-1}+s}{q^{-1/2}-q^{1/2}}(A-D). \tag{6.6}$$

Under the comultiplication

$$\begin{aligned} &\Delta(A) = A\otimes A, \quad \Delta(B) = A\otimes B + B\otimes D, \\ &\Delta(C) = A\otimes C + C\otimes D, \quad \Delta(D) = D\otimes D, \end{aligned} \tag{6.7}$$

Y_s is twisted primitive, and Y_sA is a self-adjoint element. Here, Y_sA is playing the role of the operator X for su(1,1). In a previous paper,[6] the formal eigenvectors

$$v^{(k)}(x) = \sum_{n=0}^{\infty} l_n^{(k)}(x) e_n^{(k)} \tag{6.8}$$

of Y_sA have been obtained. In order to express the results of Ref. 6, recall the definition of Askey–Wilson[17] and Al-Salam and Chihara polynomials.[18] The Askey–Wilson polynomial is defined by ($x = \cos\theta$)

$$p_m(x;a,b,c,d|q) = a^{-m}(ab,ac,ad;q)_m \, {}_4\varphi_3\left[\begin{matrix} q^{-m}, abcdq^{m-1}, ae^{i\theta}, ae^{-i\theta} \\ ab, ac, ad \end{matrix}; q, q\right], \tag{6.9}$$

and it is symmetric in its parameters a, b, c, and d. In (6.9), the notation for basic hypergeometric series and for the shifted q-factorial is taken from Ref. 12. The Al-Salam and Chihara polynomials are obtained by taking $c = d = 0$ in the Askey–Wilson polynomials:

$$s_m(x;a,b|q) = p_m(x;a,b,0,0|q) = a^{-m}(ab;q)_m \, {}_3\varphi_2\left[\begin{matrix} q^{-m}, ae^{i\theta}, ae^{-i\theta} \\ ab, 0 \end{matrix}; q, q\right]. \tag{6.10}$$

Using the abbreviation $\mu(x) = (x + x^{-1})/2$, it was shown in Ref. 6 that for

$$l_n^{(k)}(x) = \frac{1}{\sqrt{(q,q^{2k};q)_n}} s_n(\mu(x); q^k s, q^k/s|q), \tag{6.11}$$

the vectors (6.8) are formal eigenvectors of Y_sA for the eigenvalue

$$\lambda(x) = 2(\mu(s) - \mu(x))/(q^{1/2} - q^{-1/2}).$$

The tensor product of two $U_q(\mathrm{su}(1,1))$ representations $(k_1)\otimes(k_2)$ is the same as in (2.10), and the Clebsch–Gordan coefficients in

$$e_n^{(k_1k_2)k} = \sum_{n_1,n_2} C_{n_1,n_2,n}^{k_1,k_2,k} \; e_{n_1}^{(k_1)} \otimes e_{n_2}^{(k_2)} \tag{6.12}$$

have been determined, e.g., in Ref. 6. In the tensor product space, the generalized eigenvectors of $\Delta(Y_sA)$ have been considered, both in coupled and uncoupled form. It was shown that

$$v^{k_1;k_2}(x_1,x_2) = \sum_{n_1,n_2} \frac{s_{n_1}(\mu(x_1); q^{k_1}x_2, q^{k_1}/x_2|q)}{\sqrt{(q,q^{2k_1};q)_{n_1}}} \frac{s_{n_2}(\mu(x_2); q^{k_2}s, q^{k_2}/s|q)}{\sqrt{(q,q^{2k_2};q)_{n_2}}} e_{n_1}^{(k_1)} \otimes e_{n_2}^{(k_2)} \tag{6.13}$$

is a generalized eigenvector of $\Delta(Y_sA)$ for the eigenvalue $\lambda(x_1)$. In ''coupled'' form,

$$v^{(k_1k_2)k}(x_1) = \sum_{n=0}^{\infty} \frac{s_n(\mu(x_1); q^k s, q^k/s|q)}{\sqrt{(q,q^{2k};q)_n}} e_n^{(k_1k_2)k} \tag{6.14}$$

is a generalized eigenvector of $\Delta(Y_s A)$ for the same eigenvalue. The relation between the two generalized eigenvectors was given in Ref. 6:

$$v^{k_1;k_2}(x_1,x_2)=\sum_{j=0}^{\infty} C_j p_j(\mu(x_2);q^{k_1}x_1,q^{k_1}/x_1,q^{k_2}s,q^{k_2}/s|q)v^{(k_1k_2)k_1+k_2+j}(x_1), \quad (6.15)$$

where p_j is an Askey–Wilson polynomial and

$$C_j=((q,q^{2k_1},q^{2k_2},q^{2k_1+2k_2+j-1};q)_j)^{-1/2}. \quad (6.16)$$

VII. REALIZATION OF $U_q(\mathrm{su}(1,1))$

A realization of $U_q(\mathrm{su}(1,1))$ and the representation labelled by (k) is closely related to the one given for su(1,1) in Sec. III. The Hilbert space is the same (up to a rescaling of the inner product), and the basis functions are now

$$e_n^{(k)}\equiv e_n^{(k)}(z)=\sqrt{\frac{(q^{2k};q)_n}{(q;q)_n}}\,z^n. \quad (7.1)$$

The realization of the $U_q(\mathrm{su}(1,1))$ generators is given in terms of the operators

$$T_q f(z)=f(qz), \quad D_q=\frac{1-T_q}{(1-q)z}, \quad (7.2)$$

and reads

$$A=q^{k/2}T_{q^{1/2}}, \quad D=q^{-k/2}T_{q^{-1/2}}, \quad (7.3)$$

$$B=q^{(1-2k)/4}\left(z^2 D_q T_{q^{-1/2}}+\frac{1-q^{2k}}{1-q}zT_{q^{1/2}}\right), \quad (7.4)$$

$$C=-q^{(3-2k)/4}D_q T_{q^{-1/2}}. \quad (7.5)$$

For this realization, the formal vectors (6.8) become

$$v^{(k)}(x,z)=\sum_{n=0}^{\infty} s_n(\mu(x);q^k s,q^k/s|q)\,\frac{z^n}{(q;q)_n} \quad (7.6)$$

$$=\frac{(q^k zs,q^k z/s;q)_\infty}{(zx,z/x;q)_\infty}, \quad (7.7)$$

following from a known generating function for the Al-Salam and Chihara polynomials.[10] We also wish to obtain an explicit form for the coupled vectors $e_n^{(k_1k_2)k}(z_1,z_2)$. Using the proof of Lemma 4.4 of Ref. 6, one finds $(k=k_1+k_2+j)$

$$e_n^{(k_1k_2)k}(z_1,z_2)=q^{-nj-nk_1}z_2^{n+j}(q^{k_1}z_1/z_2;q)_j\sqrt{\frac{(q^{2k_1},q^{2k_2};q)_j}{(q,q^{2k_1+2k_2+j-1};q)_j}}$$
$$\times\sqrt{\frac{(q^{2k_1+2k_2+2j};q)_n}{(q;q)_n}}\,{}_3\varphi_2\left[\begin{matrix}q^{-n},q^{2k_1+j},q^{k_1+j}z_1/z_2\\ q^{2k_1+2k_2+2j},0\end{matrix};q,q\right]. \quad (7.8)$$

VIII. POISSON KERNEL FOR AL-SALAM–CHIHARA POLYNOMIALS

In the realization of the previous section, the formal vectors defined in (6.13) can be rewritten using (7.7) and they become

$$v^{k_1;k_2}(x_1,x_2,z_1,z_2)=\frac{(q^{k_1}z_1x_2,q^{k_1}z_1/x_2,q^{k_2}z_2s,q^{k_2}z_2/s;q)_\infty}{(z_1x_1,z_1/x_1,z_2x_2,z_2/x_2;q)_\infty}. \tag{8.1}$$

Note that in all the previous formulas the variables of the Al-Salam and Chihara (or Askey–Wilson) polynomials are $\mu(x_1)$ or $\mu(x_2)$; we can represent $\mu(x_l)=\cos\theta_l$ and thus $x_l=e^{i\theta_l}$ for some real θ_l $(l=1,2)$. The purpose is now to integrate Eq. (6.15) so as to obtain an explicit form for $v^{(k_1k_2)k_1+k_2+j}(x_1,z_1,z_2)$ for $j=0$. Let

$$(a,b,c,d)=(q^{k_1}x_1,q^{k_1}/x_1,q^{k_2}s,q^{k_2}/s). \tag{8.2}$$

If we assume that $q^{k_2}<|s|<q^{-k_2}$, then $\max\{|a|,|b|,|c|,|d|\}<1$, and the Askey–Wilson polynomial in (6.15) has an absolute continuous measure. Using the notation

$$h(x;a)=h(x;a;q)=(ae^{i\theta},ae^{-i\theta};q)_\infty \quad \text{for } x=\cos\theta, \tag{8.3}$$

and

$$h(x;a,b,c,d)=h(x;a)h(x;b)h(x;c)h(x;d), \tag{8.4}$$

the weight function for the Askey–Wilson polynomials reads (Ref. 12, §7.5)

$$w(x)\equiv w(x;a,b,c,d)=\frac{h(x;1,-1,q^{1/2},q^{-1/2})}{\sqrt{1-x^2}h(x;a,b,c,d)}. \tag{8.5}$$

The orthogonality relation for $p_n(x)\equiv p_n(x;a,b,c,d|q)$ is

$$\int_{-1}^{1}p_m(x)p_n(x)w(x)dx=\frac{\delta_{m,n}}{h_n}, \tag{8.6}$$

where

$$h_n=h_0\frac{(abcdq^{-1};q)_n(1-abcdq^{2n-1})}{(1-abcdq^{-1})(q,ab,ac,ad,bc,bd,cd;q)_n}, \tag{8.7}$$

and

$$h_0=\frac{(q,ab,ac,ad,bc,bd,cd;q)_\infty}{2\pi(abcd;q)_\infty}. \tag{8.8}$$

It thus follows from (6.15) that

$$v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)=h_0\int_{-1}^{1}w(\mu(x_2);a,b,c,d)v^{k_1;k_2}(x_1,x_2,z_1,z_2)d\mu(x_2). \tag{8.9}$$

Using the explicit form (8.1), the notation

$$f=z_2, \quad g=q^{k_1}z_1, \tag{8.10}$$

and writing $\mu(x_2)=t$, (8.9) becomes

$$v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)=h_0\frac{(cf,df;q)_\infty}{(g/a,g/b;q)_\infty}\int_{-1}^{1}\frac{h(t;g)}{h(t;f)}w(t;a,b,c,d)dt. \tag{8.11}$$

The last integral is a known one; in Ref. 12 it is denoted by $J(a,b,c,d,f,g)$, and shown to be equal to a ${}_8W_7$ series, i.e., a very-well-poised ${}_8\varphi_7$ series. Using eq. (6.3.8) of Ref. 12, we obtain

$$v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)$$
$$=\frac{(ag,bg,cg,df,abcf;q)_\infty}{(af,bf,g/a,g/b,abcg;q)_\infty}\,{}_8W_7(abcgq^{-1};ab,ac,bc,g/d,g/f;q,df). \quad (8.12)$$

Let us now consider the other form of $v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)$, given by (6.14). Using (7.8) and definition (6.10), one finds

$$\begin{aligned} v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)&=\sum_{n=0}^{\infty}\frac{s_n(\mu(x_1);q^{k_1+k_2}s,q^{k_1+k_2}/s|q)}{\sqrt{(q,q^{2k_1+2k_2};q)_n}}e_n^{(k_1k_2)k_1+k_2}(z_1,z_2)\\ &=\sum_{n=0}^{\infty}\frac{(q^{2k_1+2k_2};q)_n}{(q;q)_n}(q^{-2k_1-k_2}z_2/s)^n\\ &\quad\times{}_3\varphi_2\left[\begin{matrix}q^{-n},q^{k_1+k_2}sx_1,q^{k_1+k_2}s/x_1\\ q^{2k_1+2k_2},0\end{matrix};q,q\right]\\ &\quad\times{}_3\varphi_2\left[\begin{matrix}q^{-n},q^{2k_1},q^{k_1}z_1/z_2\\ q^{2k_1+2k_2},0\end{matrix};q,q\right], \end{aligned} \quad (8.13)$$

or, putting all this in the notation of (8.2) and (8.10),

$$v^{(k_1k_2)k_1+k_2}(x_1,z_1,z_2)=\sum_{n=0}^{\infty}\frac{(abcd;q)_n}{(q;q)_n}\left(\frac{f}{abc}\right)^n{}_3\varphi_2\left[\begin{matrix}q^{-n},ac,ad\\ abcd,0\end{matrix};q,q\right]{}_3\varphi_2\left[\begin{matrix}q^{-n},ab,g/f\\ abcd,0\end{matrix};q,q\right]. \quad (8.14)$$

Equating (8.14) and (8.12) gives the q-analog of (5.5). Relabelling all the variables, it can be written in the following form:

Proposition 4:

$$\sum_{n=0}^{\infty}{}_3\varphi_2\left[\begin{matrix}q^{-n},a,b\\ f,0\end{matrix};q,q\right]{}_3\varphi_2\left[\begin{matrix}q^{-n},c,d\\ f,0\end{matrix};q,q\right]\frac{(f;q)_n}{(q;q)_n}z^n$$
$$=\frac{(abcz,abdz,acdz,bcdz,fz;q)_\infty}{(acz,bcz,adz,bdz,abcdz;q)_\infty}\,{}_8W_7(abcdzq^{-1};a,b,c,d,abcdz/f;q,fz). \quad (8.15)$$

This formula is the (symmetric) Poisson kernel for Al-Salam–Chihara polynomials, and was derived earlier by classical methods in Ref. 19, (14.8), and Ref. 20. As far as we know, this is the first group theoretical derivation of it.

IX. OTHER su(1,1) REALIZATIONS AND THEIR CONSEQUENCES

In this section two different realizations of su(1,1) and its positive discrete series representations will be considered. For the realization considered so far, the typical feature is that the basis functions $e_n^{(k)}$ are simply monomials z^n. Here, the basis functions will have a more complicated form, and the q-analog will not be treated.

Before turning to the realization, it is useful to observe some realization-independent facts. In Sec. II, the operator X_c $(0<c<1)$ was introduced, and from Table I one can see that its spectrum is identical to that of J_0 apart from the factor $(c-1/c)$. So one can expect that X_c and J_0 are related through a unitary transformation, and this is indeed the case. For $0<c<1$, let $\alpha>0$ be defined through

$$e^{\alpha}=\frac{1+c}{1-c}, \quad (9.1)$$

or, inversely,

$$c=\frac{e^{\alpha}-1}{e^{\alpha}+1}. \tag{9.2}$$

Defining, as usual,

$$J_1=\frac{1}{2}(J_+ + J_-), \quad J_2=\frac{1}{2i}(J_+ - J_-) \tag{9.3}$$

in terms of the su(1,1) basis (2.1), it is well known that the following identity holds [Ref. 1, eq. (97)]:

$$\exp(i\alpha J_2)J_0 \exp(-i\alpha J_2)=(\cosh\ \alpha)J_0-(\sinh\ \alpha)J_1. \tag{9.4}$$

This equation can also be obtained by applying the ''unitary trick'' $J_x\to iJ_1$, $J_y\to iJ_2$, $J_z\to J_0$, $\theta\to i\alpha$ to the su(2) relation $\exp(-i\theta J_y)J_z \exp(i\theta J_y)=(\sin\theta)J_x+(\cos\theta)J_z$ [see Ref. 21, eq. (D.10)]. Then it follows from (2.6) and the above relation between c and α that

$$X_c=(c-1/c)\exp(i\alpha J_2)J_0 \exp(-i\alpha J_2). \tag{9.5}$$

Acting with this equation on the formal eigenvectors $v^{(k)}(m)$ of X_c yields

$$J_0 \exp(-i\alpha J_2)v^{(k)}(m)=(k+m)\exp(-i\alpha J_2)v^{(k)}(m). \tag{9.6}$$

Thus, up to a normalization factor N_m, these elements must coincide with the eigenvectors $e_m^{(k)}$ of J_0, leading to

$$\exp(i\alpha J_2)e_m^{(k)}=N_m v^{(k)}(m). \tag{9.7}$$

In other words, using (2.3), Table I, and the orthogonality for Meixner polynomials, this yields

$$\exp(i\alpha J_2)e_m^{(k)}=\phi_m\sum_{n=0}^{\infty}(1-c^2)^k c^{m+n}\sqrt{\frac{(2k)_m(2k)_n}{m!n!}}\,M_n(m;2k;c^2)e_n^{(k)}, \tag{9.8}$$

where ϕ_m is a phase factor ($|\phi_m|=1$). This relation is realization independent, and for particular realizations it leads to interesting identities. First it should be noted that for the realization of Sec. III, the explicit form of $\exp(i\alpha J_2)e_m^{(k)}$ becomes rather complicated and (9.8) does not reduce to a simple relation. Here, we shall give two different realizations for which (9.8) implies an interesting relation.

First, let $w\geqslant 1$, $r\in(0,\infty)$ and $p_r=-i(d/dr+1/r)$. In Ref. 1 eq. (75), the following realization of su(1,1) is considered:

$$\begin{aligned} J_1&=\tfrac{1}{2}(w^{-2}r^{2-w}p_r^2+\xi r^{-w}-r^w),\\ J_2&=w^{-1}(rp_r-i(w-1)/2),\\ J_0&=\tfrac{1}{2}(w^{-2}r^{2-w}p_r^2+\xi r^{-w}+r^w). \end{aligned} \tag{9.9}$$

These provide a realization of the su(1,1) basis which are Hermitian under the scalar product

$$\langle f,g\rangle=\int_0^{\infty} f^*(r)g(r)r^w\,dr. \tag{9.10}$$

More precisely, $\mathscr{D}^+(k)$ consist of functions of r that are integrable with respect to the above measure, and (9.9) forms a realization of the su(1,1) basis in this representation $\mathscr{D}^+(k)$ provided

$$\xi=k(k-1)-W(W-1), \quad \text{where} \quad W=\frac{w+1}{2w}. \tag{9.11}$$

The normalized eigenfunctions $e_m^{(k)}$ of J_0 are given by

$$e_m^{(k)}=2^W\sqrt{\frac{wm!}{\Gamma(2k+m)}}\exp(-r^w)(2r^w)^{k-W}L_m^{(2k-1)}(2r^w), \tag{9.12}$$

in terms of the Laguerre polynomials (2.7) [see Ref. 1, eq. (89)]. An advantage of this realization is that J_2 is a linear combination of rd/dr, and hence $\exp(i\alpha J_2)$ acting on a function of r takes a simple form, i.e.,

$$\exp(i\alpha J_2)f(r)=e^{\alpha W}f(e^{\alpha/w}r). \tag{9.13}$$

Using all this information, (9.8) becomes

$$\frac{m!}{(2k)_m}\exp(-e^{\alpha}r^w)e^{\alpha k}L_m^{(2k-1)}(2e^{\alpha}r^w)$$

$$=\phi_m\sum_{n=0}^{\infty}(1-c^2)^k c^{n+m}M_n(m;2k;c^2)\exp(-r^w)L_n^{(2k-1)}(2r^w). \tag{9.14}$$

Making the replacements

$$x=2r^w,\quad \rho=e^{\alpha}>1,\quad a=2k-1, \tag{9.15}$$

and keeping in mind the relation (9.2), (9.14) takes a simpler form. In fact, rather than leaving this as an infinite series expression in terms of Laguerre polynomials, it is more convenient to integrate with respect to the orthogonality measure of Laguerre polynomials and then obtain the equivalent expression:

Corollary 5: For $\rho>1$ and $a>-1$,

$$\int_0^{\infty}L_m^{(a)}(\rho x)L_n^{(a)}(x)\exp(-(\rho+1)x/2)x^a dx$$

$$=(-1)^m\frac{\Gamma(a+n+1)}{n!}\frac{(a+1)_m}{m!}\left(\frac{2}{\rho+1}\right)^{a+1}\left(\frac{\rho-1}{\rho+1}\right)^{n+m}M_n\left(m;a+1;\left(\frac{\rho-1}{\rho+1}\right)^2\right). \tag{9.16}$$

The fact that the phase factor ϕ_m is equal to $(-1)^m$ can easily be derived from the case $n=0$. Although we have not found (9.16) in the literature, it is perhaps not new, and it can probably be derived by classical methods as well. Nevertheless, in our treatment it does not require any extra work to derive it, and the appearance of a Meixner polynomial on the rhs of (9.16) has a natural explanation.

As a second example of a different realization, we shall consider the well-known boson realization of su(1,1) (see also Ref. 1). In terms of the annihilation and creation operators

$$a=\frac{1}{\sqrt{2}}\left(x+\frac{d}{dx}\right),\quad a^{\dagger}=\frac{1}{\sqrt{2}}\left(x-\frac{d}{dx}\right), \tag{9.17}$$

we have the realization

$$J_1=\tfrac{1}{4}((a^{\dagger})^2+a^2),\quad J_2=\tfrac{1}{4i}((a^{\dagger})^2-a^2),\quad J_0=\tfrac{1}{4}(aa^{\dagger}+a^{\dagger}a). \tag{9.18}$$

The representation space is $L_2(\mathbb{R})$, with inner product $\langle f,g\rangle=\int f^*(x)g(x)dx$. This corresponds to the positive discrete series representation $\mathscr{D}^+(1/4)$, so $k=1/4$. The normalized eigenvectors $e_n^{(1/4)}$ are given by

$$e_n^{(1/4)} = \frac{1}{\pi^{1/4} 2^n \sqrt{(2n)!}} \exp(-x^2/2) H_{2n}(x), \tag{9.19}$$

with $H_n(x)$ the usual notation for Hermite polynomials (Ref. 22, §22). Since also for this realization J_2 is a linear combination of xd/dx, the simple relation

$$\exp(i\alpha J_2) f(x) = e^{-\alpha/4} f(e^{-\alpha/2} x) \tag{9.20}$$

holds. Plugging all this information in (9.8) gives an expression similar to (9.14), but with Hermite polynomials instead of Laguerre polynomials. Replacing herein $e^{-\alpha/2}$ by λ, and going to the equivalent form in terms of an integral, this result can be expressed as follows:

Corollary 6:

$$\int_{-\infty}^{\infty} H_{2m}(\lambda x) H_{2n}(x) \exp(-(\lambda^2+1)x^2/2) dx$$
$$= (-1)^m \sqrt{\frac{2\pi}{1+\lambda^2}} \left(\frac{1-\lambda^2}{1+\lambda^2}\right)^{n+m} \frac{(2m)!(2n)!}{m!n!} M_n\left(m; 1/2; \left(\frac{1-\lambda^2}{1+\lambda^2}\right)^2\right). \tag{9.21}$$

The phase factor ϕ_m is again easily derived. Again we have not found (9.21) in the literature. Although it can be derived by classical methods, here it has required no extra work.

With the two different realizations considered in this section, the analysis can be continued. For example, another interesting equation is the realization of (2.13) in terms of the above Laguerre polynomials.

In the case of the quantum algebra $U_q(\mathrm{su}(1,1))$, similar developments can be made provided one can find sufficiently simple realizations in terms of q-difference operators and q-special functions.

To conclude, using the explicit knowledge of the expansion coefficients of the eigenstates of a general Hamiltonian in terms of the J_0 eigenstates, we have deduced a number of identities for orthogonal polynomials. By choosing a realization for which the J_0 eigenstates coincide with monomials z^n, generating functions or Poisson kernels were derived. Considering the q-analog of this realization, the Poisson kernel for Al-Salam–Chihara polynomials is found. Finally, we have shown that different realizations for the su(1,1) case give rise to explicit formulas for integrals over orthogonal polynomials expressed as a Meixner polynomial.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor K. Srinivasa Rao for stimulating discussions. This research was partly supported by the E.C. (Contract No. CI1*-CT92-0101).

[1] B. G. Adams, J. Cizek, and J. Paldus, Adv. Quantum Chem. **19**, 1 (1988).
[2] A. I. Solomon, J. Math. Phys. **12**, 390 (1971).
[3] C. C. Gerry and S. Silverman, J. Math. Phys. **23**, 1995 (1982).
[4] J. Van der Jeugt, J. Math. Phys. **38**, 2728 (1997).
[5] Y. I. Granovskiĭ and A. S. Zhedanov, J. Phys. A **26**, 4339 (1993).
[6] H. T. Koelink and J. Van der Jeugt, ''Convolutions for orthogonal polynomials from Lie and quantum algebra representations,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **29**, 794 (1998); see also q-alg/9607010.
[7] H. Bacry, J. Math. Phys. **31**, 2061 (1990).
[8] I. M. Burban and A. U. Klimyk, J. Phys. A **26**, 2139 (1993).
[9] N. J. Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions*, 3 volumes (Kluwer, Dordrecht, 1991, 1993).
[10] R. Koekoek and R. F. Swarttouw, ''The Askey-scheme of hypergeometric orthogonal polynomials and its q-analogue,'' Report 94-05, Technical University Delft (1994); available from ftp.twi.tudelft.nl in directory /pub/publications/tech-reports.
[11] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge U.P., Cambridge, 1966).
[12] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge U.P., Cambridge, 1990).
[13] V. K. B. Kota, J. Math. Phys. **38**, 6639 (1997).
[14] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. 1.
[15] H. T. Koelink and J. Van der Jeugt, ''Bilinear generating functions for orthogonal polynomials,'' University of Amsterdam, Report 97-03; see also q-alg/9704016.

[16] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge U.P., Cambridge, 1994).
[17] R. Askey and J. Wilson, Mem. Am. Math. Soc. **54**, 319 (1985).
[18] W. Al-Salam and T. S. Chihara, SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **7**, 16 (1976).
[19] R. A. Askey, M. Rahman, and S. K. Suslov, J. Comput. Appl. Math. **68**, 25 (1996).
[20] M. E. H. Ismail and D. Stanton, Can. J. Math. **49**, 520 (1997).
[21] U. Fano and G. Racah, *Irreducible Tensorial Sets* (Academic, New York, 1959).
[22] M. Abramowitz and I. A. Segun, *Handbook of Mathematical Functions* (Dover, New York, 1964).

Corrections to the emergent canonical commutation relations arising in the statistical mechanics of matrix models

Stephen L. Adler[a)]
Institute for Advanced Study, Princeton, New Jersey 08540

Achim Kempf
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW, United Kingdom

(Received 13 February 1998; accepted for publication 18 May 1998)

We study the leading corrections to the emergent canonical commutation relations arising in the statistical mechanics of matrix models, by deriving several related Ward identities, and give conditions for these corrections to be small. We show that emergent canonical commutators are possible only in matrix models in complex Hilbert space for which the numbers of fermionic and bosonic fundamental degrees of freedom are equal, suggesting that supersymmetry will play a crucial role. Our results simplify, and sharpen, those obtained earlier by Adler and Millard. © *1998 American Institute of Physics.* [S0022-2488(98)03310-6]

I. INTRODUCTION

It is widely believed that at distances of order the Planck length $\ell_P \sim 10^{-33}$ cm our conventional notions of the geometry of space–time break down, as a result of quantum gravity effects. One indication of the modifications in physics that might be expected is provided by string theory models of quantum gravity, in which several studies suggest a modification of the uncertainty relation of the form[1]

$$\Delta x \Delta p \geqslant \frac{\hbar}{2}[1+\beta(\Delta p)^2+\cdots], \quad \beta>0, \tag{1a}$$

implying a finite minimum uncertainty $\Delta x_0 = \hbar\beta^{1/2}$ in the vicinity of the Planck length. As discussed by Kempf and collaborators,[2] Eq. (1a) corresponds to a correction to the Heisenberg canonical commutation relations of the form

$$[x,p] = i\hbar(1+\beta p^2+\cdots). \tag{1b}$$

We wish in this paper to discuss modifications of the Heisenberg algebra arising in another context, that of the statistical mechanics of matrix models, and to compare them with Eqs. (1a, 1b). Several years ago, Adler proposed a set of rules for a generalized quantum or trace dynamics, which is a Lagrangian and Hamiltonian mechanics with arbitrary noncommutative phase space variables q, p, and this was developed in a series of papers with various collaborators.[3] For theories in which the action is constructed as the trace of a sum of matrix products of $N\times N$ matrix variables, trace dynamics gives a powerful, basis independent, way of representing the same dynamics that can also be described in terms of the N^2 individual matrix elements. A significant new result emerging from this point of view was obtained by Adler and Millard,[4] who argued that the *statistical mechanics* of trace dynamics takes the form of conventional quantum field theory, with the Heisenberg commutation relations holding for statistical averages over certain effective canonical variables obtained by projection from the original operator canonical variables. Recently, it has become clear[5] that the underlying assumptions of trace dynamics are satisfied by

a)Author to whom correspondence should be addressed. Phone: 609-734-8051; fax: 609-924-8399; electronic mail: adler@ias.edu

 © 1998 American Institute of Physics

matrix models, for which the methods of trace dynamics provide a very convenient calculational tool. Hence the results of Adler and Millard can be reinterpreted as providing a statistical mechanics of matrix models, and showing that thermal averages in this statistical mechanics can behave as Wightman functions in an emergent local quantum field theory. These results, together with recent work[6] suggesting that the underlying dynamics for string theory may be a form of a matrix model, raise in turn the question of determining the form of the leading corrections to the Heisenberg algebra implied by the statistics of matrix models, formulating conditions for these corrections to be small, and seeing whether they can be related to the string theory result of Eqs. (1a, 1b).

An investigation of these questions is the focus of this paper, which is organized as follows. In Sec. II we give a brief synopsis of the rules of trace dynamics in the context of matrix models. We show that the conservation of the operator[4,7] $\tilde{C}$ can be understood as a simple consequence of unitary invariance. We also remark that, with Grassmann fermions, $\tilde{C}$ is independent of the classical parts of the matrix phase space variables, and review the statistical mechanics[4,8] of matrix models. In Sec. III we consider the simple case of a bosonic matrix model with Hamiltonian quadratic in the canonical momenta, and, making no approximations, derive a simplified form of the Ward identity used in Ref. 4 to obtain the effective canonical algebra. This analysis shows that there are corrections to the canonical commutator quadratic in the canonical momentum. In Sec. IV we use the symplectic formalism of Ref. 4 to repeat this calculation in the case of a general commutator/anticommutator of canonical variables in a generic matrix model, that can include fermions, and we generalize the treatment of Ref. 4 to allow nonzero sources for the classical parts of the matrix variables. From the analyses of Secs. III and IV, we formulate conditions for the corrections to the emergent canonical algebra to be small. We show that these conditions require $\tilde{C}$ to be an intensive rather than extensive thermodynamic quantity, and that they can be satisfied in complex Hilbert space (if at all) only in matrix models with precisely equal numbers of bosonic and fermionic degrees of freedom. This result strongly suggests that candidate matrix models for prequantum mechanics should be supersymmetric. We conclude by generalizing the conditions to ones that permit the recovery of the full emergent quantum field theory structure derived in Ref. 4. We also compare the prequantum corrections to the canonical algebra derived in Secs. III and IV, in which (as is usual in field theories) the spatial coordinate is simply a label, and the field variables are the dynamical canonical variables, to the string theory inspired expression of Eq. (1b) in which x is a coordinate operator.

II. THE STATISTICAL MECHANICS OF MATRIX MODELS

We begin by reviewing the statistical mechanics of trace dynamics, taking into account the simplifications[5] that become possible when Grassmann algebras are employed to represent the fermion/boson distinction. Let B_1 and B_2 be two $N\times N$ matrices with matrix elements that are even grade elements of a complex Grassmann algebra, and Tr the ordinary matrix trace, which obeys the cyclic property

$$\mathrm{Tr}\ B_1B_2=\sum_{m,n}(B_1)_{mn}(B_2)_{nm}=\sum_{m,n}(B_2)_{nm}(B_1)_{mn}=\mathrm{Tr}\ B_2B_1. \tag{2a}$$

Similarly, let χ_1 and χ_2 be two $N\times N$ matrices with matrix elements that are odd grade elements of a complex Grassmann algebra, which anticommute rather than commute, so that the cyclic property for these takes the form

$$\mathrm{Tr}\ \chi_1\chi_2=\sum_{m,n}(\chi_1)_{mn}(\chi_2)_{nm}=-\sum_{m,n}(\chi_2)_{nm}(\chi_1)_{mn}=-\mathrm{Tr}\ \chi_2\chi_1. \tag{2b}$$

The cyclic/anticyclic properties of Eqs. (2a, 2b) are just those assumed for the trace operation **Tr** of trace dynamics. [In Refs. 3 and 4 the fermionic operators were realized as ordinary matrices with complex matrix elements, all of which anticommute with a grading operator $(-1)^F$ which formed part of the definition of the graded trace **Tr**, for which fermions then obeyed Eq. (2b) while bosons obeyed Eq. (2a). Since the use of Grassmann odd fermions eliminates the need for

the inclusion of the $(-1)^F$ factor, the graded trace obeying Eqs. (2a, 2b) is just the usual matrix trace, for which we use the customary notation Tr.] From Eqs. (2a, 2b), one immediately derives the trilinear cyclic identities

$$\begin{aligned}
&\mathrm{Tr}\ B_1[B_2,B_3]=\mathrm{Tr}\ B_2[B_3,B_1]=\mathrm{Tr}\ B_3[B_1,B_2],\\
&\mathrm{Tr}\ B_1\{B_2,B_3\}=\mathrm{Tr}\ B_2\{B_3,B_1\}=\mathrm{Tr}\ B_3\{B_1,B_2\},\\
&\mathrm{Tr}\ B\{\chi_1,\chi_2\}=\mathrm{Tr}\ \chi_1[\chi_2,B]=\mathrm{Tr}\ \chi_2[\chi_1,B],\\
&\mathrm{Tr}\ \chi_1\{B,\chi_2\}=\mathrm{Tr}\{\chi_1,B\}\chi_2=\mathrm{Tr}[\chi_1,\chi_2]B,
\end{aligned} \tag{2c}$$

which are used repeatedly in trace dynamics calculations.

The basic observation of trace dynamics is that given the trace of a polynomial P constructed from noncommuting matrix or operator variables (we shall use the terms ''matrix'' and ''operator'' interchangeably in the following discussion), one can define a derivative of the c-number Tr P with respect to an operator variable $\mathcal{O}$ by varying and then cyclically permuting so that in each term the factor $\delta\mathcal{O}$ stands on the right, giving the fundamental definition

$$\delta\ \mathrm{Tr}\ P=\mathrm{Tr}\ \frac{\delta\ \mathrm{Tr}\ P}{\delta\mathcal{O}}\ \delta\mathcal{O}, \tag{3a}$$

or in the condensed notation that we shall use throughout this paper, in which $\mathbf{P}\equiv\mathrm{Tr}\ P$,

$$\delta\mathbf{P}=\mathrm{Tr}\ \frac{\delta\mathbf{P}}{\delta\mathcal{O}}\ \delta\mathcal{O}. \tag{3b}$$

Letting $\mathbf{L}[\{q_r\},\{\dot{q}_r\}]$ be a trace Lagrangian that is a function of the bosonic or fermionic operators $\{q_r\}$ and their time derivatives (which are all assumed to obey the cyclic relations of Eqs. (2a-2c) under the trace), and requiring that the trace action $\mathbf{S}=\int dt\mathbf{L}$ be stationary with respect to variations of the q_r's that preserve their bosonic or fermionic type, one finds[3] the operator Euler-Lagrange equations

$$\frac{\delta\mathbf{L}}{\delta q_r}-\frac{d}{dt}\frac{\delta\mathbf{L}}{\delta\dot{q}_r}=0. \tag{3c}$$

Because, by the definition of Eq. (3a), we have

$$\left(\frac{\delta\mathbf{L}}{\delta q_r}\right)_{ij}=\frac{\partial\mathbf{L}}{\partial(q_r)_{ji}}, \tag{3d}$$

for each r the single Euler-Lagrange equation of Eq. (3c) is equivalent to the N^2 Euler-Lagrange equations obtained by regarding $\mathbf{L}$ as a function of the N^2 matrix element variables $(q_r)_{ji}$. Defining the momentum operator p_r conjugate to q_r, which is of the same bosonic or fermionic type as q_r, by

$$p_r\equiv\frac{\delta\mathbf{L}}{\delta\dot{q}_r}, \tag{4a}$$

the trace Hamiltonian $\mathbf{H}$ is defined by

$$\mathbf{H}=\mathrm{Tr}\sum_r\ p_r\dot{q}_r-\mathbf{L}. \tag{4b}$$

In correspondence with Eq. (3d), the matrix elements $(p_r)_{ij}$ of the momentum operator p_r just correspond to the momenta canonical to the matrix element variables $(q_r)_{ji}$. Performing general

same-type operator variations, and using the Euler-Lagrange equations, we find from Eq. (4b) that the trace Hamiltonian **H** is a trace functional of the operators $\{q_r\}$ and $\{p_r\}$,

$$\mathbf{H}=\mathbf{H}[\{q_r\},\{p_r\}], \tag{5a}$$

with the operator derivatives

$$\frac{\delta \mathbf{H}}{\delta q_r}=-\dot{p}_r\,,\quad \frac{\delta \mathbf{H}}{\delta p_r}=\epsilon_r \dot{q}_r\,, \tag{5b}$$

with $\epsilon_r=1(-1)$ according to whether $q_r\,,p_r$ are bosonic (fermionic). Letting **A** and **B** be two trace functions of the operators $\{q_r\}$ and $\{p_r\}$, it is convenient to define the *generalized Poisson bracket*

$$\{\mathbf{A},\mathbf{B}\}=\mathrm{Tr}\sum_r\ \epsilon_r\left(\frac{\delta \mathbf{A}}{\delta q_r}\frac{\delta \mathbf{B}}{\delta p_r}-\frac{\delta \mathbf{B}}{\delta q_r}\frac{\delta \mathbf{A}}{\delta p_r}\right). \tag{6a}$$

Then using the Hamiltonian form of the equations of motion, one readily finds that for a general trace functional $\mathbf{A}[\{q_r\},\{p_r\}]$, the time derivative is given by

$$\frac{d}{dt}\,\mathbf{A}=\{\mathbf{A},\mathbf{H}\}; \tag{6b}$$

in particular, letting **A** be the trace Hamiltonian **H**, and using the fact that the generalized Poisson bracket is antisymmetric in its arguments, it follows that the time derivative of **H** vanishes. An important property of the generalized Poisson bracket is that it satisfies[3] the Jacobi identity,

$$\{\mathbf{A},\{\mathbf{B},\mathbf{C}\}\}+\{\mathbf{C},\{\mathbf{A},\mathbf{B}\}\}+\{\mathbf{B},\{\mathbf{C},\mathbf{A}\}\}=0. \tag{6c}$$

As a consequence, if $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are two conserved charges, that is if

$$0=\frac{d}{dt}\,\mathbf{Q_1}=\{\mathbf{Q_1},\mathbf{H}\},\quad 0=\frac{d}{dt}\,\mathbf{Q_2}=\{\mathbf{Q_2},\mathbf{H}\}, \tag{6d}$$

then their generalized Poisson bracket $\{\mathbf{Q_1},\mathbf{Q_2}\}$ also has a vanishing generalized Poisson bracket with **H**, and is conserved. This has the consequence that Lie algebras of symmetries can be represented as Lie algebras of trace functions under the generalized Poisson bracket operation.

A significant feature of trace dynamics is that, as discovered by Millard,[7] the anti-self-adjoint operator[7,4]

$$\tilde{C}\equiv \sum_{r\ \text{bosons}}[q_r\,,p_r]-\sum_{r\ \text{fermions}}\{q_r\,,p_r\} \tag{7}$$

is conserved by the dynamics. As we shall now show, conservation of $\tilde{C}$ holds whenever the trace dynamics has a global unitary invariance, that is, whenever the trace Hamiltonian obeys

$$\mathbf{H}[\{U^\dagger q_r U\},\{U^\dagger p_r U\}]=\mathbf{H}[\{q_r\},\{p_r\}] \tag{8a}$$

for a constant unitary $N\times N$ matrix U, or equivalently, by Eq. (4b), whenever the trace Lagrangian obeys

$$\mathbf{L}[\{U^\dagger q_r U\},\{U^\dagger \dot{q}_r U\}]=\mathbf{L}[\{q_r\},\{\dot{q}_r\}]. \tag{8b}$$

Letting $U=\exp\Lambda$, with Λ an anti-self-adjoint bosonic generator matrix, and expanding to first order in Λ, Eq. (8a) implies that

$$\mathbf{H}[\{q_r-[\Lambda,q_r]\},\{p_r-[\Lambda,p_r]\}]=\mathbf{H}[\{q_r\},\{p_r\}]. \tag{9a}$$

But applying the definition of the variation of a trace functional given in Eq. (3b), Eq. (9a) becomes

$$\mathrm{Tr}\sum_r \left(-\frac{\delta \mathbf{H}}{\delta q_r}[\Lambda, q_r] - \frac{\delta \mathbf{H}}{\delta p_r}[\Lambda, p_r]\right) = 0, \tag{9b}$$

which by use of the trilinear cyclic identities of Eq. (2c) yields

$$\mathrm{Tr}\ \Lambda \sum_r \left(\frac{\delta \mathbf{H}}{\delta q_r} q_r - \epsilon_r q_r \frac{\delta \mathbf{H}}{\delta q_r} + \frac{\delta \mathbf{H}}{\delta p_r} p_r - \epsilon_r p_r \frac{\delta \mathbf{H}}{\delta p_r}\right) = 0. \tag{9c}$$

Since the generator Λ is an arbitrary anti-self-adjoint $N \times N$ matrix, the anti-self-adjoint matrix that multiplies it in Eq. (9c) must vanish, giving the matrix identity

$$\sum_r \left(\frac{\delta \mathbf{H}}{\delta q_r} q_r - \epsilon_r q_r \frac{\delta \mathbf{H}}{\delta q_r} + \frac{\delta \mathbf{H}}{\delta p_r} p_r - \epsilon_r p_r \frac{\delta \mathbf{H}}{\delta p_r}\right) = 0. \tag{10a}$$

But now substituting the Hamilton equations of Eq. (5b), Eq. (10a) takes the form

$$\begin{aligned} 0 &= \sum_r \ (-\dot{p}_r q_r + \epsilon_r q_r \dot{p}_r + \epsilon_r \dot{q}_r p_r - p_r \dot{q}_r) \\ &= \frac{d}{dt} \sum_r \ (-p_r q_r + \epsilon_r q_r p_r) \\ &= \frac{d}{dt} \left(\sum_{r\ \mathrm{bosons}} [q_r, p_r] - \sum_{r\ \mathrm{fermions}} \{q_r, p_r\} \right), \end{aligned} \tag{10b}$$

completing the demonstration of the conservation of $\tilde{C}$.

Corresponding to the fact that $\tilde{C}$ is conserved in any matrix model with a global unitary invariance, it is easy to see[4,8] that $\tilde{C}$ can be used to construct the generator of global unitary transformations of the Hilbert space basis. Consider the trace functional

$$\mathbf{G}_\Lambda = -\mathrm{Tr}\ \Lambda \tilde{C}, \tag{11a}$$

with Λ a fixed bosonic anti-self-adjoint operator, which can be rewritten, using cyclic invariance of the trace, as

$$\mathbf{G}_\Lambda = \mathrm{Tr}\sum_r \ [\Lambda, p_r] q_r = -\mathrm{Tr}\sum_r \ p_r [\Lambda, q_r]. \tag{11b}$$

Hence for the variations of p_r and q_r induced by $\mathbf{G}_\Lambda$ as canonical generator, which have a structure analogous to the Hamilton equations of Eq. (5b), we get

$$\delta p_r = -\frac{\delta \mathbf{G}_\Lambda}{\delta q_r} = -[\Lambda, p_r], \quad \delta q_r = \epsilon_r \frac{\delta \mathbf{G}_\Lambda}{\delta p_r} = -[\Lambda, q_r]. \tag{11c}$$

Comparing with Eqs. (8a) and (9a), we see that these have just the form of an infinitesimal global unitary transformation.

For each phase space variable q_r, p_r, let us define the *classical part* q_r^c, p_r^c and the noncommutative remainder q_r', p_r', by

$$q_r^c = \frac{1}{N}\mathrm{Tr}\ q_r, \quad p_r^c = \frac{1}{N}\mathrm{Tr}\ p_r, \quad q_r' = q_r - q_r^c, \ \ p_r' = p_r - p_r^c, \tag{12a}$$

so that bosonic q_r^c, p_r^c are *c*-numbers, fermionic q_r^c, p_r^c are Grassmann *c*-numbers (where by a *c*-number we mean a multiple of 1_N, the $N \times N$ unit matrix), and the remainders are traceless,

$$\mathrm{Tr}\ q_r' = \mathrm{Tr}\ p_r' = 0. \tag{12b}$$

Then since q_r^c, p_r^c commute (anticommute) with q_s', p_s' for r, s both bosonic (fermionic), we see that the classical parts of the phase space variables make no contribution to $\tilde{C}$, and Eq. (7) can be rewritten as

$$\tilde{C} = \sum_r \sum_{\text{bosons}} [q_r', p_r'] - \sum_r \sum_{\text{fermions}} \{q_r', p_r'\}. \tag{12c}$$

Thus $\tilde{C}$ is completely independent of the values of the classical parts of the matrix phase space variables.

Making the assumption that trace dynamics is ergodic (which undoubtedly requires an interacting as opposed to a free theory, and may presuppose taking the $N \to \infty$ limit), one can then analyze[4] the statistical mechanics of trace dynamics for the generic case in which the conserved quantities are the trace Hamiltonian **H** and the operator $\tilde{C}$. As discussed in detail in the second paper cited in Ref. 5, the analysis of Ref. 4 carries over to the case in which the fermions are represented by Grassmann matrices; the demonstration of a generalized Liouville theorem still holds, and the requirements for convergence of the partition function are much less stringent, eliminating the complexities addressed in Appendix F of Ref. 4. With Grassmann fermions, for the typical models we are studying the bosonic part of H is a positive operator, from which **H** inherits good positivity properties. The canonical ensemble then takes the simple form given in Eq. (48c) of Ref. 4,

$$\rho = Z^{-1} \exp(-\tau \mathbf{H} - \mathrm{Tr}\ \tilde{\lambda}\tilde{C}), \quad Z = \int d\mu \exp(-\tau \mathbf{H} - \mathrm{Tr}\ \tilde{\lambda}\tilde{C}), \tag{13}$$

with $d\mu$ the invariant matrix (or operator) phase space measure provided by Liouville's theorem, with τ a real number, and with $\tilde{\lambda}$ an anti-self-adjoint matrix that in the generic case (which we assume) has no zero eigenvalues. (Equation (13) can be derived directly[4] by maximizing the entropy subject to the constraints imposed by the conservation of **H** and $\tilde{C}$, or indirectly[8] by first calculating the corresponding microcanonical ensemble corresponding to these conserved quantities, and then using standard statistical physics methods to calculate the canonical ensemble from the microcanonical one.) We wish to make two points about the partition function defined in Eq. (13). First, it is not invariant under the unitary transformation of Eq. (8a) for fixed $\tilde{\lambda}$, but is invariant when $\tilde{\lambda}$ is simultaneously transformed to $U^\dagger \tilde{\lambda} U$; hence the partition function breaks unitary invariance, but has a specific form of unitary covariance. Second, the partition function contains a weighted sum over all possible commutators $[q_r, p_s]$ for bosonic variables and all possible anticommutators $\{q_r, p_s\}$ for fermionic variables; there is no restriction to the classical or quantum mechanical evaluation of these commutators/anticommutators as 0 or $i\delta_{rs}$ respectively. However, statistical integrals like Eq. (13) are typically dominated by specific regions of the integration domain, and we will see, by a study of the Ward identities following from Eq. (13), that this can lead to *effective* quantum mechanical commutators inside statistical averages. The structure of the Ward identities or equipartition theorems following from Eq. (13) will be reanalyzed in the next two sections without making approximations used in Ref. 4, so as to determine the leading corrections to the emergent canonical commutation relations. From this analysis we will infer a set of conditions for obtaining the full emergent quantum field theory structure of Ref. 4.

III. CORRECTIONS TO THE BOSONIC COMMUTATOR $[q_s, p_r]$ IN A SIMPLIFIED UNITARY INVARIANT MATRIX MODEL

We consider in this section the simplified bosonic matrix model with trace Hamiltonian

$$\mathbf{H} = \mathrm{Tr}\left[\sum_r \frac{1}{2} p_r^2 + V(\{q_r\})\right], \tag{14a}$$

with the q_r self-adjoint $N \times N$ complex matrix variables and with V a global unitary invariant potential. This form is general enough to include the matrix model forms of the bosonic field

theories of greatest interest, including the Goldstone model, non-Abelian gauge models, and the Higgs model. As we saw in the previous section, the Hamiltonian dynamics for this model conserves both the real number **H** and the matrix $\tilde{C}$, which in this case is given simply by

$$\tilde{C}=\sum_r [q_r,p_r]. \tag{14b}$$

Letting ρ and Z be respectively the canonical ensemble and partition function given in terms of **H** and $\tilde{C}$ by Eq. (13), we define the ensemble average of an arbitrary function $\mathcal{O}$ of the dynamical variables by

$$\langle \mathcal{O} \rangle_{AV}=\int d\mu \rho \mathcal{O}=Z^{-1}\int d\mu e^{-\tau \mathbf{H}-\mathrm{Tr}(\tilde{\lambda}\tilde{C})}\mathcal{O}. \tag{15}$$

Letting $\mathcal{O}$ be the conserved operator $\tilde{C}$, and noting that the right hand side of Eq. (13) can be a function only of the ensemble parameters $\tilde{\lambda}$ and τ, we have

$$\langle \tilde{C} \rangle_{AV}=f(\tilde{\lambda},\tau), \tag{16a}$$

with f an anti-self-adjoint matrix, which in general can be written as a phase matrix i_{eff} times a commuting magnitude matrix $|f|$,

$$f=i_{\mathrm{eff}}|f|, \quad i_{\mathrm{eff}}^2=-1, \quad i_{\mathrm{eff}}^\dagger=-i_{\mathrm{eff}}, \quad [i_{\mathrm{eff}},|f|]=0. \tag{16b}$$

We shall now specialize to an ensemble for which the magnitude matrix $|f|$ makes no distinction among the different bases in Hilbert space, and so takes the form of a positive real multiple (which we shall call $\hbar$) of the unit matrix. [As discussed in Appendix B of Ref. 4, when **H** can be expressed in terms of the phase space operators $\{q_r,p_r\}$ using only real number coefficients, this assumption implies that we are restricting attention to the special class of ensembles for which $\tilde{\lambda}=i_{\mathrm{eff}}\lambda$, with λ a real multiple of the unit matrix.] Equations (16a, 16b) then become

$$\langle \tilde{C} \rangle_{AV}=i_{\mathrm{eff}}\hbar. \tag{16c}$$

Since for finite N we necessarily have $\mathrm{Tr}\,\tilde{C}=0$, the phase matrix i_{eff} must have vanishing trace,

$$\mathrm{Tr}\, i_{\mathrm{eff}}=0, \tag{16d}$$

which implies that i_{eff} has $N/2$ eigenvalues i and $N/2$ eigenvalues $-i$. Thus, we are making a choice of ensemble for which the $U(N)$ symmetry of **H** is broken, by the term $\mathrm{Tr}\,\tilde{\lambda}\tilde{C}$, to $U(N/2)\times U(N/2)\times R$, with R the discrete reflection symmetry that interchanges the eigenvalues $\pm i$ of i_{eff}. This is clearly the largest symmetry group of the ensemble for which one can have $\langle \tilde{C} \rangle_{AV}\neq 0$; if one were to attempt to preserve the full $U(N)$ symmetry by taking an ensemble with $\tilde{\lambda}=i\lambda$, with λ a c-number, then in the canonical ensemble the term $\mathrm{Tr}\,\tilde{\lambda}\tilde{C}$ would vanish by virtue of the tracelessness of $\tilde{C}$, and the resulting ensemble would have $\langle \tilde{C} \rangle_{AV}=0$. Requiring the largest possible nontrivial symmetry group plays the role in our derivation of giving a single Planck constant for all pairs of canonical variables; if on the other hand, we were to sacrifice all of the $U(N)$ symmetry by allowing generic $\tilde{\lambda}$, then the emergent canonical commutation relations derived below would generically yield $N/2$ different $\hbar$'s for the $N/2$ pairs of canonical variables. It would clearly be desirable to have a deeper justification from first principles of our choice of ensemble, perhaps based on a more detailed understanding of the underlying dynamics, but at present we must simply introduce it as a postulate.

For this choice of ensemble, let us now consider the Ward identity obtained from

$$Z\langle \mathrm{Tr}\,\tilde{C}p_r \rangle_{AV}=\int d\mu e^{-\tau \mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\tilde{C}}\,\mathrm{Tr}\,\tilde{C}p_r, \tag{17a}$$

by using invariance of the measure $d\mu$ under a constant shift of p_s, which implies

$$0=\int d\mu\,\delta_{p_s}[e^{-\tau\mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\tilde{C}}\ \mathrm{Tr}\ \tilde{C}p_r]$$

$$=\int d\mu e^{-\tau\mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\tilde{C}}[(-\tau\delta_{p_s}\mathbf{H}-\mathrm{Tr}\ \tilde{\lambda}\delta_{p_s}\tilde{C})\mathrm{Tr}\ \tilde{C}p_r+\mathrm{Tr}(\delta_{p_s}\tilde{C})p_r+\mathrm{Tr}\ \tilde{C}\delta_{rs}\delta p_s]. \quad (17b)$$

Now from Eqs. (14a, 14b) we have

$$\delta_{p_s}\mathbf{H}=\mathrm{Tr}\ p_s\delta p_s,\ \ \delta_{p_s}\tilde{C}=[q_s,\delta p_s]. \quad (18)$$

Substituting these into Eq. (17b), multiplying by Z^{-1}, and using the trilinear cyclic identities of Eq. (2c), we get

$$0=\langle(-\tau\ \mathrm{Tr}\ p_s\delta p_s-\mathrm{Tr}[\tilde{\lambda},q_s]\delta p_s)\mathrm{Tr}\ \tilde{C}p_r+\mathrm{Tr}[p_r,q_s]\delta p_s+\mathrm{Tr}\ \tilde{C}\delta_{rs}\delta p_s\rangle_{AV}, \quad (19a)$$

which since δp_s is an arbitrary self-adjoint matrix, implies that the operator multiplying δp_s inside the trace must vanish,

$$0=\langle(-\tau p_s-[\tilde{\lambda},q_s])\mathrm{Tr}\ \tilde{C}p_r+[p_r,q_s]+\tilde{C}\delta_{rs}\rangle_{AV}. \quad (19b)$$

Since $\tilde{\lambda}$ is a constant matrix, it can be taken outside the ensemble average, and so the second term in Eq. (19b) takes the form

$$-[\tilde{\lambda},\langle q_s\ \mathrm{Tr}\ \tilde{C}p_r\rangle_{AV}], \quad (19c)$$

which vanishes since the ensemble average inside the commutator in Eq. (19c) is a matrix function only of $\tilde{\lambda}$ and τ, and hence commutes with $\tilde{\lambda}$. Substituting $p_r=p_r^c+p_r'$ into the term multiplied by τ in Eq. (19b), and using $\mathrm{Tr}\ \tilde{C}p_r^c=p_r^c\ \mathrm{Tr}\ \tilde{C}=0$, the traceless part of Eq. (19b) reduces to

$$0=\langle-\tau p_s'\ \mathrm{Tr}\ \tilde{C}p_r'+[p_r,q_s]+\tilde{C}\delta_{rs}\rangle_{AV}. \quad (20a)$$

Since $\langle\tilde{C}\rangle_{AV}=i_{\mathrm{eff}}\hbar$, this equation can be rewritten as the *exact* relation

$$\langle[q_s,p_r]\rangle_{AV}=i_{\mathrm{eff}}\hbar\,\delta_{rs}-\tau\langle p_s'\ \mathrm{Tr}\ \tilde{C}p_r'\rangle_{AV}, \quad (20b)$$

showing that the ensemble averaged commutator of the canonical coordinate and momentum operators has the form of the usual quantum mechanical canonical commutator, with i_{eff} playing the role of the imaginary unit, and with a correction term proportional to τ that is quadratic in the nonclassical parts of the canonical momenta. Using Eqs. (14b) and (16c), Eq. (20b) can also be written in the form

$$\sum_{t,u}\langle[q_t,p_u]\rangle_{AV}(\delta_{\mathrm{tr}}\delta_{us}-\delta_{\mathrm{tu}}\delta_{rs})=-\tau\langle p_s'\ \mathrm{Tr}\ \tilde{C}p_r'\rangle_{AV}. \quad (20c)$$

The derivation given here sharpens that given in Adler and Millard,[4] both in that here no approximations have been made, and that because we are working in complex Hilbert space, we have not had to first project out the parts of q_s and p_r that commute with i_{eff}. (As discussed in the Appendix, this projection is needed in real and quaternionic Hilbert space.)

Equation (20b) sets up a correspondence between two different matrix theories. The first is the underlying classical matrix dynamics following from the trace Hamiltonian equations of Eq. (5b), in which the $2N^2$ matrix elements of $\{q_r\},\{p_r\}$ can be regarded as dynamical variables. In this underlying matrix dynamics, $[q_s,p_r]$ is nonzero for $r\neq s$ and is a rapidly varying function over the matrix phase space. The second is the statistical thermodynamics of this underlying matrix dynamics, in which the fine details of the underlying dynamics have been averaged out. Equation (20b) shows that, when the τ term can be neglected, the *averaged* $[q_s,p_r]$ simplifies to $i_{\mathrm{eff}}\hbar\delta_{rs}$, which vanishes for $r\neq s$ and is a multiple of a constant operator over matrix phase space when $r=s$. Thus, this thermodynamics has a structure that corresponds to that of the Heisenberg matrix

mechanics form of quantum mechanics, with each canonical pair q_r, p_r corresponding to the Heisenberg picture matrices for one quantum mechanical degree of freedom.

The reduction or thinning from the $2N^2$ classical dynamical variables for each r corresponding to unrestricted matrix variables q_r, p_r in the underlying matrix theory, to the two canonical quantum matrix variables q_r, p_r in the statistical thermodynamics derived from the underlying theory, should not be surprising. It is completely analogous to the reduction in the number of degrees of freedom that occurs in the passage from classical molecular dynamics, where the molecular variables are all dynamical variables, to its thermodynamics, where the molecular dynamical variables have been averaged out. The surviving dynamics in classical thermodynamics reflects just the conservation laws of energy and particle number in the underlying molecular dynamics. Similarly, the structure of the matrix thermodynamics following from Eq. (20b) is governed by the conserved quantities $\mathbf{H}, \tilde{C}$ of the underlying matrix dynamics. As discussed at the end of Sec. IV and also on pp. 221–222 of Ref. 4, the correspondence of Eq. (20b) can be further generalized to show that, when the τ terms can be neglected, there is a one to one correspondence between thermodynamic averages over the underlying matrix dynamics with trace Hamiltonian $\mathbf{H} = \mathrm{Tr}\, H$, and the Wightman functions of the quantum field theory with operator Hamiltonian H. In this correspondence, as already noted, i_{eff} plays the role of the imaginary unit i, and the magnitude of $\langle \tilde{C} \rangle_{AV}$ plays the role of Planck's constant.

We conclude this section with several remarks on the principal result of Eq. (20b). First of all, the fluctuations about the ensemble average are fundamental to the possibility of emergent quantum behavior. The average of the commutator on the left hand side of Eq. (20b) is not the same as the commutator of the averages, which in fact vanishes since $\langle q_s \rangle_{AV}$ and $\langle p_r \rangle_{AV}$ are both functions only of τ and $\tilde{\lambda}$,

$$\langle [q_s, p_r] \rangle_{AV} \neq [\langle q_s \rangle_{AV}, \langle p_r \rangle_{AV}] = 0. \tag{21a}$$

Secondly, in order to have emergent quantum behavior, the dynamics must be such that the second term on the right hand side of Eq. (20b) is much smaller than the first term on the right hand side, that is, one must have

$$\hbar^{-1} \tau |\langle p_s' \ \mathrm{Tr}\ \tilde{C} p_r' \rangle_{AV}| \ll 1. \tag{21b}$$

We shall now show that this condition cannot be satisfied if $\tilde{C}$ is an extensive thermodynamic quantity that grows linearly with the size of the system. To see this, we note that a second Ward identity, similar in form to Eq. (20b), can be derived by starting from

$$0 = \int d\mu \, \delta_{p_s} [e^{-\tau \mathbf{H} - \mathrm{Tr}\, \tilde{\lambda} \tilde{C}} \ \mathrm{Tr}\ i_{\mathrm{eff}} \hbar p_r], \tag{22a}$$

in which the factor of $\tilde{C}$ multiplying p_r has been replaced by its ensemble average $i_{\mathrm{eff}} \hbar$, and then proceeding as in Eqs. (17b–20b) above. The resulting Ward identity is

$$0 = i_{\mathrm{eff}} \hbar \, \delta_{rs} - \tau \langle p_s' \ \mathrm{Tr}\ i_{\mathrm{eff}} \hbar p_r' \rangle_{AV}, \tag{22b}$$

and is an analog in our context of the usual equipartition theorem of classical statistical mechanics. Now if $\tilde{C}$ were an extensive quantity, the difference between $\tilde{C}$ and its ensemble average $i_{\mathrm{eff}} \hbar$ would be a fluctuation that vanishes as the system size becomes infinite, which would make it permissible to accurately approximate the right hand side of Eq. (20b) by the right hand side of Eq. (22b). This would lead to the conclusion $\langle [q_s, p_r] \rangle_{AV} = 0$, that is, the thermodynamics would give emergent classical, rather than quantum mechanical, behavior.

Thus, the inequality of Eq. (21b) can hold only if $\tilde{C}$ is not extensive; this conclusion is consistent with the observation that to have the average of $\tilde{C}$ play the role of the intensive quantity $i_{\mathrm{eff}} \hbar$, one would expect that $\tilde{C}$ should behave as a thermodynamically intensive quantity. We shall give further evidence for this conclusion in the next section, where we consider systems containing fermions as well as bosons.

Although we have first discussed the purely bosonic model of this section for expository reasons, it is easy to see from Eqs. (20b, 20c) that the inequality of Eq. (21b) cannot hold in a purely bosonic system. When the right hand side of Eq. (20c) can be neglected, multiplying by δ_{rs} and summing over r,s gives the relation

$$\sum_{t,u} \langle [q_t, p_u] \rangle_{AV} \delta_{tu}(1-N) = 0, \tag{22c}$$

which for $N>1$ implies that $\langle [q_s, p_r] \rangle_{AV} = 0$, again giving classical behavior. Thus, a purely bosonic matrix dynamics system in complex Hilbert space cannot have emergent quantum behavior. However, we shall see in the next section that in the interesting case in which the numbers of bosonic and fermionic degrees of freedom are equal, a condition that holds for supersymmetric theories, the relation of Eq. (22c) is modified, and emergent quantum behavior becomes possible.

IV. CORRECTIONS TO THE FULL CANONICAL ALGEBRA IN A GENERAL UNITARY INVARIANT MATRIX MODEL WITH CLASSICAL SOURCES

Although one can give derivations similar to that of the previous section for other canonical commutators (e.g., the $[p_r, p_s]$ and $[q_r, q_s]$ commutators, or the $[p_s, q_r]$ commutator obtained by interchanging the roles of p and q in the above derivation), and their fermionic anticommutator analogs, it is most efficient in the general case to use the symplectic formalism introduced by Adler and Millard in Ref. 4. In this notation one defines $x_1 = q_1$, $x_2 = p_1$, $x_3 = q_2$, $x_4 = p_2, \ldots$, $x_{2R-1} = q_R$, $x_{2R} = p_R$ for the matrix phase space variables; in terms of these, the Hamilton equations of Eq. (5b), the generalized Poisson bracket of Eq. (6a), and the conserved operator $\tilde{C}$ of Eq. (7) take the form

$$\dot{x}_r = \sum_{s=1}^{2R} \omega_{rs} \frac{\delta \mathbf{H}}{\delta x_s}, \quad \{\mathbf{A}, \mathbf{B}\} = \mathrm{Tr} \sum_{r,s=1}^{2R} \frac{\delta \mathbf{A}}{\delta x_r} \omega_{rs} \frac{\delta \mathbf{B}}{\delta x_s}, \quad \tilde{C} = \sum_{r,s} x_r \omega_{rs} x_s, \tag{23a}$$

with ω a numerical symplectic metric given (in terms of standard Pauli matrices $\tau_{1,2,3}$) by $i\tau_2$ for a bosonic pair of canonical variables, and by $-\tau_1$ for a fermionic canonical pair (see Eqs. (10a–10c) on p. 202, and Eqs. (10a′-10c′) on p. 224, of Ref. 4). This symplectic metric obeys[4] the useful identities

$$\omega_{sr} = -\epsilon_r \omega_{rs} = -\epsilon_s \omega_{rs}, \quad \sum_r \omega_{rs} \omega_{rt} = \sum_r \omega_{sr} \omega_{\mathrm{tr}} = \delta_{st}. \tag{23b}$$

We shall now consider a matrix dynamics generated by a general trace Hamiltonian $\mathbf{H}$, that can contain fermionic as well as bosonic degrees of freedom, and in the statistical partition function shall allow the presence of nonvanishing classical sources J_r^c for the classical parts x_r^c of the phase space variables [cf. Eqs. (12a–12c) above]. Thus we start from the ensemble

$$\rho = Z^{-1} \exp\left(-\tau \mathbf{H} - \mathrm{Tr}\, \tilde{\lambda} \tilde{C} - \sum_r J_r^c x_r^c \right), \quad Z = \int d\mu \exp\left(-\tau \mathbf{H} - \mathrm{Tr}\, \tilde{\lambda} \tilde{C} - \sum_r J_r^c x_r^c \right), \tag{24}$$

which is now used in place of the ensemble of Eq. (13) in the definition of $\langle \mathcal{O} \rangle_{AV}$ given in Eq. (15). For this ensemble, we consider the Ward identity obtained by using shift invariance of the integration measure $d\mu$ starting from

$$Z \langle \mathrm{Tr}\, \tilde{C} \sigma_t x_t' \rangle_{AV} = \int d\mu e^{-\tau \mathbf{H} - \mathrm{Tr}\, \tilde{\lambda} \tilde{C} - \Sigma_r J_r^c x_r^c} \mathrm{Tr}\, \tilde{C} \sigma_t x_t', \tag{25a}$$

where σ_t are a set of c-number auxiliary parameters that are complex for bosonic x_t, and complex Grassmann for fermionic x_t. As in Eqs. (12a, 12b) above, we use the notation x_t' to denote the noncommutative part of x_t that remains when the classical part is subtracted away, that is, $x_t' = x_t - x_t^c$.

The Ward identity derivation now proceeds exactly as in Eqs. (17b–20b). Making a constant shift of the noncommutative part x_s' of the phase space variable x_s, we have

$$0=\int d\mu\,\delta_{x_s'}[e^{-\tau\mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\tilde{C}-\Sigma_r J_r^c x_r^c}\,\mathrm{Tr}\,\tilde{C}\sigma_t x_t']$$

$$=\int d\mu e^{-\tau\mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\tilde{C}-\Sigma_r J_r^c x_r^c}$$

$$\times[(-\tau\delta_{x_s'}\mathbf{H}-\mathrm{Tr}\,\tilde{\lambda}\delta_{x_s'}\tilde{C})\mathrm{Tr}\,\tilde{C}\sigma_t x_t'+\mathrm{Tr}(\delta_{x_s'}\tilde{C})\sigma_t x_t'+\mathrm{Tr}\,\tilde{C}\sigma_t\delta_{st}\delta x_s'].\tag{25b}$$

Now from Eqs. (23a, 23b) we have

$$\delta_{x_s'}\mathbf{H}=\mathrm{Tr}\left(\frac{\delta\mathbf{H}}{\delta x_s'}\right)'\delta x_s'=\mathrm{Tr}\sum_r\,\dot{x}_r'\omega_{rs}\delta x_s',$$

$$\mathrm{Tr}\,\tilde{\lambda}\delta_{x_s'}\tilde{C}=\mathrm{Tr}\,\tilde{\lambda}\sum_r\,\omega_{rs}(x_r'\delta x_s'-\epsilon_r\delta x_s' x_r')=\mathrm{Tr}\left[\tilde{\lambda},\sum_r\,\omega_{rs}x_r'\right]\delta x_s',\tag{26}$$

$$\mathrm{Tr}(\delta_{x_s'}\tilde{C})\sigma_t x_t'=\mathrm{Tr}\left[\sigma_t x_t',\sum_r\,\omega_{rs}x_r'\right]\delta x_s'.$$

Substituting these into Eq. (25b) and multiplying by Z^{-1}, we get

$$0=\left\langle\left(-\tau\,\mathrm{Tr}\sum_r\,\dot{x}_r'\omega_{rs}\delta x_s'-\mathrm{Tr}\left[\tilde{\lambda},\sum_r\,\omega_{rs}x_r'\right]\delta x_s'\right)\mathrm{Tr}\,\tilde{C}\sigma_t x_t'\right.$$

$$\left.+\mathrm{Tr}\left[\sigma_t x_t',\sum_r\,\omega_{rs}x_r'\right]\delta x_s'+\mathrm{Tr}\,\tilde{C}\sigma_t\delta_{st}\delta x_s'\right\rangle_{AV},\tag{27a}$$

which since $\delta x_s'$ is an arbitrary traceless matrix (with the adjointness properties of x_s'), implies that the traceless part of the operator multiplying $\delta x_s'$ inside the trace must vanish,

$$0=\left\langle\left(-\tau\sum_r\,\dot{x}_r'\omega_{rs}-\left[\tilde{\lambda},\sum_r\,\omega_{rs}x_r'\right]\right)\mathrm{Tr}\,\tilde{C}\sigma_t x_t'+\left[\sigma_t x_t',\sum_r\,\omega_{rs}x_r'\right]+\tilde{C}\sigma_t\delta_{st}\right\rangle_{AV}.\tag{27b}$$

Since $\tilde{\lambda}$ is a constant matrix, as before it can be taken outside the ensemble average, and so the second term in Eq. (27b) takes the form

$$-\left[\tilde{\lambda},\sum_r\,\omega_{rs}\langle x_r'\,\mathrm{Tr}\,\tilde{C}\sigma_t x_t'\rangle_{AV}\right],\tag{27c}$$

which again vanishes since the ensemble average inside the commutator is a matrix function only of $\tilde{\lambda}$ and τ. Contracting the remainder of Eq. (27b) with $\Sigma_s\omega_{us}$, using Eqs. (16c) and (23b), and noting that $[x_u',\sigma_t x_t']=[x_u,\sigma_t x_t]$ because the classical parts do not contribute to the commutator, we get as our final result

$$\langle[x_u,\sigma_t x_t]\rangle_{AV}=i_{\mathrm{eff}}\hbar\,\omega_{ut}\sigma_t-\tau\langle\dot{x}_u'\,\mathrm{Tr}\,\tilde{C}\sigma_t x_t'\rangle_{AV}.\tag{28a}$$

Equation (28a), like Eq. (20b) of the preceding section that it generalizes, is exact.

In order to have emergent quantum behavior, it is necessary that the second term on the right hand side of Eq. (28a) be much smaller than the first term, that is, we require

$$\hbar^{-1}\tau|\langle\dot{x}_u'\,\mathrm{Tr}\,\tilde{C}\sigma_t x_t'\rangle_{AV}|\ll 1.\tag{28b}$$

Again, by replacing $\tilde{C}$ by its expectation value at the start of the derivation leading to Eq. (28a), we get a second Ward identity

$$0 = i_{\mathrm{eff}}\hbar\,\omega_{ut}\sigma_t - \tau\langle \dot{x}'_u \ \mathrm{Tr}\ i_{\mathrm{eff}}\hbar\,\sigma_t x'_t\rangle_{AV}. \tag{28c}$$

Hence the inequality of Eq. (28b) can be satisfied only if $\tilde{C}$ is not an extensive quantity, since if $\tilde{C}$ were extensive one could, in the large system limit, approximate it in Eq. (28b) by its expectation $i_{\mathrm{eff}}\hbar$, giving an expression that, by Eq. (28c), cannot be small.

Because the derivation of this section is valid for fermions as well as bosons, one can in fact make a stronger statement about the conditions for emergent quantum behavior. Letting the indices t and u in Eq. (28a) be either both bosonic or both fermionic, we get the respective relations

$$\begin{aligned} \langle [q_r, p_r]\rangle_{AV} &= i_{\mathrm{eff}}\hbar - \tau\langle \dot{q}'_r \ \mathrm{Tr}\ \tilde{C}p'_r\rangle_{AV} \quad r \ \text{bosonic}, \\ \langle \{q_r, p_r\}\rangle_{AV} &= i_{\mathrm{eff}}\hbar - \tau\langle \dot{q}'_r \ \mathrm{Tr}\ \tilde{C}p'_r\rangle_{AV} \quad r \ \text{fermionic}. \end{aligned} \tag{29a}$$

Substituting this into Eq. (7) for $\tilde{C}$, taking the ensemble average, and using Eq. (16c), we get

$$\begin{aligned} i_{\mathrm{eff}}\hbar = \langle \tilde{C}\rangle_{AV} &= \left\langle \sum_{r\ \mathrm{bosons}} [q_r, p_r] - \sum_{r\ \mathrm{fermions}} \{q_r, p_r\} \right\rangle_{AV} \\ &= \left(\sum_{r\ \mathrm{bosons}} - \sum_{r\ \mathrm{fermions}} \right) i_{\mathrm{eff}}\hbar - \tau \left(\sum_{r\ \mathrm{bosons}} - \sum_{r\ \mathrm{fermions}} \right) \langle \dot{q}'_r \ \mathrm{Tr}\ \tilde{C}p'_r\rangle_{AV}, \end{aligned} \tag{29b}$$

which on division by $\hbar$ and transposition of terms gives

$$\left(\sum_{r\ \mathrm{bosons}} - \sum_{r\ \mathrm{fermions}} \right) \hbar^{-1}\tau\langle \dot{q}'_r \ \mathrm{Tr}\ \tilde{C}p'_r\rangle_{AV} = i_{\mathrm{eff}} \left(\sum_{r\ \mathrm{bosons}} - \sum_{r\ \mathrm{fermions}} - 1 \right). \tag{29c}$$

When the condition of Eq. (28b) for emergent canonical behavior is satisfied, the left hand side of Eq. (29c) is a sum of very small terms. Assuming that this sum yields at most a finite, bounded total, let us consider the case in which r includes the spatial label of a translation invariant field theory. Then the number of bosonic and fermionic modes per unit volume contributing on the right hand side of Eq. (29c) must be equal, since if not, the right hand side of Eq. (29c) would become infinite as the spatial volume grows to infinity, contradicting the boundedness of the left hand side. Therefore, in a complex Hilbert space (for a discussion of how our arguments must be modified in real and quaternionic Hilbert space, see the Appendix), a candidate pre-quantum mechanics theory must have equal numbers of bosonic and fermionic degrees of freedom, making it plausible that such a candidate theory should be supersymmetric. When the numbers of bosonic and fermionic modes are in balance, Eq. (29c) simplifies to

$$\left(\sum_{r\ \mathrm{bosons}} - \sum_{r\ \mathrm{fermions}} \right) \hbar^{-1}\tau\langle \dot{q}'_r \ \mathrm{Tr}\ \tilde{C}p'_r\rangle_{AV} = -i_{\mathrm{eff}}, \tag{29d}$$

showing that the remainder terms in Eq. (28a), that are neglected when Eq. (28b) is satisfied, sum in Eq. (29c) to give a total of unit magnitude.

Corresponding to the Ward identity of Eq. (28a), we can derive a class of more general Ward identities by replacing $\sigma_t x'_t$ in Eq. (25b) by a general U, constructed as a Weyl ordered (i.e., symmetrized) polynomial in the products $\{\sigma_r x_r\}$, with coefficients that are c-number functions of 1 and of i_{eff}. In place of Eq. (27b), we now get

$$0 = \left\langle \left(-\tau \sum_r \dot{x}'_r \omega_{rs} - \left[\tilde{\lambda}, \sum_r \omega_{rs} x'_r \right] \right) \mathrm{Tr}\ \tilde{C}U + \left[U, \sum_r \omega_{rs} x'_r \right] + \sum_{\text{each } x_s \text{ in } U} U(\text{one } x_s \to \tilde{C})' \right\rangle_{AV}. \tag{30a}$$

As long as U has coefficients that depend only on the matrices (or operators) 1 and i_{eff}, the second term on the right in Eq. (30a), which involves a commutator with $\tilde{\lambda}$, vanishes by the same arguments as before. Contracting the remainder with ω_{us}, but making no approximations, we get as the exact general Ward identity analogous to Eq. (28a),

$$\langle [x_u, U] \rangle_{AV} = \left\langle \sum_s \omega_{us} \sum_{\text{each } x_s \text{ in } U} U(\text{one } x_s \to \tilde{C})' \right\rangle_{AV} - \tau \langle \dot{x}'_u \, \text{Tr} \, \tilde{C} U \rangle_{AV} . \tag{30b}$$

Suppose now that we can make the following two approximations, (i) we replace $\tilde{C}$ in the first term on the right hand side of Eq. (30b) by its ensemble average $i_{\text{eff}}\hbar$, and (ii) we neglect the τ term in Eq. (30b). We then are left with the relation

$$\langle [x_u, U] \rangle_{AV} = \left\langle \sum_s \omega_{us} \sum_{\text{each } x_s \text{ in } U} U(\text{one } x_s \to i_{\text{eff}}\hbar)' \right\rangle_{AV} , \tag{30c}$$

which extends the effective canonical algebra inside ensemble averages to the commutator of x_u with a general Weyl ordered polynomial U.

By the methods of Appendix E of Ref. 4, Eq. (30c) can be extended to include sources for the remainder parts x'_r of the phase space variables. Specializing the relations obtained this way to the case $U = i_{\text{eff}}$ implies that i_{eff} can be freely commuted with phase space variables inside ensemble averages. As argued in Ref. 4, the resulting set of Ward identities then yields the canonical generator structure, including the time evolution relations, of Heisenberg picture quantum field theory. The only assertion in Ref. 4 that cannot be derived this way is the claim that the time evolution equation is more exact than the other generator relations; this claim used the assumption that $\tilde{C}$ can be replaced by its ensemble average inside the τ term, which we have seen is not correct. The remainder of the conclusions of Ref. 4 rest on the two approximations that we made above, which can be rephrased as the assumptions that, (i) in the terms of Eq. (30b) that involve the *unvaried* canonical ensemble ρ of Eq. (13) with a factor of $\tilde{C}$ in the integrand, the fact that the ensemble is sharply peaked around the mean allows us to replace the integrand factor $\tilde{C}$ by its ensemble average $i_{\text{eff}}\hbar$, and (ii) the canonical ensemble displays a certain rigidity, in that terms of the form $\int d\mu \, \delta\rho \, \text{Tr} \, \tilde{C} U$ can be dropped. On the other hand, we have seen that terms of the form $\int d\mu \, \delta\rho \, \text{Tr} \, i_{\text{eff}}\hbar U$ cannot be dropped; this does not contradict our assumptions because $\delta\rho$ can be rapidly varying around the peak of the ensemble.

V. DISCUSSION

We have seen that a statistical mechanics can be formulated for a wide class of matrix models with a global unitary invariance, and that within this statistical mechanics, the ensemble averages of canonical variables obey the exact relations of Eq. (20b), (28a), (29a), and (30b). When these can be approximated by dropping the τ terms [and, in Eq. (30b), replacing the $\tilde{C}$ insertions in U' by their ensemble averages], the result is emergent quantum mechanical behavior for the statistical ensemble averages. The condition for validity of the approximation of neglecting the τ terms is rather delicate: we have argued that it requires that $\tilde{C}$ should be an intensive thermodynamic quantity, and that the numbers of bosonic and fermionic degrees of freedom should balance.

We conclude with some brief remarks:

(1) In the first two references of Ref. 3, it is shown that one can readily formulate trace dynamics models in which global unitary operator invariance is gauged to give a local unitary operator invariance. Since global unitary invariance is a special case of local unitary invariance, the considerations of this paper apply to these models.

(2) In Refs. 5, it is shown that supersymmetric Yang Mills theory, and the related ''matrix model for *M* theory,'' fit naturally into the trace dynamics framework analyzed in this paper. In these models, $\tilde{C}$ vanishes up to a surface term contribution, a behavior consistent with its being an intensive thermodynamic quantity.

(3) Although our final results of Eqs. (20b), (28a), and (29a) superficially resemble the string-inspired formula of Eq. (1b), there is an important difference. In Eq. (1b) the coordinate is a quantum operator, as is usual in nonrelativistic quantum mechanics, whereas in our results of Eqs. (20b), etc., the coordinate is merely a degree of freedom label r, as it always is in quantum field theory, and the coordinates and momenta are canonical field variables with label r. It may be possible to make a connection between the two types of modified commutation relations when the metric structure of the coordinate manifold is taken into account, using the

fact that the proper distance is related to the coordinate interval by $ds^2=g_{\mu\nu}dx^\mu dx^\nu$. In a field theoretic interpretation, dx^μ is just the change in the degree of freedom label, but the metric is a dynamical variable, and hence so is ds. This suggests that there may be an analog of Eq. (1b) involving the relativistic proper distance, and that this is the relation to be compared with our results in this paper.

(4) As is well known, in a complex Hilbert space the canonical algebra $[q,p]=i$, $[q,i]=[p,i]=0$ cannot have finite dimensional representations, since this algebra implies, for example, the relation $q^2p^2+p^2q^2-2qp^2q=-2$, which in a finite dimensional Hilbert space would have a left hand side with trace zero and a right hand side with trace nonzero. However, it is consistent for the canonical algebra to emerge as the limit $N\to\infty$ of an algebra in an N dimensional Hilbert space, which is the behavior argued for in Ref. 4 and here. Because the emergent canonical algebra involves not the imaginary unit i of the underlying complex Hilbert space, but rather the operator $i_{\rm eff}$ with Tr $i_{\rm eff}=0$, a basis for the operator algebra in the emergent theory is provided by a set of operators that commute with $i_{\rm eff}$, together with one additional operator that anticommutes with $i_{\rm eff}$, and that plays the role of the time reversal operator in the emergent complex quantum mechanics. In fact, because the condition Tr $i_{\rm eff}=0$ implies that one can find a representation in which $i_{\rm eff}$ is a real matrix (just as for Pauli matrices $\rho_{1,2,3}$ the matrix $i\rho_3$ can be given the real form $i\rho_2$ by a change of representation), the quantum mechanics emergent from matrix dynamics has the structure of a complexified real quantum mechanics, for which the operator algebra has the form just described (see, e.g., Sec. 2.6 of the second citation in Ref. 3).

(5) We have seen that the emergence of quantum mechanical behavior from matrix model statistical mechanics requires a certain ''rigidity'' of the statistical ensemble. It is easy to see[8,9] that this rigidity is a sufficient condition for the canonical and microcanonical ensembles to give the same Ward identities, and hence the same emergent quantum behavior. The need for a rigid statistical ensemble in our context suggests a possible analogy with the concept of London rigidity in the theory of superconductivity.[10] In the presence of an applied vector potential $\mathbf{A}$, the induced current density in a metal is given by

$$\langle \mathbf{j} \rangle = -\frac{ne}{m}\left\langle \mathbf{p}+\frac{e}{c}\mathbf{A}\right\rangle. \tag{31}$$

In a normal metal the two terms on the right hand side of Eq. (31) nearly cancel, leaving a small residual diamagnetism. However, in a superconductor the rigidity of the wave function leads to the vanishing of $\langle \mathbf{p} \rangle$, giving perfect diamagnetism and the Meissner effect. An analogy with the results of this paper would equate normal metal behavior with the case in which $\tilde{C}$ can be replaced by its ensemble average in the τ term; in this case the right hand side of Eq. (20b) is approximately equal to the right hand side of Eq. (22b), leading to cancellation of the emergent canonical commutator. Similarly, the analogy would equate superconducting behavior with the case in which the τ term containing $\tilde{C}$ can be dropped because of ''rigidity'' of $\delta\rho$, leading through Eqs. (20b), (28a), and (29a) to an emergent canonical commutator as an analog of the superconductive Meissner effect.

ACKNOWLEDGMENTS

The work of S.L.A. was supported in part by the Department of Energy under Grant No. DE-FG02-90ER40542, and he wishes to acknowledge a stimulating conversation with Walter Troost. He also wishes to acknowledge the hospitality of the Aspen Center for Physics, where this work was completed, and of the Department of Applied Mathematics and Theoretical Physics at Cambridge University. A.K. in turn wishes to acknowledge the hospitality of the Institute for Advanced Study in Princeton.

APPENDIX: REAL AND QUATERNIONIC HILBERT SPACE

For reasons that we now describe, in real and quaternionic Hilbert space the arguments of this paper must be modified and yield weaker conclusions. The underlying reason for this modification is that only in complex Hilbert space can one have a nonreal trace that nonetheless obeys the cyclic property. In quaternionic Hilbert space, as a consequence of the noncommutativity of the

quaternions, only the real part of the ordinary trace obeys the cyclic property. In real Hilbert space, the trace is necessarily real, and the trace of any anti-self-adjoint operator vanishes. Thus in these two cases, if one follows Refs. 3 and 4 and defines the graded trace **Tr** to be the one that obeys the cyclic property, then the definition must include taking the real part, and to get a nonzero result one must require the operator argument V of **Tr** V to be self-adjoint. As a consequence, in the general case derivation analogous to that starting from Eq. (25b), one must consider $\mathbf{Tr}\{i_{\rm eff},\widetilde{C}\}\sigma_t x_t'$ rather than $\mathrm{Tr}\,\widetilde{C}\sigma_t x_t'$. This gives the following analog of Eq. (28a):

$$\langle[x_u,\sigma_t\{i_{\rm eff},x_t\}]\rangle_{AV}=\{i_{\rm eff},i_{\rm eff}\hbar\}'\omega_{ut}\sigma_t-\tau\langle\dot{x}_u'\ \mathrm{Tr}\{i_{\rm eff},\widetilde{C}\}\sigma_t x_t'\rangle_{AV}, \tag{A1}$$

where the prime on the first term on the right hand side of Eq. (A1) indicates extraction of the traceless part. However, since the traceless part of $\{i_{\rm eff},i_{\rm eff}\hbar\}=-2\hbar$ is zero, Eq. (A1) becomes

$$\langle[x_u,\sigma_t\{i_{\rm eff},x_t\}]\rangle_{AV}=-\tau\langle\dot{x}_u'\ \mathrm{Tr}\{i_{\rm eff},\widetilde{C}\}\sigma_t x_t'\rangle_{AV}, \tag{A2}$$

which has a structure more like the equation obtained by subtracting Eq. (28c) from Eq. (28a) than like Eq. (28a) itself. In complex Hilbert space, Eqs. (A2) *and* Eqs. (28a, 28c) all hold, and conditions for emergent quantum behavior can be formulated from the latter as in the text; in the real and quaternionic Hilbert space cases, only Eq. (A2) holds, and we cannot proceed with the analysis of the text.

The remarks made here correct a subtle error made in Sec. 6 of Adler and Millard.[4] There, the classical sources introduced in this paper were set to zero, and Ward identities were derived by varying x_s, not just the noncommutative part x_s' as we have done here. This gives as the Ward identity analogous to Eq. (A1)

$$\langle[x_u,\sigma_t\{i_{\rm eff},x_t\}]\rangle_{AV}=\{i_{\rm eff},i_{\rm eff}\hbar\}\omega_{ut}\sigma_t-\tau\langle\dot{x}_u\ \mathrm{Tr}\{i_{\rm eff},\widetilde{C}\}\sigma_t x_t\rangle_{AV}. \tag{A3}$$

Equation (A3) is a valid relation, but it actually implies *two* relations, quite different in structure, for its classical or c-number part and its primed or traceless part. Separating Eq. (A3) into these parts, and ignoring interference terms between x^c and x' in the τ term, one finds that the primed part of Eq. (A3) gives Eq. (A2), while the c-number part of Eq. (A3) becomes an equipartition identity for x^c and gives no direct information about the expectation of the commutator. Hence the arguments of Ref. 4 do not, in their present form, provide evidence for emergent quantum behavior in the real and quaternionic Hilbert space cases.

[1] See, e.g., D. J. Gross and P. F. Mende, Nucl. Phys. B **303**, 407 (1988); D. Amati, M. Cialfaloni, and G. Veneziano, Phys. Lett. B **216**, 41 (1989); M. Maggiore, *ibid.* **319**, 83 (1993); recent reviews are E. Witten, Phys. Today **49**, 24 (1996); L. J. Garay, Int. J. Mod. Phys. A **10**, 145 (1995).

[2] A. Kempf, J. Math. Phys. **35**, 4483 (1994); A. Kempf, G. Mangano, and R. B. Mann, Phys. Rev. D **52**, 1108 (1995); A. Kempf and H. Hinrichsen, J. Math. Phys. **37**, 2121 (1996); A. Kempf (preprint) DAMTP/94-33, hep-th/9405067; J. Math. Phys. **38**, 1347 (1997); Phys. Rev. D **54**, 5174 (1996); A. Kempf and G. Mangano, *ibid.* **55**, 7909 (1997); A. Kempf (preprint) DAMTP/96-101, hep-th/9612082, in Proc. XXI Int. Coll. on Group Theor. Methods in Physics, Goslar, July 1996 (to be published).

[3] S. L. Adler, Nucl. Phys. B **415**, 195 (1994); *Quaternionic Quantum Mechanics and Quantum Fields* (Oxford University Press, New York, 1995), Secs. 13.5–13.7 and Appendix A; S. L. Adler, G. V. Bhanot, and J. D. Weckel, J. Math. Phys. **35**, 531 (1994); S. L. Adler and Y.-S. Wu, Phys. Rev. D **49**, 6705 (1994).

[4] S. L. Adler and A. C. Millard, Nucl. Phys. B **473**, 199 (1996).

[5] S. L. Adler, Phys. Lett. B **407**, 229 (1997); Nucl. Phys. B **499**, 569 (1997).

[6] B. de Wit, J. Hoppe, and H. Nicolai, Nucl. Phys. B **305**, 545 (1988); P. K. Townsend, Phys. Lett. B **373**, 68 (1996); T. Banks, W. Fischler, S. H. Shenker, and L. Susskind, ''*M* Theory as a Matrix Model: A Conjecture,'' hep-th/9610043. This paper gives an extensive bibliography, and has inspired a large number of very recent preprints on the relationship between string theory and matrix models. For further bibliography, see the first paper of Ref. 5.

[7] A. C. Millard, ''Noncommutative Methods in Quantum Mechanics,'' Ph.D. thesis, Princeton University, April, 1997.

[8] S. L. Adler and L. P. Horwitz, J. Math. Phys. **37**, 5429 (1996).

[9] S. L. Adler and L. P. Horwitz (unpublished).

[10] J. R. Schrieffer, *Theory of Superconductivity* (Benjamin, New York, 1964), Sec. 8-1, p. 204.

Closure of field operators, asymptotic Abelianness, and vacuum structure in hyperfunction quantum field theory

E. Brüning
Department of Mathematics and Applied Mathematics, University of Durban–Westville, Private Bag X54001, Durban 4000, Republic of South Africa

S. Nagamachi
Department of Mathematics, Faculty of Engineering, Tokushima University, Tokushima 770, Japan

(Received 2 September 1997; accepted for publication 18 May 1998)

Though, in contrast to standard quantum field theory (QFT), the tensor-algebra over the test-function space of hyperfunction quantum field theory has no local structure, the localization properties of states on this algebra can be used to derive asymptotic Abelianness in spacelike directions. Again, in contrast to standard QFT, the closure of (Hermitian) field operators can destroy localization properties. This problem is addressed in a natural modification of the definition of the closure, called the local closure. This allows one, in conjunction with asymptotic Abelianness, to define a proper reduction of the field algebra to the subspace of the translation invariant states, and to investigate the dimension of this subspace. © *1998 American Institute of Physics.* [S0022-2488(98)00810-X]

I. INTRODUCTION

In general relativistic quantum field theory one usually expects that an irreducible quantum field has a one-dimensional vacuum subspace. However, in standard quantum field theory (QFT), i.e., in the general theory of quantized fields in the sense of Gårding and Wightman,[1] Borchers[2] indicated that an irreducible quantum field can have a vacuum subspace of dimension larger than one. Usually one refers to the fact that the vacuum subspace has a dimension larger than one as *Borchers's anomaly*. Its appearance is considered as an unwanted pathology of standard QFT.

The mathematical analysis of the dimension of the vacuum subspace of an irreducible relativistic quantum field in standard QFT has to deal with a number of difficult and to a large extent technical problems. Thus a complete analysis is still missing. For a more detailed discussion of Borchers's anomaly and its (technical) difficulties we refer to Ref. 3. Therefore a discussion of this anomaly in a more general framework might offer a new perspective. Accordingly we have asked ourselves whether Borchers's anomaly can occur in hyperfunction quantum field theory (HFQFT), i.e., general relativistic quantum field theory in terms of Fourier hyperfunctions as introduced in Ref. 4 and systematically investigated and extended in Ref. 5.

Besides being more general than standard QFT, HFQFT allows ''more'' explicit models to be analyzed in a rigorous way.[6] As it is well known, the question of which type of generalized functions one should use in the formulation of relativistic quantum field theory has not yet found a final answer. Thus we like to consider our present discussion as a contribution to an answer to this question in favor of Fourier hyperfunctions. Other reasons why it might be desirable to consider quantum field theory in terms of hyperfunctions are given in Ref. 7.

Though various technical aspects are simpler in HFQFT than in standard QFT (mainly related to the fact that the sheaf of Fourier hyperfunctions is flabby while the sheaf of distributions is not flabby), HFQFT has to live with the following two difficulties: First, the tensor-algebra over the test function space of HFQFT has no local structure (this means for instance that there are no nontrivial test functions with compact support, and there are no pairs of test functions with spacelike separated supports). Nevertheless, states on this algebra can have local properties and are localizable in the strict sense just as tempered distributions (most clearly expressed by the fact that hyperfunctions form a sheaf of complex vector spaces). This is used to derive asymptotic

 © 1998 American Institute of Physics

Abelianness in spacelike directions in hyperfunction quantum field theory. Second, the closure of (Hermitian) field operators in HFQFT can destroy localization properties. This problem is solved by proposing a suitable definition of the closure, called the local closure, and this definition is applied successfully, in conjunction with asymptotic Abelianness, to define a reduction of the field algebra to the subspace of translation invariant states. This allows one to investigate the vacuum structure of the theory and to show that Borchers's anomaly[2] can indeed occur in HFQFT.

Now we indicate briefly the contents of the following subsections.

In Sec. II, we recall a few basic facts about Fourier hyperfunctions. In Sec. III, we formulate hyperfunction quantum field theory in terms of the sequence of vacuum expectation values of the field as a state $\mathscr{W}$ on the tensor algebra $\mathscr{B}$ over the test function space $Q(\mathbf{D}^4)$ of HFQFT. In Ref. 1 it is shown how this approach is related to the standard one. Since the test function space of Fourier hyperfunctions contains no nontrivial function with compact support, axioms have to be modified from those for tempered fields. In order to define the local closure a *-algebra $\mathscr{C}$ is introduced which is an extension of $\mathscr{B}$ and which contains enough information about locality. It is shown in Sec. V that the state $\mathscr{W}$ can be extended to a state on $\mathscr{C}$. In Sec. VI, we study a representation Π of $\mathscr{C}$ obtained by GNS construction of this extended state, and we prove the asymptotic Abelianness of $\Pi(\mathscr{B})$ and $\Pi(\mathscr{C})$. In Sec. VII, we define the local closure $\hat{\Pi}$ of Π and prove asymptotic Abelianness of $\hat{\Pi}(\mathscr{C})$. The domain $\hat{\mathscr{D}}$ of the local closure $\hat{\Pi}$ is such that it allows one to define a reduction of the field algebra to the space of translation invariant states, and to prove that the reduced algebra is Abelian. Thus, in Sec. VIII, we can discuss Borchers's anomaly in HFQFT.

II. FOURIER HYPERFUNCTIONS

In distribution theory, one introduces the space of tempered distributions in order to have a natural domain for the Fourier transformation. In hyperfunction theory too, we introduce the space of Fourier hyperfunctions, i.e., hyperfunctions with some growth restrictions at infinity which are needed for defining the Fourier transformation as an automorphism. A way to achieve this is to compactify the underlying space $\mathbf{R}^n$. The radial compactification $\mathbf{D}^n$ of $\mathbf{R}^n$ has proved to be very useful. It is defined in a natural way as follows: Let S_∞^{n-1} be the $(n-1)$-dimensional sphere at infinity, which is homeomorphic to the unit sphere $S^{n-1}=\{x\in\mathbf{R}^n;|x|=1\}$ by the mapping $x\to x_\infty$, where the point $x_\infty\in S_\infty^{n-1}$ lies on the ray connecting the origin with the point $x\in S^{n-1}$. The set $\mathbf{R}^n\cup S_\infty^{n-1}$ equipped with its natural topology is denoted by $\mathbf{D}^n$ (a fundamental system of neighborhoods of x_∞ is given by all open cones of arbitrary vertex generated by arbitrary open neighborhoods of x_∞ in S_∞^{n-1}, see Ref. 5).

The spaces of test functions we are going to use later are inductive limit spaces

$$\underset{\sim}{\mathscr{O}}(\mathbf{D}^n)$$

of functions on $\mathbf{D}^n$ for $n=4k$ $(k=1,2,...)$ which we describe now. For reasons which become obvious soon we introduce and study the spaces

$$\underset{\sim}{\mathscr{O}}(K)$$

for a closed subset K of $\mathbf{D}^n$.

Let $\{U_m;m\in\mathbf{N}\}$ be a fundamental sequence of neighborhoods of K in $Q^n=\mathbf{D}^n+i\mathbf{R}^n$, with $U_m\supset\supset U_{m+1}$. Then let $\mathscr{O}_c^m(U_m)$ be the Banach space of functions f analytic on $U_m\cap\mathbf{C}^n$ and continuous on $U_m\cap\mathbf{C}^n$ such that

$$\|f\|_m=\sup_{z\in U_m\cap\mathbf{C}^n}|f(z)|e^{|z|/m} \tag{2.1}$$

is finite. The space $\mathscr{O}(K)$ is now defined as the inductive limit of the Banach spaces $\mathscr{O}_c^m(U_m)$:

$$\underset{\sim}{\mathscr{O}}(K)=\operatorname{ind}\lim_{m\to\infty}\mathscr{O}_c^m(U_m). \tag{2.2}$$

$\mathcal{Q}(K)$ consists of rapidly decreasing analytic function on K. Its basic properties are collected in Proposition 2.1 (see Ref. 4).

Proposition 2.1:

(a) $\mathcal{Q}(K)$ is a DFS-space (a dual Fréchet–Schwartz space), i.e., the all embedding maps $\mathcal{O}_c^m(U_m)\to\mathcal{O}_c^{m+1}(U_{m+1})$. $m=1,2,...$, are compact.

(b) $\mathcal{Q}(\mathbf{D}^n)$ is nuclear and barreled.

(c) $\otimes^k\mathcal{Q}(\mathbf{D}^4)$ is dense in $\mathcal{Q}(\mathbf{D}^{4k})$ and the completion $\hat{\otimes}^k\mathcal{Q}(\mathbf{D}^4)$ of $\hat{\otimes}^k\mathcal{Q}(\mathbf{D}^4)$ is equal to $\mathcal{Q}(\mathbf{D}^{4k})$.

Let $\mathcal{Q}(Q^n)$ be the linear space of rapidly decreasing entire functions $f(z)$ in $\mathbf{C}^n$, i.e., for every $r>0$

$$\sup\{|f(z)|e^{\eta|z|};z\in\mathbf{C}^n,\quad |\mathrm{Im}\ z|<r\}<\infty$$

for some $\eta=\eta(r)>0$. Let V be an open set in Q^n. Define a Banach space $\mathcal{Q}_c(V;\eta)$ of functions f analytic in $V\cap\mathbf{C}^n$, continuous on $V\cap\mathbf{C}^n$ such that

$$\|f\|_{V,\eta}=\sup_{z\in V\cap\mathbf{C}^m}|f(z)|e^{\eta|z|}$$

is finite. Then we have the following proposition (see Proposition 2.6 of Ref. 8).

Proposition 2.2: Let K be a compact set in $\mathbf{D}^n$, V an open neighborhood of K in Q^n. Then there exists an open neighborhood W with $K\subset W\subset V$ such that every $f\in\mathcal{Q}_c(V;\eta)$ is approximated by a sequence $\{f_j\}\subset\mathcal{Q}(Q^n)$ in the norm $\|\cdot\|_{W,\eta/4}$. In particular, $f\in\mathcal{Q}(K)$ is approximated by a sequence $\{f_j\}\subset\mathcal{Q}(\mathbf{D}^n)$ in some norm $\|\cdot\|_m$ of $\mathcal{O}_c^m(U_m)$.

The following proposition gives the definition and the basic properties of the Fourier transformation $\mathcal{F}$ on $\mathcal{Q}(\mathbf{D}^n)$.

Proposition 2.3: The Fourier transformation $\mathcal{F}$ is well defined on $\mathcal{Q}(\mathbf{D}^n)$ by

$$(\mathcal{F}f)(p)=(2\pi)^{-n/2}\int e^{ip\cdot x}f(x)dx. \tag{2.3}$$

$\mathcal{F}$ is an isomorphism of the topological vector space $\mathcal{Q}(\mathbf{D}^n)$ with inverse

$$(\overline{\mathcal{F}}\tilde{f})(x)=(2\pi)^{-n/2}\int e^{-ip\cdot x}\tilde{f}(p)dp. \tag{2.4}$$

A Fourier hyperfunction on $\mathbf{D}^n$ is by definition an element of the topological dual $\mathcal{Q}(\mathbf{D}^n)'$ of the space $\mathcal{Q}(\mathbf{D}^n)$. Therefore we can extend the above definition of the Fourier transformation to Fourier hyperfunctions by duality, i.e., for all $f\in\mathcal{Q}(\mathbf{D}^n)$ and $\mu\in\mathcal{Q}(\mathbf{D}^n)'$ one defines

$$(\mathcal{F}\mu)(f)=\mu(\mathcal{F}f). \tag{2.5}$$

Then $\mathcal{F}$ is an isomorphism on $\mathcal{Q}(\mathbf{D}^n)'$.

If $K\subset\mathbf{D}^n$ is any closed subset then $\mathcal{Q}(\mathbf{D}^n)$ is contained in $\mathcal{Q}(K)$. For a Fourier hyperfunction μ on $\mathbf{D}^n$, denote by $C(\mu)$ the class of all those closed subsets $K\subset\mathbf{D}^n$ such that there is a continuous extension μ_K of μ to $\mathcal{Q}(K)$. Any such subset $K\in C(\mu)$ is called a carrier of the Fourier hyperfunction μ. μ has a smallest carrier, called the support of μ (see Ref. 8):

$$\mathrm{supp}\ \mu=\cap_{K\in C(\mu)}K.$$

Here we want to stress the following facts. For an open set Ω in $\mathbf{D}^n$, the set $\mathcal{R}(\Omega)$ of Fourier hyperfunction on Ω is defined by

$$\mathcal{R}(\Omega)=\mathcal{Q}(\mathbf{D}^n)'/\mathcal{Q}(\mathbf{D}^n\setminus\Omega)'. \tag{2.6}$$

The notion of support of a Fourier hyperfunction expresses their localization properties on $\mathbf{D}^n$. But a more complete description of these localization properties of Fourier hyperfunctions is contained in the fact that the correspondence $\Omega\to\mathcal{R}(\Omega)$ forms a (flabby) sheaf[9] on $\mathbf{D}^n$, and is thus well

suited to express the localization properties as needed in quantum field theory (see Ref. 5). However, in contrast to the space $\mathscr{S}'(\Omega)$ of tempered distribution on Ω, we cannot endow the space $\mathscr{R}(\Omega)$ with any reasonable topology since $\mathscr{D}(\mathbf{D}^n\backslash\Omega)'$ is dense in $\mathscr{D}(\mathbf{D}^n)'$ (see Lemma 8.6.4 of Ref. 9).

III. HYPERFUNCTION QUANTUM FIELDS

Before we can formulate concisely our basic assumptions some notation has to be introduced. Let us begin with the tensor algebra $\mathscr{B}$ over the test function space of hyperfunction quantum field theory defined as the locally convex direct sum of the spaces $\mathscr{B}_n=\mathscr{D}(\mathbf{D}^{4n})$:

$$\mathscr{B}=\oplus_{n=0}^{\infty}\mathscr{B}_n, \quad \mathscr{B}_0=\mathbf{C}, \quad \mathscr{B}_n=\mathscr{D}(\mathbf{D}^{4n}), \quad n=1.2,\ldots . \tag{3.1}$$

A continuous involution * is defined on $\mathscr{B}$ by

$$f^*=(\overline{f}^{[0]},\overline{f}^{[1]}(x_1),\ldots,\overline{f}^{[n]}(x_n,\ldots,x_1),\ldots,0,0,\ldots)\in\mathscr{B} \tag{3.2}$$

for all $f=(f^{[0]},f^{[1]}(x_1),\ldots,f^{[n]}(x_1,\ldots,x_n),\ldots,0,0,\ldots)\in\mathscr{B}$.
Furthermore, a noncommutative multiplication $\otimes$ is defined on $\mathscr{B}$ by

$$(f\otimes g)^{[n]}(x_1,\ldots,x_n)=\sum_{k=0}^{n} f^{[k]}(x_1,\ldots,x_k)g^{[n-k]}(x_{k+1},\ldots,x_n). \tag{3.3}$$

for all $f, g\in\mathscr{B}$. With these operations $\mathscr{B}$ is a graded topological *-algebra called the (completed) *tensor algebra of test functions* for hyperfunction quantum field theory. $\mathscr{B}$ is the completed tensor algebra over the test function space $\mathscr{D}(\mathbf{D}^4)$. A state $\mathscr{W}$ on $\mathscr{B}$ is a continuous linear functional on $\mathscr{B}$ which is non-negative in the sense that $\mathscr{W}(f^*\otimes f)\geqslant 0$ for all $f\in\mathscr{B}$ and which is normalized to one on the unit element $\mathbf{1}=(1,0,\ldots)$ of $\mathscr{B}$, i.e., $\mathscr{W}(\mathbf{1})=1$.

The proper Poincaré group $G=\mathscr{P}_+^{\uparrow}$ acts naturally on $\mathbf{D}^4$. This induces an action α of G as automorphisms of $\mathscr{B}$ according to the following formula:

$$\alpha_{(a,\Lambda)}(f)=f_{(a,\Lambda)} \tag{3.4}$$

for all $f\in\mathscr{B}$ and all $(a,\Lambda)\in G$ where for $n=1,2,\ldots$,

$$f^{[n]}_{(a,\Lambda)}(x_1,\ldots,x_n)=f^{[n]}(\Lambda^{-1}(x_1-a),\ldots,\Lambda^{-1}(x_n-a)). \tag{3.5}$$

Denote by $\overline{V}_+=\{p=(p^0,\mathbf{p})\in\mathbf{R}^4;\ p^2=(p^0)^2-\mathbf{p}^2\geqslant 0,\ p^0\geqslant 0\}$ the closed forward light cone and introduce the sets

$$\mathscr{S}_n=\{(p_1,\ldots,p_n)\in\mathbf{R}^{4n};p_1\in\overline{V}_+,p_1+p_2\in\overline{V}_+,\ldots,p_1+\cdots+p_n\in\overline{V}_+\}.$$

The Fourier transformation $\mathscr{F}$ on $\mathscr{D}(\mathbf{D}^n)$ extends naturally to $\mathscr{B}$. The Fourier transform $\mathscr{F}f$ of $f\in\mathscr{B}$ is denoted by $\tilde{f}$. Now we can introduce the space

$$E_{\mathrm{sp}}=\{f=(f^{[0]},\ldots,f^{[n]},\ldots,0,0,\ldots); \quad \tilde{f}^{[n]}(p_1,\ldots,p_n)\in\mathscr{D}(\overline{S}_n)\}\supset\mathscr{B} \tag{3.6}$$

where $\overline{S}_n$ is the closure of S_n in $\mathbf{D}^{4n}$. The space E_{sp} is equipped with the topology induced by the spaces $\mathscr{D}(\overline{S}_n)$ for $n=1,2,\ldots$. Next, let $\overline{K}_i^n$ denote the closure of the set $K_i^n=\{(x_1,\ldots,x_n)\in\mathbf{R}^{4n}:(x_i-x_{i+1})^2\geqslant 0\}$ in $\mathbf{D}^{4n}$. To each $g\in\mathscr{D}(\overline{K}_i^n)$ assign the function $f=c_i^n(g)$ defined by

$$f(x_1,\ldots,x_n)=g(x_1,\ldots,x_i,x_{i+1},\ldots,x_n)-g(x_1,\ldots,x_{i+1},x_i,\ldots,x_n). \tag{3.7}$$

For $n=2,3,\ldots$ introduce the vector spaces

$$E_{\diagup,n}=\mathrm{span}\{c_i^n(g_i):g_i\in\mathscr{D}(\overline{K}_i^n), \quad i=1,\ldots,n-1\} \tag{3.8}$$

and the locally convex sum of these spaces

$$E_{\mathscr{O}}=\oplus_{n=0}^{\infty}E_{\mathscr{O},n},\quad E_{\mathscr{O},0}=E_{\mathscr{O},1}=\{0\}. \tag{3.9}$$

The following properties of the space $E_{\mathscr{O}}$ are of interest to us.

Proposition 3.1:

(a) $E_{\mathscr{O}}$ is invariant under the involution *, i.e., $E_{\mathscr{O}}^{*}\subset E_{\mathscr{O}}$.

(b) $E_{\mathscr{O}}$ is a $\mathscr{B}$ module, i.e., for all $f\in E_{\mathscr{O}}$ and all $g\in\mathscr{B}$ we have $f\otimes h\in E_{\mathscr{O}}$ and $h\otimes f\in E_{\mathscr{O}}$.

Proof: For $g\in\tilde{\mathscr{O}}(\overline{K}_i^n)$ we know $g^*\in\tilde{\mathscr{O}}(K_{n-i}^n)$. Now from

$$f^{*[n]}(x_1,\ldots,x_n)=g^*(x_1,\ldots,x_i,x_{i+1},\ldots,x_n)-g^*(x_i,\ldots,x_{i+1},x_i,\ldots,x_n)$$

it follows that $f^*\in E_{\mathscr{O}}$ whenever $f\in E_{\mathscr{O}}$. Thus part (a) follows. In order to prove (b) assume that $f\in E_{\mathscr{O}}$ and $h\in\mathscr{B}$ are given. Since

$$(f\otimes h)^{[n]}(x_1,\ldots,x_n)=\sum_{k=1}^{n}f^{[k]}(x_1,\ldots,x_k)h^{[n-k]}(x_{k+1},\ldots,x_n)$$

and $f^{[k]}(x_1,\ldots,x_k)$ is a sum of elements of the form

$$g(x_1,\ldots,x_i,x_{i+1},\ldots,x_k)-g(x_1,\ldots,x_{i+1},x_i,\ldots,x_k)$$

it follows $f\otimes h\in E_{\mathscr{O}}$. In the same way we have $h\otimes f\in E_{\mathscr{O}}$. □

Hyperfunction quantum field theory, i.e., relativistic quantum field theory in terms of Fourier hyperfunctions, is now defined by the following list of conditions. We consider here only the simplest version, i.e., one neutral scalar field.

H_0 *Quantum fields as operator valued Fourier hyperfunctions:* The sequence of vacuum expectation values $\mathscr{W}$ is a state on the test function algebra $\mathscr{B}$.

H_1 *Relativistic covariance:* The state $\mathscr{W}$ is invariant under the Poincaré group $G=\mathscr{P}_+^{\uparrow}$, i.e., for all $f\in\mathscr{B}$ and all $(a,\Lambda)\in G$ we have

$$\mathscr{W}(f_{(a,\Lambda)})=\mathscr{W}(f).$$

H_2 *Spectral condition:* The energy momentum spectrum of the theory is contained in the closed forward light cone $\overline{V}_+=\{p=(p^0,\mathbf{p})\in\mathbf{R}^4;\ p^2=(p^0)^2-\mathbf{p}^2\geqslant 0,\ p^0\geqslant 0\}$. i.e., the state $\mathscr{W}\in\mathscr{B}'$ has a continuous linear extension to the space E_{sp}.

H_3 *Causality or local commutativity:* For spacelike separations field operators ''commute,'' i.e., for all $n=2,3,\ldots$ and all $i=1,\ldots,\underline{n}-1$ the functionals $\mathscr{W}\circ c_i^n:\mathscr{B}_n=\tilde{\mathscr{O}}(\mathbf{D}^{4n})\to\mathbf{C}$ have continuous linear extensions to $\tilde{\mathscr{O}}(K_i^n)$. (We say for brevity $\mathscr{W}\in\mathscr{B}'$ has a continuous linear extension to $E_{\mathscr{O}}$.)

Remark 3.2: (a) In our investigations we will actually only rely on the invariance of the state $\mathscr{W}$ under the subgroup of translations T_4.

(b) The restriction $\mathscr{W}_n=\mathscr{W}_{|\mathscr{B}_n}$ is a Fourier hyperfunction of $4n$ variables. The spectral condition $(\underline{H}_2)$ means that the support of the Fourier transform $\tilde{\mathscr{W}}_n(p_1,\ldots,p_n)$ of $\mathscr{W}_n$ is contained in S_n. Similarly, the condition of local commutativity (H_3) means that the support of

$$\mathscr{W}_n(x_1,\ldots,x_i,x_{i+1},\ldots,x_n)-\mathscr{W}_n(x_1,\ldots,x_{i+1},x_i,\ldots,x_n)$$

is contained in $\overline{K}_i^n$.

There are some immediate consequences of the above assumptions.

Proposition 3.3:

(a) Condition (H_0) implies the existence of a strongly cyclic *-representation Π of the *-algebra $\mathscr{B}$ in a complex Hilbert space $\mathscr{H}$ with strongly cyclic unit vector Ω and domain $\mathscr{D}(\Pi)=\Pi(\mathscr{B})\Omega$ such that for all $f\in\mathscr{B}$,

$$\mathscr{W}(f)=(\Omega,\Pi(f)\Omega)$$

where $(\cdot,\cdot)$ denotes the inner product of $\mathscr{H}$.

(b) If in addition condition (H_1) is taken into account there is a strongly continuous unitary

representation U of the Poincaré group G on $\mathscr{H}$ such that for all $f \in \mathscr{B}$ and all $(a,\Lambda) \in G$ one has

$$U(a,\Lambda)\Pi(f)U(a,\Lambda)^{-1} = \Pi(f_{(a,\Lambda)}).$$

(c) Conditions $(H_0)-(H_2)$ imply that the spectrum $\sigma(P)$ of the generator $P=(P^0,P^1,P^2,P^3)$ of the translations in the representation U, i.e., the energy momentum operator, is contained in the closed forward light cone $\overline{V}_+$.

(d) Assume $(H_0)-(H_3)$ and denote $A(f)=\Pi(j_1(f))$ for all $f \in \mathscr{B}_1$ where j_1 is the natural embedding of $\mathscr{B}_1$ into $\mathscr{B}$. Then, in addition to the above, we have for all $\Psi \in \mathscr{D}=\mathscr{D}(\Pi)$, as an identity for vector-valued Fourier hyperfunctions,

$$A(x)A(y)\Psi = A(y)A(x)\Psi$$

for all spacelike separated points x, y, i.e., $(x-y)^2<0$.

Proof: Since this is basically the generalized GNS construction we refer to the literature (Proposition 3.1.4 of Ref. 3). □

Compared to general quantum field theory in the sense of Wightman[1] the above list of hypotheses does not contain the condition of uniqueness of the (cyclic) vacuum state and the condition of completeness of the theory (irreducibility of the field A). In the framework given by the above hypotheses $(H_0)-(H_3)$ we intend to investigate the relation between these two conditions.

The subspace of translation invariant states in the Hilbert space $\mathscr{H}$ is

$$\mathscr{H}_0 = \{\Psi \in \mathscr{H} : U(a)\Psi = \Psi \quad \text{for all} \quad a \in T_4\}, \tag{3.10}$$

where $U(a)=U(a,1)$ denotes the representation of the translation group in $\mathscr{H}$ as given in part (b) of Proposition 3.2. By the above the cyclic unit vector Ω is contained in $\mathscr{H}_0$. Based on experiences from local quantum field theory it was believed that an irreducible quantum field theory has to have a unique cyclic vacuum state, i.e., $\dim \mathscr{H}_0 = 1$, until in 1965 Borchers[2] indicated that $d_0 = \dim \mathscr{H}_0 = \infty$ could occur since there are infinite dimensional irreducible representations of Abelian *-algebras. Since then various authors have investigated ''Borchers's anomaly'' but a comprehensive analysis is still missing.[3]

IV. ASYMPTOTIC ABELIANNESS I

Given a state $\mathscr{W}$ on the algebra $\mathscr{B}$ satisfying hypotheses H_0-H_3 we have the representations Π of $\mathscr{B}$ and $U(\cdot)$ of the translation group T_4 on the Hilbert space $\mathscr{H}$ according to Proposition 3.3. Here we intend to show that the system $\{(\Pi(\mathscr{B}),\ U(\lambda a));\ \lambda \in \mathbf{R}\}$ is asymptotically Abelian for spacelike $a \in T_4$. This proof is done in several steps.

Lemma 4.1: Let π be a permutation of $\{1,...,n\}$ such that $\pi(i) \leqslant j$ and $\pi(i+1) \geqslant j+1$. Let $a=(0,\mathbf{a})$ be a unit vector in $\mathbf{R}^4$ and $\phi^{\pi}_{\lambda a}(x_1,\ldots,x_n)=\phi(y_{\pi(1)},\ldots,y_{\pi(n)})$, where $y_k=x_k$ for $k=1,...,j$, $y_{\underline{k}}=x_k-\lambda a$ for $k=j+1,...,n$ and, $\phi \in \mathscr{O}_{\sim}(\mathbf{D}^{4n})$. Then for any $F^i \in \mathscr{O}_{\sim}(\mathbf{D}^{4n})'$ with $\operatorname{supp} F^i \subset K_i$, we have, as $\lambda \to \infty$,

$$F^i(\phi^{\pi}_{\lambda a}) \to 0. \tag{4.1}$$

Proof: For simplicity we consider $\pi=1$ (identity) and $i=j$. Let $\|\phi\|_r \leqslant N$. Then $|\phi(z)| \leqslant N e^{-|z|/r}$ for $|\operatorname{Im} z| \leqslant 1/r$. Since

$$|F^j(\phi^1_{\lambda a})| \leqslant M_{m,U} \sup_{U \cap \mathbf{C}^{4n}} |\phi^1_{\lambda a}(z)| e^{|z|/m}$$

holds for any $m \in \mathbf{N}$ and neighborhood U of $\overline{K_j}$ in Q^{4n} such that $m \geqslant r$ and $U \cap \mathbf{C}^{4n} \subset \{z; |\operatorname{Im} z| \leqslant 1/r\}$, it suffices to prove

$$\sup_{U \cap \mathbf{C}^{4n}} e^{-\Sigma_{k=1}^{j}|z_k|/r} e^{-\Sigma_{k=j+1}^{n}|z_k-\lambda a|/r} e^{|z|/m} \to 0 \tag{4.2}$$

as $\lambda\to\infty$ for some $m\in\mathbf{N}$ and neighborhood U of $\overline{K_j}$ in Q^{4n}. Let $V=\{\xi\in\mathbf{R}^4:\xi^2\geqslant 0\}$ and W be a neighborhood of V in Q^4 such that $W\cap\mathbf{C}^4\subset\{\xi\in\mathbf{R}^4;\xi_0^2>(1-\epsilon)\boldsymbol{\xi}^2-\epsilon\}\times i\{\eta\in\mathbf{R}^4;|\eta|<1/m\}$ for some $\epsilon>0$. Then $U=\{(z_1,\ldots,z_{j-1},z,z-\xi,z_{j+2},\ldots,z_n);|\mathrm{Im}\, z|,|\mathrm{Im}\, z_k|\leqslant 1/m,\xi\in W\}$ is a neighborhood of $\overline{K_j}$. Let $B_\lambda=\{z\in\mathbf{C}^4;|\mathrm{Re}\, z|\leqslant\lambda/4,|\mathrm{Im}\, z|\leqslant 1/m\}$ and $U_\lambda=\{(z_1,\ldots,z_{j-1},z,z-\xi,z_{j+2},\ldots,z_n)\in U;z\in B_\lambda\}$. Since $\mathrm{dist}(B_\lambda,W+\lambda a)\geqslant\lambda/4$, it follows

$$\begin{aligned}
&\sup_{U_\lambda\cap\mathbf{C}^{4n}} e^{-\Sigma_{k=1}^{j}|z_k|/r}e^{-\Sigma_{k=j+1}^{n}|z_k-\lambda a|/r}e^{|z|/m}\\
&\leqslant \sup_{U_\lambda\cap\mathbf{C}^{4n}} e^{-\Sigma_{k=1}^{j-1}|z_k|(1/r-1/m)}e^{-|z_j|/r}e^{-|z_{j+1}-\lambda a|/r}\\
&\quad\times e^{|z_j|/m}e^{|z_{j+1}|/m}e^{-\Sigma_{k=j+2}^{n}|z_k|(1/r-1/m)}e^{\Sigma_{k=j+2}^{n}|\lambda a|/m}\\
&\leqslant M \sup_{z\in B_\lambda\cap\mathbf{C}^4,\,\xi\in W\cap\mathbf{C}^4} e^{-|z|(1/r-1/m)}e^{-|z-\xi-\lambda a|(1/r-1/m)}e^{|\lambda a|/m}e^{\Sigma_{k=j+2}^{n}|\lambda a|/m}\\
&\leqslant Me^{-\lambda/4(1/r-1/m)}e^{(n-j)\lambda/m}=Me^{-\lambda(1/4r-(n-j-1/4)/m)}.
\end{aligned}$$

With $B_\lambda^c=\{z\in\mathbf{C}^4;|\mathrm{Re}\, z|\geqslant\lambda/4,|\mathrm{Im}\, z|\leqslant 1/m\}$ we estimate as follows:

$$\begin{aligned}
&\sup_{(U\setminus U_\lambda)\cap\mathbf{C}^{4n}} e^{-\Sigma_{k=1}^{j}|z_k|/r}e^{-\Sigma_{k=j+1}^{n}|z_k-\lambda a|/r}e^{|z|/m}\\
&\leqslant M \sup_{z\in B_\lambda^c\cap\mathbf{C}^4,\,\xi\in W\cap\mathbf{C}^4} e^{-|z|(1/r-1/m)}e^{-|z-\xi-\lambda a|(1/r-1/m)}e^{|\lambda a|/m}e^{\Sigma_{k=j+2}^{n}|\lambda a|/m}\\
&\leqslant Me^{-\lambda/4(1/r-1/m)}e^{(n-j)\lambda/m}=Me^{-\lambda(1/4r-(n-j-1/4)/m)}.
\end{aligned}$$

This yields Eq. (4.2) if we choose m sufficiently large. □

Lemma 4.2: Let $\mathscr{W}\in\mathscr{B}'$ be a state satisfying hypotheses H_0–H_3. Then one has for any spacelike vector a, any $h\in\mathscr{B}_p$, $f\in\mathscr{B}_j$, $g\in\mathscr{B}_k$, $u\in\mathscr{B}_q$, as $\lambda\to\infty$,

$$\mathscr{W}(h\otimes f\otimes g_{\lambda a}\otimes u)-\mathscr{W}(h\otimes g_{\lambda a}\otimes f\otimes u)\to 0.$$

Proof: Let $\phi=h\otimes u\otimes f\otimes g_{\lambda a}\in\mathscr{B}_{p+q+j+k}$. Then $\mathscr{W}(h\otimes f\otimes g_{\lambda a}\otimes u)-\mathscr{W}(h\otimes g_{\lambda a}\otimes f\otimes u)$ is a finite sum of terms of the form $W^i(\phi_{\lambda a}^{\pi})$ for some i and permutations π, where

$$W^i(x_1,\ldots,x_n)=\mathscr{W}_n(x_1,\ldots,x_i,x_{i+1},\ldots,x_n)-\mathscr{W}_n(x_1,\ldots,x_{i+1},x_i,\ldots,x_n),$$

which satisfies the condition for $F^i(\phi_{\lambda a}^{\pi})$ in Lemma 4.1. Now the lemma follows from Lemma 4.1. □

Finally we can conclude that the system $\{(\Pi(\mathscr{B}),U(\lambda a));\lambda\in\mathbf{R}\}$ is asymptotically Abelian.

Proposition 4.3: Suppose $\mathscr{W}\in\mathscr{B}'$ satisfies hypotheses H_0–H_3 and let Π be the representation of $\mathscr{B}$ and U the representation of the Poincaré group G as determined by Proposition 3.3. Then we have, for any $\Phi,\Psi\in\mathscr{D}=\mathscr{D}(\Pi)$, $f,g\in\mathscr{B}$ and any spacelike vector a, as $\lambda\to\infty$,

$$(\Psi,\Pi(f\,)\Pi(g_{\lambda a})\Phi)-(\Psi,\Pi(g_{\lambda a})\Pi(f\,)\Phi)\to 0.$$

Proof: Since f and g are finite sum of elements from $\mathscr{B}_n$'s, the proposition follows from Lemma 4.2. □

V. A PROBLEM OF MONOTONE CONTINUOUS EXTENSION

Consider the representation Π as given by Proposition 4.3. Following Borchers[2] we intend to investigate the subspace $\mathscr{H}_0$ of ''vacuum states'' as introduced in (3.10) by finding an Abelian reduction of the field algebra $\Pi(\mathscr{B})$ to this subspace. However, as it is well known,[3] there is a serious difficulty with this reduction since one cannot expect that the domain $\mathscr{D}$ of the represen-

tation Π is invariant under the projection Q_0 onto the subspace $\mathscr{H}_0$. Accordingly one would like to construct an extension $\hat{\Pi}$ of Π in the Hilbert space $\mathscr{H}$ with a domain $\hat{\mathscr{D}}$ such that $Q_0\hat{\Pi}(\mathscr{B})\times(\hat{\mathscr{D}}\cap\mathscr{H}_0)\subset\hat{\mathscr{D}}$. In standard QFT one takes for $\hat{\Pi}$ the closure $\overline{\Pi}$ of Π. In HFQFT however, there is an additional problem related to the remark at the end of Sec. II.

Let $K\subset L$ be a pair of compact sets in $\mathbf{D}^n$, and suppose each connected component of L intersects K. Then the natural inclusion: $\mathscr{Q}(K)'\to\mathscr{Q}(L)'$ has a dense image (see Lemma 8.6.4 of Ref. 9). Thus the ''standard closure'' (see Sec. VII) may destroy the localization properties and therefore also locality in hyperfunction quantum field theory. But locality is essential in deriving asymptotic Abelianness which in turn has been recognized to be instrumental for the reduction mentioned above. Accordingly, in the definition of the extension $\hat{\Pi}$, we propose to modify the definition of the closure in a way which takes this potential ''loss of locality'' into account. This is done by taking for $\hat{\Pi}$ the ''local closure'' of the representation Π. The definition of the local closure is naturally based on the locality module (see Definition 7.1). Our suggestion for the definition of the local closure relies on the extension of a given state $\mathscr{W}$ on $\mathscr{B}$ to a state on an enlarged *-algebra $\mathscr{C}$ which is defined as the *-algebra generated by the tensor algebra $\mathscr{B}$ and the ''locality module'' E_ℓ. The existence of this monotone continuous extension is shown in this section.

Let $F_{\ell,n}$ be the linear space generated by the elements f of the form

$$f=c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_k}(g), \tag{5.1}$$

where $g\in\mathscr{Q}(\cap_{j=1}^k\overline{K^n_{i_j}})$ and $i_j-i_{j-1}\geqslant 2$.

Proposition 5.1: Let $\mathscr{W}\in\mathscr{B}'$ have a continuous linear extension to E_ℓ. Then the mapping

$$\mathscr{W}\circ c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_k}:\mathscr{B}_n=\mathscr{Q}(\mathbf{D}^{4n})\to\mathbf{C}$$

is continuously extended to $\mathscr{Q}(\cap_{j=1}^k\overline{K^n_{i_j}})$. (We say for brevity $\mathscr{W}\in\mathscr{B}'$ has a continuous linear extension to $F_\ell=\oplus_{n=0}^\infty F_{n,\ell}$.)

Proof: Let $\mathscr{F}_n=\mathscr{F}_n(x_1,\ldots,x_n)$ be the Fourier hyperfunction defined by

$$\mathscr{Q}(\mathbf{D}^{4n})\ni g\to\mathscr{W}(c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_k}(g)).$$

Let $g\in\mathscr{Q}(\mathbf{D}^{4n})$. Then $f=c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ\check{c}^n_{i_j}\circ\cdots\circ c^n_{i_k}(g)\in\mathscr{Q}(\mathbf{D}^{4n})$. Note that $c^n_l\circ c^n_j=c^n_l\circ c^n_j$ if $|i-j|\geqslant 2$.

$$\mathscr{F}_n(g)=\mathscr{W}(c^n_{i_j}\circ c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_j}\circ\cdots\circ c^n_{i_k}(g))=\mathscr{W}(c^n_{i_j}(f\ ))$$

can be extended to $g\in\mathscr{Q}(\overline{K^n_{i_j}})$, and supp $\mathscr{F}_n\subset\overline{K^n_{i_j}}$. Therefore supp $\mathscr{F}_n\subset\cap_{j=1}^k\overline{K^n_{i_j}}$ and $\mathscr{F}_n$ can be extended to $\mathscr{Q}(\cap_{j=1}^k\overline{K^n_{i_j}})$. □

Corollary 5.2: Let $\mathscr{C}$ be the *-algebra generated by $\mathscr{B}$ and E_ℓ and let $\mathscr{W}\in\mathscr{B}'$ have a continuous linear extension to E_ℓ. Then $\mathscr{W}$ extends continuously to $\mathscr{C}$.

Proof: Let $\mathscr{C}=\oplus_{n=0}^\infty\mathscr{C}_n$. Then it follows from Proposition 3.1 that $\mathscr{C}_n\subset\mathscr{B}_n+F_{\ell,n}=\{f+g;f\in\mathscr{B}_n,g\in F_{\ell,n}\}$. Thus $\mathscr{W}$ extends continuously to $\mathscr{C}$ by Proposition 5.1. □

Lemma 5.3: (a) Let K be a compact set in $\mathbf{D}^n$, and $f\in\mathscr{Q}(K)$. Then there exist $m\in\mathbf{N}$ and a sequence $\{f_j\}\subset\mathscr{Q}(\mathbf{D}^n)$ such that $\|f_j-f\|_m\to 0$ as $j\to\infty$, where $\|\cdot\|_m$ is a norm of $\mathscr{O}^m_c(U_m)$ defined by (2.1).

(b) Let $\|g_j-g\|_m\to 0$ as $j\to\infty$ for $g\in\mathscr{Q}(L)$ with $L\subset\mathbf{D}^k$ and $g_j\in\mathscr{Q}(\mathbf{D}^k)$. Then $f_j\otimes g_j\to f\otimes g$ in $\mathscr{Q}(K\times L)$ as $j\to\infty$.

Proof: (a) follows from Proposition 2.2. For (b), we define a Banach space $\mathscr{O}^m_c(U_m\times V_m)$ with a norm

$$\|f\|_m=\sup_{z\in(U_m\times V_m)\cap\mathbf{C}^{n+k}}|f(z)|e^{|z|/m}.$$

Now we conclude by

$$\|f_j\otimes g_j - f\otimes g\|_m \leqslant \|f_j - f\|_m \|g_j\|_m + \|f\|_m \|g_j - g\|_m \to 0$$

as $j\to\infty$. □

Proposition 5.4: Let $\mathscr{W}$ be a state on $\mathscr{B}$ which satisfies condition H_3. Then $\mathscr{W}$ extends continuously to a state on $\mathscr{C}$.

Proof: Since $\mathscr{W}$ satisfies H_3 it has a continuous linear extension to E_ℓ. Proposition 5.1 implies that $\mathscr{W}$ can be extended continuously to F_ℓ. Now let $h+f\in\mathscr{C}$ with $h\in F_\ell$ and $f\in\mathscr{B}$ be given. Then it follows from Lemma 5.3 that there exists a sequence $\{h_m\}\subset\mathscr{B}$ such that $\mathscr{W}(h_m)\to\mathscr{W}(h)$ as $m\to\infty$. Since $\mathscr{W}((h_m+f)^*\otimes(h_m+f))\geqslant 0$, it suffices to prove $\mathscr{W}((h_m+f)^*\otimes(h_m+f))\to\mathscr{W}((h+f)^*\otimes(h+f))$ as $m\to\infty$. Since an element of $F_{\ell,k}$ is a sum of the elements of the form

$$h^{[k]} = c^k_{i_2}\circ c^k_{i_2}\circ\cdots\circ c^k_{i_s}(g^{[k]})$$

with $g^{[k]}\in\mathscr{C}^m_c(U_m)\subset\mathscr{Q}(\cap_{j=1}^s\overline{K^k_{i_j}})$ and $h^{[k]}$ is approximated by

$$h^{[k]}_m = c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_n}(g^{[k]}_m)$$

with $g^{[k]}_m\in\mathscr{Q}(\mathbf{D}^{4k})$, it suffices to show that for $m\to\infty$

$$\mathscr{W}((h^{[k]}_m)^*\otimes h^{[n-k]}_m)\to\mathscr{W}((h^{[k]})^*\otimes h^{[n-k]}), \tag{5.2}$$

$$\mathscr{W}((h^{[k]}_m)^*\otimes f^{[n-k]})\to\mathscr{W}((h^{[k]})^*\otimes f^{[n-k]}). \tag{5.3}$$

$$\mathscr{W}((f^{[k]})^*\otimes h^{[n-k]}_m)\to\mathscr{W}((f^{[k]})^*\otimes h^{[n-k]}). \tag{5.4}$$

Note that

$$\begin{aligned}(h^{[k]})^*\otimes h^{[n-k]} &= c^k_{i_1}\circ c^k_{i_2}\circ\cdots\circ c^k_{i_s}(g^{[k]*})\otimes c^{n-k}_{j_1}\circ c^{n-k}_{j_2}\circ\cdots\circ c^{n-k}_{j_t}(g^{[n-k]})\\ &= c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_s}\circ c^n_{i_{s+1}}\circ c^n_{i_{s+2}}\circ\cdots\circ c^n_{i_{s+t}}(g^{[k]*}\otimes g^{[n-k]}).\end{aligned}$$

where $i_{s+l}=k+j_l$, $g^{[k]}\in\mathscr{Q}(\cap_{j=1}^s\overline{K^k_{i_j}})$ and $g^{[n-k]}\in\mathscr{Q}(\cap_{i=1}^t\overline{K^{n-k}_{j_i}})$. Since

$$\mathscr{W}\circ c^n_{i_1}\circ c^n_{i_2}\circ\cdots\circ c^n_{i_s}\circ c^n_{i_{s+1}}\circ c^n_{i_{s+2}}\circ\cdots\circ c^n_{i_{s+t}}:\ \mathscr{Q}(\mathbf{D}^{4n})\to\mathbf{C}$$

is extended continuously to $\mathscr{Q}(\cap_{j=1}^{s+t}\overline{K^n_{i_j}})\supset\mathscr{Q}(\cap_{j=1}^s\overline{K^k_{i_j}})\otimes\mathscr{Q}(\cap_{i=1}^t\overline{K^{n-k}_{j_i}})$ by Proposition 5.1, Equation (5.2) follows from

$$g^{[k]}_m\otimes g^{[n-k]}_m\to g^{[k]}\otimes g^{[n-k]},$$

which in turn follows from Lemma 5.3. The relations (5.3) and (5.4) are easy to prove. □

VI. ASYMPTOTIC ABELIANNESS II

Let $\mathscr{W}$ be a state on $\mathscr{B}$ which satisfies hypothesis H_3. By Proposition 5.4 it has a monotone continuous linear extension to a state on the *-algebra $\mathscr{C}$ introduced in Corollary 5.2. After the corresponding preparations we are going to show that the GNS-representation of this state on $\mathscr{C}$ also is asymptotically Abelian.

Lemma 6.1: Let π be a permutation of $\{1,\ldots,n\}$ such that $\pi(i)\leqslant j$ and $\pi(i+1)\geqslant j+1$. Let $a=(0,\mathbf{a})$ be a unit vector in $\mathbf{R}^4$ and $\phi^\pi_{\lambda a}(x_1,\ldots,x_n)=\phi(y_{\pi(1)},\ldots,y_{\pi(n)})$, where $y_k=x_k$ for $k=1,\ldots,j$, $y_k=x_k-\lambda a$ for $k=j+1,\ldots,n$. Let $L_k=\underline{K_k}$ or $L_k=\mathbf{D}^{4n}$, and $L=\cap_{1\leqslant k\leqslant n-1}L_k$. Let $\phi\in\mathscr{Q}(L)$ and $F^i\in\mathscr{Q}(\mathbf{D}^{4n})'$ such that supp $F^i\subset L\cap K_i$. Then in the limit $\lambda\to\infty$ one has

$$F^i(\phi^\pi_{\lambda a})\to 0. \tag{6.1}$$

Proof: For simplicity assume $\pi=1$ (identity) and $i=j$. Since $\phi\in\mathscr{O}(L)$, there are positive numbers r, N and a neighborhood O of L in Q^{4n} such that $|\phi(z)|\leqslant Ne^{-|z|/r}$ for $z\in O\cap\mathbf{C}^{4n}$. Since

$$|F^j(\phi^1_{\lambda a})|\leqslant M_{m,V}\sup_{V\cap\mathbf{C}^{4n}}|\phi^1_{\lambda a}(z)|e^{|z|/m}$$

holds for any $m\in\mathbf{N}$ and neighborhood V of $L\cap\overline{K_j}$ in Q^{4n} such that $m\geqslant r$ and $V\subset O$, it suffices to prove

$$\sup_{V\cap\mathbf{C}^{4n}}\exp(-\Sigma_{k=1}^{j}|z_k|/r)\exp(-\Sigma_{k=j+1}^{n}|z_k-\lambda a|/r)\exp(|z|/m)\to 0 \tag{6.2}$$

as $\lambda\to\infty$ for some $m\in\mathbf{N}$ and neighborhood V of $L\cap\overline{K_j}$ in Q^{4n}. The difference between formulas (4.2) and (6.2) is that of U and V. U is a neighborhood of K_j and V is a neighborhood of $L\cap K_j$. Since we can choose V so that $V\subset U$, the proof of Lemma 4.1 applies. □

Now, as an extension of Proposition 4.3 asymptotic Abelianness follows for the GNS-representation Π of $\mathscr{C}$.

Proposition 6.2: As in Proposition 4.7 let Π be the *-representation of $\mathscr{C}$ associated with the state $\mathscr{W}$ which on $\mathscr{B}$ satisfies the hypotheses H_0–H_3. For any Φ, $\Psi\in\mathscr{D}=\mathscr{D}(\Pi)$, $f,g\in\mathscr{C}$ and any spacelike vector a, we have

$$(\Psi,\Pi(f)\Pi(g_{\lambda a})\Phi)-(\Psi,\Pi(g_{\lambda a})\Pi(f)\Phi)\to 0$$

as $\lambda\to\infty$ (asymptotically abelian).

Proof: The proposition follows from Lemma 6.1 by using the same argument as in Lemma 4.2. □

VII. THE LOCAL CLOSURE AND ASYMPTOTIC ABELIANNESS III

We prepare the definition of the local closure by reminding the reader of the (standard) closure of a representation of a *-algebra.

Let $\mathscr{D}$ be a dense subspace of a Hilbert space $\mathscr{H}$. Denote by $\mathscr{L}^\dagger(\mathscr{D},\mathscr{H})$ the set of all linear operators A such that $\mathscr{D}(A)=\mathscr{D}$ and $\mathscr{D}(A^*)\supset\mathscr{D}$, and put $\mathscr{L}^\dagger(\mathscr{D})=\{A\in\mathscr{L}^\dagger(\mathscr{D},\mathscr{H});A\mathscr{D}\subset\mathscr{D}$ and $A^*\mathscr{D}\subset\mathscr{D}\}$. $\mathscr{L}^\dagger(\mathscr{D})$ is a *-algebra with the involution $A\to A^\dagger=A^*_{|\mathscr{D}}$. A *-subalgebra $\mathscr{P}$ of $\mathscr{L}^\dagger(\mathscr{D})$ with identity operator is called an *O*-algebra*.

A *-representation $\pi:\mathscr{A}\to\mathscr{L}^\dagger(\mathscr{D})$ of a *-algebra $\mathscr{A}$ is called *closed* iff its domain is closed in the graph topology (see Definition 3.1.4 of Ref. 3 and Definition 2.4 of Ref. 10). We begin with the following lemma (Lemma 2.6 in Ref. 10).

Lemma: Suppose π is a *-representation of the a *-algebra $\mathscr{A}$ on a Hilbert space $\mathscr{H}$. Then, there is a unique minimal closed extension $\boldsymbol{\pi}$ of π. Furthermore, the domain of $\boldsymbol{\pi}$ is given by

$$\mathscr{D}(\boldsymbol{\pi})=\cap_{A\in\mathscr{A}}\operatorname{dom}\overline{\pi(A)}.$$

Corollary: Under the assumptions of the above lemma the following holds:

(a) $\boldsymbol{\pi}(B)\mathscr{D}(\boldsymbol{\pi})\subset\mathscr{D}(\boldsymbol{\pi})$ for all $B\in\mathscr{A}$ and $\boldsymbol{\pi}(AB)f=\boldsymbol{\pi}(A)\boldsymbol{\pi}(B)f$ for all A, $B\in\mathscr{A}$ and all $f\in\mathscr{D}(\pi)$.

(b) If $U(a,\Lambda)$ is a unitary representation of the Poincaré group such that

$$U(a,\Lambda)\mathscr{D}(\pi)\subset\mathscr{D}(\pi)\quad\text{and}\quad U(a,\Lambda)\pi(A)U(a,\Lambda)^{-1}=\pi(\alpha_{(a,\Lambda)}(A)),$$

then

$$U(a,\Lambda)\mathscr{D}(\boldsymbol{\pi})\subset\mathscr{D}(\boldsymbol{\pi})\quad\text{and}\quad U(a,\Lambda)\boldsymbol{\pi}(A)U(a,\Lambda)^{-1}=\boldsymbol{\pi}(\alpha_{(a,\Lambda)}(A)).$$

Proof: In fact, let $f\in\mathscr{D}(\boldsymbol{\pi})$. Then there exists a sequence $\{f_n\in\mathscr{D}(\pi);\ n=1,2,...\}$ such that $f_n\to f$ and $\pi(C)f_n\to\boldsymbol{\pi}(C)f$ for all $C\in\mathscr{A}$. Let $g=U(a,\Lambda)f$ and $g_n=U(a,\Lambda)^{-1}f_n$. Then $g_n\to g$ and $\pi(A)g_n=\pi(A)U(a,\Lambda)^{-1}f_n=U(a,\Lambda)^{-1}\pi(\alpha_{(a,\Lambda)}(A))f_n\to U(a,\Lambda)^{-1}\boldsymbol{\pi}(\alpha_{(a,\Lambda)}(A))f$. This shows that $g\in\mathscr{D}(\pi(A))$. Since $A\in\mathscr{A}$ is arbitrary we have

$$U(a,\Lambda)\mathscr{D}(\boldsymbol{\pi})\subset\mathscr{D}(\boldsymbol{\pi}) \quad \text{and} \quad U(a,\Lambda)\boldsymbol{\pi}(A)U(a,\Lambda)^{-1}=\boldsymbol{\pi}(\alpha_{(a,\Lambda)}(A)).$$

□

Definition 7.1: Given a state $\mathscr{W}\in\mathscr{B}'$ which satisfies hypothesis H_3 denote by Π the GNS-representation of its monotone extension to the *-algebra $\mathscr{C}$. Introduce the subspace of the representation space

$$\hat{\mathscr{D}}=\cap_{f\in\mathscr{C}}\,\mathrm{dom}\,\overline{\Pi(f)},\quad \hat{\Pi}(f)=\overline{\Pi(f)}_{|\hat{\mathscr{D}}}.$$

Then the *local closure* $\hat{\Pi}$ (of the field induced by the state $\mathscr{W}$) is defined as the mapping

$$\hat{\Pi}: f\in\mathscr{B}\to\hat{\Pi}(f)\in\mathscr{L}^{\dagger}(\hat{\mathscr{D}}).$$

As a counterpart of Proposition 6.2 we get that the local closure preserves asymptotic Abelianness.

Proposition 7.2: Let $\mathscr{W}$ be a state on $\mathscr{B}$ satisfying hypothesis H_3 and let Π be the associated GNS-representation of its extension to a state on $\mathscr{C}$. Then the local closure $\hat{\Pi}$ satisfies the following: For any $\Phi,\ \Psi\in\hat{\mathscr{D}},\ f,g\in\mathscr{C}$, and any spacelike vector a, one has, as $\lambda\to\infty$,

$$(\Psi,\hat{\Pi}(f)\hat{\Pi}(g_{\lambda a})\Phi)-(\Psi,\hat{\Pi}(g_{\lambda a})\hat{\Pi}(f)\Phi)\to 0.$$

Proof: Given $f\in\mathscr{B}_i$, $g\in\mathscr{B}_j$ and $\Phi,\ \Psi\in\hat{\mathscr{D}}$ there are sequences Φ_n, Ψ_n in $\mathscr{D}(\Pi)$ such that $\Phi_n\to\Phi$, $\Psi_n\to\Psi$, $\Pi(f^*)\Psi_n\to\hat{\Pi}(f^*)\Psi$, and $\Pi(g)\Phi_n\to\hat{\Pi}(g)\Phi$ as $n\to\infty$. The functionals

$$\tilde{\mathscr{O}}(\mathbf{D}^{4i})\times\tilde{\mathscr{O}}(\mathbf{D}^{4j})\ni(f,g)\to(\Psi_n,\Pi(f)\Pi(g)\Phi_n)\in\mathbf{C}$$

define Fourier hyperfunctions $F_n\in\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})'$, and it follows that $F_n(f\otimes g)\to(\Psi,\hat{\Pi}(f)\hat{\Pi}(g)\Phi)$ as $n\to\infty$. Since $\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})$ is a barreled space, the sequence $\{F_n\}$ is equicontinuous and converges to a Fourier hyperfunction F, uniformly on every compact subset of $\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})$ (See the Corollary of Theorem 33.1 of Ref. 11). Thus the functional

$$\tilde{\mathscr{O}}(\mathbf{D}^{4i})\times\tilde{\mathscr{O}}(\mathbf{D}^{4j})\ni(f,g)\to(\Psi,\hat{\Pi}(f)\hat{\Pi}(g)\Phi)\in\mathbf{C}$$

defines a Fourier hyperfunction $F\in\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})'$ such that

$$F(f\otimes g)=(\Psi,\hat{\Pi}(f)\hat{\Pi}(g)\Phi).$$

From the definition of the local closure ($\hat{\mathscr{D}}=\cap_{f\in\mathscr{C}}\,\mathrm{dom}\,\overline{\Pi(f)}$) it also follows

$$F_n(h)=(\Psi_n,\Pi(h)\Phi_n)\to(\Psi,\hat{\Pi}(h)\Phi)=F(h),$$

for all $h\in E_{\swarrow,i+j}$. This shows that not only F_n but also F satisfies the condition of local commutativity, i.e., the support of

$$F(x_1,\ldots,x_k,x_{k+1},\ldots,x_{i+j})-F(x_1,\ldots,x_{k+1},x_k,\ldots,x_{i+j})$$

is contained in $\overline{K_k^{i+j}}$. Since the asymptotic Abelianness follows from locality we conclude as in Propositions 4.3 respectively 6.2. □

Remark 7.3: Let $F_n\in\tilde{\mathscr{O}}(K_k^{i+j})'$, $n=1,2,\ldots$ be given. The limit $\lim_{n\to\infty}F_n(f)$ for all $f\in\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})$, if it exists, defines an element $F\in\tilde{\mathscr{O}}(\mathbf{D}^{4(i+j)})'$. However, the limit $\lim_{n\to\infty}F_n(f)$ for all $f\in\tilde{\mathscr{O}}(K_k^{i+j})$, if it exists, defines an element $F\in\tilde{\mathscr{O}}(K_k^{i+j})'$. This is the technical reason why we have defined $\hat{\mathscr{D}}$ by $\cap_{f\in\mathscr{C}}\,\mathrm{dom}\,\overline{\Pi(f)}$, not by $\cap_{f\in\mathscr{B}}\,\mathrm{dom}\,\overline{\Pi(f)}$.

Proposition 7.4: Suppose $\mathscr{W}\in\mathscr{B}'$ satisfies hypotheses H_0-H_3 and let Π be the representation of $\mathscr{B}$ and U the representation of the Poincaré group G as determined by Proposition 3.3. Denote by Q_0 the orthogonal projection operator onto the subspace $\mathscr{H}_0$ of vacuum states [Eq. (3.10)] and let $\hat{\Pi}$ be the local closure as given in Definition 5.2. Then the following holds:

(a) $Q_0\hat{\Pi}(\mathscr{C})\hat{\mathscr{D}}_0\subset\hat{\mathscr{D}}_0$ where $\hat{\mathscr{D}}_0=\hat{\mathscr{D}}\cap\mathscr{H}_0$;

(b) $\hat{\mathscr{D}}_0$ is dense in $\mathscr{H}_0$.

Proof: Given any $\Psi \in \hat{\mathscr{D}}_0$, any $f,g \in \mathscr{C}$ and any spacelike $a \in \mathbf{R}^4$, Proposition 7.2 implies

$$\|\hat{\Pi}(g)U(\lambda a)\hat{\Pi}(f\ )\Psi\|^2-(\Psi,\hat{\Pi}(g^*_{-\lambda a}g_{-\lambda a})\hat{\Pi}(f^*f\ )\Psi)$$
$$=(\Psi,\hat{\Pi}(f^*)\hat{\Pi}(g^*_{-\lambda a}g_{-\lambda a})\hat{\Pi}(f\ )\Psi)-(\Psi,\hat{\Pi}(g^*_{-\lambda a}g_{-\lambda a})\hat{\Pi}(f^*f\ )\Psi)\to 0$$

as $\lambda\to\infty$ and therefore, since for all λ,

$$|(\Psi,\hat{\Pi}(g^*_{-\lambda a}g_{-\lambda a})\hat{\Pi}(f^*f\ )\Psi)|\leq\|\hat{\Pi}(g^*g)\Psi\|\|\hat{\Pi}(f^*f\ )\Psi\|,$$

we deduce that $\{\|\hat{\Pi}(g)U(\lambda a)\hat{\Pi}(f\ )\Psi\|;\ \lambda\in\mathbf{R}\}$ is bounded. Let $\Phi\in\mathscr{D}(\hat{\Pi}(g)^*)$. Then we have

$$(\Phi,\hat{\Pi}(g)U(\lambda a)\hat{\Pi}(f\ )\Psi)=(\hat{\Pi}(g)^*\Phi,U(\lambda a)\hat{\Pi}(f\ )\Psi)\to(\hat{\Pi}(g)^*\Phi,Q_0\hat{\Pi}(f\ )\Psi)$$

as $\lambda\to\infty$ (Theorem 8 of Ref. 12). Since $\mathscr{D}(\hat{\Pi}(g)^*)$ is dense and $\{\|\hat{\Pi}(g)U(\lambda a)\hat{\Pi}(f\ )\Psi\|;\ \lambda\in\mathbf{R}\}$ is bounded, $\hat{\Pi}(g)U(\lambda a)\hat{\Pi}(f\ )\Psi$ converges weakly for $\lambda\to\infty$. Since $U(\lambda a)\hat{\Pi}(f\ )\Psi\to Q_0\hat{\Pi}(f\ )\Psi$ weakly as $\lambda\to\infty$, we conclude that $Q_0\hat{\Pi}(f\ )\Psi\in\operatorname{dom}\overline{\hat{\Pi}(g)}=\operatorname{dom}\overline{\Pi(g)}$. Since g is arbitrary this implies $Q_0\hat{\Pi}(f\ )\Psi\in\hat{\mathscr{D}}$ and thus $Q_0\hat{\Pi}(\mathscr{C})\hat{\mathscr{D}}_0\subset\hat{\mathscr{D}}_0$. Since $\hat{\Pi}(\mathscr{C})\Omega\subset\hat{\mathscr{D}}$ is dense in $\mathscr{H}$, $Q_0\hat{\Pi}(\mathscr{C})\Omega\subset Q_0\hat{\Pi}(\mathscr{C})\hat{\mathscr{D}}_0$ is dense in $\mathscr{H}_0$. Thus $\hat{\mathscr{D}}_0$ is dense in $\mathscr{H}_0$. □

Remark 7.5: Since $\mathscr{C}_{\sim}(\mathbf{D}^{4n})$ is dense in $\mathscr{C}_{\sim}(K_i^n)$ we actually know that $\hat{\Pi}(\mathscr{B})\Omega=\Pi(\mathscr{B})\Omega$ is dense in $\mathscr{H}$.

The above proposition enables us to define a reduction of the field algebra $\Pi(\mathscr{B})$ (notation of Proposition 7.4) to the subspace of vacuum states $\mathscr{H}_0$. By part (a) we know that the domain $\hat{\mathscr{D}}_0$ is left invariant under all operators $Q_0\hat{\Pi}(f\ )Q_0$, $f\in\mathscr{C}$. Hence products of these operators are well defined, and it follows easily that

$$\mathscr{P}_0=\vee\{Q_0\hat{\Pi}(f\ )Q_{0|\hat{\mathscr{D}}_0};f\in\mathscr{C}\} \tag{7.1}$$

is an algebra in $\mathscr{H}_0$ with the domain $\hat{\mathscr{D}}_0$ which is dense in the vacuum subspace $\mathscr{H}_0$. The following proposition asserts that this algebra is actually Abelian and admits a natural involution.

Proposition 7.6: Under the assumptions of Proposition 7.4 the set of operators $\mathscr{P}_0$ is an Abelian O^*-algebra in $\mathscr{H}_0$ with dense domain $\hat{\mathscr{D}}_0$.

Proof: It is also an O^*-algebra since $\operatorname{dom}(Q_0\hat{\Pi}(f\ )Q_0)^*=\operatorname{dom}Q_0\hat{\Pi}(f\ )^*Q_0\supset\operatorname{dom}Q_0\hat{\Pi}(f^*)Q_0\supset\hat{\mathscr{D}}_0$, hence $(Q_0\hat{\Pi}Q_0)^*_{|\hat{\mathscr{D}}_0}\in\mathscr{P}_0$. For the Abelianness of this algebra it suffices to show

$$(\Psi,\hat{\Pi}(f\ )Q_0\hat{\Pi}(g)\Phi)=(\Psi,\hat{\Pi}(g)Q_0\hat{\Pi}(f\ )\Phi)$$

for all Φ, $\Psi\in\mathscr{D}_0$ and $f,g\in\mathscr{C}$. And this follows easily from the relations

$$(\Psi,\hat{\Pi}(f\ )U(\lambda a)\hat{\Pi}(g)\Phi)\to(\Psi,\hat{\Pi}(f\ )Q_0\hat{\Pi}(g)\Phi).$$

$$(\Psi,\hat{\Pi}(g)U(-\lambda a)\hat{\Pi}(f\ )\Phi)\to(\Psi,\hat{\Pi}(g)Q_0\hat{\Pi}(f\ )\Phi),$$

$$(\Psi,\hat{\Pi}(f\ )U(\lambda a)\hat{\Pi}(g)\Phi)-(\Psi,\hat{\Pi}(g)U(-\lambda a)\hat{\Pi}(f\ )\Phi)\to 0$$

as $\lambda\to\infty$. □

VIII. VACUUM STRUCTURE

Suppose a HFQFT is given in terms of a state $\mathscr{W}$ on the test function algebra $\mathscr{B}$ which satisfies hypotheses H_0–H_3. According to Proposition 3.3 we have a field O^*-algebra $\Pi(\mathscr{B})$ with dense domain $\mathscr{D}=\Pi(\mathscr{B})\Omega$ in a complex Hilbert space $\mathscr{H}$, a unitary representation U of the translation group in $\mathscr{H}$ and a strongly cyclic unit vector $\Omega\in\mathscr{D}$ which is invariant under all translations. In Sec. VII, by taking the local closure, we obtained an extension $\mathscr{P}=\hat{\Pi}(\mathscr{B})$ of the field O^*-algebra $\Pi(\mathscr{B})$ with dense domain $\hat{\mathscr{D}}$ in the Hilbert space $\mathscr{H}$ which allows a reduction $\mathscr{P}_0$

[Eq. (7.1)] to the subspace $\mathscr{H}_0$ of translation invariant states since Proposition 7.4 shows $Q_0\hat{\Pi}(\mathscr{B})\hat{\mathscr{D}}_0\subset\hat{\mathscr{D}}_0=\hat{\mathscr{D}}\cap\mathscr{H}_0$ where Q_0 is the orthogonal projector onto $\mathscr{H}_0$. From here on the discussion of ''Borchers's anomaly'' is basically the same as in standard QFT. We indicate the basic steps. First one has to decide on how to express the irreducibility of the field, either in terms of the weak or the strong commutant of the field algebra to be trivial. A natural choice is the weak commutant (see Ref. 3)

$$\mathscr{P}'_w=\{X\in\mathscr{B}(\mathscr{H});\forall u,v\in\hat{\mathscr{D}},\forall A\in\mathscr{P},(X^*u,Av)=(A^*u,Xv)\}. \tag{8.1}$$

The commutant of the weak commutant of the field algebra is the von Neumann algebra $\mathscr{R}$ associated with $\mathscr{P}$:

$$\mathscr{R}=(\mathscr{P}'_w)'.$$

This algebra contains the projector Q_0 as the following lemma shows.

Lemma 8.1: Under the assumptions of Proposition 7.4 we have

(a) $Q_0\in\mathscr{R}$.

(b) $Q_0\mathscr{P}'_wQ_{0|\mathscr{H}_0}=\mathscr{P}'_{w|\mathscr{H}_0}\subset(\mathscr{P}_0)'_w$.

(c) The mapping

$$\mathscr{P}'_w\ni S\to\phi(S)=Q_0SQ_0\in\mathscr{P}'_{w|\mathscr{H}_0} \tag{8.2}$$

is an isomorphism.

Proof: For $S\in\mathscr{P}'_w$ denote $S(x)=U(x)SU(-x)$. Then we have

$$(\hat{\Pi}(f)^*\Omega,U(-x)S\Omega)=(\hat{\Pi}(f_x)^*\Omega,S\Omega)=(\Omega,S\hat{\Pi}(f_x)\Omega)=(\Omega,SU(x)\hat{\Pi}(f)\Omega).$$

The Fourier transform of this equation is

$$(\hat{\Pi}(f)^*\Omega,E(-d^4p)S\Omega)=(\Omega,SE(d^4p)\hat{\Pi}(f)\Omega).$$

Hence, by the spectral condition, its support must be contained in $\{0\}$. Thus, $(\hat{\Pi}(f)^*\Omega,U(-x)S\Omega)$ is constant and equal to $(\hat{\Pi}(f)^*\Omega,S\Omega)$ $(x=0)$. Since $\hat{\Pi}(\mathscr{B})\Omega$ is dense in $\mathscr{H}$, this implies $S(-x)\Omega=S\Omega$. Since

$$\begin{aligned}(\hat{\Pi}(f)\Omega,S(-x)\hat{\Pi}(g)\Omega)&=(\hat{\Pi}(f_x)\Omega,S\hat{\Pi}(g_x)\Omega)=(\hat{\Pi}(g_x)^*\hat{\Pi}(f_x)\Omega,S\Omega)\\&=(\hat{\Pi}(g_x)^*\hat{\Pi}(f_x)\Omega,S(x)\Omega)=(\hat{\Pi}(g)^*\hat{\Pi}(f)\Omega,S\Omega)\\&=(\hat{\Pi}(f)\Omega,S\hat{\Pi}(g)\Omega)\end{aligned}$$

and $\hat{\Pi}(\mathscr{B})\Omega$ is dense in $\mathscr{H}$, it follows $S(-x)=S$, i.e., $[U(x),S]=0$. Consequently $U(x)\in\mathscr{R}$ and therefore $Q_0\in\mathscr{R}$.

For the proof of part (b) observe that for any $T\in\mathscr{P}'_w$, $\Phi,\Psi\in\hat{\mathscr{D}}_0$, and $B\in\Pi(\mathscr{C})$ one has, denoting the local closure of B by $\overline{B}$ and observing $Q_0\in(\mathscr{P}'_w)'$,

$$(\Psi,T\overline{B}\Phi)=(\Psi,T_{|\mathscr{H}_0}Q_0\overline{B}Q_0\Phi)$$

$$(\Psi,T\overline{B}\Phi)=(\overline{B}^*\Psi,T\Phi)=((Q_0\overline{B}Q_0)^*\Psi,T_{|\mathscr{H}_0}\Phi).$$

This implies (b).

Finally consider $S\in\mathscr{P}'_w$ such that $Q_0S=0$ on $\mathscr{H}_0$, then $S\Omega=SQ_0\Omega=Q_0S\Omega=0$ and therefore, for all $f,g\in\mathscr{B}$,

$$(\hat{\Pi}(f)\Omega,S\hat{\Pi}(g)\Omega)=(\hat{\Pi}(g)^*\hat{\Pi}(f)\Omega,S\Omega)=0.$$

Since $\hat{\Pi}(\mathscr{B})\Omega$ is dense in the Hilbert space $\mathscr{H}$ this proves $S=0$. Thus we conclude. □

Let $\mathscr{P}'_s=\{X\in\mathscr{B}(\mathscr{H});\forall A\in\mathscr{P},AX\supset XA\}$ be the strong commutant of $\mathscr{P}$ and let d_0 be the dimension of $\mathscr{H}_0$. Then the following proposition is well known (see Proposition 3.1.12 of Ref. 3) in standard quantum field theory.

Proposition 8.2: Let a Wightman theory be given by a O^*-algebra $\mathscr{P}$ with strongly cyclic vector Ω such that $\mathscr{P}'_s=\mathscr{P}'_w$. Then the condition

$$(1)\quad \mathscr{P}'_w=\mathbf{C}(\mathscr{H})$$

implies the condition

$$(2)\quad d_0=1 \quad \text{or} \quad d_0=\infty.$$

The above proposition asserts that an irreducible standard quantum field theory may have not only a one-dimensional but also an infinite-dimensional vacuum subspace. This phenomenon is called sometimes Borchers's anomaly.

According to Proposition 7.6 the reduction of the field algebra $\mathscr{P}$ of HFQFT to the vacuum subspace $\mathscr{H}_0$, i.e., the set $\mathscr{P}_0$ introduced in (7.1), is an Abelian O^*-algebra. If this reduced algebra $\mathscr{P}_0$ is irreducible we know $(\mathscr{P}_0)'_w=\mathbf{C}(\mathscr{H}_0)$. Now Lemma 7.1 implies first $\mathscr{P}'_{w|\mathscr{H}_0}=\mathbf{C}(\mathscr{H}_0)$ and then $\mathscr{P}'_w=\mathbf{C}(\mathscr{H}_0)$, i.e., the field algebra $\mathscr{P}$ itself is irreducible. Recall that an Abelian O^*-algebra can be irreducible not only in a one-dimensional but also in an infinite-dimensional space, whereas an Abelian von Neumann algebra can be irreducible only in a one-dimensional space (see Ref. 10). Thus we deduce that Borchers's anomaly can occur in HFQFT.

ACKNOWLEDGMENTS

Most of the work for this paper was done during S.N.'s visit to the Department of Mathematics and Applied Mathematics, University of Durban-Westville. We are very grateful to the Foundation for Research Development (International Liaison and Open Research Programmes Departments) which promoted this collaboration through their generous funding. Finally we thank the referee of this paper for a number of very helpful suggestions and hints.

[1] N. N. Bogolubov, A. A. Logunov, A. T. Oksak, and I. T. Todorov, *General Principles of Quantum Field Theory* (Kluwer, Dordrecht, 1990).
[2] H. J. Borchers, Commun. Math. Phys. **1**, 49–56 (1965).
[3] S. S. Horuzhy, *Introduction to Algebraic Quantum Field Theory* (Kluwer, Dordrecht, 1990).
[4] S. Nagamachi and N. Mugibayashi, Commun. Math. Phys. **46**, 119–134 (1976).
[5] E. Brüning and S. Nagamachi, J. Math. Phys. **30**, 2340–2359 (1989).
[6] S. Nagamachi and N. Mugibayashi, J. Math. Phys. **27**, 832–839 (1986).
[7] A. S. Wightman, *Mathematical Analysis and Applications*, edited by L. Nachbin (Academic, New York, 1981), pp. 769–791.
[8] S. Nagamachi and T. Nishimura, Math. Japonica **35**, 293–313 (1990).
[9] A. Kaneko, *Introduction to Hyperfunctions* (Kluwer, Dordrecht, 1988).
[10] R. T. Powers, Commun. Math. Phys. **21**, 85–124 (1971).
[11] F. Treves, *Topological Vector Spaces, Distributions and Kernels* (Academic, New York, 1976).
[12] A. Petrov, Rep. Math. Phys. **4**, 183–188 (1973).

Improving the convergence and estimating the accuracy of summation approximants of 1/*D* expansions for Coulombic systems

Melchior O. Elout, David Z. Goodson,[a] Carl D. Elliston,[b] Shi-Wei Huang, and Alexei V. Sergeev[c]
Department of Chemistry, Southern Methodist University, Dallas, Texas 75275

Deborah K. Watson
Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019

(Received 30 March 1998; accepted for publication 4 May 1998)

The convergence of large-order expansions in $\delta=1/D$, where D is the dimensionality of coordinate space, for energies $E(\delta)$ of Coulomb systems is strongly affected by singularities at $\delta=1$ and $\delta=0$. Padé–Borel approximants with modifications that completely remove the singularities at $\delta=1$ and remove the dominant singularity at $\delta=0$ are demonstrated. A renormalization of the interelectron repulsion is found to move the dominant singularity of the Borel function $F(\delta)=\Sigma_j E'_j/j!$, where E'_j are the the expansion coefficients of the energy with singularity structure removed at $\delta=1$, farther from the origin and thereby accelerate summation convergence. The ground-state energies of He and H_2^+ are used as test cases. The new methods give significant improvement over previous summation methods. *Shifted* Borel summation using $F_m(\delta)=\Sigma_j E'_j/\Gamma(j+1-m)$ is considered. The standard deviation of results calculated with different values of the shift parameter m is proposed as a measure of summation accuracy. © *1998 American Institute of Physics.* [S0022-2488(98)04210-8]

I. INTRODUCTION

Large-order perturbation theory in $1/D$, where D is the dimensionality of coordinate space, was proposed by Mlodinow[1,2] as an alternative to variational methods and coupling-constant perturbation theories for solving the Schrödinger equation. This ''dimensional perturbation theory'' has some appealing features. The first-order theory gives a semiquantitative description of electron correlation[3–11] and other subtle many-body effects[12] and it is relatively easy to calculate the perturbation expansion to very high order, using recursion relations.[13,14] $1/D$ expansions have now been calculated through 30th order for two-electron atoms[15] and tenth order for three-electron atoms.[16] Large-order expansions have also been calculated for the H atom in an external field[17–21] and for the H_2^+ molecule.[19,22,23]

The expansions for energies of Coulombic systems are, in general, divergent, due to singularities in the energy function $E(\delta)$, with $\delta=1/D$ treated as a continuous complex variable. One type of singularity is a pole in $E(\delta)$ that results from the fact that for certain integer values of D the expectation value of the Coulomb potential can diverge at particle coalescences.[24] This singularity can be accurately characterized and removed either by subtraction[25] or by rescaling,[5,26] making partial sums of the expansion appear rapidly convergent at low orders. A second characteristic type of singularity is a complicated branch point or essential singularity at $\delta=0$, which leads to divergence at higher orders.[15] This singularity can in principle be explained by an instanton analysis.[27–29] The zeroth-order limit of the perturbation theory corresponds to localization at an extremum of an effective potential that is the sum of the three-dimensional Coulomb potential and a centrifugal potential that comes from the dimensional continuation of the kinetic energy operator. For applications to bound systems, one generally chooses an extremum that corresponds

[a]Electronic mail: dgoodson@mail.smu.edu
[b]Current address: Dept. of Physics, Columbia University, 538 W. 120th St., New York, NY 10027.
[c]Permanent address: S. I. Vavilov State Optical Institute, 199034 St. Petersburg, Russian Federation.

 © 1998 American Institute of Physics

to a stable minimum. However, there typically exist sub-barrier trajectories in imaginary time that allow for tunneling out of the minimum. This causes the zeroth-order limit to be singular, with the energy expansion coefficients diverging as a factorial.

Summation methods in which the Coulombic poles are removed from $E(\delta)$ and then the remainder expansion is summed with Padé approximants or Padé–Borel approximants yield apparently convergent results at all orders.[15,30,31] Our purpose here is to refine those methods to improve the rate of convergence. Our strategy is to renormalize the Hamiltonian, by redefining the dimensional continuation in such a way that the effect of the singularity at the origin is mitigated, and to improve the techniques for modeling the singularities. In addition, we develop a method for estimating the accuracy of a given summation approximant. He and H_2^+ are used as test cases.

II. SUMMATION APPROXIMANTS

The asymptotic expansion in $1/D$ for the electronic energy of an atom or molecule has the form

$$E(\delta)\sim\delta^2\sum_{j=0}^{k}E_j\delta^j, \tag{1}$$

where $\delta=1/D$. There are various algorithms for computing the expansion coefficients E_j, but the most efficient, especially for problems with more than one degree of freedom, seems to be the matrix method of Dunn *et al.*[14] The direct evaluation of Eq. (1) at $\delta=1/3$ using partial sums, from truncation at given k, typically gives slow convergence at low orders and rapid divergence at high order.

For ground states of Coulombic systems, E has an expansion about $D=1$ in the form[24]

$$E(\delta)\sim a_{-2}(D-1)^{-2}+a_{-1}(D-1)^{-1}+\cdots. \tag{2}$$

For a one-electron atom with nuclear charge Z, the first term in Eq. (2) constitutes the exact solution, with $a_{-2}=2Z^2$. For two-electron atoms a_{-2} is equal to the ground-state energy of a two-electron Schrödinger equation in which the Coulomb potentials have been replaced with delta functions.[24,32] The exact solution for this equation was obtained by Rosenthal[33] for arbitrary Z.

To remove the effects of the poles one can either rescale the expansion,[5,26] by multiplying it by $(1-\delta)^2$, or explicitly subtract out the δ expansions of the poles.[25,30] If the residues are known exactly then subtraction works best. We consider two approaches for the case where a_{-2} is known but a_{-1} is not known: subtracting the second-order pole and rescaling to remove the first-order pole,

$$E(\delta)\sim\delta^2[a_{-2}(1-\delta)^{-2}+(1-\delta)^{-1}E'(\delta)], \tag{3a}$$

$$E'(\delta)=\sum_{j=0}^{k}\delta^jE'_j,\quad E'_j=E_j-E_{j-1}-a_{-2},\quad E_{-1}=0, \tag{3b}$$

or subtraction of both poles,

$$E(\delta)\sim\delta^2[a_{-2}(1-\delta)^{-2}+a_{-1}(1-\delta)^{-1}+E'(\delta)], \tag{4a}$$

$$E'(\delta)=\sum_{j=0}^{k}\delta^jE'_j,\quad E'_j=E_j-(j+1)a_{-2}-a_{-1}, \tag{4b}$$

with

$$a_{-1}=\lim_{\delta\to1}\sum_{j=0}^{k}\delta^j(E_j-E_{j-1}-a_{-2}). \tag{5}$$

Equation (5) can be evaluated using Padé summation.[30] The use of approximants based on Eqs. (4) has been called *hybrid* summation,[25] since it is the combination of an expansion about $\delta=1$ and an expansion about $\delta=0$. We will refer to the approach in Eqs. (3) as *rescaled hybrid* summation.

It is possible to obtain convergent results by summing the expansions for E' with Padé approximants or with Padé-Borel approximants. The latter give a somewhat better convergence rate. Here we employ a more general technique[28] that will be useful not just for summing the energy expansion but also for analyzing the singularity structure at $\delta=0$ and for estimating summation accuracy. Consider a function $F_m(\delta)$ with asymptotic expansion

$$F_m(\delta)\sim\sum_{j=j_0}^{k}\delta^j F_{m,j}, \quad F_{m,j}=E'_j/\Gamma(j+1-m), \tag{6}$$

where j_0 is the smallest integer greater than or equal to m. This function is related to $E'(\delta)$ by the transform

$$E'(\delta)=\sum_{j=0}^{j_0-1}E'_j\delta^j+e^{-(m-1)i\xi}\lim_{y\to\infty}\int_0^{ye^{i\xi}}x^{-m}F_m(\delta x e^{i\xi})\exp(-xe^{i\xi})dx, \tag{7}$$

with the angle ξ chosen so that $F(\delta x e^{i\xi})$ is nonsingular along the path of integration and $-\pi/2<\xi<\pi/2$. If $m=0$ then Eq. (7) is the usual Borel sum. For the function F_m in Eq. (7) we will substitute Padé approximants of the expansion in Eq. (6). For $m\neq 0$ we will refer to the summation method as *shifted* Padé–Borel summation.

III. RENORMALIZATION

A. Charge renormalization

The standard approach to analytic continuation of the Hamiltonian to arbitrary D is to continue the kinetic energy operator to D dimensions while keeping the physical, three-dimensional, definition for the potential energy operator.[32,34] If the energy is expressed in atomic units multiplied by Z^2/D^2, then the effective potential for an S state of the two-electron atom at large D is

$$V_{\rm eff}(r_1,r_2,\theta)=V_{\rm centr}+V_{\rm Coul}, \tag{8}$$

$$V_{\rm centr}=\left(\frac{1-6\delta+8\delta^2}{8\sin^2\theta}-\frac{1}{8}\right)\left(\frac{1}{r_1^2}+\frac{1}{r_2^2}\right), \tag{9}$$

$$V_{\rm Coul}=-r_1^{-1}-r_2^{-1}+\lambda\,(r_1^2+r_2^2-2r_1r_2\cos\theta)^{-1/2}. \tag{10}$$

$V_{\rm centr}$ is the centrifugal potential from the D-dimensional kinetic energy operator while $V_{\rm Coul}$ is the three-dimensional Coulomb potential, with $\lambda=1/Z$. This effective potential has a stable global minimum, with $r_1=r_2$, as long as λ is less than a critical value $\lambda_c\approx 0.8144$. For H^-, with $\lambda=1$, this symmetric extremum is a saddle point,[35,36] and if this point is used for the zeroth order of the perturbation theory, then the coefficients of the energy expansion will be complex numbers. Similarly, the symmetric large-D extrema for the H_2^+ and H_2 molecules also correspond to saddle points.[35,37,38]

A strategy for calculating $1/D$ expansions for systems that are unstable in the large-D limit was proposed in Ref. 39, using the fact that the definition of the dimensional continuation of the Hamiltonian is arbitrary as long as it gives the correct operator at $D=3$. If we renormalize the repulsive part of the Coulomb potential, by replacing it with a continuous function of δ such that the repulsion becomes weaker as δ approaches zero, it is possible to ensure that the symmetric extremum will be a minimum. This approach is closely related to the renormalized perturbation theory developed by Killingbeck for anharmonic oscillators.[40]

Charge renormalization was applied to large-order perturbation theory for H^- in Ref. 41. The parameter λ in Eq. (10) was replaced with the linear polynomial

$$\lambda(\delta)=\lambda_0+\lambda_1\delta, \quad \lambda_1=3(Z^{-1}-\lambda_0) \tag{11}$$

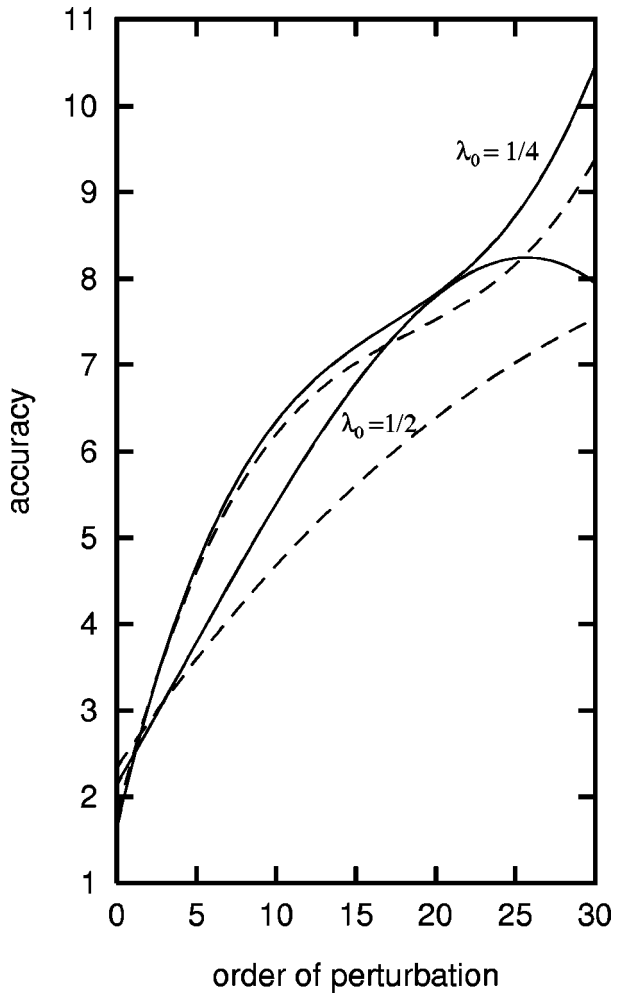

FIG. 1. Cubic polynomial fits of the number of accurate digits from summation approximants for the ground state energy of He, as a function of order, with ''accuracy'' defined as $-\log|\Delta E/E_{\text{exact}}|$, where $\Delta E = E_{\text{approx}} - E_{\text{exact}}$, with E_{exact} from Ref. 42. The dashed curves are from unshifted hybrid Padé–Borel approximants while the solid curves are from unshifted rescaled hybrid Padé–Borel approximants. Results are shown for $\lambda_0 = 1/2$ (unrenormalized) and for $\lambda_0 = 1/4$, as labeled, with $\lambda(\delta) = \lambda_0 + \lambda_1 \delta$.

with λ_0 treated as a free parameter. λ_1 is defined so that $\lambda(1/3) = 1/Z$. (This modification of λ does not affect the scale factor in the energy units. We continue to use atomic units multiplied by Z^2/D^2, where Z is independent of D.) The hybrid expansion, Eq. (4a), with Padé–Borel summation of E' gave convergent results for H^- for any value of λ_0 less than λ_c. The fastest convergence was for $\lambda_0 \approx (2Z)^{-1}$, which gave a more accurate result at given order for the ground-state energy of H^- than had been obtained[15] for the ground-state energy of He with unrenormalized λ. This rapid convergence for H^- suggests that the use of a δ-dependent potential energy operator could be used to improve the convergence even for systems that are stable at large D.

Figure 1 shows the effect on the convergence for He of setting $\lambda_0 = (2Z)^{-1}$. Since there is a fair amount of scatter in the convergence from order to order, we have plotted cubic polynomial fits to more clearly show the convergence patterns. The dashed curves correspond to hybrid approximants, based on Eq. (4a), while the solid curves correspond to rescaled hybrid approximants, based on Eq. (3a). The expansion of E' was summed with unshifted Padé–Borel approximants. We define the ''accuracy'' of a summation approximant S_k at given order k as

$$\alpha(k) = -\log_{10}|(S_k - E_{\text{exact}})/E_{\text{exact}}|, \tag{12}$$

which is a continuous measure of the number of accurate digits.

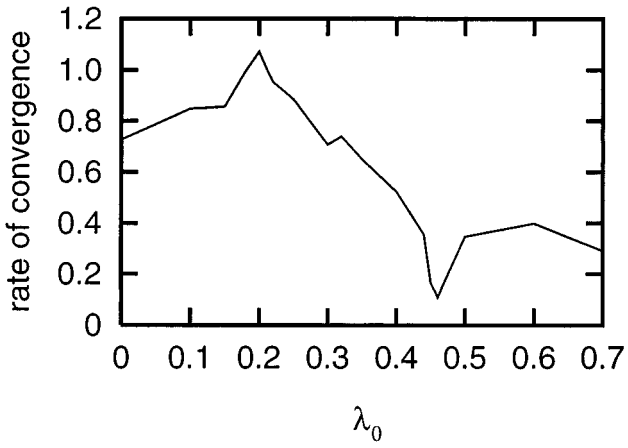

FIG. 2. The rate of convergence at low order for unshifted rescaled hybrid Padé–Borel approximants for the ground state energy of He, as a function of the renormalization parameter λ_0, using $\lambda(\delta)=\lambda_0+\lambda_1\delta$. The rate of convergence is defined here as the parameter α_1 from the three-parameter least-squares fit of the accuracy $\alpha(k)=\alpha_0+\alpha_1 k+\alpha_2 k^2+\alpha_3 k^3$ with α_0 constrained to equal the accuracy of S_0.

Several characteristic trends are evident. First, it is clear that the rescaled hybrid summation is the better of the two summation methods. It is possible to somewhat improve the accuracy of the unscaled hybrid approximants by applying Shanks extrapolation, with some subjective judgement, to the Padé-approximant sequences for a_{-1}, but the results remain less dependable than those from the rescaled hybrid method. The general shapes of the rescaled hybrid curves are representative of those over a wide range of λ_0. They begin with an almost linear increase in the number of accurate digits. Then, upon reaching an accuracy plateau of between approximately 6 and 8 digits the rate of convergence slows significantly. For $\lambda_0<0.8/Z$ there is a return to a more rapid convergence rate, beginning at around 25th order. Figure 2 shows the initial slope of $\alpha(k)$ as a function of λ_0. The increasing rate of convergence is balanced at very low order by the fact that the zeroth-order accuracy decreases for λ_0 far from the physical value. (The zeroth-order result is equal to the exact energy for $\lambda_0\approx 0.46$, which corresponds to a dip in Fig. 2.) Nevertheless, there is for the He atom a clear advantage to choosing λ_0 in the general neighborhood of $0.5/Z$.

The functional form chosen for $\lambda(\delta)$ is essentially arbitrary. In addition to the linear polynomial, Eq. (11), we have considered a quadratic polynomial,

$$\lambda(\delta)=\lambda_0+\lambda_1\delta+\lambda_2\delta^2, \tag{13}$$

$$\lambda_1=\frac{9}{2Z}-4\lambda_0-\frac{1}{2}\lambda(1), \quad \lambda_2=3\lambda_0+\frac{3}{2}\lambda(1)-\frac{9}{2Z}, \tag{14}$$

with $\lambda(1)$ and $\lambda_0=\lambda(0)$ treated as free parameters. This makes it possible to control the behavior at both of the solvable limits, $\delta=0$ and $\delta=1$. The position of the dominant singularities in the Borel function depend only on the value of λ_0. Therefore the rate of convergence of the summation approximants depends less strongly on $\lambda(1)$ than on $\lambda(0)$. We find in practice that certain choices of $\lambda(1)$ do improve the convergence at low and intermediate orders, especially for H^-, for which the linear $\lambda(\delta)$ can imply a very large value for $\lambda(1)$.

B. Effect on singularity structure

Analysis of the expansion coefficients for the ground-state energy of two-electron atoms has shown[15] that the dominant singularities of $F_0(\delta)$ (i.e., the singularities nearest to the origin of the complex δ plane) are a complex-conjugate pair of square-root branch points in the negative half-plane. Let $-\beta e^{\pm ib}$ be their locations. Then the coefficients of the asymptotic expansion in δ

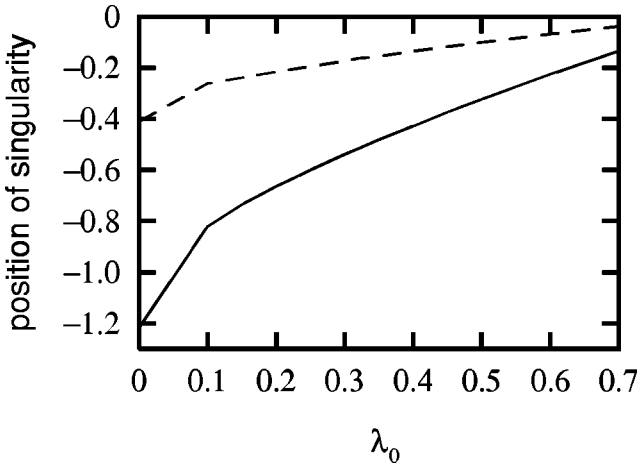

FIG. 3. Real part (solid curve) and imaginary part (dashed curve) of one of the dominant singularities in the Borel functions of the two-electron isoelectronic series, as a function of λ_0. The other dominant singularity is at the complex conjugate of this location.

of the expression $G(\delta+\beta e^{ib})^{1/2}+G^*(\delta+\beta e^{-ib})^{1/2}$, where G is a constant, will in the limit of large order be equal to the coefficients $F_{0,j}$ of the asymptotic expansion of F_0, Eq. (6). Thus, in the limit of large j we have

$$F_{0,j}\sim 2|G|\beta^{1/2}(-1)^j\,\frac{\Gamma(j-1/2)}{\Gamma(-1/2)\Gamma(j+1)}\,\beta^{-j}\,\cos\left[\left(j-\frac{1}{2}\right)b-\arg\,G\right]. \tag{15}$$

Now compare this to the function

$$f(\delta)=\gamma(\delta+\beta e^{ib})^{-1}+\gamma^*(\delta+\beta e^{-ib})^{-1}, \tag{16}$$

which has the expansion coefficients

$$f_j=2|\gamma|\beta^{-1}(-1)^j\beta^{-j}\,\cos[(j+1)b-\arg\,\gamma]. \tag{17}$$

If we set $\gamma=2^{-1}\pi^{-1/2}\beta^{3/2}|G|e^{i(\arg G+3b/2)}$, then $f_j\sim F_{3/2,j}$, which implies that the dominant singularities of $F_{3/2}(\delta)$ will be first-order poles at the same locations as the dominant branch points in $F_0(\delta)$.

The effect on convergence of the renormalization can be understood in terms of the effect on the locations of the dominant singularities of $F_m(\delta)$. A convenient method for determining these locations is to examine the singularities of the Padé approximants of F_m. Previously[15] we used the quadratic Padé approximants of F_0 for this purpose, noting the positions of the branch points nearest the origin. Once the branch points are approximately identified in this manner, the result can be refined by noting the positions of the corresponding poles in the linear Padé approximants of $F_{3/2}$, which stabilize somewhat faster.

We find that the positions of these singularities depend only on the value of λ_0. In general, there appear to be three distinct complex-conjugate pairs of singularities in the Borel functions. For λ_0 close to λ_c the dominant singularities are in the positive half plane, but as λ_0 decreases this pair rapidly moves away from the origin. For $\lambda_0<0.8$ the dominant singularities are the pair in the negative half plane. Figure 3 shows the real and imaginary parts of this pair. The steady movement away from the origin as λ_0 decreases is presumably related to the improvement in convergence. In the limit $\lambda_0\to 0$ the singularities in the Borel function must disappear, since $E'(\delta)$ is then nonsingular. It is interesting to note that the mechanism for this disappearance is for the prefactor multiplying the dominant singularities to go to zero, rather than for the positions of the singularities to move out to infinity. The third pair of singularities has an imaginary part of $\pm\pi$ with a small real part that depends weakly on λ_0. This is far enough from the origin that these can be expected to have no significant effect on the rate of convergence.

C. Effect on roundoff error

The recursive computation of the energy expansion coefficients E_k is somewhat unstable to roundoff error,[14,15] on account of their factorial divergence. However, in the hydrogenic limit $Z \to \infty$ (i.e., $\lambda \to 0$) the energy expansion is convergent, with only a linear increase in E_k with k. At small but nonzero λ the perturbation theory is still divergent but the growth of the expansion coefficients is nearly linear until very high order. This apparently comes about through massive cancellations of intermediate quantities that grow factorially. The result is extreme loss of precision near the hydrogenic limit.

The number of significant figures in the E_k decreases approximately linearly with k. It was stated previously[15,31] that roundoff error in the E_k becomes manifest in the S_k at about the value of k for which the precision of the E_k intersects the accuracy of the S_k. More careful analysis reveals that this rule is generally valid but somewhat conservative. The actual propagation of error in the summation approximants is complicated, with some approximants much less sensitive than others and thereby retaining their accuracy beyond the intersection. Such approximants can be identified empirically by adding to the E_k random error of a given magnitude.

For the renormalized calculations we find that the rate of precision loss depends mainly on the value of the parameter λ_0, increasing as λ_0 decreases. We have used quadruple-precision arithmetic (32 decimal digits) for computing the E_k in the calculations reported here, so that the roundoff error is always less than the convergence accuracy.

D. Renormalization for diatomic molecules

Within the Born–Oppenheimer approximation the Hamiltonian is not homogeneous, since the internuclear distance R is treated as a parameter rather than as a dynamical variable. Therefore, transformation to atomic units multiplied by D^{-2} replaces R with the D-dependent parameter $\tilde{R} = R/D^2$. In a recent study of H_2^+ it was shown that renormalization of this D dependence is a simple and effective way to improve the convergence of the $1/D$ expansion.[23] One replaces $\tilde{R}$ with

$$\tilde{R} = \frac{2(3+\beta)}{(D-1)(D+\beta)} R, \tag{18}$$

with β treated as a free parameter. (The factor of $D-1$ is included to give a reasonable $D \to 1$ limit.) An important effect of this renormalization is to ensure that the $D \to \infty$ limit corresponds to an effective potential with a single global minimum. Otherwise the molecule undergoes a symmetry breaking transition to a double-well problem as R increases, which significantly degrades the convergence of the large-order expansion. For many-electron diatomic molecules a renormalization of the interelectron repulsion, completely analogous to the procedure for atoms in Sec. III A, could be used to attenuate the effect on convergence of the interelectron repulsion in addition to the R renormalization that prevents the symmetry breaking.

IV. ESTIMATING THE ACCURACY OF THE SUMMATION APPROXIMANTS

We have calculated the summation approximants S_k using a range of integer and half-integer values for the shift parameter m in the Borel function F_m. The best results tend to correspond to $m = 0$, 1/2, 1, and 3/2, but which of these four is best at given order appears to be random. Therefore, to obtain an estimate of the accuracy of the summation at given order k we compute the standard deviation σ_k from the mean $\overline{S_k}$ of the four different S_k.

Figure 4 shows results for the He ground state energy with $\lambda_0 = 0.5/Z$. The dotted curve shows the actual accuracy of the mean. The solid curve shows $-\log_{10}|\sigma_k/\overline{S_k}|$. The largest disagreements between the actual accuracy and this estimate occurs at very high orders, where the mean tends to be significantly more accurate than the individual approximants and the estimate is too conservative.

V. DARBOUX–BOREL APPROXIMANTS

The singularity structure of the Borel function $F(\delta)$ for H_2^+ is similar to that for the two-electron atom except that the dominant singularity is typically a square-root branch point that lies

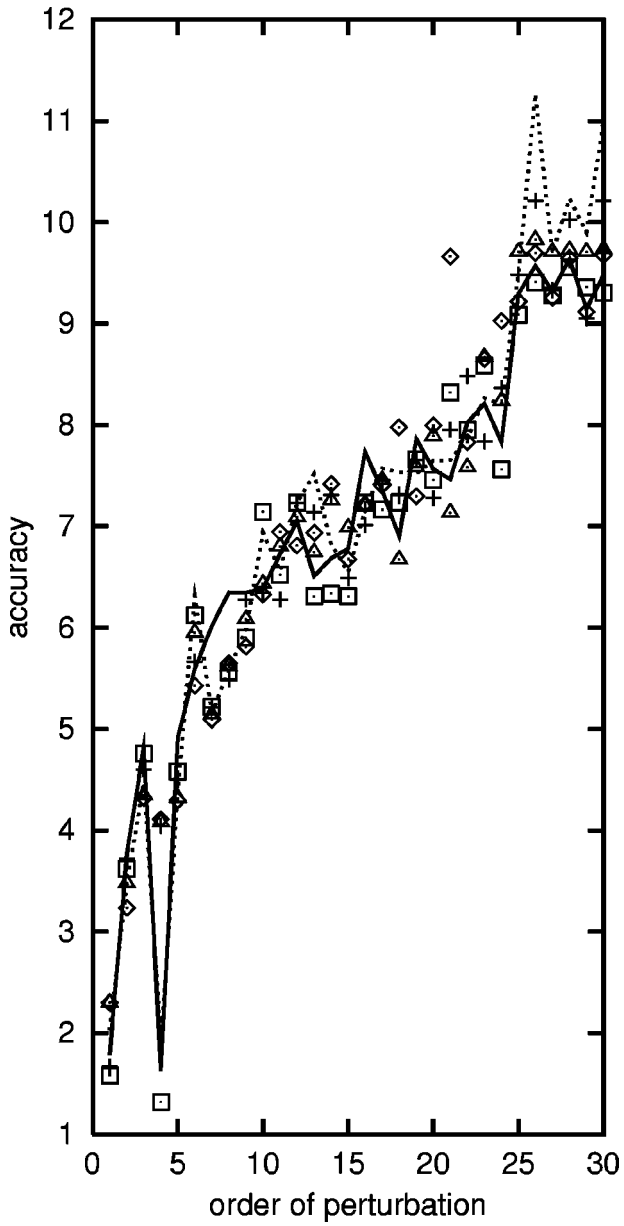

FIG. 4. Accuracy of rescaled hybrid Padé–Borel approximants for the ground state energy of He, as a function of order in the perturbation expansion. The points are determined using Borel functions F_m with the values of the shift parameter indicated as follows: $+$, $m=0$; $\square$, $m=-1/2$; $\triangle$, $m=-1$; $\diamond$, $m=-3/2$. The dotted curve shows $-\log_{10}|(\overline{S_k}-E_{\text{exact}})/E_{\text{exact}}|$, where $\overline{S_k}$ is the mean of the four summation approximants at order k. The solid curve shows $-\log_{10}|\sigma_k/\overline{S_k}|$, where σ_k is the standard deviation from the mean of the four approximants at order k.

on the negative real axis.[22,29] Because this problem is separable, in ellipsoidal coordinates, it is straightforward to derive an exact solution for the location δ_0 of the branch point.[29] Since this singularity is the nearest to the origin in the complex plane, it follows from Darboux's theorem[43] that the expression $g(\delta)(1-\delta/\delta_0)^{1/2}+h(\delta)$, where g and h are nonsingular for $|\delta|\leqslant|\delta_0|$, has the same expansion coefficients in the limit of large order as the actual expansion of F. This suggests the use of Darboux approximants,[44]

$$F_{[L,N/M]}(\delta)=\frac{R_N(\delta)}{Q_M(\delta)}(1-\delta/\delta_0)^{1/2}+\frac{P_L(\delta)}{Q_M(\delta)}, \tag{19}$$

where P_L, Q_M, and R_N are polynomials of degrees L, M, and N, respectively. These polyno-

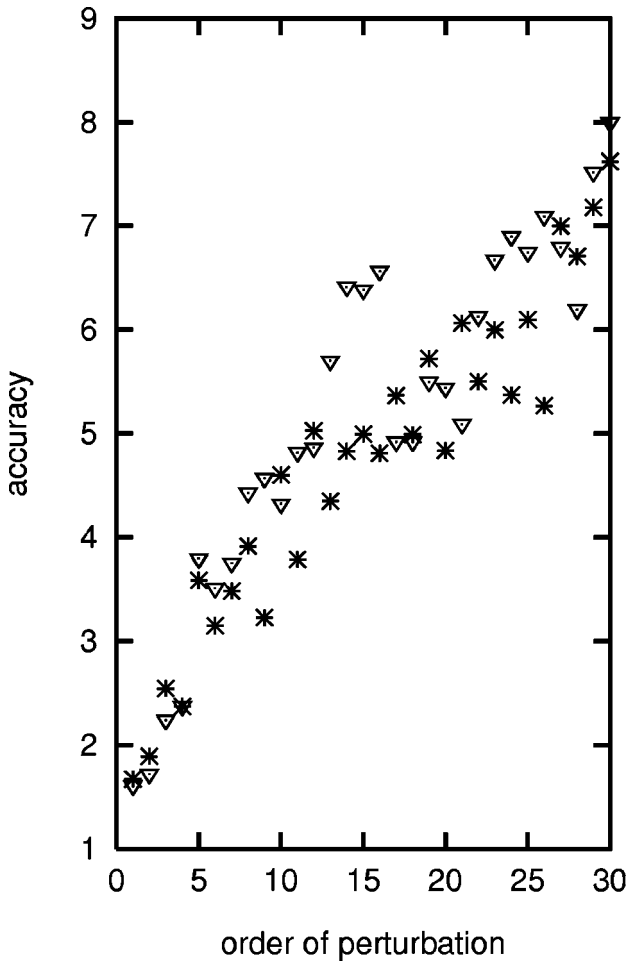

FIG. 5. Accuracy of rescaled hybrid Darboux–Borel approximants (∇) and rescaled hybrid Padé–Borel approximants (∗) for the ground state energy of H_2^+ with internuclear distance $R=2$, as a function of order in the perturbation expansion. The Borel summation is unshifted.

mials can be determined by setting $F_{[L,M/N]}$ equal to the expansion of F, expanding the square root, multiplying through by Q_M, and then solving the resulting set of linear equations.

Figure 5 compares accuracies of Darboux–Borel and Padé–Borel approximants for the ground-state energy of H_2^+. These results correspond to internuclear distance $R=2$, which is very close to the equilibrium bond length. (The renormalized D-dimensional Hamiltonian for this problem is described in Ref. 23.) There is noticeable improvement in accuracy from the Darboux approximants. This is seen as well for other values of R.

VI. DISCUSSION

Our central results are the following: (1) Summation accuracy depends significantly on removing the Coulombic poles at the $D\rightarrow 1$ limit, and rescaling removes the first-order pole more effectively than subtraction using an approximate residue. (2) Renormalization of the D-dimensional Hamiltonian, by multiplying the interelectron potential with a polynomial in δ can significantly improve the rate of convergence. (3) Comparison of results from shifted Borel transforms, given by Eqs. (6) and (7), yields a standard deviation σ_k that provides a reliable estimate of the summation accuracy at given order k. A reasonable summation procedure is to carry out a series of inexpensive moderate-order calculations using a quadratic $\lambda(\delta)$ over the full range of λ_0 values, with some arbitrary value of $\lambda(1)$, in order to determine an approximate range of λ_0 that can be expected to yield the fastest convergence. In the case of He, for example, a plot such as in Fig. 2 would suggest $0.15\leqslant\lambda_0\leqslant 0.25$. Then additional moderate-order calculations could be used to determine an optimal pair of λ_0 and $\lambda(1)$ to be used in the expensive large-order calculation.

One could smooth the convergence of the approximant sequence by using the mean $\overline{S_k}$ of various shifted approximants, eliminating $\overline{S_k}$ that have large σ_k. For example, if σ_{27} is greater than σ_{26}, as in Fig. 4, then $\overline{S_{26}}$ would be retained as the 27th-order result.

We find that the best two-parameter fit of summation accuracy versus order, for two-electron atoms and for H_2^+, is $\alpha(k)=\alpha_0+\alpha_1 k^{1/2}$. However, there appears to be a systematic deviation from this behavior, with a leveling off of accuracy at intermediate orders followed by an increase at higher orders, which led us to use cubic polynomials for the fits in Fig. 1. For He the plateau is at approximately eight significant figures. This effect was noted in Ref. 15, for unrenormalized expansions, and was attributed to the need to model more than just the dominant singularity in the Borel function. For He one obtains approximately three significant figures by explicitly including the poles at $\delta=1$, and apparently, seven figures from modeling also the dominant branch point in the Borel function. This explanation is supported by the behavior of the Darboux–Borel summation for H_2^+ in Fig. 5, which explicitly includes the square-root branch point. In that case there is rapid convergence to almost seven figures at 14th order but then it is not until 29th order that the convergence resumes. The plateau for the Padé–Borel results is less distinct, presumably because at intermediate orders the Padé approximants are still refining the description of the branch point while also attempting to decipher other singularity structure. By order 27 the Padé–Borel accuracy is about equal to the Darboux–Borel accuracy.

We have shown results only for ground states, but in general we expect similar singularity structure, and hence similar convergence behavior, for the lowest eigenstate of any given symmetry. With other excited states there are branch points in the *energy* function $E(\delta)$.[31] Although not the dominant singularities, in practice they dominate the convergence behavior until very high order. For example, it has been estimated[31] that the asymptotic behavior due to the singularity at the origin for the $1s2s\ ^1S$ state of He will not assert itself until 88th order. In that case Borel summation is not very effective, and it is more appropriate to use a summation method designed to model the apparent dominating singularities.

Recent tenth-order calculations for the three-electron atom[16] show convergence behavior almost identical to that for two-electron atoms. It remains to be seen if this will hold as well for four-electron atoms. Loeser[6] has formulated the $1/D$ expansion for S states of all atoms in the periodic table. However, with four electrons there are ten independent internal coordinates at large D but only 9 at $D=3$. Therefore it is not clear that Loeser's theory will converge to the physical solution when taken to large order. Such coordinate redundancies proliferate quadratically as the number of electrons increases. An alternative formulation that avoids this situation is ''partial'' dimensional continuation, applied recently in a study of triatomic molecular rotation spectra.[45] In that case one body-fixed axis was allowed to rotate in D dimensions while the other axis was treated as rotating about the first in only three dimensions. Some sort of related procedure could, in principle, give a D-dependent Hamiltonian with the physical number of coordinates.

Probably a more serious obstacle to applying the $1/D$ expansion to many-electron systems is the rapid increase in computational cost with increasing numbers of internal coordinates.[14] With larger systems it will be necessary to introduce approximations, such as separability assumptions[46] or basis-set truncations. The tradeoff will be between a small systematic error, from a large-order inexact theory, and a random, but possibly larger, error from truncating the exact theory at lower order.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation and the Robert. A. Welch Foundation.

[1] L. D. Mlodinow, ''The large N expansion in quantum mechanics,'' Ph.D. dissertation, University of California, Berkeley, 1981.
[2] L. D. Mlodinow, in *Progress in Particle and Nuclear Physics*, edited by D. Wilkinson (Pergamon, Oxford, 1982), Vol. 8, pp. 387–399.
[3] P. du T. van der Merwe, J. Chem. Phys. **81**, 5976 (1984); **82**, 5293 (1985).
[4] O. Goscinski and V. Mujica, Int. J. Quantum Chem. **29**, 897 (1986).
[5] D. R. Herschbach, J. Chem. Phys. **84**, 838 (1986).
[6] J. G. Loeser, J. Chem. Phys. **86**, 5635 (1987).
[7] D. Z. Goodson and D. R. Herschbach, J. Chem. Phys. **86**, 4997 (1987).

[8] D. R. Herschbach, J. G. Loeser, and D. K. Watson, Z. Phys. D **10**, 195 (1988).
[9] S. Kais and D. R. Herschbach, J. Chem. Phys. **100**, 4367 (1994).
[10] D. R. Herschbach, in *Dimensional Scaling in Chemical Physics*, edited by D. R. Herschbach, J. Avery, and O. Goscinski (Kluwer, Dordrecht, 1993), pp. 7–80, and references therein.
[11] J. G. Loeser and D. R. Herschbach, in *New Methods in Quantum Theory*, edited by C. A. Tsipis, V. S. Popov, D. R. Herschbach, and J. S. Avery (Kluwer, Dordrecht, 1996), pp. 1–33, and references therein.
[12] P. Serra and S. Kais, Chem. Phys. Lett. **260**, 302 (1996); Phys. Rev. A **55**, 238 (1997); Phys. Rev. Lett. **77**, 466 (1996).
[13] V. M. Vaĭnberg, V. D. Mur, V. S. Popov, A. V. Sergeev, and A. V. Shcheblykin, Theor. Math. Phys. **74**, 269 (1988) [Teor. Mat. Fiz. **74**, 399 (1988)].
[14] M. Dunn, T. C. Germann, D. Z. Goodson, C. A. Traynor, J. D. Morgan III, D. K. Watson, and D. R. Herschbach, J. Chem. Phys. **101**, 5987 (1994).
[15] D. Z. Goodson, M. López-Cabrera, D. R. Herschbach, and J. D. Morgan III, J. Chem. Phys. **97**, 8481 (1992).
[16] M. O. Elout and D. Z. Goodson (unpublished).
[17] C. M. Bender, L. D. Mlodinow, and N. Papanicolaou, Phys. Rev. A **25**, 1305 (1982).
[18] V. M. Vaĭnberg, V. D. Mur, V. S. Popov, and A. V. Sergeev, JETP Lett. **44**, 9 (1986) [Pis'ma Zh. Eksp. Teor. Fiz. **44**, 9 (1986)].
[19] V. D. Mur, V. S. Popov, and A. V. Sergeev, Sov. Phys. JETP **70**, 16 (1990) [Zh. Eksp. Teor. Fiz. **97**, 32 (1990)]; and references therein.
[20] V. M. Vaĭnberg, V. S. Popov, and A. V. Sergeev, Sov. Phys. JETP **71**, 470 (1990) [Zh. Eksp. Teor. Fiz. **98**, 847 (1990)].
[21] T. C. Germann, D. R. Herschbach, M. Dunn, and D. K. Watson, Phys. Rev. Lett. **74**, 658 (1995).
[22] M. López-Cabrera, D. Z. Goodson, D. R. Herschbach, and J. D. Morgan III, Phys. Rev. Lett. **68**, 1992 (1992).
[23] S.-W. Huang, D. Z. Goodson, M. López-Cabrera, and T. C. Germann, Phys. Rev. A **58**, 250 (1998).
[24] D. J. Doren and D. R. Herschbach, Phys. Rev. A **34**, 2654 (1986); J. Chem. Phys. **87**, 433 (1987).
[25] D. J. Doren and D. R. Herschbach, Chem. Phys. Lett. **118**, 115 (1985).
[26] M. López-Cabrera, A. L. Tan, and J. G. Loeser, J. Phys. Chem. **97**, 2467 (1993).
[27] L. N. Lipatov, JETP Lett. **25**, 79 (1977) [Pis'ma Zh. Eksp. Teor. Fiz. **25**, 116 (1977)].
[28] J. Zinn-Justin, Phys. Rep. **109**, 109 (1981).
[29] V. S. Popov and A. V. Sergeev, Phys. Lett. A **193**, 165 (1994); JETP Lett. **78**, 303 (1994) [Zh. Eksp. Teor. Fiz. **105**, 568 (1994)].
[30] D. Z. Goodson and D. R. Herschbach, Phys. Rev. A **46**, 5428 (1992).
[31] D. Z. Goodson and D. K. Watson, Phys. Rev. A **48**, 2668 (1993).
[32] D. R. Herrick and F. H. Stillinger, Phys. Rev. A **11**, 42 (1975).
[33] C. M. Rosenthal, J. Chem. Phys. **55**, 2474 (1971).
[34] M. Dunn and D. K. Watson, Ann. Phys. (N.Y.) **251**, 266 (1996); **251**, 319 (1996); Few-Body Syst. **21**, 187 (1996).
[35] P. du T. van der Merwe, Phys. Rev. A **36**, 3446 (1987).
[36] D. J. Doren and D. R. Herschbach, J. Phys. Chem. **92**, 1816 (1988).
[37] D. D. Frantz and D. R. Herschbach, Chem. Phys. **126**, 59 (1988).
[38] A. L. Tan and J. G. Loeser, in *Dimensional Scaling in Chemical Physics*, edited by D. R. Herschbach, J. Avery, and O. Goscinski (Kluwer Academic, Dordrecht, 1993), pp. 230–255.
[39] C. A. Traynor and D. Z. Goodson, J. Phys. Chem. **97**, 2464 (1993).
[40] J. Killingbeck, J. Phys. A **14**, 1005 (1981).
[41] D. K. Watson and D. Z. Goodson, Phys. Rev. A **51**, R5 (1995).
[42] J. D. Baker, D. E. Freund, R. N. Hill, and J. D. Morgan III, Phys. Rev. A **41**, 1247 (1990).
[43] B. W. Ninham, J. Math. Phys. **4**, 679 (1963).
[44] D. Z. Goodson and M. López-Cabrera, in *Dimensional Scaling in Chemical Physics*, edited by D. R. Herschbach, J. Avery, and O. Goscinski (Kluwer Academic, Dordrecht, 1993), pp. 275–314.
[45] A. A. Suvernev and D. Z. Goodson, J. Chem. Phys. **107**, 4099 (1997).
[46] D. Z. Goodson, Phys. Rev. A **55**, 4155 (1997).

Fermion realization of the nuclear Sp(6,*R*) model

Jutta Escher[a] and J. P. Draayer
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001

(Received 6 April 1998; accepted for publication 29 May 1998)

A fermion realization of the nuclear Sp(6,*R*) model, which complements the traditional bosonic representation, is developed. A recursive process is presented in which symplectic matrix elements of arbitrary one-body fermion operators between states of excitation $N\hbar\omega$ and $N'\hbar\omega$ in the same or in different symplectic bands are related back to valence shell matrix elements, which can be evaluated by standard shell model techniques. Matrix elements so determined may be used to calculate observables such as electron scattering form factors which carry detailed structural information on nuclear wave functions. © *1998 American Institute of Physics.* [S0022-2488(98)02110-0]

I. INTRODUCTION

Extensive effort has been devoted to developing the relevant mathematical and computational tools for a fully microscopic treatment of collective nuclear phenomena. In particular, ever since the noncompact symplectic group in three dimensions, Sp(6,*R*), was recognized as the appropriate dynamical group for a many-body theory of nuclear collective motion,[1–3] it has received careful and detailed attention. This group is of special interest because it is also the dynamical group of the three-dimensional harmonic oscillator and thus it establishes an important link between the nuclear shell model and the collective model. Generalized vector coherent state theory and boson realizations of the symplectic algebra have been employed to construct the discrete infinite-dimensional unitary irreducible representations of Sp(6,*R*).[4–13] Methods have been introduced for calculating the necessary orthonormalization factors for symplectic basis states and the requisite matrix elements of the symplectic generators.[7,14–16] In addition, a procedure has been developed for evaluating matrix elements of general two-body interactions of the type used in standard microscopic treatments of nuclear many-body systems.[17–19]

Despite practical limitations due to the large sizes of the Hilbert spaces involved, various applications have proven the symplectic approach to be successful in reproducing collective nuclear properties, such as excitation energies, quadrupole moments, and electromagnetic transition probabilities.[20–24] Unfortunately, until now it has not been possible to evaluate matrix elements of arbitrary one-body operators between symplectic basis states. These matrix elements are of particular interest since they are required for the calculation of observables, such as nuclear form factors, which carry very detailed structural information on nuclear wave functions. It is our purpose in this article to provide a new, fermionic, realization of the sp(6,*R*) algebra, which complements the traditional bosonic representation and leads to a method for determining matrix elements of a general one-body operator in a $\mathrm{Sp}(6,R)\supset\mathrm{U}(3)$ basis. Specifically, a recursive process is presented in which symplectic matrix elements of arbitrary one-body fermion operators between states of excitation $N\hbar\omega$ and $N'\hbar\omega$ in the same or in different symplectic bands are related back to valence shell matrix elements, which can be evaluated by standard shell model techniques. The fermionic realization of the symplectic algebra is particularly well suited for this approach, since the requisite valence shell matrix elements are readily available.[25]

The article is organized as follows: To establish the notation, in Sec. II we give a brief summary of the generators of the symplectic algebra, their bosonic representation, commutation relations, and matrix elements. In Sec. III the symplectic generators are recast in terms of fermi-

[a]Present address: Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel; Electronic mail: jutta@algebron.fiz.huji.ac.il

 © 1998 American Institute of Physics

onic creation and annihilation operators. The symplectic model space is reviewed in Sec. IV. In Sec. V we present the derivation of the matrix element formula for a general one-body operator acting in the symplectic space. Concluding remarks are given in Sec. VI. The derivations in Secs. III and V make use of both oscillator boson operators and fermion creation and annihilation operators. Phase conventions, conjugation properties, commutation relations, and matrix elements of the former are given in Appendix A, and conjugation properties and commutation relations of various SU(3)-coupled products of the latter are derived in Appendix B. In addition, Appendix C provides a compilation of various important SU(3) recoupling rules.

II. GENERATORS OF THE SYMPLECTIC ALGEBRA

The generators of the symplectic algebra $sp(6,R)$ can be realized in terms of bilinear products of harmonic oscillator bosons in a Cartesian scheme:[26]

$$C_{ij}=\sum_s \left(b_i^\dagger(s)b_j(s)+\frac{1}{2}\delta_{ij}\right)-\frac{1}{A}\sum_s b_i^\dagger(s)\sum_t b_j(t)-\frac{1}{2}\delta_{ij},$$

$$B_{ij}^\dagger=\frac{1}{2}\sum_s b_i^\dagger(s)b_j^\dagger(s)-\frac{1}{2A}\sum_s b_i^\dagger(s)\sum_t b_j^\dagger(t), \tag{1}$$

$$B_{ij}=\frac{1}{2}\sum_s b_i(s)b_j(s)-\frac{1}{2A}\sum_s b_i(s)\sum_t b_j(t),$$

where the sums run over all A particles in the system and the two-body $1/A$ terms effect the removal of spurious center-of-mass excitations from the A-particle system. (Refer to Appendix A for details on the boson creation and annihilation operators.) The commutation relations for the symplectic algebra in this basis are easily inferred from the commutation relations of its building blocks and are listed by Hecht.[13]

Making use of the spherical components $b_{1q}^{\dagger(10)}(s)$ and $\tilde{b}_{1q}^{(01)}(s)$ of the boson creation and annihilation operators, the symplectic generators can be written as SU(3) irreducible tensor operators:

$$C_{lm}^{(11)}=\sqrt{2}\sum_s \{b^\dagger(s)\times\tilde{b}(s)\}_{lm}^{(11)}-\frac{\sqrt{2}}{A}\sum_{st}\{b^\dagger(s)\times\tilde{b}(t)\}_{lm}^{(11)},$$

$$A_{lm}^{(20)}=\frac{1}{\sqrt{2}}\sum_s \{b^\dagger(s)\times b^\dagger(s)\}_{lm}^{(20)}-\frac{1}{\sqrt{2}A}\sum_{st}\{b^\dagger(s)\times b^\dagger(t)\}_{lm}^{(20)}, \tag{2}$$

$$B_{lm}^{(02)}=\frac{1}{\sqrt{2}}\sum_s \{\tilde{b}(s)\times\tilde{b}(s)\}_{lm}^{(02)}-\frac{1}{\sqrt{2}A}\sum_{st}\{\tilde{b}(s)\times\tilde{b}(t)\}_{lm}^{(02)},$$

transforming according to the SU(3) irreps $(\lambda\mu)=(11)$, (20), and (02), respectively. The overall normalization factor is chosen in agreement with the convention used by Rosensteel.[27] The symplectic $2\hbar\omega$ raising and lowering operators, $A_{lm}^{(20)}$ and $B_{lm}^{(02)}$, respectively, are related to each other via Hermitian conjugation: $B_{lm}^{(02)}=(-1)^{l-m}(A_{l-m}^{(20)})^\dagger$. Both $A_{lm}^{(20)}$ and $B_{lm}^{(02)}$ have $l=0$ and 2 components, whereas $C_{lm}^{(11)}$ has $l=1$ and 2 components. The relevant commutation relations between the spherical components of the symplectic raising and lowering operators are given by Rosensteel:[27]

$$[A^{(20)}_{l_1m_1},A^{(20)}_{l_2m_2}]=[B^{(02)}_{l_1m_1},B^{(02)}_{l_2m_2}]=0,$$

$$[B^{(02)}_{l_1m_1},A^{(20)}_{l_2m_2}]=\frac{1}{2}\sqrt{10}\sum_{lm}\langle(02)l_1m_1;(20)l_2m_2|(11)lm\rangle C^{(11)}_{lm}$$
$$+2\sqrt{\frac{2}{3}}(-1)^{l_2+m_2}\delta_{l_1l_2}\delta_{m_1(-m_2)}H_0. \tag{3}$$

Here $\langle -;-|-\rangle$ denotes a Wigner SU(3) coupling coefficient (see Appendix C) and $H_0=\hat{N}_b+\frac{3}{2}$, where $\hat{N}_b=\Sigma_{i=1}^{3}C_{ii}$ counts the number of harmonic oscillator bosons in the system.

The eight operators $C^{(11)}_{lm}$ generate the $su(3)$ subalgebra of $sp(3,R)$ and are related to the angular momentum operator L_q and the Elliott (algebraic) quadrupole operator $Q^a_{2m}=\sqrt{4\pi/5}\Sigma_s(r_s^2Y_{2m}(\hat{r}_s)+p_s^2Y_{2m}(\hat{p}_s))$, here given in units of $\hbar=\omega=m=1$, as follows:

$$C^{(11)}_{1q}=L_q,\quad q=0,\pm1$$
$$C^{(11)}_{2m}=\frac{1}{\sqrt{3}}Q^a_{2m},\quad m=0,\pm1,\pm2. \tag{4}$$

We can also express the collective quadrupole operator $Q^c_{2m}=\sqrt{16\pi/5}\Sigma_s r_s^2Y_{2m}(\hat{r}_s)$ as a linear combination of SU(3) irreducible tensor operators:

$$Q^c_{2m}=Q^a_{2m}+\sqrt{3}(A^{(20)}_{2m}+B^{(02)}_{2m}). \tag{5}$$

Matrix elements for $C^{(11)}$ in the standard SU(3) basis are given by[28,29]

$$\langle(\lambda'\mu')|||C^{(11)}|||(\lambda\mu)\rangle=(-1)^{\phi}\sqrt{2C_2(\lambda\mu)}\,\delta_{(\lambda'\mu')(\lambda\mu)},$$
$$\phi=\begin{cases}1, & \text{for } \mu\neq0\\ 0, & \text{for } \mu=0,\end{cases} \tag{6}$$
$$C_2(\lambda\mu)=\frac{2}{3}(\lambda^2+\lambda\mu+\mu^2+3\lambda+3\mu).$$

The reduced matrix element $\langle(\lambda'\mu')|||C^{(11)}|||(\lambda\mu)\rangle$ is related to the full SU(3) matrix element via the Wigner–Eckart theorem for SU(3) (see Appendix C). The symbol $C_2(\lambda\mu)$ denotes the second-order Casimir invariant of SU(3), and the choice of the phase is consistent with that of reference 29.

Several strategies for calculating matrix elements of the symplectic generators $A^{(20)}$ and $B^{(02)}$ have been explored. A direct way is to use the Sp(6,R) commutation relations to derive recursion formulas, as shown by Rosensteel.[27] Another approach is to start from approximate matrix elements and proceed by successive approximations, adjusting the matrix elements until the commutation relations are precisely satisfied.[3] Deenen and Quesne[16] have employed a boson mapping to obtain the generator matrix elements, and Castaños *et al.*[15] have derived simple analytical functions for some special irreps. The most elegant method, outlined by Rowe in reference 7, involves vector-valued coherent state representation theory and evaluates matrix elements of the symplectic raising and lowering operators by relating them to the matrix elements of a much simpler $u(3)\otimes$*Weyl* algebra.

III. FERMION REALIZATION OF THE SYMPLECTIC GENERATORS

In a fermion second quantization formulation, the one-body part of $C^{(11)}$ takes the form

$$\mathscr{F}(C^{(11)s=0}_{LM\sigma=0}) = \left(1-\frac{1}{A}\right)\sqrt{2}\sum_{\rho\rho'}\left\langle \underbrace{(\nu 0) lm \frac{1}{2}\sigma''}_{\rho}\middle| \{b^\dagger \times \tilde{b}\}^{(11)s=0}_{LM\sigma=0}\middle| \underbrace{(\nu' 0) l' m' \frac{1}{2}\sigma'}_{\rho'}\right\rangle \times a^\dagger_{(\nu 0) lm (1/2)\sigma''} a_{(\nu' 0) l' m' (1/2)\sigma'}, \tag{7}$$

where $a^\dagger_\rho$ and a_ρ are fermion creation and annihilation operators (see Appendix B). Note that $b^\dagger$, $\tilde{b}$, and $C^{(11)}_{LM}$ do not act on the spin part of the wave functions; thus one needs to treat these operators as $s=\sigma=0$ objects, as has been explicitly expressed in the above equation.

Utilizing the symmetry properties of the Clebsch–Gordan and SU(3) Wigner coefficients and the definition of the proper SU(3) irreducible tensor operator $\tilde{a}_{(0\nu') l' (-m') (1/2) (-\sigma')} \equiv (-1)^{\nu'+l'-m'+1/2-\sigma'} a_{(\nu' 0) l' m' (1/2)\sigma'}$, we obtain

$$\begin{aligned}\mathscr{F}(C^{(11)s=0}_{LM\sigma=0}) &= \left(1-\frac{1}{A}\right)\sqrt{2}\sum_{\nu\nu'}\left\langle (\nu 0)\frac{1}{2}\middle\|\middle\| \{b^\dagger\times\tilde{b}\}^{(11)s=0}\middle\|\middle\| (\nu' 0)\frac{1}{2}\right\rangle \sqrt{d(\nu 0)}\,\frac{1}{2} \\ &\quad \times \{a^\dagger_{(\nu 0)(1/2)} \times \tilde{a}_{(0\nu')(1/2)}\}^{(11)s=0}_{LM\sigma=0} \\ &= \left(1-\frac{1}{A}\right)\sum_{\nu} \sqrt{\frac{1}{6}\nu(\nu+1)(\nu+2)(\nu+3)} \\ &\quad \times \{a^\dagger_{(\nu 0)(1/2)} \times \tilde{a}_{(0\nu)(1/2)}\}^{(11)s=0}_{LM\sigma=0},\end{aligned} \tag{8}$$

where $d(\lambda\mu)=\frac{1}{2}(\lambda+1)(\mu+1)(\lambda+\mu+2)$ denotes the dimension of the su(3) irrep $(\lambda\mu)$ and use has been made of Eq. (A6), which expresses the triple-reduced matrix element in terms of ν and ν' only (see Appendix A). Analogously one derives the fermion realization for the one-body part of the symplectic raising and lowering operators.

The two-body parts of the symplectic generators can be obtained in the same manner. A general symmetric two-body operator for A identical particles, $\mathscr{G}=\sum_{s<t=1}^{A} g(\mathbf{x}_s,\boldsymbol{\sigma}_s,\mathbf{x}_t,\boldsymbol{\sigma}_t)$, where s and t refer to the s-th and t-th particles, respectively, takes the following form in a fermion second quantized formulation:

$$\mathscr{G} = \frac{1}{4}\sum_{\substack{\rho_1,\rho_1' \\ \rho_2,\rho_2'}} \langle \rho_1\rho_2 | g(\mathbf{x}_1,\boldsymbol{\sigma}_1,\mathbf{x}_2,\boldsymbol{\sigma}_2) | \rho_1'\rho_2'\rangle a^\dagger_{\rho_1} a^\dagger_{\rho_2} a_{\rho_2'} a_{\rho_1'}. \tag{9}$$

Here $|\rho_1\rho_2\rangle$ denotes the direct product of the single-particle wave functions $|\rho_1\rangle$ and $|\rho_2\rangle$. For the cases that are of interest here, the function $g(\mathbf{x}_1,\boldsymbol{\sigma}_1,\mathbf{x}_2,\boldsymbol{\sigma}_2)$ can be written as a product $g(1)g(2)\equiv g(\mathbf{x}_1,\boldsymbol{\sigma}_1)g(\mathbf{x}_2,\boldsymbol{\sigma}_2)$, where $g(t)$, with $t=1$ or 2, acts solely on the single-particle wave function with subscript t. Thus the two-body part of the symplectic generator $C^{(11)}$ is written as

$$\mathcal{G}(C^{(11)s=0}_{LM\sigma=0}) = -\frac{1}{\sqrt{8}A}\sum_{\substack{\rho_1,\rho_1'\\ \rho_2,\rho_2'}}\Bigg\langle \underbrace{(\nu_1 0) l_1 m_1 \frac{1}{2}\sigma_1}_{\rho_1};\underbrace{(\nu_2 0) l_2 m_2 \frac{1}{2}\sigma_2}_{\rho_2}\Bigg|$$
$$\times\{b^\dagger(1)\times\tilde{b}(2)\}^{(11)s=0}_{LM\sigma=0}\Bigg|\underbrace{(\nu_1' 0) l_1' m_1' \frac{1}{2}\sigma_1'}_{\rho_1'};\underbrace{(\nu_2' 0) l_2' m_2' \frac{1}{2}\sigma_2'}_{\rho_2'}\Bigg\rangle$$
$$\times a^\dagger_{(\nu_1 0) l_1 m_1 (1/2)\sigma_1} a^\dagger_{(\nu_2 0) l_2 m_2 (1/2)\sigma_2} a_{(\nu_2' 0) l_2' m_2' (1/2)\sigma_2'} a_{(\nu_1' 0) l_1' m_1' (1/2)\sigma_1'}$$
$$= -\frac{1}{\sqrt{8}A}\sum_{\substack{(\lambda\mu)S\\ (\lambda'\mu')S'}}\sqrt{\frac{1}{8}d(\lambda\mu)(2S+1)}\Big\langle\{(\nu_1 0)\times(\nu_2 0)\}(\lambda\mu);\Big\{\frac{1}{2}\times\frac{1}{2}\Big\}S\Big|\Big|\Big|$$
$$\times\{b^\dagger(1)\times\tilde{b}(2)\}^{(11)s=0}\Big|\Big|\Big|\{(\nu_1' 0)\times(\nu_2' 0)\}(\lambda'\mu');\Big\{\frac{1}{2}\times\frac{1}{2}\Big\}S'\Big\rangle$$
$$\times\sum_{\substack{\nu_1\nu_2\\ \nu_1'\nu_2'}}\{\{a^\dagger_{(\nu_1 0)(1/2)}\times a^\dagger_{(\nu_2 0)(1/2)}\}^{(\lambda\mu)S}\times\{\tilde{a}_{(0\nu_1')(1/2)}\times\tilde{a}_{(0\nu_2')(1/2)}\}^{(\mu'\lambda')S'}\}^{(11)s=0}_{LM\sigma=0}, \tag{10}$$

and analogously for $\mathcal{G}(A^{(20)s=0}_{LM\sigma=0})$ and $\mathcal{G}(B^{(02)s=0}_{LM\sigma=0})$. Here use has been made of the symmetry properties of the Clebsch–Gordan and Wigner coefficients and of the Wigner–Eckart theorem for both SU(2) and SU(3). Upon evaluation of the triple-reduced matrix elements of $\{b^\dagger(1)\times\tilde{b}(2)\}^{(11)}$, $\{b^\dagger(1)\times b^\dagger(2)\}^{(20)}$, and $\{\tilde{b}(1)\times\tilde{b}(2)\}^{(02)}$ (see Appendix A), we find the following fermionic expressions for the symplectic generators $C^{(11)}$, $A^{(20)}$, and $B^{(02)}$, respectively:

$$C^{(11)}_{LM} = \Big(1-\frac{1}{A}\Big)\sum_\nu\sqrt{\frac{1}{6}\nu(\nu+1)(\nu+2)(\nu+3)}\{a^\dagger_{(\nu 0)(1/2)}\times\tilde{a}_{(0\nu)(1/2)}\}^{(11)s=0}_{LM\sigma=0}$$
$$-\frac{1}{4A}\sum_{\substack{\nu\nu' S\\ (\lambda\mu)(\lambda'\mu')}}\sqrt{\frac{1}{4}d(\lambda\mu)\nu(\nu'+3)(2S+1)}\begin{Bmatrix}(\nu-1,0) & (10) & (\nu 0) & -\\ (\nu'+1,0) & (01) & (\nu' 0) & -\\ (\lambda'\mu') & (11) & (\lambda\mu) & -\\ - & - & - & \end{Bmatrix}$$
$$\times\{\{a^\dagger_{(\nu 0)(1/2)}\times a^\dagger_{(\nu' 0)(1/2)}\}^{(\lambda\mu)S}\times\{\tilde{a}_{(0,\nu-1)(1/2)}\times\tilde{a}_{(0,\nu'+1)(1/2)}\}^{(\mu'\lambda')S}\}^{(11)s=0}_{LM\sigma=0}, \tag{11}$$

$$A^{(20)}_{LM} = \Big(1-\frac{1}{A}\Big)\sum_\nu\sqrt{\frac{1}{12}(\nu+1)(\nu+2)(\nu+3)(\nu+4)}\{a^\dagger_{(\nu+2,0)(1/2)}\times\tilde{a}_{(0\nu)(1/2)}\}^{(20)s=0}_{LM\sigma=0}$$
$$-\frac{1}{4A}\sum_{\substack{\nu\nu' S\\ (\lambda\mu)(\lambda'\mu')}}\sqrt{\frac{1}{12}d(\lambda\mu)\nu\nu'(2S+1)}\begin{Bmatrix}(\nu-1,0) & (10) & (\nu 0) & -\\ (\nu'-1,0) & (10) & (\nu' 0) & -\\ (\lambda'\mu') & (20) & (\lambda\mu) & -\\ - & - & - & \end{Bmatrix}$$
$$\times\{\{a^\dagger_{(\nu 0)(1/2)}\times a^\dagger_{(\nu' 0)(1/2)}\}^{(\lambda\mu)S}\times\{\tilde{a}_{(0,\nu-1)(1/2)}\times\tilde{a}_{(0,\nu'-1)(1/2)}\}^{(\mu'\lambda')S}\}^{(20)s=0}_{LM\sigma=0}, \tag{12}$$

and

$$B^{(02)}_{LM}=\left(1-\frac{1}{A}\right)\sum_{\nu}\ \sqrt{\frac{1}{12}(\nu+1)(\nu+2)(\nu+3)(\nu+4)}\{a^{\dagger}_{(\nu 0)(1/2)}\times\tilde{a}_{(0,\nu+2)(1/2)}\}^{(02)s=0}_{LM\sigma=0}$$

$$-\frac{1}{4A}\sum_{\substack{\nu\nu' S\\(\lambda\mu)(\lambda'\mu')}}\sqrt{\frac{1}{12}d(\lambda\mu)(\nu+3)(\nu'+3)(2S+1)}\begin{Bmatrix}(\nu+1,0) & (01) & (\nu 0) & -\\ (\nu'+1,0) & (01) & (\nu' 0) & -\\ (\lambda'\mu') & (02) & (\lambda\mu) & -\\ - & - & - & \end{Bmatrix}$$

$$\times\{\{a^{\dagger}_{(\nu 0)(1/2)}\times a^{\dagger}_{(\nu' 0)(1/2)}\}^{(\lambda\mu)S}\times\{\tilde{a}_{(0,\nu+1)(1/2)}\times\tilde{a}_{(0,\nu'+1)(1/2)}\}^{(\mu'\lambda')S}\}^{(02)s=0}_{LM\sigma=0}. \tag{13}$$

Making use of the properties of the coupling coefficients and of the Hermitian conjugation properties of the generators, one can verify that $B^{(02)}_{LM}=(-1)^{L-M}(A^{(20)}_{L-M})^{\dagger}$ holds, as expected. This relation serves as a stringent test for the fermionic expressions of $A^{(20)}$ and $B^{(02)}$.

IV. SYMPLECTIC MODEL SPACE

A basis for the Hilbert space is generated by applying symmetrically coupled products of the $2\hbar\omega$ raising operator $A^{(20)}$ with itself to the usual $0\hbar\omega$ shell-model states. The $0\hbar\omega$ starting configurations are labeled by the Elliott SU(3) quantum numbers $(\lambda_\sigma\mu_\sigma)$[30,31] and by N_σ, the eigenvalue of the oscillator boson number operator which takes the minimum value consistent with the Pauli Exclusion Principle. The product of $N_n/2$ raising tensors $A^{(20)}$, each of which promotes a particle from a given shell into a higher-lying shell $2\hbar\omega$ above, generates $N_n\hbar\omega$ excitations for each starting irrep $N_\sigma(\lambda_\sigma\mu_\sigma)$. Each such product operator $\mathscr{P}^{N_n(\lambda_n\mu_n)}_{\alpha_n}$, labeled according to its SU(3) content, $(\lambda_n\mu_n)$, is then coupled with $|N_\sigma(\lambda_\sigma\mu_\sigma)\rangle$ to good SU(3) symmetry $\rho(\lambda_\omega\mu_\omega)$, with ρ denoting the multiplicity of the coupling $(\lambda_n\mu_n)\times(\lambda_\sigma\mu_\sigma)$.

It will be convenient to use the general shorthand notation, Γ, for a U(3) or SU(3) representation label, and α for an appropriate set of U(3) subgroup labels. We thus introduce, following the notation of Hecht,[32]

$$\begin{aligned}\Gamma_\sigma&\equiv[\sigma_1\sigma_2\sigma_3]\equiv N_\sigma(\lambda_\sigma\mu_\sigma)=N_\sigma(\sigma_1-\sigma_2,\sigma_2-\sigma_3),\\ \Gamma_n&\equiv[n_1n_2n_3]\equiv N_n(\lambda_n\mu_n)=N_n(n_1-n_2,n_2-n_3),\\ \Gamma_\omega&\equiv[\omega_1\omega_2\omega_3]\equiv N_\omega(\lambda_\omega\mu_\omega)=N_\omega(\omega_1-\omega_2,\omega_2-\omega_3),\end{aligned} \tag{14}$$

where $(\lambda\mu)$ are SU(3) labels and the $N_\sigma=\sigma_1+\sigma_2+\sigma_3$, $N_n=n_1+n_2+n_3$, and $N_\omega=\omega_1+\omega_2+\omega_3$ give the number of squares in the U(3) Young tableaux. With this convention, the product operators, which are defined recursively, can be written as

$$\mathscr{P}^{\Gamma_n}_{\alpha_n}(A^{(20)})=\sum_{\Gamma_{n'}\beta\alpha_{n'}}\langle(20)\beta;\Gamma_{n'}\alpha_{n'}|\Gamma_n\alpha_n\rangle X^{\Gamma_n}(\Gamma_{n'})A^{(20)}_{\beta}\mathscr{P}^{\Gamma_{n'}}_{\alpha_{n'}}(A^{(20)}), \tag{15}$$

with $A^{(20)}$ adding a $2\hbar\omega$ excitation to the $N_{n'}$ excitation ($N_{n'}=n'_1+n'_2+n'_3$) that is created by the action of the operator $\mathscr{P}^{\Gamma_{n'}}_{\alpha_{n'}}(A^{(20)})$; and the factor

$$X^{\Gamma_n}(\Gamma_{n'})=\frac{1}{n_1+n_2+n_3}(n_1n_2n_3|||\mathscr{A}^{(20)}|||n'_1n'_2n'_3) \tag{16}$$

is required to properly normalize the raising polynomials (for details see reference 32). The n_i in the above equation denote the number of oscillator bosons in the i-th direction, and we have $N_n=N_{n'}+2$. The operator $\mathscr{A}^{(20)}$ is a generator of the $u(3)\otimes Weyl$ algebra; the evaluation of its matrix elements and its relation to the symplectic generator $A^{(20)}$ are discussed in references 33, 7.

We thus obtain a basis of Sp(3,R) states that are reduced according to the subgroup chain

$$\begin{array}{ccccccccc}\mathrm{Sp}(3,R) & \supset & \mathrm{U}(3) & \supset & \mathrm{SO}(3) & \supset & \mathrm{SO}(2)\\ \Gamma_\sigma & & \Gamma_n\rho\ \ \Gamma_\omega & & \kappa\ \ L & & M.\end{array} \tag{17}$$

For each $0\hbar\omega$ SU(3) starting irrep $\Gamma_\sigma = N_\sigma(\lambda_\sigma\mu_\sigma)$ a basis for a symplectic representation is given by states of the form

$$|\Phi[\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega]\rangle \equiv \sum_{\alpha_n\alpha_\sigma} \langle \Gamma_n\alpha_n;\Gamma_\sigma\alpha_\sigma|\Gamma_\omega\alpha_\omega\rangle_\rho \mathscr{P}^{\Gamma_n}_{\alpha_n}(A^{(20)})|\Gamma_\sigma\alpha_\sigma\rangle, \tag{18}$$

where $N_n/2 = 0,1,2,\ldots$, counts the number of boson excitations, $N_\omega = N_\sigma + N_n$, $(\lambda_n\mu_n)$ ranges over the set $\Omega = \{(n_1-n_2, n_2-n_3) | n_1 \geqslant n_2 \geqslant n_3 \geqslant 0; N_n = n_1+n_2+n_3; n_1,n_2,n_3$ even integers$\}$, $\rho\Gamma_\omega$ includes all SU(3) irreps resulting from the coupling $\Gamma_n \times \Gamma_\sigma$, and $\alpha_\omega = \kappa LM$ denotes quantum numbers associated with the group chain SU(3)$\supset$SO(3)$\supset$SO(2). Alternatively, one can also choose subgroup labels $\alpha_\omega = \varepsilon\Lambda M_\Lambda$, which are associated with the chain SU(3)$\supset$SU(2)$\times$U(1)$\supset$SO(2). The states of the Sp(3,R)$\supset$SU(3) basis are thus labeled by three types of U(3) quantum numbers: $\Gamma_\sigma = N_\sigma(\lambda_\sigma\mu_\sigma)$, the symplectic bandhead or Sp(3,R) lowest weight U(3) symmetry, which specifies the Sp(3,R) irreducible representation; $\Gamma_n = N_n(\lambda_n\mu_n)$, the U(3) symmetry of the raising polynomial; and $\Gamma_\omega = N_\omega(\lambda_\omega\mu_\omega)$, the U(3) symmetry of the final state. Any given symplectic representation space $N_\sigma(\lambda_\sigma\mu_\sigma)$ is infinite dimensional, since $N_n/2$, the number of boson excitations, can take any positive integer value. In practical applications, one must therefore either truncate the symplectic Hilbert space, or restrict oneself to interactions and observables for which the matrix elements depend solely on the symplectic irrep and can be calculated analytically.

The states of Equation (18) are eigenstates of the harmonic oscillator Hamiltonian, $H_0|\Phi\rangle = E_0|\Phi\rangle$, with eigenvalues $E_0 = (N_\omega + \frac{3}{2})\hbar\omega$. Two such states with different U(3) content $\Gamma_\omega = N_\omega(\lambda_\omega\mu_\omega)$ are orthogonal, whereas two states with identical U(3) symmetry Γ_ω, but different $\rho\Gamma_n = \rho N_n(\lambda_n\mu_n)$ quantum numbers, are generally not orthogonal. The states $|\Phi(\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega)\rangle$ of Equation (18) can be related to the orthonormal basis states $|\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega\rangle$ of the unitary irreducible representation of Sp(3,R), by

$$|\Phi[\Gamma_\sigma\Gamma_{n_i}\rho_i\Gamma_\omega\alpha_\omega]\rangle = \sum_j [\mathscr{K}(\Gamma_\sigma,\Gamma_\omega)]_{\Gamma_{n_i}\rho_i,\Gamma_{n_j}\rho_j}|\Gamma_\sigma\Gamma_{n_j}\rho_j\Gamma_\omega\alpha_\omega\rangle. \tag{19}$$

Here $|\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega\rangle$, without the letter Φ stands for the orthonormal basis states, and the symbol $[\mathscr{K}(\Gamma_\sigma,\Gamma_\omega)]_{n_i\rho_i,n_j\rho_j}$ denotes the matrix elements of $\mathscr{K}$, the positive Hermitian square root of the overlap matrix $\mathscr{K}^2$, which has matrix elements

$$[\mathscr{K}^2(\Gamma_\sigma,\Gamma_\omega)]_{\Gamma_{n'}\rho',\Gamma_n\rho} \equiv \langle\Phi[\Gamma_\sigma\Gamma_{n'}\rho'\Gamma_\omega\alpha_\omega]|\Phi[\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega]\rangle. \tag{20}$$

The matrix $\mathscr{K}^2$ is diagonal in Γ_σ and Γ_ω and independent of U(3) subgroup labels α_ω, and its rows and columns are labeled by Γ_n and ρ. Due to the smallness of the off-diagonal matrix elements of $\mathscr{K}^2$ the orthonormal basis states in Equation (19) can also be tagged by those labels $\Gamma_{n_i}\rho_i$ which correspond to the dominant values of $\Gamma_n\rho$ in these states. The method for calculating the matrix elements of $\mathscr{K}^2$, and therefore of $\mathscr{K}$, is given in reference 7. An approximation formula for the matrix elements of $\mathscr{K}$ has been worked out by Hecht.[32]

V. MATRIX ELEMENTS OF ARBITRARY ONE-BODY OPERATORS

In this section we derive a recursion formula in which symplectic matrix elements of arbitrary one-body operators between states of excitation $N_{n_1}\hbar\omega$ and $N_{n_2}\hbar\omega$ in the same or in different symplectic bands are related back to valence shell matrix elements, which can be evaluated by standard shell model techniques. The derivation of the desired recursion formula makes use of the fact that the symplectic basis states are constructed by applying polynomials $\mathscr{P}^{\Gamma_n}_{\alpha_n}(A^{(20)})$ of symmetrically coupled products of the symplectic raising operator $A^{(20)}$ with itself to $0\hbar\omega$ shell model configurations $|\Gamma_\sigma\alpha_\sigma\rangle$. (As before, we will denote the orthonormal basis states by $|\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega\rangle$, without the letter Φ.) Using Equations (18) and (15) and the definition of the SU(3) Racah coefficients U (see Appendix C), we can now express a symplectic basis state $|\Phi[\Gamma_\sigma\Gamma_n\rho\Gamma_\omega\alpha_\omega]\rangle$ of excitation $N_n\hbar\omega$ in terms of basis states $|\Phi[\Gamma_\sigma\Gamma_{n'}\rho'\Gamma_{\omega'}\alpha_{\omega'}]\rangle$ of excitation $N_{n'}\hbar\omega$ with $N_{n'} = N_n - 2$:

$$|\Phi[\Gamma_\sigma \Gamma_n \rho \Gamma_\omega \alpha_\omega]\rangle = \sum_{\Gamma_{n'} \rho' \Gamma_{\omega'} \beta \alpha_{\omega'}} U[(20)\Gamma_{n'}\Gamma_\omega \Gamma_\sigma; \Gamma_{n-} \rho \Gamma_{\omega'} \rho'_-] \times \langle (20)\beta; \Gamma_{\omega'} \alpha_{\omega'} | \Gamma_\omega \alpha_\omega \rangle X^{\Gamma_n}(\Gamma_{n'}) A^{(20)}_\beta |\Phi[\Gamma_\sigma \Gamma_{n'} \rho' \Gamma_{\omega'} \alpha_{\omega'}]\rangle. \quad (21)$$

[This can be proved by decoupling the raising polynomial $\mathscr{P}^{\Gamma_n}_{\alpha_n}(A^{(20)})$ from the lowest weight state $|\Gamma_\sigma \alpha_\sigma\rangle$, inserting the definition of $\mathscr{P}^{\Gamma_n}_{\alpha_n}(A^{(20)})$ given in Equation (15), coupling the raising polynomial $\mathscr{P}^{\Gamma'_n}_{\alpha'_n}(A^{(20)})$ to $|\Gamma_\sigma \alpha_\sigma\rangle$, and making use of the property (C22) of the SU(3) Racah coefficient U.] An analogous expression can be obtained for the bra state $\langle \Phi[\Gamma_\sigma \Gamma_n \rho \Gamma_\omega \alpha_\omega]|$.

The task at hand is to evaluate matrix elements of arbitrary one-body operators. Since any one-body operator can be expanded in terms of the fermion unit operators $\{a^\dagger_{(\eta 0)(1/2)} \times \tilde{a}_{(0\eta')(1/2)}\}^{\Gamma S}_{\alpha\Sigma}$ by employing the formalism of second quantization and SU(3) recoupling techniques, it suffices to evaluate the following matrix element:

$$\langle \Gamma_{\sigma_2} \Gamma_{n_2} \rho_2 \Gamma_{\omega_2} \alpha_{\omega_2}; S_2 \Sigma_2 | \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} | \Gamma_{\sigma_1} \Gamma_{n_1} \rho_1 \Gamma_{\omega_1} \alpha_{\omega_1}; S_1 \Sigma_1 \rangle. \quad (22)$$

Here we have introduced a simplified, but unambiguous, notation for the fermion creation and annihilation operators: $a^\dagger_{(\eta 0)(1/2)} \to a^\dagger_\eta$, $\tilde{a}_{(0\eta)(1/2)} \to \tilde{a}_\eta$, and $S_1\Sigma_1$ and $S_2\Sigma_2$ denote the spin and spin projection of the ket and bra states, respectively. We need only consider the case of $\eta \geqslant \eta'$, since the matrix element of $\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}$ for $\eta < \eta'$ can be obtained through complex conjugation from the matrix element of $\{a^\dagger_{\eta'} \times \tilde{a}_\eta\}^{\tilde{\Gamma} S}_{\bar{\alpha} - \Sigma}$, where $\tilde{\Gamma} = (\mu\lambda)$ is the irrep conjugate to $\Gamma = (\lambda\mu)$ and $\bar{\alpha} = \kappa l(-m)$ for $\alpha = \kappa l m$.

Using the $\mathscr{K}$-matrix and applying the step-down procedure outlined above to the ket state $|\Phi[\Gamma_\sigma \Gamma_n \rho \Gamma_\omega \alpha_\omega]\rangle$, the matrix element can be written as

$$\begin{aligned}
&\langle \Gamma_{\sigma_2} \Gamma_{n_2} \rho_2 \Gamma_{\omega_2} \alpha_{\omega_2}; S_2 \Sigma_2 | \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} | \Gamma_{\sigma_1} \Gamma_{\hat{n}_1} \hat{\rho}_1 \Gamma_{\omega_1} \alpha_{\omega_1}; S_1 \Sigma_1 \rangle \\
&= \sum_{\Gamma_{n_1} \rho_1} [\mathscr{K}^{-1}(\Gamma_{\sigma_1}, \Gamma_{\omega_1})]_{\Gamma_{\hat{n}_1} \hat{\rho}_1, \Gamma_{n_1} \rho_1} \\
&\quad \times \langle \Gamma_{\sigma_2} \Gamma_{n_2} \rho_2 \Gamma_{\omega_2} \alpha_{\omega_2}; S_2 \Sigma_2 | \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} | \Phi[\Gamma_{\sigma_1} \Gamma_{n_1} \rho_1 \Gamma_{\omega_1} \alpha_{\omega_1}]; S_1 \Sigma_1 \rangle \\
&= \sum_{\Gamma_{n_1} \rho_1} [\mathscr{K}^{-1}(\Gamma_{\sigma_1}, \Gamma_{\omega_1})]_{\Gamma_{\hat{n}_1} \hat{\rho}_1, \Gamma_{n_1} \rho_1} \\
&\quad \times \sum_{\Gamma_{\omega'_1} \Gamma_{\hat{n}'_1} \hat{\rho}'_1} U[(20)\Gamma_{\hat{n}'_1} \Gamma_{\omega_1} \Gamma_{\sigma_1}; \Gamma_{n_1 -} \rho_1 \Gamma_{\omega'_1} \hat{\rho}'_{1-}] X^{\Gamma_{n_1}}(\Gamma_{\hat{n}'_1}) \\
&\quad \times \sum_{\Gamma_{n'_1} \rho'_1} [\mathscr{K}(\Gamma_{\sigma_1}, \Gamma_{\omega'_1})]_{\Gamma_{\hat{n}'_1} \hat{\rho}'_1, \Gamma_{n'_1} \rho'_1} \Bigg\{ \sum_{\alpha_{\omega'_1} \beta_1} \langle (20)\beta_1; \Gamma_{\omega'_1} \alpha_{\omega'_1} | \Gamma_{\omega_1} \alpha_{\omega_2} \rangle \\
&\quad \times \langle \Gamma_{\sigma_2} \Gamma_{n_2} \rho_2 \Gamma_{\omega_2} \alpha_{\omega_2}; S_2 \Sigma_2 | \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} A^{(20)}_{\beta_1} | \Gamma_{\sigma_1} \Gamma_{n'_1} \rho'_1 \Gamma_{\omega'_1} \alpha_{\omega'_1}; S_1 \Sigma_1 \rangle \Bigg\}. \quad (23)
\end{aligned}$$

Note that the symplectic raising and lowering operators do not act on the spin part of the wave functions.

We now consider the term in parentheses and express the operator $\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} A^{(20)}_{\beta_1}$ as

$$\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} A^{(20)}_{\beta_1} = A^{(20)}_{\beta_1} \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} + [\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}, A^{(20)}_{\beta_1}]. \quad (24)$$

Recalling that the fermion realization of $A^{(20)}$ is given by Equation (12), we obtain

$$\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} A^{(20)}_{\beta_1} = A^{(20)}_{\beta_1}\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}$$
$$+\left(1-\frac{1}{A}\right)\sum_{\nu}\sqrt{\frac{1}{12}(\nu+1)(\nu+2)(\nu+3)(\nu+4)}[\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}, \{a^\dagger_{\nu+2} \times \tilde{a}_{\nu}\}^{(20)s=0}_{\beta_1\sigma=0}]$$
$$-\frac{1}{4A}\sum_{\substack{\nu\nu' S' \\ \Gamma_a\Gamma_b}}\sqrt{\frac{d(\Gamma_a)}{12}\nu\nu'(2S'+1)}\begin{Bmatrix} (\nu-1,0) & (10) & (\nu 0) & - \\ (\nu'-1,0) & (10) & (\nu' 0) & - \\ \Gamma_b & (20) & \Gamma_a & - \\ - & - & - & \end{Bmatrix}$$
$$\times[\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}, \{\{a^\dagger_\nu \times a^\dagger_{\nu'}\}^{\Gamma_a S'} \times \{\tilde{a}_{\nu-1} \times \tilde{a}_{\nu'-1}\}^{\tilde{\Gamma}_b S'}\}^{(20)s=0}_{\beta_1\sigma=0}]. \tag{25}$$

Therefore, the expression in the parentheses of Equation (23) is comprised of three terms: $\{\cdots\} = C_1 + (1-1/A)C_2 + (1/4A)C_3$, where A denotes the number of nucleons in the system. The first of these is given by

$$C_1 \equiv \sum_{\alpha'_{\omega_1}\beta_1} \langle (20)\beta_1; \Gamma_{\omega'_1}\alpha'_{\omega_1} | \Gamma_{\omega_1}\alpha_{\omega_1}\rangle$$
$$\times \langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}\alpha_{\omega_2}; S_2\Sigma_2 | A^{(20)}_{\beta_1}\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma} | \Gamma_{\sigma_1}\Gamma_{n'_1}\rho'_1\Gamma_{\omega'_1}\alpha_{\omega'_1}; S_1\Sigma_1\rangle. \tag{26}$$

To evaluate C_1, we insert a complete set of states:

$$1 = \sum_{\substack{\Gamma_{\sigma''}\Gamma_{n''}\rho'' \\ \Gamma_{\omega''}\alpha_{\omega''}S''\Sigma''}} |\Gamma_{\sigma''}\Gamma_{n''}\rho''\Gamma_{\omega''}\alpha_{\omega''}; S''\Sigma''\rangle\langle\Gamma_{\sigma''}\Gamma_{n''}\rho''\Gamma_{\omega''}\alpha_{\omega''}; S''\Sigma''|, \tag{27}$$

between $A^{(20)}_{\beta_1}$ and $\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}$ and make use of the following relation:

$$\langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}\alpha_{\omega_2}; S_2\Sigma_2 | A^{(20)}_{\beta_1} | \Gamma_{\sigma''}\Gamma_{n''}\rho''\Gamma_{\omega''}\alpha_{\omega''}; S''\Sigma''\rangle$$
$$= \delta_{\Gamma_{\sigma''}\Gamma_{\sigma_2}}\delta_{N_{n''}N_2-2}\delta_{S''S_2}\delta_{\Sigma''\Sigma_2}\langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}\alpha_{\omega_2}; S_2\Sigma_2 | A^{(20)}_{\beta_1} | \Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''}\alpha_{\omega''}; S_2\Sigma_2\rangle. \tag{28}$$

The delta functions in this expression reflect the fact that the symplectic generator $A^{(20)}$ only connects states within a symplectic irrep which have the same spin and differ by $2\hbar\omega$ in their excitation. Note that these delta functions significantly restrict the sum over $\Gamma_{\sigma''}\Gamma_{n''}\rho''\Gamma_{\omega''}\alpha_{\omega''}S''\Sigma''$ of the complete set of states. Only states which are constructed from the lowest weight irrep $\Gamma_{\sigma''} = \Gamma_{\sigma_2}$ by applying a raising polynomial $\mathscr{P}^{\Gamma_{n''}}_{\alpha_{n''}}(A^{(20)})$ which is characterized by $\Gamma_{n''} = (n''_1, n''_2, n''_3)$ with $N_{n''} = n''_1 + n''_2 + n''_3 = n^{(2)}_1 + n^{(2)}_2 + n^{(2)}_3 - 2$ [where $\Gamma_{n_2} = (n^{(2)}_1, n^{(2)}_2, n^{(2)}_3)$] can yield nonvanishing contributions. Employing the (generalized) Wigner–Eckart theorem, and making use of the symmetry properties of the SU(3) coupling and recoupling coefficients, we obtain the following expression for C_1:

$$
\begin{aligned}
C_1 = (-1)^{\Gamma+\Gamma_{\omega_1}+\Gamma_{\omega_2}} & \sqrt{\frac{d(\Gamma_{\omega_1})}{d(20)d(\Gamma)}} \\
& \times \sum_{\Gamma_{n''}\rho''\Gamma_{\omega''}} \sqrt{d(\Gamma_{\omega''})} \langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}; S_2 | | | A^{(20)} | | | \Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''}; S_2 \rangle \\
& \times \sum_{\rho_3} \langle \Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''}; S_2 | | | \{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S} | | | \Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'}; S_1 \rangle_{\rho_3} \\
& \times \sum_{\rho_4\rho_5} \Phi_{\rho_3\rho_4}[\Gamma_{\omega''}\tilde{\Gamma}; \Gamma_{\omega_1'}] \Phi_{\rho_4\rho_5}[\Gamma_{\omega_1'}\tilde{\Gamma}_{\omega''}; \tilde{\Gamma}] \\
& \times \sum_{\rho_6} U[\Gamma_{\omega_2}\tilde{\Gamma}_{\omega''}\Gamma_{\omega_1}\Gamma_{\omega_1'}; (20)_{--}\tilde{\Gamma}\rho_5\rho_6] \langle \Gamma_{\omega_1}\alpha_{\omega_1}; \Gamma\alpha | \Gamma_{\omega_2}\alpha_{\omega_2} \rangle_{\rho_6} \langle S_1\Sigma_1 S\Sigma | S_2\Sigma_2 \rangle,
\end{aligned}
\tag{29}
$$

where we have introduced the abbreviation $(-1)^{\Gamma_i} \equiv (-1)^{(\lambda_i+\mu_i)}$ for $\Gamma_i = (\lambda_i \mu_i)$.

The matrix element of the second term in Equation (25) takes the form

$$
\begin{aligned}
C_2 \equiv \sum_\nu & \sqrt{\frac{1}{12}(\nu+1)(\nu+2)(\nu+3)(\nu+4)} \sum_{\alpha_{\omega_1'}\beta_1} \langle (20)\beta_1; \Gamma_{\omega_1'}\alpha_{\omega_1'} | \Gamma_{\omega_1}\alpha_{\omega_1} \rangle \\
& \times \langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}\alpha_{\omega_2}; S_2\Sigma_2 | [\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}, \{a^\dagger_{\nu+2} \times \tilde{a}_\nu\}^{(20)s=0}_{\beta_1\sigma=0}] | \Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'}\alpha_{\omega_1'}; S_1\Sigma_1 \rangle,
\end{aligned}
\tag{30}
$$

and can be evaluated by using the commutator of Equation (B8) (see Appendix B). The delta functions in Equation (B8) restrict the sum over ν to two terms only: one term with $\nu = \eta$ and one with $\nu = \eta' - 2$. Again making use of the Wigner–Eckart theorem and the symmetry properties of the coupling coefficients, we obtain

$$
\begin{aligned}
C_2 = & -(-1)^\Gamma \sqrt{\frac{1}{2}(\eta+1)(\eta+2)} \sqrt{\frac{d(\Gamma_{\omega_1})}{d(\Gamma)d(\Gamma_{\omega_1'})}} \\
& \times \sum_{\Gamma''} (-1)^{\Gamma''} \sqrt{d(\Gamma'')} U[(20)(\eta 0)\Gamma''(0\,\eta'); (\eta+2,0)\Gamma] \\
& \times \sum_{\rho_3} \langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}; S_2 | | | \{a^\dagger_{\eta+2} \times \tilde{a}_{\eta'}\}^{\Gamma'' S} | | | \Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'}; S_1 \rangle_{\rho_3} \\
& \times \sum_{\rho_4} U[\Gamma_{\omega_2}\tilde{\Gamma}''\Gamma_{\omega_1}(20); \Gamma_{\omega_1'}\rho_{3-}\tilde{\Gamma}_-\rho_4] \langle \Gamma_{\omega_1}\alpha_{\omega_1}; \Gamma\alpha | \Gamma_{\omega_2}\alpha_{\omega_2} \rangle_{\rho_4} \\
& \times \langle S_1\Sigma_1 S\Sigma | S_2\Sigma_2 \rangle \\
& + (-1)^{\eta+\eta'-\Gamma} \sqrt{\frac{1}{24}(\eta'-1)\eta'(\eta'+1)(\eta'+2)} \sqrt{\frac{d(\Gamma_{\omega_1})}{d(\eta 0)d(\Gamma_{\omega_1'})}} \\
& \times \sum_{\Gamma''} \sqrt{d(\Gamma'')} U[\Gamma(\eta' 0)\Gamma''(0, \eta'-2); (\eta 0)(20)] \\
& \times \sum_{\rho_3} \langle \Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2}; S_2 | | | \{a^\dagger_\eta \times \tilde{a}_{\eta'-2}\}^{\Gamma'' S} | | | \Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'}; S_1 \rangle_{\rho_3} \\
& \times \sum_{\rho_4} U[\Gamma_{\omega_2}\tilde{\Gamma}''\Gamma_{\omega_1}(20); \Gamma_{\omega_1'}\rho_{3-}\tilde{\Gamma}_-\rho_4] \langle \Gamma_{\omega_1}\alpha_{\omega_1}; \Gamma\alpha | \Gamma_{\omega_2}\alpha_{\omega_2} \rangle_{\rho_4} \langle S_1\Sigma_1 S\Sigma | S_2\Sigma_2 \rangle.
\end{aligned}
\tag{31}
$$

Note that the sum over Γ'' is restricted by coupling requirements $(\eta+2,0)\times(0\,\eta') \to \Gamma''$, $(20)\times\Gamma \to \Gamma''$, and $\Gamma_{\omega_1'}\times\Gamma'' \to \Gamma_{\omega_2}$ in the first term and $(\eta 0)\times(0,\eta'-2) \to \Gamma''$, $\Gamma\times(20) \to \Gamma''$, and $\Gamma_{\omega_1'}\times\Gamma'' \to \Gamma_{\omega_2}$ in the second term.

The matrix element of the third term in Equation (25) originates from the two-body center-of-mass correction in the fermion realization of $A^{(20)}$ [Equation (13)] and can be evaluated once the commutator,

$$[\{a^{\dagger}_{\eta}\times\tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma},\{\{a^{\dagger}_{\nu}\times a^{\dagger}_{\nu'}\}^{\Gamma_a S'}\times\{\tilde{a}_{\nu-1}\times\tilde{a}_{\nu'-1}\}^{\tilde{\Gamma}_b S'}\}^{(20)s=0}_{\beta_1\sigma=0}], \tag{32}$$

is worked out. Upon doing so and employing the Wigner–Eckart theorem, recoupling coefficients, their symmetry and orthogonalization properties, it turns out that (while the commutator is non-zero) the contribution to the matrix element of Equation (22) vanishes identically: $C_3=0$.

We can now combine Equations (23), (29), (31), apply the Wigner–Eckart theorem to the matrix element under consideration, and utilize the orthonormality of the Clebsch–Gordan and Wigner coupling coefficients to obtain the final expression, a recursion formula for triple-reduced matrix elements of the one-body unit operator $\{a^{\dagger}_{\eta}\times\tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}$:

$$\begin{aligned}
&\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||\{a^{\dagger}_{(\eta 0)(1/2)}\times\tilde{a}_{(0\eta')(1/2)}\}^{\Gamma S}|||\Gamma_{\sigma_1}\Gamma_{\hat{n}_1}\hat{\rho}_1\Gamma_{\omega_1};S_1\rangle_{\rho}\\
&=\sum_{\Gamma_{n_1}\rho_1}[\mathscr{K}^{-1}(\Gamma_{\sigma_1}\Gamma_{\omega_1})]_{\Gamma_{\hat{n}_1}\hat{\rho}_1,\Gamma_{n_1}\rho_1}\\
&\times\sum_{\Gamma_{\hat{n}_1'}\hat{\rho}_1'\Gamma_{\omega_1'}}U[(20)\Gamma_{\hat{n}_1'}\Gamma_{\omega_1}\Gamma_{\sigma_1};\Gamma_{n_1-}\rho_1\Gamma_{\omega_1'}\hat{\rho}_{1-}']X^{\Gamma_{n_1}}(\Gamma_{\hat{n}_1'})\\
&\times\sum_{\Gamma_{n_1'}\rho_1'}[\mathscr{K}(\Gamma_{\sigma_1}\Gamma_{\omega_1'})]_{\Gamma_{\hat{n}_1'}\hat{\rho}_1',\Gamma_{n_1'}\rho_1'}\\
&\times\Bigg\{(-1)^{\Gamma+\Gamma_{\omega_1}+\Gamma_{\omega_2}}\sqrt{\frac{d(\Gamma_{\omega_1})}{6d(\Gamma)}}\\
&\times\sum_{\Gamma_{n''}\rho''\Gamma_{\omega''}}\sqrt{d(\Gamma_{\omega''})}\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||A^{(20)}|||\Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''};S_2\rangle\\
&\times\sum_{\rho_3}\langle\Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''};S_2|||\{a^{\dagger}_{\eta}\times\tilde{a}_{\eta'}\}^{\Gamma S}|||\Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'};S_1\rangle_{\rho_3}\\
&\times\sum_{\rho_4\rho_5}\Phi_{\rho_3\rho_4}[\Gamma_{\omega''}\tilde{\Gamma};\Gamma_{\omega_1'}]\Phi_{\rho_4\rho_5}[\Gamma_{\omega_1'}\tilde{\Gamma}_{\omega''};\tilde{\Gamma}]U[\Gamma_{\omega_2}\tilde{\Gamma}_{\omega''}\Gamma_{\omega_1}\Gamma_{\omega_1'};(20)_{--}\tilde{\Gamma}\rho_5\rho]\\
&-\left(1-\frac{1}{A}\right)(-1)^{\Gamma}\sqrt{\frac{(\eta+1)(\eta+2)}{2d(\Gamma)}}\sqrt{\frac{d(\Gamma_{\omega_1})}{d(\Gamma_{\omega_1'})}}\\
&\times\sum_{\Gamma''}(-1)^{\Gamma''}\sqrt{d(\Gamma'')}U[(20)(\eta 0)\Gamma''(0\eta');(\eta+2,0)\Gamma]\\
&\times\sum_{\rho_3}\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||\{a^{\dagger}_{\eta+2}\times\tilde{a}_{\eta'}\}^{\Gamma''S}|||\Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'};S_1\rangle_{\rho_3}\\
&\times U[\Gamma_{\omega_2}\tilde{\Gamma}''\Gamma_{\omega_1}(20);\Gamma_{\omega_1'}\rho_{3-}\tilde{\Gamma}_{-}\rho]\\
&+\left(1-\frac{1}{A}\right)(-1)^{\eta+\eta'-\Gamma}\sqrt{\frac{(\eta'-1)\eta'(\eta'+1)(\eta'+2)}{24d(\eta 0)}}\sqrt{\frac{d(\Gamma_{\omega_1})}{d(\Gamma_{\omega_1'})}}\\
&\times\sum_{\Gamma''}\sqrt{d(\Gamma'')}U[\Gamma(\eta' 0)\Gamma''(0,\eta'-2);(\eta 0)(20)]\\
&\times\sum_{\rho_3}\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||\{a^{\dagger}_{\eta}\times\tilde{a}_{\eta'-2}\}^{\Gamma''S}|||\Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'};S_1\rangle_{\rho_3}\\
&\times U[\Gamma_{\omega_2}\tilde{\Gamma}''\Gamma_{\omega_1}(20);\Gamma_{\omega_1'}\rho_{3-}\tilde{\Gamma}_{-}\rho]\Bigg\}.
\end{aligned} \tag{33}$$

To evaluate this expression, a series of ingredients are necessary: First, numeric values for the SU(3) Racah coefficients $U[\Gamma_1\Gamma_2\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{23}\rho_{23}\rho_{1,23}]$ and $\Phi_{\rho\rho'}[\Gamma_1\Gamma_2;\Gamma_3] = Z[\Gamma_1(00)\Gamma_3\Gamma_2;\Gamma_{1-}\rho\Gamma_{2-}\rho']$ are required. These may be calculated with a computer code published by Akiyama and Draayer.[34] Second, matrix elements of the symplectic raising operator $A^{(20)}$ and of the $\mathscr{K}$-matrix, which effects the orthonormalization of the symplectic basis states, are needed. Both types can be obtained from the matrix elements of $\mathscr{A}^{(20)}$, a generator of the $u(3)\otimes Weyl$ algebra, as is outlined in Reference 7. Third, matrix elements of the form $\langle\Gamma_{\sigma_2}\Gamma_{n''}\rho''\Gamma_{\omega''};S_2|||\{a^\dagger_\eta\times\tilde{a}_{\eta'}\}^{\Gamma S}|||\Gamma_{\sigma_1}\Gamma_{n_1'}\rho_1'\Gamma_{\omega_1'};S_1\rangle_{\rho_3}$ between states of excitation $N_{n''}$ and $N_{n_1'}$ are requisite ingredients. Here $N_{n''}=N_{n_2}-2$ as a result of the delta function $\delta_{N_{n''},N_{n_2}-2}$, introduced through the symplectic generator $A^{(20)}$ [see also Equation (28)], and $N_{n_1'}=N_{n''}-(\eta-\eta') = N_{\hat{n}_1}-2$. Furthermore, one needs values for the matrix elements of the operators $\{a^\dagger_{\eta+2}\times\tilde{a}_{\eta'}\}^{\Gamma''S}_{\alpha''\Sigma}$ and $\{a^\dagger_\eta\times\tilde{a}_{\eta'-2}\}^{\Gamma''S}_{\alpha''\Sigma}$ between the original bra state of excitation N_{n_2} and ket states with $N_{n_1'}=N_{n_2}-(\eta-\eta'+2)=N_{\hat{n}_1}-2$. Thus the desired matrix element is expressed in terms of known quantities and unit matrix elements involving states of lower excitation. Hence, through repeated application of this recursive process matrix elements of arbitrary one-body operators between symplectic basis states may be related back to valence shell matrix elements, which in turn can be evaluated by means of standard shell model techniques. A user-friendly computer code, which calculates the latter, has been published by Bahri and Draayer.[25]

The above formula has been derived for fermionic unit operators of the form $\{a^\dagger_\eta\times\tilde{a}_{\eta'}\}^{\Gamma S}_{\alpha\Sigma}$ with $\eta\geqslant\eta'$. The recursive process also covers the case $\eta>\eta'$, since the following relation holds:

$$\begin{aligned}\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||\{a^\dagger_\eta\times\tilde{a}_{\eta'}\}^{\Gamma S}|||\Gamma_{\sigma_1}\Gamma_{n_1}\rho_1\Gamma_{\omega_1};S_1\rangle_\rho&\\ =(-1)^{\eta+\eta'-\Gamma_{\omega_1}+\Gamma_{\omega_2}+S_1+S_2}\sqrt{\frac{d(\Gamma_{\omega_1})}{d(\Gamma_{\omega_2})}}\ \sqrt{\frac{2S_1+1}{2S_2+1}}&\\ \times\langle\Gamma_{\sigma_1}\Gamma_{n_1}\rho_1\Gamma_{\omega_1};S_1|||\{a^\dagger_{\eta'}\times\tilde{a}_\eta\}^{\tilde{\Gamma}S}|||\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2\rangle^*_\rho\,.&\end{aligned}\tag{34}$$

A stringent test of Equation (33) is given by the following: One can evaluate the matrix element

$$\langle\Gamma_{\sigma_2}\Gamma_{n_2}\rho_2\Gamma_{\omega_2};S_2|||\{a^\dagger_\eta\times\tilde{a}_{\eta'}\}^{\Gamma S}|||\Gamma_{\sigma_1}\Gamma_{n_1}\rho_1\Gamma_{\omega_1};S_1\rangle_\rho\tag{35}$$

by stepping down on the bra-side, and proceeding analogously to the derivation given above. The result is a recursion formula analogous to Equation (33). Alternatively, using relation (34) in conjunction with Equation (33) yields an expression for the matrix element of Equation (35) which exactly equals the formula that is obtained by stepping down on the ket-side. Another test was carried out by encoding the recursion formula and using the one-body matrix elements so obtained to calculate expectation values of the particle number operator $\hat{N}$ and the symplectic raising and lowering operators $A^{(20)}$ and $B^{(02)}$, as well as the expectation values of $C^{(11)}$ for the valence shell. For $A\to\infty$, perfect agreement with the matrix elements of $A^{(20)}$ and $B^{(02)}$ as evaluated by means of a code that is based on the vector coherent state method outlined in reference 7 was obtained, as well as agreement with the matrix elements of $\hat{N}$ and $C^{(11)}$, which can be evaluated analytically.

VI. CONCLUDING REMARKS

We have introduced a fermion realization of the Sp(6,R) algebra. Specifically, we expressed the symplectic generators in terms of fermion creation and annihilation operators in a $\mathrm{Sp}(6,R)\supset\mathrm{U}(3)\supset\mathrm{SO}(3)$ basis. The new formalism was employed to derive a recursive process for calculating matrix elements of arbitrary one-body operators. The resulting formula requires as input SU(3) coupling coefficients, matrix elements of symplectic generators, and valence shell matrix elements of one-body unit operators. These ingredients may be calculated using published computer codes; references for those were given. The new formalism allows for the evaluation of

physical observables in the nuclear Sp(6,R) model which until now were only available, if at all, as an approximation to the exact result. First applications include symplectic calculations of longitudinal and transverse nuclear form factors,[35] which can be compared to those extracted from experiments.

One can also employ the recursion formula of Equation (33) to derive a recursion relation for $A^{(20)}$ for the special case of a large number of nucleons in the system ($A\rightarrow\infty$). Upon doing so, and comparing the result to the formula that was derived by Rosensteel using a step-down procedure analogous to the one outlined above,[27] we find that exact agreement requires the following relations to hold:

$$\begin{aligned}&(\eta+3)(\eta+4)U[(02)(\eta+2,0)(22)(0\eta);(\eta0)(20)]\\&\quad=(\eta-1)\eta U[(20)(\eta-2,0)(22)(0\eta);(\eta0)(02)],\end{aligned}\tag{36}$$

$$\begin{aligned}&(\eta+3)(\eta+4)U[(02)(\eta+2,0)(11)(0\eta);(\eta0)(20)]\\&\quad-(\eta-1)\eta U[(20)(\eta-2,0)(11)(0\eta);(\eta0)(02)]=\sqrt{10\eta(\eta+3)},\end{aligned}\tag{37}$$

$$U[(02)(\eta+2,0)(00)(0\eta);(\eta0)(20)]=U[(20)(\eta-2,0)(00)(0\eta);(\eta0)(02)]=1.\tag{38}$$

While Equation (38) was known previously,[36] the other two relations, Equation (36) and Equation (37), are new and may prove valuable for analytic work that involves SU(3) Racah coefficients.

ACKNOWLEDGMENTS

The authors wish to thank George Rosensteel for valuable input and for providing them with a code which evaluates matrix elements of symplectic raising and lowering operators. They furthermore acknowledge inspiring discussions on the subject with Jorge Hirsch and Octavio Castaños. One of the authors (J.E.) would also like to thank Amiram Leviatan and the Racah Institute of Physics, Jerusalem, for support during the final write-up of this paper.

APPENDIX A: BOSON CREATION AND ANNIHILATION OPERATORS

Given position $x_i(s)$ and momentum $p_i(s)$ coordinates of the s-th particle ($i=1,2,3$ and $s=1,\ldots,A$), one-body operators $b_i^\dagger(s)$ and $b_i(s)$, which create and annihilate, respectively, one oscillator quantum in the i-th direction of the s-th particle, can be defined as follows:

$$\begin{aligned}b_i^\dagger(s)&\equiv\frac{1}{\sqrt{2}}(x_i(s)-ip_i(s)),\\b_i(s)&\equiv\frac{1}{\sqrt{2}}(x_i(s)+ip_i(s)).\end{aligned}\tag{A1}$$

These operators are related to each other by Hermitian conjugation $b_i(s)=(b_i^\dagger(s))^\dagger$ and satisfy the standard boson commutation relations:

$$\begin{aligned}[b_i(s),b_j^\dagger(t)]&=\delta_{st}\delta_{ij},\\[b_i^\dagger(s),b_j^\dagger(t)]&=[b_i(s),b_j(t)]=0.\end{aligned}\tag{A2}$$

They may also be viewed as the components of SU(3) irreducible tensor operators, $b_{1q}^{\dagger(10)}$ and $\tilde{b}_{1q}^{(01)}$ ($q=0,\pm1$), transforming according to the $(\lambda\mu)=(10)$ and $(\lambda\mu)=(01)$ SU(3) irreps, respectively:

$$b_{1,\pm1}^{\dagger(10)}\equiv\mp\frac{1}{\sqrt{2}}(b_1^\dagger\pm ib_2^\dagger),\quad b_{1,0}^{\dagger(10)}\equiv b_3^\dagger,\tag{A3}$$

and

$$\tilde{b}^{(01)}_{1,\pm 1} \equiv \mp \frac{1}{\sqrt{2}}(b_1 \pm i b_2), \quad \tilde{b}^{(01)}_{1,0} \equiv b_3, \tag{A4}$$

where the subscript $(1,q)$ denotes the angular momentum character $(l=1;\ q=0,\pm 1)$ of the spherical components of the boson creation and annihilation operators. For the sake of simplicity the angular momentum label $l=1$ and the particle index s will be suppressed from here on, except where needed to avoid ambiguities. The operators $b^{\dagger(10)}_q$ and $\tilde{b}^{(01)}_q$ satisfy the conjugation relation $\tilde{b}^{(01)}_q = (-1)^q (b^{\dagger(10)}_{-q})^\dagger$, and their commutator is given by $[\tilde{b}^{(01)}_q, b^{\dagger(10)}_{q'}] = (-1)^q \delta_{q(-q')}$. Matrix elements of the boson creation and annihilation operators take a very simple form in the standard SU(3) basis:

$$\begin{aligned} \langle (\eta 0) S \| b^{\dagger(10)s=0} \| (\lambda \mu) S' \rangle &= \langle (\lambda+1, \mu) S \| b^{\dagger(10)s=0} \| (\eta - 1, 0) S' \rangle = \sqrt{\eta}\, \delta_{(\lambda\mu)(\eta-1,0)} \delta_{SS'}, \\ \langle (\lambda \mu) S \| \tilde{b}^{(01)s=0} \| (\eta 0) S' \rangle &= \langle (\eta - 1, 0) S \| \tilde{b}^{(01)s=0}_s \| (\lambda+1, \mu) S' \rangle = \sqrt{\eta+2}\, \delta_{(\lambda\mu)(\eta-1,0)} \delta_{SS'}. \end{aligned} \tag{A5}$$

The $b^{\dagger(10)}_q$ and $\tilde{b}^{(01)}_q$ are scalars in spin space, as has been noted explicitly in Equation (A5). [For the definition of the triple-reduced matrix element and the relevant phase conventions refer to Appendix C.]

Since the creation and annihilation operators $b^{\dagger(10)}_q$ and $\tilde{b}^{(01)}_q$ are SU(3) irreducible tensor operators, two or more of them may be coupled to form new SU(3) tensors. Matrix elements of such tensor products can be evaluated with the help of SU(3) reduction rules as given in Appendix C. A relevant example is the matrix element of the one-body operator $\{b^{\dagger(10)s=0} \times \tilde{b}^{(01)s=0}\}^{(11)S=0}$, evaluated between two single-particle states,

$$\begin{aligned} &\langle (\nu 0)\, \tfrac{1}{2} \| \{ b^{\dagger(10)s=0} \times \tilde{b}^{(01)s=0} \}^{(11)S=0} \| (\nu' 0)\, \tfrac{1}{2} \rangle \\ &\quad = \sum_{(\lambda''\mu'')S''} \Phi_{11}[(10)(01);(11)]\, U[(\nu' 0)(01)(\nu 0)(10);(\lambda''\mu'')(11)]\, U\left(\frac{1}{2}\, 0\, \frac{1}{2}\, 0; S'' 0\right) \\ &\qquad \times \langle (\nu 0)\, \tfrac{1}{2} \| b^{\dagger(10)s=0} \| (\lambda''\mu'') S'' \rangle \langle (\lambda''\mu'') S'' \| \tilde{b}^{(01)s=0} \| (\nu' 0)\, \tfrac{1}{2} \rangle \\ &\quad = \sqrt{\frac{2}{3}\nu(\nu+3)}\, \delta_{\nu\nu'}, \end{aligned} \tag{A6}$$

where $U[\cdots]$ and $\Phi[\cdots]$ are Racah recoupling coefficients (see Appendix C). For the tensor product of two raising (lowering) operators, acting on a single-particle state, one may proceed analogously to obtain

$$\begin{aligned} \langle (\nu 0)\, \tfrac{1}{2} \| \{ b^{\dagger(10)s=0} \times b^{\dagger(10)s=0} \}^{(20)S=0} \| (\nu' 0)\, \tfrac{1}{2} \rangle &= \sqrt{\nu(\nu-1)}\, \delta_{\nu,\nu'+2}, \\ \langle (\nu 0)\, \tfrac{1}{2} \| \{ \tilde{b}^{(01)s=0} \times \tilde{b}^{(01)s=0} \}^{(02)S=0} \| (\nu' 0)\, \tfrac{1}{2} \rangle &= \sqrt{(\nu+3)(\nu+4)}\, \delta_{\nu',\nu+2}. \end{aligned} \tag{A7}$$

Matrix elements of a SU(3)-coupled tensor product of two boson operators, acting on a two-particle state, can be evaluated using Millener's reduction rule[37] (see also Appendix C). For the product of a creation and an annihilation operator, coupled to $(\lambda\mu)=(11)$, this method yields

$$\begin{aligned} &\langle \{ (\nu_1 0) \times (\nu_2 0) \} (\lambda\mu); \{ \tfrac{1}{2} \times \tfrac{1}{2} \} S \| \{ b^{\dagger(10)0}(1) \times \tilde{b}^{(01)0}(2) \}^{(11)0} \| \\ &\quad \times \{ (\nu_1' 0) \times (\nu_2' 0) \} (\lambda'\mu'); \{ \tfrac{1}{2} \times \tfrac{1}{2} \} S' \rangle \\ &\quad = \sqrt{\nu_1(\nu_2+3)}\, \delta_{SS'} \delta_{\nu_1', \nu_1 - 1} \delta_{\nu_2', \nu_2 + 1} \begin{Bmatrix} (\nu_1 - 1, 0) & (10) & (\nu_1 0) & - \\ (\nu_2 + 1, 0) & (01) & (\nu_2 0) & - \\ (\lambda'\mu') & (11) & (\lambda\mu) & - \\ - & - & - & \end{Bmatrix}, \end{aligned} \tag{A8}$$

whereas for two creation operators, coupled to $(\lambda\mu)=(20)$, we obtain

$$\begin{aligned}&\langle\{(\nu_1 0)\times(\nu_2 0)\}(\lambda\mu);\{\tfrac{1}{2}\times\tfrac{1}{2}\}S|||\{b^{\dagger(10)0}(1)\times b^{\dagger(10)0}(2)\}^{(20)0}|||\\&\quad\times\{(\nu_1' 0)\times(\nu_2' 0)\}(\lambda'\mu');\{\tfrac{1}{2}\times\tfrac{1}{2}\}S'\rangle\\&\quad=\sqrt{\nu_1\nu_2}\,\delta_{SS'}\delta_{\nu_1',\nu_1-1}\delta_{\nu_2',\nu_2-1}\begin{Bmatrix}(\nu_1-1,0) & (10) & (\nu_1 0) & -\\(\nu_2-1,0) & (10) & (\nu_2 0) & -\\(\lambda'\mu') & (20) & (\lambda\mu) & -\\- & - & - & \end{Bmatrix},\end{aligned}\tag{A9}$$

and for two annihilation operators we find

$$\begin{aligned}&\langle\{(\nu_1 0)\times(\nu_2 0)\}(\lambda\mu);\{\tfrac{1}{2}\times\tfrac{1}{2}\}S|||\{\tilde{b}^{(01)0}(1)\times\tilde{b}^{(01)0}(2)\}^{(02)0}|||\{(\nu_1' 0)\times(\nu_2' 0)\}(\lambda'\mu');\{\tfrac{1}{2}\times\tfrac{1}{2}\}S'\rangle\\&=\sqrt{(\nu_1+3)(\nu_2+3)}\,\delta_{SS'}\delta_{\nu_1',\nu_1+1}\delta_{\nu_2',\nu_2+1}\begin{Bmatrix}(\nu_1+1,0) & (01) & (\nu_1 0) & -\\(\nu_2+1,0) & (01) & (\nu_2 0) & -\\(\lambda'\mu') & (02) & (\lambda\mu) & -\\- & - & - & \end{Bmatrix},\end{aligned}\tag{A10}$$

where $b^{\dagger}(t)$ and $\tilde{b}(t)$ ($t=1,2$ here) act only on that part of the wave function which refers to particle t.

APPENDIX B: FERMION SECOND QUANTIZATION

A general one-body operator that acts symmetrically on a system of A identical particles is given by

$$\mathscr{F}=\sum_s f(\mathbf{r}_s,\boldsymbol{\sigma}_s),\tag{B1}$$

where $\mathbf{r}_s$ and $\boldsymbol{\sigma}_s$ represent the position and spin (or spin–isospin) coordinates, respectively, of the s-th particle. In a fermion second quantization formulation this one-body operator takes the form

$$\mathscr{F}=\sum_{\rho,\rho'}\langle\rho'|f(\mathbf{r},\boldsymbol{\sigma})|\rho\rangle a^{\dagger}_{\rho'}a_{\rho},\tag{B2}$$

where ρ labels the available single-particle states and $a^{\dagger}_{\rho}$ and a_{ρ} are single-particle creation and annihilation operators that satisfy the usual fermion anti-commutation relations:

$$\begin{aligned}\{a_{\rho},a^{\dagger}_{\rho'}\}&=\delta_{\rho,\rho'},\\\{a_{\rho},a_{\rho'}\}&=\{a^{\dagger}_{\rho},a^{\dagger}_{\rho'}\}=0.\end{aligned}\tag{B3}$$

For fermions in a harmonic oscillator potential, ρ stands for a set of quantum numbers $\rho=\eta l m\tfrac{1}{2}\sigma$ or $\rho=\eta l\tfrac{1}{2}j m_j$, depending on whether the states are characterized by an LS- or jj-coupling scheme, respectively. Here η is the principal quantum number (major oscillator shell) of the single-particle level; l, $\tfrac{1}{2}$, and j label the orbital, spin, and total angular momenta with projections m, σ, and m_j, respectively. (In a spin–isospin formalism, one has $\rho=\eta l m\tfrac{1}{2}\sigma\tfrac{1}{2}\tau$ or $\rho=\eta l\tfrac{1}{2}j m_j\tfrac{1}{2}\tau$, respectively, where the additional $\tfrac{1}{2}$ denotes the isospin quantum number with projection τ.) For the present purposes it is most convenient to use the LS-coupling scheme.

Since the single-particle harmonic oscillator wave functions, $|\eta l m\tfrac{1}{2}\sigma\rangle=a^{\dagger}_{\eta l m(1/2)\sigma}|_-\rangle$, where $|_-\rangle$ denotes the particle vacuum, transform irreducibly under a set of physically relevant SU(3) and SU(2) symmetry group operations, the fermion creation operator $a^{\dagger}_{\eta l m(1/2)\sigma}$ is a double irreducible tensor operator of rank $(\lambda\mu)=(\eta 0)$ in SU(3), which labels its orbital character (with subgroup labels l and m), and of rank $s=\tfrac{1}{2}$ in SU(2) for the spin part (with subgroup label σ) and

should be written as $a^{\dagger}_{(\eta 0)lm(1/2)\sigma}$. Since $a_{\rho}=(a^{\dagger}_{\rho})^{\dagger}$ is not a proper irreducible tensor operator with respect to the above group transformations, it is advantageous to define $\tilde{a}_{(0\eta)lm(1/2)\sigma} \equiv(-1)^{\eta+l+m+(1/2)+\sigma} a_{(\eta 0)l(-m)(1/2)(-\sigma)}$ which is a proper irreducible tensor operator of rank $(\lambda\mu)=(0\eta)$ in SU(3) and rank $s=\frac{1}{2}$ in (spin-)SU(2).

It thus becomes possible to construct tensor products from the fermion creation and annihilation operators, such as

$$\{a^{\dagger}_{(\eta_1 0)(1/2)}\times\tilde{a}_{(0\eta_2)(1/2)}\}^{(\lambda\mu)S}_{\kappa LM\Sigma}=\sum_{\alpha_1\alpha_2\sigma_1\sigma_2}\langle(\eta_1 0)\alpha_1;(0\eta_2)\alpha_2|(\lambda\mu)\kappa LM\rangle \times\langle\tfrac{1}{2}\sigma_1\,\tfrac{1}{2}\sigma_2|S\Sigma\rangle a^{\dagger}_{(\eta_1 0)\alpha_1(1/2)\sigma_1}\tilde{a}_{(0\eta_2)\alpha_2(1/2)\sigma_2}, \tag{B4}$$

which moves a particle from the η_2-th major oscillator shell to the η_1-th shell. The possible $(\lambda\mu)$ values are given by the coupling rule $(\eta_1 0)\times(0\eta_2)=\oplus_{k=0}^{\min(\eta_1,\eta_2)}(\eta_1-k,\eta_2-k)$ (see reference 38) and α_i is an abbreviation for the set of possible subgroup labels $\alpha_i=\kappa_i l_i m_i$, where $\kappa_i=1$ must hold here. The total intrinsic spin S can take the values 0 or 1 with projection $\Sigma=0$ or $\Sigma=0$, ±1, respectively. The product $\{a^{\dagger}_{(\eta_1 0)(1/2)}\times a^{\dagger}_{(\eta_2 0)(1/2)}\}^{(\lambda\mu)S}_{\kappa LM\Sigma}$, on the other hand, creates a pair of fermions with one particle in the η_1-th shell and one in the η_2-th shell, and the pair is coupled to $(\lambda\mu)\in\{(\eta_1 0)\times(\eta_2 0)\}=\{\oplus_{k=0}^{\min(\eta_1,\eta_2)}(\eta_1+\eta_2-2k,k)\}$ and $S=0$ or 1. Similarly, the product $\{\tilde{a}_{(0\eta_1)(1/2)}\times\tilde{a}_{(0\eta_2)(1/2)}\}^{(\lambda\mu)S}_{\kappa LM\Sigma}$ annihilates a SU(3)-coupled pair of fermions with one particle in the η_1-th shell, one in the η_2-th shell, $(\lambda\mu)\in\{(0\eta_1)\times(0\eta_2)\}=\{\oplus_{k=0}^{\min(\eta_1,\eta_2)}(k,\eta_1+\eta_2-2k)\}$, and $S=0$ or 1. One can furthermore construct a SU(3) irreducible tensor which destroys a pair of fermions in a particular SU(3)-coupled configuration, and creates a new pair configuration:

$$\begin{aligned}&\{\{a^{\dagger}_{(\eta_1 0)(1/2)}\times a^{\dagger}_{(\eta_2 0)(1/2)}\}^{(\lambda_1\mu_1)S_1}\times\{\tilde{a}_{(0\nu_1)(1/2)}\times\tilde{a}_{(0\nu_2)(1/2)}\}^{(\lambda_2\mu_2)S_2}\}^{\rho(\lambda\mu)S}_{\kappa LM\Sigma}\\ &\quad=\sum_{\alpha_1\alpha_2\Sigma_1\Sigma_2}\langle(\lambda_1\mu_1)\alpha_1;(\lambda_2\mu_2)\alpha_2|(\lambda\mu)\alpha\rangle_{\rho}\langle S_1\Sigma_1 S_2\Sigma_2|S\Sigma\rangle\\ &\qquad\times\{a^{\dagger}_{(\eta_1 0)(1/2)}\times a^{\dagger}_{(\eta_2 0)(1/2)}\}^{(\lambda_1\mu_1)S_1}_{\alpha_1\Sigma_1}\{\tilde{a}_{(0\nu_1)(1/2)}\times\tilde{a}_{(0\nu_2)(1/2)}\}^{(\lambda_2\mu_2)S_2}_{\alpha_2\Sigma_2},\end{aligned} \tag{B5}$$

where ρ denotes the multiplicity of $(\lambda\mu)$ in the coupling $(\lambda_1\mu_1)\times(\lambda_2\mu_2)\to(\lambda\mu)$.

The Hermitian adjoint of the above products are given by

$$\begin{aligned}(\{a^{\dagger}_{(\eta_1 0)(1/2)}\times\tilde{a}_{(0\eta_2)(1/2)}\}^{(\lambda\mu)S}_{\kappa LM\Sigma})^{\dagger}&=(-1)^{\eta_2-\eta_1+L+M+\Sigma}\{a^{\dagger}_{(\eta_2 0)(1/2)}\times\tilde{a}_{(0\eta_1)(1/2)}\}^{(\mu\lambda)S}_{\kappa L-M-\Sigma},\\ (\{a^{\dagger}_{(\eta_1 0)(1/2)}\times a^{\dagger}_{(\eta_2 0)(1/2)}\}^{(\lambda\mu)S}_{\kappa LM\Sigma})^{\dagger}&=(-1)^{-\eta_1-\eta_2-L+M-1+\Sigma}\{\tilde{a}_{(0\eta_2)(1/2)}\times\tilde{a}_{(0\eta_1)(1/2)}\}^{(\mu\lambda)S}_{\kappa L-M-\Sigma},\end{aligned} \tag{B6}$$

and

$$\begin{aligned}&(\{\{a^{\dagger}_{(\eta_1 0)(1/2)}\times a^{\dagger}_{(\eta_2 0)(1/2)}\}^{(\lambda_1\mu_1)S_1}\times\{\tilde{a}_{(0\nu_1)(1/2)}\times\tilde{a}_{(0\nu_2)(1/2)}\}^{(\lambda_2\mu_2)S_2}\}^{\rho(\lambda\mu)S}_{\kappa LM\Sigma})^{\dagger}\\ &\quad=(-1)^{\nu_1+\nu_2-\eta_1-\eta_2+(\lambda_1+\mu_1)+(\lambda_2+\mu_2)+(\lambda+\mu)-L+M+\Sigma}\\ &\qquad\times\sum_{\rho'}(-1)^{\rho'_{\max}-\rho'}\Phi_{\rho'\rho}[(\lambda_2\mu_2)(\lambda_1\mu_1);(\lambda\mu)]\\ &\qquad\times\{\{a^{\dagger}_{(\nu_2 0)(1/2)}\times a^{\dagger}_{(\nu_1 0)(1/2)}\}^{(\mu_2\lambda_2)S_2}\times\{\tilde{a}_{(0\eta_2)(1/2)}\times\tilde{a}_{(0\eta_1)(1/2)}\}^{(\mu_1\lambda_1)S_1}\}^{\rho'(\mu\lambda)S}_{\kappa L-M-\Sigma},\end{aligned} \tag{B7}$$

where the phase matrix Φ is a special case of the recoupling coefficient Z, $\Phi_{\rho\rho'}[(\lambda_1\mu_1)\times(\lambda_2\mu_2);(\lambda\mu)]=Z[(\lambda_1\mu_1)(00)(\lambda\mu)(\lambda_2\mu_2);(\lambda_1\mu_1)_-\rho(\lambda_2\mu_2)_-\rho']$, and can be evaluated with available computer codes (see also Appendix C). The symbol $\rho'_{\max}$ denotes the maximum multiplicity in the coupling $(\mu_2\lambda_2)\times(\mu_1\lambda_1)\to(\mu\lambda)$, and ρ' takes on values $1,2,\ldots,\rho'_{\max}$.

The following commutation relations, which are required for the derivation of the fermionic expressions for the symplectic generators, can be derived by making use of the commutation rules of the uncoupled components [see Equation (B3)] and SU(3) coupling and recoupling techniques. For clarity we abbreviate $(\lambda_i \mu_i)$ with the symbol Γ_i, $(-1)^{\Gamma_i} \equiv (-1)^{\lambda_i+\mu_i}$, and $\hat{S} \equiv \sqrt{2S+1}$:

$$
\begin{aligned}
&[\{a^\dagger_\eta \times \tilde{a}_{\eta'}\}^{\Gamma_1 S_1}_{\alpha_1 \Sigma_1}, \{a^\dagger_\nu \times \tilde{a}_{\nu'}\}^{\Gamma_2 S_2}_{\alpha_2 \Sigma_2}] \\
&\quad = (-1)^{\eta+\nu} \sum_{s\sigma} \hat{S}_1 \hat{S}_2 \langle S_1 \Sigma_1 S_2 \Sigma_2 | s\sigma \rangle \begin{Bmatrix} \frac{1}{2} & \frac{1}{2} & S_1 \\ s & S_2 & \frac{1}{2} \end{Bmatrix} \\
&\qquad \times \Bigg\{ -\delta_{\eta'\nu} (-1)^{-s-\Gamma_1} \sqrt{\frac{d(\Gamma_1)}{d(\eta 0)}} \sum_{\Gamma\alpha} \{a^\dagger_\eta \times \tilde{a}_{\nu'}\}^{\Gamma s}_{\alpha\sigma} \\
&\qquad \times \sum_\rho \langle \Gamma_1 \alpha_1; \Gamma_2 \alpha_2 | \Gamma \alpha \rangle_\rho U[\Gamma_1(\nu 0)\Gamma(0\nu'); (\eta 0)_{--}\Gamma_{2-}\rho] \\
&\qquad + \delta_{\eta\nu'} (-1)^{S_1+S_2-\Gamma_2} \sqrt{\frac{d(\Gamma_2)}{d(\nu 0)}} \sum_{\Gamma\alpha} \{a^\dagger_\nu \times \tilde{a}_{\eta'}\}^{\Gamma s}_{\alpha\sigma} \\
&\qquad \times \sum_\rho \langle \Gamma_2 \alpha_2; \Gamma_1 \alpha_1 | \Gamma \alpha \rangle_\rho U[\Gamma_2(\eta 0)\Gamma(0\eta'); (\nu 0)_{--}\Gamma_{1-}\rho] \Bigg\},
\end{aligned} \tag{B8}
$$

and

$$
\begin{aligned}
&[\{a^\dagger_\eta \times a^\dagger_{\eta'}\}^{\Gamma_1 S_1}_{\alpha_1 \Sigma_1}, \{a^\dagger_\nu \times \tilde{a}_{\nu'}\}^{\Gamma_2 S_2}_{\alpha_2 \Sigma_2}] \\
&\quad = (-1)^{\eta+\nu-\Gamma_2} \sqrt{\frac{d(\Gamma_2)}{d(\nu 0)}} \sum_{s\sigma} \hat{S}_1 \hat{S}_2 \langle S_2 \Sigma_2 S_1 \Sigma_1 | s\sigma \rangle \begin{Bmatrix} \frac{1}{2} & \frac{1}{2} & S_2 \\ s & S_1 & \frac{1}{2} \end{Bmatrix} \\
&\qquad \times \Bigg\{ \delta_{\nu'\eta'} (-1)^{s+S_1-\Gamma_1} \sum_{\Gamma\alpha} \{a^\dagger_\nu \times a^\dagger_\eta\}^{\Gamma s}_{\alpha\sigma} \\
&\qquad \times \sum_\rho \langle \Gamma_2 \alpha_2; \Gamma_1 \alpha_1 | \Gamma \alpha \rangle_\rho U[\Gamma_2(\eta' 0)\Gamma(\eta 0); (\nu 0)_{--}\Gamma_{1-}\rho] \\
&\qquad + \delta_{\nu'\eta} (-1)^s \sum_{\Gamma\alpha} \{a^\dagger_\nu \times a^\dagger_{\eta'}\}^{\Gamma s}_{\alpha\sigma} \sum_\rho \langle \Gamma_2 \alpha_2; \Gamma_1 \alpha_1 | \Gamma \alpha \rangle_\rho U[\Gamma_2(\eta 0)\Gamma(\eta' 0); (\nu 0)_{--}\Gamma_{1-}\rho] \Bigg\},
\end{aligned} \tag{B9}
$$

and

$$
\begin{aligned}
&[\{\tilde{a}_\eta \times \tilde{a}_{\eta'}\}^{\Gamma_1 S_1}_{\alpha_1 \Sigma_1}, \{a^\dagger_\nu \times \tilde{a}_{\nu'}\}^{\Gamma_2 S_2}_{\alpha_2 \Sigma_2}] \\
&\quad = -\sum_{s\sigma} \hat{S}_1 \hat{S}_2 \langle S_1 \Sigma_1 S_2 \Sigma_2 | s\sigma \rangle \begin{Bmatrix} \frac{1}{2} & \frac{1}{2} & S_1 \\ s & S_2 & \frac{1}{2} \end{Bmatrix} \\
&\qquad \times \Bigg\{ \delta_{\eta'\nu} (-1)^{\eta'+\eta-\Gamma_1+s} \sqrt{\frac{d(\Gamma_1)}{d(0\eta)}} \sum_{\Gamma\alpha} \{\tilde{a}_\eta \times \tilde{a}_{\nu'}\}^{\Gamma s}_{\alpha\sigma} \\
&\qquad \times \sum_\rho \langle \Gamma_1 \alpha_1; \Gamma_2 \alpha_2 | \Gamma \alpha \rangle_\rho U[\Gamma_1(\eta' 0)\Gamma(0\nu'); (0\eta)_{--}\Gamma_{2-}\rho] \\
&\qquad + \delta_{\eta\nu} (-1)^{S_1-s} \sqrt{\frac{d(\Gamma_1)}{d(0\eta')}} \sum_{\Gamma\alpha} \{\tilde{a}_{\eta'} \times \tilde{a}_{\nu'}\}^{\Gamma s}_{\alpha\sigma} \\
&\qquad \times \sum_\rho \langle \Gamma_1 \alpha_1; \Gamma_2 \alpha_2 | \Gamma \alpha \rangle_\rho U[\Gamma_1(\eta 0)\Gamma(0\nu'); (0\eta')_{--}\Gamma_{2-}\rho] \Bigg\},
\end{aligned} \tag{B10}
$$

where $d(\Gamma)=d(\lambda\mu)=\frac{1}{2}(\lambda+1)(\mu+1)(\lambda+\mu+2)$ is the dimension of the SU(3) irrep $\Gamma=(\lambda\mu)$.

In order to derive the recursion relation (33), it is furthermore necessary to evaluate the commutator of the two-body operator $\{\{a^\dagger_\eta\times a^\dagger_{\eta'}\}^{\Gamma_1 S_1}\times\{\tilde{a}_\tau\times\tilde{a}_{\tau'}\}^{\Gamma_2 S_2}\}^{\rho_3\Gamma_3 S_3}_{\alpha_3\Sigma_3}$ with $\{a^\dagger_\nu\times\tilde{a}_{\nu'}\}^{\Gamma_4 S_4}_{\alpha_4\Sigma_4}$, where ρ_3 denotes the multiplicity of Γ_3 in the coupling $\Gamma_1\times\Gamma_2\to\Gamma_3$. We obtain the following expression:

$$
\begin{aligned}
&[\{\{a^\dagger_\eta\times a^\dagger_{\eta'}\}^{\Gamma_1 S_1}\times\{\tilde{a}_\tau\times\tilde{a}_{\tau'}\}^{\Gamma_2 S_2}\}^{\rho_3\Gamma_3 S_3}_{\alpha_3\Sigma_3},\{a^\dagger_\nu\times\tilde{a}_{\nu'}\}^{\Gamma_4 S_4}_{\alpha_4\Sigma_4}]\\
&\quad=-\sum_{\Gamma\rho}\sum_{\rho_5\Gamma_5\alpha_5}(-1)^{\Gamma_1-\Gamma_3+\Gamma_4-\Gamma}\sqrt{\frac{d(\Gamma)d(\Gamma_3)}{d(\Gamma_1)d(\Gamma_4)}}\sum_{\rho'\rho''}\Phi_{\rho\rho'}[\Gamma_2\Gamma_4;\Gamma]\Phi_{\rho'\rho''}[\Gamma\tilde{\Gamma}_2;\Gamma_4]\\
&\qquad\times\sum_{\tilde{\rho}}\langle\Gamma_3\alpha_3;\Gamma_4\alpha_4|\Gamma_5\alpha_5\rangle_{\tilde{\rho}}U[\Gamma_3\tilde{\Gamma}_2\Gamma_5\Gamma;\Gamma_1\rho_3\rho_5\Gamma_4\rho''\tilde{\rho}]\\
&\qquad\times\sum_{sS_5\Sigma_5}(-1)^{S_1+S_4-S_5+s}\hat{S}_2\hat{S}_3\hat{S}_4\hat{s}\langle S_3\Sigma_3 S_4\Sigma_4|S_5\Sigma_5\rangle\begin{Bmatrix}\frac{1}{2}&\frac{1}{2}&S_2\\ s&S_4&\frac{1}{2}\end{Bmatrix}\begin{Bmatrix}S_2&S_1&S_3\\ S_5&S_4&s\end{Bmatrix}\\
&\qquad\times\Bigg\{\delta_{\tau'\nu}(-1)^{\tau'+\tau-\Gamma_2+S_2}\sqrt{\frac{d(\Gamma_2)}{d(0\tau)}}\{\{a^\dagger_\eta\times a^\dagger_{\eta'}\}^{\Gamma_1 S_1}\times\{\tilde{a}_\tau\times\tilde{a}_{\nu'}\}^{\Gamma s}\}^{\rho_5\Gamma_5 S_5}_{\alpha_5\Sigma_5}\\
&\qquad\times U[\Gamma_2(\tau'0)\Gamma(0\nu');(0\tau)_{--}\Gamma_{4-}\rho]\\
&\qquad+\delta_{\tau\nu}\sqrt{\frac{d(\Gamma_2)}{d(0\tau')}}\{\{a^\dagger_\eta\times a^\dagger_{\eta'}\}^{\Gamma_1 S_1}\times\{\tilde{a}_{\tau'}\times\tilde{a}_{\nu'}\}^{\Gamma s}\}^{\rho_5\Gamma_5 S_5}_{\alpha_5\Sigma_5}U[\Gamma_2(\tau0)\Gamma(0\nu');(0\tau')_{--}\Gamma_{4-}\rho]\Bigg\}\\
&\qquad+\sum_{\Gamma\rho}\sum_{\rho_5\Gamma_5\alpha_5}(-1)^{\eta+\nu-\Gamma_4}\sqrt{\frac{d(\Gamma_4)}{d(\nu0)}}\sum_{\tilde{\rho}}\langle\Gamma_4\alpha_4;\Gamma_3\alpha_3|\Gamma_5\alpha_5\rangle_{\tilde{\rho}}U[\Gamma_4\Gamma_1\Gamma_5\Gamma_2;\Gamma\rho\rho_5\Gamma_3\rho_3\tilde{\rho}]\\
&\qquad\times\sum_{sS_5\Sigma_5}(-1)^{-S_2-S_4+S_5+s}\hat{S}_1\hat{S}_3\hat{S}_4\hat{s}\langle S_4\Sigma_4 S_3\Sigma_3|S_5\Sigma_5\rangle\begin{Bmatrix}\frac{1}{2}&\frac{1}{2}&S_4\\ s&S_1&\frac{1}{2}\end{Bmatrix}\begin{Bmatrix}S_1&s&S_4\\ S_5&S_3&S_2\end{Bmatrix}\\
&\qquad\times\Big\{\delta_{\nu'\eta'}(-1)^{\Gamma_1}\{\{a^\dagger_\nu\times a^\dagger_\eta\}^{\Gamma s}\times\{\tilde{a}_\tau\times\tilde{a}_{\tau'}\}^{\Gamma_2 S_2}\}^{\rho_5\Gamma_5 S_5}_{\alpha_5\Sigma_5}U[\Gamma_4(\eta'0)\Gamma(\eta0);(\nu0)_{--}\Gamma_{1-}\rho]\\
&\qquad+\delta_{\nu'\eta}(-1)^{S_1}\{\{a^\dagger_\nu\times a^\dagger_{\eta'}\}^{\Gamma s}\times\{\tilde{a}_\tau\times\tilde{a}_{\tau'}\}^{\Gamma_2 S_2}\}^{\rho_5\Gamma_5 S_5}_{\alpha_5\Sigma_5}U[\Gamma_4(\eta0)\Gamma(\eta'0);(\nu0)_{--}\Gamma_{1-}\rho]\Big\}.
\end{aligned}
\tag{B11}
$$

APPENDIX C: SU(3) COUPLING AND RECOUPLING RULES

The work presented here makes extensive use of SU(2) and SU(3) coupling techniques and the (generalized) Wigner–Eckart theorem. The relevant definitions of the SU(2) coupling coefficients, along with their phase and normalization conventions, can be found in most textbooks on quantum mechanics. Analytic expressions and numerical values for special cases of these coefficients, as well as useful relations between different coupling coefficients, are compiled in various monographs on angular momentum theory, as for instance in reference 39. In contrast, analogous results for the SU(3) case required here are distributed over about a dozen different articles. For the convenience of the reader and as a practical reference for future analytical and numeric work in this field, we list the basic definitions and relations in this appendix. We also point the reader to sources of additional information and to publicly available computer codes. Specifically, we consider the coupling of two, three, and four irreducible SU(3) representations and give the Wigner–Eckart theorem for SU(3). The results presented here are primarily taken from publications by

Hecht *et al.*,[28,36,40–42] Draayer *et al.*,[34,43–45] O'Reilly,[38] Millener,[37] and Vergados.[46] Several new relations are included as well. The phase conventions used here are those employed in reference 44.

1. SU(3) Wigner coefficients: Coupling of two SU(3) irreps

If α represents a set of labels used to distinguish orthonormal basis states within a given irreducible SU(3) representation $\Gamma=(\lambda\mu)$, the Wigner coefficients $\langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma\alpha\rangle_\rho$ are defined as the elements of a unitary transformation between coupled and uncoupled orthonormal irreps of SU(3) in the α-scheme:[44]

$$|\Gamma\alpha\rangle_\rho=\sum_{\alpha_1\alpha_2}\langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma\alpha\rangle_\rho|\Gamma_1\alpha_1\rangle|\Gamma_2\alpha_2\rangle, \tag{C1}$$

and the inverse transformation is given by

$$|\Gamma_1\alpha_1\rangle|\Gamma_2\alpha_2\rangle=\sum_{\rho\Gamma\alpha}\langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma\alpha\rangle_\rho|\Gamma\alpha\rangle_\rho\,. \tag{C2}$$

Here $\alpha=\epsilon\Lambda M_\Lambda$ for the SU(3)$\supset$SU(2)$\otimes$U(1) (canonical) group chain and $\alpha=\kappa lm$ for the SU(3)$\supset$SO(3) reduction. The subgroup chains impose certain restrictions on the above couplings, for example, $\epsilon=\epsilon_1+\epsilon_2$, $M_\Lambda=M_{\Lambda_1}+M_{\Lambda_2}$, and $\Lambda=\Lambda_1+\Lambda_2,\ldots,|\Lambda_1-\Lambda_2|$ must hold for the canonical group chain, and the usual angular momentum coupling rules, $l=l_1+l_2,\ldots,|l_1-l_2|$, and $m=m_1+m_2$ apply for the chain containing SO(3). The conjugates of the relevant SU(3) basis states are given by[44]

$$|\Gamma\alpha\rangle^*=(-1)^{\chi_\alpha}|\tilde{\Gamma}\,\overline{\alpha}\rangle, \tag{C3}$$

where $\tilde{\Gamma}=(\mu\lambda)$ is the conjugate irrep to $\Gamma=(\lambda\mu)$, and $\overline{\alpha}$ and χ_i depend on the selected subgroup chain:[47]

$$\begin{gathered}\text{for } \alpha=\epsilon\Lambda M_\Lambda, \text{ one has } \overline{\alpha}=(-\epsilon)\Lambda(-M_\Lambda) \text{ and } \chi_i=\frac{1}{3}(\lambda_i-\mu_i)-\frac{1}{6}\epsilon_i-M_{\Lambda_i};\\ \text{for } \alpha=\kappa lm \text{ one has } \overline{\alpha}=\kappa l(-m) \text{ and } \chi_i=(\lambda_i-\mu_i)+l_i-m_i\,.\end{gathered} \tag{C4}$$

The outer multiplicity label $\rho=1,2,\ldots,\rho_{\max}$ is used to distinguish multiple occurrences of a given Γ in the direct product $\Gamma_1\times\Gamma_2$: $\rho=1,2,\ldots,\rho_{\max}$, where $\rho_{\max}$ denotes the number of possible couplings $\Gamma_1\times\Gamma_2$, and the possible Γ irreps in the product can be obtained by coupling the appropriate Young diagrams.[48] O'Reilly[38] determines a closed formula for the decomposition of the outer product $(p,q)\times(r,s)$ of finite-dimensional irreps of SU(3) for arbitrary positive integers p, q, r, and s:

$$(r,s)\times(p,q)=\bigoplus_{k=0}^{\min(q,r+s)}\bigoplus_{j=0}^{\min(s,p,r+s-k)}\bigoplus_{i=\max(0,j-s+k)}^{\min(p-j+k,r)}(r+p-j-2i+k,s+q+i-j-2k), \tag{C5}$$

and furthermore derives necessary and sufficient conditions for a SU(3) irrep (m,n) to appear as summand in the product $(r,s)\times(p,q)$.

Draayer and Akiyama[44] give a prescription for the unique determination, including the phases, of all SU(3)$\supset$SU(2)$\otimes$U(1) Wigner coefficients and derive their relevant conjugation and symmetry properties. Since the SU(3)$\supset$SO(3) reduction is related to the SU(3)$\supset$SU(2)$\otimes$U(1) reduction via the coefficients of the transformation between the $\alpha=\kappa lm$ and the $\alpha=\epsilon\Lambda M_\Lambda$ schemes,[44,45] it suffices to determine the conjugation relationship and symmetry properties for the SU(3)$\supset$SU(2)$\otimes$U(1) chain only. The corresponding SU(3)$\supset$SO(3) results follow then from the known relationships among the transformation brackets between the two schemes. These relations

are given in reference 44 and a computer code which allows for a numerical determination of SU(3)⊃SU(2)⊗U(1) and SU(3)⊃SO(3) Wigner coefficients, as well as SU(3) Racah coefficients, is published in reference 34. Analytic expressions for Wigner coefficients which are of particular interest in p-shell and ds-shell nuclear shell-model calculations are tabulated in reference 28 for the canonical subgroup chain and in references 46, 41 for the SU(3)⊃SO(3) chain.

The most important of the symmetry relations of the SU(3) Wigner coefficients is the one that involves a $1\leftrightarrow 3$ interchange of the quantum labels:

$$\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_3\alpha_3\rangle_\rho=(-1)^{\varphi+\chi_2}\sqrt{\frac{\dim(\Gamma_3)}{\dim(\Gamma_1)}}\,\langle \Gamma_3\alpha_3;\tilde{\Gamma}_2\overline{\alpha_2}|\Gamma_1\alpha_1\rangle_\rho, \tag{C6}$$

where $\varphi=(\lambda_1+\mu_1)+(\lambda_2+\mu_2)-(\lambda_3+\mu_3)$. The dimension of the SU(3) irrep Γ is given by $\dim(\Gamma)=\dim(\lambda\mu)=\frac{1}{2}(\lambda+1)(\mu+1)(\lambda+\mu+2)$, and $\overline{\alpha_i}$ and χ_i are as defined in Equation (C4).

The $1\leftrightarrow 2$ interchange is more complicated:

$$\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_3\alpha_3\rangle_\rho=\sum_{\rho'}\Phi_{\rho\rho'}[\Gamma_1,\Gamma_2;\Gamma_3]\langle \Gamma_2\alpha_2;\Gamma_1\alpha_1|\Gamma_3\alpha_3\rangle_{\rho'}, \tag{C7}$$

since it requires a phase matrix Φ, which is a special case of the recoupling coefficient Z that occurs in the coupling of three SU(3) irreps:

$$\Phi_{\rho\rho'}[\Gamma_1,\Gamma_2;\Gamma_3]=Z[\Gamma_1(00)\Gamma_3\Gamma_2;\Gamma_{1-}\rho\Gamma_{2-}\rho']. \tag{C8}$$

If the SU(3) coupling $\Gamma_1\times\Gamma_2\to\Gamma_3$ is unique, that is, when $\rho_{\max}=1$, the matrix reduces to a simple phase factor: $\Phi_{11}[\Gamma_1,\Gamma_2;\Gamma_3]=(-1)^{\varphi}=(-1)^{\Gamma_1+\Gamma_2-\Gamma_3}$, where $(-1)^{\Gamma_i}\equiv(-1)^{\lambda_i+\mu_i}$ for $\Gamma_i=(\lambda_i\mu_i)$.

Another set of useful relations involves the conjugate irrep $\tilde{\Gamma}=(\mu\lambda)$ of $\Gamma=(\lambda\mu)$; for the canonical group chain,[44]

$$\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_3\alpha_3\rangle_\rho=(-1)^{\varphi+\rho_{\max}-\rho}\langle \tilde{\Gamma}_1\overline{\alpha_1};\tilde{\Gamma}_2\overline{\alpha_2}|\tilde{\Gamma}_3\overline{\alpha_3}\rangle_\rho \tag{C9}$$

holds and for $\alpha=\kappa lm$ one has[44]

$$\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_3\alpha_3\rangle_\rho=(-1)^{\varphi+\rho_{\max}-\rho+l_1+l_2-l_3}\langle \tilde{\Gamma}_1\overline{\alpha_1};\tilde{\Gamma}_2\overline{\alpha_2}|\tilde{\Gamma}_3\overline{\alpha_3}\rangle_\rho. \tag{C10}$$

With the phase and normalization conventions of reference 44, one obtains the following orthonormality relations for the Wigner SU(3) coupling coefficients:

$$\sum_{\alpha_1\alpha_2}\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma\alpha\rangle_\rho\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma'\alpha\rangle_{\rho'}=\delta_{\Gamma'\Gamma}\delta_{\rho'\rho} \tag{C11}$$

and

$$\sum_{\rho\Gamma\alpha}\langle \Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma\alpha\rangle_\rho\langle \Gamma_1\alpha_1';\Gamma_2\alpha_2'|\Gamma\alpha\rangle_\rho=\delta_{\alpha_1\alpha_1'}\delta_{\alpha_2\alpha_2'}. \tag{C12}$$

It is possible to factor out the dependence of the above SU(3)⊃SO(3) and SU(3)⊃SU(2)⊗U(1) Wigner coupling coefficients on the M or M_Λ subgroup labels, respectively, by defining so-called double-barred or ''reduced'' SU(3) coupling coefficients:

$$\langle \Gamma_1\kappa_1l_1m_1;\Gamma_2\kappa_2l_2m_2|\Gamma\kappa lm\rangle_\rho=\underbrace{\langle \Gamma_1\kappa_1l_1;\Gamma_2\kappa_2l_2||\Gamma\kappa l\rangle_\rho}_{\text{reduced Wigner coefficient}}\underbrace{\langle l_1m_1,l_2m_2|lm\rangle}_{\text{geometric part}}, \tag{C13}$$

for $\alpha=\kappa lm$, and

$$\langle \Gamma_1 \epsilon_1 \Lambda_1 M_{\Lambda_1}; \Gamma_2 \epsilon_2 \Lambda_2 M_{\Lambda_2} | \Gamma \epsilon \Lambda M_\Lambda \rangle_\rho = \underbrace{\langle \Gamma_1 \epsilon_1 \Lambda_1; \Gamma_2 \epsilon_2 \Lambda_2 || \Gamma_3 \epsilon \Lambda \rangle_\rho}_{\text{reduced Wigner coefficient}} \underbrace{\langle \Lambda_1 M_{\Lambda_1}, \Lambda_2 M_{\Lambda_2} | \Lambda M_\Lambda \rangle}_{\text{geometric part}}, \tag{C14}$$

for $\alpha = \epsilon \Lambda M_\Lambda$. The ''geometric'' part $\langle \text{--}|\text{-}\rangle$ is simply a SU(2) Clebsch–Gordan coefficient. From the unitarity of the full SU(3) Wigner and the ordinary SU(2) Clebsch–Gordan coefficients it follows that the double-bar coefficients are also unitary. With the phase convention introduced by Draayer and Akiyama[44] they become real, and therefore orthogonal.

2. SU(3) Racah coefficients: Coupling of three SU(3) irreps

Coupling of three SU(3) irreps Γ_1, Γ_2, and Γ_3 to a given resulting SU(3) irrep Γ can be accomplished in three different ways, depending on the order of the coupling: (i) $\{\Gamma_1 \otimes \Gamma_2\} \otimes \Gamma_3 \to \Gamma$, or (ii) $\Gamma_1 \otimes \{\Gamma_2 \otimes \Gamma_3\} \to \Gamma$, or (iii) $\{\Gamma_1 \otimes \Gamma_3\} \otimes \Gamma_2 \to \Gamma$. The transformation from one coupling order to another requires the introduction of a so-called SU(3)-Racah or 6-($\lambda\mu$) coefficient. More specifically, recoupling from scheme (i) to (ii) involves a unitary transformation with coefficients $U[\Gamma_1 \Gamma_2 \Gamma \Gamma_3; \Gamma_{12} \rho_{12} \rho_{12,3} \Gamma_{23} \rho_{23} \rho_{1,23}]$ (see reference 28):

$$\begin{aligned} &|\{\{\Gamma_1 \times \Gamma_2\}^{\rho_{12} \Gamma_{12}} \times \Gamma_3\}_\alpha^{\rho_{12,3} \Gamma}\rangle \\ &\quad = \sum_{\Gamma_{23} \rho_{1,23} \rho_{23}} U[\Gamma_1 \Gamma_2 \Gamma \Gamma_3; \Gamma_{12} \rho_{12} \rho_{12,3} \Gamma_{23} \rho_{23} \rho_{1,23}] |\{\Gamma_1 \times \{\Gamma_2 \times \Gamma_3\}^{\rho_{23} \Gamma_{23}}\}_\alpha^{\rho_{1,23} \Gamma}\rangle, \end{aligned} \tag{C15}$$

and the inverse transformation is given by[36]

$$\begin{aligned} &|\{\Gamma_1 \times \{\Gamma_2 \times \Gamma_3\}^{\rho_{23} \Gamma_{23}}\}_\alpha^{\rho_{1,23} \Gamma}\rangle \\ &\quad = \sum_{\Gamma_{12} \rho_{12} \rho_{12,3}} U[\Gamma_1 \Gamma_2 \Gamma \Gamma_3; \Gamma_{12} \rho_{12} \rho_{12,3} \Gamma_{23} \rho_{23} \rho_{1,23}] |\{\{\Gamma_1 \times \Gamma_2\}^{\rho_{12} \Gamma_{12}} \times \Gamma_3\}_\alpha^{\rho_{12,3} \Gamma}\rangle. \end{aligned} \tag{C16}$$

The notation is a straightforward generalization of that introduced by Racah[39] for the 6-j symbols of SU(2). Whenever a coupling is multiplicity-free, the associated ρ-label may be omitted, provided the notation remains unambiguous.

Similarly, the transformation from scheme (i) to scheme (iii) requires a transformation coefficient $Z[\Gamma_2 \Gamma_1 \Gamma \Gamma_3; \Gamma_{12} \rho_{12} \rho_{12,3} \Gamma_{13} \rho_{13} \rho_{13,2}]$ which is defined through[37]

$$\begin{aligned} &|\{\{\Gamma_1 \times \Gamma_2\}^{\rho_{12} \Gamma_{12}} \times \Gamma_3\}_\alpha^{\rho_{12,3} \Gamma}\rangle \\ &\quad = \sum_{\Gamma_{13} \rho_{13} \rho_{13,2}} Z[\Gamma_2 \Gamma_1 \Gamma \Gamma_3; \Gamma_{12} \rho_{12} \rho_{12,3} \Gamma_{13} \rho_{13} \rho_{13,2}] |\{\{\Gamma_1 \times \Gamma_3\}^{\rho_{13} \Gamma_{13}} \times \Gamma_2\}_\alpha^{\rho_{13,2} \Gamma}\rangle. \end{aligned} \tag{C17}$$

The U- and Z-functions depend only on the SU(3) representations involved in the coupling and not on the specific subgroup chain chosen to specify the states. Most of the SU(3) recoupling coefficients needed for ds-shell nuclear shell model calculations have been tabulated by Hecht,[28] and a computer code that evaluates SU(3) Racah coefficients for arbitrary couplings and multiplicity was developed by Akiyama and Draayer.[34] Special cases and symmetry properties of the SU(3) coefficients can be found in reference 36. Most notably, one has $U[\cdots] = 1$ whenever Γ_1 or Γ_2 or Γ_3 or $\Gamma = (00)$, and

$$U[\Gamma_1 \Gamma_2 \Gamma_1 \tilde{\Gamma}_2; \Gamma_{12} \rho \rho' (00)_{--}] = (-1)^{\Gamma_1 + \Gamma_2 + \Gamma_{12}} \delta_{\rho \rho'} \sqrt{\frac{d(\Gamma_{12})}{d(\Gamma_1) d(\Gamma_2)}}. \tag{C18}$$

Under conjugation the SU(3) Racah coefficients exhibit the following symmetry:

$$U[\Gamma_1\Gamma_2\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{23}\rho_{23}\rho_{1,23}] = (-1)^{\rho_{12}^{\max}-\rho_{12}+\rho_{12,3}^{\max}-\rho_{12,3}+\rho_{23}^{\max}-\rho_{23}+\rho_{1,23}^{\max}-\rho_{1,23}} U[\tilde{\Gamma}_1\tilde{\Gamma}_2\tilde{\Gamma}\tilde{\Gamma}_3;\tilde{\Gamma}_{12}\rho_{12}\rho_{12,3}\tilde{\Gamma}_{23}\rho_{23}\rho_{1,23}], \tag{C19}$$

$$Z[\Gamma_2\Gamma_1\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{13}\rho_{13}\rho_{13,2}] = (-1)^{\rho_{12}^{\max}-\rho_{12}+\rho_{12,3}^{\max}-\rho_{12,3}+\rho_{13}^{\max}-\rho_{13}+\rho_{13,2}^{\max}-\rho_{13,2}} Z[\tilde{\Gamma}_2\tilde{\Gamma}_1\tilde{\Gamma}\tilde{\Gamma}_3;\tilde{\Gamma}_{12}\rho_{12}\rho_{12,3}\tilde{\Gamma}_{13}\rho_{13}\rho_{13,2}].$$

A straightforward generalization of the relations between SU(2) unitary recoupling coefficients and SU(2) Wigner coefficients leads to the corresponding relationships between SU(3) unitary (Racah) recoupling and SU(3) Wigner coefficients.

$$\begin{aligned} U[\Gamma_1\Gamma_2\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{23}\rho_{23}\rho_{1,23}] &= \sum_{\alpha_1\alpha_2\alpha_3\alpha_{12}\alpha_{23}} \langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_{12}\alpha_{12}\rangle_{\rho_{12}}\langle\Gamma_{12}\alpha_{12};\Gamma_3\alpha_3|\Gamma\alpha\rangle_{\rho_{12,3}} \\ &\quad\times\langle\Gamma_2\alpha_2;\Gamma_3\alpha_3|\Gamma_{23}\alpha_{23}\rangle_{\rho_{23}}\langle\Gamma_1\alpha_1;\Gamma_{23}\alpha_{23}|\Gamma\alpha\rangle_{\rho_{1,23}}, \end{aligned} \tag{C20}$$

for $\alpha_i=\epsilon_i\Lambda_iM_i$ or $\alpha_i=\kappa_il_im_i$.

The Z- and U-coefficients are related to each other as follows:

$$\begin{aligned} Z[\Gamma_2\Gamma_1\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{13}\rho_{13}\rho_{13,2}] &= \sum_{\Gamma_{23}\rho_{23}\rho_{1,23}} U[\Gamma_1\Gamma_2\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{23}\rho_{23}\rho_{1,23}]\sum_{\rho'_{23}}\Phi_{\rho_{23}\rho'_{23}}[\Gamma_2,\Gamma_3;\Gamma_{23}] \\ &\quad\times U[\Gamma_1\Gamma_3\Gamma\Gamma_2;\Gamma_{13}\rho_{13}\rho_{13,2}\Gamma_{23}\rho'_{23}\rho_{1,23}], \end{aligned} \tag{C21}$$

with the geometrical phase $\Phi_{\rho\rho'}$ as defined in Equation (C8). Further useful relations for the U- and Z-recoupling coefficients are given in reference 28, for the SU(3)$\supset$SU(2)$\otimes$U(1) chain, and in references 46, 37, for the SU(3)$\supset$SO(3) chain. The following one, which holds for both cases, is especially important:

$$\sum_{\rho_{1,23}}\langle\Gamma_1\alpha_1;\Gamma_{23}\alpha_{23}|\Gamma\alpha\rangle_{\rho_{1,23}}U[\Gamma_1\Gamma_2\Gamma\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{23}\rho_{23}\rho_{1,23}] = \sum_{\alpha_2\alpha_3\alpha_{12}}\langle\Gamma_2\alpha_2;\Gamma_3\alpha_3|\Gamma_{23}\alpha_{23}\rangle_{\rho_{23}}\langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_{12}\alpha_{12}\rangle_{\rho_{12}}\langle\Gamma_{12}\alpha_{12};\Gamma_3\alpha_3|\Gamma\alpha\rangle_{\rho_{12,3}}. \tag{C22}$$

A similar relation is given by Millener for the Z-coefficient.[37]

3. SU(3) 9-($\lambda\mu$) coefficients: Coupling of four SU(3) irreps

If the coupling of four SU(3) irreps Γ_1, Γ_2, Γ_3 and Γ_4 is required, the resulting irrep Γ may be constructed in three different ways: (i) $\{\Gamma_1\otimes\Gamma_2\}\otimes\{\Gamma_3\otimes\Gamma_4\}\rightarrow\Gamma$, or (ii) $\{\Gamma_1\otimes\Gamma_3\}\otimes\{\Gamma_2\otimes\Gamma_4\}\rightarrow\Gamma$, or (iii) $\{\Gamma_1\otimes\Gamma_4\}\otimes\{\Gamma_2\otimes\Gamma_3\}\rightarrow\Gamma$. In analogy to the SU(2) Jahn–Hope symbol it is possible to define a unitary SU(3) or 9-($\lambda\mu$) symbol, which effects the transformation from one coupling order to another. In particular, for the transition from scheme (i) to (ii) one defines[36]

$$|\{\{\Gamma_1\times\Gamma_2\}^{\rho_{12}\Gamma_{12}}\times\{\Gamma_3\times\Gamma_4\}^{\rho_{34}\Gamma_{34}}\}_{\alpha}^{\rho_{12,34}\Gamma}\rangle=\sum_{\rho_{13}\Gamma_{13}\rho_{24}\Gamma_{24}\rho_{13,24}}\begin{Bmatrix}\Gamma_1&\Gamma_2&\Gamma_{12}&\rho_{12}\\ \Gamma_3&\Gamma_4&\Gamma_{34}&\rho_{34}\\ \Gamma_{13}&\Gamma_{24}&\Gamma&\rho_{13,24}\\ \rho_{13}&\rho_{24}&\rho_{12,34}&\end{Bmatrix}\times|\{\{\Gamma_1\times\Gamma_3\}^{\rho_{13}\Gamma_{13}}\times\{\Gamma_2\times\Gamma_4\}^{\rho_{24}\Gamma_{24}}\}_{\alpha}^{\rho_{13,24}\Gamma}\rangle. \tag{C23}$$

Similarly, one obtains for the transformation from scheme (i) to (iii),

$$\begin{aligned}&|\{\{\Gamma_1\times\Gamma_2\}^{\rho_{12}\Gamma_{12}}\times\{\Gamma_3\times\Gamma_4\}^{\rho_{34}\Gamma_{34}}\}_{\alpha}^{\rho_{12,34}\Gamma}\rangle\\&\quad=\sum_{\rho_{14}\Gamma_{14}\rho_{23}\Gamma_{23}\rho_{14,23}}\begin{Bmatrix}\Gamma_1&\Gamma_2&\Gamma_{12}&\rho_{12}\\ \Gamma_4&\Gamma_3&\Gamma_{34}&\rho_{34}\\ \Gamma_{14}&\Gamma_{23}&\Gamma&\rho_{14,23}\\ \rho_{14}&\rho_{23}&\rho_{12,34}&\end{Bmatrix}\\&\qquad\times\sum_{\rho'_{34}}\Phi_{\rho_{34}\rho'_{34}}(\Gamma_3,\Gamma_4;\Gamma_{34})|\{\{\Gamma_1\times\Gamma_4\}^{\rho_{14}\Gamma_{14}}\times\{\Gamma_2\times\Gamma_3\}^{\rho_{23}\Gamma_{23}}\}_{\alpha}^{\rho_{14,23}\Gamma}\rangle.\end{aligned} \tag{C24}$$

The unitary 9-($\lambda\mu$) symbols can be expressed as a sum over SU(3) Wigner coupling coefficients:

$$\begin{aligned}&\begin{Bmatrix}\Gamma_1&\Gamma_2&\Gamma_{12}&\rho_{12}\\ \Gamma_3&\Gamma_4&\Gamma_{34}&\rho_{34}\\ \Gamma_{13}&\Gamma_{24}&\Gamma&\rho_{13,24}\\ \rho_{13}&\rho_{24}&\rho_{12,34}&\end{Bmatrix}\\&\quad=\sum_{\alpha_i\alpha_{ik}}\langle\Gamma_1\alpha_1;\Gamma_2\alpha_2|\Gamma_{12}\alpha_{12}\rangle_{\rho_{12}}\langle\Gamma_3\alpha_3;\Gamma_4\alpha_4|\Gamma_{34}\alpha_{34}\rangle_{\rho_{34}}\langle\Gamma_{13}\alpha_{13};\Gamma_{24}\alpha_{24}|\Gamma\alpha\rangle_{\rho_{13,24}}\\&\qquad\times\langle\Gamma_1\alpha_1;\Gamma_3\alpha_3|\Gamma_{13}\alpha_{13}\rangle_{\rho_{13}}\langle\Gamma_2\alpha_2;\Gamma_4\alpha_4|\Gamma_{24}\alpha_{24}\rangle_{\rho_{24}}\langle\Gamma_{12}\alpha_{12};\Gamma_{34}\alpha_{34}|\Gamma\alpha\rangle_{\rho_{12,34}},\end{aligned} \tag{C25}$$

or in terms of the Racah Z and U coefficients defined above,[37]

$$\begin{aligned}&\begin{Bmatrix}\Gamma_1&\Gamma_2&\Gamma_{12}&\rho_{12}\\ \Gamma_3&\Gamma_4&\Gamma_{34}&\rho_{34}\\ \Gamma_{13}&\Gamma_{24}&\Gamma&\rho_{13,24}\\ \rho_{13}&\rho_{24}&\rho_{12,34}&\end{Bmatrix}\\&\quad=\sum_{\Gamma_0\rho_{13,2}\rho_{04}\rho_{12,3}}U[\Gamma_{13}\Gamma_2\Gamma\Gamma_4;\Gamma_0\rho_{13,2}\rho_{04}\Gamma_{24}\rho_{24}\rho_{13,24}]Z[\Gamma_2\Gamma_1\Gamma_0\Gamma_3;\Gamma_{12}\rho_{12}\rho_{12,3}\Gamma_{13}\rho_{13}\rho_{13,2}]\\&\qquad\times U[\Gamma_{12}\Gamma_3\Gamma\Gamma_4;\Gamma_0\rho_{12,3}\rho_{04}\Gamma_{34}\rho_{34}\rho_{12,34}].\end{aligned} \tag{C26}$$

Various symmetry relations and special cases of the 9-($\lambda\mu$) coefficients can be found in references 40, 36, and a computer code which provides numerical values for these recoupling coefficients is publicly available.[34]

4. Matrix elements and the Wigner–Eckart theorem for SU(3)

The Wigner–Eckart theorem for the group SU(2) yields SU(2)-reduced (double-bar) matrix elements of an SO(3) irreducible tensor operator. Analogously, the generalized Wigner–Eckart

theorem allows one to express matrix elements of SU(3) irreducible tensor operators as a sum over ρ of the product of a ρ-dependent generalized reduced matrix element multiplied by the corresponding Wigner coefficient:[44]

$$\langle \Gamma_3 \alpha_3 | T^{\Gamma_2 \alpha_2} | \Gamma_1 \alpha_1 \rangle = \sum_{\rho} \langle \Gamma_1 \alpha_1 ; \Gamma_2 \alpha_2 | \Gamma_3 \alpha_3 \rangle_{\rho} \langle \Gamma_3 ||| T^{\Gamma_2} ||| \Gamma_1 \rangle_{\rho}. \tag{C27}$$

Note that the triple-bar matrix element is independent of the chosen subgroup chain.

In analogy to the well-known reduction rules for SU(2),[39] one can derive expressions for the triple-bar matrix elements of a SU(3)-coupled tensor product acting on a one-component system:

$$\langle \Gamma ||| \{ T^{\Gamma_1} \times T^{\Gamma_2} \}^{\rho_3 \Gamma_3} ||| \Gamma' \rangle_{\rho} = \sum_{\rho_1 \rho_2 \Gamma''} \langle \Gamma ||| T^{\Gamma_1} ||| \Gamma'' \rangle_{\rho_1} \langle \Gamma'' ||| T^{\Gamma_2} ||| \Gamma' \rangle_{\rho_2} \times \sum_{\rho_3'} \Phi_{\rho_3 \rho_3'} [\Gamma_1 \Gamma_2 ; \Gamma_3] U [\Gamma' \Gamma_2 \Gamma \Gamma_1 ; \Gamma'' \rho_2 \rho_1 \Gamma_3 \rho_3' \rho]. \tag{C28}$$

Often one has to consider a quantum-mechanical system which consists of two subsystems, 1 and 2 (for example, protons and neutrons). Reduced matrix elements for the irreducible tensor product of two operators, $R^{\Gamma_r}(1)$ and $S^{\Gamma_s}(2)$, which depend only on variables of the first and second subsystem, respectively, may be evaluated with the help of the following expression:

$$\langle \{ \Gamma_1 \times \Gamma_2 \} \rho \Gamma ||| \{ R^{\Gamma_r}(1) \times S^{\Gamma_s}(2) \}^{\rho_t \Gamma_t} ||| \{ \Gamma_1' \times \Gamma_2' \} \rho' \Gamma' \rangle_{\tilde{\rho}} = \sum_{\rho_1 \rho_2} \begin{Bmatrix} \Gamma_1' & \Gamma_r & \Gamma_1 & \rho_1 \\ \Gamma_2' & \Gamma_s & \Gamma_2 & \rho_2 \\ \Gamma' & \Gamma_t & \Gamma & \tilde{\rho} \\ \rho' & \rho_t & \rho & \end{Bmatrix} \langle \Gamma_1 ||| R^{\Gamma_r}(1) ||| \Gamma_1' \rangle_{\rho_1} \langle \Gamma_2 ||| S^{\Gamma_s}(2) ||| \Gamma_2' \rangle_{\rho_2}. \tag{C29}$$

[1] G. Rosensteel and D. J. Rowe, ''Nuclear Sp(3,R) model,'' Phys. Rev. Lett. **38**, 10–14 (1977).

[2] G. Rosensteel and D. J. Rowe, ''On the algebraic formulation of collective models. III. The symplectic shell model of collective motion,'' Ann. Phys. (N.Y.) **126**, 343–370 (1980).

[3] D. J. Rowe, ''Microscopic theory of the nuclear collective model,'' Rep. Prog. Phys. **48**, 1419–1480 (1985).

[4] J. Deenen and C. Quesne, ''Dynamical group of microscopic collective states. I. One-dimensional case,'' J. Math. Phys. **23**, 878–889 (1982).

[5] J. Deenen and C. Quesne, ''Dynamical group of microscopic collective states. II. Boson representations in d dimensions,'' J. Math. Phys. **23**, 2004–2015 (1982).

[6] J. Deenen and C. Quesne, ''Boson representations of the real symplectic group and their application to the nuclear collective model,'' J. Math. Phys. **26**, 2705–2716 (1985).

[7] D. J. Rowe, ''Coherent state theory of the noncompact symplectic group,'' J. Math. Phys. **25**, 2662–2671 (1984).

[8] D. J. Rowe, G. Rosensteel, and R. Gilmore, ''Vector coherent state representation theory,'' J. Math. Phys. **26**, 2787–2791 (1985).

[9] M. Moshinsky, ''Boson realisation of symplectic algebras,'' J. Phys. A **18**, L1–L6 (1985).

[10] O. Castaños, Chacón, M. Moshinsky, and C. Quesne, ''Boson realization of sp(4). I. The matrix formulation,'' J. Math. Phys. **26**, 2107–2123 (1985).

[11] O. Castaños, P. Kramer, and M. Moshinsky, ''Boson realization of sp(4). II. The generating kernel formulation,'' J. Math. Phys. **27**, 924–935 (1985).

[12] P. Kramer, Z. Papadopolos, and W. Schweizer, ''A class of unitary representations of the Lie group Sp(3,R), its coherent states, and its map to a symplectic realization on sp*(3,R),'' J. Math. Phys. **27**, 24–28 (1986).

[13] K. T. Hecht, *The Vector Coherent State Method and Its Application to Problems of Higher Symmetry* (Springer-Verlag, Berlin, 1987).

[14] D. J. Rowe, G. Rosensteel, and R. Carr, ''Analytical expressions for the matrix elements of the non-compact symplectic algebra,'' J. Phys. A **17**, L399–L403 (1984).

[15] O. Castaños, E. Chacón, and M. Moshinsky, ''Analytic expressions for the matrix elements of generators of Sp(6) in an Sp(6)$\supset$U(3) basis,'' J. Math. Phys. **25**, 1211–1218 (1984).

[16] J. Deenen and C. Quesne, ''Determination of Sp(2d,R) generator matrix elements through a boson mapping,'' J. Phys. A **17**, L405–L409 (1984).

[17] Y. Suzuki and K. T. Hecht, ''Symplectic and cluster excitations in nuclei: Evaluation of interaction matrix elements,'' Nucl. Phys. A **455**, 315–343 (1986).

[18] Y. Suzuki and K. T. Hecht, ''Spin–orbit and tensor interactions in sp(6,R)-model calculations,'' Prog. Theor. Phys. **77**, 190–195 (1987).

[19] E. J. Reske, ''Sp(6,R) symmetry and the giant quadrupole resonance in Mg24,'' Ph.D. thesis, University of Michigan, 1984.
[20] O. Castaños and J. P. Draayer, ''Contracted symplectic model with ds-shell applications,'' Nucl. Phys. A **491**, 349–372 (1989).
[21] O. Castaños, P. O. Hess, J. P. Draayer, and P. Rochford, ''Pseudo-symplectic model for strongly deformed heavy nuclei,'' Nucl. Phys. A **524**, 469–478 (1991).
[22] J. P. Draayer, K. J. Weeks, and G. Rosensteel, ''Symplectic shell-model calculations for ^{20}Ne with horizontal configuration mixing,'' Nucl. Phys. A **413**, 215–222 (1984).
[23] G. Rosensteel, J. P. Draayer, and K. J. Weeks, ''Symplectic shell-model calculation for ^{24}Mg,'' Nucl. Phys. A **419**, 1–12 (1984).
[24] D. Troltenier, J. P. Draayer, P. O. Hess, and O. Castaños, ''Investigations of rotational nuclei via the pseudo-symplectic model,'' Nucl. Phys. A **576**, 351–386 (1994).
[25] C. Bahri and J. P. Draayer, ''SU(3) reduced matrix element package,'' Comput. Phys. Commun. **83**, 59–94 (1994).
[26] G. Rosensteel, ''Analytic formula for quadrupole–quadrupole matrix elements,'' Phys. Rev. C **42**, 2463–2468 (1990).
[27] G. Rosensteel, ''A recursion formula for sp(3,R) matrix elements,'' J. Math. Phys. **21**, 924–927 (1980).
[28] K. T. Hecht, ''SU_3 recoupling and fractional parentage in the $2s$-$1d$ shell,'' Nucl. Phys. A **62**, 1–36 (1965).
[29] J. P. Draayer and G. Rosensteel, ''U(3)→R(3) integrity-basis spectroscopy,'' Nucl. Phys. A **439**, 61–85 (1985).
[30] J. P. Elliott, ''Collective motion in the nuclear shell model: I. Classification schemes for states of mixed configurations,'' Proc. R. Soc. London, Ser. A **245**, 128–154 (1958).
[31] J. P. Elliott, ''Collective motion in the nuclear shell model: II. The introduction of intrinsic wave functions,'' Proc. R. Soc. London, Ser. A **245**, 562–581 (1958).
[32] K. T. Hecht, ''An approximation formula for the κ-matrix elements of the symplectic algebra sp(6,R),'' J. Phys. A **18**, L1003–L1008 (1985).
[33] G. Rosensteel and D. J. Rowe, ''An analytic formula for u(3)-boson matrix elements,'' J. Math. Phys. **24**, 2461–2463 (1983).
[34] Y. Akiyama and J. P. Draayer, ''A user's guide to Fortran programs for Wigner and Racah coefficients of SU_3,'' Comput. Phys. Commun. **5**, 405–415 (1973).
[35] J. Escher, *Electron scattering studies in the framework of the symplectic shell model*, Ph.D. thesis, Louisiana State University, 1997.
[36] K. T. Hecht, E. J. Reske, T. H. Seligman, and W. Zahn, ''Spectroscopic amplitudes for complex cluster systems,'' Nucl. Phys. A **356**, 146–222 (1981).
[37] D. J. Millener, ''A note on recoupling coefficients for SU(3),'' J. Math. Phys. **19**, 1513–1514 (1978).
[38] M. F. O'Reilly, ''A closed formula for the product of irreducible representations of SU(3),'' J. Math. Phys. **23**, 2022–2028 (1982).
[39] D. A. Varshalovich, A. N. Moskalev, and V. K. Kersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).
[40] K. T. Hecht and D. Braunschweig, ''Few-nucleon SU(3) parentage coefficients and α-particle spectroscopic amplitudes for core excited states in s-d shell nuclei,'' Nucl. Phys. A **244**, 365–434 (1975).
[41] K. T. Hecht and Y. Suzuki, ''Some special SU(3)⊃R(3) Wigner coefficients and their application,'' J. Math. Phys. **24**, 785–792 (1982).
[42] D. R. Peterson and K. T. Hecht, ''Sp(4,R) symmetry in light nuclei,'' Nucl. Phys. A **344**, 361–388 (1980).
[43] O. Castaños, J. P. Draayer, and Y. Leschber, ''Towards a shell model description of the low-energy structure of deformed nuclei. II. Electromagnetic properties of collective M1 bands,'' Ann. Phys. (Leipzig) **180**, 290–329 (1989).
[44] J. P. Draayer and Y. Akiyama, ''Wigner and Racah coefficients for SU_3,'' J. Math. Phys. **14**, 1904–1912 (1973).
[45] J. P. Draayer and S. A. Williams, ''Coupling coefficients and matrix elements of arbitrary tensors in the Elliott projected angular momentum basis,'' Nucl. Phys. A **129**, 647–665 (1969).
[46] J. D. Vergados, ''SU(3)⊃R(3) Wigner coefficients in the $2s$-$1d$ shell,'' Nucl. Phys. A **111**, 681–754 (1968).
[47] Note that the phase convention used here differs by $2M_\Lambda$ from that employed by Hecht,[28] where $\chi_\alpha=\frac{1}{3}(\lambda-\mu)-\frac{1}{6}\epsilon+M_\Lambda$ for $\alpha=\epsilon\Lambda M_\Lambda$.
[48] M. Hamermesh, *Group Theory and its Applications to Physical Problems* (Dover, New York, 1989).

Operator formalism for bosonic beta–gamma fields on general algebraic curves

Franco Ferrari[a)]
LPTHE, Université Pierre et Marie Curie–Paris VI and Université Denis Diderot–Paris VII, Boite 126, Tour 16, 1^er étage, 4 place Jussieu, F-75252 Paris Cedex 05, France

Jan T. Sobczyk[b)]
Institute for Theoretical Physics, Wrocław University, pl. Maxa Borna 9, 50205 Wrocław, Poland

(Received 13 November 1997; accepted for publication 7 May 1998)

An operator formalism for bosonic $\beta-\gamma$ system on arbitrary algebraic curves is introduced. The classical degrees of freedom are identified and their commutation relations are postulated. The explicit realization of the algebra formed by the fields is given in the Hilbert space equipped with a bilinear form. The construction is based on the ''Gaussian'' representation for $\beta-\gamma$ system on the complex sphere [Alvarez-Gaumé *et al.*, Nucl. Phys. B **311**, 333 (1988)]. Detailed computations are provided for two- and four-point correlation functions. © *1998 American Institute of Physics.* [S0022-2488(98)03510-5]

I. INTRODUCTION

Since the middle of the 1980s tremendous progress has been achieved in understanding the structure of $D=2$ conformal field theories (CFTs).[1] Most of that work was motivated by string theory, but CFTs have also been actively investigated because of their role in describing phase transition phenomena. In string theory it is necessary to consider CFTs defined on topologically nontrivial $D=2$ manifolds, in particular on closed higher genus Riemann surfaces (RS) of genus g.[2] The properties of CFTs on the torus are now relatively well known and, at least in the context of string theory, it turns out that the class of acceptable models are those which are modular invariant.[3]

There have been many efforts to give a satisfactory description of CFTs also on higher genus RS, for instance by means of an operator formalism.[4] Usually, an operator formalism of such kind makes it possible to introduce the notions of vacuum, creation, and annihilation operators, normal ordering, etc. In this way, theories defined on curved space–times resemble ordinary quantum field theories in flat space. A valid operator formalism for string theory application should be able to reproduce the relevant correlation functions and their analytical properties. Usually these properties (the structure of zeros and poles) are strong enough to determine these correlators up to an overall constant.[5,6]

One version of operator formalism has been developed for the $b-c$ systems in Refs. 7 and 8. In this paper we would like to extend it in such a way that also the bosonic $\beta-\gamma$ systems can effectively be treated.

These systems are in fact interesting examples of CFTs. They were extensively studied in the second half of the 1980s after they turned out to be a necessary ingredient (as Faddeev–Popov superghosts) in the perturbative approach to strings.[9–12] A treatment on hyperelliptic curves, mainly in connection with the computation of amplitudes in superstring theory at two loops, can for instance be found in Ref. 13. $\beta-\gamma$ systems have also been studied in the context of the Wess–Zumino–Witten model.[14]

The analysis of the $\beta-\gamma$ systems is in many respects analogous to that of the $b-c$ systems. In fact, the only difference lies in the statistics, since the $\beta-\gamma$ fields commute while the $b-c$ fields

a)Electronic mail: fferrari@lpthe.jussieu.fr
b)Electronic mail: jsobczyk@proton.ift.uni.wroc.pl

 © 1998 American Institute of Physics

anticommute. Following Ref. 9, many formulas characterizing both systems can be written in common. If by **b** (and **c**) one denotes either the b or β (c or γ) fields with conformal dimension λ (and $1-\lambda$), then both theories are defined by the action:

$$S=\frac{1}{\pi}\int d^2\, z(\mathbf{b}\bar{\partial}\mathbf{c}), \tag{1}$$

where $\bar{\partial}\equiv\partial/\partial\bar{z}$. It follows from (1) that the one-pion exchanges (OPEs) for the fields are

$$\mathbf{c}(z)\mathbf{b}(w)\sim\frac{1}{z-w},\quad \mathbf{b}(z)\mathbf{c}(w)\sim\frac{\epsilon}{z-w}, \tag{2}$$

where $\epsilon=1$ for $b-c$ systems and $\epsilon=-1$ for $\beta-\gamma$ systems. The energy–momentum tensor, determined by the dependence of (1) on the metric (in order to derive it one has to write the action in the covariant form) carries no dependence on ϵ:

$$T=-\lambda\mathbf{b}\partial\mathbf{c}+(1-\lambda)(\partial\mathbf{b})\mathbf{c}. \tag{3}$$

The OPEs of the elementary fields with T can be described with the same formula both in the fermionic and bosonic cases, as they are determined only by conformal dimensions:

$$T(z)\mathbf{b}(w)\sim\frac{\lambda}{(z-w)^2}\,\mathbf{b}(w)+\frac{1}{z-w}\,\partial\mathbf{b}(w)+\cdots, \tag{4}$$

$$T(z)\mathbf{c}(w)\sim\frac{1-\lambda}{(z-w)^2}\,\mathbf{c}(w)+\frac{1}{z-w}\,\partial\mathbf{c}(w)+\cdots. \tag{5}$$

The OPE with two energy–momentum tensors depends on ϵ since in calculating it one has to (anti)commute elementary fields:

$$T(z)T(w)\sim\frac{-\epsilon(6\lambda^2-6\lambda+1)}{(z-w)^4}+\frac{2T(w)}{(z-w)^2}+\frac{\partial T(w)}{z-w}+\cdots. \tag{6}$$

The correlation functions of the $\beta-\gamma$ systems were intensively studied at the end of the last decade.[11,12]

In string theory computations, one would like to compute the correlators also within the bosonized version of the theory. It turns out that in addition to the scalar fields with quadratic action one has to introduce also auxiliary fermionic fields $\eta-\xi$ of conformal dimensions one and zero. One gets an infinite number of inequivalent representations (q vacua in Ref. 9) of the basic algebra in a Hilbert space.

For $\lambda\geqslant 1$ and $g\geqslant 2$ only the zero modes in the **b** fields are present. In both bosonic and fermionic cases one has to insert delta functions for any zero mode in order to obtain a reasonable correlator. Without these insertions one gets vanishing or diverging amplitudes in the case of fermionic or bosonic systems, respectively. In the path integral approach this follows from the properties of the integrals

$$\int_{\mathbf{R}} dx\to\infty,\quad \int d\zeta=0, \tag{7}$$

where ζ is an element of a Grassmann algebra satisfying $\zeta^2=0$. In the fermionic case $\delta(\zeta)=\zeta$, so that any insertion in the amplitudes of a delta function $\delta(b)$ is equivalent to an insertion of a b field. In the bosonic case one has instead to make sense of true delta functions (or step functions) containing field operators in their argument.

The manipulations with delta function operators become straightforward in the ''Gaussian'' representation for β and γ fields.[15] In that representation one sees that inequivalent vacua can be chosen and that they cannot be interchanged under the action of a finite number of elementary excitations. Each vacuum state is determined by a choice of signs in an infinite product of

Gaussian (or, strictly speaking ''Fresnelian'') functions $\exp(\pm(i/2)x^2)$. The general property of deltalike operators is that they can map one vacuum state into another.[15] Some operations in the Gaussian representation may seem formal, e.g., formally infinite normalization factors have to be included. However, there is always a unique procedure to handle with these factors in order to obtain finite results.

The Gaussian representation makes it possible to define an operator formalism for the β–γ systems following the ideas of Ref. 8, where the case of the b–c systems has been discussed. In Ref. 8, the classical degrees of freedom are identified and the simplest possible anticommutation relations are postulated. Some of the elementary oscillators are defined as annihilation or creation operators. The remaining excitations correspond to the zero modes of the model. It has been demonstrated that these definitions enable the calculation of the correlation functions of the theory. The construction is not restricted to any particular class of RS—it can be done on arbitrary algebraic curve.

The aim of this paper is to present in detail the construction of an analogous operator formalism on RS for β–γ systems. The outcome of our computations are correlation functions with all the necessary analytic properties.

The operator formalism for the β–γ systems exhibits some similarities and also some interesting differences with respect to the operator formalism found in the case of the b–c fields in Ref. 8. The starting point is the same, one chooses the same classical degrees of freedom, and postulates the simplest possible commutation (instead of anticommutation) relations. One decomposes the elementary excitations into annihilation, creation, and zero mode sectors. A fundamental property of the normal ordering is that the difference between an ordered and an unordered product of two fields is equal to the Weierstrass kernel $K_\lambda(z,w)$, which is one of the basic building blocks in the construction of the correlation functions. $K_\lambda(z,w)$ is a λ differential in z, a $1-\lambda$ differential in w, and its only finite singularity in z is a single pole at $z=w$ (provided z and w parametrize points on the same branch of the RS). In our investigations we need to introduce multivalued fields on the complex sphere. In this sense one can think that new classes of CFTs containing multivalued models are defined on the sphere.[7] The computation of the correlation functions for β–γ systems is quite different from the case of the b–c systems. In the bosonic case many nontrivial manipulations with Fresnel integrals are required before reaching the final result. This will be shown in detail in Secs. II–IV.

The experience from the β–γ systems can hopefully be used in dealing with more complicated and important systems, such as for instance the bosonic theories with quadratic action. These theories provide the basis of the Coulomb gas formulation of minimal models. Therefore, the construction of an operator formalism for the β–γ systems can shed some light on the treatment of the minimal models on arbitrary RS represented as algebraic curve. Until now, only the case of hyperelliptic and Z_N symmetric curves have been studied in this context.[16,17]

II. $\lambda=-1$ COMPUTATIONS ON THE SPHERE

To better explain the ideas of the calculations, it is useful to start with the simpler situation of the complex sphere. We choose in this section $g=0$ and $\lambda=-1$ in order to have β rather than γ field zero modes.

The classical degrees of freedom are identified as

$$\beta(z)=\sum_{j=-\infty}^{+\infty}\beta_j z^{-j+1}dz^{\lambda}, \tag{8}$$

$$\gamma(z)=\sum_{k=-\infty}^{+\infty}\gamma_k z^{-k-2}dz^{1-\lambda}. \tag{9}$$

In quantum theory one postulates the following commutation relations:

$$[\gamma_k,\beta_j]=\delta_{j+k,0}. \tag{10}$$

The normal ordering is defined by assigning to the elementary excitations β_j and γ_k the property that they are creation or annihilation operators. Annihilation operators correspond to $j \geqslant 2$ and $k \geqslant -1$. Creation operators are those with $j \leqslant -2$ and $k \leqslant -2$. The remaining β_j operators $(j=0,\pm 1)$ correspond to the three zero modes (holomorphic vectors fields).

The above assignment implies

$$\gamma(z)\beta(w) = :\gamma(z)\beta(w): + \frac{1}{z-w}, \tag{11}$$

which should be understood as an identity in the radial ordered correlation functions.

Another important identity is [in $\exp(ip\beta(z))$] the introduction of normal ordering (makes no modification):

$$\gamma(z)\mathrm{e}^{ip\beta(w)} = :\gamma(z)\mathrm{e}^{ip\beta(w)}: + \frac{ip}{z-w}\,\mathrm{e}^{ip\beta(w)}. \tag{12}$$

Let us start calculating the propagator of the theory. It is known from the general discussion and it will also be very clearly seen in the course of the computations that the propagator requires three (the number of β field zero modes) insertions of delta functions of β fields.

Using the Wick theorem one calculates:

$$\begin{aligned}\langle 0|\gamma(z)\beta(w)\exp\left(i\sum_{r=1}^{3} p_r\beta(w_r)\right)|0\rangle = \frac{1}{z-w}\langle 0|\exp\left(i\sum_{r=1}^{3}\sum_{j=-\infty}^{+\infty} p_r\beta_j w_r^{-j+1}\right)|0\rangle \\ + \sum_{r=1}^{3}\frac{ip_r}{z-w_r}\langle 0|\beta(w)\exp\left(i\sum_{t=1}^{3}\sum_{j=-\infty}^{+\infty} p_t\beta_j w_t^{-j+1}\right)|0\rangle.\end{aligned} \tag{13}$$

Finally, the integration over $\Pi_{r=1}^{3}(dp_r/2\pi)$ has to be performed.

The computations will be done in the ''Gaussian'' representation for the β–γ systems.[15] The basic definitions of that construction are summarized in Appendix A.

In order to evaluate (13) one calculates

$$\exp\left(i\sum_{n=-\infty}^{+\infty} A_n\beta_n\right)|0\rangle \tag{14}$$

with the state $|0\rangle$ taken to be the state $\Phi^{(-1)}$ defined in (A8)

$$|0\rangle = \Phi^{(-1)} = \exp\frac{i}{2}\left(\sum_{n\geqslant -1} x_n^2 - \sum_{n<-1} x_n^2\right). \tag{15}$$

A concrete realization of the operators β_n is given by [see (A2)]

$$\beta_n = -\frac{i}{\sqrt{2}}\left(x_{-n} - i\frac{\partial}{\partial x_{-n}}\right). \tag{16}$$

The operators β_n involve only x_{-n} and acting on $\exp(-(i/2)x_{-n}^2)$ give 0. It follows that

$$\exp\left(i\sum_{n=-\infty}^{+\infty} A_n\beta_n\right)|0\rangle = \exp\left(-\frac{i}{2}\sum_{m<-1} x_m^2\right)\exp\left(i\sum_{n\leqslant 1} A_n\beta_n\right)\exp\left(\frac{i}{2}\sum_{m\geqslant -1} x_m^2\right). \tag{17}$$

Repeating steps similar to those presented in the end of Appendix A one finds

$$\exp\left(-\frac{i}{2}\sum_{m<-1}x_m^2\right)\exp\left(i\sum_{n\leqslant 1}A_n\beta_n\right)\exp\left(\frac{i}{2}\sum_{m\leqslant 1}x_{-m}^2\right)$$
$$=\exp\left(-\frac{i}{2}\sum_{m<-1}x_m^2\right)\exp\left(-\frac{i}{2}\sum_{n\leqslant 1}((A_n+i\sqrt{2}x_{-n})^2+x_{-n}^2)\right). \tag{18}$$

To define a bilinear form we introduce the right vacuum:

$$\langle 0|=\exp\frac{i}{2}\left(\sum_{m>1}x_m^2-\sum_{m\leqslant 1}x_m^2\right). \tag{19}$$

A proper normalization of scalar product is included in a (formally infinite) factor V:

$$\langle 0|\prod_{s=1}^{3}\delta(\beta(z_s)|0\rangle=\frac{1}{V}\int\prod_{s=1}^{3}dp_s\int\prod_{k=-\infty}^{+\infty}dx_k\,\exp\left(-i\sum_{m<1}x_m^2\right)$$
$$\times\exp\left(-\frac{i}{2}\sum_{m>1}(A_{-m}+i\sqrt{2}x_m)^2\right)$$
$$\times\exp\left(-\frac{i}{2}\sum_{m=0,\pm 1}((A_m+i\sqrt{2}x_{-m})^2+2x_{-m}^2))\right) \tag{20}$$

for $A_m=\Sigma_{s=1}^{3}\,p_s z_s^{1-m}$. It will soon become clear that V should be defined as

$$V=\left(\prod_{m<1}\int\exp(-ix_m^2)dx_m\right)\left(\prod_{n>1}\int\exp(ix_n^2)dx_n\right)$$
$$\times\left(\prod_{k=0,\pm 1}\int\exp\left(-\frac{i}{2}p_k^2\right)dp_k\right)\left(\prod_{l=0,\pm 1}\int\exp(-ix_l^2)dx_l\right). \tag{21}$$

It is most convenient to integrate first over $\Pi_m\,dx_m$ for $m<1$ and then for $m>1$. Both integrations are trivial and contribute only to overall normalization. Next one should integrate over $\Pi_{s=1}^{3}\,dp_s$ after a convenient change of variables that will produce an appropriate Jacobian. The trivial integration over $\Pi_{m=0,\pm 1}\,dx_m$ will be performed later.

The above-mentioned change of variables is

$$p_j\to A_m,\quad j=1,2,3,\quad m=0,\pm 1, \tag{22}$$

$$\int\prod_{j=1}^{3}dp_j=\int\prod_{m=-1}^{1}dA_m\,\det\left(\frac{\partial A_n}{\partial p_k}\right)^{-1}, \tag{23}$$

where

$$\det\left(\frac{\partial A_n}{\partial p_k}\right)=(w_1-w_2)(w_1-w_3)(w_2-w_3). \tag{24}$$

The integration over dp_s and dx_n cancel precisely the factor V and the final result is

$$\langle 0|\prod_{s=1}^{3}\delta(\beta(w_s)|0\rangle=\frac{1}{(w_1-w_2)(w_1-w_3)(w_2-w_3)}. \tag{25}$$

In the computation of the second term in (13) it is sufficient to concentrate on the most important point without repeating many of the above arguments,

$$\beta(w)\exp\left(i\sum_{n=-\infty}^{+\infty}A_n\beta_n\right)|0\rangle=-\frac{i}{\sqrt{2}}\sum_{j=-\infty}^{+\infty}w^{-j+1}\left(x_{-j}-i\frac{\partial}{\partial x_{-j}}\right)\exp\left(-\frac{i}{2}\sum_{m<-1}x_m^2\right)$$
$$\times\exp\left(-\frac{i}{2}\sum_{n\leqslant 1}((A_n+i\sqrt{2}x_{-n})^2+x_{-n}^2)\right)$$
$$=-\exp\left(-\frac{i}{2}\sum_{m<-1}x_m^2\right)\sum_{j\leqslant 1}w^{-j+1}(A_j+i\sqrt{2}x_{-j})$$
$$\times\exp\left(-\frac{i}{2}\sum_{n\leqslant 1}((A_n+i\sqrt{2}x_{-n})^2+x_{-n}^2)\right). \tag{26}$$

With the above definition of the right vacuum $\langle 0|$ one gets

$$\langle 0|\beta(w)\exp\left(i\sum_{n-\infty}^{+\infty}A_n\beta_n\right)|0\rangle=-\frac{1}{V}\int\prod_{s=1}^{3}dp_s\int\prod_{k=-\infty}^{+\infty}dx_k\sum_{j\leqslant 1}w^{-j+1}(A_j+i\sqrt{2}x_{-j})$$
$$\times\exp\left(-i\sum_{m<1}x_m^2\right)\exp\left(-\frac{i}{2}\sum_{m>1}(A_{-m}+i\sqrt{2}x_m)^2\right)$$
$$\times\exp\left(-\frac{i}{2}\sum_{m=0,\pm 1}((A_m+i\sqrt{2}x_{-m})^2+2x_{-m}^2)\right). \tag{27}$$

As before, it is most convenient to perform first the integration over $\Pi_{m<-1}\,dx_m$ and then over $\Pi_{m>1}\,dx_m$. The sum over $j\leqslant 1$ reduces to three terms with $j=0,\pm 1$. All the other terms vanish due to integration over antisymmetric function in $x_n(n>1)$.

Having in mind future generalizations, the following notation is useful:

$$A_j=\sum_{k=1}^{3}R_{jk}p_k. \tag{28}$$

Thus, in computating the second term in (13) one is left with the evaluation of

$$-i\frac{1}{V'}\int\prod_{j=0,\pm 1}dA_j\int\prod_{n=0,\pm 1}dx_n\,\det\left(\frac{\partial A_m}{\partial p_r}\right)^{-1}\sum_{s=1}^{3}\sum_{j,k=-1}^{1}\frac{R_{sk}^{-1}A_k}{z-w_s}w^{1-j}(A_j+i\sqrt{2}x_{-j})$$
$$\times\exp\left(-\frac{i}{2}\sum_{l=-1}^{1}((A_l+i\sqrt{2}x_{-l})^2+2x_l^2)\right) \tag{29}$$

with

$$V'=\left(\prod_{k=0,\pm 1}\int\exp\left(-\frac{i}{2}p_k^2\right)dp_k\right)\left(\prod_{l=0,\pm 1}\int\exp(-ix_l^2)dx_l\right). \tag{30}$$

It is useful to perform still another change of variables:

$$\tilde{A}_j=A_j+i\sqrt{2}x_{-j}. \tag{31}$$

Thus one can rewrite (29) as

$$-i\,\frac{1}{V'}\int\prod_{j=0,\pm1}d\tilde{A}_j\int\prod_{n=0,\pm1}dx_n\left(\det\left(\frac{\partial\tilde{A}_m}{\partial p_r}\right)\right)^{-1}$$

$$\times\sum_{s=1}^{3}\sum_{j,k=-1}^{1}\frac{R_{sk}^{-1}(\tilde{A}_k-i\sqrt{2}x_{-k})}{z-w_s}\,w^{1-j}\tilde{A}_j\,\exp\left(-\frac{i}{2}\sum_{l=-1}^{1}(\tilde{A}_l^2+2x_l^2)\right). \tag{32}$$

The integration over $\Pi_{j=-1}^{1}\,d\tilde{A}_j$ gives nonzero result only for $k=j$. Terms with $i\sqrt{2}x_{-k}$ drop out (they give rise to integral linear in $\tilde{A}_j$). The integration over $\Pi_{l=-1}^{1}\,dx_l$ is then trivial and one obtains

$$-\sum_{s=1}^{3}\sum_{j=-1}^{1}\frac{R_{sj}^{-1}}{z-w_s}\,w^{1-j}. \tag{33}$$

In order to get the numerical factor in (33) one has to use

$$\int_{\mathbf{R}}A^2\mathrm{e}^{-(i/2)A^2}dA=-i\int_{\mathbf{R}}\mathrm{e}^{-(i/2)A^2}dA. \tag{34}$$

The matrix R^{-1} is of the form:

$$R^{-1}=\frac{1}{(w_1-w_2)(w_1-w_3)(w_2-w_3)}\begin{pmatrix} w_2-w_3 & w_3^2-w_2^2 & w_2w_3(w_2-w_3)\\ w_3-w_1 & w_2^2-w_3^2 & w_1w_3(w_3-w_1)\\ w_1-w_2 & w_w^2-w_1^2 & w_1w_2(w_1-w_2)\end{pmatrix} \tag{35}$$

and the final result is (one has to include det $\partial\tilde{A}/\partial p$):

$$\begin{aligned}\langle 0|\gamma(z)\beta(w)\prod_{s=1}^{3}\delta(\beta(w_s))|0\rangle&=\frac{1}{(w_1-w_2)(w_1-w_3)(w_2-w_3)}\\ &\times\left(\frac{1}{z-w}-\frac{(w-w_2)(w-w_3)}{(z-w_1)(w_1-w_2)(w_1-w_3)}\right.\\ &\left.+\frac{(w-w_1)(w-w_3)}{(z-w_2)(w_1-w_2)(w_2-w_3)}-\frac{(w-w_1)(w-w_2)}{(z-w_3)(w_1-w_3)(w_2-w_3)}\right)\\ &=\frac{1}{z-w}\prod_{s=1}^{3}\frac{w-w_s}{z-w_s}\prod_{m<n}(w_m-w_n)^1.\end{aligned} \tag{36}$$

It is clear that the above expression has all the analytical properties that the propagator of the bosonic β–γ systems should possess.

III. THE PROPAGATOR ON ARBITRARY ALGEBRAIC CURVES

In this section the β–γ system with $\lambda\geqslant 1$ given by (1) is defined on arbitrary RS represented by means of the algebraic equation

$$F(z,y)=y^NP_N(z)+y^{N-1}P_{N-1}(z)+\cdots+P_0(z)=0, \tag{37}$$

where $P_j(z)$ are polynomials in z. y can be viewed either as single-valued functions on the RS or as multivalued function on the complex sphere. The monodromy properties of $y(z)$ then define the RS.

The classical degrees of freedom are identified as

$$\beta(z)=\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty}\beta_{s,j}z^{-j-\lambda}f_s(z), \tag{38}$$

$$\gamma(z)=\sum_{p=0}^{N-1}\sum_{k=-\infty}^{+\infty}\gamma_{p,k}z^{-k+\lambda-1}\phi_p(z), \tag{39}$$

where

$$f_k(z)=\frac{y^{N-1-k}(z)dz^{\lambda}}{(F_y(z,y(z)))^{\lambda}} \tag{40}$$

and

$$\phi_k(z)=\frac{dz^{1-\lambda}}{(F_y(z,y(z)))^{1-\lambda}}\,(y^k(z)+y^{k-1}(z)P_{N-1}(z)+\cdots+P_{N-k}(z)). \tag{41}$$

The basic commutation relations are

$$[\gamma_{p,k},\beta_{s,j}]=\delta_{s,p}\delta_{j+k,0}. \tag{42}$$

The price paid for the simple commutation relations (42) is the complicated definition of basis (41).

The normal ordering is introduced by assigning to the elementary excitations $\beta_{s,j}$ and $\gamma_{s,k}$ the property that they are creation or annihilation operators. Annihilation operators are those with $j\geqslant 1-\lambda$ and $k\geqslant\lambda$ and creation operators with $j\leqslant-\lambda-M_s$ and $k\leqslant\lambda-1$ (M_s is a number of zero modes in the s sector of the theory). The remaining operators $\beta_{s,j}$ with $-\lambda-M_s<j<1-\lambda$ correspond to zero modes.

From the above assignment it follows that

$$\gamma(z)\beta(w)=:\gamma(z)\beta(\psi):+\frac{1}{z-w}\sum_{n=0}^{N-1}f_n(z)\phi_n(w)=:\gamma(z)\beta(w):+K_{\lambda}(z,w). \tag{43}$$

It is an identity in radial ordered correlation functions. $K_{\lambda}(z,w)$ is the Weierstrass kernel (see Sec. I).

A simple implication of (43) is

$$\gamma(z)e^{ip\beta(w)}=:\gamma(z)e^{ip\beta(w)}:+ipK_{\lambda}(z,w)e^{ip\beta(w)}. \tag{44}$$

We wish to calculate the propagator for the β–γ system. In order to have finite correlators one has to insert a number M_{tot} of delta functions in the β fields, where M_{tot} is the total number of β zero modes:

$$M_{\mathrm{tot}}=\sum_{s=0}^{N-1}M_s=(g-1)(2\lambda-1). \tag{45}$$

The Wick theorem implies

$$\begin{aligned}&\langle 0|\gamma(z)\beta(w)\exp\left(i\sum_{r=1}^{M_{\mathrm{tot}}}p_r\beta(w_r)\right)|0\rangle\\&=K_{\lambda}(z,w)\langle 0|\exp\left(i\sum_{r=1}^{M_{\mathrm{tot}}}\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty}p_r\beta_{s,j}w_r^{-j-\lambda}f_s(w_r)\right)|0\rangle\\&\quad+\sum_{r=1}^{M_{\mathrm{tot}}}ip_rK_{\lambda}(z,w_r)\langle 0|\beta(w)\exp\left(i\sum_{r=1}^{M_{\mathrm{tot}}}\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty}p_t\beta_{s,j}w_t^{-j-\lambda}f_s(w_t)\right)|0\rangle.\end{aligned} \tag{46}$$

The propagator is obtained after integration over $\Pi_{r=1}^{M_{\mathrm{tot}}}(dp_r/2\pi)$.

As in Sec. II, the computations will be performed in the ''Gaussian'' representation for the β–γ system.[15] However, complications can be expected on a Riemann surface due to the appearance of M_s zero modes in the sector s of the theory.

A first attempt to accommodate these zero modes is to modify the way in which the $\beta_{s,j}$ and $\gamma_{p,k}$ excitations are realized:

$$\beta_{s,j}=-\frac{i}{\sqrt{2}}\left(x_{s,-j-\nu_s}-i\,\frac{\partial}{\partial x_{s,-j-\nu_s}}\right), \tag{47}$$

$$\gamma_{p,n}=\frac{1}{\sqrt{2}}\left(x_{p,n+\nu_p}+i\,\frac{\partial}{\partial x_{p,n+\nu_p}}\right), \tag{48}$$

where ν_s is to be related to the number M_s of zero modes in the s sector.

Let us suppose that the vacuum state is taken as

$$\Phi^{\bar{\kappa}}(x)=\exp\left(\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n\geqslant\kappa_s}x_{s,n}^2-\sum_{n<\kappa_s}x_{s,n}^2\right)\right), \tag{49}$$

where κ_s is again chosen in an appropriate way. Unfortunately, it turns out that nothing can be achieved in this way. The conditions that $\beta_{s,j}$ and $\gamma_{s,k}$ are annihilation operators only for $j\geqslant 1-\lambda$, $k\geqslant\lambda$ imply in fact that

$$\kappa_s=\lambda, \quad \nu_s=0. \tag{50}$$

We conclude that the only way to accommodate the generic structure of zero modes is to redefine the scalar product in an appropriate way. We shall come back to this point shortly.

In calculating (46) one follows most of steps described in Sec. II. The vacuum state is that introduced in (49) with $\kappa_s=\lambda$. One finds

$$\begin{aligned}\exp\left(i\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty}A_{s,j}\beta_{s,j}\right)|0\rangle&=\exp\left(i\sum_{s=0}^{N-1}\sum_{j\leqslant\lambda}A_{s,j}\beta_{s,j}\right)|0\rangle\\&=\exp\left(-\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n<\lambda}x_{s,n}^2+\sum_{j\leqslant-\lambda}((A_{s,j}+i\sqrt{2}x_{s,-j})^2+x_{s,-j}^2)\right)\right)\end{aligned} \tag{51}$$

with

$$A_{s,j}=\sum_{r=1}^{M_{\rm tot}}p_r w_r^{-j-\lambda}f_s(w_r). \tag{52}$$

The correct definition of $\langle 0|$ is

$$\langle 0|=\exp\left(\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n\geqslant\lambda+M_s}x_{s,n}^2-\sum_{n<\lambda+M_s}x_{s,n}^2\right)\right). \tag{53}$$

A possible way of thinking about $\langle 0|$ is that under the conjugation

$$x_{s,n}^{\dagger}=x_{s,-n+2\lambda+M_s-1}. \tag{54}$$

The conjugation on elementary excitations can be deduced from (47) and (48) to be

$$\beta_{s,n}^{\dagger}=-\beta_{s,-n-2\lambda-M_s+1}, \quad \gamma_{s,n}^{\dagger}=\gamma_{s,-n+2\lambda+M_s-1}. \tag{55}$$

This definition of conjugation is analogous to that obtained for the $b-c$ system on arbitrary RS.[18] For $g=0$ and $\lambda=-1$ (the situation discussed in Sec. II) $M_{\rm tot}=3$ and $\beta_n^\dagger=-\beta_{-n}$, $\gamma_n^\dagger=\gamma_{-n}$.

The first term in (46) is

$$\frac{1}{\mathbf{V}_\lambda}\int\prod_{s=0}^{N-1}\prod_{j=-\infty}^{+\infty}dx_{s,j}\int\prod_{r=1}^{M_{\rm tot}}dp_r\,\exp\left(\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n\geqslant\lambda+M_s}x_{s,n}^2-\sum_{n<\lambda+M_s}x_{s,n}^2\right)\right)$$
$$\times\exp\left(-\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n<\lambda}x_{s,n}^2+\sum_{j\geqslant\lambda}((A_{s,-j}+i\sqrt{2}x_{s,j})^2+x_{s,j}^2)\right)\right). \tag{56}$$

The normalization factor (formally infinite) $\mathbf{V}_\lambda$ is

$$\mathbf{V}_\lambda=\left(\int\prod_{s=0}^{N-1}\prod_{n<\lambda}\exp(-ix_{s,n}^2)dx_{s,n}\right)\left(\int\prod_{s=0}^{N-1}\prod_{n\geqslant\lambda+M_s}\exp(ix_{s,n}^2)dx_{s,n}\right)\left(\int\prod_{r=1}^{M_{\rm tot}}\right.$$
$$\left.\times\exp\left(-\frac{i}{2}p_r^2\right)dp_j\right)\left(\int\prod_{s=0}^{N-1}\prod_{n=\lambda}^{\lambda+M_s-1}\exp(-ix_{s,n}^2)dx_{s,n}\right). \tag{57}$$

One integrates first over $\Pi_{s=0}^{N-1}\,\Pi_{n<\lambda}\,dx_{s,n}$ and then over $\Pi_{s=0}^{N-1}\,\Pi_{n\geqslant\lambda+M_s}\,dx_{s,n}$. What remains is

$$\frac{1}{\mathbf{V}'_\lambda}\int\prod_{s=0}^{N-1}\prod_{n=\lambda}^{\lambda+M_s-1}dx_{s,n}\int\prod_{r=1}^{M_{\rm tot}}\exp\left(-\frac{i}{2}p_r^2\right)dp_j$$
$$\times\exp\left(-\frac{i}{2}\sum_{s=0}^{N-1}\sum_{n=\lambda}^{\lambda+M_s-1}((A_{s,-n}+i\sqrt{2}x_{s,n})^2+2x_{s,n}^2)\right) \tag{58}$$

with

$$\mathbf{V}'_\lambda=\left(\int\prod_{r=1}^{M_{\rm tot}}\exp\left(-\frac{i}{2}p_r^2\right)dp_j\right)\left(\int\prod_{s=0}^{N-1}\prod_{n=\lambda}^{\lambda+M_s-1}\exp(-ix_{s,n}^2)dx_{s,n}\right). \tag{59}$$

The number of integrations to be done over $dx_{s,n}$ is equal $M_{\rm tot}$. It is very convenient to introduce an index L which runs over all the allowed pairs $(s,n)L=1,...,M_{\rm tot}$. There is still minor technical complications due to the fact that in (58) expressions with $x_{s,n}$ are combined with $A_{s,-n}$. The following definition will be adopted: If the index J denotes a pair (s,n) then

$$x_J\equiv x_{s,n},\quad A_J\equiv A_{s,-n}. \tag{60}$$

A change of variables is to be done

$$p_j\to A_{r,-k}=\sum_{n=1}^{M_{\rm tot}}p_n w_n^{k-\lambda}f_r(w_n)=\sum_{n=1}^{M_{\rm tot}}p_n\Omega_L(w_n). \tag{61}$$

$\Omega_L(w)$ are β field zero modes, i.e., holomorphic λ differentials. Introducing a similar notation as in Sec. II one can write

$$A_L=\sum_{j=1}^{M_{\rm tot}}R_{Lj}p_j, \tag{62}$$

where

$$R_{Lj}=\begin{pmatrix} \Omega_1(w_1) & \Omega_1(w_2) & \cdots & \Omega_1(w_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_1) & \Omega_{M_{\rm tot}}(w_2) & \cdots & \Omega_{M_{\rm tot}}(w_{M_{\rm tot}}) \end{pmatrix} \tag{63}$$

is a matrix of zero modes. The result is that only the Jacobian arising due to the change of variables (61) gives nontrivial contribution. The remaining integrations cancel the normalization factor $\mathbf{V}'_\lambda$ and the first term in (46) is

$$\frac{K_\lambda(z,w)}{\det\,\Omega_L(w_j)}. \tag{64}$$

In order to calculate the next term in (46) one should follow the procedure presented in Sec. II with the above-explained modifications due to the presence of N sectors and M_s zero modes in each of them. The result is

$$-\det(\Omega_L(w_j))\sum_{s,N=1}^{M_{\rm tot}} K_\lambda(z,w_s)R^{-1}_{sN}\Omega_N(w). \tag{65}$$

The final result can be written in compact form

$$\langle 0|\,\gamma(z)\beta(w)\prod_{r=1}^{M_{\rm tot}}\delta(\beta(w_r))|0\rangle=\det\begin{pmatrix} K_\lambda(z,w) & K_\lambda(z,w_1) & \cdots & K_\lambda(z,w_{M_{\rm tot}}) \\ \Omega_1(w) & \Omega_1(w_1) & \cdots & \Omega_1(w_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w) & \Omega_{M_{\rm tot}}(w_1) & \cdots & \Omega_{M_{\rm tot}}(w_{M_{\rm tot}}) \end{pmatrix} \times(\det\,\Omega_L(w_j))^{-2}. \tag{66}$$

It has all the necessary analytical properties of the propagator: poles as $z\to w$, $z\to w_r$, $w_{r_1}\to w_{r_2}$ $(r_1\neq r_2)$ and zeroes as $w\to w_j$.

One can also define

$$\frac{\langle 0|\,\gamma(z)\beta(w)\Pi_{r=1}^{M_{\rm tot}}\delta(\beta(w_r))|0\rangle}{\langle 0|\Pi_{r=1}^{M_{\rm tot}}\delta(\beta(w_r))|0\rangle}=\det\begin{pmatrix} K_\lambda(z,w) & K_\lambda(z,w_1) & \cdots & K_\lambda(z,w_{M_{\rm tot}}) \\ \Omega_1(w) & \Omega_1(w_1) & \cdots & \Omega_1(w_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w) & \Omega_{M_{\rm tot}}(w_1) & \cdots & \Omega_{M_{\rm tot}}(w_{M_{\rm tot}}) \end{pmatrix} \times(\det\,\Omega_L(w_j))^{-1}. \tag{67}$$

IV. THE FOUR-POINT CORRELATOR

The operator formalism also allows the calculation of higher order correlation functions. In this section the computation of the four-point correlator

$$\langle 0|\,\gamma(z_1)\gamma(z_2)\beta(w_1)\beta(w_2)\prod_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle \tag{68}$$

is presented. The Wick theorem implies

$$\langle 0|\,\gamma(z_1)\gamma(z_2)\beta(w_1)\beta(w_2)\prod_{s=1}^{M_{\rm tot}}\exp(ip_s\beta(u_s))|0\rangle$$

$$=(K_\lambda(z_1,w_1)K_\lambda(z_2,w_2)+K_\lambda(z_1,w_2)K_\lambda(z_2,w_1))\langle 0|\prod_{s=1}^{M_{\rm tot}}\exp(ip_s\beta(u_s))|0\rangle$$

$$+K_\lambda(z_1,w_1)\sum_{t=1}^{M_{\rm tot}} ip_tK_\lambda(z_2,u_t)\langle 0|\beta(w_2)\prod_{s=1}^{M_{\rm tot}} \exp(ip_s\beta(u_s))|0\rangle$$

$$+K_\lambda(z_1,w_2)\sum_{t=1}^{M_{\rm tot}} ip_tK_\lambda(z_2,u_t)\langle 0|\beta(w_1)\prod_{s=1}^{M_{\rm tot}} \exp(ip_s\beta(u_s))|0\rangle$$

$$+K_\lambda(z_2,w_1)\sum_{t=1}^{M_{\rm tot}} ip_tK_\lambda(z_1,u_t)\langle 0|\beta(w_2)\prod_{s=1}^{M_{\rm tot}} \exp(ip_s\beta(u_s))|0\rangle$$

$$+K_\lambda(z_2,w_2)\sum_{t=1}^{M_{\rm tot}} ip_tK_\lambda(z_1,u_t)\langle 0|\beta(w_1)\prod_{s=1}^{M_{\rm tot}} \exp(ip_s\beta(u_s))|0\rangle$$

$$-\sum_{t_1,t_2=1}^{M_{\rm tot}} p_{t_1}p_{t_2}K_\lambda(z_1,u_{t_1})K_\lambda(z_2,u_{t_2})\langle 0|\beta(w_1)\beta(w_2)\prod_{s=1}^{M_{\rm tot}} \exp(ip_s\beta(u_s))|0\rangle. \quad (69)$$

Only the last term in (69) requires a new analysis.
It is necessary to compute

$$\beta(w_1)\beta(w_2)\exp\left(i\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty} A_{s,j}\beta_{s,j}\right)|0\rangle$$

$$=-\frac{1}{2}\sum_{r,s=0}^{N-1}\sum_{j,k=-\infty}^{+\infty}\Psi_{r,-j}\Psi_{s,-k}\left(x_{r,j}-i\,\frac{\partial}{\partial x_{r,j}}\right)\left(x_{s,k}-i\,\frac{\partial}{\partial x_{s,k}}\right)\exp\left(-\frac{i}{2}\sum_{s'=0}^{N-1}\sum_{n'<\lambda}x^2_{s',n'}\right)$$

$$\times\exp\left(-\frac{i}{2}\sum_{q=0}^{N-1}\sum_{m\geqslant\lambda}((A_{q,-m}+i\sqrt{2}x_{q,m})^2+x^2_{q,m})\right)$$

$$=\sum_{r,s=0}^{N-1}\sum_{j,k\geqslant\lambda}\Psi_{r,-j}\Psi_{s,-k}((A_{r,-j}+i\sqrt{2}x_{r,j})(A_{s,-k}+i\sqrt{2}x_{s,k})+i\,\delta_{jk})$$

$$\times\exp\left(-\frac{i}{2}\sum_{s'=0}^{N-1}\sum_{n'<\lambda}x^2_{s',n'}\right)\exp\left(-\frac{i}{2}\sum_{q=0}^{N-1}\sum_{m\geqslant\lambda}((A_{q,m}+i\sqrt{2}x_{q,-m})^2+x^2_{q,-m})\right), \quad (70)$$

where

$$\beta(w)=\sum_{s=0}^{N-1}\sum_{j=-\infty}^{+\infty}\Psi_{s,j}\beta_{s,j}, \quad \Psi_{s,j}(w)=f_s(w)w^{-j-\lambda} \quad (71)$$

so that

$$A_{r,j}=\sum_{s=1}^{M_{\rm tot}} p_s\Psi_{r,j}(u_s). \quad (72)$$

In the correlator

$$\langle 0|\beta(w_1)\beta(w_2)\exp\left(i\sum_{r=1}^{M_{\rm tot}} p_r\beta(u_r)\right)|0\rangle$$

$$=\frac{1}{\mathbf{V}_\lambda}\int\prod_{s=0}^{N-1}\prod_{j=-\infty}^{+\infty}dx_{s,j}\int\prod_{r=1}^{M_{\rm tot}}dp_r\,\exp\left(\frac{i}{2}\sum_{t=0}^{N-1}\left(\sum_{n\geqslant\lambda+M_s}x_{s,n}^2-\sum_{n<\lambda+M_s}x_{s,n}^2\right)\right)$$
$$\times\sum_{r,s=1}^{N-1}\sum_{j,k\geqslant\lambda}\Psi_{r,-j}\Psi_{s,-k}((A_{r,-j}+i\sqrt{2}x_{r,j})(A_{s,-k}+i\sqrt{2}x_{s,k})+i\,\delta_{j,k}\delta_{r,s})$$
$$\times\exp\left(-\frac{i}{2}\sum_{s=0}^{N-1}\left(\sum_{n<\lambda}x_{s,n}^2+\sum_{j\geqslant\lambda}((A_{s,-j}+i\sqrt{2}x_{s,j})^2+x_{s,j}^2)\right)\right) \tag{73}$$

the integrations over $dx_{s,n}$ for $n<\lambda$ are trivial. In the integrations over $dx_{s,n}$ for $n\geqslant\lambda+M_s$ a crucial observation is that in the sum over $j,\,k>\lambda$ the only terms that survive are those for which $j,\,k\leqslant\lambda+M_s-1$. The reason is that (34) holds. One obtains therefore

$$\langle 0|\beta(w_1)\beta(w_2)\exp\left(i\sum_{r=1}^{M_{\rm tot}} p_r\beta(u_r)\right)$$

$$=\frac{1}{\mathbf{V}'_\lambda}\int\prod_{s=0}^{N-1}\prod_{n=\lambda}^{\lambda+M_s-1}dx_{s,n}\int\prod_{r=1}^{M_{\rm tot}}dp_r\sum_{r,s=1}^{N-1}\sum_{j,k=\lambda}^{\lambda+M_s-1}\Psi_{r,-j}\Psi_{s,-k}((A_{r,-j}+i\sqrt{2}x_{r,j})$$
$$\times(A_{s,-k}+i\sqrt{2}x_{s,k})+i\,\delta_{j,k}\delta_{r,s})$$
$$\times\exp\left(-\frac{i}{2}\sum_{s=0}^{N-1}\sum_{n=\lambda}^{\lambda+M_s-1}((A_{s,-n}+i\sqrt{2}x_{s,n})^2+2x_{s,n}^2)\right). \tag{74}$$

At this point it is very convenient to rewrite the above expression using the indices L introduced in Sec. III.

$$\langle 0|\beta(w_1)\beta(w_2)\exp\left(i\sum_{r=1}^{M_{\rm tot}} p_r\beta(u_r)\right)|0\rangle$$

$$=\frac{1}{\mathbf{V}'_\lambda}\int\prod_{J=1}^{M_{\rm tot}}\int\prod_{r=1}^{M_{\rm tot}}dp_r\sum_{K,L=1}^{M_{\rm tot}}\Psi_K\Psi_L((A_K+i\sqrt{2}x_K)$$
$$\times(A_L+i\sqrt{2}x_L)+i\,\delta_{K,L})\exp\left(-\frac{i}{2}\sum_{J=1}^{M_{\rm tot}}((A_J+i\sqrt{2}x_J^2+2x_J^2)\right). \tag{75}$$

One should not forget that in (69) this expression is multiplied by $(-)p_{t_1}p_{t_2}$. Performing the standard changes of variables $p_r\to A_L$ and $A_J\to\tilde{A}_J=A_J+i\sqrt{2}x_J$, the remaining integrals are

$$\frac{1}{\mathbf{V}'_\lambda}\int\prod_{K=1}^{M_{\rm tot}}dx_K\int\prod_{J=1}^{M_{\rm tot}}d\tilde{A}_J\,\exp\left(-\frac{i}{2}\sum_{J=1}^{M_{\rm tot}}(\tilde{A}_J^2+2x_J^2)\right)R^{-1}_{t_1M}R^{-1}_{t_2N}(\tilde{A}_M-i\sqrt{2}x_M)$$
$$\times(\tilde{A}_N-i\sqrt{2}x_N)\sum_{K,L=1}^{M_{\rm tot}}\Psi_K\Psi_L(\tilde{A}_K\tilde{A}_L+i\,\delta_{K,L}). \tag{76}$$

One can replace $\Psi_J(z)$ with $\Omega_J(z)$ as for the allowed values of J only the zero modes appear in the above expression, which we denote with the symbol Ω_J.

The integrals to be evaluated are

$$\int \prod_{Q=1}^{M_{\rm tot}} d\tilde{A}_Q(\tilde{A}_M - i\sqrt{2}x_M)(\tilde{A}_N - i\sqrt{2}x_N)(\tilde{A}_K\tilde{A}_L + i\,\delta_{K,L})\exp\left(-\frac{i}{2}\sum_{S=1}^{M_{\rm tot}}\tilde{A}_S^2\right). \tag{77}$$

It is clear that terms with x_M drop out. Terms linear in x_M give rise to terms linear or of the third order in $\tilde{A}_P$ which yield no contribution after integration. Also the quadratic term containing $x_M x_N$ gives rise to integral with $\tilde{A}_K\tilde{A}_L + i\,\delta_{K,L}$ which integrated gives 0 because of (34).

The tensor structure of (77) implies that it must be of the form

$$\text{const}(\,\delta_{N,K}\delta_{M,L} + \delta_{N,L}\delta_{M,K}) \int \prod_{S=1}^{M_{\rm tot}} d\tilde{A}_S \exp\left(-\frac{i}{2}\sum_{S=1}^{M_{\rm tot}}\tilde{A}_S^2\right).$$

Using

$$\int dp p^4 \exp\left(-\frac{i}{2}p^2\right) = -3\int dp \exp\left(-\frac{i}{2}p^2\right) \tag{78}$$

and

$$\int dp p^2 \exp\left(-\frac{i}{2}p^2\right) = -i\int dp \exp\left(-\frac{i}{2}p^2\right) \tag{79}$$

it is easy to verify that const$=-1$.

After performing the remaining trivial integrations the final expression for the last term in (69) is (apart from the Jacobian $\det(\partial A_J/p_r)$)

$$\sum_{t_1,t_2=1}^{M_{\rm tot}} K_\lambda(z_1,u_{t_1})K_\lambda(z_2,u_{t_2})(R^{-1}_{t_1L}\Omega_L(w_2)R^{-1}_{t_2K}\Omega_K(w_1) + R^{-1}_{t_1K}\Omega_K(w_1)R^{-1}_{t_2L}\Omega_L(w_2)). \tag{80}$$

Altogether (69) is equal to

$$\begin{aligned}
&\frac{1}{\det\,\Omega_I(u_J)}\,(K_\lambda(z_1,w_1)K_\lambda(z_2,w_2) + K_\lambda(z_1,w_2)K_\lambda(z_2,w_1)\\
&\quad + -\sum_{t=1}^{M_{\rm tot}}\sum_{K=1}^{M_{\rm tot}}(K_\lambda(z_1,w_1)K_\lambda(z_2,u_t)R^{-1}_{tK}\Omega_K(z_2) + K_\lambda(z_1,w_2)K_\lambda(u_t,w_1)R^{-1}_{tK}\Omega_K(z_2)\\
&\quad + K_\lambda(z_2,w_1)K_\lambda(u_t,w_2)R^{-1}_{tK}\Omega_K(z_1) + K_\lambda(z_1,w_1)K_\lambda(u_t,w_2)R^{-1}_{tK}\Omega_K(z_1))\\
&\quad + \sum_{t_1,t_2=1}^{M_{\rm tot}}\sum_{L,K=1}^{M_{\rm tot}}(K_\lambda(u_{t_1},w_1)K_\lambda(u_{t_2},w_2)R^{-1}_{t_1L}\Omega_L(z_2)R^{-1}_{t_2K}\Omega_K(z_1))\\
&\quad + K_\lambda(u_{t_1},w_1)K_\lambda(u_{t_2},w_2)R^{-1}_{t_1K}\Omega_K(z_1)R^{-1}_{t_2}\Omega_L(z_2)).
\end{aligned} \tag{81}$$

With manipulations analogous to those used in the derivation of (65) it is possible to write down (81) in a compact form. The details are given in Appendix B. The four-point correlator is

$$\begin{aligned}
&\frac{\langle 0|\beta(z_1)\beta(z_2)\gamma(w_1)\gamma(w_2)\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}{\langle 0|\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}\\
&\quad = \frac{\langle 0|\beta(z_1)\gamma(w_1)\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}{\langle 0|\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}\,\frac{\langle 0|\beta(z_2)\gamma(w_2)\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}{\langle 0|\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}\\
&\qquad + \frac{\langle 0|\beta(z_1)\gamma(w_2)\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}{\langle 0|\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}\,\frac{\langle 0|\beta(z_2)\gamma(w_1)\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}{\langle 0|\Pi_{s=1}^{M_{\rm tot}}\delta(\beta(u_s))|0\rangle}.
\end{aligned} \tag{82}$$

APPENDIX A: GAUSSIAN REPRESENTATION

The aim of this Appendix is to provide a short but self-consistent presentation of the construction of ''Gaussian'' representation for the β–γ system proposed in Ref. 15. The notation corresponds to that used in Sec. II where the RS is a complex sphere.

The elementary excitations for β and γ satisfy the commutation relations:

$$[\gamma_n, \beta_m] = \delta_{n+m,0}. \tag{A1}$$

They are represented as

$$\beta_n = -\frac{i}{\sqrt{2}}\left(x_{-n} - i\,\frac{\partial}{\partial x_{-n}}\right), \tag{A2}$$

$$\gamma_n = \frac{1}{\sqrt{2}}\left(x_n + i\,\frac{\partial}{\partial x_n}\right). \tag{A3}$$

The operators β_n and γ_m act on functions of x_n. The vector space they act on is a product of ''Gaussian'' factors $\exp(\pm(i/2)x_n^2)$ multiplied by polynomials in x_m. In this vector ''Fock space'' the scalar product is defined in the following way:

$$\langle \Phi_1 | \Phi_2 \rangle \equiv \frac{1}{V_\lambda} \int \prod_n dx_n (\Phi_1(x))^* \Phi_2(x). \tag{A4}$$

It is assumed that $x_n^* = x_{-n}$ (but in the case of general RS this rule will have to be modified). V_λ is a formally infinite normalization factor equal to

$$V_\lambda = \left(\frac{\int dp\, \exp\left(-\frac{i}{2}p^2\right)}{2\pi}\right)^{2\lambda-1} \left(\prod_{m<-\lambda} \int \exp(-ix_m^2)dx_m\right)\left(\prod_{n\geqslant 1-\lambda} \int \exp(ix_n^2)dx_n\right). \tag{A5}$$

It is understood that (A4) is defined in such a way that the two formally infinite terms in numerator and denominator cancel each other.

A choice of the vaccum state is a choice of ''Bose sea level.'' A vacuum state $\Phi^{(\lambda)}$ can be constructed such that

$$\langle \Phi^{(\lambda)} | \Phi^{(\lambda)} \rangle = \infty \tag{A6}$$

(the explanation for that infinity can be found in Sec. I) but (assume that $\lambda \geqslant 2$)

$$\langle \Phi^{(\lambda)} | \prod_{j=1-\lambda}^{\lambda-1} \delta(\gamma_j) | \Phi^{(\lambda)} \rangle = 1. \tag{A7}$$

The correct definition is

$$\Phi^\lambda(x) = \exp\frac{i}{2}\left(\sum_{n\geqslant\lambda} x_n^2 - \sum_{n<\lambda} x_n^2\right). \tag{A8}$$

It satisfies

$$(\Phi^{(\lambda)}(x))^* = \Phi^{1-\lambda}(x) \tag{A9}$$

so that

$$\Phi_0^{(\lambda)}(\Phi_0^{(\lambda)})^* = \exp i\left(\sum_{n\geqslant\lambda} x_n^2 - \sum_{n<-\lambda} x_n^2\right). \tag{A10}$$

The integration over $dx_{\lambda-1}\cdots dx_{1-\lambda}$ in (A4) lead to (A6).
On the other hand,

$$\begin{aligned}\delta(\gamma_n)\exp\left(-\frac{i}{2}x_n^2\right)&=\frac{1}{2\pi}\int dp\ \exp(ip\gamma_n)\exp\left(-\frac{i}{2}x_n^2\right)\\&=\frac{1}{2\pi}\int dp\ \exp\left(\frac{i}{\sqrt{2}}p\left(x_n+i\frac{\partial}{\partial x_n}\right)\right)\exp\left(-\frac{i}{2}x_n^2\right)\\&=\frac{1}{2\pi}\int dp\ \exp\left(\frac{i}{2}x_n^2-\frac{i}{2}(p-i\sqrt{2}x_n)^2\right)\\&=\frac{1}{2\pi}\int dp\ \exp\left(-\frac{1}{2}p^2\right)\exp\left(\frac{i}{2}x_n^2\right).\end{aligned}\tag{A11}$$

Therefore, thanks to the appropriate definition of the normalization factor V_λ, (A7) holds.
An important identity used above is

$$\exp\left(a\frac{\partial}{\partial x_{-n}}\right)\exp\left(\frac{i}{2}x_{-n}^2\right)=\exp\left(iax_{-n}+\frac{i}{2}a^2\right)\exp\left(\frac{i}{2}x_{-n}^2\right).\tag{A12}$$

APPENDIX B: DERIVATION OF (82)

The aim of this Appendix is to provide a proof that the four-point correlation function can be represented by (82). The starting point is (81) and the basic trick is that used in the derivation of (65):

$$\sum_{s,N=1}^{M_{\rm tot}}K_\lambda(z,w_s)R_{sN}^{-1}\Omega_N(w)=K_\lambda(z,w)-\det\begin{pmatrix}K_\lambda(z,w)&K_\lambda(z,w_1)&\cdots&K_\lambda(z,w_{M_{\rm tot}})\\\Omega_1(w)&\Omega_1(w_1)&\cdots&\Omega_1(w_{M_{\rm tot}})\\\cdots&\cdots&\cdots&\cdots\\\Omega_{M_{\rm tot}}(w)&\Omega_{M_{\rm tot}}(w_1)&\cdots&\Omega_{M_{\rm tot}}(w_{M_{\rm tot}})\end{pmatrix}\times(\det\Omega_L(w_j))^{-1}.\tag{B1}$$

Thanks to (B1) all the summations in (81) can be eliminated. One obtains (everything has to be multiplied by $(1/\det\Omega_I(u_J))$

$$\begin{aligned}&K_\lambda(z_1,w_1)K_\lambda(z_2,w_2)+K_\lambda(z_1,w_2)K_\lambda(z_2,w_1)+-K_\lambda(z_1,w_1)K_\lambda(z_2,w_2)\\&+\frac{1}{(\det\Omega_I(u_J))}K_\lambda(z_1,w_1)\det\begin{pmatrix}K_\lambda(z_2,w_2)&K_\lambda(z_2,u_1)&\cdots&K_\lambda(z_2,u_{M_{\rm tot}})\\\Omega_1(w_2)&\Omega_1(u_1)&\cdots&\Omega_1(u_{M_{\rm tot}})\\\cdots&\cdots&\cdots&\cdots\\\Omega_{M_{\rm tot}}(w_2)&\Omega_{M_{\rm tot}}(u_1)&\cdots&\Omega_{M_{\rm tot}}(u_{M_{\rm tot}})\end{pmatrix}\\&-K_\lambda(z_1,w_2)K_\lambda(z_2,w_1)\\&+\frac{1}{(\det\Omega_I(u_J))}K_\lambda(z_1,w_2)\det\begin{pmatrix}K_\lambda(z_2,w_1)&K_\lambda(z_2,u_1)&\cdots&K_\lambda(z_2,u_{M_{\rm tot}})\\\Omega_1(w_1)&\Omega_1(u_1)&\cdots&\Omega_1(u_{M_{\rm tot}})\\\cdots&\cdots&\cdots&\cdots\\\Omega_{M_{\rm tot}}(w_1)&\Omega_{M_{\rm tot}}(u_1)&\cdots&\Omega_{M_{\rm tot}}(u_{M_{\rm tot}})\end{pmatrix}\\&-K_\lambda(z_2,w_1)K_\lambda(z_1,w_2)\end{aligned}$$

$$+\frac{1}{(\det\,\Omega_I(u_J))}\,K_\lambda(z_2,w_1)\det\begin{pmatrix} K_\lambda(z_1,w_2) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_2) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}$$

$$-K_\lambda(z_2,w_2)K_\lambda(z_1,w_1)$$

$$+\frac{1}{(\det\,\Omega_I(u_J))}\,K_\lambda(z_2,w_2)\det\begin{pmatrix} K_\lambda(z_1,w_1) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_1) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_1) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}$$

$$+\left(K_\lambda(z_1,w_2)-\frac{1}{(\det\,\Omega_I(u_J))}\det\begin{pmatrix} K_\lambda(z_1,w_2) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_2) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right)$$

$$\times\left(K_\lambda(z_2,w_1)-\frac{1}{(\det\,\Omega_I(u_J))}\det\begin{pmatrix} K_\lambda(z_2,w_1) & K_\lambda(z_2,u_1) & \cdots & K_\lambda(z_2,u_{M_{\rm tot}}) \\ \Omega_1(w_1) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_1) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right)$$

$$+\left(K_\lambda(z_1,w_1)-\frac{1}{(\det\,\Omega_I(u_J))}\det\begin{pmatrix} K_\lambda(z_1,w_1) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_1) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right)$$

$$\times\left(K_\lambda(z_2,w_2)-\frac{1}{(\det\,\Omega_I(u_J))}\det\begin{pmatrix} K_\lambda(z_2,w_2) & K_\lambda(z_2,u_1) & \cdots & K_\lambda(z_2,u_{M_{\rm tot}}) \\ \Omega_1(w_2) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right). \tag{B2}$$

All except two terms cancel each other and (81) turns out to be equal

$$\frac{1}{(\det\,\Omega_I(u_J))^3}\left(\det\begin{pmatrix} K_\lambda(z_1,w_2) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_2) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right)$$

$$\times\left(\det\begin{pmatrix} K_\lambda(z_2,w_1) & K_\lambda(z_2,u_1) & \cdots & K_\lambda(z_2,u_{M_{\rm tot}}) \\ \Omega_1(w_1) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_1) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix}\right)$$

$$+\left\langle \det\begin{pmatrix} K_\lambda(z_1,w_1) & K_\lambda(z_1,u_1) & \cdots & K_\lambda(z_1,u_{M_{\rm tot}}) \\ \Omega_1(w_1) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix} \right\rangle$$

$$\times\left\langle \det\begin{pmatrix} K_\lambda(z_2,w_2) & K_\lambda(z_2,u_1) & \cdots & K_\lambda(z_2,u_{M_{\rm tot}}) \\ \Omega_1(w_2) & \Omega_1(u_1) & \cdots & \Omega_1(u_{M_{\rm tot}}) \\ \cdots & \cdots & \cdots & \cdots \\ \Omega_{M_{\rm tot}}(w_2) & \Omega_{M_{\rm tot}}(u_1) & \cdots & \Omega_{M_{\rm tot}}(u_{M_{\rm tot}}) \end{pmatrix} \right\rangle . \tag{B3}$$

By looking at (67) one immediately gets (82).

[1] P. Di Francesco, P. Mathieu, and D. Sénéchal, *Conformal Field Theory* (Springer, Berlin, 1996).
[2] A. M. Polyakov, Phys. Lett. B **103**, 207 (1981); **103**, 207 (1981).
[3] J. Cardy, Nucl. Phys. B **270**, 186 (1986).
[4] L. Bonora, A. Lugo, M. Matone, and J. Russo, Commun. Math. Phys. **123**, 329 (1989); C. Vafa, Phys. Lett. B **190**, 47 (1987); L. Alvarez-Gaumé, C. Gomez, G. Moore, and C. Vafa, Nucl. Phys. B **303**, 455 (1988); A. M. Semikhatov, Phys. Lett. B **212**, 357 (1988).
[5] E. Verlinde and H. Verlinde, Nucl. Phys. B **288**, 357 (1987).
[6] A. K. Raina, Helv. Phys. Acta **63**, 694 (1990); Commun. Math. Phys. **140**, 373 (1991).
[7] F. Ferrari and J. Sobczyk, Int. J. Mod. Phys. A **11**, 2213 (1996).
[8] F. Ferrari and J. Sobczyk, J. Geom. Phys. **19**, 287 (1996).
[9] D. Friedan, E. Martinec, and S. Shenker, Nucl. Phys. B **271**, 93 (1986).
[10] J. Atick and A. Sen, Nucl. Phys. B **293**, 317 (1987).
[11] E. Verlinde and H. Verlinde, Phys. Lett. B **192**, 95 (1987).
[12] A. Losev, Phys. Lett. B **226**, 62 (1989); O. Lechtenfeld, *ibid.* **232**, 193 (1989); P. di Vecchia, *ibid.* **248**, 329 (1990); U. Carow-Watamura, Z. F. Ezawa, K. Harada, A. Tezuku, and S. Watamura, *ibid.* **227**, 73 (1989).
[13] D. Montano, Nucl. Phys. B **297**, 125 (1988); E. Gava (unpublished); E. Gava, R. Iengo, and G. Sotkov, Phys. Lett. B **207**, 283 (1988); D. Lebedev and A. Morozov, Nucl. Phys. B **302**, 163 (1988); R. Iengo and C-J Zhu, Phys. Lett. B **212**, 313 (1988).
[14] A. Gerasimov *et al.*, Int. J. Mod. Phys. A **5**, 2495 (1990).
[15] L. Alvarez-Gaumé, C. Gomez, P. Nelson, G. Sierra, and C. Vafa, Nucl. Phys. B **311**, 333 (1988).
[16] C. Crnkovic, G. M. Sotkov, and M. Stanishkov, Phys. Lett. B **220**, 397 (1989).
[17] S. A. Apikyan and C. J. Efthimiou, Report, hep-th/9610051.
[18] F. Ferrari and J. T. Sobczyk, Report, hep-th/9709162.

Application of the discrete Wentzel–Kramers–Brillouin method to spin tunneling

Anupam Garg[a)]
Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60208

(Received 9 December 1997; accepted for publication 17 June 1998)

A discrete version of the Wentzel–Kramers–Brillouin (WKB) method is developed and applied to calculate the tunnel splittings between classically degenerate states of spin Hamiltonians. The results for particular model problems are in complete accord with those previously found using instanton methods. The discrete WKB method is more elementary and also yields wave functions. © *1998 American Institute of Physics.* [S0022-2488(98)03710-4]

I. INTRODUCTION

A large variety of problems in physics can be modeled in terms of a spin of large magnitude that is then described by a Hamiltonian which is a polynomial in the spin components. It is natural in many of these cases to seek a semiclassical description. This is especially so when the system possesses a number of degenerate states in the classical limit, and one wishes to see how this degeneracy is lifted by quantum tunneling. Examples include the Lipkin–Meshov–Glick model in nuclear physics,[1–3] the rotational spectrum of polyatomic molecules,[4] and the Hamamoto–Mottleson model[5] for the $\Delta I=4$ staggering of the rotational spectra of certain superdeformed nuclei.

Perhaps the simplest example of such a problem is provided by the Hamiltonian

$$\mathscr{H}=-k_1S_z^2+k_2S_x^2, \tag{1.1}$$

where $k_1>0$, $k_2>0$, and S_α $(\alpha=x,y,z)$ are components of the spin operator $\mathbf{S}$ obeying the usual commutation rules

$$[S_\alpha,S_\beta]=i\epsilon_{\alpha\beta\gamma}S_\gamma. \tag{1.2}$$

Further, the spin has a magnitude S, i.e., the operator $\mathbf{S}\cdot\mathbf{S}$ has eigenvalue $S(S+1)$. As $S\to\infty$, the system is more and more classical, and in the limit it has two degenerate ground states corresponding to $S_z/S=\pm1$. For large but finite S, the quantal analogs of these states will be admixed by a tunneling matrix element $\Delta_0/2$. The two lowest energy eigenstates will then be split by an amount Δ_0. The goal of a tunneling calculation is to find an expression for Δ_0 that is asymptotically valid as $S\to\infty$. In analogy with the result for particles with a position like coordinate in a potential well, we expect this expression to take the form

$$\Delta_0=e^{-(B_0+B_1+B_2+\cdots)}, \tag{1.3}$$

where

$$B_0\gg B_1\gg B_2\gg\cdots \tag{1.4}$$

in the sense that B_1/B_0, $B_2/B_1,\ldots\to0$ as $S\to\infty$. By looking at the commutation relations for the scaled operators S_α/S, we can see that $1/S$ plays the same role as $\hbar$ in conventional Wentzel–Kramers–Brillouin (WKB) theory, and we thus expect $B_0=O(S)$. As is often done for a particle in a well, it is convenient to call e^{-B_0} the Gamow factor, or the WKB exponent, and the remaining terms $e^{-(B_1+B_2+\cdots)}$ the prefactor.

a)Electronic address: agarg@nwu.edu

 © 1998 American Institute of Physics

In many physical problems, it is only sensible or possible to calculate Δ_0 to ''exponential accuracy,'' i.e., to find only the leading term B_0 in Eq. (1.3). As a general point of formalism, however, it is clearly desirable to have a proper asymptotic approximation Δ_0' to Δ_0 itself and not just to $\ln \Delta_0$. In other words, we demand that $\Delta_0'/\Delta_0 \to 1$ as $S \to \infty$. For this one must know all B_n that are of $O(S^0)$ or greater. The first correct calculation to such accuracy that we know of was done by Enz and Schilling,[6,7] who mapped the spin onto a canonically conjugate variable pair via a Villain transformation, and then applied an instanton method[8] to the resulting problem. Great care must be taken in this approach with operator ordering problems, leading to a rather complex calculation. A more natural approach is one based on Klauder's spin-coherent-state path integral.[9] A naive application of instanton methods for the tunnel splitting to this path integral,[10] however, only gives the leading S dependence correctly. In particular, terms of $O(S^0)$ in $\ln \Delta_0$ are incorrectly obtained.[11,12] A correct calculation within this formalism has only recently been done by Belinicher, Providencia, and Providencia,[13] who show that it is necessary to consider nondifferentiable paths by a careful examination of the discrete time version of the path integral. Once again, the simplicity of the instanton approach is lost.

It is therefore desirable to have a means of calculating the tunnel splitting using only elementary methods of analysis. Such an approach is provided by the discrete WKB method. (An alternative approach, used in Ref. 14, also meets the criterion of being elementary.) In the present context, the method consists of writing the Schrödinger equation as a recursion relation for the expansion coefficients of an energy eigenstate in the S_z basis, and solving this recursion relation under the assumption that the coefficients in the recursion relation vary slowly with m, the S_z eigenvalue. By and large, the method is a straightforward generalization of the continuous WKB method,[15] and it may also be viewed as equivalent to the use of semiclassical dynamics for Bloch electrons.

The earliest appearance of this method in a physics context of which we are aware is by Schulten and Gordon,[16] who use it to find semiclassical approximations for the $3j$ (Wigner) and $6j$ (Racah) symbols. A cogent review of the method in the context of atomic physics has been given by Braun,[17] who also cites its application in a variety of other physics settings, as well as earlier discussions in Russian mathematics texts. It has also been previously applied to spin tunneling problems by van Hemmen and Sütő.[18,19] These authors have not, however, utilized the method to its full potential. Specifically, they calculate Δ_0 for a class of Hamiltonians including

$$\mathscr{H} = -\gamma S_z^2 - \alpha S_x, \tag{1.5}$$

but only the leading S dependence is correctly found. One of the goals of this paper is to show that the prefactor can also be found correctly to order S^0 with little extra labor. (See Sec. V.)

We will illustrate the discrete WKB method by applying it to Eqs. (1.1) and (1.5). We will find the splittings Δ_k, $k=0,1,2,\ldots$ for all low lying pairs of energy levels for Eq. (1.1), and the lowest splitting Δ_0 for Eq. (1.5). Our answers for Δ_0 agree completely with those of Enz and Schilling,[7] or Belinicher, Providencia, and Providencia,[13] and we believe they are obtained with much less effort. While the discrete WKB method does not have the nice geometrical structure of the spin-coherent-state path integral approach, it provides compensation in that it yields wave functions in the course of the calculation, which may be of further use; for evaluating matrix elements of perturbations, see, e.g., Refs. 20 and 21. Other problems commonly solved by the continuous WKB method, such as escape from a metastable well, are also amenable to the method, but we do not discuss these here.

The plan of the paper is as follows. In Sec. II we briefly review the discrete WKB method. Our discussion is heuristic and physically motivated. See Braun[17] for one that is more formal. In Sec. III we derive a formula [see Eq. (3.11)] for the tunnel splitting for a symmetric double well problem analogous to that for the continuous case.[22,23] We apply the method to the Hamiltonian (1.1) in Sec. IV, and to (1.5) in Sec. V. We compare our answers for the splittings with numerical results, and in the case of Eq. (1.5) to answers given by van Hemmen and Sütő,[19] and Scharf, Wreszinski, and van Hemmen.[14]

II. THE DISCRETE WKB METHOD

To introduce the discrete WKB method, let us consider the Hamiltonian (1.1), and focus on an energy eigenfunction $|\psi\rangle$. Denoting S_z eigenfunctions $|m\rangle$ with $S_z|m\rangle = m|m\rangle$ as usual, and writing $C_m = \langle m|\psi\rangle$, it is easy to see that Schrödinger's equation takes the form of a three-term recursion relation connecting C_{m-2}, C_m, and C_{m+2}. The step $\Delta m = 2$ in the index is inconvenient, so it is preferable to change the notation somewhat, and relabel the states which are connected to each other by an index j, with $j = 1,2,\ldots,N$, where N equals S or $S+1$. The Schrödinger equation then takes the form

$$w_j C_j + t_{j,j+1} C_{j+1} + t_{j,j-1} C_{j-1} = E C_j, \tag{2.1}$$

where $w_j = \mathscr{H}_{jj}$ and $t_{j,j\pm1} = \mathscr{H}_{j,j\pm1}$, with $t_{N,N+1}$ and $t_{N+1,N}$ being understood as zero. We may also assume that $t_{j,j+1} \neq 0$ for any j, since otherwise the matrix for $\mathscr{H}$ is block diagonal, and Eq. (2.1) may be regarded as applying to one block. We can also choose the phases of the basis states so that all $t_{j,j\pm1}$ are real. Hermiticity then yields $t_{j+1,j} = t^*_{j,j+1} = t_{j,j+1}$.

The above problem is equivalent to a tight-binding model for an electron on a one-dimensional lattice with on-site energies w_j, and hopping matrix elements $t_{j,j\pm1}$, and indeed our notation is chosen to reflect this analogy. If w_j and $t_{j,j+1}$ vary slowly enough with j (how slowly will be seen below) we may use the approximation of semiclassical electron dynamics. To do this we first note that for the uniform case where $w_j = w$ and $t_{j,j+1} = t$ for all j, the eigenstates are Bloch states with wavefunction e^{iqj} and energy $E(q) = w + 2t\cos q$. For the nonuniform case, we first extend j to be a continuous variable, and define $w(j)$ and $t(j)$ to be extensions of w_j and $t_{j,j\pm1}$ via the relations

$$w(j) = w_j, \tag{2.2}$$

$$t(j) = (t_{j,j+1} + t_{j,j-1})/2, \tag{2.3}$$

which are required to hold whenever j is an integer. We further demand that $w(j)$ and $t(j)$ be smooth enough that if j/N is regarded as a quantity of order N^0, then

$$\frac{dw}{dj} = O\left(\frac{w(j)}{N}\right), \tag{2.4}$$

and likewise for $t(j)$. This equation is the formal statement of ''slowly varying'' coefficients. Whether or not it can be satisfied depends, of course, on the problem under consideration, and it is this condition which determines whether or not the problem is amenable to solution by the discrete WKB method. It is easy to see that it holds for Eq. (1.1) whenever $S \gg 1$.

We next define a local wave vector $q(j)$ by the relation

$$\cos q(j) = (E - w(j))/2t(j). \tag{2.5}$$

Further, since the electron velocity is given by $\partial E/\partial q$ in semiclassical dynamics, we define

$$v(j) = -2t(j)\sin q(j). \tag{2.6}$$

Then by a straightforward repeat of the arguments employed in the continuous WKB method (see e.g., Ref. 15), it is a simple matter to show that two linearly independent solutions to Eq. (1.1) are given by

$$C_j \sim \frac{1}{\sqrt{v(j)}} \exp\left(\pm i \int^j q(j')\,dj'\right). \tag{2.7}$$

A general solution is obtained by taking linear combinations of these two solutions.

A formal proof of Eq. (2.7) is given by Braun,[17] along with connection formulas, Bohr–Sommerfeld quantization rules, etc. We will therefore limit ourselves to a few key comments and comparisons with the continuous WKB method.

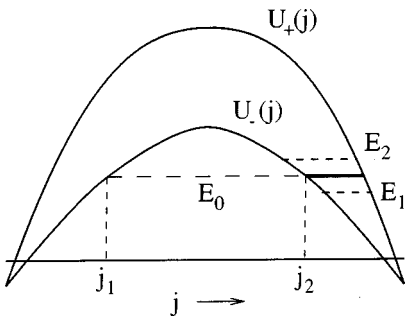

FIG. 1. Discrete WKB potential energy diagram for tunnel splitting problem. For a particle with energy E_0, the region between j_1 and j_2 is classically forbidden.

The first point is that Eq. (2.7) represents a development of $\log C_j$ in powers of N^{-1}, in which only the leading two terms are kept. To this accuracy, we need only keep the first two terms of $w(j)$ and $t(j)$ in an expansion in powers of $1/N$. Alternatively, we can use any expression for these functions which will reproduce the first two terms correctly. In particular, Eqs. (2.2) and (2.3) need only hold to this order. In considering this point, it should be remembered that j/N is regarded as being of $O(1)$.

The second point is that the factor $1/\sqrt{v}$ in Eq. (2.7) has the effect of normalizing each of the two solutions so that the probability flux is conserved and has a value of unity. This follows from noting that $|C_j|^2 \sim 1/v$, so that the particle spends a time inversely proportional to its velocity in any given coordinate segment dj. Since this is exactly what we expect for a classical particle, the quantum mechanical wave function is correctly adjusted to give a conserved probability flux.

The third point is that for a given energy E, the classically accessible region of motion is defined by the inequalities

$$U^-(j) \leqslant E \leqslant U^+(j), \tag{2.8}$$

where

$$U^{\pm} = w(j) \pm 2|t(j)|. \tag{2.9}$$

This follows from Eqs. (2.5) and (2.6) because whenever the condition (2.8) is violated, $q(j)$ and $v(j)$ acquire imaginary parts. The expression (2.7) (with $\sqrt{v(j)}$ replaced by $\sqrt{|v(j)|}$ as usual) then describes an exponentially decaying or growing wave function as opposed to an oscillatory one. It may also be noted that for uniform w_j and $t_{j,j\pm1}$, the quantities $U^{\pm}$ would be the limits of the allowable range of Bloch state energies. More generally they may be regarded as local, j-dependent band edges.

Finally, let us note the conditions for the validity of Eq. (2.7). The physical requirement is that the wavelength of the particle should not change by very much over dimensions of the wavelength itself, which is equivalent to demanding that

$$|dq(j)/dj| \ll 1. \tag{2.10}$$

In addition to failing if $w(j)$ and $t(j)$ do not vary slowly (as defined above), it is easy to see that Eq. (2.10) is violated whenever $q \approx 0$ or $\pm\pi$. In light of Eqs. (2.8) and (2.9), this is hardly surprising, since these are turning points of the classical motion. Solutions of the form (2.7) on opposite sides of these points must be related to each other by the above-mentioned connection formulas, as in the usual WKB method.

III. GENERAL FORMULA FOR TUNNEL SPLITTING

In this section we consider the problem where the functions w_j and $t_{j,j+1}$ possess reflection symmetry about some point, and the central region is classically forbidden for some energies. The band-edge functions $U^{\pm}(j)$ then have the form shown in Fig. 1. We will derive a formula [see Eq. (3.11)] for the tunnel splitting between states formed from symmetrical states localized in the two

potential wells. Our approach is completely analogous to that for a continuous coordinate.[22] A formula very close to Eq. (3.11) also appears in van Hemmen and Wreszinski,[23] but our method and notation are sufficiently different, and the argument sufficiently brief, that it is worth giving it in toto.

It is convenient to shift the index j so that it runs over the values

$$-\frac{N-1}{2},-\frac{N-3}{2},\ldots,\frac{N-1}{2}. \tag{3.1}$$

Note that j takes on integer values when N is odd, and half-integer values when N is even. With this labeling, we have

$$t_{j,k}=t_{-k,-j}=t_{k,j}. \tag{3.2}$$

It is also convenient, for future use, to define $j_0=1$ or 1/2, depending on whether N is odd or even, respectively. In either case, j_0 labels the first site to the right of the midpoint. We also define $J=(N-1)/2$, and $t_{\rm mid}=t_{10}$ or $t_{1/2,-1/2}$ for odd or even N.

Let C_j be the normalized wave function localized in the right well, and which satisfies the Schrödinger equation (2.1), neglecting the possibility of tunneling, with some energy E_0. Further, let a_j and s_j be the exact associated antisymmetric and symmetric wave functions, respectively, and let them have energies E_1 and E_2, where $(E_1+E_2)/2=E_0$. These functions are given to very good accuracy by

$$a_j=\frac{1}{\sqrt{2}}(C_j-C_{-j}),\qquad s_j=\frac{1}{\sqrt{2}}(C_j+C_{-j}). \tag{3.3}$$

Since the product C_jC_{-j} is everywhere exponentially small, these functions are properly normalized to unity.

The Schrödinger equations obeyed by C_j and a_j are

$$(w_j-E_0)C_j+t_{j,j-1}C_{j-1}+t_{j,j+1}C_{j+1}=0, \tag{3.4}$$

$$(w_j-E_1)a_j+t_{j,j-1}a_{j-1}+t_{j,j+1}a_{j+1}=0. \tag{3.5}$$

We multiply Eq. (3.4) by a_j, Eq. (3.5) by C_j, subtract the latter from the former, and sum the result from $j=j_0$ to $J\equiv(N-1)/2$. This yields

$$(E_1-E_0)\sum_{j\geqslant j_0}C_ja_j+\Sigma_1-\Sigma_2=0, \tag{3.6}$$

where

$$\Sigma_1=\sum_{j\geqslant j_0}a_j(t_{j,j-1}C_{j-1}+t_{j,j+1}C_{j+1}), \tag{3.7}$$

$$\Sigma_2=\sum_{j\geqslant j_0}C_j(t_{j,j-1}a_{j-1}+t_{j,j+1}a_{j+1}). \tag{3.8}$$

By shifting the index of summation in Eq. (3.8) we can see that most of the terms in the sums Σ_1 and Σ_2 are identical. Recalling that $t_{J,J+1}=0$, and that $a_0=0$ for odd N, we obtain

$$\Sigma_1-\Sigma_2=\begin{cases}a_1t_{\rm mid}C_0 & (\text{odd } N)\\ a_{1/2}t_{\rm mid}(C_{1/2}+C_{-1/2}) & (\text{even } N).\end{cases} \tag{3.9}$$

To evaluate the sum $\Sigma_j a_jC_j$ in Eq. (3.6), we use Eq. (3.3). Since C_{-j} is exponentially small everywhere in the right-hand well, and since C_j is normalized,

$$\sum_{j\geqslant j_0} C_j a_j \approx \frac{1}{\sqrt{2}} \sum_{j\geqslant j_0} C_j^2 = \frac{1}{\sqrt{2}}. \tag{3.10}$$

Substituting Eqs. (3.3), (3.9), and (3.10) in Eq. (3.6), and recalling that $E_1 - E_2 = 2(E_1 - E_0)$, we obtain[22]

$$E_1 - E_2 = \begin{cases} -2t_{\text{mid}} C_0 (C_1 - C_{-1}) & (\text{odd } N) \\ -2t_{\text{mid}} (C_{1/2}^2 - C_{-1/2}^2) & (\text{even } N). \end{cases} \tag{3.11}$$

Equation (3.11) is the sought after formula for the splitting. Note that unlike the case of a particle with a continuous coordinate degree of freedom, it is not now necessary for the antisymmetric state to have the higher energy. In fact, it is not hard to show that

$$\text{sgn}(E_1 - E_2) = \begin{cases} 1 & (\text{odd } N) \\ -\text{sgn } t & (\text{even } N). \end{cases} \tag{3.12}$$

IV. TUNNEL SPLITTINGS FOR MODEL HAMILTONIAN (1.1)

We now apply our general formalism to the Hamiltonian (1.1). Writing a general eigenstate of $\mathscr{H}$ as

$$|\psi\rangle = \sum_m C_m |m\rangle, \tag{4.1}$$

where $S_z|m\rangle = m|m\rangle$, Schrödinger's equation becomes

$$(w_m - E) C_m + t_{m,m+2} C_{m+2} + t_{m,m-2} C_{m-2} = 0, \tag{4.2}$$

with

$$w_m = (k_1 + \tfrac{1}{2} k_2)(S(S+1) - m^2), \tag{4.3}$$

$$t_{m,m+2} = \tfrac{1}{4} k_2 [(S(S+1) - m(m+1))(S(S+1) - (m+1)(m+2))]^{1/2}. \tag{4.4}$$

For convenience, we have added a constant $k_1 S(S+1)$ to the Hamiltonian, and thus to w_m.

As noted earlier, Eq. (4.2) is of the same form as Eq. (2.1) with the minor difference that the index m jumps in steps of 2 rather than 1. We can take care of this point by changing the integral in the exponential in Eq. (2.7) to $\int q(m')dm'/2$, leaving the other formulas unaffected.

One elementary point should be noted at this stage. For half-integer S, Kramers's theorem ensures an exact double degeneracy of all energy eigenstates of the Hamiltonian (1.1). This means that Δ_k vanishes. In the spin-coherent-state path integral approach, this result comes about because the kinetic term in the action is a Berry phase, and different symmetry related paths acquire phases which give rise to destructive interference for half-integer S.[24–26] In the discrete WKB method, this result comes about from an almost trivial separation of the eigenvalue problem into two isomorphic disjoint subspaces.

There are two steps in the calculation of Δ_k via Eq. (3.11). First, one must find the wave function C_m in the classically allowed and disallowed regions separately. In the present case, an expression for the former can be found exactly and in such a way that it actually holds in part of the disallowed region. This allows one to match on to the quasiclassical wave function given by Eq. (2.7) without using connection formulas. The second step consists of substituting the quasiclassical wave function into Eq. (3.11).

To execute the first step, let us consider a wave function $C_m^{(k)}$ with energy $E_0^{(k)}$ which corresponds to the kth state localized in the well near $m = -S$, and begin by finding it near this well. The coefficients w_m and $t_{m,m+2}$ are evaluated by writing $m = -S + n$, and expanding Eqs. (4.3) and (4.4) in powers of $1/S$, with n regarded as a quantity of order S^0. This yields

$$w_{-S+n}=(k_1+\tfrac{1}{2}k_2)(2n+1)S, \tag{4.5}$$

$$t_{-S+n,-S+n+2}=\tfrac{1}{2}k_2S((n+1)(n+2))^{1/2}. \tag{4.6}$$

Rewriting $C_{-S+n}=b_n$, the recursion relation (4.2) reads

$$(2k_1+k_2)S(n+\tfrac{1}{2})b_n+\tfrac{1}{2}k_2S\sqrt{n(n+1)}b_{n-2}+\tfrac{1}{2}k_2S\sqrt{(n+1)(n+2)}b_{n+2}=Eb_n, \tag{4.7}$$

which is immediately recognizable as arising from the harmonic oscillator Hamiltonian

$$\mathscr{H}_{\rm osc}=(2k_1+k_2)S(a^\dagger a+\tfrac{1}{2})+\tfrac{1}{2}k_2S(a^2+(a^\dagger)^2), \tag{4.8}$$

where a and $a^\dagger$ are the usual annihilation and creation operators obeying $[a,a^\dagger]=1$. To see this let us denote the kth energy state of $\mathscr{H}_{\rm osc}$ by $|\psi_k\rangle$, and the number eigenstates by $|n\rangle$ with $a^\dagger a|n\rangle=n|n\rangle$. It is then apparent that $b_n^{(k)}=\langle n|\psi_k\rangle$ obeys the eigenvalue equation (4.7).

To find $b_n^{(k)}$, we transform to position and momentum operators x and p obeying $[x,p]=i$ via $a=(x+ip)/\sqrt{2}$, $a^\dagger=(x-ip)/\sqrt{2}$. In terms of these,

$$\mathscr{H}_{\rm osc}=(k_1+\tfrac{1}{2}k_2)S(x^2+p^2)+\tfrac{1}{2}k_2S(x^2-p^2). \tag{4.9}$$

It follows that the kth state $|\psi_k\rangle$ has an energy

$$E_0^{(k)}=(k+\tfrac{1}{2})\omega_0, \tag{4.10}$$

$$\omega_0=2S(k_1(k_1+k_2))^{1/2}. \tag{4.11}$$

The wave function of this state is given by a textbook formula:

$$\langle x|\psi_k\rangle=(2^{2k}(k!)^2\pi\xi^2)^{-1/4}e^{-x^2/2\xi^2}H_k(x/\xi), \tag{4.12}$$

where H_k is the kth Hermite polynomial, and

$$\xi^2=\left(\frac{k_1}{k_1+k_2}\right)^{1/2}\equiv\tanh(\kappa/2). \tag{4.13}$$

It is also useful to write $\xi=e^{-\theta}$, in terms of which we have the symmetrical relations

$$e^{-2\theta}=\tanh(\kappa/2), \qquad e^{-\kappa}=\tanh\,\theta. \tag{4.14}$$

The number states $|n\rangle$ on the other hand, have wave functions

$$\langle x|n\rangle=(2^{2n}(n!)^2\pi)^{-1/4}e^{-x^2/2}H_n(x). \tag{4.15}$$

We can thus evaluate the overlap $\langle n|\psi_k\rangle=b_n^{(k)}$ in the x representation. The requisite integrals are most simply found by using the generating function for the Hermite polynomials,

$$e^{2xz-z^2}=\sum_{n=0}^{\infty}\frac{z^n}{n!}H_n(x). \tag{4.16}$$

The final result for $b_n^{(k)}$ is

$$b_n^{(k)}=\cos(\pi D)\left(\frac{n!k!}{2^{n+k}\cosh\,\theta}\right)^{1/2}\sum_{j=0}^{[k/2]}\frac{(-1)^j(\tanh\,\theta)^{D+2j}(2\,\mathrm{sech}\,\theta)^{k-2j}}{(D+j)!(k-2j)!j!}. \tag{4.17}$$

Here $D\equiv(n-k)/2$, and $[x]$ denotes the largest integer less than or equal to x. Note that the $\cos(\pi D)$ factor accounts for the vanishing of $b_n^{(k)}$ if $n-k$ is odd. The latter fact is a simple

consequence of the parity relation $H_n(-x)=(-1)^n H_n(x)$. It can be seen as a restatement of the separation of $\mathscr{H}$ into two subspaces, or of the $\Delta m=2$ step in Eq. (4.2), and rigorously implies $C_m^{(k)}=0$ for odd $m+S-k$ for the entire range of m.

Next, let us find $C_m^{(k)}$ in the central or classically forbidden region where $S-m$ and $S+m$ are both $O(S)$. In this region, we have to sufficient accuracy

$$w(m)=(k_1+\tfrac{1}{2}k_2)((S+\tfrac{1}{2})^2-m^2)+O(S^0), \tag{4.18}$$

$$t(m)=\tfrac{1}{4}k_2((S+\tfrac{1}{2})^2-m^2)+O(S^0). \tag{4.19}$$

It is easy to see that $d(w/t)/dm=O(S^{-1})$, so the quasiclassicality condition (2.10) holds if $S\gg 1$. Equation (2.5) then yields for the local wave vector

$$\begin{aligned}\cos\, q(m) &= \frac{2E_0^{(k)}-(2k_1+k_2)((S+\frac{1}{2})^2-m^2)}{k_2((S+\frac{1}{2})^2-m^2)} \\ &= -\left(\frac{2k_1+k_2}{k_2}\right)+\frac{2E_0^{(k)}}{k_2(2S+1)}\left(\frac{1}{S+\frac{1}{2}-m}+\frac{1}{S+\frac{1}{2}+m}\right).\end{aligned} \tag{4.20}$$

For the discrete WKB method to work, the energy levels which are split must lie well below the barrier, i.e., $E_0^{(k)}\ll(k_1+k_2/2)S^2$. The second term in Eq. (4.20) is then of order S^{-1} relative to the first, and we can solve for $q(m)$ as an expansion in powers of $1/S$. Equation (4.13) implies that

$$1+2k_1/k_2=\cosh\,\kappa, \tag{4.21}$$

so that

$$q(m)=\pi+i\left[\kappa-\frac{E_0^{(k)}}{\omega_0}\left(\frac{1}{S+\frac{1}{2}-m}+\frac{1}{S+\frac{1}{2}+m}\right)+O\left(\frac{E_0^{(k)}}{\omega_0 S^2}\right)\right]. \tag{4.22}$$

The exponent in Eq. (2.7) therefore equals (recalling the extra factor of 1/2 due to the step $\Delta m=2$)

$$\phi(m)=i\int^m q(m')\frac{dm'}{2}=\frac{i\pi m}{2}-\frac{\kappa m}{2}+\frac{E_0^{(k)}}{2\omega_0}\ln\left(\frac{S+\frac{1}{2}+m}{S+\frac{1}{2}-m}\right)+\cdots. \tag{4.23}$$

Similarly, correct to leading order in S, we have

$$|v(m)|=\frac{\omega_0}{2S}\left(\left(S+\frac{1}{2}\right)^2-m^2\right). \tag{4.24}$$

We now substitute Eqs. (4.23) and (4.24) into Eq. (2.7), changing the normalization from unit flux to total unit probability in accord with the discussion in Sec. III. Writing the result only for the nonzero $C_m^{(k)}$, i.e., for $m+S-k$ even, we have

$$C_m^{(k)}\approx A_k\frac{(S+\frac{1}{2}+m)^{(2k-1)/4}}{(S+\frac{1}{2}-m)^{(2k+3)/4}}e^{i\pi(m-k)/2}e^{-\kappa m/2}, \tag{4.25}$$

where A_k is the normalization constant, and we have used $E_0^{(k)}/\omega_0=k+\frac{1}{2}$.

It is now possible to match Eqs. (4.25) and (4.17) directly, since both results are valid in the region $S\gg n\gg 1$. We first find $b_n^{(k)}$ for $S\gg n\gg k$, with $k=O(1)$:

$$b_n^{(k)}=\frac{\cos(\pi(n-k)/2)}{\sqrt{k!}}\left(\frac{2}{\pi}\right)^{1/4}(\mathrm{sech}\,\theta)^{(2k+1)/2}(\tanh\,\theta)^{(n-k)/2}n^{(2k-1)/4}. \tag{4.26}$$

To find $C_m^{(k)}$ in the overlap region, we put $m=-S+n$ in Eq. (4.25), recall that $e^{-\kappa}=\tanh\theta$, and in this way obtain for $n\gg 1$,

$$C_{-S+n}^{(k)}=\frac{A_k}{(2S)^{(2k+3)/4}}\cos(\pi(n-k)/2)e^{-i\pi S/2}(\tanh\theta)^{(n-S)/2}n^{(2k-1)/4}. \tag{4.27}$$

Comparing with Eq. (4.26) we obtain

$$A_k=e^{i\pi S/2}\left(\frac{2}{\pi}\right)^{1/4}\frac{1}{\sqrt{k!}}(\operatorname{sech}\theta)^{(2k+1)/2}(\tanh\theta)^{(S-k)/2}(2S)^{(2k+3)/4}. \tag{4.28}$$

The last step is to apply Eq. (3.11) for the splitting. The only issue requiring caution is that of even versus odd N, the number of points on which the wave function $C_m^{(k)}$ is nonzero. We have $N=S+1$ for even k and $N=S$ for odd k. Thus the parity of N is that of $S+k-1$. Both cases can be easily handled at the same time, however. From Eq. (4.25), we see that for $|m|\ll S$, the nonvanishing $C_m^{(k)}$'s are given by $(A_k/S)e^{-\kappa m/2}$ up to an overall phase. Thus, the factor multiplying $2t_{\mathrm{mid}}$ in Eq. (3.11) equals $2\sinh\kappa|A_k/S|^2$ in magnitude in both cases. Similarly, we see from Eq. (4.4) that t_{mid} equals $k_2S^2/4$ in both cases. Thus,

$$\Delta_k=k_2\sinh\kappa|A_k|^2 \tag{4.29}$$

for all k and S.

It remains to substitute Eq. (4.28) into Eq. (4.29). We quote the answer in terms of the parameters ω_0 and θ:

$$\Delta_k=\frac{1}{k!}\left(\frac{4S}{\sinh 2\theta}\right)^k\Delta_0, \tag{4.30}$$

$$\Delta_0=4\omega_0\left(\frac{S}{\pi}\right)^{1/2}\operatorname{sech}\theta(\tanh\theta)^S. \tag{4.31}$$

For some purposes, it may be desirable to write these results in terms of the ratio ξ defined in Eq. (4.13). The conversion is easily done using the formulas

$$\tanh\theta=(1-\xi^2)/(1+\xi^2), \tag{4.32}$$

$$\operatorname{sech}\theta=2\xi/(1+\xi^2), \tag{4.33}$$

$$(1/\sinh 2\theta)=2\xi/(1-\xi^4). \tag{4.34}$$

As mentioned before, the result (4.31) for the ground state splitting agrees with that of Enz and Schilling,[7] and Belinicher, Providencia, and Providencia.[13] A comparison with numerically obtained exact answers is given in Table I. As is evident, the agreement is quite good, and improves with increasing S.

The formula (4.30) for the higher splittings is, to our knowledge, new.

V. TUNNEL SPLITTING FOR MODEL HAMILTONIAN (1.5)

In this section we apply our method to the Hamiltonian (1.5), limiting ourselves to finding the ground state tunnel splitting Δ_0. This time the recursion relation connects contiguous coefficients C_{m-1}, C_m, and C_{m+1}, and

$$w(m)=-\gamma m^2, \tag{5.1}$$

$$t(m)=(-\alpha/2)(S'^2-m^2)^{1/2}, \tag{5.2}$$

where we have introduced the convenient abbreviation

TABLE I. Comparison between numerical and analytical [Eq. (4.31)] results for the ground state tunnel splitting for the Hamiltonian (1.1). Numbers in parentheses give the power of 10 multiplying the answer. We have chosen $k_1=5.0/S$, $k_2=20.0/S$. This scaling keeps ω_0 fixed as $S\to\infty$.

S	Δ_0 (numerical)	Δ_0 (analytic)	Error(%)
10	9.282(−3)	9.749(−3)	5.0
11	3.738(−3)	3.906(−3)	4.5
12	1.497(−3)	1.558(−3)	4.1
13	5.974(−4)	6.195(−4)	3.7
14	2.375(−4)	2.455(−4)	3.4
15	9.412(−5)	9.708(−5)	3.1
16	3.721(−5)	3.830(−5)	2.9
17	1.468(−5)	1.508(−5)	2.7
18	5.778(−6)	5.927(−6)	2.6
19	2.271(−6)	2.326(−6)	2.4
20	8.910(−7)	9.115(−7)	2.3

$$S'=S+1/2. \tag{5.3}$$

Let us begin, as in Sec. IV, by finding the wave function for the state localized in the right-hand well, C_m, in the classically accessible region. This time the answer depends on how the parameters α and γ are taken to vary as $S\to\infty$. To see this note that the lower band-edge function $U^-(m)$ now has a quadratic as opposed to linear form near its minima (see Fig. 2). If we define

$$m=S'\cos\theta, \tag{5.4}$$

then

$$U^-(m)=-\gamma S'^2\cos^2\theta-\alpha S'\sin\theta. \tag{5.5}$$

The mimima are located at $\theta=\theta_1$, $\pi-\theta_1$, where

$$\sin\theta_1=\alpha/2\gamma S'. \tag{5.6}$$

It follows that if α and γ remain fixed as $S\to\infty$, the minima approach the points $m=\pm S$, which gives us a problem similar to that of Sec. IV. For novelty's sake, let us therefore take α and γ to scale so that θ_1 remains fixed and not close to 0 as $S\to\infty$. We shall see that our final answer for Δ_0 will hold even when this condition is not satisfied.

For future use, let us define θ_0 as the angle at which the classical expression for the energy (not $U^-(m)$) has a minimum. We have

$$\sin\theta_0=\alpha/2\gamma S. \tag{5.7}$$

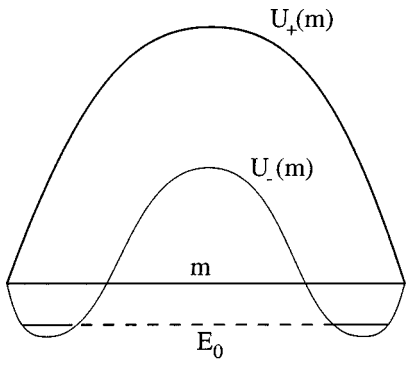

FIG. 2. Sketch of the band-edge functions for the Hamiltonian (1.5). The key point is that $U^-(m)$ now has a quadratic minimum.

Further,

$$\sin\ \theta_1=(1-(1/2S)+\cdots)\sin\ \theta_0,\quad \cos\ \theta_1=\cos\ \theta_0-\sin^2\ \theta_0/2S\ \cos\ \theta_0+\cdots. \tag{5.8}$$

With the above conditions on θ_1, it is easy to find C_m near the potential minima. Let us define

$$m_1=S'\ \cos\ \theta_1. \tag{5.9}$$

The conditions on α and γ imply that $S-m_1=O(S)$. Since the wave function we seek is essentially that for the ground state of a harmonic oscillator, its spatial extent, Δm, on the other hand, is of order $S^{1/2}$. This can be seen by expanding the semiclassical Hamiltonian in powers of q and $m-m_1$. We have

$$\begin{aligned}\mathscr{H}_{\rm sc}(q,m)&=w(m)+2t(m)\cos\ q\\ &\approx-\gamma S'^2(1+\sin^2\ \theta_1)+\gamma S'^2\ \sin^2\ \theta_1 q^2+\gamma\ \cot^2\ \theta_1(m-m_1)^2+\cdots.\end{aligned} \tag{5.10}$$

The leading neglected terms are, ignoring S' independent constants, of order $S'(m-m_1)q^2 \equiv h_1$, $S'^2q^4\equiv h_2$, and $(m-m_1)^3/S'\equiv h_3$. It follows that the small oscillation frequency ω_0 and the ground state energy E_0 (in the absence of tunneling, needless to say) are given by

$$\omega_0=2\gamma S'\ \cos\ \theta_1, \tag{5.11}$$

$$E_0=-\gamma S'^2(1+\sin^2\ \theta_1)+\gamma S'\ \cos\ \theta_1. \tag{5.12}$$

The wave function can be written down directly as for a harmonic oscillator with a continuous spatial coordinate,

$$C_m=\left(\frac{\cos\ \theta_1}{\pi S'\ \sin^2\ \theta_1}\right)^{1/4}\exp\left(-\frac{\cos\ \theta_1}{2S'\ \sin^2\ \theta_1}(m-m_1)^2\right), \tag{5.13}$$

which is already normalized to unity inside the well.

What is the range of validity of Eq. (5.13)? This can be found by demanding that the terms ignored, h_1, h_2, and h_3, be much smaller than $\omega_0/2$, the value of the q^2 and $(m-m_1)^2$ terms in Eq. (5.10). Since the kinetic and potential energies are equal for a harmonic oscillator, we have $q^2\propto(m-m_1)^2/S'^2$ in the region of validity. Thus, $h_1=O(m-m_1)^3/S'=O(h_3)$, and $h_2=O(m-m_1)^4/S'^2\ll h_1,h_3$. Demanding that $\omega_0/2\gg h_1$, we find that Eq. (5.13) is valid provided

$$|m-m_1|\ll(S')^{2/3}. \tag{5.14}$$

Note that this range keeps us clear of the end point $m=S$.

The second step is to find C_m in the classically forbidden region, especially near the point $m=0$ (if S is integral) or $m=1/2$ (if S is half-integral). Since $q(m)$ is imaginary, we define

$$q(m)=i\kappa(m), \tag{5.15}$$

with $\kappa(m)>0$. Using the basic definition (2.5) with the energy (5.12), we obtain

$$\cosh\ \kappa(m)=\frac{\sin^2\ \theta+\sin^2\ \theta_1}{2\sin\ \theta\ \sin\ \theta_1}-\frac{\cos\ \theta_1}{2S'\ \sin\ \theta\ \sin\ \theta_1}+O(S')^{-2}. \tag{5.16}$$

Writing

$$\kappa(m)=\kappa_0(m)+\kappa_1(m)+\cdots, \tag{5.17}$$

where κ_j is of order $1/S'^j$, we obtain

$$\kappa_0(m)=\frac{1}{2}\log\left(\frac{S'^2-m^2}{S'^2\ \sin^2\ \theta_1}\right), \tag{5.18}$$

$$\kappa_1(m)=\frac{1}{2}\left[\frac{1}{m-m_1}-\frac{1}{m+m_1}\right]. \tag{5.19}$$

The exponent in Eq. (2.7) therefore equals

$$\begin{aligned}\phi(m)&=\phi_0(m)+\phi_1(m)\\ &=\tfrac{1}{2}[(S'+m)\log(S'+m)-(S'-m)\log(S'-m)\\ &\quad -2m\ \log(eS'\ \sin\ \theta_1)]\\ &\quad +\tfrac{1}{2}[\log(m_1-m)-\log(m_1+m)].\end{aligned} \tag{5.20}$$

Here, ϕ_0 and ϕ_1 are the quantities in the first and second square brackets above, and correspond, respectively, to κ_0 and κ_1.

The velocity is now easy to find. To leading order in S, we get

$$|v(m)|=2\gamma S'^2\ \sin\ \theta_1\ \sin\ \theta\times\sinh\ \kappa_0=\gamma(m_1^2-m^2). \tag{5.21}$$

The wave function in the central region therefore equals

$$C_m=\frac{A}{m+m_1}\exp\ \phi_0(m), \tag{5.22}$$

where A is a constant that must be found by matching the solution with Eq. (5.13). This in turn is done by expanding $\phi_0(m)$ in powers of $m-m_1$, which is permissible as $|m-m_1|\ll S^{2/3}$ in the matching region. We find

$$\phi_0(m)=\phi_{00}-\frac{\cos\ \theta_1}{2S'\ \sin^2\ \theta_1}(m-m_1)^2; \tag{5.23}$$

$$\phi_{00}=S'[\log\ \cot(\theta_1/2)-\cos\ \theta_1]. \tag{5.24}$$

Putting $m=m_1$ in the pre-exponential factor in Eq. (5.22), we see that the forbidden region C_m has exactly the structure required to match on to Eq. (5.13). Comparing the two expressions, we obtain

$$A=2m_1\left(\frac{\cos\ \theta_1}{\pi S'\ \sin^2\ \theta_1}\right)^{1/4}e^{-\phi_{00}}. \tag{5.25}$$

The final step is to apply the formula (3.11) for the splitting. In doing this, the following simplifications are useful. First,

$$\phi_0(m)=-m\ \log\ \sin\ \theta_1\quad \text{for}\ \ m=O(1). \tag{5.26}$$

Second, to leading order in S,

$$\omega_0=2\gamma S\ \cos\ \theta_0. \tag{5.27}$$

Collecting together all these facts, we obtain, for both integral and half-integral S,

$$\Delta_0=4\omega_0\left(\frac{S}{\pi}\right)^{1/2}\frac{\cos^{3/2}\ \theta_0}{\sin\ \theta_0}\left(\frac{1-\cos\ \theta_0}{1+\cos\ \theta_0}\right)^{S+1/2}e^{2S\ \cos\ \theta_0}. \tag{5.28}$$

This is the main result of this section. It can be seen to be identical to Eq. (16) of Enz and Schilling,[7] via the substitutions $A=\gamma$, $a=\sin\ \theta_0$ and $b=0+$. It is almost, but not quite identical to Eq. (4.11) of Garg and Kim,[12] which errs by a factor of $O(S^0)$.

Several comments about Eq. (5.28) are in order. First, although we have only derived it assuming that $\sin\ \theta_0\gg 1/S$, it holds in fact for all $\sin\ \theta_0$, provided of course that the WKB approximation itself is valid. This can be seen from the numerical comparison we give below, and also analytically, by assuming that $\sin\ \theta_0=O(1/S)$ from the very start. This suggests that Eq. (5.28) holds uniformly for all $\sin\ \theta_0$ not too close to 1, but we have not investigated this point carefully. As a subcomment, note that in the limit $\theta_0\to 0$, i.e., vanishing α, $\Delta_0\sim\alpha^{2S}$. This is of course,

TABLE II. Comparison between numerical and various analytical and semianalytical results for the ground state tunnel splitting of the Hamiltonian (1.5). The parameters are $\alpha=\gamma=1$. As in Table I, $x(n)$ denotes the number $x\times 10^n$. The quantities Δ_n, Δ_{asymp}, and Δ_{int} are taken from Table I of van Hemmen and Sütő (Ref. 19), and are, respectively, the numerically computed splitting, the result (5.31), and the numerically evaluated integral expression (5.32). The quantity Δ_0 is given by Eq. (5.28) of this work.

	Numerics	Analytic/semianalytic results from Ref. 19				This work	
S	Δ_n	Δ_{asymp}[a]	$\Delta_{\text{asymp}}/\Delta_n$	Δ_{int}	$\Delta_{\text{int}}/\Delta_n$	Δ_0	Δ_0/Δ_n
3	1.44(−3)	4.12(−3)	2.86	2.01(−3)	1.40	1.48(−3)	1.03
4	1.18(−5)	6.10(−5)	5.17	1.64(−5)	1.39	1.20(−5)	1.02
5	5.18(−8)	5.12(−7)	9.88	7.25(−8)	1.40	5.24(−8)	1.01
6	1.43(−10)	2.76(−9)	19.3	2.00(−10)	1.40	1.44(−10)	1.01
7	2.68(−13)	1.03(−11)	38.5	3.77(−13)	1.41	2.70(−13)	1.01
8	3.66(−16)	2.84(−14)	77.7	5.15(−16)	1.41	3.68(−16)	1.01
9	3.79(−19)	6.00(−17)	158	5.36(−19)	1.41	3.81(−19)	1.01
10	3.09(−22)	1.00(−19)	324	4.37(−22)	1.41	3.10(−22)	1.00
11	2.03(−25)	1.35(−22)	666	2.87(−25)	1.41	2.03(−25)	1.00

[a]The numbers have been given to greater accuracy and corrected in a few places.

exactly what is expected from perturbation theory—it takes $2S$ applications of the perturbation $-\alpha S_x$ to connect the states with $S_z=\pm S$.[27] Second, the WKB validity restriction means that the field α in Eq. (1.5) cannot be too large. Quantitatively, we require that the energy barrier be much greater than ω_0, the level spacing in the well. This condition can be shown to be equivalent to

$$1-\sin\,\theta_0 \gg S^{-2/3}. \tag{5.29}$$

Third, as long as Eq. (5.29) holds, in the limit where the field is large and the barrier is small, Eq. (5.28) reduces to

$$\Delta_0=4\sqrt{3}\,\omega_0(B_0/2\pi)^{1/2}e^{-B_0}, \tag{5.30}$$

where $B_0=(2/3)S\,\cos^3\theta_0=(16/3)\Delta U/\omega_0$, ΔU being the energy barrier. This form is the well-known answer for the tunnel splitting of a massive particle moving in a quartic potential double well. The spin problem can be transformed into such a problem by noting that (i) in this limit $S_x\approx S$, so that the commutator $[S_y,S_z]=iS_x$ can be approximately reproduced by writing $S_z=x$, $S_y=-Sp$, where p and x are momentum and position operators with the commutator $[p,x]=-i$, and (ii) expanding the square root in the relation $S_x=[S^2-(S_y^2+S_z^2)]^{1/2}$, and rewriting S_y and S_z in terms of p and x as indicated above. We leave it to the reader to check that the frequency ω_0 and energy barrier ΔU of this particle problem are given by $2\gamma S\,\cos\theta_0$ and $\gamma S^2\,\cos^3\theta_0/4$, so that the answer for Δ_0 is indeed given by Eq. (5.30).

Fourth, let us compare our result for Δ_0 with the numerical and analytic answers obtained by van Hemmen and Sütő,[19] which apply when $\sin\theta_0=O(1/S)$. This comparison is shown in Table II. The parameters are $\gamma=\alpha=1$, and the two approximate forms for the splitting are given by

$$\Delta_{\text{asymp}}=2\gamma S\,\exp[-2S\,\log(\gamma S)], \tag{5.31}$$

$$\Delta_{\text{int}}=\frac{\omega_0}{2}\,\exp\left[-\int_{-m_t}^{m_t}\cosh^{-1}\left(\frac{-E_0-\gamma m^2}{\alpha[S(S+1)-m^2]^{1/2}}\right)\right]. \tag{5.32}$$

In Eq. (5.32), $\pm m_t$ are the classical turning points, where $E_0=w(m)$, and E_0 is the numerically obtained ground state energy. van Hemmen and Sütő describe Δ_{asymp} as an ''asymptotic'' result, and Δ_{int} as the ''exact WKB'' answer. As can be seen from the columns containing the ratios of the splittings given by these formulas to the numerically found splittings, neither result asymptotes to the correct answer. [Recall that two functions $f(x)$ and $g(x)$ are said to be asymptotically equal as $x\to x_0$, i.e., $f(x)\sim g(x)$, if $f(x)/g(x)\to 1$ as $x\to x_0$.] Δ_{asymp} actually gets worse with increasing S, which indicates that a multiplicative term such as e^{-cS} has been omitted (c is some constant), while Δ_{int} is off by a term of S^0.[28] Neither result therefore meets the goal set in Sec. I

of our paper: to find all terms B_n in Eq. (1.3) that are of order S^0 or greater. Equation (5.28), on the other hand, does just that. It is also obvious that the recipe given by van Hemmen and Sütő [see Eq. (5.19) of Ref. 19] cannot really be exact in general as regards the prefactor.

Last, Eq. (5.28) can be compared with the answer obtained by Scharf, Wreszinski, and van Hemmen,[14] which is valid when $\sin\theta_0 = O(1/S)$. Their result—see their Eqs. (1.7) and (1.8), and Table I—is also off by $O(S^0)$. In fact the discrepancy is again $(e/\pi)^{1/2}$, which is to be expected in light of their method.[28]

ACKNOWLEDGMENT

This work is supported by the NSF via Grant No. DMR-9616749.

[1] H. J. Lipkin, N. Meshov, and A. J. Glick, Nucl. Phys. **62**, 188 (1965).
[2] R. Shankar, Phys. Rev. Lett. **45**, 1088 (1980).
[3] A. Klein and C. T. Li, Phys. Rev. Lett. **46**, 895 (1981).
[4] See, e.g., W. G. Harter and C. W. Patterson, J. Chem. Phys. **80**, 4241 (1984).
[5] I. Hamamoto and B. Mottelson, Phys. Lett. B **333**, 294 (1994); Phys. Scr. **T56**, 27 (1995).
[6] M. Enz and R. Schilling, J. Phys. C **19**, L711 (1986).
[7] M. Enz and R. Schilling, J. Phys. C **19**, 1765 (1986).
[8] S. Coleman, Phys. Rev. D **15**, 2929 (1977); *Aspects of Symmetry* (Cambridge University Press, Cambridge, England), Chap. 7.
[9] J. R. Klauder, Phys. Rev. D **19**, 2349 (1979).
[10] E. M. Chudnovsky and L. Gunther, Phys. Rev. Lett. **60**, 661 (1988).
[11] H. Kuratsuji and Y. Mizobuchi, J. Math. Phys. **22**, 757 (1981).
[12] A. Garg and G.-H. Kim, Phys. Rev. B **45**, 12 921 (1992).
[13] V. I. Belinicher, C. Providencia, and J. da Providencia, J. Phys. A **30**, 5633 (1997).
[14] G. Scharf, W. F. Wreszinski, and J. L. van Hemmen, J. Phys. A **20**, 4309 (1987). These authors obtain tunnel splittings for the Hamiltonian (1.5) by constructing a generating function for the expansion coefficients C_m. They then apply the usual continuous WKB method to this generating function. The splittings found by them, however, are again in error by terms of order $O(S^0)$. See the end of Sec. V for more discussion of this point.
[15] See, e.g., L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, New York, 1977), 3rd ed., Chap. VII.
[16] K. Schulten and R. G. Gordon, J. Math. Phys. **16**, 1971 (1975).
[17] P. A. Braun, Rev. Mod. Phys. **65**, 115 (1993).
[18] J. L. van Hemmen and A. Sütő, Europhys. Lett. **1**, 481 (1986).
[19] J. L. van Hemmen and A. Sütő, Physica B & C **141**, 37 (1986).
[20] A. Garg, Phys. Rev. B **51**, 15161 (1995).
[21] A. Garg, ''Coherent magnetic moment reversal in small particles with nuclear spins,'' in *Tunneling in Complex Systems*, edited by S. Tomsovic (World Scientific, Singapore, 1998).
[22] Reference 15, Sec. 50, Problem 3. Landau and Lifshitz proceed to rewrite the analog of Eq. (3.11) in terms of the integral of $q(j)$ over the entire classically forbidden region. We find it preferable to stop at the third-last displayed formula in their solution, since applying it is no less tractable. This formula is sometimes named after C. Herring, who found a similar result for the energy difference of the g and u states of the H_2^+ ion, for which, see the solved problem in Sec. 81 in Ref. 15.
[23] J. L. van Hemmen and W. F. Wreszinski, Commun. Math. Phys. **119**, 213 (1988).
[24] D. Loss, D. P. DiVincenzo, and G. Grinstein, Phys. Rev. Lett. **69**, 3232 (1992).
[25] J. von Delft and C. L. Henley, Phys. Rev. Lett. **69**, 3236 (1992).
[26] A. Garg, Europhys. Lett. **22**, 205 (1993).
[27] F. Hartmann-Boutron, J. Phys. I **5**, 1281 (1995).
[28] Readers may wonder if all that is wrong with van Hemmen and Sütő's formula is the typographical omission of a factor of $2^{-1/2}$. This is not so. An inspection of the arguments accompanying their Eqs. (5.19) and (C.10-12) reveals that there is no scope for such a factor and it is purely accidental that the error seems to be $2^{1/2}$ in this case. Indeed, use of a similar formula for potential problems with massive particles in one dimension also yields incorrect prefactors; for the quartic double well, e.g., the error is known to be $(e/\pi)^{1/2}$. The same error is present in Scharf, Wreszinski, and van Hemmen's result (Ref. 14).

Quantum coordinates of an event in local quantum physics

R. Giannitrapani
Dipartimento di Fisica, Università di Trento, Povo (TN) I-38050, Italy

(Received 8 April 1998; accepted for publication 19 May 1998)

Recently Toller [quant-ph/9702060 (1997)] has proposed, using the formalism of positive-operator-valued measures, a possible definition of quantum coordinates for events in the context of quantum mechanics. In this short note we analyze this definition from the point of view of local algebras in the framework of local quantum theories. © *1998 American Institute of Physics.* [S0022-2488(98)00710-5]

I. INTRODUCTION

Local quantum theory is now a well established framework for physical concepts and theories.[1,2] An interesting problem is to see whether a given local quantum theory admits a sharp localization of a space–time event; this problem is intrinsically related to the definition of space-time at small scale and so to the quantum gravity problem.

In a recent work[3] Toller has proposed defining the localizability of an event using quantum coordinates via positive operator valued (POV) measures;[4] such a useful formalism has been used[5–8] to investigate the long-standing problem[9–14] of time in quantum mechanics and the quantum coordinates of Toller are a generalization of these works. In this short note we want to stress the fact that such quantum observable for the coordinates of an event cannot be built with quasilocal operators.

II. POV MEASURES IN THE LOCAL FRAMEWORK

Let $\mathcal{M}$ be the Minkowsky space–time and $\mathcal{F}(\mathcal{M})$ be the σ-algebra of borel subsets of $\mathcal{M}$; to the physical quantum object that defines the event is associated an Hilbert space $\mathcal{H}$ and we call $\mathcal{B}(\mathcal{H})$ the algebra of bounded linear operators on $\mathcal{H}$ and with $\mathcal{B}(\mathcal{H})^+$ the positive ones. A POV measure on $\mathcal{M}$ with value in $\mathcal{B}(\mathcal{H})^+$ is a map

$$\tau:\mathcal{F}(\mathcal{M})\rightarrow\mathcal{B}(\mathcal{H})^+$$

such that τ is σ-additive in the weak topology, $\tau(\emptyset)=0$ and $\tau(\mathcal{M})\leqslant 1$. If τ has to fix a quantum event we require that it is covariant with respect to a unitary representation of the Poincaré group. We remark on the fact that we cannot, in general, require the normalization $\tau(\mathcal{M})=1$ for the POV measure, as discussed in Ref. 1.

Let us now consider the formulation of a local quantum theory[3] in terms of local operators algebras; it is defined by a net of C^*-algebras, i.e., by an inclusion preserving assignment to every open, bounded region $\mathcal{O}$ of $\mathcal{M}$ of a unital C^*-algebra $\mathcal{A}(\mathcal{O})$; the closure in the norm topology of their union is the algebra of quasilocal observables

$$\mathcal{A}=\overline{\bigcup_{\mathcal{O}\subset\mathcal{M}}\mathcal{A}(\mathcal{O})}.$$

It is apparent that $\tau(\mathcal{T})$, with $\mathcal{T}\in\mathcal{F}(\mathcal{M})$ and $\mathcal{T}\neq\emptyset$, is not a local operator, i.e., it does not belong to one of the local algebras $\mathcal{A}(\mathcal{O})$. To see this, note that the vacuum state Ω is invariant under the translations on $\mathcal{M}$ and so

$$(\Omega,\tau(\mathcal{T})\Omega)=(\Omega,U(x)\tau(\mathcal{T})U^{-1}(x)\Omega)=(\Omega,\tau(\mathcal{T}_x)\Omega),$$

where $U(x)$ is an unitary representation on $\mathcal{H}$ of the group of spacetime translations and $\mathcal{T}_x$ is the region $\mathcal{T}$ transformed with the translation by x. If for $\mathcal{T}$ we have

 © 1998 American Institute of Physics

$$(\Omega, \tau(\mathscr{T})\Omega) = \epsilon$$

with $\epsilon > 0$, there does exist an integer n such that $n\epsilon > 1$. Now we can choose $n-1$ vectors $\{x_1, \ldots, x_{n-1}\}$ such that $\mathscr{T} \cap \mathscr{T}_{x_i} = \varnothing$ and $\mathscr{T}_{x_i} \cap \mathscr{T}_{x_j} = \varnothing \ \forall i, j;\ i \neq j$; for the additivity of τ and the invariance of the vacuum we have

$$(\Psi, \tau(\mathscr{T} \cup \mathscr{T}_{x1} \cup \ldots \cup \mathscr{T}_{x_{n-1}})\Omega) = n\epsilon > 1$$

that is impossible. So $(\Omega, \tau(\mathscr{T})\Omega) = 0$ and, since $\tau(\mathscr{T})$ is bounded and positive, for the Reeh–Schlieder[15] theorem it cannot be a local operator.

With a more complicated demonstration we can see that it is not a quasilocal operator. To see this we need the following trivial lemma.

Lemma 1: Given a POV measure τ on $\mathscr{M}$ covariant with respect to a unitary representation of the translations group, if $\tau(\mathscr{T})$ is a quasilocal operator for some region $\mathscr{T}$ then it is a quasilocal operator for every translated regions $\mathscr{T}_x$.

A second necessary lemma is a standard result of local quantum physics (see Refs. 3, 16, and 17); let us consider n quasilocal operators

$$A_1 \ldots A_n \in \mathscr{A}$$

and the automorphism of $\mathscr{A}$ associated to a translation of x given by $\alpha_x A \equiv U(x) A U^{-1}(x)$. Let us divide the set of the indices $\{1 \ldots n\}$ in two disjoint subsets I and J and define n new quasi-local operator $A'_1 \ldots A'_n$ by

$$A'_i = A_i, \quad i \in I,$$

$$A'_i = A_i(x), \quad i \in J,$$

where $A_i(x) = \alpha_x A_i$. If we indicate $x = (t, \mathbf{x})$ then we have

Lemma 2 (Cluster property): For $|\mathbf{x}| \to \infty$

$$(\Omega, A'_1 \cdots A'_n \Omega) \to \left(\Omega, \prod_{i \in I} A'_i \Omega \right) \cdot \left(\Omega, \prod_{j \in J} A'_j \Omega \right).$$

With these lemmas we can now demonstrate the following

Proposition 1: If τ is a POV measure on $\mathscr{M}$ covariant with respect to the space–time translations group, then $\tau(\mathscr{T}) \notin \mathscr{A}$ for all $\mathscr{T} \in \mathscr{F}(\mathscr{M})$ with $\tau(\mathscr{T}) \neq 0$.

Proof:

Let us suppose that $\tau(\mathscr{T})$ is a quasilocal operator and let ψ be a normalized vector such that

$$(\psi, \tau(\mathscr{T})\psi) = p$$

with $0 < p \leqslant 1$; since the subset of vectors obtained by applying quasilocal operators to Ω is dense in $\mathscr{H}$ one can assume that

$$\psi = A\Omega,$$

where A is a quasilocal operator. Defining the vector $\psi_x = A(x)\Omega$, where $x = (t, \mathbf{x})$, for the covariance property of τ we have

$$(\psi_x, \tau(\mathscr{T}_x)\psi_x) = p.$$

We can now show that a vector sequence $\{\tilde{\psi}_x\}$ exists such that, for $|\mathbf{x}| \to \infty$,

$$(\tilde{\psi}_x, \tau(\mathscr{T} \cup \mathscr{T}_x)\tilde{\psi}_x) \to 2p$$

and this is obviously impossible since $\tau(\mathscr{M}) \leqslant 1$.

In fact, let us define the normalized vector

$$\tilde{\psi}_x = \frac{A(x)A\Omega}{\|A(x)A\Omega\|};$$

using the clustering lemma one can see that

$$\|A(x)A\Omega\|^2 \to \|A(x)\Omega\|^2 \cdot \|A\Omega\|^2 = 1$$

and that

$$(\tilde{\psi}_x, \tau(\mathscr{T})\tilde{\psi}_x) \to (\Omega, A^*(x)A(x)\Omega) \cdot (\Omega, A^*\tau(\mathscr{T})A\Omega) = (\psi, \tau(\mathscr{T})\psi) = p$$

for $|\mathbf{x}| \to \infty$. Similarly, in the same limit,

$$(\tilde{\psi}_x, \tau(\mathscr{T}_x)\tilde{\psi}_x) \to (\Omega, A^*A\Omega) \cdot (\Omega, A^*(x)\tau(\mathscr{T}_x)A(x)\Omega) = (\psi_x, \tau(\mathscr{T}_x)\psi_x) = p.$$

So

$$(\tilde{\psi}_x, \tau(\mathscr{T} \cup \mathscr{T}_x)\tilde{\psi}_x) = (\tilde{\psi}_x, \tau(\mathscr{T})\tilde{\psi}_x) + (\tilde{\psi}_x, \tau(\mathscr{T}_x)\tilde{\psi}_x) \to 2p.$$

If $2p > 1$ the demonstration is over; otherwise we can repeat the same argument finding normalized states such that the preceding probability tends toward $4p$ and so on. □

So we see that the existence of a normalized POV measure covariant with respect to a unitary representation of the group of translations on $\mathscr{M}$ is incompatible with the local principle if we require that $\rho(\mathscr{T})$ is quasilocal.

III. CONCLUSIONS

The use of POV measures as observables is motivated by some deep considerations for the foundations of quantum mechanics.[18,19] In particular, the use of positive bounded operators is very interesting, not necessarily projectors, as generalized propositions; it is an open problem to see what kind of observables it is possible to build with them. In this note we have shown that a localizability for events in space–time cannot be described by one of these observables if the generalized propositions have to belong to the algebra of quasilocal operators; further investigations are necessary in this direction.

ACKNOWLEDGMENTS

I wish to thank Marco Toller whose ideas inspired this work and Valter Moretti for some useful discussions on the subject.

[1] R. Haag and D. Kastler, J. Math. Phys. **5**, 848 (1964).
[2] R. Haag, *Local Quantum Physics*, 2nd ed. (Springer, Berlin, 1996).
[3] M. Toller, quant-ph/9702060 (1997).
[4] P. Busch, P. J. Lathi, and P. Mittelstaedt, *The Quantum Theory of Measurement* (Springer, Berlin, 1991).
[5] A. S. Holevo, *Probabilistic and Statistical Aspect of Quantum Theory* (North-Holland, Amsterdam, 1982).
[6] P. Busch, M. Grabowski, and P. J. Lathi, Phys. Lett. A **191**, 357 (1994).
[7] R. Giannitrapani, Int. J. Theor. Phys. **36**, 1575 (1997).
[8] H. Atmanspacher and A. Amann (unpublished).
[9] N. Grot, C. Rovelli, and R. S. Tate, Phys. Rev. A **54**, 4676 (1996).
[10] V. Delgado and J. G. Muga, Phys. Rev. A **56**, 3425 (1997).
[11] V. Delgado, quant-ph/9709037 (1997).
[12] J. Leon, J. Phys. A **30**, 4791 (1997).
[13] J. G. Muga, R. Sala, and J. P. Palao, Superlattices Microstruct. **24**, 833 (1998).
[14] Y. Aharonov, J. Oppenheim, S. Popescu, B. Reznik, and W. G. Unruh, quant-ph/9709031 (1997).
[15] H. Reeh and S. Schlieder, Nuovo Cimento **22**, 1051 (1961).
[16] R. Jost, *The General Theory of Quantized Fields*, Lectures in Applied Mathematics (Benjamin-Cummings, New York, 1964).
[17] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That*, Mathematical Physics Monograph Series (American Mathematical Society, Providence, RI, 1977).
[18] R. Giles, J. Math. Phys. **11**, 2139 (1970).
[19] G. Ludwig, Commun. Math. Phys. **9**, 1 (1968).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 10 OCTOBER 1998

Kontsevich integral for Vassiliev invariants from Chern–Simons perturbation theory in the light-cone gauge

J. M. F. Labastida[a)]
Theory Division, CERN, CH-1211 Geneva 23, Switzerland and Departamento de Física de Partículas, Universidade de Santiago de Compostela, E-15706 Santiago de Compostela, Spain

Esther Pérez
Departamento de Física de Partículas, Universidade de Santiago de Compostela, E-15706 Santiago de Compostela, Spain

(Received 21 January 1998; accepted for publication 27 February 1998)

We analyze the structure of the perturbative series expansion of Chern–Simons gauge theory in the light-cone gauge. After introducing a regularization prescription that entails the consideration of framed knots, we present the general form of the vacuum expectation value of a Wilson loop. The resulting expression turns out to give the same framing dependence as the one obtained using nonperturbative methods and perturbative methods in covariant gauges. It also contains the Kontsevich integral for Vassiliev invariants of framed knots. © *1998 American Institute of Physics.* [S0022-2488(98)02909-0]

I. INTRODUCTION

Chern–Simons gauge theory has provided a very useful tool to study different aspects of the theory of knot and link invariants. Since its formulation by Witten in 1988[1] it has been studied from both perturbative and nonperturbative points of view. While nonperturbative methods[1–6] have led to its connection to polynomial invariants as the Jones polynomial[7] and its generalizations,[8–10] perturbative ones[11–14] have provided representations of Vassiliev invariants.

One of the advantages of the Chern–Simons approach to Vassiliev invariants is that it provides different representations of them because the perturbative analysis of the theory can be carried out in different gauges. Most of the perturbative analysis has been performed in covariant gauges, mostly in the Landau gauge. The loop structure in this gauge has been extensively analyzed in Refs. 11 and 15–17. Vacuum expectation values of Wilson loops have also been studied in a variety of papers.[11–14,18,19] These analyses lead to a representation of Vassiliev invariants for knots and links, which is equivalent to the one proposed by Bott and Taubes in Ref. 20.

The perturbative Chern–Simons series expansion in noncovariant gauges has not been studied so extensively. The loop structure has been analyzed in the light-cone gauge in Refs. 21 and 22. A first step into the study of vacuum expectation values of Wilson loops was presented in Ref. 23, where a second-order analysis was carried out in the axial gauge. No systematic study of the perturbative series expansion corresponding to Wilson loops has been carried out in the light-cone gauge. The main goal of this paper is to begin with this study. In the process we find that the light-cone gauge leads to the Kontsevich integral[24,25] for Vassiliev invariants of framed knots.

Chern–Simons gauge theory was first studied in the light-cone gauge in Ref. 4. That paper points out a close relation between the vacuum expectation value of a Wilson line and the Knizhnik–Zamolodchikov equations.[26] This fact is used to prove that, for SU(N) as gauge group, vacuum expectation values of Wilson loops are related to the HOMFLY polynomial[8] for links. From a perturbative point of view, vacuum expectation values of Wilson loops have been considered only in Ref. 27, where it is conjectured that the corresponding perturbative series expansion

a)Electronic mail: labasti@nxth04.cern.ch

 © 1998 American Institute of Physics

is related to the Kontsevich integral[24] for Vassiliev invariants. In the present paper we prove this conjecture.

Vacuum expectation values of Wilson loops are ill-defined in the light-cone gauge because of the presence of singularities. To have a well-defined perturbative expansion we introduce framed Wilson loops. The fact that the right object to be studied in Chern–Simons gauge theory is a framed Wilson loop has been known since the theory was first formulated in Ref. 1. The reason to introduce a framing might result differently in each approach, but in general it is related to the presence of singularities in the n-point correlation functions at coincident points. This will indeed be the case in the light-cone gauge. A version of the Kontsevich integral for Vassiliev invariants adapted to the case of framed knots and links was presented in Ref. 25. In this paper we will consider the case of framed knots but a similar construction holds for framed links. We show that the vacuum expectation value of a Wilson loop contains the Kontsevich integrals introduced in Ref. 25.

The paper is organized as follows. In Sec. II we discuss the quantization of Chern–Simons gauge theory in the light-cone gauge. In Sec. III we analyze the general features of the perturbative series expansion of Wilson loops. In Sec. IV, after introducing a regularization that involves framed knots, we prove the finiteness of the perturbative series expansion. In addition, we extract the framing dependence and show that the perturbative series contains the Kontsevich integral for Vassiliev invariants of framed knots. Finally, in Sec. V we state our conclusions and we discuss some open problems and future work.

II. CHERN–SIMONS GAUGE THEORY IN THE LIGHT-CONE GAUGE

In this section we introduce Chern–Simons gauge theory in the light-cone gauge from a perturbation theory point of view. Let us consider a semisimple compact Lie group G and a connection A. The action of the Chern–Simons theory over a three-dimensional Minkowski space $\mathbf{M}^3$ is defined by the integral:

$$S_{\mathrm{CS}}(A)=\frac{k}{4\pi}\int_{\mathbf{M}^3}\mathrm{Tr}\left(A\wedge dA+\frac{2}{3}A\wedge A\wedge A\right), \tag{2.1}$$

where Tr denotes the trace over the fundamental representation of G, and k is a real parameter. This action is invariant under gauge transformation

$$A_\mu\rightarrow h^{-1}A_\mu h+h^{-1}\partial_\mu h, \tag{2.2}$$

where h is a map from $\mathbf{M}^3$ to G, which is connected to the identity map. For maps that are not connected to the identity, (2.1) transforms into itself plus a term of the form $2\pi kn$, n being the winding number of the map. If k is an integer, the exponential of the action, $\exp(iS_{\mathrm{CS}})$, which is what enters the functional integral, is invariant under both types of gauge transformations.

Relative to the action (2.1) and to the gauge transformation (2.2), we choose the following conventions. The gauge group generators, T^a, $a=1\cdots\dim(G)$, will be taken anti-Hermitian, satisfying

$$[T^a,T^b]=-f_{abc}T^c, \tag{2.3}$$

f_{abc} being the structure constants of the group G. They are normalized so that

$$\mathrm{Tr}(T^aT^b)=-\tfrac{1}{2}\delta^{ab} \tag{2.4}$$

in the fundamental representation. A point in $\mathbf{M}^3$ will be denoted by coordinates x^μ, with μ running from zero to two, $\mu=2$ being the time coordinate.

In order to define the perturbative series expansion associated to the action (2.1) we must make a gauge choice. This choice is defined by a gauge-fixing condition, which in our case will be corresponding to the light-cone gauge:

$$n^\mu A_\mu=0, \tag{2.5}$$

where n^μ is a constant vector satisfying $n^2=0$. The corresponding gauge-fixing term to be added to the action (2.1) is

$$S_{\rm gf}=\int_{\mathbf{M}^3} d^3x\ \mathrm{Tr}(dn^\mu A_\mu+bn^\mu D_\mu c), \tag{2.6}$$

where d is an auxiliary field, and c and b are ghost fields; D_μ stands for the covariant derivative, $D_\mu c=\partial_\mu c+[A_\mu,c]$. In order to define the perturbative series expansion it is convenient to rescale the fields by $A\to gA$, where $g=\sqrt{4\pi/k}$. The quantum action, $S=S_{\rm CS}+S_{\rm gf}$, becomes

$$S=-\frac{1}{2}\int_{\mathbf{M}^3} d^3x\,\epsilon^{\mu\nu\rho}\left(A^a_\mu\partial_\nu A^a_\rho-\frac{g}{3}f_{abc}A^a_\mu A^b_\nu A^c_\rho\right)-\frac{1}{2}\int_{\mathbf{M}^3} d^3x(d^aA^a_\mu n^\mu+b^an^\mu D^{ab}_\mu c^b), \tag{2.7}$$

where

$$D^{ab}_\mu=\partial_\mu\delta^{ab}-gf_{abc}A^c_\mu. \tag{2.8}$$

Following Ref. 4 we introduce light-cone coordinates:

$$x^+=x^1+x^2,\quad x^-=x^1-x^2, \tag{2.9}$$

and light-cone components for the gauge connection:

$$A_+=A_1+A_2,\quad A_-=A_1-A_2. \tag{2.10}$$

Choosing the vector n^μ as $(0,1,-1)$ the gauge-fixing condition (2.5) implies $A_-=0$. Notice that the gauge condition (2.5) does not fix the gauge completely. For example, in our particular choice of n^μ we still have gauge invariance under gauge transformations (2.2), in which the gauge parameter h depends only on x^0 and x^+. The quantum action (2.7) takes the following form:

$$S=\int_{\mathbf{M}^3} d^3x(A^a_+\partial_-A^a_0-A^a_0\partial_-A^a_+-b^a\partial_-c^a). \tag{2.11}$$

Actually, to consider this action as the quantum action of the theory constitutes a rather simplified version of the full story. The perturbative higher-loop analysis in axial-type gauges is a very delicate issue, which requires us to take into consideration some specific prescription to regulate unphysical poles. Fortunately, this analysis has been done for the case of Chern–Simons gauge theory in Ref. 21. In these works it is shown that the effect of higher-loop contributions is a shift of the parameter k entering the Chern–Simons action (2.1) by the quantity c_v, which denotes the value of the quadratic Casimir in the adjoint representation of the gauge group. Though strictly speaking this has been proved at one loop, it is believed that, as in the case of covariant gauges, it holds at any order in higher loops. In this paper we will assume that higher-loop effects just account for the shift of the parameter k in (2.1), and we will work with (2.11) as the quantum action of the theory.

The quantum action (2.11) has three important properties. First, it does not have derivatives in the transverse direction, second, it is quadratic in the fields, and third, the ghost fields are not coupled to the gauge fields. This last property implies that the ghost fields can be integrated out trivially. The second property implies that there are no interaction vertices and therefore that all the correlation functions are determined by the two-point correlation functions. Certainly, this property notably simplifies the structure of the perturbative series expansion.

The two-point correlation functions corresponding to the gauge fields entering (2.11) have been computed in Ref. 4. Their result is more conveniently expressed after performing a Wick rotation to Euclidean space $\mathbf{R}\times\mathbf{C}$. A point in Euclidean space will be denoted as (t,z), where $z=x^1+ix^2$. After introducing $A_z=A_1+iA_2$ and $A_{\bar z}=A_1-iA_2$ one finds:[4]

$$\langle A^a_z(x)A^b_m(x')\rangle=0,\quad \langle A^a_m(x)A^b_n(x')\rangle=\delta^{ab}\epsilon_{mn}\frac{1}{2\pi i}\frac{\delta(t-t')}{z-z'}, \tag{2.12}$$

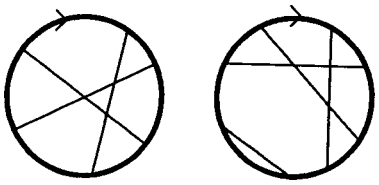

FIG. 1. Examples of Feynman diagrams.

with m, $n=\{0,z\}$, and ϵ_{mn} is antisymmetric with $\epsilon_{0z}=1$.

III. PERTURBATIVE EXPANSION OF THE WILSON LOOP

Wilson loops are gauge-invariant operators of Chern–Simons gauge theory labeled by a loop C embedded in $\mathbf{M}^3$ and a representation R of the gauge group G. They are defined by the holonomy along the loop C of the gauge connection A:

$$\mathscr{W}_R(C,G)=\mathrm{Tr}\left[P_R\ \exp\ g\ \oint A\right], \tag{3.1}$$

where P_R denotes that the integral is path-ordered and that A must be considered in the representation R of G. In Chern–Simons gauge theory one considers vacuum expectation values of products of Wilson-line operators:

$$\begin{aligned}&\langle \mathscr{W}_{R_1}(C_1,G)\mathscr{W}_{R_2}(C_2,G)\cdots\mathscr{W}_{R_n}(C_n,G)\rangle\\ &\quad=\frac{1}{Z_k}\int [DA]\mathscr{W}_{R_1}(C_1,G)\mathscr{W}_{R_2}(C_2,G)\cdots\mathscr{W}_{R_n}(C_n,G)e^{iS_k(A)},\end{aligned} \tag{3.2}$$

where Z_k is the partition function of the theory:

$$Z_k=\int [DA]e^{iS_{\mathrm{CS}}(A)}. \tag{3.3}$$

As shown in Ref. 1 the quantity (3.2) is a link invariant associated to a colored n-component link whose j-component C_j carries the representation R_j of the gauge group G. In this paper we will consider only the vacuum expectation value of a single loop $\langle W_R(C,G)\rangle$. This quantity can be expressed as a perturbative series expansion in the coupling constant g, which is the result of evaluating all the corresponding Feynman diagrams. The structure of these diagrams is obtained using standard field theory methods. A typical diagram of order i consists of a solid thick circle representing the oriented path C, with i propagators attached to it in a certain order (see some examples in Fig. 1). A term in the sum of the perturbative series may be regarded as constructed out of two Feynman rules: the one corresponding to the propagator (2.12):

$$D^{ab}_{mn}(x-x')=\delta^{ab}\epsilon_{mn}\frac{1}{2\pi i}\frac{\delta(t-t')}{z-z'}, \tag{3.4}$$

while all other components of $D_{\mu\nu}$ vanish; and the one corresponding to the vertex between the end point of a propagator and the oriented path C:

$$V^{j\mu a}_i(x)=g(T^a_{(R)})^j_i\int dx^\mu. \tag{3.5}$$

Notice that, as discussed in the previous section, there are not three-vertices in the light-cone gauge. The Feynman rules are depicted in Fig. 2.

The power series expansion corresponding to $\langle \mathscr{W}_R(C,G)\rangle$ can be written as:[13]

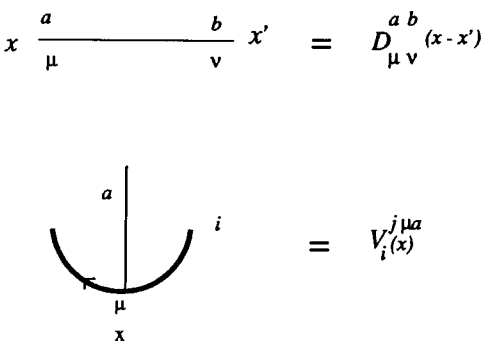

FIG. 2. Feynman rules.

$$\langle \mathscr{W}_R(C,G)\rangle = \dim R \sum_{i=0}^{\infty} \sum_{j=1}^{d_i} \alpha_{ij}(C) r_{ij}(R) x^i, \tag{3.6}$$

where $x = ig^2/2$ is the expansion parameter. The quantities $\alpha_{ij}(C)$, or geometrical factors, are combinations of path integrals of some kernels along the loop C, and the r_{ij} are traces of products of generators of the Lie algebra associated to the gauge group G. The index i corresponds to the order in perturbation theory, and j labels independent contributions at a given order, d_i being the number of these at order i. In (3.6) dim R denotes the dimension of the representation R. Notice the convention $\alpha_{01}(C) = 1$. For a given order in perturbation theory, $\{r_{ij}\}_{\{j=1\cdots d_i\}}$ represents a basis of independent group factors.

The choice of basis in (3.6) is not unique. As shown in Ref. 19, there is a special type of basis, called canonical, for which the series (3.6) satisfies a factorization theorem. Let us introduce some notation to recall the definition of a canonical basis. A group factor is an element in the center of the universal enveloping algebra of G, its general form being a trace over products of generators and structure constants with all indices contracted. It may also be represented in terms of diagrams that look like Feynman diagrams in which propagators and three-vertices are attached to a circle in a certain order, which represents the trace. In this way one considers three group-theoretical Feynman rules as the ones depicted in Fig. 3. A canonical basis consists of an independent set of group factors made out of connected diagrams or products of nonoverlapping connected subdiagrams (see Ref. 19 for more details).

The factorization theorem[19] states that (3.6) may be written as an exponentiation over the elements of a canonical basis, which are not the product of connected subdiagrams. These elements are called primitive and (3.6) takes the form:

$$\langle \mathscr{W}_R(C,G)\rangle = \dim R \, \exp\left\{ \sum_{i=0}^{\infty} \sum_{j=1}^{\hat{d}_i} \alpha^c_{ij}(C) r^c_{ij}(R) x^i \right\}. \tag{3.7}$$

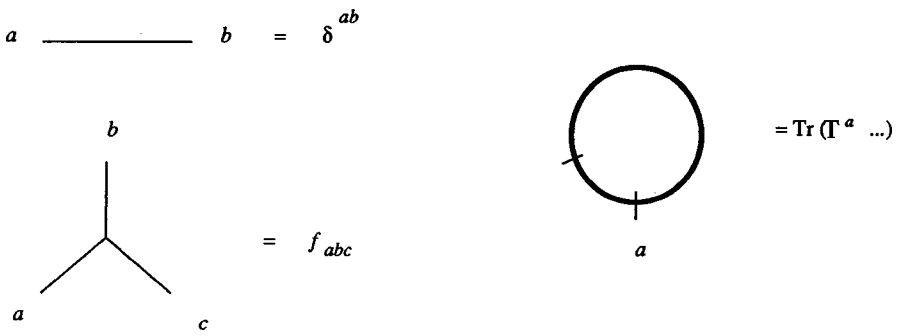

FIG. 3. Group-theoretical rules.

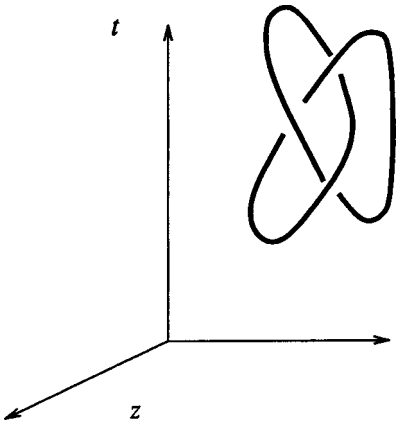

FIG. 4. Example of a Morse knot.

In this expression the superindex c denotes the primitive connected diagrams. The quantities $\alpha_{ij}^{c}(C)$ are their corresponding geometrical factors, which in general are linear combinations of the $\alpha_{ij}(C)$ entering (3.6). In (3.7), $\hat{d}_i$ denotes the number of independent primitive group factors at order i. For a more detailed discussion about the properties of (3.7) we refer the reader to Ref. 19. Here we will recall only a few facts. In the series expansion (3.6) of $\langle \mathscr{W}_R(C,G) \rangle$ there is a class of diagrams that contain at least one collapsible propagator. A propagator is called collapsible whenever its two legs are attached to two points in the oriented path C, which can be considered as the end points of a part of C in which no other propagator is attached. Notice that in the right-hand side of (3.7) all these diagrams have factorized out as the exponentiation of the diagram of order 1, its geometrical factor being $\alpha_{11}^{c}(C)$. The perturbative analysis carried out in the Landau gauge in Ref. 13 demonstrated that if one chooses a canonical basis all the dependence on the framing is contained in $\alpha_{11}^{c}(C)$. In this paper we will show that, as expected, this also holds when the theory is analyzed in the light-cone gauge.

As shown in Refs. 28 and 29, each term of the perturbative series expansion (3.6) is a Vassiliev invariant. The quantities $\alpha_{ij}^{c}(C)$, $i>1$, in the exponential of (3.7) represent a particular choice of a basis of primitive ones. The analysis of the perturbative series expansion of the theory in the light-cone gauge will provide integral expressions for these invariants. Actually, in order to make contact with the Kontsevich integral we are not going to construct the integral expression for the primitive elements $\alpha_{ij}^{c}(C)$, $i>1$. We will not make the choice of a canonical basis except in our discussion on the framing dependence. Instead, we will analyze the perturbative series expansion as it results from the application of the Feynman rules without selecting a particular basis of group factors. As will be shown in the next section, this will lead us to the Kontsevich integral for Vassiliev invariants.

Before entering into the analysis of the perturbative series expansion corresponding to a Wilson loop, we must discuss the potential problems that might be encountered because of the particular form of the gauge-field propagator (3.4). This propagator is singular when its two end points coincide. Actually, it is particularly singular in this situation, because both the numerator and the denominator lead to divergences. This fact tells us that, as in other gauges, one must consider Wilson loops with oriented paths C with no self-intersections. In other words, one must consider knots. However, contrary to covariant gauges, in the light-cone gauge we have two special kinds of singularities, there may be situations in which only one, the numerator or the denominator, leads to a divergence. In order to avoid singularities from the numerators one is forced to avoid paths with sections in which t, the first component of a generic point (t,z) in $\mathbf{R}\times\mathbf{C}$, is constant. This constraint, together with the fact that one is only allowed to consider the non-self-intersecting paths, implies that one must consider paths C, which correspond to Morse knots. A Morse knot in $\mathbf{R}\times\mathbf{C}$ is a knot in which t is a Morse function on it. A Morse knot is characterized by $2n$ extrema, half of them maxima, and the other half minima. An example of a Morse knot is depicted in Fig. 4.

The third potential problem due to the structure of (3.4) comes from situations in which the

two end points of the propagator are close to one of the extrema of a Morse knot since the denominator then vanishes. To solve this problem we will introduce a regularization procedure based on the introduction of a framing for the knot. The resulting invariants will correspond to invariants of framed knots.

IV. THE KONTSEVICH INTEGRAL FOR FRAMED KNOTS

In this section we will apply the Feynman rules derived in the previous section to construct the perturbative series expansion corresponding to a Wilson loop for a Morse knot. As argued before, this requires the introduction of a regularization procedure; this will be achieved by considering framed knots. In the first part of this section we will describe this procedure in detail. In three subsequent subsections we will first prove the finiteness of the resulting perturbative series expansion, then we will extract the framing dependence, and, finally, we will obtain the Kontsevich integral for framed knots.

Let us consider a framed oriented Morse knot K on $\mathbf{R}\times\mathbf{C}$. The framing of K is defined by a vector $m^{\mu}(K)$ normal to K such that along the coordinates $(t(a),z(a))$, $a\in S^1$, which define K, we have another set of coordinates $(t(a),z(a))+\varepsilon(m^0(t,z),m^z(t,z))$, which define a companion knot K_{ε}. For small ε, K_{ε} is also a Morse knot, which does not intersect K and becomes K in the limit $\varepsilon\to 0$. We assign to K_{ε} the orientation that is naturally inherited from K. The framing is characterized by the vector $m^{\mu}(K)$ or, equivalently, by the companion knot K_{ε}. It is natural to associate to each choice of framing the integer number that corresponds to the linking number, l, between K and K_{ε}.

Our regularization of the vacuum expectation value of a Wilson loop is based on the replacement of the propagator (3.4), which is attached to two points on K, by a propagator attached to a point in K and another in K_{ε}. As it is described below, there exists a precise way to carry out this replacement. Certainly, the resulting regularized integrals entering the power series expansion are finite for Morse knots. We have to show that they remain also finite in the limit $\varepsilon\to 0$. Before doing this, let us give full details of how the propagator replacement is carried out.

Let us assume that the Morse knot K possesses $2n$ extrema. There are $2n$ curves k^i, $i=1,\ldots,2n$, joining the different maxima and minima of K. For each curve k^i there is a one-to-one correspondence between points on k^i and the values that the variable t takes. The Morse knot can therefore be regarded as a complex multivalued function of the variable t with $2n$ components, each one corresponding to a curve k^i, which is completely labeled by the values that the complex variable z takes in k^i as a function of t. One can think of two different parametrizations of K, the previous one: $(t(a),z(a))$, $a\in S^1$, and a new one: $z^i(t)$, $t\in I^i$, $i=1,\ldots,2n$, where $I^i=[t_i^-,t_i^+]$ is the segment of $\mathbf{R}$ whose end points are the values that the coordinate t takes at the two extrema joined by the curve k^i. A similar analysis can be done for the companion knot K_{ε}. It also possesses $2n$ extrema and $2n$ curves k_{ε}^i, parametrized by $z_i'(t)$, $t\in I_{\varepsilon}^i$, $i=1,\ldots,2n$.

It is convenient to trade the integrations along the S^1 parameter a of K by integrations over the height parameter t. Let us consider a propagator attached to the curves k^i and k^j. Using (3.4), it is easy to show that for such a propagator:

$$dx^{\mu}dx^{\nu}\langle A_{\mu}^{a}(x)A_{\nu}^{b}(x')\rangle\to\frac{1}{2\pi i}\,dt_i dt_j\delta(t_i-t_j)p_{ij}\,\frac{\dot{z}_i(t_i)-\dot{z}_j(t_j)}{z_i(t_i)-z_j(t_j)},\tag{4.1}$$

where

$$p_{ij}=\begin{cases}1 & \text{if } k_i \text{ and } k_j \text{ have the same orientation}\\ -1 & \text{if } k_i \text{ and } k_j \text{ have opposite orientations.}\end{cases}\tag{4.2}$$

The propagators (4.1) may appear attached to two points lying on different curves k^i and k^j, $i\neq j$, or to two points lying on the same curve k^i. In the second case the regularization consists of just replacing one of the points on k^i of the propagator by a point on k_{ε}^i. Since the delta function in the propagator (4.1) implies that its two end points must be at the same height, there is no ambiguity in doing this. Notice, however, that in this case, since we have a path-ordered integration, when evaluating the delta function in (4.1) we must take into account the appearance of a factor 1/2. In the first case, $i\neq j$, one uses the same procedure, namely one of the end points is

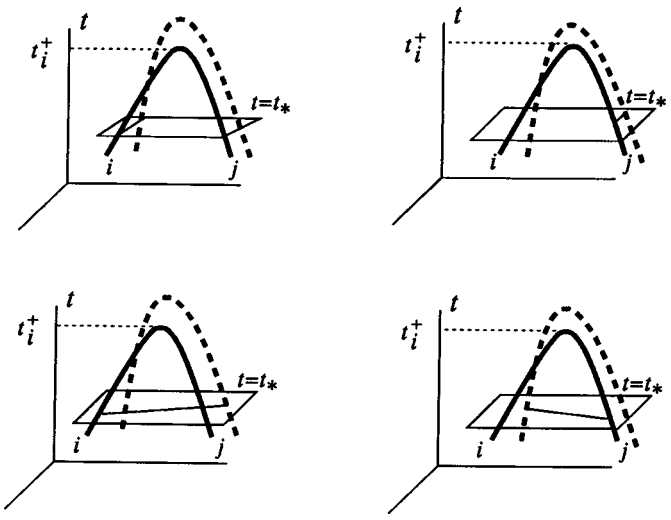

FIG. 5. Generic divergent contributions.

attached to the curve corresponding to the companion knot. However, there are now two possible ways of doing this. We will use the following prescription: we will add both possibilities and multiply the result by a factor 1/2. For the same reasons as in the second case, this last step is well defined. Notice that in this case there is no additional factor 1/2 after the evaluation of the delta functions, since the variables of integration in each curve are now free.

The regularization prescription leads to the following formulas for the propagator (4.1). For $i=j$:

$$\frac{1}{2\pi i}\frac{1}{2}\,ds\,\frac{\dot{z}_i(s)-\dot{z}_i'(s)}{z_i(s)-z_i'(s)}, \tag{4.3}$$

while for $i\neq j$:

$$\frac{1}{2\pi i}\frac{1}{2}\,ds\left(\frac{\dot{z}_i(s)-\dot{z}_j'(s)}{z_i(s)-z_j'(s)}+\frac{\dot{z}_i'(s)-\dot{z}_j(s)}{z_i'(s)-z_j(s)}\right)p_{ij}\,. \tag{4.4}$$

A. Finiteness

We will now prove that our prescription is finite in the limit $\varepsilon\to 0$. There are four sources of singularities. All originate when the two end points of a propagator are near an extremum. We will describe the situation for the case in which this extremum is a maximum, but it will become clear that a similar analysis can be carried out in the case in which it corresponds to a minimum. The four configurations that lead to divergences are depicted in Fig. 5. We have drawn only the part of K around the maximum. We will assume that the rest of the diagram is the same in the four cases. Let us compute the four contributions after integrating from t_* to t_i^+. Using (4.3) and (4.4) one finds, after adding them up:

$$\frac{1}{2\pi i}\frac{1}{2}\log\frac{(z_i(s)-z_i'(s))(z_j(s)-z_j'(s))}{(z_i(s)-z_j'(s))(z_i'(s)-z_j(s))}\bigg|_{s=t}^{s=t_i^+}. \tag{4.5}$$

The most dangerous contribution is the one coming from the upper limit. However, since $|z_i(t_i^+)-z_i'(t_i^+)|=|z_j(t_i^+)-z_i'(t_i^+)|=\varepsilon$, and $|z_i(t_i^+)-z_j'(t_i^+)|=|z_j(t_i^+)-z_j'(t_i^+)|=\varepsilon$, one finds a cancellation of divergences in the limit $\varepsilon\to 0$ and therefore the contribution is finite. One also encounters divergences from the lower limit of (4.5) in the $\varepsilon\to 0$ limit. However, these divergences cancel against the ones originated after integration under the height t_*. One is therefore left with a term of the form:

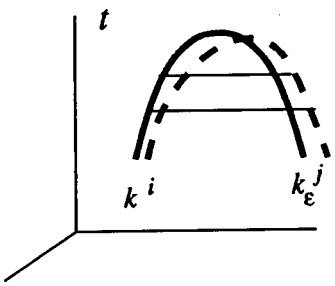

FIG. 6. Higher-order divergent contributions.

$$-\frac{1}{2\pi i}\frac{1}{2}\log((z_i(t_*)-z_j'))(z_i'(t_*)-z_j(t_*))), \tag{4.6}$$

which has to be taken into account when integrating over t_*. Notice that (4.6) is finite in the limit $\varepsilon\to 0$ for values of t_* away from t_i^+, but it diverges when t_* approaches t_i^+. However, this divergence is too soft and it does not generate singularities in the integration, except in situations of the kind depicted in Fig. 6. All the contributions of this type will be treated below when analyzing the factorization of the framing dependence, and it will be shown that all possible divergences cancel out. We can therefore affirm that the perturbative series expansion corresponding to a Morse knot is finite in the limit $\varepsilon\to 0$.

B. The framing contribution

In this section we are going to compute the lowest-order contribution to the perturbative series expansion of the vacuum expectation value of a Wilson loop. Using the factorization theorem discussed in Sec. III, we will obtain the full dependence on the framing. Our result can be stated very simply: if K is a framed Morse knot:

$$\langle \mathscr{W}_R(K,G)\rangle = e^{2\pi i l h}\langle \mathscr{W}_R(K,G)\rangle', \tag{4.7}$$

where $h=\mathrm{Tr}(T^a_{(R)}T^a_{(R)})/k$, l is the linking number between K and its companion knot K_e, and $\langle \mathscr{W}_R(K,G)\rangle'$ is a framing-independent quantity. This result agrees with the one found nonperturbatively,[1] and in covariant gauges.[13]

The lowest-order contribution to the perturbative series expansion has the form:

$$ig^2\frac{1}{2\pi i}\frac{1}{2}\mathrm{Tr}(T^a_{(R)}T^a_{(R)})\sum_{i,j=1}^{2n}\int_{\max\{t_i^-,t_j^-\}}^{\min\{t_i^+,t_j^+\}} ds\,\frac{\dot z_i(s)-\dot z_j'(s)}{z_i(s)-z_j'(s)}\,p_{ij}. \tag{4.8}$$

Notice that, owing to the prescription (4.4), the sum is over all possible values of i and j and not only for $i\leqslant j$. The factor i in front of this expression is due to the fact that one is considering the functional integral of i times the action (2.7). In general, in the perturbative series expansion, one must include a factor i for each power of g^2. According to our previous arguments, the integral (4.8) is certainly finite in the limit $\varepsilon\to 0$. Actually it can be computed very easily. The key observation is that an integral of the form

$$\int_{t_*}^{t'_*} ds\,\frac{\dot z_i(s)-\dot z_j'(s)}{z_i(s)-z_j'(s)} \tag{4.9}$$

develops an imaginary part, which counts the number of times that the curve k_e^j winds around the curve k^i, times 2π. The real parts just cancel with each other, following a mechanism similar to the one described in our proof of finiteness. After summing over all the values of i, the contribution from the integral in (4.8) for $i=j$ is just $2\pi i$ times the twist tw of the band made by the two knots K and K_e, i.e., the number of times that the band twists around itself. The basic contributions to tw are shown in Fig. 7. The analysis of the contributions from $i\neq j$ is very similar. The real parts cancel among themselves and some of the real contributions from the case $i=j$, as

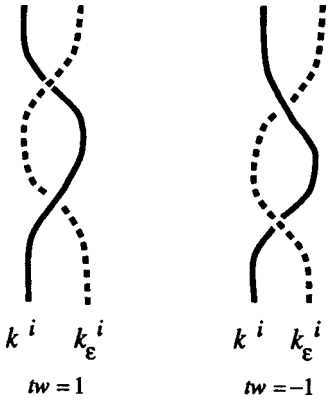

FIG. 7. Basic contributions to the twist of a framed knot.

described in the proof of finiteness. The imaginary parts now count the number of times that the curve k^i twists around the curve k^j, times 2π. Certainly, for this counting one can take the limit $\varepsilon \to 0$. If the Morse knot K is viewed in such a way that its projection in the plane corresponding to Im $z=0$ contains only single crossings, one can compute the contribution just assigning values $\pm 1/2$ to the two types of crossings shown in Fig. 8. The result of performing the integral in (4.8) is precisely $2\pi i$ times the writhe w of the knot K for that given view. Notice that although we get factors of modulus 1/2 from each crossing, they are counted twice in (4.8). Thus the full contribution from (4.8) is

$$ig^2 \tfrac{1}{2} \operatorname{Tr}(T^a_{(R)} T^a_{(R)})(w+tw) = 2\pi i h l \tag{4.10}$$

since $l = w + tw$. In obtaining (4.10) we have used $g^2 = 4\pi/k$. Recall that $h = \operatorname{Tr}(T^a_{(R)} T^a_{(R)})/k$. Notice that, although w and tw depend on the view of the knot K, their combination l, the linking number between K and K_ε, is independent of it.

As shown in previous works,[13,19] the contribution found at first order exponentiates if one considers a canonical basis. Let us describe the reasons for this exponentiation by analyzing what occurs at next order in perturbation theory. At second order one encounters the two group factors corresponding to the Feynman diagrams shown in Fig. 9. Diagrams *a* and *b* share the same group factor:

$$\operatorname{Tr}(T^a_{(R)} T^a_{(R)} T^b_{(R)} T^b_{(R)}), \tag{4.11}$$

while diagram *c* possesses a different one:

$$\operatorname{Tr}(T^a_{(R)} T^b_{(R)} T^a_{(R)} T^b_{(R)}). \tag{4.12}$$

If one decomposes this group factor in terms of the first one and a new group factor using the commutation relations (2.3):

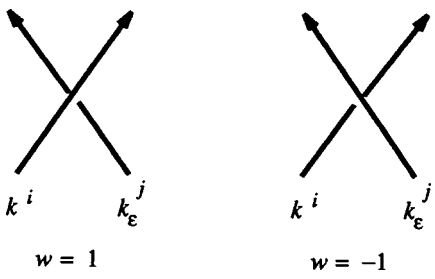

FIG. 8. Basic contributions to the writhe.

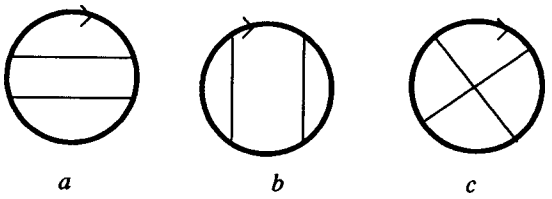

FIG. 9. Feynman diagrams contributing at second order.

$$\mathrm{Tr}(T^a_{(R)}T^b_{(R)}T^a_{(R)}T^b_{(R)}) = \mathrm{Tr}(T^a_{(R)}T^b_{(R)}T^b_{(R)}T^a_{(R)}) - f_{abc}\ \mathrm{Tr}(T^a_{(R)}T^b_{(R)}T^c_{(R)}), \tag{4.13}$$

the full sum of the geometrical terms multiplying the first group factor can be written as 1/2 times a product of integrations over contributions coming from a single propagator. It precisely gives the second-order term of the exponential of (4.10). A similar rearrangement can be done at any order. Notice that in order to extract the terms building the exponential of (4.10) one modifies the group factors that remain in the rest. This mechanism is encoded in the factorization theorem.[19] Recall that in our analysis of finiteness we postponed the discussion on the cancellation of divergences from diagrams as the one in Fig. 6. Now is the moment to discuss that issue. These diagrams are indeed the ones having group factors as (4.11). Together with the ones coming after rearranging group factors as done in (4.13), after adding them up, one obtains an expression that can be written as a product of the contribution from the first order. Since that first-order contribution is finite, the contribution at any arbitrary higher order is also finite. Notice also that this argument shows that all the dependence on the framing is contained in the exponential factor and therefore we can affirm that (4.7) holds. Indeed all the potential dependence on the framing must come from diagrams that lead to divergences. As we have seen, these diagrams only contribute to the exponential factor in (4.7). This behavior is entirely similar to that occurring in the Landau gauge.[13]

C. The Kontsevich integral

In this section we will prove that the perturbative series expansion for the vacuum expectation value of the Wilson loop contains the Kontsevich integral for framed knots as presented in Ref. 25.

Let us begin by writing all the contributions to a given order m. To carry this out we must consider all possible ways of connecting $2m$ points on the Morse knot by m propagators, following the regularization prescription described in the previous section, i.e., with one point of the propagator attached to K and the other to its companion knot K_ε, and then path-order integrating. The path-order integration can be split into a sum such that enters in each term a path-ordered integration along $2m$ curves among the set k^i, k^i_ε, $i=1,...,2n$. This set of curves builds the Morse knot under consideration. A given term in this sum might contain propagators joining k^i and k^i_ε. In this case one must introduce a factor 1/2 as explained in our discussion leading to (4.3). The contributions coming from propagators joining k^i and k^j_ε, with $i\neq j$, have also a factor 1/2 due to the double counting, as explained in the discussion of Eq. (4.4). Accordingly, propagators joining different curves must be replaced by 1/2 the sum of their two possible choices of attaching their end points.

To each rearrangement of the m propagators corresponds a group factor. These are easily obtained using the group-theoretical Feynman rules shown in Fig. 3. To fix ideas we will present in detail the second-order contribution, $m=2$, for a particular group factor. For $m=2$ one must take into consideration the two group factors (4.11) and (4.12). As discussed above, the group factor (4.11) is associated to the framing dependence. We will analyze the contribution corresponding to the group factor (4.12). This contribution is of the form

$$(ig)^2\int_0^1 ds_1\int_{s_1}^1 ds_2\int_{s_2}^1 ds_3\int_{s_3}^1 ds_4\dot{x}^{\mu_1}(s_1)\dot{x}^{\mu_2}(s_2)\dot{x}^{\mu_3}(s_3)\dot{x}^{\mu_4}(s_4\times\langle A^{a_1}_{\mu_1}(x(s_1))A^{a_3}_{\mu_3}(x(s_3))\rangle$$
$$\times\langle A^{a_2}_{\mu_2}(x(s_2))A^{a_4}_{\mu_4}(x(s_4))\rangle\mathrm{Tr}(T^{a_1}T^{a_2}T^{a_3}T^{a_4}). \tag{4.14}$$

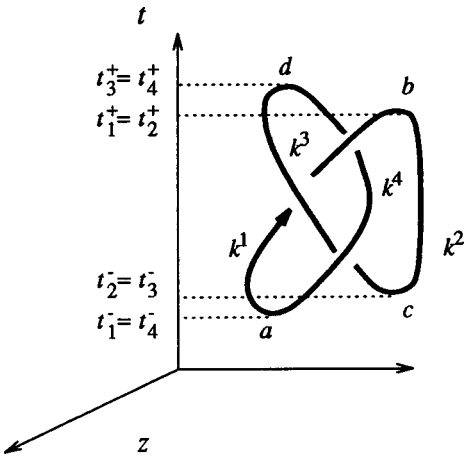

FIG. 10. Labeled trefoil knot.

We will now write more explicitly this multiple integral taking into account the form of the propagator (3.4). The delta function in this propagator imposes very strong restrictions on the possible contributions. Its presence implies that the only nonvanishing configurations are those in which the two end points of each propagator are at the same height. To be more concrete let us consider the computation of (4.14) for the trefoil knot shown in Fig. 10. This knot is made out of four curves, k^1, k^2, k^3, and k^4, whose end points are the four critical points a, b, c, and d. The heights of these critical points are

$$a \to t_1^- = t_4^- , \quad b \to t_1^+ = t_2^+ , \quad c \to t_2^- = t_3^- , \quad d \to t_3^+ = t_4^+ . \tag{41.5}$$

They are depicted in Fig. 10.

To obtain all the contributions we will divide in four parts the circle that represents the knot in Fig. 10. Then we will join these parts by lines representing the propagators, taking into account the ordering of the four points to which they are attached. This ordering and the delta function in the height imply that no line can have its two end points attached to the same part. They also imply that there are no contributions in which two end points of different lines are attached to one part and the other two to another part. The only possibilities are shown in Fig. 11. There is a total of eight contributions. Notice that this result is general for any Morse knot made out of four curves. The contributions are easily depicted on the knot itself as shown in Fig. 12. For each contribution,

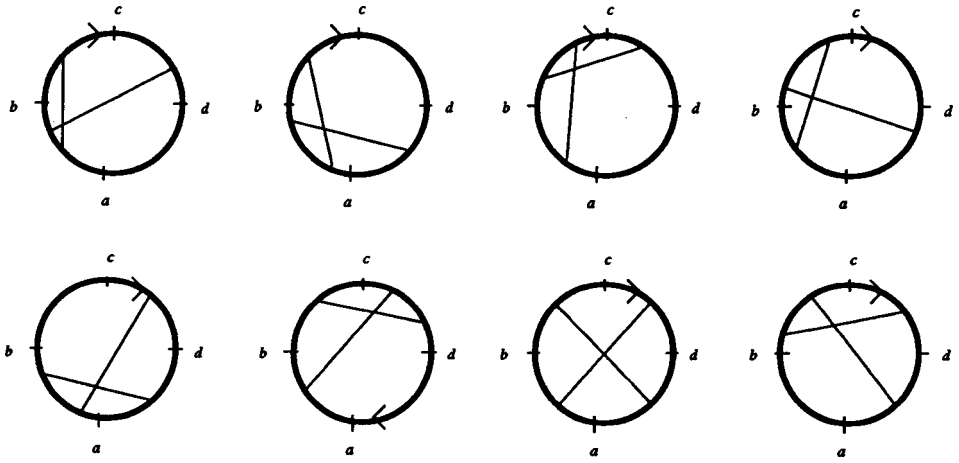

FIG. 11. Contributions to a Morse knot made out of four curves.

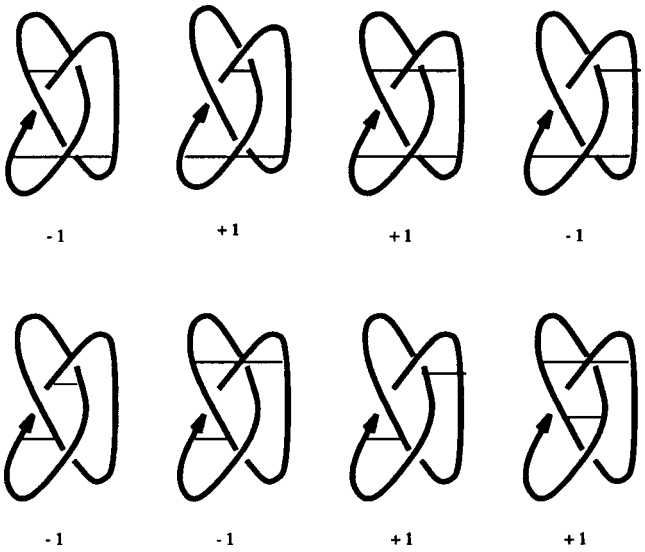

FIG. 12. Contributions corresponding to the trefoil knot.

one must compute a sign, which is the product of the p_{ij} in (4.2). The resulting signs are displayed in Fig. 12. To be more explicit, let us write, for example, the integral associated to the first contribution. It takes the form

$$(ig)^2 \frac{1}{(2\pi i)^2} \frac{1}{2^2} \int_{t_2^- < s_1 < s_2 < t_1^+} ds_1 ds_2 \times \left(\frac{\dot{z}_1(s_1) - \dot{z}_2'(s_1)}{z_1(s_1) - z_2'(s_1)} + \frac{\dot{z}_1'(s_1) - \dot{z}_2(s_1)}{z_1'(s_1) - z_2(s_1)} \right) \left(\frac{\dot{z}_3(s_2) - \dot{z}_1'(s_2)}{z_3(s_2) - z_1'(s_2)} + \frac{\dot{z}_3'(s_2) - \dot{z}_1(s_2)}{z_3'(s_2) - z_1(s_2)} \right). \quad (4.16)$$

Similar expressions can be easily written for the rest of the contributions depicted in Fig. 12. The data entering in the double integral (4.16) are shown in Fig. 13. Notice that this integral is not divergent if we take the limit $\varepsilon \to 0$ before performing the integration. This feature is common to all the contributions corresponding to the group factor under consideration. As shown in our discussion on finiteness, only the contributions related to framing are potentially divergent. One could therefore remove in (4.16) the terms with primes and the factor $1/2^2$. The integral to be computed takes the form:

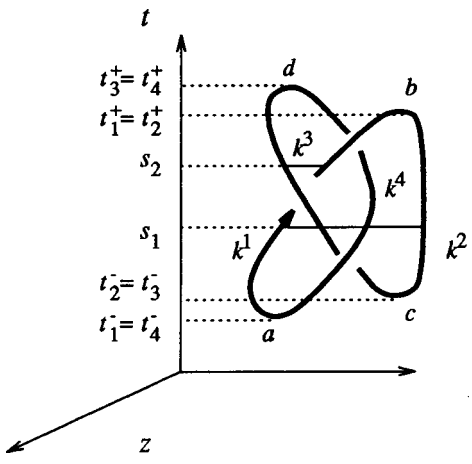

FIG. 13. Details on the labeling of one of the contributions.

$$(ig)^2 \frac{1}{(2\pi i)^2} \int_{t_2^- < s_1 < s_2 < t_1^+} ds_1 ds_2 \frac{\dot{z}_1(s_1) - \dot{z}_2(s_1)}{z_1(s_1) - \dot{z}_2(s_1)} \frac{\dot{z}_3(s_2) - \dot{z}_1(s_2)}{z_3(s_2) - z_1(s_2)}. \tag{4.17}$$

One of the two integrations can easily be performed, leading to

$$(ig)^2 \frac{1}{(2\pi i)^2} \int_{t_2^- < s_2 < t_1^+} ds_2 \log\left(\frac{z_1(s_2) - z_2(s_2)}{z_1(t_2^-) - z_2(t_2^-)} \right) \frac{\dot{z}_3(s_2) - \dot{z}_1(s_2)}{z_3(s_2) - z_1(s_2)}. \tag{4.18}$$

Notice that, as argued before, this integral is finite. Although z_1 and z_2 get close to each other when $s_2 \to t_1^+$, the singularity in the integrand, being logarithmic, is too mild to lead to a divergent result. Expressions similar to (4.18) can easily be obtained for the rest of the contributions depicted in Fig. 12.

We are now in a position to write the form of the general contribution originated from the Feynman rules of the theory. Notice that the most significant fact of our previous discussion is the presence of the delta function in the height of the propagator (3.4). It implies that the only nonvanishing configurations of the propagators are those in which their two end points have the same height; in other words, only contributions in which the line representing the propagator is horizontal do not vanish. This observation allows us to rearrange the contributions to the perturbative series expansion in the following way. Consider all possible pairings $\{z_i(s), z_j'(s)\}$ of curves k^i and k_ε^j, $i,j = 1,...,2n$, where $2n$ is the number of extrema of the Morse knot under consideration. A contribution at order m in perturbation theory will involve a path-ordered integral in the heights $s_1 < \cdots < s_r < \cdots < s_m$ of a product of m propagators of the form (4.3) and (4.4):

$$\prod_{r=1}^{m} \frac{dz_{i_r}(s_r) - dz'_{j_r}(s_r)}{z_{i_r}(s_r) - z'_{j_r}(s_r)}. \tag{4.19}$$

This product is characterized by a set of m ordered pairings, each one labeled by a pair of numbers (i_r, j_r) with $r = 1,...,m$. We will denote an ordered pairing of m propagators generically by P_m. One must take into account all possible ordered pairings, i.e., one must sum over all the possible P_m. The group factor that corresponds to each ordered pairing P_m is simply obtained by placing the group generators at the end points of the propagators and taking the trace of the product, which results after traveling along the knot. The resulting group factor will be denoted by $R(P_m)$.

Another ingredient in (4.4) that must be taken into account is the factor p_{ij}. For each pairing $P_m = \{(i_r, j_r), r = 1,...,m\}$ there will be a contribution from their product. Certainly, the result will be a sign that will depend on the ordered pairing P_m. We will denote such a product by

$$s(P_m) = \prod_{r=1}^{m} p_{i_r j_r}. \tag{4.20}$$

We are now in a position to write the full expression for the contribution to the perturbative series expansion at order m. It takes the form

$$(ig^2)^m \left(\frac{1}{2\pi i} \right)^m \frac{1}{2^m} \sum_{P_m} \int_{t_{P_m}^- < t_1 < \cdots < t_r < \cdots < t_m < t_{P_m}^+} s(P_m) \prod_{r=1}^{m} \frac{dz_{i_r}(t_r) - dz'_{j_r}(t_r)}{z_{i_r}(t_r) - z'_{j_r}(t_r)} R(P_m), \tag{4.21}$$

where $t_{P_m}^+$ and $t_{P_m}^-$ are highest and lowest heights, which can be reached by the last and first propagators of a given ordered pairing P_m. This expression corresponds to the Kontsevich integral for framed knots as presented in Ref. 25.

V. CONCLUSIONS AND OPEN PROBLEMS

In this paper we have presented an analysis of Chern–Simons gauge theory in the light-cone gauge from a perturbative point of view. As it became clear in our discussion in Sec. III, only vacuum expectation values associated to Morse knots are suitable for calculation in the light-cone

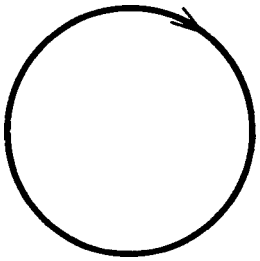

FIG. 14. Unknot with two critical points.

gauge, at least if no additional regularization to the one considered in the paper is introduced. We have shown that the regularization prescription leads to the consideration of framed knots, and that the vacuum expectation value possesses a dependence on the framing that agrees with the ones appearing in other gauges and in the nonperturbative analysis of the theory. We have also shown that the perturbative series expansion contains the Kontsevich integral for framed knots.

Our results, however, demonstrate that something has been missed in the perturbative series expansion. Indeed, contrary to what is obtained in other gauges or in nonperturbative approaches, according to our result (4.21), the vacuum expectation value corresponding to the unknot shown in Fig. 14 carrying a representation R of the gauge group is just dim R (in the trivial framing). In other words, all the contributions in (4.21) vanish and one is left with the zeroth-order contribution, which is just dim R. Actually, the expression (4.21) does not lead to knot invariants. It is well known[24] that the Kontsevich integral is only invariant under deformations of the knot which preserve the number of critical points. To obtain a truly invariant quantity one must take into account a correction to the sum of the contributions (4.21). If we denote by $Z_m(K,K_\varepsilon)$ the contribution shown in (4.21) (order m), the quantity that leads to a knot invariant is[24]

$$\frac{\dim R+\Sigma_{m=1}^{\infty} Z_m(K,K_\varepsilon)}{(1+(1/\dim R)\Sigma_{m=1}^{\infty} Z_m(U,U_\varepsilon))^{n/2}}, \tag{5.1}$$

where U and U_ε are the unknots shown in Fig. 15, and n is the number of critical points of the knot K. The coefficients of the powers of g in this expression are Vassiliev invariants.

The fact that the perturbative series expansion corresponding to the terms (4.21) is not invariant under all kinds of deformations of the knot is something that one could have expected. As discussed in Sec. III, the perturbative construction fails for knots that are not of the Morse type. In a deformation of the knot in which the number of critical points is changed, one has to consider at some intermediate step a knot that is not of the Morse type. The analysis presented here is not applicable to that situation and it is therefore consistent to find that the final expression for the perturbative expansion is not invariant under those types of deformations. A very interesting problem is to understand, from the point of view of Chern–Simons gauge theory, why the de-

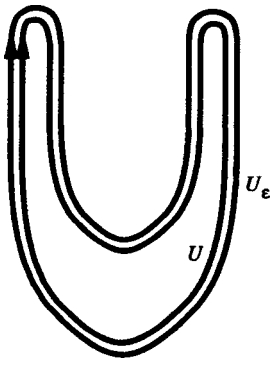

FIG. 15. Unknots U and U_ϵ with four critical points.

nominator of (5.1) has to be included there. Presumably, this would reconcile the results obtained here with the ones from covariant gauges and from nonperturbative approaches. This and other related issues are at present under investigation.

ACKNOWLEDGMENTS

We would like to thank M. Alvarez for helpful discussions. This work was supported in part by DGICYT under Grant No. PB93-0344.

[1] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
[2] M. Bos and V. P. Nair, Phys. Lett. B **223**, 61 (1989); Int. J. Mod. Phys. A **5**, 959 (1990); S. Elitzur, G. Moore, A. Schwimmer, and N. Seiberg, Nucl. Phys. **B326**, 108 (1989).
[3] J. M. F. Labastida and A. V. Ramallo, Phys. Lett. B **227**, 92 (1989); **228**, 214 (1989); J. M. F. Labastida, P. M. Llatas, and A. V. Ramallo, Nucl. Phys. **B348**, 651 (1991); J. M. F. Labastida and M. Mariño, Int. J. Mod. Phys. A **10**, 1045 (1995); J. M. F. Labastida and E. Pérez, J. Math. Phys. **37**, 2013 (1996).
[4] J. Frohlich and C. King, Commun. Math. Phys. **126**, 167 (1989).
[5] S. Martin, Nucl. Phys. **B338**, 244 (1990).
[6] R. K. Kaul, Commun. Math. Phys. **162**, 289 (1994); R. K. Kaul and T. R. Govindarajan, Nucl. Phys. **B380**, 293 (1992); **B393**, 392 (1993); P. Ramadevi, T. R. Govindarajan and R. K. Kaul, Mod. Phys. Lett. A **10**, 1635 (1995).
[7] V. F. R. Jones, Bull. Am. Math. Soc. **12**, 103 (1985); Ann. Math. **126**, 335 (1987).
[8] P. Freyd, D. Yetter, J. Hoste, W. B. R. Lickorish, K. Millet, and A. Ocneanu, Bull. Am. Math. Soc. **12**, 239 (1985).
[9] L. H. Kauffman, Trans. Am. Math. Soc. **318**, 417 (1990).
[10] Y. Akutsu and M. Wadati, J. Phys. Soc. Jpn. **56**, 839 (1987); **56**, 3039 (1987).
[11] E. Guadagnini, M. Martellini, and M. Mintchev, Phys. Lett. B **227**, 111 (1989); **228**, 489 (1989); Nucl. Phys. **B330**, 575 (1990).
[12] D. Bar-Natan, ''Perturbative aspects of Chern-Simons topological quantum field theory,'' Ph.D. thesis, Princeton University, 1991.
[13] M. Alvarez and J. M. F. Labastida, Nucl. Phys. **B395**, 198 (1993); **B433**, 555 (1995); **B441**, 403 (1995).
[14] D. Altschuler and L. Freidel, Commun. Math. Phys. **187**, 261 (1997).
[15] L. Alvarez-Gaumé, J. M. F. Labastida, and A. V. Ramallo, Nucl. Phys. **B334**, 103 (1990); G. P. Korchemsky, Mod. Phys. Lett. A **6**, 727 (1991).
[16] A. Blasi, O. Piguet, and S. P. Sorella, Nucl. Phys. **B356**, 154 (1991).
[17] C. P. Martin, Phys. Lett. B **241**, 513 (1990); G. Giavarini, C. P. Martin, and F. Ruiz Ruiz, Nucl. Phys. **B381**, 222 (1992).
[18] M. Alvarez and J. M. F. Labastida, J. Knot Theory Ramifications **5**, 779 (1996); M. Alvarez, J. M. F. Labastida, and E. Pérez, Nucl. Phys. **B488**, 677 (1997).
[19] M. Alvarez and J. M. F. Labastida, ''Primitive Vassiliev invariants and factorization in Chern-Simons gauge theory,'' Commun. Math. Phys. **189**, 641 (1997).
[20] R. Bott and C. Taubes, J. Math. Phys. **35**, 5247 (1994).
[21] G. Leibbrandt and C. P. Martin, Nucl. Phys. **B377**, 593 (1992); **B416**, 351 (1994).
[22] A. Brandhuber, S. Emery, M. Langer, O. Piguet, M. Schweda, and S. P. Sorella, Helv. Phys. Acta **66**, 551 (1993).
[23] J. F. W. H. van de Wetering, Nucl. Phys. **B379**, 172 (1992).
[24] M. Kontsevich, Adv. Sov. Math. **16**, 137 (1993).
[25] T. Q. T. Let and J. Murakami, ''The universal Vassiliev-Kontsevich invariant for framed oriented links,'' Max-Plank Institut fur Mathematik preprint, 1994.
[26] V. G. Knizhnik and A. B. Zamolodchikov, Nucl. Phys. **B247**, 83 (1984).
[27] A. S. Cattaneo, P. Cotta-Ramusino, J. Frohlich, and M. Martellini, J. Math. Phys. **36**, 6137 (1995).
[28] J. S. Birman and X. S. Lin, Invent. Math. **111**, 225 (1993); J. S. Birman, Bull. Am. Math. Soc. **28**, 253 (1993).
[29] D. Bar-Natan, Topology **34**, 423 (1995).

A Lie algebra for closed strings, spin chains, and gauge theories

C.-W. H. Lee[a] and S. G. Rajeev[b]
Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627

(Received 23 January 1998; accepted for publication 21 May 1998)

We consider quantum dynamical systems whose degrees of freedom are described by $N\times N$ matrices, in the planar limit $N\to\infty$. Examples are gauge theories and the M(atrix)-theory of strings. States invariant under U(N) are ''closed strings,'' modeled by traces of products of matrices. We have discovered that the U(N)-invariant operators acting on both open and closed string states form a remarkable new Lie algebra which we will call the heterix algebra. (The simplest special case, with one degree of freedom, is an extension of the Virasoro algebra by the infinite-dimensional general linear algebra.) Furthermore, these operators acting on closed string states only form a quotient algebra of the heterix algebra. We will call this quotient algebra the cyclix algebra. We express the Hamiltonian of some gauge field theories (like those with adjoint matter fields and dimensionally reduced pure QCD models) as elements of this Lie algebra. Finally, we apply this cyclix algebra to establish an isomorphism between certain planar matrix models and quantum spin chain systems. Thus we obtain some matrix models solvable in the planar limit; e.g., matrix models associated with the Ising model, the XYZ model, models satisfying the Dolan–Grady condition and the chiral Potts model. Thus our cyclix Lie algebra describes the dynamical symmetries of quantum spin chain systems, large-N gauge field theories, and the M(atrix)-theory of strings. © *1998 American Institute of Physics.* [S0022-2488(98)01810-6]

I. NTRODUCTION

The string is emerging as a verstile concept in physics, second only to the notion of a particle in its usefulness. The instantaneous configuration of a string can be thought of roughly as a curve in space, with an energy proportional to its length. The curve may be closed, or open with some extra degrees of freedom stuck at its endpoints. The string is being intensely studied as the fundamental object in the quantum theory of gravity and perhaps even the unified theory of all forces of nature.

The modern notion of a relativistic string theory originated in attempts to understand hadron dynamics, in the 1960s and early 1970s. However, it is now established that quantum chromodynamics (QCD) is the fundamental theory of strong interactions: the hadrons are not elementary particles but instead are bound states of quarks and gluons. In spite of this, attempts to understand hadron dynamics in terms of strings have not altogether ceased. Many features of hadron dynamics which originally prompted people to construct the string theory still do not have a satisfactory explanation. For example, the squares of the masses of hadrons increase linearly with their angular momenta, a characteristic property of strings. Indeed, it seems very plausible that there is a version of string theory which is equivalent to QCD but that directly describes hadrons. (Such a theory has been constructed in two dimensions.[1]) Since the gluon degrees of freedom are represented by matrices, this suggests that at least this type of string can be thought of as made of more elementary entities described by products of matrices. Connections to noncommutative geometry are also natural, as suggested in Ref. 2.

Meanwhile, in recent years the strings that appear in theories of quantum gravity have also

[a]Electronic mail: cwhlee@pas.rochester.edu
[b]Electronic mail: rajeev@pas.rochester.edu

 © 1998 American Institute of Physics

come to be seen as made of more elemental objects (''string bits''[3,4]), again described by matrices.[5,6] Indeed the underlying mathematical structures have much in common with earlier work on QCD with some extra symmetries (such as supersymmetry) added. These M(atrix)-theories are thought to describe not only stringlike excitations but also M(embranes) and are collectively known as the M-theory. The underlying symmetries and geometrical significance of this theory still remain M(ysterious).

Both of the above applications of string theory involve fundamental degrees of freedom represented by $N\times N$ matrices. In both cases the limit $N\to\infty$ is important, as it simplifies many issues. There are, in fact, many inequivalent ways of letting N go to infinity, the simplest one is the so-called planar large-N limit. It is called as such because it was originally[7] constructed by summing Feynman diagrams which have planar topology. (Other limits such as the double scaling limit which gives the continuum string theory are of interest as well.) We view this planar limit as a linear approximation to the general large-N limit discussed in Ref. 8.

Thus, the understanding of the quantum matrix models is of fundamental importance to modern theoretical physics. The two great challenges of theoretical physics — that of understanding hadron dynamics and that of developing a quantum theory of gravity — are both intertwined with the dynamics of quantum matrix models.

We have discovered the fundamental dynamical symmetry of matrix models in the large-N limit: it is a remarkable new Lie algebra which we will call the heterix algebra. It is thus the algebra of symmetries that will help us solve both of the two great problems of theoretical physics. In addition, we will find a connection to a third branch of theoretical physics: we can understand the solvability of quantum spin chain models within the formalism of the heterix algebra.

The outline of this paper is as follows. In Sec. II, we will introduce heterix algebra axiomatically. We will also discuss aspects of its structure such as its Cartan subalgebra and root vectors. In particular, we will show that this Lie algebra is an extension of another Lie algebra by $gl_{+\infty}$.[9] In Sec. III, we will study the simplest heterix algebra in which there is only one degree of freedom. It will turn out that the noncommutative algebra is simplified to the Virasoro algebra. Thus our algebra can be viewed as a generalization of the Virasoro algebra. In Sec. IV, we will show that gauge-invariant observables in the planar large-N limit can be formulated in terms of the elements of this algebra. Moreover, closed string bit states in this large-N limit provide a representation for this Lie algebra. However, this representation is not faithful and thus there are many relations among some elements of this algebra. We will quotient out these relations and get another Lie algebra — the cyclix algebra.[10]

In Sec. V, we will discuss the applications of the cyclix algebra to various gauge field theories, which, in the planar large-N limit, are multi-matrix models. In Sec. VI, we will construct more tractable multi-matrix models with the aid of quantum spin chain models. In addition, we will find that we can understand the solvability of quantum spin chain models within the formalism of the cyclix algebra.

II. THE HETERIX LIE ALGEBRA

In this section we will give a self-contained definition of a certain Lie algebra, which we will call the heterix algebra. In a latter section we will see that closed string operators provide a representation for this algebra, albeit an unfaithful one. The defining representation in this section is on the combined state space of open and closed strings and is, of course, faithful.

Define a complex vector space $\mathcal{T}_{op}$ of *finite* linear combinations of orthogonal states of an open string: $s^L=s^{l_1l_2\cdots l_d}$. (See Appendix A for a more detailed explanation of the notation.) This is just the space of all tensors (with no particular symmetry properties) on the complex vector space C^Λ:

$$\mathcal{T}_{op}=\oplus_{r=1}^{\infty}\otimes^r C^\Lambda. \tag{1}$$

This vector space should be thought of as the space of states of an open string propagating in a discrete model of space–time with Λ points. In addition define $\mathcal{T}_c$ of finite linear combinations of orthogonal states $\Psi^{(K)}=\Psi^{(k_1\cdots k_c)}$ labelled by cyclically symmetric indices. Thus $\mathcal{T}_c$ is the space of *cyclically symmetric tensors* on C^Λ and should be viewed as the space of states of a closed string propagating in discretized space. Our Lie algebra will be defined by the commutators of some

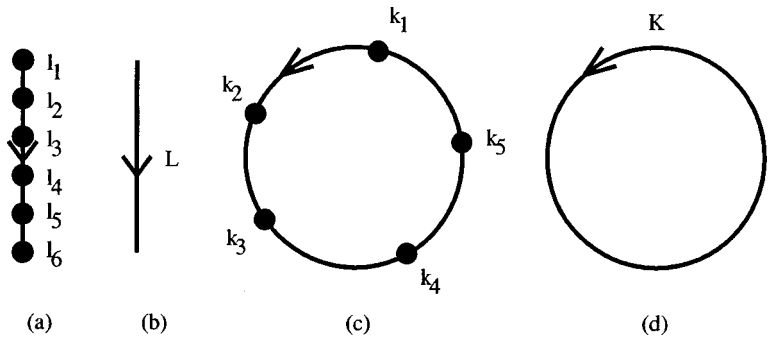

FIG. 1. (a) A typical open string state s^L in detail. There is a series of 6 indices l_1, l_2, ..., and l_6. Each index is represented by a solid bead. These solid beads are connected by thick lines (in contrast with the thin lines in latter figures). The arrow indicates the direction of the integer sequence L. (b) An open string state in brief. The beads are omitted. They will be consistently omitted in all ensuing brief figures. The integer sequence is represented by the letter L. The length of the line does not indicate the number of indices the line carries. (c) A typical closed string state $\Psi^{(K)}$ with five indices in detail. There is a series of indices k_1, k_2, ..., and k_5. Notice the cyclic symmetry of this state. (d) A closed string state in brief. The state is represented by the equivalence class (K), but in the figure we simply label the state as K. The size of the circle does not indicate the number of indices the circle carries.

operators on the total space of closed and open strings $\mathcal{T}=\mathcal{T}_c\oplus\mathcal{T}_{op}$. Thus $\mathcal{T}$ has an orthogonal basis $\{\Psi^{(K)},s^L\}$ as (K) ranges over (nonempty) equivalence classes of cyclically symmetric sequences and L over all nonempty sequences. Typical open and closed string states are illustrated in Fig. 1. [The reader can compare Figs. 1(a) and 1(b) with Figs. 1(a) and 1(b) in Ref. 11.]

Now let us introduce some operators that act on these states. Define the operator $\tilde{f}^{(I)}_{(J)}$ by its action on $\mathcal{T}$ as follows:

$$\tilde{f}^{(I)}_{(J)}\Psi^{(K)}=\delta^K_{(J)}\Psi^{(I)} \quad \text{and} \quad \tilde{f}^{(I)}_{(J)}s^L=0. \tag{2}$$

This operator just converts the closed string state $\Psi^{(K)}$ to $\Psi^{(I)}$, with a multiplicative factor which is the number of cyclic permutations of K that is equal to J. It gives zero if there are no such permutations or if it acts on an open string state. Clearly $\tilde{f}^{(I)}_{(J)}$ is a finite rank operator. The operator $\tilde{f}^{(I)}_{(J)}$ is shown pictorially in Fig. 2. Also shown in the same figure is the action of $\tilde{f}^{(I)}_{(J)}$ on $\Psi^{(K)}$.

In addition let us define the operator γ^I_J by its actions on closed strings,

$$\gamma^I_J\Psi^{(K)}=\delta^K_{(J)}\Psi^{(I)}+\sum_{K_1K_2=(K)}\delta^{K_1}_J\Psi^{(IK_2)}, \tag{3}$$

and on open strings as shown:

$$\gamma^I_Js^L=\delta^L_Js^I+\sum_{L_1L_2=L}\delta^{L_1}_Js^{IL_2}+\sum_{L_1L_2=L}\delta^{L_2}_Js^{L_1I}+\sum_{L_1L_2L_3=L}\delta^{L_2}_Js^{L_1IL_3}. \tag{4}$$

Clearly, it is the direct sum of an operator g^I_J acting on closed strings and σ^I_J acting on open strings: $\gamma^I_J=g^I_J\oplus\sigma^I_J$. Here σ^I_J has been defined in a previous paper.[11] These operators are not finite rank, unlike the $\tilde{f}^{(I)}_{(J)}$; indeed they are not even bounded.

Let us describe Eq. (3) in words. Regard (K) as a circle of points labelled by positive integers. If J has more indices than K, then the action of γ^I_J on $\Psi^{(K)}$ gives zero. When J is shorter than K, each segment of K that is identical to J is replaced with I and then a sum over all the resultant terms is taken. This is done even if some of the segments agreeing with J overlap partially with each other. If J does not overlap with any segment of K, then we get zero.

Next, consider the case when the length of J is the same as that of K, i.e., $\#(J)=\#(K)=c$. Then we get zero if no cyclic permutation of J agrees with K; otherwise, we replace K with I and multiply $\Phi^{(I)}$ by the number of cyclic permutatations of K that agree with J.

Equation (3) is illustrated in Fig. 3. Using the identities in Appendix A, Eq. (3) can be rewritten as

$$\gamma_J^I \Psi^{(K)} = \delta_{(J)}^K \Psi^{(I)} + \sum_A \delta_{(JA)}^K \Psi^{(IA)}. \tag{5}$$

This form is more convenient for some calculations.

We can understand Eq. (4) analogously. Again, if J is longer than K, we get zero. Otherwise, each segment of K that agrees with J is replaced with I and a sum over all such resultants is taken. The first term represents the case when J is identical to K; the second when J agrees with the beginning of K; the third when J agrees with the end of K; and the last term, which is the ''generic'' one, describes the case when J agrees with some segment in the middle of K. Of course we get zero if there is no segment of K that agrees with J. Equation (4) is depicted in Fig. 4.

The reader can notice from the figures an analogy between the action of our operators on states and the action of some viruses on DNA molecules.

Altogether, $\{\tilde{f}_{(J)}^{(I)}, \gamma_J^I\}$ is a linearly independent set of operators. This fact, the proof of which can be found in Appendix B, is of some importance, as we will define a Lie algebra with this set as a basis soon. The operators g_J^I acting on closed string states alone do not form a linearly independent set, however. We will derive some relations among them in a later section.

The product of two operators $\tilde{f}_{(J)}^{(I)}, \tilde{f}_{(L)}^{(K)}$ is a finite linear combination of such operators. Moreover, the $\tilde{f}_{(J)}^{(I)}$'s are linearly independent so that they span an associative algebra. Their commuta-

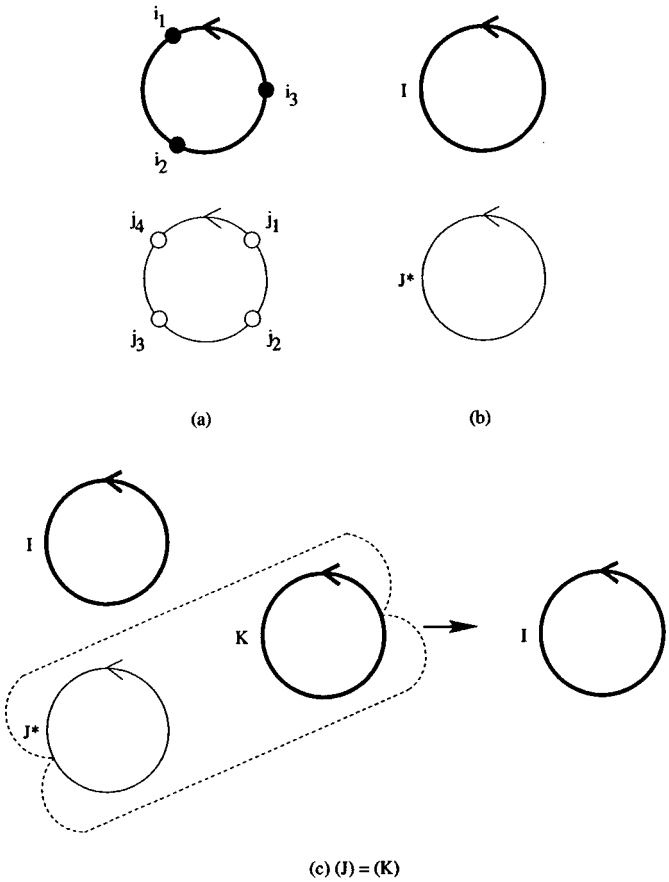

FIG. 2. (a) A detailed diagrammatic representation of a typical $\tilde{f}_{(J)}^{(I)}$. The upper circle represents (I), whereas the lower circle represents (J). In this example, there are three indices in I and four indices in J. An index in I is represented by a solid bead, whereas an index in J is represented by a hollow bead. Moreover, the solid beads are joined by thick lines, while the hollow ones are joined by thin lines. Note the sequence J is put in the clockwise instead of the counter-clockwise direction. (b) $\tilde{f}_{(J)}^{(I)}$ in brief. An asterisk is added to J to signify the fact that J is in the counter-clockwise direction. (c) The action of $\tilde{f}_{(J)}^{(I)}$ on $\Psi^{(K)}$. The dotted lines connect the line segments to be ''annihilated'' together. The figure on the right of the arrow is the resultant closed string state.

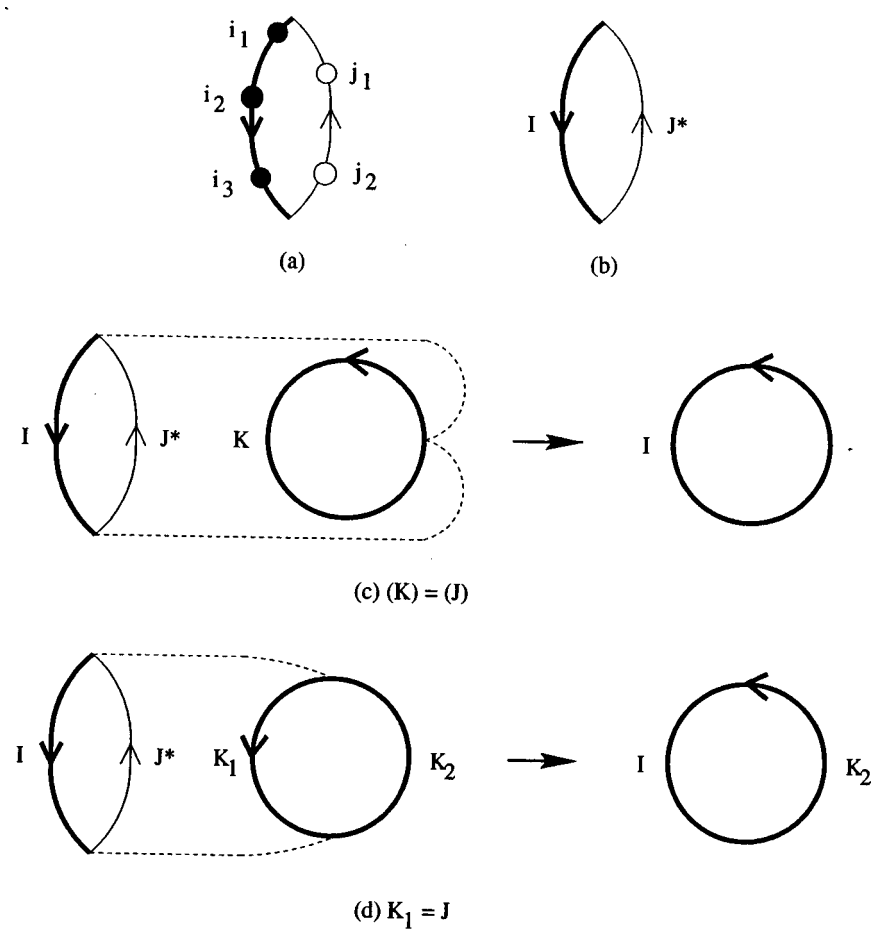

FIG. 3. (a) A typical γ_J^I. There are three indices in I and two indices in J. Note that J is in reverse. (b) A concise representation of γ_J^I. Again an asterisk is added to J to signify the fact that J is put in reverse. (c),(d) The action of γ_J^I on $\Psi^{(K)}$. Only the first two terms of Eq. (3) are shown.

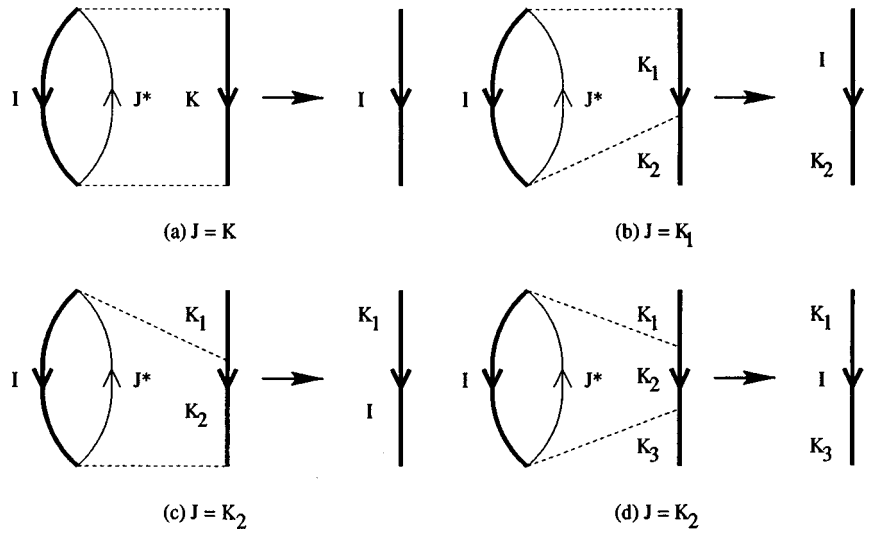

FIG. 4. (a)–(d) The action of γ_J^I on s^L (c.f. Fig 8 in Ref. 11). Only the first four terms are shown.

tors span a Lie algebra which we will also call $\tilde{F}'_\Lambda$. However, the product $\gamma^I_J \gamma^K_L$ cannot be written, in general, as a finite linear combination of the γ's and $\tilde{f}$'s. A proof of this statement can be found in Appendix C.

Nevertheless, it is a remarkable fact that the commutators of any two stringlike operators defined above can, in fact, be written as finite linear combinations of themselves. In fact we see, by a straightforward (but tedious) computation, that the commutator between two γ's reads

$$
\begin{aligned}
[\gamma^I_J, \gamma^K_L] = {} & \delta^K_J \gamma^I_L + \sum_{J_1 J_2 = J} \delta^K_{J_2} \gamma^I_{J_1 L} + \sum_{K_1 K_2 = K} \delta^{K_1}_J \gamma^{IK_2}_L + \sum_{\substack{J_1 J_2 = J \\ K_1 K_2 = K}} \delta^{K_1}_{J_2} \gamma^{IK_2}_{J_1 L} + \sum_{J_1 J_2 = J} \delta^K_{J_1} \gamma^I_{L J_2} \\
& + \sum_{K_1 K_2 = J} \delta^{K_2}_J \gamma^{K_1 I}_L + \sum_{\substack{J_1 J_2 = J \\ K_1 K_2 = K}} \delta^{K_2}_{J_1} \gamma^{K_1 I}_{L J_2} + \sum_{J_1 J_2 J_3 = J} \delta^K_{J_2} \gamma^I_{J_1 L J_3} + \sum_{K_1 K_2 K_3 = K} \delta^{K_2}_J \gamma^{K_1 I K_3}_L \\
& + \sum_{\substack{J_1 J_2 = J \\ K_1 K_2 = K}} \delta^{K_1}_{J_2} \delta^{K_2}_{J_1} \tilde{f}^{(I)}_{(L)} + \sum_{\substack{J_1 J_2 J_3 = J \\ K_1 K_2 = K}} \delta^{K_1}_{J_3} \delta^{K_2}_{J_1} \tilde{f}^{(I)}_{(J_2 L)} + \sum_{\substack{J_1 J_2 = J \\ K_1 K_2 K_3 = K}} \delta^{K_1}_{J_2} \delta^{K_3}_{J_1} \tilde{f}^{(I K_2)}_{(L)} \\
& + \sum_{\substack{J_1 J_2 J_3 = J \\ K_1 K_2 K_3 = K}} \delta^{K_1}_{J_3} \delta^{K_3}_{J_1} \tilde{f}^{(I K_2)}_{(J_2 L)} - (I \leftrightarrow K, J \leftrightarrow L). \qquad (6)
\end{aligned}
$$

Moreover, the commutators between a γ and an $\tilde{f}$, and between two $\tilde{f}$'s, are

$$
[\gamma^I_J, \tilde{f}^{(K)}_{(L)}] = \delta^K_{(J)} \tilde{f}^{(I)}_{(L)} + \sum_{K_1 K_2 = (K)} \delta^{K_1}_J \tilde{f}^{(I K_2)}_{(L)} - \delta^I_{(L)} \tilde{f}^{(K)}_{(J)} - \sum_{L_1 L_2 = (L)} \delta^I_{L_2} \tilde{f}^{(K)}_{(L_1 J)} \qquad (7)
$$

and

$$
[\tilde{f}^{(I)}_{(J)}, \tilde{f}^{(K)}_{(L)}] = \delta^K_{(J)} \tilde{f}^{(I)}_{(L)} - \delta^I_{(L)} \tilde{f}^{(K)}_{(J)}, \qquad (8)
$$

respectively. The proof of the above three equations will be relegated to Appendix D. Some terms in Eq. (6) are depicted in Figs. 5–7. Illustrated in Fig. 8 are the two Lie algebraic relations Eqs. (7) and (8). Let us call the Lie algebra with the basis $\{\tilde{f}^{(I)}_{(J)}, \gamma^I_J\}$ and the above commutation relations the *heterix algebra*, $\hat{\Gamma}_\Lambda$.

Let us explore the structure of the heterix algebra. First of all, we see from Eq. (7) that $\tilde{F}'_\Lambda$ is a proper ideal of the cyclix algebra. Next, consider the subspace $\mathcal{H}$ spanned by all vectors of the forms γ^I_I and $\tilde{f}^{(I)}_{(I)}$, where I is an arbitrary finite integer sequence of integers between 1 and Λ inclusive. It turns out that $\mathcal{H}$ is a Cartan subalgebra of the heterix algebra. (We define, following Ref. 12, a Cartan subalgebra to be a nilpotent subalgebra which is its own normalizer.) (The proof can be found in Appendix E.) In addition, we have the following special cases of Eq. (7) and (8):

$$
[\gamma^I_I, \tilde{f}^{(K)}_{(L)}] = \left(\delta^K_{(I)} + \sum_{K_1 K_2 = (K)} \delta^{K_1}_I - \delta^I_{(L)} - \sum_{L_1 L_2 = (L)} \delta^I_{L_1} \right) \tilde{f}^{(K)}_{(L)} \qquad (9)
$$

and

$$
[\tilde{f}^{(I)}_{(I)}, \tilde{f}^{(K)}_{(L)}] = (\delta^K_{(I)} - \delta^I_{(L)}) \tilde{f}^{(K)}_{(L)}. \qquad (10)
$$

Therefore, every $\tilde{f}^{(K)}_{(L)}$ is a root vector of the cyclix algebra with respect to the subalgebra $\mathcal{H}$.

Suppose we define some operators as follows:

$$
l^I_J \equiv \gamma^I_J - \sum_{i=1}^{\Lambda} \gamma^{iI}_{iJ} - \tilde{f}^{(I)}_{(J)} \quad \text{and} \quad r^I_J \equiv \gamma^I_J - \sum_{j=1}^{\Lambda} \gamma^{Ij}_{Jj} - \tilde{f}^{(I)}_{(J)}. \qquad (11)
$$

The actions of these operators on $\mathcal{T}$ are

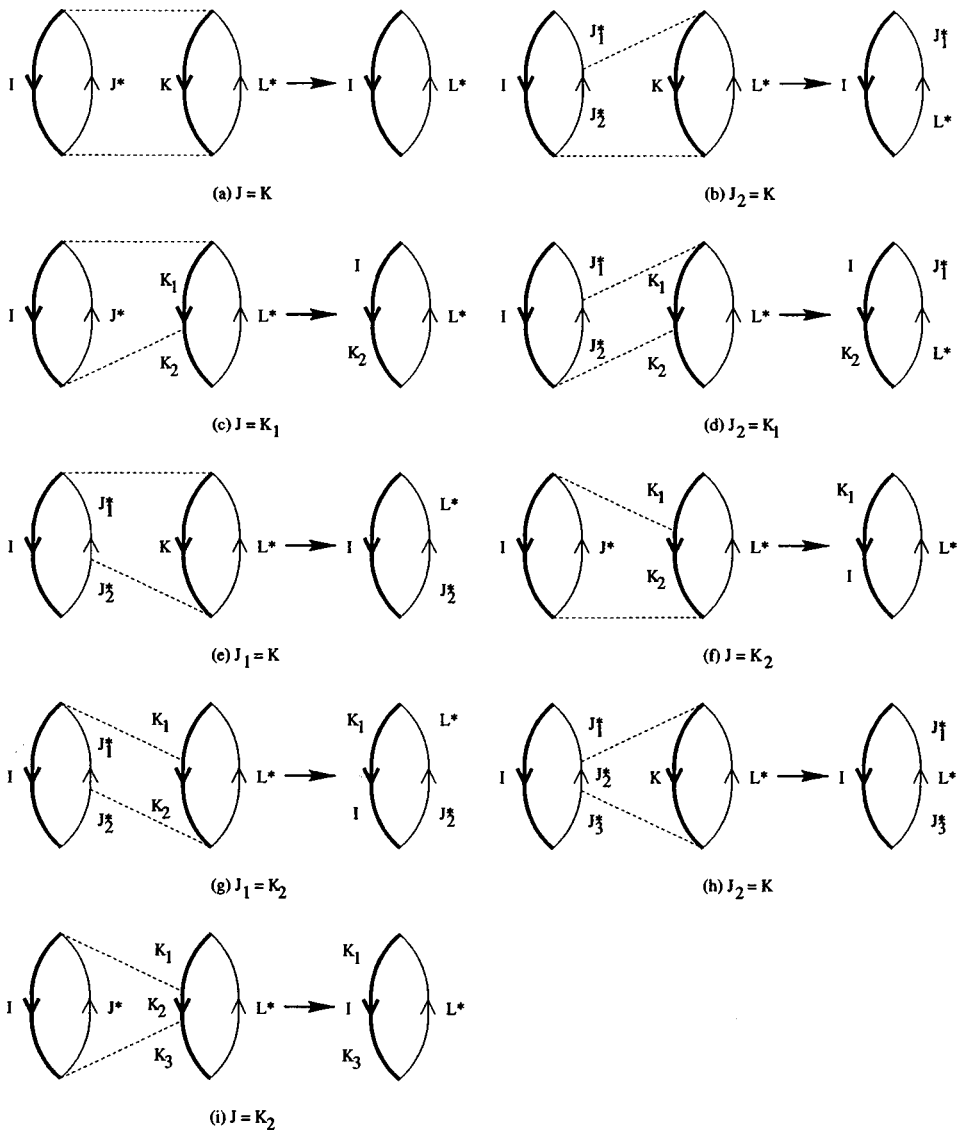

FIG. 5. Diagrammatic representations of the first nine terms on the rhs of Eq. (6).

$$l_J^I s^K = \delta_J^K s^I + \sum_{K_1 K_2 = K} \delta_J^{K_1} s^{I K_2}, \quad l_J^I \Psi^{(K)} = 0, \tag{12}$$

$$r_J^I s^K = \delta_J^K s^I + \sum_{K_1 K_2 = K} \delta_J^{K_2} s^{K_1 I}, \quad r_J^I \Psi^{(K)} = 0.$$

Hence they act only on the open string states: l_J^I and r_J^I are precisely the operators defined in a previous paper[11] on open strings. From the results of that paper, we deduce that the algebra consisting of finite linear combinations of l_J^I's, $\hat{L}'_\Lambda$, is a subalgebra of the heterix algebra. Similarly, the algebra consisting of finite linear combinations of r_J^I's, $\hat{R}'_\Lambda$, is also a subalgebra of the cyclix algebra.

Now define

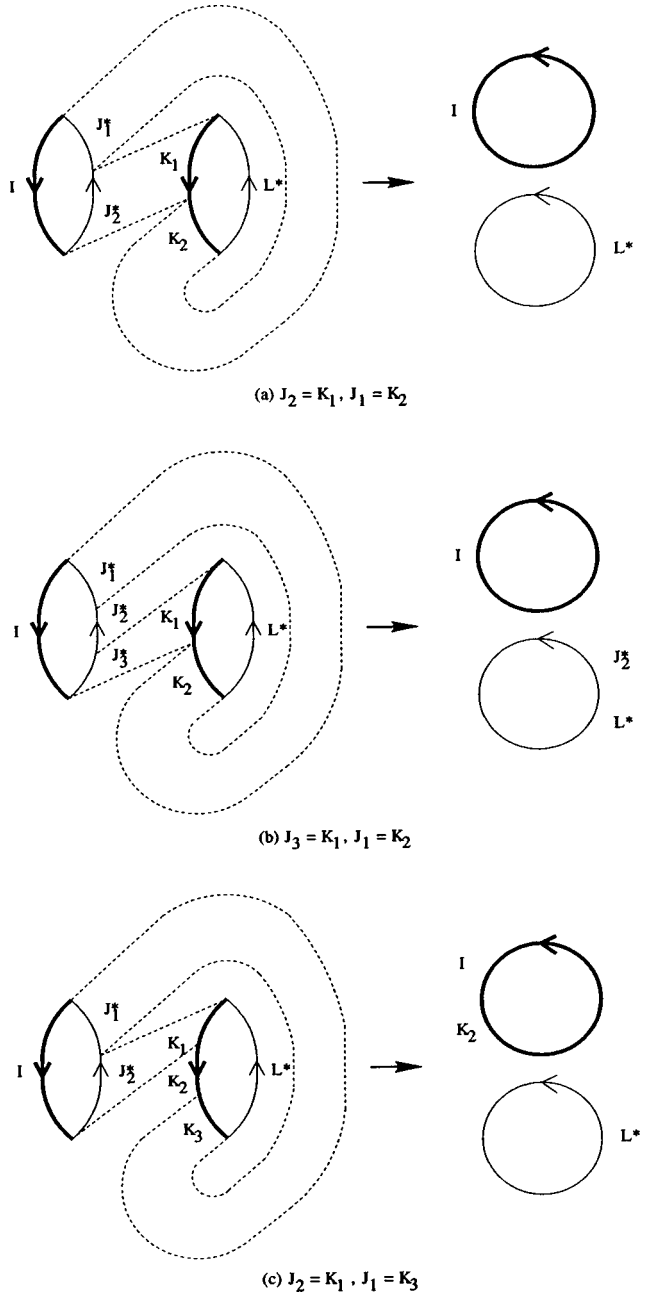

FIG. 6. Diagrammatic representations of the tenth to twelfth terms on the rhs of Eq. (6).

$$f^I_J \equiv \gamma^I_J - \sum_{i=1}^{\Lambda} \gamma^{iI}_{iJ} - \sum_{j=1}^{\Lambda} \gamma^{Ij}_{Jj} + \sum_{i,j=1}^{\Lambda} \gamma^{iIj}_{iJj} - \tilde{f}^{(I)}_{(J)} + \sum_{i=1}^{\Lambda} \tilde{f}^{(iI)}_{(iJ)}$$
$$= l^I_J - \sum_{j=1}^{\Lambda} l^{Ij}_{Jj} = r^I_J - \sum_{i=1}^{\Lambda} r^{iI}_{iJ}. \tag{13}$$

Its action on $\mathcal{T}$ reads

$$f^I_J s^K = \delta^K_J s^I \quad \text{and} \quad f^I_J \Psi^{(K)} = 0. \tag{14}$$

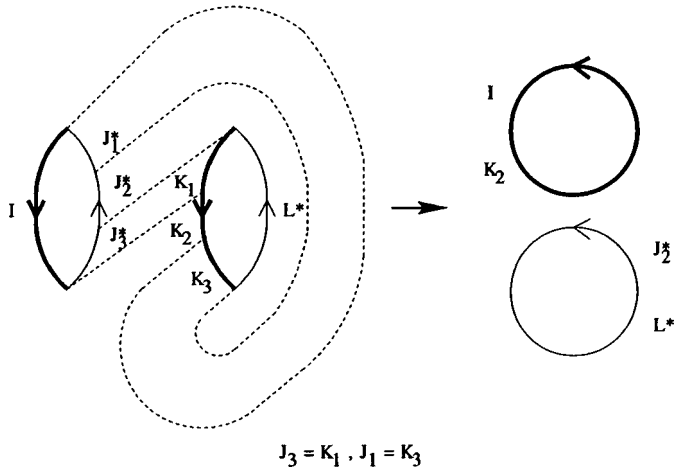

FIG. 7. Diagrammatic representations of the thirteenth term on the rhs of Eq. (6).

Hence this f^I_J also acts only on open strings; it too was introduced in Ref. 11. The algebra consisting of finite linear combinations of f^I_J's, F'_Λ, is a subalgebra of $\hat{L}'_\Lambda$, $\hat{R}'_\Lambda$ and the heterix algebra. We deduce from Eqs. (2), (3), (4) and (14) that

$$\begin{aligned}[\gamma^I_J, f^K_L] &= \delta^K_J f^I_L + \sum_{K_1K_2=K} \delta^{K_1}_J f^{IK_2}_L + \sum_{K_1K_2=K} \delta^{K_2}_J f^{K_1I}_L + \sum_{K_1K_2K_3=K} \delta^{K_2}_J f^{K_1IK_3}_L - \delta^I_L f^K_J \\ &\quad - \sum_{L_1L_2=L} \delta^I_{L_1} f^K_{JL_2} - \sum_{L_1L_2=L} \delta^I_{L_2} f^K_{L_1J} - \sum_{L_1L_2L_3=L} \delta^I_{L_2} f^K_{L_1JL_3}, \\ [\tilde{f}^{(I)}_{(J)}, f^K_L] &= 0. \end{aligned} \tag{15}$$

In particular,

$$\begin{aligned}[\gamma^I_I, f^K_L] &= \Bigg(\delta^K_I + \sum_{K_1K_2=K} \delta^{K_1}_I + \sum_{K_1K_2=K} \delta^{K_2}_I + \sum_{K_1K_2K_3=K} \delta^{K_2}_I \\ &\quad - \delta^I_L - \sum_{L_1L_2=L} \delta^I_{L_1} - \sum_{L_1L_2=L} \delta^I_{L_2} - \sum_{L_1L_2L_3=L} \delta^I_{L_2} \Bigg) f^K_L, \\ [\tilde{f}^{(I)}_{(I)}, f^K_L] &= 0. \end{aligned} \tag{16}$$

Thus every f^K_L is also a root vector with respect to the Cartan subalgebra $\mathcal{H}$. In fact, a root vector with respect to $\mathcal{H}$ is either of the form $\tilde{f}^{(K)}_{(L)}$ or f^K_L. The proof of this assertion can be found in Appendix F. Hence every root space is one-dimensional.

The reader can refer to Ref. 11 for a more detailed discussion of the Lie algebraic relations among the f's, l's, r's and γ's. Also shown in detail in the same article are various diagrammatic representations of f's, l's, r's, and the Lie algebraic relations pertaining to them.

Now let us consider various quotient algebras from the heterix algebra. Since $\tilde{F}'_\Lambda$ is a proper ideal, we can make the set of all cosets $\gamma^I_J + \tilde{F}'_\Lambda$ into a quotient algebra. This quotient algebra is nothing but the centrix algebra $\hat{\Sigma}_\Lambda$ (Ref. 11) defined previously. The union of $\hat{L}'_\Lambda$ and $\hat{R}'_\Lambda$, the multix algebra $\hat{M}'_\Lambda$,[11] is another proper ideal, thus we can make the set of all cosets $\gamma^I_J + \hat{M}'_\Lambda$ and $\tilde{f}^{(I)}_{(J)} + \hat{M}'_\Lambda$ a quotient algebra as well. Note that every operator in $\hat{M}'_\Lambda$ acts on $\mathcal{T}_c$ to produce 0. We conjecture that $\mathcal{T}_c$ provides a faithful representation for the quotient algebra of all cosets $\gamma^I_J + \hat{M}'_\Lambda$ and $\tilde{f}^{(I)}_{(J)} + \hat{M}'_\Lambda$. As we will see in the next section, the corresponding $\Lambda = 1$ quotient algebra is precisely the Virasoro algebra without any central extension. Therefore the quotient algebra with $\Lambda > 1$ can be regarded as a generalization of the Virasoro algebra.

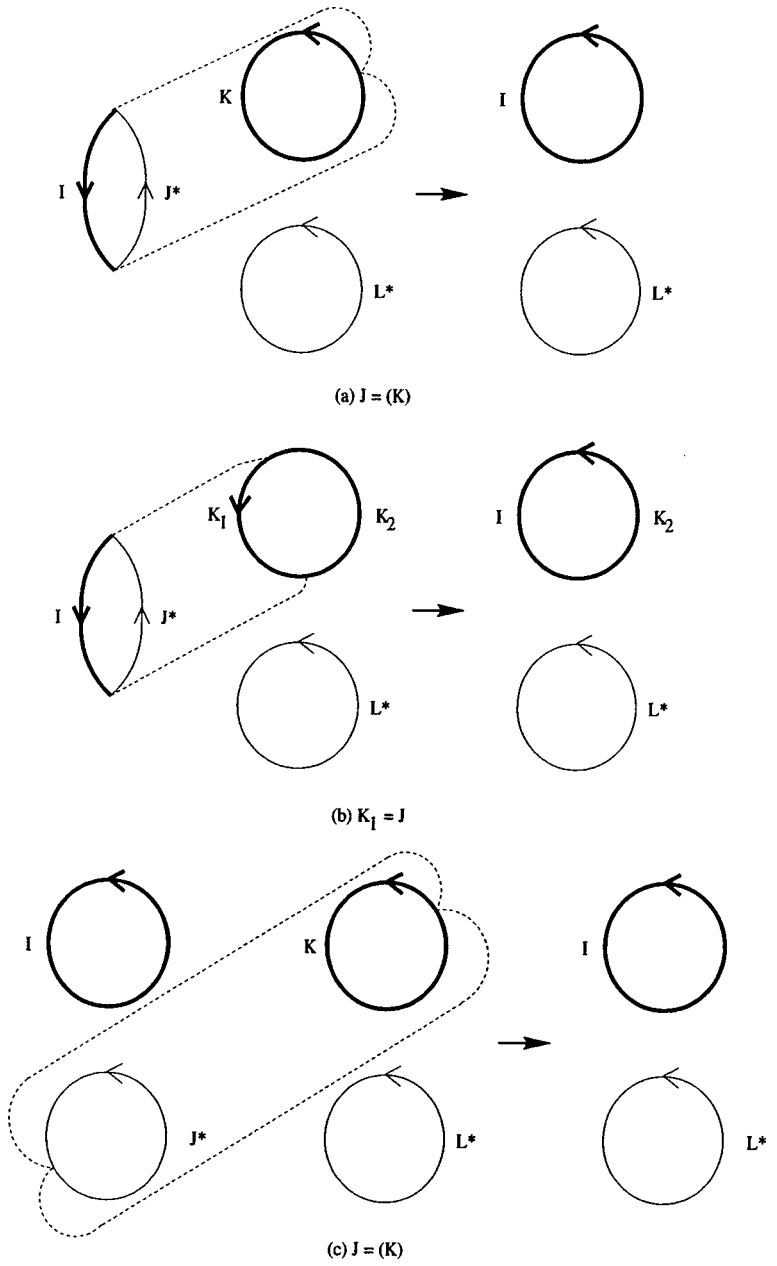

FIG. 8. Diagrammatic representations of the rhs of Eqs. (7) and (8). Only the first two terms and the first term are depicted in these two equations, respectively.

The Lie bracket of the subalgebra $\tilde{F}_\Lambda$ is given by Eq. (8). If we rescale each vector by defining

$$f^{(I)}_{(J)} \equiv \frac{\tilde{f}^{(I)}_{(J)}}{\sqrt{\delta^I_{(I)}\delta^J_{(J)}}}, \tag{17}$$

then the Lie bracket between two $f^{(I)}_{(J)}$'s is similar to that in Eq. (8), except that the delta function there with the generic form $\delta^I_{(J)}$ should be replaced with another delta function such that this new delta function yields simply 1 if $(I)=(J)$, and remains 0 otherwise. Then there is a one-to-one correspondence between each $f^{(I)}_{(J)}$ and each complex matrix $(a_{ij})_{i,j}\in Z_+$ such that all but one a_{ij} is nonzero, and that nonzero entry is 1. Thus the Lie algebras $\tilde{F}'_\Lambda$, for any Λ, are all isomorphic to $gl_{+\infty}$, the inductive limit of the general linear algebras. $gl_{+\infty}$ and the Kac–Moody algebra associated with it were thoroughly studied, and their properties can be found in, e.g., Kac's work.[9]

III. THE SPECIAL CASE $\Lambda=1$

Let us consider the cyclix algebra for the simplest special case $\Lambda=1$: this is useful as it will connect our Lie algebra with more familiar algebras such as the Virasoro algebra. Then all the sequences are repetitions of the number 1 some numbers of times. We can simplify the notations and write $\Psi^{(K)}$ as $\Psi^{(\#(K))}$, s^K as $s^{\#(K)}$, γ_J^I as $\gamma_{\#(J)}^{\#(I)}$, and $\tilde{f}_{(J)}^{(I)}$ as $\tilde{f}_{(\#(J))}^{(\#(I))}$. We can deduce from Eqs. (2), (3) and (4) that the actions of $\tilde{f}_{(b)}^{(a)}$ and γ_b^a, where a and b are the number of integers in the various sequences, on $\Psi^{(c)}$ and s^c, where c is also a positive integer, are given by

$$\tilde{f}_{(b)}^{(a)}\Psi^{(c)}=c\,\delta_b^c\Psi^{(a)}, \tag{18}$$

$$\tilde{f}_{(b)}^{(a)}s^c=0, \tag{19}$$

$$\gamma_b^a\Psi^{(c)}=c\,\theta(b\leqslant c)\Psi^{(a+c-b)}, \tag{20}$$

and

$$\gamma_b^a s^c=(c-b+1)\,\theta(b\leqslant c)s^{a+c-b}. \tag{21}$$

where θ(condition) is 1 if the condition holds, and 0 otherwise. The Lie brackets for the $\Lambda=1$ cyclix algebra are

$$\begin{aligned}[\gamma_b^a,\gamma_d^c]=&\,\theta(c\leqslant b)[2\gamma_{b+d-1}^{a+c-1}+2\gamma_{b+d-2}^{a+c-2}+\cdots+2\gamma_{b+d-c+1}^{a+1}\\&+(b-c+1)\gamma_{b+d-c}^a]+\theta(2\leqslant c\leqslant b)[\tilde{f}_{(b+d-2)}^{(a+c-2)}+2\tilde{f}_{(b+d-3)}^{(a+c-3)}\\&+\cdots+(c-1)\tilde{f}_{(b+d-c)}^{(a)}]+\theta(b<c)[2\gamma_{b+d-1}^{a+c-1}+2\gamma_{b+d-2}^{a+c-2}\\&+\cdots+2\gamma_{d+1}^{a+c-b+1}+(c-b+1)\gamma_d^{a+c-b}]+\theta(2\leqslant b<c)[\tilde{f}_{(b+d-2)}^{(a+c-2)}\\&+\tilde{f}_{(b+d-3)}^{(a+c-3)}+\cdots+(b-1)\tilde{f}_{(b)}^{(a+c-d)}]-(a\leftrightarrow c,b\leftrightarrow d),\end{aligned} \tag{22}$$

$$[\gamma_b^a,\tilde{f}_{(d)}^{(c)}]=c\,\theta(b\leqslant c)\tilde{f}_{(d)}^{(a+c-b)}-d\,\theta(a\leqslant d)\tilde{f}_{(b+d-a)}^{(c)} \tag{23}$$

and

$$[\tilde{f}_{(b)}^{(a)},\tilde{f}_{(d)}^{(c)}]=b\,\delta_b^c\tilde{f}_{(d)}^{(a)}-a\,\delta_d^a\tilde{f}_{(b)}^{(c)}. \tag{24}$$

These three equations can be deduced from Eqs. (6), (7) and (8), respectively.

The Cartan subalgebra $\mathcal{H}$ is now spanned by vectors of the forms γ_a^a and $\tilde{f}_{(a)}^{(a)}$. A root vector reads $\tilde{f}_{(d)}^{(c)}$ or $f_d^c=\gamma_d^c-2\gamma_{d+1}^{c+1}+\gamma_{d+2}^{c+2}-\tilde{f}_{(d)}^{(c)}+\tilde{f}_{(d+1)}^{(c+1)}$. The expression for f_d^c can be deduced from Eq. (13). The eigenequations are

$$[\gamma_a^a,\tilde{f}_{(d)}^{(c)}]=(c\,\theta(a\leqslant c)-d\,\theta(a\leqslant d))\tilde{f}_{(d)}^{(c)},\quad [\tilde{f}_{(a)}^{(a)},\tilde{f}_{(d)}^{(c)}]=a(\delta_a^c-\delta_d^a)\tilde{f}_{(d)}^{(c)}, \tag{25}$$

$$[\gamma_a^a,f_d^c]=(\theta(a\leqslant c)(c-a+1)-\theta(a\leqslant d)(d-a+1))f_d^c,$$

and

$$[\tilde{f}_{(a)}^{(a)},f_d^c]=0. \tag{26}$$

Let us consider the quotient algebra of all cosets $\gamma_b^a+\hat{M}_1'$ and $\tilde{f}_{(b)}^{(a)}+\hat{M}_1'$. *In the rest of this section and in the accompanying appendices, we will write* $\gamma_b^a+\hat{M}_1'$ *and* $\tilde{f}_{(b)}^{(a)}+\hat{M}_1'$ *simply as* g_b^a

and $\tilde{f}^{(a)}_{(b)}$, *respectively*. $\mathcal{T}_c$ still provides a representation for this quotient algebra — the actions of g^a_b and $\tilde{f}^{(a)}_{(b)}$ on $\mathcal{T}_c$ are still described by Eqs. (18) and (20). However, there is now a relation among $\tilde{f}$'s and g's:

$$\tilde{f}^{(a)}_{(b)} = g^a_b - g^{a+1}_{b+1}. \tag{27}$$

Hence,

$$g^a_b = \begin{cases} -\tilde{f}^{(a)}_{(b)} - \tilde{f}^{(a-1)}_{(b-1)} - \cdots - \tilde{f}^{(a-b+1)}_{(1)} + g^{a-b+1}_1 & \text{if } a \geqslant b, \\ -\tilde{f}^{(a)}_{(b)} - \tilde{f}^{(a-1)}_{(b-1)} - \cdots - \tilde{f}^{(1)}_{(b-a+1)} + g^1_{b-a+1}, & \text{if } a < b. \end{cases} \tag{28}$$

A simplification brought about by setting Λ to 1 is that we can write a basis for the quotient algebra easily. Indeed, the set of all $\tilde{f}^{(a)}_{(b)}$'s, g^1_b's and g^a_1's where a and b are arbitrary positive integers form a basis for the $\Lambda = 1$ quotient algebra. The proof of this statement will be given in Appendix G. The same proof also shows that $\mathcal{T}_c$ provides not only a representation, but also a *faithful* representation for this quotient algebra. From Eq. (23), we deduce that the subspace $\tilde{F}'_1$ spanned by all the vectors of the form $\tilde{f}^{(c)}_{(d)}$ forms a proper ideal of this quotient algebra.

The set of all $\tilde{f}^{(a)}_{(a)}$'s, where a is arbitrary, and g^1_1 form a basis for a Cartan subalgebra. The proof of this statement will be seen in Appendix H. In addition, since $\tilde{F}'_1$ is a proper ideal of the quotient algebra, the same proof reveals that any root vector with respect to this Cartan subalgebra must be a linear combination of a number of $\tilde{f}^{(a)}_{(b)}$'s. Therefore, all root vectors are of the form $\tilde{f}^{(c)}_{(d)}$'s with $c \neq d$. The corresponding eigenequations, which can be deduced from Eqs. (9) and (10), are

$$[g^1_1, \tilde{f}^{(c)}_{(d)}] = (c-d)\tilde{f}^{(c)}_{(d)} \tag{29}$$

and

$$[\tilde{f}^{(a)}_{(a)}, \tilde{f}^{(c)}_{(d)}] = a(\delta^c_a - \delta^a_d)\tilde{f}^{(c)}_{(d)}. \tag{30}$$

As $\tilde{F}'_1$ is a proper ideal of this $\Lambda = 1$ quotient algebra, we can form yet another quotient algebra of cosets of the form $v + \tilde{F}'_1$ where v is an arbitrary vector of the previous quotient algebra. This new quotient algebra is spanned by the cosets $g^a_1 + \tilde{F}_\Lambda$ and $g^1_b + \tilde{F}'_1$, where a and b run over all positive integers. It is a straightforward matter to show that the following Lie brackets are true:

$$[g^a_1 + \tilde{F}'_1, g^c_1 + \tilde{F}'_1] = (c-a)(g^{a+c-1}_1 + \tilde{F}'_1),$$

$$[g^a_1 + \tilde{F}'_1, g^1_d + \tilde{F}'_1] = \begin{cases} (2-a-d)(g^{a-d+1}_1 + \tilde{F}'_1), & \text{if } d \leqslant a, \\ (2-a-d)(g^1_{d-a+1} + \tilde{F}'_1), & \text{if } a \leqslant d, \end{cases} \tag{31}$$

$$[g^1_b + \tilde{F}'_1, g^1_d + \tilde{F}'_1] = (b-d)(g^1_{b+d-1} + \tilde{F}'_1).$$

Let us define

$$L_a = \begin{cases} -g^{a+1}_1 + \tilde{F}'_1, & \text{if } a \geqslant 0, \\ -g^1_{1-a} + \tilde{F}'_1, & \text{if } a \leqslant 0. \end{cases} \tag{32}$$

Note that the L here is *not* an integer sequence. Then Eq. (31) becomes

$$[L_a, L_b] = (a-b)L_{a+b}. \tag{33}$$

The reader can notice at once that this is the Lie algebraic relation of the Virasoro algebra without any central element.[13] Consequently, the $\Lambda=1$ quotient algebra of all cosets γ_b^a and $\tilde{f}_{(b)}^{(a)}$ can be regarded as an extension of the Virasoro algebra by an algebra isomorphic to $gl_{+\infty}$. The quotient algebras $\gamma_J^I+\hat{M}'_\Lambda+\tilde{F}_\Lambda$ for $\Lambda>1$ can then be regarded as generalizations of the Virasoro algebra, as stated in the previous section.

IV. CANONICAL REALIZATION

We have presented the operators acting on strings somewhat abstractly in the last section. Although the diagrammatic descriptions give the definitions of the operators a natural motivation, it would be better to have a more direct derivation of these relations in terms of more familiar operators such as those satisfying canonical commutation relations.

Let us define some bosonic operators which satisfy the canonical commutation relations:

$$[a_{\mu_2}^{\mu_1}(k_1),a^{\dagger\mu_3}_{\mu_4}(k_2)]=\delta_{k_1k_2}\delta^{\mu_1}_{\mu_4}\delta^{\mu_3}_{\mu_2},$$
$$[a_{\mu_2}^{\mu_1}(k_1),a_{\mu_4}^{\mu_3}(k_2)]=0,\quad [a^{\dagger\mu_1}_{\mu_2}(k_1),a^{\dagger\mu_3}_{\mu_4}(k_2)]=0. \tag{34}$$

These operators create and annihilate ''gluons,'' the most fundamental entities in our theory. The name ''gluons'' comes from the application of our theory to regularized QCD, in which case these operators act on gluons. The indices μ,ν and $\upsilon=1,2,\ldots,N$ label a quantum number we will call ''color.'' In this context, Λ is the possible number of distinct quantum states of a gluon excluding color. These quantum states are denoted by the numbers 1, 2, ..., Λ.

We identify the closed string states with a ring of gluons:

$$\Psi^{(K)}|0\rangle\equiv N^{-c/2}a^{\dagger\upsilon_2}_{\upsilon_1}(k_1)a^{\dagger\upsilon_3}_{\upsilon_2}(k_2)\cdots a^{\dagger\upsilon_1}_{\upsilon_c}(k_c)|0\rangle, \tag{35}$$

where $c=\#(K)$. (The summation convention for color indices is implicit in the above and following expressions.) A factor of $N^{-c/2}$ is inserted so that the state has a finite norm in the large-N limit. [Now the beads and connecting lines in Fig. 1(c) acquire some extra meanings — a solid bead represents a creation operator. Two beads connected by a line represents that the two corresponding creation operators share a common color index. This index is being summed over.]

We will now look at the operators which can map the above single glueball states to linear combinations of such states. In general they could also create states with several glueballs, but their amplitudes are suppressed by terms of order at least $1/N$: this fact is rather well known[3] and is a reflection in the Hamiltonian formalism of the planarity of Feynman diagrams in perturbation theory.[7] As an alternative to reading the relevant papers cited above, the reader can also consult and adapt an argument in a paper by one of us,[14] where an operator in the large-N limit is shown to propagate a single meson state to a linear combination of single meson states only.

There are two classes of such operators:

$$\tilde{f}^{(I)}_{(J)}\equiv N^{-(a+b)/2}a^{\dagger\mu_2}_{\mu_1}(i_1)a^{\dagger\mu_3}_{\mu_2}(i_2)\cdots a^{\dagger\mu_1}_{\mu_a}(i_a)a^{\nu_b}_{\nu_1}(j_b)a^{\nu_{b-1}}_{\nu_b}(j_{b-1})\cdots a^{\nu_1}_{\nu_2}(j_1) \tag{36}$$

and

$$\gamma_J^I\equiv N^{-(a+b-2)/2}a^{\dagger\mu_2}_{\mu_1}(i_1)a^{\dagger\mu_3}_{\mu_2}(i_2)\cdots a^{\dagger\nu_b}_{\mu_a}(i_a)a^{\nu_{b-1}}_{\nu_b}(j_b)a^{\nu_{b-2}}_{\nu_{b-1}}(j_{b-1})\cdots a^{\mu_1}_{\nu_1}(j_1). \tag{37}$$

These operators can act on a gluon segment of a glueball. (Now Figs. 2 and 3 can be regarded as illustrations of these operators. The solid beads represent creation operators, whereas the hollow beads represent annihilation operators. A connecting line, regardless of thickness, symbolizes the fact that the two corresponding operators share a common color index, and this index is being summed over.)

In the large-N limit, such operators propagate single glueball states to linear combinations of single glueball states:

$$\tilde{f}^{(I)}_{(J)}\Psi^{(K)}=\delta^{K}_{(J)}\Psi^{(I)}$$

and

$$\gamma^{I}_{J}\Psi^{(K)}=\delta^{K}_{(J)}\Psi^{(I)}+\sum_{K_1K_2=(K)}\delta^{K_1}_{J}\Psi^{(IK_2)}.$$

These are just the formulas [Eqs. (2) and (3)] we had in the previous section. Thus we have a representation of the Lie algebra of the last section on glueball states.

The representation we have obtained, however, is not faithful: we are only considering closed string states here. To emphasize this fact we will denote the operator acting on closed string states alone by g^{I}_{J}. (We would have a faithful representation if we included the open string states as well, but in this paper we will not consider them.) Indeed, using Eqs. (3) and (2), the reader can show that the following relations hold:

$$\begin{gathered}
\tilde{f}^{(I)}_{(J)}=g^{I}_{J}-\sum_{k=1}^{\Lambda}g^{Ik}_{Jk},\quad \tilde{f}^{(I)}_{(J)}=g^{I}_{J}-\sum_{k=1}^{\Lambda}g^{kI}_{kJ},\quad \tilde{f}^{(I)}_{(J)}=g^{I_2I_1}_{J}-\sum_{k=1}^{\Lambda}g^{I_2I_1k}_{Jk}\quad \text{if } I_1I_2=I,\\
\tilde{f}^{(I)}_{(J)}=g^{I_2I_1}_{J}-\sum_{k=1}^{\Lambda}g^{kI_2I_1}_{kJ}\quad \text{if } I_1I_2=I,\quad \tilde{f}^{(I)}_{(J)}=g^{I}_{J_2J_1}-\sum_{k=1}^{\Lambda}g^{Ik}_{J_2J_1k}\quad \text{if } J_1J_2=J,\\
\tilde{f}^{(I)}_{(J)}=g^{I}_{J_2J_1}-\sum_{k=1}^{\Lambda}g^{kI}_{kJ_2J_1}\quad \text{if } J_1J_2=J,\\
\tilde{f}^{(I)}_{(J)}=g^{I_2I_1}_{J_2J_1}-\sum_{k=1}^{\Lambda}g^{I_2I_1k}_{J_2J_1k}\quad \text{if } I_1I_2=I\quad \text{and } J_1J_2=J,\\
\tilde{f}^{(I)}_{(J)}=g^{I_2I_1}_{J_2J_1}-\sum_{k=1}^{\Lambda}g^{kI_2I_1}_{kJ_2J_1}\quad \text{if } I_1I_2=I\quad \text{and } J_1J_2=J.
\end{gathered}\tag{38}$$

These relations are generalizations of Eq. (27).

The quotient of the heterix algebra by the kernel of this representation was called the cyclix algebra[10] $\hat{C}_{\Lambda}$. As has been stated in a previous section, we conjecture that $\hat{C}_{\Lambda}=\hat{\Gamma}_{\Lambda}/\hat{M}'_{\Lambda}$. The results in Sec. III show that this conjecture is true for $\Lambda=1$.

V. APPLICATION TO GAUGE FIELD THEORY

We are going to apply the cyclix algebra to physical models which can be formulated as matrix models in the large-N limit. In this section, we will concentrate on gauge field theories in the large-N limit. We will find that we can express the physical observables of these theories in terms of the elements of the cyclix algebra. (Indeed, we will see in a future paper[15] that closed superstring models and M-theory can be formulated in terms of the supersymmetric generalization of the cyclix algebra in a similar manner.) However, they will look complicated. In the next section, we will construct multi-matrix models *integrable* in the large-N limit associated with integrable quantum spin chain models satisfying the periodic boundary condition. Studying the ways these multi-matrix models are solved can shed light on how the more complicated gauge field theories are solved.

In this section we will let the regulator $\Lambda\to\infty$, so that the momentum indices will take an infinite number of values. In order that the previous discussions of our algebra apply directly here, we will need to regularize the field theory such that the momentum variables can take only a finite number Λ of distinct values. But this is mostly a technicality, since we will only talk of field theories without divergences for which the limit $\Lambda\to\infty$ should exist: dimensional reductions of gauge theory to $1+1$ dimesnions. The deeper problem of describing renormalizable field theories this way is being studied.

The first gauge field model is a $(1+1)$-dimensional SU(N) gauge theory coupled to bosonic matter in the adjoint representation. This model has been studied by Dalley and Klebanov[16] in detail, and so we will only give a brief account here. Let g be the strong coupling constant, α and β be ordinary space-time indices, A_α be a gauge potential and ϕ be a scalar field in the adjoint representation of the gauge group U(N). Both ϕ and A_α are $N\times N$ Hermitian matrix fields. (We can regard both A_α and ϕ as describing the gluons of a Yang–Mills theory in three dimensions, dimensionally reduced to $1+1$ dimensions.) Let the covariant derivative $D_\alpha\phi=\partial_\alpha\phi+\mathrm{i}[A_\alpha,\phi]$ and the Yang–Mills field $F_{\alpha\beta}=\partial_\alpha A_\beta-\partial_\beta A_\alpha+\mathrm{i}[A_\alpha,A_\beta]$. Lastly, let m be the mass of a boson and g be the gauge field coupling strength. Then the Minkowski space action is

$$S=\int d^2x\ \mathrm{Tr}\left[\frac{1}{2}D_\alpha\phi D^\alpha\phi-\frac{1}{2}m^2\phi^2-\frac{1}{4g^2}F_{\alpha\beta}F^{\alpha\beta}\right]. \tag{39}$$

We then transform the Minkowski coordinates to light-cone coordinates, and choose the light-cone gauge $A_-=(A_0-A_3)/\sqrt{2}=0$. Then A_+ is not dynamical and can be eliminated by the constrained equations. We canonically quantize the adjoint matter field:

$$\phi^\mu_\nu(x^+=0)=\frac{1}{\sqrt{2\pi}}\int_0^\infty\frac{dk^+}{\sqrt{2k^+}}[a^\mu_\nu(k^+)\mathrm{e}^{-\mathrm{i}k^+x^-}+a^{\dagger\mu}_\nu(k^+)\mathrm{e}^{\mathrm{i}k^+x^-}], \tag{40}$$

where μ and ν are color indices. The creation and annihilation operators satisfy the canonical commutation relation equation (34). One of the constraints dictates that the Hilbert space on which the creation and annihilation operators act be spanned by color singlet states:

$$N^{-c/2}\mathrm{Tr}[a^\dagger(k_1^+)\cdots a^\dagger(k_c^+)]|0\rangle. \tag{41}$$

Equation (41) is exactly in the form of a single glueball state $\Psi^{(K)}|0\rangle$ defined in Eq. (35). Thus the creation operator can be interpreted as a creation operator for a gluon. Furthermore, we can obtain the light-cone momentum and energy in terms of the elements of the cyclix algebra:

$$P^+=\int_0^\infty dk\ kg^k_k, \tag{42}$$

and

$$P^-=\frac{1}{2}m^2\int_0^\infty\frac{dk}{k}g^k_k+\frac{g^2N}{4\pi}\int_0^\infty\frac{dk}{k}Cg^k_k+\frac{g^2N}{8\pi}\int_0^\infty\frac{dk_1dk_2dk_3dk_4}{\sqrt{k_1k_2k_3k_4}}\{A\,\delta(k_1+k_2-k_3-k_4)g^{k_1k_2}_{k_3k_4}$$
$$+B\,\delta(k_1-k_2-k_3-k_4)g^{k_1}_{k_2k_3k_4}+B\,\delta(k_1+k_2+k_3-k_4)g^{k_1k_2k_3}_{k_4}\}, \tag{43}$$

where

$$A=\frac{(k_2-k_1)(k_4-k_3)}{(k_1+k_2)^2}-\frac{(k_3+k_1)(k_4+k_2)}{(k_4-k_2)^2},$$

$$B=\frac{(k_1+k_4)(k_3-k_2)}{(k_3+k_2)^2}+\frac{(k_3+k_4)(k_1-k_2)}{(k_1+k_2)^2},$$

and

$$C=\int_0^k dp\,\frac{(k+p)^2}{p(k-p)^2}.$$

This adjoint matter model can be used to study glueball spectrum.[17] Consider next a $(3+1)$-dimensional QCD model with only gluons and no quarks. If we apply the dimensional reduction

to it, we will obtain an effective Yang–Mills model in $1+1$ dimensions with two adjoint matter fields ϕ_1 and ϕ_2, which are constant multiples of transverse gluon fields. Rewrite the adjoint matter fields in the helicity basis

$$\phi_{\pm}=\frac{\phi_1 \pm i\phi_2}{\sqrt{2}}, \tag{44}$$

and canonically quantize $\phi_{\pm}$:

$$\phi^{\mu}_{\pm\,\nu}(x^+=0)=\frac{1}{\sqrt{2\pi}}\int_0^{\infty}\frac{dk^+}{\sqrt{2k^+}}[a^{\mu}_{\nu}(k^+,\pm)e^{-ik^+x^-}+a^{\dagger\mu}_{\nu}(k^+,\pm)e^{ik^+x^-}]. \tag{45}$$

Then the light-front momentum and energy are

$$P^+=\int_0^{\infty}dk\;\; k\sum_{j=+,-} g^{k,j}_{k,j},$$

and

$$\begin{aligned}
P^-=&\frac{m^2_{\rm ind}}{2}\int_0^{\infty}\frac{dk}{k}g^{k,+}_{k,+}+\frac{g^2}{8\pi}\int_0^{\infty}\frac{dk_1dk_2dk_3dk_4}{\sqrt{k_1k_2k_3k_4}}\delta(k_1+k_2-k_3-k_4)\\
&\cdot\{(D[k_1,k_2,k_3]-1)g^{k_3,+;k_4,-}_{k_1,-;k_2,+}+(-E[k_1,k_2,k_3]-1)g^{k_3,+;k_4,+}_{k_1,+;k_2,+}\\
&+(D[k_1,k_2,k_3]-E[k_1,k_2,k_3]+2)g^{k_3,+;k_4,-}_{k_1,+;k_2,-}\}\\
&+\frac{g^2}{8\pi}\int_0^{\infty}\frac{dk_1dk_2dk_3dk_4}{\sqrt{k_1k_2k_3k_4}}\delta(k_1+k_2+k_3-k_4)\\
&\cdot\{(F[k_1,k_2,k_3]+1)\\
&\times(g^{k_1,+;k_2,-;k_3,+}_{k_4,+}+g^{k_3,+;k_2,-;k_1,+}_{k_4,+}+g^{k_4,+}_{k_1,+;k_2,-;k_3,+}+g^{k_4,+}_{k_3,+;k_2,-;k_1,+})\\
&+(F[k_1,k_2,k_3]-1)(g^{k_3,-;k_2,-;k_1,+}_{k_4,-}+g^{k_1,+;k_2,-;k_3,-}_{k_4,-}+g^{k_4,+}_{k_3,+;k_2,+;k_1,-}\\
&+g^{k_4,+}_{k_1,-;k_2,+;k_3,+})\}+(\text{positive helicity}\leftrightarrow\text{negative helicity}),
\end{aligned} \tag{46}$$

where

$$E[x_1,x_2,x_3]=\frac{(x_1+x_3)(x_1+2x_2-x_3)}{(x_1-x_3)^2},\quad D[x_1,x_2,x_3]=E[x_2,-x_3,-x_1],$$

and

$$F[x_3,x_2,x_1]=E[x_3,x_1,-x_2],$$

and $m_{\rm ind}$ is the induced mass due to normal ordering the creation and annihilation operators.

The various algebras described in this paper can also be modified to study supersymmetric models with fermionic adjoint matter fields. Not long ago there were studies on $d>2$ noncritical superstrings using a matrix model.[18] The superfields were presented as matrices. Canonical quantization yielded a number of operators satisfying various canonical commutation and anticommutation relations. The Hilbert space on which these operators acted were composed of closed strings of bosonic and fermionic creation operators. The terms in the supercharge operators could be written in a way analogous to the gluonic operators, except that there were fermionic operators in addition to bosonic operators, which are the only operators present in the gluonic operators defined in Eq. (37). The various algebras described in this paper can be generalized to superalgebras[15] to

study this noncritical superstring model. A closely related model is a supersymmetric Yang–Mills theory with a fermionic matter field.[16,19] Moreover, these superalgebras can also be employed to study M-theory, which is being conjectured as a matrix model.[5]

VI. SOLVABLE MATRIX AND SPIN CHAIN MODELS

Now that we have learned how to formulate gauge field theories in the large-N limit in the language of the cyclix algebra, let us turn to integrable periodic quantum spin chain models and the associated, more tractable multi-matrix models. Understanding the properties of these solvable matrix models may shed new light on the mathematics of the more difficult gauge field theories.

Consider the Hamiltonian of a matrix model in the form of the linear combination $H=\Sigma_{IJ}h_I^J g_J^I$, where $h_I^J\neq 0$ only if I and J have the same number of indices. (This means that the ''parton number'' is conserved. Note that the gauge field theories described in the previous section are not of this type.) Such linear combinations form a subalgebra of the cyclix algebra. Let us call this $\hat{C}_\Lambda^0$.

There is an isomorphism between multi-matrix models whose Hamiltonians are in $\hat{C}_\Lambda^0$ and periodic quantum spin chains. Consider a spin chain with c sites. At any site $p=1,2,\ldots$ or c there is a variable i_p called *spin* that describes the quantum state of that site, and that can take the value 1, 2, . . . , or Λ. We impose the periodic boundary condition. We will show that a subspace of these states, namely, the states with zero total momentum, provides a (nonfaithful) representation of the algebra $\hat{C}_\Lambda^0$.

A basis of states is given by the set of states of the form $|k_1k_2\ldots k_c\rangle$. Define the operator $X_j^i(p)$ by its action on these collective spin chain states by the formula

$$X_j^i(p)|k_1k_2\ldots k_c\rangle\equiv\delta_j^{k_p}|k_1k_2\ldots k_{p-1}ik_{p+1}\ldots k_c\rangle. \tag{47}$$

[Those who are familiar with the theory of the Hubbard model can notice at once that $X_j^i(p)$ is the Hubbard operator at the site p, which is conventionally written as X_p^{ij}.[20]] Furthermore, we impose the periodic boundary condition $X_j^i(p+c)=X_j^i(p)$. It can be verified that if I and J have the same length $a\leqslant c$, then

$$r_c(g_J^I)\equiv\sum_{p=1}^{c}X_{j_1}^{i_1}(p)X_{j_2}^{i_2}(p+1)\cdots X_{j_a}^{i_a}(p+a-1) \tag{48}$$

satisfies the commutation relations of the algebra $\hat{C}_\Lambda^0$. If we further define $r_c(g_J^I)=0$ for $a>c$, we will have a representation r_c of $\hat{C}_\Lambda^0$. The cyclically symmetric states of the matrix model correspond to the states of the periodic spin chain with zero total momentum. The representation is not faithful as it is sets all generators of length a greater than the number of spins c to zero; if we take spin chains of all possible lengths, we will get a faithful representation.

We are now ready to associate the matrix model where the Hamiltonian $H=\Sigma_{IJ}h_I^J g_J^I$ is in $\hat{C}_\Lambda^0$, with the quantum spin chain with the Hamiltonian

$$H^{\text{spin}}=\sum_{IJ}h_I^J\sum_{p=1}^{c}X_{j_1}^{i_1}(p)X_{j_2}^{i_2}(p+1)\cdots X_{j_a}^{i_a}(p+a-1). \tag{49}$$

Therefore, matrix models conserving the parton number correspond to quantum spin systems with interactions involving neighborhoods of spins at the sites p, $p+1,\ldots,$ and $p+a-1$.

Recall that a convenient way of solving spin chain systems[21] is via the Bethe ansatz. This ansatz is applicable whenever the spin wave S-matrix satisfies the Yang–Baxter equation, in which case we say that the spin chain system is integrable. Then this integrable spin chain system yields an integrable multi-matrix model.

Let us look at the simplest example, the quantum Ising spin chain.[22,23] The Hamiltonian $H_{\text{Ising}}^{\text{spin}}$ of this model is

$$H^{\text{spin}}_{\text{Ising}}(\tau,\lambda)=\sum_{p=1}^{c}\tau^z(p)+\lambda\sum_{p=1}^{c}\tau^x(p)\tau^x(p+1). \tag{50}$$

Here λ is a constant, and $\tau_p^{x,y,z}$ are Pauli matrices at site p. Moreover, $\tau_{c+1}^{x,y,z}=\tau^{x,y,z}$. Two Pauli matrices at different sites (i.e., with different subscripts) commute with each other. Let us rewrite the Ising spin chain as a two-matrix model. We set the quantum state 1 for a boson in the matrix model to correspond to the spin-up state in the Ising spin chain, and the quantum state 2 to correspond to the spin-down state. Since

$$\tau_p^z=X_1^1(p)-X_2^2(p),\quad \tau_p^x+\mathrm{i}\tau_p^y=2X_2^1(p)\quad\text{and}\quad \tau_p^x-\mathrm{i}\tau_p^y=2X_1^2(p), \tag{51}$$

we obtain the corresponding element in $\hat{C}_2^0$:

$$H^{\text{matrix}}_{\text{Ising}}=H_0+\lambda V, \tag{52}$$

where

$$H_0=g_1^1-g_2^2\quad\text{and}\quad V=\lambda[g_{11}^{22}+g_{12}^{21}+g_{21}^{12}+g_{22}^{11}]. \tag{53}$$

This is a two-matrix model with the Hamiltonian

$$\begin{aligned}H^{\text{matrix}}_{\text{Ising}}=&\,\mathrm{tr}[a^\dagger(1)a(1)-a^\dagger(2)a(2)]+(\lambda/N)\mathrm{tr}[a^\dagger(2)a^\dagger(2)a(1)a(1)\\&+a^\dagger(2)a^\dagger(1)a(2)a(1)+a^\dagger(1)a^\dagger(2)a(1)a(2)\\&+a^\dagger(1)a^\dagger(1)a(2)a(2)].\end{aligned} \tag{54}$$

Our results, along with known results of the Ising spin chain,[23] give the spectrum of this matrix model in the large N limit:

$$E(n_p,c,\lambda)=-2\sum_{p=-c}^{c}\left(1+2\lambda\,\cos\left[\frac{2\pi p}{2c+1}\right]+\lambda^2\right)^{1/2}n_p, \tag{55}$$

where c is any positive integer and $n_p=0$ or 1. Also, we must impose the condition $\Sigma_{p=-c}{}^{c}n_p p=0$ to get cyclically symmetric states. Let us underscore that *the matrix model defined by the Hamiltonian* $H^{\text{matrix}}_{\text{Ising}}$ *in Eqs. (53) or (54) is an integrable matrix model in the large-N limit.*

Equation (55) manifests the self-duality of the Ising model:

$$E(n_p,c,\lambda)=\lambda E(n_p,c,1/\lambda). \tag{56}$$

Indeed, under the operator dual transformation

$$\tilde{\tau}^x(p)\equiv\tau^z(p)\tau^z(p+1)\quad\text{and}\quad\tilde{\tau}^z(p)\equiv\tau^x(1)\tau^x(2)\cdots\tau^x(p), \tag{57}$$

the Hamiltonian is changed to

$$H^{\text{spin}}_{\text{Ising}}(\tilde{\tau},\lambda)=\sum_{p=1}^{c}\tilde{\tau}^x(p)\tilde{\tau}^x(p+1)+\lambda\sum_{p=1}^{c}\tilde{\tau}^z(p). \tag{58}$$

Since $\tilde{\tau}^x(p)$'s and $\tilde{\tau}^z(p)$'s satisfy the same algebra as $\tau^x(p)$'s and $\tau^z(p)$'s,

$$H^{\text{spin}}_{\text{Ising}}(\tilde{\tau},\lambda)=\lambda H^{\text{spin}}_{\text{Ising}}(\tau,1/\lambda), \tag{59}$$

i.e., the Ising model is self-dual.

It is possible to understand the solvability of the Ising model in terms of the Dolan–Grady conditions[24,25] and the Onsager algebra. Let us digress to summarize this method. Suppose we have a system whose Hamiltonian can be written as

$$H = H_0 + V \tag{60}$$

with the two terms in the Hamiltonian satisfying the Dolan–Grady conditions:

$$[H_0,[H_0,[H_0,V]]] = 16[H_0,V] \tag{61}$$

and

$$[V,[V,[V,H_0]]] = 16[V,H_0]. \tag{62}$$

Then we can construct operators satisfying an infinite-dimensional Lie algebra

$$[A_m,A_n] = 4G_{m-n}, \quad [G_m,A_n] = 2A_{n+m} - 2A_{n-m}, \quad \text{and } [G_m,G_n] = 0 \tag{63}$$

by the following recursion relations:

$$\begin{gathered} A_0 = H_0, \quad A_1 = V, \quad A_{n+1} - A_{n-1} = \tfrac{1}{2}[G_1,A_n], \\ G_1 = \tfrac{1}{4}[A_1,A_0], \quad \text{and } G_n = \tfrac{1}{4}[A_n,A_0]. \end{gathered} \tag{64}$$

This Lie algebra is called the Onsager Lie algebra. It is known to be isomorphic to an infinite direct sum of su(2) algebras. In particular, the system will admit an infinite number of conserved quantities:[26]

$$Q_m = -\tfrac{1}{2}(A_m + A_{-m} + \lambda A_{m+1} + \lambda A_{-m+1}). \tag{65}$$

Thus any such system should be integrable.

In the case of the Ising model we choose H_0 and V as above. The Lie brackets of the g_J^I then allow us to verify easily that the first of the Dolan–Grady conditions is satisfied. This, together with the self-duality of the Ising model, guarantee that the other Dolan–Grady condition is satisfied also. [Equation (62) could also be verified directly. The only caveat is that as the operators act on closed string states only, we have to treat all $\tilde{f}$'s as linear combinations of g's in order for Eq. (62) to hold true.] Moreover, we see that the Onsager algebra is a subalgebra of our algebra $\hat{C}_\Lambda^0$: all the conserved quantities are just linear combinations of our g_J^I and $\tilde{f}_J^I$. This suggests that there may be other models that are integrable by this method: we need to identify pairs of elements in our algebra that satisfy the Dolan–Grady conditions.

We can transcribe other integrable quantum spin chain models with the periodic boundary condition to integrable matrix models in the large-N limit. Some examples are listed below:

(i) *A generalization of the Ising model*.[26] The Hamiltonian is

$$H_{\mathrm{GI}}^{\mathrm{spin}} = \sum_{p=1}^{c} \tau_p^z + \lambda \sum_{p=1}^{c} [\tau_p^x \tau_{p+1}^x + v\{\tau_p^x \tau_{p+1}^y - \tau_p^y \tau_{p+1}^x\}], \tag{66}$$

where λ and v are constants. The corresponding Hamiltonian in the matrix model is

$$H_{\mathrm{GI}}^{\mathrm{matrix}} = g_1^1 - g_2^2 + \lambda[g_{11}^{22} + (1-2iv)g_{12}^{21} + (1+2iv)g_{21}^{12} + g_{22}^{11}]. \tag{67}$$

This model satisfies the Dolan–Grady conditions and hence is integrable using the Onsager subalgebra.

(ii) *The XYZ model*.[27] This is a generalization of the Ising model in another direction. It does not satisfy the Dolan–Grady condition, but is integrable by methods using the Yang–Baxter equation. The Hamiltonian is

$$H_{\mathrm{XYZ}}^{\mathrm{spin}} = \sum_{p=1}^{c} \tau_p^z \tau_{p+1}^z - \lambda \sum_{p=1}^{c} [\tau_p^x \tau_{p+1}^x + v\,\tau_p^y \tau_{p+1}^y]. \tag{68}$$

The corresponding Hamiltonian in the matrix model is

$$H_{\mathrm{XYZ}}^{\mathrm{matrix}}=g_{11}^{11}-g_{12}^{12}-g_{21}^{21}+g_{22}^{22}-\lambda[(1-v)(g_{11}^{22}+g_{22}^{11})+(1+v)(g_{12}^{21}+g_{21}^{12})]. \tag{69}$$

(iii) *The chiral Potts model.*[28] This is a model in which the number of quantum states available for a site is not restricted to 2 but is any finite positive integer. The Hamiltonian is

$$H_{\mathrm{CP}}^{\mathrm{spin}}=\sum_{p=1}^{c}\sum_{k=1}^{\Lambda-1}[\tilde{\alpha}_k Q_p^k+\lambda\alpha_k P_p^k P_{p+1}^{\Lambda-k}], \tag{70}$$

where

$$\alpha_k=\frac{e^{i\phi(2k/n-1)}}{\sin\,\pi k/n},\quad \tilde{\alpha}_k=\frac{e^{i\varphi(2k/n-1)}}{\sin\,\pi k/n},\quad \cos\,\varphi=\lambda\,\cos\,\phi,\quad \omega=e^{2\pi i/n},$$

$$\sigma_j=\begin{pmatrix}1&0&0&\dots&0\\0&\omega&0&\dots&0\\0&0&\omega^2&\dots&0\\&&\dots&&\\0&0&0&\dots&\omega^{n-1}\end{pmatrix},$$

and

$$\Gamma_j=\begin{pmatrix}0&0&0&\dots&0&1\\1&0&0&\dots&0&0\\0&1&0&\dots&0&0\\&&&\dots&&\\0&0&0&\dots&1&0\end{pmatrix}. \tag{71}$$

This model is exactly solvable by the Yang–Baxter method. The Hamiltonian of the associated solvable multi-matrix model is

$$H_{\mathrm{CP}}^{\mathrm{matrix}}=\sum_{k=1}^{\Lambda-1}\left[\tilde{\alpha}_k\sum_{j=1}^{\Lambda}\omega^{k(j-1)}g_j^j+\lambda\alpha_k\sum_{j_1,j_2=1}^{\Lambda}g_{j_1,j_2}^{j_1+k,j_2-k}\right], \tag{72}$$

where j_1+k should be replaced with $j_1+k-\Lambda$ if $j_1+k>\Lambda$ and j_2-k should be replaced with $j_2+\Lambda-k$ if $j_2-k\leqslant 0$ in $g_{j_1,j_2}^{j_1+k,j_2-k}$ in the above equation.

ACKNOWLEDGMENTS

We thank O. T. Turgut for discussions in an early stage of this work. SGR thanks the I.H.E.S., where part of this work was done, for hospitality. We were supported in part by funds provided by the U.S. Department of Energy under Grant No. DE-FG02-91ER40685.

APPENDIX A: NOTATION FOR MULTI-INDICES

Much of our work involves manipulating tensors carrying multiple indices. For the convenience of the reader, we give here a summary of the notations used in this paper for multi-indices. We have attempted to match the notations with our other papers[10,11] and to make this Appendix self-contained. More details can be found in Ref. 11.

We will use lower case Latin letters such as i, j, i_1, and j_2 to denote indices which are positive integers 1, 2, . . . , and Λ. Here Λ itself is a fixed positive integer, denoting the number of degrees of freedom of gluons. A *nonempty* sequence of indices $i_1 i_2 i_3 \cdots i_a$ will be denoted by the corresponding uppercase letter I. The length of the sequence I will be denoted by $\#(I)$.

A capital letter I denotes a *nonempty* sequence of indices $i_1 i_2 \cdots i_a$, each taking values from the set $\{1,2,\ldots,\Lambda\}$. The length $\#(I)$ of I is just the number of elements in the sequence. Two sequences are equal if they have the same entries. Concatenation of two sequences will be denoted by

$$IJ = i_1 i_2 \cdots i_a j_1 j_2 \cdots j_b .$$

In particular,

$$Ij = i_1 i_2 \cdots i_a j, \tag{A1}$$

when only a single index is added at the end.

The Kronecker symbol has the obvious definition:

$$\delta^I_J \equiv \begin{cases} 1 & \text{if } I = J, \\ 0 & \text{if } I \neq J. \end{cases}$$

Thus,

$$\sum_{I_1 I_2 = I} X^{I_1} Y^{I_2}_K \equiv \sum_{I_1 I_2} \delta^{I_1 I_2}_I X^{I_1} Y^{I_2}_K , \tag{A2}$$

where X^{I_1} is a function dependent on I_1 and $Y^{I_2}_K$ is another function dependent on I_2 and K, denotes the sum over all the ways in which a given index I can be split into two *nonempty* subsequences I_1 and I_2. If there is no way to split I as required, then the sum simply yields 0.

We define (I) to be the equivalence class of all cyclic permutations of I. In fact (I) can be viewed as a discrete model for a closed loop, or closed string. The corresponding Kronecker delta function is defined by the following relation:

$$\delta^I_{(J)} \equiv \delta^I_J + \sum_{J_1 J_2 = J} \delta^I_{J_2 J_1} . \tag{A3}$$

Equation (A3) means that the delta function returns the number of different cyclic permutations of J such that each permuted sequence is identical with I. *Thus $\delta^I_{(J)}$ can take any non-negative integer as its value, not just 0 or 1.* The reader can verify from this definition that

$$\delta^I_{(J)} = \delta^J_{(I)} . \tag{A4}$$

Next, the expression

$$\sum_{(I)} X^I ,$$

where X^I is dependent on the equivalence class (I), i.e., $X^I = X^J$ if $(I) = (J)$, means that *all possible distinct equivalence classes* (I) are summed. Note that each equivalence class appears only *once* in the sum. In all cases of interest to us, it turns out that there are only a finite number of (I)'s such that $X^I \neq 0$.

Now we can introduce the formula that defines the following summation:

$$\sum_{I_1 I_2 \cdots I_n = (I)} X^{I_1, I_2, \ldots, I_n} \equiv \sum_{I_1, I_2, \ldots, I_n} \delta^{I_1 I_2 \cdots I_n}_{(I)} X^{I_1, I_2, \ \ldots \ , I_n}. \tag{A5}$$

In words, in Eq. (A5) we sum over all distinct ways of cyclically permuting I, and then all *distinct sets* of n nonempty sequences $I_1, I_2, \ldots, I_n$ (but note that within a particular set of $I_1, I_2, \ldots, I_n$, some of the sequences can be identical) such that $I_1 I_2 \cdots I_n$ is the same as this permuted sequence. This equation, together with Eq. (A4), then leads to

$$\delta_{(I)}^{I_1 I_2 \cdots I_n} = \sum_{I_1' I_2' \cdots I_n' = (I)} \delta_{I_1'}^{I_1} \delta_{I_2'}^{I_2} \cdots \delta_{I_n'}^{I_n}. \tag{A6}$$

A direct consequence of Eq. (A6) is

$$\delta_{(I)}^{I_1 I_2 \cdots I_n} = \delta_{(I)}^{I_2 I_3 \cdots I_n I_1}.$$

In addition, the reader can verify from Eq. (A5) together with Eqs. (A2), (A3) and (A4) that

$$\begin{aligned} \sum_{I_1 I_2 \cdots I_n = (I)} = \sum_{I_1, I_2, \ldots, I_n} \Bigg(& \delta_I^{I_1 I_2 \cdots I_n} + \sum_{I_{11} I_{12} = I_1} \delta_I^{I_{12} I_2 I_3 \cdots I_n I_{11}} + \delta_I^{I_2 I_3 \cdots I_n I_1} \\ & + \sum_{I_{21} I_{22} = I_2} \delta_I^{I_{22} I_3 \cdots I_n I_1 I_{21}} + \cdots + \delta_I^{I_n I_1 I_2 \cdots I_{n-1}} \\ & + \sum_{I_{n1} I_{n2} = I_n} \delta_I^{I_{n2} I_1 I_2 \cdots I_{n-1} I_{n1}} \Bigg). \end{aligned} \tag{A7}$$

APPENDIX B: LINEAR INDEPENDENCE OF STRINGLIKE OPERATORS

In this Appendix, we are going to show that the set of all $\tilde{f}_{(J)}^{(I)}$'s defined by Eq. (2) and all γ_J^I's defined by Eqs. (3) and (4) is linearly independent. This can be proved by *ad absurdum* as follows. Assume on the contrary that the set is linearly dependent. Then there exists an equation

$$\sum_{p=1}^{r} \alpha_p \tilde{f}_{(J_p)}^{(I_p)} + \sum_{p=r+1}^{s} \alpha_p \gamma_{J_p}^{I_p} = 0 \tag{B1}$$

such that r and s are positive integers with $r \leqslant s$ ($r = s$ means that the second sum vanishes), and all α_p's for $p = 1, 2, \ldots,$ and s are nonzero complex constants. In addition, in the above equation either $(I_p) \neq (I_q)$ or $(J_p) \neq (J_q)$ if $1 \leqslant p \leqslant r$, $1 \leqslant q \leqslant r$ and $p \neq q$, and either $I_p \neq I_q$ or $J_p \neq J_q$ if $r+1 \leqslant p \leqslant s$, $r+1 \leqslant q \leqslant s$ and $p \neq q$.

Assume that $s > r$. We can assume without loss of generality that

(1) $J_{r+1} = J_{r+2} = \cdots = J_{r+x}$;
(2) $J_{r+1} \neq J_{r+x+1}$, $J_{r+1} \neq J_{r+x+2}$, ..., and $J_{r+1} \neq J_s$; and
(3) $\#(J_{r+1}) \leqslant \#(J_p)$ for all $p = r+1, r+2, \ldots,$ and s for some integer x such that $r < x \leqslant s$.

Consider the action of the lhs of Eq. (B1) on $s^{J_{r+1}}$. We get

$$\left(\sum_{p=1}^{r} \alpha_p \tilde{f}_{(J_p)}^{(I_p)} + \sum_{p=r+1}^{s} \alpha_p \gamma_{J_p}^{I_p} \right) s^{J_{r+1}} = \sum_{p=r+1}^{r+x} \alpha_p s^{I_p}. \tag{B2}$$

Combining Eqs. (B1) and (B2) yields

$$\sum_{p=r+1}^{r+x} \alpha_p s^{I_p} = 0,$$

which is impossible because s^{I_r}, $s^{I_{r+1}}$, ..., $s^{I_{r+x}}$ are linearly independent. Therefore, $s = r$ and Eq. (B1) can be simplified to

$$\sum_{p=1}^{r} \alpha_p \tilde{f}_{(J_p)}^{(I_p)} = 0. \tag{B3}$$

Again we can assume without loss of generality that

(1) $(J_1) = (J_2) = \cdots = (J_y)$; and

(2) $(J_1)\neq(J_{y+1})$, $(J_1)\neq(J_{y+2})$, ..., and $(J_1)\neq(J_r)$ for some integer y such that $1\leqslant y\leqslant r$.

Consider the action of the lhs of Eq. (B3) on $\tilde{f}^{(I_1)}_{(J_1)}$:

$$\left(\sum_{p=1}^{r}\alpha_p\tilde{f}^{(I_p)}_{(J_p)}\right)\Psi^{(I_1)}_{(J_1)}=\sum_{p=1}^{y}\delta^{J_1}_{(J_1)}\alpha_p\Psi^{(I_p)}.$$

However, the rhs of this equation is impossible to vanish because $\Psi^{(I_1)}$, $\Psi^{(I_2)}$, ..., $\Psi^{(I_y)}$ are linearly independent. Thus the set of all $\tilde{f}^{(I)}_{(J)}$'s together with all γ^I_J's is linearly independent. Q.E.D.

APPENDIX C: MULTIPLICATION OF TWO STRINGLIKE OPERATORS

The assertion that the product $\gamma^I_J\gamma^K_L$ cannot be written in general as a finite linear combination of γ's and $\tilde{f}$'s can be proved by contradiction as follows. Consider the case when $\Lambda=1$. Let $\gamma^a_b=\gamma^{11\cdots1}_{11\cdots1}$, where the number 1 shows up a times in the superscript and b times in the subscript of γ. Moreover, let $\Psi^{(c)}=\Psi^{(11\cdots1)}$, where the number 1 shows up c times, and $s^d=s^{11\cdots1}$, where the number 1 shows up d times.

Assume that $\gamma^1_1\gamma^1_1=\Sigma_{p=1}{}^r\alpha_p\gamma^p_p+\Sigma_{q=1}{}^s\beta_q\tilde{f}^{(q)}_{(q)}$, where α_1, α_2, ..., α_r, β_1, β_2, ..., and β_s are nonzero complex numbers for some positive integers r and s. Then from the equations $\gamma^1_1\gamma^1_1(s^1)=\gamma^1_1(\gamma^1_1s^1)=1^2s^1$, $\gamma^1_1\gamma^1_1(s^2)=\gamma^1_1(\gamma^1_1s^1)=2^2s^1$, ..., and $\gamma^1_1\gamma^1_1(s^r)=\gamma^1_1(\gamma^1_1s^1)=r^2s^1$, we deduce that $\alpha_1=1$ and $\alpha_2=\alpha_3=\cdots=\alpha_r=2$. Hence $\gamma^1_1\gamma^1_1=\gamma^1_1+2\Sigma_{p=2}{}^r\gamma^p_p+\Sigma_{q=1}{}^s\beta_q\tilde{f}^{(q)}_{(q)}$. However, $\gamma^1_1\gamma^1_1(s^{r+1})=(r+1)^2s^{r+1}$ and $\gamma^1_1+2\Sigma_{p=2}{}^r\gamma^p_p+\Sigma_{q=1}{}^s\beta_q\tilde{f}^{(q)}_{(q)}(s^{r+1})=(r^2+2r-1)s^{r+1}$, leading to a contradiction.

Thus we assume instead $\gamma^1_1\gamma^1_1=\Sigma_{q=1}{}^s\beta_q\tilde{f}^{(q)}_{(q)}$, where the β_q's are nonzero complex numbers. However, $\gamma^1_1\gamma^1_1(\Psi^{(s+1)})=(s+1)^2\Psi^{(s+1)}$ whereas $\Sigma_{q=1}{}^s\beta_q\tilde{f}^{(q)}_{(q)}\Psi^{(s+1)}=0$, leading to a contradiction, too. Consequently, it is impossible to write $\gamma^1_1\gamma^1_1$ as a finite linear combination of γ's and $\tilde{f}$'s.

This proof can be easily generalized to the case $\Lambda>1$. Q.E.D.

APPENDIX D: THE COMMUTATION RELATIONS OF THE STRING-LIKE OPERATORS

What we need to do is to show that the three equations satisfy

$$[\gamma^I_J,\gamma^K_L]\Psi^{(P)}=\gamma^I_J(\gamma^K_L\Psi^{(P)})-\gamma^K_L(\gamma^I_J\Psi^{(P)}),\tag{D1}$$

$$[\gamma^I_J,\tilde{f}^{(K)}_{(L)}]\Psi^{(P)}=\gamma^I_J(\tilde{f}^{(K)}_{(L)}\Psi^{(P)})-\tilde{f}^{(K)}_{(L)}(\gamma^I_J\Psi^{(P)}),\tag{D2}$$

$$[\tilde{f}^{(I)}_{(J)},\tilde{f}^{(K)}_{(L)}]\Psi^{(P)}=\tilde{f}^{(I)}_{(J)}(\tilde{f}^{(K)}_{(L)}\Psi^{(P)})-\tilde{f}^{(K)}_{(L)}(\tilde{f}^{(I)}_{(J)}\Psi^{(P)}),\tag{D3}$$

$$[\gamma^I_J,\gamma^K_L]s^M=\gamma^I_J(\gamma^K_Ls^M)-\gamma^K_L(\gamma^I_Js^M),\tag{D4}$$

$$[\gamma^I_J,\tilde{f}^{(K)}_{(L)}]s^M=\gamma^I_J(\tilde{f}^{(K)}_{(L)}s^M)-\tilde{f}^{(K)}_{(L)}(\gamma^I_Js^M),\tag{D5}$$

and

$$[\tilde{f}^{(I)}_{(J)},\tilde{f}^{(K)}_{(L)}]s^M=\tilde{f}^{(I)}_{(J)}(\tilde{f}^{(K)}_{(L)}s^M)-\tilde{f}^{(K)}_{(L)}(\tilde{f}^{(I)}_{(J)}s^M).\tag{D6}$$

for any integer sequences I, J, K, L, M, and P. Equations (D5) and (D6) are trivially true. That Eq. (8) satisfies Eq. (D3) is also straightforward. What is remaining is whether Eq. (7) satisfies Eq. (D2), and whether Eq. (6) satisfies Eqs. (D4) and (D1).

Consider Eq. (7). The action of the Lie bracket operator on the lhs of this equation on $\Psi^{(P)}$, where P is arbitrary, can be evaluated using Eqs. (5) and (2), and we get

$$[\gamma^I_J,\tilde{f}^{(K)}_{(L)}]\Psi^{(P)}=\sum_{(Q)}\left(\delta^I_{(Q)}\delta^K_{(J)}\delta^P_{(L)}+\sum_A\delta^{IA}_{(Q)}\delta^K_{(JA)}\delta^P_{(L)}\right.$$

$$- \delta^K_{(Q)} \delta^I_{(L)} \delta^P_{(J)} - \sum_{A'} \delta^K_{(Q)} \delta^{IA'}_{(L)} \delta^P_{(JA')} \Bigg) \Psi^{(Q)}. \tag{D7}$$

On the other hand, the action of the operators on the rhs of Eq. (7) (let us call this linear combination of operators $g\tilde{f}^{I,K}_{J,L}$) on $\Psi^{(P)}$ is

$$g\tilde{f}^{I,K}_{J,L}\Psi^{(P)} = \sum_{(Q)} \Bigg(\delta^K_{(J)} \delta^I_{(Q)} \delta^P_{(L)} + \sum_{K_1 K_2 = (K)} \delta^{K_1}_J \delta^{IK_2}_{(Q)} \delta^P_{(L)} - \delta^I_{(L)} \delta^K_{(Q)} \delta^P_{(J)} - \sum_{L_1 L_2 = (L)} \delta^I_{L_2} \delta^K_{(Q)} \delta^P_{(L_1 J)} \Bigg) \Psi^{(Q)}. \tag{D8}$$

The rhs of Eqs. (D7) and (D8) can be seen to be the same by using the delta function defined in Eq. (A3).

Let us determine the correctness of Eq. (6). The properties of the delta functions discussed in Appendix A will be extensively used. To verify Eq. (D4), consider the action of the Lie bracket operator on the lhs of Eq. (6) on s^P, where P is arbitrary:

$$\begin{aligned}
[\gamma^I_J, \gamma^K_L] s^P = \sum_Q \Bigg(& \delta^I_Q \delta^K_J \delta^P_L + \sum_C \delta^I_Q \delta^{CK}_J \delta^P_{CL} + \sum_D \delta^I_Q \delta^{KD}_J \delta^P_{LD} \\
& + \sum_{C,D} \delta^I_Q \delta^{CKD}_J \delta^P_{CLD} + \sum_A \delta^{AI}_Q \delta^K_{AJ} \delta^P_L + \sum_{A,C} \delta^{AI}_Q \delta^{CK}_{AJ} \delta^P_{CL} \\
& + \sum_{A,D} \delta^{AI}_Q \delta^{KD}_{AJ} \delta^P_{LD} + \sum_{A,C,D} \delta^{AI}_Q \delta^{CKD}_{AJ} \delta^P_{CLD} + \sum_B \delta^{IB}_Q \delta^K_{JB} \delta^P_L \\
& + \sum_{B,C} \delta^{IB}_Q \delta^{CK}_{JB} \delta^P_{CL} + \sum_{B,D} \delta^{IB}_Q \delta^{KD}_{JB} \delta^P_{LD} + \sum_{B,C,D} \delta^{IB}_Q \delta^{CKD}_{JB} \delta^P_{CLD} \\
& + \sum_{A,B} \delta^{AIB}_Q \delta^K_{AJB} \delta^P_L + \sum_{A,B,C} \delta^{AIB}_Q \delta^{CK}_{AJB} \delta^P_{CL} + \sum_{A,B,D} \delta^{AIB}_Q \delta^{KD}_{AJB} \delta^P_{LD} \\
& + \sum_{A,B,C,D} \delta^{AIB}_Q \delta^{CKD}_{AJB} \delta^P_{CLD} \Bigg) s^Q - (I \leftrightarrow K, J \leftrightarrow L).
\end{aligned} \tag{D9}$$

After a tedious calculation (for more details on the intermediate steps, please see Ref. 14), Eq. (D9) leads to

$$\begin{aligned}
[\gamma^I_J, \gamma^K_L] s^P = \sum_Q \Bigg\{ & \delta^K_J \Bigg(\delta^I_Q \delta^P_L + \sum_E \delta^{EI}_Q \delta^P_{EL} + \sum_F \delta^{IF}_Q \delta^P_{LF} + \sum_{E,F} \delta^{EIF}_Q \delta^P_{ELF} \Bigg) \\
& + \sum_{J_1 J_2 = J} \delta^K_{J_2} \Bigg(\delta^I_Q \delta^P_{J_1 L} + \sum_E \delta^{EI}_Q \delta^P_{EJ_1 L} + \sum_F \delta^{IF}_Q \delta^P_{J_1 LF} + \sum_{E,F} \delta^{EIF}_Q \delta^P_{EJ_1 LF} \Bigg) \\
& + \sum_{K_1 K_2 = K} \delta^{K_1}_J \Bigg(\delta^{IK_2}_Q \delta^P_L + \sum_E \delta^{EIK_2}_Q \delta^P_{EL} + \sum_F \delta^{IK_2 F}_Q \delta^P_{LF} + \sum_{E,F} \delta^{EIK_2 F}_Q \delta^P_{ELF} \Bigg) \\
& + \sum_{J_1 J_2 = J} \delta^K_{J_1} \Bigg(\delta^I_Q \delta^P_{LJ_2} + \sum_E \delta^{EI}_Q \delta^P_{ELJ_2} + \sum_F \delta^{IF}_Q \delta^P_{LJ_2 F} + \sum_{E,F} \delta^{EIF}_Q \delta^P_{ELJ_2 F} \Bigg) \\
& + \sum_{K_1 K_2 = K} \delta^{K_2}_J \Bigg(\delta^{K_1 I}_Q \delta^P_L + \sum_E \delta^{EK_1 I}_Q \delta^P_{EL} + \sum_F \delta^{K_1 IF}_Q \delta^P_{LF} + \sum_{E,F} \delta^{EK_1 IF}_Q \delta^P_{ELF} \Bigg)
\end{aligned}$$

$$+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\delta^{K_1}_{J_2}\Big(\delta^{IK_2}_{Q}\delta^{P}_{J_1L}+\sum_{E}\delta^{EIK_2}_{Q}\delta^{P}_{EJ_1L}+\sum_{F}\delta^{IK_2F}_{Q}\delta^{P}_{J_1LF}+\sum_{E,F}\delta^{EIK_2F}_{Q}\delta^{P}_{EJ_1LF}\Big)$$

$$+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\delta^{K_2}_{J_1}\Big(\delta^{K_1I}_{Q}\delta^{P}_{LJ_2}+\sum_{E}\delta^{EK_1I}_{Q}\delta^{P}_{ELJ_2}+\sum_{F}\delta^{K_1IF}_{Q}\delta^{P}_{LJ_2F}+\sum_{E,F}\delta^{EK_1IF}_{Q}\delta^{P}_{ELJ_2F}\Big)$$

$$+\sum_{J_1J_2J_3=J}\delta^{K}_{J_2}\Big(\delta^{I}_{Q}\delta^{P}_{J_1LJ_3}+\sum_{E}\delta^{EI}_{Q}\delta^{P}_{EJ_1LJ_3}+\sum_{F}\delta^{IF}_{Q}\delta^{P}_{J_1LJ_3F}+\sum_{E,F}\delta^{EIF}\delta^{P}_{EJ_1LJ_3F}\Big)$$

$$+\sum_{K_1K_2K_3=K}\delta^{K_2}_{J}\Big(\delta^{K_1IK_3}_{Q}\delta^{P}_{L}+\sum_{E}\delta^{EK_1IK_3}_{Q}\delta^{P}_{EL}$$

$$+\sum_{F}\delta^{K_1IK_3F}_{Q}\delta^{P}_{LF}+\sum_{E,F}\delta^{EK_1IK_3F}_{Q}\delta^{P}_{ELF}\Big)+\Gamma_1\Bigg\}s^{Q}-(I\leftrightarrow K,J\leftrightarrow L),\tag{D10}$$

where

$$\Gamma_1=\delta^{IK}_{Q}\delta^{P}_{JL}+\sum_{E}\delta^{EIK}_{Q}\delta^{P}_{EJL}+\sum_{F}\delta^{IFK}_{Q}\delta^{P}_{JFL}+\sum_{G}\delta^{IKG}_{Q}\delta^{P}_{JLG}$$

$$+\sum_{E,F}\delta^{EIFK}_{Q}\delta^{P}_{EJFL}+\sum_{F,G}\delta^{IFKG}_{Q}\delta^{P}_{JFLG}+\sum_{E,G}\delta^{EIKG}_{Q}\delta^{P}_{EJLG}$$

$$+\sum_{E,F,G}\delta^{EIFKG}_{Q}\delta^{P}_{EJFLG}+(I\leftrightarrow K,J\leftrightarrow L).\tag{D11}$$

If we substitute Eq. (D10) without Γ_1 into Eq. (D9), we will obtain exactly the action of operators on the rhs of Eq. (7) on s^P. Γ_1 is reproduced when I and K are interchanged with J and L respectively in Eq. (D11) and so it is cancelled. Consequently, Eq. (D4) is indeed satisfied.

Now let us verify Eq.(D1). Consider the Lie bracket operator on the lhs of Eq. (6) on $\Psi^{(P)}$, where P is again arbitrary. We then obtain

$$[\gamma^{I}_{J},\gamma^{K}_{L}]\Psi^{(P)}=\sum_{(Q)}\Big(\delta^{I}_{(Q)}\delta^{K}_{(J)}\delta^{P}_{(L)}+\sum_{A}\delta^{IA}_{(Q)}\delta^{K}_{(JA)}\delta^{P}_{(L)}+\sum_{B}\delta^{I}_{(Q)}\delta^{KB}_{(J)}\delta^{P}_{(LB)}$$

$$+\sum_{A,B}\delta^{IA}_{(Q)}\delta^{KB}_{(JA)}\delta^{P}_{(LB)}\Big)\Psi^{(Q)}-(I\leftrightarrow K,J\leftrightarrow L).\tag{D12}$$

The first three summations on the rhs of this equation can be turned into the following expressions:

$$\delta^{I}_{(Q)}\delta^{K}_{(J)}\delta^{P}_{(L)}=\delta^{K}_{J}\delta^{I}_{(Q)}\delta^{P}_{(L)}+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\delta^{K_1}_{J_2}\delta^{K_2}_{J_1}\delta^{I}_{(Q)}\delta^{P}_{(L)},\tag{D13}$$

$$\sum_{A}\delta^{IA}_{(Q)}\delta^{K}_{(JA)}\delta^{P}_{(L)}=\sum_{K_1K_2=K}\delta^{K_1}_{J}\delta^{IK_2}_{(Q)}\delta^{P}_{(L)}+\sum_{K_1K_2=K}\delta^{K_2}_{J}\delta^{IK_1}_{(Q)}\delta^{P}_{(L)}$$

$$+\sum_{K_1K_2K_3=K}\delta^{K_2}_{J}\delta^{K_1IK_3}_{(Q)}\delta^{P}_{(L)}+\sum_{\substack{J_1J_2=J\\K_1K_2K_3=K}}\delta^{K_1}_{J_2}\delta^{K_3}_{J_1}\delta^{IK_2}_{(Q)}\delta^{L}_{(L)},\tag{D14}$$

and

$$\sum_{B}\delta^{I}_{(Q)}\delta^{KB}_{(J)}\delta^{P}_{(LB)}=\sum_{J_1J_2=J}\delta^{K}_{J_2}\delta^{I}_{(Q)}\delta^{P}_{(J_1L)}+\sum_{J_1J_2=J}\delta^{K}_{J_1}\delta^{I}_{(Q)}\delta^{P}_{(J_2L)}$$

$$+\sum_{J_1J_2J_3=J}\delta^K_{J_2}\delta^I_{(Q)}\delta^P_{(J_1LJ_3)}+\sum_{\substack{J_1J_2J_3=J\\K_1K_2=K}}\delta^{K_1}_{J_3}\delta^{K_2}_{J_1}\delta^I_{(Q)}\delta^P_{(J_2L)}. \tag{D15}$$

The fourth summation is more complicated. This can be manipulated to be

$$\begin{aligned}\sum_{A,B}\delta^{IA}_{(Q)}\delta^{KB}_{(JA)}\delta^P_{(LB)}=\delta^K_J\sum_C\delta^{IC}_{(Q)}\delta^P_{(LC)}&+\sum_{J_1J_2=J}\sum_C\delta^K_{J_1}\delta^{IC}_{(Q)}\delta^P_{(LJ_2C)}\\&+\sum_{K_1K_2=K}\sum_C\delta^{K_1}_J\delta^{IK_2C}_{(Q)}\delta^P_{(LC)}+\sum_{J_1J_2=J}\sum_C\delta^K_{J_2}\delta^{IC}_{(Q)}\delta^P_{(J_1LC)}\\&+\sum_{K_1K_2=K}\sum_C\delta^{K_2}_J\delta^{K_1IC}_{(Q)}\delta^P_{(LC)}+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\delta^{K_2}_{J_1}\delta^{IK_1}_{(Q)}\delta^P_{(J_2L)}\\&+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\delta^{K_1}_{J_2}\delta^{IK_2}_{(Q)}\delta^P_{(J_1L)}+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\sum_C\delta^{K_1}_{J_2}\delta^{IK_2C}_{(Q)}\delta^P_{(J_1LC)}\\&+\sum_{\substack{J_1J_2=J\\K_1K_2=K}}\sum_C\delta^{K_2}_{J_1}\delta^{K_1IC}_{(Q)}\delta^P_{(LJ_2C)}+\sum_{J_1J_2J_3=J}\sum_C\delta^K_{J_2}\delta^{IC}_{(Q)}\delta^P_{(J_1LJ_3C)}\\&+\sum_{K_1K_2K_3=K}\sum_C\delta^{K_2}_J\delta^{K_1IK_3C}_{(Q)}\delta^P_{(LC)}+\sum_{\substack{J_1J_2J_3=J\\K_1K_2K_3=K}}\delta^{K_1}_{J_3}\delta^{K_3}_{J_1}\delta^{IK_2}_{(Q)}\delta^P_{(J_2L)}+\Gamma_2,\end{aligned} \tag{D16}$$

where

$$\begin{aligned}\Gamma_2=&\sum_{\substack{P_1P_2=(P)\\Q_1Q_2=(Q)}}\delta^I_{Q_1}\delta^K_{Q_2}\delta^{P_1}_L\delta^{P_2}_J+\sum_{\substack{P_1P_2P_3=(P)\\Q_1Q_2Q_3=(Q)}}\delta^K_{Q_1}\delta^I_{Q_2}\delta^{P_1}_{Q_3}\delta^{P_2}_L\delta^{P_3}_J\\&+\sum_{\substack{P_1P_2P_3=(P)\\Q_1Q_2Q_3=(Q)}}\delta^I_{Q_1}\delta^K_{Q_2}\delta^{P_1}_{Q_3}\delta^{P_2}_J\delta^{P_3}_L+\sum_{\substack{P_1P_2P_3P_4=(P)\\Q_1Q_2Q_3Q_4=(Q)}}\delta^I_{Q_1}\delta^K_{Q_3}\delta^{P_1}_J\delta^{P_3}_L\delta^{P_2}_{Q_2}\delta^{P_4}_{Q_4}.\end{aligned} \tag{D17}$$

If we substitute Eqs. (D13), (D14), (D15) and (D16) without Γ_2 into Eq. (D12), we will obtain exactly the action of the operators on the rhs of Eq. (6) on $\Psi^{(P)}$. Γ_2 is reproduced when I and K are interchanged with J and L, respectively, in Eq. (D17) and so it is cancelled. Hence Eq. (D1) is also satisfied. Consequently, Eq. (6) is true. Q.E.D.

APPENDIX E: CARTAN SUBALGEBRA OF THE HETERIX ALGEBRA

First of all, it is obvious that $\mathcal{H}$ is Abelian. *A fortiori*, $\mathcal{H}$ is nilpotent. To proceed on, let us digress and state the following two lemmas:

Lemma 1: Let

$$[\gamma^I_I,\gamma^K_L]=\sum_{k=1}^{p}\alpha^{N_k}_{M_k}\gamma^{M_k}_{N_k}+\sum_{k=p+1}^{q}\alpha^{N_k}_{M_k}\tilde{f}^{(M_k)}_{(N_k)},$$

$$[\gamma^I_I,\tilde{f}^{(K)}_{(L)}]=\sum_{k=1}^{q}\alpha^{N_k}_{M_k}\tilde{f}^{(M_k)}_{(N_k)},\quad [\tilde{f}^{(I)}_{(I)},\gamma^K_L]=\sum_{k=1}^{q}\alpha^{N_k}_{M_k}\tilde{f}^{(M_k)}_{(N_k)},$$

and

$$[\tilde{f}^{(I)}_{(I)},\tilde{f}^{(K)}_{(L)}]=(\delta^{K}_{(I)}-\delta^{I}_{(L)})\tilde{f}^{(K)}_{(L)},$$

where p and q are finite non-negative integers such that $q\geqslant p$, M_k's *and* N_k's *are positive integer sequences such that* $\gamma^{M_k}_{N_k}\neq\gamma^{M'_k}_{N'_k}$ *and* $\tilde{f}^{(M_k)}_{(N_k)}\neq\tilde{f}^{(M'_k)}_{(N'_k)}$ *for* $k\neq k'$, *and* $\alpha^{N_k}_{M_k}$'s *are nonzero numerical coefficients. Then*

$$\#(M_k)-\#(N_k)=\#(K)-\#(L)$$

for every $k=1,2,\ldots,p$ or q.

This lemma can be proved by using Eqs. (6), (7) and (8) with $J=I$.

Lemma 2: With the same assumptions as in the previous lemma, we have

$$\#(M_k)+\#(N_k)\geqslant\#(K)+\#(L)$$

for every $k=1,2,\ldots,p$ *or* q.

This lemma can also be proved by using the same equations with $J=I$.

Let r and s be *positive* integers such that $s\geqslant r$. We are now ready to show for arbitrary nonzero complex numbers $\xi^{L_i}_{K_i}$'s where $i=1,2,\ldots,s$ and arbitrary integer sequences L_i's and K_i's such that $\gamma^{K_i}_{L_i}\neq\gamma^{K'_i}_{L'_i}$ and $\tilde{f}^{(K_i)}_{(L_i)}\neq\tilde{f}^{(K'_i)}_{(L'_i)}$ for $i\neq i'$, and $K_i\neq L_i$ for at least one i in $\{1,2,\ldots,r\}$ that there exisits a sequence I such that

$$\left[\gamma^{I},\sum_{i=1}^{r}\xi^{L_i}_{K_i}\gamma^{K_i}_{L_i}+\sum_{i=r+1}^{s}\xi^{L_i}_{K_i}\tilde{f}^{(K_i)}_{(L_i)}\right]$$

does not belong to $\mathcal{H}$. Indeed, let j be an integer such that

(1) $K_j\neq L_j$;
(2) $\#(K_j)-\#(L_j)\geqslant\#(K_i)-\#(L_i)$ for all $i=1,2,\ldots$ and r; and
(3) $\#(K_j)+\#(L_j)\leqslant\#(K_i)+\#(L_i)$ for any $i=1,2,\ldots$ or r such that $\#(K_j)-\#(L_j)=\#(K_i)-\#(L_i)$.

If $\#(K_j)\geqslant\#(L_j)$, then consider

$$\begin{aligned}\left[\gamma^{K_j}_{K_j},\sum_{i=1}^{r}\xi^{L_i}_{K_i}\gamma^{K_i}_{L_i}+\sum_{i=r+1}^{s}\xi^{L_i}_{K_i}\tilde{f}^{(K_i)}_{(L_i)}\right]&=\xi^{L_j}_{K_j}\gamma^{K_j}_{L_j}+\xi^{L_j}_{K_j}\left(\sum_{k=2}^{p_j}\alpha^{N_{jk}}_{M_{jk}}\gamma^{M_{jk}}_{N_{jk}}+\sum_{k=p_j+1}^{q_j}\alpha^{N_{jk}}_{M_{jk}}\tilde{f}^{(M_{jk})}_{(N_{jk})}\right)\\&\quad+\sum_{\substack{i=1\\ i\neq j}}^{r}\xi^{L_i}_{K_i}\left(\sum_{k=1}^{p_i}\alpha^{N_{ik}}_{M_{ik}}\gamma^{M_{ik}}_{N_{ik}}+\sum_{k=p_i+1}^{q_i}\alpha^{N_{ik}}_{M_{ik}}\tilde{f}^{(M_{ik})}_{(N_{ik})}\right)\\&\quad+\sum_{i=r+1}^{s}\xi^{L_i}_{K_i}\sum_{k=1}^{q'_i}\alpha^{N_{ik}}_{M_{ik}}\tilde{f}^{(M_{ik})}_{(N_{ik})},\end{aligned}\tag{E1}$$

where we have set $\alpha^{N_{j1}}_{M_{j1}}=1$, and each p_i, q_i for $i=1,2,\ldots,r$ and q'_i for $i=r+1,2,\ldots,s$ are dependent on i. Let us assume that

$$\gamma^{M_{ik}}_{N_{ik}}=\gamma^{K_i}_{L_i}\tag{E2}$$

for some $i\in\{1,2,\ldots,r\}$ but $i\neq j$ and $k\in\{1,2,\ldots,p_i\}$. Then $\#(M_{ik})=\#(K_j)$ and $\#(N_{ik})=\#(L_j)$. By Lemmas 1 and 2, we get $\#(K_i)=\#(K_j)$ and $\#(L_i)=\#(L_j)$. However, we also know that $K_i\neq K_j$ or $L_i\neq L_j$ and so there is no $k\in\{1,2,\ldots,p_i\}$ such that Eq. (E2) holds. This is a contradiction and so we conclude that

$$\gamma^{M_{ik}}_{N_{ik}}\neq\gamma^{K_i}_{L_i}$$

for *all* $i \in \{1,2,\ldots,r\}$ but $i \neq j$ and $k \in \{1,2,\ldots,p_i\}$. From Eq. (E1), we deduce that

$$\left[\gamma_{K_j}^{K_j}, \sum_{i=1}^{r} \xi_{K_i}^{L_i} \gamma_{L_i}^{K_i} + \sum_{i=r+1}^{s} \xi_{K_i}^{L_i} \tilde{f}_{(L_i)}^{(K_i)}\right]$$

does not belong to $\mathcal{H}$. Similarly, if $\#(K_j) \leqslant \#(L_j)$, then

$$\left[\gamma_{L_j}^{L_j}, \sum_{i=1}^{r} \xi_{K_i}^{L_i} \gamma_{L_i}^{K_i} + \sum_{i=1}^{s} \xi_{K_i}^{L_i} \tilde{f}_{(L_i)}^{(K_i)}\right]$$

does not belong to $\mathcal{H}$. Consequently,

$$\sum_{i=1}^{r} \xi_{K_i}^{L_i} \gamma_{L_i}^{K_i} + \sum_{i=r+1}^{s} \xi_{K_i}^{L_i} \tilde{f}_{(L_i)}^{(K_i)}$$

does not belong to the normalizer (see Ref. 12 for the definition of a normalizer) of $\mathcal{H}$ if $r \neq 0$ and $K_i \neq L_i$ for at least one $i = 1,2,\leqslant r$.

Now, consider

$$\sum_{i=1}^{r} \xi_{K_i}^{L_i} \gamma_{K_i}^{K_i} + \sum_{i=r+1}^{s} \xi_{K_i}^{L_i} \tilde{f}_{(L_i)}^{(K_i)},$$

where r is a non-negative integer and s a positive integer, $\xi_{K_i}^{L_i}$'s are complex constants and $\tilde{f}_{(L_i)}^{(K_i)} \neq \tilde{f}_{(L_i')}^{(K_i')}$ for $i \neq i'$, and $(K_i) \neq (L_i)$ for at least one $i \in \{r+1, r+2, \ldots, s\}$. We may well assume without loss of generality that $(K_{r+1}) \neq (L_{r+1})$. Then

$$\left[\tilde{f}_{(K_{r+1})}^{(K_{r+1})}, \sum_{i=1}^{r} \xi_{K_i}^{L_i} \gamma_{K_i}^{K_i} + \sum_{i=r+1}^{s} \xi_{K_i}^{L_i} \tilde{f}_{(L_i)}^{(K_i)}\right]$$

does not belong to $\mathcal{H}$, too. As a result, the normalizer of $\mathcal{H}$ is $\mathcal{H}$ itself. It is therefore that $\mathcal{H}$ is a Cartan subalgebra of the heterix algebra. Q.E.D.

APPENDIX F: ROOT VECTORS OF THE HETERIX ALGEBRA

Since the set of all $\tilde{f}_{(L)}^{(K)}$'s is linearly independent and the root for which $\tilde{f}_{(L)}^{(K)}$ is a root vector is distinct from the root for which $\tilde{f}_{(L')}^{(K')}$, where $(K) \neq (K')$ or $(L) \neq (L')$, is a root vector, every root vector which is a finite linear combination of $\tilde{f}_{(L)}^{(K)}$'s only must be a constant multiple of one $\tilde{f}_{(L)}^{(K)}$ only. Hence we only need to show that if a root vector is of the form

$$f = \sum_{P,Q} a_P^Q \gamma_Q^P + \sum_{(P),(Q)} b_{(P)}^{(Q)} \tilde{f}_{(Q)}^{(P)},$$

where a finite but *nonzero* number of a_P^Q's are nonzero, and a finite number of $b_{(P)}^{(Q)}$'s are nonzero as well, then f must be a constant multiple of a f_L^K.

Indeed, let

$$[\gamma_M^M, f] = \lambda_M f \quad \text{and} \quad [\tilde{f}_{(M)}^{(M)}, f] = \lambda_{(M)} f,$$

where each λ_M and each $\lambda_{(M)}$ are complex constants for any integer sequence M. Since

$$[\tilde{f}_{(M)}^{(M)}, f] \in \tilde{F}_\Lambda$$

but $f \notin \tilde{F}_\Lambda$, $\lambda_{(M)} = 0$ for any (M). Now obtain an expression for λ_M starting with the equation

$$[\gamma_M^M, f]s^K = \lambda_M f s^K \tag{F1}$$

as follows. From Eq. (4), we deduce that

$$\sigma_Q^P s^J = \sum_I \left(\delta_Q^J \delta_I^P + \sum_{J_1 J_2 = J} \delta_Q^{J_2} \delta_I^{J_1 P} + \sum_{J_1 J_2 = J} \delta_Q^{J_1} \delta_I^{P J_2} \sum_{J_1 J_2 J_3 = J} \delta_Q^{J_2} \delta_I^{J_1 P J_3} \right) s^I.$$

Therefore,

$$\begin{aligned}
[\sigma_M^M, f]s^K = \sum_{I,P,Q} \Bigg(& \delta_I^M \delta_Q^K \delta_M^P + \sum_{K_1 K_2 = K} \delta_I^M \delta_Q^{K_2} \delta_M^{K_1 P} + \sum_{K_1 K_2 = K} \delta_I^M \delta_Q^{K_1} \delta_M^{P K_2} \\
& + \sum_{K_1 K_2 K_3 = K} \delta_I^M \delta_Q^{K_2} \delta_M^{K_1 P K_3} + \sum_{I_1 I_2 = I} \delta_{I_2}^M \delta_Q^K \delta_I^P + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_2}^M \delta_Q^{K_2} \delta_I^{K_1 P} \\
& + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_2}^M \delta_Q^{K_1} \delta_I^{P K_2} + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 K_3 = K}} \delta_{I_2}^M \delta_Q^{K_2} \delta_I^{K_1 P K_3} + \sum_{I_1 I_2 = I} \delta_{I_1}^M \delta_Q^K \delta_I^P \\
& + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_1}^M \delta_Q^{K_2} \delta_I^{K_1 P} + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_1}^M \delta_Q^{K_1} \delta_I^{P K_2} + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 K_3 = K}} \delta_{I_1}^M \delta_Q^{K_2} \delta_I^{K_1 P K_3} \\
& + \sum_{I_1 I_2 I_3 = I} \delta_{I_2}^M \delta_Q^K \delta_I^P + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 = K}} \delta_{I_2}^M \delta_Q^{K_2} \delta_I^{K_1 P} + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 = K}} \delta_{I_2}^M \delta_Q^{K_1} \delta_I^{P K_2} \\
& + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 K_3 = K}} \delta_{I_2}^M \delta_Q^{K_2} \delta_I^{K_1 P K_3} \Bigg) a_P^Q s^I - \sum_{I,P,Q} \Bigg(\delta_I^P \delta_M^K \delta_Q^M + \sum_{K_1 K_2 = K} \delta_I^P \delta_M^{K_2} \delta_Q^K \\
& + \sum_{K_1 K_2 = K} \delta_I^P \delta_M^{K_1} \delta_Q^K + \sum_{K_1 K_2 K_3 = K} \delta_I^P \delta_M^{K_2} \delta_Q^K + \sum_{I_1 I_2 = I} \delta_{I_2}^P \delta_M^K \delta_{I_1 Q} + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_2}^P \delta_M^{K_2} \delta_{I_1 Q}^K \\
& + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_2}^P \delta_M^{K_1} \delta_{I_1 Q}^K + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 K_3 = K}} \delta_{I_2}^P \delta_M^{K_2} \delta_{I_1 Q}^K + \sum_{I_1 I_2 = I} \delta_{I_1}^P \delta_M^K \delta_{Q I_2} + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_1}^P \delta_M^{K_2} \delta_{Q I_2}^K \\
& + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 = K}} \delta_{I_1}^P \delta_M^{K_1} \delta_{Q I_2}^K + \sum_{\substack{I_1 I_2 = I \\ K_1 K_2 K_3 = K}} \delta_{I_1}^P \delta_M^{K_2} \delta_{Q I_2}^K + \sum_{I_1 I_2 I_3 = I} \delta_{I_2}^P \delta_M^K \delta_{I_1 Q I_3} \\
& + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 = K}} \delta_{I_2}^P \delta_M^{K_2} \delta_{I_1 Q I_3}^K + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 = K}} \delta_{I_2}^P \delta_M^{K_1} \delta_{I_1 Q I_3}^K + \sum_{\substack{I_1 I_2 I_3 = I \\ K_1 K_2 K_3 = K}} \delta_{I_2}^P \delta_M^{K_2} \delta_{I_1 Q I_3}^K \Bigg) a_P^Q s^I.
\end{aligned} \tag{F2}$$

As a result, we can combine Eqs. (F1) and (F2) together to obtain an equation which is too long to be written down here for *any* integer sequences I, K, and M.

Let us find an a_R^S in f such that $R \neq S$, $a_R^S \neq 0$, $a_{R_1}^{S_1} = a_{R_2}^{S_2} = 0$ for all R_1's, S_1's, R_2's and S_2's such that $R_1 R_2 = R$ and $S_1 S_2 = S$, and $a_{R_2}^{S_2} = 0$ for all R_2's and S_2's such that $R_1 R_2 R_3 = R$ and $S_1 S_2 S_3 = S$ for some R_1, R_3, S_1 and S_3. The reader can easily convince himself or herself that such an a_R^S always exists. Let us choose $I = R$ and $K = S$ in Eq. (F2). Then when we combine Eqs. (F1) and (F2), we get

$$\lambda_M=\delta_M^R+\sum_{R_1R_2=R}\delta_M^{R_1}+\sum_{R_1R_2=R}\delta_M^{R_2}+\sum_{R_1R_2R_3=R}\delta_M^{R_2}-\delta_S^M-\sum_{S_1S_2=S}\delta_{S_1}^M-\sum_{S_1S_2=S}\delta_{S_2}^M-\sum_{S_1S_2S_3=S}\delta_{S_2}^M. \tag{F3}$$

Therefore, we obtain after some manipulation that

$$\lambda_M-\sum_{j=1}^{\Lambda}\lambda_{Mj}-\sum_{i=1}^{\Lambda}\lambda_{iM}+\sum_{i,j=1}^{\Lambda}\lambda_{iMj}=\delta_M^R-\delta_S^M.$$

This means

$$[f_R^R,f]\neq 0 \quad \text{and} \quad [f_S^S,f]\neq 0.$$

Thus $f\in F'_\Lambda$. Since we know that every f_L^K is a root vector, and the set of all f_L^K's is linearly independent, $f=f_L^K$ for some K and L. Q.E.D.

APPENDIX G: A BASIS FOR A $\Lambda=1$ QUOTIENT LIE ALGEBRA

We are going to prove that the set of all $\tilde{f}_{(b)}^{(a)}$, g_b^1 and g_1^a where a and b are arbitrary positive integers form a basis for the quotient algebra of all *cosets* γ_b^a and $\tilde{f}_{(b)}^{(a)}$. Indeed, consider the equation

$$\sum_{a,b=1}^{\infty}\alpha_{(a)}^{(b)}\tilde{f}_{(b)}^{(a)}+\sum_{d=1}^{\infty}\alpha^d g_d^1+\sum_{c=2}^{\infty}\alpha_c g_1^c=0, \tag{G1}$$

where only a finite number of the α's are nonzero. Let n be an integer such that for all a's and b's such that $\alpha_{(a)}^{(b)}\neq 0$, we have $n>b$ and for all d's such that $\alpha^d\neq 0$, we have $n>d$ also. Then

$$\left(\sum_{a,b=1}^{\infty}\alpha_{(a)}^{(b)}\tilde{f}_{(b)}^{(a)}+\sum_{d=1}^{\infty}\alpha^d g_d^1+\sum_{c=2}^{\infty}\alpha_c g_1^c\right)\Psi^{(n)}=\sum_{d=1}^{\infty}n\alpha^d\Psi^{(n-d+1)}+\sum_{c=2}^{\infty}n\alpha_c\Psi^{(n+c-1)}.$$

For the rhs of this equation to vanish, we need $\alpha^d=0$ and $\alpha_c=0$ for all c's and d's. Hence Eq. (G1) becomes

$$\sum_{a,b=1}^{\infty}\alpha_{(a)}^{(b)}\tilde{f}_{(b)}^{(a)}=0.$$

Then

$$\sum_{a,b=1}^{\infty}\alpha_{(a)}^{(b)}\tilde{f}_{(b)}^{(a)}\Psi^{(e)}=\sum_{a=1}^{\infty}e\,\alpha_{(a)}^{(e)}\Psi^{(a)}$$

for any positive integer e. Thus $\alpha_{(a)}^{(e)}=0$ also for any positive a and e. Consequently, the set of all $\tilde{f}_{(b)}^{(a)}$'s, g_b^1's and g_1^a's where a and b are arbitrary integers is linearly independent. Q.E.D.

APPENDIX H: CARTAN SUBALGEBRA OF A $\Lambda=1$ QUOTIENT LIE ALGEBRA

This can be seen as follows. From the results of Sec. II, we know that the subspace spanned by all $\tilde{f}_{(a)}^{(a)}$'s and g_1^1 forms an Abelian subalgebra. Moreover, consider a vector v of the form

$$v=\sum_{a,b=1}^{\infty}\alpha_{(a)}^{(b)}\tilde{f}_{(b)}^{(a)}+\sum_{d=2}^{\infty}\alpha^d g_d^1+\sum_{c=2}^{\infty}\alpha_c g_1^c$$

where only a finite number of the α's are not equal to 0, and where $\alpha_{(a)}^{(a)}=0$ for all $a=1,2,\ldots,\infty$. If all the α^d's and α_c's vanish, then choose a particular a_0 such that there exists a b with $\alpha_{(a_0)}^{(b)}\neq 0$. Then

$$[\tilde{f}^{(a_0)}_{(a_0)},v]=\sum_{b=1}^{\infty} a_0\alpha^{(b)}_{(a_0)}\tilde{f}^{(a_0)}_{(b)}-\sum_{a=1}^{\infty} a_0\alpha^{(a_0)}_{(a)}\tilde{f}^{(a)}_{(a_0)}\neq 0.$$

Hence v does not commute with the subspace spanned by all $\tilde{f}^{(a)}_{(a)}$'s and g^1_1. If there exists at least one nonzero α^d or α_c, set m to be the maximum of all a's and b's such that $\alpha^{(b)}_{(a)}\neq 0$. Then $\alpha^{(b')}_{(a')}=0$ if either b' or $a'>m$. Use Eq. (27) to rewrite each g^1_d and g^c_1 such that $\alpha^d\neq 0$ and $\alpha_c\neq 0$ as

$$g^1_d=\tilde{f}^{(1)}_{(d)}+\tilde{f}^{(2)}_{(d+1)}+\cdots+\tilde{f}^{(m)}_{(m+d-1)}+g^{m+1}_{d+m}$$

and

$$g^c_1=\tilde{f}^{(c)}_{(1)}+\tilde{f}^{(c+1)}_{(2)}+\cdots+\tilde{f}^{(m+c-1)}_{(m)}+g^{c+m}_{m+1}.$$

Then

$$v=\sum_{a,b=1}^{\infty}\alpha'^{(b)}_{(a)}\tilde{f}^{(a)}_{(b)}+\sum_{d=2}^{\infty}\alpha^d g^{m+1}_{d+m}+\sum_{c=2}^{\infty}\alpha_c g^{c+m}_{m+1},$$

where

$$\alpha'^{(b)}_{(a)}=\begin{cases}\alpha^{(b)}_{(a)}+\alpha_{a-b+1} & \text{if } a>b,\\ \alpha^{(b)}_{(a)}+\alpha^{b-a+1} & \text{if } b>a.\end{cases}$$

It is possible for $\alpha'^{(b)}_{(a)}\neq 0$ only if $b\leqslant m$ or $a\leqslant m$. If $\alpha'^{(b_0)}_{(a_0)}\neq 0$ for a particular pair of numbers b_0 and a_0, then

$$[\tilde{f}^{(b_0)}_{(b_0)},v]\neq 0 \quad \text{if } b_0\leqslant m \quad \text{or} \quad [\tilde{f}^{(a_0)}_{(a_0)},v]\neq 0 \quad \text{if } a_0\leqslant m$$

because

$$\left[\tilde{f}^{(b_0)}_{(b_0)},\sum_{d=2}^{\infty}\alpha^d g^{m+1}_{d+m}\right]=0 \quad \text{and} \quad \left[\tilde{f}^{(b_0)}_{(b_0)},\sum_{c=2}^{\infty}\alpha_c g^{c+m}_{m+1}\right]=0$$

if $b_0\leqslant m$, and

$$\left[\tilde{f}^{(a_0)}_{(a_0)},\sum_{d=2}^{\infty}\alpha^d g^{m+1}_{d+m}\right]=0 \quad \text{and} \quad \left[\tilde{f}^{(a_0)}_{(a_0)},\sum_{c=2}^{\infty}\alpha_c g^{c+m}_{m+1}\right]=0$$

if $a_0\leqslant m$. If all $\alpha'^{(b)}_{(a)}$'s vanish, then

$$v=\sum_{d=2}^{\infty}\alpha^d g^{m+1}_{d+m}+\sum_{c=2}^{\infty}\alpha_c g^{c+m}_{m+1}$$

and so

$$[\tilde{f}^{(m+1)}_{(m+1)},v]=\sum_{d=2}^{\infty}(m+1)\alpha^d\tilde{f}^{(m+1)}_{(d+m)}-\sum_{c=2}^{\infty}(m+1)\alpha_c\tilde{f}^{(c+m)}_{(m+1)}\neq 0.$$

Q.E.D.

[1] S. G. Rajeev, Int. J. Mod. Phys. A **9**, 5583 (1994).
[2] S. G. Rajeev, Phys. Lett. B **209**, 53 (1988); Syracuse 1989, Proceedings, 11th Annual Montreal–Rochester–Syracuse–Toronto Meeting, p. 78; Phys. Rev. D **42**, 2779 (1990); **44**, 1836 (1991); G. Ferretti and S. G. Rajeev, Phys. Lett. B **244**, 265 (1990).
[3] C. B. Thorn, Phys. Rev. D **20**, 1435 (1979).
[4] O. Bergman and C. B. Thorn, Phys. Rev. D **52**, 5980 (1995); Phys. Rev. Lett. **76**, 2214 (1995).
[5] T. Banks, W. Fischler, S. H. Shenkar, and L. Susskind, Phys. Rev. D **55**, 5112 (1997); T. Banks, hep-th/9710231; D. Bigatti and L. Susskind, hep-th/9712072.
[6] T. Eguchi and H. Kawai, Phys. Rev. Lett. **48**, 1063 (1982).
[7] G. 't Hooft, Nucl. Phys. B **72**, 461 (1974); E. Witten, *ibid.* **160**, 57 (1979).
[8] S. G. Rajeev and O. T. Turgut, J. Math. Phys. **37**, 637 (1996).
[9] V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd ed. (Cambridge U. P., Cambridge, 1990), p. 112. gl_∞ is also mentioned by, for example, V. G. Kac and D. H. Peterson, in *Lectures on the Infinite Wedge-Representation and the MKP Hierarchy*, Système dynamiques nonlińeaires: intégrabilité et comportement qualitatif, Sem. Math. Sup. 102 (Press Univ. Montréal, Montreal, Quebec, 1986).
[10] C.-W. H. Lee and S. G. Rajeev, Phys. Rev. Lett. **80**, 2285 (1998).
[11] C.-W. H. Lee and S. G. Rajeev, hep-th/9712090 to be published in Nucl. Phys. B.
[12] J. E. Humphreys, *Introduction to Lie Algebras And Representation Theory* (Springer-Verlag, New York, 1972).
[13] P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).
[14] C.-W. H. Lee, Ph.D. thesis, in preparation.
[15] C.-W. H. Lee and S. G. Rajeev, to be published.
[16] S. Dalley and I. R. Klebanov, Phys. Rev. D **47**, 2517 (1993).
[17] F. Antonuccio and S. Dalley, Nucl. Phys. B **461**, 275 (1996).
[18] A. Hashimoto and I. R. Klebanov, Mod. Phys. Lett. A **10**, 2639 (1995).
[19] D. Kutasov, Nucl. Phys. B **414**, 33 (1994).
[20] P.-A. Bares, G. Blatter, and M. Ogata, Phys. Rev. B **44**, 130 (1991).
[21] C. Gómez, M. Ruiz-Altaba, and G. Sierra, *Quantum Groups in Two-dimensional Physics* (Cambridge U. P., Cambridge, 1996).
[22] L. Onsager, Phys. Rev. **65**, 117 (1944); E. Fradkin and L. Susskind, Phys. Rev. D **17**, 2637 (1978).
[23] J. B. Kogut, Rev. Mod. Phys. **51**, 659 (1979).
[24] L. Dolan and M. Grady, Phys. Rev. D **25**, 1587 (1982).
[25] B. Davies, J. Math. Phys. **32**, 2945 (1991).
[26] A. Honecker, Ph.D. thesis, hep-th/9503104.
[27] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic, London, 1982).
[28] H. Au-Yang, B. M. McCoy, J. H. H. Perk, S. Tang, and M. L. Yan, Phys. Lett. A **123**, 219 (1987); H. Au-Yang, R. J. Baxter, and J. H. H. Perk, *ibid.* **128**, 138 (1988); Ph. Christe and M. Henkel, *Introduction to Conformal Invariance and its Applications to Critical Phenomena*, Lecture Notes in Physics Mono. 16 (Springer-Verlag, New York, 1993).

Operator formulation of Wigner's *R*-matrix theories for the Schrödinger and Dirac equations

Radosław Szmytkowski
Faculty of Applied Physics and Mathematics, Technical University of Gdańsk, ul. Narutowicza 11/12, PL 80-952 Gdańsk, Poland

(Received 21 January 1998; accepted for publication 1 May 1998)

The *R*-matrix theories for the Schrödinger and Dirac equations are formulated in the language of integral operators. In the nonrelativistic theory the central role is played by an integral operator $\hat{\mathscr{R}}_{\mathsf{b}}(E)$ relating function values to normal derivatives on a surface $\mathscr{S}$ of a closed volume $\mathscr{V}$, inside which the function satisfies the Schrödinger equation at energy E. In the relativistic theory, the same role is played by two integral operators, $\hat{\mathscr{R}}_{\mathsf{b}}^{(+)}(E)$ and $\hat{\mathscr{R}}_{\mathsf{b}}^{(-)}(E)$, linking on the surface $\mathscr{S}$ values of upper and lower components of spinor wave functions satisfying in the volume $\mathscr{V}$ the Dirac equation at energy E. Systematic procedures for constructing the operators $\hat{\mathscr{R}}_{\mathsf{b}}(E)$ and $\hat{\mathscr{R}}_{\mathsf{b}}^{(\pm)}(E)$, generalizing the methods due to Kapur and Peierls and to Wigner, are presented. © *1998 American Institute of Physics.* [S0022-2488(98)00210-2]

I. INTRODUCTION

The *R*-matrix theory originated in the late 1930s[1] within the framework of the nuclear reactions theory. At the early stages of the evolvement of the method, the most significant contributions were made by Wigner[2–6] and since then the method has been invariably associated with his name. In the field of nuclear physics the theory achieved the most spectacular successes in the 1950s and the early 1960s.[7,8] Later, the interest of the nuclear physics community in the *R*-matrix theory diminished but at nearly the same time it was recognized that the method was ideally suited for applications in atomic and molecular physics. Over the last 30 years the *R*-matrix theory has evolved into one of the most popular and accurate methods used in theoretical studies of such atomic and molecular phenomena as electron collisions with atoms, molecules and ions, atomic and molecular photoionization and spectra of Rydberg systems.[9–11]

To summarize the most important features of the *R*-matrix theory, let us consider scattering of a particle of energy E in a field of a real local potential $V(\mathbf{r})$. In the *R*-matrix approach one divides the physical space into two regions separated by an imagined closed surface $\mathscr{S}$. It is assumed that in the external region the potential $V(\mathbf{r})$ is so simple that a general solution of a time-independent wave equation governing the dynamics of the particle in that region is known either analytically or at least in a numerical form. To the contrary, in the internal region $\mathscr{V}$ the potential $V(\mathbf{r})$ may be much more complicated. It was pointed out by Kapur and Peierls[1] and by Wigner and co-workers[2–6] that from the point of view of the external region, the domain $\mathscr{V}$ may be treated as a ''black box'' and the collision matrix $\mathsf{U}(E)$ may still be found provided one knows a natural boundary condition obeyed on the surface $\mathscr{S}$ by all solutions of the wave equation considered. It was shown by Teichmann and Wigner[6] that if an orthonormal set of basis functions $\{\Phi_i\}$ spanning the surface $\mathscr{S}$ is given, a general matrix form of such a natural boundary condition for the Schrödinger equation is

$$\mathsf{P}(E)=\mathsf{R}_{\mathsf{b}}(E)[\mathsf{D}(E)-\mathsf{b}\mathsf{P}(E)]. \tag{1}$$

In Eq. (1) $\mathsf{P}(E)$ and $\mathsf{D}(E)$ are column matrices containing, respectively, projections of the wave function Ψ and its normal derivative $\nabla_n\Psi$ at the surface $\mathscr{S}$ onto the surface basis functions $\{\Phi_i\}$, b is an arbitrary Hermitian matrix which may be chosen at one's will. The goal of the theory is to find the energy-dependent matrix $\mathsf{R}_{\mathsf{b}}(E)$. Once this goal is achieved, the collision matrix $\mathsf{U}(E)$ may be expressed in terms of the *R*-matrix $\mathsf{R}_{\mathsf{b}}(E)$.[6–8]

 © 1998 American Institute of Physics

Some years ago Nesbet[12,13] pointed out that since the Schrödinger equation is not a matrix equation but a partial differential equation, it is more fundamental to formulate the theory in terms of integral operators rather than in terms of matrices. According to Nesbet, in the nonrelativistic theory the central role is played not by the matrix $\mathsf{R}_{\mathfrak{b}}(E)$ itself but by an integral operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ such that

$$\Psi(E,\boldsymbol{\rho})=\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)[\nabla_n\Psi(E,\boldsymbol{\rho})-\hat{\mathfrak{b}}\Psi(E,\boldsymbol{\rho})]. \tag{2}$$

Here $\boldsymbol{\rho}$ is an arbitrary point on the surface $\mathscr{S}$ and $\hat{\mathfrak{b}}$ is a Hermitian integral operator acting on functions defined on the surface $\mathscr{S}$ which may be chosen at our will. The operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ is independent of the choice of a particular basis set spanning the surface $\mathscr{S}$. However, once such a basis set is chosen, matrix elements of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ between basis functions form the matrix $\mathsf{R}_{\mathfrak{b}}(E)$ in that particular representation.

In this work we present the thorough discussion of the operator formulation of the version of the *R*-matrix theory due to Kapur and Peierls[1,14] and to Wigner.[2,3,4,6] We consider both the nonrelativistic case when the wave equation of the particle is the Schrödinger equation as well as the relativistic case of a particle whose dynamics are governed by the Dirac equation. There are two main motivations for undertaking this work. First, the operator formulation of the *R*-matrix theory is more general and elegant than the matrix approach. It is more transparent and therefore better suited for examining mathematical aspects of the method. Such a formulation is also in accord with the permanent trend in theoretical physics to make a mathematical formalism underlying any theoretical method as general and (to a reasonable extent) as rigorous as possible. Even if providing a general mathematical formulation of a particular method does not result in an immediate breakthrough in practical applications, usually it leads to better understanding of the essence of the method and may stimulate its further development. The second motivation is of a more practical nature. It appears that the operator formulation of the *R*-matrix theory relatively simply yields results which are valid for an arbitrarily shaped volume $\mathscr{V}$. This is of particular significance in the case of the *R*-matrix theory for the Dirac equation since previous attempts to develop such a theory based on a matrix approach[15–18] led to difficulties: the resulting theories either contained an error (this was shown by Szmytkowski and Hinze[17,18]) or were restricted solely to a spherically symmetric volume $\mathscr{V}$.[17,18]

The plan of the paper is as follows. In Sec. II the notation used subsequently is introduced. Section III deals with the nonrelativistic *R*-matrix theory. In Sec. III A it is shown how the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ arises in the theory. A method of constructing the integral kernel of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ is discussed in Sec. III B. The operator formulation of the *R*-matrix theory for the Dirac equation is the subject of Sec. IV. In Sec. IV A two operators $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ and $\hat{\mathscr{R}}^{(-)}_{\hat{\mathfrak{b}}}(E)$ are introduced. The operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ is, to some extent, an analog of the nonrelativistic operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$. The nonrelativistic limit of the operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ is discussed in Sec. IV B. In Sec. IV C the explicit forms of the operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$ are found. Finally, conclusions are presented in Sec. V.

II. PRELIMINARIES

Before we proceed to the matter, we acquaint the reader with the notation used throughout the paper. In the following, we shall restrict our considerations to a finite volume $\mathscr{V}$ bounded by a surface $\mathscr{S}$. Everywhere in the paper $\mathbf{r}$ stands for a position vector of a point in the volume $\mathscr{V}$. If the point $\mathbf{r}$ lies on the boundary $\mathscr{S}$, we use the symbol $\boldsymbol{\rho}$ instead of $\mathbf{r}$. Moreover, $\mathbf{n}(\boldsymbol{\rho})$ means an outward unit vector normal to the surface $\mathscr{S}$ at the point $\boldsymbol{\rho}$. Since this should not cause any misunderstanding, we shall use the same symbols to denote inner products of scalar or spinor functions. Thus, if $\Phi(\mathbf{r})$ and $\Phi'(\mathbf{r})$ are two scalar functions (such functions appear in the nonrelativistic theory), their volume and surface scalar products are

$$\langle\Phi|\Phi'\rangle\equiv\int_{\mathscr{V}}d^3\mathbf{r}\;\Phi^*(\mathbf{r})\Phi'(\mathbf{r}),\quad(\Phi|\Phi')\equiv\int_{\mathscr{S}}d^2\boldsymbol{\rho}\;\Phi^*(\boldsymbol{\rho})\Phi'(\boldsymbol{\rho}), \tag{3}$$

respectively, where the asterisk denotes the complex conjugation. If, in turn, $\Phi(\mathbf{r})$ and $\Phi'(\mathbf{r})$ are two spinor functions (encountered in the Dirac theory), their volume and surface scalar products are

$$\langle\Phi|\Phi'\rangle\equiv\int_{\mathscr{V}} d^3\mathbf{r}\ \Phi^\dagger(\mathbf{r})\Phi'(\mathbf{r}),\quad (\Phi|\Phi')\equiv\int_{\mathscr{S}} d^2\boldsymbol{\rho}\ \Phi^\dagger(\boldsymbol{\rho})\Phi'(\boldsymbol{\rho}),\tag{4}$$

respectively, where the dagger denotes the matrix Hermitian conjugation. In Eqs. (3) and (4) $d^3\mathbf{r}$ is an infinitesimal volume element around the point $\mathbf{r}$ and $d^2\boldsymbol{\rho}$ is an infinitesimal scalar surface element around the point $\boldsymbol{\rho}$.

The following notation concerning the operators will be used: if $\hat{\mathscr{O}}$ is an integral operator on the surface $\mathscr{S}$ and if $\mathscr{O}(\boldsymbol{\rho},\boldsymbol{\rho}')$ is its integral kernel, then for any well behaving scalar or spinor function $\Phi(\boldsymbol{\rho})$ we have

$$\hat{\mathscr{O}}\Phi(\boldsymbol{\rho})=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \mathscr{O}(\boldsymbol{\rho},\boldsymbol{\rho}')\Phi(\boldsymbol{\rho}'),\tag{5}$$

$$\Phi^\dagger(\boldsymbol{\rho})\hat{\mathscr{O}}=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \Phi^\dagger(\boldsymbol{\rho}')\mathscr{O}(\boldsymbol{\rho}',\boldsymbol{\rho})\tag{6}$$

[here the dagger denotes either the complex or the matrix Hermitian conjugation according to the type of the function $\Phi(\boldsymbol{\rho})$ used].

We also define a family of delta functions, $\delta^{(1)}_{\mathscr{S}}(\mathbf{r})$, $\delta^{(2)}(\boldsymbol{\rho}-\boldsymbol{\rho}')$ and $\delta^{(3)}(\mathbf{r}-\mathbf{r}')$, such that for any reasonable function $\Phi(\mathbf{r})$, either scalar or spinor, one has

$$\int_{\mathscr{V}} d^3\mathbf{r}\ \delta^{(1)}_{\mathscr{S}}(\mathbf{r})\Phi(\mathbf{r})=\int_{\mathscr{S}} d^2\boldsymbol{\rho}\ \Phi(\boldsymbol{\rho}),\tag{7}$$

$$\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \delta^{(2)}(\boldsymbol{\rho}'-\boldsymbol{\rho})\Phi(\boldsymbol{\rho}')=\Phi(\boldsymbol{\rho}),\tag{8}$$

$$\int_{\mathscr{V}} d^3\mathbf{r}'\ \delta^{(3)}(\mathbf{r}'-\mathbf{r})\Phi(\mathbf{r}')=\Phi(\mathbf{r}),\quad \int_{\mathscr{V}} d^3\mathbf{r}'\ \delta^{(3)}(\mathbf{r}'-\boldsymbol{\rho})\Phi(\mathbf{r}')=\Phi(\boldsymbol{\rho}).\tag{9}$$

It may be inferred from Eqs. (7), (8) and the second of Eqs. (9) that, formally,

$$\delta^{(3)}(\mathbf{r}'-\boldsymbol{\rho})=\delta^{(2)}(\boldsymbol{\rho}'-\boldsymbol{\rho})\,\delta^{(1)}_{\mathscr{S}}(\mathbf{r}'),\quad \delta^{(1)}_{\mathscr{S}}(\mathbf{r})=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \delta^{(3)}(\boldsymbol{\rho}'-\mathbf{r}).\tag{10}$$

III. *R*-MATRIX THEORY FOR THE SCHRÖDINGER EQUATION

A. The operator $\hat{\mathscr{R}}_{\mathfrak{b}}(E)$

In this section we shall be concerned with a nonrelativistic pointlike and spinless particle of mass m moving with energy E in the field of force described by a real, local, in general noncentral potential $V(\mathbf{r})$. We restrict our considerations to the volume $\mathscr{V}$ bounded by the surface $\mathscr{S}$. [It is to be emphasized that the particle's movement is not constrained to the volume $\mathscr{V}$. The particle may leave and enter this volume across the surface $\mathscr{S}$ freely but we focus our attention on its movement when it penetrates the domain $\mathscr{V}$.] The Schrödinger equation describing the dynamics of the particle is

$$[\hat{H}-E]\Psi(E,\mathbf{r})=0,\tag{11}$$

where the Hamiltonian $\hat{H}$ is

$$\hat{H}=-\frac{\hbar^2}{2m}\,\boldsymbol{\nabla}^2+V(\mathbf{r}). \tag{12}$$

It will be assumed hereafter that energy E is real.

Consider now two wave functions $\Psi(E,\mathbf{r})$ and $\Psi'(E,\mathbf{r})$ which are solutions to the wave equation (11) corresponding to the same energy E. Premultiplying the equation for $\Psi(E,\mathbf{r})$ by $\Psi'^*(E,\mathbf{r})$ and the complex conjugate of the equation for $\Psi'(E,\mathbf{r})$ by $\Psi(E,\mathbf{r})$, subtracting, integrating the result over the volume $\mathscr{V}$, and making use of the Green integration theorem, we obtain

$$\langle\hat{H}\Psi'|\Psi\rangle-\langle\Psi'|\hat{H}\Psi\rangle=\frac{\hbar^2}{2m}\,(\Psi'|\nabla_n\Psi)-\frac{\hbar^2}{2m}\,(\nabla_n\Psi'|\Psi), \tag{13}$$

where $\nabla_n\Psi(E,\boldsymbol{\rho})$ denotes the outward normal derivative of the function $\Psi(E,\mathbf{r})$ at the surface point $\boldsymbol{\rho}$. By virtue of Eq. (11) and the reality of E, the left-hand side of Eq. (13) is zero. Consequently, we have

$$(\Psi'|\nabla_n\Psi)=(\nabla_n\Psi'|\Psi). \tag{14}$$

Equation (14) may be formally interpreted that the normal derivative operator ∇_n, when acting on functions satisfying in the interior of the volume $\mathscr{V}$ the Schrödinger equation (11) at the fixed energy E, is Hermitian with respect to the surface scalar product $(|)$.

We shall introduce now an auxiliary energy-dependent linear Hermitian integral operator $\hat{\mathscr{B}}(E)$ defined on $L^2_{(|)}(\mathscr{S})$ such that

$$\nabla_n\Psi(E,\boldsymbol{\rho})=\hat{\mathscr{B}}(E)\Psi(E,\boldsymbol{\rho}) \tag{15}$$

for any solution to Eq. (11) at energy E. The operator $\hat{\mathscr{B}}(E)$ is represented by an integral kernel $\mathscr{B}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ in terms of which Eq. (15) may be rewritten in the form

$$\nabla_n\Psi(E,\boldsymbol{\rho})=\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\;\mathscr{B}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\Psi(E,\boldsymbol{\rho}'). \tag{16}$$

The existence of the operator $\hat{\mathscr{B}}(E)$ has been proved by Szmytkowski[19] who has demonstrated that the kernel $\mathscr{B}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ has the representation

$$\mathscr{B}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_i\;\Psi_i(E,\boldsymbol{\rho})b_i(E)\Psi_i^*(E,\boldsymbol{\rho}'), \tag{17}$$

where $\{\Psi_i(E,\mathbf{r})\}$ are those solutions to the Schrödinger equation (11) at energy E which have constant normal derivatives on $\mathscr{S}$,

$$\nabla_n\Psi_i(E,\boldsymbol{\rho})=b_i(E)\Psi_i(E,\boldsymbol{\rho}) \tag{18}$$

and are normalized according to $(\Psi_i|\Psi_i)=1$ (notice that the surface, *not* volume, scalar product has been used for the normalization purposes). It has been shown[19] that the numbers $\{b_i(E)\}$ are real and that the surface functions $\{\Psi_i(E,\boldsymbol{\rho})\}$ are such that $(\Psi_i|\Psi_j)=\delta_{ij}$.

It should be emphasized that the operators ∇_n and $\hat{\mathscr{B}}(E)$ are not identical. If $\Phi(\mathbf{r})$ is any sufficiently regular function defined in the volume $\mathscr{V}$, in general one has

$$\nabla_n\Phi(\boldsymbol{\rho})\neq\hat{\mathscr{B}}(E)\Phi(\boldsymbol{\rho}) \tag{19}$$

unless $\Phi(\mathbf{r})$ obeys in $\mathscr{V}$ the Schrödinger equation (11) at energy E.

We are now prepared to introduce an integral operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ defined on $L^2_{(|)}(\mathscr{S})$, with the integral kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$, a matrix representation of which plays the central role in the R-matrix theory. Let $\hat{\mathfrak{b}}$ be some linear Hermitian integral operator defined on $L^2_{(|)}(\mathscr{S})$, with the integral kernel $\mathfrak{b}(\boldsymbol{\rho},\boldsymbol{\rho}')$. The operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ is then defined as an operator reciprocal to the operator $\hat{\mathscr{B}}(E)-\hat{\mathfrak{b}}$ in the sense of

$$\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)[\hat{\mathscr{B}}(E)-\hat{\mathfrak{b}}]\hat{\mathscr{P}}(E)=\hat{\mathscr{P}}(E), \tag{20}$$

where $\hat{\mathscr{P}}(E)$ is the projector on the manifold of surface parts of solutions to the Schrödinger equation (11) at energy E with the kernel

$$\mathscr{P}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_i \Psi_i(E,\boldsymbol{\rho})\Psi_i^*(E,\boldsymbol{\rho}'). \tag{21}$$

The functions $\{\Psi_i(E,\mathbf{r})\}$ have been defined below Eq. (17). In principle, the operator $\hat{\mathfrak{b}}$ used in the definition (20) may be arbitrary and the only restriction we impose on $\hat{\mathfrak{b}}$ is that the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ should exist. In terms of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ the surface relation (15) may be rewritten in the form

$$\Psi(E,\boldsymbol{\rho})=\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)[\nabla_n-\hat{\mathfrak{b}}]\Psi(E,\boldsymbol{\rho}) \tag{22}$$

or, equivalently, in terms of the kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$, in the form

$$\Psi(E,\boldsymbol{\rho})=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')[\nabla_n'-\hat{\mathfrak{b}}]\Psi(E,\boldsymbol{\rho}'). \tag{23}$$

To relate the operator and matrix formulations of the theory, we assume that some set of surface functions $\{\Phi_i(\boldsymbol{\rho})\}\in L^2_{(|)}(\mathscr{S})$ orthonormal under the surface scalar product $(|)$ and spanning the surface $\mathscr{S}$ is given. In this basis the kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ has a bilinear expansion

$$\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_{i,j} \Phi_i(\boldsymbol{\rho})(\Phi_i|\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}\Phi_j)\Phi_j^*(\boldsymbol{\rho}'), \tag{24}$$

where the energy-dependent expansion coefficients

$$(\Phi_i|\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}\Phi_j)\equiv\int_{\mathscr{S}} d^2\boldsymbol{\rho}\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \Phi_i^*(\boldsymbol{\rho})\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\Phi_j(\boldsymbol{\rho}'), \tag{25}$$

matrix elements of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ between the basis functions $\{\Phi_i(\boldsymbol{\rho})\}$, form an R-matrix $\mathsf{R}_{\mathsf{b}}(E)$ in that particular representation. The matrix representation of Eq. (22), obtained by projecting the latter from the left on the basis functions $\{\Phi_i(\boldsymbol{\rho})\}$ is

$$\mathsf{P}(E)=\mathsf{R}_{\mathsf{b}}(E)[\mathsf{D}(E)-\mathsf{b}\mathsf{P}(E)], \tag{26}$$

where $\mathsf{P}(E)$ and $\mathsf{D}(E)$ are column matrices with elements $\{P_i(E)=(\Phi_i|\Psi)\}$ and $\{D_i(E)=(\Phi_i|\nabla_n\Psi)\}$, respectively, and b is a square matrix with elements $\{b_{ij}=(\Phi_i|\hat{\mathfrak{b}}\Phi_j)\}$. The relation (26) is identical with the relation (1) found by Teichmann and Wigner.[6] We have thus succeeded to formulate the R-matrix theory for the Schrödinger equation in terms of integral operators and to relate this formulation to the standard matrix approach to the theory. It remains to find the explicit form of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$. In Sec. III B we shall construct its kernel following the method due to Kapur and Peierls and to Wigner.

B. Construction of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$

In this section it will be our goal to find the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ defined by Eq. (20) by constructing its integral kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$. We shall follow, with improvements whenever they are necessary, an approach due to Kapur and Peierls[1] and to Wigner and co-workers.[2–6] For further purposes we shall need a set of functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ defined as solutions of the boundary-value problem consisting of the Schrödinger differential equation

$$[\hat{H}-E_{\hat{\mathfrak{b}}k}]\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})=0 \tag{27}$$

and the boundary condition

$$\nabla_n \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = \hat{\mathfrak{b}} \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \tag{28}$$

with $\{E_{\hat{\mathfrak{b}}k}\}$ being eigenvalues. Because of the Hermicity of the operator $\hat{\mathfrak{b}}$ under the surface scalar product $(|)$, the boundary-value problem (27) and (28) is self-adjoint with respect to the volume scalar product $\langle|\rangle$ and therefore the eigenvalues $\{E_{\hat{\mathfrak{b}}k}\}$ are real and eigenfunctions belonging to different eigenvalues are orthogonal in the sense of

$$\langle \Psi_{\hat{\mathfrak{b}}k} | \Psi_{\hat{\mathfrak{b}}k'} \rangle = 0 \quad (E_{\hat{\mathfrak{b}}k} \neq E_{\hat{\mathfrak{b}}k'}). \tag{29}$$

In what follows, we shall assume that the functions belonging to degenerate eigenvalues (if there are any) have been also orthogonalized and that the normalization factors of the eigenfunctions have been chosen so that

$$\langle \Psi_{\hat{\mathfrak{b}}k} | \Psi_{\hat{\mathfrak{b}}k'} \rangle = \delta_{kk'}\,. \tag{30}$$

We assume also that the boundary condition operator $\hat{\mathfrak{b}}$ is such that the set of eigenfunctions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ is complete in the volume $\mathscr{V}$. The corresponding closure relation is then

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r}') = \delta^{(3)}(\mathbf{r}-\mathbf{r}') \quad (\mathbf{r},\mathbf{r}' \in \mathscr{V} \setminus \mathscr{S}). \tag{31}$$

[Notice that, because the functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ are constrained to obey the boundary condition (28), it is not known *a priori* whether the relation (31) holds also for the points $\mathbf{r}$ and $\mathbf{r}'$ lying on the surface $\mathscr{S}$ or not. We shall study this problem later in this section.]

By virtue of the assumed completeness of the set $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$, one may use these functions for expanding the wave function $\Psi(E,\mathbf{r})$ satisfying the Schrödinger equation (11). Denoting the expansion by $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})$, one has

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r}) \equiv \sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \langle \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle \tag{32}$$

or, explicitly,

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r}) \equiv \int_{\mathscr{V}} d^3\mathbf{r}' \left\{ \sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r}') \right\} \Psi(E,\mathbf{r}'). \tag{33}$$

The expansion coefficients $\langle \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle$ in Eq. (32) may be found in the standard way: One premultiplies Eq. (11) by $\Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r})$, postmultiplies the complex conjugate of Eq. (27) by $\Psi(E,\mathbf{r})$, subtracts and integrates the result over the volume $\mathscr{V}$ obtaining

$$\langle \hat{H} \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle - \langle \Psi_{\hat{\mathfrak{b}}k} | \hat{H} \Psi \rangle = [E_{\hat{\mathfrak{b}}k} - E] \langle \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle \tag{34}$$

which, after utilizing the explicit form (12) of the Hamiltonian $\hat{H}$ and the assumed Hermicity of the potential $V(\mathbf{r})$, becomes

$$\frac{\hbar^2}{2m} \langle \Psi_{\hat{\mathfrak{b}}k} | \boldsymbol{\nabla}^2 \Psi \rangle - \frac{\hbar^2}{2m} \langle \boldsymbol{\nabla}^2 \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle = [E_{\hat{\mathfrak{b}}k} - E] \langle \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle. \tag{35}$$

The left-hand side of Eq. (35) may be simplified by using the Green integration theorem which states that

$$\langle \Psi_{\hat{\mathfrak{b}}k} | \boldsymbol{\nabla}^2 \Psi \rangle - \langle \boldsymbol{\nabla}^2 \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle = (\Psi_{\hat{\mathfrak{b}}k} | \nabla_n \Psi) - (\nabla_n \Psi_{\hat{\mathfrak{b}}k} | \Psi). \tag{36}$$

Transforming the second surface integral on the right-hand side of Eq. (36) with the aid of the boundary condition (28) and substituting the resulting equation to Eq. (35) one arrives at

$$\langle \Psi_{\hat{\mathfrak{b}}k} | \Psi \rangle = \frac{\hbar^2}{2m} \frac{(\Psi_{\hat{\mathfrak{b}}k} | [\nabla_n - \hat{\mathfrak{b}}] \Psi)}{E_{\hat{\mathfrak{b}}k} - E}, \tag{37}$$

hence, it follows that

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \mathbf{r}) = \frac{\hbar^2}{2m} \sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \frac{(\Psi_{\hat{\mathfrak{b}}k} | [\nabla_n - \hat{\mathfrak{b}}] \Psi)}{E_{\hat{\mathfrak{b}}k} - E}. \tag{38}$$

For the sake of convenience, we shall rewrite the latter equation in the explicit form

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \mathbf{r}) = \int_{\mathscr{S}} d^2\boldsymbol{\rho}' \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E} [\nabla'_n - \hat{\mathfrak{b}}] \Psi(E, \boldsymbol{\rho}'). \tag{39}$$

Equation (39) holds for an arbitrary point $\mathbf{r}$ in the volume $\mathscr{V}$ and in particular it is valid also for points on the surface $\mathscr{S}$, i.e., one has

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}) = \int_{\mathscr{S}} d^2\boldsymbol{\rho}' \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E} [\nabla'_n - \hat{\mathfrak{b}}] \Psi(E, \boldsymbol{\rho}'). \tag{40}$$

To proceed further, we shall define an integral operator $\hat{\overline{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E)$ relating the surface functions $\Psi(E, \boldsymbol{\rho})$ and $\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho})$ in the following way:

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}) = \hat{\overline{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E) [\nabla_n - \hat{\mathfrak{b}}] \Psi(E, \boldsymbol{\rho}). \tag{41}$$

The operator $\hat{\overline{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E)$ is represented by its integral kernel $\overline{\mathscr{R}}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}, \boldsymbol{\rho}')$ which, from Eqs. (40) and (41), is found to be

$$\overline{\mathscr{R}}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}, \boldsymbol{\rho}') = \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E}. \tag{42}$$

Consider now the problem of convergence of the expansion $\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \mathbf{r})$. If the point $\mathbf{r}$ lies in the interior of the volume $\mathscr{V}$, by virtue of the closure relation (31) and Eq. (33) we have

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E, \mathbf{r}) \equiv \int_{\mathscr{V}} d^3\mathbf{r}' \left\{ \sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r}') \right\} \Psi(E, \mathbf{r}') = \Psi(E, \mathbf{r}) \quad (\mathbf{r} \in \mathscr{V} \setminus \mathscr{S}). \tag{43}$$

On utilizing this result in Eq. (39) we obtain

$$\Psi(E, \mathbf{r}) = \int_{\mathscr{S}} d^2\boldsymbol{\rho}' \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E} [\nabla'_n - \hat{\mathfrak{b}}] \Psi(E, \boldsymbol{\rho}') \quad (\mathbf{r} \in \mathscr{V} \setminus \mathscr{S}). \tag{44}$$

The relation (44) is valid for points $\mathbf{r}$ lying arbitrarily close to the surface $\mathscr{S}$. The function $\Psi(E, \mathbf{r})$ obeys the second-order differential equation (11) and therefore it is continuous across the surface $\mathscr{S}$. Hence, we must have

$$\Psi(E, \boldsymbol{\rho}) = \lim_{\mathbf{r} \to \boldsymbol{\rho}} \int_{\mathscr{S}} d^2\boldsymbol{\rho}' \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E} [\nabla'_n - \hat{\mathfrak{b}}] \Psi(E, \boldsymbol{\rho}') \quad (\mathbf{r} \in \mathscr{V} \setminus \mathscr{S}) \tag{45}$$

and comparison of Eqs. (22) and (45) yields the following representation of the kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}, \boldsymbol{\rho}')$:

$$\mathscr{R}_{\hat{\mathfrak{b}}}(E, \boldsymbol{\rho}, \boldsymbol{\rho}') = \lim_{\mathbf{r} \to \boldsymbol{\rho}} \left\{ \frac{\hbar^2}{2m} \sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}) \Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k} - E} \right\}. \tag{46}$$

The question which arises now is: Can the operations $\underline{\lim}_{\mathbf{r}\to\boldsymbol{\rho}}$ and Σ_k be interchanged in Eq. (46)? Or, in other words, are the kernels $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ and $\overline{\mathscr{R}}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ identical? We shall show now that in the nonrelativistic case discussed in the present section the answer is positive. To this end, we observe that the question posed is closely related to the following problem: How does the closure relation for the functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ look whenever any of the points $\mathbf{r},\mathbf{r}'$ lies on the surface $\mathscr{S}$? On the ground of the analysis of the form of Eq. (31) and the first of Eqs. (10) we assume that the following relations hold:

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r}') = \mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\,\delta^{(1)}_{\mathscr{S}}(\mathbf{r}') \quad (\mathbf{r}'\in\mathscr{V}), \tag{47}$$

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}') = \mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\,\delta^{(1)}_{\mathscr{S}}(\mathbf{r}) \quad (\mathbf{r}\in\mathscr{V}), \tag{48}$$

where the kernel $\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')$ defining the surface integral operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ is to be found. It stems from the definitions (33) and (47) that the kernel $\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')$ relates the surface functions $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})$ and $\Psi(E,\boldsymbol{\rho})$ according to

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho}) = \int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\Psi(E,\boldsymbol{\rho}'). \tag{49}$$

Equivalently, in an abstract form Eq. (49) reads

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho}) = \hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi(E,\boldsymbol{\rho}). \tag{50}$$

To determine $\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')$ we postmultiply the kernel

$$\sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^*_{\hat{\mathfrak{b}}k'}(\mathbf{r}')$$

with an arbitrary eigenfunction from the basis set $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}')\}$ and integrate the result over the volume $\mathscr{V}$. On one hand, from the orthogonality relation (30) one obtains

$$\int_{\mathscr{V}} d^3\mathbf{r}' \left\{ \sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^*_{\hat{\mathfrak{b}}k'}(\mathbf{r}') \right\} \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}') = \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{51}$$

On the other hand, from Eq. (47) one finds

$$\int_{\mathscr{V}} d^3\mathbf{r}' \left\{ \sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^*_{\hat{\mathfrak{b}}k'}(\mathbf{r}') \right\} \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}') = \hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{52}$$

From Eqs. (51) and (52) one infers that

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{53}$$

This relation holds for an arbitrary scalar eigenfunction $\Psi_{\hat{\mathfrak{b}}k}$ and therefore we conclude that

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}} = \hat{\mathscr{T}} \Leftrightarrow \mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}') = \delta^{(2)}(\boldsymbol{\rho}-\boldsymbol{\rho}') \tag{54}$$

(here $\hat{\mathscr{T}}$ denotes the unit operator) which, in conjunction with Eq. (50), implies that

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho}) = \Psi(E,\boldsymbol{\rho}), \tag{55}$$

i.e., the eigenfunction expansion $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})$ converges to the function $\Psi(E,\mathbf{r})$ also at the surface $\mathscr{S}$. From Eqs. (41) and (56) we infer that

$$\Psi(E,\boldsymbol{\rho}) = \hat{\bar{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E)[\nabla_n - \hat{\mathfrak{b}}]\Psi(E,\boldsymbol{\rho}), \tag{56}$$

hence it follows that the operators $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ and $\hat{\tilde{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E)$ are identical,

$$\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)=\hat{\tilde{\mathscr{R}}}_{\hat{\mathfrak{b}}}(E) \tag{57}$$

and therefore [cf. Eq. (42)] one has

$$\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\frac{\hbar^2}{2m}\sum_k \frac{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k}-E}, \tag{58}$$

which means that the operations Σ_k and $\lim_{\mathbf{r}\to\boldsymbol{\rho}}$ in Eq. (46) do commute. We have thus completed the task of finding the kernel of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$. It is to be noticed that, in virtue of the reality of the eigenvalues $\{E_{\hat{\mathfrak{b}}k}\}$, it follows from Eq. (58) that the kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ possesses the property

$$\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\mathscr{R}^*_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho}',\boldsymbol{\rho}), \tag{59}$$

which means that the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ is Hermitian.

Once the kernel $\mathscr{R}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ has been determined, the R-matrix $\mathsf{R}_{\mathfrak{b}}(E)$ may be found. Projecting Eq. (58) from the left and from the right on the surface basis functions $\{\Phi_i(\boldsymbol{\rho})\}$ we get the well known result for $\mathsf{R}_{\mathfrak{b}}(E)$ (cf., for instance, Ref. 9)

$$\mathsf{R}_{\mathfrak{b}}(E)=\frac{\hbar^2}{2m}\sum_k \frac{\mathsf{P}_{\mathfrak{b}k}\mathsf{P}^{\dagger}_{\mathfrak{b}k}}{E_{\mathfrak{b}k}-E}, \tag{60}$$

where $\{\mathsf{P}_{\mathfrak{b}k}\}$ are column vectors with elements $\{P_{i,\mathfrak{b}k}=(\Phi_i|\Psi_{\hat{\mathfrak{b}}k})\}$ and $\{\mathsf{P}^{\dagger}_{\mathfrak{b}k}\}$ are row vectors with elements $\{P^*_{i,\mathfrak{b}k}=(\Psi_{\hat{\mathfrak{b}}k}|\Phi_i)\}$. Equation (60) is to be used in Eq. (25). Thus, we have completed the task of finding the matrix $\mathsf{R}_{\mathfrak{b}}(E)$ using the integral operators formalism.

In addition, it is to be noticed that as a by-product of the determination of the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$, from Eqs. (47), (48) and the second of Eqs. (54) we obtain the following extensions of the closure relation (31) for the cases when the point $\mathbf{r}'$ lies on the surface $\mathscr{S}$,

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\Psi^*_{\hat{\mathfrak{b}}k}(\mathbf{r})=\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^*_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')=\delta^{(3)}(\mathbf{r}-\boldsymbol{\rho}') \quad (\mathbf{r}\in\mathscr{V}). \tag{61}$$

IV. *R*-MATRIX THEORY FOR THE DIRAC EQUATION

A. The operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$

In this section we shall present an operator formulation of the R-matrix theory for a Dirac particle moving in a real, local, in general noncentral potential $V(\mathbf{r})$. As in the nonrelativistic case discussed above, we restrict our considerations to the closed volume $\mathscr{V}$ surrounded by the surface $\mathscr{S}$ [in this connection, see a remark preceding Eq. (11)]. The time-independent Dirac equation describing the particle is

$$[\hat{H}-E]\Psi(E,\mathbf{r})=0, \tag{62}$$

where E is the total energy of the particle (including its rest energy mc^2) and the Hamiltonian $\hat{H}$ is

$$\hat{H}=-ic\hbar\,\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}+\beta mc^2+V(\mathbf{r}), \tag{63}$$

where the Hermitian 4×4 Dirac matrices $\boldsymbol{\alpha}$ and β are defined as usual.[20]

Before proceeding further, we have to define a matrix function

$$\alpha_n(\boldsymbol{\rho})=\mathbf{n}(\boldsymbol{\rho})\cdot\boldsymbol{\alpha} \tag{64}$$

and a set of matrices

$$\beta^{(\pm)}=\frac{I\pm\beta}{2},\quad \alpha_n^{(\pm)}(\boldsymbol{\rho})=\beta^{(\pm)}\alpha_n(\boldsymbol{\rho}), \tag{65}$$

where I is the unit 4×4 matrix. The matrices $\alpha_n^{(\pm)}(\boldsymbol{\rho})$ and $\beta^{(\pm)}$ satisfy the following relations:

$$\alpha_n^{(+)}(\boldsymbol{\rho})+\alpha_n^{(-)}(\boldsymbol{\rho})=\alpha_n(\boldsymbol{\rho}),\quad \beta^{(+)}+\beta^{(-)}=I, \tag{66}$$

$$\alpha_n^{(\pm)\dagger}(\boldsymbol{\rho})=\alpha_n^{(\mp)}(\boldsymbol{\rho}),\quad \beta^{(\pm)\dagger}=\beta^{(\pm)}. \tag{67}$$

For the sake of later applications, we list also the following properties of the matrices $\beta^{(\pm)}$ and $\alpha_n^{(\pm)}(\boldsymbol{\rho})$:

$$\beta^{(\pm)}\beta^{(\pm)}=\beta^{(\pm)},\quad \beta^{(\pm)}\beta^{(\mp)}=0,\quad \alpha_n^{(\pm)}(\boldsymbol{\rho})\alpha_n^{(\pm)}(\boldsymbol{\rho})=0,\quad \alpha_n^{(\pm)}(\boldsymbol{\rho})\alpha_n^{(\mp)}(\boldsymbol{\rho})=\beta^{(\pm)}, \tag{68}$$

$$\alpha_n^{(\pm)}(\boldsymbol{\rho})\beta^{(\pm)}=0,\quad \beta^{(\pm)}\alpha_n^{(\pm)}(\boldsymbol{\rho})=\alpha_n^{(\pm)}(\boldsymbol{\rho}),\quad \beta^{(\pm)}\alpha_n^{(\mp)}(\boldsymbol{\rho})=0,\quad \alpha_n^{(\pm)}(\boldsymbol{\rho})\beta^{(\mp)}=\alpha_n^{(\pm)}(\boldsymbol{\rho}), \tag{69}$$

which may be easily derived from the definitions (64) and (65) and the anticommutation relations satisfied by the matrices $\boldsymbol{\alpha}$ and β.[20] Henceforth, whenever the matrices $\alpha_n^{(\pm)}(\boldsymbol{\rho})$ multiply an integral operator, we shall omit their argument $\boldsymbol{\rho}$ and assume that the following notational convention holds: If $\hat{\mathscr{O}}$ is an operator with a kernel $\mathscr{O}(\boldsymbol{\rho},\boldsymbol{\rho}')$, then $\alpha_n^{(\pm)}\hat{\mathscr{O}}$ and $\hat{\mathscr{O}}\alpha_n^{(\pm)}$ are operators with kernels $\alpha_n^{(\pm)}(\boldsymbol{\rho})\mathscr{O}(\boldsymbol{\rho},\boldsymbol{\rho}')$ and $\mathscr{O}(\boldsymbol{\rho},\boldsymbol{\rho}')\alpha_n^{(\pm)}(\boldsymbol{\rho}')$, respectively.

After that preparation we are now ready to attack the problem posed. Let $\Psi(E,\mathbf{r})$ and $\Psi'(E,\mathbf{r})$ be two solutions to Eq. (62) at the same energy E. On applying the Gauss divergence theorem we find

$$\langle\hat{H}\Psi'|\Psi\rangle-\langle\Psi'|\hat{H}\Psi\rangle=(\Psi'|ic\hbar\alpha_n\Psi). \tag{70}$$

By virtue of the reality of E, the left-hand side of this equation vanishes and we obtain

$$(\Psi'|i\alpha_n\Psi)=0. \tag{71}$$

On utilizing Eqs. (66) and (67) we may rewrite this relation in two equivalent forms,

$$(i\alpha_n^{(+)}\Psi'|\Psi)=(\Psi'|i\alpha_n^{(+)}\Psi),\quad (i\alpha_n^{(-)}\Psi'|\Psi)=(\Psi'|i\alpha_n^{(-)}\Psi). \tag{72}$$

Relations (72) may be formally interpreted that $i\alpha_n^{(\pm)}$, when acting on solutions of the Dirac equation (62) at fixed energy E, are Hermitian with respect to the surface scalar product $(|)$.

In the next step, we introduce two auxiliary energy-dependent linear Hermitian integral operators $\hat{\mathscr{B}}^{(\pm)}(E)$ defined on $L^2_{(|)}(\mathscr{S})$ such that

$$i\alpha_n^{(\pm)}(\boldsymbol{\rho})\Psi(E,\boldsymbol{\rho})=\gamma^{(\pm)}\hat{\mathscr{B}}^{(\pm)}(E)\Psi(E,\boldsymbol{\rho}) \tag{73}$$

for any solution of Eq. (62) at energy E. In Eq. (73) and hereafter

$$\gamma^{(\pm)}=\pm\left(\frac{\hbar}{2mc}\right)^{\pm1} \tag{74}$$

are numerical constants introduced for later convenience, primarily to facilitate a later discussion of the nonrelativistic limit [notice that $\gamma^{(\mp)}=-(\gamma^{(\pm)})^{-1}$]. The operators $\hat{\mathscr{B}}^{(\pm)}(E)$ are represented by their integral kernels $\mathscr{B}^{(\pm)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ and Eq. (73) may be equivalently rewritten in the form

$$i\alpha_n^{(\pm)}(\boldsymbol{\rho})\Psi(E,\boldsymbol{\rho})=\gamma^{(\pm)}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ \mathscr{B}^{(\pm)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\Psi(E,\boldsymbol{\rho}'). \tag{75}$$

The existence of the operators $\hat{\mathscr{B}}^{(\pm)}(E)$ has been proved by Szmytkowski[21] who has shown that the kernels $\mathscr{B}^{(\pm)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ may be written in the forms

$$\mathscr{B}^{(+)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_i \beta^{(+)}\Psi_i(E,\boldsymbol{\rho})b_i(E)\Psi_i^\dagger(E,\boldsymbol{\rho}')\beta^{(+)}, \tag{76}$$

$$\mathscr{B}^{(-)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=(\gamma^{(-)})^2\sum_i \beta^{(-)}\Psi_i(E,\boldsymbol{\rho})b_i^{-3}(E)\Psi_i^\dagger(E,\boldsymbol{\rho}')\beta^{(-)}, \tag{77}$$

where $\{\Psi_i(E,\mathbf{r})\}$ are those solutions to the Dirac equation (62) at energy E which on the surface $\mathscr{S}$ satisfy the relations

$$i\alpha_n^{(+)}(\boldsymbol{\rho})\Psi_i(E,\boldsymbol{\rho})=\gamma^{(+)}b_i(E)\beta^{(+)}\Psi_i(E,\boldsymbol{\rho}) \tag{78}$$

and are normalized according to $(\Psi_i|\beta^{(+)}\Psi_i)=1$. It may be shown[21] that, as in the nonrelativistic case, the numbers $\{b_i(E)\}$ are real and that the surface spinor functions $\{\Psi_i(E,\boldsymbol{\rho})\}$ are orthogonal in the sense of $(\Psi_i|\beta^{(+)}\Psi_j)=\delta_{ij}$.

It is to be noticed that Eqs. (76) and (77) imply the following property of the operators $\hat{\mathscr{B}}^{(\pm)}(E)$:

$$\beta^{(\pm)}\hat{\mathscr{B}}^{(\pm)}(E)\beta^{(\pm)}=\hat{\mathscr{B}}^{(\pm)}(E). \tag{79}$$

Moreover, the operators $\hat{\mathscr{B}}^{(+)}(E)$ and $\hat{\mathscr{B}}^{(-)}(E)$ are not independent. Indeed, operating on Eq. (73) from the left with $-i\alpha_n^{(\mp)}(\boldsymbol{\rho})$ yields

$$\beta^{(\mp)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathscr{B}}^{(\pm)}(E)\Psi(E,\boldsymbol{\rho}). \tag{80}$$

Operating then on this equation from the left with $\alpha_n^{(\pm)}\hat{\mathscr{B}}^{(\mp)}(E)$ and transforming the resulting equation with the aid of Eq. (80) gives

$$\beta^{(\pm)}\Psi(E,\boldsymbol{\rho})=\alpha_n^{(\pm)}\hat{\mathscr{B}}^{(\mp)}(E)\alpha_n^{(\mp)}\hat{\mathscr{B}}^{(\pm)}(E)\Psi(E,\boldsymbol{\rho}), \tag{81}$$

which implies that the operators $\hat{\mathscr{B}}^{(\pm)}(E)$ and $\alpha_n^{(\pm)}\hat{\mathscr{B}}^{(\mp)}(E)\alpha_n^{(\mp)}$ are reciprocal in the sense of

$$[\alpha_n^{(\pm)}\hat{\mathscr{B}}^{(\mp)}(E)\alpha_n^{(\mp)}]\hat{\mathscr{B}}^{(\pm)}(E)\hat{\mathscr{P}}^{(\pm)}(E)=\hat{\mathscr{P}}^{(\pm)}(E), \tag{82}$$

where the projectors $\hat{\mathscr{P}}^{(\pm)}(E)$ are related to the operator $\hat{\mathscr{P}}(E)$ projecting on the surface parts of solutions to the Dirac equation (62) at energy E in the following way:

$$\hat{\mathscr{P}}^{(\pm)}(E)=\beta^{(\pm)}\hat{\mathscr{P}}(E)\beta^{(\pm)}. \tag{83}$$

Integral kernels of the projectors $\hat{\mathscr{P}}^{(\pm)}(E)$ are

$$\mathscr{P}^{(+)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_i \beta^{(+)}\Psi_i(E,\boldsymbol{\rho})\Psi_i^\dagger(E,\boldsymbol{\rho}')\beta^{(+)}, \tag{84}$$

$$\mathscr{P}^{(-)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=(\gamma^{(-)})^2\sum_i \beta^{(-)}\Psi_i(E,\boldsymbol{\rho})b_i^{-2}(E)\Psi_i^\dagger(E,\boldsymbol{\rho}')\beta^{(-)}, \tag{85}$$

where the numbers $\{b_i(E)\}$ and the functions $\{\Psi_i(E,\mathbf{r})\}$ have been defined below Eqs. (76) and (77).

A note analogous to that remarked in the paragraph following Eq. (18) is in order here. Namely, we emphasize that the operators $i\alpha_n^{(\pm)}$ and $\gamma^{(\pm)}\hat{\mathscr{B}}^{(\pm)}(E)$ are not identical and for an arbitrary four-component spinor function $\Phi(\boldsymbol{\rho})$ defined on $\mathscr{S}$ in general one has

$$i\alpha_n^{(\pm)}(\boldsymbol{\rho})\Phi(\boldsymbol{\rho})\neq\gamma^{(\pm)}\hat{\mathscr{B}}^{(\pm)}(E)\Phi(\boldsymbol{\rho}) \tag{86}$$

unless $\Phi(\boldsymbol{\rho})$ coincides with a surface part of some solution to the Dirac equation (62) at energy E.

Let $\hat{\mathfrak{b}}^{(+)}$ and $\hat{\mathfrak{b}}^{(-)}$ be two linear Hermitian integral operators defined on $L^2_{(|)}(\mathscr{S})$, with the integral kernels $\mathfrak{b}^{(+)}(\boldsymbol{\rho},\boldsymbol{\rho}')$ and $\mathfrak{b}^{(-)}(\boldsymbol{\rho},\boldsymbol{\rho}')$, respectively. It will be assumed that the operators $\hat{\mathfrak{b}}^{(\pm)}$ are such that

$$\beta^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}\beta^{(\pm)}=\hat{\mathfrak{b}}^{(\pm)} \tag{87}$$

[cf. Eq. (79)] and such that

$$\hat{\mathfrak{b}}^{(\pm)}\alpha_n^{(\pm)}\hat{\mathfrak{b}}^{(\mp)}\alpha_n^{(\mp)}=\alpha_n^{(\pm)}\hat{\mathfrak{b}}^{(\mp)}\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}=\beta^{(\pm)} \tag{88}$$

[cf. Eq. (82)]. We define the operators $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ and $\hat{\mathscr{R}}^{(-)}_{\hat{\mathfrak{b}}}(E)$, possessing the property

$$\beta^{(\pm)}\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)\beta^{(\pm)}=\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E), \tag{89}$$

as the operators reciprocal to the operators $\hat{\mathscr{B}}^{(+)}(E)-\hat{\mathfrak{b}}^{(+)}$ and $\hat{\mathscr{B}}^{(-)}(E)-\hat{\mathfrak{b}}^{(-)}$, respectively, in the sense of

$$\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)[\hat{\mathscr{B}}^{(\pm)}(E)-\hat{\mathfrak{b}}^{(\pm)}]\hat{\mathscr{P}}^{(\pm)}(E)=\hat{\mathscr{P}}^{(\pm)}(E). \tag{90}$$

In Sec. IV B we shall show how the operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ is related to the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ encountered in the nonrelativistic theory.

In terms of the operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$ the surface relation (73) may be rewritten as

$$\beta^{(\pm)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(\mp)}\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)[i\alpha_n^{(\pm)}-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}) \tag{91}$$

or, explicitly,

$$\beta^{(\pm)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(\mp)}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ \mathscr{R}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}'). \tag{92}$$

Assume now that a complete set of spinor functions $\{\Phi_i(\boldsymbol{\rho})\}\in L^2_{(|)}(\mathscr{S})$ orthonormal with respect to the surface scalar product $(|)$, spanning the surface $\mathscr{S}$, is given. In the basis $\{\Phi_i(\boldsymbol{\rho})\}$ the kernels $\mathscr{R}^{(\pm)}_{\mathfrak{b}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ have bilinear expansions

$$\mathscr{R}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\sum_{i,j}\ \beta^{(\pm)}\Phi_i(\boldsymbol{\rho})(\Phi_i|\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}\Phi_j)\Phi_j^\dagger(\boldsymbol{\rho}')\beta^{(\pm)}, \tag{93}$$

where the coefficients

$$(\Phi_i|\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}\Phi_j)\equiv\int_{\mathscr{S}}d^2\boldsymbol{\rho}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ \Phi_i^\dagger(\boldsymbol{\rho})\mathscr{R}^{(+)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\Phi_j(\boldsymbol{\rho}') \tag{94}$$

and

$$(\Phi_i|\hat{\mathscr{R}}^{(-)}_{\hat{\mathfrak{b}}}\Phi_j)\equiv\int_{\mathscr{S}}d^2\boldsymbol{\rho}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ \Phi_i^\dagger(\boldsymbol{\rho})\mathscr{R}^{(-)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\Phi_j(\boldsymbol{\rho}') \tag{95}$$

form R-matrices $\mathsf{R}^{(+)}_{\mathfrak{b}}(E)$ and $\mathsf{R}^{(-)}_{\mathfrak{b}}(E)$, respectively. To relate the matrix and operator formulations of the theory, we need a matrix representation of Eq. (91). Projecting the latter equation from the left on the basis spinor functions $\{\Phi_i(\boldsymbol{\rho})\}$ gives

$$\mathsf{P}^{(\pm)}(E)=\mathsf{R}^{(\pm)}_{\mathfrak{b}}(E)\left[\pm\left(\frac{2mc}{\hbar}\right)^{\pm1}\mathsf{Q}^{(\pm)}(E)-\mathsf{b}^{(\pm)}\mathsf{P}^{(\pm)}(E)\right], \tag{96}$$

where $\mathsf{P}^{(\pm)}(E)$ and $\mathsf{Q}^{(\pm)}(E)$ are column matrices with elements $\{P_i^{(\pm)}(E)=(\Phi_i|\beta^{(\pm)}\Psi)\}$ and $\{Q_i^{(\pm)}(E)=(\Phi_i|i\alpha_n^{(\pm)}\Psi)\}$, respectively, and $\mathsf{b}^{(\pm)}$ are square matrices with elements

$\{b_{ij}^{(\pm)}=(\Phi_i|\hat{\mathfrak{b}}^{(\pm)}\Phi_j)\}$. Equations (96) are the counterparts of the nonrelativistic formula (1). The relation which corresponds to the choice of the upper sign in Eq. (96) was found before by Chang.[16]

In Sec. IV C we shall show how the integral kernels of the operators $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}^{(\pm)}(E)$ may be constructed.

B. The nonrelativistic limit

It is of interest to investigate the nonrelativistic limit of Eq. (91) in order to relate the relativistic and nonrelativistic formulations of the operator versions of the *R*-matrix theory. To accomplish this goal, we proceed in the standard manner and express the relativistic four-component wave function $\Psi(E,\mathbf{r})$ in terms of two-component spinors $\psi_U(E,\mathbf{r})$ and $\psi_L(E,\mathbf{r})$ in the following way:

$$\Psi(E,\mathbf{r})=\begin{pmatrix}\psi_U(E,\mathbf{r})\\ \psi_L(E,\mathbf{r})\end{pmatrix}. \tag{97}$$

It is a standard problem in introductory quantum mechanics to show that in the nonrelativistic limit the function $\psi_U(E,\mathbf{r})$ obeys the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2+V(\mathbf{r})-E\right]\psi_U(E,\mathbf{r})=0 \quad (c\to\infty) \tag{98}$$

[cf. Eqs. (11) and (12)] while

$$\psi_L(E,\mathbf{r})\xrightarrow{c\to\infty}-\left(\frac{\hbar}{2mc}\right)i\boldsymbol{\sigma}\cdot\boldsymbol{\nabla}\psi_U(E,\mathbf{r}), \tag{99}$$

where $\boldsymbol{\sigma}$ is a vector composed of the Pauli matrices. The relation (99) holds for arbitrary $\mathbf{r}$, in particular it is also valid on the surface $\mathscr{S}$, i.e., for $\mathbf{r}=\boldsymbol{\rho}$. Next, we notice that, as a consequence of Eqs. (87) and (89), the operators $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}^{(+)}(E)$ and $\hat{\mathfrak{b}}^{(+)}$ act on an arbitrary function

$$\Phi(\boldsymbol{\rho})=\begin{pmatrix}\phi_U(\boldsymbol{\rho})\\ \phi_L(\boldsymbol{\rho})\end{pmatrix} \tag{100}$$

from $L^2_{(|)}(\mathscr{S})$ in the following way:

$$\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}^{(+)}(E)\Phi(\boldsymbol{\rho})\equiv\begin{pmatrix}\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)\,\phi_U(\boldsymbol{\rho})\\ 0\end{pmatrix}, \quad \hat{\mathfrak{b}}^{(+)}\Phi(\boldsymbol{\rho})\equiv\begin{pmatrix}\hat{\mathfrak{b}}'\,\phi_U(\boldsymbol{\rho})\\ 0\end{pmatrix}, \tag{101}$$

where $\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)$ and $\hat{\mathfrak{b}}'$ are operators acting on two-component functions. Making use of Eq. (101), the surface relation (91) may be rewritten in terms of the upper component of $\Psi(E,\boldsymbol{\rho})$ and of the operators $\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)$ and $\hat{\mathfrak{b}}'$ in the following way:

$$\psi_U(E,\boldsymbol{\rho})=\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)\left[\left(\frac{2mc}{\hbar}\right)i\boldsymbol{\sigma}\cdot\mathbf{n}(\boldsymbol{\rho})\,\psi_L(E,\boldsymbol{\rho})-\hat{\mathfrak{b}}'\,\psi_U(E,\boldsymbol{\rho})\right]. \tag{102}$$

The relation (102) is exact. Passing in this relation to the nonrelativistic limit, making use of Eq. (99) and utilizing well-known properties of the Pauli matrices, which give

$$[\boldsymbol{\sigma}\cdot\mathbf{n}(\boldsymbol{\rho})][\boldsymbol{\sigma}\cdot\boldsymbol{\nabla}]=\nabla_n+i\boldsymbol{\sigma}\cdot[\mathbf{n}(\boldsymbol{\rho})\times\boldsymbol{\nabla}], \tag{103}$$

we obtain

$$\psi_U(E,\boldsymbol{\rho})=\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)[\nabla_n+i\boldsymbol{\sigma}\cdot[\mathbf{n}(\boldsymbol{\rho})\times\boldsymbol{\nabla}]-\hat{\mathfrak{b}}']\psi_U(E,\boldsymbol{\rho}) \quad (c\to\infty). \tag{104}$$

It is understood that in Eq. (104) $\psi_U(E,\boldsymbol{\rho})$, $\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E)$ and $\hat{\mathfrak{b}}'$ stand for nonrelativistic limits of the quantities which have been denoted with the same symbols before the passage to the limit has been done. Comparison between Eqs. (22) and (104) shows that in the nonrelativistic limit

$$\hat{\mathscr{R}}'_{\hat{\mathfrak{b}}}(E) \xrightarrow{c\to\infty} \hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E).\hat{\mathscr{T}} \tag{105}$$

(here $\hat{\mathscr{T}}$ is the unit 2×2 operator) provided that the operator $\hat{\mathfrak{b}}^{(+)}$ is chosen so that

$$\hat{\mathfrak{b}}' \xrightarrow{c\to\infty} \hat{\mathfrak{b}}.\hat{\mathscr{T}} + i\boldsymbol{\sigma}\cdot[\mathbf{n}(\boldsymbol{\rho})\times\boldsymbol{\nabla}], \tag{106}$$

where $\hat{\mathfrak{b}}$ is the operator used in the formulation of the theory for the Schrödinger equation (cf. Sec. III A). It is interesting and somewhat surprising that the operators $\hat{\mathfrak{b}}'$ and $\hat{\mathfrak{b}}.\hat{\mathscr{T}}$ do not coincide. This observation is, however, consistent with earlier findings of Norrington and Grant[22,23] and of Thumm and Norcross[24] who investigated the nonrelativistic limit of the matrix formulation of the theory in the specific case of the spherically symmetric volume $\mathscr{V}$.

C. Construction of the operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$

Thus far we have discussed some properties of the operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$ but have not touched the problem of construction of integral kernels of these operators. This will be the subject of the present section.

To achieve the goal, we need a set of auxiliary spinor functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ defined as eigenfunctions of the boundary value problem consisting of the differential equation [cf. Eq. (62)]

$$[\hat{H}-E_{\hat{\mathfrak{b}}k}]\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})=0 \tag{107}$$

and of the homogeneous boundary condition

$$i\alpha_n^{(\pm)}(\boldsymbol{\rho})\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = \gamma^{(\pm)}\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \mathfrak{b}^{(\pm)}(\boldsymbol{\rho},\boldsymbol{\rho}')\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}'). \tag{108}$$

In Eq. (107) $\hat{H}$ is the Dirac Hamiltonian defined by Eq. (63) and $E_{\hat{\mathfrak{b}}k}$ is an eigenvalue for the problem belonging to an eigenfunction $\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})$. In the more compact operator notation the boundary condition (108) reads

$$i\alpha_n^{(\pm)}(\boldsymbol{\rho})\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = \gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{109}$$

At first sight it might seem that Eqs. (107) and (109) constitute two eigenvalue problems depending on which of the two possible signs is chosen in Eq. (109). Yet this is not the case: It is easy to verify that, because of relation (88), both boundary conditions (109) are essentially equivalent and consequently we have only one eigenvalue problem. For the sake of later applications, we notice also here that the boundary condition (109) may be equivalently rewritten, after operating on it from the left with $-i\alpha_n^{(\mp)}(\boldsymbol{\rho})$, in the form

$$\beta^{(\mp)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = -\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \tag{110}$$

hence, it follows that

$$\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \equiv \beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) + \beta^{(\mp)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) = [\beta^{(\pm)} - \gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}]\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{111}$$

Henceforth we shall assume that the integral operators $\hat{\mathfrak{b}}^{(\pm)}$ are such that the eigenfunctions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ form a complete set in the interior of the volume $\mathscr{V}$. It may be inferred from Eqs. (107) and (109) and from properties of the operators $\hat{\mathfrak{b}}^{(\pm)}$ that eigenvalues $\{E_{\hat{\mathfrak{b}}k}\}$ are real and eigenfunctions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ belonging to different eigenvalues are orthogonal. Indeed, after premultiplying Eq. (107) (with k replaced by k') by $\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r})$, postmultiplying the matrix Hermitian conjugate of Eq.

(107) by $\Psi_{\hat{\mathfrak{b}}k'}(\mathbf{r})$, subtracting, integrating the result over the volume $\mathscr{V}$, employing the Gauss divergence theorem, making use of the boundary condition (107) and of the Hermicity of $\hat{\mathfrak{b}}^{(\pm)}$, one arrives at

$$[E^{*}_{\hat{\mathfrak{b}}k}-E_{\hat{\mathfrak{b}}k'}]\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi_{\hat{\mathfrak{b}}k'}\rangle=0, \tag{112}$$

hence the validity of the assertion follows immediately. Hereafter we shall assume that functions belonging to degenerate eigenvalues (if there are any) have also been orthogonalized and that the functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ have been normalized so that

$$\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi_{\hat{\mathfrak{b}}k'}\rangle=\delta_{kk'}\,. \tag{113}$$

On the ground of the assumed completeness, the eigenfunctions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ obey the closure relation

$$\sum_{k}\ \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')=\delta^{(3)}(\mathbf{r}-\mathbf{r}')I\quad(\mathbf{r},\mathbf{r}'\in\mathscr{V}\setminus\mathscr{S}) \tag{114}$$

[cf. the remark following Eq. (31)].

Consider now a problem of expanding the function $\Psi(E,\mathbf{r})$ satisfying the Dirac equation (62) in the series of eigenfunctions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$. Denoting the expansion by $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})$, we have

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})\equiv\sum_{k}\ \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi\rangle \tag{115}$$

or equivalently

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})\equiv\int_{\mathscr{V}}d^3\mathbf{r}'\left\{\sum_{k}\ \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')\right\}\Psi(E,\mathbf{r}'). \tag{116}$$

The expansion coefficients $\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi\rangle$ may be found in the routine way. We premultiply Eq. (62) by $\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r})$, postmultiply the matrix Hermitian conjugate of Eq. (107) by $\Psi(E,\mathbf{r})$, subtract and integrate over the volume $\mathscr{V}$ obtaining

$$ic\hbar\int_{\mathscr{V}}d^3\mathbf{r}\ \ \boldsymbol{\nabla}\cdot[\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r})\boldsymbol{\alpha}\Psi(E,\mathbf{r})]=[E_{\hat{\mathfrak{b}}k}-E]\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi\rangle. \tag{117}$$

Employing the Gauss divergence theorem, the volume integral on the left-hand side may be converted into a surface integral

$$\int_{\mathscr{V}}d^3\mathbf{r}\ \boldsymbol{\nabla}\cdot[\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r})\boldsymbol{\alpha}\Psi(E,\mathbf{r})]=\int_{\mathscr{S}}d^2\boldsymbol{\rho}\ \Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\alpha_n(\boldsymbol{\rho})\Psi(E,\boldsymbol{\rho}). \tag{118}$$

This yields

$$\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi\rangle=\frac{(\Psi_{\hat{\mathfrak{b}}k}|ic\hbar\alpha_n\Psi)}{E_{\hat{\mathfrak{b}}k}-E}. \tag{119}$$

Making use of the properties of the matrices $\alpha_n(\boldsymbol{\rho})$ and $\alpha^{(\pm)}_n(\boldsymbol{\rho})$, of the boundary condition (109) and of the Hermicity property of the operators $\hat{\mathfrak{b}}^{(\pm)}$ we transform this expression to the form

$$\langle\Psi_{\hat{\mathfrak{b}}k}|\Psi\rangle=\frac{c\hbar(\Psi_{\hat{\mathfrak{b}}k}|[i\alpha^{(\pm)}_n-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi)}{E_{\hat{\mathfrak{b}}k}-E} \tag{120}$$

and consequently we arrive at

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})=\sum_k\ \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\ \frac{c\hbar(\Psi_{\hat{\mathfrak{b}}k}|[i\alpha_n^{(\pm)}-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi)}{E_{\hat{\mathfrak{b}}k}-E} \tag{121}$$

or equivalently

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})=\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ c\hbar\sum_k\ \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k}-E}[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}'). \tag{122}$$

In particular, if the point $\mathbf{r}$ lies on the surface $\mathscr{S}$, then we have

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})=\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ c\hbar\sum_k\ \frac{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k}-E}[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}'). \tag{123}$$

At that moment, it is convenient to define two integral operators $\hat{\bar{\mathscr{R}}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$ (notice the bar) such that [cf. Eq. (91)]

$$\beta^{(\pm)}\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})=-\gamma^{(\mp)}\hat{\bar{\mathscr{R}}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)[i\alpha_n^{(\pm)}-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}). \tag{124}$$

Comparison of Eqs. (123) and (124) shows that the operators $\hat{\bar{\mathscr{R}}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$ have the integral kernels

$$\bar{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=c\hbar\,\gamma^{(\pm)}\sum_k\ \frac{\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\beta^{(\pm)}}{E_{\hat{\mathfrak{b}}k}-E}. \tag{125}$$

By virtue of the closure relation (114), in the interior of the volume $\mathscr{V}$ one has

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})\equiv\int_{\mathscr{V}}d^3\mathbf{r}'\left\{\sum_k\ \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')\right\}\Psi(E,\mathbf{r}')=\Psi(E,\mathbf{r})\quad(\mathbf{r}\in\mathscr{V}\backslash\mathscr{S}), \tag{126}$$

i.e., the expansion $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\mathbf{r})$ does converge to the function $\Psi(E,\mathbf{r})$ and consequently

$$\Psi(E,\mathbf{r})=\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ c\hbar\sum_k\ \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k}-E}[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}')\quad(\mathbf{r}\in\mathscr{V}\backslash\mathscr{S}). \tag{127}$$

This relation is valid for the point $\mathbf{r}$ laying as close to the boundary $\mathscr{S}$ as we please and therefore, in view of the continuity of the function $\Psi(E,\mathbf{r})$ across $\mathscr{S}$, one has

$$\Psi(E,\boldsymbol{\rho})=\lim_{\mathbf{r}\to\boldsymbol{\rho}}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ c\hbar\sum_k\ \frac{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')}{E_{\hat{\mathfrak{b}}k}-E}[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}')\quad(\mathbf{r}\in\mathscr{V}\backslash\mathscr{S}) \tag{128}$$

and consequently

$$\begin{aligned}\beta^{(\pm)}\Psi(E,\boldsymbol{\rho})=\lim_{\mathbf{r}\to\boldsymbol{\rho}}\int_{\mathscr{S}}d^2\boldsymbol{\rho}'\ c\hbar\sum_k\ &\frac{\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\beta^{(\pm)}}{E_{\hat{\mathfrak{b}}k}-E}[i\alpha_n^{(\pm)}(\boldsymbol{\rho}')\\&-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}')\quad(\mathbf{r}\in\mathscr{V}\backslash\mathscr{S}),\end{aligned} \tag{129}$$

which, in conjunction with Eq. (92), implies

$$\mathscr{R}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')=\lim_{\mathbf{r}\to\boldsymbol{\rho}}\left\{c\hbar\,\gamma^{(\pm)}\sum_k\ \frac{\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\beta^{(\pm)}}{E_{\hat{\mathfrak{b}}k}-E}\right\}. \tag{130}$$

This is to be compared with Eq. (125).

We now ask the question: Are the operations $\lim_{\mathbf{r}\to\boldsymbol{\rho}}$ and Σ_k in Eq. (130) interchangeable freely? Relying on the apparent similarity of the relativistic theory discussed here to the nonrelativistic theory exposed in Sec. III B one might expect that the answer to this question is positive. However, this is *not* the case. In the relativistic theory the operations $\lim_{\mathbf{r}\to\boldsymbol{\rho}}$ and Σ_k do *not* commute and accordingly one finds $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')\neq\overline{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$ and $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)\neq\hat{\bar{\mathscr{R}}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$. To show this (and find the consequences of interchanging the two operations) we need to find a direct relation between the surface functions $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})$ and $\Psi(E,\boldsymbol{\rho})$. We begin with an observation that the closure relation (114) cannot hold whenever any of the points $\mathbf{r}$ or $\mathbf{r}'$ lies on the surface $\mathscr{S}$ since the upper and the lower components of the surface spinor functions $\{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\}$ are related by the boundary condition (109). Therefore, we assume that

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')=\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\,\delta^{(1)}_{\mathscr{S}}(\mathbf{r}') \quad (\mathbf{r}'\in\mathscr{S}), \tag{131}$$

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')=\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\,\delta^{(1)}_{\mathscr{S}}(\mathbf{r}) \quad (\mathbf{r}\in\mathscr{S}) \tag{132}$$

[cf. Eq. (114) and the first of Eqs. (9)] and hope that, in the course of the following considerations, we shall be able to find the surface kernel $\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')$. Substituting the relation (131) into the definition (116) we get

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\ \mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\Psi(E,\boldsymbol{\rho}'), \tag{133}$$

which implies that the kernel $\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')$ defines the integral operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ transforming the surface function $\Psi(E,\boldsymbol{\rho})$ into $\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})$,

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})=\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi(E,\boldsymbol{\rho}). \tag{134}$$

An important property of the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ may be deduced from the definitions (131) and (132) and the relation (111). On one hand, it follows from Eq. (111) that

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')=[\hat{\mathscr{I}}-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}]\beta^{(\pm)}\left\{\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\right\}, \tag{135}$$

where $\hat{\mathscr{I}}$ is the unit operator. On the other hand, if $\Phi(\boldsymbol{\rho})$ is any reasonable spinor function defined on the surface $\mathscr{S}$, from Eq. (132) and the Hermicity of the operators $\hat{\mathfrak{b}}^{(\pm)}$ under the surface scalar product $(|)$ we readily obtain

$$\begin{aligned}
&\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\left\{\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\right\}\Phi(\boldsymbol{\rho}')\\
&\quad=\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\{[\beta^{(\pm)}-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}]\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\}^{\dagger}\Phi(\boldsymbol{\rho}')\\
&\quad=\int_{\mathscr{S}} d^2\boldsymbol{\rho}'\left\{\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\right\}\beta^{(\pm)}[\hat{\mathscr{I}}+\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}i\alpha_n^{(\pm)}]\Phi(\boldsymbol{\rho}').
\end{aligned} \tag{136}$$

Equations (131), (132), (135) and (136) imply that the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ may be sought in any of the two equivalent forms

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=[\hat{\mathscr{I}}-\gamma^{(+)}i\alpha_n^{(-)}\hat{\mathfrak{b}}^{(+)}].\hat{\mathscr{A}}^{(+)}_{\hat{\mathfrak{b}}}[\hat{\mathscr{I}}+\gamma^{(+)}\hat{\mathfrak{b}}^{(+)}i\alpha_n^{(+)}], \tag{137}$$

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=[\hat{\mathscr{I}}-\gamma^{(-)}i\alpha_n^{(+)}\hat{\mathfrak{b}}^{(-)}].\hat{\mathscr{A}}^{(-)}_{\hat{\mathfrak{b}}}[\hat{\mathscr{I}}+\gamma^{(-)}\hat{\mathfrak{b}}^{(-)}i\alpha_n^{(-)}], \tag{138}$$

where the operators

$$\hat{\mathscr{A}}^{(\pm)}_{\hat{\mathfrak{b}}}=\beta^{(\pm)}\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\beta^{(\pm)} \tag{139}$$

have the integral kernels $\mathscr{A}^{(\pm)}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')=\beta^{(\pm)}\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\beta^{(\pm)}$ defined formally by

$$\sum_k \beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r}')\beta^{(\pm)}=\mathscr{A}^{(\pm)}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')\,\delta^{(1)}_{\mathscr{T}}(\mathbf{r}') \quad (\mathbf{r}'\in\mathscr{T}). \tag{140}$$

The operators $\hat{\mathscr{A}}^{(+)}_{\hat{\mathfrak{b}}}$ and $\hat{\mathscr{A}}^{(-)}_{\hat{\mathfrak{b}}}$ are related; it is readily derivable from Eqs. (137) to (139) that

$$\hat{\mathscr{A}}^{(\pm)}_{\hat{\mathfrak{b}}}=(\gamma^{(\pm)})^2\alpha_n^{(\pm)}\hat{\mathfrak{b}}^{(\mp)}\hat{\mathscr{A}}^{(\mp)}_{\hat{\mathfrak{b}}}\hat{\mathfrak{b}}^{(\mp)}\alpha_n^{(\mp)}. \tag{141}$$

To find the explicit form of the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$, we postmultiply the kernel

$$\sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k'}(\mathbf{r}')$$

by any function $\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}')$ from the complete set $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}')\}$ and integrate over the volume $\mathscr{T}$. From the orthogonality relation (113) we infer that

$$\int_{\mathscr{T}} d^3\mathbf{r}'\left\{\sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k'}(\mathbf{r}')\right\}\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}')=\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{142}$$

On the other hand, from Eq. (131) we have

$$\int_{\mathscr{T}} d^3\mathbf{r}'\left\{\sum_{k'} \Psi_{\hat{\mathfrak{b}}k'}(\boldsymbol{\rho})\Psi^{\dagger}_{\hat{\mathfrak{b}}k'}(\mathbf{r}')\right\}\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r}')=\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{143}$$

Equating the right-hand sides of Eqs. (142) and (143) we find

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})=\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{144}$$

At first sight Eq. (144) might seem to imply that, as in the nonrelativistic theory, the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ is the unit operator. Yet this is *not* the case. The reason is that in the relativistic theory the functions $\{\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})\}$ are multicomponent and their upper and lower components are related by the algebraic condition (109). Rewriting Eq. (144) with the aid of Eqs. (137) and (138) we obtain

$$[\hat{\mathscr{T}}-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}]\hat{\mathscr{A}}^{(\pm)}_{\hat{\mathfrak{b}}}[\hat{\mathscr{T}}+\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}i\alpha_n^{(\pm)}]\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})=\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}). \tag{145}$$

Upon operating on both sides of this equation with $\beta^{(\pm)}$ and utilizing the boundary condition (109) we get

$$\hat{\mathscr{A}}^{(\pm)}_{\hat{\mathfrak{b}}}[\hat{\mathscr{T}}+(\gamma^{(\pm)})^2(\hat{\mathfrak{b}}^{(\pm)})^2]\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})=\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}) \tag{146}$$

and accordingly

$$\hat{\mathscr{A}}^{(\pm)}_{\hat{\mathfrak{b}}}=\beta^{(\pm)}[\hat{\mathscr{T}}+(\gamma^{(\pm)})^2(\hat{\mathfrak{b}}^{(\pm)})^2]^{-1}\beta^{(\pm)} \tag{147}$$

since the relation (136) holds for an arbitrary eigenfunction $\Psi_{\hat{\mathfrak{b}}k}$. Hence, it follows that

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=[\beta^{(\pm)}-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}][\hat{\mathscr{T}}+(\gamma^{(\pm)})^2(\hat{\mathfrak{b}}^{(\pm)})^2]^{-1}[\beta^{(\pm)}+\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}i\alpha_n^{(\pm)}] \tag{148}$$

and consequently Eq. (134) may be written in the form

$$\overline{\Psi}_{\hat{\mathfrak{b}}}(E,\boldsymbol{\rho})=[\beta^{(\pm)}-\gamma^{(\pm)}i\alpha_n^{(\mp)}\hat{\mathfrak{b}}^{(\pm)}][\hat{\mathscr{T}}+(\gamma^{(\pm)})^2(\hat{\mathfrak{b}}^{(\pm)})^2]^{-1}[\beta^{(\pm)}+\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}i\alpha_n^{(\pm)}]\Psi(E,\boldsymbol{\rho}). \tag{149}$$

Substituting this relation to Eq. (124), after straightforward manipulations we arrive at

$$\beta^{(\pm)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(\mp)}[\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)-\hat{\mathfrak{b}}^{(\pm)}[(\gamma^{(\mp)})^2\hat{\mathscr{T}}+(\hat{\mathfrak{b}}^{(\pm)})^2]^{-1}][i\alpha_n^{(\pm)}-\gamma^{(\pm)}\hat{\mathfrak{b}}^{(\pm)}]\Psi(E,\boldsymbol{\rho}) \tag{150}$$

hence, on comparing Eqs. (91) and (150), we find the relationship between the operators $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ and $\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(\pm)}(E)$,

$$\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)=\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(\pm)}(E)-\hat{\mathfrak{b}}^{(\pm)}[(\gamma^{(\mp)})^2\hat{\mathscr{T}}+(\hat{\mathfrak{b}}^{(\pm)})^2]^{-1}. \tag{151}$$

Equation (150) may be perceived as a set of two homogeneous coupled equations for the functions $\beta^{(+)}\Psi(E,\boldsymbol{\rho})$ and $\beta^{(-)}\Psi(E,\boldsymbol{\rho})$. Indeed, upon utilizing the properties of the matrices $\alpha_n^{(\pm)}(\boldsymbol{\rho})$ and $\beta^{(\pm)}$ the two relations constituting Eq. (150) may be explicitly rewritten in the form

$$\beta^{(+)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(-)}\hat{\mathscr{R}}_{\mathfrak{b}}^{(+)}(E)i\alpha_n^{(+)}\beta^{(-)}\Psi(E,\boldsymbol{\rho})-\hat{\mathscr{R}}_{\mathfrak{b}}^{(+)}(E)\hat{\mathfrak{b}}^{(+)}\beta^{(+)}\Psi(E,\boldsymbol{\rho}), \tag{152}$$

$$\beta^{(-)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(+)}\hat{\mathscr{R}}_{\mathfrak{b}}^{(-)}(E)i\alpha_n^{(-)}\beta^{(+)}\Psi(E,\boldsymbol{\rho})-\hat{\mathscr{R}}_{\mathfrak{b}}^{(-)}(E)\hat{\mathfrak{b}}^{(-)}\beta^{(-)}\Psi(E,\boldsymbol{\rho}), \tag{153}$$

where the operators $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ have been given by Eq. (151). The following problem arises now: Since Eqs. (152) and (153) are homogeneous and coupled, are they consistent? This should be the case if the relations (151) are correct; therefore seeking an answer to the question posed provides a test for correctness of the method of construction of the operators $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ presented above. [It is to be emphasized that testing the theory in that manner is by no means insipid. This is just the way in which Szmytkowski and Hinze[17,18] recognized an error in earlier attempts[15,16] to formulate the *R*-matrix theory for the Dirac equation within the framework of the matrix language.] To clear up the problem we solve Eq. (153) for $\beta^{(-)}\Psi(E,\boldsymbol{\rho})$ obtaining

$$\beta^{(-)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(+)}\beta^{(-)}[\hat{\mathscr{T}}+\hat{\mathscr{R}}_{\mathfrak{b}}^{(-)}(E)\hat{\mathfrak{b}}^{(-)}]^{-1}\hat{\mathscr{R}}_{\mathfrak{b}}^{(-)}(E)i\alpha_n^{(-)}\beta^{(+)}\Psi(E,\boldsymbol{\rho}). \tag{154}$$

Substituting this relation into the right-hand side of Eq. (152), making use of Eq. (151) and of the relations

$$\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(\pm)}(E)=-\alpha_n^{(\pm)}\hat{\mathfrak{b}}^{(\mp)}\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(\mp)}(E)\hat{\mathfrak{b}}^{(\mp)}\alpha_n^{(\mp)} \tag{155}$$

which stem from the explicit form (125) of the kernels $\bar{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E,\boldsymbol{\rho},\boldsymbol{\rho}')$, the conditions (110) and the Hermicity of the operators $\hat{\mathfrak{b}}^{(\pm)}$ [please notice that the relations analogous to Eq. (155) but with $\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(\pm)}(E)$ replaced by $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ do not hold], utilizing the properties of the matrices $\alpha_n^{(\pm)}(\boldsymbol{\rho})$ and $\beta^{(\pm)}$ and employing Eq. (88) transforms the right-hand side of Eq. (152) to $\beta^{(+)}\Psi(E,\boldsymbol{\rho})$. Similarly, determining $\beta^{(+)}\Psi(E,\boldsymbol{\rho})$ from Eq. (152)

$$\beta^{(+)}\Psi(E,\boldsymbol{\rho})=-\gamma^{(-)}\beta^{(+)}[\hat{\mathscr{T}}+\hat{\mathscr{R}}_{\mathfrak{b}}^{(+)}(E)\hat{\mathfrak{b}}^{(+)}]^{-1}\hat{\mathscr{R}}_{\mathfrak{b}}^{(+)}(E)i\alpha_n^{(+)}\beta^{(-)}\Psi(E,\boldsymbol{\rho}) \tag{156}$$

and substituting the result to the right-hand side of Eq. (153) transforms the latter, after some manipulations, to $\beta^{(-)}\Psi(E,\boldsymbol{\rho})$. Since in both cases the final results obtained are identical with the left-hand sides of the respective equations, one is led to the conclusion that Eqs. (152) and (153) are indeed consistent.

Given the relation (151) between the operators $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ and $\hat{\bar{\mathscr{R}}}_{\mathfrak{b}}^{(+)}(E)$, one may find the matrix representation of the operators $\hat{\mathscr{R}}_{\mathfrak{b}}^{(\pm)}(E)$ in the surface basis $\{\Phi_i(\boldsymbol{\rho})\}$. Projecting Eq. (151) from the left and from the right onto the spinor functions from the set $\{\Phi_i(\boldsymbol{\rho})\}$ one obtains

$$\mathsf{R}_{\mathfrak{b}}^{(\pm)}(E)=\frac{\hbar^2}{2m}\sum_k\frac{\mathsf{P}_{\mathfrak{b}k}^{(\pm)}\mathsf{P}_{\mathfrak{b}k}^{(\pm)\dagger}}{E_{\mathfrak{b}k}-E}-\frac{\mathfrak{b}^{(\pm)}}{(\gamma^{(\mp)})^2\mathsf{I}+(\mathfrak{b}^{(\pm)})^2}, \tag{157}$$

where $\{\mathsf{P}_{\mathfrak{b}k}^{(\pm)}\}$ are column vectors with elements $\{P_{i,\mathfrak{b}k}^{(\pm)}=(\Phi_i|\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k})\}$, $\{\mathsf{P}_{\mathfrak{b}k}^{(\pm)\dagger}\}$ are row vectors with elements $\{P_{i,\mathfrak{b}k}^{(\pm)*}=(\beta^{(\pm)}\Psi_{\hat{\mathfrak{b}}k}|\Phi_i)\}$, $\mathsf{b}^{(\pm)}$ are square matrices with elements $\{b_{ij}^{(\pm)}=(\Phi_i|\hat{\mathfrak{b}}^{(\pm)}\Phi_j)\}$ and I is the matrix of the unit operator with elements $\{I_{ij}=\delta_{ij}\}$. The result (157) is to be used in Eq. (96).

Let us examine somewhat closer the right-hand side of Eq. (157). Its first constituent is the matrix representation of the operators $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}^{(\pm)}(E)$ in the basis $\{\Phi_i(\boldsymbol{\rho})\}$; its form is formally identical with that of the nonrelativistic matrix $\mathsf{R}_{\mathfrak{b}}(E)$ given by Eq. (61) (cf., however, Sec. IV B). To the contrary, the second term on the right-hand side of Eq. (157) does not have its counterpart in the nonrelativistic theory. The origin of the occurrence of this term is the structure of the Dirac equation which is the set of coupled first-order equations while the Schrödinger equation is the second-order equation. The additional term in the formula for the matrix $\mathsf{R}_{\mathfrak{b}}^{(+)}(E)$ vanishes in the event when $\gamma^{(+)}=0$ (i.e., $\gamma^{(-)}=\infty$, which corresponds to the nonrelativistic limit $c\to\infty$) and in the exceptional case $\mathsf{b}^{(+)}=0$. With regard to the matrix $\mathsf{R}_{\mathfrak{b}}^{(-)}(E)$, the second term in Eq. (157) vanishes in the case $\mathsf{b}^{(-)}=0$. It is to be mentioned that the particular case of Eq. (157) was obtained, in a different way, by Szmytkowski and Hinze[18] who considered a spherically symmetric volume $\mathscr{V}$ and used a very particular surface basis $\{\Phi_i(\boldsymbol{\rho})\}$.

Finally, it is to be noticed that once the operator $\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}$ has been determined [cf. Eq. (148)], the explicit forms of the relations (131) and (132) may be given. We shall not discuss the general case when the operators $\hat{\mathfrak{b}}^{(\pm)}$ are arbitrary [but still related by Eq. (88)] and consider instead what happens if $\gamma^{(+)}=0$, $\hat{\mathfrak{b}}^{(+)}=\hat{0}$ or $\hat{\mathfrak{b}}^{(-)}=\hat{0}$ (here $\hat{0}$ denotes the null operator).

The first case, $\gamma^{(+)}=0$, corresponds to the nonrelativistic limit $c\to\infty$. From Eq. (148) one finds

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=\beta^{(+)}\Rightarrow\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')=\delta^{(2)}(\boldsymbol{\rho}-\boldsymbol{\rho}')\beta^{(+)}\quad(c\to\infty)\tag{158}$$

and, after substitution to Eqs. (131) and (132),

$$\sum_k\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\Psi_{\hat{\mathfrak{b}}k}^{\dagger}(\mathbf{r})=\sum_k\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi_{\hat{\mathfrak{b}}k}^{\dagger}(\boldsymbol{\rho}')=\delta^{(3)}(\mathbf{r}-\boldsymbol{\rho}')\beta^{(+)}\quad(\mathbf{r}\in\mathscr{V},c\to\infty).\tag{159}$$

This result is in accord with the relation (61) found in the nonrelativistic case.

In the event when $\hat{\mathfrak{b}}^{(+)}=\hat{0}$, from the boundary condition (109) and the properties of the matrices $\alpha_n^{(\pm)}(\boldsymbol{\rho})$ one infers that the basis functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ possess the property

$$\beta^{(-)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})=0\quad(\hat{\mathfrak{b}}^{(+)}=\hat{0}),\tag{160}$$

i.e., their lower components vanish identically on $\mathscr{S}$. From Eq. (148) one finds

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=\beta^{(+)}\Rightarrow\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')=\delta^{(2)}(\boldsymbol{\rho}-\boldsymbol{\rho}')\beta^{(+)}\quad(\hat{\mathfrak{b}}^{(+)}=\hat{0})\tag{161}$$

and consequently

$$\sum_k\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\Psi_{\hat{\mathfrak{b}}k}^{\dagger}(\mathbf{r})=\sum_k\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi_{\hat{\mathfrak{b}}k}^{\dagger}(\boldsymbol{\rho}')=\delta^{(3)}(\mathbf{r}-\boldsymbol{\rho}')\beta^{(+)}\quad(\mathbf{r}\in\mathscr{V},\hat{\mathfrak{b}}^{(+)}=\hat{0}).\tag{162}$$

It is to be emphasized that although the relations (159) and (162) are seemingly identical, they refer to completely different situations. In the case of Eq. (159) the operator $\hat{\mathfrak{b}}^{(+)}$ is arbitrary and $c=\infty$ while in the case of Eq. (162) the operator $\hat{\mathfrak{b}}^{(+)}$ is the null operator and the speed of light c is finite.

It remains to consider the third special case, namely that when $\hat{\mathfrak{b}}^{(-)}=\hat{0}$. Now one finds that the upper components of the functions $\{\Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\}$ vanish identically on $\mathscr{S}$,

$$\beta^{(+)}\Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho})=0\quad(\hat{\mathfrak{b}}^{(-)}=\hat{0}),\tag{163}$$

and that

$$\hat{\mathscr{A}}_{\hat{\mathfrak{b}}}=\beta^{(-)}\Rightarrow\mathscr{A}_{\hat{\mathfrak{b}}}(\boldsymbol{\rho},\boldsymbol{\rho}')=\delta^{(2)}(\boldsymbol{\rho}-\boldsymbol{\rho}')\beta^{(-)}\quad(\hat{\mathfrak{b}}^{(-)}=\hat{0}) \tag{164}$$

and consequently

$$\sum_k \Psi_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\mathbf{r})=\sum_k \Psi_{\hat{\mathfrak{b}}k}(\mathbf{r})\Psi^{\dagger}_{\hat{\mathfrak{b}}k}(\boldsymbol{\rho}')=\delta^{(3)}(\mathbf{r}-\boldsymbol{\rho}')\beta^{(-)}\quad(\mathbf{r}\in\mathscr{V},\hat{\mathfrak{b}}^{(-)}=\hat{0}). \tag{165}$$

The reader should compare the relations (159), (162) and (165) with the closure relation (114).

V. CONCLUDING REMARKS

In this paper we have formulated Wigner's *R*-matrix theory in the language of integral operators rather than matrices. In the first part of the work, following the ideas of Nesbet,[12,13] we have shown that in the nonrelativistic theory for particles described by the Schrödinger equation the central role is played by the integral operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ which relates function values to normal derivatives on a surface $\mathscr{S}$ of a closed volume $\mathscr{V}$ inside which the function satisfies the Schrödinger equation at energy E. Matrix elements of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ between functions from an orthonormal scalar basis set spanning the surface $\mathscr{S}$ form an *R*-matrix $\mathsf{R}_{\mathfrak{b}}(E)$. The method due to Kapur and Peierls[1] and to Wigner,[2–4,6] has been used, with necessary modifications and improvements, to construct an integral kernel of the operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$.

In the second part of the work we have developed the operator formulation of Wigner's *R*-matrix theory for the Dirac equation. It has been found that in the relativistic theory a counterpart of the nonrelativistic operator $\hat{\mathscr{R}}_{\hat{\mathfrak{b}}}(E)$ is an integral operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ relating on the enclosing surface $\mathscr{S}$ values of upper and lower components of spinor wave functions satisfying in the volume $\mathscr{V}$ the Dirac equation at energy E. Besides the operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$, another operator $\hat{\mathscr{R}}^{(-)}_{\hat{\mathfrak{b}}}(E)$ arises in the relativistic theory in the natural way. By generalizing Wigner's method,[2–4,6] we have constructed explicit forms of the operators $\hat{\mathscr{R}}^{(\pm)}_{\hat{\mathfrak{b}}}(E)$. It has been shown that although there are apparent similarities between the nonrelativistic and relativistic theories, the explicit form of the operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ contains a component which does not have its counterpart in the nonrelativistic theory. This component vanishes in the nonrelativistic limit. This agrees with a result obtained earlier by Szmytkowski and Hinze[17,18] who found such a correction in the matrix representation of the operator $\hat{\mathscr{R}}^{(+)}_{\hat{\mathfrak{b}}}(E)$ in a particular case of the spherically symmetric volume $\mathscr{V}$ and for a very special choice of a spinor basis spanning the surface $\mathscr{S}$.

ACKNOWLEDGMENTS

I am grateful to Professor Cz. Szmytkowski for commenting on the manuscript. The work was supported in part by the Polish State Committee for Scientific Research under Grant No. 950/P03/97/12. Financial support rendered by the Alexander von Humboldt Foundation is also gratefully acknowledged.

[1] P. L. Kapur and R. Peierls, Proc. R. Soc. London, Ser. A **166**, 277 (1938).
[2] E. P. Wigner, Phys. Rev. **70**, 15 (1946).
[3] E. P. Wigner, Phys. Rev. **70**, 606 (1946).
[4] E. P. Wigner and L. Eisenbud, Phys. Rev. **72**, 29 (1947).
[5] E. P. Wigner, Phys. Rev. **73**, 1002 (1948).
[6] T. Teichmann and E. P. Wigner, Phys. Rev. **87**, 123 (1952).
[7] A. M. Lane and R. G. Thomas, Rev. Mod. Phys. **30**, 257 (1958).
[8] G. Breit, in *Handbuch der Physik*, edited by S. Flügge (Springer, Berlin, 1959), Vol. XLI/1 part C.
[9] P. G. Burke and W. D. Robb, Adv. At. Mol. Phys. **11**, 143 (1975).
[10] *Atomic and Molecular Processes: An R-matrix Approach*, edited by P. G. Burke and K. A. Berrington (IOP, Bristol, 1993).
[11] M. Aymar, C. H. Greene, and E. Luc-Koenig, Rev. Mod. Phys. **68**, 1015 (1996).
[12] R. K. Nesbet, Phys. Rev. B **30**, 4230 (1984).
[13] R. K. Nesbet, Phys. Rev. A **38**, 4955 (1988).
[14] R. Peierls, Proc. Cambridge Philos. Soc. **44**, 242 (1948).
[15] G. Goertzel, Phys. Rev. **73**, 1463 (1948).
[16] J.-J. Chang, J. Phys. B **8**, 2327 (1975).
[17] R. Szmytkowski and J. Hinze, J. Phys. B **29**, 761 (1996), erratum **29**, 3800 (1996).

[18] R. Szmytkowski and J. Hinze, J. Phys. A **29**, 6125 (1996).
[19] R. Szmytkowski, J. Phys. A **30**, 4413 (1997).
[20] L. I. Schiff, *Quantum Mechanics*, 3rd ed. (McGraw–Hill, New York, 1968).
[21] R. Szmytkowski, Phys. Rev. A **57**, 4351 (1998).
[22] P. H. Norrington and I. P. Grant, J. Phys. B **14**, L261 (1981).
[23] P. H. Norrington and I. P. Grant, J. Phys. B **20**, 4869 (1987).
[24] U. Thumm and D. W. Norcross, Phys. Rev. A **45**, 6349 (1992).

Extension of Bertrand's theorem and factorization of the radial Schrödinger equation

Zuo-Bing Wu and Jin-Yan Zeng
Department of Physics, Peking University, Beijing 100871, People's Republic of China

(Received 2 April 1998; accepted for publication 25 May 1998)

The Bertrand's theorem is extended, i.e., closed orbits still may exist for central potentials other than the power law Coulomb potential and isotropic harmonic oscillator. It is shown that for the combined potential $V(r)=W(r)+b/r^2(W(r)=ar^\nu)$, when (and only when) $W(r)$ is the Coulomb potential or isotropic harmonic oscillator, closed orbits still exist for suitable angular momentum. The correspondence between the closeness of classical orbits and the existence of raising and lowering operators derived from the factorization of the radial Schrödinger equation is investigated. © *1998 American Institute of Physics.*
[S0022-2488(98)00110-8]

I. INTRODUCTION

The orbit of a classical particle in a central field, due to the angular momentum conservation, must lie in a plane perpendicular to the angular momentum. However, the orbit is, in general, not closed. In classical mechanics there is a famous Bertrand's theorem,[1,2] which says that the only central forces that result in closed orbits for all bound particles are the inverse square law and Hooke's law. For the two central potentials, apart from the energy and angular momentum, there exist an additional conserved quantity, which implies a higher dynamical symmetry than the geometrical symmetry (space isotropy).

Over fifty years ago, Schrödinger introduced the factorization method[3] to treat the eigenvalue problem of a one-dimensional harmonic oscillator and the quantized energy eigenvalues of a harmonic oscillator are connected by energy raising and lowering operators. The factorization method was generalized in the supersymmetry quantum mechanics (SSQM)[4] with the help of the concept of supersymmetry and shape invariance. SSQM mainly focuses on the factorization of one-dimensional Schrödinger equation (including the radial Schrödinger equation of a particle in a central field)[5] and the relation of eigenstates between two quantum systems (supersymmetric partner).

It was shown[6,7] that only for the Coulomb potential and isotropic harmonic oscillator the radial Schrödinger equation can be factorized and both the angular momentum and energy raising and lowering operators can be constructed. This reminds us that there may exist a certain connection between the factorization of radial Schrödinger equation and the closeness of classical orbits. Careful examination shows that in the derivation of Bertrand's theorem and in Ref. 6, a power law central potential ($V(r)=ar^\nu$) was assumed. For such a power law central potential, the Bertrand's theorem does hold. However, more careful analysis shows that if the restriction of power law form central potential is relaxed, the Bertrand's theorem needs reexamination. In Sec. II, it is shown that for the combined type of central potential $V(r)=W(r)+b/r^2(W(r)=ar^\nu)$, when $W(r)$ is the Coulomb potential or isotropic harmonic potential, there still exists closed orbits for suitable angular momenta. The factorization of the corresponding radial Schrödinger equation and the connection between the closeness of classical orbits and the existence of raising and lowering operators are investigated in Sec. III. A brief summary is given in Sec. IV.

II. EXTENSION OF BERTRAND'S THEOREM

It will be shown that if the restriction of a pure power law central potential is relaxed, the Bertrand's theorem may be extended. In particular, when an additional central potential b/r^2 is added to $W(r)=ar^\nu$,

 © 1998 American Institute of Physics

$$V(r)=W(r)+b/r^2, \tag{1}$$

we will show that when and only when $W(r)$ is the Coulomb potential or isotropic harmonic potential, there still exists closed orbits for some suitable angular momenta.

The equations of motion of a particle in the central potential (1) is expressed as ($\mu=1$)

$$\begin{aligned} &\dot{r}=p_r, \quad \dot{\theta}=L/r^2, \\ &\dot{p}_r=L^2/r^3+f(r), \quad \dot{L}=0, \end{aligned} \tag{2}$$

where L is the angular momentum and $f(r)=-dV/dr=g(r)+2b/r^3$, $g(r)=-dW/dr=-a\nu r^{\nu-1}$. It was shown[2] that for a circular orbit at radius r_0 to be possible, the force must be attractive and

$$E=L^2/2r_0^2+V(r_0), \quad f(r_0)=-L^2/r_0^3. \tag{3}$$

To guarantee the stability of a circular orbit $r=r_0$, it is required[2] that the second derivative of the effective potential $U(r)=V(r)+L^2/2r^2$ must be positive, i.e.,

$$-f'(r_0)-3f(r_0)/r_0>0. \tag{4}$$

It is noted that the inclusion of the additional force $2b/r^3$ does not alter the stability condition (4), so we have

$$-g'(r_0)-3g(r_0)/r_0>0, \quad \text{i.e.,} \quad a\nu(\nu+2)>0. \tag{5}$$

Thus $\nu<-2$ or $\nu>0$ for $a>0$, and $-2<\nu<0$ for $a<0$. Since we are interested in the bounded motion, the case $\nu<-2$ for $a>0$ is excluded. Thus, we obtain the stability condition for circular orbits: $a\nu>0$, i.e., $g(r)<0$.

To seek the condition for closeness of an orbit, let us consider the orbit equation[2] for potential (1)

$$\frac{d^2u}{d\theta^2}+\left(1+\frac{2b}{L^2}\right)u=-\frac{1}{L^2}\frac{d}{du}W\left(\frac{1}{u}\right), \quad u=\frac{1}{r}. \tag{6}$$

Introducing $\kappa=\sqrt{1+2b/L^2}$, $\xi=\kappa u$ and $\phi=\kappa\theta$, we get

$$\frac{d^2\xi}{d\phi^2}+\xi=J(\xi), \quad J(\xi)\equiv-\frac{1}{L^2}\frac{d}{d\xi}W\left(\frac{\kappa}{\xi}\right). \tag{7}$$

Using (3), it can be shown that at $\xi=\xi_0=\kappa/r_0$, $J(\xi_0)=\xi_0$. In addition, using (5) and $g(r)<0$, one gets

$$\beta^2\equiv 1-J'(\xi_0)=3+\frac{r}{g}\frac{dg}{dr}\bigg|_{r=r_0}>0,$$

then $\beta^2=\nu+2$. Thus, the stability condition for circular orbits also may be expressed as: $\beta^2>2$ for $a>0$, and $2>\beta^2>0$ for $a<0$.

Using the Taylor expansion of $J(\xi)$ at ξ_0, a deviation of an orbit from a circle ξ_0, $\delta\xi=\xi-\xi_0$, is governed by

$$\frac{d^2\delta\xi}{d\phi^2}+\delta\xi=J'(\xi_0)\delta\xi+\frac{1}{2}J''(\xi_0)(\delta\xi)^2+\frac{1}{6}J'''(\xi_0)(\delta\xi)^3+O((\delta\xi)^4). \tag{8}$$

For a small deviation of an orbit from circularity, to the first order of $\delta\xi$, Eq. (8) is reduced to

$$\frac{d^2\delta\xi}{d\phi^2}+\beta^2\delta\xi=0. \tag{9}$$

The solution to (9) can be expressed as

$$\delta\xi = a_1 \cos(\beta\phi), \quad \text{i.e.,} \quad \delta u = \frac{a_1}{\kappa}\cos(\beta\kappa\theta). \tag{10}$$

It can be seen that for $\beta\kappa = q/p$ (a rational number, q and p are integers), when $\theta = \theta_0 + 2\pi p$, $\delta u = \delta u_0$, i.e., the orbit remains closed. Thus, we get the condition for closeness of orbits against a small deviation from circularity: $\beta\kappa$ must be a rational number.

To seek the condition for a closed orbit when the deviation from circularity is considerable, $\delta\xi$ can be expressed as the Fourier expansion, i.e., in addition to the fundamental term in (10), more terms in the Fourier expansion should be included in $\delta\xi$. Substituting $\delta\xi$ in (8) and using the arguments similar to those in Ref. 2, one gets $\beta^2 = 0$, 1 and 4 for closed orbits. Thus, we obtain the conclusion: when (and only when) $\beta^2 = 1(\nu = -1)$ for $a < 0$, and $\beta^2 = 4(\nu = 2)$ for $a > 0$, the orbit is closed. Then, κ must be a rational number. Therefore, for the potential (1), when (and only when) $W(r)$ is the Coulomb potential or isotropic harmonic oscillator, closed orbits still exist for suitable angular momentum $L = \sqrt{2b/(\kappa^2 - 1)}$, where κ is a rational number.

Example: For the valence electron in an alkalis atom $V(r)$ may be approximately expressed as $(e = \mu = \hbar = 1)$[8]

$$V(r) = -\frac{1}{r} - \frac{\lambda}{r^2} \quad (0 < \lambda \ll 1). \tag{11}$$

In this case, the orbit equation (6) is reduced to

$$d\theta = -du/\sqrt{2E/L^2 + 2u/L^2 - \kappa^2 u^2}, \tag{12}$$

where $\kappa = \sqrt{1 - 2\lambda/L^2}$. Integrating (12), we get

$$u = \frac{1}{r} = \frac{1}{L^2\kappa^2}\left[1 + \sqrt{1 + 2EL^2\kappa^2}\cos\kappa(\theta - \theta_0)\right], \tag{13}$$

where $\sqrt{1 + 2EL^2\kappa^2} = \sqrt{1 + 2E(L^2 - 2\lambda)} \geqslant 0$ to make the solution (13) having sense. For $1 + 2EL^2\kappa^2 = 0$, the orbits are circular. Because $\kappa < 1$ for $\lambda > 0$, no elliptic orbits exist. However, when κ is a rational number, closed orbits (other than ellipses) still exist. Some illustrative examples are displayed in Fig. 1. In Fig. 1(a), $\kappa = 1/2(L = 2/3\sqrt{6\lambda})$, Fig. 1(b), $\kappa = 2/3(L = 3/5\sqrt{10\lambda})$ and Fig. 1(c), $\kappa = 3/4(L = 4/7\sqrt{14\lambda})$, there exists closed orbits around the center with period 2, 3 and 4, respectively. The geometry of closed orbits is dependent on $\kappa(L)$, but irrelevant to the energy E.

III. FACTORIZATION OF THE RADIAL SCHRÖDINGER EQUATION

For a particle in the central field $V(r)$, the energy eigenfunction may be chosen as the simultaneous eigenstate of the complete set of conserved observables $(H, \mathbf{l}^2, l_z)$, and the radial wave function $R_{n,l}(r) = \chi_{n,l}(r)/r$, ($l$ are nonnegative integers) satisfies the radial Schrödinger equation $(\hbar = \mu = 1)$

$$\begin{aligned} H\chi_{n,l}(r) &= E_n\chi_{n,l}(r), \\ H &= -\tfrac{1}{2}\, d^2/dr^2 + l(l+1)/2r^2 + V(r). \end{aligned} \tag{14}$$

For the potential (1), (14) is reduced to

$$\left[\frac{d^2}{dr^2} - \frac{l(l+1)}{r^2} - 2W(r) - \frac{2b}{r^2}\right]\chi_{n,l}(r) = -2E_n\chi_{n,l}(r). \tag{15}$$

Let $l'(l'+1) = l(l+1) + 2b$, i.e., $l' = -1/2 + (l + 1/2)\sqrt{1 + 2b/(l+1/2)^2}$, (15) can be rewritten as

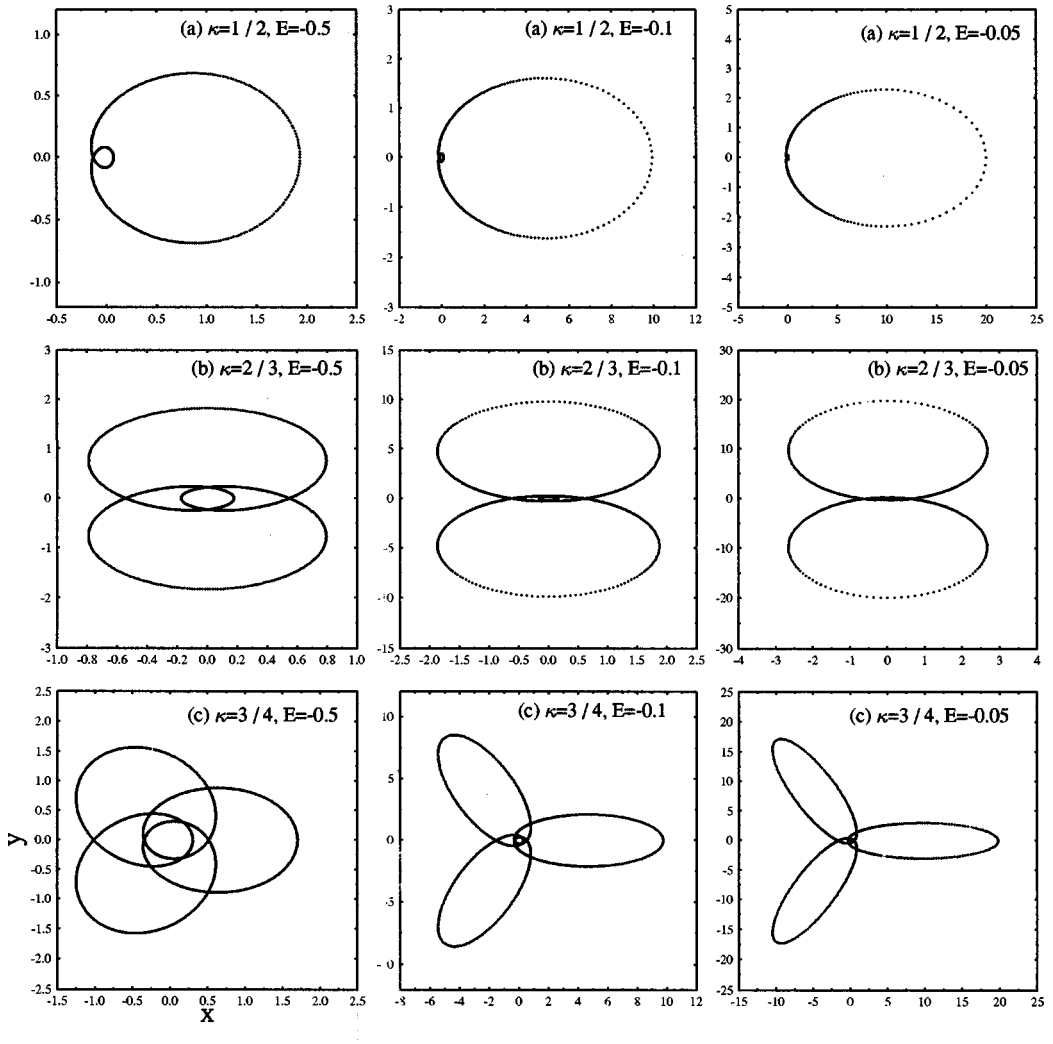

FIG. 1. The closed orbits of the valence electron in an alkalis atom ($\lambda = 0.2$).

$$\mathscr{D}_{l'}(r)\chi_{n,l'}(r) = -2E_n\chi_{n,l'}(r),$$
$$\mathscr{D}_{l'}(r) = d^2/dr^2 - l'(l'+1)/r^2 - 2W(r), \tag{16}$$

which has the same form as the usual radial Schrödinger equation with the central potential $W(r)$, but the nonnegative integer quantum number l is replaced by a noninteger l'. It has been shown[6,7] that if the power law central potential $W(r)$ is assumed, only for the Coulomb potential or isotropic harmonic oscillator the radial Schrödinger equation can be factorized to derive the raising and lowering operators of $l'(\Delta l' = \pm 1)$. However, for $\Delta l' = \pm 1$, the shift in l is by no means ± 1. Therefore, though the radial Schrödinger equation can be factorized for the central potential (1), no angular momentum raising and lowering operators ($\Delta l = \pm 1$) can be constructed, which is intimately connected with the breaking of the dynamical symmetry of Coulomb potential and isotropic harmonic oscillator due to the added term b/r^2 in (1).

For the Coulomb potential and isotropic harmonic oscillator, whose exact eigenvalues are well known, the energy raising and lowering operators have been derived by using different approaches in Refs. 6 and 9. For a particle in a more general central potential, the exact energy eigenvalues are usually unknown. To derive the energy raising and lowering operators, we may use the WKB approximation. According to the WKB approximation, the eigenvalues of bounded states of (16) take the form[10] (note: here the usual angular momentum l has been replaced by l')

$$E_n=\alpha_\nu n^{2\nu/(\nu+2)}, \tag{17}$$

where

$$\alpha_\nu=a^{2/(\nu+2)}2^{\nu/(\nu+2)}\left[\nu\sqrt{\pi}\,\frac{\Gamma\left(\frac{1}{\nu}+\frac{3}{2}\right)}{\Gamma\left(\frac{1}{\nu}\right)}\right]^{2\nu/(\nu+2)}>0, n=n_r+\frac{l'}{2}+\frac{3}{4}\ \text{for}\ a>0,$$

and

$$\alpha_\nu=-(-a)^{2/(\nu+2)}2^{\nu/(\nu+2)}\left[(-\nu)a\sqrt{\pi}\,\frac{\Gamma\left(1-\frac{1}{\nu}\right)}{\Gamma\left(-\frac{1}{2}-\frac{1}{\nu}\right)}\right]^{2\nu/(\nu+2)}<0,$$

$$n=n_r+\frac{2l'+\nu+3}{2(\nu+2)}\ \text{for}\ a<0.$$

Thus, (16) may be recast into

$$\begin{aligned}&\mathscr{D}_n(r)\chi_{n,l'}(r)=l'(l'+1)\chi_{n,l'}(r),\\ &\mathscr{D}_n(r)=r^2d^2/dr^2-2ar^{\nu+2}+2\alpha_\nu n^{2\nu/(\nu+2)}r^2.\end{aligned} \tag{18}$$

(1) For $a>0$, the factorizability of (18) requires

$$(rd/dr-\sqrt{2a}r^{\nu/2+1}+A)(rd/dr+\sqrt{2a}r^{\nu/2+1}+B)\chi_{n,l'}(r)=\mathscr{D}_n(r)\chi_{n,l'}(r)+AB\chi_{n,l'}(r), \tag{19}$$

then we have

$$\begin{aligned}&A+B+1=0,\\ &\sqrt{2a}r^{\nu/2+1}(A-B+\nu/2+1)=2\alpha_\nu n^{2\nu/(\nu+2)}r^2.\end{aligned} \tag{20}$$

The solution to (20) is $\nu=2$, $\alpha_2=2\sqrt{2a}$, $A=2n-3/2$, and $B=-(2n-1/2)$. Thus, (19) is reduced to

$$\begin{aligned}&[rd/dr-\sqrt{2a}r^2+2n-3/2][rd/dr+\sqrt{2a}r^2-(2n-1/2)]\chi_{n,l'}(r)\\ &\quad=\mathscr{D}_n(r)\chi_{n,l'}(r)-(2n-3/2)(2n-1/2)\chi_{n,l'}(r).\end{aligned} \tag{21}$$

Similarly, (18) also can be factorized as

$$\begin{aligned}&[rd/dr+\sqrt{2a}r^2-(2n+3/2)][rd/dr-\sqrt{2a}r^2+(2n+1/2)]\chi_{n,l'}(r)\\ &\quad=\mathscr{D}_n(r)\chi_{n,l'}(r)-(2n+1/2)(2n+3/2)\chi_{n,l'}(r).\end{aligned} \tag{22}$$

Using (21) and (22), one may obtain the energy raising and lowering operators

$$\begin{aligned}&\chi_{n+2,l'}(r)\sim[rd/dr-\sqrt{2a}r^2+(2n+1/2)]\chi_{n,l'}(r),\\ &\chi_{n-2,l'}(r)\sim[rd/dr+\sqrt{2a}r^2-(2n-1/2)]\chi_{n,l'}(r).\end{aligned} \tag{23}$$

(2) For $a<0$, the factorizability of (18) requires

$$(rd/dr-\sqrt{-2\alpha_\nu}n^{\nu/(\nu+2)}r+A)(rd/dr+\sqrt{-2\alpha_\nu}n^{\nu/(\nu+2)}r+B)\chi_{n,l'}(r)$$
$$=\mathscr{D}_n(r)\chi_{n,l'}(r)+AB\chi_{n,l'}(r), \tag{24}$$

then we have

$$\begin{aligned} A+B+1&=0,\\ \sqrt{-2\alpha_\nu}n^{\nu/(\nu+2)}r(A-B+1)&=-2ar^{\nu+2}. \end{aligned} \tag{25}$$

The solution to (25) is $\nu=-1$, $\alpha_{-1}=-a^2/2$, $A=n-1$, and $B=-n$. Thus, (24) is reduced to

$$[rd/dr+ar/n+n-1][rd/dr-ar/n-n]\chi_{n,l'}(r)=\mathscr{D}_n(r)\chi_{n,l'}(r)-(n-1)n\chi_{n,l'}(r). \tag{26}$$

Similarly, (18) also can be factorized as

$$[rd/dr-ar/n-(n+1)][rd/dr+ar/n+n]\chi_{n,l'}(r)=\mathscr{D}_n(r)\chi_{n,l'}(r)-n(n+1)\chi_{n,l'}(r). \tag{27}$$

Using (26) and (27) and introducing the scaling operator $\mathscr{M}(k)$ defined by $\mathscr{M}(k)f(r)=f(kr)$, one obtains the energy raising and lowering operators

$$\begin{aligned} \chi_{n+1,l'}(r)&\sim\mathscr{M}\left(\frac{n}{n+1}\right)(rd/dr+ar/n+n)\chi_{n,l'}(r)=[rd/dr+ar/(n+1)+n]\mathscr{M}\left(\frac{n}{n+1}\right)\chi_{n,l'}(r),\\ \chi_{n-1,l'}(r)&\sim\mathscr{M}\left(\frac{n}{n-1}\right)(rd/dr-ar/n-n)\chi_{n,l'}(r)=[rd/dr-ar/(n-1)-n]\mathscr{M}\left(\frac{n}{n-1}\right)\chi_{n,l'}(r). \end{aligned} \tag{28}$$

Thus, we arrive at the conclusion that, for the potential (1), only when $W(r)$ is the Coulomb potential or isotropic harmonic oscillator, the radial Schrödinger equation can be factorized to derive the energy raising and lowering operators.

For a classical particle in the Coulomb potential $W(r)=-1/r$, the orbit is always closed for any negative energy ($E<0$) and positive angular momentum L, i.e., an ellipse, of which the length of semimajor axis is $a=1/2|E|$ and the eccentricity $e=\sqrt{1-2|E|L^2}$. This is guaranteed by the existence of an additional conserved quantity—the Runge–Lenz vector $\mathbf{R}=\mathbf{p}\times\mathbf{L}-\mathbf{r}/r$, which determines the direction of semimajor axis and the eccentricity $e=|R|$. In quantum mechanics, the angular momentum and bound energy are quantized. Therefore, it is not surprising that for the Coulomb potential the radial Schrödinger equation can be factorized and there exist four kinds of raising and lowering operators connecting various eigenstates of energy and angular momentum.[6]

However, for the combined central potential (11), the dynamical symmetry O_4 is broken and $\mathbf{R}$ no longer remains constant, and the orbit, in general, is not closed. It has been shown in Sec. II that for suitable angular momenta $L=\sqrt{2\lambda/(1-\kappa^2)}$ (κ is a rational number), the orbits are still closed (but not ellipses). Therefore, it is understandable why no angular momentum raising and lowering operators ($\Delta l=\pm 1$) exist, but energy raising and lowering operators for a fixed l' [see (28)] still can be constructed. In fact, for $\lambda=0$, (28) is just the raising and lowering operators $B(l,n\uparrow)$ and $B(l,n\downarrow)$ given in Eq. (33) of Ref. 6. It is noted that the angular momentum raising and lowering operators A, C and D in Ref. 6 no longer exist for the combined potential (11). For the isotropic harmonic oscillator, the situation is similar.

IV. SUMMARY

The Bertrand's theorem is extended in this paper. For a classical particle in a power law central potential $W(r)$, the Bertrand's theorem does hold, i.e., only for the Coulomb potential or isotropic harmonic oscillator, closed orbits always exist for continuous energy and angular momentum. Accordingly, in quantum mechanics there exist both energy and angular momentum raising and lowering operators connecting neighboring discrete energy and angular momentum eigenstates. For the combined potential (1), it is shown that when $W(r)$ is the Coulomb potential or isotropic harmonic oscillator, classical closed orbits (other than ellipse) still exist for continuous

energy, but only for suitable angular momenta. From this one can understand why in this case there exist only energy raising and lowering operators in quantum mechanics, but no angular momentum raising and lowering operators.

ACKNOWLEDGMENTS

This work was supported in part by the Postdoctoral Foundation of China (Z.B.W.) and the National Natural Science Foundation of China (J.Y.Z.).

[1] J. Bertrand, Comptes Rendus **77**, 849 (1873).
[2] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, New York, 1980), sec. 3.5 and Appendix A.
[3] E. Schrödinger, Proc. R. Ir. Acad. Sect. A, Math. Astron. Phys. Sci. **46**, 183 (1940), **47**, 53 (1942).
[4] For review papers, see F. Cooper, A. Khare, and U. P. Sukhatme, Phys. Rep. **251**, 1 (1995); R. Dutt, A. Khare, and U. P. Sukhatme, Am. J. Phys. **56**, 163 (1988).
[5] R. W. Haymaker and A. R. P. Rau, Am. J. Phys. **54**, 928 (1986); V. A. Kostelecky and M. M. Nieto, Phys. Rev. Lett. **53**, 2285 (1984).
[6] Y. F. Liu, Y. A. Lei, and J. Y. Zeng, Phys. Lett. A **231**, 9 (1997).
[7] J. D. Newmarch and R. M. Golding, Am. J. Phys. **46**, 658 (1978).
[8] L. D. Landau and M. E. Lifshitz, *Quantum Mechanics, Nonrelativistic Theory* (Pergamon, Oxford, 1977); B. C. Qian and J. Y. Zeng, *Problem in Quantum Mechanics* (Science Press, Beijing, 1988).
[9] O. L. DeLange and R. E. Raab, *Operator Methods in Quantum Mechanics* (Clarendon, Oxford, 1991).
[10] C. Quigg and J. L. Rosner, Phys. Rep. **56**, 167 (1979).

A constructive index theorem, phase operators, and phase-difference operators

Sixia Yu[a)]
Atominstitut der Österreichischen Universitäten Stadionallee 2 A-1020, Vienna, Austria

Yongde Zhang
Department of Modern Physics, University of Science and Technology of China, Hefei, Anhui, 230027, People's Republic of China

(Received 17 November 1997; accepted for publication 23 January 1998)

A constructive reversed index theorem concerning the polar decomposition of an operator into a product of a unitary exponential phase operator and a Hermitian amplitude operator is investigated. Its applications to the construction of Hermitian phase operators of fermions and phase-difference operators between bosons or fermions are presented. Specifically, a Hermitian operator is constructed for the phase difference between a single-mode fermion and boson, which is justified by the fact that it is exactly the interaction part of the Hamiltonian of the Jaynes-Cummings Model. Furthermore, a Hermitian phase operator of a single-mode boson can also be defined referring to a single-mode fermion. All those quantized phases and phase-differences are found to obey a quantum addition rule instead of the ordinary commutative addition rule. © *1998 American Institute of Physics.* [S0022-2488(98)00107-8]

I. INTRODUCTION

Phase is one of the most important and perplexing aspects in quantum physics. The concept of phase is most important because, experimentally, it plays an essential role in the interference phenomena, through which many quantum properties can be revealed and, theoretically, it leads to the gauge theory. It is also a most perplexing aspect because it enters the experiments merely as a parameter without a quantum description. The nonexistence of a proper operator for a simple quantum system such as a single harmonic oscillator or single-mode bosonic field was known by Dirac seventy years ago.[1]

The difficulty of quantizing the phase variable of a quantum harmonic oscillator can most elegantly be illustrated by an index theorem. The index[2] of an operator O in a Hilbert space is defined as

$$\mathrm{index}(O)\equiv \dim \ker(O)-\dim \ker(O^\dagger), \tag{1}$$

in which $\dim \ker(O)$ denotes the dimension of the kernel space $\ker(O)=\{|\psi\rangle|O|\psi\rangle=0\}$ of the operator O. Because of the relation $\ker(O)=\ker(O^\dagger O)$, the index can equivalently be defined by

$$\mathrm{index}(O)\equiv \dim \ker(O^\dagger O)-\dim \ker(OO^\dagger). \tag{2}$$

This index of an operator is invariant under a very general deformation of the Hilbert space. For two arbitrary reversible coordinate transformations P and Q in the Hilbert space of the system, the index value cannot be changed due to the relation

$$\dim \ker(POQ)=\dim \ker(O). \tag{3}$$

[a)]Current address: Institute of Theoretical Physics, Academia Sinica, Beijing 100080, P.R. China.

 © 1998 American Institute of Physics

If an operator O has a polar decomposition, i.e., it can be decomposed into a product $O=\mathscr{E}A$ (or $O=A\mathscr{E}$) of a unitary exponential phase operator $\mathscr{E}$ and a Hermitian amplitude operator A, then its index according to the second definition, Eq. (2),

$$\text{index}(O)=\dim\ker(A^2)-\dim\ker(\mathscr{E}A^2\mathscr{E}^\dagger), \tag{4}$$

vanishes because of Eq. (3). In other words, if an operator does not have a zero index, it cannot have a polar decomposition as long as it is restricted to the original Hilbert space. This may be referred to as a no-go index theorem.[3]

As is well known, in the Hilbert space $\mathscr{H}_b$ of a single harmonic oscillator or a single-mode bosonic field described by ladder operators $b^\dagger$ and b satisfying canonical commutation relation $[b,b^\dagger]=1$, there is a state, vacuum state that is annihilated by the destruction operator while there is no state that can be destructed by the creation operator. As a result, according to the first definition, Eq. (1), we have an index condition[2–4]

$$\text{index}(b)\equiv\dim\ker(b)-\dim\ker(b^\dagger)=1. \tag{5}$$

Therefore a Hermitian phase operator defined by the polar decomposition is unattainable in the original Hilbert space $\mathscr{H}_b$. However, when the Hilbert space is modified so that the operator carrying the information of phase has a vanishing index, a Hermitian operator can exist for the description of phase properties because of a reversed index theorem discussed in the next section.

As the first application, the phase difference operators for bosons are constructed and a quantum addition rule of quantized phase differences is introduced in Sec. III. In Sec. IV its applications to the constructions of phase and phase difference operators for fermions are discussed. Finally, in Sec. V a phase difference operator is constructed to describe the phase difference between a single-mode boson and fermion.

II. REVERSED INDEX THEOREM

Generally, for an operator O in a certain Hilbert space, if the following commutation relation holds (which is satisfied by all the operators concerned later on)

$$[O^\dagger O,OO^\dagger]=0, \tag{6}$$

then it can always be decomposed into a product $O=EA$ of a Hermitian amplitude operator $A=\sqrt{O^\dagger O}$ as a function of the operator $O^\dagger O$ and a *nonunitary* exponential phase operator E satisfying

$$E^\dagger E=1-P_O,\quad EE^\dagger=1-P_{O^\dagger}, \tag{7}$$

with P_O and $P_{O^\dagger}$ being the projectors on the kernel space $\ker(O)$ and $\ker(O^\dagger)$, respectively. Outside the kernel space $\ker(O)$ the nonunitary exponential phase operator can be written as $E=OA^{-1}$, which is supposed to be identically zero inside the kernel space $\ker(O)$. For example, the annihilation operator b of a quantum harmonic oscillator can be decomposed into a product of nonunitary exponential phase operator $E=b(\sqrt{N})^{-1}$ (Ref. 5) and a Hermitian amplitude operator $A=\sqrt{N}$. If the operator O possesses a zero index, however, it can have unitary polar decompositions.

The vanishing index means that the kernel space $\ker(O)$ has the same dimension as the kernel space of its Hermitian conjugate $\ker(O^\dagger)$ according to the definition, Eq. (1). Since we are mainly dealing with separable Hilbert space with a countable number of bases, these two kernel spaces are also separable Hilbert spaces. As a result they can be in a one-to-one correspondence with each other. In this sense the kernel spaces are not necessarily restricted to finite-dimensional spaces. The orthonormal bases of these two kernel space are denoted as $\{|\phi_i\rangle|i\in S\}$ for $\ker(O)$ and $\{|\psi_i\rangle|i\in S\}$ for $\ker(O^\dagger)$, respectively, with S being a subset of the natural number set, finite or infinite. An isomorphism $Y:\ker(O)\mapsto\ker(O^\dagger)$ can be established as

$$Y=\sum_{i\in S}|\psi_i\rangle\langle\phi_i|, \tag{8}$$

which is reversible and its reverse is exactly its Hermitian conjugate because we have $YY^\dagger = P_{O^\dagger}$ and $Y^\dagger Y = P_O$ where the projectors, i.e., the identity operators, on those two kernel spaces are

$$P_O = \sum_{i \in S} |\phi_i\rangle\langle\phi_i|, \quad P_{O^\dagger} = \sum_{i \in S} |\psi_i\rangle\langle\psi_i|. \tag{9}$$

Now we are ready for the reversed index theorem: It is possible to decompose an operator O with vanishing index satisfying a commutation relation, Eq. (6), into a product of a unitary operator and a Hermitian operator.

This is because relation Eq. (6) ensures a nonunitary exponential phase operator E and the vanishing index ensures an isomorphism Y between kernel spaces $\ker(O)$ and $\ker(O^\dagger)$ and operator

$$\mathscr{E} = E + Y, \tag{10}$$

defined in the whole Hilbert space, is unitary and satisfies the polar decomposition $O = \mathscr{E}A$. Therefore there exists a Hermitian phase operator defined by the exponential function.

This polar decomposition depends obviously on the choice of the isomorphism Y or the orthonormal bases of kernel spaces $\ker(O)$ and $\ker(O^\dagger)$. Therefore the Hermitian phase operator is not unique which depends on the experimental setup of measuring the Hermitian phase operator.[6] This degree of freedom is somewhat similar to the gauge freedom and the isomorphism plays a role of connection to make the comparison of phases possible.

In the following sections, as the applications of the above reversed index theorem, the Hermitian phase difference operators between boson and boson, fermion and fermion, boson and fermion are constructed, which enable a complete quantum description of the phase variable.

III. PHASE-DIFFERENCE OPERATORS OF BOSONS

As shown in the introductory section, the no-go index theorem excludes a Hermitian phase operator for a single-mode boson or a single harmonic oscillator in its original Hilbert space. Many efforts have been tried to solve this problem by introducing another degree of freedom[7,8] or truncating the original Hilbert space to a finite-dimensional one.[9] However, these additional structures are rather artificial so they are not a satisfactory quantum description of phase.[10]

Taking account of the fact that what is physical and measurable is only the difference of an observable, a natural way to circumvent the no-go index theorem, however, is to consider the phase *difference* between two bosons,[11,12] which are supposed to be described by ladder operators $\{b_i, b_i^\dagger | i = 1,2\}$ satisfying canonical commutation rules $(i,j = 1,2)$,

$$[b_i, b_j^\dagger] = \delta_{ij}, \quad [b_i, b_j] = [b_i^\dagger, b_j^\dagger] = 0.$$

The Hilbert space $\mathscr{H}_{2b}$ is spanned by the common eigenstates $|n_1, n_2\rangle$ $(n_1, n_2 = 0,1,2,\ldots)$ of the number operators $N_i = b_i^\dagger b_i$ $(i = 1,2)$. From the correspondence between classical and quantum physics, the information of the phase difference between these two modes of boson is carried by operator $J_+ = b_1^\dagger b_2$. In the Hilbert space $\mathscr{H}_{2b}$ both the kernel space $\ker(J_+)$ spanned by bases $|n,0\rangle$ $(n = 0,1,2,\ldots)$ and the kernel space $\ker(J_+^\dagger)$ spanned by bases $|0,n\rangle$ $(n = 0,1,2,\ldots)$ are countably infinite-dimensional Hilbert spaces and they can be in a one-to-one correspondence with each other. Therefore we have a vanishing index,

$$\mathrm{index}(J_+) = 0. \tag{11}$$

Thus a polar decomposition of operator J_+ is possible according to the reversed index theorem discussed above and a Hermitian phase difference operator is attainable. The most general isomorphism $Y_{12}: \ker(J_+) \mapsto \ker(J_+^\dagger)$ is given by

$$Y_{12} = U_{12}\, P_{10}\, X_{12}, \tag{12}$$

where U_{12}, called the bosonic connection operator, is an arbitrary unitary operator commuting with number operator N_1, $P_{10}=|0\rangle_1\langle 0|_1$ is the vacuum projector of the first mode, and X_{12} is a linear quantum transformation,[13]

$$X_{12}=e^{i(\pi/2)(b_1^\dagger b_2+b_2^\dagger b_1)}, \tag{13}$$

which exchanges the Hilbert spaces of two modes of the boson, $X_{12}b_1=b_2X_{12}$, for example. Unlike the case of a single harmonic oscillator, where the Hermitian phase operator defined by polar decomposition does not exist, there is an infinite number of Hermitian phase-difference operators for two-mode boson.

The most important choice of the isomorphism Y_{12} is to require that it commutes with the total number operator N_1+N_2, which reduces the bosonic connection operator U_{12} in Eq. (12) to a diagonal matrix. Therefore the exponential phase difference operator is

$$\mathscr{E}_{12}=E_1^\dagger E_2+e^{i\phi_{12}(N_2)}\,P_{10}\,X_{12}, \tag{14}$$

where $E_i=b_i(\sqrt{N_i})^{-1}$ is the nonunitary phase operator for the ith mode ($i=1,2$) and the phase-difference function $\phi_{12}(x)$ is an arbitrary real function. The Hermitian phase-difference operator defined by the exponential function $\mathscr{E}_{12}=\exp\{i\mathscr{P}_{12}\}$, which is called the direct phase-difference operator, possesses a new type of eigenvalue, namely a countable and dense type.[12] There is a countably infinite number of eigenvalues, which shows its conjugation relationship with the number operator, and it is dense in a 2π window, which shows its periodic nature. The phase-difference states, eigenstates of the phase-difference operator $\mathscr{P}_{12}$ form an orthonormal and complete basis of the Hilbert space $\mathscr{H}_{2b}$.

Another important choice of the the isomorphism Y'_{12} leads to the indirect phase-difference operator. Classically, for three harmonic oscillators there are three direct phase differences Φ_{ij}. Trivially, we can have an indirect phase difference between first two oscillators by comparing them with the third oscillator, $\Phi'_{12}=\Phi_{13}+\Phi_{32}$, which is obviously equal to the direct phase difference Φ_{12} because of the identity

$$\Phi_{12}+\Phi_{23}+\Phi_{31}=0. \tag{15}$$

When three quantum oscillators are concerned, there are three different direct phase-difference operators $\mathscr{P}_{ij}$ $(i,j=1,2,3,i\neq j)$ which are not commuting with one another describing the quantum phase differences. The ordinary addition of two of them, $\mathscr{P}_{12}+\mathscr{P}_{23}$, for example, does not satisfy the polar decomposition and is therefore not a quantum phase difference at all. However, one can introduce a quantum addition rule[12] of two phase operators $\mathscr{P}_1$ and $\mathscr{P}_2$ according to

$$\mathscr{P}_1\dot{+}\mathscr{P}_2\equiv -i\,\ln\{e^{i\mathscr{P}_1}e^{i\mathscr{P}_2}\}\quad(\mathrm{mod}\,2\pi), \tag{16}$$

which reduces to the ordinary addition rule when these two operators are commuting and may not be commutative when they are not commuting with each other. The indirect phase-difference operator is defined as

$$\mathscr{P}'_{12}=\mathscr{P}_{13}\dot{+}\mathscr{P}_{32}. \tag{17}$$

Its exponential function, the indirect exponential phase-difference operator satisfies the polar decomposition, which corresponds to a choice of the bosonic connection operator,

$$U'_{12}=e^{i\{\phi_{12}(N_2)+\phi_{23}(N_3)\}}\,X_{23}. \tag{18}$$

Similar to the classical identity, Eq. (15), fulfilled by those three classical phase differences according to the normal addition rule, the quantum combinations of three quantized phase differences according to the quantum addition rule have the following properties $(i,j,k=1,2,3)$:

$$\mathrm{spec}(\mathscr{P}_{ij}\dot{+}\mathscr{P}_{jk}\dot{+}\mathscr{P}_{ki})=\{0,\pi\}, \tag{19a}$$

$$\mathrm{spec}(\mathscr{P}_{ij}\dot{+}\mathscr{P}_{ki}\dot{+}\mathscr{P}_{jk})=\left\{0,\frac{\pi}{2},\pi,\frac{3\pi}{2}\right\}, \tag{19b}$$

where $\mathrm{spec}(O)$ denotes the spectrum of the operator O and all those phase-difference functions ϕ_{ij} are taken as zero for simplicity. This is because for the exponential phase difference operators we have

$$(\mathscr{E}_{ij}\mathscr{E}_{jk}\mathscr{E}_{ki})^2=1, \quad (\mathscr{E}_{ij}\mathscr{E}_{ki}\mathscr{E}_{jk})^4=1,$$

considering the fact that the permutators X_{ij} $(i,j=1,2,3)$ defined as in Eq. (13) which are the generators of the permutation group P_3. The existence of additional nonzero eigenvalues in spectra, Eq. (19), is a purely quantum feature.

Generally for a K-mode boson let us consider K quantized phase difference $\mathscr{P}_{ij}$ in which the subscript of each mode appears exactly twice. When they are added up in such a way that their subscripts are arranged in a chainlike order as in Eq. (19a) we have the following spectrum:

$$\mathrm{spec}(\mathscr{P}_{\delta,K}\equiv\mathscr{P}_{12}\dot{+}\mathscr{P}_{23}\dot{+}\cdots\dot{+}\mathscr{P}_{K-1,K}\dot{+}\mathscr{P}_{K1})=\left\{\frac{2m\pi}{K-1}\middle| m=0,1,\ldots,K-2\right\}. \tag{20}$$

We may change the orders of those noncommuting phase-difference operators in the above spectrum to obtain another spectrum. Let us denote the resulting operator as $\mathscr{P}'_{\delta,K}$ after the rearrangement of the subscripts. In order to obtain a chainlike ordering of the subscripts again, let use suppose we need to exchange at *least* L times the phase-difference operators. Then the exponential phase-difference operator is found to satisfy the relation

$$\{e^{i\mathscr{P}'_{\delta,K}}\}^{(L+1)(K-1)}=1,$$

from which the spectrum of $\mathscr{P}'_{\delta,K}$ is obtained as

$$\frac{2m\pi}{(L+1)(K-1)}, \tag{21}$$

with $m=0,1,2,\ldots,(L+1)(K-1)-1$.

IV. PHASE AND PHASE-DIFFERENCE OPERATORS OF FERMIONS

The second application of the above reversed index theorem is to construct a Hermitian phase operator for a single-mode fermion and phase-difference operator between two modes of a fermion.

A single-mode fermion or a single fermionic oscillator can be described by fermionic creation and annihilation operators $f^\dagger$ and f satisfying anticommutator $\{f,f^\dagger\}=1$. Its Hilbert space $\mathscr{H}_f$ is spanned by the eigenstates $|\sigma\rangle$ $(\sigma=0,1)$ of the fermionic number operator $N_f=f^\dagger f$. The index of the fermionic annihilation operator has a vanishing index $(f)=0$ in Hilbert space $\mathscr{H}_f$. Therefore a polar decomposition of the annihilation operator f is possible and a Hermitian phase operator can be therefore well defined.

The most general isomorphism from a one-dimensional kernel space $\ker(f)$ spanned by the ground state $|0\rangle$ to another one-dimensional kernel space $\ker(f^\dagger)$ spanned by the exited state $|1\rangle$ is

$$Y_f=\exp(-i2\alpha)f^\dagger, \tag{22}$$

with α being an arbitrary real number that determines the limits of a 2π window of phase variable. Therefore the unitary exponential phase operator $\mathscr{E}_f$ and the Hermitian phase operator defined by the exponential function are, respectively,

$$\mathscr{E}_f=f+f^\dagger e^{-i2\alpha}=e^{-i\mathscr{P}_f}, \tag{23a}$$

$$\mathscr{P}_f=\alpha+\frac{\pi}{2}(1-e^{i\alpha}\mathscr{E}). \tag{23b}$$

This Hermitian phase operator can be and is always being measured in the interferometer. In fact the essential element of the interferometer is a single-mode fermion by identifying its two paths with two fermionic number states. The number operator carries the which-way information while the phase operator shows the interference pattern. The complementariness of the which-way information showing the particle property and interference pattern showing the wave property is just a manifestation of the noncommuting property of the phase operator and the number operator.[6]

Let us consider a two-mode fermion, which is described by ladder operators $\{f_i, f_i^\dagger | i=1,2\}$ satisfying anticommutation rules $(i,j=1,2)$,

$$\{f_i, f_j^\dagger\}=\delta_{ij}, \quad \{f_i, f_j\}=\{f_i^\dagger, f_j^\dagger\}=0.$$

Its Hilbert space $\mathscr{H}_{2f}$ is spanned by the common eigenstates of two fermionic number operators $N_{fi}=f_i^\dagger f_i$ $(i=1,2)$. Similar to the case of a two-mode boson, the phase difference is contained in the operator $f_1^\dagger f_2$, whose index is obviously zero in the Hilbert space $\mathscr{H}_{2f}$. Therefore we have polar decomposition,

$$f_1^\dagger f_2=\mathscr{E}_{12}^f N_2(1-N_1), \tag{24}$$

in which the unitary exponential phase-difference operator is $\mathscr{E}_{12}^f=f_1^\dagger f_2+Y_{12}^f$ with the isomorphism $Y_{12}^f:\ker(f_1^\dagger f_2)\mapsto\ker(f_2^\dagger f_1)$ given by

$$Y_{12}^f=U_{12}^f(f_1 f_2+f_2^\dagger), \tag{25}$$

where the fermionic connection operator U_{12}^f is an arbitrary 3×3 matrix belonging to a representation of group $U(3)$ in the kernel space $\ker(f_2^\dagger f_1)$.

There are two well defined phase operators $\mathscr{P}_1^f$ and $\mathscr{P}_2^f$ for two modes. As it turns out, the exponential function of their normal subtraction $\mathscr{P}_1^f-\mathscr{P}_2^f$ does not satisfy the polar decomposition, Eq. (24), as a phase difference factor should be. However, when they are summed up,

$$\mathscr{P}_{12}^f=\mathscr{P}_1^f \dot{-} \mathscr{P}_2^f \equiv \mathscr{P}_1^f \dot{+} (-\mathscr{P}_2^f), \tag{26}$$

according to the quantum addition rule, Eq. (16), its exponential function satisfies polar decomposition, Eq. (25), which corresponds to the following fermionic connection operator:

$$U_{12}^{f\prime}=e^{-2i\alpha_2}\mathscr{E}_{f1}N_{f2}+e^{2i\alpha_1}(1-N_{f1})(1-N_{f2}). \tag{27}$$

The eigenvalues of this fermionic phase-difference operator $\mathscr{P}_{12}^f$ are found to be $\alpha_1-\alpha_2 \pm \pi/2$ which are both two-fold degenerated. Those four fermionic phase-difference states form an orthonormal and complete basis of Hilbert space $\mathscr{H}_{2f}$. In the fermionic number state basis $\{|\sigma_1,\sigma_2\rangle | \sigma_1,\sigma_2=0,1\}$, by assuming that $\alpha_1=\pi$ and $\alpha_2=\pi/2$, one of the eigenstates corresponding to the eigenvalue π is

$$\frac{1}{\sqrt{2}}\{|0,1\rangle-|1,0\rangle\}, \tag{28}$$

which is exactly an EPR state of two fermions. This provides a different view point of an EPR entangled state: It is just one of the eigenstates of the phase difference operator of these two fermions or two-level systems.

When a three-mode fermion is concerned, there are three different quantum phase differences $\mathscr{P}_{12}^f$, $\mathscr{P}_{23}^f$, and $\mathscr{P}_{31}^f$ between each of two modes. When they are added up by using the quantum addition rule, we have identities

$$\mathscr{P}_{ij}^f \dot{+} \mathscr{P}_{jk}^f \dot{+} \mathscr{P}_{ki}^f = 0, \tag{29a}$$

$$\mathscr{P}_{ij}^f \dot{+} \mathscr{P}_{ki}^f \dot{+} \mathscr{P}_{jk}^f = \pi, \tag{29b}$$

with $i,j,k=1,2,3$. However, when they are added up by using the ordinary addition rule we cannot have similar identities. Similarly, for a finite-mode of a fermion, when the subscripts of the fermionic phase-difference operators are arranged in an even permutation of the chainlike order like in Eq. (29a), their quantum addition yields zero and when they are arranged in an odd permutation of the chainlike order, their quantum addition yields the additional phase π, which has no classical correspondence.

V. PHASE-DIFFERENCE OPERATOR BETWEEN BOSON AND FERMION

There are two distinct kinds of systems, namely bosons and fermions. The phase difference between two modes of a boson can be described by the direct and indirect phase difference operators as seen in Sec. III and the phase of a single-mode fermion and the phase difference between two modes of fermions can also be described by Hermitian operators as shown in Sec. IV. To complete the quantization of phase, we shall construct a Hermitian operator describing the phase difference between a single-mode fermion and a single-mode boson.

Because there is no phase operator for a single-mode boson, the phase difference between a single-mode fermion and boson cannot be defined by the difference of phases of different systems either. Similar to the two-mode boson case, the phase difference information must be contained in operator $b^\dagger f$ whose kernel space is spanned by the basis

$$\{|n,0\rangle \equiv |n\rangle_b|0\rangle_f | n=0,1,2,\ldots\}. \tag{30}$$

The kernel space $\ker(bf^\dagger)$ of its Hermitian conjugate is spanned by the basis

$$\{|n,1\rangle | n=0,1,2,\ldots\} \cup \{|0,\sigma\rangle | \sigma=0,1\}, \tag{31}$$

which is seemingly a little larger than the kernel space $\ker(b^\dagger f\ )$. But when the property of countable infinity is taken into account, these two kernel spaces have exactly the same dimension and the index of operator $b^\dagger f$ vanishes,

$$\mathrm{index}(b^\dagger f\ ) = 0. \tag{32}$$

The isomorphism Y_{bf} from kernel space $\ker(b^\dagger f\ )$ to kernel space $\ker(bf^\dagger)$ is most generally found to be

$$Y_{bf} = U_0 P_0 (1-N_f) + U_{bf} E f^\dagger, \tag{33}$$

with U_{bf} and U_0 being two unitary operators commuting with the fermionic operators. By requiring the phase-difference factor $\mathscr{E}_{bf}$ to commute with the total number operator $\Delta = N + N_f$, an exponential phase-difference operator and the Hermitian phase-difference operator defined by exponential function for bosonic and fermionic systems are, respectively,

$$\mathscr{E}_{bf} = E^\dagger f + e^{2i\phi(\Delta)}(P_0 f + E) f^\dagger = e^{i\mathscr{P}_{bf}}, \tag{34a}$$

$$\mathscr{P}_{bf} = \phi(\Delta) + \frac{\pi}{2}(1 - e^{-i\phi(\Delta)}\mathscr{E}_{bf}), \tag{34b}$$

where the phase-difference function $\phi(x)$ is an arbitrary real function. The eigenvalues of the phase-difference operator $\mathscr{P}_{bf}$ is found to be $\phi(n)+\sigma\pi$ with $n=0,1,2,\ldots$, and $\sigma=0,1$. In the following discussions the phase-difference function is supposed to be zero for simplicity.

As a physical justification of considering the phase difference between a boson and fermion, let us consider the Jaynes-Cummings model (JCM),[14] through which many quantum features can be revealed. It is also the simplest model describing the interaction between light and matter and attracts much attention. The most generally generalized JCM was investigated from an algebraic point of view and an su(2) structure was revealed.[15] Under the rotating wave approximation, the interaction part of the Hamiltonian can be rewritten in terms of the phase-difference operator between the light and matter as

$$H_I = g\frac{2}{\pi}\sqrt{\Delta}\mathscr{P}_{bf}. \tag{35}$$

Therefore the phase-difference operator of a boson and fermion gives a rather intuitive physical meaning to this model: It just compares the phase difference between a fermion, the matter, and a boson, the light. Therefore a two-level atom in a cavity field can be regarded as an interferometer between a fermion and a boson.

At this point one is able to define a Hermitian phase operator for single-mode boson as follows despite the no-go index theorem. At first we compare this single-mode boson with a single-mode fermion and we have a Hermitian phase-difference operator $\mathscr{P}_{bf}$. Since there exists a Hermitian phase operator $\mathscr{P}_f$ for a single-mode fermion one can subtract the contribution of the phase of the fermion to obtain a Hermitian phase operator for the single-mode boson. But again these two quantized phases are not commuting with each other and therefore they must be combined by using the quantum addition rule, Eq. (16). The resulting phase operator,

$$\mathscr{P}_N \equiv \mathscr{P}_{bf} \dot{-} \mathscr{P}_f, \tag{36}$$

turns out to be exactly the Hermitian phase operator introduced by Newton[7] by adding an additional fermionic degree of freedom to the original single-mode boson, which is unitarily equivalent to the infinite number states approach[16] and the relative number state approach.[8,17] We see that this kind of approach to the phase operator can be interpreted as a phase operator of single-mode boson referring to a single-mode fermion.

ACKNOWLEDGMENTS

One of the authors (S. Yu) would like to thank H. Rauch for providing a favorable researching environment at the Atominstitut in Vienna and his guidance through the years. This research was supported by Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (Project No. 10491).

[1] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **114**, 243 (1927).
[2] M. F. Atiyah and I. M. Singer, Ann. Math. **87**, 484 (1968).
[3] K. Fujikawa, Phys. Rev. A **52**, 3299 (1995).
[4] Z. Tang, Phys. Rev. A **54**, 154 (1996).
[5] L. Susskind and J. Glogower, Physics (Long Island City, NY) **1**, 49 (1964).
[6] S. Yu, S. Liu, J. Summhammer, Y. D. Zhang, G. Badurek, and H. Rauch, Phys. Rev. Lett. (to be sumitted).
[7] R. G. Newton, Ann. Phys. (N.Y.) **124**, 327 (1980).
[8] M. Ban, J. Math. Phys. **32**, 3077 (1991).
[9] D. T. Pegg and S. M. Barnett, Phys. Rev. A **39**, 1665 (1989).
[10] R. Lynch, Phys. Rep. **367**, 256 (1995).
[11] A. Luis and L. L. Sànchez-Soto, Phys. Rev. A **48**, 4702 (1993).
[12] S. Yu, Phys. Rev. Lett. **79**, 780 (1997).
[13] S. Yu and Y. Zhang, Commun. Theor. Phys. **24**, 185 (1995).
[14] E. T. Jaynes and F. W. Cummings, Proc. IEEE **51**, 89 (1963).
[15] S. Yu, H. Rauch, and Y. Zhang, Phys. Rev. A **52**, 2585 (1995).
[16] J. A. Vaccaro, Phys. Rev. A **51**, 3309 (1995).
[17] S. Yu, Phys. Rev. A **56**, 3464 (1997).

Integrable and solvable many-body problems in the plane via complexification

F. Calogero[a),b)]
Dipartimento di Fisica, Universitá di Roma I ''La Sapienza,'' 00185 Roma, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Roma

(Received 21 January 1998; accepted for publication 8 June 1998)

A simple prescription allows us to transform, by appropriate complexification, any *one-dimensional* many-body problem (describing ''motions on a line'') with Newtonian (''acceleration equals forces'') equations of motion featuring forces which depend *analytically* on the positions, and possibly on the velocities, of the particles, and which are moreover *scaling-* or *translation-invariant*, into a *two-dimensional* many-body problem (describing ''motions in the plane'') with *rotation-invariant* equations of motion. If the original (one-dimensional) model is Hamiltonian, the two-dimensional model is also Hamiltonian (in fact, bi-Hamiltonian). In this manner, starting from known integrable or solvable one-dimensional many-body problems, integrable or solvable many-body problems *in the plane*, generally featuring much richer dynamics, are obtained. Several examples are exhibited. Finally another, more direct, prescription is outlined, to transform by complexification almost *any* one-dimensional many-body problem into a *rotation-invariant* many-body problem *in the plane*. © *1998 American Institute of Physics.*
[S0022-2488(98)02310-X]

I. INTRODUCTION

Consider a one-dimensional many-body problem characterized by Newtonian equations of motion,

$$\ddot{z}_j = f_j^{(1)}(\dot{z}_j; z_j) + \sum_{k=1, k\neq j}^{n} f_{jk}^{(2)}(\dot{z}_j, \dot{z}_k; z_j, z_k). \tag{1.1}$$

Here $z_j \equiv z_j(t)$ is the coordinate of the j-th particle, and, for simplicity, we only consider models with autonomous one- and two-body forces, although what follows holds as well in, or is easily extendable to, more general cases. As usual, superimposed dots denote time-derivatives.

If, as we hereafter assume, the functions $f^{(1)}$ and $f^{(2)}$ depend *analytically* on their arguments, the equations of motion (1.1) can be assumed to hold for complex values of the coordinates z_j; thereby one-dimensional motions occurring on the real line are extended to two-dimensional motions extending throughout the complex plane. Trivial as this extension is, it generally entails a substantial enrichment of the qualitative features of the motion. It is, moreover, clear that, if the original model is integrable or solvable (see below), the motions in the complex domain are also integrable or solvable.

By identifying the complex plane with the real physical plane (see below), one can identify in this manner integrable or solvable two-dimensional models, which, in our opinion, do qualify as genuine *many-body problems in the plane* iff the corresponding equations of motion are *rotation-invariant*.

Recently several such integrable or solvable many-body problems have been exhibited;[1–3] several others are displayed below. Indeed, as we point out below, any model characterized by Newtonian equations of motion, see (1.1), which are either *scaling-* or *translation-invariant*, yields

[a)]On leave while serving as Chairman of the Pugwash Council, Pugwash Conferences on Science and World Affairs.
[b)]Electronic mail: calogero@uniroma1.it

 © 1998 American Institute of Physics

by appropriate transformations a *many-body problem in the plane* characterized by *rotation-invariant* equations of motion.

In the following, Sec. II, we collect several useful formulas, and thereby also establish our notation. The result outlined above is given in Sec. III, and examples of integrable or solvable many-body problems in the plane are displayed in Sec. IV, which constitutes the core of this paper. Future developments are outlined in Sec. V.

II. NOTATION AND PRELIMINARIES

Two-vectors in the plane are denoted by a superimposed arrow. We use for them a formally three-dimensional notation, by writing $\vec{r}\equiv(x,y,0)$ and $\hat{z}\wedge\vec{r}\equiv(-y,x,0)$. The symbol $\wedge$ is defined by this formula, which corresponds, see below, to the usual (''three-dimensional'') vector product. Indeed here and always below $\hat{z}$ denotes the unit vector orthogonal to the plane: $\hat{z}\equiv(0,0,1)$. Of course these formulas entail $\hat{z}\wedge(\hat{z}\wedge\vec{r})=-\vec{r}$.

Circular coordinates in the plane are defined as usual by setting $x=r\cos\theta, y=r\sin\theta$, $r^2=x^2+y^2, \tan\theta=x/y$, so that $\vec{r}=r(\cos\theta,\sin\theta,0)$, $\hat{z}\wedge\vec{r}=r(\cos(\theta+\pi/2),\sin(\theta+\pi/2),0)=r(-\sin\theta,\cos\theta,0)$.

The scalar and pseudoscalar products of 2 two-vectors are defined in the standard manner:

$$\vec{r}_j\cdot\vec{r}_k\equiv c_{jk}=x_jx_k+y_jy_k=r_jr_k\cos(\theta_j-\theta_k), \tag{2.1}$$

$$\hat{z}\cdot\vec{r}_j\wedge\vec{r}_k\equiv s_{jk}=x_jy_k-y_jx_k=r_jr_k\sin(\theta_j-\theta_k). \tag{2.2}$$

These definitions entail the following identities (and many others!):

$$c_{12}=c_{21}, c_{11}=r_1^2, \quad s_{12}=-s_{21}, s_{11}=0, \quad \theta_1-\theta_2=\arctan(s_{12}/c_{12}), \tag{2.3}$$

$$\vec{r}_1=(c_{12}\vec{r}_2-s_{12}\hat{z}\wedge\vec{r}_2)/r_2^2, \quad r_1^2r_2^2\equiv c_{11}c_{22}=c_{12}^2+s_{12}^2, \tag{2.4}$$

$$(c_{11}c_{22}-c_{12}^2)\vec{r}_3=(c_{13}c_{22}-c_{23}c_{12})\vec{r}_1+(c_{11}c_{23}-c_{12}c_{13})\vec{r}_2, \quad s_{12}\vec{r}_3=c_{23}\hat{z}\wedge\vec{r}_1-c_{13}\hat{z}\wedge\vec{r}_2, \tag{2.5}$$

$$c_{11}c_{22}c_{33}-c_{11}c_{23}^2-c_{22}c_{31}^2-c_{33}c_{12}^2+2c_{12}c_{23}c_{31}=0, \tag{2.6}$$

$$s_{12}s_{34}=c_{13}c_{24}-c_{14}c_{23}=r_1r_2r_3r_4\sin(\theta_1-\theta_2)\sin(\theta_3-\theta_4), \tag{2.7}$$

$$c_{12}s_{34}-c_{34}s_{12}=c_{13}s_{24}-c_{24}s_{13}=r_1r_2r_3r_4\sin(\theta_1-\theta_2-\theta_3+\theta_4). \tag{2.8}$$

In the following we occasionally use the convenient short-hand notation $\vec{r}_{jk}\equiv\vec{r}_j-\vec{r}_k$, entailing $r_{jk}^2=r_j^2+r_k^2-2\vec{r}_j\cdot\vec{r}_k=c_{jj}+c_{kk}-2c_{jk}$.

The one-to-one correspondence (hereafter denoted by the convenient symbol $\doteq$) among the *physical plane* spanned by the *real* two-vector $\vec{r}$, and the *complex plane* spanned by the *complex* number $z=x+iy$, is given by the following self-evident formulas: $z\doteq\vec{r}, iz\doteq\hat{z}\wedge\vec{r}, |z|^2=r^2$, $\theta=\arg(z)$. They entail the following relations:

$$z_1/z_2=r_2^{-2}(c_{12}-is_{12}), \tag{2.9}$$

$$z_1z_2/z_3\doteq r_3^{-2}(\vec{r}_1c_{23}+\vec{r}_2c_{13}-\vec{r}_3c_{12}), \tag{2.10}$$

$$z_1z_2/(z_3z_4)=r_3^{-2}r_4^{-2}[c_{13}c_{24}+c_{14}c_{23}-c_{12}c_{34}-i(c_{14}s_{23}+c_{23}s_{14})]. \tag{2.11}$$

The left-hand side of this formula is evidently invariant under the interchange of the indices 1 and 2, or 3 and 4; this holds as well for the right-hand side [this is evident for the real part; it is less evident, but of course also true, for the imaginary part, see (2.8)].

$$z_1z_2z_3/z_4z_5=(1/3)r_4^{-2}r_5^{-2}\cdot\{\vec{r}_1[-c_{23}c_{45}+2(c_{24}c_{35}+c_{25}c_{34})]$$
$$+\vec{r}_2[-c_{13}c_{45}+2(c_{14}c_{35}+c_{15}c_{34})]+\vec{r}_3[-c_{12}c_{45}+2(c_{14}c_{25}+c_{15}c_{24})]$$
$$-\vec{r}_4[c_{12}c_{35}+c_{13}c_{25}+c_{23}c_{15}]-\vec{r}_5[c_{12}c_{34}+c_{13}c_{24}+c_{23}c_{14}]\}. \tag{2.12}$$

Let $f(z)$ be an *analytic* function of the complex variable $z=x+isy$ with $s=1$ or $s=-1$, so that $\partial f/\partial z=\partial f/\partial x=-is\partial f/\partial y$. Then set

$$f(z)\equiv f(x+isy)=F(\vec{r})+i\hat{F}(\vec{r}), \tag{2.13a}$$

with $F(\vec{r})$ and $\hat{F}(\vec{r})$ *real* functions of the real two-vector $\vec{r}\equiv(x,sy,0)$ (note that $\vec{r}\doteq z$ if $s=+1$,$\vec{r}\doteq z^\star$ if $s=-1$). Then

$$(\partial/\partial\vec{r})F(\vec{r})=s(\hat{z}\wedge\partial/\partial\vec{r})\hat{F}(\vec{r}),(\partial/\partial\vec{r})\hat{F}(\vec{r})=-s(\hat{z}\wedge\partial/\partial\vec{r})F(\vec{r}). \tag{2.13b}$$

Consider now a system characterized by the Hamiltonian equations of motion

$$\dot{z}_j=\partial h(\mathbf{z},\mathbf{Z})/\partial Z_j,\dot{Z}_j=-\partial h(\mathbf{z},\mathbf{Z})/\partial z_j, \tag{2.14}$$

with $h(\mathbf{z},\mathbf{Z})$ an *analytic* function of the n complex ''canonical coordinates'' z_j and of the corresponding ''canonical momenta'' Z_j. (Throughout this paper we use for convenience lower and upper cases for the canonical coordinates and the corresponding momenta.) Now set

$$z_j=x_j+iy_j,Z_j=X_j-iY_j, \tag{2.15a}$$

namely,

$$z_j\doteq\vec{r}_j,Z_j^\star\doteq\vec{R}_j, \tag{2.15b}$$

and

$$h(\mathbf{z},\mathbf{Z})=H(\vec{\mathbf{r}},\vec{\mathbf{R}})+i\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}). \tag{2.16}$$

Here of course $H(\mathbf{r},\mathbf{R})$ and $\tilde{H}(\mathbf{r},\mathbf{R})$ are two real functions of the (real) two-vectors $\vec{r}_j\equiv(x_j,y_j,0)$ and $\vec{R}_j\equiv(X_j,Y_j,0)$. Then (2.14) yield

$$\dot{\vec{r}}_j=+(\partial/\partial\vec{R}_j)H(\vec{\mathbf{r}},\vec{\mathbf{R}}),\quad \dot{\vec{R}}_j=-(\partial/\partial\vec{r}_j)H(\vec{\mathbf{r}},\vec{\mathbf{R}}), \tag{2.17}$$

$$\dot{\vec{r}}_j=+[\hat{z}\wedge(\partial/\partial\vec{R}_j)]\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}),\quad \dot{\vec{R}}_j=+[\hat{z}\wedge(\partial/\partial\vec{r}_j)]\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}). \tag{2.18}$$

The equations (2.17) are *real* Hamiltonian equations for the (real) two-vector canonical coordinates $\vec{r}_j$ and the corresponding (real) two-vector canonical momenta $\vec{R}_j$. The equations (2.18), instead, are not Hamiltonian (for these choices of canonical coordinates and momenta): note the positive sign in the right-hand-sides of both these equations. They are instead Hamiltonian for the canonical coordinates $\vec{\hat{r}}_j$ and the (new) canonical momenta $\vec{\hat{R}}_j$ defined as follows:

$$\vec{\hat{r}}_j=\vec{r}_j,\quad \vec{\hat{R}}_j=\hat{z}\wedge\vec{R}_j;\quad \hat{H}(\vec{\mathbf{r}},\vec{\mathbf{R}})=\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}), \tag{2.19}$$

which indeed transform (2.18) into

$$\dot{\vec{\hat{r}}}_j=+(\partial/\partial\vec{\hat{R}}_j)\hat{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}),\quad \dot{\vec{\hat{R}}}_j=-(\partial/\partial\vec{\hat{r}}_j)\hat{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}). \tag{2.20}$$

It is thereby seen that the complex Hamiltonian system (2.14), with analytic Hamiltonian function $h(\mathbf{z},\mathbf{Z})$, yields two distinct *real* Hamiltonian structures, see (2.17) and (2.20); and let us emphasize that the three Hamiltonian evolutions (2.14), (2.17) and (2.20) are completely equivalent.

Likewise, consider a system characterized by the (complex) Lagrangian evolution,

$$(d/dt)[\partial l(\mathbf{z},\dot{\mathbf{z}})/\partial \dot{z}_j] = \partial l(\mathbf{z},\dot{\mathbf{z}})/\partial z_j, \tag{2.21}$$

with $l(\mathbf{z},\dot{\mathbf{z}})$ an *analytic* function of all its arguments z_j and $\dot{z}_j$, $j=1,\ldots,n$, and set

$$l(\mathbf{z},\dot{\mathbf{z}}) = \mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}) + i\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}), \tag{2.22}$$

with $\mathcal{L}(\vec{\mathbf{r}},\dot{\mathbf{r}})$ and $\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}})$ *real* functions of the *real* two-vectors $\vec{r}_j \doteq z_j, \dot{\vec{r}}_j \doteq \dot{z}_j$. It is then easily seen that two equivalent (real) Lagrangian evolutions correspond to (2.21):

$$(d/dt)[(\partial/\partial \dot{\vec{r}}_j)\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}})] = (\partial/\partial \vec{r}_j)\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}), \tag{2.23a}$$

$$(d/dt)[(\partial/\partial \dot{\vec{r}}_j)\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}})] = (\partial/\partial \vec{r}_j)\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}). \tag{2.23b}$$

The ''complex plane'' versions of the connection between the Lagrangian and Hamiltonian avatars of the equations of motion (2.14) and (2.21) read as follows:

$$h(\mathbf{z},\mathbf{Z}) = \sum_{j=1}^{n} Z_j \dot{z}_j - l(\mathbf{z}_j,\dot{\mathbf{z}}_j), \tag{2.24a}$$

$$Z_j = \partial l(\mathbf{z},\dot{\mathbf{z}})/\partial \dot{z}_j. \tag{2.24b}$$

The corresponding ''real plane'' relations read as

$$H(\vec{\mathbf{r}},\vec{\mathbf{R}}) = \sum_{j=1}^{n} (\vec{R}_j \cdot \dot{\vec{r}}_j) - \mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}), \tag{2.25a}$$

$$\vec{R}_j = (\partial/\partial \dot{\vec{r}}_J)\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}); \tag{2.25b}$$

$$\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}}) = \sum_{j=1}^{n} (\hat{z} \cdot \vec{R}_j \wedge \dot{\vec{r}}_j) - \tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}), \tag{2.26a}$$

$$\vec{R}_j = [\hat{z} \wedge (\partial/\partial \dot{\vec{r}}_j)]\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}). \tag{2.26b}$$

Finally, for future reference below, we report here a Proposition and its Corollary.[3] Consider the dynamical system characterized by equations of motion of Newtonian type having the following structure:

$$\zeta_j'' = \sum_{k,m=1}^{n} \zeta_k' \zeta_m' \varphi_{jkm}(\boldsymbol{\zeta}), \tag{2.27}$$

where $\zeta_j \equiv \zeta_j(\tau)$ and the primes denote differentation with respect to τ.

Then set

$$z_j(t) = \zeta_j(\tau), \quad \tau = [\exp(at) - 1]/a, \tag{2.28}$$

assuming of course these equations make sense, and in particular $\zeta_j(\tau)$ is well defined for complex values of τ if a is complex, see below.

Proposition: The $z_j(t)$'s evolve according to the equations of motion,

$$\ddot{z}_j = a\dot{z}_j + \sum_{k,m=1}^{n} \dot{z}_k \dot{z}_m \varphi_{jkm}(\mathbf{z}). \tag{2.29}$$

Note that, for the validity of this Proposition, no restriction is *a priori* required on the arbitrary functions $\varphi_{jkm}(\mathbf{z})$ [except the nontrivial requirement that (2.28) be well defined, see above and below].

Corollary: If the system (2.27) is *integrable* [a property which of course only holds for special choices of the functions $\varphi_{jkm}(\mathbf{z})$] so that the solutions $\zeta_j(\tau)$ are globally defined for all (complex) values of τ, and $a = \pm i\Omega$, with $\Omega > 0$ a real nonvanishing constant, *all* solutions of (2.29) are *completely periodic* (note that these motions extend then necessarily into the *complex* plane), the generic solution having period $T = 2\pi/\Omega$.

III. ROTATION-INVARIANT TWO-DIMENSIONALIZATIONS VIA COMPLEXIFICATION

It is obvious that, if the (complex) Newtonian equations of motion (1.1) are *invariant under the rescaling transformation* $z_j \to c z_j$, with c an arbitrary (complex) constant, the real equations in the plane, obtained from (1.1) via the correspondence introduced above among real two-vectors $\vec{r}$ and complex numbers z, are *invariant under rotations*, since for $c = \exp(i\theta)$, with θ a real arbitrary ''angle,'' the *rescaling transformation*,

$$z_j \to z_j \exp(i\theta), \tag{3.1}$$

corresponds to a *rotation in the plane*, namely, in real Cartesian coordinates,

$$x_j \to x_j \cos\theta - y_j \sin\theta, \quad y_j \to x_j \sin\theta + y_j \cos\theta, \tag{3.2}$$

or equivalently, in real circular coordinates,

$$r_j \to r_j, \quad \theta_j \to \theta_j + \theta. \tag{3.3}$$

Various integrable and solvable many-body problems in the plane were obtained in this manner.[1–3]

Suppose now that the (complex) Newtonian equations of motion (1.1) are *translation-invariant*, namely that, after having being rewritten (for convenience, see below) with z_j replaced by u_j, they are invariant under the translation $u_j \to u_j + u_0$, with u_0 an arbitrary complex constant, so that they read as

$$\ddot{u}_j = \sum_{k=1}^{n} f_{jk}(\dot{u}_j, \dot{u}_k; u_j - u_k). \tag{3.4}$$

It is then clear that, by setting

$$u_j = \log(z_j), \tag{3.5a}$$

which of course entails

$$\dot{u}_j = \dot{z}_j/z_j, \quad \ddot{u}_j = [\ddot{z}_j - \dot{z}_j^2/z_j]/z_j, \tag{3.5b}$$

one obtains new equations of motion for the new coordinates z_j which are *scaling-invariant*:

$$\ddot{z}_j = \dot{z}_j^2/z_j + z_j \sum_{k=1}^{n} f[\dot{z}_j/z_j, \dot{z}_k/z_k; \log(z_j/z_k)]. \tag{3.6}$$

Hence, by complexification, one obtains from these equations real many-body problems in the plane which do possess the essential property of being *rotation-invariant*. This opens the possibility to generate new integrable or solvable many-body problems in the plane; several examples

are displayed below, and of course several others can be produced by the diligent reader. Note, however, that these new models need not be *translation-invariant*, and that they generally contain velocity-dependent forces.

If there exists a Lagrangian $l(\mathbf{u},\dot{\mathbf{u}})$ which yields the equations of motion (3.4), via (3.5) one immediately gets the corresponding Lagrangian that yields (3.6). Likewise, if $h(\mathbf{u},\mathbf{U})$ is a Hamiltonian that yields (3.4), the corresponding Hamiltonian that yields (3.6) is obtained via (3.5a) and the position

$$U_j = Z_j z_j\,, \tag{3.7}$$

which introduces the canonical momenta Z_j conjugated to the canonical coordinates z_j. After complexification, one then obtains, via (2.16), (2.19) and (2.22), new *real* Hamiltonians and Lagrangians that yield, via (2.17), (2.20) and (2.23), new *real rotation-invariant* equations of motion which are interpretable as the equations of motions of *many-body problems in the plane*.

Let us end this Section by emphasizing that our requirement of *rotation-invariance* applies here to the (second-order) equations of motion of Newtonian type, not to the (first-order) Hamiltonian equations. At the end of this paper (see Sec. V), we outline another (more general, but rather trivially so) version of the complexification prescription — not employed in the examples presented in the following, Sec. IV — which allows us to transform *any* one-dimensional problem for which complexification makes sense, into a corresponding *rotation-invariant* real two-dimensional problem. Again, if the original one-dimensional problem is Hamiltonian, the two-dimensional problem obtained via this trick is also Hamiltonian, and, in fact, even its Hamiltonian equations are then *rotation-invariant*. And of course solvability or integrability is again preserved by such an extension from one- to two-dimensional space.

IV. INTEGRABLE AND SOLVABLE MANY-BODY PROBLEMS IN THE PLANE

In this Section we present an overview of integrable and solvable many-body problems in the plane.

By overview we mean a terse presentation, consisting mainly of an identification of each system, including the display of its Newtonian equations of motion, as well as their Hamiltonian and Lagrangian versions. We do not analyze the corresponding motions in the plane, except for some occasional remarks, such as the identification of special solutions, or special models which feature only completely periodic motions. For systems with ''nearest-neighbor'' interactions — that is, systems in which the j-th particle interacts only with the $(j-1)$-th and $(j+1)$-th ones; of course these particles need not be close to each other as they move in the plane — we always ignore the ''end-point'' issue, namely the additional prescription needed to specify the interaction experienced by the first and last particles (if their number is finite). Hence, each of the models described below requires further study, to analyze in detail its motions: a worthwhile task, in view of the much greater richness of available motions in the plane, as compared to those on the line (see, for instance, Ref. 1).

In each case we reveal the reason that underpins the ''integrability'' or ''solvability'' of the model under display. Let us emphasize in this connection that the definition and distinction among *integrable* and *solvable* models is heuristic: useful but not quite precise. We call here ''integrable'' those models which possess at least as many (globally defined) integrals of motion as degrees of freedom, generally as a consequence of the possibility to recast their equations of motion (nontrivially!)[4] in ''Lax-pair form.'' We identify instead as ''solvable'' those models for which a technique of solution is known — a technique which may be implemented quite explicitly or only partly so, and which is generally based on an appropriate change of variables. According to this terminology, ''solvability'' is generally more stringent than (namely, it entails) ''integrability;'' although in fact, for finite-dimensional systems, these two notions tend to overlap altogether.

All the integrable or solvable many-body problems in the plane presented below feature velocity-dependent one- and two-body (''pair'' or ''nearest-neighbor'') forces. As we already emphasized, all of them are invariant under rotations in the plane. Most of them are, instead, not

invariant under translations. All the new systems are Hamiltonian; we generally also display, as mentioned above, their Hamiltonian and Lagrangian structures (in more or less explicit detail, as the case may be).

Each of the Subsections below is devoted to a different system. All of these systems but one are new, at least in their two-dimensional avatars, and most of them have been obtained using the results reported in the preceding Section. We also report one case which is not quite new,[1] to exhibit a similarity solution not previously given (''model three,'' see Sec. IV C). The criteria for selecting the models displayed, among all those which can be obtained from known integrable and solvable systems using the results of the preceding Sec. III, as well as for deciding the amount of explicit detail in each case and the format of the results being exhibited, are based on ''aesthetic'' considerations and the related effort to maximize the ''user-friendly'' character of our presentation.

Finally, let us mention the somewhat special status of the last of the many-body systems reported (''model eleven''), as presented in Sec. IV L below: in this case the equations of motion can be exhibited only in Hamiltonian form (not in Newtonian form), and the *general solution* is given.

A. Model one

The Newtonian equations of motion for this system read as follows:

$$\ddot{\vec{r}}_j = r_j^{-2}[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j) - \vec{r}_j|\dot{\vec{r}}_j|^2] - (\alpha + \hat{\alpha}\hat{z}\wedge)\sum_{k=1,k\neq j}^{n} r_{jk}^{-6}$$
$$\cdot\{\vec{r}_j r_k^2[r_{jk}^4 - r_{jk}^2(r_j^2 - 3r_k^2) + 2(r_j^2 - r_k^2)^2] + \vec{r}_k r_j^2[r_{jk}^2(r_j^2 + r_k^2) - 2(r_j^2 - r_k^2)^2]\}. \tag{4.1}$$

Here α and $\hat{\alpha}$ are two (real) arbitrary ''coupling constants.''

This model features one-body velocity-dependent forces (quadratic in the velocities), and two-body velocity-independent forces. It is obviously rotation-invariant; it is not translation invariant. It is clearly invariant under the scale transformation $\vec{r}_j \to c\vec{r}_j$ with c an arbitrary constant.

The equations of motion (4.1) can be cast into the Lagrangian form (2.23). There are indeed two (families of) Lagrangians that yield (4.1):

$$\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}) = (A^2+\hat{A}^2)^{-1}\sum_{j=1}^{n} r_j^{-4}(A\lambda_j + \hat{A}\hat{\lambda}_j) + \sum_{j,k=1;j\neq k}^{n} r_{jk}^{-4}[(GA - \hat{G}\hat{A})\mu_{jk} - (G\hat{A} + \hat{G}A)\hat{\mu}_{jk}], \tag{4.2}$$

$$\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}}) = (A^2+\hat{A}^2)^{-1}\sum_{j=1}^{n} r_j^{-4}(A\hat{\lambda}_j - \hat{A}\lambda_j) + \sum_{j,k=1;j\neq k}^{n} r_{jk}^{-4}[(GA - \hat{G}\hat{A})\hat{\mu}_{jk} + (G\hat{A} + \hat{G}A)\mu_{jk}], \tag{4.3}$$

$$\lambda_j \equiv (1/8)[2(\dot{\vec{r}}_j\cdot\vec{r}_j)^2 - |\dot{\vec{r}}_j|^2 r_j^2], \quad \hat{\lambda}_j \equiv (1/4)[2(\dot{\vec{r}}_j\cdot\vec{r}_j)(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j), \tag{4.4}$$

$$\mu_{jk} \equiv 2r_j^2 r_k^2 - (r_j^2 + r_k^2)(\vec{r}_j\cdot\vec{r}_k), \quad \hat{\mu}_{jk} \equiv -(r_j^2 - r_k^2)(\hat{z}\cdot\vec{r}_j\wedge\vec{r}_k). \tag{4.5}$$

Each of these two Lagrangians depends on the 4 *real* constants $A,\hat{A},G,\hat{G}$, which are related to the 2 *real* constants α and $\hat{\alpha}$ [see (4.1)] as follows:

$$\alpha = 8[(A^2 - \hat{A}^2)G - 2A\hat{A}\hat{G}], \quad \hat{\alpha} = 8[(A^2 - \hat{A}^2)\hat{G} + 2A\hat{A}G]. \tag{4.6}$$

Likewise, there are (families of) Hamiltonians that yield, via (2.17) and (2.18) with (2.19) and (2.20), the Newtonian equations of motion (4.1):

$$H(\vec{\mathbf{r}},\vec{\mathbf{R}}) = \sum_{j=1}^{n}(A\Lambda_j + \hat{A}\hat{\Lambda}_j) - \sum_{j,k=1;j\neq k}^{n} r_{jk}^{-4}[(GA - \hat{G}\hat{A})\mu_{jk} - (G\hat{A} + \vec{G}A)\hat{\mu}_{jk}], \tag{4.7}$$

$$\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}})=\sum_{j=1}^{n}(-A\hat{\Lambda}_j+\hat{A}\Lambda_j)-\sum_{j,k=1;j\neq k}^{n} r_{jk}^{-4}[(GA-\hat{G}\hat{A})\hat{\mu}_{jk}+(G\hat{A}+\hat{G}A)\mu_{jk}], \tag{4.8}$$

$$\Lambda_j\equiv 2[2(\vec{r}_j\cdot\vec{R}_j)^2-r_j^2R_j^2], \quad \hat{\Lambda}_j\equiv 4(\vec{r}_j\cdot\vec{R}_j)(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_j), \tag{4.9}$$

of course again with (4.5) and (4.6).

The Hamiltonian (4.7) is related to the Lagrangian (4.2) via (2.25); likewise $\tilde{H}$ (4.8) is related to $\tilde{\mathcal{L}}$ (4.3) via (2.26).

Both (4.2) and (4.3) yield, via (2.25b) and (2.26b),

$$\vec{R}_j=\frac{1}{4}(A^2+\hat{A}^2)^{-1}r_j^{-4}\cdot\{A[2\vec{r}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\dot{\vec{r}}_jr_j^2]+\hat{A}[\vec{r}_j(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j)-\hat{z}\wedge\vec{r}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)]\}, \tag{4.10a}$$

and both (4.7), respectively (4.8) yield, via (2.20a), respectively (2.21a),

$$\dot{\vec{r}}_j=4\{A[2\vec{r}_j\cdot(\vec{r}_j\cdot\vec{R}_j)-\vec{R}_jr_j^2]+\hat{A}[-2\hat{z}\wedge\vec{r}_j(\vec{r}_j\cdot\vec{R}_j)+\vec{z}\wedge\vec{R}_jr_j^2]\}, \tag{4.10b}$$

which is in fact equivalent to (4.10a).

The equations of motion (4.1) possess the two constants of motion,

$$C=\sum_{j=1}^{n}(\dot{\vec{r}}_j\cdot\vec{r}_j)/r_j^2, \quad \hat{C}=\sum_{j=1}^{n}(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j)/r_j^2. \tag{4.11}$$

A simple proof of this fact, as well as of the existence of $2n-2$ additional (real) constants of motion entailing the *integrability* of the system (4.1), are implied by the results given below.

The system (4.1) possesses the ''similarity solution'' (written using polar coordinates)

$$r_j(t)=r_0\,\exp(ut), \quad \theta_j(t)=\theta_0+2\pi j/n+vt, \tag{4.12}$$

which features the 4 arbitrary (real) constants r_0,θ_0,u and v. The corresponding picture sees n particles sitting equispaced on a circular ring centered at the origin, which rotates with the constant (arbitrary) angular velocity v and whose radius evolves exponentially in time: for $u>0$ it diverges as $t\to\infty$, for $u<0$ it tends to zero, while for $u=0$ it remains constant and the motion is periodic with period $T=2\pi r/|v|$ (if $v\neq 0$; if v also vanishes, the system is at equilibrium). There are in fact $n!$ such solutions, obtained by replacing the index j in the right-hand side of the second equation (4.12) with k, and by then assigning to these indices as their values an arbitrary permutation of the n numbers $1,2,\ldots,n$.

Note moreover that, if $r_j(t),\theta_j(t)$ is, in circular coordinates, a solution of (4.1), then $r_j'(t)=c\,\exp(ut)r_j(t)$, $\theta_j'(t)=\theta_j(t)+\theta_0+vt$ is also a solution (with c,u,θ_0 and v four arbitrary constants); while $\tilde{r}_j(t)=r_j(t)\exp[ut+\beta(t)]$, $\tilde{\theta}_j(t)=\theta_j(t)+\theta_0+vt+\hat{\beta}(t)$ provides a solution of the new many-body system characterized by equations of motion which differ from (4.1) only due to the additional presence, in the left-hand-side, of the term $[\ddot{\beta}(t)+\dddot{\hat{\beta}}(t)\hat{z}\wedge]\vec{r}_j$.

The complete integrability of the equations of motion (4.1), as well as the other results mentioned above, are due to the equivalence of (4.1) to the complex equations of motion,

$$\ddot{z}_j=\dot{z}_j^2/z_j+8a^4g^2\sum_{k=1,k\neq j}^{n} z_j^2z_k(z_j+z_k)/(z_j-z_k)^3, \tag{4.13}$$

with

$$a^2=A+i\hat{A}, g^2=G+i\hat{G}, \tag{4.14}$$

entailing, via (4.6),

$$8a^4g^2=\alpha+i\hat{\alpha}. \tag{4.15}$$

Dividing (4.12) by z_j and summing over j from 1 to n yields

$$(d/dt)\sum_{j=1}^{n}(\dot{z}_j/z_j)=0, \tag{4.16}$$

because the double sum in the right-hand side vanishes due to the antisymmetry of the summand. This yields (4.11).

The complex Lagrangian that yields, via (2.21), the equations of motion (4.13) reads as

$$l(\mathbf{z},\dot{\mathbf{z}})=(8a^2)^{-1}\sum_{j=1}^{n}(\dot{z}_j/z_j)^2-a^2g^2\sum_{j,k=1;j\neq k}^{n}z_jz_k/(z_j-z_k)^2, \tag{4.17}$$

and the corresponding [see (2.24)] complex Hamiltonian reads as

$$h(\mathbf{z},\mathbf{Z})=a^2\left\{2\sum_{j=1}^{n}z_j^2Z_j^2+g^2\sum_{j,k=1;j\neq k}^{n}z_jz_k/(z_j-z_k)^2\right\}. \tag{4.18}$$

They yield, via (2.22) as well as (2.16), the real Lagrangians (4.2)–(4.5) as well as two real Hamiltonians [see (4.7)–(4.9) and (2.19), (2.20)].

The similarity solution (4.12) reads, in complex coordinates, as

$$z_j(t)=\exp[\varphi(t)+2\pi ij/n],\quad \varphi(t)=\log(r_0)+i\theta_0+(u+iv)t, \tag{4.19}$$

and it satisfies (4.13) since the quantities

$$\epsilon_j\equiv\sum_{k=1,k\neq j}^{n}\exp[2\pi i(j+k)/n][\exp(2\pi ij/n)+\exp(2\pi ik/n)]/[\exp(2\pi ij/n)-\exp(2\pi ik/n)]^3 \tag{4.20}$$

vanish.

[*Proof*: by dividing the numerator and denominator in the right-hand side of (4.20) by $\exp(6\pi ij/n)$, and then setting $k=j+l \bmod (n)$, it is easily seen that $\epsilon_j=\epsilon$ is independent of j. But by summing (4.20) over j from 1 to n one then gets $n\epsilon=0$, since the right-hand side vanishes due to the antisymmetry of the summand under the exchange of the dummy indices j and k in the double sum.]

Finally, the complete integrability of the complex equations of motion (4.13) is demonstrated by setting

$$z_j=\exp(2aq_j), \tag{4.21}$$

transforming thereby (4.13) into

$$\ddot{q}_j=g^2a^3\sum_{k=1,k\neq j}^{n}\cosh[a(q_j-q_k)]/\{\sinh[a(q_j-q_k)]\}^3, \tag{4.22}$$

which are the equations of motion corresponding to the well-known *integrable* Hamiltonian,[5,6]

$$\mathcal{H}(\mathbf{q},\mathbf{Q})=\frac{1}{2}\sum_{j=1}^{n}Q_j^2+(g^2a^2/4)\sum_{j,k=1;j\neq k}^{n}\{\sinh[a(q_j-q_k)]\}^{-2}. \tag{4.23}$$

Hence the system reported in this Subsection corresponds, up to an appropriate (complex) rescaling, to the application of the technique of Sec. III to the integrable translation-invariant system associated with this Hamiltonian.

B. Model two

In this case, which features one- and two-body forces, we write the equations of motion, and the corresponding Hamiltonian and Lagrangian, only in the complex version. They read as

$$\ddot{z}_j = a\dot{z}_j + \dot{z}_j^2/z_j + 4\sum_{k=1,k\neq j}^{n} \dot{z}_j\dot{z}_k z_j(z_j+z_k)\cdot\{(z_j-z_k)[4z_jz_k+b(z_j-z_k)^2]\}^{-1}, \tag{4.24}$$

$$h(\mathbf{z},\mathbf{Z}) = \sum_{j=1}^{n}\left\{\exp(sz_jZ_j)\prod_{k=1,k\neq j}^{n}\{[b(z_j^2+z_k^2)-2(2+b)z_jz_k]^{1/2}/(z_j-z_k)\}-(a/s)\log z_j\right\}, \tag{4.25}$$

$$l(\mathbf{z},\mathbf{Z}) = -\sum_{j=1}^{n}\left\{[\dot{z}_j/(sz_j)]\left(1+\log[\dot{z}_j/(sz_j)]-\sum_{k=1,k\neq j}^{n}\log\{[b(z_j^2+z_k^2)-2(2+b)z_jz_k]^{1/2}/(z_j-z_k)\}\right)-(a/s)\log z_j\right\}, \tag{4.26}$$

$$Z_j = (sz_j)^{-1}\left(\log[\dot{z}_j/(sz_j)]-\sum_{k=1,k\neq j}^{n}\log\{[b(z_j^2+z_k^2)-2(2+b)z_jz_k]^{1/2}/(z_j-z_k)\}\right). \tag{4.27}$$

In these formulas the (complex) constants a, b and s are arbitrary ($s\neq 0$). Note that s does not appear in the Newtonian equations of motion (4.24), which are invariant under the rescaling $z_j \to cz_j$, hence yield, via the transition from the complex to the real plane (see Secs. II and III), Newtonian equations of motion for the real 2-vectors $\vec{r}_j(t)$ which are rotation-invariant, contain 4 arbitrary real constants and can be written in Hamiltonian and Lagrangian form. The derivation of the relevant formulas is an easy task using the formulas of Sec. II; we forsake here their explicit display.

If the constant a is imaginary, $a=\pm i\Omega, \Omega>0$, the generic solution of (4.24) is completely periodic with period $T=2\pi/\Omega$.

By dividing (4.24) by z_j and summing over j from 1 to n it is easily seen that

$$w(t) = \sum_{j=1}^{n}[\dot{z}_j(t)/z_j(t)] \tag{4.28}$$

satisfies the linear equation

$$\dot{w} = aw, \tag{4.29a}$$

implying

$$w(t) = w(0)\exp(at). \tag{4.29b}$$

Likewise, it is easily seen that the system (4.24) possesses $n!$ ''similarity solutions'' of the form

$$z_j(t) = \varphi(t)\exp(2\pi ik/n), \tag{4.30}$$

with the values of the index j on the left-hand side matched to the values of k on the right-hand side by an arbitrary permutation of the n positive integers $1,2,\ldots,n$. This corresponds to a configuration of the n particles equally spaced on a circle centered at the origin, with radius $r(t)=|\varphi(t)|$ and angular phase $\theta(t)=\arg[\varphi(t)]$. The (complex) function $\varphi(t)$ satisfies the equation

$$n\dot{\varphi} = w, \tag{4.31a}$$

implying, via (4.29b),

$$\varphi(t)=r_0\ \exp(i\theta_0)\exp\{c[\exp(at)-1]\}, \tag{4.31b}$$

with

$$\varphi(0)=r_0\ \exp(i\theta_0), \quad \dot{\varphi}(0)=ac\varphi(0). \tag{4.31c}$$

Hence, with $a=\alpha+i\hat{\alpha}, c=\gamma\ \exp(i\eta)$,

$$r(t)=r_0\ \exp\{\gamma[\exp(\alpha t)\cos(\hat{\alpha}t+\eta)-\cos\ \eta]\}, \tag{4.32a}$$

$$\theta(t)=\theta_0+\gamma[\exp(\alpha t)\sin(\hat{\alpha}t+\eta)-\sin\ \eta], \tag{4.32b}$$

with the 4 real constants r_0, γ, η and θ_0 *a priori* arbitrary. Clearly if $\alpha>0$ the behavior as $t\to+\infty$ is rather unwieldy, while if $\alpha<0$ it is more reasonable. Of course if $\alpha=0$ this solution is periodic with period $T=2\pi/|\hat{\alpha}|$.

The system (4.24) is *integrable* because, via the position $z_j=\exp(2u_j)$, it gets transformed into the ''Ruijsenaars-type'' system,

$$\ddot{u}_j=a\dot{u}_j+2\sum_{k=1,k\neq j}^{n}\dot{u}_j\dot{u}_k\mathrm{cotanh}(u_j-u_k)/[1+b\sinh^2(u_j-u_k)] \tag{4.33}$$

[see, for instance, Eq. (2.5a) with (2.5f) of Ref. 3, or Ref. 7 from which we obtained (4.25)].

C. Model three

The Newtonian equations of motion for this system read as follows:

$$\begin{aligned}\ddot{\vec{r}}_j=&(\alpha+\hat{\alpha}\hat{z}\wedge)\dot{\vec{r}}_j+(\beta+\hat{\beta}\hat{z}\wedge)\vec{r}_j+\sum_{k=1,k\neq j}^{n}r_{jk}^{-2}\cdot(2[\dot{\vec{r}}_j(\dot{\vec{r}}_k\cdot\vec{r}_{jk})+\dot{\vec{r}}_k(\dot{\vec{r}}_j\cdot\vec{r}_{jk})-\vec{r}_j(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_k)]\\&+(\lambda+\hat{\lambda}\hat{z}\wedge)\{(\dot{\vec{r}}_j+\dot{\vec{r}}_k)[r_j^2-(\vec{r}_j\cdot\vec{r}_k)]-\vec{r}_j[\vec{r}_k\cdot(\dot{\vec{r}}_j+\dot{\vec{r}}_k)]+\vec{r}_k[\vec{r}_j\cdot(\dot{\vec{r}}_j+\dot{\vec{r}}_k)]\}\\&+(\mu+\hat{\mu}\hat{z}\wedge)\{\vec{r}_j[r_j^2-2(\vec{r}_j\cdot\vec{r}_k)]+\vec{r}_kr_j^2\}),\end{aligned} \tag{4.34a}$$

or equivalently, in complex notation,

$$\ddot{z}_j=A\dot{z}_j+Bz_j+\sum_{k=1,k\neq j}^{n}(z_j-z_k)^{-1}[2\dot{z}_j\dot{z}_k+L(\dot{z}_j+\dot{z}_k)z_j+Mz_j^2], \tag{4.34b}$$

with $A=\alpha+i\hat{\alpha},\ B=\beta+i\hat{\beta},\ L=\lambda+i\hat{\lambda},\ M=\mu+i\hat{\mu}$.

This model is not new;[1] it contains 8 real (arbitrary) coupling constants, hence it features a rich gamut of behaviors.[1] Here we exhibit (in complex notation) a simple ''similarity solution'' (in fact, $n!$ of them), namely

$$z_j(t)=\varphi(t)\exp(2\pi ik/n), \tag{4.35}$$

with the usual [see above, after (4.30)] match by permutation among the values of the indexes j and k. Here the (complex) function $\varphi(t)$ satisfies the equation

$$\ddot{\varphi}=A\dot{\varphi}+\left[B+\frac{1}{2}(n-1)M\right]\varphi+(1-n)\dot{\varphi}^2/\varphi, \tag{4.36a}$$

entailing, via $\varphi=\psi^{1/n}$,

$$\ddot{\psi}=A\dot{\psi}-n\left[B+\frac{1}{2}(n-1)M\right]\psi. \tag{4.36b}$$

Hence

$$\varphi(t)=[c_+\ \exp(\eta_+ t)+c_-\ \exp(\eta_- t)]^{1/n}, \tag{4.37a}$$

with $\eta_\pm$ the two roots of the equation

$$\eta^2-A\eta-n\left[B+\frac{1}{2}(n-1)M\right]=0. \tag{4.37b}$$

In writing (4.37a) we have, for simplicity, assumed these two (generally complex) roots η_+ and η_- to be different. It is clear from (4.36b) how (4.37a) must be modified when this is not the case.

To obtain (4.36a) we used the following identities:

$$\sum_{k=1}^{n-1}[1-\exp(2\pi ik/n)]^{-1}=(n-1)/2, \tag{4.38a}$$

$$\sum_{k=1}^{n-1}[1+\exp(2\pi ik/n)]/[1-\exp(2\pi ik/n)]=0, \tag{4.38b}$$

$$\sum_{k=1}^{n-1}\exp(2\pi ik/n)/[1-\exp(2\pi ik/n)]=(1-n)/2. \tag{4.38c}$$

D. Model four

The Newtonian equations of motion for this *integrable many-body problem in the plane* read as follows:

$$\ddot{\vec{r}}_j=[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2+[g_j(\vec{\mathbf{r}})+\hat{g}_j(\vec{\mathbf{r}})\hat{z}\wedge]\vec{r}_j\,, \tag{4.39}$$

or equivalently,

$$\ddot{r}_j=\dot{r}_j^2/r_j+r_j g_j(\vec{\mathbf{r}}), \tag{4.40a}$$

$$\ddot{\theta}_j=\hat{g}_j(\vec{\mathbf{r}}), \tag{4.40b}$$

with

$$g_j(\vec{\mathbf{r}})=\sum_{k=1,k\neq j}^{n}[\alpha\ \log(r_j/r_k)-\hat{\alpha}(\theta_j-\theta_k)+[\beta\ \log(r_j/r_k)\{[\log(r_j/r_k)]^2-3(\theta_j-\theta_k)^2\}$$
$$-\hat{\beta}(\theta_j-\theta_k)\{-3[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}]/\{[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}^3], \tag{4.41a}$$

$$\hat{g}_j(\vec{\mathbf{r}})=\sum_{k=1,k\neq j}^{n}[\hat{\alpha}\ \log(r_j/r_k)+\alpha(\theta_j-\theta_k)+[\hat{\beta}\ \log(r_j/r_k)\{[\log(r_j/r_k)]^2-3(\theta_j-\theta_k)^2\}$$
$$+\beta(\theta_j-\theta_k)\{-3[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}]/\{[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}^3]. \tag{4.41b}$$

This Hamiltonian (see below) model features one-body velocity-dependent forces, and two-body velocity-independent forces: see (4.39), and note that to write the two-body contribution, see (4.41), as well as the equivalent version (4.40) of the equations of motion, we employed circular coordinates; see Sec. II.

These equations of motion are clearly rotation-invariant; they are not translation-invariant. They are invariant under the rescaling transformation $\vec{r}_j \to c\vec{r}_j$, with c an arbitrary constant. In fact, it is clear (see also below) that, if $r_j(t)$, $\theta_j(t)$, $j=1,\ldots,n$, is, in circular coordinates, a solution of (4.40),

$$\tilde{r}_j(t) = r_0 \exp(ut) r_j(t), \tag{4.42a}$$

$$\tilde{\theta}_j(t) = \theta_0 + vt + \theta_j(t), \tag{4.42b}$$

is also a solution; here r_0, u, θ_0 and v are 4 arbitrary (real) constants ($r_0 > 0$).

These equations of motion feature the 4 arbitrary real ''coupling constants'' $\alpha, \hat{\alpha}, \beta, \hat{\beta}$. If $\alpha < 0$ and $\hat{\alpha} = 0$, the generic solution is, up to a transformation of type (4.42), completely periodic with period $T = 2\pi(-\alpha)^{-1/2}$ (see below for a justification of this statement).

If the quantity b/a is *real* and *negative*, with a and b defined as follows:

$$a = \alpha + i\hat{\alpha}, \quad b = \beta + i\hat{\beta}, \tag{4.43}$$

this model possesses the following similarity solutions (written in circular coordinates):

$$r_j(t) = r_0 \exp(ut), \tag{4.44a}$$

$$\theta_j = \theta_0 + vt + [-b/(2an)]^{1/4} \xi_k . \tag{4.44b}$$

Here r_0, θ_0 and v are 4 arbitrary real constants ($r_0 > 0$), the real root should of course be taken in the right-hand side of (4.44b), the indices k there are related to the indices j by any permutation of the n positive integers from 1 to n, and the n real numbers ξ_k are the n roots of the Hermite polynomial of order n,

$$H_n(\xi_k) = 0, \tag{4.45a}$$

so that[8]

$$\sum_{j=1}^{n} \xi_j = 0, \tag{4.45b}$$

$$\xi_j = 2 \sum_{k=1, k\neq j}^{n} (\xi_j - \xi_k)^{-3}. \tag{4.45c}$$

These similarity solutions feature the n particles sitting, appropriately spaced, on a circle centered at the origin which rotates uniformly and whose radius evolves exponentially in time. The verification that (4.44) with (4.45) satisfies (4.40) is an easy task (see also below).

The ''complex-plane'' avatar of these equations of motion reads as

$$\ddot{z}_j = \dot{z}_j^2/z_j + z_j \sum_{k=1, k\neq j}^{n} \{a \log(z_j/z_k) + b[\log(z_j/z_k)]^{-3}\}, \tag{4.46}$$

with the complex ''coupling constants'' a and b related to $\alpha, \hat{\alpha}, \beta$ and $\hat{\beta}$ by (4.43). The Hamiltonian $h(\mathbf{z}, \mathbf{Z})$ and the Lagrangian $l(\mathbf{z}, \dot{\mathbf{z}})$ that yield these equations read as

$$h(\mathbf{z}, \mathbf{Z}) = \frac{1}{2} \sum_{j=1}^{n} z_j^2 Z_j^2 + V(\mathbf{z}), \tag{4.47a}$$

$$l(\mathbf{z}, \dot{\mathbf{z}}) = \frac{1}{2} \sum_{j=1}^{n} \dot{z}_j^2/z_j^2 - V(\mathbf{z}), \tag{4.47b}$$

with

$$V(\mathbf{z})=\frac{1}{4}\sum_{k,j=1;k\neq j}^{n}\{-a[\log(z_j/z_k)]^2+b[\log(z_j/z_k)]^{-2}\}. \tag{4.47c}$$

From these formulas, using the results of Sec. II, one can easily obtain the real Hamiltonians and Lagrangians that yield the equations (4.39) or (4.40) with (4.41).

The integrability of these equations of motion, as well as more straightforward proofs of some of the results reported above, follow from well-known properties of the integrable system[6] defined by the Hamiltonian:

$$H(\mathbf{u},\mathbf{U})=\frac{1}{2}\sum_{j=1}^{n}U_j^2+\frac{1}{4}\sum_{j,k=1;j\neq k}^{n}\{-a(u_j-u_k)^2+b(u_j-u_k)^{-2}\}, \tag{4.48}$$

whose equations of motion,

$$\ddot{u}_j=\sum_{k=1,j\neq j}^{n}[a(u_j-u_k)+b(u_j-u_k)^{-3}], \tag{4.49}$$

yield (4.46) via (3.5).

E. Model five

The Newtonian equations of motion for this system read as follows:

$$\ddot{\vec{r}}_j=[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2+(\alpha+\hat{\alpha}\hat{z}\wedge)\dot{\vec{r}}_j+2\sum_{k=1,k\neq j}^{n}(\rho_{jk}-\hat{\rho}_{jk}\hat{z}\wedge)$$
$$\times[\dot{\vec{r}}_j(\dot{\vec{r}}_k\cdot\vec{r}_j)+\dot{\vec{r}}_k(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_k)], \tag{4.50}$$

with

$$\rho_{jk}=\log(r_j/r_k)\{[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}^{-1}, \tag{4.51a}$$

$$\hat{\rho}_{jk}=(\theta_j-\theta_k)\{[\log(r_j/r_k)]^2+(\theta_j-\theta_k)^2\}^{-1}. \tag{4.51b}$$

Hence they contain velocity-dependent one- and two-body forces [to write the latter we also employed circular coordinates; see (4.51) and Sec. II].

These equations of motion are clearly invariant under rotations in the plane ($\theta_j\to\theta_j+\theta_0$, with θ_0 an arbitrary real constant), as well as under rescaling ($\vec{r}_j\to c\vec{r}_j$, with c an arbitrary real constant). They contain 2 arbitrary (real) ''coupling constants,'' α and $\hat{\alpha}$; for $\alpha=0$ and $\hat{\alpha}\neq 0$, the generic motion is *completely periodic*, with period $T=2\pi/|\hat{\alpha}|$ (this is implied by the Corollary reported at the end of Sec. II; see below).

The ''complex plane'' version of the equations of motion (4.50) reads as

$$\ddot{z}_j=\dot{z}_j^2/z_j+a\dot{z}_j+2\sum_{k=1,k\neq j}^{n}\dot{z}_j\dot{z}_k/[z_j\log(z_j/z_k)], \tag{4.52}$$

with

$$a=\alpha+i\hat{\alpha}. \tag{4.53}$$

It admits the (complex) similarity solution $z_j(t)=z_0\exp\{c[b+\exp(at)]^{1/n}\exp(2\pi ik/n)\}$, with z_0, c and b arbitrary complex constants ($z_0\neq 0$), and the index k related to the index j via an arbitrary permutation of the n positive integers from 1 to n.

The many-body system (4.50) is Hamiltonian. Indeed the (complex) Hamiltonian,

$$h(\mathbf{z},\mathbf{Z})=\sum_{j=1}^{n}\left\{\exp(sz_jZ_j)/\left[\prod_{k=1,k\neq j}^{n}\log(z_j/z_k)\right]-(a/s)\log(z_j)\right\}, \tag{4.54}$$

yields, via (2.14), the (complex) equations of motion (4.52). The corresponding expressions for the Lagrangian $l(\mathbf{z},\dot{\mathbf{z}})$ and for the canonical momenta Z_j read as

$$l(\mathbf{z},\dot{\mathbf{z}})=\sum_{j=1}^{n}\left\{[\dot{z}_j/(sz_j)]\left[-1+\log[\dot{z}_j/(sz_j)]+\sum_{k=1,k\neq j}^{n}\log\log(z_j/z_k)\right]+(a/s)\log(z_j)\right\}, \tag{4.55}$$

$$Z_j=(sz_j)^{-1}\left\{\log[\dot{z}_j/(sz_j)]+\sum_{k=1,k\neq j}^{n}\log\log(z_j/z_k)\right\}. \tag{4.56}$$

Note the presence in these expressions of the arbitrary (complex) constant s, which does not appear in the equations of motion (4.52).

The solvability of this model is demonstrated via (3.5), which transforms (4.52) into the solvable[8,1,4,7] equations of motion,

$$\ddot{u}_j=a\dot{u}_j+2\sum_{k=1,k\neq j}^{n}\dot{u}_j\dot{u}_k/(u_j-u_k). \tag{4.57}$$

Note that, except for notational changes, these equations of motion correspond to (the special case of) (4.34b) with $B=L=M=0$.

F. Model six

The Newtonian equations of motion for this system read as follows:

$$\ddot{\vec{r}}_j=[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2+(\alpha+\hat{\alpha}\hat{z}\wedge)\{\vec{r}_{j+1}-[2\vec{r}_j(\vec{r}_j\cdot\vec{r}_{j-1})-\vec{r}_{j-1}r_j^2]/r_{j-1}^2\}. \tag{4.58}$$

Hence they feature velocity-dependent one-body forces and velocity-independent ''nearest-neighbor'' forces.

These equations of motion are clearly invariant under rotation and rescaling; they are not invariant under translations. They contain (rather trivially, see below) the 2 arbitrary (real) ''coupling constants'' α and $\hat{\alpha}$.

They possess the following ''similarity solution'' (written in circular coordinates; see Sec. II):

$$r_j(t)=r_0\,\exp[(u+ju')t],\quad \theta_j(t)=\theta_0+(v+jv')t. \tag{4.59}$$

Here r_0,θ_0,u,u',v and v' are 6 arbitrary (real) constants ($r_0>0$).

The equations of motion (4.58) are Hamiltonian and Lagrangian; see (2.17), (2.20) and (2.23), with

$$H(\vec{\mathbf{r}},\vec{\mathbf{R}})=(1/2)\sum_j[(\vec{R}_j\cdot\vec{r}_j)^2-(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_j)^2]+V(\vec{\mathbf{r}}), \tag{4.60a}$$

$$\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}})=\sum_j(\vec{R}_j\cdot\vec{r}_j)(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_j)+\hat{V}(\vec{\mathbf{r}}), \tag{4.60b}$$

$$\mathcal{L}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}})=(1/2)\sum_j[(\dot{\vec{r}}_j\cdot\vec{r}_j)^2-(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j)^2]/r_j^4-V(\vec{\mathbf{r}}), \tag{4.61a}$$

$$\tilde{\mathcal{L}}(\vec{\mathbf{r}},\dot{\vec{\mathbf{r}}})=-\sum_j(\dot{\vec{r}}_j\cdot\vec{r}_j)(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j)/r_j^4-\hat{V}(\vec{\mathbf{r}}), \tag{4.61b}$$

$$V(\vec{\mathbf{r}})=(\alpha^2+\hat{\alpha}^2)^{-1}\sum_j\ [-\alpha(\vec{r}_j\cdot\vec{r}_{j-1})+\hat{\alpha}(\hat{z}\cdot\vec{r}_j\wedge\vec{r}_{j-1})]/r_{j-1}^2, \tag{4.62a}$$

$$\hat{V}(\vec{\mathbf{r}})=(\alpha^2+\hat{\alpha}^2)^{-1}\sum_j\ [\hat{\alpha}(\vec{r}_j\cdot\vec{r}_{j-1})+\alpha(\hat{z}\cdot\vec{r}_j\wedge\vec{r}_{j-1})]/r_{j-1}^2, \tag{4.62b}$$

$$\vec{R}_j=[(\dot{\vec{r}}_j\cdot\vec{r}_j)+(\hat{z}\cdot\dot{\vec{r}}_j\wedge\vec{r}_j)\hat{z}\wedge]\vec{r}_j/r_j^4. \tag{4.63}$$

The ''complex-plane'' avatar of (4.58) reads as follows:

$$\ddot{z}_j=\dot{z}_j^2/z_j+a(z_{j+1}-z_j^2/z_{j-1}), \tag{4.64}$$

with $a=\alpha+i\hat{\alpha}$. It yields, via the position

$$z_j=a^j\ \exp(u_j), \tag{4.65}$$

the *integrable* ''Toda'' equations,[6]

$$\ddot{u}_j=\exp(u_{j+1}-u_j)-\exp(u_j-u_{j-1}). \tag{4.66}$$

G. Model seven

The Newtonian equations of motion for this system read as follows:

$$\begin{aligned}
\ddot{\vec{r}}_j=&(\alpha+\hat{\alpha}\hat{z}\wedge)\dot{\vec{r}}_j+[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2\\
&-\{\dot{\vec{r}}_j[\dot{\vec{r}}_{j+1}\cdot(\vec{r}_j-\vec{r}_{j+1})]+\dot{\vec{r}}_{j+1}[\dot{\vec{r}}_j\cdot(\vec{r}_j-\vec{r}_{j+1})]-(\vec{r}_j-\vec{r}_{j+1})(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_{j+1})\}/|\vec{r}_j-\vec{r}_{j+1}|^2\\
&-(\dot{\vec{r}}_j\{(\dot{\vec{r}}_{j-1}\cdot\vec{r}_j)[(\vec{r}_j\cdot\vec{r}_{j-1})+r_{j-1}^2]+2(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1})[r_j^2-2(\vec{r}_j\cdot\vec{r}_{j-1})]\}\\
&+\dot{\vec{r}}_{j-1}\{(\dot{\vec{r}}_j\cdot\vec{r}_j)[(\vec{r}_j\cdot\vec{r}_{j-1})+r_{j-1}^2]+2(\dot{\vec{r}}_j\cdot\vec{r}_{j-1})[r_j^2-2(\vec{r}_j\cdot\vec{r}_{j-1})]\}\\
&+\vec{r}_j\{(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_{j-1})[r_{j-1}^2-2(\vec{r}_j\cdot\vec{r}_{j-1})+(\dot{\vec{r}}_j\cdot\vec{r}_j)(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1})+(\dot{\vec{r}}_j\cdot\vec{r}_{j-1})[(\dot{\vec{r}}_{j-1}\cdot\vec{r}_j)\\
&-4(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1})]\}+\vec{r}_{j-1}\{(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_{j-1})[-r_j^2+2(\vec{r}_j\cdot\vec{r}_{j-1})]+(\dot{\vec{r}}_j\cdot\vec{r}_j)[(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1})-(\dot{\vec{r}}_{j-1}\cdot\vec{r}_j)]\\
&+(\dot{\vec{r}}_j\cdot\vec{r}_{j-1})[(\dot{\vec{r}}_j\cdot\vec{r}_{j-1})+(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1})]\})/(3r_{j-1}^2|\vec{r}_j-\vec{r}_{j-1}|^2).
\end{aligned} \tag{4.67}$$

Hence they feature velocity-dependent one-body and ''nearest-neighbor'' forces.

These equations of motion are clearly invariant under rotation and rescaling; they are not invariant under translation. They contain the 2 arbitrary (real) ''coupling constants'' α and $\hat{\alpha}$; if $\alpha=0$ and $\hat{\alpha}\neq 0$, the generic solution of (4.67) is *completely periodic*, with period $T=2\pi/|\hat{\alpha}|$ (this remarkable fact is implied by the results written below, together with the Corollary reported at the end of Sec. II).

The ''complex-plane'' version of (4.67) reads as follows:

$$\ddot{z}_j=a\dot{z}_j+\dot{z}_j^2/z_j-\dot{z}_j\dot{z}_{j+1}/(z_j-z_{j+1})-\dot{z}_j\dot{z}_{j-1}z_j/[z_{j-1}(z_j-z_{j-1})], \tag{4.68}$$

with $a=\alpha+i\hat{\alpha}$, and it admits the ''similarity solution''

$$z_j(t)=z_0[\varphi(t)]^{\lambda+j}, \tag{4.69a}$$

with $\varphi(t)$ any (complex) solution of the (complex) transcendental equation,

$$\varphi(t)[1-\varphi(t)]^{-1}\ \exp\{[1-\varphi(t)]^{-1}\}=c\ \exp[b\ \exp(at)]. \tag{4.69b}$$

Here z_0, λ, c and b are 4 arbitrary complex constants.

There are two Hamiltonians, and correspondingly two Lagrangians, that yield the equations of motion (4.68) via (2.14) and (2.21):

$$h^{(+)}(\mathbf{z},\mathbf{Z}) = \sum_j \{\exp(sz_jZ_j)[1-(z_{j+1}/z_j)] - (a/s)\log(z_j)\}, \tag{4.70a}$$

$$h^{(-)}(\mathbf{z},\mathbf{Z}) = \sum_j \{\exp(sz_jZ_j)[1-(z_j/z_{j+1})] - (a/s)\log(z_j)\}, \tag{4.71a}$$

$$l^{(+)}(\mathbf{z},\dot{\mathbf{z}}) = \sum_j [\dot{z}_j/(sz_j)][-1+\log\{\dot{z}_j/[s(z_j-z_{j+1})\} + (a/s)\log(z_j)], \tag{4.70b}$$

$$l^{(-)}(\mathbf{z},\dot{\mathbf{z}}) = \sum_j [\dot{z}_j/(sz_j)][-1+\log\{\dot{z}_j z_{j-1}/[s(z_j(z_{j+1}-z_j)]\} + (a/s)\log(z_j)]. \tag{4.71b}$$

Note that these expressions contain the arbitrary (complex) constant s, that does not appear in the equations of motion (4.68). The main hint to find them came from Refs. 4, 7 and especially 9.

It is a simple exercise to obtain from these formulas, using the results reported in Sec. II, explicit expressions for real Hamiltonians and Lagrangians that yield, via (2.17), (2.20) and (2.23), the real equations of motion in the plane (4.67).

The *integrability* of the equations of motion (4.67) and (4.68) is entailed by the fact that, via (3.5), the latter read as follows:

$$\ddot{u}_j = a\dot{u}_j - \dot{u}_j\dot{u}_{j+1}/[1-\exp(u_j-u_{j+1})] + \dot{u}_j\dot{u}_{j-1}/[1-\exp(u_{j-1}-u_j)], \tag{4.72}$$

namely they coincide with (a special case of) the equations of motion of the integrable ''Ruijsenaars–Toda model'' [see e.g., Eq. (3.5.4a) of Ref. 3, or Eq. (1.3) of Ref. 9 and the Proposition of Sec. II].

Note that ''more general'' versions of (4.72) or (4.68) are easily obtained via simple transformations such as $u_j \to u_j + bt + j\log(c)$, $z_j \to cj\exp(bt)z_j$, with b and c arbitrary complex constants. The diligent reader may easily write down the generalized versions of (4.72) and (4.68) entailed by such simple transformations, and with somewhat more effort the corresponding versions of (4.67), as well as the Hamiltonians and Lagrangians that yield these generalized equations.

H. Model eight

The Newtonian equations of motion for this system read as follows:

$$\ddot{\vec{r}}_j = (1+\gamma+\hat{\gamma}\hat{z}\wedge)[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j) - \vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2 + (\alpha+\hat{\alpha}\hat{z}\wedge)\dot{\vec{r}}_j + (\beta+\hat{\beta}\hat{z}\wedge)\vec{r}_j - \sum_{k=j\pm1} r_{jk}^{-2}[(\gamma+\hat{\gamma}\hat{z}\wedge)$$
$$\times[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_{jk}) - \vec{r}_{jk}\dot{r}_j^2] + (\alpha+\hat{\alpha}\hat{z}\wedge)\{\dot{\vec{r}}_j[r_j^2-(\vec{r}_j\cdot\vec{r}_k)] - \vec{r}_j(\dot{\vec{r}}_j\cdot\vec{r}_k) + \vec{r}_k(\dot{\vec{r}}_j\cdot\vec{r}_j)\} + (\beta+\hat{\beta}\hat{z}\wedge)$$
$$\times\{\vec{r}_j[r_j^2-2(\vec{r}_j\cdot\vec{r}_k)] + \vec{r}_k r_j^2\}]. \tag{4.73}$$

Hence they feature velocity-dependent one-body and ''nearest-neighbor'' forces. The 6 (real) coupling constants $\alpha, \hat{\alpha}, \beta, \hat{\beta}, \gamma, \hat{\gamma}$ are arbitrary.

This system is rotation-invariant. It is not translation-invariant. It is invariant under the rescaling $\vec{r}_j \to c\vec{r}_j$.

The special case with $\gamma=-1$, $\hat{\gamma}=\alpha=\hat{\alpha}=0$ was already known [see Eq. (3.7.12) of Ref. 3)].

The system (4.73) possesses the similarity solution (written in circular coordinates, Sec. II)

$$r_j(t) = r_0\rho^j\exp(ut), \quad \theta_j(t) = \theta_0 + j\eta + vt. \tag{4.74}$$

Here $r_0, \rho, u, \theta_0, \eta$ and v are 6 arbitrary real constants ($r_0>0,\ \rho>0$).

The system

$$\ddot{\vec{r}}_j=(1+\gamma+\hat{\gamma}\hat{z}\wedge)[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2+(\alpha+\hat{\alpha}\hat{z}\wedge)\dot{\vec{r}}_j$$
$$-\sum_{k=j\pm1}\sum r_{jk}^{-2}(\gamma+\hat{\gamma}\hat{z}\wedge)[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_{jk})-\vec{r}_{jk}|\dot{\vec{r}}_j|^2] \tag{4.75}$$

can be obtained from (4.73) by first setting $\alpha=\hat{\alpha}=\beta=\hat{\beta}=0$, and by then applying the Proposition reported at the end of Sec. II (see below). Hence this integrable system, in the special case $\alpha=0,\hat{\alpha}\neq0$, features only completely periodic solutions (the generic solution has period $T=2\pi/|\hat{\alpha}|$). In the special (translation-invariant) case $\gamma=-1,\hat{\gamma}=0$, this system was already known [see Eq. (3.7.11) of Ref. 3]. The equations of motion (4.75) possess the similarity solution

$$r_j(t)=r_0\rho^j\ \exp\{\exp(\alpha t)[\lambda\ \cos(\hat{\alpha}t)-\mu\ \sin(\hat{\alpha}t)]\}, \tag{4.76a}$$

$$\theta_j(t)=\theta_0+j\,\eta+\exp(\alpha t)[\mu\ \cos(\hat{\alpha}t)+\lambda\ \sin(\hat{\alpha}t)], \tag{4.76b}$$

featuring the 6 arbitrary real constants $r_0,\rho,\lambda,\mu,\theta_0$ and η ($r_0>0,\rho>0$). The corresponding motion is, as $t\to+\infty$, rather unwieldy if $\alpha>0$ and $\hat{\alpha}\neq0$; it is more reasonable if $\alpha<0$, and it is of course completely periodic with period $T=2\pi/|\hat{\alpha}|$ if $\alpha=0,\ \hat{\alpha}\neq0$.

The ''complex-plane'' version of (4.73) reads as follows:

$$\ddot{z}_j=(1+c)\dot{z}_j^2/z_j+a\dot{z}_j+bz_j-(c\dot{z}_j^2+a\dot{z}_jz_j+bz_j^2)[(z_j-z_{j+1})^{-1}+(z_j-z_{j-1})^{-1}], \tag{4.77a}$$

$$a=\alpha+i\hat{\alpha},\quad b=\beta+i\hat{\beta},\quad c=\gamma+i\hat{\gamma}. \tag{4.77b}$$

These equations of motion correspond, via (2.14) and (2.21), to the following Hamiltonian $h(\mathbf{z},\mathbf{Z})$ and Lagrangian $l(\mathbf{z},\dot{\mathbf{z}})$:

$$h(\mathbf{z},\mathbf{Z})=\sum_j\ \{\lambda z_jZ_j+\nu[\mu\ \log\sinh(z_jZ_j/\nu)+\rho\ \log(1-z_{j+1}/z_j)]\}, \tag{4.78a}$$

$$l(\mathbf{z},\dot{\mathbf{z}})=\nu\sum_j\ \left\{\frac{1}{2}\mu[(w_j+1)\log(w_j+1)-(w_j-1)\log(w_j-1)]-\rho\ \log(1-z_{j+1}/z_j)\right\}, \tag{4.78b}$$

$$w_j=[(\dot{z}_j/z_j)-\lambda]/\mu, \tag{4.78c}$$

entailing [see (2.24b)]

$$Z_j=(\nu/z_j)\,\mathrm{arc\ cotanh}(w_j)=\frac{1}{2}(\nu/z_j)\log[(w_j+1)/(w_j-1)]. \tag{4.78d}$$

The 4 (complex) constants λ,μ,ν,ρ are related to the 3 (complex) constants a,b,c [see (4.77b)] as follows:

$$c=-\rho/\mu,\quad a=2\rho\lambda,\quad b=\rho\mu(1-\lambda^2). \tag{4.79}$$

The integrability of the system (4.77) is demonstrated via the position

$$z_j=\exp(-cu_j). \tag{4.80}$$

Thereby the equations of motion (4.77) take the form

$$\ddot{u}_j=R(\dot{u}_j)[g(u_{j+1}-u_j)-g(u_j-u_{j-1})], \tag{4.81a}$$

with

$$R(v)=Ev^2+Av+B, \tag{4.81b}$$

$$g(u)=[g_1-g_2\ \exp(-cu)]/[1-\exp(-cu)], \tag{4.81c}$$

entailing

$$g'(u)=E[g(u)-g_1][g(u)-g_2], \tag{4.81d}$$

and

$$c=E(g_2-g_1), \quad a=A(g_2-g_1), \quad b=B(g_2-g_1). \tag{4.81e}$$

These equations have been identified as integrable ones by Yamilov (on the basis of previous work with Shabat).[10] The Hamiltonian (4.78) has also been obtained from Ref. 10.

I. Models nine and ten

The Newtonian equations of motion of ''model nine'' read as follows:

$$\begin{aligned}\ddot{\vec{r}}_j=&[2\dot{\vec{r}}_j(\dot{\vec{r}}_j\cdot\vec{r}_j)-\vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2+(\beta+\hat{\beta}\hat{z}\wedge)\cdot\{-[2\dot{\vec{r}}_j(-a_j^{(+)}r_j^2+4b_jc_j^{(+)})+\dot{\vec{r}}_{j+1}(-|\dot{\vec{r}}_j|^2r_j^2\\&+4b_j^{(+)2})-2\vec{r}_j(|\dot{\vec{r}}_j|^2c^{(+)}+2a^{(+)}b^{(+)})]/(3r_j^4)+[2\dot{\vec{r}}_j(-a_j^{(-)}r_{j-1}^2+4b_j^{(-)}c_j^{(-)})\\&+\dot{\vec{r}}_{j-1}(-|\dot{\vec{r}}_j|^2r_{j-1}^2+4b_j^{(-)2})-2\vec{r}_{j-1}(|\dot{\vec{r}}_{j-1}|^2c^{(-)}+2a^{(-)}b^{(-)})]/(3r_{j-1}^4)\},\end{aligned} \tag{4.82a}$$

with the short-hand notation

$$\begin{gathered}a_j^{(\pm)}\equiv(\dot{\vec{r}}_j\cdot\dot{\vec{r}}_{j\pm1}); \quad b_j^{(+)}\equiv(\dot{\vec{r}}_j\cdot\vec{r}_j), \quad b_j^{(-)}\equiv(\dot{\vec{r}}_j\cdot\vec{r}_{j-1});\\ c_j^{(+)}\equiv(\dot{\vec{r}}_{j+1}\cdot\vec{r}_j), \quad c_j^{(-)}\equiv(\dot{\vec{r}}_{j-1}\cdot\vec{r}_{j-1}).\end{gathered} \tag{4.82b}$$

Hence they feature one-body forces quadratic in the velocities and ''nearest-neighbor'' forces cubic in the velocities, and they contain the 2 arbitrary (real) coupling constants β and $\hat{\beta}$.

These equations of motion are rotation- and scaling-invariant. They are not translation-invariant.

They possess the ''similarity solution'' (4.74). The ''complex plane'' version of (4.82) reads as

$$\ddot{z}_j=\dot{z}_j^2/z_j-b\dot{z}_j^2[(\dot{z}_{j+1}/z_j^2)-(\dot{z}_{j-1}/z_{j-1}^2)], \tag{4.83a}$$

with

$$b=\beta+i\hat{\beta}. \tag{4.83b}$$

These equations are Hamiltonian and Lagrangian [see (2.14) and (2.21)], with

$$h(\mathbf{z},\mathbf{Z})=-\sum_j \log\{z_j[Z_j-b(z_j/z_{j-1})]\}, \tag{4.84a}$$

$$l(\mathbf{z},\dot{\mathbf{z}})=\sum_j [b(\dot{z}_j/z_{j-1})-\log(\dot{z}_j/z_j)], \tag{4.84b}$$

$$Z_j=-(\dot{z}_j)^{-1}+b/z_{j-1}. \tag{4.84c}$$

The integrability of this model is demonstrated via the position

$$z_j = b^{-j} \exp(u_j), \tag{4.85}$$

which transforms (4.83a) into

$$\ddot{u}_j = -\dot{u}_j^2[\dot{u}_{j+1} \exp(u_{j+1}-u_j) - \dot{u}_{j-1} \exp(u_j - u_{j-1})]. \tag{4.86}$$

The integrability of these evolution equations has been recently demonstrated by Suris.[9] The Hamiltonian and Lagrangian (4.84) have also been obtained from Ref. 9.

The Newtonian equations of motion of ''model ten'' read as follows:

$$\begin{aligned}\ddot{\vec{r}}_j =& [2\dot{\vec{r}}_j(\dot{\vec{r}}_r \cdot \vec{r}_j) - \vec{r}_j|\dot{\vec{r}}_j|^2]/r_j^2 + (\beta + \hat{\beta}\hat{z}\wedge)\{\dot{\vec{r}}_{j+1} + \{\dot{\vec{r}}_{j-1}[r_j^2 r_{j-1}^2 - 4(\vec{r}_j \cdot \vec{r}_{j-1})^2] \\ &+ 2\vec{r}_j[(\dot{\vec{r}}_{j-1} \cdot \vec{r}_j) r_{j-1}^2 - 4(\dot{\vec{r}}_{j-1}\vec{r}_{j-1})(\vec{r}_j \cdot \vec{r}_{j-1})] + 2\vec{r}_{j-1}[2(\dot{\vec{r}}_{j-1} \cdot \vec{r}_j)(\vec{r}_j \cdot \vec{r}_{j-1}) \\ &+ (\dot{\vec{r}}_{j-1} \cdot \vec{r}_{j-1}) r_j^2]\}/(3r_{j-1}^4)\} + (\beta^2 - \hat{\beta}^2 + 2\beta\hat{\beta}\hat{z}\wedge)\{[\vec{r}_j r_{j+1}^2 - 2\vec{r}_{j+1}(\vec{r}_{j+1} \cdot \vec{r}_j)]/r_j^2 \\ &+ \{\vec{r}_j[4(\vec{r}_j \cdot \vec{r}_{j-1})^2 - r_j^2 r_{j-1}^2] - 2\vec{r}_{j-1}(\vec{r}_j \cdot \vec{r}_{j-1}) r_j^2\}/(3r_{j-1}^4)\},\end{aligned} \tag{4.87}$$

or equivalently, in complex notation,

$$\ddot{z}_j = \dot{z}_j^2/z_j + b[\dot{z}_{j+1} - \dot{z}_{j-1}(z_j/z_{j-1})^2] + b^2[-z_{j+1}^2/z_j + z_j^3/z_{j-1}^2], \tag{4.88}$$

with $b = \beta + i\hat{\beta}$.

The forces appearing in the right-hand side of the equations of motion (4.87) feature a one-body term quadratic in the velocities, a ''nearest-neighbor'' term linear in the velocities and another ''nearest-neighbor'' velocity-independent term. These equations contain (linearly and quadratically) the 2 arbitrary real constants β and $\hat{\beta}$.

These equations of motion possess the similarity solution (4.74).

They are Hamiltonian. As for the preceding ''model nine,'' we write the corresponding Hamiltonian and Lagrangian only in complex notation, see (2.14) and (2.21), leaving to the diligent reader the task to derive, via (2.15), (2.16) (2.19) and (2.22), the explicit expressions of the real Hamiltonians and Lagrangians that yield directly (4.87):

$$h(\mathbf{z}, \mathbf{Z}) = \sum_j (s z_j^2 Z_j^2 + a z_{j+1} Z_j), \tag{4.89a}$$

$$l(\mathbf{z}, \dot{\mathbf{z}}) = (4s)^{-1} \sum_j [(\dot{z}_j/z_j)^2 - 2a\dot{z}_j z_{j+1}/z_j^2 + a^2(z_{j+1}/z_j)^2], \tag{4.89b}$$

$$Z_j = (\dot{z}_j - a z_{j+1})/(2s z_j^2). \tag{4.89c}$$

Here s is an arbitrary complex constant.

The *integrability* of this model is demonstrated via the position (4.85), which transforms (4.88) into

$$\ddot{u}_j = \dot{u}_{j+1} \exp(u_{j+1}-u_j) - \dot{u}_{j-1} \exp(u_j - u_{j-1}) - \exp[2(u_{j+1}-u_j)] + \exp[2(u_j - u_{j-1})], \tag{4.90}$$

a dynamical system whose integrability has been recently demonstrated by Suris.[9] The Hamiltonian and the Lagrangians (4.89) have also been obtained from Ref. 9.

J. Model eleven

This model is Hamiltonian and completely solvable; we exhibit below its *general solution*. It is, however, a somewhat degenerate system, as indeed demonstrated by its complete solvability, although its motions are far from trivial, see below; and there is no simple way, in the general n-body case, to write explicitly its Newtonian equations of motion.

Its Hamiltonians in the plane, and the corresponding equations of motion, read, in the notation of Sec. II, as follows:

$$H(\vec{\mathbf{r}},\vec{\mathbf{R}})=(1/2)\sum_{j,k=1}^{n}\{\alpha[(\vec{R}_j\cdot\vec{r}_{jk})(\vec{R}_k\cdot\vec{r}_{jk})-(\hat{z}\cdot\vec{R}_j\cdot\wedge\vec{r}_{jk})(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})]-\hat{\alpha}[(\vec{R}_j\cdot\vec{r}_{jk})(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})$$
$$+(\vec{R}_k\cdot\vec{r}_{jk})(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_{jk})]\}+\sum_{j=1}^{n}[\beta(\vec{r}_j\cdot\vec{R}_j)-\hat{\beta}(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_j)], \tag{4.91}$$

$$\tilde{H}(\vec{\mathbf{r}},\vec{\mathbf{R}})=(1/2)\sum_{j,k=1}^{n}\{\hat{\alpha}[(\vec{R}_j\cdot\vec{r}_{jk})(\vec{R}_k\cdot\vec{r}_{jk})-(\hat{z}\cdot\vec{R}_j\cdot\wedge\vec{r}_{jk})(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})]+\alpha[(\vec{R}_j\cdot\vec{r}_{jk})(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})$$
$$+(\vec{R}_k\cdot\vec{r}_{jk})(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_{jk})]\}+\sum_{j=1}^{n}[\hat{\beta}(\vec{r}_j\cdot\vec{R}_j)+\beta(\hat{z}\cdot\vec{R}_j\wedge\vec{r}_j)], \tag{4.92}$$

$$\dot{\vec{r}}_j=(\beta+\hat{\beta}\hat{z}\wedge)\vec{r}_j+(\alpha+\hat{\alpha}\hat{z}\wedge)\sum_{k=1}^{n}[\vec{r}_{jk}(\vec{R}_k\cdot\vec{r}_{jk})+\hat{z}\wedge\vec{r}_{jk}(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})], \tag{4.93a}$$

$$\dot{\vec{R}}_j=-(\beta+\hat{\beta}\hat{z}\wedge)\vec{R}_j-2(\alpha+\hat{\alpha}\hat{z}\wedge)\sum_{k=1}^{n}[\vec{R}_j(\vec{R}_k\cdot\vec{r}_{jk})+\hat{z}\wedge\vec{R}_j(\hat{z}\cdot\vec{R}_k\wedge\vec{r}_{jk})]. \tag{4.93b}$$

These equations contain the 4 arbitrary coupling constants $\alpha,\hat{\alpha},\beta,\hat{\beta}$; the solvability of the model would not be marred by an (arbitrary) time-dependence of these quantities (see below), but we hereafter assume, for simplicity, that they are all time-independent.

Note that these Hamiltonians, and (of course) the corresponding equations of motion, are rotation-invariant. They are moreover translation-invariant, namely invariant under the transformation $\vec{r}_j=\vec{r}_j+\vec{r}_0$ with $\vec{r}_0$ an arbitrary constant two-vector, iff $\beta=\hat{\beta}=0$.

A somewhat more general, but still solvable, version of this model exists, which features appropriate combinations of the vectors $\vec{r}_j$ and $\vec{r}_k$ in place of $\vec{r}_{jk}\equiv\vec{r}_j-\vec{r}_k$, and therefore lacks translation-invariance even when $\beta=\hat{\beta}=0$. We forsake to display and treat this case, which the diligent reader might easily retrieve from the results reported below.

The complex-plane versions of (4.91)–(4.93) read as follows (with $a=\alpha+i\hat{\alpha},b=\beta+i\hat{\beta}$):

$$h(\mathbf{z},\mathbf{Z})=(1/2)a\sum_{j,k=1}^{n}Z_jZ_k(z_j-z_k)^2+b\sum_{j=1}^{n}z_jZ_j, \tag{4.94}$$

$$\dot{z}_j=bz_j+a\sum_{k=1}^{n}Z_k(z_j-z_k)^2, \tag{4.95a}$$

$$\dot{Z}_j=-bZ_j-2aZ_j\sum_{k=1}^{n}Z_k(z_j-z_k). \tag{4.95b}$$

Hereafter we focus on these equations; the transformation of the results, presented below in complex-plane notation, into the corresponding results in real-plane language, is a trivial task; see Sec. II.

The equations of motion (4.95) can be generalized or simplified via the following transformations:

$$z_j(t)=\varphi(t)\tilde{z}_j(\tau),\quad Z_j(t)=\Phi(t)\tilde{Z}_j(\tau),\quad \tau\equiv\tau(t), \tag{4.96}$$

which yield

$$\tilde{z}_j' = \tilde{b}\tilde{z}_j + \tilde{a}\sum_{k=1}^{n} \tilde{Z}_k(\tilde{z}_j - \tilde{z}_k)^2, \tag{4.97a}$$

$$\tilde{Z}_j' = \tilde{c}\tilde{Z}_j - 2\tilde{a}\tilde{Z}_j\sum_{k=1}^{n} \tilde{Z}_k(\tilde{z}_j - \tilde{z}_k), \tag{4.97b}$$

with

$$\tilde{b} = (b - \dot{\varphi}/\varphi)/\dot{\tau}, \quad \tilde{c} = (b + \dot{\Phi}/\Phi)/\dot{\tau}, \quad \tilde{a} = a\varphi\Phi/\dot{\tau}. \tag{4.98}$$

Here of course the prime denotes differentiation with respect to the variable τ.

It is thus seen that, by appropriate choice of the 3 *a priority* arbitrary functions φ, Φ and τ, one can control the more general version (4.97) of the equations of motion, where $\tilde{a}(\tau)$, $\tilde{b}(\tau)$ and $\tilde{c}(\tau)$ are 3 essentially arbitrary functions of the independent variable τ, by reducing them to the simpler version

$$\dot{z}_j = \sum_{k=1}^{n} Z_k(z_j - z_k)^2, \tag{4.99a}$$

$$\dot{Z}_j = -2Z_j\sum_{k=1}^{n} Z_k(z_j - z_k). \tag{4.99b}$$

In particular, (4.95) can be reduced (up to a trivial notational change) to (4.99), by setting $\varphi(t) = \exp(bt)$, $\Phi(t) = \exp(-bt)$, $\tau = at$. Hence hereafter our attention will be confined to solving (4.99).

It is now convenient to note that these equations of motion (4.99) can be rewritten in the form

$$\dot{z}_j = c_0 z_j^2 - 2c_1 z_j + c_2, \tag{4.100a}$$

$$\dot{Z}_j = -2Z_j(c_0 z_j - c_1), \tag{4.100b}$$

by setting

$$c_m = \sum_{k=1}^{n} Z_k z_k^m, \quad m = 0,1,2. \tag{4.101}$$

Note the degeneracy of the equations of motion (4.100a), which imply that, for $n > 3$, the quantities $z_j(0)$ and $\dot{z}_j(0)$ cannot be all assigned independently.

It is moreover easy to verify that (4.101) and (4.100) entail that the 3 quantities c_m are constants of motion:

$$\dot{c}_m = 0, \quad m = 0,1,2. \tag{4.102}$$

It is then a trivial matter to solve (4.100):

$$z_j(t) = (c_1/c_0)\{1 - \gamma\,\mathrm{cotanh}[2\gamma c_1 t + \alpha_j]\}, \tag{4.103a}$$

$$Z_j(t) = Z_j(0)\sinh(\alpha_j)/\sinh(2\gamma c_1 t + \alpha_j), \tag{4.103b}$$

$$\alpha_j = \mathrm{arctanh}\{[1 - (c_0/c_1)z_j(0)]/\gamma\}, \tag{4.103c}$$

$$\gamma = (1 - c_0 c_2/c_1^2)^{1/2}. \tag{4.103d}$$

Here of course all the constants $c_0, c_1, c_2, \gamma, \alpha_j, z_j(0)$ and $Z_j(0)$ are complex; this accounts for the richness of the motions entailed by (103), whose detailed analysis is left as an easy task for the diligent reader.

Let us end by emphasizing that the findings reported in this Subsection are closely connected, for instance via the results of Sec. III, with the results of Ref. 11.

V. OUTLOOK

In this paper we have exhibited several *integrable* or *solvable* dynamical systems, whose equations of motion are naturally interpretable as the *Newtonian equations of motion for a many-body problem in the plane*. It is clear from the techniques we have introduced and employed how other models sharing these properties could be obtained. As mentioned above, to select those we have displayed, we have been guided mainly by ''aesthetic'' considerations, trying to strike a fair balance between the contradictory requirements to avoid overwhelming the reader with too many examples yet to exhibit a representative sample of the systems that are treatable. Analogous ''aesthetic'' criteria have also been employed to balance the contradictory requirements, in presenting each model, to seek maximal generality or *vice versa* maximal simplicity. We have tried in this respect to make our presentation user-friendly, having also in mind potential users interested rather in applications than in the underlying theory. It is clear that many of the models presented above could be further generalized by appropriate changes of variables, such as $z_j \to \tilde{z}_j = (z_j)^\lambda$ or $z_j \to \tilde{z}_j = f(t)\zeta_j(\tau) + g(t), \tau \equiv \tau(t)$.

Our main purpose in this paper has been to present a simple technique suitable to manufacture integrable or solvable many-body problems in the plane, and to exhibit several such models, rather than to study their solutions. We have, however, in most cases exhibited special (''similarity'') solutions, to give at least an idea of the motions entailed by these models. But these results give only a glimpse of the behaviors of these systems, and are certainly inadequate to display all the reachness of the motions exhibited by their solutions (see, for instance, Ref. 1). Hence much work remains to be done, in the guise of a detailed study of the solutions of each model, possibly including their display by appropriate computer graphics.

Let us end by mentioning three additional directions for further research, and two final remarks, which also open vistas of additional research.

Three directions for further research. (i) The study of quantized versions of these models, in particular of those which only feature *completely periodic motions*, for which it is natural to conjecture the existence of quantized versions featuring *equispaced energy spectra*. (ii) The investigation of the limit in which the number n of particles becomes infinite, including in particular the cases in which the particles give rise to continuous distributions (strings in the plane, and perhaps also blobs; for seminal progress in that direction see Ref. 12). (iii) The introduction of related *discretized* models, in which the *real* space and/or time variables are replaced by lattice entries (much work has already been done in this direction for ''one-dimensional'' integrable and solvable models, see e.g., Refs. 9 and 13).

First remark. There is of course a simple trick to transform by complexification *any* analytical problem for the n complex variables $z_j, j = 1, \ldots, n,$ into a *rotation-invariant* problem in the plane for $n+1$ real two-vectors $\vec{r}_j, j = 0,1, \ldots, n$. The trick is to set, instead of $z_j = x_j + iy_j$ (see Sec. II),

$$z_j = \hat{r}_0 \cdot \vec{r}_j + i\hat{z} \cdot \hat{r}_0 \wedge \vec{r}_j, j = 1, \ldots, n, \tag{5.1}$$

with $\hat{r}_0 \equiv \vec{r}_0 / r_0$ the unit two-vector in the direction of $\vec{r}_0$, and to supplement the equations of motion for $\vec{r}_j(t), j = 1, \ldots, n$, obtained via this position from those for $z_j(t)$, with the additional *rotation-invariant* (and trivial!) equation,

$$\dot{\vec{r}}_0 = 0. \tag{5.2}$$

Note that this equation is automatically yielded by any Hamiltonian problem, where the original (complex) Hamiltonian $h(z_j, Z_j; j = 1, \ldots, n)$ generates new (real) Hamiltonians $H(\vec{r}_0, \vec{r}_j, \vec{R}_j; j = 1, \ldots, n)$ and $\hat{H}(\vec{r}_0, \vec{\hat{r}}_j, \vec{\hat{R}}_j; j = 1, \ldots, n)$ via the technique of Sec. II, using (5.1) and correspondingly [instead of (2.15)]

$$z_j=\hat{r}_0\cdot\vec{r}_j+i\hat{z}\cdot\hat{r}_0\wedge\vec{r}_j\,,\quad Z_j=\hat{r}_0\cdot\vec{R}_j-i\hat{z}\cdot\hat{r}_0\wedge\vec{R}_j\,,\quad j=1,\ldots,n. \tag{5.3}$$

Second remark. Let us reemphasize that the main message of this paper is to advocate that the study of integrable systems be extended to the complex plane (''go complex, man!''), especially by appropriately interpreting such an extension, which generally entails a substantial enrichment of the dynamical behaviors displayed by the systems under consideration, in terms of ''physical'' (*real* and *rotation-invariant*) models *in the plane*.

ACKNOWLEDGMENTS

Much of the work reported in this paper was done while visiting the UFR 920 at the Université Paris 6. It is a pleasure to thank professor Jean-Pierre Françoise for the pleasant hospitality there and for many useful discussions, as well as the institutions that have facilitated such visits, including the CNRS in France and the CNR in Italy.

[1] F. Calogero, ''A solvable *n*-body problem in the plane. I,'' J. Math. Phys. **37**, 1735–1759 (1996).
[2] F. Calogero, ''Three solvable many-body problems in the plane,'' Acta Appl. Math. **51**, 93–111 (1998).
[3] F. Calogero, ''Tricks of the trade: relating and deriving solvable and integrable dynamical systems,'' in *Proceedings of the Workshop on Calogero–Moser–Sutherland Models*, Montreal, 10–15 March 1997 (in press).
[4] F. Calogero and J.-P. Françoise, ''Solution of certain integrable dynamical systems with completely periodic trajectories,'' Ann. Inst. H. Poincaré (submitted).
[5] F. Calogero, ''Exactly solvable one-dimensional many-body problems,'' Lett. Nuovo Cimento **13**, 507–511 (1975).
[6] M. A. Olshanetsky and A. M. Perelomov, ''Classical integrable finite-dimensional systems related to Lie algebras,'' Phys. Rep. **71**, 313–400 (1981).
[7] F. Calogero, ''A class of integrable Hamiltonian systems whose solutions are (perhaps) all completely periodic,'' J. Math. Phys. **38**, 5711–5719 (1997).
[8] F. Calogero, ''Motion of poles and zeros of special solutions of nonlinear and linear partial differential equations and related ''solvable'' many-body problems,'' Nuovo Cimento B **43**, 177–241 (1978).
[9] Yu. B. Suris, ''New integrable systems related to the relativistic Toda lattice,'' J. Phys. A **30**, 1745–1761 (1997).
[10] R. I. Yamilov: ''Classification of Toda type scalar lattices,'' in *Nonlinear Evolution Equations and Dynamical Systems-NEEDS' 92*, edited by V. Makhankov, I. Puzynin, and O. Pashaev (World Scientific, Singapore, 1993), pp. 423–431.
[11] F. Calogero, ''An integrable Hamiltonian system,'' Phys. Lett. A **201**, 306–310 (1995).
[12] F. Calogero, ''Motion of strings in the plane: A solvable model. I,'' J. Math. Phys. **38**, 821–829 (1997).
[13] F. W. Nijhoff, V. Kuznetsov, and O. Ragnisco, ''Integrable discretization of the Ruijsenaars–Schneider system,'' Commun. Math. Phys. **176**, 681–700 (1996).

Two families of nonstandard Poisson structures for Newton equations

Krzysztof Marciniak[a] and Stefan Rauch-Wojciechowski[b]
Department of Mathematics, Linköping University, S 581-83 Linköping, Sweden

(Received 10 March 1998; accepted for publication 30 April 1998)

Two families of nonstandard two-dimensional Poisson structures for systems of Newton equations are studied. They are closely related either with separable systems or with the so-called quasi-Lagrangian systems. A theorem characterizing the general form of bi-Hamiltonian formulation for separable systems in two and in n dimensions is formulated and proved. © *1998 American Institute of Physics.* [S0022-2488(98)03810-9]

I. INTRODUCTION

In this paper we study two different types of nonstandard Poisson operators related to systems of Newton equations, i.e., second-order ordinary differential equations (ODEs) of the form $q_{,xx}=F(q)$, where $q=(q_1,\ldots,q_n)$, q_i, $x\in\mathbf{R}$, and with a function F not depending explicitly on the velocity $q_{,x}$ (throughout the whole article *comma* followed by a subscript denotes derivation with respect to the subscript variable, so for example $q_{,xx}=d^2q/dx^2$ and $x_{,t}=dx/dt$). Such systems originate in physics as equations of motion of type acceleration=external force. The study and efforts to solve these equations were over centuries major driving factors for development of classical mechanics and resulted in introducing the Lagrangian and Hamiltonian formalisms.

The majority of Newton equations are not solvable in any explicit sense, but those which are are of considerable interest in applications (Kepler problem, integrable rigid body motions) as well as in theoretical understanding of the formalism of classical mechanics.

If the force F is potential, i.e., if $F=-\partial V/\partial q$, then the Newton equations $q_{,xx}=F(q)$ naturally admit the energy integral of motion and possibly some extra integrals of motion so that they can be solved by quadratures. The most complete characterization of classes of Newton equations solvable by quadratures is given in the framework of their Hamiltonian formulation by the Liouville theorem about completely integrable systems.[1]

Completely integrable systems are very rare and several different tools were developed for their characterization; for instance, the classical separability theory for the Hamilton–Jacobi equation,[1] Kac–Moody algebraic constructions of complete sets of commuting integrals of motion,[2,3] the classical r-matrix approach,[4] or the bi-Hamiltonian approach.[5] All these are quite general and powerful construction schemes which lead to large classes of completely integrable Hamiltonian systems, but very few of these systems correspond to some physically interpretable sets of equations like Newton equations or equations of rigid body motion. It is a highly nontrivial task to construct completely integrable sets of Newton equations.

In the present paper we study the relationship between systems of Newton equations and Poisson operators. More precisely, we present two families of Poisson operators of relatively general form, inspired by two illustrative examples of completely integrable Newton equations. We show how their Poisson property inevitably leads to two classes of Newton equations.

The first family of Poisson operators is related to the classically (pointwise) separable natural Hamiltonian systems $H=\frac{1}{2}p^2+V(q)$ which also admit a second (nonstandard) Hamiltonian formulation. We need to mention that the bi-Hamiltonian formulation for the most important cases of separable systems (elliptic and parabolic separable) has been determined before,[6,7] where for given classes of potentials a bi-Hamiltonian formulation has been found. Here we proceed in the oppo-

[a]On leave of absence from Department of Physics, A. Mickiewicz University, Poznań, Poland. Electronic mail: krmar@mai.liu.se
[b]Electronic mail: strau@mai.liu.se

 © 1998 American Institute of Physics

site direction: we show that a certain type of assumption about the Poisson operator inevitably leads to the class of separable systems.

The second family of Poisson operators describes a completely new class of Newton equations. This new family consists of systems with nonpotential forces and so they have no Lagrangian formulation and thus they have no standard Hamiltonian formulation. Instead, they have what we call a *quasi-Lagrangian* formulation and a certain nonstandard Hamiltonian formulation.

This paper is organized as follows. We start with the mentioned assumptions about the structure of a Poisson operator in a five-dimensional phase space. These assumptions are inspired by two examples (Examples 2.1 and 2.2 below) and by the previous results[6,7] about the bi-Hamiltonian structure of separable natural Hamiltonian systems. In Sec. III we formulate a theorem about bi-Hamiltonian formulation for all two-dimensional separable natural Hamiltonians. This theorem admits a generalization to arbitrary dimension, which is done in Sec. IV, where we give a complete characterization of bi-Hamiltonian formulation for all natural Hamiltonians separable in generalized elliptic coordinates or any of its degenerations[8] which correspond to indecomposable metrics. What happens in the decomposable case is illustrated by Example 4.3. In the last section we introduce and discuss new class of systems—the already mentioned quasi-Lagrangian systems; we explain how they lead to sets of nonpotential Newton equations and how they generalize Example 2.2 coming from the Harry–Dym hierarchy. We also find for them a completely new Hamiltonian formulation.

II. STARTING POINT: A POISSON OPERATOR

Let us consider a five-dimensional phase space $\mathscr{M}$ with variables denoted by (x,y,π,ρ,c) or by (r,w,s,z,d). We remind the reader that an operator Π mapping covector fields (gradients) on $\mathscr{M}$ into vector fields on $\mathscr{M}$ is called a *Poisson operator*, if the corresponding bilinear mapping, defined for any pair of functions $f,g\colon \mathscr{M}\to\mathbf{R}$ by

$$\{f,g\}_\Pi=(\nabla f,\Pi\nabla g)$$

[where $(\cdot,\cdot)$ is the dual map between the cotangent and tangent spaces and ∇ is the gradient operator in M] yields a Poisson bracket, that is, if

(1) it is antisymmetric: $\{f,g\}_\Pi=-\{g,f\}_\Pi$;
(2) it satisfies the Leibniz rule: $\{fg,h\}_\Pi=f\{g,h\}_\Pi+\{f,h\}_\Pi g$; and
(3) it satisfies the Jacobi identity: if h is a third function on $\mathscr{M}$, then

$$\{\{f,g\}_\Pi,h\}_\Pi+\{\{g,h\}_\Pi,f\}_\Pi+\{\{h,f\}_\Pi,g\}_\Pi=0.$$

Thus, an operator Π is Poisson if its matrix is antisymmetric and if the Jacobi identities $\{\{x,y\}_\Pi,\pi\}_\Pi+\text{cycl}=0$, etc. hold for all triples of coordinates (cycl means the cyclic permutation of variables).

In what follows we will not distinguish between Poisson operators and their matrices.

Now let $\mathscr{M}$ be endowed with a standard (degenerate) Poisson operator

$$\Pi_0=\left[\begin{array}{cc|c} 0 & I & 0\\ -I & 0 & 0\\ \hline 0 & 0 & 0\end{array}\right], \tag{2.1}$$

where

$$I=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

is a 2×2 identity matrix. Thus, if $\{\cdot,\cdot\}_{\Pi_0}$ denotes the Poisson bracket induced by (2.1), then $\{x,\pi\}_{\Pi_0}=\{y,\rho\}_{\Pi_0}=\{r,s\}_{\Pi_0}=\{w,z\}_{\Pi_0}=1$ and the remaining elementary Poisson brackets are equal to zero. We introduce two different labelings of the phase space variables since they will correspond to two different families of Poisson structures which we present below.

Let us now consider an *antisymmetric* operator of a quite general form:

$$\left[\begin{array}{cc|cc|c} 0 & 0 & \mu+f_{11} & f_{12} & \\ 0 & 0 & f_{21} & \mu+f_{22} & \\ \hline -\mu-f_{11} & -f_{21} & 0 & F & \mathscr{A} \\ -f_{12} & -\mu-f_{22} & -F & 0 & \\ \hline \multicolumn{4}{c|}{*} & \end{array}\right] \tag{2.2}$$

where $f_{ij}=f_{ij}(x,y)$ [or $f_{ij}=f_{ij}(r,w)$], $i,j=1,2$, are arbitrary functions of the position coordinates, F is an arbitrary function on the phase space not depending on the last coordinate (c or d, respectively), and $\mathscr{A}$ is a column vector of functions on the phase space with the last entry equal to zero due to the antisymmetry of (2.2). The parameter $\mu\in\mathbf{R}$ is introduced in order to study the compatibility of (2.2) with the canonical Poisson operator (2.1). Here and in what follows $*$ denotes those matrix elements, which are determined by the requirement of antisymmetry of a given matrix.

In this paper we investigate the following question: what conditions must be satisfied in order to turn (2.2) into a Poisson operator? It will be shown that the study of this problem leads to new facts about systems of Newton equations, in particular about separable systems.

To motivate further development, let us present two examples:

Example 2.1: The well-known Garnier system,[9]

$$\begin{aligned} 0&=x_{,tt}+(x^2+y^2+c-\lambda_1)x,\\ 0&=y_{,tt}+(x^2+y^2+c-\lambda_2)y, \end{aligned} \tag{2.3}$$

with $x=x(t)$, $y=y(t)$, $\lambda_1,\lambda_2,c=\text{const}$, has the bi-Hamiltonian form[6]

$$\frac{d}{dt}\begin{bmatrix} x\\ y\\ \pi\\ \rho\\ c\end{bmatrix}=\Pi_0\nabla\left(\tfrac{1}{2}\pi^2+\tfrac{1}{2}\rho^2+V(x,y,c)\right)=\Pi_1\nabla c \tag{2.4}$$

(where $\pi=x_{,t}$, $\rho=y_{,t}$) with the second Poisson operator

$$\Pi_1=\left[\begin{array}{cc|cc|c} 0 & 0 & -\frac{1}{2}x^2+\lambda_1 & -\frac{1}{2}xy & \pi\\ 0 & 0 & -\frac{1}{2}xy & -\frac{1}{2}y^2+\lambda_2 & \rho\\ \hline \frac{1}{2}x^2-\lambda_1 & \frac{1}{2}xy & 0 & \frac{1}{2}y\pi-\frac{1}{2}x\rho & -V_{,x}\\ \frac{1}{2}xy & \frac{1}{2}y^2-\lambda_2 & \frac{1}{2}x\rho-\frac{1}{2}y\pi & 0 & -V_{,y}\\ \hline \multicolumn{4}{c|}{*} & 0 \end{array}\right],$$

where $V(x,y,c)=V(x,y)+\frac{1}{2}c(x^2+y^2)$ and $V(x,y)=\frac{1}{4}(x^2+y^2)^2-\frac{1}{2}(\lambda_1x^2+\lambda_2y^2)$. It can be easily shown that the above potential $V(x,y)$ is separable, since it satisfies the appropriate criterion of separability (see, for instance, Refs. 10 and 11). In the present paper we will show that it is not a coincidence: there is a close relationship between Poisson operators of the type (2.2) and separable systems.

Example 2.2: The stationary Harry Dym equation

$$\tfrac{1}{4}(u^{-1/2})_{,xxx}-\kappa(u^{-1/2})_{,x}=0$$

for $u=u(x)$, $\kappa\in\mathbf{R}$, after integrating it once, takes the following Newton form:

$$0=r_{,xx}-\kappa r+wr^{-5}, \tag{2.5}$$

$$0 = w_{,xx} - 4\kappa w + d, \tag{2.6}$$

in terms of the new variables $u = r^{-4}$, $w = \kappa r^6 - r^5 r_{,xx}$, with an integration constant d. Observe that it is a *driven* system, which means that the equation (2.6) is independent of (2.5), while the x evolution of r is driven by the solution of (2.6). The above system is not Lagrangian, but it can be shown that it has a nonstandard Hamiltonian formulation, which is important if one wants to consider complete integrability of the equations (2.5)–(2.6):

$$\frac{d}{dx}\begin{bmatrix} r \\ w \\ s \\ z \\ d \end{bmatrix} = \Pi_1 \nabla d, \quad \text{with } \Pi_1 = \left[\begin{array}{cccc|c} 0 & 0 & -\frac{1}{2}r^2 & 0 & \mathscr{E}_{,s} \\ 0 & 0 & 0 & -\frac{1}{2}r^2 & \mathscr{E}_{,z} \\ \frac{1}{2}r^2 & 0 & 0 & rz & \mathscr{E}_{,r} \\ 0 & \frac{1}{2}r^2 & -rz & 0 & \mathscr{E}_{,w} \\ \hline -\mathscr{E}_{,s} & -\mathscr{E}_{,z} & -\mathscr{E}_{,r} & -\mathscr{E}_{,w} & 0 \end{array}\right],$$

where

$$s = r w_{,x} - 2 w r_{,x}, \quad z = r r_{,x} \tag{2.7}$$

are the momenta conjugate to r, w through the generating function

$$E(r,w,d) = r r_{,x} w_{,x} - w r_{,x}^2 - \kappa w r^2 + \tfrac{1}{2} d r^2 + \frac{w^2}{2r^4}. \tag{2.8}$$

The function $\mathscr{E}$ is the Legendre transform of E,

$$\mathscr{E} = s r_{,x}(r,w,s,z) + z w_{,x}(r,w,s,z) - E = \frac{zs}{r} + \frac{w z^2}{r^2} + \kappa w r^2 - \tfrac{1}{2} d r^2 - \frac{w^2}{2r^4}, \tag{2.9}$$

where the functions $r_{,x}(r,w,s,z)$, $w_{,x}(r,w,s,z)$ are obtained by inverting the formulas in (2.7). The origin and meaning of the function E will be explained later.

Let us thus require the operator (2.2) to be a Poisson operator. Motivated by the above examples we can see that it makes sense to investigate (at least) two possible choices of the last column $\mathscr{A}$ (the superscript ^{T} denotes transposition of matrices):

Case A: $\mathscr{A} = \Pi_0 \nabla H \equiv (H_{,\pi}, H_{,\rho}, H_{,x}, H_{,y}, 0)^T$, $H = \frac{1}{2}(\pi^2 + \rho^2) + V(x,y,c)$ (cf. Example 2.1).

Case B: $\mathscr{A} = \Theta_0 \nabla \mathscr{E} \equiv (\mathscr{E}_{,s}, \mathscr{E}_{,z}, \mathscr{E}_{,r}, \mathscr{E}_{,w}, 0)^T$, with $\mathscr{E} = 1/\Delta(-az^2 + bsz - cs^2) - k(r,w,d)$, where Θ_0 is a symmetric operator of the form

$$\Theta_0 = \left[\begin{array}{cc|c} 0 & I & 0 \\ I & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right], \tag{2.10}$$

$\Delta = b^2 - 4ac$, and where the coefficients a, b, c are given as

$$a = Aw^2 + Bw + \alpha_1,$$

$$b = -2Arw - Br - Cw + \alpha_2,$$

$$c = Ar^2 + Cr + \alpha_3$$

(A, B, C, α_i, $i = 1,2,3$ are constants). This choice of $\mathscr{A}$ may seem peculiar, but the system in Example 2.2 is a particular realization of the above system. Moreover, this assumption, being sufficiently general, leads to new interesting types of Newton equations—the already mentioned quasi-Lagrangian systems.

It is also possible to consider multi-dimensional variants of the above cases. An n-dimensional generalization of case A is investigated in Sec. III.

III. CASE A—TOWARDS SEPARABLE SYSTEMS

Let us thus consider an antisymmetric operator of the form

$$\left[\begin{array}{cccc|c} 0 & 0 & \mu+f_{11} & f_{12} & \pi \\ 0 & 0 & f_{21} & \mu+f_{22} & \rho \\ -\mu-f_{11} & -f_{21} & 0 & F & -V_{,x} \\ -f_{12} & -\mu-f_{22} & -F & 0 & -V_{,y} \\ \hline -\pi & -\rho & V_{,x} & V_{,y} & 0 \end{array}\right] \tag{3.1}$$

acting in the phase space of variables x,y,π,ρ,c, with $f_{ij}=f_{ij}(x,y)$, $i,j=1,2$, $V=V(x,y,c)$, $F=F(x,y,\pi,\rho)$, and with some real parameter μ. Let us require the above operator to be a Poisson operator. The Jacobi identity applied to all triples of x,y,π,ρ yields then

$$\begin{bmatrix} f_{11}+\mu & f_{12} \\ f_{21} & f_{22}+\mu \end{bmatrix}\begin{bmatrix} F_{,\pi} \\ F_{,\rho} \end{bmatrix}=\begin{bmatrix} \{f_{11},\rho\}-\{f_{12},\pi\} \\ \{f_{21},\rho\}-\{f_{22},\pi\} \end{bmatrix}, \tag{3.2}$$

where $\{\cdot,\cdot\}$ denotes the Poisson bracket associated with (3.1). If the determinant of the 2×2 matrix on the left-hand side of (3.2) is different from zero, the above equation shows that F is linear in momenta, that is

$$F=g_1(x,y)\pi+g_2(x,y)\rho+\Lambda(x,y), \tag{3.3}$$

and that the functions $g_1=F_{,\pi}$ and $g_2=F_{,\rho}$ are determined by the functions f_{ij}. Further, $\{\{x,y\},c\}+\text{cycl}=0$ implies the symmetry of f_{ij}: $f_{12}=f_{21}$. The conditions $\{\{x,\pi\},c\}+\text{cycl}=0$ and $\{\{y,\rho\},c\}+\text{cycl}=0$ yield

$$f_{11,y}=f_{22,x}=0 \tag{3.4}$$

and

$$f_{11,x}+V_{,xc}=0, \quad f_{22,y}+V_{,yc}=0. \tag{3.5}$$

Using (3.3), we get from $\{\{x,\rho\},c\}+\text{cycl}=0$ and from $\{\{y,\pi\},c\}+\text{cycl}=0$

$$f_{21,x}+g_1=0, \quad f_{12,y}-g_2=0 \tag{3.6}$$

and

$$f_{12,x}+V_{,yc}-g_1=0, \quad f_{21,y}+V_{,c}+g_2=0. \tag{3.7}$$

Moreover, we can conclude that $\Lambda(x,y)\equiv0$. The last Jacobi identity $\{\{\pi,\rho\},c\}+\text{cycl}=0$ yields

$$g_{1,x}=g_{1,y}+g_{2,x}=g_{2,y}=0, \tag{3.8}$$

and the following second-order partial differential equation for V:

$$f_{12}V_{,xx}-f_{21}V_{,yy}+(f_{22}-f_{11})V_{,xy}+V_{,xc}V_{,y}-V_{,yc}V_{,x}-g_1V_{,x}-g_2V_{,y}=0. \tag{3.9}$$

Notice that μ does not appear explicitly in any of the above equations.

Let us now solve the equations (3.4)–(3.9). First, (3.8) implies that $g_1=g_1(y)$, $g_2=g_2(x)$ and thus $\partial g_1(y)/\partial y+\partial g_2(x)/\partial x=0$ integrates to

$$g_1=-Ay-D, \quad g_2=Ax+B$$

for some constants A,B,D. This, combined with (3.6) and with $f_{12}=f_{21}$ gives in turn

$$f_{12}=f_{21}=Axy+Dx+By-C$$

with a new constant C. The equations (3.7) now determine the dependence of V on c

$$V_{,c}=-A(x^2+y^2)-2(Bx+Dy)-K, \tag{3.10}$$

where K is also a constant. This makes it possible to solve (3.5) and we obtain

$$f_{11}=Ax^2+2Bx+\alpha, \quad f_{22}=Ay^2+2Dy+\beta$$

(α,β=const). By inserting the above solutions into (3.2) we obtain an identity which shows that the equations (3.2), though used only partly, do not impose any restrictions on these solutions.

At this stage the equation (3.9) reads

$$\begin{aligned}&(V_{,yy}-V_{,xx})(Axy+By+Dx-C)-V_{,xy}(Ay^2-Ax^2+2Dy-2Bx+\beta-\alpha)\\&\quad+3(Ax+B)V_{,y}-3(Ay+D)V_{,x}=0.\end{aligned} \tag{3.11}$$

The equation (3.10) gives

$$V(x,y,c)=W(x,y)-c(A(x^2+y^2)+2(Bx+Dy)+K). \tag{3.12}$$

After inserting (3.12) in (3.11), we easily check that the function $W(x,y)$ also satisfies the equation (3.11) with the same coefficients. The remarkable fact is that this is the well-known criterion of separability of a natural Hamiltonian system (Bertrand–Darboux equation[10,11]) generated in the canonical way by the Hamiltonian $H=(\pi^2+\rho^2)/2+W(x,y)$. Thus, we have proved the following theorem.

Theorem 3.1: *Let the operator of the form*

$$\Pi_1=\left[\begin{array}{cccc|c} 0 & 0 & f_{11} & f_{12} & \pi \\ 0 & 0 & f_{21} & f_{22} & \rho \\ -f_{11} & -f_{21} & 0 & F & -V_{,x} \\ -f_{12} & -f_{22} & -F & 0 & -V_{,y} \\ \hline -\pi & -\rho & V_{,x} & V_{,y} & 0 \end{array}\right],$$

acting in the five-dimensional phase space with variables x,y,π,ρ,c, with $f_{ij}=f_{ij}(x,y)$, $i,j=1,2$, $\det(f_{ij})\neq 0$, $V=V(x,y,c)$ and with $F=F(x,y,\pi,\rho)$, be a Poisson operator. Then

(1) Π_1 *must have the form*

$$\left[\begin{array}{cc|cc|c} 0 & 0 & Ax^2+2Bx+\alpha & Axy+By+Dx-C & \pi \\ 0 & 0 & Axy+By+Dx-C & Ay^2+2Dy+\beta & \rho \\ \hline & & 0 & (Ax+B)\rho-(Ay+D)\pi & -V_{,x} \\ \multicolumn{2}{c|}{*} & (Ay+D)\pi-(Ax+B)\rho & 0 & -V_{,y} \\ \hline -\pi & -\rho & V_{,x} & V_{,y} & 0 \end{array}\right]$$

with $A,B,C,D,\alpha,\beta=$const;

(2) *V must have the form*

$$V(x,y,c)=W(x,y)-c(A(x^2+y^2)+2(Bx+Dy)+h) \tag{3.13}$$

($h=$const) *where the function $W=W(x,y)$ satisfies the separability criterion for two-dimensional natural Hamiltonian systems (Bertrand–Darboux equation[10])*

$$\begin{aligned}&(W_{,yy}-W_{,xx})(Axy+By+Dx-C)-W_{,xy}(Ay^2-Ax^2+2Dy-2Bx+\beta-\alpha)\\&\quad+3(Ax+B)W_{,y}-3(Ay+D)W_{,x}=0;\end{aligned} \tag{3.14}$$

and

(3) Π_1 *is compatible with* Π_0.

The last part of the above theorem is true since its proof is valid for any value of μ.
Conclusion (2) of the above theorem yields the following corollary.
Corollary 3.2: The natural two-dimensional Hamiltonian system

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ \pi \\ \rho \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \nabla\left(\tfrac{1}{2}\pi^2 + \tfrac{1}{2}\rho^2 + W(x,y)\right),$$

with W(x,y) satisfying the criterion of separability (3.14), has in the extended (five-dimensional) phase space of variables x,y,π,ρ,c a bi-Hamiltonian formulation of the form

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ \pi \\ \rho \\ c \end{bmatrix} = \Pi_0 \nabla\left(\tfrac{1}{2}\pi^2 + \tfrac{1}{2}\rho^2 + V(x,y,c)\right) = \Pi_1 \nabla c \tag{3.15}$$

with V determined by (3.13), and with Π_1 *as in Conclusion (1) of Theorem 3.1.* Since $\Pi_0 \nabla c = 0$ the natural question arises: is it possible to extend the system (3.15) to a bi-Hamiltonian chain? In other words, we have to check if there exists a function $K = K(x,y,\pi,\rho)$ on our phase space such that $\Pi_0 \nabla K = \Pi_1 \nabla H$ with $H = (\frac{1}{2}\pi^2 + \frac{1}{2}\rho^2 + V(x,y,c))$ and such that K is a Casimir function for Π_1. A direct integration of $\Pi_0 \nabla K = \Pi_1 \nabla H$ leads to

$$K = \tfrac{1}{2}\pi^2 f_{22} + \tfrac{1}{2}\rho^2 f_{11} - \rho\pi f_{12} + \Psi(x,y,c), \tag{3.16}$$

where Ψ has to satisfy the equations

$$\begin{bmatrix} \Psi_{,x} \\ \Psi_{,y} \end{bmatrix} = \begin{bmatrix} f_{22} - (\alpha + \beta - h) & -f_{12} \\ -f_{21} & f_{11} - (\alpha + \beta - h) \end{bmatrix} \begin{bmatrix} V_{,x} \\ V_{,y} \end{bmatrix} \tag{3.17}$$

with f_{ij}, α, β as in Theorem 3.1. The demand $\Pi_1 \nabla K = 0$ leads further to the relation $h = \alpha + \beta$ (thus showing that the sought chain exists only for a special choice of integration constants) and, moreover, to $\Psi_{,c} = f_{12} f_{21} - f_{11} f_{22} = -\det(f)$. Combining all this with (3.17) we finally get

$$K = \tfrac{1}{2}\pi^2 f_{22} + \tfrac{1}{2}\rho^2 f_{11} - \rho\pi f_{12} - c \det(f) + \overline{\Psi}(x,y), \tag{3.18}$$

where the function $\overline{\Psi}$ can be established from the relation

$$\begin{bmatrix} \overline{\Psi}_{,x} \\ \overline{\Psi}_{,y} \end{bmatrix} = \begin{bmatrix} f_{22} & -f_{12} \\ -f_{21} & f_{11} \end{bmatrix} \begin{bmatrix} W_{,x} \\ W_{,y} \end{bmatrix}, \tag{3.19}$$

which is solvable since the compatibility condition $\overline{\Psi}_{xy} = \overline{\Psi}_{yx}$ for (3.19) has exactly the form of the separability criterion (3.14), which is satisfied by the function W. Thus the bi-Hamiltonian vector field (3.15) extends to the following bi-Hamiltonian chain:

$$\Pi_0 \nabla c = 0,$$

$$\Pi_0 \nabla H = \Pi_1 \nabla c,$$

$$\Pi_0 \nabla K = \Pi_1 \nabla H,$$

$$\Pi_1 \nabla K = 0,$$

with K given by (3.18) and by (3.19).

IV. GENERALIZATION TO *N* DIMENSIONS

The results of Sec. III can be generalized from two to n dimensions so that a simple assumption about the form of a $(2n+1)$-dimensional Poisson operator completely determines its structure. Moreover, it appears again to be the structure of a Poisson pencil for separable systems.[12]

Consider a $2n+1$-dimensional phase space $\mathscr{M}$ endowed with coordinates (q,p,c) where $q=(q_1,\ldots,q_n)$, $p=(p_1,\ldots,p_n)$, $c\in\mathbf{R}$. For the purposes of this section we temporarily modify our notation: for a function $F\colon \mathscr{M}\to\mathbf{R}$ we denote its derivative with respect to q_j by $F_{,j}$ so that $F_{,j}\equiv F_{,q_j}=\partial F/\partial q_j$. Similarly, $F_{,c}:=\partial F/\partial c$. The derivative of F with respect to p_j will be denoted by the standard expression $\partial F/\partial p_j$.

Consider also an antisymmetric operator of the form

$$\left[\begin{array}{c|c|c} 0 & f(q)+\mu I_n & p^T \\ \hline * & B(q,p) & -V_{,q}^T \\ \hline -p & V_{,q} & 0 \end{array}\right], \tag{4.1}$$

where f,B are $n\times n$ matrices with entries $f_{ij}(q)$ and $B_{ij}(q,p)$, respectively, $i,j=1,\ldots,n$, $B_{ij}=-B_{ji}$, I_n is $n\times n$ identity matrix, $\mu\in\mathbf{R}$, $V=V(q,c)$, $V_q=(V_{,1},\ldots,V_{,n})$, and where the superscript ^{T} denotes the transpose of matrix. As in Sec. III let us require the above operator to be a Poisson operator. Let us denote also the Poisson bracket induced by (4.1) again by $\{\cdot,\cdot\}$. The equations $\{\{q_i,q_j\},q_k\}+\text{cycl}=0$ and $\{\{q_i,q_j\},p_k\}+\text{cycl}=0$ are satisfied identically, while the equations $\{\{q_i,q_j\},c\}+\text{cycl}=0$ yield symmetry of f: $f_{ij}=f_{ji}$. The equations $\{\{q_i,p_j\},p_k\}+\text{cycl}=0$ give

$$\sum_{s=1}^{n}(f_{is}+\mu\delta_{is})\frac{\partial B_{jk}}{\partial p_s}=\{f_{ij},p_k\}-\{f_{ik},p_j\},\quad i,j,k=1,\ldots,n, \tag{4.2}$$

with $\delta_{ii}=1$, $\delta_{ij}=0$, $i\neq j$. If $\det(f+\mu\delta)\neq 0$, then (4.2) implies that B_{jk} are linear in momenta, that is

$$B_{jk}=\sum_{s=1}^{n}g_{jk}^s p_s+\Lambda_{jk} \tag{4.3}$$

with some functions $g_{jk}^s=g_{jk}^s(q)$, $\Lambda_{jk}=\Lambda_{jk}(q)$. Due to the skew-symmetry of B_{jk}, we have also $g_{jk}^s=-g_{kj}^s$, $\Lambda_{jk}=-\Lambda_{kj}$.

The next set of Jacobi identities $\{\{q_i,p_j\},c\}+\text{cycl}=0$ gives $\Lambda_{ij}\equiv 0$ and

$$f_{ij,s}-g_{ij}^s+V_{,jc}\delta_{si} \tag{4.4}$$

with $V_{,jc}=\partial^2 V/\partial q_i\partial c$. We will treat the equation $\{\{p_i,p_j\},p_k\}+\text{cycl}=0$ later. The last set $\{\{p_i,p_j\},c\}+\text{cycl}=0$ leads to

$$\sum_{r,s=1}^{n}g_{ij,s}^r p_s p_r-\sum_{r=1}^{n}g_{ij}^r V_{,r}+\sum_{r=1}^{n}V_{,ir}f_{rj}+V_{,ic}V_{,j}-\sum_{r=1}^{n}V_{,jr}f_{ri}-V_{,jc}V_{,i}=0. \tag{4.5}$$

By taking the coefficients at linearly independent powers of momenta we get the following conditions:

$$g_{ij,s}^s=0, \tag{4.6}$$

$$g_{ij,s}^r+g_{ij,r}^s=0, \tag{4.7}$$

and

$$\sum_{r=1}^{n}g_{ij}^r V_{,r}-\sum_{r=1}^{n}V_{,ir}f_{rj}-V_{,ic}V_{,j}+\sum_{r=1}^{n}V_{,jr}f_{ri}+V_{,jc}V_{,i}=0. \tag{4.8}$$

The system of equations (4.4)–(4.8) is analogous to the system (3.4)–(3.9), but the procedure of its integration is more complicated. We outline it shortly below.

First, assume $i\neq j$. For $s\neq i$ (4.4) reads $f_{ij,s}-g^s_{ij}=0$. If, moreover, $s\neq j$, we also have $f_{ji,s}-g^s_{ji}=0$. Due to the symmetry of f, $f_{ij}=f_{ji}$, we obtain $g^s_{ij}=g^s_{ji}$, and because g^s_{ij} are antisymmetric in the lower indices we get $g^s_{ij}=0$ for $s\neq i,j$. The equation (4.7) for $r=i$ reads thus $g^i_{ij,s}=0$, $s\neq i,j$, and, by virtue of (4.6), we see that g^i_{ij} is a function of q_j only: $g^i_{ij}=g^i_{ij}(q_j)$. Similarly, $g^j_{ij}=g^j_{ij}(q_i)$. The equation (4.7) for $r=i$, $s=j$ then yields $g^i_{ij,j}=-g^j_{ij,i}=-A_{ij}$ with some constants A_{ij}. By separating variables in this system we integrate it to

$$g^j_{ij}=A_{ij}q_i+b_{ij},\quad g^i_{ij}=-A_{ij}q_j-d_{ij} \tag{4.9}$$

with the integration constants b_{ij} and d_{ij}, and therefore

$$B_{ij}=g^i_{ij}p_i+g^i_{ij}p_j=-(A_{ij}q_j+d_{ij})p_i+(A_{ij}q_i+b_{ij})p_j.$$

Due to the skew-symmetry $g^j_{ij}=-g^j_{ji}$, the equation (4.9) implies $A_{ij}=A_{ji}$ and $b_{ij}=d_{ji}$.

Finally we can calculate f_{ij} (still for $i\neq j$). By (4.4) $f_{ij,s}=g^s_{ij}=0$ and therefore $f_{ij}=f_{ij}(q_i,q_j)$. By using (4.4) and the symmetry of f, we conclude that

$$f_{ij}=A_{ij}q_iq_j+b_{ij}q_j+d_{ij}q_i+c_{ij}$$

with some integration constants $c_{ij}=c_{ji}$.

Let us now turn to the case $i=j$. The equation (4.4) implies that $f_{ii}=f_{ii}(q_i)$ and that $f_{ii,i}+V_{,ic}=0$. By using (4.4) again one can show that $f_{ii}=A_{ij}q_i^2+2b_{ij}q_i+\alpha_i$, which proves that $A_{ij}=A_i$ and that $b_{ij}=b_i$. However, $A_{ij}=A_{ji}$ and so $A_{ij}=A$ for all $i,j=1,...,n$. Moreover, $d_{ij}=b_{ji}$ and so $d_{ij}=d_j=b_j$. To summarize our calculations let us collect the above results:

$$g^s_{ij}=0,\quad s\neq i,j, \tag{4.10}$$

$$g^j_{ij}=Aq_i+b_i,\quad g^i_{ij}=-Aq_j-b_j, \tag{4.11}$$

$$B_{ij}=-(Aq_j+b_j)p_i+(Aq_i+b_i)p_j, \tag{4.12}$$

$$f_{ij}=Aq_iq_j+b_iq_j+b_jq_i+c_{ij}, \tag{4.13}$$

with the constants $A,b_i,c_{ij}=c_{ji}$, $i,j=1,...,n$. Observe also that (4.13) is valid even for $i=j$ with $\alpha_i=c_{ii}$.

Due to (4.13) the equation (4.4) yields now $V_{,ci}=V_{,ic}=-f_{ii,i}=-2(Aq_i+b_i)$ and so

$$V(q,c)=W(q)-c\left(\sum_{i=1}^{n}(Aq_i^2+2b_iq_i)+h\right) \tag{4.14}$$

with $h=\text{const}$.

Up to now we did not use the equations $\{\{p_i,p_j\},p_k\}+\text{cycl}=0$. One can, however, show that by virtue of (4.12) it becomes an identity. As in the case $n=2$ considered in Sec. III one can also check, again by using (4.12) and (4.13), that (4.2) are also satisfied identically and so they do not impose any other restrictions on our solutions (4.10)–(4.13).

Let us now turn to the equation (4.8). Due to (4.10)–(4.13) it reads as

$$\begin{aligned}&(V_{,jj}-V_{,ii})(Aq_iq_j+b_iq_j+b_jq_i+c_{ij})\\&\quad+\sum_{r\neq i,j}(V_{,jr}(Aq_rq_i+b_rq_i+b_iq_r+c_{ri})-V_{,ir}(Aq_rq_j+b_rq_j+b_jq_r+c_{rj}))\\&\quad-V_{,ij}(Aq_j^2-Aq_i^2+2b_jq_j-2b_iq_i+c_{jj}-c_{ii})+3(Aq_i+b_i)V_{,j}-3(Aq_j+b_j)V_{,i}=0.\end{aligned} \tag{4.15}$$

Inserting (4.14) into (4.15) we easily check that the function $W(q)$ satisfies exactly the same equation (4.15) with the same coefficients. This equation has been found previously[13] and serves

as a criterion of separability (in generalized elliptic coordinates or any of its degenerations) of natural Hamiltonian systems generated in the canonical way by $H=\frac{1}{2}\Sigma_{i=1}^{n}p_i^2+W(q)$.

The results of the above derivation shall be formulated as the following theorem.

Theorem 4.1: *Let the antisymmetric operator of the form*

$$\Pi_1=\left[\begin{array}{c|c|c} 0 & f(q) & p^T \\ \hline -f^T(q) & B(q,p) & -V_{,q}^T \\ \hline -p & V_{,q} & 0 \end{array}\right]$$

acting in the $2n+1$*-dimensional phase space with variables* (q,p,c)*, where* $q=(q_1,\ldots,q_n)$*,* $p=(p_1,\ldots,p_n)$*,* $c\in\mathbf{R}$*, with* $n\times n$ *matrices* f *and* B *with entries* $f_{ij}(q)$ *and* $B_{ij}(q,p)$ *respectively, with* $B_{ij}=-B_{ji}$*,* $\det(f)\neq 0$ *and with* $V=V(q,c)$*,* $V_{,q}=(V_{,1},\ldots,V_{,n})$*,* $V_{,i}=\partial V/\partial q_i$*, be a Poisson operator. Then*

(1) Π_1 *must have the form*

$$\left[\begin{array}{c|c|c} 0 & Aq_iq_j+b_iq_j+b_jq_i+c_{ij} & p^T \\ \hline * & (Aq_i+b_i)p_j-(Aq_j+b_j)p_i & -V_{,q}^T \\ \hline -p & V_{,q} & 0 \end{array}\right], \tag{4.16}$$

with constants $A,b_i,c_{ij}=c_{ji}$*,* $i,j=1,\ldots,n$;

(2) $V(q,c)$ *must have the form*

$$V(q,c)=W(q)-c\left(\sum_{i=1}^{n}(Aq_i^2+2b_iq_i)+h\right) \tag{4.17}$$

with $h=$*const where the function* $\mathrm{W}(q)$ *satisfies the separability criterion (4.15) for n-dimensional natural Hamiltonian systems*; and

(3) Π_1 *is compatible with the standard Poisson operator*

$$\left[\begin{array}{cc|c} 0 & I_n & 0 \\ -I_n & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$$

in the extended phase space of variables (q,p,c).

As in Sec. III, this theorem also yields the following corollary:

Corollary 4.2: The natural n-dimensional Hamiltonian system

$$\frac{d}{dt}\begin{bmatrix} q^T \\ p^T \end{bmatrix}=\begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}\nabla\left(\frac{1}{2}\sum_{i=1}^{n}p_i^2+W(q)\right)$$

with $q=(q_1,\ldots,q_n)$*,* $p=(p_1,\ldots,p_n)$ *and with W(q) satisfying the criterion of separability (4.15) has in the extended* ($2n+1$-dimensional) phase space of variables (q,p,c), $\mathrm{c}\in\mathbf{R}$ *a bi-Hamiltonian formulation of the form*

$$\frac{d}{dt}\begin{bmatrix} q^T \\ p^T \\ c \end{bmatrix}=\left[\begin{array}{cc|c} 0 & I_n & 0 \\ -I_n & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right]\nabla\left(\frac{1}{2}\sum_{i=1}^{n}p_i^2+V(q,c)\right)=\Pi_1\nabla c$$

with V(q,c) determined by (4.17) and with Π_1 *as in (4.16).*

It has to be mentioned that the equations (4.15) can also be satisfied by decomposable potentials, that is, potentials of the type $V(\overline{q},\hat{q})=V_1(\overline{q})+V_2(\hat{q})$, where $\overline{q}$ and $\hat{q}$ denote coordinates on subspaces $\mathbf{R}_r^n$ and $\mathbf{R}_s^n$ of $\mathbf{R}^n$, respectively (with $\mathbf{R}^n=\mathbf{R}_r^n\oplus\mathbf{R}_s^n$ being their direct sum), whereas neither of V_i has to be separable on the appropriate subspace. Moreover, the above theorem does not encompass all separable potentials, only those which are indecomposable, for if $V=V(\overline{q},\hat{q})=V_1(\overline{q})+V_2(\hat{q})$ with $\overline{q}=(q_1,\ldots,q_r)$, $\hat{q}=(q_{r+1},\ldots,q_n)$ and if V_1 is separable on the subspace

spanned by the coordinates $\overline{q}$ and V_2 is separable on the subspace spanned by $\hat{q}$, then V has a Hamiltonian structure which does not have the form (4.16), as the example below shows.

Example 4.3: Consider a $(2n+2)$-dimensional phase space of variables $(\overline{q},\hat{q},\overline{p},\hat{p},c,d)$, $\overline{q}=(q_1,\ldots,q_r)$, $\hat{q}=(q_{r+1},\ldots,q_n)$, $\overline{p}=(p_1,\ldots,p_r)$, $\hat{p}=(p_{r+1},\ldots,p_n)$, $c,d\in\mathbf{R}$, with the antisymmetric Poisson operator

$$\left[\begin{array}{cccc|c|c} 0 & 0 & \overline{f}(\overline{q}) & 0 & \overline{p}^T & 0 \\ 0 & 0 & 0 & \hat{f}(\hat{q}) & 0 & \hat{p}^T \\ -\overline{f}(\overline{q}) & 0 & \overline{B}(\overline{q},\overline{p}) & 0 & -V_{,\overline{q}}^T & 0 \\ 0 & -\hat{f}(\hat{q}) & 0 & \hat{B}(\hat{q},\hat{p}) & 0 & -V_{,\hat{q}}^T \\ \hline -\overline{p} & 0 & V_{,\overline{q}} & 0 & 0 & 0 \\ \hline 0 & -\hat{p} & 0 & V_{,\hat{q}} & 0 & 0 \end{array}\right], \tag{4.18}$$

where $\overline{f}=\overline{f}(\overline{q})$ and $\overline{B}=\overline{B}(\overline{q},\overline{p})$ are $r\times r$ matrices with entries $\overline{f}_{ij}(\overline{q})$ and $\overline{B}_{ij}(\overline{q},\overline{p})$, respectively; $\overline{B}_{ij}=-\overline{B}_{ji}$, where $\hat{f}=\hat{f}(\hat{q})$ and $\hat{B}=\hat{B}(\hat{q},\hat{p})$ are $(n-r)\times(n-r)$ matrices with entries $\hat{f}_{ij}(\hat{q})$ and $\hat{B}_{ij}(\hat{q},\hat{p})$, respectively; $\hat{B}_{ij}=-\hat{B}_{ji}$; and where $V=V(\overline{q},\hat{q},c,d)=\overline{V}(\overline{q},c)+\hat{V}(\hat{q},d)$, $V_{,\overline{q}}=(V_{,1},\ldots,V_{,r})$, and $V_{,\hat{q}}=(V_{,r+1},\ldots,V_{,n})$. One can check that the Poisson identities for (4.18) split into two separate classes, one involving only bar variables $\overline{q},\overline{p}$ and the other one involving only hat variables $\hat{q},\hat{p}$, each class leading to the equations of exactly the same form as (4.2)–(4.8). It means that $\overline{f}_{ij}$, $\hat{f}_{ij}$ are both of the form (4.13) (with two different sets of constants) and that both $\overline{B}_{ij}$ and $\hat{B}_{ij}$ are given by (4.12) (again, with two different sets of constants). It also means that both $\overline{V}(\overline{q})$ and $\hat{V}(\hat{q})$ satisfy the separability criterion analogous to (4.15). It is, however, impossible to transform this operator to the form (4.16).

One can also consider decomposable potentials which decompose into more than two independent parts. Such potentials correspond (in the sense described above) to Poisson operators which generalize (4.18) by having more blocks and more than two last columns depending ov V.

V. CASE B—TOWARDS QUASI-LAGRANGIAN SYSTEMS

Let us now consider an antisymmetric operator of the form

$$\left[\begin{array}{cccc|c} 0 & 0 & f_{11} & f_{12} & \mathscr{E}_{,s} \\ 0 & 0 & f_{21} & f_{22} & \mathscr{E}_{,z} \\ -f_{11} & -f_{21} & 0 & F & \mathscr{E}_{,r} \\ -f_{12} & -f_{22} & -F & 0 & \mathscr{E}_{,w} \\ \hline -\mathscr{E}_{,s} & -\mathscr{E}_{,z} & -\mathscr{E}_{,r} & -\mathscr{E}_{,w} & 0 \end{array}\right] \tag{5.1}$$

in the five-dimensional phase space of variables r,w,s,z,d, where $f_{ij}=f_{ij}(r,w)$, $i,j=1,2$, $F=F(r,w,s,z)$, and where

$$\mathscr{E}=\frac{1}{\Delta}(-az^2+bsz-cs^2)-k(r,w,d), \tag{5.2}$$

with $\Delta=b^2-4ac$ and with the polynomials a,b,c given by

$$\begin{aligned} a&=Aw^2+Bw+\alpha_1, \\ b&=-2Arw-Br-Cw+\alpha_2, \\ c&=Ar^2+Cr+\alpha_3 \end{aligned} \tag{5.3}$$

$(A,B,C,\alpha_i=\text{const},\ i=1,2,3)$.

The structures of the above type originate in the framework of the theory of a special type of Newton systems, the so-called quasi-Lagrangian systems. Here we shall present only those fea-

tures of quasi-Lagrangian systems which are relevant for understanding the origins of Poisson operators of type B. A more complete theory of quasi-Lagrangian structures is being developed in a separate paper.[14]

Consider the configuration space of variables r,w and a function on its tangent bundle given by

$$E=ar_{,x}^2+br_{,x}w_{,x}+cw_{,x}^2+k(r,w,d) \tag{5.4}$$

with a,b,c given by (5.3). The *quasi-Lagrangian* equations for E are *defined* as

$$\frac{\partial E}{\partial r}+\frac{d}{dx}\frac{\partial E}{\partial r_{,x}}=0, \quad \frac{\partial E}{\partial w}+\frac{d}{dx}\frac{\partial E}{\partial w_{,x}}=0. \tag{5.5}$$

They differ from the classical Lagrange equations by having a plus sign instead of a minus sign between derivatives. This simple modification leads to equations which do not follow from any variational principle but nonetheless possess many interesting properties which make them worth studying (see Ref. 14). They explicitly read

$$2ar_{,xx}+bw_{,xx}+k_{,r}=0, \quad br_{,xx}+2cw_{,xx}+k_{,w}=0,$$

so they do not have any first-order x derivatives. This is a nontrivial fact, which is a consequence of our special choice of the coefficients a,b,c, in (5.3). If the determinant $\Delta=b^2-4ac$ of the above system is different from zero, then they can be written as a system of Newton equations

$$r_{,xx}=\frac{1}{\Delta}(2ck_{,r}-bk_{,w}), \quad w_{,xx}=\frac{1}{\Delta}(2ak_{,w}-bk_{,r}). \tag{5.6}$$

They usually do not have any Lagrangian formulation, and thus any standard Hamiltonian formulation. However, if we introduce momentum type variables

$$s=\frac{\partial E}{\partial r_{,x}}=2ar_{,x}+bw_{,x}, \quad z=\frac{\partial E}{\partial w_{,x}}=br_{,x}+2cw_{,x}, \tag{5.7}$$

then it can be shown that Eqs. (5.6) in the variables r,w,s,z,d attain the form

$$\frac{d}{dx}\begin{bmatrix} r \\ w \\ s \\ z \\ d \end{bmatrix}=\Theta_0\nabla\mathscr{E}, \tag{5.8}$$

with the symmetric operator Θ_0 given by (2.10) and with $\mathscr{E}$ given exactly by (5.2). The function $\mathscr{E}$ is the Legendre transform of E with the momenta defined by (5.7):

$$\mathscr{E}=sr_{,x}(r,w,s,z)+zw_{,x}(r,w,s,z)-E(r,w,r_{,x}(r,w,s,z),w_{,x}(r,w,s,z),d),$$

where $r_{,x}(r,w,s,z)$ and $w_{,x}(r,w,s,z)$ are obtained by inverting the formulas (5.7), while E itself in terms of variables r,w,s,z,d attains the form

$$E=E(r,w,r_{,x}(r,w,s,z),w_{,x}(r,w,s,z),d)=\frac{1}{\Delta}(-az^2+bsz-cs^2)+k(r,w,d).$$

The function E appears to be an integral of motion for (5.6) and will thus be referred to as *energy* of the system (5.6). The systems of type (5.8) may be called *quasi-gradient* systems, since $\Theta_0\nabla\mathscr{E}=(\mathscr{E}_{,s},\mathscr{E}_{,z},\mathscr{E}_{,r},\mathscr{E}_{,w},0)$ is obtained by permutation of the entries of the four-dimensional gradient $(\mathscr{E}_{,r},\mathscr{E}_{,w},\mathscr{E}_{,s},\mathscr{E}_{,z})$ of $\mathscr{E}$. It is also important to observe that in Example 2.2 the energy E is a Casimir of Π_1, that is, $\Pi_1\nabla E=0$.

Similar to case A, let us require the operator (5.1) to be Poisson. We can now completely determine the structure of this operator if we assume that the energy E, corresponding to $\mathscr{E}$ as described above, is a Casimir function of (5.1) (like in Example 2.2).

Let us start with the Casimir property of E. The requirement $\Pi_1 \nabla E = 0$ immediately leads to two sets of linear homogeneous equations for the functions f_{ij},

$$\begin{aligned} &0=(f_{11}+k_{,d})(-2\gamma)+f_{12}\beta, \quad 0=f_{21}(-2\gamma)+(f_{22}+k_{,d})\beta, \\ &0=(f_{11}+k_{,d})\beta+f_{12}(-2\alpha), \quad 0=f_{21}\beta+(f_{22}+k_{,d})(-2\alpha), \end{aligned} \tag{5.9}$$

where $\alpha=a/\Delta$, $\beta=b/\delta$, $\gamma=c/\Delta$, and, moreover, to

$$F=-2\frac{T_{,r}}{T_{,z}}k_{,d}, \quad F=-2\frac{T_{,w}}{T_{,s}}k_{,d}, \tag{5.10}$$

$$T_{,w}T_{,z}+T_{,s}T_{,r}=0, \tag{5.11}$$

where T is the kinetic energy part of $\mathscr{E}$,

$$T=\frac{1}{\Delta}(-az^2+bsz-cs^2).$$

Since we assume that $\Delta\neq 0$, Eqs. (5.9) have only zero solutions and we immediately get

$$f_{11}=f_{22}=-k_{,d}, \quad f_{12}=f_{21}=0. \tag{5.12}$$

One can also easily see that (5.11) is the condition of compatibility for (5.10) and that it is also satisfied identically due to our choice of a,b,c in (5.3).

Let us now use our main assumption that the operator (5.1) is Poisson. The formula (3.2) (with $\mu=0$ and with new labeling of variables) is still valid and, due to (5.12), it reads

$$\begin{bmatrix} F_{,s} \\ F_{,z} \end{bmatrix} = \frac{1}{k_{,d}} \begin{bmatrix} \{k_{,d},z\} \\ -\{k_{,d},s\} \end{bmatrix}. \tag{5.13}$$

Let us denote $k_{,d}$ by $G=G(r,w)$ [so that $G=-f_{11}=-f_{22}$ by virtue of (5.12)]. Then, due to (5.13), we get

$$F=G_{,r}z-G_{,w}s. \tag{5.14}$$

Furthermore, all Jacobi identities involving the variable d imply the following, heavily overdetermined set of equations,

$$\begin{gathered} 0=G_{,w}\beta+2G\gamma_{,r}, \\ 0=2\alpha G_{,w}+\beta G_{,r}+2\beta_{,r}G, \\ 0=\alpha G_{,r}+\alpha_{,r}G, \\ 0=2\beta_{,w}G+2\gamma G_{,r}+\beta G_{,w}, \\ 0=\gamma_{,w}G+\gamma G_{,w}, \\ 0=2\alpha_{,w}G+\beta G_{,r}, \\ 0=2\gamma G_{,wr}-\beta G_{,ww}-2\gamma_{,r}G_{,w}+\beta_{,w}G_{,w}, \\ 0=\beta G_{,rr}-2\alpha G_{,rw}-G_{,r}\beta_{,r}+2\alpha_{,w}G_{,r}, \\ 0=2\alpha G_{,ww}-2\gamma G_{,rr}+\beta_{,r}G_{,w}+2\gamma_{,r}G_{,r}-2\alpha_{,w}G_{,w}-\beta_{,w}G_{,r}, \end{gathered} \tag{5.15}$$

which is certainly solvable since we have already seen one of its solutions in Example 2.2. Let us solve this system. Observe that the third and the fifth of the above equations are total derivatives

$$(\alpha G)_{,r}=0, \quad (\gamma G)_{,w}=0.$$

So, for $\alpha\neq 0$ and $\gamma\neq 0$, $G(r,w)=\xi(w)/\alpha=\eta(r)/\gamma$ for some functions ξ and η depending only on w and r, respectively. Since $\alpha=a/\Delta$ and $\gamma=c/\Delta$, it means that $\xi(w)=\chi a(w)$ and $\eta(r)=\chi c(r)$ for some constant χ. This yields immediately

$$G(r,w)=\chi\Delta(r,w). \tag{5.16}$$

For $\alpha=0$ or for $\gamma=0$ the above solution remains valid. One can check, through a lengthy calculation, that G given by (5.16) satisfies also all the remaining equations (5.15). Thus, it is in fact their general solution. We conclude that

$$k(r,w,d)=\chi\Delta d+\mathscr{F}(r,w) \tag{5.17}$$

with some arbitrary function $\mathscr{F}$.

One can also check that comparison of the first of the equations in (5.10) with (5.14),

$$F=-2\frac{T_{,r}}{T_{,z}}\,G=G_{,r}z-G_{,w}s,$$

leads exactly to the first three equations of (5.15), while comparison of the second equation in (5.10) with (5.14),

$$F=-2\frac{T_{,w}}{T_{,s}}\,G=G_{,r}z-G_{,w}s,$$

leads to the next three equations of (5.15). This shows that the conditions (5.10) are satisfied automatically whenever k is given by (5.17). Thus we have proved the following theorem.

Theorem 5.1: *Let Π_1 be an antisymmetric operator of the form*

$$\Pi_1=\left[\begin{array}{cccc|c} 0 & 0 & f_{11} & f_{12} & \mathscr{E}_{,s} \\ 0 & 0 & f_{21} & f_{22} & \mathscr{E}_{,z} \\ -f_{11} & -f_{21} & 0 & F & \mathscr{E}_{,r} \\ -f_{12} & -f_{22} & -F & 0 & \mathscr{E}_{,w} \\ \hline -\mathscr{E}_{,s} & -\mathscr{E}_{,z} & -\mathscr{E}_{,r} & -\mathscr{E}_{,w} & 0 \end{array}\right]$$

acting in the five-dimensional phase space of variables r,w,s,z,d, and let

$$\mathscr{E}=\frac{1}{\Delta}\,(-az^2+bsz-cs^2)-k(r,w,d)$$

with $\Delta=b^2-4ac$ and with the polynomials a,b,c given by

$$a=Aw^2+Bw+\alpha_1,$$

$$b=-2Arw-Br-Cw+\alpha_2,$$

$$c=Ar^2+Cr+\alpha_3$$

with some constants A,B,C,α_i, $i=1,2,3$. Further, let $f_{ij}=f_{ij}(r,w)$, $i,j=1,2$ and let $F=F(r,w,s,z)$. If Π_1 is a Poisson operator, and if the energy E associated with $\mathscr{E}$ through $E=(1/\Delta)(-az^2+bsz-cs^2)+k(r,w,d)$ is a Casimir function of Π_1 (i.e., $\Pi_1\nabla E=0$), then we have the following:

(1) *The function k must be of the form* $k(r,w,d)=\chi\Delta d+\mathscr{F}(r,w)$ *with some constant* χ *and with* $\mathscr{F}$ *arbitrary function of r and w.*

(2) *The operator* Π_1 *must have the form*

$$\Pi_1=\left[\begin{array}{cccc|c} 0 & 0 & -\chi\Delta & 0 & \mathscr{E}_{,s} \\ 0 & 0 & 0 & -\chi\Delta & \mathscr{E}_{,z} \\ \chi\Delta & 0 & 0 & F & \mathscr{E}_{,r} \\ 0 & \chi\Delta & -F & 0 & \mathscr{E}_{,w} \\ \hline -\mathscr{E}_{,s} & -\mathscr{E}_{,z} & -\mathscr{E}_{,r} & -\mathscr{E}_{,w} & 0 \end{array}\right]$$

with $F=\chi(\Delta_{,r}z-\Delta_{,w}s)$.

(3) *The system (5.8) also has a Hamiltonian formulation, so that*

$$\frac{d}{dx}\begin{bmatrix} r \\ w \\ s \\ z \\ d \end{bmatrix}=\Theta_0\nabla\mathscr{E}=\Pi_1\nabla d.$$

VI. CONCLUSIONS

In this paper we have shown how relatively general assumptions about a Poisson operator lead either to the class of separable potential Newton equations or to a new class of systems—quasi-Lagrangian systems. In this way we have found a bi-Hamiltonian characterization for all separable potentials in two dimensions and for indecomposable potentials in *n* dimensions. In order to characterize bi-Hamiltonian formulations for decomposable potentials one has to describe all possible ways of splitting $\mathbf{R}^n$ into its subspaces as in the classification scheme of Kalnins and Miller.[8] Since it is technically complex we have chosen to illustrate it with a simple example.

The quasi-Lagrangian family of Newton equations and their Hamiltonian (but not bi-Hamiltonian) formulation is completely new, discovered only very recently, and is currently under investigation. The results will be published in a separate paper,[14] where the question of their complete integrability will be thoroughly discussed.

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, 1989).
[2] M. Adler and P. van Moerbecke, Adv. Math. **38**, 267 (1980).
[3] A. Reiman and M. A. Semenov Tian-Shansky, Invent. Math. **54**, 81 (1979).
[4] M. A. Semenov Tian-Shansky, Funct. Anal. Appl. **17**, 259 (1983).
[5] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
[6] M. Antonowicz and S. Rauch-Wojciechowski, Phys. Lett. A **147**, 455 (1990).
[7] S. Rauch-Wojciechowski, Phys. Lett. A **160**, 149 (1991).
[8] E. G. Kalnins and W. Miller, Jr., J. Math. Phys. **27**, 1721 (1986).
[9] R. Garnier, Circolo Mat. Palermo **43**, 155 (1919).
[10] E. T. Whittaker, *A Treatise on the Analytic Dynamics of Particles and Rigid Bodies* (Cambridge U.P., Cambridge, 1959), Chap. 12, Sec. 152.
[11] I. Marshall and S. Wojciechowski, J. Math. Phys. **29**, 1338 (1988).
[12] S. Rauch-Wojciechowski, Phys. Lett. A **160**, 149 (1991).
[13] S. Rauch-Wojciechowski, ''A criterion of separability for natural hamiltonian systems,'' in preparation.
[14] S. Rauch-Wojciechowski, K. Marciniak, and H. Lundmark, ''Quasi-Lagrangian systems of Newton equations,'' in preparation.

Stochastic construction of a Feynman path integral representation for Green's functions in radiative transfer

Steven D. Miller[a)]
Johnson Foundation, University of Pennsylvania and University City Science Center, Philadelphia, Pennsylvania 19104

(Received 23 September 1997; accepted for publication 16 June 1998)

A construction of a stochastic, Lagrangian path integral representation is presented for classical Green's functions or propagators for radiative transfer in random scattering medias. As Fermat rays comprising the initial collimated pulse (e.g., from a laser) enter the medium, they evolve into a bundle of random paths or trajectories due to scattering. Stochastically, this can be interpreted as a bundle of randomly evolving Markov trajectories traced out by a gas or ensemble of ''Brownian-type'' classical point photons undergoing multiple scatterings. The path integral is structurally of the same form as a Wick rotated Euclidean quantum Feynman integral with direct optical analogs of the Hamiltonian and Lagrangian emerging. However, the optical stochastic integral is real and is defined in real time becoming exponentially damped rather than oscillatory. The calculation also constitutes an alternative mathematical derivation of the time dependent diffusion equation from the radiative transfer theory. The limits of the path integral representation give a maximally diffused Gaussian distribution of the heat kernal form (the large scatter limit) and a Beer's law exponential decay corresponding to the extremal Fermat rays obeying Euler–Lagrange equations (the zero scatter limit). The approach also highlights the direct structural analogs between classical radiative transfer and optical diffusion in real time (t) and ordinary quantum mechanics in Euclidean time ($\sqrt{(-1)}t=it$).
© *1998 American Institute of Physics.* [S0022-2488(98)03910-3]

I. INTRODUCTION

Multiple scattering of light in random medias can be described by the following levels of theoretical description. The macroscopic approach resides in solving an effective hyperbolic diffusion equation, which gives spatiotemporal variations in the scalar diffuse intensity or fluence $\phi(\mathbf{x},t)$ at points $\{\mathbf{x}\}$ at time t, or $\phi(\mathbf{x})$ in the steady state.[1,2] This approach is excellent in the bulk of the medium on length scales much larger than the mean free path l, and when the particles comprising the medium scatter strongly and isotropically with little or no absorption.[1]

The mesoscopic approach[1,3,4] commonly utilized in astrophysics and nuclear engineering, is based on the heuristic radiative transfer theory (RTT). Spatiotemporal variations in optical radiance or intensity $R(\mathbf{x},\mathbf{\Omega},t)$ at points $\mathbf{x}$, from propagation direction $\mathbf{\Omega}$, are described by an integrodifferential balance equation, structurally similar to the Boltzmann equation of kinetic theory. Although the formalism is flexible, the general analytical solutions are unknown even in one dimension; most approaches to the equation are therefore a compromise between computational tractability and accuracy. The macroscopic diffusion approach is recovered in the limit of length scales larger than 1 and when the particles are highly scattered.[1] In both descriptions, we quantify scattering and absorption properties of the medium by the bulk parameters μ_s and μ_a, respectively, which are the scattering probability per unit length and the absorption probability per unit length (in cm^{-1}). The macroscopic approach is recovered from the mesoscopic approach in the diffusion approximation when $\mu_s \gg \mu_a$.

The microscopic or analytical approach begins with Maxwell's equations or the wave equations and the scattering–absorbing properties of individual particles. Differential and integral equations are developed for statistical quantities such as correlation functions and field variances.

[a)]Address for all correspondance: Steven Miller, Ph.D., 142 Cameron Drive, Kilmarnock KA3 7PL, Scotland, United Kingdom. Electronic mail: 101551.1243@Compuserve.com

 © 1998 American Institute of Physics

This approach has a mathematically rigorous structure and can (in principle) include all possible wave related phenomenology.[1] However, the various theories which can provide solutions are all approximate, being viable only within specific ranges of parameters. The Dyson and Bethe–Salpeter equations and Twersky's theory are examples of this type.[1,5]

Each of these viewpoints essentially represents a qualitative improvement on the previous one. In this paper, we will concentrate on the more pragmatic transport and diffusion approaches where wave properties are ignored, and interpret the classical probabilistic multiple scattering of light as a fundamental stochastic process with Markovian structure. A collimated light pulse (from a laser, for example) is incident at a point on a turbid media. As Fermat rays comprising the pulse enter the medium, they evolve irreversibly into a bundle of random paths or trajectories due to scattering. This can be interpreted as a bundle of Markov chains traced out by a statistical ensemble of Brownian-type classical point photons being randomly scattered and absorbed.

A simple stochastic analysis for the Brownian-type motions of the ensemble of classical Markov photons is then used to construct a stochastic path integral representation for the propagator of the scattered light pulse. Structurally, the path integral representation is of the same form as the Wick rotated Feynman path integral for the quantum amplitude. Recently, interest has also been renewed in stochastic analysis as a means of interpreting quantization and Feynman path sums.[6] However, the stochastic integral representation for the scattered light pulse is exponentially damped and real (describing a probability or Green's function in real time) rather than complex and oscillatory. Optical analogs for the Hamiltonian and Lagrangian are also identified. The path integral itself can be interpreted as a partition function analog for the statistical ensemble of Brownian optical paths. The limit $\mu_s \gg \mu_a$, give a Gaussian (heat) kernal for maximally diffused light. The zero-scattering ($\mu_s \to 0$) limit reduces the integral to a time dependent Beer's law for a collimated pulse. This last point is analogous to the classical limit of quantum mechanics where the classical Euler–Lagrangian path is recovered in the limit as Planck's constant goes to zero. Solutions and interpretations of radiative transfer as a sum over paths have been considered in Refs. 7–11 in connection with some technological applications. However, the path integral representations here were not derived and were presented heuristically. In this paper we derive and justify a path integral representation using a more rigorous stochastic based approach not previously considered. The approach also highlights some direct analogs between classical radiative transfer in real time and ordinary quantum mechanics in Euclidean time.

II. CONSTRUCTION OF A STOCHASTIC FEYNMAN PATH INTEGRAL REPRESENTATION

We begin with the fundamental equation of radiative transfer.[1,3,4] Let $\mathscr{M}$ be a semi-infinite scattering medium and let F be a region of free space with smooth boundary $\partial\mathscr{M} \equiv (\mathscr{M} \cap F)$ at $z=0$, since $\mathbf{x}=(x,y,z)=(x,y,0)$ on the boundary. For a slab, the boundaries are $\partial\mathscr{M}_1 = (\mathscr{M} \cap F)_1$ at $z=0$ and $\partial\mathscr{M}_2 = (\mathscr{M} \cap F)_2$ at $z=d$. At each $\mathbf{x}=(x,y,z) \in \mathscr{M}$, we have homogeneous coefficients $\mu_s(\mathbf{x})=\mu_s$ and $\mu_a(\mathbf{x})=\mu_a$. If a light pulse (e.g., from a laser or fiber) comprised of collimated or Fermat rays $\{\mathscr{R}\}$, is incident normal to $(\mathscr{M} \cap F)$ or $z=0$ at $t_0=0$, the evolving scattered light distribution within $\mathscr{M}$ for $t > t_0 = 0$ is given by the radiative transfer equation,

$$\frac{1}{c}\frac{\partial R(\mathbf{x},\mathbf{\Omega},t)}{\partial t} + [\mathbf{\Omega}.\nabla + (\mu_s+\mu_a)]R(\mathbf{x},\mathbf{\Omega},t) = \left(\frac{\mu_s+\mu_a}{4\pi}\right)\int \mathscr{P}(\mathbf{\Omega}|\mathbf{\Omega}')R(\mathbf{x},\mathbf{\Omega}',t)\,d\mathbf{\Omega}', \quad (1)$$

where $R(\mathbf{x},\mathbf{\Omega},t)$ is the radiance at $\mathbf{x}$ from direction $\mathbf{\Omega}$ at time t, and $\mathbf{x} \in \mathscr{M}$ and $t \in \mathbb{R}$. The differential $d\mathbf{\Omega}'$ is a solid angle with $d\mathbf{\Omega}' \equiv \sin\theta' d\theta' d\phi'$. The function $\mathscr{P}(\mathbf{\Omega}|\mathbf{\Omega}')$ is the phase function or scattering probability for all scatterings $\mathbf{\Omega}' \to \mathbf{\Omega}$,[1] which equals unity for isotropic scatter. In this paper, we will initially consider isotropic scattering. A solution $R(\mathbf{x},\mathbf{\Omega},t)$ can be considered on the basis of initial data. The initial data can be the incident radiance $R(\mathbf{x}_0,\mathbf{\Omega}_0,t_0)=R(z=z_0;\theta_0,\phi_0,t_0)$ at boundary $\mathbf{x}_0=(x_0,y_0,z_0)=(0,0,z_0)=(0,0,0) \in (\mathscr{M} \cap F)_1$ at time t_0, where the unit vector $\mathbf{\Omega}_0$ is normal to the plane $z=z_0$ for the incident pulse and points into the medium along ($\theta_0=\pi/2$). The radiance $R(\mathbf{x},\mathbf{\Omega},t)$ at any point below the initial data point at time $t > t_0$, arises from all possible scattered optical paths that pass through $\mathbf{x}$ at time t in the direction $\mathbf{\Omega}$. We will present an independent derivation of a stochastic path integral representation

for these Brownian-type paths. Given initial data at some point $\mathbf{x}_0=(x_0,y_0,z_0)$, the radiance $R(\mathbf{x},\mathbf{\Omega},t)$ is given by Huygen's principle in terms of the Green's function $G(\mathbf{x},\mathbf{\Omega},t;\mathbf{x}_0,\mathbf{\Omega}_0,t_0)$,

$$R(\mathbf{x},\mathbf{\Omega},t)=\int d^3x\int dt\int d\mathbf{\Omega}G(\mathbf{x},\mathbf{\Omega},t;\mathbf{x}_0,\mathbf{\Omega}_0,t_0)R(\mathbf{x}_0,\mathbf{\Omega}_0,t_0). \tag{2}$$

If we integrate over all scatterings of order n over all paths, then

$$R(\mathbf{x},\mathbf{\Omega},t)=\int\int\int\prod_{n>0} d^3x_n d\mathbf{\Omega}_n dt_n G(\mathbf{x_{n+1}},\mathbf{\Omega}_{n+1},t_{n+1};\mathbf{x_n},\mathbf{\Omega}_n,t_n)R(\mathbf{x_0},\mathbf{\Omega}_0,t_0). \tag{3}$$

This highlights the fundamental Markovian structure of radiative transport within the macroscopic and mesoscopic approaches. The classical Green's function or propagator $G(\mathbf{x}_{n+1},t_{n+1};\mathbf{x}_n,t_n)$ therefore must satisfy a Markov–Kolmogorov evolution equation,[12,13]

$$G(\mathbf{x_{n+1}},t_{n+1};\mathbf{x_{n-1}},t_{n-1})=\int\int d^3x_n dt_n G(\mathbf{x_{n+1}},t_{n+1};\mathbf{x_n},t_n)G(\mathbf{x_n},t_n;\mathbf{x_{n-1}},t_{n-1}), \tag{4}$$

where the angular integral is now absorbed into $\int d^3x$ (we are only concerned with the probability of light arriving at $\mathbf{x}_{n+1}$ at $t=t_{n+1}$, regardless of the direction of its approach). The probability of reaching $\mathbf{x}_{n+1}$ from the nth scattering at $\mathbf{x}_n$, does not depend on the previous history or orientations of the path which brought it to $\mathbf{x}_n$. There is therefore no memory of the past for a given present which defines a Markov process.[12]

The first step of the derivation requires that we consider the Green's function or transition probability $G(\mathbf{x}_{n+1},t_{n+1};\mathbf{x}_n,t_n)$ over an infinitesimally small time interval δt. For convenience we will set $\mathbf{x}'=\mathbf{x}_n$, $t=t_n$, $\mathbf{x}=\mathbf{x}_{n+1}$, and $t_n=t+\delta t$, giving a propagator $G(\mathbf{x},t+\delta t;\mathbf{x}',t)$. The Kolmogorov integral in this infinitesimal limit can be written as

$$G(\mathbf{x},t+\delta t)=\int d^3x' G(\mathbf{x},t+\delta t;\mathbf{x}',t)G(\mathbf{x}',t), \tag{5}$$

where any $G(\mathbf{x},t)$ reduces to $G(\mathbf{x},t;\mathbf{x}_0,t_0)$ when $G(\mathbf{x},t_0)=\delta(\mathbf{x}-\mathbf{x}_0)$. For small δt, the moments of the Markovian Green's function are expandable in powers $(\delta t)^\beta$ for all integer $\beta>0$. We can define a set $\{\mathscr{U}_\beta(x,y,x,t)\}$ or $\{\mathscr{U}_\beta(\mathbf{x},t)\}$ for all $\beta\in Z$ by

$$\mathscr{U}_\beta(\mathbf{x}',t)=\frac{1}{\delta t}\left(\frac{\beta}{\beta!}\right)\int G(\mathbf{x},t+\delta t;\mathbf{x}',t)(\mathbf{x}-\mathbf{x}')^\beta d^3x. \tag{6}$$

For a homogeneous, linear Markov process $G(\mathbf{x},t+\delta t;\mathbf{x}',t)$, the integral on the right over the Green's function or kernel, should be constant for all $\mathbf{x}\in\mathbb{R}^3$ and $t\in\mathbb{R}$. For example, at order $\beta=2$, then $\mathscr{U}_2(\mathbf{x},t)=\mathscr{U}_2=\text{const}$. We will return to this point (and verify it) later. Note also, this definition is strictly valid only in the limit as $\delta t\to 0$ which is true for $\mu_s\gg\mu_a$, the diffusion approximation. The propagator or kernel $G(\mathbf{x},t+\delta t;\mathbf{x}',t)$ is also normalized as

$$\int d^3x G(\mathbf{x},t+\delta t;\mathbf{x}',t)=1. \tag{7}$$

Given any arbitrary smooth real function $\psi(\mathbf{x})$ one can form a Taylor expansion of $\psi(\mathbf{x})$ with respect to $\psi(\mathbf{x}')$,

$$\psi(\mathbf{x})=\psi(\mathbf{x}')+\sum_{\beta=0}^{\infty}\frac{(\mathbf{x}-\mathbf{x}')^\beta}{\beta!}\frac{\partial^{(\beta)}\psi(\mathbf{x}')}{\partial\mathbf{x}^\beta}=\left((\mathbf{x}-\mathbf{x}')\frac{\partial\psi(\mathbf{x}')}{\partial\mathbf{x}}\right)+\left(\frac{(\mathbf{x}-\mathbf{x}')^2}{2}\frac{\partial^2\psi(\mathbf{x}')}{\partial\mathbf{x}^2}\right)+\cdots. \tag{8}$$

An integrodifferential equation for the evolution of the statistical average $\langle\psi(\mathbf{x})\rangle$ can then be derived. Equation (5) is multiplied by $\psi(\mathbf{x}')$ with $\psi(\mathbf{x})$ placed under $\int d^3x'$ and then integrated over d^3x using the right-hand side of Eq. (8). A straightforward calculation using the definition for $\mathscr{U}_\beta(\mathbf{x},t)$ and the propagator normalization (7) gives

$$\langle \psi(\mathbf{x})\rangle_{t+\delta t} = \int d^3x'\, \psi(\mathbf{x}')G(\mathbf{x}',t+\delta t)$$

$$= \int\int \sum_{\beta=1}^{\infty} \frac{(\mathbf{x}-\mathbf{x}')^\beta}{\beta!}\left[\left(\frac{\partial}{\partial \mathbf{x}}\right)^\beta \psi(\mathbf{x}')\right] G(\mathbf{x},t+\delta t;\mathbf{x}',t)G(\mathbf{x}',t)d^3x$$

$$= \langle \psi(\mathbf{x}')\rangle_t + \delta t \sum_{\beta=1}^{\infty} \int d^3x' G(\mathbf{x}',t)\left[\left(\frac{\partial}{\partial \mathbf{x}}\right)^\beta \psi(\mathbf{x}')\right]\left[\frac{1}{\beta}\,\mathscr{U}_\beta(\mathbf{x}',t)\right], \tag{9}$$

where the expectation value or average of $\psi(\mathbf{x})$ at time $t\in\mathbb{R}$ associated with $G(\mathbf{x},t)$ is given by the usual statistical mechanical definition,

$$\langle \psi(\mathbf{x})\rangle_t = \int d^3x\, \psi(\mathbf{x})G(\mathbf{x},t). \tag{10}$$

Given a function $\psi(\mathbf{x})$, the differential equation for the statistical average $\langle \psi(\mathbf{x})\rangle$, in the limit as $\delta t\to 0$ is

$$\lim_{\delta t-\to 0}|\langle \psi(\mathbf{x})\rangle_{t+\delta t} - \langle \psi(\mathbf{x})\rangle_t|/\delta t = \frac{\partial}{\partial t}\langle \psi(\mathbf{x})\rangle = \sum_{\beta=1}^{\infty}\left\langle \frac{1}{\beta}\,\mathscr{U}_\beta(\mathbf{x},t)\left(\frac{\partial}{\partial \mathbf{x}}\right)^\beta \psi(\mathbf{x})\right\rangle$$

$$= \sum_{\beta=1}^{\infty}\left[\int d^3x\left(\frac{1}{\beta}\,\mathscr{U}_\beta(\mathbf{x},t)\left(\frac{\partial}{\partial \mathbf{x}}\right)^\beta \psi(\mathbf{x})\right)G(\mathbf{x},t)\right]$$

$$\equiv -\langle \xi\psi(\mathbf{x})\rangle, \tag{11}$$

where the operator ξ is defined as

$$\xi \equiv -\sum_{\beta=1}^{\infty}\frac{1}{\beta}\,\mathscr{U}_\beta(\mathbf{x},t)(\partial/\partial \mathbf{x})^\beta = \left(\mathscr{U}_1(\mathbf{x},t)\,\frac{\partial}{\partial \mathbf{x}}\right) + \left(\frac{1}{2}\,\mathscr{U}_2(\mathbf{x},t)\,\frac{\partial^2}{\partial \mathbf{x}^2}\right) + \cdots. \tag{12}$$

If we now set $\psi(\mathbf{x})=\mathbf{x}$, $\mathbf{x}^2$, $\mathbf{x}^3$ in (11), where $\mathbf{x}\in\mathscr{R}$, lies on a collimated or Fermat ray $\mathscr{R}$ within medium $\mathscr{M}$, we obtain time dependent fluctuation, dispersion or moments relations. These describe to each order β, the time dependent deviation or dispersion of the incident collimated beam due to scattering, away from the Fermat or straight ray propagation,

$$\partial\langle \mathbf{x}\rangle/\partial t = \langle \mathscr{U}_1\rangle,$$

$$\partial\langle \mathbf{x}^2\rangle/\partial t = \langle \mathbf{x}\mathscr{U}_1\rangle + \langle \mathscr{U}_2\rangle, \tag{13}$$

$$\partial\langle \mathbf{x}^3\rangle/\partial t = \langle \mathbf{x}^2\mathscr{U}_1\rangle + 2\langle \mathbf{x}\mathscr{U}_2\rangle + 2\langle \mathscr{U}_3\rangle,$$

and so on for all integers $\beta\in Z$. Zero dispersion (with time) around $\mathbf{x}\in\mathscr{R}$ at order 2 for example, requires $\langle \mathbf{x}^2\rangle = t\langle \mathbf{x}\mathscr{U}_1\rangle + \langle \mathscr{U}_2\rangle = 0$ or $\mathscr{U}_1=0$ and $\mathscr{U}_2=0$. The $\mathscr{U}_2$ will later be identified as a diffusion coefficient.

If the choice $\psi(\mathbf{x})=\delta(\mathbf{x}-\mathbf{x}')$ is made then $G(\mathbf{x},t)=\langle \delta(\mathbf{x}-\mathbf{x}')\rangle$. If this is now inserted back into Eq. (11), then integrating by parts gives a generic differential equation for $G(\mathbf{x}',t)$ to all orders in β,

$$\frac{\partial G(\mathbf{x}',t)}{\partial t} = \sum_{\beta=1}^{\infty}\frac{(-1)^\beta}{\beta}\left[\frac{\partial}{\partial \mathbf{x}'}\right]^\beta \mathscr{U}_\beta(\mathbf{x}',t)G(\mathbf{x}',t) = -\mathscr{H}G(\mathbf{x}',t), \tag{14}$$

where $\mathscr{H}$ is given by

$$-\mathscr{H} = \sum_{\beta=1}^{\infty}\frac{(-1)^\beta}{\beta}\left[\frac{\partial}{\partial \mathbf{x}'}\right]^\beta \mathscr{U}_\beta(\mathbf{x}',t) \equiv \sum_{\beta=1}^{\infty}\mathscr{H}_\beta. \tag{15}$$

Note that $\mathbf{x}$ and $\mathbf{x}'$ are just equivalent stochastic spatial vector labels. Linear, scalar optical diffusions will emerge from truncation at some β. For now, we keep the calculation defined to all orders in β until we have the final result.

Given any $G(\mathbf{x},t)$ we can define the Fourier transform $\mathscr{F}(\mathbf{x},t)$ at fixed $t\in\mathbb{R}$ and its inverse as

$$\mathscr{F}(\mathbf{c},t)=\langle\exp(i\mathbf{c}\cdot\mathbf{x})\rangle=\int d^3x\ \exp(i\mathbf{c}\cdot\mathbf{x})G(\mathbf{x},t) \tag{16a}$$

$$G(\mathbf{x},t)=\int d^3c\ \exp(-i\mathbf{c}\cdot\mathbf{x})\mathscr{F}(\mathbf{c},t). \tag{16b}$$

The Fourier transform $\mathscr{F}$ then essentially takes all spatial vectors $\{\mathbf{x}(t)\}$ at fixed $t\in\mathbb{R}$ to a conjugate set of variables $\{\mathbf{c}(t)\}$ at the same $t\in\mathbb{R}$. The βth order fluctuations or moments $[\langle\mathbf{x}\rangle]^\beta$ can then be expressed as the βth derivatives of $\mathscr{F}$ since

$$\mathscr{F}(\mathbf{c},t)=\sum_{\beta=1}^{\infty}(i\mathbf{c})^\beta\frac{\langle\mathbf{x}\rangle^\beta}{\beta!}=\left((i\mathbf{c})\langle\mathbf{x}\rangle+(i\mathbf{c})^2\frac{\langle\mathbf{x}\rangle^2}{2}\right)+\cdots. \tag{17}$$

If we expand $\mathscr{F}$ as

$$\langle\exp(i\mathbf{c}\cdot\mathbf{x})\rangle=\exp\left[\sum_{\beta=1}^{\infty}(1/\beta!)(i\mathbf{c})^\beta\mathscr{S}_\beta\right]=\exp\left[i\mathbf{c}\mathscr{S}_1-\frac{1}{2}\mathbf{c}^2\mathscr{S}_2\right], \tag{18}$$

then the $\mathscr{S}(\mathbf{x})$ can be identified within this context as statistical Thiele semi-invariants or ''linked moments.'' [14] For independent stochastic variables $\{\mathbf{e}\}$ and $\{\mathbf{f}\}$ we have the linearity $\mathscr{S}_\beta(\mathbf{e}+\mathbf{f})=\mathscr{S}_\beta(\mathbf{e})+\mathscr{S}_\beta(\mathbf{f})$. We can apply the notation,

$$\mathscr{S}_\beta(\mathbf{x})\equiv\langle\mathbf{x}^\beta{:}\,\text{link}\rangle. \tag{19}$$

To express linked moments (at fixed t) in terms of ordinary moments we expand both sides of (18) so that we obtain intrinsic correlators,

$$\langle\mathbf{x}{:}\,\text{link}\rangle=\langle\mathbf{x}\rangle, \tag{20a}$$

$$\langle\mathbf{x}^2{:}\,\text{link}\rangle=\langle\mathbf{x}^2\rangle-\langle\mathbf{x}\rangle\langle\mathbf{x}\rangle, \tag{20b}$$

$$\langle\mathbf{x}^3{:}\,\text{link}\rangle=\langle[\mathbf{x}-\langle\mathbf{x}\rangle]^3\rangle, \tag{20c}$$

$$\langle\mathbf{x}^4{:}\,\text{link}\rangle=\langle(\mathbf{x}-\langle\mathbf{x}\rangle)^4\rangle-3[\langle(\mathbf{x}-\langle\mathbf{x}\rangle)^2\rangle]^2, \tag{20d}$$

and so on. This already suggests that $\beta=2$ is the maximum order for diffuse light with second order correlations $\mathscr{S}_2(\mathbf{x})=\langle\mathbf{x}^2{:}\,\text{link}\rangle$. We will return to this point. The Fourier transform $\mathscr{F}(\mathbf{c},t)$ can now be written in the compact symbolic form

$$\langle\exp(i\mathbf{c}\cdot\mathbf{x})\rangle=\exp[\langle\exp(i\mathbf{c}\cdot\mathbf{x})-1]\rangle{:}\,\text{link}]. \tag{21}$$

One can then derive a Feynman-type stochastic path integral from these results. The operator $\mathscr{H}$ can be written as

$$-\mathscr{H}=\sum_{\beta=1}^{\infty}\frac{(-1)^\beta}{\beta}\left[\frac{\partial}{\partial\mathbf{x}}\right]^\beta\mathscr{U}_\beta(\mathbf{x},t)=\sum_{\beta=1}^{\infty}\mathscr{H}_\beta. \tag{22}$$

Associated with this operator, we define $\mathscr{H}(\mathbf{c},\mathbf{x},t)$ at fixed t,

$$-\mathscr{H}(\mathbf{c},\mathbf{x},t)=\sum_{\beta=1}^{\infty}[i\mathbf{c}]^\beta\frac{1}{\beta}\mathscr{U}_\beta(\mathbf{x},t), \tag{23}$$

where we have used $\mathbf{c}=i\partial/\partial\mathbf{x}$. This follows from differentiation of the inverse of Fourier transform $\mathscr{F}$ so that $\partial G/\partial\mathbf{x}=-i\mathbf{c}G$. The diffusion coefficients $\mathscr{U}_\beta$ can be expressed as moments strictly in the limit as $\delta t\to 0$,[12]

$$\mathscr{U}_\beta(\mathbf{x},t)=\lim_{\delta t-\to 0}\frac{1}{\delta t}\langle[\mathbf{x}(t+\delta t)-\mathbf{x}(t))]^\beta\rangle \tag{24}$$

or as the linked moments previously defined,

$$\mathscr{U}_\beta(\mathbf{x},t)=\lim_{\delta t-\to 0}\frac{1}{\delta t}\langle[\mathbf{x}(t+\delta t)-\mathbf{x}(t)]^\beta:\text{link}\rangle. \tag{25}$$

If we set $\delta\mathbf{x}\equiv\mathbf{x}(t+\delta t)-\mathbf{x}(t)$, then using (23) we can write the operator $\mathscr{H}$ as

$$-\mathscr{H}(\mathbf{x},\mathbf{c},t)=\frac{1}{\delta t}\langle[\exp(i\mathbf{c}\cdot\delta\mathbf{x})-1]:\text{link}\rangle. \tag{26}$$

The Fourier transform can then be written as

$$\mathscr{F}=\langle\exp(i\mathbf{c}\cdot\delta\mathbf{x})\rangle=\int d[\delta\mathbf{x}]\exp(i\mathbf{c}\cdot\delta\mathbf{x})G(\mathbf{x}+\delta\mathbf{x},t+\delta t;\mathbf{x},t)=\exp[-\mathscr{H}(\mathbf{c},\mathbf{x},t)]. \tag{27}$$

This is correct to order δt. The inverse Fourier transform gives an expression for the transition probability or Green's function,

$$\begin{aligned}G(\mathbf{x}+\delta\mathbf{x},t+\delta t;\mathbf{x},t)&=\int d^3c\ \exp[-i\mathbf{c}\cdot\delta\mathbf{x}-\mathscr{H}(\mathbf{c},\mathbf{x})\,\delta t]\\&=\int d^3c\ \exp\left[-i\,\delta t\left(\mathbf{c}\,\frac{\delta\mathbf{x}}{\delta t}-\mathscr{H}(\mathbf{c},\mathbf{x})\right)\right].\end{aligned} \tag{28}$$

We can see at once that this structure is essentially a ''piece'' of a Feynman path sum[15,16] if we identify $\mathbf{c}$ as a conjugate momentum analog and $\mathscr{H}$ as a Hamiltonian analog. However, let us first consider the truncation order β required to describe noninteracting classical Brownian-type diffusing photon paths. Since light is massless and chargeless there are no interactions or forces between the individual photon paths in any ensemble of paths. The equations describing classical light diffusion away from the Fermat propagation, must then be linear. It is also not possible to apply an exterior field or perturbation that would induce vectorial drift of the paths. We therefore set $\mathscr{U}_1=0$ and $\mathscr{U}_2(\mathbf{x},t)=\mathscr{U}_2=2\mathscr{U}=\text{constant}$, which is Eq. (6) to the second order for a homogeneous, linear Markov process, with all $\mathscr{U}_\beta=0$ for $\beta>2$. Equation (14) is then linear and hyperbolic and describes a free homogeneous, diffusing light distribution due to a Dirac function point source profile,

$$\frac{\partial}{\partial t}G(\mathbf{x},t)=\frac{1}{2}\mathscr{U}_2\left[\frac{\partial}{\partial\mathbf{x}}\right]^2G(\mathbf{x},t)\equiv\frac{1}{2}\mathscr{U}_2\nabla^2G(\mathbf{x},t)=0, \tag{29a}$$

$$\frac{\partial}{\partial t}G(\mathbf{x},t)-\mathscr{U}\nabla^2G(\mathbf{x},t)=0. \tag{29b}$$

This is the free, linear optical diffusion equation. We will briefly return to the kernel solution of this equation and its relation to (6) in Sec. III B. The correlations are second order as defined in (20b). Note that δt is infinitesimal only when $\mu_s\gg\mu_a$ (the diffusion approximation condition) so that this method constitutes an alternative derivation of the time dependent optical diffusion equation from the transfer equation of (1). The ''Hamiltonian'' $\mathscr{H}$ is then quadratic in $\mathbf{c}$ when $\beta=2$. Equation (28) can be substituted into the Markov–Kolmogorov equations in (4) to give

$$\int \prod_{n>0} d^3\mathbf{x_n} \int \prod_{n>0} d^3\mathbf{c_n} \exp\left[-i\int_0^t \delta t\left(\mathbf{c_n}\frac{\delta \mathbf{x_n}}{\delta t}-\mathscr{H}_2\right)\right]$$
$$=\int \mathscr{D}x \int \mathscr{D}c \exp\left[-i\int_0^t \delta t\left(\mathbf{c}\frac{\delta \mathbf{x}}{\delta t}-\mathscr{H}_2\right)\right]. \tag{30}$$

This is a Hamiltonian path integral structure in a ''phase space'' $(\mathbf{x}(t),\mathbf{c}(t))=(x(t),y(t),c_1(t),c_2(t),c_3(t))$ at fixed t. Note that $\delta t=\int \delta t$ with limits t and 0. The result is essentially the same as the Feynman path measure in quantum mechanics for the evolution kernal of a wave function for a particle of mass m (Ref. 15) in phase space $(\mathbf{x}(t),\mathbf{p}(t))$, where the Hamiltonian is $H=p^2/2m$ and the Lagrangian is $L=\frac{1}{2}m(d\mathbf{x}(t)/dt)^3$. From Eq. (22) or (23) to order $\beta=2$ (which is now the lowest order in the absence of drift) and $\frac{1}{2}\mathscr{U}_2=\mathscr{U}=$constant, the quadratic Hamiltonian analog is

$$\mathscr{H}(\mathbf{c},t)\equiv\mathscr{H}=\tfrac{1}{2}|\mathbf{c}(t)|^2(\mathscr{U}_2)^2\equiv c^2\mathscr{U}_2. \tag{31}$$

The ''conjugate momentum'' $\mathbf{c}(t)$ is identified in the optical diffusion description with the speed of light in the medium and is a constant so that $d\mathbf{c}(t)/dt=0$ for elastic scattering. However we do not fix $|\mathbf{c}(t)|=c=$const. or perform a functional analytical continuation until the functional integrals over all d^3c_n are performed. We first perform a functional analytic continuation $(i\to 1)$ to get a viable classical (real) integral in Lagrangian form in real time, since the effect of the Fourier transform was to introduce Euclidean time $(i\,\delta t)$. The stochastic evolution of the bundle of random optical (Markov) paths due to scattering, has a real Feynman path measure representation in $(1+3)$ Minkowski space,

$$G(\mathbf{x},t;\mathbf{x}',0)=\int \prod_{n=0} d^3x_n \exp\left(-\int_0^t \mathscr{L}dt'\right)\equiv\int \prod_{n>0} d^3x_n \exp[-S(\mathbf{x}(t))]=\int \mathscr{D}x \exp[-S], \tag{32}$$

where the ''momentum'' integral $\int d^3c$ is performed to yield the Lagrangian form with the action $S(\mathbf{x}(t))=\int \mathscr{L}dt$. The analytically continued integral is exponentially damped rather than oscillatory which ensures the correct attenuative optical decay into the random medium. We can therefore identify the following direct analogies or dualities between classical scalar optical diffusion or radiative transfer and ordinary quantum mechanics,

i	1
it	t
$\mathbf{p}$	$\mathbf{c}$
$1/2m$	$\mathscr{U}$
$H=p^2/2m$	$\mathscr{H}_2=c^2\mathscr{U}=\mathscr{H}$
$L=\frac{1}{2}m[d\mathbf{x}(t)/dt]^2$	$\mathscr{L}=(1/4\mathscr{U})[d\mathbf{x}(t)/dt]^2$.

Given the standard Feynman path integral for quantum mechanics, these substitutions will recover the classical physics of radiative transfer and optical diffusion theory.

III. LIMITS OF THE OPTICAL PATH INTEGRAL

In this section we consider the limits of the integral as (1) $\mu_s\to 0$ or $n=0$ and (2) $n\to\infty$ (or $\mu_s\gg\mu_a$).

A. Zero scattering limit: $\mu_s\to 0$ or $n=0$

We consider the limit as scattering vanishes. Consider the action $S[\mathbf{x}'(t)]$ at t, where $\mathbf{x}'\in\mathscr{R}$ and where $\mathscr{R}$ is an interior Fermat ray such that $\mathbf{x}$ does not lie on the Fermat ray. The Taylor expansion of $S[\mathbf{x}]$ with respect to $S[\mathbf{x}']$ as in (8) is

$$S[\mathbf{x}]=S[\mathbf{x}']+\delta\mathbf{x}\partial_{\mathbf{x}}S[\mathbf{x}']+\tfrac{1}{2}\delta\mathbf{x}\delta\mathbf{x}\partial_{\mathbf{x}}\partial_{\mathbf{x}}S[\mathbf{x}']+\tfrac{1}{6}\delta\mathbf{x}\delta\mathbf{x}\delta\mathbf{x}\partial_{\mathbf{x}}\partial_{\mathbf{x}}\partial_{\mathbf{x}}S[\mathbf{x}']+\cdots. \tag{33}$$

Cubic and higher order terms ($\beta>2$) would correspond to classical photon–photon interactions or forces which do not exist. However we require $\partial S[\mathbf{x}']/\partial\mathbf{x}=0$ since points $\{\mathbf{x}'\}$ on the Fermat rays $\mathscr{R}$ are critical points for the action. Again, this is consistent with setting $\mathscr{U}_1=0$ and $\mathscr{U}_\beta=0$ for all $\beta>2$. There is a ''classical limit'' analog as $\mu_s\to 0$, and the diffuse light pulse reduces to a collimated pulse as the Brownian paths collapse to the minimum (extremal) path of Fermat propagation between two points.

The Euler–Lagrange equations here simply state that Fermat ray propagation within the medium in the absense of scattering is an extrema (that is, the speed of light is a constant). The Euler Lagrange equations with $\mathscr{L}=(1/4\mathscr{U})(d\mathbf{x}/dt)^2$ are

$$\frac{d}{dt}\left(\frac{\partial\mathscr{L}}{\partial\mathbf{x}}\right)=\frac{1}{4\mathscr{U}}\left(\frac{d^2\mathbf{x}(t)}{dt^2}\right)=\frac{\partial\mathscr{L}}{\partial\mathbf{x}}=0 \tag{34}$$

so that const.$=2\mathscr{U}\mathbf{c}(t)$ and $|\mathbf{c}(t)|=c$ is the speed of light in a medium requiring $d\mathbf{c}(t)/dt=0$. The distribution of paths must collapse to the single Fermat path $\mathscr{R}$. In the limit as $\mu_s\to 0$ with no Brownian scatter paths (paths with $n>0$) connecting any fixed points $\mathbf{x}'$ and $\mathbf{x}$ for $n=0$,

$$G(\mathbf{x},t;\mathbf{x}',0)=\exp\left[-\int_0^t dt(2\mathscr{U})^{-1}\left(\frac{d\mathbf{x}(t)}{dt}\right)^2\right]=\exp\left[-\int_0^t dt(2\mathscr{U})^{-1}c\left(\frac{d\mathbf{x}(t)}{dt}\right)\right], \tag{35}$$

where $\mathbf{x}$ now lies on a straight line from $\mathbf{x}'$. Since $(d\mathbf{x}(t)/dt)=\mathbf{c}$ and the arc length $(d\mathbf{x}(t)/dt)^2dt^2$ is now a line segment of a straight Fermat geodesic, the integral is trivial and gives

$$G(\mathbf{x},t;\mathbf{x}',0)=\exp[-(2\mathscr{U})^{-1}c|\mathbf{x}-\mathbf{x}'|) \tag{36}$$

since $t=|\mathbf{x}-\mathbf{x}'|/c$ or $dt=dx/c$. The homogeneous diffusion constant $\mathscr{U}$ in this formalism is related to the usual diffusion constant definition $\mathscr{D}$ by a factor of $\frac{2}{3}$ so that $\mathscr{U}=3\mathscr{D}/2$, where $\mathscr{D}=c[3(\mu_a+\mu_s)]^{-1}$ is the form given in the literature as Refs. 2, 17, 18. Since $\mathscr{D}$ and $\mathscr{U}$ are constants, they are simply rescaled by a factor of $\frac{3}{2}$ to make both formalisms match, hence $(3/2\mathscr{U})c=(\mu_s+\mu_a)$. With hindsight, one can always redefine (6) to within this factor.

It is important to note the time dependence is now removed from the coherent Green's function in (36). Since the two points $\mathbf{x}'$ and $\mathbf{x}$ are fixed, all Fermat rays $\mathscr{R}$ propagate from $\mathbf{x}'$ to $\mathbf{x}$ at the *same* time. The probability is then simply $G(\mathbf{x};\mathbf{x}')$ for all Fermat rays $\mathscr{R}$ in a coherent or collimated pulse. It is only when the Fermat rays scatter and evolve into random Brownian-type trajectories that there exist probabilities $\{G(\mathbf{x},t)\}$ for the random paths to arrive at $\mathbf{x}$ in different times t, having started from the same $\mathbf{x}'$. However, the coherent intensity at $\mathbf{x}$ varies in time with the source. For example, if the source intensity at $\mathbf{x}'$ has the profile $I_0(t)$ then the intensity of unscattered light at $\mathbf{x}$ is simply

$$I(\mathbf{x},t+|\mathbf{x}-\mathbf{x}'|/c)=G(\mathbf{x},\mathbf{x}')I_0(t)=I_0(t)\exp[-\mu_a|\mathbf{x}-\mathbf{x}'|], \tag{37}$$

which is Beer's law with time dependence. The effect of the phase function $\mathscr{P}(\mathbf{\Omega}|\mathbf{\Omega}')$ is built into the (unknown) kernel $G(\mathbf{x},t;\mathbf{x}',t')$. The anisotropy is $g=\langle\cos\theta\rangle=\int\int\cos\theta\mathscr{P}(\theta|\theta')d\theta d\varphi$, the average of the azimuthal scatter angle θ (the polar angle is φ). In the standard diffusion approximation,[1] $g=0$ or g is small (i.e., $g>0$ but $g\ll 1$) and $\mathscr{P}=1$. Via the similarity relation,[1,4,17] an anisotropic scattering medium with nonzero anisotropy factor $g\in[0,1]$ and scatter coefficient μ_s', is equivalent to an isotropic scattering medium with $g=0$ and the scatter coefficient μ_s such that $\mu_s=\mu_s'(1-g)$. The diffusion coefficient becomes $(3/2\mathscr{U})c=[\mu_a+\mu_s(1-g)]$. Therefore, increasing $g\to 1$ for an anisotropic medium, has the same effect as reducing μ_s to zero for an isotropic medium. The distribution of scattered light paths between two points $(\mathbf{x},\mathbf{x}')$ becomes more ellipsoidal and collapses back to the Fermat straight line connecting the points, as $g\to 1$. This is again described by Eq. (36).

B. Large scatter limit: $n\gg 0$ or $\mu_s\gg\mu_a$

When $n\to\infty$ (or becomes large) one can perform the path integral using Gaussian integrals as is done in ordinary quantum mechanics.[16] We have $\int\Pi_n d^3x\to\int\mathscr{D}x$ for large n, where $\mathscr{D}x$ is the path measure. Very few functional path integrals can be performed, but free linear systems gen-

erally have a Gaussian representation. The solution is the well known homogeneous, heat kernal form which describes free linear diffusion of light out from a point source or impulse,

$$G(\mathbf{x},t;\mathbf{x}',0)=(4\pi\mathscr{U}t)^{-1/2}\exp[-(\mathbf{x}-\mathbf{x}')^2(4\mathscr{U}t)^{-1}], \tag{38}$$

which automatically satisfies the linear, optical diffusion equation (29) with $\langle(\mathbf{x}-\mathbf{x}')(\mathbf{x}-\mathbf{x}')\rangle = \mathscr{U}_2 t \equiv 2\mathscr{U}t$. The transformations $it\leftrightarrow t$ and $\mathscr{U}\leftrightarrow 1/2m$ give the free particle wave function of quantum theory which has the same Gaussian form. As a final check on the consistency of the description, we substitute the Markovian kernel (38) into Eq. (6) at second order ($\beta=2$), and integrate over the semi-infinite region $[0,\infty)$. We should then find that $\mathscr{U}_2(\mathbf{x},t)=\mathscr{U}_2=\mathscr{U}$ for all $\mathbf{x}\in\mathscr{M}$ and $t\in\mathbb{R}$,

$$\begin{aligned}\mathscr{U}_2(\mathbf{x}',t)&=\frac{1}{2t}\int G(\mathbf{x},t+\delta t;\mathbf{x}',t)(\mathbf{x}-\mathbf{x}')^2 d\mathbf{x}\\ &=\frac{1}{t}\frac{1}{(4\pi\mathscr{U}t)^{1/2}}\int_0^\infty(\mathbf{x}-\mathbf{x}')^2\exp[-(4\mathscr{U}t)^{-1}(\mathbf{x}-\mathbf{x}')^2]d\mathbf{x},\end{aligned} \tag{39}$$

where we have set $t=0$. The right-hand side is simply a standard Gaussian integral,

$$\int_0^\infty v^2\exp[-av^2]dv=\frac{1}{4}\left(\frac{\pi}{a^3}\right)^{1/2} \tag{40}$$

with $a=(4\mathscr{U}t)^{-1}$ and $v=(\mathbf{x}-\mathbf{x}')$ and $dv=d\mathbf{x}$. Hence,

$$\mathscr{U}_2(\mathbf{x}',t)=\left(\frac{1}{(t^{3/2})(4)^{1/2}(\pi)^{1/2}(\mathscr{U})^{1/2}}\right)\left(\frac{(\pi)^{1/2}(4)^{3/2}(\mathscr{U})^{3/2}t^{3/2}}{4}\right)=\mathscr{U} \tag{41}$$

as required.

IV. CONCLUDING REMARKS

Beginning with the transfer equation, we have presented a simple stochastic derivation and interpretation of the scalar optical diffusion equation and a path integral representation of the Green's function or propagator of radiative transfer. Although optical path integral representations for radiative transfer have been presented heuristically in the literature, this approach and interpretation has not previously been considered, and justifies their use from a more rigorous and fundamental viewpoint. A direct duality can also be made with classical radiative transfer in real time and ordinary quantum theory in Euclidean time.

ACKNOWLEDGMENTS

I would like to thank Professor Britton Chance and David Miller.

[1] I. Ishimaru, *Wave Propagation in Random Medias* (Academic, New York, 1978), Vols. I and II.
[2] L. O Reynolds, C. Johnson, and I. Ishimaru, J. Opt. Soc. Am. A **5**, 813 (1988).
[3] S. Chandrasekhar, *Radiative Transfer Theory* (Dover, New York, 1960).
[4] G. I. Bell and G. Glasston, *Nuclear Reactor Theory* (Van Nostrand Reinhold, Princeton, 1970).
[5] V. Twersky, J. Opt. Soc. Am. **60**, 908 (1970); **60**, 1084 (1970).
[6] *Stochastic Quantization*, edited by P. H. Damgaard and H. Huffel (World Scientific, Singapore, 1987).
[7] J. Tessendorf, Phys. Rev. A **35**, 872 (1986).
[8] J. Tessendorf, J. Opt. Soc. Am. A **5**, 1410 (1988).
[9] J. C. Schotland, J. C. Haselgrove, and J. S. Leigh, J. Appl. Opt. **32**, 448 (1993).
[10] L. Perelman, J. Wu, I. Itzka, and M. S. Feld, Phys. Rev. Lett. **72**, 1341 (1994).
[11] L. Perelman and M. S. Feld, Phys. Rev. E **51**, 6134 (1995).
[12] J. Medhi, *Stochastic Processes*, 2nd ed. (Wiley, New York, 1981).
[13] M. Lax, Rev. Mod. Phys. **38**, 359 (1966).
[14] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics* (Charles Griffen, London, 1958), Chap. 3.
[15] R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).
[16] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw–Hill, New York, 1965).
[17] S. A. Prahl, Ph.D. thesis, University of Texas at Austin, 1988.
[18] D. A. Boas, M. A. O'Leary, B. Chance, and A. G. Yodh, Proc. Natl. Acad. Sci. USA **91**, 4887 (1994).

The monotony criterion for a finite size scaling analysis of phase transitions

Hildegard Meyer-Ortmanns[a)]
Institut für Theoretische Physik, Bergische Universität Wuppertal, Gauss strasse 20, D-42097 Wuppertal, Germany

Thomas Reisz[b)]
Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany

(Received 25 February 1998; accepted for publication 25 May 1998)

We propose a new criterion to analyze the order of phase transitions within a finite size scaling analysis. It refers to response functions like order parameter susceptibilities and the specific heat and states different monotony behavior in volume for first- and second-order transitions close to the transition point. The criterion applies to analytical and numerical studies of phase diagrams including tricritical behavior.
© *1998 American Institute of Physics.* [S0022-2488(98)00310-7]

I. INTRODUCTION

Finite size scaling analysis is particularly useful in numerical simulations to distinguish first- from second-order phase transitions by their different onset in a large but finite volume. Usually it is the specific scaling behavior of the peak in a susceptibility, the shift in the critical coupling/temperature, and the width of the critical region that anticipates the nature of the transition in the thermodynamic limit.[1–4] In this letter we formulate a new criterion in a finite volume to distinguish first- from second-order phase transitions. For a certain interval of the scaling region response functions with a nonanalytic behavior in the infinite volume limit show different monotony behavior for first- and second-order transitions. Examples for such functions are the specific heat and order parameter susceptibilities. They are increasing in volume in a certain neighborhood of T_c for second-order transitions, and decreasing for first-order transitions for some range in the scaling region, which will be specified below.

For definiteness we fix the notation in terms of an order parameter susceptibility $\chi(T,L)$, considered as a function of the temperature T and the spatial size parameter L, so that a symmetric D-dimensional volume is of size L^D. By $T_c(L)$ we denote the location of the maximum of $\chi(\cdot,L)$ in volume L^D associated with the phase transition at $T_c \equiv T_c(\infty)$. In the infinite volume

$$\chi(t+T_c(\infty),L=\infty)<\infty \quad \text{as} \quad t\to 0 \tag{1}$$

for a first-order transition at T_c, with a possible discontinuity at $t=0$ in the associated order parameter, with t denoting the scaling field $t=(T-T_c(L))/T_c(L)$, whereas

$$\chi(t+T_c(\infty),L=\infty)\simeq \mathscr{A}|t|^{-\gamma} \quad \text{as} \quad t\to 0 \tag{2}$$

for a second-order transition with critical exponent $\gamma>0$. On the other hand, in the opposite order of limits, $\chi(T_c(L),L)$ diverges in both cases as L approaches infinity. Stated differently, at $T_c(L)$, in the limit $L\to\infty$ χ has a δ-function singularity or a power law type of singularity for a first- or second-order transition, respectively. It is this difference that is responsible for the different monotony properties for $t\neq 0$ in the finite volume.

a)Electronic mail: ortmanns@theorie.physik.uni-wuppertal.de
b)Supported by a Heisenberg Fellowship. Electronic mail: reisz@thphys.uni-heidelberg.de

 © 1998 American Institute of Physics

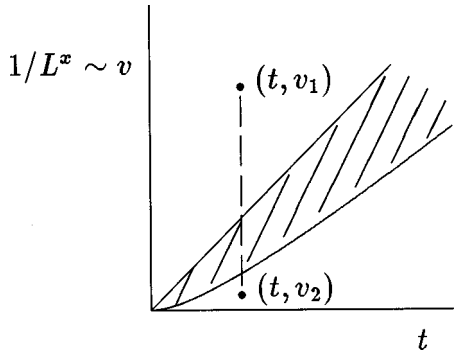

FIG. 1. (t,v) plane in the vicinity of a phase transition at $t=v=0$. t denotes the scaling field $t=(T-T_c(L))/T_c(L)$, v is inverse to some power of the volume L^x with $x>0$. Let us consider the susceptibility as a function of t and v. Then, for a first-order transition, $\chi(t,v_1)>\chi(t,v_2)$, whereas for a second-order transition $\chi(t,v_1)<\chi(t,v_2)$. The shaded part is the crossover region between different asymptotic behavior as $(t,v)\to 0$.

The essential statement of the monotony criterion is the following (details will be given below): For a sufficiently large size $L_s<\infty$, there is always a size L_l with $L_l>L_s$ and such that (cf. Fig. 1)

$$\chi(\delta+T_c(L),L\geqslant L_l)<\chi(\delta+T_c(L_s),L_s) \quad \text{for first order,} \tag{3}$$

$$\chi(\delta+T_c(L),L\geqslant L_l)>\chi(\delta+T_c(L_s),L_s) \quad \text{for second order,} \tag{4}$$

that is, $\chi(\delta+T_c(L),L)$ is decreasing or increasing in volume for a first- or second-order phase transition, respectively. The susceptibility is measured at fixed distance δ to the volume dependent maximum, constrained by

$$c_1\sigma(L)^{1/(1+\epsilon)}<|\delta|<c_2\sigma(L_s). \tag{5}$$

Here $\sigma(L)=L^{-D}$ or $\sigma(L)=L^{-1/\nu}$ for first- or second-order transition, respectively, with ν being the critical exponent of the correlation length, $\xi\sim|T-T_c|^{-\nu}$ as $T\to T_c$. Hence $\sigma(L)$ is the width of the critical region in the volume L^D, c_1, c_2, and ϵ are positive constants, typically $\epsilon=1$. Beyond the general constraint that both linear sizes L_s and L_l have to be sufficiently large, in addition L_l has to be sufficiently larger than L_s, so that $\sigma(L_l)$ is considerably smaller than $\sigma(L_s)$. The monotony behavior of χ does *not* refer to values of L_l close to L_s. In terms of Fig. 1, L_s and L_l are separated by the shaded area, for which we do not make any predictions on the behavior of χ.

In Sec. II we make these statements more precise. We specify the conditions under which the criterion holds and give the proof. As it turns out, the criterion applies whenever the standard conjectures on finite size scaling behavior hold. Rigorous proofs of these standard conjectures are missing in general, and we do not attempt to give such proofs in the treatment below. In Sec. III we conclude with some remarks on testing the criterion in analytical and numerical calculations.

II. STATEMENTS OF THE CRITERION

Let t denote the scaling field, i.e., $t=(T-T_c(L))/T_c(L)$. $T_c(L)$ locates the maximum of the susceptibility in the volume L^D. Furthermore we set $v=L^{-m}$ with some $m>0$. The infinite volume limit is obtained as $v\to 0$ from above.

The transition region is given by small t and v. Let

$$H^2:=\{(t,v)\in\mathbf{R}^2|v\geqslant 0\} \tag{6}$$

denote a half plane, $\mathcal{U}\subseteq H^2$ the intersection of an open neighborhood of $0\in\mathbf{R}^2$ with H^2 and $\mathcal{U}^*=\mathcal{U}\setminus\{0\}$.

A. First-order transitions

The typical behavior of a susceptibility close to a first-order transition is summarized by the following definition (with $v=L^{-D}$). As usual, $C(\mathscr{U}^*)$ denotes the set of real valued continuous functions in $\mathscr{U}^*$.

Definition 1: Let $\omega>0$. $\Psi_1^\omega(\mathscr{U})$ *is the set of functions* $\chi\in C(\mathscr{U}^*)$ *with the following properties.*

(1) $\chi\in C^1(\mathscr{U}^*\setminus\{(\mathbf{R},0)\})$, *that is,* χ *is continuously differentiable for* $v\neq 0$.

(2) There exists $\nu(t)>0$ *such that for all* $v<|t|/\nu(t)$

$$|\chi(t,v)|\leqslant\omega.$$

(3) With appropriate positive constants c, K_1, K_2 *and* ϵ *we have in* $\mathscr{U}^*$ *for* $v\neq 0$

$$\left|\chi(0,v)-\frac{c}{v}\right|<\frac{K_1}{v^{1-\epsilon}},$$

$$\left|\frac{\partial}{\partial t}\,\chi(t,v)\right|<\frac{K_2}{v^2}.$$

It follows already from this definition that $\nu(t)$ is bounded from below by some positive constant. This is part of the following proposition. The main statement is the monotony behavior of a function that obeys Definition 1.

Proposition 1: Let $\omega>0$, $\chi\in\Psi_1^\omega(\mathscr{U})$. *There exist positive* δ, ϵ *such that the following statements hold.*

(1) For all t *with* $|t|<\epsilon\delta$,

$$\nu(t)>\epsilon. \tag{7}$$

(2) For all (t,v), $(t,w)\in\mathscr{U}^*$ *with* $\nu(t)w<|t|<\epsilon v$ *and* $v<\delta$, *we have* $w<v$ *and*

$$\chi(t,v)>|\chi(t,w)|. \tag{8}$$

The proposition includes the case $w=0$ of infinite volume. In this case, in the critical region $|t|/v<\epsilon$, the susceptibility χ is always larger in the finite volume than in the infinite one.

Proof: Let $\chi\in\Psi_1^\omega(\mathscr{U})$, $\omega>0$.

(α) There are ϵ, $\delta>0$ such that for all t, v with $v<\delta$, $|t|<\epsilon v$

$$\chi(t,v)\geqslant 2\omega. \tag{9}$$

For

$$\chi(t,v)=\chi(0,v)+t\int_0^1 ds\,\frac{\partial}{\partial\eta}\,\chi(\eta,v)\bigg|_{\eta=st},$$

and there are c_0, $K>0$ so that

$$\chi(0,v)>\frac{c_0}{v},$$

$$\left|\frac{\partial}{\partial\eta}\,\chi(\eta,v)\right|<\frac{K}{v^2}$$

in $\mathscr{U}^*$. Hence, with $\epsilon=c_0/(2K)$,

$$\chi(t,v) \geqslant \frac{c_0}{v} - \frac{|t|K}{v^2} > \frac{c_0 - \epsilon K}{v} = \frac{c_0}{2v}.$$

Choosing $\delta = c_0/(4\omega)$, we get for $v < \delta$ the lower bound (9) on $\chi(t,v)$.

(β) By definition of $\nu(t)$, we have whenever $\nu(t)w < |t|$ the upper bound

$$|\chi(t,w)| \leqslant \omega. \tag{10}$$

(γ) Summarizing α and β, it follows that for all w, v with $\nu(t)w < |t| < \epsilon v$, $v < \delta$, the strict bound

$$|\chi(t,w)| < \chi(t,v)$$

holds. Furthermore, (10) implies that $|\chi(t,w')| \leqslant \omega$ for all $0 \leqslant w' \leqslant w$. Hence it must hold that $w < v$. It follows that

$$\frac{|t|}{\nu(t)} \leqslant \frac{|t|}{\epsilon}$$

and hence $\nu(t) > \epsilon$ for $|t| < \epsilon\delta$. □

The remainder of this section is devoted to the function class $\Psi_1^{\omega}(\mathscr{U})$. As an example of such a function we consider the following typical representation of the magnetic susceptibility in the volume L^D close to a first-order phase transition:

$$\chi_2(T,L) = c_2 L^D \exp(-f_2 L^{2D}(T - T_c(L))^2) + \eta_2(T,L). \tag{11}$$

Here c_2, f_2 are positive real constants and $\eta_2(\cdot,L)$ is analytic for $L < \infty$, locally uniformly convergent as $L \nearrow \infty$ [so that $\eta_2(\cdot,\infty)$ is analytic]. Furthermore, $T_c(L) = T_c - d_2 L^{-m} + o(L^{-m})$, with $m \geqslant 0$.

We can view χ_2 as a function of some Ψ_1^{ω} in essentially two ways. The first one corresponds to the point of view given in Sec. I, in which the scaling field t measures the distance of the temperature T to the volume dependent location of the peak $T_c(L)$. It does not refer to a particular volume dependence of $T_c(L)$ itself. That is, m does not need further specification except that $m \geqslant 0$.

We set $v = L^{-D}$, $t = (T - T_c(L))/T_c(L)$, and

$$\chi(t,v) = \chi_2(T,L) = c v^{-1} \exp(-f(v) v^{-2} t^2) + \eta(t,v), \tag{12}$$

where $\eta(t,v) \equiv \eta_2(T,L)$ and $f(v) = f_2 T_c(L)^2$, $c = c_2$.

Lemma 1: $\chi \in \Psi_1^{\omega}(\mathscr{U})$ *for some* $\omega > 0$ *and sufficiently small* $\mathscr{U}$. $\nu(t)$ of *Definition 1 satisfies* $\nu(t)^{-1} \to 0$ *as* $|t| \to 0$.

Proof: For α, $\beta > 0$ and $x \in \mathbf{R}$ the function

$$f_{\alpha,\beta}(x) = x^{\alpha} \exp(-\beta x^2)$$

attains its maximum at $x^2 = \alpha/(2\beta)$, is positive and monotone decreasing for $x > (\alpha/(2\beta))^{1/2}$, and satisfies on $\mathbf{R}$ the bound

$$|f_{\alpha,\beta}(x)| \leqslant \left(\frac{\alpha}{2\beta e}\right)^{\alpha/2}.$$

We prove that $\chi(t,v)$ match the criteria of Definition 1 to belong to some $\Psi_1^{\omega}(\mathscr{U})$. There is $\delta > 0$ such that for all $v < \delta$ we have $\hat{f} \leqslant f(v) \leqslant \bar{f}$ for some $\hat{f}, \bar{f} > 0$. We choose $\mathscr{U}$ such that $(t,v) \in \mathscr{U}$ implies that $v < \delta$ and $|t| < \hat{f}^{1/2}$ and put

$$K_{\eta} = \max_{(t,v) \in \mathscr{U}} (|\eta(t,v)|, |\partial_t \eta(t,v)|).$$

Condition (1) is obviously satisfied. To show condition (2) we notice that for

$$\frac{|t|}{v} > \nu(t) > \left(\frac{1}{2\hat{f}}\right)^{1/2}$$

one obtains

$$|\chi(t,v)| < \left|\frac{c}{|t|}\frac{|t|}{v}\exp\left(-f(v)\left(\frac{|t|}{v}\right)^2\right)+\eta\right| \leqslant c\,\frac{\nu(t)}{|t|}\exp(-f(v)\nu(t)^2)+|\eta|.$$

We choose $\nu(t)=|t|^{-1}$ and obtain with $\omega = c/(\hat{f}e)+K_\eta$,

$$|\chi(t,v)| \leqslant \omega.$$

To verify the third condition, we compute

$$|\chi(0,v)-cv^{-1}| = |\eta(t,v)| \leqslant K_\eta,$$

and notice that there is $K>0$ such that

$$\left|\frac{\partial}{\partial t}\chi(t,v)\right| \leqslant \frac{2|t|}{v^2}f|\chi| + |\partial_t\eta(t,v)| \leqslant \frac{2\overline{f}}{v^2}\left[c\left(\frac{1}{2\hat{f}e}\right)^{1/2}+\hat{f}^{1/2}K_\eta\right]+K_\eta \leqslant \frac{K}{v^2}.$$

Hence, $\chi \in \Psi_1^\omega(\mathscr{U})$. □

The next lemma states that the susceptibility χ_2 matches the criteria of Definition 1 also if the scaling field t measures the distance of T from the phase transition point, i.e., to $T_c(L=\infty)$. The only additional constraint then is that the peak $T_c(L)$ of χ_2 converges sufficiently fast to $T_c(\infty)$, namely that

$$|T_c(L)-T_c(\infty)| \leqslant \frac{c}{L^m} \quad \text{with} \quad m \geqslant D. \tag{13}$$

We set $v=L^{-D}$, $t=(T-T_c(\infty))/T_c(\infty)$, and

$$\chi(t,v)=\chi_2(T,L)=cv^{-1}\exp(-fv^{-2}(t+dv^n)^2)+\eta(t,v), \tag{14}$$

where now $\eta(t,v)\equiv\eta_2(T,L)$ and $f=f_2T_c(\infty)^2$, $d=d_2/T_c(\infty)$, and $n=m/D$. Similar to the above, we get

Lemma 2: Let $n \geqslant 1$. *Then* $\chi \in \Psi_1^\omega(\mathscr{U})$ *for some* $\omega>0$ *and sufficiently small* $\mathscr{U}$. $\nu(t)$ *of Definition 1 satisfies* $\nu(t)^{-1}\to 0$ *as* $|t|\to 0$.

Proof: For a, b, $\epsilon \in \mathbf{R}$, $\epsilon \neq 0$,

$$(a+b)^2 \geqslant a^2\left(1-\frac{1}{\epsilon^2}\right)+b^2(1-\epsilon^2).$$

We choose δ, $\eta>0$ sufficiently small such that the following estimates on χ are true for all t, v with $|t|<\eta$, $0<v<\delta$, identifying $\mathscr{U}$.

χ obviously satisfies condition (1) of Definition 1. To show condition (2) we estimate

$$|\chi(t,v)| \leqslant cv^{-1}\exp(-fv^{-2}(\tfrac{1}{2}t^2-(dv^n)^2))+|\eta| \leqslant 2cv^{-1}\exp\left(-\frac{f}{2}v^{-2}t^2\right)+|\eta|. \tag{15}$$

Proceeding as in the proof of the previous lemma we obtain condition (2), with appropriate ω. To realize the third condition, we first notice that

$$|\chi(0,v)-cv^{-1}| \leqslant cv^{-1}|e^{-fd^2v^{2(n-1)}}-1|+|\eta| \leqslant \frac{K}{v^{1-\epsilon}}$$

for some positive constant K, with $\epsilon=\min(1,2(n-1)) \geqslant 0$. Finally there is $K'>0$ such that

$$\left|\frac{\partial}{\partial t}\chi(t,v)\right| = 2fv^{-2}|t+dv^n||\chi| + |\partial_t\eta(t,v)| \leqslant 2fcv^{-2}\left(\frac{1}{2fe}\right)^{1/2}\cdot 2 = \frac{K'}{v^2}. \tag{16}$$

Hence $\chi \in \Psi_1^{\omega}(\mathscr{U})$. □

Finally, we state without proof that every function $\chi: \mathscr{U}^* \to \mathbf{R}$ of the form

$$\chi(t,v) = \frac{1}{v} f\left(\frac{t}{v}\right) + \tilde{\chi}(t,v) \tag{17}$$

belongs to $\Psi_1^{\omega}(\mathscr{U})$ for some $\mathscr{U}$ and with appropriate $\omega > 0$, if the following conditions are satisfied.

(1a) $\tilde{\chi}(t,v) \in C^1(\mathscr{U}^*)$.

(1b) $\tilde{\chi}(t,v)$ together with its (first) partial derivatives are uniformly bounded in $\mathscr{U}^*$.

(2a) $f \in C^1(\mathbf{R})$ is a non-negative function with $f(0) > 0$.

(2b) $\lim_{x \to \pm\infty} |x|^{1+\epsilon} f(x) = 0$ for some $\epsilon > 0$.

(2c) $(d/dx)f(x)$ is uniformly bounded on $\mathbf{R}$.

Any such function has the property that it ''approaches δ'' locally, i.e., there is $c > 0$ so that for any test function $g(t)$ supported in a small neighborhood of $t = 0$,

$$\lim_{v \to 0+} [\chi(t,v)]g = \lim_{v \to 0+} \int dt\ \chi(t,v)g(t) = cg(0).$$

The example discussed above corresponds to

$$g(x) = \left(\frac{c}{\pi}\right)^{1/2} \exp(-cx^2), \quad c > 0.$$

B. Second-order transitions

In contrast to the first-order case, at a second-order transition the susceptibility χ can be divergent in the infinite volume limit as the critical temperature is approached. (It need not be divergent, since another second derivative of a suitable thermodynamic potential may be divergent instead.) If χ is divergent with a critical exponent, it is described by the following definition (with $v = L^{-1/\nu}$, $\nu > 0$ being the critical exponent of the correlation length ξ).

Definition 2: Let $\gamma > 0$. $\Psi_2^{\gamma}(\mathscr{U})$ is the set of functions $\chi \in C(\mathscr{U}^*)$ that satisfy the following conditions.

(1) There are constants $\mathscr{A}$, K, $\epsilon > 0$ such that in $\mathscr{U}^*$,

$$|\chi(t,0) - \mathscr{A}|t|^{-\gamma}| \leqslant K|t|^{-\gamma+\epsilon}.$$

Furthermore, there exist ν, $\mathscr{C} > 0$ so that for $|t| > \nu v$,

$$\chi(t,v) \geqslant \mathscr{C}\chi(t,0).$$

(2) The are constants η, $\mathscr{B} > 0$ such that for $|t| < \eta v$,

$$|\chi(t,v)| < \mathscr{B}v^{-\gamma}.$$

Functions that obey Definition 2 are always increasing in volume close to the transition.

Proposition 2: Let $\gamma > 0$ and $\chi \in \Psi_2^{\gamma}(\mathscr{U})$. There are constants ν, $\epsilon > 0$ such that for all t, v, w with $\nu w < |t| < \epsilon v$ we have $w < v$ and

$$|\chi(t,v)| < \chi(t,w).$$

The inequality includes the case $w = 0$, corresponding to infinite volume. That is, the susceptibility is always smaller in finite volume than in the infinite volume as long as we are in the critical region $|t|/v < \epsilon$.

Proof: Let $\chi \in \Psi_2^{\gamma}(\mathscr{U})$, $\gamma > 0$. There are numbers $\mathscr{C}$, $\mathscr{D}$, $\nu > 0$ such that in $\mathscr{U}^*$ for $|t| > \nu w$,

$$\chi(t,w) > \mathscr{C}\chi(t,0) > \mathscr{D}|t|^{-\gamma}.$$

Furthermore, there are $\eta, \mathscr{B} > 0$ such that for $|t| < \eta v$,

$$|\chi(t,v)| < \mathscr{B}v^{-\gamma}.$$

We choose $\epsilon = \min(\eta, (\mathscr{D}/\mathscr{B})^{1/\gamma})$ and get for $\nu w < |t| < \epsilon v$,

$$|\chi(t,v)| < \mathscr{B}v^{-\gamma} < \mathscr{B}\left(\frac{\epsilon}{|t|}\right)^{\gamma} < \frac{\mathscr{B}\epsilon^{\gamma}}{\mathscr{D}}\,\chi(t,w) < \chi(t,w).$$

This proves the lemma. □

The specific property for a second-order transition that the singular part of the free energy density behaves as a generalized homogeneous function implies for the susceptibility in a volume L^D a typical form like

$$\chi_2(T,L) = |T-T_c|^{-\gamma} Q((T-T_c)L^{1/\nu}) + \eta(T,L), \tag{18}$$

with some $\gamma > 0$. Here $\eta(\cdot, L)$ has similar analyticity properties as in (11) above, $\nu > 0$ is the critical exponent of the correlation length $\xi \sim |T-T_c|^{-\nu}$. Q is continuous and behaves as

$$\lim_{x\to 0} |x|^{-\gamma} Q(x) = K > 0,$$

$$\lim_{x\to \pm\infty} Q(x) = C > 0.$$

The first equation expresses the absence of a divergence of χ_2 for finite L, the second one its presence in the infinite volume case. With $t = (T-T_c)/T_c$, $v = L^{-1/\nu}$, and

$$\chi(t,v) := \chi_2(T,L)$$

it is straightforward to verify that χ belongs to $\Psi_2^{\gamma}(\mathscr{C})$.

More generally, every function χ: $\mathscr{C}^* \to \mathbf{R}$ of the form

$$\chi(t,v) = \frac{1}{v^{\gamma}} f\left(\frac{t}{v}\right) + \tilde{\chi}(t,v) \tag{19}$$

with $\gamma > 0$ belongs to some $\Psi_2^{\gamma}(\mathscr{C})$, if the following conditions are satisfied:
(1) $\tilde{\chi}(t,v) \in C^1(\mathscr{C}^*)$.
(2a) $f \in C^1(\mathbf{R})$, and $f(0) > 0$,
(2b) $\lim_{x\to\pm\infty} |x|^{\gamma} f(x) = C$ for some finite $C > 0$.

Compared to the first-order case (17) the essential difference comes from property (2b).

As an example,

$$\chi(t,v) = (t^2+v^2)^{-(m/2)}, \quad m > 0,$$

belongs to the class Ψ_2^m.

III. APPLICATIONS

The monotony criterion has been applied in the framework of convergent series expansions of $O(N)$ symmetric scalar models on the lattice, with ϕ^4 and ϕ^6 interactions, performed to the twentieth order in the hopping parameter (corresponding to the inverse temperature) both in the infinite and finite volume.[5] This class of models admits both first- and second-order transition regions. By means of the series representation it is hard to compute the scaling behavior of a susceptibility $\chi(T,L)$ at the maximum $T_c(L)$, since $T_c(L)$ as well as the transition at $T_c(\infty)$ are extrapolated as the convergence radius of the series. In contrast to other criteria that have been proposed so far (cf. Refs. 1–3), the monotony criterion circumvents this difficulty in that it allows

us to determine the order of the transition by the different L-dependence of $\chi(T,L)$ already close to, but not at $T_c(L)$, that is by an increase in volume for a second-order transition, and a decrease for a first-order transition. This feature is always needed in cases where it is not possible to work directly at the pseudocritical temperature $T_c(L)$.

In series representations the larger volume L_l^D is conveniently chosen to be infinite so that (3)–(5), with $L=\infty$ and $\sigma(L)=0$, give strong criteria in the whole scaling region. In Monte Carlo simulations both volumes L_s^D and L_l^D must be finite. The monotony criterion works equally well in this case. Both volumes have to be sufficiently large in order to match the conditions of the criterion (which reflect the standard assumptions of finite size scaling). Some care is needed to ensure that the susceptibility is actually measured at a temperature that satisfies (5), i.e., out of a region where the difference in the volume dependence between first- and second-order transitions appears. In particular this concerns the lower bound on δ in (5) because for $L_l<\infty$, also in case of a first-order transition, a small neighborhood around the peak at $T_c(L)$ exists, where $\chi(T,L)$ is increasing in volume because of the rounding of the δ-singularity.

To summarize, let λ denote a generic coupling constant. Application of the monotony criterion amounts to a choice of two sizes L_s and L_l with L_s, L_l, and L_l/L_s sufficiently large. Let us define the ratio r,

$$r(L_s,L_l;\lambda) := 1-\frac{\chi(T_c(L_s)+\delta,L_s;\lambda)}{\chi(T_c(L_l)+\delta,L_l;\lambda)}, \tag{20}$$

with δ satisfying the constraint (5). The monotony criterion then says that

$$r(L_s,L_l;\lambda)\begin{cases} >0, & \text{second order} \\ <0, & \text{first order} \\ =0, & \text{tricritical point for } r \text{ changing sign at } \lambda. \end{cases}$$

In (20) the susceptibility is measured at fixed distance to the volume-dependent location of its maximum. A reasonable alternative is to measure $\chi(T,L)$ for both volumes at the same temperature T regardless of L, but T chosen as $T=T_c(\infty)+\delta$. This requires that $T_c(L)$ approaches $T_c=T_c(\infty)$ according to

$$|T_c(L)-T_c(\infty)| \leqslant c\,\sigma(L), \tag{21}$$

cf. (5) and (13). The behavior (21) meets the standard finite size scaling behavior.

Other applications of the monotony criterion so far have been the three-dimensional Gross-Neveu model[6] and the electroweak phase transition of the early universe.[7]

[1] M. N. Barber, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and J. L. Lebowitz (Academic, New York), Vol. 8, p. 145.
[2] K. Binder and D. P. Landau, Phys. Rev. B **30**, 1477 (1984); K. Binder, Rep. Prog. Phys. **50**, 783 (1987).
[3] C. Borgs and J. Z. Imbrie, Commun. Math. Phys. **123**, 305 (1989); C. Borgs and R. Kotecky, J. Stat. Phys. **61**, 79 (1990); Phys. Rev. Lett. **68**, 1734 (1992); C. Borgs, R. Kotecky, and S. Miracle-Sole, J. Stat. Phys. **62**, 529 (1991).
[4] P. M. C. Oliveira, S. M. Moss de Oliveira, C. E. Cordeiro, and D. Stauffer, e-print archive cond-mat/9503102.
[5] H. Meyer-Ortmanns and T. Reisz, J. Stat. Phys. **87**, 755 (1997).
[6] T. Reisz, in *Field Theoretical Tools for Polymer and Particle Physics*, edited by H. Meyer-Ortmanns and A. Klümper (Springer, Heidelberg, 1998) Lecture Notes in Physics Vol. 508, p. 192; e-print archive hep-lat/9712017.
[7] H. Meyer-Ortmanns, K. Pinn, and T. Reisz (unpublished).

Continuous iteration of dynamical maps

R. Aldrovandi[a] and L. P. Freitas[b]
Instituto de Física Teórica, Universidade Estadual Paulista,
Rua Pamplona, 145, 01405-900 São Paulo SP, Brazil

(Received 10 November 1997; accepted for publication 16 June 1998)

A precise meaning is given to the notion of continuous iteration of a mapping. Usual discrete iterations are extended into a dynamical flow which is a homotopy of them all. The continuous iterate reveals that a dynamic map is formed by independent component modes evolving without interference with each other. An application to turbulent flow suggests that the velocity field assumes nonseparable values. © *1998 American Institute of Physics.* [S0022-2488(98)03410-0]

I. INTRODUCTION

There are two main approaches to describe the evolution of a dynamical system.[1] The first has its roots in classical mechanics—the solutions of the dynamical differential equations provide the continuous motion of the representative point in phase space.[2] The second takes a quite different point of view: it models evolution by the successive iterations of a well-chosen map, so that the state is known after each step, as if the ''time'' parameter of the system were only defined at discrete values.[3] It is possible to go from the first kind of description to the second through the snapshots leading to a Poincaré map. Our aim here is to present the first steps into a converse procedure, going from a discrete to a continuous description *while preserving the idea of iteration.* This is possible if we are able to ''interpolate'' between the discrete values in such a way that the notion of iteration keeps its meaning in the intervals.

Iteration is a particular case of function composition: given a basic map $f(x)=f^{\langle 1\rangle}(x)$, its first iterate is $f^{\langle 2\rangle}(x)=[f\circ f](x)=f[f(x)]$, its nth iterate is $f^{\langle n\rangle}(x)=f[f^{\langle n-1\rangle}(x)]=f^{\langle n-1\rangle}[f(x)]$, etc. The question is whether or not, given the set of functions $f^{\langle n\rangle}(x)$, an interpolation $f^{\langle t\rangle}(x)$ with real values of t can be found which represents the one-parameter continuous group (or semigroup) describing the dynamical flow of the system. In order to do it, $f^{\langle t\rangle}$ should satisfy the conditions

$$f^{\langle t\rangle}[f^{\langle t'\rangle}(x)]=f^{\langle t'\rangle}[f^{\langle t\rangle}(x)]=f^{\langle t+t'\rangle}(x), \tag{I.1}$$

$$f^{\langle 0\rangle}(x)=\mathrm{Id}(x)=x. \tag{I.2}$$

We shall find in what follows a map interpolation $f^{\langle t\rangle}(x)$ with these properties. It is a well-known fact that Taylor series $f(x)$, $g(x)$ satisfying the conditions $f(0)=0$, $g(0)=0$, and with nonvanishing coefficients of x are invertible and constitute a group by composition.[4] The neutral element is the identity function $e(x)=\mathrm{Id}(x)=x$ and two functions f, g are inverse to each other if it is true that $f[g(x)]=g[f(x)]=\mathrm{Id}(x)=x$. In this case, $g=f^{\langle -1\rangle}$ and $f=g^{\langle -1\rangle}$ (we are taking the liberty of using the word ''function'' even for purely formal series and multivalued maps). The clue to the question lies in the formalism of Bell polynomials, which attributes to every such function f a matrix $B[f]$, whose inverse represents the inverse function and such that the matrix product represents the composition operation. In other words, these matrices provide a linear representation of the group formed by the functions with the operation of composition. Composition is thus represented by matrix product and, consequently, iterations are represented by matrix powers. Furthermore, the representation is faithful, and the function f is completely determined by $B[f]$. Now, in the matrix group there does exist a clear interpolation of discrete powers by real powers and the inverse way, going from matrices to functions, yields a map interpolation with the desired properties.

[a]Electronic mail: RA@AXP.IFT.UNESP.BR
[b]Electronic mail: LFREITAS@IFT.UNESP.BR

 © 1998 American Institute of Physics

Section II is a short presentation of Bell polynomials, with only the minimum information necessary for our present objective. It is shown how a matrix $B[f]$ can be found which represents each formal series f, and that the composition $f\circ g$ of two functions is represented by the (right-) product of the respective matrices: $B[f\circ g]=B[g]B[f]$. The identity matrix corresponds to the identity function, $B[\mathrm{Id}]=I$, the matrix $B[f^{\langle n\rangle}]$ corresponding to the nth iterate, $f^{\langle n\rangle}$ is the nth power $B^n[f]$, and the Lagrange inverse $f^{\langle -1\rangle}$ to a series f is represented by the respective inverse matrix, $B[f^{\langle -1\rangle}]=B^{-1}[f]$. The necessity of finding $B^t[f]$ for noninteger ''t'' leads to the problem of defining functions of matrices, succinctly discussed in Sec. III. Given a matrix B, there exists a very convenient basis of projectors in terms of which any function of B is defined in a simple way. A method is given to obtain the members of this basis in a closed form, in terms of powers of B. The procedure is applied to Bell matrices in Sec. IV to obtain $B^t[f]=B[f^{\langle t\rangle}]$ for any value of t, from which the function $f^{\langle t\rangle}(x)$ can be extracted and shown to satisfy conditions (I.1) and (I.2). It turns out that, though it is quite natural to call ''time'' the continuous label t, this ''time'' is related to a certain class of flows, among all those leading to a specific Poincaré map. There is an extra bonus: The matrix decomposition in terms of projectors is reflected in a decomposition of the original map, and of its iterate, into a sum in terms of certain ''elementary functions,'' each one with an independent and well-defined time evolution. An application, relating a simplified turbulence model to nonseparable solutions of the double harmonic oscillator, is given in Sec. V.

II. BELL MATRICES

Given a formal series with vanishing constant term,

$$g(x)=\sum_{j=1}^{\infty}\frac{g_j}{j!}\,x^j, \tag{II.1}$$

its Bell polynomials $B_{nk}[g]$ are certain polynomials[5] in the Taylor coefficients g_i, defined by

$$B_{nk}(g_1,g_2,\ldots,g_{n-k+1})=\frac{1}{k!}\left\{\frac{d^n}{dt^n}\,[g(t)]^k\right\}_{t=0}. \tag{II.2}$$

Their properties are in general obtained from their appearance in the multinomial theorem, which reads

$$\frac{1}{k!}\left(\sum_{j=1}^{\infty}\frac{g_j}{j!}\,t^j\right)^k=\sum_{n=k}^{\infty}\frac{t^n}{n!}\,B_{nk}(g_1,g_2,\ldots,g_{n-k+1}). \tag{II.3}$$

Depending on the situation, one or another of the notations

$$B_{nk}[g]=B_{nk}(g_1,g_2,\ldots,g_{n-k+1})=B_{nk}\{g_j\}, \tag{II.4}$$

is more convenient. The symbol $\{g_j\}$ represents the Taylor coefficient list of g, with g_j a typical member. Some properties coming immediately from the multinomial theorem are the following:

$$B_{n1}[g]=g_n, \tag{II.5}$$

$$B_{nn}[g]=(g_1)^n, \tag{II.6}$$

$$B_{nk}[cg(t)]=c^kB_{nk}[g(t)], \tag{II.7}$$

$$B_{nk}[g(ct)]=c^nB_{nk}[g(t)], \tag{II.8}$$

where c is a constant. Given two formal Taylor series,

$$f(u)=\sum_{j=1}^{\infty}\frac{f_j}{j!}\,u^j,\quad g(t)=\sum_{j=1}^{\infty}\frac{g_j}{j!}\,t^j, \tag{II.9}$$

their composition

$$F(t)=[f\circ g](t)=f[g(t)]=\sum_{j=1}^{\infty}\frac{t^n}{n!}F_n[f;g] \tag{II.10}$$

will have the Taylor coefficients F_n given by the Faà di Bruno formula,

$$F_n[f;g]=\sum_{k=1}^{n} f_k B_{nk}\{g_j\}. \tag{II.11}$$

Other properties can be obtained from the double generating function

$$e^{ug(t)}-1=\sum_{n=1}^{\infty}\frac{t^n}{n!}\sum_{j=1}^{n}u^j B_{nj}[g]. \tag{II.12}$$

Series like (II.1) constitute a group under the composition operation $(f\circ g)(x)=f[g(x)]$. The identity series ''e'' such that $e(x)=\mathrm{Id}(x)=x$ plays the role of the neutral element and each series g possesses an inverse $g^{\langle -1\rangle}$, satisfying $g^{\langle -1\rangle}\circ g=g\circ g^{\langle -1\rangle}=e$ and given by the Lagrange inversion formula. The simplest example of a Bell matrix and its inverse is given by the well-known case of the Stirling numbers: matrices formed by the first and second kind of Stirling numbers are inverse to each other, because they correspond to functions inverse to each other. In effect, consider the series

$$g(x)=\ln(1+x)=\sum_{j=1}^{\infty}\frac{(-)^{j-1}}{j}x^j, \tag{II.13}$$

whose inverse is

$$f(u)=e^u-1=\sum_{j=1}^{\infty}\frac{1}{j!}u^j. \tag{II.14}$$

A generating function for the Stirling numbers of the first kind $s_k{}^{(j)}$ is

$$\frac{1}{k!}(\ln(1+x))^k=\sum_{n=k}^{\infty}\frac{x^n}{n!}s_n{}^{(k)}. \tag{II.15}$$

It follows from (II.3) that

$$B_{nk}[\ln(1+x)]=B_{nk}(0!,-1!,2!,-3!,\ldots)=B_{nk}\{(-)^{j-1}(j-1)!\}=s_n{}^{(k)}. \tag{II.16}$$

For the Stirling numbers of the second kind $S_k{}^{(j)}$, the generating function is

$$\frac{1}{k!}(e^u-1)^k=\sum_{n=k}^{\infty}\frac{u^n}{n!}S_n{}^{(k)}, \tag{II.17}$$

from which

$$B_{nk}[e^u-1]=B_{nk}(1,1,1,\ldots,1)=B_{nk}\{1\}=S_n{}^{(k)}. \tag{II.18}$$

The inverse property of Stirling numbers is $\Sigma_{k=j}^{n}\, s_n{}^{(k)}\, S_k{}^{(j)}=\delta_n{}^j$, the same as

$$\sum_{k=j}^{n} B_{nk}[\ln(1+x)]B_{kj}[e^u-1]=\delta_{nj}. \tag{II.19}$$

The polynomials $B_{nk}[g]$ are the entries of a (lower-)triangular matrix $B[g]$, with n as row index and k as the column index. From (II.5), the function coefficients constitute the first column,

so that actually $B[g]$ is an overcomplete representative of g. From (II.6), the eigenvalues of $B[g]$ are $(g_1)^j$. Triangular matrices form a group, of which the set of matrices (B_{nk}) constitutes a subgroup. Hereby comes the most fascinating property of Bell polynomials: The matrices $B[g]=(B_{nk}[g])$, with the operation of matrix product, provide a representation of the series composition group:

$$B[g]B[f]=B[f\circ g]. \tag{II.20}$$

It is in reality an antirepresentation because of the inverse order, but this does not represent any problem. This property comes easily by using twice (II.3), as

$$\begin{aligned}\frac{1}{k!}\,(f[g(t)])^k&=\frac{1}{k!}\left(\sum_{j=1}^{\infty}\frac{f_j}{j!}\,g^j\right)^k\\&=\sum_{n=k}^{\infty}\frac{g(t)^n}{n!}\,B_{nk}[f]\\&=\sum_{n=k}^{\infty}\sum_{j=n}^{\infty}\frac{t^j}{j!}\,B_{jn}[g]B_{nk}[f]\\&=\sum_{j=k}^{\infty}\frac{t^j}{j!}\sum_{n=k}^{j}B_{jn}[g]B_{nk}[f],\end{aligned}$$

from which

$$B_{jk}[f(g(t))]=\sum_{n=k}^{j}B_{jn}[g]B_{nk}[f], \tag{II.21}$$

which is just (II.20). Associativity can be easily checked, and it is trivial to see that the ''identity'' series $e(x)=x$ has the representative $B_{nk}[e]=\delta_{nk}$, so that $B[e]=I$. The series $g(t)$ with $g_1=0$ can be attributed to a matrix, but a singular one and, consequently, outside the group. Summing up, infinite Bell matrices constitute a linear representation of the group of invertible formal series. If we consider only the first N rows and columns, what we have is an approximation, but it is important to notice that the group properties hold at each order N. The general aspect of a Bell matrix can be illustrated by the case $N=5$:

$$B[g]=\begin{pmatrix} g_1 & 0 & 0 & 0 & 0\\ g_2 & {g_1}^2 & 0 & 0 & 0\\ g_3 & 3g_1g_2 & {g_1}^3 & 0 & 0\\ g_4 & 4g_1g_3+3{g_2}^2 & 6{g_1}^2g_2 & {g_1}^4 & 0\\ g_5 & 10g_2g_3+5g_1g_4 & 15g_1{g_2}^2+10{g_1}^2g_3 & 10{g_1}^3g_2 & {g_1}^5\end{pmatrix}. \tag{II.22}$$

The result (II.19) is the best example of the general property

$$B[f]B[f^{\langle -1\rangle}]=I. \tag{II.23}$$

It is evident that, given the series f, its inverse series can be obtained from

$$B[f^{\langle -1\rangle}]=B^{-1}[f] \tag{II.24}$$

by simple matrix inversion. The inversion properties of Bell matrices have been used in the study of cluster expansions for real gases.[6] Because $B[g^{\langle n\rangle}]=B^n[g]$, Bell matrices convert function iteration into matrix power and provide a linearization of the process of iteration.

Suppose now that we are able to obtain the matrix B^t, with t an arbitrary real number. The continuous iteration of $g(x)$ will then be that function $g^{\langle t\rangle}$ such that $B[g^{\langle t\rangle}]=B^t[g]$. By (II.5), its Taylor coefficients are fixed by $g^{\langle t\rangle}{}_n=B[g^{\langle t\rangle}]_{n1}=B^t[g]_{n1}$. To arrive at B^t, let us make a short preliminary incursion into the subject of matrix functions.

III. MATRIX FUNCTIONS

Suppose a function $F(\lambda)$ is given which can be expanded as a power series $F(\lambda)=\Sigma_{k=0}^{\infty}\, c_k(\lambda-\lambda_0)^k$ inside the convergence circle $|\lambda-\lambda_0|<r$. Then the function $F(B)$, whose argument is now a given $N\times N$ matrix B, is defined by $F(B)=\Sigma_{k=0}^{\infty}\, c_k(B-\lambda_0)^k$ and has a sense whenever the eigenvalues of B lie within the convergence circle. Given the eigenvalues $x_1,x_2,\ldots,x_N$, the set of eigenprojectors $\{Z_j[B]=|x_j\rangle\langle x_j|\}$ constitutes a basis in which B is written

$$B=\sum_{j=1}^{N} x_j Z_j[B], \tag{III.1}$$

and the function $F(B)$, defined as above, can also be written[7] as the matrix

$$F(B)=\sum_{j=1}^{N} F(x_j) Z_j[B]. \tag{III.2}$$

Thus, for example, $e^B=\Sigma_{j=1}^{N}\, e^{x_j} Z_j$ and $B^\alpha=\Sigma_{j=1}^{N}\, x_j{}^\alpha\, Z_j[B]$. The basis $\{Z_j[B]\}$ depends on B, but is the same for every function F. The Z_j's, besides being projectors (that is, idempotents, $Z_j^2=Z_j$), can be normalized so that $\mathrm{tr}(Z_j)=1$ for each j and are then orthonormal by the trace, $\mathrm{tr}(Z_iZ_j)=\delta_{ij}$. Other properties follow easily, for example, $\mathrm{tr}[F(B)]=\Sigma_{j=1}^{N}\, F(x_j)Z_j$ and $\mathrm{tr}[B^kZ_j]=(x_j)^k$. If B is a normal matrix diagonalized by a matrix U, $UBU^{-1}=B_{\mathrm{diagonal}}$, then the entries of Z_k are given by $(Z_k)_{rs}=U_{rk}^{-1}U_{ks}$ (no summation, of course).

A set of N powers of B is enough to fix the projector basis. Using for $F(B)$ in (III.2) the power functions $B^0=I$, $B^1,B^2,\ldots,B^{N-1}$, we have $I=\Sigma_{j=1}^{N}\, Z_j$; $B=\Sigma_{j=1}^{N}\, x_j\, Z_j$; $B^2=\Sigma_{j=1}^{N}\, x_j{}^2Z_j;\ldots;$ $B^k=\Sigma_{j=1}^{N}\, x_j{}^kZ_j;\ldots;B^{N-1}=\Sigma_{j=1}^{N}\, x_j{}^{N-1}Z_j$. For $k\geqslant N$, the B^k's are no more independent. This comes from the Cayley–Hamilton theorem,[8] by which B satisfies its own secular equation

$$\Delta(x)=\det[xI-B]=(x-x_1)(x-x_2)(x-x_3)\cdots(x-x_N)=0.$$

$\Delta(B)=0$ will give B^N in terms of lower powers of B, so that the higher powers of B can be computed from the lower powers.

Inversion of the above expressions for the powers of B in terms of the Z_j's leads to a closed form for each Z_j,

$$Z_j[B]=\frac{(B-x_1)(B-x_2)\cdots(B-x_{j-1})(B-x_{j+1})\cdots(B-x_{N-1})(B-x_N)}{(x_j-x_1)(x_j-x_2)\cdots(x_j-x_{j-1})(x_j-x_{j+1})\cdots(x_j-x_{N-1})(x_j-x_N)}. \tag{III.3}$$

The function $F(B)$ is consequently given by

$$F(B)=\sum_j \left\{\prod_{k\neq j}\frac{B-x_k}{x_j-x_k}\right\}F(x_j). \tag{III.4}$$

Thus, in order to obtain $F(B)$, it is necessary to find the eigenvalues of B and the detailed form of its first $(N-1)$ powers. Though for N not too large the $Z_j[B]$'s can be directly computed, we shall give closed expressions for them. These expressions involve some symmetric functions of the eigenvalues.

Let us examine the spectrum $\{x_k\}$ of B in some more detail. The eigenvalues x_k will be called ''letters'' and indicated collectively by the ''alphabet'' $\mathbf{x}=\{x_1,x_2,x_3,\ldots,x_N\}$. A monomial is a ''word.'' It will be convenient to consider both the alphabet $\mathbf{x}$ and its ''reciprocal,'' the alphabet $\mathbf{x}^*=\{x^*{}_1,x^*{}_2,x^*{}_3,\ldots,x^*{}_N\}$ where each $x^*{}_j=-1/x_j$. Notice that taking the reciprocal is an involution, $\mathbf{x}^{**}=\mathbf{x}$. A symmetric function in the variables $x_1,x_2,x_3,\ldots,x_N$ is any polynomial

$P(x_1,x_2,x_3,\ldots,x_N)$ which is invariant under all the permutations of the x_k's. Only one kind of them will be needed here, the ''jth elementary symmetric functions,'' σ_j = sum of all words with j distinct letters:

$$\sigma_0[\mathbf{x}]=1(\text{by convention}),$$

$$\sigma_1[\mathbf{x}]=x_1+x_2+x_3+\cdots+x_N,$$

$$\sigma_2[\mathbf{x}]=x_1x_2+x_1x_3+\cdots+x_1x_N+\cdots+x_2x_3+x_2x_4+\cdots+\cdots+x_{N-1}x_N,$$

$$\vdots$$

$$\sigma_N[\mathbf{x}]=x_1x_2x_3\cdots x_{N-1}x_N.$$

The symmetric functions of $\mathbf{x}$ and $\mathbf{x}^*$ are related by

$$(-)^j\sigma_{N-j}[\mathbf{x}]=\sigma_N[\mathbf{x}]\sigma_j[\mathbf{x}^*]. \tag{III.5}$$

Their generating function is

$$\sum_{j=0}^{N}\sigma_j[\mathbf{x}]t^j=\sum_{j=1}^{N}(1+x_jt)=\prod_{j=1}^{N}(1-t/x^*{}_j)=\frac{1}{\sigma_N[\mathbf{x}^*]}\prod_{j=1}^{N}(x^*{}_j-t),$$

so that

$$\prod_{j=1}^{N}(x^*{}_j-t)=\sigma_N[\mathbf{x}^*]\sum_{j=0}^{N}\sigma_j[\mathbf{x}]t^j.$$

We use the involution property and (III.5) to write the general expression

$$\prod_{j=1}^{N}(x_j-t)=\sigma_N[\mathbf{x}]\sum_{j=0}^{N}\sigma_j[\mathbf{x}^*]t^j=\sum_{j=0}^{N}(-)^j\sigma_{N-j}[\mathbf{x}]t^j. \tag{III.6}$$

The jth eigenvalue is absent in the numerator of expression (III.3) for Z_j. We shall need some results involving an alphabet with one missing letter. Let $\sigma_{ji}[\mathbf{x}]$ be the sum of all j-products of the alphabet $\mathbf{x}$, but excluding x_i. For example, $\sigma_{Ni}[\mathbf{x}]=\Pi_{k\neq i}^{N}x_k$. We put by convention $\sigma_{0i}=1$ and find that

$$\sigma_{ki}[\mathbf{x}]=\sum_{p=0}^{k}(-)^px_i{}^p\sigma_{k-p}[\mathbf{x}]=\sum_{j=0}^{k}(-x_i)^{k-j}\sigma_j[\mathbf{x}]. \tag{III.7}$$

In the absence of the ith letter, (III.6) becomes

$$\prod_{j=1;j\neq i}^{N}(x_j-t)=\sigma_{Ni}[\mathbf{x}]\sum_{j=0}^{N}\sigma_{ji}[\mathbf{x}^*]t^j. \tag{III.8}$$

The projectors (III.3) are then

$$Z_j[B]=\prod_{k=1;k\neq j}^{N}\frac{x_k-B}{x_k-x_j}=\frac{\Sigma_{k=0}^{N}\sigma_{kj}[\mathbf{x}^*]B^k}{\Sigma_{k=0}^{N}\sigma_{kj}[\mathbf{x}^*]x_j{}^k}, \tag{III.9}$$

clearly written in the basis $\{I,B,B^2,\ldots,B^{N-1}\}$. In our application to Bell matrices, it will be convenient to use instead the basis $\{B,B^2,\ldots,B^N\}$. This is due to the fact that we shall prefer to start from the matrix $B[g]$, corresponding to g, and not from the matrix I, corresponding to the identity function. The mappings of interest, like the logistic map for example, have a general ''$\cap$'' aspect and the identity map does not have the same end points: the identity map has Brouwer[9]

degree 1, while the dynamical maps of interest can be characterized by their degree 0. The Cayley–Hamilton theorem implies $\Sigma_{j=0}^{N}\ \sigma_j[\mathbf{x}^*]B^j=0$, from which we obtain the identity B^0 as

$$I=-\sum_{j=1}^{N}\ \sigma_j[\mathbf{x}^*]B^j. \tag{III.10}$$

Replacing this identity in (III.9), the projectors are recast into another form,

$$Z_i[B]=\frac{I+\Sigma_{k=1}^{N}\sigma_{ki}[\mathbf{x}^*]B^k}{1+\Sigma_{k=1}^{N}\sigma_{ki}[\mathbf{x}^*]x_i{}^k}=\frac{\Sigma_{k=1}^{N}\{\sigma_{ki}[\mathbf{x}^*]-\sigma_k[\mathbf{x}^*]\}B^k}{\Sigma_{k=1}^{N}\{\sigma_{ki}[\mathbf{x}^*]-\sigma_k[\mathbf{x}^*]\}x_i{}^k}. \tag{III.11}$$

Using (III.7) and (III.5) we get

$$Z_i[B]=\frac{\Sigma_{k=1}^{N}\{\Sigma_{j=0}^{k-1}(x_i)^{j-k}\sigma_j[\mathbf{x}^*]\}B^k}{\Sigma_{j=0}^{N-1}(N-j)(x_i)^j\sigma_j[\mathbf{x}^*]}=\frac{\Sigma_{k=1}^{N}\{\Sigma_{j=0}^{k-1}(x_i)^{j-k}(-)^j\sigma_{N-j}[\mathbf{x}]\}B^k}{\Sigma_{j=0}^{N-1}(N-j)(x_i)^j(-)^j\sigma_{N-j}[\mathbf{x}]}, \tag{III.12}$$

where we have also used $\Sigma_{r=1}^{N}\{\Sigma_{j=0}^{r-1}(x_i)^j\sigma_j[\mathbf{x}^*]\}=\Sigma_{j=0}^{N-1}(N-j)(x_i)^j\sigma_j[\mathbf{x}^*]$. There is a good immediate check: Replacing B by the eigenvalues we find, as expected,

$$Z_i[x_k]=\delta_{ik}I. \tag{III.13}$$

The projectors are now clearly in the basis $\{B,B^2,\ldots,B^N\}$. Actually, each Z_i is now just that given in the basis $\{I,B,B^2,\ldots,B^{N-1}\}$ multiplied by (B/x_i): instead of (III.3),

$$Z_j[B]=\frac{(B-x_1)(B-x_2)\cdots(B-x_{j-1})B(B-x_{j+1})\cdots(B-x_{N-1})(B-x_N)}{(x_j-x_1)(x_j-x_2)\cdots(x_j-x_{j-1})x_j(x_j-x_{j+1})\cdots(x_j-x_{N-1})(x_j-x_N)}. \tag{III.14}$$

The Z_i's and the powers B^k can be seen as components of two formal column ''vectors.'' The linear conditions $B^n=\Sigma_{j=1}^{N}\ x_j{}^nZ_j$ are then represented by a matrix $\Lambda=[x_j{}^n]$,

$$\begin{pmatrix} B^1\\ B^2\\ .\\ .\\ .\\ B^N\end{pmatrix}=\begin{pmatrix} x_1 & x_2 & x_3 & . & . & x_N\\ x_1{}^2 & x_2{}^2 & x_3{}^2 & . & . & x_N{}^2\\ . & . & . & . & . & .\\ . & . & . & . & . & .\\ . & . & . & . & . & .\\ x_1{}^N & x_2{}^N & x_3{}^N & . & . & x_N{}^N\end{pmatrix}\begin{pmatrix} Z_1\\ Z_2\\ .\\ .\\ .\\ Z_N\end{pmatrix}, \tag{III.15}$$

and what we have done has been to obtain its inverse:

$$Z_i[B]=\sum_{k=1}^{N}\ [\Lambda^{-1}]_{ik}B^k, \tag{III.16}$$

$$[\Lambda^{-1}]_{ik}=\frac{\Sigma_{j=0}^{k-1}(x_i)^{j-k}(-)^j\sigma_{N-j}[\mathbf{x}]}{\Sigma_{j=0}^{N-1}(N-j)(x_i)^j(-)^j\sigma_{N-j}[\mathbf{x}]}. \tag{III.17}$$

It seems a difficult task to improve the above expressions, as it would mean knowing a closed analytical expression for the recurrent summation of the form $\Sigma_{j=0}^{k-1}\ u^j\sigma_j[\mathbf{x}^*]$. A closed expression for $\sigma_j[\mathbf{x}^*]$ would be necessary and, even for the simple alphabet consisting of powers of a fixed letter a, which we shall find in the application to Bell matrices, this would be equivalent to solving an as yet unsolved problem in Combinatorics. In effect, in terms of such an alphabet $\{a^j\}$ with N letters, the symmetric function is given by $\sigma_k=\Sigma_{j=1}^{N(N+1)/2}\ q_{j,k,N}a^j$, where $q_{j,k,N}$ = number of partitions of j into k unequal summands, each one $\leqslant N$. These conditional partition numbers have the generating function $\Pi_{r=1}^{N}[1+ua^r]=1+\Sigma_{j\text{ and }k\geqslant 1}q_{j,k,N}a^ju^k$, but have no known closed expression. They are calculated, one by one, just in this way.[10]

Bell matrices are not normal, that is, they do not commute with their transposes. Normality is the condition for diagonalizability. This means that Bell matrices cannot be put into diagonal form by a similarity transformation. As it happens, this will not be a difficulty because we know their eigenvalues. That functions of matrices are completely determined by their spectra is justified on much more general grounds. Matrix algebras are very particular kinds of von Neumann algebras and it is a very strong result of the still more general theory of Banach algebras[11] that functions on such spaces, as long as they can be defined, are fixed by the spectra. Another point worth mentioning is that the infinite Bell matrices which constitute the true, complete representation of the group of invertible series will belong, as limits for $N\to\infty$ of $N\times N$ matrices, to a hyperfinite von Neumann algebra. Our considerations here are purely formal from the mathematical point of view, as we are only discussing formal series. We are not concerned with the topological intricacies involved in the convergence problems, though they surely deserve a detailed study. By the way, the infinite algebra generated by Bell matrices would provide a good guide in the study of function algebras with the composition operation.

IV. THE CONTINUOUS ITERATE

We are now in a position to find, given a function g, the matrix $B^t[g]$ and its corresponding function, the continuous iterate $g^{\langle t\rangle}(x)$. Since many things—the $B[g]$ spectrum, for example—will depend only on the first-order Taylor coefficient, we shall put $g_1=a$. By (II.6), the letters in the eigenvalue-alphabet of $B[g]$ will be simple powers of a and the alphabet $\mathbf{a}=(a^1,a^2,\ldots,a^N)$ will have the reciprocal $\mathbf{a}^*=(-a^{-1},-a^{-2},\ldots,-a^{-N})$. The matrix Λ has entries $(\Lambda_{ik})=(a^{ik})$ and the projectors

$$Z_i[B]=\sum_{k=1}^{N}\Lambda^{-1}{}_{ik}B^k[g] \tag{IV.1}$$

now have the coefficients

$$[\Lambda^{-1}]_{ik}=\frac{\Sigma_{j=0}^{k-1}a^{i(j-k)}(-)^j\sigma_{N-j}[\mathbf{a}]}{\Sigma_{r=1}^{N}\Sigma_{j=0}^{r-1}a^{ij}(-)^j\sigma_{N-j}[\mathbf{a}]}.$$

We can verify easily that $\mathrm{tr}\, Z_i[B]=1$ and $BZ_i=a^iZ_i$. A consequence of the latter is

$$f(B)Z_i=f(a^i)Z_i, \tag{IV.2}$$

whose particular case

$$B^tZ_i=a^{it}Z_i \tag{IV.3}$$

will be helpful later on. To give an idea of their aspect, we show the projector matrices $Z_i^{(N)}[B[g]]$ for $N=3$:

$$Z_1{}^{(3)}=\begin{pmatrix}1 & 0 & 0\\ \dfrac{g_2}{g_1(1-g_1)} & 0 & 0\\ \dfrac{3g_2{}^2+g_3(1-g_1)}{g_1(1-g_1)^2(1+g_1)} & 0 & 0\end{pmatrix},\quad Z_2{}^{(3)}=\begin{pmatrix}0 & 0 & 0\\ \dfrac{-g_2}{g_1(1-g_1)} & 1 & 0\\ \dfrac{-3g_2{}^2}{g_1{}^2(1-g_1)^2} & \dfrac{3g_2}{g_1(1-g_1)} & 0\end{pmatrix},$$

$$Z_3{}^{(3)}=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \dfrac{3g_2{}^2-g_1g_3(1-g_1)}{g_1{}^2(1-g_1)^2(1+g_1)} & \dfrac{-3g_2}{g_1(1-g_1)} & 1\end{pmatrix}.$$

Because they do not have the form (II.22), they cannot be the Bell matrices of any function. They inherit, however, a good property of the Bell matrices: for each N, the projectors $Z_j{}^{(N)}$ contain, in their higher rows, the projectors $Z_j{}^{(k)}$ for all $k<N$. The upper-left 2×2 submatrices of the $Z_k{}^{(3)}$'s above are just $Z_k{}^{(2)}$.

We can take the first-column entries of the matrix (IV.1) as Taylor coefficients defining the functions

$$R_i{}^{(N)}(x)=\sum_{r=1}^{\infty}\frac{x^r}{r!}\,[Z_i[B]]_{r1} \tag{IV.4}$$

$$=\sum_{k=1}^{N}\,[\Lambda^{-1}{}_{ik}g^{\langle k\rangle}(x)]. \tag{IV.5}$$

To each projector Z_i corresponds such an ''elementary function'' $R_i{}^{(N)}(x)$, a relationship reflecting in part that between the series and their Bell matrices. Taking the summation of all $R_i{}^{(N)}$'s in (IV.4) and using $I=\Sigma_{i=1}^{N}\,Z_i$, we find

$$\sum_{i=1}^{N}\,R_i{}^{(N)}(x)=\sum_{r=1}^{\infty}\,\frac{x^r}{r!}\left[\sum_{i=1}^{N}\,Z_i[B]\right]_{r1}=\sum_{r=1}^{\infty}\,\frac{x^r}{r!}\,\delta_{r1}=x. \tag{IV.6}$$

Thus, just as the projectors give a decomposition of the identity matrix, the elementary functions provide a decomposition of the identity function,

$$\mathrm{Id}=\sum_{i=1}^{N}\,R_i{}^{(N)}. \tag{IV.7}$$

Applying to $B[g]$ the general form (III.2) for the function of a matrix, we have

$$B^t=\sum_{i=1}^{N}\,a^{it}Z_i[B]=\sum_{k=1}^{N}\,C^{(N)}{}_k(t)B^k, \tag{IV.8}$$

where

$$C^{(N)}{}_k(t)=\sum_{i=1}^{N}\,a^{it}\Lambda^{-1}{}_{ik}\,. \tag{IV.9}$$

The coefficients of the continuum iterate function

$$g^{\langle t\rangle}(x)=\sum_{r=1}^{\infty}\,g_r{}^{\langle t\rangle}\,\frac{x^r}{r!} \tag{IV.10}$$

will be

$$g_r{}^{\langle t\rangle}=[B^t[g]]_{r1}=\sum_{i=1}^{N}\,a^{it}[Z_i[B]]_{r1}=\sum_{k=1}^{N}\,C^{(N)}{}_k(t)B^k[g]_{r1}=\sum_{k=1}^{N}\,C^{(N)}{}_k(t)g_r{}^{\langle k\rangle}. \tag{IV.11}$$

Therefore,

$$g^{\langle t\rangle}(x)=\sum_{k=1}^{N}\,C^{(N)}{}_k(t)g^{\langle k\rangle}(x)=\sum_{k=1}^{N}\,\left[\sum_{i=1}^{N}\,a^{it}\Lambda^{-1}{}_{ik}\right]g^{\langle k\rangle}(x). \tag{IV.12}$$

Time dependence is factorized in the alternative form

$$g^{\langle t\rangle}(x)=\sum_{k=1}^{N}\,a^{kt}R_k{}^{(N)}(x). \tag{IV.13}$$

One of the announced properties, (I.2), follows immediately: $g^{\langle 0\rangle}(x)=\mathrm{Id}(x)$. For $t=1$, a suggestive decomposition of the function comes up:

$$g(x)=\sum_{k=1}^{N} a^k R_k{}^{(N)}(x). \tag{IV.14}$$

It remains to show that (IV.13) does satisfy property (I.1). Notice first that, from the definition (IV.4) of $R_i{}^{(N)}$ and the multinomial theorem (II.3),

$$\begin{aligned} R_i{}^{(N)}[g^{\langle t\rangle}(x)] &= \sum_{r=1}^{\infty} [Z_i[B]]_{r1} \frac{[g^{\langle t\rangle}(x)]^r}{r!} \\ &= \sum_{r=1}^{\infty} \sum_{j\geqslant r} \frac{x^j}{j!} B_{jr}[g^{\langle t\rangle}][Z_i[B]]_{r1} \\ &= \sum_{j=1}^{\infty} \frac{x^j}{j!} \sum_{r=1}^{j} B^t_{jr}[g][Z_i[B]]_{r1} \\ &= \sum_{j=1}^{\infty} \frac{x^j}{j!} [B^t Z_i]_{j1}. \end{aligned} \tag{IV.15}$$

Therefore,

$$\begin{aligned} g^{\langle t'\rangle}(g^{\langle t\rangle}(x)) &= \sum_{i=1}^{N} a^{it'} R_i{}^{(N)}[g^{\langle t\rangle}(x)] \\ &= \sum_{r=1}^{\infty} \frac{x^r}{r!} \sum_{i=1}^{N} a^{it'} [B^t Z_i]_{r1} \\ &= \sum_{r=1}^{\infty} \frac{x^r}{r!} \left[B^t \sum_{i=1}^{N} a^{it'} Z_i \right]_{r1} \\ &= \sum_{r=1}^{\infty} \frac{x^r}{r!} [B^t B^{t'}]_{r1} = \sum_{r=1}^{\infty} \frac{x^r}{r!} g^{\langle t+t'\rangle}{}_{r1} = g^{\langle t+t'\rangle}(x), \end{aligned}$$

just the result looked for. Notice also that, using (IV.3) and (IV.4) in (IV.15), we obtain

$$R_k{}^{(N)}[g^{\langle t\rangle}(x)] = a^{kt} R_k{}^{(N)}(x) \tag{IV.16}$$

and, consequently,

$$R_k{}^{(N)}[g^{\langle t+t'\rangle}(x)] = a^{k(t+t')} R_k{}^{(N)}(x) = a^{kt'} R_k{}^{(N)}[g^{\langle t\rangle}(x)]. \tag{IV.17}$$

The function decomposition is preserved in time, as we can write

$$g(x)=\sum_{k=1}^{N} R_k{}^{(N)}[g(x)], \quad g^{\langle t\rangle}(x)=\sum_{k=1}^{N} R_k{}^{(N)}[g^{\langle t\rangle}(x)]. \tag{IV.18}$$

All the above expressions hold for each value of N and give approximations to that order. Of course, exact results would only really come out when $N\to\infty$. The exact expressions would, as long as they are defined, take forms like

$$R_j(x)=\sum_{k\geqslant 1} [\Lambda^{-1}{}_{jk} g^{\langle k\rangle}(x)], \tag{IV.19}$$

$$g(x)=\sum_{k\geqslant 1} a^k R_k(x), \tag{IV.20}$$

$$g^{\langle t\rangle}(x)=\sum_{k\geqslant 1} a^{kt} R_k(x). \tag{IV.21}$$

It is tempting to conjecture that the elementary functions R_i play, in function space and with the operation of composition, the role of projectors. In fact, this is not so. Their properties mirror only partially those of the projectors Z_i. If we calculate $R_i[R_j(x)]$ using (IV.4) and the multinomial theorem, we find $R_i[R_j(x)]=R_i(x^j/j!)$. The composition gives rise to a change in the variable. Thus, the functions R_i fail even to satisfy the defining idempotent property: in general, $R_i\circ R_i\neq R_i$. An explicit example of $R_j(x)$, for which all these properties can be tried out, is given below [see (V.6)].

We can introduce $\epsilon=\ln a$ and rewrite (IV.13) as

$$x(t)=g^{\langle t\rangle}(x_0)=\sum_{k\geqslant 1} e^{k\epsilon t} R_k(x_0), \tag{IV.22}$$

a decomposition of $g^{\langle t\rangle}(x)$ into a sum of modes, each one evolving independently according to

$$x_k(t)=e^{k\epsilon t}x_k(0)=e^{k\epsilon t}R_k(x_0). \tag{IV.23}$$

The imaginary ''frequency'' $k\epsilon$ plays the role of a ''modular Lyapunov exponent'' of the kth mode. We have thus a ''multi-Hamiltonian'' flow $x(t)$ with one Hamiltonian for each projector component: Equations (IV.16) and (IV.17) show $R_k{}^{(N)}$ as a representation of the one-dimensional group engendered by the kth dynamical flow

$$R_k[g^{\langle t\rangle}(x)]=e^{k\epsilon t}R_k[g^{\langle 0\rangle}(x)]. \tag{IV.24}$$

The function $g^{\langle t\rangle}(x)$ is actually a Lagrange interpolation in the convenient variable $(g_1)^t=a^t$, which coincides with the discrete iterates at each integer value of t and *keeps the meaning of a continuum iterate in between*. It is a homotopy of all the usual discrete iterations. As announced in Sec. III, we have used the basis $\{B,B^2,B^3,\ldots,B^N\}$, instead of $\{I,B,B^2,\ldots,B^{N-1}\}$, because in the latter the matrix ''I'' would correspond to the identity function $\mathrm{Id}(x)=x$, which does not have the same end points of the maps of interest.

The variable t has been given the sense of a ''time.'' If it is really time, $g^{\langle n\rangle}(x)$ will give the nth point of a Poincaré map. Now, there are in principle (infinitely) many dynamical flows corresponding to a given Poincaré map. The function $g^{\langle t\rangle}(x)$ as given above would correspond to a class of them, that of the flows with equal intervals of time between successive points on the Poincaré section. To consider other cases, we recall that any strictly monotonous function of a first-given ''time'' is another ''time.''

Finally, we cannot resist exhibiting a ''thing of beauty.'' By the very manner in which it has been obtained, the expression (IV.12) for the continuum iterate is equivalent to the highly mnemonic determinantal equation

$$\begin{vmatrix} g^{\langle 1\rangle}(x) & a & a^2 & \ldots & \ldots & a^N \\ g^{\langle 2\rangle}(x) & a^2 & a^4 & \ldots & \ldots & a^{2N} \\ g^{\langle 3\rangle}(x) & a^3 & a^6 & \ldots & \ldots & a^{3N} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ g^{\langle N\rangle}(x) & a^N & a^{2N} & \ldots & \ldots & a^{N^2} \\ g^{\langle t\rangle}(x) & a^t & a^{2t} & \ldots & \ldots & a^{Nt} \end{vmatrix} = 0. \tag{IV.25}$$

Expansion along the first column or the last row and comparison with (IV.11) and (IV.13) will give determinant expressions for $C^{(N)}{}_k(t)$ and $R_k{}^{(N)}(x)$.

V. AN APPLICATION TO TURBULENCE

Let us describe an application of our basic result (IV.21), which shows that a relationship may exist between harmonic oscillators and turbulence. Consider the special case of the logistic map

$$g(x)=4x(1-x), \tag{V.1}$$

which is helpful in modeling the velocity field in fully developed turbulence (the ''poor man's Navier–Stokes equation'').[12] It takes the interval $[0, 1]$ into itself. The coordinate transformation $x'=2x-1$, $y'=2y-1$, takes it into a map $[-1,1]\to[-1,1]$. We shall write $v(x)=2x-1$, in terms of which the logistic map takes the form

$$g(x)=\tfrac{1}{2}[1-T_2(2x-1)], \tag{V.2}$$

where $T_2(z)=2z^2-1=\cos[2\arccos z]$ is the second-order Chebyshev polynomial of the first kind. Using $v^{\langle -1\rangle}(u)=\frac{1}{2}(u+1)$, we notice that

$$T_2(z)=-v\circ g\circ v^{\langle -1\rangle}(z).$$

Iterating once this expression, we have $[T_2\circ T_2](z)=-v\circ g\circ v^{\langle -1\rangle}[-v]\circ g\circ v^{\langle -1\rangle}(z)$. As $v^{\langle -1\rangle}[-v(u)]=1-u$, but $g(1-u)=g(u)$, we arrive at $T_2{}^{\langle 2\rangle}(z)=-v\circ g^{\langle 2\rangle}\circ v^{\langle -1\rangle}(z)$. In the same way we obtain the nth iterate

$$T_2{}^{\langle n\rangle}(z)=-v\circ g^{\langle n\rangle}\circ v^{\langle -1\rangle}(z)=1-2g^{\langle n\rangle}[\tfrac{1}{2}(1+z)]. \tag{V.3}$$

Now, these Chebyshev polynomials[13] have a very interesting property, easily proved from the definition: they transform composition into index product:

$$T_n[T_m(z)]=[T_n\circ T_m](z)=T_{nm}(z).$$

Therefore, $T_2{}^{\langle n\rangle}(z)=T_{2^n}(z)=\cos[2^n\arccos z]$, and we obtain a closed expression for the n-iterate of g: $g^{\langle n\rangle}=v^{\langle -1\rangle}\circ[-T_{2^n}(z)]\circ v$, or

$$g^{\langle n\rangle}(x)=\tfrac{1}{2}\{1-T_{2^n}(z)(2x-1)\}=\tfrac{1}{2}\{1-\cos[2^n\ \arccos(2x-1)]\}=\sin^2[2^{n-1}\ \arccos(2x-1)]. \tag{V.4}$$

We proceed to find a continuum version of this iterate. The result,

$$g^{\langle \alpha\rangle}(x)=\tfrac{1}{2}\{1-\cos[2^\alpha\ \arccos(2x-1)\}, \tag{V.5}$$

can be naively anticipated, but we shall prove it. In the present case $a=4$, the eigenvalue alphabet is $(4^1,4^2,4^3,\ldots)$ and matrix $\Lambda^{-1}{}_{jn}$ in (III.17) is such that $\Sigma_{n\geqslant 1}\ \Lambda^{-1}{}_{kn}4^{np}=\delta_k^p$. We use the second expression in (V.4) in the form $g^{\langle n\rangle}(x)=-(1/2)\Sigma_{p\geqslant 1}[(-)^p/(2p)!]\ 4^{np}[\arccos(2x-1)]^{2p}$ and calculate (IV.19):

$$R_k(x)=\sum_{n\geqslant 1}\ \Lambda^{-1}{}_{kn}g^{\langle n\rangle}(x)=-\frac{1}{2}\,\frac{(-)^k}{(2k)!}\,[\arccos(2x-1)]^{2k} \tag{V.6}$$

[notice that (IV.6) is trivially satisfied]. Equation (IV.21) gives then

$$g^{\langle \alpha\rangle}(x)=\sum_{k\geqslant 1}\ 4^{k\alpha}R_k(x)=-\frac{1}{2}\sum_{k\geqslant 1}\ \frac{(-)^k}{(2k)!}\,[2^\alpha\ \arccos(2x-1)]^{2k},$$

which is the same as (V.5).

Let us now recall some facts about the double oscillator. Given $x=\cos\omega t$ and $y=\cos(n\omega t)$, we have of course $y=T_n[x]$, which solves the differential equation $(1-x^2)y''-xy'-n^2y=0$. This leads to the usual, ''isolated,'' finite Lissajous curves. When the two frequencies are not commensurate, $y=\cos(\nu\omega t)$, the differential equation becomes

$$(1-x^2)y''-xy'-\nu^2y=0, \tag{V.7}$$

whose solution $y=\cos[\nu \arccos x]$ is not a polynomial in the general case. It is a well-known fact that, when ν is not a rational number, each one of these transcendental, infinite Lissajous curves is dense in the (x,y) square,[14] and that the topology of the space of such trajectories is not even first separable.[15]

Now, the related functions

$$T_{2^\alpha}(z)=\cos[2^\alpha \arccos(2x-1)]=1-2g^{\langle\alpha\rangle}[\tfrac{1}{2}(1+z)]$$

are just the continuum versions of the Chebyshev polynomials obtained from (V.5). For each value of α, $T_{2^\alpha}(z)$ solves (V.7) with $\nu=2^\alpha$, and constitutes one of the Lissajous curves. The continuum parameter α, which can be interpreted as ''time'' for the logistic map, is a label for Lissajous curves from the point of view of the double oscillator. It is just the set of its values which constitutes a nonseparable space: all trajectories with α irrational are ultimately indistinguishable. However rough the ''poor man's'' model may be, the result suggests that the velocity field in turbulent flow can assume a large amount of nonseparable values.

VI. CONCLUDING REMARKS

In the passage from functions to Bell matrices, composition is translated into matrix product and iteration into matrix powers. Continuous powers of matrices have a sound meaning, which is translated back into continuous iteration and makes of it also a sound concept, respecting properties (I.1)–(I.2). The usual discrete iterations are thus extended into a continuous flow. The term-by-term factorization of the x and t dependencies of $g^{\langle t\rangle}(x)$ reveals itself as a decomposition into independently evolving modes, one for each projector of the corresponding Bell matrix.

Once it is established that iteration can have a continuous meaning and is, furthermore, a homotopy, new possibilities are made open to study, such as the use of the conservation of degrees like Brouwer's and Morse's. It is our contention that, with the continuum version of iterations, it will be possible to get a better understanding of the detailed unfolding of bifurcations and of the general relationship between the differential and the mapping approaches to chaotic dynamics.

ACKNOWLEDGMENTS

This work was supported by CNPg, Brasilia. The authors are grateful to Uriel Frisch for many a hint given during his inspiring lectures at I.F.T. They are also greatly indebted to Dr. Keith M. Briggs for a very useful mail interchange. Dr. Briggs suggested an interesting relationship to the Schröder equation, which is under study.

[1] J. L. McCauley, *Classical Mechanics* (Cambridge University Press, Cambridge, 1997).
[2] See for instance V. I. Arnold, *Les Méthodes Mathématiques de la Méchanique Classique* (MIR, Moscow, 1976).
[3] See for instance S. N. Rasband, *Chaotic Dynamics of Nonlinear Systems* (Wiley, New York, 1989).
[4] P. Henrici, *Applied and Computational Complex Analysis* (Wiley, New York, 1974).
[5] L. Comtet, *Advanced Combinatorics* (Reidel, Dordrecht, 1974), whose notations we have adopted.
[6] R. Aldrovandi and I. Monte Lima, J. Phys. A **13**, 3685 (1980); Astrophys. Space Sci. **90**, 179 (1983).
[7] See F. R. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1990), Vol. I. The simple treatment sketched here holds only for matrices with all eigenvalues distinct.
[8] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra* (Freeman, San Francisco, 1963).
[9] See for instance R. Aldrovandi and J. G. Pereira, *An Introduction to Geometrical Physics* (World Scientific, Singapore, 1995).
[10] See Ref. 5, Chap. II.3.
[11] A. Kirillov, *Élements de la Théorie des Représentations* (MIR, Moscow, 1974); O. Bratelli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I* (Springer, New York, 1979).
[12] U. Frisch, *Turbulence* (Cambridge University Press, Cambridge, 1996).
[13] T. J. Rivlin, *The Chebyshev Polynomials* (Wiley, New York, 1974).
[14] See, for example, R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics* (Wiley, New York, 1975) (appendix).
[15] A. Connes, *Noncommutative Geometry* (Academic, London, 1994).

Vertex operator representation of the soliton tau functions in the $A_n^{(1)}$ Toda models by dressing transformations

H. Belich,[a] G. Cuba,[b] and R. Paunov[c]
Centro Brasileiro de Pesquisas Fisicas, Rua Dr. Xavier Sigaud 150, Rio de Janeiro, Brazil

(Received 16 January 1998; accepted for publication 5 May 1998)

We study the relation between the group-algebraic approach and the dressing symmetry one to the soliton solutions of the $A_n^{(1)}$ Toda field theory in $1+1$ dimensions. Originally, solitons in the affine Toda models were found by Olive, Turok, and Underwood. Single solitons are created by exponentials of elements which addiagonalize the principal Heisenberg subalgebra. Alternatively, Babelon and Bernard exploited the dressing symmetry to reproduce the known expressions for the fundamental tau functions in the sine-Gordon model. In this paper we show the equivalence between these two methods to construct solitons in the $A_n^{(1)}$ Toda models. © *1998 American Institute of Physics.* [S0022-2488(98)01010-X]

I. INTRODUCTION

Exact localizable solutions of a relativistic field theory with finite energy are known in the literature as (depending on the dimension of the space–time) solitons (kinks) or monopoles.[1] Such solutions play an important role, since they are interpreted as new particles, which appear in the spectrum of the theory. As argued in Refs. 1 and 2, the particles corresponding to solitons (or monopoles) are of nonperturbative nature. Usually, the solitons (monopoles) are also characterized by nontrivial topological charges. This provides a connection between the field theory and the geometry.[3] There exists[4] a remarkable duality which exchanges the sine-Gordon model with the massive Thirring model. Under this duality transformation perturbative particles are mapped into solitons and vice versa. The duality also exchanges Noether currents with topological ones. Recently,[5] Seiberg and Witten have shown that there exist supersymmetric field theories in four dimensions which exhibit exact duality.

Within the inverse scattering method (ISM),[6] soliton equations are required to admit Lax or zero curvature representation. This representation guarantees that the spectrum of the Lax operator is constant in time. The leading idea of the ISM is to consider the time evolution as an evolution of the scattering data of the corresponding Lax operator. In view of the zero curvature condition, the equations of motion for the associated scattering data can be found explicitly. In the framework of the ISM, solitons correspond to vanishing reflection coefficients. Imposing this condition, the inverse spectral transformation reduces to a linear algebraic system. The integrability (in the sense of Liouville) also ensures that the interaction between the single solitons is elastic. This intriguing property survives the quantization and was used to get exact quantum S matrices in various integrable models.[7]

The propagation of waves in one-dimensional lattices with exponential nearest-neighbor interaction was studied by Toda.[8] In this pioneering paper, exact soliton solutions were also studied. The Toda lattice equations admit a field theoretical analog in $1+1$ dimensions[9] which exhibit both the integrability and the relation to the Lie algebras. Due to the last sequence of papers, it became clear that (generalized) Cartan matrices can be used to construct integrable exponential interactions in two dimensions. It was also shown that the field equations admit zero curvature representation of a Lax connection whose components belong to a certain Lie algebra. A deep relation between integrable hierarchies and the Kac–Moody (or affine) Lie algebras[10] has been clarified by

[a]Electronic mail: belich@cbpfsu1.cat.cbpf.br
[b]Electronic mail: gcubac@cbpfsu1.cat.cbpf.br
[c]Electronic mail: paunov@cbpfsu1.cat.cbpf.br

 © 1998 American Institute of Physics

Drinfeld and Sokolov.[11] In the last paper the crucial role of Heisenberg subalgebras was also explained and the gradations related to them in constructing integrable evolution equations.[12,13]

Soliton solutions of the sine-Gordon model have been known since the early days of the ISM.[6] N-soliton solutions in the $A_n^{(1)}$ Toda models are found by the use of the Hirota method in Ref. 14. It also became clear that physical observables, when evaluated on solitons, take finite real quantities for imaginary values of the coupling constant. These results have been extended by using group-algebraic methods,[15,16] which also provide a bridge between the integrable models and the Kac–Moody algebras. In particular, the soliton tau functions corresponding to a fundamental highest weight state $|\Lambda\rangle$ of the affine Lie algebra admit the representation

$$\frac{\tau_\Lambda(\Phi)}{\tau_\Lambda(\Phi_0)}=\langle\Lambda|\prod_{i=1}^{N}(1+X_iF^{r_i}(\mu_i))|\Lambda\rangle, \tag{I.1}$$

where Φ_0 is the vacuum solution, X_i are numerical factors depending exponentially on the light cone coordinates, and F^{r_i} are elements of the affine Lie algebra which diagonalize the adjoint action of the principal Heisenberg subalgebra (for details see Refs. 10 and 15) and N is the number of the solitons. The above expression is analogous to the representation obtained by the Kyoto group[17] for the tau function of the KP hierarchy.

Integrable models possess a dressing symmetry.[18] The dressing group acts by gauge transformations on the Lax connection and preserves its form. An important property of the dressing symmetry is that it is a Lie–Poisson group. Applications of this symmetry to the Toda field theories has been done in Ref. 19. Using the dressing symmetry, an alternative expression for the soliton tau functions arises,

$$\frac{\tau_\Lambda(\Phi)}{\tau_\Lambda(\Phi_0)}=\langle\Lambda|g_-^{-1}(x^+,x^-).g_+(x^+,x^-)|\Lambda\rangle, \tag{I.2}$$

where g_-^{-1} and g_+ are triangular elements of the affine group which generate solitons from the vacuum. Babelon and Bernard demonstrated that the expressions (I.1) and (I.2) are equivalent, at least for affine sine-Gordon solitons.[20] In our previous paper[21] we obtained explicit expressions for the elements $g_\pm$ (I.2). In the present work, by using the results of Ref. 21, we show that the vertex operator representation of the tau functions (I.1) corresponding to fundamental representations of the affine Lie algebra is a consequence of the dressing group expression (I.2) for arbitrary $A_n^{(1)}$ Toda solitons.

The paper is organized as follows: In Sec. II we introduce the $A_n^{(1)}$ Toda models, their soliton solutions and briefly comment on the results of Ref. 21. Section III is also complementary: the vertex operator representation of the $A_n^{(1)}$ algebras in the principal gradation is derived. In Sec. IV we obtain the vertex operator representation of the monosoliton tau functions starting from the dressing symmetry. Section V generalizes this result to generic soliton solutions. The Appendix contains a brief summary of the Lie algebraic background used in the paper.

II. $A_n^{(1)}$ TODA SOLITONS AND THE RELATED DRESSING TRANSFORMATIONS

Solitons in the $A_n^{(1)}$ Toda models have been found by Hollowood,[14] who used the Hirota equations. Physically relevant solutions appear for imaginary values of the coupling constant, nevertheless the components of the Toda field are complex. Affine Toda field theories and their soliton solutions are also studied from the point of view of the ISM.[22] The underlying Jost solutions[6] as well as the elements of the transition matrix lose, in general, their nice analyticity properties. A similar phenomenon occurs when one considers the scattering data related to a linear differential operator of an arbitrary order.[23] In Ref. 21 we developed an elegant method to get $A_n^{(1)}$ Toda solitons. It strongly exploits the fact that the Lax connection belongs to the $A_n^{(1)}$ Lie algebra in the *principal* gradation (for a summary of the Lie algebraic background, see the Appendix). The dynamical variables which we use to describe the soliton dynamics, appeared previously in the study of the periodic solutions of the Korteweg–de Vries (KdV) equation and of the periodic Toda chain.[24] The same approach, applied to the sine-Gordon solitons, was recently exploited to compute form factors in the quantum theory.[25]

To get the $A_n^{(1)}$ Toda equations, we impose the zero curvature condition

$$\partial_+ A_- - \partial_- A_+ + [A_+, A_-] = 0 \tag{II.1a}$$

on the Lax connection

$$A_\pm = \pm \partial_\pm \Phi + m e^{\pm ad\Phi} \mathscr{E}_\pm, \quad \partial_\pm = \frac{\partial}{\partial x^\pm}, \tag{II.1b}$$

which belongs to the loop Lie algebra $\tilde{\mathscr{G}} = \widetilde{\mathrm{SL}(n+1)}$; $x^\pm = x \pm t$ are the light cone coordinates in the two-dimensional Minkowski space. The elements $\mathscr{E}_\pm = \mathscr{E}_{\pm 1}$ are generators of grade ± 1 of the principal Heisenberg subalgebra (A7), (A23),

$$\mathscr{E}_k = \lambda^k \sum_{p \in \mathbb{Z}_{n+1}} |p\rangle\langle p+k| \tag{II.1c}$$

and Φ is a diagonal matrix

$$\Phi = \frac{1}{2} \sum_{i \in \mathbb{Z}_{n+1}} \varphi_i E^{ii}. \tag{II.1d}$$

The $A_n^{(1)}$ Toda models have a conformally invariant extension[26] known as the A_n conformal affine Toda (CAT) model. It also admits a zero curvature representation (II.1a) for a connection of the form (II.1b) in the *affine* Lie algebra $\hat{\mathscr{G}} = \hat{\mathrm{sl}}(n+1) = A_n^{(1)}$.[10] The affine algebra analog of (II.1d) is

$$\Phi \to \Phi + \eta \hat{d} + \frac{\hat{c}}{2(n+1)} \zeta, \tag{II.2}$$

where (see the Appendix) $\hat{d}$ and $\hat{c}$ are the derivation and the central element, respectively. The A_n CAT field equations are

$$\begin{aligned} \partial_+ \partial_- \varphi_i &= m^2 e^{2\eta} (e^{\varphi_i - \varphi_{i+1}} - e^{\varphi_{i-1} - \varphi_i}), \\ \partial_+ \partial_- \eta &= 0, \\ \partial_+ \partial_- \zeta &= m^2 e^{2\eta} \sum_{i \in \mathbb{Z}_{n+1}} e^{\varphi_i - \varphi_{i+1}}. \end{aligned} \tag{II.3}$$

The first of the above equations coincides with the $A_n^{(1)}$ Toda equations provided that $\eta = 0$. As noted in Ref. 20, affine Toda solitons arise after imposing the last restriction. In such a case (II.3) admit a Hirota bilinear representation[14]

$$\partial_+ \tau_k \partial_- \tau_k - \tau_k \partial_+ \partial_- \tau_k = m^2 (\tau_{k+1} \tau_{k-1} - \tau_k^2), \tag{II.4a}$$

$$e^{-\varphi_k} = \frac{\tau_k}{\tau_{k-1}}, \quad k \in \mathbb{Z}_{n+1}, \tag{II.4b}$$

$$e^{\zeta_0 - \zeta} = \prod_{k \in \mathbb{Z}_{n+1}} \tau_k, \tag{II.4c}$$

where

$$\zeta_0 = (n+1) m^2 x^+ x^-. \tag{II.4d}$$

The above value of the field ζ together with $\varphi_k = 0$, $k \in \mathbb{Z}_{n+1}$ corresponds to the vacuum solution of the CAT model. The $A_n^{(1)}$ Toda vacuum is obtained by ignoring the additional field ζ.

In Ref. 21 we used an alternative procedure to get $A_n^{(1)}$ Toda solitons. The approach we used strongly relies on the work[27] where several soliton equations, including the sine-Gordon equation, have been studied. The dynamics of the $A_n^{(1)}$ Toda solitons is governed by the following *algebraic* equations:

$$\prod_{l=1}^{N} \frac{\epsilon_{kl}+\omega^{r_j}\mu_j}{\epsilon_{kl}+\mu_j}=c_j\omega^{r_j(1-k)}\frac{e(\omega^{r_j}\mu_j)}{e(\mu_j)},$$
$$e(\lambda)=\exp\left\{m\left(\lambda x^{+}+\frac{x^{-}}{\lambda}\right)\right\}, \tag{II.5a}$$

$$\epsilon^{-\varphi_k}=(-)^N\prod_{j=1}^{N}\frac{\epsilon_{kj}}{\mu_j}, \quad k\in\mathbb{Z}_{n+1}, \tag{II.5b}$$

where the integer N stands for the number of solitons; μ_j, $j=1,...,N$ are (complex) parameters related to the soliton rapidities, and r_j are discrete parameters which take nonvanishing values in the cyclic group $\mathbb{Z}_{n+1}$. The latter are known in the literature[14–16] as *soliton species*. The relation between solitons and N-body integrable relativistic systems was studied in detail.[28,29] In order to simplify our analysis, only solitons with $|\mu_i|\neq|\mu_j|$ $(i\neq j)$ will be considered. Certain particular solutions which violate this restriction are also known.[30] Let us note that in contrast to Refs. 14–16, we are working with a real value of the coupling constant. This implies that the solutions (II.5a), (II.5b) are solitons in the *algebraic* sense.

Transformation groups for soliton equations were introduced in Ref. 17. They have also been studied in relation to the underlying Riemann problem[31] and are known as groups of dressing transformations. In the present paper we will not comment on the Poisson–Lie properties[18,19] of the dressing group. For dressing group elements which create monosoliton solutions from the vacuum in the sine-Gordon model this problem has been discussed in Ref. 32. One believes that[20] in the limit $N\to\infty$, the N-soliton solutions are dense in the dressing group orbit of the vacuum, and therefore, the soliton parameters are expected to provide a convenient coordinate system on the dressing group.

To introduce the group of dressing transformations we first recall that the zero curvature condition (II.1a) can be equivalently written as the compatibility condition of the linear system

$$(\partial_{\pm}+A_{\pm})T=0. \tag{II.6}$$

We shall also impose the normalization condition $T|_{x^+=x^-=0}=1$. An element of the dressing group is represented by a pair of triangular elements $(g_+(x),g_-(x))$ of the corresponding loop (or affine) Lie group. This is to say that $g_{\pm}(x)=e^{\mathscr{H}}e^{\mathscr{N}_{\pm}}$ where $\mathscr{H}$ is the Cartan subalgebra and $\mathscr{N}_{\pm}$ contain all the elements of positive (negative) grade. The gradation is introduced by the derivation $\hat{d}$ of the loop (affine) Lie algebra (A18) and (A19). The dressing group acts on the components of the Lax connection by gauge transformations

$$A_\mu\to A_\mu^g=-\partial_\mu g_{\pm}g_{\pm}^{-1}+g_{\pm}A_\mu g_{\pm}^{-1}, \tag{II.7a}$$

or equivalently

$$T(x)\to T(x)^g=g_{\pm}(x)T(x)g_{\pm}^{-1}(0). \tag{II.7b}$$

The space–time independent factors $g_{\pm}^{-1}(0)$ are added to ensure the normalization condition $T|_{x^+=x^-=0}=1$. The gauge transformations are also required to preserve the form of the Lax connection (II.1b). Therefore, the dressing transformations form a symmetry group of the corresponding field equations (II.3). Since the group elements g_+ and g_- produce the same result (II.7a), it turns out that they are a solution of the factorization problem

$$g_-^{-1}(x)g_+(x)=T(x)g_-^{-1}(0)g_+(0)T^{-1}(x). \tag{II.8}$$

Comparing (II.1b) with (II.7a), one observes that g_+ and g_- have opposite components on the Cartan torus $e^{\mathscr{H}}$. Therefore, the solution of (II.8) is unique. The multiplication in the dressing group is the same as in the dual of the loop group.[18,19] In particular, the map $(g_+,g_-) \to g_-^{-1}g_+$ is not an isomorphism of Lie groups.

In the CAT models based on arbitrary simple Lie algebras, one associates a group-algebraic tau function to each highest weight vector $|\Lambda\rangle$,

$$\tau_\Lambda(\Phi)=\langle\Lambda|e^{-2\Phi}|\Lambda\rangle. \tag{II.9}$$

Suppose that a solution Φ (II.2), (II.3) with $\eta=0$ is related to the vacuum solution $\phi_0=(m^2/2)x^+x^-\hat{c}$ by dressing transformation (g_+,g_-). The following relation[19,20]

$$\frac{\tau_\Lambda(\Phi)}{\tau_\Lambda(\phi_0)}=\langle\Lambda|g_-^{-1}(x)g_+(x)|\Lambda\rangle \tag{II.10}$$

between the tau functions of the two solutions is valid. In Ref. 20 it was observed that, after factorizing out the contribution which belongs to the center of the affine group, one can perform the calculation of the element $g_\pm$ in the corresponding loop group:

$$g_\pm(x)=\exp\left(\pm\frac{\zeta_0-\zeta}{2(n+1)}\hat{c}\right)\tilde{g}_\pm(x), \tag{II.11}$$

where $\tilde{g}_\pm$, considered as elements of the loop group, generate the $A_n^{(1)}$ Toda solution $\Phi=\frac{1}{2}\Sigma_{i\in\mathbb{Z}_{n+1}}\varphi_i|i\rangle\langle i|$ from the vacuum $\Phi_0=0$. We recall that the above identity has been established in Ref. 20 for the affine Lie algebra $A_1^{(1)}=\hat{\mathrm{sl}}(2)$. However, it is easy to check that the proof can be easily generalized for an arbitrary affine Lie algebra.

The representation (II.5a), (II.5b) of the N-soliton solutions was used[21] to calculate the elements $\tilde{g}_\pm$ which by (II.7a) generate these solutions from the vacuum. It turns out that the dressing group elements $\tilde{g}_\pm$ can be factorized into a product of monosoliton factors

$$\tilde{g}_\pm=\tilde{g}_\pm(N)\tilde{g}_\pm(N-1)\cdots\tilde{g}_\pm(1), \tag{II.12}$$

where

$$\tilde{g}_\pm(i)=e^{K(F_i)+P_i}e^{W_\pm(i)}. \tag{II.13a}$$

In the above expression, F_i, $K(F_i)$, and P_i are diagonal traceless matrices

$$F_i=\frac{1}{2}\sum_{k\in\mathbb{Z}_{n+1}}f_{ki}|k\rangle\langle k|, \quad P_i=\frac{1}{2}\sum_{k\in\mathbb{Z}_{n+1}}p_{ki}|k\rangle\langle k|,$$

$$K(F_i)=\sum_{k\in\mathbb{Z}_{n+1}}K_k(F_i)|k\rangle\langle k|, \tag{II.13b}$$

$$\sum_{k\in\mathbb{Z}_{n+1}}f_{ki}=\sum_{k\in\mathbb{Z}_{n+1}}p_{ki}=\sum_{k\in\mathbb{Z}_{n+1}}K_k(F_i)=0.$$

The quantities $K_k(F_i)$ obey the relations

$$k_k(F_i)-K_{k+1}(F_i)=\frac{f_{ki}+f_{ki+1}}{2}, \tag{II.13c}$$

which agree with the periodicity property $K_k(F_i)=K_{k+n+1}(F_i)$ since F_i is traceless. The loop algebra elements $W_\pm(i)$ are given by

$$W_{\pm}(i) = -K(F_i) + \sum_{k=1}^{n} f_{ki} \mathscr{S}_{\pm}^{k}(\mu_i),$$

$$\mathscr{S}_{\pm}^{k}(\mu_i) = \mathscr{B}_{\pm}^{k}(\mu_i) - \mathscr{B}_{\pm}^{n+1}(\mu_i), \tag{II.13d}$$

$$\mathscr{B}_{\pm}^{k}(\mu_i) = \mp \left(\frac{1}{2} + \frac{\left(\frac{\lambda}{\mu_i}\right)^{\pm(n+1)}}{1-\left(\frac{\lambda}{\mu_i}\right)^{\pm(n+1)}} \right) E^{kk} \mp \sum_{l \gtrless k} \frac{\left(\frac{\lambda}{\mu_i}\right)^{l-k\pm(n+1)}}{1-\left(\frac{\lambda}{\mu_i}\right)^{\pm(n+1)}} E^{kl} \mp \sum_{l \gtrless k} \frac{\left(\frac{\lambda}{\mu_i}\right)^{l-k}}{1-\left(\frac{\lambda}{\mu_i}\right)^{\pm(n+1)}} E^{kl}.$$

In the above expressions λ stands for the loop (or spectral) parameter (see the Appendix). Note that the elements $\mathscr{B}_{\pm}^{k}(\mu_i)$, in contrast to $\mathscr{S}_{\pm}^{k}(\mu_i)$, are not in the loop algebra $\mathrm{sl}(n+1)$. Due to the diagonal (λ-depending) contributions with nonvanishing trace, $\mathscr{B}_{\pm}^{k}(\mu_i)$ are in the loop algebra $\widetilde{\mathrm{gl}}(n+1)$. Taking into account the loop algebra analog of (A13) together with (A23) one gets

$$\mathscr{S}_{\pm}^{k}(\mu_i) = \mp \sum_{r \in \mathbb{Z}_{n+1}} \omega^{r}(\omega^{-rk}-1)\left(\frac{F_0^r}{2} + \sum_{p \gtrless 0} \frac{F_p^r}{\mu_i^p} \right). \tag{II.14}$$

The elements of the Cartan subalgebra $K(F_i)$ (II.13b), (II.13c) can also be expanded in the alternative basis (A13). Therefore, from (II.13d) and the above expansions, we obtain

$$W_{\pm}(i) = \sum_{k \in \mathbb{Z}_{n+1}} f_{ki} W_{\pm}^{k}(\mu_i), \tag{II.15a}$$

where

$$W_{+}^{k}(\mu_i) = -\sum_{r=1}^{n} \left(\frac{1-\omega^{-rk}}{1-\omega^{-r}} F_0^r + \omega^{r(1-k)} \sum_{p \geqslant 1} \frac{F_p^r}{\mu_i^p} \right), \tag{II.15b}$$

$$W_{-}^{k}(\mu_i) = \sum_{r=1}^{n} \left(-\frac{1-\omega^{r(1-k)}}{1-\omega^{-r}} F_0^r + \omega^{r(1-k)} \sum_{p \leqslant -1} \frac{F_p^r}{\mu_i^p} \right). \tag{II.15c}$$

Note that neither $W_{+}^{k}(\mu_i)$ nor $W_{-}^{k}(\mu_i)$ contain contributions belonging to the principal Heisenberg subalgebra. Moreover, the identities

$$W_{+}^{k}(\mu_i) - W_{-}^{k}(\mu_i) = -\sum_{r=1}^{n} \omega^{r(1-k)} \sum_{p \in \mathbb{Z}} \frac{F_p^r}{\mu_i^p} \tag{II.16}$$

are valid. The above expression together with (II.15a)–(II.15c) provides a hint of how to relate the dressing group approach to the group-algebraic methods, developed in Refs. 15 and 16. This relation has been conjectured for general integrable hierarchies which admit a vacuum solution.[13]

Up to now we discussed the properties of the factors $\tilde{g}_{\pm}(k)$ (II.12) which produce single solitons without specifying their relation to the N-soliton solution (II.5a), (II.5b). This relation has been discussed in detail in Ref. 21. Here we only list the results. First of all, it has been shown that the diagonal matrices F_l, $l=1,\ldots,N$ (II.13b) satisfy the identities

$$\begin{aligned} &\sum_{k \in \mathbb{Z}_{n+1}} \omega^{r(1-k)} \left(\frac{\rho_{kl}(\omega^r \mu_{l+1})}{\rho_{kl}(\mu_{l+1})} - \omega^{-r} \frac{\rho_{k+1l}(\omega^r \mu_{l+1})}{\rho_{k+1l}(\mu_{l+1})} \right) \beta_k(F_{l+1}) \\ &\quad = \delta_{r,r_{l+1}}(1-\omega^r) \prod_{a \neq l+1} \frac{\omega^r \mu_{l+1} - \mu_a}{\mu_{l+1} - \mu_a} \prod_{a=1}^{N} \frac{\mu_{+1} + \epsilon_{1a}}{\omega^r \mu_{l+1} + \epsilon_{1a}} \\ &\quad \times \sum_{k \in \mathbb{Z}_{n+1}} \left(1 + \mu_{l+1} \frac{d}{d\lambda} \ln \frac{\rho_{kl}}{\rho_{k+1l}}(\mu_{l+1}) \right) \beta_k(F_{l+1}). \end{aligned} \tag{II.17a}$$

In the above system r is a discrete parameter which belongs to $\mathbb{Z}_{n+1}$; $r_i = 1,\ldots,n$ stand for the

soliton species (II.5a), (II.5b) related to the soliton with the rapiditylike parameter μ_i for $i=1,...,N$. It was noted in Ref. 21 that (II.17a) is a system of algebraic equations which determines the variables

$$\beta_k(F_l)=\exp\left(K_k(F_l)-\frac{f_{kl}}{2}\right). \tag{II.17b}$$

The quantities $K_k(F_l)$ and f_{kl} which appear in the above definition are introduced by (II.13a) and (II.13c). The functions ρ_{al} (II.17a) on λ, $a\in\mathbb{Z}_{n+1}$, $l=1,...,N$ can be considered as components of $n+1$-dimensional column vectors $\rho_l(\lambda)$. The following relations are valid:

$$\begin{aligned}\rho_l(\lambda)&=\mathscr{D}^{(ll-1)}(\lambda)\rho_{l-1}(\lambda),\\ \rho_{l-1}(\lambda)&=\mathscr{D}^{(l-1l)}(\lambda)\rho_l(\lambda),\end{aligned} \tag{II.17c}$$

where $\mathscr{D}^{(ll-1)}(\lambda)$ and $\mathscr{D}^{(l-1l)}(\lambda)$ are $(n+1)\times(n+1)$ matrices $(\mathscr{D}^{(ll-1)}(\lambda)\mathscr{D}^{(l-1l)}(\lambda)=1)$. Their matrix elements are given by

$$\begin{aligned}\mathscr{D}^{(ll-1)}_{ab}(\lambda)&=\frac{\gamma_l\beta_b(F_l)}{(n+1)\rho_{bl-1}(\mu_l)}\sum_{q\in\mathbb{Z}_{n+1}}\omega^{q(a-b)}\frac{\lambda-\omega^q e^{f_{bl}}\mu_l}{\lambda-\omega^q\mu_l}\\ \mathscr{D}^{(l-1l)}_{ab}(\lambda)&=\frac{\rho_{al-1}(\mu_l)}{(n+1)\gamma_l\beta_b(F_l)}\sum_{q\in\mathbb{Z}_{n+1}}\omega^{q(a-b)}\frac{\lambda-\omega^q e^{-f_{bl}}\mu_l}{\lambda-\omega^q\mu_l}\\ \gamma_l&=\left(\prod_{p\in\mathbb{Z}_{n+1}}\rho_{pl-1}(\mu_l)\right)^{1/n+1}.\end{aligned} \tag{II.17d}$$

The first relation (II.17c) together with

$$\rho_{j0}(\lambda)=\frac{1}{n+1} \tag{II.17e}$$

determines uniquely the vectors $\rho_l(\lambda)$. The entries of the diagonal matrices P_l are fixed by the relations[21]

$$\exp\left[K_a(F_{l-1})+\frac{p_{al-1}}{2}\right]=\frac{\rho_{al-1}(\mu_l)}{\gamma_l}. \tag{II.17f}$$

Therefore, Eqs. (II.17a)–(II.17f) provide a recursive method to obtain the factorized expression (II.12) for a dressing group element in the loop group which generate the N-soliton solution (II.5a), (II.5b) from the vacuum. In view of the close relation (II.11) between the dressing group elements in the affine group and in the loop group, the decomposition (II.12) together with the recurrence relations (II.17a)–(II.17f), will play a crucial role in obtaining the vertex operator representation for the soliton tau function. Note also that due to (II.13c), the quantities $\beta_k(F_i)$ (II.17b) satisfy the relation

$$e^{-f_{kl}}=\frac{\beta_k(F_l)}{\beta_{k-1}(F_l)}, \tag{II.18}$$

which resembles (II.4b). However β_k, as it can be seen from (II.17a), (II.17c), and (II.17d), do not satisfy the Hirota bilinear equations (II.4a) in general. This is to say that the fields F_l (II.13b) *do not* satisfy the $A^{(1)}_n$ Toda equations for arbitrary values of l.

The components $\rho_{al}(\lambda)$ of the vectors $\rho_l(\lambda)$ (II.17a), (II.17c), (II.17d) satisfy algebraic and differential equations which will be important in what follows. First of all, combining (II.17c) with (II.17d), we get

$$\sum_{a,r\in\mathbb{Z}_{n+1}} \frac{\omega^{r(a'-a)}}{\lambda-\omega^r\mu_l}\left(\frac{\lambda-\omega^r e^{-f_{al}}\mu_l}{\lambda-\omega^r\mu_l}\frac{\rho_{al}(\lambda)}{\gamma_l\beta_a(F_l)}-\frac{\rho_{al-1}(\lambda)}{\rho_{al-1}(\mu_l)}\right)=0 \tag{II.19}$$

for any $a'\in\mathbb{Z}_{n+1}$. From the above identity and the second equation (II.17d) we deduce the relations

$$\sum_{a\in\mathbb{Z}_{n+1}} \frac{d\mathscr{D}^{(l-1l)}_{a'a}}{d\lambda}(\lambda)\rho_{al}(\lambda)=\frac{\rho_{a'l-1}(\mu_l)}{n+1}\sum_{a,r\in\mathbb{Z}_{n+1}} \frac{\omega^{r(a'-a)}}{\lambda-\omega^r\mu_l}\left(\frac{\rho_{al}(\lambda)}{\gamma_l\beta_a(F_l)}-\frac{\rho_{al-1}(\lambda)}{\rho_{al-1}(\mu_l)}\right). \tag{II.20}$$

In view of (II.17c), the following representation,

$$\rho_j(\lambda)=\mathscr{D}^{(jj-1)}\cdots\mathscr{D}^{(10)}\rho_0(\lambda), \tag{II.21a}$$

takes place. Differentiating the above identity with respect to λ and taking into account that $\rho_0(\lambda)$ (II.17e) does not depend on the spectral parameter, one obtains

$$\frac{d\rho_j}{d\lambda}(\lambda)=\sum_{l=1}^{j}\mathscr{D}^{(jl)}(\lambda)\frac{d\mathscr{D}^{(ll-1)}}{d\lambda}(\lambda)\rho_{l-1}(\lambda), \tag{II.21b}$$

where the matrices $\mathscr{D}^{(jl)}(\lambda)$ for $j\geqslant l$ are given by

$$\begin{aligned}\mathscr{D}^{(jl)}(\lambda)&=\mathscr{D}^{(jj-1)}(\lambda)\cdots\mathscr{D}^{(l+1l)}(\lambda),\\ \rho_j(\lambda)&=\mathscr{D}^{(jl)}(\lambda)\rho_l(\lambda),\\ \mathscr{D}^{(jj)}(\lambda)&=1.\end{aligned} \tag{II.21c}$$

Since $\mathscr{D}^{(l-1l)}$ is the inverse of $\mathscr{D}^{(ll-1)}$, we can rewrite (II.21b) as follows:

$$\frac{d\rho_j}{d\lambda}(\lambda)=-\sum_{l=1}^{j}\mathscr{D}^{(jl-1)}(\lambda)\frac{d\mathscr{D}^{(l-1l)}}{d\lambda}(\lambda)\rho_l(\lambda). \tag{II.21d}$$

Inserting (II.20) into the above identity we obtain

$$\frac{d\rho_{kj}}{d\lambda}(\lambda)=-\frac{1}{n+1}\sum_{l=1}^{j}\sum_{a,a',r\in\mathbb{Z}_{n+1}}\frac{\omega^{r(a'-a)}}{\lambda-\omega^r\mu_l}\mathscr{D}^{(jl-1)}_{ka'}(\lambda)\rho_{a'l-1}(\lambda)\left(\frac{\rho_{al}(\lambda)}{\gamma_l\beta_a(F_l)}-\frac{\rho_{al-1}(\lambda)}{\rho_{al-1}(\mu_l)}\right). \tag{II.22}$$

This result will be used in Sec. V to demonstrate the equivalence between expressions (I.1) and (I.2) for the N-soliton tau functions.

III. FREE FIELD REALIZATION OF THE $A_n^{(1)}$ LIE ALGEBRAS IN THE PRINCIPAL GRADATION

Realizations of infinite dimensional Lie algebras in terms of harmonic oscillators play a crucial role in the representation theory. They are also important in the applications to the string theory and the two-dimensional conformal models (for a review, see Ref. 33). In the simplest case of the affine Lie algebra $A_1^{(1)}=\hat{\mathrm{sl}}(2)$, the free field (or vertex operator) construction was obtained in Ref. 34. This result has been generalized for affine Lie algebras $\hat{\mathscr{G}}$ in the principal gradation[35] when the underlying classical Lie algebras $\mathscr{G}$ are simply laced and the level of the corresponding representation is one. The vertex operator realization of $\hat{\mathscr{G}}$ with $\mathscr{G}$ being nonsimply laced as well as for twisted affine Lie algebras, is also known.[15,16] In the present section we limit our attention on the $A_n^{(1)}$ Lie algebras in the principal gradation only. A detailed discussion of the general case is given in Ref. 10.

As is seen from (II.10), general solutions of the CAT models (II.1a), (II.2), and (II.3) can be written by using the highest weight representations of the affine Lie algebras $\hat{\mathscr{G}}$. The representa-

tion spaces are generated by the action of arbitrary products of negative grade elements (A19) on the highest-weight state $|\Lambda\rangle$. The latter is annihilated by the elements of positive grade of the affine Lie algebra

$$F_k^r|\Lambda\rangle=0, \quad k\geqslant 1, \quad r\in\mathbb{Z}_{n+1}, \tag{III.1a}$$

where we have used the notation (A22). We also recall that $F^0_{p(n+1)}\equiv 0$, $p\in\mathbb{Z}$ since these elements are not in the affine algebra. The highest-weight state is an eigenvector of the subalgebra $\hat{\mathscr{G}}_0$ of the elements of $\mathbb{Z}$ grade zero (A18) and (A19),

$$F_0^r|\Lambda\rangle=\Lambda(F_0^r)|\Lambda\rangle, \quad \hat{d}|\Lambda\rangle=\Lambda(\hat{d})|\Lambda\rangle, \quad \hat{c}|\Lambda\rangle=\Lambda(\hat{c})|\Lambda\rangle. \tag{III.1b}$$

In view of (A23), the generators $\mathscr{E}_k=(n+1)F_k^0$ $(k\neq 0 \bmod(n+1))$ form a Heisenberg subalgebra. It is known as the *principal* Heisenberg subalgebra. Since $\mathscr{E}_k$ are bosonic oscillators, they admit a Fock space representation, built up on the Fock vacuum $|0\rangle$,

$$\begin{aligned} \mathscr{E}_k|0\rangle&=0, \quad k\geqslant 1,\\ \langle 0|\mathscr{E}_k&=0, \quad k\leqslant -1. \end{aligned} \tag{III.2}$$

It is known that when the value of the central charge $\hat{c}$ is 1, all the irreducible highest representations of the Lie algebras $A_n^{(1)}$ are expressed in terms of the elements of the (principal) Heisenberg subalgebra only.[10,34,35] In this section we give an elementary demonstration of this statement.

We start by introducing the notations

$$\mathscr{E}(\mu)=\sum_{\substack{k\in\mathbb{Z}\\ k\neq 0 \bmod(n+1)}} \frac{\mathscr{E}_k}{\mu^k}, \tag{III.3a}$$

$$\begin{aligned} \theta_r(\mu)&=i\sum_{\substack{k\in\mathbb{Z}\\ k\neq 0 \bmod(n+1)}} \frac{\omega^{-rk}-1}{k}\frac{\mathscr{E}_k}{\mu^k},\\ \theta_r(\mu)&=\theta_{r+n+1}(\mu), \quad \theta_0(\mu)=0, \end{aligned} \tag{III.3b}$$

where i is the imaginary unit $i^2=-1$. The above expressions are related by the identity

$$\mathscr{E}(\mu)=-\frac{i}{n+1}\,\mu\,\frac{d}{d\mu}\sum_{r\in\mathbb{Z}_{n+1}}\theta_r(\mu). \tag{III.3c}$$

We also define the bosonic normal product $::$ in the standard way: $\mathscr{E}_k$ with $k\geqslant 1$ are moved to the right, while $\mathscr{E}_k$ with $k\leqslant -1$ are moved to the left in each normal ordered monomial containing generators of the Heisenberg subalgebra only. Since the Heisenberg subalgebra does not contain elements of grade zero, it turns out that each normal ordered monomial of positive (negative) grade annihilates the Fock vacuum $|0\rangle$ ($\langle 0|$) (III.2). Taking into account (III.2) and the Heisenberg commutation relations (A23) we get the contraction identities

$$\begin{aligned} \theta_r(\mu_1)\theta_s(\mu_2)&=\langle\theta_r(\mu_1)\theta_s(\mu_2)\rangle+:\theta_r(\mu_1)\theta_s(\mu_2):,\\ \mathscr{E}(\mu_1)\theta_s(\mu_2)&=\langle\mathscr{E}(\mu_1)\theta_s(\mu_2)\rangle+:\mathscr{E}(\mu_1)\theta_s(\mu_2):,\\ |\mu_1|&>|\mu_2|. \end{aligned} \tag{III.4a}$$

The corresponding vacuum expectation values are given by

$$\langle\theta_r(\mu_1)\theta_s(\mu_2)\rangle=\ln\frac{(\mu_1-\omega^{-r}\mu_2)(\mu_1-\omega^s\mu_2)}{(\mu_1-\mu_2)(\mu_1-\omega^{s-r}\mu_2)}, \tag{III.4b}$$

$$\langle \mathscr{E}(\mu_1)\theta_r(\mu_2)\rangle = i\,\frac{(1-\omega^r)\mu_1\mu_2}{(\mu_1-\mu_2)(\mu_1-\omega^r\mu_2)}.$$

To obtain a realization of the basic representations of the affine Lie algebras $A_n^{(1)}$ we introduce the *vertex* operators

$$V^r(\mu) = :e^{i\theta_r(\mu)}: = \exp\left(\sum_{k\leqslant -1}\ \frac{1-\omega^{-kr}}{k}\,\frac{\mathscr{E}_k}{\mu^k}\right)\exp\left(\sum_{k\geqslant 1}\ \frac{1-\omega^{-kr}}{k}\,\frac{\mathscr{E}_k}{\mu^k}\right)\ . \qquad r=1,\ldots,n. \tag{III.5}$$

Using the Wick theorem one deduces the following operator products:

$$\mathscr{E}(\mu_1)V^r(\mu_2) = \frac{(\omega^r-1)\mu_1\mu_2}{(\mu_1-\mu_2)(\mu_1-\omega^r\mu_2)}\,V^r(\mu_2) + :\mathscr{E}(\mu_1)V^r(\mu_2):$$

$$V^r(\mu_1)V^s(\mu_2) = \frac{(\mu_1-\mu_2)(\mu_1-\omega^{s-r}\mu_2)}{(\mu_1-\omega^{-r}\mu_2)(\mu_1-\omega^s\mu_2)}:e^{i(\theta_r(\mu_1)+\theta_s(\mu_2))}: \qquad |\mu_1|>|\mu_2|. \tag{III.6}$$

In view of (A23) and (III.2) we also get

$$\mathscr{E}(\mu_1)\mathscr{E}(\mu_2) = \frac{\mu_1\mu_2}{(\mu_1-\mu_2)^2} - (n+1)\,\frac{\mu_1^{n+1}\mu_2^{n+1}}{(\mu_1^{n+1}-\mu_2^{n+1})^2} + :\mathscr{E}(\mu_1)\mathscr{E}(\mu_2):, \qquad |\mu_1|>|\mu_2|. \tag{III.7}$$

By using the notion of radial ordering[33] which is an analog of the standard time ordering in the field theory, one can extend the operator products (III.6) and (III.7) to the region $|\mu_1|<|\mu_2|$ in accordance with the ''local'' commutativity relations

$$A(\mu_1)B(\mu_2) = B(\mu_2)A(\mu_1), \quad |\mu_1|\neq|\mu_2|,$$

for $A(\mu)$, $B(\mu)$ being $\mathscr{E}(\mu)$ (III.3a) or $V^r(\mu)$ (III.5). This enables us to use the Laurent expansions

$$V^r(\mu) = \sum_{l\in\mathbb{Z}}\ \frac{V_l^r}{\mu^l}, \quad V_l^r = \oint_{\mathscr{S}^1}\frac{d\mu}{2\pi i\mu}\,\mu^l V^r(\mu). \tag{III.8}$$

Note also that the normal ordered terms which appear on the right-hand side of (III.6) and (III.7) have finite matrix elements when evaluated between normalizable states in the bosonic Fock space.

To compute the commutators of the Laurent coefficients $\mathscr{E}_k$, $k\in\mathbb{Z}\backslash(n+1)\mathbb{Z}$ and V_k^r we shall use the *contour* deformation technique.[33] Let us briefly recall it: Suppose that $A(\mu)$ and $B(\mu)$ are meromorphic operator valued functions on μ which *locally* commute $A(\mu_1)B(\mu_2) = A(\mu_2)B(\mu_1)$ for $|\mu_1|\neq|\mu_2|$. Then the commutator between the Laurent coefficients

$$A_k = \oint_{\mathscr{S}^1}\frac{\mu^k d\mu}{2\pi i\mu}\,A(\mu), \quad B_l = \oint_{S^1}\frac{\mu^l d\mu}{2\pi i\mu}\,B(\mu) \tag{III.9a}$$

is given by

$$[A_k,B_l] = \left(\oint_{|\mu_1|>|\mu_2|} - \oint_{|\mu_1|<|\mu_2|}\right)\mu_1^k\,\frac{d\mu_1}{2\pi i\mu_1}\,\mu_2^l\,\frac{d\mu_2}{2\pi i\mu_2}\,A(\mu_1)B(\mu_2). \tag{III.9b}$$

Using the above identity, the Cauchy theorem, and taking into account (III.6) one gets

$$[\mathscr{E}_k, V_l^r] = \left(\oint_{|\mu_1|>|\mu_2|} - \oint_{|\mu_1|<|\mu_2|} \right) \mu_1^k \frac{d\mu_1}{2\pi i \mu_1} \mu_2^l \frac{d\mu_2}{2\pi i \mu_2} \mathscr{E}(\mu_1) V^r(\mu_2)$$

$$= \oint \mu_2^{l+1} \frac{d\mu_2}{2\pi i \mu_2} \left(\oint_{C_{\mu_2}} - \oint_{C_{\omega^r \mu_2}} \right) \mu_1^{k+1} \frac{d\mu_1}{2\pi i \mu_1} (\omega^r - 1) \frac{V^r(\mu_2)}{(\mu_1 - \mu_2)(\mu_1 - \omega^r \mu_2)}$$

$$= (\omega^{kr} - 1) V_{k+l}^r, \tag{III.10a}$$

where C_μ stands for a small counterclockwise oriented contour which surrounds the point μ. Similarly from the second equation (III.6) we obtain the commutators

$$[V_k^r, V_l^s] = \frac{(\omega^r - 1)(\omega^s - 1)}{(\omega^{r+s} - 1)} (\omega^{sk} - \omega^{rl}) V_{k+l}^{r+s}, \qquad r + s \neq 0 \bmod(n+1), \tag{III.10b}$$

$$[V_k^r, V_l^{-r}] = -\omega^r (1 - \omega^{-r})^2 ((\omega^{-rk} - \omega^{rl}) \mathscr{E}_{k+l} + k \omega^{-rk} \delta_{k+l,0}). \tag{III.10c}$$

In deriving the above equations we have also used the identity

$$\theta_r(\omega^p \mu) = \theta_{p+r}(\mu) - \theta_p(\mu),$$

which is a straightforward consequence from (III.3b). Using the contour deformation technique one also concludes that (III.7) is equivalent to the Heisenberg commutators (A23). Therefore we conclude that $\mathscr{E}_k$, $k \in \mathbb{Z}(n+1)\mathbb{Z}$ together with V_k^r, $r = 1,...,n$, $k \in \mathbb{Z}$ form a Lie algebra. The last is *isomorphic* to the Lie algebra $A_n^{(1)}$ in the principal gradation. To see this it is sufficient to set

$$F_k^r = \frac{\omega^{rl} V_k^r}{(n+1)(\omega^r - 1)}, \quad l \in \mathbb{Z}_{n+1}, r = 1,..., n, \tag{III.11}$$

and to verify that (III.10a)–(III.10c) together with (A15) coincide with (A21). In the above equation, $l \in \mathbb{Z}_{n+1}$ is an additional parameter. Note that the map

$$\mathscr{E}_k \to \mathscr{E}_k, \quad V_k^r \to \omega^r V_k^r, \quad r = 1,..., n \tag{III.12}$$

is an automorphism of order $(n+1)$ of the Lie algebra (III.10a)–(III.10c), (A23). It is clear that this automorphism is *outer*, and therefore the representations of $A_n^{(1)}$ corresponding to different values of the additional parameter l (III.11) are *inequivalent*. Due to (III.2), (III.5), and (III.11) we obtain highest weight representations characterized by

$$F_0^r |\Lambda_l\rangle = \frac{\omega^{rl}}{(n+1)(\omega^r - 1)} |\Lambda_l\rangle, \quad \hat{c} = |\Lambda_l\rangle = |\Lambda_l\rangle, \quad \hat{d} |\Lambda_l\rangle = 0, \tag{III.13}$$

where the highest-weight state coincides with the Fock vacuum (III.2). Since $\hat{c} = 1$, we conclude that representations constructed by us are the $n+1$ *inequivalent fundamental* representations[10] of the Lie algebra $A_n^{(1)}$ in the principal gradation.

We proceed by calculating the tau functions (II.9) in terms of the components of the field (II.1d), (II.2) with $\eta = 0$. Also taking into account the last identity (A13), one obtains

$$\Phi = \frac{1}{2} \sum_{\substack{1 \leqslant r \leqslant n \\ k \in \mathbb{Z}_{n+1}}} \omega^{r(1-k)} \varphi_k F_0^r + \frac{\zeta}{2(n+1)} \hat{c}. \tag{III.14}$$

Combining the above expression with (III.13) we derive

$$\langle \Lambda_l | \Phi | \Lambda_l \rangle = \frac{\zeta}{2(n+1)} + \frac{1}{2(n+1)} \sum_{\substack{1 \leqslant r \leqslant n \\ k \in \mathbb{Z}_{n+1}}} \frac{\omega^{r(l+1-k)}}{\omega^r - 1} \varphi_k. \tag{III.15}$$

Therefore the group-algebraic tau functions (II.9) satisfy the identities

$$\frac{\tau_{\Lambda_k}(\Phi)}{\tau_{\Lambda_{k-1}}(\Phi)} = e^{-\varphi_k}, \quad k \in \mathbb{Z}_{n+1}, \tag{III.16a}$$

$$\prod_{k \in \mathbb{Z}_{n+1}} \tau_{\Lambda_k}(\Phi) = e^{-\zeta}. \tag{III.16b}$$

Comparing the above identities with the definition (II.4a)–(II.4c) of the Hirota tau functions we get the relations

$$\tau_{\Lambda_k}(\Phi) = \exp\left(-\frac{\zeta_0}{n+1}\right)\tau_k(\Phi), \tag{III.17}$$

where ζ_0 is the vacuum value (II.4d) of the field ζ (II.2).

IV. DERIVATION OF THE VERTEX OPERATOR REPRESENTATION FOR ONE-SOLITON TAU FUNCTIONS

In the present section we focus our attention on the equivalence between the representations (I.1) and (I.2) for one-soliton solutions of the A_n CAT equations. The existence of one-solitons, i.e., solitary waves which propagate without changing their shape, is not a distinguishing property of the integrable evolution equations. As a counterexample, one can quote the existence of kink solution of the φ^4 model in $1+1$ dimensions.[1] Monosoliton solutions of the affine Toda equations based on arbitrary simple Lie algebra are also known.[36]

In accordance with (II.5a), (II.5b) the evolution of a single soliton is dictated by

$$e^{-\varphi_k} = \frac{1+X_k}{1+X_{k-1}}, \quad X_k = -\frac{\omega^{rk}}{c}\frac{e(\mu)}{e(\omega^r\mu)}, \tag{IV.1}$$

where μ is the unique rapidity like parameter, $r=1,...,n$ is the soliton species, and c stands for the constant which appear on the right-hand side of (II.5a). From the above identities it follows that the corresponding tau functions (II.4a), (II.4b) are given by

$$\tau_k = 1 + X_k. \tag{IV.2}$$

In this section we shall demonstrate that the representations (I.1) and (I.2) for the one-soliton tau functions associated with the fundamental representations of the Lie algebra $A_n^{(1)}$ are equivalent. Denote by $\tilde{g}_{\pm}(\mu)$ the dressing group elements which generate the one-soliton solution (IV.1) of the $A_n^{(1)}$ Toda model from the vacuum. Taking into account (II.12) and (II.13a) we get

$$\begin{gathered}\tilde{g}(\mu) = \tilde{g}_-^{-1}(\mu)\tilde{g}_+(\mu) = e^{-W_-(\Phi)}.e^{W_+(\Phi)}, \\ \Phi = \frac{1}{2}\sum_{k \in \mathbb{Z}_{n+1}} \varphi_k E^{kk},\end{gathered} \tag{IV.3}$$

where φ_k are the components of the $A_n^{(1)}$ Toda field and $W_{\pm}(\Phi) = \Sigma_{k \in \mathbb{Z}_{n+1}}\, \varphi_k W_{\pm}^k(\mu)$ according to (II.15a). The elements $W_{\pm}(\mu)$ has been introduced by (II.15b), (II.15c). In what follows we shall use the representations

$$W_{\pm}(\Phi) = \sum_{k=1}^{n} \varphi_k \tilde{W}_{\pm}^k(\mu), \quad \tilde{W}_{\pm}^k(\mu) = W_{\pm}^k(\mu) - W_{\pm}^{n+1}(\mu), \tag{IV.4a}$$

$$\tilde{W}^k_{\pm}(\mu)=-\sum_{k=1}^{n}\ \frac{1-\omega^{-rk}}{1-\omega^{-r}}\ \oint_{\lambda\gtrless|\mu|}\frac{d\lambda}{2\pi i\lambda}\,\frac{\omega^r\lambda-\mu}{\lambda-\mu}\,F^r(\lambda), \tag{IV.4b}$$

where

$$F^r(\lambda)=\sum_p\ \frac{F^r_p}{\lambda^p},\quad F^r_p=\oint\frac{\lambda^p d\lambda}{2\pi i\lambda}\,F^r(\lambda). \tag{IV.4c}$$

To calculate the element (IV.3) in the fundamental representation (III.11), (III.13) of the affine Lie algebra $A^{(1)}_n$ we first introduce the notation

$$h(\alpha)=e^{-\alpha W_-(\Phi)}.e^{\alpha W_+(\Phi)},\quad h(\alpha)=\sum_{l=0}^{\infty}\ \frac{\alpha^l}{l!}\,\frac{d^l}{d\alpha^l}\,h(0), \tag{IV.5a}$$

$$\frac{d^l}{d\alpha^l}\,h(0)=-W_-(\Phi)\,\frac{d^{l-1}}{d\alpha^{l-1}}\,h(0)+\frac{d^{l-1}}{d\alpha^{l-1}}\,h(0).W_+(0)=\sum_{k=0}^{l}\,(-)^k\binom{l}{k}W^k(\Phi)W^{l-k}_+(\Phi). \tag{IV.5b}$$

Using (IV.4a), (IV.4b) and since the matrix (II.1d) is traceless, one can also write

$$W_{\pm}(\Phi)=\sum_{r=1}^{n}\ \frac{\hat{\varphi}_r}{1-\omega^{-r}}\ \oint_{|\lambda|\gtrless|\mu|}\frac{d\lambda}{2\pi i\lambda}\,\frac{\omega^r\lambda-\mu}{\lambda-\mu}\,F^r(\lambda), \tag{IV.6}$$

where we have introduced the discrete Fourier transform of φ_k ($\varphi_k=\varphi_{k+n+1}$) according to

$$\hat{\varphi}_r=\sum_{k\in\mathbb{Z}_{n+1}}\omega^{-kr}\varphi_k,\quad \varphi_k=\frac{1}{n+1}\sum_{r\in\mathbb{Z}_{n+1}}\omega^{kr}\hat{\varphi}_r. \tag{IV.7a}$$

Let us also recall the identity

$$\sum_{k\in\mathbb{Z}_{n+1}}\omega^{k(i-j)}=(n+1)\,\delta^{(n+1)}_{i,j}, \tag{IV.7b}$$

where the $\delta^{(n+1)}_{i,j}$ is the Kronecker symbol on the cyclic group: It vanishes for $i-j\neq 0 \bmod (n+1)$ and $\delta^{(n+1)}_{i,j}=1$ if $i-j=0 \bmod (n+1)$.

We proceed by calculating explicitly the Taylor expansion (IV.5a). First of all, we note that

$$\begin{aligned}\frac{dh}{d\alpha}(0)&=W_+(\Phi)-W_-(\Phi)\\&=-\sum_{r=1}^{n}\ \frac{\hat{\varphi}_r}{1-\omega^{-r}}\ \oint_{C_\mu}\frac{d\lambda}{2\pi i\lambda}\,\frac{\omega^r\lambda-\mu}{\lambda-\mu}\,F^r(\lambda)\\&=-\sum_{r=1}^{n}\ \hat{\varphi}_r\omega^r F^r(\mu).\end{aligned} \tag{IV.8}$$

The above identities agree with (II.15a), (II.16). To calculate the higher derivatives of $h(\alpha)$ (IV.5a) we first observe that due to (III.6) one gets

$$
\begin{aligned}
&W_-(\Phi)V^r(\mu)-V^r(\mu)W_+(\Phi)\\
&\quad=\frac{1}{n+1}\sum_{s=1}^{n}\frac{\omega^{s(k+1)}\hat{\varphi}_s}{(\omega^s-1)^2}\left(\oint_{|\lambda|>|\mu|}-\oint_{|\lambda|<|\mu|}\right)\frac{d\lambda}{2\pi i\lambda}\,\frac{\omega^s\lambda-\mu}{\lambda-\mu}\,V^s(\lambda)V^r(\mu)\\
&\quad=\frac{1}{n+1}\sum_{s=1}^{n}\frac{\omega^{s(k+2)}\hat{\varphi}_s}{(\omega^s-1)^2}\left(\oint_{|\lambda|>|\mu|}-\oint_{|\lambda|<|\mu|}\right)\frac{d\lambda}{2\pi i\lambda}\,\frac{\lambda-\omega^{r-s}\mu}{\lambda-\omega^r\mu}:e^{i(\theta_s(\lambda)+\theta_r(\mu))}\\
&\quad=\frac{1}{n+1}\sum_{s=1}^{n}\frac{\omega^{s(k+1)}}{\omega^s-1}\,\hat{\varphi}_sV^{r+s}(\mu),\quad k\in\mathbb{Z}_{n+1},
\end{aligned}
\tag{IV.9}
$$

in the irreducible representation of $A_n^{(1)}$ with a highest-weight $|\Lambda_k\rangle$ (III.11)–(III.13). Using the above identity together with (IV.5b), it is easy to prove inductively that

$$
\frac{d^lh}{d\alpha^l}(0)=(-)^l\sum_{s_1,\ldots,s_l=1}^{n}\prod_{j=1}^{l}\Psi_{s_j}(\Phi)\cdot V^{s_1+\cdots+s_l}(\mu),
\tag{IV.10a}
$$

$$
\Psi_s(\Phi)=\frac{1}{n+1}\,\frac{\omega^{s(k+1)}}{\omega^s-1}\,\hat{\varphi}_s,
\tag{IV.10b}
$$

where we have omitted the dependence on the parameter $k\in\mathbb{Z}_{n+1}$ which label the different fundamental representations of the affine Lie algebra. Substituting the last result into the Taylor expansion (IV.5a), it turns out that (IV.3) is given by the following intermediate expression:

$$
\tilde{g}(\mu)=h(1)=\sum_{l=0}^{\infty}\frac{(-1)^l}{l!}\sum_{s_1,\ldots,s_l=1}^{n}\prod_{j=1}^{l}\Psi_{s_j}(\Phi)\cdot V^{s_1+\cdots+s_l}(\mu).
\tag{IV.11}
$$

In order to write the right-hand side of the above identity in a more convenient form, we recall that in view of (III.3b) and (III.5), the vertex operators are periodic functions on the upper case index $V^s(\mu)=V^{s+n+1}(\mu)$ with $V^0(\mu)=1$. This observation suggests that we introduce the commutative associative algebra $\mathscr{F}$ of (complex) dimension $n+1$. It is generated by the elements V^s, $s\in\mathbb{Z}_{n+1}$. The multiplication $\star$ in $\mathscr{F}$ is defined by

$$
V^r\star V^s=V^{r+s}.
\tag{IV.12}
$$

Due to the periodicity condition, the element $V^0=V^{n+1}$ is the unity of $\mathscr{F}$. It is worthwhile to note that (IV.12) describes the ''*fusion rules*'' of a class of rational conformal field theories.[37] They correspond to a theory of a free massless scalar field in two dimensions which is compactified on circle with a rational value of the square of the radius. The algebra $\mathscr{F}$ can be endowed with a symmetric bilinear invariant form

$$
\begin{aligned}
&\langle a,b\rangle_{\mathscr{F}}=\langle b,a\rangle_{\mathscr{F}},\\
&\langle a,b\star c\rangle_{\mathscr{F}}=\langle a\star b,c\rangle_{\mathscr{F}},\quad a,b,c\in\mathscr{F},
\end{aligned}
\tag{IV.13a}
$$

which is uniquely fixed by

$$
\langle V^r,V^s\rangle_{\mathscr{F}}=\delta_{r+s,0}^{(n+1)}.
\tag{IV.13b}
$$

The algebra $\mathscr{F}$ together with the above bilinear form is an example of a Fröbenius algebra. Fröbenius algebras and their deformations appear in the study of the topological field theory.[38] We go back to the problem of simplifying expression (IV.11). Using the multiplication rules (IV.12) we can rewrite (IV.11) in the following form:

$$
\tilde{g}=\mathscr{F}\exp\left\{-\sum_{s=1}^{n}\Psi_s(\Phi)V^s\right\},
\tag{IV.14}
$$

where the symbol $\mathscr{F}$ means that the exponential is taken in the Fröbenius algebra (IV.12)–(IV.13b). In the above expression the dependence on μ of the elements V^s, $s=1,...,n$ was skipped. Note that the Fourier transformation (IV.7a) diagonalizes the ''fusion rules'' (IV.12),

$$V^r \star \hat{V}^s = \omega^{rs}\hat{V}^s, \quad \hat{V}^p \star \hat{V}^q = (n+1)\,\delta_{p,q}^{(n+1)}\hat{V}^p. \tag{IV.15}$$

Substituting the above expression into (IV.14) we obtain

$$\tilde{g} = \frac{1}{n+1}\sum_{p\in\mathbb{Z}_{n+1}} e^{-\Sigma_{s=1}^{n}\omega^{ps}\Psi_s(\Phi)}\hat{V}^p = \frac{1}{n+1}\sum_{p,r\in\mathbb{Z}_{n+1}} \omega^{-pr}\cdot e^{\chi_p(\Phi)}V^r, \tag{IV.16a}$$

$$\chi_p(\Phi) = -\sum_{s=1}^{n}\omega^{ps}\Psi_s(\Phi), \quad p\in\mathbb{Z}_{n+1}. \tag{IV.16b}$$

Due to the identity (IV.7b), the above quantities satisfy the restriction

$$\sum_{p\in\mathbb{Z}_{n+1}}\chi_p(\Phi) = 0. \tag{IV.17a}$$

Moreover, taking into account (IV.10b), we get the recurrence relations

$$\chi_p(\Phi) - \chi_{p+1}(\Phi) = \varphi_{p+k+1}, \tag{IV.17b}$$

where $k\in\mathbb{Z}_{n+1}$ is the parameter which labels the $n+1$ unequivalent fundamental representations of $A_n^{(1)}$ (III.11), (III.13). Comparing (IV.17a) and (IV.17b) with the definition (II.4a)–(II.4c) of the Hirota tau functions, we get

$$e^{\chi_p(\Phi)} = \frac{\tau_{p+k}(\Phi)}{(\Pi_{l\in\mathbb{Z}_{n+1}}\tau_l(\Phi))1/n+1}, \quad p\in\mathbb{Z}_{n+1}. \tag{IV.18}$$

Substituting back the above expression into (IV.16a) and taking into account (IV.1) as well as the vertex operator realization of the affine Lie algebra $A_n^{(1)}$ (III.5), (III.11), and (III.8), we conclude that

$$\begin{aligned}\tilde{g}(\mu) &= \frac{1}{n+1}\sum_{p,q\in\mathbb{Z}_{n+1}}\omega^{-pq}e^{\chi_p(\Phi)}V^q(\mu)\\ &= \frac{1}{(n+1)\Pi_{l\in\mathbb{Z}_{n+1}}\tau_l^{1/(n+1)}(\Phi)}\sum_{p,q\in\mathbb{Z}_{n+1}}(\omega^{-pq}+\omega^{p(r-q)+rk}X_0)V^q(\mu)\\ &= \frac{1}{\Pi_{l\in\mathbb{Z}_{n+1}}\tau_l^{1/(n+1)}(\Phi)}\,(1+(n+1)(\omega^r-1)X_0F^r(\mu)).\end{aligned} \tag{IV.19a}$$

From the above identity, (II.10), (II.11), and (II.4c) one derives

$$\begin{aligned}\frac{\tau_{\Lambda_k}(\Phi)}{\tau_{\Lambda_k}(\Phi_0)} &= \exp\left(\frac{\zeta_0-\zeta}{n+1}\right)\langle\Lambda_k|\tilde{g}_-(\mu)\tilde{g}_+(\mu)|\Lambda_k\rangle\\ &= \frac{\exp\left(\dfrac{\zeta_0-\zeta}{n+1}\right)}{\Pi_{l\in\mathbb{Z}_{(n+1)}}\tau_l^{1/(n+1)}(\Phi)}\,\langle\Lambda_k|(1+(n+1)(\omega^r-1)X_0F^r(\mu))|\Lambda_k\rangle\\ &= \langle\Lambda_k|(1+(n+1)(\omega^r-1)X_0F^r(\mu))|\Lambda_k\rangle.\end{aligned} \tag{IV.19b}$$

Therefore, the representations (I.1) and (I.2) are equivalent in the one-soliton case.

V. *N*-SOLITON TAU FUNCTIONS AND VERTEX OPERATORS

The existence of *N*-soliton solutions is a peculiar property of the integrable field theories. From another point of view,[29] soliton solutions provide a relation of classes of solutions of integrable nonlinear partial differential equations with certain dynamical systems with a finite number of degrees of freedom. The latter are known as *N*-body integrable systems (for a review, see Ref. 28). The *N*-soliton solutions, in the in- and out-limit, become asymptotically a superposition of monosolitons separated in the space. Moreover, the transformation from the in-variables to the out-variables is canonical. The underlying generating function of this transformation is known as the classical *S*-matrix.[6] An intriguing property of the classical *S*-matrix is that it is represented as a sum of terms, each of them representing a two-particle scattering. It is well known that the last property admits a generalization valid within the quantum theory: The quantum *S*-matrix which describes collision of solitons in a given quantum integrable model is a product of factors describing two particle interactions.[7] *N*-soliton solutions of the Toda field theories have been obtained both within the ISM[22] and by using the Hirota method.[39]

We recall that the algebraic *N*-soliton solutions (II.5a), (II.5b) in the $A_n^{(1)}$ Toda models are generated from the vacuum by the dressing group elements (II.12)–(II.15c). Due to the factorized expression (II.12), we can also write

$$\begin{aligned}\tilde{g}=\tilde{g}_-^{-1}\tilde{g}_+&=\tilde{g}_-^{-1}(1)\ldots\tilde{g}_-^{-1}(N)\cdot\tilde{g}_+(N)\cdots\tilde{g}_+(1)\\ &=\tilde{g}(1)\cdot \mathrm{Ad}\,\tilde{g}_+^{-1}(1)(\tilde{g}(2))\cdots\mathrm{Ad}(\tilde{g}_+^{-1}(1))\cdots\tilde{g}_+^{-1}(N-1))\cdot(\tilde{g}(N)),\\ \tilde{g}(i)&=\tilde{g}_-^{-1}(i)\tilde{g}_+(i).\end{aligned} \tag{V.1}$$

In the above expression and in what follows we shall assume that the rapiditylike parameters μ_i, $i=1,\ldots,N$ corresponding to the factors (II.13a)–(II.13d) are radially ordered $|\mu_1|>|\mu_2|>\cdots>|\mu_N|$. In view of (IV.16a), (IV.16b), and the vertex operator construction of the affine Lie algebra $A_n^{(1)}$ (III.5), (III.8), (III.11), to calculate (V.1) in the fundamental representations, one first has to obtain an expression for the adjoint action of $\tilde{g}_+(i)$ on the elements of the alternative basis (A22). To do that we first note that diagonal traceless $(n+1)\times(n+1)$ matrices $D=\Sigma_{k\in\mathbb{Z}_{n+1}}\,d_k|k\rangle\langle k|$ are written in the alternative basis (A13), (A15) as follows:

$$D=\sum_{s\in\mathbb{Z}_{n+1}}\omega^s\tilde{d}_sF_0^s, \tag{V.2}$$

where $\hat{d}_s$, $s\in\mathbb{Z}_{n+1}$ is the discrete Fourier transformation (IV.7a) of the (diagonal) entries of D. Note that F_0^0 is *not* an element of the Lie algebra A_n. However, since D is traceless, it is clear that the coefficient $\hat{d}_0$ which multiplies F_0^0 vanishes. The above expression can be considered as an element either of the classical Lie algebra A_n or of the affine algebra $A_n^{(1)}$. Taking into account the commutation relations (A21) together with (A23), it is easy to get

$$[F^r(\mu),D]=\frac{1}{n+1}\sum_{\substack{k\in\mathbb{Z}\\ s\in\mathbb{Z}_{n+1}}}\frac{\omega^s(\omega^{sk}-1)\hat{d}_s}{\mu^k}F_k^{r+s} \tag{V.3a}$$

from where it is easy to get

$$\begin{aligned}e^{-D}F^r(\mu)e^D&=\sum_{l=0}^{\infty}\frac{(-1)^l}{l!}\,\mathrm{ad}^lD\cdot F^r(\mu)\\ &=\sum_{k\in\mathbb{Z}}\frac{1}{\mu^k}\sum_{l=0}^{\infty}\frac{1}{l!}\sum_{s_1,\ldots,s_l\in\mathbb{Z}_{n+1}}\prod_{j=1}^{l}\omega^{s_j}\frac{\omega^{s_jk}-1}{n+1}\,\hat{d}_{s_j}F_k^{r+s_1+\cdots+s_l}.\end{aligned} \tag{V.3b}$$

The right-hand side of Eq. (V.3b) is of the same form as (IV.11). This suggests using the approach developed in Sec. IV to write (V.3b) explicitly as a linear combination of the affine algebra elements. In order to do that we first introduce the following (reducible) representation of the algebra $\mathscr{F}$ (IV.12):

$$V^s(F_k^r)=F_k^{r+s}, \quad r,s\in\mathbb{Z}_{n+1}, \quad k\in\mathbb{Z}. \tag{V.4}$$

In complete analogy with the derivation of the expression (IV.14), we conclude that

$$e^{-D}F^r(\mu)e^D=\sum_{k\in\mathbb{Z}_{n+1}}\frac{1}{\mu^k}\mathscr{F}\exp\left(\sum_{s\in\mathbb{Z}_{n+1}}\frac{\omega^s(\omega^{sk}-1)}{n+1}\hat{d}_sV^s\right)(F_k^r). \tag{V.5}$$

As in Sec. IV, we diagonalize the operators V^s (V.4),

$$V^s(\hat{F}_k^p)=\omega^{ps}\hat{F}_k^p, \quad \hat{F}_k^p=\sum_r\omega^{-pr}F_k^r, \quad F_k^r=\frac{1}{n+1}\sum_p\omega^{pr}\hat{F}_k^p. \tag{V.6}$$

Combining (V.5) with (V.6) we obtain

$$\begin{aligned}
e^{-D}F^r(\mu)e^D&=\frac{1}{n+1}\sum_{\substack{k\in\mathbb{Z}\\ p\in\mathbb{Z}_{n+1}}}\frac{\omega^{pr}}{\mu^k}e^{d_{k+p+1}-d_{p+1}}\hat{F}_k^p\\
&=\frac{1}{n+1}\sum_{\substack{k\in\mathbb{Z}\\ p,s\in\mathbb{Z}_{n+1}}}\frac{\omega^{p(r-s)}}{\mu^k}e^{d_{k+p+1}-d_{p+1}}F_k^s\\
&=\frac{1}{n+1}\sum_{l,p,s\in\mathbb{Z}_{n+1}}\omega^{p(r-s)}\frac{e^{d_{l+p+1}}}{e^{d_{p+1}}}\sum_{k\in\mathbb{Z}}\frac{F^s_{k(n+1)+l}}{\mu^{k(n+1)+l}}\\
&=\frac{1}{(n+1)^2}\sum_{l,p,q,s\in\mathbb{Z}_{n+1}}\omega^{p(r-s)+lq}\frac{e^{d_{l+p+1}}}{e^{d_{p+1}}}F^s(\omega^q\mu).
\end{aligned} \tag{V.7}$$

As a consequence of the above identity and (II.17f), one obtains

$$e^{-\mathrm{ad}(K(F_i)+P_i)}F^r(\omega^a\lambda)=\sum_{s,q\in\mathbb{Z}_{n+1}}U^{ra}_{sq}(i)F^s(\omega^q\lambda), \tag{V.8a}$$

where $K(F_i)$ and P_i are diagonal traceless matrices (II.13b) and

$$\begin{aligned}
U^{ra}_{sq}(i)&=M^{r+a}_{s+q}(i)N^a_q(i),\\
M^r_s(i)&=\frac{\omega^{s-r}}{n+1}\sum_{p\in\mathbb{Z}_{n+1}}\frac{\omega^{p(r-s)}}{\rho_{pi}(\mu_i+1)},\\
N^a_q(i)&=\frac{\omega^{a-q}}{n+1}\sum_{p\in\mathbb{Z}_{n+1}}\omega^{p(q-a)}\rho_{pi}(\mu_{i+1}).
\end{aligned} \tag{V.8b}$$

In what follows we shall also need the commutators $[W^k_+(\mu_i),F^s(\lambda)]$ where the affine algebra elements $W^k_+(\mu_i)$ and $F^s(\lambda)$ are given by (II.15b) and (IV.4c), respectively. Using (A21) with $\hat{c}=1$ we get

$$[W_+^k(\mu_i),F^s(\lambda)]=-\frac{1}{n+1}\sum_{r=1}^{n}\frac{(1-\omega^{-rk})\mu_i-\omega^s(1-\omega^{r(1-k)})\lambda}{(1-\omega^{-r})(\mu_i-\omega^s\lambda)}F^{r+s}(\lambda)$$
$$+\frac{1}{n+1}\sum_{r=1}^{n}\frac{(1-\omega^{-rk})\mu_i-\omega^{-r}(1-\omega^{r(1-k)})\lambda}{(1-\omega^{-r})(\mu_i-\omega^{-r}\lambda)}F^{r+s}(\omega^{-r}\lambda)$$
$$-\frac{1}{(n+1)^2}\sum_{r=1}^{n}\frac{\omega^{-rk}\lambda\mu_i}{(\mu_i-\omega^{-r}\lambda)^2}\delta_{r+s,0}^{(n+1)},\quad |\mu_i|>|\lambda|. \tag{V.9a}$$

From the above expression, (II.13b), and (II.15a) we also derive the commutator

$$[W_+(i),F^s(\lambda)]=\sum_{k\in\mathbb{Z}_{n+1}}f_{ki}[W_+^k(\mu_i),F^s(\lambda)]$$
$$=\frac{1}{n+1}\sum_{r=1}^{n}\frac{\hat{f}_{ri}}{1-\omega^{-r}}\left(\frac{\mu_i-\omega^{r+s}\lambda}{\mu_i-\omega^s\lambda}F^{r+s}(\lambda)-\frac{\mu_i-\lambda}{\mu_i-\omega^{-r}\lambda}F^{r+s}(\omega^{-r}\lambda)\right)$$
$$-\frac{\hat{f}_{-si}}{(n+1)^2}\frac{\lambda\mu_i}{(\mu_i-\omega^{-s})^2},\quad |\mu_i|>|\lambda|,\quad s\in\mathbb{Z}_{n+1}, \tag{V.9b}$$

where $\hat{f}_{ri}$ stands for the discrete Fourier transformation (IV.7a) of f_{ki}, $k\in\mathbb{Z}_{n+1}$. Using (V.9b) one proves inductively

$$\mathrm{ad}^l\ W_+(i)F^s(\lambda)=\frac{\mu_i-\lambda}{\mu_i-\omega^s\lambda}\sum_{k=0}^{l}(-)^{l-k}\binom{l}{k}$$
$$\times\sum_{r_1,\ldots,r_l=1}^{n}\frac{\mu_i-\omega^{r_1+\cdots+r_k+s}\lambda}{\mu_i-\omega^{-r_{k+1}-\cdots-r_l}\lambda}\prod_{p=1}^{l}\psi_{r_p i}F^{r_1+\cdots+r_l+s}(\omega^{-r_{k+1}-\cdots-r_l}\lambda)$$
$$+\frac{\mu_i(\mu_i-\lambda)}{(n+1)^2(\mu_i-\omega^s\lambda)}\sum_{k=0}^{l-1}(-)^{l-k}\binom{l-1}{k}$$
$$\times\sum_{r_1,\ldots,r_{l-1}=1}^{n}\frac{\lambda\omega^{-r_{k+1}-\cdots-r_{l-1}}}{(\mu_i-\omega^{-r_{k+1}-\cdots-r_{l-1}}\lambda)(\mu_i-\omega^{r_l+\cdots+r_k+s}\lambda)}$$
$$\cdot\hat{f}_{-r_1\cdots-r_{l-1}-si}\prod_{p=1}^{l-1}\psi_{r_p i},\quad l\geqslant 1,$$

$$\psi_{ri}=\frac{\hat{f}_{ri}}{(n+1)(1-\omega^{-r})},\quad r=1,\ldots,n. \tag{V.10}$$

Note that after the change $\hat{f}_{ri}\to\hat{\varphi}_i$, the quantities ψ_{ri} coincide up to phase with (IV.10b). To write the right-hand side of the above identity in a compact form, we consider the commutative associative algebra $\mathscr{A}=\mathscr{F}\times\mathscr{T}$. We recall that $\mathscr{F}$ is the Fröbenius algebra (IV.12)–(IV.13b) while $\mathscr{T}$ is spanned on the elements $\mathbb{T}^r$, $r\in\mathbb{Z}_{n+1}$. The multiplication in $\mathscr{A}$ is introduced by

$$V^r*V^s=V^{r+s},\quad \mathbb{T}^r*\mathbb{S}^s=\mathbb{T}^{r+s},\quad V^r*\mathbb{T}^s=\mathbb{T}^s*V^r. \tag{V.11a}$$

We shall also need specific representations of $\mathscr{A}$. They are defined by the relations

$$V^r(g(\lambda)F^s(\lambda))=g(\lambda)V^r(F^s(\lambda))=g(\lambda)F^{r+s}(\lambda),$$
$$\mathbb{T}^r(g(\lambda)F^s(\lambda))=g(\omega^{-r}\lambda)\mathbb{T}^r(F^s(\lambda))=g(\omega^{-r}\lambda)F^s(\omega^{-r}\lambda), \tag{V.11b}$$

where g is a function on the complex parameter λ. It is also assumed that the algebra $\mathscr{A}$ acts trivially on μ_i (V.10). Using (V.11a) and (V.11b), it is easy to check that (V.10) can be written as follows:

$$
\begin{aligned}
(-)^l \mathrm{ad}^l\ W_+(i)F^s(\lambda) = \frac{\mu_i-\lambda}{\mu_i-\omega^s\lambda}\Bigg(\Bigg(\sum_{r-1}^{n}\psi_{ri}V^r(\mathbb{T}^r-1)\Bigg)_*^l \frac{\mu_i F^s(\lambda)}{\mu_i-\lambda} \\
-\omega^s\Bigg(\sum_{r=1}^{n}\omega^r\psi_{ri}V^r(\mathbb{T}^r-1)\Bigg)_*^l \frac{\lambda F^s(\lambda)}{\mu_i-\lambda}\Bigg) + \frac{\mu_i(\mu_i-\lambda)}{(n+1)^2(\mu_i-\omega^s\lambda)} \\
\times\Bigg\langle \sum_{p\in\mathbb{Z}_{n+1}}\hat{f}_{pi}V^p, \Bigg(\sum_{r=1}^{n}\psi_{ri}V^r(1\otimes_{\mathscr{F}}\mathbb{T}^r-\mathbb{T}^{-r}\otimes_{\mathscr{F}}1)\Bigg)_*^{l-1} V^s\Bigg\rangle_{\mathscr{F}} \\
\times\frac{1}{\mu_i-\omega^s\lambda}\otimes\frac{\lambda}{\mu_i-\lambda},
\end{aligned}
\tag{V.12}
$$

where $\langle\ ,\ \rangle_{\mathscr{F}}$ is the invariant bilinear form in the Fröbenius algebra $\mathscr{F}$ (IV.12)–(IV.13b). We have also used the notation

$$
(\mathscr{P})_*^l = \underbrace{\mathscr{P}*\ldots*\mathscr{P}}_{l}
\tag{V.13a}
$$

to indicate that the lth power of $\mathscr{P}\in\mathscr{A}$ is taken with respect to the multiplication (V.11a). In what follows we shall use the operators

$$
\mathscr{A}\ \exp(\mathscr{P}) = \sum_{l=0}^{\infty}\frac{(\mathscr{P})_*^l}{l!}, \quad \mathscr{A}\ \frac{\exp(\mathscr{P})-1}{\mathscr{P}} = \sum_{l=1}^{\infty}\frac{(\mathscr{P})_*^{l-1}}{l!}.
\tag{V.13b}
$$

Note that the second of the above operators has eigenvalue one when applied on the zero modes of $\mathscr{P}$. Taking into account (V.12) together with (V.13a), (V.13b) we obtain

$$
\begin{aligned}
e^{-\mathrm{ad}\ W_+(i)\cdot}F^s(\lambda) = \frac{\mu_i-\lambda}{\mu_i-\omega^s\lambda}\Bigg(\mathscr{A}\ \exp\Bigg(\sum_{r=1}^{n}\psi_{ri}V^r(\mathbb{T}^r-1)\Bigg)\frac{\mu_i F^s(\lambda)}{\mu_i-\lambda} - \omega^s.\mathscr{A} \\
\times\exp\Bigg(\sum_{r=1}^{n}\omega^r\psi_{ri}V^r(\mathbb{T}^r-1)\Bigg)\frac{\lambda F^s(\lambda)}{\mu_i-\lambda}\Bigg) + \frac{\mu_i(\mu_i-\lambda)}{(n+1)^2(\mu_i-\omega^s\lambda)} \\
\times\Bigg\langle \sum_{p\in\mathbb{Z}_{n+1}}\hat{f}_{p_i}V^p, \mathscr{A}\ \frac{\exp(\Sigma_{r=1}^{n}\psi_{ri}V^r(1\otimes_{\mathscr{F}}\mathbb{T}^r-\mathbb{T}^{-r}\otimes_{\mathscr{F}}1))-1}{\Sigma_{r=1}^{n}\psi_{ri}V^r(1\otimes_{\mathscr{F}}\mathbb{T}^r-\mathbb{T}^{-r}\otimes_{\mathscr{F}}1)}\ V^s\Bigg\rangle_{\mathscr{F}} \\
\times\frac{1}{\mu_i-\omega^s\lambda}\otimes\frac{\lambda}{\mu_i-\lambda}.
\end{aligned}
\tag{V.14}
$$

In view of the above equation, we shall need the (reducible) representations (V.11b) of $\mathscr{A}$ (V.11a) spanned on the vectors $g(\omega^r\lambda)F^s(\omega^r\lambda)$, $r,s\in\mathbb{Z}_{n+1}$ for g being an arbitrary meromorphic function on λ. It is easy to check that

$$
G^{pq}(\lambda) = \sum_{r,s\in\mathbb{Z}_{n+1}}\omega^{-ps+rq}g(\omega^r\lambda)F^s(\omega^r\lambda)
\tag{V.15a}
$$

are eigenvectors of $\mathscr{T}$,

$$
V^kG^{pq}(\lambda) = \omega^{pk}G^{pq}(\lambda), \quad \mathbb{T}^kG^{pq}(\lambda) = \omega^{kq}G^{pq}(\lambda).
\tag{V.15b}
$$

Due to the identity (IV.7b), the inverse of (V.15a) is given by

$$g(\omega^r\lambda)F^s(\omega^r\lambda)=\frac{1}{(n+1)^2}\sum_{p,q\in\mathbb{Z}_{n+1}}\omega^{ps-rq}G^{pq}(\lambda). \tag{V.15c}$$

Note also that

$$\sum_{r=1}^{n}\omega^{rp}\psi_{ri}=\frac{f_{pi}}{2}-K_p(F_i), \tag{V.16a}$$

where the diagonal traceless matrices F_i and $K(F_i)$ have been defined by (II.13b). To obtain the above identity we have also used the relations (II.13c). Taking into account (II.17b) we can write

$$\exp\left(-\sum_{r=1}^{n}\omega^{rp}\psi_{ri}\right)=\beta_p(F_i) \tag{V.16b}$$

substituting back (V.15a)–(V.16b) into (V.14) one gets

$$\begin{aligned}e^{-\mathrm{ad}\,W_+(i)}\cdot F^s(\lambda)=&\frac{\mu_i-\lambda}{(n+1)^2(\mu_i-\omega^s\lambda)}\\&\times\sum_{p,q,r',s'\in\mathbb{Z}_{n+1}}\omega^{p(s-s')+r'q}\,\frac{\mu_i-\omega^{r'+s'}\lambda}{\mu_i-\omega^{r'}\lambda}\,\frac{\beta_p(F_i)}{\beta_{p+q}(F_i)}\,F^{s'}(\omega^{r'}\lambda)\\&+\frac{\lambda\mu_i(\mu_i-\lambda)}{(n+1)^4(\mu_i-\omega^s\lambda)}\sum_{a,q,q'p,p'\in\mathbb{Z}_{n+1}}f_{ai}\,\frac{\omega^{as+pq+p'(q'+1)}}{(\mu_i-\omega^{p+s}\lambda)(\mu_i-\omega^{p'}\lambda)}\\&\times\frac{\beta_{a-q}(F_i)-\beta_{a+q'}(F_i)}{\beta_{a+q'}(F_i)(\ln\,\beta_{a-q}(F_i)-\ln\,\beta_{a+q'}(F_i))}.\end{aligned} \tag{V.17}$$

Therefore, the adjoint action of the affine group element $e^{-W_+(i)}$ on the affine algebra, can be written as follows:

$$\begin{aligned}e^{-\mathrm{ad}\,W_+(i)}F^s(\omega^c\lambda)=&\sum_{q,v\in\mathbb{Z}_{n+1}}W^{sc}_{qv}[F_i](i;\lambda)F^q(\omega^v\lambda)+Z^{sc}[F_i](i,\lambda)\\&\text{for}\quad|\mu_i|>|\lambda|.\end{aligned} \tag{V.18a}$$

Due to (II.15a), $W_+(i)$ are linear in the entries f_{ki} of the diagonal matrices F_i (II.13b). This is to say that the relations

$$\begin{aligned}&\sum_{q,v\in\mathbb{Z}_{n+1}}W^{sc}_{qv}[F_i](i;\lambda)W^{qv}_{kl}[-F_i](i;\lambda)=\delta^{(n+1)}_{s,k}\delta^{(n+1)}_{c,l},\\&\sum_{q,v\in\mathbb{Z}_{n+1}}W^{sc}_{qv}[F_i](i;\lambda)Z^{qv}[-F_i](i;\lambda)+Z^{sc}[F_i](i;\lambda)=0\end{aligned} \tag{V.18b}$$

are valid. Comparing (V.17) with (V.18a) we obtain

$$\begin{aligned}&W^{sc}_{qv}[F_i](i;\lambda)=K^{s+c}_{q+v}[F_i](i;\lambda)L^c_v[F_i](i;\lambda),\\&K^s_q[F_i](i;\lambda)=\frac{1}{n+1}\,\frac{\mu_i-\omega^q\lambda}{\mu_i-\omega^s\lambda}\sum_{p\in\mathbb{Z}_{n+1}}\omega^{p(s-q)}\beta_p(F_i),\\&L^c_v[F_i](i;\lambda)=\frac{1}{n+1}\,\frac{\mu_i-\omega^c\lambda}{\mu_i-\omega^v\lambda}\sum_{p\in\mathbb{Z}_{n+1}}\omega^{p(v-c)}\beta_p(-F_i),\end{aligned} \tag{V.18c}$$

and

$$Z^{sc}[F_i](i,\lambda)=\frac{\omega^c\lambda\mu_i(\mu_i-\omega^c\lambda)}{(n+1)^4(\mu_i-\omega^{s+c}\lambda)}\sum_{\substack{a,b,p\\ v,v'}} f_{pi}\,\frac{\omega^{sa+(p-a)v+(b+1-p)v'+c(a-b-1)}}{(\mu_i-\omega^v\lambda)(\mu_i-\omega^{v'}\lambda)}$$
$$\times\frac{\beta_a(F_i)-\beta_b(F_i)}{\beta_b(F_i)(\ln\,\beta_a(F_i)-\ln\,\beta_b(F_i))}. \tag{V.18d}$$

Taking into account that the matrices F_i (II.13b) are traceless and the identity

$$\sum_{\substack{p\in\mathbb{Z}_{n+1}\\ 1\leqslant r\leqslant n}}\frac{\omega^{(p-a)r}}{1-\omega^r}\,f_{pi}=-(n+1)\ln\,\beta_a(F_i),$$

we conclude that (V.18d) can be alternatively written in the form

$$Z^{sc}[F_i](i,\lambda)=\frac{\lambda(\mu_i-\omega^c\lambda)}{(n+1)^3(\mu_i-\omega^{s+c}\lambda)}\sum_{a,b,v\in\mathbb{Z}_{n+1}}\frac{\omega^{a(s+c+v)-b(v+c)}}{\lambda-\omega^v\mu_i}\,\frac{\beta_a(F_i)}{\beta_b(F_i)}$$
$$+\frac{\lambda\omega^c\delta^{(n+1)}_{s,0}}{(n+1)(\mu_i-\omega^{s+c}\lambda)}. \tag{V.19a}$$

After certain trivial algebraic manipulations involving the first term of the above expression and using (II.18), (II.17d) we arrive at the result

$$Z^{sc}[-F_i](i,\lambda)=\frac{\lambda\omega^c}{\gamma_i(n+1)^2(\omega^{s+c}\lambda-\mu_i)}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(s+c)a-bc}\mathscr{D}^{(ii-1)}_{ab}(\lambda)\,\frac{\rho_{bi-1}(\mu_i)}{\beta_a(F_i)}$$
$$+\frac{\lambda\omega^c\delta^{(n+1)}_{s,0}}{(n+1)(\mu_i-\omega^{s+c}\lambda)}, \tag{V.19b}$$

where γ_i are the constants defined by (II.17d).

Equations (V.8a), (V.8b), and (V.18a) provide an expression for the adjoint action of the element (II.13a) on the affine Lie algebra

$$\tilde{g}_+^{-1}(i)F^s(\omega^c\lambda)\tilde{g}_+(i)=\sum_{r,v\in\mathbb{Z}_{n+1}}(R^{sc}_{rv}(i;\lambda)F^r(\omega^v\lambda)+U^{sc}_{rv}(i)Z^{rv}[F_i](i;\lambda))$$
$$=\sum_{r,v\in\mathbb{Z}_{n+1}}R^{sc}_{rv}(i;\lambda)(F^r(\omega^v\lambda)-Z^{rv}[-F_i](i;\lambda)),\quad |\mu_i|>|\lambda|, \tag{V.20a}$$

where $R^{sc}_{rv}(i;\lambda)$ are the elements of the matrix $R(j;\mu_i)=U(j)W(j;\mu_i)$ which acts on tensor product $\mathbb{C}^{n+1}\otimes\mathbb{C}^{n+1}$,

$$R^{sc}_{rv}(i;\lambda)=\sum_{p,q\in\mathbb{Z}_{n+1}}U^{sc}_{pq}(i)W^{pq}_{rv}(i;\lambda)=P^c_v(i;\lambda)Q^{s+c}_{r+v}(i;\lambda),$$

$$P^c_v(i;\lambda)=\frac{1}{n+1}\sum_{p\in\mathbb{Z}_{n+1}}\omega^{(p-1)(v-c)}\,\frac{\lambda-\omega^{-v}e^{-f_{pi}}\mu_i}{\lambda-\omega^{-v}\mu_i}\,\frac{\rho_{pi}(\mu_{i+1})}{\beta_p(F_i)}, \tag{V.20b}$$

$$Q^s_r(i;\lambda)=\frac{\mu_i-\omega^r\lambda}{(n+1)^2}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(a-1)s-br}\,\frac{\beta_b(F_i)}{\rho_{ai}(\mu_{i+1})}\sum_{p\in\mathbb{Z}_{n+1}}\frac{\omega^{(b+1-a)p}}{\mu_i-\omega^p\lambda}.$$

Note that the second identity (V.20a) is a consequence of (V.18b). Comparing the above equations with (II.17d) and taking into account (II.18) we obtain the following alternative expressions:

$$P_v^c(i;\lambda)=\frac{\gamma_i}{n+1}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(a-1)v-(b-1)c}\,\frac{1}{\rho_{ai-1}(\mu_i)}\,\mathscr{D}_{ab}^{(i-1i)}(\lambda)\rho_{bi}(\mu_{i+1})$$

$$Q_r^s(i;\lambda)=\frac{1}{(n+1)\gamma_i}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(a-1)s-(b-1)r}\,\frac{1}{\rho_{ai}(\mu_{i+1})}\,\mathscr{D}_{ab}^{(ii-1)}(\lambda)\rho_{bi-1}(\mu_i), \tag{V.21}$$

which together with (IV.7b), (II.21c) yield

$$(P(j;\lambda)\cdots P(k;\lambda))_v^c=\frac{\gamma_j\cdots\gamma_k}{n+1}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(a-1)v-(b-1)c}\,\frac{1}{\rho_{ak-1}(\mu_k)}\,\mathscr{D}_{ab}^{(k-1j)}(\lambda)\rho_{bj}(\mu_{j+1})$$

$$(Q(j;\lambda)\cdots Q(k;\lambda))_r^s=\frac{1}{(n+1)\gamma_j\cdots\gamma_k}\sum_{a,b\in\mathbb{Z}_{n+1}}\omega^{(a-1)s-(b-1)r}\,\frac{1}{\rho_{aj}(\mu_{j+1})}\,\mathscr{D}_{ab}^{(jk-1)}(\lambda)\rho_{bk-1}(\mu_k).$$

$$\text{for}\quad k\leqslant j. \tag{V.22a}$$

Therefore, as a consequence of (V.20b), (II.21c) and the above identities we get

$$(R(j;\mu_{j+1})\cdots R(k;\mu_{j+1}))_{rv}^{p0}=\frac{1}{(n+1)^2}\sum_{a,a',b'\in\mathbb{Z}_{n+1}}\omega^{(a-1)v+(a'-1)p-(b'-1)(r+v)}$$
$$\times\frac{\rho_{ak-1}(\mu_{j+1})}{\rho_{ak-1}(\mu_k)}\,\mathscr{D}_{a'b'}^{(jk-1)}(\mu_{j+1})\,\frac{\rho_{b'k-1}(\mu_k)}{\rho_{a'j}(\mu_{j+1})}. \tag{V.22b}$$

On the other hand, using (II.17d) and (II.21c), it is not difficult to show that

$$\mathscr{D}_{ab}^{(jk)}(\omega^r\lambda)=\omega^{r(a-b)}\mathscr{D}_{ab}^{(jk)}(\lambda).$$

Inserting this identity into (V.22b) and taking into account (II.17e) one obtains

$$(R(j;\mu_{j+1})\cdots R(1;\mu_{j+1}))_{rv}^{p0}=\frac{\delta_{v,0}^{(n+1)}}{n+1}\sum_{a\in\mathbb{Z}_{n+1}}\omega^{(a-1)(p-r)}\,\frac{\rho_{aj}(\omega^r\mu_{j+1})}{\rho_{aj}(\mu_{j+1})}$$

and therefore

$$\sum_{p\in\mathbb{Z}_{n+1}}\omega^{p(1-k)}(R(j;\mu_{j+1})\cdots R(1;\mu_{j+1}))_{rv}^{p0}=\omega^{r(1-k)}\delta_{v,0}^{(n+1)}\,\frac{\rho_{kj}(\omega^r\mu_{j+1})}{\rho_{kj}(\mu_{j+1})}. \tag{V.23a}$$

Using similar arguments and (II.17b)–(II.22) we also prove the identity

$$\sum_{\substack{l<l<j\\ p\in\mathbb{Z}_{n+1}}}\omega^{p(1-k)}(R(j;\mu_{j+1})\cdots R(l;\mu_{j+1})Z[-F_l](l;\mu_{j+1}))^{p0}=-\frac{\mu_{j+1}}{n+1}\cdot\frac{d\,\ln\rho_{kj}}{d\lambda}(\mu_{j+1}), \tag{V.23b}$$

where the functions $Z^{sc}[\pm F_i]$ have been defined by (V.18a), (V.18b).

Now we are ready to calculate the product (V.1). Inserting back (V.18a) and (V.18b) into (V.1) we get

$$\mathrm{Ad}(\tilde{g}_+^{-1}(1)\cdots\tilde{g}_+^{-1}(i-1))F^p(\mu_i)=\sum_{r,v\in\mathbb{Z}_{n+1}}(R(i-1;\mu_i)\cdots R(1;\mu_i))_{rv}^{p0}F^r(\omega^v\mu_i)$$
$$-\sum_{j=1}^{i-1}(R(i-1;\mu_i)\cdots R(j;\mu_i)Z[-F_j](j;\mu_i))^{p0}. \tag{V.24}$$

On the other hand one can repeat the procedure developed in Sec. IV to obtain a result analogous to (IV.16a), (IV.16b),

$$\tilde{g}(i)=\tilde{g}_{-}^{-1}(i)\tilde{g}_{+}(i)=\frac{\hat{\beta}_0(F_i)}{n+1}+\sum_{p\in\mathbb{Z}_{n+1}}(\omega^p-1)\hat{\beta}_p(F_i)F^p(\mu_i) \tag{V.25}$$

where $\hat{\beta}_k(F_i)$, $k\in\mathbb{Z}_{n+1}$ is the discrete Fourier transformation (IV.7a) of $\beta_k(F_i)$ (II.17b). Inserting back the above expression into (V.24) and taking into account (V.23a), (V.23b), and (II.17a) we conclude that

$$\mathrm{Ad}(\tilde{g}_{+}^{-1}(1)\cdots\tilde{g}_{+}^{-1}(i-1))\tilde{g}(i)=Y_i(1+X_iF^{r_i}(\mu_i)),$$

$$X_i=(n+1)(1-\omega^{r_i})\prod_{a\neq i}\frac{\omega^{r_i}\mu_i-\mu_a}{\mu_i-\mu_a}\prod_{a=1}^{N}\frac{\mu_i+\epsilon_{1a}}{\omega^{r_i}\mu_i+\epsilon_{1a}}, \tag{V.26a}$$

$$Y_i=\frac{1}{n+1}\sum_{k\in\mathbb{Z}_{n+1}}\left(1+\mu_i\frac{d}{d\lambda}\ln\frac{\rho_{ki-1}}{\rho_{k+1i-1}}(\mu_i)\right)\beta_k(F_i).$$

Note that due to (II.5a), the quantities X_i depend exponentially on the light cone variables x^+ and x^-. Combining (V.26a), (II.10), and (II.11) we deduce the relation

$$\frac{\tau_\Lambda(\Phi)}{\tau_\Lambda(\Phi_0)}=\exp\left(\frac{\zeta_0-\zeta}{a+1}\right)\prod_{i=1}^{N}Y_i\langle\Lambda|(1+X_1F^{r_1}(\mu_1))\cdots(1+X_NF^{r_N}(\mu_N))|\Lambda\rangle \tag{V.26b}$$

for any fundamental representation with highest weight vector $|\Lambda\rangle$ of the affine Lie algebra $A_n^{(1)}$. Therefore, the representations (I.1) and (I.2) coincide provided that

$$\exp\left(\frac{\zeta-\zeta_0}{n+1}\right)=\prod_{i=1}^{N}Y_i. \tag{V.27}$$

We hope to go back to the proof of the above equation elsewhere.

APPENDIX: THE AFFINE LIE ALGEBRA $A_n^{(1)}$ IN THE PRINCIPAL GRADATION

The purpose of this Appendix is to resume some basic facts from the theory of the Lie algebras. We will focus our attention on the simple Lie algebras A_n and on the untwisted affine Lie algebras $A_n^{(1)}$. More detailed exposition can be found in Refs. 10, 15, and 16.

We recall that the Lie algebra sl$(n+1)$, $n\geqslant 1$ is the set of the traceless $(n+1)\times(n+1)$ matrices. Within the Cartan classification they are known as A_n Lie algebras. The Cartan subalgebra consists of all traceless linear combinations of the diagonal elementary matrices $E^{ii}=|i\rangle\langle i|$ $(i=1\cdots n+1)$. The root system can be embedded into the $(n+1)$-dimensional Euclidean space. Fixing an orthonormalized basis e_i, the roots are $\alpha_{ij}=e_i-e_j$, $i\neq j$ $(i,j=1,...,n+1)$. To each root one associates a step operator $E^{\alpha_{ij}}=E^{ij}=|i\rangle\langle j|$ $(i\neq j)$ which is an eigenvector of the adjoint action of the Cartan subalgebra $\mathscr{H}$. As simple roots one chooses the vectors $\alpha_i=e_i-e_{i+1}$, $i=1,...,n$. The related step operators satisfy the commutation relations

$$[H_\xi,E^{\pm\alpha_i}]=\pm\alpha_i\cdot\xi E^{\pm\alpha_i}=\pm(\xi_i-\xi_{i+1})E^{\pm\alpha_i},$$

$$[E^{\alpha_i}E^{-\alpha_j}]=\delta_{ij}H_{\alpha_i}, \tag{A.1}$$

$$H_\xi=\sum_i\xi_i|i\rangle\langle i|.$$

The generic step operators are obtained by successive commutators of E^{α_i} and their transposed $E^{-\alpha_i}$.

In the theory of the Lie algebras it is important to study their finite order inner automorphisms. The general theory has been developed by Kac[10] and reviewed in Ref. 12. In this paper we shall only use a special inner automorphism σ of the simple Lie algebra A_n of order $n+1$ ($\sigma^{n+1}=1$). Before introducing it, we recall that the fundamental weights are

$$\lambda_i=\sum_{k=1}^{i} e_k-\frac{i}{n+1}\sum_{k=1}^{n+1} e_k,$$
$$2\frac{\alpha_i\cdot\lambda_j}{\alpha_i\cdot\alpha_i}=\delta_{ij},\quad i,j=1,...,n. \tag{A2}$$

We also set $\rho=\Sigma_{i=1}^{n}\lambda_i$ and define[10,11]

$$\sigma(X)=SXS^{-1},$$
$$S=\exp\left(2\pi i\,\frac{H_\rho}{n+1}\right), \tag{A3}$$

for an arbitrary element X of the Lie algebra A_n. Note that in the defining representation, the element S which implements the automorphism σ takes the following form:

$$S=\omega^{n/2}\sum_{k=1}^{n+1}\omega^{1-k}E^{kk},$$
$$\omega=\exp\left(\frac{2\pi i}{n+1}\right). \tag{A4}$$

From the commutation relations (A1) and the above identity, one concludes that σ acts trivially on the elements of $\mathscr{H}$ and as a multiplication by phase on the step operators:

$$\sigma(H_\xi)=H_\xi,$$
$$\sigma(E^{\alpha_{kl}})=\omega^{\alpha_{kl}\cdot\rho}E^{\alpha_{kl}}=\omega^{l-k}E^{\alpha_{kl}}. \tag{A5}$$

The Lie algebra $\mathscr{G}=A_n$ together with the automorphism σ is a $\mathbb{Z}_{n+1}$ graded algebra

$$\mathscr{G}=\oplus_{k\in Z_{n+1}}\mathscr{G}_k,$$
$$\sigma(\mathscr{G}_k)=\omega^k\mathscr{G}_k, \tag{A6}$$
$$[\mathscr{G}_k,\mathscr{G}_l]\subseteq\mathscr{G}_{k+l}.$$

It is well known that for a given simple Lie algebra $\mathscr{G}$, its Cartan subalgebra is fixed up to a conjugation by elements belonging to the corresponding Lie group. In particular, instead of $\mathscr{H}$, one can introduce an *alternative* Cartan subalgebra $\mathscr{H}'$, spanned on the mutually commuting generators:

$$\mathscr{E}_i=\sum_{k=1}^{n+1-i}E^{kk+i}+\sum_{k=1}^{i}E^{n+1+k-ik}=\sum_{k\in Z_{n+1}}|k\rangle\langle k+i|. \tag{A7}$$

To show that the above elements actually generate certain Cartan subalgebra, it suffices to note that the matrix with entries

$$\Omega_{ij}=\omega^{(i-1)(j-1)},$$
$$\Omega_{ij}^{-1}=\frac{\omega^{-(i-1)(j-1)}}{n+1} \tag{A8}$$

diagonalizes $\mathscr{E}_i$ (A7),

$$\Omega^{i-1}\mathscr{E}_i\Omega=\sum_{k=1}^{n+1}\omega^{i(k-1)}E^{kk}. \tag{A9}$$

It is worthwhile to note that the alternative Cartan subalgebra generators are eigenvectors of the inner automorphism σ,

$$\sigma(\mathscr{E}_i)=\omega^i\mathscr{E}_i. \tag{A10}$$

For general simple Lie algebras, it is known[10,15] that the eigenvalues of the corresponding automorphism σ, when restricted to the alternative Cartan algebra, are in correspondence to the Betti numbers. To complete the alternative basis of the Lie algebra A_n, we introduce the generators

$$\begin{aligned} F^i&=\Omega E^{i+1\,i}\Omega^{-1}, \quad i=1,...,n,\\ [\mathscr{E}_i,F^j]&=(\omega^{ij}-1)F^j. \end{aligned} \tag{A11}$$

Due to the grade decomposition (A6) and the above commutation relations one gets

$$\begin{aligned} F^i&=\sum_{k\in Z_{n+1}}F^i_k,\\ \sigma(F^i_k)&=\omega^kF^i_k,\\ [\mathscr{E}_i,F^j_k]&=(\omega^{ij}-1)F^j_{k+i}. \end{aligned} \tag{A12}$$

We thus end up with a basis which is formed by the alternative Cartan generators $\mathscr{E}_i$ (A7) and F^i_k, $i=1,...,n$, $k\in\mathbb{Z}_{n+1}$. Due to (A8) and (A11) the transformation which gives back the Cartan–Weyl basis is

$$\begin{aligned} E^{il}&=\frac{\mathscr{E}_{l-i}}{n+1}+\sum_{r=1}^{n}\omega^{r(1-i)}F^r_{l-i}, \quad i<l,\\ E^{il}&=\frac{\mathscr{E}_{n+1+l-i}}{n+1}+\sum_{r=1}^{n}\omega^{r(1-i)}F^r_{n+1+l-i}, \quad i>l,\\ E^{ii}-E^{n+1\,n+1}&=\sum_{r=1}^{n}\omega^r(\omega^{-ri}-1)F^r_0. \end{aligned} \tag{A13}$$

The commutation relations in the alternative basis are completed by

$$[F^r_i,F^s_l]=\frac{\omega^{sk}-\omega^{rl}}{n+1}F^{r+s}_{k+l}. \tag{A14}$$

Introducing notation

$$F^r_k=\begin{cases} F^r_k, & r=1,...,n, \quad k\in\mathbb{Z}_{n+1}\\ \dfrac{1}{n+1}\mathscr{E}_k & r=0, \quad k=1,...,n, \end{cases} \tag{A15}$$

we see that the commutation relations in the alternative basis assume the uniform expression (A14).

The A_n Lie algebras are equipped with a nondegenerated invariant scalar product $(X,Y)=\mathrm{tr}(X\cdot Y)$. The trace is taken in the defining representation. In the alternative basis (A15) this scalar product is given by

$$(F_k^r, F_l^s) = \frac{\omega^{sk}}{n+1}\,\delta_{k+l,0}^{(n+1)}, \tag{A16}$$

where $\delta_{k,l}^{(n+1)}$ is the delta function in the cyclic group $\mathbb{Z}_{n+1}$. To treat integrable evolution equations, one has to extend the classical Lie algebras by introducing the spectral (or loop) parameter.[11] This implies that the Lax connection belongs to the loop Lie algebra $\tilde{\mathcal{G}} = \mathbb{C}[\lambda, \lambda^{-1}] \otimes \mathcal{G}$. In other words, the loop algebra is the set of the Laurent series with coefficients in the corresponding (classical) Lie algebra $\mathcal{G}$. Therefore $\tilde{\mathcal{G}}$ is spanned on the elements $X_k = \lambda^k X$, $k \in \mathbb{Z}$, $X \in \mathcal{G}$. The Lie bracket is

$$[X_k, Y_l] = [X, Y]_{k+l}. \tag{A17}$$

Loop algebras $\tilde{\mathcal{G}}$ possess central extension,[10] known as affine (or Kac–Moody) algebras $\hat{\mathcal{G}} = \tilde{\mathcal{G}} \oplus \mathbb{C}\hat{c} \oplus \mathbb{C}\hat{d}$,

$$\begin{gathered}
[X_k, Y_l] = [X, Y]_{k+l} + \frac{k}{(n+1)}\,\hat{c}\,\delta_{k+l,0}(X, Y), \\
[\hat{d}, X_k] = kX_k, \\
[\hat{c}, \hat{\mathcal{G}}] = 0.
\end{gathered} \tag{A18}$$

The normalization factor which multiplies the central element $\hat{c}$ is chosen for convenience. The derivation $\hat{d}$ can be used to define a $\mathbb{Z}$ gradation in $\hat{\mathcal{G}}$:

$$\begin{gathered}
\hat{\mathcal{G}} = \oplus_{k \in \mathbb{Z}} \hat{\mathcal{G}}_k, \\
[\hat{d}, \hat{\mathcal{G}}_k] = k\hat{\mathcal{G}}_k.
\end{gathered} \tag{A19}$$

Since σ is an automorphism of the underlying classical Lie algebra $\mathcal{G} = A_n = \mathrm{sl}(n+1)$, it is clear that the commutators (A18) together with the restriction

$$\begin{gathered}
X(\omega\lambda) = \sigma X(\lambda), \\
X(\lambda) = \sum_{l \in \mathbb{Z}} \lambda^l X_l, \\
\sigma X(\lambda) = \sum_l \lambda^l \sigma(X_l),
\end{gathered} \tag{A20}$$

still define a Lie algebra. In the literature[10,12,15] it is known as the affine Lie algebra $A_n^{(1)}$ in the principle gradation. Taking into account (A14), (A16), (A18), together with the above restrictions, we obtain the commutation relations

$$\begin{gathered}
[F_k^r, F_l^s] = \frac{\omega^{sk} - \omega^{rl}}{n+1}\,F_{k+l}^{r+s} + \frac{k\omega^{rl}}{(n+1)^2}\,\hat{c}\,\delta_{r+s,0}^{(n+1)}\,\delta_{k+l,0}, \\
r, s \in \mathbb{Z}_{n+1}, \quad k, l \in \mathbb{Z}.
\end{gathered} \tag{A21}$$

Note that the generators F_k^0 with $k = 0 \bmod(n+1)$ are not in the affine Lie algebra $A_n^{(1)}$. We thus end up with the following alternative basis:

$$\begin{gathered}
F_k^r, \quad r = 1, \ldots, n, \quad k \in \mathbb{Z}, \\
F_k^0, \quad k \in \mathbb{Z}, \quad k \neq 0 \mod(n+1)
\end{gathered} \tag{A22}$$

of the $A_n^{(1)}$ Lie algebra in the principal gradation. The generators F_k^0, $k \neq 0 (\mathrm{mod})$ form a basis of Heisenberg subalgebra in the *principal* gradation.[10–15] In analogy with (A15), we will also use the following generators:

$$\mathscr{E}_k = (n+1) F_k^0,$$
$$[\mathscr{E}_k, \mathscr{E}_l] = k \hat{c} \, \delta_{k+l,0}, \quad k \neq 0 (\mathrm{mod}\ \ n+1). \tag{A23}$$

From the above commutation relations we conclude that $\mathscr{E}_{\pm k}$, $k \geqslant 1$ form an infinite set of noninteracting bosonic oscillators.

ACKNOWLEDGMENTS

It is a pleasure to thank L. A. Ferreira and M. A. C. Kneipp for discussions on the subject. R.P. is also grateful to J. P. Zubelli for bringing to his attention the monograph.[23] Two of us, H.B. and G.C., are supported by CNPq-Brazil while R.P. is supported by FAPERJ-Rio de Janeiro. We deeply acknowledge both foundations for invaluable financial support.

[1] R. Rajaraman, *Solitons and Instantons* (North-Holland, Amsterdam, 1982).
[2] L. D. Faddeev and V. E. Korepin, Phys. Rep. **42**, 1 (1978).
[3] C. Nash and S. Sen, *Topology and Geometry for Physicists* (Academic, New York, 1983); M. Nakahara, *Geometry, Topology and Physics* (IUP, Bristol, 1992); M. Göckeler and T. Sohucker, *Differential Geometry, Gauge Theories, and Gravity* (Cambridge University Press, Cambridge, 1990).
[4] S. Coleman, Phys. Rev. D **11**, 2088 (1975); S. Mandelstam, *ibid.* **11**, 3026 (1975).
[5] N. Seiberg and E. Witten, Nucl. Phys. B **426**, 19 (1994); erratum **430**, 485 (1994); **431**, 484 (1994).
[6] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Method* (SIAM, Philadelphia, 1981); S. Novikov, S. Manakov, L. P. Pitaevsky, and V. E. Zakharov, *Theory of Solitons* (Consultants Bureau, New York, 1984); L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer, New York, 1987).
[7] A. B. Zamolodchikov and Al. B. Zamolodchikov, Ann. Phys. (N.Y.) **120**, 253 (1979).
[8] M. Toda, Phys. Rep. **18**, 1 (1975).
[9] D. I. Olive and N. Turok, Nucl. Phys. B **215**, 470 (1983); **220**, 491 (1983); **257**, 277 (1985), **265**, 469 (1986).
[10] V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1990).
[11] V. G. Drinfeld and V. Sokolov, Sov. Math. Dokl. **23**, 457 (1981).
[12] N. Burroghs, M. de Groot, T. Hollowood, and L. Miramontes, Phys. Lett. B **277**, 89 (1992); Commun. Math. Phys. **153**, 187 (1993); M. de Groot, T. Hollowood, and L. Miramontes, *ibid.* **145**, 57 (1992).
[13] L. A. Ferreira, L. Miramontes, and J. S. Guillén, J. Math. Phys. **38**, 882 (1997).
[14] T. Hollowood, Nucl. Phys. B **384**, 523 (1992).
[15] D. I. Olive, M. V. Saveliev, and J. W. Underwood, Phys. Lett. B **311**, 117 (1993); D. I. Olive, N. Turok, and J. W. Underwood, Nucl. Phys. B **401**, 663 (1993); **409**, 509 (1993); M. A. C. Kneipp and D. I. Olive, *ibid.* **408**, 565 (1993).
[16] M. A. C. Kneipp and D. I. Olive, Commun. Math. Phys. **177**, 561 (1996).
[17] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, in *Non-linear Integrable Systems*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983); Publ. RIMS **18**, 1077 (1982).
[18] M. Semenov-Tian-Shansky, Publ. RIMS **21**, 1237 (1985); Poisson Lie Groups, Quantum Duality Principle, and the Quantum Double, hep-th/9304042.
[19] O. Babelon and D. Bernard, Phys. Lett. B **260**, 81 (1991); Commun. Math. Phys. **149**, 279 (1992).
[20] O. Babelon and D. Bernard, Int. J. Mod. Phys. A **8**, 507 (1993).
[21] H. Belich and R. Paunov, J. Math. Phys. **38**, 4108 (1997).
[22] M. Niedermaier, Commun. Math. Phys. **160**, 391 (1994).
[23] R. Beals, P. Deift, and C. Tomei, *Direst and Inverse Scattering on the Line* (Mathematical Surveys and Monographs No. 28, 1988).
[24] H. Flaschka and D. W. McLaughlin, Prog. Theor. Phys. **55**, 438 (1976).
[25] O. Babelon, D. Bernard, and F. A. Smirnov, hep-th/9603010; Commun. Math. Phys. **182**, 319 (1996); **186**, 601 (1997).
[26] O. Babelon and L. Bonora, Phys. Lett. B **244**, 220 (1990); H. Aratyn, L. A. Ferreira, J. F. Gomes, and A. H. Zimerman, *ibid.* **254**, 372 (1991); C. P. Constantinidis, L. A. Ferreira, J. F. Gomes, and A. H. Zimerman, *ibid.* **298**, 88 (1993).
[27] E. Date, Osaka J. Math. **19**, 125 (1982).
[28] M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **71**, 313 (1981).
[29] O. Babelon and D. Bernard, Phys. Lett. B **317**, 363 (1993); H. Braden and A. N. W. Hone, *ibid.* **380**, 296 (1996).
[30] E. J. Beggs and P. R. Johnson, Nucl. Phys. B **484**, 653 (1997).
[31] V. Zakharov and A. Shabat, Funct. Anal. Appl. **13**, 166 (1979).
[32] G. Cuba and R. Paunov, Phys. Lett. B **381**, 255 (1996).
[33] P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986); P. Ginsparg, *Applied Conformal Field Theory*, edited by E. Brézin and J. Zinn-Justin (Les Houches, 1988), Session XLIX.
[34] J. Lepowsky and R. L. Wilson, Commun. Math. Phys. **62**, 43 (1978).
[35] V. G. Kac, D. A. Kazhdan, J. Lepowsky, and R. L. Wilson, Adv. Math. **42**, 83 (1981).
[36] N. J. Mackay and W. A. McGhee, Int. J. Mod. Phys. A **8**, 2791 (1993).
[37] E. Verlinde, Nucl. Phys. B: Field Theory Stat. Syst. **300**, 360 (1988).
[38] B. Dubrovin, Nucl. Phys. B **379**, 627 (1992); Commun. Math. Phys. **152**, 539 (1993).
[39] H. Aratyn, C. P. Constantinidis, L. A. Ferreira, J. F. Gomes, and A. H. Zimerman, Nucl. Phys. B **406**, 727 (1993).

Higher dimensional Painlevé integrable models from the Kadomtsev–Petviashvili equation

Sen-yue Lou[a)]
CCAST (World Laboratory), P.O. Box 8730,
Beijing 100080, People's Republic of China
and Institute of Modern Physics, Normal College, Ningbo University,
Ningbo 315211, People's Republic of China[b)]
and International Centre for Theoretical Physics, P.O. Box 586, 34100 Trieste, Italy

Jian-jun Xu
Department of Physics, Fudan University,
200433 Shanghai, People's Republic of China
and International Centre for Theoretical Physics, P.O. Box 586, 34100 Trieste, Italy

(Received 17 October 1997; accepted for publication 1 June 1998)

After embedding the Kadomtsev–Petviashvili equation in higher dimensions and extending the Painlevé analysis approach to a new form such that the coefficients of the expansion around the singular manifold possess conformal invariance and contain explicit new space variables, we can get infinitely many Painlevé integrable models in $(3+1)$-dimensions and higher dimensions. Some concrete higher dimensional modified Korteweg–de Vries type of extensions are given. Whether the models are Lax integrable or integrable under other meanings remain still open. © *1998 American Institute of Physics.* [S0022-2488(98)00510-6]

I. INTRODUCTION

Integrable models in $(1+1)$- and $(2+1)$-dimensions have been deeply studied by many mathematicians and physicists and widely applied in many physical fields like the classical and quantum field theory,[1] condense matter physics,[2] fluid mechanics,[3] plasma physics,[4] optics,[5] particle and nuclear physics,[6] universe,[7] geophysics,[8] planetary and space science.[9] Actually, the existence of the localized excitations (solitons or solitary waves) is one of the basic nonlinear phenomena in many other natural sciences, for instance, communications,[10] chemistry,[11] and biology.[12]

However, there is no one known real significant $(n+1)$-dimensional $(n\geqslant 3)$ nontrivial nonlinear integrable model except the ''$2+2$'' dimensional self-dual Yang–Mills field equation[13] though the real physical space is $(3+1)$-dimensional. Because the real physical space is $(3+1)$-dimensional, various physicists and mathematicians have tried to find some nontrivial $(3+1)$-dimensional integrable models[13,14] for a long time. But there is little progress in this direction.

Recently, we have proposed some possible directions to find some nontrivial integrable models in higher dimensions.

Based on the fact that all the known $(2+1)$-dimensional integrable models possess a common generalized Virasoro type symmetry algebra,[15]

$$[\sigma(f_1),\sigma(f_2)]=\sigma(\dot{f}_2 f_1-\dot{f}_1 f_2), \tag{1}$$

where $\sigma(f)$ with f being an arbitrary function of a single independent variable is a symmetry of the integrable model, we proposed a general method to get some integrable models under the meaning that they possess the generalized Virasoro symmetry algebra (1) in Refs. 16 and 17.

$(2+1)$-dimensional breaking soliton equations[18] are a special type of integrable extension of the $(1+1)$-dimensional integrable models. A $(2+1)$-dimensional breaking soliton equation and

[a)]Electronic mail: sylou@public.nbptt.zj.cn
[b)]Corresponding address.

 © 1998 American Institute of Physics

its corresponding $(1+1)$-dimensional integrable model possess a common recursion operator.[19,20] Using the same common recursion operator, the breaking soliton equation can be extended to any dimensions.[20,17]

Usually, using the symmetry constraint[21–23] (or reduction[24–26]) approaches, one can always get some lower dimensional integrable models from higher dimensional ones. In Refs. 27 and 28, we pointed out that using the inner parameter dependent symmetry constraints, one can also get some types of integrable models in the same or even higher dimensions.

It is also known that for every known $(1+1)$-dimensional and $(2+1)$-dimensional integrable model, there is an invariant form with conformal invariance (invariance under the Möbious transformation). And the conformal invariance also plays a very important role in the study of other integrable properties like infinitely many symmetries and the Darboux transformation.[29,28] These facts hint to us that if we want to find a higher dimensional integrable model, starting from a conformal invariant form may be one of the best convenient ways. In Refs. 17, 30, and 31, we can see that some quite general higher dimensional conformal invariant models are integrable under the meaning that they possess the Painlevé property.

Getting some lower dimensional integrable models from a higher dimensional one is quite easy, say, one can use some limit procedures or symmetry reduction approaches.[21–26] In Ref. 32, we have also pointed out that the inverse procedure of the above (i.e., to obtain higher dimensional integrable models from the lower dimensional ones) is possible. On the other hand, from the symmetry reductions of $(1+1)$-dimensional and $(2+1)$-dimensional integrable models, we know that the ordinary differential equation (ODE) reductions can usually be casted to several forms, say, the Riccati equation and the Painlevé I–VI equations.[33] This fact shows us that all the higher dimensional integrable models can be considered as the deformations of a small set of integrable ODEs. For instance, some types of $(1+1)$-dimensional and $(2+1)$-dimensional sinh-Gordon equations and the Mikhailov–Dodd–Bullough (MDB) equations have been obtained by deforming a simple Riccati equation[32]

$$\phi_t = \phi^2. \tag{2}$$

In this paper, we would like to propose another approach to get higher dimensional integrable models from lower dimensional ones. The singularity analysis is one of the most powerful methods to study the integrabilities of integrable models. For instance, using the singularity analysis of an integrable model proposed by Weiss, Tabor, and Carnevale (WTC),[34] one can obtain the Painlevé property, Bäcklund transformation, Lax pair, symmetries, bilinear form and some types of exact solutions[33–37] for some kinds of integrable models. However, to our knowledge, no one uses the Painlevé analysis approach to get some new integrable models under any meanings.

In Sec. II, after extending the Painlevé approach to a general form, we propose a possible general approach to get some higher dimensional Painlevé integrable (*P*-integrable) models from lower dimensional ones by means of the extended Painlevé analysis approach. In Sec. III, taking the Kadomtsev–Petviashvili (KP) equation as a seed integrable model and using the method outlined in Sec. II, we obtain various higher dimensional *P*-integrable models especially in $(3+1)$-dimensions. Section IV is a summary and discussion.

II. GENERAL THEORY

For a given *n*-dimensional *N* order PDE,

$$F(x_0 \equiv t, x_1, x_2, \ldots, x_n, u, u_{x_i}, u_{x_i x_j}, \ldots, u_{x_{i_1} x_{i_2}, \ldots, x_{i_N}}) = 0, \tag{3}$$

where the function F is a polynomial function of the field u and its derivatives, the model is said to possess the Painlevé property if all the movable singularities of its solution with respect to an arbitrary singular manifold $\phi \equiv \phi(x_0, x_1, x_2, \ldots, x_n) = 0$ are poles. That is to say, if expanding the solution of (3) near the singular manifold ϕ, we should have the form[34]

$$u = \sum_{j=0}^{\infty} u_j \phi^{j+\alpha} \tag{4}$$

with $N-1$ further arbitrary functions $u_{j1},u_{j2},\ldots,u_{j_{N-1}}$ and negative integer α. Substituting (4) into (3), we can get a recursion relation to determine the expansion coefficients u_j,

$$(j+1)(j-j_1)(j-j_2)\cdots(j-j_{N-1})u_j=f_j(x_i,\phi_{x_i},\phi_{x_{i_1}x_{i_2}},\ldots,u_1,u_2,\ldots,u_{j-1})\equiv f_j, \tag{5}$$

where $j_1,j_2,\ldots,j_{N-1}$ should be positive integers and the resonance conditions

$$f_{j_i}=0 \quad (i=1,2,...,N-1), \tag{6}$$

should be satisfied identically.

If we take

$$u_j=0 \quad (j>-\alpha), \tag{7}$$

then we get a Bäcklund transformation of (3), i.e., if $u_{-\alpha}$ is a solution of (3), then the truncated expansion u given by (4) with (6) is also a solution of (3). Furthermore, if $u=0$ is a solution of (3), the single soliton solution of (3) can be obtained from the Bäcklund transformation and the trivial solution $u_{-\alpha}=0$. Usually, when one uses the truncated expansions to discuss the Bäcklund transformation or exact solutions, no one discusses any properties of the equations $u_j=0$, $(j\geqslant-\alpha)$ i.e.,

$$f_j=0 \quad (j\geqslant-\alpha, j\neq j_1,j_2,\ldots,j_{N-1}). \tag{8}$$

Because equations (7) for different j must be consistent when we use the truncation expansion to study the integrable property of (3), we believe that many equations of (7) or some combinations of u_j $(j\geqslant-\alpha)$ may be integrable.

The models obtained from (7) possess the same dimensions as the original model (3). In order to get some higher-dimensional integrable models from (3), we may take the following two steps. First, we should embed the lower dimensional integrable model (3) in higher dimensions. In other words, we should consider u as not only dependent on the explicit independent variables $\{x_1,\ldots,x_n,t\}$ but also some implicit ones, say, $\{x_{n+1},\ldots,x_{n+m}\}$. Second, the Painlevé expansion (4) should be extended to a different resummation form such that the implicit independent variables $\{x_{n+1},\ldots,x_{n+m}\}$ appear explicitly in the new expansion coefficients and the same number of arbitrary functions are still included in the new expansion form. The second step can be realized because of the singular manifold ϕ being arbitrary. For instance, we can take

$$\xi\equiv\left(\frac{\phi_{x_{n+1}}}{\phi}-\frac{\phi_{x_{n+1}x_{n+1}}}{2\phi_{x_{n+1}}}\right)^{-1} \tag{9}$$

as a new expansion variable, i.e., we change the expansion (4) as

$$u=\sum_{j=0}^{\infty}u'_j\xi^{j+\alpha} \tag{10}$$

with arbitrary $\xi, u'_{j_1},u'_{j_2},\ldots,u'_{j_{N-1}}$ and the same integers $j_1,j_2,\ldots,j_{N-1}$, and α as in (4). From Eq. (9) one can easily prove the following identities:

$$\xi_{x_i}=P_i-P_{ix_{n+1}}\xi+\tfrac{1}{2}(P_iS+P_{ix_{n+1}x_{n+1}})\xi^2, \quad i=0,1,2,...,n, \tag{11}$$

where the functions

$$P_i\equiv\frac{\phi_{x_i}}{\phi_{x_{n+1}}}, \quad S\equiv\frac{\phi_{x_{n+1}x_{n+1}x_{n+1}}}{\phi_{x_{n+1}}}-\frac{3}{2}\left(\frac{\phi_{x_{n+1}x_{n+1}}}{\phi_{x_{n+1}}}\right)^2\equiv\{\phi,x_{n+1}\} \tag{12}$$

are all conformal invariant. In other words, P_i and S are invariant under the Möbious transformation

$$\phi \to \frac{a+b\phi}{c+d\phi}, \quad ad \neq bc. \tag{13}$$

It is straightforward to see that all the expansion coefficients u_j' in (10) are all conformal invariant because they are functions of P_i, S and $\{u_{j_1}', u_{j_2}', \ldots, u_{j_{N-1}}'\}$. Now a further independent variable x_{n+1} has been included explicitly to the expansion coefficients u_j' though the original equation (3) is not x_{n+1} dependent explicitly.

If taking $n=1$, $x_1=x$, and $x_2=x_1=x$, the expansion function (9) and the extended approach proposed above reduced to Conte's simplification for the usual Painlevé analysis approach.[36]

On the other hand, the conformal invariance plays very important roles in integrable theory. For instance, starting from the conformal invariance of the Korteweg–de Vries (KdV) and KP equations, one can obtain the infinitely many symmetries, Darboux transformations, sine-Gordon extensions, etc.[29] Starting from a quite general conformal form one can get infinitely many *P*-integrable models.[31] Now it is reasonable to assume that:

Assumption. If Eq. (3) is integrable, then the equations obtained from vanishing the coefficients of the extended Painlevé expansion (10), for $u_j'=0$, $j \geqslant -\alpha$, are integrable.

Though we have not yet proved this assumption and/or even the assumption may not be completely true, we can still obtain various higher-dimensional integrable models under some special meanings, say, under the Painlevé meaning (the model possesses a variant form with Painlevé property), from lower ones by checking the assumption for some given lower dimensional integrable models and small j. It is interesting that if the seed equation (3) is $(2+1)$-dimensional, then the integrable models obtained from the assumption are $(3+1)$-dimensional. Furthermore, the assumption can be used in many times to get some integrable models in arbitrary dimensions no matter how many dimensions the seed model has. For instance, if the seed model is $(1+1)$-dimensional, then using the assumption once, one may get some $(2+1)$-dimensional models (after some necessary integrability verification of the obtained models because the assumption has not yet been proved). Using the obtained $(2+1)$-dimensional integrable models as seeds, one may obtain some $(3+1)$-dimensional integrable models and so on.

Now a natural and important question is whether the new independent variables introduced in this way are genuinely new independent variables? In our experience, the new independent variables are genuinely entered into every coefficient of the series expansion (10). The new variables can be absorbed only when we considered all the terms of the expansion (10).

III. HIGHER-DIMENSIONAL *P*-INTEGRABLE MODELS FROM THE KP EQUATION

A. *P*-integrable models from the KP equation directly

The $(2+1)$-dimensional KP equation ($x_0=t$, $x_1=x$, $x_2=y$, $x_3=z$)

$$(u_t - 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0 \tag{14}$$

is one of the most important integrable physical models which is widely used in many physical fields such as fluid mechanics, plasma physics, optics, condense matter physics, geophysics, planetary and space science, etc. And the model is widely studied by various mathematicians and physicists. The Painlevé property of the KP equation was first proved by WTC.[34] In order to obtain some nontrivial $(3+1)$-dimensional model from the KP equation, we restudy the Painlevé property of the model by using the new expansion (10) and considering u of (14) is not only a function of $\{x,y,t\}$ but also a function of z. As in the usual case,[34] by means of the leading order analysis, we have

$$\alpha = -2, \quad u_0' = 2P_1^2 \tag{15}$$

in (10) for the KP equation (14). Substituting (10) with (11) and (15) into (14) we get the recursion relation of the coefficients u_j'

$$(j+1)(j-4)(j-5)(j-6)u_j' = f_j'(S, P_i, P_{ix_i}, \ldots, u_0', \ldots, u_{j-1}') \equiv f_j', \tag{16}$$

where f_j' is a complicated function of $u_0',\ldots,u_{j-1}'$ and the conformal invariants P_i, S and their derivatives. After solving (16) out one by one, f_j' becomes only a function of the conformal invariants. The first three of u_j' read

$$u_1' = -2P_{1x} - 2P_1P_{1z}, \tag{17}$$

$$u_2' = \frac{1}{6P_1^2}(4P_1P_{1xx} + 4P_1^2P_{1xz} + 2P_1P_{1x}P_{1z} - 3P_{1x}^2 + 4P_1^4S + P_1P_0 + P_1^2P_{1z}^2 + 4P_1^3P_{1zz} + 3\sigma^2P_2^2), \tag{18}$$

$$\begin{aligned} u_3' = \frac{1}{6P_1^4}(&P_1^2P_{0x} - 2P_1^2P_{1t} - P_1^4P_{1xzz} - P_1^3P_{1xxz} - P_1^5S_x - 2P_1^3P_{0z} - 15\sigma^2P_1^2P_2P_{2z} \\ &+ P_1P_{1x}P_0 + 2P_1^2P_{1z}P_0 + P_1^3P_{1xz}P_{1z} + 4P_1P_{1x}P_{1xx} - 2P_1^4P_{1x}S - 3P_{1x}^3 \\ &+ 3\sigma^2P_{1x}P_2^2 + P_1^2P_{1x}P_{1xz} - P_1^3P_{1x}P_{1zz} - P_1^2P_{1xxx} + 15\sigma^2P_1P_2^2P_{1z} - P_1P_{1z}P_{1x}^2 \\ &+ 12\sigma^2P_1P_2P_{2x} - 3\sigma^2P_1^2P_{2y} + P_1^2P_{1z}P_{1xx} - 12\sigma^2P_1P_2P_{1y}). \end{aligned} \tag{19}$$

Using the relations (12), (15), (17)–(19), and the computer algebra, say, MAPLE or MATHEMATICA, it is easy to prove that the following three resonant conditions (16) for $j=4,5,6$ are satisfied identically. The Painlevé property of the KP equation is reobtained under the new configuration.

Now using the assumption, we may obtain a set of Painlevé integrable models

$$f_j' = 0, \quad j = 2,3,7,8,..., \tag{20}$$

where $j=4$, 5, 6 disappear due to the resonance conditions as these values are satisfied identically. It is difficult to prove the Painlevé property for whole set. We can only check the integrabilities for small j of (20).

After substituting (12) into (20) for $j=3$ [i.e., (19) with $u_3'=0$], we have

$$\phi_{xxxx}\phi_x^2 + \phi_x^2\phi_{xt} + \phi_{xx}^3 - \phi_t\phi_x\phi_{xx} - 4\phi_x\phi_{xx}\phi_{xxx} + 3\sigma^2\phi_{yy}\phi_x^2 - 3\sigma^2\phi_y^2\phi_{xx} = 0, \tag{21}$$

i.e.,

$$\left(\frac{\phi_t}{\phi_x} + \{\phi,x\} + \frac{3}{2}\sigma^2\frac{\phi_y^2}{\phi_x^2}\right)_x + 3\sigma^2\left(\frac{\phi_y}{\phi_x}\right)_y = 0. \tag{22}$$

From (21) we see that it is not explicitly z dependent. The integrability of (21) is trivial because it is really the original $(2+1)$-dimensional KP equation in its Schwarz form[34] and Eqs. (14) and (22) are related by[32,35]

$$u = \frac{1}{2}\{\phi,x\} + \frac{\phi_{xx}^2}{2\phi_x^2} + \sigma^2\frac{\phi_y^2}{4\phi_x^2} - \frac{\sigma^2}{2}\int^x\left(\frac{\phi_y}{\phi_x}\right)_y dx. \tag{23}$$

The usual Schwarz KP equation appears as a special case of our assumption, $u_3'=0$, which is natural because we obtain (20) from the KP equation.

Though we can prove the Painlevé property of (20) for $j=7$, we do not write it down because (20) with $j=7$ is too complicated after substituting (12) in (about two or more printed pages). So we discuss only for $j=2$ here. Substituting $j=2$ and (12) into (20) [i.e., (18) with $u_2'=0$] yields a nontrivial $(3+1)$-dimensional model

$$4\phi_x\phi_z^4\phi_{xxx} - 3\phi_z^4\phi_{xx}^2 + 3\phi_x^4\phi_{zz}^2 + \phi_x\phi_t\phi_z^4 + 3\sigma^2\phi_y^2\phi_z^4 - 6\phi_x^2\phi_z^2\phi_{xx}\phi_{zz} = 0. \tag{24}$$

Making the following transformation

$$\phi = e^f, \tag{25}$$

Eq. (24) can be rewritten as

$$f_t+4f_{xxx}-2f_x^3-3f_{xx}^2f_x^{-1}+3\sigma^2f_y^2f_x^{-1}+3f_x^3f_{zz}^2f_z^{-4}-6f_xf_{xx}f_{zz}f_z^{-2}=0. \tag{26}$$

Equation (26) is a $(3+1)$-dimensional extension of the potential modified KdV (PMKdV) equation. The first four $\{y,z\}$ independent terms are equivalent to the PMKdV equation. The fifth term and the sixth and seventh terms of (26) are the y and z part modification of the PMKdV equation.

In order to prove the integrability of (26) [then (24)], we change (26) to an equivalent equation system

$$4UW^4U_{xx}-3W^4U_x^2-2W^4U^4+3U^4W_z^2+UGW^4-6U^2W^2U_xW_z+3\sigma^2V^2W^4=0, \tag{27}$$

$$U_t=G_x, \tag{28}$$

$$V_t=G_y, \tag{29}$$

$$W_t=G_z, \tag{30}$$

by using further transformations

$$f_x=U,\ \ f_y=V,\ \ f_z=W,\ \ f_t=G. \tag{31}$$

Equations (28)–(30) are the consistent conditions of the transformation (31). To prove the Painlevé property of the equation system (27)–(30), we may use both the traditional WTC approach[34] or its extended form proposed in Sec. II. Using the extended WTC approach we can not only prove the Painlevé property of the system (27)–(30) but also get some more integrable models both in $(4+1)$-dimensions and in $(3+1)$-dimensions (combining with some suitable reduction approaches). So we would like take the extended approach to prove the Painlevé property of (27)–(30) again in Sec. III B.

B. *P*-integrable models from the system (27)–(30)

Expanding U,V,W,G as

$$U=\sum_{j=0}^{\infty}U_j\xi^{j+\alpha_1},\quad V=\sum_{j=0}^{\infty}V_j\xi^{j+\alpha_2},\quad W=\sum_{j=0}^{\infty}W_j\xi^{j+\alpha_3},\quad G=\sum_{j=0}^{\infty}G_j\xi^{j+\alpha_4}, \tag{32}$$

where ξ is given by (9) with $n=3$, $x_0=t$, $x_1=x$, $x_2=y$, $x_3=z$, and $\alpha_1=\alpha_2=\alpha_3=\alpha_4=-1$,

$$U_0=\pm P_1,\quad V_0=\pm P_2,\quad W_0=\pm P_3,\quad G_0=\pm P_0 \tag{33}$$

are determined by the leading order analysis and the relation (11) for $n=3$.

Substituting (32) with $\alpha_i=-1$, (33) and (11) into (27)–(30), one can see that all the needed five resonances are located at

$$j=-1,1,1,1,1. \tag{34}$$

The resonance at $j=-1$ is related to ϕ (and then ξ) being arbitrary. The resonance conditions at $j=1$ read

$$\begin{aligned}&3U_0(P_1W_0-U_0P_3)(W_0^2P_{1x_4}-U_0(W_0P_{3x_4}+W_{0z}))+2W_0^3U_0P_{1x}\\&+W_0^2(W_0P_1-3U_0P_3)U_{0x}-2W_0[(2U_0^2W_0^2-3U_0^2P_3^2\\&+3U_0W_0P_1P_3-2W_0^2P_1^2)U_1-U_0(5P_1^2W_0-2W_0U_0^2-3P_1P_3U_0)W_1]=0,\end{aligned} \tag{35}$$

$$U_{0t}-G_{0x}+U_0P_{0x_4}-G_0P_{1x_4}=0, \tag{36}$$

$$V_{0t}-G_{0y}+V_0P_{0x_4}-G_0P_{2x_4}=0, \tag{37}$$

$$W_{0t}-G_{0z}+W_0P_{0x_4}-G_0P_{3x_4}=0. \tag{38}$$

It is straightforward to see that all the resonance conditions (35)–(38) are satisfied identically because of the relations (33) and (12) with $n=3$. So the equation system (27)–(30) possesses the Painlevé property and then Eq. (26) [and (24)] is a $(3+1)$-dimensional *P*-integrable model under the meaning that it possesses the conformal invariance and the Painlevé property.

Using the further assumption to the system (27)–(30), some $(4+1)$-dimensional models with the Painlevé property can be obtained. For instance, from $U_2=V_2=W_2=G_2=0$, we get

$$\begin{aligned}&(3\sigma^2P_2^2+P_0P_1-2P_1^4S+P_1^2P_{1x_4}^2+2P_1P_{1x}P_{1x_4}-3P_{1x}^2)P_3^4+3(P_3^2P_{3x_4}^2+2P_3P_{3z}P_{1x_4}+P_{3z}^2)P_1^4\\&\quad-6P_1^2P_3^2P_{1x}P_{3z}-6P_1^3P_3^3P_{1x_4}P_{3x_4}-6P_1^2P_3^3P_{1x}P_{3x_4}-6P_3^2P_1^3P_{3x}P_{1x_4}=0,\end{aligned} \tag{39}$$

i.e.,

$$\begin{aligned}&8\phi_z^4\phi_x\phi_{x_4}^3\phi_{xx}\phi_{xx_4}+4\phi_z^4\phi_x^2\phi_{x_4}^2\phi_{xx}\phi_{x_4x_4}-3\phi_z^4\phi_{x_4}^4\phi_{xx}^2+3\phi_x^4\phi_{x_4}^4\phi_{zz}^2-6\phi_z^2\phi_x^2\phi_{x_4}^4\phi_{xx}\phi_{zz}\\&\quad+\phi_t\phi_x\phi_z^4\phi_{x_4}^4-2\phi_x^4\phi_z^4\phi_{x_4x_4x_4}\phi_{x_4}-4\phi_z^4\phi_x^2\phi_{xx_4}\phi_{x_4}^2+3\sigma^2\phi_y^2\phi_z^4\phi_{x_4}^4+\phi_x^4\phi_z^4\phi_{x_4x_4}^2=0.\end{aligned} \tag{40}$$

The Painlevé property of (40) can be proved in the same way. Using the classical and nonclassical Lie approach, some kinds of $(3+1)$-dimensional Painlevé integrable reductions can be obtained from (40). The simplest three $(3+1)$-dimensional reductions are

$$6\phi_z^4\phi_{xx}^2+3\phi_x^4\phi_{zz}^2-6\phi_z^2\phi_x^2\phi_{xx}\phi_{zz}+\phi_t\phi_x\phi_z^4-2\phi_x\phi_z^4\phi_{xxx}+3\sigma^2\phi_y^2\phi_z^4=0, \tag{41}$$

$$\begin{aligned}&8\phi_z^4\phi_x\phi_{x_4}^3\phi_{xx}\phi_{xx_4}+4\phi_z^4\phi_x^2\phi_{x_4}^2\phi_{xx}\phi_{x_4x_4}-3\phi_z^4\phi_{x_4}^4\phi_{xx}^2+3\phi_x^4\phi_{x_4}^4\phi_{zz}^2-6\phi_z^2\phi_x^2\phi_{x_4}^4\phi_{xx}\phi_{zz}\\&\quad+\phi_t\phi_x\phi_z^4\phi_{x_4}^4-2\phi_x^4\phi_z^4\phi_{x_4x_4x_4}\phi_{x_4}-4\phi_z^4\phi_x^2\phi_{xx_4}\phi_{x_4}^2+\phi_x^4\phi_z^4\phi_{x_4x_4}^2=0,\end{aligned} \tag{42}$$

and

$$\begin{aligned}&8\phi_z^4\phi_x\phi_{x_4}^3\phi_{xx}\phi_{xx_4}+4\phi_z^4\phi_x^2\phi_{x_4}^2\phi_{xx}\phi_{x_4x_4}-3\phi_z^4\phi_{x_4}^4\phi_{xx}^2+3\phi_x^4\phi_{x_4}^4\phi_{zz}^2-6\phi_z^2\phi_x^2\phi_{x_4}^4\phi_{xx}\phi_{zz}\\&\quad-2\phi_x^4\phi_z^4\phi_{x_4x_4x_4}\phi_{x_4}-4\phi_z^4\phi_x^2\phi_{xx_4}\phi_{x_4}^2+3\sigma^2\phi_y^2\phi_z^4\phi_{x_4}^4+\phi_x^4\phi_z^4\phi_{x_4x_4}^2=0.\end{aligned} \tag{43}$$

Equations (41)–(43) can be obtained by using the symmetry constraints

$$\phi_x-\phi_{x_4}=0, \tag{44}$$

$$\phi_y=0, \tag{45}$$

and

$$\phi_t=0, \tag{46}$$

respectively. One can easily see that the difference between the reduction (41) and the original equation (24) is only the ratio of the coefficients of the terms $\phi_z^4\phi_x\phi_{xxx}$ and $\phi_z^4\phi_{xx}^2$. Equations (42) and (43) can also be obtained directly from the KdV equation and the Boussinesq equation by using the method proposed here.

C. The Painlevé property of (40) and further Painlevé integrable models

Now we use the extended Painlevé analysis and the assumption further on (40) to prove its Painlevé property and get some more higher dimensional *P*-integrable models.

Using the transformation (25), Eq. (40) becomes

$$f_t f_x f_z^4 f_{x_4}^4 + 3\sigma^2 f_y^2 f_z^4 f_{x_4}^4 + 8 f_z^4 f_x f_{x_4}^3 f_{xx} f_{xx_4} + f_z^4 f_x^4 f_{x_4}^4 + 4 f_z^4 f_x^2 f_{x_4}^2 f_{xx} f_{x_4x_4} - 3 f_z^4 f_{x_4}^4 f_{xx}^2 + 3 f_x^4 f_{x_4}^4 f_{zz}^2$$
$$- 6 f_z^2 f_x^2 f_{x_4}^4 f_{xx} f_{zz} - 2 f_x^4 f_z^4 f_{x_4} f_{x_4x_4x_4} - 4 f_z^4 f_x^2 f_{x_4}^2 f_{xx_4}^2 + f_x^4 f_z^4 f_{x_4x_4}^2 = 0. \quad (47)$$

Substituting the further transformations

$$f_x = U, \quad f_y = V, \quad f_z = W, \quad f_t = G, \quad f_{x_4} = H \quad (48)$$

into (47) yields an equivalent system

$$GUW^4H^4 + 3\sigma^2 V^2 W^4 H^4 + 8 W^4 U H^3 H_x U_{x_4} + 4 W^4 U^2 H^2 U_x H_{x_4} - 3 W^4 H^4 U_x^2 + U^4 W^4 H^4$$
$$+ U^4 W^4 H_{x_4}^2 + 3 U^4 H^4 W_z^2 - 6 W^2 U^2 H^4 U_x W_z - 2 U^4 W^4 H H_{x_4x_4} - 4 W^4 U^2 H^2 U_{x_4}^2 = 0, \quad (49)$$

$$U_t = G_x, \quad (50)$$

$$V_t = G_y, \quad (51)$$

$$W_t = G_z, \quad (52)$$

$$H_t = G_{x_4}. \quad (53)$$

Equations (50)–(53) are the consistent conditions of the transformation (48).

To prove the Painlevé property of the equation system (49)–(53), we expand U, V, W, G, H as

$$U = \sum_{j=0}^{\infty} U_j \xi^{j+\alpha_1}, \quad V = \sum_{j=0}^{\infty} V_j \xi^{j+\alpha_2}, \quad W = \sum_{j=0}^{\infty} W_j \xi^{j+\alpha_3},$$

$$G = \sum_{j=0}^{\infty} G_j \xi^{j+\alpha_4}, \quad H = \sum_{j=0}^{\infty} H_j \xi^{j+\alpha_4}, \quad (54)$$

where ξ is given by (9) with $n=4$, $x_0=t$, $x_1=x$, $x_2=y$, $x_3=z$ and $\alpha_1=\alpha_2=\alpha_3=\alpha_4=-1$,

$$U_0 = \pm P_1, \quad V_0 = \pm P_2, \quad W_0 = \pm P_3, \quad G_0 = \pm P_0, \quad H_0 = \pm P_4 \quad (55)$$

are fixed by the leading order analysis and the relation (11) for $n=4$.

Substituting (54) with $\alpha_i = -1$, (55) and (11) into (49)–(53), we see that all the required six resonances are located at

$$j = -1, 1, 1, 1, 1, 1. \quad (56)$$

The resonance at $j=-1$ is related to ϕ (and then ξ) being arbitrary. The resonance conditions at $j=1$ require that the expansion coefficients $\{U_1, V_1, W_1, G_1, H_1\}$ should be arbitrary. The related resonance conditions are

$$(P_0^2 - G_0^2)(2P_4P_3U_1 + 2P_1P_4W_1 + 2P_1P_3H_1) + P_1P_3(G_{0x_4}P_0 - G_0P_{0x_4}) = 0, \quad (57)$$

$$U_{0t} - G_{0x} + U_0 P_{0x_4} - G_0 P_{1x_4} = 0, \quad (58)$$

$$V_{0t} - G_{0y} + V_0 P_{0x_4} - G_0 P_{2x_4} = 0, \quad (59)$$

$$H_{0t} - G_{0x_4} + H_0 P_{0x_4} - G_0 P_{4x_4} = 0, \quad (60)$$

$$W_{0t} - G_{0z} + W_0 P_{0x_4} - G_0 P_{3x_4} = 0. \quad (61)$$

It is straightforward to see that all the resonance conditions (57)–(61) are satisfied identically because of the relations (55) and (12) with $n=4$. So the equation system (49)–(53) possesses the Painlevé property.

Using the assumption further to the system (49)–(53), some $(5+1)$-dimensional models with the Painlevé property can be obtained. The simplest five of them read

$$\begin{aligned}
&P_3^4P_1^4P_{4x_4}^2-3P_3^4P_4^4P_{1x}^2+3P_1^4P_4^4P_{3z}^2+P_1^4P_4^4P_3^4S+4P_3^4P_1^4P_4^3P_{4x_5x_5}-2P_3^4P_1^4P_4P_{4x_4x_4}\\
&\quad-6P_1^3P_4^4P_3^2P_{1x_5}P_{3z}-6P_1^3P_4^4P_3^3P_{1x_5}P_{3x_5}+4P_1^3P_4^2P_3^4P_{1x_5}P_{4x_4}+3P_1^4P_4^4P_3^2P_{3x_5}^2\\
&\quad+6P_1^4P_4^4P_3P_{3z}P_{3x_5}+12P_1^2P_4^3P_3^4P_{1x}P_{4x_5}-6P_1^2P_4^4P_3^2P_{1x}P_{3z}-5P_3^4P_1^4P_4^2P_{4x_5}^2\\
&\quad-4P_3^4P_1^2P_4^2P_{1x_4}^2-8P_3^4P_3^3P_1^2P_4^2P_{1x_4}P_{4x_5}-6P_1^2P_4^4P_3^3P_{1x}P_{3x_5}+4P_1^2P_4^2P_3^4P_{1x}P_{4x_4}\\
&\quad-2P_3^4P_1^4P_4P_{4x_4}P_{4x_5}+12P_1^3P_4^3P_3^4P_{1x_5}P_{4x_5}+8P_1P_3^4P_4^3P_{1x}P_{1x_4}-6P_3^4P_4^4P_1P_{1x}P_{1x_5}\\
&\quad-2P_3^4P_1^4P_4^2P_{4x_4x_5}+8P_1^2P_3^4P_4^3P_{1x_5}P_{1x_4}-3P_1^3P_3^4P_4^4P_{1x_5x_5}-3P_3^4P_4^4P_1^2P_{1x_5}^2+3\sigma^2P_2^2P_3^4P_4^4\\
&\quad+P_0P_1P_3^4P_4^4=0,
\end{aligned}\tag{62}$$

$$\begin{aligned}
&3\sigma^2P_2^3P_3^4P_4^4+P_2P_3^4P_1^4P_{4x_4}^2+P_2P_1^4P_4^4P_3^4S-3P_1^4P_3^4P_4^4P_{2x_5x_5}+12P_2P_1^2P_4^3P_3^4P_{1x}P_{4x_5}\\
&\quad-2P_2P_3^4P_1^4P_4^2P_{4x_4x_5}-6P_2P_1^3P_4^4P_3^2P_{1x_5}P_{3z}-3P_2P_3^4P_4^4P_1^2P_{1x_5}^2-2P_2P_3^4P_1^4P_4P_{4x_4x_4}\\
&\quad+8P_2P_1^2P_3^4P_4^3P_{1x_5}P_{1x_4}+3P_2P_1^4P_4^4P_3^2P_{3x_5}^2-6P_2P_1^3P_4^4P_3^3P_{1x_5}P_{3x_5}-5P_2P_3^4P_1^4P_4^2P_{4x_5}^2\\
&\quad+4P_2P_3^4P_1^4P_4^3P_{4x_5x_5}-3P_2P_3^4P_4^4P_1^2+8P_2P_1P_3^4P_4^3P_{1x}P_{1x_4}-6P_2P_1^2P_4^4P_3^2P_{1x}P_{3z}\\
&\quad-6P_2P_3^4P_4^4P_1P_{1x}P_{1x_5}-2P_2P_3^4P_1^4P_4P_{4x_4}P_{4x_5}-8P_2P_3^4P_1^3P_4^2P_{1x_4}P_{4x_5}\\
&\quad+12P_2P_1^3P_3^4P_4^4P_{1x_5}P_{4x_5}+6P_2P_1^4P_4^4P_3P_{3z}P_{3x_5}+4P_2P_1^3P_4^2P_3^4P_{1x_5}P_{4x_4}-4P_2P_3^4P_1^2P_4^2P_{1x_4}^2\\
&\quad-6P_2P_1^2P_4^4P_3^3P_{1x}P_{3x_5}+3P_2P_1^4P_4^4P_{3z}^2+4P_2P_1^2P_4^2P_3^4P_{1x}P_{4x_4}+P_2P_0P_1P_3^4P_4^4=0,
\end{aligned}\tag{63}$$

$$\begin{aligned}
&P_3^4P_1^4P_{4x_4}^2-3P_3^4P_4^4P_{1x}^2+3P_1^4P_4^4P_{3z}^2+P_1^4P_4^4P_3^4S+4P_3^4P_1^4P_4^3P_{4x_5x_5}-2P_3^4P_1^4P_4P_{4x_4x_4}\\
&\quad-6P_1^3P_4^4P_3^2P_{1x_5}P_{3z}-6P_1^3P_4^4P_3^3P_{1x_5}P_{3x_5}+4P_1^3P_4^2P_3^4P_{1x_5}P_{4x_4}+3P_1^4P_4^4P_3^2P_{3x_5}^2\\
&\quad+12P_1^2P_4^3P_3^4P_{1x}P_{4x_5}-6P_1^2P_4^4P_3^2P_{1x}P_{3z}-5P_3^4P_1^4P_4^2P_{4x_5}^2-4P_3^4P_1^2P_4^2P_{1x_4}^2-3P_3^4P_4^4P_1^2P_{1x_5}^2\\
&\quad-6P_3^4P_4^4P_1P_{1x}P_{1x_5}-2P_3^4P_1^4P_4^2P_{4x_4x_5}+3\sigma^2P_2^2P_3^4P_4^4+P_0P_1P_3^4P_4^4-3P_{3x_5x_5}P_1^4P_3^3P_4^4\\
&\quad-2P_3^4P_1^4P_4P_{4x_4}P_{4x_5}+12P_1^3P_4^3P_3^4P_{1x_5}P_{4x_5}+8P_1P_3^4P_4^3P_{1x}P_{1x_4}+8P_1^2P_3^4P_4^3P_{1x_5}P_{1x_4}\\
&\quad+6P_1^4P_4^4P_3P_{3z}P_{3x_5}-8P_3^4P_1^3P_4^2P_{1x_4}P_{4x_5}-6P_1^2P_4^4P_3^3P_{1x}P_{3x_5}+4P_1^2P_4^2P_3^4P_{1x}P_{4x_4}=0,
\end{aligned}\tag{64}$$

$$\begin{aligned}
&P_3^4P_1^4P_{4x_4}^2-3P_3^4P_4^4P_{1x}^2+3P_1^4P_4^4P_{3z}^2+P_1^4P_4^4P_3^4S+P_3^4P_1^4P_4^3P_{4x_5x_5}-2P_3^4P_1^4P_4P_{4x_4x_4}\\
&\quad-6P_1^3P_4^4P_3^2P_{1x_5}P_{3z}-6P_1^3P_4^4P_3^3P_{1x_5}P_{3x_5}+4P_1^3P_4^2P_3^4P_{1x_5}P_{4x_4}+6P_1^4P_4^4P_3P_{3z}P_{3x_5}\\
&\quad+3P_1^4P_4^4P_3^2P_{3x_5}^2+12P_1^2P_4^3P_3^4P_{1x}P_{4x_5}-6P_1^2P_4^4P_3^2P_{1x}P_{3z}-5P_3^4P_1^4P_4^2P_{4x_5}^2-4P_3^4P_1^2P_4^2P_{1x_4}^2\\
&\quad-8P_3^4P_1^3P_4^2P_{1x_4}P_{4x_5}-6P_1^2P_4^4P_3^3P_{1x}P_{3x_5}+4P_1^2P_4^2P_3^4P_{1x}P_{4x_4}-2P_3^4P_1^4P_4P_{4x_4}P_{4x_5}\\
&\quad+12P_1^3P_4^3P_3^4P_{1x_5}P_{4x_5}+8P_1P_3^4P_4^3P_{1x}P_{1x_4}-6P_3^4P_4^4P_1P_{1x}P_{1x_5}-2P_3^4P_1^4P_4^2P_{4x_4x_5}\\
&\quad-3P_3^4P_4^4P_1^2P_{1x_5}^2+8P_1^2P_3^4P_4^3P_{1x_5}P_{1x_4}+3\sigma^2P_2^2P_3^4P_4^4+P_0P_1P_3^4P_4^4=0,
\end{aligned}\tag{65}$$

$$P_0^3P_1P_3^4P_4^4+P_1^4P_4^4P_3^4SP_0^2+P_3^4P_1^4P_{4x_4}^2P_0^2+12P_1^3P_4^3P_3^4P_{1x_5}P_{4x_5}P_0^2+12P_1^2P_4^3P_3^4P_{1x}P_{4x_5}P_0^2$$
$$-2P_3^4P_1^4P_4P_{4x_4}P_{4x_5}P_0^2-8P_3^4P_1^3P_4^2P_{1x_4}P_{4x_5}P_0^2-6P_1^2P_4^4P_3^3P_{1x}P_{3x_5}P_0^2$$
$$+4P_1^2P_4^2P_3^4P_{1x}P_{4x_4}P_0^2+6P_1^4P_4^4P_3P_{3z}P_{3x_5}P_0^2+4P_3^4P_1^4P_4^3P_{4x_5x_5}P_0^2-5P_3^4P_1^4P_4^2P_{4x_5}^2P_0^2$$
$$-6P_1^2P_4^4P_3^2P_{1x}P_{3z}P_0^2-2P_3^4P_1^4P_4^2P_{4x_4x_5}P_0^2-3P_3^4P_4^4P_1^2P_{1x_5}^2P_0^2-2P_3^4P_1^4P_4P_{4x_4x_4}P_0^2$$
$$+3P_1^4P_4^4P_{3z}^2P_0^2+4P_1^3P_4^2P_3^4P_{1x_5}P_{4x_4}P_0^2-3P_3^4P_4^4P_{1x}^2P_0^2-6P_3^4P_4^4P_1P_{1x}P_{1x_5}P_0^2$$
$$-6P_1^3P_4^4P_3^3P_{1x_5}P_{3x_5}P_0^2-6P_1^3P_4^4P_3^2P_{1x_5}P_{3z}P_0^2+3P_1^4P_4^4P_3^2P_{3x_5}^2P_0^2+8P_1^2P_3^4P_3^4P_{1x_5}P_{1x_4}P_0^2$$
$$+8P_1P_3^4P_4^3P_{1x}P_{1x_4}P_0^2-4P_3^4P_1^2P_4^2P_{1x_4}^2P_0^2-3P_3^4P_1^4P_4^4P_0P_{0x_5x_5}+3\sigma^2P_2^2P_3^4P_4^4P_0^2=0, \quad (66)$$

where P_0, P_1, P_2, P_3, P_4, and S are given by (12) with $n=4$. Equations (62)–(66) (which come from $U_2=0$, $V_2=0$, $W_2=0$, $G_2=0$, and $H_2=0$) are all conformal invariant because S, P_i, $i=0,1,...,4$ and their derivatives are all conformal invariant.

Using the classical and nonclassical Lie approach, some kinds of $(3+1)$-dimensional P-integrable reductions can be obtained from (62)–(66). Here we list some simple ones. Substituting

$$x_4=x_5=x$$

which is equivalent to take the symmetry constraints

$$\phi_{x_4}=\phi_x, \quad \phi_{x_5}=\phi_\eta, (\eta=x+x_4)$$

into (62)–(66), we get four independent $(3+1)$-dimensional Painlevé integrable models:

$$6\phi_x^4\phi_{zz}^2+2\phi_t\phi_x\phi_z^4+2\phi_z^4\phi_{xxx}\phi_x+3\phi_z^4\phi_{xx}^2+6\sigma^2\phi_y^2\phi_z^4-12\phi_x^2\phi_z^2\phi_{xx}\phi_{zz}=0, \quad (67)$$

$$6\phi_y\phi_x^4\phi_{zz}^2-6\phi_z^4\phi_{xxy}\phi_x^2+8\phi_z^4\phi_x\phi_y\phi_{xxx}+12\phi_z^4\phi_{xx}\phi_{xy}\phi_x-9\phi_z^4\phi_y\phi_{xx}^2$$
$$-12\phi_y\phi_z^2\phi_x^2\phi_{zz}\phi_{xx}+2\phi_y\phi_t\phi_z^4\phi_x+6\sigma^2\phi_y^3\phi_z^4=0, \quad (68)$$

$$6\phi_x^4\phi_{zz}^2+2\phi_t\phi_x\phi_z^4+8\phi_z^4\phi_{xxx}\phi_x-9\phi_z^4\phi_{xx}^2+6\sigma^2\phi_y^2\phi_z^4$$
$$+12\phi_z^3\phi_{xx}\phi_x\phi_{xz}-12\phi_x^2\phi_z^2\phi_{xx}\phi_{zz}-6\phi_z^3\phi_x^2\phi_{xxz}=0, \quad (69)$$

$$-6\phi_x^4\phi_t\phi_{zz}^2-2\phi_t^2\phi_x\phi_z^4-6\sigma^2\phi_y^2\phi_t\phi_z^4-8\phi_z^4\phi_t\phi_{xxx}\phi_x+9\phi_t\phi_z^4\phi_{xx}^2$$
$$+12\phi_x^2\phi_t\phi_z^2\phi_{xx}\phi_{zz}+6\phi_z^4\phi_{txx}\phi_x^2-12\phi_z^4\phi_{xx}\phi_{tx}\phi_z=0. \quad (70)$$

Substituting $x_5=x_4=z$ into (62)–(66), we get only three independent $(3+1)$-dimensional reductions:

$$-6\phi_z^4\phi_{xx}^2+16\phi_z^3\phi_{xx}\phi_x\phi_{xz}-8\phi_z^2\phi_x^2\phi_{xz}^2+6\sigma^2\phi_y^2\phi_z^4+2\phi_x^4\phi_{zzz}\phi_z$$
$$-\phi_x^4\phi_{zz}^2+2\phi_t\phi_x\phi_z^4-4\phi_x^2\phi_z^2\phi_{xx}\phi_{zz}=0, \quad (71)$$

$$6\sigma^2\phi_y^2\phi_z^4+16\phi_x\phi_z^3\phi_{xx}\phi_{xz}-4\phi_x^2\phi_z^2\phi_{xx}\phi_{zz}-8\phi_x^2\phi_z^2\phi_{xz}^2+12\phi_x^3\phi_z\phi_{xz}\phi_{zz}$$
$$+8\phi_x^4\phi_{zzz}\phi_z-13\phi_x^4\phi_{zz}^2+2\phi_t\phi_x\phi_z^4-6\phi_x^3\phi_{xzz}\phi_z^2-6\phi_z^4\phi_{xx}^2=0, \quad (72)$$

and

$$2\phi_t^2\phi_x\phi_z^4-6\phi_t\phi_z^4\phi_{xx}^2+16\phi_t\phi_z^3\phi_{xx}\phi_x\phi_{xz}-8\phi_t\phi_x^2\phi_x^2\phi_{xz}^2-4\phi_x^2\phi_t\phi_z^2\phi_{xx}\phi_{zz}-6\phi_x^4\phi_{tzz}\phi_z^2$$
$$+8\phi_x^4\phi_z\phi_t\phi_{zzz}+12\phi_x^4\phi_{zz}\phi_{tz}\phi_z-13\phi_x^4\phi_t\phi_{zz}^2+6\sigma^2\phi_y^2\phi_t\phi_z^4=0. \quad (73)$$

In the same way, substituting $x_5=x_4=y$ into (62)–(66) leads to four further $(3+1)$-dimensional reductions:

$$-12\phi_x^2\phi_z^2\phi_y^4\phi_{xx}\phi_{zz}+8\phi_x^2\phi_z^4\phi_y^2\phi_{xx}\phi_{yy}+2\phi_t\phi_x\phi_z^4\phi_y^4+6\sigma^2\phi_z^4\phi_y^6-6\phi_x^3\phi_z^4\phi_{xyy}\phi_y^2$$
$$+8\phi_x^4\phi_z^4\phi_y\phi_{yyy}+12\phi_x^3\phi_z^4\phi_{yy}\phi_{xy}\phi_y+16\phi_z^4\phi_y^3\phi_{xx}\phi_x\phi_{xy}-8\phi_z^4\phi_y^2\phi_x^2\phi_{xy}^2$$
$$-19\phi_x^4\phi_z^4\phi_{yy}^2-6\phi_z^4\phi_y^4\phi_{xx}^2+6\phi_x^4\phi_y^4\phi_{zz}^2=0, \tag{74}$$

$$-6\phi_x^4\phi_z^3\phi_{yyz}\phi_y^2+6\sigma^2\phi_z^4\phi_y^6+8\phi_x^2\phi_z^4\phi_y^2\phi_{xx}\phi_{yy}-19\phi_x^4\phi_z^4\phi_{yy}^2+2\phi_t\phi_x\phi_z^4\phi_y^4$$
$$+6\phi_x^4\phi_y^4\sigma_{zz}^2-8\phi_z^4\phi_y^2\phi_x^2\phi_{xy}^2+16\phi_z^4\phi_y^3\phi_{xx}\phi_x\phi_{xy}-6\phi_z^4\phi_y^4\phi_{xx}^2$$
$$+12\phi_x^4\phi_z^3\phi_y\phi_{yz}\phi_{yy}-12\phi_x^2\phi_z^2\phi_y^4\phi_{xx}\phi_{zz}+8\phi_x^4\phi_z^4\phi_y\phi_{yyy}=0, \tag{75}$$

$$6\sigma^2\phi_z^4\phi_y^6+8\phi_x^2\phi_z^4\phi_y^2\phi_{xx}\phi_{yy}-7\phi_x^4\phi_z^4\phi_{yy}^2+2\phi_t\phi_x\phi_z^4\phi_y^4+6\phi_x^4\phi_y^4\phi_{zz}^2-8\phi_z^4\phi_y^2\phi_x^2\phi_{xy}^2$$
$$+16\phi_z^4\phi_y^3\phi_{xx}\phi_x\phi_{xy}-6\phi_z^4\phi_y^4\phi_{xx}^2-12\phi_x^2\phi_z^2\phi_y^4\phi_{xx}\phi_{zz}+2\phi_x^4\phi_z^4\phi_y\phi_{yyy}=0, \tag{76}$$

$$2\phi_t^2\phi_x\phi_z^4\phi_y^4+6\sigma^2\phi_z^4\phi_t\phi_y^6+8\phi_x^4\phi_z^4\phi_t\phi_{yyy}\phi_y-19\phi_x^4\phi_z^4\phi_t\phi_{yy}^2-8\phi_z^4\phi_t\phi_y^2\phi_x^2\phi_{xy}^2$$
$$-12\phi_x^2\phi_z^2\phi_t\phi_y^4\phi_{xx}\phi_{zz}+6\phi_x^4\phi_t\phi_y^4\phi_{zz}^2-6\phi_z^4\phi_t\phi_y^4\phi_{xx}^2+16\phi_z^4\phi_t\phi_y^3\phi_{xx}\phi_x\phi_{xy}$$
$$-6\phi_x^4\phi_z^4\phi_{tyy}\phi_y^2+12\phi_x^4\phi_z^4\phi_{yy}\phi_{ty}\phi_y+8\phi_z^4\phi_x^2\phi_t\phi_y^2\phi_{xx}\phi_{yy}=0. \tag{77}$$

The contributions of the z, y and t part among Eqs. (24), (41), and (67) are same. The only difference in these three equations is the coefficients of the terms which possess the factors $\phi_x\phi_{xxx}$ and ϕ_{xx}^2. This difference cannot be ruled out by rescaling the independent and dependent variables x and ϕ. Using some different symmetry constraints on (62)–(66), we can also get various other types of $(3+1)$-dimensional Painlevé integrable models. For instance, using the constraints $\{\phi_{x_5}=\phi_x,\ \phi_y=0\}$ on (62)–(66) leads to four further different $(3+1)$-dimensional P-integrable models.

All the models (62)–(77) can now be considered as new seed models. If we use the extended Painlevé analysis approach and the assumption on these seeds, we can get various P-integrable models in $(6+1)$-dimensions, $(5+1)$-dimensions, $(4+1)$-dimensions, and $(3+1)$-dimensions with help of the symmetry reduction approach. But we do not discuss this problem further.

IV. SUMMARY AND DISCUSSION

Though the $(1+1)$-dimensional and $(2+1)$-dimensional integrable models have been widely applied in many scientific fields and deeply studied in mathematics, there is little progress in the study of the higher dimensional space especially in $(3+1)$ dimensions. One has not yet even found any $(3+1)$-dimensional completely integrable model while the real physics space-time is $(3+1)$-dimensional.

In this paper, after embedding the lower dimensional integrable models in higher dimensions and extending the standard WTC Painlevé analysis approach to a more general form such that the new expansion coefficients are conformal invariant, we have proposed an assumption to get some more P-integrable models in higher dimensions from a known lower dimensional one. Starting from the $(2+1)$-dimensional KP equation, some $(4+1)$-dimensional, $(5+1)$-dimensional and $(3+1)$-dimensional P-integrable models are given explicitly. All the equations are conformal invariant at the same time. Possessing the conformal invariance is also an important property of an integrable model. For example, infinitely many symmetries and the Darboux transformation of the KdV equation can be obtained from the conformal invariance of its Schwarz form.[29,17] Some quite general conformal invariant forms can be changed to the corresponding forms with the Painlevé property.[17,31] The $(3+1)$-dimensional Painlevé integrable models (24), (41), (42), (67)–(69), (71), (72), and (74)–(76) are all the generalizations of the PMKdV equation while Eqs. (42), (70), and (77) are not.

Though the assumption is based on that the original model is integrable and the result equations come from vanishing the consistent Painlevé expansion coefficients with the conformal

invariance, we still have to check the Painlevé property of the result models one by one because we have not yet proved the assumption strictly. Even if the assumption is not completely true, it is still possible to get some *P*-integrable models from the approach after checking the Painlevé property of the result equations.

To reduce a higher dimensional model to some lower ones, one can use some possible limiting procedure, the classical and nonclassical Lie approaches or the direct method.[21–23] Now we know that it is also possible to get some higher dimensional models from lower ones. For instance, one can get some suitable deformation relations,[32] inner parameter dependent symmetry constraints[32] and from the extended Painlevé analysis proposed in this paper.

Though we have obtained some $(3+1)$-dimensional *P*-integrable models from the well known KP equation, more about these equations should be studied further because the real physical space is $(3+1)$-dimensional and there is no knowledge on the $(3+1)$-dimensional soliton solutions. Especially, whether the obtained $(3+1)$-dimensional *P*-integrable models obtained here are really integrable under other meaning (say, whether the models possess Lax pairs, infinitely many symmetries or conservation laws) is still open. We hope that some mathematicians and physists will join in with us to solve this important problem in future studies.

ACKNOWLEDGMENTS

The work was supported partially by the National Nature Science Foundation of China and the Nature Science Foundation of Zhejiang province in China. We thank Professor Q.-P. Liu, Professor X.-B. Hu, Professor G.-X. Huang, and Professor G.-J. Ni for their helpful discussions.

[1] L. Dolan, Nucl. Phys. B **489**, 245 (1997); J. Distler and A. Hanany, *ibid.* **496**, 75 (1997); J. Ellis, N. E. Maoromatos, and D. V. Nanopoulos, Int. J. Mod. Phys. A **12**, 2639 (1997).
[2] I. Loutsenko and D. Roubtsov, Phys. Rev. Lett. **78**, 3011 (1997); M. W. Coffey, Phys. Rev. B **54**, 1279 (1996); R. Siddhartham and B. S. Shastry, *ibid.* **55**, 12196 (1997).
[3] M. Tajiri and H. Maesono, Phys. Rev. E **55**, 3351 (1997).
[4] G. C. Das, Phys. Plasmas **4**, 2095 (1997).
[5] M. Gedalin, T. C. Scott, and Y. B. Band, Phys. Rev. Lett. **78**, 448 (1997); N. Ne and D. Drummond, *ibid.* **78**, 4311 (1997); T. Georges, Opt. Lett. **22**, 679 (1997).
[6] H. Weigel, L. Gamberg, and H. Reinhardt, Phys. Lett. B **399**, 287 (1997); Phys. Rev. D **55**, 6910 (1997).
[7] T. Chiueh and T. P. Woo, Phys. Rev. E **55**, 1048 (1997).
[8] G. L. Huang, D. Y. Wand, D. J. Wu, H. de Feraudy, D. Le Queau, and M. Volwerk, J. Geophys. Res. **102**, 7217 (1997); O. A. Pokhotelov, D. O. Pokhotelov, M. B. Gokhberg, F. Z. Feygin, L. Stenflo, and P. K. Shukla, *ibid.* **101**, 7913 (1996).
[9] G. C. Das, M. Gogoi, and P. Deka, Planet. Space Sci. **42**, 993 (1994); G. C. Das and K. M. Sen, *ibid.* **42**, 41 (1994).
[10] A. Niiyama and M. Koshiba, IEICE Trans. Commun. **E80-B**, 522 (1997); A. M. Dunlop, W. J. Firth, and E. M. Wright, Opt. Commun. **138**, 211 (1997).
[11] B. A. Kalinikos, N. G. Kovshikov, and E. C. Patton, Phys. Rev. Lett. **78**, 2827 (1997).
[12] A. S. Shcherbakov, A. Yu Kosarsky, and V. N. Zvegintsev, Opt. Mem. Neural Netw. **5**, 313 (1996); M. Kinoshita, Y. Hirano, M. Kuwabara, and Y. Ono, J. Phys. Soc. Jpn. **66**, 703 (1997); L. Mutsson, J. Theor. Biol. **180**, 93 (1996).
[13] M. J. Ablowitz and P. A. Clarkson, *Solitons Nonlinear Evolution Equations and Inverse Scattering*, London Mathematical Society Lecture Note Series 149 (Cambridge University Press, Cambridge, 1991); R. S. Ward, Philos. Trans. R. Soc. London, Ser. A **315**, 451 (1985).
[14] R. Newton, Phys. Rev. Lett. **43**, 541 (1976); J. Math. Phys. **23**, 693 (1982); **23**, 594 (1982); A. S. Fokas, Phys. Rev. Lett. **51**, 3 (1983); **57**, 159 (1986); L. D. Faddeev, Sov. Phys. Dokl. **11**, 209 (1966); J. Math. Phys. **5**, 334 (1976); A. I. Nachman and M. J. Ablowitz, Stud. Appl. Math. **71**, 243 (1984); **71**, 251 (1984); M. J. Ablowitz and A. I. Nachman, Physica D **18**, 223 (1986); R. Beals and R. R. Coifman, *ibid.* **18**, 242 (1986); W. Weder, Inverse Probl. **6**, 267 (1990); **7**, 461 (1991); C. H. Gu, Lett. Math. Phys. **26**, 199 (1992).
[15] D. David, N. Kamran, D. Levi, and P. Winternitz, J. Math. Phys. **27**, 1225 (1986); B. Champagne and P. Winternitz, *ibid.* **29**, 1 (1988); K. M. Tamizhmani and B. Gramaticos, *ibid.* **32**, 2635 (1991); S. Lou and X. Hu, J. Phys. A **27**, L207 (1994); S. Lou, Phys. Rev. Lett. **71**, 4099 (1993).
[16] S. Lou, J. Yu, and J. Lin, J. Phys. A **28**, L191 (1995); S. Lou, J. Lin, and J. Yu, Phys. Lett. A **201**, 47 (1995).
[17] S. Lou, Sci. China A **34**, 1317 (1997).
[18] F. Calogero and A. Degasperis, Nuovo Cimento B **31**, 201 (1976); **39**, 54 (1997); O. I. Bogoyovlenskii, Usp. Mat. Nauk **53**, 234 (1989); **53**, 907 (1989); **54**, 1123 (1989).
[19] Y. Li and Y. Zhang, J. Phys. A **26**, 7487 (1993).
[20] S-y Lou, Commun. Theor. Phys. **28**, 41 (1997).
[21] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Applications in Mathematical Science, Vol. 81 (Springer, Berlin, 1989); P. J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Graduate Texts Mathematics, Vol. 107 (Springer, New York, 1993).
[22] P. A. Clarkson and M. D. Kruskal, J. Math. Phys. **30**, 2201 (1989); P. A. Clarkson and P. Winternitz, Physica D **49**, 257 (1991); P. A. Clarkson and E. L. Mansfield, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **54**, 1693 (1994); P. A. Clarkson, Chaos Solitons Fractals **5**, 2261 (1995).

[23] S. Lou, J. Phys. A **23**, L649 (1990); Phys. Lett. A **151**, 133 (1990). Sci. Math. Methods Appl. Sci. **18**, 789 (1995); S. Lou, H. Ruan, D. Chen, and W. Chen, J. Phys. A **24**, 1455 (1991).
[24] C. W. Cao, Henan Sci. **5**, 1 (1987); Chin. Quart. J. Math. **3**, 90 (1988); Sci. China A **33**, 528 (1990); C. W. Cao and X. G. Geng, J. Phys. A **23**, 4117 (1990); J. Math. Phys. **32**, 2323 (1991).
[25] Y. B. Zeng and Y. S. Li, J. Math. Phys. **30**, 1617 (1989); J. Phys. A **23**, L89 (1990); Y. B. Zeng, Phys. Lett. A **160**, 541 (1991).
[26] Y. Cheng and Y. S. Li, Phys. Lett. A **157**, 22 (1991); J. Phys. A **25**, 419 (1992); B. Konopenlcnenko, J. Sidorenko, and W. Strampp, Phys. Lett. A **157**, 17 (1991).
[27] S. Lou, Commun. Theor. Phys. **27**, 249 (1997); S. Lou, L. Chen, and Q. Wu, Chin. Phys. Lett. **14**, 1 (1997).
[28] S. Lou and X. Hu, J. Math. Phys. **38**, 6401 (1997).
[29] S. Lou, J. Phys. A **30**, 4803 (1997).
[30] S. Lou, Commun. Theor. Phys. **28**, 129 (1997).
[31] S. Lou, J. Math. Phys. **39**, 2112 (1998).
[32] S. Lou, J. Phys. A **30**, 7259 (1997).
[33] A. Ramani, B. Gramaticos, and T. Bountis, Phys. Rep. **180**, 159 (1989).
[34] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[35] M. C. Nucci, J. Phys. A **22**, 2897 (1989).
[36] R. Conte, Phys. Lett. A **140**, 383 (1989).
[37] N. A. Kudryashov, J. Phys. A **30**, 5445 (1997).

Rational solutions of the KP hierarchy and the dynamics of their poles. II. Construction of the degenerate polynomial solutions

Dmitry Pelinovsky[a)]
Department of Mathematics, University of Cape Town,
Rondebosch 7701, Western Cape, South Africa

(Received 15 September 1997; accepted for publication 3 June 1998)

A general approach to constructing the polynomial solutions satisfying various reductions of the Kadomtsev–Petviashvili (KP) hierarchy is described. Within this approach, the reductions of the KP hierarchy are equivalent to certain differential equations imposed on the τ-function of the hierarchy. In particular, the *l*-reduction and the *k*-constraint as well as their generalized counterparts are considered. A general construction of the rational solutions to these reductions is found and the particular solutions are explicitly derived for some typical examples including the KdV and Gardner equations, the Boussinesq and classical Boussinesq systems, the NLS and Yajima–Oikawa equations. It is shown that the degenerate rational solutions of the KP hierarchy are related to stationary manifolds of the Calogero–Moser (CM) hierarchy of dynamical systems. The scattering dynamics of interacting particles in the CM systems may become complicated due to an anomalously slow fractional-power rate of the particle motion along the stationary manifolds.
© *1998 American Institute of Physics.* [S0022-2488(98)03110-7]

I. INTRODUCTION

It is generally believed that all known hierarchies of $(1+1)$-dimensional equations integrable by means of the inverse scattering method can be represented as certain reductions of a universal Kadomtsev–Petviashvili (KP) hierarchy of $(2+1)$-dimensional equations and/or of its extentions to modified and multicomponent cases. This hypothesis originates from a unifying Sato theory[1–3] that describes the KP hierarchy in terms of the pseudodifferential operator $\mathscr{L}=\partial_x+u_2\partial_x^{-1}+u_3\partial_x^{-2}+\cdots$. The infinite set of functions u_n for $n\geqslant 3$ can be expressed through the scalar function $u_2=u(t_1,t_2,\ldots)$ that depends on $t_1=x$ and on an infinite sequence of the time variables t_k, $k\geqslant 2$. Then, equations of the KP hierarchy arise from the generalized Lax equation,

$$\frac{\partial \mathscr{L}}{\partial t_k}=[B_k,\mathscr{L}], \quad k\geqslant 1, \tag{1.1}$$

where B_k is a differential part of $\mathscr{L}^k$. A typical example is the KP equation (see Ref. 3),

$$-4\frac{\partial^2 u}{\partial t_1 \partial t_3}+6\frac{\partial^2 u^2}{\partial t_1^2}+\frac{\partial^4 u}{\partial t_1^4}+3\frac{\partial^2 u}{\partial t_2^2}=0. \tag{1.2}$$

Equations of the KP hierarchy are associated with the isospectral deformation of the eigenvalue problems $\mathscr{L}\psi=\lambda\psi$ and $\mathscr{L}^*\bar{\psi}=\lambda\bar{\psi}$, where λ is a spectral parameter, $\mathscr{L}^*$ is a formal operator adjoint, and ψ and $\bar{\psi}$ are eigenfunctions satisfying the time evolution problems,

$$\frac{\partial \psi}{\partial t_k}=B_k\psi, \quad \frac{\partial \bar{\psi}}{\partial t_k}=-B_k^*\bar{\psi}, \quad k\geqslant 1. \tag{1.3}$$

[a)]Present address: Department of Mathematics, University of Toronto, Toronto, Ontario M5S 3G3, Canada. Electronic mail: dmpeli@math.toronto.edu

 © 1998 American Institute of Physics

The KP hierarchy is closely related to the modified KP hierarchy which starts with the modified KP equation,

$$-4\frac{\partial^2 v}{\partial t_1 \partial t_3}-2\frac{\partial^2 v^3}{\partial t_1^2}+\frac{\partial^4 v}{\partial t_1^4}+3\frac{\partial^2 v}{\partial t_2^2}+6\frac{\partial}{\partial t_1}\left(\frac{\partial v}{\partial t_1}\,\partial_{t_1}^{-1}\,\frac{\partial v}{\partial t_2}\right)=0. \tag{1.4}$$

Within the Sato approach, the modified KP hierarchy can be constructed by means of the gauge transformation of the KP hierarchy which results in modification of the pseudodifferential operator $\mathscr{L}$.[4] On the other hand, solutions of the modified KP hierarchy are formally expressed through the eigenfunction ψ satisfying Eqs. (1.3) according to the relation, $v=\psi^{-1}\partial\psi/\partial t_1$.

Besides its own meaning, the KP hierarchy is useful for generating numerous integrable hierarchies of $(1+1)$-dimensional equations by means of a certain reductive procedure. The conventional l-reduction arises under the condition imposed to the operator $\mathscr{L}$,

$$\mathscr{L}^l=B_l, \quad l\geqslant 2, \tag{1.5}$$

so that the whole set of functions u_n become independent on t_l as well as on t_{nl} (see, e.g., Ref. 3). For $l=2$ this reduction leads to the KdV hierarchy, for $l=3$ it produces the Boussinesq hierarchy, and so on.

Another reductive procedure was recently proposed through a symmetry constraint[5] which was also formulated as the k-constraint imposed on the operator $\mathscr{L}$.[6] The k-constraint occurs under the condition,

$$\mathscr{L}^k=B_k+\psi\partial_x^{-1}\bar{\psi}, \quad k\geqslant 1, \tag{1.6}$$

and leads to coupled nonlinear equations between the functions $u_2,u_3,\ldots,u_{k-1}$ of the KP hierarchy and the eigenfunctions ψ and $\bar{\psi}$. Typical examples of the k-constrained KP hierarchy include the NLS equation for $k=1$, the Yajima–Oikawa equations for $k=2$, and so on (see Refs. 5, 6).

The reductive procedure applied to the KP hierarchy typically gives an idea as to which exact solutions, such as soliton and rational solutions, remain invariant under the reduction and hence represent the exact solutions of the $(1+1)$-dimensional equations.[7–11] This search for the explicit solutions is especially straightforward if one deals with the functions u_n of the KP hierarchy and the eigenfunctions ψ and $\bar{\psi}$ under the zero boundary conditions at infinity. In contrast, a structure of explicit solutions with nonzero boundary conditions is not so obvious and is not usually under consideration. Some results concerned with multisoliton solutions of the k-constrained KP hierarchy under the nonzero boundary conditions for the fields ψ and $\bar{\psi}$ were obtained quite recently by Loris and Willox[12,13] while the existence of rational solutions for this reduction remains an open question. Therefore, a natural problem is how to detect those rational solutions of the KP hierarchy that satisfy a generalized reduction or constraint imposed to the operator $\mathscr{L}$.

The generalized l-reduction of the KP hierarchy can be formulated by means of a modification of the condition (1.5) given by

$$\mathscr{L}^l-\alpha\mathscr{L}=B_l-\alpha B_1, \quad l\geqslant 2, \tag{1.7}$$

where the parameter α is arbitrary. This reduction describes solutions of the KP hierarchy which depend on t_l according to the form, $u_m=u_m(t_1+\alpha t_l,t_2,\ldots,t_{l-1},t_{l+1},\ldots)$. A typical example is the Boussinesq equation for acoustic waves[14] which arises from the condition (1.7) for $l=3$.

The generalized k-constraint was formulated by Loris and Willox[13] as the condition,

$$\mathscr{L}^k+\beta\mathscr{L}^{-1}=B_k+\psi\partial^{-1}\bar{\psi}, \quad k\geqslant 1, \tag{1.8}$$

where the constant β is related to the asymptotic values of the eigenfunctions ψ and $\bar{\psi}$ at infinity, i.e., $\psi\bar{\psi}\to\beta$ as $t_1\to\infty$. In particular, the NLS equation at the continuous-wave background[15] arises from the condition (1.8) for $k=1$.

This paper is devoted to studies of the rational solutions satisfying the generalized reductions (1.7) and (1.8). The main ideas and methods of this paper follow those of the previous paper[16] referred to henceforth as paper I. The aims of both papers are twofold; (i) to analyze and classify

various sets of rational solutions of the single-component KP hierarchy through the construction of the equivalent polynomial τ-function and (ii) to relate the different structure of the polynomials to different dynamics of particles in a many-body Calogero–Moser (CM) system and its higher commuting flows.[17,18]

The τ-function representation of solutions of the KP hierarchy appears naturally within the framework of the Hirota bilinear transformation (see, e.g., Ref. 3),

$$u=\frac{\partial^2}{\partial t_1^2}\log\,\tau,\quad \psi=c\,\frac{\tau^+}{\tau},\quad \bar{\psi}=\bar{c}\,\frac{\tau^-}{\tau},\quad v=\frac{\partial}{\partial t_1}\log\frac{\tau^+}{\tau}, \tag{1.9}$$

where c and $\bar{c}$ are arbitrary constants, the function $\tau=\tau(t_1,t_2,\ldots)$ is referred to as the τ-function of the KP hierarchy while the associated functions $\tau^{\pm}$ are expressed through τ as follows:

$$\tau^{\pm}=\tau\left(t_1\mp\frac{1}{\lambda},t_2\mp\frac{1}{2\lambda^2},\ldots\right)\exp\left(\pm\sum_{n=1}^{\infty}\lambda^n t_n\right). \tag{1.10}$$

It was recently proved by Shiota[19] (following the pioneer paper by Krichever[20]) that zeros of the polynomial τ-function of the KP hierarchy describe motion of particles in the whole hierarchy of the CM dynamical systems.

In paper I it was shown that a general polynomial τ-function of the KP hierarchy can be factorized through a set of partial polynomial solutions generated by derivatives of certain exponential functions with only one value of the spectral parameter λ. These partial solutions were referred to as the degenerate polynomials of the KP hierarchy. Different types of scattering of particles in the CM hierarchy were analyzed in paper I with the help of the factorizing identities and the degenerate polynomials were shown to be related to anomalous scattering of particles accompanying slow rates of the particle motion. The first example of such an anomalous scattering was presented by Gorshkov *et al.*[21] for the KP1 equation and later by Ward[22] and Ioannidou[23] for an integrable chiral model. Recently, it was shown by Ablowitz and Villarroel[24] that the degenerate rational solutions of the KP1 equation naturally appear in the inverse scattering formalism through a multiple pole expansion of the associated eigenfunction ψ (which is equivalent to the pole expansion of the Baker–Akhiezer function used in a geometric method).[20] However, in previous works a complete classification of the degenerate rational solutions of the KP hierarchy has not been proposed.

In this paper we develop a unifying approach to constructing the degenerate polynomial solutions of the KP hierarchy by using the theory of generalized Schur polynomials and vertex operators. This approach enables us to characterize all degenerate polynomials in terms of stationary manifolds of the CM hierarchy[25] and to find a simple representation of the rational solutions satisfying the generalized l-reduction and k-constraint of the KP hierarchy. A general scheme for constructing the degenerate polynomials is described and particular applications of the scheme to some physically meaningful nonlinear evolution equations are also given.

The paper is organized as follows. The basic notation and properties of the generalized Schur polynomials and the vertex operators acting on these polynomials are described in Sec. II. A connection of these polynomials with the stationary manifolds of the CM system is discussed in Sec. III. A general classification of the stationary manifolds and of the related rational solutions to the l-reduced KP hierarchy is given in Sec. IV. The rational solutions to the generalized l-reduction and k-constraint of the KP hierarchy are constructed in Sec. V and VI, respectively. A classification of different dynamics of the anomalous scattering of particles in the CM hierarchy is considered in Sec. VII. Finally, Sec. VIII concludes the paper.

II. THE GENERALIZED SCHUR POLYNOMIALS

Exact solutions of the KP hierarchy can be expressed through one of the conventional representations of the τ-function. Following paper I, only the Wronskian representation[26] is considered here but it enables us to generate all classes of rational solutions without loss of generality. Within the Wronskian representation, the τ-function is

$$\tau = W_N[\phi_1, \phi_2, \ldots, \phi_N] = \det \left| \frac{\partial^{k-1} \phi_n}{\partial t_1^{k-1}} \right|_{1 \leqslant n,k \leqslant N}, \tag{2.1}$$

where the entries of the Wronskian determinant $\phi_n = \phi_n(t_1, t_2, \ldots)$, $n = \overline{1,N}$ satisfy the differential equations

$$\frac{\partial \phi_n}{\partial t_k} = \frac{\partial^k \phi_n}{\partial t_1^k}, \quad k \geqslant 1. \tag{2.2}$$

A degenerate polynomial τ-function is generated by a partial solution of Eq. (2.2),

$$\phi_n = \left(\sum_{j=1}^{n} \alpha_j P_{m_j}(p) \right) e^{\Phi(p)}, \tag{2.3}$$

where α_j are arbitrary coefficients, p is a spectral parameter, the phase factor is given by

$$\Phi(p) = \sum_{k=1}^{\infty} p^k t_k, \tag{2.4}$$

and the functions $P_n(p)$ are defined by a generating exponential function,

$$P_n(p) = e^{-\Phi(p)} \frac{\partial^n}{\partial p^n} e^{\Phi(p)}. \tag{2.5}$$

These functions are quasipolynomials of the time variables t_k because they contain infinite series of t_k. However, the time variables t_k enter the quasipolynomials $P_n(p)$ only in the form of a finite number of the generalized time variables $\theta_k(p)$ given by

$$\theta_k(p) = \frac{1}{k!} \frac{\partial^k \Phi(p)}{\partial p^k} = \sum_{m=k}^{\infty} \binom{m}{k} p^{m-k} t_m. \tag{2.6}$$

Hence we refer to $P_n(p)$ as the generalized Schur polynomials. In the limit $p \to 0$ the polynomial $P_n(p)$ reduces to a conventional Schur polynomial $p_n(t_1, t_2, \ldots) = P_n(0)$ arising from the expansion $e^{\Phi(p)} = \sum_{n=1}^{\infty} (1/n!) p_n(t_1, t_2, \ldots) p^n$. The first few generalized polynomials have the form,

$$P_1 = \theta_1, \quad P_2 = \theta_1^2 + 2\theta_2, \quad P_3 = \theta_1^3 + 6\theta_1\theta_2 + 6\theta_3,$$

$$P_4 = \theta_1^4 + 12\theta_1^2\theta_2 + 12\theta_2^2 + 24\theta_1\theta_3 + 24\theta_4,$$

$$P_5 = \theta_1^5 + 20\theta_1^3\theta_2 + 60\theta_1\theta_2^2 + 60\theta_1^2\theta_3 + 120\theta_2\theta_3 + 120\theta_1\theta_4 + 120\theta_5.$$

The key property of the Schur polynomials $P_n(p)$ is that they satisfy the original system (2.2) rewritten in terms of the generalized variables θ_k and also the following relation:

$$\frac{\partial P_n}{\partial \theta_k} = \frac{\partial^k P_n}{\partial \theta_1^k} = \frac{n!}{(n-k)!} P_{n-k}. \tag{2.7}$$

Using this key property, Matveev[27] proved the following lemma.

Lemma 2.1: Suppose that $m_1 < m_2 < \cdots < m_n$ and $\alpha_n = 1$ in Eq. (2.3). Then, there exists constants θ_{k0}, $k = \overline{1,m_n}$ so that the superposition of the Schur polynomials in Eq. (2.3) is reducible to a polynomial of the highest degree, i.e.,

$$\sum_{j=1}^{n} \alpha_j P_{m_j}(p) = \widetilde{P}_{m_n}(p) \equiv P_{m_n}(p; \theta_1 - \theta_{10}, \theta_2 - \theta_{20}, \ldots, \theta_{m_n} - \theta_{m_n 0}). \tag{2.8}$$

According to Lemma 2.1, the degenerate rational solutions of the KP hierarchy can be expressed through the leading-order polynomials $\tilde{P}_{m_n}(p)$ with different values of the phase constants θ_{k0}. Then, the rational solution of this type has the form,

$$\tau = \tau_p(\theta_1, \theta_2, \ldots) e^{N\Phi(p)}, \tag{2.9}$$

where $\tau_p(\theta_1, \theta_2, \ldots)$ is a polynomial (or, equivalently, quasipolynomial) with respect to the time variables $\theta_n \equiv \theta_n(p)$ or t_n, respectively. This polynomial τ-function is given by the Wronskian determinant,

$$\tau_p(\theta_1, \theta_2, \ldots) = W_N[\tilde{P}_{m_1}(p), \tilde{P}_{m_2}(p), \ldots, \tilde{P}_{m_N}(p)]. \tag{2.10}$$

Using Eqs. (1.10) and (2.10), we find the associated functions $\tau^{\pm}$ for the polynomial representation,

$$\tau^{\pm} = \left(\frac{\lambda - p}{\lambda}\right)^{\pm N} \tau_p\left(\theta_1 \mp \frac{1}{\lambda - p}, \theta_2 \mp \frac{1}{2(\lambda - p)^2}, \ldots\right) \exp(N\Phi(p) \pm \Phi(\lambda)), \tag{2.11}$$

where the spectral parameter λ may be arbitrary. However, the rational representations for the eigenfunctions ψ and $\bar{\psi}$ arise only in the limit $\lambda \to 0$ provided that the constants c and $\bar{c}$ are chosen appropriately. Then, the rational solutions to the KP hierarchy can be expressed through a scalar polynomial function $\tau_p = \tau_p(\theta_1, \theta_2, \ldots)$ as follows:

$$u = \frac{\partial^2}{\partial \theta_1^2} \log \tau_p, \quad \psi = \rho \frac{S(\mu)\tau_p}{\tau_p}, \quad \bar{\psi} = \bar{\rho} \frac{S^{-1}(\mu)\tau_p}{\tau_p}, \quad v = \frac{\partial}{\partial \theta_1} \log \frac{S(\mu)\tau_p}{\tau_p}. \tag{2.12}$$

Here the parameters ρ and $\bar{\rho}$ define the boundary conditions of the eigenfunctions ψ and $\bar{\psi}$ as $t_1 \to \infty$, the parameter μ is $\mu = -p^{-1}$, and the generalized vertex operator $S(\mu)$ is introduced to characterize shifts of the phase variables $\theta_k \equiv \theta_k(p)$,

$$S(\mu) = \exp\left[-\sum_{k=1}^{\infty} \frac{1}{k} \mu^k \frac{\partial}{\partial \theta_k}\right]. \tag{2.13}$$

The action of the vertex operator on the generalized Schur polynomials $P_n(p)$ is described by the following lemma.

Lemma 2.2: If $P_n(p)$ satisfies Eq. (2.7) for any k and $S(\mu)$ is defined by Eq. (2.13), then the following identities are valid:

$$S^k(\mu) P_n(p) = \sum_{j=0}^{k} (-\mu)^j \binom{k}{j} \frac{\partial P_n(p)}{\partial \theta_j}, \tag{2.14}$$

$$S^{-k}(\mu) P_n(p) = \sum_{j=0}^{\infty} \mu^j \binom{j+k-1}{k-1} \frac{\partial P_n(p)}{\partial \theta_j}. \tag{2.15}$$

Lemma 2.2 has a classic analog in the theory of group characters (see, e.g., Ref. 28). Within the given context, the formulas (2.14) and (2.15) have already been used for the construction of integrable hierarchies[11] and that of the rational solutions.[14] Importantly, these formulas are not applicable to the action of vertex operators on the polynomial function $\tau_p(\theta_1, \theta_2, \ldots)$ because the polynomial function τ_p does not generally satisfy the key relations (2.7).

The multicomponent generalization of the vertex operator can be defined through the product of the vertex operators (2.13),

$$S(\nu_1, \ldots, \nu_k) = \prod_{i=1}^{k} S(\mu_i), \tag{2.16}$$

where the parameters ν_i are related to μ_j by the partial sums,

$$\nu_i=(-1)^{i-1}\sum_{z_1+\ldots+z_k=i}\mu_1^{z_1}\ldots\mu_k^{z_k}, \tag{2.17}$$

and the indices z_l are either zeros or unities. Using the formulas (2.14) and (2.16) one can prove the following lemma.

Lemma 2.3: If $P_n(p)$ satisfies Eq. (2.7) for any k and $S(\nu_1,\ldots,\nu_k)$ is defined by Eqs. (2.16) and (2.17), then the following identity is satisfied:

$$S(\nu_1,\ldots,\nu_k)P_n(p)=1-\sum_{i=1}^{k}\nu_i\frac{\partial P_n(p)}{\partial\theta_i}. \tag{2.18}$$

III. CONNECTION WITH THE CM HIERARCHY

Properties of the degenerate rational solutions of the KP hierarchy are described by the construction of the polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ within the form (2.10). The following theorem proved by Shiota[19] clarifies a relationship between this polynomial function and the dynamics of particles in the CM hierarchy of dynamical systems.

Theorem 3.1: Let the function $\tau_p(t_1,t_2,\ldots)$ be a monic polynomial in t_1 having R zeros $x_j(t_2,\ldots)$, $j=\overline{1,R}$ according to the representation,

$$\tau_p(t_1,t_2,\ldots)=\prod_{j=1}^{R}(t_1-x_j(t_2,\ldots)). \tag{3.1}$$

This function is the τ-function of the KP hierarchy if and only if the motion of zeros of τ_p is governed by the hierarchy of the CM dynamical systems,

$$\frac{\partial x_j}{\partial t_n}=\frac{\partial H_n}{\partial p_j},\quad \frac{\partial p_j}{\partial t_n}=-\frac{\partial H_n}{\partial x_j},\quad n\geqslant 2, \tag{3.2}$$

where $H_n=(-1)^n\mathrm{Sp}(L)^n$ and the elements of the matrix L are $L_{ij}=p_j\delta_{ij}+(1-\delta_{ij})(x_i-x_j)^{-1}$. In the particular case $n=2$, this system is just the Calogero–Moser system of particles,[17,18]

$$\frac{\partial^2 x_j}{\partial t_2^2}=8\sum_{i\neq j}\frac{1}{(x_i-x_j)^3}, \tag{3.3}$$

where x_j are coordinates of particles, while the velocities of particles are $v_j=\partial x_j/\partial t_2=2p_j$.

Collorary 3.2: The degree of the polynomial τ-function with respect to t_1 denoted by R coincides with the number of particles of the corresponding CM system.

Collorary 3.3: The number of linearly independent time variables $\theta_n(p)$ of the polynomial τ-function denoted by G coincides with the dimension of a dynamical manifold of the corresponding CM system.

A general solution describing normal scattering of R particles along a manifold of the dimension $G=2R$ is well known for the CM systems (3.2) and (3.3) (see, e.g., Ref. 19, and references therein). In paper I it was shown that besides a general solution there exists a wide set of degenerate solutions which describe anomalous scattering of particles with asymptotically equal velocities, i.e. when $v_j\to 2p$ as $t_2\to\pm\infty$. A complete picture of dynamical processes occurring in the CM systems can be explained as a superposition of elementary processes of either normal or anomalous scattering.[16] Here we study only the anomalous scattering in the CM hierarchy which is described by zeros of the degenerate polynomials $\tau_p(\theta_1,\theta_2,\ldots)$ according to Theorem 3.1.

Proposition 3.4: The degenerate polynomial $\tau_p(\theta_1,\theta_2,\ldots)$ given by Eq. (2.10) corresponds to the CM system of R particles, where

$$R=\sum_{j=1}^{N}m_j-\frac{N(N-1)}{2}. \tag{3.4}$$

Proof: Suppose that $0<m_1<m_2<\ldots<m_N$ in Eq. (2.10). The polynomials $P_n(p)$ at the leading order approximation with respect to θ_1 (or, equivalently, to t_1) are $P_n(p)\sim\theta_1^n+O(\theta_1^{n-2})$. Expanding the Wronskian $W_N[\tilde{P}_{m_1}(p),\ldots,\tilde{P}_{m_N}(p)]$ within this approximation, we find the result (3.4) for the degree of the polynomial $\tau_p(\theta_1,\theta_2,\ldots)$ which is equivalent to the number of particles R according to Collorary 3.2. □

The degenerate polynomials of the KP hierarchy and the corresponding dynamics of the CM systems can be classified into two different groups. The first group consists of the polynomials which can be reduced to the generalized Schur polynomials $P_R(p;\theta_1,\ldots,\theta_R)$. It is clear from the parametrization of these polynomials that a dynamical manifold of the CM system described by $P_R(p)$ has the dimension $G=R$ with the ''additional'' parameter p determining the asymptotic velocity of the particles as $t_2\to\infty$. The other group of the polynomial τ-functions which includes the rest of the polynomials τ_p describes the anomalous scattering occurring near one of the stationary manifolds of the CM hierarchy.[25]

Definition 3.5: A manifold of the CM hierarchy is stationary with respect to the time t_l if

$$\frac{\partial x_j}{\partial t_l}=0, \text{ for all } j \text{ and } l\geqslant 2. \tag{3.5}$$

Stationary manifolds of the CM hierarchy form certain embeddings into dynamical manifolds describing the anomalous dynamics of the CM particles. To be specific, the stationary manifolds have the dimension G which is less than the number of particles R, i.e., although all asymptotic velocities of particles are specified by a single parameter p, the asymptotic phases of the coordinates $x_j(t_2,\ldots)$ in the limit $t_2\to\infty$ are functionally dependent. The following result proves that all polynomial functions of the form (2.10) excepting those reducible to the polynomials $P_n(p)$ describe, within a particular limit, a certain stationary manifold of the CM hierarchy.

Proposition 3.6: Suppose that the polynomial τ-function of the KP hierarchy is given by

$$\tau_p(t_1,t_2,\ldots)=W_N[p_{m_1},p_{m_2},\ldots,p_{m_N}], \tag{3.6}$$

where $p_n\equiv p_n(t_1,t_2,\ldots)=P_n(0)$. Zeros of this polynomial function describe one of the stationary manifolds of the CM hierarchy if $N>1$ and the set $\{m_j\}_{j=1}^N$ is different from the ordered set $\{j\}_{j=1}^N$.

Proof: In the case $m_N<R$ it is clearly seen from Eq. (3.6) that the function $\tau_p(t_1,t_2,\ldots)$ is independent of at least the variables $t_{m_N+1},\ldots,t_R$ and hence, zeros of the polynomial function τ_p do not depend on these times.

In the case $m_N\geqslant R$ it follows from Eq. (3.4) that the only possible solution is $\tau_p(t_1,t_2,\ldots)=W_N[p_1,p_2,\ldots,p_{N-1},p_R]$ which is realized for $m_N=R$ and $R>N$. Then, it can be shown that the polynomial function τ_p does not depend on time t_R, i.e., corresponds to the stationary manifold with respect to t_R. In the exceptional case $R=N$ the condition of the proposition is not satisfied, and, indeed, the function τ_p is now reducible to the Schur polynomial $p_N(t_1,t_2,\ldots)$ (see paper I). □

In the next section we classify degenerate polynomial τ-functions of the KP hierarchy into different subgroups corresponding to different stationary manifolds of the CM hierarchy. This classification is directly related to the construction of rational solutions of the l-reduced KP hierarchy.

IV. THE *l*-REDUCED KP HIERARCHY

The following lemma can be readily proved by using the definition of the l-reduction (1.5) (see Ref. 3 for details).

Lemma 4.1: The polynomial τ-function of the l-reduced KP hierarchy satisfies the differential equation,

$$\frac{\partial\tau_p}{\partial t_l}=0, \qquad l\geqslant 2. \tag{4.1}$$

Collorary 4.2: Zeros of the function $\tau_p(t_1,t_2,\ldots)$ satisfying the l-reduced KP hierarchy define the stationary manifold of the CM hierarchy with respect to the time t_l.

Collorary 4.3: The τ-function of the l-reduced KP hierarchy does not depend on variables $t_l,t_{2l},\ldots,t_{nl}$.

Collorary 4.3 follows from the identity $L^{nl}=(B_l)^n=B_{nl}$ and leads to the fact that the τ-function of the l-reduced KP hierarchy also gives a subset of solutions of the nl-reduced hierarchies. In order to present a closed classification scheme for the rational solutions of the l-reduced KP hierarchy we assume that the function $\tau_p(t_1,t_2,\ldots)$ satisfies the complimentary constraints,

$$\frac{\partial\tau_p}{\partial t_j}\neq 0, \qquad j=\overline{1,l-1}. \tag{4.2}$$

Then, the following result defines a structural element of the polynomial τ-function of the l-reduced KP hierarchy.

Proposition 4.4: The differential equation (4.1) is satisfied by the $(l-1)$ families of the polynomial τ-functions,

$$\tau_p(t_1,t_2,\ldots)=W_{n_j}[p_jp_{j+l},\ldots,p_{j+(n_j-1)l}], \qquad j=\overline{1,l-1}, \tag{4.3}$$

where $p_n\equiv p_n(t_1,t_2,\ldots)=P_n(0)$ and n_j are positive integers enumerating the families of the polynomials.

A proof can be given by means of a direct differentiation of $\tau_p(t_1,t_2,\ldots)$ with respect to t_l, use the formula (2.7) for $p=0$, and elimination of two identical columns in the Wronskian determinants. This result has been widely used in previous literature (see, e.g., Ref. 11). A less known fact is that the $(l-1)$ families of the polynomials given by Eq. (4.3) do not cover all possible rational solutions of the l-reduced KP hierarchy except for the simple case $l=2$. Indeed, a more general solution can still be constructed through a combination of the particular solutions (4.3) in the following form:

$$\tau_p(t_1,t_2,\ldots)=\tau_l^{n_1n_2\ldots n_{l-1}}=W_N[p_1,p_{1+l},\ldots,p_{1+(n_1-1)l};\ldots;p_{l-1},p_{2l-1},\ldots,p_{n_{l-1}l-1}], \tag{4.4}$$

where $N=\Sigma_{j=1}^{l-1}n_j$. As follows from Collorary 4.2, zeros of this polynomial solution define a general construction of the stationary manifold of the CM hierarchy with respect to the time t_l. By using Eq. (3.4) we find that the number of particles forming a stationary configuration is not arbitrary but given by the expression,

$$R=\sum_{j=1}^{l-1}n_j\left(j+\frac{1}{2}(n_j-l+1)-\frac{1}{2}\sum_{i=1}^{l-1}(n_i-n_j)\right). \tag{4.5}$$

The dimension of the stationary manifold G is less than the number of particles R because the time variables t_l, t_{2l}, and so on, do not parametrize the function $\tau_l^{n_1n_2\ldots n_{l-1}}$. It seems to be difficult to find G for a general solution (4.4) (see, e.g., Ref. 14). Still for the particular solutions (4.3) we find the result,

$$R=\tfrac{1}{2}n_j(2j+(l-1)(n_j-1)), \quad G=j+(l-1)(n_j-1). \tag{4.6}$$

Example 4.5: The KdV hierarchy $(l=2)$.

The KdV hierarchy starts with the KdV equation,

$$-4\frac{\partial u}{\partial t_3}+12u\frac{\partial u}{\partial t_1}+\frac{\partial^3u}{\partial t_1^3}=0. \tag{4.7}$$

A complete set of the polynomial τ-functions for the KdV hierarchy[29,30] is given by Eq. (4.4) for $l=2$ and $n_1=n$, i.e.,

$$\tau_p(t_1,t_3,\ldots)=\tau_2^n=W_n[p_1,p_3,\ldots,p_{2n-1}]. \tag{4.8}$$

Zeros of these polynomials define the stationary manifolds for the CM dynamical system (3.3) with respect to t_2.[25] It follows from Eq. (4.6) that these stationary manifolds exist only for $R=n(n+1)/2$ particles and have $G=n$ parameter subspace.[29] For reference, we reproduce here the first polynomials of this family, omitting a constant factor in front of τ_2^n,

$$\tau_2^1 \cong t_1,$$

$$\tau_2^2 \cong t_1^3 - 3t_3,$$

$$\tau_2^3 \cong t_1^6 - 15t_1^3 t_3 - 45t_3^2 + 45t_1 t_5,$$

$$\tau_2^4 \cong t_1^{10} - 45t_1^7 t_3 - 4725t_1 t_3^3 + 315t_1^5 t_5 + 4725t_1^2 t_3 t_5 - 4725t_5^2 - 1575t_1^3 t_7 + 4725t_3 t_7.$$

Example 4.6: The Boussinesq hierarchy ($l=3$).

The Boussinesq hierarchy starts with the Boussinesq equation,

$$3\frac{\partial^2 u}{\partial t_2^2} + 6\frac{\partial^2 u^2}{\partial t_1^2} + \frac{\partial^4 u}{\partial t_1^4} = 0. \tag{4.9}$$

A general rational solution of the Boussinesq equation[14] is expressed through the polynomial τ-function (4.4) for $l=3$, $n_1=n$, and $n_2=m$, i.e.,

$$\tau_p(t_1,t_2,\ldots) = \tau_3^{nm} = W_{n+m}[p_1,p_4,\ldots,p_{3n-2};p_2,p_5,\ldots,p_{3m-1}]. \tag{4.10}$$

Zeros of these polynomials define the stationary manifold of the first higher-order commuting flow of the CM hierarchy (3.2), i.e., that with respect to t_3. It was found by Galkin *et al.*[14] that the number of particles R forming the stationary configuration and the dimension of the manifolds G are specified as follows:

$$R = n^2 + m(m+1) - nm, \tag{4.11}$$

$$G = \begin{cases} 2m-n & m \geqslant 2n \\ m+n & 2n > m \geqslant n/2. \\ 2n-m-1 & n/2 > m \end{cases} \tag{4.12}$$

The first polynomials are

$$\tau_3^{1,0} \cong t_1, \quad \tau_3^{0,1} \cong t_1^2 + 2t_2,$$

$$\tau_3^{2,0} \cong t_1^4 + 4t_1^2 t_2 - 4t_2^2 - 8t_4, \quad \tau_3^{1,1} \cong t_1^2 - 2t_2,$$

$$\tau_3^{0,2} \cong t_1^6 + 10t_1^4 t_2 + 20t_1^2 t_2^2 + 40t_2^3 - 40t_1^2 t_4 + 80t_2 t_4 - 80t_1 t_5,$$

$$\begin{aligned}\tau_3^{3,0} \cong\ & t_1^9 + 16t_1^7 t_2 + 56t_1^5 t_2^2 - 560t_1 t_2^4 - 112t_1^5 t_4 - 2240t_1 t_2^2 t_4 - 2240t_1 t_4^2 - 280t_1^4 t_5 + 1120t_1^2 t_2 t_5 \\ & + 1120t_2^2 t_5 - 2240t_4 t_5 + 1120t_1^2 t_7 + 2240t_2 t_7,\end{aligned}$$

$$\tau_3^{2,1} \cong t_1^4 - 4t_1^2 t_2 - 4t_2^2 + 8t_4, \quad \tau_3^{1,2} \cong t_1^5 - 20t_1 t_2^2 + 20t_5,$$

$$\begin{aligned}\tau_3^{0,3} \cong\ & t_1^{12} + 28t_1^{10} t_2 + 260t_1^8 t_2^2 + 1120t_1^6 t_2^3 + 2800t_1^4 t_2^4 + 11200t_1^2 t_2^5 + 11200t_2^6 - 280t_1^8 t_4 - 2240t_1^6 t_2 t_4 \\ & - 11200t_1^4 t_2^2 t_4 + 44800t_1^2 t_2^3 t_4 + 67200t_2^4 t_4 - 11200t_1^4 t_4^2 + 134400t_1^2 t_2 t_4^2 + 44800t_2^2 t_4^2 - 89600t_4^3 \\ & - 960t_1^7 t_5 - 4480t_1^5 t_2 t_5 - 44800t_1^3 t_2^2 t_5 - 89600t_1 t_2^3 t_5 - 89600t_1^3 t_4 t_5 + 179200t_1 t_2 t_4 t_5 \\ & - 89600t_1^2 t_5^2 + 179200t_2 t_5^2 + 8960t_1^5 t_7 - 179200t_1 t_2^2 t_7 + 179200t_5 t_7 + 22400t_1^4 t_8 + 89600t_1^2 t_2 t_8 \\ & - 89600t_2^2 t_8 - 179200t_4 t_8.\end{aligned}$$

We notice from these explicit formulas that the polynomial solutions related to the symmetry of the KP hierarchy with respect to the transformation $t_k \to (-1)^{k-1} t_k$ appear naturally from the general formula (4.10) for $m, n \neq 0$. □

A straightforward generalization of the polynomial solutions (4.3) which satisfies the k-constraint (1.6) of the KP hierarchy was recently presented by Loris and Willox.[11] Unfortunately, the rational solutions of the l-reduced and k-constrained KP hierarchies usually turn out to be physically meaningless since they are either singular or complex. However, the results on the existence and dimension of stationary manifolds of the CM hierarchy serve as an useful tool for constructing and analyzing the degenerate rational solutions which satisfy the generalized reductions of the KP hierarchy. The latter solutions often describe physically interesting phenomena within the underlying nonlinear evolution equations. In the next two sections we construct the rational solutions satisfying the generalized l-reduction and k-constraint of the KP hierarchy and present a few examples where the rational solutions are physically meaningful.

V. THE GENERALIZED l-REDUCED KP HIERARCHY

Here we consider the generalized l-reduction of the KP hierarchy defined by the condition (1.7). Using a simple modification of the methods of Ref. 3 one can prove the following lemma.

Lemma 5.1: The polynomial τ-function of the generalized l-reduced KP hierarchy satisfies the differential equation,

$$\frac{\partial \tau_p}{\partial t_l} = \alpha \frac{\partial \tau_p}{\partial t_1}, \quad l \geqslant 2. \tag{5.1}$$

As in Lemma 4.1, we assume here that there are no differential relations similar to Eq. (5.1) for the evolution of $\tau_p(t_1, t_2, \ldots)$ with respect to the time variables t_j for $j = \overline{1, l-1}$. The differential equation (5.1) implies that functions of the KP hierarchy depend on t_l in the form of a stationary phase, i.e., $u_k = u_k(t_1 + \alpha t_l, t_2, \ldots, t_{l-1}, t_{l+1}, \ldots)$. It is important that no rational solutions to the generalized l-reduction could be constructed by means of the Schur polynomials $p_n(t_1, t_2, \ldots)$ and, therefore, we have to consider the generalized Schur polynomials $P_n(p)$ given by Eq. (2.5) at $p \neq 0$. Thus, a set of the rational solutions of the generalized l-reduced KP hierarchy is prescribed by the following result.

Proposition 5.2: The differential equation (5.1) is satisfied by the $(l-1)$ families of the polynomial τ-functions,

$$\tau_p(\theta_1, \theta_2, \ldots) = W_n[P_1(p), S^{-1}(\nu_1, \ldots, \nu_{l-2}) P_3(p), \ldots, S^{-(n-1)}(\nu_1, \ldots, \nu_{l-2}) P_{2n-1}(p)]. \tag{5.2}$$

Here n is a positive integer, $P_n(p)$ is the generalized Schur polynomial (2.5), and the generalized vertex operator $S(\nu_1, \ldots, \nu_{l-2})$ is defined by Eqs. (2.16) and (2.17). The parameters ν_i for $i = \overline{1, l-2}$ are given by

$$\nu_i = -\binom{l}{2}^{-1} \binom{l}{i+2} p^{-i} \tag{5.3}$$

and the parameter p is one of the $(l-1)$ roots of the algebraic equation,

$$l p^{l-1} = \alpha. \tag{5.4}$$

Proof: First, we find that the polynomial function $\tau_p(\theta_1, \theta_2, \ldots)$ defined by Eq. (5.2) satisfies a certain differential equation with respect to the generalized time variables $\theta_k \equiv \theta_k(p)$. This can be done by differentiating each column of the Wronskian (5.2) with respect to θ_2, by subsequent subtracting of the nearbouring column multiplied by a certain factor and by using the formulas (2.7) and (2.18). As a result, we find the following differential equation for $\tau_p(\theta_1, \theta_2, \ldots)$,

$$\frac{\partial \tau_p}{\partial \theta_2} = \sum_{i=3}^{l} \nu_{i-2} \frac{\partial \tau_p}{\partial \theta_i}. \tag{5.5}$$

This relation should be compared with the differential equation (5.1) To do this, we rewrite Eq. (5.1) in terms of the variables θ_k as follows:

$$(lp^{l-1}-\alpha)\frac{\partial\tau_p}{\partial\theta_1}+\sum_{i=2}^{l}\binom{l}{i}p^{l-i}\frac{\partial\tau_p}{\partial\theta_i}=0. \tag{5.6}$$

The leading-order term $\partial\tau_p/\partial\theta_1$ is always dominant in Eq. (5.6) because the time variable θ_1 always has a different weight in the polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ compared to the other time variables. Therefore, we have to remove this term and specify p as a root of the algebraic equation (5.4). Then, by comparing the coefficients in Eqs. (5.5) and (5.6) we find the parameters ν_i in the form (5.3). □

It is obvious that the expression (5.2) generalizes the polynomial solutions (4.8) for the 2-reduced KP hierarchy (the KdV hierarchy) (see Example 4.5). The polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ given by Eq. (5.2) has the degree $R=n(n+1)/2$ and is parametrized by $G=n$ linear superpositions of the time variables θ_k (see examples below). This polynomial function generates a rational solution of the generalized l-reduced KP hierarchy according to the representation (2.12). A more general rational solution can still be constructed within the framework of the original Wronskian representation (2.1) where the functions ϕ_n are specified according to the particular $(l-1)$ solutions (5.2) [recall that the parameter p may have $(l-1)$ values]. This general rational solution can always be analyzed by applying the superposition formula found in paper I.

Example 5.4: The Gardner equation $(l=2)$.

The reduction $v=v(t_1+\alpha t_2,t_3,\ldots)$ transforms the modified KP hierarchy to a hierarchy of the Gardner equation. The Gardner equation follows from (1.4) in the form,

$$-4\frac{\partial v}{\partial t_3}+3(\alpha^2+2\alpha v-2v^2)\frac{\partial v}{\partial t_1}+\frac{\partial^3 v}{\partial t_1^3}=0. \tag{5.7}$$

Equivalently, the Gardner equation can be thought as the modified KdV equation under the nonzero boundary conditions at infinity. Indeed, the modified KdV equation,

$$-4\frac{\partial w}{\partial t_3}+\frac{9}{2}\alpha^2\frac{\partial w}{\partial t_1}-6w^2\frac{\partial w}{\partial t_1}+\frac{\partial^3 w}{\partial t_1^3}=0, \tag{5.8}$$

reduces to the Gardner equation (5.7) by means of the transformation, $w=v-\alpha/2$. A single family of the rational solutions of the Gardner equation is expressed through the representation (2.12) with the polynomial τ-function given by Eq. (5.2) for $l=2$, i.e.,

$$\tau_p(\theta_1,\theta_2,\ldots)=W_n[P_1(p),P_3(p),\ldots,P_{2n-1}(p)], \tag{5.9}$$

where $p=\alpha/2$. The first polynomials are just the polynomials of the KdV hierarchy (see Example 4.5) with t_k replaced by θ_k, i.e., $\tau_p=\tau_2^n(\theta_1,\theta_3,\ldots)$. These polynomials can further be used to construct the real nonsingular solutions of the (focusing) Gardner equation which follows from Eq. (5.7) under the transformation $\alpha\to i\tilde{\alpha}$ and $v\to i\tilde{v}$. To do this, we just shift phases of the variables θ_k in the function $\tau_p(\theta_1,\theta_2,\ldots)$ and rewrite Eq. (2.12) for v in the equivalent form,

$$v=\frac{\partial}{\partial\theta_1}\log\frac{S^{1/2}(\mu)\tau_p}{S^{-1/2}(\mu)\tau_p}, \tag{5.10}$$

where $\mu=-p^{-1}=-2\alpha^{-1}$. Then, the transformation given above ensures the function $\tilde{v}$ be real and nonsingular. Within this context, the rational solutions (5.9) and (5.10) describe an algebraic soliton of the Gardner equation for $n=1$,[31] a structural instability of the algebraic soliton due to a resonance with a long wave shelf for $n=2$,[32] and the scattering dynamics of two and more algebraic solitons complicated by excitations of the wave shelfs for $n\geqslant 3$.

Example 5.5: The Boussinesq equation for acoustic waves $(l=3)$.

The reduction $u=u(t_1+\alpha t_3,t_2,t_4,\ldots)$ transforms the KP hierarchy to a hierarchy of the Boussinesq equation for acoustic waves. The latter equation follows from Eq. (1.2) in the form,

$$3\frac{\partial^2 u}{\partial t_2^2}-4\alpha\frac{\partial^2 u}{\partial t_1^2}+6\frac{\partial^2 u^2}{\partial t_1^2}+\frac{\partial^4 u}{\partial t_1^4}=0. \tag{5.11}$$

Equivalently, the Boussinesq equation for acoustic waves can be thought of as the Boussinesq equation (4.9) under the nonzero boundary conditions at infinity. Indeed, the transformation, $u \to u-\alpha/3$ reduces Eq. (4.9) to the form (5.11). Two particular families of the rational solutions of the Boussinesq equation (5.11) are expressed through the polynomial τ-function (5.2) for $l=3$, i.e.,

$$\tau_p(\theta_1,\theta_2,\ldots)=W_n[P_1(p),S^{-1}(\nu)P_3(p),\ldots,S^{-(n-1)}(\nu)P_{2n-1}(p)], \tag{5.12}$$

where $p=\pm\sqrt{\alpha/3}$, and $S(\nu)$ is given by Eq. (2.13) with the parameter $\nu=-(3p)^{-1}$.[14] The first polynomials (5.12) can be transformed by means of the phase translations to the polynomials of the KdV hierarchy (see Example 4.5) with the residue terms, i.e., $\tau_p=\tau_2^n(z_1,z_3,\ldots)+\Delta\tau_2^n(z_1,z_3,\ldots)$, where z_k are new variables, i.e.,

$$z_1=\theta_1, \quad z_3=\theta_3+\nu\theta_2+\tfrac{1}{4}\nu^2\theta_1,$$

$$z_5=\theta_5+2\nu\theta_4+\tfrac{3}{2}\nu^2\theta_3+\tfrac{3}{2}\nu^3\theta_2+\tfrac{9}{16}\nu^4\theta_1,$$

$$z_7=\theta_7+3\nu\theta_6+\tfrac{15}{4}\nu^2\theta_5+\tfrac{9}{2}\nu^3\theta_4+\tfrac{63}{16}\nu^4\theta_3+\tfrac{63}{16}\nu^5\theta_2+\tfrac{105}{64}\nu^6\theta_1,$$

and $\Delta\tau_2^n$ are the residue terms, i.e.,

$$\Delta\tau_2^1=0, \quad \Delta\tau_2^2=0, \quad \Delta\tau_2^3=\tfrac{45}{4}\nu^6,$$

$$\Delta\tau_2^4=-\frac{42525}{8}\nu^8 z_1^2+\frac{14175}{4}\nu^6 z_1 z_3+\frac{382725}{64}\nu^{10}.$$

These polynomials can further be used to construct the real nonsingular rational solutions to the (elliptic) Boussinesq equation which follows from Eq. (5.11) under the transformation $t_2 \to iy$. This equation describes stationary wave processes in a medium with positive dispersion. The rational solutions to this equation correspond to multilump wave structures propagating with a constant velocity.[14,21]

VI. THE GENERALIZED *k*-CONSTRAINED KP HIERARCHY

Here we consider the generalized k-constraint of the KP hierarchy defined by the condition (1.8). Loris and Willox[13] proved the following result by using the technique based on a bilinear formulation and identities (see also Ref. 11).

Lemma 6.1: The τ-function of the generalized k-constrained KP hierarchy defined by Eq. (1.8) admits the Wronskian solutions in the form (2.1) with the functions $\phi_n(t_1,t_2,\ldots)$, $n=\overline{1,N}$ satisfying Eqs. (2.2) and the complimentary equation,

$$\frac{\partial\phi_n}{\partial t_k}+\beta\partial_{t_1}^{-1}\phi_n=c_n\phi_n, \tag{6.1}$$

where c_n is arbitrary and $\partial_{t_1}^{-1}$ is a formal operator of integration with respect to t_1.

This lemma is a modification of Theorem 3 by Loris and Willox.[13] Although they formulate the theorem for a particular exponential representation of the functions ϕ_n which is suitable for soliton solutions, the proof is based on the pseudodifferential equation (6.1) rather than on the explicit representation of the functions ϕ_n. In order to proceed with the rational solutions of the generalized k-constraint we apply a simple trick which essentially simplifies the analysis. Let us introduce an auxillary ''time'' variable t_{-1} according to the pseudo-differential equation,

$$\frac{\partial\phi_n}{\partial t_{-1}}=\partial_{t_1}^{-1}\phi_n. \tag{6.2}$$

Then, we can construct the polynomial function $\tau_p(t_1,t_2,\ldots;t_{-1})$ of the generalized k-constrained KP hierarchy according to the formulas (2.1)–(2.10) taking into account the auxillary equation (6.2). The only modification of this analysis is an extension of the generalized variables $\theta_k(p)$ to the form,

$$\theta_k(p)\to\tilde{\theta}_k(p)=\theta_k(p)+\frac{(-1)^k}{p(1+k)}\,t_{-1}. \tag{6.3}$$

As a result, the generalized k-constraint (1.8) can be reformulated for the polynomial solutions of the KP hierarchy as follows.

Lemma 6.2: The polynomial τ-function of the generalized k-constrained KP hierarchy satisfies the differential equation,

$$\frac{\partial\tau_p}{\partial t_k}+\beta\,\frac{\partial\tau_p}{\partial t_{-1}}=0,\quad k\geqslant 1. \tag{6.4}$$

For the particular case $k=1$ the differential equation (6.4) follows from the analysis of the nonlocal Boussinesq equation (see formulas (19) and (20) in Ref. 12) and also from the analysis of the Davey–Stewartson system at the continuous-wave background as a generalized reduction of the KP hierarchy (see formulas (19) and (31) in Ref. 15). The result of Lemma 6.2 enables us to construct a set of the rational solutions of the generalized k-constrained KP hierarchy.

Proposition 6.3: The differential equation (6.4) is satisfied by the $(k-1)$ families of the polynomial τ-functions,

$$\tau_p=W_n[P_1(p),S^{-1}(\nu_1,\ldots,\nu_{k-1})S(\mu)P_3(p),\ldots,S^{-(n-1)}(\nu_1,\ldots,\nu_{k-1})S^{n-1}(\mu)P_{2n-1}(p)], \tag{6.5}$$

where $S(\mu)$ and $S(\nu_1,\ldots,\nu_{k-1})$ are defined by Eqs. (2.13) and (2.16), respectively. The parameter μ is $\mu=-p^{-1}$, while the parameters ν_i for $i=\overline{1,k-1}$ are given by

$$\nu_i=-\binom{k+1}{2}^{-1}\binom{k+1}{i+2}p^{-i}. \tag{6.6}$$

Last, the parameter p is one of the $(k+1)$ roots of the algebraic equation,

$$kp^{k+1}=\beta. \tag{6.7}$$

Proof: First, we express Eq. (6.4) through the derivatives of $\tau_p(\tilde{\theta}_1,\tilde{\theta}_2,\ldots)$ with respect to the variables $\tilde{\theta}_k$ given by Eqs. (2.6) and (6.3). This leads to the differential equation for τ_p,

$$\left(kp^{k-1}-\frac{\beta}{p^2}\right)\frac{\partial\tau_p}{\partial\theta_1}+\sum_{i=2}^{k}\left[\binom{k}{i}p^{k-i}+\frac{(-1)^i\beta}{p^{i+1}}\right]\frac{\partial\tau_p}{\partial\theta_i}=\beta\sum_{i=k+1}^{\infty}\frac{(-1)^{i+1}}{p^{i+1}}\frac{\partial\tau_p}{\partial\theta_i}, \tag{6.8}$$

where we have omited the sign ''tilde'' for convenience. Moreover, after the transformation of the differential equation (6.4) to the form (6.8) we can treat the extra variable t_{-1} as a phase constant since the equations of the KP hierarchy are only expressed through the time variables t_k for $k\geqslant 1$.

The first term in Eq. (6.8) is removed if the parameter p is a root of the algebraic equation (6.7). The rest of terms in Eq. (6.8) disappear due to a special structure of the polynomial function $\tau(\theta_1,\theta_2,\ldots)$ given by (6.5). To show this, we apply a standard analysis described in the proof of Proposition 5.2 and find from Eq. (6.5) the differential relation for $\tau_p(\theta_1,\theta_2\ldots)$,

$$\frac{\partial\tau_p}{\partial\theta_2}=-\sum_{i=3}^{k}\mu^{i-2}\frac{\partial\tau_p}{\partial\theta_i}\left(1-\sum_{j=1}^{i-2}\frac{\nu_j}{\mu^j}\right)-\sum_{i=k+1}^{\infty}\mu^{i-2}\frac{\partial\tau_p}{\partial\theta_i}\left(1-\sum_{j=1}^{k-1}\frac{\nu_j}{\mu_j}\right). \tag{6.9}$$

Substituting Eq. (6.9) into Eq. (6.8) and taking into account Eq. (6.7) we find that the parameter p can be eliminated if μ is given $\mu=-p^{-1}$, while the left-hand side and right-hand-side of Eq. (6.8) reduce the system of linear equations for the coefficients $\nu_1,\ldots,\ \nu_{k-1}$,

$$1-\sum_{m=1}^{i-2}\frac{\nu_m}{\mu^m}=\frac{2}{k+1}\left[1+\frac{(k-1)!(-1)^i}{i!(k-i)!}\right] \tag{6.10}$$

for $3\leqslant i\leqslant k$ and

$$1-\sum_{m=1}^{k-1}\frac{\nu_m}{\mu^m}=\frac{2}{k+1}. \tag{6.11}$$

This linear algebraic system is not degenerate and the only solution is given by Eq. (6.6). Indeed, a direct substitution of Eq. (6.6) into Eqs. (6.10) and (6.11) produces the system of combinatorial identities,

$$\sum_{m=0}^{i}(-1)^m\binom{k+1}{m+2}=k+(-1)^i\binom{k}{i+2} \tag{6.12}$$

for $1\leqslant i\leqslant k-2$ and

$$\sum_{m=0}^{k-1}(-1)^m\binom{k+1}{m+2}=k. \tag{6.13}$$

For $i=1$ the equality (6.12) is obviously satisfied. Furthermore, the equality (6.12) is satisfied for any $(i+1)$ provided it is valid for i. This statement is a consequence of the Pascal' triangle identity,

$$\binom{k+1}{i+3}=\binom{k}{i+2}+\binom{k}{i+3}. \tag{6.14}$$

Finally, Eq. (6.13) is satisfied due to the combinatorial identity,

$$\sum_{m=-2}^{k-1}(-1)^m\binom{k+1}{m+2}=0. \qquad \square \tag{6.15}$$

By comparing the results of Propositions 5.2 and 6.3, we conclude that the only difference in the rational solutions of the generalized l-reduced and k-constrained KP hierarchies is the appearance of the vertex operator $S(\mu)$ in the columns of the Wronskian determinant (6.5). This additional operator provides an annihilation of the infinite sum in the right-hand side of Eq. (6.8) due to the remarkable property (2.15).

Example 6.4: The NLS hierarchy at the CW background $(k=1)$.

The NLS hierarchy at the continuous-wave (CW) background starts with the NLS equation written in the form,

$$-\frac{\partial\psi}{\partial t_2}+\frac{\partial^2\psi}{\partial t_1^2}+2(\psi\overline{\psi}-\beta)\psi=0,\quad \frac{\partial\overline{\psi}}{\partial t_2}+\frac{\partial^2\overline{\psi}}{\partial t_1^2}+2(\psi\overline{\psi}-\beta)\overline{\psi}=0, \tag{6.16}$$

where the functions ψ and $\overline{\psi}$ are supposed to have the nonzero boundary conditions at infinity, i.e., $\psi\overline{\psi}\to\beta$ as $t_1\to\infty$. Two particular families of rational solutions of the NLS hierarchy at the CW background are expressed by means of the formula (2.12) through the function τ_p given by Eq. (6.5) for $k=1$, i.e.,

$$\tau_p(\theta_1,\theta_2,\ldots)=W_n[P_1(p),S(\mu)P_3(p),\ldots,S^{n-1}(\mu)P_{2n-1}(p)], \tag{6.17}$$

where $p=\pm\sqrt{\beta}$ and $\mu=-p^{-1}$.[15] The first polynomials (6.17) can be transformed by means of the phase translations exactly to the polynomials of the KdV hierarchy (see Example 4.5), i.e., $\tau_p=\tau_2^n(z_1,z_3,\ldots)$, where z_k are new variables, i.e.,

$$z_1=\theta_1, \quad z_3=\theta_3-\mu\theta_2+\tfrac{1}{4}\mu^2\theta_1,$$

$$z_5=\theta_5-2\mu\theta_4+\tfrac{3}{2}\mu^2\theta_3-\tfrac{1}{2}\mu^3\theta_2+\tfrac{1}{16}\mu^4\theta_1,$$

$$z_7=\theta_7-3\mu\theta_6+\tfrac{15}{4}\mu^2\theta_5-\tfrac{5}{2}\mu^3\theta_4+\tfrac{15}{16}\mu^4\theta_3-\tfrac{3}{16}\mu^5\theta_2+\tfrac{1}{64}\mu^6\theta_1.$$

These polynomials can further be used to construct the real nonsingular rational solutions to the (focusing) NLS equation which arises under the transformation $t_2\to it$ and the reduction $\overline{\psi}=\psi^*$, where ψ^* is complex conjugate to ψ. These rational solutions describe the modulational instability of a CW background under a localized perturbation of the algebraic profile.[33,34]

In addition, the same set of the polynomial τ-function also gives the rational solutions of another physically important equation which is the Kaup equation (or, equivalently, the classical Boussinesq system).[35] The polynomials $\tau_p(\theta_1,\theta_2,\ldots)$ given by Eq. (6.17) generate the rational solutions of the Kaup equation written in the form,

$$\frac{\partial^2 v}{\partial t_2^2}-4\beta\frac{\partial^2 v}{\partial t_1^2}-\frac{\partial^4 v}{\partial t_1^4}-2\frac{\partial v}{\partial t_1}\frac{\partial^2 v}{\partial t_1\partial t_2}+2\frac{\partial}{\partial t_1}\left[\left(\frac{\partial v}{\partial t_1}\right)^3-\frac{\partial v}{\partial t_1}\frac{\partial v}{\partial t_2}\right]=0, \tag{6.18}$$

within an equivalent bilinear representation (5.10).[36]

Example 6.5: The Yajima–Oikawa hierarchy ($k=2$).

The Yajima–Oikawa hierarchy starts with the system,

$$\frac{\partial u}{\partial t_2}=\frac{\partial\psi\overline{\psi}}{\partial t_1}, \quad \frac{\partial\psi}{\partial t_2}=\frac{\partial^2\psi}{\partial t_1^2}+2u\psi, \quad -\frac{\partial\overline{\psi}}{\partial t_2}=\frac{\partial^2\overline{\psi}}{\partial t_1^2}+2u\overline{\psi}, \tag{6.19}$$

where we impose the boundary conditions, $u\to 0$ and $\psi\overline{\psi}\to\beta$ as $t_1\to\infty$. Three particular families of the rational solutions of the Yajima–Oikawa hierarchy are expressed by means of the formula (2.12) through the function τ_p given by Eq. (6.5) for $k=2$, i.e.,

$$\tau_p(\theta_1,\theta_2,\ldots)=W_n[P_1(p),S^{-1}(\nu)S(\mu)P_3(p),\ldots,S^{-(n-1)}(\nu)S^{n-1}(\mu)P_{2n-1}(p)], \tag{6.20}$$

where p may have three values, $p=(\beta/2)^{1/3}$ and $p=(\beta/2)^{1/3}(-1\pm\sqrt{3}i)/2$, $S(\mu)$ and $S(\nu)$ are both given by Eq. (2.13) with the parameters $\mu=-p^{-1}$ and $\nu=-(3p)^{-1}$. The first polynomials (6.20) can be transformed by means of the phase translations to the polynomials of the KdV hierarchy (see Example 4.5) with the residue terms, i.e., $\tau_p=\tau_2^n(z_1,z_3,\ldots)+\Delta\tau_2^n(z_1,z_3,\ldots)$, where z_k are new variables, i.e.,

$$z_1=\theta_1, \quad z_3=\theta_3+(\nu-\mu)\theta_2+\tfrac{1}{4}(\nu-\mu)^2\theta_1,$$

$$\begin{aligned}z_5=\theta_5&+2(\nu-\mu)\theta_4+\tfrac{3}{2}(\nu-\mu)^2\theta_3+\tfrac{1}{2}(3\nu^3-5\nu^2\mu+3\nu\mu^2-\mu^3)\theta_2\\&+\tfrac{1}{16}(9\nu^4-20\nu^3\mu+14\nu^2\mu^2-4\nu\mu^3+\mu^4)\theta_1,\end{aligned}$$

$$\begin{aligned}z_7=\theta_7&+3(\nu-\mu)\theta_6+\tfrac{15}{4}(\nu-\mu)^2\theta_5+\tfrac{1}{2}(9\nu^3-19\nu^2\mu+15\nu\mu^2-5\mu^3)\theta_4+\tfrac{1}{16}(63\nu^4-156\nu^3\mu\\&+138\nu^2\mu^2-60\nu\mu^3+15\mu^4)\theta_3+\tfrac{1}{16}(63\nu^5-147\nu^4\mu+130\nu^3\mu^2-58\nu^2\mu^3+15\nu\mu^4-3\mu^5)\theta_2\\&+\tfrac{1}{64}(105\nu^6-294\nu^5\mu+303\nu^4\mu^2+359\nu^3\mu^3+39\nu^2\mu^4-6\nu\mu^5+\mu^6)\theta_1,\end{aligned}$$

and $\Delta\tau_2^n$ are the residue terms, i.e.,

$$\Delta\tau_2^1=0, \quad \Delta\tau_2^2=0, \quad \Delta\tau_2^3=\tfrac{45}{4}\nu^4(\nu-\mu)^2,$$

$$\Delta\tau_2^4=\frac{798525}{64}\,\nu^3\mu^3 z_1^4-\frac{4725}{8}\,\nu^4(9\nu^4-24\nu^3\mu+22\nu^2\mu^2-8\nu\mu^3+\mu^4)z_1^2$$
$$+\frac{14175}{64}\,\nu^3(16\nu^3-32\nu^2\mu+16\nu\mu^2-169\mu^3)z_1z_3$$
$$+\frac{4725}{64}\,\nu^4(81\nu^6-270\nu^5\mu+351\nu^4\mu^2-228\nu^3\mu^3+79\nu^2\mu^4-14\nu\mu^5+\mu^6).$$

Within a physical context, the Yajima–Oikawa system (6.19) transformed according to the reduction $t_2\to it$ and $\psi=i\psi^*$ describes a resonance of a mean flow and short dispersive waves. The rational solutions to this system have not yet been analyzed. However, by an analogy with the NLS hierarchy at the CW background (see Example 6.4), these solutions seem to be relevant for description of the modulational instability of a CW background under a localized perturbation of the algebraic profile.

VII. ANOMALOUS SCATTERING IN THE CM HIERARCHY

We have shown that all degenerate rational solutions of the KP hierarchy excepting those reducible to the generalized Schur polynomials $P_n(p)$ describe, within a particular limit, the stationary manifolds of the CM hierarchy (see Proposition 3.6). Hence, in a general case, these solutions describe the scattering dynamics of interacting particles in a neighborhood of the stationary manifolds. The stationary manifolds of the CM system (3.3), i.e., those with respect to the time t_2, are especially important because they are relevant for the rational solutions satisfying the generalized reductions of the KP hierarchy. Here we study the dynamical processes of scattering of the CM particles associated with this family of the rational solutions.

Proposition 7.1: Suppose that the τ-function of the KP hierarchy is given by

$$\tau_p(\theta_1,\theta_2,\ldots)=W_N[\widetilde{P}_1(p),\widetilde{P}_3(p),\ldots,\widetilde{P}_{2N-1}(p)], \tag{7.1}$$

where the tilde means that the translations of the time variables $\theta_k=\theta_k(p)$ may be arbitrary. Define the index s for the scattering rate of R particles according to the asymptotic representation $x_j-2pt_2\sim t_2^s$ as $t_2\to\infty$ for $j=\overline{1,R}$, where $x_j(t_2,\ldots)$ are zeros of the τ-function given by Eq. (3.1). Then, the scattering dynamics described by the polynomials (7.1) occur near a stationary manifold of the CM system (3.3) of $R=N(N+1)/2$ particles, and the scattering rate s may only have the particular values,

$$s=s_{k,m}=\frac{k}{2k+m}, \tag{7.2}$$

where for each integer $k=\overline{1,N-1}$ there exists a set of integers $m=\{2l-1\}_{l=1}^{N-k}$.

Proof: Noting that the τ-function given by Eq. (7.1) can be reduced, within a particular limit, to the form $\tau(t_1,t_3,\ldots)=W_N[p_1,p_3,\ldots,p_{2N-1}]$ which describes a stationary manifold of the CM system (3.3) of $R=N(N+1)/2$ particles according to Collorary (4.2) and Proposition 4.4. Therefore, a more general form (7.1) of the τ-function describes the scattering dynamics of R particles occurring near this stationary manifold. In order to find the scattering rate of the particle dynamics, we need to show that zeros of the polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ have the asymptotic representation $\theta_1\sim\theta_2^s$ as $\theta_1,\ \theta_2\to\infty$, where the index s is given by Eq. (7.2).

The polynomial solution (7.1) has generally $G=R=N(N+1)/2$ parameters. Among these parameters, $G_1=N$ parameters correspond to the phases $\theta_1,\ \theta_3,\ldots,\theta_{2N-1}$ which define the configuration of the stationary manifold of the CM system while $G_2=N(N-1)/2$ parameters can be regarded as amplitudes of elementary excitations of particles near the stationary manifold. We show that an individual elementary excitation can be described by the particular polynomial function τ_p following from Eq. (7.1),

$$\tau_p(\theta_1,\theta_2,\ldots)=\tau_N^{m,k}=W_N[P_1(p),\ldots,S_m^{-1}(a)P_{1+2k}(p),\ldots,S_m^{-2}(a)P_{1+4k}(p),\ldots], \tag{7.3}$$

where k and m are the same integers as below Eq. (7.2), while the operator $S_m(a)$ is given by a modification of Eq. (2.13) in the form,

$$S_m(a)=\exp\left[-\sum_{i=1}^{\infty}\frac{1}{i}\,a^i\,\frac{\partial}{\partial\theta_{mi}}\right]. \tag{7.4}$$

Using the standard analysis, one can readily show that the polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ given by Eq. (7.3) depends on the phases θ_{2k+m} and θ_{2k} only in the form of a linear superposition $(\theta_{2k+m}+a\,\theta_{2k})$. Next, because the time variables θ_{2k+m} define the stationary manifold of the CM system, zeros of the polynomial $\tau_N^{m,k}$ have the asymptotic representation $\theta_1\sim\theta_{2k+m}^{1/(2k+m)}$ as θ_1, $\theta_{2k+m}\to\infty$. Substituting the linear superposition in this asymptotic representation and recalling that $\theta_{2k}\sim\theta_2^k$ as θ_2, $\theta_{2k}\to\infty$, we conclude that the (m,k) elementary excitation of the stationary manifold of the CM system prescribed by Eq. (7.4) displays the fractional scattering rate s defined by Eq. (7.2). Finally, it is clear that there exists exactly $G_2=N(N-1)/2$ particular representations (7.4) for each polynomial function (7.1). Therefore, all possible asymptotical scattering rates are listed in the formula (7.2). □

Example 7.2: The stationary manifold of 6 particles $(N=3)$.

The following polynomial τ-function of the KP hierarchy follows from Eq. (7.1) at $N=3$ and describes the scattering dynamics of $R=6$ particles occurring near the corresponding stationary manifold of the CM system (3.3),

$$\tau_p=z_1^6-15z_1^3z_3-45z_3^2+45z_1z_5+\Delta, \tag{7.5}$$

where

$$z_1=\theta_1,\quad z_3=\theta_3+b\,\theta_2+\tfrac{1}{64}(16b^2+9a^2)\,\theta_1,$$

$$z_5=\theta_5+2b\,\theta_4+\frac{3}{32}\,(16b^2-a^2)\,\theta_3+a\,\theta_2^2+c\,\theta_2-\frac{3}{512}\,a(16b^2+9a^2)\,\theta_1^2-\frac{3}{8}\,\alpha\theta_1\theta_3$$
$$+\frac{5}{8}\,ab\,\theta_1\theta_2-\frac{1}{2048}\,(81a^4+1360a^2b^2+384b^4-1024bc)\,\theta_1,$$

and

$$\Delta=-\frac{225}{8}\,az_5-\frac{135}{512}\,a(19a^2+256b^2)z_3+45a^2\theta_2^2-\frac{45}{32}\,a(9a^2b+16b^3-32c)\,\theta_2$$
$$-\frac{45}{262144}\,(2781a^6+32704a^4b^2+129024a^2b^4-16384b^6$$
$$+36864a^2bc+65536b^3c-65536c^2).$$

Here the variables z_1, z_3, and z_5 define the $G_1=3$ parameters of the stationary manifold, while the constants a, b, and c correspond to the $G_2=3$ elementary excitations of the particles near the manifold. It follows from zeros of the polynomial function (7.5) that if $a\neq0$ the scattering rate s is $s_{2,1}=2/5$ according to Eq. (7.2). If $a=0$ but $b\neq0$, then the scattering rate is slower, i.e., $s_{1,1}=1/3$. At last, if $a=b=0$ but $c\neq0$ then the rate is still slower, i.e., $s_{1,2}=1/5$. For a, b, and c all equal to zero, the particles are not excited and lie on the stationary manifold of the CM system.□

Thus, we conclude that the existence of the stationary manifolds of the CM system leads to a slowing down of the anomalous scattering of interacting particles according to a hierarchy of characteristic scattering rates (7.2). It is worthwhile to notice that the rates of anomalous scattering occurring far from the stationary manifolds of the CM systems are higher than those given by Eq. (7.2). Indeed, zeros of the degenerate polynomials of the KP hierarchy including the generalized Schur polynomials $P_n(p)$ have generally an asymptotic representation $\theta_1\sim\theta_2^{1/2}$ as θ_1, $\theta_2\to\infty$, i.e., the scattering rate is generally $s=1/2$. Only if the degenerate polynomials reduce to the form (7.1) the scattering rate becomes slower with respect to the time t_2, i.e., $s_{k,m}<1/2$. A first example

of the anomalous scattering occurring near the stationary manifolds of the CM system was presented by Gorshkov *et al.*[21] and recently reproduced by Ablowitz and Villarroel for the lump solutions in the KP1 equation.[24] Although we do not consider here the rational solutions of the KP1 equation, the results presented can further be generalized to describe the same scattering dynamics of the lump solitons as the particles in the CM system (see also Ref. 21).

VIII. CONCLUSION

In this paper we have developed a general approach to construct and analyze the degenerate rational solutions of equations of the KP hierarchy and its reductions. In particular, we have presented the rational solutions to the generalized l-reduction and k-constraint of the KP hierarchy which correspond to the reductions of the KP hierarchy under the nonzero boundary conditions at infinity. We have also shown that these solutions describe anomalously slow dynamical processes of scattering occurring near the stationary manifolds of the CM dynamical system.

Since the paper does not cover all possible generalized reductions of the KP hierarchy (see, e.g., Refs. 5, 6 for other reductions) it is worthwhile to summarize our approach and present a general scheme of a search for the polynomial τ-function satisfying a generalized reduction of the KP hierarchy. Suppose that the reduction is formulated in terms of a differential equation imposed on the τ-function of the KP hierarchy. Then, in order to find the particular form of the polynomial function $\tau_p(\theta_1,\theta_2,\ldots)$ (2.10) satisfying this reduction, one needs

(i) to rewrite the differential equation in terms of the generalized time variables $\theta_k(p)$;
(ii) to choose p such that the leading-order term $\partial\tau_p/\partial\theta_1$ is removed;
(iii) to find the first nonzero coefficient in front of the derivative $\partial\tau_p/\partial\theta_m$, then the polynomials τ_p generalize the polynomials of the m-reduced KP hierarchy given by (4.4);
(iv) to introduce the generalized vertex operators (2.13) and (2.16) which displace phases of the polynomials $P_n(p)$ in each subsequent column of the Wronskian determinant (2.10);
(v) to find the parameters of the vertex operators by comparing the differential equation derived for the polynomial function (2.10) with that of the given reduction.

This scheme and the approaches used in both papers I and II can be developed for analysis of the rational solutions of multicomponent KP hierarchies which include such an important example as the Davey–Stewartson system (see, e.g., Ref. 15). The τ-function of the multicomponent KP hierarchies can be expressed through a multicomponent Wronskian[26] which can be treated in the same manner as that in a single-component case. The classification and construction of rational solutions of the multicomponent KP hierarchies remain open for further studies.

ACKNOWLEDGMENTS

This work was inspired by discussions with Professor M. J. Ablowitz and Dr. I. Loris and was completed during the author's visit to the Vrije Universiteit Brussel in February, 1997. The author is very grateful to Professor F. Lambert for his warm hospitality and encouragement. The author acknowledges the financial support by the DWTC of the Belgium Government within the framework of IUAP-IV.

[1] M. Sato and Y. Sato, ''Soliton equations as dynamical systems on infinite dimensional Grassmann manifold,'' in Proceedings of the U.S.–Japan seminars, *Nonlinear Partial Differential Equations in Applied Science, Tokyo, 1982*, edited by P. D. Lax and H. Fujita (North-Holland, Kinokuniya, 1982), pp. 259–271.
[2] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, ''Transformation groups for soliton equations,'' in *Proceedings of the RIMS Symposium on Nonlinear Integrable Systems: Classical Theory and Quantum Theory, Kyoto*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983), pp. 39–119.
[3] Y. Ohta, J. Satsuma, D. Takahashi, and T. Tokihiro, ''An elementary introduction to Sato theory,'' Prog. Theor. Phys. Suppl. **94**, 210–241 (1988).
[4] W. Oevel and C. Rogers, ''Gauge transformations and reciprocal links in $2+1$ dimensions,'' Rev. Math. Phys. **5**, 299–330 (1993).
[5] B. Konopelchenko and W. Strampp, ''New reductions of the Kadomtsev–Petviashvili and two-dimensional Toda lattice hierarchies via symmetry constraints,'' J. Math. Phys. **33**, 3676–3686 (1992).
[6] Y. Cheng, ''Constraints of the KP hierarchy,'' J. Math. Phys. **33**, 3774–3782 (1992).
[7] Y. Zhang and Y. Cheng, ''Solutions of the vector k-constrained KP hierarchy,'' J. Math. Phys. **35**, 5869–5884 (1994).
[8] Y. Zhang and Y. Cheng, ''The Wronskian structure for solution of the k-constrained KP hierarchy,'' J. Phys. A **27**, 4581–4588 (1994).

[9] W. Oevel and W. Strampp, ''Wronskian solutions of the constrained KP hierarchy,'' J. Math. Phys. **37**, 6213–6219 (1996).
[10] I. Loris and R. Willox, ''Bilinear form and solutions of the *k*-constrained KP hierarchy,'' Inverse Probl. **13**, 411–420 (1997).
[11] I. Loris and R. Willox, ''On solutions of constrained KP equations,'' J. Math. Phys. **38**, 283–291 (1997).
[12] I. Loris and R. Willox, ''Soliton solutions of Wronskian form to the nonlocal Boussinesq equation,'' J. Phys. Soc. Jpn. **65**, 383–388 (1996).
[13] I. Loris and R. Willox, ''KP symmetry reductions and a generalized constraint,'' J. Phys. A **30**, 6925–6938 (1997).
[14] V. M. Galkin, D. E. Pelinovsky, and Yu. A. Stepanyants, ''The structure of the rational solutions to the Boussinesq equation,'' Physica D **80**, 246–255 (1995).
[15] D. Pelinovsky, ''On a structure of the explicit solutions to the Davey–Stewartson equations,'' Physica D **87**, 115–122 (1995).
[16] D. Pelinovsky, ''Rational solutions of the KP hierarchy and the dynamics of their poles. I. New form of a general rational solution,'' J. Math. Phys. **35**, 5820–5830 (1994) (Paper I).
[17] F. Calogero, ''Exactly solvable one-dimensional many-body problems,'' Lett. Nuovo Cimento **13**, 411–416 (1975).
[18] J. Moser, ''Dynamical systems, finitely many mass points on the line under the influence of an exponential potential-an integrable system,'' in *Dynamical Systems, Theory and Applications*, Lecture Notes in Physics, edited by J. Moser (Springer, New York, 1975), Vol. 38.
[19] T. Shiota, ''Calogero–Moser hierarchy and KP hierarchy,'' J. Math. Phys. **35**, 5844–5849 (1994).
[20] I. Krichever, ''Rational solutions of Kadomtsev–Petviashvili equation and integrable systems of *n* particles on a line,'' Funct. Anal. Appl. **12**, 59–61 (1978).
[21] K. A. Gorshkov, D. E. Pelinovsky, and Yu. A. Stepanyants, ''Normal and anomalous scattering, formation and decay of bound states of two-dimensional solitons described by the Kadomtsev–Petviashvili equation,'' JETP **77**, 237–245 (1993).
[22] R. S. Ward, ''Nontrivial scattering of localized solitons in a $(2+1)$-dimensional integrable system,'' Phys. Lett. A **208**, 203–208 (1995).
[23] T. Ioannidou, ''Soliton solutions and nontrivial scattering in an integrable chiral model in $(2+1)$ dimensions,'' J. Math. Phys. **37**, 3422–3441 (1996).
[24] M. J. Ablowitz and J. Villarroel, ''Solutions to the time dependent Schrödinger and the Kadomtsev–Petviashvili equations,'' Phys. Rev. Lett. **78**, 570–573 (1997).
[25] H. Airault, H. P. McKean, and J. K. Moser, ''Rational and elliptic solutions of the Korteweg–de Vries equation and a related many-body problem,'' Commun. Pure Appl. Math. **30**, 95–148 (1977).
[26] R. Hirota, Y. Ohta, and J. Satsuma, ''Wronskian structures of solutions for soliton equations,'' Prog. Theor. Phys. Suppl. **94**, 59–72 (1988).
[27] V. B. Matveev, ''Some comments on the rational solutions of the Zakharov–Shabat equations,'' Lett. Math. Phys. **3**, 503–512 (1979).
[28] D. E. Littlewood, *The Theory of Group Characters* (Clarendon, Oxford, 1950).
[29] M. Adler and J. Moser, ''On a class of polynomials connected with the Korteweg–de Vries equation,'' Commun. Math. Phys. **61**, 1–30 (1978).
[30] J. J. C. Nimmo and N. C. Freeman, ''Rational solutions of the Korteweg–de Vries equation in Wronskian form,'' Phys. Lett. A **96**, 443–445 (1983).
[31] M. J. Ablowitz and J. Satsuma, ''Solitons and rational solutions of nonlinear evolution equations,'' J. Math. Phys. **19**, 2180–2186 (1978).
[32] D. E. Pelinovsky and R. H. J. Grimshaw, ''Structural transformation of eigenvalues for a perturbed algebraic soliton potential,'' Phys. Lett. A **229**, 165–172 (1997).
[33] N. N. Akhmediev, V. M. Eleonsky, and N. E. Kulagin, ''Generation of periodic trains of picosecond pulses in an optical fiber: Exact solutions,'' Sov. Phys. JETP **62**, 894–899 (1985).
[34] M. J. Ablowitz and B. M. Herbst, ''On homoclinic structure and numerically induced chaos for the nonlinear Schrödinger equation,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **50**, 339–351 (1990).
[35] R. Hirota, ''Classical Boussinesq equation is a reduction of the modified KP equation,'' J. Phys. Soc. Jpn. **54**, 2409–2415 (1985).
[36] D. Pelinovsky, J. Springael, F. Lambert, and I. Loris, ''On modified NLS, Kaup, and NLBq equations: Differential transformations and bilinearization,'' J. Phys. A **30**, 8705–8717 (1997).

On local inverse transforms

Pierre C. Sabatier[a)]
Laboratoire de Physique Mathématique et Théorique-URA CNRS 768
Université Montpellier II, 34095 Montpellier cedex 05, France

(Received 4 May 1998; accepted for publication 9 May 1998)

The direct linearization of a nonlinear partial differential equation, when it is possible, is a purely algebraic process, which should not imply strong conditions on the solutions at infinity. Thus, it should be possible to manage it in such a way that the relevant inversion equation is not of the Marchenko form, but others, e.g., the true Gelfand–Levitan form, which preserves the local character of the approach. This is demonstrated on the Korteweg–de Vries case. © *1998 American Institute of Physics.* [S0022-2488(98)02610-3]

I. INTRODUCTION

Processing the direct linearization of a nonlinear partial differential equation (E) is giving a linear integral equation, called ''Spectral'' because ''spectral variables'' only are involved, the space variables being fixed, and whose solutions are those of a ''spectral'' problem depending on solutions of (E) as parameters.[1,2] This process is possible when special algebraic conditions are fulfilled, and so, it is a purely algebraic and ''local'' process,[2] that should not need or imply boundary conditions on the solutions of E. Yet, common experience shows that the so-called ''inverse scattering transform'' works only for a special class of solutions of E, usually with a vanishing condition at ∞. This is satisfactory for people of scattering theory but, in light of a direct linearization, it is very surprising. We study here the most simple process of direct linearization, that of Korteweg–de Vries equation. In classical and very well-known results,[3] this process is handled in such a way that it reduces the research of solutions to what we call[4] a Faddeev–Marchenko inversion equation, and the necessary vanishing of the solutions of (E) at infinity is related to the structure of this equation. In the present paper, we show that the process can be handled in other ways, which rather correspond to a Gelfand–Levitan equation, or other inversion equations with the same ''local'' significance. The reader understands that we use the words ''Gelfand–Levitan equations'' for inversion equations whose kernel is sitting on a finite support, ''Marchenko equations'' if the support is semi-infinite, as they are called in books on inverse problems[4,5] and not in some papers on inverse methods!

In Sec. II, a very general spectral integral equation is given, and shown to yield solutions, in the general case, of a matricial 2×2 Korteweg–de Vries equation. This equation has been obtained, and its properties are derived, as simple applications of the linearization method we previously described[6] for matricial cases. However, in this scalar case it is easier to handle it with $(k+\lambda)$ in the kernel denominator, the kernel numerator again being constructed from the Korteweg–de Vries dispersion relations. The way we handle ''parasite'' and ''free'' terms follows the lines of our previous method, so that, here, we do not give details on why we do so: we only state and prove the results.

In Sec. III, we first study the most general matricial case when the spectral measure is supported by one point only on the imaginary axis, and we show that a geometric transformation reduces here the matricial case to the scalar Korteweg–de Vries equation. The solutions we obtain generalize, in some way, solitons. We study then the case where the ''spectral measure'' has its support on the real axis, where it is even, and we show that again the matricial Korteweg–de Vries equation diagonalizes and that the inversion equation in this case is nothing but the familiar Gelfand–Levitan equation of the radial Schrodinger problem!

The main interest of this paper in our opinion is the following.

a)Electronic mail: sabatier@LPM.univ-montp2.fr

 © 1998 American Institute of Physics

(1) A local approach may be a key to solutions fulfilling prescribed conditions at a finite distance and more generally to new classes of solutions, both for Korteweg–de Vries and for related nonlinear equations (e.g., Painlevé ones).
(2) The method seems open to many generalizations, both by using algebraic direct linearizations for other nonlinear partial differential equations and by looking for other inversion equations.

II. THE SPECTRAL INTEGRAL EQUATION AND KORTEWEG–DE VRIES LINEARIZATION

We start from the spectral integral equation,

$$f_1(k,x,t)=\sin[kx+k^3t]+\lim_{\epsilon\to 0+}\int \frac{d\mu(\lambda)}{k+\lambda+i\epsilon}\sin[(k+\lambda)x+(k^3+\lambda^3)t+\nu(\lambda)]f_1(\lambda,x,t), \tag{2.1}$$

which will be written for short as

$$f_1(k,x,t)=\sin[kx+k^3t]+\oint T(k,\lambda,x,t)f_1(\lambda,x,t), \tag{2.2}$$

and gathered with its ''sister'' equation,

$$f_2(k,x,t)=\cos[kx+k^3t]+\oint T(k,\lambda,x,t)f_2(\lambda,x,t). \tag{2.3}$$

We assume throughout that the support of $d\mu(\lambda)$ is contained in $\mathrm{Im}\,\lambda\geqslant 0$. ϵ and the limit process are used if part of the support is on the real axis. We assume also that the properties of this measure are such that (2.2) and (2.3) are Fredholm equations, for which the Fredholm alternative holds. In the generic cases, and/or at generic values of x and t, we use the working assumption that the homogeneous equation,

$$f(k,x,t)=\oint T(k,\lambda,x,t)f(\lambda,x,t), \tag{2.4}$$

has no solution but $f=0$. Let us now introduce the following functions, which we call ''the prepotentials:''

$$u_{11}(x,t)=\int d\mu(\lambda)\sin[\lambda x+\lambda^3t+\nu(\lambda)]f_1(\lambda,x,t), \tag{2.5}$$

$$u_{12}(x,t)=\int d\mu(\lambda)\cos[\lambda x+\lambda^3t+\nu(\lambda)]f_1(\lambda,x,t), \tag{2.6}$$

$$u_{21}(x,t)=\int d\mu(\lambda)\sin[\lambda x+\lambda^3t+\nu(\lambda)]f_2(\lambda,x,t), \tag{2.7}$$

$$u_{22}(x,t)=\int d\mu(\lambda)\cos[\lambda x+\lambda^3t+\nu(\lambda)]f_2(\lambda,x,t). \tag{2.8}$$

It is a matter of straightforward calculations, by applying the differential operator to both sides of (2.2) or (2.3), and reducing parasite terms, to show the following equations:

$$\frac{\partial}{\partial x}f_1(k,x,t)=\cos(kx+k^3t)[k+u_{12}(x,t)]-\sin(kx+k^3t)u_{11}(x,t)$$
$$+\oint T(k,\lambda,x,t)\frac{\partial f_1}{\partial x}(\lambda,x,t), \tag{2.9}$$

$$\frac{\partial}{\partial x} f_2(k,x,t) = -\sin(kx+k^3t)[k+u_{21}(x,t)] + \cos(kx+k^3t)u_{22}(x,t) + \oint T(k,\lambda,x,t) \frac{\partial}{\partial x} f_2(\lambda,x,t), \tag{2.10}$$

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right) f_1(k,x,t) = 2\cos(kx+k^3t) \frac{\partial}{\partial x} u_{12}(x,t) - 2\sin(kx+k^3t) \frac{\partial}{\partial x} u_{11}(x,t) + \oint T(k,\lambda,x,t)\left(\frac{\partial^2}{\partial x^2}+\lambda^2\right) f_1(\lambda,x,t), \tag{2.11}$$

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right) f_2(k,x,t) = 2\cos(kx+k^3t) \frac{\partial}{\partial x} u_{22}(x,t) - 2\sin(kx+k^3t) \frac{\partial}{\partial x} u_{21}(x,t) + \oint T(k,\lambda,x,t)\left(\frac{\partial^2}{\partial x^2}+\lambda^2\right) f_2(\lambda,x,t). \tag{2.12}$$

A word of explanation for a reader not familiar with the author's words;[2] ''*parasite*'' terms are terms which appear when the operator applies inside the integral *and* which are not *similar* to the left-hand side of the equation [contrary to the one in (2.12), for instance]. Parasite terms can be reduced if algebraic calculations can kill for them the denominator $k+\lambda$, so that, thanks to the separability of $\sin[(k+\lambda)x+(k^3+\lambda^3)t+\nu(\lambda)]$, the former parasite term becomes a combination of ''free'' terms. In the four equations above [and in (2.2) and (2.3)], you can see only the left-hand side, free terms, and an integral term similar to the left-hand side of the equation—as a matter of fact, you see in the six equations the same equation with different free terms. Again, you can obtain the same equation by using a function of x and t only to multiply both sides (a ''*compatible*'' operation). Thus, it is easy to combine (2.11) or (2.12) and convenient linear combinations of f_1 and f_2, as given by (2.2) and (2.3), to obtain an expression that satisfies the homogeneous equation (2.4) and therefore vanishes, the key trick in direct linearization.

The result is the Schrodinger matricial problem:

$$\left(\frac{\partial^2}{\partial x^2}+k^2-\mathbf{V}(x,t)\right)\mathbf{f}(k,x,t)=0, \tag{2.13}$$

where

$$\mathbf{f}(k,x,t)=\begin{pmatrix} f_1(k,x,t) \\ f_2(k,x,t) \end{pmatrix} \tag{2.14}$$

$$\mathbf{V}(x,t)=\begin{pmatrix} -2\dfrac{\partial}{\partial x}u_{11} & 2\dfrac{\partial}{\partial x}u_{12} \\ -2\dfrac{\partial}{\partial x}u_{21} & 2\dfrac{\partial}{\partial x}u_{22} \end{pmatrix}. \tag{2.15}$$

Now, we wish to obtain the time operator of the Lax pair.

Again, it is a matter of lengthy but straightforward calculations to show first, from (2.5)–(2.8) and (2.13), and using ''primes'' to denote x derivatives:

$$\frac{\partial^2 u_{11}}{\partial x^2} = 2\frac{\partial}{\partial x}\int d\mu(\lambda)\cos[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_1(\lambda,x,t) - 2u'_{11}u_{11}+2u'_{12}u_{21}, \tag{2.16}$$

$$\frac{\partial^2 u_{12}}{\partial x^2} = -2\frac{\partial}{\partial x}\int d\mu(\lambda)\sin[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_1(\lambda,x,t) - 2u'_{11}u'_{12}+2u'_{12}u_{22}, \tag{2.17}$$

$$\frac{\partial^2 u_{21}}{\partial x^2}=2\frac{\partial}{\partial x}\int d\mu(\lambda)\cos[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_2(\lambda,x,t)-2u'_{21}u_{11}+2u'_{22}u_{21}, \tag{2.18}$$

$$\frac{\partial^2 u_{22}}{\partial x^2}=-2\frac{\partial}{\partial x}\int d\mu(\lambda)\sin[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_2(\lambda,x,t)-2u'_{21}u_{12}+2u'_{22}u_{22}. \tag{2.19}$$

Then we apply $(\partial/\partial t-k^2(\partial/\partial x))$ to both sides of (2.2), and do the standard work on parasite and free terms, using (2.16) and (2.17), to obtain

$$\begin{aligned}\left(\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}\right)f_1(k,x,t)&=\cos(kx+k^3t)\left[-ku'_{11}+\left(u'_{12}u_{22}-u'_{11}u_{12}-\frac{1}{2}u''_{12}\right)\right]\\&+\sin(kx+k^3t)\left[-ku'_{12}+\left(u'_{11}u_{11}-u'_{12}u_{21}+\frac{1}{2}u''_1\right)\right]\\&+\oint T(k,\lambda,x,t)\left(\frac{\partial}{\partial t}-\lambda^2\frac{\partial}{\partial x}\right)f_1(\lambda,x,t),\end{aligned} \tag{2.20}$$

and, with similar derivations,

$$\begin{aligned}\left(\frac{\partial}{\partial t}-k^2\frac{\partial}{\partial x}\right)f_2(k,x,t)&=\cos(kx+k^3t)\left[-ku'_{21}+\left(u'_{22}u_{22}-u'_{21}u_{12}-\frac{1}{2}u''_{22}\right)\right]\\&+\sin(kx+k^3t)\left[-ku'_{22}+\left(u'_{21}u_{11}-u'_{22}u_{21}+\frac{1}{2}u''_{21}\right)\right]\\&+\oint T(k,x,t)\left(\frac{\partial}{\partial t}-\lambda^2\frac{\partial}{\partial x}\right)f_2(\lambda,x,t),\end{aligned} \tag{2.21}$$

whereas it follows from (2.9) (2.10), (2.2), and (2.3), that

$$\frac{\partial f_1}{\partial x}+u_{11}f_1-u_{12}f_2=k\cos+\oint T(k,\lambda,x,t)\left[\frac{\partial f_1}{\partial x}+u_{11}f_1-u_{12}f_2\right], \tag{2.22}$$

$$\frac{\partial f_2}{\partial x}+u_{21}f_1-u_{22}f_2=-k\sin+\oint T(k,\lambda,x,t)\left[\frac{\partial f_1}{\partial x}+u_{21}f_1-u_{22}f_2\right], \tag{2.23}$$

where we used condensed notations to make it easier to grasp. With the same care, we write down the result obtained after combining (2.20)–(2.23) into the homogeneous vectorial equation:

$$\frac{\partial \mathbf{f}}{\partial t}-k^2\frac{\partial \mathbf{f}}{\partial x}-\frac{1}{2}\mathbf{V}\frac{\partial \mathbf{f}}{\partial x}+\frac{1}{4}\mathbf{V}'\mathbf{f}=0. \tag{2.24}$$

The asymptotic behavior of f as $k\to\pm\infty$ on the real axis can be derived from (2.2), (2.3), and the values of prepotentials and their derivatives. Setting

$$\mathbf{f}^0=\begin{pmatrix}\sin(kx+k^3t)\\ \cos(kx+k^3t)\end{pmatrix}, \tag{2.25}$$

we obtain

$$\mathbf{f}=\mathbf{f}^{(0)}+k^{-1}\mathbf{f}^{(1)}+k^{-2}\mathbf{f}^{(2)}+0(k^{-3})\quad(k\to\pm\infty), \tag{2.26}$$

where

$$\mathbf{f}^{(1)}=\mathbf{P}\mathbf{f}^{(0)}, \tag{2.27}$$

$$\mathbf{P}=\begin{bmatrix} u_{12}(x,t) & u_{11}(x,t) \\ u_{22}(x,t) & u_{21}(x,t) \end{bmatrix}, \tag{2.28}$$

$$\mathbf{f}^{(2)}=-\mathbf{Q}\mathbf{f}^{(0)}. \tag{2.29}$$

$$\mathbf{Q}=\begin{bmatrix} v_{12}(x,t) & v_{11}(x,t) \\ v_{22}(x,t) & v_{21}(x,t) \end{bmatrix} \tag{2.30}$$

$$v_{11}(x,t)=\int d\mu(\lambda)\sin[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_1(\lambda,x,t), \tag{2.31}$$

$$v_{12}(x,t)=\int d\mu(\lambda)\cos[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_1(\lambda,x,t), \tag{2.32}$$

$$v_{21}(x,t)=\int d\mu(\lambda)\sin[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_2(\lambda,x,t), \tag{2.33}$$

$$v_{22}(x,t)=\int d\mu(\lambda)\cos[\lambda x+\lambda^3 t+\nu(\lambda)]\lambda f_2(\lambda,x,t). \tag{2.34}$$

It follows from (2.16)–(2.19) that

$$v'_{11}=u'_{12}u_{22}-u'_{11}u_{12}-\tfrac{1}{2}u''_{12}, \tag{2.35}$$

$$v'_{12}=u'_{11}u_{11}-u'_{12}u_{21}+\tfrac{1}{2}u''_{11}, \tag{2.36}$$

$$v'_{21}=u'_{22}u_{22}-u'_{21}u_{12}-\tfrac{1}{2}u''_{22}, \tag{2.37}$$

$$v'_{22}=u'_{21}u_{11}-u'_{22}u_{21}+\tfrac{1}{2}u''_{21}. \tag{2.38}$$

Now, let us write down the commutator of the linear operators acting on **f** in the spectral problem (2.13) and in the auxiliary problem (2.24), and seek the asymptotic behavior of the result. After two terms identically vanishing at positive orders of k, we get at order 0 in k the Korteweg–de Vries matricial equation for **V**:

$$\frac{\partial \mathbf{V}}{\partial t}+\frac{1}{4}\mathbf{V}'''-\frac{3}{4}(\mathbf{V}\mathbf{V}'+\mathbf{V}'\mathbf{V})=0. \tag{2.39}$$

III. THE SCALAR KORTEWEG–DE VRIES EQUATION

There are several cases where **V** can be diagonalized, so that (2.39) yields two families of solutions for the familiar Korteweg–de Vries equation:

$$\frac{\partial V}{\partial t}+\frac{1}{4}\frac{\partial^3 V}{\partial x^3}-\frac{3}{2}V\frac{\partial V}{\partial x}=0, \tag{3.1}$$

where V is a (scalar) function of x and t. We pinpoint two examples, both showing a degree of generality such that we cannot call them special cases.

A. First example—solitons

We choose the spectral measure $d\mu(\lambda)$ in the starting equations as

$$d\mu(\lambda)=\mu_0\delta(\lambda-i\kappa), \quad \kappa>0. \tag{3.2}$$

Setting

$$\cos\ \nu(i\kappa)=\alpha(\kappa); \quad \sin\ \nu(i\kappa)=i\beta(\kappa), \tag{3.3}$$

We obtain

$$f_1(i\kappa,x,t)=\frac{i\ \sinh(\kappa x-\kappa^3 t)}{1+S}, \tag{3.4}$$

$$f_2(i\kappa,x,t)=\frac{\cosh(\kappa x-\kappa^3 t)}{1+S}, \tag{3.5}$$

where

$$S=-\frac{\mu_0}{2\kappa}\left[\alpha(\kappa)\sinh X+\beta(\kappa)\cosh X\right], \tag{3.6}$$

$$X=2(\kappa x-\kappa^3 t). \tag{3.7}$$

Notice that if μ_0 is real and $\nu(\lambda)$ is an odd function, f_2 and if_1 are real. From (2.15) we derive the matrix elements of $\mathbf{V}$:

$$V_{11}=\mu_0(1+S)^{-2}[\mu_0(\alpha^2-\beta^2)+(2\kappa\alpha-\beta\alpha\mu_0)\sinh X+(2\kappa\beta-\mu_0\alpha^2)\cosh X], \tag{3.8}$$

$$V_{12}=i\mu_0(1+S)^{-2}(2\kappa-\beta\mu_0)(\alpha\ \cosh X+\beta\ \sinh X), \tag{3.9}$$

$$V_{12}=-i\mu_0(1+S)^{-2}(2\kappa+\beta\mu_0)(\alpha\ \cosh X+\beta\ \sinh X), \tag{3.10}$$

$$V_{22}=\mu_0(1+S)^{-2}[\mu_0(\alpha^2-\beta^2)+(2\kappa\alpha+\beta\alpha\mu_0)\sinh X+(2\kappa\beta+\mu_0\alpha^2)\cosh X]. \tag{3.11}$$

It is easy to see that

$$V_{22}-V_{11}=2\mu_0^2\alpha(1+S)^{-2}(\alpha\ \cosh X+\beta\ \sinh X), \tag{3.12}$$

and setting

$$\gamma=\kappa/\mu_0\alpha;\quad \delta=\beta/2\alpha, \tag{3.13}$$

we get

$$V_{12}=i(\gamma-\delta)(V_{22}-V_{11});\quad V_{21}=-i(\gamma+\delta)(V_{22}-V_{11}). \tag{3.14}$$

It follows that V can be diagonalized on a basis of eigenvectors that do not depend on x and t. If U is the matrix made of them, inserting $\mathbf{U}^*\mathbf{\Lambda}\mathbf{U}$ for $\mathbf{V}$ in the matricial Korteweg–de Vries equation readily yields the equation (3.1) for either of the two diagonal components of $\mathbf{\Lambda}$:

$$\lambda^{\pm}=\tfrac{1}{2}[(V_{22}+V_{11})\pm(V_{22}-V_{11})\Gamma], \tag{3.15}$$

where

$$\Gamma=[1+4(\gamma^2-\delta^2)]^{1/2}. \tag{3.16}$$

For a proof, calculate $\lambda^{\pm}$ as zeros of $\det(\mathbf{V}-\lambda\mathbf{I})$, then check that $\mathbf{V}-\lambda^{\pm}\mathbf{I}$ is proportional to $V_{22}-V_{11}$, so that, in the equation for eigenvectors, this factor, which is the only term depending on x and t, can be dropped.

From (3.8), (3.11), (3.12), and (3.13), we obtain the more attractive form

$$\lambda^{\pm}=4\kappa^2\ \frac{1-4\delta^2+2\gamma(\sinh X+2\delta\ \cosh X)\pm\Gamma(\cosh X+2\delta\sinh X)}{(\sinh X+2\delta\ \cosh X-2\gamma)^2}. \tag{3.17}$$

These families of solutions are, of course, solitary waves depending only on X. They are generally singular at one value of X.

They are real for μ_0 real and $\nu(\lambda)$ odd. Since they are functions of X only, they should be solutions, as a function of X, of

$$\frac{d}{dX}[4\kappa^2(\Lambda''-\Lambda)-3\Lambda^2]=0. \tag{3.18}$$

As a matter of fact, it is not difficult to check that they are solutions of

$$2\kappa^2(\Lambda'^2-\Lambda^2)-\Lambda^3=0, \tag{3.19}$$

which is more restrictive than (3.18).

B. Second example: Gelfand–Levitan inversion

We choose the spectral measure on the real axis [this is why $i\epsilon$ was necessary in (2.1)], and ''symmetric,'' meaning that if $d\mu(\lambda)$ is a Stieltjes measure $m(\lambda)d\lambda$, $m(\lambda)$ is even.

We also assume that $\nu(\lambda)$ is an odd function of λ. Now suppose a solution f of the homogeneous equation (2.4) be written as a sum of its even part and its odd part as a function of k:

$$f(k,\cdot)=f^+(k,\cdot)+f^-(k,\cdot). \tag{3.20}$$

Substituting (3.20) into (3.4) yields for f^+ and f^- the equations

$$f^+(k,x,t)=\int d\mu(\lambda)T^+(k,\lambda,x,t)f^+(\lambda,x,t), \tag{3.21}$$

$$f^-(k,x,t)=\int d\mu(\lambda)T^-(k,\lambda,x,t)f^-(\lambda,x,t), \tag{3.22}$$

where

$$2T^+(k,\lambda,x,t)=\frac{\sin[(k+\lambda)x+(k^3+\lambda^3)t+\nu(\lambda)]}{k+\lambda}+\frac{\sin[(k-\lambda)x+(k^3-\lambda^3)t-\nu(\lambda)]}{k-\lambda}, \tag{3.23}$$

$$2T^-(k,\lambda,x,t)=\frac{\sin[(k+\lambda)x+(k^3+\lambda^3)t+\nu(\lambda)]}{k+\lambda}-\frac{\sin[(k-\lambda)x+(k^3-\lambda^3)t-\nu(\lambda)]}{k-\lambda}. \tag{3.24}$$

It follows that our assumption on (2.4) (no solution but 0) implies the same for (3.21) and (3.22). Now, let us study the problem at $t=0$. Our remarks imply that $f_1(k,x)$ is an odd function of k, and the unique solution of

$$f_1(k,x)=\sin kx+\int d\mu(\lambda)T^-(k,\lambda,x,0)f_1(\lambda,x). \tag{3.25}$$

From now on we assume to be simple that $\nu(\lambda)=0$. Hence T^- is equal to a simple integral, which for $x>0$ reads

$$T^-(k,\lambda,x,0)=-\int_0^x \sin ky\ \sin \lambda y\ dy. \tag{3.26}$$

Inserting it in (3.25), we get, after defining the kernels,

$$L_1(x,y)=-\int d\mu(\lambda)f_1(\lambda,x)\sin \lambda y, \tag{3.27}$$

$$G_1(x,y)=\int d\mu(\lambda)\sin \lambda x\ \sin \lambda y, \tag{3.28}$$

the equations

$$f_1(k,x)=\sin\, kx+\int_0^x L_1(x,y)\sin\, ky\; dy, \tag{3.29}$$

$$L_1(x,z)+G_1(x,z)+\int_0^z L_1(x,y)G_1(y,z)dy=0. \tag{3.30}$$

On the other hand, a glance at Eqs. (2.5)–(2.8) shows that because of the parity properties, only u_{11} and u_{22} do not vanish, so that the potential $\mathbf{V}$, given by (2.15), is diagonal for any x and t. Comparing (2.15) and (3.27) shows that

$$V_{11}=2\frac{d}{dx}L_1(x,x). \tag{3.31}$$

By the same token, we successively obtain

$$f_2(k,x)=\cos\, kx+\int d\mu(\lambda)\left[\int_0^z \cos\, ky\, \cos\, \lambda y\; dy\right] f_2(\lambda,x), \tag{3.32}$$

$$L_2(x,y)=\int d\mu(\lambda) f_2(\lambda,x)\cos\, \lambda y, \tag{3.33}$$

and, setting

$$G_2(x,z)=-\int d\mu(\lambda)\cos\, \lambda x\, \cos\, \lambda z, \tag{3.34}$$

$$f_2(k,x)=\cos\, kx+\int_0^z \cos\, kyL_2(x,y)dy, \tag{3.35}$$

$$L_2(x,z)+G_2(x,z)+\int_0^z L_2(x,y)G_2(y,z)dy=0, \tag{3.36}$$

$$V_{22}=2\frac{d}{dx}L_2(x,x). \tag{3.37}$$

Hence, in this case, the Gelfand–Levitan equations are reproduced with their usual relations to potentials. If t is different from zero, the parity properties remain, and f_1 is still an odd function of k, the unique solution of

$$f_1(k,x,t)=\sin(kx+k^3t)+\int d\mu(\lambda)T^-(k,\lambda,x,t)f_1(\lambda,x,t), \tag{3.38}$$

and f_2 still an even function of k, and the unique solution of

$$f_2(k,x,t)=\cos(kx+k^3t)+\int d\mu(\lambda)T^+(k,\lambda,x,t)f_2(\lambda,x,t). \tag{3.39}$$

Because of parity properties, $\mathbf{V}$ remains diagonal. Hence V_{11} and V_{22} define two families of solutions of the Korteweg–de Vries equation (3.1). So as to describe them, we can solve (3.38) and (3.39). It is also possible to construct a system of equations of the Gelfand–Levitan form with kernels which are integrated from 0 to x or 0 to t. This is done in the Appendix. However, we think that for solving the Korteweg–de Vries equation the best management is done by directly solving (2.2) and (2.3).

Remarks:

(1) Following the lines above, it is not difficult to add Dirac measures on the imaginary axis to the Gelfand–Levitan case.

(2) Setting in (2.2) and (2.3),

$$\nu(\lambda)=-ip;\quad d\mu(\lambda)=2e^{-p}\,d\rho(\lambda), \tag{3.40}$$

then letting $p\to+\infty$, and gathering the two equations, we obtain for $f=f_2+if_1$ the equation

$$f(k,x,t)=e^{i(kx+k^3t)}-i\lim_{\epsilon\to 0^+}\int\frac{d\rho(\lambda)}{k+\lambda+i\epsilon}\,e^{i[(k+\lambda)x+(k^3+\lambda^3)t]}f(\lambda,x,t), \tag{3.41}$$

which is commonly used[3] to linearize the Korteweg–de Vries equation, and can easily be used to derive the Faddeev–Marchenko equations as we used (3.25) to derive the Gelfand–Levitan equation. Setting (3.40), and

$$d\rho(\lambda)=\mu\,\delta(\lambda-i\kappa),\quad e^{-X_0}=2\kappa/\mu, \tag{3.42}$$

can also be done directly in (3.17). The limit of $\lambda^{\pm}$ as p goes to $+\infty$ is then 0 for λ^+ and

$$\lambda^-=-2\kappa^2/\cosh^2[\tfrac{1}{2}(X-X_0)], \tag{3.43}$$

which is the usual soliton.

(3) When spectral equations like (2.1) are used, a problem sometimes arises: is it possible to construct $d\mu(\lambda)$ from the potential values at a given t? This problem is solved if a Gelfand–Levitan or a Marchenko equation exists, since $d\mu(\lambda)$ can be identified with the corresponding ''measures'' (respectively, spectral measure or reflection coefficient). In other cases, one should study separately this previously ''direct'' problem, now an inverse one.

APPENDIX: INVERSION EQUATIONS (GELFAND–LEVITAN EXAMPLE) FOR $t\neq 0$

With the symmetry assumptions of the Gelfand–Levitan example, and $t\neq 0$, Eq. (3.25) is replaced by (3.38) [with $\nu(\lambda)$ set to zero in $T^-(k,\lambda,x,t)$].

We easily derive, from Eq. (3.24),

$$T^-(k,\lambda,x,t)=\int_0^x\sin(ky+k^3t)\sin(\lambda y+\lambda^3t)dy$$
$$-(k^2+\lambda^2)\int_0^t\sin k^3s\,\sin\lambda^3s\,ds-k\lambda\int_0^t\cos k^3s\,\cos\lambda^3s\,ds. \tag{A1}$$

Now, setting

$$L(x,y,t,u)=-\int d\mu(\lambda)f_1(\lambda,x,t)\sin(\lambda y+\lambda^3u), \tag{A2}$$

we can derive from (A1) the representation:

$$f_1(k,x,t)=\sin(kx+k^3t)+\int_0^x\sin(kz+k^3t)L(x,z,t,t)dz$$
$$+k^2\int_0^t\sin k^3sL(x,0,t,s)ds-\int_0^t\sin k^3s\left[\frac{\partial^2}{\partial y^2}L(x,y,t,s)\right]_{y=0}ds$$
$$+k\int_0^t\cos k^3s\left[\frac{\partial}{\partial y}L(x,y,t,s)\right]_{y=0}ds. \tag{A3}$$

Setting then

$$G(x,y,t,u)=\int d\mu(\lambda)\sin(\lambda x+\lambda^3 t)\sin(\lambda y+\lambda^3 u), \tag{A4}$$

we multiply both sides of (A3) by $\sin(ky+k^3u)d\mu(k)$ and integrate, obtaining

$$\begin{aligned} &L(x,y,t,u)+G(x,y,t,u)+\int_0^x L(x,z,t,t)G(y,z,t,u)dz \\ &-\int_0^t\left[\frac{\partial^2}{\partial x^2}G(x,y,s,u)\right]_{x=0}L(x,0,t,s)ds \\ &-\int_0^t G(0,y,s,u)\left[\frac{\partial^2}{\partial y^2}L(x,y,t,u)\right]_{y=0}ds \\ &+\int_0^t\left[\frac{\partial}{\partial x}G(x,y,s,u)\right]_{x=0}\left[\frac{\partial}{\partial y}L(x,y,t,s)\right]_{y=0}ds=0 \end{aligned} \tag{A5}$$

whereas Eq. (2.15) yields the expected relation between $V(x,t)$ and $L(x,x,t,t)$. The equation (A5) is readily transformed into a linear integral Fredholm equation for the vectorial function,

$$\mathbf{L}(\mathbf{x},\mathbf{y},\mathbf{t},\mathbf{u})=\begin{pmatrix} L(x,y,t,u) \\ L(x,0,t,u) \\ \left[\frac{\partial}{\partial y}L(x,y,t,u)\right]_{y=0} \\ \left[\frac{\partial^2}{\partial y^2}L(x,y,t,u)\right]_{y=0} \end{pmatrix}, \tag{A6}$$

G being known from $d\mu(\lambda)$, as well as its derivatives.

Thus, the generalization at $t\neq 0$ of the Gelfand–Levitan formulas is more complicated than a direct use of (3.38). Similar results, with similar remarks, apply in the even case.

ACKNOWLEDGMENTS

It is a pleasure for me to acknowledge the information that I received on previous studies of Korteweg–de Vries from Marco Boiti and Boris Konopelcenko, and the very helpful suggestions from the referee.

[1] P. C. Sabatier, ''Spectral transform for nonlinear evolution equations with N space-dimensions,'' Phys. Lett. A **161**, 345–351 (1992); M. Boiti, F. Pempinelli, and P. C. Sabatier, ''First and second order nonlinear partial differential evolution equations from an inverse problem,'' Inverse Probl. **9**, 1–37 (1993).

[2] P. C. Sabatier, Inverse problems related to ''Integrable Nonlinear Partial Differential Equations,'' pp. 193–202, in *Inverse Problems in Mathematical Physics*, Lecture Notes in Physics, edited by L. Paivarinta and E. Somersalo (Springer-Verlag, Berlin, 1993), Vol. 422; P. C. Sabatier, ''Constructing nonlinear integrable equations by direct linearization,'' Teor. Mat. Fiz. **99**, 517–527 (1994).

[3] A. S. Fokas and M. J. Ablowitz, ''Linearization of the Korteweg de Vries and Painlevé II equations,'' Phys. Rev. Lett. **47**, 1096–1100 (1981); R. S. Rosales, ''Exact solutions of some nonlinear evolution equations,'' Stud. Appl. Math. **59**, 117–151 (1978); and see other references in *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, by M. J. Ablowitz and P. A. Clarkson (Cambridge University Press, Cambridge, 1991).

[4] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 2nd ed. revised and expanded (Springer-Verlag, Berlin, 1989).

[5] D. N. Ghosh Roy, *Methods of Inverse Problems in Physics* (CRC Press, Boca Raton, 1991); G. M. L. Gladwell, *Inverse Problems in Scattering* (Kluwer, Dordrecht, 1993).

[6] P. C. Sabatier, ''On algebra of direct linearization,'' pp. 208–215 in *Nonlinear Evolution Equations and Dynamical Systems*, edited by V. Makhankov, A. Bishop, and D. Holm (World Scientific, Singapore, 1995).

The Lanczos potential for vacuum space–times with an Ernst potential

P. Dolan[a] and B. D. Muratori[b]
Department of Mathematics, Huxley Building, Imperial College, 180 Queen's Gate, London SW7 2BZ, England

(Received 24 March 1998; accepted for publication 27 May 1998)

By a particular choice of tetrad and coordinate systems, we find the Lanczos potential for general vacuum stationary axisymmetric space–times and for general vacuum cylindrically symmetric solutions. This result is particularly elegant and unifies the treatment of the Lanczos potential and the Ernst potential for vacuum stationary axially symmetric and cylindrically symmetric space–times. The Lanczos potential for such large classes of vacuum space–times is expressed very simply in terms of spin-coefficients for a null tetrad. © *1998 American Institute of Physics.* [S0022-2488(98)00910-4]

I. INTRODUCTION

Let a space–time M endowed with a metric of signature $(+,-,-,-)$ with the usual Levi-Civita connection for covariant differentiation have the Riemann tensor defined as

$$v_{c;ab}-v_{c;ba}=R_{abc}{}^{h}v_{h}, \tag{1}$$

where $a,b,c,h=0,1,2,3$ and we can decompose the Riemann tensor into its irreducible parts as follows:

$$R_{abcd}=G_{abcd}+E_{abcd}+C_{abcd}, \tag{2}$$

where

$$G_{abcd}=\tfrac{1}{12}R(g_{ac}g_{bd}-g_{ad}g_{bc}), \tag{3}$$

$$E_{abcd}=\tfrac{1}{2}(g_{ac}S_{bd}+g_{bd}S_{ac}-g_{ad}S_{bc}-g_{bc}S_{ad}), \tag{4}$$

$$S_{ab}=R_{ab}-\tfrac{1}{4}Rg_{ab}, \tag{5}$$

$$R_{ab}=R_{ahb}{}^{h}, \quad R=R^{e}{}_{e}. \tag{6}$$

The tensor C_{abcd} is the Weyl conformal curvature tensor: It satisfies the same algebraic symmetries as the Riemann tensor and is traceless in all pairs of indices. Because G_{abcd} and E_{abcd} can both be shown to be derivable from simpler irreducible tensors with fewer indices, namely S_{ab} and R, Lanczos[1] thought he could make the Weyl tensor also derivable from a simpler tensor field. This tensor field is now known as the *Lanczos potential* and it satisfies the following index symmetries:

$$L_{abc}=-L_{bac}, \tag{7}$$

together with the algebraic conditions

$$L_{a}{}^{e}{}_{e}=0, \quad L_{abc}+L_{bca}+L_{cab}=0, \tag{8}$$

[a]Electronic mail: pdolan@ma.ic.ac.uk
[b]Electronic mail: b.muratori@ma.ic.ac.uk

 © 1998 American Institute of Physics

where the last two conditions (8) can be seen to be duals of each other on the first two indices ($L_a{}^e{}_e = {}^*L_a{}^e{}_e = 0$). We have thus reduced L_{abc} to at most 16 independent components but to have a perfect match with the Weyl tensor, Lanczos further suggested the imposition of the six equations

$$L_{ab}{}^e{}_{;e} = 0 \tag{9}$$

known as the *Lanczos differential gauge conditions* so as to have only ten effective degrees of freedom. For simplicity we follow the conventions and methods of Dolan and Kim[2,3] rather than Lanczos.[1]

The potential-field relations of the Lanczos potential to the Weyl tensor are

$$\begin{aligned} C_{abcd} = {} & L_{abc;d} - L_{abd;c} + L_{cda;b} - L_{cdb;a} \\ & + L_{(ad)}g_{bc} + L_{(bc)}g_{ad} - L_{(ac)}g_{bd} - L_{(bd)}g_{ac} + \tfrac{2}{3}L^{ef}{}_{e;f}(g_{ac}g_{bd} - g_{ad}g_{bc}), \end{aligned} \tag{10}$$

where $L_{bd} = L_b{}^c{}_{d;c} - L_b{}^c{}_{c;d}$, (10) are known as the *Weyl–Lanczos relations*.

With the introduction of such a potential tensor for the conformal curvature tensor Lanczos achieved in Eq. (10) the gravitational analogue of the potential 4-vector A_a for the electromagnetic field $F_{ab} = A_{a;b} - A_{b;a}$.

Starting with the 20 Bianchi identities

$$R_{abcd;e} + R_{abde;c} + R_{abec;d} = 0, \tag{11}$$

which decompose into the 4 irreducible equations

$$G^{ae}{}_{;e} = 0, \tag{12}$$

where

$$G_{ab} = R_{ab} - \tfrac{1}{2}Rg_{ab} \tag{13}$$

is the Einstein tensor, and the 16 irreducible equations

$$C_{abc}{}^e{}_{;e} = J_{abc}, \tag{14}$$

where

$$J_{abc} = R_{c[a;b]} - \tfrac{1}{6}g_{c[a}R_{;b]} \tag{15}$$

is known as the Schouten tensor. We see that, like the potential field L_{abc}, the tensor field J_{abc} is irreducible with at most 16 independent components because of its symmetries and only 10 effective degrees of freedom because of the six identities

$$J_{ab}{}^s{}_{;s} = 0. \tag{16}$$

We note that the right-hand side of (15) can be rewritten entirely in terms of covariant derivatives of the *irreducible* tensor field S_{ab}, i.e., the traceless part of Ricci and Einstein tensors.

If we substitute (10) into (14), we obtain the following tensor wave equation:[2]

$$\begin{aligned} & \Box L_{abc} + R^n{}_{asc}L_b{}^s{}_n - R^n{}_{bsc}L_a{}^s{}_n - R^n{}_{abs}L_n{}^s{}_c + R_{nc}L_{ab}{}^n \\ & \quad - R_{nb}L_c{}^n{}_a + R_{na}L_c{}^n{}_b + R^l{}_{bns}L_l{}^{ns}g_{ac} - R^l{}_{ans}L_l{}^{ns}g_{bc} = J_{abc}. \end{aligned} \tag{17}$$

Dolan and Kim[2] further simplified this wave equation and made it *linear* with the use of the following identity (valid only in four dimensions):

$$C^n{}_{asc}L_b{}^s{}_n - C^n{}_{bsc}L_a{}^s{}_n - C^n{}_{abs}L_n{}^b{}_c + C^m{}_{bns}L_m{}^{sn}g_{ac} - C^m{}_{ans}L_m{}^{sn}g_{bc} = 0, \tag{18}$$

resulting in

$$\Box L_{abc}+2R_c{}^s L_{abs}-R_a{}^s L_{bcs}+R_b{}^s L_{acs}-g_{ac}R^{ls}L_{lbs}+g_{bc}R^{ls}L_{las}-\tfrac{1}{2}RL_{abc}=J_{abc}. \tag{19}$$

In empty space, this wave equation reduces to the remarkably simple equation

$$\Box L_{abc}=0. \tag{20}$$

There are several ways of solving the Weyl–Lanczos relations, though none are ever as straightforward as one would like them to be. The three most familiar ways are by a method of general observers,[4] a method of Killing vectors[3] (not necessarily timelike), or the most obvious of all is a simple comparison of the Weyl–Lanczos relations (WL) with the Newman–Penrose (NP) field equations in order to try to identify some sort of equivalence between the spin coefficients and the Lanczos coefficients. This approach may be inappropriate and is not at all intuitive unless one has made an appropriate choice of tetrad, canonical or otherwise, where there are special relations between the spin coefficients and many of them are made to vanish. This is exactly what we aim to do in this paper. We want to show that, given the appropriate choice of tetrad, we can reduce the Lanczos coefficients (in the cases of stationary axial symmetry and cylindrical symmetry) to being simple combinations of the spin coefficients. We begin with a very quick review of some of the known ways of determining the Lanczos tensor for some space–times. It is hoped that, in finding the Lanczos potential for vacuum space–times with an Ernst potential, we can relate the two and thereby have a tensor form of the Ernst potential. The full physical significance of the Lanczos tensor is thought to lie in its suitability for the quantization of the gravitational field[5–7] in the same way as quantum electrodynamics relies on the electromagnetic four vector A_a. Just as the wave equation for the electromagnetic field is derivable from the wave equation for the electromagnetic four vector, so is the Penrose curved space wave equation for the Weyl tensor derivable from the wave equation for the Lanczos tensor (19). So both classical theories are formally consistent in the same way.

General observers. Solutions of the Lanczos potential can be found in tensor form through the method of general observers described by a unit timelike vector u_a (Ref. 4) as follows:

$$u^e u_e=1.$$

We define the electric and the magnetic parts of the Weyl tensor as

$$E_{ac}=C_{abcd}u^b u^d, \tag{21}$$

$$H_{ac}={}^*C_{abcd}u^b u^d=\tfrac{1}{2}\eta_{ab}{}^{ef}C_{efcd}u^b u^d, \tag{22}$$

where η_{abcd} is the completely antisymmetric Levi-Civita tensor, thus, both E_{ac} and H_{ac} are spacelike, symmetric and traceless, i.e.,

$$\begin{aligned} &E_{ac}=E_{ca}, \quad E_{ae}u^e=0, \quad E^e{}_e=0,\\ &H_{ac}=H_{ca}, \quad H_{ae}u^e=0, \quad H^e{}_e=0, \end{aligned} \tag{23}$$

and the relation which enables us to go back from these to the original Weyl tensor is

$$\begin{aligned} C_{abcd}=&2u_a u_c E_{bd}+2u_b u_d E_{ac}-2u_a u_d E_{bc}-2u_b u_c E_{ad}\\ &+g_{ad}E_{bc}+g_{bc}E_{ad}-g_{ac}E_{bd}-g_{bd}E_{ac}\\ &+\eta_{ab}{}^{pq}u_p u_d H_{qc}-\eta_{ab}{}^{pq}u_p u_c H_{qd}\\ &+\eta_{cd}{}^{pq}u_b u_p H_{aq}-\eta_{cd}{}^{pq}u_a u_p H_{bq}. \end{aligned} \tag{24}$$

The projection tensor h_{ab} is defined as

$$h_{ab}=g_{ab}-u_a u_b, \tag{25}$$

and has the following properties:

$$h_a{}^e h_{eb} = h_{ab}, \quad h_{ab} = h_{ba}, \quad h^e{}_e = 3, \quad h_{ae} u^e = 0.$$

The covariant derivative of u_a may now be separated into its irreducible parts

$$u_{a;b} = \sigma_{ab} + \tfrac{1}{3}\theta h_{ab} + \omega_{ab} + a_a u_b, \tag{26}$$

where the shear tensor is symmetric and traceless and is given by

$$\sigma_{ab} = h_a{}^e h_b{}^f u_{(e;f)} - \tfrac{1}{3}\theta h_{ab}, \tag{27}$$

the vorticity tensor is antisymmetric and is given by

$$\omega_{ab} = h_a{}^e h_b{}^f u_{[e;f]}, \tag{28}$$

the expansion is

$$\theta = u^e{}_{;e}, \tag{29}$$

the acceleration vector is

$$a_a = u_{a;e} u^e, \tag{30}$$

and all the subtensors $\sigma_{ab}, \omega_{ab}, a_a$ are spacelike, i.e.,

$$\sigma_{ae} u^e = 0, \quad \sigma^e{}_e = 0, \quad \omega_{ae} u^e = 0, \quad a_e u^e = 0. \tag{31}$$

The vorticity tensor is equivalent to a vorticity vector

$$\omega^a = \tfrac{1}{2}\eta^{aefg}\omega_{ef} u_g, \quad \Rightarrow \omega_{ab} = \eta_{abef}\omega^e u^f. \tag{32}$$

Using the three dynamical equations (33), (34), (35) and the three equations of constraint (36), (37), (38) obeyed by these four irreducible quantities, namely

$$\dot{\theta} + \frac{\theta^2}{3} + \sigma^2 - \omega^2 - a^e{}_{;e} = R_{gh} u^g u^h, \tag{33}$$

where $\dot{\theta} = \theta_{;e} u^e$, $\sigma^2 = \sigma_{ef}\sigma^{ef}$, $\omega^2 = \omega_{ef}\omega^{ef}$, known as Raychaudhuri's equation, together with

$$\begin{aligned} &h_a{}^e h_b{}^f \dot{\sigma}_{ef} - \tfrac{1}{3} h_{ab}(\sigma^2 - \omega^2 - a^h{}_{;h}) + a_a a_b - h_a{}^e h_b{}^f a_{(e;f)} + \tfrac{2}{3}\theta\sigma_{ab} + \sigma_{ae}\sigma^e{}_b + \omega_{ae}\omega^e{}_b \\ &\quad = R_{aebf} u^e u^f - \tfrac{1}{3} R_{ef} u^e u^f h_{ab}, \end{aligned} \tag{34}$$

where $\dot{\sigma}_{ab} = \sigma_{ab;e} u^e$, and

$$h_a{}^e h_b{}^f \dot{\omega}_{ef} - h_a{}^e h_b{}^f a_{[e;f]} + \tfrac{2}{3}\theta\omega_{ab} + \sigma_{ae}\omega^e{}_b - \sigma_{be}\omega^e{}_a = 0, \tag{35}$$

where $\dot{\omega}_{ab} = \omega_{ab;e} u^e$, governing the behavior of the shear and the vorticity, and

$$\tfrac{2}{3}\theta_{,e} h^e{}_a - (\sigma^f{}_e + \omega^f{}_e)_{;f} h^e{}_a - a^e(\sigma_{ae} + \omega_{ae}) = R_{ef} u^e h^f{}_a, \tag{36}$$

$$\omega^e{}_{;e} + 2\omega^e a_e = 0, \tag{37}$$

$$h_{(a}{}^e h_{b)}{}^f \eta_e{}^{ghi} u_i (\sigma_{fg} + \omega_{fg})_{;h} + a_{(a}\omega_{b)} = H_{ab}, \tag{38}$$

Novello and Velloso have shown that the Lanczos tensors of several space–times are also expressible in terms of these quantities, the only restriction being that the space–time has to be a perfect fluid[8] with

$$T_{ab} = \rho u_a u_b - p h_{ab}. \tag{39}$$

Their results correspond to the following three lemmas which we quote below.

Lemma 1: If in a given space–time, there is a field of observers u^a that is shear-free irrotational and expansion-free ($\sigma_{ab}=\omega_{ab}=\theta=0$), *then the Lanczos potential is given by*

$$L_{abc}=a_a u_b u_c-a_b u_a u_c-\tfrac{1}{3}(a_a g_{bc}-a_b g_{ac}). \tag{40}$$

Lemma 2: If in a given space–time there is a field of observers u^a which is geodetic and irrotational ($a_a=\omega_{ab}=0$), *and such that either*
$H_{ab}\neq 0$ *and*

$$(i)\quad \sigma_a{}^e\sigma_{eb}-\sigma^2 h_{ab}-\tfrac{1}{3}\theta\sigma_{ab}=0, \tag{41}$$

or $H_{ab}=0$ *and*

$$(ii)\quad \dot{\sigma}_{ab}+\theta\sigma_{ab}=0, \tag{42}$$

then the Lanczos potential is given by

$$(i)\quad L_{abc}=\sigma_{ac}u_b-\sigma_{bc}u_a\,, \tag{43}$$

and

$$(ii)\quad L_{abc}=\tfrac{1}{3}(\sigma_{ac}u_b-\sigma_{bc}u_a), \tag{44}$$

respectively.

Lemma 3: If in a given space–time there is a field of observers u^a which is geodetic, shear-free and expansion-free such that the vorticity vector ω_a is covariantly constant

$$\omega_{a;b}=0, \tag{45}$$

then the Lanczos potential is given by

$$L_{abc}=\tfrac{1}{9}[2\omega_{ab}u_c+\omega_{ac}u_b-\omega_{bc}u_a]. \tag{46}$$

A fourth lemma was introduced by Dolan and Kim[3] and consists of a generalization of the first lemma above to any vacuum space–time with a hypersurface-orthogonal Killing vector.

Lemma 4: If a given space–time has a non-null hypersurface-orthogonal Killing vector K_a, $K_{[a;b}K_{c]}=0$, *and,*

$$u_a=\frac{K_a}{K},\quad K=\sqrt{K^eK_e},\quad u^eu_e=\epsilon,\quad \epsilon=\pm 1,$$

then the Lanczos tensor is given by

$$L_{abc}=a_a u_b u_c-a_b u_a u_c-\tfrac{1}{3}\epsilon(a_a g_{bc}-a_b g_{ac}). \tag{47}$$

Using *Lemmas 1* and *4* Dolan and Kim[3] were able to produce two Lanczos tensors for the Schwarzschild space–time, the difference between the two being attributable to a gauge transformation. Also, using *Lemma 4*, they found a Lanczos tensor for the Bondi axially symmetric space–time[9] which is exact and not approximate or given asymptotically.

Now Ernst[10,11] together with Plebañski[12] in examining the work of Papapetrou[13] on stationary axisymmetric space–times have introduced a complex potential $\mathscr{E}$ related to the norm K^2 and the vorticity vector associated with the timelike Killing vector K_a,[14]

$$\mathscr{E}_{,a}=2K^s(K_{s;a}+\tfrac{1}{2}i\,\eta_{astu}K^{t;u}). \tag{48}$$

This should give a very wide class of vacuum space–times for which to generate Lanczos potentials in a straightforward manner.

II. STATIONARY AXIALLY SYMMETRIC VACUUM SPACE–TIMES

The line element for stationary axisymmetric space–times can be expressed as follows in Weyl canonical coordinates (t,φ,ρ,z), and signature $(+,-,-,-)$:

$$ds^2=f(dt-wd\varphi)^2-f^{-1}[P^{-2}(d\rho^2+dz^2)+R^2d\varphi^2], \tag{49}$$

where f, P, R and w are functions of ρ and z only.

For a comprehensive review of the definitions of axisymmetric space–times together with the vacuum field equations we refer the reader to Ref. 15. We can introduce four orthogonal 1-forms in order to describe the line element (49) as follows:[16]

$$\begin{aligned}\sigma^1&=f^{1/2}(dt-wd\varphi),\quad \sigma^2=f^{-1/2}Rd\varphi,\\ \sigma^3&=f^{-1/2}P^{-1}d\rho,\quad \sigma^4=f^{-1/2}P^{-1}dz,\end{aligned} \tag{50}$$

$$g_{ab}dx^adx^b=(\sigma^1)^2-(\sigma^2)^2-(\sigma^3)^2-(\sigma^4)^2. \tag{51}$$

From (50), we can define a conventional null tetrad system

$$\begin{aligned}l_a&=(1/\sqrt{2})(f^{1/2},-(f^{-1/2}R+f^{1/2}w),0,0)=(1/\sqrt{2})(\sigma^1-\sigma^2),\\ n_a&=(1/\sqrt{2})(f^{1/2},f^{-1/2}R-f^{1/2}w,0,0)=(1/\sqrt{2})(\sigma^1+\sigma^2),\\ m_a&=(1/\sqrt{2})f^{-1/2}P^{-1}(0,0,1,i)=(1/\sqrt{2})(\sigma^3+i\sigma^4),\\ \overline{m}_a&=(1/\sqrt{2})f^{-1/2}P^{-1}(0,0,1,-i)=(1/\sqrt{2})(\sigma^3-i\sigma^4).\end{aligned} \tag{52}$$

Stationary axially symmetric space–times are best described in terms of their Ernst potential $\mathscr{E}$ explained below. If we consider a coordinate transformation on (49) to prolate spheroidal coordinates of the form

$$\rho=k(x^2-1)^{1/2}(1-y^2)^{1/2}, \tag{53}$$

$$z=kxy, \tag{54}$$

where k is an arbitrary constant, then the line element becomes

$$ds^2=f(dt-wd\varphi)^2-\frac{1}{f}\left[P^{-2}\left(\frac{dx^2}{x^2-1}+\frac{dy^2}{1-y^2}\right)+R^2d\varphi^2\right]. \tag{55}$$

Ernst[10,11] showed that the field equations governing the gravitational field of a uniformly rotating axially symmetric source may be derived from a single potential $\mathscr{E}$, now known as the Ernst potential. The expression of the metric functions in terms of the Ernst potential can be described as follows.

We rewrite f in terms of two complex functions $\hat{\alpha}$ and $\hat{\beta}$,

$$f=\frac{\hat{\alpha}\overline{\hat{\alpha}}-\hat{\beta}\overline{\hat{\beta}}}{(\hat{\alpha}+\hat{\beta})(\overline{\hat{\alpha}}+\overline{\hat{\beta}})}, \tag{56}$$

$$P^2=\frac{\hat{P}^2}{\hat{\alpha}\overline{\hat{\alpha}}-\hat{\beta}\overline{\hat{\beta}}}, \tag{57}$$

where $\hat{P}=\hat{P}(x,y)$, $\hat{\alpha}=\hat{\alpha}(x,y)$, $\hat{\beta}=\hat{\beta}(x,y)$, and neither $\hat{\alpha}$ nor $\hat{\beta}$ bear any relation to the NP spin coefficients. If we now consider the function $\mathscr{E}$, defined below

$$\mathscr{E}=\frac{\hat{\alpha}-\hat{\beta}}{\hat{\alpha}+\hat{\beta}}, \tag{58}$$

we see that this can be rewritten as

$$\mathscr{E}=\frac{(\hat{\alpha}-\hat{\beta})(\overline{\hat{\alpha}}+\overline{\hat{\beta}})}{(\hat{\alpha}+\hat{\beta})(\overline{\hat{\alpha}}+\overline{\hat{\beta}})}=\frac{\hat{\alpha}\overline{\hat{\alpha}}-\hat{\beta}\overline{\hat{\beta}}}{(\hat{\alpha}+\hat{\beta})(\overline{\hat{\alpha}}+\overline{\hat{\beta}})}+\frac{\hat{\alpha}\overline{\hat{\beta}}-\hat{\beta}\overline{\hat{\alpha}}}{(\hat{\alpha}+\hat{\beta})(\overline{\hat{\alpha}}+\overline{\hat{\beta}})}, \tag{59}$$

and so we can write the defining equation of f in terms of $\mathscr{E}$ as

$$f=\mathrm{Re}[\mathscr{E}]=\mathrm{Re}\left[\frac{\hat{\alpha}-\hat{\beta}}{\hat{\alpha}+\hat{\beta}}\right]=\mathrm{Re}\left[\frac{\xi-1}{\xi+1}\right], \tag{60}$$

where $\mathscr{E}$ is the Ernst potential and the function $\xi=\hat{\alpha}/\hat{\beta}$ is a related potential which is often simpler to use.

The remaining functions w, R, and P in the metric can be best looked at if we go back to cylindrical coordinates and redefine P as being an exponential by

$$P=e^{\hat{\gamma}},$$

where $\hat{\gamma}=\hat{\gamma}(x,y)$, and the metric now is

$$ds^2=f(dt-wd\varphi)^2-f^{-1}[e^{-2\hat{\gamma}}(d\rho^2+dz^2)+R^2d\varphi^2]. \tag{61}$$

In this notation, Ernst[10,11] showed that the remaining functions satisfy the following differential equations:

$$\frac{\partial w}{\partial \rho}=-\frac{2\rho}{(\xi\overline{\xi}-1)^2}\,\mathrm{Im}\left[(\overline{\xi}+1)^2\,\frac{\partial \xi}{\partial z}\right], \tag{62}$$

$$\frac{\partial w}{\partial z}=-\frac{2\rho}{(\xi\overline{\xi}-1)^2}\,\mathrm{Im}\left[(\overline{\xi}+1)^2\,\frac{\partial \xi}{\partial \rho}\right], \tag{63}$$

$$\frac{\partial \gamma}{\partial \rho}=\frac{\rho}{(\xi\overline{\xi}-1)^2}\left[\frac{\partial \xi}{\partial \rho}\frac{\partial \overline{\xi}}{\partial \rho}-\frac{\partial \xi}{\partial z}\frac{\partial \overline{\xi}}{\partial z}\right], \tag{64}$$

$$\frac{\partial \gamma}{\partial z}=\frac{2\rho}{(\xi\overline{\xi}-1)^2}\,\mathrm{Re}\left[\frac{\partial \xi}{\partial \rho}\frac{\partial \overline{\xi}}{\partial z}\right]. \tag{65}$$

Thus any stationary axially symmetric line element can be fully described (up to constants of integration) by (60)–(65), and from now on we can consider the Ernst potential as defining the whole space–time in question.

Ernst[16] and Höenselaers[17] further showed that the Weyl scalars for all vacuum stationary axially symmetric space–times are given by

$$\Psi_0=\frac{1}{8}\left[2\eth\eth\mathscr{E}+\frac{1}{f}\,(\eth\mathscr{E})^2\right], \tag{66}$$

$$\Psi_2=-\frac{1}{8}\left[-2\eth^*\eth\mathscr{E}+\frac{1}{f}\,(\eth\mathscr{E})(\eth^*\mathscr{E})\right], \tag{67}$$

$$\Psi_4=\frac{1}{8}\left[2\eth^*\eth^*\mathscr{E}+\frac{1}{f}\,(\eth^*\mathscr{E})^2\right], \tag{68}$$

$$\Psi_1=\Psi_3=0, \tag{69}$$

where the differential operator $\eth$ is defined as follows:

Given a field η of spin weight s we have

$$\eth\eta=P^{1-s}\nabla(P^s\eta), \quad \eth^*\eta=P^{1+s}\nabla^*(P^{-s}\eta), \tag{70}$$

and in the ρ, z coordinate system we had earlier $\nabla=\partial/\partial\rho+i\,\partial/\partial z$ and ∇^* is its complex conjugate.

Thus the operators $\eth$ and $\eth^*$ raise and lower the spin-weight by one, respectively. In our case *the Ernst potential $\mathscr{E}$ is of spin-weight* 0. It should be noted that, though related to the operator $\eth$ as defined by Penrose,[18] this operator is to be considered as completely distinct.

Using the formalism developed above, it is also possible to give the Newman–Penrose spin coefficients *almost entirely* in terms of derivatives of the Ernst potential,[16]

$$\alpha=-\frac{1}{2\sqrt{2}}f^{1/2}\eth(\ln P)-\frac{1}{4\sqrt{2}}f^{-1/2}\eth\mathscr{E}, \tag{71}$$

$$\beta=\frac{1}{\sqrt{2}}f^{1/2}\eth^*(\ln P)+\frac{1}{4\sqrt{2}}f^{-1/2}\eth^*\mathscr{E}, \tag{72}$$

$$\kappa=-\frac{1}{2\sqrt{2}}f^{1/2}R^{-1}\eth R+\frac{1}{2\sqrt{2}}f^{-1/2}\eth\mathscr{E}, \tag{73}$$

$$\nu=\frac{1}{2\sqrt{2}}f^{1/2}R^{-1}\eth^*R-\frac{1}{2\sqrt{2}}f^{-1/2}\eth^*\mathscr{E}, \tag{74}$$

$$\tau=\frac{1}{2\sqrt{2}}f^{1/2}R^{-1}\eth R, \tag{75}$$

$$\pi=-\frac{1}{2\sqrt{2}}f^{1/2}R^{-1}\eth^*R. \tag{76}$$

It should be noted that because only Ψ_0, Ψ_2 and Ψ_4 can be nonzero, stationary axially symmetric space–times can only be Petrov types I, II, D or N.

A. Construction of the Lanczos tensor

Because of our choice of tetrad, and the fact that the metric components depend only on two variables (ρ and z), we were able to discard any terms involving the D or the Δ operators in the Newman–Penrose[19] field equations as well as being helped by the vanishing of six of the spin coefficients. Following the method introduced by Sibgatulin,[20] of expressing the Weyl scalars solely in terms of spin coefficients and their derivatives, we were able to express the field equations in the following form:

$$\epsilon=\gamma=\sigma=\lambda=\rho=\mu=0, \quad \tau=-\overline{\pi}, \tag{77}$$

$$\begin{aligned} D\zeta&=\frac{1}{\sqrt{2}}\left[(f^{-1/2}+f^{1/2}\omega R^{-1})\frac{\partial}{\partial t}+f^{1/2}R^{-1}\frac{\partial}{\partial\varphi}\right]\zeta=0, \quad \forall\zeta=\zeta(\rho,z),\\ \Delta\zeta&=\frac{1}{\sqrt{2}}\left[(f^{-1/2}-f^{1/2}\omega R^{-1})\frac{\partial}{\partial t}-f^{1/2}R^{-1}\frac{\partial}{\partial\varphi}\right]\zeta=0, \quad \forall\zeta=\zeta(\rho,z), \end{aligned} \tag{78}$$

$$\Psi_0=-(\delta-3\beta-\tau-\overline{\alpha}+\overline{\pi})\kappa, \tag{79}$$

$$\Psi_2=(\overline{\delta}+\overline{\beta}-\overline{\tau}-\alpha)\tau+\kappa\nu, \tag{80}$$

$$\Psi_2=(-\delta+\overline{\alpha}-\overline{\pi}-\beta)\pi+\kappa\nu, \tag{81}$$

$$\Psi_4=(\overline{\delta}+\pi-\overline{\tau}+3\alpha+\overline{\beta})\nu, \tag{82}$$

$$\Psi_2=\nu\kappa-\tau\pi, \tag{83}$$

$$\Psi_1=\Psi_3=0. \tag{84}$$

We now examine the Weyl–Lanczos relations in scalar form as given by Dolan and Kim[3] including a factor of (-1) on the left-hand side which had been erroneously omitted, giving the gauge conditions $L_{ab}{}^{e}{}_{;e}=0$ separately,

$$-\tfrac{1}{2}\Psi_0=DL_4-(\delta-3\beta-\overline{\alpha}+\overline{\pi})L_0+3\kappa L_5, \tag{85}$$

$$-2\Psi_1=3DL_5-3(\delta-\beta-\overline{\alpha}+\overline{\pi})L_1-3\pi L_4+6\kappa L_6+(\overline{\delta}-3\alpha+\overline{\beta}-\overline{\tau})L_4-\Delta L_0-3\tau L_1, \tag{86}$$

$$-\Psi_2=DL_6-(\delta+\beta-\overline{\alpha}+\overline{\pi})L_2-2\pi L_5+\kappa L_7+(\overline{\delta}-\alpha+\overline{\beta}-\overline{\tau})L_5-\Delta L_1+\nu L_0-2\tau L_2, \tag{87}$$

$$-2\Psi_3=3(\overline{\delta}+\alpha+\overline{\beta}-\overline{\tau})L_6-3\Delta L_2+6\nu L_1-3\tau L_3+DL_7-(\delta+3\beta-\overline{\alpha}+\overline{\pi})L_3-3\pi L_6, \tag{88}$$

$$-\tfrac{1}{2}\Psi_4=(\overline{\delta}+3\alpha+\overline{\beta}-\overline{\tau})L_7-\Delta L_3+3\nu L_2, \tag{89}$$

where

$$\begin{gathered}\Psi_0=C_{abcd}l^a m^b l^c m^d,\quad \Psi_1=C_{abcd}l^a m^b l^c n^d,\quad \Psi_2=C_{abcd}l^a m^b \overline{m}^c n^d,\\ \Psi_3=C_{abcd}l^a n^b \overline{m}^c n^d,\quad \Psi_4=C_{abcd}\overline{m}^a n^b \overline{m}^c n^d,\end{gathered} \tag{90}$$

$$\begin{gathered}L_0=L_{abc}l^a m^b l^c,\quad L_1=L_{abc}l^a m^b \overline{m}^c,\quad L_2=L_{abc}\overline{m}^a n^b l^c,\\ L_3=L_{abc}\overline{m}^a n^b \overline{m}^c,\quad L_4=L_{abc}l^a m^b m^c,\quad L_5=L_{abc}l^a m^b n^c,\\ L_6=L_{abc}\overline{m}^a n^b m^c,\quad L_7=L_{abc}\overline{m}^a n^b n^c.\end{gathered} \tag{91}$$

The gauge conditions are

$$DL_5-(\delta-\beta-\overline{\alpha}+\overline{\pi}-3\tau)L_1+2\kappa L_6-(\overline{\delta}-3\alpha+\overline{\beta}-\overline{\tau}+\pi)L_4+\Delta L_0=0, \tag{92}$$

$$DL_6-(\delta+\beta-\overline{\alpha}+\overline{\pi}-2\tau)L_2+\kappa L_7-(\overline{\delta}-\alpha+\overline{\beta}-\overline{\tau}+2\pi)L_5+\Delta L_1-\nu L_0=0, \tag{93}$$

$$(\overline{\delta}+\alpha+\overline{\beta}-\overline{\tau}+3\pi)L_6-\Delta L_2+2\nu L_1-DL_7+(\delta+3\beta-\overline{\alpha}+\overline{\pi}-\tau)L_3=0. \tag{94}$$

Because all the tensors involved depend only on the two coordinates ρ and z, we can reasonably assume the same for the Lanczos scalars $L_0,\ldots,L_7$, hence, $DL_n=\Delta L_n=0$ for $n=0-7$. Four equations [(86), (88), (92), (94)] may be solved by setting $L_1=L_4=L_6=L_3=0$, thus leaving us with four equations [(85), (87), (89), (93)] still to be satisfied. There are only five relevant NP equations required to do this, namely (79), (80), (81), (82) and the purely algebraic relation (83). Again, Eqs. (85) and (89) can be solved by comparison with the NP equations and we can then verify that this solution is also in agreement with (87) and (93) thus giving us the simple answer

$$\begin{gathered}L_0=-\tfrac{1}{2}\kappa,\quad L_5=-\tfrac{1}{6}\tau,\\ L_7=-\tfrac{1}{2}\nu,\quad L_2=-\tfrac{1}{6}\pi.\end{gathered} \tag{95}$$

To illustrate (95), we give some well-known examples of stationary axially symmetric vacuum space–times, but it should be stressed that all line elements which can be written in the form (55) and satisfy the vacuum Einstein equations will have (95) as a solution for their Lanczos tensor to within a gauge transformation.

B. Tomimatsu–Sato, Kerr and Schwarzschild solutions

If we now return to the equation for axially symmetric stationary line elements (55), we can obtain various well-known space–times via specification of the Ernst potential $\mathscr{E}$. We can write the associated function ξ in a general way as

$$\xi=\frac{\hat{\alpha}(x,y;p,q,\delta)}{\hat{\beta}(x,y;p,q,\delta)}, \tag{96}$$

where δ is a constant (not necessarily an integer) and both the functions $\hat{\alpha}$ and $\hat{\beta}$ can be given in terms of polynomials of order δ^2 and δ^2-1, respectively, and p and q are scalar polynomials in δ satisfying the following algebraic relation:

$$p^2+q^2=1.$$

Thus both the functions $\hat{\alpha}$ and $\hat{\beta}$ can be expressed in terms of polynomials in x, y, p and q for various values of δ as follows:

1. $\delta=1$

$$\hat{\alpha}=px-iqy, \quad \hat{\beta}=1, \tag{97}$$

$$\xi=px-iqy, \tag{98}$$

so f is determined by (56) or (60) and $e^{\hat{\gamma}}$ is given by

$$e^{\hat{\gamma}}=\frac{A[x^2p^2+q^2y^2-1]}{x^2-y^2}, \tag{99}$$

where A is an arbitrary constant of integration of (65) and (64). w is given by

$$w=\frac{2qp^{-1}(1-y^2)(px+1)}{x^2p^2+q^2y^2-1}+w_0, \tag{100}$$

where w_0 is again a constant of integration of (62) and (63).

This choice of ξ gives us the Kerr metric[21] in prolate spheroidal coordinates and the Lanczos tensor is then given by

$$L_5=-\tfrac{1}{6}\tau, \quad L_2=-\tfrac{1}{6}\pi. \tag{101}$$

The reasons for this simple solution lie entirely in the choice of coordinates and of tetrad. Had we decided, for example, to use more well known coordinates such as the Boyer–Lindquist coordinates[22] for the Kerr metric, together with the Kinnersley tetrad,[23] things would not have simplified themselves in the desired manner and Eq. (78) would not have been satisfied. We also note that, due to (77), the solution of the Lanczos tensor for the Kerr metric only depends on one of the spin coefficients π or τ.

The Schwarzschild solution is contained as the nonrotating subset of the above solution; this is most easily seen if we transform (98) to the better known Boyer–Lindquist coordinates mentioned above via the following local coordinate transformation:[15]

$$px+1=\frac{r}{m}, \quad qy=\frac{a}{m}\cos\,\theta, \quad \varphi'=\varphi, \quad t'=t. \tag{102}$$

And the Schwarzschild solution is given by the $a=0$ case of (102), or, equivalently $q=0$.

2. $\delta=2$

$$\hat{\alpha}=p^2x^4+q^2y^4-1-2ipqxy(x^2-y^2), \quad \hat{\beta}=2px(x^2-1)-2iqy(1-y^2), \tag{103}$$

$$\xi=\frac{p^2x^4+q^2y^4-1-2ipqxy(x^2-y^2)}{2px(x^2-1)-2iqy(1-y^2)}. \tag{104}$$

Because of the increasing complexity of the Ernst function, ξ, and all the surrounding equations, it is profitable to introduce three new functions A, B, and C which enable us to write f, w, and $e^{\hat{\gamma}}$ as

$$f=\frac{A}{B}, \quad w=\frac{2p^{-1}q(1-y^2)C}{A}, \quad e^{\hat{\gamma}}=\frac{A}{p^4(x^2-y^2)^4}, \tag{105}$$

where

$$A=[p^2(x^2-1)^2+q^2(1-y^2)^2]^2-4p^2q^2(x^2-1)(1-y^2)(x^2-y^2)^2, \tag{106}$$

$$B=(p^2x^4+q^2y^4-1+2px^3-2px)^2+4q^2y^2(px^3-pxy^2+1-y^2)^2, \tag{107}$$

$$\begin{aligned}C=&p^2(x^2-1)[(x^2-1)(1-y^2)-4x^2(x^2-y^2)]\\&-p^3x(x^2-1)[2(x^4-1)+(x^2+3)(1-y^2)+q^2(1+px)(1-y^2)^2].\end{aligned} \tag{108}$$

This choice of Ernst potential leads to the well known $\delta=2$ Tomimatsu–Sato[24,25] solution which is algebraically general and is thought to describe the field of a spinning mass with a ring singularity on the equatorial plane outside the event horizon as well as inside it (as is the case for the Kerr metric). The Lanczos potential may be computed from (95) and has all four Lanczos scalars nonvanishing.

3. General integer δ

Tomimatsu and Sato[25] gave an approximate expression for the function ξ for any integer value of δ in the limit $q\ll 1$,

$$\xi\simeq\frac{(x+1)^\delta+(x-1)^\delta}{(x+1)^\delta-(x-1)^\delta}+iq\xi_1(x,y)+\cdots, \tag{109}$$

and

$$\xi_1(x,y)=\frac{1}{[(x+1)^\delta-(x-1)^\delta]^2}\sum_{l=1}^{\delta}a_{2l-1}(x)P_{2l-1}(y), \tag{110}$$

where $P_l(y)$ is a Legendre polynomial and $a_l(x)$ is a polynomial of at most order $2\delta-l-1$ and satisfying the following differential equation:

$$(x^2-1)\frac{d^2a_l}{dx^2}-(4\delta-2)x\frac{da_l}{dx}+(4\delta^2-l(l+1))a_l=0, \tag{111}$$

and $e^{\hat{\gamma}}$ is given by

$$e^{\hat{\gamma}}=\frac{A}{p^{2\delta}(x^2-y^2)^{\delta^2}}. \tag{112}$$

Again the Lanczos scalars may be computed in terms of (95). It is interesting to notice that for the case $q=0$, all the Tomimatsu solutions with the Ernst potential given above reduce to the Weyl–Papapetrou solutions, where we only have the first term in the expansion of ξ in Eq. (109) above, namely

$$\xi=\frac{(x+1)^{\delta}+(x-1)^{\delta}}{(x+1)^{\delta}-(x-1)^{\delta}}. \tag{113}$$

Recently, Hori[26–29] further generalized the Tomimatsu–Sato solutions to include the cases of *noninteger* δ as well as the cases where δ is *complex*. Hori found that *only* the cases where the deformation parameter δ is an integer lead to asymptotic flatness. The Weyl–Lanczos relations in these cases are also solved by the Lanczos scalars (95). We note that the solution (95) is also valid in nonempty space–times provided $\Phi_{11}=\Lambda=0$.

C. *C*-metric

To see that the *C*-metric is also a stationary axially symmetric metric and that therefore the appropriate nonvanishing Lanczos scalars may be given by (95), we first have to consider a coordinate transformation. We start with the line element as[30]

$$ds^2=A^{-2}(x+y)^{-2}\left[Fdt^2-\frac{1}{F}\,dy^2-\frac{1}{G}\,dx^2-Gdp^2\right], \tag{114}$$

in coordinates (t,x,y,p), signature $(+,-,-,-)$, and where

$$F=-1+y^2-2mAy^3, \quad G=1-x^2-2mAx^3, \tag{115}$$

and m and A are constants, then consider a transformation of the form

$$z=A^{-2}(x+y)^{-2}[1+mAxy(x-y)+xy], \tag{116}$$

$$r=A^{-2}(x+y)^{-2}F^{1/2}G^{1/2}, \tag{117}$$

and $\varphi=p$, $t=t$. This gives us the line element

$$ds^2=e^{\mu}dt^2-r^2e^{-\mu}d\varphi^2-e^{\omega}(dz^2+dr^2), \tag{118}$$

where ω and μ are functions of z and r.

We recognize this line element as being of the form (49) and thus stationary axially symmetric. And so the Lanczos scalars will be given by (95).

III. VACUUM CYLINDRICALLY SYMMETRIC SPACE–TIMES

We can find the solution for the Lanczos tensor-potential in a cylindrically symmetric space–time in a way analogous to the one we followed in finding the solution for stationary axially symmetric space–times in Sec. II. The line element for these space–times is[14]

$$ds^2=f^{-1}[e^{2\hat{\gamma}}(dt^2-dr^2)-r^2d\varphi^2]-f(dz-wd\varphi)^2, \tag{119}$$

in coordinates (t,r,φ,z) and signature $(+,-,-,-)$ and where f, $\hat{\gamma}$, w are all functions of r only.

We choose our tetrad to be

$$\begin{aligned} l_a&=(1/\sqrt{2})f^{-1/2}e^{\hat{\gamma}}(1,1,0,0),\\ n_a&=(1/\sqrt{2})f^{-1/2}e^{\hat{\gamma}}(1,-1,0,0),\\ m_a&=(1/\sqrt{2})(0,0,f^{-1/2}r+if^{1/2}w,-if^{1/2}),\\ \overline{m}_a&=(1/\sqrt{2})(0,0,f^{-1/2}r-if^{1/2}w,if^{1/2}). \end{aligned} \tag{120}$$

In this frame, the spin coefficients and the Weyl scalars satisfy

$$\alpha=\beta=\kappa=\nu=\tau=\pi=0, \quad \rho=\overline{\rho}, \quad \mu=\overline{\mu}, \tag{121}$$

and

$$\Psi_1=\Psi_3=0. \tag{122}$$

Construction of the Lanczos tensor. Following (120), it is also possible to show

$$\delta\zeta=\overline{\delta}\zeta=0, \quad \forall\,\zeta=\zeta(t,r). \tag{123}$$

Therefore the Newman–Penrose field equations[19] reduce, *in vacuo*, to

$$\Psi_0=D\sigma+(\rho+\overline{\rho}+\overline{\epsilon}-3\epsilon)\sigma, \tag{124}$$

$$\Psi_2=D\mu+(\epsilon+\overline{\epsilon}-\overline{\rho})\mu)-\sigma\lambda, \tag{125}$$

$$\Psi_2=-\Delta\rho+(\gamma+\overline{\gamma}-\overline{\mu})-\sigma\lambda, \tag{126}$$

$$\Psi_4=-\Delta\lambda-(\mu+\overline{\mu}-\overline{\gamma}+3\gamma)\lambda, \tag{127}$$

$$\Psi_2=\mu\rho-\sigma\lambda, \tag{128}$$

$$\Psi_1=\Psi_3=0. \tag{129}$$

The Weyl–Lanczos relations, when all the vanishing quantities have been removed, become

$$-\tfrac{1}{2}\Psi_0=(D-3\epsilon+\overline{\epsilon}-\overline{\rho})L_4-3\sigma L_1, \tag{130}$$

$$-\tfrac{1}{2}\Psi_2=(D+\epsilon+\overline{\epsilon}-\overline{\rho})L_6+2\mu L_1-\sigma L_3, \tag{131}$$

$$-\tfrac{1}{2}\Psi_2=-(\Delta-\gamma-\overline{\gamma}+\overline{\mu})L_1-\lambda L_4+2\rho L_6, \tag{132}$$

$$-\tfrac{1}{2}\Psi_4=-(\Delta+3\gamma-\overline{\gamma}+\overline{\mu})L_3-3\lambda L_6, \tag{133}$$

where we have already assumed that we can apply condition (123) to the Lanczos scalars and we have set L_0, L_2, L_4, and L_7 to zero in order to satisfy all equations involving Ψ_1 and Ψ_3 trivially. We are thus left with a strictly analogous system of equations to (79)–(84) and (85)–(89) whose solution is

$$\begin{aligned} &L_4=-1/2\sigma, \quad L_1=-1/6\rho, \\ &L_3=-1/2\lambda, \quad L_6=-1/6\mu. \end{aligned} \tag{134}$$

IV. CONCLUSION

We have given a closed form for the Lanczos tensor for some space–times with an Ernst potential. Through the completeness relation

$$L_{abc}=T_{abc}+\overline{T}_{abc}, \tag{135}$$

between the Lanczos tensor and the Lanczos scalars where

$$\begin{aligned} T_{abc}=&L_0U_{ab}n_c+L_1(M_{ab}n_c-U_{ab}m_c) \\ &+L_2(V_{ab}n_c-M_{ab}m_c)-L_3V_{ab}m_c \\ &-L_4U_{ab}\overline{m}_c+L_5(U_{ab}l_c-M_{ab}\overline{m}_c)+L_6(M_{ab}l_c-V_{ab}\overline{m}_c)+L_7V_{ab}l_c, \end{aligned} \tag{136}$$

and

$$\begin{aligned} &V_{ab}=l_am_b-l_bm_a, \quad U_{ab}=-n_a\overline{m}_b+n_b\overline{m}_a, \\ &M_{ab}=l_an_b-l_bn_a+m_a\overline{m}_b-m_b\overline{m}_a, \end{aligned} \tag{137}$$

we can express the Weyl tensor as coming from a three index tensor L_{abc}.

We note that our solution (95) is not entirely expressible in terms of the Ernst potential. This is due to the presence of the R^2 factor (respectively, r^2 for cylindrical symmetry) in g_{33}, without which we would have had $\tau=\pi=0$ (respectively, $\rho=\mu=0$) and so the Lanczos scalars would have been entirely expressible in terms of derivatives of the Ernst potential as follows:

$$L_0=\frac{1}{4\sqrt{2}}\,f^{-1/2}\eth\mathscr{E},\quad L_7=\frac{1}{4\sqrt{2}}\,f^{-1/2}\eth^*\mathscr{E}, \tag{138}$$

with corresponding expressions for the cylindrically symmetric case. In many articles in the literature (see, e.g., Refs. 31 and 32) the factor of R^2 (respectively, r^2) is set to one for simplicity since it has to satisfy Laplace's equation.

Future work includes examining whether the Ernst potential is still related to the Lanczos potential, in a similar way to the above description, in more general cases, namely those with metric,

$$g_{ab}=\begin{pmatrix} f & -f\omega_1 & -f\omega_2 & -f\omega_3 \\ -f\omega_1 & f\omega_1^2-f^{-1} & f\omega_1\omega_2 & f\omega_1\omega_3 \\ -f\omega_2 & f\omega_1\omega_2 & f\omega_2^2-f^{-1} & f\omega_2\omega_3 \\ -f\omega_3 & f\omega_1\omega_3 & f\omega_2\omega_3 & f\omega_3^2-f^{-1} \end{pmatrix}, \tag{139}$$

where the functions f and ω_α, $\alpha=1,2,3$ depend on all three space variables only and not the time. These spacetimes are no longer axially symmetric, but only stationary, though, as indicated earlier, an Ernst potential may still be defined for them.[32]

ACKNOWLEDGMENT

We thank the referee for very helpful comments and suggestions.

[1] C. Lanczos, ''The splitting of the Riemann tensor,'' Rev. Mod. Phys. **34**, 379–389 (1962).
[2] P. Dolan and C. W. Kim, ''The wave equation for the Lanczos potential. I,'' Proc. R. Soc. London, Ser. A **447**, 557–575 (1994).
[3] P. Dolan and C. W. Kim, ''Some solutions of the Lanczos vacuum wave equation,'' Proc. R. Soc. London, Ser. A **447**, 577–585 (1994).
[4] M. Novello and V. Velloso, ''The connection between general observers and the Lanczos potential,'' Gen. Relativ. Gravit. **19**, 1251–1265 (1987).
[5] M. Novello and L. M. C. S. Rodrigues, ''A unified model for gravity and electroweak interactions,'' Lett. Nuovo Cimento **43**, 292–296 (1985).
[6] M. Novello, L. R. de Freitas, N. Pinto-Neto, and N. F. Svaiter, ''Quantization of spin-two field in terms of Fierz variables—The linear case,'' Fortschr. Phys. **40**, 195–209 (1992).
[7] L. R. de Freitas, M. Novello, and N. Pinto-Neto, ''Nonlinear theories of spin-2 field in terms of Fierz variables,'' J. Math. Phys. **35**, 734–747 (1994).
[8] G. F. R. Ellis, ''Relativistic cosmology'' in *General Relativity and Cosmology*, Proceedings of Course 47 of the International school of Physics 'Enrico Fermi' (Academic, New York, 1971).
[9] H. Bondi, M. G. J. Van der Burg, and A. W. K. Metzner, ''Gravitational waves in general relativity. VII. Waves from axisymmetric isolated systems,'' Proc. R. Soc. London, Ser. A **269**, 21–52 (1962).
[10] F. J. Ernst, ''New formulation of the axially symmetric gravitational field. Problem I.,'' Phys. Rev. **167**, 1175–1178 (1968).
[11] F. J. Ernst, ''New formulation of the axially symmetric gravitational field. Problem II,'' Phys. Rev. **168**, 1415–1417 (1968).
[12] F. J. Ernst and J. F. Plebañski, ''Killing structures and complex $\mathscr{E}$ potentials,'' Ann. Phys. (N.Y.) **107**, 266–282 (1976).
[13] A. Papapetrou, ''Eine rotationssymmetrische Lösung in der allgemeinen Relativitätstheorie,'' Ann. Phys. (Leipzig) **12**, 309–315 (1953).
[14] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *The Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[15] C. Reina and A. Treves, ''Axisymmetric gravitational fields,'' Gen. Relativ. Gravit. **7**, 817–838 (1976).
[16] F. J. Ernst, ''Complex potential formulation of the axially symmetric gravitational field problem,'' J. Math. Phys. **15**, 1409–1412 (1974).
[17] C. Höenselaers, ''Weyl conform tensor of the Tomimatsu–Sato $\delta=3$ metric,'' Gen. Relativ. Gravit. **11**, 325–327 (1977).
[18] P. Geroch, A. Held, and R. Penrose, ''A space–time calculus based on pairs of null directions,'' J. Math. Phys. **14**, 874–881 (1973).

[19] E. T. Newman and R. Penrose, ''An approach to gravitational radiation by a method of spin coefficients,'' J. Math. Phys. **3**, 566–578 (1962).
[20] N. R. Sibgatulin, *Oscillations and Waves in Strong Gravitational and Electromagnetic Fields* (Springer, Berlin, 1991).
[21] R. P. Kerr, ''Gravitational field of a spinning mass as an example of algebraically special metrics,'' Phys. Rev. Lett. **11**, 237 (1963).
[22] R. H. Boyer and R. W. Lindquist, ''Maximal analytic extension of the Kerr metric,'' J. Math. Phys. **8**, 265–281 (1967).
[23] W. Kinnersley, ''Type D vacuum metrics,'' J. Math. Phys. **10**, 1195–2003 (1969).
[24] A. Tomimatsu and H. Sato, ''New exact solution for the gravitational field of a spinning mass,'' Phys. Rev. Lett. **29**, 1344–1345 (1972).
[25] A. Tomimatsu and H. Sato, ''New series of exact solutions for gravitational fields of spinning masses,'' Prog. Theor. Phys. **50**, 95–110 (1973).
[26] S. Hori, ''Generalisation of Tomimatsu solutions. I.'' Prog. Theor. Phys. **95**, 65–70 (1996).
[27] S. Hori, ''Generalisation of Tomimatsu solutions. II,'' Prog. Theor. Phys. **95**, 557–564 (1996).
[28] S. Hori, ''Generalisation of Tomimatsu solutions. III,'' Prog. Theor. Phys. **95**, 1097–1120 (1996).
[29] S. Hori, ''Generalisation of Tomimatsu solutions. IV,'' Prog. Theor. Phys. **96**, 327–345 (1996).
[30] W. Kinnersley and M. Walker, ''Uniformly accelerating charged mass in general relativity,'' Phys. Rev. D **2**, 1359–1370 (1970).
[31] F. J. Ernst, ''Exterior-algebraic derivation of Einstein field equations employing a generalised basis,'' J. Math. Phys. **12**, 2395–2397 (1971).
[32] F. J. Ernst, ''Weyl conform tensor for stationary gravitational fields,'' J. Math. Phys. **15**, 2027–2030 (1974).

Static stringlike solutions in five-dimensional relativity

Pierre Gravel[a)]
Département de Mathématiques et Informatique, Collège Militaire Royal du Canada, Kingston, Ontario K7K 5LO, Canada

(Received 5 November 1997; accepted for publication 3 June 1998)

We exhibit a family of static cylindrically symmetric solutions to $(4+1)$ Kaluza–Klein gravity which are separable functions of the spatial radial coordinate $x^2=r$ and the extra coordinate $x^5=\psi$. These contain solutions other than radiative and, when interpreted as string solutions, display behaviors which differ from the standard $(3+1)$-dimensional static strings. All of those solutions with nonvanishing mass density obey the relativistic matter equation of state. © *1998 American Institute of Physics.* [S0022-2488(98)02010-6]

I. INTRODUCTION

In this paper, the conventions about the metric, curvature tensor, and indices will be those of the book by Adler *et al.*[1] and the work of Wesson *et al.* (see Ref. 2 and references therein). In particular, the metric is of signature $(+,-,-,-,\pm)$, greek indices run from 1 to 5 and latin indices from 1 to 4, and the eigenvalues of the four-dimensional stress tensor will be $(\rho,-p_2,-p_3,-p_4)$.

The essence of the induced-matter approach to the Kaluza–Klein formalism is that ''matter'' should come from the behavior of the $x^5=\psi$ dependent part of the five-dimensional vacuum field equations without the fifth dimension being necessarily compactified. This, of course, is at variance with the original Kaluza–Klein idea, where the extra dimension is either actually absent or compactified on a small circle. In this nonprojective, noncylindrical reinterpretation of higher-dimensional pure gravity, the five-dimensional vacuum Einstein equations are each rewritten as a strictly four-dimensional part on the one side (the four-dimensional curvature) with the other, ''matter,'' side containing all terms depending on the $g_{\alpha 5}$'s and their derivatives, and the ψ derivatives of the $g_{\alpha\beta}$'s. Thus in this formalism, four-dimensional matter is a manifestation of five-dimensional pure gravity.

The standard tests of general relativity do not falsify the theory[3–5] and the formalism has been extended in a number of directions, including cosmology,[6,7] wave propagation,[8,9] inflation,[10] solitons/black holes,[11–13] and other special solutions.[14,15] The formalism has also been extended to the process of obtaining D-dimensional matter-gravity equations from $(D+1)$-dimensional empty space–time Einstein equations.[16] Here we shall restrict ourselves to the original $D=4$ case. For a recent review of the field, one should consult Ref. 17.

Our goal here is to find new solutions with cylindrical symmetry. Such solutions where the metric coefficients depend only on the radius and time coordinates already exist.[18] However, it is known that metrics which are independent of the fifth coordinate can lead to radiation only.[19] By allowing the metric to depend on the fifth coordinate, new, nonradiative solutions with cylindrical symmetry can be found. In this paper, new solutions of this type are found using the separability assumption and some of their properties are discussed.

II. FINDING SEPARABLE CYLINDRICAL METRICS

We are interested in studying diagonal metrics which extend Vilenkin's cylindrical form[20] in two ways: they have to describe five-dimensional spaces and their coefficients are to be allowed to vary with the radius r and the extra coordinate ψ. The axis of cylindrical symmetry is chosen to

[a)]Electronic mail: gravelp@sv2.rmc.ca

 © 1998 American Institute of Physics

be that of the z variable and the metrics are assumed to be invariant under boosts in that direction. If we also assume separability of the coefficients with respect to the radial and the extra variable, then the form of such metrics will be

$$ds^2=A(r)^2a(\psi)^2(dt^2-dz^2)-B(r)^2b(\psi)^2dr^2-C(r)^2c(\psi)^2d\theta^2+\epsilon F(r)^2f(\psi)^2d\psi^2$$

with $\epsilon=\pm 1$. However, we shall assume that $\epsilon=-1$ in the interpretation part of this paper. The usual $\cos^2\theta$ factor in front of $d\theta^2$ has been absorbed into the differentials. It is also possible to absorb $B(r)$ and $f(\psi)$ into their associated differentials. This will make the computations come easier. We will keep $b(\psi)$ and $c(\psi)$ distinct to study conical defects in Sec. IV.

So for our metric we will use

$$ds^2=A(r)^2a(\psi)^2(dt^2-dz^2)-b(\psi)^2dr^2-C(r)^2c(\psi)^2d\theta^2+\epsilon F(r)^2d\psi^2. \tag{1}$$

Then the five-dimensional Einstein equations which are not identically satisfied separate to give

$$\begin{aligned}
R^1{}_1=R^4{}_4=0:\quad & F^2\left(\frac{A''}{A}+\frac{A'^2}{A^2}+\frac{A'C'}{AC}+\frac{A'F'}{AF}\right)=\frac{m}{2}=\epsilon b^2\left(\frac{a''}{a}+\frac{a'2}{a^2}+\frac{a'b'}{ab}+\frac{a'c'}{ac}\right),\\
R^2{}_2=0:\quad & F^2\left(2\frac{A''}{A}+\frac{C''}{C}+\frac{F''}{F}\right)=n=\epsilon b^2\left(\frac{b''}{b}+2\frac{b'a'}{ba}+\frac{b'c'}{bc}\right),\\
R^3{}_3=0:\quad & F^2\left(\frac{C''}{C}+2\frac{C'A'}{CA}+\frac{C'F'}{CF}\right)=p=\epsilon b^2\left(\frac{c''}{c}+2\frac{c'a'}{ca}+\frac{c'b'}{cb}\right),\\
R^5{}_5=0:\quad & F^2\left(\frac{F''}{F}+2\frac{F'A'}{FA}+\frac{F'C'}{FC}\right)=q=\epsilon b^2\left(2\frac{a''}{a}+\frac{b''}{b}+\frac{c''}{c}\right),\\
R^2{}_5=0:\quad & -2\frac{A'b'}{Ab}-2\frac{F'a'}{Fa}+2\frac{A'a'}{Aa}+\frac{C'c'}{Cc}-\frac{F'c'}{Fc}-\frac{C'b'}{Cb}=0,
\end{aligned} \tag{2}$$

where m, n, p, and q are separation constants. The prime denotes the derivative taken with respect to the variable on which a function depends and to carry out full separation of the equations we have assumed that none of the functions A, C, F, a, b, or c vanishes. At any rate, these would give degenerate metrics in five dimensions. We restrict ourselves in this work to the case: $m=n=p=q=0$. We look first at the left-hand sides (the r set of equations).

Multiplying the $R^1{}_1=0$ equation by A/A' (assuming A is not a constant), leads to the relation $(A^2)'BF=c_1$ (all the c_i's will be constants). Similarly, $R^3{}_3=0$ will give $A^2B'F=c_3$, while $R^5{}_5=0$ gives $A^2BF'=c_5$. Thus $(A^2BF)'=c_1+c_3+c_5=c$ [a constant, not to be confused with $c(\psi)$] and we have

$$A^2BF=cr+c_6.$$

Using this combination in the relation defining c_1 we get $(A^2)'/A^2=c_1/(cr+c_6)$ so that after integration we have

$$A^2=c_7(cr+c_6)^{c_1/c}.$$

Similar work on the equations defining c_3 and c_5 leads to

$$C=c_8(cr+c_6)^{c_3/c},$$

$$F=c_9(cr+c_6)^{c_5/c}.$$

From this point on we will use the symbols c_i to represent what was understood up until now to be c_i/c, so that in particular we now have $c_1+c_3+c_5=1$.

Substituting these solutions in $R^2{}_2=0$ we obtain the following relation linking the constants of the r set:

$$c_1{}^2+4(c_1c_3+c_1c_5+c_3c_5)=0. \tag{3}$$

This implies in particular that c_1, c_3, and c_5 cannot all be of the same sign if $c_1\neq 0$ (which, as we shall see later on, has to be satisfied if the density is not to vanish).

Solving the ψ equations is accomplished through a similar process, the equations being similar in form. We get the functions

$$a^2=d_5(d\psi+d_4)^{d_1/d},$$

$$b=d_6(d\psi+d_4)^{d_2/d},$$

$$c=d_7(d\psi+d_4)^{d_3/d},$$

where $d=d_1+d_2+d_3$, together with the relation

$$d_1{}^2+4(d_1d_2+d_1d_3+d_2d_3)=0 \tag{4}$$

and we rename from now on the (d_i/d)'s, as we did the (c_i/c)'s above.

Finally, the equation $R^2{}_5=0$ leads to the following compatibility equation between the c_i's and the d_i's:

$$c_1d_1+2c_3d_3-2c_1d_2-2c_3d_2-2c_5d_1-2c_5d_3=0. \tag{5}$$

To sum up, when c_6 and d_4 are absorbed in the differentials, when the separation constants all vanish, and provided we take Eqs. (3)–(5) into account, the possible static separable metrics in five dimensions which are cylindrically symmetric about the z coordinate axis are

$$\begin{aligned} ds^2=&(ar\pm 1)^{c_1}(b\psi\pm 1)^{d_1}(dt^2-dz^2)-(b\psi\pm 1)^{2d_2}dr^2\\ &-(ar\pm 1)^{2c_3}(b\psi\pm 1)^{2d_3}d\theta^2+\epsilon(ar\pm 1)^{2c_5}d\psi^2, \end{aligned} \tag{6}$$

where the constants a and b are positive. We shall keep only solutions with the negative sign in the following, as they exhibit interesting singularities in four dimensions which are particular to the ''five to four formalism.''

III. THE STRESS TENSOR

In the Ponce de Leon–Wesson formalism, the stress-energy-momentum tensor induced in four dimensions is given by those terms in $G^{\alpha}{}_{\beta}$ which contain the coefficient g_{55} or derivatives with respect to ψ of the metric coefficients. In the case of a general cylindrical metric of the form

$$ds^2=B_1{}^2(dt^2-dz^2)-B_2{}^2dr^2-B_3{}^2d\theta^2-B_5{}^2d\psi^2,$$

with

$$B_i=B_i(r,\psi),$$

this separation leads to the nonvanishing stress tensor components

$$\begin{aligned} T_1{}^1=T_4{}^4=&-\frac{B_{5,rr}}{B_2{}^2B_5}+\frac{B_{2,r}B_{5,r}}{B_2{}^3B_5}-\frac{B_{3,r}B_{5,r}}{B_3B_2{}^2B_5}-\frac{B_{1,r}B_{5,r}}{B_1B_2{}^2B_5}\\ &+\epsilon\left(\frac{B_{1,\psi\psi}}{B_1B_5{}^2}+\frac{B_{2,\psi\psi}}{B_2B_5{}^2}+\frac{B_{3,\psi\psi}}{B_3B_5{}^2}-\frac{B_{3,\psi}B_{5,\psi}}{B_3B_5{}^3}-\frac{B_{1,\psi}B_{5,\psi}}{B_1B_5{}^3}-\frac{B_{2,\psi}B_{5,\psi}}{B_2B_5{}^3}\right.\\ &\left.+\frac{B_{1,\psi}B_{3,\psi}}{B_1B_3B_5{}^2}+\frac{B_{1,\psi}B_{2,\psi}}{B_1B_2B_5{}^2}+\frac{B_{2,\psi}B_{3,\psi}}{B_2B_3B_5{}^2}\right), \end{aligned}$$

$$T_2{}^2=-2\frac{B_{1,r}B_{5,r}}{B_1B_2{}^2B_5}-\frac{B_{3,r}B_{5,r}}{B_3B_2{}^2B_5}$$
$$+\epsilon\left(2\frac{B_{1,\psi\psi}}{B_1B_5{}^2}+2\frac{B_{3,\psi\psi}}{B_3B_5{}^2}+\frac{B_{1,\psi}{}^2}{B_1{}^2B_5{}^2}-2\frac{B_{1,\psi}B_{5,\psi}}{B_1B_5{}^3}-\frac{B_{3,\psi}B_{5,\psi}}{B_3B_5{}^3}+2\frac{B_{1,\psi}B_{3,\psi}}{B_1B_3B_5{}^2}\right),$$

$$T_3{}^3=-\frac{B_{5,rr}}{B_2{}^2B_5}-2\frac{B_{1,r}B_{5,r}}{B_1B_2{}^2B_5}+\frac{B_{2,r}B_{5,r}}{B_2{}^3B_5}$$
$$+\epsilon\left(2\frac{B_{1,\psi\psi}}{B_1B_5{}^2}+\frac{B_{2,\psi\psi}}{B_2B_5{}^2}+\frac{B_{1,\psi}{}^2}{B_1{}^2B_5{}^2}-2\frac{B_{1,\psi}B_{5,\psi}}{B_1B_5{}^3}-\frac{B_{2,\psi}B_{5,\psi}}{B_2B_5{}^3}+2\frac{B_{1,\psi}B_{2,\psi}}{B_1B_2B_5{}^2}\right), \quad (7)$$

$$T_5{}^5=\frac{\epsilon}{B_5{}^2}\left(\frac{B_{1,\psi}{}^2}{B_1{}^2}+\frac{B_{2,\psi}B_{3,\psi}}{B_2B_3}+2\frac{B_{1,\psi}B_{3,\psi}}{B_1B_3}-2\frac{B_{1,\psi}B_{2,\psi}}{B_1B_2}\right),$$

$$T_5{}^2=2\frac{B_{1,r\psi}}{B_1B_2{}^2}+\frac{B_{3,r\psi}}{B_2{}^2B_3}-2\frac{B_{1,r}B_{2,\psi}}{B_1B_2{}^3}-\frac{B_{3,r}B_{2,\psi}}{B_3B_2{}^3}-2\frac{B_{5,r}B_{1,\psi}}{B_1B_2{}^2B_5}-\frac{B_{5,r}B_{3,\psi}}{B_2{}^2B_3B_5}.$$

In the separable case with vanishing separation constants, this means that the stress tensor becomes, after some computation

$$T^1{}_1=\frac{c_1}{2}\,q(r,\psi),$$

$$T^2{}_2=-(c_1+c_3)q(r,\psi),$$

$$T^3{}_3=c_3q(r,\psi),$$

$$T^1{}_1=T^4{}_4,$$

where

$$q(r,\psi)=\frac{ac_5}{(ar-1)^2(b\psi-1)^{2d_2/d}}. \quad (8)$$

We see that this stress tensor vanishes at r-infinity, which is desirable. The density and pressure of whatever fluid surrounds the z-axis are given by the values of the diagonal components of $T^\alpha{}_\beta$ in a comoving system where the unit velocity of a fluid element reads (1,0,0,0).[21] Since the stress tensor is already in diagonal form, the change to such a system is readily made, and we get

$$\rho=c_1q/2,$$

$$p_r=(c_1+c_3)q,$$

$$p_\theta=-c_3q, \quad (9)$$

$$p_z=-c_1q/2.$$

Thus this stress tensor cannot be matched with the usual perfect fluid stress tensor since its spatial eigenvalues are not equal (unless they all vanish).

About the best we can do to extract an equation of state from the stress tensor is to interpret the average of its spatial components as a perfect fluid isotropic pressure. Then the tracelessness of the stress tensor gives us the equation of state of radiation or ultrarelativistic matter since

$$p_{\text{avg}}=\frac{1}{3}\,(-T^2{}_2-T^3{}_3-T^4{}_4)=\frac{c_1}{6}\,q(r,\psi)=\frac{1}{3}\,\rho, \quad (10)$$

and thus $\rho=3p_{\text{avg}}$.

Let us assume that $(b\psi-1)$ is positive and that two of the pressures are equal (they cannot all be of the same sign, as the form of the stress tensor shows). Then in order for the density ρ to be positive at large r values (the most natural choice), the combination ac_1c_5 must be positive. We first consider the case where ac_5 is positive. Then c_1 must be positive, there is tension along the z-axis in all possible cases and these are as follows:

$p_r=p_\theta$: pressure along dr and $d\theta$:

$$ds^2=(ar-1)(b\psi-1)^{(10\pm2\sqrt{3})/11}(dt^2-dz^2)-(b\psi-1)^{-4(-1\pm2\sqrt{3})/11}dr^2$$
$$-(ar-1)^{-1}(b\psi-1)^{2(-1\pm2\sqrt{3})/11}d\theta^2+\epsilon(ar-1)d\psi^2. \tag{11}$$

$p_z=p_\theta$: pressure along dr and tension along $d\theta$:

$$ds^2=(ar-1)(b\psi-1)^{2(1\mp\sqrt{3})/5}(dt^2-dz^2)-(b\psi-1)^{4/5}dr^2$$
$$-(ar-1)(b\psi-1)^{(2\pm4\sqrt{3})/5}d\theta^2+\epsilon(ar-1)^{-1}d\psi^2.$$

The solution with $p_r=p_z$ would contradict the assumption of positivity of the mass density as $c_1=-1/3$ in this case. Finally, the solution with vanishing pressures and density, namely

$$ds^2=(dt^2-dz^2)-(b\psi-1)^2dr^2-d\theta^2+\epsilon(ar-1)d\psi^2, \tag{12}$$

is common to all three cases. This can be considered to be a class of Minkowski spaces in four dimensions, its members being parametrized by ψ. In these, the c_i's and d_i's have been solved for using the pressure conditions and relations (3)–(5). We shall not cover the case where ac_5 is negative except to say that the spatial coordinates exchange their tension and pressures roles (that is, if we insist on keeping the density positive). We are mainly interested in the case where there is tension along the z coordinate and the density is positive.

Given the form of the metric, there exists a hypersurface at which the four-dimensional space becomes singular since the stress tensor is singular at $r=1/a$. Note that although this is a true singularity in four dimensions, it is not in five, since the full five-dimensional empty space–time Einstein equations are everywhere satisfied. This indicates that the validity of the space–time split between the set $(x_1,\ldots,x_4)$ and x_5 must be questioned in the neighborhood of $r=1/a$, and that whatever symmetry breaking process allows us to restrict our considerations to a $GL(4,\mathbf{R})\times G$ does not work close to this hypersurface.

IV. METRICS WITH IDENTIFIABLE CONICAL DEFECTS

We may look in another direction. The general solution can be specialized so as to put into explicit form the angular dependence on θ. This happens if the c_i's are given the values $c_1=c_5=0$ and $c_3=1$. In that case we have $d_1=1(-1/3)$, $d_2=-1/2(5/6)$, and $d_3=1/2(1/2)$ and we have either of two radiation solutions

$$ds^2=(b\psi-1)(dt^2-dz^2)-(b\psi-1)^{-1}dr^2-(ar-1)^2(b\psi-1)d\theta^2+\epsilon d\psi^2$$

or

$$ds^2=(b\psi-1)^{-1/3}(dt^2-dz^2)-(b\psi-1)^{5/3}dr^2-(ar-1)^2(b\psi-1)d\theta^2+\epsilon d\psi^2$$

and the pressure is along the r direction while the tension is along the θ direction; there is no z pressure or tension. On hypersurfaces of constant ψ such metrics can be interpreted, after proper rescaling of r, t, and z, as those of infinite strings

$$ds^2=(d\bar{t}^2-d\bar{z}^2)-d\bar{r}^2-(a\sqrt{b\psi-1}\bar{r}-1)^2(b\psi-1)d\theta^2$$

and

$$ds^2=(d\bar{t}^2-d\bar{z}^2)-d\bar{r}^2-(a(b\psi-1)^{-5/6}\bar{r}-1)^2(b\psi-1)d\theta^2$$

with conical angular defects[22,23] at large $\bar{r}$ values of $2\pi[1-(a(b\psi-1))^{-1}]$ and $2\pi[1-(b\psi-1)^{2/3}/a]$, respectively. Hence the lensing effect would be ''mass-dependent.''

The fact that we have obtained radiation only is due to our insisting that the $d\theta^2$ factor be exactly quadratic in $\bar{r}$. If we want $\rho\neq 0$ solutions, then $c_1\neq 0$ and since the equation of state is of no help in solving for the c_i's (since it is identically given by the tracelessness of the stress tensor) we can only find the relation between c_1 and c_3 [assuming c_5 has been eliminated from (3) using $c_1+c_2+c_5=1$]; what we get is that the solution pair (c_1,c_3) lies on an ellipse, in the plane of these variables, which is tangent to the $c_3=1$ line at the point (0,1), which shows that solutions with non-negligible matter density are possible with the spatial part remaining nearly flat.

V. DISCUSSION

A few final remarks are in order. First, as in other solutions based on the Wesson–Ponce de Leon ansatz which were obtained under the separability condition, no study of the stability of these solutions under perturbation has been carried out here. Nor have we tried to find other solutions from these using the generating techniques of Geroch[24] to see whether this would yield solutions which could be physically interesting.

Then, because four-dimensional solutions internal and external to the singular hypersurface are completely specified as a matching pair by the nonsingular five-dimensional solution, the matter outside the hypersurface is as exotic as the one inside. Hence in order to match such solutions to external ordinary matter, the asymptotic flatness of the four-dimensional solution has to be used in an essential way. It also shows that it would be interesting to analyze the behavior of the four- and five-dimensional geodesics close to the $r=1/a$ hypersurface to classify the 5-D ones according to their behavior in 4-D. This could possibly lead to an interpretation of the extra coordinate different from all those which have been suggested before.

Finally, we still have to show what the solutions with nonvanishing separation constants look like. At a similar level, the introduction of either the t variable (nonstatic cylindrical solutions) or the z variable (static radial solutions) in the metric coefficients could possibly remove some of the exotism from these string solutions. We intend to take a closer look at these avenues in coming publications.

[1] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965).
[2] P. S. Wesson, Gen. Relativ. Gravit. **22**, 707 (1990).
[3] P. H. Lim and P. S. Wesson, Astrophys. J. **397**, L91 (1992).
[4] D. Kalligas, P. S. Wesson, and C. W. F. Everitt, Astrophys. J. **439**, 548 (1995).
[5] P. H. Lim, J. M. Overduin, and P. S. Wesson, J. Math. Phys. **36**, 6907 (1995).
[6] D. J. McManus, J. Math. Phys. **35**, 4889 (1994).
[7] P. S. Wesson and H. Liu, Astrophys. J. **440**, 1 (1995).
[8] A. Billyard and P. S. Wesson, Fields Inst. Commun. **15**, 161 (1997).
[9] A. Billyard and P. S. Wesson, Gen. Relativ. Gravit. **28**, 129 (1996).
[10] Ø. Grøn, Astron. Astrophys. **193**, 1 (1988).
[11] P. S. Wesson and H. Liu, Class. Quantum Grav. **11**, 1341 (1994).
[12] P. S. Wesson and H. Liu, Class. Quantum Grav. **14**, 1651 (1997).
[13] H. Liu and P. S. Wesson, Phys. Lett. B **381**, 420 (1996).
[14] S. Chatterjee, Astron. Astrophys. **230**, 1 (1990).
[15] H. Liu and P. S. Wesson, J. Math. Phys. **33**, 3888 (1992).
[16] S. Rippl, C. Romero, and R. Tavakol, Class. Quantum Grav. **12**, 2411 (1995).
[17] J. M. Overduin and P. S. Wesson, Phys. Rep. **283**, 303 (1997).
[18] S. Chatterjee, P. S. Wesson, and A. P. Billyard, Nuovo Cimento B **112**, 1667 (1997).
[19] P. S. Wesson and J. Ponce de Leon, J. Math. Phys. **33**, 3883 (1992).
[20] A. Vilenkin, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, England, 1987).
[21] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[22] P. Schneider, J. Ehlers, and E. E. Falco, *Gravitational Lenses* (Springer-Verlag, Berlin, 1992).
[23] P. J. E. Peebles, *Principles of Physical Cosmology* (Princeton University Press, Princeton, NJ, 1993).
[24] R. P. Geroch, J. Math. Phys. **12**, 918 (1971).

Universality of the future chronological boundary

Steven G. Harris[a)]
Department of Mathematics, Saint Louis University, St. Louis, Missouri 63103

(Received 20 October 1997; accepted for publication 3 June 1998)

The purpose of this note is to establish, in a categorical manner, the universality of the Geroch-Kronheimer-Penrose (GKP) causal boundary when considering the types of causal structures that may profitably be put on any sort of boundary for a space–time. Actually, this can only be done for the future causal boundary (or the past causal boundary) separately; furthermore, only the chronology relation, not the causality relation, is considered, and the GKP topology is eschewed. The final result is that there is a unique map, with the proper causal properties, from the future causal boundary of a space–time onto any ''reasonable'' boundary which supports some sort of chronological structure and which purports to consist of a future completion of the space–time. Furthermore, the future causal boundary construction is categorically unique in this regard. © *1998 American Institute of Physics.* [S0022-2488(98)01910-0]

I. INTRODUCTION AND DISCUSSION

The causal boundary of Geroch, Kronheimer, and Penrose[1] (also explicated in Ref. 2, Sec. 6.8) has played a role in helping to make clear the purely causal aspects of space–times, especially at timelike or null infinity. Being defined in terms of the pasts and futures of points, it seems to be a ''natural'' object for focusing on such matters. However, it has generally been considered somewhat opaque in application and difficult to calculate. Is it still worth pursuing in spite of its somewhat formidable reputation? An affirmative answer may be forthcoming if the intuitive sense of naturalness accruing to the causal boundary can be justified in a precise and general manner. It is the purpose of this note to do exactly that: to show that the causal boundary is a ''universal object'' in the strict categorical sense, that is to say, that in an appropriate mathematical category, the causal boundary construction is the essentially unique way of completing a space–time in such a way as to yield a model to which any other causal completion process can naturally be compared (by means of a map from the causal boundary to the other boundary).

However, some caveats are in order: First of all, the categorical constructions performed here operate only on the future causal boundary (or, alternatively, only the past causal boundary), not on the melding together of the future and past boundaries that constitutes one of the more complicated issues in the full Geroch-Kronheimer-Penrose (GKP) construction. Effectively, this is the portion of the GKP construction which places an endpoint on each future-endless causal curve in a space–time, leaving past endpoints unaccounted for. Correspondingly, the categorical comparison will be made only with any ''reasonable'' future completion of a space–time—a completion of all future causal curves only. One way to understand why this is necessary is by noting that although the future causal boundary completes all future causal curves, and the past causal boundary completes all past causal curves, the combination of the two creates additional causal relations which can result in an object which is no longer complete in either future or past. (An example is given in the Appendix.)

Second, in this presentation, only the chronology relation will be addressed, not the causality relation; this is more a matter of simplification than anything else: Categorical results can be obtained by working with only the one relation, so that is the path followed here, though there is no barrier to including both relations. Accordingly, the boundary construction here is called the future chronological boundary, allowing a distinction from the GKP (future) causal boundary.

Finally, all topology is eschewed from the approach here, with notions of convergence being

[a)]Electronic mail: harrissg@slu.edu

 © 1998 American Institute of Physics

treated by purely chronological means. This allows an enormous savings of complexity, as well as the avoidance of one of the more difficult elements of the GKP causal boundary. (A topology can be inferred from the chronological notion of convergence developed herein, but this will not play an explicit role.)

As a model for the construction followed here, consider the problem of completing a Riemannian manifold. There is more than one way to do this: One could take the one-point compactification, the Cauchy completion, or some arbitrary isometric embedding into a complete metric space. Part of the issue is that one is liable to end up with a nonmanifold as one's complete object: The category of Riemannian manifolds and differentiable functions is not big enough to contain all the completions we may wish to consider. So the first thing to do is to expand to a larger category, the category **Met** of metric spaces and continuous functions. We will use Cauchy completion as our method of completing not just any Riemannian manifold, but any metric space at all. This is actually a functor $\mathbf{C}:\mathbf{Met}\to\mathbf{Met}_0$ from **Met** to the subcategory of complete metric spaces, $\mathbf{Met}_0$; this means that not only does each metric space X have a Cauchy completion $\mathbf{C}(X)$, but also that for every map $f:X\to Y$ in **Met**, there is an extension of f to the completions, $\mathbf{C}(f):\mathbf{C}(X)\to\mathbf{C}(Y)$, which works with compositions and with identity maps, i.e., $\mathbf{C}(f\circ g)=\mathbf{C}(f)\circ\mathbf{C}(g)$ and $\mathbf{C}(\mathrm{id}_X)=\mathrm{id}_{\mathbf{C}(X)}$. We need to take note not only of the completion of each metric space X but also of its embedding into its completion, $i_X:X\to\mathbf{C}(X)$. This is a natural transformation $\mathbf{i}:\mathbf{I}_{\mathbf{Met}}\dot{\to}\mathbf{U}\circ\mathbf{C}$, where $\mathbf{I}_{\mathbf{Met}}$ is the identity functor on **Met** and $\mathbf{U}:\mathbf{Met}_0\to\mathbf{Met}$ is the ''forgetful'' (i.e., inclusion) functor; this means that for any map $f:X\to Y$, the embeddings on X and Y work with the extension to f, i.e., $i_Y\circ f=\mathbf{C}(f)\circ i_X$ [also, as a natural transformation, i_X is considered as technically mapping X to $\mathbf{U}(\mathbf{C}(X))$]. The importance of the Cauchy construction for completion is that it is universal in the sense of being a left adjoint to the forgetful functor; this means that for any map from a metric space into a complete metric space, there is a unique extension of that map to the Cauchy completion: For any X in **Met**, any Y in $\mathbf{Met}_0$, and any map $f:X\to\mathbf{U}(Y)$, there is a unique map $f_0:\mathbf{C}(X)\to Y$ such that $\mathbf{U}(f_0)\circ i_X=f$ (more simply, ignoring the forgetful functor: for any $f:X\to Y$ with Y complete, there is a unique $f_0:\mathbf{C}(X)\to Y$ such that $f_0\circ i_X=f$). Since left adjoints are unique up to natural isomorphism (a natural transformation consisting of isomorphisms) (Ref. 3, Corollary IV.1), this tells us that Cauchy completion is *the* correct construction for the ''free'' completion of a metric space, in exact analogy with the free vector space on a set: It is functorial, natural, and universal. (But note that we had to expand from Riemannian manifolds to the larger setting of metric spaces in order to obtain a category in which the functoriality of the construction comes into play.)

The general idea here is to define a category of objects which includes both space–times and space–times that have boundaries attached to them, and to show that the future chronological boundary is a functorial, natural, and universal construction. The category, called chronological sets, consists, roughly speaking, of objects that are space–time-like in that they possess a relation that works like the chronology relation in a space–time. There is a great deal of freedom in choosing which properties of the space–time relation to abstract and axiomatize for our new category. The ones chosen here seem to be the minimum ones needed to carry enough structure to be able to replace topology as the means for defining convergence for appropriate sequences of points. The approach here is somewhat reminiscent of the causal spaces axiomatized in Ref. 4, and some of the language (such as ''full'' and ''chain'') from that paper is borrowed for this one; but there is a great deal of simplification achieved by concentrating solely on the chronology relation.

What is the practical impact of establishing the universality of the future chronological boundary? How does this help in understanding space–times? The answer lies in being able to relate any proposed future boundary for a space–time to what is essentially the GKP future (pre-)boundary points, in a manner respecting the chronological relations—so long as the proposed boundary includes an extension of the chronology relation of the space–time and does so in a ''reasonable'' manner (such as being past-distinguishing). Some examples at this point may help:

First an artificial example: Consider a space–time M obtained from Minkowski two-space ($\mathbb{L}^2$) by excising a timelike line segment L (such as $\{(0,t)|0\leqslant t\leqslant 1\}$). (A closely related example is considered in detail at the end of Sec. IV.) The most obvious way to put a helpful boundary on M is to embed it back into $\mathbb{L}^2$ in the obvious way, yielding L as the boundary points for M. However, this is not the only useful possibility. Let us expand our conception of M to include anything conformal to $\mathbb{L}^2-L$. Then, considering that the metric may become inherently singular

as the points of L are approached, we may wish to choose a boundary construction that respects L as a physical blockage between its left and right sides; typically, we may choose as boundary two copies of L, one for the left side and one for the right side of the slit in M, joined at the top and the bottom, i.e., a boundary of the form $\{(0^-,t)|0\leqslant t\leqslant 1\}\cup\{(0^+,t)|0\leqslant t\leqslant 1\}$, with $(0^-,0)$ and $(0^+,0)$ identified, as well as $(0^-,1)$ and $(0^+,1)$. (By ''physical blockage'' here I refer to the chronology relation in the space–time-*cum*-boundary: We have, for instance, $(-0.1,0.3)\ll(0^-,0.5)$, and $(0.1,0.3)\ll(0^+,0.5)$, but no chronology relation between $(-0.1,0.3)$ and $(0^+,0.5)$. This reflects an impossibility of passage from one side of the slit to the other.) The GKP causal boundary produces precisely this latter boundary, in addition to the standard points at future and past null- and timelike-infinity associated with Minkowski space. The future (pre-)boundary points of the GKP construction are similar, with the only difference being that there is a ''double-point'' at the top of the boundary: $(0^-,1)$ and $(0^+,1)$ are not identified but kept distinct. This last is the future chronological boundary.

What, then, does the universality of the future chronological boundary have to say in this example? It says that the future chronological boundary covers any other (past-distinguishing) boundary. For instance, the full GKP causal boundary identifies the double-point of the future chronological boundary into a single point. The first boundary mentioned, consisting of L itself, identifies all the pairs of points $(0^-,t)$ and $(0^+,t)$. Other boundaries can be imagined, such as identifying $(0^-,t)$ and $(0^+,t)$ for some but not all t. [Note that past-distinguishment in the boundary necessitates retaining the identification of $(0^-,0)$ and $(0^+,0)$.] The point is that *any* appropriate boundary construction for M, intending to add future end points for all timelike curves, can be mapped onto by the future chronological boundary, in a manner that appropriately preserves the chronological relation; that this mapping is unique; and that the entire construction is categorical in nature.

For a more realistic example, let our space–time be Schwarzschild space. The standard picture of the (future) singularity is of $\mathbb{S}^2\times\mathbb{R}^1$: a two-sphere (parametrized by ϕ and θ) at each point of $r=0$ and any $t\in\mathbb{R}$, using Schwarzschild coordinates. We also have future null-infinity ($\mathscr{I}^+=\mathbb{S}^2\times\mathbb{R}^1{:}r=\infty$, other coordinates arbitrary) and future timelike-infinity (i^+, a single point). What other reasonable options are there for future boundary points in Schwarzschild? The boundary just described is the future chronological boundary (external Schwarzschild is conformal to a standard static space–time, which makes it easy to find its causal boundary). It follows that any past-distinguishing way of putting a future boundary on Schwarzschild must be a quotient of this: We can collapse the singularity to, for instance, a projective space or lens space at each value of t, and we can make identifications in $\mathscr{I}^+$, but we cannot blow up i^+ into a sphere or anything other than a single point.

It may be helpful to outline some of the complexities encountered in this presentation.

The first step is to define a category which contains both reasonable (strongly causal) space–times and any such space–time combined with a boundary of any sort (including an extension of the chronology relation to the boundary). The next step is to define a ''future-completion'' construction which will apply to any object in this category; it should replicate the GKP future (pre-)boundary construction when applied to space–times, and should yield nothing new when applied to an object which is already ''future-complete.'' It is also intended that this future-completion construction not change anything in the original object, but just add additional points.

A problem arises in this, however: The condition used to define the chronology relation for the new (future boundary) points—the same criterion as in the GKP construction—is not necessarily compatible with the preexisting chronology relation in the original object. (Among space–times, global hyperbolicity prevents this from happening, but the idea is to provide a construction for all strongly causal space–times.) The problem is that if the same criterion is applied to the original object, then more pairs of points ought to be considered as chronologically related than was originally the case. For instance, if the object is the space–time M from the example above, then the criterion applied to M would have $(-1,-1)\ll(0.5,1)$; but there is no timelike curve between those points in M. Among other things, this incompatibility interferes with the functoriality of the future-completion construction.

The solution is to define an additional construction (''past-determination'') on these objects whose sole purpose is to add in precisely those additional chronology relations so that when the future-completion construction is applied, the result is compatible with the chronology relation in

the past-determined object. This defeats the intent of preserving the original chronology relation for the object, but there is no getting around this point. Indeed, this is precisely the tack chosen by GKP: The chronology relation in the GKP construction results in additional relations within the space–time, exactly the same ones as the past-determination construction here. The exposition here essentially separates out this portion of the GKP construction for its own categorical and universal explication.

One still has to contend with the question of whether it makes a difference if one first past-determines and then future-completes, or does it the other way around. As it turns out, these operations are categorically commutative; but that takes some doing to establish.

Finally, one needs to make a comparison between the future chronological boundary construction here—the GKP future (pre-)boundary—and the full GKP causal boundary. A difficulty arises in that the simple definition of future-complete used here is not generally satisfied by the GKP causal boundary; a slight generalization is needed, allowing, for instance, for the identification of points such as in the causal boundary of the example M above. But all goes through, as exemplified in that example.

In summation: Section II introduces the basic category and the future-completion operation; this applies, without further modification, to globally hyperbolic space–times. Section III defines the past-determination operation, extending the results of Sec. II to, for instance, all strongly causal space–times. Section IV addresses the generalization that is necessary in order to include such completions as those deriving from the full GKP construction.

A word on what this paper addresses and what it does not:

The intent here is to provide a justification for the use of the causal boundary construction. This is accomplished largely though showing how at least the future portion of that construction provides a universal standard of comparison for any other future-completing boundary, in terms of the chronology relation.

One obvious failing of this enterprise is to evince a similar standard of comparison for the full causal boundary. This, however, I believe to be innately impossible: As evidenced in the Appendix, the full causal boundary simply does not result in a complete object. This does not necessarily mean that the full causal boundary is an inappropriate tool, but it does detract from hopes of its having a universal role in boundary considerations.

Another element missing here is the causality relation. I believe it possible to do much of the same work here with the causality relation included, but it seems an unnecessary complication at this point. The chronology relation suffices to cover much of the necessary structures, and it yields a satisfactory universality result.

Topology is probably the most complex element missing from this development. It would be desireable to have a universality result which included full topological information, not just chronological. This is a matter under investigation and the subject for another paper.

II. CHRONOLOGICAL SETS AND FUTURE COMPLETION

The intent here is to abstract and axiomatize the chronology relation in a space–time: A point x chronologically precedes a point y, written $x \ll y$, if and only if there is a future-directed timelike curve from x to y. This is the relation that must be emulated for the new category.

We start with a subset of the needed properties:

A *prechronological set* is a set X together with a relation $\ll$ satisfying the following conditions:

(1) $\ll$ is transitive: $x \ll y \ll z$ implies $x \ll z$;
(2) $\ll$ is nonreflexive: for all x, *not* $(x \ll x)$;
(3) X has no isolates: for all x, $I^-(x) = \{y | y \ll x\}$, the *past* of x, or $I^+(x) = \{y | x \ll y\}$, the *future* of x, is nonempty; and
(4) $\ll$ is full: for any $x \ll y$, there is some z with $x \ll z \ll y$.

For any nonempty subset A of a prechronological set X, the *past* of A, $I^-[A]$, is $\cup_{a \in A} I^-(a)$ (the *future* of A, and all subsequent dualizations, are defined analogously). A nonempty subset P

is a *past set* if $I^-[P]=P$. [In general, ( ) will be used to indicate application of an operator to an entity to be regarded as a single-element argument, [ ] will be used to indicate application to each of the various elements making up the argument.]

1. Proposition: In a prechronological set, any nonempty set of the form $I^-[A]$ is a past set.

Proof: By transitivity, $I^-[I^-[A]]\subset I^-[A]$. Now suppose $x\ll a$ for $a\in A$; then, by fullness, there is some y with $x\ll y\ll a$. Thus $I^-[I^-[A]]=I^-[A]$. □

A past set is *decomposable* if it is the union of two proper subsets that are past sets. An indecomposable past set is often called an IP.

2. Theorem: *In a prechronological set, a past set P is indecomposable if and only if for every x and y in P, there is some $z\in P$ with $x\ll z$ and $y\ll z$.*

Proof: (This theorem is due to Geroch, Kronheimer, and Penrose; the proof here is adapted from Ref. 2, in the proof of Proposition 6.8.1.)

Suppose P is decomposable as $P_1\cup P_2$ (proper past subsets). Then neither $P_1\subset P_2$ nor $P_2\subset P_1$, so we can find $x_1\in P_1-P_2$ and $x_2\in P_2-P_1$. Then there can be no $y\in P$ with $x_1\ll y$ and $x_2\ll y$: If y is in, say, P_1, then since $x_2\ll y$, $x_2\in P_1$ also, which is false.

Suppose P is indecomposable. For any x_1 and x_2 in P, let $Q_i=\{q\in P|\,not(x_i\ll q)\}$, and let $P_i=I^-[Q_i]$ (for each i). Note that $P_i\subset Q_i$: For $x\in P_i$, there is some $q\in P$ with $x\ll q$ and $not(x_i\ll q)$; thus, $x\in P$ and $not(x_i\ll x)$ (lest $x_i\ll q$), so $x\in Q_i$. Therefore P_i cannot be all of P, since Q_i is not (since x_i is in P and $P=I^-[P]$, there must be some $p_i\in P$ with $x_i\ll p_i$). Since P_1 and P_2 are both past sets, it follows from P being an IP that $P\neq P_1\cup P_2$, i.e., there is some $p\in P$ with $p\notin I^-[Q_1]$ and $p\notin I^-[Q_2]$. Now, since $P=I^-[P]$, there is some $z\in P$ with $p\ll z$; it then follows that z must be in neither Q_i, i.e., that $x_1\ll z$ and $x_2\ll z$. □

A *future chain* in a prechronological set X is a sequence $\{x_n|n\geqslant 1\}$ of points in X such that for all n, $x_n\ll x_{n+1}$.

A subset S of a prechronological set X is called *dense* in X if for all x and y in X with $x\ll y$, there is some s in S with $x\ll s\ll y$. X is *separable* (in a chronological sense) if there is a countable dense set.

The following theorem shows how future chains in a chronological set replace future timelike curves in a space–time; this is adapted from Proposition 6.8.1 in Ref. 2, due to Geroch, Kronheimer, and Penrose, which makes essentially the same statement with ''timelike curve'' in the place of ''chain.''

3. Theorem: *A past set in a separable prechronological set is indecomposable if and only if it is the past of a future chain.*

Proof: Let P be $I^-[\{x_n|n\geqslant 1\}]$ for $\{x_n\}$ a future chain; then for any p_1 and p_2 in P, there are numbers n_1 and n_2 with $p_1\ll x_{n_1}$ and $p_2\ll x_{n_2}$. Let $k=1+\max\{n_1,n_2\}$; then x_k is in P and has each p_i in its past.

Let P be an IP. If S is the countable dense set for the prechronological set, then $P\cap S=\{p_n|n\geqslant 1\}$ is a dense set in P. Since $p_1\epsilon P$ and P is a past set, there is some $x_1\epsilon P$ with $p_1\ll x_1$. Suppose we have defined $\{x_1,\ldots,x_k\}$ with $x_i\in P$ and $x_{i-1}\ll x_i$ and $p_i\ll x_i$ for each $i\leqslant k$. Then by Theorem 2 there is some $x_{k+1}\in P$ with $x_k\ll x_{k+1}$ and $p_{k+1}\ll x_{k+1}$. This defines for us the future chain $\{x_n\}$ in P with $p_n\ll x_n$ for each n. Clearly, $I^-[\{x_n\}]\subset P$. For any $x\in P$, since P is a past set, there is some $y\in P$ with $x\ll y$; then there is some k with $x\ll p_k\ll y$, so $p_k\ll x_k$ shows us $x\ll x_k$. Thus $I^-[\{x_n\}]=P$. □

We will say that an IP P is *generated* by any future chain c for which $P=I^-[c]$.

A prechronological set is called *past-distinguishing* if for points x and y, $I^-(x)=I^-(y)$ implies $x=y$. (Unlike nearly all the other terms defined here, this one is traditional, following the usage in Ref. 2.)

For a future chain $c=\{x_n|n\geqslant 1\}$, a point x is called a *future limit* of c if $I^-(x)=I^-[c]$. Note that if x is a future limit of c, then for all n, $x_n\ll x$, since $I^-(x)$ includes $I^-(x_{n+1})$. Also, if x is a future limit of c, then it is also a future limit of every subchain of c; and, conversely, if x is a future limit of some subchain of c, then it is also a future limit of c. Clearly, in a past-distinguishing prechronological set, a future chain can have at most one future limit. [We could define a topology from this notion of convergence—a set is (future-) closed if it contains all future limits of future chains within it—but this will not be pursued here.]

A prechronological set that is separable is called a *chronological set*. A chronological set for which every future chain has a future limit is called *future-complete*.

It will prove useful to have the $\ll$ relation determined, in a certain sense, by I^-: Specifically, call a prechronological set X *past-determined* if for any x and y in X, if $I^-(x)$ is nonempty and for some $w \ll y$, $I^-(x) \subset I^-(w)$, then $x \ll y$. [We can also say that x is past-determined if $I^-(x)$ is nonempty and whenever $I^-(x) \subset I^-(w)$, $x \ll y$ for all $y \gg w$.]

4. Theorem: *A strongly causal space–time is a past-distinguishing chronological set; if it has the property that the past of x is contained in the past of y implies that x causally precedes y (for instance, any globally hyperbolic space–time), then it is past-determined. A future limit of a future chain in a space–time is the same as a topological limit.*

Proof: A space–time is clearly a prechronological set; it is separable in the chronological sense because it is separable in the topological sense. A strongly causal space–time is past-distinguishing.[2]

If $I^-(x) \subset I^-(w)$ implies $x \in J^-(w)$ (points causally preceding w), then with $w \ll y$ also, we have $x \ll y$. Globally hyperbolic space–times are examples of this: For $I^-(x) \subset I^-(w)$, consider any $z \ll x$; then since $z \in I^-(x)$, z is also in $I^-(w)$, so $z \ll w$. It follows that the space $C(z,w)$ of causal curves from z to w is compact in the C^0 topology. Let σ be a future timelike curve with $\sigma(0)=z$ and $\sigma(1)=x$; for all $t<1$, there is also a future timelike curve τ_t from $\sigma(t)$ to w. Let c_t be the concatenation of $\sigma|_{[0,t]}$ with τ_t; then the family of causal curves $\{c_t\}$ has a causal limit curve c_1 in $C(z,w)$; since c_1 must pass through x, this shows $x \in J^-(w)$.

In a strongly causal space–time, each point has a fundamental neighborhood system $\{U_n\}$ such that for each n, U_n is geodesically convex and for any point $x \in U_n$, any other point $y \in U_n$ is timelike-related to x iff it is joined to x by a timelike geodesic in U_n ($x \ll y$ but not joined by a timelike geodesic in U_n implies that the timelike geodesic connecting them must exit and re-enter U_n, which can be prevented with correct choice of U_n). Let $c=\{x_n\}$ be a future chain. First suppose that c has a topological limit x; then c eventually enters all such neighborhoods U_n of x as just described. For any $y \ll x$, we have $x \in I^+(y)$, a neighborhood of x; thus, eventually, $x_n \in I^+(y)$, i.e., $y \ll x_n$: $I^-(x) \subset I^-[c]$. If any x_m fails to be in the past of x, the same is true for all x_k with $k>m$. Consider such an x_k in a neighborhood U_n of x as above; due to the chronological relation in U_n being given by the geodesic structure, this is essentially the same causal relations as exist in Minkowski space: In particular, x_{k+1} is not in *closure*$(I^-(x))$, and all x_j for $j>k+1$ are in $I^+(x_{k+1})$, whose closure does not include x. This violates x being the limit of c; thus we must have $I^-(x)=I^-[c]$. Conversely, suppose that $I^-(x)=I^-[c]$; then we have $x_n \ll x$ for each n (since $x_n \in I^-[c]$). Using the neighborhoods U_n of x from above, we can find a past chain $\{y_i\}$ in $I^+(x)$ and a future chain $\{z_i\}$ in $I^-(x)$ with x the topological limit of both chains and the sets $\{V_i=I^-(y_i) \cap I^+(z_i)\}$ a fundamental neighborhood system for x. Then for each i, since $z_i \in I^-(x)=I^-[c]$, $x_n \gg z_i$ for sufficiently large n. Since $x_n \ll x \ll y_i$, this gives $x_n \in V_i$ for sufficiently large n, i.e., x is the topological limit of c. □

For a chronological set X, let $\hat{\partial}(X)=\{P \mid P$ is an IP in X such that P is not $I^-(x)$ for any point $x \in X\}$, the *future chronological boundary* of X. Let $\hat{X}=X \cup \hat{\partial}(X)$, the *future chronological completion* of X. Define an extension of $\ll$ on X to $\hat{X}$ by the following:

(1) among points in X, there is no change in $\ll$;
(2) for $x \in X$ and $P \in \hat{\partial}(X)$, $x \ll P$ iff $x \in P$;
(3) for $x \in X$ and $P \in \hat{\partial}(X)$, $P \ll x$ iff for some $y \in X$ with $y \ll x$, $P \subset I^-(y)$; and
(4) for $P \in \hat{\partial}(X)$ and $Q \in \hat{\partial}(X)$, $P \ll Q$ iff for some $y \in Q$, $P \subset I^-(y)$.

Where it is necessary to distinguish between the relation on X and that on $\hat{X}$, $\hat{I}^-$ will be used for the latter. [Thus $x \in \hat{I}^-(y)$ iff $x \in I^-(y)$, and so on.] Note that $\hat{I}^-(P)=P \cup \{Q \in \hat{\partial}(X) \mid Q \subset I^-(w)$ for some $w \in P\} = \hat{I}^-[P]$ (the first being the past of a single point in $\hat{X}$, the last being the past of a subset of $\hat{X}$); in particular, $\hat{I}^-(P) \cap X = P$. Also note that P is a future limit of a future chain c (in $\hat{X}$) if and only if P is generated by c ($I^-[c]=P$ iff $I^-[c]=I^-[P]$ iff $\hat{I}^-[c]=\hat{I}^-[P]$ iff $\hat{I}^-[c]=\hat{I}^-(P)$).

5. Theorem: *Let X be a chronological set. Then*

(1) *$\hat{X}$ is a chronological set;*
(2) *if X is past-distinguishing, then so is $\hat{X}$;*

(3) *if X is past-determined, then so is $\hat{X}$; and*
(4) *$\hat{X}$ is future-complete.*

Proof: (1) $\hat{X}$ is a chronological set:

This is a matter of routine checking of various cases. For instance, to show transitivity, one of the cases to be established is that for x in X and P and Q in $\hat{\partial}(X)$, if $x \ll P \ll Q$, then $x \ll Q$: We have $x \in P$ and $P \subset I^-(w)$ for some $w \in Q$; then $x \ll w$, and, since Q is a past set, $x \in Q$, i.e., $x \ll Q$.

A set which is dense in X is dense in $\hat{X}$ as well (routine to establish).

Note that $\hat{X}$ has no isolates, since for any $P \in \hat{\partial}(X)$, $\hat{I}^-(P)$ includes the elements of P, which is nonempty.

(2) If X is past-distinguishing, then $\hat{X}$ is also (x, P, and Q as above, with $y \in X$):

Suppose $\hat{I}^-(x) = \hat{I}^-(y)$. Then, in particular, $I^-(x) = I^-(y)$, so $x = y$.

Suppose $\hat{I}^-(x) = \hat{I}^-(P)$. Then, in particular, $\hat{I}^-(P) \cap X = I^-(x)$; but this means $P = I^-(x)$, which is forbidden by the definition of $\hat{\partial}(X)$.

Suppose $\hat{I}^-(P) = \hat{I}^-(Q)$. Then, in particular, $\hat{I}^-(P) \cap X = \hat{I}^-(Q) \cap X$, which just means precisely that $P = Q$.

(3) If X is past-determined, then $\hat{X}$ is also (notation as above):

First note that if $\hat{I}^-(x)$ is nonempty, then so is $I^-(x)$ ($P \ll x$ requires some $w \ll x$), and $\hat{I}^-(P)$ is always nonempty.

Suppose $\hat{I}^-(x) \subset \hat{I}^-(w)$ with $w \ll y$; then $I^-(x) \subset I^-(w)$, so $x \ll y$. Suppose $\hat{I}^-(x) \subset \hat{I}^-(Q)$ with $Q \ll y$, i.e., $Q \subset I^-(w)$ and $w \ll y$; then $I^-(x) \subset Q$, so $I^-(x) \subset I^-(w)$, so $x \ll y$.

Suppose $\hat{I}^-(x) \subset \hat{I}^-(w)$ with $w \in R$ and $\hat{I}^-(x)$ nonempty; then we can find $z \in R$ with $w \ll z$, so the previous result gives us $x \ll z$, so $x \in R$, so $x \ll R$. Suppose $\hat{I}^-(x) \subset \hat{I}^-(Q)$ with $Q \subset I^-(w)$ and $w \in R$; then pick z the same, and the previous result gives $x \ll z$, so $x \in R$, i.e., $x \ll R$.

Suppose $\hat{I}^-(P) \subset \hat{I}^-(w)$ with $w \ll y$; then $P \subset I^-(w)$, so $P \ll y$. Suppose $\hat{I}^-(P) \subset \hat{I}^-(Q)$ with $Q \subset I^-(w)$ and $w \ll y$; then $P \subset Q$, so $P \subset I^-(w)$, so $P \ll y$.

Suppose $\hat{I}^-(P) \subset \hat{I}^-(w)$ with $w \in R$; then pick $z \in R$ (i.e., $z \ll R$) with $w \ll z$, and the previous result yields $P \ll z$, so $P \ll R$. Suppose $\hat{I}^-(P) \subset \hat{I}^-(Q)$ with $Q \subset I^-(w)$ and $w \in R$ (i.e., $w \ll R$); then the previous result yields $P \ll w$, so $P \ll R$.

(4) $\hat{X}$ is future-complete:

First consider a future chain $c = \{x_n\}$ made up of points of X. Let $P = I^-[c]$; then P is an IP in X. If there is a point $x \in X$ with $I^-(x) = P$, then x is a future limit of c: Since each $x_n \ll x$, for any $\alpha \in \hat{I}^-(x_n)$, $\alpha \ll x$ also; for $y \in I^-(x)$, we have $y \in I^-[c]$; and for $Q \ll x$ with $Q \in \hat{\partial}(X)$, we have $Q \subset I^-(w)$ for some $w \ll x$, hence $w \ll x_n$ for some n, so $Q \ll x_n$. Now suppose, on the other hand, that $P \in \hat{\partial}(X)$; then P is a future limit of c, as noted previously.

Now consider a future chain c in which there is some subchain c' consisting of points in X. By the paragraph above, c' has a future limit, which is then also a future limit for c.

For a future chain consisting solely of elements of $\hat{\partial}X$, just interpolate elements of S (the countable dense subset of X) into the chain, resulting in a new future chain to which the previous paragraph applies; this has the same future limits as the original one. □

Let $f: X \to Y$ be a set map between prechronological sets. We will call f *chronological* if $x \ll y$ in X implies $f(x) \ll f(y)$ in Y; f will be called *future-continuous* if, in addition, whenever x is a future limit of a future chain c in X, $f(x)$ is a future limit of $f[c]$ (necessarily a future chain in Y). For any chronological set X, we have the future-continuous map $\hat{\iota}_X: X \to \hat{X}$ given by $\hat{\iota}_X(x) = x$; this is the *standard future injection* of X (that it is future-continuous follows directly from the definitions).

For M and N strongly causal space–times, a function $f: M \to N$ is both past- and future-continuous if and only if it is (topologically) continuous and carries timelike curves to timelike curves, preserving time-orientation. This can be seen by examining, for a point $x \in M$ with sequence $\{x_n\}$ topologically converging to x, a fundamental neighborhood system of the form described in the proof of Theorem 4, $\{V_i = I^-(y_i) \cap I^+(z_i)\}$ for $\{y_i\}$ a past chain and $\{z_i\}$ a future chain, both with x as topological limit. With f both past- and future-continuous, we have $f(x)$ is

the topological limit of the chains $\{f(y_i)\}$ and $\{f(z_i)\}$, so $\{W_i = I^-(f(y_i)) \cap I^+ f(z_i))\}$ is a fundamental neighborhood system for $f(x)$. Eventually, $x_n \in V_i$, whence $f(x_n) \in W_i$; thus f is topologically continuous. However, a merely future-continuous function need not be continuous: Consider $f:\mathbb{L}^2 \to \mathbb{L}^2$ defined by $f(x,t) = (x,t)$ for $t > 0$ and $f(x,t) = (x,t-1)$ for $t \leqslant 0$ ($\mathbb{L}^n$ denotes Minkowski n-space).

Let $f:X \to Y$ be a chronological map between chronological sets with Y past-distinguishing; we will define a map $\hat{f}:\hat{X} \to \hat{Y}$, the *future completion* of f. First we define $f:\mathscr{P}(X) \to \mathscr{P}(Y)$ (the power sets) by $f(S) = \underline{I^-}[f[S]]$. For any point $x \in X$, we just let $\hat{f}(x) = f(x)$. For an IP $P \in \hat{\partial}(X)$, first note that $f(P)$ is an IP in Y: This is because for any future chain c generating P, $f(P) = \underline{I^-}[f[c]]$. If there is some $y \in Y$ with $I^-(y) = f(P)$, then we let $\hat{f}(P) = y$; otherwise, we let $\hat{f}(P) = f(P)$. [Note that $\hat{f}$ can be defined even if Y is not past-distinguishing, so long as there is never more than a single point $y \in Y$ for which $I^-(y)$ is the same as a given $f(P)$.]

6. Proposition: Let $f:X \to Y$ be a chronological map between chronological sets with Y past-determined and past-distinguishing (or, more generally: for any future chain in X with no future limit in X, its image under f does not have more than one future limit in $\hat{Y}$); then $\hat{f}$ is also chronological. If f is future-continuous, then $\hat{f}$ is the unique future-continuous map such that $\hat{f} \circ \hat{\imath}_X = \hat{\imath}_Y \circ f$.

Proof: That $\hat{f} \circ \hat{\imath}_X = \hat{\imath}_Y \circ f$ is clear. From this and f being chronological it follows that $x \ll y$ in X implies $\hat{f}(x) \ll \hat{f}(y)$. Suppose $x \ll P$ for $x \in X$ and $P \in \hat{\partial}(X)$. Let P be generated by a future chain $c = \{x_n\}$, so that $\underline{f}[c]$ generates $f(P)$. We have $x \in P$, so $x \ll x_n$ for some n; thus $f(x) \ll f(x_n)$, so $\underline{f}(x) \in f(P)$. If $f(P) \in \hat{\partial}(Y)$, this is all we need for $\hat{f}(x) \ll \hat{f}(P)$; otherwise, for some $y \in Y$, $f(P) = I^-(y)$, so $f(x) \in I^-(y)$, or $f(x) \ll y$, so $\hat{f}(x) \ll \hat{f}(P)$. If $P \ll x$ (same notation), then $P \subset I^-(w)$ for some $w \ll x$. With the future chain c as before, we have for all n, $x_n \ll w \ll x$, so $f(x_n) \ll f(w) \ll \underline{f}(x)$. This gives us $f(P) \subset I^-(f(w))$. If $f(P) \in \hat{\partial}(Y)$, then $\hat{f}(P) \ll \hat{f}(x)$; otherwise, for some $y \in Y$, $f(P) = I^-(y)$, so $y \ll f(x)$ by Y being past-determined. Finally, suppose $P \ll Q$ for P and Q in $\hat{\partial}(X)$; then there is some $x \in X$ with $P \ll x \ll Q$, so the previous results apply.

Suppose f is future-continuous. Clearly, for a point $x \in X$ being a future limit of a future chain of points $c = \{x_n\}$ in X, the same is true for $\hat{f}(x)$ and $\{\hat{f}(x_n)\}$. For $P \in \hat{\partial}(X)$ being a future limit of the chain, since P is generated by c, $f(P)$ is generated by $\underline{f}[c]$. Thus if $f(P) \in \hat{\partial}(Y)$, then it is the future limit of $f[c]$, and we are done, since $\hat{f}(P) = f(P)$. Otherwise, for some $y \in Y$, $f(P) = I^-(y)$ and again we are done, since $\hat{f}(P) = y$ and y is a future limit of $f[c] = \hat{f}[c]$. We need concern ourselves solely with chains of points in X, for any other sort can have points of X interpolated into it with no change in future limits.

The restriction of $\hat{f}$ to X is uniquely determined by $\hat{f} \circ \hat{\imath}_X = \hat{\imath}_Y \circ f$. For $\hat{f}$ to be future-continuous, for any $P \in \hat{\partial}(X)$, since P is a future limit of any chain c generating it, we must have $\hat{f}(P)$ a future limit of $f[c]$. With Y past-distinguishing (or, more generally, with the definition of $\hat{f}$ above yielding a well-defined function), this is unique. □

Another way to express the previous result is more closely related to adjunctions:

7. Corollary: Let $f:X \to Y$ be a future-continuous function between chronological sets with Y past-determined, past-distinguishing, and future-complete (or, more generally: for any future chain in X, its image under f has exactly one future limit in Y, which is past-determined); then there is a unique future-continuous function $\hat{f}:\hat{X} \to Y$ such that $\hat{f} \circ \hat{\imath}_X = f$.

Proof: We need but note that with Y being future complete, $\hat{Y} = Y$ and $\hat{\imath}_Y$ is the identity map on Y; then Proposition 6 applies. □

Conversely, we could have started with Corollary 7 and deduced Proposition 6: Given future-continuous $f:X \to Y$ with Y past-determined and past-distinguishing, we have $\hat{\imath}_Y \circ f:X \to \hat{Y}$ fulfilling the hypotheses of Corollary 7, yielding $\hat{f}:\hat{X} \to \hat{Y}$ as the unique future-continuous map such that $\hat{f} \circ \hat{\imath}_X = \hat{\imath}_Y \circ f$.

These results allow us to define some categories, some functors, a natural transformation, and an adjunction: Let **Chron** be the category of chronological sets with morphisms all chronological functions; since the composition of chronological functions is clearly chronological, this is a category. Let **PdetPdisChron** be the subcategory of past-determined, past-distinguishing chronological sets, with morphisms all future-continuous functions; again, a composition of future-

continuous functions is future-continuous, so this is a category. Let **FcplPdetPdisChron** be the full subcategory of **PdetPdisChron** of objects which are also future-complete. Then future completion is a functor $\hat{}$**:PdetPdisChron**$\to$**FcplPdetPdisChron** (we need only check that $\widehat{g\circ f}=\hat{g}\circ\hat{f}$; this follows from the uniqueness part of Proposition 6, since $\hat{g}\circ\hat{f}$ has the requisite properties); while Theorem 5 gives us the right properties for the future completion of any chronological set, Proposition 6 gives us functoriality only for this subcategory. We also have the forgetful functor $\hat{\mathbf{U}}$:**FcplPdetPdisChron**$\to$**PdetPdisChron**. The maps $\hat{\iota}_X$ define a natural transformation $\hat{\boldsymbol{\iota}}$:**I**$\dot{\to}\hat{\mathbf{U}}\circ\hat{}$ (where **I** is the identity functor on **PdetPdisChron**). Finally, Proposition 7 gives us the universality condition on $\hat{\boldsymbol{\iota}}$ making $\hat{}$ left adjoint to $\hat{\mathbf{U}}$. From category theory we obtain

8. Theorem: *The future completion functor and the standard future injection are, up to natural isomorphism, the unique construction yielding a future-continuous map from any past-determined, past-distinguishing chronological set to a future-complete, past-determined, past-distinguishing chronological set, consistent with all future-continuous maps and yielding a unique extension for any future-continuous map into a future-complete, past-determined, past-distinguishing chronological set.* □

To restate the previous three theorems in summary form:

(1) For X in **Chron**, Y in **PdetPdisChron**, and $f:X\to Y$ future-continuous, $\hat{f}:\hat{X}\to\hat{Y}$ is the unique future-continuous map with $\hat{f}\circ\hat{\iota}_X=\hat{\iota}_Y\circ f$.
(2) For X in **Chron**, Y in **FcplPdetPdisChron**, and $f:X\to Y$ future-continuous, $\hat{f}:\hat{X}\to Y$ is the unique future-continuous map with $\hat{f}\circ\hat{\iota}_X=f$.
(3) Restricted to **PdetPdisChron**, the construction of $\hat{X}$, $\hat{f}$, and $\hat{\iota}_X$ is functorial, natural, and universal, hence unique as a way to fulfill (1) and (2) in such a manner.

III. PAST-DETERMINATION FUNCTOR

It is quite common for a space–time not to be past-determined, so there is a limited application of the functor $\hat{}$ to space–times. However, this can be remedied if we are willing to create new $\ll$ relations in the space–time, in order to make it past-determined. We will do this generally:

Consider a prechronological set X; using the same set X, let us define a new relation, $\ll^p$: For any x and y in X, set $x\ll^p y$ if $x\ll y$ or if $I^-(x)$ is nonempty and for some $w\ll y$, $I^-(x)\subset I^-(w)$; we will use I^{-p} to denote the past of a point (or set) using the $\ll^p$ relation. Clearly, $\ll^p$ is an extension of $\ll$. Note that if X is past-determined, then $\ll^p$ is identical to $\ll$.

Two properties of $\ll^p$ should be noted: First, for any x, $I^-[I^{-p}(x)]=I^-(x)$: If $y\ll z\ll^p x$, then $z\ll x$ (and so $y\ll x$) or $I^-(z)\subset I^-(w)$ for some $w\ll x$; we have $y\in I^-(z)$, so we get $y\in I^-(w)$, i.e., $y\ll w$, so $y\ll x$. On the other hand, if $y\ll x$, then by fullness there is some z with $y\ll z\ll x$; this gives us $z\ll^p x$, so $y\in I^-[I^{-p}(x)]$. [Also, $I^{-p}[I^-(x)]=I^{-p}(x)$; this is done similarly.] Second, $I^-(x)\subset I^-(y)$ if and only if $I^{-p}(x)\subset I^{-p}(y)$: Given $I^-(x)\subset I^-(y)$, if $z\ll^p x$, then we have $z\ll x$ [so $z\in I^-(x)\subset I^-(y)\subset I^{-p}(y)$] or we have $I^-(z)\neq\varnothing$ and $I^-(z)\subset I^-(w)$ for some $w\ll x$; in the latter case, we have $w\ll y$ also, so $z\ll^p y$. On the other hand, given $I^{-p}(x)\subset I^{-p}(y)$, we obtain, from the first property noted, $I^-(x)=I^-[I^{-p}(x)]\subset I^-[I^{-p}(y)]=I^-(y)$. In particular, we have $I^-(x)=I^-(y)$ if and only if $I^{-p}(x)=I^{-p}(y)$.

For a space–time, the physical significance of the $\ll^p$ relation can be viewed this way: Consider any observer as compiling an encyclopedia of all events witnessed; the space–time is to be thought of also as the set of encyclopedias, with the event p corresponding to the history of all events in $I^-(p)$. Then $p\ll^p q$ means that not only has q transcribed all the events that are written at p, but, moreover, has had a chance to think about them (i.e., an observer has traveled from some event w, which has recorded all events that are written at p, and then continued on to event q).

We will use X^p to denote X with the relation $\ll^p$; call it the *past-determination* of X.

If $c=\{x_n\}$ is a future chain in X^p, then we have for all n, $I^-(x_n)\subset I^-(w_n)$ for some sequence $\{w_n\}$ with $w_n\ll x_{n+1}$. This gives us $w_n\in I^-(x_{n+1})\subset I^-(w_{n+1})$, so $w_n\ll w_{n+1}$:$c'=\{w_n\}$ is a future chain in X. Note that $I^-[c']=I^-[c]$, due to the interweavings $I^-(x_n)\subset I^-(w_n)$ and $I^-(w_n)\subset I^-(x_{n+1})$; in general, when this interweaving obtains, we will say that c' (in X) is an *associated* chain to c (in X^p). This works well for looking at future limits:

9. Lemma: Let c' be a future chain in X associated to a future chain c in X^p. Then a point x is a future limit of c in X^p if and only if it is a future limit of c' in X.

Proof: Suppose $I^{-p}(x)=I^{-p}[c]$; then $I^-(x)=I^-[c]=I^-[c']$. Now suppose $I^-(x)=I^-[c']$; then $I^-(x)=I^-[c]$. Let $c=\{x_n\}$. For $z\ll^p x$, we have $I^-(z)$ is nonempty and $I^-(z)\subset I^-(w)$ for some $w\ll x$; since $I^-(x)=I^-[c]$, we have $w\ll x_n$ for some n, so $z\in I^{-p}[c]$. On the other hand, if $z\in I^{-p}[c]$, then $I^-(z)$ is nonempty and $I^-(z)\subset I^-(w)$ for some $w\ll x_n$ for some n; then [since $I^-[c]=I^-(x)$] $w\ll x$ also, so $z\ll^p x$. □

10. Proposition: For any prechronological set X, X^p is a past-determined prechronological set. If X is a chronological set, so is X^p; if X is past-distinguishing, so is X^p; and if X is future-complete, so is X^p.

Proof: This is largely routine. For example, to show X^p is past-determined, consider $I^{-p}(x)\subset I^{-p}(w)$ with $I^{-p}(x)$ nonempty and $w\ll^p y$; we want to show $x\ll^p y$. From $w\ll^p y$ we know that for some $v\ll y$, $I^-(w)\subset I^-(v)$ (this is true even if $w\ll y$). We have $I^-(x)\subset I^-(w)$, so $I^-(x)\subset I^-(v)$. Since $I^{-p}(x)$ is nonempty, so is $I^-(x)$. Thus $x\ll^p y$.

Suppose X is future-complete. Let $c=\{x_n\}$ be a future chain in X^p. Let x be a future limit of an associated future chain c' in X; then by Lemma 9, x is also a future limit of c in X^p. □

Past-determination works well with future-continuous functions, so long as the domain has the property that every point is a future limit of some future chain; let us call this property being *past-connected*. Clearly, any space–time is past-connected, and future completion preserves being past-connected (since an element of the future chronological boundary is always a future limit of any of its generating chains). Also, forming the past-determination of a prechronological set preserves being past-connected, in virtue of Lemma 9.

11. Proposition: If $f:X\to Y$ is a future-continuous map between prechronological sets, and X is past-connected, then the same function $f^p=f:X^p\to Y^p$ between the past-determinations is also future-continuous.

Proof: First we will show that f^p is chronological; we must show that if $x\ll^p y$, then $f(x)\ll^p f(y)$: We have $I^-(x)$ is nonempty and $I^-(x)\subset I^-(w)$ for some $w\ll y$. We have some $z\ll x$, so $f(z)\ll f(x)$: $I^-(f(x))$ is nonempty. With x a future limit of a future chain c, $f(x)$ is a future limit of $f[c]$, so $I^-(f(x))=I^-[f[c]]$. We have $c\subset I^-(w)$, so $f[c]\subset I^-(f(w))$, so $I^-[f[c]]\subset I^-(f(w))$. Therefore $I^-(f(x))\subset I^-(f(w))$ and $f(w)\ll f(y)$: $f(x)\ll^p f(y)$.

Now suppose that $c=\{x_n\}$ is a future chain in X^p with future limit x. For c' an associated chain in X, we know by Lemma 9 that x is also a future limit in X of c'; then $f(x)$ is a future limit in Y of $f[c']$. Furthermore, $f[c']$ is an associated chain of $f[c]$ (the relations defining association are preserved by a chronological function), so Lemma 9 also gives us that $f(x)$ is a future limit in Y^p of $f[c]$. □

For any prechronological set X, let us denote by $\iota^p_X:X\to X^p$ the map which, on the set level, is the identity map. This is clearly chronological, and it is also future-continuous: If, for a future chain c and point x in X, $I^-(x)=I^-[c]$, then $I^{-p}(x)=I^{-p}[c]$ also (using the second property of $\ll^p$ mentioned at the beginning of this section).

12. Corollary: Let $f:X\to Y$ be a future-continuous map between prechronological sets, with X past-connected and Y past-determined; then there is a unique future-continuous map $f^p:X^p\to Y$ such that $f^p\circ\iota^p_X=f$.

Proof: By Proposition 10, $Y^p=Y$. Then the map $f^p:X^p\to Y^p$ given by Proposition 11 fills the required condition, and uniquely does so because of what ι^p_X is on the set level. □

Proposition 11 and Corollary 12 have the same categorical properties as Proposition 6 and Corollary 7, and we derive similar categorical constructions: We have a functor **p**: **PconChron** $\to$**PdetPconChron**, where **-Pcon-** denotes the subcategory of past-connected objects. We also have a natural transformation $\boldsymbol{\iota}^{\mathbf{p}}:\mathbf{I}\dot{\to}\mathbf{U}^{\mathbf{p}}\circ\mathbf{p}$, where $\mathbf{U}^{\mathbf{p}}$: **PdetPconChron**$\to$**PconChron** is the forgetful functor and **I** is the identity functor on **PconChron**. Then Corollary 12 is precisely the statement that **p** (via $\boldsymbol{\iota}^{\mathbf{p}}$) is a left adjoint to $\mathbf{U}^{\mathbf{p}}$. (The same applies to the restriction, in all cases, to the respective subcategories of past-distinguishing objects, **-Pdis-**, which will be used for composition with the future completion functor.) Thus, similar to Theorem 8, we can conclude that the past-determination functor is the unique way to create a past-determined chronological set (at least, if one starts with all points being future limits), in a categorical, natural, and universal manner. Summarizing as before, we have

(4) For X in **PconPrechron**, Y in **Prechron**, and $f:X\to Y$ future-continuous, $f^p:X^p\to Y^p$ is the unique future-continuous map with $f^p\circ\iota^p_X=\iota^p_Y\circ f$.

(5) For X in **PconPrechron**, Y in **PdetPrechron**, and $f:X\to Y$ future-continuous, $f^p:X^p\to Y$ is the unique future-continuous map with $f^p\circ\iota_X^p=f$.
(6) Restricted to **PconChron** (or, more generally, **PconPrechron**), the construction of X^p, f^p, and ι_X^p is functorial, natural, and universal, hence the unique way to fulfill (4) and (5) in such a manner. The same is true for the further restriction to **PconPdisChron**.

We can combine these results with those of the previous section, composing the categorical items: Restricting the functor $\hat{}$ to **PdetPconPdisChron**, we have the composition $\hat{}\circ\mathbf{p}$:**PconPdisChron**$\to$**FcplPdetPconPdisChron** and the forgetful functor $\mathbf{U}^+=\hat{\mathbf{U}}\circ\mathbf{U}^{\mathbf{p}}$ in the opposite direction. For any X in **PconChron**, let $\iota_X^\uparrow=\hat{\iota}_{X^p}\circ\iota_X^p:X\to\widehat{X^p}$; then the collection of maps $\{\iota_X^\uparrow\}$ forms a natural transformation $\boldsymbol{\iota}^\uparrow:\mathbf{I}\dot{\to}\mathbf{U}^+\circ\hat{}\circ\mathbf{p}$ ($\mathbf{I}$ as above), and this yields $\hat{}\circ\mathbf{p}$ as left adjoint to $\mathbf{U}^+$. Summarizing as before, we have

(7) For X in **PconChron**, Y in **PdisChron**, and $f:X\to Y$ future-continuous, $\widehat{f^p}:\widehat{X^p}\to\widehat{Y^p}$ is the unique future-continuous map with $\widehat{f^p}\circ\iota_X^\uparrow=\iota_Y^\uparrow\circ f$.
(8) For X in **PconChron**, Y in **FcplPdetPdisChron**, and $f:X\to Y$ future-continuous, $\widehat{f^p}:\widehat{X^p}\to Y$ is the unique future-continuous map with $\widehat{f^p}\circ\iota_X^\uparrow=f$.
(9) Restricted to **PconPdisChron**, the construction of $\widehat{X^p}$, $\widehat{f^p}$, and $\iota_X^\uparrow$ is the functorial, natural, and universal, hence the unique way to fulfill (7) and (8) in such a manner.

We need to consider how past-determination interacts with future completion. First, let us compare the past of a point x of X in $\hat{X}^p$ with its past in $\widehat{X^p}$: For $y\in X$, we have $y\in\hat{I}^{-p}(x)$ if $y\ll x$ or if $\hat{I}^-(y)$ is nonempty and $\hat{I}^-(y)\subset\hat{I}^-(w)$ for some $w\ll x$; while we have $y\in\widehat{I^{-p}}(x)$ if $y\ll x$ or if $I^-(y)$ is nonempty and $I^-(y)\subset I^-(w)$ for some $w\ll x$. These are equivalent conditions: For instance, suppose $I^-(y)\subset I^-(w)$; then $P\ll y$ means $P\subset I^-(z)$ for some $z\ll y$, which then yields $z\ll w$, so $P\ll w$, also. For $Q\in\hat{\partial}(X)$, we have $Q\in\hat{I}^{-p}(x)$ if $Q\ll x$—i.e., if $Q\subset I^-(w)$ for some $w\ll x$—or if $\hat{I}^-(Q)$ is nonempty and $\hat{I}^-(Q)\subset\hat{I}^-(w)$ for some $w\ll x$; while for $Q\in\hat{\partial}(X^p)$, we have $Q\in\widehat{I^{-p}}(x)$ if $Q\subset I^{-p}(w)$ for some $w\ll^p x$. But to compare these two conditions directly is awkward, and we will instead use a pair of maps to aid in the analysis.

For a chronological set X, let us define the maps $i:\hat{X}^p\to\widehat{X^p}$ and $j:\widehat{X^p}\to\hat{X}^p$ as follows: Both maps are to be the identity on X. For $P\in\hat{\partial}(X)$, let c be any future chain in X generating P; then $i(P)=I^{-p}[c]$. For $Q\in\hat{\partial}(X^p)$, let c be a future chain in X^p generating Q; let c' be an associated chain in X, and let $j(Q)=I^-[c']$.

13. Proposition: For any chronological set X, the maps i and j above are future-continuous and give an isomorphism between $\hat{X}^p$ *and* $\widehat{X^p}$.

Proof: We will use $\hat{\ll}$ to denote the relation in $\hat{X}$, $\hat{\ll}^p$ to denote the relation in $\hat{X}^p$ and $\widehat{\ll^p}$ to denote that in $\widehat{X^p}$.

First we will treat $i:\hat{X}^p\to\widehat{X^p}$. We have that for $x\in X(=X^p)$, $i(x)=x$, and for $P\in\hat{\partial}(X)$, $i(P)=I^{-p}[c]$ for a future chain c in X generating P; we need to see that this is well-defined. Clearly, $I^{-p}[c]$ is an IP in X^p; it is in $\hat{\partial}(X^p)$ because if $I^{-p}[c]=I^{-p}(x)$, then $I^-[c]=I^-(x)$, which is impossible with $P\in\hat{\partial}(X)$. We can also write it as $I^{-p}[P]$, which shows that it is independent of the chain taken as the generator for P.

To show i is chronological, consider x and y in X, P and Q in $\hat{\partial}(X)$, and future chains $c=\{x_n\}$ and $d=\{y_n\}$ generating P and Q, respectively:

It is largely routine to show that if $x\hat{\ll}^p y$, then $x\widehat{\ll^p}y$. Since $i(x)=x$ and $i(y)=y$, we are done.

If $x\hat{\ll}^p Q$, then we have $x\hat{\ll}Q$ or we have $\hat{I}^-(x)\neq\varnothing$ and $\hat{I}^-(x)\subset\hat{I}^-(w)$ for some $w\hat{\ll}Q$ (as above, we can take $w\in X$). In the first case, we have $x\in Q=I^-[d]\subset I^{-p}[d]$. In the second case, we have $I^-(x)\neq\varnothing$ and $I^-(x)\subset I^-(w)$ and $w\ll y_n$ for some n (since $w\in Q$), so $x\ll^p y_n$, so again $x\in I^{-p}[d]$. In either case, this gives us $x\in i(Q)$, so $i(x)\widehat{\ll^p}i(Q)$, and we are done.

If $P\hat{\ll}^p y$ (or Q), then we have that $P\hat{\ll}y$ (or Q)—in which case we have $P\subset I^-(w)$ for some $w\ll y$ (or $w\in Q$)—or we have that $\hat{I}^-(P)\neq\varnothing$ and $\hat{I}^-(P)\subset\hat{I}^-(w)$ for some $w\hat{\ll}y$ (or Q)—in which case we again have $P\subset I^-(w)$ and $w\ll y$ (or $w\in Q$). Then in either case we have

$I^-[c]\subset I^-(w)$, so $i(P)=I^{-p}[c]\subset I^{-p}(w)$, with $w\ll^p i(y)$ [or $w\in i(Q)$], so $i(P)\widehat{\ll^p} i(y)$ [or $i(Q)$], and we are done.

Now we consider $j:\widehat{X^p}\to\hat{X}^p$. For $x\in X^p$, $j(x)=x$, and for $Q\in\hat{\partial}(X^p)$, $j(Q)=I^-[c']$ for c' a future chain in X associated to a chain c in X^p generating Q. Clearly $I^-[c']$ is an IP in X; it is in $\hat{\partial}(X)$ because if x is a future limit in X for c', then it is also a future limit in X^p for c, which is impossible for $Q\in\hat{\partial}(X^p)$. We can also write it as $I^-[Q]$, which shows it is independent of the chain taken as the generator for Q and of the chain in X associated to that one.

To show j is chronological, consider x and y in X, P and Q in $\hat{\partial}(X^p)$, and $c=\{x_n\}$ and $d=\{y_n\}$ future chains in X^p generating, respectively, P and Q:

It is routine to show that if $x\widehat{\ll^p}y$, then $x\hat{\ll}^p y$. Thus $j(x)\hat{\ll}^p j(y)$.

If $x\widehat{\ll^p}Q$, then $x\in Q$. To show $j(x)\hat{\ll}^p j(Q)$, we must show either that $x\hat{\ll}I^-[Q]$ or that $\hat{I}^-(x)\neq\varnothing$ and $\hat{I}^-(x)\subset\hat{I}^-(w)$ for some $w\hat{\ll}I^-[Q]$, i.e., $w\in\hat{I}^-[Q]$. If $x\in I^-[Q]$, then we have $x\hat{\ll}\hat{I}^-[Q]$. Otherwise, the only way for x to be in Q is for there to be some $q\in Q$ with $x\ll^p q$, i.e., $I^-(x)\neq\varnothing$ and $I^-(x)\subset I^-(w)$ for some $w\ll q$; then $\hat{I}^-(x)\subset\hat{I}^-(w)$, also. Since w is in $I^-[Q]$, this shows $j(x)\hat{\ll}^p j(Q)$.

If $P\widehat{\ll^p}y$ (or Q), then $P\subset I^{-p}(w)$ for some $w\ll^p y$ (or in Q). We then have $j(P)=I^-[P]\subset I^-[I^{-p}(w)]=I^-(w)$. Taking the case of y, we need only [since $I^-(w)\neq\varnothing$] look at $I^-(w)\subset I^-(z)$ for some $z\ll y$. We have $j(P)\subset I^-(z)$ [and $I^-(j(P))=I^-[P]$ contains points in P, so is nonempty], so $j(P)\hat{\ll}^p y$. For the case of Q, we have $w\in Q$, so for some $q\in Q$, $w\ll^p q$; again, we consider only $I^-(w)\subset I^-(z)$ for some $z\ll q$. We have $j(P)\subset I^-(z)$ with $z\in I^-[Q]$, i.e., $z\hat{\ll}j(Q)$; thus, $j(P)\hat{\ll}^p j(Q)$.

Now we consider the compositions: For $P\in\hat{\partial}(X)$, $j(i(P))=I^-[I^{-p}[P]]=I^-[P]=P$. Thus we have $j\circ i=1_{\hat{X}^p}$. For $Q\in\hat{\partial}(X^p)$, $i(j(Q))=I^{-p}[I^-[Q]]=I^{-p}[Q]=Q$ [using $I^{-p}[I^-(x)]=I^{-p}(x)$]. This gives us $i\circ j=1_{\widehat{X^p}}$.

Now we know that i and j provide an isomorphism of sets-with-a-relation between $\hat{X}^p$ and $\widehat{X^p}$. Since future limits are defined purely in terms of the relations, we automatically get that both maps are future-continuous. □

It is tempting to think of the maps i_X and j_X, as collections for all X in **PconPdisChron**, as forming a pair of natural transformations $\mathbf{i}:\mathbf{p}\circ\hat{}\dot{\to}\hat{}\circ\mathbf{p}$ and $\mathbf{j}:\hat{}\circ\mathbf{p}\dot{\to}\mathbf{p}\circ\hat{}$; since they are inverses of one another, we would then have a natural isomorphism between the two constructions. This would then establish that all that was done with the construction of $\widehat{X^p}$ can also be done with the construction of $\hat{X}^p$. Since this latter is the construction that precisely reproduces the GKP future causal boundary, this is a desirable outcome. However, there is a problem: One cannot compose the functors in the order $\mathbf{p}\circ\hat{}$, since Proposition 6 allows us to conclude that $\hat{f}:\hat{X}\to\hat{Y}$ is future-continuous (starting with $f:X\to Y$ future-continuous) only if Y is already past-determined. But there is a way around this difficulty, thereby establishing the categorical nature of the GKP construction:

We will use X^+ to denote $\hat{X}^p$, for X any chronological set. For a future-continuous function $f:X\to Y$ between chronological sets, we cannot in general consider $\hat{f}^p$, since $\hat{f}$ is not in general future-continuous unless Y is past-determined. But so long as Y is past-distinguishing and X is past-connected, we can use $\widehat{f^p}:\widehat{X^p}\to\widehat{Y^p}$, which is future-continuous; then we can define $f^+=j_Y\circ\widehat{f^p}\circ i_X:X^+\to Y^+$. Finally, we define $\iota_X^+=j_X\circ\iota_X^\uparrow:X\to X^+$.

14. Theorem: *For any chronological set X, X^+ is past-determined and future-complete, and $\iota_X^+:X\to X^+$ is future-continuous; if X is past-connected (or, respectively, past-distinguishing), then so is X^+. If X is past-connected and Y is a past-distinguishing chronological set, then for any future-continuous map $f:X\to Y$, $f^+:X^+\to Y^+$ is the unique future-continuous map such that $f^+\circ\iota_X^+=\iota_Y^+\circ f$. If Y is also past-determined and future-complete, than $f^+:X^+\to Y$ is the unique future-continuous map such that $f^+\circ\iota_X^+=f$.*

Proof: These claims (especially in the first sentence) are largely automatic consequences of previous results. However, it is worth some effort to explore in finer detail how the maps are related, such as how to express $\iota_X^\uparrow$ and ι_X^+ (defined by compositions through X^p) as compositions through $\hat{X}$. To do this we need to establish that several diagrams involving the injections ι commute. Some of these follow from categorical principles when in the category **PconChron**, but

others involve the particularities of these maps (especially when past-connectedness is not assumed).

One noncategorical result is the application of future-completion using the full strength of Proposition 6: The purely categorical use of future completion applies only to the category of past-determined, past-distinguishing chronological sets [as per summary statement (3)], but we need to apply it to the map $\iota^p_X : X \to X^p$, where it would be pointless to assume X past-determined. But Proposition 6 gives us the result anyway, as displayed in Diagram 1:

$$\begin{array}{ccc} X & \xrightarrow{\iota^p_X} & X^p \\ \downarrow{\scriptstyle \hat{\iota}_X} & & \downarrow{\scriptstyle \hat{\iota}_{X^p}} \\ \hat{X} & \xrightarrow{\widehat{\iota^p_X}} & \widehat{X^p} \end{array}$$

Diagram 1. Composition is $\iota^{\uparrow}_X$.

The map ι^p_X is always future-continuous, as is the standard future injection $\hat{\iota}$, whether for X or for X^p. By Proposition 6, we can define future completion of ι^p_X, and have it be future-continuous, even without X^p being past-distinguishing, so long as for any future chain c in X generating an element of $\hat{\partial}(X)$, $\iota^p_X[c]$ has no more than one future-limit in $\widehat{X^p}$; but this is manifestly the case, since $\hat{\partial}(X^p)$ (or $\hat{\partial}$ in general) is defined in such a way that no two elements can have identical pasts. Proposition 6 also yields that Diagram 1 commutes, and, as indicated, the composition is $\iota^{\uparrow}_X$.

For X past-connected, an entirely categorical result comes from starting with $\hat{\iota}_X : X \to \hat{X}$ and applying the past-determination functor and the natural transformation ι^P, yielding the commutative diagram of future-continuous maps shown in Diagram 2.

$$\begin{array}{ccc} X & \xrightarrow{\iota^p_X} & X^p \\ \downarrow{\scriptstyle \hat{\iota}_X} & & \downarrow{\scriptstyle (\hat{\iota}_X)^p} \\ \hat{X} & \xrightarrow{\iota^p_{\hat{X}}} & \hat{X}^p \end{array}$$

Diagram 2. Composition is ι^+_X (see Diagram 4).

Even if X is not past-connected, Diagram 2 still commutes (since each composition is just inclusion), and $(\hat{\iota}_X)^p$ is still future-continuous: For chronologicality, we need to show that if $x \ll^p y$ (in X), then $x \hat{\ll}^p y$ (in $\hat{X}$): If $I^-(x) \subset I^-(w)$ for $w \ll y$, then $\hat{I}^-(x) \subset \hat{I}^-(w)$ and $w \hat{\ll} y$. For future-continuity, we need to show that if x is a future limit of a future chain c in X^p, then x is also a future limit of c in $\hat{X}^p$: If x is a future limit of c in X^p, then by Lemma 9 x is a future limit of an associated future chain c' in X; application of the future-continuous $\hat{\iota}_X$ shows that x is a future limit of c' also in $\hat{X}$; then, applying Lemma 9 once more (with c' being associated to c in $\hat{X}$ as well as in X), we have x a future limit of c in $\hat{X}^p$.

The identification of the composition in Diagram 2 as ι^+_X will follow from Diagram 4. But before that we must consider Diagram 3: This commutes because both $i_X \circ (\hat{\iota}_X)^p$ and $\hat{\iota}_{X^p}$ are just inclusion on the set-level. Commuting in the other direction, with j_X, then follows automatically.

$$\begin{array}{ccc} X^p & = & X^p \\ \downarrow{\scriptstyle (\hat{\iota}_X)^p} & & \downarrow{\scriptstyle \hat{\iota}_{X^p}} \\ \hat{X}^p & \underset{j_X}{\overset{i_X}{\rightleftarrows}} & \widehat{X^p} \end{array}$$

Diagram 3.

The final result involving commuting injections is shown in Diagram 4. To establish the commutativity, we first combine Diagrams 2 and 3 into Diagram 5 and compare with Diagram 1; we use the uniqueness property of future completion from Proposition 6 [summary statement (1)] to see that that the bottom lines of Diagrams 1 and 5 must be the same, and that is the content of Diagram 4. (As before, establishing the result with i_X automatically yields the result for j_X.) Combining this with Diagram 1 and the definition of ι_X^+ yields the identification alluded to in Diagram 2.

$$\begin{array}{ccc} \hat{X} & = & \hat{X} \\ \downarrow \iota^p_{\hat{X}} & & \downarrow \widehat{\iota^p_X} \\ \hat{X}^p & \underset{j_X}{\overset{i_X}{\rightleftarrows}} & \widehat{X^p} \end{array}$$

Diagram 4.

$$\begin{array}{ccccc} X & \xrightarrow{\iota^p_X} & X^p & = & X^p \\ \downarrow \hat{\iota}_X & & \downarrow (\hat{\iota}_X)^p & & \downarrow \hat{\iota}_{X^p} \\ \hat{X} & \xrightarrow{\iota^p_{\hat{X}}} & \hat{X}^p & \xrightarrow{i_X} & \widehat{X^p} \end{array}$$

Diagram 5. Diagrams 2 and 3 combined.

Finally, in Diagram 6 we consider the diagram derived from future-continuous $f:X\to Y$ for X past-connected and Y past-distinguishing. The top two portions of Diagram 6 commute due to previously established results [summary statements (1) and (4)]. The bottom portion commutes, as that is the definition of f^+. The overall structure of Diagram 6 can be expressed as Diagram 7.

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ \downarrow \iota^p_X & & \downarrow \iota^p_Y \\ X^p & \xrightarrow{f^p} & Y^p \\ \downarrow \hat{\iota}_{X^p} & & \downarrow \hat{\iota}_{Y^p} \\ \widehat{X^p} & \xrightarrow{\widehat{f^p}} & \widehat{Y^p} \\ \cong \uparrow i_X & & \cong \downarrow j_Y \\ \hat{X}^p = X^+ & \xrightarrow{f^+} & Y^+ = \hat{Y}^p \end{array}$$

Diagram 6.

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ \downarrow \iota^+_X & & \downarrow \iota^+_Y \\ X^+ & \xrightarrow{f^+} & Y^+ \end{array}$$

Diagram 7. Overall structure of Diagram 6.

The map f^+ is unique for fulfilling its role in Diagram 7, since for any other future-continuous map $g:X^+\to Y^+$ making Diagram 7 commute, $i_Y\circ g\circ j_X$ would fill the role of $\widehat{f^p}$ in Diagram 6, and the uniqueness portion of summary statement (4) would show $i_Y\circ g\circ j_X=\widehat{f^p}$, so $g=f^+$.

If Y is also past-determined and future-complete, then $Y^+=Y$ and $\iota_Y^+=1_Y$, yielding the last statement in the theorem. □

Theorem 14 enables us to define the functor $+$: **PconPdisChron**$\to$ **FcplPdetPconPdisChron**, with $X^+=\hat{X}^p$ and $f^+:X^+\to Y^+$, even though we cannot define f^+ as $\hat{f}^p$. We also have the natural transformation $\boldsymbol{\iota}^+:\mathbf{I}\dot{\to}\mathbf{U}^+\circ+$ (where **I** denotes the identity on **PconPdisChron**) yielding $+$ as left adjoint to $\mathbf{U}^+$. Since a functor can have only one left adjoint, up to natural isomorphism, there must be a natural isomorphism, then, between $+$ and $\hat{}\circ\mathbf{p}$. This is provided, in fact, by $\mathbf{i}:+\dot{\to}\hat{}\circ\mathbf{p}$ and $\mathbf{j}:\hat{}\circ\mathbf{p}\dot{\to}+$, whose naturality is a triviality.

Expressing this in summary form as before, we have

(10) For X in **PconChron**, Y in **PdisChron**, and $f:X\to Y$ future-continuous, $f^+:X^+\to Y^+$ is the unique future-continuous map with $f^+\circ\iota_X^+=\iota_Y^+\circ f$.

(11) For X in **PconChron**, Y in **FcplPdetPdisChron**, and $f:X\to Y$ future-continuous, $f^+:X^+\to Y$ is the unique future-continuous map with $f^+\circ\iota_X^+=f$.

(12) Restricted to **PconPdisChron**, the construction of X^+, f^+, and ι_X^+ is functorial, natural, and universal, hence the unique way to fulfill (10) and (11).

For a space–time M, $\hat{M}^p$ is precisely the addition to M of the GKP future causal boundary: $\hat{M}$ is identified with the IPs of M [the TIPs—IPs not of the form $I^-(x)$ for any point x—being $\hat{\partial}(M)$], and the $\ll^p$ relation is exactly that as specified by GKP. Since any strongly causal space–time M is in **PconPdisChron**, any ''reasonable'' way of future-completing M—an embedding of M into a past-determined, past-distinguishing, future-complete object—must contain the GKP future boundary in the sense of Theorem 14: there is a unique future-continuous extension of the embedding to the GKP future completion of M.

Typically, one applies these ideas by starting with a strongly causal space–time M and a purported future completion of it, such as a topological embedding of M into another space–time N, $j:M\to N$, which preserves the chronological relations of M and which includes future end points for all the endless timelike curves in M. The actual object which is the purported future completion of M is the closure of $j(M)$ in N; call this object Y, with a chronology relation inherited from N. Then, since M is past-connected and Y (typically) is past-distinguishing, we have, from Theorem 14, $j^+:M^+\to Y^+$; assuming Y is itself future-complete, we have $Y^+=Y^p$. Then, in particular, j^+ maps $\hat{\partial}(M)$ to the boundary of $j[M]$ in N. If we had the good sense to choose the enveloping space–time N in such a way that every point in the closure of $j[M]$ was actually a future end point of a timelike curve in M (mediated by j)—or if we restrict Y to be just such points—then Y is mapped onto by j^+, with the boundary-points of Y being covered by $j^+[\hat{\partial}(M)]$. This places strong restrictions on just what Y can be.

As an example, consider a space–time M which is conformal to $K\times\mathbb{L}^1$, where K is any compact Riemannian manifold (recall $\mathbb{L}^n$ denotes Minkowski n-space). M is globally hyperbolic (any $K\times\{t\}$ is a Cauchy surface), hence past-determined. The future chronological boundary of M consists of a single point i^+ which is the future limit of every endless timelike curve. (This can be seen by observing that for any point $p\in M$ and any future-endless timelike curve c, c eventually enters the future of p; this is best noted by looking at projections into the factors K and $\mathbb{L}^1$.) Thus for any past-distinguishing, past-determined, future-complete chronological set Y and any future-continuous map $f:M\to Y$, f has a unique future-continuous extension to $M\cup\{i^+\}$; this means precisely that there is some point $y^+\in Y$ such that for every endless timelike curve c in M, y^+ is the future limit of $f\circ c$. In other words, any ''reasonable'' way of putting a future boundary on M essentially replicates the GKP procedure, producing i^+ and nothing more (any other points in Y which are neither in $f[M]$ nor y^+ are not the future limit of anything in $f[M]$ and are in that sense disconnected from M).

More generally, for M any strongly causal space–time, let $f:M\to Y$ be any future-continuous map into a past-distinguishing, future-complete, chronological set. Applying the past-determination functor we have $f^p:M^p\to Y^p$, a future continuous map into a past-determined, past-distinguishing, future-complete chronological set. Thus we have the future extension of f^p to $f^+:M^+\to Y^p$. Since $\hat{\partial}(M)$ (the GKP future causal boundary of M) contains the future limit of

every future-endless timelike curve in M, the same is true of its image in Y^p under f^+; assuming that f is an embedding of M into Y, this means that the portion of Y which is directly connected to $f[M]$—i.e., the points of Y which are future limits of chains coming from M—consists, aside from $f[M]$ itself, of a set future-continuously mapped onto by the GKP future causal boundary of M.

IV. GENERALIZED CONDITIONS AND RESULTS

In a space–time, for every point x, $I^-(x)$ is an IP; this is not necessarily so in a chronological set X. A future chain c may not have a future limit because the point x which one might nominate for the future limit of c, has a decomposable past. However, this can still play the role of a future limit in a more generalized sense. Let us order by inclusion the IPs contained within $I^-(x)$ (for any $x \in X$); if P is maximal in this ordering, call it a *past component of* x [so if $I^-(x)$ is an IP, there is only that one past component]. Let us call x a *generalized future limit* of a future chain c if $I^-[c]$ is a past component of x. As before, if x is a generalized future limit of c, then it is a generalized future limit of any subchain of c; conversely, if it is a generalized future limit of some subchain of c, then it is a generalized future limit of c. Call X *generalized past-distinguishing* if whenever x and y in X share a past component, $x=y$ (this is a stronger condition than being past-distinguishing). If X is generalized past-distinguishing, a future chain can have no more than one generalized future limit. Call X *generalized future-complete* if every future chain has a generalized future limit (a weaker condition than being future-complete). Call a chronological function $f:X\to Y$ *generalized future-continuous* if for every generalized future limit x of a future chain c in X, $f(x)$ is a generalized future limit of $f[c]$ (this is neither a stronger nor a weaker property than being future-continuous); a composition of generalized future-continuous functions is generalized future-continuous. Finally, call a point $x \in X$ *generalized past-determined* if $I^-(x)$ is nonempty and for any w with $I^-(w)$ containing some past component of x, $x \ll y$ for all $y \gg w$ (a stronger property than being past-determined); and X is generalized past-determined if all of its points are.

These generalized notions are what are needed to provide the mappings into chronological sets with decomposable pasts for some of the points; a typical example would be starting with a space–time and embedding it into another manifold to create a boundary for the space–time, with some of the boundary points possibly having decomposable pasts. We still get useful information, such as analogs of Proposition 7, Corollary 12, and Theorem 14—allowing a comparison with the GKP future causal boundary—so long as the envelopment obeys these generalized notions:

15. Theorem: *For any generalized future-continuous map $f:X\to Y$ with X a past-connected chronological set and Y a generalized past-determined, generalized past-distinguishing, generalized future-complete chronological set, there are unique generalized future-continuous maps $\hat{f}:\hat{X}\to Y$, $f^p:X^p\to Y$, and $f^+:X^+\to Y$ such that $\hat{f}\circ\hat{\iota}_X = f^p\circ\iota_X^p = f^+\circ\iota_X^+ = f$.*

Proof: First consider $\hat{f}$: For any $P \in \hat{\partial}(X)$, with c a future chain generating P, $\hat{f}[c]=f[c]$ must have a unique generalized future limit $y \in Y$, so $\hat{f}(P)=y$ is forced. For $x \ll P$, with P generated by c, we have $x \in P$, so $\hat{f}(x)=f(x)\in f[P]=f[I^-[c]]\subset I^-[f[c]]$, which is a past component of $\hat{f}(P)$; that implies $I^-[f[c]]\subset I^-(\hat{f}(P))$, so $\hat{f}(x)\ll\hat{f}(P)$. For $P\ll x$, with P generated by c, we have $I^-[c]\subset I^-(w)$ for some $w\ll x$, which implies $c\subset I^-(w)$ (as in the proof of Proposition 6); therefore, $I^-[f[c]]\subset I^-(f(w))$ and $f(w)\ll f(x)$. With $\hat{f}(P)$ the generalized future limit of $f[c]$, we know that $I^-[f[c]]$ is a past component of $\hat{f}(P)$; thus, since $\hat{f}(P)$ is generalized past-determined, we have $\hat{f}(P)\ll f(x)=\hat{f}(x)$. For $P\ll Q$, we interpolate $P\ll x\ll Q$. Thus $\hat{f}$ is chronological. Generalized future-continuity follows from that of f, since the only new future limits are those in $\hat{\partial}(X)$.

Next consider f^p: For $x\ll^p y$ in X, we have $I^-(x)\subset I^-(w)$ for some $w\ll y$. We also have, since X is past-connected, that x is a future limit of some future chain c, so $I^-[c]\subset I^-(w)$. Thus $I^-[f[c]]\subset I^-(f(w))$ and $f(w)\ll f(y)$. Since f is generalized future-continuous, $f(x)$ is the generalized future limit of $f[c]$, i.e., $I^-[f[c]]$ is a past component of $f(x)$. Since $f(x)$ is generalized past-determined, this gives us $f(x)\ll f(y)$. Therefore f^p is chronological. For generalized future-continuity, the proof in Proposition 11 suffices, once we obtain an extension of Lemma 9, that x is a generalized future limit of a future chain c in X^p if and only if x is a generalized future limit of any associated chain c' in X:

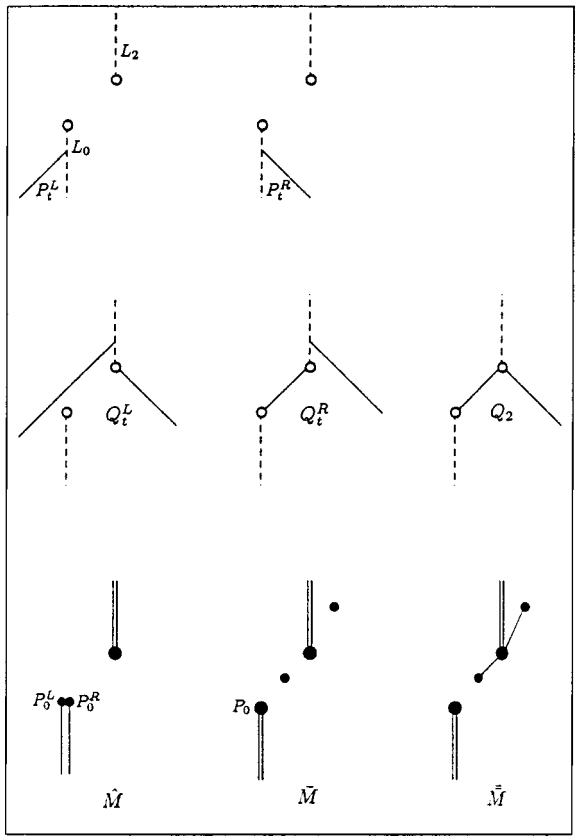

FIG. 1. IPs in M and three embeddings for it.

With $I^-[c']$ a past component of x in X, we need to have $I^{-p}[c]$ a past component of x in X^p (the other way being easy). The proof of Lemma 9 shows that $I^{-p}[c]\subset I^{-p}(x)$, we just need to show it a maximal IP with that property. Suppose d is a chain in X^p with $I^{-p}[c]\subset I^{-p}[d]\subset I^{-p}(x)$; then $I^-[c']\subset I^-[d']\subset I^-(x)$, where d' is a chain in X associated to d. Since $I^-[c']$ is maximal with the property of being an IP contained in the past of x, we have $I^-[c']=I^-[d']$, and it follows that $I^{-p}[c]=I^{-p}[d]$.

Finally, f^+ comes from the previous two constructions, just as in Theorem 14. □

As an example, let us consider the space–time M formed from $\mathbb{L}^2$ by deleting two timelike half-lines: $L_0=\{(0,t)|t\leqslant 0\}$ and $L_2=\{(2,t)|t\geqslant 2\}$. Then $\hat{\partial}(M)$ consists of these IPs (illustrated in Fig. 1): for each $t\leqslant 0$, P_t^L, bounded by L_0 and the null line from $(0,t)$ going to the left in the past, and P_t^R, bounded by L_0 and the null line from $(0,t)$ going to the right in the past; for each $t>2$, Q_t^L, bounded by L_2, the null line from $(2,t)$ going to the left in the past, and the null line from (2,2) going to the right in the past, and Q_t^R, bounded by L_2, the null line from $(2,t)$ going to the right in the past, the null line going from (2,2) to (0,0), and L_0; Q_2, bounded by the null line from (2,2) going to the right in the past, the null line going from (2,2) to (0,0), and L_0; and the usual future causal boundary for $\mathbb{L}^2$, consisting of the pasts of null lines going out to future infinity, as well as the entire space–time. In $\hat{M}$, we have $P_0^R\ll(3,4)$ [because $\hat{I}^-(P_0^R)\subset I^-(3,3)$], but there is no relation between P_0^L and (3,4). Also, since there is no relation between (1,1) and (3,4), the space is not past-determined (same for M). The past-determination of $\hat{M}$ adds some relations: $(s,s)\ll^p(3,4)$ for $0<s<2$; however, even in $\hat{M}^p$, there is no relation between P_0^L [or (x,t) for $x<0$ and $t<x$] and (3,4).

A typical embedding of M to create a boundary for it would be to use $\hat{M}$, except to coalesce P_0^R and P_0^L into a point P_0, with P_0 not related to (3,4); call this $\bar{M}$. This is not past-determined, since (1,1) is not related to (3,4). We can add relations to $\bar{M}$ by saying that two points are related if there is a curve between them, causal in $\mathbb{L}^2$, somewhere timelike, allowed to include ''one side'' of L_0 or L_2 but not to cross either of them (the points P and Q being identified with one side or

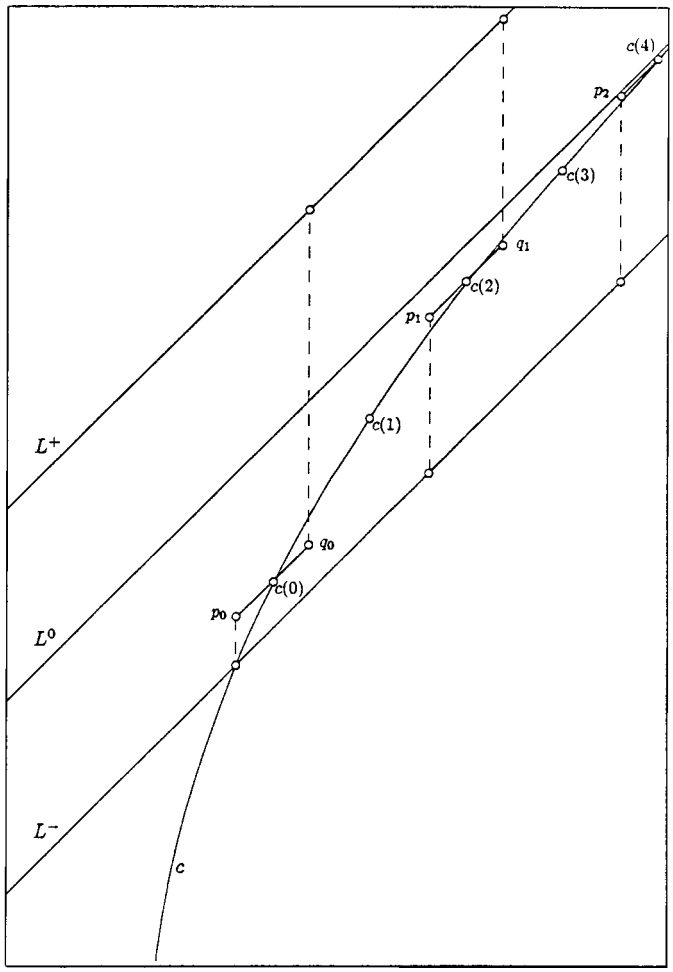

FIG. 2. Space–time with incomplete full causal boundary.

the other of those two lines). This space, $\bar{\bar{M}}$, is past-determined, past-distinguishing, and generalized future-complete (since P_0 has two past components, it is not future-complete); hence we get a generalized future-continuous map from M^+ to $\bar{\bar{M}}$; this is just the map coalescing P_0^R and P_0^L into P_0. Similarly, any means of adding complete boundaries to the slits in M will end up being a quotient of the boundaries from $\hat{M}$.

ACKNOWLEDGMENTS

This research was supported in part by the Department of Physics and Theoretical Physics of the Australian National University, and in part by NSF Grant No. DMS-9310477 to Saint Louis University.

APPENDIX: AN EXAMPLE WITH THE FULL CAUSAL BOUNDARY

Why has the foregoing structure been involved solely with the future chronological boundary and not with some combination of future and past, such as the complete causal boundary of the GKP construction? In a nutshell, the reason is that the combined causal boundary, future with past, is, in general, neither future- nor past-complete: It is as if one were to attempt a completion of a Riemannian manifold by adding end points to all finite-length endless curves, and then one found, for instance, that the added end points themselves constituted a finite-length endless curve. Here is an example of how that can happen with the GKP causal boundary and a rather ordinary space–time (illustrated in Fig. 2):

Our space–time M will be a subset of Minkowski two-space, $\mathbb{L}^2$, in which we will cut an infinite number of slits and poke an infinite collection of holes. First we select three parallel null lines as reference objects, say $L^+=\{t=x+1\}$, $L^0=\{t=x\}$, and $L^-=\{t=x-1\}$; we also select a future-timelike curve c asymptotic (in the future) to the middle null line, say $c(s)=(\sinh s,\cosh s)$. For each positive integer k, consider a short null segment centered at $c(2k)$, parallel to the reference lines, say from $(\sinh(2k)-0.5,\cosh(2k)-0.5)$ to $(\sinh(2k)+0.5,\cosh(2k)+0.5)$, and denote the past and future end points of these null segments by, respectively, p_k and q_k. To form M, delete from $\mathbb{L}^2$ each vertical segment from p_k down to L^-, each vertical segment from q_k up to L^+, and each point $c(n)$, all k and all n.

The future causal boundary for M includes an IP P_n corresponding to each of the deleted points $c(n)$, while the past causal boundary includes an IF F_n for the same. (The future causal boundary also has, for each of the slits, IPs for each point of the left and right sides of the slit, except a single IP for the bottom point; an IP for future timelike infinity, i^+; and IPs for future null infinity, $\mathscr{I}^+$—but with points corresponding to the null lines between L^- and L^+ missing from the usual $\mathscr{I}^+$ for $\mathbb{L}^2$, due to the slits. The past causal boundary also has, for each slit, IFs for each point of the left and right sides, except a single IF for the top; and past timelike and null infinity, i^- and $\mathscr{I}^-$. However, we shall be largely unconcerned with these other points in the causal boundary.) We have each $P_{2k}\subset P_{2k+1}$, so that within $\hat{M}$, $P_{2k}\ll P_{2k+1}$, but no other relations obtain among these boundary points, due to the slits; and the future chronological completion of M is precisely the same as adding the GKP future causal boundary to M. Similarly, within $\check{M}$, the past chronological completion of M (the same as adding the GKP past causal boundary), we have $F_{2k-1}\ll F_{2k}$ (since $F_{2k-1}\supset F_{2k}$) but no other relations among those particular boundary points.

There is more than one scheme for combining future and past causal boundaries into a single causal boundary (see, for instance, Refs. 5 and 6 for a good alternative to the GKP prescription), but they all agree in a space–time as simple as this one: Each P_n is identified with F_n, which we will call z_n (and the two copies of the left side of each slit—the one from the future causal boundary and the one from the past causal boundary—are identified, and the same with the right sides, with a single boundary point for the top and one for the bottom of each slit); this is called M^* in Ref. 2. Thus in any choice of chronology relation on M^* which makes the obvious mappings from $\hat{M}$ and $\check{M}$ to M^* chronological, we must have $z_{2k-1}\ll z_{2k}$ (from $\check{M}$) and $z_{2k}\ll z_{2k+1}$ (from $\hat{M}$), so that $\{z_n\}$ is a future chain in M^*. However, this chain has no future limit in M^*—for instance, the past of L^0 is not an IP, since the slits decompose that past into an infinite number of past sets: In essence, the chain $\{z_n\}$ ''escapes'' our intended completion of M by aiming for the ''hole'' in $\mathscr{I}^+$ left by the missing points corresponding to the null lines between L^- and L^+.

Szabados, in Sec. 6 of Ref. 5, proposes a chronology relation on M^* that, in this case, would leave no relations at all among the $\{z_n\}$; this construction has the advantage of not adding any additional relations among the points of M—unlike the GKP construction in the future or past causal boundaries, or the future and past completions defined here—but the disadvantage of not preserving relations already defined in $\hat{M}$ or $\check{M}$.

[1] R. P. Geroch, E. H. Kronheimer, and R. Penrose, ''Ideal points in space-time,'' Proc. R. Soc. London, Ser. A **327**, 545–567 (1972).
[2] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge, England 1973).
[3] S. Mac Lane, *Categories for the Working Mathematician* (Springer-Verlag, New York, 1971).
[4] E. H. Kronheimer and R. Penrose, ''On the structure of causal spaces,'' Proc. Cambridge Philos. Soc. **63**, 481–501 (1967).
[5] L. B. Szabados, ''Causal boundary for strongly causal spacetimes,'' Class. Quantum Grav. **5**, 121–134 (1988).
[6] L. B. Szabados, ''Causal boundary for strongly causal spacetimes: II,'' Class. Quantum Grav. **6**, 77–91 (1989).

Symplectic gravity models in four, three and two dimensions

Oleg Kechkin[a] and Maria Yurova[b]
Institute of Nuclear Physics, Moscow State University, Moscow 119899, Russia

(Received 1 May 1997; accepted for publication 4 February 1998)

A class of the $D=4$ gravity models describing a coupled system of n Abelian vector fields and the symmetric $n\times n$ matrix generalizations of the dilaton and Kalb–Ramond fields is considered. It is shown that the Peccei–Quinn axion matrix can be entered and the resulting equations of motion possess the $\mathrm{Sp}(2n,R)$ symmetry in four dimensions. The stationary case is studied. It is established that the theory allows a σ-model representation with a target space which is invariant under the $\mathrm{Sp}[2(n+1),R]$ group of isometry transformations. The chiral matrix of the coset $\mathrm{Sp}[2(n+1),R]/\mathrm{U}(n+1)$ is constructed. A Kähler formalism based on the use of the Ernst $(n+1)\times(n+1)$ complex symmetric matrix is developed. The stationary axisymmetric case is considered. The Belinsky–Zakharov chiral matrix depending on the original field variables is obtained. The Kramer–Neugebauer transformation, which algebraically maps the original variables into the target space ones, is presented. © *1998 American Institute of Physics.*
[S0022-2488(98)04310-2]

I. INTRODUCTION

In recent years much attention has been paid to the study of symmetries for gravity models arising as the low energy limit of superstring theory.[1–7] This activity was motivated by the problems of supersting theory, as well as by the remarkable results obtained for the Einstein and Einstein–Maxwell theories in three and two dimensions. Namely, it established the chiral matrix formulations for the latter theories in the stationary case, and their complete integrability in the stationary axisymmetric case.[8–21]

Recently it was shown that the simplest string gravity model, the Einstein–Maxwell–Dilaton–Axion (EMDA) system, can be represented as the 2×2 matrix generalization of the pure Einstein theory.[22,23] It means that one can exchange the complex symmetric 2×2 matrix, describing the EMDA theory in three dimensions, by the Ernst potential of the Einstein theory to obtain this last one from the former system.[22] Thus the mentioned matrix (the Ernst matrix potential) generalizes the Ernst potential to the EMDA case, and both the Einstein and EMDA theories belong to the same class of gravity models. In Ref. 23 it was shown that the established analogy is the more complete one, and can be naturally prolonged to the stationary axisymmetric case. It had been shown that the EMDA system, as the Einstein theory, allows both the target space representation and the formulation based on the use of the original field components. These two formulations are differentially related in three dimensions and can be algebraically mapped one into another in two dimensions.[23] This map is known as the Kramer–Neugebauer transformation for the stationary axisymmetric Einstein equations;[24] its existence is an important property of the theory.[25]

The formal properties of the EMDA theory do not actually depend on the Ernst matrix potential dimension and it is possible to consider a series of similar gravity models in three and two dimensions. All these models exist as the chiral ones and possess the remarkable features of the Einstein theory. But first of all one must answer the question: what kind of four-dimensional theories when reduced to three and two dimensions lead to such models?

In this paper we present such a class of four-dimensional theories. Any theory generalizes the EMDA system by the following way: here the dilaton and Kalb–Ramond fields become the n

[a]Electronic mail: kechkin@monet.npi.msu.su
[b]Electronic mail: yurova@depni.npi.msu.su

 © 1998 American Institute of Physics

$\times n$ symmetric matrices, and instead of the Maxwell field one has the column of n Maxwell fields (thus, the EMDA theory corresponds to the case of $n=1$).

It is shown that one can introduce the Peccei–Quinn axion matrix for an arbitrary $n \geqslant 1$ and that the resulting equations of motion possess the $\mathrm{Sp}(2n,R)$ hidden symmetry. This fact generalizes the one for the EMDA theory which $\mathrm{Sp}(2,R) \sim \mathrm{SL}(2,R)$ symmetry is known as the S-duality.[26,27] Also it is shown that the model admits the discrete symmetry transformation which had been named as the ''strong-weak coupling duality'' for the low energy heterotic string theory with moduli fields in four dimensions.[3]

In the next section we consider the stationary case and reduce the four-dimensional model with an arbitrary n to three dimensions. We use an approach which was first formulated by Israel and Wilson for the Einstein–Maxwell theory in Ref. 28, and recently had been generalized to the EMDA theory in Ref. 29. It is shown that the system under consideration allows the σ-model representation with a symmetric target space, which is a natural matrix generalization of the EMDA one.

Then, the Kähler formalism based on the use of the $(n+1)\times(n+1)$ Ernst matrix potential is developed. It is shown that the motion equations become the matrix-valued Ernst equation. The chiral $2(n+1)\times 2(n+1)$ matrix constructed on the real and imaginary parts of the Ernst matrix potential is presented. This matrix belongs to the coset $\mathrm{Sp}[2(n+1),R]/\mathrm{U}(n+1)$, and the established chiral formulation possesses the $\mathrm{Sp}[2(n+1),R]$ group of symmetry transformations. The corresponding transformations for the Ernst matrix potential have the ''matrix-valued $\mathrm{SL}(2,R)$ form'' and directly generalize the transformations for the Einstein theory. Also it is shown that the $\mathrm{Sp}(2n,R)$ symmetry, corresponding to one of the four-dimensional motion equations, exists as the subgroup of the complete $\mathrm{Sp}[2(n+1),R]$ symmetry group. At the end of the section the discrete transformation, which is the analogy of the strong-weak coupling duality for the low energy heterotic string theory in three dimensions,[4] is presented.

In the following section we demonstrate some new features of the model, which arise in the stationary axisymmetric case. It is established that the theory allows the alternative Ernst-like formulation using two real $(n+1)\times(n+1)$ matrices, which are the nondualized (i.e., depended on the original nontarget space field components) analogies of the Ernst matrix and its complex conjugation. These two real matrices are used for the construction of the new $2(n+1)\times 2(n+1)$ chiral matrix which coincides with the Belinsky–Zakharov one for the case of the Einstein theory (see also Ref. 23 for the EMDA theory case). This new matrix undergoes the same $\mathrm{Sp}[2(n+1),R]$ transformations as the coset chiral matrix under the action of the isometry group. The above mentioned real Ernst-like matrix potentials transform as their complex Ernst analogies.

At the end of the paper we present the Kramer–Neugebauer map for an arbitrary n. It turned out that this transformation has an especially simple form in terms of the real and complex pairs of the Ernst matrix potentials.

II. FOUR-DIMENSIONAL EQUATIONS

In this paper we consider a class of gravity models with the action

$$ {}^4S = \int d^4x |g|^{1/2} \{ -R + \mathrm{Tr}[\tfrac{1}{2}(\partial p p^{-1})^2 - pFF^T + \tfrac{1}{3}(pH)^2] \}, \tag{1} $$

where R is the Ricci scalar for the metric $g_{\mu\nu}$, $(\mu = 0,\ldots,3)$; $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ and

$$ H_{\mu\nu\lambda} = \partial_\mu B_{\nu\lambda} - \tfrac{1}{2}(A_\mu F^T_{\nu\lambda} + F_{\nu\lambda} A^T_\mu) + \text{cyclic}. \tag{2} $$

Here the symmetric $n\times n$ matrix p is constructed on the scalar field components ($p = e^{-2\phi}$ for the EMDA case, where ϕ is the dilaton field), $B_{\mu\nu}$ is the symmetric $n\times n$ matrix which contains the antisymmetric tensor Kalb–Ramond fields (i.e., $B^T_{\mu\nu} = B_{\mu\nu}$ and $B_{\mu\nu} = -B_{\nu\mu}$), and A_μ is the $n \times 1$ column of the Abelian vector fields. This action describes the Einstein and EMDA theories in the cases of $n=0$ and $n=1$, respectively, and provides their natural generalization for an arbitrary n. The models of such type arise in the low energy limit of the heterotic string theory after the compactification of extra dimensions on a torus.[1–7]

As it has been done for the EMDA theory and for the string gravity with moduli fields, one can establish an alternative formulation of the problem which is connected with the dualization of the axion field $H_{\mu\nu\lambda}$ in four dimensions. Namely, using the first equation of the Euler–Lagrange ones,

$$\nabla^{\mu}(pH^{\mu\nu\lambda}p)=0, \tag{3}$$

$$\nabla_{\mu}(pF^{\mu\nu})+\tfrac{1}{2}pH^{\mu\nu\lambda}F_{\mu\nu}=0, \tag{4}$$

$$\nabla J^{p}+pFF^{T}-\tfrac{2}{3}(pH)^{2}=0, \tag{5}$$

$$R_{\mu\nu}=\mathrm{Tr}[\tfrac{1}{2}J^{p}_{\mu}J^{p}_{\nu}-2p(F_{\mu\lambda}F^{T^{\nu\lambda}}-\tfrac{1}{4}g_{\mu\nu}FF^{T})+pH_{\mu\lambda\sigma}pH^{\lambda\sigma}_{\nu}-\tfrac{1}{3}g_{\mu\nu}(pH)^{2}], \tag{6}$$

where $J^{p}=(\nabla p)p^{-1}$, it is possible to introduce the symmetric pseudoscalar matrix q,

$$H^{\mu\nu\lambda}=\tfrac{1}{2}E^{\mu\nu\lambda\sigma}p^{-1}q_{,\sigma}p^{-1} \tag{7}$$

generalizing the Peccei-Quinn axion field. It satisfies the relation

$$\nabla J^{q}-J^{p}J^{q}+p\tilde{F}F^{T}=0, \tag{8}$$

where $J^{q}=(\nabla q)p^{-1}$ and $\tilde{F}^{\mu\nu}=\tfrac{1}{2}E^{\mu\nu\lambda\sigma}F_{\lambda\sigma}$, which together with Eqs. (4)–(6), rewritten using Eq. (7) as

$$\nabla_{\mu}(pF^{\mu\nu}+q\tilde{F}^{\mu\nu})=0, \tag{9}$$

$$\nabla J^{p}+(J^{q})^{2}+pFF^{T}=0, \tag{10}$$

$$R_{\mu\nu}=\mathrm{Tr}[\tfrac{1}{2}(J^{p}_{\mu}J^{p}_{\nu}+J^{q}_{\mu}J^{q}_{\nu})-2p(F_{\mu\lambda}F^{T\lambda}_{\ \ \nu}-\tfrac{1}{4}g_{\mu\nu}FF^{T})] \tag{11}$$

form the complete system for the action

$${}^{4}S=\int d^{4}x|g|^{1/2}\{-R+\mathrm{Tr}[\tfrac{1}{2}((J^{p})^{2}+(J^{q})^{2})-pFF^{T}-q\tilde{F}F^{T}]\}. \tag{12}$$

The action of this type was considered for the first time by Breitenlohner *et al.* in Ref. 30, where some remarkable properties of the model had been established.

The main purpose of this paper is to show a complete formal analogy between the series of models described by the action (1), and the Einstein theory in three and two dimensions. However, the simplest representative of the series, which possesses all the typical properties in four dimensions, is the EMDA theory. Namely, let us establish the symmetries of the theory for an arbitrary $n\geqslant 1$. To do it one can introduce the complex variables

$$z=q+ip, \tag{13}$$

$$\mathscr{F}=\tfrac{1}{2}(F+i\tilde{F}). \tag{14}$$

Then the set of motion equations obtains the form:

$$\nabla_{\mu}(z.\mathscr{F}^{\mu\nu}-\overline{z.\mathscr{F}}^{\mu\nu})=0, \tag{15}$$

$$\nabla J^{z}-J^{z}(J^{z}-\overline{J^{z}})-\frac{i}{2}(z-\overline{z}).\overline{\mathscr{F}}\mathscr{F}^{+}=0, \tag{16}$$

$$R_{\mu\nu}=\mathrm{Tr}\{2J^{z}_{(\mu}\overline{J^{z}_{\nu)}}+i(z-\overline{z})(.\mathscr{F}_{\mu\lambda}\mathscr{F}^{+\lambda}_{\ \ \nu.}+\overline{\mathscr{F}}_{\mu\lambda}\mathscr{F}^{T\lambda}_{\ \ \nu.})\}, \tag{17}$$

where $J^{z}=\nabla z(z-\overline{z})^{-1}$ (see also Ref. 26 for the EMDA case).

It is not difficult to prove using a straightforward calculation that these equations are invariant under the transformation

$$z \to s^T (z^{-1} + l)^{-1} s + r, \quad \mathcal{F} \to s^{-1}(I + lz)\mathcal{F}, \tag{18}$$

where the constant matrices $r^T = r$, $l^T = l$ and s are real. The former formula from (18) also can be rewritten using the following $2n \times 2n$ matrices

$$m = \begin{pmatrix} p^{-1} & p^{-1}q \\ qp^{-1} & p + qp^{-1}q \end{pmatrix} \tag{19}$$

(see the similar matrix for the Einstein–Dilaton–Axion theory in Ref. 31),

$$g = \begin{pmatrix} (s^T)^{-1} & (s^T)^{-1} r \\ l(s^T)^{-1} & s + l(s^T)^{-1} r \end{pmatrix}, \tag{20}$$

as

$$m \to g^T m g. \tag{21}$$

The above introduced matrix g satisfies the relation $g^T j g = j$, where

$$j = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}, \tag{22}$$

and I is the unit matrix, so that $g \in \mathrm{Sp}(2n,R)$ (also from the relations $m^T j m = j$ and $m^T = m$ it follows that $m \in \mathrm{Sp}(2n,R)/\mathrm{U}(n)$, see the third chapter for details). Thus the motion equations possess the $\mathrm{Sp}(2n,R)$ symmetry for an arbitrary $n \geqslant 1$.

Not all the symplectic transformations (21) can be represented in the Gauss decomposition form (20). For example, one can prove that $j \in \mathrm{Sp}(2n,R)$, but only the limit procedure defined by the matrices $l = -r = s \to \infty$ allows one to obtain the matrix j from Eq. (20). The corresponding map is $m \to m^{-1}$; and using complex variables one has:

$$z \to -z^{-1}, \quad \mathcal{F} \to z\mathcal{F}. \tag{23}$$

It is not difficult to prove that this transformation is actually the symmetry transformation for the equations (15)–(17). The map (23) generalizes the strong-weak coupling duality of the EMDA theory to the case of an arbitrary n. It has the same form as the transformation established for the low energy heterotic string theory with moduli in four dimensions.[3]

It is easy to see that only the $\mathrm{GL}(n,R)$ subgroup of the complete $\mathrm{Sp}[2(n+1),R]$ group, defined by a matrix s, preserves the action (12) which can be rewritten as

$${}^4S = \int d^4x |g|^{1/2} \{ -R + \mathrm{Tr}[2 J^z \overline{J^z} + i(z\mathcal{F}\mathcal{F}^T - \overline{z\mathcal{F}}\mathcal{F}^+)] \}. \tag{24}$$

The situation is the same for the EMDA theory which possesses the $\mathrm{Sp}(2,R) \sim \mathrm{SL}(2,R)$ symmetry only for the motion equations.[26,27] This symmetry is connected with the notable S-duality for the low energy heterotic string theory. The subgroup $\mathrm{SL}(2,Z)$ of the group $\mathrm{SL}(2,R)$ is an exact symmetry for the heterotic string in four dimensions.[3]

Also one can see that the discussed models do not possess any analogy with the T-duality transformations,[3] thus the established $\mathrm{Sp}(2n,R)$ symmetry is the single nontrivial one.

III. REDUCTION TO THREE DIMENSIONS

In this section we study the stationary case using the classical approach of Israel and Wilson.[28] This method also had been explored for the EMDA theory analysis;[29] one can find its application to the string gravity with moduli fields in Ref. 4.

First of all, it is convenient to parametrize the four-dimensional line element as

$$ds_4^2 = g_{\mu\nu}dx^\mu dx^\nu = f(dt - \omega_m dx^m)^2 - f^{-1}h_{mn}dx^m dx^n, \tag{25}$$

where $m = 1,2,3$. (In this section all vector variables, as well as the operator ∇, will be related with the three-metric h_{mn}.)

Then, the $\mu = m$ component of Eq. (9) allows one to introduce the magnetic potential u according to

$$pF^{mn} + q\tilde{F}^{mn} = \frac{1}{\sqrt{2}} fE^{mnk}u_{,k}, \tag{26}$$

while for the electric potential v we put

$$F_{m0} = \frac{1}{\sqrt{2}} v_{,m}. \tag{27}$$

Using the relations[28] $F^{0m} = \omega_n F^{mn} + h^{mn}F_{n0}$ and $F_{mn} = f^{-2}h_{mk}h_{nl}F^{kl} + 2F_{0[m}\omega_{n]}$, it is possible to rewrite the $\mu = 0$ component of Eq. (9) as

$$\nabla[f^{-1}(p\nabla v - qp^{-1}(\nabla u - q\nabla v))] = -f^{-2}\boldsymbol{\tau}\nabla u, \tag{28}$$

where $\boldsymbol{\tau} = -f^2\nabla\times\boldsymbol{\omega}$. The second electromagnetic equation can be obtained from the Bianchi identity $F_{[mn,k]} = 0$; the result is

$$\nabla[f^{-1}p^{-1}(\nabla u - q\nabla v)] = f^{-2}\boldsymbol{\tau}\nabla v. \tag{29}$$

The equations (8) and (10) transform to

$$\nabla J^q - J^p J^q - f^{-1}[(\nabla u - q\nabla v)\nabla v^T + p\nabla v(\nabla u - q\nabla v)^T p^{-1}) = 0, \tag{30}$$

$$\nabla J^p + (J^p)^2 + f^{-1}[p\nabla v\nabla v^T - (\nabla u - q\nabla v)(\nabla u - q\nabla v)^T p^{-1}] = 0. \tag{31}$$

Then, the $\binom{m}{0}$-component of the Einstein equations (11) means the existence of the rotational potential χ

$$\nabla\chi + v^T\nabla u - u^T\nabla v = \boldsymbol{\tau}, \tag{32}$$

which accordingly with the definition of $\boldsymbol{\tau}$ satisfies

$$\nabla[f^{-2}(\nabla\chi + v^T\nabla u - u^T\nabla v)] = 0; \tag{33}$$

while the $(_{00})$-component of Eq. (11) leads to the following equation for the function f:

$$f\nabla^2 f - (\nabla f)^2 + (\nabla\chi + v^T\nabla u - u^T\nabla v)^2 - f[\nabla v^T p\nabla v + (\nabla u - q\nabla v)^T p^{-1}(\nabla u - q\nabla v)] = 0. \tag{34}$$

Finally, the $(^{mn})$-component of Eq. (11) after the lowering of the three-dimensional indexes takes the form:

$${}^3R_{mn} = \tfrac{1}{2}\mathrm{Tr}[J_m^p J_n^p + J_m^q J_n^q] + \tfrac{1}{2}f^{-2}[f_{,m}f_{,n} + \boldsymbol{\tau}_m\boldsymbol{\tau}_n] - f^{-1}[v_{,m}^T p v_{,n} + (u_{,m} - qv_{,m})^T p^{-1}(u_{,n} - qv_{,n})], \tag{35}$$

where ${}^3R$ is constructed using the metric h_{mn} and $\boldsymbol{\tau}_m$ is given by Eq. (32).

One can prove that Eqs. (28)–(31) and (33)–(35) are the Euler–Lagrange ones for the action

$$\begin{aligned} {}^3S = \int d^3x h^{1/2}\{ &-{}^3R + \tfrac{1}{2}\,\mathrm{Tr}[(J^p)^2 + (J^q)^2] + \tfrac{1}{2}f^{-2}[(\nabla f)^2 + (\nabla\chi + v^T\nabla u - u^T\nabla v)^2] \\ &- f^{-1}[\nabla v^T p\nabla v + (\nabla u - q\nabla v)^T p^{-1}(\nabla u - q\nabla v)]\}. \end{aligned} \tag{36}$$

Hence this system allows the sigma-model representation in the stationary case.

This result contains the one for the EMDA theory (see Refs. 29 and 32), which admits some compact matrix formulations. To establish them for an arbitrary n, it is convenient to introduce the complex potentials

$$\Phi = u - zv, \qquad \mathscr{E} = if - \chi + v^T \Phi, \tag{37}$$

which together with the matrix z, entered before (13), form the set of Ernst-like potentials[33,34] for this problem (see also Ref. 22 for the EMDA case). These potentials can be combined into the following $(n+1)\times(n+1)$ symmetric matrix

$$E = \begin{pmatrix} \mathscr{E} & \Phi^T \\ \Phi & -z \end{pmatrix}, \tag{38}$$

which provides the Kähler representation of the model. Namely, it is easy to check that

$$^3S = \int d^3x h^{1/2} \{ -{}^3R + 2\ \mathrm{Tr}(J^E \overline{J^E}) \}, \tag{39}$$

where $J^E = \nabla E (E - \overline{E})^{-1}$ and the two-form $\mathscr{K} = 2i\ \mathrm{Tr}(\mathscr{J}^E \wedge \overline{\mathscr{J}^E})$, corresponding to the target space metric $dl^2 = 2\ \mathrm{Tr}(\mathscr{J}^E \overline{\mathscr{J}^E})$ with $\mathscr{J}^E = dE(E - \overline{E})^{-1}$, is exact. The Kähler formulations of the Einstein and Einstein-Maxwell theories are well known in the literature.[16] Another example of the Kähler systems can be found in Ref. 35.

The equations following from the action (39)

$$\nabla J^E = J^E (J^E - \overline{J^E}), \tag{40}$$

$$^3R_{mn} = 2\ \mathrm{Tr}(J^E_{(m} \overline{J^E}_{n)}) \tag{41}$$

coincide with the Ernst ones[33] in the case of $n = 0$. Thus it is natural to name the matrix function E as the ''matrix Ernst potential'' for an arbitrary n. This Ernst-like representation for the model provides the above mentioned complete formal analogy between it and the Einstein theory in the stationary case.

Then, using the real and imaginary parts of $E = Q + iP$,

$$P = \begin{pmatrix} f - v^T p v & -v^T p \\ -pv & -p \end{pmatrix}, \quad Q = \begin{pmatrix} v^T w - \chi & w^T \\ w & -q \end{pmatrix}, \tag{42}$$

where $w = u - qv$, one can rewrite the action (39) as

$$^3S = \int d^3x h^{1/2} \{ -{}^3R + \tfrac{1}{2}\ \mathrm{Tr}[(J^P)^2 + (J^Q)^2] \}, \tag{43}$$

with $J^P = (\nabla P) P^{-1}$ and $J^Q = (\nabla Q) P^{-1}$.

Now let us establish an alternative matrix formulation of the problem based on the use of the original nontarget space variables. One can see that the first equation of motion corresponding to the action (43),

$$\nabla J^Q - J^P J^Q = 0, \tag{44}$$

$$\nabla J^P + (J^Q)^2 = 0, \tag{45}$$

$$^3R_{mn} = \tfrac{1}{2} \mathrm{Tr}[J^P_m J^P_n + J^Q_m J^Q_n], \tag{46}$$

being written as $\nabla[P^{-1}(\nabla Q) P^{-1}] = 0$, ensures the compatibility condition for the relation

$$\nabla\times\mathbf{\Omega}=P^{-1}(\nabla Q)P^{-1} \tag{47}$$

which defines the vector matrix $\mathbf{\Omega}$. Then, using the relations (7), (26), (32) and the definition of $\boldsymbol{\tau}$ one obtains the explicit form of this matrix:

$$\mathbf{\Omega}=\begin{pmatrix} \boldsymbol{\omega} & -\sqrt{2}(\mathbf{A}+A_0\boldsymbol{\omega})^T \\ -\sqrt{2}(\mathbf{A}+A_0\boldsymbol{\omega}) & 2\mathbf{B}+(\vec{A}+A_0\boldsymbol{\omega})A_0^T+A_0(\mathbf{A}+A_0\boldsymbol{\omega})^T \end{pmatrix}, \tag{48}$$

where $B_m=2B_{0m}$. It is easy to see that the matrix $\mathbf{\Omega}$ consists of the original field variables from Eqs. (1) and (2) which had been named as ''nondualized'' in Ref. 23 for the case of the EMDA theory.

The remaining components of the Kalb–Ramond field $B_{mn}\equiv\frac{1}{2}E_{mnk}C^k$ satisfy the nondynamical equation

$$\nabla\mathbf{C}=\{\nabla[\mathbf{B}\times\boldsymbol{\omega}]+[\mathbf{B}-(\mathbf{A}+A_0\boldsymbol{\omega})A_0^T-A_0(\mathbf{A}+A_0\boldsymbol{\omega})^T]\nabla\times\boldsymbol{\omega}+(\mathbf{A}+A_0\boldsymbol{\omega})\nabla\times(\mathbf{A}+A_0\boldsymbol{\omega})^T\}$$
$$+\nabla\times(\mathbf{A}+A_0\boldsymbol{\omega})(\mathbf{A}+A_0\boldsymbol{\omega})^T, \tag{49}$$

as it follows from Eqs. (7) and the t-independance of the q.

Then, it is easy to see from Eqs. (47) that the matrices P and $\mathbf{\Omega}$ satisfy the relation $\nabla\times[P\nabla\times(\mathbf{\Omega})P]=0$ which also can be rewritten using the matrix current $J^{\mathbf{\Omega}}=P\nabla\times\mathbf{\Omega}$ as

$$\nabla\times J^{\mathbf{\Omega}}-J^{\mathbf{\Omega}}\times J^P=0. \tag{50}$$

This relation together with Eqs. (45) and (46), expressed in terms of the matrices P and $\mathbf{\Omega}$,

$$\nabla J^P+(J^{\mathbf{\Omega}})^2=0, \tag{51}$$

$${}^3R_{mn}=\tfrac{1}{2}\mathrm{Tr}[J^P_mJ^P_n+J^{\mathbf{\Omega}}_mJ^{\mathbf{\Omega}}_n], \tag{52}$$

form the set of motion equations for the action

$${}^3S=\int d^3xh^{1/2}(-{}^3R+\tfrac{1}{2}\,\mathrm{Tr}[(J^P)^2-(J^{\mathbf{\Omega}})^2]), \tag{53}$$

thus these matrices provide an alternative Lagrange formulation of the theory under consideration.

To establish the symmetry group of the system, it is convenient to introduce the $2(n+1)\times2(n+1)$ symmetric matrix

$$M=\begin{pmatrix} P^{-1} & P^{-1}Q \\ QP^{-1} & P+QP^{-1}Q \end{pmatrix}, \tag{54}$$

which possesses the symplectic property

$$M^TJM=J,\quad \text{where}\quad J=\begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}. \tag{55}$$

Then the action (43) obtains the form

$${}^3S=\int d^3xh^{1/2}\{-{}^3R+\tfrac{1}{4}\,\mathrm{Tr}[(J^M)^2]\}, \tag{56}$$

where $J^M=\nabla MM^{-1}$, and the corresponding equations of motion are

$$\nabla J^M=0, \tag{57}$$

$${}^3R_{mn}=\tfrac{1}{4}\mathrm{Tr}[J^M_mJ^M_n]. \tag{58}$$

It is known that in the case of $n=0$ (for the Einstein theory) a matrix M belongs to the coset $\mathrm{Sp}(2,R)/\mathrm{U}(1)$ [or, equivalently, to the $\mathrm{SL}(2,R)/\mathrm{O}(2)$].[12] Also, it has been established that for the EMDA theory one has $M\in\mathrm{Sp}(4,R)/\mathrm{U}(2)$.[22,36] Hence it is natural to suppose that $M\in\mathrm{Sp}[2(n+1),R]/\mathrm{U}(n+1)$ for an arbitrary n.

This supposition really takes place and the local isomorphism can be established as follows. One can see that a matrix M remains symplectic and symmetric under the transformation

$$M\to G^T M G \tag{59}$$

with an arbitrary symplectic matrix G. It means that the group of the isometry transformations for the action (56) is isomorphic to $\mathrm{Sp}[2(n+1),R]$. The set of the corresponding generators $\Gamma=G-I$ can be obtained from (59); the result is

$$\Gamma=\begin{pmatrix} -\Gamma_S^T & \Gamma_R \\ \Gamma_L & \Gamma_S \end{pmatrix}, \tag{60}$$

where $(n+1)\times(n+1)$ infinitesimal matrices Γ_R and Γ_L are symmetric, and Γ_S is an arbitrary matrix of the same dimension. Then, to obtain the coset, one must remove all the antisymmetric generators from Eq. (60), i.e., to exclude the set of generators which is defined by the matrix

$$\tilde{\Gamma}=\begin{pmatrix} \tilde{\Gamma}_S & \tilde{\Gamma}_R \\ -\tilde{\Gamma}_R & \tilde{\Gamma}_S \end{pmatrix}, \tag{61}$$

where $\tilde{\Gamma}_S^T=-\tilde{\Gamma}_S$ and $\tilde{\Gamma}_R^T=\tilde{\Gamma}_R$.

From the other hand, the generators u of the group $\mathrm{U}(n+1)$ can be found from the relation $\mathrm{U}^+\mathrm{U}=I$ in the infinitesimal case of $\mathrm{U}=I+u$. The result is

$$u=u_1+iu_2, \tag{62}$$

where $u_1^T=-u_1$ and $u_2^T=u_2$ are the real $(n+1)\times(n+1)$ matrices. Now one can prove using a straightforward calculation of the commutators for the corresponding generators that the waiting isomorphism can be defined as

$$\tilde{\Gamma}_S\equiv u_1, \quad \tilde{\Gamma}_R\equiv u_2. \tag{63}$$

Now let us establish the action of the $\mathrm{Sp}[2(n+1),R]$ group of the isometry transformations on the Ernst matrix potential E. One can see that the matrix

$$G=\begin{pmatrix} (S^T)^{-1} & (S^T)^{-1}R \\ L(S^T)^{-1} & S+L(S^T)^{-1}R \end{pmatrix}, \tag{64}$$

with $R^T=R$ and $L^T=L$ possesses the symplectic property and describes all the belonging to $\mathrm{Sp}[2(n+1),R]$ transformations which can be continuously connected with the identical one. It is easy to show that the corresponding transformation for the matrix E is

$$E\to S^T(E^{-1}+L)^{-1}S+R. \tag{65}$$

Thus the matrix E undergoes the ''matrix-valued $\mathrm{SL}(2,R)$'' transformations which are the matrix generalizations of the $\mathrm{SL}(2,R)$ ones for the Einstein theory. Namely, the transformation defined by the matrix L is the matrix generalization of the Ehlers one,[37] while the matrices S and R define the transformations which generalize the rescaling and gauge shift correspondingly for the stationary Einstein system.[12]

The transformation (65) has the same form as the transformation for the matrix z in Eq. (18). It is not difficult to prove that one can identify the $\mathrm{Sp}(2n,R)$ subgroup of the complete group $\mathrm{Sp}[2(n+1),R]$, defined by the matrices

$$L=\begin{pmatrix} 0 & 0 \\ 0 & l \end{pmatrix}, \quad S=\begin{pmatrix} 1 & 0 \\ 0 & s \end{pmatrix}, \quad R=\begin{pmatrix} 0 & 0 \\ 0 & r \end{pmatrix}, \tag{66}$$

where $l^T=l$ and $r^T=r$, with the $\mathrm{Sp}(2n,R)$ symmetry transformations (18) of the motion equations (15)–(17) in the stationary case.

As an example of a symplectic transformation which cannot be represented in the form of (64), one can take the matrix J. It is easy to see that the corresponding map is $M\to M^{-1}$ and its complex representation has the form $E\to -E^{-1}$. This transformation can be obtained from the decomposition formula (64) using the limit procedure as it has been done in the four-dimensional case. It also provides an analogy with the strong-weak coupling duality transformation of the low energy heterotic string theory with moduli in three dimensions.[4]

At the end of this section we want to note that all the transformations discussed here preserve both the actions and the corresponding sets of motion equations.

IV. TWO-DIMENSIONAL CASE

In this section we consider some additional properties of the theory which arise in the stationary and axisymmetric case.

Here one can parametrize the three-dimensional line element in the Lewis–Papapetrou form:

$$ds_3^2=e^{2\gamma}(d\rho^2+dz^2)+\rho^2 d\varphi^2. \tag{67}$$

Then Eq. (57) transforms to

$$\nabla(\rho J^M)=0 \tag{68}$$

which is the Euler–Lagrange equation for the action

$$^2S=\frac{1}{4}\int d\rho\ dz\ \rho\ \mathrm{Tr}[(J^M)^2], \tag{69}$$

while the Einstein equations (58) become the relations defining the function γ:

$$\begin{aligned} \gamma_{,z}&=\frac{\rho}{4}\mathrm{Tr}[J^M_\rho J^M_z], \\ \gamma_{,\rho}&=\frac{\rho}{8}\mathrm{Tr}[(J^M_\rho)^2-(J^M_z)^2] \end{aligned} \tag{70}$$

(the operator ∇ is connected with the flat metric δ_{ab} in this section).

The remarkable fact is that one can obtain another chiral matrix for the stationary axisymmetric equations of the problem. To make it, we first consider the formalism based on the use of the matrices P and Q. It is easy to see that the φ-independent equations of motion (44)–(45) take the form

$$\nabla(\rho J^Q)-\rho J^P J^Q=0, \tag{71}$$

$$\nabla(\rho J^P)+\rho(J^Q)^2=0, \tag{72}$$

and can be derived from the action

$$^2S=\frac{1}{2}\int d\rho\ dz\ \rho\ \mathrm{Tr}[(J^P)^2+(J^Q)^2]. \tag{73}$$

Also, from Eq. (46) one obtains:

$$\gamma_{,z}=\frac{\rho}{2}\mathrm{Tr}[J_\rho^P J_z^P+J_\rho^Q J_z^Q],$$
$$\gamma_{,\rho}=\frac{\rho}{4}\mathrm{Tr}[(J_\rho^P)^2-(J_z^P)^2+(J_\rho^Q)^2-(J_z^Q)^2]. \tag{74}$$

The equation (71), being written as $\nabla[\rho P^{-1}(\nabla Q)P^{-1}]=0$, provides the compatibility condition for the relation

$$\nabla\Omega=\rho P^{-1}(\tilde{\nabla}Q)P^{-1} \tag{75}$$

which defines the symmetric matrix Ω (here, accordingly with Ref. 12, $\tilde{\nabla}_\rho=\nabla_z$ and $\tilde{\nabla}_z=-\nabla_\rho$). Then, using a straightforward calculation one obtains that

$$\Omega=\mathbf{\Omega}_\varphi. \tag{76}$$

It is easy to see that the compatibility condition for the relation (75), rewritten as $\nabla Q=-\rho^{-1}P(\tilde{\nabla}\Omega)P$, and Eq. (72), expressed in terms of the matrices P and Ω,

$$\nabla(\rho^{-1}J^\Omega)+\rho^{-1}J^\Omega J^P=0, \tag{77}$$

$$\nabla(\rho^{-1}J^P)+\rho^{-1}(J^\Omega)^2=0, \tag{78}$$

form the Euler–Lagrange set of equations for the action

$${}^2S=\frac{1}{2}\int d\rho\ dz\ \mathrm{Tr}[\rho(J^P)^2-\rho^{-1}(J^\Omega)^2], \tag{79}$$

where $J^\Omega=P\nabla\Omega$. Also one can see from Eq. (49), that $\nabla\mathbf{C}=0$ and, fixing a gauge, it is possible to put $\mathbf{C}=0$ without lose of generality. Then, for the function γ one has

$$\gamma_{,z}=\tfrac{1}{2}\mathrm{Tr}[\rho J_\rho^P J_z^P-\rho^{-1}J_\rho^\Omega J_z^\Omega],$$
$$\gamma_{,\rho}=\tfrac{1}{4}\mathrm{Tr}[\rho((J_\rho^P)^2-(J_z^P)^2)+\rho^{-1}((J_\rho^\Omega)^2-(J_z^\Omega)^2)]. \tag{80}$$

The new chiral matrix N can be defined as follows

$$N=\begin{pmatrix} P & -P\Omega \\ -\Omega P & \Omega P\Omega-\rho^2P^{-1}\end{pmatrix}. \tag{81}$$

It satisfies the equation

$$\nabla[\rho J^N]=0, \tag{82}$$

where $J^N=\nabla NN^{-1}$, while the relations for the function γ are

$$\Gamma_{,z}=\frac{\rho}{4}\mathrm{Tr}[J_\rho^N J_z^N],$$
$$\Gamma_{,\rho}=\frac{\rho}{8}\mathrm{Tr}[(J_\rho^N)^2-(J_z^N)^2], \tag{83}$$

with

$$\Gamma=\gamma-\frac{1}{2}\ln|\det P|+\frac{n+1}{2}\ln\rho. \tag{84}$$

It is easy to check that the introduced symmetric matrix N satisfies the nongroup relation

$$NJN=-\rho^2 J \tag{85}$$

which is equivalent to the normalization condition $\det N=-\rho^2$ in the case of the Einstein theory (see also Ref. 23 for the EMDA case).

An analogy between the model under consideration and the Einstein theory allows us to establish one additional property of the system. Namely, the relation (75) provides the differential correspondence between the formulations of the problem based on the use of the dualized matrix Q and the original one Ω. The remarkable fact is that also there is a pure algebraical map

$$P\to\rho P^{-1}, \quad Q\to i\Omega, \tag{86}$$

which directly transforms Eqs. (71) and (72) into (77) and (78). The corresponding transformation for the function γ is

$$e^{2\gamma^{\Omega}}=\frac{\rho^{(n+1)/2}}{|\det P|}e^{2\gamma^{Q}}, \tag{87}$$

where γ^{Ω} is connected with matrices P and Ω, and γ^{Q} is related with the formulation using P and Q. The map (86) coincides with the Kramer–Neugebauer transformation in the cases of $n=0$ (Ref. 24) and $n=1$ (Ref. 23) and can naturally preserve this name for an arbitrary n.

Now let us establish the action of the $\mathrm{Sp}[2(n+1),R]$ group on the matrices P and Ω. To make it one can introduce the matrices

$$E_{\pm}=\Omega\pm\rho P^{-1}. \tag{88}$$

It is easy to check that the matrix N remains symmetric and preserves the property (85) under the transformation $N\to G^{T}NG$ with an arbitrary symplectic matrix G. Then, in the case of the matrix G, possessing the decomposition property (64), one has

$$E_{\pm}\to S^{T}(E_{\pm}^{-1}+L)^{-1}S+R. \tag{89}$$

Thus the real nondualized matrix variables $E_{\pm}$ transform under the $\mathrm{Sp}[2(n+1),R]$ group of transformations as the matrix Ernst potentials E and $\overline{E}$.

The entered matrices $E_{\pm}$, being the analogy for the Ernst ones E and $\overline{E}$, provide also the compact Lagrange formulation of the problem. Namely, one can check that the action (79) can be rewritten as

$${}^{2}S=-2\int d\rho\, dz\, \rho\, \mathrm{Tr}[J^{E_+}J^{E_-}], \tag{90}$$

where $J^{E_{\pm}}=(E_{\pm}-E_{\mp})^{-1}\nabla E_{\pm}-\frac{1}{2}\nabla\rho$.

Also one can see that the above established Kramer–Neugebauer map has an especially simple form in terms of the Ernst matrices $E,\overline{E}$ and their real analogies E_{+},E_{-}:

$$E\to iE_{+}, \quad \overline{E}\to iE_{-}. \tag{91}$$

V. CONCLUSION

In this paper we have considered a class of gravity models with the Einstein theory as the first representative, the EMDA theory as the second one, etc. It is shown that all the models have the same formal properties which are the natural matrix generalizations of the ones for the Einstein theory in three and two dimensions. Namely, the substitution $P\to f$, $Q\to\chi$ and $\mathbf{\Omega}\to\boldsymbol{\omega}$ (or $\Omega\to\omega$ in the stationary axisymmetric case) directly transforms all the relations for $n\geqslant 1$ to the Einstein theory relations. An analogy established here is even a more complete one, and it is possible to develop the Hauser–Ernst formalism[17,19] and to construct the Geroch group[8,9] in the stationary axisymmetric case for the model with an arbitrary n.

ACKNOWLEDGMENT

We would like to thank our colleagues for their encouragement.

[1] J. Maharana and J. H. Schwarz, Nucl. Phys. B **390**, 3 (1993).
[2] J. H. Schwarz and A. Sen, Nucl. Phys. B **411**, 35 (1993).
[3] A. Sen, Int. J. Mod. Phys. A **9**, 3707 (1993).
[4] A. Sen, Nucl. Phys. B **434**, 179 (1995).
[5] A. Sen, Nucl. Phys. B **447**, 62 (1995).
[6] J. Maharana, Phys. Rev. Lett. **75**, 208 (1995).
[7] A. K. Biswas, A. Kumar, and K. Ray, Nucl. Phys. B **453**, 181 (1995).
[8] R. Geroch, J. Math. Phys. **12**, 918 (1971).
[9] R. Geroch, J. Math. Phys. **13**, 394 (1972).
[10] V. A. Belinsky and V. E. Zakharov, Sov. Phys. JETP **48**, 985 (1978) [Zh. Exp. Teor. Fiz. **75**, 1955 (1978)].
[11] V. A. Belinsky and V. E. Zakharov, Sov. Phys. JETP **50**, 1 (1979) [Zh. Exp. Teor. Fiz. **77**, 3 (1979)].
[12] W. Kinnersley, J. Math. Phys. **18**, 529 (1977).
[13] W. Kinnersley and D. M. Chitre, J. Math. Phys. **18**, 1538 (1977).
[14] W. Kinnersley and D. M. Chitre, J. Math. Phys. **19**, 1926 (1978).
[15] W. Kinnersley and D. M. Chitre, J. Math. Phys. **19**, 2037 (1978).
[16] P. O. Mazur, Acta Phys. Pol. B **14**, 219 (1983).
[17] I. Hauser and F. J. Ernst, Phys. Rev. D **20**, 362 (1979).
[18] I. Hauser and F. J. Ernst, Phys. Rev. D **20**, 1783 (1979).
[19] I. Hauser and F. J. Ernst, J. Math. Phys. **21**, 1126 (1980).
[20] I. Hauser and F. J. Ernst, J. Math. Phys. **21**, 1418 (1980).
[21] C. M. Cosgrove, J. Math. Phys. **22**, 2624 (1981).
[22] D. V. Gal'tsov and O. V. Kechkin, Phys. Lett. B **361**, 52 (1995).
[23] O. Kechkin and M. Yurova, Phys. Rev. D **54**, 6135 (1996).
[24] G. Neugebauer and D. Kramer, Ann. Phys. (N.Y.) **24**, 62 (1969).
[25] P. Breitenlohner and D. Maison, Ann. Inst. Henri Poincaré Phys. Theor. **46**, 215 (1987).
[26] A. Shapere, S. Trivedi, and F. Wilczec, Mod. Phys. Lett. A **6**, 2677 (1991).
[27] R. Kallosh and T. Ortin, Phys. Rev. D **48**, 742 (1993).
[28] W. Israel and G. A. Wilson, J. Math. Phys. **13**, 865 (1972).
[29] D. V. Gal'tsov, A. A. Garcia, and O. V. Kechkin, J. Math. Phys. **36**, 5023 (1995).
[30] P. Breitenlohner, D. Maison, and G. Gibbons, Commun. Math. Phys. **120**, 295 (1988).
[31] I. Bakas, Nucl. Phys. B **428**, 374 (1994).
[32] D. V. Gal'tsov and O. V. Kechkin, Phys. Rev. D **50**, 7394 (1994).
[33] F. J. Ernst, Phys. Rev. **167**, 1175 (1968).
[34] F. J. Ernst, Phys. Rev. **168**, 1415 (1968).
[35] P. Fre, Nucl. Phys. B, Proc. Suppl. **45**, 59 (1996).
[36] D. V. Gal'tsov and O. V. Kechkin, Phys. Rev. D **54**, 1656 (1996).
[37] J. Ehlers, in *Les Theories de la Gravitation* (CNRS, Paris, 1959).

Full causal bulk-viscous cosmological models

M. K. Mak[a)]
Department of Physics, University of Hong Kong, Pokfulam, Hong Kong, People's Republic of China

T. Harko
Oncological Institute, 3400 Cluj-Napoca, Romania

(Received 9 December 1997; accepted for publication 13 May 1998)

New exact solutions of the gravitational field equations for a homogeneous flat Friedmann–Robertson–Walker Universe filled with a causal bulk-viscous fluid with a bulk viscosity coefficient proportional to the Hubble function are obtained in the framework of the full causal Israel–Stewart–Hiscock (ISH) theory. The evolution equation of the Universe is reduced to a first kind Abel-type first-order differential equation and new classes of exact solutions of the field equations are generated from some particular solutions of this equation leading to two classes of general solutions of the Einstein field equations corresponding to particular values of the parameters entering in the physical model. The obtained solutions are represented mathematically in an exact parametric form and are interpreted physically as describing cosmological particle production. The evolution of the temperature, entropy, deceleration parameter, and bulk viscosity coefficient is also considered for each model. © *1998 American Institute of Physics.* [S0022-2488(98)01710-1]

I. INTRODUCTION

Dissipative processes are supposed to play a fundamental role in the evolution and dynamics of the early Universe. Thirty years ago Misner[1] suggested that the observed large scale isotropy of the Universe is due to the action of the neutrino viscosity which was effective when the Universe was about 1 s old. Bulk viscosity may arise in a different context during the evolution of the early Universe like in the evolution of cosmic strings due to their interaction with each other and the surrounding matter, in classical description of the (quantum) particle production phase, interaction between matter and radiation, quark and gluon plasma viscosity, different components of dark matter, and in many other physical phenomena.[2] The first attempts at creating a theory of relativistic dissipative fluids were those of Eckart[3] and Landau and Lifshitz.[4] These theories are now known to be pathological in several respects. Regardless of the choice of equation of state, all equilibrium states in these theories are unstable and in addition signals may be propagated through the fluid at velocities exceeding the speed of light.[5] These problems arise due to the first-order nature of the theory, i.e., it considers only first-order deviations from the equilibrium leading to parabolic differential equations, hence to infinite speeds of propagation for heat flow and viscosity in contradiction with the principle of causality. While such paradoxes appear particularly glaring in relativistic theory, infinite propagation speeds already constitute a difficulty at the classical level, since one does not expect thermal disturbances to be carried faster than some (suitably defined) mean molecular speed. Conventional theory is thus applicable only to phenomena which are ''quasistationary,'' i.e., slowly varying on space and time scales characterized by mean free path and mean collision time.[5] This is inadequate for many phenomena in high-energy astrophysics and relativistic cosmology involving steep gradients or rapid variations. These deficiencies can be traced to the fact that the conventional theories (both classical and relativistic) make overly restrictive hypothesis concerning the relation between the fluxes and densities of entropy, energy, and particle number. In the case of a simple fluid, for instance, the conventional theory postulates that entropy flux is simply proportional to heat flux and that entropy density is independent of heat flow and viscosity.[5] This postulate is indeed correct to first order in deviations from equilibrium;

[a)]Electronic mail: mkmak@vtc.edu.hk

 © 1998 American Institute of Physics

however, in order to infer the correct linear phenomenological laws from the expression for the entropy production, the latter is required to second order. The neglected second-order terms are necessary to prevent noncausal and unstable behavior.[5]

A relativistic second-order theory was found by Israel[5] and developed by Israel and Stewart[6] into what is called ''transient'' or ''extended'' irreversible thermodynamics in which the deviations from equilibrium (bulk stress, heat flow, and shear stress) are treated as independent dynamical variables, leading to a total of 14 dynamical fluid variables to be determined. However, Hiscock and Lindblom[7] and Hiscock and Salmonson[8] have shown that most versions of the causal second-order theories omit certain divergence terms. The truncated causal thermodynamics of bulk viscosity leads to pathological behavior in the late Universe while the solutions of the full causal theory are well behaved for all times. Therefore the best currently available theory for analyzing dissipative processes in the Universe is the full Israel–Stewart–Hiscock (ISH) causal thermodynamics.

Very recently, Chimento and Jakubi[9,10] have found the exact general solution to the Einstein gravitational field equations in a homogeneous Universe filled with a full causal viscous fluid source obeying the relation $\xi\sim\rho^{1/2}$, corresponding to two different choices of the state equations for pressure, bulk viscosity coefficient, temperature, and bulk relaxation time. The equation describing the dynamics of the Universe takes in this case the form of a particular Painleve–Ince equation which can be linearized by means of an appropriate transformation. The exact solution of the gravitational field equations is expressed in an exact parametrical form as a two-parameter family. Depending on the values of the parameters Chimento and Jakubi classified their solutions according to their number of singularities and obtained several scenarios for the dynamics of the Universe. So, in the case[9] $p=(\gamma-1)\rho$, $\xi\sim\rho^s$, $T\sim\rho^r$ and $\tau\sim\xi/\rho\sim\rho^{s-1}$, the evolution of the Universe begins at a singularity asymptotically approaching a Minkowskian space–time or is asymptotically Friedmann or de Sitter. Chimento[11] considered in detail the properties of the corresponding Painleve–Ince equation whose invariant form was reduced to a linear inhomogeneous ordinary second-order differential equation with constant coefficients by means of a nonlocal transformation. Exact singular and late time inflationary solutions of the gravitational field equations with bulk viscosity coefficient proportional to the Hubble factor have been obtained by Chimento *et al.*[12] in the framework of a nonlinear generalization of the causal linear thermodynamics of bulk viscosity proposed by Maartens and Mendez[13] in order to obtain a thermodynamically consistent viscous fluid inflation.

Our intention in the present paper is to obtain some new classes of exact solutions of the gravitational field equations for a flat Friedmann–Robertson–Walker (FRW) Universe filled with a bulk-viscous imperfect fluid under the framework of the full Israel–Stewart–Hiscock causal theory. By supposing that the bulk viscosity coefficient is proportional to the square root of the density, $\xi\sim\rho^{1/2}$ or equivalently proportional with the Hubble factor H, $\xi\sim H$, with the use of a mathematical formalism different from that of Chimento and Jakubi,[9–11] we reduce the equation describing the evolution of the dissipative processes dominated Universe to a first kind Abel-type first-order differential equation. Particular solutions of the Abel-type equation are used to generate new general solutions of the evolution equation (and, consequently, of the Einstein field equations), corresponding to some particular values of the parameters involved in our physical model. The obtained solutions of the gravitational field equations are expressed in a parametric form and they describe inflationary evolution of a FRW Universe filled with a bulk-viscous fluid. The linear theory is supposed to be applicable at least locally and for short intervals of time and from a physical point of view, we interpret these inflationary solutions as describing particle production in the early Universe with creation pressure modeled by means of bulk-viscous pressure. The behavior of the energy density, temperature, bulk viscosity coefficient, deceleration parameter, and entropy is analyzed for each case.

The present paper is organized as follows. In Sec. II we present the field equations for a bulk-viscous fluid-filled FRW Universe in the framework of the full Israel–Stewart–Hiscock theory and the first-order first kind Abel-type evolution equation that has to be integrated is also obtained. The first class of solutions, corresponding to fixed values of parameters is presented in Sec. III. In Sec. IV we generate the second class of solutions. Certain physical properties of the obtained solutions are considered in Sec. V and the equivalence between causal bulk-viscous and

particle production models is explicitly considered. Finally, in Sec. VI we summarize and conclude our results.

II. GEOMETRY, THERMODYNAMICS, FIELD EQUATIONS, AND CONSEQUENCES

For a spatially flat Friedmann–Robertson–Walker (FRW) Universe with line element

$$ds^2=dt^2-a^2(t)(dx_1^2+dx_2^2+dx_3^2) \tag{1}$$

filled with a bulk-viscous cosmological fluid the energy momentum tensor is given by

$$T_i^k=(\rho+p+\Pi)u_iu^k-(p+\Pi)\delta_i^k, \tag{2}$$

where ρ is the energy density, p the thermodynamic pressure, Π the bulk-viscous pressure, and u_i the four velocity satisfying the condition $u_iu^i=1$.

Using (1) and (2) the gravitational field equations take the form:

$$3H^2=\rho, \tag{3}$$

$$2\dot{H}+3H^2=-p-\Pi. \tag{4}$$

From the continuity equation $T^k_{i;k}=0$ or from Eqs. (3) and (4), we obtain

$$\dot{\rho}+3(\rho+p)H=-3\Pi H. \tag{5}$$

The causal evolution equation for the bulk viscous pressure in the most general case including nonlinear terms is given by:[13]

$$\tau\dot{\Pi}=-3\xi H-\frac{1}{2}\,\tau\Pi\left(3H+\frac{\dot{\tau}}{\tau}-\frac{\dot{\xi}}{\xi}-\frac{\dot{T}}{T}\right)+\xi\chi(\Pi), \tag{6}$$

where T is the temperature, ξ the bulk viscosity coefficient, τ the relaxation time, $\chi(\Pi)$ the thermodynamic ''force'' associated to the thermodynamic ''flux'' Π and we have denoted $H=\dot{a}/a$ (Hubble factor). Equation (6) arises as the simplest way to satisfy the H theorem (i.e., for the entropy production to be non-negative), $S^i_{;i}=-(\Pi/T)\chi(\Pi)\geqslant 0$, where $S^i=sN^i-(\tau\Pi^2/2\xi T)u^i$ is the entropy flow vector and $N^i=nu^i$ is the particle flow vector[5,6] and to maintain quadratic dissipative contributions to the effective nonequilibrium specific entropy[13] $S_{\rm eff}=S-(\tau/2nT\xi)\Pi^2$. For $\chi=-\Pi/\xi$ we obtain the linear Israel–Stewart theory[5,6] describing small deviations from equilibrium.

Maartens and Mendez[13] have considered a nonlinear generalization of the linear causal thermodynamics leading to a thermodynamically consistent inflation with $\chi=-\Pi/(\xi+\tau_*\Pi)$, where $\tau_*=k^2\tau$, $k=$const, is a characteristic nonlinear time scale. For $\tau_*\ll 1/\Pi$ the usual Israel–Stewart theory is recovered.

In order to describe the general evolution of a bulk-viscous dissipative cosmological fluid with arbitrary nonlinear viscosity we shall consider that it is characterized by an iterative structure whose single steps are based on the possibility of neglecting in the field equations the effects due to nonlinearity. Thus the time evolution of a nonlinear viscous Universe consists of an infinite number of linear epochs each of which being related to the previous one by well-defined transformations laws for some physical quantities necessary to individualize a given dissipative linear regime.

From a mathematical point of view this means that we consider for the thermodynamic ''force'' $\chi(\Pi)$ a ''piecewise linear approximation'' of the form

$$\chi(\Pi)\approx\chi^i(\Pi)\approx-\frac{\Pi^{(i)}}{\xi^{(i)}_{\rm eff}},\quad t_i\leqslant t\leqslant t_{i+1},\quad i=1,...,n \tag{7}$$

and we suppose that at least locally and for a short interval of time, $t_i\leqslant t\leqslant t_{i+1}, |t_{i+1}-t_i|\ll 1$, $i=1,...,n$, the relations $\xi^{(i)}_{\rm eff}\approx\xi(t^{(i)}_m)$, and $\Pi^{(i)}\approx\Pi(t^{(i)}_m), \forall t^{(i)}_m\in[t_i,t_{i+1}], i=1,...,n$ hold.

In order to close the system of equations (3)–(6) we have to give the equation of state for p and specify T, τ, and ξ. We shall assume the following laws:[14]

$$p=(\gamma-1)\rho, \tag{8}$$

$$\xi=\alpha\rho^{s}, \tag{9}$$

$$T=\beta\rho^{r}, \tag{10}$$

$$\tau=\frac{\xi}{\rho}=\alpha\rho^{s-1}, \tag{11}$$

where $1\leqslant\gamma\leqslant 2$, $\alpha\geqslant 0$, $\beta\geqslant 0$, $r\geqslant 0$, and $s\geqslant 0$ are constants. Equations (8)–(10) are standard in cosmological models whereas (11) is a simple procedure to ensure that the speed of viscous pulses does not exceed the speed of light.[14]

In the context of irreversible thermodynamics p, ρ, T, and the number density n are equilibrium magnitudes which are related by equations of state of the form $\rho=\rho(T,n)$ and $p=p(T,n)$. From the general requirement that the entropy is a state function we obtain the equation $(\partial\rho/\partial n)_T=(\rho+p)/n-(T/n)(\partial p/\partial T)_n$. In the present model this relation would impose the constraint $r=(\gamma-1)/\gamma$ so that $0\leqslant r\leqslant 1/2$ for $1\leqslant\gamma\leqslant 2$, a range of values which is usually considered in the physical literature.[9–11] With this choice for $\gamma=4/3$ we obtain the usual relationship for thermal radiation relating the energy density ρ to temperature T, $\rho\sim T^4$. But this equation is valid only for equilibrium and leads to a vanishing bulk-viscosity coefficient.[15] On the other hand the evolution law for the temperature[16]

$$\frac{\dot{T}}{T}=-3H\left(\frac{\dfrac{\partial p}{\partial T}}{\dfrac{\partial\rho}{\partial T}}+\frac{\Pi}{\left[T\dfrac{\partial\rho}{\partial T}\right]}\right)$$

leads, in the case of a FRW Universe and in the framework of the full causal theory to a nonbarotropic equation for temperature of the form $T\sim\rho^{a_1}a^{3(\gamma a_1-b_1)}$, where $a_1\equiv\rho/nTc_\nu$, $b_1=(1/nc_\nu)(\partial p/\partial T)$, where we have denoted $c_\nu=(1/n)(\partial\rho/\partial T)$, and n is the particle number.[16] In fact any reliable expression linking ρ to T should depend very much on the different kinds of particle species making up the early cosmological fluid and their proportions in it, something that no one knows. In the first moments of the Universe's expansion the material content is completely unknown and in that nonequilibrium environment even the concept of an equation of state is inadequate, the energy density and pressure being independent functions of the expansion scale factors and time. So, we shall analyze in the present paper a generalized physical model by considering that r, α, and γ are independent parameters and we shall not impose *a priori* any limitations on the range of values of r.

The growth of the total comoving entropy over a proper time interval (t_0,t_1) is given by:[14]

$$\Sigma_1-\Sigma_0=-\frac{3}{k_B}\int_{t_0}^{t_1}\frac{H\Pi a^3}{T}\,dt, \tag{12}$$

where k_B is the Boltzmann's constant.

The Israel–Stewart theory is derived under the assumption that the thermodynamical state of the fluid is close to equilibrium which means that the nonequilibrium bulk-viscous pressure should be small when compared to the local equilibrium pressure, that is[14] $|\Pi|\ll p=(\gamma-1)\rho$. If this condition is violated then one is effectively assuming that the linear theory holds also in the nonlinear regime far from equilibrium. For a fluid description of the matter, the condition of the smallness of the bulk-viscous pressure ought to be satisfied[14] but in the nonlinear theory this constraint is not necessary[13] and it is not maintained without affecting the thermodynamical consistency of the theory.

To see if a cosmological model inflates or not it is convenient to introduce the deceleration parameter q,

$$q=-H^{-2}(\dot{H}+H^2)=\frac{d}{dt}\left(\frac{1}{H}\right)-1. \tag{13}$$

The positive sign of the deceleration parameter corresponds to standard decelerating models whereas the negative sign indicates inflation.

By using the gravitational field equations (3) and (4) it is easy to check that the deceleration parameter can be expressed in the form

$$q=\frac{\rho+3p+3\Pi}{2\rho}. \tag{14}$$

With these assumptions and supposing that locally and for a short interval of time t so that $t_i\leqslant t\leqslant t_{i+1}$ and $|t_{i+1}-t_i|\ll 1$, $i=1,...,n$ the dynamics of the bulk-viscous fluid is linear, $\chi(\Pi)\approx-\Pi/\xi$, the field equations, and the causal evolution equation for the bulk viscosity lead to the following evolution equation[14] for the Hubble factor H:

$$\ddot{H}+\tfrac{3}{2}[1+(1-r)\gamma]H\dot{H}+3^{1-s}\alpha^{-1}H^{2-2s}\dot{H}-(1+r)H^{-1}\dot{H}^2$$
$$+\tfrac{9}{4}(\gamma-2)H^3+\tfrac{1}{2}3^{2-s}\alpha^{-1}\gamma H^{4-2s}=0. \tag{15}$$

We shall consider exact solutions of Eq. (15) corresponding to a bulk-viscous fluid with a bulk viscosity coefficient proportional to the square root of the density or to the Hubble factor, that is we choose $s=1/2$. In this case Eq. (15) takes the form:

$$\ddot{H}+\left\{\frac{3}{2}[1+(1-r)\gamma]+\frac{\sqrt{3}}{\alpha}\right\}H\dot{H}-(1+r)H^{-1}\dot{H}^2+\frac{9}{4}\left[\left(1+\frac{2}{\sqrt{3}\alpha}\right)\gamma-2\right]H^3=0. \tag{16}$$

By means of the transformations[17]

$$H^2=y,\quad \eta=\frac{3}{2}\left[1+(1-r)\gamma+\frac{2}{\sqrt{3}\alpha}\right]\int H dt \tag{17}$$

we obtain the equation

$$\frac{d^2y}{d\eta^2}+\frac{dy}{d\eta}-\frac{1+r}{2}\frac{1}{y}\left(\frac{dy}{d\eta}\right)^2+\frac{2b}{1-r}y=0, \tag{18}$$

where we have denoted

$$b=\frac{(1-r)\left[\left(1+\frac{2}{\sqrt{3}\alpha}\right)\gamma-2\right]}{\left[1+(1-r)\gamma+\frac{2}{\sqrt{3}\alpha}\right]^2}.$$

We shall consider in the following that r and α are arbitrary independent (positive) parameters (independent of γ), so we will not impose any special conditions on the range of values of the parameter b, but we shall consider both cases $b=1/4$ and $b=2/9$ in the present paper.

With the use of the mathematical substitution $dy/d\eta=y^{(1+r)/2}/w$, Eq. (18) is transformed into the following (first kind) Abel-type first-order differential equation for the unknown function w:

$$\frac{dw}{dy}-y^{-(1+r)/2}w^2-\frac{2b}{1-r}y^{-r}w^3=0. \tag{19}$$

By introducing a new independent variable $z=2/(1-r)y^{(1-r)/2}$, $r\neq 1$, Eq. (19) becomes

$$\frac{dw}{dz}-w^2-bzw^3=0. \tag{20}$$

By means of the substitution $w=-(1/\psi)\ (d\psi/dz)$ Eq. (20) takes the form of an Euler-type linear second-order differential equation:

$$\psi^2\ \frac{d^2z}{d\psi^2}+bz=0 \tag{21}$$

and has the general solutions

$$z=C_1\psi^{m_1}+C_2\psi^{m_2},\quad b<\tfrac{1}{4}, \tag{22}$$

$$z=(C_3+C_4\ \ln|\psi|)\psi^{1/2},\quad b=\tfrac{1}{4}, \tag{23}$$

where C_i, $i=1,2,3,4$ are constants of integration and we have denoted $m_1=(1+\sqrt{1-4b})/2$ and $m_2=(1-\sqrt{1-4b})/2$. So, by an appropriate choice of values of integration constants we obtain the following two classes of particular solutions to Eq. (20):

$$w_i^{(p)}=-\frac{1-r}{2m_i}\,y^{-(1-r)/2},\quad i=1,2\ \text{ and }\ b<\frac{1}{4}, \tag{24}$$

$$w_3^{(p)}=-(1-r)y^{-(1-r)/2},\quad b=\tfrac{1}{4}. \tag{25}$$

We shall consider now that Eq. (19) has two general solutions $w_i(y)$, $i=1,2$ which can be generated from the particular solutions obtained above by means of the transformation

$$w_i(y)=\frac{E_i(y)}{u_i(y)}+w_i^{(p)},\quad i=1,2,3 \tag{26}$$

(usual in the theory of the Abel-type differential equations), with

$$E_i(y)=y^{[(1-r)/m_i](3b/2m_i-1)},\quad i=1,2\ \text{ and }\ b<\tfrac{1}{4}, \tag{27}$$

$$E_3(y)=y^{(r-1)/2},\quad b=\tfrac{1}{4} \tag{28}$$

with $u_i(y)$, $i=1,2,3$ some arbitrary functions to be determined and satisfying the following first-order differential equations:

$$\frac{du_i}{dy}+\frac{By^{M_i}}{u_i}+\omega_i y^{(1/2)(M_i-1)}=0,\quad i=1,2\ \text{ and }\ b<\frac{1}{4}, \tag{29}$$

$$\frac{du_3}{dy}+\left(\frac{1}{(1-r)u_3}-1\right)\frac{1}{2y}=0,\quad b=\frac{1}{4}, \tag{30}$$

where we have denoted $B=2b/(1-r)$, $\omega_i=1-(3b/m_i)$, and $M_i=-r+[2(1-r)/m_i](3b/2m_i-1)$, $i=1,2$.

Equations (29) and (30) represent the basic equations describing the dynamics of our bulk-viscous fluid-filled FRW Universe in the framework of the full ISH theory.

III. THE FIRST CLASS OF SOLUTIONS

We shall start our study of bulk-viscous Universes in the causal ISH theory by considering that the parameter b entering in the evolution equations satisfies the condition $b<1/4$. In this case the equations (29) describing the evolution and dynamics of a bulk viscous fluid-filled FRW Universe contain two constants denoted M_i and ω_i, $i=1,2$. In the present section we shall

consider some classes of exact general solutions of the field equations obtained by imposing some mathematical conditions on these parameters. A possible and simple choice leading to an exact analytical solution of equations (29) can be obtained if we choose the value of parameter b so that the condition

$$\omega_i = 1 - \frac{3b}{m_i} = 0, \quad i = 1,2 \tag{31}$$

holds. Equation (31) is satisfied for $b=0$ and $b=2/9$. The case $b=0$ will not be considered here. The unique value of m_i, $i=1,2$ corresponding to $b=2/9$ and satisfying Eq. (31) is $m=m_1=(1+\sqrt{1-4b})/2=2/3$. In view of the definition of b, taking into account that $b=2/9$ and denoting $f=1+(1-r)\gamma$, we obtain the following algebraic parametric compatibility relation between the positive constants α, γ, and r:

$$\left(\frac{2f^2}{9} - f + 3 - 2r\right)\alpha^2 + \frac{2}{9\sqrt{3}}(9-5f)\alpha + \frac{8}{27} = 0. \tag{32}$$

The condition of α being real leads to the following condition for r:

$$r > 1. \tag{33}$$

For these values of parameters it follows from (29) that the function $u=u_1$ satisfies the equation:

$$\frac{du}{dy} + \frac{4}{9(1-r)}\frac{y^{(r-3)/2}}{u} = 0. \tag{34}$$

The general solution of equation (34) can be easily expressed as

$$y = \left[\frac{9(r-1)^2}{16}u^2 - K\right]^{2/(r-1)}, \tag{35}$$

where K is an arbitrary constant of integration. Using (35), (27), and (26) we obtain the following class of exact solutions of Eq. (19), corresponding to our particular choice of parameter b:

$$w(u) = \left(\frac{9(r-1)^2}{16}u^2 - K\right)\left[\frac{\left(\frac{9(r-1)^2}{16}u^2 - K\right)^{1/2}}{u} + \frac{3}{4}(r-1)\right]. \tag{36}$$

By introducing a new parameter $\nu > 0$ by means of the transformation

$$\nu = \left[\frac{9(r-1)^2}{16}u^2 - K\right]^{-1/(r-1)} \tag{37}$$

we can express the first classes of solutions of the gravitational field equations for a bulk-viscous fluid-filled FRW Universe with bulk viscosity coefficient proportional to the Hubble factor and with the parameter $b=2/9$ in the following exact parametrical form [here we have written σ in (38) for the variable of integration to distinguish it from the independent variable ν].

$$t - t_0 = \frac{1-r}{\beta_1}\int_{\nu_0}^{\nu}[(1+K\sigma^{r-1})^{-1/2} + 1]\,d\sigma; \tag{38}$$

$$H = \frac{1}{\nu}, \tag{39}$$

$$a = a_0 \nu^{2(1-r)/\beta_1}[1 + (1+K\nu^{r-1})^{1/2}]^{2/\beta_1}, \tag{40}$$

$$\rho=\frac{3}{\nu^2}, \tag{41}$$

$$p=\frac{3(\gamma-1)}{\nu^2}, \tag{42}$$

$$\xi=\frac{\sqrt{3}\alpha}{\nu}, \tag{43}$$

$$T=\frac{3^r\beta}{\nu^{2r}}, \tag{44}$$

$$\tau=\frac{\alpha}{\sqrt{3}}\,\nu, \tag{45}$$

$$q=\frac{\beta_1}{1-r}\,\frac{1}{(1+K\nu^{r-1})^{-1/2}+1}-1, \tag{46}$$

$$\Delta\Sigma=-S_0\int_{\nu_0}^{\nu}\sigma^{2r-3+[6(1-r)/\beta_1]}[1+(1+K\sigma^{r-1})^{1/2}]^{6/\beta_1}\times\left[\frac{3\gamma(r-1)[(1+K\sigma^{r-1})^{-1/2}+1]}{2\beta_1}+1\right]d\sigma, \tag{47}$$

where a_0, t_0, and $\Sigma(\nu_0)$ are constants of integration and we have denoted $S_0=2*3^{1-r}a_0^3/\beta k_B$, $\Delta\Sigma=\Sigma(\nu)-\Sigma(\nu_0)$, and $\beta_1=f+2/\alpha\sqrt{3}$.

IV. THE SECOND CLASS OF SOLUTIONS

We shall consider now a second class of general solutions of the gravitational field equations describing a FRW Universe filled with a bulk-viscous cosmological fluid in the framework of the full causal ISH theory, generated by the general solution of Eq. (30) corresponding to the value 1/4 of the parameter b. In view of the definition of b and taking into account that $b=1/4$, we obtain the following algebraic parametric compatibility relation between the positive constants α, γ, and r:

$$\left(\frac{f^2}{4}-f+3-2r\right)\alpha^2+\left(\frac{2-f}{\sqrt{3}}\right)\alpha+\frac{1}{3}=0 \tag{48}$$

and we have also denoted $f=1+(1-r)\gamma$ in this case. The condition of α being real leads to the condition

$$r>1. \tag{49}$$

For the sake of simplicity we denote u_3 by u, $u_3=u$. The solution generating function $u(y)$ to be determined satisfies the following first-order differential equation:

$$\frac{du}{dy}+\left(\frac{1}{(1-r)u}-1\right)\frac{1}{2y}=0 \tag{50}$$

having the general solution

$$y=K^2e^{2u}[(r-1)u+1]^{2/(1-r)} \tag{51}$$

with K an arbitrary integration constant.

By using the mathematical substitution $F(u)=(r-1)u+1$, denoting $\Delta\Sigma=(\Sigma-\Sigma_0)/S_0$, $\beta_0=K(3f/2+\sqrt{3}/\alpha)$, $S_0=3^{1-r}a_0^3K^{2(1-r)}/(r-1)\beta k_B$, $F_0=(r-1)u_0+1$, and taking $u>0$ as pa-

rameter we can express the general solution of the gravitational field equations for a FRW space–time filled with a causal bulk-viscous cosmological fluid with bulk viscosity coefficient proportional to the Hubble factor in the following exact parametrical form:

$$t-t_0=\frac{2(r-1)}{\beta_0}\int_{u_0}^{u}e^{-\sigma}F^{1/(r-1)}(\sigma)d\sigma, \tag{52}$$

$$H=Ke^{u}F^{1/(1-r)}(u), \tag{53}$$

$$a=a_0e^{2(r-1)Ku/\beta_0}, \tag{54}$$

$$\rho=3K^2e^{2u}F^{2/(1-r)}(u), \tag{55}$$

$$p=3(\gamma-1)K^2e^{2u}F^{2/(1-r)}(u), \tag{56}$$

$$\xi=\alpha\sqrt{3}Ke^{u}F^{1/(1-r)}(u), \tag{57}$$

$$T=\beta 3^rK^{2r}e^{2ur}F^{2r/(1-r)}(u), \tag{58}$$

$$\tau=\frac{\alpha}{\sqrt{3}K}e^{-u}F^{-1/(1-r)}(u), \tag{59}$$

$$q=\frac{\beta_0}{2(1-r)K}(F^{-1}(u)-1)-1, \tag{60}$$

$$\begin{aligned}\Delta\Sigma=\beta_2^{-3}&[e^{\beta_2(F-1)}[\beta_3\beta_2^2F^2+2(\beta_2+\beta_3)(1-\beta_2F)]\\&-e^{\beta_1(F_0-1)}[\beta_3\beta_2^2F_0^2+2(\beta_2+\beta_3)(1-\beta_2F_0)]],\end{aligned} \tag{61}$$

where t_0, a_0, and u_0 are constants of integration and we have denoted $\beta_2=2[(3K/\beta_0)-1]$ and $\beta_3=2\{1+[3\gamma(r-1)K]/\beta_0\}$.

V. SOME PHYSICAL ASPECTS

The behavior of the dissipative causal bulk-viscous cosmological solutions described by Eqs. (38)–(47) and (52)–(61) essentially depends on the considered range of values of the parameters ν and u. For the first class of solutions (38)–(47) it is obvious that the parameter ν cannot be negative because $\nu<0$ would lead to $H(\nu)<0$ and consequently to $\xi(\nu)<0,\forall\nu<0$. To obtain a thermodynamically consistent cosmological evolution it is necessary that $\nu\geqslant 0$. For the first solution the direction (arrow) of time is inverse to the sense of variation of the parameter ν: to increasing time corresponds a decreasing parameter. In the limit of small values of the time $t\to 0$, corresponding to $\nu\to\infty$, we have from Eq. (37) that $\nu\to\infty$ for $u\to 4\sqrt{K}/3(r-1)$ leading to $y\to 0$ and consequently to $H\to 0$ and from Eq. (40) to $a\to 0$ for $t\to 0$. The Universe described by the first solution starts its evolution from a state of zero energy and a singularity in the space–time geometry. This model implies the appearance of the Universe from the large quantum fluctuation of a quantum vacuum (from ''nothing'') which must be present prior to the existence of the particles and matter in the space–time. But for $a\to 0$ due to the presence of large quantum effects only the quantum cosmology can provide a coherent description of the dynamics of the Universe. So we shall consider that the description of the evolution of the Universe according to the classical Einstein gravitational equations begins only at a large enough scale factor a when due to a specific particle content of the early cosmological fluid, bulk-viscous type dissipative processes begin to play an important role and consequently, we shall restrict the range of variation of the parameter ν by introducing an upper finite limit $\nu=\nu_m<\infty$ leading to a history of the Universe starting at a time $t=t_0>0$ (corresponding to the value $\nu=\nu_m$ of the parameter) from a nonsingular state, characterized by constant and finite values of the scale factor $a(t=t_0)=a_0\neq 0$, the energy density $\rho(t=t_0)=\rho_0\neq 0$, and consequently with constant nonzero pressure, temperature, and bulk viscosity coefficient. The solution is free of singularity in the entire range $t\geqslant t_0$.

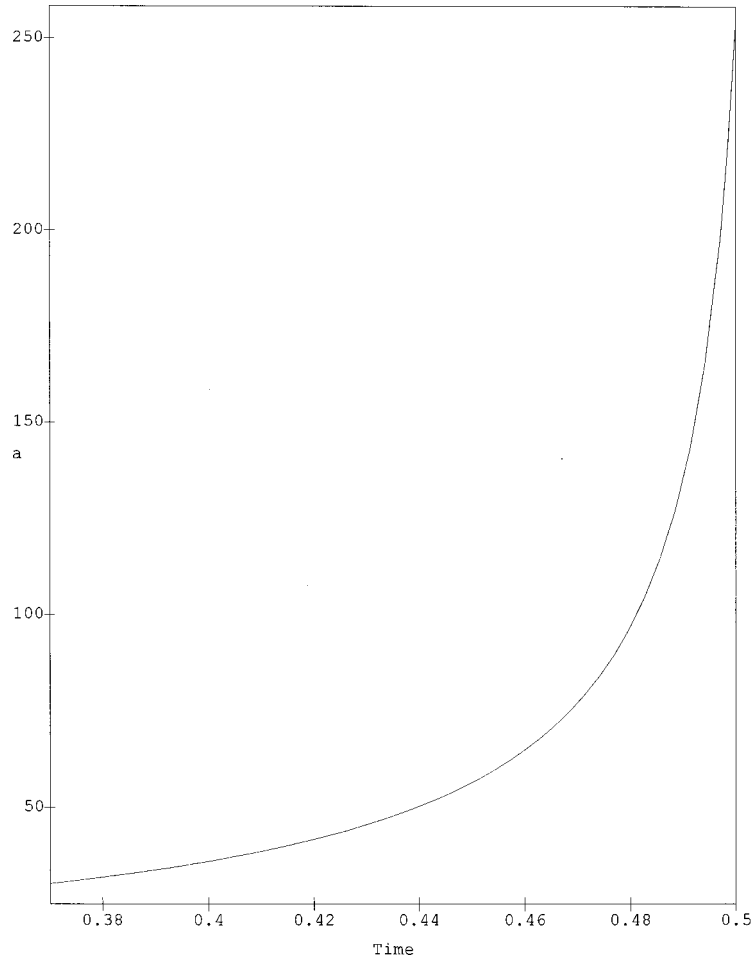

FIG. 1. Variation of the scale factor against the time for $b=2/9$, $r=1.3$, and $\gamma=4/3$.

The Universe shows an inflationary behavior with $q<0$ for all $t>t_0$ so this type of causal bulk-viscous Universe will experience an inflationary era from the beginning of its classical period of existence and its dynamics is reduced to a very rapidly expanding phase. During this period of evolution of the Universe the Hubble factor, energy density, temperature, and bulk viscosity coefficient are rapidly increasing. The Universe enters into a period of accelerated heating and is dominated by thermal processes. A large amount of comoving entropy is produced as a result of dissipative processes. As this thermal Universe expands, the intensity of the bulk-viscous type processes is explosively increasing. The time dependence of the scalar factor a, the energy density ρ, and the deceleration parameter q for the first class of solutions are represented in Figs. 1–3.

In order to obtain the behavior of the Universe in the large time limit $t\to\infty$ ($\nu\to 0$) we observe from Eq. (37) that $\nu\to 0$ in the limit $u\to\infty$ and consequently we obtain from Eq. (26) and for $b=2/9$ that $\lim_{u\to\infty} w=w^{(p)}=[3(r-1)/4]y^{(r-1)/2}$. So the particular solution (24) generating the first general solution (38)–(47) of the gravitational field equations represents the solution of the causal bulk-viscous cosmological fluid in the limit of small parameter ν, corresponding to large times t, and describing the large time dynamics of the model. In this limit we immediately obtain from the field equations

$$H=-\frac{(r-1)}{[1+(1-r)\gamma+2/\sqrt{3}\alpha]}\frac{1}{t}<0$$

and $\xi(t)<0$, $t\to\infty$ leading to a negative bulk viscosity coefficient and consequently a thermodynamically consistent cosmological model can be obtained from the first solution only for positive

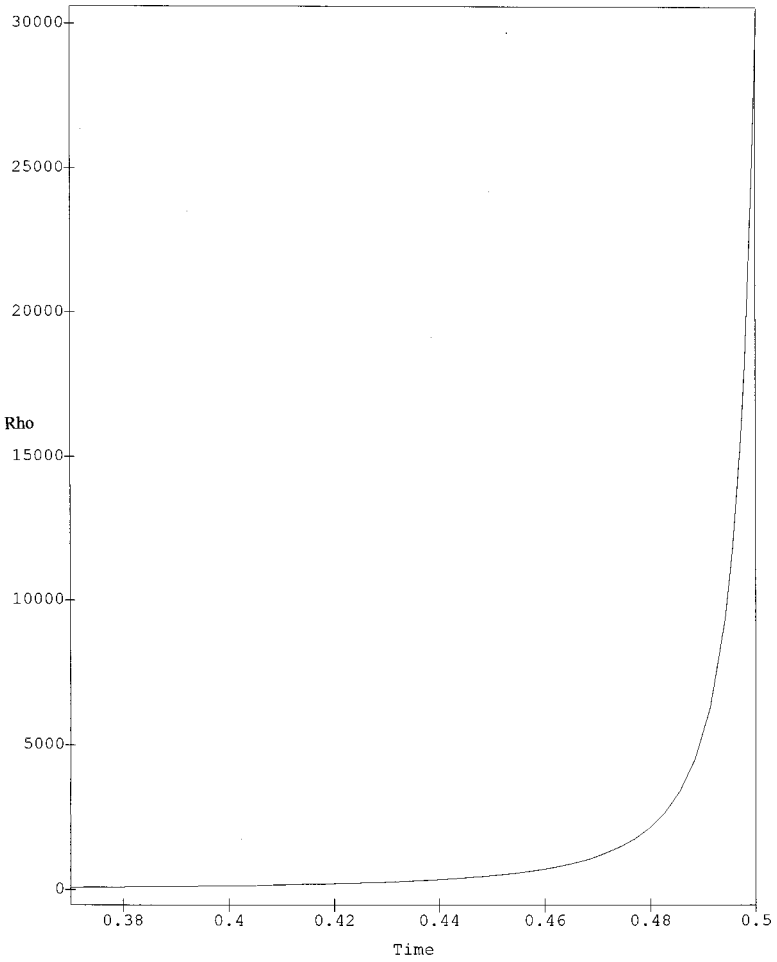

FIG. 2. Variation of the energy density against the time for $b=2/9$, $r=1.3$, and $\gamma=4/3$.

nonzero values of the parameter ν, $\nu>\nu_0>0$. So, a physically consistent evolution of the Universe described by the first solution is obtained only for a finite period of time $t_0\leqslant t\leqslant t_c$, corresponding to a finite range of variation of the parameter $\nu\in[\nu_0,\nu_m]$.

The second class of solutions (52)–(61) leads to physically consistent behavior of the corresponding cosmological model and is defined for all values of r satisfying the condition (49), $r>1$, only for positive values of the parameter u, $u\geqslant 0$. For negative values of the parameter u and for some values of r ($r=1.4$, for example) the time function (52) has imaginary values and for negative u and some classes of r values the Hubble factor $H(u)$ becomes negative leading to a negative bulk viscosity coefficient and consequently, we shall consider the behavior of this class of solutions only in the range of the positive values of the parameter $u\in[0,\infty)$. The Universe starts its evolution at the moment $t=t_0$ (corresponding to the value $u=0$ of the parameter) from a nonsingular state characterized by constant and finite values of the scale factor $a(t=t_0)=a_0\neq 0$, the energy density $\rho(t=t_0)=\rho_0\neq 0$, and consequently with constant nonzero pressure, temperature, and bulk viscosity coefficient. The solutions are free of singularity in the entire range $t_0\leqslant t<\infty$. From the beginning the evolution of the Universe is inflationary and is associated with a rapid increase of the energy density, temperature, and bulk viscosity coefficient. The Universe is heated due to the action of the dissipative processes and a large amount of comoving entropy is produced. In the large time limit the Universe will end in a state characterized by constant values of the scale factor, energy density, temperature, and comoving entropy and its inflationary evolution is stopped. So, this solution can describe the evolution of the Universe only for a finite interval of time $t_0\leqslant t\leqslant t_c$ thus showing a very similar evolution with the Universe described by the first class of solutions. The time dependence of the scalar factor a, the energy density ρ, the

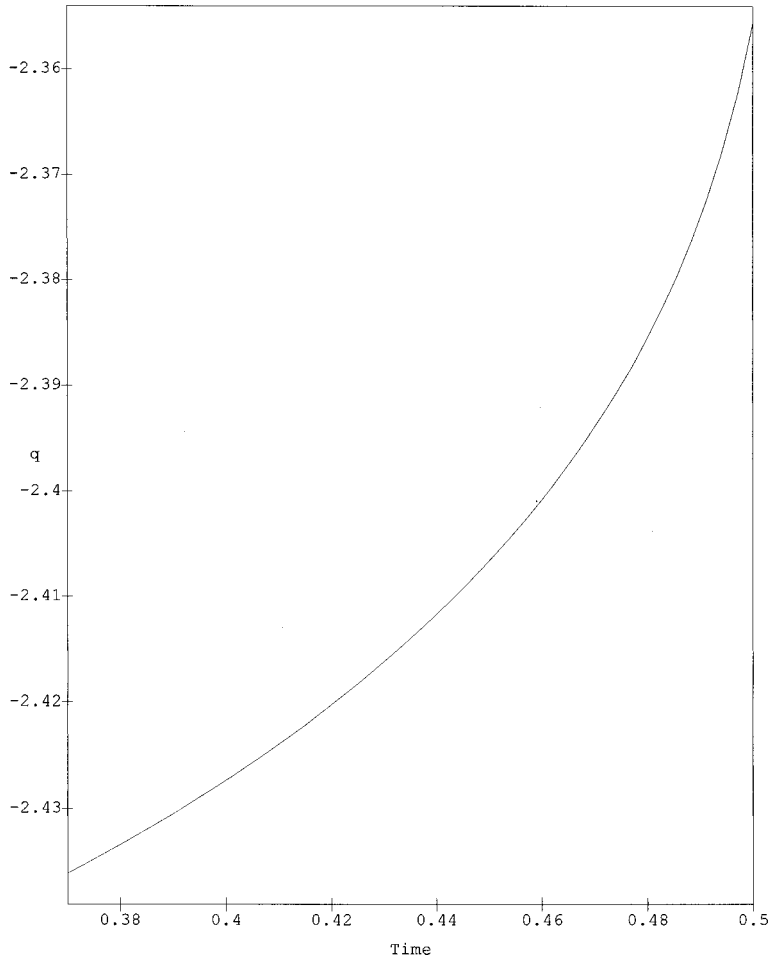

FIG. 3. Variation of the decelerating parameter against the time for $b=2/9$, $r=1.3$, and $\gamma=4/3$.

deceleration parameter q, the entropy Σ, and the Kretschmann scalar $R_{ijkl}R^{ijkl}$ for the second class of solutions are represented in Figs. 4–8.

The nonsingular character of the solutions in the range $t_0 \leqslant t \leqslant t_c$ is consistent with the finite character of the invariants $R_{ij}R^{ij}$ and $R_{ijkl}R^{ijkl}$ for the case of $b=1/4$.

$$R_{ij}R^{ij}=12\left[\left(\frac{\ddot{a}}{a}\right)^2+\left(\frac{\ddot{a}}{a}\right)\left(\frac{\dot{a}}{a}\right)^2+\left(\frac{\dot{a}}{a}\right)^4\right] \tag{62}$$

$$=12(Ke^uF^{1/(1-r)})^4\left[\left[\frac{\beta_0}{2(1-r)K}\,(F^{-1}-1)\right]^2+\frac{3\beta_0}{2(1-r)K}\,(F^{-1}-1)+3\right], \tag{63}$$

$$R_{ijkl}R^{ijkl}=12\left[\left(\frac{\ddot{a}}{a}\right)^2+\left(\frac{\dot{a}}{a}\right)^4\right] \tag{64}$$

$$=12(Ke^uF^{1/(1-r)})^4\left[\left[1+\frac{\beta_0}{2(1-r)K}\,(F^{-1}-1)\right]^2+1\right]. \tag{65}$$

We have also obtained the invariants for the first solution ($b=2/9$) after a tedious but straightforward calculation. We shall not present them here in order to avoid lengthy and cumbersome expressions. These invariants are all finite for all values of the time $t \geqslant t_0$. The presence of a heating period in the framework of the causal viscous theory has been already mentioned in the literature[13,16] and it is always associated with the presence of an inflationary era. The association

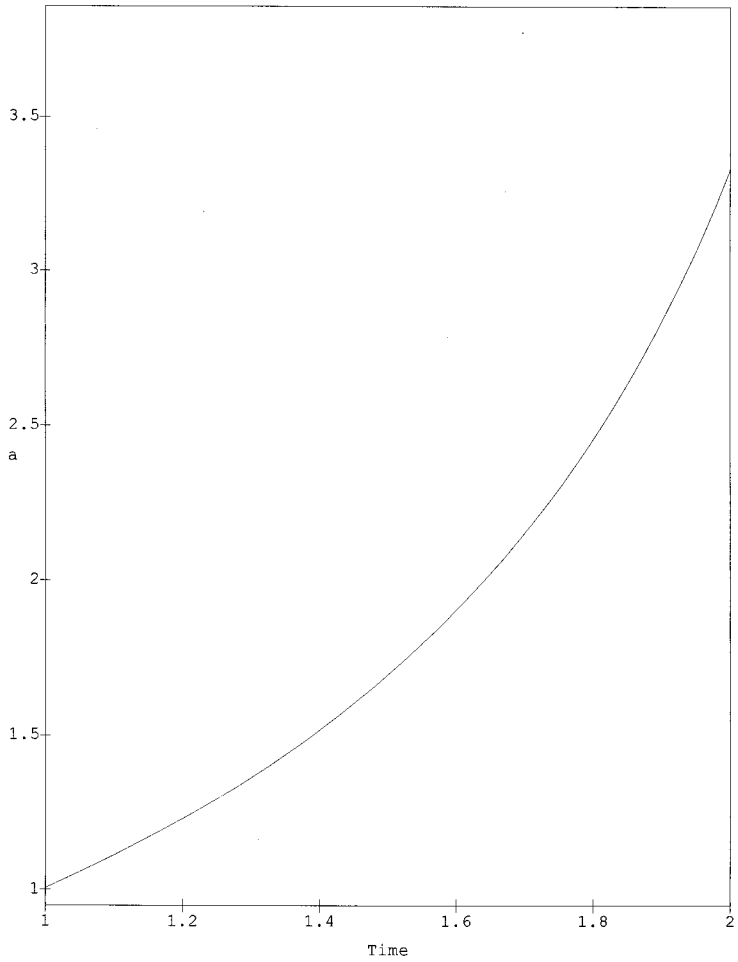

FIG. 4. Variation of the scale factor against the time for $b=1/4$, $r=1.3$, and $\gamma=4/3$.

between heating period, entropy production, and inflation is a specific characteristic of the general solution of the gravitational field equations for a bulk-viscous cosmological fluid with a bulk viscosity coefficient proportional to H and $0\leqslant r\leqslant 1/2$.

From Eq. (14) it follows that, because $q<0$ for all $t\geqslant t_0$,

$$\rho+3p_{\text{eff}}<0 \quad \text{for all} \quad t\geqslant t_0=0 \tag{66}$$

(where $p_{\text{eff}}=p+\Pi$) and

$$\rho+3p<3H^2[3\gamma-2(q+1)]. \tag{67}$$

For periods of time for which the condition

$$q>\frac{3\gamma}{2}-1>0 \quad \text{for all} \quad \gamma\in[1,2] \tag{68}$$

holds, the strong energy condition $\rho+3p>0$ is violated by causal bulk-viscous (noninflationary) cosmological models. On the other hand, from (14) it follows that during a causal bulk-viscous inflationary period $p\ll|\Pi|$ which clearly contradicts the condition of smallness of bulk-viscous pressure. In the nonlinear generalization of the full causal theory describing far from equilibrium evolution of a viscous cosmological fluid this condition no longer holds.

The solutions obtained in the present paper describe bulk-viscosity-driven inflationary Universes and are thermodynamically consistent only in the framework of the nonlinear generalization

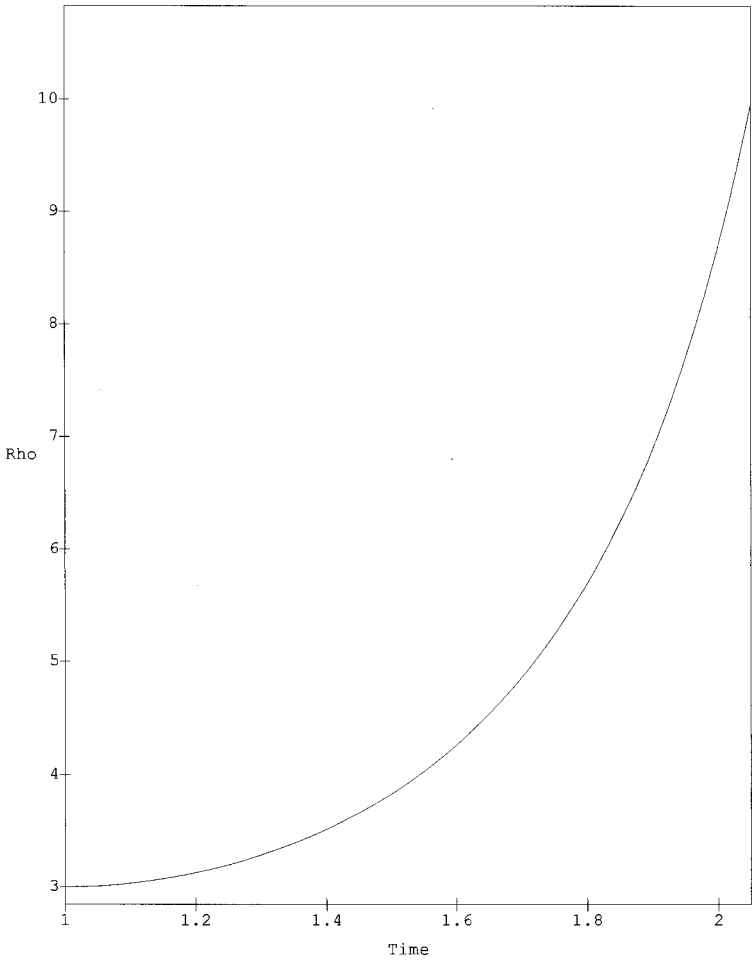

FIG. 5. Variation of the energy density against the time for $b=1/4$, $r=1.3$, and $\gamma=4/3$.

of the linear causal thermodynamics or in our approximation only for periods of time for which $\xi_{\rm eff}^{(i)}\approx\xi$, $i=1,...,n$. In the case of the nonlinear theory of Maartens and Mendez[13] this relation can be expressed as $\xi_{\rm eff}^{(i)}(t)\approx(1+k_i^2(\Pi^{(i)}/\rho))\xi(t),\forall t\in[t_i,t_{i+1}]$, $i=1,...,n$. Consequently, the obtained solutions can describe far from equilibrium inflationary evolution only for periods of time for which the condition

$$k_i^2\,\frac{\Pi^{(i)}}{\rho}\ll 1,\quad i=1,...,n \tag{69}$$

is satisfied. So in this model by an appropriate choice of the constants k_i, $i=1,...,n$ entering in the expression of the nonlinear relaxation time $\tau_*^{(i)}=k_i^2\tau$ it is always possible to obtain (69) for successive short time intervals and the evolution of a nonlinear far from equilibrium causal bulk-viscous Universe can be reduced to a sequence of a large number of linear epochs, $n\to\infty$, so that in each era the nonlinear relaxation time τ_* is related to the linear one by means of different constants k_i, $i=1,...,n$, so that the equation (69) is satisfied during the corresponding bulk-viscous regime.

The causal bulk viscous inflationary solutions we have obtained in the present paper lead to a rapid increase of the energy density, temperature, and entropy of the cosmological fluid filling the early Universe due to the presence of intense dissipative processes. But these results can be interpreted also from the alternative point of view of another physical theory. In fact the cosmological behavior of our solutions is exactly what is expected from a theory describing particle production in the early Universe.

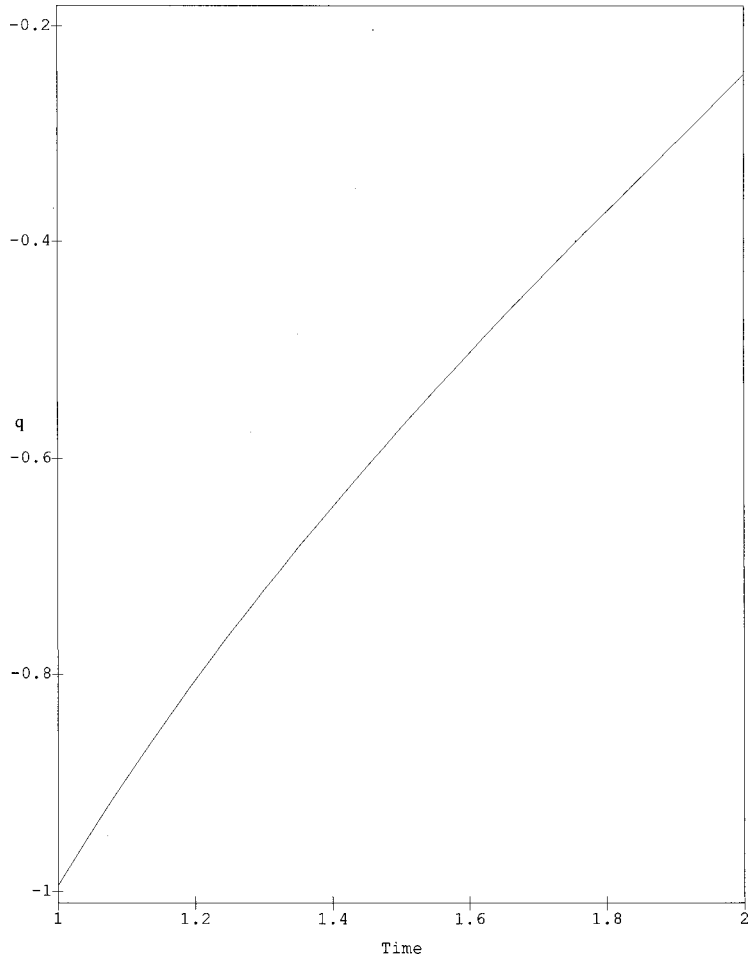

FIG. 6. Variation of the decelerating parameter against the time for $b=1/4$, $r=1.3$, and $\gamma=4/3$.

As was observed first by Zeldovich[18] and later on confirmed by Murphy[19] and Hu[20] a nonvanishing particle production rate is equivalent to a bulk-viscous pressure in the cosmological fluid or from a quantum point of view with a viscosity of the vacuum. This is due to the simple circumstance that any source term in the energy balance of a relativistic fluid may be formally rewritten in terms of an effective bulk viscosity. Barrow[21] has considered creation processes in string-driven inflation within a fluid model and concluded that this process may be described phenomenologically in terms of an effective bulk viscosity. From the point of view of the kinetic theory based on a Boltzmann type equation it follows that this simple phenomenological approach is compatible with the kinetic theory in homogeneous space–times but not in inhomogeneous ones.[22] Consequently if it is possible to describe particle production processes consistently by means of an effective viscous pressure one is able to study, at least on a simple classical phenomenological level, the impact of these processes on the early dynamics of the Universe.

The possibility that the bulk-viscous pressure of the full ISH theory may also be interpreted as an effective description for particle production processes has been considered in detail by Zimdahl.[16] The creation process leads to considerable changes in the thermodynamical behavior of the Universe and if the chemical potential of the newly created particles is zero, $\mu=0$, then the nonvanishing bulk pressure Π associated with an increase in the number of fluid particles satisfies formally the same equation (6) as in the case of presence of a real dissipative bulk viscosity.[16]

In order to analyze the equivalence between the full ISH theory and particle production we shall make the fundamental assumption that immediately after creation all newly created particles are (approximately) amenable to an adiabatic perfect fluid description; that is we suppose that the energy density of the created cosmological fluid obeys an equation of state of the form

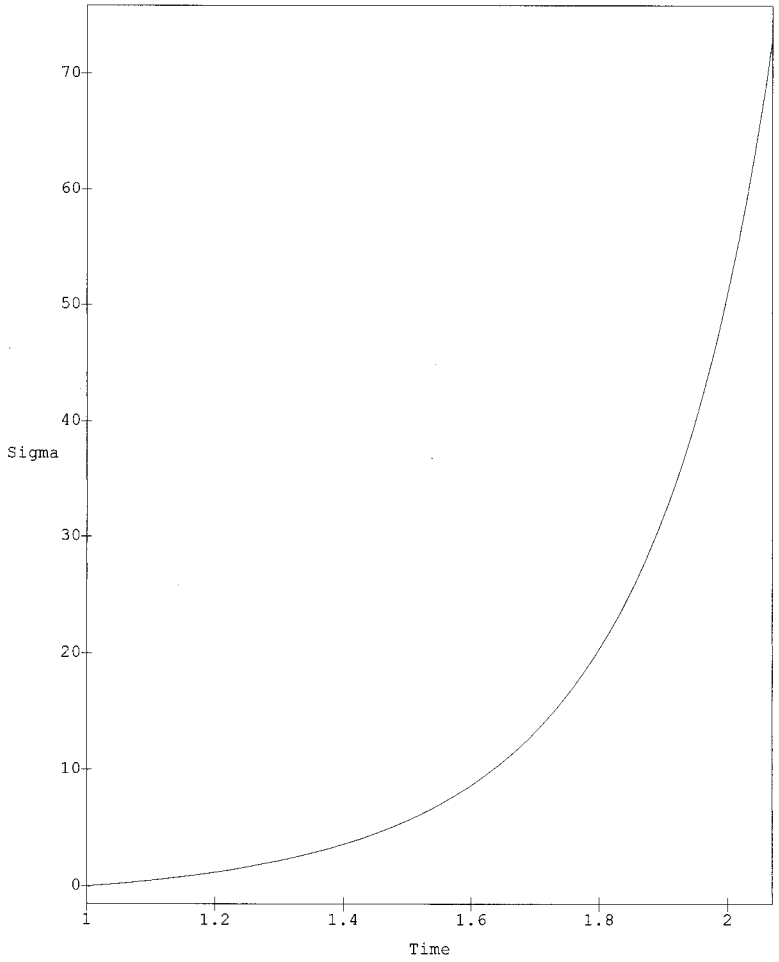

FIG. 7. Variation of the entropy $\Delta\Sigma=(\Sigma-\Sigma_0)/S_0$ against the time for $b=1/4$, $r=1.3$, and $\gamma=4/3$.

$$\rho=\rho_0\left(\frac{n}{n_0}\right)^{\gamma}=kn^{\gamma}, \tag{70}$$

where n is the particle number density and $k=\rho_0/n_0^{\gamma}>0$ is a constant. We suppose that the pressure is given by $p=(\gamma-1)\rho$. By using (70) Eq. (5) takes the form of a particle number balance equation

$$\dot{n}+3Hn=\Gamma n, \tag{71}$$

where

$$\Gamma=-\frac{\Pi}{\gamma H}=\frac{H}{\gamma}[3\gamma-2(q+1)]\geqslant 0 \tag{72}$$

is the particle production rate.

The condition of non-negativity of Γ implies $H>0$ (particle creation takes place only in an expanding Universe) and

$$q<\frac{3\gamma}{2}-1. \tag{73}$$

Condition (73) is automatically satisfied by all inflationary Universes (with $q<0$).

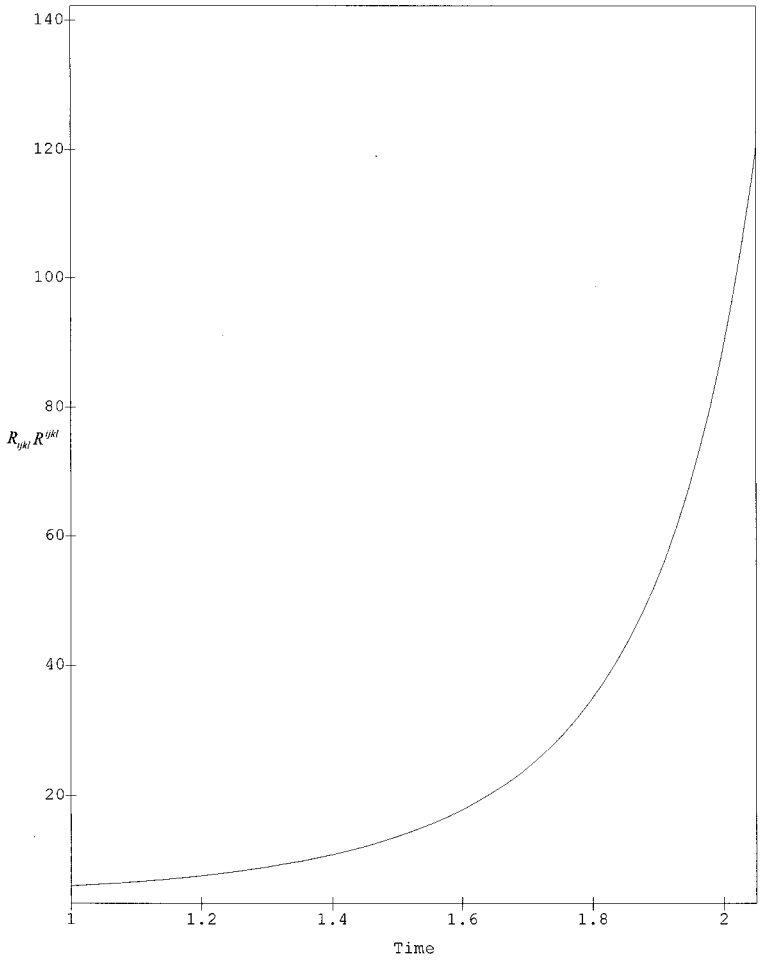

FIG. 8. Variation of the Kretschmann scalar $R_{ijkl}R^{ijkl}$ against the time for $b=1/4$, $r=1.3$, and $\gamma=4/3$.

From the causal evolution equation of the bulk-viscous pressure (6) it follows that the dynamics of the particle production rate is described in the general nonlinear case by the following linear first-order differential equation:

$$\tau\dot{\Gamma}+\left[\left(\frac{\dot{H}}{H}+\frac{1}{2}\left(3H+\frac{\dot{\tau}}{\tau}-\frac{\dot{\xi}}{\xi}-\frac{\dot{T}}{T}\right)\right)\tau\right]\Gamma=-\frac{1}{\gamma}\frac{\chi(-\gamma H\Gamma)}{H}+\frac{s_0}{\gamma^2}\left(\frac{3}{k}\right)^{1/\gamma}\xi TH^{2(1/\gamma-1)}, \quad (74)$$

where s_0 is the (constant) entropy of the newly produced particles. The entropy production, entirely due to particle creation, is given by $S^i_{;j}=-(\Pi/T)\chi(\Pi)>0$. In the case of a linear bulk-viscous regime (which we consider that can model inflationary viscous far from equilibrium dynamics for small time intervals) we obtain for the particle creation rate the evolution equation

$$\tau\dot{\Gamma}+\left[1+\left(\frac{\dot{H}}{H}+\frac{1}{2}\left(3H+\frac{\dot{\tau}}{\tau}-\frac{\dot{\xi}}{\xi}-\frac{\dot{T}}{T}\right)\right)\tau\right]\Gamma=\frac{s_0}{\gamma^2}\left(\frac{3}{k}\right)^{1/\gamma}\xi TH^{2(1/\gamma-1)}. \quad (75)$$

So a causal bulk viscous inflationary cosmological model is mathematically equivalent with a particle creation model with particle production rate satisfying Eqs. (72) or (75). The particle creation rate is a complicated function of the Hubble factor, temperature, relaxation time, and of a generalized bulk-viscosity coefficient related to the bulk-viscous pressure (that plays the role of a ''creation pressure''). During the particle creation era the temperature of the newly created particle is described by a law of the form $T\sim n^{\gamma r}$ which taking into account that r has arbitrary values leads in the present model to a different thermal behavior of a gas of particles during creation comparatively to an ideal gas.[16]

VI. CONCLUSIONS AND FINAL REMARKS

In the present paper we have presented two new classes of solutions of the Einstein gravitational field equations for a causal bulk-viscous fluid-filled FRW space–time corresponding to a relaxation of the constraints imposed to the parameters entering in the physical mode. We have supposed that the linear model of Israel and Stewart[5,6] can accurately describe bulk viscous far from equilibrium inflation only for very short intervals of time and that for small periods it is a reasonable approximation to the more general nonlinear bulk-viscous type inflationary theories.[13] Both obtained solutions are inflationary, show a similar behavior, and lead to a rapid increase of the energy density, temperature, and comoving entropy of the Universe. This physical behavior naturally fits into the framework of the models describing particle creation in the early Universe and that is why we interpret them as describing on a phenomenological classical level, particle production in a homogeneous FRW space–time (which in fact is essentially a quantum process). Due to the presence of a negative bulk-viscous type particle creation pressure the strong energy condition is violated in the present models. We suppose that in our solutions the entire inflation driving bulk-viscous pressure is due to particle production on a classical level, the effect of the ''real'' viscosity being neglected. With this physical interpretation the inflationary classes of solutions which frequently appear in near- or far- from-equilibrium bulk-viscous cosmologies can be used to study the essential influence of particle creation on the dynamics of the Universe which is always modified compared with the standard FRW behavior as long as there is a substantial change in the number of relevant particle species. On the other hand, inflation and particle creation have been analyzed mainly in space–times with de Sitter or power-law type scale factors. The solutions presented in this paper show that inflation, a general and very attractive feature of many cosmological models, can be described mainly by complicated mathematical models.

One of the most attractive features of the inflationary scenarios is that they might account for the remarkable large scale homogeneity of the Universe as evidenced by the isotropy of the cosmic microwave background. The observed structures consisting of galaxies, clusters of galaxies, etc., are assumed to have developed from initial small fluctuations and one of the key problems in the theory of the galaxy formation is the explanation of a sufficient growth of density inhomogeneities about a homogeneous and isotropic background. Cosmological perturbations in viscous cosmological fluids, having both bulk and shear viscosity have been considered by Zimdahl[23] using the fluctuation dissipation theorems of Landau–Lifshitz type while Zimdahl *et al.*[24] analyzed perturbations in a viscous Universe with particle production. In these papers noncausal thermodynamics has been used to describe dissipative processes. In the presence of particle creation it has been shown[24] that the gauge invariantly defined fractional adiabatic energy-density perturbations are unstable on large cosmic scales and they grow quadratically with the cosmic time. The gravitational stability against small perturbations of the cosmological models obtained in the present paper will be the subject of future research.

Whether one of these models actually describes a well-determined period in the very early dynamics and evolution of our Universe can only be decided once reliable knowledge about the quantum mechanisms that characterize particle or strings creation out of the quantum vacuum close to the Planck time becomes available from the underlying quantum physics.

ACKNOWLEDGMENTS

M. K. Mak is grateful to the Vocational Training Council, Hong Kong, China for continuous support. The authors would like to thank the anonymous referee for several comments and suggestions on an earlier version of the manuscript. Finally thanks to Mak's wife, Sandra, for her patience while Mak was buried in this paper.

[1] C. W. Misner, Phys. Rev. Lett. **19**, 533 (1966).
[2] L. P. Chimento and A. S. Jakubi, Phys. Lett. A **212**, 320 (1996).
[3] C. Eckart, Phys. Rev. **58**, 919 (1940).
[4] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics* (Butterworth Heinemann, London, 1987).
[5] W. Israel, Ann. Phys. (N.Y.) **100**, 310 (1976).
[6] W. Israel and J. M. Stewart, Phys. Lett. **A58**, 213 (1976).
[7] W. A. Hiscock and L. Lindblom, Ann. Phys. (N.Y.) **151**, 466 (1989).
[8] W. A. Hiscock and J. Salmonson, Phys. Rev. D **43**, 3249 (1991).
[9] L. P. Chimento and A. S. Jakubi, Class. Quantum. Grav. **14** 1811 (1997).
[10] L. P. Chimento and A. S. Jakubi, Int. J. Mod. Phys. D **7**, 177 (1998).

[11] L. P. Chimento, J. Math. Phys. **38**, 2565 (1997).
[12] L. P. Chimento, A. S. Jakubi, V. Mendez, and R. Maartens, Class. Quantum. Grav. **14**, 3363 (1997).
[13] R. Maartens and V. Mendez, Phys. Rev. D **55**, 1937 (1997).
[14] R. Maartens, Class. Quantum Grav. **12**, 1455 (1995).
[15] V. Romano and D. Pavon, Phys. Rev. D **47**, 1396 (1993).
[16] W. Zimdahl, Phys. Rev. D **53**, 5483 (1996).
[17] L. P. Chimento and A. J. Jakubi, Class. Quantum Grav. **10**, 2047 (1993).
[18] Ya. B. Zeldovich, JETP Lett. **12**, 307 (1970).
[19] G. L. Murphy, Phys. Rev. D **8**, 4231 (1973).
[20] B. L. Hu, Phys. Lett. A **90**, 375 (1982).
[21] J. D. Barrow, Nucl. Phys. B **310**, 743 (1988).
[22] J. Triginer, W. Zimdahl, and D. Pavon, Class. Quantum Grav. **13**, 403 (1996).
[23] W. Zimdahl, Phys. Rev. D **48**, 2431 (1993).
[24] W. Zimdahl, D. Pavon, and D. Jou, Class. Quantum Grav. **10**, 1775 (1993).

On a certain formulation of the Einstein equations

Paweł Nurowski[a)]
Dipartimento di Scienze Matematiche, Universita degli Studi di Trieste, Trieste, Italy

(Received 13 April 1998; accepted for publication 8 June 1998)

We define a certain differential system on an open set of $\mathbf{R}^6$. The system locally defines a Lorentzian four-manifold satisfying the Einstein equations. The converse statement is indicated and its details are postponed to the forthcoming paper.
© *1998 American Institute of Physics.* [S0022-2488(98)01310-3]

Let $\mathscr{U}$ be an open subset of $\mathbf{R}^6$. Suppose that on $\mathscr{U}$ we have six one-forms $(F,\bar{F},T,\Lambda,E,\bar{E})$ which satisfy the following conditions:

(i) T, Λ are real- and F, E are complex-valued one-forms;

(ii) $F\wedge F\wedge T\wedge\Lambda\wedge E\wedge E\neq 0$ at each point p of $\mathscr{U}$;

(iii) there exist complex-valued one-forms Ω and Γ on $\mathscr{U}$, and a certain complex function α on $\mathscr{U}$ such that

$$\begin{aligned} dF&=(\Omega-\bar{\Omega})\wedge F+E\wedge T+\bar{\Gamma}\wedge\Lambda,\\ dT&=\Gamma\wedge F+\bar{\Gamma}\wedge\bar{F}-(\Omega+\bar{\Omega})\wedge T,\\ d\Lambda&=\bar{E}\wedge F+E\wedge\bar{F}+(\Omega+\bar{\Omega})\wedge\Lambda,\\ dE&=2\Omega\wedge E+\bar{F}\wedge T+\alpha\Lambda\wedge F. \end{aligned} \tag{1}$$

The aim of this paper is to prove that we can associate to such $\mathscr{U}$ a Lorentzian four-manifold that satisfies the Einstein equations. The proof will be a direct consequence of the following sequence of Lemmas.

Lemma 1: If $(F,\bar{F},T,\Lambda,E,\bar{E})$ *satisfy* (1) *then there exist complex functions a, h on* $\mathscr{U}$ *and a real constant* λ *such that*

$$\begin{aligned} d\Gamma&=2\Gamma\wedge\Omega+\alpha T\wedge\bar{F}+a(T\wedge\Lambda+F\wedge\bar{F})+h\Lambda\wedge F,\\ d\Omega&=E\wedge\Gamma-\left(\alpha+\frac{\lambda}{2}\right)(T\wedge\Lambda+F\wedge\bar{F})+a\Lambda\wedge F. \end{aligned} \tag{2}$$

Sketch of the proof: Condition (iii) can be understood as follows. We consider the forms $(F,\bar{F},T,\Lambda,E,\bar{E})$ as the basic objects. We assume that their differentials have a special form (iii) with certain auxiliary forms $(\Omega,\bar{\Omega},\Gamma,\bar{\Gamma})$. Suppose that we have the system of ten forms $(F,\bar{F},T,\Lambda,E,\bar{E},\Omega,\bar{\Omega},\Gamma,\bar{\Gamma})$ satisfying (1). Then

$$d^2\Lambda=d^2T=d^2F=d^2E=0 \tag{3}$$

imply that the differentials $d\Gamma$ and $d\Omega$ have a special form. The proof that $d\Gamma$ and $d\Omega$ have the form (2) is a direct (but lengthy) calculation which uses (3) and the independence condition (ii). To obtain information about the differentials of the functions appearing in (2) one has to use

$$d^2\Gamma=d^2\Omega=0. \tag{4}$$

These equations show, in particular, that λ must be a real constant. □

[a)]Permanent address: Instytut Fizyki Teoretycznej, Wydzial Fizyki, Uniwersytet Warszawski, ul. Hoża 69, Warszawa, Poland. Electronic mail: nurowski@fuw.edu.pl

 © 1998 American Institute of Physics

There are also other equations that are implied by (3) and (4). They carry information about the differentials of α, a, h and about the decompositions of Γ and Ω onto the basis one-forms $(F,\overline{F},T,\Lambda,E,\overline{E})$. Explicitly we have

$$\begin{aligned} d\alpha &= \alpha_1 F + \gamma_4 \overline{F} + \gamma_1 T + \alpha_4 \Lambda - 2aE, \\ da &= a_1 F + \alpha_4 \overline{F} + \alpha_1 T + a_4 \Lambda + hE - (3\alpha + \lambda)\Gamma - 2a\Omega, \\ dh &= h_1 F - a_4 \overline{F} - a_1 T + h_4 \Lambda + 4a\Gamma - 4h\Omega, \end{aligned} \tag{5}$$

where the possible forms of Ω and Γ are

$$\begin{aligned} \Omega &= \omega_1 F + \omega_2 \overline{F} + \omega_3 T + \omega_4 \Lambda, \\ \Gamma &= \gamma_1 F - 4\omega_4 \overline{F} - 4\omega_1 T + \gamma_4 \Lambda - (3\alpha + \lambda)E \end{aligned} \tag{6}$$

and $\gamma_1, \gamma_4, \alpha_1, \alpha_4, a_1, a_4, h_1, h_4, \omega_1, \omega_2, \omega_3, \omega_4$ are complex functions on $\mathscr{U}$. The differentials of these functions can be further analyzed by looking at equations such as $d^2\alpha = 0$, etc.

Reformulating the above statements in the language of the theory of differential systems we say that the differential system (1), (2), (5), (6) is still not closed. For example, the equations of the sort $d^2\alpha = 0$ should still be added.

Lemma 2: $\mathscr{U}$ is locally foliated by two-dimensional manifolds $\mathscr{S}_x$, which are tangent to the real distribution $\mathscr{V}$ defined by

$$\Lambda(\mathscr{V}) = F(\mathscr{V}) = T(\mathscr{V}) = 0.$$

Here the proof is an immediate application of the Froebenius theorem, since the forms $(F,\overline{F},T,\Lambda)$ form a closed differential ideal due to the relations (1). □

Let us define

$$G = F\overline{F} - T\Lambda \tag{7}$$

on $\mathscr{U}$. G constitutes a degenerate metric on $\mathscr{U}$. It has the signature $(+,+,+,-,0,0)$.

Lemma 3: G is constant along any leaf $\mathscr{S}_x$ of the foliation $\{\mathscr{S}_x\}$.

To prove this we define a basis of vector fields $(f,\overline{f},X,Y,e,\overline{e})$, which is the respective dual of $(F,\overline{F},T,\Lambda,E,\overline{E})$. Then we notice that the distribution $\mathscr{V}$ is spanned by vector fields of the form

$$V = Ue + \overline{U}\overline{e},$$

where U is any complex function on $\mathscr{U}$. Using an arbitrary V and the explicit form of G one easily finds that $\mathscr{L}_V G = 0$ due to Eq. (1) and the properties of the Lie derivative $\mathscr{L}_V$. □

Now we introduce an equivalence relation $\sim$ on $\mathscr{U}$ which identifies points lying on the same leave of $\{\mathscr{S}_x\}$. We assume that the quotient space $\mathscr{M} = \mathscr{U}/\sim$ is a four-manifold. According to Lemma 3, G projects down to a well-defined nondegenerate Lorentzian metric g on $\mathscr{M}$.

Theorem 1: *The Lorentzian metric g on $\mathscr{M} = \mathscr{U}/\sim$ satisfies the Einstein equations $R_{ij} = \lambda g_{ij}$ and is not conformally flat.*

Proof: To prove the theorem, we first consider a four-dimensional submanifold $\mathscr{M}'$ of $\mathscr{U}$ that is transversal to the leaves of $\{\mathscr{S}_x\}$. We have a natural inclusion $\iota: \mathscr{M}' \hookrightarrow \mathscr{U}$. $\mathscr{M}'$ may be equipped with a Lorentzian metric $g' = \iota^* G$. It is clear that $(\mathscr{M},g)$ and $(\mathscr{M}',g')$ are locally isometrically equivalent. This statement does not depend on the choice of $\mathscr{M}'$. Thus, we may represent $(\mathscr{M},g)$ by $(\mathscr{M}',g')$. In this way, it is enough to show that g' satisfies the Einstein equations. To do this, we first observe that $g' = \iota^*(F)\iota^*(\overline{F}) - \iota^*(T)\iota^*(\Lambda)$. Thus, a set of forms $\theta^i = (\iota^*(F), \iota^*(\overline{F}), \iota^*(T), \iota^*(\Lambda))$, $(i = 1,2,3,4)$ constitutes a null cotetrad for g'. To calculate the curvature we need to know $d\theta^i$. But these, due to the fact that $\iota^* d = d\iota^*$, are given by the relations (1). To simplify the notation, we will omit the signs of the pullback ι^* in all of the formulas. On doing that we find that the connection one-forms for g' determined by

$$dF=-\Gamma^1{}_1\wedge F-\Gamma^1{}_3\wedge T-\Gamma^1{}_4\wedge\Lambda,$$

$$d\overline{F}=-\Gamma^2{}_2\wedge\overline{F}-\Gamma^2{}_3\wedge T-\Gamma^2{}_4\wedge\Lambda,$$

$$dT=-\Gamma^3{}_1\wedge F-\Gamma^3{}_2\wedge\overline{F}-\Gamma^3{}_3\wedge T,$$

$$d\Lambda=-\Gamma^4{}_1\wedge F-\Gamma^4{}_2\wedge\overline{F}-\Gamma^4{}_4\wedge\Lambda,$$

$$g_{ik}\Gamma^k{}_j+g_{jk}\Gamma^k{}_i=0,\quad g_{ij}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&-1\\0&0&-1&0\end{pmatrix}$$

read

$$\Gamma^i{}_j=\begin{pmatrix}\overline{\Omega}-\Omega&0&-E&-\overline{\Gamma}\\0&\Omega-\overline{\Omega}&-\overline{E}&-\Gamma\\-\Gamma&-\overline{\Gamma}&\Omega+\overline{\Omega}&0\\-\overline{E}&-E&0&-\Omega-\overline{\Omega}\end{pmatrix}.$$

The curvature $\mathscr{R}^i{}_j=d\Gamma^i{}_j+\Gamma^i{}_k\wedge\Gamma^k{}_j$ of this connection can be calculated using the relations (1) and their consequences (2). Modulo the obvious reality conditions its components read as follows (we lower indices by means of the metric g_{ij}).

$$\mathscr{R}_{23}=-\overline{F}\wedge T-\alpha\Lambda\wedge F,$$

$$\mathscr{R}_{14}=\overline{\alpha}\overline{F}\wedge T+a(\Lambda\wedge T-F\wedge\overline{F})-h\Lambda\wedge F,$$

$$\tfrac{1}{2}(\mathscr{R}_{43}-\mathscr{R}_{12})=-(\alpha+\tfrac{1}{2}\lambda)(\Lambda\wedge T-F\wedge\overline{F})-a\Lambda\wedge F.$$

These should be compared with the definitions of the spinorial coefficients of the Weyl tensor Ψ_μ, $\mu=0,1,2,3,4$, the traceless Ricci tensor S_{ij}, and the Ricci scalar R given by

$$\begin{aligned}\mathscr{R}_{23}=&\Psi_4\overline{F}\wedge T+\Psi_3(\Lambda\wedge T-F\wedge\overline{F})+(\Psi_2+\tfrac{1}{12}R)\Lambda\wedge F\\&+\tfrac{1}{2}S_{33}F\wedge T+\tfrac{1}{2}S_{32}(\Lambda\wedge T+F\wedge\overline{F})+\tfrac{1}{2}S_{22}\Lambda\wedge\overline{F},\end{aligned}$$

$$\begin{aligned}\mathscr{R}_{14}=&(-\Psi_2-\tfrac{1}{12}R)\overline{F}\wedge T-\Psi_1(\Lambda\wedge T-F\wedge\overline{F})-\Psi_0\Lambda\wedge F\\&-\tfrac{1}{2}S_{11}F\wedge T-\tfrac{1}{2}S_{41}(\Lambda\wedge T+F\wedge\overline{F})-\tfrac{1}{2}S_{44}\Lambda\wedge\overline{F},\end{aligned}$$

$$\begin{aligned}\tfrac{1}{2}(\mathscr{R}_{43}-\mathscr{R}_{12})=&\Psi_3\overline{F}\wedge T+(\Psi_2-\tfrac{1}{24}R)(\Lambda\wedge T-F\wedge\overline{F})+\Psi_1\Lambda\wedge F\\&+\tfrac{1}{2}S_{31}F\wedge T+\tfrac{1}{4}(S_{12}+S_{34})(\Lambda\wedge T+F\wedge\overline{F})+\tfrac{1}{2}S_{42}\Lambda\wedge\overline{F}.\end{aligned}$$

Looking at these equations we easily get $S_{ij}=R_{ij}-\frac{1}{4}Rg_{ij}\equiv 0$. This proves the Einstein property of the metric.

The spinorial coefficients for the Weyl tensor are also easy to obtain. They read

$$\Psi_0=h,\quad \Psi_1=-a,\quad \Psi_2=-\alpha-\frac{\lambda}{3},\quad \Psi_3=0,\quad \Psi_4=-1.$$

Due to the nonvanishing of Ψ_4 the metric is never conformally flat.

We can also interpret the constant λ which for the first time appeared in Eq. (2). One easily reads from the above that it is proportional to the Ricci scalar $R=4\lambda$. This concludes the proof of the Theorem. □

We, additionally, note that the metric is of the Cartan–Petrov–Penrose type D iff

$$a=0; \quad h=-(3\alpha+\lambda)^2 \tag{8}$$

and of the type N iff

$$h=a=0, \quad \lambda=-3\alpha. \tag{9}$$

The metric is algebraically special iff $I^3=27J^2$, where

$$I=-h+\tfrac{1}{3}(3\alpha+\lambda)^2, \quad J=a^2+\tfrac{1}{27}(3\alpha+\lambda)^3+\tfrac{1}{3}h(3\alpha+\lambda).$$

The converse of the present paper can be stated in the following theorem (see Ref. 1 for details).

Theorem 2: *Let $\mathscr{M}$ be a four-dimensional manifold with a Lorentzian conformally nonflat metric g satisfying the Einstein equations $R_{ij}=\lambda g_{ij}$. Then, there exists a double branched cover $\widetilde{\mathscr{P}}$ of the bundle of null directions $\mathscr{P}$ over $\mathscr{M}$*[2,3]

(i) which is a fibration over $\mathscr{M}$ with fibers being two-dimensional tori (or, in algebraically special cases, their degenerate counterparts[1]*)*

and

(ii) on which there exists a unique system of one-forms $(F,\overline{F},T,\Lambda,E,\overline{E})$ which satisfies the following three conditions:

(i) T, Λ are real- and F, E are complex-valued one-forms,
(ii) $F\wedge\overline{F}\wedge T\wedge\Lambda\wedge E\wedge\overline{E}\neq 0$ at each point p of $\widetilde{\mathscr{P}}$;
(iii)

$$\begin{aligned} dF&=(\Omega-\overline{\Omega})\wedge F+E\wedge T+\overline{\Gamma}\wedge\Lambda,\\ dT&=\Gamma\wedge F+\overline{\Gamma}\wedge\overline{F}-(\Omega+\overline{\Omega})\wedge T,\\ d\Lambda&=\overline{E}\wedge F+E\wedge\overline{F}+(\Omega+\overline{\Omega})\wedge\Lambda,\\ dE&=2\Omega\wedge E+\overline{F}\wedge T+\alpha\Lambda\wedge F \end{aligned}$$

with certain complex-valued one-forms Ω and Γ, and a certain complex function α on $\widetilde{\mathscr{P}}$.

I warmly thank Andrzej Trautman for stimulating my interest in Refs. 4–6. I am also very grateful to Mike Crampin, Simonetta Fritelli, Jerzy Lewandowski, Lionel Mason, Ted Newman, David Robinson, and Helmuth Urbantke for helpful discussions. This research was supported in part by Komitet Badań Naukowych (Grant No. 2P03B 017 12), Consorzio per lo Sviluppo Internazionale dell'Universita degli Studi di Trieste, and Erwin Schrödinger International Institute for Mathematical Physics.

[1] P. Nurowski, ''Elliptic fibrations over the Einstein space–times,'' J. Math. Phys. **39**, 5481 (1998).
[2] P. Nurowski, ''Optical geometries and related structures,'' J. Geom. Phys. **18**, 335–348 (1996).
[3] P. Nurowski, ''Twistor bundles, Einstein equations and real structures,'' in Class. Quantum Grav. **14**, The special volume edited by R. Penrose, I. Robinson and J. Tafel to celebrate A. Trautman's 64th birthday, A261–A290 (1997).
[4] E. Cartan, ''Sur la geometrie pseudo-conforme des hypersurfaces de deux variables complexes. I,'' Ann. Math. Pura Appl. **11**, 17–90 (1932).
[5] E. Cartan, ''Sur la geometrie pseudo-conforme des hypersurfaces de deux variables complexes. II,'' Ann. Scuola Norm. Sup. Pisa **1**, 333–354 (1932).
[6] R. Penrose, ''The twistor programme,'' Rep. Math. Phys. **12**, 65–76 (1967).

Elliptic fibrations associated with the Einstein space–times

Paweł Nurowski[a)]
Dipartimento di Scienze Matematiche, Universita degli Studi di Trieste, Trieste, Italy, and Department of Mathematics, King's College, London, United Kingdom

(Received 13 April 1998; accepted for publication 20 April 1998)

Given a conformally nonflat Einstein space–time we define a fibration $\tilde{\mathscr{P}}$ over it. The fibers of this fibration are elliptic curves (two-dimensional tori) or their degenerate counterparts. Their topology depends on the algebraic type of the Weyl tensor of the Einstein metric. The fibration $\tilde{\mathscr{P}}$ is a double branched cover of the bundle $\mathscr{P}$ of null direction over the space–time and is equipped with six linearly independent one-forms which satisfy a certain relatively simple system of equations. © *1998 American Institute of Physics.* [S0022-2488(98)01210-9]

I. DEFINITIONS

In Ref. 1 we defined a certain differential system $\mathscr{T}$ on an open set U of $\mathbf{R}$.[6] We showed that a pair $(U,\mathscr{T})$ naturally defines a four-dimensional conformally nonflat Lorentzian space–time $(\mathscr{M},g)$ which satisfies the Einstein equations $R_{ij}=\lambda g_{ij}$. In this paper we prove the converse statement. In particular, we give a construction which associates a certain six-dimensional elliptic fibration $\tilde{\mathscr{P}}$ with any conformally nonflat Lorentzian Einstein space–time. Moreover, we show how $\tilde{\mathscr{P}}$ may be equipped with a unique differential system which has all the properties of the system $\mathscr{T}$ on U.

We briefly recall the definitions of the geometrical objects we need in the following. Let $\mathscr{M}$ be a four-dimensional-oriented and time-oriented manifold equipped with a Lorentzian metric g of signature $(+,+,+,-)$. It is convenient to introduce a null frame $(m,\bar{m},k,l)$ on $\mathscr{M}$ with a coframe $\theta^i=(\theta^1,\theta^2,\theta^3,\theta^4)=(M,\bar{M},K,L)$ so that

$$g=g_{ij}\theta^i\theta^j=M\bar{M}-KL. \tag{1}$$

[Such expressions as $\theta^i\theta^j$ mean the symmetrized tensor product, e.g., $\theta^i\theta^j=\frac{1}{2}(\theta^i\otimes\theta^j+\theta^j\otimes\theta^i)$. Also, we will denote by round (respectively, square) brackets the symmetrization (respectively, antisymmetrization) of indices, e.g., $a_{(ik)}=\frac{1}{2}(a_{ik}+a_{ki})$, $a_{[ik]}=\frac{1}{2}(a_{ik}-a_{ki})$, etc.]

The Lorentz group $\mathbf{L}$ consists of matrices $\lambda^i_j\in\mathbf{GL}(4,\mathbf{C})$ such that

$$g_{jl}=g_{ik}\lambda^i{}_j\lambda^k{}_l,$$

$$\lambda^2{}_2=\overline{\lambda^1{}_1},\quad \lambda^2{}_1=\overline{\lambda^1{}_2},\quad \lambda^2{}_3=\overline{\lambda^1{}_3},\quad \lambda^2{}_4=\overline{\lambda^1{}_4}, \tag{2}$$

$$\lambda^3{}_2=\overline{\lambda^3{}_1},\quad \lambda^3{}_3=\overline{\lambda^3{}_3},\quad \lambda^3{}_4=\overline{\lambda^3{}_4},\quad \lambda^4{}_2=\overline{\lambda^4{}_1},\quad \lambda^4{}_3=\overline{\lambda^4{}_3},\quad \lambda^4{}_4=\overline{\lambda^4{}_4}.$$

We will denote the inverse of the Lorentz matrix $\lambda^i{}_j$ by $\tilde{\lambda}^i{}_j$.

The connected component of the identity element of $\mathbf{L}$ is the proper ortochroneous Lorentz group, which we denote by $\mathbf{L}^{\uparrow}_{+}$.

Given g and θ^i the connection one-forms $\Gamma_{ij}=g_{ik}\Gamma^k{}_j$ are uniquely defined by

$$d\theta^i=-\Gamma^i{}_j\wedge\theta^j,\quad \Gamma_{ij}+\Gamma_{ji}=0. \tag{3}$$

a)Permanent address: Instytut Fizyki Teoretycznej, Wydział Fizyki, Uniwersytet Warszawski, ul. Hoża 69, Warszawa, Poland. Electronic mail: nurowski@fuw.edu.pl

 © 1998 American Institute of Physics

The connection coefficients Γ_{ijk} are determined by $\Gamma_{ij}=\Gamma_{ijk}\theta^k$ (we lower and raise indices by means of the metric and its inverse). Using them we define the curvature two-forms $\mathscr{R}^k{}_i$, the Riemann tensor $R^i{}_{jkl}$, the Ricci tensor R_{ij}, and its scalar R by

$$\mathscr{R}^k{}_i=\tfrac{1}{2}R^k{}_{imj}\theta^m\wedge\theta^j=d\Gamma^k{}_i+\Gamma^k{}_j\wedge\Gamma^j{}_i, \quad R_{ij}=R^k{}_{ikj}, \quad R=g^{ij}R_{ij}.$$

We also introduce the traceless Ricci tensor by

$$S_{ij}=R_{ij}-\tfrac{1}{4}g_{ij}R.$$

Note that the vanishing of S_{ij} is equivalent to the Einstein equations $R_{ij}=\lambda g_{ij}$ for the metric g. We define the Weyl tensor $C^i{}_{jkl}$ by

$$C_{ijkl}=R_{ijkl}+\tfrac{1}{3}Rg_{i[k}g_{l]j}+R_{j[k}g_{l]i}+R_{i[l}g_{k]j},$$

and its spinorial coefficients Ψ_μ by

$$\begin{aligned}\mathscr{R}_{23}&=\Psi_4\overline{M}\wedge K+\Psi_3(L\wedge K-M\wedge\overline{M})+(\Psi_2+\tfrac{1}{12}R)L\wedge M\\&\quad+\tfrac{1}{2}S_{33}M\wedge K+\tfrac{1}{2}S_{32}(L\wedge K+M\wedge\overline{M})+\tfrac{1}{2}S_{22}L\wedge\overline{M},\\ \mathscr{R}_{14}&=(-\Psi_2-\tfrac{1}{12}R)\overline{M}\wedge K-\Psi_1(L\wedge K-M\wedge\overline{M})-\Psi_0L\wedge M\\&\quad-\tfrac{1}{2}S_{11}M\wedge K-\tfrac{1}{2}S_{41}(L\wedge K+M\wedge\overline{M})-\tfrac{1}{2}S_{44}L\wedge\overline{M},\\ \tfrac{1}{2}(\mathscr{R}_{43}-\mathscr{R}_{12})&=\Psi_3\overline{M}\wedge K+(\Psi_2-\tfrac{1}{24}R)(L\wedge K-M\wedge\overline{M})+\Psi_1L\wedge M\\&\quad+\tfrac{1}{2}S_{31}M\wedge K+\tfrac{1}{4}(S_{12}+S_{34})(L\wedge K+M\wedge\overline{M})+\tfrac{1}{2}S_{42}L\wedge\overline{M}.\end{aligned}$$

II. THE BUNDLE OF NULL COFRAMES

Let $\mathscr{F}(\mathscr{M})$ denote the bundle of oriented and time oriented null coframes over $\mathscr{M}$. This means that $\mathscr{F}(\mathscr{M})$ is the set of all equally oriented null coframes θ^i at all points of $\mathscr{M}$. The mapping $\pi:\mathscr{F}(\mathscr{M})\to\mathscr{M}$, which maps a coframe θ^i at $x\in\mathscr{M}$ onto x, gives the canonical projection. A fiber $\pi^{-1}(x)$ in $\mathscr{F}(\mathscr{M})$ consists of all the null coframes at point x which have the same orientation and time orientation. If θ^j is a null coframe at $x\in\mathscr{M}$, then any other equivalently oriented null coframe at x is given by $\theta'^i=\lambda^i{}_j\theta^j$, where $\lambda^i{}_j$ is a certain element of $\mathbf{L}^\uparrow_+$. This defines an action of $\mathbf{L}^\uparrow_+$ on $\mathscr{F}(\mathscr{M})$. Thus, $\mathscr{F}(\mathscr{M})$ is a ten-dimensional principal fiber bundle with $\mathbf{L}^\uparrow_+$ as its structural group.

It is well known that the bundle $\mathscr{F}(\mathscr{M})$ is equipped with a natural four-covector-valued one-form e^i, $i=1,2,3,4$, the Cartan soldering form, which is defined as follows. Take any vector v_c tangent to $\mathscr{F}(\mathscr{M})$ at a point c. Let c be in the fiber $\pi^{-1}(x)$ over a point $x\in\mathscr{M}$. This means that c may be identified with a certain null coframe θ^i_c at x.

Then, the formal definition of e^i reads: $e^i(v_c)=\theta^i_c(\pi_*v_c)$. The first two components of e^i are complex and mutually conjugated. The remaining two are real. Altogether, they constitute a system of four well-defined linearly independent one-forms on $\mathscr{F}(\mathscr{M})$. In the following we will denote them by

$$F=e^1=\overline{e^2}, \quad T=\overline{T}=e^3, \quad \Lambda=\overline{\Lambda}=e^4. \tag{4}$$

A theorem which we present below is a null coframe reformulation of the Elie Cartan theorem on affine connections.

Theorem 1: *Let $e^i=(F,\overline{F},T,\Lambda)$ be the soldering form on $\mathscr{F}(\mathscr{M})$. Then*

(i) the system of equations

$$de^i+\omega^i{}_j\wedge e^j=0, \quad g_{ki}\omega^i{}_j+g_{ji}\omega^i{}_k=0 \tag{5}$$

for a matrix of complex-valued one-forms $\omega^i{}_j$ $(i,j=1,2,3,4)$ on $\mathscr{F}(\mathscr{M})$ has a unique solution,

(ii) $\omega^i{}_j$ *uniquely defines six complex-valued one-forms* $(E,\overline{E},\Gamma,\overline{\Gamma},\Omega,\overline{\Omega})$ *by*

$$\omega^i{}_j=\begin{pmatrix} \overline{\Omega}-\Omega & 0 & -E & -\overline{\Gamma} \\ 0 & \Omega-\overline{\Omega} & -\overline{E} & -\Gamma \\ -\Gamma & -\overline{\Gamma} & \Omega+\overline{\Omega} & 0 \\ -\overline{E} & -E & 0 & -\Omega-\overline{\Omega} \end{pmatrix}, \tag{6}$$

(iii) the forms $(F,\overline{F},T,\Lambda,E,\overline{E},\Gamma,\overline{\Gamma},\Omega,\overline{\Omega})$ *are linearly independent at each point of* $\mathscr{F}(\mathscr{M})$.

For completeness we sketch the proof.

First, we show that if there is a solution to (5) then it is unique. To do this we assume the existence of two solutions—$\omega^i{}_j$ and $\hat{\omega}^i{}_j$.

Subtracting $de^i+\hat{\omega}^i{}_j\wedge e^j=0$ from $de^i+\omega^i{}_j\wedge e^j=0$ we get

$$(\omega^i{}_j-\hat{\omega}^i{}_j)\wedge e^j=0. \tag{7}$$

Now, let e^μ, $\mu=1, 2, 3, 4, 5, 6$, be a system of one-forms such that the ten one-forms (e^i,e^μ) constitute a basis of one-forms on $\mathscr{F}(\mathscr{M})$. Let $\omega^i{}_j=\omega^i{}_{jk}e^k+\omega^i{}_{j\mu}e^\mu$ and $\hat{\omega}^i{}_j=\hat{\omega}^i{}_{jk}e^k+\hat{\omega}^i{}_{j\mu}e^\mu$ be the corresponding decompositions of the solutions. Then (7) easily yields $\omega^i{}_{j\mu}=\hat{\omega}^i{}_{j\mu}$ and $\omega_{i[jk]}=\hat{\omega}_{i[jk]}$. The defining properties of the solutions give also $\omega_{(ij)k}=0=\hat{\omega}_{(ij)k}$. Now, due to the identity $A_{ijk}=A_{i[jk]}-A_{j[ik]}-A_{k[ij]}$, which is true for any A_{ijk} such that $A_{(ij)k}=0$, we get $\omega_{ijk}=\hat{\omega}_{ijk}$. This shows that $\omega^i{}_j=\hat{\omega}^i{}_j$, hence the uniqueness.

We proceed to the construction of a solution.

Given a sufficiently small neighborhood $\mathscr{O}$ in $\mathscr{M}$ we identify $\pi^{-1}(\mathscr{O})$ with $\mathscr{O}\times\mathbf{L}^\uparrow_+$. Then, the soldering form may be written as

$$e^i=\lambda^i{}_j\theta^j. \tag{8}$$

Taking de^i and using Eq. (3) one easily finds that

$$\omega^i{}_j=\lambda^i{}_k\Gamma^k{}_m\tilde{\lambda}^m_j-d\lambda^i{}_k\tilde{\lambda}^k{}_j \tag{9}$$

is a solution to (5). Given this solution one defines the forms (E,Γ,Ω) by

$$E=-\omega^1{}_3,\quad \Gamma=-\omega^2{}_4,\quad \Omega=\tfrac{1}{2}(\omega^2{}_2+\omega^3{}_3). \tag{10}$$

This is in accordance with (6) due to $\omega_{(ij)}=0$ and the reality properties of e^i.

To prove the linear independence of the system $(F,\overline{F},T,\Lambda,E,\overline{E},\Gamma,\overline{\Gamma},\Omega,\overline{\Omega})$ it is enough to observe that the six one-forms $d\lambda^1{}_k\tilde{\lambda}^k{}_3$, $d\lambda^2{}_k\tilde{\lambda}^k{}_3$, $d\lambda^1{}_k\tilde{\lambda}^k{}_4$, $d\lambda^2{}_k\tilde{\lambda}^k{}_4$, $d\lambda^1{}_k\tilde{\lambda}^k{}_1$, and $d\lambda^3{}_k\tilde{\lambda}^k{}_3$ constitute a basis of right invariant forms on $\mathbf{L}^\uparrow_+$.

The theorem is proven.

III. THE STRUCTURE EQUATIONS

Consider the ten well-defined forms $(F,\overline{F},T,\Lambda,E,\overline{E},\Omega,\overline{\Omega},\Gamma,\overline{\Gamma})$ on $\mathscr{F}(\mathscr{M})$ given by (4) and (10). The differentials of the first four forms are given by (5) and (6). In the basis $(F,\overline{F},T,\Lambda,E,\overline{E},\Omega,\overline{\Omega},\Gamma,\overline{\Gamma})$ they assume the form

$$dF=(\Omega-\overline{\Omega})\wedge F+E\wedge T+\overline{\Gamma}\wedge\Lambda, \tag{11}$$

$$dT=\Gamma\wedge F+\overline{\Gamma}\wedge\overline{F}-(\Omega+\overline{\Omega})\wedge T, \tag{12}$$

$$d\Lambda=\overline{E}\wedge F+E\wedge\overline{F}+(\Omega+\overline{\Omega})\wedge\Lambda. \tag{13}$$

The differentials of the other forms may be easily calculated using the local representation (9) and the well-known structure equations

$$d\omega^i{}_j = -\omega^i{}_k \wedge \omega^k{}_j + \tfrac{1}{2} R^k{}_{smn} \lambda^i{}_k \tilde{\lambda}^s{}_j \tilde{\lambda}^m{}_l \tilde{\lambda}^n{}_p e^l \wedge e^p. \tag{14}$$

These differentials are by far much more complicated than the differentials of dF, dT, and $d\Lambda$. In particular, the decompositions of dF, dT, and $d\Lambda$ onto the basis of two-forms associated with $(F,\bar{F},T,\Lambda,E,\bar{E},\Omega,\bar{\Omega},\Gamma,\bar{\Gamma})$ have only constant coefficients. It turns out that in the differentials dE, $d\Omega$, and $d\Gamma$ coefficients which are functions appear. The zero sets of these functions have a well-defined geometrical meaning and define certain subsets of $\mathscr{F}(\mathscr{M})$. Now, the hope is that when we restrict ourselves to such subsets then the differentials of E, Ω, and Γ will have a much simpler form than their differentials on the whole $\mathscr{F}(\mathscr{M})$. Our aim now is to study this possibility.

IV. EXPLICIT EXPRESSIONS FOR THE BASIC FORMS ON $\mathscr{F}(\mathscr{M})$

We concentrate on the analysis of $dE = -d\omega^1{}_3$.

Let us introduce the matrices $\lambda^i{}_j(w,z,y)$ and $\lambda^i{}_j(w',z',y')$ such that

$$\lambda^i{}_j(w,z,y) = \begin{pmatrix} |w|w^{-1}(1+\overline{yz}) & |w|w^{-1}\overline{yz} & |w|w^{-1}(1+\overline{yz})z & |w|w^{-1}\overline{y} \\ w|w|^{-1}\overline{yz} & w|w|^{-1}(1+yz) & w|w|^{-1}(1+yz)\overline{z} & w|w|^{-1}y \\ |w|(1+\overline{yz})y & |w|(1+yz)\overline{y} & |w||1+\overline{yz}|^2 & |w||y|^2 \\ \overline{z}|w|^{-1} & z|w|^{-1} & |z|^2|w|^{-1} & |w|^{-1} \end{pmatrix},$$

$$\lambda^i{}_j(w',z',y') = \begin{pmatrix} |w'|w'^{-1}\overline{y'}z' & |w'|w'^{-1}(1+\overline{y'}\overline{z'}) & |w'|w'^{-1}\overline{y'} & |w'|w'^{-1}(1+\overline{y'}\overline{z'})z' \\ w'|w'|^{-1}(1+y'z') & w'|w'|^{-1}y'\overline{z'} & w'|w'|^{-1}y' & w'|w'|^{-1}(1+y'z')\overline{z'} \\ |w'|(1+y'z')\overline{y'} & |w'|(1+\overline{y'}\overline{z'}) & |w'||y'|^2 & y'|w'||1+\overline{y'}\overline{z'}|^2 \\ z'|w'|^{-1} & \overline{z'}|w'|^{-1} & |w'|^{-1} & |z'|^2|w'|^{-1} \end{pmatrix}.$$

Then, it is well known that $\mathbf{L}^{\uparrow}_{+}$ can be represented by

$$\mathbf{L}^{\uparrow}_{+} = \mathscr{U} \cup \mathscr{U}', \tag{15}$$

where

$$\mathscr{U} = \{\lambda^i{}_j(w,z,y) \text{ such that } (w,z,y) \in \mathbf{C}^3, w \neq 0\}, \tag{16}$$

$$\mathscr{U}' = \{\lambda^i{}_j(w',z',y') \text{ such that } (w',z',y') \in \mathbf{C}^3, w' \neq 0\}. \tag{17}$$

On the intersection $\mathscr{U} \cap \mathscr{U}'$, the coordinates (w,z,y) and (w',z',y') shall be related by

$$w = -\frac{w'}{z'^2}, \quad z = \frac{1}{z'}, \quad y = -z'(1+y'z'), \tag{18}$$

$$w' = -\frac{w}{z^2}, \quad z' = \frac{1}{z}, \quad y' = -z(1+yz). \tag{19}$$

Thus, we can cover any $\pi^{-1}(\mathscr{O}) \cong (\mathscr{O} \times \mathbf{L}^{\uparrow}_{+})$ by the two charts $\mathscr{O} \times \mathscr{U}$ and $\mathscr{O} \times \mathscr{U}'$. Now, on $\mathscr{O}$ consider the coframe θ^i of (1). Inserting $\lambda^i{}_j = \lambda^i{}_j(w,z,y)$ or $\lambda^i{}_j(w',z',y')$ to the formulas (8), (9), (14) and using the definitions (1), (4), (10) we easily obtain the following two lemmas.

Lemma 1: On $\mathscr{O} \times \mathscr{U}$ the forms F, T, Λ, E, Ω and Γ read

$$F = \frac{|w|}{w}[(1+\overline{yz})(M+zK) + \overline{yz}\overline{M} + \overline{y}\overline{L}], \tag{20}$$

$$T = |w|[|1+\overline{yz}|^2 K + y(1+\overline{yz})M + \overline{y}(1+yz)\overline{M} + y\overline{y}\overline{L}], \tag{21}$$

$$\Lambda=\frac{1}{|w|}\,[L+z\overline{z}K+z\overline{M}+\overline{z}M], \tag{22}$$

$$E=\frac{1}{w}\,[dz+\Gamma_{32}+z(\Gamma_{21}+\Gamma_{43})+z^2\Gamma_{14}], \tag{23}$$

$$\Omega=-\frac{1}{2}\,\frac{dw}{w}-ydz-y\Gamma_{32}-\frac{1}{2}\,(1+2yz)(\Gamma_{21}+\Gamma_{43})-z(1+yz)\Gamma_{14}, \tag{24}$$

$$\Gamma=w[dy-y^2dz-y^2\Gamma_{32}-y(1+yz)(\Gamma_{43}-\Gamma_{12})-(1+yz)^2\Gamma_{14}]. \tag{25}$$

Lemma 2: On $\mathscr{O}\times\mathscr{U}'$ the forms F, T, Λ, E, Ω and Γ read

$$F=\frac{|w'|}{w'}\,[(1+\overline{y'}\overline{z'})(\overline{M}+z'L)+\overline{y'}z'M+\overline{y'}K], \tag{26}$$

$$T=|w'|[|1+\overline{y'}\overline{z'}|^2L+y'(1+\overline{y'}\overline{z'})\overline{M}+\overline{y'}(1+y'z')M+y'\overline{y'}K], \tag{27}$$

$$\Lambda=\frac{1}{|w'|}\,[K+z'\overline{z'}L+z'M+\overline{z'}\overline{M}], \tag{28}$$

$$E=\frac{1}{w'}\,[dz'+\Gamma_{41}+z'(\Gamma_{12}+\Gamma_{34})+z'^2\Gamma_{23}], \tag{29}$$

$$\Omega=-\frac{1}{2}\,\frac{dw'}{w'}-y'dz'-y'\Gamma_{41}-\frac{1}{2}\,(1+2y'z')(\Gamma_{12}+\Gamma_{34})-z'(1+y'z')\Gamma_{23}, \tag{30}$$

$$\Gamma=w'[dy'-y'^2dz'-y'^2\Gamma_{41}-y'(1+y'z')(\Gamma_{34}-\Gamma_{21})-(1+y'z')^2\Gamma_{23}]. \tag{31}$$

Note that this Lemma follows from the previous one by applying transformations $M\leftrightarrow\overline{M}$, $K\leftrightarrow L$, $1\leftrightarrow2$, and $3\leftrightarrow4$, where the last two transformations refer to the tetrad indices.

Now one can easily find the differential of *dE*.

Lemma 3: The decomposition of dE onto the basis $(F,\overline{F},T,\Lambda,E,\overline{E},\Omega,\overline{\Omega},\Gamma,\overline{\Gamma})$ of one-forms on $\pi^{-1}(\mathscr{O})$ reads

$$dE=2\Omega\wedge E+\phi T\wedge F+b(\Lambda\wedge T+F\wedge\overline{F})+\beta F\wedge\Lambda+\psi T\wedge\overline{F}+a(\Lambda\wedge T-F\wedge\overline{F})+\alpha\Lambda\wedge F,$$

where $\phi,b,\beta,\psi,a,\alpha$ are well-defined functions on $\pi^{-1}(\mathscr{O})$.

The functions $\phi,b,\beta,\psi,a,\alpha$ are given by

$$\phi=\frac{1}{|w|^2}\,\Phi,\quad b=\overline{w}(\tfrac{1}{2}\phi_{\overline{z}}+\overline{y}\phi),\quad \beta=\overline{w}^2(\tfrac{1}{2}\phi_{\overline{z}\overline{z}}+\overline{y}\phi_{\overline{z}}+\overline{y}^2\phi),$$

$$\psi=\frac{1}{w^2}\,\Psi,\quad a=w(\tfrac{1}{4}\psi_z+y\psi),\quad \alpha=-w^2(\tfrac{1}{12}\psi_{zz}+y^2\psi+\tfrac{1}{2}y\psi_z)-\tfrac{1}{12}R,$$

$$\Phi=\tfrac{1}{2}S_{33}-\overline{z}S_{23}-zS_{13}+z\overline{z}(S_{12}+S_{34})+\tfrac{1}{2}\overline{z}^2S_{22}+\tfrac{1}{2}z^2S_{11}-\overline{z}^2zS_{24}-z^2\overline{z}S_{14}+\tfrac{1}{2}z^2\overline{z}^2S_{44},$$

and

$$\Psi=\Psi_4-4\Psi_3z+6\Psi_2z^2-4\Psi_1z^3+\Psi_0z^4$$

on $\mathscr{O}\times\mathscr{U}$ and by

$$\phi=\frac{1}{|w'|^2}\,\Phi',\quad b=\overline{w}'(\tfrac{1}{2}\phi_{\overline{z}'}+\overline{y}'\phi),\quad \beta=\overline{w}'^2(\tfrac{1}{2}\phi_{\overline{z}'\overline{z}'}+\overline{y}'\phi_{\overline{z}'}+\overline{y}'^2\phi),$$

$$\psi=\frac{1}{w'^2}\,\Psi',\quad a=w'(\tfrac{1}{4}\psi_{z'}+y'\psi),\quad \alpha=-w'^2(\tfrac{1}{12}\psi_{z'z'}+y'^2\psi+\tfrac{1}{2}y'\psi_{z'})-\tfrac{1}{12}R,$$

$$\Phi'=\tfrac{1}{2}S_{44}-\overline{z}'S_{14}-z'S_{24}+z'\overline{z}'(S_{12}+S_{34})+\tfrac{1}{2}\overline{z}'^2S_{11}$$
$$+\tfrac{1}{2}z'^2S_{22}-\overline{z}'^2z'S_{13}-z'^2\overline{z}'S_{23}-\tfrac{1}{2}\overline{z}'^2\overline{z}'^2S_{33}$$

and

$$\Psi'=\Psi_0-4\Psi_1z'+6\Psi_2z'^2-4\Psi_3z'^3+\Psi_4z'^4$$

on $\mathscr{O}\times\mathscr{U}'$.

The following three cases are of particular interest.

(a) The metric g of the four-manifold $\mathscr{M}$ satisfies the Einstein equations $R_{ij}=\lambda g_{ij}$ and is not conformally flat. This case is characterized by $\Phi\equiv 0$ and $\Psi\not\equiv 0$.
(b) The metric g is conformally flat but not Einstein. This case corresponds to $\Psi\equiv 0$, $\Phi\not\equiv 0$.
(c) The metric g is of constant curvature. This means that $\Psi\equiv\Phi\equiv 0$.

In the first two cases there is a canonical choice of certain six-dimensional subsets in $\mathscr{F}(\mathscr{M})$. This is defined by the demand that on such sets certain components of dE should identically vanish. This approach is impossible in case (c) since this implies an immediate reduction of dE to the form

$$dE=2\Omega\wedge E+\tfrac{1}{12}R\Lambda\wedge F. \tag{32}$$

V. DISTINGUISHED SUBSET OF $\mathscr{F}(\mathscr{M})$

From now on we consider case (a). This is the most interesting generic Einstein case.

Imposing the restrictions (a) on dE we immediately see that

$$dE=2\Omega\wedge E+\psi T\wedge\overline{F}+a(\Lambda\wedge T-F\wedge\overline{F})+\alpha\Lambda\wedge F,$$

where ψ,a,α are the same as in the Lemma 1. Since we are in the not conformally flat case (a) we have $\psi\neq 0$. This makes possible the restriction to such a set $\mathscr{W}\subset\pi^{-1}(\mathscr{O})$ in which a identically vanish. Thus we consider

$$\mathscr{W}=\mathscr{W}_1\cup\mathscr{W}_2, \tag{33}$$

where

$$\mathscr{W}_1=\{(x;w,z,y)\in(\mathscr{O}\times\mathscr{U})\ \text{ such that }\ \psi y+\tfrac{1}{4}\psi_z=0\},$$

$$\mathscr{W}_2=\{(x;w',z',y')\in(\mathscr{O}\times\mathscr{U}')\ \text{ such that }\ \psi y'+\tfrac{1}{4}\psi_{z'}=0\},$$

or (what is the same due to the nonvanishing of w and w')

$$\mathscr{W}_1=\{(x;w,z,y)\in(\mathscr{O}\times\mathscr{U})\ \text{ such that }\ \Psi y+\tfrac{1}{4}\Psi_z=0\}, \tag{34}$$

$$\mathscr{W}_2=\{(x;w',z',y')\in(\mathscr{O}\times\mathscr{U}')\ \text{ such that }\ \Psi' y'+\tfrac{1}{4}\Psi'_{z'},=0\}. \tag{35}$$

On $\mathscr{W}$ we have

$$dE=2\Omega\wedge E+\psi T\wedge\overline{F}+\alpha\Lambda\wedge F.$$

One can still simplify this relation by restricting oneself to a subset $\tilde{\mathscr{P}}_0$ of $\mathscr{W}$ in which $\psi=-1$. Then, $\tilde{\mathscr{P}}_0$ is a subset of $\pi^{-1}(\mathscr{O})$, which is given by

$$\tilde{\mathscr{P}}_0=\tilde{\mathscr{P}}_{01}\cup\tilde{\mathscr{P}}_{02}, \tag{36}$$

where

$$\tilde{\mathscr{P}}_{01}=\{(x;w,z,y)\in(\mathscr{O}\times\mathscr{U}) \text{ such that } \Psi y+\tfrac{1}{4}\Psi_z=0,\ \Psi+w^2=0\} \tag{37}$$

and

$$\tilde{\mathscr{P}}_{02}=\{(x;w',z',y')\in(\mathscr{O}\times\mathscr{U}') \text{ such that } \Psi' y'+\tfrac{1}{4}\Psi'_{z'}=0,\ \Psi'+w'^2=0\}. \tag{38}$$

It follows from the construction that on $\tilde{\mathscr{P}}_0$ we have

$$dE=2\Omega\wedge E+\overline{F}\wedge T+\alpha F\wedge\Lambda.$$

VI. ELLIPTIC FIBRATION

We study the geometry and topology of the set $\tilde{\mathscr{P}}_0$.

The equations defining $\tilde{\mathscr{P}}_0$ may be written as

$$y=-\frac{1}{4}\frac{\Psi_z}{\Psi},\quad y'=-\frac{1}{4}\frac{\Psi'_{z'}}{\Psi'}, \tag{39}$$

$$\Psi+w^2=0,\quad \Psi'+w'^2=0. \tag{40}$$

We see that the first pair of equations uniquely subordinates y to z and y' to z'. The second pair gives a relation between w and z and w' and z'. Thus, locally among the parameters $(x;w,z,y)$ in $\mathscr{O}\times\mathscr{U}$ [respectively, $(x;w',z',y')$ in $\mathscr{O}\times\mathscr{U}'$], only x and z (respectively, x and z') are free. This shows that $\tilde{\mathscr{P}}_0$ is six dimensional. Moreover, $\tilde{\mathscr{P}}_0$ is fibered over $\mathscr{O}$ with two-dimensional fibers. These are locally parametrized by z or z'. To discuss the topology of fibers one observes that the relation between w and z (respectively, w' and z') is purely polynomial. Over every point of $\mathscr{O}$ it has the form

$$w^2=-\Psi_4+4\Psi_3 z-6\Psi_2 z^2+4\Psi_1 z^3-\Psi_0 z^4 \tag{41}$$

or

$$w'^2=-\Psi_0+4\Psi_1 z'-6\Psi_2 z'^2+4\Psi_3 z'^3-\Psi_4 z'^4. \tag{42}$$

Assume for a moment (a) that $w=0$ and $w'=0$ are the allowed values of the parameters and (b) that equations

$$-\Psi_4+4\Psi_3 z-6\Psi_2 z^2+4\Psi_1 z^3-\Psi_0 z^4=0, \tag{43}$$

$$-\Psi_0+4\Psi_1 z'-6\Psi_2 z'^2+4\Psi_3 z'^3-\Psi_4 z'^4=0 \tag{44}$$

for z and z' have only distinct roots. Then, the relations (41)–(42), as being fourth order in the parameters z and z', describe a two-dimensional torus. (This is a well-known fact of classical algebraic geometry, see, e.g., Ref. 2. I am very grateful to Roger Penrose for clarifying this for me).[3]

Let us comment on (a) and (b).

(a) We know that w and w' cannot be zero by their definitions. So to have a torus fibration over the Einstein space–time we need to accept that w and w' may vanish. A price paid for this is that some of the forms $(F,\overline{F},T,\Lambda,E,\overline{E},\Omega,\overline{\Omega},\Gamma,\overline{\Gamma})$ will be singular on the resulting fibration at

these values of w and w'. With these remarks, from now on, we accept that w and w' may vanish. This enables us to introduce the fibration $\tilde{\mathscr{P}}$ over $\mathscr{M}$ which, over the neighborhood $\mathscr{O}\subset\mathscr{M}$ is given by

$$\tilde{\mathscr{P}}=\tilde{\mathscr{P}}_1\cup\tilde{\mathscr{P}}_2, \tag{45}$$

where

$$\tilde{\mathscr{P}}_1=\{(x;w,z,y)\in(\mathscr{O}\times\mathbf{C}^3)\ \text{ such that }\ \Psi y+\tfrac{1}{4}\Psi_z=0,\ \Psi+w^2=0\}, \tag{46}$$

$$\tilde{\mathscr{P}}_2=\left\{(x;w',z',y')\in(\mathscr{O}\times\mathbf{C}^3)\ \text{ such that }\ \Psi' y'+\frac{1}{4}\Psi'_{z'}=0,\ \Psi'+w'^2=0\right\} \tag{47}$$

and the transition functions between (w,z,y) and (w',z',y') coordinates are given by (18).

(b) The discussion in (a) means that $\tilde{\mathscr{P}}$ is a torus fibration over $\mathscr{O}$ provided that Eqs. (43) and (44) have distinct roots. It is well known that the number of distinct roots in (43) and (44) is directly related to the algebraic (Cartan–Petrov–Penrose[4–6]) classification of space–times. Thus, if the space–time $\mathscr{M}$ is algebraically general in $\mathscr{O}$, then $\tilde{\mathscr{P}}$ is a torus fibration over $\mathscr{O}$. In the algebraically special cases the fibers of $\tilde{\mathscr{P}}$ are degenerate tori. These topologically are:

(II) a torus with one vanishing cycle in the Cartan–Petrov–Penrose type II,

(D) two spheres touching each other in two different points in the Cartan–Petrov–Penrose type D,

(III) a sphere with one singular point in the Cartan–Petrov–Penrose type III,

(N) two spheres touching each other in one point in the Cartan–Petrov–Penrose type N.

The pure situations considered so far may be a bit more complicated when the Cartan–Petrov–Penrose type of the Einstein space–time varies from point to point. Imagine, for example, that along a continuous path from x to x' in $\mathscr{O}$ the Cartan–Petrov–Penrose type of the Einstein space–time changes from I to II. Then the fiber of $\tilde{\mathscr{P}}$ over x' is only a torus with one vanishing cycle although the fiber over the starting point x was a torus. It is clear that more complicated situations may occur, and that the fibers of $\tilde{\mathscr{P}}$ over different points of $\mathscr{M}$ can have topologies II, D, III, and N. Fibrations of this type are widely used in algebraic geometry. They are called elliptic, since their fibers can be any kind (even degenerate) of an elliptic curve.

VII. THE MAIN THEOREM

In Secs. IV–VI, for the clarity of presentation, we restricted ourselves to the neighborhood $\mathscr{O}$ of $\mathscr{M}$. We ended up with an elliptic fibration $\tilde{\mathscr{P}}$ over $\mathscr{O}$. This, however, can be easily prolonged to an elliptic fibration over the whole $\mathscr{M}$. To see this it is enough to observe that a fiber over any point in $\mathscr{O}$ is essentially defined by the Weyl tensor. Since the Weyl tensor is uniquely defined on the whole $\mathscr{M}$ then we can use equations like (41) to uniquely define the elliptic fibers over all $\mathscr{M}$.

Summing up the information from Secs. IV–VI we have the following theorem.

Theorem 2: *Given a four-dimensional conformally nonflat space–time $\mathscr{M}$ satisfying the Einstein equations $R_{ij}=\lambda g_{ij}$ one naturally defines a fibration $\Pi:\tilde{\mathscr{P}}\to\mathscr{M}$ with the following properties.*

(1) A fiber $\Pi^{-1}(x)$ over a point $x\in\mathscr{M}$ is a (possibly degenerate) elliptic curve $\mathscr{C}$ given by

$$\mathscr{C}=\mathscr{C}_1\cup\mathscr{C}_2,$$

$$\mathscr{C}_1=\{(w,z)\in\mathbf{C}^2\ \text{ such that }\ w^2=-\Psi_4+4\Psi_3 z-6\Psi_2 z^2+4\Psi_1 z^3-\Psi_0 z^4\},$$

$$\mathscr{C}_2=\{(w',z')\in\mathbf{C}^2\ \text{ such that }\ w'^2=-\Psi_0+4\Psi_1 z'-6\Psi_2 z'^2+4\Psi_3 z'^3-\Psi_4 z'^4\},$$

where the transition functions between (w,z) and (w',z') coordinates are given by

$$w=-\frac{w'}{z'^2},\quad z=\frac{1}{z'}.$$

(2) The degeneracy of a fiber depends on the algebraic type of the space–time metric and may change from point to point.

(3) There is a unique construction of a certain surface $\tilde{\mathscr{P}}_0$ of dimension six immersed in the bundle of null coframes $\mathscr{F}(\mathscr{M})$. $\tilde{\mathscr{P}}_0$ is fibered over $\mathscr{M}$ and $\tilde{\mathscr{P}}$ may be viewed as an extension of $\tilde{\mathscr{P}}_0$ achieved by adding to each fiber of $\tilde{\mathscr{P}}_0$ at most four points.

(4) There are ten one-forms $(F,\overline{F},T,\Lambda,E,\overline{E},\Omega,\overline{\Omega},\Gamma,\overline{\Gamma})$ on $\tilde{\mathscr{P}}$ with the following properties:

(a) Forms T,Λ are real, all the other are complex valued.

(b) The forms are defined in two steps. First, by restricting the soldering form components e^i and the Levi–Civita connection components ω^i_j from $\mathscr{F}(\mathscr{M})$ to $\tilde{\mathscr{P}}_0$ and second, by extending the restrictions to $\tilde{\mathscr{P}}$.

(c) $(F,\overline{F},T,\Lambda,E,\overline{E})$ constitute the basis of one-forms on $\tilde{\mathscr{P}}$.

(d) The forms satisfy the following equations on $\tilde{\mathscr{P}}$:

$$\begin{aligned} dF&=(\Omega-\overline{\Omega})\wedge F+E\wedge T+\overline{\Gamma}\wedge\Lambda,\\ dT&=\Gamma\wedge F+\overline{\Gamma}\wedge\overline{F}-(\Omega+\overline{\Omega})\wedge T,\\ d\Lambda&=\overline{E}\wedge F+E\Lambda\overline{F}+(\Omega+\overline{\Omega})\wedge\Lambda,\\ dE&=2\Omega\wedge E+\overline{F}\wedge T+\alpha\Lambda\wedge F, \end{aligned} \tag{48}$$

with a certain function α on $\tilde{\mathscr{P}}$.

The explicit formulas for the forms on $\tilde{\mathscr{P}}_1$ (respectively, on $\tilde{\mathscr{P}}_2$) may be obtained from the expressions of Lemma 1 (respectively, Lemma 2) by inserting the relations $y=-\Psi_z/(4\Psi)$ and $w^2+\Psi=0$ [respectively, $y'=-\Psi'_{z'}/(4\Psi')$ and $w'^2+\Psi'=0$].

VIII. RELATION BETWEEN THE ELLIPTIC FIBRATION AND THE BUNDLE OF NULL DIRECTIONS

Finally we note that the bundle $\tilde{\mathscr{P}}$ constitutes a double branched cover of the Penrose bundle $\mathscr{P}$ of null directions over the space–time.
To see this consider the map f given by

$$\tilde{\mathscr{P}}_1\ni(x;w,z,y)\overset{f}{\to}(x,z)\in\mathscr{O}\times\mathbf{C},$$

$$\tilde{\mathscr{P}}_2\ni(x;w',z',y')\overset{f}{\to}(x,z')\in\mathscr{O}\times\mathbf{C}.$$

Since on the intersection $\tilde{\mathscr{P}}_1\cap\tilde{\mathscr{P}}_2$ the coordinates z and z' are related by $z'=1/z$, then the two copies of $\mathbf{C}$, which appear in the above relations may be considered as two coordinate charts (say around the North and the South pole, respectively) on the two-dimensional sphere. This sphere is the sphere of null directions at a given point of $\mathscr{O}$ which can be seen as follows.

Consider directions of all the one-forms $L(z)=L+z\overline{z}K+z\overline{M}+\overline{z}M$ at $x\in\mathscr{O}$ for all the values of the complex parameter z. These directions are in one to one correspondence with null directions $k(z)=k+z\overline{z}\overline{l}-zm-\overline{zm}$ via $L(z)=-g(k(z))$. These null directions do not form a sphere yet, since the direction corresponding to the vector l is missing. But the missing direction is included in the family of null directions corresponding to the directions of one-forms $K(z')=K+z'\overline{z'}L+z'\overline{M}+\overline{z'}M$ at x. Thus the space parametrized by z and z' subject to the relation $z'=1/z$ is in one to one correspondence with the sphere of null directions at x. This proves the assertion that f is a cover of $\mathscr{P}$.

The map f is a double cover since if z (or z') is not a root of (43) [respectively, (44)] then $f^{-1}(z)$ is a two-point set. In at most four cases when z is a root $f^{-1}(z)$ is a one-point set. This proves that f is a singular cover.

As we know the singular points are branch points, which proves the statement from the beginning of this section.

ACKNOWLEDGMENTS

I am deeply indebted to Andrzej Trautman who focused my attention on the problem of encoding Einstein equations on natural bundles associated with a space–time. Without his influence this work would never have been completed. I would also like to express my gratitude to Paolo Budinich, Mike Crampin, Ludwik Dąbrowski, Jerzy Lewandowski, Roger Penrose, David Robinson, and Jacek Tafel for helping me at various stages of the present work. This research was partially supported by: Komitet Badań Naukowych (Grant No. 2 P03B 017 12), Consorzio per lo Sviluppo Internazionale dell Universita degli Studi di Trieste and the Erwin Schrödinger International Institute for Mathematical Physics.

[1] P. Nurowski, ''On a certain formulation of the Einstein equations,'' J. Math. Phys. **39**, 5477 (1998).
[2] D. Mumford, ''Tata lectures on Theta II,'' Prog. Math. **43**, 3.12–3.13 (1984).
[3] R. Penrose (private communications).
[4] E. Cartan, ''Sur les espaces conformes generalises at l'universe optique,'' Compt. Rendus Acad. Sci. Paris **174**, 857–859 (1922).
[5] R. Penrose, ''A spinor approach to general relativity,'' Ann. Phys. (N.Y.) **10**, 171–201 (1960).
[6] A. Z. Petrov, ''Classification of spaces defining gravitational fields,'' Sci. Not. Kazan State Univ. **114**, 55–69 (1954).

Perturbations in inflationary cosmologies with smooth exit

Daniel Tilley and Roy Maartens
School of Computer Science and Mathematics, University of Portsmouth, Portsmouth PO1 2EG, Britain

(Received 24 March 1998; accepted for publication 4 June 1998)

We give a comprehensive analysis of how scalar and tensor perturbations evolve in cosmologies with a smooth transition from power-law-like and de Sitter-like inflation to a radiation era. Analytic forms for the superhorizon and subhorizon perturbations in the inflationary and radiation dominated eras are found. © *1998 American Institute of Physics.* [S0022-2488(98)01110-4]

I. INTRODUCTION

In Ref. 1, the evolution of perturbations for a decaying vacuum cosmology given in Ref. 2, with smooth exit from inflation to radiation domination, was found exactly on superhorizon scales. The solution confirmed the standard result,[3–7] within gauge invariant perturbation theory that superhorizon density perturbations are strongly amplified through the transition from inflation to radiation. This transition is often approximated as an instantaneous jump,[3,4,7] since superhorizon modes change on a time scale that is much greater than the transition time. To avoid problems in the matching conditions for a jump transition, we consider only smooth transitions.

Given the variety of methods and models that have been used, we believe it is useful to do a comparative study using the same method on two broad classes of transition models, i.e., those based on power-law-like and de Sitter-like inflation. We use the class of power-law-like models first given in Ref. 8. The second class of models is a generalization of the models of Refs. 5 and 8 with a de Sitter-like inflationary era. The models, together with the necessary equations in perturbation theory, are discussed in Sec. II. In Sec. III we present our results for scalar and tensor perturbations on superhorizon scales, and show how they confirm the standard predictions. In Sec. IV we present solutions for scalar and tensor perturbations on subhorizon scales. However, toy models that evolve smoothly from inflation to radiation dominated eras are only applicable for superhorizon modes around the time of transition, since on subhorizon scales, the dynamics of the reheating era have a significant effect. The solutions for subhorizon scales therefore do not apply during the transition era. Section V gives some concluding remarks.

Throughout this paper we use units in which $8\pi G=1=c$.

II. THE MODELS AND PERTURBATION EQUATIONS

A flat Friedmann–Lemaitre–Robertson–Walker (FLRW) background has metric

$$ds^2=-dt^2+a^2(t)[dx^2+dy^2+dz^2], \tag{1}$$

where t is the physical time and $a(t)$ is the expansion scale factor. The background pressure p, energy density ρ, and Hubble parameter $H=\dot{a}/a$ (where a dot denotes a derivative with respect to cosmic time t) satisfy the field equations

$$p=-2\dot{H}-3H^2, \tag{2}$$

$$\rho=3H^2. \tag{3}$$

The pressure index and effective adiabatic sound speed are defined by

$$w=\frac{p}{\rho}, \tag{4}$$

 © 1998 American Institute of Physics

$$c_s^2=\frac{dp}{d\rho}=\frac{\dot{p}}{\dot{\rho}} \tag{5}$$

Inflation is characterized by $\ddot{a}>0$, so that the epoch t_e of exit is given by $\ddot{a}(t_e)=0$. The radiation dominated era is characterized by $p\approx\frac{1}{3}\rho$, $H\approx 1/2t$.

Scalar and tensor perturbations lead to a metric of the form

$$ds^2=a^2(\eta)\{-(1+2\Phi)d\eta^2+[(1-2\Psi)\gamma_{ij}+h_{ij}]dx^i dx^j\}, \tag{6}$$

in the longitudinal gauge, where η is conformal time $(d\eta=dt/a)$.[4] Here Φ and Ψ are gauge-invariant amplitudes of the scalar metric perturbations. The field equations require $\Phi=\Psi$ when the spatial part of the energy–momentum tensor is diagonal. The tensor perturbations are gauge invariantly described by $h_{ij}=hQ_{ij}$, where Q_{ij} is a symmetric traceless three tensor and h represents the amplitude of a gravitational wave.

The evolution of adiabatic scalar perturbations Φ is given by[4]

$$\frac{d^2\Phi}{d\eta^2}+3aH(1+c_s^2)\frac{d\Phi}{d\eta}-\nabla^2\Phi+3a^2H^2(c_s^2-w)\Phi=0. \tag{7}$$

Decomposing $\Phi(\eta,\vec{x})$ into eigenmodes $\tilde{\Phi}(\eta,\vec{k})$ of the comoving Laplacian, and then using the cosmic time t and scale factor a as the dynamical variables, respectively, gives

$$\frac{d^2\tilde{\Phi}}{dt^2}+3H\left(\frac{4}{3}+c_s^2\right)\frac{d\tilde{\Phi}}{dt}+\left[3H^2(c_s^2-w)+\left(\frac{k}{a}\right)^2\right]\tilde{\Phi}=0, \tag{8}$$

$$\frac{d^2\tilde{\Phi}}{da^2}+\frac{1}{a}\left(\frac{7}{2}+3c_s^2-\frac{3}{2}w\right)\frac{d\tilde{\Phi}}{da}+\left[\frac{3}{a^2}(c_s^2-w)+\frac{k^2}{a^4H^2}\right]\tilde{\Phi}=0, \tag{9}$$

where k is the comoving wave number. The decomposition allows us to follow the evolution of a single mode through the inflationary and radiation dominated eras. The physical length scale λ corresponding to the comoving wave number k is $\lambda=2\pi a/k$. This will equal the Hubble radius H^{-1} when

$$k=2\pi aH. \tag{10}$$

Superhorizon modes are characterized by $k\ll aH$. For these modes one can neglect the k^2 term in Eqs. (8) and (9) (and subsequent equations). The solutions of the truncated equations will be a good approximation to the limit of the solutions of the full equations, provided[9]

$$k^2\Phi\to 0 \quad \text{as} \quad k\to 0. \tag{11}$$

For adiabatic superhorizon scalar perturbations, the growing modes have a conserved quantity[10] which can be used to determine $\tilde{\Phi}$:

$$\zeta\equiv\tilde{\Phi}+\frac{2}{3(1+w)}\left(\tilde{\Phi}+\frac{1}{H}\frac{d\tilde{\Phi}}{dt}\right)\Rightarrow\tilde{\Phi}=-\zeta\frac{H}{a}\int\frac{a\dot{H}}{H^2}dt. \tag{12}$$

Equation (12) can be used to express the growing perturbations at late times in terms of their early time forms.[4] In Refs. 5–7, ζ is used to estimate the amplification of perturbations through a smooth transition. This amplification has been disputed in Ref. 11 (see also Refs. 12 and 13).

The density perturbations in the longitudinal gauge are given by[4]

$$\frac{\delta\rho}{\rho}=-2\left(\frac{k^2}{3a^2H^2}+1\right)\tilde{\Phi}-\frac{2}{H}\frac{d\tilde{\Phi}}{dt}. \tag{13}$$

The Fourier mode $\tilde{h}$ of the amplitude h of the tensor perturbations satisfies[4]

$$\frac{d^2\tilde{h}}{dt^2}+3H\frac{d\tilde{h}}{dt}+\left(\frac{k}{a}\right)^2\tilde{h}=0, \tag{14}$$

which may also be rewritten with a as the dynamical variable:

$$a^2\frac{d^2\tilde{h}}{da^2}+\left(4a+\frac{a^2}{H}\frac{dH}{da}\right)\frac{d\tilde{h}}{da}+\frac{k^2}{a^4H^2}\tilde{h}=0. \tag{15}$$

For superhorizon modes one can reduce to quadrature:

$$\tilde{h}=C+D\int\frac{dt}{a^3}=C+D\int\frac{da}{a^4H}, \tag{16}$$

where C and D are constant. It follows that D corresponds to the decaying mode and therefore there is no amplification in long-wavelength tensor perturbations.[3,4]

A. Power-law-like models

In Ref. 8 a simple class of power-law-like inflation models, which exit smoothly to a radiation era, is given by

$$a(t)=\frac{a_e}{\sqrt{t_e}}(1+m)^{(2n-1)/2}t^n(t+mt_e)^{(1-2n)/2}, \tag{17}$$

where the parameter n satisfies $n>1$ to achieve inflationary expansion, and

$$m=\frac{n+\sqrt{n(2n-1)}}{2n(n-1)}.$$

For $t\ll mt_e$,

$$a\approx a_e\left(1+\frac{1}{m}\right)^{(2n-1)/2}\left(\frac{t}{t_e}\right)^n,$$

showing the early power-law-like inflation era. For $t\gg mt_e$,

$$a\approx a_e(1+m)^{(2n-1)/2}\left(\frac{t}{t_e}\right)^{1/2},$$

as in a radiation dominated era. Thus we have a cosmology which evolves smoothly from power-law-like inflation to a radiation dominated era. The Hubble rate follows on differentiating Eq. (17):

$$H(t)=\frac{t+2nmt_e}{2t(t+mt_e)}. \tag{18}$$

Using Eqs. (2) and (3), Eq. (18) gives

$$p=\frac{t^2+4nmt_et-4n(3n-2)m^2t_e^2}{4t^2(t+mt_e)^2}, \tag{19}$$

$$\rho=\frac{3(t+2nmt_e)^2}{4t^2(t+mt_e)^2}. \tag{20}$$

It follows that the pressure index and effective adiabatic sound speed given by Eqs. (4) and (5) are

$$w=\frac{t^2+4nmt_et-4n(3n-2)m^2t_e^2}{3t^2+12nmt_et+12n^2m^2t_e^2}, \tag{21}$$

$$c_s^2=\frac{t(t+mt_e)(t+2nmt_e)-[t^2+4nmt_et-4n(3n-2)m^2t_e^2](2t+mt_e)}{3t(t+mt_e)(t+2nmt_e)-3(t+2nmt_e)^2(2t+mt_e)}. \tag{22}$$

The evolution of scalar and tensor perturbations for this model is given by Eqs. (8), (13), and (14). Using Eqs. (17), (18), (21), and (22), we get

$$t^2\frac{d^2\tilde{\Phi}}{dt^2}+t\left[\frac{t+2nmt_e}{2(t+mt_e)}\right]U(t)\frac{d\tilde{\Phi}}{dt}+\left[\frac{t+2nmt_e}{2(t+mt_e)}\right]^2V(t)\tilde{\Phi}=0, \tag{23}$$

where

$$U(t)=\frac{5t(t+mt_e)(t+2nmt_e)-(5t^2+20nmt_et+4n(n+2)m^2t_e^2)(2t+mt_e)}{t(t+mt_e)(t+2nmt_e)-(t+2nmt_e)^2(2t+mt_e)},$$

$$V(t)=\frac{t(t+mt_e)(t+2nmt_e)-[t^2+4nmt_et-4n(3n-2)m^2t_e^2](2t+mt_e)}{t(t+mt_e)(t+2nmt_e)-(t+2nmt_e)^2(2t+mt_e)}$$
$$-\frac{t^2+4nmt_et-4n(3n-2)m^2t_e^2}{(t+2nmt_e)^2}+k^2\frac{t_e(1+m)^{1-2n}t^{2-2n}(t+mt_e)^{2n+1}}{a_e^2(t+2nmt_e)^2},$$

and

$$\frac{\delta\rho}{\rho}=-2\left[\frac{4k^2t_e(1+m)^{1-2n}t^{2-2n}(t+mt_e)^{2n+1}}{3a_e^2(t+2nmt_e)^2}+1\right]\tilde{\Phi}-\left[\frac{4t(t+mt_e)}{t+2nmt_e}\right]\frac{d\tilde{\Phi}}{dt}, \tag{24}$$

and

$$t^2\frac{d^2\tilde{h}}{dt^2}+3t\left[\frac{t+2nmt_e}{2(t+mt_e)}\right]\frac{d\tilde{h}}{dt}+\left[\frac{k^2t_e(1+m)^{1-2n}(t+mt_e)^{2n-1}t^{2-2n}}{a_e^2}\right]\tilde{h}=0. \tag{25}$$

B. de Sitter-like models

In Ref. 8 a new exact inflationary solution is given, using a as the effective dynamical variable. The Hubble rate is given by

$$H(a)=\frac{2H_ea_e^2}{a^2+a_e^2}, \tag{26}$$

where H_e is the Hubble rate at exit. A similar model [but one which does not produce an exact solution for superhorizon scalar perturbations, unlike (26)[1]] is proposed in Ref. 5:

$$H(a)=\frac{2^{1/2}H_ea_e^2}{(a^4+a_e^4)^{1/2}}. \tag{27}$$

For both Hubble rates, when $a\ll a_e$, H is approximately constant as in a de Sitter inflation era, and for $a\gg a_e$, the Hubble rate decays like a^{-2}, as in a radiation dominated era. Thus for both models we have a cosmology which evolves smoothly from inflation to a radiation dominated era.

Here we generalize these models by considering the class of Hubble rates

$$H(a)=\frac{2^{1/\ell}a_e^2H_e}{(a^{2\ell}+a_e^{2\ell})^{1/\ell}}, \tag{28}$$

parametrized by ℓ ($\ell>0$). For Eq. (26), $\ell=1$, and for (27), $\ell=2$. Using Eq. (28), $H\approx$constant for $a\ll a_e$ (inflation), and $H\propto a^{-2}$ for $a\gg a_e$ (radiation domination), confirming that the model has the desired properties for all positive ℓ.

Using Eqs. (2) and (3), Eq. (28) gives

$$p=\frac{2^{2/\ell}H_e^2a_e^4(a^{2\ell}-3a_e^{2\ell})}{(a^{2\ell}+a_e^{2\ell})^{1+2/\ell}}, \tag{29}$$

$$\rho=\frac{2^{2/\ell}3H_e^2a_e^4}{(a^{2\ell}+a_e^{2\ell})^{2/\ell}}. \tag{30}$$

The pressure index and effective adiabatic sound speed given by Eqs. (4) and (5) are

$$w=\frac{1}{3}\left(\frac{a^{2\ell}-3a_e^{2\ell}}{a^{2\ell}+a_e^{2\ell}}\right), \tag{31}$$

$$c_s^2=\frac{1}{3}\left[\frac{a^{2\ell}-(2\ell+3)a_e^{2\ell}}{a^{2\ell}+a_e^{2\ell}}\right]. \tag{32}$$

Applying Eq. (10) to Eq. (28) and taking $k_e=2\pi a_eH_e$ (the comoving wave number of the Hubble radius at exit), the epoch a_- of leaving and a_+ of re-entering the Hubble radius are given exactly in this class of models by

$$\frac{a_\pm}{a_e}=\frac{k_e}{k}\left[1\pm\sqrt{1-\left(\frac{k}{k_e}\right)^{2\ell}}\right]^{1/\ell}. \tag{33}$$

Scales with $k\geqslant k_e$ never cross the Hubble radius and remain subhorizon. For superhorizon scales with $k\ll k_e$, Eq. (33) reduces to

$$\frac{a_\pm}{a_e}\approx\left(\frac{2^{1/\ell}k_e}{k}\right)^{\pm1}.$$

The scalar and tensor perturbations for this model are governed by Eqs. (9), (13), and (15). Using Eqs. (28), (31), and (32) we get

$$a^2\frac{d^2\tilde{\Phi}}{da^2}+a\left[\frac{4a^{2\ell}-(2\ell-2)a_e^{2\ell}}{a^{2\ell}+a_e^{2\ell}}\right]\frac{d\tilde{\Phi}}{da}+\left[\pi^2\left(\frac{k}{k_e}\right)^2\frac{(a^{2\ell}+a_e^{2\ell})^{2/\ell}}{2^{2/\ell-2}a_e^2a^2}-\frac{2\ell a_e^{2\ell}}{(a^{2\ell}+a_e^{2\ell})}\right]\tilde{\Phi}=0, \tag{34}$$

$$\frac{\delta\rho}{\rho}=-2\left[\frac{1}{3}\left(\frac{k}{k_e}\right)^2\frac{(a^{2\ell}+a_e^{2\ell})^{2/\ell}}{2^{2/\ell}a_e^2a^2}+1\right]\tilde{\Phi}-2a\frac{d\tilde{\Phi}}{da}, \tag{35}$$

and

$$a^2\frac{d^2\tilde{h}}{da^2}+2a\left[\frac{a^{2\ell}+2a_e^{2\ell}}{a^{2\ell}+a_e^{2\ell}}\right]\frac{d\tilde{h}}{da}+\left[\pi^2\left(\frac{k}{k_e}\right)^2\frac{(a^{2\ell}+a_e^{2\ell})^{2/\ell}}{2^{2/\ell-2}a_e^2a^2}\right]\tilde{h}=0. \tag{36}$$

III. SUPERHORIZON PERTURBATIONS

For the modes which leave the Hubble radius (during inflation), while they remain outside the Hubble radius, we can neglect the k-term in Eqs. (23)–(25), (34), (35) and (36). Only Eqs. (25) and (36), using (16), give exact solutions for the truncated equations. However, using Refs. 14 and 15 we can give analytic forms for the solutions in the inflationary and radiation dominated eras for Eqs. (23), (24), (34), and (35), since they reduce to Bessel forms. For all solutions, A_j, B_j, $j=1,2,...$ are used to denote arbitrary constants. For Eq. (11) to be satisfied it is required that for scalar perturbations A_j, $B_j \sim k^q$ as $k \to 0$, where $q > -2$. For all scalar perturbation solutions this is assumed to be the case.

A. Power-law-like model

For $t \ll mt_e$ Eq. (23) has solution

$$\tilde{\Phi} \approx A_1 + B_1 \left(\frac{t_e}{t} \right)^{n+1}, \tag{37}$$

describing modes that leave well before exit, while for $t \gg mt_e$

$$\tilde{\Phi} \approx A_2 + B_2 \left(\frac{t_e}{t} \right)^{3/2}, \tag{38}$$

describing modes that remain superhorizon in the radiation era. In Eqs. (37) and (38), A_j correspond to the constant modes, and B_j to the decaying modes.

The conserved quantity given by Eq. (12), applied to the case of the model given by (17), leads to

$$\tilde{\Phi} = \zeta \, \frac{(t+mt_e)^{(2n-3)/2}(t+2nmt_e)}{t^{n+1}} \int \frac{t^n(t^2+4nmt_e t+2nm^2 t_e^2)}{(t+2nmt_e)^2(t+mt_e)^{(2n-1)/2}} \, dt, \tag{39}$$

where we neglect the decaying contributions (and therefore no constant of integration arises from the integral). Amplification of the scalar growing modes according to Eqs. (37) and (38) is given by the factor $\alpha = A_2/A_1$. Using Eq. (39), we find that

$$A_1 = \frac{\zeta}{(n+1)},$$

and

$$A_2 = \frac{2}{3} \, \zeta.$$

Therefore

$$\alpha = \tfrac{2}{3}(n+1) > \tfrac{4}{3}, \tag{40}$$

so that amplification is only significant for large n. This is in agreement with results based on instantaneous transition models.[3,4,7]

The density perturbations in the inflationary and radiation dominated eras are found analytically using Eqs. (24), (37), and (38). For modes that leave the Hubble radius well before exit ($t \ll mt_e$)

$$\frac{\delta\rho}{\rho} \approx -2A_1 + \frac{2}{n} \, B_1 \left(\frac{t_e}{t} \right)^{n+1}, \tag{41}$$

while for modes that are still superhorizon during radiation domination ($t \gg mt_e$)

$$\frac{\delta\rho}{\rho}\approx -2A_2+4B_2\left(\frac{t_e}{t}\right)^{3/2}. \tag{42}$$

The density perturbations undergo the same amplification (40) as the potential perturbations.

The tensor perturbations are given by Eq. (16):

$$\tilde{h}=A_3+B_3\int t^{-3n}(t+mt_e)^{3(2n-1)/2}dt. \tag{43}$$

During power-law inflation ($t\ll mt_e$)

$$\tilde{h}\approx A_3+B_3\,\frac{m^{(6n-3)/2}t_e^{-1/2}}{1-3n}\left(\frac{t_e}{t}\right)^{3n-1}, \tag{44}$$

while for those modes that remain superhorizon during radiation domination ($t\gg mt_e$)

$$\tilde{h}\approx A_3-2B_3t^{-1/2}. \tag{45}$$

The results (43)–(45) imply that tensor perturbations do not grow.

B. de Sitter-like model

For modes that leave the Hubble radius when $a\ll a_e$, Eq. (34) has superhorizon solutions

$$\tilde{\Phi}\approx A_4\left(\frac{a}{a_e}\right)^{2\ell}+B_4\,\frac{a_e}{a}. \tag{46}$$

For superhorizon modes during radiation domination ($a\gg a_e$)

$$\tilde{\Phi}\approx\left(\frac{a_e}{a}\right)^{3/2}Z_{-3/2\ell}\left[-i\sqrt{\frac{2}{\ell}}\left(\frac{a_e}{a}\right)^{\ell}\right], \tag{47}$$

where Z_ν denotes a linear combination of the Bessel functions J_ν and Y_ν. Using the asymptotic forms of the Bessel functions,[14,15] Eq. (47) reduces to

$$\tilde{\Phi}\approx A_5+B_5\left(\frac{a_e}{a}\right)^3. \tag{48}$$

It follows that in the inflationary era, $|\tilde{\Phi}|$ grows as $a^{2\ell}$, while in the radiation dominated era, $|\tilde{\Phi}|$ is approximately constant (while the scales are still superhorizon). Thus growing superhorizon scalar perturbations are strongly amplified during inflation and then remain approximately constant after inflation. The amplification factor is clearly much greater than that for the power-law-like model.

We can calculate the amplification using Eq. (12), which gives

$$\tilde{\Phi}=\frac{2\zeta}{a(a^{2\ell}+a_e^{2\ell})^{1/\ell}}\int\frac{a^{2\ell}(a^{2\ell}+a_e^{2\ell})^{1/\ell}}{(a^{2\ell}+a_e^{2\ell})}\,da. \tag{49}$$

Evaluating the integral in the inflation and radiation eras and using Eqs. (46) and (48), we find

$$A_4=\frac{2\zeta}{2\ell+1},$$

and

$$A_5=\frac{2\zeta}{3}.$$

The amplification from the epoch a_i ($a_i \ll a_e$) to the radiation era is thus

$$\alpha = \frac{1}{3}(2\ell+1)\left(\frac{a_e}{a_i}\right)^{2\ell}. \tag{50}$$

The explicit dependence in (50) of the amplification on the initial epoch a_i shows that amplification, and thus the spectrum, is not scale invariant.

The density perturbations are given using Eqs. (35), (46), and (47). For modes beyond the Hubble radius during inflation

$$\frac{\delta\rho}{\rho} \approx -2(2\ell+1)A_4\left(\frac{a}{a_e}\right)^{2\ell}, \tag{51}$$

while for superhorizon modes during the radiation dominated era ($a \gg a_e$)

$$\frac{\delta\rho}{\rho} \approx 2i\sqrt{2\ell}\left(\frac{a_e}{a}\right)^{\ell+3/2} Z_{(2\ell-3)/2\ell}\left[-i\sqrt{\frac{2}{\ell}}\left(\frac{a_e}{a}\right)^{\ell}\right] - 2\left(\frac{a_e}{a}\right)^{3/2} Z_{-3/2\ell}\left[-i\sqrt{\frac{2}{\ell}}\left(\frac{a_e}{a}\right)^{\ell}\right], \tag{52}$$

which leads to

$$\frac{\delta\rho}{\rho} \approx -2A_5 + 4B_5\left(\frac{a_e}{a}\right)^3. \tag{53}$$

During inflation ($a \ll a_e$), $\delta\rho/\rho$ grows as $a^{2\ell}$, while in the radiation dominated era ($a \gg a_e$), $\delta\rho/\rho$ is approximately constant. The amplification is the same as (50).

The evolution of the tensor perturbations for superhorizon modes is given by Eq. (16):

$$\tilde{h} = A_6 + B_6 \int \frac{(a^{2\ell} + a_e^{2\ell})^{1/\ell} a_e}{a^4}\, da. \tag{54}$$

For the inflationary era ($a \ll a_e$) Eq. (54) gives

$$\tilde{h} \approx A_6 - \frac{1}{3} B_6\left(\frac{a_e}{a}\right)^3, \tag{55}$$

while for the radiation dominated era ($a \gg a_e$)

$$\tilde{h} \approx A_6 - B_6\left(\frac{a_e}{a}\right). \tag{56}$$

Unlike the scalar the tensor perturbations do not grow during inflation.

IV. SUBHORIZON PERTURBATIONS

For perturbations on scales inside the Hubble radius [where $k/(aH)$ is not negligible], Eqs. (23)–(25) and (34)–(36) cannot be solved exactly. We can, however, give analytic forms for the solutions in inflation and radiation domination using Refs. 14 and 15, since the equations reduce to Bessel forms.

A. Power-law-like model

Defining the comoving wave numbers

$$k_1 \equiv \frac{a_e(1-n)}{t_e(1+1/m)^{(1-2n)/2}}, \quad k_2 \equiv \frac{a_e}{2t_e(1+m)^{(1-2n)/2}},$$

we find the following. During power-law-like inflation ($t \ll m t_e$), the scalar and tensor perturbations evolve as

$$\tilde{\Phi}\approx\left(\frac{t_e}{t}\right)^{(n+1)/2} Z_{(n+1)/2(1-n)}\left[\frac{k}{k_1}\left(\frac{t}{t_e}\right)^{1-n}\right], \tag{57}$$

$$\frac{\delta\rho}{\rho}\approx\left[\frac{-2}{3}\left(\frac{k}{k_1}\right)^2(1-n)^2\left(\frac{t}{t_e}\right)^{(2-2n)}-2\right]\left(\frac{t_e}{t}\right)^{(n+1)/2} Z_{(n+1)/2(1-n)}\left[\frac{k}{k_1}\left(\frac{t}{t_e}\right)^{(1-n)}\right]$$
$$+\frac{2(1-n)}{n}\frac{k}{k_1}\left(\frac{t}{t_e}\right)^{(1-3n)/2} Z_{(3-n)/2(1-n)}\left[\frac{k}{k_1}\left(\frac{t}{t_e}\right)^{(1-n)}\right], \tag{58}$$

$$\tilde{h}\approx\left(\frac{t}{t_e}\right)^{(1-3n)/2} Z_{(3n-1)/2(1-n)}\left[\frac{k}{k_1}\left(\frac{t}{t_e}\right)^{1-n}\right]. \tag{59}$$

During radiation domination ($t\gg mt_e$), we find the following:

$$\tilde{\Phi}\approx\left(\frac{t_e}{t}\right)^{3/4} Z_{3/2}\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]$$
$$=\left(\frac{t_e}{t}\right)^{3/2}\left\{A_7\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]-B_7\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}$$
$$-\frac{k}{k_2}\left(\frac{t_e}{t}\right)\left\{A_7\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]+B_7\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}, \tag{60}$$

$$\frac{\delta\rho}{\rho}\approx\left[\frac{-8}{3}\left(\frac{k}{k_2}\right)^2\frac{t}{t_e}-2\right]\left(\frac{t_e}{t}\right)^{3/4} Z_{3/2}\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]+2\frac{k}{k_2}\left(\frac{t_e}{t}\right)^{3/4} Z_{5/2}\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]$$
$$=\left[\frac{-8}{3}\left(\frac{k}{k_2}\right)^2\frac{t}{t_e}-2\right]\left(\frac{t_e}{t}\right)^{3/2}\left\{A_7\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]-B_7\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}$$
$$+\left[\frac{8}{3}\left(\frac{k}{k_2}\right)^2\frac{t}{t_e}+2\right]\frac{t_e}{t}\frac{k}{k_2}\left\{A_7\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]+B_7\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}$$
$$-\frac{k}{k_2}A_8\left(\frac{t_e}{t}\right)^2\left\{3\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]-3\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right.$$
$$\left.-\left(\frac{k}{k_2}\right)^2\frac{t}{t_e}\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}-\frac{k}{k_2}B_8\left(\frac{t_e}{t}\right)^2\left\{3\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]+3\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right.$$
$$\left.-\left(\frac{k}{k_2}\right)^2\frac{t}{t_e}\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}, \tag{61}$$

$$\tilde{h}\approx\left(\frac{t_e}{t}\right)^{1/4} Z_{1/2}\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]=\left(\frac{t_e}{t}\right)^{1/2}\left\{A_9\sin\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]-B_9\cos\left[\frac{k}{k_2}\left(\frac{t}{t_e}\right)^{1/2}\right]\right\}. \tag{62}$$

As expected, density perturbations oscillate after re-entry, since radiation pressure balances gravitational infall. Note that while there are nondecaying oscillations in the density perturbations, the tensor perturbation oscillations are purely decaying. At late times, the density perturbation (61) has the form

$$\frac{\delta\rho}{\rho}\approx A\cos\left(\frac{k}{k_2}\frac{a}{a_e}\right)+B\sin\left(\frac{k}{k_2}\frac{a}{a_e}\right),$$

in agreement with the standard results for the radiation era (Ref. 16, p. 152).

B. de Sitter-like model

Subhorizon perturbations during inflation ($a \ll a_e$) have the form

$$\tilde{\Phi} \approx \left(\frac{a}{a_e}\right)^{(2\ell-1)/2} Z_{-(2\ell+1)/2}\left(-\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right), \tag{63}$$

$$\frac{\delta\rho}{\rho} \approx -\frac{2}{3}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\right)^2\left(\frac{a}{a_e}\right)^{(2\ell-5)/2} Z_{-(2\ell+1)/2}\left(-\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right)$$
$$-\frac{2\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\left(\frac{a}{a_e}\right)^{(2\ell-3)/2} Z_{-(2\ell+3)/2}\left(-\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right), \tag{64}$$

$$\tilde{h} \approx \left(\frac{a_e}{a}\right)^{3/2} Z_{-3/2}\left(-\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right)$$
$$= A_{10}\left[\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right) + \frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right)\right]$$
$$-B_{10}\left[\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right) + \frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a_e}{a}\right)\right]. \tag{65}$$

Subhorizon perturbations during radiation domination ($a \gg a_e$) are given by

$$\tilde{\Phi} \approx \left(\frac{a_e}{a}\right)^{3/2} Z_{3/2}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)$$
$$= A_{11}\left(\frac{a_e}{a}\right)^3\left[\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right) - \frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k}\frac{a}{a_e}\right)\right]$$
$$-B_{11}\left(\frac{a_e}{a}\right)^3\left[\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right) + \frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right.$$
$$\left.\times\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right], \tag{66}$$

$$\frac{\delta\rho}{\rho} \approx \left[-\frac{2}{3}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\right)^2\left(\frac{a}{a_e}\right)^2 - 2\right]\left(\frac{a_e}{a}\right)^{3/2} Z_{3/2}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)$$
$$+\frac{2\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\left(\frac{a_e}{a}\right)^{1/2} Z_{5/2}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)$$
$$= \left[\frac{-2}{3}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\right)^2\left(\frac{a}{a_e}\right)^2 - 2\right]\left(\frac{a_e}{a}\right)^3 A_{11}\left[\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right.$$
$$\left.-\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right] + \left[\frac{2}{3}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\right)^2\left(\frac{a}{a_e}\right)^2 + 2\right]$$
$$\times\left(\frac{a_e}{a}\right)^3 B_{11}\left[\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right) + \frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right]$$
$$-A_{12}\left(\frac{a_e}{a}\right)^3\left[\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)^2\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right.$$

$$-3\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)$$

$$+3\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\Bigg]+B_{12}\left(\frac{a_e}{a}\right)^3\Bigg[\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)^2$$

$$\times\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)-3\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)-\frac{3\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}$$

$$\times\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\Bigg], \tag{67}$$

$$\tilde{h}\approx\left(\frac{a_e}{a}\right)^{1/2}Z_{1/2}\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)=\frac{a_e}{a}\left[A_{13}\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right.$$

$$\left.-B_{13}\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)\right]. \tag{68}$$

Once again, we see that the density perturbations contain nondecaying oscillations while the tensor perturbation oscillations are purely decaying. At late times, the density perturbation (67) has the form

$$\frac{\delta\rho}{\rho}\approx A\cos\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right)+B\sin\left(\frac{\pi}{2^{(1-\ell)/\ell}}\frac{k}{k_e}\frac{a}{a_e}\right),$$

as in the power-law-like case.

V. CONCLUSIONS

Our main result is a comprehensive presentation of analytic solutions for the scalar and tensor perturbations on super- and subhorizon scales for two classes of cosmology, where the scale factor evolves smoothly from an inflationary era to a radiation dominated era. These classes encompass parametrized power-law-like and de Sitter-like inflationary behavior. Our results confirm the amplification of superhorizon scalar perturbations for all values of the parameters which define the classes of models. Explicit forms, given in Eq. (40) and (50), of the amplification for scalar perturbations show how the de Sitter-like inflation is much more effective in amplification than the power-law-like inflation. On subhorizon scales, the detailed form of oscillating modes after reentry was found, showing that there are nondecaying modes for density perturbations, but not for tensor perturbations. Taken together, these results provide a unified catalog of perturbation solutions in what can be taken as the two main types of inflationary models with smooth exit.

[1] R. Maartens and D. Tilley, Gen. Relativ. Gravit. **30**, 289 (1998).
[2] E. Gunzig, R. Maartens, and A. Nesteruk, Class. Quantum Grav. **15**, 923 (1998).
[3] N. Deruelle and V. Mukhanov, Phys. Rev. D **52**, 5549 (1995).
[4] V. Mukhanov, H. Feldman, and R. Brandenberger, Phys. Rep. **215**, 203 (1992).
[5] R. Caldwell, Class. Quantum Grav. **13**, 2437 (1996).
[6] M. Götz, Mon. Not. R. Astron. Soc. (to be published).
[7] J. Martin and D. Schwarz, Phys. Rev. D **57**, 3302 (1998).
[8] R. Maartens, D. Taylor, and N. Roussos, Phys. Rev. D **52**, 3358 (1995).
[9] H. Kodama and T. Hamazaki, Report, gr-qc/9712045 (1997).
[10] D. Lyth, Phys. Rev. D **31**, 1792 (1985).
[11] L. Grishchuk, Phys. Rev. D **50**, 7154 (1994).
[12] L. Grishchuk, Report, gr-qc/9801011 (1998).
[13] W. Unruh, Report, astro-ph/9802323 (1998).
[14] E. Kamke, *Differentialgleichen: LösungsMethoden Und Lösungen* (Teubner, Stuttgart, 1983).
[15] I. Gradshteyn and I, Ryzhik, *Tables of Integrals, Series, and Products* (Academic, London, 1980).
[16] T. Padmanabhan, *Structure Formation in the Universe* (Cambridge University Press, Cambridge, 1993).

Algebraic solutions for point groups: The tetrahedral group for the group chain $T \supset D_2 \supset C_2$

Jin-Quan Chen[a] and Peng-Dong Fan
Department of Physics, Nanjing University,
Nanjing 210008, The People's Republic of China

(Received 24 December 1997; accepted for publication 13 May 1998)

A factorization lemma is discovered for the irreducible symmetry operators (ISO's) of a semidirect product group of two abelian groups. By means of it, simple algebraic expressions of the $T \supset D_2$ single-valued irreducible matrices, ISO's and symmetry adapted functions are derived. Using the double-induced technique, the algebraic solutions for the double-valued representations of group T in the group chain $T \supset C_2$ are also obtained. The simplicity and elegance of the algebraic solution of the point group can be compared with that of the analytic solution of the rotation group. © *1998 American Institute of Physics.* [S0022-2488(98)02510-9]

I. INTRODUCTION

The importance of working with symmetry-adapted functions (SAF's) was realized by many early researchers of solid-state physics (e.g., von der Laga and Bethe[1] in the 1930's and 1940's). In later work, Altmann, and Cracknell[2–4] extended some of these early ideas to generally practical and computational levels. By using the projection operator method, Bradley and Cracknell[4] calculated the SAF's for both single-valued (s.v.) and double-valued (d.v.) reps of all crystal point groups. However, the d.v. SAF's are given in the uncoupled form (i.e., a linear combination of the products of Y_{lm} and the spin wave functions), which is not convenient. By using the building up method, Butler[5] calculated both the s.v. and d.v. SAF's for the dihedral group D_6, as well as the octahedral and icosahedral groups. The interest in the symmetry adapted functions has persisted to the present. Recently, by using a Gaussian integration technique Fernando *et al.*[6] obtained the symmetry adapted functions for identity reps of all crystal point groups.

Until recently the solutions to the SAF's problem were restricted to the numerical, and the results were tabulated for each irrep of each group chain for angular momentum up to a certain value (12 in Ref. 4 and 8 in Ref. 5). It is highly desirable to seek algebraic expressions for the SAF's of all point groups so that they have closed and simple formulas, and the SAF's for any high angular momentum can be readily obtained.

The lack of algebraic solutions to the SAF's problem is due to the lack of algebraic solutions to the irreducible symmetry operators (ISO's) (or the generalized projection operators), with which the SAF's are obtained by the projection operator method. One of the obstacles for obtaining algebraic expressions of the ISO's is associated with the conventional labeling schemes for irreps of point groups, such as the Mulliken, Koster,[7] or Butler[5] notations. Obviously a precondition for obtaining algebraic solutions is the use of quantum numbers to label irreps, just as the use of j for the irreps of SO(3) is. In 1977 a new approach to group representation theory was proposed,[8,9] in which an analogy with the three operators $(J^2, J_z, \overline{J_z})$ for the rotation group, three sets of the operators, $(C, C(s), \overline{C}(s))$ are introduced, whose eigenvalues $(\lambda, \rho, \overline{\rho})$, the counterpart of the quantum numbers $(j, m, \overline{m})$, are used for labeling or distinguishing the irreps, the components of irreducible basis vectors, and the multiple occurrence of the irrep λ in a given reducible representation space. The following three sets of operators, $C, (C, C(s)), (C, C(s), \overline{C}(s))$, are called the CSCO-I, CSCO-II, and CSCO-III, which are the complete set of commuting operators (CSCO) in the class space, any representation space which is multiplicity free, and the group space, respectively. The new approach not only put the representation theory of finite groups

[a]Electronic mail: jqchen@chenwang.nju.edu.cn

 © 1998 American Institute of Physics

(including the space groups) and compact Lie groups on the same footing, but also provides a powerful method, the *eigenfunction method*, for finding irreps and various coupling coefficients for finite and Lie groups.[9,10]

Very recently, a so-called double-induced technique was proposed and used, in conjunction with the *eigenfunction method*, for obtaining algebraic solutions of the tetrahedral group in the group chain $T \supset C_3$.[11] Simple algebraic expressions of the $T \supset C_3$ irreducible matrices, ISO's and SAF's are derived for both single-valued and double-valued reps in a unified way without using any knowledge of the irreducible matrices.

The purpose of this paper is to derive the algebraic solutions of the group T in the group chain $T \supset D_2$ for the s.v. reps and $T \supset C_2$ for the d.v. reps. For the former, the double-induced technique is not suitable, and a factorization lemma is first introduced for the ISO's of a semidirect product group of two abelian groups, and applied to the group chain $T \supset D_2$. For the $T \supset C_2$ d.v. reps, the double-induced technique can be used with slight extension. Extremely simple expressions are derived for the ISO's and SAF's, which are functions of only the quantum numbers characterizing the group chain instead of all the matrix elements. The high symmetry of the SAF's are disclosed for the first time and the time reversal property of the SAF's are also discussed.

II. A FACTORIZATION LEMMA FOR ISO'S OF A SEMIDIRECT PRODUCT GROUP

For the s.v. reps, the group chain $T \supset D_2$ is canonical and the subgroup C_2 is redundant in the group chain $T \supset D_2 \supset C_2$. Since the group T is a semidirect product of D_2 and C_3

$$T = D_2 \wedge C_3, \tag{1}$$

and both D_2 and C_3 are abelian, instead of using the more sophisticated double-induced representation technique, we can use a factorization lemma to be introduced below for obtaining its ISO in a straightforward way. For this purpose we need to first introduce a generalized ISO by

$$\mathcal{P}^{(\lambda)}_{\rho,\nu} = \frac{h_\lambda}{|G|} \sum_{a=1}^{|G|} \mathcal{D}^{(\lambda)}_{\rho\nu}(R_a)^* R_a, \tag{2a}$$

where h_λ is the dimension of the irrep λ, $|G|$ is the order of G, and $\mathcal{D}^{(\lambda)}_{\rho\nu}(R_a)$ is a generalized irreducible matrix element of R_a defined in Eqs. (3-212) of Ref. 9 between two irreducible basis vectors in two different classification schemes, the $G \supset G(s)$ basis ψ^λ_ρ and $G \supset G(s')$ basis φ^λ_ν, with $\rho(\nu)$ specifying the irreps of $G(s)(G(s'))$

$$\mathcal{D}^{(\lambda)}_{\rho\nu}(R_a) = \langle \psi^\lambda_\rho | R_a | \varphi^\lambda_\nu \rangle. \tag{2b}$$

The group chains $G \supset G(s)$ and $G \supset G(s')$ are assumed to be canonical so that the labeling (λ,ρ) or (λ,ν) for irreducible basis vectors is unique.

The generalized ISO obeys the eigenvalue equations[9]

$$\begin{pmatrix} C \\ C(s) \\ \overline{C}(s') \end{pmatrix} \mathcal{P}^{(\lambda)}_{\rho,\nu} = \begin{pmatrix} \lambda \\ \rho \\ \nu \end{pmatrix} \mathcal{P}^{(\lambda)}_{\rho,\nu}, \tag{2c}$$

where $(C, C(s))$ is the CSCO of the group chain $G \supset G(s)$, the analogy of (J^2, J_z) of the group chain $\mathrm{SO}(3) \supset \mathrm{SO}(2)$, while $\overline{C}(s')$ is obtained by replacing the elements R_a in the CSCO $C(s')$ of $G(s')$ by their intrinsic operators $\overline{R}_a$, which are defined in the group space by the following equation:

$$\overline{R}_a R_b = R_b R_a, \quad a,b = 1,2,\ldots,|G|, \tag{3}$$

which implies that

$$\overline{R}_a R_b R_c \cdots R_d = R_b R_c \cdots R_d R_a. \tag{4}$$

Notice that $\overline{C}(s')$ is the analogy of the third component of the angular momentum in the intrinsic (or body-fixed) coordinate system.[9]

The generalized ISO satisfies

$$\mathcal{P}^{(\lambda)}_{\rho,\nu}\mathcal{P}^{(\lambda')}_{\nu',\rho'}=\delta_{\lambda\lambda'}\delta_{\nu\nu'}\mathcal{P}^{(\lambda)}_{\rho,\rho'}, \tag{5a}$$

$$(\mathcal{P}^{(\lambda)}_{\rho,\nu})^{\dagger}=\mathcal{P}^{(\lambda)}_{\nu,\rho}. \tag{5b}$$

When $G(s')\to G(s), \overline{C}(s')\to\overline{C}(s), \varphi^{\lambda}_{\nu}\to\psi^{\lambda}_{\rho}, \mathcal{D}^{(\lambda)}_{\rho\nu}(R_a)\to D^{(\lambda)}_{\rho\rho}(R_a)$, the generalized ISO goes over to the conventional ISO $\mathcal{P}^{(\lambda)}_{\rho,\rho}$ and Eqs. (2) and (5) remain valid.

The generalized and conventional ISO's differ only in the intrinsic quantum numbers ν and $\overline{\rho}$, which serve as the multiplicity label. Since the multiplicity separation can be chosen freely, it gives us a flexibility in choosing the intrinsic quantum number. For example, it can be chosen in such a way so as to make the factorization of the ISO's as simple as possible.

Since the numerical factor $h_\lambda/|G|$ in the ISO $\mathcal{P}^{(\lambda)}_{\rho,\nu}$ or $\mathcal{P}^{(\lambda)}_{\rho,\overline{\rho}}$ is unimportant in applications except in the case of reading out the irreducible matrices from the ISO's, for the sake of clarity we are sloppy about the numerical factor of the ISO, but careful about its phase. For reading out the irreducible matrix elements from the ISO's it is convenient to write the ISO's in the normalized form

$$\mathcal{P}^{(\lambda)\overline{\rho}}_{\rho}=\sqrt{\frac{h_\lambda}{|G|}}\sum_{a=1}^{|G|}D^{(\lambda)}_{\rho\overline{\rho}}(R_a)^{*}R_a, \tag{6}$$

The normalization is easily found according to the assumption that the scalar product of two elements is $(R_a,R_b)=\delta_{ab}$ and will be inserted into the final expression of the ISO's.

Lemma: If H and K are two abelian groups with the ISO's P^ρ and $\wp^\nu$, respectively, then the ISO's of the semidirect product group $G=H\wedge K$ for four possible classifications are factored in the following way:

$$G\supset H, \overline{G}\supset\overline{K}: \quad \mathcal{P}^{(\lambda)}_{\rho,\nu}=P^\rho\wp^\nu, \tag{7a}$$

$$G\supset K, \overline{G}\supset\overline{H}: \quad \mathcal{P}^{(\lambda)}_{\nu,\rho}=\wp^\nu P^\rho, \tag{7b}$$

$$G\supset H, \overline{G}\supset\overline{H}: \quad \mathcal{P}^{(\lambda)}_{\rho,\rho'}=P^\rho\wp^\nu P^{\rho'}, \tag{7c}$$

$$G\supset K, \overline{G}\supset\overline{K}: \quad \mathcal{P}^{(\lambda)}_{\nu,\nu'}=\wp^\nu P^\rho\wp^{\nu'}. \tag{7d}$$

Proof: Since P^ρ and $\wp^\nu$ are the ISO of H and K, respectively, they are eigenvectors of the CSCO's of H and K, denoted by $C(s)$ and $C(s')$

$$C(s)P^\rho=\rho P^\rho, \quad C(s')\wp^\nu=\nu\wp^\nu. \tag{8}$$

Since $C(s)$ and $C(s')$ are the CSCO of H and K, they have $|H|$ and $|K|$ distinct eigenvalues of ρ and ν, respectively. The operators $C(s)$ and $C(s')$ can be easily found, for example, the generators of the abelian groups H and K can be chosen as $C(s)$ and $C(s')$, respectively. Now let us derive the eigenvalue equations satisfied by $P^\rho\wp^\nu$. Using (4), (8) and the fact that $\wp^\nu$ commutes with $C(s')$ we can show that $P^\rho\wp^\nu$ is a common eigenvector of $C(s)$ and $\overline{C}(s')$

$$\begin{pmatrix} C(s) \\ \overline{C}(s') \end{pmatrix}P^\rho\wp^\nu=\begin{pmatrix} C(s)P^\rho\wp^\nu \\ P^\rho\wp^\nu C(s') \end{pmatrix}=\begin{pmatrix} \rho \\ \nu \end{pmatrix}P^\rho\wp^\nu. \tag{9a}$$

It is to be noted that the set of operators $(C(s),\overline{C}(s'))$ has $|G|=|H|\times|K|$ sets of eigenvalues (ρ,ν) and for a given (ρ,ν) we have a unique eigenvector $P^\rho\wp^\nu$. In other words,$(C(s),\overline{C}(s'))$ is already a CSCO-III of G, and C is redundant. This can also be shown in the following way. Since

C commutes with $C(s)$ and $\overline{C}(s')$, $CP^{\rho}\wp^{\nu}$ remains an eigenvector of $(C(s),\overline{C}(s'))$ with the same eigenvalue set (ρ,ν). Due to the uniqueness of the solution in the eigenvalue Eq. (9a), $CP^{\rho}\wp^{\nu}$ is necessarily proportional to $P^{\rho}\wp^{\nu}$, i.e.,

$$CP^{\rho}\wp^{\nu}=\lambda P^{\rho}\wp^{\nu}. \tag{9b}$$

Thus $P^{\rho}\wp^{\nu}$ is a simultaneous eigenvector of $(C,C(s),\overline{C}(s'))$. Comparing (9) with (2c) we know that $P^{\rho}\wp^{\nu}$ is necessarily the ISO $\mathcal{P}^{(\lambda)}_{\rho,\nu}$ of G. It confirms the statement that for a semidirect product of two abelian groups, the operator C is redundant for determining the ISO of G. Actually, according to a theorem due to Dirac (see Theorem 3 in p. 78 of Ref. 12: A linear operator that commutes with each of a complete set of commuting observables is a function of those observables), C is necessarily a function of $C(s)$ and $\overline{C}(s')$, and thus λ is a function of ρ and ν.

Equation (7b) can be proved in the same way, while Eqs. (7c) and (7d) can be derived from Eq.(5a).

Remark 1: By comparing (7a) and (7b) with (7c) and (7d) it is seen that the factorization of the generalized ISO's is simpler than the conventional ones.

Remark 2: The ISO's $\mathcal{P}^{(\lambda)}_{\rho,\rho'}=P^{\rho}\wp^{\nu}P^{\rho'}$ with different values of ν differ at most by a ν-dependent factor so long as the irrep λ of G can be induced from ν. Therefore, to obtain $\mathcal{P}^{(\lambda)}_{\rho,\rho'}$ through using $P^{\rho}\wp^{\nu}P^{\rho'}$ we can choose any value ν which is compatible with λ. Similar remarks apply to $\mathcal{P}^{(\lambda)}_{\nu,\nu'}=\wp^{\nu}P^{\rho}\wp^{\nu'}$.

Remark 3: We have a unified expression $P^{\rho}\wp^{\nu}$ for the $G\supset H$ generalized ISO's for both one-dimensional and multidimensional irreps, but with the essential difference that the factors P^{ρ} and $\wp^{\nu}$ are interchangeable for one-dimensional irreps but not for multidimensional irreps. In other words, for one-dimensional irreps the generators of both the subgroups H and K have definite values ρ and ν, respectively, while for multi-dimensional irreps only the generator of H (not K) has the definite value ρ. Therefore, we can use (ρ,ν) or $A_{\rho\nu}$ as labels for the one-dimensional irreps, and write

$$\mathcal{P}^{A_{\rho\nu}}=P^{\rho}\wp^{\nu}=\wp^{\nu}P^{\rho}. \tag{10}$$

Remark 4: The factorization Eq. (10) for one dimensional irreps apply to any semidirect product group $G=H\wedge K$ without the restriction that the groups H and K be abelian.

To show this we need to introduce the concept of the so-called *representation group*. A *representation group* $\mathcal{G}$ is defined[9] as the matrix group formed by all representation matrices of an abstract group G in a given representation space (reducible or irreducible). Suppose $G=H\wedge K$, where H and K are nonabelian groups. Obviously in any one-dimensional irrep of G, the *representation group* $\mathcal{G}$ is abelian and is a *direct* product of the *representation groups* $\mathcal{H}$ and $\mathcal{K}$

$$\mathcal{G}=\mathcal{H}\times\mathcal{K}, \tag{11}$$

and thus any one-dimensional ISO of G can be expressed as a product of the one-dimensional ISO P^{ρ} of H and the one-dimensional ISO $\wp^{\nu}$ of K. Therefore (10) is valid in general cases and can be used, for example, to write down immediately the ISO's of the one-dimensional irreps of the octahedral group $O=T\wedge C_2$.

Notice that the reverse is not true. In other words, the product of the one-dimensional ISO's of nonabelian groups H and K is in general not an ISO of G.

III. THE $T\supset D_2$ s.v. REPRESENTATIONS

The starting point for finding algebraic solutions to the double point group $G^{\dagger}$ is the construction of its group table. Inconsistencies of the group table for the double point group $O^{\dagger}$ given in Ref. 4 have been pointed out in Refs. 13 and 11. A self consistent group table for the group $T^{\dagger}$ is constructed in Ref. 11 and given in Table I with the Bradley–Cracknell notation[4] for the group elements, which is identical to the group table given in Ref. 14, constructed using Altermann's quaternion realization of rotations.[13]

TABLE I. The group table of $T^{\dagger}$.

E	1	2	3	4	5	6	7	8	9	10	11	12
C_{2x}	2	$\tilde{1}$	4	$\tilde{3}$	$\tilde{8}$	7	$\tilde{6}$	5	10	$\tilde{9}$	12	$\widetilde{11}$
C_{2y}	3	$\tilde{4}$	$\tilde{1}$	2	$\tilde{6}$	5	8	$\tilde{7}$	11	$\widetilde{12}$	$\tilde{9}$	10
C_{2z}	4	3	$\tilde{2}$	$\tilde{1}$	$\tilde{7}$	$\tilde{8}$	5	6	12	11	$\widetilde{10}$	$\tilde{9}$
C_{31}^{+}	5	$\tilde{6}$	$\tilde{7}$	$\tilde{8}$	$\tilde{9}$	12	10	11	1	3	4	2
C_{32}^{+}	6	5	8	$\tilde{7}$	11	$\widetilde{10}$	12	9	$\tilde{3}$	1	$\tilde{2}$	4
C_{33}^{+}	7	$\tilde{8}$	5	6	12	9	$\widetilde{11}$	10	$\tilde{4}$	2	1	$\tilde{3}$
C_{34}^{+}	8	7	$\tilde{6}$	5	10	11	9	$\widetilde{12}$	$\tilde{2}$	$\tilde{4}$	3	1
C_{31}^{-}	9	12	10	11	1	$\tilde{2}$	$\tilde{3}$	$\tilde{4}$	$\tilde{5}$	7	8	6
C_{32}^{-}	10	$\widetilde{11}$	$\tilde{9}$	12	2	1	$\tilde{4}$	3	8	$\tilde{6}$	5	7
C_{33}^{-}	11	10	$\widetilde{12}$	$\tilde{9}$	3	4	1	$\tilde{2}$	6	8	$\tilde{7}$	5
C_{34}^{-}	12	$\tilde{9}$	11	$\widetilde{10}$	4	$\tilde{3}$	2	1	7	5	6	$\tilde{8}$

A. Irreducible symmetry operators

According to (7a), the generalized ISO, or the $T\supset D_2$ and $\overline{T}\supset\overline{C}_3$ ISO of the group $T=D_2\wedge C_3$, $\mathcal{P}^{(\lambda)}_{\nu\nu',\rho}\propto\mathcal{P}^{(\lambda)\rho}_{\nu\nu'}$, is a product of the projection operators $\wp^{\nu\nu'}$ and P^{ρ} for the groups D_2 and C_3

$$\mathcal{P}^{(\lambda)\rho}_{\nu\nu'}=\wp^{\nu\nu'}P^{\rho} \tag{12}$$

where

$$\wp^{\nu\nu'}=(e+\nu C_{2x})(e+\nu' C_{2y}),\nu,\nu'=\pm 1, \tag{13a}$$

$$P^{\rho}\equiv P^{\rho_\mu}=\sum_{k=0}^{2}(\rho_\mu^{*})^k(C_{31}^{+})^k,\quad \rho\equiv\rho_\mu=\exp\left(-\frac{2\pi\mu i}{3}\right),\quad \mu=0,\pm 1, \tag{13b}$$

The irrep λ in (12) is decided from the quantum numbers ν,ν', and ρ of C_{2x},C_{2y}, and C_{31}^{+}. As mentioned in Sec. II, the one-dimensional irreps are those in which all the operators, and hence the generators of T can have definite values. In order to find the values of ν,ν' and ρ in the one-dimensional reps of T, we examine the commutation relations between D_2 and C_3. Using Table I, it is easy to get the following results:

$$C_{31}^{+}C_{2i}=C_{2i+1}C_{31}^{+},\quad i=1,2,3(\text{mod } 3), \tag{14a}$$

where

$$C_{21}\equiv C_{2x},\quad C_{22}\equiv C_{2y},\quad C_{23}\equiv C_{2z}.$$

By taking the Hermitian conjugate of (14a), we get

$$C_{31}^{-}C_{2i}=C_{2i-1}C_{31}^{-},\quad i=1,2,3(\text{mod } 3). \tag{14b}$$

From the above we know that when $C_{2x}=C_{2y}=C_{2z}=1$, C_{31}^{+} and C_{31}^{-} commute with C_{2x},C_{2y},C_{2z}, and thus all elements of T can be diagonalized simultaneously. Therefore, we know that in the one-dimensional irreps of T, $\nu=\nu'=1$, and $C_{31}^{+}=\rho_\mu,\mu=0,\pm 1$, which correspond to three one-dimensional irreps of T, designated as A_μ in the modified Mulliken notation introduced in Ref. 11, with the ISO's

$$\mathcal{P}^{(A_\mu)}=\frac{1}{\sqrt{12}}\wp^{++}P^{\rho_\mu},\quad \mu=0,\pm 1, \tag{15a}$$

where $\wp^{++}=1+C_{2x}+C_{2y}+C_{2z}$ is the projection operator for the identity representation of the group D_2. The irreps A_0,A_1,A_{-1} correspond to the irrep $(A,{}^1E,{}^2E)$ in Mulliken notation, or $(\Gamma_1,\Gamma_2,\Gamma_3)$ in Koster notation.

Excluding the value $(+,+)$, the set of operators (C_{2x},C_{2y}) takes three possible values, $(+-,-+,--)$. Since for the group T there is only a single three-dimensional irrep, the irrep F,we must have

$$\mathcal{P}^{(F)\rho}_{\nu\nu'}=\wp^{\nu\nu'}P^{\rho}, \tag{15b}$$

where $\nu\nu'$ is the component index, while ρ is the index which distinguishes the equivalent irreps. Usually only one set of the ISO's are needed, which can be conveniently chosen as $\rho=1$, thus the $T\supset D_2$ generalized ISO's are

$$\mathcal{P}^{(F)1}_{\nu\nu'}=\wp^{\nu\nu'}(1+C^{+}_{31}+C^{-}_{31}),\quad (\nu,\nu')=(+-,-+,--). \tag{15c}$$

The $T\supset D_2$ conventional ISO can be obtained from the generalized one as

$$\mathcal{P}^{(F)\overline{\nu\nu'}}_{\nu\nu'}=\mathcal{P}^{(F)1}_{\nu\nu'}(\mathcal{P}^{(F)1}_{\nu\nu'})^{\dagger}=\wp^{\nu\nu'}(1+C^{+}_{31}+C^{-}_{31})\wp^{\overline{\nu\nu'}}. \tag{15d}$$

Equation (15) gives the algebraic expressions of the s.v. $T\supset D_2$ ISO's. In order to compare our results with those derived by Folland,[15] we will transform Eq. (15d) into another form. Using (14) we have

$$C^{+}_{31}\wp^{\nu,\nu'}=\wp^{\nu\nu',\nu}C^{+}_{31},\quad C^{-}_{31}\wp^{\nu,\nu'}=\wp^{\nu',\nu\nu'}C^{-}_{31}. \tag{16a}$$

Using (15d) and (16a)

$$\mathcal{P}^{(F)\overline{\nu\nu'}}_{\nu\nu'}=\wp^{\nu\nu'}(\wp^{\overline{\nu\nu'}}+\wp^{\overline{\nu\nu'},\overline{\nu}}C^{+}_{31}+\wp^{\overline{\nu'},\overline{\nu\nu'}}C^{-}_{31}). \tag{16b}$$

If the ISO's $\mathcal{P}^{(F)\overline{\nu\nu'}}_{\nu\nu'}$ with proper normalization are regarded as the elements of a matrix $\mathbf{P}^{(F)}$ with rows $\nu\nu'$ and columns $\overline{\nu\nu'}$ in the order of $+-,-+,--$, then from (16b) we have

$$\mathbf{P}^{(F)}=\frac{1}{2}\begin{pmatrix} \wp_{+-}, & \wp_{+-}C^{-}_{31}, & \wp_{+-}C^{+}_{31} \\ \wp_{-+}C^{+}_{31}, & \wp_{-+}, & \wp_{-+}C^{-}_{31} \\ \wp_{--}C^{-}_{31}, & \wp_{--}C^{+}_{31}, & \wp_{--} \end{pmatrix}. \tag{16c}$$

Equation (16c) has been obtained by using the induction method[15] in a less transparent way.

B. The irreducible matrices

Since (ν,ν') or $(C_{2x},C_{2y})=(1,-1),(-1,1),(-1,-1)$, the matrix for $C_{2z}=C_{2x}C_{2y}$ is diagonal with elements $(-1,-1,1)$. From (15c) and (16a) we can obtain the irreducible matrix of C^{+}_{31}. Putting them together we have

$$\mathcal{D}^{F}(C_{2z})=\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\quad \mathcal{D}^{F}(C^{+}_{31})=\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{17}$$

C. The $T\supset D_2$ s.v. SAF's

We choose the coordinate system for the group chain $T\supset D_2$ as shown in Fig. 1. The Euler angles (α,β,γ) for $C_{2x},C_{2y},C_{2z},C^{+}_{31}$, and C^{-}_{31} are given in Table II of Ref. 11 and listed below:

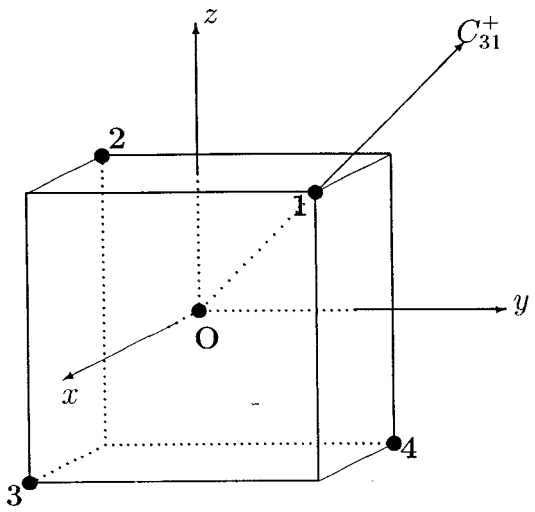

FIG. 1. The coordinate system xyz for $T\supset D_2$, with **1, 2, 3**, and **4** indicating the four vertices of the terahedral. The axes of C_{31}^{+}–C_{34}^{+} are from the origin to the vertices **1**–**4**.

$$\begin{array}{ccccc} C_{2x} & C_{2y} & C_{2z}^{-} & C_{31}^{+} & C_{31}^{-} \\ (0,\pi,\pi) & (0,\pi,0) & (0,0,\pi) & \left(0,\frac{\pi}{2},\frac{\pi}{2}\right) & \left(\frac{\pi}{2},\frac{\pi}{2},3\pi\right) \end{array}.$$

Applying $\mathcal{P}^{(A_\mu)}$ in (15a) and $\mathcal{P}^{(F)--}_{\nu\nu'}$ (noting that the choice of $\overline{\nu\nu'}$ is arbitrary), which are given by the third column of (16c), to Y_{lm} and using the following results[16]

$$C_{2x}|jm\rangle=(-1)^{-j}|j-m\rangle,\quad C_{31}^{+}|j\bar{m}\rangle=\sum_m d^j_{mm}\left(\frac{\pi}{2}\right)i^{-\bar{m}}|jm\rangle,$$

$$C_{2y}|jm\rangle=(-1)^{j-m}|j-m\rangle,\ C_{31}^{-}|j\bar{m}\rangle=(-1)^{-\bar{m}}\sum_m i^{-m}d^j_{mm}\left(\frac{\pi}{2}\right)|jm\rangle,$$

as well as the fact that $i^{-m}=i^m, i^{-\bar{m}}=i^{\bar{m}}$ for even m and $\bar{m}$, we obtain the $T\supset D_2$ s.v. unnormalized SAF's:

One-dimensional irrep

$$\varphi^{(A_\mu)\bar{m}}=Y_{l\bar{m}}+(-1)^lY_{l-\bar{m}}+(1+(-1)^l)[\rho_\mu^* i^{\bar{m}}+\rho_\mu]d^l_{0\bar{m}}\left(\frac{\pi}{2}\right)Y_{l0}$$

$$+2\sum_{m=2,4,6,\ldots}[\rho_\mu^* i^{\bar{m}}+\rho_\mu i^m]d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)[Y_{lm}+(-1)^lY_{l-m}],\ \ \bar{m}=\text{even}. \tag{18a}$$

Three-dimensional irrep

With the conventional labeling (x,y,z) in place of $(+-,-+,--)$ for the components of the irrep F, we have

$$\varphi_x^{(F)\bar{m}}=i^{\bar{m}}\sum_{m=1,3,5,\ldots}d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)[Y_{lm}+(-1)^lY_{l-m}],$$

$$\varphi_y^{(F)\bar{m}}=-\sum_{m=1,3,5,\ldots}i^m d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)[Y_{lm}-(-1)^lY_{l-m}],\quad \bar{m}=\text{even} \tag{18b}$$

$$\varphi_z^{(F)\bar{m}}=Y_{l\bar{m}}-(-1)^lY_{l-\bar{m}}.$$

Notice that the quantum number $\bar{m}$ in (18) serves as the multiplicity label (for more detail, see Sec. IV E in Ref. 11). Equation (18) is the sought-for algebraic expression for the s.v. $T\supset D_2$ SAF's.

The s.v. $T\supset D_2$ SAF's have been calculated by Bradley and Cracknell with the projection operator method and tabulated in pp. 66–70 of Ref. 4 with accuracy to 12 digits for l up to 12. To facilitate a comparison with these tables, we use the real functions Y^c_{lm} and Y^s_{lm} defined in Ref. 4

$$Y^{(+)}_{lm}\equiv Y^c_{lm}=\frac{1}{\sqrt{2}}(Y_{lm}+Y_{l,-m})_{B-C},\quad Y^{(-)}_{lm}\equiv Y^s_{lm}=\frac{-i}{\sqrt{2}}(Y_{lm}-Y_{l,-m})_{B-C},\ \ m>0. \tag{19}$$

The $(Y_{lm})_{B-C}$ defined in Ref. 4 is related to the Y_{lm}[16] in Condon–Shortley phase convention as

$$(Y_{lm})_{B-C}=(-1)^{(m+|m|)/2}Y_{lm}. \tag{20}$$

Equation (18) can be rewritten as

$$\varphi^{(A_\mu)\bar{m}}=\sqrt{1+\delta_{\bar{m}0}}\,Y^{(\pi_l)}_{l\bar{m}}+\frac{1}{\sqrt{2}}(1+\pi_l)(i^{\bar{m}}\rho^*_\mu+\rho_\mu)d^l_{0\bar{m}}\left(\frac{\pi}{2}\right)Y^{(\pi_l)}_{l0}$$

$$+2\sum_{m=2,4,6,\ldots}(i^{\bar{m}}\rho^*_\mu+i^m\rho_\mu)d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)Y^{(\pi_l)}_{lm},\quad \mu=0,\pm1, \tag{21a}$$

$$\varphi^{(F)\bar{m}}_x=-i^{\bar{m}}\sum_{m=1,3,5,\ldots}d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)Y^{(\pi_{l+1})}_{lm},$$

$$\varphi^{(F)\bar{m}}_y=(-1)^l\sum_{m=1,3,5,\ldots}i^{m-1}d^l_{m\bar{m}}\left(\frac{\pi}{2}\right)Y^{(\pi_l)}_{lm},\quad \varphi^{(F)\bar{m}}_z=Y^{(\pi_{l+1})}_{l\bar{m}}, \tag{21b}$$

with $\pi_l=+(-)$ for $l=$ even (odd) and

$$\bar{m}=\begin{cases}0,2,4,\ldots\\ 2,4,6,\ldots\end{cases}\quad\text{for}\quad\begin{cases}A_\mu,\ l=\text{even};\quad F,\ l=\text{odd},\\ A_\mu,\ l=\text{odd};\quad F,\ l=\text{even}.\end{cases} \tag{21c}$$

It is noted that the SAF's $\varphi^{(A_0)}$ and$\varphi^{(F)}$ are real, while $\varphi^{(A_1)}=(\varphi^{(A_{-1})})^*$, i.e., under the time reversal $\hat{T}$ we have

$$\hat{T}\varphi^{(F)}_k=\varphi^{(F)}_k,\ \ k=x,y,z,$$
$$\hat{T}\varphi^{(A_0)}=\varphi^{(A_0)},\quad \hat{T}\varphi^{(A_1)}=(\varphi^{(A_1)})^*=\varphi^{(A_{-1})}. \tag{22}$$

With these algebraic expressions, it is easy to find the exact numerical expression of the SAF's $\psi^{(\Gamma)\bar{m}}_\mu$ for any l with the help of some software, say the Maple software. The $T\supset D_2$ SAF's for $j=1$–3, and a part for $j=6,8,13$ are listed in Table II, which is produced automatically. Notice that the SAF's $\psi^{(\Gamma)\bar{m}}_\mu$ with different multiplicity quantum numbers $\bar{m}$ may be neither orthogonal nor linearly independent. However, the SAF's in (21) happen to be orthogonal in the quantum number $\bar{m}$, and their numerical values are identical to the results in Ref. 4.

D. The relation between the $T\supset C_3$ and $T\supset D_2$ bases

Now we are going to find the transformation matrix between the bases of $T\supset C_3$ and $T\supset D_2$. From the $T\supset C_3$ irreducible matrices of C^+_{31} and C_{2z} given in (37) of Ref. 11 and the corresponding matrices for the group chain $T\supset D_2$ given in (17) we can easily find the similarity transformation matrix between the two bases

TABLE II. The $T\supset D_2$ s.v. SAF's. The superscript is the multiplicity label. For $l=6$, 8, and 13 only the SAF's of one-dimensional irreps with multiplicity are listed, and the SAF's $|A_{-1}\rangle(=|A_1\rangle^*)$ are not listed.

$$\bar{m}=0 \quad |F,x\rangle = |11,c\rangle$$
$$|F,y\rangle = |11,s\rangle$$
$$|F,z\rangle = |10,c\rangle$$

$$\bar{m}=0 \quad |A_1\rangle = \frac{\sqrt{2}}{2}|20,c\rangle + i\frac{\sqrt{2}}{2}|22,c\rangle$$

$$\bar{m}=2 \quad |F,x\rangle = |21,s\rangle$$
$$|F,y\rangle = |21,c\rangle$$
$$|F,z\rangle = |22,s\rangle$$

$$\bar{m}=2 \quad |A_0\rangle = |32,s\rangle$$
$$\bar{m}=0 \quad |F,x\rangle^1 = -\frac{\sqrt{2\cdot 3}}{4}|31,c\rangle + \frac{\sqrt{2\cdot 5}}{4}|33,c\rangle$$
$$|F,y\rangle^1 = -\frac{\sqrt{2\cdot 3}}{4}|31,s\rangle - \frac{\sqrt{2\cdot 5}}{4}|33,s\rangle$$
$$|F,z\rangle^1 = |30,c\rangle$$

$$\bar{m}=2 \quad |F,x\rangle^2 = -\frac{\sqrt{2\cdot 5}}{4}|31,c\rangle - \frac{\sqrt{2\cdot 3}}{4}|33,c\rangle$$
$$|F,y\rangle^2 = \frac{\sqrt{2\cdot 5}}{4}|31,s\rangle - \frac{\sqrt{2\cdot 3}}{4}|33,s\rangle$$
$$|F,z\rangle^2 = |32,c\rangle$$

$$\bar{m}=0 \quad |A_0\rangle^1 = \frac{\sqrt{2}}{4}|60,c\rangle - \frac{\sqrt{2\cdot 7}}{4}|64,c\rangle$$

$$\bar{m}=2 \quad |A_0\rangle^2 = \frac{\sqrt{11}}{4}|62,c\rangle - \frac{\sqrt{5}}{4}|66,c\rangle$$

$$\bar{m}=0$$
$$|A_1\rangle^1 = \frac{\sqrt{2\cdot 31}}{16}|80,c\rangle - i\frac{\sqrt{3\cdot 5\cdot 7\cdot 31}}{124}|82,c\rangle - \frac{\sqrt{2\cdot 7\cdot 11\cdot 31}}{248}|84,c\rangle - i\frac{\sqrt{11\cdot 13\cdot 31}}{124}|86,c\rangle - \frac{\sqrt{2\cdot 5\cdot 11\cdot 13\cdot 31}}{496}|88,c\rangle$$
$$\bar{m}=2$$
$$|A_1\rangle^2 = i\frac{\sqrt{3\cdot 5\cdot 7}}{32}|80,c\rangle + \frac{\sqrt{2}}{2}|82,c\rangle + i\frac{\sqrt{3\cdot 5\cdot 11}}{48}|84,c\rangle - i\frac{\sqrt{3\cdot 7\cdot 11\cdot 13}}{96}|88,c\rangle$$

$$\bar{m}=2$$
$$|A_0\rangle^1 = \frac{\sqrt{2\cdot 3\cdot 5\cdot 19}}{48}|13\ 2,s\rangle - \frac{\sqrt{3\cdot 11\cdot 17}}{48}|13\ 6,s\rangle - \frac{\sqrt{3\cdot 17\cdot 23}}{48}|13\ 10,s\rangle$$
$$\bar{m}=4$$
$$|A_0\rangle^2 = \frac{5\sqrt{3\cdot 19}}{48}|13\ 4,s\rangle + \frac{\sqrt{2\cdot 3\cdot 5}}{24}|13\ 8,s\rangle - \frac{\sqrt{3\cdot 11\cdot 23}}{48}|13\ 12,s\rangle$$
$$\bar{m}=2$$
$$|A_1\rangle^1 = \frac{17\sqrt{3}}{48}|13\ 2,s\rangle + i\frac{3\sqrt{2\cdot 5\cdot 11\cdot 17}}{544}|13\ 4,s\rangle + \frac{\sqrt{2\cdot 3\cdot 5\cdot 11\cdot 17\cdot 19}}{1632}|13\ 6,s\rangle + i\frac{\sqrt{11\cdot 17\cdot 19}}{136}|13\ 8,s\rangle$$
$$+\frac{\sqrt{2\cdot 3\cdot 5\cdot 17\cdot 19\cdot 23}}{1632}|13\ 10,s\rangle + i\frac{\sqrt{2\cdot 5\cdot 17\cdot 19\cdot 23}}{544}|13\ 12,s\rangle$$
$$\bar{m}=4$$
$$|A_1\rangle^2 = -i\frac{3\sqrt{5\cdot 11\cdot 17}}{16\sqrt{293}}|13\ 2,s\rangle + \frac{293\sqrt{2\cdot 3}}{96\sqrt{293}}|13\ 4,s\rangle - i\frac{53\sqrt{2\cdot 19}}{32\sqrt{293}}|13\ 6,s\rangle - \frac{5\sqrt{3\cdot 5\cdot 19}}{24\sqrt{293}}|13\ 8,s\rangle$$
$$+i\frac{\sqrt{2\cdot 11\cdot 19\cdot 23}}{32\sqrt{293}}|13\ 10,s\rangle + \frac{5\sqrt{2\cdot 3\cdot 11\cdot 19\cdot 23}}{96\sqrt{293}}|13\ 12,s\rangle$$

$$\begin{pmatrix} \varphi_x^F \\ \varphi_y^F \\ \varphi_z^F \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} \epsilon & 1 & \epsilon^* \\ \epsilon^* & 1 & \epsilon \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \psi_1^F \\ \psi_0^F \\ \psi_{-1}^F \end{pmatrix}. \tag{23}$$

IV. THE $T\supset C_2$ d.v. REPRESENTATIONS

For the d.v. reps, the group chain $T\supset C_2$ is canonical and D_2 is redundant in the group chain $T\supset D_2\supset C_2$. To find the algebraic solutions for the group chain $T\supset C_2$ we need to use the

double-induced technique. We follow the six steps in Sec. II of Ref. 11 to find the algebraic solution of $T \supset C_2$. We first note that the CSCO of T is

$$C = C_{31}^{+} + C_{32}^{+} + C_{33}^{+} + C_{34}^{+} = \mathbf{5+6+7+8}, \tag{24}$$

where (and in the following) we use boldface integers to denote the group element listed in Table I, e.g., **5** represents the element $R_5 = C_{31}^{+}$. In the d.v. reps, C has three distinct eigenvalues $(2, 2\epsilon, 2\epsilon^*)$, $\epsilon = e^{-2\pi i/3}$, corresponding to the three irreps $(E_{1/2}, E_{3/2}, E_{-3/2})$ in the modified Mulliken notation,[11] or $(\Gamma_5, \Gamma_6, \Gamma_7)$ in Koster notation, respectively.

A. The C_2 double-induced representation

We choose the subgroup $C_2 = (e, C_{2z})$, and the CSCO of C_2 is

$$C_{2z} = \exp(-\pi J_z i). \tag{25}$$

Its eigenvalue is denoted by ϱ,

$$\varrho \equiv \varrho_\mu = \exp(-\pi\mu i) = \begin{cases} -i \\ i \end{cases}, \quad \text{for} \quad \mu = \begin{cases} \frac{1}{2} \\ -\frac{1}{2} \end{cases}. \tag{26}$$

Notice the subtle difference between the notations for the quantum number $\rho_\mu = \exp[-2\pi\mu i/3]$ of C_{31}^{+} and the quantum number $\varrho_\mu = \exp(-\mu\pi i)$ of C_{2z}.

The projection operators of C_2 is defined as

$$P^{\varrho} \equiv P^{\varrho_\mu} \equiv P^{\mu} = e + \varrho^* C_{2z}. \tag{27}$$

The group T can be expanded in double cosets of C_2

$$T = C_2(e + C_{2x} + C_{31}^{-} + C_{31}^{+})C_2. \tag{28}$$

According to (9c) in Ref. 11, sandwiching the four coset generators between P^{ϱ} and $P^{\bar{\varrho}}$, and inserting the normalization coefficients we get four common eigenvectors of C_{2z} and $\overline{C}_{2z}$

$$\begin{aligned} \varphi^1_{\varrho,\bar{\varrho}} &= \frac{1}{2\sqrt{2}} P^{\varrho} e P^{\bar{\varrho}}, \quad \varphi^2_{\varrho,\bar{\varrho}} = \frac{1}{2\sqrt{2}} P^{\varrho} C_{2x} P^{\bar{\varrho}}, \\ \varphi^3_{\varrho,\bar{\varrho}} &= \tfrac{1}{2} P^{\varrho} C_{31}^{-} P^{\bar{\varrho}}, \quad \varphi^4_{\varrho,\bar{\varrho}} = \tfrac{1}{2} P^{\varrho} C_{31}^{+} P^{\bar{\varrho}}. \end{aligned} \tag{29a}$$

Using

$$P^{\varrho} P^{\varrho'} = 3\delta_{\varrho,\varrho'} P^{\varrho}, \quad C_{2x} P^{\varrho} = P^{\varrho^*} C_{2x}$$

we have

$$\varphi^1_{\varrho\bar{\varrho}} = \delta_{\varrho\bar{\varrho}} \varphi^1_{\varrho}, \quad \varphi^2_{\varrho\bar{\varrho}} = \delta_{\varrho\bar{\varrho}^*} \varphi^2_{\varrho}, \tag{29b}$$

with

$$\varphi^1_{\varrho} = \frac{1}{\sqrt{2}} P^{\varrho}, \quad \varphi^2_{\varrho} = \frac{1}{\sqrt{2}} P^{\varrho} C_{2x}. \tag{29c}$$

Using Table I, the four basis vectors can be written explicitly

$$\varphi^1_{\varrho} = \frac{1}{\sqrt{2}}[\mathbf{1} - \varrho\mathbf{4}], \quad \varphi^2_{\varrho} = \frac{1}{\sqrt{2}}[\mathbf{2} - \varrho\mathbf{3}],$$

$$\varphi^3_{\varrho\bar{\varrho}}=\tfrac{1}{2}[\mathbf{9}-\varrho\mathbf{12}-\bar{\varrho}\mathbf{11}-\varrho\bar{\varrho}\mathbf{10}],\quad \varphi^4_{\varrho\bar{\varrho}}=\tfrac{1}{2}[\mathbf{5}+\varrho\mathbf{7}+\bar{\varrho}\mathbf{8}-\varrho\bar{\varrho}\mathbf{6}], \tag{30}$$

where we used the fact that for d.v. reps, $\varrho^*=-\varrho$ and $\varrho\bar{\varrho}$ is real.

We choose the elements **1, 2, 9**, and **5** as the representatives of $\varphi^1_{\varrho\bar{\varrho}}-\varphi^4_{\varrho\bar{\varrho}}$, respectively. The four basis vectors $\varphi^1_{\varrho\bar{\varrho}}-\varphi^4_{\varrho\bar{\varrho}}$ form the double-induced rep. The task of finding the eigenvectors of the CSCO-III in the 12-dimensional group space of T is reduced to finding the eigenvector of the CSCO-I in the four-dimensional double-induced rep space.

Following the same steps as in Sec. IV A of Ref. 11 we can find the rep matrix M of the operator C, the CSCO of T, in the basis (30), which is defined as $C\varphi^i_{\rho\bar{\rho}}=\Sigma_j M_{ji}\varphi^j_{\rho\bar{\rho}}$.

(1) Highlighting the elements **1**, **2**, **9**, and **5** in rows 5–8 of Table I which correspond to the elements $C^+_{31}-C^+_{34}$. (2) Collecting the highlighted numbers in each column, we obtain the following table:

R_i	1	2	3	4	5	6	7	8	9	10	11	12
CR_i	5	5	5	5	−9	9	9	9	1−2	1+2	1−2	1+2

The table shows e.g., $CR_5=-R_9+\cdots$. The terms involving elements other than R_1, R_2, R_5, and R_9 are ignored. Using the above table and (30) we have

$$C\varphi^1_{\varrho}=\frac{1}{\sqrt{2}}[\mathbf{5}-\varrho\mathbf{5}]=\sqrt{2}(1-\varrho)\varphi^4_{\varrho\bar{\varrho}},$$

$$C\varphi^2_{\varrho}=\frac{1}{\sqrt{2}}[\mathbf{5}-\varrho\mathbf{5}]=\sqrt{2}(1-\varrho)\varphi^4_{\varrho\varrho^*}, \tag{31}$$

$$C\varphi^3_{\varrho\bar{\varrho}}=\tfrac{1}{2}[\mathbf{1}-\mathbf{2}-\varrho(\mathbf{1}+\mathbf{2})-\bar{\varrho}(\mathbf{1}-\mathbf{2})-(\varrho\bar{\varrho})(\mathbf{1}+\mathbf{2})]=f_{\varrho\bar{\varrho}}\varphi^1_{\varrho}+g_{\varrho\bar{\varrho}}\varphi^2_{\varrho},$$

$$C\varphi^4_{\varrho\bar{\varrho}}=\tfrac{1}{2}[-\mathbf{9}+\varrho\mathbf{9}+\bar{\varrho}\mathbf{9}-\varrho\bar{\varrho}\mathbf{9}]=-(1-\varrho-\bar{\varrho}+\varrho\bar{\varrho})\varphi^3_{\varrho\bar{\varrho}},$$

where

$$f_{\varrho\bar{\varrho}}=\frac{1}{\sqrt{2}}(1-\varrho-\bar{\varrho}-\varrho\bar{\varrho}), \tag{32a}$$

$$g_{\varrho\bar{\varrho}}=-\frac{1}{\sqrt{2}}(1+\varrho-\bar{\varrho}+\varrho\bar{\varrho}).$$

It is easy to show that

$$f_{\varrho\bar{\varrho}}=\sqrt{2}(1-\varrho)\delta_{\varrho\bar{\varrho}},\quad g_{\varrho\bar{\varrho}}=-\sqrt{2}(1+\varrho)\delta_{\varrho\bar{\varrho}^*}. \tag{32b}$$

From (31) and (32) we obtain the matrix M.

$$M=\sqrt{2}\begin{pmatrix} 0 & 0 & (1-\varrho)\delta_{\varrho\bar{\varrho}} & 0 \\ 0 & 0 & -(1+\varrho)\delta_{\varrho\bar{\varrho}^*} & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}}(1-\varrho-\bar{\varrho}+\varrho\bar{\varrho}) \\ (1-\varrho)\delta_{\varrho\bar{\varrho}} & (1-\varrho)\delta_{\varrho\bar{\varrho}^*} & 0 & 0 \end{pmatrix}. \tag{33}$$

(There are three misprints in the eigenvalue equations in Ref. 11: the $\tilde{M}$ in Eqs. (9g), (24), and (27) should be replaced by M.) The eigenvectors of the matrix in (33) are divided into two cases:

Case (i): $\bar{\varrho}=\varrho$.

In this case the matrix M becomes three-dimensional. In the basis $\{\varphi^1_\varrho,\varphi^3_{\varrho\varrho},\varphi^4_{\varrho\varrho}\}$, the rep matrix of C is

$$M=\begin{pmatrix} 0 & \sqrt{2}(1-\varrho) & 0 \\ 0 & 0 & 2\varrho \\ \sqrt{2}(1-\varrho) & 0 & 0 \end{pmatrix}. \tag{34}$$

Its eigenvalue is found as $\lambda=2,2\epsilon,2\epsilon^*\{\epsilon=\exp(-[2\pi i/3])\}$, and its eigenvectors are

$$\mathcal{P}^{(\lambda)\varrho}_\varrho=\frac{1}{\sqrt{3}}\left[\varphi^1_\varrho+\frac{1+\varrho}{\sqrt{2}}\left(\frac{\lambda}{2}\right)\varphi^3_{\varrho\varrho}+\frac{1-\varrho}{\sqrt{2}}\left(\frac{\lambda}{2}\right)^*\varphi^4_{\varrho\varrho}\right],\quad \varrho=\pm i, \tag{35a}$$

where we used the fact that $(\lambda/2)^2=\lambda^*/2$.

Case (ii): $\bar{\varrho}=\varrho^*$.

In the basis $\{\varphi^2_\varrho,\varphi^3_{\varrho,\varrho^*},\varphi^4_{\varrho\varrho^*}\}$ the rep matrix of C is

$$M=\begin{pmatrix} 0 & -\sqrt{2}(1+\varrho) & 0 \\ 0 & 0 & -2 \\ \sqrt{2}(1-\varrho) & 0 & 0 \end{pmatrix}. \tag{35b}$$

Its eigenvalue $\lambda=2,2\epsilon,2\epsilon^*$ with the eigenvectors

$$\mathcal{P}^{(\lambda)-\varrho}_\varrho=i^{1+\mu}\frac{1}{\sqrt{3}}\left[\varphi^2_\varrho-\frac{1-\varrho}{\sqrt{2}}\left(\frac{\lambda}{2}\right)\varphi^3_{\varrho\varrho^*}+\frac{1-\varrho}{\sqrt{2}}\left(\frac{\lambda}{2}\right)^*\varphi^4_{\varrho\varrho^*}\right],\quad \varrho=\pm i, \tag{35c}$$

where $\varrho=e^{-\pi\mu i},\mu=\pm\frac{1}{2}$. The phase factor $\sqrt{i}$ for $\mu=-\frac{1}{2}$ is chosen so that the SAF's will be real, while the phase factor $(\sqrt{i})^3$ for $\mu=\frac{1}{2}$ is determined from $\mathcal{P}^{(\lambda)\varrho'}_\varrho=(\mathcal{P}^{(\lambda)\varrho}_{\varrho'})^\dagger$.

B. The $T\supset C_2$ d.v. ISO's

In the following it is convenient to change the notation from $\mathcal{P}^{(\lambda)\bar{\varrho}}_\varrho$ to $\mathcal{P}^{(\lambda)\bar{\mu}}_\mu$ with $\varrho=e^{-\pi\mu i},\bar{\varrho}=e^{-\pi\bar{\mu}i}$. Using (35) and (29) we have the $T\supset C_2$ ISO's

$$\begin{aligned}
\mathcal{P}^{(\lambda)\mu}_\mu&=\frac{1}{\sqrt{6}}\left[P^\mu+\frac{1}{2\sqrt{2}}P^\mu(i^{-\mu}\lambda C^-_{31}+i^\mu\lambda^* C^+_{31})P^\mu\right],\\
\mathcal{P}^{(\lambda)\mu}_{-\mu}&=\frac{1}{\sqrt{6}}\left[i^{1-\mu}P^{-\mu}C_{2x}+\frac{(-1)^{1/2+\mu}}{2\sqrt{2}}P^{-\mu}(\lambda C^-_{31}-\lambda^* C^+_{31})P^\mu\right],
\end{aligned}\qquad \mu=\pm\tfrac{1}{2} \tag{36}$$

Choosing $\mu=\frac{1}{2}$, and ignoring the normalization, we obtain the $T\supset C_2$ d.v. generalized projection operators

$$\begin{aligned}
\mathbf{P}^{(\lambda)}_{1/2}&=P^{1/2}+\frac{\sqrt{2}}{4}P^{1/2}(\sqrt{i^*}\lambda C^-_{31}+\sqrt{i}\lambda^* C^+_{31})P^{1/2},\\
\mathbf{P}^{(\lambda)}_{-1/2}&=\sqrt{i}P^{-1/2}C_{2x}-\frac{\sqrt{2}}{4}P^{-1/2}(\lambda C^-_{31}-\lambda^* C^+_{31})P^{1/2}.
\end{aligned} \tag{37}$$

Recall that $\lambda=2,2\epsilon,2\epsilon^*$ for the irreps $E_{1/2},E_{3/2},E_{-3/2}$, respectively.

The simplicity of the algebraic expression for the $T\supset C_2$ ISO's $\mathcal{P}_{\mu}^{(\lambda)\bar{\mu}}$ is to be noted. It is a function of only three quantum numbers $\lambda,\mu,\bar{\mu}$, and contains only three terms, while in the original form (6), the ISO's of T for the three d.v. two-dimensional irreps involve altogether $3\times 2^2\times 12=144$ irreducible matrix elements, and each ISO has 12 terms.

C. $T\supset C_2$ irreducible matrices

The matrices of C_{2z} are diagonal, while the representation matrices of the coset generators $\hat{k}_i=C_{2x},C_{31}^{+},C_{31}^{-}$ are obtained in the following way. According to Sec. IV C in Ref. 11, we know that the coefficients of $P^{\varrho}\hat{k}_iP^{\bar{\varrho}}$ in (36) are equal to

$$\sqrt{h_\lambda/12}D^{\Gamma}_{\varrho\bar{\varrho}}(\hat{k}_i)^*.$$

Consequently from (36) we immediately get the algebraic expression for the $T\supset C_2$ irreducible matrices in the order of $\mu=\frac{1}{2},-\frac{1}{2}$:

$$\begin{aligned} D^{(\lambda)}(C_{2z})&=\begin{pmatrix} -i & 0\\ 0 & i\end{pmatrix},\quad D^{(\lambda)}(C_{31}^{+})=\frac{\lambda}{2\sqrt{2}}\begin{pmatrix}\sqrt{i^*}, & -1\\ 1, & \sqrt{i}\end{pmatrix}.\\ D^{(\lambda)}(C_{2x})&=\begin{pmatrix} 0 & i\sqrt{i}\\ \sqrt{i} & 0\end{pmatrix}^*,\quad D^{(\lambda)}(C_{31}^{-})=\frac{\lambda^*}{2\sqrt{2}}\begin{pmatrix}\sqrt{i}, & 1\\ -1, & \sqrt{i^*}\end{pmatrix}.\end{aligned}\tag{38}$$

The matrices for other elements are obtained through

$$D^{(\lambda)}_{\varrho,\bar{\varrho}}(C_{2z}^{j}C_{2x})=\varrho^{j}D^{\lambda}_{\varrho,\bar{\varrho}}(C_{2x}),D^{(\lambda)}_{\varrho\bar{\varrho}}(C_{2z}^{j}C_{31}^{\pm}C_{2z}^{k})=\varrho^{j}D^{(\lambda)}_{\varrho\bar{\varrho}}(C_{31}^{\pm})\bar{\varrho}^{k}.$$

The elements of $C_{2z}^{j}C_{2x}$ and $C_{2z}^{j}C_{31}^{\pm}C_{2z}^{k}$ are listed below:

j	0	1		
$C_{2z}^{j}C_{2x}$	C_{2x}	C_{2y}		
(j,k)	(00)	(01)	(10)	(11)
$C_{2z}^{j}C_{31}^{+}C_{2z}^{k}$	C_{31}^{+}	$\tilde{C}_{34}^{+}$	$\tilde{C}_{33}^{+}$	$\tilde{C}_{32}^{+}$
$C_{2z}^{j}C_{31}^{-}C_{2z}^{k}$	C_{31}^{-}	C_{33}^{-}	C_{34}^{-}	$\tilde{C}_{32}^{-}$

D. The $T\supset C_2$ d.v. SAF's

In order to obtain real SAF's, we choose the coordinate system $x''y''z$ as shown in Fig. 2, which is obtained from rotating the coordinate system xyz in Fig. 1 through angle $-\pi/4$ about the z axis. Using (11) of Ref. 11 and the Euler angles in the system xyz shown in Sec. III C, we obtain the Euler angles for C_{2x} and $C_{31}^{\pm}$ in the $x''y''z$ system

$$C_{2x}=R\left(0,\pi,\frac{\pi}{2}\right),\quad C_{31}^{+}=R\left(\frac{\pi}{4},\frac{\pi}{2},\frac{\pi}{4}\right),\quad C_{31}^{-}=R\left(\frac{3\pi}{4},\frac{\pi}{2},\frac{11\pi}{4}\right).$$

It should be mentioned that if for the group chain $T\supset D_2$ we also choose the coordinate system $x''y''z$, then the $T\supset D_2$ SAF's $\varphi^{(A_\mu)}$ will become real, however, the SAF's $\varphi^{(F)}$ will become complex.

Applying (37) to $|j\bar{m}\rangle$ and using $P^{\mu}|j\bar{m}\rangle=2\delta_{\mu,\bar{m}(\mathrm{mod}\,2)}|j\bar{m}\rangle$, we obtain the $T\supset C_2$ d.v. SAF

$$\varphi_{1/2}^{(E_\mu)\bar{m}}=|j\bar{m}\rangle+\sqrt{2}\sum_{m=1/2}^{(2)}\mathrm{Re}(\lambda\sqrt{i}^{m+\bar{m}-1})d^{j}_{m\bar{m}}\left(\frac{\pi}{2}\right)|jm\rangle,\tag{39}$$

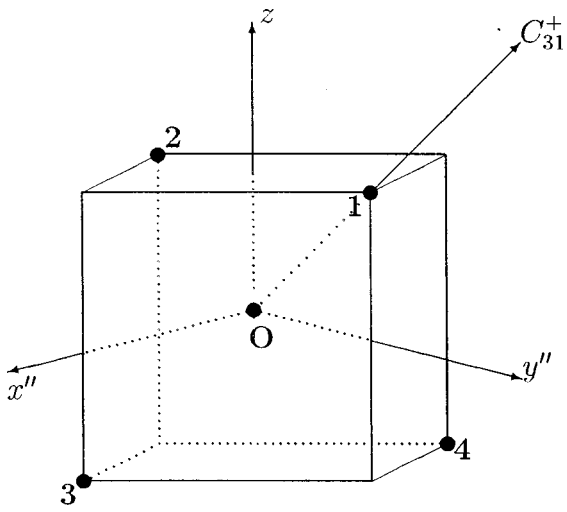

FIG. 2. The coordinate system $x''y''z$ for $0 \supset T \supset C_2$, $xyz \xrightarrow{R(\omega)} x''y''z$, $R(\omega) \equiv R(0,0,-\pi/4)$.

$$\varphi_{-1/2}^{(E_\mu)\bar{m}} = (-1)^{j-1/2+\kappa_{\bar{m}}} |j-\bar{m}\rangle + \sqrt{2} \sum_{m=-1/2}^{(2)} \mathrm{Re}(\lambda \sqrt{i}^{m+\bar{m}}) d^j_{m\bar{m}}\left(\frac{\pi}{2}\right) |jm\rangle,$$

where $\bar{m} = \frac{1}{2} (\mathrm{mod}\, 2)$, and $\lambda = 2, 2\epsilon, 2\epsilon^*$ ($\epsilon = e^{-(2\pi/3)i}$), for $\mu = \frac{1}{2}, \frac{3}{2}, -\frac{3}{2}$, respectively, $\kappa_{\bar{m}}$ is an integer

$$\kappa_{\bar{m}} = \tfrac{1}{2}(\bar{m} - \tfrac{1}{2}).$$

The summation symbol in (39) stands for

$$\sum_{m=\mu}^{(2)} = \sum_{m=\mu,\mu+2,\mu-2,\cdots}.$$

Since for $E_{1/2}$ the eigenvalue $\lambda = 2$, and the condition that $\lambda \sqrt{i}^{m+\bar{m}-1}$ be real imposes an extra condition for the summation index m in (39): $m = \bar{m} \pmod 4$ for $\varphi_{1/2}^{(E_{1/2})\bar{m}}$ and similarly $m = -\bar{m} \pmod 4$ for $\varphi_{-1/2}^{(E_{1/2})\bar{m}}$.

Notice that all the coefficients in (39) are real and they have the following symmetries:

$$\begin{aligned} \langle \varphi_{1/2}^{(E_\nu)\bar{m}} | jm \rangle &= (-1)^{j-1/2+\kappa_m} \langle \varphi_{-1/2}^{(E_\nu)\bar{m}} | j-m \rangle, \quad \nu = \tfrac{1}{2}, \pm\tfrac{3}{2}, \\ \langle \varphi_{\mu}^{(E_{3/2})\bar{m}} | jm \rangle &= (-1)^{\kappa_m + \kappa_{\bar{m}}} \langle \varphi_{\mu}^{(E_{-3/2})\bar{m}} | jm \rangle, \quad \mu = \pm\tfrac{1}{2}. \end{aligned} \tag{40}$$

Under the time reversal $\hat{T} = -i\sigma_y K$ the SAF's transform as

$$\begin{aligned} \hat{T} \varphi_{\mu}^{(E_{1/2})\bar{m}} &= (-1)^{1/2-\mu} (-1)^{\kappa_{\bar{m}}} \varphi_{-\mu}^{(E_{1/2})\bar{m}}, \quad \mu = \pm\tfrac{1}{2} \\ \hat{T} \varphi_{\mu}^{(E_\nu)\bar{m}} &= (-1)^{1/2-\mu} (-1)^{\kappa_{\bar{m}}} \varphi_{-\mu}^{(E_{-\nu})-\bar{m}}, \quad \nu = \pm\tfrac{3}{2}, \quad \mu = \pm\tfrac{1}{2}. \end{aligned} \tag{41}$$

Again the quantum number $\bar{m}$ in (39) serves as the multiplicity label. The exact numerical expression of the SAF's for $j = \frac{1}{2}$-$\frac{9}{2}$ are shown in Table III. Notice that the listed SAF's with different $\bar{m}$ are linearly independent but in general not orthogonal. The high symmetries (40) of the coefficients in Table III are to be noted. The $T \supset C_2$ d.v. SAF's for the lowest two angular momenta $j = \frac{1}{2}, \frac{3}{2}$ have been given by Damhus.[17] No other systematic tables for the $T \supset C_2$ SAF's are available. Even if such tables do exist, the symmetries of the SAF's would be buried in the vast numerical data and cannot be easily discovered.

TABLE III. The $T\supset C_2$ d.v. SAF's for $j=\frac{1}{2}$-$\frac{9}{2}$. For $j=\frac{9}{2}$ only the SAF's with multiplicity are listed.

$\bar{m}=\frac{1}{2}$

$$|E_{1/2}\tfrac{1}{2}\rangle = |\tfrac{1}{2}\tfrac{1}{2}\rangle$$
$$|E_{1/2}-\tfrac{1}{2}\rangle = |\tfrac{1}{2}-\tfrac{1}{2}\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{3/2}\frac{1}{2}\right\rangle = -\frac{\sqrt{2}}{2}\left|\frac{3}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{3}{2}\frac{1}{2}\right\rangle$$
$$\left|E_{3/2}-\frac{1}{2}\right\rangle = -\frac{\sqrt{2}}{2}\left|\frac{3}{2}-\frac{1}{2}\right\rangle-\frac{\sqrt{2}}{2}\left|\frac{3}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{-3/2}\frac{1}{2}\right\rangle = \frac{\sqrt{2}}{2}\left|\frac{5}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{5}{2}\frac{1}{2}\right\rangle$$
$$\left|E_{-3/2}-\frac{1}{2}\right\rangle = -\frac{\sqrt{2}}{2}\left|\frac{3}{2}-\frac{1}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{3}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{5}{2}$

$$\left|E_{1/2}\frac{1}{2}\right\rangle = -\frac{\sqrt{2\cdot3\cdot5}}{6}\left|\frac{5}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2\cdot3}}{6}\left|\frac{5}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{1/2}-\frac{1}{2}\right\rangle = -\frac{\sqrt{2\cdot3}}{6}\left|\frac{5}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{6}\left|\frac{5}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{-3/2}\frac{1}{2}\right\rangle = \frac{\sqrt{3}}{6}\left|\frac{5}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{5}{2}\frac{1}{2}\right\rangle-\frac{\sqrt{3\cdot5}}{6}\left|\frac{5}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{-3/2}-\frac{1}{2}\right\rangle = \frac{\sqrt{3\cdot5}}{6}\left|\frac{5}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{5}{2}-\frac{1}{2}\right\rangle-\frac{\sqrt{3}}{6}\left|\frac{5}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{3/2}\frac{1}{2}\right\rangle = -\frac{\sqrt{3}}{6}\left|\frac{5}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{5}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{3\cdot5}}{6}\left|\frac{5}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{3/2}-\frac{1}{2}\right\rangle = -\frac{\sqrt{3\cdot5}}{6}\left|\frac{5}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{5}{2}-\frac{1}{2}\right\rangle+\frac{\sqrt{3}}{6}\left|\frac{5}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{1/2}\frac{1}{2}\right\rangle^{1} = -\frac{\sqrt{3\cdot5}}{6}\left|\frac{7}{2}-\frac{7}{2}\right\rangle+\frac{\sqrt{3\cdot7}}{6}\left|\frac{7}{2}\frac{1}{2}\right\rangle$$
$$\left|E_{1/2}-\frac{1}{2}\right\rangle^{1} = -\frac{\sqrt{3\cdot7}}{6}\left|\frac{7}{2}-\frac{1}{2}\right\rangle+\frac{\sqrt{3\cdot5}}{6}\left|\frac{7}{2}\frac{7}{2}\right\rangle$$

$\bar{m}=\frac{5}{2}$

$$\left|E_{1/2}\frac{1}{2}\right\rangle^{2} = \frac{1}{2}\left|\frac{7}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{3}}{2}\left|\frac{7}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{1/2}-\frac{1}{2}\right\rangle^{2} = \frac{\sqrt{3}}{2}\left|\frac{7}{2}-\frac{5}{2}\right\rangle+\frac{1}{2}\left|\frac{7}{2}\frac{3}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{3/2}\frac{1}{2}\right\rangle = \frac{\sqrt{2\cdot3\cdot7}}{12}\left|\frac{7}{2}-\frac{7}{2}\right\rangle+\frac{\sqrt{2\cdot3}}{4}\left|\frac{7}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{7}{2}\frac{1}{2}\right\rangle-\frac{\sqrt{2}}{4}\left|\frac{7}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{3/2}-\frac{1}{2}\right\rangle = -\frac{\sqrt{2}}{4}\left|\frac{7}{2}-\frac{5}{2}\right\rangle-\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{7}{2}-\frac{1}{2}\right\rangle+\frac{\sqrt{2\cdot3}}{4}\left|\frac{7}{2}\frac{3}{2}\right\rangle-\frac{\sqrt{2\cdot3\cdot7}}{12}\left|\frac{7}{2}\frac{7}{2}\right\rangle$$

$\bar{m}=\frac{1}{2}$

$$\left|E_{-3/2}\frac{1}{2}\right\rangle = \frac{\sqrt{2\cdot3\cdot7}}{12}\left|\frac{7}{2}-\frac{7}{2}\right\rangle-\frac{\sqrt{2\cdot3}}{4}\left|\frac{7}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{7}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{2}}{4}\left|\frac{7}{2}\frac{5}{2}\right\rangle$$
$$\left|E_{-3/2}-\frac{1}{2}\right\rangle = \frac{\sqrt{2}}{4}\left|\frac{7}{2}-\frac{5}{2}\right\rangle-\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{7}{2}-\frac{1}{2}\right\rangle-\frac{\sqrt{2\cdot3}}{4}\left|\frac{7}{2}\frac{3}{2}\right\rangle-\frac{\sqrt{2\cdot3\cdot7}}{12}\left|\frac{7}{2}\frac{7}{2}\right\rangle$$

V. SUMMARY

A factorization lemma for the direct-product group of two abelian groups is discovered, with which the ISO's for the group T can be written down without any derivation.

The algebraic expressions of the ISO's and SAF's of the group T are given in (15) and (18) for the group chain $T\supset D_2$, and in (36) and (39) for the group chain $T\supset C_2$.

The simplicity of the expressions (36) for the $T\supset C_2$ ISO's $\mathcal{P}_{\mu}^{(\lambda)\bar{\mu}}$ is remarkable that they are functions of the quantum numbers λ,μ, and $\bar{\mu}$ only without involving any irreducible matrix elements, and they are simpler than the algebraic expressions for the $T\supset C_3$ ISO's given in (35) of Ref. 11. The simplicity can be appreciated more clearly if they are compared with the usual ''numerical'' expressions of the ISO's involving all the $36(=3\times12)$ two by two matrices.

The SAF's $\psi_{\underline{\mu}}^{(\lambda)\bar{m}}$ are functions of the quantum numbers j,λ,μ, and $\bar{m}$ only. The intrinsic quantum number $m$plays an important role as the multiplicity label for the SAF's, and as an index for specifying the transformation property of the SAF's under the time reversal, despite the fact

TABLE III. (*Continued*).

$\tilde{m}=\frac{1}{2}$

$$|E_{3/2\,\frac{1}{2}}\rangle^1 = \frac{\sqrt{3\cdot5\cdot7}}{60}\left|\frac{9}{2}-\frac{7}{2}\right\rangle+\frac{\sqrt{3\cdot5}}{10}\left|\frac{9}{2}-\frac{3}{2}\right\rangle-\frac{\sqrt{5\cdot7}}{10}\left|\frac{9}{2}\frac{5}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{9}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{3\cdot5\cdot7}}{20}\left|\frac{9}{2}\frac{9}{2}\right\rangle$$

$$|E_{3/2-\frac{1}{2}}\rangle^1 = \frac{\sqrt{3\cdot5\cdot7}}{20}\left|\frac{9}{2}-\frac{9}{2}\right\rangle+\frac{\sqrt{5\cdot7}}{10}\left|\frac{9}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{9}{2}-\frac{1}{2}\right\rangle-\frac{\sqrt{3\cdot5}}{10}\left|\frac{9}{2}\frac{3}{2}\right\rangle+\frac{\sqrt{3\cdot5\cdot7}}{60}\left|\frac{9}{2}\frac{7}{2}\right\rangle$$

$\tilde{m}=\frac{5}{2}$

$$|E_{3/2\,\frac{1}{2}}\rangle^2 = -\frac{5\sqrt{2\cdot3}}{24}\left|\frac{9}{2}-\frac{7}{2}\right\rangle-\frac{\sqrt{3\cdot7}}{12}\left|\frac{9}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{9}{2}\frac{5}{2}\right\rangle-\frac{\sqrt{2\cdot3}}{8}\left|\frac{9}{2}\frac{9}{2}\right\rangle$$

$$|E_{3/2-\frac{1}{2}}\rangle^2 = -\frac{\sqrt{2\cdot3}}{8}\left|\frac{9}{2}-\frac{9}{2}\right\rangle-\frac{\sqrt{2}}{2}\left|\frac{9}{2}-\frac{5}{2}\right\rangle-\frac{\sqrt{3\cdot7}}{12}\left|\frac{9}{2}-\frac{1}{2}\right\rangle-\frac{5\sqrt{2\cdot3}}{24}\left|\frac{9}{2}\frac{7}{2}\right\rangle$$

$\tilde{m}=\frac{1}{2}$

$$|E_{-3/2\,\frac{1}{2}}\rangle^1 = \frac{\sqrt{3\cdot5\cdot7}}{60}\left|\frac{9}{2}-\frac{7}{2}\right\rangle-\frac{\sqrt{3\cdot5}}{10}\left|\frac{9}{2}-\frac{3}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{9}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{5\cdot7}}{10}\left|\frac{9}{2}\frac{5}{2}\right\rangle+\frac{\sqrt{3\cdot5\cdot7}}{20}\left|\frac{9}{2}\frac{9}{2}\right\rangle$$

$$|E_{-3/2-\frac{1}{2}}\rangle^1 = \frac{\sqrt{3\cdot5\cdot7}}{20}\left|\frac{9}{2}-\frac{9}{2}\right\rangle-\frac{\sqrt{5\cdot7}}{10}\left|\frac{9}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{2\cdot3\cdot5}}{12}\left|\frac{9}{2}-\frac{1}{2}\right\rangle+\frac{\sqrt{3\cdot5}}{10}\left|\frac{9}{2}\frac{3}{2}\right\rangle+\frac{\sqrt{3\cdot5\cdot7}}{60}\left|\frac{9}{2}\frac{7}{2}\right\rangle$$

$\tilde{m}=\frac{5}{2}$

$$|E_{-3/2\,\frac{1}{2}}\rangle^1 = \frac{5\sqrt{2\cdot3}}{24}\left|\frac{9}{2}-\frac{7}{2}\right\rangle+\frac{\sqrt{3\cdot7}}{12}\left|\frac{9}{2}\frac{1}{2}\right\rangle+\frac{\sqrt{2}}{2}\left|\frac{9}{2}\frac{5}{2}\right\rangle+\frac{\sqrt{2\cdot3}}{8}\left|\frac{9}{2}\frac{9}{2}\right\rangle$$

$$|E_{-3/2-\frac{1}{2}}\rangle^2 = \frac{\sqrt{2\cdot3}}{8}\left|\frac{9}{2}-\frac{9}{2}\right\rangle-\frac{\sqrt{2}}{2}\left|\frac{9}{2}-\frac{5}{2}\right\rangle+\frac{\sqrt{3\cdot7}}{12}\left|\frac{9}{2}-\frac{1}{2}\right\rangle+\frac{5\sqrt{2\cdot3}}{24}\left|\frac{9}{2}\frac{7}{2}\right\rangle$$

that the SAF's are in general not orthogonal in $\bar{m}$. A similar situation occurs for the rotation group, the irreducible basis vectors of which are in general not orthogonal in the intrinsic quantum number K, the third component of the total angular momentum in the intrinsic (body-fixed) coordinate system.[18]

From the results of our previous paper[11] and this paper it is seen that the elegance of the algebraic solutions of the point group T in both the group chains $T\supset C_3$ and $T\supset D_2\supset C_2$ can be compared with that of the analytic solutions of the rotation group. An application of the double-induced technique to all other point groups is straightforward and will be reported elsewhere.

Finally it should be mentioned that an extremely simple relation between the SAF's of the octahedral group O and tetrahedral group T has been found,[19] namely the $T\supset G(s)$ SAF's are just the $O\supset T\supset G(s)$ SAF's, where $G(s)$ can be any subgroup of T, e.g., $G(s)=C_3,C_2,D_2$. In other words, the SAF's found in Ref. 11 and this paper are also the SAF's of the group O. The SAF's of the groups O and T listed in pp. 66–70 of Ref. 4 offered a verification of the above conclusion.

ACKNOWLEDGMENTS

This work was supported in part by the Fundamental Research Fund of State Science and Technology Committee of China and the Research Fund of the National Educational Committee of China.

[1] F. C. von der Lage and H. A. Bethe, Phys. Rev. **71**, 612 (1947).
[2] S. L. Altmann, Proc. Cambridge Philos. Soc. **53**, 343 (1957); S. L. Altmann and A. P. Cracknell, Rev. Mod. Phys. **37**, 19 (1965); S. L. Altmann and C. J. Bradley, *ibid.* **37**, 33 (1965).
[3] A. P. Cracknell, Proc. Cambridge Philos. Soc. **65**, 567 (1969); A. P. Cracknell and S. J. Joshua, Proc. Cambridge Philos. Soc. **67**, 647 (1970).
[4] C. J. Bradley and A. P. Cracknell, *Symmetry in Solids* (Oxford University Press, London, 1972).
[5] P. H. Butler, *Point Group Symmetry Applications: Methods and Table* (Plenum, New York, 1981).
[6] G. W. Fernando, M. Weinert, R. E. Watson, and J. W. Davenport, J. Comput. Phys. **112**, 282 (1994).

[7] G. F. Koster, J. O. Dimmock, R. G. Wheeler, and H. Statz, *Properties of the thirty-two point groups* (M.I.T., Cambridge, MA, 1963).
[8] J. Q. Chen, F. Wang, and M. J. Gao, Acta Phys. Sin. **26**, 307 (1997); Chin. Phys. **1**, 533 (1981) (English translation).
[9] J. Q. Chen, *Group Representation Theory for Physicists* (World Scientific, Singapore, 1989); J. Q. Chen, M. J. Gao, and G. Q. Ma, Rev. Mod. Phys. **57**, 211 (1985).
[10] J. L. Ping, Q. R. Zheng, B. Q. Chen, and J. Q. Chen, Comput. Phys. Commun. **52**, 355 (1989).
[11] J. Q. Chen and P. D. Fan, J. Math. Phys. **39**, 467 (1998).
[12] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Claredon, Oxford, 1958), Sec 19.
[13] S. L. Altmann, Mol. Phys. **38**, 489 (1979).
[14] P. Herzig, Int. J. Quantum Chem. **26**, 297 (1984).
[15] N. O. Folland, J. Math. Phys. **18**, 31 (1977).
[16] M. E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[17] T. Damhus, S. E. Harnung, and C. E. Schaffer, Theor. Chim. Acta (Berl) **65**, 405 (1984).
[18] J. P. Elliott, Proc. R. Soc. London, Ser. A **245**, 562 (1958).
[19] J. Q. Chen and P. D. Fan, Algebraic solutions for point groups: The octahedral group for the group chain $O \supset T \supset C_3$ (preprint).

Algebraic expressions of the Clebsch-Gordon coefficients of the point group $T^\dagger$

Jin-Quan Chen[a] and Peng-Dong Fan
Department of Physics, Nanjing University, Nanjing, 210008, People's Republic of China

(Received 24 March 1998; accepted for publication 5 May 1998)

A method for finding algebraic expressions of the Clebsch-Gordan (CG) coefficients of point groups is proposed and applied to the tetrahedral group. It is shown that in constructing the CG coefficients the irreducible symmetry operator (ISO) of a double group $G^\dagger$ can be replaced by an effective ISO which is much simpler than the usual ISO. The effective ISO for the group chain $T^\dagger \supset C_3^\dagger$ is $\mathscr{P}^{(\lambda)}_{\mu,\mu} = \delta_{\mu\mu} + 3d^{(\lambda)}_{\mu\mu}(C_{2z})^* C_{2z}$, where $d^{(\lambda)}(C_{2z})$ is the matrix of C_{2z} in the irrep λ of $T^\dagger$. With this effective ISO and the algebraic expression of $d^{(\lambda)}(C_{2z})$, the algebraic expressions are derived for the real CG coefficients of $T^\dagger$ in the group chain $T^\dagger \supset C_3^\dagger$. The algebraic expressions for the complex (real) CG coefficients of the group chain $T^\dagger \supset D_2^\dagger \supset C_2^\dagger$ ($T^\dagger \supset C_2^\dagger$) have also been obtained. © *1998 American Institute of Physics.* [S0022-2488(98)04110-3]

I. INTRODUCTION

The irreducible representations of point groups are well known, along with their Clebsch-Gordan (CG) coefficients, subduction coefficients, etc. However, all the solutions to the point groups are numerical.[1–11] Recently, by using the double-induced technique, the algebraic expressions of the irreducible symmetry operators (ISO's), irreducible matrices, and symmetry adapted functions (SAF's) are found for the double group $T^\dagger$ in the group chain $T \supset C_3$ and $T \supset D_2 \supset C_2$.[12,13] What makes the algebraic solution appealing is the fact that the irreducible matrices, the ISO's and SAF's are merely functions of the quantum numbers (λ,μ) characterizing the group chain [the analogy of (j,m) for $SO_3 \supset SO_2$], and these expressions are valid for both single-valued and double-valued reps. The remaining task in seeking the algebraic solutions of the point group T is to get the algebraic expressions of its Clebsch-Gordan (CG) coefficients, which is the main purpose of the present paper. In Refs. 12 and 13 it has been shown that with the available algebraic expression of the ISO's, the projection operator method for constructing the SAF's becomes extremely simple. In this paper we will show that the same applies to the construction of the CG coefficients.

Since in this paper we always deal with both the single-valued (s.v.) and double-valued (d.v.) reps simultaneously, the symbol for the double group $G^\dagger$ will be replaced simply by G.

The numerical tables of the $T \supset D_2 \supset C_2$ CG coefficients have been first calculated in Ref. 3, the 3-Γ (or $3jm$) symbols for the group chain $T \supset C_3$ have been tabulated in Refs. 14–16 and for the group chain $T \supset C_2$ in Ref. 16. The CG coefficients for the group chains $T \supset D_2 \supset C_2$ and $T \supset C_3$ can also be easily obtained from the code PGIR.[17]

Various methods were used for computing the CG coefficients numerically, which include:

(1) The projection operator method[18] and its variations (see Sec. VII.E in Ref. 19).

(2) The Griffith method[2] based on the transformation properties of the direct product under the generators of the group.

(3) The Golding method.[15] First calculate the subduction coefficients $\langle jm|\Gamma\mu\rangle$ for a set of minimum j values, and then obtain the CG coefficients of G from those of $SO_3 \supset SO_2$. If there is a multiplicity in $\Gamma_1 \times \Gamma_2 \to \Gamma$, then use the calculated CG coefficients to find the subduction coefficients $\langle j'm'|\Gamma\mu\rangle$ for a higher j' value, and repeat the above process for obtaining another set of CG coefficients.

[a]Electronic mail: jqchen@chenwang.nju.edu.cn

 © 1998 American Institute of Physics

(4) The Butler-Wybourne method,[7] or the building up method which uses only character theory results.

(5) The eigenfunction method[8,19] by which the CG coefficients are obtained by diagonalizing a set of commuting operators, the analogy of (J^2,J_z), in the Kronecker product representation of G.

Though straightforward in principle, the projection operator method in its original form is quite tedious and is avoided by many people for constructing the CG coefficients of the high-order point groups O and I.[20,21] It will be shown in this paper that the projection operator for constructing the $G\supset C_n$ CG coefficients, where C_n is the maximal cyclic group of G, can be replaced by an effective projection operator, which can be written, up to a multiplicative factor, as

$$\mathscr{P}_{\mu}^{(\Gamma)\bar{\mu}}=\sum_{i=1}^{q} M_i^{-1} d_{\mu\mu}^{(\Gamma)}(\beta_i)^{*}\beta_i, \tag{1}$$

where M_i is the number of times an element of G appears in the double coset $C_n\beta_i C_n$, q is the number of double cosets, β_i is the double coset generator, and $d_{\mu\mu}^{(\Gamma)}(\beta_i)$ is the irreducible matrix elements. Therefore the effective projection operator of a point group of any order is almost as simple as the projection operator of a group of order q, which is very small, $q=2$, 2, 3, 4 for the groups D_n,T,O,I, respectively. With the effective projection operator, the algebraic expressions for the CG coefficients of point groups can be derived. As examples, the method is used for obtaining the algebraic expressions of the $T\supset C_3$ CG coefficients. With slight modifications, the method is also used for finding the algebraic expressions of the $T\supset D_2\supset C_2$ and $T\supset C_2$ CG coefficients. Using these algebraic expressions numerical tables of the CG coefficients are easily generated and tabulated.

II. THE EFFECTIVE ISO'S OF $G\supset C_n$

Before discussing the group chain $T\supset C_3$, let us consider a more general group chain $G\supset C_n$. The maximal cyclic groups C_n of G are

$$\begin{aligned}&\text{Dihedral groups} && D_n\supset C_n,\\ &\text{Tetrahedral group} && T\supset C_3,\\ &\text{Octahedral group} && O\supset C_4,\\ &\text{Icosahedral group} && I\supset C_5.\end{aligned} \tag{2}$$

Since the number of double cosets, q, is the smallest for the maximal cyclic group, the group chains (2) are the simplest ones, and will be referred to as the natural group chains. The axis of C_n is chosen as the z axis, and the generator of C_n is

$$C_{nz}=\exp\left(-i\frac{2\pi}{n}J_z\right), \tag{3a}$$

and its eigenvalue is denoted by

$$\rho\equiv\rho_{\mu}=\exp\left(-i\frac{2\mu\pi}{n}\right), \tag{3b}$$

with the values of μ given below:

n	2	3	4	5
μ	$0,1;\pm\frac{1}{2}$	$0,\pm1;\pm\frac{1}{2},\frac{3}{2}$	$0,\pm1,2;\pm\frac{1}{2},\pm\frac{3}{2}$	$0,\pm1,\pm2;\pm\frac{1}{2},\pm\frac{3}{2},\frac{5}{2}$

(3c)

where $\mu=$ integers (half integers) are for the s.v. (d.v.) reps.

The un-normalized projection operator of C_n is

$$P^{\rho}\equiv P^{\mu}=\sum_{s=0}^{n-1}\rho^{*s}(C_{nz})^{s}. \tag{3d}$$

The normalized $G\supset C_n$ ISO is defined as

$$\mathscr{P}_{\rho}^{(\lambda)\bar{\rho}}=\sqrt{\frac{h_\lambda}{|\mathbf{g}|}}\sum_{a=1}^{|\mathbf{g}|}D_{\rho\rho}^{(\lambda)}(R_a)^{*}R_a, \tag{4}$$

where h_λ is the dimension of the irrep λ, $|\mathbf{g}|$ the order of G, R_a the ath group element, and $D_{\rho\rho}^{(\lambda)}(R_a)$ is the irreducible matrix element of R_a,

$$D_{\rho\rho}^{(\lambda)}(R_a)=\langle\psi_{\rho}^{\lambda}|R_a|\psi_{\rho}^{\lambda}\rangle.$$

In the double-induced method,[12,13] the group G is decomposed into double cosets,

$$G=\sum_{i=1}^{q}\oplus C_n\beta_i C_n, \tag{5a}$$

where β_i are coset generators. Let $M(\beta_i)\equiv M_i$ be the number of times an element appears in the double coset $C_n\beta_i C_n$. For the natural group chains (2) M_i is either n or 1 and is decided as following. In general, the frequency M_i for an element in the double coset $C_n\beta_i C_n$ is equal to the intersection of the right coset $C_n\beta_i$ and the left coset $\beta_i C_n$, see, for example, Ref. 22. In the decomposition (5a) some double cosets, say the first p double cosets, are coincidental with the right or left cosets,

$$C_n\beta_i C_n=\beta_i C_n=C_n\beta_i, \quad i=1,2,\ldots,p, \tag{5b}$$

then obviously

$$M_i=\begin{cases} n, & i=1,\ldots,p \\ 1, & i=p+1,\ldots,q. \end{cases}$$

Any element of G can be expressed as

$$R_a=C_{nz}^{s}\beta_i C_{nz}^{t}, \quad s,t=0,1,\ldots,n-1. \tag{5c}$$

In the $G\supset C_n$ basis $\psi_{\rho}^{(\lambda)}$, the irreducible matrix elements can be expressed as

$$D_{\rho\rho}^{(\lambda)}(C_{nz}^{s}\beta_k C_{nz}^{t})=\rho^{s}d_{\rho\rho}^{(\lambda)}(\beta_k)\bar{\rho}^{t}. \tag{6}$$

Due to (3b) there is a one to one correspondence between ρ and μ, and since the quantum number μ is additive, it is more convenient to use μ instead of ρ as the component index for the irreps. Therefore in the following we switch to the μ notation.

From (5c) and (6) we know that the $G\supset C_n$ ISO in (4) can be factored as

$$\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}=\sqrt{\frac{h_\lambda}{|\mathbf{g}|}}\,\dot{\mathscr{P}}_{\mu}^{(\lambda)\bar{\mu}}, \tag{7a}$$

$$\dot{\mathscr{P}}_{\mu}^{(\lambda)\bar{\mu}}=\sum_{i=1}^{q}M_i^{-1}P^{\mu}d_{\mu\mu}^{(\lambda)}(\beta_i)^{*}\beta_i P^{\bar{\mu}}. \tag{7b}$$

The normalized ISO satisfies the relation

$$(\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}})^{\dagger}=\mathscr{P}_{\mu}^{(\lambda)\mu}, \tag{8a}$$

$$\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}\mathscr{P}_{\mu'}^{(\lambda')\bar{\mu}'}=\delta_{\lambda\lambda'}\delta_{\bar{\mu}\bar{\mu}'}\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}'}. \tag{8b}$$

Let us consider the Kronecker product $D^{(\lambda_1)}\times D^{(\lambda_2)}$. Applying the ISO's to a basis vector $|\bar{\mu}_1\bar{\mu}_2\rangle\equiv|\lambda_1\bar{\mu}_1\lambda_2\bar{\mu}_2\rangle$, referred to as the initial or trial state, in the Kronecker product space, and using (8) we get a normalized symmetry adapted state, referred to as the final state,

$$\Psi_{\mu}^{(\lambda)\bar{\mu}_1\bar{\mu}_2}=\dot{\mathscr{N}}\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}|\bar{\mu}_1\bar{\mu}_2\rangle,\quad \dot{\mathscr{N}}^{-1}=\langle\bar{\mu}_1\bar{\mu}_2|\dot{\mathscr{P}}_{\mu}^{(\lambda)\bar{\mu}}|\bar{\mu}_1\bar{\mu}_2\rangle^{1/2}. \tag{9a}$$

From it we obtain an expression for the CG coefficients first derived in Ref. 23,

$$C_{\lambda_1\mu_1,\lambda_2\mu_2}^{(\lambda)\bar{\mu},\mu}=\frac{\langle\mu_1\mu_2|\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}|\bar{\mu}_1\bar{\mu}_2\rangle}{\langle\bar{\mu}_1\bar{\mu}_2|\dot{\mathscr{P}}_{\mu}^{(\lambda)\bar{\mu}}|\bar{\mu}_1\bar{\mu}_2\rangle^{1/2}}=\sqrt{\frac{h_\lambda}{|\mathbf{g}|}}\,\dot{\mathscr{N}}\sum_{a=1}^{|\mathbf{g}|}D_{\mu_1\mu_1}^{(\lambda_1)}(R_a)D_{\mu_2\mu_2}^{(\lambda_2)}(R_a)D_{\mu\mu}^{(\lambda)}(R_a)^*. \tag{9b}$$

Using

$$\frac{1}{n}\sum_{k=0}^{n-1}(\rho_\mu^*\rho_m)^k=\dot{\delta}_{\mu,m},$$

where

$$\dot{\delta}_{\mu,m}=\begin{cases}1, & \mu=m(\mathrm{mod}\ n)\\ 0, & \text{otherwise}\end{cases} \tag{9c}$$

we have

$$P^{\mu}|\mu_1\mu_2\rangle=\dot{\delta}_{\mu,\mu_1+\mu_2}n|\mu_1\mu_2\rangle. \tag{9d}$$

Using (9a) and (9d) we get

$$\Psi_{\mu}^{(\lambda)\bar{\mu}_1\bar{\mu}_2}=\mathscr{N}\sum_{\mu_1\mu_2}{}'\sum_{i=1}^{q}M_i^{-1}d_{\mu\mu}^{(\lambda)}(\beta_i)^*d_{\bar{\mu}_1\bar{\mu}_1}^{(\lambda_1)}(\beta_i)d_{\bar{\mu}_2\bar{\mu}_2}^{(\lambda_2)}(\beta_i)|\mu_1\mu_2\rangle, \tag{10a}$$

where Σ' means that the summation is subjected to the condition $\mu\doteq\mu_1+\mu_2$, and the quantum numbers with bars have to obey $\bar{\mu}\doteq\bar{\mu}_1+\bar{\mu}_2$. The $\mathscr{N}$ is the normalization factor given by

$$\mathscr{N}=n^2\dot{\mathscr{N}}\sqrt{\frac{h_\lambda}{|\mathbf{g}|}}, \tag{10b}$$

$$\dot{\mathscr{N}}^{-1}=\left[\sum_{i=1}^{q}M_i^{-1}d_{\mu\mu}^{(\lambda)}(\beta_i)^*d_{\bar{\mu}_1\bar{\mu}_1}^{(\lambda_1)}(\beta_i)d_{\bar{\mu}_2\bar{\mu}_2}^{(\lambda_2)}(\beta_i)\right]^{1/2}. \tag{10c}$$

From (10a) we get a simple expression for the CG coefficients

$$C_{\lambda_1\mu_1,\lambda_2\mu_2}^{(\lambda)\bar{\mu},\mu}=\mathscr{N}\sum_{i=1}^{q}M_i^{-1}d_{\bar{\mu}_1\bar{\mu}_1}^{(\lambda_1)}(\beta_i)d_{\bar{\mu}_2\bar{\mu}_2}^{(\lambda_2)}(\beta_i)d_{\mu\mu}^{(\lambda)}(\beta_i)^*, \tag{11a}$$

$$\mu\doteq\mu_1+\mu_2,\quad \bar{\mu}\doteq\bar{\mu}_1+\bar{\mu}_2. \tag{11b}$$

Equation (11b) will be referred to as the magnetic quantum number conservation rule.

The quantum number $\bar{\mu}$ in (11) serves as the multiplicity label. The CG coefficients are in general not orthogonal with respect to $\bar{\mu}$. Obviously the expression (11a) is a great simplification over (9b), which involves only a summation of q terms instead of $|\mathbf{g}|$ terms.

Equation (10a) suggests that acting on $|\bar{\mu}_1\bar{\mu}_2\rangle$ the ISO's $\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}$ in (7) can be replaced by the following effective ISO's

$$\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}=n^2\sqrt{\frac{h_\lambda}{|\mathbf{g}|}}\sum_{i=1}^{q} M_i^{-1}d_{\mu\mu}^{(\lambda)}(\beta_i)^*\beta_i, \tag{12}$$

so long as we imposed the conservation rules of the magnetic quantum numbers $\bar{\mu}_1+\bar{\mu}_2\doteq\bar{\mu}$ for the initial state and $\mu_1+\mu_2\doteq\mu$ for the final state.

Remark 1: Applying the ISO (7) to the $SO_3\supset SO_2$ basis $|jm\rangle$, we get what is usually called the symmetry adapted function (SAF), or the $SO_3\supset G\supset C_n$ irreducible basis,

$$\psi_{\mu}^{\lambda}=\mathscr{N}\sum_{m}{}'\sum_{i=1}^{q} M_i^{-1}d_{\mu\mu}^{(\lambda)}(\beta_i)^*D_{mm}^{(j)}(\beta_i)|jm\rangle, \quad \bar{m}\doteq\bar{\mu}. \tag{13}$$

where Σ' means the summation is subjected to the condition $m=\mu\ (\mathrm{mod}\ n)$.

Equation (13) shows that acting on $|jm\rangle$ the ISO's $\mathscr{P}_{\mu}^{(\lambda)\bar{\mu}}$ in (7) can also be replaced by the effective ISO's (12), but with the magnetic quantum number conservation rule replaced by $m\doteq\mu$, $\bar{m}\doteq\bar{\mu}$.

Remark 2. The ISO's in the form (12) are a drastic simplification over the usual form (4). It contains only q instead of $|\mathbf{g}|$ terms. Obviously, for a given group, the larger the n, the smaller the number q of the double cosets, and the simpler the effective ISO. For the group chains $T\supset C_3$, $O\supset C_4$, and $I\supset C_5$, $q=2$, 3, and 4, respectively. Therefore the effective ISO's of the group I_h of order 120 contains only four terms.

Remark 3. An important discovery is that only the irreducible matrices of the coset generators are needed for constructing the SAF's and CG coefficients. Therefore the usual way of listing the matrices for all elements is a waste of time. For all point groups we only need to list the irreducible matrices of at most three operators (since $q_{\max}=4$, and one of the generators is the identity operator). For example, for the group I, we do not need the matrices of all 60 elements, but only three. In case one needs the matrices of the elements other than the coset generators, one can construct them easily with the help of the group table and (6).

Remark 4. Among the $|\mathbf{g}|$ elements of a group G, only the q double-coset generators play a key role, while the role played by all the remaining $|\mathbf{g}|-q$ elements, the vast majority, is merely to ensure the magnetic-quantum-number-conservation rules for the initial (trial) state and for the final (projected) state.

III. THE CG COEFFICIENTS OF $T\supset C_3$

The conventional irrep labels, the Mulliken notation or Koster notation are inadequate for writing algebraic expressions of point groups. In deriving algebraic expressions it is vital to use quantum numbers as irrep labels. The quantum number of the group T is the eigenvalue λ of the CSCO (complete set of commuting operators), $C=C_{31}^{+}+C_{32}^{+}+C_{33}^{+}+C_{34}^{+}$, of T.[12] However, in listing the CG coefficients it is better to use some kind of labeling scheme that is more familiar to readers. It is what is called the modified Mulliken notation introduced in Refs. 12 and 24 according to the following recipe:

An irrep of G is labeled by $\Gamma=A,E,F,G,H,I$ for the irreps of dimension 1,2,3,4,5,6, respectively, as in the Mulliken notation. If more than one irrep has the same dimension, we use a subscript μ_m to distinguish them, where $\mu_m=\mu_{\max}$, the largest possible quantum number of C_n, in all cases with the exception that for the irrep $E_{-3/2}$ of T, for which $\mu_m=\mu_{\min}=-\frac{3}{2}$. In analogy with the Lie group, μ_m can be regarded as the ''highest weight'' of the irrep, and the possible values of μ of an irrep are called the ''weights'' of the irrep.

For example, for the group T there are three one-dimensional irreps A_μ, $\mu=0,\pm1$, three two-dimensional irreps E_ν, $\nu=\frac{1}{2},\frac{3}{2},-\frac{3}{2}$, and one three-dimensional irrep F. The correspondence between the modified Mulliken notation, Koster notation, the eigenvalues λ, as well as the ''weights'' for each irrep of T are shown in Table I.

TABLE I. The eigenvalues λ,ρ_μ of the CSCO-II, (C,C_{3z}), of T.

	A_0	A_1	A_{-1}	F	$E_{1/2}$	$E_{3/2}$	$E_{-3/2}$
	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6	Γ_7
λ	4	4ϵ	$4\epsilon^*$	0	2	2ϵ	$2\epsilon^*$
μ (''the weights'')	0	1	-1	$(1,0,-1)$	$(\frac{1}{2},-\frac{1}{2})$	$(\frac{3}{2},\frac{1}{2})$	$(-\frac{3}{2},-\frac{1}{2})$
$\rho_\mu=\exp\left(-\frac{2\pi\mu i}{3}\right)$	1	ϵ	ϵ^*	$(\epsilon,1,\epsilon^*)$			

[a] $\epsilon=\exp(-2\pi i/3)$ and $\frac{3}{2}\doteq-\frac{3}{2}$.

We now apply the method introduced in Sec. II to the tetrahedral group. The s. v. and d. v. ISO's of $T\supset C_3$ have been obtained in (31) and (35) of Ref. 12, which can be written as (ignoring the normalizations),

$$\mathscr{P}_\mu^{(\lambda)\bar{\mu}}=P^\mu\delta_{\mu\bar{\mu}}+d_{\mu\bar{\mu}}^{(\lambda)*}(C_{2z})P^\mu C_{2z}P^{\bar{\mu}}, \tag{14}$$

where P^μ is the projection operator of C_3, λ stands for the irrep A_μ ($\mu=0,\pm1$), F, E_ν ($\nu=\frac{1}{2},\pm\frac{3}{2}$), and $d_{\mu\bar{\mu}}^{(\lambda)}(C_{2z})$ is the matrix element of irrep λ of C_{2z}. The subgroup C_3 is generated from C_{31}^+,

$$C_{31}^+=\exp\left(-\frac{2\pi J_z i}{3}\right). \tag{15}$$

Its eigenvalue is denoted by $\rho\equiv\rho_\mu$, $\rho=\rho_\mu=\exp(-2\pi\mu i/3)$, with the range of μ given in (3c). Notice that the irrep $\frac{3}{2}$ of C_3 is equivalent to $-\frac{3}{2}$, i.e., $\frac{3}{2}\doteq-\frac{3}{2}$.

The effective ISO's for $T\supset C_3$ are

$$\mathscr{P}_\mu^{(\lambda)\bar{\mu}}=\delta_{\mu\bar{\mu}}+3d_{\mu\bar{\mu}}^{(\lambda)*}(C_{2z})C_{2z}. \tag{16}$$

The algebraic expressions for the irreducible matrix of the generator C_{2z} has been found in (37) and (38) of Ref. 12,

$$d^{(A_\mu)}(C_{2z})=1,\quad d_{\mu\bar{\mu}}^{(F)}(C_{2z})=\frac{2}{3}-\delta_{\mu\bar{\mu}},$$

$$d_{\mu\bar{\mu}}^{(E_\nu)}(C_{2z})=\frac{\nu}{|\nu|}i(-1)^{\delta_{\mu\nu}\delta_{\mu\bar{\mu}}}\sqrt{\frac{2-\delta_{\mu\bar{\mu}}}{3}}, \tag{17}$$

where $\nu=\frac{1}{2}$, $\mu,\bar{\mu}=\pm\frac{1}{2}$; $\nu=\frac{3}{2}$, $\mu,\bar{\mu}=\frac{1}{2},\frac{3}{2}$; and $\nu=-\frac{3}{2}$, $\mu,\bar{\mu}=-\frac{1}{2},-\frac{3}{2}$.

In Ref. 12 an ISO with a chosen intrinsic quantum number $\bar{\mu}$ is called the generalized projection operator, which is to be used as projection operator for constructing the CG coefficients. Corresponding to the $T\supset C_3$ generalized projection operators in (40) of Ref. 12, we have the effective generalized projection operators

$$\begin{aligned}
&\mathbf{P}^{A_\mu}=1+3C_{2z}, && \bar{\mu}_1+\bar{\mu}_2\doteq\mu,\\
&\mathbf{P}_\mu^{F}=\delta_{\mu0}-(-1)^\mu C_{2z}, && \bar{\mu}_1+\bar{\mu}_2\doteq0,\\
&\mathbf{P}_\mu^{E_\nu}=\delta_{\mu\nu}+(-1)^{\mu-1/2+|\nu|}\sqrt{3}C_{2z}, && \bar{\mu}_1+\bar{\mu}_2\doteq\nu,
\end{aligned} \tag{18}$$

where the condition $\mu_1+\mu_2\doteq\mu$ has been omitted for simplicity.

A. The CG series

Instead of using the character theory, we can use a method similar to the one used in Lie groups[8,25] to determine the CG series of T. As mentioned before, in the modified Mulliken

notation Γ_{μ_m},μ_m can be regarded as the ''highest weight'' of the irrep. However the ''highest weight'' is not unique for some irreps. For example, the irreps A_1 and F have the same ''highest weight'' 1. However, the nonuniqueness of the ''highest weight'' causes so much trouble for us.

First we construct the weight system of the product representation $\lambda_1\times\lambda_2$, which is an algebraic sum of the weights of the irreps λ_1 and λ_2 modulo 3, and then use the subtraction method to determine the CG series.

Example: Let us consider $F\times F$. The weights of F are $(1,0,-1)$ and the weight system of $F\times F$ is $(2,1,0,-1,-2;1,0,-1;0)\doteq(1^3,0^3,-1^3)$. Since the irrep F is self-time-reversal, the identity irrep A_0 occurs once and only once in the product $F\times F$. Consulting Table I we get the CG series $F\times F=A_0+A_1+A_{-1}+2F$.

In this way we can easily obtain the following results.

$$\begin{gathered}A_{\mu_1}\times A_{\mu_2}=A_{\mu_1+\mu_2},\qquad A_\mu\times F=F,\quad \mu=0,\pm1,\\ F\times F=A_0+A_1+A_{-1}+2F,\\ F\times E_\nu=E_{1/2}+E_{3/2}+E_{-3/2},\quad \nu=\frac{1}{2},\pm\frac{3}{2},\\ A_{\pm1}\times(E_{1/2},E_{\pm3/2},E_{\mp3/2})=(E_{\pm3/2},E_{\mp3/2},E_{1/2}),\\ E_{\nu_1}\times E_{\nu_2}=A_\mu+F,\end{gathered}\tag{19a}$$

with μ given by

(ν_1,ν_2)	$(\frac{1}{2},\frac{1}{2})$	$(\frac{3}{2},-\frac{3}{2})$	$(\frac{1}{2},\frac{3}{2})$	$(-\frac{3}{2},-\frac{3}{2})$	$(\frac{1}{2},-\frac{3}{2})$	$(\frac{3}{2},\frac{3}{2})$
μ	0	0	1	1	-1	-1

(19b)

Using the eigenvalue λ as the irrep label, the last CG series can be written in a concise form:

$$(\lambda_1)\times(\lambda_2)=A_\mu+F,\quad \rho_\mu=\frac{1}{4}\lambda_1\lambda_2,\tag{20}$$

where $\lambda=2,\ 2\exp(-\frac{2}{3}\pi i),\ 2\exp(\frac{2}{3}\pi i)$ for the irrep $E_{1/2},E_{3/2},E_{-3/2}$.

B. The $T\supset C_3$ CG coefficients

Applying (18) to an appropriate initial state $|\lambda_1\overline{\mu_1}\lambda_2\overline{\mu_2}\rangle$ and using the CG series (19) we can obtain the CG coefficients easily. For the group T there is no multiplicity except that the irrep F occurs twice in the product $F\times F$. Therefore we will omit the multiplicity label $\overline{\mu_1}\overline{\mu_2}$ except for the irrep F in $F\times F$.

(1) $A_{\mu_1}\times A_{\mu_2}=A_{\mu_1+\mu_2}$,

$$\Psi^{A_{\mu_1+\mu_2}}=|A_{\mu_1},A_{\mu_2}\rangle,\tag{21a}$$

(2) $A_{\mu_1}\times F=F$,

$$\Psi^{F}_{\mu_1+\mu_2}=|A_{\mu_1},F\mu_2\rangle,\tag{21b}$$

(3) $A_{\mu_1}\times E_\nu=E_{\nu'}$,

$$\Psi^{E_{\nu'}}_{\mu_1+\mu_2}=|A_{\mu_1},E_\nu\mu_2\rangle,\tag{21c}$$

(4) $F\times F=A_1+A_0+A_{-1}+2F$,

TABLE II. The trial states $|\bar{\mu}_1\bar{\mu}_2\rangle$ used in (22b) and (22c).

$F\times E_{1/2}$:	$E_{1/2}$	$\|0,\frac{1}{2}\rangle$,	$E_{3/2}$	$\|1,\frac{1}{2}\rangle$, $E_{-3/2}\|1,\frac{1}{2}\rangle$
$F\times E_{3/2}$:	$E_{1/2}$	$\|0,\frac{1}{2}\rangle$,	$E_{3/2}$	$\|0,\frac{3}{2}\rangle$, $E_{-3/2}\|0,\frac{3}{2}\rangle$
$F\times E_{-3/2}$:	$E_{1/2}$	$\|1,-\frac{1}{2}\rangle$,	$E_{3/2}$	$\|0,\frac{3}{2}\rangle$, $E_{-3/2}\|0,\frac{3}{2}\rangle$
$E_{1/2}\times E_{1/2}$:	A_0	$\|\frac{1}{2},-\frac{1}{2}\rangle$,	F	$\|\frac{1}{2},-\frac{1}{2}\rangle$
$E_{1/2}\times E_{3/2}$:	A_1	$\|\frac{1}{2},\frac{1}{2}\rangle$,	F	$\|-\frac{1}{2},\frac{1}{2}\rangle$
$E_{1/2}\times E_{-3/2}$:	A_{-1}	$\|-\frac{1}{2},-\frac{1}{2}\rangle$,	F	$\|\frac{1}{2},-\frac{1}{2}\rangle$
$E_{3/2}\times E_{3/2}$:	A_{-1}	$\|\frac{1}{2},\frac{3}{2}\rangle$,	F	$\|\frac{3}{2},\frac{3}{2}\rangle$
$E_{3/2}\times E_{-3/2}$:	A_0	$\|\frac{1}{2},-\frac{1}{2}\rangle$,	F	$\|\frac{1}{2},-\frac{1}{2}\rangle$
$E_{-3/2}\times E_{-3/2}$:	A_1	$\|-\frac{1}{2},\frac{3}{2}\rangle$,	F	$\|\frac{3}{2},\frac{3}{2}\rangle$

$$\Psi^{A_\mu}=|F0,F\mu\rangle+3\sum_{\mu_1,\mu_2}{}'d^{(F)}_{\mu_1 0}d^{(F)}_{\mu_2\mu}|F\mu_1,F\mu_2\rangle,$$

$$\Psi^{(F)\bar{\mu}_1\bar{\mu}_2}_{\mu}=|F\bar{\mu}_1,F\bar{\mu}_2\rangle\delta_{\mu 0}-(-1)^{\mu}\sum_{\mu_1\mu_2}{}'d^{(F)}_{\mu_1\bar{\mu}_1}d^{(F)}_{\mu_2\bar{\mu}_2}|F\mu_1,F\mu_2\rangle, \tag{22a}$$

where for simplicity we ignore the normalization factor and $d^{(F)}_{\mu\bar{\mu}}=d^{(F)}_{\mu\bar{\mu}}(C_{2z})$. Since the irrep F occurs twice, we need take two initial states, which are chosen as $|F0,F0\rangle$ and $|F1,F-1\rangle$. The two sets of the CG coefficients are not orthogonal. The orthogonalized CG coefficients are listed in Table III.

(5) $F\times E_\nu=\Sigma_{\nu'}E_{\nu'}$

$$\Psi^{E_{\nu'}}_{\mu}=|F\bar{\mu}_1,E_\nu\bar{\mu}_2\rangle\delta_{\mu\nu'}+(-1)^{\mu-1/2+|\nu'|}\sqrt{3}\sum_{\mu_1\mu_2}{}'d^{(F)}_{\mu_1\bar{\mu}_1}d^{(E_\nu)}_{\mu_2\bar{\mu}_2}|F\mu_1,E_\nu\mu_2\rangle,$$

$$\bar{\mu}_1+\bar{\mu}_2\doteq\nu'. \tag{22b}$$

(6) $E_{\nu_1}\times E_{\nu_2}=A_\mu+F$, or $(\lambda_1)\times(\lambda_2)=A_\mu+F$.

$$\Psi^{A_\mu}=|\lambda_1\bar{\mu}_1,\lambda_2\bar{\mu}_2\rangle+3\sum_{\mu_1\mu_2}{}'d^{(\lambda_1)}_{\mu_1\bar{\mu}_1}d^{(\lambda_2)}_{\mu_2\bar{\mu}_2}|\lambda_1\mu_1,\lambda_2\mu_2\rangle,$$

$$\mu\doteq\bar{\mu}_1+\bar{\mu}_2,$$

$$\Psi^{F}_{\mu}=|\lambda_1\bar{\mu}_1,\lambda_2\bar{\mu}_2\rangle\delta_{\mu 0}-(-1)^{\mu}\sum_{\mu_1\mu_2}{}'d^{(\lambda_1)}_{\mu_1\bar{\mu}_1}d^{(\lambda_2)}_{\mu_2\bar{\mu}_2}|\lambda_1\mu_1,\lambda_2\mu_2\rangle,$$

$$\bar{\mu}_1+\bar{\mu}_2\doteq 0. \tag{22c}$$

The initial states $|F\bar{\mu}_1,E_\nu\bar{\mu}_2\rangle$ and $|\lambda_1\bar{\mu}_1,\lambda_2\bar{\mu}_2\rangle$ in cases 5 and 6 are listed in Table II.

From the algebraic expression (22) we easily constructed the numerical table shown in Table III, where only the CG coefficients with $\Gamma_1\leqslant\Gamma_2$ are listed. The CG coefficients have the property that

$$C^{(\Gamma)r,\mu}_{\Gamma_2\mu_2,\Gamma_1\mu_1}=(-1)^{\Gamma_1+\Gamma_2+\Gamma+r}C^{(\Gamma)r,\mu}_{\Gamma_1\mu_1,\Gamma_2\mu_2}, \tag{23}$$

where r is the multiplicity label, $r=0$ for multiplicity free cases, or $r=0,1$ for $F\times F\to F$, and the phase $(-1)^{\Gamma}=1,-1,i$ for the irreps A_μ,F,E_ν, respectively.[14]

It is to be noted that all the $T\supset C_3$ CG coefficients are real.

TABLE III. The CG coefficients of the group chain $T \supset C_3$.

$F \times F$

$$\Psi^{A_0} = \frac{1}{\sqrt{3}}[|00\rangle + |1\,-1\rangle + |-\,11\rangle], \qquad \Psi^{A_1} = \frac{1}{\sqrt{3}}[|01\rangle + |10\rangle + |-\,-1\,1\rangle]$$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{3}}[|0\,-1\rangle + |-\,10\rangle + |11\rangle],$$

$$\Psi_0^{(F)0} = \sqrt{\tfrac{2}{3}}|00\rangle - \frac{1}{\sqrt{6}}|1\,-1\rangle - \frac{1}{\sqrt{6}}|-\,11\rangle, \qquad \Psi_0^{(F)1} = \frac{1}{\sqrt{2}}[|1\,-1\rangle - |-\,11\rangle]$$

$$\Psi_1^{(F)0} = \sqrt{\tfrac{2}{3}}\,|-1\,-1\rangle - \frac{1}{\sqrt{6}}|10\rangle - \frac{1}{\sqrt{6}}|01\rangle, \qquad \Psi_1^{(F)1} = \frac{1}{\sqrt{2}}[|01\rangle - |10\rangle]$$

$$\Psi_{-1}^{(F)0} = \sqrt{\tfrac{2}{3}}\,|-10\rangle - \frac{1}{\sqrt{6}}|11\rangle - \frac{1}{\sqrt{6}}|0\,-1\rangle, \qquad \Psi_{-1}^{(F)1} = \frac{1}{\sqrt{2}}[|-10\rangle - |0\,-1\rangle]$$

$F \times E_{1/2}$

$$\Psi_{1/2}^{E_{1/2}} = \sqrt{\tfrac{1}{3}}|0\,\tfrac{1}{2}\rangle - \sqrt{\tfrac{2}{3}}|1\,-\tfrac{1}{2}\rangle, \qquad \Psi_{-1/2}^{E_{1/2}} = -\sqrt{\tfrac{2}{3}}|-1\,\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|0\,-\tfrac{1}{2}\rangle$$

$$\Psi_{1/2}^{E_{3/2}} = \sqrt{\tfrac{2}{3}}|0\,\tfrac{1}{2}\rangle + \sqrt{\tfrac{1}{3}}|1\,-\tfrac{1}{2}\rangle, \qquad \Psi_{3/2}^{E_{3/2}} = -\sqrt{\tfrac{1}{3}}|1\,\tfrac{1}{2}\rangle + \sqrt{\tfrac{2}{3}}|-1\,-\tfrac{1}{2}\rangle$$

$$\Psi_{-1/2}^{E_{-3/2}} = \sqrt{\tfrac{2}{3}}|0\,-\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|-1\,\tfrac{1}{2}\rangle, \qquad \Psi_{3/2}^{E_{-3/2}} = -\sqrt{\tfrac{2}{3}}|1\,\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|-1\,-\tfrac{1}{2}\rangle$$

$F \times E_{3/2}$

$$\Psi_{1/2}^{E_{1/2}} = \sqrt{\tfrac{2}{3}}|0\,\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|-1\,\tfrac{3}{2}\rangle, \qquad \Psi_{-1/2}^{E_{1/2}} = \sqrt{\tfrac{1}{3}}|-1\,\tfrac{1}{2}\rangle + \sqrt{\tfrac{2}{3}}|1\,\tfrac{3}{2}\rangle$$

$$\Psi_{1/2}^{E_{3/2}} = \sqrt{\tfrac{1}{3}}|0\,\tfrac{1}{2}\rangle + \sqrt{\tfrac{2}{3}}|-1\,\tfrac{3}{2}\rangle, \qquad \Psi_{3/2}^{E_{3/2}} = \sqrt{\tfrac{2}{3}}|1\,\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|0\,\tfrac{3}{2}\rangle$$

$$\Psi_{-1/2}^{E_{-3/2}} = -\sqrt{\tfrac{2}{3}}|-1\,\tfrac{1}{2}\rangle + \sqrt{\tfrac{1}{3}}|1\,\tfrac{3}{2}\rangle, \qquad \Psi_{3/2}^{E_{-3/2}} = \sqrt{\tfrac{2}{3}}|0\,\tfrac{3}{2}\rangle + \sqrt{\tfrac{1}{3}}|1\,\tfrac{1}{2}\rangle$$

$F \times E_{-3/2}$

$$\Psi_{1/2}^{E_{1/2}} = -\sqrt{\tfrac{1}{3}}|1\,-\tfrac{1}{2}\rangle - \sqrt{\tfrac{2}{3}}|-1\,\tfrac{3}{2}\rangle, \qquad \Psi_{-1/2}^{E_{1/2}} = \sqrt{\tfrac{2}{3}}|0\,-\tfrac{1}{2}\rangle - \sqrt{\tfrac{1}{3}}|1\,\tfrac{3}{2}\rangle$$

$$\Psi_{1/2}^{E_{3/2}} = \sqrt{\tfrac{1}{3}}|-1\,\tfrac{3}{2}\rangle - \sqrt{\tfrac{2}{3}}|1\,-\tfrac{1}{2}\rangle, \qquad \Psi_{3/2}^{E_{3/2}} = \sqrt{\tfrac{2}{3}}|0\,\tfrac{3}{2}\rangle + \sqrt{\tfrac{1}{3}}|-1\,-\tfrac{1}{2}\rangle$$

$$\Psi_{-1/2}^{E_{-3/2}} = -\sqrt{\tfrac{2}{3}}|1\,\tfrac{3}{2}\rangle - \sqrt{\tfrac{1}{3}}|0\,-\tfrac{1}{2}\rangle, \qquad \Psi_{3/2}^{E_{-3/2}} = \sqrt{\tfrac{1}{3}}|0\,\tfrac{3}{2}\rangle - \sqrt{\tfrac{2}{3}}|-1\,-\tfrac{1}{2}\rangle$$

$E_{1/2} \times E_{1/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi_0^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle], \qquad \Psi_1^{(F)} = |\tfrac{1}{2}\,\tfrac{1}{2}\rangle, \qquad \Psi_{-1}^{(F)} = -|-\tfrac{1}{2}\,-\tfrac{1}{2}\rangle$$

$E_{1/2} \times E_{3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{3}{2}\rangle],$$

$$\Psi_0^{(F)} = -|-\tfrac{1}{2}\,\tfrac{1}{2}\rangle, \qquad \Psi_1^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{3}{2}\rangle], \qquad \Psi_{-1}^{(F)} = |\tfrac{1}{2}\,\tfrac{3}{2}\rangle$$

$E_{1/2} \times E_{-3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|-\tfrac{1}{2}\,-\tfrac{1}{2}\rangle + |\tfrac{1}{2}\,\tfrac{3}{2}\rangle],$$

$$\Psi_0^{(F)} = |\tfrac{1}{2}\,-\tfrac{1}{2}\rangle, \qquad \Psi_1^{(F)} = |-\tfrac{1}{2}\,\tfrac{3}{2}\rangle, \qquad \Psi_{-1}^{(F)} = \frac{1}{\sqrt{2}}[|-\tfrac{1}{2}\,-\tfrac{1}{2}\rangle - |\tfrac{1}{2}\,\tfrac{3}{2}\rangle]$$

$E_{3/2} \times E_{3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,\tfrac{3}{2}\rangle - |\tfrac{3}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi_0^{(F)} = -|\tfrac{3}{2}\,\tfrac{3}{2}\rangle, \qquad \Psi_1^{(F)} = |\tfrac{1}{2}\,\tfrac{1}{2}\rangle, \qquad \Psi_{-1}^{(F)} = -\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,\tfrac{3}{2}\rangle + |\tfrac{3}{2}\,\tfrac{1}{2}\rangle]$$

$E_{3/2} \times E_{-3/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,-\tfrac{1}{2}\rangle + |\tfrac{3}{2}\,\tfrac{3}{2}\rangle],$$

$$\Psi_0^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}\,-\tfrac{1}{2}\rangle - |\tfrac{3}{2}\,\tfrac{3}{2}\rangle], \qquad \Psi_1^{(F)} = |\tfrac{3}{2}\,-\tfrac{1}{2}\rangle, \qquad \Psi_{-1}^{(F)} = |\tfrac{1}{2}\,\tfrac{3}{2}\rangle$$

$E_{-3/2} \times E_{-3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|-\tfrac{1}{2}\,\tfrac{3}{2}\rangle - |\tfrac{3}{2}\,-\tfrac{1}{2}\rangle],$$

$$\Psi_0^{(F)} = -|\tfrac{3}{2}\,\tfrac{3}{2}\rangle, \qquad \Psi_1^{(F)} = -\frac{1}{\sqrt{2}}[|\tfrac{3}{2}\,-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{3}{2}\rangle], \qquad \Psi_{-1}^{(F)} = |-\tfrac{1}{2}\,-\tfrac{1}{2}\rangle$$

IV. THE $T\supset D_2\supset C_2$ CG COEFFICIENTS

The group chain $T\supset D_2\supset C_2$ is a little bit more complicated than $T\supset C_3$ in that for the s. v. reps, the subgroup C_2 is redundant, while for the d. v. reps, the subgroup D_2 is redundant. In other words, we actually use two different group chains, $T\supset D_2$ and $T\supset C_2$, for the s. v. and d. v. reps of T, respectively. In this section we need to use the ISO's, instead of the effective ISO's. Now the axis of C_{2z} is chosen as the z axis. The $T\supset D_2$ generalized projection operators have been found in (14a) and (15c) of Ref. 13 and are cited in the following,

1. One-dimensional s. v. irreps:

$$\mathbf{P}^{(A_\mu)}=P^{\rho_\mu}\wp^{++},\quad \mu=0,\pm 1, \tag{24a}$$

where P^{ρ_μ} and $\wp^{\nu\nu'}$ are the projection operators of C_3 and D_2, respectively,

$$\wp^{\nu\nu'}=(e+\nu C_{2x})(e+\nu' C_{2y}),\quad \nu,\nu'=\pm 1. \tag{24b}$$

The set of eigenvalues $(\nu\nu')=(++),(+-),(-+),(--)$ can be used to label the four s. v. irreps of D_2.

2. Three-dimensional s. v. irrep:

From the third column in the matrix given in (15c) of Ref. 13 and using (15a) in Ref. 13 we have

$$\begin{cases}\mathbf{P}_x^{(F)}=C_{31}^{+}\wp^{--},\\ \mathbf{P}_y^{(F)}=C_{31}^{-}\wp^{--},\\ \mathbf{P}_z^{(F)}=\wp^{--},\end{cases} \tag{24c}$$

where the components x,y,z have the quantum numbers $(\nu\nu')=(+-,-+,--)$, respectively.

The irreducible matrix of C_{31}^{+} is given in (16) of Ref. 13,

$$\mathscr{D}^F(C_{31}^{+})=\tilde{\mathscr{D}}^F(C_{31}^{-})=\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}. \tag{24d}$$

3. Two-dimensional d. v. irreps:

From (36) in Ref. 13 we have the $T\supset C_2$ generalized projection operators,

$$\begin{cases}\mathbf{P}_{1/2}^{(\lambda)}=P^{1/2}+\dfrac{1}{2\sqrt{2}}P^{1/2}(i^{-1/2}\lambda C_{31}^{-}+i^{1/2}\lambda^{*}C_{31}^{+})P^{1/2},\\ \mathbf{P}_{-1/2}^{(\lambda)}=\sqrt{i}C_{2x}P^{1/2}-\dfrac{1}{2\sqrt{2}}P^{-1/2}(\lambda C_{31}^{-}-\lambda^{*}C_{31}^{+})P^{1/2},\end{cases} \tag{25}$$

where $P^{\pm 1/2}$ are the projection operators of C_2, $P^{\sigma}=1+\varrho_\sigma^{*}C_{2z}$, with $\varrho_\sigma=\exp(-\pi\sigma i)$, and $\lambda=2$, $2\epsilon\equiv 2\exp(-\frac{2}{3}\pi i)$, $2\epsilon^{*}$ for the irreps $E_{1/2}$, $E_{3/2}$ and $E_{-3/2}$, respectively. Recall that μ, ν, ν', and σ are the quantum numbers of C_{31}^{+}, C_{2x}, C_{2y}, and C_{2z}, respectively.

The irreducible matrices of $C_{31}^{\pm}$, C_{2x}, C_{2y} are listed in Eq. (37) of Ref. 13

$$D^{(\lambda)}(C_{31}^{+})=\tilde{D}^{(\lambda)}(C_{31}^{-})^{*}=\frac{\lambda}{2\sqrt{2}}\begin{pmatrix}\sqrt{i^{*}}, & -1\\ 1, & \sqrt{i}\end{pmatrix}, \tag{26a}$$

$$D^{(\lambda)}(C_{2x})=-\tilde{D}^{(\lambda)}(C_{2y})=\begin{pmatrix}0 & i\sqrt{i}\\ \sqrt{i} & 0\end{pmatrix}^{*}. \tag{26b}$$

The CG coefficients for the three cases are found in the following way.

A. s. v.×s. v.→s. v. reps

The projection operators for the s. v. irreps of T in (24a) and (24c) can be written as $\mathbf{P}^{(\Gamma)}=O^{\Gamma}\wp^{\nu\nu'}$, where O^{Γ} is an operator depending on Γ, and $(\nu\nu')=(++)$ and $(--)$ for $\Gamma=A_{\mu}$ and $\Gamma=F$, respectively. The quantum numbers of the initial state $|\Gamma_1\nu_1\nu_1',\Gamma_2\nu_2\nu_2'\rangle$ have to obey the following conditions:

$$\nu_1\nu_2=\nu, \quad \nu_1'\nu_2'=\nu'. \tag{27}$$

Equation (27) is crucial for choosing the initial state. Notice that the quantum numbers $(\nu\nu')=(++)$ for the states $|A_{\mu}\rangle$ and $\nu\nu'=+-,-+,--$ for the states $|Fx\rangle,|Fy\rangle,|Fz\rangle$, respectively.

(1) $A_{\mu_1}\times A_{\mu_2}=A_{\mu_1+\mu_2}$. Equation (21a) still holds.

(2) $A_{\mu}\times F=F,\ \mu=0,\pm1$.

Applying (24c) to the state $|A_{\mu},Fz\rangle$ which has $(\nu\nu')=(--)$ and using (24d), we obtain

$$\Psi_k^F=\rho_{\mu}^{\kappa}|A_{\mu},Fk\rangle, \quad k=x,y,z, \tag{28a}$$

where $\kappa=1,2,3$ for $k=x,y,z$, respectively.

(3) $F\times F=A_0+A_1+A_{-1}+2F$.

Applying (24a) to the state $|Fx,Fx\rangle$ with $(\nu\nu')=(++)$, and (24c) to the states $|Fx,Fy\rangle \pm|Fy,Fx\rangle$ which has $(\nu\nu')=(--)$, we obtain the CG coefficients of irrep A_{μ} and F, respectively,

$$\Psi^{A_{\mu}}=\frac{1}{\sqrt{3}}[|Fx,Fx\rangle+\rho_{\mu}^{*}|Fy,Fy\rangle+\rho_{\mu}|Fz,Fz\rangle],$$

$$\Psi_z^{(F)r}=\sqrt{\frac{1}{2}}\,[|Fx,Fy\rangle+(-1)^r|Fy,Fx\rangle], \quad r=0,1; \quad x,y,z \text{ in cyclic.} \tag{28b}$$

B. s. v.×d. v.→d. v. reps

The projection operators in (25) have the intrinsic quantum number[13] $\bar{\sigma}=\frac{1}{2}$, therefore the quantum numbers of the initial state $|\Gamma_1\bar{\sigma}_1,\Gamma_2\bar{\sigma}_2\rangle$ have to obey

$$\bar{\sigma}_1+\bar{\sigma}_2=\frac{1}{2}.$$

The quantum number σ of C_{2z} is not explicitly given for the s. v. irreducible basis vectors and needs to be found out in the following way: Since in the states $|A_{\mu}\rangle$ $(\mu=0,\pm1)$ and $|Fz\rangle$, the operator C_{2z} has the eigenvalue $\varrho_{\sigma}=\exp(-\pi\sigma i)=1$, these states have $\sigma=0$.

(1) $A_{\mu}\times E_{\nu}=E_{\nu'}$.

We choose the state $|A_{\mu},E_{\nu\frac{1}{2}}\rangle$ which has $\sigma=\frac{1}{2}$ as the initial state for projection. Using (25) we have

$$\Psi_{\sigma}^{(E_{\nu'})}=|A_{\mu},E_{\nu}\sigma\rangle. \tag{29a}$$

(2) $F\times E_{\nu}=\Sigma_{\nu'}E_{\nu'}$, or $F\times(\lambda)=\Sigma_{\lambda'}(\lambda')$

Similarly, choosing $|Fz,\lambda\frac{1}{2}\rangle$ as the initial state for projection we obtain,

$$\Psi_{1/2}^{(\lambda')}=\frac{1}{\sqrt{3}}\left[\left|Fz,\lambda\,\frac{1}{2}\right\rangle-\frac{1}{4}\left(i^{-1/2}\lambda'\lambda^{*}\left|Fy,\lambda-\frac{1}{2}\right\rangle-i^{1/2}\lambda'^{*}\lambda\left|Fx,\lambda-\frac{1}{2}\right\rangle\right)\right],$$

$$\Psi_{-1/2}^{(\lambda')}=\frac{1}{\sqrt{3}}\left[-\left|Fz,\lambda-\frac{1}{2}\right\rangle-\frac{1}{4}\left(i^{1/2}\lambda'\lambda^{*}\left|Fy,\lambda\,\frac{1}{2}\right\rangle-i^{-1/2}\lambda'^{*}\lambda\left|Fx,\lambda\,\frac{1}{2}\right\rangle\right)\right]. \tag{29b}$$

C. d. v.×d. v.→s. v. reps

We now consider $(\lambda_1)\times(\lambda_2)=A_\mu+F$. From (26) we have

$$C_{2x}\left|\lambda_1\,\frac{1}{2},\lambda_2-\frac{1}{2}\right\rangle=-\left|\lambda_1-\frac{1}{2},\lambda_2\,\frac{1}{2}\right\rangle,\quad C_{2y}\left|\lambda_1\,\frac{1}{2},\lambda_2-\frac{1}{2}\right\rangle=-\left|\lambda_1-\frac{1}{2},\lambda_2\,\frac{1}{2}\right\rangle. \tag{29c}$$

Applying (24a) to the state $|\lambda_1\,\frac{1}{2},\lambda_2-\frac{1}{2}\rangle$ and using (24b) and (29c) we obtain compact expressions for the CG coefficients of the irrep A_μ and F,

$$\begin{cases}\Psi^{A_\mu}=\dfrac{1}{\sqrt{2}}\left(\left|\lambda_1\,\dfrac{1}{2},\lambda_2-\dfrac{1}{2}\right\rangle-\left|\lambda_1-\dfrac{1}{2},\lambda_2\,\dfrac{1}{2}\right\rangle\right)\delta_{\rho_\mu,\lambda_1\lambda_2/4},\\ \Psi_x^F=-\dfrac{\lambda_1\lambda_2}{4\sqrt{2}}i^{-1/2}\left(\left|\lambda_1\dfrac{1}{2},\lambda_2\dfrac{1}{2}\right\rangle-i\left|\lambda_1-\dfrac{1}{2},\lambda_2-\dfrac{1}{2}\right\rangle\right),\\ \Psi_y^F=\dfrac{\lambda_1^*\lambda_2^*}{4\sqrt{2}}i^{1/2}\left(\left|\lambda_1\,\dfrac{1}{2},\lambda_2\,\dfrac{1}{2}\right\rangle+i\left|\lambda_1-\dfrac{1}{2},\lambda_2-\dfrac{1}{2}\right\rangle\right),\\ \Psi_z^F=\dfrac{1}{\sqrt{2}}\left(\left|\lambda_1\,\dfrac{1}{2},\lambda_2-\dfrac{1}{2}\right\rangle+\left|\lambda_1-\dfrac{1}{2},\lambda_2\,\dfrac{1}{2}\right\rangle\right).\end{cases} \tag{29d}$$

Equations (28) and (29) are the algebraic expressions for the $T\supset D_2\supset C_2$ CG coefficients. From (29) we can easily obtain the numerical results as listed in Table IV. It is seen that some of the CG coefficients are imaginary, in agreement with the assertion of Butler[26] that some $3jm$ factors of $T\supset D_2$ must be complex. However, the $T\supset C_2$ CG coefficients can be chosen to be all real, as shown in Ref. 16. In the next section, we are going to derive the algebraic expressions for the $T\supset C_2$ real CG coefficients.

V. THE $T\supset C_2$ CG COEFFICIENTS

The group chain $T\supset C_2$ is not canonical for the s. v. reps, as is seen from the fact that both the projection operators $\mathbf{P}_x^{(F)}$ and $\mathbf{P}_y^{(F)}$ in (24c) have $C_{2z}=-1$. They can be linearly combined into the two new operators $\mathbf{P}_1^{(F)}$ and $\mathbf{P}_2^{(F)}$ so that $\mathbf{P}_1^{(F)},\mathbf{P}_2^{(F)},\mathbf{P}_3^{(F)}(=\mathbf{P}_z^{(F)})$ carry the irrep F which is identical to that of Ref. 16.

$$\mathbf{P}_1^{(F)}=\frac{1}{\sqrt{2}}[i^{3/2}C_{31}^{+}+i^{-3/2}C_{31}^{-}]\wp^{--},$$

$$\mathbf{P}_2^{(F)}=-\frac{i}{\sqrt{2}}[i^{3/2}C_{31}^{+}-i^{-3/2}C_{31}^{-}]\wp^{--},\qquad \mathbf{P}_3^{(F)}=\wp^{--}, \tag{30}$$

where the projection operators of components 1,2,3 have the quantum numbers $C_{2z}=-1,-1,1$ or $\sigma=1,1,0$, respectively [recall $\varrho_\sigma=\exp(-\pi\sigma i)$].

The corresponding irreducible matrices of C_{2z}, C_{2x}, C_{31}^{+} and C_{31}^{-} are

$$D^{(F)}(C_{2z})=\begin{pmatrix}-1&0&0\\0&-1&0\\0&0&1\end{pmatrix},\quad D^{(F)}(C_{2x})=\begin{pmatrix}0&-i&0\\i&0&0\\0&0&-1\end{pmatrix},$$
$$D^{(F)}(C_{31}^{+})=\tilde{D}^{(F)*}(C_{31}^{-})=\frac{1}{2}\begin{pmatrix}-i&-1&-\sqrt{2i}\\-1&i&\sqrt{2i^*}\\-\sqrt{2i}&\sqrt{2i^*}&0\end{pmatrix}. \tag{31}$$

We also change the phase factor of the d. v. projection operator $\mathbf{P}_{-1/2}^{(\lambda)}$ in (25) by multiplying it with i, i.e., the new d. v. projection operators are

TABLE IV. The CG coefficients of $T \supset D_2 \supset C_2$.

$F \times E_{1/2}$

$$\Psi_{1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}|y-\tfrac{1}{2}\rangle + i^{1/2}|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}|y\tfrac{1}{2}\rangle + i^{-1/2}|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon^*|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon^*|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon^*|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon^*|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon|x\tfrac{1}{2}\rangle]$$

$F \times E_{3/2}$

$$\Psi_{1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon^*|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon^*|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}|y-\tfrac{1}{2}\rangle + i^{1/2}|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}|y\tfrac{1}{2}\rangle + i^{-1/2}|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon^*|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon^*|x\tfrac{1}{2}\rangle]$$

$F \times E_{-3/2}$

$$\Psi_{1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon^*|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{1/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon^*|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}\epsilon^*|y-\tfrac{1}{2}\rangle + i^{1/2}\epsilon|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}\epsilon^*|y\tfrac{1}{2}\rangle + i^{-1/2}\epsilon|x\tfrac{1}{2}\rangle]$$

$$\Psi_{1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[|z\tfrac{1}{2}\rangle - i^{-1/2}|y-\tfrac{1}{2}\rangle + i^{1/2}|x-\tfrac{1}{2}\rangle], \qquad \Psi_{-1/2}^{(E_{-3/2})} = \frac{1}{\sqrt{3}}[-|z-\tfrac{1}{2}\rangle - i^{1/2}|y\tfrac{1}{2}\rangle + i^{-1/2}|x\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{1/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\epsilon\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\epsilon^*\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{-3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\epsilon^*\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\epsilon\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$E_{3/2} \times E_{3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\epsilon^*\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\epsilon\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$E_{3/2} \times E_{-3/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$E_{-3/2} \times E_{-3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\tfrac{1}{2}\rangle], \qquad \Psi_x^{(F)} = -i^{-1/2}\epsilon\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle - i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle]$$

$$\Psi_y^{(F)} = i^{1/2}\epsilon^*\frac{1}{\sqrt{2}}[|\tfrac{1}{2}\tfrac{1}{2}\rangle + i|-\tfrac{1}{2}-\tfrac{1}{2}\rangle], \qquad \Psi_z^{(F)} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\tfrac{1}{2}\rangle]$$

$\epsilon = \exp(-2\pi i/3)$.

$$\mathbf{P}^{(\lambda)}_{1/2}=P^{1/2}+\frac{1}{2\sqrt{2}}P^{1/2}(i^{-1/2}\lambda C^{-}_{31}+i^{1/2}\lambda^{*}C^{+}_{31})P^{1/2},$$
$$\mathbf{P}^{(\lambda)}_{-1/2}=i^{3/2}C_{2x}P^{1/2}-\frac{i}{2\sqrt{2}}P^{-1/2}(\lambda C^{-}_{31}-\lambda^{*}C^{+}_{31})P^{1/2}. \tag{32}$$

The irreducible matrices of C_{2x}, C^{+}_{31} and C^{-}_{31} in (26) are changed into

$$D^{(\lambda)}(C^{+}_{31})=\tilde{D}^{(\lambda)*}(C^{-}_{31})=\frac{\lambda}{2\sqrt{2}}\begin{pmatrix}\sqrt{i^{*}} & -i\\ -i & \sqrt{i}\end{pmatrix} \tag{33a}$$

$$D^{(\lambda)}(C_{2x})=-\tilde{D}^{(\lambda)}(C_{2y})=\begin{pmatrix}0, & \sqrt{i}\\ i\sqrt{i}, & 0\end{pmatrix}^{*}. \tag{33b}$$

From (31) and (33a) it is seen that the irreducible matrices of C^{+}_{31} and C^{-}_{31} are symmetric now.

Applying the generalized ISO's to appropriate trial states, we can obtain the CG coefficients easily.

A. s. v.×s. v.→s. v. reps

(1) $A_{\mu_1}\times A_{\mu_2}$ remains the same as (21a).

(2) $A_{\mu}\times F=F$.

Applying (30) to the trial state $|A_{\mu},F3\rangle$ with $\nu\nu'=--$ and using (31), we obtain

$$\Psi^{(F)}_{1}=\cos\left(\frac{2\mu\pi}{3}\right)|A_{\mu},F1\rangle+\sin\left(\frac{2\mu\pi}{3}\right)|A_{\mu},F2\rangle,$$
$$\Psi^{(F)}_{2}=-\sin\left(\frac{2\mu\pi}{3}\right)|A_{\mu},F1\rangle+\cos\left(\frac{2\mu\pi}{3}\right)|A_{\mu},F2\rangle,\qquad \Psi^{(F)}_{3}=|A_{\mu},F3\rangle. \tag{34a}$$

(3) $F\times F=A_0+A_1+A_{-1}+2F$.

Applying $\mathbf{P}^{(A_\mu)}$ in (24a) to the trial states $|33\rangle\equiv|F3,F3\rangle$ with $\nu\nu'=++$, we obtain

$$\Psi^{(A_\mu)}=\frac{1}{\sqrt{3}}[|33\rangle-\Re e(i\rho_\mu)(|11\rangle-|22\rangle)-\Re e(\rho_\mu)(|12\rangle+|21\rangle)]. \tag{34b}$$

From (31) we have

$$C_{2x}|F1\rangle=i|F2\rangle,\quad C_{2x}|F2\rangle=-i|F1\rangle,$$
$$C_{2y}|F1\rangle=-i|F2\rangle,\quad C_{2y}|F2\rangle=i|F1\rangle. \tag{34c}$$

From (34c) we know that both the states $|12\rangle-|21\rangle$ and $|11\rangle+|22\rangle$ have $\nu\nu'=--$. Applying (30) to $|11\rangle+|22\rangle$ and $|12\rangle-|21\rangle$, we obtain

$$\Psi^{(F)0}_{1}=-\frac{1}{\sqrt{2}}[|23\rangle+|32\rangle],\quad \Psi^{(F)1}_{1}=\frac{1}{\sqrt{2}}[|13\rangle-|31\rangle],$$
$$\Psi^{(F)0}_{2}=-\frac{1}{\sqrt{2}}[|13\rangle+|31\rangle],\quad \Psi^{(F)1}_{2}=\frac{1}{\sqrt{2}}[|32\rangle-|23\rangle], \tag{34d}$$
$$\Psi^{(F)0}_{3}=\frac{1}{\sqrt{2}}[|11\rangle+|22\rangle],\quad \Psi^{(F)1}_{3}=\frac{1}{\sqrt{2}}[|12\rangle-|21\rangle].$$

TABLE V. The CG coefficients of $T \supset C_2$.

$F \times F$

$$\Psi^{A_1} = \frac{1}{\sqrt{3}}[|33\rangle - \frac{\sqrt{3}}{2}(|11\rangle - |22\rangle) + \tfrac{1}{2}(|12\rangle + |21\rangle)],$$

$$\Psi^{A_0} = \frac{1}{\sqrt{3}}[|33\rangle - |12\rangle - |21\rangle],$$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{3}}[|33\rangle + \frac{\sqrt{3}}{2}(|11\rangle - |22\rangle) + \tfrac{1}{2}(|12\rangle + |21\rangle)],$$

$F \times E_{1/2}$

$$\Psi^{(E_{1/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle - \sqrt{2}|1-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{1/2})}_{-1/2} = \frac{1}{\sqrt{3}}[-|3-\tfrac{1}{2}\rangle + \sqrt{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{3/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{-3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{-3/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$F \times E_{3/2}$

$$\Psi^{(E_{1/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{1/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle - \sqrt{2}|1-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{3/2})}_{-1/2} = \frac{1}{\sqrt{3}}[-|3-\tfrac{1}{2}\rangle + \sqrt{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{-3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{-3/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$F \times E_{-3/2}$

$$\Psi^{(E_{1/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{1/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|1-\tfrac{1}{2}\rangle - \frac{\sqrt{6}}{2}|2-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{3/2})}_{-1/2} = -\frac{1}{\sqrt{3}}[|3-\tfrac{1}{2}\rangle + \frac{\sqrt{6}}{2}|1\,\tfrac{1}{2}\rangle + \frac{\sqrt{2}}{2}|2\,\tfrac{1}{2}\rangle]$$

$$\Psi^{(E_{-3/2})}_{1/2} = \frac{1}{\sqrt{3}}[|3\,\tfrac{1}{2}\rangle - \sqrt{2}|1-\tfrac{1}{2}\rangle],$$

$$\Psi^{(E_{-3/2})}_{-1/2} = \frac{1}{\sqrt{3}}[-|3-\tfrac{1}{2}\rangle + \sqrt{2}|2\,\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{1/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = |\tfrac{1}{2}\,\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = |-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = -\tfrac{1}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle + \frac{\sqrt{3}}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = -\frac{\sqrt{3}}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \tfrac{1}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

$E_{1/2} \times E_{-3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = -\tfrac{1}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \frac{\sqrt{3}}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = \frac{\sqrt{3}}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \tfrac{1}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

$E_{3/2} \times E_{3/2}$

$$\Psi^{A_{-1}} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = -\tfrac{1}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \frac{\sqrt{3}}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = \frac{\sqrt{3}}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \tfrac{1}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

$E_{3/2} \times E_{-3/2}$

$$\Psi^{A_0} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = |\tfrac{1}{2}\,\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = |-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

$E_{-3/2} \times E_{-3/2}$

$$\Psi^{A_1} = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle - |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle],$$

$$\Psi^{(F)}_1 = -\tfrac{1}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle + \frac{\sqrt{3}}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle$$

$$\Psi^{(F)}_2 = -\frac{\sqrt{3}}{2}|\tfrac{1}{2}\,\tfrac{1}{2}\rangle - \tfrac{1}{2}|-\tfrac{1}{2}-\tfrac{1}{2}\rangle,$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}[|\tfrac{1}{2}-\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\,\tfrac{1}{2}\rangle]$$

B. s. v.×d. v.→d. v. reps

(1) $A_\mu \times E_\nu$ is the same as Eq. (29a).
(2) $F \times E_\nu = \Sigma_{\nu'} E_{\nu'}$, or $F \times (\lambda) = \Sigma_{\lambda'}(\lambda')$.
Applying (32) to the initial state $|F3,\lambda\,\frac{1}{2}\rangle$ we obtain,

$$\Psi^{(\lambda')}_{1/2} = \sqrt{\frac{1}{3}}\left|F3,\lambda\ \frac{1}{2}\right\rangle + \sqrt{\frac{2}{3}}\left[\Re e(-\lambda'\lambda^*/4)\left|F1,\lambda-\frac{1}{2}\right\rangle + \Re e(i\lambda'\lambda^*/4)\left|F2,\lambda-\frac{1}{2}\right\rangle\right],$$

$$\Psi^{(\lambda')}_{-1/2} = -\sqrt{\frac{1}{3}}\left|F3,\lambda-\frac{1}{2}\right\rangle + \sqrt{\frac{2}{3}}\left[\Re e(i\lambda'\lambda^*/4)\left|F1,\lambda\ \frac{1}{2}\right\rangle + \Re e(\lambda'\lambda^*/4)\left|F2,\lambda\ \frac{1}{2}\right\rangle\right]. \tag{35a}$$

C. d. v.×d. v.→s. v. reps

From (33a) we again have (29c). Similar to the case of $T \supset D_2 \supset C_2$ we have

$$\Psi^{(F)}_1 = \Re e(\lambda_1\lambda_2/4)\left|\lambda_1\ \frac{1}{2},\lambda_2\ \frac{1}{2}\right\rangle + \Re e(i\lambda_1\lambda_2/4)\left|\lambda_1-\frac{1}{2},\lambda_2-\frac{1}{2}\right\rangle,$$

$$\Psi^{(F)}_2 = \Im m(\lambda_1\lambda_2/4)\left|\lambda_1\ \frac{1}{2},\lambda_2\ \frac{1}{2}\right\rangle + \Im m(i\lambda_1\lambda_2/4)\left|\lambda_1-\frac{1}{2},\lambda_2-\frac{1}{2}\right\rangle, \tag{35b}$$

$$\Psi^{(F)}_3 = \frac{1}{\sqrt{2}}\left[\left|\lambda_1\ \frac{1}{2},\lambda_2-\frac{1}{2}\right\rangle + \left|\lambda_1-\frac{1}{2},\lambda_2\ \frac{1}{2}\right\rangle\right],$$

and $\Psi^{(A_\mu)}$ is identical to the one in Eq. (29d). From (34) and (35) we see that all the CG coefficients are real, and the numerical results for (34b) and (35) are listed in the Table V.

VI. SUMMARY

Using the effective ISO's or the factored ISO's the algebraic expressions of the CG coefficients, (21) and (22), (28) and (29), (34) and (35), for the group chain $T \supset C_3$, $T \supset D_2 \supset C_2$, $T \supset C_2$ are derived for the first time. Although they are not as elegant as the analytic expression for the $SO_3 \supset SO_2$ CG coefficients, their compactness, especially for the d. v. × d. v. CG coefficients, is remarkable. The projection operator method is often quoted as extremely difficult for large-order groups. From the sequent papers[12,13] and the present one, it is seen that this ''accusation'' is not totally fair. With the effective ISO's the projection operator method becomes extremely simple and powerful, which enable us to find algebraic expressions for the SAF's and CG coefficients. *Therefore it is the time for a rehabilitation of the projection operator method.*

ACKNOWLEDGMENTS

This work was supported in part by the Fundamental Research Fund of State Science and Technology Committee of China and the Research Fund of the National Educational Committee of China.

[1] M. A. Melvin, Rev. Mod. Phys. **28**, 18 (1956).
[2] J. S. Griffith, *Theory of Transition Metal Ions* (Cambridge University, Cambridge, England, 1961).
[3] G. F. Koster, J. O. Dimmock, R. G. Wheeler, and H. Statz, *Properties of the Thirty-Two Point Groups* (MIT, Cambridge, MA, 1963).
[4] R. McWeeny, *Symmetry—An Introduction to Group Theory and Its Applications* (Pergamon, New York, 1963).
[5] H. V. McIntosh, J. Mol. Spectrosc. **5**, 269 (1960); **10**, 51 (1963).
[6] C. J. Bradley and A. P. Cracknell, *Symmetry in Solids* (Oxford University Press, London, 1972).
[7] P. H. Butler, *Point Group Symmetry Applications: Methods and Table* (Plenum, New York, 1981).
[8] J. Q. Chen, *Group Representation Theory for Physicists* (World Scientific, Singapore, 1989).
[9] W. G. Harter, *Principle of Symmetry, Dynamics and Spectroscopy* (Wiley, New York, 1993).
[10] S. L. Altmann, *Rotations, Quaternions and Double Groups* (Clarendon, Oxford, 1986).
[11] S. L. Altmann and P. Herzig, *Point Group Theory Tables* (Clarendon, Oxford, 1994).
[12] J. Q. Chen and P. D. Fan, J. Math. Phys. **39**, 467 (1998).

[13] J. Q. Chen and P. D. Fan, J. Math. Phys. **39**, 5502 (1998).
[14] P. H. Butler and B. G. Wybourne, Int. J. Quantum Chem. **10**, 615 (1976).
[15] R. M. Golding and J. D. Newmarch, Mol. Phys. **33**, 1301 (1977).
[16] T. Damhus, S. E. Harnung, and C. E. Schaffer, Theor. Chim. Acta **65**, 405 (1984).
[17] J. L. Ping, Q. R. Zheng, B. Q. Chen, and J. Q. Chen, Comput. Phys. Commun. **52**, 355 (1989); J. L. Ping and J. Q. Chen, Commun. Comput. Phys. (submitted). The code ''PGIR'' is available upon request.
[18] M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw-Hill, New York, 1964), p. 41.
[19] J. Q. Chen, M. J. Gao, and G. Q. Ma, Rev. Mod. Phys. **57**, 211 (1985).
[20] R. M. Golding and J. D. Newmarch, Mol. Phys. **21**, 157 (1971).
[21] R. M. Golding, Mol. Phys. **26**, 661 (1973)
[22] Q. Zhang and X. Z. Li, Int. J. Quantum Chem. **29**, 293 (1986).
[23] R. Dirl, J. Math. Phys. **20**, 659 (1979).
[24] J. Q. Chen, P. D. Fan, L. McAven, and P. H. Butler (preprint).
[25] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[26] P. H. Butler, Aust. J. Phys. **32**, 137 (1979).

Addendum: Overcoming dissipative distortions by a waveform restorer and designer [J. Math. Phys. 36, 2324 (1995)]

S. K. Foong[a)]
Department of Natural Sciences, Faculty of Education, University of the Ryukyus, Okinawa 903-0213, Japan

(Received 27 May 1998; accepted for publication 5 June 1998)

If observation is restricted to just a single point in space, then Kac's path integral solution of the telegrapher equation, a wave equation with a damping term, gives the same solution for an infinite number of inputs consisting of the solutions of the wave equation without damping. While this was implicit in a previous publication by the author [J. Math. Phys. **36**, 2324 (1995)], it is made explicit here along with examples. © *1998 American Institute of Physics.* [S0022-2488(98)04610-6]

In Ref. 1, I gave an algorithm for a waveform restorer and a designer motivated by Kac's path integral solution of the telegrapher equation, a wave equation with a damping term. This addendum is motivated by comments and questions I received after the paper was published. To facilitate the discussion, recall that the (generalized) Kac's solution to the following telegrapher type equation with completely general initial conditions:

$$\frac{1}{v^2}\frac{\partial^2 F}{\partial t^2}+\frac{2a}{v^2}\frac{\partial F}{\partial t}-O_{\mathbf{x}}F=0, \quad t\geqslant 0,$$
$$F(\mathbf{x},0)=\kappa(\mathbf{x}), \quad \frac{\partial F}{\partial t}(\mathbf{x},0)=\xi(\mathbf{x}), \tag{1}$$

is given by an average over the history of an undamped wave ϕ at the point

$$F(\mathbf{x},t)=\int_{-t}^{t}\phi(\mathbf{x},r)g(t,r)dr, \quad t\geqslant 0. \tag{2}$$

In Eq. (1), a is assumed to be a positive constant, and $O_{\mathbf{x}}$ is *any* linear and spatial operator such as the Laplacian ∇^2. The undamped wave $\phi(\mathbf{x},t)$ in Eq. (2) satisfies, for all time,

$$\frac{1}{v^2}\frac{\partial^2 \phi}{\partial t^2}-O_{\mathbf{x}}\phi=0,$$
$$\phi(\mathbf{x},0)=\kappa(\mathbf{x}), \quad \frac{\partial \phi}{\partial t}(\mathbf{x},0)=\xi(\mathbf{x}). \tag{3}$$

$\phi(\mathbf{x},t)$ was called the would-have-been wave since if a vanishes, then $F(\mathbf{x},t)=\phi(\mathbf{x},t)$ for $t\geqslant 0$. $g(t,r)$, parametrized by t, denotes the distribution of Kac's randomized time $S(t)$, defined by

$$S(t)=\int_0^t(-1)^{N(\tau)}\,d\tau, \tag{4}$$

where $N(t)$, with $N(0)=0$, is a random variable Poisson distributed with intensity a:

[a)]Electronic-mail: foong@edu.u-ryukyu.ac.jp

 © 1998 American Institute of Physics

$$P(N(t)-N(t')=k)=\begin{cases} e^{-a(t-t')}(a(t-t'))^k/k!, & k\geqslant 0; \\ 0, & k<0. \end{cases} \tag{5}$$

The value of $S(t)$ is denoted by r. In Eq. (2), there is no averaging of ϕ over space, and while F and ϕ are the waveforms at the same arbitrary point $\mathbf{x}$ in space, the argument t in F is assumed positive but in ϕ t may assume both positive and negative values.

The density distribution $g(t,r)$ is given by, with $u\equiv\sqrt{t^2-r^2}$,

$$g(t,r)=e^{-at}\delta(t-r)+\frac{ae^{-at}}{2}\theta(t-|r|)\left[I_0(au)+\frac{t+r}{u}I_1(au)\right]. \tag{6}$$

Since g is known, Eq. (2) may be paraphrased as the following: Knowing ϕ yields F. It is natural to ask if the converse, knowing F yields ϕ, is also true. Equation (2) may be viewed as a matrix multiplication, and so the question amounts to asking whether a left inverse g^{-1} of g exists, such that

$$\int_0^t d\tau g^{-1}(t,\tau)g(\tau,r)=\delta(t-r), \tag{7}$$

which would evidently enable the following restoration of ϕ from F:

$$\int_0^t d\tau g^{-1}(t,\tau)F(\mathbf{x},\tau)=\int_\infty^\infty dr\,\phi(\mathbf{x},r)\int_0^t d\tau g^{-1}(t,\tau)g(\tau,r)=\phi(\mathbf{x},t). \tag{8}$$

Strong evidence that the answer to the above question is negative is provided by the fact that when the time t is discretized, the discretized version of $g(t,r)$ is a rectangular matrix (g_{ij}). This can be seen from Eq. (4): with $t=m\Delta t$, r takes a value from $-(m-1)\Delta t$ to $m\Delta t$, i.e., $0\leqslant i\leqslant m$ and $-(m-1)\leqslant j\leqslant m$, and so (g_{ij}) is an $m\times(2m-1)$ matrix. Since a $p\times q$ matrix has a left inverse if and only if the rank of the matrix is q, it is necessary that $p\geqslant q$, and hence (g_{ij}) does not have a left inverse, a fact that remains to be true as we let $\Delta t\to 0$, and it appears that the same would remain true for the limit $g(t,r)$.

An ansatz consisting of a power series for the inverse also fails. Let

$$g^{-1}(t,\tau)=e^{a\tau}\left\{\delta(t-\tau)+\theta(\tau)\,\theta(t-\tau)\sum_{k=1}^{\infty}c_k^i(t,\tau)a^k\right\}. \tag{9}$$

By expanding the modified Bessel function in powers of a, g takes the form

$$g(\tau,r)=e^{-a\tau}\left[\delta(\tau-r)+\theta(\tau-|r|)\frac{a}{2}+\cdots\right]. \tag{10}$$

Hence, up to order a, we have

$$\int_0^t g^{-1}(t,\tau)g(\tau,r)d\tau=\delta(t-r)+\theta(r)\,\theta(t-r)c_1(t,r)a+\theta(t-|r|)\frac{a}{2}+O(a^2). \tag{11}$$

On the right-hand side of Eq. (11), we require the coefficients of all orders of a to vanish. Hence, for the term first order in a, we have

$$\theta(r)\,\theta(t-r)c_1(t,r)+\theta(t-|r|)\tfrac{1}{2}=0, \tag{12}$$

which has no solution for $-t<r<0$. The generality of the ansatz, Eq. (9), is another strong hint pointing to the nonexistence of the inverse of g.

The nonuniqueness of $\phi(\mathbf{x},t)$ for $|t|\leqslant T$, given $F(\mathbf{x},t)$ for $0\leqslant t\leqslant T$ (T is some arbitrary time interval) at an arbitrary *fixed* point $\mathbf{x}=\mathbf{x}_f$, say, was implicit in Ref. 1. There are actually an infinite

number of possible undamped waves $\phi(\mathbf{x}_f,t)$, which, when substituted into Eq. (2), give the same observed $F(\mathbf{x}_f,t)$. This can be easily seen as follows. It was shown in Ref. 1 that $F(\mathbf{x},t)$ may be written as a sum of two components:

$$F(\mathbf{x},t)=F_1(\mathbf{x},t)+F_2(\mathbf{x},t), \tag{13}$$

where the F_i are given by

$$F_i(\mathbf{x},t)\equiv\int_0^t dr\,h_i(t,r)\,\phi_i(\mathbf{x},r), \quad i=1,2, \tag{14}$$

where $\phi_1(\mathbf{x},t)=[\phi(\mathbf{x},t)+\phi(\mathbf{x},-t)]/2$ and $\phi_2(\mathbf{x},t)=[\phi(\mathbf{x},t)-\phi(\mathbf{x},-t)]/2$, namely the symmetric part and the antisymmetric part, with respect to t, of ϕ, respectively, and $h_1(t,r)\equiv g(t,r)+g(t,-r)$ and $h_2(t,r)\equiv g(t,r)-g(t,-r)$, the symmetric and antisymmetric parts (up to a factor of 2), with respect to r, of $g(t,r)$. These h_i have left inverses h_i^{-1} expressible in closed form:

$$h_1^{-1}(t,\tau)=e^{a\tau}\left\{\delta(t-\tau)-\theta(t-\tau)\left[ae^{a\tau}-\frac{2}{\pi}\int_0^a dy\,\frac{\sin ty}{y}[w(y)\cosh(\tau w(y))+a\,\sinh(\tau w(y))]\right]\right\}, \tag{15}$$

where $w(y)\equiv\sqrt{a^2-y^2}$, and

$$h_2^{-1}(t,\tau)=e^{a\tau}\left[\delta(t-\tau)-\theta(t-\tau)\,\frac{a\tau}{\sqrt{t^2-\tau^2}}J_1(a\sqrt{t^2-\tau^2})\right], \tag{16}$$

satisfying, for each value of i,

$$\int_r^t d\tau\,h_i^{-1}(t,\tau)h_i(\tau,r)=\delta(t,r), \quad r>0. \tag{17}$$

Hence, the ϕ can be restored from the F_i as follows: From Eq. (14), we obtain the ϕ_i by multiplying the observed F_i by h_i^{-1} and performing an integration—

$$\phi_i(\mathbf{x},t)=\int_0^t d\tau\,h_i^{-1}(t,\tau)F_i(\mathbf{x},\tau), \tag{18}$$

and then taking the linear combination of ϕ_i—

$$\begin{aligned}\phi(\mathbf{x},t)&=\phi_1(\mathbf{x},t)+\phi_2(\mathbf{x},t),\\ \phi(\mathbf{x},-t)&=\phi_1(\mathbf{x},t)-\phi_2(\mathbf{x},t),\end{aligned} \tag{19}$$

for $t>0$.

However, there are infinite choices for the pairs $F_i(\mathbf{x},t)$ for $0\leqslant t\leqslant T$, and each pair F_i gives rise to a unique pair of ϕ_i via Eq. (18), yielding a different $\phi(\mathbf{x},t)$ for $|t|\leqslant T$. All these different ϕ at $\mathbf{x}_f$ would yield the same F at $\mathbf{x}_f$. Hence, knowing F yields ϕ is false, and the left inverse of g does not exist.

We now consider an explicit example:

$$F(t)=\begin{cases}1, & \text{for } 0\leqslant t\leqslant 1,\\ 0, & \text{for } t\geqslant 1,\end{cases} \tag{20}$$

an observed pulse of unit height lasting for a unit time. Since the validity of the following discussion is independent of $\mathbf{x}_f$, it will be suppressed. Let us consider the following two choices out of an infinite possible choices for F_i: (1) $F_1(t)=F(t)$ and $F_2(t)=0$, and (2) $F_1(t)=0$ and $F_2(t)=F(t)$.

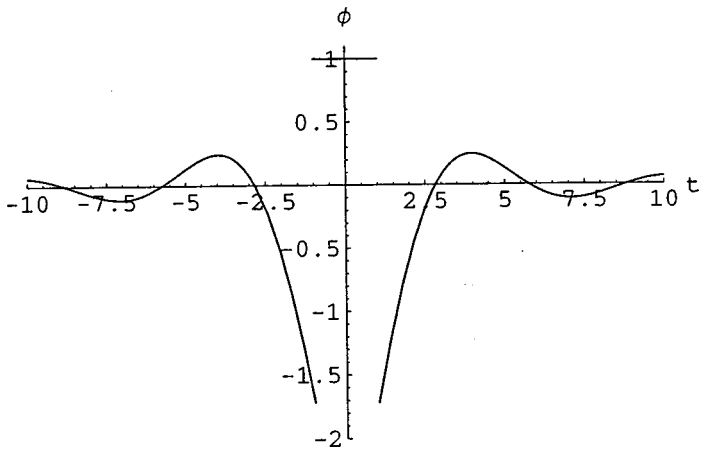

FIG. 1. A $\phi(t)$ that would give the observed pulse F of unit height lasting for a unit time of Eq. (20).

Choice (1). It follows from Eq. (18) that

$$\phi_2(t)=0, \tag{21}$$

and

$$\phi_1(t)=\begin{cases} \int_0^t d\tau h_1^{-1}(t,\tau), & \text{for } 0\leqslant t\leqslant 1, \\ \int_0^1 d\tau h_1^{-1}(t,\tau), & \text{for } t\geqslant 1. \end{cases} \tag{22}$$

The integral in Eq. (22) corresponding to $0\leqslant t\leqslant 1$ can be done explicitly with the help of Ref. 2 (item 2.481) to yield

$$\phi_1(t)=\frac{1}{2}+e^{at}-\frac{1}{2}e^{2at}+\frac{2}{\pi}e^{at}\int_0^a dy\,\frac{\sin ty}{y}\sinh(t\sqrt{a^2-y^2}). \tag{23}$$

The integral in Eq. (23) is of the form $I(\alpha)=\int_0^\alpha d\beta\,\sin\beta\sinh(\sqrt{\alpha^2-\beta^2})/\beta$. Letting $\beta=\alpha\sin\theta$, the integral becomes

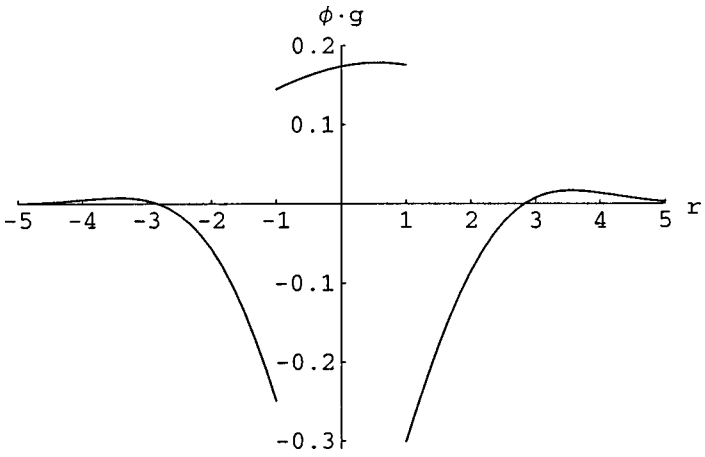

FIG. 2. This graph shows the contribution to F of the $\phi(t)$, $|t|<5$, given in Fig. 1.

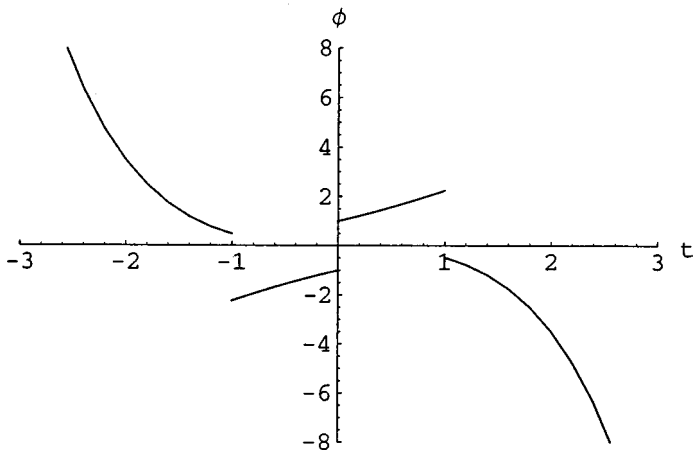

FIG. 3. Another $\phi(t)$ that would give the observed pulse F of unit height lasting for a unit time of Eq. (20).

$$I(\alpha)=\int_0^{\pi/2}\cot\,\theta\,\sin(\alpha\,\sin\,\theta)\sinh(\alpha\,\cos\,\theta)d\theta=\frac{\pi}{2}(\cosh\,\alpha-1), \tag{24}$$

which may be evaluated with the help of Ref. 2 (items 3.933 and 3.934). Hence, we have

$$\phi_1(\mathbf{x},t)=1, \quad \text{for } 0\leqslant t\leqslant 1, \tag{25}$$

and it follows from Eqs. (19) and (21) that

$$\phi(\mathbf{x},t)=1, \quad \text{for } -1\leqslant t\leqslant 1. \tag{26}$$

This ϕ clearly yields $F(t)=1$ for $0\leqslant t\leqslant 1$ upon application of Eq. (2).

For $|t|>1$, Eq. (22) yields

$$\phi(t)=-\frac{1}{2}(e^{2a}-1)+\frac{2}{\pi}e^{a}\int_0^{a}dy\,\frac{\sin|t|y}{y}\,\sinh(\sqrt{a^2-y^2}). \tag{27}$$

By using the software MATHEMATICA, $\phi(t)$ with $a=1$ is plotted in Fig. 1, and $\phi(r)g(t,r)$, without considering the contribution from the delta function in g, is plotted for $t=5$ in Fig. 2. The graph shows the contribution to F of ϕ at times $|t|<5$. The contribution to $F(t)$ coming from $e^{-at}\int\phi(r)\,\delta(t-r)dr=e^{-at}\phi(t)$ at $t=5$ amounts to 8.863×10^{-4}, which when added to the area under the graph in Fig. 2 gives $F(t=5)=0$, as required by Eq. (20).

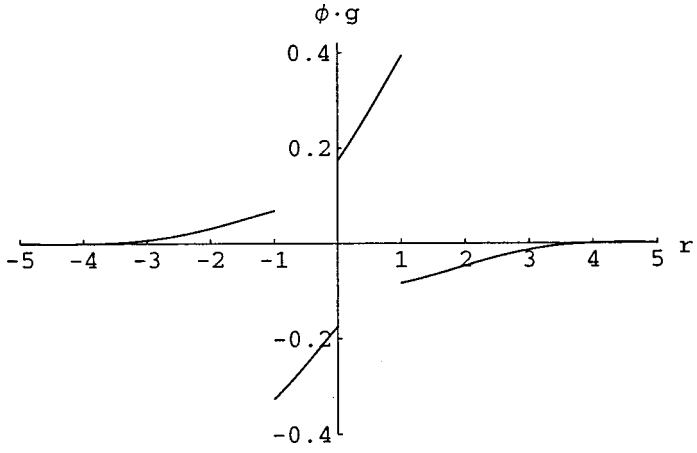

FIG. 4. This graph shows the contribution to F of the $\phi(t)$, $|t|<5$, given in Fig. 3.

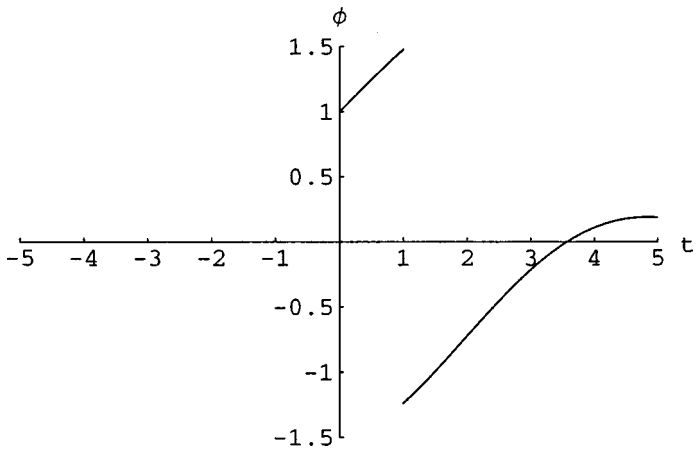

FIG. 5. A $\phi(t)$, where $\phi(t)=0$ for $t<0$, that would give the observed pulse F of Eq. (20).

Choice (2). It follows from Eq. (18) that

$$\phi_1(t)=0, \tag{28}$$

and hence

$$\phi(t)=\begin{cases} \int_0^t d\tau h_2^{-1}(t,\tau), & \text{for } 0\leqslant t\leqslant 1, \\ \int_0^1 d\tau h_2^{-1}(t,\tau), & \text{for } t\geqslant 1. \end{cases} \tag{29}$$

Substituting h_2^{-1} given by Eq. (16) into the above gives, with $\alpha\equiv at$,

$$\begin{aligned}\phi(t)&=e^{at}-a\int_0^t d\tau \frac{\tau e^{a\tau}}{\sqrt{t^2-\tau^2}} J_1(a\sqrt{t^2-\tau^2}) \\ &=e^{\alpha}-\alpha\int_0^{\pi/2} d\theta \sin\theta e^{\alpha\sin\theta} J_1(\alpha\cos\theta), \quad \text{for } 0\leqslant t\leqslant 1\end{aligned} \tag{30}$$

and

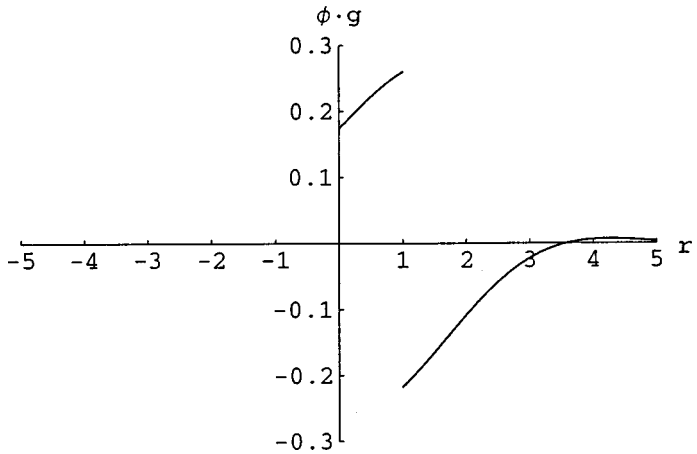

FIG. 6. This graph shows the contribution to F of the $\phi(t)$, $|t|<5$, given in Fig. 5.

$$\phi(t) = -a\int_0^1 d\tau \frac{\tau e^{a\tau}}{\sqrt{t^2-\tau^2}} J_1(a\sqrt{t^2-\tau^2}), \quad \text{for } t \geqslant 1. \tag{31}$$

Since $\phi_1 = 0$, $\phi(t)$ is antisymmetric with respect to t, $\phi(-t) = -\phi(t)$. Figure 3 is a plot of $\phi(t)$ with $a=1$. Figure 4 is a plot of $\phi(r)g(t,r)$, excluding the contribution from the delta function in g, for $t=5$.

Both ϕ, as plotted in Fig. 1 and Fig. 3, would give rise to the same observed pulse F of unit height lasting for a unit time of Eq. (20). Hence, we conclude that, if he knows only F at the point, he would not be able to obtain a unique would-have-been wave ϕ at the same point. This result is independent of the Kac's formalism which we used to derive it. We wish to emphasize that for a given Cauchy data, both the $F(x,t)$ and $\phi(x,t)$ are unique. However, it is assumed here that the observer (restorer) has available to him only the distorted waves F or F_i at one point in space only, and has absolutely no knowledge of their values elsewhere.

We summarize the discussion up to this point: Given only that two waves $F(\mathbf{x}_f,t)$, $0 \leqslant t \leqslant T$ and $\phi(\mathbf{x}_f,t)$, $-T \leqslant t \leqslant T$ at an arbitrary chosen point $\mathbf{x}_f$ share the *same* Cauchy data at $t=0$ (but the data are not given), propagate under Eq. (1) and Eq. (3), respectively, then $F(\mathbf{x}_f,t)$, $0 \leqslant t \leqslant T$ can be obtained from $\phi(\mathbf{x}_f,t)$, $-T \leqslant t \leqslant T$ via Eq. (2). However, it is impossible, without further information, to obtain $\phi(\mathbf{x}_f,t)$ for any value of t (except the trivial case of $t=0$) from $F(\mathbf{x}_f,t)$, $0 \leqslant t \leqslant T$. An example of the ''further information'' is the knowledge of the two components $F_1(x_f,t)$ and $F_2(x_f,t)$ of $F(x_f,t)$, which would allow the restoration of $\phi(\mathbf{x}_f,t)$, $-T \leqslant t \leqslant T$.

In closing, another example of the ''further information'' should be mentioned, namely, $\phi(\mathbf{x}_f,t)=0$ for $t<0$. In this case, g^{-1}, which satisfies

$$\int_0^t g^{-1}(t,\tau)g(\tau,r)\,d\tau = \delta(t-r), \quad r>0, \tag{32}$$

is conjectured to be[3,4]

$$g^{-1}(t,\tau) = e^{a\tau}\left\{\delta(t-\tau) - \theta(t-\tau)\left[\left(\frac{t-\tau}{t+\tau} + a\tau\right)\frac{J_1(aw)}{w} + \frac{a\tau}{t+\tau}J_o(aw)\right]\right\}, \quad t,\tau>0, \tag{33}$$

where $w=\sqrt{t^2-\tau^2}$. Hence, $\phi(\mathbf{x},t)$ can be restored as follows:

$$\phi(\mathbf{x},t) = \int_0^t g^{-1}(t,\tau)F(\mathbf{x},\tau)\,d\tau. \tag{34}$$

Given only $F(\mathbf{x}_f,t)$ and $\phi(\mathbf{x}_f,t)=0$ for $t<0$, without the knowledge of F_i, the restoration can be carried out.

For F of Eq. (20), $\phi(t)$ resulting from Eq. (34) with $a=1$ is plotted in Fig. 5, and $\phi(r)g(t,r)$, without considering the contribution from the delta function in g, is plotted for $t=5$ in Fig. 6.

ACKNOWLEDGMENT

I wish to thank Professor S. Matsumoto for a discussion which has improved the presentation of this paper.

[1] S. K. Foong, ''Overcoming dissipative distortions by a waveform restorer and designer,'' J. Math. Phys. **36**, 2324–2340 (1995).
[2] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, corrected and enlarged edition (Academic, New York, 1980).
[3] S. K. Foong, ''Functional integration and wave propagation,'' in *Functional Integration: Basics and Applications*, edited by C. DeWitt-Morette, P. Cartier, and A. Folacci (Plenum, New York, 1997), pp. 97–130.
[4] S. K. Foong and S. Kanno, ''A new method for restoration of telegraphic waveform,'' preprint, 1998.

Decomposition of time-ordered products and path-ordered exponentials

C. S. Lam[a)]
Department of Physics, McGill University,
3600 University Street, Montreal, QC H3A 2T8, Canada

(Received 29 April 1998; accepted for publication 21 May 1998)

We present a decomposition formula for U_n, an integral of time-ordered products of operators, in terms of sums of products of the more primitive quantities C_m, which are the integrals of time-ordered commutators of the same operators. The resulting factorization enables a summation over n to be carried out to yield an explicit expression for the time-ordered exponential, an expression which turns out to be an exponential function of C_m. The Campbell–Baker–Hausdorff formula and the non-Abelian eikonal formula obtained previously are both special cases of this result. © *1998 American Institute of Physics.* [S0022-2488(98)00610-0]

I. INTRODUCTION

The path-ordered exponential $U(T,T')=\mathscr{P}\exp(\int_{T'}^{T}H(t)dt)$ is the solution of the first order differential equation $dU(T,T')/dT=H(T)U(T,T')$ with the initial condition $U(T',T')=\mathbf{1}$. The function $H(T)$ can be a finite-dimensional matrix or an infinite-dimensional operator. In the latter case all concerns of domain and convergence will be ignored. The parameter t labeling the path shall be referred to as ''time,'' so path-ordering and time-ordering are synonymous in the present context. The path-ordered exponential is usually computed from its power series expansion, $U(T,T')=\Sigma_{n=0}^{\infty}U_n$, in terms of the time-ordered products $U_n=\mathscr{P}(\int_{T'}^{T}H(t)dt)^n/n!$.

Path-ordered exponential can be found in many areas of theoretical physics. It is the time-evolution operator in quantum mechanics if iH is the Hamiltonian. It is the nonintegrable phase factor (Wilson line) of the Yang–Mills theory if $-iH=A$ is the non-Abelian vector potential. It defines an element of a Lie group connected to the identity via a path whose tangent vector at time t is given by the Lie-algebra element $-iH(t)$. It can be used to discuss canonical transformation and classical particle trajectories.[1] It also gives rise to useful formulas in many other areas of mathematical physics by suitable choices of $H(t)$, some of which will be mentioned later. Likewise, time-ordered products are ubiquitous. For example, free-field matrix elements of time-ordered products of an interaction Hamiltonian give rise to perturbation diagrams.

The main result of this paper is a *decomposition theorem* for the time-ordered products U_n. It provides a formula equating them to sums of products of the more primitive quantities C_m. The factorization thus achieved enables U_n to be summed up and $U(T,T')$ expressed as an exponential function of the C_m's.

This result grew out of a non-Abelian eikonal formula. The Abelian version of the eikonal formula[2] is well known. It gives rise to a geometrical interpretation for elastic scattering at high energies,[3] besides being a useful tool in dealing with infrared divergence[4] in QED. The non-Abelian version was originally developed to deal with the consistency problem of baryonic amplitudes in large-N_c QCD,[5,6] but it proves to be useful also in understanding gluon reggeization and reggeon factorization.[7,8] It is expected to be valuable also in the discussion of geometrical pictures and infrared problems in QCD.

The generalization in this paper can be used to take the non-Abelian eikonal formula a step further to include small transverse motions of the high-energy trajectory. Such corrections to the usual eikonal formula are known to be crucial in obtaining the Landau–Pomeranchuk–Migdal effect[9] in QED and so are expected to be important as well in the case of QCD. However, we shall postpone all such physical applications and concentrate in this paper on the development of mathematical formulas.

[a)]Electronic mail: Lam@physics.mcgill.ca

 © 1998 American Institute of Physics

On the mathematical side, one can derive the Campbell–Baker–Hausdorff formula[10] as a corollary of the formulas developed in these pages.

Statements, explanations, and simple illustrations of the results will be found in the main text, while proofs and other details will be relegated to the Appendices.

The formulas are combinatorial in character. They cannot be adequately explained without a suitable set of notations, which we develop in Sec. II. The main result of the paper, the decomposition theorem, will be discussed in Secs. III and IV. This theorem can be stated for a more general time-ordered product U_n, in which all the $H_i(t)$ are different. This will be dealt with in Sec. III. The special case when these $H_i(t)=H(t)$ are identical will be taken up in Sec. IV. This in turn leads to the exponential formula for $U(T,T')$ in Sec. V. The final section, Sec. VI, is put in mainly to illustrate the versatility of our results in deriving other formulas useful in mathematics and physics, by suitable choices of $H_i(t)$. Among them are the Campbell–Baker–Hausdorff formula for the multiplication of group elements, and the non-Abelian generalization of the eikonal formula in physics.

II. DEFINITIONS

We start by generalizing the definition of U_n to cases when the operators $H(t)$ are all different.

Let $[s]=[s_1s_2\cdots s_n]$ be a permutation of the n numbers $[12\cdots n]$, and S_n the corresponding permutation group. We define the *time-ordered product* $U[s]$ to be the integral

$$U[s]=U[s_1s_2\cdots s_n]=\int_{R[s]}dt_1dt_2\cdots dt_nH_{s_1}(t_{s_1})H_{s_2}(t_{s_2})\cdots H_{s_n}(t_{s_n}), \tag{2.1}$$

taken over the hyper-triangular region $R[s]=\{T\geqslant t_{s_1}\geqslant t_{s_2}\geqslant\cdots\geqslant t_{s_n}\geqslant T'\}$, with the operator $H_{s_i}(t_{s_i})$ standing to the left of $H_{s_j}(t_{s_j})$ if $t_{s_i}>t_{s_j}$. The operator $U_n=U_n[T,T']$ is then defined to be the average of $U[s]$ over all permutations $s\in S_n$:

$$U_n=\frac{1}{n!}\sum_{s\in S_n}U[s]. \tag{2.2}$$

In the special case when all $H_i(t)=H(t)$ are identical, this U_n coincides with the one in Sec. I.

The *decomposition theorem* expresses U_n as sums of products of the *time-ordered commutators* $C[s]=C[s_1s_2\cdots s_n]$. These are defined analogous to $U[s_1s_2\cdots s_n]$, but with the products of H_i's changed to nested multiple commutators:

$$C[s]=\int_{R[s]}dt_1dt_2\cdots dt_n[H_{s_1}(t_{s_1}),[H_{s_2}(t_{s_2}),[\cdots,[H_{s_{n-1}}(t_{s_{n-1}}),H_{s_n}(t_{s_n})]\cdots]]]. \tag{2.3}$$

For $n=1$, we let $C[s_i]=U[s_i]$ by definition. Similarly, the operator $C_n=C_n[T,T']$ is defined to be the average of $C[s]$ over all permutations $s\in S_n$:

$$C_n=\frac{1}{n!}\sum_{s\in S_n}C[s]. \tag{2.4}$$

It is convenient to use a ''*cut*'' (a vertical bar) to denote products of $C[\cdots]$'s. For example, $C[31|2]\equiv C[31]C[2]$, and $C[71|564|2|3]\equiv C[71]C[564]C[2]C[3]$.

We are now in a position to state the main theorem.

III. GENERAL DECOMPOSITION THEOREM

$$n!U_n=\sum_{s\in S_n}C[s_1\cdots|\cdots|\cdots\qquad\cdots s_n]\equiv\sum_{s\in S_n}C[s]_c. \tag{3.1}$$

A cut (vertical bar) is placed after s_i if and only if $s_j>s_i$ for all $j>i$. An element $s\in S_n$ with cuts placed this way will be denoted by $[s]_c$.

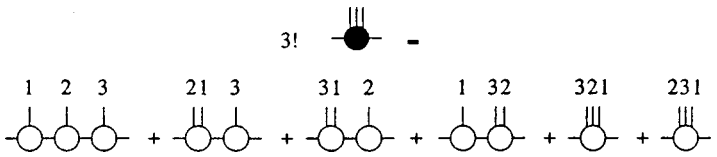

FIG. 1. The decomposition of $3!U_3$ (filled circle) in terms of $C[s]$'s.

In view of the fundamental nature of this theorem, two separate proofs shall be provided for it in Appendix A.

A. Examples

For illustrative purposes here are explicit formulas for $n=1,2,3,$ and 4:

$$1!U_1=C[1], \qquad 2!U_2=C[1|2]+C[21],$$
$$3!U_3=C[1|2|3]+C[21|3]+C[31|2]+C[1|32]+C[321]+C[231], \tag{3.2}$$

$$\begin{aligned}4!U_4=&C[1|2|3|4]+C[321|4]+C[231|4]+C[421|3]+C[241|3]+C[431|2]+C[341|2]\\&+C[1|432]+C[1|342]+C[21|43]+C[31|42]+C[41|32]+C[21|3|4]+C[31|2|4]\\&+C[41|2|3]+C[1|32|4]+C[1|42|3]+C[1|2|43]+C[4321]+C[3421]+C[4231]\\&+C[3241]+C[2341]+C[2431].\end{aligned}$$

The formula (3.1) can be displayed graphically, with a filled circle with n lines on top indicating U_n, and an open circle with n tops on top indicating $C[s]$. This is shown in Fig. 1 for $n=3$.

B. Factorization

In the special case when all the $H_i(t)$'s mutually commute, the only surviving $C[\cdots]$'s are those with one argument, so Eq. (3.1) reduces to a *factorization theorem*:

$$U_n=\frac{1}{n!}\prod_{i=1}^{n} C[i]. \tag{3.3}$$

Thus the general decomposition theorem in (3.1) may be thought of as a *non-Abelian generalization of this factorization theorem.*

IV. SPECIAL DECOMPOSITION THEOREM

Great simplification occurs when all $H_i(t)=H(t)$ are identical, for then $U[s]$ and $C[s]$ depend only on n but not on the particular $s\in S_n$. Hence all $U[s]=U_n$ and all $C[s]=C_n$. In this case the decomposition theorem becomes

$$U_n=\sum \xi(m_1m_2\cdots m_k)C_{m_1}C_{m_2}\cdots C_{m_k},$$
$$\xi(m)\equiv\xi(m_1m_2\cdots m_k)=\prod_{i=1}^{k}\left[\sum_{j=i}^{k} m_j\right]^{-1}. \tag{4.1}$$

The sum in the first equation is taken over all k, and all $m_i>0$ such that $\Sigma_{i=1}^{k} m_i=n$. Note that $\xi(m)=\xi(m_1m_2\cdots m_k)$ is *not* symmetric under the interchange of the m_i's. It is this asymmetry that produces the commutator terms in the formulas for K_n in Eq. (5.4).

See Appendix B for a proof of this special decomposition theorem.
We list below this special decomposition up to $n=5$:

$$1!U_1=C_1, \qquad 2!U_2=C_1^2+C_2,$$

$$3!U_3=C_1^3+2C_2C_1+C_1C_2+2C_3, \tag{4.2}$$

$$4!U_4=C_1^4+6C_3C_1+2C_1C_3+3C_2^2+3C_2C_1^2+2C_1C_2C_1+C_1^2C_2+6C_4,$$

$$5!U_5=C_1^5+24C_4C_1+6C_1C_4+12C_3C_2+8C_2C_3+12C_3C_1^2+6C_1C_3C_1+2C_1^2C_3+8C_2^2C_1$$
$$+4C_2C_1C_2+3C_1C_2^2+4C_2C_1^3+3C_1C_2C_1^2+2C_1^2C_2C_1+C_1^3C_2.$$

The case for $n=3$ is explicitly shown in Fig. 2.

V. EXPONENTIAL FORMULA FOR $U(T,T')$

The factorization character in (4.1) and (4.2) suggests that it may be possible to sum up the power series U_n to yield an explicit exponential function of the C_n's. This is indeed the case.

A. Commutative C_i's

First assume all the C_{m_i} in (4.1) commute with one another. Then, as will be shown in Appendix C,

$$U[T,T']=1+\sum_{n=1}^{\infty} U_n=\prod_{j=1}^{\infty}\sum_{m=0}^{\infty}\frac{1}{j^m m!}C_j^m=\exp\left[\sum_{j=1}^{\infty}\frac{C_j}{j}\right]. \tag{5.1}$$

Explicit calculation of low-order terms can also be obtained from (4.2) for further verification. This yields

$$U_0=1, \qquad U_1=C_1, \qquad U_2=\tfrac{1}{2}(C_1^2+C_2), \qquad U_3=\frac{1}{3!}C_1^3+\frac{1}{2}C_1C_2+\frac{1}{3}C_3,$$

$$U_4=\frac{1}{4!}C_1^4+\frac{1}{3}C_1C_3+\frac{1}{8}C_2^2+\frac{1}{4}C_1^2C_2+\frac{1}{4}C_4, \tag{5.2}$$

$$U_5=\frac{1}{5!}C_1^5+\frac{1}{4}C_1C_4+\frac{1}{6}C_2C_3+\frac{1}{6}C_1^2C_3+\frac{1}{8}C_1C_2^2+\frac{1}{12}C_1^3C_2+\frac{1}{5}C_5,$$

$$\sum_{i=0}^{5} U_n=\exp\left[C_1+\frac{1}{2}C_2+\frac{1}{3}C_3+\frac{1}{4}C_4+\frac{1}{5}C_5\right]+R,$$

where R contains products of C's whose subscript indices add up to 6 or more.

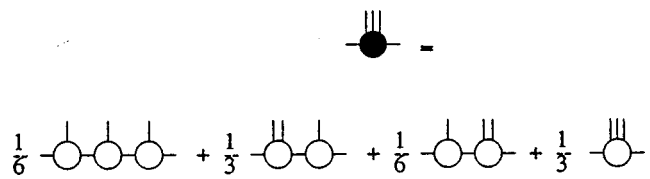

FIG. 2. Graphical form for the decomposition of U_3 (filled circle) in Eq. (4.2).

B. General exponential formula

In general the C_j's do not commute with one another so the exponent in (5.1) must be corrected by terms involving commutators of the C_j's. The exponent K can be computed by taking the logarithm of $U(T,T')$:

$$U(T,T')=1+\sum_{n=1}^{\infty} U_n=\exp[K]\equiv\exp\left[\sum_{i=1}^{\infty} K_i\right],$$
$$K=\ln\left[1+\sum_{n=1}^{\infty} U_n\right]=\sum_{\ell=1}^{\infty}\frac{(-)^{\ell-1}}{\ell}\left[\sum_{n=1}^{\infty} U_n\right]^{\ell}\equiv\sum_{n=1}^{\infty}\frac{C_n}{n}+\sum_{m}\eta(m)C_m, \tag{5.3}$$

in which the last sum is taken over *all* $m=(m_1m_2\cdots m_k)$ for $k\geqslant 2$, and $C_m\equiv C_{m_1}C_{m_2}\cdots C_{m_k}$. The resulting expression must be expressible as (multiple-)commutators of the C's. Only commutators of $H(t)$, in the form of C_m and their commutators, enter into K. This is so because in the special case when $H(t)$ is a member of a Lie algebra, $U(T,T')$ is a member of the corresponding Lie group and so K must also be a member of a Lie algebra.

By definition, K_i contains i factors of $H(t)$. Calculation for the first five is carried out in Appendix D. The result is

$$K_1=C_1,\qquad K_2=\tfrac{1}{2}C_2,\qquad K_3=\tfrac{1}{3}C_3+\tfrac{1}{12}[C_2,C_1],\qquad K_4=\tfrac{1}{4}C_4+\tfrac{1}{12}[C_3,C_1],$$
$$K_5=\tfrac{1}{5}C_5+\tfrac{3}{40}[C_4,C_1]+\tfrac{1}{60}[C_3,C_2]+\tfrac{1}{360}[C_1,[C_1,C_3]]+\tfrac{1}{240}[C_2,[C_2,C_1]]$$
$$+\tfrac{1}{720}[C_1,[C_1,[C_1,C_2]]]. \tag{5.4}$$

K_n consists of C_n/n, plus the compensating terms in the form of commutators of the C's. By counting powers of $H(t)$ it is clear that the subscripts of these C's must add up to n, but beyond that all independent commutators and multiple commutators may appear. For that reason it is rather difficult to obtain an explicit formula valid for all K_n, if for no other reason than the fact that new commutator structures appear in every new n. It is however very easy to compute K_n using (5.3) in a computer. This is actually how K_5 was obtained.

Nevertheless, when we stick to commutators of a definite structure, their coefficients in K can be computed as follows.

C. Coefficient $\eta(m_1m_2\cdots m_k)$

It is not difficult to compute $\eta(m_1\cdots m_k)$ for small k (but arbitrary m_i). The computation for $k\leqslant 4$ will be given below. In order to avoid excessive subscripts we shall use variables like w,x,y,z to denote the positive integers m_i.

1. $\eta(xy)$

According to (5.3), C_xC_y can come from U_{x+y}, or U_xU_y. The former corresponds to $\ell=1$, and the latter $\ell=2$. Using (4.2), we obtain

$$\eta(xy)=\frac{1}{(x+y)y}-\frac{1}{2xy}=\frac{y-x}{2xy(x+y)}. \tag{5.5}$$

The antisymmetry under x,y exchange shows explicitly that it is the commutator $\eta(xy)\times[C_x,C_y]$ that enters into K.

Using this formula, we can verify the coefficients of the single commutator terms appearing in (5.4): $\eta(2,1)=1/12$, $\eta(3,1)=1/12$, $\eta(4,1)=3/40$, $\eta(3,2)=1/60$.

2. $\eta(xyz)$

According to (5.3), $C_xC_yC_z$ can come from $U_{x+y+z}(\ell=1)$, U_xU_{y+z} and $U_{x+y}U_z(\ell=2)$, and $U_xU_yU_z(\ell=3)$. Using (4.2), we get

$$\eta(xyz)=\frac{1}{(x+y+z)(y+z)z}-\frac{1}{2}\left[\frac{1}{x(y+z)z}+\frac{1}{(x+y)yz}\right]+\frac{1}{3}\frac{1}{xyz}$$

$$=\frac{y^2(x+z)+y(2x^2+2z^2-3xz)-xz(x+z)}{6xyz(x+y)(y+z)(x+y+z)}. \tag{5.6}$$

Because of the Jacobi identity there are only two independent double commutators in K. They can be taken to be $\alpha[C_x,[C_y,C_z]]+\beta[C_y,[C_x,C_z]]$, if $y\neq z$. We may take $\beta=0$ if $x=y$. In any case, we have $\alpha=\eta(xyz)$.

We can use this formula to verify the double commutator terms in (5.4): $\eta(1,1,2)=1/360$, $\eta(2,2,1)=1/240$.

3. $\eta(wxyz)$

$C_wC_xC_yC_z$ may come from $U_{w+x+y+z}$, $U_{w+x+y}U_z$, $U_{w+x}U_{y+z}$, U_wU_{x+y+z}, $U_{w+x}U_yU_z$, $U_wU_{x+y}U_z$, $U_wU_xU_{y+z}$, and $U_wU_xU_yU_z$. Thus

$$\eta(wxyz)=[(w+x+y+z)(x+y+z)(y+z)z]^{-1}-[2(w+x+y)(x+y)yz]^{-1}$$
$$-[2(w+x)x(y+z)z]^{-1}-[2w(x+y+z)(y+z)z]^{-1}+[3(w+x)yz]^{-1}$$
$$+[3w(x+y)yz]^{-1}+[3wx(y+z)z]^{-1}-[4wxyz]^{-1}. \tag{5.7}$$

It can be shown (Appendix E) that this is also the coefficient of the triple commutator $[C_w,[C_x,[C_y,C_z]]]$.

This formula leads to $\eta(1,1,1,2)=1/720$, agreeing with the coefficient of the last term in (5.4).

D. Gauge invariance

Under the ''gauge transformation,''

$$\delta H(t)=\frac{d\Lambda(t)}{dt}+[\Lambda(t),H(t)]\equiv\delta_{-1}H(t)+\delta_0H(t), \tag{5.8}$$

where $\Lambda(t)$ is an arbitrary operator vanishing at $t=T$ and $t=T'$, $U(T,T')$ is gauge invariant. This means that $K=K(T,T')$ must also be gauge invariant. Since δ_{-1} decreases the power of H by one, and δ_0 leaves it unchanged, the gauge invariance must be implemented order-by-order as

$$\delta_{-1}U_{n+1}+\delta_0U_n=0, \tag{5.9}$$

$$\delta_{-1}K_{n+1}+\delta_0K_n=0. \tag{5.10}$$

Equation (5.9) is easy to verify but Eq. (5.10) is not. See Appendix F. The reason is that U_n transforms simply under a gauge change, as in (5.9), but C_n does not. It is this complexity of gauge behavior of C_n that makes the verification of (5.10) complicated, but we will do it explicitly for the first few orders in Appendix F. From this point of view, the reason why the dependence of K_n on C_m is so complicated is simply because the complexity is necessary to offset the complexity of the gauge behavior of C_m, so that the gauge change of K_n becomes simple again, as in (5.10).

VI. FORMULAS RESULTING FROM SPECIFIC CHOICES OF $H_I(T)$

It is well known that the path-ordered exponential $U[T,T']$ obeys the composition law

$$U[T,T']=U[T,T'']U[T'',T'] \tag{6.1}$$

if $T\geqslant T''\geqslant T'$. We can combine this with a judicious choice of $H(t)$ to obtain many mathematical formulas. Three of them are presented below for illustrative purposes. The first two are purely mathematical results; the third is a non-Abelian eikonal formula useful for high-energy scattering amplitudes.

A. Campbell–Baker–Hausdorff formula

Let $T=2$, $T''=1$, and $T'=0$. Let $H(t)=P$ for $2\geqslant t\geqslant 1$ and $H(t)=Q$ for $1>t\geqslant 0$, where P and Q are arbitrary matrices or operators. Using (6.1) and (5.4), we obtain in this case (see Appendix G for details)

$$C_m=\frac{1}{(m-1)!}(\text{ad } P)^{m-1}Q, \tag{6.2}$$

where as usual $(\text{ad } P)V\equiv[P,V]$ for any operator V. Substituting this into (5.4) we get

$$\begin{gathered}\exp(P)\cdot\exp(Q)=\exp[K_1+K_2+K_3+K_4+K_5+\cdots],\\ K_1=P+Q,\qquad K_2=\tfrac{1}{2}[P,Q],\\ K_3=\tfrac{1}{12}[P,[P,Q]]+\tfrac{1}{12}[Q,[Q,P]],\qquad K_4=-\tfrac{1}{24}[P,[Q,[P,Q]]],\\ K_5=-\tfrac{1}{720}[P,[P,[P,[P,Q]]]]-\tfrac{1}{720}[Q,[Q,[Q,[Q,P]]]]+\tfrac{1}{360}[P,[Q,[Q,[Q,P]]]]\\ +\tfrac{1}{360}[Q,[P,[P,[P,Q]]]]+\tfrac{1}{120}[P,[Q,[P,[Q,P]]]]+\tfrac{1}{120}[Q,[P,[Q,[P,Q]]]],\end{gathered} \tag{6.3}$$

which is the Campbell–Baker–Hausdorff formula. The case when $[P,Q]$ commutes with P and Q is well known. In that case all K_n for $n\geqslant 3$ are zero. Otherwise, up to and including K_4 this formula can be found in Eq. (15), Sec. 6.4, Chap. II of Ref. 10.

If necessary, we may also use the general knowledge for the coefficient $\eta(m)$ obtained in Sec. V C to deduce information on higher-order terms.

B. Translational operator

A simple and trivial example to illustrate the commutative formula (5.1) is obtained by choosing $T=3$, $T'=0$, $H(t)=a(d/dx)$ for $t\in[2,3]$, $H(t)=f(x)$ for $t\in[1,2]$, and $H(t)=-a(d/dx)$ for $t\in[0,1]$. $f(x)$ is an arbitrary function of x and a is a constant. In this case $C_n/n=(a^n/n!)(d^nf(x)/dx^n)$ (see Appendix H), so all commutators of C_i vanish and (5.1) is valid. We get

$$\begin{aligned}U[T,T']&=\exp(\text{ad}/dx)\cdot\exp(f(x))\cdot\exp(-\text{ad}/dx)\\ &=\exp[K]=\exp\left[\sum_{n=0}^{\infty}a^nf^{(n)}(x)/n!\right]=\exp[f(x+a)].\end{aligned} \tag{6.4}$$

This formula, expressing the fact that $\exp(\text{ad}/dx)$ is a translational operator, is of course well-known. There is absolutely no need to derive it with this heavy apparatus. We include it here just to illustrate how different choices of $H(t)$ can lead to different formulas.

C. Non-Abelian eikonal formula

A useful formula in high-energy scattering is obtained from (3.1) by choosing $H_i(t)=\exp(ik_i\cdot pt)V_i\equiv\exp(i\omega_it)$, where V_i are time-independent operators, and p,k_i are arbitrary four-momenta. We should also choose $[T,T']$ to be $[\infty,-\infty]$ to correspond to ''on-shell amplitudes'' and $[T,T']=[\infty,0]$ to correspond to ''off-shell amplitudes.''

In that case, (3.1) reads

$$\begin{gathered}U[123\cdots n]=a[123\cdots n]V_1V_2V_3\cdots V_n,\\ C[123\cdots n]=a[123\cdots n]\cdot\\ [V_1,[V_2,[V_3,[\cdots,[V_{n-1},V_n]\cdots]]]],\end{gathered} \tag{6.5}$$

$$a[123\cdots n]=i^{n-1}\left(\prod_{j=1}^{n-1}\frac{1}{\Sigma_{i=1}^{j}\omega_i+i\epsilon}\right)L_{[T,T']},$$

$$L_{[\infty,-\infty]}=2\pi\delta\left(\sum_{i=1}^{n}\omega_i\right),\qquad L_{[\infty,0]}=\frac{1}{\Sigma_{i=1}^{n}\omega_i+i\epsilon}.$$

Similar formulas can be obtained for $U[s]$ and $C[s]$ for any $s\in S_n$ by permutation. They are the tree amplitudes with vertex V_i governing the emission of n bosons with momenta k_i from a common source particle with momentum p. The bosons are arranged along the tree in the order dictated by the permutation s. This formula is valid in the high-energy approximation where $p^0\gg k_i^\mu$, whence spins can be ignored and the product of propagators along the source line can be approximated by $a[s]$. $n!U_n$ is then the complete tree amplitude obtained by summing the $n!$ tree diagrams. The decomposition theorem for U_n in this case has been derived directly.[5] Its on-shell version exhibits spacetime and color interference, and can be used to prove the self-consistency of the baryonic scattering amplitudes in large-N_c QCD[6]. It also explains the emergence of reggeized gluons and the factorization of high-energy near-forward scattering amplitudes into multiple-reggeon exchanges.[7,8]

ACKNOWLEDGMENTS

I am grateful to Herbert Fried, John Klauder, Cheuk-Yin Wong, and Tung-Mow Yan for informative discussions and suggestions. This research is supported by the Natural Sciences and Engineering Research Council of Canada.

APPENDIX A: TWO PROOFS OF THE GENERAL DECOMPOSITION THEOREM

On account of the importance of the decomposition theorem (3.1) to the rest of this paper, we shall provide two separate proofs for it. Actually a third proof exists, along the lines given in Ref. 5 for the ''multiple commutator formula'' [which is simply the ''non-Abelian eikonal formula'' (6.5)]. However, this requires the expansion of every operator $H_i(t)$ into a sum over a complete set of time-independent operators, thus introducing lots of indices and even more complicated notations. So we shall skip that third proof here.

The combinatorics needed for a general proof are unfortunately rather involved, though the basic idea in either case is really quite simple. In order not to be bogged down with complicated notations, we will start with the simplest case, $n=2$, which already contains most of the basic ideas.

1. $n=2$

a. The decomposition theorem

In that case, the decomposition theorem (3.1) reads

$$2!U_2=U[12]+U[21]=C[1|2]+C[21]=C[1]C[2]+C[21], \tag{A1}$$

$$U[12]=\int_{R[12]}dt_1\,dt_2H_1(t_1)H_2(t_2)=\int_{T'}^{T}dt_1dt_2\theta(t_1-t_2)H_1(t_1)H_2(t_2), \tag{A2}$$

$$U[21]=\int_{R[21]}dt_1\,dt_2H_2(t_2)H_1(t_1)=\int_{T'}^{T}dt_1dt_2\theta(t_2-t_1)H_2(t_2)H_1(t_1) \tag{A3}$$

$$C[1|2]=C[1]C[2]=\int_{T'}^{T}dt_1dt_2H_1(t_1)H_2(t_2), \tag{A4}$$

$$C[21]=\int_{R[21]}dt_1dt_2[H_2(t_2),H_1(t_1)]. \tag{A5}$$

b. First proof for n=2

The first proof makes use of the simple observation that the union of the two triangular integration regions, $R[12]=\{T\geqslant t_1\geqslant t_2\geqslant T'\}$, and $R[21]=\{T\geqslant t_2\geqslant t_1\geqslant T'\}$, is the square region $\{T\geqslant t_1,t_2\geqslant T'\}$. Thus

$$U[12]+U[21]=\int_{R[12]\cup R[21]} H_1(t_1)H_2(t_2)+\int_{R[21]}[H_2(t_2),H_1(t_1)]$$
$$=C[1]C[2]+C[21]\equiv C[1|2]+C[21]. \tag{A6}$$

c. Second proof for n=2

Equation (3.1) is clearly true when $T=T'$, for both sides then vanish. To prove its general validity for any T', it is sufficient to show the T'-derivatives of both sides to be equal. For $n=2$, we have to show that

$$\frac{d}{dT'}\{U[12]+U[21]\}=\frac{d}{dT'}\{C[1|2]+C[21]\}. \tag{A7}$$

This identity is true because

$$\frac{d}{dT'}U[12]=-\int_{T'}^{T}H_1(t)dt\ H_2(T'),\qquad \frac{d}{dT'}U[21]=-\int_{T'}^{T}H_2(t)dt\ H_1(T'),$$
$$\frac{d}{dT}C[1|2]=-H_1(T')\int_{T'}^{T}H_2(t)dt-\int_{T'}^{T}H_1(t)dt\ H_2(T'), \tag{A8}$$
$$\frac{d}{dT'}C[21]=-\left[\int_{T'}^{T}H_2(t)dt,H_1(T')\right].$$

We shall now proceed to the general proofs.

2. First proof for arbitrary *n*

In the case of $n=2$, the ''first proof'' involves two crucial steps. Step one is to recognize that the union of the two triangular regions, $R[12]$ and $R[21]$, is a square, thus allowing the first term on the right-hand side of (A6) to be factorized. The generalization of this to an arbitrary n is the subject of discussions that immediately follow. The second crucial step is to introduce commutators to rewrite $H_2(t_2)H_1(t_1)$ as a sum of two terms: $H_1(t_1)H_2(t_2)+[H_2(t_2),H_1(t_1)]$. The generalization of this step to arbitrary n will then be discussed. In a final subsection, these two steps will be assembled to complete the proof of the general decomposition theorem.

a. Sums and Cartesian products of hyper-triangular regions

Let $[s]$ be a sequence of n natural numbers, not necessarily consecutive, and $[s]=[\tilde{s}_1\tilde{s}_2\cdots\tilde{s}_p]$ be a partition of this sequence into p subsequences. For example, $[s]=[3254167]$ could be partitioned into three subsequences $[\tilde{s}_1]=[32]$,$[\tilde{s}_2]=[5416]$, and $[\tilde{s}_3]=[7]$. We define the *set* $\{\tilde{s}_1;\tilde{s}_2;\cdots;\tilde{s}_p\}$ to consist of all sequences of these n natural numbers, obtained by merging and interleaving the numbers in $\tilde{s}_1,\tilde{s}_2,\cdots,\tilde{s}_p$ in all possible ways, subject to the condition that the original orderings of numbers within each $\tilde{s}_i$ be kept fixed. This set then contains $n!/(k_1!k_2!\cdots k_p!)$ sequences, where k_i is the number of numbers in the subsequence $\tilde{s}_i$, and $\Sigma_{i=1}^{p}k_i=n$. In the explicit examples above, the set $\{32;5416;7\}$ will consist of $7!/(2!4!1!)=105$ sequences, which among others include $[3254167]$,$[3524176]$, and $[7534162]$, but *not* $[2354167]$ and $[6541732]$ because the original orderings are not kept in these last two cases: the numbers 2 and 3 are reversed in the first instance and the number 6 appears before the numbers 5,4,1 in the second.

Given a permutation $s \in S_n$, recall that the hyper-triangular integration region $R[s]$ is defined to be $\{T \geqslant t_{s_1} \geqslant t_{s_2} \geqslant \cdots \geqslant t_{s_n} \geqslant T'\}$. We shall now define the integration region $R\{\tilde{s}_1;\tilde{s}_2;\cdots;\tilde{s}_p\}$ to be the union of all the regions $R[t]$ with $[t] \in \{\tilde{s}_1;\tilde{s}_2;\cdots;\tilde{s}_p\}$. A moment's thought reveals that this is nothing but the Cartesian product of the individual hyper-triangular regions $R[\tilde{s}_1], R[\tilde{s}_2], \cdots, R[\tilde{s}_p]$. This leads to a *factorization theorem* for the following sums of $U[t]$:

$$\sum_{[t] \in \{\tilde{s}_1;\tilde{s}_2;\cdots;\tilde{s}_p\}} U[t] = \int_{R\{\tilde{s}_1;\tilde{s}_2;\cdots;\tilde{s}_p\}} dt_1\, dt_2 \cdots dt_n H_{s_1}(t_{s_1}) H_{s_2}(t_{s_2}) \cdots H_{s_n}(t_{s_n})$$
$$= U[\tilde{s}_1] U[\tilde{s}_2] \cdots U[\tilde{s}_p]. \tag{A9}$$

A factorization theorem similar to this will be used for the proof of the decomposition theorem.

b. Canonical ordering

From now on we shall denote $H_{s_i}(t_{s_i})$ simply as H_{s_i}, and $H_{s_1} H_{s_2} \cdots H_{s_n}$ as $H[s_1|s_2|\cdots|s_n] \equiv H_c[s]$. The subscript "$c$" of H serves to indicate that a vertical bar is to be put after each number in the sequence $[s]$ to indicate that products of operators should be taken.

A similar symbol without the vertical bar will denote *nested multiple commutators*. Thus $H[1|2|3|4] = H_1 H_2 H_3 H_4 = H_c[1234]$, but $H[12|34] = [H_1, H_2] \cdot [H_3, H_4]$ and $H[1234] = [H_1, [H_2, [H_3, H_4]]]$, etc.

Suppose $[s] = [s_1 s_2 \cdots s_n]$ is a sequence of n numbers, not necessarily consecutive. Then $[s]_c$ will be used to denote the same sequence, but with *vertical bars inserted after* s_i *if* $s_i < s_j$ *for all* $i < j$. For example, if $[s] = [7125]$, then $[s]_c = [71|2|5]$, and if $[s] = [21345]$, then $[s]_c = [21|3|4|5]$. An operator of the form $H[s]_c$ will be called a *canonical operator*. Please note the difference between $H_c[s]$ and $H[s]_c$. The former is generally not a canonical operator, but the latter is, by definition.

Let $s \in S_n$. From (2.1), the integrand in $U[s]$ is given by $H_c[s]$, which as mentioned above is generally *not* a canonical operator. By introducing commutators, it is possible to *canonically order* this operator, meaning that it can be written as a linear combination of canonical operators. I shall illustrate how this is done by looking at a simple example, $H[3|2|1] = H_3 H_2 H_1 = H_c[321]$:

$$\begin{aligned} H[3|2|1] &= H_3 H_2 H_1 \\ &= H_3 [H_2, H_1] + [H_3, H_1] H_2 + H_1 H_3 H_2 = [H_2, H_1] H_3 + [H_3, [H_2, H_1]] \\ &\quad + [H_3, H_1] H_2 + H_1 H_3 H_2 \\ &= H[21|3] + H[321] + H[31|2] + H[1|3|2] \qquad \text{(A10)} \\ &= H[21|3] + H[321] + H[31|2] + H[1|2|3] + H[1|32]. \qquad \text{(A11)} \end{aligned}$$

Note that the first three terms of (A10) are canonical operators but not the fourth. This is then fixed in (A11) so that all its five operators are now canonical.

With this example in mind, we can now discuss how to canonically order an arbitrary operator of the form $H_c[s]$, where $[s]$ is a sequence of numbers, not necessarily consecutive.

Locate the smallest number s_0 in the sequence. By introducing commutators and commutators of commutators, we shall try to move H_{s_0}, and commutators involving it, to the extreme left. This results in a sum of a number of terms which can be described as follows.

Suppose s_0 is situated somewhere in the middle of $[s]$ so that $[s] = [\tilde{s}_1 s_0 \tilde{s}_2]$, with a subsequence $[\tilde{s}_1]$ before s_0 and a subsequence $[\tilde{s}_2]$ after. Let $[\tilde{s}_1']$ and $[\tilde{s}_1'']$ be two complementary subsequences of numbers in $[\tilde{s}_1]$ in the sense that all numbers in $[\tilde{s}_1]$ appear either in $[\tilde{s}_1']$ or $[\tilde{s}_1'']$. For example, if $[\tilde{s}_1] = [7352]$, then $[\tilde{s}_1'], [\tilde{s}_1'']$ may be $[3], [752]$, or $[75], [32]$, etc., but not $[7], [52]$ (3 is left out), nor $[37], [52]$ (order of 7,3 are reversed). If $[\tilde{s}_1]$ has q numbers, then there are altogether 2^q pairs of complementary subsequences like that.

Now we are ready to describe the terms obtained by the move of H_{s_0} and its commutators to the extreme left in order to achieve canonical ordering. They involve all terms of the form $H[\tilde{s}_1' s_0]H_c[t]$, where $[t]=[\tilde{s}_1''\tilde{s}_2]$ is the sequence obtained by prepending the subsequence $[\tilde{s}_1'']$ to the subsequence $[\tilde{s}_2]$. Since s_0 is the smallest number in $[s]$, the operator $H[\tilde{s}_1' s_0]$ is canonical, though the remaining factor $H_c[t]$ may not be. If not, we will then repeat the same procedure, find the smallest number t_0 of $[t]$, and move H_{t_0} and its commutators to the extreme left of $H_c[t]$ to render sections of $H_c[t]$ canonical. Continuing thus, eventually $H_c[s]$ can be rendered into sums of canonical operators.

We can describe this result in another way, more useful for our subsequent proof. Given a canonical operator $H[u]_c$, with $[u]\in S_n$, this result states that it will appear in the canonical ordering of $H_c[s]$ iff $[s]\in\{u\}$, where $\{u\}$ is the *set* of permutations obtained by changing the vertical bars in $[u]_c$ into semicolons '';'', and the square brackets $[\cdots]_c$ into curly brackets $\{\cdots\}$. For example, if $[u]_c=[21|3]$, then $\{u\}=\{21;3\}=\{[321],[231],[213]\}$, and the operators $H_c[s]$ containing the canonical operator $H[21|3]$ are $H_c[321]=H[3|2|1]$, $H_c[231]=H[2|3|1]$, and $H_c[213]=H[2|1|3]$.

c. Proof

We can now assemble these two ingredients into a proof of the general decomposition theorem (3.1). According to the rule for canonical ordering, the integrand $H[u]_c$ will be contained in $U[s]$ iff $[s]\in\{u\}\equiv\{\tilde{u}_1;\tilde{u}_2;\cdots;\tilde{u}_q\}$. The integration region for the integrand $H[u]_c$ is therefore $\cup_{[s]\in\{u\}}R[s]=R\{u\}$. The resulting integral is therefore

$$\int_{R\{u\}} dt_1\cdots dt_n\ H[\tilde{u}_1|\tilde{u}_2|\cdots|\tilde{u}_q]=C[\tilde{u}_1]C[\tilde{u}_2]\cdots C[\tilde{u}_q]=C[\tilde{u}_1|\tilde{u}_2|\cdots|\tilde{u}_q]. \tag{A12}$$

The last step is similar to the one used to obtain the factorization theorem (A9). Summing over all possible $[u]\in S_n$ is equivalent to summing over all possible $s\in S_n$; by doing so we will get from (A12) the general decomposition theorem (3.1).

3. Second proof for arbitrary *n*

The decomposition theorem is trivially true when $T=T'$, thus its general validity would follow if the T'-derivatives of (3.1) are obeyed:

$$\frac{d}{dT'}\sum_{[u]\in S_n} U[u]=\frac{d}{dT'}\sum_{[u]\in S_n} C[u]_c. \tag{A13}$$

Our second proof consists of proving this equation, with the help of the identities

$$\frac{d}{dT'}U[u]=-U[u']H_r(T'), \tag{A14}$$

$$\frac{d}{dT'}C[\tilde{u}_j]=-[C[\tilde{u}_j'],H_s(T')], \tag{A15}$$

where r is the last element in $[u]\in S_n$ and s is the last element of $[\tilde{u}_j]$. In other words, $[u]=[u'r]$ and $[\tilde{u}_j]=[\tilde{u}_j's]$. $[\tilde{u}_j]$ is one of the sequences between vertical bars in $[u]_c$, and by definition $C[u]_c$ is given by a product of $C[\tilde{u}_j]$ over different j.

We have to be careful with (A14) and (A15) when $[u]$ or $[\tilde{u}_j]$ contains only one number. In that case $[u']$ and $[\tilde{u}_j']$ are void and $U[u']=1$. Moreover,

$$\frac{d}{dT'}C[s]=-H_s(T'). \tag{A16}$$

Let $S_{n-1}[m]$ be the permutation group of the first n natural numbers with the number m removed. Using (A14), the left-hand side of (A13) becomes

$$\frac{d}{dT'}\sum_{[u]\in S_n} U[u] = -\sum_{m=1}^{n}\sum_{[u']\in S_{n-1}[m]} U[u']H_m(T'). \tag{A17}$$

In order for (A13) to be true, the right-hand side must also be given by (A17).

Now in (A17) the operators $H_m(T')$ always appear at the extreme right. On the other hand, if we apply (A15), (A16) to compute the right-hand side of (A13), the operators $H_m(T')$ may also appear at other positions. In fact, when $H_m(T')$ appears m is always the last element of $[\tilde{u}_j] = [\tilde{u}_j' m]$ for some j. To determine for what $[u]_c$ this occurs we consider an arbitrary element $[u'] \in S_{n-1}[m]$. Suppose $[u']_c = [\tilde{u}_1'|\tilde{u}_2'|\cdots|\tilde{u}_k']$. We will now discuss where in $[u']_c$ we may insert the number m to obtain a legitimate $[u]_c$ for some $[u]\in S_n$ in which m appears just before a vertical bar, or the right-hand square bracket]. If m is to be inserted at the end of a subsequence $\tilde{u}_j'$, then m must be smaller than all the numbers in $\tilde{u}_j'$. Since the last element of $\tilde{u}_j'$ increases with j, if m may appear at the end of $\tilde{u}_j'$ it may also appear at the end of $\tilde{u}_\ell'$ for all $\ell > j$. On the other hand, there may be a smallest $j=j_0$. In that case, it is not allowed to insert m at the end of $[\tilde{u}_j']$ for $j<j_0$, but the insertion $[\tilde{u}_1'|\cdots|m|\tilde{u}_{j_0}'|\cdots|\tilde{u}_k]$ is allowed.

So consider a fixed $[u'] \in S_{n-1}[m]$ and all the resulting $[u]\in S_n$ with these insertions of the number m. Using (A15) and (A16), it is now easy to see that $dC[u]_c/dT'$, summed over all such $[u]$'s, will give $-C[u']_c H_m(T']$. Summing over all $[u']$ and all m, we therefore regain the right-hand side of (A17), and hence proving (A13), provided

$$\sum_{[u']\in S_{n-1}[m]} U[u'] = \sum_{[u']\in S_{n-1}[m]} C[u']_c. \tag{A18}$$

This last identity follows from the induction hypothesis (on n), which we may invoke, because (3.1) is true for $n=2$. This completes the second proof of the decomposition theorem.

APPENDIX B: SPECIAL DECOMPOSITION THEOREM

When the operator functions $H_i(t)$ are all identical, the decomposition theorem reduces to Eq. (4.1), as we shall show by deriving the expression for $\xi(m)$. This is equal to the number of $[s]_c = [\tilde{s}_1|\tilde{s}_2|\cdots|\tilde{s}_k]$ with $[\tilde{s}_j]$ containing m_j numbers ($\Sigma_j\, m_j = n$), divided by $n!$.

We remind the readers how to construct $[s]_c$ for every $s\in S_n$. A vertical bar is placed behind a number s_i iff $s_i < s_j$ for all $i<j$.

Thus the last element of $[\tilde{s}_1]$ is always the smallest of the n numbers, i.e., the number ''1.'' There are therefore $[(n-1)!/(m_1-1)!(n-m_1)!](m_1-1)! = n!/(n-m_1)!$ ways of constructing $[\tilde{s}_1]$.

Similarly, the last element of $[\tilde{s}_2]$ is the smallest number in the residue sequence $[s]/[\tilde{s}_1] = [\tilde{s}_2\tilde{s}_3\cdots\tilde{s}_k]$. The remaining m_2-1 numbers in $\tilde{s}_2$ can be chosen arbitrarily from the $n-1-m_1$ elements in the residue sequence, so there are $(n-1-m_1)!/(n-m_1-m_2)!$ ways of doing so. Continuing thus, the total number of ways is simply the product of these numbers. After dividing by $n!$, we get

$$\xi(m_1 m_2\cdots m_k) = \left[(n-m_1)(n-m_1-m_2)\cdots\left(n-\sum_{i=1}^{k-1} m_i\right)\right]^{-1}, \tag{B1}$$

which is the same as the formula given in (4.1).

APPENDIX C: EXPONENTIAL FORMULA FOR COMMUTING C_i'S

To obtain (5.1) when all the C_i's commute, we need to know the number of $[s]_c$ whose vertical bars separate the sequence into m_j subsequences of length j ($j=1,2,3,\cdots$).

The number of ways that n numbers can be divided into m_j groups of j numbers is $n!/\Pi_j(m_j!j!^{m_j})$. From any one of these groupings we can produce $\Pi_j(m_j-1)!$ viable $[s]_c$'s in the following way. Recall that the vertical bars of $[s]_c$ are put behind a number s_i iff $s_i<s_k$ for all $i<k$. Stated differently, this means that (i) the last number in each group of numbers separated by the vertical bars is always the smallest number of that group, and that (ii) the groups must be arranged in such a way that their smallest numbers increase from left to right. For a given grouping of numbers, there are $\Pi_j(j-1)!^{m_j}$ ways of arranging the orderings within each group to satisfy (i). Once this is done, we can order the different groups in a unique way so that (ii) is satisfied. Hence the number of $[s]_c$ is given by $[n!/\Pi_j(m_j!j!^{m_j})][\Pi_j(j-1)!^{m_j}]=n!/(\Pi_j m_j!j^{m_j})$. From this the rest of (5.1) follows easily.

APPENDIX D: GENERAL EXPONENTIAL FORMULA

(5.4) can be computed from (5.3) and (4.1) as follows. If $P(U_n)$ is a polynomial of U_n, then $[P(U_n)]_m$ is obtained from $P(U_n)$ by discarding all products of U_i's whose ''degree'' (sum of indices i) is not equal to m. With this convention we have $K_n=\Sigma_{\ell=1}^{n}(-)^{\ell-1}[U^{\ell}]_n$, so

$$K_1=U_1=C_1, \qquad K_2=U_2-\tfrac{1}{2}U_1^2=\tfrac{1}{2}C_2,$$

$$K_3=U_3-\tfrac{1}{2}(U_1U_2+U_2U_1)+\tfrac{1}{3}U_1^3=\tfrac{1}{3}C_3+\tfrac{1}{12}[C_2,C_1], \tag{D1}$$

$$K_4=U_4-\tfrac{1}{2}[U_1U_3+U_3U_1+U_2^2]+\tfrac{1}{3}[U_1^2U_2+U_1U_2U_1+U_2U_1^2]-\tfrac{1}{4}U_1^4,$$

$$K_5=U_5-\tfrac{1}{2}[U_1U_4+U_4U_1+U_2U_3+U_3U_2]+\tfrac{1}{3}[U_1^2U_3+U_1U_3U_1+U_3U_1^2+U_2^2U_1$$
$$+U_2U_1U_2+U_1U_2^2]-\tfrac{1}{4}[U_1^3U_2+U_1^2U_2U_1+U_1U_2U_1^2+U_2U_1^3]+\tfrac{1}{5}U_1^5.$$

Substituting in the expressions given in (4.2) for U_i into the last two equations, we obtain the expression for K_4 and K_5 given in (5.4).

APPENDIX E: NESTED MULTIPLE COMMUTATORS

Given n operators $C_{m_i}\equiv B_i$, we denoted their nested multiple commutators by

$$B[s]\equiv B[s_1s_2\cdots s_n]\equiv[B_{s_1},[B_{s_2},[\cdots,[B_{s_{n-1}},B_{s_n}]\cdots]]], \tag{E1}$$

where $s\in S_n$. We want to show that the $(n-1)!$ nested operators $B[s'1]$, with $s'\in S_{n-1}[1]$ [this is defined just below eq. (A16)] forms a linear basis for all the n-tuple multiple commutators. In other words, if V_n is the vector space spanned by these $(n-1)!$ nested commutators, then V_n contains all n-tuple multiple commutators.

This means that (i) multiple commutators not of the nested form can be written as linear combinations of the multiple commutators of the nested form, and (ii) nested multiple commutators with B_1 not at the last place can be written as linear combinations of those with B_1 at the last place.

We shall use M_n to denote the vector space generated by all n-tuple multiple commutators. We shall now proceed to show that $M_n=V_n$.

To denote a multiple commutator not of the nested type, we use parentheses to single out those factors within it that are nested. For example,

$$B[((123)4(567))8]=[B[(123)4(567)],B_8],$$

$$B[(123)4(567)]=B[(123)4567]=[B[(123)],[B_4,[B_5,[B_6,B_7]]]], \tag{E2}$$

$$B[(123)]=B[123]=[B_1,[B_2,B_3]].$$

Note that if $[s]$ is any sequence of numbers, then $B[\cdots(s)]=B[\cdots s]$. In other words, the pair of parentheses located at the end can be dropped.

To prove (i), we must show that all the parentheses can be removed. This can be accomplished via the Jacobi identity

$$[A_1,[A_2,A_3]]=[A_3,[A_2,A_1]]-[A_2,[A_3,A_1]], \tag{E3}$$

with A_1 being the operator contained in the leftmost pair of parentheses. The identity will be used to move A_1 to the right. If it reaches the end then the parentheses can be dropped. Repeating this procedure eventually we can gradually move all the parentheses to the end and have them all dropped. This shows that $M_n=V_n$.

We shall now prove (ii) by induction.

For $n=2$ this is obvious because

$$[B_2,B_1]=-[B_2,B_1]. \tag{E4}$$

For $n=3$ this follows from (E4) and the Jacobi identity (E3). We merely have to take $B_i=A_i$.

Assuming this can be done for n up to $n=m-1$, we must now show it to be true for $n=m$.

Unless B_1 is at the beginning position there is nothing to prove, for otherwise B_1 and operators to its right are located in an $M_n=V_n$ for $n\leqslant m-1$, so B_1 can be moved to the end.

If B_1 is located at the first position, so that the nested multiple commutator is of the form $B[1s]$ for some sequence $[s]$ with $n-1$ numbers, then using (E3) with $B_1=A_1$ we can move B_1 to the right. Again it and the operators to its right now belongs to M_n with $n\leqslant m-1$, so by the induction hypothesis B_1 can be moved to the end. This completes the proof of (ii).

In particular, V_4 is spanned by the 6 nested commutators $B[4321],B[4231],B[3421]$, $B[3241],B[2431]$, and $B[2341]$. Only the first nested commutator contains the term $B_4B_3B_2B_1$, so the coefficient of this nested commutator in K is identical to the coefficient of this term in K, a fact that has been used in Sec. V C 3 to compute the coefficient of a nested commutator.

This result can also be used to understand why only the particular nested commutators of H_i defined in C_m occur in K.

APPENDIX F: GAUGE INVARIANCE

We want to study in this Appendix the behavior of U_n, C_n, and K_n under the ''gauge transformation''

$$\delta H(t)=\frac{d\Lambda(t)}{dt}+[\Lambda(t),H(t)]\equiv\delta_{-1}H(t)+\delta_0H(t), \tag{F1}$$

where $\Lambda(t)$ is an arbitrary function of t vanishing at the boundaries: $\Lambda(T)=\Lambda(T')=0$.

Recall that $U_n=U[123\cdots n]$ and $C_n=C[123\cdots n]$, that time is ordered according to the sequence of indices in the square bracket, and that an operator $H(t_i)$ is present inside the integral at the position of the index i. The only difference between $U[\cdots]$ and $C[\cdots]$ is that this operator appears as a factor in a straight product in the former, and it appears inside a nested multiple commutator in the latter.

We shall now introduce an index with a prime to denote the operator Λ. At the position where i' appears in the sequence will be the operator $\Lambda(t_i)$. If the primed index is inside $U[\cdots]$, then Λ is a member of a straight product as before. If it appears inside $C[\cdots]$, then it is a part of the nested commutators.

We will also introduce parentheses $(i'i)$ to denote commutators. Thus at the position this appears should stand the operator $[\Lambda(t_i),H(t_i)]$.

With these notations we are now ready to discuss the gauge properties of the various quantities.

For $U_n=U[123\cdots n]$ it is simple:

$$\delta_0U_{n+1}=\sum_{j=1}^{n}\{-U[12\cdots j'j\cdots n]+U[12\cdots jj'\cdots n]\}=-\sum_{j=1}^{n}U[12\cdots(j'j)\cdots n], \tag{F2}$$

$$\delta_{-1}U_n=\sum_{j=1}^{n} U[12\cdots(j'j)\cdots n]. \tag{F3}$$

Clearly (5.9) is satisfied and U is gauge invariant.

Now for C_n:

$$\delta_0 C_{n+1}=\sum_{j=1}^{n}\{-C[12\cdots j'j\cdots n]+C[12\cdots jj'\cdots n]\}, \tag{F4}$$

$$\delta_{-1}C_n=\sum_{j=1}^{n} C[12\cdots(j'j)\cdots n], \tag{F5}$$

so they look deceptively similar to the equations for U_n. However, since the operators in $C[\cdots]$ appear in *nested* commutators, the two terms in (F4) can be combined into $-C[12\cdots(j'j)\cdots n]$ only with the help of the Jacobi identity, and this we can do only for $j\leqslant n-1$. Jacobi identity involves double commutators and those are absent at $j=n$. Instead, we can use the antisymmetry of a single commutator to add up these two terms. Hence

$$\delta_0 C_{n+1}=-\sum_{j=1}^{n-1} C[12\cdots(j'j)\cdots n]-2C[12\cdots(n'n)]. \tag{F6}$$

The parentheses in the last term may be dropped; we kept it there just for the uniformity of notation with the other terms. The main difference between this and (F2) is that this last term now has a factor of 2, which makes $\delta_{-1}C_{n+1}+\delta_0 C_n\neq 0$. It is this ''slight'' difference that eventually makes K very complicated just in order to keep its gauge invariance according to (5.10)!

We shall now use (D1) and (F6) to verify (5.10) for the first few orders. Since K_n contains the term C_n/n, it is useful to compute

$$\frac{1}{n+1}\delta_{-1}C_{n+1}+\frac{1}{n}\delta_0 C_n=\frac{1}{n(n+1)}\left\{\sum_{j=1}^{n-1} C[12\cdots(j'j)\cdots n]-(n-1)C[12\cdots(n'n)]\right\}. \tag{F7}$$

Referring back to (5.4), commutator terms are not present at K_n for $n=1,2$, which means that gauge invariance of K demands (F7) to be zero for $n=1$, which it is. For $n=2$,

$$\begin{aligned}\delta_{-1}K_3+\delta_0 K_2&=\tfrac{1}{3}\delta_{-1}C_3+\tfrac{1}{2}\delta_0 C_2+\tfrac{1}{12}\delta_{-1}([C_2,C_1])\\&=\tfrac{1}{6}(C[(1'1)2]-C[1(2'2)]-[C[(1'1)],C_1]).\end{aligned} \tag{F8}$$

The last expression is arrived at by using (F7), $\delta_{-1}C_1=0$, and $\delta_{-1}C_2=-2C[(1'1)]$. Now use the observation just before Eq. (A6) to obtain the identity

$$[C_1,C[(1'1)]]=C[1(2'2)]-C[(1'1)2]. \tag{F9}$$

Substituting this back into (F8) we conclude that $\delta_{-1}K_3+\delta_0 K_2=0$.

I have verified (5.10) to one higher order. The verification becomes increasingly more difficult at larger n.

APPENDIX G: CAMPBELL–BAKER–HAUSDORFF FORMULA

Consider a special case in which $T=2$, $T'=0$, $H(t)=P$ for $2\geqslant t\geqslant 1$, and $H(t)=Q$ for $1\geqslant t\geqslant 0$. The operators P and Q are completely arbitrary.

We consider the ramification of (5.4) in this special case. Using (6.1) with $T''=1$, the left-hand-side of (5.4) is just $\exp(P)\exp(Q)$. To compute the right-hand side we must first compute the C_m's in this special situation using (2.3).

Clearly $C_1=P+Q$. For C_m with $m\geqslant 2$, in order for the commutators inside (2.3) not to vanish we must have t_m between 0 and 1, and all the other $t_i(1\leqslant i\leqslant m-1)$ between 1 and 2. Thus

$C_m=(\text{ad } P)^{m-1}Q/(m-1)!$, where as usual $(\text{ad } P)V=[P,V]$ for any operator V. Substituting this back into (5.4) we get the Campbell–Baker–Hausdorff formula shown in (6.3).

APPENDIX H: TRANSLATIONAL OPERATOR

Consider now the special case when $T=3$, $T'=0$, $H(t)=\text{ad}/dx$ for t between 3 and 2, $H(t)=f(x)$ for t between 2 and 1, and $H(t)=-\text{ad}/dx$ for t between 1 and 0. Here $f(x)$ is an arbitrary function and a is a constant.

By using (6.1) the left-hand side of (5.4) becomes $\exp(\text{ad}/dx)\cdot\exp(f(x))\cdot\exp(-\text{ad}/dx)$. To compute the right-hand side we need to compute C_m using (2.3). It is clear that $C_1=f(x)$. To compute C_{m+1} for $m\geqslant 1$, note the following. In order for the commutators in (2.3) not to vanish, t_{m+1} must either be between 0 and 1, or between 1 and 2. We shall denote the former contribution by C'_{m+1}, and the latter by C''_{m+1}.

To compute C'_{m+1}, we must have t_m to be between 1 and 2, and t_i for $i<m$ to be between 2 and 3. Hence $C'_{m+1}=a^m(d^mf(x)/dx^m)/(m-1)!$. For C''_{m+1}, we must have all t_i for $i\leqslant m$ to be between 2 and 3, hence $C''_{m+1}=a^m(d^mf(x)/dx^m)/m!$. Adding up the two, we get $C_{m+1}=C'_{m+1}+C''_{m+1}=(m+1)a^m(d^mf(x)/dx^m)/m!$. Hence all the C_i's commute with one another, so (5.1) can be used instead. (6.4) then follows immediately.

[1] A. J. Dragt and J. M. Finn, J. Math. Phys. **17**, 2215 (1976); A. J. Dragt and E. Forest, *ibid.* **24**, 2734 (1983); L. M. Healy, A. J. Dragt, and I. V. Gjala, *ibid.* **33**, 1948 (1992).
[2] R. Torgerson, Phys. Rev. **143**, 1194 (1966); H. Cheng and T. T. Wu, *ibid.* **182**, 1868 (1969); **182**, 1899 (1969); M. Levy and J. Sucher, *ibid.* **186**, 1656 (1969).
[3] For a review, see H. Cheng and T. T. Wu, *Expanding Protons: Scattering at High Energies* (MIT, Cambridge, MA, 1987); R. J. Glauber, in *Lectures in Theoretical Physics*, edited by W. E. Brittin and L. G. Dunham (Interscience, New York, 1959), Vol. 1.
[4] D. Yennie, S. Frautschi, and H. Suura, Ann. Phys. (N.Y.) **13**, 379 (1961); S. Weinberg, Phys. Rev. **140**, B516 (1965); G. Grammer, Jr. and D. R. Yennie, Phys. Rev. D **8**, 4332 (1973).
[5] C. S. Lam and K. F. Liu, Nucl. Phys. B **483**, 514 (1997).
[6] C. S. Lam and K. F. Liu, Phys. Rev. Lett. **79**, 597 (1997).
[7] Y. J. Feng, O. Hamidi-Ravari, and C. S. Lam, Phys. Rev. D **54**, 3114 (1996).
[8] Y. J. Feng and C. S. Lam, Phys. Rev. D **55**, 4016 (1997).
[9] L. D. Landau and I. J. Pomeranchuk, Dokl. Akad. Nauk SSSR **92**, 535 (1953); **92**, 735 (1953); A. B. Migdal, Phys. Rev. **103**, 1811 (1956); R. Blankenbecler and S. D. Drell, Phys. Rev. D **53**, 6265 (1996); R. Baier, Yu. L. Dokshitzer, A. H. Mueller, S. Peigné, and D. Schiff, hep-ph/9604327; P. L. Anthony *et al.*, Phys. Rev. Lett. **75**, 1949 (1995); SLAC-PUB-7413 (February, 1997).
[10] N. Bourbaki, *Lie Groups and Lie Algebras* (Hermann, Paris, 1975).

Finite-dimensional irreducible representations of twisted Yangians

A. I. Molev[a)]
School of Mathematics and Statistics, Carslaw Building (F07), University of Sydney, Sydney, New South Wales 2006, Australia

(Received 1 December 1997; accepted for publication 29 April 1998)

We study quantized enveloping algebras called twisted Yangians. They are analogs of the Yangian $Y(gl(N))$ for the classical Lie algebras of *B*, *C*, and *D* series. The twisted Yangians are subalgebras in $Y(gl(N))$ and coideals with respect to the coproduct in $Y(gl(N))$. We give a complete description of their finite-dimensional irreducible representations. Every such representation is highest weight and we give necessary and sufficient conditions for an irreducible highest weight representation to be finite dimensional. The result is analogous to Drinfeld's theorem for the ordinary Yangians. Its detailed proof for the *A* series is also reproduced. For the simplest twisted Yangians we construct an explicit realization for each finite-dimensional irreducible representation in tensor products of representations of the corresponding Lie algebras. © *1998 American Institute of Physics.*
[S0022-2488(98)01610-7]

I. INTRODUCTION

The Yangian Y(a) associated with a simple complex Lie algebra a is a deformation of the universal enveloping algebra $U(a[t])$ in the class of Hopf algebras.[1,2] In the case of the *A* series one can also define the Yangian $Y(N)=Y(gl(N))$ for the reductive complex Lie algebra $a=gl(N)$ (see, e.g., Ref. 3). The Yangian $Y(sl(N))$ can be identified with a (Hopf) subalgebra in $Y(gl(N))$. The Yangians are closely related to rational solutions of the Yang–Baxter equation[1] and play the role of quantum symmetry groups of certain integrable systems; see, e.g., Refs. 4–7. Finite-dimensional irreducible representations of the Yangians were classified by Drinfeld.[2] In the case of $a=gl(2)$ this had been done by Tarasov;[8,9] see also Ref. 10. [In these papers the language of the quantum inverse scattering theory and integrable models was used. Representations of the Yangian Y(2) appear there in the form of the *monodromy matrix* for the XXX-model]. This result plays the main role in the proof in the general case. Representations of the Yangians were further studied in Refs. 11–18.

For any a the Yangian Y(a) contains the universal enveloping algebra U(a). However, the *A* series case seems to be exceptional since only in this case does there exists an algebra homomorphism $Y(a)\to U(a)$ identical on U(a). This homomorphism is important for applications of the ''quantum theory'' to the representation theory of the Lie algebra $gl(N)$. In particular, it can be used to construct Laplace operators for $gl(N)$, to derive characteristic identities and Capelli identities.[15,19–23]

It was shown by Olshanski[24] that the Yangian $Y(N)$ is isomorphic to a certain projective limit algebra. Namely, there exist natural algebra homomorphisms $A_N(M+1)\to A_N(M)$, where $A_N(M)$ denotes the $gl(M)$-centralizer in $U(gl(N+M))$, and the projective limit A_N of the chain $\{A_N(M)\}$ as $M\to\infty$ contains $Y(N)$ as a subalgebra. This construction can be applied to other series of classical Lie algebras.[25] However, as a result one obtains not the corresponding Yangian Y(a), but different algebras which were called twisted Yangians;[25] see also Ref. 26. Defining relations for the twisted Yangian can be written in the form of a reflection-type equation [see (2.16) below] which allows one to use a special *R*-matrix technique for their study. Some other algebras associated with the reflection equation were studied in Refs. 27–29.

Let $g(N)$ denote the orthogonal Lie algebra $o(N)$ or symplectic Lie algebra $sp(N)$. We regard $g(N)$ as a fixed point subalgebra in $gl(N)$ with respect to an involutive automorphism σ. The twisted Yangian is a deformation of the universal enveloping algebra $U(gl(N)[t]^\sigma)$, where

a)Electronic-mail: alexm@maths.usyd.edu.au

 © 1998 American Institute of Physics

$$\mathrm{gl}(N)[t]^{\sigma}=\{A(t)\in \mathrm{gl}(N)[t]\,|\,\sigma(A(t))=A(-t)\}.$$

We denote by $\mathrm{Y}^{+}(N)$ and $\mathrm{Y}^{-}(N)$ the twisted Yangians corresponding to the orthogonal and symplectic Lie algebras, respectively. Similar to the A series case the algebra $\mathrm{Y}^{\pm}(N)$ contains the corresponding universal enveloping algebra $\mathrm{U}(\mathrm{g}(N))$ and one has a natural algebra homomorphism $\mathrm{Y}^{\pm}(N)\to\mathrm{U}(\mathrm{g}(N))$ identical on $\mathrm{U}(\mathrm{g}(N))$.

A detailed description of the algebraic structure of the Yangian $\mathrm{Y}(N)$ and the twisted Yangians $\mathrm{Y}^{\pm}(N)$ is contained in Ref. 30. An explicit formula for the formal series (the Sklyanin determinant) whose coefficients generate the center of $\mathrm{Y}^{\pm}(N)$ was given in Ref. 31. Bethe-type commutative subalgebras in $\mathrm{Y}^{\pm}(N)$ were constructed in Ref. 32. These results were applied in Refs. 19, 31, and 33 to construct Laplace operators for the orthogonal and symplectic Lie algebras and to get characteristic identities and Capelli identities.

In this paper we give a complete description of the finite-dimensional irreducible representations for all twisted Yangians. Our approach is very close to the $\mathrm{gl}(N)$-case and we use some ideas and results of that case in our constructions.

Section II contains definitions and preliminary results on the Yangian $\mathrm{Y}(N)$ and the twisted Yangians $\mathrm{Y}^{\pm}(N)$.

In Sec. III we give detailed proofs of Drinfeld's and Tarasov's theorems describing the finite-dimensional irreducible representations of $\mathrm{Y}(N)$. These representations can be constructed as follows. First, using the homomorphism $\mathrm{Y}(N)\to\mathrm{U}(\mathrm{gl}(N))$ one extends any finite-dimensional irreducible representation $L(\lambda)$ of the Lie algebra $\mathrm{gl}(N)$ to the Yangian $\mathrm{Y}(N)$. Using the Hopf algebra structure on $\mathrm{Y}(N)$ one can consider representations of the form

$$L(\lambda^{(1)})\otimes\cdots\otimes L(\lambda^{(k)}). \tag{1.1}$$

Then any finite-dimensional irreducible representation of $\mathrm{Y}(\mathrm{sl}(N))$ [or $\mathrm{Y}(N)$, up to tensoring with one-dimensional representations of the center of $\mathrm{Y}(N)$] can be realized as the irreducible quotient of the cyclic span in (1.1) of the tensor product of the highest weight vectors of the $L(\lambda^{(i)})$. In the case $N=2$ there is no need to use quotients as any finite-dimensional irreducible representation of $\mathrm{Y}(\mathrm{sl}(2))$ is isomorphic to a tensor product (1.1). A criterion of irreducibility of the representation (1.1) of $\mathrm{Y}(\mathrm{sl}(2))$ in terms of ''strings'' (subsets in $\mathbb{C}$ of the form $\{a,a+1,...,b\}$) was given by Chari and Pressley.[11,13] We give a proof of this result below (Corollary 3.11).

Alternatively, every finite-dimensional irreducible representation of $\mathrm{Y}(N)$ is highest weight (see Definition 3.1). It is isomorphic to a representation of the form $L(\lambda(u))$, where $\lambda(u)=(\lambda_1(u),...,\lambda_N(u))$ is a collection of formal series in u^{-1}. Moreover, given a set $\lambda(u)$ there exists a unique (up to an isomorphism) irreducible highest weight module $L(\lambda(u))$. Drinfeld's theorem[2] (see Theorem 3.12 below) provides necessary and sufficient conditions for $L(\lambda(u))$ to be finite-dimensional.

In Secs. IV–VII we give analogs of these descriptions for the twisted Yangians. Namely, using the homomorphism $\mathrm{Y}^{\pm}(N)\to\mathrm{U}(\mathrm{g}(N))$ we extend any finite-dimensional irreducible representation $V(\mu)$ of the Lie algebra $\mathrm{g}(N)$ to the twisted Yangian $\mathrm{Y}^{\pm}(N)$. The subalgebra $\mathrm{Y}^{\pm}(N)\subset\mathrm{Y}(N)$ is a left coideal,

$$\Delta(\mathrm{Y}^{\pm}(N))\subset\mathrm{Y}(N)\otimes\mathrm{Y}^{\pm}(N), \tag{1.2}$$

and so, we may construct representations of $\mathrm{Y}^{\pm}(N)$ of the form

$$L(\lambda^{(1)})\otimes\cdots\otimes L(\lambda^{(k)})\otimes V(\mu). \tag{1.3}$$

As in the $\mathrm{gl}(N)$-case, any finite-dimensional irreducible representation of $\mathrm{Y}^{\pm}(N)$ (up to tensoring with one-dimensional representations of the center) can be realized as the irreducible quotient of the cyclic span in (1.3) of the tensor product of the highest weight vectors of the $L(\lambda^{(i)})$ and $V(\mu)$.

On the other hand, we prove that any finite-dimensional irreducible representations of $\mathrm{Y}^{\pm}(N)$ is highest weight, that is, isomorphic to a representation of the form $V(\mu(u))$ (see Definition 4.1), and we find necessary and sufficient conditions for $V(\mu(u))$ to be finite-dimensional; see Theorems 5.8, 6.9, 7.7 for the cases of C, D, B series, respectively.

The key part of the proof is a detailed study of the simplest cases $\mathrm{Y}^{\pm}(2)$ and $\mathrm{Y}^{+}(3)$. We use a modified Tarasov's argument[8,9] and also use some ideas from Refs. 11 and 13, as well as Refs. 34 and 35 where representations of the super-Yangians were studied. We give explicit realizations for all finite-dimensional irreducible representations of the simplest twisted Yangians $\mathrm{Y}^{-}(2)$ and $\mathrm{Y}^{+}(2)$ and prove analogs of Chari–Pressley's theorem for these algebras; see Corollaries 5.5, 5.7 and Corollaries 6.5, 6.6, respectively. The *B* series case requires an extra care. We construct raising and lowering operators in a $\mathrm{Y}^{+}(3)$-module $V(\mu(u))$ which preserve the subspace of $\mathrm{Y}^{+}(2)$-singular vectors. To get the necessary conditions for $V(\mu(u))$ to be finite-dimensional we examine the weights of the $\mathrm{Y}^{+}(2)$-singular vectors and use the results for the case of $\mathrm{Y}^{+}(2)$.

Some of the results of this paper [mainly for the case of $\mathrm{Y}^{-}(2n)$] were announced without complete proofs in Ref. 36.

II. DEFINITIONS AND PRELIMINARY RESULTS

Our main reference for the algebraic structure of the Yangian $\mathrm{Y}(N)$ and the twisted Yangians $\mathrm{Y}^{\pm}(N)$ is the expository paper—Ref. 30. In this section we give necessary definitions and formulate some results to be used in the subsequent sections.

We shall assume that given positive integer N, indices i,j run through the set $\{-n,\dots,-1,0,1,\dots,n\}$ if $N=2n+1$ and run through the set $\{-n,\dots,-1,1,\dots,n\}$ if $N=2n$.

The *Yangian* $\mathrm{Y}(N)=\mathrm{Y}(\mathrm{gl}(N))$ is the complex associative algebra with the generators $t_{ij}^{(1)},t_{ij}^{(2)},\dots$ where $-n\leqslant i,j\leqslant n$, and the defining relations

$$[t_{ij}^{(r+1)},t_{kl}^{(s)}]-[t_{ij}^{(r)},t_{kl}^{(s+1)}]=t_{kj}^{(r)}t_{il}^{(s)}-t_{kj}^{(s)}t_{il}^{(r)},$$

where $r,s=0,1,2,\dots$ and $t_{ij}^{(0)}:=\delta_{ij}$. For any i,j define the formal power series

$$t_{ij}(u)=\delta_{ij}+t_{ij}^{(1)}u^{-1}+t_{ij}^{(2)}u^{-2}+\cdots\in\mathrm{Y}(N)[[u^{-1}]].$$

Then the defining relations can be also written as follows:

$$[t_{ij}(u),t_{kl}(v)]=\frac{1}{u-v}\,(t_{kj}(u)t_{il}(v)-t_{kj}(v)t_{il}(u)). \tag{2.1}$$

They can also be written as a single *ternary relation* (see, e.g., Ref. 30, Sec. 1) for the *T-matrix*,

$$T(u):=\sum_{i,j}t_{ij}(u)\otimes E_{ij}\in\mathrm{Y}(N)[[u^{-1}]]\otimes\mathrm{End}\mathbb{C}^N,$$

where the E_{ij} are the standard matrix units. It has the form

$$R(u-v)T_1(u)T_2(v)=T_2(v)T_1(u)R(u-v). \tag{2.2}$$

Here $T_1(u)$ and $T_2(u)$ are regarded as elements of $\mathrm{Y}(N)[[u^{-1}]]\otimes\mathrm{End}\mathbb{C}^N\otimes\mathrm{End}\mathbb{C}^N$, the subindex of $T(u)$ indicates to which copy of $\mathrm{End}\mathbb{C}^N$ this matrix corresponds; $R(u)=1-u^{-1}P$, where P is the permutation operator in $\mathbb{C}^N\otimes\mathbb{C}^N$.

The following analog of the Poincaré–Birkhoff–Witt theorem holds for the algebra $\mathrm{Y}(N)$ (Ref. 30, Corollary 1.23):

Given an arbitrary linear order on the set of generators $\{t_{ij}^{(r)}\}$, *any element of the algebra* $\mathrm{Y}(N)$ *is uniquely written as a linear combination of ordered monomials in the generators.*

In what follows we shall always assume that a linear order $\prec$ is chosen in such a way that

$$t_{i_1j_1}^{(r_1)}\prec t_{i_2j_2}^{(r_2)}\prec t_{i_3j_3}^{(r_3)} \tag{2.3}$$

as soon as $i_1>j_1$, $i_2=j_2$, and $i_3<j_3$.

We shall regard the set of the matrix units $\{E_{ij}\}$ as a basis of the Lie algebra $\mathrm{gl}(N)$. The mapping

$$E_{ij}\mapsto t_{ij}^{(1)} \tag{2.4}$$

defines an inclusion $\mathrm{U}(\mathrm{gl}(N))\hookrightarrow \mathrm{Y}(N)$. The mapping

$$t_{ij}(u)\mapsto \delta_{ij}+E_{ij}u^{-1} \tag{2.5}$$

defines an algebra homomorphism $\mathrm{Y}(N)\to \mathrm{U}(\mathrm{gl}(N))$.

The *quantum determinant* $\mathrm{qdet}\, T(u)$ of the matrix $T(u)$ is a formal series in u^{-1} with coefficients from $\mathrm{Y}(N)$ defined as follows:

$$\mathrm{qdet}\, T(u)=\sum_{p\in \mathfrak{S}_N} \mathrm{sgn}(p)t_{p(-n),-n}(u)\cdots t_{p(n),n}(u-N+1),$$

where $\mathfrak{S}_N$ is the group of permutations of the indices $\{-n,\ldots,n\}$. The coefficients of the quantum determinant $\mathrm{qdet}\, T(u)$ are algebraically independent generators of the center of the algebra $\mathrm{Y}(N)$.

The *quantum comatrix* $\hat{T}(u)=(\hat{t}_{ij}(u))$ is defined by the following formula:

$$\mathrm{qdet}\, T(u)=\hat{T}(u)T(u-N+1). \tag{2.6}$$

The matrix element $\hat{t}_{ij}(u)$ equals $(-1)^{i+j}$ times the quantum determinant of the submatrix of $T(u)$ obtained by removing the ith column and jth row.

Each of the following mappings defines an antiautomorphism of the algebra $\mathrm{Y}(N)$: the change of sign in u,

$$\sigma: T(u)\mapsto T(-u); \tag{2.7}$$

a matrix transposition,

$$T(u)\mapsto T^{t}(u); \tag{2.8}$$

the comatrix transformation,

$$T(u)\mapsto \hat{T}(u). \tag{2.9}$$

The shift in u,

$$T(u)\mapsto T(u+a),\quad a\in\mathbb{C} \tag{2.10}$$

and the multiplication by a formal series

$$T(u)\mapsto \varphi(u)T(u),\quad \varphi(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]] \tag{2.11}$$

define automorphisms of $\mathrm{Y}(N)$.

The *Yangian* $\mathrm{Y}(\mathrm{sl}(N))$ for the Lie algebra $\mathrm{sl}(N)$ can be defined as the subalgebra of $\mathrm{Y}(N)$ consisting of the elements fixed by all automorphisms of the form (2.11). One has a tensor product decomposition

$$\mathrm{Y}(N)\simeq \mathrm{Z}(N)\otimes \mathrm{Y}(\mathrm{sl}(N)), \tag{2.12}$$

where $\mathrm{Z}(N)$ denotes the center of $\mathrm{Y}(N)$.

Using the automorphism of $\mathrm{Y}(N)$ of the form $T(u)\mapsto T^{t}(-u)$, one can show that the following identity holds:[5]

$$\mathrm{qdet}\, T(u)=\hat{T}^{t}(u-1)T^{t}(u). \tag{2.13}$$

The coproduct $\Delta:\mathrm{Y}(N)\to \mathrm{Y}(N)\otimes \mathrm{Y}(N)$ is defined by

$$\Delta(t_{ij}(u))=\sum_{a} t_{ia}(u)\otimes t_{aj}(u). \tag{2.14}$$

Let us now define the *twisted Yangians* $\mathrm{Y}^+(N)$ and $\mathrm{Y}^-(N)$ corresponding to the orthogonal Lie algebra $\mathrm{o}(N)$ and symplectic Lie algebra $\mathrm{sp}(N)$, respectively. To consider both cases simultaneously, it will be convenient to use the symbol θ_{ij} which is defined as follows:

$$\theta_{ij} := \begin{cases} 1, & \text{in the orthogonal case} \\ \mathrm{sgn}(i)\mathrm{sgn}(j), & \text{in the symplectic case.} \end{cases}$$

Whenever the double sign $\pm$ or $\mp$ occurs, the upper sign corresponds to the orthogonal case and the lower sign to the symplectic case. By $X \mapsto X^t$ we will denote the matrix transposition such that $(E_{ij})^t = \theta_{ij} E_{-j,-i}$. Let us introduce the *S-matrix* $S(u) = (s_{ij}(u))$ by setting $S(u) := T(u)T^t(-u)$, or, in terms of matrix elements,

$$s_{ij}(u) = \sum_a \theta_{aj} t_{ia}(u) t_{-j,-a}(-u). \tag{2.15}$$

Write $s_{ij}(u) = \delta_{ij} + s_{ij}^{(1)} u^{-1} + s_{ij}^{(2)} u^{-2} + \cdots$. The twisted Yangian $\mathrm{Y}^{\pm}(N)$ is the subalgebra of $\mathrm{Y}(N)$ generated by the elements $s_{ij}^{(1)}, s_{ij}^{(2)}, \ldots$, where $-n \leqslant i,j \leqslant n$. One can show that the *S*-matrix satisfies the following *quaternary relation* and *symmetry relation* (see Ref. 30, Sec. 3):

$$\begin{gathered} R(u-v)S_1(u)R^t(-u-v)S_2(v) = S_2(v)R^t(-u-v)S_1(u)R(u-v), \\ S^t(-u) = S(u) \pm \frac{S(u) - S(-u)}{2u}. \end{gathered} \tag{2.16}$$

Here we use the same notation as for the ternary relation (2.2), where $R^t(u)$ is obtained from $R(u)$ by applying the transposition t in either of the two copies of $\mathrm{End}\mathbb{C}^N$. The relations (2.16) are defining relations for the algebra $\mathrm{Y}^{\pm}(N)$ and they can be rewritten in terms of the generating series $s_{ij}(u)$ as follows:

$$\begin{aligned} [s_{ij}(u), s_{kl}(v)] = {} & \frac{1}{u-v} \left(s_{kj}(u)s_{il}(v) - s_{kj}(v)s_{il}(u) \right) \\ & - \frac{1}{u+v} \left(\theta_{k,-j} s_{i,-k}(u) s_{-j,l}(v) - \theta_{i,-l} s_{k,-i}(v) s_{-l,j}(u) \right) \\ & + \frac{1}{u^2 - v^2} \left(\theta_{i,-j} s_{k,-i}(u) s_{-j,l}(v) - \theta_{i,-j} s_{k,-i}(v) s_{-j,l}(u) \right), \end{aligned} \tag{2.17}$$

and

$$\theta_{ij} s_{-j,-i}(-u) = s_{ij}(u) \pm \frac{s_{ij}(u) - s_{ij}(-u)}{2u}. \tag{2.18}$$

The elements

$$s_{ij}^{(2k)}, \quad i+j \geqslant 0, \qquad s_{ij}^{(2k-1)}, \quad i+j > 0, \qquad k = 1,2,..., \tag{2.19}$$

in the case of $\mathrm{Y}^+(N)$, and the elements

$$s_{ij}^{(2k)}, \quad i+j > 0, \qquad s_{ij}^{(2k-1)}, \quad i+j \geqslant 0, \qquad k = 1,2,..., \tag{2.20}$$

in the case of $\mathrm{Y}^-(N)$, form a system of linearly independent generators. An analog of the Poincaré–Birkhoff–Witt theorem for the algebra $\mathrm{Y}^{\pm}(N)$ can be formulated as follows (Ref. 30, Remark 3.14):

Given an arbitrary linear order on the set of the above generators, any element of the algebra $\mathrm{Y}^{\pm}(N)$ *is uniquely written as a linear combination of ordered monomials in the generators.*

Similar to the case of $\mathrm{Y}(N)$ we shall always assume that a linear order $\prec$ is chosen in such a way that

$$s_{i_1 j_1}^{(r_1)} \prec s_{i_2 j_2}^{(r_2)} \prec s_{i_3 j_3}^{(r_3)} \tag{2.21}$$

as soon as $i_1 > j_1$, $i_2 = j_2$, and $i_3 < j_3$.

Introduce the following elements of the Lie algebra $\mathrm{gl}(N)$:

$$F_{ij} = E_{ij} - \theta_{ij} E_{-j,-i}, \quad -n \leqslant i,j \leqslant n.$$

Denote by $\mathrm{g}(N)$ the Lie subalgebra of $\mathrm{gl}(N)$ spanned by the elements F_{ij}. Then $\mathrm{g}(N)$ is isomorphic to $\mathrm{o}(2n)$ or $\mathrm{sp}(2n)$ if $N = 2n$, and to $\mathrm{o}(2n+1)$ if $N = 2n+1$.

The mapping

$$F_{ij} \mapsto s_{ij}^{(1)} \tag{2.22}$$

defines an inclusion $\mathrm{U}(\mathrm{g}(N)) \hookrightarrow \mathrm{Y}^{\pm}(N)$. The mapping

$$s_{ij}(u) \mapsto \delta_{ij} + F_{ij}(u \pm \tfrac{1}{2})^{-1} \tag{2.23}$$

defines an algebra homomorphism $\mathrm{Y}^{\pm}(N) \to \mathrm{U}(\mathrm{g}(N))$.

There is an analog of the quantum determinant for the twisted Yangians. It is denoted by $\mathrm{sdet}\, S(u)$ and is called the *Sklyanin determinant*. This is a formal series in u^{-1} and its coefficients generate the center of the algebra $\mathrm{Y}^{\pm}(N)$. The Sklyanin determinant is related to the quantum determinant by

$$\mathrm{sdet}\, S(u) = \gamma_N(u)\ \ \mathrm{qdet}\, T(u)\ \ \mathrm{qdet}\, T(-u+N-1), \tag{2.24}$$

where $\gamma_N(u) \equiv 1$ for $\mathrm{Y}^{+}(N)$ and $\gamma_N(u) = (2u+1)/(2u-N+1)$ for $\mathrm{Y}^{-}(N)$. An explicit determinant-type expression for $\mathrm{sdet}\, S(u)$ in terms of the generators $s_{ij}(u)$ is given in Ref. 31.

Define the *Sklyanin comatrix* $\hat{S}(u) = (\hat{s}_{ij}(u))$ by

$$\mathrm{sdet}\, S(u) = \hat{S}(u) S(u-N+1). \tag{2.25}$$

An explicit expression for $\hat{s}_{ij}(u)$ can also be found from Ref. 31, Sec. 6.

The restriction of the antiautomorphism (2.7) to the subalgebra $\mathrm{Y}^{\pm}(N)$ yields an antiautomorphism

$$\sigma : S(u) \mapsto S^t(u). \tag{2.26}$$

We shall need the following result.

Proposition 2.1: The mapping

$$S(u) \mapsto \gamma_N(u) \hat{S}\left(-u + \frac{N}{2} - 1\right), \tag{2.27}$$

defines an automorphism of the algebra $\mathrm{Y}^{\pm}(N)$.

Proof: Note that the mapping $T(u) \mapsto \widetilde{T}(u)$ with

$$\widetilde{T}(u) = \hat{T}^t\left(u + \frac{N}{2} - 1\right) = \mathrm{qdet}\, T\left(u + \frac{N}{2}\right)\left(T^t\left(u + \frac{N}{2}\right)\right)^{-1}$$

defines an automorphism of the algebra $\mathrm{Y}(N)$. To see this, it suffices to present this map as a composition of antiautomorphisms of the form (2.8), (2.9) and an automorphism of the form (2.10). Denote by $\widetilde{S}(u)$ the image of the matrix $S(u) = T(u)T^t(-u)$ under this automorphism. That is,

$$\widetilde{S}(u)=\widetilde{T}(u)\widetilde{T}^{t}(-u)=\hat{T}^{t}\Big(u+\frac{N}{2}-1\Big)\hat{T}\Big(-u+\frac{N}{2}-1\Big).$$

By (2.6) and (2.13) we have

$$\widetilde{S}\Big(-u+\frac{N}{2}-1\Big)S(u-N+1)=\hat{T}^{t}(-u+N-2)\hat{T}(u)T(u-N+1)T^{t}(-u+N-1)$$

$$=\mathrm{qdet}\ T(u)\hat{T}^{t}(-u+N-2)T^{t}(-u+N-1)=\mathrm{qdet}\ T(u)\ \ \mathrm{qdet}\ T(-u+N-1),$$

which equals $\gamma_N^{-1}(u)$ sdet $S(u)$ by (2.24). Note that

$$\gamma_N(u)\,\gamma_N\Big(-u+\frac{N}{2}-1\Big)=1,$$

and so, using (2.25) we obtain that

$$\widetilde{S}(u)=\gamma_N(u)\hat{S}\Big(-u+\frac{N}{2}-1\Big).$$

Thus, the subalgebra $\mathrm{Y}^{\pm}(N)\subset\mathrm{Y}(N)$ is stable under the automorphism $T(u)\mapsto\widetilde{T}(u)$ and its restriction to $\mathrm{Y}^{\pm}(N)$ yields the automorphism (2.27) which completes the proof. □

We shall use the following commutation relations between matrix elements of the matrices $\hat{S}(u)$ and $S(u)$:

$$[\hat{s}_{ij}(u),s_{kl}(v)]=-\frac{1}{u-v-N+1}\sum_a\ (\delta_{kj}\hat{s}_{ia}(u)s_{al}(v)-\delta_{il}s_{ka}(v)\hat{s}_{aj}(u))$$

$$+\frac{1}{u+v+1}\sum_a\ (\delta_{i,-k}\theta_{ak}\hat{s}_{-a,j}(u)s_{al}(v)-\delta_{j,-l}\theta_{al}s_{ka}(v)\hat{s}_{i,-a}(u))$$

$$+\frac{1}{(u-v-N+1)(u+v+1)}\sum_a\ (\delta_{i,-k}\theta_{jk}\hat{s}_{-j,a}(u)s_{al}(v)$$

$$-\delta_{j,-l}\theta_{il}s_{ka}(v)\hat{s}_{a,-i}(u)). \tag{2.28}$$

They easily follow from the quaternary relation; see (2.16). Indeed, inverting its both sides and multiplying from the left and right by $S_2(v)$ we obtain

$$R^{t}(u+v+N)S_1^{-1}(u)R(v-u)S_2(v)=S_2(v)R(v-u)S_1^{-1}(u)R^{t}(u+v+N).$$

To get (2.28) rewrite this first in terms of the matrix elements of $S^{-1}(u)$ and $S(u)$, then use (2.25) and the centrality of the Sklyanin determinant in $\mathrm{Y}^{\pm}(N)$. (In a slightly different form this calculation was performed in Ref. 30, Sec. 7).

Multiplication by an even formal series

$$S(u)\mapsto\psi(u)S(u),\quad \psi(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]] \tag{2.29}$$

defines an automorphism of $\mathrm{Y}^{\pm}(N)$.

The *special twisted Yangian* $\mathrm{SY}^{\pm}(N)$ is defined by the equality $\mathrm{SY}^{\pm}(N)=\mathrm{Y}(\mathrm{sl}(N))\cap\mathrm{Y}^{\pm}(N)$. In other words, $\mathrm{SY}^{\pm}(N)$ is the subalgebra of $\mathrm{Y}^{\pm}(N)$ consisting of elements which are stable under all automorphisms of the form (2.29). One has the following tensor product decomposition analogous to (2.12):

$$\mathrm{Y}^{\pm}(N)\simeq\mathrm{Z}^{\pm}(N)\otimes\mathrm{SY}^{\pm}(N), \tag{2.30}$$

where $\mathrm{Z}^{\pm}(N)$ denotes the center of $\mathrm{Y}^{\pm}(N)$.

The images of the generators $s_{ij}(u)$ under coproduct Δ are given by

$$\Delta(s_{ij}(u))=\sum_{a,b}\ \theta_{bj}t_{ia}(u)t_{-j,-b}(-u)\otimes s_{ab}(u). \tag{2.31}$$

This implies that the subalgebra $\mathrm{Y}^{\pm}(N)\subset\mathrm{Y}(N)$ is a left coideal; see (1.2).

III. FINITE-DIMENSIONAL IRREDUCIBLE REPRESENTATIONS OF Y(N) and Y(sl(N))

For this section only we adopt a more standard enumeration of the generators $t_{ij}(u)$ of the Yangian $\mathrm{Y}(N)$: We let the indices i,j run through the set $\{1,...,N\}$. All formulas from the previous section should be understood accordingly. In the following definitions we follow Ref. 2 and Refs. 11–13.

Definition 3.1: A representation L of the Yangian $\mathrm{Y}(N)$ is called *highest weight* if there exists a nonzero vector $\xi\in L$ such that L is generated by ξ,

$$t_{ij}(u)\xi=0 \quad \text{for } 1\leqslant i<j\leqslant N, \tag{3.1}$$

and

$$t_{ii}(u)\xi=\lambda_i(u)\xi \quad \text{for } 1\leqslant i\leqslant N \tag{3.2}$$

for some formal series $\lambda_i(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$. The vector ξ is called the *highest weight vector* of L and the set $\lambda(u):=(\lambda_1(u),\ldots,\lambda_N(u))$ is the *highest weight* of L. □

Sometimes we shall also refer to L as a representation *with the highest weight* $\lambda(u)$. The definition of a *lowest weight representation* is obtained by replacing (3.1) with the relations

$$t_{ji}(u)\xi=0 \quad \text{for } 1\leqslant i<j\leqslant N.$$

Definition 3.2: Let $\lambda(u)=(\lambda_1(u),\ldots,\lambda_N(u))$ be any set of formal series where $\lambda_i(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$. The *Verma module* $M(\lambda(u))$ is the quotient of $\mathrm{Y}(N)$ by the left ideal generated by all the coefficients of the series $t_{ij}(u)$ for $1\leqslant i<j\leqslant N$ and $t_{ii}(u)$-$\lambda_i(u)$ for $1\leqslant i\leqslant N$. □

The following properties of $M(\lambda(u))$ are immediate from the Poincaré–Birkhoff–Witt theorem for $\mathrm{Y}(N)$. First, ordered monomials in the coefficients of the series $t_{ij}(u)$ with $i>j$ [assuming that (2.3) is satisfied] form a basis of $M(\lambda(u))$. The Verma module $M(\lambda(u))$ is obviously a representation of $\mathrm{Y}(N)$ with the highest weight $\lambda(u)$ and any representation with the highest weight $\lambda(u)$ is isomorphic to a quotient of $M(\lambda(u))$. Further, $M(\lambda(u))$ has a unique irreducible quotient which we denote by $L(\lambda(u))$. Indeed, the sum of all submodules of $M(\lambda(u))$ which do not contain the highest weight vector 1 is a unique maximal proper submodule K, and $L(\lambda(u))$ is the quotient $M(\lambda(u))/K$. Moreover, if an irreducible $\mathrm{Y}(N)$-module L contains a nonzero vector ξ satisfying (3.1) and (3.2) then L is isomorphic to $L(\lambda(u))$.

Using the inclusion (2.4) we may regard any $\mathrm{Y}(N)$-module as a $\mathrm{gl}(N)$-module. Relations (2.1) imply that

$$[E_{ij},t_{kl}(u)]=\delta_{kj}t_{il}(u)-\delta_{il}t_{kj}(u), \tag{3.3}$$

where we identify the elements $E_{ij}\in\mathrm{gl}(N)$ with their images in $\mathrm{Y}(N)$. Let h be the diagonal subalgebra in $\mathrm{gl}(N)$ with the basis $E_{11},\ldots,E_{NN}$ and let $\epsilon_1,\ldots,\epsilon_N$ be the dual basis in h*. Choose the standard system of positive roots R_+ for $\mathrm{gl}(N)$ corresponding to the upper triangular Borel subalgebra. Then the elements $\alpha_1=\epsilon_1-\epsilon_2,\ldots,\alpha_{N-1}=\epsilon_{N-1}-\epsilon_N$ form a basis in R_+. If L is a representation of $\mathrm{Y}(N)$ with the highest weight $\lambda(u)$, and ξ is its highest weight vector, then ξ is also a $\mathrm{gl}(N)$-highest weight vector of the $\mathrm{gl}(N)$-weight $\lambda=(\lambda_1,\ldots,\lambda_N)$ where λ_i is the coefficient at u^{-1} in the series $\lambda_i(u)$. By (3.3) all $\mathrm{gl}(N)$-weights of L have the form $\lambda-\omega$, where ω is a linear combination of the α_i with coefficients from $\mathbb{Z}_+$. Moreover, the subspace in L of vectors of $\mathrm{gl}(N)$-weight λ if one-dimensional and spanned by ξ.

The following result is contained in Ref. 2.

Theorem 3.3: *Every finite-dimensional irreducible representation L of $\mathrm{Y}(N)$ is highest weight. Moreover, L contains a unique (up to scalar multiples) highest weight vector.*

Proof: The $\mathrm{gl}(N)$-module L admits a weight space decomposition

$$L=\bigoplus_{\lambda\in \mathrm{h}^*} L_\lambda, \quad L_\lambda=\{\eta\in L\,|\,E_{ii}\eta=\lambda_i\eta,\ \ i=1,\ldots,N\}. \tag{3.4}$$

To see this, we note that by (3.3) the direct sum L' of all weight subspaces in L is a $\mathrm{Y}(N)$-submodule. This submodule is nontrivial because L is finite-dimensional and so it contains an eigenvector for the commutative subalgebra h. Since L is irreducible as a $\mathrm{Y}(N)$-module we conclude that $L'=L$.

Introduce the subspace in L

$$L^0:=\{\eta\in L\,|\,t_{ij}(u)\,\eta=0 \quad \text{for } 1\leqslant i<j\leqslant N\}. \tag{3.5}$$

By (3.3) the weight space decomposition (3.4) induces that of L^0:

$$L^0=\bigoplus_{\lambda\in \mathrm{h}^*} L^0_\lambda, \quad L^0_\lambda=L^0\cap L_\lambda. \tag{3.6}$$

Since the $\mathrm{gl}(N)$-module L is finite-dimensional there exists a nonzero vector $\eta_0\in L$ of a $\mathrm{gl}(N)$-weight λ' such that $\lambda'+\alpha$ is not a weight of L for any $\alpha\in R_+$. We see from (3.3) that the vectors $t_{ij}^{(r)}\eta_0$ with $i<j$ and $r\geqslant 1$ have the $\mathrm{gl}(N)$-weight $\lambda'+\epsilon_i-\epsilon_j$ and therefore are equal to zero. This proves that the subspace L^0 is nontrivial.

Relations (2.1) and (3.3) imply that $\mathrm{Y}(N)$ can be equipped with a $\mathbb{Z}^N$-grading

$$\mathrm{Y}(N)=\bigoplus_{\nu\in\mathbb{Z}^N} \mathrm{Y}(N)_\nu,$$

where

$$\mathrm{Y}(N)_\nu=\{y\in \mathrm{Y}(N)\,|\,[E_{ii},y]=\nu_i y,\ \ i=1,\ldots,N\}.$$

In particular, $\mathrm{Y}(N)_0$ is a subalgebra in $\mathrm{Y}(N)$. Denote by I and I' the left and right ideals in $\mathrm{Y}(N)$, respectively, generated by the coefficients of the series $t_{ij}(u)$ and $t_{ji}(u)$, where $1\leqslant i<j\leqslant N$. Set $I_0=\mathrm{Y}(N)_0\cap I$. Exactly as in the case of semisimple Lie algebras (see, e.g., Ref. 37, Sec. 7.4) one derives from the Poincaré–Birkhoff–Witt theorem for $\mathrm{Y}(N)$ that $I_0=I'\cap Y(N)_0$ and that I_0 is a two-sided ideal in $\mathrm{Y}(N)_0$. Moreover, $\mathrm{Y}(N)_0=\mathrm{Y}\oplus I_0$, where Y is the subspace in $\mathrm{Y}(N)_0$ with the basis formed by ordered monomials in the generators $t_{ii}^{(r)}$. Consider the algebra $\mathscr{Y}=Y(N)_0/I_0$. We see from (2.1) that $[t_{ii}(u),t_{ii}(v)]=0$ and

$$[t_{ii}(u),t_{jj}(v)]=\frac{1}{u-v}\,(t_{ji}(u)t_{ij}(v)-t_{ji}(v)t_{ij}(u)).$$

So, if $i<j$ each commutator $[t_{ii}^{(r)},t_{jj}^{(s)}]$ belongs to I_0. This implies that the algebra $\mathscr{Y}$ is commutative and freely generated by the elements $t_{ii}^{(r)}$ mod I_0.

Further, we note that the subspace L^0 has a natural structure of a $\mathrm{Y}(N)_0$-module and hence also of a $\mathscr{Y}$-module because L^0 is annihilated by I_0. Therefore, L^0 contains a nonzero vector ξ which is a common eigenvector for all operators $t_{ii}^{(r)}$. So, ξ generates a highest weight $\mathrm{Y}(N)$-submodule in L. Since L is irreducible this submodule coincides with L which proves that L is highest weight.

Finally, if ξ has a $\mathrm{gl}(N)$-weight λ_0 then nonzero weight subspaces in (3.6) can only correspond to the $\mathrm{gl}(N)$-weights λ of the form $\lambda_0-\omega$, where ω is a $\mathbb{Z}_+$-linear combination of the α_i. Moreover, the subspace $L^0_{\lambda_0}$ is one-dimensional. However, each subspace L^0_λ is $\mathscr{Y}$-invariant. If it is nonzero for a certain $\lambda\neq\lambda_0$ then a common eigenvector in L^0_λ for all $t_{ii}^{(r)}$ generates a proper submodule is L, which makes a contradiction. □

Remark 3.4: It follows from the proof that if L is an irreducible highest weight representation of $\mathrm{Y}(N)$ (not necessary finite-dimensional) then the subspace (3.5) is one-dimensional and spanned by the highest weight vector ξ of L. □

Due to Theorem 3.3, to describe the finite-dimensional irreducible representations of $\mathrm{Y}(N)$ it suffices to find necessary and sufficient conditions for an irreducible highest weight representation

$L(\lambda(u))$ to be finite-dimensional. The case of the simplest Yangian Y(2) plays a key role in the proof. We shall be assuming now that $N=2$. The following result is due to Tarasov.[8,9]

Proposition 3.5: If $\dim L(\lambda(u))<\infty$ *then there exists a formal series* $\varphi(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$ *such that* $\varphi(u)\lambda_1(u)$ *and* $\varphi(u)\lambda_2(u)$ *are polynomials in* u^{-1}.

Proof: Let ξ denote the highest weight vector of $L(\lambda(u))$. Since $L(\lambda(u))$ is finite-dimensional there exists a non-negative integer k such that the vector $t_{21}^{(k+1)}\xi$ is a linear combination of $t_{21}^{(1)}\xi,\dots,t_{21}^{(k)}\xi$. Take the minimum k with this property. Let us show that then for any vector $t_{21}^{(r)}\xi$ with $r\geqslant k+1$ we have

$$t_{21}^{(r)}\xi=a_1^{(r)}\xi_1+\cdots+a_k^{(r)}\xi_k \tag{3.7}$$

for some complex coefficients $a_i^{(r)}$, where $\xi_i:=t_{21}^{(i)}\xi$, $i=1,\dots,k$. Indeed, this is true for $r=k+1$ by our choice of k. Relations (2.1) imply that

$$t_{11}^{(2)}t_{21}(u)=t_{21}(u)(t_{11}^{(2)}-t_{11}^{(1)}-u)+t_{21}^{(1)}t_{11}(u), \tag{3.8}$$

and taking the coefficient at u^{-p} we obtain

$$t_{11}^{(2)}t_{21}^{(p)}=-t_{21}^{(p+1)}+t_{21}^{(1)}t_{11}^{(p)}+t_{21}^{(p)}(t_{11}^{(2)}-t_{11}^{(1)}). \tag{3.9}$$

Write $\lambda_1(u)=1+\lambda_1^{(1)}u^{-1}+\lambda_1^{(2)}u^{-2}+\cdots$. Then by (3.9), for any $p\geqslant 1$,

$$t_{21}^{(p+1)}\xi=-t_{11}^{(2)}t_{21}^{(p)}\xi+(\lambda_1^{(2)}-\lambda_1^{(1)})t_{21}^{(p)}\xi+\lambda_1^{(p)}t_{21}^{(1)}\xi. \tag{3.10}$$

Hence, for $p=1,\dots,k-1$ we have

$$t_{11}^{(2)}\xi_p=\lambda_1^{(p)}\xi_1+(\lambda_1^{(2)}-\lambda_1^{(1)})\xi_p-\xi_{p+1} \tag{3.11}$$

and applying (3.7) with $r=k+1$, we get

$$t_{11}^{(2)}\xi_k=\lambda_1^{(k)}\xi_1+(\lambda_1^{(2)}-\lambda_1^{(1)})\xi_k-(a_1^{(k+1)}\xi_1+\cdots+a_k^{(k+1)}\xi_k). \tag{3.12}$$

Now (3.7) follows from (3.10) to (3.12) by an obvious induction on r.

Rewriting (3.7) in terms of generating series we obtain

$$t_{21}(u)\xi=a_1(u)\xi_1+\cdots+a_k(u)\xi_k, \tag{3.13}$$

where $a_i(u)=u^{-i}+a_i^{(k+1)}u^{-k-1}+\cdots$. Applying both sides of (3.8) to ξ and using (3.13) we obtain

$$(t_{11}^{(2)}-\lambda_1^{(2)}+\lambda_1^{(1)}+u)(a_1(u)\xi_1+\cdots+a_k(u)\xi_k)=\lambda_1(u)\xi_1. \tag{3.14}$$

Using (3.11) and (3.12) and taking the coefficient at ξ_p, $p=2,\dots,k$ in this relation we get

$$-a_{p-1}(u)+ua_p(u)-a_p^{(k+1)}a_k(u)=0.$$

This implies that for any $p=1,\dots,k$ one has $a_p(u)=A_p(u)a_k(u)$ for a monic polynomial $A_p(u)$ of degree $k-p$ in u. Further, taking the coefficient at ξ_1 in (3.14) we obtain $\lambda_1(u)=B(u)a_k(u)$ for a monic polynomial $B(u)$ of degree k in u.

Similarly, repeating this argument with the use of

$$t_{22}^{(2)}t_{21}(u)=t_{21}(u)(t_{22}^{(2)}+t_{22}^{(1)}+u)-t_{21}^{(1)}t_{22}(u)$$

instead of (3.8) we obtain that $\lambda_2(u)=C(u)a_k(u)$ for a monic polynomial $C(u)$ of degree k in u. Thus, the desired series is $\varphi(u)=(a_k(u)u^k)^{-1}$. □

Given $\lambda=(\lambda_1,\dots,\lambda_N)\in\mathfrak{h}^*\simeq\mathbb{C}^N$ we shall denote by $L(\lambda)$ the irreducible $\mathrm{gl}(N)$-module with the highest weight λ with respect to the basis $\{E_{11},\dots,E_{NN}\}$ in $\mathfrak{h}$. The representation $L(\lambda)$ is finite-dimensional if and only if $\lambda_i-\lambda_{i+1}\in\mathbb{Z}_+$ for all $i=1,\dots,N-1$. The homomorphism (2.5) allows us to consider any representation $L(\lambda)$ as a $\mathrm{Y}(N)$-module. It is obviously highest weight

with $\lambda_i(u)=1+\lambda_i u^{-1}$, $i=1,...,N$. The Hopf algebra structure on $\mathrm{Y}(N)$ [see (2.14)] allows one to naturally equip any tensor product $L_1\otimes L_2$ of $\mathrm{Y}(N)$-modules with a $\mathrm{Y}(N)$-action by the rule

$$y\cdot(\xi_1\otimes\xi_2):=\Delta(y)(\xi_1\otimes\xi_2),\quad y\in Y(N),\quad \xi_i\in L_i. \tag{3.15}$$

Let us now return to the case $N=2$. A basis in the irreducible gl(2)-module $L(\alpha,\beta)$ with the highest weight (α,β) is formed by the elements $(E_{21})^r\xi$, where ξ is the highest weight vector and $r=0,...,\alpha-\beta$ if $\alpha-\beta\in\mathbb{Z}_+$; or $r\in\mathbb{Z}_+$ otherwise. Using (3.15) we can construct representations of $\mathrm{Y}(2)$ of the form

$$L=L(\alpha_1,\beta_1)\otimes\cdots\otimes L(\alpha_k,\beta_k). \tag{3.16}$$

In accordance with (2.14) the generators of $\mathrm{Y}(2)$ act by the rule

$$t_{ij}(u)(\eta_1\otimes\cdots\otimes\eta_k)=\sum_{a_1,\ldots,a_{k-1}} t_{ia_1}(u)\eta_1\otimes t_{a_1a_2}(u)\eta_2\otimes\cdots\otimes t_{a_{k-1}j}(u)\eta_k, \tag{3.17}$$

where $\eta_i\in L(\alpha_i,\beta_i)$. By (2.5) all generators $t_{ab}^{(r)}$ with $r\geqslant 2$ act as zero operators in each $L(\alpha_i,\beta_i)$ and so, by (3.17) we have for any $\eta\in L$,

$$t_{ij}^{(r)}\eta=0\quad\text{for } r\geqslant k+1. \tag{3.18}$$

For a multiset of complex numbers $\{\alpha_1,\ldots,\alpha_k,\beta_1,\ldots,\beta_k\}$ (repetitions of elements are allowed) we can find a re-enumeration of the α_i and a re-enumeration of the β_i such that the following condition is satisfied:

$$\begin{gathered}\text{for every } i=1,...,k \text{ we have: if the multiset}\\ \{\alpha_p-\beta_q\,|\,i\leqslant p,q\leqslant k\} \text{ contains non-negative integers}\\ \text{then } \alpha_i-\beta_i \text{ is minimal among them.}\end{gathered} \tag{3.19}$$

Indeed, choose first a minimal non-negative integer difference among all differences $\alpha_p-\beta_q$ if it exists. Re-enumerating the indices if necessary, we may assume that this difference is $\alpha_1-\beta_1$. Then we proceed by induction, considering the differences $\alpha_p-\beta_q$ with $p,q\geqslant 2$ and so on. The following proposition was proved in Ref. 9.

Proposition 3.6: Let $L(\lambda_1(u),\lambda_2(u))$ be an irreducible highest weight representation of $\mathrm{Y}(2)$ *and $\lambda_1(u)$ and $\lambda_2(u)$ be polynomials of degree $\leqslant$k in u^{-1}. Fix decompositions*

$$\lambda_1(u)=(1+\alpha_1u^{-1})\cdots(1+\alpha_ku^{-1}), \tag{3.20}$$

$$\lambda_2(u)=(1+\beta_1u^{-1})\cdots(1+\beta_ku^{-1}), \tag{3.21}$$

and assume that condition (3.19) is satisfied. Then $L(\lambda_1(u),\lambda_2(u))$ is isomorphic to the tensor product (3.16).

Proof: Let ξ_i be the highest weight vector of $L(\alpha_i,\beta_i)$. We have

$$t_{11}(u)\xi_i=(1+\alpha_iu^{-1})\xi_i,\quad t_{22}(u)\xi_i=(1+\beta_iu^{-1})\xi_i,\quad t_{12}(u)\xi_i=0. \tag{3.22}$$

Therefore, for the vector $\xi:=\xi_1\otimes\cdots\otimes\xi_k$ there holds

$$t_{11}(u)\xi=\lambda_1(u)\xi,\quad t_{22}(u)\xi=\lambda_2(u)\xi,\quad t_{12}(u)\xi=0, \tag{3.23}$$

which follows immediately from (3.17). Thus, the cyclic $\mathrm{Y}(2)$-span of the vector ξ in L is a highest weight module with the highest weight $(\lambda_1(u),\lambda_2(u))$. Therefore, to prove Proposition 3.6 we only need to show that the representation L is irreducible.

We use induction on k. For $k=1$ this is obvious, because $L(\alpha,\beta)$ is an irreducible gl(2)-module and hence it is also irreducible as a $\mathrm{Y}(2)$-module. Suppose now that $k>1$. Let us prove first that any nonzero submodule $\widetilde{L}$ of L contains the vector ξ. Note that $\widetilde{L}$ contains a nonzero

vector η such that $t_{12}(u)\,\eta=0$. Indeed, gl(2)-weights of L have the form $(\Sigma_i\ \alpha_i-r,\ \Sigma_i\ \beta_i+r)$ with $r\in\mathbb{Z}_+$. Therefore, there exists a gl(2)-weight (μ_1,μ_2) of $\widetilde{L}$ such that (μ_1+1,μ_2-1) is not a weight. Then a vector η of the gl(2)-weight (μ_1,μ_2) is annihilated by $t_{12}(u)$. Write η in the form

$$\eta=\sum_{r=0}^{p}(E_{21})^r\xi_1\otimes\eta_r,$$

with $p\geqslant 0$, where $\eta_r\in L(\alpha_2,\beta_2)\otimes\cdots\otimes L(\alpha_k,\beta_k)$; and if $\alpha_1-\beta_1\in\mathbb{Z}_+$ then $p\leqslant\alpha_1-\beta_1$. We may assume that $\eta_p\neq 0$. Applying $t_{12}(u)$ to η with the use of (2.14) and (3.15) we get

$$\sum_{r=0}^{p}\left(t_{11}(u)(E_{21})^r\xi_1\otimes t_{12}(u)\,\eta_r+t_{12}(u)(E_{21})^r\xi_1\otimes t_{22}(u)\,\eta_r\right)=0. \tag{3.24}$$

Using (2.5) and the commutation relations in gl(2) we can write

$$t_{11}(u)(E_{21})^r\xi_1=(1+(\alpha_1-r)u^{-1})(E_{21})^r\xi_1 \tag{3.25}$$

and

$$t_{12}(u)(E_{21})^r\xi_1=u^{-1}r(\alpha_1-\beta_1-r+1)(E_{21})^{r-1}\xi_1. \tag{3.26}$$

Taking the coefficient at $(E_{21})^p\xi_1$ in (3.24) we get

$$(1+(\alpha_1-p)u^{-1})t_{12}(u)\,\eta_p=0.$$

This implies that $t_{12}(u)\,\eta_p=0$. On the other hand, by the induction hypotheses, the representation $L(\alpha_2,\beta_2)\otimes\cdots\otimes L(\alpha_k,\beta_k)$ is irreducible, and so

$$\eta_p=\text{const}\cdot\xi_2\otimes\cdots\otimes\xi_k, \tag{3.27}$$

where ''const'' is a nonzero constant; see Remark 3.4. Suppose now that $p\geqslant 1$ and take the coefficient at $(E_{21})^{p-1}\xi_1$ in (3.24). By (3.25) and (3.26) we have

$$(1+(\alpha_1-p+1)u^{-1})t_{12}(u)\,\eta_{p-1}+u^{-1}p(\alpha_1-\beta_1-p+1)t_{22}(u)\,\eta_p=0. \tag{3.28}$$

Relations (3.27) and (3.17) imply that

$$t_{22}(u)\,\eta_p=(1+\beta_2u^{-1})\cdots(1+\beta_ku^{-1})\,\eta_p. \tag{3.29}$$

Multiply both sides of (3.28) by u^k:

$$(u+\alpha_1-p+1)u^{k-1}t_{12}(u)\,\eta_{p-1}+p(\alpha_1-\beta_1-p+1)u^{k-1}t_{22}(u)\,\eta_p=0. \tag{3.30}$$

The operators $u^{k-1}t_{ij}(u)$ in $L(\alpha_2,\beta_2)\otimes\cdots\otimes L(\alpha_k,\beta_k)$ are polynomials in u by (3.18) and we may put $u=-\alpha_1+p-1$ in (3.30). Taking into account (3.29) we get

$$p(\alpha_1-\beta_1-p+1)(\alpha_1-\beta_2-p+1)\cdots(\alpha_1-\beta_k-p+1)=0.$$

However, this is impossible because of condition (3.19). Therefore, p has to be equal to 0 and so, η is a multiple of ξ.

To complete the proof we have to show that the submodule of L generated by ξ coincides with L. Using the antiautomorphism (2.7) we equip the vector space V^* dual to a Y(2)-module V with a structure of Y(2)-module. Namely, set

$$(y\cdot f)(v)=f(\sigma(y)v),\quad y\in Y(2),\quad f\in V^*,\quad v\in V.$$

Then one can easily see that $L(\alpha_i,\beta_i)^* \simeq L'(-\alpha_i,-\beta_i)$, where $L'(-\alpha_i,-\beta_i)$ is the lowest weight representation of gl(2) whose lowest weight vector is the vector ξ_i' dual to ξ_i. This implies that

$$L^* \simeq L'(-\alpha_1,-\beta_1)\otimes\cdots\otimes L'(-\alpha_k,-\beta_k). \tag{3.31}$$

Similar to the case of representation L, any nonzero submodule of L^* contains a nonzero vector η' such that $t_{21}(u)\,\eta' = 0$. Modifying the previous argument for lowest weight representations we prove that η' coincides (up to a nonzero factor) with the vector $\xi' := \xi_1' \otimes \cdots \otimes \xi_k'$. Now, if the submodule in L generated by ξ is proper, then its annihilator in L^* is a nonzero submodule which does not contain ξ'. Contradiction. □

Proposition 3.6 allows one to get necessary and sufficient conditions for $L(\lambda(u))$ to be finite-dimensional.

Proposition 3.7: Let $L(\lambda(u))$ be an irreducible representation of Y(2) *with the highest weight $\lambda(u)=(\lambda_1(u),\lambda_2(u))$. Then $L(\lambda(u))$ is finite-dimensional if and only if there exists a formal series $\varphi(u) \in 1+u^{-1}\mathbb{C}[[u^{-1}]]$ such that $\varphi(u)\lambda_1(u)$ and $\varphi(u)\lambda_2(u)$ are polynomials in u^{-1} with decompositions*

$$\varphi(u)\lambda_1(u) = (1+\alpha_1 u^{-1})\cdots(1+\alpha_k u^{-1}), \tag{3.32}$$

$$\varphi(u)\lambda_2(u) = (1+\beta_1 u^{-1})\cdots(1+\beta_k u^{-1}), \tag{3.33}$$

and $\alpha_i-\beta_i \in \mathbb{Z}_+$, $i=1,...,k$ for certain re-enumerations of the α_i and β_i.

Proof: Let $\dim L(\lambda(u))<\infty$. By Proposition 3.5, taking the composition of our representation with an automorphism of the form (2.11) we may assume that (3.32) and (3.33) are satisfied for a certain series $\varphi(u)$. Re-enumerate the α_i and β_i in such a way that condition (3.19) is satisfied. Now use Proposition 3.6. Since the representation (3.16) is finite-dimensional this gives the relations $\alpha_i-\beta_i \in \mathbb{Z}_+$ for all $i=1,...,k$.

Conversely, let (3.32) and (3.33) hold with $\alpha_i-\beta_i \in \mathbb{Z}_+$, $i=1,...,k$. Then the representation (3.16) is finite-dimensional and the composition of $L(\lambda(u))$ with the automorphism (2.11) corresponding to the series $\varphi(u)$ is isomorphic to the irreducible quotient of the Y(2)-cyclic span in L of the vector ξ; see (3.23). Therefore, $\dim L(\lambda(u))<\infty$. □

The following reformulation of this result is due to Drinfeld.[2]

Theorem 3.8: *The irreducible highest weight representation $L(\lambda_1(u),\lambda_2(u))$ of* Y(2) *is finite-dimensional if and only if there exists a monic polynomial $P(u) \in \mathbb{C}[u]$ such that*

$$\frac{\lambda_1(u)}{\lambda_2(u)} = \frac{P(u+1)}{P(u)}. \tag{3.34}$$

In this case $P(u)$ is unique.

Proof: If $\dim L(\lambda_1(u),\lambda_2(u))<\infty$ then by Proposition 3.7 relations (3.32) and (3.33) hold for a certain series $\varphi(u)$. If $\alpha_i-\beta_i \in \mathbb{Z}_+$ for all i we put

$$P(u) = \prod_{i=1}^{k} (u+\beta_i)\cdots(u+\alpha_i-1).$$

Conversely, suppose (3.34) is satisfied for $P(u)=(u+\gamma_1)\cdots(u+\gamma_s)$. Introduce the polynomials

$$\lambda_1'(u) = (1+(\gamma_1+1)u^{-1})\cdots(1+(\gamma_s+1)u^{-1})$$

and

$$\lambda_2'(u) = (1+\gamma_1 u^{-1})\cdots(1+\gamma_s u^{-1}).$$

By Proposition 3.7, the irreducible highest weight representation $L(\lambda_1'(u),\lambda_2'(u))$ is finite-dimensional. However, $\lambda_1'(u)$ and $\lambda_2'(u)$ satisfy (3.34) and so, there exists an automorphism of

$\mathrm{Y}(2)$ of the form (2.11) such that its composition with the representation $L(\lambda_1'(u),\lambda_2'(u))$ is isomorphic to $L(\lambda_1(u),\lambda_2(u))$. Thus, the latter is also finite-dimensional.

To complete the proof of the theorem, suppose that $Q(u)$ is another monic polynomial in u and

$$\frac{P(u+1)}{P(u)}=\frac{Q(u+1)}{Q(u)}.$$

This means that $P(u)/Q(u)$ is periodic in u which is only possible for $P(u)=Q(u)$. □

The polynomial $P(u)$ defined by (3.34) is called the *Drinfeld polynomial*; see also Corollary 3.13 below.

Relation (3.34) motivates the following notation which will be used in the sequel. For two formal series $\lambda_1(u)$ and $\lambda_1(u)$ in u^{-1} we shall write

$$\lambda_1(u)\to\lambda_2(u), \tag{3.35}$$

if there exists a monic polynomial $P(u)$ in u such that (3.34) holds.

Due to the decomposition (2.12) we can establish a correspondence between representations of $\mathrm{Y}(N)$ and $\mathrm{Y}(\mathrm{sl}(N))$ analogous to the correspondence between the representations of the Lie algebras $\mathrm{gl}(N)$ and $\mathrm{sl}(N)$. Consider the following *similarity classes* of finite-dimensional irreducible representations of $\mathrm{Y}(N)$: Two representations belong to the same class if one can be obtained from the other by the composition with an automorphism (2.11). Since the subalgebra $\mathrm{Y}(\mathrm{sl}(N))$ is stable with respect to all automorphisms of the form (2.11), any representation from such a class, restricted to the subalgebra $\mathrm{Y}(\mathrm{sl}(N))$ gives the same representation of $\mathrm{Y}(\mathrm{sl}(N))$. Moreover, all finite-dimensional irreducible representations of $\mathrm{Y}(\mathrm{sl}(N))$ can be obtained in this way. In particular, using Theorem 3.8 we obtain a complete description of finite-dimensional irreducible representations of the Yangian $\mathrm{Y}(\mathrm{sl}(2))$.

Corollary 3.9: There is a one-to-one correspondence between finite-dimensional irreducible representations of the Yangian $\mathrm{Y}(\mathrm{sl}(2))$ *and monic polynomials in u. Every such representation is isomorphic to a representation of the form (3.16).*

Proof: The first claim follows from the fact that the ratios $\lambda_1(u)/\lambda_2(u)$ take the same value for all representations $L(\lambda_1(u),\lambda_2(u))$ of $\mathrm{Y}(2)$ of a given similarity class and hence by Theorem 3.8 the similarity classes are parametrized by monic polynomials in u.

Further, given a monic polynomial $P(u)$ we may find polynomials $\lambda_1(u)$ and $\lambda_2(u)$ in u^{-1} such that relation (3.34) holds; see the proof of Theorem 3.8. Then by Proposition 3.6, the representation $L(\lambda_1(u),\lambda_2(u))$ of $\mathrm{Y}(2)$ is isomorphic to a tensor product (3.16) and its restriction to $\mathrm{Y}(\mathrm{sl}(2))$ gives the finite-dimensional irreducible representation corresponding to $P(u)$. □

Remark 3.10: Tensoring $L(\lambda_1(u),\lambda_2(u))$ by factors of the form $L(\gamma,\gamma)$ is equivalent to multiplying both $\lambda_1(u)$, and $\lambda_2(u)$ simultaneously by $\gamma(u)=1+\gamma u^{-1}$. However, the representations $L(\lambda_1(u),\lambda_2(u))$ and $L(\gamma(u)\lambda_1(u),\gamma(u)\lambda_2(u))$ of $\mathrm{Y}(2)$ belong to the same similarity class. Therefore, we might add to Corollary 3.9 the condition that in (3.16) all differences $\alpha_i-\beta_i$ are positive integers. □

Since all finite-dimensional irreducible representation of $\mathrm{Y}(\mathrm{sl}(2))$ have the form of (3.16) it is useful to have a criterion of irreducibility of such representations.

Following Ref. 11 we define the *string* corresponding to a pair of complex numbers (α,β) with $\alpha-\beta\in\mathbb{Z}_+$ as the set

$$S(\alpha,\beta)=\{\beta,\beta+1,\ldots,\alpha-1\}\subset\mathbb{C}.$$

If $\alpha=\beta$ the set $S(\alpha,\beta)$ is considered to be empty. We say that two strings S_1 and S_2 are *in general position* if either

(i) $S_1\cup S_2$ is not a string, or

(ii) $S_1\subseteq S_2$, or $S_2\subseteq S_1$.

The following result is contained in Refs. 11 and 13.

Corollary 3.11: Consider the representation L of $\mathrm{Y}(2)$ *or* $\mathrm{Y}(\mathrm{sl}(2))$ *given by (3.16) where all differences $\alpha_i-\beta_i$ are non-negative integers. Then L is irreducible if and only if the strings $S(\alpha_1,\beta_1),\dots,S(\alpha_1,\beta_k)$ are pairwise in general position.*

Proof: Suppose that the strings are pairwise in general position and assume first that $\alpha_1-\beta_1\leqslant\cdots\leqslant\alpha_k-\beta_k$. Then one easily checks that (3.19) holds and so, by Proposition 3.6 representation (3.16) is irreducible. To complete the proof of the ''if'' part note that if representation (3.16) is irreducible, then any permutation of the tensor factors yields an isomorphic representation L'. Indeed, the submodule in L' generated by the tensor product of the highest weight vectors of the $L(\alpha_i,\beta_i)$ is highest weight with the same $\lambda_1(u)$ and $\lambda_2(u)$ given by (3.20) and (3.21). Therefore, L is isomorphic to a subquotient of L'. As L and L' have the same dimension, they have to be isomorphic.

Conversely, let $k=2$ and let $L(\alpha_1,\beta_1)\otimes L(\alpha_2,\beta_2)$ be irreducible. Suppose that the strings $S(\alpha_1,\beta_1)$ and $S(\alpha_2,\beta_2)$ are not in general position. Then $\alpha_2-\beta_1$, $\alpha_1-\beta_2\in\mathbb{Z}_+$ and either both of $\beta_1-\beta_2$ and $\alpha_1-\alpha_2$, or both of $\beta_2-\beta_1$ and $\alpha_2-\alpha_1$ are positive integers. The strings $S(\alpha_1,\beta_2)$ and $S(\alpha_2,\beta_1)$ are in general position and so, the representation $L(\alpha_1,\beta_2)\otimes L(\alpha_2,\beta_1)$ is irreducible. Since it has the same highest weight as the irreducible representation $L(\alpha_1,\beta_1)\otimes L(\alpha_2,\beta_2)$ these two representations have to be isomorphic and, in particular, have the same dimension:

$$(\alpha_1-\beta_1+1)(\alpha_2-\beta_2+1)=(\alpha_1-\beta_2+1)(\alpha_2-\beta_1+1).$$

This implies that $(\alpha_1-\alpha_2)(\beta_1-\beta_2)=0$, which is impossible.

Now consider the general case. Suppose that L is irreducible but two strings $S(\alpha_i,\beta_i)$ and $S(\alpha_j,\beta_j)$ are not in general position. Permuting the factors in (3.16) if necessary, we may assume that i and j are adjacent. However, the representation $L(\alpha_i,\beta_i)\otimes L(\alpha_j,\beta_j)$ is reducible. Contradiction. □

We are now able to prove Drinfeld's classification theorem for representations of the Yangian $\mathrm{Y}(N)$, N is an arbitrary positive integer; see Ref. 2. Theorem 3.3 and the following theorem provide a complete description of finite-dimensional irreducible representations of $\mathrm{Y}(N)$. We use notation (3.35).

Theorem 3.12: *The irreducible highest weight representation $L(\lambda(u))$, $\lambda(u)=(\lambda_1(u),\dots,\lambda_N(u))$ of* $\mathrm{Y}(N)$ *is finite-dimensional if and only if the following relation holds:*

$$\lambda_1(u)\to\lambda_2(u)\to\cdots\to\lambda_N(u). \tag{3.36}$$

Proof: Suppose that $\dim L(\lambda(u))<\infty$. By the Poincaré–Birkhoff–Witt theorem for $\mathrm{Y}(N)$, for any $i\in\{1,\dots,N-1\}$ the subalgebra in $\mathrm{Y}(N)$ generated by the coefficients of the series $t_{ii}(u)$, $t_{i,i+1}(u)$, $t_{i+1,i}(u)$, $t_{i+1,i+1}(u)$ is isomorphic to $\mathrm{Y}(2)$. The cyclic span of the highest weight vector of $L(\lambda(u))$ with respect to this subalgebra is a representation with the highest weight $(\lambda_i(u),\lambda_{i+1}(u))$. Its irreducible quotient is finite-dimensional and so, applying Theorem 3.8 we get (3.36).

Conversely, let

$$\frac{\lambda_i(u)}{\lambda_{i+1}(u)}=\frac{P_i(u+1)}{P_i(u)} \tag{3.37}$$

for certain polynomials $P_i(u)$, $i=1,\dots,N-1$ given by

$$P_i(u)=(u+\gamma_1^{(i)})\cdots(u+\gamma_{s_i}^{(i)}).$$

Consider first the irreducible highest weight representation $L(\lambda'(u))$ of $\mathrm{Y}(N)$ with the highest weight $\lambda'(u)=(\lambda_1'(u),\dots,\lambda_N'(u))$, where

$$\lambda_i'(u)=\prod_{a=1}^{i-1}(1+\gamma_1^{(a)}u^{-1})\cdots(1+\gamma_{s_a}^{(a)}u^{-1})\prod_{a=i}^{N-1}(1+(\gamma_1^{(a)}+1)u^{-1})\cdots(1+(\gamma_{s_a}^{(a)}+1)u^{-1}).$$

These polynomials have the form

$$\lambda_i'(u)=(1+\alpha_1^{(i)}u^{-1})\cdots(1+\alpha_k^{(i)}u^{-1})$$

with $k=\Sigma\, s_i$ and the condition $\alpha_r^{(i)}-\alpha_r^{(i+1)}\in\mathbb{Z}_+$ is satisfied for each $r=1,...,k$ and $i=1,...,N-1$. Consider the tensor product

$$L(\alpha_1)\otimes\cdots\otimes L(\alpha_k), \tag{3.38}$$

where $L(\alpha_r)$ is the irreducible representation of $\mathrm{gl}(N)$ with the highest weight $\alpha_r=(\alpha_r^{(1)},\ldots,\alpha_r^{(N)})$. The generators $t_{ij}(u)$ act in the representation (3.38) by the rule (3.17). So, the tensor product of the highest weight vectors of the $L(\alpha_r)$ generates a representation of $\mathrm{Y}(N)$ with the highest weight $\lambda'(u)$. Hence, its irreducible quotient is isomorphic to $L(\lambda'(u))$. Representation (3.38) is finite-dimensional, and therefore $L(\lambda'(u))$ is. Finally, note that the $\lambda_i'(u)$ satisfy (3.37) and so, there exists an automorphism of $\mathrm{Y}(N)$ of the form (2.11) such that its composition with the representation $L(\lambda'(u))$ is isomorphic to $L(\lambda(u))$ which completes the proof. □

We conclude this section with a description of finite-dimensional irreducible representations of the Yangian $\mathrm{Y}(\mathrm{sl}(N))$ implied by Theorem 3.12; see Ref. 2.

Corollary 3.13: There is a one-to-one correspondence between finite-dimensional irreducible representations of the Yangian $\mathrm{Y}(\mathrm{sl}(N))$ *and the families of monic polynomials* $\{P_1(u),\ldots,P_{N-1}(u)\}$ *in* u. *Every such representation is isomorphic to a subquotient of a representation of the form (3.38).* □

The $P_i(u)$ are called the *Drinfeld polynomials* associated with the representation $L(\lambda(u))$ of $\mathrm{Y}(\mathrm{sl}(N))$. Note that our definition of $P_i(u)$ given by (3.37) differs from that of Ref. 2 by a shift in u.

Remark 3.14: Contrary to the case $N=2$, it is not true for $N\geqslant 3$ that every finite-dimensional irreducible representation of $\mathrm{Y}(\mathrm{sl}(N))$ is isomorphic to a tensor product (3.38). For example, the $\mathrm{Y}(\mathrm{sl}(3))$-module $L(\lambda(u))$ with

$$\lambda_1(u)=(1+3u^{-1})(1+u^{-1}),\quad \lambda_2(u)=1+3u^{-1},\quad \lambda_3(u)=1+2u^{-1}$$

can be shown to have dimension 8 (see, e.g., Ref. 18) and it is not isomorphic to a tensor product of type (3.38). However, this is still true for ''generic'' representations of $\mathrm{Y}(\mathrm{sl}(N))$; see, e.g., Refs. 16–18.

IV. HIGHEST WEIGHT REPRESENTATIONS OF THE TWISTED YANGIANS

Here we define highest weight representations of the algebras $\mathrm{Y}^{\pm}(N)$ and prove an analog of Theorem 3.3 for $\mathrm{Y}^{\pm}(N)$. We also find some necessary conditions for an irreducible highest weight representation to be finite-dimensional and construct a family of representations of $\mathrm{Y}^{\pm}(N)$.

From now until the end of the paper we shall be assuming that the generators $t_{ij}(u)$ of $\mathrm{Y}(N)$ as well as the basis elements E_{ij} of the Lie algebra $\mathrm{gl}(N)$ are parameterized by the indices $i,j\in\{-n,-n+1,\ldots,n\}$, where $n=[N/2]$ and the index 0 is skipped in $N=2n$ (see Sec. II).

Definition 4.1: A representation V of the twisted Yangian $\mathrm{Y}^{\pm}(N)$ is called *highest weight* if there exists a nonzero vector $\xi\in V$ such that V is generated by ξ,

$$s_{ij}(u)\xi=0\quad\text{for}\quad -n\leqslant i<j\leqslant n, \tag{4.1}$$

and

$$s_{ii}(u)\xi=\mu_i(u)\xi \tag{4.2}$$

for some formal series $\mu_i(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$, where $i=1,\ldots,n$ if $N=2n$ and $i=0,\ldots,n$ if $N=2n+1$. The vector ξ is called the *highest weight vector* of V and the set $\mu(u):=(\mu_1(u),\ldots,\mu_n(u))$ or $\mu(u):=(\mu_0(u),\ldots,\mu_n(u))$ is the *highest weight* of V. □

Due to the symmetry relation (2.18), the vector ξ is also an eigenvector for the generators $s_{-i,-i}(u)$, $i=1,\ldots,n$. Moreover, if $N=2n+1$ the component $\mu_0(u)$ of the highest weight has to be even: $\mu_0(u)=\mu_0(-u)$.

If we replace (4.1) by the relations

$$s_{ji}(u)\xi=0 \quad \text{for } -n\leqslant i<j\leqslant n.$$

we get the definition of a *lowest weight representation.*

Definition 4.2: Let $\mu(u)=(\mu_1(u),\dots,\mu_n(u))$ or $\mu(u)=(\mu_0(u),\dots,\mu_n(u))$ be any set of formal series where $\mu_i(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$ for $i=1,\dots,n$ and $\mu_0(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]]$. The *Verma module* $M(\mu(u))$ is the quotient of $\mathrm{Y}^{\pm}(N)$ by the left ideal generated by all the coefficients of the series $s_{ij}(u)$ for $-n\leqslant i<j\leqslant n$ and $s_{ii}(u)-\mu_i(u)$ for $i=1,\dots,n$ or $i=0,\dots,n$, respectively. □

The following properties of $M(\mu(u))$ are immediate from the Poincaré–Birkhoff–Witt theorem for $\mathrm{Y}^{\pm}(N)$ (see Sec. II). First, ordered monomials in the generators (2.19) or (2.20) with the additional condition $i>j$ [assuming that (2.21) is satisfied] form a basis of $M(\mu(u))$. Further, $M(\mu(u))$ is a representation of $\mathrm{Y}^{\pm}(N)$ with the highest weight $\mu(u)$ and any representation of the highest weight $\mu(u)$ is isomorphic to a quotient of $M(\mu(u))$. The module $M(\mu(u))$ has a unique irreducible quotient which we denote by $V(\mu(u))$. Moreover, if an irreducible $\mathrm{Y}^{\pm}(N)$-module V contains a nonzero vector ξ satisfying (4.1) and (4.2) then V is isomorphic to $V(\mu(u))$.

Using the inclusion (2.22) we may regard any $\mathrm{Y}^{\pm}(N)$-module as a $\mathrm{g}(N)$-module. Relations (2.17) imply that

$$[F_{ij},s_{kl}(u)]=\delta_{kj}s_{il}(u)-\delta_{il}s_{kj}(u)-\theta_{k,-j}\delta_{i,-k}s_{-j,l}(u)+\theta_{i,-l}\delta_{-l,j}s_{k,-i}(u), \tag{4.3}$$

where we identify the elements $F_{ij}\in\mathrm{g}(N)$ with their images in $\mathrm{Y}^{\pm}(N)$. Set $\mathrm{h}=$span of $\{F_{11},\dots,F_{nn}\}$. Then h is a Cartan subalgebra in $\mathrm{g}(N)$. Choose the standard system of positive roots R_+ for $\mathrm{g}(N)$ with respect to h corresponding to the upper triangular Borel subalgebra and let $\alpha_1,\dots,\alpha_n$ be a basis in R_+. If V is a representation of $\mathrm{Y}^{\pm}(N)$ with the highest weight $\mu(u)$ then its highest weight vector ξ is also a $\mathrm{g}(N)$-highest weight vector of the $\mathrm{g}(N)$-weight $\mu=(\mu_1,\dots,\mu_n)$ where μ_i is the coefficient at u^{-1} in the series $\mu_i(u)$. By (4.3) all $\mathrm{g}(N)$-weights of V have the form $\mu-\omega$, where ω is a $\mathbb{Z}_+$-linear combination of the α_i. Moreover, the subspace in V of vectors of $\mathrm{g}(N)$-weight μ if one-dimensional and spanned by ξ.

Theorem 4.3: *Every finite-dimensional irreducible representation V of $\mathrm{Y}^{\pm}(N)$ is highest weight. Moreover, V contains a unique (up to scalar multiples) highest weight vector.*

Proof: The proof follows the same steps as that of Theorem 3.3. We set

$$V^0:=\{\eta\in V\,|\,s_{ij}(u)\,\eta=0 \quad \text{for } -n\leqslant i<j\leqslant n\}. \tag{4.4}$$

By (4.3) we have the weight space decomposition

$$V^0=\bigoplus_{\mu\in\mathrm{h}^*} V^0_\mu, \tag{4.5}$$

where

$$V^0_\mu=\{\eta\in V^0\,|\,F_{ii}\,\eta=\mu_i\,\eta\},\quad \mu_i=\mu(F_{ii}).$$

Introduce the h-centralizer in $\mathrm{Y}^{\pm}(N)$:

$$\mathrm{Y}^{\pm}(N)_0=\{y\in Y^{\pm}(N)\,|\,[F_{ii},y]=0 \ \text{ for all } \ i\}.$$

Denote by I the left ideal in $\mathrm{Y}^{\pm}(N)$ generated by the coefficients of the series $s_{ij}(u)$, where $-n\leqslant i<j\leqslant n$ and set $I_0=\mathrm{Y}^{\pm}(N)_0\cap I$. We derive from the Poincaré–Birkhoff–Witt theorem for $\mathrm{Y}^{\pm}(N)$ that I_0 is a two-sided ideal in $\mathrm{Y}^{\pm}(N)_0$. Next, using (2.17) we verify that the algebra $\mathscr{Y}:=\mathrm{Y}^{\pm}(N)_0/I_0$ is commutative and generated by the elements $s_{ii}^{(r)}$ mod I_0.

Further, V^0 is a nonzero $\mathscr{Y}$-module (cf. the proof of Theorem 3.3) and the weight spaces V^0_μ are $\mathscr{Y}$-invariant. If $V^0_{\mu_0}\neq 0$ for a certain μ_0 then there exists a nonzero element $\xi\in V^0_{\mu_0}$ which is a common eigenvector for the generators $s_{ii}^{(r)}$. So, ξ generates a highest weight $\mathrm{Y}^{\pm}(N)$-submodule in V. Since V is irreducible this submodule coincides with V which proves that V is highest weight.

Finally, the subspace $V^0_{\mu_0}$ is one-dimensional and nonzero weight subspaces in (4.5) can only correspond to the g(N)-weights μ of the form $\mu_0-\omega$, where ω is a $\mathbb{Z}_+$-linear combination of the α_i. If V^0_μ is nonzero for a certain $\mu\neq\mu_0$ then a common eigenvector in V^0_μ for all $s^{(r)}_{ii}$ generates a proper submodule in V, which makes a contradiction. □

Remark 4.4: The above proof implies that if V is an irreducible highest weight representation of $\mathrm{Y}^{\pm}(N)$ then the subspace (4.4) is one-dimensional and spanned by the highest weight vector ξ of V. □

Due to Theorem 4.3, to describe the finite-dimensional irreducible representations of $\mathrm{Y}^{\pm}(N)$ it suffices to find necessary and sufficient conditions for an irreducible highest weight representation $V(\mu(u))$ to be finite-dimensional. Some necessary conditions are provided by the following proposition. We use notation (3.35).

Proposition 4.5: If the irreducible highest weight representation $V(\mu(u))$ of the algebra $\mathrm{Y}^{\pm}(N)$ is finite-dimensional then the following relation holds:

$$\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u).$$

Proof: Let J be the left ideal of $\mathrm{Y}^{\pm}(N)$ generated by the coefficients of the series $s_{-i,j}(u)$, where $i,j=1,\ldots,n$. Denote by Norm J the normalizer of J:

$$\mathrm{Norm}\ J=\{y\in\mathrm{Y}^{\pm}(N)\,|\,Jy\subseteq J\}.$$

The subspace

$$V(\mu(u))^J=\{v\in V(\mu(u))\,|\,Jv=0\}$$

has a natural structure of a Norm J/J-module. Note that the highest weight vector ξ of $V(\mu(u))$ is contained in $V(\mu(u))^J$. Defining relations (2.17) show that the mapping

$$t^{(r)}_{ij}\mapsto s^{(r)}_{ij}\ \mathrm{mod}\ J,\quad 1\leqslant i,j\leqslant n$$

defines an algebra homomorphism $\mathrm{Y}(n)\to\mathrm{Norm}\ J/J$. By making use of this homomorphism we may regard $V(\mu(u))^J$ as a $\mathrm{Y}(n)$-module. The cyclic span of the vector ξ is a finite-dimensional representation of $\mathrm{Y}(n)$ with the highest weight $(\mu_1(u),\ldots,\mu_n(u))$. Applying Theorem 3.12 we complete the proof. □

Given $\mu=(\mu_1,\ldots,\mu_n)\in\mathrm{h}^*$ consider the irreducible g(N)-module $V(\mu)$ with the highest weight μ with respect to the basis $\{F_{11},\ldots,F_{nn}\}$ in h. It is finite-dimensional if and only if the following conditions are satisfied (see, e.g., Ref. 37, Sec. 7.2):

$$\mu_i-\mu_{i+1}\in\mathbb{Z}_+\qquad\text{for } i=1,\ldots,n-1 \tag{4.6}$$

and

$$-\mu_1-\mu_2\in\mathbb{Z}_+\quad\text{for o}(2n), \tag{4.7}$$

$$-\mu_1\in\mathbb{Z}_+\quad\text{for sp}(2n), \tag{4.8}$$

$$-2\mu_1\in\mathbb{Z}_+\quad\text{for o}(2n+1). \tag{4.9}$$

The homomorphism (2.23) allows us to extend $V(\mu)$ to a $\mathrm{Y}^{\pm}(N)$-module. This module is obviously highest weight with

$$\mu_i(u)=\frac{1+(\mu_i\pm 1/2)u^{-1}}{1\pm 1/2u^{-1}},\quad i=1,\ldots,n, \tag{4.10}$$

and $\mu_0(u)=1$ if $N=2n+1$.

Using (2.31) we can equip any tensor product $L\otimes V$ of a $\mathrm{Y}(N)$-module L and a $\mathrm{Y}^{\pm}(N)$-module V with a $\mathrm{Y}^{\pm}(N)$-action by the rule

$$y\cdot(\xi\otimes\eta):=\Delta(y)(\xi\otimes\eta),\quad y\in \mathrm{Y}^{\pm}(N),\quad \xi\in L,\quad \eta\in V. \tag{4.11}$$

In particular, we can construct representations of $\mathrm{Y}^{\pm}(N)$ of the form (1.3).

Decomposition (2.30) enables one to establish a correspondence between representations of the twisted Yangian $\mathrm{Y}^{\pm}(N)$ and the special twisted Yangian $\mathrm{SY}^{\pm}(N)$ analogous to the *A* series case (see Sec. III). Consider the *similarity classes* of finite-dimensional irreducible representations of $\mathrm{Y}^{\pm}(N)$: Two representations belong to the same class if one can be obtained from the other by the composition with an automorphism (2.29). Since the subalgebra $\mathrm{SY}^{\pm}(N)$ is stable with respect to all automorphisms of the form (2.29), any representation from such a class, restricted to the subalgebra $\mathrm{SY}^{\pm}(N)$ gives the same representation of the latter. Moreover, all finite-dimensional irreducible representations of $\mathrm{SY}^{\pm}(N)$ can be obtained in this way.

In the next three sections we consider separately the twisted Yangians corresponding to the series *C*, *D*, and *B*, respectively. We obtain a description of finite-dimensional irreducible representations for the algebras $\mathrm{Y}^{\pm}(N)$ and $\mathrm{SY}^{\pm}(N)$.

V. REPRESENTATIONS OF $\mathrm{Y}^{-}(2n)$

As for the Yangian $\mathrm{Y}(N)$ the representations of the simplest twisted Yangian $\mathrm{Y}^{-}(2)$ play an important role in the general case.

Irreducible highest weight representations $V(\mu(u))$ of $\mathrm{Y}^{-}(2)$ are parametrized by formal series $\mu(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$. The following is an analog of Proposition 3.5 for the algebra $\mathrm{Y}^{-}(2)$.

Proposition 5.1: If $\dim V(\mu(u))<\infty$ *then there exists an even formal series* $\psi(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]]$ *such that* $\psi(u)\mu(u)$ *is a polynomial in* u^{-1}.

Proof: The symmetry relation (2.18) implies that

$$(2u-1)s_{1,-1}(u)=(-2u-1)s_{1,-1}(-u),$$

and so, for any $r\geqslant 1$,

$$s_{1,-1}^{(2r-1)}=2s_{1,-1}^{(2r)}. \tag{5.1}$$

Denote by ξ the highest weight vector of $V(\mu(u))$. Take the minimum non-negative integer k such that $s_{1,-1}^{(2k+2)}\xi$ is a linear combination of the vectors $s_{1,-1}^{(2)}\xi,\ldots,s_{1,-1}^{(2k)}\xi$; such k exists because $V(\mu(u))$ is finite-dimensional. Let us prove that for any $r\geqslant k+1$ we have

$$s_{1,-1}^{(2r)}\xi=a_1^{(2r)}\xi_1+\cdots+a_k^{(2r)}\xi_k \tag{5.2}$$

for some complex coefficients $a_i^{(2r)}$, where $\xi_i:=s_{1,-1}^{(2i)}\xi$, $i=1,\ldots,k$. This is true for $r=k+1$ by our choice of *k*. Taking in (2.17) $i=k=l=1$, $j=-1$, and exchanging u and v we get

$$[s_{11}(u),s_{1,-1}(v)]=\frac{2u+1}{u^2-v^2}\,s_{1,-1}(v)s_{11}(u)-\frac{u+v+1}{u^2-v^2}\,s_{1,-1}(u)s_{11}(v)$$
$$+\frac{1}{u+v}\,s_{1,-1}(u)s_{-1,-1}(v). \tag{5.3}$$

Take the coefficient at $u^{-3}v^{-2p}$ in this relation. Using (5.1) and the fact that the series $s_{11}(u)+s_{-1,-1}(u)$ is even in u [see (2.18)], we obtain

$$[s_{11}^{(3)},s_{1,-1}^{(2p)}]=2s_{1,-1}^{(2p+2)}+s_{1,-1}^{(2p)}(2s_{11}^{(2)}+s_{11}^{(1)})-s_{1,-1}^{(2)}(3s_{11}^{(2p)}+s_{-1,-1}^{(2p)}). \tag{5.4}$$

Write $\mu(u)=1+\mu^{(1)}u^{-1}+\mu^{(2)}u^{-2}+\cdots$. Then applying both sides of (5.4) to ξ we get

$$2s_{1,-1}^{(2p+2)}\xi=s_{11}^{(3)}s_{1,-1}^{(2p)}\xi-(\mu^{(3)}+2\mu^{(2)}+\mu^{(1)})s_{1,-1}^{(2p)}\xi+(3\mu^{(2p)}+\mu_{-}^{(2p)})s_{1,-1}^{(2)}\xi, \tag{5.5}$$

where $\mu_{-}^{(r)}$ is the eigenvalue of $s_{-1,-1}^{(r)}$ at ξ. In particular, for $p=1,\ldots,k-1$ we have

$$s_{11}^{(3)}\xi_p=2\xi_{p+1}+(\mu^{(3)}+2\mu^{(2)}+\mu^{(1)})\xi_p-(3\mu^{(2p)}+\mu_-^{(2p)})\xi_1. \tag{5.6}$$

The use of (5.2) with $r=k+1$ gives

$$s_{11}^{(3)}\xi_k=2(a_1^{(2k+2)}\xi_1+\cdots+a_k^{(2k+2)}\xi_k)+(\mu^{(3)}+2\mu^{(2)}+\mu^{(1)})\xi_k-(3\mu^{(2k)}+\mu_-^{(2k)})\xi_1. \tag{5.7}$$

Now (5.2) follows from (5.5) to (5.7) by an induction on r. Rewriting (5.2) in terms of generating series and using (5.1) we obtain

$$s_{1,-1}(u)\xi=(2u+1)(a_1(u)\xi_1+\cdots+a_k(u)\xi_k), \tag{5.8}$$

where the $a_i(u)$ are even series in u^{-1}: $a_i(u)=u^{-2i}+a_i^{(2k+2)}u^{-2k-2}+\cdots$. Taking the coefficient at u^{-3} in (5.3) and replacing v with u we get

$$\begin{aligned}s_{11}^{(3)}s_{1,-1}(u)=&\,s_{1,-1}(u)(s_{11}^{(3)}+2s_{11}^{(2)}+s_{11}^{(1)}+2u^2)\\&-(s_{1,-1}^{(2)}+(u+1)s_{1,-1}^{(1)})s_{11}(u)+(s_{1,-1}^{(2)}-us_{1,-1}^{(1)})s_{-1,-1}(u).\end{aligned}$$

Apply both sides to the vector ξ. Then using (5.8) and the symmetry relation (2.18) we find that

$$\begin{aligned}&(s_{11}^{(3)}-\mu^{(3)}-2\mu^{(2)}-\mu^{(1)}-2u^2)(a_1(u)\xi_1+\cdots+a_k(u)\xi_k)\\&\quad=-\left(\frac{2u+1}{2u}\,\mu(u)+\frac{2u-1}{2u}\,\mu(-u)\right)\xi_1.\end{aligned} \tag{5.9}$$

Using (5.6) and (5.7) and taking the coefficient at ξ_p, $p=2,\ldots,k$ in this relation we get

$$a_{p-1}(u)-u^2a_p(u)+\text{const}\cdot a_k(u)=0.$$

Therefore, for any $p=1,\ldots,k$ one has $a_p(u)=A_p(u)a_k(u)$ for a monic polynomial $A_p(u)$ of degree $k-p$ in u^2. Hence, taking the coefficient at ξ_1 in (5.9) we obtain that

$$\left(\frac{2u+1}{2u}\,\mu(u)+\frac{2u-1}{2u}\,\mu(-u)\right)=B(u)a_k(u) \tag{5.10}$$

for a polynomial $B(u)$ of degree k in u^2.

Similarly, taking the coefficient at u^{-2} in (5.3) we get

$$s_{11}^{(2)}s_{1,-1}(u)=s_{1,-1}(u)(s_{11}^{(2)}+2s_{11}^{(1)}+1)-s_{1,-1}^{(1)}(s_{11}(u)-s_{-1,-1}(u)). \tag{5.11}$$

Applying both sides to ξ, using (5.8) and the symmetry relation (2.18) we obtain

$$(s_{11}^{(2)}-\mu^{(2)}-2\mu^{(1)}-1)(a_1(u)\xi_1+\cdots+a_k(u)\xi_k)=-u^{-1}(\mu(u)-\mu(-u))\xi_1. \tag{5.12}$$

In particular, $s_{11}^{(2)}\xi_p$ is a linear combination of ξ_p and ξ_1. Finally, taking the coefficient at ξ_1 in (5.12) we find that $u^{-1}(\mu(u)-\mu(-u))=C(u)a_k(u)$ for a polynomial $C(u)$ of degree $k-1$ in u^2. Together with (5.10) this proves that the desired series is $\psi(u)=(u^{2k}a_k(u))^{-1}$. □

We shall consider (3.16) as a representation of $\mathrm{Y}^-(2)$ obtained by restriction. Equivalently, using (4.11) one may regard this representation as the tensor product of Y(2)-module (3.16) with the trivial $\mathrm{Y}^-(2)$-module.

Let $\mu(u)$ be a polynomial in u^{-1} of degree $\leqslant 2k$. Consider an expansion

$$\mu(u)=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k}u^{-1}). \tag{5.13}$$

Re-enumerating the numbers γ_i if necessary, we may assume that the following condition is satisfied [cf. (3.19)]:

for every $i=1,\ldots,k$ we have: if the multiset

$$\{\gamma_p+\gamma_q \mid 2i-1\leqslant p<q\leqslant 2k\} \text{ contains non-negative integers} \tag{5.14}$$

then $\gamma_{2i-1}+\gamma_{2i}$ is minimal among them.

The following is an analog of Proposition 3.6.

Proposition 5.2: Let $V(\mu(u))$ be an irreducible highest weight representation of $\mathrm{Y}^-(2)$ where $\mu(u)$ is a polynornial of degree $\leqslant 2k$ in u^{-1}. Suppose that decomposition (5.13) is given and condition (5.14) is satisfied. Then $V(\mu(u))$ is isomorphic to the tensor product

$$L(\gamma_1,-\gamma_2)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k}), \tag{5.15}$$

regarded as $\mathrm{Y}^-(2)$-module.

Proof: Denote the representation (5.15) by V. Let ξ_i be the highest weight vector of $L(\gamma_{2i-1},-\gamma_{2i})$. We obtain from (3.23) that the vector $\xi:=\xi_1\otimes\cdots\otimes\xi_k$ satisfies the relations

$$\begin{gathered} t_{-1,1}(u)\xi=0, \\ t_{11}(u)\xi=(1-\gamma_2u^{-1})\cdots(1-\gamma_{2k}u^{-1})\xi, \\ t_{-1,-1}(u)\xi_i=(1+\gamma_1u^{-1})\cdots(1+\gamma_{2k-1}u^{-1})\xi. \end{gathered} \tag{5.16}$$

Hence, by (2.15)

$$s_{11}(u)\xi=\mu(u)\xi, \quad s_{-1,1}(u)\xi=0. \tag{5.17}$$

Thus, the cyclic $\mathrm{Y}^-(2)$-span of the vector ξ in V is a highest weight module with the highest weight $\mu(u)$. Therefore, to prove Proposition 5.2 it suffices to show that the representation V of $\mathrm{Y}^-(2)$ is irreducible.

We use induction on k. For $k=1$ this is obvious, because the restriction of $L(\gamma_{2i-1},-\gamma_{2i})$ to sp(2) is irreducible and so its extension to $\mathrm{Y}^-(2)$ is.

Suppose now that $k>1$. If $\widetilde{V}$ is a nonzero submodule of V then $\widetilde{V}$ contains a nonzero vector η such that $s_{-1,1}(u)\,\eta=0$. Write η in the form

$$\eta=\sum_{r=0}^{p}(E_{1,-1})^r\xi_1\otimes\eta_r, \tag{5.18}$$

for some $p\geqslant 0$, where $\eta_r\in L(\gamma_3,-\gamma_4)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})$; and if $\gamma_1+\gamma_2\in\mathbb{Z}_+$ then $p\leqslant\gamma_1+\gamma_2$. We may assume that $\eta_p\neq 0$. By (2.31) we have

$$\begin{aligned} s_{-1,1}(u)((E_{1,-1})^r\xi_1\otimes\eta_r)=&-t_{-1,-1}(u)t_{-1,1}(-u)(E_{1,-1})^r\xi_1\otimes s_{-1,-1}(u)\,\eta_r \\ &+t_{-1,-1}(u)t_{-1,-1}(-u)(E_{1,-1})^r\xi_1\otimes s_{-1,1}(u)\,\eta_r \\ &-t_{-1,1}(u)t_{-1,1}(-u)(E_{1,-1})^r\xi_1\otimes s_{1,-1}(u)\,\eta_r \\ &+t_{-1,1}(u)t_{-1,-1}(-u)(E_{1,-1})^r\xi_1\otimes s_{1,1}(u)\,\eta_r. \end{aligned} \tag{5.19}$$

Rewriting (3.25) and (3.26) we obtain

$$t_{-1,-1}(u)(E_{1,-1})^r\xi_1=(1+(\gamma_1-r)u^{-1})(E_{1,-1})^r\xi_1 \tag{5.20}$$

and

$$t_{-1,1}(u)(E_{1,-1})^r\xi_1=u^{-1}r(\gamma_1+\gamma_2-r+1)(E_{1,-1})^{r-1}\xi_1. \tag{5.21}$$

Taking the coefficient at $(E_{1,-1})^p\xi_1$ in the relation $s_{-1,1}(u)\,\eta=0$ gives

$$(1-(\gamma_1-p)u^{-2})s_{-1,1}(u)\,\eta_p=0.$$

So, $s_{-1,1}(u)\,\eta_p=0$. By the induction hypotheses, the representation $L(\gamma_3,-\gamma_4)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})$ of $\mathrm{Y}^-(2)$ is irreducible, and hence

$$\eta_p=\mathrm{const}\cdot\xi_2\otimes\cdots\otimes\xi_k\,, \tag{5.22}$$

where ''const'' is a nonzero constant; see Remark 4.4. Suppose now that $p\geqslant 1$ and take the coefficient at $(E_{1,-1})^{p-1}\xi_1$ in the relation $s_{-1,1}(u)\,\eta=0$. Using (5.19)–(5.21) we obtain

$$\begin{aligned}&(1-(\gamma_1-p+1)^2u^{-2})s_{-1,1}(u)\,\eta_{p-1}\\&\quad+u^{-1}p(\gamma_1+\gamma_2-p+1)(1+(\gamma_1-p+1)u^{-1})s_{-1,-1}(u)\,\eta_p\\&\quad+u^{-1}p(\gamma_1+\gamma_2-p+1)(1-(\gamma_1-p)u^{-1})s_{11}(u)\,\eta_p=0.\end{aligned} \tag{5.23}$$

The symmetry relation (2.18) gives

$$s_{-1,-1}(u)=\frac{2u+1}{2u}\,s_{11}(-u)-\frac{1}{2u}\,s_{11}(u) \tag{5.24}$$

and

$$s_{-1,1}^{(2r-1)}=2s_{-1,1}^{(2r)},\quad r\geqslant 1. \tag{5.25}$$

Using (5.24) and multiplying (5.23) by $u^{2k}/(2u+1)$ we bring it to the form

$$\begin{aligned}&(u^2-(\gamma_1-p+1)^2)u^{2k-2}\,\frac{s_{-1,1}(u)}{2u+1}\,\eta_{p-1}+p(\gamma_1+\gamma_2-p+1)\\&\times\frac{1}{2u}\,((u+\gamma_1-p+1)u^{2k-2}s_{11}(-u)+(u-\gamma_1+p-1)u^{2k-2}s_{11}(u))\,\eta_p=0.\end{aligned} \tag{5.26}$$

Due to (3.18) and (2.15) the elements $s_{ij}^{(r)}$ with $r>2k-2$ act as zero operators in $L(\gamma_3,-\gamma_4)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})$. Hence, the operators $u^{2k-2}s_{ij}(u)$ are polynomials in u, as well as $u^{2k-2}s_{-1,1}(u)/(2u+1)$ is, because of (5.25). Further, by (5.22) and (5.16),

$$u^{2k-2}s_{11}(u)\,\eta_p=(u-\gamma_3)\cdots(u-\gamma_{2k})\,\eta_p\,.$$

Putting $u=\gamma_1-p+1$ in (5.26) we get

$$p(\gamma_1+\gamma_2-p+1)(\gamma_1+\gamma_3-p+1)\cdots(\gamma_1+\gamma_{2k}-p+1)=0.$$

However, this is impossible because of the condition (5.14). Therefore, p has to be equal to 0 and so, η is a multiple of ξ.

To complete the proof we have to show that the submodule of V generated by ξ coincides with V. The representation of $\mathrm{Y}(2)$ dual to (5.15) is isomorphic to

$$L'(-\gamma_1,\gamma_2)\otimes\cdots\otimes L'(-\gamma_{2k-1},\gamma_{2k}); \tag{5.27}$$

see (3.31). On the other hand, using the antiautomorphism (2.26) we can equip the vector space V^* dual to V with an action of $\mathrm{Y}^-(2)$:

$$(y\cdot f)(v)=f(\sigma(y)v),\quad y\in\mathrm{Y}^-(2),\quad f\in V^*,\quad v\in V.$$

Since the antiautomorphism (2.26) is a restriction of (2.7), the representation V^* is isomorphic to the restriction of (5.27) to $\mathrm{Y}^-(2)$. Similarly to the case of representation V, any nonzero submodule of V^* contains a nonzero vector η' such that $s_{1,-1}(u)\,\eta'=0$. Modifying the previous argument for lowest weight representations of $\mathrm{Y}^-(2)$ we prove that η' coincides (up to a nonzero factor) with the vector $\xi'=\xi'_1\otimes\cdots\otimes\xi'_k$, where ξ'_i is the lowest weight vector of

$L'(-\gamma_{2i-1},\gamma_{2i})$. Now, if the submodule in V generated by ξ is proper, then its annihilator in V^* is a nonzero submodule which does not contain ξ'. Contradiction. □

We are now able to give necessary and sufficient conditions for $V(\mu(u))$ to be finite-dimensional (cf. Proposition 3.7).

Proposition 5.3: Let $V(\mu(u))$ be an irreducible representation of $\mathrm{Y}^-(2)$ with the highest weight $\mu(u)$. Then $V(\mu(u))$ is finite-dimensional if and only if there exists a formal series $\psi(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]]$ such that $\psi(u)\mu(u)$ is a polynomial in u^{-1} with a decomposition

$$\psi(u)\mu(u)=(1-\gamma_1 u^{-1})\cdots(1-\gamma_{2k}u^{-1}), \tag{5.28}$$

and $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$, $i=1,...,k$ for a certain re-enumeration of the γ_i.

Proof: Let $\dim V(\mu(u))<\infty$. By Proposition 5.1, taking the composition of our representation with an automorphism of the form (2.29) we may assume that (5.28) is satisfied for a certain series $\psi(u)$. Re-enumerate the γ_i in such a way that condition (5.14) be satisfied. Now use Proposition 5.2. Since representation (5.15) is finite-dimensional this gives the relations $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$ for all $i=1,...,k$.

Conversely, let (5.28) hold with $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$, $i=1,...,k$. Then representation (5.15) is finite-dimensional and the composition of $V(\mu(u))$ with the automorphism (2.29) corresponding to the series $\psi(u)$ is isomorphic to the irreducible quotient of the $\mathrm{Y}^-(2)$-cyclic span in V of the vector ξ; see (5.17). Therefore, $\dim V(\mu(u))<\infty$. □

Proposition 5.3 can be reformulated with the use of monic polynomials in u as follows (cf. Theorem 3.8).

Theorem 5.4: *The irreducible highest weight representation $V(\mu(u))$ of $\mathrm{Y}^-(2)$ is finite-dimensional if and only if there exists a monic polynomial $P(u)\in\mathbb{C}[u]$ such that $P(u)=P(-u+1)$ and*

$$\frac{\mu(-u)}{\mu(u)}=\frac{P(u+1)}{P(u)}. \tag{5.29}$$

In this case $P(u)$ is unique.

Proof: If $\dim V(\mu(u))<\infty$ then by Proposition 5.3 relation (5.28) holds for a certain even series $\psi(u)$ and $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$ for all i. Set

$$\begin{aligned}\sigma(u)&=(1+\gamma_1 u^{-1})(1+\gamma_3 u^{-1})\cdots(1+\gamma_{2k-1}u^{-1}),\\ \tau(u)&=(1-\gamma_2 u^{-1})(1-\gamma_4 u^{-1})\cdots(1-\gamma_{2k}u^{-1}).\end{aligned} \tag{5.30}$$

Then there exists a monic polynomial $Q(u)$ in u such that

$$\frac{\sigma(u)}{\tau(u)}=\frac{Q(u+1)}{Q(u)}; \tag{5.31}$$

see the proof of Theorem 3.8. Define $P(u)$ by

$$P(u)=Q(u)Q(-u+1)(-1)^{\deg Q}. \tag{5.32}$$

Then $P(u)$ is monic and $P(u)=P(-u+1)$. Moreover,

$$\frac{\tau(-u)}{\sigma(-u)}=\frac{Q(-u)}{Q(-u+1)}. \tag{5.33}$$

Therefore, since $\psi(u)\mu(u)=\sigma(-u)\tau(u)$, relation (5.29) holds due to (5.31) and (5.33).

Conversely, let (5.29) be satisfied for a monic polynomial $P(u)$ such that $P(u)=P(-u+1)$. Then it has an even degree and the multiset of its roots is invariant with respect to the transformation $u\mapsto -u+1$ (the symmetry in $\mathbb{C}$ with the center at 1/2). Therefore, we may write the roots in the form $\{-\delta_1,\ldots,-\delta_s,\delta_1+1,...,\delta_s+1\}$ for a certain s. Set

$$\mu'(u)=(1-(\delta_1+1)u^{-1})\cdots(1-(\delta_s+1)u^{-1})(1+\delta_1 u^{-1})\cdots(1+\delta_s u^{-1}).$$

By Proposition 5.3 the representation $V(\mu'(u))$ is finite-dimensional. Moreover, the polynomial $\mu'(u)$ satisfies (5.29) and so, the ratio $\mu(u)/\mu'(u)$ is an even formal series in u^{-1}. This proves that $V(\mu(u))$ is also finite-dimensional.

The uniqueness of $P(u)$ is established in the proof of Theorem 3.8. □

The following notation is motivated by Theorem 5.4 [cf. (3.35)]. For two formal series $\mu(u)$ and $\nu(u)$ in u^{-1} we shall write

$$\mu(u) \Rightarrow \nu(u), \tag{5.34}$$

if there exists a monic polynomial $P(u)$ in u such that $P(u)=P(-u+1)$ and

$$\frac{\mu(u)}{\nu(u)} = \frac{P(u+1)}{P(u)}.$$

Obviously, this is only possible if $\mu(u)$ and $\nu(u)$ satisfy the relation $\mu(u)\mu(-u) = \nu(u)\nu(-u)$.

Proposition 5.2 and Theorem 5.4 imply the following parametrization of representations of the special twisted Yangian $\mathrm{SY}^-(2)$ (cf. Corollary 3.9).

Corollary 5.5. There is a one-to-one correspondence between finite-dimensional irreducible representations of the special twisted Yangian $\mathrm{SY}^-(2)$ *and the monic polynomials* $P(u)$ *in u such that* $P(u)=P(-u+1)$. *Every such representation is isomorphic to a representation of the form (5.15).* □

Remark 5.6: Tensoring $V(\mu(u))$ by factors of the form $L(\gamma,\gamma)$ is equivalent to multiplying the highest weight $\mu(u)$ by $1-\gamma^2u^{-2}$. However, the representations $V(\mu(u))$ and $V((1-\gamma^2u^{-2})\mu(u))$ of $\mathrm{Y}^-(2)$ belong to the same similarity class. Therefore, we might add to Corollary 5.5 the condition that in (5.15) all $\gamma_{2i-1}+\gamma_{2i}$ are positive integers. □

The following corollary provides a criterion of irreducibility for the representations of $\mathrm{Y}^-(2)$ [or $\mathrm{SY}^-(2)$] of the form (5.15) (cf. Corollary 3.11). We formulate it in terms of the strings $S(\alpha,\beta)$ introduced in Sec. III. Note that two strings $S(\alpha,-\beta)$ and $S(\beta,-\alpha)$ are symmetric to each other with respect to the center $-1/2$ in $\mathbb{C}$. Therefore, $S(\alpha_1,-\beta_1)$ and $S(\alpha_2,-\beta_2)$ are in general position whenever $S(\beta_1,-\alpha_1)$ and $S(\beta_2,-\alpha_2)$ are.

Corollary 5.7: Consider an irreducible finite-dimensional representation of $\mathrm{Y}(2)$ *of the form*

$$V = L(\gamma_1,-\gamma_2)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k}). \tag{5.35}$$

Then its restriction to $\mathrm{Y}^-(2)$ *is irreducible if and only if for each pair* $i<j$ *the strings* $S(\gamma_{2i}, -\gamma_{2i-1})$ *and* $S(\gamma_{2j-1},-\gamma_{2j})$ *are in general position.*

Proof: By Corollary 3.11, the strings $S(\gamma_{2i-1},-\gamma_{2i})$, $i=1,\ldots,k$ are pairwise in general position. Permuting the tensor factors in (5.35) if necessary, we may assume that $\gamma_1+\gamma_2\leqslant\cdots\leqslant\gamma_{2k-1}+\gamma_{2k}$. Suppose that the strings $S(\gamma_{2i},-\gamma_{2i-1})$ and $S(\gamma_{2j-1},-\gamma_{2j})$ are in general position for all $i<j$. Then one easily checks that (5.14) is satisfied and so, by Proposition 5.2 representation (5.35) of $\mathrm{Y}^-(2)$ is irreducible.

Conversely, let V be irreducible as a $\mathrm{Y}^-(2)$-module. Consider the module V' obtained by replacement of some of the tensor factors $L(\gamma_{2i-1},-\gamma_{2i})$, respectively, by $L(\gamma_{2i},-\gamma_{2i-1})$. Then V' is isomorphic to V. Indeed, $\dim V=\dim V'$ and one easily checks using (2.15) that V is isomorphic to the irreducible quotient of the $\mathrm{Y}^-(2)$-submodule in V' generated by the tensor product of the highest weight vectors of the factors. So, V' is irreducible as a $\mathrm{Y}^-(2)$-module and hence, as a $\mathrm{Y}(2)$-module. By Corollary 3.11 the corresponding strings are pairwise in general position. □

The following theorem together with Theorem 4.3 provides a description of finite-dimensional irreducible representations of $\mathrm{Y}^-(2n)$ (cf. Theorem 3.12). We use notations (3.35) and (5.34).

Theorem 5.8: *The irreducible highest weight representation* $V(\mu(u))$, $\mu(u)=(\mu_1(u),\ldots,\mu_n(u))$ *of* $\mathrm{Y}^-(2n)$ *is finite-dimensional if and only if the following relation holds:*

$$\mu_1(-u)\Rightarrow\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u). \tag{5.36}$$

Proof: Suppose that $\dim V(\mu(u))<\infty$. By Proposition 4.5 we have

$$\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u). \tag{5.37}$$

The subalgebra in $\mathrm{Y}^-(2n)$ generated by the coefficients of the series $s_{ij}(u)$ with $i,j=-1,1$ is isomorphic to $\mathrm{Y}^-(2)$. The cyclic span of the highest weight vector of $V(\mu(u))$ with respect to this subalgebra is a representation with the highest weight $\mu_1(u)$. Its irreducible quotient is finite-dimensional and so, by Theorem 5.4 we have $\mu_1(-u)\Rightarrow\mu_1(u)$.

Conversely, let (5.36) hold. Then

$$\frac{\mu_1(-u)}{\mu_1(u)}=\frac{P_1(u+1)}{P_1(u)}, \tag{5.38}$$

and

$$\frac{\mu_{i-1}(u)}{\mu_i(u)}=\frac{P_i(u+1)}{P_i(u)} \tag{5.39}$$

for certain polynomials $P_i(u)$ of the form

$$P_i(u)=(u+\delta_1^{(i)})\cdots(u+\delta_{s_i}^{(i)}),\quad i=2,...,n \tag{5.40}$$

and

$$P_1(u)=(u+\delta_1^{(1)})\cdots(u+\delta_{s_1}^{(1)})(u-\delta_1^{(1)}-1)\cdots(u-\delta_{s_1}^{(1)}-1); \tag{5.41}$$

see the proof of Theorem 5.4. Consider the irreducible highest weight representation $L(\lambda(u))$ with $\lambda(u)=(\lambda_{-n}(u),\cdots,\lambda_n(u))$ of the Yangian $\mathrm{Y}(2n)$ where for $i=1,...,n$,

$$\lambda_i(u)=\prod_{a=1}^{i}(1+\delta_1^{(a)}u^{-1})\cdots(1+\delta_{s_a}^{(a)}u^{-1})\prod_{a=i+1}^{n}(1+(\delta_1^{(a)}+1)u^{-1})\cdots(1+(\delta_{s_a}^{(a)}+1)u^{-1}) \tag{5.42}$$

and

$$\lambda_{-i}(u)=\prod_{a=1}^{n}(1+(\delta_1^{(a)}+1)u^{-1})\cdots(1+(\delta_{s_a}^{(a)}+1)u^{-1}). \tag{5.43}$$

Then $L(\lambda(u))$ is finite-dimensional by Theorem 3.12. One easily derives from relations (2.1) and (2.15) that the cyclic $\mathrm{Y}^-(2n)$-span of its highest weight vector is a representation with the highest weight $\mu'(u)=(\mu_1'(u),...,\mu_n'(u))$ where $\mu_i'(u)=\lambda_i(u)\lambda_{-i}(-u)$. So, the representation $V(\mu'(u))$ of $\mathrm{Y}^-(2n)$ is finite-dimensional. However, the polynomials $\mu_i'(u)$ satisfy (5.38) and (5.39). Hence, there exists an automorphism of $\mathrm{Y}^-(2n)$ of the form (2.29) such that its composition with the representation $V(\mu'(u))$ is isomorphic to $V(\mu(u))$ which completes the proof. □

Theorem 5.8 implies the following parametrization of representations of the special twisted Yangian $\mathrm{SY}^-(2n)$ (cf. Corollary 3.13).

Corollary 5.9: There is a one-to-one correspondence between finite-dimensional irreducible representations of the special twisted Yangian $\mathrm{SY}^-(2n)$ *and the families* $\{P_1(u),...,P_n(u)\}$ *of monic polynomials in u with* $P_1(u)=P_1(-u+1)$. *Every such representation is isomorphic to a subquotient of a representation of the form L(λ(u)).* □

VI. REPRESENTATIONS OF $\mathrm{Y}^+(2n)$

The results for the *D* series are very similar to the previous case. However, some differences take place for $n=1$ since the twisted Yangian $\mathrm{Y}^+(2)$ corresponds to the Abelian Lie algebra o(2). In particular, contrary to the case of $\mathrm{Y}^-(2)$, it is not true that every finite-dimensional irreducible representations of $\mathrm{Y}^+(2)$ is obtained by restriction of a representation of Y(2). Also, the step from $\mathrm{Y}^+(2)$ to $\mathrm{Y}^+(2n)$, $n\geqslant 2$ requires an extra care. We omit most of the proofs in this section which are parallel to the case of $\mathrm{Y}^-(2)$ but provide all necessary additional arguments.

We start with the case $n=1$. Irreducible highest weight representations $V(\mu(u))$ of $\mathrm{Y}^+(2)$ are parametrized by formal series $\mu(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$.

Proposition 6.1: If $\dim V(\mu(u))<\infty$ *then there exists an even formal series* $\psi(u)\in 1+u-2\mathbb{C}[[u-2]]$ *such that* $(1+1/2u^{-1})\psi(u)\mu(u)$ *is a polynomial in* u^{-1}.

Proof: We repeat the argument of the proof of Proposition 5.1 almost word by word and only point out a few differences. First, here the symmetry relation (2.18) gives $s_{1,-1}(u)=s_{1,-1}(-u)$ and so, for any $r\geqslant 1$ we have $s_{1,-1}^{(2r-1)}=0$. Relation (5.3) is replaced by

$$[s_{11}(u),s_{1,-1}(v)]=\frac{2u+1}{u^2-v^2}\,s_{1,-1}(v)s_{11}(u)$$
$$-\frac{u+v+1}{u^2-v^2}\,s_{1,-1}(u)s_{11}(v)-\frac{1}{u+v}\,s_{1,-1}(u)s_{-1,-1}(v). \tag{6.1}$$

Instead of (5.8) we obtain here $s_{1,-1}(u)\xi=a_1(u)\xi_1+\cdots+a_k(u)\xi_k$. The counterpart of relation (5.10) has exactly the same form. However, for the remaining part of the proof instead of (5.11) we need to consider the following formula obtained from (6.1) by taking the coefficient at u^{-4}:

$$s_{11}^{(4)}s_{1,-1}(u)=s_{1,-1}(u)(s_{11}^{(4)}+2s_{11}^{(3)}+s_{11}^{(2)}+2s_{11}^{(1)}u^2+u^2)-s_{1,-1}^{(2)}((u+1)s_{11}(u)-us_{-1,-1}(u)).$$

Applying both sides to ξ and using the symmetry relation (2.18) we find that

$$(u+\tfrac{1}{2})\mu(u)-(u-\tfrac{1}{2})\mu(-u)=C(u)a_k(u)$$

for a polynomial $C(u)$ of degree k in u^2. Together with (5.10) this proves that for $\psi(u)=(u^{2k}a_k(u))^{-1}$ the series $(1+1/2u^{-1})\psi(u)\mu(u)$ is a polynomial in u^{-1} of degree $\leqslant 2k+1$. □

Note that the irreducible representation of $\mathrm{Y}^+(2)$ with the highest weight $\mu(u)=(1+\gamma u^{-1})(1+1/2u^{-1})^{-1}$, where $\gamma\in\mathbb{C}$ is one-dimensional. It is obtained by extending the one-dimensional representation of the Abelian Lie algebra o(2) given by $F_{11}\xi=(\gamma-1/2)\xi$, where ξ is the basis vector; see (4.10). We shall denote this representation of $\mathrm{Y}^+(2)$ by $V(\gamma)$.

Let $(1+1/2u^{-1})\mu(u)$ be polynomial in u^{-1} of degree $\leqslant 2k+1$. Consider an expansion

$$\mu(u)=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k+1}u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1}. \tag{6.2}$$

Re-enumerating the numbers γ_i if necessary, we may assume that the following condition is satisfied [cf. (5.14)]:

$$\text{for every } i=1,\ldots,k \text{ we have: if the multiset}$$
$$\{\gamma_p+\gamma_q\,|\,2i-1\leqslant p<q\leqslant 2k+1\} \text{ contains non-negative integers} \tag{6.3}$$
$$\text{then } \gamma_{2i-1}+\gamma_{2i} \text{ is minimal among them.}$$

In the following proposition we use tensor products of representations constructed in accordance with (4.11). As before, we denote by $L(\alpha,\beta)$ a highest weight representation of the Lie algebra gl(2) which is also regarded as a Y(2)-module.

Proposition 6.2: Let $V(\mu(u))$ *be an irreducible highest weight representation of* $\mathrm{Y}^+(2)$ *where* $(1+1/2u^{-1})\mu(u)$ *is a polynomial of degree* $\leqslant 2k+1$ *in* u^{-1}. *Suppose that decomposition (6.2) is given and condition (6.3) is satisfied. Then* $V(\mu(u))$ *is isomorphic to the tensor product*

$$L(\gamma_1,-\gamma_2)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})\otimes V(-\gamma_{2k+1}). \tag{6.4}$$

Proof: Denote representation (6.4) by V. Let ξ_i be the highest weight vector of $L(\gamma_{2i-1},-\gamma_{2i})$ and ξ_{k+1} be the basis vector of the one-dimensional represetation $V(-\gamma_{2k+1})$. Using (3.23) and (2.31) we find that the vector $\xi:=\xi_1\otimes\cdots\otimes\xi_{k+1}$ satisfies

$$s_{11}(u)\xi=\mu(u)\xi,\qquad s_{-1,1}(u)\xi=0.$$

Thus, the cyclic $\mathrm{Y}^+(2)$-span of the vector ξ in V is a highest weight module with the highest weight $\mu(u)$. So, it suffices to show that the representation V is irreducible.

We use induction on k. For $k=0$ this is obvious. Suppose that $k\geqslant 1$. Proceeding exactly as in the proof of Proposition 5.2 we obtain that if the vector η is given by (5.18) with

$$\eta_r \in L(\gamma_3,-\gamma_4)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})\otimes V(-\gamma_{2k+1})$$

then

$$\eta_p=\mathrm{const}\cdot\xi_2\otimes\cdots\otimes\xi_{k+1}, \tag{6.5}$$

where ''const'' is a nonzero constant. Further, instead of (5.26) we get here

$$(u^2-(\gamma_1-p+1)^2)u^{2k-1}s_{-1,1}(u)\,\eta_{p-1}+p(\gamma_1+\gamma_2-p+1)u^{2k-1}$$
$$\times\left(\frac{2u+1}{2u}\,(u-\gamma_1+p-1)s_{11}(u)-\frac{2u-1}{2u}\,(u+\gamma_1-p+1)s_{11}(-u)\right)\eta_p=0. \tag{6.6}$$

The elements $s^{(r)}_{-1,1}$ and $s^{(r)}_{1,-1}$ act trivially in $V(-\gamma_{2k+1})$. Hence, by (2.31) and (4.11) we have

$$s_{-1,1}(u)(\zeta\otimes\xi_{k+1})=t_{-1,1}(u)t_{-1,-1}(-u)\zeta\otimes s_{11}(u)\xi_{k+1}+t_{-1,-1}(u)t_{-1,1}(-u)\zeta\otimes s_{-1,-1}(u)\xi_{k+1},$$

where $\zeta\in L(\gamma_3,-\gamma_4)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})$. Using the defining relations (2.1) and the symmetry relation (2.18) we may rewrite this as follows:

$$s_{-1,1}(u)(\zeta\otimes\xi_{k+1})=t_{-1,-1}(-u)t_{-1,1}(u)\zeta\otimes\frac{2u+1}{2u}\,s_{11}(u)\xi_{k+1}$$
$$+t_{-1,-1}(u)t_{-1,1}(-u)\zeta\otimes\frac{2u-1}{2u}\,s_{11}(-u)\xi_{k+1}. \tag{6.7}$$

Now, the definition of $V(\gamma)$ and (3.18) imply that $u^{2k-1}s_{-1,1}(u)\,\eta_{p-1}$ is a polynomial in u. Further, by (6.5) we have

$$\frac{2u+1}{2u}\,u^{2k-1}s_{11}(u)\,\eta_p=(u-\gamma_3)\cdots(u-\gamma_{2k+1})\,\eta_p\,.$$

So, putting $u=\gamma_1-p+1$ in (6.6) we get

$$p(\gamma_1+\gamma_2-p+1)(\gamma_1+\gamma_3-p+1)\cdots(\gamma_1+\gamma_{2k+1}-p+1)=0.$$

However, this is impossible because of condition (6.3). Therefore, p has to be equal to 0 and so, η is a multiple of ξ.

To show that the submodule of V generated by ξ coincides with V we consider the dual representation V^* defined by

$$(y\cdot f)(v)=f(\sigma(y)v),\quad y\in\mathrm{Y}^+(2),\quad f\in V^*,\quad v\in V,$$

where σ is the antiautomorphism of $\mathrm{Y}^+(2)$ given by (2.26). One easily verifies that $V(\gamma)^*\simeq V(1-\gamma)$ and so, the representation V^* is isomorphic to

$$L'(-\gamma_1,\gamma_2)\otimes\cdots\otimes L'(-\gamma_{2k-1},\gamma_{2k})\otimes V(\gamma_{2k+1}+1),$$

where $L'(-\gamma_{2i-1},\gamma_{2i})$ is the lowest weight representation of gl(2); see (3.31). Repeating the previous argument we prove that any nonzero submodule in V^* contains the tensor product of the lowest weight vectors of the $L'(-\gamma_{2i-1},\gamma_{2i})$ and the basis vector of $V(\gamma_{2k+1}+1)$. The proof is completed exactly as that of Proposition 5.2. □

The following proposition is proved in the same way as Proposition 5.3.

Proposition 6.3: Let $V(\mu(u))$ be an irreducible representation of $\mathrm{Y}^+(2)$ with the highest weight $\mu(u)$. Then $V(\mu(u))$ is finite-dimensional if and only if there exists a formal series $\psi(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]]$ such that $(1+1/2u^{-1})\psi(u)\mu(u)$ is a polynomial in u^{-1} with a decomposition

$$(1+\tfrac{1}{2}u^{-1})\psi(u)\mu(u)=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k+1}u^{-1}), \tag{6.8}$$

and $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$, $i=1,...,k$ for a certain re-enumeration of the γ_i. □

By analogy with Theorem 5.4 we can reformulate these conditions in terms of monic polynomials in u.

Theorem 6.4: *The irreducible hiqhest weight representation $V(\mu(u))$ of $\mathrm{Y}^+(2)$ is finite-dimensional if and only if there exists a pair $(P(u),\gamma)$, where $P(u)$ is a monic polynomial in u with $P(u)=P(-u+1)$ and $\gamma\in\mathbb{C}$ with $P(-\gamma)\neq 0$ such that*

$$\frac{\mu(-u)}{\mu(u)}=\frac{P(u+1)}{P(u)}\,\frac{(u+\gamma)(2u+1)}{(u-\gamma)(2u-1)}. \tag{6.9}$$

In this case the pair $(P(u),\gamma)$ is unique.

Proof: If $\dim V(\mu(u))<\infty$ then by Proposition 6.3 relation (6.8) holds for a certain even series $\psi(u)$. Re-enumerate the numbers γ_i to satisfy condition (6.3). By Proposition 6.2 we have $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$ for $i=1,...,k$ since $V(\mu(u))$ is finite-dimensional. Define $P(u)$ as in the proof of Theorem 5.4 by (5.30)–(5.32) and set $\gamma=\gamma_{k+1}$. The relation $P(-\gamma)\neq 0$ is implied by (6.3).

Conversely, let (6.9) be satisfied for a monic polynomial $P(u)$ such that $P(u)=P(-u+1)$ and $\gamma\in\mathbb{C}$ with $P(-\gamma)\neq 0$. Let $\{-\delta_1,\ldots,-\delta_s,\delta_1+1,...,\delta_s+1\}$ be the roots of $P(u)$; see the proof of Theorem 5.4. Set

$$\mu'(u)=(1-(\delta_1+1)u^{-1})\cdots(1-(\delta_s+1)u^{-1})\\ \times(1+\delta_1u^{-1})\cdots(1+\delta_su^{-1})(1-\gamma u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1}.$$

By Proposition 6.3 the representation $V(\mu'(u))$ is finite-dimensional. Since $\mu'(u)$ satisfies (6.9) the ratio $\mu(u)/\mu'(u)$ is an even formal series in u^{-1}. This proves that $V(\mu(u))$ is also finite-dimensional.

To prove the uniqueness of the pair $(P(u),\gamma)$ suppose that $(R(u),\delta)$ is another pair, where $R(u)$ is a monic polynomial in u such that $R(u)=R(-u+1)$, $R(-\delta)\neq 0$ and

$$\frac{P(u+1)}{P(u)}\,\frac{(u+\gamma)}{(u-\gamma)}=\frac{R(u+1)}{R(u)}\,\frac{(u+\delta)}{(u-\delta)}. \tag{6.10}$$

If $\gamma=\delta$ then we get $P(u)=R(u)$; see the proof of Theorem 3.8. Let $\gamma\neq\delta$. We prove by induction on the sum of the degrees of the polynomials $P(u)$ and $R(u)$ that (6.10) is impossible. If $P(u)=R(u)=1$ then this is obvious. Further, suppose $P(u)$ is of degree $\geqslant 2$. Take a root u_0 of $P(u)$ such that u_0+1 is not a root. Rewriting (6.10) in the form

$$P(u)R(u+1)(u+\delta)(u-\gamma)=P(u+1)R(u)(u-\delta)(u+\gamma) \tag{6.11}$$

we find that either $R(u_0)=0$, or $u_0=\delta$. If $R(u_0)=0$ then we may write

$$P(u)=P'(u)(u-u_0)(u+u_0-1),\quad R(u)=R'(u)(u-u_0)(u+u_0-1)$$

and use the induction hypotheses for the polynomials $P'(u)$ and $R'(u)$. If $u_0=\delta$ but $R(u_0)\neq 0$ then we write $P(u)=P'(u)(u-u_0)(u+u_0-1)$ and obtain from (6.11) that

$$P'(u)R(u+1)(u+u_0-1)(u-\gamma)=P'(u+1)R(u)(u-u_0+1)(u+\gamma).$$

Denote $\delta'=u_0-1$. Then we have $\delta'\neq\gamma$ because $P(-\delta')=0$. Moreover, $R(-\delta')=R(u_0)\neq 0$ and we may apply the induction hypotheses to the polynomials $P'(u)$ and $R(u)$. □

Corollary 6.5: There is a one-to-one correspondence between finite-dimensional irreducible representations of the special twisted Yangian $\mathrm{SY}^+(2)$ and the pairs $(P(u),\gamma)$, where $P(u)$ is a

monic polynomial in u such that $P(u)=P(-u+1)$ *and* $\gamma\in\mathbb{C}$ *with* $P(-\gamma)\neq 0$. *Every such representation is isomorphic to a representation of the form (6.4).* □

The following analog of Corollary 5.7 holds.

Corollary 6.6: Consider an irreducible finite-dimensional representation of $\mathrm{Y}(2)$ *of the form*

$$L(\gamma_1,-\gamma_2)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k}). \tag{6.12}$$

Then the representation

$$L(\gamma_1,-\gamma_2)\otimes\cdots\otimes L(\gamma_{2k-1},-\gamma_{2k})\otimes V(-\gamma) \tag{6.13}$$

of $\mathrm{Y}^+(2)$ *is irreducible if and only if for each pair* $1\leqslant i<j\leqslant k$ *the strings* $S(\gamma_{2i},-\gamma_{2i-1})$ *and* $S(\gamma_{2j-1},-\gamma_{2j})$ *are in general position, and for each* $i=1,\ldots,k$ *one has*

$$\gamma\notin S(\gamma_{2i-1},-\gamma_{2i}),\quad \gamma\notin S(\gamma_{2i},-\gamma_{2i-1}). \tag{6.14}$$

Proof: By Corollary 3.11, the strings $S(\gamma_{2i-1},-\gamma_{2i})$, $i=1,\ldots,k$ are pairwise in general position. Permuting the tensor factors in (6.12) if necessary, we may assume that $\gamma_1+\gamma_2\leqslant\cdots\leqslant\gamma_{2k-1}+\gamma_{2k}$. If the condition on the strings holds then one easily checks that (6.3) is satisfied for $\gamma_{2k+1}=\gamma$ and so, by Proposition 6.2 representation (6.13) is irreducible.

To prove the ''only if'' part consider first the case $k=1$. Let the representation $L(\gamma_1,-\gamma_2)\otimes V(-\gamma)$ be irreducible but condition (6.14) is violated. Note that $L(\gamma_2,-\gamma_1)\otimes V(-\gamma)$ is also irreducible. Indeed, these two representations have the same dimension and the tensor product of the highest weight vector of $L(\gamma_2,-\gamma_1)$ and the basis vector of $V(-\gamma)$ generates a highest weight module with the same highest weight as $L(\gamma_1,-\gamma_2)\otimes V(-\gamma)$. So we may assume that $\gamma\in S(\gamma_1,-\gamma_2)$ and if $\gamma+\gamma_1\in\mathbb{Z}_+$ then $\gamma+\gamma_2\leqslant\gamma+\gamma_1$. Then by Proposition 6.2 the representation $L(\gamma,-\gamma_2)\otimes V(-\gamma_1)$ is irreducible. It is isomorphic to $L(\gamma_1,-\gamma_2)\otimes V(-\gamma)$ because they have the same highest weight. Comparing the dimensions we come to a contradiction.

Consider now the general case. Let representation (6.13) be irreducible. If (6.14) is violated then we move the corresponding tensor factor in (6.13) to the right to get the tensor product of the form $L(\gamma_{2i-1},-\gamma_{2i})\otimes V(-\gamma)$. However, this representation is reducible. Contradiction.

Finally, replacing some of the tensor factors $L(\gamma_{2i-1},-\gamma_{2i})$ in (6.13), respectively, by $L(\gamma_{2i},-\gamma_{2i-1})$ we get an isomorphic representation of $\mathrm{Y}^+(2)$ (see the proof of Corollary 5.7). Hence, the representation of $\mathrm{Y}(2)$ obtained from (6.12) by this replacement is also irreducible. Therefore, by Corollary 3.11 the corresponding strings are pairwise in general position. □

Note that $V(\gamma)$ for $\gamma=1/2$ is the trivial representation of $\mathrm{Y}^+(2)$. Thus, in this particular case Corollary 6.6 provides a criterion of irreducibility of the restriction of (6.12) to $\mathrm{Y}^+(2)$.

Example 6.7: Consider the restriction of the representation $L(\gamma_1,-\gamma_2)$ of $\mathrm{Y}(2)$ with $\gamma_1+\gamma_2\in\mathbb{Z}_+$ to $\mathrm{Y}^+(2)$. By Corollary 6.6, it is reducible if and only if the string $S(\gamma_1,-\gamma_2)$ contains $-1/2$. In this case $\gamma_1,\ \gamma_2\in 1/2+\mathbb{Z}_+$. Denote the highest weight vector of $L(\gamma_1,-\gamma_2)$ by ξ and set $\xi_p=(E_{1,-1})^p\xi$. Suppose first that $\gamma_1\geqslant\gamma_2$. Then the following are $\mathrm{Y}^+(2)$-submodules in $L(\gamma_1,-\gamma_2)$:

$$L_1=\text{span of }\{\xi_0,\ldots,\xi_{\gamma_2-1/2}\},$$

$$L_2=\text{span of }\{\xi_{\gamma_1+1/2},\ldots,\xi_{\gamma_1+\gamma_2}\}.$$

By Proposition 6.2 one has $L_1\simeq L(\gamma_2,-1/2)\otimes V(-\gamma_1)$ and $L_2\simeq L(\gamma_2,-1/2)\otimes V(\gamma_1+1)$. The quotient $L(\gamma_1,-\gamma_2)/L$, where $L=L_1\oplus L_2$, is isomorphic to $L(\gamma_1,\gamma_2+1)$. If $\gamma_1\leqslant\gamma_2$ then $L(\gamma_1,-\gamma_2)$ contains the submodule

$$L=\text{span of }\{\xi_{\gamma_1+1/2},\ldots,\xi_{\gamma_2-1/2}\}$$

which is isomorphic to $L(\gamma_2,\gamma_1+1)$ and the quotient $L(\gamma_1,-\gamma_2)/L$ is isomorphic to the direct sum of $L(\gamma_1,-1/2)\otimes V(-\gamma_2)$ and $L(\gamma_1,-1/2)\otimes V(\gamma_2+1)$. □

Note that the mapping

$$s_{ij}(u)\mapsto s_{i'j'}(u), \tag{6.15}$$

where $i'=-i$ for $i=-1,1$ and $i'=i$ otherwise, defines an automorphism of the algebra $\mathrm{Y}^+(2n)$. Consider the case $n=1$ and denote the composition of this automorphism with a finite-dimensional irreducible representation $V(\mu(u))$ of $\mathrm{Y}^+(2)$ by $V(\mu(u))^\sharp$. By Theorem 4.3, this is a representation with the highest weight which we denote by $\mu(u)^\sharp$. We shall say that $\mu(u)^\sharp$ is *well-defined* if $\mu(u)$ satisfies the conditions of Proposition 6.3. Since the automorphisms (2.29) and (6.15) commute, the ratio $\mu(u)/\mu(u)^\sharp$ takes the same value for all representations from the similarity class containing $V(\mu(u))$.

Equivalently, $V(\mu(u))$ contains a unique (up to scalar multiples) nonzero vector ζ such that $s_{1,-1}(u)\zeta=0$ (cf. Remark 4.4). Then ζ is an eigenvector for $s_{-1,-1}(u)$ and $\mu(u)^\sharp$ is defined by

$$s_{-1,-1}(u)\zeta=\mu(u)^\sharp\zeta. \tag{6.16}$$

Proposition 6.8: Let $\mu(u)$ be a series given by (6.2) such that condition (6.3) is satisfied and $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$ for $i=1,\dots,k$. Then

$$\mu(u)^\sharp=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k}u^{-1})(1+(\gamma_{2k+1}+1)u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1}. \tag{6.17}$$

Proof: By Proposition 6.2 the representation $V(\mu(u))$ is isomorphic to the tensor product (6.4). In the notation of the proof of Proposition 6.2 set

$$\zeta=(E_{1,-1})^{\gamma_1+\gamma_2}\xi_1\otimes\cdots\otimes(E_{1,-1})^{\gamma_{2k-1}+\gamma_{2k}}\xi_k\otimes\xi_{k+1}.$$

Then using (2.31) we verify that ζ satisfies the relations

$$s_{-1,-1}(u)\zeta=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k}u^{-1})(1+(\gamma_{2k+1}+1)u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1}\zeta$$

and $s_{1,-1}(u)\zeta=0$. Thus, $V(\mu(u))^\sharp$ is isomorphic to $V(\mu(u)^\sharp)$ with $\mu(u)^\sharp$ given by (6.17) □.

The following theorem together with Theorem 4.3 provides a description of finite-dimensional irreducible representations of $\mathrm{Y}^+(2n)$ (cf. Theorem 5.8). We use notations (3.35) and (5.34).

Theorem 6.9: *The irreducible highest weight representation $V(\mu(u))$, $\mu(u)=(\mu_1(u),\dots,\mu_n(u))$ of $\mathrm{Y}^+(2n)$, $n\geqslant 2$ is finite-dimensional if and only if $\mu_1(u)^\sharp$ is well-defined and either one of the following four relations holds:*

$$\mu_1(-u)\Rightarrow\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u), \tag{6.18}$$

$$\frac{2u-1}{2u+1}\,\mu_1(-u)\Rightarrow\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u), \tag{6.19}$$

$$\mu_1(-u)^\sharp\Rightarrow\mu_1(u)^\sharp\to\mu_2(u)\to\cdots\to\mu_n(u), \tag{6.20}$$

$$\frac{2u-1}{2u+1}\,\mu_1(-u)^\sharp\Rightarrow\mu_1(u)^\sharp\to\mu_2(u)\to\cdots\to\mu_n(u). \tag{6.21}$$

Proof: Suppose that $\dim V(\mu(u))<\infty$. Then by Proposition 4.5 we have

$$\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u). \tag{6.22}$$

Denote the composition of $V(\mu(u))$ with the automorphism (6.15) by $V(\mu(u))^\sharp$. By Theorem 4.3 the representation $V(\mu(u))^\sharp$ is highest weight. To find its highest weight vector consider the cyclic span $V=\mathrm{Y}^+(2)\xi$, where ξ is the highest weight vector of $V(\mu(u))$, and $\mathrm{Y}^+(2)$ is identified with the subalgebra in $\mathrm{Y}^+(2n)$ generated by the coefficients of the series $s_{ij}(u)$ with $i,j=-1,1$. We claim that V is an irreducible $\mathrm{Y}^+(2)$-module. Indeed, let J be the left ideal in $\mathrm{Y}^+(2n)$ generated by the $s_{kl}^{(r)}$ with $k<l$ and $(k,l)\neq(-1,1)$. Then V is annihilated by J. This follows from the fact that $J\xi=0$ and the relations

$$[s_{ij}(u),s_{kl}(v)]\equiv 0 \bmod J, \tag{6.23}$$

where $i,j=\pm 1$, and $k<l$ with $(k,l)\neq(-1,1)$. To verify (6.23) we may assume, applying (2.18) if necessary, that $l\geqslant 2$. However, in this case (6.23) is immediate from (2.17). Now, if $\widetilde{V}$ is a nonzero submodule in V then there exists a nonzero vector $\eta\in\widetilde{V}$ such that $s_{-1,1}(u)\,\eta=0$. Then $s_{kl}(u)\,\eta=0$ for all $k<l$ since V is annihilated by J. So, η is a multiple of ξ; see Remark 4.4. This proves that $\widetilde{V}=V$.

Thus, V is isomorphic to the representation $V(\mu_1(u))$ of $\mathrm{Y}^+(2)$. Since $\dim V<\infty$ there exists a nonzero vector $\zeta\in V$ such that

$$s_{-1,-1}(u)\zeta=\mu_1(u)^{\sharp}\zeta,\quad s_{1,-1}(u)\zeta=0;$$

see (6.16). Moreover, we have $s_{kk}(u)\zeta=\mu_k(u)\zeta$ for $k=2,\ldots,n$. This is implied by the following relations which are immediate from (2.17):

$$[s_{ij}(u),s_{kk}(v)]\equiv 0 \bmod J,$$

where $i,j=\pm 1$ and $k\geqslant 2$. Since ζ is annihilated by J we may conclude that ζ is the highest weight vector of $V(\mu(u))^{\sharp}$ and so, the latter is a representation with the highest weight $(\mu_1(u)^{\sharp},\mu_2(u),\ldots,\mu_n(u))$. Applying Proposition 4.5 again, we find that

$$\mu_1(u)^{\sharp}\to\mu_2(u)\to\cdots\to\mu_n(u). \tag{6.24}$$

Further, since the representation $V(\mu_1(u))$ of $\mathrm{Y}^+(2)$ is finite-dimensional, then by Proposition 6.3 we may assume, applying an automorphism of the form (2.29) if necessary, that $(1+1/2u^{-1})\mu_1(u)$ is a polynomial in u^{-1}. By (6.22) we have

$$\frac{\mu_1(u)}{\mu_2(u)}=\frac{P(u+1)}{P(u)} \tag{6.25}$$

for a monic polynomial $P(u)$ in u. Hence, applying an automorphism (2.29) again, we may assume that both $(1+1/2u^{-1})\mu_1(u)$ and $(1+1/2u^{-1})\mu_2(u)$ are polynomials in u^{-1}:

$$\mu_1(u)=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k+1}u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1},$$

$$\mu_2(u)=(1-\delta_1u^{-1})\cdots(1-\delta_{2k+1}u^{-1})(1+\tfrac{1}{2}u^{-1})^{-1}.$$

Re-enumerating the γ_i if necessary, we may assume that condition (6.3) holds. Moreover, $\gamma_{2i-1}+\gamma_{2i}\in\mathbb{Z}_+$ for $i=1,\ldots,k$ since $\dim V(\mu_1(u))<\infty$; see Proposition 6.2. Further, (6.25) implies that there exists a re-enumeration of the δ_i such that

$$\delta_i-\gamma_i\in\mathbb{Z}_+\quad\text{for } i=1,\ldots,2k+1. \tag{6.26}$$

To see this, choose a string-type subset $\{a+1,a+2,\ldots,b\}$ of roots of $P(u)$ such that a and $b+1$ are not roots. Then we find from (6.25) that $\gamma_i=a$ for a certain i, and b equals one of the δ's which we re-denote by δ_i. Therefore, $\delta_i-\gamma_i\in\mathbb{Z}_+$ and we can write

$$\mu_1(u)=(1-au^{-1})\mu_1'(u),\quad \mu_2(u)=(1-bu^{-1})\mu_2'(u)$$

and

$$\frac{P(u+1)}{P(u)}=\frac{P'(u+1)}{P'(u)}\cdot\frac{u-a}{u-b},$$

where $P(u)=(u-a-1)\cdots(u-b)P'(u)$. Then we proceed by induction applying the same argument to the series $\mu_1'(u)$, $\mu_2'(u)$ and the polynomial $P'(u)$.

Now we use Proposition 6.8. Exactly as above, we obtain from (6.24) that there exists a permutation $\pi\in\mathfrak{S}_{2k+1}$ such that

$$\delta_{\pi(i)}-\gamma_i\in\mathbb{Z}_+,\ \ i=1,\ldots,2k,\qquad \delta_{\pi(2k+1)}+\gamma_{2k+1}+1\in\mathbb{Z}_+. \tag{6.27}$$

Now, (6.26) and (6.27) imply that $\delta_{\pi(i)}-\delta_i\in\mathbb{Z}$ for $i=1,...,2k$. But then this has to be true for $i=2k+1$ as well. Thus, $2\gamma_{2k+1}+1\in\mathbb{Z}$, that is, the number $\gamma:=\gamma_{2k+1}$ belongs to $\frac{1}{2}\mathbb{Z}$.

Suppose first that $\gamma\in\frac{1}{2}+\mathbb{Z}$ and $\gamma\geqslant-\frac{1}{2}$. Then exactly as in the proof of Theorem 5.4 we obtain that $\mu_1(-u)\Rightarrow\mu_1(u)$. Similarly, if $\gamma\in\frac{1}{2}+\mathbb{Z}$ and $\gamma\leqslant-\frac{1}{2}$ then $\mu_1(-u)^\sharp\Rightarrow\mu_1(u)^\sharp$. Further, let $\gamma\in\mathbb{Z}$ and $\gamma\geqslant 0$. Then writing

$$(1+\tfrac{1}{2}u^{-1})\mu_1(u)=(1-\gamma_1u^{-1})\cdots(1-\gamma_{2k}u^{-1})(1-\gamma u^{-1})(1-0u^{-1})$$

we see that $(1-1/2u^{-1})\mu_1(-u)\Rightarrow(1+1/2u^{-1})\mu_1(u)$, or, equivalently,

$$\frac{2u-1}{2u+1}\,\mu_1(-u)\Rightarrow\mu_1(u).$$

Similarly, if $\gamma\in\mathbb{Z}$ and $\gamma\leqslant-1$ then using Proposition 6.8 we obtain

$$\frac{2u-1}{2u+1}\,\mu_1(-u)^\sharp\Rightarrow\mu_1(u)^\sharp.$$

Combining these relations with (6.22) or (6.24) we complete the proof of the ''only if'' part.

Suppose now that relation (6.18) is satisfied. Then we repeat the corresponding argument of the proof of Theorem 5.8 to show that $V(\mu(u))$ is finite-dimensional. Similarly, if (6.20) holds then the irreducible representation with the highest weight $(\mu_1(u)^\sharp,\mu_2(u),\ldots,\mu_n(u))$ is finite-dimensional, and so is $V(\mu(u))$, since the automorphism (6.15) is involutive. If (6.19) holds then

$$\frac{\mu_1(-u)}{\mu_1(u)}=\frac{P_1(u+1)}{P_1(u)}\cdot\frac{2u+1}{2u-1} \tag{6.28}$$

for a monic polynomial $P_1(u)$ of the form (5.41), and relations (5.39) hold for polynomials $P_i(u)$ of the form (5.40). Define the $\lambda_j(u)$ by (5.42) and (5.43). Then the representation $L(\lambda(u))$ of $\mathrm{Y}(2n)$ is finite-dimensional by Theorem 3.12.

Let $V(\mu_0)$ denote the irreducible representation of the Lie algebra $\mathrm{o}(2n)$ with the highest weight $\mu_0=(-1/2,...,-1/2)$. It is finite-dimensional; see (4.6) and (4.7). Extend $V(\mu_0)$ to a representation of $\mathrm{Y}^+(2n)$ by using (2.23) and consider the tensor product $L(\lambda(u))\otimes V(\mu_0)$. The tensor product of the highest weight vectors of $L(\lambda(u))$ and $V(\mu_0)$ generates a $\mathrm{Y}^+(2n)$-submodule with the highest weight $\mu'(u)=(\mu_1'(u),\ldots,\mu_n'(u))$ where $\mu_i'(u)=\lambda_i(u)\lambda_{-i}(-u)(1+1/2u^{-1})^{-1}$; see (4.10) and (4.11). So the representation $V(\mu'(u))$ of $\mathrm{Y}^+(2n)$ is finite-dimensional. However, the polynomials $\mu_i'(u)$ satisfy (6.28) and (5.39). Hence, there exists an automorphism of $\mathrm{Y}^+(2n)$ of the form (2.29) such that its composition with the representation $V(\mu'(u))$ is isomorphic to $V(\mu(u))$ which completes the proof in this case. Finally, if relation (6.21) holds then the same argument shows that the representation $V(\mu_1(u)^\sharp,\mu_2(u),\ldots,\mu_n(u))$ is finite-dimensional, and so $V(\mu(u))$ is. □

We conclude this section with the following description of representations of the special twisted Yangian $\mathrm{SY}^+(2n)$ implied by Theorem 6.9 (cf. Corollary 5.9).

Corollary 6.10: There is a one-to-one correspondence between finite-dimensional irreducible representations of the special twisted Yangian $\mathrm{SY}^+(2n)$, $n\geqslant 2$ *and the families* $\{P_1(u),\ldots,P_n(u),\epsilon\}$, *where the* $P_i(u)$ *are monic polynomials in* u *with* $P_1(u)=P_1(-u+1)$, *and the parameter* ϵ *takes values in* $\{1,2,3,4\}$; *the values* $\epsilon=1$ *and* $\epsilon=3$ *are identified if* $P_1(1/2)\neq 0$. *Every such representation is isomorphic to a subquotient of a representation of the form* $L(\lambda(u))$ *or* $L(\lambda(u))\otimes V(\mu_0)$.

Proof: Each of the four conditions (6.18)–(6.21) allows one to naturally associate a family $\{P_1(u),\ldots,P_n(u)\}$ of monic polynomials to the corresponding similarity class of finite-dimensional irreducible representations of $\mathrm{Y}^+(2n)$; see the proof of Theorem 6.9. We associate the values of ϵ, respectively, to conditions (6.18)–(6.21). It follows from the proof of Theorem 6.9 that two families of monic polynomials with values of $\epsilon=1$ and $\epsilon=3$ correspond to the same similarity class if and only if $P_1(1/2)\neq 0$. One easily checks that identifying these values we obtain the desired one-to-one correspondence. The second claim also follows from the proof of Theorem 6.9. □

VII. REPRESENTATIONS OF $\mathrm{Y}^+(2n+1)$

The description of finite-dimensional irreducible representations of $\mathrm{Y}^+(2n+1)$ is based on the simplest nontrivial case $n=1$. Our main instrument in this case is an investigation of the restriction of an irreducible highest weight representation of $\mathrm{Y}^+(3)$ to the subalgebra $\mathrm{Y}^+(2)$ and the use of the results of Sec. VI.

Irreducible highest weight representations $V(\mu(u))$ of $\mathrm{Y}^+(3)$ are parametrized by pairs of formal series $\mu(u)=(\mu_0(u),\mu_1(u))$, where $\mu_0(u)\in 1+u^{-2}\mathbb{C}[[u^{-2}]]$ and $\mu_1(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$.

Proposition 7.1: If $\dim V(\mu(u))<\infty$ *then there exists a formal series* $\varphi(u)\in 1+u^{-1}\mathbb{C}[[u^{-1}]]$ *such that* $\varphi(u)\mu_0(u)$ *and* $\varphi(u)\mu_1(u)$ *are polynomials in* u^{-1}.

Proof: Denote by ξ the highest weight vector of $V(\mu(u))$. Let k denote the minimum nonnegative integer such that $s_{10}^{(k+1)}\xi$ is a linear combination of the vectors $\xi_1:=s_{10}^{(1)}\xi,\dots,\xi_k:=s_{10}^{(k)}\xi$. Relations (2.17) give

$$[s_{11}(u),s_{10}(v)]=\frac{1}{u-v}\,(s_{10}(v)s_{11}(u)-s_{10}(u)s_{11}(v))+\frac{u-v+1}{u^2-v^2}\,s_{1,-1}(v)s_{01}(u)$$
$$-\frac{1}{u+v}\,s_{1,-1}(u)s_{-1,0}(v)-\frac{1}{u^2-v^2}\,s_{1,-1}(u)s_{01}(v). \tag{7.1}$$

Since the vector ξ is annihilated by $s_{01}(u)$ [and hence by $s_{-1,0}(u)$] this makes the consideration here very similar to the case of Y(2); see Proposition 3.5. In particular, repeating the corresponding argument of the proof of Proposition 3.5 we can show that

$$s_{10}(u)\xi=a_1(u)\xi_1+\cdots+a_k(u)\xi_k$$

for certain series $a_i(u)$; cf. (3.13). By (2.17) we have

$$[s_{00}(u),s_{10}(v)]=\frac{u+v+1}{u^2-v^2}\,s_{10}(u)s_{00}(v)-\frac{1}{u+v}\,s_{0,-1}(u)s_{00}(v)-\frac{2v+1}{u^2-v^2}\,s_{10}(v)s_{00}(u). \tag{7.2}$$

Taking the coefficients at u^{-2} in (7.1) and (7.2) and using again the argument of the proof of Proposition 3.5 we find that the components of $\mu(u)$ can be written in the form

$$\mu_0(u)=B(u)a_k(u),\quad \mu_1(u)=C(u)a_k(u),$$

where $B(u)$ and $C(u)$ are monic polynomials in u of degree k. Thus, the series $\varphi(u)$ can be defined by $\varphi(u)=(a_k(u)u^k)^{-1}$. □

Write the polynomials $\varphi(u)\mu_0(u)$ and $\varphi(u)\mu_1(u)$ in the form

$$\begin{aligned}\varphi(u)\mu_0(u)&=(1+\alpha_1u^{-1})\cdots(1+\alpha_ku^{-1}),\\ \varphi(u)\mu_1(u)&=(1+\beta_1u^{-1})\cdots(1+\beta_ku^{-1}),\end{aligned} \tag{7.3}$$

with $\alpha_i,\ \beta_i\in\mathbb{C}$. Since $\mu_0(u)$ is an even series in u^{-1} we can find an appropriate automorphism of the form (2.29) such that its composition with $V(\mu(u))$ is an irreducible representation with the following highest weight which we shall again denote by $\mu(u)$:

$$\begin{aligned}\mu_0(u)&=(1-\alpha_1^2u^{-2})\cdots(1-\alpha_k^2u^{-2}),\\ \mu_1(u)&=(1-\alpha_1u^{-1})\cdots(1-\alpha_ku^{-1})(1+\beta_1u^{-1})\cdots(1+\beta_ku^{-1}).\end{aligned} \tag{7.4}$$

Our next task is to find necessary conditions for the α_i and β_i so that the representation $V(\mu(u))$ with the highest weight given by (7.4) be finite-dimensional.

Consider the following representation of Y(3):

$$L(\alpha_1,\alpha_1,\beta_1)\otimes\cdots\otimes L(\alpha_k,\alpha_k,\beta_k). \tag{7.5}$$

It follows from (2.15) that the $\mathrm{Y}^+(3)$-cyclic span of the tensor product of the highest weight vectors of the $L(\alpha_i,\alpha_i,\beta_i)$ is a representation of $\mathrm{Y}^+(3)$ with the highest weight $\mu(u)=(\mu_0(u),\mu_1(u))$ given by (7.4). The generators $t_{ij}^{(r)}$ of $\mathrm{Y}(3)$ with $r\geqslant k+1$ act trivially in representation (7.5) [cf. (3.18)]. Therefore, by (2.15) the generators $s_{ij}^{(r)}$ of $\mathrm{Y}^+(3)$ with $r\geqslant 2k+1$ act trivially in $V(\mu(u))$. The series

$$S_{ij}(u):=u^{2k}s_{ij}(u) \tag{7.6}$$

is then a polynomial operator in u and we may evaluate u to get well-defined operators in $V(\mu(u))$. Set

$$\begin{aligned} a_0(u)&=u^{2k}\mu_0(u)=(u^2-\alpha_1^2)\cdots(u^2-\alpha_k^2),\\ a_1(u)&=u^{2k}\mu_1(u)=(u-\alpha_1)\cdots(u-\alpha_k)(u+\beta_1)\cdots(u+\beta_k). \end{aligned} \tag{7.7}$$

Using Proposition 2.1 we shall now equip the vector space $V:=V(\mu(u))$ with another structure of $\mathrm{Y}^+(3)$-module. The matrix elements of the Sklyanin comatrix $\hat{S}(u)$ can be found from Ref. 31, Sect. VI. In particular,

$$\begin{aligned} \hat{s}_{00}(u)&=s_{11}(-u)s_{11}(u-1)-s_{1,-1}(-u)s_{-1,1}(u-1),\\ \hat{s}_{-1,0}(u)&=s_{0,-1}(-u)s_{-1,1}(u-1)-s_{01}(-u)s_{11}(u-1),\\ \hat{s}_{11}(u)&=s_{11}(-u)s_{00}(u-1)-s_{10}(-u)s_{-1,0}(u-1). \end{aligned} \tag{7.8}$$

Let us take the composition of the representation $V(\mu(u))$ with the automorphism $S(u)\mapsto \hat{S}(-u+1/2)$ of $\mathrm{Y}^+(3)$; see (2.27). One can easily see from (7.8) that the resulting representation is highest weight and generated by the vector ξ. The highest weight of this representation is

$$(\mu_1(u-1/2)\mu_1(-u-1/2),\ \ \mu_1(u-1/2)\mu_0(-u-1/2)). \tag{7.9}$$

We shall use the notation

$$\alpha_i^*=\tfrac{1}{2}-\beta_i,\quad \beta_i^*=\tfrac{1}{2}-\alpha_i,\quad \text{for } i=1,\ldots,k. \tag{7.10}$$

Set

$$\begin{aligned} \mu_0^*(u)&=(1-(\alpha_1^*)^2u^{-2})\cdots(1-(\alpha_k^*)^2u^{-2}),\\ \mu_1^*(u)&=(1-\alpha_1^*u^{-1})\cdots(1-\alpha_k^*u^{-1})(1+\beta_1^*u^{-1})\cdots(1+\beta_k^*u^{-1}). \end{aligned} \tag{7.11}$$

Take the composition of the irreducible representation of $\mathrm{Y}^+(3)$ with the highest weight (7.9) and the automorphism (2.29) corresponding to the series

$$\psi_0(u)=\frac{\mu_0^*(u)}{\mu_1(u-1/2)\mu_1(-u-1/2)}.$$

The resulting representation is isomorphic to $V(\mu^*(u))$, $\mu^*(u)=(\mu_0^*(u),\mu_1^*(u))$.

Thus, the vector space V carries two representations of $\mathrm{Y}^+(3)$ isomorphic to $V(\mu(u))$ and $V(\mu^*(u))$, respectively. To distinguish them we shall denote by $s_{ij}^*(u)$ the action of the generators $s_{ij}(u)$ in $V(\mu^*(u))$. By the above definition we have

$$s_{ij}^*(u)=\psi_0(u)\hat{s}_{ij}(-u+\tfrac{1}{2}). \tag{7.12}$$

Denote by $S_{ij}^*(u)$ the corresponding polynomial operator (7.6) in $V(\mu^*(u))$. By analogy with (7.7) set

$$a_0^*(u)=u^{2k}\mu_0^*(u), \quad a_1^*(u)=u^{2k}\mu_1^*(u).$$

For the proof of the next proposition fix an index $i\in\{1,...,k\}$ and set $\alpha:=\alpha_i$, $\beta^*:=\frac{1}{2}-\alpha_i$. Given non-negative integer p introduce the following vector in V:

$$\eta_p=S_{10}(-\alpha+p-1)\cdots S_{10}(-\alpha)\xi.$$

Proposition 7.2: We have the relations

$$S_{00}(u)\,\eta_p=a_0(u)\,\frac{u^2-(\alpha-p)^2}{u^2-\alpha^2}\,\eta_p, \tag{7.13}$$

$$S_{-1,0}(u)\,\eta_p=pa_0(u)a_1(-\alpha+p-1)\,\frac{u+\alpha-p}{u^2-\alpha^2}\,\eta_{p-1}, \tag{7.14}$$

$$S^*_{-1,1}(u)\,\eta_p=0, \tag{7.15}$$

$$S^*_{11}(u)\,\eta_p=a_1^*(u)\,\frac{u+\beta^*+p}{u+\beta^*}\,\eta_p. \tag{7.16}$$

Proof: We use induction on p to prove (7.13)–(7.16) simultaneously. For $p=0$ all the relations are obvious. Suppose now that $p\geqslant 1$. Using the induction hypotheses and (7.2) we obtain

$$S_{00}(u)\,\eta_p=S_{00}(u)S_{10}(-\alpha+p-1)\,\eta_{p-1}=\frac{u^2-(\alpha-p)^2}{u^2-(\alpha-p+1)^2}\,S_{10}(-\alpha+p-1)S_{00}(u)\,\eta_{p-1},$$

which proves (7.13). Further, (2.28) gives

$$[\hat{s}_{11}(u),s_{10}(v)]=-\frac{1}{u-v-2}\,(\hat{s}_{1,-1}(u)s_{-1,0}(v)+\hat{s}_{10}(u)s_{00}(v)+\hat{s}_{11}(u)s_{10}(v)). \tag{7.17}$$

Hence, using the induction hypotheses and (7.12) we can write

$$S^*_{11}(u)\,\eta_p=S^*_{11}(u)S_{10}(-\alpha+p-1)\,\eta_{p-1}=\frac{u+\beta^*+p}{u+\beta^*+p-1}\,S_{10}(-\alpha+p-1)S^*_{11}(u)\,\eta_{p-1},$$

which proves (7.16). The proof of (7.15) is the same; it suffices to use the expansion for $[\hat{s}_{-1,1}(u),s_{10}(v)]$ implied by (2.28). Finally, to prove (7.14) consider the following relation implied by (2.17):

$$\begin{aligned}[s_{-1,0}(u),s_{10}(v)]=&\frac{1}{u-v}\,(s_{10}(u)s_{-1,0}(v)-s_{10}(v)s_{-1,0}(u))\\&+\frac{u-v-1}{u^2-v^2}\,s_{11}(v)s_{00}(u)-\frac{1}{u+v}\,s_{-1,-1}(u)s_{00}(v)\\&+\frac{1}{u^2-v^2}\,s_{11}(u)s_{00}(v).\end{aligned} \tag{7.18}$$

Using the induction hypotheses we obtain

$$S_{-1,0}(u)\,\eta_p = S_{-1,0}(u)S_{10}(-\alpha+p-1)\,\eta_{p-1}$$
$$= \frac{u+\alpha-p}{u+\alpha-p+1}\,S_{10}(-\alpha+p-1)S_{-1,0}(u)\,\eta_{p-1}$$
$$+\frac{u+\alpha-p}{u^2-(\alpha-p+1)^2}\,S_{11}(-\alpha+p-1)S_{00}(u)\,\eta_{p-1}. \tag{7.19}$$

Using (7.13) and applying further (7.19) to $S_{-1,0}(u)\,\eta_{p-1}$, etc., we find that

$$S_{-1,0}(u)\,\eta_p = a_0(u)\,\frac{u+\alpha-p}{u^2-\alpha^2}\,\eta_p^{(1)}, \tag{7.20}$$

where

$$\eta_p^{(1)} = \sum_{q=1}^{p} S_{10}(-\alpha+p-1)\cdots S_{11}(-\alpha+q-1)\cdots S_{10}(-\alpha)\,\xi$$

and $S_{11}(-\alpha+q-1)$ takes the qth position from the right.

On the other hand, by (7.8)

$$\hat{s}_{11}(u+1) = s_{11}(-u-1)s_{00}(u) - s_{10}(-u-1)s_{-1,0}(u).$$

Let us multiply both sides of this relation by $(u(u+1))^{2k}$ and apply them to the vector η_{p-1}. By (7.12) we have

$$\hat{s}_{11}(u+1) = s_{11}^{*}(-u-\tfrac{1}{2})\,\psi_0(u+\tfrac{1}{2})^{-1}.$$

So, we can use (7.16) to calculate $(u(u+1))^{2k}\hat{s}_{11}(u+1)\,\eta_{p-1}$. Performing this calculation and using (7.13) and (7.20) we come to the relation

$$a_1(-u-1)\,\eta_{p-1} = \frac{u-\alpha+p-1}{u-\alpha}\,S_{11}(-u-1)\,\eta_{p-1} - \frac{1}{u-\alpha}\,S_{10}(-u-1)\,\eta_{p-1}^{(1)}. \tag{7.21}$$

Note that

$$\eta_p^{(1)} = S_{11}(-\alpha+p-1)\,\eta_{p-1} + S_{10}(-\alpha+p-1)\,\eta_{p-1}^{(1)}.$$

Therefore, putting $u=\alpha-p$ into (7.21) we get

$$\eta_p^{(1)} = p\,a_1(-\alpha+p-1)\,\eta_{p-1}. \tag{7.22}$$

Together with (7.20) this completes the proof of (7.14). □

Given non-negative integers $p_1,\ldots,p_k$ introduce vectors in V of a more general form:

$$\eta_{p_1,\ldots,p_k} = \left(\prod_{i=1,\ldots,k}^{\rightarrow} S_{10}(-\alpha_i+p_i-1)\cdots S_{10}(-\alpha_i)\right)\xi. \tag{7.23}$$

Set

$$S_{-1,0}^{\natural}(u) = \frac{S_{-1,0}(u)}{(u+\alpha_1-p_1)\cdots(u+\alpha_k-p_k)}.$$

Proposition 7.3: (i) *The series* $S_{-1,0}^{\natural}(u)\,\eta_{p_1,\ldots,p_k}$ *is a polynomial in u with values in V and for every* $i=1,\ldots,k$,

$$S^{\natural}_{-1,0}(\alpha_i-p_i)\,\eta_{p_1,\dots,p_k}=-a_1(-\alpha_i+p_i-1)\prod_{j=1}^{k}(\alpha_i-\alpha_j-p_i)\,\eta_{p_1,\dots,p_i-1,\dots,p_k}. \tag{7.24}$$

(ii) *One has the relation*

$$S_{00}(u)\,\eta_{p_1,\dots,p_k}=(u^2-(\alpha_1-p_1)^2)\cdots(u^2-(\alpha_k-p_k)^2)\,\eta_{p_1,\dots,p_k}. \tag{7.25}$$

Proof: First we verify (ii) by using (7.2) and repeating the corresponding argument of the proof of Proposition 7.2. Next, by an obvious induction with the use of (7.18) and (7.25) [cf. the proof of (7.20)] we show that

$$S_{-1,0}(u)\,\eta_{p_1,\dots,p_k}=(u+\alpha_1-p_1)\cdots(u+\alpha_k-p_k)\sum_{j=1}^{k}(u-\alpha_1)\cdots(u-\alpha_{j-1})$$
$$\times(u-\alpha_{j+1}+p_{j+1})\cdots(u-\alpha_k+p_k)\,\eta^{(j)}_{p_1,\dots,p_k}, \tag{7.26}$$

where $\eta^{(j)}_{p_1,\dots,p_k}$ is the vector in V obtained from $\eta_{p_1,\dots,p_k}$ by replacing the product

$$S_{10}(-\alpha_j+p_j-1)\cdots S_{10}(-\alpha_j) \tag{7.27}$$

in (7.23) with the sum

$$\sum_{q=1}^{p_j}S_{10}(-\alpha_j+p_j-1)\cdots S_{11}(-\alpha_j+q-1)\cdots S_{10}(-\alpha_j);$$

here $S_{11}(-\alpha_j+q-1)$ takes the qth position from the right. This proves that the series $S^{\natural}_{-1,0}(u)\,\eta_{p_1,\dots,p_k}$ is a polynomial in u. Moreover, (7.26) and (7.22) imply (7.24) for $i=k$. To complete the proof we show that $\eta_{p_1,\dots,p_k}$ is not changed if we arbitrarily permute the products (7.27) in the definition (7.23). We only need to show this for permutations of adjacent products. By (7.25) it suffices to verify the following auxiliary claim: if a vector $\eta\in V$ satisfies

$$S_{00}(u)\,\eta=(u^2-\gamma_1^2)\cdots(u^2-\gamma_k^2)\,\eta$$

then

$$S_{10}(\gamma_1)S_{10}(\gamma_2)\,\eta=S_{10}(\gamma_2)S_{10}(\gamma_1)\,\eta. \tag{7.28}$$

Indeed, relations (2.17) give

$$[s_{10}(u),s_{10}(v)]=-\frac{1}{u+v}\,(s_{1,-1}(u)s_{00}(v)-s_{1,-1}(v)s_{00}(u)). \tag{7.29}$$

Therefore, if $\gamma_1+\gamma_2\neq 0$ then (7.28) holds because $S_{00}(\gamma_i)\,\eta=0$. Further, we obtain from (7.29) that

$$[S_{10}(-u),S_{10}(u)]=\frac{dS_{1,-1}(u)}{du}\,S_{00}(u)-S_{1,-1}(u)\,\frac{dS_{00}(u)}{du}.$$

So, if $\gamma_1+\gamma_2=0$ then (7.28) holds since in this case η is annihilated by both the values of $S_{00}(u)$ and its derivative at $u=\gamma_2$. □

We suppose, as before, that the p_i are non-negative integers.

Proposition 7.4: The vector $\eta_{p_1,\dots,p_k}\in V$ satisfies the relations

$$S^{*}_{-1,1}(u)\,\eta_{p_1,\dots,p_k}=0, \tag{7.30}$$

$$S^{*}_{11}(u)\,\eta_{p_1,\dots,p_k}=(u-\alpha_1^{*})\cdots(u-\alpha_k^{*})(u+\beta_1^{*}+p_1)\cdots(u+\beta_k^{*}+p_k)\,\eta_{p_1,\dots,p_k}. \tag{7.31}$$

Proof: Relation (7.31) follows from (7.17) and Proposition 7.3 by an easy induction; cf. the proof of (7.16). The proof of (7.30) is the same with the use of (7.12) and the expansion for $[\hat{s}_{-1,1}(u),s_{10}(v)]$ given by (2.28). □

Recall that the representation $L(\gamma,\gamma,\delta)$ of Y(3) is finite-dimensional if and only if $\gamma-\delta \in \mathbb{Z}_+$; see Sec. III.

Proposition 7.5: Let the irreducible representation $V(\mu(u))$ of $\mathrm{Y}^+(3)$ with the components of $\mu(u)$ given by (7.4) be finite-dimensional and let $\widetilde{\mu}(u)$ be defined by (7.4) with the corresponding parameters $\widetilde{\alpha}_i$ and $\widetilde{\beta}_i$, where

$$\widetilde{\alpha}_i=\alpha_i+l_i, \quad \widetilde{\beta}_i=\beta_i-m_i, \quad i=1,...,k \tag{7.32}$$

and l_i, $m_i\in\mathbb{Z}_+$. Then the representation $V(\widetilde{\mu}(u))$ is also finite-dimensional.

Proof: For each i the representation $L(\alpha_i+l_i,\alpha_i+l_i,\alpha_i)\otimes V(\mu(u))$ of $\mathrm{Y}^+(3)$ is finite-dimensional. The tensor product of the highest weight vectors of $L(\alpha_i+l_i,\alpha_i+l_i,\alpha_i)$ and $V(\mu(u))$ generates a $\mathrm{Y}^+(3)$-module with the highest weight $\mu'(u)$, where

$$\mu_0'(u)=\mu_0(u)(1-(\alpha_i+l_i)^2u^{-2},$$

$$\mu_1'(u)=\mu_1(u)(1-(\alpha_i+l_i)u^{-1})(1+\alpha_i\mu^{-1});$$

see (2.31) and (4.11). The composition of this module and the automorphism (2.29) with the series $\psi(u)=(1-\alpha_i^2u^{-2})^{-1}$ yields a finite-dimensional representation with the highest weight obtained from $\mu(u)$ by replacing α_i with α_i+l_i.

Similarly, considering tensor products of the form $L(\beta_i,\beta_i,\beta_i-m_i)\otimes V(\mu(u))$ we show that the representation with the highest weight obtained from $\mu(u)$ by replacing β_i with β_i-m_i is also finite-dimensional. □

We are now in a position to establish necessary and sufficient conditions for $V(\mu(u))$ to be finite-dimensional.

Proposition 7.6: The irreducible highest weight representation $V(\mu(u))$ of $\mathrm{Y}^+(3)$ with the components of $\mu(u)$ given by (7.4) is finite-dimensional if and only if there exist a re-enumeration of the α_i and a re-enumeration of the β_i such that either one of the following two conditions is satisfied:

$$\alpha_i-\beta_i\in\mathbb{Z}_+ \quad \text{for } i=1,...,k; \tag{7.33}$$

$$\alpha_i-\beta_i\in\mathbb{Z}_+ \quad \text{for } i=1,...,k-1 \quad \text{and } \alpha_k\in\tfrac{1}{2}+\mathbb{Z}_+, \quad \beta_k\in-\mathbb{Z}_+. \tag{7.34}$$

Proof: Suppose that $\dim V(\mu(u))<\infty$. As before, we equip the vector space $V=V(\mu(u))$ with one more structure of $\mathrm{Y}^+(3)$-module. This module is isomorphic to $V(\mu^*(u))$ with $\mu^*(u)$ given by (7.11). Our basic idea is to construct nonzero vectors which generate highest weight $\mathrm{Y}^+(2)$-submodules in $V(\mu^*(u))$ and then use the conditions on the highest weight imposed by Proposition 6.3.

Consider the multiset $M=\{\alpha_1^*,\dots,\alpha_k^*,\beta_1^*,\dots,\beta_k^*\}$. Let us represent it as a disjoint union

$$M=\bigsqcup_{z\in\mathbb{C}/\mathbb{Z}} M_z, \quad M_z=(z+\mathbb{Z})\cap M.$$

Let

$$M_z=\{\alpha_{i_1}^*,\dots,\alpha_{i_r}^*,\beta_{j_1}^*,\dots,\beta_{j_s}^*\} \tag{7.35}$$

with both sequences written in the increasing order of the real parts, where $I_z=\{i_1,\dots,i_r\}$ and $J_z=\{j_1,\dots,j_s\}$ are subsets of $\{1,...,k\}$ depending on z.

Assume first that for each M_z the following conditions are satisfied:

$$\alpha_{i_1}^*=\cdots=\alpha_{i_r}^*, \quad \beta_{j_1}^*=\cdots=\beta_{j_s}^*, \quad \text{Re } \beta_{j_1}^*<0,$$
$$\text{and if } \beta_{j'}^*\in M_{-z} \text{ then } \alpha_{i_1}^*+\beta_{j'}^*>0. \tag{7.36}$$

Consider the vector $\eta_{p_1,\dots,p_k}$ [see (7.23)] with the numbers p_i defined as follows. For M_z given by (7.35) set $p_{j_b}=\alpha_{j_b}+\alpha_{j'}$, $b=1,\dots,s$ if there exists $\beta_{j'}^*\in M_{-z}$; otherwise set $p_{j_b}=0$. Starting with $\eta_{p_1,\dots,p_k}$ we shall apply repeatedly appropriate operators $S_{-1,0}^{\natural}(\alpha_i-q_i)$ to the vectors of the form $\eta_{q_1,\dots,q_k}$ to get the highest weight vector $\xi=\eta_{0,\dots,0}$ with a certain coefficient. By conditions (7.36) and the choice of the numbers p_i this can be done in such a way that the coefficient would be nonzero, as follows from (7.24). This proves that $\eta_{p_1,\dots,p_k}\neq 0$.

On the other hand, by Proposition 7.4 the vector $\eta_{p_1,\dots,p_k}$ generates a $\mathrm{Y}^+(2)$-submodule in $V(\mu^*(u))$ with the highest weight

$$(1-\alpha_1^*u^{-1})\cdots(1-\alpha_k^*u^{-1})(1+(\beta_1^*+p_1)u^{-1})\cdots(1+(\beta_k^*+p_k)u^{-1}). \tag{7.37}$$

This submodule is finite-dimensional. However, for all i we have $-\beta_i^*-p_i-\frac{1}{2}\notin\mathbb{Z}_+$ and for all $i<j$ we have $-\beta_i^*-p_i-\beta_j^*-p_j\notin\mathbb{Z}_+$. So, Proposition 6.3 implies that there exist re-enumerations of the α_i^* and of the β_i^* such that either

$$\alpha_i^*-\beta_i^*\in\mathbb{Z}_+ \quad \text{for } i=1,\dots,k; \tag{7.38}$$

or

$$\alpha_i^*-\beta_i^*\in\mathbb{Z}_+ \quad \text{for } i=1,\dots,k-1; \quad \text{and } \alpha_k^*\in\tfrac{1}{2}+\mathbb{Z}_+. \tag{7.39}$$

Let now the parameters α_i^* and β_i^* be arbitrary. We can choose non-negative integers l_i and m_i in such a way that conditions (7.36) are satisfied for any z by the shifted parameters $\alpha_i^*+l_i$ and $\beta_i^*-m_i$. Therefore, by Proposition 7.5 either (7.38) or (7.39) with $\mathbb{Z}_+$ replaced by $\mathbb{Z}$ will still hold.

Let us consider the multisets M_{z_0} for different z_0. Suppose first that $z_0\not\equiv 0,\frac{1}{2} \bmod \mathbb{Z}$. Then in (7.35) the index s may only be equal to r or $r+1$. Let us prove that

$$\alpha_{i_t}^*-\beta_{j_t}^*\in\mathbb{Z}_+ \quad \text{for } t=1,\dots,r. \tag{7.40}$$

Indeed, if this condition is violated then there exists an index $m\in\{1,..,r\}$ such that the number of the elements $\beta_{j_t}^*$ satisfying $\alpha_{i_m}^*-\beta_{j_t}^*\in\mathbb{Z}_+$ is less than m. Now we use Proposition 7.5 again. Applying shifts of the form (7.32), if necessary, we may assume that conditions (7.36) are satisfied for all $z\not\equiv z_0 \bmod \mathbb{Z}$ while for the elements of M_{z_0} we have:

$$\alpha_{i_1}^*=\cdots=\alpha_{i_m}^*, \quad \alpha_{i_{m+1}}^*=\cdots=\alpha_{i_r}^*,$$
$$\beta_{j_1}^*=\cdots=\beta_{j_l}^*, \quad \beta_{j_{l+1}}^*=\cdots=\beta_{j_s}^*, \tag{7.41}$$

where $\text{Re } \beta_{j_1}^*<0$, and if $\beta_{j'}^*\in M_{-z_0}$ then $\alpha_{i_r}^*+\beta_{j'}^*>0$. Here $l<m$ and the differences $\alpha_{i_1}^*-\beta_{j_1}^*$, $\beta_{j_s}^*-\alpha_{i_1}^*$ and $\alpha_{i_r}^*-\beta_{j_s}^*$ are positive integers.

Consider the vector $\eta_{p_1,\dots,p_k}$ with the p_i defined as in the previous argument except for $i\in J_{z_0}\cup J_{-z_0}$. In the latter case set $p_{j_b}=\alpha_{j_b}+\alpha_{j'}$ for $b=l+1,\dots,s$ if there exists $\beta_{j'}^*\in M_{-z_0}$; otherwise set $p_{j_b}=0$ (shifting further $\beta_{j'}^*\to\beta_{j'}^*-m_{j'}$ we may suppose that the p_{j_b} are non-negative integers). Let each of the remaining p_i with $i\in J_{z_0}\cup J_{-z_0}$ be equal to zero. Using (7.24) again we find that $\eta_{p_1,\dots,p_k}\neq 0$ and so by Proposition 7.4 it generates a finite-dimensional $\mathrm{Y}^+(2)$-submodule in $V(\mu^*(u))$ with the highest weight given by (7.37). However, one easily checks that the condition of Proposition 6.3 is not satisfied. Contradiction.

Further, if $z_0 \equiv 0 \mod \mathbb{Z}$ the same argument shows that (7.40) holds. The only difference is that here $M_{z_0} = M_{-z_0}$ and one should take j' to be equal to j_1 or j_s depending on whether $\beta^*_{j_s} \in 1 + \mathbb{Z}_+$ or $\beta^*_{j_s} \in -\mathbb{Z}_+$.

Finally, let $z_0 \equiv \frac{1}{2} \mod \mathbb{Z}$. In (7.35) the index s must be equal to r or $r-1$. We shall prove that either condition (7.40) holds (then $s = r$), or

$$\alpha^*_{i'_t} - \beta^*_{j_t} \in \mathbb{Z}_+ \quad \text{for } t = 1,...,r-1 \quad \text{and } \alpha^*_{i'_r} \in \tfrac{1}{2} + \mathbb{Z}_+ \tag{7.42}$$

for a permutation $(i'_1,\dots,i'_r)$ of the elements of I_{z_0}. Suppose that (7.40) is violated. Then we can again represent M_{z_0} in the form (7.41). If either $\beta^*_{j_s} \in \frac{3}{2} + \mathbb{Z}_+$ with $\alpha^*_{i_1} \in -\frac{1}{2} - \mathbb{Z}_+$, or $\beta^*_{j_s} \in \frac{1}{2} - \mathbb{Z}_+$ then we come to a contradiction exactly as above. So, we must have $\beta^*_{j_s} \in \frac{3}{2} + \mathbb{Z}_+$, $\alpha^*_{i_1} \in \frac{1}{2} + \mathbb{Z}_+$. Taking $j' = j_1$ and repeating the previous argument we obtain the inequality $s - l \leqslant r - m$. It can only be satisfied if $s = r-1$ and $l = m-1$. Note that this conclusion is valid for all possible values of the parameter m. Let us now assume that m takes the minimum value. Then (7.42) will be satisfied for the indices $i'_1,\dots,i'_r$ defined as follows:

$$i'_t = i_t \quad \text{for } t = 1,...,m-1,$$

$$i'_t = i_{t+1} \quad \text{for } t = m,\dots,r-1,$$

$$i'_r = i_m .$$

Thus, we have proved that if $V(\mu(u))$ is finite-dimensional then there exist re-enumerations of the α_i and of the β_i such that either (7.38), or (7.39) holds. Conditions (7.38) are equivalent to (7.33), while (7.39) means that

$$\alpha_i - \beta_i \in \mathbb{Z}_+ \quad \text{for } i = 1,...,k-1; \quad \text{and } \beta_k \in -\mathbb{Z}_+ ; \tag{7.43}$$

see (7.10). Applying this result to the representation $V(\mu^*(u))$ instead of $V(\mu(u))$ we obtain that either relation (7.33) holds, or there exist permutations $\sigma, \tau \in \mathfrak{S}_k$ such that

$$\alpha_{\sigma(i)} - \beta_{\tau(i)} \in \mathbb{Z}_+ \quad \text{for } i = 1,...,k-1; \quad \text{and } \alpha_{\sigma(k)} \in \tfrac{1}{2} + \mathbb{Z}_+ . \tag{7.44}$$

If $\sigma(k) = k$ then (7.34) is satisfied. Otherwise, set $l = \sigma(k)$ and rewrite (7.44) as follows:

$$\alpha_i - \beta_{\rho(i)} \in \mathbb{Z}_+ \quad \text{for } i \neq l; \quad \text{and } \alpha_l \in \tfrac{1}{2} + \mathbb{Z}_+ ,$$

where $\rho \in \mathfrak{S}_k$. Consider the cycle of ρ of the form $k \to i_1 \to \cdots \to i_s \to k$. Suppose that it does not contain l. Then setting $\beta'_i = \beta_{\rho(i)}$ if i belongs to the cycle, and $\beta'_i = \beta_i$ otherwise, we obtain by (7.43) that $\alpha_i - \beta'_i \in \mathbb{Z}_+$ for all $i = 1,...,k$, that is, (7.33) holds.

Suppose now that the cycle contains l: $k \to i_1 \to \cdots \to i_t \to l$. Set $\alpha'_k = \alpha_l$, $\alpha'_{\rho(i)} = \alpha_i$ for $i = k, i_1, \dots, i_t$ and $\alpha'_i = \alpha_i$ for the remaining indices i. Then by (7.43) relation (7.34) is satisfied for the α'_i and β_i which completes the proof of the ''only if'' part of the proposition.

Now let (7.33) hold. Then representation (7.5) of $\mathrm{Y}^+(3)$ is finite-dimensional. However, $V(\mu(u))$ is isomorphic to its subquotient and therefore $\dim V(\mu(u)) < \infty$.

Finally, let (7.34) hold. Consider the highest weight representation $V(\mu)$ of the Lie algebra $\mathrm{o}(3)$ and extend it to $\mathrm{Y}^+(3)$ by using the homomorphism (2.23). We obtain a representation with the highest weight given by (4.10). By (4.9) the representation $V(\mu)$ is finite-dimensional for $\mu = \beta_k - \frac{1}{2}$. Therefore the representation of $\mathrm{Y}^+(3)$ of the form

$$L(\alpha_1,\alpha_1,\beta_1) \otimes \cdots \otimes L(\alpha_{k-1},\alpha_{k-1},\beta_{k-1}) \otimes L(\alpha_k,\alpha_k,\tfrac{1}{2}) \otimes V(\beta_k - \tfrac{1}{2}) \tag{7.45}$$

is also finite-dimensional. However, we find from (2.31) and (4.11) that $V(\mu(u))$ is isomorphic to a subquotient of (7.45) and so, $\dim V(\mu(u)) < \infty$. □

The following theorem together with Theorem 4.3 provides a description of finite-dimensional irreducible representations of $\mathrm{Y}^+(2n+1)$ (cf. Theorems 5.8 and 6.9). We use notation (3.35).

Theorem 7.7: *The irreducible highest weight representation* $V(\mu(u)),\mu(u)=(\mu_0(u),\ldots,\mu_n(u))$ *of* $\mathrm{Y}^+(2n+1)$ *is finite-dimensional if and only if either one of the following two relations holds:*

$$\mu_0(u)\to\mu_1(u)\to\cdots\to\mu_n(u), \tag{7.46}$$

$$\frac{2u}{2\mu+1}\,\mu_0(u)\to\mu_1(u)\to\cdots\to\mu_n(u). \tag{7.47}$$

Proof: Suppose that $\dim V(\mu(u))<\infty$. Then by Proposition 4.5 we have

$$\mu_1(u)\to\mu_2(u)\to\cdots\to\mu_n(u).$$

The subalgebra in $\mathrm{Y}^+(2n+1)$ generated by the coefficients of the series $s_{ij}(u)$ with $i,j=-1,0,1$ is isomorphic to $\mathrm{Y}^+(3)$. The cyclic span of the highest weight vector of $V(\mu(u))$ with respect to this subalgebra is a representation with the highest weight $(\mu_0(u),\mu_1(u))$. Its irreducible quotient is finite-dimensional and so, by Proposition 7.6 we have $\mu_0(u)\to\mu_1(u)$ or $(2u/2u+1)\mu_0(u)\to\mu_1(u)$ depending on whether condition (7.33) or (7.34) is satisfied; see the proof of Theorem 3.8.

Conversely, let relation (7.46) hold. Then for $i=1,\ldots,n$ we have

$$\frac{\mu_{i-1}(u)}{\mu_i(u)}=\frac{P_i(u+1)}{P_i(u)} \tag{7.48}$$

for monic polynomials $P_i(u)$. Let

$$P_i(u)=(u+\delta_1^{(i)})\cdots(u+\delta_{s_i}^{(i)}),\quad i=1,\ldots,n. \tag{7.49}$$

Consider the irreducible highest weight representation $L(\lambda(u))$ with $\lambda(u)=(\lambda_{-n}(u),\ldots,\lambda_n(u))$ of the Yangian $\mathrm{Y}(2n+1)$ where the $\lambda_i(u)$ are defined by (5.42) for $i=1,\ldots,n$ and by (5.43) for $i=0,\ldots,n$. Then $L(\lambda(u))$ is finite-dimensional by Theorem 3.12 and we conclude by repeating the corresponding argument of the proof of Theorem 5.8.

Finally, let (7.47) hold. Then there exist monic polynomials (7.49) such that (7.48) holds for $i=2,\ldots,n$ and

$$\frac{\mu_0(u)}{\mu_1(u)}=\frac{P_1(u+1)}{P_1(u)}\,\frac{2u+1}{2u}. \tag{7.50}$$

Let $V(\mu_0)$ denote the irreducible representation of the Lie algebra $\mathrm{o}(2n+1)$ with the highest weight $\mu_0=(-1/2,\ldots,-1/2)$. It is finite-dimensional; see (4.6) and (4.9). Extend $V(\mu_0)$ to a representation of $\mathrm{Y}^+(2n+1)$ by using (2.23) and consider the tensor product $L(\lambda(u))\otimes V(\mu_0)$, where $L(\lambda(u))$ is the defined above representation of $\mathrm{Y}(2n+1)$. The tensor product of the highest weight vectors of $L(\lambda(u))$ and $V(\mu_0)$ generates a $\mathrm{Y}^+(2n+1)$-submodule with the highest weight $\mu'(u)=(\mu_0'(u),\ldots,\mu_n'(u))$ where $\mu_i'(u)=\lambda_i(u)\lambda_{-i}(-u)(1+1/2\mu^{-1})^{-1}$ for $i=1,\ldots,n$ and $\mu_0'(u)=\lambda_0(u)\lambda_0(-u)$; see (2.31) and (4.11). Therefore the representation $V(\mu'(u))$ of $\mathrm{Y}^+(2n+1)$ is finite-dimensional. Moreover, the $\mu_i'(u)$ satisfy both (7.48) for $i=2,\ldots,n$, and (7.50). It remains to repeat the corresponding argument of the proof of Theorem 6.9. □

Theorem 7.7 implies the following parametrization of representations of the special twisted Yangian $\mathrm{SY}^+(2n+1)$ (cf. Corollaries 5.9 and 6.10).

Corollary 7.8: There is a one-to-one correspondence between finite-dimensional irreducible representations of the special twisted Yangian $\mathrm{SY}^+(2n+1)$ *and the families* $\{P_1(u),\ldots,P_n(u),\epsilon\}$, *where the* $P_i(u)$ *are monic polynomials in u and the parameter* ϵ *takes values in* $\{1,2\}$. *Every such representation is isomorphic to a subquotient of a representation of the form* $L(\lambda(u))$ *or* $L(\lambda(u))\otimes V(\mu_0)$. □

Remark 7.9: One can use Propositions 7.3 and 7.4 to get a branching rule for the restriction of a ''generic'' finite-dimensional irreducible representation V of $\mathrm{Y}^+(3)$ to the subalgebra $\mathrm{Y}^+(2)$ and thus to construct an analog of Gelfand-Tsetlin basis in V; cf. Refs. 17 and 18. □

ACKNOWLEDGMENTS

During this work I benefited from collaboration with G. Olshanski to whom I would like to express my sincere gratitude. I would also like to thank M. Nazarov, V. Tarasov and R. Zhang for valuable discussions.

[1] V. G. Drinfeld, ''Hopf algebras and the quantum Yang–Baxter equation,'' Sov. Math. Dokl. **32**, 254–258 (1985).
[2] V. G. Drinfeld, ''A new realization of Yangians and quantized affine algebras,'' Sov. Math. Dokl. **36**, 212–216 (1988).
[3] L. A. Takhtajan and L. D. Faddeev, ''Quantum inverse scattering method and the Heisenberg XYZ-model,'' Russ. Math. Surv. **34**, 11–68 (1979).
[4] D. Bernard and A. LeClair, ''Quantum group symmetries and non-local currents in 2D QFT,'' Commun. Math. Phys. **142**, 99–138 (1991).
[5] P. P. Kulish and E. K. Sklyanin, ''Quantum spectral transform method: Recent developments,'' in *Integrable Quantum Field Theories*, Lecture Notes in Physics Vol. 151 (Springer, Berlin, 1982), pp. 61–119.
[6] K. Takemura and D. Uglov, ''The orthogonal eigenbasis and norms of eigenvectors in the Spin Calogero–Sutherland Model,'' J. Phys. A **30**, 3685–3717 (1997).
[7] D. B. Uglov and V. E. Korepin, ''The Yangian symmetry of the Hubbard model,'' Phys. Lett. A **190**, 238–242 (1994).
[8] V. O. Tarasov, ''Structure of quantum L-operators for the R-matrix of the XXZ-model,'' Theor. Math. Phys. **61**, 1065–1071 (1984).
[9] V. O. Tarasov, ''Irreducible monodromy matrices for the R-matrix of the XXZ-model and lattice local quantum Hamiltonians,'' Theor. Math. Phys. **63**, 440–454 (1985).
[10] V. E. Korepin, ''The analysis of the bilinear relation of the six-vertex model,'' Sov. Phys. Dokl. **27**, 612–613 (1982).
[11] V. Chari and A. Pressley, ''Yangians and R-matrices,'' L'Enseign. Math. **36**, 267–302 (1990).
[12] V. Chari and A. Pressley, ''Fundamental representations of Yangians and singularities of R-matrices,'' J. Reine Angew. Math. **417**, 87–128 (1991).
[13] V. Chari and A. Pressley, ''Yangians: Their representations and characters,'' Acta Appl. Math. **44**, 39–58 (1996).
[14] I. V. Cherednik, ''A new interpretation of Gelfand–Tzetlin bases,'' Duke Math. J. **54**, 563–577 (1987).
[15] I. V. Cherednik, ''Quantum groups as hidden symmetries of classic representation theory,'' in *Differential Geometric Methods in Physics*, edited by A. I. Solomon (World Scientific, Singapore, 1989), pp. 47–54.
[16] A. N. Kirillov and N. Yu. Reshetikhin, ''Yangians, Bethe ansatz and combinatorics,'' Lett. Math. Phys. **12**, 199–208 (1986).
[17] A. Molev, ''Gelfand–Tsetlin basis for representations of Yangians,'' Lett. Math. Phys. **30**, 53–60 (1994).
[18] M. Nazarov and V. Tarasov, ''Representations of Yangians with Gelfand–Zetlin bases,'' J. Reine Angew. Math. **496**, 181–212 (1998).
[19] A. Molev, ''Noncommutative symmetric functions and Laplace operators for classical Lie algebras,'' Lett. Math. Phys. **35**, 135–143 (1995).
[20] M. Nazarov and V. Tarasov, ''Yangians and Gelfand–Zetlin bases,'' Publ. RIMS, Kyoto Univ. **30**, 459–478 (1994).
[21] M. Nazarov, ''Quantum Berezinian and the classical Capelli identity,'' Lett. Math. Phys. **21**, 123–131 (1991).
[22] M. Nazarov, ''Yangians and Capelli identities,'' in *Kirillov's Seminar on Representation Theory*, edited by G. I. Olshanski, American Mathematical Society Translations Vol. 181 (American Mathematical Society, Providence, RI, 1998), pp. 139–164.
[23] A. Okounkov, ''Quantum immanants and higher Capelli identities,'' Transformation Groups **1**, 99–126 (1996).
[24] G. Olshanski, ''Representations of infinite-dimensional classical groups, limits of enveloping algebras, and Yangians,'' in *Topics in Representation Theory*, edited by A. A. Kirillov, Advances in Soviet Mathematics Vol. 2 (American Mathematical Society, Providence, RI, 1991), pp. 1–66.
[25] G. Olshanski, ''Twisted Yangians and infinite-dimensional classical Lie algebras,'' in *Quantum Groups*, edited by P. P. Kulish, Lecture Notes in Mathematics Vol. 1510 (Springer, Berlin, 1992), pp. 103–120.
[26] E. K. Sklyanin, ''Boundary conditions for integrable quantum systems,'' J. Phys. A **21**, 2375–2389 (1988).
[27] T. H. Koornwinder and V. B. Kuznetsov, ''Gauss hypergeometric function and quadratic R-matrix algebras,'' St. Petersburg Math. J. **6**, 161–184 (1994).
[28] P. P. Kulish and E. K. Sklyanin, ''Algebraic structures related to reflection equations,'' J. Phys. A **25**, 5963–5975 (1992).
[29] V. B. Kuznetsov, M. F. Jørgensen, and P. L. Christiansen, ''New boundary conditions for integrable lattices,'' J. Phys. A **28**, 4639 (1995).
[30] A. Molev, M. Nazarov, and G. Olshanski, ''Yangians and classical Lie algebras,'' Russ. Math. Surveys **51**, 205–282 (1996).
[31] A. Molev, ''Sklyanin determinant, Laplace operators, and characteristic identities for classical Lie algebras,'' J. Math. Phys. **36**, 923–943 (1995).
[32] M. Nazarov and G. Olshanski, ''Bethe subalgebras in twisted Yangians,'' Commun. Math. Phys. **178**, 483–506 (1996).
[33] A. Molev and M. Nazarov (unpublished); q-alg/9712021.
[34] R. B. Zhang, ''Representations of super Yangian,'' J. Math. Phys. **36**, 3854–3865 (1995).
[35] R. B. Zhang, ''The gl$(M|N)$ super Yangian and its finite-dimensional representations,'' Lett. Math. Phys. **37**, 419–434 (1996).
[36] A. Molev, ''Representations of twisted Yangians,'' Lett. Math. Phys. **26**, 211–218 (1992).
[37] J. Dixmier, *Algèbres Enveloppantes* (Gauthier-Villars, Paris, 1974).

Invariant tensors and Casimir operators for simple compact Lie groups

Arthur J. Mountain[a)]
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, United Kingdom

(Received 17 April 1998; accepted for publication 8 June 1998)

The Casimir operators of a Lie algebra are in one-to-one correspondence with the symmetric invariant tensors of the algebra. There is an infinite family of Casimir operators whose members are expressible in terms of a number of primitive Casimirs equal to the rank of the underlying group. A systematic derivation is presented of a complete set of identities expressing nonprimitive symmetric tensors in terms of primitive tensors. Several examples are given including an application to an exceptional Lie algebra. © *1998 American Institute of Physics.*
[S0022-2488(98)01510-2]

I. INTRODUCTION

The Casimir operators for a simple Lie algebra are constructed from symmetric tensors invariant under the adjoint action of the group. Given a simple compact Lie group of rank l, it is known that there are l such primitive symmetric invariant tensors of order m_i, $i=1\cdots l$. The orders m_i are in fact also known because of the correspondence of the primitive symmetric invariant tensors with cocycles of order $(2m_i-1)$ which are representatives of the de Rham cohomology classes of G. However, most methods of constructing symmetric invariant tensors give rise naturally to an infinite family of them, of which only the finite number l can be primitive.

In this paper, we consider the set of invariant tensors arising from symmetrized traces of products of Lie algebra basis elements. These correspond to Casimir operators and arise naturally in many physical theories. It is often useful to be able to express this (infinite) family of tensors in terms of the l primitive tensors (see above). We present a method which systematically generates a complete set of identities expressing nonprimitive symmetric invariant tensors of this form in terms of primitive tensors. We illustrate our methods for the classical series of Lie algebras and perform explicit calculations for the examples A_2, B_2, and G_2.

A. Notation and conventions

In this paper, we will follow the ''physical'' convention for Lie groups and use Hermitian generators $\{X_i\}$ and the normalization $\mathrm{Tr}(X_iX_j)=2\delta_{ij}$. When the generators are assumed to be in *matrix* form, we take them to be in the defining representation of the algebra. We will assume there is no distinction between upper and lower indices, their position being dictated by notational convenience. We will take $\epsilon_{i_1i_2\cdots i_n}$ to be the standard totally antisymmetric tensor, with $\epsilon_{123\cdots n}=1$. We will define from this the unit-weight symbol

$$\epsilon^{j_1j_2\cdots j_n}_{i_1i_2\cdots i_n}=\frac{1}{n!}\epsilon_{i_1i_2\cdots i_n}\epsilon^{j_1j_2\cdots j_n}=\frac{1}{n!}\sum_{\sigma\in S_n}(-1)^{\pi(\sigma)}\delta^{j_{\sigma(1)}}_{i_1}\cdots\delta^{j_{\sigma(n)}}_{i_n}, \tag{1.1}$$

where $\pi(\sigma)$ is the parity of the permutation σ.

II. INVARIANT SYMMETRIC TENSORS AND CASIMIRS

Let $\mathcal{G}$ be a simple algebra of rank l with basis $\{X_i\}$, $[X_i,X_j]=f^k_{ij}X_k$, $i=1\cdots\dim\mathcal{G}$, and let G be its (compact) associated Lie group. Let $\{\omega^j\}$ be the dual basis in $\mathcal{G}^*$, $\omega^j(X_i)=\delta^j_i$, and consider a G-invariant symmetric tensor h of order m

a)Electronic mail: A.J.Mountain@damtp.cam.ac.uk

 © 1998 American Institute of Physics

$$h = h_{i_1 \cdots i_m} \omega^{i_1} \otimes \cdots \otimes \omega^{i_m}. \tag{2.1}$$

The G-invariance (ad-invariance) of h means that

$$\sum_{s=1}^{m} f^{\rho}_{\nu i_s} h_{i_1 \cdots \hat{i}_s \rho i_{s+1} \cdots i_m} = 0. \tag{2.2}$$

This is the case for the symmetric tensors $k^{(m)}$ given by the coordinates

$$k^{(m)}_{i_1 \cdots i_m} = \mathrm{sTr}(X_{i_1} \cdots X_{i_m}) \equiv \mathrm{Tr}(X_{(i_1} \cdots X_{im)}). \tag{2.3}$$

where sTr is the unit-weight symmetric trace,

$$\mathrm{sTr}(X_{i_1} \cdots X_{i_m}) = \frac{1}{m!} \Sigma_{\sigma \in S_m} \mathrm{Tr}(X_{i_{\sigma(1)}} \cdots X_{i_{\sigma(m)}}).$$

Indices inside round brackets $(i_1, \ldots, i_m)$ will also be understood as symmetrized with unit weight. In fact,[1] a complete set of l primitive (see below) invariant tensors may be constructed in this way by selecting suitable representations.

Let G be compact so that the Killing tensor may be taken as the unit matrix and let $h_{i_1 \cdots i_m}$ be an arbitrary symmetric invariant tensor. Then the order m element in the enveloping algebra $\mathcal{U}(\mathcal{G})$ defined by

$$\mathcal{C}^{(m)} = h^{i_1 \cdots i_m} X_{i_1} \cdots X_{i_m}, \tag{2.4}$$

commutes with all elements of $\mathcal{G}$. This is so because the commutator $[X_\rho, \mathcal{C}^{(m)}]$ may be written as

$$[X_\rho, \mathcal{C}^{(m)}] = \sum_{s=1}^{m} f^{i_s}_{\rho \nu} h^{i_1 \cdots \hat{i}_s \nu \cdots i_m} X_{i_1} \cdots X_{i_m} = 0, \tag{2.5}$$

which is indeed zero as a result of the invariance condition (2.2). In fact, the only conditions for the m-tensor h to generate a Casimir operator $\mathcal{C}^{(m)}$ of $\mathcal{G}$ of order m are its symmetry (nonsymmetric indices would allow us to reduce the order m of $\mathcal{C}^{(m)}$ by replacing certain products of generators by commutators) and its invariance [Eq. (2.5)]; h does not need to be obtained from a symmetric trace (2.3). Thus for any invariant symmetric tensor h of order m, $\mathcal{C}^{(m)} = h^{i_1 \cdots i_m} X_{i_1} \cdots X_{i_m}$ is a Casimir of $\mathcal{G}$ of the same order m.

It is well known[1–8] that a simple algebra of rank l has l independent (primitive) Casimir–Racah operators of order $m_1, \cdots, m_l$, the first of them given by the familiar quadratic Casimir[9] operator $k_{ij} X^i X^j$ obtained from the Killing tensor $(m_1 = 2)$. Thus there must be (Cayley–Hamilton) relations among the invariant tensors obtained from (2.3) for $m \neq m_l$ or otherwise one would obtain an arbitrary number of primitive Casimirs. We shall demonstrate this explicitly for several examples.

III. RELATIONS AMONG SYMMETRIC TENSORS

The results in this section involve an important result that, for a matrix A, we have the relationship

$$\det A = \exp(\mathrm{Tr} \log A). \tag{3.1}$$

This is clearly true for diagonal A and can be seen to be true generally by conjugation of A. We apply this to the case $A = \mathbf{1} + \alpha F$, with $\mathbf{1}$ the identity matrix and α an arbitrary parameter. We can treat the right hand side of (3.1) as a power series in α, giving

$$\det(\mathbf{1}+\alpha F)=1-\frac{\alpha^2}{2}\mathrm{Tr}\,F^2+\frac{\alpha^3}{3}\mathrm{Tr}\,F^3+\alpha^4\left[-\frac{1}{4}\mathrm{Tr}\,F^4+\frac{1}{8}(\mathrm{Tr}\,F^2)^2\right]$$
$$+\alpha^5\left[\frac{1}{5}\mathrm{Tr}\,F^5-\frac{1}{6}\mathrm{Tr}\,F^2\mathrm{Tr}\,F^3\right]$$
$$+\alpha^6\left[-\frac{1}{6}\mathrm{Tr}\,F^6+\frac{1}{18}(\mathrm{Tr}\,F^3)^2+\frac{1}{8}\mathrm{Tr}\,F^2\mathrm{Tr}\,F^4-\frac{1}{48}(\mathrm{Tr}\,F^2)^3\right]$$
$$+\alpha^7\left[\frac{1}{7}\mathrm{Tr}\,F^7-\frac{1}{10}\mathrm{Tr}\,F^2\mathrm{Tr}\,F^5-\frac{1}{12}\mathrm{Tr}\,F^3\mathrm{Tr}\,F^4+\frac{1}{24}(\mathrm{Tr}\,F^2)^2\mathrm{Tr}\,F^3\right]$$
$$+\alpha^8\left[-\frac{1}{8}\mathrm{Tr}\,F^8+\frac{1}{12}\mathrm{Tr}\,F^2\mathrm{Tr}\,F^6+\frac{1}{15}\mathrm{Tr}\,F^3\mathrm{Tr}\,F^5+\frac{1}{32}(\mathrm{Tr}\,F^4)^2\right.$$
$$\left.-\frac{1}{32}\mathrm{Tr}\,F^4(\mathrm{Tr}\,F^2)^2-\frac{1}{36}(\mathrm{Tr}\,F^3)^2\mathrm{Tr}\,F^2+\frac{1}{16.4!}(\mathrm{Tr}\,F^2)^4\right]+\cdots. \tag{3.2}$$

In the above, $\mathrm{Tr}\,F^n$ is defined as $F_{i_1i_2}F_{i_2i_3}\cdots F_{i_ni_1}$. All the terms to $\mathcal{O}(\alpha^8)$ are given as these are the terms we shall use below. It is easy to calculate the terms in higher powers of α. We will now show how to express nonprimitive symmetric invariant tensors in terms of primitive tensors.

A. The algebras A_l

We consider a matrix representation of SU(n) with matrix generators $\{X_i\}$. For our purposes we will always work with the fundamental representation so that X_i are $n\times n$ matrices. The rank of the group is $l=n-1$ and its Lie algebra is $A_l=\mathrm{su}(l+1)$. A basis of the symmetric invariant tensors of A_l is given by the order m symmetric traces $k^{(m)}, m=2,3,\ldots$ from Eq. (2.3). The algebra A_l has l primitive symmetric invariant tensors. A basis of these is given by $k^{(m)}$, $m=2\cdots n$.

We now define the matrix F [see Eq. (3.2)] to be a linear combination of the X_i, introducing a vector $\mathbf{y}$, as $(F)_{ab}=y^i(X_i)_{ab}$. As $\{X_i\}$ are $n\times n$ matrices, no terms of $\mathcal{O}(\alpha^m)$ can appear on the right hand side of (3.2) for $m>n$. Thus these terms must vanish identically, giving us one equation at each order in α. This gives us one equation for each of the nonprimitive tensors of the form (2.3) in terms of primitive tensors.

B. The algebras B_l and C_l

The fundamental matrix representations of the algebras $B_l=\mathrm{so}(2l+1)$ and $C_l=\mathrm{sp}(2l)$ are in terms of $(2l+1)\times(2l+1)$ and $2l\times 2l$ matrices, respectively. In their defining representations these groups preserve the euclidean or symplectic metric η, respectively, and thus the generators X_i of these algebras satisfy

$$X_i\eta=-\eta X_i^t, \tag{3.3}$$

where $i=1,\ldots,l(2l+1)$. The symmetrized products of an *odd* number of generators of the orthogonal and symplectic groups also satisfies (3.3); hence, they are a member of their respective algebras. In particular, as the generators of both B_l and C_l are traceless, we find that the odd-order tensors $k^{(2m+1)}$ [see Eq. (2.3)], arising from symmetrized traces, vanish identically. For both B_l and C_l, we have the property that a basis for the symmetric invariant tensors is given by $k^{(m)}$ for $m=2,4,6,\ldots$. Thus for these examples, all odd powers of α in the expansion (3.2) vanish identically. A basis for the primitive symmetric invariant tensors is given by $k^{(m)}$, $m=2,4,6,\ldots 2l$. As above, we can write $F_{ab}=y^i(X_i)_{ab}$, where $\{X_i\}$ is a basis for the required Lie algebra in the fundamental representation. Substituting this into (3.2), we see that all terms of $\mathcal{O}(\alpha^m)$ must vanish for $m>2l$. This gives precisely one equation for each of the nonprimitive tensors of the form (2.3) in terms of primitive tensors.

C. The algebras D_l

The algebra $D_l = \mathrm{so}(2l)$ is the algebra of antisymmetric $(2l)\times(2l)$ matrices. The odd-order tensors $k^{(2m+1)}$ vanish identically using an identical argument to that used above for B_l. There are l primitive symmetric invariant tensors of which only $(l-1)$ can be expressed in the form (2.3). These are $k^{(m)}$, $m=2,4,6,\ldots,2(l-1)$. The final primitive tensor is the Pfaffian, given by

$$Pf_{i_1 i_2 \cdots i_l} = \frac{1}{2^l l!}\epsilon^{a_1 a_2 \cdots a_{2l}}(X_{i_1})_{a_1 a_2}(X_{i_2})_{a_3 a_4}\cdots(X_{i_l})_{a_{(2l-1)}a_{2l}}. \tag{3.4}$$

As before, we introduce a vector $\mathbf{y}$ and write $F_{ab} = y_i (X^i)_{ab}$. We rewrite the expression (3.2) in the form

$$\begin{aligned}\det(\mathbf{1}+\alpha F) = 1 &+ \alpha^2 \epsilon^{b_1 b_2}_{a_1 a_2} F^{a_1}{}_{b_1} F^{a_2}{}_{b_2} + \alpha^4 \epsilon^{b_1 b_2 b_3 b_4}_{a_1 a_2 a_3 a_4} F^{a_1}{}_{b_1} F^{a_2}{}_{b_2} F^{a_3}{}_{b_3} F^{a_4}{}_{b_4} + \cdots \\ &+ \alpha^{2n} \epsilon^{b_1 b_2 \cdots b_{2n}}_{a_1 a_2 \cdots a_{2n}} F^{a_1}{}_{b_1} F^{a_2}{}_{b_2} \ldots F^{a_{2n}}{}_{b_{2n}} + \cdots .\end{aligned} \tag{3.5}$$

The terms of $\mathcal{O}(\alpha^{2m})$ for $m>l$ express the nonprimitive symmetric tensors $k^{(2m)}$ [see Eq. (2.3)] in terms of primitive tensors. For the $\mathcal{O}(\alpha^{2l})$ terms, we must use the identity

$$\begin{aligned}\epsilon^{b_1 b_2 \cdots b_{2l}}_{a_1 a_2 \cdots a_{2l}} F^{a_1}{}_{b_1} F^{a_2}{}_{b_2} \cdots F^{a_{2l}}{}_{b_{2l}} &= \left(\frac{1}{2^l l!} \epsilon_{a_1 a_2 \cdots a_{2l}} F_{a_1 a_2} F_{a_3 a_4} \cdots F_{a_{(2l-1)} a_{2l}} \right)^2 \\ &= Pf_{(i_1 \cdots i_l} Pf_{i_{l+1} \cdots i_{2l})} y^{i_1} \cdots y^{i_{2l}}.\end{aligned} \tag{3.6}$$

This expresses the nonprimitive tensor $k^{(2l)}$ in terms of the Pfaffian Pf and the $(l-1)$ tensors $k^{(2m)}$, $m<l$, which together are the l primitive symmetric tensors of D_l.

IV. EXAMPLES

Below are given three explicit examples of construction of nonprimitive tensors for simple Lie algebras in terms of primitive tensors. Connections are also made to the geometry of the associated Lie groups. In particular, simple compact Lie groups can be expressed as products of spheres of odd dimension, as far as the cohomology is concerned. Decomposing a Lie group into such a product gives the de Rham cohomology as the Betti numbers[10] for a sphere S^n are known to be

$$b^{(0)}=1,\ b^{(n)}=1,\ b^{(k)}=0 \quad \text{otherwise.} \tag{4.1}$$

We can also associate with each symmetric invariant tensor k of order m an antisymmetric invariant tensor of order $(2m-1)$ according to

$$\Omega_{i_1 \cdots i_{2m-2}\sigma} = f^{j_1}_{[i_1 i_2} \cdots f^{j_{m-1}}_{i_{2m-3} i_{2m-2}} k_{\sigma] j_1 \cdots j_{m-1}}. \tag{4.2}$$

If k is primitive, this gives a $(2m-1)$-form on G representing a Lie algebra cocycle of order $(2m-1)$. If k is nonprimitive, the expression is identically zero.[11]

A. The algebras A_l

The group $\mathrm{SU}(l+1)$ has algebra $\mathrm{su}(l+1)=A_l$. The cohomology of the group manifold is the cohomology of l spheres as $\mathrm{SU}(l+1)\to S^{2l+1}\times S^{2l-1}\times\cdots\times S^3$. Thus there are l nontrivial cocycles and l corresponding primitive symmetric invariant tensors as in (4.2). There is a basis for the symmetric invariant tensors of A_l given by Sudbery,[12] consisting of one symmetric tensor $d^{(k)}$ at each order $k=2,3,\ldots$. Given a normalized basis $\{\lambda_i\}$ for $\mathrm{su}(l+1)$, we have the multiplication law

$$\lambda_i \lambda_j = \frac{2}{l+1}\delta_{ij} + (d_{ijk} + i f_{ijk})\lambda_k. \tag{4.3}$$

The tensors f and d are, respectively, the totally antisymmetric and symmetric structure constants. The two lowest-order symmetric invariant tensors are δ_{ij} and d_{ijk}. In terms of these, the Sudbery basis for symmetric tensors is

$$\begin{aligned} d^{(4)}_{i_1i_2i_3i_4} &= d_{(i_1i_2}{}^x d_{i_3i_4)x}, \\ d^{(5)}_{i_1i_2i_3i_4i_5} &= d_{(i_1i_2}{}^x d^x{}_{i_3}{}^y d_{i_4i_5)y}, \\ d^{(6)}_{i_1i_2i_3i_4i_5i_6} &= d_{(i_1i_2}{}^x d^x{}_{i_3}{}^y d^y{}_{i_4}{}^z d_{i_5i_6)z}. \\ &\vdots \end{aligned} \tag{4.4}$$

We can choose a basis $d^{(k)}$, $k=2,3,\ldots,l+1$ of the primitive symmetric invariant tensors of A_l. These will give rise to representatives of the l primitive cocycles of A_l via (4.2).

B. The algebra A_2

The fundamental representation of the algebra $A_2=\mathrm{su}(3)$ is in terms of eight (3×3) Hermitian matrix generators $\{\lambda_i\}$. The rank of A_2 is two so we have two primitive invariant symmetric tensors; these are given by

$$\tfrac{1}{2}\,\mathrm{Tr}(\lambda_i\lambda_j)=\delta_{ij}, \quad \tfrac{1}{2}\,\mathrm{sTr}(\lambda_i\lambda_j\lambda_k)=d_{ijk}. \tag{4.5}$$

As far as the cohomology is concerned, the group manifold SU(3) can be decomposed into $S^5\times S^3$ and there are three- and five-cocycles associated with δ_{ij} and d_{ijk}, respectively, [see (4.2)]. Defining an arbitrary eight-vector $\mathbf{y}$, writing $(F)_{ab}=y^i(X_i)_{ab}$ and substituting this into (3.2) gives expressions for $d^{(m)}$, $m>3$ in terms of the tensors (4.5). Defining $k_{i_1\cdots i_m}=\mathrm{sTr}(X_{i_1}\cdots X_{i_m})$ as in (2.3), the first few of these are

$$k_{i_1i_2i_3i_4}=2\,\delta_{(i_1i_2}\delta_{i_3i_4)}, \tag{4.6}$$

$$k_{i_1i_2i_3i_4i_5}=\tfrac{10}{3}\,\delta_{(i_1i_2}d_{i_3i_4i_5)}, \tag{4.7}$$

$$k_{i_1i_2i_3i_4i_5i_6}=\tfrac{4}{3}\,d_{(i_1i_2i_3}d_{i_4i_5i_6)}+2\,\delta_{(i_1i_2}\delta_{i_3i_4}\delta_{i_5i_6)}. \tag{4.8}$$

The first three nonprimitive terms in the Sudbery basis for the symmetric invariant tensors of A_2 can then be computed as

$$\begin{aligned} d^{(4)}_{i_1i_2i_3i_4} &= d_{(i_1i_2}{}^x d_{i_3i_4)x}=\tfrac{1}{3}\,\delta_{(i_1i_2}\delta_{i_3)i_4}, \\ d^{(5)}_{i_1i_2i_3i_4i_5} &= d_{(i_1i_2}{}^x d^x{}_{i_3}{}^y d_{i_4i_5)y}=\tfrac{1}{3}\,\delta_{(i_1i_2}d_{i_3i_4i_5)}, \end{aligned} \tag{4.9}$$

$$d^{(6)}_{i_1i_2i_3i_4i_5i_6}=d_{(i_1i_2}{}^x d^x{}_{i_3}{}^y d^y{}_{i_4}{}^z d_{i_5i_6)z}=\tfrac{2}{15}\,d_{(i_1i_2i_3}d_{i_4i_5i_6)}+\tfrac{1}{15}\,\delta_{(i_1i_2}\delta_{i_3i_4}\delta_{i_5i_6)}. \tag{4.10}$$

C. The algebras B_n

The fundamental representation of the algebra $B_n=\mathrm{so}(2n+1)$ is in terms of $(2n+1)\times(2n+1)$ antisymmetric matrices. We define a basis $\{X_i\}$, normalized as $\mathrm{Tr}(X_iX_j)=2\,\delta_{ij}$. There will be n primitive symmetric invariant tensors (equal to the rank of the group). The first two of these are $\delta_{ij}=\frac{1}{2}\mathrm{Tr}(X_iX_j)$ and the order four totally symmetric tensor[13] given by

$$\mathrm{Tr}(X_{(i}X_jX_{k)})=d^{(4)}_{ijkl}X_l. \tag{4.11}$$

We notice that the tensors $k^{(m)}$ [given by (2.3)] vanish for m odd and we have a basis for the symmetric tensors of B_n in analogy with Sudbery's basis for A_n, i.e.,

$$d^{(6)}_{i_1i_2i_3i_4i_5i_6} = d_{(i_1i_2i_3}{}^x d_{i_4i_5i_6)x},$$

$$d^{(8)}_{i_1i_2i_3i_4i_5i_6i_7i_8} = d_{(i_1i_2i_3}{}^x d^x{}_{i_4i_5}{}^y d^y{}_{i_6i_7i_8)}.$$

$$\vdots \tag{4.12}$$

The algebra B_n will have n primitive symmetric tensors of orders $2,4,\ldots,2n$, giving rise to cocycles of orders $3,7,\ldots,(4n-1)$ via (4.2).

D. The algebra B_2

The fundamental representation of the algebra $B_2=\mathrm{so}(5)$ is in terms of (5×5) antisymmetric matrices with a basis $\{X_i\}$, normalized as above. As far as the cohomology is concerned, the group manifold decomposes as $\mathrm{SO}(5)\sim S^3\times S^7$, giving nontrivial three- and seven-cocycles. The three-cocycle is associated with the symmetric tensor $\delta_{ij}=\frac{1}{2}\mathrm{Tr}(X_iX_j)$ and the seven-cocycle with the order four symmetric tensor $d^{(4)}_{ijkl}=\frac{1}{2}\mathrm{Tr}(X_{(i}X_jX_{k)}X_l)$ according to (4.2). As above, we define an arbitrary vector $\mathbf{y}$, write $(F)_{ab}=y^i(X_i)_{ab}$ and substitute this into (3.2). The antisymmetry of the algebra elements means that all traces of odd powers of F vanish identically, giving the result at $\mathcal{O}(\alpha^6)$

$$\mathrm{Tr}\ F^6=\tfrac{3}{4}\mathrm{Tr}\ F^2\mathrm{Tr}\ F^4-\tfrac{1}{8}(\mathrm{Tr}\ F^2)^3, \tag{4.13}$$

or equivalently

$$k^{(6)}_{i_1i_2i_3i_4i_5i_6}=\tfrac{3}{2}\,\delta_{(i_1i_2}k^{(4)}_{i_3i_4i_5i_6)}-\delta_{(i_1i_2}\delta_{i_3i_4}\delta_{i_5i_6)}. \tag{4.14}$$

Expressions for the higher-order symmetric tensors can be obtained easily.

E. The algebra G_2

The algebra G_2 can be embedded into so(7); the generators of $SO(7)$ decompose into representations $\mathbf{21}\rightarrow\mathbf{7}\oplus\mathbf{14}$ under the action of G_2. We will work with the seven-dimensional representation of G_2 embedded into the fundamental representation of so(7). This has generators $\{X_i\}$, normalized as $\mathrm{Tr}(X_iX_j)=2\,\delta_{ij}$. As for the orthogonal groups, the terms odd in F in the expansion (3.2) will vanish identically. It is known[13] that there are two primitive symmetric invariant tensors of G_2, of order two and six, given by the tensors $k^{(2)}$ and $k^{(6)}$ defined in (2.3). We note the result from[14]

$$\mathrm{sTr}(X_iX_jX_kX_l)=\tfrac{1}{4}\,\mathrm{Tr}(X_{(i}X_j)\mathrm{Tr}(X_kX_{l)}). \tag{4.15}$$

Using this and the methods described above, we can now write any symmetric invariant tensor of G_2 in terms of the two primitive symmetric tensors

$$k_{i_1i_2}=\mathrm{Tr}(X_{i_1}X_{i_2}),\quad k_{i_1i_2i_3i_4i_5i_6}=\mathrm{sTr}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}X_{i_5}X_{i_6}). \tag{4.16}$$

V. CONCLUSION

A simple Lie group has a number of primitive symmetric invariant tensors equal to the rank of the group. Other symmetric invariant tensors must be expressible in terms of these. This is equivalent to expressing nonprimitive Casimirs in terms of primitive Casimirs. We have presented a systematic method for doing this which has been illustrated with three examples including one example of an exceptional Lie algebra.

ACKNOWLEDGMENTS

I thank Alan Macfarlane, Jonathan Evans, and Niall Mackay for interesting and helpful discussions. I acknowledge the financial assistance of a PPARC studentship.

[1] B. Gruber and L. O'Raifeartaigh, ''*s*-theorem and construction of the invariants of the semisimple compact Lie algebras,'' J. Math. Phys. **5**, 1796–1804 (1964).
[2] G. Racah, *Sulla caratterizzazione della rappresentazioni irreducibili dei gruppi semisimplici di Lie,*'' Lincei-Rend. Sc. fis. mat. e nat. **VIII**, 108–112 (1950).
[3] I. M. Gel'fand, ''The center of an infinitesimal group ring'' (english transl.: Los Alamos Sci. Lab. AEC-TR-6133 (1963)), Math. Sbornik. **26**, 103–112, Princeton Lectures, CERN-61-8 (Springer-Verlag, Berlin, 1950); reprinted in Ergeb. Exact Naturwiss. **37**, 28–84, 1965).
[4] L. C. Biedenharn, ''On the representations of the semisimple Lie groups I,'' J. Math. Phys. **4**, 436–445 (1963).
[5] A. Klein, ''Invariant operators of the unimodular group in *n* dimensions,'' J. Math. Phys. **4**, 1283–1284 (1963).
[6] A. M. Perelomov and V. S. Popov, ''Casimir operators for semisimple groups,'' Math. USSR-Izvestija **2**, 1313–1335 (1968).
[7] S. Okubo and J. Patera, ''General indices of representations and Casimir invariants,'' J. Math. Phys. **25**, 219–227 (1983).
[8] S. Okubo, ''Modified fourth-order Casimir invariants and indices for simple Lie algebras,'' J. Math. Phys. **23**, 8–20 (1982).
[9] H. B. G. Casimir, Proc. R. Acad. Sci. Amsterdam **34**, 844 (1931).
[10] E. Betti, Ann. Mat. Pura Appl. **4**, 140–158 (1871).
[11] J. A. de Azcárraga, A. J. Macfarlane, A. J. Mountain, and J. C. Pérez Bueno, ''Invariant tensors for simple groups,'' Nucl. Phys. B **510**, 657–687 (1998).
[12] A. Sudbery, Ph.D. thesis, Cambridge Univ., 1970; ''Computer-friendly *d*-tensor identities for SU(n),'' J. Phys. A **23**, L705–L710 (1990).
[13] J. A. de Azcárraga and J. C. Pérez Bueno, ''Higher order simple Lie algebras,'' Commun. Math. Phys. **184**, 669–681 (1997).
[14] S. Okubo, ''Casimir invariants and vector operators in simple and classical Lie algebras,'' J. Math. Phys. **18**, 2382–1294 (1977).

Twisting cocycles in fundamental representation and triangular bi-crossproduct Hopf algebras

A. I. Mudrov[a)]
Department of Theoretical Physics, Institute of Physics, St. Petersburg State University, Ulyanovskaya 1, Stary Petergof, St. Petersburg 198904, Russia

(Received 6 April 1998; accepted for publication 18 May 1998)

We find the general solution to the twisting equation in the tensor bialgebra $T(\mathbf{R})$ of an associative unital ring $\mathbf{R}$ viewed as that of fundamental representation for a universal enveloping Lie algebra and its quantum deformations. We suggest a procedure of constructing twisting cocycles belonging to a given quasitriangular sub-bialgebra $\mathscr{H}\subset T(\mathbf{R})$. This algorithm generalizes Reshetikhin's approach, which involves cocycles fulfilling the Yang–Baxter equation. Within this framework we study a class of quantized inhomogeneous Lie algebras related to associative rings in a certain way, for which we build twisting cocycles and universal R-matrices. Our approach is a generalization of the methods developed for the case of commutative rings in our recent work including such well-known examples as Jordanian quantization of the Borel subalgebra of sl(2) and the null-plane quantized Poincaré algebra by Ballesteros *et al.* We reveal the role of special group 1-cocycles in this process and establish the bi-crossproduct structure of the examples studied. © *1998 American Institute of Physics.* [S0022-2488(98)01410-8]

I. INTRODUCTION

Quantum deformations of Lie groups and algebras are at present a subject of intensive studies from the viewpoints of collecting facts and crystallizing mathematical concepts as well as of searching for new physical applications. Among the established notions of the quantum group theory one should mention Drinfeld's twisting[1,2] and Majid's bi-crossproduct and double crossproduct constructions.[3] Twisting, realizing a specific equivalence between two Hopf algebras, plays an important role for the geometrical and physical reasons because it controls deformation not only of the symmetry algebra of a manifold but of its whole geometry coherently. Therefore a classification of quantum deformations of, say, a universal enveloping Lie algebra ought to provide the answer about twist-equivalence between its different types. Majid's double crossproduct construction has a close connection with twisting and the quantum double in particular.[4–6] As for the bi-crossproduct, its relation to quasitriangularity and twisting is not so well understood, despite numerous examples including quasitriangular Hopf algebras. The most significant step in that direction was made in Ref. 7, where the double of the algebra $\mathbb{C}(\mathbf{M})\bowtie\mathbb{C}\mathbf{G}$, built on a matched pair of groups $\mathbf{M}$ and $\mathbf{G}$, was shown to be a bi-crossproduct itself. As examples of bi-crossproducts we would like to mention the κ deformation of the Poincaré algebra,[8,9] the canonical example of the Jordanian quantization of the Borel subalgebra of sl(2),[3] and the null-plane quantized Poincaré algebra.[10,11] The last two are the results of twisting of classical universal enveloping algebras and are therefore triangular. As we have shown in Ref. 12, they are associated with commutative rings (throughout the paper by ring we mean a finite-dimensional algebra over a field $\mathbf{K}$ so as to reserve the word ''algebra'' for Hopf algebras), one dimensional in the first case and that spanned by matrices

$$\alpha_1=\begin{pmatrix}0&0&1\\0&0&0\\0&0&0\end{pmatrix},\quad \alpha_2=\begin{pmatrix}0&0&0\\0&0&1\\0&0&0\end{pmatrix},\quad \alpha_3=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$$

in the second. Along this line we have found a generalization of the examples mentioned for an arbitrary commutative ring. The algebras studied in Ref. 12 were twisted classical universal

a)Electronic mail: aimudrov@DG2062.spb.edu

 © 1998 American Institute of Physics

enveloping algebras and also bi-crossproduct Hopf algebras, although not considered in that context. In the present paper we formulate a generalization of our approach to an arbitrary associative ring providing a new class of quasitriangular bi-crossproduct Hopf algebras. The Hopf operations are explicitly written out in terms of generators and the twisting 2-cocycles and universal R-matrices are presented. The class under investigation arises as an example of twisting which may be regarded as a generalized Reshetikhin's procedure[13–15] involving a solution to the Yang–Baxter equation as a twisting cocycle. We come to this generalization analyzing solutions to the twisting equation in the tensor bialgebra $T(\mathbf{R})$ of an associative unital ring $\mathbf{R}$, considered as that of the fundamental representation of a given Hopf algebra $\mathscr{H}$ (we call a homomorphism $\mathscr{H} \to \mathbf{R}$ fundamental if its lifting to whole $T(\mathbf{R})$, which always exists, is nondegenerate).

The paper is organized as follows. Section II is auxiliary and contains a general description of the tensor bialgebra $T(\mathbf{R})$ structure, its sub-bialgebras and homomorphisms. This part may be considered as the ''noncoordinate'' formulation of the Faddeev–Reshetikhin–Takhtajan (FRT) method[16] suitable for an arbitrary associative unital ring. Since rings with identities have exact matrix realizations, e.g., by the regular representations on themselves, such a reformulation does not supply us with particularly new information compared with the traditional matrix approach. Nevertheless, it provides certain technical convenience, so we find it possible to present this formulation here. Section III is devoted to solving the twisting equation in $T(\mathbf{R})$. Therein we develop a procedure of constructing ''universal'' cocycles starting from elements of $\mathbf{R}^{\otimes 2}$ obeying certain conditions. This algorithm is illustrated in Sec. IV on inhomogeneous Lie algebras related in a sense to associative rings. We build the deformed coproduct, quantum commutation relations, find twisting cocycles and universal R-matrices. In Sec. V the connection between the investigated algebras and the bi-crossproduct construction is established.

II. BIALGEBRA $T(\mathbf{R})$

To perform algebraic manipulations it is convenient to formulate the algorithm by Faddeev, Reshetikhin, and Takhtajan[16] in the part of constructing quantum algebra $\mathrm{Fun}_q(\mathbf{R})$ of functions on matrix rings for an arbitrary associative ring $\mathbf{R}$. We need some information concerning its structure and the structure of homomorphisms from a Hopf algebra or, more generally, bialgebra $\mathscr{H}$ into $\mathrm{Fun}_q(\mathbf{R})$. Let μ be the multiplication in $\mathbf{R}$. We choose a basis $(x^\alpha) \subset \mathbf{R}$ such that x^0 is the identity of $\mathbf{R}$. The dual basis in $\mathbf{R}^*$ will be marked with subscripts. Denote $F(\mathbf{R})$ the algebra over a field $\mathbf{K}$ freely generated by 1 and $(x_\alpha) \in \mathbf{R}^*$. Introduce the coproduct and the counit defining them on the generators as

$$\Delta(x_\alpha) = \mu_\alpha^{\rho\sigma} x_\rho \otimes x_\sigma, \quad \Delta(1) = 1 \otimes 1,$$

$$\epsilon(1) = 1, \quad \epsilon(x_0) = 1, \quad \epsilon(x_i) = 0, \quad i \neq 0$$

and extending over whole $F(\mathbf{R})$ homomorphically. The dual bialgebra $T(\mathbf{R}) = F^*(\mathbf{R})$ appears to be a direct sum of its ideals $\Sigma_{n=0}^{\infty} \mathbf{R}^{\otimes n}$, where $\mathbf{R}^0$ coincides the field $\mathbf{K}$ of scalars. The multiplication in $T(\mathbf{R})$ is characterized by the property $\mathbf{R}^{\otimes n} \mathbf{R}^{\otimes m} = 0$ for $m \neq n$. The identity of $T(\mathbf{R})$ is expanded as the sum $\Sigma_{n=0}^{\infty} e^n$ of idempotents, where e^0 is the unity of $\mathbf{K}$, and e^n with $n > 0$ are those of $\mathbf{R}^{\otimes n}$. Multiplying by e^n carries out the projection homomorphism

$$\pi^m : T(\mathbf{R}) \to \mathbf{R}^{\otimes m},$$

and for $n = 0$ this is just the bialgebra counit. The coproduct in $T(\mathbf{R})$ is determined by the product of its dual algebra $F(\mathbf{R})$ and on the basis elements is defined by the formula

$$\Delta(x^{i_1 \cdots i_n}) = e^0 \otimes x^{i_1 \cdots i_n} + \cdots + x^{i_1 \cdots i_k} \otimes x^{i_{k+1} \cdots i_n} + \cdots + x^{i_1 \cdots i_n} \otimes e^0, \quad x^{i_1 \cdots i_n} \in \mathbf{R}^{\otimes n}.$$

It follows from here that the composite mapping

$$\mathbf{R}^{\otimes (i+j)} \to T(\mathbf{R}) \xrightarrow{\Delta} T(\mathbf{R}) \otimes T(\mathbf{R}) \to \mathbf{R}^{\otimes i} \otimes \mathbf{R}^{\otimes j},$$

where the left arrow means the injection and the right one is the projection homomorphism, turns out to be a ring isomorphism. Quotient of $F(\mathbf{R})$ by the ideal J generated by quadratic relations of

the form $x_\alpha x_\beta = B^{\rho\sigma}_{\beta\alpha} x_\rho x_\sigma$ inherits the coproduct if and only if the subspace $\mathbf{S} \in \mathbf{R}^{\otimes 2}$ of functionals annihilating these relations is a subalgebra in $\mathbf{R}^{\otimes 2}$. In particular, such a subalgebra can be determined as the set of solutions to the equation

$$RzR^{-1} = \tau z, \quad z \in \mathbf{R}^{\otimes 2},$$

(τ permutes the factors in $\mathbf{R}^{\otimes 2}$), and then the relations in the dual algebra will look as

$$R_{\gamma\nu} \mu^{\gamma\rho}_\alpha \mu^{\nu\sigma}_\beta x_\rho x_\sigma = x_\sigma x_\rho \mu^{\rho\gamma}_\alpha \mu^{\sigma\nu}_\beta R_{\gamma\nu}.$$

The bialgebra $\mathscr{U}$ dual to the factor-bialgebra $\mathscr{A} \equiv F(\mathbf{R})/J$ is decomposed into the direct sum of its ideals $\Sigma_{n=0}^{\infty} \pi^n(\mathscr{U})$, where $\pi^1(\mathscr{U})$ is isomorphic to the ring $\mathbf{R}$ itself and each addend $\pi^n(\mathscr{U})$ at $n>1$ is a subring in $\mathbf{R}^{\otimes n}$ characterized by the condition $z_{\dots\alpha\beta\dots} = B^{\rho\sigma}_{\beta\alpha} z_{\dots\rho\sigma\dots}$ for $z \in \pi^n(\mathscr{U})$. Evidently, $\pi^n(\mathscr{U})$ is the intersection of all possible subalgebras $\mathbf{R}^{\otimes i} \otimes \pi^2(\mathscr{U}) \otimes \mathbf{R}^{\otimes j}$, such that $i + j + 2 = n$.

Let us describe the structure of a homomorphism ϕ of an arbitrary bialgebra $\mathscr{H}$ to $\mathscr{U}$. The composition of ϕ with the projector π^n is an algebraic mapping. We set $\rho \equiv \pi^1 \circ \phi$ and $\rho^n \equiv \pi^n \circ \phi$ and also introduce the notations

$$\Delta^1 \equiv id : \mathscr{H} \to \mathscr{H}, \quad \Delta^2 \equiv \Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}, \quad \Delta^3 \equiv (\Delta \otimes id)\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H} \otimes \mathscr{H}, \dots .$$

Let $h \in \mathscr{H}$ and $x_{\vec{\alpha}} \in \mathscr{A}$, where $\vec{\alpha}$ is a multi-index of length n. From the chain of equalities

$$\begin{aligned}
\langle \rho^n(h), x_{\vec{\alpha}} \rangle &= \langle \pi^n \circ \phi(h), x_{\vec{\alpha}} \rangle \\
&= \langle \phi(h), x_{\vec{\alpha}} \rangle \\
&= \langle \Delta^n \circ \phi(h), x_{\alpha_1} \otimes \cdots \otimes x_{\alpha_n} \rangle \\
&= \langle \phi^{\otimes n} \circ \Delta^n(h), x_{\alpha_1} \otimes \cdots \otimes x_{\alpha_n} \rangle \\
&= \langle (\pi^1 \circ \phi)^{\otimes n} \circ \Delta^n(h), x_{\alpha_1} \otimes \cdots \otimes x_{\alpha_n} \rangle = \langle \rho^{\otimes n} \circ \Delta^n(h), x_{\alpha_1} \otimes \cdots \otimes x_{\alpha_n} \rangle
\end{aligned}$$

we find that conditions

(1) $\rho^n = \rho^{\otimes n} \circ \Delta^n$

and

(2) $\rho^2(\mathscr{H}) \subset \mathbf{S} = \pi^2(\mathscr{U})$

are satisfied, and these completely specify ϕ. Namely, if an algebraic mapping $\rho : \mathscr{H} \to \mathbf{R}$ fulfills condition 2, there exists the unique bialgebra homomorphism of $\phi : \mathscr{H} \to \mathscr{U}$ such that $\pi^n \circ \phi = \rho^n$. Indeed, mapping $\rho^{\otimes n} \circ \Delta^n$ respects the algebraic structure for every n. Obviously, $\rho^{\otimes(n+2)} \circ \Delta^{n+2}(\mathscr{H}) \subset \mathbf{R}^{\otimes i} \otimes \mathbf{S} \otimes \mathbf{R}^{\otimes(n-i)}$ for every $i \leqslant n$, because of the coassociativity of the coproduct. We define ϕ via the formula $\phi(h) = e^0 \varepsilon(h) + \Sigma_{n>0} \rho^{\otimes n} \circ \Delta^n(h)$ and verify that it is also a coalgebra homomorphism (its algebraic property is evident). For two multi-indices $\vec{\alpha}$ and $\vec{\beta}$ of lengths m and n we have

$$\begin{aligned}
\langle (\phi \otimes \phi)\Delta(h), x_{\vec{\alpha}} \otimes x_{\vec{\beta}} \rangle &= \langle (\rho \otimes \cdots \otimes \rho)\Delta^{m+n}(h), x_{\alpha_1} \otimes \cdots \otimes x_{\alpha_m} \otimes x_{\beta_1} \otimes \cdots \otimes x_{\beta_n} \rangle \\
&= \langle \phi(h), x_{\vec{\alpha}} x_{\vec{\beta}} \rangle = \langle \Delta \circ \phi(h), x_{\vec{\alpha}} \otimes x_{\vec{\beta}} \rangle
\end{aligned}$$

that proves what is required. Conditions 1 and 2 are convenient practical criteria for checking out the homomorphic properties of mappings from $\mathscr{H}$ to $\mathscr{U}$.

III. TWISTING EQUATION IN $T(\mathbf{R})$

Algebras we are interested in, namely, universal enveloping Lie algebras and their quantum deformations appear to be embedded into $T(\mathbf{R})$ associated with their fundamental representation ring $\mathbf{R}$. Since twisting of a subalgebra induces that of the whole algebra,[15] the problem can be put

forward of describing all solutions to the twisting equations in $T(\mathbf{R})$ with the hope of further selecting among them those belonging to required subalgebras. We can point out the following advantages of such an approach. As far as the composition of two twistings is concerned, we avoid the inconvenience of dealing with a bialgebra different from the original after performing the first deformation. Another remarkable feature of such a description is a possibility to reduce the elaborate task of constructing the universal twistor to the much easier problem of solving a system of equations, rather nonlinear, in a finite-dimensional ring. Our study resulted in finding the general solution to the twisting equation in $T(\mathbf{R})$. We have managed to formulate conditions more general than those employed in Reshetikhin's approach[13,14] which ensure that the twisting cocycle built on its image in $\mathbf{R}^{\otimes 2}$ would lie in the required subalgebra.

Consider the twisting equation

$$(\Delta\otimes id)(\Phi)\Phi_{12}=(id\otimes\Delta)(\Phi)\Phi_{23}, \tag{1}$$

in the tensor cube of a bialgebra $\mathscr{H}$ where the subscripts determine the embeddings $\mathscr{H}^{\otimes 2}\to\mathscr{H}^{\otimes 3}$. An invertible solution to this equation can be normalized in such a way that

$$(\epsilon\otimes id)(\Phi)=(id\otimes\epsilon)(\Phi)=1. \tag{2}$$

Such a solution takes part in transforming bialgebra $\mathscr{H}$ into a new one with the same multiplication and the coproduct $\widetilde{\Delta}(h)=\Phi^{-1}\Delta(h)\Phi$, $h\in\mathscr{H}$. Other objects, e.g., the counit, antipode, universal $\mathscr{R}$-matrix, if any, are connected with the old ones via the well-known formulas which can be found in Ref. 2. We are going to prove the following assertion.

Theorem 1: For every set of invertible elements $\Phi^{1,k}\in\mathbf{R}\otimes\mathbf{R}^k$, there exists the unique solution $\Phi\in T(\mathbf{R})\otimes T(\mathbf{R})$ to the twisting equation, such that $(\pi^1\otimes\pi^k)(\Phi)=\Phi^{1,k}$.

Having applied the projector $\pi^m\otimes\pi^n\otimes\pi^k$ to the both sides of equality (1), we come to the equation in $\mathbf{R}^{\otimes(m+n+k)}$:

$$\Phi^{m+n,k}\Phi_b^{m,n}=\Phi^{m,n+k}\Phi_e^{n,k}, \tag{3}$$

where $\Phi^{m,n}$ is the image of the twisting cocycle $(\pi^m\otimes\pi^n)(\Phi)$. Letters b and e indicate that the elements $\Phi_b^{m,n}$ and $\Phi_e^{n,k}$ are embedded into $\mathbf{R}^{\otimes(m+n+k)}$ from the ''beginning'' and from the ''end,'' respectively: $\Phi_b^{m,n}\in\mathbf{R}^{\otimes(m+n)}\otimes e^k\subset\mathbf{R}^{\otimes(m+n+k)}$, $\Phi_e^{n,k}\in e^m\otimes\mathbf{R}^{\otimes(n+k)}\subset\mathbf{R}^{\otimes(m+n+k)}$. Because of (2), for every $k\geqslant 0$ we have $\Phi^{0,k}=e^0\otimes e^k$ and $\Phi^{k,0}=e^k\otimes e^0$. Suppose now that the elements $\Phi^{1,k}$ are known. Then, using Eq. (3), $\Phi^{m,k}$ can be defined recursively for all m and k greater than 1:

$$\Phi^{m+1,k}\equiv\Phi^{m,1+k}\Phi_e^{1,k}\overline{\Phi}_b^{m,1} \tag{4}$$

(the bar stands for the inverse). This implies the uniqueness of the solution. Obviously, Eq. (3) is true when one of the numbers m, n, and k are equal to zero. By construction, it is fulfilled for $n=1$ and all m as well. So we must show that Eq. (3) is satisfied for arbitrary m, n, and k. Assume the required property proved for all m and n which sum is less than $N>2$. Then for $m+n=N$ and $k>1$ we have

$$\Phi^{m+n,k}=\Phi^{m+n-1,1+k}\Phi_e^{1,k}\overline{\Phi}_b^{m+n-1,1}.$$

Within the assumption made, we decompose the first factor on the right-hand side of this equality, according to (4), and rewrite (3) in the equivalent form

$$\Phi^{m,n+k}\Phi_c^{n-1,1+k}\overline{\Phi}_b^{m,n-1}\Phi_e^{1,k}\overline{\Phi}_b^{m+n-1,1}=\Phi^{m,n+k}\Phi_e^{n,k}\overline{\Phi}_b^{m,n}.$$

Dividing both sides by the first factor we come to condition

$$\Phi_e^{n-1,1+k}\overline{\Phi}_b^{m,n-1}\Phi_e^{1,k}\overline{\Phi}_b^{m+n-1,1}=\Phi_e^{n,k}\overline{\Phi}_b^{m,n}.$$

Again, decomposing the first factor on the right-hand side according to the recursion assumption we find

$$\Phi_e^{n-1,1+k}\overline{\Phi}_b^{m,n-1}\Phi_e^{1,k}\overline{\Phi}_b^{m+n-1,1}=\Phi_e^{n-1,1+k}\Phi_e^{1,k}\overline{\Phi}_{\{m+1\}}^{n-1,1}\overline{\Phi}_b^{m,n},$$

where the subscript in $\overline{\Phi}^{n-1,1}_{\{m+1\}}$ means that it is embedded into $\mathbf{R}^{\otimes(m+n+k)}$ beginning from $\{m+1\}$th place. It is important that the factor $\overline{\Phi}^{m,n-1}_b$ on the left-hand side can be permuted with $\Phi^{1,k}_e$. Division by $\Phi^{n-1,1+k}_e\Phi^{1,k}_e$ yields

$$\overline{\Phi}^{m,n-1}_b\overline{\Phi}^{m+n-1,1}_b=\overline{\Phi}^{n-1,1}_{\{m+1\}}\overline{\Phi}^{m,n}_b$$

or

$$\Phi^{m+n-1,1}_b\Phi^{m,n-1}_b=\Phi^{m,n}_b\Phi^{n-1,1}_{\{m+1\}}.$$

This is exactly twisting equation (3) for $k\to 1$ and $n\to n-1$. According to the induction principle we consider the theorem proved.

Thus the family of solutions to Eq. (1) in the bialgebra $T(\mathbf{R})$ turns out to be very large: it is parametrized by an arbitrary set of invertible elements $\Phi^{1,k}$, $k>0$. On the other hand, we are interested only in those Φ which belong to the subalgebra $\mathscr{U}\otimes\mathscr{U}$. We cannot propose the general method to build such solutions. However, we can point out an algorithm which can help to solve the problem at least for quasitriangular $\mathscr{U}$ and which can be interpreted as a generalization of Reshetikhin's approach. Starting form $\Phi^{1,1}$ as known, set $\Phi^{1,k}\equiv\Phi_{12}\Phi_{13}\cdots\Phi_{1(k+1)}$. Here Φ_{1i} are the images of $\Phi^{1,1}$ via the corresponding embeddings $\mathbf{R}^{\otimes 2}\to\mathbf{R}\otimes\mathbf{R}^{\otimes k}$. For $\Phi^{1,k}$ to belong to $\mathscr{U}\otimes\mathscr{U}$, it is necessary and sufficient to require

$$R_{23}\Phi_{12}\Phi_{13}=\Phi_{13}\Phi_{12}R_{23}. \tag{5}$$

Further, element $\Phi^{2,1}$ lies in $\mathscr{U}\otimes\mathscr{U}$ if and only if $R_{12}\Phi^{2,1}=(\tau\otimes id)(\Phi^{2,1})R_{12}$. Having expressed $\Phi^{2,1}$ through $\Phi^{1,1}$ we come to equation $R_{12}\Phi_{12}\Phi_{13}\Phi_{23}\overline{\Phi}_{12}=\Phi_{21}\Phi_{23}\Phi_{13}\Phi_{21}R_{12}$ or, the matrix $\widetilde{R}=\overline{\Phi}_{21}R\Phi$ introduced, to equation

$$\widetilde{R}_{12}\Phi_{13}\Phi_{23}=\Phi_{23}\Phi_{13}\widetilde{R}_{12}. \tag{6}$$

Verification of the condition $R_{12}\Phi^{2,k}=(\tau\otimes id)(\Phi^{2,k})R_{12}$ for $k>1$ boils down to equality $\widetilde{R}_{12}\Phi_{13}\cdots\Phi_{1k+2}\Phi_{23}\cdots\Phi_{2k+2}=\Phi_{23}\cdots\Phi_{2k+2}\Phi_{13}\cdots\Phi_{1k+2}\widetilde{R}_{12}$, which, as can be easily seen, follows from (6). Now, with further use of identity (3) one can see that $\Phi^{m,n}$ belongs to $\mathscr{U}\otimes\mathscr{U}$ for every m and n, indeed.

IV. TWISTING OF INHOMOGENEOUS LIE ALGEBRAS

One can notice that the example of solution to the twisting equation built at the end of the previous section satisfies the identities

$$(id\otimes\Delta)(\Phi)=\Phi_{12}\Phi_{13}, \tag{7}$$

$$(\widetilde{\Delta}\otimes id)(\Phi)=\Phi_{13}\Phi_{23}, \tag{8}$$

where $\widetilde{\Delta}$ is the twisted coproduct: $\widetilde{\Delta}(h)=\overline{\Phi}\Delta(h)\Phi$. Reshetikhin's conditions are obtained from here if Φ solves the Yang–Baxter equation and, besides, $(\Delta\otimes id)(\Phi)=\Phi_{23}\Phi_{13}$. This generalization of Reshetikhin's twisting is nontrivial, the nonstandard quantization of the Borel subalgebra of sl(2) taken into account. The twisting cocycle for $U(b(2))$ has the form $\exp(X\otimes H)$, where H is the primitive element in $U(b(2))$ and X is the primitive element in $U_h(b(2))$.[17] This example was generalized in Ref. 12 for an arbitrary commutative ring which in the case of $U(b(2))$ coincides with the field of scalars. In the present section we shall formulate the analogous generalization for an arbitrary associative ring $\mathbf{L}$, not necessarily commutative. Let the multiplication in $\mathbf{L}$ be defined by the structure constants $B^{\sigma}_{\mu\nu}$. Consider a Lie algebra built on $H_\mu\in\mathbf{L}$ and $X^\nu\in\mathbf{L}^*$ subjected to the commutation relations

$$[H_\mu,H_\nu]=(B^{\sigma}_{\mu\nu}-B^{\sigma}_{\nu\mu})H_\sigma,\quad [H_\mu,X^\nu]=-B^{\nu}_{\mu\sigma}X^\sigma.$$

The subalgebra $\mathbf{L}^*$ generated by X^ν is assumed to be commutative. It is easy to see that the element $H_\nu\otimes X^\nu-X^\nu\otimes H_\nu$ satisfies the classical Yang–Baxter equation. The subspace $\mathbf{L}^*$ is a

right module over ring **L**. Let us affiliate the identity to $\mathbf{L}\triangleright\mathbf{L}^*$ and denote the resulting ring **R**. The product in **R** is evaluated according to the rules

$$\hat{H}_\mu\hat{H}_\nu=B^\sigma_{\mu\nu}\hat{H}_\sigma,\quad \hat{X}^\nu\hat{H}_\mu=B^\nu_{\mu\sigma}\,\hat{X}^\sigma,\quad \hat{X}^\mu\hat{X}^\nu=0,\quad \hat{H}_\mu\hat{X}^\nu=0,$$

plus evident expressions involving identity $\hat{E}$. Starting from this multiplication, one can see that the element

$$\tilde{R}=\hat{E}\otimes\hat{E}+\hat{H}_\nu\otimes\hat{X}^\nu-\hat{X}^\nu\otimes\hat{H}_\nu \tag{9}$$

is a solution to the quantum Yang–Baxter equation, and the element

$$\Phi^{1,1}=\hat{E}\otimes\hat{E}-\hat{X}^\nu\otimes\hat{H}_\nu \tag{10}$$

obeys (5) and (6) where R should be set to $\hat{E}\otimes\hat{E}$. Correspondence $1\to\hat{E}$, $X\to\hat{X}$, $H\to\hat{H}$ is extended to a homomorphism (nondegenerate) of the universal enveloping algebra $U(\mathbf{R})$ into $T(\mathbf{R})$.

Theorem 2: Twisting cocycle Φ expanding $\Phi^{1,1}$ [Eq. (10)] by modified Reshetikhin's procedure belongs to $U(\mathbf{L}^*)\otimes U(\mathbf{L})$ and has the form $\exp(-\tilde{X}^\nu\otimes H_\nu)$, where $\tilde{X}^\nu$ are expressed by series in X^ν.

Indeed, $\Phi^{1,k}\in U(\mathbf{L}^*)\otimes U(\mathbf{L})$ by construction. From defining formula (4), recursively using the facts that $\mathrm{Span}(\hat{E},\hat{H}_\mu)$ and $\mathrm{Span}(\hat{X}^\mu)$ form a subring and an ideal in **R**, respectively, we get the first assertion of the theorem. Now the announced form of the twisting cocycle follows from (7).

Unemployed so far identity (8) enables us to determine the twisted coproduct on the generators $\tilde{X}^\nu$. It turns out that $\tilde{\Delta}(\tilde{X}^\nu)=D^\nu(1\otimes\tilde{X},\tilde{X}\otimes 1)$, where $D(a,b)$ is the Campbell–Hausdorff series: $e^{D(a,b)}=e^ae^b$ for arbitrary a and b from the Lie algebra of the ring **L**. Thus the commutative algebra $U_q(\mathbf{L}^*)$ is isomorphic to the function algebra on the group $\exp(\mathbf{L})$ taken with the opposite coproduct. We have yet to evaluate the twisted coproduct on H_ν and to determine commutation relations $[H_\nu,\tilde{X}^\mu]=f(\tilde{X})^\mu_\nu$. What can be said about functions $f(x)^\mu_\nu$ is that they are subject to the ''boundary'' conditions

$$f(0)^\mu_\nu=0,\quad \partial_\sigma f(0)^\mu_\nu=-B^\mu_{\nu\sigma}. \tag{11}$$

It is accounted for the following. First of all, knowing the image $\Phi^{1,1}$ of the cocycle Φ in $\mathbf{R}^{\otimes 2}$ we conclude that $\partial\tilde{X}^\mu/\partial X^\nu|_{x=0}=\delta^\mu_\nu$. Now the required properties of $f(x)^\mu_\nu$ are conditioned by the homomorphism from $U(\mathbf{R})$ to **R** and the strong nilpotence of $\hat{X}^\mu$.

Having introduced matrices of the left and right regular representations $L(X)^\mu_\nu\equiv B^\mu_{\sigma\nu}\,X^\sigma$, $R(X)^\mu_\nu\equiv B^\mu_{\nu\sigma}\,X^\sigma$, from the definition of $\tilde{\Delta}$ we find

$$\begin{aligned}\tilde{\Delta}(H_\mu)&=\exp(\tilde{X}\otimes H)(H_\mu\otimes 1+1\otimes H_\mu)\exp(-\tilde{X}\otimes H)\\&=H_\mu\otimes 1-\left(\frac{e^{L(\tilde{X})-R(\tilde{X})}-1}{L(\tilde{X})-R(\tilde{X})}\,f(\tilde{X})\right)^\nu_\mu\otimes H_\nu+(e^{L(\tilde{X})-R(\tilde{X})})^\nu_\mu\otimes H_\nu=H_\mu\otimes 1+g(\tilde{X})^\nu_\mu\otimes H_\nu.\end{aligned} \tag{12}$$

The coassociativity requirement imposed, formula (12) implies $g(D(1\otimes\tilde{X},\tilde{X}\otimes 1))=g(1\otimes\tilde{X})g(\tilde{X}\otimes 1)$, and this with necessity entails $g(\tilde{X})=e^{A(\tilde{X})}$, where A represents a left action of **L** on itself. Resolving $g(\tilde{X})$ with respect to $f(\tilde{X})$ and using conditions (11) we come finally to $A(\tilde{X})^\mu_\nu=L(\tilde{X})^\mu_\nu=B^\mu_{\sigma\nu}\,X^\sigma$. The resulting formulas describing the Hopf structure of the twisted algebra $U_q(\mathbf{R})$ read

$$\begin{gathered}\tilde{\Delta}(\tilde{X}^\mu)=D(1\otimes\tilde{X},\tilde{X}\otimes 1),\qquad \tilde{\Delta}(H_\mu)=H_\mu\otimes 1+(e^{L(\tilde{X})})^\nu_\mu\otimes H_\nu,\\ [H_\mu,H_\nu]=(B^\sigma_{\mu\nu}-B^\sigma_{\nu\mu})H_\sigma,\qquad [H_\nu,\tilde{X}^\mu]=\left(\frac{L(\tilde{X})-R(\tilde{X})}{e^{L(\tilde{X})-R(\tilde{X})}-1}\,(e^{L(\tilde{X})-R(\tilde{X})}-e^{L(\tilde{X})})\right)^\mu_\nu.\end{gathered} \tag{13}$$

The antipode is easily found from the coproduct:

$$S(\tilde{X}^\mu)=-\tilde{X}^\mu, \quad S(H_\mu)=-(e^{-L(\tilde{X})})^\nu_\mu H_\nu.$$

The expressions obtained generalize formulas deduced for commutative rings in Ref. 12. In the latter case the value of the commutator $[H_\nu,\tilde{X}^\mu]$ is simplified because of $L(\tilde{X})=R(\tilde{X})$ and turns into $(1-e^{L(\tilde{X})})$, as for the well-known example of the Borel subalgebra of sl(2). The universal R-matrix is expressed through the twisting cocycle by the standard formula $\mathscr{R}=\Phi_{21}^{-1}\Phi:^1$

$$\mathscr{R}=\exp(H_\nu\otimes\tilde{X}^\nu)\exp(-\tilde{X}^\nu\otimes H_\nu) \tag{14}$$

and has the form familiar from the theories of the Jordanian quantization of sl(2)[17] and the null-plane quantized Poincaré algebra.[12,18] Thus we obtain a closed and complete description of the deformed algebra $U_q(\mathbf{R})$, although it would be desirable to find the relation between $\tilde{X}^\mu$ and the classical generators X^μ. To this end, let us calculate the twisted coproduct on elements X^μ:

$$\tilde{\Delta}(X^\mu)=\exp(\tilde{X}\otimes H)(1\otimes X^\mu+X^\mu\otimes 1)\exp(-\tilde{X}\otimes H)=X^\mu\otimes 1+(e^{-L(\tilde{X})})^\mu_\nu\otimes X^\nu,$$

that results in the functional equation

$$\varphi(\Delta(\tilde{X}))=\varphi(\tilde{X}')+e^{-L(\tilde{X}')}\varphi(\tilde{X}''), \tag{15}$$

where the primes distinguish the tensor components, and φ is the transformation connecting the quantum and classical generators: $X^\mu=\varphi^\mu(\tilde{X})$. This equation is well known from the theory of the group cohomologies[19] and its solution is

$$\varphi(\tilde{X})=\frac{e^{-L(\tilde{X})}-1}{-L(\tilde{X})}\,\tilde{X}.$$

This formula solves the problem of proceeding to the classical basis of $U_q(\mathbf{R})$.

The analysis of the solution found allows us to perform a further generalization of the examples considered above in the following direction. Let $\mathbf{G}$ be a Lie group and $\mathbf{L}$ its Lie algebra with the basis elements H_μ. Assume a left action $H_\mu\triangleright H_\nu=B^\sigma_{\mu\nu}H_\sigma$ of $\mathbf{L}$ on itself, which, as a rule, does not coincide with the adjoint representation. Let function φ: $\mathbf{G}\to\mathbf{L}$ be a group 1-cocycle, that is $\varphi(ba)=\varphi(a)+a^{-1}\triangleright\varphi(b)$, $a,b\in\mathbf{G}$. It can be viewed as a mapping defined in some neighborhood of the origin in $\mathbf{L}$. We suppose φ to be invertible and denote its inverse ψ. By the left conjugate action $H_\mu\triangleright X^\nu=-B^\nu_{\mu\sigma}X^\sigma$ on the dual space we build the semidirect sum $\mathbf{L}\triangleright\mathbf{L}^*$, where $\mathbf{L}^*$ is considered as an Abelian subalgebra.

Theorem 3: Element $\Phi=\exp(-\psi^\nu(X)\otimes H_\nu)$ is a twisting cocycle for the universal enveloping Hopf algebra $U(\mathbf{L}\triangleright\mathbf{L}^*)$.

Notice that $\exp(-\psi^\nu(X)\otimes H_\nu)$ satisfies identity (7). Hence in order to prove the theorem the second identity (8) should be stated. Making use of the fact that $\tilde{\Delta}$ is an algebraic mapping for arbitrary Φ, we evaluate $\tilde{\Delta}(\tilde{X})$ on the elements $\tilde{X}^\mu=\psi^\mu(X)$ and come to the equation (15), where operator $L(\tilde{X})$ is defined via tensor B as above. Because of the invertibility of φ, this implies $\Delta(\tilde{X}^\mu)=D^\mu(1\otimes\tilde{X},\tilde{X}\otimes 1)$. Then identity (8) is obeyed as well.

V. BI-CROSSPRODUCT STRUCTURE

To exhibit the bi-crossproduct structure of the examples considered let us use the FRT method and recover the quantum groups with the solution to the quantum Yang–Baxter equation (9). Modulo the order of the factors, they are isomorphic to the quantum algebras, and the isomorphism is realized via the universal R-matrix (14).

In terms of the basis (e,x_μ,h^ν) dual to the basis $(\hat{E},\hat{H}_\mu,\hat{X}^\nu)$ of the ring $\mathbf{R}$ the coproduct, according to the scheme rendered in Sec. II, has the form

$$\Delta(h^\sigma)=h^\sigma\otimes e+e\otimes h^\sigma+B^\sigma_{\mu\nu}h^\mu\otimes h^\nu,\quad \Delta(x_\sigma)=x_\sigma\otimes e+e\otimes x_\sigma+B^\nu_{\mu\sigma}x_\nu\otimes h^\mu,$$

$$\Delta(e)=e\otimes e.$$

The counit is determined by the rule $\varepsilon(e)=1$, $\varepsilon(x_\mu)=0$, $\varepsilon(h^\mu)=0$. Imposing RTT-type relations with the matrix R given by (9), we come to the following permutation rules

$$[x_\mu,x_\nu]=(B^\sigma_{\nu\mu}-B^\sigma_{\mu\nu})x_\sigma e, \tag{16}$$

$$[x_\mu,h^\nu]=B^\nu_{\mu\sigma}h^\sigma e+B^\nu_{\sigma\alpha}B^\sigma_{\beta\mu}h^\alpha h^\beta. \tag{17}$$

Other commutation relations are trivial and, in particular, the element e belongs to the center of the algebra. Note that the ideal $(e-1)$ is a Hopf one and set $e=1$. Introduce quantities η^μ starting from equality $h^\mu\otimes\hat{H}_\mu=e^{\eta^\mu\otimes\hat{H}_\mu}-1\otimes\hat{E}$. In terms of new generators (η^μ,x_μ) the coproduct turns out to be opposite to that of the algebra $\tilde{U}(\mathbf{L})$, and that is seen through the substitution $\eta^\mu\to\tilde{X}^\mu$, $x_\mu\to-H_\mu$. Commutation relation (16) is thus recovered exactly. It has yet to be shown that relation (17) goes over into the last expression in (13). This is guaranteed by the uniqueness of the value of the commutator (17) as a function in h^μ compatible with the given coproduct and fulfilling $\partial[x_\mu,h^\nu]/\partial h^\sigma|_{h=0}=B^\nu_{\mu\sigma}$. This boundary condition is determined by the homomorphism from the quantum group into the ring $\mathbf{L}$ via the ''square'' matrix R (9).

According to Ref. 3, bi-crossproduct $\mathscr{A}\bowtie\mathscr{B}$ of Hopf algebras $\mathscr{A}$ and $\mathscr{B}$ is defined via a left action $\triangleright$ of $\mathscr{A}$ on $\mathscr{B}$ and a right coaction β of $\mathscr{B}$ on $\mathscr{A}$. The latter is a mapping from $\mathscr{A}$ into the tensor product $\mathscr{A}\otimes\mathscr{B}$. The conjugate mapping to β realizes a right action $\mathscr{A}^*\triangleleft\mathscr{B}^*$. These operations are subjected to the set of consistency conditions.[3] Multiplication and comultiplication on $\mathscr{A}\bowtie\mathscr{B}$ are evaluated via

$$(a\otimes h)(b\otimes g)=a(h_{(1)}\triangleright b)\otimes h_{(2)}g,\qquad \Delta(a\otimes h)=(a_{(1)}\otimes\bar{h}^{(1)}_{(1)})\otimes(a_{(2)}\bar{h}^{(2)}_{(1)}\otimes h_{(2)}),$$

where $a,b\in\mathscr{A}$, $h,g\in\mathscr{B}$, $\Delta(h)=h_{(1)}\otimes h_{(2)}$, and $\beta(h)=\bar{h}^{(1)}\otimes\bar{h}^{(2)}$. Turning to $U_q(\mathbf{R})$ constructed in the previous section we see that algebraically it has the same structure as $U^*(\mathbf{L})\triangleleft U(\mathbf{L})$, whereas its dual quantum group has the form $U(\mathbf{L})\triangleright U^*(\mathbf{L})$. In terms of generators h^μ and x_ν the left action is $x_\mu\triangleright h^\nu=B^\nu_{\mu\sigma}h^\sigma+B^\nu_{\sigma\alpha}B^\sigma_{\beta\mu}h^\alpha h^\beta$, and the right coaction β can be computed from the coproduct formula

$$1\otimes\beta(h)\otimes1=(id\otimes id\otimes id\otimes\varepsilon)\circ\Delta(1\otimes h).$$

This yields

$$\beta(x_\sigma)=x_\sigma\otimes1+B^\nu_{\mu\sigma}x_\nu\otimes h^\mu$$

for the generators x^μ thus stating the bi-crossproduct structure of $U^*_q(\mathbf{R})$ and $U_q(\mathbf{R})$.

VI. DISCUSSIONS

The problem of explicitly calculating twisting 2-cocycles for Hopf algebras is a nontrivial one even if their existence is *a priori* known. Difficulties arise already in the simplest case of classical universal enveloping Lie algebras, despite advanced Drinfeld's theory on quantizing triangular Lie bialgebras.[1] This explains why examples of explicitly given twisting 2-cocycles are in short supply. So, it seems quite natural to reduce the problem to studying ''quadratic'' 2-cocycles which are images of ''universal'' ones in fundamental representations, provided there exists some ''fusion'' procedure to expand those matrix solutions over representations of higher spins. There are two algorithms of this kind,[14,20–22] both based on factorization properties of twisting elements[5,6,13,23] $(\Delta\otimes id)(\Phi)=\Phi_{13}\Phi_{23}$ or $(\Delta\otimes id)(\Phi)=\Phi_{23}\Phi_{13}$ (and appropriate identities involving $id\otimes\Delta$). Depending on the order of the factors on the right-hand side, additional requirements like $\Phi_{12}\Phi_{23}=\Phi_{23}\Phi_{12}$ or the Yang–Baxter equation are imposed on Φ. Although most of explicitly known universal twisting 2-cocycles are due to these two options, it is clear that they cannot cover all possibilities. The idea of proceeding to fundamental representation in a finite-

dimensional ring **R** in studying twist-equivalences among various quantizations seems yet more fruitful because the associated bialgebra $T(\mathbf{R})$ plays the role of a container, in a general situation, for all the deformations of a Hopf algebra. Thus every twisting cocycle of a sub-bialgebra undergoing deformation remains so for whole $T(\mathbf{R})$. This makes it reasonable to consider twisting equation in $T(\mathbf{R})$ and then try to select solutions belonging to the given sub-bialgebra. The first part of this program has been completely solved in the present paper, while for the second we have suggested a new kind of fusion procedure which appears to be close to Reshetikhin's twisting. The novelty is that the conditions imposed on Φ employ both twisted and untwisted coproducts. To demonstrate effectiveness of the proposed scheme we have considered a class of inhomogeneous universal enveloping Lie algebras related to associative rings in a special way and quantized them along that line. The technique can be viewed as a generalization of the theory developed in our previous work[12] dealing exclusively with commutative rings, which was motivated by the Jordanian quantization of sl(2) and the null-plane quantized Poincaré algebra. We have also exposed the bi-crossproduct structure of the objects investigated thus providing new examples of quasitriangular bi-crossproduct Hopf algebras. A remarkable fact is that the class considered may be treated directly with the use of special Lie group 1-cocycles which take part in building twisting elements. Relevance of Hopf algebra cohomologies to the twisting procedure and bi-crossproduct construction was already pointed out in Ref. 3, so the present study gives a new insight to their role in the theory. It is interesting to generalize cohomological methods applied here for classical universal enveloping algebras to Hopf algebras of a more general nature.

[1] V. G. Drinfeld, ''On constant quasiclassical solutions to the quantum Yang-Baxter equation,'' Dokl. Akad. Nauh. USSR **273**, 531–535 (1983).
[2] V. G. Drinfeld, ''Quasi-Hopf algebras,'' Leningrad Math. J. **1**, 1419–1457 (1990).
[3] S. Majid, ''Crossproduct quantization, nonabelian cohomology and twisting of Hopf algebras,'' hep-th/9311184.
[4] V. G. Drinfeld, ''Quantum groups,'' in *Proceedings of the International Congress of Mathematicians, Berkeley, 1986*, edited by A. V. Gleason (American Mathematical Society, Providence, 1987), pp. 798–820.
[5] N. Yu. Reshetikhin and M. A. Semenov-Tian-Shansky, ''Quantum *R*-matrices and factorization problems,'' J. Geom. Phys. **5**, 533–550 (1988).
[6] S. Majid, ''Some remarks on the Quantum Double,'' Czech. J. Phys. **44**, 1059–1071 (1994).
[7] E. Beggs and S. Majid, ''Quasitriangular and differential structures on bicrossproduct Hopf algebras,'' q-alg/9701041.
[8] J. Lukierski, H. Ruegg, V. N. Tolstoy, and A. Nowicki, ''*q*-Deformation of Poincaré algebra,'' Phys. Lett. B **264**, 331–338 (1991).
[9] S. Majid and H. Ruegg, ''Bicrossproduct structure of κ-Poincaré group and non-commutative geometry,'' hep-th/9405107.
[10] A. Ballesteros, F. J. Herranz, V. A. del Olmo, and M. Santander, ''A new 'null-plane' quantization of quantum Poincaré algebra,'' Phys. Lett. B **351**, 137–145 (1995).
[11] O. Arratia, F. J. Herranz, and M. A. del Olmo, ''Bicrossproduct structure of the null-plane quantum Poincaré algebra,'' J. Phys. A: Math. Gen. **31**, L1–L7 (1998).
[12] A. I. Mudrov, ''Twisting cocycle for null-plane quantized Poincaré algebra,'' J. Phys. A **31**, 6219–6224 (1998).
[13] N. Yu. Reshetikhin, ''Multiparametric quantum groups and twisted quasitriangular Hopf algebras,'' Lett. Math. Phys. **20**, 331–335 (1990).
[14] A. Kempt, ''Multiparameter *R*-matrices sub-quantum groups and generalized twisting method,'' in *Proceedings of the XXth International Conference on Differential Geometric Methods in Theoretical Physics, June 1991, New York*, edited by S. Catto and A. Rocha (World Scientific, Singapore, 1991), p. 546.
[15] R. A. Engeldinger and A. Kempf, ''Sub-Hopf-algebra-induced twists of quantum enveloping algebras,'' J. Math. Phys. **35**, 1931–1938 (1994).
[16] N. Yu. Reshetikhin, L. A. Takhtajan, and L. D. Faddeev, ''Quantization of Lie groups and Lie algebras,'' Leningrad Math. J. **1**, 193–225 (1990).
[17] O. Ogievetski, ''Hopf structures on the Borel subalgebra of sl(2),'' Max-Planck-Institut report No. VPI-Ph/92-99.
[18] A. Ballesteros, F. J. Herranz, and C. M. Perena, ''Null-plane quantum universal $\mathscr{R}$-matrix,'' Phys. Lett. B **351**, 71 (1997).
[19] A. Guichardet, *Cohomologie des groupes topologiques et des algebres de Lie* (Cedic/Fernand Nathan, Paris, 1980).
[20] T. Hodges, ''On the Cremmer-Gervais quantization of SL(n),'' q-alg/9506018.
[21] T. Hodges, ''Nonstandard quantum groups associated to certain Belavin–Drinfeld triples,'' q-alg/9603029.
[22] A. D. Jacobs and J. F. Cornwell, ''Twisting 2-cocycles for the construction of new non-standard quantum groups,'' q-alg/9702028.
[23] A. I. Mudrov, ''Quantum deformations of the Lorentz algebra,'' Phys. At. Nucl. **60**, 848–859 (1997).

Bäcklund transformation of partial differential equations from the Painlevé–Gambier classification. I. Kaup–Kupershmidt equation

Micheline Musette
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, B–1050 Brussel, Belgium

Robert Conte
Service de physique de l'état condensé, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France

(Received 27 January 1998; accepted for publication 17 June 1998)

Among the 50 nonlinear second-order differential equations of Painlevé and Gambier, those which are linearizable provide a natural scheme for deriving the Lax pair and the Darboux transformation of a nonlinear partial differential equation when the order of the scattering problem is three. This new method allows one to obtain, from singularity analysis only, the Bäcklund transformation of the Kaup–Kupershmidt equation, a result which was missing until now. © *1998 American Institute of Physics.* [S0022-2488(98)03610-X]

I. INTRODUCTION

Among the numerous definitions of the *integrability* of a partial differential equation (PDE), one of the most widely accepted is the existence of a Bäcklund transformation (BT).[1,2] A BT between two given PDEs, taken for simplicity in two independent variables x,t,

$$E_1(u,x,t)=0, \quad E_2(U,X,T)=0 \tag{1}$$

is by definition (Ref. 3, Vol. III, Chap. XII, Ref. 4) a pair of relations

$$F_j(u,u_x,u_t,\ldots,U,U_x,U_t,\ldots,x,t,X,T)=0, \quad j=1,2, \tag{2}$$

with a transformation law between (x,t) and (X,T), such that the elimination of u (respectively, U) between (F_1,F_2) implies $E_2(U,X,T)=0$ [respectively, $E_1(u,x,t)=0$]. In case the two PDEs are the same, the BT is called auto-BT.

There exist several direct methods to obtain the BT of a PDE, among them the singularity approach initiated by Weiss, Tabor, and Carnevale (WTC).[5,6] Their ''truncation procedure'' and its homographically invariant version[7] have later been extended to process some equations which escaped the original method: first by the ''two-singular manifold method'' [8,9] to handle the modified Korteweg–de Vries, sine-Gordon and Broer-Kaup equations, second by taking account of involutions[10] to handle the nonlinear Schrödinger equation. The notion of ''singular manifold equation'' of Weiss has also been clarified[11] by implementing an additional degree of freedom of the invariant Painlevé analysis.

Very few well known $1+1$-dimensional PDEs still fail to provide, with the truncation procedure and its extensions, the auto-Bäcklund transformation. One of them is the Kaup–Kupershmidt equation.[12,13] This gap is filled in the present article. Preliminary results have been announced in Cargèse lecture notes.[14]

The paper is organized as follows. In Sec. II, we rephrase the method for PDEs with one family of movable singularities, mainly dissociating the role of two entire functions τ and ψ; the application to the Boussinesq equation allows one to extend an old result of Hirota and Satsuma.

Section III is the core of the paper: We introduce a link between the unknown BT and the nonlinear ordinary differential equations (ODEs) which possess two properties, namely the Painlevé property and the linearizability; this restricts the ODE part of the BT to a very small number

 © 1998 American Institute of Physics

of possibilities, namely the Riccati equation for scattering problems of second order, the Gambier equations numbered 5 and 25 for scattering problems of third order, and it also provides the explicit link between τ and ψ.

In Sec. IV, we recall the singularity structure of Sawada–Kotera (SK) and Kaup–Kupershmidt (KK) equations, two equations deeply interconnected as well as their hierarchies[13,15,16] and argue that, for each of them, only one of their two families is relevant, thus fitting the one-family scheme.

In Sec. V, for both the SK and KK equations, we try the three possibilities allowed at second and third scattering orders, and obtain the BT of both PDEs.

II. METHOD FOR A ONE-FAMILY EQUATION

Let us rephrase, more simply than in our previous work,[7] the procedure for PDEs with only one family of movable singularities, for which the *singular part operator* is denoted $\mathscr{D}$. This will be followed by an example.

First step. Assume a Darboux transformation (DT)[17] defined as

$$u = U + \mathscr{D} \log \tau, \quad E(u) = E(U) = 0, \tag{3}$$

with u, U two different solutions of the PDE under consideration and τ some ''entire'' function. A consequence of this assumption is the existence of the involution

$$(u, U, \tau) \to (U, u, \tau^{-1}), \tag{4}$$

since the operator $\mathscr{D}$ is linear.

Second step. Choose the order two, then three, then ..., for the unknown scalar Lax pair and represent this Lax pair by an equivalent, multicomponent, Riccati pseudopotential $\mathbf{Y} = (Y_1, \dots)$. A second-order scalar Lax pair in canonical form, written here in the case of two independent variables,

$$L_1\psi \equiv \psi_{xx} + \frac{S}{2}\psi = 0, \tag{5}$$

$$L_2\psi \equiv \psi_t + C\psi_x - \frac{C_x}{2}\psi = 0, \tag{6}$$

$$2[L_1, L_2] \equiv S_t + C_{xxx} + CS_x + 2C_xS = 0, \tag{7}$$

is thus represented by the one-component Riccati pseudopotential $\mathbf{Y} = \psi_x/\psi = \chi^{-1}$,

$$(\chi^{-1})_x = -\chi^{-2} - \frac{S}{2}, \tag{8}$$

$$(\chi^{-1})_t = C\chi^{-2} - C_x\chi^{-1} + \frac{CS + C_{xx}}{2}. \tag{9}$$

A third-order scalar Lax pair in canonical form

$$L_1\psi \equiv \psi_{xxx} - a\psi_x - b\psi = 0, \tag{10}$$

$$L_2\psi \equiv \psi_t - c\psi_{xx} - d\psi_x - e\psi = 0, \tag{11}$$

$$[L_1, L_2] \equiv X_0 + X_1\partial_x + X_2\partial_x^2, \tag{12}$$

$$X_0 \equiv -b_t - ae_x + e_{xxx} + b_{xx}c + 3bc_{xx} + 3b_xc_x + 3bd_x + b_xd = 0, \tag{13}$$

$$X_1 \equiv -a_t + 3e_{xx} + 2b_xc + a_{xx}c + d_{xxx} + 3ac_{xx} + 2ad_x + 3a_xc_x + 3bc_x + a_xd = 0, \tag{14}$$

$$X_2 \equiv (2ac + c_{xx} + 3d_x + 3e)_x = 0, \tag{15}$$

is similarly represented by the two-component Riccati pseudopotential $\mathbf{Y} = (Y_1, Y_2)$ of projective type[18,7,19]

$$Y_1 = \frac{\psi_x}{\psi}, \quad Y_2 = \frac{\psi_{xx}}{\psi}, \tag{16}$$

$$Y_{1,x} = -Y_1^2 + Y_2, \tag{17}$$

$$Y_{2,x} = -Y_1 Y_2 + aY_1 + b, \tag{18}$$

$$Y_{1,t} = -(dY_1 + cY_2)Y_1 + (ac + d_x)Y_1 + (c_x + d)Y_2 + e_x + bc \tag{19}$$

$$= (cY_2 + dY_1 + e)_x, \tag{20}$$

$$Y_{2,t} = -(dY_1 + cY_2)Y_2 + (2ac_x + a_x c + bc + d_{xx} + ad + 2e_x)Y_1 + (c_{xx} + 2d_x + ac)Y_2 + 2bc_x + b_x c + bd + e_{xx}, \tag{21}$$

$$(Y_{1,t})_x - (Y_{1,x})_t = X_1 + X_2 Y_1, \tag{22}$$

$$(Y_{2,t})_x - (Y_{2,x})_t = -X_0 + X_2 Y_2. \tag{23}$$

Remark: In case the PDE has more than one family of movable singularities, it is necessary to introduce an additional degree of freedom in the scattering problem, such as a term $A\psi_x$ in (5).[11]

Third step. Choose an explicit link

$$\mathscr{D} \log \tau = f(\psi), \tag{24}$$

between the function τ and the solution ψ of a scalar Lax pair. For most PDEs, this is simply

$$\tau = \psi, \tag{25}$$

but it is more complicated for some tricky PDEs like Kaup–Kupershmidt (KK), see (63) below. It will be shown in Sec. III that, at each scattering order, there exists only a finite number of choices (24).

Fourth step. Define the ''truncation'' and solve it, that is to say: With the assumptions (3) for a DT, (24) for a link between τ and ψ, (8)–(9) or (17)–(21) for the gradient of $\mathbf{Y}$, express $E(u) - E(U)$ as a polynomial in $\mathbf{Y}$,

$$E(u) - E(U) = \sum_{j=0}^{-q} E_j(S, C, U)\chi^{j+q} \quad \text{(for second order)}, \tag{26}$$

$$E(u) - E(U) = \sum_{k \geqslant 0} \sum_{l \geqslant 0} E_{k,l}(a, b, c, d, e, U) Y_1^k Y_2^l \quad \text{(for third order)}, \tag{27}$$

and solve the set of *determining equations*

$$\forall j E_j(S, C, U) = 0 \quad \text{(for second order)}, \tag{28}$$

$$\forall k \forall l E_{k,l}(a, b, c, d, e, U) = 0 \quad \text{(for third order)}, \tag{29}$$

for the unknown coefficients (S, C) or (a, b, c, d, e) as functions of U. Our previous formulas (39) and (40) in Ref. 7 contained a useless restriction to the above written formulas (27) and (29).

The second, third and fourth steps must be repeated until a success occurs. The process is successful if and only if the following three conditions are met:

(1) U comes out unconstrained, apart from being a solution of the PDE,

(2) the vanishing of the commutator $[L_1,L_2]$ is equivalent to the vanishing of $E(U)$,
(3) in the $1+1$-dimensional case only, the coefficients depend on an arbitrary constant λ, the spectral or Bäcklund parameter.

At this stage, one has obtained the DT and the Lax pair.

Fifth step. Obtain the two equations for the BT by eliminating ψ^{20} between the DT and the scalar Lax pair. This sometimes uneasy operation may become elementary by eliminating the appropriate component of the multicomponent pseudopotential $\mathbf{Y}$ rather than ψ, and this is the only reason for introducing $\mathbf{Y}$ in the second step. Assume for instance that $\tau=\psi$ and $\mathscr{D}=\partial_x$. Then Eq. (3) reads

$$Y_1=u-U \tag{30}$$

and the BT is computed as follows: Eliminate all the components of $\mathbf{Y}$ but Y_1 between the equations for the gradient of $\mathbf{Y}$, then in the resulting equations substitute Y_1 as defined in (30).

Remark: In this fifth step, the elimination of ψ is made between three equations and results in two equations, as expected for a BT. For an nth order Lax pair with $n-1$ components in $\mathbf{Y}$, the elimination of all components of $\mathbf{Y}$ but Y_1 results in n equations instead of two, but, at least in the case of evolution equations to which we restrict ourselves in this article, $n-2$ of them must be discarded, namely those defining $Y_{k,t}, k=2,\ldots,n-1$. The two resulting equations in Y_1 are then one nonlinear ODE and one PDE. For $n=2$ these are

$$Y_{1,xx}+3Y_1Y_{1,x}+Y_1^3-aY_1-b=0, \tag{31}$$

$$Y_{1,t}-(cY_{1,x}+cY_1^2+dY_1+e)_x=0, \tag{32}$$

$$(Y_{1,xx})_t-(Y_{1,t})_{xx}=X_0+X_1Y_1+X_2Y_1^2=0. \tag{33}$$

To fix the ideas, let us take the Boussinesq equation (Bq) in its two-component evolution form[21]

$$\mathrm{sBq}(u,r)\equiv\begin{cases} u_t-r_x=0,(\alpha,\beta,\varepsilon)\text{constant},\\ r_t+\varepsilon^2((u+\alpha)^2+(\beta^2/3)u_{xx})_x=0, \end{cases} \tag{34}$$

and consider its one-component ''potential'' form

$$\mathrm{pBq}(v)\equiv v_{tt}+\varepsilon^2((v_x+\alpha)^2+(\beta^2/3)v_{xxx})_x=0,\quad u=v_x,\quad r=v_t. \tag{35}$$

Equation (35) has only one family of movable singularities

$$p=-1,\quad q=-5,\quad \text{indices}\ -1,1,4,6,\quad \mathscr{D}=2\beta^2\partial_x \tag{36}$$

and it passes the Painlevé test.[22] Since (35) is a conservation law, the computations to come can be reduced by considering the ''second potential Bq'' equation

$$\mathrm{ppBq}(w)\equiv w_{tt}+\varepsilon^2((w_{xx}+\alpha)^2+(\beta^2/3)w_{xxxx})=0,\quad u=v_x=w_{xx}, \tag{37}$$

whose single family is of the logarithmic type $w\sim 2\beta^2\log\chi$,

$$p=0^-,\quad q=-4,\quad \text{indices}\ -1,0,1,6,\quad \mathscr{D}=2\beta^2. \tag{38}$$

The assumptions (3) and (25)

$$w=2\beta^2\log\tau+W,\quad \log\tau=\log\psi,\quad \mathrm{ppBq}(w)=0,\quad \mathrm{ppBq}(W)=0, \tag{39}$$

with ψ solution of the unknown second-order scalar Lax pair (5)–(6), generate the usual WTC truncation in the invariant formalism[23]

$$\mathrm{ppBq}(w)-\mathrm{ppBq}(W)\equiv\sum_{j=0}^{4}E_j\chi^{j-4}=0, \tag{40}$$

$$E_2\equiv(4/3)\beta^2\varepsilon^2S-2C^2-4\varepsilon^2(W_{xx}+\alpha)=0, \tag{41}$$

$$E_3\equiv-2(C_t+CC_x-(\beta^2\varepsilon^2/3)S_x)=0, \tag{42}$$

$$E_4\equiv(SE_2-E_{3,x})/2+C_x^2=0. \tag{43}$$

Their general solution depends on three arbitrary constants

$$S=3(c_2-c_0c_1t-c_1^2t^2/2+c_1x)/(\beta\varepsilon)^2, \tag{44}$$

$$C=c_1t+c_0, \tag{45}$$

$$W_{xx}=U=-\alpha+(c_1x-c_1^2t^2-2c_0c_1t+c_2-c_0^2/2)/\varepsilon^2, \tag{46}$$

and, because of the constraint on U, it cannot define a Lax pair and just defines the particular solution

$$u=-\alpha+(c_1x-c_1^2t^2-2c_0c_1t+c_2-c_0^2/2)/\varepsilon^2+2\beta^2\partial_x^2\ \log\ \mathrm{Ai}(S), \tag{47}$$

where Ai is the Airy function, solution of the ODE (5). In case $c_1=0$, this is the one-soliton solution.

Going to third order, the assumption (39), with ψ solution of the unknown scalar Lax pair (10)–(11), generates

$$\mathrm{ppBq}(w)-\mathrm{ppBq}(W)\equiv\sum_{k=0}^{2}\sum_{l=0}^{2}E_{k,l}Y_1^kY_2^l,\quad k+l\leqslant 2. \tag{48}$$

These six determining equations $E_{k,l}=0$, plus the three compatibility conditions $X_j=0$, $j=0,1,2$, are solved as follows:[19]

$$\begin{aligned}
&E_{02}\equiv(\beta\varepsilon)^2-c^2=0 && \Rightarrow c=\beta\varepsilon,\\
&E_{11}\equiv d=0 && \Rightarrow d=0,\\
&E_{20}\equiv 3(V_x+\alpha)+2\beta^2a=0 && \Rightarrow a=-3(V_x+\alpha)/(2\beta^2),\\
&E_{10}\equiv\varepsilon V_{xx}-\beta e_x=0 && \Rightarrow e_x=\beta^{-1}\varepsilon V_{xx},\\
&X_1\equiv 3V_{xt}+3\beta\varepsilon V_{xxx}+4\beta^3\varepsilon b_x=0 && \Rightarrow b=g(t)-3(\beta^{-2}V_{xx}+\beta^{-3}\varepsilon^{-1}V_t)/4,\\
&E_{00}\equiv 2\beta^2g'(t)=0 && \Rightarrow g(t)=\lambda,
\end{aligned} \tag{49}$$

in which λ is an arbitrary constant. The coefficients a,b,c,d,e are

$$a=-(3/2)\beta^{-2}(V_x+\alpha),\quad b=\lambda-(3/4)\beta^{-2}V_{xx}-(3/4)\beta^{-3}\varepsilon^{-1}V_t,$$

$$c=\beta\varepsilon,\quad d=0,\quad e=\beta^{-1}\varepsilon(V_x+\alpha), \tag{50}$$

$$X_0=(3/(4\varepsilon\beta^2))\mathrm{pBq}(V),\quad X_1=0,\quad X_2=0, \tag{51}$$

and they define a third-order Lax pair of the potential Boussinesq equation (35).[24,21,25]

The BT is just (31)–(32) or equivalently, after substitution of $Y_1=(v-V)/(2\beta^2)$,

$$(v-V)_{xx}+3\beta^{-1}\varepsilon^{-1}(v+V)_t+3\beta^{-2}(v-V)((v+V)_x+2\alpha)+\beta^{-4}(v-V)^3-8\beta^2\lambda=0, \tag{52}$$

$$(v+V)_{xx}-\beta^{-1}\varepsilon^{-1}(v-V)_t+\beta^{-2}(v-V)(v-V)_x=0. \tag{53}$$

Written in bilinear notation,

$$(\beta\varepsilon D_x^3+3D_xD_t+(6\alpha\beta^{-1}\varepsilon-3K)D_x-4\beta\varepsilon\lambda)f.g=0, \tag{54}$$

$$(\beta\varepsilon D_x^2-D_t+K)f.g=0, \tag{55}$$

$$v=2\beta^2\partial_x \log f, \quad V=2\beta^2\partial_x \log g, \tag{56}$$

this result, in which K is an irrelevant arbitrary integration constant, is an extension to $\lambda\neq 0$ of the bilinear BT given by Hirota and Satsuma.[26,27]

III. THE HELP FROM GAMBIER

In the case of $1+1$-dimensional evolution equations to which we restrict ourselves in this article, one of the two equations defining the BT to be found is an ordinary differential equation, e.g., (8) or (31). This nonlinear ODE for Y, with coefficients depending on U and an arbitrary constant λ, has two properties. First, it is linearizable since it results from the Lax pair, a linear system, and the Darboux transformation by an elimination process.[20] Second, it has the Painlevé property since it is linearizable. Therefore, if its order is small (at most three), it belongs to the appropriate finite list established by the Painlevé school between 1900 and 1910.

These very special nonlinear ODEs provide a link to both the Lax pair, via their linearizing transformation, and the Darboux transformation, via an involution which leaves them invariant.

The only ODE of first order and degree one with the Painlevé property is the Riccati equation, it is linearizable into a second-order linear equation and this defines a unique choice (24) for describing scattering problems which have order two.

Next, the ODEs of order two and degree one with the Painlevé property which are also linearizable have been listed by Gambier himself, page 21 of his thesis,[28] these are the equations numbered 5, 6, 14, 24, 25, 27, in his classification of 50 equations inequivalent under the homographic group of transformations, with the respective orders for the associated linear equations, i.e., in our context for the unknown scattering problem: 3, 2, 2, 2, 3, 2 (and 4 or 3 for the case $n=2$ in equation 27).

Therefore, the only two generic choices for describing scattering problems of third order are the two classes of equivalence numbered 5 and 25 by Gambier. The representative equation of interest to us in each class of equivalence is what Painlevé calls the ''complete equation'' (in the present section, the $(')$ symbol means ∂_x)

$$Y''+3YY'+Y^3+rY+q=0, \tag{G5}$$

$$Y''-3Y'^2/(4Y)+3YY'/2+Y^3/4-q'(Y'+Y^2)/(2q)-rY-q=0, \tag{G25}$$

in which q and r are two arbitrary functions. These two classes are equivalent under the birational group, as shown by the explicit transformation between (G5) $(y;q,r)$ and (G25) $(Y;Q,R)$

$$Y=\frac{Q}{2z'+z^2-(Q'/Q)z-R}, \quad z=y+\frac{Q'}{2Q}, \quad 2y=\frac{Y'}{Y}+Y+\frac{Q'}{Q}, \tag{57}$$

$$r=-R+\frac{Q''}{Q}-\frac{5Q'^2}{4Q^2}, \quad q=-\frac{Q}{2}-\frac{R'}{2}+\frac{Q'''}{2Q}-\frac{7Q'Q''}{4Q^2}+\frac{5Q'^3}{4Q^3}. \tag{58}$$

Equation (G5) for $Y(x)$ is identical to (31) and it is linked to the fifth equation chosen by Gambier for $\tilde{Y}(\tilde{x})$ by (Ref. 29 Eq. (1) Tableau IV)

$$\tilde{Y}''+3\tilde{Y}\tilde{Y}'+\tilde{Y}^3-a=0, \tag{59}$$

$$Y=\lambda(x)\tilde{Y}+\mu(x), \quad \tilde{x}=\xi(x), \tag{60}$$

$$\xi' = \lambda, \quad \lambda'/\lambda = -\mu, \quad 2\mu' + \mu^2 + r = 0, \quad a = (r'/2 - q)\lambda^{-3}. \tag{61}$$

The linearizing transformations are

$$\text{(G5)} \quad Y = \frac{\tau'}{\tau} = \frac{\psi'}{\psi}, \quad \psi''' + r\psi' + q\psi = 0, \tag{62}$$

$$\text{(G25)} \quad \begin{cases} Y = \dfrac{\tau'}{\tau} = \dfrac{q}{2z' + z^2 - (q'/q)z - r}, \quad z = \dfrac{\psi'}{\psi}, \\ \psi''' - \dfrac{3q'}{2q}\psi'' - \left(r + \dfrac{q''}{q} - \dfrac{q'^2}{q^2} \right)\psi' - \left(\dfrac{r'}{2} + \dfrac{q}{2} - \dfrac{q'r}{2q} \right)\psi = 0. \end{cases} \tag{63}$$

The corresponding involutions

$$\text{(G5)} \quad (Y, q, r) \to (-Y, -q, r + 6Y'), \tag{64}$$

$$\text{(G25)} \quad (Y, q, r) \to (-Y, -q, r - 3Y' - (q'/q)Y) \tag{65}$$

will be seen to define the Darboux transformation.

The method exposed in Sec. II remains unchanged, apart from its third step: The assumption for a link $\mathscr{D} \log \tau = f(\psi)$ becomes, at the third scattering order, a choice between the two *and only two* above linearization formulas (62) and (63), which thus replace the assumption (10). Consequently, in the determining equations of the fourth step, the unknowns (a, b) are replaced by the unknowns (q, r) of the Gambier equations numbered 5 and 25.

These two and only two possibilities were rediscovered in 1980 in the context of scattering theory by Caudrey[30] and Kaup.[12]

Let us apply this to the Kaup–Kupershmidt equation.

IV. SINGULARITY STRUCTURE OF SK AND KK EQUATIONS

Because of their duality,[12,15] it is convenient to handle at the same time SK and KK. They are defined as

$$\mathrm{SK}(u) \equiv \beta u_t + \left(u_{xxxx} + \frac{30}{\alpha} u u_{xx} + \frac{60}{\alpha^2} u^3 \right)_x = 0, \tag{66}$$

$$\mathrm{pSK}(v) \equiv \beta v_t + v_{xxxxx} + \frac{30}{\alpha} v_x v_{xxx} + \frac{60}{\alpha^2} v_x^3 = 0, \tag{67}$$

$$\mathrm{KK}(u) \equiv \beta u_t + \left(u_{xxxx} + \frac{30}{\alpha} u u_{xx} + \frac{45}{2\alpha} u_x^2 + \frac{60}{\alpha^2} u^3 \right)_x = 0, \tag{68}$$

$$\mathrm{pKK}(v) \equiv \beta v_t + v_{xxxxx} + \frac{30}{\alpha} v_x v_{xxx} + \frac{45}{2\alpha} v_{xx}^2 + \frac{60}{\alpha^2} v_x^3 = 0, \tag{69}$$

in which u denotes the conservative field and v the potential one, with $u = v_x$.

KK has escaped all previous singularity-based attempts: Weiss,[15] Musette and Conte,[7] Conte, Musette, and Pickering (see page 71 in Ref. 10), Pickering.[11]

The t part of the BT of KK has never been published. The x part of the BT of KK obtained by Rogers and Carillo[16] is restricted to a zero Bäcklund parameter λ, and the introduction of such a parameter is probably uneasy since it has never been published.

The DT of KK found by Levi and Ragnisco[31] allows one in principle to deduce the BT by an elimination process between the Lax pair and the DT, but in practice (contrary to the case of SK) this is so difficult that it has not yet been done, to our knowledge.

Both equations have the Painlevé property.[15] Each of them has two families of movable singularities, whose leading order (in potential form) $v \sim v_0 \chi^p$ and Fuchs indices are the following:[15]

$$\text{pSK, F1: } p=-1, \quad v_0=\alpha, \quad \text{indices } -1,1,2,3,10, \tag{70}$$

$$\text{pSK, F2: } p=-1, \quad v_0=2\alpha, \quad \text{indices } -2,-1,1,5,12, \tag{71}$$

$$\text{pKK, F1: } p=-1, \quad v_0=\alpha/2, \quad \text{indices } -1,1,3,5,7, \tag{72}$$

$$\text{pKK, F2: } p=-1, \quad v_0=4\alpha, \quad \text{indices } -7,-1,1,10,12. \tag{73}$$

The singular part operator $\mathscr{D}$ attached to a given family is $\mathscr{D}=v_0\partial_x$.

The two families have residues which are not opposite, which makes the two-singular manifold method inapplicable.[8]

Hopefully, each potential equation possesses in its hierarchy a ''minus-one'' equation[32]

$$\text{pSK}_{-1}: v_{xxt}+\frac{6}{\alpha}v_x v_t=0, \tag{74}$$

$$\text{pKK}_{-1}: v_t v_{xxt}-\frac{3}{4}v_{xt}^2+\frac{6}{\alpha}v_x v_t^2=0. \tag{75}$$

These equations have only one family, and the first one, initially written by Hirota and Satsuma,[33] has already been processed successfully[7,19] by the one-family method recalled in Sec. II. This is a strong indication that the same method should also succeed for the KK_{-1} and the KK equations, by restricting to their unique common family F1 Eq. (72).

V. ONE-FAMILY TRUNCATION OF THE KK EQUATION

We follow the method of Sec. II and process the potential equations.

First step. The assumption (3) is

$$v=v_0\partial_x \log \tau+V, \tag{76}$$

with (v,V) two solutions pSK or pKK, and v_0 the residue for family F1.

Second step. One successively chooses the second-order Lax pair (5)–(6) and the third order one (10)–(11).

Third step. If the Lax order is two, the only choice is Riccati, i.e., $\tau=\psi$, since this is a one-family truncation. If the order is three, the two choices are (G5) and (G25), i.e., (62) and (63).

Fourth step. For each PDE, there are only three possibilities to examine: order two and Riccati, order three and (G5), order three and (G25).

A. First possibility: Order two and Riccati, for both PDEs

The one-family truncation (26) with $\tau=\psi$ generates the three equations for SK:[34]

$$E_4 \equiv \beta C-4S^2+9S_{xx}+60SV_x/\alpha-180(V_x/\alpha)^2-30V_{xxx}/\alpha=0, \tag{77}$$

$$E_5 \equiv -\beta C_x-2SS_x+S_{xxx}+30S_xV_x/\alpha=0, \tag{78}$$

$$E_6 \equiv (SE_4-E_{5,x})/2+5S_x(3V_{xx}/\alpha-S_x/2)=0, \tag{79}$$

and the three ones for KK:

$$E_2 \equiv 15(S/4-3V_x/\alpha)=0, \tag{80}$$

$$E_4 \equiv \beta C/2+7S^2/4+3S_{xx}/4-15SV_x/\alpha-90(V_x/\alpha)^2-15V_{xxx}/\alpha=0, \tag{81}$$

$$E_6 \equiv (S/2)E_4/\alpha + (4\beta(CS + C_{xx}) - S^3 + (21/2)S_x^2 + 14SS_{xx} - 4S_{xxxx})/16 + (15/4)(3S^2 - 2S_{xx})V_x/\alpha - 45S(V_x/\alpha)^2 = 0. \quad (82)$$

The constraints $S_x(6V_x - \alpha S)_x = 0$ for SK and $12V_x - \alpha S = 0$ for KK imply that only particular solutions are available, meaning a failure to obtain the Lax pair at this second order. For SK, one obtains the two particular solutions,

$$V = 2\alpha(\zeta - 12g_3 t/\beta), \quad (83)$$

$$S = -12\wp - 6K, \quad (84)$$

$$C = 144(\wp^2 - K\wp + K^2)/\beta - 2c, (K, c, g_3, x_0) \text{arbitrary constants}, \quad (85)$$

in which the Weierstrass functions $\wp(z, g_2, g_3)$ and $\zeta(z, g_2, g_3)$ are taken at point $(x - ct - x_0, \beta c/12, g_3)$, and

$$V = \alpha\left(\zeta(x - (c - c_0)t - x_0, \frac{k^4}{12} + \beta\frac{c_0}{9}, g_3) - \frac{k^2}{12}x + (5(k^4 + \beta c_0)\frac{k^2}{36} - 12g_3)\frac{t}{\beta}\right), \quad (86)$$

$$S = -k^2/2, \quad (87)$$

$$C = k^4/\beta + 5c_0/3, \quad (88)$$

$$c = k^4/\beta + 2c_0/3, (k^2, c_0, g_3, x_0) \text{arbitrary constants}. \quad (89)$$

For KK, one obtains the particular solution

$$V = \alpha\left(-\frac{k^2}{24}(x - x_0) + \frac{5}{1152}k^6 t/\beta\right), \quad (90)$$

$$S = -k^2/2, \quad (91)$$

$$C = \frac{k^4}{16\beta}, (k^2, x_0) \text{arbitrary constants}. \quad (92)$$

Given these values of the homographic invariants (S, C), the integration of the system (5)–(6) provides another solution v defined by Eq. (76). For the second above listed solution of SK, one thus obtains a degeneracy of the more general solution

$$v/\alpha = \zeta(x - ct - x_0, g_2, g_3) + \zeta(x - c't - x_0', g_2', g_3') - 12(g_3 + g_3')t/\beta, \quad (93)$$

$$\beta c = 15g_2' - 3g_2, \quad (94)$$

$$\beta c' = 15g_2 - 3g_2', \quad (95)$$

depending on six arbitrary constants $(x_0, g_2, g_3, x_0', g_2', g_3')$; this is an extrapolation to two different velocities c, c' of the solitary wave found by Chazy in 1910.[35]

In the case of KK, one does not even obtain the one-soliton solution.

Remarks:

(1) Had we left V unconstrained instead of requiring it to satisfy the PDE, we would have recovered the result of Weiss,[15] namely: from SK, Miura transformations to KK and to the Fordy–Gibbons equation[13] and a second-order Lax pair of KK without spectral parameter; from KK, Miura transformations to SK and to the Fordy–Gibbons equation and a second-order Lax pair of SK without spectral parameter.

(2) If the PDE were KdV5, the method would succeed at this point.

B. Second possibility: Order three and G5, for both PDEs

This generates 6 determining equations (29) for SK and 13 for KK, which are solved exactly as was done for Boussinesq in Sec. II.

The process is successful for SK:[15,36]

$$q=-\lambda, \tag{96}$$

$$r=6U/\alpha, \tag{97}$$

$$L_1=\partial_x^3+6\frac{U}{\alpha}\partial_x-\lambda, \tag{98}$$

$$L_2=\beta\partial_t+\left(18\frac{U_x}{\alpha}-9\lambda\right)\partial_x^2+\left(36\frac{U^2}{\alpha^2}-6\frac{U_{xx}}{\alpha}\right)\partial_x-36\lambda\frac{U}{\alpha}, \tag{99}$$

$$\beta[L_1,L_2]=(6/\alpha)\mathrm{SK}(U). \tag{100}$$

This is the Lax pair given by Satsuma and Kaup.[37]

For KK, this truncation fails and provides no solution at all, not even the one-soliton solution. Indeed, the one-soliton solution of Kaup corresponds to constant coefficients for the scalar Lax pair (10)–(11), and, with the above procedure, the only way to obtain it[38,11] is to enforce the two first integrals K_1 and K_2 which result from the zero value of b,

$$K_1=\psi_{xx}-a\psi, \quad K_2=\psi_x^2-a\psi^2-2K_1\psi. \tag{101}$$

C. Third possibility: Order three and G25

This has to be done for KK only. We do not assume $\mathrm{pKK}(V)=0$, so we suppress the second term $-\mathrm{pKK}(V)$ in the lhs of Eq. (27). This generates 38 determining equations (29) in the basis $(\psi_x/\psi,\psi_{xx}/\psi-((\psi_x/\psi)^2+q_x\psi_x/(q\psi)+r)/2)$, out of which only 15 are linearly independent; they are solved as follows (g_k denotes an arbitrary integration function, λ an arbitrary constant):

$$\begin{gathered}
E_{4,5}: q=2g_1(t),\\
E_{4,4}: c=9g_1/\beta,\\
E_{3,5}: r=-6V_x/\alpha,\\
E_{1,5}: d=(3V_{xxx}/\alpha-(6V_x/\alpha)^2)/\beta,\\
E_{0,5}: g_1=\lambda,\\
E_{0,6}: \mathrm{pKK}(V)=0,\\
E_{1,4}: e=g_2(t)+(36\lambda V_x/\alpha+72V_xV_{xx}/\alpha^2+3V_{xxxx}/\alpha)/\beta,\\
X_2: a=g_3(t)-6V_x/\alpha,\\
X_1: b=g_4(t)-3V_{xx}/\alpha+g_3\lambda^{-1}((2V_x/\alpha)^2+V_{xxx}/(3\alpha))+\beta x g_3'/(18\lambda),\\
X_0: g_3=0 \ \text{ and } \ g_4=\lambda,
\end{gathered} \tag{102}$$

and the result is, with $U=V_x$,

$$q=2\lambda, \tag{103}$$

$$r=-6U/\alpha, \tag{104}$$

$$\partial_x \log \tau=\frac{\lambda}{\psi_{xx}/\psi-(1/2)(\psi_x/\psi)^2+3(U/\alpha)}, \tag{105}$$

$$L_1=\partial_x^3+6\frac{U}{\alpha}\partial_x+3\frac{U_x}{\alpha}-\lambda, \tag{106}$$

$$L_2=\beta\partial_t-9\lambda\partial_x^2+\left(3\frac{U_{xx}}{\alpha}+36\frac{U^2}{\alpha^2}\right)\partial_x-3\frac{U_{xxx}}{\alpha}-72\frac{UU_x}{\alpha^2}-36\lambda\frac{U}{\alpha}, \tag{107}$$

$$\beta[L_1,L_2]=(6/\alpha)\mathrm{KK}(U)\partial_x+(3/\alpha)\mathrm{KK}(U)_x. \tag{108}$$

This is the Lax pair given by Kaup.[12] The integration of the first-order ODE (105) *modulo* the Lax pair yields the DT given by Levi and Ragnisco:[31]

$$\tau=\psi\psi_{xx}-(1/2)\psi_x^2+3(U/\alpha)\psi^2,\quad \tau_x=\lambda\psi^2, \tag{109}$$

which shows that τ is an entire function, as expected.

Although the relation $\tau_x/\psi^2=$constant is the same as in the case of KdV (see Eq. (4.14) in Ref. 5), it cannot be taken as an *a priori* assumption, it is the result of the method.

Starting from the *vacuum* solution $U=0$, the general solution ψ of $L_1\psi=0, L_2\psi=0$,

$$\psi=c_1e^{Kx+9K^5t/\beta}+c_2e^{jKx+9j^2K^5t/\beta}+c_3e^{j^2Kx+9jK^5t/\beta},\quad j^3=1,\quad K^3=\lambda, \tag{110}$$

in which c_1,c_2,c_3,K are arbitrary complex constants, leads by (3) to the one-soliton solution of Kaup[12]

$$u=(\alpha/2)\partial_x^2\ \log(2+\cosh(k/2)(x-(k/2)^4t/\beta)),\quad k\in\mathscr{R} \tag{111}$$

for the choice $(c_1,c_2,c_3)=(0,j^2,-j)$, $K^2=-k^2/12$, which corresponds to the entire function

$$\tau=-(k^2/12)(2+\cosh(k/2)(x-(k/2)^4t/\beta))e^{(k/2)(x+(k/2)^4t/\beta)}. \tag{112}$$

The fifth step is performed in Sec. VI.

VI. BÄCKLUND TRANSFORMATION

This is the fifth and last step.

The two equations which define the BT result from the elimination of ψ between the DT and the two equations of the scalar Lax pair.

In order to perform this elimination easily, it is convenient to choose one of the two components of the pseudopotential $\mathbf{Y}$ so as to characterize the DT,

$$\mathrm{SK}:\frac{v-V}{\alpha}=\frac{\psi_x}{\psi}=Y_1, \tag{113}$$

$$\mathrm{KK}:\frac{2(v-V)}{\alpha}=\frac{\tau_x}{\tau}=Z. \tag{114}$$

The chosen equivalent systems are the canonical projective Riccati system (17)–(21) in (Y_1,Y_2) for SK, and the following system in (Y_1,Z) for KK:

$$Y_1=\frac{\psi_x}{\psi},\quad Z=\frac{\tau_x}{\tau}, \tag{115}$$

$$Y_{1,x}=-Y_1^2/2+\lambda Z^{-1}-3U/\alpha, \tag{116}$$

$$Z_x=2Y_1Z-Z^2, \tag{117}$$

$$\beta Y_{1,t}=[9\lambda Y_1^2/2-(3U_{xx}/\alpha+36(U/\alpha)^2)Y_1+9\lambda^2Z^{-1}+3U_{xxx}/\alpha+72UU_x/\alpha^2+9\lambda U/\alpha]_x, \tag{118}$$

$$\beta Z_t=[18\lambda U/\alpha+9\lambda^2 Z^{-1}+9\lambda Y_1^2 + (45(U/\alpha)^2+6(U_{xx}/\alpha)-18(U_x/\alpha)Y_1+27(U/\alpha)Y_1^2+(9/4)Y_1^4)Z]_x. \quad (119)$$

The BT of SK results from the elimination of Y_2, which provides Eqs. (31)–(32) for $Y_1=Y$,

$$Y_{xx}+3YY_x+Y^3+6(U/\alpha)Y-\lambda=0, \quad (120)$$

$$\beta Y_t-9[(\lambda-2U_x/\alpha)(Y_x+Y^2)+4(\lambda U/\alpha-(U/\alpha)^2Y)+(2/3)(U_{xx}/\alpha)Y]_x=0, \quad (121)$$

$$\beta((Y_{xx})_t-(Y_t)_{xx})/Y=-(6/\alpha)\mathrm{SK}(U), \quad (122)$$

followed by the substitution $Y=(v-V)/\alpha$,

$$(v-V)_{xx}/\alpha+3(v-V)(v+V)_x/\alpha^2+(v-V)^3/\alpha^3-\lambda=0, \quad (123)$$

$$\beta(v-V)_t/\alpha-(3/2)[(v-V)_{xxxx}/\alpha+(5(v-V)(v+V)_{xxx} + 15(v+V)_x(v-V)_{xx})/\alpha^2+(15(v-V)^2(v-V)_{xx}+30(v-V)(v+V)_x^2)/\alpha^3 + 30(v-V)^3(v+V)_x/\alpha^4+6(v+V)^5/\alpha^5]_x=0, \quad (124)$$

a result due to Satsuma and Kaup.[37]

The BT of KK arises from the elimination of Y_1 between (116), (117) and (119) [Eq. (118) must be discarded], which results in the two equations for $Z=Y$,

$$Y_{xx}-(3/4)Y_x^2/Y+3YY_x/2+Y^3/4+6(U/\alpha)Y-2\lambda=0, \quad (125)$$

$$\beta Y_t-(3/16)[3Y^5+15YY_x^2+30Y^2Y_{xx}+8Y_{xxxx}+30(Y^3+2Y_{xx})(Y_x+4V_x/\alpha) + 60Y(Y_x+4V_x/\alpha)^2+30Y_x(Y_{xx}+4V_{xx}/\alpha)+20Y(Y_{xxx}+4V_{xxx}/\alpha)]_x=0, \quad (126)$$

$$\beta((Y_{xx})_t-(Y_t)_{xx})/Y=-(6/\alpha)\mathrm{KK}(U), \quad (127)$$

followed by the substitution $Y=2(v-V)/\alpha$,

$$(v-V)_{xx}/\alpha-(3/4)(v-V)_x^2/(\alpha(v-V))+3(v-V)(v+V)_x/\alpha^2+(v-V)^3/\alpha^3-\lambda=0, \quad (128)$$

$$\beta(v-V)_t/\alpha-(3/2)[2(v-V)_{xxxx}/\alpha+60(v-V)^3(v+V)_x/\alpha^4+12(v-V)^5/\alpha^5 + (10(v-V)(v+V)_{xxx}+30(v+V)_x(v-V)_{xx}+15(v-V)_x(v+V)_{xx})/\alpha^2 + (30(v-V)^2(v-V)_{xx}+60(v-V)(v+V)_x^2+15(v-V)(v-V)_x^2)/\alpha^3]_x=0. \quad (129)$$

The simple form of the conservative equations (124) and (129) results from the addition of suitable differential consequences of (123) and (128).

If we write (116)–(117) in the variables $(Y_1, Z_2=Z^{-1})$,

$$Y_{1,x}=-Y_1^2/2+\lambda Z_2-3U/\alpha, \quad (130)$$

$$Z_{2,x}=-2Y_1Z_2+1, \quad (131)$$

both systems (17)–(18) and (130)–(131) are coupled Riccati systems, with the difference that the transformation from Y_1 to ψ_x/ψ is a point transformation while the one from Z_2 to ψ_x/ψ is a contact one. Thus, the Riccati system (17)–(18) is in the classification of linearizable coupled Riccati systems given by Lie (this is the projective one), while the Riccati system (130)–(131) is outside it.

The bilinear representation of the BT of SK

$$(D_x^3-\lambda)f.g=0, \tag{132}$$

$$(D_x^5+5\lambda D_x^2-(2/3)\beta D_t)f.g=0, \tag{133}$$

$$v=\alpha\partial_x \log f, \quad V=\alpha\partial_x \log g, \tag{134}$$

was already known;[37] the one for KK requires four fields

$$(D_x^4-8\lambda D_x)f.g-(1/2)(gh_1+fh_2)=0, \tag{135}$$

$$(D_x^5-(16\beta/9)D_t)f.g-(5/6)(D_xg.h_1-D_xf.h_2)=0, \tag{136}$$

$$D_x^4f.f-fh_1=0, \tag{137}$$

$$D_x^4g.g-gh_2=0, \tag{138}$$

$$v=(\alpha/2)\partial_x \log f, \quad V=(\alpha/2)\partial_x \log g, \tag{139}$$

and this is a new result, for which the help of Johan Springael is warmly acknowledged. Let us recall, for reference, the bilinear representation of the PDEs[37,39]

$$\text{SK:}(D_x^6+\beta D_tD_x)f.f=0, \quad v=\alpha\partial_x \log f, \tag{140}$$

$$\text{KK:}(D_x^6+16\beta D_xD_t)f.f+15D_x^2f.h=0, \quad D_x^4f.f=fh, \quad v=(\alpha/2)\partial_x \log f. \tag{141}$$

VII. CONCLUSION

Finding the BT of a PDE is equivalent to finding its Lax pair and its DT. At every order of the assumed Lax pair with unknown coefficients, there exists only a finite list of admissible Darboux transformations, which makes the process of finding the BT a finite one given some upper bound for the Lax order. In the case of $1+1$-dimensional evolution equations, these DTs are the involutions, like (64) and (65), of nonlinear ODEs classified by the Painlevé school. The Kaup–Kupershmidt equation has offered the simplest, nontrivial application of this new method.

ACKNOWLEDGMENTS

It is a pleasure to thank A. Michel Grundland, Hu Xing-biao, Decio Levi, Andrew Pickering and Johan Springael for numerous discussions. The authors acknowledge the financial support of the Tournesol Grant No. T 95/004. M.M. acknowledges the financial support extended within the framework of the IUAP Contract No. P4/08 funded by the Belgian government.

[1] A. V. Bäcklund, ''Om ytor med konstant negativ krökning,'' Lunds Universitets Arsskrift Avd. **2**, 19 (1883).
[2] C. Rogers and W. F. Shadwick, *Bäcklund Transformations and their Applications* (Academic, New York, 1982).
[3] G. Darboux, *Leçons sur la théorie générale des surfaces et les applications géométriques du calcul infinitésimal* (Gauthier-Villars, Paris, 1894), Vol. III; *Théorie générale des surfaces* (Chelsea, New York, 1972), reprint; (Gabay, Paris, 1993), reprint, see original title.
[4] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons* (Springer, Berlin, 1991).
[5] J. Weiss, M. Tabor, and G. Carnevale, ''The Painlevé property for partial differential equations,'' J. Math. Phys. **24**, 522–526 (1983).
[6] J. Weiss, ''The Painlevé property for partial differential equations. II. Bäcklund transformation, Lax pairs, and the Schwarzian derivative,'' J. Math. Phys. **24**, 1405–1413 (1983).
[7] M. Musette and R. Conte, ''Algorithmic method for deriving Lax pairs from the invariant Painlevé analysis of nonlinear partial differential equations,'' J. Math. Phys. **32**, 1450–1457 (1991).
[8] M. Musette and R. Conte, ''The two-singular manifold method. I. Modified KdV and sine-Gordon equations,'' J. Phys. A **27**, 3895–3913 (1994).
[9] R. Conte, M. Musette, and A. Pickering, ''The two-singular manifold method. II. Classical Boussinesq system,'' J. Phys. A **28**, 179–187 (1995).
[10] R. Conte and M. Musette, ''Beyond the two-singular manifold method, in *Nonlinear Physics: Theory and Experiment*, edited by E. Alfinito, M. Boiti, L. Martina and F. Pempinelli (World Scientific, Singapore, 1996), pp. 67–74.
[11] A. Pickering, ''The singular manifold method revisited,'' J. Math. Phys. **37**, 1894–1927 (1996).
[12] D. J. Kaup, ''On the inverse scattering problem for cubic eigenvalue problems of the class $\psi_{xxx}+6Q\psi_x+6R\psi=\lambda\psi$,'' Stud. Appl. Math. **62**, 189–216 (1980).

[13] A. P. Fordy and J. Gibbons, ''Some remarkable nonlinear transformations,'' Phys. Lett. **75A**, 325–325 (1980).
[14] M. Musette, ''Painlevé analysis for nonlinear partial differential equations,'' in *The Painlevé Property, One Century Later*, edited by R. Conte (Springer, Berlin, 1998), 54 pages.
[15] J. Weiss, ''On classes of integrable systems and the Painlevé property,'' J. Math. Phys. **25**, 13–24 (1984).
[16] C. Rogers and S. Carillo, ''On reciprocal properties of the Caudrey–Dodd–Gibbon and Kaup-Kupershmidt hierarchies,'' Phys. Scr. **36**, 865–869 (1987).
[17] G. Darboux, ''Sur une proposition relative aux équations linéaires,'' C. R. Acad. Sci. Paris **94**, 1456–1459 (1882).
[18] R. Anderson, J. Harnad, and P. Winternitz, ''Systems of ordinary differential equations with nonlinear superposition principles,'' Physica D **4**, 164–182 (1982).
[19] M. Musette and R. Conte, ''Solitary waves and Lax pairs from polynomial expansions of nonlinear differential equations,'' in *Nonlinear Evolution Equations and Dynamical Systems*, edited by M. Boiti, L. Martina, and F. Pempinelli (World Scientific, Singapore, 1992), pp. 161–170.
[20] H. H. Chen, ''General derivation of Bäcklund transformations from inverse scattering problems,'' Phys. Rev. Lett. **33**, 925–928 (1974).
[21] V. E. Zakharov and A. B. Shabat, ''A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem. I.'' Funktsional'nyi Analiz i Ego Prilozheniya **8**, 43–53 (1974) [English: Funct. Anal. Appl. **8**, 226–235 (1974)].
[22] J. Weiss, ''The Painlevé property and Bäcklund transformations for the sequence of Boussinesq equations,'' J. Math. Phys. **26**, 258–269 (1985).
[23] R. Conte, ''Invariant Painlevé analysis of partial differential equations,'' Phys. Lett. A **140**, 383–390 (1989).
[24] V. E. Zakharov, ''On stochastization of one-dimensional chains of nonlinear oscillators,'' Zh. Eksp. Teor. Fiz. **65**, 219–225 (1973) [English: Sov. Phys. JETP **38**, 108–110 (1974)].
[25] H. C. Morris, ''Prolongation structures and a generalized inverse scattering problem,'' J. Math. Phys. **17**, 1867–1869 (1976).
[26] R. Hirota and J. Satsuma, ''Nonlinear evolution equations generated from the Bäcklund transformation for the Boussinesq equation,'' Prog. Theor. Phys. **57**, 797–807 (1977).
[27] R. Hirota and J. Satsuma, ''A simple structure of superposition formula of the Bäcklund transformation,'' J. Phys. Soc. Jpn. **45**, 1741–1750 (1978).
[28] B. Gambier, ''Sur les équations différentielles du second ordre et du premier degré dont l'intégrale générale est à points critiques fixes,'' thèse, Paris, 1909; Acta Math. **33**, 1–55 (1910).
[29] P. Painlevé, ''Sur les équations différentielles du second ordre et d'ordre supérieur dont l'intégrale générale est uniforme,'' Acta Math. **25**, 1–85 (1902).
[30] P. J. Caudrey, ''The inverse problem for the third order equation $u_{xxx}+q(x)u_x+r(x)u=-i\zeta^3 u$,'' Phys. Lett. **79A**, 264–268 (1980).
[31] D. Levi and O. Ragnisco, ''Non-isospectral deformations and Darboux transformations for the third-order spectral problem,'' Inverse Probl. **4**, 815–828 (1988).
[32] R. Willox (private communication).
[33] R. Hirota and J. Satsuma, ''N-soliton solutions of model equations for shallow water waves,'' J. Phys. Soc. Jpn. **40**, 611–612 (1976).
[34] R. Conte, ''Painlevé singular manifold equation and integrability,'' in *Inverse Methods in Action*, edited by P. C. Sabatier (Springer, Berlin, 1990), pp. 497–504.
[35] J. Chazy, ''Sur les équations différentielles du troisième ordre et d'ordre supérieur dont l'intégrale générale a ses points critiques fixes,'' thèse, Paris, 1910; Acta Math. **34**, 317–385 (1911). Table des matières commentée avec index, R. Conte, 6 pages (1991).
[36] R. Conte, ''Towards the equivalence between integrability and Painlevé test for partial differential equations,'' in *Nonlinear and Chaotic Phenomena in Plasmas, Solids and Fluids*, edited by W. Rozmus and J. A. Tuszyński (World Scientific, Singapore, 1991), pp. 94–101.
[37] J. Satsuma and D. J. Kaup, ''A Bäcklund transformation for a higher order Korteweg-de Vries equation,'' J. Phys. Soc. Jpn. **43**, 692–697 (1977).
[38] R. Conte and M. Musette, ''Linearity inside nonlinearity: Exact solutions to the complex Ginzburg-Landau equation,'' Physica D **69**, 1–17 (1993).
[39] M. Jimbo and T. Miwa, ''Solitons and infinite dimensional Lie algebras,'' Publ. RIMS, Kyoto **19**, 943–1001 (1983).

Complementary group resolution of the SU(n) outer multiplicity problem. I. The Littlewood rules and a complementary $\mathcal{U}(2n-2)$ group structure

Feng Pan[a] and J. P. Draayer
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001

(Received 28 April 1997; accepted for publication 8 June 1998)

A complementary group to SU(n) is found that realizes all features of the Littlewood rules for Kronecker products of SU(n) representations. This is accomplished by considering a state of SU(n) to be a special Gel'fand state of the complementary group $\mathcal{U}(2n-2)$ with labels of the latter used to distinguish multiple occurrences of irreducible representations of SU(n) (irreps) in the SU(n)$\times$SU(n)$\downarrow$SU(n) decomposition that is obtained from the Littlewood rules. Furthermore, this realization also helps us to determine SU(n)$\supset$SU($n-1$)$\times$U(1) Reduced Wigner Coefficients (RWCs, frequently called Isoscalar Factors) and Clebsch–Gordan Coefficients [CGCs, or full (nonreduced) Wigner Coefficients] of SU(n), using algebraic or numeric methods, in either the canonical or a noncanonical basis. New explicit formulas for the SU(3) and SU(4) multiplicities are obtained by using this technique. © *1998 American Institute of Physics.* [S0022-2488(98)04410-7]

I. INTRODUCTION

Reduced Wigner Coefficients (RWCs) of SU(n)$\supset$SU($n-1$)$\times$U(1) are of importance in many physical applications. Except for those of SU(2), which have been discussed extensively and expressed in various forms, RWCs of SU(n)$\supset$SU($n-1$)$\times$U(1), which can be used to evaluate Clebsch–Gordan Coefficients (CGCs) of SU(n) in the canonical basis according to the Racah Factorization Lemma, have only been given analytically for some special cases. [Reduced Wigner Coefficients are often called Isoscalar Factors and Clebsch–Gordan Coefficients as used in this article are also known as full (nonreduced) Wigner Coefficients or simply Wigner Coefficients.] The biggest challenge involves the outer multiplicity in the decomposition of Kronecker products of SU(n)$\times$SU(n)$\downarrow$SU(n). The first nontrivial but simplest $n=3$ case was studied as part of the first applications of nonmultiplicity-free CGCs of SU(3) in nuclear and particle physics. There are several distinct approaches to the problem: (i) a tensor operator method; (ii) an infinitesimal generator approach, in which matrix elements of SU(n) generators are used to determine recursion relations for the RWCs and CGCs; (iii) a polynomial basis and generating invariants, in which a convenient model space is used to realize a basis of irreps; and (iv) use of the Schur–Weyl duality relation between SU(n) and the symmetric group S_f. Among these are various means of addressing the problem; indeed, sometimes a combination of two or more methods is used. There are also different schemes for handling the outer multiplicity, especially for SU(3), and these are usually referred to as either the canonical or a noncanonical labeling scheme.

A thoroughly discussed approach to this problem is the canonical unit tensor operator method developed by Biedenharn and collaborators in a series of publications.[1–8] The unit tensor operator approach is particularly useful for deriving multiplicity-free CGCs of U(n). The techniques that are part of this method have also proven to be useful in other approaches, but this methodology has not been used to produce a closed algebraic solution to the general outer multiplicity problem. This method was revisited in the late 1980s in a Bargmann–Hilbert space representation using Vector Coherent State (VCS) theory.[9–11] Although the results seem no simpler than those found earlier, they do show that there is a relationship between U(3)$\supset$U(2) RWCs and $3nj$ coefficients

[a]On leave from the Department of Physics, Liaoning Normal University, Dalian 116029, People's Republic of China. Electronic mail: feng@rouge.phys.lsu.edu

 © 1998 American Institute of Physics

of SU(2), with some of these being a consequence of the Schur–Weyl duality relation between the unitary and symmetric groups given by Ališuaskas *et al.*[12–14]

Noncanonical definitions of SU(n) outer multiplicity labels, especially of SU(3), have also been discussed rather extensively, for example, by Moshinsky *et al.*,[15,16] Derome and Sharp,[17,18] Resnikoff,[19] and Pluhař *et al.*[20,21] A wider class of RWCs has been considered by Hecht,[22] Klimyk and Gavrilik,[23] and Le Blanc and Rowe,[24] who used definitions related to the canonical scheme. Generally, however, these results are for noncanonical labeling schemes. A further example is the extensive work of Ališauskas,[25–28] who investigated paracanonical coupling relations and symmetries and various pseudocanonical coupling schemes, which lead to biorthogonalities among the corresponding coefficients. It should be stated that noncanonical definitions for SU(n) coupling coefficients normally lead to nonorthogonality with respect to the outer multiplicity. In such cases, a Gram–Schmidt process can be adopted to recover orthonormality, but this procedure includes an arbitrary choice in ordering the orthogonalized elements. In addition, except for a few simple cases where analytical expressions are available, only numerical implementations of the corresponding algorithm is possible.[24–28]

The Schur–Weyl duality relation between SU(n) and S_f is another method that has been used by several authors. This method was studied first by Moshinsky,[16] Kramer,[29] and Alisauskas and Jucy,[12–14] who were able to show that the scheme works for nonmultiplicity-free as well as multiplicity-free cases. For nonmultiplicity-free couplings, however, numerical orthogonalizition is again required. This approach is illustrated for some simple cases in the work of Chen *et al.*,[30] and by Pan and Chen for the $U_q(n)$ generalization of U(n).[31]

Based on these methods, several packages have been developed for numerically evaluating CGCs of U(n), especially of SU(3). The earliest one is the well-known Akiyama–Draayer code for SU(3) based on a combination of the tensor operator and infinitesimal generator methods.[32,33] Another is Chen's code for various couplings of U(n) based on symmetric group techniques.[30] Still another is the RWC and CGC code for SU(3) developed by Kaeding and Williams.[34–36]

Very recently, Parkash and Sharatchandra worked out an algebraic formula for the general CGCs of SU(3).[37] The method used in their paper is based on a polynomial realization in Bargmann space using generating functions, which was first studied by Shelepin and Karasev for the multiplicity-free case.[38,39] Up to a normalization factor, the final results are expressed in terms of a restricted sum over 33 variables. To determine the value of a single CGC within this formulation is not easy; neither the algebraic nor numeric results are simple. Nevertheless, it is the first algebraic expression for CGCs of SU(3) with multiplicity. It should be noted, however, that to extend this method to $n \geqslant 4$ cases will be much more complicated. Therefore, another simpler and more direct approach to a resolution of the outer multiplicity problem for SU(n) is desirable.

The present paper is the first (I) in a series which has this as its objective. First of all, a complementary group $\mathcal{U}(2n-2)$ realization of the Kronecker product SU(n)$\times$SU(n)$\downarrow$SU(n) is found according to the well-known Littlewood rules. The scheme gives a simple exact resolution of the outer multiplicity. An analysis of the Littlewood rules is also used to derive a new multiplicity formula for SU(n). Examples are given for the SU(3) and SU(4) cases, which can, in principle, be extended to SU(n). A procedure for evaluating CGCs or RWCs of SU(n)$\supset$SU($n-1$)$\times$U(1) will be given for the SU(3) and SU(4) cases in Parts (II) and (III) of the series, respectively. It should be noted that the RWCs or CGCs obtained in this way are orthogonal with respect to the outer multiplicity labels, and in this sense the scheme is canonical.

II. LITTLEWOOD RULES AND THE COMPLEMENTARY GROUP

The Littlewood rules[40] for determining Kronecker products of SU(n) in SU(n)$\times$SU(n)$\supset$SU(n) is a reflection of the Schur–Weyl duality relation between SU(n) and the symmetric group S_f. According to the Schur–Weyl duality relation, an irrep $[\lambda]$ of SU(n) can also be regarded as the same irrep of S_f with $\Sigma_{i=1}^{n}\lambda_i = f$. Therefore, the Kronecker product of two SU(n) irreps $[\lambda]\times[\mu]$ in the decomposition SU(n)$\times$SU(n)$\downarrow$SU(n) can be obtained from the product of two S functions of the corresponding symmetric groups:[41]

$$[\lambda]\times[\mu] = \sum_{\nu} \{\lambda\mu\nu\}[\nu], \tag{2.1}$$

where $\{\lambda\mu\nu\}$ is the number of occurrences of $[\nu]$ in the product. To determine all the irreps that appear on the rhs of (2.1), one can use the well-known Littlewood rules:[40] First fill in the Young diagram $[\mu]=[\mu_1,\mu_2,\ldots,\mu_n]$ with μ_1 symbols a_1 in the first row, μ_2 symbols a_2 in the second row, μ_3 symbols a_3 in the third row,..., and μ_n symbols a_n in the nth row. Then, the final irrep denoted by a Young diagram $[\nu]$ can be obtained by augmenting the Young diagram $[\lambda]$ with the μ_1 a_1 symbols, μ_2 a_2 symbols,..., and μ_n a_n symbols, respectively, in ways specified by the following three conditions.

(a) No identical symbols should appear in the same column of the diagram.

(b) If the $a_1,a_2,\ldots,a_n$ symbols are counted from right to left starting at the top, then at each stage the number of a_1 symbols must not be less than the number of a_2 symbols, which must not be less than the number of a_3 symbols, and so on.

(c) The Young diagram $[\nu]$ obtained after the addition of each symbol must be standard, that is, $\nu_1\geqslant\nu_2\geqslant\cdots\geqslant\nu_n$.

The Young diagram $[\nu]$ filled with symbols $a_1,a_2,\ldots,a_n$ under restrictions (a)–(c) can be regarded as a special Weyl tableau of a unitary group. Recall some basic definitions for the Weyl tableau: A Weyl tableau is a Young diagram with the boxes filled by a set of ordered indices $a_1,a_2,\ldots,a_n$, with the filling done such that (i) no identical symbols should appear in the same column; (ii) the symbols must be in nondecreasing order from left to right in any row and in increasing order from top to bottom in any column.

Every irrep $[\nu]$ of U(n) can be characterized by a partition $[\nu]\equiv[m_{in}]\equiv[m_{1n},m_{2n},\ldots,m_{nn}]$ of n non-negative integers obeying the relation $m_{in}\geqslant m_{i+1n}$; conversely, every such partition denotes a unique irrep. It is well known that any basis vector of U(n) in its canonical basis, i.e., the basis adapted to $\mathrm{U}(n)\supset\mathrm{U}(n-1)\supset\cdots\supset\mathrm{U}(2)\supset\mathrm{U}(1)$ can be labeled by the Gel'fand symbol,

$$\begin{pmatrix}[\nu]\\(m)\end{pmatrix}=\begin{pmatrix} m_{1n} & m_{2n} & \cdots & m_{n-1n} & m_{nn}\\ m_{1n-1} & m_{2n-1} & \cdots & m_{n-1n-1} & \\ & \cdots & & & \\ & m_{12} & m_{22} & & \\ & & m_{11} & & \end{pmatrix}.$$

Equivalently, sometimes the Weyl tableau is also convenient in labelling the canonical basis of U(n). The one-to-one correspondence between the Gel'fand symbol and the Weyl tableau is realized in the following way:

$$\begin{pmatrix}[\nu]\\(m)\end{pmatrix}\Longleftrightarrow W^{[\nu]}=\begin{array}{|l|l|l|l|l|}\hline f_{11}a_1\text{'s} & f_{12}a_2\text{'s} & \cdots & \cdots & f_{1n}a_n\text{'s}\\\hline f_{22}a_2\text{'s} & f_{23}a_3\text{'s} & \cdots & f_{2n}a_n\text{'s} & \\\cline{1-4} \cdots\cdots\cdots & & & & \\\cline{1-1} f_{nn}a_n\text{'s} & & & & \\\cline{1-1}\end{array} \tag{2.2}$$

where

$$\begin{aligned} &f_{1k}=m_{1k}-m_{1k-1},\quad f_{2k}=m_{2k}-m_{2k-1},\ldots,\\ &f_{k-1k}=m_{k-1k}-m_{k-1k-1},\quad f_{kk}=m_{kk}. \end{aligned} \tag{2.3}$$

In other words, a Weyl tableau $W^{[\nu]}$ filled with $a_1,a_2,\ldots,a_n$, corresponds to the n partitions, $[\nu](=[m_{in}]),[m_{in-1}],\ldots,[m_{i2}]$ and $[m_{11}]$ of a Gel'fand symbol, where $[m_{ik}]$ is the Young diagram resulting from deleting all the boxes in the Weyl tableau occupied by the symbols $a_n,a_{n-1},\ldots,a_{k+1}$.

It is clear that the definition of a Weyl tableau and the rules for placing symbols in a Young diagram given by Littlewood rules are the same except for some of the restrictions given by (b), that is, a_k's can appear in the ith rows with $i<k$, and the number of a_k's can be greater than the number of a_i's with $i<k$ counted from right to left and from top to bottom, while these cases are forbidden by the restriction (b) of the Littlewood rules. Therefore, it is obvious that the Littlewood

rules for placing symbols in a Young diagram can be regarded as a special Weyl tableau for a unitary group. Hence, under the restrictions of Littlewood rules given by (b), one obtains a special Gel'fand basis of a corresponding unitary group, which is called the complementary group for Kronecker products of SU(n).

Assume the irrep $[\lambda]$ has p_1 rows, while $[\mu]$ has p_2 rows. Then, the final irrep $[\nu]$ has, at most, p_1+p_2 rows, with $p_1+p_2 \leqslant n$. Therefore, the complementary group corresponding to the Kronecker product of SU(n) is $\mathcal{U}(p_1+p_2)$. A general SU(n) irrep has, at most, $n-1$ rows because one can always use the equivalence condition $[m_{1n}m_{2n}\cdots m_{nn}]=[m_{1n}-m_{nn},m_{2n}-m_{nn},\ldots,m_{1n-1}-m_{nn}]$ to remove the nth row if it exists. From this, it follows that the minimum complementary group is $\mathcal{U}(2n-2)$ for general Kronecker products of SU(n).

Using the correspondence between the Weyl tableau and a Gel'fand symbol, one can easily find the following relations among coupled and uncoupled state labels of $\mathcal{U}(2n-2)$.

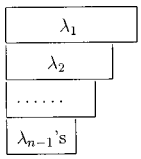

$$\Rightarrow \begin{pmatrix} [\lambda_1\lambda_2\cdots\lambda_{n-1}\dot{0}] \\ \Gamma_1 \end{pmatrix} = \begin{pmatrix} [\lambda_1\lambda_2\cdots\lambda_{n-1}\dot{0}] & \mathcal{U}(2n-2) \\ \cdots\cdots & \\ [\lambda_1\lambda_2\cdots\lambda_{n-1}\dot{0}] & \mathcal{U}(n-1) \\ \rho & \end{pmatrix} \tag{2.4a}$$

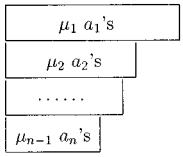

$$\Rightarrow \begin{pmatrix} [\mu_1\mu_2\cdots\mu_{n-1}\dot{0}] \\ \Gamma_2 \end{pmatrix} = \begin{pmatrix} [\mu_1\mu_2\cdots\mu_{n-1}\dot{0}] & \mathcal{U}(2n-2) \\ [\mu_1\mu_2\cdots\mu_{n-2}\dot{0}] & \mathcal{U}(2n-3) \\ \cdots\cdots & \\ [\mu_1\mu_2\dot{0}] & \mathcal{U}(n+1) \\ [\mu_1\dot{0}] & \mathcal{U}(n) \\ [\dot{0}] & \mathcal{U}(n-1) \end{pmatrix} \tag{2.4b}$$

while the final coupled $\mathcal{U}(2n-2)$ basis is

$$\begin{pmatrix} [\nu_1\nu_2\cdots\nu_n\dot{0}] \\ \Gamma(\tau) \end{pmatrix} = \begin{pmatrix} [\nu_1\nu_2\cdots\nu_n\dot{0}] & \mathcal{U}(2n-2) \\ (\tau) & \\ [\lambda_1\lambda_2\cdots\lambda_{n-1}\dot{0}] & \mathcal{U}(n-1) \\ \rho & \end{pmatrix}, \tag{2.4c}$$

where (τ) stands for intermediate sublabels between $\mathcal{U}(2n-2)$ and $\mathcal{U}(n-1)$ given by the Littlewood rules, which can be regarded simultaneously as the outer multiplicity label of both $\mathcal{U}(2n-2)$ and SU(n), and ρ represents sublabels of $\mathcal{U}(n-1)$, which is unimportant for labeling the multiplicity of SU(n)$\times$SU(n).

In order to make this point clear, let us give an example for SU(3)$\times$SU(3)$\downarrow$SU(3) for the coupling [21]$\times$[21]$\downarrow$[321]. According to our procedure, uncoupled basis vectors of SU(3)$\times$SU(3) for the irreps [21]$\times$[21] can be labeled as

$$\left|\begin{matrix} \Gamma_1 \\ [210] \\ (\nu_1) \end{matrix}\right\rangle \left|\begin{matrix} \Gamma_2 \\ [210] \\ (\nu_2) \end{matrix}\right\rangle = \left|\begin{matrix} \rho \\ [21] \\ [210] \\ [210] \\ (\nu_1) \end{matrix}\right\rangle \left|\begin{matrix} 0 \\ [00] \\ [200] \\ [210] \\ (\nu_2) \end{matrix}\right\rangle, \tag{2.5a}$$

where (ν_1), (ν_2) are sublabels for two SU(3) irreps [210], respectively,

$$\Gamma_1 = \begin{pmatrix} [210] \\ [21] \\ \rho \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} [200] \\ [00] \\ 0 \end{pmatrix}, \tag{2.5b}$$

are $\mathcal{U}(4)$ sublabels for the same irreps [2100] required from the Littlewood rule, and ρ stands for unimportant sublabels of $\mathcal{U}(2)$. While the final coupled basis vectors for [321] is denoted by

$$\left|\begin{matrix}\Gamma(\eta)\\ [321]\\ (\nu)\end{matrix}\right\rangle = \left|\begin{matrix}\rho\\ [21]\\ [32-\eta\eta]\\ [321]\\ (\nu)\end{matrix}\right\rangle, \tag{2.5c}$$

where $\eta=0$, or 1, which is determined by the Littlewood rule directly, can be understood as the multiplicity label.

From this procedure, it can be seen that the $\mathcal{U}(2n-2)$ labels are twofold. On the one hand, the final coupled $\mathcal{U}(2n-2)$ sublabels $\Gamma(\eta)$ can be used to label the outer multiplicity of $\mathrm{SU}(n)\times\mathrm{SU}(n)\downarrow\mathrm{SU}(n)$, while the uncoupled $\mathcal{U}(2n-2)$ sublabels Γ_1 and Γ_2 just indicate how the final coupled sublabels of $\mathcal{U}(2n-2)$, which gives the multiplicity labels, can be obtained. General cases for SU(3) and SU(4) will be shown in the next section. Therefore, this situation is similar to the upper Gel'fand pattern introduced by Biedenharn *et al.*,[1–8] in which the upper Gel'fand patterns Γ_2 of $\mathcal{U}(n)$ are used to label the outer multiplicity of $\mathrm{U}(n)\times\mathrm{U}(n)$, while other two upper Gel'fand patterns Γ_1 and Γ of $\mathcal{U}(n)$ are set to the highest weight. While, in our scheme, the sublabels of uncoupled $\mathcal{U}(2n-2)$ labels are fixed according to the Littlewood rule, the final coupled $\mathcal{U}(2n-2)$ sublabels $\Gamma(\eta)$, which can be taken as different values according to the Littlewood rule, are used as outer multiplicity labels for $\mathrm{SU}(n)\times\mathrm{SU}(n)$. However, in the SU(3) case, for example, in their upper pattern scheme, there are some restrictions on the ranges of Γ_{22} in the upper pattern $[\Gamma_{12},\Gamma_{22}]=[m_1+m_2-\lambda_1-2\mu_1-\Gamma_{22},\Gamma_{22}]$ for the coupling $(\lambda_1\mu_1)\times(\lambda_2\mu_2)\downarrow[m_1m_2m_3]$ that come from the Littlewood rules for SU(3). Actually, the ranges of Γ_{22} should be the same as those of η, to be determined in the next section, which cannot be derived directly from restrictions on upper pattern labels. For example, [422] occurs only once in the decomposition $[310]\times[310]$, but there are two sets of upper pattern labels:

$$\begin{pmatrix}\Gamma_{11}\\ \Gamma_{12}\quad\Gamma_{22}\end{pmatrix}=\begin{pmatrix}1\\ 2\quad 0\end{pmatrix};\quad\begin{pmatrix}1\\ 1\quad 1\end{pmatrix}$$

are both allowed by the upper pattern labeling scheme, but according to the Littlewood rules the

$$\begin{pmatrix}1\\ 1\quad 1\end{pmatrix}$$

result should be eliminated. Therefore, in addition to the usual rules for determining allowed upper Gel'fand patterns, restrictions that follow from an application of Littlewood rules must be applied, a result not noted in the papers of Refs. 1–8 and, indeed, one that cannot be deduced directly from their upper Gel'fand pattern methodology. While, in our scheme, the possible values of η gives exactly the outer multiplicity of SU(3). In this sense, the complementary group is meaningless, but it provides us with the labeling scheme for the outer multiplicity of $\mathrm{SU}(n)\times\mathrm{SU}(n)$, and tells us how these labels can be obtained. On the other hand, in order to evaluate the CGCs of $\mathrm{SU}(n)$, one can consider the following $\mathrm{SU}(n)\star\mathcal{U}(2n-2)$ coupling coefficients, for example, in the canonical chain $(\mathrm{SU}(n)\supset\mathrm{SU}(n-1)\supset\cdots\supset\mathrm{U}(1))\star(\mathcal{U}(2n-2)\supset\mathcal{U}(2n-3)\supset\cdots\supset\mathcal{U}(1))$:

$$\left(\begin{matrix}(\Gamma_1) & (\Gamma_2)\\ [\lambda] & [\mu]\\ (\kappa_1) & (\kappa_2)\end{matrix}\middle|\begin{matrix}(\Gamma(\eta))\\ [\nu]\\ (\kappa)\end{matrix}\right), \tag{2.6}$$

where $[\lambda]$, $[\mu]$, and $[\nu]$ are irreps of $\mathrm{SU}(n)$ and $\mathcal{U}(2n-2)$, simultaneously, (Γ_1), (Γ_2), and $(\Gamma(\eta))$ are sublabels of $\mathcal{U}(2n-2)$, and κ_1, κ_2, and κ are sublabels of $\mathrm{SU}(n)$. Using the recoupling method, which will be discussed in detail in paper II, we have

$$\left(\begin{array}{cc|c} (\Gamma_1) & (\Gamma_2) & (\Gamma(\eta)) \\ [\lambda] & [\mu] & [\nu] \\ (\kappa_1) & (\kappa_2) & (\kappa) \end{array}\right) = \sum_{\eta'} \left(\begin{array}{cc|c} \Gamma_1 & \Gamma_2 & \Gamma(\eta') \\ [\lambda] & [\mu] & [\nu] \\ (\kappa_1) & (\kappa_2) & (\kappa) \end{array}\right) \left(\begin{array}{cc|c} \Gamma_1 & \Gamma_2 & \Gamma(\eta') \\ [\lambda] & [\mu] & [\nu] \\ (\Gamma_1) & (\Gamma_2) & (\Gamma(\eta)) \end{array}\right), \tag{2.7}$$

where the coupling coefficient of $SU(n) \star \mathcal{U}(2n-2)$ on the lhs of (2.7) can be evaluated by using many methods, which will be discussed in (II), while, on the rhs of (2.7), there is a sum over the outer multiplicity labels η' of a products of $SU(n)$ and $\mathcal{U}(2n-2)$ CG coefficients, the symbols (Γ_i) with $i=1,2$, and $(\Gamma(\eta))$ are used to denote sublabels of $\mathcal{U}(2n-2)$, which are the same as those derived from the Littlewood rule discussed above, and the same symbols without parentheses are used to label the outer multiplicities in the coupling. Equation (2.7) can be used to derive CGCs of $SU(n)$ algebraically or numerically. A general procedure for SU(3) and SU(4) will be outlined in papers (II) and (III), respectively. From this point of view, the complementary group $\mathcal{U}(2n-2)$ is indeed active in obtaining the CGCs of $SU(n)$.

III. OUTER MULTIPLICITY PROBLEM OF SU(3) AND SU(4)

As noted above, the outer multiplicity in the decomposition of the Kronecker products of $SU(n) \times SU(n) \downarrow SU(n)$ is the main obstacle in applications of algebraic methods to physical problems. There are a lot of articles devoted to this subject. In order to resolve the problem for SU(3), Hecht[22] proposed an external labeling operator of third order, an operator that may be related to the one proposed by Moshinsky[16] in terms of the complementary $U(4) \supset U(2) \times U(2)$ chain. Alisauskas and Kulish[42] have also proposed an external labeling operator, a fourth-order form suggested by Sharp[43] in a study of Yang–Baxter equations. There are also other articles on this subject. For example, new Casimir operators, the so-called chiral Casimirs, were introduced in Refs. 16, and 44–45. Also, various formulas[19,46–49] for the multiplicity of SU(3) exist in the literature, however, such expressions are normally not linked to the SU(3) coupling and recoupling coefficients problem. There is still no general formula for the outer multiplicity of $SU(n)$ with $n \geqslant 4$. In this article and forthcoming papers, the complementary group $\mathcal{U}(2n-2)$ to the $SU(n) \times SU(n) \downarrow SU(n)$ will be shown to be a powerful tool for deriving both multiplicity formulas and coupling and recoupling coefficients of $SU(n)$. Multiplicity formulas for SU(3) and SU(4) are considered below.

(1) *SU(3) case*. Consider the general Keronecker product $(\lambda_1 \mu_1) \times (\lambda_2 \mu_2)$, where the well-known notation for SU(3) in physics is adopted. The irrep $(\lambda \mu)$ can be expressed in terms of a two-rowed Young diagram $[\nu_1 \nu_2]$ with $\nu_1 = \lambda + \mu$ and $\nu_2 = \mu$. Using the Littlewood rules, the decomposition of $(\lambda_1 \mu_1) \times (\lambda_2 \mu_2)$ can be expressed in terms of a quintuple sum,

$$\begin{aligned} (\lambda_1 \mu_1) \times (\lambda_2 \mu_2) = & \sum_{k_1=0}^{\lambda_2+\mu_2} \sum_{k_2=0}^{\min(\lambda_1, \lambda_2+\mu_2-k_1)} \sum_{k_3=0}^{\min(\mu_1, \lambda_2+\mu_2-k_1-k_2)} \sum_{n_1=0}^{\min(\lambda_1+k_1-k_2, \mu_2, k_1)} \\ & \times \sum_{n_2=0}^{\min(\mu_2-n_1, \mu_1+k_2-k_3, k_1+k_2-n_1)} [\lambda_1+\mu_1+k_1, \mu_1+k_2+n_1, k_3+n_2], \end{aligned} \tag{3.1}$$

where the constraints

$$\sum_{i=1}^{3} k_i = \lambda_2 + \mu_2, \quad \sum_{i=1}^{2} n_i = \mu_2 \tag{3.2}$$

apply in the summation. Expression (3.1) can be further simplified, for example, to O'Reilly's formula[48] in which only a triple sum appears. However, (3.1) can be used to help determine a multiplicity formula and determine the multiplicity labels of the complementary group.

Consider a Young diagram of the resultant irrep $[m_1m_2m_3]$ according to (3.1) with conditions given by (3.2):

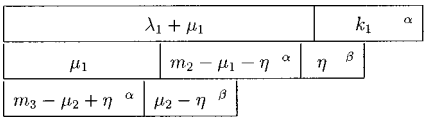

(3.3)

where α and β are the a_i symbols of the Littlewood rules for SU(3). The labels in (3.3) have been arranged to accommodate the constraints of (3.2) and to yield a multiplicity formula in a straightforward manner. In this form it is obvious that a diagram with the same number of boxes in each row can only appear repeatedly when η is not a fixed integer. Therefore, η can be regarded as the multiplicity label of SU(3).

According to Littlewood rules (a)–(c), it is easy to derive the following limits on η:

$$\eta_{\min}\leqslant\eta\leqslant\eta_{\max}, \tag{3.4}$$

where

$$\begin{aligned}\eta_{\max}&=\min(m_1-\lambda_1-\mu_1,\mu_2,m_2-\mu_1,\lambda_2+\mu_2-m_3,\mu_1+\mu_2-m_3,m_2-m_3),\\ \eta_{\min}&=\max(0,\mu_2-m_3,m_2-\lambda_1-\mu_1).\end{aligned} \tag{3.5}$$

Hence, the multiplicity of $[m_1m_2m_3]\equiv(m_1-m_3,m_2-m_3)$ occurring in the Kronecker product $(\lambda_1\mu_1)\times(\lambda_2\mu_2)$ is given by

$$\mathrm{Multi}(\mathrm{SU}_3)=\eta_{\max}-\eta_{\min}+1. \tag{3.6}$$

This expression is very simple and more transparent than others found in the literature.

In this case, the complementary group is $\mathcal{U}(4)$. The Gel'fand symbol of $\mathcal{U}(4)$ corresponding to the resultant irrep of SU(3) given in (3.3) is

$$\begin{pmatrix} & [m_1m_2m_30] & \\ [m_1 & m_2-\eta \quad m_3-\mu_2+\eta] & \\ & [\lambda_1+\mu_1\mu_1] & \\ & \rho & \end{pmatrix}, \quad \eta=\eta_{\min}, \quad \eta_{\min}+1,\ldots,\eta_{\max}, \tag{3.7}$$

where ρ is the intrinsic label for $\mathcal{U}(1)$, which is not important for our purpose. Some conditions in (3.5) can also be easily obtained from the betweenness conditions of the entries in the Gel'fand symbol (3.7). However, the remaining conditions in (3.5) can only be deduced from the Littlewood rules (b), and cannot be obtained from the betweenness conditions. Hence, only one outer multiplicity label is needed in the decomposition of $\mathrm{SU}(3)\times\mathrm{SU}(3)\downarrow\mathrm{SU}(3)$. This is why the CGCs of SU(3) can be determined numerically by using only one type of tensor operator.[32–36]

In contrast with the so-called canonical labeling scheme proposed by Biedenharn *et al.*, in which three independent shifts determined by an upper pattern are introduced, the complementary $\mathcal{U}(4)$ group provides only one outer multiplicity label in the SU(3) case. The complementary group labeling scheme is therefore a very economical way to label the outer multiplicity of SU(3), and by extension of SU(n).

(2) *SU(4) case*. A general SU(4) irrep has three rows. Using the Littlewood rules, the following formula for the decomposition of $\mathrm{SU}(4)\times\mathrm{SU}(4)\downarrow\mathrm{SU}(4)$ can be determined:

$$\begin{aligned}[\lambda_1\lambda_2\lambda_3]\times[\mu_1\mu_2\mu_3]=&\sum_{k_1=0}^{\mu_1}\sum_{k_2=0}^{\min(\mu_1-k_1,\lambda_1-\lambda_2)}\sum_{k_3=0}^{\min(\lambda_2-\lambda_3,\mu_1-k_1-k_2)}\sum_{k_4=0}^{\min(\lambda_3,\mu_1-k_1-k_2-k_3)}\\ &\times\sum_{l_1=0}^{\min(\lambda_1+k_1-\lambda_2-k_2,\mu_2,k_1)}\sum_{l_2=0}^{\min(\lambda_2+k_2-\lambda_3-k_3,\mu_2-l_1,k_1+k_2-l_1)}\end{aligned}$$

$$\times \sum_{l_3=0}^{\min(\lambda_3+k_3-k_4,\mu_2-l_1-l_2,k_1+k_2+k_3-l_1-l_2)} \sum_{n_1=0}^{\min(\lambda_2+k_2+l_1-\lambda_3-k_3-l_2,\mu_3,l_1)}$$

$$\times \sum_{n_2=0}^{\min(\lambda_3+k_3+l_2-k_4-l_3,\mu_3-n_1,l_1+l_2-n_1)}$$

$$\times[\lambda_1+k_1,\lambda_2+k_2+l_1,\lambda_3+k_3+l_2+n_1,k_4+l_3+n_2], \tag{3.8}$$

where the following constraints:

$$\sum_{i=1}^{4} k_i=\mu_1, \quad \sum_{i=1}^{3} l_1=\mu_2, \quad \sum_{i=1}^{2} n_i=\mu_3, \tag{3.9}$$

apply in the summation. In the resultant irrep $[\nu_1\nu_2\nu_3\nu_4]\equiv[\lambda_1+k_1,\lambda_2+k_2+l_1,\lambda_3+k_3+l_2+n_1,k_4+l_3+n_2]$, with the restrictions given by (3.9), there may be six ways to relabel the configuration that leave the irrep unchanged:

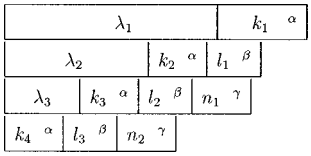

(3.10a)

where

$$k_1=\nu_1-\lambda_1, \qquad k_2=\nu_2-\lambda_2-\xi_1-\xi_2,$$

$$k_3=\nu_3-\lambda_3-\mu_2+\xi_1-\xi_3-\xi_5-\xi_6,$$

$$k_4=\nu_4-\mu_3+\xi_2+\xi_3+\xi_5+\xi_6, \tag{3.10b}$$

$$l_1=\xi_1+\xi_2, \qquad l_2=\mu_2-\xi_1-\xi_4+\xi_5, \qquad l_3=\xi_4-\xi_2-\xi_5,$$

$$n_1=\xi_3+\xi_4+\xi_6, \qquad n_2=\mu_3-\xi_3-\xi_4-\xi_6,$$

and α, β, and γ are the symbols filling in each box according to the Littlewood rules. However, by using the following transformation:

$$\eta_1=\xi_1+\xi_2, \quad \eta_2=\xi_3+\xi_4+\xi_6, \quad \eta_3=\xi_4-\xi_2-\xi_5, \tag{3.11}$$

it can be shown that only three variables η_i with $i=1$, 2, and 3 are independent. Therefore, (3.10) can be relabeled in terms of these three variables:

$$k_1=\nu_1-\lambda_1, \qquad k_2=\nu_2-\lambda_2-\eta_1,$$

$$k_3=\nu_3-\lambda_3-\mu_2+\eta_1-\eta_2+\eta_3,$$

$$k_4=\nu_4-\mu_3+\eta_2-\eta_3, \tag{3.12}$$

$$l_1=\eta_1, \qquad l_2=\mu_2-\eta_1-\eta_3,$$

$$l_3=\eta_3, \qquad n_1=\eta_2, \qquad n_2=\mu_3-\eta_2.$$

Applying the Littlewood rules to this result yields the following boundary conditions for the outer multiplicity labels η_1, η_2, and η_3:

$$\eta_{1\,\min}\leqslant\eta_1\leqslant\eta_{1\,\max}, \quad \eta_{2\,\min}\leqslant\eta_2\leqslant\eta_{2\,\max}, \quad \eta_{3\,\min}\leqslant\eta_3\leqslant\eta_{3\,\max}, \tag{3.13}$$

where

$$\begin{aligned}
&\eta_{1\,\min}=\max(0,\nu_2-\lambda_1),\quad \eta_{1\,\max}=\min(\nu_2-\lambda_2,\nu_1-\lambda_1),\\
&\eta_{2\,\min}=\max(\nu_3-\nu_2+\eta_1,0),\quad \eta_{2\,\max}=\min(\eta_1,\mu_3,\nu_3-\nu_4),\\
&\eta_{3\,\min}=\max(2\eta_2-\eta_1+\mu_2+\nu_4-\nu_3-\mu_3,\eta_2-\eta_1+\lambda_3+\mu_2-\nu_3,\eta_2+\nu_4-\lambda_3\\
&\quad-\mu_3,\eta_2+\lambda_1+\lambda_2+\lambda_3+2\mu_2-\nu_1-\nu_2-\nu_3,\eta_1+\lambda_1+\lambda_2+\mu_2-\nu_1\\
&\quad-\nu_2,0,\mathrm{Int}[(\eta_2+\lambda_1+\lambda_2+\lambda_3+2\mu_2-\nu_1-\nu_2-\nu_3)/2]),\\
&\eta_{3\,\max}=\min(\mu_2-\eta_1,\nu_4-\mu_3+\eta_2,\mu_2-\mu_3,\lambda_2-\nu_3+\mu_2-\eta_1+\eta_2,\mu_2-\eta_2),
\end{aligned}\tag{3.14}$$

where $\mathrm{Int}[x]$ is the integer part of x. Thus, the multiplicity of $[\nu_1,\nu_2,\nu_3,\nu_4]\equiv[\nu_1-\nu_4,\nu_2-\nu_4,\nu_3-\nu_4]$ appearing in the Kronecker product $[\lambda_1\lambda_2\lambda_3]\times[\mu_1\mu_2\mu_3]$ can be calculated by

$$\mathrm{Multi}(\mathrm{SU}_4)=\sum_{\eta_1=\eta_{1\,\min}}^{\eta_{1\,\max}}\ \sum_{\eta_2=\eta_{2\,\min}(\eta_1)}^{\eta_{2\,\max}(\eta_1)}\ \sum_{\eta_3=\eta_{3\,\min}(\eta_1,\eta_2)}^{\eta_{3\,\max}(\eta_1,\eta_2)}1. \tag{3.15}$$

The complementary group of the Kronecker product $[\lambda_1\lambda_2\lambda_3]\times[\mu_1\mu_2\mu_3]$ of SU(4) is $\mathcal{U}(6)$, with the following special Gel'fand labels:

$$\left\langle\begin{array}{cc}
[\nu_1\nu_2\nu_3\nu_4] & \mathcal{U}(6)\\
[\nu_1,\nu_2,\nu_3-\eta_2,\nu_4-\mu_3+\eta_2,0] & \mathcal{U}(5)\\
[\nu_1,\nu_2-\eta_1,\nu_3-\mu_2+\eta_1+\eta_3-\eta_2,\nu_4-\mu_3+\eta_2-\eta_3] & \mathcal{U}(4)\\
[\lambda_1\lambda_2\lambda_3] & \mathcal{U}(3)\\
\rho &
\end{array}\right\rangle. \tag{3.16}$$

The Gel'fand labels of $[\lambda_1\lambda_2\lambda_3 0]$ and $[\mu_1\mu_2\mu_3\dot{0}]$ for $\mathcal{U}(6)$ are

$$\left\langle\begin{array}{cc}
[\lambda_1\lambda_2\lambda_3\dot{0}] & \mathcal{U}(6)\\
[\lambda_1\lambda_2\lambda_3\dot{0}] & \mathcal{U}(5)\\
[\lambda_1\lambda_2\lambda_3\dot{0}] & \mathcal{U}(4)\\
[\lambda_1\lambda_2\lambda_3] & \mathcal{U}(3)\\
\rho &
\end{array}\right\rangle \quad\text{and}\quad \left\langle\begin{array}{cc}
[\mu_1\mu_2\mu_3\dot{0}] & \mathcal{U}(6)\\
[\mu_1\mu_2\dot{0}] & \mathcal{U}(5)\\
[\mu_1 0] & \mathcal{U}(4)\\
[\dot{0}] & \mathcal{U}(3)
\end{array}\right\rangle. \tag{3.17}$$

Again, most of the boundary conditions for the multiplicity labels η_1, η_2, and η_3 can be obtained from the betweenness conditions for the Gel'fand symbol shown in (3.16). However, the remaining conditions can only be deduced from the Littlewood rules because (3.16) is a special Gel'fand basis for the canonical chain $\mathcal{U}(6)\supset\mathcal{U}(5)\supset\cdots\supset(2)\supset\mathcal{U}(1)$. From this development it is clear that there are at most, three quantum numbers needed to label the outer multiplicity for the decomposition $\mathrm{SU}(4)\times\mathrm{SU}(4)\downarrow\mathrm{SU}(4)$. In the canonical unit tensor approach proposed by Biedenharn *et al.* for the SU(4) case, there are four shifts out of six upper labels, of which only three labels are independent.[1–8] As for the SU(3) case, restrictions that follow from the Littlewood rules must be applied to eliminate spurious upper Gel'fand patterns.

It should be noted that any SU(n) function, for example, CGCs, RWCs, or Racah coefficients, etc., is rank n independent, and only depends on boxes contained in the Young diagrams of the corresponding irreps because of the Schur-Weyl duality relation between SU(n) and S_f. For example, the multiplicity of the Kronecker product for two two-rowed irreps of SU(n) is the same as that of SU(3), and that for two three-rowed irreps is the same as that of SU(4), and so on. Hence, the results for SU(3) and SU(4) apply for the general SU(n) case as well. As a trivial example, note that the SU(3) multiplicity expression follows from the one for SU(4) in the two-rowed limit ($\eta_1\rightarrow\eta$, $\eta_2=0$, and $\eta_3=0$) of the theory, though this fact cannot be clearly seen from (3.14).

IV. CONCLUSIONS

In this paper, a complementary group $U(2n-2)$ to $SU(n)$ is found that gives a complete realization of all the features of the Littlewood rules in the Kronecker product decomposition of $SU(n)\times SU(n)\downarrow SU(n)$. By using this scheme, the outer multiplicity labels for $SU(n)$ can be easily assigned, being nothing other than a set of sublabels of the special Gel'fand basis of the complementary $\mathcal{U}(2n-2)$ group. Furthermore, within this framework, most of the boundary conditions on the multiplicity labels can be easily obtain from the betweenness conditions of the Gel'fand symbols of $\mathcal{U}(2n-2)$, while the remaining conditions must be deduced from the Littlewood rules. The method was used to obtain simple multiplicity formulas for SU(3) and SU(4). In addition, in the coupling of two $SU(n)$ irreps, the basis for $SU(n)$ can further be labeled by the final $\mathcal{U}(2n-2)$ sublabels η_i obtained from the coupling of two uncoupled basis vectors of the corresponding special Gel'fand basis of $\mathcal{U}(2n-2)$, which are missing within the $SU(n)$ group. This situation is similar to the canonical unit tensor approach proposed by Biedenharn *et al.* However, in the canonical unit tensor approach, there are n independent shifts indicated by the upper pattern of $U(n)$ from the $n(n-1)/2$ upper labels. While these upper indices can be used to label the outer multiplicity of $U(n)$, there may very well be spurious degrees of freedom among the labels, and these have to be eliminated according to restrictions of the Littlewood rules of $SU(n)$.

It should be noted that the same complementary group for a resolution of $SU(n)$ was considered in Ref. 15. However, the method used and the final outcome are all different from those of this article. First, in Ref. 15, the complementary group structure was derived through boson realizations,[16] whereas in the present case it comes directly from the Littlewood rules. Second, according to Ref. 15, the complementary group should be labeled in terms of a noncanonical chain $\mathcal{U}(2n-2)\supset\mathcal{U}(n-1)\times\mathcal{U}(n-1)$. In this way, the RWCs of $SU(n)$ cannot be derived easily because new inner multiplicity occurs in the decomposition $\mathcal{U}(2n-2)\downarrow\mathcal{U}(n-1)\times\mathcal{U}(n-1)$. In order to overcome this problem, another type of Wigner coefficients, so-called auxiliary Wigner coefficients, was introduced, which is different from the standard definition and which satisfies other orthogonality conditions. These special Wigner coefficients will be considered in the next paper, where it will also be shown that one can derive analytical expressions for some simple cases and algorithms for $SU(n)$ RWCs or CGCs with general multiplicity in canonical as well as noncanonical bases if the multiplicity-free coefficients in these bases are known.

The complementary $\mathcal{U}(2n-2)$ group scheme for labeling outer multiplicities in Kronecker products of $SU(n)$ is canonical because the basis of $SU(n)$ labeled in this way is orthogonal with respect to the outer multiplicity labels. A general procedure for evaluating RWCs or CGCs for SU(3) and SU(4) will be given in forthcoming papers.

ACKNOWLEDGMENT

The project was supported in part by the U.S. National Science Foundation and the State Education Commission of China.

[1] G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963); **5**, 1723 (1964); **5**, 1730 (1964); **6**, 1847 (1965).
[2] L. C. Biedenharn, A. Giovannini, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).
[3] L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1968).
[4] L. C. Biedenharn, J. D. Louck, E. Chacon, and M. Ciftan, J. Math. Phys. **13**, 1957 (1972).
[5] E. Chacon, M. Ciftan, and L. C. Biedenharn, J. Math. Phys. **13**, 577 (1972).
[6] L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **93**, 143 (1984).
[7] L. C. Biedenharn, M. A. Lohe, and J. D Louck, J. Math. Phys. **26**, 1458 (1985).
[8] L. C. Biedenharn, M. A. Lohe, and H. T. Williams, J. Math. Phys. **35**, 6672 (1994).
[9] R. Le Blanc and K. T. Hecht, J. Phys. A **20**, 4613 (1987).
[10] R. Le Blanc and L. C. Biedenharn, J. Phys. A **22**, 4613 (1989).
[11] K. T. Hecht and L. C. Biedenharn, J. Math. Phys. **31**, 2781 (1990).
[12] S. J. Ališauskas, A. A. Jucy, and A. P. Jucy, J. Math. Phys. **13**, 1329 (1972).
[13] S. J. Ališauskas, V. V. Vanagas, and J. P. Jucy, Dokl. Akad. Nauk SSSR **197**, 804 (1971).
[14] S. J. Ališauskas, Sov. J. Part. Nucl. **14**, 563 (1983).
[15] T. A. Brody, M. Moshinsky, and I. Renero, J. Math. Phys. **6**, 1540 (1965).
[16] M. Moshinsky, J. Math. Phys. **4**, 1128 (1963); Rev. Mod. Phys. **34**, 813 (1962).
[17] J. R. Derome and W. T. Sharp, J. Math. Phys. **7**, 612 (1966).
[18] J. R. Derome, J. Math. Phys. **8**, 714 (1967).
[19] M. Resnikoff, J. Math. Phys. **8**, 63 (1967).

[20] Z. Pluhař, Yu F. Smirnov, and V. N. Tolstoy, J. Phys. A **19**, 21 (1986).
[21] Z. Pluhař, L. J. Weigert, and P. Holan, J. Phys. A **19**, 29 (1986).
[22] K. T. Hecht, Nucl. Phys. **62**, 1 (1965).
[23] A. U. Klimyk and A. M. Gavrilik, J. Math. Phys. **20**, 1624 (1979).
[24] R. Le Blanc and D. J. Rowe, J. Phys. A **19**, 2913 (1986).
[25] S. J. Ališauskas, J. Math. Phys. **29**, 2351 (1988).
[26] S. J. Ališauskas, J. Math. Phys. **31**, 1325 (1990).
[27] S. J. Ališauskas, J. Math. Phys. **33**, 1983 (1992).
[28] S. J. Ališauskas, J. Phys. A **29**, 2687 (1996).
[29] P. Kramer, Z. Phys. **216**, 68 (1968); **205**, 181 (1967).
[30] J. Q. Chen, P. N. Wang, Z. M. Lü, and X. B. Wu, *Tables of the CG, Racah, and Subduction Coefficients of SU(n) Groups* (World Scientific, Singapore, 1987).
[31] F. Pan and J. Q. Chen, J. Phys. A **26**, 4299 (1993); J. Math. Phys. **34**, 4305 (1993); **34**, 4316 (1993).
[32] J. P. Draayer and Y. Akiyama, J. Math. Phys. **14**, 1904 (1973).
[33] Y. Akiyama and J. P. Draayer, Comput. Phys. Commun. **5**, 405 (1973).
[34] T. A. Kaeding, Comput. Phys. Commun. **85**, 82 (1995).
[35] T. A. Kaeding and H. T. Williams, Comput. Phys. Commun. **98**, 398 (1996).
[36] H. T. Williams, J. Math. Phys. **37**, 4187 (1996).
[37] J. S. Prakash and H. S. Sharatchandra, J. Math. Phys. **37**, 6530 (1996).
[38] L. A. Shelepin and V. P. Karasev, Sov. J. Nucl. Phys. **5**, 156 (1967).
[39] V. P. Karasev and L. A. Shelepin, Sov. J. Nucl. Phys. **7**, 678 (1968).
[40] D. E. Littlewood, *The Theory of Group Characters*, 2nd ed. (Claredom, Oxford, 1950).
[41] B. G. Wybourne, *Symmetry Principles and Atomic Spectroscopy* (Wiley, New York, 1970).
[42] S. J. Ališauskas and P. P. Kulish, J. Sov. Math. **35**, 2653 (1986).
[43] R. T. Sharp, J. Math. Phys. **16**, 2050 (1975).
[44] A. J. Macfarlane, L. O. O'raifertaigh, and P. S. Rao, J. Math. Phys. **8**, 536 (1967).
[45] G. H. Gadiyar and H. S. Sharatchandra, J. Phys. A **25**, L85 (1992).
[46] C. K. Chew and R. T. Sharp, Can. J. Phys. **44**, 2789 (1966).
[47] P. Jasselette, J. Phys. A **19**, 2261 (1986).
[48] M. F. O'Reilly, J. Math. Phys. **23**, 2022 (1982).
[49] B. Preziosi, A. Simoni, and B. Vitale, Nuovo Cimento **34**, 10 (1964).

Complementary group resolution of the SU(n) outer multiplicity problem. II. Recoupling approach for SU(3)⊃U(2) reduced Wigner coefficients

Feng Pan[a] and J. P. Draayer
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001

(Received 6 August 1997; accepted for publication 8 June 1998)

A general procedure for the derivation of SU(3)⊃U(2) reduced Wigner coefficients (RWCs) for the coupling $(\lambda_1\mu_1)\times(\lambda_2\mu_2)\downarrow(\lambda\mu)^{\eta}$, where η is the outer multiplicity label required in the decomposition, is proposed based on a recoupling approach that follows the complementary group technique for a resolution of the outer multiplicity of SU(n) introduced in Part (I) of this series. RWCs of SU(n) are not unique under a canonical resolution of the outer multiplicity; the transformation from one set to another are elements of SO(m), where m is the number of occurrences of the $(\lambda\mu)$ irrep in the decomposition $(\lambda_1\mu_1)\times(\lambda_2\mu_2)\downarrow(\lambda\mu)$. A special resolution of the multiplicity is identified that leads to a recursive procedure for the determination of RWCs. New features of these special RWCs and differences from those obtained with other choices are discussed. The method can be applied to the derivation of general SU(n) Wigner or RWCs. Algebraic expressions for another kind of RWCs, the so-called reduced auxiliary Wigner coefficients for SU(3)⊃U(2), are also obtained. © *1998 American Institute of Physics.* [S0022-2488(98)03210-1]

I. INTRODUCTION

Wigner coefficients (WCs) or reduced Wigner coefficients (RWCs) of SU(3) in the SU(3)⊃U(2) reduction have been discussed by many authors, including, for example, Biedenharn *et al.*[1–10] using a canonical unit tensor operator method, Moshinsky *et al.*[11–14] using an infinitesimal generator approach along with a complementary group prescription, and most recently Ališauskas *et al.* using a symmetric group approach[15–17] and paracanonical and pseudocanonical coupling schemes.[18–21] Certain classes of RWCs have been considered by Hecht,[22] Resnikoff,[23] Shelepin and Karasev,[24,25] Klimyk and Gavrilik,[26] Le Blanc and Rowe,[27] and others. Among these approaches, only the unit tensor operator method is canonical in the sense that it leads directly to the usual RWC orthogonality relations. Other methods for labeling the outer multiplicity are noncanonical, i.e., the RWCs produced are in general nonorthogonal with respect to the outer multiplicity label. In this case a Gram–Schmidt process must be adopted, which then depends upon an arbitrary choice of order for carrying out the orthogonalization. In this case a numerical algorithm is required such as the one used by Draayer and Akiyama[28,29] or Kaeding and Williams.[30–32] In Ref. 33, RWCs associated with the 27-plet operator, which has a maximum multiplicity of three, are studied in full detail. Recently, Parkash and Sharatchandra[34] worked out an algebraic formula for general WCs for SU(3) in the canonical basis. However, the final result, which involves a summation over 33 variables with some restrictions, requires a separate normalization procedure. This means the scheme is neither algebraically or numerically easy to implement.

In this paper, the complementary group representation proposed in (I) is used to determine RWCs of SU(3)⊃U(2) from known multiplicity-free RWCs through simple recoupling procedures. This new approach, in principle, is labeling scheme independent. For example, the RWCs of SU(3)⊃U(2) in the scheme proposed by Biedenhran *et al.* can be derived using the present method. However, values of the RWCs with different canonical outer multiplicity labeling

[a]On leave from Department of Physics, Liaoning Normal University, Dalian 116029, People's Republic of China.

 © 1998 American Institute of Physics

schemes will be related to one another through an SO(m) group transformation within the multiplicity space.

In Sec. II, the nonuniqueness of the concept of a canonical resolution of the multiplicity is addressed with the transformation from one set of RWCs to another shown to be elements of SO(m), where m is the number of occurrences of the resultant irrep in the decomposition $(\lambda_1\mu_1)\times(\lambda_2\mu_2)$. A special choice for a resolution of the outer multiplicity is then identified, one that yields a recursive procedure for the determination of RWCs with multiplicity. In Sec. III, a new recoupling procedure is proposed for the evaluation of the RWCs from known multiplicity-free RWCs of SU(n)$\supset$U($n-1$) given previously by Ališauskas *et al.*[15] New features of these RWCs, different from ones with other choices for the outer multiplicity, are discussed in Sec. IV. In Sec. V, an analytical expression for another kind of RWCs, the so-called reduced auxiliary WCs proposed by Brody, Moshinsky, and Renero[11] is derived using the recoupling approach. Further observations are given in Sec. VI.

II. RELATIONS AMONG DIFFERENT CHOICES FOR THE MULTIPLICITY LABELS

Let $|[\lambda]\xi,\rho\rangle\equiv|[\lambda_1][\lambda_2][\lambda]\xi,\rho\rangle$ be coupled basis vectors of U(n)$\times$U(n)$\downarrow$U(n), where ξ denotes multiplicity labels that are needed in the decomposition $[\lambda_1]\times[\lambda_2]\downarrow[\lambda]$, and ρ denotes sublabels of the resultant irrep $[\lambda]$. The completeness condition for $|[\lambda]\xi,\rho\rangle$ is

$$\sum_{\xi[\lambda]\rho}|[\lambda]\xi,\rho\rangle\langle[\lambda]\xi,\rho|=1. \tag{2.1}$$

The choice for the multiplicity labels is not unique. If there are m distinct ξ values, denoted $\xi=\xi_1,\xi_2,\ldots,\xi_m$, the transformation from one scheme to another is given by

$$|[\lambda]\eta,\rho\rangle=\sum_{\xi}y(\xi,\eta)|[\lambda]\xi,\rho\rangle, \tag{2.2}$$

where η denotes another set of multiplicity labels for U(n) and $y(\xi,\eta)$ is an element of an SU(m) transformation between the two. This can be seen by noting that

$$\begin{aligned}\sum_{\xi}y(\xi,\eta)y^*(\xi,\eta')&=\delta_{\eta\eta'},\\ \sum_{\eta}y(\xi,\eta)y^*(\xi',\eta)&=\delta_{\xi\xi'}.\end{aligned} \tag{2.3}$$

The $\{y(\xi,\eta)\}$ define a unitary transformation of the basic representation of SU(m). A unitary transformation $\mathbf{Y}\in$SU(m), where m is the number of occurrences of $[\lambda]$ in the decomposition $[\lambda_1]\times[\lambda_2]$, that transforms from one set of outer multiplicity labels to another always exists. Usually, the WCs of U(n) are taken to be real in which case the internal symmetry group for the transformation of the outer multiplicity labels can then be chosen to be SO(m). It is clear that the dimension of the transformation group is the number of occurrences of the resultant irrep $[\lambda]$ in the decomposition $[\lambda_1]\times[\lambda_2]\downarrow[\lambda]$, and the basis vectors of $[\lambda]$ remain orthonormal with respect to one another under the transformation. This freedom accounts for the fact that there exists different forms for RWCs or WCs with respect to the outer multiplicity labels. In what follows, RWCs are considered to be real.

Although the discussion will now be restricted to SU(3)$\supset$U(2), it can be extended easily to the general SU(n) case. The common notation for SU(3) irreps $(\lambda\mu)\equiv[\lambda+\mu,\mu]$ will also be used, where $[\lambda+\mu,\mu]$ is a usual two-rowed label corresponding to the irrep described by the two-rowed Young diagram with $\lambda+\mu$ boxes in the first row, and μ boxes in the second row. In (I) it was shown that there is only one multiplicity label needed in the decomposition $(\lambda_1\mu_1)\times(\lambda_2\mu_2)\downarrow[m_1m_2m_3]$. The multiplicity is labeled ξ with $\xi=\xi_1,\xi_2,\ldots,\xi_m$, where m is the number of occurrences of $[m_1m_2m_3]$ in the decomposition.

The key relation in our evaluation of SU(3) WCs or RWCs is the transformation

$$\sum_{\xi}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle y(\xi,\eta)=\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\eta[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle, \quad (2.4)$$

where ρ_1, ρ_2, and ρ are sublabels of SU(3) if

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle$$

is a WC or if a RWC then the corresponding U(2) labels, and the $y(\xi,\eta)$ are a special set of matrix elements of SO(m) chosen in a manner that will now be specified.

Assume

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle, \quad \xi=\xi_1,\xi_2,\ldots,\xi_m, \quad (2.5)$$

is a set of WCs satisfying the orthogonality relation

$$\sum_{\rho_1\rho_2}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi'[m_1'm_2'm_3']\\ \rho'\end{matrix}\right\rangle=\delta_{m_im_i'}\delta_{\xi\xi'}\delta_{\rho\rho'}, \quad (2.6a)$$

$$\sum_{\xi\rho m_i}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1' & \rho_2'\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle=\delta_{\rho_1\rho_1'}\delta_{\rho_2\rho_2'}, \quad (2.6b)$$

or RWCs satisfying

$$\sum_{\xi m_i}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1' & \rho_2'\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle=\delta_{\rho_1\rho_1'}\delta_{\rho_2\rho_2'}, \quad (2.6c)$$

$$\sum_{\rho_1\rho_2}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi'[m_1'm_2'm_3']\\ \rho\end{matrix}\right\rangle=\delta_{m_im_i'}\delta_{\xi\xi'}. \quad (2.6d)$$

According to the Schur–Weyl duality relation, the RWCs for SU(3)⊃U(2) given by (2.5) are also RWCs for SU(4)⊃U(3) for the same coupling with the same set of outer multiplicity labels $\{\xi_i\}$. Also, according to the complementary group technique outlined in (I), the complementary group for the SU(3) coupling is $\mathcal{U}(4)$. Therefore, consider the same coupling of $\mathcal{U}(4)$ in the special Gel'fand basis assigned according to the Littlewood rule, namely,

$$\left\langle \begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle. \quad (2.7)$$

Next recall that there is an $m\times m$ orthonormal matrix Y that transforms the RWCs or WCs between two sets of multiplicity labels (ξ) and ($\tilde{\eta}$), where the range of ξ and $\tilde{\eta}$ is the same as the sublabels η in the SU(3) subirrep $[m_1,m_2-\eta,m_3-\mu_2+\eta]$. Specifically, in (I) it is shown that the number of occurrences of $[m_1,m_2,m_3]$ in the Kronecker product $[\lambda_1+\mu_1,\mu_1]\times[\lambda_2+\mu_2,\mu_2]$ can be described exactly by η within the range

$$\eta_{\min}\leqslant\eta\leqslant\eta_{\max} \quad (2.8a)$$

where

$$\eta_{\max}=\min(m_1-\lambda_1-\mu_1,\mu_2,m_2-\mu_1,\lambda_2+\mu_2-m_3,\mu_1+\mu_2-m_3,m_2-m_3), \quad (2.8b)$$

$$\eta_{\min}=\max(0,\mu_2-m_3,m_2-\lambda_1-\mu_1).$$

Thus, we require

$$\sum_{\xi} y(\xi,\eta)\left\langle \begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle| \begin{matrix}\xi[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle$$
$$=\left\langle \begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle| \begin{matrix}\tilde{\eta}[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle', \tag{2.9}$$

where $\tilde{\eta},\ \eta=\eta_1,\ \eta_2,\ldots,\eta_m$, and the prime indicates a new RWC.

The RWCs,

$$\left\langle \frac{\xi}{\eta}\right\rangle \equiv \left\langle \begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle| \begin{matrix}\xi[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle \tag{2.10}$$

with fixed η can be regarded as a vector in $\mathbf{R}^m$ space.

$$\left\langle \frac{\xi}{\eta}\right\rangle, \quad \xi=\xi_1,\xi_2,\ldots,\xi_m. \tag{2.11}$$

Now consider a special transformation,

$$\mathbf{Y}\begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1}\right\rangle \\ \left\langle \frac{\eta_2}{\eta_1}\right\rangle \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_1}\right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1}\right\rangle' \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{2.12}$$

so all of the components of the old vectors $\langle \xi/\eta_1\rangle$ can be expressed by the following relation:

$$\left\langle \frac{\xi}{\eta_1}\right\rangle = y(\eta_1,\xi)\left\langle \frac{\eta_1}{\eta_1}\right\rangle. \tag{2.13}$$

The other $m-1$ vectors $\langle \xi/\eta_i\rangle$, $i=2,3,...,m$, undergo the same transformation. The complete transformation is specified by requiring

$$\mathbf{Y}\begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2}\right\rangle \\ \left\langle \frac{\eta_2}{\eta_2}\right\rangle \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_2}\right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2}\right\rangle' \\ \left\langle \frac{\eta_2}{\eta_2}\right\rangle' \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{Y}\begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3}\right\rangle \\ \left\langle \frac{\eta_2}{\eta_3}\right\rangle \\ \left\langle \frac{\eta_3}{\eta_3}\right\rangle \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_3}\right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3}\right\rangle' \\ \left\langle \frac{\eta_2}{\eta_3}\right\rangle' \\ \left\langle \frac{\eta_3}{\eta_3}\right\rangle' \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ldots, \tag{2.14}$$

$$\mathbf{Y}\begin{pmatrix} \left\langle \frac{\eta_1}{\eta_{m-1}} \right\rangle \\ \left\langle \frac{\eta_2}{\eta_{m-1}} \right\rangle \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_{m-1}} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_{m-1}} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_{m-1}} \right\rangle' \\ \vdots \\ \left\langle \frac{\eta_{m-1}}{\eta_{m-1}} \right\rangle' \\ 0 \end{pmatrix}, \quad \mathbf{Y}\begin{pmatrix} \left\langle \frac{\eta_1}{\eta_m} \right\rangle \\ \left\langle \frac{\eta_2}{\eta_m} \right\rangle \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_m} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_m} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_m} \right\rangle' \\ \vdots \\ \left\langle \frac{\eta_m}{\eta_m} \right\rangle' \end{pmatrix},$$

where the zero components (the special RWCs) have clearly been written out after the transformation. Other components are in general nonzero.

In the following, an example for the $m=3$ case is given to show that such a transformation is always possible. In the $m=3$ case, one can make the following special rotation A_1 around the third axis such that

$$\begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & \\ \sin\theta_1 & \cos\theta_1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle \\ \left\langle \frac{\eta_2}{\eta_1} \right\rangle \\ \left\langle \frac{\eta_3}{\eta_1} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle' \\ 0 \\ \left\langle \frac{\eta_3}{\eta_1} \right\rangle \end{pmatrix},$$

$$\begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & \\ \sin\theta_1 & \cos\theta_1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle \\ \left\langle \frac{\eta_3}{\eta_2} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle' \\ \left\langle \frac{\eta_3}{\eta_2} \right\rangle' \end{pmatrix}, \tag{2.15}$$

$$\begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & \\ \sin\theta_1 & \cos\theta_1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle \\ \left\langle \frac{\eta_3}{\eta_3} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle' \\ \left\langle \frac{\eta_3}{\eta_3} \right\rangle' \end{pmatrix}.$$

Then, make another rotation A_2 around the second axis with

$$\begin{pmatrix} \cos\theta_2 & & -\sin\theta_2 \\ 0 & 1 & 0 \\ \sin\theta_2 & & \cos\theta_2 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle' \\ 0 \\ \left\langle \frac{\eta_3}{\eta_1} \right\rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle'' \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} \cos\theta_2 & & -\sin\theta_2 \\ 0 & 1 & 0 \\ \sin\theta_2 & & \cos\theta_2 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle' \\ \left\langle \frac{\eta_3}{\eta_2} \right\rangle' \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle'' \end{pmatrix}, \tag{2.16}$$

$$\begin{pmatrix} \cos\theta_2 & & -\sin\theta_2 \\ 0 & 1 & 0 \\ \sin\theta_2 & & \cos\theta_2 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle' \\ \left\langle \frac{\eta_3}{\eta_3} \right\rangle' \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle'' \end{pmatrix}.$$

Finally, make a special rotation A_3 around the first axis with

$$\begin{pmatrix} 1 & & \\ & \cos\theta_3 & -\sin\theta_3 \\ & \sin\theta_3 & \cos\theta_3 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle'' \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_1} \right\rangle'' \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} 1 & & \\ & \cos\theta_3 & -\sin\theta_3 \\ & \sin\theta_3 & \cos\theta_3 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle'' \\ \left\langle \frac{\eta_3}{\eta_2} \right\rangle'' \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_2} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_2} \right\rangle''' \\ 0 \end{pmatrix}, \tag{2.17}$$

$$\begin{pmatrix} 1 & & \\ & \cos\theta_3 & -\sin\theta_3 \\ & \sin\theta_3 & \cos\theta_3 \end{pmatrix} \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle'' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle'' \\ \left\langle \frac{\eta_3}{\eta_3} \right\rangle'' \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\eta_1}{\eta_3} \right\rangle''' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle''' \\ \left\langle \frac{\eta_2}{\eta_3} \right\rangle''' \end{pmatrix}.$$

Hence, the special orthogonal matrix $\mathbf{Y}$ is

$$Y = A_3 A_2 A_1. \tag{2.18}$$

It should be clear that the angle θ_1 is fixed by $\langle \eta_1/\eta_1 \rangle$ and $\langle \eta_2/\eta_1 \rangle$, θ_2 by $\langle \eta_1/\eta_1 \rangle'$ and $\langle \eta_3/\eta_1 \rangle$, and θ_3 by $\langle \eta_2/\eta_2 \rangle''$ and $\langle \eta_3/\eta_2 \rangle''$. One can extend this procedure to the m-dimensional space and prove that there are indeed unique solutions for m rotational angles if the final form (2.14) is selected.

It is important to note that the choice of the rotation is not unique because there is an infinite number of solutions for RWCs or WCs with outer multiplicity, all related through SO(m) transformations. However, once the special rotation given in (2.14) is chosen, the resolution of the outer multiplicity is fixed; there remains no other arbitrariness in specifying the RWCs with multiplicity except for an over all phase factor. We will now show that the overall phase factor can be chosen as

$$\left\langle \frac{\eta_i}{\eta_i} \right\rangle \geqslant 0 \tag{2.19}$$

for $i=1,2,...,m$. The RWCs are then uniquely determined.

It is easy to prove that the orthogonality conditions for the WCs or RWCs given by (2.6) remain valid after the transformation for both the $\mathcal{U}(4)\supset\mathcal{U}(3)$ and SU(3)$\supset$U(2) cases. First of all note that the $\mathcal{U}(4)\supset\mathcal{U}(3)$ or SU(3)$\supset$U(2) RWCs all undergo the same transformation $\mathbf{Y}$. One can easily verify that the orthogonality conditions are valid for the $\mathcal{U}(4)\supset\mathcal{U}(3)$ case. Since, according to the Schur–Weyl duality relation, SU(3) WCs or SU(3)$\supset$U(2) RWCs are a subset of those of $\mathcal{U}(4)$ or $\mathcal{U}(4)\supset\mathcal{U}(3)$, the same conclusion applies to WCs of SU(3) or RWCs of SU(3)$\supset\mathcal{U}(2)$ as well.

Although the orthogonality conditions survive, other properties such as Biedneharn's $1\leftrightarrow 2$ exchange symmetry, is lost in this new scheme. Questions of this type will be considered further in Sec. IV.

Finally, we want to show what is actually achieved by the special rotation $\mathbf{Y}$. If the RWCs of $\mathcal{U}(4)\supset\mathcal{U}(3)$ in terms of an $m\times m$ matrix are arranged such that the columns are set by the outer multiplicity label $\tilde{\eta}=\eta_1,\ \eta_2,\ldots,\eta_m$ and the rows by the label η in the (irrep) $[m_1,m_2-\eta,m_3-\mu_2+\eta]$ for $\mathcal{U}(3)$, the RWCs have the following structure:

$$\left(\left\langle \frac{\tilde{\eta}}{\eta} \right\rangle\right)=\begin{pmatrix} \times & 0 & \cdots & \cdots & 0 \\ \times & \times & 0 & \cdots & 0 \\ \times & \times & \times & 0\cdots & 0 \\ \cdots & \cdots & & & \\ \times & \times & \cdots & \times & 0 \\ \times & \times & \cdots & \cdots & \times \end{pmatrix}, \tag{2.20}$$

i.e., one achieves a lower triangular structure for the RWCs with the multiplicity postulated by Braunschweig.[35] This development shows that it is always possible to choose such a structure. Hecht, in Ref. 22, argued that one can resolve the SU(3) multiplicity problem simply by requiring a similar lower triangular structure. Le Blanc and Rowe also pointed out that such a resolution becomes *ipso facto* equivalent to a canonical labeling scheme.[27] Actually, one can also choose an upper triangular structure for the RWCs. (Under row-column interchange there are clearly a total of $m(m-1)/2$ ''equivalent'' choices.) The structure of the RWCs is not unique, it depends on what kind of special transformation is chosen.

Now recall Biedenharn's definition for a canonical resolution to the outer multiplicity problem. In Ref. 6, ''canonical'' implies there no free choices involved in the resolution of the multiplicity. This definition must be understood within the context of the procedure introduced for resolving the multiplicity problem. As shown here, there is not a single ''canonical'' resolution of the multiplicity; ''canonical'' has to be understood within the context of the methodology offered for a resolution of the multiplicity. The same applies to the definition of a canonical basis for U(n). A ''canonical'' construction has to be understood in terms of an equivalent class associated with a particular U(1), out of the set of all equivalent U(1) subgroups, at each stage of the decomposition.

The special transformation given by (2.14) makes it possible to evaluate RWCs of SU(3) in any basis using only recoupling procedures. In the following RWCs for SU(3) in the canonical SU(3)$\supset$U(2) basis will be considered; RWCs for SU(3) in the noncanonical SU(3)$\supset$O(3) basis will be discussed elsewhere.

III. RECOUPLING APPROACH FOR SU(3)$\supset$U(2) RWCs

In this section, a new procedure for evaluating RWCs with multiplicity for SU(3)$\supset$U(2), or SU(n)$\supset$U($n-1$) in general, is introduced. It uses known results, including the analytic expression for RWCs of U(n)$\supset$U($n-1$) with one symmetric irrep[15] given by Ališauskas *et al.*, and a special choice for the multiplicity transformation. U(3)$\supset$U(2) RWCs of the type given by Ališauskas *et al.* can also be found in the work of Chacon *et al.*,[5] who derived their results using the

canonical unit tensor operator method proposed by Biedenharn *et al.* The result of Chacon and that of Ališauskas *et al.* are the same, including the choice of phase. To make every step clear, this section is divided into several subsections.

A. A recoupling approach

Using analytic expressions for multiplicity free RWCs given, for example, in Ref. 15, one can construct an expression for the product of general $\mathcal{U}(4)$ or SU(3) RWCs and their respective special recoupling coefficients,[36,37]

$$\begin{aligned}\sum_{\xi} U_{\xi}\begin{pmatrix}(\lambda_{\iota}\mu_1) & [\lambda_2+\mu_2\ 0] & [\overline{m}]\\ [\mu_2 0] & [m_1m_2m_3] & (\lambda_2\mu_2)\end{pmatrix}\left\langle\begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\\ =\sum\left\langle\begin{matrix}(\lambda_1\mu_1) & [\lambda_2+\mu_2 0]\\ \rho_1 & \rho_2'\end{matrix}\middle|\begin{matrix}[\overline{m}]\\ \overline{\rho}\end{matrix}\right\rangle\left\langle\begin{matrix}[\overline{m}] & [\mu_2 0]\\ \overline{\rho} & \rho_2''\end{matrix}\middle|\begin{matrix}[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle\\ \times\left\langle\begin{matrix}[\lambda_2+\mu_2 0] & [\mu_2 0]\\ \rho_2' & \rho_2''\end{matrix}\middle|\begin{matrix}(\lambda_2\mu_2)\\ \rho_2\end{matrix}\right\rangle.\end{aligned} \tag{3.1}$$

In this equation, U is the unitary form of a Racah coefficient for $\mathcal{U}(4)$ or SU(3), respectively, and the sum on the rhs is over ρ_2', ρ_2'', and $\overline{\rho}$.

It is convenient to introduce the following abbreviated notation:

$$U_{\xi}([\overline{m}])\equiv U_{\xi}\begin{pmatrix}(\lambda_1\mu_1) & [\lambda_2+\mu_2\ 0] & [\overline{m}]\\ [\mu_2 0] & [m_1m_2m_3] & (\lambda_2\mu_2)\end{pmatrix}$$

for the Racah coefficients,

$$\left\langle\frac{\tilde{\eta}}{\eta}\right\rangle\equiv\left\langle\begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle|\begin{matrix}\tilde{\eta}[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle$$

for a special set of $\mathcal{U}(4)\supset$U(3) RWCs,

$$\left\langle\frac{\tilde{\eta}}{\rho_1\rho_2\rho}\right\rangle\equiv\left\langle\begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ \rho_1 & \rho_2\end{matrix}\middle|\begin{matrix}\tilde{\eta}[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle$$

for SU(3) WCs, or SU(3)$\supset$U(2) RWCs if ρ_1, ρ_2, and ρ are the corresponding U(2) labels,

$$\begin{pmatrix}\eta'\\ \eta\end{pmatrix}\equiv\begin{pmatrix}[\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2 0] & [m_1m_2-\eta' m_3-\mu_2+\eta']\\ [\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2\mu_2] & [m_1m_2m_3]\\ [\lambda_1+\mu_1] & [\lambda_2+\mu_2 0] & [m_1m_2-\eta m_3-\mu_2+\eta]\end{pmatrix}$$

for $(\mathcal{U}4\supset\mathcal{U}(3))\star(\mathcal{U}(4)\supset\mathcal{U}(3))$ reduced coupling coefficients, and

$$\begin{pmatrix}\eta'\\ \rho_1\rho_2\rho\end{pmatrix}\equiv\begin{pmatrix}[\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2 0] & [m_1m_2-\eta' m_3-\mu_2+\eta']\\ [\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2\mu_2] & [m_1m_2m_3]\\ \rho_1 & \rho_2 & \rho\end{pmatrix}$$

for $(\mathcal{U}(4)\supset\mathcal{U}(3))\star$SU(3) coupling coefficients.

For the $\mathcal{U}(4)\supset\mathcal{U}(3)$ case, only the simpler case

$$\sum_{\xi} U_{\xi}([\overline{m}])\left\langle\frac{\xi}{\eta}\right\rangle=G([\overline{m}],\eta), \tag{3.2}$$

needs to be considered, where

$$G([\bar{m}],\eta)=\sum_{p_1p_2[\bar{m}']}\left\langle\begin{matrix}[\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2 0] & \\ [\lambda_1+\mu_1\mu_1] & p_1 & [\bar{m}'] \\ [\lambda_1+\mu_1\mu_1] & 0 & [\lambda_1+\mu_1\mu_1]\end{matrix}\right\rangle \times\left\langle\begin{matrix}[\bar{m}] & [\mu_2 0] & \\ [\bar{m}'] & p_2 & [m_1m_2m_3] \\ [\lambda_1+\mu_1\mu_1] & 0 & [m_1m_2-\eta m_3-\mu_2+\eta] \\ & & [\lambda_1+\mu_1\mu_1]\end{matrix}\right\rangle \times\left\langle\begin{matrix}[\lambda_2+\mu_2 0] & [\mu_2 0] & [\lambda_2+\mu_2\mu_2] \\ p_1 & p_2 & [\lambda_2+\mu_2 0]\end{matrix}\right\rangle \times\left\langle\left\langle\begin{matrix}[\lambda_1+\mu_1\mu_1] & [\lambda_2+\mu_2 0] & \\ [\lambda_1+\mu_1\mu_1] & 0 & [m_1m_2-\eta m_3-\mu_2+\eta] \\ & & [\lambda_1+\mu_1\mu_1]\end{matrix}\right\rangle\right\rangle^{-1}. \tag{3.3}$$

While for SU(3), the following expression is important:

$$\sum_{\xi} U_{\xi}([\bar{m}])\left\langle\begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) & \\ \rho_1 & \rho_2 & \xi[m_1m_2m_3] \\ & & \rho\end{matrix}\right\rangle=G([\bar{m}],\rho_1\rho_2\rho), \tag{3.4}$$

where

$$\left\langle\begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) & \\ \rho_1 & \rho_2 & \xi[m_1m_2m_3] \\ & & \rho\end{matrix}\right\rangle$$

is the WCs of SU(3), and $G([\bar{m}],\rho_1\rho_2\rho)$ is the expression given by the rhs of (3.1).

Using these G polynomials, one can easily construct the following coupling coefficients:

$$\left(\begin{matrix}\eta' \\ \eta\end{matrix}\right)=\sum_{[\bar{m}]} G([\bar{m}],\eta)G([\bar{m}],\eta'), \tag{3.5}$$

which is a symmetric function with respect to the labels η, η'. What (3.5) gives is nothing but reduced coupling coefficient of $\mathcal{U}(4)\star\mathcal{U}(4)$,

$$\left\langle\left(\begin{matrix}[\lambda_1+\mu_1\mu_1] \\ [\lambda_1+\mu_1\mu_1] \\ [\lambda_1+\mu_1\mu_1]\end{matrix}\right);\left(\begin{matrix}[\lambda_2+\mu_2 0] \\ [\lambda_2+\mu_2\mu_2] \\ [\lambda_2+\mu_2 0]\end{matrix}\right)\middle|\left(\begin{matrix}[m_1m_2-\eta' m_3-\mu_2+\eta'] \\ [m_1m_2m_3] \\ [m_1m_2-\eta m_3-\mu_2+\eta]\end{matrix}\right)\right\rangle. \tag{3.6}$$

Similarly,

$$\left(\begin{matrix}\eta \\ \rho_1\rho_2\rho\end{matrix}\right)=\sum_{[\bar{m}]} G([\bar{m}],\eta)G([\bar{m}],\rho_1\rho_2\rho), \tag{3.7}$$

which is the following $\mathcal{U}(4)\star$SU(3) coupling coefficient

$$\left\langle\left(\begin{matrix}[\lambda_1+\mu_1\mu_1] \\ [\lambda_1+\mu_1\mu_1] \\ \rho_1\end{matrix}\right);\left(\begin{matrix}[\lambda_2+\mu_2 0] \\ [\lambda_2+\mu_2\mu_2] \\ \rho_2\end{matrix}\right)\middle|\left(\begin{matrix}[m_1m_2-\eta m_3-\mu_2+\eta] \\ [m_1m_2m_3] \\ \rho\end{matrix}\right)\right\rangle. \tag{3.8}$$

As pointed out by Biedenharn *et al.*,[2] there are close relations between RWCs of U(n) and the coupling coefficients of U(n)$\star$U(n). They have proved that such coupling coefficients can be expressed in terms of the product of two U(n) WCs with summation over the outer multiplicity labels. One can directly verify by using the recoupling technique that

$$\left(\begin{matrix}\eta' \\ \eta\end{matrix}\right)=\sum_{[\bar{m}]} G([\bar{m}],\eta)G([\bar{m}],\eta')=\sum_{\xi}\left\langle\frac{\xi}{\eta}\right\rangle\left\langle\frac{\xi}{\eta'}\right\rangle \tag{3.9}$$

and

$$\left(\begin{matrix}\eta\\ \rho_1\rho_2\rho\end{matrix}\right)=\sum_{[\bar{m}]} G([\bar{m}],\eta)G([\bar{m}],\rho_1\rho_2\rho)=\sum_{\xi}\left\langle\frac{\xi}{\eta}\right\rangle\left\langle\frac{\xi}{\rho_1\rho_2\rho_3}\right\rangle. \tag{3.10}$$

B. Explicit expressions of *G* polynomials

Using the recoupling technique, it is possible to evaluate explicit expressions for the G polynomials defined in the above subsection from the known multiplicity-free RWCs of $\mathrm{U}(n)\supset\mathrm{U}(n-1)$ given by Ališauskas *et al.*,

$$\begin{aligned}G(k_1k_2,\eta)=&\left[\frac{(m_3-\mu_2+\eta)!^2(m_2-\eta+1)!^2(m_1+2)!\lambda_1!(\lambda-\mu_1-2k_1-k_2+2)(\mu_1+k_1-k_2+1)}{(\lambda_2+\mu_2)!^2(\mu_1+\mu_2-m_3-\eta)!(\lambda_2+\mu_2+1)}\right.\\
&\times\frac{(\lambda-k_1-k_2-\lambda_1-\mu_1)!(\lambda-k_1-k_2-\mu_1+1)!(m_1-m_2+1)!(m_1-m_3+2)!(m_2-m_3+1)}{(\lambda_1-k_1)!(\lambda_1+\mu_1-k_2+1)!(\mu_1-k_2)!(\mu_2-\eta)!(\lambda_1+\mu_1-m_2+\eta)!(\lambda_1+\mu_1+\mu_2-m_3-\eta+1)!}\\
&\times\frac{(\lambda-k_1-k_2-m_2)!(\lambda-k_1-k_2-m_3+1)!(\lambda-k_1-k_2+2)!(\mu_1+k_1-m_3)!(\mu_1+k_1+1)!k_1!k_2!}{(m_1-\lambda_1+k_1+k_2)!(m_1-\mu_1-k_1+1)!(m_1-k_2+2)!(m_2-\mu_1-k_1)!(m_2+1)!m_3!(m_3-k_2)!}\\
&\left.\times\frac{(\lambda_2+1)!\mu_2!(m_2-m_3-\eta)!(m_1-\lambda_1-\mu_1)!(m_1-\mu_1+1)!(m_2-\mu_1-\eta)!}{(m_1-m_2+\eta+1)!(m_1-m_3+\mu_2+2)!\,\eta!(m_2-m_3+\mu_2-\eta+1)!}\right]^{1/2}\\
&\times\sum_{p=0}^{\mu_2}\sum_{m_{23}=\min(m_{23})}^{\max(m_{23})}\sum_{m_{33}=\min(m_{33})}^{\max(m_{33})}(-)^p\frac{p!(\lambda_1-m_{23})!(\lambda_1+\mu_1-m_{33}+1)!(\mu_1-m_{33})!(\lambda_2+\mu_2-p)!}{(\lambda-p-m_{23}-m_{33}-\lambda_1-\mu_1)!(\lambda-p-m_{23}-m_{33}-\mu_1+1)!m_{23}!}\\
&\times\frac{(\lambda-p-2m_{23}-m_{33}-\mu_1+1)(\lambda-p-m_{23}-2m_{33}+2)(m_{23}-m_{33}+\mu_1+1)}{(\lambda-p-m_{23}-m_{33}+2)!(\mu_1+m_{23}+1)!m_{33}!}\\
&\times F_3\left(\begin{matrix}[\lambda-k_1-k_2,\mu_1+k_1,k_2]\\ [\lambda-p-m_{23}-m_{33},m_{23}+\mu_1,m_{33}]\end{matrix}\;;\;\begin{matrix}[m_1m_2m_3]\\ [m_1m_2-\eta\ m_3-\mu_2+\eta]\end{matrix}\right),\end{aligned} \tag{3.11}$$

where

$$\lambda=\lambda_1+\mu_1+\lambda_2+\mu_2, \tag{3.12a}$$

$$[\bar{m}]=[\lambda-k_1-k_2,\mu_1+k_1,k_2]. \tag{3.12b}$$

Boundary conditions for m_{23} and m_{33} in these expressions can be obtained by using the Littlewood rule for the Kronecker products involved and the betweenness conditions for the decomposition $\mathrm{U}(n)\downarrow\mathrm{U}(n-1)$,

$$\begin{aligned}\max(m_{23})&=\min(\lambda_1,k_1,\lambda_2+\mu_2-p,m_2-\eta-\mu_1),\\ \min(m_{23})&=\max(0,m_2-\eta-\mu_1-p,m_3-\mu_1-\mu_2+\eta),\end{aligned} \tag{13.12c}$$

for fixed p, while

$$\begin{aligned}\max(m_{33})=&\min(\lambda-p-m_{23}-m_2+\eta,m_3-\mu_2+\eta,\mu_1,\\ &\lambda_2+\mu_2-p-m_{23},\ \lambda-p-m_{23}-\mu_1-k_1,\ \lambda-p-m_{23}-\lambda_1-\mu_1),\\ \min(m_{33})=&\max(0,m_2+m_3+\mu_1-\mu_2-p+m_{23},k_1+k_2-p-m_{23},k_2)\end{aligned} \tag{3.12d}$$

for fixed p and m_{23}. Furthermore,

$$
F_3\left(\begin{matrix}[h_1h_2h_3] \\ [q_1q_2q_3]\end{matrix} \quad ; \quad \begin{matrix}[m_1m_2m_3] \\ [n_1n_2n_3]\end{matrix}\right)
$$

$$
\begin{aligned}
&=\sum_{xyz}(-)^{x+y+z-q_1-q_2-q_3} \\
&\times\frac{(x-y+1)(x-z+2)(y-z+1)(x-h_3+1)!}{(h_2-y)!(n_2-z+1)!(h_3-z)!(h_1-z+2)!(h_1-y+1)!} \\
&\times\frac{(q_1-y)!(q_1-z+1)!(q_2-z)!(x-n_2)!(x-n_3+1)!(y-n_3)!(y-h_3)!}{(x-q_1)!(x-q_2+1)!(x-q_3+2)!(y-q_2)!(y-q_3+1)!(z-q_3)!(n_1-x)!(n_1-y+1)!(n_1-z+2)!} \\
&\times\frac{(m_1-x)!(m_1-y+1)!(m_1-z+2)!(m_2-y)!(m_2-z+1)!(m_3-z)!(x-h_2)!}{(n_2-y)!(n_2-z+1)!(n_3-z)!(x-m_2)!(x-m_3+1)!(y-m_3)!(h_1-x)!}.
\end{aligned}
\tag{3.13}
$$

Similarly,

$$
\begin{aligned}
&G(k_1k_2;[m_{12}m_{22}][m'_{12}m'_{22}][m''_{12}m''_{22}]) \\
&=\sum_{\xi}U_{\xi}(k_1k_2)\left\langle\begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ [m_{12}m_{22}] & [m'_{12}m'_{22}]\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3] \\ [m''_{12}m''_{22}]\end{matrix}\right\rangle \\
&=\Bigg[\frac{(\lambda_2+1)!(\mu_2-m'_{22})!(m_1-m_2+1)(m_1-m_3+2)(m_2-m_3+1)(m''_{12}-m_2)!}{(\lambda_2+\mu_2-m'_{12})!(\lambda_2+\mu_2-m'_{22}+1)!(m'_{12}-\mu_2)!(m_2-k_2+1)!(m_1-m''_{12})!(m_2-m''_{22})!} \\
&\times\frac{(\lambda-k_1-k_2-m_2)!(\mu_1+k_1-m_3)!(\lambda-k_1-k_2-m_3+1)!(m''_{22}-m_3)!(m''_{12}-m_3+1)!}{(m_1-\lambda+k_1+k_2)!(m_2-\mu_1-k_1)!(m_3-k_2)!(m_1-\mu_1-k_1+1)!(m_1-k_2+2)!} \\
&\times\frac{(\mu_1-k_2)!(\lambda_1+\mu_1-k_2+1)!(\lambda_1+\mu_1-m_{12})!(\mu_1-m_{22})!(\lambda_1+\mu_1-m_{22}+1)!}{(\mu_1+k_1+1)!(m_{12}-\mu_1)!m_{22}!(m_{12}+1)!(m_1-m''_{22}+1)!}\Bigg]^{1/2} \\
&\times\sum_{p=0}^{\mu_2}\sum_{q=\min(q)}^{\max(q)}(-1)^{p-m'_{22}}(\lambda_2+\mu_2+p-m'_{12}-m'_{22})! \\
&\times((m'_{12}-p)!(m''_{12}+m''_{22}-p-q-m_{12})!(q-m_{22})!)^{1/2} \\
&\times\left(\frac{(m''_{12}+m''_{22}-p-q-m_{22}+1)!(m''_{12}+m''_{22}-p-2q+1)(p+q-m''_{22})!(m''_{22}-q+1)!}{(p-m'_{22})!(m''_{12}-p-q)!(m''_{12}+m''_{22}-p-q-\mu_1-k_1)!(m_{12}-q)!}\right)^{1/2} \\
&\times U\left(\begin{matrix}[m_{12}m_{22}] & [m'_{12}+m'_{22}-p0] & [m''_{12}+m''_{22}-p-q,q] \\ [p0] & [m''_{12}m''_{22}] & [m'_{12}m'_{22}]\end{matrix}\right) \\
&\times F_2\left(\begin{matrix}[\lambda-k_1-k_2,\mu_1+k_1,k_2] & [m_1m_2m_3] \\ [m''_{12}+m''_{22}-p-q,q] & [m''_{12}m''_{22}]\end{matrix}\right) \\
&\times F_2\left(\begin{matrix}[\lambda_1+\mu_1\mu_1] & [\lambda-k_1-k_2,\mu_1+k_1,k_2] \\ [m_{12}+m_{22}0] & [m''_{12}+m''_{22}-p-q,q]\end{matrix}\right),
\end{aligned}
\tag{3.14}
$$

where

$$
\begin{aligned}
\max(q)&=\min(\mu_1+k_1,m''_{12}+m''_{22}-p-\mu_1-k_1), \\
\min(q)&=\max(k_2,m''_{12}+m''_{22}-p-\lambda+k_1+k_2),
\end{aligned}
\tag{3.15a}
$$

U is an SU(2) Racah coefficient in unitary form, and

$$F_2\begin{pmatrix}[h_1h_2h_3] & [m_1m_2m_3]\\ [q_1q_2] & [n_1n_2]\end{pmatrix}=\sum_{xy}(-)^{x+y-q_1-q_2}(x-y+1)$$
$$\times\frac{(x-n_2)!(m_1-x)!(x-h_2)!(x-h_3+1)!}{(x-q_1)!(x-q_2+1)!(n_1-x)!(x-m_2)!(x-m_3+1)!(h_1-x)!}$$
$$\times\frac{(q_1-y)!(m_1-y+1)!(m_2-y)!(y-h_3)!}{(y-q_2)!(n_2-y)!(n_1-y+1)!(y-m_3)!(h_1-y+1)!(h_2-y)!}. \tag{3.15b}$$

Using these expressions, (3.9) and (3.10) can be expressed explicitly as

$$\begin{pmatrix}\eta'\\ \eta\end{pmatrix}=\sum_{k_1=\min(k_1)}^{\max(k_1)}\sum_{k_2=\min(k_2)}^{\max(k_2)}G(k_1,k_2,\eta)G(k_1,k_2,\eta'), \tag{3.16a}$$

$$\begin{pmatrix}\eta\\ \rho_1\rho_2\rho\end{pmatrix}=\sum_{k_1=\min(k_1)}^{\max(k_1)}\sum_{k_2=\min(k_2)}^{\max(k_2)}G(k_1,k_2,\eta)G(k_1,k_2,\rho_1\rho_2\rho), \tag{3.16b}$$

where

$$\begin{aligned}\max(k_1)&=\min(m_2-\mu_1,\lambda_1,\lambda_2+\mu_2),\\ \min(k_1)&=\max(m_3-\mu_1,m_2-\mu_1-\mu_2,0),\\ \max(k_2)&=\min(\mu_1,\lambda_2+\mu_2-k_1,\lambda_2+\mu_2-k_1+\lambda_1+\mu_1-m_2),\\ \min(k_2)&=\max(m_2+m_3-\mu_1-\mu_2-k_1).\end{aligned} \tag{3.16c}$$

C. Recursive procedure for the evaluation of all SU(3)⊃U(2) RWCs

The G polynomials can be used to construct $\mathcal{U}(4)\star\mathcal{U}(4)$ as well as $\mathcal{U}(4)\star\mathrm{SU}(3)$ reduced coupling coefficients as given by (3.16a) and (3.16b), and these, in turn, can be used to evaluate all RWCs of SU(3)⊃U(2) after the special transformation given in Sec. II. It can be proved that

$$\sum_{\xi}\left\langle\frac{\xi}{\eta}\right\rangle\left\langle\frac{\xi}{\eta}\right\rangle\neq 0 \tag{3.17}$$

for any η. First, if there exists an η such that

$$\sum_{\xi}\left\langle\frac{\xi}{\eta}\right\rangle\left\langle\frac{\xi}{\eta}\right\rangle=0, \tag{3.18}$$

then

$$\left\langle\frac{\xi}{\eta}\right\rangle=0 \tag{3.19}$$

for $\xi=\eta_1,\eta_2,\ldots,\eta_m$ because the condition (3.18) can be written as $\boldsymbol{\eta}\cdot\boldsymbol{\eta}=0$ in $\mathbf{R}^m$ space, which then implies its components are all zero, i.e.,

$$\left\langle\begin{matrix}[\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2]\\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0]\end{matrix}\middle|\begin{matrix}\xi[m_1m_2m_3]\\ [m_1,m_2-\eta,m_3-\mu_2+\eta]\end{matrix}\right\rangle=0 \tag{3.20}$$

for any multiplicity label ξ and fixed η. However, using the build-up principle, one can deduce that (3.20) is valid if and only if

$$G([\bar{m}],\eta)=0, \tag{3.21}$$

where the expression of $G([\overline{m}],\eta)$ is given by (3.3). In this case, the unitarity condition for $U(n)\supset U(n-1)$ RWCs to prove that (3.21) is satisfied only when one type of the WCs or RWCs involved are zero. However, one can verify from the explicit expressions for these WCs or RWCs given by Ališauskas *et al.*[15] that (3.21) is satisfied only when the irreps involved in the coupling do not satisfy the Littlewood rule for the Kronecker products involved or betweenness conditions for the decomposition. But these are all trivial cases and need not be considered. Hence, (3.17) is valid in general for all nontrivial cases.

Hence, after applying the special transformation given in Sec. II, the rhs of (3.9) has, in general, k nonzero terms in the summation when a smaller η which is a label in $[m_1 m_2-\eta\ m_3 -\mu_2+\eta]$ equals to η_k, namely

$$\left\langle \begin{matrix}\eta_k\\ \eta_l\end{matrix}\right\rangle=\sum_{\xi=\eta_1}^{\eta_k}\left\langle \frac{\xi}{\eta_k}\right\rangle\left\langle \frac{\xi}{\eta_l}\right\rangle,\quad \text{for } k\leqslant l,\ \ k\leqslant m. \tag{3.22}$$

Using (3.22), it can be shown that

$$\left\langle \frac{\eta_k}{\eta_k}\right\rangle=+\left[\left\langle \begin{matrix}\eta_k\\ \eta_k\end{matrix}\right\rangle-\sum_{\xi=\eta_1}^{\eta_{k-1}}\left\langle \frac{\xi}{\eta_k}\right\rangle^2\right]^{1/2}, \tag{3.23}$$

where the sign is chosen positive for any k, fixing the overall phase. Then,

$$\left\langle \frac{\eta_k}{\eta_k}\right\rangle\left\langle \frac{\eta_k}{\eta_l}\right\rangle=\left\langle \begin{matrix}\eta_k\\ \eta_l\end{matrix}\right\rangle-\sum_{\xi=\eta_1}^{\eta_{k-1}}\left\langle \frac{\xi}{\eta_k}\right\rangle\left\langle \frac{\xi}{\eta_l}\right\rangle \tag{3.24}$$

for $l>k$. It can also be proven that (3.23) cannot be zero. First, $\langle \eta_k/\eta_k\rangle^2$ is a square of the component of the vector $\boldsymbol{\eta}_k$ in $\mathbf{R}^{\mathrm{m}}$. Therefore,

$$\left\langle \frac{\eta_k}{\eta_k}\right\rangle^2\geqslant 0,\quad \text{for } k\leqslant m. \tag{3.25a}$$

Second, if (3.23) is zero, (3.24) should also be zero for any l values. This only occurs when the multiplicity equals to $k-1$. However, an initial assumption was that $m\geqslant k$. Hence, (3.24) cannot be zero for $k\leqslant m$. Hence,

$$\left\langle \frac{\eta_k}{\eta_k}\right\rangle^2>0,\quad \text{for } k\leqslant m. \tag{3.25b}$$

Thus, (3.23) and (3.24) allow us to calculate all of the special RWCs of $\mathcal{U}(4)\supset\mathcal{U}(3)$ recursively,

$$\left\langle \frac{\eta_1}{\eta_1}\right\rangle=\left\langle \begin{matrix}\eta_1\\ \eta_1\end{matrix}\right\rangle^{1/2},$$

$$\left\langle \frac{\eta_1}{\eta}\right\rangle=\left\langle \begin{matrix}\eta_1\\ \eta_1\end{matrix}\right\rangle^{-1/2}\left\langle \begin{matrix}\eta_1\\ \eta\end{matrix}\right\rangle, \tag{3.26a}$$

$$\left\langle \frac{\eta_2}{\eta}\right\rangle=\left[\left\langle \begin{matrix}\eta_2\\ \eta\end{matrix}\right\rangle-\left\langle \begin{matrix}\eta_1\\ \eta_1\end{matrix}\right\rangle^{-1}\left\langle \begin{matrix}\eta_1\\ \eta_2\end{matrix}\right\rangle\left\langle \begin{matrix}\eta_1\\ \eta\end{matrix}\right\rangle\right]\Big/\left\langle \frac{\eta_2}{\eta_2}\right\rangle, \tag{3.26b}$$

where

$$\left\langle \frac{\eta_2}{\eta_2}\right\rangle=\left[\left\langle \begin{matrix}\eta_2\\ \eta_2\end{matrix}\right\rangle-\left\langle \begin{matrix}\eta_1\\ \eta_1\end{matrix}\right\rangle^{-1}\left\langle \begin{matrix}\eta_1\\ \eta_2\end{matrix}\right\rangle^2\right]^{1/2}. \tag{3.26c}$$

$$\cdots\ \cdots$$

Once the $\langle(\eta_{k-1})/\eta\rangle$ for any η are known from the $(k-1)$th step, $\langle\eta_k/\eta\rangle$ can be obtained by using (3.23) and (3.24). Thus, one obtains all the special RWCs of $\mathcal{U}(4)\supset\mathcal{U}(3)$, which are important in determining the RWCs of SU(3)⊃U(2).

Using (3.10) and known special RWCs of $\mathcal{U}(4)\supset\mathcal{U}(3)$, SU(3) WCs or SU(3)⊃U(2) RWCs can be determined from

$$\left\langle\frac{\eta_1}{\rho_1\rho_2\rho}\right\rangle=\left\langle\frac{\eta_1}{\eta_1}\right\rangle^{-1}\begin{pmatrix}\eta_1\\ \rho_1\rho_2\rho\end{pmatrix},$$

$$\left\langle\frac{\eta_2}{\rho_1\rho_2\rho}\right\rangle=\left[\begin{pmatrix}\eta_2\\ \rho_1\rho_2\rho\end{pmatrix}-\left\langle\frac{\eta_1}{\eta_1}\right\rangle^{-1}\left\langle\frac{\eta_1}{\eta_2}\right\rangle\begin{pmatrix}\eta_1\\ \rho_1\rho_2\rho\end{pmatrix}\right]\Big/\left\langle\frac{\eta_2}{\eta_2}\right\rangle, \tag{3.27}$$

$$\cdots\ \cdots$$

$$\left\langle\frac{\eta_k}{\rho_1\rho_2\rho}\right\rangle=\left[\begin{pmatrix}\eta_k\\ \rho_1\rho_2\rho\end{pmatrix}-\sum_{\eta'=\eta_1}^{\eta_{k-1}}\left\langle\frac{\eta'}{\rho_1\rho_2\rho}\right\rangle\left\langle\frac{\eta'}{\eta_k}\right\rangle\right]\Big/\left\langle\frac{\eta_k}{\eta_k}\right\rangle,\quad\text{for } k\leqslant m. \tag{3.28}$$

It should be noted that (3.28) determines not only SU(3) WCs in the canonical basis, and SU(3)⊃U(2) RWCs, but also SU(4)⊃U(3) RWCs for the same coupling.

D. Some algebraic expressions

In this subsection, some algebraic expressions for SU(3)⊃U(2) RWCs and related Racah coefficients of SU(3) are worked out. Starting from $\eta=\eta_1$, and using (3.23) and (3.24), one obtains

$$\left\langle\frac{\eta_1}{\eta_1}\right\rangle=\sum_{k_1k_2}G^2(k_1,k_2,\eta_1),$$

$$\left\langle\frac{\eta_1}{\eta_{1+i}}\right\rangle=\left\langle\frac{\eta_1}{\eta_1}\right\rangle^{-1}\sum_{k_1k_2}G(k_1k_2,\eta_{i+1})G(k_1k_2,\eta_1),$$

$$\left\langle\frac{\eta_2}{\eta_2}\right\rangle^2=\left\langle\frac{\eta_1}{\eta_1}\right\rangle^{-2}\sum_{i=0}^{1}\sum_{k_1k_2k_1'k_2'}(-)^{[i/2]}G(k_1k_2,\eta_2)G(k_1k_2,\eta_{2-i})G(k_1'k_2',\eta_{1+i})G(k_1'k_2',\eta_1), \tag{3.29}$$

$$\cdots\ \cdots$$

$$\begin{aligned}\left\langle\frac{\eta_3}{\eta_3}\right\rangle^2=&\left\langle\frac{\eta_1}{\eta_1}\right\rangle^{-5}\left\langle\frac{\eta_2}{\eta_2}\right\rangle^{-1}\sum_{i=0}^{2}{}'\sum_{j=0}^{1}{}'\sum_{k=0}^{1}{}'\sum_{k_l,p_l,q_l,n_l}(-)^{[i/2]+[j/2]+[k/2]}G(k_1k_2,\eta_3)\\ &\times G(k_1k_2,\eta_{3-i-j})\\ &\times G(p_1p_2,\eta_{1+i})G(p_1p_2,\eta_1)G(q_1q_2,\eta_{2+j})\\ &\times G(q_1q_2,\eta_{2-k})G(n_1n_2,\eta_{1+k})G(n_1n_2,\eta_1),\end{aligned}$$

where the primes on the summation signs indicate that the sums should be restricted by

$$i+j+k\leqslant 2, \tag{3.30}$$

and

$$[x]=\begin{cases}x & \text{if } x \text{ is an integer,}\\ 2x & \text{if } x \text{ is a half-integer.}\end{cases} \tag{3.31}$$

The expression becomes more complicated with increasing k in η_k. Once $\langle \eta_k / \eta_k \rangle$ for $k=1,2,...,m$ are known, one can similarly get the WCs or RWCs of SU(3)⊃U(2) with multiplicity,

$$\left\langle \frac{\eta_k}{\rho_1\rho_2\rho} \right\rangle = \left(\prod_{t=0}^{k-1} \left\langle \frac{\eta_t}{\eta_t} \right\rangle^{-1} \sum_{i_t=0}^{k-t}{}' \right) \sum_{k_1 k_2} G(k_1,k_2,\rho_1\rho_2\rho) G(k_1,k_2,\eta_{k-\Sigma_{j=1}^{k-1} i_j}) \times (-)^{\Sigma_{p=1}^{k-1}[i_p/2]} \prod_{q=1}^{k-1} \left\langle \frac{\eta_q}{\eta_{q+i_p}} \right\rangle, \tag{3.32}$$

where the prime on the summation sum indicates that the condition

$$\sum_{t=0}^{k-1} i_t \leqslant k-1 \tag{3.33}$$

should be satisfied. Similarly, define

$$V\left(\frac{\eta_i}{\eta_j} \right) = \left\langle \frac{\eta_i}{\eta_i} \right\rangle^{-1} \left\langle \frac{\eta_i}{\eta_j} \right\rangle. \tag{3.34a}$$

Then the Racah coefficient of SU(3) can be expressed as

$$\begin{aligned} R_{\eta_1}([\bar{m}]) \left\langle \frac{\eta_1}{\eta_1} \right\rangle &= G([\bar{m}],\eta_1), \\ R_{\eta_2}([\bar{m}]) \left\langle \frac{\eta_2}{\eta_2} \right\rangle &= G([\bar{m}],\eta_2) - V\left(\frac{\eta_1}{\eta_2} \right) G([\bar{m}],\eta_1), \\ &\cdots \cdots \\ R_{\eta_j}([\bar{m}]) \left\langle \frac{\eta_j}{\eta_j} \right\rangle &= G([\bar{m}],\eta_j) + \sum_{i_1=1}^{j-1} \sum_{k=1}^{j-1} (-)^k \sum_{i_1 \leqslant i_2 \leqslant \cdots \leqslant i_k < j} G([\bar{m}],\eta_{i_1}) V\left(\frac{\eta_{i_k}}{\eta_j} \right) \prod_{l=1}^{k-1} V\left(\frac{\eta_{i_l}}{\eta_{i_{l+1}}} \right). \end{aligned} \tag{3.34b}$$

IV. Some features of the new SU(3)⊃U(2) RWCs

In this section, symmetry properties of WCs of SU(3) in the canonical basis will be considered. In Sec. III, the phase

$$\left\langle \begin{matrix} [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,\mu_2] \\ [\lambda_1+\mu_1,\mu_1] & [\lambda_2+\mu_2,0] \end{matrix} \middle| \begin{matrix} \eta[m_1m_2m_3] \\ [m_1,m_2-\eta,m_3-\mu_2+\eta] \end{matrix} \right\rangle \geqslant 0 \tag{4.1}$$

was chosen. Then, the relative phase is completely determined by the recursion relations (3.23) and (3.24). When $\mu_2=0$ one can check that this phase choice is consistent with that of (Ref. 15), which is the same as that of (Ref. 5) with the same phase structure defined by Biedenharn *et al.*[10] as it should be because the multiplicity-free RWCs of (Refs. 5 and 15) are used. Hence, in order to discuss symmetry properties of the SU(3) WCs, one can expand the WCs obtained in this paper in terms of those defined in (Ref. 10),

$$\left\langle \begin{matrix} (\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2 \end{matrix} \middle| \begin{matrix} \eta[m_1m_2m_3] \\ \rho \end{matrix} \right\rangle = \sum_{\xi} y(\xi,\eta) \left[\begin{matrix} (\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2 \end{matrix} \middle| \begin{matrix} \xi[m_1m_2m_3] \\ \rho \end{matrix} \right], \tag{4.2a}$$

where the coupling coefficients on the rhs in square brackets are the WCs defined in (Ref. 10), and $y(\xi,\eta)$ is the corresponding special transformation matrix element. Using the orthogonality relations for $y(\xi,\eta)$ and symmetry properties of the WCs discussed in (Ref. 10), one can prove that

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2\end{matrix} \middle| \begin{matrix}\eta'[m_1m_2m_3] \\ \rho\end{matrix} \right\rangle$$

$$=\sum_{\eta}\sum_{\xi} y(\xi,\eta)y(\xi,\eta')(-)^{\phi(\xi)}\left\langle \begin{matrix}(\lambda_2\mu_2) & (\lambda_1\mu_1) \\ \rho_2 & \rho_1\end{matrix} \middle| \begin{matrix}\eta[m_1m_2m_3] \\ \rho\end{matrix} \right\rangle, \tag{4.2b}$$

where the phase factor $(-)^{\phi(\xi)}$ comes from

$$\left[\begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2\end{matrix} \middle| \begin{matrix}\xi[m_1m_2m_3] \\ \rho\end{matrix} \right]=(-)^{\phi(\xi)}\left[\begin{matrix}(\lambda_2\mu_2) & (\lambda_1\mu_1) \\ \rho_2 & \rho_1\end{matrix} \middle| \begin{matrix}\xi[m_1m_2m_3] \\ \rho\end{matrix} \right] \tag{4.3}$$

given in Ref. 10,

$$\phi(\xi)=\Gamma_{12}-\Gamma_{11}+\mu_1+\mu_2-m_2+m_1. \tag{4.4}$$

The multiplicity label ξ in this case can be regarded as

$$\xi=\Gamma_{12}-\Gamma_{11}, \tag{4.5}$$

where Γ_{ij} are multiplicity labels from the upper SU(3) pattern defined by Biedenharn *et al.*

It is obvious that

$$Z_{\eta\eta'}=Z\left(\begin{matrix}(\lambda_1\mu_1) & (0) & (\lambda_1\mu_1)\eta' \\ (\lambda_2\mu_2) & [m_1m_2m_3] & (\lambda_2\mu_2)\eta\end{matrix} \right)=\sum_{\xi} y(\xi,\eta)y(\xi,\eta')(-)^{\phi(\xi)}, \tag{4.6}$$

where $Z_{\eta\eta'}$ is a special Z coefficient[38] defined by Millener, which transforms the coupling coefficients from the coupling $(\lambda_1\mu_1)\times(\lambda_2\mu_2)$ to $(\lambda_2\mu_2)\times(\lambda_1\mu_1)$. Equation (4.2) can only be simplified when the coupling is multiplicity-free. In this case, the transformation matrix Y is 1×1 with $y(\xi,\eta)=1$ for fixed ξ and η, and

$$Z_{\eta\eta'}=\delta_{\eta\eta'}(-)^{\phi(\xi)}, \tag{4.7}$$

where ξ can be expressed in terms of λ_1, μ_1, λ_2, μ_2, and m_i with $i=1, 2, 3$.

Similarly, one can deduce the following symmetry properties for the WCs of SU(3):

$$\left\langle \begin{matrix}(\mu_1\lambda_1) & (\mu_2\lambda_2) \\ \tilde{\rho}_1 & \tilde{\rho}_2\end{matrix} \middle| \begin{matrix}\eta'[-m_3-m_2-m_1] \\ \tilde{\rho}\end{matrix} \right\rangle$$

$$=\sum_{\eta} Z_{\eta\eta'}(-)^{m_1-m_3}\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2\end{matrix} \middle| \begin{matrix}\eta[m_1m_2m_3] \\ \rho\end{matrix} \right\rangle, \tag{4.8a}$$

where $\tilde{\rho}$ is the conjugation of ρ defined by

$$|\tilde{\rho}\rangle\equiv|\tilde{m}\rangle=\left| \begin{matrix}-m_{22}-m_{12} \\ -m_{11}\end{matrix} \right\rangle, \tag{4.8b}$$

and

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ \rho_1 & \rho_2\end{matrix} \middle| \begin{matrix}\eta'[m_1m_2m_3] \\ \rho\end{matrix} \right\rangle=\left[\frac{\dim([m_1m_2m_3])}{\dim((\lambda_1\mu_1))}\right]^{1/2}(-)^{\phi_3-\phi_1}$$

$$\times\sum_{\eta} Z_{\eta\eta'}\left\langle \begin{matrix}[m_1m_2m_3] & (\mu_2\lambda_2) \\ \rho & \tilde{\rho}_2\end{matrix} \middle| \begin{matrix}\eta(\lambda_1\mu_1) \\ \rho_1\end{matrix} \right\rangle, \tag{4.9}$$

where

$$\phi_3=m''_{11}-m''_{12}-m''_{22}+m_1, \tag{4.10}$$

TABLE I. RWCs of SU(3)⊃U(2), $\langle {}^{[21]}_{[q_1q_2]}\ {}^{[21]}_{[q_1'q_2']}|\ {}^{\eta[m_1m_2m_3]}_{[32]}\rangle$.

$\downarrow[q_1q_2][q_1'q_2']\backslash m_1m_2m_3$	[420]	$[321]_1$	$[321]_2$	[300]
[21] [20]	$-\sqrt{\frac{1}{20}}$	$-\sqrt{\frac{9}{20}}$	$\sqrt{\frac{1}{4}}$	$\sqrt{\frac{1}{4}}$
[20] [21]	$\sqrt{\frac{1}{20}}$	$\sqrt{\frac{9}{20}}$	$\sqrt{\frac{1}{4}}$	$\sqrt{\frac{1}{4}}$
[21] [11]	$\sqrt{\frac{9}{20}}$	$-\sqrt{\frac{1}{20}}$	$-\sqrt{\frac{1}{4}}$	$\sqrt{\frac{1}{4}}$
[11] [21]	$\sqrt{\frac{9}{20}}$	$-\sqrt{\frac{1}{20}}$	$\sqrt{\frac{1}{4}}$	$-\sqrt{\frac{1}{4}}$
[21] [20]	$-\sqrt{\frac{1}{20}}$	$\sqrt{\frac{7}{10}}$	0	$\sqrt{\frac{1}{4}}$
[20] [21]	$\sqrt{\frac{1}{20}}$	$-\sqrt{\frac{2}{35}}$	$\sqrt{\frac{9}{14}}$	$\sqrt{\frac{1}{4}}$
[21] [11]	$\sqrt{\frac{9}{20}}$	$-\sqrt{\frac{1}{70}}$	$-\sqrt{\frac{2}{7}}$	$\sqrt{\frac{1}{4}}$
[11] [21]	$\sqrt{\frac{9}{20}}$	$\sqrt{\frac{8}{35}}$	$\sqrt{\frac{1}{14}}$	$-\sqrt{\frac{1}{4}}$
$\uparrow[q_1q_2][q_1'q_2']\backslash[m_1m_2m_3]$	[420]	$[321]_{\eta=0}$	$[321]_{\eta=1}$	[300]

$$\phi_1 = m_{11} - m_{12} - m_{22} + \lambda_1 + \mu_1,$$

and

$$\rho = \begin{pmatrix} [m''_{12} m''_{22}] \\ m''_{11} \end{pmatrix}, \quad \rho_1 = \begin{pmatrix} [m_{12} m_{22}] \\ m_{11} \end{pmatrix}. \tag{4.11}$$

Explicit examples for the coupling $[21]\times[21]\downarrow[321]$, which illustrate main features of the new RWCs, will now be given. The following notation will be used:

$$\begin{aligned} \left\langle \frac{0}{0} \right\rangle &= \left\langle \begin{matrix} [21] & [21] \\ [21] & [20] \end{matrix} \middle| \begin{matrix} \eta=0 \\ \end{matrix} \begin{matrix} [321] \\ [32] \end{matrix} \right\rangle, \quad \left\langle \frac{1}{0} \right\rangle = \left\langle \begin{matrix} [21] & [21] \\ [21] & [20] \end{matrix} \middle| \begin{matrix} \eta=1 \\ \end{matrix} \begin{matrix} [321] \\ [32] \end{matrix} \right\rangle \\ \left\langle \frac{0}{1} \right\rangle &= \left\langle \begin{matrix} [21] & [21] \\ [21] & [20] \end{matrix} \middle| \begin{matrix} \eta=0 \\ \end{matrix} \begin{matrix} [321] \\ [311] \end{matrix} \right\rangle, \quad \left\langle \frac{1}{1} \right\rangle = \left\langle \begin{matrix} [21] & [21] \\ [21] & [20] \end{matrix} \middle| \begin{matrix} \eta=1 \\ \end{matrix} \begin{matrix} [321] \\ [311] \end{matrix} \right\rangle. \end{aligned} \tag{4.12}$$

Using the recursion relations given by (3.23) and (3.24), it is simple to show that

$$\begin{pmatrix} \left\langle \frac{0}{0} \right\rangle & \left\langle \frac{1}{0} \right\rangle \\ \left\langle \frac{0}{1} \right\rangle & \left\langle \frac{1}{1} \right\rangle \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{7}{10}} & 0 \\ -\sqrt{\frac{1}{42}} & \sqrt{\frac{10}{21}} \end{pmatrix}. \tag{4.13}$$

With this result in hand, Tables I and II can be worked out by using (3.28). In the tables, the upper parts are taken from de Swart,[39] while the lower parts are derived by the new method. These two types of RWCs with multiplicity two can be transformed with each other by a two dimensional rotation. Furthermore, one can check that the new RWCs satisfy the orthogonality conditions for SU(3) RWCs given by (2.6).

V. REDUCED AUXILIARY WIGNER COEFFICIENTS FOR SU(3)⊃U(2)

The complementary U(4) group is also used in Ref. 11 to label the multiplicity of SU(3). However, in that case the U(4) group is labeled in a noncanonical U(2)×U(2) chain. The authors also point out that the so-called auxiliary Wigner coefficient (AWC) of SU(3) can be calculated from a U(4)⋆SU(3) scalar product. The reduced AWCs satisfy

$$\sum_{\lambda_2\mu_2 q_1' q_i''} \left\langle \begin{matrix} (\lambda_1\mu_1) & (\lambda_2\mu_2) \\ [q_1'q_2'] & [q_1''q_2''] \end{matrix} \middle| \begin{matrix} [m_1m_2m_3] \\ [q_1q_2] \end{matrix} ; [u_1u_2u_3] \right\rangle \left\langle \begin{matrix} (\lambda_1\mu_1) & (\lambda_2\mu_2) \\ [q_1'q_2'] & [q_1''q_2''] \end{matrix} \middle| \begin{matrix} [\bar{m}_1\bar{m}_2\bar{m}_3] \\ [q_1q_2] \end{matrix} ; [\bar{u}_1\bar{u}_2\bar{u}_3] \right\rangle$$

$$= \prod_i \delta_{\bar{u}_i u_i} \delta_{m_i m_i}, \tag{5.1}$$

TABLE II. RWCs of SU(3)⊃U(2), $\langle {}^{[21]}_{[q_1q_2]}\ {}^{[21]}_{[q_1'q_2']}|{}^{\eta[m_1m_2m_3]}_{[22]}\rangle$.

$\downarrow[q_1q_2][q_1'q_2']\backslash[m_1m_2m_3]$	[420]	$[321]_1$	$[321]_2$	[330]
[10] [21]	$\sqrt{\frac{3}{20}}$	$\sqrt{\frac{1}{10}}$	$\sqrt{\frac{1}{2}}$	$\sqrt{\frac{1}{4}}$
[21] [10]	$-\sqrt{\frac{3}{20}}$	$-\sqrt{\frac{1}{10}}$	$\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{4}}$
[20] [20]	$-\sqrt{\frac{1}{40}}$	$-\sqrt{\frac{3}{5}}$	0	$\sqrt{\frac{3}{8}}$
[11] [11]	$\sqrt{\frac{27}{40}}$	$-\sqrt{\frac{1}{5}}$	0	$-\sqrt{\frac{1}{8}}$
[10] [21]	$\sqrt{\frac{3}{20}}$	$\sqrt{\frac{16}{35}}$	$-\sqrt{\frac{1}{7}}$	$\sqrt{\frac{1}{4}}$
[21] [10]	$-\sqrt{\frac{3}{20}}$	$\sqrt{\frac{1}{35}}$	$-\sqrt{\frac{4}{7}}$	$-\sqrt{\frac{1}{4}}$
[20] [20]	$-\sqrt{\frac{1}{40}}$	$-\sqrt{\frac{27}{70}}$	$-\sqrt{\frac{3}{14}}$	$\sqrt{\frac{3}{8}}$
[11] [11]	$\sqrt{\frac{27}{40}}$	$-\sqrt{\frac{9}{70}}$	$-\sqrt{\frac{1}{14}}$	$-\sqrt{\frac{1}{8}}$
$\uparrow[q_1q_2][q_1'q_2']\backslash[m_1m_2m_3]$	[420]	$[321]_{\eta=0}$	$[321]_{\eta=1}$	[330]

$$\sum_{u_i m_i} \left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ [q_1'q_2'] & [q_1''q_2'']\end{matrix}\middle| \begin{matrix}[m_1m_2m_3]\\ [q_1q_2]\end{matrix};[u_1u_2u_3]\right\rangle \left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2)\\ [\overline{q_1'}\overline{q_2'}] & [\overline{q_1''}\overline{q_2''}]\end{matrix}\middle| \begin{matrix}[m_1m_2m_3]\\ [q_1q_2]\end{matrix};[u_1u_2u_3]\right\rangle$$
$$= \prod_i \delta_{\overline{q_i'}q_i''}\delta_{\overline{q_i}q_i'}, \tag{5.2}$$

where $[u_1u_2u_3]$ are labels of $\mathcal{U}(3)$ which is the subgroup of U(4). The AWCs can be shown to be equivalent to those constructed using the recoupling approach. First of all, assume the irrep $(\lambda_2\mu_2)$ is a coupled irrep, namely,

$$\left|\begin{matrix}(\lambda_2\mu_2)\\ \rho_2\end{matrix}\right\rangle = \sum_{\rho_2'\rho_2''} \left\langle \begin{matrix}(\lambda_2'+\mu_2'0) & (\mu_2'0)\\ \rho_2' & \rho_2''\end{matrix}\middle| \begin{matrix}(\lambda_2\mu_2)\\ \rho_2\end{matrix}\right\rangle \left|\begin{matrix}\lambda_2'+\mu_2' & ; & \mu_2'\\ \rho_2' & & \rho_2''\end{matrix}\right\rangle, \tag{5.3}$$

where ρ_2, ρ_2', ρ_2'' are sublabels of SU(3) in the canonical basis. Special final coupled basis vectors can then be written as

$$\left|((\lambda_1\mu_1),(\lambda_2'+\mu_2'0))[u_1u_2u_3],(\mu_2'0),{}^{[m_1m_2m_3]}_{\rho}\right\rangle$$
$$\equiv \left|{}^{[m_1m_2m_3]}_{\rho};[u_1u_2u_3]\right\rangle$$
$$= \sum \left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2'+\mu_2'0)\\ \rho_1 & \rho_2'\end{matrix}\middle| \begin{matrix}[u_1u_2u_3]\\ \overline{\rho}\end{matrix}\right\rangle \left\langle \begin{matrix}[u_1u_2u_3] & (\mu_2'0)\\ \overline{\rho} & \rho_2''\end{matrix}\middle| \begin{matrix}[m_1m_2m_3]\\ \rho\end{matrix}\right\rangle$$
$$\times \left\langle \begin{matrix}(\lambda_2'+\mu_2'0) & (\mu_2'0)\\ \rho_2' & \rho_2''\end{matrix}\middle| \begin{matrix}(\lambda_2\mu_2)\\ \rho_2\end{matrix}\right\rangle \left|\begin{matrix}(\lambda_1\mu_1) & ; & (\lambda_2'+\mu_2'0) & ; & (\mu_2'0)\\ \rho_1 & & \rho_2' & & \rho_2''\end{matrix}\right\rangle, \tag{5.4}$$

where the sum is over ρ_1, ρ_2', ρ_2'', $\overline{\rho}$, and the WCs involved in the summation are all given by Ališauskas *et al.*[15] and Chacon *et al.*[5]

The AWCs can be evaluated by the overlap

$$\left\langle \begin{matrix}(\lambda_1\mu_1) \\ \rho_1\end{matrix} ; \begin{matrix}(\lambda_2\mu_2) \\ \rho_2\end{matrix} \middle| \begin{matrix}[m_1m_2m_3] \\ \rho\end{matrix} ; [u_1u_2u_3] \right\rangle = \sum_{\rho_2'\rho_2''\overline{\rho}} \left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2'+\mu_2'0) \\ \rho_1 & \rho_2'\end{matrix} \middle| \begin{matrix}[u_1u_2u_3] \\ \overline{\rho}\end{matrix} \right\rangle \times \left\langle \begin{matrix}[u_1u_2u_3] & (\mu_2'0) \\ \overline{\rho} & \rho_2''\end{matrix} \middle| \begin{matrix}[m_1m_2m_3] \\ \rho\end{matrix} \right\rangle \times \left\langle \begin{matrix}(\lambda_2'+\mu_2'0) & (\mu_2'0) \\ \rho_2' & \rho_2''\end{matrix} \middle| \begin{matrix}(\lambda_2\mu_2) \\ \rho_2\end{matrix} \right\rangle. \tag{5.5}$$

One can check that the labels u_1, u_2, u_3 take the same values as those given by Brody *et al.*[11] One can also get reduced AWCs as follows:

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2\mu_2) \\ [q_1q_2] & [q_1'q_2']\end{matrix} \middle| \begin{matrix}[m_1m_2m_3] \\ [q_1''q_2'']\end{matrix} ; [u_1u_2u_3] \right\rangle = \sum_{pq_1q_2} \left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2'+\mu_2'0) \\ [q_1q_2] & [\lambda_2'+\mu_2'-p0]\end{matrix} \middle| \begin{matrix}[u_1u_2u_3] \\ [\overline{q_1}\overline{q_2}]\end{matrix} \right\rangle \left\langle \begin{matrix}[u_1u_2u_3] & (\mu_2'0) \\ [\overline{q_1}\overline{q_2}] & [p0]\end{matrix} \middle| \begin{matrix}[m_1m_2m_3] \\ [q_1''q_2'']\end{matrix} \right\rangle \times \left\langle \begin{matrix}(\lambda_2'+\mu_2'0) & (\mu_2'0) \\ [\lambda_2'+\mu_2'-p0] & [p0]\end{matrix} \middle| \begin{matrix}(\lambda_2\mu_2) \\ [q_1'q_2']\end{matrix} \right\rangle U\left(\begin{matrix}[q_1q_2] & [\lambda_2'+\mu_2'-p0] & [\overline{q_1}\overline{q_2}] \\ [p0] & [q_1''q_2''] & [q_1'q_2']\end{matrix} \right). \tag{5.6}$$

One can verify that the reduced AWCs given by (5.5) indeed satisfy the orthogonality relations (5.1) and (5.2). Using the explicit expression of the multiplicity-free RWCs involved in the sum, one can get a closed algebraic expression for the reduced AWCs. Such AWCs may also be useful, especially in two-particle coupling problems. It should be pointed out that though the result (5.6) is the same as required by Brody *et al.*, the labeling scheme is not the same. The multiplicity labels of their AWCs are specified by a U(4) subgroup U(3), while they are now specified by the irrep of SU(3) coupled from the first two irreps.

Using analytic expressions for the multiplicity-free RWCs given in Ref. 15, the following algebraic expression for the reduced AWCs of SU(3)⊃U(2) can be obtained:

$$\left\langle \begin{matrix}(\lambda_1\mu_1) & (\lambda_2'+\mu_2';\mu_2')(\lambda_2\mu_2) \\ [m_1m_2] & [m_1'm_2']\end{matrix} \middle| \begin{matrix}[m_1m_2m_3] \\ [m_1''m_2'']\end{matrix} ; [u_1u_2u_3] \right\rangle$$

$$= \delta_{\lambda_2'+2\mu_2',\lambda_2+2\mu_2}\delta_{\lambda_1+2\mu_1+\lambda_2+2\mu_2,m_1+m_2+m_3}$$

$$\times\left[\frac{(\lambda_2+1)(\lambda_2'+\mu_2'-\mu_2)!\mu_2!(m_1-m_2+1)(m_1-m_3+2)(m_2-m_3+1)(m_{12}''-m_2)!}{(\lambda_2+\mu_2-\lambda_2'-\mu_2')!(\mu_2-\mu_2')!(\lambda_2+\mu_2-m_{12}')!(m_2-k_2+1)!(m_1-m_{12}'')!(m_2-m_{22}'')!}\right.$$

$$\times\frac{(\lambda-k_1-k_2-m_2)!(\mu_1'+k_1-m_3)!(\lambda-k_1-k_2-m_3+1)!(m_{22}''-m_3)!(m_{12}''-m_3+1)!}{(m_1-\lambda+k_1+k_2)!(m_2-\mu_1-k_1)!(m_3-k_2)!(m_1-\mu_1-k_1+1)!(m_1-k_2+2)!}$$

$$\left.\times\frac{(m_{12}'-\mu_2)!(\mu_1-k_2)!(\lambda_1+\mu_1-k_2+1)!(\lambda_1+\mu_1-m_{12})!(\mu_1-m_{22})!(\lambda_1+\mu_1-m_{22}+1)!}{(\mu_2-m_{22}')!(\lambda_2+\mu_2-m_{22}'+1)!(\mu_1+k_1+1)!(m_{12}-\mu_1)!m_{22}!(m_{12}+1)!(m_1-m_{22}''+1)!}\right]^{1/2}$$

$$\times\sum_{p=0}^{\mu_2'}\sum_{q=\min(q)}^{\max(q)}(\mu_2'-p)!(\lambda_2'+\mu_2'+p-m_{12}'-m_{22}')!$$

$$\times\left(\frac{(m_{12}''+m_{22}''-p-q-m_{12})!(p-m_{22}')!(q-m_{22})!}{(m_{12}'-p)!}\right)^{1/2}$$

$$\times\left(\frac{(m_{12}''+m_{22}''-p-q-m_{22}+1)!(m_{12}''+m_{22}''-p-2q+1)(p+q-m_{22}'')!(m_{22}''-q+1)!}{(m_{12}''-p-q)!(m_{12}''+m_{22}''-p-q-\mu_1-k_1)!(m_{12}-q)!}\right)^{1/2}$$

$$\times U\left(\begin{matrix}[m_{12}m_{22}] & [m_{12}'+m_{22}'-p0] & [m_{12}''+m_{22}''-p-q,q] \\ [p0] & [m_{12}''m_{22}''] & [m_{12}'m_{22}']\end{matrix} \right)\times F_2\left(\begin{matrix}[\lambda-k_1-k_2,\mu_1+k_1,k_2] & [m_1m_2m_3] \\ [m_{12}''+m_{22}''-p-q,q] & [m_{12}''m_{22}'']\end{matrix} \right)$$

$$\times F_2\begin{pmatrix}[\lambda_1+\mu_1\mu_1] & [\lambda-k_1-k_2,\mu_1+k_1,k_2]\\ [m_{12}+m_{22}0] & [m''_{12}+m''_{22}-p-q,q]\end{pmatrix}$$
$$\times F\begin{pmatrix}\lambda'_2+\mu'_2 & ; & \mu'_2 & ; & [\lambda_2+\mu_2\mu_2]\\ m'_{12}+m'_{22}-p & & p & & [m'_{12}m'_{22}]\end{pmatrix},$$

(5.7)

where the $(\lambda_2\mu_2)$ irrep is written in coupled form $(\lambda'_2 0)\times(\mu'_2 0)$ with $\lambda'_2+2\mu'_2=\lambda_2+2\mu_2$,

$$\lambda=\lambda_1+\mu_1+\lambda'_2+\mu'_2,$$

$$[u_1,u_2,u_3]=[\lambda-k_1-k_2,\mu_1+k_1,k_2],$$
$$\min(q)=\max(k_2,m''_{12}+m''_{22}-p-\lambda+k_1+k_2),$$
$$\max(q)=\min(\mu_1+k_1,m''_{12}+m''_{22}-p-\mu_1-k_1),$$

(5.8)

and

$$F\begin{pmatrix}\lambda'_2+\mu'_2 & ; & \mu'_2 & ; & [\lambda_2+\mu_2\mu_2]\\ m'_{12}+m'_{22}-p & & p & & [m'_{12}m'_{22}]\end{pmatrix}$$
$$=\sum_t (-)^t \frac{(m'_{12}-p+t)!(\lambda_2+\mu_2-m'_{12}-m'_{22}+p-t)!}{t!(\lambda'_2+\mu'_2-m'_{12}-m'_{22}+p-t)!(p-m'_{22}-t)!(m'_{12}+m'_{22}-p-\mu_2+t)!}.$$

(5.9)

VI. DISCUSSION

In this paper, the complementary group technique proposed in (I) for a resolution of outer multiplicity problem of SU(n) is used to obtain a general procedure for the derivation of SU(3)⊃U(2) RWCs with multiplicity. The procedure uses a recoupling formula together with a special transformation within the multiplicity space. The outer multiplicity labeling scheme is shown to be nonunique; one can transform from one scheme to another under SO(m), where m is the multiplicity of $[m_1m_2m_3]$ in the decomposition $(\lambda_1\mu_1)\times(\lambda_2\mu_2)$. From this perspective, Biedenharn's resolution for the outer multiplicity of SU(n) can also be regarded as a complementary group resolution. In their work, the complementary group is U(n). Unlike the resolution proposed in (I) where only special Gel'fand basis are considered, here the subirreps of the complementary group U(n) are all considered in order to resolve the multiplicity of SU(n)×SU(n). This is why the multiplicity formulae cannot be worked out easily. The canonical schemes all belong to an equivalent class.

Using the method proposed, RWCs of SU(3)⊃U(2) with multiplicity can be derived recursively. A computer algorithm based on this procedure can be easily developed, which will make numerical calculation possible in practical applications. Furthermore, one can also obtain closed algebraic expressions for these RWCs for small m values. However, the expression remains cumbersome with summation over many variables. The complexity increases with increasing the multiplicity.

Only RWCs of SU(3) in the canonical chain SU(3)⊃U(2) are discussed. Nonetheless, this method can also be applied to noncanonical basis of SU(3), for example SU(3)⊃SO(3), by using RWCs of SU(3)⊃SO(3) with one symmetric irrep, and therefore multiplicity-free.[40,41] In addition, the procedure outlined in this paper can be extended to the general SU(n) case. The SU(4)⊃U(3) RWCs will be the topic of the next paper.

A closed algebraic expression of reduced AWCs proposed in Ref. 11 is obtained using the new approach. These coefficients satisfy different orthogonality relations and therefore may be useful in many-particle coupling problems.

ACKNOWLEDGMENTS

The authors are grateful to Professor S. Ališauskas for helpful suggestions and comments. This project was supported by the National Science Foundation under Cooperative Agreement No. EPS-9550481 and Grant No. 9603006, which includes matching from the Louisiana Board of Regents Support Fund, and by the State Education Commission of China.

[1] G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963;); **5**, 1723 (1964;); **5**, 1730 (1964;); **6**, 1847 (1965).
[2] L. C. Biedenharn, A. Giovannini, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).
[3] L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1968).
[4] L. C. Biedenharn, J. D. Louck, E. Chacon, and M. Ciftan, J. Math. Phys. **13**, 1957 (1972).
[5] E. Chacon, M. Ciftan, and L. C. Biedenharn, J. Math. Phys. **13**, 577 (1972).
[6] L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **93**, 143 (1984).
[7] L. C. Biedenharn, M. A. Lohe, and J. D. Louck, J. Math. Phys. **26**, 1458 (1985).
[8] R. Le Blanc and L. C. Biedenharn, J. Phys. A **22**, 4613 (1989).
[9] K. T. Hecht and L. C. Biedenharn, J. Math. Phys. **31**, 2781 (1990).
[10] L. C. Biedenharn, M. A. Lohe, and H. T. Williams, J. Math. Phys. **35**, 6672 (1994).
[11] T. A. Brody, M. Moshinsky, and I. Renero, J. Math. Phys. **6**, 1540 (1965).
[12] M. Moshinsky, J. Math. Phys. **4**, 1128 (1963).
[13] M. Moshinsky, Rev. Mod. Phys. **34**, 813 (1962).
[14] M. Moshinsky, J. Math. Phys. **7**, 691 (1966).
[15] S. Ališauskas, A. A. Jucys, and A. P. Jucys, J. Math. Phys. **13**, 1329 (1972).
[16] S. Ališauskas, V. V. Vanagas, and J. P. Jucy, Dokl. Akad. Nauk SSSR **197**, 804 (1971).
[17] S. Ališauskas, Sov. J. Part. Nucl. **14**, 563 (1983).
[18] S. Ališauskas, J. Math. Phys. **29**, 2351 (1988).
[19] S. Ališauskas, J. Math. Phys. **31**, 1325 (1990).
[20] S. Ališauskas, J. Math. Phys. **33**, 1983 (1992).
[21] S. Ališauskas, J. Phys. A **29**, 2687 (1996).
[22] K. T. Hecht, Nucl. Phys. **62**, 1 (1965).
[23] M. Resnikoff, J. Math. Phys. **8**, 63 (1967).
[24] L. A. Shelepin and V. P. Karasev, Sov. J. Nucl. Phys. **5**, 156 (1967).
[25] V. P. Karasev and L. A. Shelepin, Sov. J. Nucl. Phys. **7**, 678 (1968).
[26] A. U. Klimyk and A. M. Gavrilik, J. Math. Phys. **20**, 1624 (1979).
[27] R. Le Blanc and D. J. Rowe, J. Phys. A **19**, 2913 (1986).
[28] J. P. Draayer and Y. Akiyama, J. Math. Phys. **14**, 1904 (1973).
[29] Y. Akiyama and J. P. Draayer, Comput. Phys. Commun. **5**, 405 (1973).
[30] T. A. Kaeding, Comput. Phys. Commun. **85**, 82 (1995).
[31] T. A. Kaeding and H. T. Williams, Comput. Phys. Commun. **98**, 398 (1996).
[32] H. T. Williams, J. Math. Phys. **37**, 4187 (1996).
[33] J. A. Castilho Alcaras, L. C. Biedenharn, K. T. Hecht, and G. Neely, Ann. Phys. (N.Y.) **60**, 85 (1970).
[34] J. S. Prakash and H. S. Sharatchandra, J. Math. Phys. **37**, 6530 (1996).
[35] D. Braunschweig, Comput. Phys. Commun. **14**, 109 (1978).
[36] K. T. Hecht, Nucl. Phys. **62**, 1 (1965); **63**, 177 (1965); **A102**, 11 (1967).
[37] K. T. Hecht and S. C. Pang, J. Math. Phys. **10**, 1571 (1969).
[38] D. Millener, J. Math. Phys. **19**, 1513 (1978).
[39] J. J. De Swart, Rev. Mod. Phys. **35**, 916 (1963).
[40] J. D. Vargados, Nucl. Phys. **A111**, 681 (1968).
[41] S. Ališauskas and E. Norvaišas, J. Phys. A **22**, 5177 (1989).

Comment on "An exactly solvable two-body problem with retarded interactions and radiation reaction in classical electrodynamics" [J. Math. Phys. 38, 5690 (1997)]

V. Hnizdo
Department of Physics and Schonland Research Centre for Nuclear Sciences, University of the Witwatersrand, Johannesburg, 2050 South Africa

(Received 28 January 1998; accepted for publication 7 April 1998)

[S0022-2488(98)02710-8]

A recent paper under the above title by Rivera and Villarroel[1] considers the problem of energy conservation in a system of two identical charges moving uniformly and diametrically opposite each other along a circle. Apart from the external forces needed to maintain their circular motion, the charges are subject to their retarded electromagnetic interaction and the Lorentz–Dirac radiation reaction force. We would like to point out that this interesting problem had already been investigated, in numerical calculations[2] as well as analytically,[3] with findings and conclusions similar to those of Ref. 1—namely that the rate of radiation from such a system

$$4\beta^4 \sum_{n=2,4,6,\ldots} n^2 \int_0^{\pi} \left[\left(\frac{dJ_n(n\beta \sin\theta)}{d(n\beta \sin\theta)} \right)^2 + n^2 \cos^2\theta \left(\frac{J_n(n\beta \sin\theta)}{n\beta \sin\theta} \right)^2 \right] \sin\theta \, d\theta, \tag{1}$$

equals the rate of work of the external forces against those of the Lorentz–Dirac radiation reaction and the retarded Lienard–Wiechert interaction[4]

$$\frac{4\beta^4}{3(1-\beta^2)^2} - \frac{\beta[(\beta \cos 2\psi - \sin\psi)(1-\beta^2) + 2\psi^2(\sin\psi + \beta)]}{2(1+\beta \sin\psi)^3 \cos^2\psi}. \tag{2}$$

Here, $J_n(x)$ are the Bessel functions and $\beta = \psi/\cos\psi$ is the speed of the circling motion, with ψ being the angle subtended at the position of each charge by the retarded and current positions of the other charge; units such that the speed of light, radius of the circle and magnitude of the charges are all unity are used. The above expressions also hold for a system of unlike charges of equal magnitude when the summation in Eq. (1) is changed to run over n odd and the second term in Eq. (2) is added to, instead of subtracted from, the first term. The analytical proof[5] of the equality of the expressions (1) and (2) dispenses with the task that Rivera and Villarroel set themselves to undertake in another paper.

It should be noted also that the mistake in a paper by Briggs[6] which Rivera and Villarroel have found had been corrected in an erratum,[7] and that, in a remarkable coincidence, a paper by Gordeyev[8] addressing the same problem and obtaining similar findings appeared in the literature concurrently with Ref. 1. Gordeyev has derived an analytical expression for the angle ψ in terms of the speed β

$$\psi = 2 \sum_{m=0}^{\infty} \frac{(-1)^m}{2m+1} J_{2m+1}[(2m+1)\beta]. \tag{3}$$

[1] R. Rivera and D. Villarroel, J. Math. Phys. **38**, 5690 (1997).
[2] V. Hnizdo, Phys. Lett. A **129**, 426 (1988); K. Briggs, Aust. J. Phys. **41**, 629 (1988).
[3] V. Hnizdo, SIAM (Soc. Ind. Appl. Math.) Rev. **32**, 149 (1990); W. B. Jordan, *ibid.* **32**, 150 (1990); misprints corrected in *ibid.* **34**, 126 (1992).
[4] Thus refuting a claim to the contrary, see E. Comay, Phys. Lett. A **126**, 155 (1987); **129**, 424 (1988).
[5] W. B. Jordan, Ref. 3; see also Ref. 8.
[6] K. Briggs, Phys. Rev. A **37**, 977 (1988).
[7] K. Briggs, Phys. Rev. A **39**, 914 (1989).
[8] A. N. Gordeyev, Am. J. Phys. **65**, 1060 (1997).

 © 1998 American Institute of Physics

Response to ''Comment on 'An exactly solvable two-body problem with retarded interactions and radiation reaction in classical electrodynamics' '' [J. Math. Phys. 39, 5663 (1998)]

R. Rivera and D. Villarroel
Departmento de Física, Universidad Técnica Federico Santa Maria, Casilla 110-V, Valparaiso, Chile

(Received 28 January 1998; accepted for publication 7 April 1998)

[S0022-2488(98)02710-8]

In his comment[1] Hnizdo brings to our knowledge a previous work[2] in which the rate of radiation of two identical charges moving with constant angular speed at opposite ends of a diameter is numerically evaluated. Although this calculation is related to part of our paper, our work is based on a more general framework, since it deals with the construction of an exact solution of the Lorentz–Dirac equation, where the external fields that make the motion possible, and their sources are studied in detail.

Regarding the calculation of the rate of radiation, Hnizdo refers to a paper by Jordan.[3] This work starts from a formula given in Jackson's book,[4] where the rate of radiation is expressed in terms of Bessel functions as a sum of the power radiated in the different harmonics. On the other hand, we work directly with the Liénard–Wiechert retarded fields and obtain an alternative representation for the power radiated [see our Eq. (3.27)]. Our method seems to be simpler, since it avoids tedious manipulations of Bessel functions in order to evaluate the rate of radiation.

[1] V. Hnizdo, J. Math. Phys. **39**, 5663 (1998).
[2] V. Hnizdo, Phys. Lett. A **129**, 426 (1988).
[3] W. B. Jordan, SIAM (Soc. Ind. Appl. Math.) Rev. **32**, 150 (1990); misprints corrected in **34**, 126 (1992).
[4] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), Problem 14.12.

 © 1998 American Institute of Physics

Erratum: ''Particles in singular magnetic field'' [J. Math. Phys. 39, 818 (1998)]

Władysław Marcinek
Institute of Theoretical Physics, University of Wrocław, Pl. Maxa Borna 9, 50-024 Wrocław, Poland

(Received 16 March 1998; accepted for publication 31 March 1998)

[S0022-2488(98)00410-1]

The factor $-(-1)^{\Omega_{ij}}$ in the second and third line of the multilinear formula (21) should be replaced by $-(-1)^{N}$. The block of formulas (22) should be replaced by (23) and vice versa.

 © 1998 American Institute of Physics

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 11 NOVEMBER 1998

Two-form gauge fields in the Becchi–Rouet–Stora–Tyutin superspace

J. Barcelos-Neto[a] and R. Thibes[b]
Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ 21945-970, Brazil

(Received 6 November 1997; accepted for publication 26 June 1998)

We consider a geometrical description for tensor gauge fields. Based on this geometrical treatment, we develop the theory involving one- and two-form gauge fields by means of the Becchi–Rouet–Stora–Tyutin (BRST) superfield formalism. This permits us to directly obtain invariant Lagrangians for both BRST and anti-BRST transformations and we shall see that all the ingredients of the formalism (ghosts, ghost-for-ghosts and all the auxiliary fields) naturally occur. We introduce collective fields to construct the field–antifield quantum action in a generic gauge. We deal with both Abelian and non-Abelian cases. In this last case, the BRST superspace formulation sheds more light on this still open problem. © *1998 American Institute of Physics.* [S0022-2488(98)00211-4]

I. INTRODUCTION

The geometrical interpretation of the Becchi–Rouet–Stora–Tyutin (BRST)[1] and anti-BRST[2] symmetries leads to a particular superspace formulation of gauge theories,[3] where the corresponding superfields are objects that naturally contain gauge, ghost, and auxiliary fields. The importance of formulating a gauge theory in terms of these superfields is that it is naturally invariant under BRST and anti-BRST transformations. As examples of the use of this formalism, we mention the usual Yang–Mills theory[3,4] and the path integral quantization of relativistic strings.[5] It can also be used in the field–antifield quantization prescription due to Batalin–Vilkovisly (BV),[6] where a set of fields called *collective fields* is introduced.[7–9] We also mention that in the case of Yang–Mills and non-Abelian two-form gauge fields, one may use the duality property to identify the antifields.[10]

Our purpose here is to apply the superfield formalism for gauge theories containing one- and two-form gauge fields, coupled in a topological way. We show that this directly leads to the quantum BV actions. We shall consider both Abelian and non-Abelian cases. We mention that the particular Abelian case without topological term has already been treated by Marchetti and Tonin, but without any relation with the BV formalism.[11]

It might be opportune to mention that the interest for tensor gauge fields dates back to more than 20 years. They were considered by Kalb and Ramond[12] with the motivation that they could carry the force in the string interaction. At the same time, Cremmer and Scherk[13] considered one- and two-form gauge fields coupled in a topological way with the purpose of obtaining a kind of dynamical symmetry breaking for the gauge symmetry and a consequent mass generation for the vector field. This same problem has been considered nowadays in a version where the mass generation is carried out as an effective theory when the tensor field is conveniently eliminated.[14–16] We also mention that tensor gauge fields appear in recent developments of string theories and supergravity models.[17]

The formulation of two-form gauge fields in the context of the BRST superspace leads to interesting systems whose characteristics are not generally found in the one-form case. We mention, for example, the reducibility condition of the Abelian formulation. This means that the first-class constraints[18] related to the gauge symmetry are not all independent. In consequence, the covariant quantization of these theories leads to the presence of ghost-for-ghosts and many aux-

[a]Electronic mail: barcelos@if.ufrj.br
[b]Electronic mail: thibes@if.ufrj.br

 © 1998 American Institute of Physics

iliary fields.[15,19] The BRST superfield formulation naturally contains all these ingredients. More than that, it also naturally incorporates both BRST and anti-BRST symmetries and what does not happen when the superfield formulation is not used. We shall also see that the coefficient parameter of the gauge-fixing term also affects the ghost-for-ghost terms and the ones containing the corresponding auxiliary fields (a fact that is not present when the two symmetries are considered separately).

Concerning the non-Abelian formulation for tensor gauge fields, it is opportune to say that a complete theory, that can be accepted without restrictions, still does not exist. In fact, the obtainment of such a theory is not a simple task. Let us briefly explain why this occurs.

In the Abelian formulation, we have that the theory is invariant by the following gauge transformation of the (antisymmetric) tensor gauge field

$$\delta B_{\mu\nu}=\partial_\mu\xi_\nu-\partial_\nu\xi_\mu, \tag{I.1}$$

where ξ_μ is the characteristic parameter of the gauge transformation. As usual, it shows that not all components of the gauge potential $B_{\mu\nu}$ are independent. Two of them disappear by virtue of the massless condition. So, if all components of the vector gauge parameter ξ_μ were independent, the theory would exhibit zero degrees of freedom. However, this is nonetheless the case because the Abelian theory is reducible, in a sense that if one writes ξ_μ as $\xi_\mu=\partial_\mu\xi$, where ξ is some scalar quantity, Eq. (I.1) does not lead to any gauge transformation. In other words, the four components of ξ_μ are not independent and we can conclude that the tensor gauge field $B_{\mu\nu}$ has one physical degree of freedom. Incidentally, we also mention that it is precisely this degree of freedom which is used to give mass to the vector gauge in a coupled vector–tensor gauge theory[14] and the same number of physical degrees of freedom is kept after and before the mass generation, as it should be.

In the non-Abelian case, the corresponding gauge transformation reads

$$\delta B^a_{\mu\nu}=(D_\mu\xi_\nu)^a-(D_\nu\xi_\mu)^a. \tag{I.2}$$

Now, we observe that if one replaces the gauge parameter ξ^a_μ by $\xi^a_\mu=(D_\mu\xi)^a$, we get

$$\begin{aligned}\delta B^a_{\mu\nu}&=([D_\mu,D_\nu]\xi)^a\\&=gf^{abc}F^b_{\mu\nu}\xi^c\\&\equiv g(F_{\mu\nu}\times\xi)^a,\end{aligned} \tag{I.3}$$

where f^{abc} is the structure constant of the gauge symmetry group and g is a coupling parameter. The quantity $F^a_{\mu\nu}$ is the usual non-Abelian strength tensor for the vector field. We notice that the non-Abelian tensor theory would be reducible just in the case where $F^a_{\mu\nu}=0$ (on-shell), what can be obtained from the interesting model given by Freedman and Townsend.[20] We mention that the quantization of this model has been studied in a wide variety of ways.[21] Unfortunately, besides the fact of the model to be reducible just on-shell, the Abelian limit of this formulation does not correspond to the one given by Cremmer and Scherk and consequently it would not be compatible with a complete non-Abelian treatment.

An alternative way of circumventing these problems is by introducing a kind of a Stuckelberg field.[22] This permits us to have a non-Abelian theory with vector and tensor gauge fields coupled in a topological way which has an Abelian limit compatible with the Cremmer and Scherk paper.[23] Even though this non-Abelian theory is in general not reducible, the interesting point is that the Stuckelberg field disappears in the Abelian limit, what gives the same number of physical degrees of freedom in both sectors of the theory. We also mention that it is possible to obtain a reducible theory if the transformations involve vector and tensor fields in the same relation.[24] The superfield formalism of such an irreducible model has already been worked out by Baulieu and Thierry-Mieg,[25] but without identifying the antifields of the BV quantization scheme. We also address our attention to this problem. We take this opportunity to mention the Chapline–Manton model,[26] which is a kind of mixed case since it couples non-Abelian one-form gauge fields with Abelian two-form ones, where its superfield treatment was done by Baulieu.[27]

Our paper is organized as follows: In Sec. II, we briefly review the BRST and anti-BRST superspace formalism applied to the Yang–Mills fields. We develop this section in the way we

intend to use the formalism for the two-form case. In Sec. III, we consider two-form gauge fields, both Abelian and non-Abelian, from a geometrical point of view. We shall see that even though the concept of curvature is meaningless for the two-form case, it is possible to construct another object that makes the geometrical formulation consistent. The development of the theory in this way will permit us to introduce some ingredients that will be useful in the development of the next sections and we shall also see how the problems of the non-Abelian formulation can be circumvented, based on the ideas introduced in the paper of Ref. 23. In Sec. IV, we formulate the Abelian theory involving one- and two-form gauge fields, coupled in a topological way, by using the formalism developed in the two previous sections. In Sec. V we generalize the results of Sec. IV to the non-Abelian case through the introduction of the Stuckelberg field. Section VI is devoted to the Batalin–Vilkovisky quantization prescription, where we apply the technique of collective fields to obtain the BV action directly from the known superfields of the theory. We left Sec. VII for some concluding remarks and also included two appendices to contain details of the calculations.

II. BRIEF REVIEW OF THE BRST SUPERSPACE FORMALISM FOR THE YANG–MILLS THEORY

The development of the BRST and anti-BRST symmetries in the language of superfields is done in a space of four space–time coordinates x_μ and two Grassmannian ones, θ and $\bar{\theta}$, that we suppose to be real. We then introduce the following quantities (also real)

$$\mathscr{A}^a_\mu(x,\theta,\bar{\theta})=A^a_\mu(x)+i\theta A^{\theta a}_\mu(x)+i\bar{\theta}\bar{A}^{\bar{\theta}a}_\mu(x)+i\bar{\theta}\theta A^{\bar{\theta}\theta a}_\mu(x), \tag{II.1}$$

$$\mathscr{C}^a(x,\theta,\bar{\theta})=c^a(x)+\theta C^{\theta a}(x)+\bar{\theta}\bar{C}^{\bar{\theta}a}(x)+i\bar{\theta}\theta C^{\bar{\theta}\theta a}(x), \tag{II.2}$$

$$\bar{\mathscr{C}}^a(x,\theta,\bar{\theta})=\bar{c}^a(x)+\theta b^a(x)+\bar{\theta}\bar{b}^a(x)+i\bar{\theta}\theta\bar{C}^{\bar{\theta}\theta a}(x). \tag{II.3}$$

We associate the ghost numbers 0, 1, and -1, for the superfields $\mathscr{A}^a_\mu$, $\mathscr{C}^a$, and $\bar{\mathscr{C}}^a$, respectively. With these superfields, one constructs the following connection one-form.

$$\Gamma=-ig(dx^\mu\mathscr{A}_\mu+id\theta\mathscr{C}+id\bar{\theta}\bar{\mathscr{C}}), \tag{II.4}$$

where $\mathscr{A}_\mu=\mathscr{A}^a_\mu T^a$, $\mathscr{C}=\mathscr{C}^a T^a$, and $\bar{\mathscr{C}}=\bar{\mathscr{C}}^a T^a$. The quantities T^a are the matrix generators of the gauge symmetry group and the relative positions between $d\theta$ and $\mathscr{C}$, as well as between $d\bar{\theta}$ and $\mathscr{C}$, because we are using the left derivative convention with respect to the Grassmannian coordinates.

The curvature tensor is introduced by using its well-known definition,

$$\begin{aligned}F&=\mathscr{D}\Gamma\\&=d\Gamma+\Gamma\wedge\Gamma,\end{aligned} \tag{II.5}$$

where the exterior derivative d is here defined by

$$d=dx^\mu\partial_\mu+d\theta\partial_\theta+d\bar{\theta}\partial_{\bar{\theta}}. \tag{II.6}$$

The combination of (II.4)–(II.6) leads to

$$\begin{aligned}F=&-\frac{i}{2}g\ dx^\mu\wedge dx^\nu(\partial_\mu\mathscr{A}_\nu-\partial_\nu\mathscr{A}_\mu-ig[\mathscr{A}_\mu,\mathscr{A}_\nu])-ig\ dx^\mu\wedge d\theta(i\partial_\mu\mathscr{C}-\partial_\theta\mathscr{A}_\mu+g[\mathscr{A}_\mu,\mathscr{C}])\\&-ig\ dx^\mu\wedge d\bar{\theta}(i\partial_\mu\bar{\mathscr{C}}-\partial_{\bar{\theta}}\mathscr{A}_\mu+g[\mathscr{A}_\mu,\bar{\mathscr{C}}])-g\ d\theta\wedge d\theta(\partial_\theta\mathscr{C}+g\mathscr{C}\mathscr{C})-g\ d\bar{\theta}\wedge d\bar{\theta}\\&\times(\partial_{\bar{\theta}}\bar{\mathscr{C}}+g\bar{\mathscr{C}}\bar{\mathscr{C}})-g\ d\theta\wedge d\bar{\theta}(\partial_\theta\bar{\mathscr{C}}+\partial_{\bar{\theta}}\mathscr{C}+g\mathscr{C}\bar{\mathscr{C}}+g\bar{\mathscr{C}}\mathscr{C}).\end{aligned} \tag{II.7}$$

The crucial point of the superfield formulation of the BRST and anti-BRST symmetries is to assume the so-called horizontally condition, that is, the geometry is flat along the θ and $\bar{\theta}$ directions.[3,4] Imposing this condition in the curvature tensor above, we obtain

$$\begin{aligned}
A_\mu^{\theta a} &= (D_\mu c)^a, \\
A_\mu^{\bar{\theta} a} &= (D_\mu \bar{c})^a, \\
C^{\theta a} &= -\frac{i}{2} g(c \times c)^a, \\
\bar{b}^a &= -\frac{i}{2} g(\bar{c} \times \bar{c})^a, \\
C^{\bar{\theta} a} &= -b^a - ig(c \times \bar{c})^a, \\
C^{\bar{\theta}\theta a} &= g[c \times (b + igc \times \bar{c})]^a, \\
\bar{C}^{\bar{\theta}\theta} &= -g(b \times \bar{c})^a, \\
A_\mu^{\bar{\theta}\theta a} &= (D_\mu b)^a + ig(D_\mu c \times \bar{c})^a,
\end{aligned} \tag{II.8}$$

plus many other equations that one can show to be just consistency relations for the equations above (we shall present some details of a similar calculation for the tensor case).

Using the expressions (II.8) into (II.1)–(II.3), we obtain the superfield expressions

$$\mathscr{A}_\mu = A_\mu + i\theta D_\mu c + i\bar{\theta} D_\mu \bar{c} + i\bar{\theta}\theta [D_\mu b + ig(D_\mu c) \times \bar{c}], \tag{II.9}$$

$$\mathscr{C} = c - \frac{i}{2} g\theta c \times c - \bar{\theta}(b + igc \times \bar{c}) + ig\bar{\theta}\theta c \times (b + igc \times \bar{c}), \tag{II.10}$$

$$\bar{\mathscr{C}} = \bar{c} + \theta b - \frac{i}{2} g\bar{\theta}\bar{c} \times \bar{c} - ig\bar{\theta}\theta b \times \bar{c}. \tag{II.11}$$

Let us now see how the BRST and anti-BRST symmetries are embodied in the superfields above. We generically denote by Φ any superfield and suppose it has the following general form

$$\Phi(x, \theta, \bar{\theta}) = A(x) + \theta B(x) + \bar{\theta}\bar{B}(x) + i\bar{\theta}\theta D(x). \tag{II.12}$$

Under translations λ and $\bar{\lambda}$ of the coordinates θ and $\bar{\theta}$, respectively, we obtain that the superfield Φ transforms as

$$\delta_\lambda \Phi = \lambda B + i\bar{\theta}\lambda D, \tag{II.13}$$

$$\delta_{\bar{\lambda}} \Phi = \bar{\lambda}\bar{B} - i\theta\bar{\lambda} D. \tag{II.14}$$

This leads to the following transformations for the component fields

$$\begin{aligned}
\delta_\lambda A &= \lambda B, \\
\delta_\lambda B &= 0, \\
\delta_\lambda \bar{B} &= i\lambda D, \\
\delta_\lambda D &= 0,
\end{aligned} \tag{II.15}$$

and

$$\delta_{\bar{\lambda}} A = \overline{\lambda B},$$

$$\delta_{\bar{\lambda}} B = -i\overline{\lambda D},$$

$$\delta_{\bar{\lambda}} \bar{B} = 0, \tag{II.16}$$

$$\delta_{\bar{\lambda}} D = 0.$$

In this superspace formalism, (II.15) and (II.16) are identified as the BRST and anti-BRST transformations for the component fields. In fact, considering the superfields given by (II.9)–(II.11), we obtain

$$\delta_\lambda A_\mu = i\lambda D_\mu c,$$

$$\delta_\lambda c = -\frac{i}{2} g\lambda c \times c,$$

$$\delta_\lambda \bar{c} = \lambda b, \tag{II.17}$$

$$\delta_\lambda b = 0,$$

and

$$\delta_{\bar{\lambda}} A_\mu = i\overline{\lambda} D_\mu \bar{c},$$

$$\delta_{\bar{\lambda}} c = -\overline{\lambda}(b + igc \times c),$$

$$\delta_{\bar{\lambda}} \bar{c} = \frac{i}{2} g\overline{\lambda c} \times \bar{c}, \tag{II.18}$$

$$\delta_{\bar{\lambda}} b = ig\overline{\lambda} b \times \bar{c}.$$

that are precisely the well-known BRST and anti-BRST transformations for the Yang–Mills fields.

In order to construct invariant Lagrangians under BRST and anti-BRST transformations, we look at the set of transformations (II.15) and (II.16). We then notice, for example, that this is the case when the Lagrangian is the $\theta\theta$ component of a superfield. Another example is any Lagrangian that is a singlet (a superfield independent of θ and $\bar{\theta}$). There are two more possibilities: when the Lagrangian is the θ or the $\bar{\theta}$ component of a superfield without the $\theta\theta$ term. Besides these requirements, there are others that have to be taken into account in order to obtain invariant Lagrangians. For example, the mass dimension ought to be less or equal to four for the question of renormalizeability and the ghost and parity numbers ought to be zero.

In the case of the Yang–Mills theory, one directly sees that

$$\mathscr{L}_I = \text{const tr } \mathscr{F}_{\mu\nu}\mathscr{F}^{\mu\nu} \tag{II.19}$$

is a singlet, where $\mathscr{F}_{\mu\nu}$ is a superfield defined by $\mathscr{F}_{\mu\nu} = \partial_\mu \mathscr{A}_\nu - \partial_\nu \mathscr{A}_\mu + g\mathscr{A}_\mu \times \mathscr{A}_\nu$ (we point out that the product of two superfields is a superfield also). This singlet corresponds to the usual term of the Yang–Mills theory, $-\frac{1}{4}$ tr $F_{\mu\nu}F^{\mu\nu}$.

Considering the superfields $\mathscr{C}$ and $\bar{\mathscr{C}}$, the following combinations have parities and ghost numbers zero: tr $\partial_\theta \bar{c} \partial_\theta \bar{c}$, tr $\partial_{\bar{\theta}} c \partial_{\bar{\theta}} c$, tr $\partial_\theta \bar{c} \partial_{\bar{\theta}} c$, and tr $\partial_{\bar{\theta}} \bar{c} \partial_\theta c$. One can see that the first one is a singlet related to tr bb. The second one is a superfield without the $\theta\theta$ component, so its term in $\bar{\theta}$ is invariant. However, this component cannot be used as a Lagrangian because its ghost and parity numbers are not zero. The last two terms are complete superfields. Consequently, their $\theta\theta$ components are also invariant. Thus the following Lagrangians are possible

$$\mathscr{L}_{II}^{(1)} = \text{const tr } \partial_\theta \bar{c} \partial_\theta \bar{c},$$

$$\mathscr{L}_{II}^{(2)} = \text{const tr } \partial_\theta \partial_{\bar{\theta}} (\partial_\theta \bar{c} \partial_{\bar{\theta}} c), \tag{II.20}$$

$$\mathscr{L}_{II}^{(3)} = \text{const tr } \partial_\theta \partial_{\bar{\theta}} (\partial_{\bar{\theta}} \bar{c} \partial_\theta c).$$

All of them are related to gauge-fixing Lagrangians. The first one corresponds to the usual Lorentz term. The other ones are more complicated gauge-fixing conditions.

Finally, the $\bar{\theta}\theta$ component of the superfield $\mathscr{A}_\mu \mathscr{A}^\mu$ is the term that is lacking in this formulation. In fact,

$$\mathscr{L}_{III} = \frac{i}{2} \partial_\theta \partial_{\bar{\theta}} \text{tr}(\mathscr{A}_\mu \mathscr{A}^\mu)$$

$$= 2 \text{ tr}(\partial_\mu \bar{c} D^\mu c + i A_\mu \partial^\mu b). \tag{II.21}$$

III. GEOMETRIC TREATMENT FOR ABELIAN AND NON-ABELIAN VECTOR–TENSOR GAUGE THEORIES

Let us first consider the Abelian case. Remembering, for one-form gauge field the connection one-form reads (for question of simplicity, we make the coupling parameters equal to one from now on)

$$\Gamma = -iA_\mu dx^\mu, \tag{III.1}$$

and the curvature in the Abelian case is just obtained by taking

$$F = d\Gamma$$

$$= -\frac{i}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu. \tag{III.2}$$

For the two-form gauge fields we similarly write

$$\Lambda = -\frac{i}{2} B_{\mu\nu} dx^\mu \wedge dx^\nu. \tag{III.3}$$

Of course, this quantity cannot be considered as a connection. But, in spite of this, we define a geometric quantity H in the same way that F was done, i.e.,

$$H = d\Lambda. \tag{III.4}$$

We emphasize that even though H has been defined similarly to F, it cannot be identified as a curvature tensor because Λ is not a connection. We also see that this procedure may be extended for gauge tensors of any rank. The motivation of introducing H by the above definition is that a similar Lagrangian to the vector gauge field, namely

$$\mathscr{L} = -\tfrac{1}{2} * (H * H), \tag{III.5}$$

leads to the well-known Lagrangian introduced by Kalb and Ramond[12]

$$\mathscr{L} = \tfrac{1}{12} H_{\mu\nu\rho} H^{\mu\nu\rho}, \tag{III.6}$$

where $H_{\mu\nu\rho}$ is defined by $H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \partial_\rho B_{\mu\nu} + \partial_\nu B_{\rho\mu}$.

The gauge transformations for these objects are

$$\delta_V \Gamma = -d\epsilon,$$

$$\delta_V F = 0, \tag{III.7}$$

$$\delta_V\Lambda=0,$$

$$\delta_V H=0,$$

and

$$\begin{aligned}\delta_T\Gamma&=0,\\ \delta_T F&=0,\\ \delta_T\Lambda&=-d\xi,\\ \delta_T H&=0.\end{aligned}\tag{III.8}$$

The subscripts V and T stand for the vector and tensor gauge transformations. The quantity ξ is a one-form parameter related to the quantity introduced in Eq. (I.1).

Finally, we mention that the gauge invariant Lagrangian corresponding to the Cremmer and Scherk theory is given by[13]

$$\mathscr{L}=\tfrac{1}{2}*(F*F)-\tfrac{1}{2}*(H*H)-4m*(F\wedge\Lambda).\tag{III.9}$$

We now consider the non-Abelian case. For the one- and two-form gauge fields we have

$$\Gamma=-iT^aA^a_\mu dx^\mu,\tag{III.10}$$

$$\Lambda=-\frac{i}{2}\,T^aB^a_{\mu\nu}dx^\mu\wedge dx^\nu.\tag{III.11}$$

Now, the vector and tensor gauge transformations read

$$\begin{aligned}\delta_V\Gamma&=-D\epsilon\\ &=-d\epsilon-\Gamma\epsilon+\epsilon\Gamma,\\ \delta_V\Lambda&=\epsilon\Lambda-\Lambda\epsilon,\\ \delta_T\Gamma&=0,\\ \delta_T\Lambda&=-D\xi\\ &=-d\xi-\Gamma\wedge\xi-\xi\wedge\Gamma.\end{aligned}\tag{III.12}$$

We have considered that Λ transforms as a vector under the vector gauge transformations. This is so because tensor gauge fields are also charged in the color space. The curvature related to the vector gauge field is given by the covariant derivative of the connection and we consider, suggested by the Abelian case, that the corresponding quantity for the two-form gauge field is given in the same way, that is to say, by taking the covariant derivative over Λ. Using the transformations above, we get

$$\begin{aligned}\delta_V F&=\epsilon F-F\epsilon,\\ \delta_V H&=\epsilon H-H\epsilon,\\ \delta_T F&=0,\\ \delta_T H&=\xi\wedge F-F\wedge\xi.\end{aligned}\tag{III.13}$$

We observe from the last relation (III.13) that H is not a geometrical object under the tensor gauge transformation (it is just in the Abelian limit). Thus a Lagrangian like $-\frac{1}{2}*(H*H)$ is just invariant

for the vector gauge transformation but is not for the tensor one. Further, it is not difficult to conclude that there is no other Lagrangian we can construct by using H that can be invariant under the tensor gauge transformation.

The way we envisage having a gauge invariant Lagrangian is by redefining Λ conveniently as[28]

$$\hat{\Lambda}=\Lambda+D\Omega, \tag{III.14}$$

where Ω is a kind of Stuckelberg field one-form. Since it is also charged, we consider that it is also a vector under the vector gauge transformation, namely,

$$\delta_V\Omega=\epsilon\Omega-\Omega\epsilon. \tag{III.15}$$

If one considers

$$\delta_T\Omega=\xi, \tag{III.16}$$

we obtain

$$\delta_T\hat{\Lambda}=0. \tag{III.17}$$

Consequently,

$$\begin{aligned}\delta_T\hat{H}&=D(\delta_T\hat{\Lambda})\\&=0.\end{aligned} \tag{III.18}$$

In this case, a Lagrangian like

$$\mathscr{L}=-\tfrac{1}{2}\ \mathrm{tr}^*(\hat{H}^*\hat{H}) \tag{III.19}$$

is invariant for both vector and tensor gauge transformations. It is opportune to say that the present non-Abelian formulation is still not reducible, but we emphasize that there is no problem with respect of the number of degrees of freedom in the Abelian and non-Abelian sectors. This is so by virtue of the Stuckelberg field that does not appear in the Abelian limit because

$$\begin{aligned}\hat{H}&=d\hat{\Lambda}\\&=d\Lambda+dd\Omega\\&=H.\end{aligned} \tag{III.20}$$

For the Chern–Simons (CS) Lagrangian, we have that the introduction of the Stuckelberg field is not necessary to obtain a gauge invariant Lagrangian. One can directly see that

$$\mathscr{L}=-4m\ \mathrm{tr}^*(F\wedge\Lambda). \tag{III.21}$$

is invariant for both vector and tensor gauge transformations. Incidentally we mention that this geometric formulation has permitted us to obtain a simpler CS Lagrangian comparing with the one of Ref. 23.

Putting all these terms together, we write down the complete form of the gauge invariant Lagrangian for a vector–tensor non-Abelian theory

$$\mathscr{L}=\tfrac{1}{2}\ \mathrm{tr}^*(F^*F)-\tfrac{1}{2}\ \mathrm{tr}^*(\hat{H}^*\hat{H})-4m\ \mathrm{tr}^*(F\wedge\Lambda). \tag{III.22}$$

IV. BRST SUPERSPACE FOR THE ABELIAN TWO-FORM GAUGE FIELDS

Our goal is to use the formalism discussed in Sec. II for two-form gauge fields. It might be instructive to consider the Abelian case first. This is not only because it is simpler than the non-Abelian one, but because it has the reducibility property that is not found in the non-Abelian counterpart. Let us then introduce the two-form quantity

$$\Lambda = -i(dx^\mu \wedge dx^\nu \mathscr{B}_{\mu\nu} + d\theta \wedge dx^\mu \mathscr{G}_\mu + d\bar{\theta} \wedge dx^\mu \bar{\mathscr{G}}_\mu + d\theta \wedge d\theta \mathscr{G} + d\bar{\theta} \wedge d\bar{\theta} \bar{\mathscr{G}} + d\theta \wedge d\bar{\theta} \mathscr{K}). \tag{IV.1}$$

We identify the following ghost numbers for the quantities that appear in the equation above

$$\begin{gathered} \mathrm{gh}[\mathscr{B}_{\mu\nu}] = 0, \quad \mathrm{gh}[\mathscr{G}_\mu] = 1, \quad \mathrm{gh}[\bar{\mathscr{G}}_\mu] = -1, \\ \mathrm{gh}[\mathscr{G}] = 2, \quad \mathrm{gh}[\bar{\mathscr{G}}] = -2, \quad \mathrm{gh}[\mathscr{K}] = 0. \end{gathered} \tag{IV.2}$$

In terms of component fields, we may conveniently write

$$\begin{aligned} \mathscr{B}_{\mu\nu} &= B_{\mu\nu} + i\theta B^\theta_{\mu\nu} + i\bar{\theta} B^{\bar{\theta}}_{\mu\nu} + i\bar{\theta}\theta B^{\bar{\theta}\theta}_{\mu\nu}, \\ \mathscr{G}_\mu &= g_\mu + \theta \bar{e}_\mu + \bar{\theta} e_\mu + i\bar{\theta}\theta G^{\bar{\theta}\theta}_\mu, \\ \bar{\mathscr{G}}_\mu &= \bar{g}_\mu + \theta \bar{G}^\theta_\mu + \bar{\theta} G^{\bar{\theta}}_\mu + i\bar{\theta}\theta \bar{G}^{\bar{\theta}\theta}_\mu, \\ \mathscr{G} &= G + i\theta G^\theta + i\bar{\theta} G^{\bar{\theta}} + i\bar{\theta}\theta G^{\bar{\theta}\theta}, \\ \bar{\mathscr{G}} &= \bar{G} + i\theta \bar{G}^\theta + i\bar{\theta} \bar{G}^{\bar{\theta}} + i\bar{\theta}\theta \bar{G}^{\bar{\theta}\theta}, \\ \mathscr{K} &= K + i\theta f + i\bar{\theta} \bar{f} + i\bar{\theta}\theta K^{\bar{\theta}\theta}. \end{aligned} \tag{IV.3}$$

We now define a three-form H as it was done in the previous section, i.e.,

$$H = d\Lambda. \tag{IV.4}$$

Introducing (IV.1) into the expression above and also imposing that the components along θ and $\bar{\theta}$ are zero, we obtain the equations:

$$\begin{gathered} \partial_\theta \mathscr{B}_{\mu\nu} - \tfrac{1}{2}(\partial_\mu \mathscr{G}_\nu - \partial_\nu \mathscr{G}_\mu) = 0, \\ \partial_{\bar{\theta}} \mathscr{B}_{\mu\nu} - \tfrac{1}{2}(\partial_\mu \bar{\mathscr{G}}_\nu - \partial_\nu \bar{\mathscr{G}}_\mu) = 0, \\ \partial_\theta \mathscr{G}_\mu - \partial_\mu \mathscr{G} = 0, \\ \partial_{\bar{\theta}} \bar{\mathscr{G}}_\mu - \partial_\mu \bar{\mathscr{G}} = 0, \\ \partial_\mu \mathscr{K} - \partial_\theta \bar{\mathscr{G}}_\mu - \partial_{\bar{\theta}} \mathscr{G}_\mu = 0, \\ \partial_\theta \mathscr{G} = 0, \\ \partial_{\bar{\theta}} \bar{\mathscr{G}} = 0, \\ \partial_\theta \bar{\mathscr{G}} + \partial_{\bar{\theta}} \mathscr{K} = 0, \\ \partial_{\bar{\theta}} \mathscr{G} + \partial_\theta \mathscr{K} = 0. \end{gathered} \tag{IV.5}$$

Considering the expansion expressions given by (IV.3), we obtain

$$
\begin{aligned}
B^{\theta}_{\mu\nu} &= -\frac{i}{2}(\partial_\mu g_\nu - \partial_\nu g_\mu),\\
B^{\bar{\theta}}_{\mu\nu} &= -\frac{i}{2}(\partial_\mu \bar{g}_\nu - \partial_\nu \bar{g}_\mu),\\
B^{\bar{\theta}\theta}_{\mu\nu} &= -\frac{i}{2}(\partial_\mu e_\nu - \partial_\nu e_\mu),\\
\bar{e}_\mu &= \partial_\mu G,\\
\bar{G}^{\theta}_{\mu} &= -e_\mu + \partial_\mu K,\\
\bar{G}^{\bar{\theta}}_{\mu} &= \partial_\mu \bar{G},\\
G^{\bar{\theta}\theta}_{\mu} &= \partial_\mu f,\\
\bar{G}^{\bar{\theta}\theta}_{\mu} &= -\partial_\mu \bar{f},\\
G^{\theta} &= 0,\\
G^{\bar{\theta}} &= -f,\\
\bar{G}^{\theta} &= -\bar{f},\\
\bar{G}^{\bar{\theta}} &= 0,\\
G^{\bar{\theta}\theta} &= 0,\\
\bar{G}^{\bar{\theta}\theta} &= 0,\\
K^{\bar{\theta}\theta} &= 0,
\end{aligned}
\tag{IV.6}
$$

plus many other equations that are just consistency relations for the results above (see Appendix A). The combination of (IV.3) and (IV.6) gives the superfield expressions

$$
\begin{aligned}
\mathscr{B}_{\mu\nu} &= B_{\mu\nu} + \tfrac{1}{2}\theta(\partial_\mu g_\nu - \partial_\nu g_\mu) + \tfrac{1}{2}\bar{\theta}(\partial_\mu \bar{g}_\nu - \partial_\nu \bar{g}_\mu) - \tfrac{1}{2}\bar{\theta}\theta(\partial_\mu e_\nu - \partial_\nu e_\mu),\\
\mathscr{G}_\mu &= g_\mu + \theta\partial_\mu G + \bar{\theta} e_\mu + i\bar{\theta}\theta\partial_\mu f,\\
\bar{\mathscr{G}}_\mu &= \bar{g}_\mu + \theta(\partial_\mu K - e_\mu) + \bar{\theta}\partial_\mu \bar{G} - i\bar{\theta}\theta\partial_\mu \bar{f},\\
\mathscr{G} &= G - \bar{\theta} f,\\
\bar{\mathscr{G}} &= \bar{G} - \theta \bar{f},\\
\mathscr{K} &= K + i\theta f + i\bar{\theta}\bar{f}.
\end{aligned}
\tag{IV.7}
$$

The fields that appear in the expressions above are the net fields of the theory. We notice that they appear naturally in BRST superspace formalism. It would be instructive to compare them with conventional treatments,[19] where a lot of careful arguments have to be taken into account in order to introduce all of them.

We list below the ghost and parity numbers for these fields.

$$\begin{gathered}
\mathrm{gh}[B_{\mu\nu},e_\mu,K]=0,\\
\mathrm{gh}[g_\mu,f]=1,\\
\mathrm{gh}[\bar{g}_\mu,\bar{f}]=-1,\\
\mathrm{gh}[G]=2,\\
\mathrm{gh}[\bar{G}]=-2,\\
\epsilon[B_{\mu\nu},e_\mu,K,G,\bar{G}]=0,\\
\epsilon[g_\mu,\bar{g}_\mu,f,\bar{f}]=1.
\end{gathered} \tag{IV.8}$$

Considering expressions (II.15), (II.16), and (IV.7), we also naturally obtain the BRST and anti-BRST transformations for these fields

$$\begin{gathered}
\delta_\lambda B_{\mu\nu}=\tfrac{1}{2}\lambda(\partial_\mu g_\nu-\partial_\nu g_\mu),\\
\delta_\lambda g_\mu=\lambda\partial_\mu G,\\
\delta_\lambda \bar{g}_\mu=\lambda(\partial_\mu K-e_\mu),\\
\delta_\lambda G=0,\\
\delta_\lambda \bar{G}=-\lambda\bar{f},\\
\delta_\lambda K=i\lambda f,\\
\delta_\lambda e_\mu=-\lambda\partial_\mu f,\\
\delta_\lambda f=0,\\
\delta_\lambda \bar{f}=0,
\end{gathered} \tag{IV.9}$$

and

$$\begin{gathered}
\delta_{\bar{\lambda}} B_{\mu\nu}=\tfrac{1}{2}\bar{\lambda}(\partial_\mu \bar{g}_\nu-\partial_\nu \bar{g}_\mu),\\
\delta_{\bar{\lambda}} g_\mu=\bar{\lambda}e_\mu,\\
\delta_{\bar{\lambda}} \bar{g}_\mu=\bar{\lambda}\partial_\mu \bar{G},\\
\delta_{\bar{\lambda}} G=-\bar{\lambda}f,\\
\delta_{\bar{\lambda}} \bar{G}=0,\\
\delta_{\bar{\lambda}} K=i\bar{\lambda}\bar{f},\\
\delta_{\bar{\lambda}} e_\mu=0,\\
\delta_{\bar{\lambda}} f=0,\\
\delta_{\bar{\lambda}} \bar{f}=0.
\end{gathered} \tag{IV.10}$$

We observe that contrary to the Abelian vector formulation, the BRST and anti-BRST transformations for the tensor case are not symmetric.

To obtain invariant Lagrangians under these transformations, we follow the same steps of the vector treatment. First, one can directly show that $\mathscr{H}_{\mu\nu\rho}\mathscr{H}^{\mu\nu\rho}$ is a singlet, where $\mathscr{H}_{\mu\nu\rho}$ is defined by

$$\mathscr{H}_{\mu\nu\rho}=\partial_\mu\mathscr{B}_{\nu\rho}+\partial_\rho\mathscr{B}_{\mu\nu}+\partial_\nu\mathscr{B}_{\rho\mu}. \tag{IV.11}$$

So, a term involving the product above is both BRST and anti-BRST invariant. We then write[12]

$$\mathscr{L}_I=\tfrac{1}{12}\mathscr{H}_{\mu\nu\rho}\mathscr{H}^{\mu\nu\rho}. \tag{IV.12}$$

Considering that the $\bar{\theta}\theta$ component of a superfield is BRST and anti-BRST invariant, we conveniently write another invariant Lagrangian

$$\mathscr{L}_{II}=\tfrac{1}{2}\partial_\theta\partial_{\bar{\theta}}\mathscr{B}_{\mu\nu}\mathscr{B}^{\mu\nu}, \tag{IV.13}$$

that in components fields corresponds to

$$\mathscr{L}_{II}=-B_{\mu\nu}\partial^\mu e^\nu-\tfrac{1}{2}(\partial_\mu\bar{g}_\nu-\partial_\nu\bar{g}_\mu)\partial^\mu g^\nu. \tag{IV.14}$$

The gauge-fixing term is given by

$$\mathscr{L}_{III}=\beta\partial_\theta\partial_{\bar{\theta}}\bar{\mathscr{G}}_\mu\mathscr{G}^\mu. \tag{IV.15}$$

In components, we have

$$\mathscr{L}_{III}=\beta[i\bar{g}_\mu\partial^\mu f+(e_\mu-\partial_\mu K)e^\mu+\partial_\mu\bar{G}\partial^\mu G-i\partial_\mu\bar{f}g^\mu]. \tag{IV.16}$$

We observe that the gauge-fixing term related to the auxiliary field, namely $e_\mu e^\mu$, does not appear isolated. This would occur just for the anti-BRST symmetry. It is also important to emphasize that the ghost-for-ghost term also depends on the gauge-fixing parameter β. We mention that in the cases treated in Refs. 19 and 15 this does not occur because just one of these two symmetries was considered.

It is just a matter of algebraic calculation to show that other invariant terms we can construct with the superfields (IV.7), that have parity and ghost numbers equal to zero, do not lead to any new independent result.

Finally we deal with the topological term. One can directly show that the combination $\epsilon^{\mu\nu\rho\kappa}\partial_\mu\mathscr{A}_\nu\mathscr{B}_{\rho\kappa}$ is a singlet in a sense that the terms along θ and $\bar{\theta}$ are total derivatives. So the topological Lagrangian reads

$$\mathscr{L}_{IV}=-m\epsilon^{\mu\nu\rho\kappa}\mathscr{T}_{\mu\nu}\mathscr{B}_{\rho\kappa}. \tag{IV.17}$$

V. THE NON-ABELIAN CASE

For the non-Abelian case, we start from the two-form quantity (IV.3), but now considering that

$$\mathscr{B}_{\mu\nu}=\mathscr{B}^a_{\mu\nu}T^a,$$

$$\mathscr{G}_\mu=\mathscr{G}^a_\mu T^a, \quad \text{etc.}, \tag{V.1}$$

and defining the three-form H by taking the covariant derivative over Λ, i.e.,

$$\begin{aligned} H&=\mathscr{D}\Lambda \\ &=d\Lambda+\Gamma\wedge\Lambda-\Lambda\wedge\Gamma, \end{aligned} \tag{V.2}$$

where Γ is the connection one-form given by (II.4). The requirement that the components of H along θ and $\bar{\theta}$ directions are zero leads us to the equations

$$
\begin{gathered}
\partial_\theta \mathscr{B}_{\mu\nu} - \tfrac{1}{2}(\mathscr{D}_\mu \mathscr{G}_\nu - \mathscr{D}_\nu \mathscr{G}_\mu) + i\mathscr{C} \times \mathscr{B}_{\mu\nu} = 0, \\
\partial_{\bar\theta} \mathscr{B}_{\mu\nu} - \tfrac{1}{2}(\mathscr{D}_\mu \bar{\mathscr{G}}_\nu - \mathscr{D}_\nu \bar{\mathscr{G}}_\mu) + i\bar{\mathscr{C}} \times \mathscr{B}_{\mu\nu} = 0, \\
\partial_\theta \mathscr{G}_\mu - \mathscr{D}_\mu \mathscr{G} + i\mathscr{C} \times \mathscr{G}_\mu = 0, \\
\partial_{\bar\theta} \bar{\mathscr{G}}_\mu - \mathscr{D}_\mu \bar{\mathscr{G}} + i\bar{\mathscr{C}} \times \bar{\mathscr{G}}_\mu = 0, \\
\partial_\theta \bar{\mathscr{G}}_\mu + \partial_{\bar\theta} \mathscr{G}_\mu - \mathscr{D}_\mu \mathscr{K} + i\mathscr{C} \times \bar{G}_\mu + i\bar{\mathscr{C}} \times \mathscr{G}_\mu = 0, \\
\partial_\theta \mathscr{G} + i\mathscr{C} \times \mathscr{G} = 0, \\
\partial_{\bar\theta} \bar{\mathscr{G}} + i\bar{\mathscr{C}} \times \bar{\mathscr{G}} = 0, \\
\partial_\theta \bar{\mathscr{G}} + \partial_{\bar\theta} \mathscr{K} + i\mathscr{C} \times \bar{\mathscr{G}} + i\bar{\mathscr{C}} \times \mathscr{K} = 0, \\
\partial_{\bar\theta} \mathscr{G} + \partial_\theta \mathscr{K} + i\bar{\mathscr{C}} \times \mathscr{G} + i\mathscr{C} \times \mathscr{K} = 0.
\end{gathered}
\tag{V.3}
$$

We reinforce the fact that we are making the coupling constants equal to one. Making similar developments as in the previous section, we obtain similar relations also involving the component fields plus other equations. But here an interesting fact occurs. These last equations are not all consistency relations as in the Abelian case. Instead, they permit us to conclude that the entire superfields $\mathscr{G}$, $\bar{\mathscr{G}}$, and $\mathscr{K}$ are zero (see Appendix B). This shows the consistency of the formalism we are using because the non-Abelian version of the theory we are dealing with is not reducible, and it is not necessary with the presence of these terms here.

The final form of the superfields for the non-Abelian case reads

$$
\begin{aligned}
\mathscr{B}_{\mu\nu} = {} & B_{\mu\nu} + \theta(\tfrac{1}{2} D_\mu g_\nu - \tfrac{1}{2} D_\nu g_\mu - ic \times B_{\mu\nu}) + \bar\theta(\tfrac{1}{2} D_\mu \bar g_\nu - \tfrac{1}{2} D_\nu \bar g_\mu - i\bar c \times B_{\mu\nu}) \\
& + \bar\theta\theta\Big[-\frac{1}{2} D_\mu e_\nu + \frac{1}{2} D_\nu e_\mu - \frac{i}{2} c \times (D_\mu \bar g_\nu - D_\nu \bar g_\mu) - \frac{i}{2} (D_\mu \bar c) \times g_\nu \\
& + \frac{i}{2} (D_\nu \bar c) \times g_\mu - ib \times B_{\mu\nu} - \bar c \times (c \times B_{\mu\nu}) \Big],
\end{aligned}
\tag{V.4}
$$

$$
\mathscr{G}_\mu = g_\mu - i\theta c \times g_\mu + \bar\theta e_\mu + \bar\theta\theta[(c \times \bar c) \times g_\mu - ic \times e_\mu - ib \times g_\mu],
\tag{V.5}
$$

$$
\begin{aligned}
\bar{\mathscr{G}}_\mu = {} & \bar g_\mu - \theta(e_\mu + ic \times \bar g_\mu + i\bar c \times g_\mu) - i\bar\theta\bar c \times \bar g_\mu \\
& + \bar\theta\theta[\bar c \times (\bar c \times g_\mu) + \bar c \times (c \times \bar g_\mu) - ib \times \bar g_\mu - i\bar c \times e_\mu].
\end{aligned}
\tag{V.6}
$$

We stress the fact that, as can be seen in Appendix B, to arrive at the solution (V.4)–(V.6) above, one must use the previous results for the one-form gauge field (II.8). The BRST and anti-BRST transformations for the non-Abelian theory read

$$
\begin{gathered}
\delta_\lambda B_{\mu\nu} = \lambda(\tfrac{1}{2} D_\mu g_\nu - \tfrac{1}{2} D_\nu g_\mu - ic \times B_{\mu\nu}), \\
\delta_\lambda g_\mu = -i\lambda c \times g_\mu, \\
\delta_\lambda e_\mu = \lambda[(c \times \bar c) \times g_\mu - i\bar c \times e_\mu - ib \times g_\mu], \\
\delta_\lambda \bar g_\mu = -\lambda(e_\mu + ic \times \bar g_\mu + i\bar c \times g_\mu),
\end{gathered}
\tag{V.7}
$$

and

$$\begin{aligned}
\delta_{\bar\lambda} B_{\mu\nu} &= \bar\lambda\left(\tfrac{1}{2} D_\mu \bar g_\nu - \tfrac{1}{2} D_\nu \bar g_\mu - i\bar c \times B_{\mu\nu}\right),\\
\delta_{\bar\lambda} g_\mu &= \bar\lambda e_\mu,\\
\delta_{\bar\lambda} e_\mu &= 0,\\
\delta_{\bar\lambda} \bar g_\mu &= -i\bar\lambda \bar c \times \bar g_\mu .
\end{aligned} \tag{V.8}$$

Let us finally discuss the obtainment of invariant Lagrangians under BRST and anti-BRST transformations for the non-Abelian tensor theory. First we have that a term that would correspond to $\mathscr{L}_I$ of the Abelian case, namely tr $\mathscr{H}_{\mu\nu\rho}\mathscr{H}^{\mu\nu\rho}$ is not a singlet here. So, it cannot be directly used as an invariant Lagrangian. We shall discuss this term at the end of the section. Let us consider the other terms first.

The gauge-fixing Lagrangians are

$$\begin{aligned}
\mathscr{L}_{II} &= \tfrac{1}{2}\,\partial_\theta \partial_{\bar\theta}\,\mathrm{tr}\,\mathscr{B}_{\mu\nu}\mathscr{B}^{\mu\nu}\\
&= -\,\mathrm{tr}\,B_{\mu\nu} D^\mu (e^\nu + i\bar c \times g^\nu) - \tfrac{1}{2}\,\mathrm{tr}(D_\mu \bar g_\nu - D_\nu \bar g_\mu) D^\mu \bar g^\nu,
\end{aligned} \tag{V.9}$$

$$\begin{aligned}
\mathscr{L}_{III} &= \beta\,\partial_\theta \partial_{\bar\theta} \bar{\mathscr{G}}_\mu \mathscr{G}^\mu\\
&= \beta\,\mathrm{tr}(e_\mu + i\bar c \times g_\mu)(e^\mu + i\bar c \times g^\mu).
\end{aligned} \tag{V.10}$$

Concerning the topological Lagrangian, we have that the quantity $\epsilon^{\mu\nu\rho\kappa}\mathscr{F}_{\mu\nu}\mathscr{B}_{\rho\kappa}$ is a singlet, in a sense that the components along θ and $\bar\theta$ are total derivatives. Thus the topological Lagrangian reads

$$\begin{aligned}
\mathscr{L}_{IV} &= -\,m\,\epsilon^{\mu\nu\rho\kappa}\,\mathrm{tr}\,\mathscr{F}_{\mu\nu}\mathscr{B}_{\rho\kappa}\\
&= -\,m\,\epsilon^{\mu\nu\rho\kappa}\,\mathrm{tr}\,F_{\mu\nu}B_{\rho\kappa}.
\end{aligned} \tag{V.11}$$

Let us finally consider the $\mathscr{L}_I$ Lagrangian. As it was told in the beginning, the quantity tr $\mathscr{H}_{\mu\nu\rho}\mathscr{H}^{\mu\nu\rho}$ is not a singlet and consequently, it cannot be directly considered as an invariant Lagrangian. Considering what was developed in Sec. III, the way we have to obtain an invariant Lagrangian is introducing a kind of Stuckelberg field. Here we consider the superfield $\hat{\mathscr{H}}_{\mu\nu\rho}$ defined by

$$\tfrac{1}{3}\hat{\mathscr{H}} = \mathscr{D}_{[\mu}\mathscr{B}_{\nu\rho]} + \mathscr{D}_{[\mu}\mathscr{D}_\nu \mathscr{S}_{\rho]}, \tag{V.12}$$

where in order to simplify the notation we are using the subscript $[\cdots]$ to denote antisymmetry among the indices. Here $\mathscr{S}_\mu$ is a (Stuckelberg) superfield that has the purpose to make $\hat{\mathscr{H}}_{\mu\nu\rho}\hat{\mathscr{H}}^{\mu\nu\rho}$ a singlet. This is achieved after a long algebraic calculation and we conclude that the superfield $\mathscr{S}_\mu$ must be

$$\begin{aligned}
\mathscr{S}_\mu &= \Omega_\mu - \theta(g_\mu + ic \times \Omega_\mu) - \bar\theta(\bar g_\mu + i\bar c \times \Omega_\mu)\\
&\quad + \bar\theta\theta[e_\mu + c \times (i\bar g_\mu - \bar c \times \Omega_\mu) - i(b + c \times \bar c) \times \Omega_\mu].
\end{aligned} \tag{V.13}$$

The BRST and anti-BRST invariant $\mathscr{L}_I$ Lagrangian then reads

$$\begin{aligned}
\mathscr{L}_I &= \tfrac{1}{12}\,\mathrm{tr}\,\hat{\mathscr{H}}_{\mu\nu\rho}\hat{\mathscr{H}}^{\mu\nu\rho}\\
&= \tfrac{3}{4}\,\mathrm{tr}(D_{[\mu}B_{\nu\rho]} + D_{[\mu}D_\nu \Omega_{\rho]})^2.
\end{aligned} \tag{V.14}$$

VI. THE FIELD–ANTIFIELD FORMALISM

In this section we are going to use the field–antifield formalism of Batalin–Vilkowisky to deal with one- and two-form gauge fields in the superfield language. To do this, we have to introduce

collective fields,[7] from where the antifields[8,9] are identified. Since the set of fields in the Abelian and non-Abelian cases are different, due to their different reducibility characteristics, we consider both cases separately.

Let us start from the Abelian case. The coupling of the tensor gauge field $B_{\mu\nu}$ and the ghosts with the vector gauge one is a trivial matter. To fix the superfields given at Eq. (IV.7), we introduce the collective fields

$$\mathscr{B}^1_{\mu\nu},\ \mathscr{G}^1_\mu,\ \overline{\mathscr{G}}^1_\mu,\ \mathscr{G}^1,\ \overline{\mathscr{G}}^1,\ \mathscr{K}^1. \tag{VI.1}$$

The index 1 stands for the BRST symmetry which we have privileged in the BV formalism. If we were interested in maintaining both symmetries at the same level we would introduce also collective fields with indices 2 and arrive at a Batalin–Lavranov–Tyutin (BLT) action.[29] The next step is to define the shifted two-form $\tilde{\Lambda}$ as in Eq. (IV.1) but here written in terms of the shifted superfields

$$\begin{aligned}
\tilde{\mathscr{B}}_{\mu\nu} &= \mathscr{B}_{\mu\nu} - \mathscr{B}^1_{\mu\nu},\\
\tilde{\mathscr{G}}_\mu &= \mathscr{G}_\mu - \mathscr{G}^1_\mu,\\
\tilde{\bar{\mathscr{G}}}_\mu &= \overline{\mathscr{G}}_\mu - \overline{\mathscr{G}}^1_\mu,\\
\tilde{\mathscr{G}} &= \mathscr{G} - \mathscr{G}^1,\\
\tilde{\bar{\mathscr{G}}} &= \overline{\mathscr{G}} - \overline{\mathscr{G}}^1,\\
\tilde{\mathscr{K}} &= \mathscr{K} - \mathscr{K}^1.
\end{aligned} \tag{VI.2}$$

We impose again the horizontal condition $\tilde{H} = d\tilde{\Lambda}$, that is to say, its components along the θ and $\overline{\theta}$ directions must vanish. We naturally obtain the same conditions (IV.5) with the shifted superfields in place of the original superfields. Now we have an ambiguity in determining the components of the individual superfields. As an example, for the superfield $\tilde{\mathscr{G}}_\mu$ we obtain from Eq. (IV.7)

$$\tilde{\mathscr{G}}_\mu = \tilde{g}_\mu + \theta\partial_\mu\tilde{G} + \overline{\theta} e_\mu + i\overline{\theta}\partial_\mu f, \tag{VI.3}$$

and a possible choice for $\mathscr{G}_\mu$ and $\mathscr{G}^1_\mu$ is

$$\begin{aligned}
\mathscr{G}_\mu &= g_\mu + \theta\Pi_\mu + \overline{\theta}(\overline{g}^*_\mu + e_\mu) + i\overline{\theta}\theta(N_\mu + \partial_\mu f\ ),\\
\mathscr{G}^1_\mu &= g^1_\mu + \theta(\Pi_\mu - \partial_\mu\tilde{G}) + \overline{\theta}\overline{g}^*_\mu + i\overline{\theta}\theta N_\mu,
\end{aligned} \tag{VI.4}$$

where Π_μ, $\overline{g}^*_\mu$, and N_μ are new fields which arise from the additional shift symmetry. Observe that it is not necessary to introduce fields like f^1 or e^1_μ since f and e_μ are already auxiliary fields.

The remaining set of fields is chosen to be

$$\begin{aligned}
\mathscr{B}_{\mu\nu} &= B_{\mu\nu} + \theta\Pi_{\mu\nu} + \overline{\theta}(B^*_{\mu\nu} + \partial_{[\mu}\tilde{\bar{g}}_{\nu]}) + \overline{\theta}\theta(N_{\mu\nu} - \partial_{[\mu}e_{\nu]}),\\
\mathscr{B}^1_{\mu\nu} &= B^1_{\mu\nu} + \theta(\Pi_{\mu\nu} - \partial_{[\mu}\tilde{g}_{\nu]}) + \overline{\theta}B^*_{\mu\nu} + \overline{\theta}\theta N_{\mu\nu},\\
\overline{\mathscr{G}}_\mu &= \overline{g}_\mu + \theta\overline{\Pi}_\mu + \overline{\theta}(g^*_\mu + \partial_\mu\tilde{\bar{G}}) + i\overline{\theta}\theta(\overline{N}_\mu - \partial_\mu\overline{f}\ ),\\
\overline{\mathscr{G}}^1_\mu &= \overline{g}^1_\mu + \theta(\overline{\Pi}_\mu - \partial_\mu\tilde{K} + e_\mu) + \overline{\theta}g^*_\mu + i\overline{\theta}\theta\overline{N}_\mu,\\
\mathscr{G} &= G + \theta\Pi + \overline{\theta}(\overline{G}^* - f\ ) + \overline{\theta}\theta N,\\
\mathscr{G}^1 &= G^1 + \theta\Pi + \overline{\theta\overline{G}}^* + \overline{\theta}\theta N,
\end{aligned} \tag{VI.5}$$

$$\bar{\mathscr{G}}=\bar{G}+\theta\bar{\Pi}+\bar{\theta}G^*+\bar{\theta}\theta\bar{N},$$

$$\bar{\mathscr{G}}^1=\bar{G}^1+\theta(\bar{\Pi}+\bar{f})+\bar{\theta}G^*+\bar{\theta}\theta\bar{N},$$

$$\mathscr{K}=K+i\theta\pi+i\bar{\theta}(K^*+\bar{f})+\bar{\theta}\theta M,$$

$$\mathscr{K}^1=K^1+i\theta(\pi-f)+i\bar{\theta}K^*+\bar{\theta}\theta M.$$

The natural interpretation for the new fields appearing in (VI.4) and (VI.5) is the following. The quantum action describing the two-form gauge fields has an additional trivial shift symmetry with respect to the tilde fields defined in Eq. (VI.2). Consequently we have to introduce the ghosts: $\Pi_{\mu\nu}$, Π_μ, $\bar{\Pi}_\mu$, Π, $\bar{\Pi}$, and π; the antighosts: $B^*_{\mu\nu}$, g^*_μ, $\bar{g}^*_\mu$, G^*, $\bar{G}^*$, and K^*, and auxiliary Nakanish–Lautrup fields $N_{\mu\nu}$, N_μ, $\bar{N}_\mu$, N, $\bar{N}$, and M.

We shall see that with the choice given by Eqs. (VI.4) and (VI.5), these antighosts will play the role of the antifields in the BV formalism. In order to fix this shift symmetry we impose that all indices 1 fields vanish. This can be done with the aid of the following BRST and anti-BRST invariant terms

$$\begin{aligned}
&\tfrac{1}{2}\,\partial_\theta\partial_{\bar{\theta}}(\mathscr{B}^1_{\mu\nu}\mathscr{B}^{1\mu\nu})=B^1_{\mu\nu}N^{\mu\nu}+(\Pi_{\mu\nu}-\partial_{[\mu}\tilde{g}_{\nu]})B^{*\mu\nu},\\
&\partial_\theta\partial_{\bar{\theta}}(\mathscr{G}^1_\mu\bar{\mathscr{G}}^{1\mu})=iN_\mu\bar{g}^{1\mu}+ig^1_\mu\bar{N}^\mu-(\Pi_\mu-\partial_\mu\tilde{G})g^{*\mu}+\bar{g}^*_\mu(\bar{\Pi}^\mu-\partial^\mu\tilde{K}+e^\mu),\\
&\partial_\theta\partial_{\bar{\theta}}(\mathscr{G}^1\bar{\mathscr{G}}^1)=G^1\bar{N}+N\bar{G}^1+\Pi G^*-\bar{G}^*(\bar{\Pi}+\bar{f}),\\
&\tfrac{1}{2}\,\partial_\theta\partial_{\bar{\theta}}(\mathscr{K}^1\mathscr{K}^1)=K^1M-(\pi-f)K^*.
\end{aligned}\tag{VI.6}$$

To fix the original gauge symmetry of the theory, we choose a gauge-fixing superboson

$$\Phi=\phi+\theta\delta\phi+\bar{\theta}\psi+\bar{\theta}\theta\delta\psi,\tag{VI.7}$$

with $\psi=\bar{\delta}\phi$. Then we have

$$\begin{aligned}
\partial_\theta\partial_{\bar{\theta}}\Phi&=\delta\psi\\
&=\delta B_{\mu\nu}\frac{\delta\psi}{\delta B_{\mu\nu}}+\delta g_\mu\frac{\delta\psi}{\delta g_\mu}+\delta\bar{g}_\mu\frac{\delta\psi}{\delta\bar{g}_\mu}\\
&\quad+\delta G\frac{\delta\psi}{\delta G}+\delta\bar{G}\frac{\delta\psi}{\delta\bar{G}}+\delta K\frac{\delta\psi}{\delta K},
\end{aligned}\tag{VI.8}$$

where we have considered $\psi=\psi(B_{\mu\nu},g_\mu,\bar{g}_\mu,G,\bar{G},K)$. The δ variations of the fields are obtained from the θ components of (VI.4) and (VI.5) and we have, for example,

$$\begin{aligned}
\delta B_{\mu\nu}&=\Pi_{\mu\nu},\\
\delta K&=i\pi.
\end{aligned}\tag{VI.9}$$

The use of the equations of motion for the fields $N_{\mu\nu}$, N_μ, $\bar{N}_\mu$, N, $\bar{N}$, and M, as well as for the ones with indices 1, makes these last ones vanish. This corresponds to a functional integration over the same fields. Once this is done, the sum of Eqs. (VI.6) and (VI.8), and the singlet $\mathscr{S}_I$ in Eq. (IV.12) give us the two-form field-dependent part of the quantum action

$$S_B \equiv \int dx^\mu \Bigg[\frac{1}{12} H_{\mu\nu\rho} H^{\mu\nu\rho} + B^*_{\mu\nu} \partial^{[\mu} g^{\nu]} + g^{*\mu} \partial_\mu G + \overline{g}^*_\mu (e^\mu - \partial^\mu K) - G^* f - \overline{G}^* \overline{f} + K^* \pi$$
$$+ \Pi_{\mu\nu} \left(B^{*\mu\nu} + \frac{\delta\psi}{\delta B_{\mu\nu}} \right) + \Pi_\mu \left(- g^{*\mu} + \frac{\delta\psi}{\delta g_\mu} \right) + \overline{\Pi}_\mu \left(\overline{g}^{*\mu} + \frac{\delta\psi}{\delta \overline{g}_\mu} \right) + \Pi \left(G^* + \frac{\delta\psi}{\delta G} \right)$$
$$+ \overline{\Pi} \left(\overline{G}^* + \frac{\delta\psi}{\delta \overline{G}} \right) + \pi \left(- K^* + i \frac{\delta\psi}{\delta K} \right) \Bigg]. \tag{VI.10}$$

If we add the action

$$S_A = \int dx^\mu \left[- \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i A^*_\mu \partial^\mu c + \overline{c}^* b \right], \tag{VI.11}$$

with a convenient gauge choice for A^*_μ, and $\overline{c}^*$, and the interacting singlet term $\mathscr{L}_{IV}$ of Eq. (IV.17), to S_B, we obtain the full BV action which agrees with the one obtained in Ref. 15.

We now turn our attention to the non-Abelian theory described in Secs. III and V. The procedure is quite similar but now the schedule of superfields is different since we do not have ghosts-for-ghosts and we have to introduce the Stuckelberg superfield. Another point is that the coupling of the one-form gauge field with its corresponding ghosts is not so trivial as in the Abelian case. Let us start from the one-form Yang–Mills fields by introducing new index 1 superfields through the following shift

$$\tilde{\mathscr{A}}_\mu \equiv \mathscr{A}_\mu - \mathscr{A}^1_\mu,$$
$$\tilde{\mathscr{C}} \equiv \mathscr{C} - \mathscr{C}^1, \tag{VI.12}$$
$$\tilde{\overline{\mathscr{C}}} \equiv \overline{\mathscr{C}} - \overline{\mathscr{C}}^1.$$

Imposing the horizontal condition on the shifted curvature, we arrive at expressions similar to (II.9)–(II.11) for the shifted superfields (VI.12), but written in terms of the shifted components

$$\tilde{A}_\mu \equiv A_\mu - A^1_\mu,$$
$$\tilde{c} \equiv c - c^1, \tag{VI.13}$$
$$\tilde{\overline{c}} \equiv \overline{c} - \overline{c}^1.$$

A field like b^1 does not need to be introduced here since b is already an auxiliary field. The solution for the individual superfields contains some arbitrariness and we choose

$$\mathscr{A}_\mu = A_\mu + i\theta \Pi^A_\mu + i\overline{\theta}(A^*_\mu + D^{A-A^1}_\mu (\overline{c} - \overline{c}^1)) + i\overline{\theta}\theta [N^A_\mu + D^{A-A^1}_\mu b + i D^{A-A^1}_\mu (c - c^1) \times (\overline{c} - \overline{c}^1)],$$
$$\mathscr{A}^1_\mu = A^1_\mu + i\theta [\Pi^A_\mu - D^{A-A_1}_\mu (c - c^1)] + i\overline{\theta} A^*_\mu + i\overline{\theta}\theta N^A_\mu,$$
$$\mathscr{C} = C + i\theta \Pi^c + i\overline{\theta}[\overline{c}^* - b + (c - c^1) \times (\overline{c} - \overline{c}^1)] + i\overline{\theta}\theta [N^c + (c - c^1) \times (b + i(c - c^1) \times (\overline{c} - \overline{c}^1))], \tag{VI.14}$$
$$\mathscr{C}^1 = c^1 + i\theta [\Pi^c + \tfrac{1}{2} g (c - c^1) \times (c - c^1)] + i\overline{\theta} c^* + i\overline{\theta}\theta N^c,$$
$$\overline{\mathscr{C}} = \overline{c} + i\theta \Pi^{\overline{c}} + i\overline{\theta}[c^* - \tfrac{1}{2} (\overline{c} - \overline{c}^1) \times (\overline{c} - \overline{c}^1)] + i\overline{\theta}\theta [N^{\overline{c}} - b \times (\overline{c} - \overline{c}^1)],$$
$$\overline{\mathscr{C}}^1 = \overline{c}^1 + i\theta (\Pi^c + ib) + i\overline{\theta}\overline{c}^* + i\overline{\theta}\theta N^{\overline{c}}.$$

Concerning the two-form gauge fields sector, we introduce four index 1 superfields corresponding to $\mathscr{B}_{\mu\nu}$, $\mathscr{S}_\mu$, $\mathscr{G}_\mu$, $\bar{\mathscr{G}}_\mu$ and impose the horizontal condition on $\tilde{H}=\tilde{\mathscr{D}}\Gamma$ defined through (V.2). One of the possible solutions that is in agreement with Eqs. (V.4)–(V.6) and (V.13), is given by (this one leads to the familiar BV action)

$$
\begin{aligned}
\mathscr{B}_{\mu\nu}&=B_{\mu\nu}+\theta\Pi_{\mu\nu}+\bar{\theta}(B^*_{\mu\nu}+\tilde{B}^{\bar{\theta}}_{\mu\nu})+\bar{\theta}\theta(N_{\mu\nu}+\tilde{B}^{\bar{\theta}\theta}_{\mu\nu}),\\
\mathscr{B}^1_{\mu\nu}&=B^1_{\mu\nu}+\theta(\Pi_{\mu\nu}-\tilde{B}^{\theta}_{\mu\nu})+\bar{\theta}B^*_{\mu\nu}+\bar{\theta}\theta N_{\mu\nu},\\
\mathscr{S}_\mu&=\Omega_\mu+\theta s_\mu+\bar{\theta}(\Omega^*_\mu+\tilde{\Omega}^{\bar{\theta}}_\mu)+\bar{\theta}\theta(M_\mu+\tilde{\Omega}^{\bar{\theta}\theta}_\mu),\\
\mathscr{S}^1_\mu&=\Omega^1_\mu+\theta(s_\mu-\tilde{\Omega}^{\theta}_\mu)+\bar{\theta}\Omega^*_\mu+\bar{\theta}\theta M_\mu,\\
\mathscr{G}_\mu&=g_\mu+\theta\Pi_\mu+\bar{\theta}(\bar{g}^*_\mu+\tilde{g}^{\bar{\theta}}_\mu)+\bar{\theta}\theta(N_\mu+\tilde{g}^{\bar{\theta}\theta}_\mu),\\
\mathscr{G}^1_\mu&=g^1_\mu+\theta(\Pi_\mu-\tilde{g}^{\theta}_\mu)+\bar{\theta}\bar{g}^*_\mu+\bar{\theta}\theta N_\mu,\\
\bar{\mathscr{G}}_\mu&=\bar{g}_\mu+\theta\bar{\Pi}_\mu+\bar{\theta}(g^*_\mu+\tilde{g}^{\bar{\theta}}_\mu)+\bar{\theta}\theta(\bar{N}_\mu+\tilde{\bar{g}}^{\bar{\theta}}_\mu),\\
\bar{\mathscr{G}}^1_\mu&=\bar{g}^1_\mu+\theta(\bar{\Pi}_\mu-\tilde{\bar{g}}^{\theta}_\mu)+\bar{\theta}g^*_\mu+\bar{\theta}\theta\bar{N}_\mu,
\end{aligned}
\tag{VI.15}
$$

where

$$
\begin{aligned}
\tilde{B}^{\theta}_{\mu\nu}&=\tilde{D}_{[\mu}\tilde{g}_{\nu]}-i\tilde{c}\times\tilde{B}_{\mu\nu},\\
\tilde{B}^{\bar{\theta}}_{\mu\nu}&=\tilde{D}_{[\mu}\tilde{\bar{g}}_{\nu]}-i\tilde{\bar{c}}\times\tilde{B}_{\mu\nu},\\
\tilde{B}^{\bar{\theta}\theta}_{\mu\nu}&=-\tilde{D}_{[\mu}e_{\nu]}-i\tilde{c}\times\tilde{D}_{[\mu}\tilde{\bar{g}}_{\nu]}-i(\tilde{D}_{[\mu}\tilde{\bar{c}})\times\tilde{g}_{\nu]}-ib\times\tilde{B}_{\mu\nu}-\tilde{\bar{c}}\times(\tilde{c}\times\tilde{B}_{\mu\nu}),\\
\tilde{\Omega}^{\theta}_\mu&=-\tilde{g}_\mu-i\tilde{c}\times\tilde{\Omega}_\mu,\\
\tilde{\Omega}^{\bar{\theta}}_\mu&=-\tilde{\bar{g}}_\mu-i\tilde{\bar{c}}\times\tilde{\Omega}_\mu,\\
\tilde{\Omega}^{\bar{\theta}\theta}_\mu&=e_\mu+\tilde{c}\times(i\tilde{\bar{g}}_\mu-\tilde{\bar{c}}\times\tilde{\Omega}_\mu)-i(b+\tilde{c}\times\tilde{\bar{c}})\times\tilde{\Omega}_\mu,\\
\tilde{g}^{\theta}_\mu&=-i\tilde{c}\times\tilde{g}_\mu,\\
\tilde{g}^{\bar{\theta}}_\mu&=e_\mu,\\
\tilde{g}^{\bar{\theta}\theta}_\mu&=(\tilde{c}\times\tilde{\bar{c}})\times\tilde{g}_\mu-ic\times e_\mu-ib\times\tilde{g}_\mu,\\
\tilde{\bar{g}}^{\theta}_\mu&=-e_\mu-i\tilde{c}\times\tilde{\bar{g}}_\mu-i\tilde{\bar{c}}\times\tilde{g}_\mu,\\
\tilde{\bar{g}}^{\bar{\theta}}_\mu&=-i\tilde{\bar{c}}\times\tilde{\bar{g}}_\mu,\\
\tilde{\bar{g}}^{\bar{\theta}\theta}_\mu&=\tilde{\bar{c}}\times(\tilde{\bar{c}}\times\tilde{g}_\mu)+\tilde{\bar{c}}\times(\tilde{c}\times\tilde{\bar{g}}_\mu)-ib\times\tilde{\bar{g}}_\mu-i\tilde{\bar{c}}\times e_\mu.
\end{aligned}
\tag{VI.16}
$$

In Eqs. (VI.14) and (VI.15) some new auxiliary fields have appeared which may be interpreted as ghosts ($\Pi^A_\mu,\Pi^c,\Pi^c,\Pi_{\mu\nu},s_\mu,\Pi_\mu,\bar{\Pi}_\mu$), antighosts ($\underline{A}^*_\mu,c^*,\bar{c}^*,B^*_{\mu\nu},\Omega^*_\mu,\bar{g}^*_\mu,g^*_\mu$), and Nakanish–Lautrup fields (N^A_μ, N^c, N^c, $N_{\mu\nu}$, M_μ, N_μ, and $\bar{N}_\mu$).

As in the Abelian case, the antighosts above will play the role of the BV antifields. At this point we can fix the shift symmetry using the following (anti-) BRST invariant terms:

$$\tfrac{1}{2}\ \mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}(\mathscr{A}^1_\mu\mathscr{A}^{1\mu})=\mathrm{tr}\{iA^1_\mu B^\mu+(D^{A-A^1}_\mu(c-c^1)-\Pi^A_\mu)A^{*\mu}\},$$

$$\mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}(\mathscr{C}^1\overline{\mathscr{C}}^1)=\mathrm{tr}\{iN^c\overline{c}^1+[\Pi^c+\tfrac{1}{2}\ (c-c^1)\times(c-c^1)]c^*-\overline{c}^*(\Pi^{\overline{c}}+ib)+ic^1N^{\overline{c}}\},$$

$$\tfrac{1}{2}\ \mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}(\mathscr{B}^1_{\mu\nu}\mathscr{B}^{1\mu\nu})=\mathrm{tr}\{B^1_{\mu\nu}N^{\mu\nu}+(\Pi_{\mu\nu}-\tilde{B}^\theta_{\mu\nu})B^{*\mu\nu}\},\tag{VI.17}$$

$$\tfrac{1}{2}\ \mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}(\mathscr{S}^1_\mu\mathscr{S}^{1\mu})=\mathrm{tr}\{\Omega^1_\mu M^\mu+(s_\mu-\tilde{\Omega}^\theta_\mu)\Omega^{*\mu}\},$$

$$\mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}(\mathscr{G}^1_\mu\overline{\mathscr{G}}^{1\mu})=\mathrm{tr}\{N_\mu\overline{g}^{1\mu}+g^1_\mu\overline{N}^\mu-(\Pi_\mu-\tilde{g}^\theta_\mu)g^{*\mu}+\overline{g}^*_\mu(\overline{\Pi}^\mu-\tilde{\bar{g}}^{\theta\mu})\}.$$

In order to fix the original gauge symmetry we choose a superboson

$$\Phi=\phi+\theta\delta\phi+\overline{\theta\delta}\phi+\overline{\theta}\theta\delta\overline{\delta}\phi,\tag{VI.18}$$

which generates the invariant term

$$\begin{aligned}\mathrm{tr}\ \partial_\theta\partial_{\overline{\theta}}\Phi&=\mathrm{tr}\ \delta\overline{\delta}\phi\\&=\mathrm{tr}\Bigg[\delta A_\mu\frac{\delta(\overline{\delta}\phi)}{\delta A_\mu}+\delta c\ \frac{\delta(\overline{\delta}\phi)}{\delta c}+\delta\overline{c}\frac{\delta(\overline{\delta}\phi)}{\delta\overline{c}}\\&\quad+\delta B_{\mu\nu}\frac{\delta(\overline{\delta}\phi)}{\delta B_{\mu\nu}}+\delta\Omega_\mu\frac{\delta(\overline{\delta}\phi)}{\delta\Omega_\mu}+\delta g_\mu\frac{\delta(\overline{\phi})}{\delta g_\mu}+\delta\overline{g}_\mu\frac{\delta(\overline{\delta}\phi)}{\delta\overline{g}_\mu}\Bigg],\end{aligned}\tag{VI.19}$$

provided that $\overline{\delta}\phi=\overline{\delta}\phi(A_\mu,B_{\mu\nu},\Omega_\mu,c,\overline{c},g_\mu,\overline{g}_\mu)$. The BV gauge-fixing fermion is naturally identified with $\overline{\delta}\phi$. With this choice, the gauge-fixed Lagrangian density reads

$$\begin{aligned}\mathscr{L}_{\mathrm{gf}}=\mathrm{tr}\Bigg[&iA^1_\mu B^\mu+iN^c\overline{c}^1+ic^1N^{\overline{c}}-A^{*\mu}D^{A-A_1}(c-c_1)+\overline{c}^*[(c-c^1)\times(c-c^1)]-ic^*b\\&+\Pi^A_\mu\Bigg(-A^{*\mu}+i\ \frac{\delta(\overline{\delta}\phi)}{\delta A_\mu}\Bigg)+\Pi^c\Bigg(c^*+i\ \frac{\delta(\overline{\delta}\phi)}{\delta c}\Bigg)+\Pi^{\overline{c}}\Bigg(-\overline{c}^*+i\ \frac{\delta(\overline{\delta}\phi)}{\delta\overline{c}}\Bigg)+\Omega^1_\mu M^\mu+B^1_{\mu\nu}N^{\mu\nu}\\&+N_\mu\overline{g}^{1\mu}+g^1_\mu\overline{N}^\mu+B^{*\mu\nu}\tilde{B}^\theta_{\mu\nu}+\Omega^{*\mu}\tilde{\Omega}^\theta_\mu+g^{*\mu}\tilde{g}^\theta_\mu-\overline{g}^*_\mu\tilde{\bar{g}}^{\theta\mu}+s_\mu\Bigg(\Omega^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta\Omega_\mu}\Bigg)\\&+\Pi_{\mu\nu}\Bigg(B^{*\mu\nu}+\frac{\delta(\overline{\delta}\phi)}{\delta B_{\mu\nu}}\Bigg)+\Pi_\mu\Bigg(-g^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta g_\mu}\Bigg)+\overline{\Pi}_\mu\Bigg(\overline{g}^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta\overline{g}_\mu}\Bigg)\Bigg].\end{aligned}\tag{VI.20}$$

The replacement of the equations of motion for $N_{\mu\nu}$, M_μ, N_μ, $\overline{N}_\mu$, and the index 1 fields leads to the action

$$\begin{aligned}S=S_0+\mathrm{tr}\int dx^\mu\Bigg[&B^*_{\mu\nu}(D^{[\mu}g^{\nu]}-ic\times B_{\mu\nu})+\Omega^*_\mu(-g^\mu-ic\times B^{\mu\nu})-g^*_\mu(ic\times g^\mu)\\&+\overline{g}^*_\mu(e^\mu+ic\times\overline{g}^\mu+i\overline{c}\times g^\mu)+s_\mu\Bigg(\Omega^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta\Omega_\mu}\Bigg)+\Pi_{\mu\nu}\Bigg(B^{*\mu\nu}+\frac{\delta(\overline{\delta}\phi)}{\delta B_{\mu\nu}}\Bigg)\\&+\Pi_\mu\Bigg(-g^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta g_\mu}\Bigg)+\overline{\Pi}_\mu\Bigg(\overline{g}^{*\mu}+\frac{\delta(\overline{\delta}\phi)}{\delta\overline{g}_\mu}\Bigg)\Bigg],\end{aligned}\tag{VI.21}$$

where the classical part S_0 is given by Eqs. (II.19), (V.11), and (V.14), namely

$$S_0=\int dx\Bigg[-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}-m\epsilon^{\mu\nu\rho\kappa}\ \mathrm{tr}\ F_{\mu\nu}B_{\rho\kappa}+\frac{3}{4}\ \mathrm{tr}(D_{[\mu}B_{\nu\rho]}+D_{[\mu}D_\nu\Omega_{\rho]})^2\Bigg].\tag{VI.22}$$

Action S is exactly the Batalin–Vilkovisky one with gauge-fixing fermion $\overline{\delta}\phi$.

VII. CONCLUSION

We have used the BRST superspace formalism for two-form gauge fields. We have considered the Abelian and the non-Abelian cases. The first one is of course much simpler algebraically, but has a richer structure by virtue of the reducibility condition. We have shown that the treatment in the BRST superspace language naturally contains all the ingredients of the theory and has permitted us to obtain a deeper comprehension of the theory where both BRST and anti-BRST symmetries are present. In the non-Abelian case, we have considered an off-shell treatment, compatible with the Abelian limit where both vector and tensor fields are coupled in a topological manner. This was achieved by the introduction of a kind of a Stuckelberg field which we believe is the only way we can make the non-Abelian formalism compatible with this Abelian limit. We have also explicitly shown that the superfield treatment of one- and two-form gauge fields leads directly to the field–antifield BV action.

ACKNOWLEDGMENT

This work is supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq, Financiadora de Estudos e Projetos—FINEP and Fundação Universitária José Bonifácio—FUJB (Brazilian Research Agency).

APPENDIX A

The superfield equations given by (IV.5) and the superfield expressions given by (IV.3) lead, in components, to the following set of equations [each set corresponds to one of the equations (IV.5)]

$$
\begin{aligned}
&B^{\theta}_{\mu\nu}+\frac{i}{2}\,(\partial_\mu g_\nu-\partial_\nu g_\mu)=0,\\
&B^{\bar\theta\theta}_{\mu\nu}-\frac{i}{2}\,(\partial_\mu e_\nu-\partial_\nu e_\mu)=0,\\
&\partial_\mu \bar e_\nu-\partial_\nu \bar e_\mu=0,\\
&\partial_\mu G^{\bar\theta\theta}_{\nu}-\partial_\nu G^{\bar\theta\theta}_{\mu}=0,
\end{aligned}
\tag{A1}
$$

$$
\begin{aligned}
&B^{\bar\theta}_{\mu\nu}+\frac{i}{2}\,(\partial_\mu \bar g_\nu-\partial_\nu \bar g_\mu)=0,\\
&B^{\bar\theta\theta}_{\mu\nu}+\frac{i}{2}\,(\partial_\mu \bar G^{\theta}_{\nu}-\partial_\nu \bar G^{\theta}_{\mu})=0,\\
&\partial_\mu \bar G^{\bar\theta}_{\nu}-\partial_\nu \bar G^{\bar\theta\theta}_{\mu}=0,\\
&\partial_\mu \bar G^{\bar\theta}_{\nu}-\partial_\nu \bar G^{\bar\theta\theta}_{\mu}=0,
\end{aligned}
\tag{A2}
$$

$$
\begin{aligned}
&\bar e_\mu-\partial_\mu G=0,\\
&G^{\bar\theta\theta}_{\mu}+\partial_\mu G^{\bar\theta}=0,\\
&\partial_\mu G^{\bar\theta}=0,\\
&\partial_\mu G^{\bar\theta\theta}=0,
\end{aligned}
\tag{A3}
$$

$$\begin{gathered}\overline{\overline{G}^{\theta}_{\mu}}-\partial_{\mu}\overline{G}=0,\\ \overline{\overline{G}^{\theta}_{\mu}}-\partial_{\mu}\overline{G}^{\theta}=0,\\ \partial_{\mu}\overline{\overline{G}^{\theta}}=0,\\ \partial_{\mu}\overline{\overline{G}^{\theta\theta}}=0,\end{gathered}\tag{A4}$$

$$\begin{gathered}\overline{G}^{\theta}_{\mu}+e_{\mu}-\partial_{\mu}K=0,\\ G^{\overline{\theta\theta}}_{\mu}-\partial_{\mu}f=0,\\ \overline{G}^{\overline{\theta\theta}}_{\mu}+\partial_{\mu}\overline{f}=0,\\ \partial_{\mu}K^{\overline{\theta\theta}}=0,\end{gathered}\tag{A5}$$

$$\begin{gathered}G^{\theta}=0,\\ G^{\overline{\theta\theta}}=0,\end{gathered}\tag{A6}$$

$$\begin{gathered}\overline{\overline{G}^{\theta}}=0,\\ \overline{\overline{G}^{\theta\theta}}=0,\end{gathered}\tag{A7}$$

$$\begin{gathered}\overline{G}^{\theta}+\overline{f}=0,\\ \overline{\overline{G}^{\theta\theta}}=0,\\ K^{\overline{\theta\theta}}=0,\end{gathered}\tag{A8}$$

$$\begin{gathered}G^{\overline{\theta}}+f=0,\\ G^{\overline{\theta\theta}}=0,\\ K^{\overline{\theta\theta}}=0.\end{gathered}\tag{A9}$$

As one observes, some of the equations above correspond to the relations already displayed in (IV.6). One can also verify that the remaining ones do not lead to any new result besides what is shown in (IV.6). For example, the third of Eqs. (A1) is in agreement with the fourth Eq. (IV.6). The same occurs with respect the fourth Eq. (A1) and the sixth Eq. (IV.6), and so on.

APPENDIX B

Each one of the equations (V.3) leads, respectively, to the following sets below

$$\begin{gathered}B^{\theta}_{\mu\nu}+iD_{[\mu}g_{\nu]}+c\times B_{\mu\nu}=0,\\ D_{[\mu}\overline{e}_{\nu]}+iA^{\theta}_{[\mu}\times g_{\nu]}-c\times B^{\theta}_{\mu\nu}-iC^{\theta}\times B_{\mu\nu}=0,\\ \overline{B}^{\theta\theta}_{\mu\nu}-iD_{[\mu}e_{\nu]}+\overline{A}^{\theta}_{[\mu}\times g_{\nu]}-c\times\overline{B}^{\theta}_{\mu\nu}-\overline{c}^{\theta}\times B_{\mu\nu}=0,\end{gathered}\tag{B1}$$

$$D_{[\mu}\bar{G}^{\theta\theta}_{\nu]}+A^{\theta}_{[\mu}\times e_{\nu]}-\bar{A}^{\theta}_{[\mu}\times\bar{e}_{\nu]}+\bar{A}^{\theta\theta}_{[\mu}\times g_{\nu]}-ic\times\bar{B}^{\theta\theta}_{\mu\nu}+ic^{\theta}\times\bar{B}^{\theta}_{\mu\nu}-i\bar{c}^{\theta}\times B^{\theta}_{\mu\nu}-i\bar{c}^{\theta\theta}\times B_{\mu\nu}=0,$$

$$\bar{B}^{\theta}_{\mu\nu}+iD_{[\mu}\bar{g}_{\nu]}+\bar{c}\times B_{\mu\nu}=0,$$

$$D_{[\mu}\bar{G}^{\theta}_{\nu]}+i\bar{A}^{\theta}_{[\mu}\times\bar{g}_{\nu]}-\bar{c}\times\bar{B}^{\theta}_{\mu\nu}-i\bar{b}\times B_{\mu\nu}=0,$$

$$\bar{B}^{\theta\theta}_{\mu\nu}+iD_{[\mu}\bar{G}^{\theta}_{\nu]}-A^{\theta}_{[\mu}\times\bar{g}_{\nu]}-i\bar{c}\times B^{\theta}_{\mu\nu}+b\times B_{\mu\nu}=0,\tag{B2}$$

$$D_{[\mu}\bar{G}^{\theta\theta}_{\nu]}+A^{\theta}_{[\mu}\times\bar{G}^{\theta}_{\nu]}-\bar{A}^{\theta}_{[\mu}\times\bar{G}^{\theta}_{\nu]}+\bar{A}^{\theta\theta}_{[\mu}\times\bar{g}_{\nu]}-i\bar{c}\times\bar{B}^{\theta\theta}_{\mu\nu}+ib\times\bar{B}^{\theta}_{\mu\nu}-i\bar{b}\times B^{\theta}_{\mu\nu}-i\bar{C}^{\theta\theta}\times B_{\mu\nu}=0,$$

$$D_{\mu}G-\bar{e}_{\mu}-ic\times g_{\mu}=0,$$

$$D_{\mu}G^{\theta}+A^{\theta}_{\mu}\times G+c\times\bar{e}_{\mu}-C^{\theta}\times g_{\mu}=0,$$

$$G^{\theta\theta}_{\mu}+D_{\mu}\bar{G}^{\theta}+\bar{A}^{\theta}_{\mu}\times G+c\times e_{\mu}-\bar{C}^{\theta}\times g_{\mu}=0,\tag{B3}$$

$$D_{\mu}\bar{G}^{\theta\theta}+iA^{\theta}_{\mu}\times\bar{G}^{\theta}-iA^{\theta}_{\mu}\times G^{\theta}-\bar{A}^{\theta\theta}_{\mu}\times G+ic\times\bar{G}^{\theta\theta}_{\mu}-C^{\theta}\times e_{\mu}+\bar{C}^{\theta}\times\bar{e}_{\mu}+iC^{\theta\theta}\times g_{\mu}=0,$$

$$D_{\mu}\bar{G}-\bar{G}^{\theta}_{\mu}-i\bar{c}\times\bar{g}_{\mu}=0,$$

$$D_{\mu}\bar{G}^{\theta}+A^{\theta}_{\mu}\times\bar{G}+\bar{c}\times\bar{g}^{\theta}_{\mu}-\bar{b}\times\bar{G}_{\mu}=0,$$

$$\bar{G}^{\theta\theta}_{\mu}-D_{\mu}\bar{G}^{\theta}-A^{\theta}_{\mu}\times\bar{G}-\bar{c}\times\bar{G}^{\theta}_{\mu}-b\times\bar{g}_{\mu}=0,\tag{B4}$$

$$D_{\mu}\bar{G}^{\theta\theta}-iA^{\theta}_{\mu}\times\bar{G}^{\theta}+iA^{\theta}_{\mu}\times\bar{G}^{\theta}+\bar{A}^{\theta\theta}_{\mu}\times\bar{G}-i\bar{c}\times\bar{G}^{\theta\theta}_{\mu}-\bar{b}\times\bar{G}^{\theta}_{\mu}+b\times\bar{G}^{\theta}_{\mu}-i\bar{C}^{\theta\theta}\times\bar{g}_{\mu}=0,$$

$$D_{\mu}K-\bar{G}^{\theta}_{\mu}-e_{\mu}-ic\times\bar{g}_{\mu}-i\bar{c}\times g_{\mu}=0,$$

$$D_{\mu}f-G^{\theta\theta}_{\mu}+c\times\bar{G}^{\theta}_{\mu}-C^{\theta}\times\bar{g}_{\mu}+\bar{c}\times\bar{e}_{\mu}-b\times g_{\mu}+A^{\theta}_{\mu}\times K=0,$$

$$D_{\mu}\bar{f}+\bar{G}^{\theta\theta}_{\mu}+c\times\bar{G}^{\theta}_{\mu}-\bar{C}^{\theta}\times\bar{g}_{\mu}+\bar{c}\times e_{\mu}-\bar{b}\times g_{\mu}+\bar{A}^{\theta}_{\mu}\times K=0,\tag{B5}$$

$$D_{\mu}K^{\theta\theta}-ic\times\bar{G}^{\theta\theta}_{\mu}+C^{\theta}\times\bar{G}^{\theta}_{\mu}-\bar{C}^{\theta}\times\bar{G}^{\theta}_{\mu}-iC^{\theta\theta}\times\bar{g}_{\mu}-i\bar{c}\times G^{\theta\theta}_{\mu}+b\times e_{\mu}-\bar{b}\times\bar{e}_{\mu}-i\bar{C}^{\theta\theta}\times g_{\mu}+iA^{\theta}_{\mu}\times\bar{f}-i\bar{A}^{\theta}_{\mu}\times f+\bar{A}^{\theta\theta}_{\mu}\times K=0,$$

$$\bar{G}^{\theta}+f+\bar{c}\times G+c\times K=0,$$

$$\bar{G}^{\theta\theta}-i\bar{c}\times G^{\theta}+b\times G-ic\times f+C^{\theta}\times K=0,$$

$$K^{\theta\theta}+i\bar{c}\times\bar{G}^{\theta}-\bar{b}\times G+ic\times\bar{f}-\bar{C}^{\theta}\times K=0,\tag{B6}$$

$$\bar{c}\times G^{\theta\theta}+b\times\bar{G}^{\theta}-\bar{b}\times G^{\theta}+\bar{C}^{\theta\theta}\times G+c\times K^{\theta\theta}-C^{\theta}\times\bar{f}+\bar{C}^{\theta}\times f+C^{\theta\theta}\times K=0,$$

$$\bar{G}^{\theta}+\bar{f}+c\times\bar{G}+\bar{c}\times K=0,$$

$$K^{\theta\theta}-ic\times\bar{G}^{\theta}+C^{\theta}\times\bar{G}-i\bar{c}\times f+b\times K=0,$$

$$\bar{G}^{\bar{\theta\theta}}+ic\times\bar{G}^{\bar{\theta}}-C^{\bar{\theta}}\times\bar{G}+i\bar{c}\times\bar{f}-\bar{b}\times K=0, \tag{B7}$$

$$c\times\bar{G}^{\bar{\theta\theta}}-C^{\theta}\times\bar{G}^{\bar{\theta}}+C^{\bar{\theta}}\times\bar{G}^{\theta}+C^{\bar{\theta\theta}}\times\bar{G}+\bar{c}\times K^{\bar{\theta\theta}}-b\times\bar{f}+\bar{b}\times f+\bar{C}^{\bar{\theta\theta}}\times K=0,$$

$$\begin{gathered}
G^{\theta}+c\times G=0,\\
c\times G^{\theta}+iC^{\theta}\times G=0,\\
G^{\bar{\theta\theta}}+ic\times G^{\bar{\theta}}-C^{\bar{\theta}}\times G=0,\\
c\times G^{\bar{\theta\theta}}-C^{\theta}\times G^{\bar{\theta}}-iC^{\bar{\theta}}\times G^{\theta}+C^{\bar{\theta\theta}}\times G=0,
\end{gathered} \tag{B8}$$

$$\begin{gathered}
\bar{G}^{\bar{\theta}}+\bar{c}\times\bar{G}=0,\\
\bar{c}\times\bar{G}^{\bar{\theta}}+i\bar{b}\times\bar{G}=0,\\
\bar{G}^{\bar{\theta\theta}}-i\bar{c}\times\bar{G}^{\theta}-b\times\bar{G}=0,\\
\bar{c}\times\bar{d}^{\bar{\theta\theta}}-b\times\bar{G}^{\bar{\theta}}+\bar{b}\times\bar{G}^{\theta}+\bar{C}^{\bar{\theta\theta}}\times\bar{G}=0.
\end{gathered} \tag{B9}$$

We observe that the component relations that permitted us to write the superfield expressions (V4)–(V6) are contained in the first expressions of the set (B1)–(B9). As usual in this formalism, some of the remaining expressions are consistency relations. As an example, we mention the second and fourth equations of (B3) and (B4), the last three equations of (B5), and others. The new fact here is that it is possible to show that the superfields related to ghost-for-ghost are zero. For example, the combination of (II.8) and the two first equation (B1) and (B2) leads to $D_{[\mu}D_{\nu]}G=0$ and $D_{[\mu}D_{\nu]}G=0$. Since our formulation is not on-shell, one concludes that $G=0$ and $G=0$. Similarly, after hard algebraic work, one obtains that all components of the superfields related to ghost-for-ghosts are zero.

[1] C. Becchi, A. Rouet, and R. Stora, Ann. Phys. (N.Y.) **98**, 287 (1976); I. V. Tyutin, Lebedev preprint 39/1975 (unpublished).
[2] G. Curci and R. Ferrari, Phys. Lett. **63B**, 51 (1976).
[3] S. Ferrara, O. Piguet, and M. Schweda, Nucl. Phys. **B119**, 493 (1977); K. Fujikawa, Prog. Theor. Phys. **59**, 2045 (1977); L. Bonora and M. Tonin, Phys. Lett. **98B**, 48 (1981); R. Delbourgo, P. D. Jarvis, and G. Thomson, Phys. Rev. D **26**, 775 (1982); Phys. Lett. **109B**, 25 (1982); R. Delbourgo and P. D. Jarvis, J. Phys. A **15**, 611 (1982); R. Delbourgo, P. D. Jarvis, and R. C. Warner, J. Math. Phys. **34**, 3616 (1993).
[4] K. Fujikawa, Prog. Theor. Phys. **63**, 1364 (1980).
[5] K. Fujikawa, Phys. Rev. D **25**, 2584 (1982).
[6] I. A. Batalin and G. A. Vilkovisky, Phys. Lett. **102B**, 27 (1981); Phys. Rev. D **28**, 2567 (1983).
[7] J. Alfaro and P. H. Dangaard, Nucl. Phys. **B404**, 751 (1993).
[8] N. R. F. Braga and A. Das, Nucl. Phys. **B442**, 669 (1995).
[9] N. R. F. Braga and S. M. de Souza, Phys. Rev. D **53**, 916 (1996); E. M. C. Abreu and N. R. F. Braga, *ibid.* **54**, 4080 (1996).
[10] L. Baulieu, hep-th/9512026.
[11] P. A. Marchetti and M. Tonin, Nuovo Cimento A **63**, 459 (1981).
[12] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1974).
[13] E. Cremmer and J. Scherk, Nucl. Phys. **B72**, 117 (1974).
[14] T. J. Allen, M. J. Bowick, and A. Lahiri, Mod. Phys. Lett. A **6**, 559 (1991).
[15] R. Amorim and J. Barcelos-Neto, Mod. Phys. Lett. A **10**, 917 (1996).
[16] J. Barcelos-Neto and S. Rabello, Z. Phys. C **74**, 715 (1997).
[17] See, for example, M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, England, 1987).
[18] P. A. M. Dirac, Can. J. Math. **2**, 129 (1950); *Lectures on Quantum Mechanics* (Yeshiva University, New York, 1964).
[19] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, Princeton, NJ, 1992).
[20] D. Z. Freedman and P. K. Townsend, Nucl. Phys. **B177**, 282 (1981).
[21] See, for example, S. P. Alwis, M. T. Grisaru, and L. Mezincescu, Phys. Lett. B **190**, 122 (1987); A. H. Diaz, *ibid.* **203**, 408 (1988); L. Baulieu, E. Bergshoeff, and E. Sezgin, Nucl. Phys. **B307**, 348 (1988); J. Thierry-Meig, *ibid.* **B335**, 334 (1990); G. Barnich, R. Constantinescu, and P. Grégoire, Phys. Lett. **B292**, 353 (1992). For a superspace treatment along similar lines of the present article see J. Thierry-Mieg and L. Baulieu, Nucl. Phys. **B228**, 259 (1983).

[22] S. E. C. G. Stuckelberg, Helv. Phys. Acta **30**, 209 (1957).
[23] J. Barcelos-Neto, A. Cabo, and M. B. D. Silva, Z. Phys. C **72**, 345 (1996).
[24] A. Lahiri, Phys. Rev. D **55**, 5045 (1997).
[25] L. Baulieu and J. Thierry-Mieg, Phys. Lett. B **144**, 221 (1984).
[26] G. F. Chapline and N. Manton, Santa Barbara preprint NSF-ITP 82/85, 1982.
[27] L. Baulieu, Phys. Lett. B **126**, 455 (1983); Nucl. Phys. **B227**, 157 (1983).
[28] See Ref. 23 for a treatment without differential forms.
[29] I. A. Batalin, P. M. Lavrov, and I. V. Tyutin, J. Math. Phys. **31**, 1487 (1990).

Expansion of the hydrogen-like wave functions in the inverse principal quantum number, for the discrete and continuous spectra

Bruno Blaive[a] and Michel Cadilhac
CNRS, Ecole Nat. Sup. Chimie, Faculty of St-Jerome, 13397 Marseille Cedex 20, France

(Received 29 April 1998; accepted for publication 12 August 1998)

An expansion with respect to the inverse principal quantum number n^{-1} has been derived recently for the hydrogen-like wave functions of the discrete spectrum. The basic functions of this expansion were Bessel functions multiplied by a power r^k of the distance r to the origin. We show that this expansion remains valid for the continuous spectrum, with minor changes. For the wave function corresponding to the energy $E=+1/(2n^2)$ a.u. of the continuum, the pure imaginary number $-in$ (positive real n) must be substituted for the integer n in the expansion for the discrete spectrum, and a normalizing factor must be introduced. © *1998 American Institute of Physics.* [S0022-2488(98)03811-0]

I. INTRODUCTION

The energy $E=0$ is an upper limit for the eigenenergy corresponding to the atomic electronic Schrödinger wave functions of the hydrogen discrete spectrum ($E=-1/(2n^2)\rightarrow 0^-$), and a lower limit for those of the continuous spectrum ($E=+1/(2n^2)\rightarrow 0^+$). In a recent paper,[1] we derived an expansion of the discrete hydrogen-like wave functions $\psi_{nlm}(r,\theta,\varphi)=R_{nl}(r)Y_l^m(\theta,\varphi)$ in the vicinity of $E=0^-$, with respect to the inverse principal quantum number n^{-1}, for fixed quantum numbers l,m:

$$R_{nl}(r)=\frac{2}{n^{3/2}}\left[\prod_{k=1}^{l}\left(1-\frac{k^2}{n^2}\right)^{-1/2}\right]e^{-x/(2n)} \times\left[x^{-1/2}J_{2l+1}(2\sqrt{x})-\sum_{q=1}^{\infty}\frac{1}{n^q}\sum_{k=q}^{2q}s_{q,k}(l)x^{(k-1)/2}J_{k-2l-1}(2\sqrt{x})\right], \tag{1}$$

where

$$x=2r. \tag{2}$$

The coefficients $s_{q,k}(l)$ $(q\leqslant k\leqslant 2q)$ of the expansion are polynomials of l that satisfy the recursion formula,

$$s_{1,1}=-l, \quad s_{1,2}=-\tfrac{1}{2}, \tag{3}$$

$$s_{q,k}=\left(1-\frac{l+1}{k}\right)s_{q-1,k-1}-\frac{1}{k}s_{q-1,k-2}. \tag{4}$$

Equation (1), which has no connection with previous expansions involving Bessel functions,[2–5] is founded on a development of the Laguerre polynomial $L_n^{\alpha}(x/n)$ with respect to n^{-1}. The idea of the expansion (1) originated from the calculation of quantum integrals involving discrete atomic wave functions. However, one also finds in the literature cases where such an expansion would be useful for the continuous spectrum, as in the theory of collisions, where a zeroth order expansion is used.[6,7] In the present article we shall examine how to obtain an equation analogous to Eq. (1) for the wave functions of the continuum.

[a]Electronic mail: blaive2spi-chim.u-3mrs.fr

 © 1998 American Institute of Physics

II. RECALL: THE RADIAL WAVE FUNCTION OF THE CONTINUUM, AND THE CORRESPONDENCE WITH THE DISCRETE SPECTRUM

The Schrödinger radial equation is a second-order linear differential equation, whose complex solutions $R(r)$ belong to a two-dimensional complex space. Since the wave functions are also submitted to a regularity condition,[8] they must be proportional to a particular solution R_{El}:

$$R(r)=cR_{El}(r), \tag{5}$$

where c is any complex number. Being real, the Schrödinger radial equation has a real normalized solution (the real part, or the imaginary part, of a solution is a solution), which is unique if it is chosen to be positive for small r. This solution can be expressed as [Ref. 9, Eq. (4.23)]

$$R_{El}(r)=\frac{2}{\sqrt{1-e^{-2\pi|n|}}}\prod_{q=1}^{l}(n^2+q^2)^{1/2}\frac{(2r/n)^l}{(2l+1)!}e^{-ir/n}F(in+l+1,2l+2,2ir/n), \tag{6}$$

where $n=(2E)^{-1/2}$, and all the other complex radial wave functions differ from $R_{El}(r)$ by a uniform phase factor $c=e^{i\alpha}$.

A correspondence can be established[2] with the discrete wave functions,[8]

$$R_{nl}(r)=N_{nl}\left(\frac{2r}{n}\right)^l e^{-r/n}(n+l)!L_{n-l-1}^{2l+1}\left(\frac{2r}{n}\right), \tag{7}$$

$$=\frac{2}{(2l+1)!}n^{-3/2}\prod_{q=1}^{l}(n^2-q^2)^{1/2}\left(\frac{2r}{n}\right)^l e^{-r/n}F\left(-n+l+1,2l+2,\frac{2r}{n}\right). \tag{8}$$

Equation (8) devolves from the proportionality between the Laguerre polynomial L and a finite confluent hypergeometric series F, and from the identity

$$\binom{n+l}{2l+1}\frac{2}{n^2}\sqrt{\frac{(n-l-1)!}{(n+l)!}}=\frac{2}{(2l+1)!}n^{-3/2}\prod_{q=1}^{l}(n^2-q^2)^{1/2}. \tag{9}$$

When we perform in Eq. (8), the change of variable,

$$n(n\text{natural integer})\mapsto -in(n\text{positive real number}),$$

the coefficients i^l arising from different factors cancel, yielding the function

$$R_{-in,l}(r)=e^{-i\pi/4}n^{-3/2}\sqrt{1-e^{-2\pi n}}R_{El}(r). \tag{10}$$

In the following we consider only positive real numbers n, since opposite values of n correspond to the same energy E. R_{El} is normalized with respect to the energy $E=+1/(2n^2)$; let R'_{El} be the same function renormalized with respect to n. R'_{El} is defined in such a way that the integral over the spectrum,

$$\int_0^\infty dE\, R_{El}^* R_{El}=\int_0^\infty \frac{dn}{n^3}R_{El}^* R_{El}, \tag{11}$$

is written in the form

$$\int_0^\infty dn\, R_{El}^{\prime *} R'_{El}, \tag{12}$$

i.e.,

$$R'_{El}(r)=R_{El}(r)/n^{3/2}. \tag{13}$$

Equations (10) and (13) show that the change of variable $n \mapsto -in$ made in the discrete wave function yields the continuous wave function R'_{El} multiplied by the coefficient

$$e^{-i\pi/4}\sqrt{1-e^{-2\pi n}}. \tag{14}$$

This correspondence between the two spectra suggests that expansion (1) could remain valid for the continuum when n is replaced by $-in$, apart from a factor equal to (14).

III. EXPANSION OF THE CONFLUENT HYPERGEOMETRIC FUNCTION $F(in+l+1,2l+2,2ir/n)$

In the case of the discrete spectrum, the expansion of ψ_{nlm} with respect to n^{-1} has been obtained by using a Laplace transform, applied to the Laguerre polynomial of Eq. (7). In this section, the same method will be generalized to the confluent hypergeometric function of Eq. (6). We shall refer to Ref. 1 for the steps of the derivation that will remain unchanged. Several additional difficulties due to infinite series and complex variables will be described.

The hypergeometric function F involved in Eq. (6) is defined by

$$F\left(in+l+1,2l+2,\frac{ix}{n}\right)=\frac{\Gamma(2l+2)}{\Gamma(in+l+1)}\sum_{k=0}^{\infty}\frac{\Gamma(in+l+k+1)}{\Gamma(2l+k+2)k!}\left(\frac{ix}{n}\right)^k. \tag{15}$$

The Laplace transform of (15), with respect to the variable x, is given by

$$(\mathcal{L}F)(p)=\frac{\Gamma(2l+2)}{\Gamma(in+l+1)}\sum_{k=0}^{\infty}\frac{\Gamma(in+l+k+1)}{\Gamma(2l+k+2)}\left(\frac{i}{n}\right)^k\frac{1}{p^{k+1}}. \tag{16}$$

The p domain of convergence of the series (16) is given by

$$\left|\frac{i}{np}\right|<1, \tag{17}$$

i.e., $|p|>1/n$, a domain that will allow us to define the incoming inverse Laplace transform. Inequality (17) is deduced easily from the Sterling formula, which gives an equivalent of the series coefficient when k tends to infinity:

$$\Gamma(k+\alpha)/\Gamma(k+\gamma)\sim k^{\alpha-\gamma}. \tag{18}$$

Using the identity

$$\Gamma(x)\Gamma(1-x)=(-1)^{x-y}\Gamma(y)\Gamma(1-y), \tag{19}$$

which is valid for complex numbers x and y whose difference $x-y$ is an integer, we obtain

$$(\mathcal{L}F)(p)=\Gamma(2l+2)p^{2l}(in)^{2l+1}\sum_{k=0}^{\infty}\frac{\Gamma(-in-l)}{\Gamma(-in-l-k)\Gamma(2l+k+2)}\left(\frac{-i}{np}\right)^{2l+1+k}. \tag{20}$$

In Eq. (20), let us set $q=2l+1+k$, and insert the identity

$$\frac{\Gamma(-in+l+1)}{\Gamma(-in-l)}=(-1)^{l+1}in\prod_{j=1}^{l}(n^2+j^2). \tag{21}$$

We then obtain

$$(\mathcal{L}F)(p)=Cp^{2l}\sum_{q=2l+1}^{\infty}\frac{\Gamma(-in+l+1)}{\Gamma(-in+l+1-q)q!}\left(\frac{-i}{np}\right)^q, \tag{22}$$

where C represents the constant

$$C \equiv -(2l+1)!\, n^{2l} \prod_{j=1}^{l} (n^2+j^2)^{-1}. \tag{23}$$

Equation (22) can be rewritten in the form

$$(\mathcal{L}F)(p) = Cp^{2l} D(p) - \sum_{q=0}^{2l} b_q p^{2l-q}, \tag{24}$$

with

$$D(p) = \sum_{q=0}^{\infty} \frac{\Gamma(-in+l+1)}{\Gamma(-in+l+1-q)q!} \left(\frac{-i}{np}\right)^q; \tag{25}$$

hence

$$D(p) = \exp\left[(-in+l)\mathrm{Log}\left(1-\frac{i}{np}\right)\right], \tag{26}$$

where Log represents the principal determination of the complex logarithm (the imaginary part comprised between $-\pi$ and π). Equation (26), which generalizes the Newton binomial formula to a nonreal power $-in+l$, can be readily verified in the following manner: if we consider expression (26) as a function $f(z)$ of the variable $z=1-i/(np)$, its nth derivatives give the coefficients of the series (25). This property, together with the limit $D(p)\to 1$ when $p\to+\infty$, shows that the series (25) is merely the Taylor series of $f(z)$.

The constants b_q in the second term of (24) are unimportant. In fact, in the inverse Laplace transform that will be performed later, the second term of (24) will bring only a delta distribution and its derivatives $\delta^{(k)}$. Since F is a function, these terms must be compensated for in the final result by terms arising from $p^{2l}D(p)$.

With the view to expanding $D(p)$ into powers of n^{-1}, we first expand its complex logarithm,

$$\mathrm{Log}\, D(p) = (-in+l)\mathrm{Log}\left(1-\frac{i}{np}\right) + 2iK\pi \tag{27}$$

(K integer). We have

$$\mathrm{Log}\, D(p) = i\sum_{k=1}^{\infty} \frac{1}{kn^{k-1}} \left(\frac{i}{p}\right)^k - l\sum_{k=1}^{\infty} \frac{1}{kn^k} \left(\frac{i}{p}\right)^k + 2iK\pi, \tag{28}$$

$$= -\frac{1}{p} - \sum_{k=1}^{\infty} \frac{1}{n^k} i^k a_k(p) + 2iK\pi, \tag{29}$$

where we have introduced the same coefficients,

$$a_k(p) \equiv \frac{l}{kp^k} + \frac{1}{(k+1)p^{k+1}}, \tag{30}$$

as in Ref. 1. Apart from the coefficient i^k, Eqs (29), (30) coincide with those obtained for the discrete spectrum. The derivation can, therefore, be ended in a similar manner. Let us expose only the differences due to the new coefficient i^k: in the equation,

$$D(p) = e^{-1/p}\left(1 + \sum_{q=1}^{\infty} \frac{1}{n^q} \sum_{k=q}^{2q} \frac{s_{q,k}(l)}{p^k}\right) \quad (q \geqslant 1), \tag{31}$$

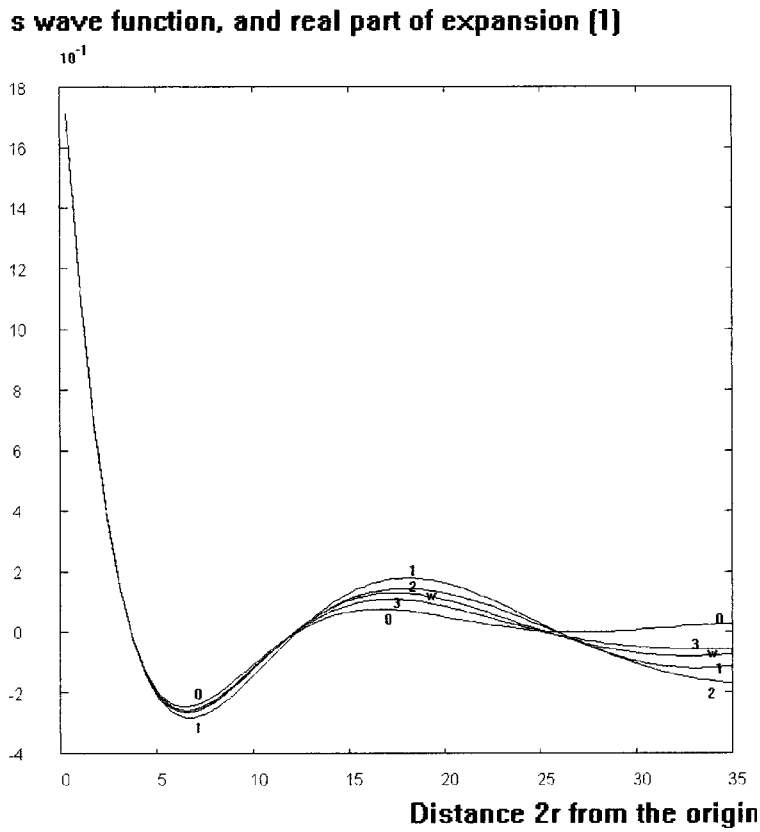

FIG. 1. A comparison between the R_{Es} $(n=9,\ l=0)$ radial wave function of the continuous spectrum of the hydrogen atom, and its expansion (1) to order 0, 1, 2, 3. The following radial functions are drawn: Curve 0: zeroth-order expansion (term $q=0$), $\mathrm{Re}[2e^{-ix/(2n)}J_1(2\sqrt{x})/\sqrt{x}]=2\cos(x/(2n))\,J_1(2\sqrt{x})/\sqrt{x}$. Curve 1: first-order expansion $(q=0,1)$, $\mathrm{Re}[2e^{-ix/(2n)}\times[1/\sqrt{x}+(i/2n)\sqrt{x}]J_1(2\sqrt{x})]$. Curve 2: second-order expansion $(q=0,1,2)$, $idem+\mathrm{Re}[2/n^2e^{-ix/(2n)}[-(x/3)J_2(2\sqrt{x})+x^{3/2}J_3(2\sqrt{x})/8]]$. Curve 3: third-order. Curve w: wave function, $2e^{-ix/(2n)}F(in+1,2,ix/n)$.

which has been derived formerly for the discrete spectrum, the polynomials $s_{q,k}(l)$ should now be replaced by $i^q s_{q,k}(l)$. This means that Eq. (31) remains valid unchanged, provided that n is replaced by $-in$. Like in Ref. 1, the inverse Laplace transform of $p^{2l}D(p)$ in Eq. (24) is a sum of terms of the type

$$\mathcal{L}^{-1}(e^{-1/p}p^k)=(-1)^{k+1}\left[x^{-(k+1)/2}J_{k+1}(2\sqrt{x})\Upsilon(x)-\sum_{q=0}^{k}(-1)^q\frac{\delta^{(q)}}{(k-q)!}\right], \tag{32}$$

and this leads to Eq. (1).

IV. CONCLUSION

We have shown that expansion (1) can be used to compute the wave function R'_{El} of the continuum, corresponding to the energy $E=+1/(2n^2)$, normalized with respect to n, provided that the following changes are made. First, all the occurrences of the integer n in Eq. (1) must be replaced by $-in$, where $n=1/\sqrt{2E}$ now is a positive real number. Second, the rhs of Eq. (1) must be multiplied by the coefficient

$$c_n=e^{i\pi/4}(1-e^{-2\pi n})^{-1/2}, \tag{33}$$

i.e.,

$$R'_{El}(r)=c_nR_{-in,l}(r). \tag{34}$$

The wave function R_{El} normalized with respect to the energy E is related to R'_{El} by Eq. (13).

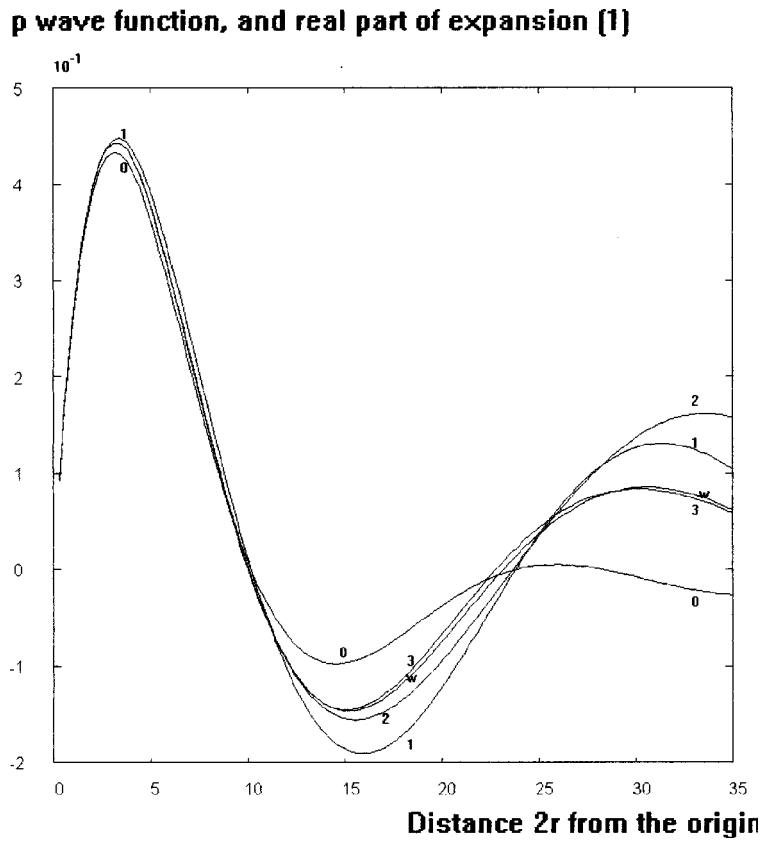

FIG. 2. A comparison between the R_{Ep} $(n=9,l=1)$ radial wave function and its expansion (1) to order 0, 1, 2, 3. The following radial functions are drawn: Curve 0: zeroth-order expansion (term $q=0$), $\mathrm{Re}[2e^{-ix/(2n)}(1+1/n^2)^{-1/2}J_3(2\sqrt{x})/\sqrt{x}]$. Curve 1: first-order expansion $(q=0,1)$, $\mathrm{Re}[2e^{-ix/(2n)}(1+1/n^2)^{-1/2}[1/\sqrt{x}+i/(2n)\sqrt{x}]J_3(2\sqrt{x})]$. Curve 2: second-order expansion $(q=0,1,2)$, $idem+\mathrm{Re}[(2/n^2)e^{-ix/(2n)}(1+1/n^2)^{-1/2}[(x/6)J_0(2\sqrt{x})+x^{3/2}J_1(2\sqrt{x})/8]]$. Curve 3: third-order. Curve w: wave function, $e^{-ix/(2n)}\sqrt{1+n^2}x/(3n)F(in+2,4,ix/n)$.

For practical physical calculations, $R'_{El}(r)$, given by (34) and the series (1), will always be approximated by $R'_{El,Q}(r)$, its partial sum truncated at $q=Q$. The partial sums are real for the discrete spectrum, and complex for the continuous spectrum. In the latter case, their limit is real, thanks to the phase factor chosen in (33). One can use either expansion (1) itself, or its real part, both of them being equal to R'_{El}. We have shown that

$$\lim_{Q\to+\infty}[\mathrm{Re}\,R'_{El,Q}(r)+i\,\mathrm{Im}\,R'_{El,Q}(r)]=R'_{El}(r), \tag{35}$$

which implies

$$\lim_{Q\to+\infty}[\mathrm{Im}\,R'_{El,Q}(r)]=0, \tag{36}$$

since R'_{El} is real. The order of magnitude of the imaginary part is a clue to estimate the error made in the real part of the expansion.

The convergence of expansion (1) is uniform on a bounded interval $0<r<r_{\max}$. Practically, this means that the order Q of the expansion that is necessary for a good approximation of $R'_{El}(r)$ increases with $r_{\max}$. Conversely, for a given order Q, arbitrarily large discrepancies between $R'_{El}(r)$ and its approximation $R'_{El,Q}(r)$ are observed for large radii r. This is exemplified for the third-order approximation $(Q=3)$ of s and p states, for level $n=9$ (Figs. 1–3).

In this study, we also verify that the wave functions are analytical functions of the parameter $\zeta=1/n$ (discrete spectrum) or $\zeta=i/n$ (continuous spectrum). Therefore, the transition between the two types of spectra $(\zeta=0)$ is completely regular.

Equation (1) suggests interesting mathematical properties of expansions on the family of complex functions,

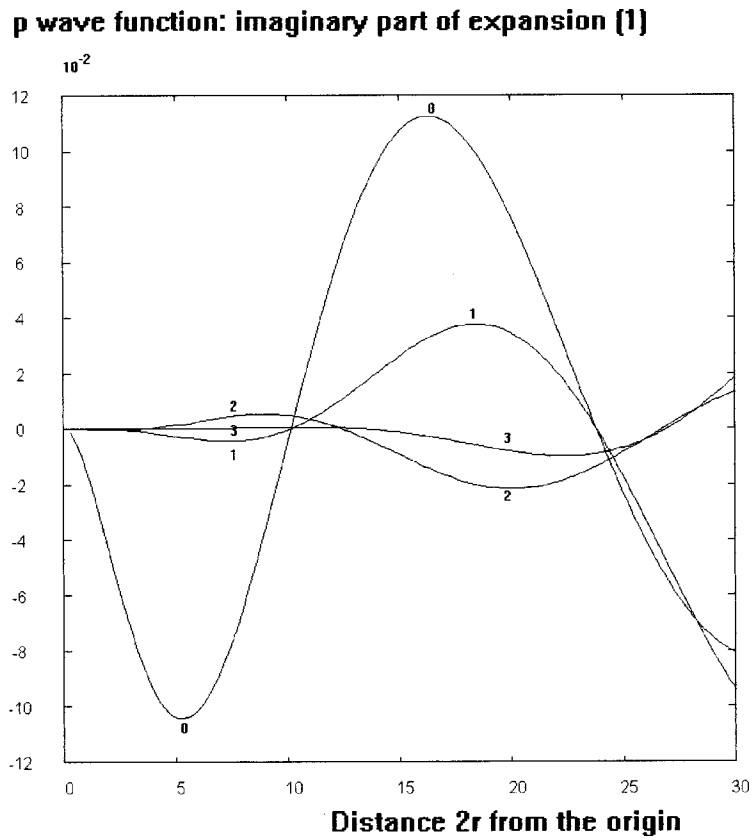

FIG. 3. Imaginary part of expansion (1) for R_{Ep} $(n=9,l=1)$. The same functions as in Fig. 2, with Im instead of Re. Curves 0, 1, 2, 3 correspond to zeroth, first, second, and third order expansions.

$$z^{n/2}J_n(2\sqrt{z}), \tag{37}$$

in connection with the Nielsen series.[10,11] These properties will be dealt with separately.

APPENDIX: GLOSSARY

$F(a,b,z)=\Gamma(b)/\Gamma(a)\Sigma_{k=0}^{\infty}\Gamma(a+k)/(\Gamma(b+k)k!)z^k$, confluent hypergeometric or Kummer function ${}_1F_1$

$J_n(z)=(z/2)^n\Sigma_{k=0}^{\infty}(-1)^k(z/2)^{2k}/(k!(n+k)!)$, Bessel function

$L_n^{\alpha}(x)=\Sigma_{k=0}^{n}(-1)^k\binom{n+\alpha}{n-k}x^k/k!$, Laguerre polynomial [this definition[12] is the definition of Ref. 8 divided by $(n+\alpha)!$].

$\mathcal{L}$, Laplace transform $(\mathcal{L}f)(p)=\int_0^{\infty}dx\,e^{-px}f(x)$

$Y_l^m(\theta,\varphi)$, spherical harmonics

$\delta^{(k)}$, kth derivative of the Dirac distribution δ

$\Gamma(z)$, Euler Gamma function

$\Upsilon(x)$, Heaviside function [$\Upsilon(x)=0$ for $x<0$, 1 for $\geqslant 0$].

[1] B. Blaive and M. Cadilhac, J. Math. Phys. **38**, 6061 (1997).
[2] V. Dose and C. Semini, Helv. Phys. Acta **47**, 307 (1974).
[3] E. Filter and E. O. Steinborn, Phys. Rev. A **18**, 1 (1978).
[4] D. Antolovic and J. Delhalle, Phys. Rev. A **21**, 1815 (1980).
[5] A. Holas, S. Olszewski, and A. Wierzbicki, Phys. Rev. B **32**, 7913 (1985).
[6] T. Mukoyama and L. Sarkadi, Bull. Inst. Chem. Res., Kyoto Univ. **64**, 307 (1986).
[7] T. Mukoyama and L. Sarkadi, Bull. Inst. Chem. Res., Kyoto Univ. **66**, 11 (1988).
[8] A. Messiah, *Quantum Mechanics* (Elsevier, Amsterdam, 1961).
[9] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Plenum, New York, 1977).
[10] N. Nielsen, *Handbuch der Theorie der Cylinderfunktionen* (Teubner, Leipzig, 1904).
[11] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed. (Cambridge University Press, Cambridge, 1944).
[12] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980).

Complex Hamiltonian evolution equations and field theory

Theodore Bodurov[a)]
Eugene, Oregon 97405

(Received 22 August 1997; accepted for publication 23 July 1998)

The form of the complex Hamiltonian evolution equations (linear and nonlinear) is deduced here from that of the classical Hamiltonian systems. Their properties and those of the associated continuous Poisson bracket are studied. It is shown that if any space-localized solution of the above type equations is expanded in any Hilbert space basis the expansion coefficients are canonically conjugate variables and obey Hamiltonian systems of complex equations. The last is proved also for sets of ''canonically conjugate'' functionals of general type. As an application, commutation relations which are equivalent in form to Dirac's quantization rules (the constant in place of $\hbar$ is undetermined) are obtained. The question of how a nonlinear field theory may contain quantum mechanics as a special case is discussed in view of this paper's results. © *1998 American Institute of Physics.*
[S0022-2488(98)03211-3]

I. INTRODUCTION

Since the time of quantum mechanics' first successes many physicists, led by diverse motivations, have proposed nonlinear extensions for quantum mechanics and quantum field theory. Among the most prominent were de Broglie[1] and Heisenberg.[2] A more recent contribution which has attracted much attention is from Weinberg.[3]

The underlying idea in Weinberg[3] may be stated as follows: *To study the role of possible nonlinearities in quantum mechanics, and field theory in general, one should study the mathematical properties of those nonlinear field equations whose form is as close as possible to that of the currently accepted fundamental equations, and which form contains the latter as special cases*. This is also the guiding principle for the present work. However, unlike Weinberg, we do not assume that the nonlinearities are small. Instead, we assume that the nonlinear field equations are such that they possess *space-localized* solutions. By doing this we not only extend the validity of Weinberg's general formalism, but we also find other results which would have been inaccessible otherwise.

Just as the physics motivations of the various contributors differ substantially, the nonlinear field equations or the formalisms which are assumed in the respective contributions also differ. Illustrations for this state of affairs are the following samples from the pertinent literature: Finkelstein and Fronsdal,[4] Soler,[5] Bialynicki-Birula and Mycielski,[6] Freidberg and Lee,[7] Cooperstock and Rosen,[8] Enz,[9] Reinisch,[10] Bodurov,[11] Schwartz,[12] in addition to the references given earlier. It is apparent, even from this short list, that many physicists share the view that it makes a lot of sense to represent particles by self-localized solutions of nonlinear field equations, although no one is sure what these equations should be beyond the model-making stage.

The work of Berestycki and Lions,[13] which does not seem to be well known, is cited as an illustration of how rich the nonlinear field equations are. There it is proved that even simple field equations like some of the free nonlinear Schrödinger equations (NLS), in three space dimensions, possess infinitely many space-localized solutions starting from a ground-state solution, all of which have finite Hermitian norms and are singularity free. Thus, there is a complete analogy with the solutions of the Schrödinger equation proper for a particle trapped in some potential well. But while the solutions of the fundamental equations in quantum mechanics (QM), due to their linearity, are space-localized (bound states) only if an appropriate external potential is present, the above NLS equations admit solutions which are space-localized (or self-localized) solely by the field's self-interaction.

A generalization of Ehrenfest's theorem which applies to the self-localized solutions of all nonlinear Schrödinger equations, and a corollary of it (pertaining to the motion of the region of

a)Electronic mail: bodt@efn.org

 © 1998 American Institute of Physics

localization) are proved and discussed in a recent paper, Bodurov.[14] The generalized Ehrenfest theorem establishes a link between those field equations and classical mechanics/electrodynamics, just as the Ehrenfest theorem proper establishes the relation of quantum mechanics to classical physics. The question now arises: *What is the generic form of the field equations for which such results apply?* One objective of the present paper is to offer a possible answer and to discuss its implications.

In this paper, the term ''field'' will mean exclusively a function of the space coordinates and the time whose components are complex-valued, in general, and obey certain transformation rules. That is, the fields considered here are not quantized.

The organization of the paper is as follows. Section II introduces the form of a family of field equations, here called *complex Hamiltonian evolution* (CHE) equations, by deducing it from the form of the classical Hamiltonian systems and shows that it contains Schrödinger and Dirac equations. Section III defines the *continuous complex Poisson bracket* and uses it to derive some of the general properties of the complex Hamiltonian evolution equations. Section IV proves that the expansion coefficients of *any* space-localized solution of a CHE equation in terms of *any* Hilbert space basis satisfy complex Hamiltonian systems of equations. The expansion coefficients and their conjugates, treated as linear functionals, obey canonical Poisson bracket relations. Further, it is found that any set of ''canonically conjugate'' functionals also satisfies a Hamiltonian systems of equations, thus providing an equivalent description for the time evolution of space-localized fields. Section V shows, as an application of the previous results, that the same lead to commutation relations whose form is that of Dirac's quantization rules. Consequently, the *complementarity* of the classical canonical variables with the operators of quantum mechanics may be seen as a property of the field dynamics which is given by the complex Hamiltonian evolution equations. Section VI discusses the following: ''Can a QM extension which assumes the presence of only small nonlinearities be compatible with a nonlinear field theory based on equations which possess self-localized solutions?'' This inquiry leads to the key question: ''When may a nonlinear field theory be said to contain QM as a special case?'' and to its possible answer. This new, and rather unexpected, answer may be seen as the conclusion of the entire paper.

Notation: Here, all functionals will be denoted with Roman capital letters, all densities (the integrands of functionals) with script capital letters, all operators with bold roman capital letters, all fields with the Greek letters: $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}$. The rest of the symbols will be defined at their first appearance. The summation convention of repeated indexes is assumed for the entire paper. All integrals' domains are $\mathbb{R}^n$.

II. COMPLEX HAMILTONIAN EVOLUTION EQUATIONS FROM CLASSICAL HAMILTONIAN SYSTEMS

It is known that any Hamiltonian system

$$\frac{dq_i}{dt}=\frac{\partial}{\partial p_i}H(q,p,t),\quad \frac{dp_i}{dt}=-\frac{\partial}{\partial q_i}H(q,p,t), \tag{2.1}$$

with $i=1,2,...,s$ may be transformed into a system of complex equations

$$\frac{dz_i}{dt}=-i\frac{\partial}{\partial z_i^*}H(z^*,z,t)\quad \text{where}\quad \frac{\partial}{\partial z_i^*}=\frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial q_i}+i\frac{\partial}{\partial p_i}\right)\quad \text{or} \tag{2.2}$$

$$\frac{dz_i^*}{dt}=i\frac{\partial}{\partial z_i}H(z^*,z,t)\quad \text{where}\quad \frac{\partial}{\partial z_i}=\frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial q_i}-i\frac{\partial}{\partial p_i}\right), \tag{2.3}$$

whose form is the same as that of the original system, when the Hamiltonian function is expressed in terms of the complex variables

$$z_i=\frac{1}{\sqrt{2}}(q_i+ip_i)\quad \text{and}\quad z_i^*=\frac{1}{\sqrt{2}}(q_i-ip_i). \tag{2.4}$$

Thus, (2.4) may be considered to be a type of *canonical* transformation. For real-valued Hamiltonian functions $H=H(z^*,z,t)$ the systems (2.2) and (2.3) are equivalent. Then, one of them may be dropped without loss of information.

The first to point out the significance of the transformation (2.4) and its properties for quantum mechanics appears to be Strocchi.[15]

Actually, this transformation is permissible only if the *dimensionalities* of the variables q_i and p_i are the same. Otherwise, they cannot be combined into the complex variables z_i and z_i^*. Hence, the following simple canonical transformation must be applied first:

$$q_i'=c_i q_i \quad \text{and} \quad p_i'=\frac{p_i}{c_i} \quad (i=1,2,...,s, \quad \text{no summation}) \tag{2.5}$$

which will make the dimensionalities of q_i' and p_i' the same, if the dimensionalities of the constants c_i are chosen according to: $\mathrm{dms}(c_i)=\sqrt{\mathrm{dms}(p_i)/\mathrm{dms}(q_i)}$.

The transformation (2.5) together with (2.4) erase all distinctions between the two types of canonical variables. Since (2.5) does not alter the form of the Hamiltonian system (2.1), we may assume that it has already been applied when the complex transformation (2.4) is performed. Consequently, the dimensionalities of the squared magnitudes of all variables $z_i^*z_i=(q_i^2+p_i^2)/2$ are [*action*], in all complex Hamiltonian systems. The significance of this observation will become apparent at the end of Sec. V.

One finds easily that the transformation (2.4) preserves, within a multiplication constant, the Poisson bracket of any two real or complex-valued functions *R* and *S*, that is

$$\{R,S\}=\frac{\partial R}{\partial q_i}\frac{\partial S}{\partial p_i}-\frac{\partial S}{\partial q_i}\frac{\partial R}{\partial p_i}=\frac{1}{i}\left(\frac{\partial R}{\partial z_i}\frac{\partial S}{\partial z_i^*}-\frac{\partial S}{\partial z_i}\frac{\partial R}{\partial z_i^*}\right). \tag{2.6}$$

(The summation convention of repeated indexes is assumed for the entire paper.)

The next step, from where Strocchi[15] left, is to convert the complex system (2.2) into an evolution equation by letting the number of variables *s* in (2.2) become infinitely large. Then, in place of a system of infinitely many ordinary differential equations one will have a system of finitely many partial differential equations, which will describe the evolution of a multicomponent field.

A discussion of how to perform this transition, may be found in Olver[16] (Chap. 7). There it is shown that (in this paper notation)

$$\frac{\partial \eta_j}{\partial t}=\sum_k \mathbf{D}_{jk}\frac{\delta}{\delta \eta_k}\mathrm{H}[\boldsymbol{\eta};t] \tag{2.7}$$

is the most general form of the Hamiltonian evolution equations which are obtained from Hamiltonian systems with real variables. In this expression $\eta_j=\eta_j(x,t)$ are the components of the real field $\boldsymbol{\eta}$, $\mathbf{D}_{jk}$ are the entries of a skew-adjoint matrix-differential operator which may depend on $\boldsymbol{\eta}$, $\mathrm{H}[\boldsymbol{\eta};t]$ is the Hamiltonian functional for the field $\boldsymbol{\eta}$, and $\delta/\delta\eta_k$ is the variational derivative with respect to η_k.

Clearly, the Hamiltonian functional must remain finite during the entire time interval of interest, which necessitates the following

Definition 2.1: A scalar, vector or spinor-valued function $\boldsymbol{\psi}=\boldsymbol{\psi}(x)$ *of the coordinates* $x\in\mathbb{R}^n$ *will be called space-localized, or* $\boldsymbol{\psi}$*-field, if it is smooth* (C^∞), *hence singularity-free, and if its norm*

$$\mathrm{N}=\langle\boldsymbol{\psi},\boldsymbol{\psi}\rangle=\int_{\mathbb{R}^n}\psi_k^*\psi_k\,d^nx, \tag{2.8}$$

with the integral taken over all space, is finite.

Definition 2.2: A $\boldsymbol{\psi}$ *field will be called self-localized if it is a solution of a free (no interaction with external fields) field equation and if it remains space-localized for all time.*

Definition 2.3: All functionals considered in this paper (admissible functionals) are assumed to be finite when evaluated with a space-localized field.

Note that, the linear free field equations do not have space-localized solutions. The physics aspect of Definition 2.1 requires some comments, which are included in Sec. VI.

The form (2.7) is too general for the purposes of this paper. We would like to have a form which is as specific as possible, and yet, is general enough to contain as special cases the fundamental equations of quantum mechanics, together with the nonlinear equations which in some respect are related to the QM equations. Since the former are complex, such a form then has to represent a class of complex Hamiltonian evolution equations, both linear and nonlinear. It turns out that, the desired equation form may easily be deduced from the complex Hamiltonian system (2.2) by following the same process with which (2.7) was deduced. This is done by replacing (Olver,[16] Chap. 7) the set of all variables z_i with the $\boldsymbol{\psi}$ field, and the z_i^* with the $\boldsymbol{\psi}^*$ field; the Hamiltonian function $H(z^*,z,t)$ with the Hamiltonian functional $\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$; the ordinary time derivative d/dt with the partial time derivative $\partial/\partial t$; and the partial derivatives $\partial/\partial z_i$, $\partial/\partial z_i^*$ with the variational derivatives $\delta/\delta\psi_k$, $\delta/\delta\psi_k^*$, where $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_r)$ is an r-component field and $\boldsymbol{\psi}^*$ is its complex conjugate.

The result is the form of the *complex Hamiltonian evolution* (CHE) equations

$$\frac{\partial\psi_k}{\partial t}=-i\,\frac{\delta}{\delta\psi_k^*}\,\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t],\quad k=1,2,\ldots,r \tag{2.9}$$

for $\boldsymbol{\psi}=\boldsymbol{\psi}(x,t)$ which may be a *scalar, vector*, or *spinor* field. Here, the Hamiltonian functional H is an integral of some real-valued *Hamiltonian* density $\mathscr{H}$ over all space

$$\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]=\int_{\mathbb{R}^n}\mathscr{H}(x,t,\boldsymbol{\psi}^*,\boldsymbol{\psi},\partial\boldsymbol{\psi}^*,\partial\boldsymbol{\psi},\ldots)d^nx.$$

$\partial\boldsymbol{\psi}$ and $\partial\boldsymbol{\psi}^*$ stand for derivatives with respect to the space-coordinates x_i, $i=1,\ldots,n$. The latter are assumed to be Cartesian in order to avoid complications, unnecessary for the present purposes. The dimension of the space on which the fields are defined is n. Generally, $\mathscr{H}$ may contain derivatives of orders higher than the first. The variational derivative of H with respect to ψ_k^* is given with the *Euler–Lagrange expressions*

$$\frac{\delta\mathrm{H}}{\delta\psi_k^*}=\frac{\partial\mathscr{H}}{\partial\psi_k^*}-\frac{d}{dx_i}\,\frac{\partial\mathscr{H}}{\partial(\partial_i\psi_k^*)}+\frac{d^2}{dx_idx_j}\,\frac{\partial\mathscr{H}}{\partial(\partial_{ij}\psi_k^*)}-\cdots=\mathbf{H}_k(\boldsymbol{\psi}^*,\boldsymbol{\psi}), \tag{2.10}$$

where $\mathbf{H}_k(\boldsymbol{\psi}^*,\boldsymbol{\psi})$ denotes operators which may depend on x and t, and which act on $\boldsymbol{\psi}$, $\boldsymbol{\psi}^*$ nonlinearly in general.

Unlike the multiparticle wave functions of quantum mechanics, here, the $\boldsymbol{\psi}$ fields will never depend on more than three space coordinates, that is $n=1$, 2, or 3. This departure from the QM description of multiparticle systems is briefly discussed in Sec. VI.

Schrödinger and Dirac equations, written as evolution equations, have the forms

$$i\,\frac{\partial\psi}{\partial t}=\mathbf{H}\psi,\quad\text{and}\quad i\,\frac{\partial\psi_k}{\partial t}=\mathbf{H}_{kj}\,\psi_j,\quad k,j=1,\ldots,r, \tag{2.11}$$

with $\mathbf{H}$ being a scalar Hermitian operator and $\mathbf{H}_{kj}$ denoting the entries of a matrix-differential Hermitian operator. Then, we see immediately that the two fundamental equations of quantum mechanics are CHE equations, since the forms (2.11) are special cases of the form (2.9) obtained with the real-valued bilinear Hamiltonian functionals

$$\mathrm{H}=\int\psi^*\mathbf{H}\psi\,d^nx\quad\text{and}\quad\mathrm{H}=\int\psi_k^*\mathbf{H}_{kj}\,\psi_j\,d^nx, \tag{2.12}$$

respectively. This shows that the generic form (2.9) is of the desired type.

It is simple to verify that, Eq. (2.9) are the necessary conditions, i.e., the Euler–Lagrange equations, for the extremums of the real-valued functional

$$\int \mathrm{L}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]\,dt=\iint \mathscr{L}\,d^nx\,dt=\iint\left(\frac{i}{2}\left(\psi_k^*\,\frac{\partial\psi_k}{\partial t}-\frac{\partial\psi_k^*}{\partial t}\,\psi_k\right)-\mathscr{H}\right)d^nx\,dt.$$

This establishes that $\mathrm{L}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]$ is the *Lagrangian functional*, and that the relation of the *Lagrangian density* $\mathscr{L}$ of the $\boldsymbol{\psi}$ field to its *Hamiltonian* density $\mathscr{H}$ is

$$\mathscr{L}=\frac{i}{2}\left(\psi_k^*\,\frac{\partial\psi_k}{\partial t}-\frac{\partial\psi_k^*}{\partial t}\,\psi_k\right)-\mathscr{H}. \tag{2.13}$$

With this relation it is very simple to show that the *$\boldsymbol{\psi}$-field energy density*, i.e., *the $\mathscr{T}_{00}$ component of the stress-energy 4-tensor is equal to the Hamiltonian density $\mathscr{H}$* as follows:

$$\mathscr{T}_{00}=\frac{\partial\pi_i}{\partial t}\,\frac{\partial\mathscr{L}}{\partial(\partial_t\pi_i)}-\mathscr{L}=\frac{i}{2}\left(\psi_k^*\,\frac{\partial\psi_k}{\partial t}-\frac{\partial\psi_k^*}{\partial t}\,\psi_k\right)-\mathscr{L}=\mathscr{H}. \tag{2.14}$$

Here, the field $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_{2r})=(\psi_1,\ldots,\psi_r,\psi_1^*,\ldots,\psi_r^*)$ with $i=1,2,...,2r$ and $k=1,2,...,r$ is only a notational device.

The physics interpretation of an important result at the end of Sec. V depends on the validity of identifying the $\boldsymbol{\psi}$-field total energy with the value of the Hamiltonian functional. Equation (2.14) assures us that such an identification is well justified.

III. THE COMPLEX CONTINUOUS POISSON BRACKET

Now, we apply the same transition $s\to\infty$ (a finite to a continuous set of variables) on the complex Poison bracket (2.6) and obtain

$$\{\mathrm{R},\mathrm{S}\}=\frac{1}{i}\int_{\mathbb{R}^n}\left(\frac{\delta\mathrm{R}}{\delta\psi_k}\,\frac{\delta\mathrm{S}}{\delta\psi_k^*}-\frac{\delta\mathrm{S}}{\delta\psi_k}\,\frac{\delta\mathrm{R}}{\delta\psi_k^*}\right)d^nx, \tag{3.1}$$

which defines the *continuous complex Poisson bracket* of two real or complex-valued functionals $\mathrm{R}=\mathrm{R}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$ and $\mathrm{S}=\mathrm{S}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$. Both the discrete and the continuous Poisson brackets will be denoted by the same symbol $\{\cdot,\cdot\}$, unless noted otherwise, since it will be clear from the context to which type it refers.

Just as for the finite-dimensional Poisson bracket (2.6), so also for (3.1) it is straightforward but tedious to verify that it satisfies *Jacobi's identity*. Of course, the second variations of the functionals R and S are assumed to exist. The rest of the properties which qualify (3.1) to be a Poisson bracket are obvious.

The formalism proposed in Weinberg[3] is based on an expressly assumed product of homogeneous functions of degree one in their arguments ψ_k and ψ_k^*. Since this product is noncommutative it defines a bracket which is essentially the discrete counterpart of (3.1) and structurally the same as (2.6). The difference is that neither the functions $R(z^*,z)$ and $S(z^*,z)$ in (2.6) nor the functionals $\mathrm{R}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]$ and $\mathrm{S}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]$ in (3.1) are required to be homogeneous in their arguments. The physics implications of dropping or retaining this *homogeneity condition* will be discussed in Sec. VI.

Expression (3.1) is an exact form-replica of the ''classical'' continuous Poisson bracket

$$\{\mathrm{R},\mathrm{S}\}_{\mathrm{cl}}=\int\left(\frac{\delta\mathrm{R}}{\delta\eta_k}\,\frac{\delta\mathrm{S}}{\delta\pi_k}-\frac{\delta\mathrm{S}}{\delta\eta_k}\,\frac{\delta\mathrm{R}}{\delta\pi_k}\right)d^nx, \tag{3.2}$$

where η_k and $\pi_k=\partial\mathscr{L}/\partial\dot{\eta}_k$ are the components of two ''canonically conjugate'' fields. However, the complex Poisson bracket (3.1) and the one proposed in Weinberg[3] have an important advantage over the ''classical'' bracket. The problem with (3.2) is that its properties under Lorentz transformations are not well defined since the fields $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ behave differently under change of reference frame. This criticism does not apply to the brackets discussed here because $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$ transform, of course, as the same entity regardless of whether they are scalars, vectors, or spinors. Some details and a discussion of the difficulties with (3.2) may be found in Goldstein[17] (Sec. 12.14).

The time-derivative of any functional $\mathrm{R}=\mathrm{R}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$ is given in terms of the complex Poisson bracket by

$$\frac{d\mathrm{R}}{dt}=\int\left(\frac{\delta\mathrm{R}}{\delta\psi_k}\frac{\partial\psi_k}{\partial t}+\frac{\delta\mathrm{R}}{\delta\psi_k^*}\frac{\partial\psi_k^*}{\partial t}\right)d^nx+\frac{\partial\mathrm{R}}{\partial t}$$

$$=\frac{1}{i}\int\left(\frac{\partial\mathrm{R}}{\partial\psi_k}\frac{\delta\mathrm{H}}{\delta\psi_k^*}-\frac{\delta\mathrm{H}}{\delta\psi_k}\frac{\delta\mathrm{R}}{\delta\psi_k^*}\right)d^nx+\frac{\partial\mathrm{R}}{\partial t}=\{\mathrm{R},\mathrm{H}\}+\frac{\partial\mathrm{R}}{\partial t},\tag{3.3}$$

which is a direct consequence of (2.9). If the Hamiltonian functional $\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]$ does not depend explicitly on the time, then its value is constant, as seen from (3.3).

All bilinear functionals (and trivially, all linear field equations) share a common property, that is, they are invariant under the *gauge type I* transformation:

$$\psi_k'=\psi_k\mathbf{e}^{i\epsilon}\quad\text{and}\quad\psi_k'^*=\psi_k^*\mathbf{e}^{-i\epsilon},\tag{3.4}$$

with ϵ its parameter. Then, the nonlinear CHE equations whose Hamiltonian functionals are gauge type I invariant should have properties which are similar to those of the linear CHE equations. The rest of this section will show that, indeed, this is the case.

Theorem 3.1: *Let* $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_r)$ *be a field on a n-dimensional space, whose components* $\psi_k=\psi_k(x_1,\ldots,x_n,t)$ *are complex-valued and may depend on the time t. The continuous Poisson bracket (3.1) of the Hermitian norm* N *(2.8) of* $\boldsymbol{\psi}$ *with any real or complex-valued functional*

$$\mathrm{F}=\int\mathscr{F}(x,t,\boldsymbol{\psi},\boldsymbol{\psi}^*,\partial\boldsymbol{\psi},\partial\boldsymbol{\psi}^*,\ldots)d^nx$$

is zero if F *is invariant under the gauge type I transformation (3.4).*

Proof: Since N is also a functional its continuous Poisson bracket (3.1) with F is

$$\{\mathrm{N},\mathrm{F}\}=\frac{1}{i}\int\left(\frac{\delta\mathrm{N}}{\delta\psi_k}\frac{\delta\mathrm{F}}{\delta\psi_k^*}-\frac{\delta\mathrm{F}}{\delta\psi_k}\frac{\delta\mathrm{N}}{\delta\psi_k^*}\right)d^nx$$

$$=\frac{1}{i}\int\left(\psi_k^*\left(\frac{\partial\mathscr{F}}{\partial\psi_k^*}-\frac{d}{dx_i}\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k^*)}+\cdots\right)-\left(\frac{\partial\mathscr{F}}{\partial\psi_k}-\frac{d}{dx_i}\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k)}+\cdots\right)\psi_k\right)d^nx$$

$$=\frac{1}{i}\int\left(\psi_k^*\frac{\partial\mathscr{F}}{\partial\psi_k^*}-\psi_k\frac{\partial\mathscr{F}}{\partial\psi_k}+\partial_i\psi_k^*\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k^*)}-\partial_i\psi_k\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k)}+\cdots\right)d^nx,\tag{3.5}$$

where all terms with total derivatives were integrated by parts. The invariance condition for F, with $\mathscr{F}'=\mathscr{F}(x,t,\boldsymbol{\psi}',\boldsymbol{\psi}'^*,\partial\boldsymbol{\psi}',\partial\boldsymbol{\psi}'^*,\ldots)$ being its transformed density, is

$$\frac{d}{d\epsilon}\mathscr{F}'\bigg|_{\epsilon=0}=\boldsymbol{\nabla}\cdot\mathscr{C}$$

where $\mathscr{C}$ is an undetermined vector-valued density which becomes constant at infinity. Performing the differentiation and setting $\epsilon=0$ into the resulting expression we obtain

$$\frac{d}{d\epsilon}\mathscr{F}'\bigg|_{\epsilon=0}=i\left(\psi_k\frac{\partial\mathscr{F}}{\partial\psi_k}-\psi_k^*\frac{\partial\mathscr{F}}{\partial\psi_k^*}\right)+i\left(\partial_i\psi_k\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k)}-\partial_i\psi_k^*\frac{\partial\mathscr{F}}{\partial(\partial_i\psi_k^*)}\right)+\cdots=\boldsymbol{\nabla}\cdot\mathscr{C},$$

which when inserted into (3.5) concludes the proof by showing that

$$\{\mathrm{N},\mathrm{F}\}=\int\boldsymbol{\nabla}\cdot\mathscr{C}\,d^nx=0.\qquad\square$$

A result, structurally the same as the above, is obtained in Weinberg[3] but under a much more restrictive condition—there F is required to be homogeneous of degree one in its arguments (instead of gauge type I invariant).

The *stationary solutions* of the CHE equations (2.9) are, just as in QM, those whose components absolute values $|\psi_k(x)|$ are time independent. Hence, a solution is stationary when it can be written as $\boldsymbol{\psi}(t,x)=\boldsymbol{\varphi}(x)\exp(-i\omega t)$, where ω is a real constant. It is simple to verify that this is the case if the Hamiltonian functional in (2.9) is gauge type I (3.4) invariant. Then, the components $\varphi_k=\varphi_k(x)$ of $\boldsymbol{\varphi}(x)$ are solutions of

$$\frac{\delta}{\delta\varphi_k^*}\,\mathrm{H}[\boldsymbol{\varphi}^*,\boldsymbol{\varphi}]=\omega\varphi_k, \quad k=1,2,...,r, \tag{3.6}$$

where $\mathrm{H}[\boldsymbol{\varphi}^*,\boldsymbol{\varphi}]=\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi}]$ by the invariance assumption. Again, there is an equivalent result in Weinberg[3] which, however, is derived from the homogeneity assumption.

The form (3.6) is structurally equivalent to that of the *time-independent* Schrödinger and Dirac equations, although it is nonlinear in general.

The following corollary, to Theorem 3.1, extends a fundamental property of the QM equations to the gauge type I invariant nonlinear CHE equations:

Corollary 3.2: The Hermitian norm $\mathrm{N}=\langle\boldsymbol{\psi},\boldsymbol{\psi}\rangle$ *of any space-localized solution of a nonlinear CHE equation is constant in time if the equation is gauge type I invariant.*

Proof: A CHE equation is gauge type I invariant if its associated Hamiltonian functional H is gauge type I invariant. Hence, by Theorem 3.1 we have $\{\mathrm{N},\mathrm{H}\}=0$. Then from the relation (3.3) the corollary follows. □

IV. COMPLEX POISSON BRACKETS IN HILBERT SPACE

In quantum field theory the Fourier expansion of the fields plays, of course, a key role. This fact together with the previous results raises the question: ''If we expand a space-localized solution of a complex Hamiltonian evolution equation in terms of *a Hilbert space basis* defined on $\mathbb{R}^n$ then what system of equations do the expansion coefficients satisfy?'' In its answer and those of several other associated questions are encountered some new mathematical relations.

Let $\eta_\mu=\eta_\mu(x)$ be a set of scalar complex-valued functions on $\mathbb{R}^n$ which define an orthonormal basis in the Hilbert space $\mathbb{H}(\mathbb{R}^n)$, that is

$$\langle\eta_\mu,\eta_\nu\rangle=\int_{\mathbb{R}^n}\eta_\mu^*(x)\,\eta_\nu(x)\,d^nx=\delta_{\mu\nu}, \quad \mu,\nu=1,2,...,\infty, \tag{4.1}$$

where $x\in\mathbb{R}^n$. (Lower case Greek indices will be used for numbering the elements of all *infinite* sets and their subsets.) Any space-localized $\boldsymbol{\psi}$ field belongs to $\mathbb{H}(\mathbb{R}^n)$ by Definition 2.1. Hence, its components ψ_k may be expressed as linear combinations of the basis functions:

$$\psi_k(x,t)=\sum_{\mu=1}^{\infty}\mathrm{A}_{k\mu}(t)\,\eta_\mu(x). \tag{4.2}$$

The expansion coefficients $\mathrm{A}_{k\mu}$ are given by

$$\mathrm{A}_{k\mu}(t)=\int_{\mathbb{R}^n}\psi_k(x,t)\,\eta_\mu^*(x)\,d^nx. \tag{4.3}$$

Clearly, (4.3) and its complex conjugate $\mathrm{A}_{k\mu}^*(t)$ are linear functionals of $\psi_k(x,t)$ and $\psi_k^*(x,t)$. The following theorem shows that they provide an equivalent description of the time evolution of a space-localized field

Theorem 4.1: *Let* $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_r)$ *be a space-localized solution of a complex Hamiltonian evolution equation with the form (2.9). Then its expansion coefficients* $\mathrm{A}_{k\mu}$ *and* $\mathrm{A}_{k\mu}^*$, *given by (4.3), satisfy the infinite-dimensional Hamiltonian system of equations*

$$\frac{d\mathrm{A}_{k\mu}}{dt}=-i\,\frac{\partial}{\partial A_{k\mu}^*}\,\mathrm{H}(\mathrm{A}^*,\mathrm{A},t) \tag{4.4}$$

with $k=1,2,...,r$ and $\mu=1,2,...,\infty$. $H=H(A^*,A,t)$ *is the Hamiltonian function of the system obtained by inserting (4.2) and its complex conjugate into the functional* $H[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$ *and performing the integration. The converse is true with any set of* $A_{k\mu}$ *provided the series* $\Sigma_\mu A_{k\mu}^* A_{k\mu}$ *converges for all* $k=1,2,...,r$.

Proof: Since $H(A^*,A,t)$ is a function of functionals its variational derivatives $\delta H/\delta\psi_j$ and $\delta H/\delta\psi_j^*$ are well defined. From the definition of the variational derivative and with the use of (4.2) we get

$$\frac{\partial H}{\partial A_{k\mu}^*}=\int\left(\frac{\delta H}{\delta\psi_j}\frac{\partial\psi_j}{\partial A_{k\mu}^*}+\frac{\delta H}{\delta\psi_j^*}\frac{\partial\psi_j^*}{\partial A_{k\mu}^*}\right)d^nx=\int\eta_\mu^*\delta_{jk}\frac{\delta H}{\delta\psi_j^*}\,d^nx.$$

Then, applying (2.9) and taking the time-derivative of (4.2) gives the desired result:

$$\frac{\partial H}{\partial A_{k\mu}^*}=i\int\eta_\mu^*\frac{\partial\psi_k}{\partial t}\,d^nx=i\,\frac{dA_{k\nu}}{dt}\int\eta_\mu^*\eta_\nu\,d^nx=i\,\frac{dA_{k\mu}}{dt}.$$

To prove the converse observe that the convergence of the series $\Sigma_\mu A_{k\mu}^* A_{k\mu}$ ensures that the field $\boldsymbol{\psi}$, given by (4.2), is space-localized. Then, with the use of (4.4) we have

$$\frac{\partial\psi_k}{\partial t}=\frac{dA_{k\mu}}{dt}\,\eta_\mu=-i\,\frac{\partial H(A^*,A,t)}{\partial A_{k\mu}^*}\,\eta_\mu=-i\,\frac{\partial H(A^*,A,t)}{\partial A_{k\mu}^*}\frac{\delta A_{k\mu}^*}{\delta\psi_k^*}=-i\,\frac{\delta}{\delta\psi_k^*}H[\boldsymbol{\psi}^*,\boldsymbol{\psi};t].$$

□

A corollary of Theorem 4.1, which does not require any further proofs, is: *All unitary transformations which perform change of basis in the Hilbert space* $\mathbb{H}(\mathbb{R}^n)$ *are canonical transformations since they preserve the form (4.4) of the Hamiltonian systems.*

Theorem 4.1 implies that the expansion coefficients $A_{k\mu}$ and $A_{k\nu}^*$ constitute an infinite set of complex canonical variables. To show that this is so we need to distinguish between the continuous complex Poisson bracket $\{\cdot,\cdot\}^c$ (3.1) and the *discrete* complex Poisson bracket $\{\cdot,\cdot\}^d$ defined by (2.6) in which for the summation index, now, we use $\alpha=1,2,...,\infty$. $R=R(z^*,z,t)$ and $S=S(z^*,z,t)$ are real or complex-valued functions of z_α, z_α^* and possibly t. With these definitions the next result is stated as:

Theorem 4.2: *The continuous Poisson bracket (3.1) is identically equal to the discrete Poisson bracket (2.6) when* $z_\alpha=A_{k\mu}$ *are the expansion coefficients of the space-localized* $\boldsymbol{\psi}$ *field in any Hilbert space basis.*

Proof: By inserting (4.2) into a functional $F=F[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]$ and performing the required integration one obtains its value $F=F(A^*,A,t)$ as a function of the functionals $A_{k\mu}[\boldsymbol{\psi}]$ and $A_{k\mu}^*[\boldsymbol{\psi}^*]$. Then, the identity

$$\frac{\delta F[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]}{\delta\psi_k}=\frac{\partial F(A^*,A,t)}{\partial A_{j\mu}}\frac{\delta A_{j\mu}[\boldsymbol{\psi}]}{\delta\psi_k}=\frac{\partial F(A^*,A,t)}{\partial A_{k\mu}}\,\eta_\mu^* \tag{4.5}$$

is the result of taking the variational derivative of a function of functionals. An alternative derivation of the same identity is

$$\begin{aligned}\frac{\partial F(A^*,A,t)}{\partial A_{k\mu}}\,\eta_\mu^*(x)&=\int\frac{\delta F[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]}{\delta\psi_k(x')}\frac{\partial\psi_k(x')}{\partial A_{k\mu}}\,\eta_\mu^*(x)\,d^nx'\\&=\int\frac{\delta F[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]}{\delta\psi_k(x')}\,\delta(x'-x)\,d^nx'=\frac{\delta F[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]}{\delta\psi_k(x)},\end{aligned}$$

where use was made of the *closure* relation $\eta_\mu^*(x)\,\eta_\mu(x')=\delta(x'-x)$. With the identity (4.5) the continuous Poisson bracket is transformed into the discrete one as follows:

$$\{\mathrm{R},\mathrm{S}\}^c=\frac{1}{i}\int\left(\frac{\delta \mathrm{R}}{\delta\psi_k}\frac{\delta \mathrm{S}}{\delta\psi_k^*}-\frac{\delta \mathrm{S}}{\delta\psi_k}\frac{\delta \mathrm{R}}{\delta\psi_k^*}\right)d^nx$$
$$=\frac{1}{i}\left(\frac{\partial \mathrm{R}}{\partial \mathrm{A}_{k\mu}}\frac{\partial \mathrm{S}}{\partial \mathrm{A}_{k\nu}^*}-\frac{\partial \mathrm{S}}{\partial \mathrm{A}_{k\mu}}\frac{\partial \mathrm{R}}{\partial \mathrm{A}_{k\nu}^*}\right)\int \eta_\mu^*\eta_\nu\, d^nx$$
$$=\frac{1}{i}\left(\frac{\partial \mathrm{R}}{\partial \mathrm{A}_{k\mu}}\frac{\partial \mathrm{S}}{\partial \mathrm{A}_{k\mu}^*}-\frac{\partial \mathrm{S}}{\partial \mathrm{A}_{k\mu}}\frac{\partial \mathrm{R}}{\partial \mathrm{A}_{k\mu}^*}\right)=\{\mathrm{R},\mathrm{S}\}^d. \qquad \square$$

Now, if the discrete bracket form (4.5) is used the canonical relations

$$\{\mathrm{A}_{j\mu}^*,\mathrm{A}_{k\nu}\}=i\,\delta_{jk}\delta_{\mu\nu},\quad \{\mathrm{A}_{j\mu}^*,\mathrm{A}_{k\nu}\}=0,\quad \{\mathrm{A}_{j\mu},\mathrm{A}_{k\nu}\}=0 \tag{4.6}$$

follow trivially. To see that they hold also with the continuous bracket forms (without relying on Theorem 4.2) note that $\mathrm{A}_{k\mu}$ and $\mathrm{A}_{k\mu}^*$ are functionals, and hence

$$\{\mathrm{A}_{j\mu}^*,\mathrm{A}_{k\nu}\}=\frac{1}{i}\int\left(\frac{\delta \mathrm{A}_{j\mu}^*}{\delta\psi_i}\frac{\delta \mathrm{A}_{k\nu}}{\delta\psi_i^*}-\frac{\delta \mathrm{A}_{k\nu}}{\delta\psi_i}\frac{\delta \mathrm{A}_{j\mu}^*}{\delta\psi_i^*}\right)d^nx=i\,\delta_{ij}\delta_{ik}\int \eta_\nu^*\eta_\mu\, d^nx=i\,\delta_{jk}\delta_{\mu\nu}.$$

Similarly, one finds that $\{\mathrm{A}_{j\mu},\mathrm{A}_{k\nu}\}=\{\mathrm{A}_{j\mu}^*,\mathrm{A}_{k\nu}^*\}=0$.

With the expansion coefficients $\mathrm{A}_{k\mu}$ one can define real-valued functionals as follows:

$$\mathrm{X}_{k\mu}=\frac{1}{\sqrt{2}}\,(\mathrm{A}_{k\mu}^*+\mathrm{A}_{k\mu})=\frac{1}{\sqrt{2}}\int(\psi_k^*\eta_\mu+\psi_k\eta_\mu^*)\,d^nx, \tag{4.7}$$

$$\mathrm{Y}_{k\mu}=\frac{i}{\sqrt{2}}\,(\mathrm{A}_{k\mu}^*-\mathrm{A}_{k\mu})=\frac{i}{\sqrt{2}}\int(\psi_k^*\eta_\mu-\psi_k\eta_\mu^*)\,d^nx. \tag{4.8}$$

Proposition 4.3: The values $\mathrm{X}_{k\mu}(t)$ *and* $\mathrm{Y}_{k\mu}(t)$ *of the functionals (4.7) and (4.8) are real canonical variables which satisfy a classical Hamiltonian system if the* $\boldsymbol{\psi}$ *field is a solution of a complex Hamiltonian evolution equation of the form (2.9).*

Proof: Use the Poisson bracket relations (4.6) to see that

$$\{\mathrm{X}_{j\mu},\mathrm{Y}_{k\nu}\}=i\{\mathrm{A}_{j\mu}^*,\mathrm{A}_{k\nu}\}=\delta_{jk}\delta_{\mu\nu},\quad \{\mathrm{X}_{j\mu},\mathrm{X}_{k\nu}\}=0,\quad \{\mathrm{Y}_{j\mu},\mathrm{Y}_{k\nu}\}=0,$$

which shows that $\mathrm{X}_{k\mu}(t)$ and $\mathrm{Y}_{k\mu}(t)$ are canonical variables. To show that they are solutions of a Hamiltonian system of equations take the time derivative of $\mathrm{X}_{k\mu}(t)$ and use (2.9) with $\mathrm{H}=\mathrm{H}[\boldsymbol{\psi}^*,\boldsymbol{\psi};t]=\mathrm{H}(\mathrm{A}^*,\mathrm{A},t)$. Then

$$\frac{d\mathrm{X}_{k\mu}}{dt}=\frac{1}{\sqrt{2}}\left(\frac{d\mathrm{A}_{k\mu}}{dt}+\frac{d\mathrm{A}_{k\mu}^*}{dt}\right)=\frac{i}{\sqrt{2}}\left(\frac{\partial \mathrm{H}}{\partial A_{k\mu}}-\frac{\partial \mathrm{H}}{\partial \mathrm{A}_{k\mu}^*}\right)=\frac{\partial \mathrm{H}(\mathrm{X},\mathrm{Y},t)}{\partial \mathrm{Y}_{k\mu}}.$$

The Hamiltonian $\mathrm{H}(\mathrm{X},\mathrm{Y},t)$ is obtained by inserting (4.7) and (4.8) into $\mathrm{H}(\mathrm{A}^*,\mathrm{A},t)$. A similar derivation produces

$$\frac{d\mathrm{Y}_{k\mu}}{dt}=-\frac{\partial \mathrm{H}(\mathrm{X},\mathrm{Y},t)}{\partial \mathrm{X}_{k\mu}}. \qquad \square$$

Proposition 4.3 shows that there is at least one class of real-valued functionals which contains infinitely many sets of *canonically conjugate functionals* and that any well behaved functional can be expressed as their function. But more important, one observes that according to Theorem 4.2 the functionals (4.7) and (4.8) define a transformation from the *continuous set* of variables $\boldsymbol{\psi}(x,t)$ to the *discrete set* of variables $\mathrm{X}_{k\mu}(t)$ and $\mathrm{Y}_{k\mu}(t)$ which preserves the value of the Poisson bracket, while changing only its type. Its inverse, of course, is given by

$$\psi_k=\frac{1}{\sqrt{2}}\sum_{\alpha=1}^{\infty}(X_{k\alpha}+iY_{k\alpha})\eta_\alpha. \tag{4.9}$$

Thus, one arrives at the following natural extension of the well known classical definition:

Definition 4.1: An invertible transformation between a continuous and a discrete sets of variables will be called infinite-dimensional canonical transformation if it preserves the value of the Poisson bracket (while changing its type) of any two admissible functionals.

Once the above link, between the expansions of functions in orthonormal basis and the canonical transformations, is established we become aware that while the classical canonical transformations are nonlinear, in general, the infinite-dimensional canonical transformations whose form is given by (4.7) and (4.8) are linear. Hence, we are led to inquire: Are there infinite-dimensional canonical transformations which are nonlinear?

The remaining portion of this section presents the answer and some of its implications. Before we state the next result we note that a set of functionals is called *independent* if no one of its members can be expressed as a function of the rest of its members.

Theorem 4.4: *Let the countably infinite set of real-valued functionals*

$$\mathrm{Q}_\mu=\mathrm{Q}_\mu[\boldsymbol{\psi}^*,\boldsymbol{\psi}] \quad \text{and} \quad \mathrm{P}_\mu=\mathrm{P}_\mu[\boldsymbol{\psi}^*,\boldsymbol{\psi}] \tag{4.10}$$

(not necessarily linear or bilinear) be independent. If the relations

$$\{\mathrm{Q}_\mu,\mathrm{P}_\nu\}^c=\delta_{\mu\nu}, \quad \{\mathrm{Q}_\mu,\mathrm{Q}_\nu\}^c=0, \quad \{\mathrm{P}_\mu,\mathrm{P}_\nu\}^c=0, \quad \mu,\nu=1,2,...,\infty \tag{4.11}$$

hold when evaluated with any space-localized function $\boldsymbol{\psi}$ then, the transformation (4.10), from the continuous set of variables $\boldsymbol{\psi}(x,t)$ to the discrete set of variables Q_μ and P_μ, is an infinite-dimensional canonical transformation.

Proof: First, we will show that the transformation (4.10) is invertible by expanding the components ψ_k of the localized field $\boldsymbol{\psi}(x,t)$ in an orthonormal Hilbert space basis $\eta_\alpha(x)$, according to (4.2), with the complex expansion coefficients $\mathrm{A}_{k\alpha}$ given by (4.3). When the series (4.2) and its complex conjugate are inserted into the functionals Q_μ, P_μ and the integrations are performed the resulting expressions give the functionals values

$$\mathrm{Q}_\mu=\mathrm{Q}_\mu(\mathrm{A}^*,\mathrm{A}), \quad \mathrm{P}_\mu=\mathrm{P}_\mu(\mathrm{A}^*,\mathrm{A}) \tag{4.12}$$

as functions of the variables $\mathrm{A}_{k\alpha}$ and $\mathrm{A}^*_{k\alpha}$.

Consider, now, the finite-dimensional case when $\alpha=1,2,...,\rho$ and denote by ψ_k^ρ the field obtained by truncating the expansion series for ψ_k (4.2) to ρ terms

$$\psi_k^\rho=\sum_{\alpha=1}^{\rho}\mathrm{A}_{k\alpha}\eta_\alpha. \tag{4.13}$$

By inserting (4.13) into the set of independent functionals (4.10) one finds that the last splits into infinitely many subsets each containing $2r\rho$ independent functions

$$\mathrm{Q}_\mu^\rho=\mathrm{Q}_\mu^\rho(\mathrm{A}^{\rho *},\mathrm{A}^\rho) \quad \text{and} \quad \mathrm{P}_\mu^\rho=\mathrm{P}_\mu^\rho(\mathrm{A}^{\rho *},\mathrm{A}^\rho), \quad \mu=1,2,...,r\rho,$$

where the collection of $r\rho$ expansion coefficients $\mathrm{A}_{11},...,\mathrm{A}_{r\rho}$ is denoted by A^ρ and that of their complex conjugates—by $\mathrm{A}^{\rho *}$. Consequently, each subset determines a solution which expresses the coefficients $\mathrm{A}_{k\alpha}(\mathrm{Q}^\rho,\mathrm{P}^\rho)$ as functions of Q_μ^ρ and P_μ^ρ with $\mu=1,2,...,r\rho$. For a finite ρ these solutions are distinct, in general.

However, we are concerned only with the limits when $\rho\to\infty$. And these limits must all consist of the same set of functions

$$\mathrm{A}_{k\alpha}=\mathrm{A}_{k\alpha}(Q,P), \quad \alpha=1,2,...,\infty, \tag{4.14}$$

since otherwise the original set of functionals $\mathrm{Q}_\mu^\rho\to\mathrm{Q}_\mu$ and $\mathrm{P}_\mu^\rho\to\mathrm{P}_\mu$ which obtains as a result of $\rho\to\infty$ will not be independent, contradicting the theorem's assumption.

The *uniqueness* of the set of functions (4.14) as a solution of the system (4.12) will be demonstrated by an alternative argument, as follows: Consider the identities

$$\{Q_\mu, A_{k\alpha}\}^c = \frac{\partial A_{k\alpha}}{\partial Q_\nu}\{Q_\mu, Q_\nu\}^c + \frac{\partial A_{k\alpha}}{\partial P_\nu}\{Q_\mu, P_\nu\}^c = \frac{\partial A_{k\alpha}(Q,P)}{\partial P_\mu}, \tag{4.15}$$

$$\{P_\mu, A_{k\alpha}\}^c = \frac{\partial A_{k\alpha}}{\partial Q_\nu}\{P_\mu, Q_\nu\}^c + \frac{\partial A_{k\alpha}}{\partial P_\nu}\{P_\mu, P_\nu\}^c = -\frac{\partial A_{k\alpha}(Q,P)}{\partial Q_\mu}, \tag{4.16}$$

which are derived by regarding $A_{k\alpha}(Q,P)$ as functions of independent functionals and applying the canonical relations (4.11). For a given space-localized $\boldsymbol{\psi}$ field the values of the functionals $\{Q_\mu, A_{k\alpha}\}^c$, $\{P_\mu, A_{k\alpha}\}^c$, Q_μ and P_μ are well defined and unique. Then, the identities (4.15) and (4.16) show that all partial derivatives of all functions $A_{k\alpha}(Q,P)$ are well defined and unique which, in turn, shows that the functions themselves are unique.

The inverse of the transformation (4.10) is determined by inserting the so obtained functions $A_{k\alpha}(Q,P)$ into (4.2) which becomes

$$\psi_k = \sum_{\mu=1}^{\infty} A_{k\mu}(Q,P)\,\eta_\mu(x) = \psi_k(Q,P,x). \tag{4.17}$$

This series is convergent since the series (4.2) is convergent and the expansion coefficients remain the same in value, while they are being expressed as functions of Q_μ and P_μ.

In (4.17) we have the most general form of the infinite-dimensional canonical transformations from a discrete set of real variables Q_μ, P_μ to a continuous set of complex variables $\boldsymbol{\psi}$. Accordingly, when the field is given with this form its time-dependence is determined only by the time-dependence of $Q_\mu(t)$ and $P_\mu(t)$. If complex discrete variables Z_μ are used in place of the real ones the form of the inverse transformation will be $\psi_k = \psi_k(Z^*,Z,x)$ according to the relevant relations in Sec. II.

It remains to show that the transformations (4.10) preserve the Poisson bracket of any two admissible functionals R and S. For this purpose we substitute their arguments with the inverse transformation functions $\psi_k = \psi_k(Q,P,x)$ and perform the integrations. The results are functions which give the functionals values

$$R = R[\boldsymbol{\psi}^*, \boldsymbol{\psi}] = R(Q,P) \quad \text{and} \quad S = S[\boldsymbol{\psi}^*, \boldsymbol{\psi}] = S(Q,P)$$

in terms of the discrete variables. Now, both the continuous Poisson bracket (3.1) and the discrete Poisson bracket

$$\{R,S\}^d = \frac{\partial R}{\partial Q_\mu}\frac{\partial S}{\partial P_\mu} - \frac{\partial S}{\partial Q_\mu}\frac{\partial R}{\partial P_\mu} \tag{4.18}$$

are well defined. Then, as a consequence of (4.11) we have

$$\begin{aligned}
\{R,S\}^c &= \frac{1}{i}\int\left(\frac{\delta R}{\delta\psi_k}\frac{\delta S}{\delta\psi_k^*} - \frac{\delta R}{\delta\psi_k}\frac{\delta S}{\delta\psi_k^*}\right)d^n x \\
&= \frac{1}{i}\int\left(\frac{\partial R}{\partial Q_\mu}\frac{\delta Q_\mu}{\delta\psi_k} + \frac{\partial R}{\partial P_\mu}\frac{\delta P_\mu}{\delta\psi_k}\right)\left(\frac{\partial S}{\partial Q_\nu}\frac{\delta Q_\nu}{\delta\psi_k^*} + \frac{\partial S}{\partial P_\nu}\frac{\delta P_\nu}{\delta\psi_k^*}\right)d^n x \\
&\quad - \frac{1}{i}\int\left(\frac{\partial S}{\partial Q_\mu}\frac{\delta Q_\mu}{\delta\psi_k} + \frac{\partial S}{\partial P_\mu}\frac{\delta P_\mu}{\delta\psi_k}\right)\left(\frac{\partial R}{\partial Q_\nu}\frac{\delta Q_\nu}{\delta\psi_k^*} + \frac{\partial R}{\partial P_\nu}\frac{\delta P_\nu}{\delta\psi_k^*}\right)d^n x \\
&= \frac{\partial R}{\partial Q_\mu}\frac{\partial S}{\partial Q_\nu}\{Q_\mu,Q_\nu\}^c + \frac{\partial R}{\partial Q_\mu}\frac{\partial S}{\partial P_\nu}\{Q_\mu,P_\nu\}^c + \frac{\partial R}{\partial P_\mu}\frac{\partial S}{\partial P_\nu}\{P_\mu,P_\nu\}^c + \frac{\partial R}{\partial P_\mu}\frac{\partial S}{\partial Q_\nu}\{P_\mu,Q_\nu\}^c \\
&= \frac{\partial R}{\partial Q_\mu}\frac{\partial S}{\partial P_\mu} - \frac{\partial S}{\partial Q_\mu}\frac{\partial R}{\partial P_\mu} = \{R,S\}^d.
\end{aligned}$$

This establishes that $\{R,S\}^c=\{R,S\}^d$ is an identity, since it holds for all space-localized ψ fields, completing the proof. □

In the finite-dimensional case a canonical transformation of $2s$ variables preserves the form of any Hamiltonian system with the same number of variables. In a complete correspondence, we have as a consequence from the above theorem the following:

Corollary 4.5: An infinite-dimensional canonical transformation of the general type (4.10) converts any Hamiltonian evolution equation of the form (2.9) for the ψ field into a Hamiltonian system of infinitely many equations for the variables Q_μ, P_μ. *The converse holds for the inverse transformation (4.17).*

Proof: Express the Hamiltonian functional $H[\psi^*,\psi;t]=H(Q,P,t)$ of the evolution equation as a function of Q_μ and P_μ using the inverse transformation functions $\psi=\psi(Q,P,x)$ (which exist according to Theorem 4.4). Then, the desired result follows from (3.3) for the time-derivative of a functional and from the identity of Theorem 4.4:

$$\frac{dQ_\mu}{dt}=\{Q_\mu,H\}^c=\{Q_\mu,H\}^d=\frac{\partial H}{\partial P_\mu},$$

$$\frac{dP_\mu}{dt}=\{P_\mu,H\}^c=\{P_\mu,H\}^d=-\frac{\partial H}{\partial Q_\mu}.$$

To establish the converse, represent $\psi_k(x)$ with the functional $\int\psi_k(x')\,\delta(x'-x)d^nx'$ and denote $\psi_k'=\psi(x')$. Then, we have

$$\frac{\partial\psi_k}{\partial t}=\frac{\partial\psi_k}{\partial Q_\mu}\frac{dQ_\mu}{dt}+\frac{\partial\psi_k}{\partial P_\mu}\frac{dP_\mu}{dt}=\{\psi_k,H\}^d.$$

Applying the same identity completes the proof by showing that

$$\frac{\partial\psi_k}{\partial t}=\{\psi_k,H\}^c=\frac{1}{i}\int\frac{\delta H}{\delta\psi_j'^*}\,\delta_{jk}\,\delta(x'-x)\,d^nx'=-i\,\frac{\delta H}{\delta\psi_k^*}.$$ □

V. ON THE QUANTIZATION RULES

This section is an application of the previous results.

By very general physics arguments, one expects that the characteristics of a localized ψ field *as a whole*, like its position, velocity, linear, and angular momentum, etc., are to be given with real-valued bilinear functionals even when the ψ field is assumed to obey some nonlinear evolution equation. Or, to use the QM term, the *observables* are to remain linear Hermitian operators and their values are to be given with bilinear functionals (except for the Hamiltonian) even if the wave function was to obey a nonlinear equation. Assuming that this is true, then these functionals must be *normalized bifunctionals*, that is, divided by N from (2.8). For, the function ψ cannot be normalized, as done in quantum mechanics, since by the above assumption it is a solution of a nonlinear evolution equation.

In place of ''normalized bilinear functional'' one could use, of course, the QM term ''expectation value'' of the associated linear operator. But no probabilistic interpretation should be attached to it since, here, such an interpretation would be entirely unnecessary, and the controversies accompanying it may distract us from the subject at hand.

After these preliminaries we pose the question: ''What can be said for the Poisson bracket of the expectation values of two linear operators if they are evaluated with a space-localized solution of a *nonlinear* field equation?'' There is at least one statement which, as we will see, is surprisingly similar to the quantum mechanics' answer in the linear case, namely:

Theorem 5.1: *The continuous Poisson bracket of two normalized bilinear real or complex-valued functionals* R *and* S *which are associated with the linear matrix-differential operators* **R** *and* **S** *according to the ''rule''*

$$\mathrm{R}=\frac{1}{\mathrm{N}}\int \psi_j^*\mathbf{R}_{jk}\psi_k\,d^nx=\frac{1}{\mathrm{N}}\int \boldsymbol{\psi}^\dagger\mathbf{R}\boldsymbol{\psi}\,d^nx \quad \text{and}$$

$$\mathrm{S}=\frac{1}{\mathrm{N}}\int \psi_j^*\mathbf{S}_{jk}\,\psi_k\,d^nx=\frac{1}{\mathrm{N}}\int \boldsymbol{\psi}^\dagger\mathbf{S}\boldsymbol{\psi}\,d^nx$$

is equal to the normalized bilinear functional which is associated, precisely by the same ''rule,'' with the linear matrix-differential operator

$$[\mathbf{R},\mathbf{S}]_{\mathrm{QM}}=\frac{1}{i\mathrm{N}}\,(\mathbf{RS}-\mathbf{SR}). \tag{5.1}$$

($\mathrm{N}=\langle\boldsymbol{\psi},\boldsymbol{\psi}\rangle$ and $\boldsymbol{\psi}^\dagger=(\boldsymbol{\psi}^*)^t$ denotes the transpose of $\boldsymbol{\psi}^*$.)

Proof: First, we observe that any bilinear functional is gauge-type I invariant. Then, from Theorem 3.1 it follows that $\{\mathrm{N},\mathrm{R}\}=\{\mathrm{N},\mathrm{S}\}=0$ since for any functional F, not necessarily bilinear, which is gauge type I invariant we have

$$\left\{\mathrm{N},\frac{\mathrm{F}}{\mathrm{N}}\right\}=\frac{1}{i}\int\left(\frac{\delta\mathrm{N}}{\delta\psi_k}\frac{\delta}{\delta\psi_k^*}\left(\frac{\mathrm{F}}{\mathrm{N}}\right)-\frac{\delta}{\delta\psi_k}\left(\frac{\mathrm{F}}{\mathrm{N}}\right)\frac{\delta\mathrm{N}}{\delta\psi_k^*}\right)d^nx=\frac{1}{\mathrm{N}}\{\mathrm{N},\mathrm{F}\}-\frac{\mathrm{F}}{\mathrm{N}^2}\{\mathrm{N},\mathrm{N}\}=0,$$

where $\{\mathrm{N},\mathrm{F}\}=0$ according to Theorem 3.1. Hence, it follows that

$$\begin{aligned}\{\mathrm{R},\mathrm{S}\}&=\frac{1}{i}\int\left(\frac{\delta\mathrm{R}}{\delta\psi_k}\frac{\delta\mathrm{S}}{\delta\psi_k^*}-\frac{\delta\mathrm{S}}{\delta\psi_k}\frac{\delta\mathrm{R}}{\delta\psi_k^*}\right)d^nx\\&=\frac{1}{i\mathrm{N}^2}\int\psi_i^*(\mathbf{R}_{ik}\mathbf{S}_{kj}-\mathbf{S}_{ik}\mathbf{R}_{kj})\psi_j\,d^nx\\&\quad+\frac{\{\mathrm{N},\mathrm{R}\}}{\mathrm{N}^2}\int\boldsymbol{\psi}^\dagger\mathbf{S}\boldsymbol{\psi}\,d^nx-\frac{\{\mathrm{N},\mathrm{S}\}}{\mathrm{N}^2}\int\boldsymbol{\psi}^\dagger\mathbf{R}\boldsymbol{\psi}\,d^nx=\frac{1}{\mathrm{N}}\int\boldsymbol{\psi}^\dagger[\mathbf{R},\mathbf{S}]_{\mathrm{QM}}\,\boldsymbol{\psi}\,d^nx\end{aligned}\qquad\square$$

which completes the proof.

The above result deserves a discussion, and this is what follows:

Theorem 5.1 holds regardless of what the nonlinear equation for $\boldsymbol{\psi}$ is. Moreover, N is not required to be constant in time, as long as it remains finite and nonzero.

We also see that when the operators **R** and **S** are Hermitian the operator $[\mathbf{R},\mathbf{S}]_{\mathrm{QM}}$ in (5.1) not only plays the role of, but also has the form of the *quantum commutator* or the ''quantum Poisson bracket,'' as P. M. Dirac[18] called it. The single, and remarkable, difference is that the Hermitian norm N appears in (5.1) in place of Planck's constant $\hbar$.

But first, we must make sure that the *dimensionality* of N is right. Certainly, N is not dimensionless, as it is in quantum mechanics, since $\boldsymbol{\psi}$ is not normalized. From the identity (2.14) it follows that the dimensionality of any Hamiltonian functional for a CHE equation is always [*energy*]. Then, the form (2.9) together with expression (2.10) tell us that the dimensionality of all the products $\psi_k^*\psi_k$ is [*action*] per unit volume. Referring to the definition (2.8) we conclude that the dimensionality of the Hermitian norm N is always [*action*] which, of course, is the dimensionality of Planck's constant. The only condition for the validity of the above is that $\boldsymbol{\psi}$ must be a solution of a CHE equation.

The same conclusion is reached by recalling, from the beginning of Sec. II, that the dimensionality of $\Sigma_{i=1}^{s}\,z_i^*z_i$ is always [*action*] if z_i are solutions of a complex Hamiltonian system. Obviously, the transition $s\to\infty$ cannot change the dimensionality of the sum, whose limit is precisely N.

There is an entirely different argument discussed in Bodurov[11] which shows that if $\boldsymbol{\psi}$ is a solution of a *Galilei-invariant* nonlinear CHE equation then, again, N plays the role of Planck's constant when $\boldsymbol{\psi}$ is Galilei transformed.

Next, let us assume that we have found a set of Hermitian operators $\mathbf{Q}_\mu$ and $\mathbf{P}_\mu$ for whose commutators the following conditions hold:

$$[\mathbf{Q}_\mu,\mathbf{P}_\nu]=i\mathrm{N}\delta_{\mu\nu},\quad [\mathbf{Q}_\mu,\mathbf{Q}_\nu]=0,\quad [\mathbf{P}_\mu,\mathbf{P}_\nu]=0 \tag{5.2}$$

and whose expectation values Q_μ and P_μ are defined by the ''rule'' of Theorem 5.1. Then, by the same theorem their continuous Poisson brackets are

$$\{\mathrm{Q}_\mu,\mathrm{P}_\nu\}=\delta_{\mu\nu},\quad \{\mathrm{Q}_\mu,\mathrm{Q}_\nu\}=0,\quad \{\mathrm{P}_\mu,\mathrm{P}_\nu\}=0. \tag{5.3}$$

The last relations allow us to apply Theorem 4.4 and Corollary 4.5 which, in turn, tell us *that the time development of the expectation values* Q_μ *and* P_μ *is governed by a classical Hamiltonian system of equations*.

The preceeding sequence of arguments demonstrates that the form of the *fundamental quantization rules* which where postulated by P. M. Dirac,[18] and which are equivalent to the commutation relations (5.2) with $\mathrm{N}=\hbar$, are obtainable without any ''external'' assumptions. This claim is made only with respect to the form of the quantization rules because the value of N in (5.1) is left undetermined by the field equation. The investigation of a possible condition with which the value of N can be ''fixed'' will be the subject of a forthcoming paper.

The commutation relations (5.2) and the canonical relations (5.3) are not equivalent: (5.2) obtains uniquely from (5.2), but (5.2) does not follow uniquely from (5.3). The most general commutation relations which follow from (5.3) contain (5.2) as a special case. This observation may be taken as an evidence that there is a nonlinear field theory which contains quantum mechanics as a special case.

We remark that the results of this section allow us to view the *complementarity* between classical variables and the operators of quantum mechanics as a manifestation of infinite-dimensional canonical transformations between discrete and continuous variables.

VI. A CONCLUDING DISCUSSION

The previous sections were intended to be as free of speculative arguments as possible. This section, on the other hand, is free of such restrictions. The history of QM, alone, is evidence that speculative physics arguments may be essential in finding key questions and nonspeculative answers. To those who share this view the following may be of interest.

It is natural to assume, as done in Weinberg[3] and in other publications, that if any nonlinearities exist in the equations of quantum mechanics they must be only small perturbations. Could it be otherwise, without having been known already? But then, all the proposed models which are based on nonlinear self-localized fields would be no more than futile exercises, including the work of W. Heisenberg[2] (and the references there).

These two approaches, the one critically relying on the assumption of *perturbationally small* nonlinearity and the other exploiting the possibilities of self-localized fields, seem to be mutually excluding. Indeed, only particular types of nonlinear field equation may possess self-localized solutions, and certainly none if the degree of their nonlinearity is arbitrarily small. (See, for example, Berestycki and Lions.[13])

Yet, a third alternative is possible, and the following will explain what it is.

But first, we need to define the meaning of the ''*linearized version*'' of a ψ-field equation, which is the equation obtained by letting $|\psi|\to 0$ and retaining only the terms linear in ψ.

Now, consider a self-localized solution $\psi(x,t)$ of a particular nonlinear *free field* CHE equation, and let ρ_1 be an appropriately defined radius of the localization region which, of course, will depend only on the equation parameters. Next, assume that the initially free moving localization region enters a domain in space where it interacts with an external potential $U(x)$ according to the QM prescription:

$$\frac{\partial}{\partial t}\to\frac{\partial}{\partial t}+igU(x) \tag{6.1}$$

where g is a suitably chosen interaction constant. Then, as long as the variation of $|U(x)|$ *within the localization region* is sufficiently small the motion of the ψ field *as a whole* will be identical to the motion of a point-mass in the same potential. This is a rigorous result, and it is derived in Bodurov[19] for a large class of *gauge I invariant nonlinear* field equations, including the CHE.

When the external field $-\nabla U$ reach sufficiently high intensity the initial ψ-field configuration, of course, will become unstable. Consequently, ψ will expand in space (rather than contract to a point), because otherwise either N would vanish, or ψ would become singular. However, for certain potential wells space-localization (that is, bound state) may occur again, but this time, due to the presence of $U(x)$, and not because of nonlinearity.

This can happen if the following three conditions are met: First, the ''linearized version'' of the initial CHE equation possesses a bound state solution with a suitably defined radius ρ_2. Second (this is the only speculative part in this argument), the norm N of the ψ field remains the same, or at least within the same order of magnitude, in both cases of localization. And third, ρ_2 is sufficiently larger than ρ_1.

One way to satisfy the first condition is to assume that the ''linearized version'' is an equation of the Schrödinger or of the Dirac-type. The second condition is easy to justify as long as the ''linearized version'' provides a sufficiently accurate description of the bound state, since then we may chose N to have any real value we like. Then, the ratio of the maximum field magnitudes $|\psi_2|_{\max}$ and $|\psi_1|_{\max}$ in the respective cases is

$$\frac{|\psi_2|_{\max}}{|\psi_1|_{\max}}=\left(\frac{\rho_1}{\rho_2}\right)^{3/2}, \tag{6.2}$$

where ρ_1 and ρ_2 are defined by (for $n=3$)

$$\rho_i^3=\frac{3\mathrm{N}}{4\pi|\psi_i|^2_{\max}} \quad \text{with} \quad i=1,2. \tag{6.3}$$

According to (6.2) the effect of nonlinearity will necessarily become perturbationally small when the ratio ρ_2/ρ_1 is sufficiently large, that is, when the third condition is satisfied. Then, the ψ-field equation will become essentially the same as its ''linearized version.'' Hence, at least when the last is a Schrödinger or a Dirac-type equation, we are assured that, indeed, there are potential wells for which the assumed bound state does exist.

Thus, the assumption that the field equations contain only *small nonlinearity*, and the assumption of *strongly nonlinear* field equations are not necessarily incompatible. On the contrary, we see that the two approaches must be linked together if each of them is to meet the expectations of its proponents.

Let us illustrate these arguments by an example. Any nonlinear field theory ''X'' which represents, say, the electron with a self-localized field must give for the radius ρ_1 a number which is much smaller than the *the classical electron radius*, according to the experimental results. But, as a ''worst case'' assumption let us take $\rho\approx 10^{-13}$ cm. On the other hand, *Bohr radius* provides a good estimate $\rho_2\approx 10^{-8}$ cm of what the radius of localization is when the electron (or more precisely, its wave function) is trapped in a potential well. Even this very conservative estimate gives the ratio $\rho_2/\rho_1\approx 10^5$. Then, (6.2) tells us that the ratio of the maximum magnitude of the externally localized field to that of the self-localized field is $|\psi_2|_{\max}/|\psi_1|_{\max}\approx 3\times 10^{-8}$. Certainly, this is enough to make any (physically meaningful) nonlinearity perturbationally small and to allow a linear treatment of the trapped electron by the same theory ''X.''

This discussion was intended also to illustrate how a ''strongly'' nonlinear field theory may reduce to quantum mechanics, and thus contain it as a special case.

Next, the reasons for limiting n to three space-dimensions (Sec. II) are briefly discussed: Quantum mechanics represents systems consisting of i particles with wave functions which depend on $3i$ space-coordinates. The resulting equations, then, are solved by perturbation methods because there are no known methods for obtaining exact solutions.

In a nonlinear field theory there is a natural alternative, not available to quantum mechanics, which is: If we assume that a region of localization in a ψ field represents a particle, then a system of i particles should be represented by a ψ field with i regions of localization. For, if a nonlinear field equation admits a solution with a single localization region, then it admits solutions with any number of localization regions. In this approach the field equation which describes a system of

particles is the same equation which describes one particle, only the initial conditions differ. If such a unification in the treatment of single and multiparticle systems is accepted as correct, then the *homogeneity assumption* (Sec. III) becomes unnecessary.

As remarked in Sec. II Definition 2.1 requires a few comments: If a space-localized field is to represent a particle then we expect, on the basis of physics considerations, that the field localization should be much ''sharper'' than what is demanded by Definition 2.1, alone. It is very satisfying that this expectation is met (without any additional assumptions or restrictions) by all pertinent results which are known to me. For example: It was shown by Berestycki and Lions[13] that all space-localized solutions of all nonlinear Schrödinger and of all nonlinear Klein–Gordon equations (which admit such solutions) *decay exponentially* as $|x|\to\infty$. The localization of the solutions (''gaussons'') to the logarithmic NLS equation[6] is even sharper, that is, they *decay exponentially* as $|x|^2\to\infty$.

Thus, even for the purposes of representing particles with localized fields there is no need for more restrictive definition. On the other hand, we do want Definition 2.1 to give *as generally as possible* the mathematical meaning of a term which is needed both for the preciseness and for the clarity of the preceeding discussions.

It is hoped that this paper will contribute toward the development of a nonlinear field theory whose compatibility with the principles of quantum mechanics will be ensured not by inventing ''quantization rules for nonlinear fields,'' but by the theory's ability to reduce to conventional quantum mechanics as a special case. Accordingly, this paper's results are to be seen not as arrived-at destinations, but rather as destination pointers.

[1] L. de Broglie, *Nonlinear Wave Mechanics, Causal Interpretation* (Elsevier, Amsterdam, 1960).
[2] W. Heisenberg, ''Quantum Theory of Fields and Elementary Particles,'' Rev. Mod. Phys. **29**, 269–278 (1957).
[3] S. Weinberg, ''Testing Quantum Mechanics,'' Ann. Phys. (N.Y.) **194**, 336–386 (1989).
[4] R. Finkelstein and C. Fronsdal, ''Nonlinear Spinor Fields,'' Phys. Rev. **103**, 1571–1579 (1956).
[5] M. Soler, ''Nonlinear Spinor Fields,'' Phys. Rev. D **1**, 2766–2769 (1970).
[6] I. Bialynicki-Birula and J. Mycielski, ''Nonlinear Wave Mechanics,'' Ann. Phys. (N.Y.) **100**, 62–93 (1976).
[7] R. Friedberg, T. D. Lee, and A. Sirlin, ''Class of Scalar-field Soliton Solutions in Three Space Dimensions,'' Phys. Rev. D **13**, 2739–2761 (1976).
[8] F. Cooperstock and N. Rosen, ''A Nonlinear Gauge-invariant Field Theory of Leptons,'' Int. J. Theor. Phys. **28**, 423–440 (1989).
[9] U. Enz, ''The Nonlinear Extended Oscillator: A Particle Concept Beyond Quantum Mechanics,'' Ann. Fond. Louis Broglie **15**, 19–34 (1990).
[10] G. Reinisch, ''Nonlinear Quantum Mechanics,'' Physica A **206**, 229–252 (1994).
[11] T. Bodurov, ''De Broglie-type Relations from Nonlinear Evolution Equations,'' Int. J. Theor. Phys. **36**, 1771–1785 (1997).
[12] C. Schwartz, ''Nonlinear operators and their propagators. I,'' J. Math. Phys. **38**, 484–500 (1997); ''Nonlinear operators and their propagators. II,'' J. Math. Phys. **38**, 3841–3862 (1997).
[13] H. Berestycki and P. Lions, ''Nonlinear Scalar Field Equations. I. Existence of Ground State, II. Existence of Infinitely Many Solutions,'' Arch. Ration. Mech. Anal. **82**, 313–375 (1983).
[14] T. Bodurov, ''Generalized Ehrenfest Theorem for Nonlinear Schrödinger Equations,'' Int. J. Theor. Phys. **37**, 1299–1306 (1998).
[15] F. Strocchi, ''Complex Coordinates and Quantum Mechanics,'' Rev. Mod. Phys. **38**, 36–40 (1966).
[16] P. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed. (Springer, New York, 1993).
[17] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, Reading, MA, 1980).
[18] P. M. Dirac, *The Principles of Quantum Mechanics*, Fourth ed. (Oxford University Press, New York, 1958).
[19] T. Bodurov, ''Solitary Waves Interacting with an External Field,'' Int. J. Theor. Phys. **35**, 2489–2499 (1996).

Canonical and quasicanonical realizations of the osp($D|d$) superalgebra and the Coulomb problem in superspaces

Michel Daumens and Pierre Minnaert[a)]
Centre de Physique Théorique et de Modélisation, Unité associée au CNRS URA 1537
Université de Bordeaux I, Rue du Solarium, 33175 Gradignan CEDEX, France

(Received 1 April 1998; accepted for publication 27 May 1998)

Canonical and quasi-canonical realizations of the osp($D|d$) superalgebra are presented. In these realizations there is only one independent Casimir operator: The quadratic one C, the eigenvalues of which allow specification of the involved representations. Pauli's method for solving the spectrum of the Coulomb problem through quantization of the Runge–Lenz vector is extended to an arbitrary dimension superspace $E(D|d)$ with D bosonic coordinates and d (even) fermionic coordinates. The symmetry algebra of the Hamiltonian is the superalgebra osp($D+1|d$). It is shown how the energy spectrum is related to the Casimir operator C' of osp($D+1|d$) in its quasicanonical realization. Then the energy levels are given in terms of the eigenvalues of the operator C'. © *1998 American Institute of Physics.* [S0022-2488(98)02210-5]

I. INTRODUCTION

It is well-known that a few months before the event of Schrödinger equation, Pauli[1,2] solved the quantum-mechanical hydrogen atom in the framework of Born, Heisenberg, Jordan matrix mechanics, by a purely algebraic method through quantization of the Runge–Lenz vector. Very soon, it appeared that Schrödinger wave mechanics is much more efficient than Heisenberg matrix mechanics for solving many problems in atomic physics, especially through perturbation theory. And this explains why, with the exception of some work by Fock,[3] Pauli's method was completely forgotten for about 40 years, and was not even mentioned in the classical textbooks on quantum mechanics and atomic physics. However, Pauli's technique was revived in the mid 1960's with the discovery[4] of the so(4,2) dynamical symmetry of the Coulomb problem. Since then, numerous articles[5,6] and books[7,8] have reviewed the original work of Pauli and the more recent developments of the algebraic method. Before proceeding, let us also recall that the spectrum of the Coulomb problem in arbitrary dimension Euclidean space had been obtained a long time ago by solving the Schrödinger equation in D dimensions and that more recently it has been found very fruitful to consider the reciprocal dimension $1/D$ as a perturbation parameter for solving some problems in $D=3$ quantum mechanics.[9] Finally it should also be mentioned that recently, Horwitz[10] has generalized Pauli's method to the relativistic Kepler problem.

In this paper, we show that Pauli's method can be generalized to arbitrary dimension superspaces. More precisely we show that besides the generalized angular momentum that is a constant of motion for the Coulomb problem whatever the dimension of the superspace, one can define another set of constants of motion in the form of a supervector that generalizes the Runge–Lenz vector. In addition, it is shown that, exactly as in the three-dimensional case, the whole set of constants of motion forms an orthosymplectic algebra and that the Hamiltonian of the Coulomb problem can be written in terms of the quadratic Casimir operator of this symmetry algebra, which provides an algebraic solution to the problem.

In order to generalize Pauli's method, we consider an arbitrary dimension superspace $E(D|d)$ with D bosonic coordinates and d (even) fermionic coordinates. To each position coordinate we associate a conjugate momentum that is realized as a differential operator acting on functions defined on the superspace $E(D|d)$. Then, by means of the antisupersymmetric tensor product of the coordinates and the momenta, we build the canonical realization of the superalgebra

[a)]Electronic mail: minnaert@bortibm4.in2p3.fr

 © 1998 American Institute of Physics

osp$(D|d)$. Next, we define the Runge–Lenz supervector, constant of motion for the Coulomb problem, and we show that this supervector allows us to construct the quasicanonical realization of the superalgebra osp$(D+1|d)$. Finally, we derive the energy spectrum for the Coulomb problem from the eigenvalues of the Casimir operators of the osp$(D+1|d)$ superalgebra. Note that the general case of a (bosonic) D-dimensional space is obtained by setting $d=0$.

The organization of the paper is the following. In Sec. II, we review the original work of Pauli for the three-dimensional Euclidean space $E(3)$ and we write it in a tensorial form that is more suitable for generalization. Section III is devoted to the construction of the canonical and quasicanonical realizations of the osp$(D|d)$ superalgebra. Finally, in Sec. IV we generalize the Pauli's method to an arbitrary dimension orthosymplectic superspace $E(D|d)$.

II. THE COULOMB PROBLEM IN THREE-DIMENSIONS

A. Pauli's method

Let us first recall the basics and results of Pauli's method for the three-dimensional Coulomb problem.[1,5] The Hamiltonian of the Hydrogen atom (in units with $m=e^2=\hbar=1$) is

$$H=\frac{1}{2}\,\mathbf{p}^2-\frac{1}{r}, \tag{2.1}$$

where $r=|\mathbf{r}|$. From the canonical commutation relation $[r_i,p_j]=i\,\delta_{ij}$, it is easy to verify that the angular momentum vector operator $\mathbf{L}=\mathbf{r}\times\mathbf{p}$ satisfies the commutation relations

$$[L_i,L_j]=i\,\epsilon_{ijk}L_k, \tag{2.2}$$

and is a constant of motion

$$[H,\mathbf{L}]=0. \tag{2.3}$$

Pauli introduced another constant of motion, the quantum-mechanical Runge–Lenz vector operator, defined by

$$\mathbf{A}=\frac{1}{2}\,(\mathbf{p}\times\mathbf{L}-\mathbf{L}\times\mathbf{p})-\frac{\mathbf{r}}{r}. \tag{2.4}$$

Use of the commutation relations for $\mathbf{r}$, $\mathbf{p}$, and $\mathbf{L}$ leads to

$$[H,\mathbf{A}]=0. \tag{2.5}$$

Furthermore, one derives the constraints

$$\mathbf{A}\cdot\mathbf{L}=\mathbf{L}\cdot\mathbf{A}=0, \tag{2.6}$$

and the commutation relations between the angular momentum and the Runge–Lenz vector components

$$[L_i,A_j]=i\,\epsilon_{ijk}A_k, \tag{2.7}$$

$$[A_i,A_j]=i(-2H)\,\epsilon_{ijk}L_k. \tag{2.8}$$

The clue to an algebraic solution of the problem is the remark that, in the quantum case as well as in the classical case, the Hamiltonian can be written in terms of the constants of motion $\mathbf{L}^2$ and $\mathbf{A}^2$. Indeed, one has the relation

$$\mathbf{A}^2=1+(2H)(\mathbf{L}^2+1), \tag{2.9}$$

and, therefore, the spectrum of H is related to the eigenvalues of $\mathbf{A}^2$ and $\mathbf{L}^2$. In order to exhibit more precisely this relation, it is interesting to define two new vector operators $\mathbf{M}$ and $\mathbf{N}$, through the combinations (assuming $H<0$)

$$\mathbf{M}=\frac{1}{2}\left(\mathbf{L}+\frac{\mathbf{A}}{\sqrt{-2H}}\right), \tag{2.10}$$

$$\mathbf{N}=\frac{1}{2}\left(\mathbf{L}-\frac{\mathbf{A}}{\sqrt{-2H}}\right). \tag{2.11}$$

They are constants of motion

$$[H,\mathbf{M}]=[H,\mathbf{N}]=0, \tag{2.12}$$

and from Eqs. (2.3), (2.7), and (2.8), it is easy to show that they satisfy the commutation relations of su(2)⊕su(2)=so(4)

$$[M_i,N_j]=0, \tag{2.13}$$

$$[M_i,M_j]=i\epsilon_{ijk}M_k, \tag{2.14}$$

$$[N_i,N_j]=i\epsilon_{ijk}N_k. \tag{2.15}$$

From Eq. (2.9), the expression of the Hamiltonian in terms of these new operators is

$$H=-\frac{1}{2}\frac{1}{2(\mathbf{M}^2+\mathbf{N}^2)+1}. \tag{2.16}$$

But, because of Eq. (2.6), one has the constraint

$$\mathbf{M}^2=\mathbf{N}^2, \tag{2.17}$$

and since **M** satisfies the commutation relations (2.14) of an angular momentum, the eigenvalues of $\mathbf{M}^2$ are $j(j+1)$ with j integer or half-integer, and therefore, the eigenvalues of H are

$$E_n=-\frac{1}{2}\frac{1}{4j(j+1)+1}=-\frac{1}{2}\frac{1}{n^2}, \tag{2.18}$$

with $n=2j+1=1,2,\ldots$. This is the well-known spectrum of the hydrogen atom.

The main result of Pauli's method is Eq. (2.16) that relates the Hamiltonian H of the three-dimensional problem to the quadratic Casimir operator $C_4=2(\mathbf{M}^2+\mathbf{N}^2)$ of the so(4) algebra.

Let us finally recall that the method provides not only the spectrum but also the degeneracy of each level. Indeed, for fixed n, i.e., for fixed j, the number of eigenstates for each of the angular momentum operators **M** and **N** is $2j+1$; therefore, the total degeneracy is $(2j+1)^2=n^2$.

B. Tensorial form of Pauli's method

Instead of the vectorial form recalled in the previous section, Pauli's method can be rewritten in a tensorial form[11] that is more suitable for generalizations. Let us associate to the angular momentum operator **L** the skew symmetric tensor M_{ij} defined by

$$M_{ij}=\epsilon_{ijk}L_k=r_ip_j-r_jp_i \quad (i,j=1,2,3). \tag{2.19}$$

The components of this tensor satisfy the SO(3) commutation relations

$$[M_{ij},M_{kl}]=\frac{1}{i}(\delta_{jk}M_{il}+\delta_{il}M_{jk}-\delta_{jl}M_{ik}-\delta_{ik}M_{jl}), \tag{2.20}$$

and are constants of motion

$$[H,M_{ij}]=0. \tag{2.21}$$

The Runge–Lenz vector **A** is built with the tensor M_{ij} and the vectors r_i and p_j as follows:

$$A_i=\frac{1}{2}\{M_{ij},p_j\}-\frac{r_i}{r}, \tag{2.22}$$

where {,} is set for the usual anticommutator. Using the definitions and the commutation relations, one verifies that this vector is a constant of motion

$$[H,A_i]=0, \tag{2.23}$$

and that the commutation relations between the M_{ij} and the A_i components are

$$[M_{ij},A_k]=\frac{1}{i}(\delta_{jk}A_i-\delta_{ik}A_j), \tag{2.24}$$

$$[A_i,A_j]=i(-2H)M_{ij}. \tag{2.25}$$

In addition, the vector A_i satisfies constraints that result from the orthogonality relations (2.6), namely

$$M_{ij}A_k+M_{jk}A_i+M_{ki}A_j=A_iM_{jk}+A_jM_{ki}+A_kM_{ij}=0. \tag{2.26}$$

Equation (2.16) yields the following relation between the Hamiltonian H and the constants of motion $\mathrm{Tr}(M^2)$ and $\mathbf{A}^2$:

$$\mathbf{A}^2=1+(2H)(1-\tfrac{1}{2}\mathrm{Tr}(M^2)). \tag{2.27}$$

Then, to define the so(4) algebra, we introduce the skew symmetric tensor operators $M'_{ij}=-M'_{ji}$, $(i,j=1,2,3,4)$, such that

$$M'_{ij}=M_{ij}, \quad \text{for } i,j=1,2,3, \tag{2.28}$$

$$M'_{k,4}=-M'_{4,k}=\frac{A_k}{\sqrt{-2H}}, \quad \text{for } k=1,2,3. \tag{2.29}$$

One readily verifies that these six independent operators M'_{ij} satisfy the commutation relations of the so(4) algebra

$$[M'_{ij},M'_{kl}]=\frac{1}{i}(\delta_{jk}M'_{il}+\delta_{il}M'_{jk}-\delta_{jl}M'_{ik}-\delta_{ik}M'_{jl}). \tag{2.30}$$

The quadratic Casimir operator of this algebra reads

$$C'=-\frac{1}{2}\mathrm{Tr}(M'^2)=-\frac{1}{2}\left(\mathrm{Tr}(M^2)+\frac{A^2}{H}\right), \tag{2.31}$$

and Eq. (2.27) implies that the Hamiltonian H can be written

$$H=-\frac{1}{2}\frac{1}{C'+1}. \tag{2.32}$$

In the particular realization of the so(4) algebra represented by the expressions (2.19) and (2.29), because of the relations between traces of the powers of M'

$$\mathrm{Tr}(M'^3)=\frac{1}{2i}\mathrm{Tr}(M'^2), \quad \mathrm{Tr}(M'^4)=\frac{1}{2}[\mathrm{Tr}(M'^2)]^2, \tag{2.33}$$

there is only one algebraically independent Casimir operator, namely the quadratic operator C'. The eigenvalues of C' are

$$C' = k(k+2), \quad k = 0,1,\ldots . \tag{2.34}$$

By substitution of this value in Eq. (2.32) and setting $k = n-1$, we find again the energy spectrum (2.18).

The realization (2.19) of the so(3) algebra is called the canonical realization because the generators M_{ij} are functions of the operators r_i and p_j that satisfy the canonical commutation relations $[r_i, p_j] = i\delta_{ij}$. By analogy, we call quasicanonical realization, the realization of the so(4) algebra that is given by Eqs. (2.28) and (2.29). Indeed, it is built from the canonical realization of the so(3) subalgebra completed with a set of three generators A_i that are also functions of the operators r_i and p_j.

In the following section, we shall generalize this construction of canonical and quasicanonical realizations to orthosymplectic superalgebras and this will provide us a way to generalize Pauli's method for the resolution of the Coulomb problem in arbitrary dimension spaces.

III. CANONICAL AND QUASICANONICAL REALIZATIONS OF ORTHOSYMPLECTIC SUPERALGEBRAS

A. The superspace $E(D|d)$

The elements of the orthosymplectic superspace $E(D|d)$ (d even) are called supervectors. The position and the momentum supervector are denoted, respectively, by $R_a = (r_i \equiv \mathbf{r}, \chi_\alpha)$ and $P_a = (p_i \equiv \mathbf{p}, \pi_\alpha)$. The range of the collective index a is: $a = \{i = 1,\ldots,D; \alpha = 1,\ldots,d\}$. The components r_i (or p_i) are bosonic, i.e., they commute, while the components χ_α (or π_α) are fermionic, i.e., they anticommute: $\{\chi_\alpha, \chi_\beta\} = 0$ (or $\{\pi_\alpha, \pi_\beta\} = 0$).

The contravariant components of the metric tensor are $g^{ab} = (-1)^{ab} g^{ba} = (\delta^{ij}, \Gamma^{\alpha\beta})$ where Γ is the $d \times d$ antisymmetric matrix with $(1, -1, 1, -1, \ldots, -1)$ on the second diagonal. The covariant components are $g_{ab} = (\delta_{ij}, -\Gamma_{\alpha\beta})$. The scalar product $R \cdot P$ of the supervectors R and P is defined by

$$R \cdot P = g^{ab} R_a P_b = \mathbf{r} \cdot \mathbf{p} + \Gamma^{\alpha\beta} \chi_\alpha \pi_\beta . \tag{3.1}$$

The square R^2 of the position supervector, and its norm R are

$$R^2 = \mathbf{r}^2 + \Gamma^{\alpha\beta} \chi_\alpha \chi_\beta , \quad R = r + \frac{1}{2r} \Gamma^{\alpha\beta} \chi_\alpha \chi_\beta , \tag{3.2}$$

In order to define a quantum mechanics in the orthosymplectic superspace $E(D|d)$, we define canonical supercommutation relations (CSR) between the position and the momentum components as

$$[R_a, P_b]_s \equiv R_a P_b - (-1)^{ab} P_b R_a = i g_{ab} , \tag{3.3}$$

i.e., one has $[r_i, p_j] = i\delta_{ij}$ and $\{\chi_\alpha, \pi_\beta\} = -i\Gamma_{\alpha\beta}$.

These CSR can be realized in terms of operators acting on functions defined on the $E(D|d)$ superspace. The components of the position supervector act as multiplication by r_i or χ_α and the components of the momentum supervector are represented by differential operators

$$p_i = \frac{1}{i} \frac{\partial}{\partial r_i}, \quad \pi_\alpha = \frac{1}{i} \frac{\partial}{\partial \chi^\alpha}. \tag{3.4}$$

The Laplace operator on the superspace $E(D|d)$ is defined by $\Delta \equiv -P^2$ and it writes

$$\Delta = \frac{\partial^2}{\partial R^2} + (D - d - 1) \frac{1}{R} \frac{\partial}{\partial R} - \frac{C}{R^2}, \tag{3.5}$$

where C is a positive operator, involving the nonradial variables. The eigenvalues of this operator are determined in the following way. Let $Z_l(R_a)$ be a degree l homogeneous polynomial, solution of the Laplace equation, i.e.,

$$Z_l(\lambda R_a)=\lambda^l Z_l(R_a), \quad \Delta Z_l=0. \tag{3.6}$$

We define reduced variables by $\rho_a=R_a/R$ so that $Z_l(R_a)=R^l Z_l(\rho_a)$. Then, the functions $Z_l(\rho_a)$ are eigenfunctions of the operator C with eigenvalues $l(l+D-d-2)$

$$CZ_l(\rho_a)=l(l+D-d-2)Z_l(\rho_a). \tag{3.7}$$

B. Canonical realization of the osp($D|d$) superalgebra

In the superspace $E(D|d)$, the super angular momentum M is built with the position and momentum supervectors in the following way:

$$M_{ab}=R_aP_b-(-1)^{ab}R_bP_a. \tag{3.8}$$

From the canonical supercommutation relations (3.3), it is easy to verify that the operator M is a super antisymmetric tensor

$$M_{ab}=-(-1)^{ab}M_{ba}, \tag{3.9}$$

and that its components satisfy the super commutation relations of the orthosymplectic superalgebra osp($D|d$)

$$\begin{aligned}[M_{ab},M_{cd}]_s &\equiv M_{ab}M_{cd}-(-1)^{(a+b)(c+d)}M_{cd}M_{ab}\\ &=\frac{1}{i}\,(g_{bc}M_{ad}-(-1)^{cd}g_{bd}M_{ac}\\ &\quad -(-1)^{ab}g_{ac}M_{bd}+(-1)^{ab+cd}g_{ad}M_{bc}).\end{aligned} \tag{3.10}$$

This realization of the superalgebra osp($D|d$) is called the canonical realization. Its essential characteristic is to have only one independent Casimir operator. This fact can be proved as follows. First, it is well known that for any representation of osp($D|d$), the traces of the odd powers of M depend linearly on the traces of the even powers,[12,13] for example

$$\mathrm{Tr}(M^3)=-\tfrac{1}{2}(D-d-2)\mathrm{Tr}(M^2). \tag{3.11}$$

Then, in the canonical realization, one has the relation

$$(M^4)_{ab}=\tfrac{1}{2}\mathrm{Tr}(M^2)(M^2)_{ab}+(\mathrm{td}\leqslant 3), \tag{3.12}$$

where (td$\leqslant$3) means ''terms of degree less than or equal to 3.'' By taking the trace of this relation, we get precisely

$$\mathrm{Tr}(M^4)=\tfrac{1}{2}\mathrm{Tr}(M^2)[\mathrm{Tr}(M^2)-(D-d-2)(D-d-3)]. \tag{3.13}$$

By induction, we find

$$\mathrm{Tr}(M^{2n})=\frac{1}{2^{2n-1}}\,\mathrm{Tr}(M^2)^n+Q_{n-1}(\mathrm{Tr}\,M^2), \tag{3.14}$$

where Q_{n-1} is a polynomial of degree $n-1$.

Hence, we can choose as the unique algebraically independent Casimir operator, the quadratic one

$$\begin{aligned}C&=-\tfrac{1}{2}\mathrm{Tr}(M^2)\equiv-\tfrac{1}{2}g^{ab}M_{ac}g^{cd}M_{db},\\ &=R^2P^2-(R\cdot P)^2+i(D-d-2)R\cdot P.\end{aligned} \tag{3.15}$$

This equation is equivalent to Eq. (3.5) that gives the form of the Laplace operator. The eigenvalues of the operator C were determined in the previous subsection. By using the fact that C is the square of a super-Hermitian operator and must be positive definite, these eigenvalues can be written as

$$\text{for } D \geqslant d+2: \quad C = l(l+D-d-2), \tag{3.16}$$

$$\text{for } D < d+2: \quad C = l(l-D+d+2). \tag{3.17}$$

with $l = 0,1,2...$, and where for the second relation we have performed a change of integer variable.

C. Quasicanonical realization of the osp($D+1|d$) superalgebra

Let us consider a supervector K belonging to the $E(D|d)$ superspace and that satisfies the following relations:

$$[M_{ab}, K_b]_s = \frac{1}{i}(g_{bc}K_a - (-1)^{ab} g_{ac}K_b), \tag{3.18}$$

$$[K_a, K_b]_s = iM_{ab}, \tag{3.19}$$

$$M_{ab}K_c + (-1)^{a(b+c)}M_{bc}K_a + (-1)^{c(a+b)}M_{ca}K_b = 0,$$

$$K_a M_{bc} + (-1)^{a(b+c)}K_b M_{ca} + (-1)^{c(a+b)}K_c M_{ab} = 0, \tag{3.20}$$

where M is the canonical realization of osp($D|d$).

It is possible to gather the operators K_a and M_{bc} into a new super antisymmetric tensor operator $M'_{ab} = -(-1)^{ab}M'_{ba}$, the indices of which take their values in $E(D+1|d)$, i.e., $a = \{i = 1,2,...,D+1; \alpha = 1,...,d\}$. The tensor M'_{ab} is defined by

$$M'_{ab} = M_{ab} \quad \text{for} \quad a,b \in E(D|d),$$

$$M'_{a,D+1} = -M'_{D+1,a} = K_a. \tag{3.21}$$

The components of this tensor satisfy the super commutation relations of the osp($D+1|d$) superalgebra. We call it quasicanonical realization of osp($D+1|d$) because the same relations exist between the traces of the powers as for the canonical realization. Indeed, conditions (3.18)–(3.20) imply the identities

$$(K \cdot M^2)_a = -\tfrac{1}{2}\mathrm{Tr}(M^2)K_a + (\mathrm{td} \leqslant 2), \tag{3.22}$$

$$(M^2 \cdot K)_a = -\tfrac{1}{2}\mathrm{Tr}(M^2)K_a + (\mathrm{td} \leqslant 2), \tag{3.23}$$

$$(M \cdot K)_a (K \cdot M)_b = K^2 (M^2)_{ab} - \tfrac{1}{2}\mathrm{Tr}(M^2)K_a K_b + (\mathrm{td} \leqslant 3). \tag{3.24}$$

These identities allow us to prove the following relation:

$$(M'^4)_{ab} = \tfrac{1}{2}\mathrm{Tr}(M'^2)(M'^2)_{ab} + (\mathrm{td} \leqslant 3), \tag{3.25}$$

that is similar to Eq. (3.12) for the canonical realization. Then, by the same method, one proves that there is only one algebraically independent Casimir operator, the quadratic one

$$C' = -\tfrac{1}{2}\mathrm{Tr}(M'^2). \tag{3.26}$$

Its eigenvalues are denoted by

$$\text{for } D \geqslant d+1: \quad C' = k(k+D-d-1), \tag{3.27}$$

$$\text{for } D < d+1: \quad C' = k(k-D+d+1), \tag{3.28}$$

with $k=0,1,2...$. Note that the relation

$$C'=C+K^2, \tag{3.29}$$

where K^2 is positive definite, implies the inequality $k \geqslant l$ between the eigenvalues of C' and C.

IV. THE COULOMB PROBLEM IN SUPERSPACE $E(D|d)$

In this section we show that Pauli's method can be extended to an arbitrary superspace in the form of the quasicanonical realization of the superalgebra $\mathrm{osp}(D+1|d)$. The Coulomb problem in this superspace is defined by the Hamiltonian

$$H=\frac{1}{2}P^2-\frac{1}{R}, \tag{4.1}$$

where the inverse $1/R$ of the norm of R is

$$\frac{1}{R}=\frac{1}{r}-\frac{1}{2r^3}\Gamma^{\alpha\beta}\chi_\alpha\chi_\beta, \quad \text{where } r=|\mathbf{r}|. \tag{4.2}$$

The super angular momentum operator M in this superspace defined by

$$M_{ab}=R_aP_b-(-1)^{ab}R_bP_a, \tag{4.3}$$

forms the canonical realization of the $\mathrm{osp}(D|d)$ superalgebra. One can show that it is a constant of motion: $[H,M_{ab}]=0$, which means that the Coulomb problem has an $\mathrm{osp}(D|d)$ symmetry.

The extension of Pauli's method is made possible because of the existence of another constant of motion, namely the Runge–Lenz supervector defined by[14]

$$A_a=\frac{1}{2}\{M_{ab},P^b\}_s-\frac{R_a}{R}, \tag{4.4}$$

where $\{,\}_s$ is the super anticommutator. Use of the canonical supercommutation relation leads to $[H,A_a]=0$. Furthermore, one derives the supercommutation relations between the super angular momentum and the Runge–Lenz supervector components, namely

$$[M_{ab},A_c]_s=\frac{1}{i}(g_{bc}A_a-(-1)^{ab}g_{ac}A_b), \tag{4.5}$$

$$[A_a,A_b]_s=i(-2H)M_{ab}, \tag{4.6}$$

and the following relations that generalize the classical orthogonality condition:

$$\begin{aligned} &M_{ab}A_c+(-1)^{a(b+c)}M_{bc}A_a+(-1)^{c(a+b)}M_{ca}A_b=0,\\ &A_aM_{bc}+(-1)^{a(b+c)}A_bM_{ca}+(-1)^{c(a+b)}A_cM_{ab}=0. \end{aligned} \tag{4.7}$$

As in the classical case, there exists a relation between H and the constants of motion. A long calculation yields

$$A^2=1+(2H)(C+\tfrac{1}{4}(D-d-1)^2), \tag{4.8}$$

where C is the quadratic Casimir operator of $\mathrm{osp}(D|d)$ superalgebra. This expression relates the spectrum of H to the dimensions of the superspace and to the eigenvalues of A^2 and C. It is the generalization of Eq. (2.27).

It is now interesting to gather A_a and M_{ab} into a new super antisymmetric tensor operator $M'_{ab}=-(-1)^{ab}M'_{ba}$ the indices of which take their values in $E(D+1|d)$, i.e., $a=\{i=1,2,...,D+1;\alpha=1,...,d\}$. Assuming $H<0$, the tensor M'_{ab} is defined by

$$M'_{ab}=M_{ab} \quad \text{for} \quad a,b\in E(D|d),$$
$$M'_{a,D+1}=-M'_{D+1,a}=\frac{A_a}{\sqrt{-2H}}. \tag{4.9}$$

The tensor M'_{ab} forms the quasicanonical realization of the osp$(D+1|d)$ superalgebra described in the previous section and it is a constant of motion: $[H,M'_{ab}]=0$. The quadratic Casimir operator C' of this algebra is

$$C'=-\frac{1}{2}\,\mathrm{Tr}(M'^2)=C-\frac{A^2}{2H}. \tag{4.10}$$

Finally, relation (4.8) between H and the constants of motion yields the expression

$$H=-\frac{1}{2}\,\frac{1}{C'+(1/4)(D-d-1)^2}, \tag{4.11}$$

This relation is the direct generalization of Eq. (2.32). It yields the spectrum of H, when the Casimir operator C' is replaced by its eigenvalues given by Eqs. (3.27) and (3.28).

For $D\neq d+1$, in order to obtain expressions similar to the usual ones, we set $n=k+1$, with $n=1,2,...$ and we get

(i) for $D>d+1$

$$H=-\frac{1}{2}\,\frac{1}{(n+(1/2)(D-d-3))^2}, \tag{4.12}$$

and the eigenvalues of the square of the angular momentum operator are

$$C=l(l+D-d-2), \quad l=0,...,n-1. \tag{4.13}$$

(ii) for $D<d+1$

$$H=-\frac{1}{2}\,\frac{1}{(n+(1/2)(d-D-1))^2}, \tag{4.14}$$

and the eigenvalues of the square of the angular momentum operator are

$$C=l(l+2+d-D), \quad l=0,...,n-1. \tag{4.15}$$

For $D=d+1$, we set $n=k$ and we get

$$H=-\frac{1}{2}\,\frac{1}{n^2}, \tag{4.16}$$

and the eigenvalues of the square of the angular momentum operator are

$$C=l(l+1), \quad l=0,...,n-1. \tag{4.17}$$

For $d=0$, Eq. (4.12) yields the energy spectrum of the Coulomb problem in an ordinary D dimensional space.

V. CONCLUSION

In this paper we have shown how Pauli's method for the resolution of Coulomb problem can be extended to an arbitrary dimensional superspace. First we have written Pauli's method in a tensorial form suitable for generalization. Then, the canonical realization of the osp$(D|d)$ superalgebra has been built by means of differential operators and the quasicanonical realization of the osp$(D+1|d)$ superalgebra has been constructed by extension of the previous realization by a supervector operator. For the Coulomb problem, this supervector is the generalization of the

Runge–Lenz vector for the hydrogen atom. Finally, we have deduced the energy spectrum of the Coulomb problem in terms of the eigenvalues of the quadratic operator C' of the $\mathrm{osp}(D+1|d)$ superalgebra. This result is valid for spaces and superspaces with arbitrary dimension.

It is interesting to note that for the particular cases $D=d-1$, $d+1$, $d+3$ the energy spectrum is identical to that of the hydrogen atom. For superspaces, this last result has been confirmed by a direct calculation for the simplest superspace $E(1|2)$,[15] and for the smallest superspace that contains the ordinary $E(3)$ space, namely the superspace $E(3|2)$.[16]

[1] W. Pauli, Jr., Z. Phys. **36**, 336 (1926).
[2] B. L. Van der Waerden, *Sources of Quantum Mechanics* (North Holland, Amsterdam, 1967).
[3] V. A. Fock, Z. Phys. **98**, 145 (1935).
[4] A. O. Barut, Phys. Rev. **135**, B839 (1964).
[5] C. E. Wulfman, ''Dynamical groups in Atomic and Molecular Physics'' in *Group Theory and its Applications, Vol. II*, edited by E. M. Loebl (Academic, New York, 1971), p. 145.
[6] H. V. McIntosh, ''Symmetry and Degeneracy'' in *Group Theory and its Applications, Vol. II*, edited by E. M. Loebl (Academic, New York, 1971).
[7] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[8] M. J. Englefield, *Group Theory and the Coulomb Problem* (Wiley, New York, 1972).
[9] E. Witten, Phys. Today **33**, 33 (1980).
[10] L. P. Horwitz, J. Math. Phys. **34**, 645 (1993).
[11] This method was used in Ref. 10 to write the Pauli problem in relativistic form.
[12] G. Racah, Lincei Rend. Sci. Fis. Mat. Nat. **8**, 108 (1950).
[13] P. D. Jarvis and H. S. Green, J. Math. Phys. **20**, 2115 (1979).
[14] M. Daumens and P. Minnaert, ''Coulomb problem in superspaces,'' in *Quantum Groups: Formalism and Applications. XXX Karpacz Winter School of Theoretical Physics, Poland (1994)*, edited by J. Lukierski, Z. Popowicz, and J. Sobczyk (Polish Scientific Publishers, Warsawa, 1995).
[15] M. Daumens and Y. Noirot, Nuovo Cimento A **110**, 587 (1997).
[16] M. Daumens, Nuovo Cimento B (in press).

Wick's theorem for nonsymmetric normal ordered products and contractions

T. S. Evans[a] and T. W. B. Kibble[b]
Theoretical Physics, Blackett Laboratory, Imperial College, Prince Consort Road, London, SW7 2BZ, United Kingdom

D. A. Steer[c]
D.A.M.T.P., Silver Street, Cambridge, CB3 9EW, United Kingdom

(Received 4 February 1998; accepted for publication 1 July 1998)

We consider arbitrary splits of field operators into two parts; $\psi=\psi^{+}+\psi^{-}$, and use the corresponding definition of normal ordering introduced earlier [T. S. Evans and D. A. Steer, Nucl. Phys. B **474**, 481 (1996)]. In this case the normal ordered products and contractions have none of the special symmetry properties assumed in existing proofs of Wick's theorem. Despite this, we prove that Wick's theorem still holds in its usual form as long as the contraction is a *c*-number. Wick's theorem is thus shown to be much more general than existing derivations suggest, and we discuss possible simplifying applications of this result. © *1998 American Institute of Physics.* [S0022-2488(98)01911-2]

I. INTRODUCTION

Wick's theorem is at the core of perturbative calculations in the canonical operator approach to quantum field theory: on expansion of the *S*-matrix, it is used to relate time ordered products of field operators to normal ordered products and contractions.[1] Within the traditional definition of normal ordered products, the interaction picture field operators, $\psi(x)$, are split into positive $\psi^{(0+)}$ and negative $\psi^{(0-)}$ energy waves containing annihilation and creation operators, respectively:

$$\psi(x)=\psi^{(0+)}(x)+\psi^{(0-)}(x). \tag{1.1}$$

Normal ordering, $N^{(0)}[\bullet]$, then places the annihilation (creation) operators to the right (left), otherwise leaving the order of the operators unchanged. Hence the vacuum expectation values of normal ordered products are guaranteed to vanish,

$$\langle 0|N^{(0)}[\bullet]|0\rangle=0. \tag{1.2}$$

Calculations therefore proceed in a straightforward manner by simply taking the vacuum expectation value of the operator identity called Wick's theorem.

In the literature, Wick's theorem has been proved either through the use of a generating functional,[2–4] or by the manipulation of products of operators.[1,5–7] In both cases, an essential ingredient of the proof is the fact that the normal ordered products defined above are *symmetric* in the following sense. Let us denote the normal ordered product of n (possibly different) field operators $\psi_i:=\psi(x_i)$ evaluated at different space time points x_i by

$$N^{(0)}_{1,2,\ldots,n}:=N^{(0)}[\psi_1\psi_2\ldots\psi_n]. \tag{1.3}$$

Then since $[\psi^{(0+)}(x),\psi^{(0+)}(x')]_\sigma=[\psi^{(0-)}(x),\psi^{(0-)}(x')]_\sigma=0 \quad \forall x,x'$ (where $\sigma=+1$ for bosonic operators, and $\sigma=-1$ for fermionic ones), it follows that

$$N^{(0)}_{1,2,\ldots,n}=(-1)^p N^{(0)}_{a,b,\ldots,t}\forall n>0. \tag{1.4}$$

[a]Electronic mail: T.Evans@ic.ac.uk, WWW: http://euclid.tp.ph.ic.ac.uk/~time
[b]Electronic mail: t.kibble@ic.ac.uk
[c]Electronic mail: D.A.Steer@damtp.cam.ac.uk

 © 1998 American Institute of Physics

Here $a,b,\ldots,t$ is any permutation of 1,2,$\ldots,n$ and p is the number of times pairs of fermionic fields are interchanged in moving from an order 1,2,$\ldots,n$ to an $a,b,\ldots,t$ order of fields. Normal ordered products satisfying (1.4) are said to be *symmetric*. The reason why this symmetry is central to the existing derivations of Wick's theorem will be explained in detail in Sec. III below where we work with the generating functional approach. However, the following comments may help illustrate the reason. Recall that Wick's theorem relates time ordered products, normal ordered products and contractions. One way to prove this theorem[1,6] is to begin by considering a given order of times. The resulting (single) contribution from the time ordered product may then be shown to be equivalent to a combination of normal ordered products and contractions. *If* these normal ordered products and contractions are symmetric, then each contribution to the time ordered product takes the same form, and thus Wick's theorem is proved.[1,6] However, if they are *not symmetric*, this is no longer the case (see Sec. III).

In general, however, normal ordering with respect to annihilation and creation operators does *not* simplify calculations. As an example, consider how, in the presence of symmetry breaking, one must normal order with respect to new annihilation and creation operators related by a Bogoliubov transformation to the ones appearing in the field.[5,6]. This is equivalent to choosing to split the field in a different way from (1.1). A more recent example[8] was provided by a system in thermal equilibrium at a temperature T. There the thermal expectation value of a normal ordered product of creation and annihilation operators ($\hat{a}^\dagger$ and $\hat{a}$, respectively) no longer vanishes in the way that it did at $T=0$:

$$\langle\langle N^{(0)}[\hat{a},\hat{a}^\dagger]\rangle\rangle = n_{be}, \quad n_{fd} \neq 0. \tag{1.5}$$

(Here n_{be}, n_{fd} are the Bose–Einstein and Fermi–Dirac distributions, respectively.) As a result, perturbative calculations at finite T would not seem to be simplified by the use of Wick's theorem. This point has caused some confusion in the literature[6,9,10] and it was addressed in Ref. 8. There it was shown that Wick's theorem could be used in the usual way at finite temperature provided one split the field in a different way from the usual one into creation and annihilation operators (1.1).

Thus, in general, one must consider arbitrary splits of the field operators. The precise split used will reflect the physical environment of the problem if it is to simplify practical calculations in the usual way. We denote the arbitrary split by

$$\psi(x) = \psi^+(x) + \psi^-(x), \tag{1.6}$$

where $\psi^+(x)$ and $\psi^-(x)$ are unspecified in this paper, so that we can accommodate any physical problem. We still call the two parts ''positive'' and ''negative'' even though the split is general and no longer necessarily into positive and negative energy waves. The corresponding (natural) generalization of normal ordering is then expressed in terms of $\psi^+(x)$ and $\psi^-(x)$, and is defined as follows:

Normal ordering of products of fields, $N[\bullet]$, places all the ''positive'' parts to the right and the ''negative'' parts to the left, otherwise leaving the order of the fields unchanged. A sign change occurs whenever two fermion field operators are interchanged.

Note that this reduces to the usual definition, $N^{(0)}[\bullet]$, when the split is into the annihilation and creation operators appearing in the field. We have used the (0) superscript to denote this traditional and very special case. So, for example, in the two-point case one has

$$N[\psi_1\psi_2] := N_{1,2} = \psi_1^+\psi_2^+ + \psi_1^-\psi_2^+ + \sigma\psi_2^-\psi_1^+ + \psi_1^-\psi_2^-. \tag{1.7}$$

Such a generalization of normal ordering was successfully used for thermal field theory in Ref. 8, and is equivalent to discussions of normal ordering involving Bogoliubov transformations and symmetry breaking.[5,6] Contractions are then defined in the usual way through the two-point time ordered products which, being independent of the split, take the canonical form

$$T_{1,2} := T[\psi_1\psi_2] = \psi_1\psi_2\theta(t_1-t_2) + \sigma\psi_2\psi_1\theta(t_2-t_1). \tag{1.8}$$

(In general the θ functions may be defined for complex times and hence be contour dependent—as, for example, in Euclidean quantum field theory and thermal field theory.[2,11] Thus the time ordered products are also contour dependent. However, the analysis presented here holds regardless of contour.) Using (1.7) and (1.8) the contraction $D_{1,2}$ is

$$D_{1,2}=D[\psi_1,\psi_2]:=T_{1,2}-N_{1,2}=\theta(t_1-t_2)[\psi_1^+,\psi_2^-]_\sigma \tag{1.9}$$

$$-\theta(t_2-t_1)\{[\psi_1^+,\psi_2^+]_\sigma+[\psi_1^-,\psi_2^+]_\sigma+[\psi_1^-,\psi_2^-]_\sigma\}. \tag{1.10}$$

Now the key point is that for such arbitrary splits, the contraction and normal product are *not* generally symmetric: from (1.7) and (1.10) the condition for symmetry is

$$D_{1,2}=\sigma D_{2,1}, N_{1,2}=\sigma N_{2,1}\Leftrightarrow[\psi_1^+,\psi_2^+]_\sigma+[\psi_1^-,\psi_2^-]_\sigma=0, \tag{1.11}$$

which is not necessarily satisfied. If these two-point relations are true, then all higher order normal ordered products are symmetric in the sense of (1.4). This is only easy to prove if one has already proved that Wick's theorem holds independent of the symmetry of all normal ordered products!

It is conceivable that in some situations, for example systems out of equilibrium, calculations may be simplified by the use of a split for which the normal ordered products are not symmetric [i.e., (1.11) is not satisfied]. Indeed, out of equilibrium it may be possible to find a split such that (1.11) is true for specific times, but without time-translation symmetry it is doubtful that this split will ensure (1.11) will be true for all times. Wick's theorem, which relates normal ordered products (1.7), time ordered products (1.8), and contractions (1.10), has only been proved for the special case where the normal ordered products and contractions are symmetric. The question this paper asks then is: Does the usual form of Wick's theorem still hold when normal ordered products are not symmetric?

Our answer is yes. Wick's theorem is, in fact, more general than the existing proofs suggest, namely: *if the contraction is a c-number (and hence the split is linear in creation and annihilation operators), Wick's theorem holds for both symmetric and nonsymmetric normal ordered products.* Possible practical applications of this result, in particular to time dependent nonequilibrium systems, are discussed in the conclusions.

In Sec. II we derive the generating functional for Wick's theorem in terms of the generalized definition of normal ordering. We stress the places in which the nonsymmetry of products means that our derivation differs from others existing in the literature. In Sec. III we use the generating functional to obtain Wick's theorem. For symmetric products the proof is immediate—for nonsymmetric ones it is considerably more complicated. We finally summarize our work and also comment on a claim made in the original paper on thermal field theory by Matsubara.[9] Applications for our work are then discussed.

II. THE GENERATING FUNCTIONAL FOR WICK'S THEOREM

In this section we derive the generating functional for Wick's Theorem – that is, we prove the identity

$$T\left[\exp\left\{-i\int d^4x j(x)\psi(x)\right\}\right]=N\left[\exp\left\{-i\int d^4x j(x)\psi(x)\right\}\right]$$
$$\times\exp\left\{-\frac{1}{2}\int d^4x d^4y j(x)D[\psi(x)\psi(y)]j(y)\right\}. \tag{2.1}$$

Here $\int d^4x=\int dt\int d^3\mathbf{x}$, and the precise range of integration does not affect our arguments. For simplicity, we just consider bosonic fields. (Comments regarding fermionic fields will be made later.) Thus the $j(x)$'s are *c*-number sources and since Wick's theorem is used in perturbation theory, the field operators $\psi(x)$ are in the interaction picture. The normal ordered products $N[\bullet]$ and contraction $D[\psi(x)\psi(y)]$ have been defined above in terms of the arbitrary split of the fields into positive and negative parts; $\psi(x)=\psi^+(x)+\psi^-(x)$. Following the discussion in the previous section, we therefore make no assumptions about their symmetry properties. The first part of the proof presented here follows closely that of Refs. 2 and 3.

Recall that for two operators A and B

$$e^{A+B}=e^{A}e^{B}e^{(1/2)[B,A]} \tag{2.2}$$

if and only if

$$[A,[A,B]]=[B,[A,B]]=0. \tag{2.3}$$

Recall, too, that time ordering puts the larger times to the left, and so, for example, in the case of two operators $A(t_1)$ and $A(t_2)$ with $t_2>t_1$,

$$T[e^{A(t_1)+A(t_2)}]=e^{A(t_2)}e^{A(t_1)}=T[e^{A(t_1)}e^{A(t_2)}]. \tag{2.4}$$

We comment that a similar relation holds for any operation which places the two operators in a given order. This will be used for normal ordered products below in Eq. (2.11).

Consider the lhs of (2.1),

$$T[e^{-i\int dt\int d^3x j(x)\psi(x)}]:=T[e^{-i\int dt O(t)}]. \tag{2.5}$$

Let t_i and t_f denote the initial and final time. Divide the total time interval into N equal intervals with $\Delta t=(t_f-t_i)/N$ and

$$t_f=t_N>t_{N-1}>\cdots>t_1>t_0=t_i\,. \tag{2.6}$$

By definition,

$$\int_{t_i}^{t_f}dt\,O(t)=\lim_{N\to\infty}\Delta t\sum_{j=1}^{N}O(t_j). \tag{2.7}$$

So here,

$$\begin{aligned}T[e^{-i\int_{t_i}^{t_f}dtO(t)}]&=\lim_{N\to\infty}T[e^{-i\Delta t\Sigma_{j=1}^{N}O(t_j)}]\\&=\lim_{N\to\infty}[e^{-i\Delta tO(t_N)}e^{-i\Delta tO(t_{N-1})}\cdots e^{-i\Delta tO(t_1)}] \qquad (2.8)\\&=\lim_{N\to\infty}[e^{-i\Delta t\Sigma_{k=1}^{N}O(t_k)}e^{-(1/2)(\Delta t)^2\Sigma_{1<k<l<N}[O(t_l),O(t_k)]}]. \qquad (2.9)\end{aligned}$$

In going from (2.8) to (2.9) we have used (2.2), which is legitimate since $[O(t_l),O(t_k)]$ is a c-number. Note also that in this exponential $t_l>t_k$. Thus using the definition of $O(t)$, (2.9) gives

$$T[e^{-i\int d^4x j(x)\psi(x)}]=e^{-i\int d^4x j(x)\psi(x)}e^{-(1/2)\int d^4x\int d^4y\,\theta(x_0-y_0)j(x)[\psi(x),\psi(y)]j(y)}. \tag{2.10}$$

In order to relate the time ordered product in (2.10) to a normal ordered product, consider an arbitrary split of the fields into ''positive'' and ''negative'' parts introduced in (1.6) and the corresponding definition of normal ordering given in the previous section. Using (2.2), the first exponential on the rhs of (2.10) may be rewritten as

$$\begin{aligned}e^{-i\int d^4x j(x)\psi(x)}&=e^{-i\int d^4x j(x)\{\psi^-(x)+\psi^+(x)\}}\\&=e^{-i\int d^4x j(x)\psi^-(x)}e^{-i\int d^4x j(x)\psi^+(x)}e^{(1/2)\int d^4x\int d^4y j(x)[\psi^-(x),\psi^+(y)]j(y)}\\&=N[e^{-i\int d^4x j(x)\psi(x)}]e^{1/2\int d^4x\int d^4y j(x)[\psi^-(x),\psi^+(y)]j(y)}. \qquad (2.11)\end{aligned}$$

In obtaining (2.11) we have used (2.4) for normal ordered products, and (2.2), which requires that

$$[\psi^{\pm},[\psi^+,\psi^-]]=0. \tag{2.12}$$

This condition is satisfied if ψ^+ and ψ^- are linear in annihilation and creation operators, and in turn it ensures that the contraction is a *c*-number. Substituting (2.11) into (2.10) gives

$$T[e^{-i\int d^4x j(x)\psi(x)}]=N[e^{-i\int d^4x j(x)\psi(x)}]e^{-(1/2)\int d^4x\int d^4y j(x)\{\theta(x_0-y_0)[\psi(x),\psi(y)]-[\psi^-(x),\psi^+(y)]\}j(y)}. \tag{2.13}$$

Our proof now diverges from that given in the usual texts[2,3] since we are considering the possibility of having an *arbitrary* (unsymmetric) split of the fields rather than a (symmetric) one into annihilation and creation operators. Look at the part in curly brackets in the second exponential of (2.13); it is

$$\begin{aligned}
&\theta(x_0-y_0)[\psi(x),\psi(y)]-[\psi^-(x),\psi^+(y)]\\
&\quad=[\psi^+(y),\psi^-(x)]+\theta(x_0-y_0)\{[\psi^+(x),\psi^+(y)]+[\psi^-(x),\psi^+(y)]\\
&\qquad+[\psi^+(x),\psi^-(y)]+[\psi^-(x),\psi^-(y)]\}\\
&\quad=\theta(y_0-x_0)[\psi^+(y),\psi^-(x)]-\theta(x_0-y_0)\\
&\qquad\times\{[\psi^+(y),\psi^+(x)]+[\psi^-(y),\psi^+(x)]+[\psi^-(y),\psi^-(x)]\}\\
&\quad=D[\psi(y)\psi(x)],
\end{aligned} \tag{2.14}$$

where we have used the definition of the contraction in (1.10).

Thus, now substitute (2.14) into the final exponential of (2.13), relabel $x\leftrightarrow y$ and use the fact that the $j(x),j(y)$ are *c*-numbers to switch their positions. This enables (2.13) to be rewritten as

$$\begin{aligned}
T\left[\exp\left\{-i\int d^4x j(x)\psi(x)\right\}\right]&=N\left[\exp\left\{-i\int d^4x j(x)\psi(x)\right\}\right]\\
&\quad\times\exp\left\{-\frac{1}{2}\int d^4x d^4y j(x)D[\psi(x)\psi(y)]j(y)\right\},
\end{aligned} \tag{2.15}$$

thus proving (2.1). This equation is also true for fermionic operators. There the sources are Grassmann variables, and the contraction, defined in (1.10), contains anticommutators.

III. WICK'S THEOREM FOR NONSYMMETRIC PRODUCTS

Wick's theorem is obtained by functionally differentiating (2.15) with respect to the sources $j(x)$. Observe that we need only work with an even number of fields as the perturbation is about the expectation value of the fields, and hence the expectation value of an odd number of fields will always vanish.

Functionally differentiate (2.15) with respect to the sources $2n$ times and then set every $j(x_a)=j_a$ to zero. Since time-ordered products of fields evaluated at the same time points are undefined, the indices *a* refer to different space–time points and are therefore all different. The lhs of (2.15) gives

$$T\left[\frac{\delta}{\delta j_1\cdots\delta j_{2n}}\exp\{-ij\psi\}\right]\Big|_{j_a=0}=T\left[\frac{\delta}{\delta j_1\cdots\delta j_{2n}}\frac{(-ij\psi)^{2n}}{(2n)!}\right] \tag{3.1}$$

$$=\sum_{\text{perm}\{a\}}(-1)^p\frac{T_{a_1,a_2,\ldots,a_{2n}}}{(2n)!} \tag{3.2}$$

$$=T_{1,2,3,\ldots,2n}. \tag{3.3}$$

In (3.2) the sum takes $a_1,\ldots,a_{2n}$ through all permutations of 1,…,2*n*;

$$\text{perm}\{a\}=\begin{pmatrix}1&2&3&\cdots&2n\\a_1&a_2&a_3&\cdots&a_{2n}\end{pmatrix} \tag{3.4}$$

and p is the number of times one has to interchange pairs of fermions in moving from 1,…,2n to $a_1,\ldots,a_{2n}$ order. The last identity [going from (3.2) to (3.3)] follows from the fact that by definition time ordered products are symmetric in their indices, and that there are $(2n)!$ permutations of the numbers 1,2,…,2n. This type of manipulation will *not* hold for normal ordered products and contractions since they are *not*, in general, symmetric in their indices.

Carrying out a similar operation for the rhs of (2.15) gives a general form of Wick's theorem for nonsymmetric products:

$$T_{1,2,\ldots,2n}=\sum_{m=0}^{n}\sum_{\mathrm{perm\{a\}}}(-1)^p\left[\frac{N_{a_1,a_2,\ldots,a_{2m}}}{(2m)!}\cdot\frac{1}{(n-m)!}\left(\prod_{j=m+1}^{n}\frac{1}{2}D_{a_{2j-1},a_{2j}}\right)\right]. \tag{3.5}$$

In the traditional split of the fields into annihilation and creation operators, the normal ordered products $N^{(0)}_{1,2,\ldots}$ and contraction $D^{(0)}_{1,2}$ are *symmetric* as was discussed in Sec. I. As a result, the sum over permutations on the rhs of (3.5) can easily be evaluated and thus (3.5) rapidly reduces to the well known form of Wick's theorem. For two bosonic fields this is

$$T_{1,2}=N^{(0)}_{1,2}+D^{(0)}_{1,2}, \tag{3.6}$$

and for four bosonic fields

$$\begin{aligned}T_{1,2,3,4}=&N^{(0)}_{1,2,3,4}+D^{(0)}_{1,2}N^{(0)}_{3,4}+D^{(0)}_{1,3}N^{(0)}_{2,4}+D^{(0)}_{1,4}N^{(0)}_{2,3}+D^{(0)}_{2,3}N^{(0)}_{1,4}+D^{(0)}_{2,4}N^{(0)}_{1,3}\\&+D^{(0)}_{3,4}N^{(0)}_{1,2}+D^{(0)}_{1,2}D^{(0)}_{3,4}+D^{(0)}_{1,3}D^{(0)}_{2,4}+D^{(0)}_{1,4}D^{(0)}_{2,3}.\end{aligned} \tag{3.7}$$

Higher order relations follow the same pattern. Note that in the above expressions we have *chosen* to arrange the indices on any one given normal ordered product and contraction such that the smaller indices are to the left of the larger ones. In the case of symmetric products, this ordering is, of course, not important. However, it will also be used below where, starting from (3.5), we show that the above expressions [Eqs. (3.6) and (3.7)] hold even for *nonsymmetric* products N and D, *as long as the contraction is a c-number*. We again consider bosonic fields. The results below can be extended to fermionic fields by inserting a minus sign every time the positions of two fermionic fields are interchanged. The following relations and lemmas are useful for the proof.

A. Lemmas and relations

Lemma 1: Two-point normal ordered products and contractions are related by

$$N_{2,1}-N_{1,2}=D_{1,2}-D_{2,1}. \tag{3.8}$$

Proof: By definition of the contraction (1.9)

$$\begin{aligned}D_{1,2}-D_{2,1}=&\theta_{1,2}[\psi_1^+,\psi_2^-]-\theta_{2,1}\{[\psi_1^+,\psi_2^+]+[\psi_1^-,\psi_2^+]+[\psi_1^-,\psi_2^-]\}-\theta_{2,1}[\psi_2^+,\psi_1^-]\\&+\theta_{1,2}\{[\psi_2^+,\psi_1^+]+[\psi_2^-,\psi_1^+]+[\psi_2^-,\psi_1^-]\}\\=&[\psi_2^+,\psi_1^+]+[\psi_2^-,\psi_1^-]=N_{2,1}-N_{1,2},\end{aligned} \tag{3.9}$$

where $\theta_{1,2}=\theta(t_1-t_2)$ and where the last line follows directly from the definition of normal ordering (1.7).

Lemma 2: Let X and Y be arbitrary operators. Then if the contraction is a c-number,

$$N[X\psi_1\psi_2Y]=N[X\psi_2\psi_1Y]+(D_{2,1}-D_{1,2})N[XY]. \tag{3.10}$$

Proof: By definition

$$N[X\psi_1^+\psi_2^-Y]=N[X\psi_2^-\psi_1^+Y]\forall X,Y. \tag{3.11}$$

Therefore

$$
\begin{aligned}
N[X\psi_1\psi_2Y] &= N[X\psi_1^+\psi_2^+Y]+N[X\psi_1^+\psi_2^-Y]+N[X\psi_1^-\psi_2^+Y]+N[X\psi_1^-\psi_2^-Y] \\
&= N[X\psi_2^+\psi_1^+Y]+N[X\psi_2^-\psi_1^+Y]+N[X\psi_2^+\psi_1^-Y] \\
&\quad +N[X\psi_2^-\psi_1^-Y]+N[X[\psi_1^+,\psi_2^+]Y]+N[X[\psi_1^-,\psi_2^-]Y] \\
&= N[X\psi_2\psi_1Y]+(D_{2,1}-D_{1,2})N[XY], \qquad (3.12)
\end{aligned}
$$

provided $[\psi_1^+,\psi_2^+]$ and $[\psi_1^-,\psi_2^-]$ are *c*-numbers. This implies that both ψ^+ and ψ^- are linear in the creation and annihilation operators, and hence that *the contraction is a c-number*.

Relation 3: The number of ways of filling p indistinguishable bags each with exactly two marbles labeled $1,2,\ldots,2q$ *is precisely*

$$
[(2q)(2q-1)][(2q-2)(2q-3)]\cdots[(2q-2p+2)(2q-2p+1)]\frac{1}{p!}\frac{1}{2^p}=\frac{(2q)!}{(2q-2p)!2^p p!}. \qquad (3.13)
$$

Consequence 3.1: Let labels a_i's take values from $1,2,3,\ldots,2q$. The number of distinct configurations of p contractions $D_{a_1,a_2}D_{a_3,a_4}\cdots D_{a_{2p-1},a_{2p}}$ with unordered indices is given by (3.13).

Consequence 3.2: For a product of p contractions and one normal ordered product, with a total number of $2q$ indices; $N_{a_1,a_2,\ldots,a_{2q-2p}}D_{a_{2q-2p+1},a_{2q-2p+2}}\cdots D_{a_{2q-1},a_{2q}}$, the number of distinct configurations with unordered indices is again given by (3.13).

Using these, we now show that (3.6), (3.7) and the usual higher order version of Wick's theorem holds even for nonsymmetric normal ordered products and contractions, as long as the contraction is a *c*-number.

B. Proof for $n=1$

This is the easiest case. For $n=1$, (3.5) gives

$$
T_{1,2}=\tfrac{1}{2}(N_{1,2}+N_{2,1}+D_{1,2}+D_{2,1}). \qquad (3.14)
$$

Substituting (3.8) into (3.14) gives

$$
T_{1,2}=N_{1,2}+D_{1,2}, \qquad (3.15)
$$

showing that for $n=1$ (3.6) holds even for nonsymmetric normal ordered products and contractions. [This result is, of course, consistent with the definition of the contraction; from (1.9), $D_{1,2}:=T_{1,2}-N_{1,2}$ and similarly $D_{2,1}:=T_{2,1}-N_{2,1}=T_{1,2}-N_{2,1}$. Adding these two equations gives (3.14).]

C. Proof for $n=2$

For $n=2$, (3.5) gives

$$
\begin{aligned}
T_{1,2,3,4} &= \sum_{\text{perm}\{a\}}\sum_{m=0}^{2}\left[\frac{N_{a_1,a_2,\ldots,a_{2m}}}{(2m)!}\cdot\frac{1}{(2-m)!}\left(\prod_{j=m+1}^{2}\frac{1}{2}D_{a_{2j-1},a_{2j}}\right)\right] \\
&= \sum_{\text{perm}\{a\}}\left(\frac{1}{4!}N_{a_1,a_2,a_3,a_4}+\frac{1}{2!}\frac{1}{2}N_{a_1,a_2}D_{a_3,a_4}+\frac{1}{2!}\frac{1}{2}\frac{1}{2}D_{a_1,a_2}D_{a_3,a_4}\right) \\
&= \sum_{\text{perm}\{a\}}\frac{1}{4!}(N_{a_1,a_2,a_3,a_4}+6N_{a_1,a_2}D_{a_3,a_4}+3D_{a_1,a_2}D_{a_3,a_4}) \qquad (3.16) \\
&= \sum_{\text{perm}\{a\}}\frac{1}{4!}S_{a_1,a_2,a_3,a_4}, \qquad (3.17)
\end{aligned}
$$

where in (3.16) the first term corresponds to $m=2$, the second to $m=1$, and the third to $m=0$, and in (3.17) S_{a_1,a_2,a_3,a_4} will be defined later. Provided we choose S_{a_1,a_2,a_3,a_4} correctly (see below),

proof that $T_{1,2,3,4}$ is given by the usual expression [i.e., (3.7)] even for nonsymmetric N and D will then follow directly from (3.17) if we can show that the chosen S_{a_1,a_2,a_3,a_4} is symmetric under interchange of any two indices.

To do that, consider initially the second term of (3.16); $\Sigma_{\text{perm}\{a\}} 6N_{a_1,a_2}D_{a_3,a_4}$. The first important point to notice is that because of the sum over permutations, the order in which the a_i indices are written is irrelevant; for example,

$$\sum_{\text{perm}\{a\}} N_{a_1,a_2}D_{a_3,a_4} = \sum_{\text{perm}\{a\}} N_{a_1,a_3}D_{a_2,a_4}. \tag{3.18}$$

Second, the factor of 6 can be understood in the following way. From Consequence 3.2 we know that for a product of one normal ordered product and one contraction with a total of $2q=4$ indices, there are $4!/2!2=6$ distinct configurations with unordered indices. Thus we can rewrite the second term of (3.16) as a sum over the six distinct configurations with unordered indices:

$$\begin{aligned}\sum_{\text{perm}\{a\}} 6N_{a_1,a_2}D_{a_3,a_4} = \sum_{\text{perm}\{a\}} [&D_{a_1,a_2}N_{a_3,a_4}+D_{a_1,a_3}N_{a_2,a_4}+D_{a_1,a_4}N_{a_2,a_3}\\ &+D_{a_2,a_3}N_{a_1,a_4}+D_{a_2,a_4}N_{a_1,a_3}+D_{a_3,a_4}N_{a_1,a_2}].\end{aligned} \tag{3.19}$$

Note that since the indices on the contractions and normal ordered products are unordered, there was still some freedom in writing down the above expression. We have made the particular choice to write the a_i indices on each individual N and each individual D with the smallest value of i to the left. The reason for this will become clear later.

Now carry out a similar procedure for the third term of (3.16) involving the product of two contractions: the three terms arise from the three distinct configurations of two contractions with unordered indices (see Consequence 3.1), and we then chose to write the indices on each individual contraction such that the smallest value of i is to the left. Thus S_{a_1,a_2,a_3,a_4} in (3.17) is defined by

$$\begin{aligned}S_{a_1,a_2,a_3,a_4} := &N_{a_1,a_2,a_3,a_4}+D_{a_1,a_2}N_{a_3,a_4}+D_{a_1,a_3}N_{a_2,a_4}+D_{a_1,a_4}N_{a_2,a_3}\\ &+D_{a_2,a_3}N_{a_1,a_4}+D_{a_2,a_4}N_{a_1,a_3}+D_{a_3,a_4}N_{a_1,a_2}+D_{a_1,a_2}D_{a_3,a_4}\\ &+D_{a_1,a_3}D_{a_2,a_4}+D_{a_1,4}D_{a_2,a_3}.\end{aligned} \tag{3.20}$$

Hence we see that S_{a_1,a_2,a_3,a_4} is given by the rhs of (3.7), where instead of the symmetric normal ordered products and contractions one uses the more general nonsymmetric normal ordered products and contractions: this was the reason why (3.16) was decomposed as above.

To show that S_{a_1,a_2,a_3,a_4} is symmetric under the interchange of any two indices, consider a particular order of the indices, say, a,b,c,d, and then the effect of $a\leftrightarrow b$ say. Now

$$\begin{aligned}S_{a,b,c,d} = &N_{a,b,c,d}+D_{a,b}N_{c,d}+D_{c,d}(N_{a,b}+D_{a,b})+[D_{a,c}N_{b,d}+D_{b,c}N_{a,d}]\\ &+[D_{a,d}N_{b,c}+D_{b,d}N_{a,c}]+[D_{a,c}D_{b,d}+D_{a,d}D_{b,c}].\end{aligned} \tag{3.21}$$

The pairs of terms in square brackets are invariant under interchange $a\leftrightarrow b$. Now use (3.8) and (3.10) to show that the first line of (3.21) is also invariant under $a\leftrightarrow b$:

$$\begin{aligned}&N_{a,b,c,d}+D_{a,b}N_{c,d}+D_{c,d}(N_{a,b}+D_{a,b})\\ &\quad=N_{b,a,c,d}+(D_{b,a}-D_{a,b})N_{c,d}+D_{a,b}N_{c,d}+D_{c,d}(N_{b,a}+D_{b,a})\\ &\quad=N_{b,a,c,d}+D_{b,a}N_{c,d}+D_{c,d}(N_{b,a}+D_{b,a}).\end{aligned} \tag{3.22}$$

Thus $S_{a,b,c,d}=S_{b,a,c,d}$. One can prove more generally that $S_{a,b,c,d}$ is invariant under the interchange of any two adjacent indices and hence under any of the 4! permutations of the indices. Thus from (3.18) and (3.21) it follows that

$$T_{1,2,3,4}=N_{1,2,3,4}+D_{1,2}N_{3,4}+D_{1,3}N_{2,4}+D_{1,4}N_{2,3}+D_{2,3}N_{1,4}+D_{2,4}N_{1,3}$$
$$+D_{3,4}N_{1,2}+D_{1,2}D_{3,4}+D_{1,3}D_{2,4}+D_{1,4}D_{2,3}. \tag{3.23}$$

Hence, both at the two- and four-point levels, the usual form of Wick's theorem holds even if the contraction and normal ordered product are not symmetric.

D. Proof for arbitrary *n*

At the *n*th order, the proof is a generalization of that given above for $n=2$. However, before proceeding, we use some of the expressions encountered in the previous subsection to motivate the introduction of two new pieces of notation.

The first enables us to write the terms appearing in square brackets on the rhs of (3.19) in a much more compact way. Recall that in this square bracket (which we now denote by Q), the i subscript on *each* normal ordered product and on each contraction is arranged in increasing order going from left to right. The simplest case is when $a_1=1$, $a_2=2$, $a_3=3$ and $a_4=4$. Then Q is given by

$$Q=D_{1,2}N_{3,4}+D_{1,3}N_{2,4}+D_{1,4}N_{2,3}+D_{2,3}N_{1,4}+D_{2,4}N_{1,3}+D_{3,4}N_{1,2}. \tag{3.24}$$

However, we can rewrite this as

$$Q=\sum_{\mathrm{perm}\{b\}\in B} D_{b_1,b_2}N_{b_3,b_4}, \tag{3.25}$$

where the sum over permutations of $\{b\}$ is *restricted* to include only those permutations which are an element of the set B defined by

$$B=\{\mathrm{perm}\{b\}:b_1<b_2,b_3<b_4\}. \tag{3.26}$$

[It is possible to check that (3.25) indeed reduces to (3.24).] If we now return to the case in which the a_i's are arbitrary, then Q is given by

$$Q=\sum_{\mathrm{perm}\{b\}\in B} D_{a_{b_1},a_{b_2}}N_{a_{b_3},a_{b_4}}, \tag{3.27}$$

where B is defined as in (3.26).

We now introduce the second piece of notation whose aim is to enable one to rewrite expressions such as (3.27) in a slightly simpler way (with fewer subscripts). Define

$$N(a_1,a_2,\ldots,a_{2n}):=N_{a_1,a_2,\ldots,a_{2n}},$$
$$D(a_1,a_2):=D_{a_1,a_2}. \tag{3.28}$$

As an example of the use of this notation, use (3.27) and (3.28) to rewrite (3.19) as

$$\sum_{\mathrm{perm}\{a\}} 6D(a_1,a_2)N(a_3,a_4)=\sum_{\mathrm{perm}\{a\}}\sum_{\mathrm{perm}\{b\}\in B} D(a_{b_1},a_{b_2})N(a_{b_3},a_{b_4}). \tag{3.29}$$

We now continue with the proof of Wick's theorem to *n*th order.

Step 1. For a given value of *m*, the rhs of (3.5) can be rewritten as a sum over distinct configurations of *N*'s and *D*'s, with the indices on each individual *N* and each individual *D* arranged in increasing order going from left to right [compare with (3.29)]. To do that, use Lemma 4.

Lemma 4: The following identity holds,

$$\frac{(2n)!}{(2m)!(n-m)!2^{n-m}}\sum_{\mathrm{perm}\{a\}} N(a_1,a_2,\ldots,a_{2m})\prod_{j=m+1}^{n} D(a_{2j-1},a_{2j})$$

$$=\sum_{\mathrm{perm}\{a\}}\sum_{\mathrm{perm}\{b\}\in B} N(a_{b_1},a_{b_2},\ldots,a_{b_{2m}})\prod_{j=m+1}^{n} D(a_{b_{2j-1}},a_{b_{2j}}), \tag{3.30}$$

where

$$\mathrm{perm}\{b\}=\begin{pmatrix}1 & 2 & 3 & \ldots & 2n\\ b_1 & b_2 & b_3 & \ldots & b_{2n}\end{pmatrix}, \tag{3.31}$$

and the restriction on the sum of permutations of $\{b\}$ *is such that (1) it arranges the indices on each individual normal ordered product and on each individual contraction in increasing order going from left to right, and (2) it orders the contractions by size of the leading element so as to avoid repetition of* $D_{1,2}D_{3,4}\cdots$ *as* $D_{3,4}D_{1,2}\cdots$:

$$B=\left\{\mathrm{perm}\{b\}:\begin{bmatrix} b_1<b_2<\cdots<b_{2m}\\ b_{2m+1}<b_{2m+2}\\ b_{2m+3}<b_{2m+4}\\ \ldots\\ b_{2n-1}<b_{2n}\end{bmatrix};b_{2m+1}<b_{2m+3}<\ldots<b_{2n-1}\right\}. \tag{3.32}$$

Proof: From Relation 3, for a product of $n-m$ contractions and one normal ordered product with a total of $2n$ indices, there are $(2n)!/(2m)!(n-m)!2^{n-m}$ distinct configurations with unordered indices. These distinct configurations are represented by the sum over permutations of the b labels. On a given contraction or normal ordered product, the order of the indices is arbitrary, but above the choice has been made by the restriction.

Step 2. Using Lemma 4, rewrite the rhs of (3.5) as

$$T_{1,2,\ldots,2n}=\frac{1}{(2n)!}\sum_{\mathrm{perm}\{a\}}\left[\sum_{m=0}^{n}\sum_{\mathrm{perm}\{b\}\in B} N(a_{b_1},a_{b_2},\ldots,a_{b_{2m}})\prod_{j=m+1}^{n} D(a_{b_{2j-1}},a_{b_{2j}})\right]$$

$$=:\frac{1}{(2n)!}\sum_{\mathrm{perm}\{a\}} S(\{a\}). \tag{3.33}$$

This is a generalization of (3.20) to arbitrary n.

Step 3. Use Lemmas 1 and 2 to show the following lemma.

Lemma 5: For any two permutations, $\{a\}$ *and* $\{a'\}$

$$S(\{a\})=S(\{a'\}). \tag{3.34}$$

Proof: Consider two permutations related by $\{a\}\to\{a'\}$ with

$$a_i=a'_{i+1},\quad a_{i+1}=a'_i,\quad a_j=a'_j\,\forall\, j\neq i,i+1. \tag{3.35}$$

The only nonzero contributions to $\Delta:=S(\{a\})-S(\{a'\})$ arise when *both* the indices a_i and a_{i+1} are *either* on the same normal ordered product, or on the *same* contraction. The reason is that when these indices occur on different normal ordered products or contractions, the restriction on the sum of permutations of $\{b\}$ orders the terms in exactly the same way in both the $\{a\}$ and $\{a'\}$ permutation thus making Δ vanish. [These terms are the generalization of those in square brackets in (3.21).]

Thus we may write

$$S(\{a\})-S(\{a'\})=\sum_{m=0}^{n}{\sum_{\mathrm{perm}\{b\}\in B}}'[N(a_{b_1},a_{b_2},\ldots,a_{b_{2m}})$$

$$-N(a'_{b_1},a'_{b_2},\ldots,a'_{b_{2m}})]\prod_{j=m+1}^{n}D(a_{b_{2j-1}},a_{b_{2j}}) \tag{3.36}$$

$$+\sum_{m=0}^{n}{\sum_{\mathrm{perm}\{b\}\in B}}''N(a_{b_1},a_{b_2},\ldots,a_{b_{2m}})[D(a_i,a_{i+1})$$

$$-D(a'_i,a'_{i+1})]\prod_{\substack{j=m+1\\ j\neq k}}^{n}D(a_{b_{2j-1}},a_{b_{2j}}). \tag{3.37}$$

Here the prime on the sum in (3.36) means that the sum over permutations of $\{b\}$ is further restricted by the condition that $(i,i+1)\in\{b_1,b_2,\ldots,b_{2m}\}$. (Thus it guarantees that the indices a_i and a_{i+1} only occur on the N and *not* on the D's.) In (3.37), the double prime means that this sum is further restricted to have $\{i,i+1\}=\{b_{2k-1},b_{2k}\}$ for some $k>m$, and this is the reason for which the contribution with $(j=k)$ is excluded in the product. (Hence the double prime on the sum ensures that the indices a_i and a_{i+1} both occur on a single D.)

Noting that when $m=0$ there are no contributions to (3.36) (since it involves no normal ordered products), and using Lemma 2, (3.36) may be rewritten as

$$\begin{aligned}
&\sum_{m=0}^{n}{\sum_{\mathrm{perm}\{b\}\in B}}'[N(a_{b_1},a_{b_2},\ldots,a_i,a_{i+1},\ldots,a_{b_{2m}})\\
&\quad-N(a'_{b_1},a'_{b_2},\ldots,a'_i,a'_{i+1},\ldots,a'_{b_{2m}})]\prod_{j=m+1}^{n}D(a_{b_{2j-1}},a_{b_{2j}})\\
&\quad=\sum_{m=1}^{n}\sum_{\mathrm{perm}\{c\}\in B}N(a_{c_1},a_{c_2},\ldots,a_{c_{2m-2}})[D(a_{i+1},a_i)-D(a_i,a_{i+1})]\prod_{j=m}^{n-1}D(a_{c_{2j-1}},a_{c_{2j}})\\
&\quad=\sum_{p=0}^{n-1}\sum_{\mathrm{perm}\{c\}\in B}N(a_{c_1},\ldots,a_{c_{2p}})[D(a_{i+1},a_i)-D(a_i,a_{i+1})]\prod_{j=p+1}^{n-1}D(a_{c_{2j-1}},a_{c_{2j}}),
\end{aligned} \tag{3.38}$$

where $p=m-1$, $c_j=b_j$ for $j<i$ and $c_j=b_{j+2}$ for $j\geqslant i$.

The second term of Eq. (3.37) vanishes for $m=n$ (as then it contains no contractions) and thus it can be rewritten as

$$\begin{aligned}
&\sum_{m=0}^{n}{\sum_{\mathrm{perm}\{b\}\in B}}''N(a_{b_1},a_{b_2},\ldots,a_{b_{2m}})[D(a_i,a_{i+1})-D(a'_i,a'_{i+1})]\prod_{\substack{j=m+1\\ j\neq k}}^{n}D(a_{b_{2j-1}},a_{b_{2j}})\\
&\quad=\sum_{m=0}^{n-1}{\sum_{\mathrm{perm}\{b\}\in B}}''N(a_{b_1},\ldots,a_{b_{2m}})[D(a_i,a_{i+1})-D(a_{i+1},a_i)]\prod_{\substack{j=m+1\\ j\neq k}}^{n}D(a_{b_{2j-1}},a_{b_{2j}}).
\end{aligned} \tag{3.39}$$

Recalling the definition of the c's, it is now possible to see that (3.38) and (3.39) just differ by a sign, and thus from (3.37) we have proved that $\Delta=S(\{a\})-S(\{a'\})=0$. Hence Δ vanishes for *any* two permutations.

Step 4. Since $S(\{a\})=S(\{a'\})$ for any two permutations $\{a\}$ and $\{a'\}$, then from (3.33) one has

$$T_{1,2,\ldots,2n}=\frac{1}{(2n)!}\sum_{\mathrm{perm}\{a\}} S(\{a\})=S(\{a\}), \tag{3.40}$$

where $S(\{a\})$ is just given by the usual form of Wick's theorem with the ordered indices. This finishes the proof that Wick's theorem holds in its usual form for nonsymmetric normal ordered products and contractions, *provided* the contraction is a *c*-number.

IV. CONCLUSIONS

In this paper we have worked with a general split of the field operators into two parts and the corresponding nonsymmetric normal ordered products and contractions. We have shown that, somewhat surprisingly, Wick's theorem still holds in its usual form for such nonsymmetric products as long as the contraction is a *c*-number. As a result, Wick's theorem—an important calculational tool in field theory—is more general than was originally thought.

We now comment on another interesting idea inspired by the discussion of Wick's theorem in the original paper on thermal field theory by Matsubara.[9] There, the definition of the contraction given for bosonic relativistic phonon field is *always symmetric* for all splits, and consistent with another definition of normal ordering $N^{(T)}[\bullet]$;

$$N^{(T)}[\psi_1\psi_2\ldots\psi_n]=N[T[\psi_1\psi_2\ldots\psi_n]]. \tag{4.1}$$

Note that $N^{(T)}[\bullet]$ is distinct from $N[\bullet]$, the definition of normal ordering which was given in Sec. I and was used throughout this paper: by definition $N^{(T)}$ is always symmetric in the sense of (1.4) because of the way it is built out of the time-ordered product [hence our designation with a (T) superscript[12]]. It therefore gives a corresponding symmetric contraction

$$D^{(T)}[\psi_1,\psi_2]=T_{12}-N^{(T)}_{12}=\theta(t_1-t_2)[\psi_1^+,\psi_2^-]_\sigma+\sigma\theta(t_2-t_1)[\psi_2^+,\psi_1^-]_\sigma. \tag{4.2}$$

It is clear that there are, in fact, many possible splits of the field which may be used to define normal ordering, and even then the definition of $N^{(T)}$ shows that one can think of other definitions of normal ordering different from the N given in Sec. I. Whichever definition is used, we have shown that even if the normal ordered product is not symmetric and the contraction is a *c*-number, Wick's theorem can be used so that calculations are not too cumbersome.

In practice the only definite cases we know of have symmetric normal ordered products. Symmetry breaking problems expressed in terms of Bogoliubov transformations[5,6] and equilibrium thermal field theory, whether using path-ordered methods[8] or thermo field dynamics,[13] all use our definition of N but have symmetric products. It is easy to see why symmetric products are so common; it is because the most useful normal ordered product is one which has zero expectation value. Thus taking the expectation value of Wick's theorem implies that the *c*-number contraction must then be symmetric (since it is then equal to the expectation value of the two-point time ordered function). In turn the operator version of Wick's theorem now shows that the normal ordered operator must itself be symmetric.

However, there is at least one situation where we might envisage the appearance of nonsymmetric products, and that is nonequilibrium field theory. In this case the environment is changing with time so we should expect the split, and hence the normal ordered products, also to change in time. It is then not at all obvious whether one can ever define normal ordered products of fields at different times which have zero physical expectation values. Thus, reversing the argument of the previous paragraph, it is not at all clear that symmetric products will simplify the calculations, and therefore it is reasonable that nonsymmetric products may prove to be a useful tool.

ACKNOWLEDGMENTS

We thank Brian Steer for many useful comments and for correcting a number of mistakes in the original manuscript. T.S.E. thanks the Royal Society for their support. D.A.S. is supported by P.P.A.R.C. through a research fellowship and is a member of Girton College, Cambridge. This work was supported in part by the European Commission under the Human Capital and Mobility program, Contract No. CHRX-CT94-0423.

[1] G. C. Wick, Phys. Rev. **80**, 268 (1950).
[2] M. Le Bellac, *Thermal Field Theory* (Cambridge U.P., Cambridge, 1996).
[3] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw–Hill, Singapore, 1985).
[4] J. W. Negele and H. Orland, *Quantum Many-Particle Systems* (Addison–Wesley, Redwood City, CA, 1988).
[5] A. A. Abrikosov, L. P. Gor'kov, and I. Ye. Dzyaloshinskii, *Methods of Quantum Field Theory in Statistical Physics* (Pergamon, Oxford, 1965).
[6] A. Fetter and J. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw–Hill, New York, 1971).
[7] J. L. Anderson, Phys. Rev. **94**, 703 (1954).
[8] T. S. Evans and D. A. Steer, Nucl. Phys. B **474**, 481 (1996).
[9] T. Matsubara, Prog. Theor. Phys. **14**, 4 (1955).
[10] M. Gaudin, Nucl. Phys. **15**, 89 (1960).
[11] N. P. Landsman and Ch. G. van Weert, Phys. Rep. **145**, 141 (1987).
[12] Only by using $N^{(T)}$ rather than our N for the normal ordering can the derivation of Wick's theorem given in the appendix of Matsubara (Ref. 9) be of use, as it is only given for a particular time ordering and requires symmetry arguments if one is to apply it for all possible time orderings. However, the definition of normal ordering for phonons given in the main text of Ref. 9 is, in fact, our $N[\bullet]$! Thus it is *not* consistent with the definition of the contraction in the main text, nor with the proof of Wick theorem proof in the appendix, whereas the $N^{(T)}$ definition (4.1) is consistent.
[13] Y. Takahshi and H. Umezawa, Collect. Phenom. **2**, 55 (1975); H. Umezawa, H. Matsumoto, and M. Tachiki, *Thermo Field Dynamics and Condensed States* (North Holland, Amsterdam, 1982); P. A. Henning, Phys. Rep. **253**, 235 (1995).

Algebraic description of S_N-invariant oscillator states

Roberto Floreanini
Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, Dipartimento di Fisica Teorica, Università di Trieste, Strada Costiera 11, 34014 Trieste, Italy

Luc Lapointe and Luc Vinet
Centre de recherches mathématiques, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada

(Received 12 June 1997; accepted for publication 10 April 1998)

A group-theoretical description of N identical harmonic oscillators on a line is presented. It provides a scheme for labeling the energy eigenstates that are invariant under the permutation group S_N and for obtaining the symmetric operators that transform these degenerate eigenfunctions among themselves. The symmetry algebra that these generators form is in general polynomial. The 2- and 3-particle cases are considered in detail. For the simple 2-body problem the invariance algebra is found to be the cubic SU(2) algebra: $[J_0,J_\pm]=\pm J_\pm$, $[J_+,J_-]=2J_0-\alpha J_0^3$. In the 3-body case, the permutational invariant states are characterized with the help of the subgroup chain $U(3)\supset U(2)\supset O(2)$. The labeling and step operators are obtained from determining an integrity basis for the S_3 scalar in $\mathcal{U}(U(3))$. Generating functions techniques are used to that end; an eight-dimensional basis is found whose elements span the symmetry algebra of the three identical oscillator problem. These constants of motion are seen to generate a nonlinear algebra whose representation on the symmetric states is provided. © *1998 American Institute of Physics.* [S0022-2488(98)02007-6]

I. INTRODUCTION

There is sustained interest for integrable N-particle systems, in particular because of the light they throw on quantum many-body physics. In such models, the statistical properties of the systems are accounted for by the transformation laws of the wave functions under permutations of the N-particle coordinates. In the case of bosons for instance, these functions should be invariant under all the elements of the symmetric group S_N of permutations of N objects.

Quite often, the Hamiltonians of such integrable models admit a fairly large invariance algebra. It is then natural to ask how can this dynamical algebra be used on the one hand, to obtain quantum numbers that characterize those states of the system behaving appropriately under S_N and, how can it be exploited on the other hand, to find operators that transform those states among themselves. This is the problem that we consider here in a most simple and fundamental context, namely, that of a system of N-identical Bose oscillators on the line. As is well known, the Hamiltonian for a collection of independent harmonic oscillators is invariant under SU(N). We shall only consider the case of bosonic quantum excitations. The physical wave functions of the system will therefore be those that are completely symmetric under the exchange of the N-particle coordinates. Now, it so happens that the generators of the SU(N) invariance group (except for one) do not transform the symmetric states onto themselves. These states thus cannot be adequately labeled by the eigenvalues of a subset of the SU(N) algebra elements, say the Cartan generators, as is customarily done to identify the basis vectors of SU(N) representation spaces. Here one turns instead to the enveloping algebra $\mathcal{U}(SU(N))$ of SU(N) to find, as we shall show, a complete set of commuting S_N-invariant operators whose eigenvalues provide the desired quantum numbers. We shall see in addition that $\mathcal{U}(SU(N))$ also contains S_N-invariant ladder operators that transform irreducibly among themselves the symmetric wave functions with common eigenenergies. These operators are moreover found to generate polynomial algebras; for instance, when $N=2$, they realize the following SU(2) polynomial algebra:[1–6]

$$[J_0,J_\pm]=\pm J_\pm, \quad [J_+,J_-]=2J_0-\alpha J_0^3. \tag{1.1}$$

 © 1998 American Institute of Physics

Owing to the ubiquity of harmonic oscillators in the description of physical phenomena, these basic results should prove useful.

Analyses bearing some connection with the one presented here have been carried out in the past. Some 30 years ago, the kinematics of nonrelativistic free particles was studied[7–9] in a fashion close to the approach used here for oscillators. The aim was to construct complete sets of observables that would lead to a classification of states where all the N particles are treated on an equal footing. The key observation there is that in d dimensions, after separation of the center of mass, the reduced Hamiltonian is invariant under $\mathrm{SO}(dN-d)$. This was then exploited to obtain labeling operators with definite transformation laws under S_N and such that their common eigenvectors enjoy simple symmetry properties under permutations. In the case of three particles, this has led[7] to a classification of states according to SU(3) which proved most useful to perform harmonic analyses on the Dalitz plot and to describe 3-body decays such as $K^+ \to \pi^+ \pi^+ \pi^-$. An important difference that distinguishes these studies from the one presented in this paper is that the dynamical group of a collection of oscillators is much larger than the kinematical group of a set of free particles.

Closer to our analysis are the extensive studies by nuclear theorists of the group theory of harmonic oscillators in the context of the shell, collective, and cluster model. In Ref. 10 for instance, the problem of constructing N-particle oscillator states with permutational symmetry is reviewed. The authors discuss the physical situation where the states of N three-dimensional oscillators are characterized by irreducible representations of $\mathrm{U}(3)\times\mathrm{U}(N)$. In the case where the particles interact via 2-body harmonic forces, they use the chain of groups $\mathrm{U}(N)\supset\mathrm{U}(N-1)\supset\mathrm{O}(N-1)$ to classify the states with permutational symmetry. While this chain will also arise in our work, the context is rather different: We do not consider the translational invariant system of N 3-D oscillators but instead, the one consisting of N 1-D particles in harmonic potentials. More importantly, the focus is not the same as we aim to find the ladder operators that transform the symmetric N-oscillator states among themselves in order to identify the dynamical algebra of these N-identical bosons on the line.

The paper is organized as follows. In Sec. II, we provide a general presentation of how the symmetric states of N-harmonic oscillators on the line can be described with the help of $\mathrm{SU}(N)$. Sections III and IV are, respectively, devoted to the specifics of the 2- and 3-particle special cases. The paper ends with a discussion of how the present work could be extended and generalized.

II. THE GENERAL SCHEME

A. *N*-harmonic oscillators on the line and SU(*N*)

We shall consider systems consisting of N-identical harmonic oscillators with coordinate x_i, $i=1,...,N$ in one dimension whose quantum Hamiltonians are given by

$$H^{(N)}=\frac{1}{2}\sum_{i=1}^{N}\left[-\frac{\partial^2}{\partial x_i^2}+x_i^2\right]. \tag{2.1}$$

Their masses and frequencies have been set to one (the superscript on H specify the number of oscillators). Upon introducing the creation and annihilation operators a_i^+ and a_i, $i=1,...,N$, respectively

$$a_i=\frac{1}{\sqrt{2}}\left(x_i+\frac{\partial}{\partial x_i}\right), \quad a_i^+=\frac{1}{\sqrt{2}}\left(x_i-\frac{\partial}{\partial x_i}\right). \tag{2.2}$$

$H^{(N)}$ is rewritten in the familiar quadratic form

$$H^{(N)}=\sum_{i=1}^{N} a_i^+ a_i+\frac{N}{2}. \tag{2.3}$$

The operators a_i and a_i^+, $i=1,...,N$, obey the commutation relations

$$[a_i,a_j^+]=\delta_{ij}, \quad [a_i,a_j]=[a_i^+,a_j^+]=0 \tag{2.4}$$

and verify

$$[H^{(N)},a_i]=-a_i, \quad [H^{(N)},a_i^+]=a_i^+ . \tag{2.5}$$

The invariance algebra of $H^{(N)}$ is readily identified as follows. Let

$$J_{ij}=a_i^+ a_j, \quad i,j=1,...,N. \tag{2.6}$$

Owing to the equations of motion (2.5), these N^2 operators are immediately seen to be constants of motion

$$[H^{(N)},J_{ij}]=0, \tag{2.7}$$

and they are found to realize the commutation relations of the U(N) algebra

$$[J_{ij},J_{kl}]=\delta_{jk}J_{il}-\delta_{il}J_{kj}. \tag{2.8}$$

Apart from the Hamiltonian $H^{(N)}=\Sigma_{i=1}^N J_{ii}+N/2$ which is central, we thus have N^2-1 independent operators generating a SU(N) invariance group. The eigenstates of $H^{(N)}$ are built through repeated action of the creation operators on the ground state wave function ϕ_0 which satisfy

$$a_i\phi_0=0; \quad i=1,...,N; \quad H^{(N)}\phi_0=\frac{N}{2}\,\phi_0. \tag{2.9}$$

They read

$$\phi(n_1,\ldots,n_N)=\prod_{i=1}^N \frac{(a_i^+)^{n_i}}{\sqrt{n_i!}}\,\phi_0 \tag{2.10}$$

and in view of (2.5) obey

$$H^{(N)}\phi(n_1,\ldots,n_N)=\left[|n|_N+\frac{N}{2}\right]\phi(n_1,\ldots,n_N) \tag{2.11}$$

with

$$|n|_N=n_1+\cdots+n_N. \tag{2.12}$$

Clearly, the U(N) generators transform the eigenstates with the same energy among themselves and thus account for the degeneracies:

$$J_{kl}\phi(n_1,\ldots,n_k,\ldots,n_l,\ldots,n_N)=\sqrt{(n_k+1)n_l}\,\phi(n_1,\ldots,n_k+1,...,n_l-1,...,n_N). \tag{2.13}$$

The irreducible representations of U(N) (like those of the symmetric group S_N) are labeled by partitions, that is sets $\lambda=(\lambda_1,\lambda_2,\ldots,)$ of ordered non-negative integers: $\lambda_1\geqslant\lambda_2\geqslant\cdots$. It is customary to represent partitions by Young diagrams which are of the form

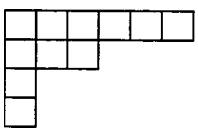

with λ_1 boxes in the first row, λ_2 boxes in the second, etc. The representations of U(N) that the oscillators realize are the fully symmetrized tensor products of the fundamental representation. These irreducible representations are associated to partitions $\lambda=(|n|_N)$ with only one part, or to diagrams of the form

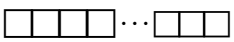 (2.14)

The vectors of the canonical bases for U(N) modules are eigenfunctions of the Casimir operators of U(N) and of all the Casimir operators of the following chain of subgroups:

$$\mathrm{U}(N)\supset \mathrm{U}(N-1)\supset\cdots\supset \mathrm{U}(1). \tag{2.15}$$

This ensures that the basis vectors also span representation spaces for all the subgroups in the chain. In the case of the oscillators, since the representation of U(N) is specified by only one number $|n|_N$, it is superfluous to diagonalize all Casimir operators. The reduction to subgroups is characterized by diagonalizing the Hamiltonians $H^{(N-1)},H^{(N-2)},\ldots$ which specify through the partitions $|n|_{N-1},|n|_{N-2},\ldots$, the irreducible representations of U($N-1$), U($N-2$),… that are contained in the original U(N) representation. Note that this is tantamount to giving the occupation numbers $n_1,n_2,\ldots,n_N$ since $n_1=|n|_1$, $n_2=|n|_2-|n|_1$, $n_3=|n|_3-|n|_2$, etc., and that this does therefore characterize the states (2.11).

It should be stressed immediately though, that the wave functions (2.10) are not appropriate for a system of N indistinguishable bosons since most of these do not have the required symmetry property under the exchange of the particle coordinates.

At times, we shall use the Bargmann–Fock representation in which the creation and annihilation operators act on holomorphic functions of the variables z_i, $i=1,...,N$ and are realized by

$$a_i=\frac{\partial}{\partial z_i},\quad a_i^+=z_i. \tag{2.16}$$

In this realization, $H^{(N)}$ takes the form of the Euler operator

$$H^{(N)}=\sum_{i=1}^{N} z_i\,\frac{\partial}{\partial z_i}+\frac{N}{2}, \tag{2.17}$$

its eigenvalues are monomials

$$\phi(n_1,\ldots,n_N)=\prod_{i=1}^{N}\frac{z_i^{n_i}}{\sqrt{n_i!}} \tag{2.18}$$

and the U(N) generators read $J_{ij}=z_i(\partial/\partial z_j)$.

B. S_N-invariant oscillator states

The symmetric group S_N acts in a natural way on the dynamical variables; under the transpositions $\sigma_{ij}=\sigma_{ji}$, $(\sigma_{ij})^2=1$, which generate S_N, a_i and a_i^+ transform as follows:

$$\sigma_{ij}a_i=a_j\sigma_{ij},\quad \sigma_{ij}a_i^+=a_j^+\sigma_{ij}. \tag{2.19}$$

In the framework of the bosonic realization of U(N), this entails a representation of S_N on this algebra. Clearly, the Hamiltonian is invariant under S_N: $\sigma_{ij}H=H\sigma_{ij}\forall i,j$, and hence, the Hilbert space of state vectors decomposes into irreducible S_N modules. We are interested in the eigensubspaces consisting of states that are totally symmetric under the exchanges of particle coordinates and have the same energy. We want to find the symmetry generators that transform these degenerate symmetric states among themselves and for that, wish to exploit the fact that SU(N) is the symmetry algebra of the Hamiltonian when no symmetry restrictions are imposed.

We first need to label the eigenstates conveniently. A standard characterization is achieved by arranging the occupation numbers $n_1,\ldots,n_N$ in a nondecreasing sequence $n_{\kappa_1}\geqslant n_{\kappa_2}\geqslant\cdots\geqslant n_{\kappa_N}$, that is, in a partition of $|n|_N=n_{\kappa_1}+\cdots+n_{\kappa_N}$. The partitions $n=(n_{\kappa_1},\ldots,n_{\kappa_N})$ of $|n|_N$ (or their associated diagrams) can then be used to enumerate the vectors of a basis for the symmetric N-oscillators states as follows:

$$\phi_n=\sum_{\substack{\text{distinct}\\ \text{permutations}}}\prod_{i=1}^{N}\frac{(a_i^+)^{n_{\kappa_i}}}{\sqrt{n_{\kappa_i}!}}\,\phi_0. \tag{2.20}$$

Let $I_l=\Sigma_i(J_{ii})^l$, $l=1,...,N$. The symmetric functions ϕ_n are characterized by the fact that they are the simultaneous eigenstates of these operators

$$I_l\phi_n=\left(\sum_i n_{\kappa_i}^l\right)\phi_n. \tag{2.21}$$

Natural as it might be, this basis does not, nevertheless, lend itself easily to the algebraic analysis that we want to perform since the transformation properties under U(N) of the vectors ϕ_n are not well known. In fact, representations of U(N) are typically constructed using basis states that are labeled by the eigenvalues of the Casimir operators associated with chains of continuous subgroups of U(N). The functions ϕ_n are not of that type. We shall therefore use another basis and a different set of quantum numbers for the S_N-invariant oscillator states.

What follows is a general description of how a basis which is convenient for our purposes can be characterized, and of how the corresponding ladder operators can be obtained. (The cases $N=2,3$ will be worked out in detail in Secs. III and IV, respectively.) We need a complete set of commuting symmetric operators to label the basis states. The first place to look is in the U(N) algebra itself. Usually, one diagonalizes the Cartan generators, this will not do here, however, since these operators are not symmetric. In fact, there are only two S_N-invariant operators in the bosonic realization of U(N): the Hamiltonian $H^{(N)}$ and

$$K=\sum_{\substack{i,j\\ i\neq j}} J_{ij}. \tag{2.22}$$

We shall demand that the basis states be eigenfunctions of this Hermitian operator. Given the symmetry requirements, this choice will pretty much determine the set of labeling operators. To supplement K with operators that commute with it, we consider the centralizer $C_K(\mathrm{U}(N))$ of K in U(N), in other words, the set of U(N) elements that commute with K. None of these operators are symmetric though (excluding of course H and K), so we turn to the enveloping algebra $\mathscr{U}(C_K(\mathrm{U}(N)))$ of C_K and look there for the required S_N-invariant elements (which are guaranteed to commute with K). The labeling generators that we will obtain in this way will provide quantum numbers related to those of familiar U(N) modules. As for the ladder operators, they will be taken among the S_N-invariant operators in $\mathscr{U}(\mathrm{U}(N))$ that do not lie in $\mathscr{U}(C_K)$.

In order to describe C_K, it is best to perform a change of variables so as to isolate the center of mass (CM) excitations from the relative ones. Let $\{O_{ij}\}$ be an $N\times N$ orthogonal matrix with the property

$$O_{Ni}=N^{-1/2},\quad i=1,...,N \tag{2.23}$$

implying that $\Sigma_{j=1}^N O_{ij}=0$ for $i\neq N$. We define the new variables as well as new creation and annihilation operators by

$$\hat{x}_i=\sum_j O_{ij}x_j,\quad \hat{a}_i=\sum_j O_{ij}a_j,\quad \hat{a}_i^+=\sum_j O_{ij}a_j^+. \tag{2.24}$$

Owing to (2.23)

$$\hat{a}_N=N^{-1/2}\sum_i a_i,\quad \hat{a}_N^+=N^{-1/2}\sum_i a_i^+, \tag{2.25}$$

$\hat{a}_N$ and $\hat{a}_N^+$ are the annihilation and the creation operators of the CM oscillator excitations. Obviously, the operators $\hat{J}_{ij}=\hat{a}_i^+\hat{a}_j$ obey the same commutation relations as the operators J_{ij} and thus generate also the U(N) algebra. The Jacobi variables are a special case of (2.23) with

$$O_{ji}=\begin{cases} -1/[j(j+1)]^{1/2} & i\leqslant j\\ j/[j(j+1)]^{1/2} & i=j+1\\ 0 & i>j+1\end{cases} \tag{2.26}$$

for $j=1,...,N-1$.

Let us clarify at this point the nature of the N-dimensional representation of S_N under which the variables a_i and a_i^+ $i=1,...,N$ transform when the particles are permuted. This representation is reducible since the CM operator $\hat{a}_N$ (or $\hat{a}_N^+$) is invariant and thus associated to the one-dimensional identity representation. It can be shown, however, that the remaining relative operators $\hat{a}_i$, $i=1,...,N-1$ (or $\hat{a}_i^+$) transform irreducibly under the $N-1$ dimensional representation

$$(N-1,1) \rightarrow \begin{array}{|c|c|c|c|c|}\hline \multicolumn{1}{|c|}{} & & & & \\ \hline & & & & \\ \cline{2-2}\end{array}\cdots \square \tag{2.27}$$

of S_N. This can be done, for instance, by computing explicitly the characters or comparing directly the transformation matrices with the Yamanouchi symbols[11] of the representation (2.27). Let us also remark that the matrices representing the permutations of the relative operators can be viewed as improper rotations and thus embedded in $O(N-1)$.

The generators of C_K are readily identified when K is expressed in terms of the operators $\hat{J}_{ij}$. First note that $H^{(N)}=\Sigma_{i=1}^N J_{ii}=\Sigma_{i=1}^N \hat{J}_{ii}=\hat{H}^{(N)}$. For K we have,

$$\begin{aligned} K&=\sum_{\substack{i,j\\ i\neq j}}\left(\sum_{k,l\neq N} O_{ki}O_{lj}\hat{J}_{kl}+\frac{1}{N}\hat{J}_{NN}\right)\\ &=-\sum_{k,l\neq N}\sum_{i=1}^{N} O_{ki}O_{li}\hat{J}_{kl}+(N-1)\hat{J}_{NN}=-\sum_{k=1}^{N-1}\hat{J}_{kk}+(N-1)\hat{J}_{NN} \end{aligned} \tag{2.28}$$

in other words,

$$K=-\hat{H}^{(N-1)}+(N-1)\hat{J}_{NN}. \tag{2.29}$$

Upon performing the change of coordinate (2.24)–(2.25) we see that K is a combination of the CM phonon number operator and of the $(N-1)$-dimensional Hamiltonian for the relative oscillators. From (2.29), the centralizer of K is immediately recognized to be

$$C_K \simeq U(N-1)\times U(1)_{CM} \tag{2.30}$$

with $U(1)_{CM}$ generated by $\hat{J}_{NN}$ and $U(N-1)$ by the symmetries $\hat{J}_{ij}$ $i,j=1,...,N-1$ of $\hat{H}^{(N-1)}$. Note first that we can equivalently diagonalize $\hat{H}^{(N-1)}$ and $\hat{J}_{NN}$ instead of $H^{(N)}$ and K. This is what we shall do in the following. In order to find the complementary labeling operators, we shall therefore be looking for S_N-invariant operators in $\mathcal{U}(U(N-1)\otimes U(1)_{CM})$. We already remarked that the action of S_N on the relative variables is by $O(N-1)$ matrices of S_N, hence the Casimir operators of the $O(N-1)$ subgroup of $U(N-1)$ will *a fortiori* be permutational invariant and admissible labeling operators. We are thus led naturally to the following chain of subgroups:

$$U(N)\supset U(N-1)\supset O(N-1)\supset\cdots. \tag{2.31}$$

Basis vectors for representations of $U(N)$ corresponding to the reduction (2.31) have been relatively well studied. They are eigenstates of various Casimir operators of all the groups in the chain. It is known, however, that in general this does not suffice because irreducible representations of $O(N-1)$ can occur with multiplicities bigger than one in irreducible representations of $U(N-1)$. One thus faces a notorious missing label problem. This degeneracy can be resolved in various ways. It will not be our concern here, however, since this problem does not arise when one deals with the $U(N-1)$ representations that the oscillators realize i.e., the completely symmetrized tensor products of the fundamental representation. In such cases the representations of $O(N-1)$ all have multiplicity 0 or 1.

Once a complete set of commuting symmetric operators has been found, it will thus be useful to start from the $U(N)$ representation space bases associated to the chain (2.31) in order to diagonalize the labeling operators. This done, there remains the task of finding among these joint

eigenfunctions those that are actually invariant under S_N. This requires restricting the range of the quantum numbers and taking appropriate linear combination. (This will be worked out in detail for $N=2$ and $N=3$ in the next two sections.)

C. Raising and lowering operators

The above group theoretical characterization of the symmetric wave functions sets up nicely the problem of determining the dynamical algebra of this system of N identical bosonic oscillators. To proceed we need the ladder operators that transform the symmetric degenerate eigenstates among themselves. As already mentioned, the symmetric polynomials involving the generators associated with the coset $\mathrm{U}(N)/\mathrm{U}(N-1)$ will do that job. A fundamental fact about the space of these invariant polynomials is that it is finitely generated. It will thus suffice to consider a minimal set whose elements generate all the symmetric ladder operators. Such a minimal set is called an integrity basis. At this point, the task that remains in order to specify the dynamical algebra, is to understand the structure of the algebra that the labeling operators form with the integrity basis of ladder operators. We would also want ideally, to construct the representations of this algebra. It is clear that it is of polynomial type: The commutators of the generators will be polynomial expressions in the algebra elements.

We have not found it possible to proceed from here in full generality, that is, for arbitrary N. On the one hand, the study of polynomial algebras has just begun and not much is known of them yet. On the other hand, although the dimension of the integrity basis for the S_N-invariant operators in $\mathrm{U}(N)$ is finite, its cardinality grows very quickly with N. In the following sections therefore we shall focus the discussion on the interesting $N=2$ and 3 cases. This will illustrate the procedure that one could follow to work out the cases involving larger numbers of particles.

An essential part of the program is to obtain an integrity basis for the set $\mathscr{I}$ of the polynomials in the generators of $\mathrm{U}(N)$ that are invariant under S_N. The elements of this basis that belong to $\mathrm{U}(N-1)$ are used as labeling operators and the others act as ladder operators. Let $p(\{J_{kl}\})$ denote a homogeneous polynomial in the N^2 $\mathrm{U}(N)$ generators J_{kl}, $k,l=1,...,N$. To be precise, a S_N-invariant polynomial is one which is such that $p(\{J_{kl}\})=p(\{\sigma_{ij}J_{kl}\sigma_{ij}\})$ for all transpositions σ_{ij}, $i,j=1,...,N$. The procedure for finding an integrity basis for these polynomial operators is the following.[12] One first assumes that the J_{kl} are ''classical'' quantities and obtain an integrity basis $I=p_1(\{J_{kl}\}),\ldots,p_q(\{J_{kl}\})$ in the variables J_{kl} momentarily taken to commute. Upon restoring the noncommutative character of the J_{kl}, one sees that the set $\{p_1,\ldots,p_q\}$ is not an integrity basis of $\mathscr{I}$ but rather that a subset of it is. This is because there may be algebraic relations between the polynomials $p_1,\ldots,p_q$ with operator arguments that do not exist when the variables are c-numbers. Indeed if p_1 and p_2 do not commute, then $p=[p_1,p_2]$ is also an invariant. This invariant p could belong to the integrity basis for functions of the commuting variables, if so, it should be discarded when the group generators are reinstated as noncommuting variables since $p=p_1p_2-p_2p_1$. (Classically of course this relation could not be obtained.) Therefore, in order to avoid this overcounting, one should form all the possible commutators $[p_s,p_t]$ out of the set $\{p_1,\ldots,p_q\}$ so as to determine a minimal set of independent p_κ.

In the situation of interest here there is another cause for possible overcounting which has to do with the specific bosonic realization of the $\mathrm{U}(N)$-algebra we are using, namely, $J_{ij}=a_i^+a_j$ and the fact that there are in this case quadratic relations of the form

$$J_{ij}J_{kl}=J_{il}J_{kj}-\delta_{kl}J_{ij}+\delta_{jk}J_{il} \tag{2.32}$$

between the generators. These could be responsible for the fact that two apparently distinct generators only differ by a lower degree invariant (already included in the basis).

D. Generating functions

When facing the problem of obtaining a minimal set $\{p_1,\ldots,p_q\}$ of invariants in the classical J_{kl} or an integrity basis for these, it is extremely useful to know how many basis elements there are and their degree. This information is provided by a generating function. (Much of what follows is taken from Ref. 13.)

Let Γ and Γ' be representations of a group G (for us G will be S_N). We shall call (Γ,Γ') tensors those tensors whose components transform according to the representation Γ and are polynomials in the components of another tensor transforming according to the representation Γ'. We know that the generators of U(N) transform according to the representation $\Gamma'=[(N-1,1)+(N)]\times[(N-1,1)+(N)]$ which admits the following Clebsch–Gordan series:

$$\Gamma'=2\Gamma_1+\Gamma_2+3\Gamma_3+\Gamma_4, \tag{2.33}$$

where

$$\Gamma_1=(N), \quad \Gamma_2=(N-2,1^2), \quad \Gamma_3=(N-1,1), \quad \Gamma_4=(N-2,2). \tag{2.34}$$

Our problem thus amounts to finding an integrity basis for the (Γ_1,Γ') tensor with Γ_1 the identity representation and Γ' given by (2.33). With Γ_m the irreducible representations of G, the generating function for the multiplicities of the (Γ_r,Γ') tensors of given degree is denoted by $B(\Gamma_r,\Gamma';\lambda)$. We shall first give formulas for $B(\Gamma_r,\Gamma_m;\lambda)$ when both representations Γ_r and Γ_m are irreducible. We shall then provide a composition rule that will allow us to find $B(\Gamma_r,\Gamma_m+\Gamma_{m'};\lambda,\lambda')$ from $B(\Gamma_r,\Gamma_m;\lambda)$ and $B(\Gamma_r,\Gamma_{m'};\lambda')$. A general expression for $B(\Gamma_r,\Gamma_m;\lambda)$ is

$$B(\Gamma_r,\Gamma_m;\lambda)=\frac{1}{N}\sum_s \frac{N_s\chi^*_{sr}}{\det(1-\lambda A_s)}, \tag{2.35}$$

where the sum runs over the classes s of G, N is the order of the group G, N_s is the order of the class s of G, A_s the matrix in the representation Γ_m of an element of G belonging to the class s and χ^*_{sr} the complex conjugate of the character χ_{sr} of s in Γ_r. Upon expanding the right-hand side of (2.35) as a power series in λ

$$B(\Gamma_r,\Gamma_m;\lambda)=\sum_{n=0}^{\infty} c_n\lambda^n, \tag{2.36}$$

the coefficients c_n are seen to count the number of times the irreducible representation Γ_r occurs in the symmetrized product $\{\Gamma_m\}$.

The generating functions for reducible representations can be derived with the help of the following formula:

$$B(\Gamma_r,\Gamma_{m_1}+\Gamma_{m_2};\lambda_1,\lambda_2)=\sum_{r_1,r_2} B(\Gamma_{r_1},\Gamma_{m_1};\lambda_1)B(\Gamma_{r_2},\Gamma_{m_2};\lambda_2)C^r_{r_1r_2}, \tag{2.37}$$

where the constants $C^r_{r_1r_2}$ are defined by the Clebsch–Gordan series

$$\Gamma_{r_1}\times\Gamma_{r_2}=\sum C^r_{r_1r_2}\Gamma_r. \tag{2.38}$$

The coefficient of $\lambda_1^{n_1}\lambda_2^{n_2}$ in the expansion of $B(\Gamma_r,\Gamma_{m_1}+\Gamma_{m_2};\lambda_1,\lambda_2)$ is the number of linearly independent $(\Gamma_r,\Gamma_{m_1}+\Gamma_{m_2})$ tensors of degree n_1 and n_2, respectively, in the components of two tensors transforming, respectively, according to Γ_{m_1} and Γ_{m_2}. When $\Gamma_r=\Gamma_1$, the identity representation, (2.37) reduces to

$$B(\Gamma_1,\Gamma_{m_1}+\Gamma_{m_2};\lambda_1,\lambda_2)=\sum_r B(\Gamma_r,\Gamma_{m_1};\lambda_1)B(\Gamma^*_r,\Gamma_{m_2};\lambda_2), \tag{2.39}$$

where Γ^*_r denotes the representation which is the complex conjugate of Γ_r. One obtains the generating function $B(\Gamma_r,\Gamma';\lambda_1,\lambda_2,\ldots)$ for any reducible representation Γ' by iterating formula (2.38).

The generating function $B(\Gamma_r,\Gamma';\lambda)$ contains more than just the multiplicities of the (Γ_r,Γ') tensors of given degree, it in fact allows one to find the dimensions of the integrity basis for the (Γ_r,Γ') tensors. To see this, the finite sum in (2.35) has to be performed and the result expressed as a ratio of two polynomials in λ

$$B(\Gamma_r,\Gamma_m;\lambda)=\frac{N_{rm}(\lambda)}{D_m(\lambda)}. \tag{2.40}$$

It turns out that

$$N_{rm}(\lambda)=\sum_{\alpha} k_\alpha \lambda^\alpha, \quad D_m(\lambda)=\prod_{\beta}(1-\lambda^\beta), \tag{2.41}$$

where the coefficients k_α are positive integers and the sum and product taken over finite sets of integers $\alpha_1,\alpha_2,\ldots,$ and $\beta_1,\beta_2,\ldots,$ respectively. The denominator $D_m(\lambda)$ is common to all generating functions with the same Γ_m and G. In view of (2.38), the generating function $B(\Gamma_r,\Gamma';\lambda)$ associated with the reducible representation $\Gamma'=\Sigma\Gamma_i$ will have a similar form

$$B(\Gamma_r,\Gamma';\{\lambda_i\})=\frac{N_r'(\lambda_i)}{D'(\lambda_i)} \tag{2.42}$$

with

$$N_r'(\lambda_i)=\sum_{\{\alpha_i\}} k_{\{\alpha_i\}}\prod_i \lambda_i^{\alpha_i}, \quad D'(\lambda_i)=\prod_i\prod_{\beta_i}(1-\lambda_i^{\beta_i}). \tag{2.43}$$

The number of elements in an integrity basis for the invariants or the (Γ_1,Γ') tensors can now be determined as follows. To each factor $(1-\lambda_i^{\beta_i})$ of $D'(\lambda_i)$ corresponds one invariant of degree β_i in the components of the Γ_i tensor. To each term $k_{\{\alpha_i\}}\Pi_i\lambda_i^{\alpha_i}$ of $N_1'(\lambda_i)$ corresponds $k_{\{\alpha_i\}}$ independent invariants of degree α_1 in the components of the Γ_1 tensor, α_2 in the components of the Γ_2 tensor and so on, as long as the monomials $\Pi_i\,\lambda_i^{\alpha_i}$ cannot be formed (in equal number) out of the monomials $\Pi_i\,\lambda_i^{\beta_i}$ and of monomials $\Pi_i\,\lambda_i^{\alpha_i'}$ of lower degree already considered.

The generating function $B(\Gamma_1,\Gamma';\lambda_1,\ldots,\lambda_7)$ with Γ' as in (2.33)–(2.34) thus provides the number of elements in a minimal set of invariants in the classical J_{kl}; it also gives the degree of the elements in the set. To construct all these invariants one first assumes that they are arbitrary polynomials of the given degree in various irreducible components of the tensor J_{kl}, then impose that they remain unchanged when the components on which these polynomials depend are transformed according to their respective representation. After that, one should not forget to model out the realization dependent relations (2.32) and to compute the commutators between these invariants in order to eliminate them from the set the symmetric polynomials that are no longer independent from the others when the functions are made to depend on the U(N) generators. A concrete illustration of how all this works will be found in Sec. IV.

III. THE $N=2$ CASE

We shall now apply the program that we described in the previous section to the simplest possible situation: the $N=2$ case. We shall find, remarkably, that the invariance algebra of two identical bosonic harmonic oscillators on the line is the cubic deformation of SU(2) whose defining commutation relations were given in (1.1). We shall show that the symmetric eigenstates of the 2-oscillator system belong, depending on the energy, to irreducible modules of this algebra that are of two types and shall thus explain the spectrum degeneracies.

A. The symmetry algebra

In accordance with the general discussion, we shall characterize the states of the system as the joint eigenfunctions of the Hamiltonian $H^{(2)}=a_1^+a_1+a_2^+a_2$ and the symmetric observable K

$=J_{12}+J_{21}$. [The reader will recall that the operators $J_{ij}=a_i^+ a_j$, $i,j=1,2$, generate the U(2) invariance algebra of $H^{(2)}$.] In order to diagonalize these operators simultaneously, it is best to pass to the relative and center of mass coordinates

$$\hat{x}_i=\sum_{j=1}^{2} M_{ij}x_j, \quad M=\frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{3.1}$$

and to use correspondingly, the creation and annihilation operators

$$\hat{a}_i=\sum_{j=1}^{2} M_{ij}a_j, \quad \hat{a}_i^+=\sum_{j=1}^{2} M_{ij}a_j^+ . \tag{3.2}$$

In terms of these, $H^{(2)}$ and K read

$$H^{(2)}=\hat{a}_1^+\hat{a}_1+\hat{a}_2^+\hat{a}_2, \quad K=-\hat{a}_1^+\hat{a}_1+\hat{a}_2^+\hat{a}_2. \tag{3.3}$$

It is then manifest that they are diagonal on the states

$$|n_1,n_2\rangle=\frac{(\hat{a}_1^+)^{n_1}}{\sqrt{n_1!}}\frac{(\hat{a}_2^+)^{n_2}}{\sqrt{n_2!}}|0\rangle \quad n_1,n_2\in\mathbb{N}, \tag{3.4}$$

and in fact

$$H^{(2)}|n_1,n_2\rangle=(n_1+n_2)|n_1,n_2\rangle, \quad K|n_1,n_2\rangle=(n_2-n_1)|n_1,n_2\rangle. \tag{3.5}$$

Of these eigenstates we want to retain only those that are invariant under the exchange of the oscillators. The group S_2 is generated by the transposition $\sigma_{12}\equiv\sigma$. Under this operation, the creation and annihilation operators transform as follows:

$$\begin{aligned} &\sigma\hat{a}_1\sigma=-\hat{a}_1, \quad \sigma\hat{a}_2\sigma=\hat{a}_2, \\ &\sigma\hat{a}_1^+\sigma=-\hat{a}_1^+, \quad \sigma\hat{a}_2^+\sigma=\hat{a}_2^+ . \end{aligned} \tag{3.6}$$

The ground state moreover, is symmetric $\sigma|0\rangle=|0\rangle$. Hence the S_2-invariant states are $|2n_1,n_2\rangle$, that is, the states (3.4) with n_1 even. The energy spectrum is thus

$$E_{n_1,n_2}=2n_1+n_2\equiv n \tag{3.7}$$

and shows at level n a degeneracy of degree $[n/2]+1$ with $[n/2]$ standing for the integer part of $n/2$.

In this simple case and in view of (3.6), an integrity basis for the S_2 scalars in $\mathscr{U}(\mathrm{U}(2))$ is readily found by inspection. One indeed checks that the set $\{H^{(2)},K,(\hat{J}_{12})^2,(\hat{J}_{21})^2\}$ forms such a basis. The elements $(\hat{J}_{12})^2$ and $(\hat{J}_{21})^2$ must thus be the ladder operators for the eigenvalues of K and a short computation confirms that this is indeed the case

$$[K,(\hat{J}_{12})^2]=8(\hat{J}_{12})^2, \quad [K,(\hat{J}_{21})^2]=-8(\hat{J}_{21})^2. \tag{3.8}$$

The question now is to identify the algebra that the constants K, $(\hat{J}_{12})^2$, and $(\hat{J}_{21})^2$ realize. In this connection note that the commutator of $(\hat{J}_{12})^2$ and $(\hat{J}_{21})^2$ is a cubic expression in $H^{(2)}$ and K

$$[(\hat{J}_{12})^2,(\hat{J}_{21})^2]=K^3-K[(H^{(2)})^2+2H^{(2)}-2]. \tag{3.9}$$

Let

$$\alpha=\frac{32}{[(H^{(2)})^2+2H^{(2)}-2]} \tag{3.10}$$

and redefine the symmetry generators as follows:

$$J_0 = -\tfrac{1}{4}K = \tfrac{1}{4}(\hat{a}_1^+\hat{a}_1 - \hat{a}_2^+\hat{a}_2),$$

$$J_+ = \frac{\sqrt{\alpha}}{8}(\hat{J}_{12})^2 = \frac{\sqrt{\alpha}}{8}(\hat{a}_1^+)^2(\hat{a}_2)^2, \tag{3.11}$$

$$J_- = \frac{\sqrt{\alpha}}{8}(\hat{J}_{21})^2 = \frac{\sqrt{\alpha}}{8}(\hat{a}_2^+)^2(\hat{a}_1)^2 = (J_+)^\dagger.$$

It is straightforward to verify that these operators satisfy the commutation relations

$$[J_0, J_\pm] = \pm J_\pm, \tag{3.12a}$$

$$[J_+, J_-] = 2J_0 - \alpha J_0^3 \equiv f(J_0), \tag{3.12b}$$

and thus generate a cubic deformation of the SU(2) Lie algebra if we treat α as a constant which can always be done on energy eigensubspaces. This generalized algebra is the symmetry algebra of the two identical oscillator systems. It remains to determine the irreducible representations according to which the degenerate eigenstates of $H^{(2)}$ transform.

B. Physical representations

Since there are finitely many degenerate levels, we are interested in highest-weight irreducible representations of dimension $d+1<\infty$. Let m be the eigenvalues of J_0 and j denote the highest value that m takes in one such representation. We shall use these numbers j and m to label the basis states. In view of (3.12a), we can posit that J_0, J_+ and J_- act as follows on the vectors $|j,m\rangle$:

$$J_0|j,m\rangle = m|j,m\rangle, \tag{3.13a}$$

$$J_+|j,m\rangle = c^*(j,m+1)|j,m+1\rangle, \tag{3.13b}$$

$$J_-|j,m\rangle = c(j,m)|j,m-1\rangle, \tag{3.13c}$$

if we request that the representation be unitary, in other words that $J_+ = (J_-)^\dagger$. Here $c^*(j,m+1)$ is the complex conjugate of the coefficient $c(j,m+1)$. From the definition of j, we must have

$$c(j,j+1) = 0. \tag{3.14}$$

We now define the function $h(J_0)$ through [see (3.12b)]

$$f(J_0) = h(J_0) - h(J_0 - 1). \tag{3.15}$$

From (3.12b) we have

$$[J_+, J_-]|j,m\rangle = f(J_0)|j,m\rangle \tag{3.16}$$

which gives, using (3.13) and (3.15),

$$|c(j,m)|^2 - |c(j,m+1)|^2 = h(m) - h(m-1). \tag{3.17}$$

With the help of (3.14), it is easily seen that this entails the relation

$$|c(j,m)|^2 = h(j) - h(m-1). \tag{3.18}$$

Without loss of generality, the phases of the vectors $|j,m\rangle$ can be chosen so that $c(j,m)=[h(j)-h(m-1)]^{1/2}$ and these coefficients therefore are real and non-negative for all m. As a matter of fact, for this representation to be $(d+1)$-dimensional, we must also have

$$c(j,j-d)=0, \tag{3.19}$$

implying that m takes the values $j,j-1,...,j-d+1$. The Casimir operator of the algebra (3.12) is given by

$$C=J_-J_++h(J_0). \tag{3.20}$$

It is readily checked that this element is central. It is found moreover from (3.18) that the basis vectors $|j,m\rangle$ are all eigenvectors of C with eigenvalues $h(j)$. This confirms that j properly labels the irreducible representations of this polynomial algebra.

We now want to determine which of these modules are realized by the symmetric oscillator states $\{|2n_1,n_2\rangle;n_1,n_2\in\mathbb{N}\}$. In this context, the irreducible subspaces are identified by the total energy, in other words by the integers

$$n=2n_1+n_2. \tag{3.21}$$

To identify the irreducible representations of the symmetry algebra, one needs to obtain the relation between n and the highest eigenvalue j of J_0 on the states $|2n_1,n_2\rangle$. This is easily achieved. From $J_0|2n_1,n_2\rangle=\frac{1}{4}(2n_1-n_2)|2n_1,n_2\rangle$ we see that

$$m=\tfrac{1}{4}(2n_1-n_2)=\tfrac{1}{4}(n-2n_2). \tag{3.22}$$

From the last equality we see that for a given n, the maximum value j of m is attained when n_2 takes its minimum value. Since n_1 and n_2 are non-negative integers, this minimum value is 0 when n is even and 1 when n is odd. Thus

$$j(n)=\begin{cases} n/4 & \text{for } n \text{ even} \\ (n-2)/4 & \text{for } n \text{ odd.} \end{cases} \tag{3.23}$$

These values of j can also be obtained by looking at the action of $J_-=\sqrt{\alpha}/8(\hat{a}_2^+)^2(\hat{a}_1)^2$ on the oscillator states

$$|2n_1,n_2\rangle=|j(n),m\rangle=\frac{(\hat{a}_1^+)^{(n/2)+2m}(\hat{a}_2^+)^{(n/2)-2m}}{\sqrt{((n/2)+2m)!((n/2)-2m)!}}\,|0\rangle. \tag{3.24}$$

An easy computation shows that

$$J_-|j(n),m\rangle=\frac{\sqrt{\alpha}}{8}\left\{\left(\frac{n}{2}-2m+1\right)\left(\frac{n}{2}-2m+2\right)\left(\frac{n}{2}+2m\right)\left(\frac{n}{2}+2m-1\right)\right\}|j(n),m-1\rangle. \tag{3.25}$$

From the representation theory we know [see (3.13c)] that $J_-|j,m\rangle=c(j,m)|j,m-1\rangle$, where [see (3.18)] $c(j,m)=[h(j)-h(m-1)]^{1/2}$. In the specific case at hand, $f(J_0)=2J_0-\alpha J_0^3$ [see (3.12b)], and the function $h(J_0)$ which solves (3.15) is found to be

$$h(J_0)=J_0(J_0+1)\left[1-\frac{\alpha}{4}J_0(J_0+1)\right]. \tag{3.26}$$

Combining (3.18), (3.24), and (3.26), one thus gets the following relation:

$$\frac{\alpha}{64}\left(\frac{n}{2}-2m+1\right)\left(\frac{n}{2}-2m+2\right)\left(\frac{n}{2}+2m\right)\left(\frac{n}{2}+2m-1\right)$$
$$=j(j+1)\left[1-\frac{\alpha}{4}j(j+1)\right]-m(m-1)\left[1-\frac{\alpha}{4}m(m-1)\right], \tag{3.27}$$

which connects n and j. Of course, when $m=j+1$, the right-hand side of (3.27) vanishes and for the equality to be valid j must take either one of the values given in (3.23), that is $n/4$ or $(n-2)/4$. Upon replacing α by $32/(n^2+2n-2)$ according to (3.10), one may check that (3.27) is satisfied for arbitrary m provided that j is as specified above.

We may now determine the dimensions $(=d+1)$ of the representations from the condition $J_-|j(n),j(n)-d\rangle=0$. This equation is immediately found from (3.25) to possess the two solutions $d=n/2$ and $d=(n-1)/2$ for each of the two values of j. Of these solutions, in each case, only one will be admissible since d must be an integer. The situation will be the following:

$$n \text{ even: } j=n/4, \quad d+1=2j+1=\frac{n}{2}+1,$$
$$n \text{ odd: } j=(n-2)/4, \quad d+1=2j+\frac{3}{2}=\frac{1}{2}(n-1)+1. \tag{3.28}$$

In summary we found that the degenerate symmetric energy eigenstates of the 2-oscillator system fall into irreducible representations of the cubic deformation (3.12) of SU(2) that are of two types. The dimensions of these representations are $[n/2]+1$ and we have thus algebraically explained the observed degeneracies of the spectrum.

It is instructive to write down explicitely a few wave functions $\phi_{n,m}(z_1,z_2)=\langle z_1,z_2|n,m\rangle$ in the holomorphic realization where

$$\hat{a}_i=\frac{\partial}{\partial z_i}, \quad \hat{a}_i^+=z_i, \quad i=1,2. \tag{3.29}$$

For the two-dimensional representation we have

$$\phi_{2,-1/2}=\frac{z_2^2}{\sqrt{2}}, \quad \phi_{2,1/2}=\frac{z_1^2}{\sqrt{2}},$$
$$\phi_{3,-3/4}=\frac{z_2^3}{\sqrt{3!}}, \quad \phi_{3,1/4}=\frac{z_1^2 z_2}{\sqrt{2}}. \tag{3.30}$$

In this realization the symmetry algebra generators take the form

$$J_0=\frac{1}{4}\left(z_1\frac{\partial}{\partial z_1}-z_2\frac{\partial}{\partial z_2}\right),$$
$$J_+=\frac{\sqrt{\alpha}}{8}z_1^2\frac{\partial^2}{\partial z_2^2}, \tag{3.31}$$
$$J_-=\frac{\sqrt{\alpha}}{8}z_2^2\frac{\partial^2}{\partial z_1^2}.$$

These operators are easily seen to transform the wave functions (3.21) with the same n among themselves and according to (3.13). Note that since α depends on n, the normalizations of J_+ and J_- change in each irreducible representation. Since the dimension of the representations is given by $[n/2]+1$, it increases by one with every even number. The three-dimensional representations, for instance, are obtained when $n=4$ and $n=5$. The corresponding normalized basis functions are

$$\phi_{4,-1}=\frac{z_2^4}{\sqrt{4!}},\quad \phi_{4,0}=\frac{z_1^2z_2^2}{2},\quad \phi_{4,1}=\frac{z_1^4}{2\sqrt{3!}},$$
$$\phi_{5,-5/4}=\frac{z_2^5}{\sqrt{5!}},\quad \phi_{5,-1/4}=\frac{z_1^2z_2^3}{2\sqrt{3}},\quad \phi_{5,3/4}=\frac{z_1^4z_2}{2\sqrt{3!}}, \tag{3.32}$$

and one may again check that the action of the operators (3.22) on these functions is in agreement with the formulas obtained from the abstract representation theory.

IV. THE $N=3$ CASE

We now present the $N=3$ case. Although the 2-particle problem proved surprisingly interesting, it remains somewhat special. The system consisting of three identical oscillators on the line is more generic and its analysis will require all the tools introduced in Sec. II.

We are dealing here with the Hamiltonian $H^{(3)}=\Sigma_{i=1}^3 a_i^+ a_i$ whose U(3) invariance algebra is generated by the operators $J_{ij}=a_i^+ a_j$, $i,j=1,2,3$. Our goal is to characterize the symmetric eigenstates of $H^{(3)}$ and to find the operators that will transform these degenerate states among themselves. In order to accomplish the first part of this task, we shall (i) obtain a complete set of commuting observables that are symmetric, (ii) diagonalize these operators simultaneously, and (iii) impose that the joint eigenfunctions be invariant under permutations. As for the ladder operators, we shall compute the generating function from which they can be enumerated, provide their action on the basis vectors, and give the commutation relations for a minimal generating set.

A. The symmetric 3-oscillator states

Following the plan of Sec. II, we shall take

$$K=J_{12}+J_{13}+J_{21}+J_{23}+J_{31}+J_{32} \tag{4.1}$$

as one of the labeling operators. As already explained, it will also be best to use relative and CM coordinates and we shall choose those of Jacobi

$$\hat{x}_1=\frac{1}{\sqrt{2}}(x_1-x_2),$$
$$\hat{x}_2=-\frac{1}{\sqrt{6}}(x_1+x_2-2x_3), \tag{4.2}$$
$$\hat{x}_3=\frac{1}{\sqrt{3}}(x_1+x_2+x_3).$$

The corresponding annihilation and creation operators are $\hat{a}_i=\Sigma_{j=1}^3 O_{ij}a_j$, $\hat{a}_i^+=\Sigma_{j=1}^3 O_{ij}a_j^+$ with the orthogonal matrix O given by (2.23) and (2.26) for $i,j=1,2,3$. The operators $\hat{J}_{ij}=\hat{a}_i^+\hat{a}_j=\Sigma_{k,l=1}^3 O_{ik}O_{jl}J_{kl}$ form an equivalent basis for the symmetry algebra of $H^{(3)}=\hat{H}^{(3)}$ whose eigenvalues we shall denote by $|n|_3$. Moreover, K reads

$$K=-\hat{H}^{(2)}+2\hat{J}_{33}+1 \tag{4.3}$$

with $\hat{H}^{(2)}=\hat{a}_1^+\hat{a}_1+\hat{a}_2^+\hat{a}_2+1$. Instead of using the eigenvalues of $H^{(3)}=\hat{H}^{(2)}+\hat{J}_{33}+1/2$ and K as labels for the state vectors, we shall use equivalently those of $\hat{H}^{(2)}$ and $\hat{J}_{33}$ to which a third quantum number has to be added. The operators that commute with K and $H^{(3)}$ (or $\hat{H}^{(2)}$ and $\hat{J}_{33}$) are $\hat{J}_{ij}$ $i,j=1,2$; they generate the U(2) invariance algebra of the two-dimensional oscillator in the relative coordinates with Hamiltonian $\hat{H}^{(2)}$. In terms of

$$\hat{J}_x=-\frac{1}{2}(\hat{J}_{12}+\hat{J}_{21}),\quad \hat{J}_y=-\frac{1}{2i}(\hat{J}_{12}-\hat{J}_{21}),\quad \hat{J}_z=-\frac{1}{2}(\hat{J}_{11}-\hat{J}_{22}), \tag{4.4}$$

we have $[\hat{J}_x,\hat{J}_y]=i\hat{J}_z$ and the other cyclic combinations. Note also that since

$$\hat{J}^2=\hat{J}_x^2+\hat{J}_y^2+\hat{J}_z^2=\left(\frac{\hat{H}^{(2)}-1}{2}\right)\left(\frac{\hat{H}^{(2)}-1}{2}+1\right), \tag{4.5}$$

the eigenvalues of $\hat{H}^{(2)}$ could also be replaced by those of $\hat{J}^2$. In addition to commuting with $\hat{H}^{(2)}$ and $\hat{J}_{33}$, the additional labeling operator we are looking for must have the property of being invariant under the permutations of the three oscillators. We must therefore find a polynomial in $\hat{J}_x$, $\hat{J}_y$, and $\hat{J}_z$ with this feature. To that end, we shall now examine how the symmetry generators $\hat{J}_{ij}$ $i,j=1,2,3$ transform under S_3.

The symmetric group S_3 is generated by the transposition σ_{12} and the cyclic permutation σ_{123} whose action on the coordinates x_i, $i=1,2,3$ are

$$\begin{aligned}&\sigma_{12}: \quad x_1\to x_2 \quad x_2\to x_1 \quad x_3\to x_3,\\ &\sigma_{123}: \quad x_1\to x_2 \quad x_2\to x_3 \quad x_3\to x_1.\end{aligned} \tag{4.6}$$

The irreducible content of the corresponding 3-dimensional representation of S_3 is

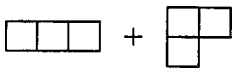

and the decomposition effected by passing to the Jacobi coordinates:

$$\sigma_{12}: \quad \hat{x}_3\to\hat{x}_3, \quad \sigma_{123}: \ \hat{x}_3\to\hat{x}_3, \tag{4.7}$$

$$\sigma_{12}: \ \hat{x}_i\to\hat{x}_i^I=\sum_{j=1}^{2} I_{ij}\hat{x}_j, \quad \sigma_{123}: \ \hat{x}_i\to\hat{x}_i^R=\sum_{j=1}^{2} R_{ij}\hat{x}_j, \tag{4.8}$$

where

$$I=\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}, \quad R=\begin{pmatrix}-1/2 & \sqrt{3}/2\\ -\sqrt{3}/2 & -1/2\end{pmatrix}. \tag{4.9}$$

The operators $\hat{J}_{ij}$ behave as a second-rank tensor and their transformation properties under σ_{12} and σ_{123} are easily obtained. We introduce the following basis for U(3):

$$\{\hat{H}^{(2)}, \quad \hat{J}_{33}, \quad \hat{J}_y, \quad \hat{J}_\pm=\hat{J}_x\pm i\hat{J}_z, \quad \hat{J}_{\pm 3}=\hat{J}_{13}\pm i\hat{J}_{23}, \quad \hat{J}_{3\pm}=\hat{J}_{31}\pm i\hat{J}_{32}\}. \tag{4.10}$$

Its elements are seen to transform as follows:

$$\begin{array}{lll}\sigma_{12}: & \hat{H}^{(2)}\to\hat{H}^{(2)} & \sigma_{123}: \quad \check{H}^{(2)}\to\check{H}^{(2)}\\ & \hat{J}_{33}\to\hat{J}_{33} & \hat{J}_{33}\to\hat{J}_{33}\\ & \hat{J}_y\to-\hat{J}_y & \hat{J}_y\to\hat{J}_y\\ & \hat{J}_\pm\to-\hat{J}_\mp & \hat{J}_\pm\to e^{\pm i4\pi/3}\hat{J}_\pm\\ & \hat{J}_{\pm 3}\to-\hat{J}_{\mp 3} & \hat{J}_{\pm 3}\to e^{\pm i4\pi/3}\hat{J}_{\pm 3}\\ & \hat{J}_{3\pm}\to-\hat{J}_{3\mp} & \hat{J}_{3\pm}\to e^{i4\pi/3}\hat{J}_{3\pm}.\end{array} \tag{4.11}$$

In order to obtain an extra labeling operator, we set out to find an S_3-invariant polynomial in $\hat{J}_x$, $\hat{J}_y$, and $\hat{J}_z$. From (4.11), it is readily seen that the lowest order invariant is $\hat{J}_y^2$. The operator $\hat{J}_+\hat{J}_-+\hat{J}_-\hat{J}_+$ is also invariant but in view of (4.5) can be expressed in terms of $\hat{H}^{(2)}$ and $\hat{J}_y^2$. The operators $\hat{H}^{(2)}$, $\hat{J}_y^2$, and $\hat{J}_{33}$ will thus form a natural set of commuting and symmetric observables for labeling the permutational invariant 3-oscillator states

$$\hat{H}^{(2)}|j,l,n_3\rangle=(2j+1)|j,l,n_3\rangle,$$

$$\hat{J}_y^2|j,l,n_3\rangle=l^2|j,l,n_3\rangle, \tag{4.12}$$

$$\hat{J}_{33}|j,l,n_3\rangle=n_3|j,l,n_3\rangle.$$

It proves convenient to set $|n|_2=2j$.

The eigenstates $|j,l,n_3\rangle$ are straightforwardly expressed in terms of the vectors $\hat{\phi}(n_1,n_2,n_3)=\Pi_{i=1}^3((\hat{a}_i^+)^{n_i}/\sqrt{n_i!})\phi_0$. One first observes that $\hat{\phi}(j-l,j+l,n_3)$ is an eigenvector of $\hat{H}^{(2)}$, $\hat{J}_z$ and $\hat{J}_{33}$ with eigenvalue $2j+1$, l, and n_3, respectively. Let $U(\alpha,\beta,\gamma)=e^{-i\alpha\hat{J}_z}e^{-i\beta\hat{J}_y}e^{-i\gamma\hat{J}_z}$; one further notes that $\hat{J}_y=U(\pi/2,\pi/2,0)\hat{J}_zU^{-1}(\pi/2,\pi/2,0)$. It thus follows that the states $e^{i\pi j/2}U(\pi/2,\pi/2,0)\hat{\phi}(j-l,j+l,n_3)$ are eigenvectors of $\hat{J}_y$ with eigenvalue l. The phase $e^{i\pi j/2}$ is introduced for convenience. We shall denote by $|\widetilde{j,l,n_3}\rangle$ the eigenstates of $\hat{H}^{(2)}$, $\hat{J}_y$ and $\hat{J}_{33}$.

In terms of the Wigner matrix elements, we have

$$|\widetilde{j,l,n_3}\rangle=e^{i\pi j/2}\sum_{m=-j}^{j}D_{m,l}^j\left(\frac{\pi}{2},\frac{\pi}{2},0\right)\phi(j-m,j+m,n_3), \tag{4.13}$$

where for m, $l=-j,\ldots,j$

$$\begin{aligned}D_{m,l}^j\left(\frac{\pi}{2},\frac{\pi}{2},0\right)&=e^{-(1/2)i\pi m}d_{m,l}^j\left(\frac{\pi}{2}\right)\\&=e^{-(1/2)i\pi m}\sqrt{\frac{(j+l)!(j-l)!}{(j+m)!(j-m)!}}\left(\frac{1}{2}\right)^j\sum_{\substack{0\le s\le j-m\\0\le s\le j-l}}(-)^{j-l+s}\binom{j-m}{s}\binom{j+m}{j-l-s},\end{aligned} \tag{4.14}$$

with $\binom{m}{s}$, the binomial coefficient. The eigenvectors of $\hat{J}_y^2$ with eigenvalue l^2 will thus be of the form

$$|j,l,n_3\rangle=\frac{1}{\sqrt{2}}\,(|\widetilde{j,l,n_3}\rangle+e^{i\alpha}|\widetilde{j,-l,n_3}\rangle) \tag{4.15}$$

with $e^{i\alpha}$ an arbitrary phase at this point. There remains to characterize the states that are invariant under the permutations of the oscillators and this is easily carried out with our choice of quantum numbers. We must determine which of the states $|j,l,n_3\rangle$ are stable against the action of σ_{12} and σ_{123}. From the transformations it induces on the observables (4.10), we see that the cyclic permutation is realized by the $4\pi/3$ rotations generated by $\hat{J}_y$. The symmetric states must satisfy

$$e^{-i(4\pi/3)\hat{J}_y}|j,l,n_3\rangle=|j,l,n_3\rangle. \tag{4.16}$$

Clearly, this imposes that l be a multiple of 3/2

$$l=\pm 3k/2,\quad k\in\mathbb{N}. \tag{4.17}$$

Similarly, the transposition σ_{12} is seen to correspond to the inversion $\mathscr{P}$ which acts as follows on the states $\hat{\phi}$: $\mathscr{P}\hat{\phi}(j-m,j+m,n_3)=(-1)^{j-m}\hat{\phi}(j-m,j+m,n_3)$. Hence, using the fact that

$$d_{m,-l}^j\left(\frac{\pi}{2}\right)=(-1)^{j+m}d_{m,l}^j\left(\frac{\pi}{2}\right), \tag{4.18}$$

we have

$$\mathscr{P}\widetilde{|j,l,n_3\rangle} = e^{i\pi j/2}\sum_{m=-j}^{j} D^{i}_{m,l}\left(\frac{\pi}{2},\frac{\pi}{2},0\right)(-1)^{j-m}\phi(j-m,j+m,n_3) = (-1)^{2j}\widetilde{|j,-l,n_3\rangle}. \tag{4.19}$$

Therefore, upon demanding the invariance of the states $|j,l,n_3\rangle$ under σ_{12}, i.e.,

$$\mathscr{P}|j,l,n_3\rangle = |j,l,n_3\rangle, \tag{4.20}$$

we find that the relative phase $e^{i\alpha}$ in (4.15) is $(-1)^{2j}$. This leads to the following formula for the symmetric 3-oscillator states:

$$\begin{aligned}|j,l,n_3\rangle &= \frac{1}{\sqrt{2}}\left(\widetilde{|j,l,n_3\rangle} + (-1)^{2j}\widetilde{|j,-l,n_3\rangle}\right)\\ &= \frac{e^{i\pi j/2}}{\sqrt{2}}\sum_{m=-j}^{j}(1+(-1)^{j-m})e^{-(1/2)i\pi m}d^{j}_{m,l}\left(\frac{\pi}{2}\right)\hat{\phi}(j-m,j+m,n_3)\end{aligned} \tag{4.21}$$

with l a multiple of 3. [We used the fact that $(-1)^{4j}=1$.] Note that the vectors $|j,l,n_3\rangle$ and $|j,-l,n_3\rangle$ only differ by a phase

$$|j,-l,n_3\rangle = (-1)^{2j}|j,l,n_3\rangle; \tag{4.22}$$

they hence describe the same physical state and should be identified. Furthermore, notice that the vectors $|j,l,n_3\rangle$ are normalized to one except when $l=0$; in this case, $\langle j,0,n_3|j,0,n_3\rangle = 2$.

B. The raising and lowering operators

We now aim to find raising and lowering operators that transform the degenerate eigenstates (4.21) of $H^{(3)}$ among themselves; ultimately we want to characterize the algebra that these constants of motion generate. The procedure that we shall follow has been outlined in Sec. II and amounts to obtaining an integrity basis for S_3 scalars in the envelopping algebra $\mathscr{U}(\mathrm{U}(3))$.

We shall denote by

$\Gamma_1 =$ 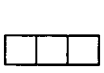 $\Gamma_2 =$ $\Gamma_3 =$ (4.23)

the three irreducible representations of S_3 of dimension 1, 1, and 2, respectively. The coefficients C^{c}_{ab} entering in the Clebsch–Gordan series $\Gamma_a\times\Gamma_b = \Sigma C^{c}_{ab}\Gamma_c$, $a,b,c=1,2,3$, are[11]

$$\begin{array}{lll} C^{c}_{1b}=C^{c}_{b1}=\delta^{c}_{b} & & \\ C^{1}_{22}=1 & C^{2}_{22}=C^{3}_{22}=0 & \\ C^{a}_{33}=1 & C^{3}_{23}=1 & C^{1}_{23}=C^{2}_{23}=0. \end{array} \tag{4.24}$$

The generators of U(3) transform under S_3 according to the representation $\Gamma' = (\Gamma_1+\Gamma_3)\times(\Gamma_1+\Gamma_3)$ whose irreducible content, according to (4.24), is

$$\Gamma' = \Gamma_1+\Gamma_1+\Gamma_2+\Gamma_3+\Gamma_3+\Gamma_3. \tag{4.25}$$

This decomposition is in effect exhibited in the basis (4.10) of U(3) which we shall keep using: the two scalars are $\hat{H}^{(2)}$ and $\hat{J}_{33}$, the pseudoscalar is $\hat{J}^{2}_{\gamma}$, and the 3 Γ_3 doublets, respectively, formed by $\hat{J}_{\pm}$, $\hat{J}_{\pm 3}$, $\hat{J}_{3\pm}$. We now wish to find a ''minimal'' basis for polynomials in the components of this Γ' tensor [i.e., elements of $\mathscr{U}(\mathrm{U}(3))$] that are invariant under S_3 (i.e., that transform according to Γ_1). We shall first determine the maximum number of polynomials in this basis by computing the generating function $B(\Gamma_1,\Gamma')$. One proceeds using formulas (2.36) and (2.38) repeatedly. The generating functions involving the irreducible representations are evaluated straightforwardly from the definition (2.35) and then cast in the form (2.39). One finds

$$
\begin{aligned}
&B(\Gamma_1,\Gamma_1)=\frac{1}{1-\lambda}, \quad B(\Gamma_1,\Gamma_2)=\frac{1}{1-\lambda^2},\\
&B(\Gamma_2,\Gamma_2)=\frac{\lambda}{1-\lambda^2}, \quad B(\Gamma_1,\Gamma_3)=\frac{1}{(1-\lambda^2)(1-\lambda^3)},\\
&B(\Gamma_2,\Gamma_3)=\frac{\lambda^3}{(1-\lambda^2)(1-\lambda^3)}, \quad B(\Gamma_3,\Gamma_3)=\frac{\lambda+\lambda^2}{(1-\lambda^2)(1-\lambda^3)}.
\end{aligned}
\tag{4.26}
$$

(Recall that S_3 has 1 cycle of length 1, 3 of length 2, and 2 of length 3. Γ_1 is the identity representation and Γ_2 the alternating one in which the transpositions are represented by -1. As for Γ_3, I and R given in (4.9) can be taken to play the role of A_s in (2.35) for transpositions and cyclic permutations, respectively.) Combining everything, we find that $B(\Gamma_1,\Gamma')$ takes the form

$$
B(\Gamma_1,\Gamma_1+\Gamma_1+\Gamma_2+\Gamma_3+\Gamma_3+\Gamma_3;\lambda_1,\lambda_2,\ldots,\lambda_6)=\frac{N(\lambda_1,\ldots,\lambda_6)}{D(\lambda_1,\ldots,\lambda_6)},
\tag{4.27}
$$

where

$$
D(\lambda_1,\ldots,\lambda_6)=(1-\lambda_1)(1-\lambda_2)(1-\lambda_3^2)(1-\lambda_4^2)(1-\lambda_4^3)(1-\lambda_5^2)(1-\lambda_5^3)(1-\lambda_6^2)(1-\lambda_6^3)
\tag{4.28}
$$

and

$$
\begin{aligned}
N(\lambda_1,\ldots,\lambda_6)=1&+\lambda_4\lambda_5+\lambda_4\lambda_6+\lambda_5\lambda_6+\lambda_4^2\lambda_5+\lambda_4\lambda_5^2+\lambda_4^2\lambda_6+\lambda_4\lambda_6^2+\lambda_5^2\lambda_6\\
&+\lambda_5\lambda_6^2+\lambda_4\lambda_5\lambda_6+\lambda_4^2\lambda_5^2+\lambda_4^2\lambda_6^2+\lambda_5^2\lambda_6^2+\lambda_4^2\lambda_5\lambda_6+\lambda_4\lambda_5^2\lambda_6\\
&+\lambda_4\lambda_5\lambda_6^2+\lambda_4^3\lambda_5\lambda_6+\lambda_4\lambda_5^3\lambda_5+\lambda_4\lambda_5\lambda_6^3+\lambda_4^2\lambda_5^2\lambda_6+\lambda_4^2\lambda_5\lambda_6^2\\
&+\lambda_4\lambda_5^2\lambda_6^2+\lambda_4^3\lambda_5^3+\lambda_4^3\lambda_6^3+\lambda_5^3\lambda_6^3+\lambda_4^3\lambda_5^2\lambda_6+\lambda_4^3\lambda_5\lambda_6^2+\lambda_4^2\lambda_5^3\lambda_6\\
&+\lambda_4^2\lambda_5\lambda_6^3+\lambda_4\lambda_5^3\lambda_6^2+\lambda_4\lambda_5^2\lambda_6^3+\lambda_4^2\lambda_5^2\lambda_6^2+\lambda_4^3\lambda_5^2\lambda_6^2+\lambda_4^2\lambda_5^3\lambda_6^2\\
&+\lambda_4^2\lambda_5^2\lambda_6^3+\lambda_3\lambda_4\lambda_5+\lambda_3\lambda_4\lambda_6+\lambda_3\lambda_5\lambda_6+\lambda_3\lambda_4^3\\
&+\lambda_3\lambda_5^3+\lambda_3\lambda_6^3+\lambda_3\lambda_4^2\lambda_5+\lambda_3\lambda_4\lambda_5^2+\lambda_3\lambda_4^2\lambda_6+\lambda_3\lambda_4\lambda_6^2\\
&+\lambda_3\lambda_5^2\lambda_6+\lambda_3\lambda_5\lambda_6^2+\lambda_3\lambda_4\lambda_5\lambda_6+\lambda_3\lambda_4^2\lambda_5^2+\lambda_3\lambda_4^2\lambda_6^2+\lambda_3\lambda_5^2\lambda_6^2\\
&+\lambda_3\lambda_4^2\lambda_5\lambda_6+\lambda_3\lambda_4\lambda_5^2\lambda_6+\lambda_3\lambda_4\lambda_5\lambda_6^2+\lambda_3\lambda_4^3\lambda_5\lambda_6+\lambda_3\lambda_4\lambda_5^3\lambda_6\\
&+\lambda_3\lambda_4\lambda_5\lambda_6^3+\lambda_3\lambda_4^2\lambda_5^2\lambda_6+\lambda_3\lambda_4^2\lambda_5\lambda_6^2+\lambda_3\lambda_4\lambda_5^2\lambda_6^2+\lambda_3\lambda_4^3\lambda_5^2\lambda_6\\
&+\lambda_3\lambda_4^3\lambda_5\lambda_6^2+\lambda_3\lambda_4^2\lambda_5^3\lambda_6+\lambda_3\lambda_4^2\lambda_5\lambda_6^3+\lambda_3\lambda_4\lambda_5^3\lambda_6^2+\lambda_3\lambda_4\lambda_5^2\lambda_6^3\\
&+\lambda_3\lambda_4^2\lambda_5^2\lambda_6^2+\lambda_3\lambda_4^3\lambda_5^2\lambda_6^2+\lambda_3\lambda_4^2\lambda_5^3\lambda_6^2+\lambda_3\lambda_4^2\lambda_5^2\lambda_6^3.
\end{aligned}
\tag{4.29}
$$

We can now count directly the number of elements that we would have in our integrity basis if the components of the Γ' tensor were classical objects with no further properties. The denominator reflects the invariants that can be formed by mixing together components that belong to the same irreducible modules only. For each factor $(1-\lambda_i^{\beta_i})$ there is one such invariant giving, in the present case, a total of nine. These invariants are

$$
\begin{aligned}
&\hat{H}^{(2)}, \quad \hat{J}_{33}, \quad \hat{J}_y^2, \quad \hat{J}_+\hat{J}_-+\hat{J}_-\hat{J}_+, \quad \hat{J}_{+3}\hat{J}_{-3}+\hat{J}_{-3}\hat{J}_{+3}, \quad \hat{J}_{3+}\hat{J}_{3-}+\hat{J}_{3-}\hat{J}_{3+},\\
&\qquad \hat{J}_+^3-\hat{J}_-^3, \quad \hat{J}_{+3}^3-\hat{J}_{-3}^3, \quad \hat{J}_{3+}^3-\hat{J}_{3-}^3.
\end{aligned}
\tag{4.30}
$$

The numerator $N(\lambda_1,\ldots,\lambda_6)$ shows the number of invariants that can be constructed by multiplying components belonging to different irreducible singlets or doublets. It also allows us to determine the number of elements in the integrity basis. In the case of reducible representations there is in general some overcounting in N in this respect.

Consider an example: The presence of $\lambda_4\lambda_5$ in (4.29), indicates that there exists a quadratic invariant in both sets of components $\hat{J}_\pm$ and $\hat{J}_{\pm 3}$, call it $Y_6^{(2)}=-2(\hat{J}_+\hat{J}_{-3}+\hat{J}_-\hat{J}_{-3})$. Meanwhile, the terms λ_4^2 and λ_5^2 in the denominator have already told us that there are two quadratic invariants say $Y_4^{(2)}=2(\hat{J}_+\hat{J}_-+\hat{J}_-\hat{J}_+)$ and $Y_2^{(2)}=(\hat{J}_{+3}\hat{J}_{-3}+\hat{J}_{-3}\hat{J}_{+3})/2$ that can be formed using separately the 2 Γ_3-doublets $\hat{J}_\pm$ and $\hat{J}_{3\pm}$. It is then checked that the quartic invariant associated with the term $\lambda_4^2\lambda_5^2$ in N can be built out of the invariants $Y_6^{(2)}$, $Y_4^{(2)}$, and $Y_2^{(2)}$ already included in the integrity basis. It thus follows, as a rule, that in counting the number of elements in the integrity basis, one should discard from N the monomials that can be formed from those of the form $\lambda_i^{\beta_i}$ occuring in the factors of D and those of lower degree in N. When this is done in (4.29) we are left with the following 23 terms:

$$\begin{aligned}&1+\lambda_4\lambda_5+\lambda_4\lambda_6+\lambda_5\lambda_6+\lambda_4^2\lambda_5+\lambda_4\lambda_5^2+\lambda_4^2\lambda_6+\lambda_4\lambda_6^2+\lambda_5^2\lambda_6+\lambda_5\lambda_6^2\\&\quad+\lambda_4\lambda_5\lambda_6+\lambda_3\lambda_4\lambda_5+\lambda_3\lambda_4\lambda_6+\lambda_3\lambda_5\lambda_6+\lambda_3\lambda_4^3+\lambda_3\lambda_5^3+\lambda_3\lambda_6^3\\&\quad+\lambda_3\lambda_4^2\lambda_5+\lambda_3\lambda_4\lambda_5^2+\lambda_3\lambda_4^2\lambda_6+\lambda_3\lambda_4\lambda_6^2+\lambda_3\lambda_5^2\lambda_6+\lambda_3\lambda_5\lambda_6^2+\lambda_3\lambda_4\lambda_5\lambda_6.\end{aligned}\tag{4.31}$$

We thus have a total of 32 invariants. In terms of the basis elements (4.10), they read

$$\begin{aligned}
&Y_1^{(1)}=\hat{J}_{33}, && Y_2^{(1)}=\hat{H}^{(2)}-1\\
&Y_1^{(2)}=4\hat{J}_y^2, && Y_2^{(2)}=\tfrac{1}{2}\,(\hat{J}_{+3}\hat{J}_{-3}+\hat{J}_{-3}\hat{J}_{+3}),\\
&Y_3^{(2)}=\tfrac{1}{2}\,(\hat{J}_{3+}\hat{J}_{3-}+\hat{J}_{3-}\hat{J}_{3+}), && Y_4^{(2)}=2(\hat{J}_+\hat{J}_-+\hat{J}_-\hat{J}_+),\\
&Y_5^{(2)}=\hat{J}_{3+}\hat{J}_{-3}+\hat{J}_{3-}\hat{J}_{+3}, && Y_6^{(2)}=-2(\hat{J}_+\hat{J}_{-3}+\hat{J}_-\hat{J}_{+3})\\
&Y_7^{(2)}=-2(\hat{J}_{3-}\hat{J}_++\hat{J}_{3+}\hat{J}_-), && Y_1^{(3)}=4\hat{J}_y(\hat{J}_+\hat{J}_{3-}-\hat{J}_-\hat{J}_{3+}),\\
&Y_2^{(3)}=4\hat{J}_y(\hat{J}_+\hat{J}_{-3}-\hat{J}_-\hat{J}_{+3}), && Y_3^{(3)}=-2\hat{J}_y(\hat{J}_{3+}\hat{J}_{-3}-\hat{J}_{3-}\hat{J}_{+3}),\\
&Y_4^{(3)}=4i(\hat{J}_+^2\hat{J}_{3+}-\hat{J}_-^2\hat{J}_{3-}), && Y_5^{(3)}=4i(\hat{J}_+^2\hat{J}_{+3}-\hat{J}_-^2\hat{J}_{-3}),\\
&Y_6^{(3)}=i(\hat{J}_{3+}^2\hat{J}_{+3}-\hat{J}_{3-}^2\hat{J}_{-3}), && Y_7^{(3)}=-2i(\hat{J}_+\hat{J}_{3+}^2-\hat{J}_-\hat{J}_{3-}^2),\\
&Y_8^{(3)}=-2i(\hat{J}_+\hat{J}_{+3}^2-\hat{J}_-\hat{J}_{-3}^2), && Y_9^{(3)}=i(\hat{J}_{3+}\hat{J}_{+3}^2-\hat{J}_{3-}\hat{J}_{-3}^2),\\
&Y_{10}^{(3)}=-2i(\hat{J}_+\hat{J}_{3+}\hat{J}_{+3}-\hat{J}_-\hat{J}_{3-}\hat{J}_{-3}), && Y_{11}^{(3)}=-8i(\hat{J}_+^3-\hat{J}_-^3),\\
&Y_{12}^{(3)}=i(\hat{J}_{3+}^3-\hat{J}_{3-}^3), && Y_{13}^{(3)}=i(\hat{J}_{+3}^3-\hat{J}_{-3}^3),\\
&Y_1^{(4)}=-8i\hat{J}_y(\hat{J}_+^2\hat{J}_{3+}+\hat{J}_-^2\hat{J}_{3-}), && Y_2^{(4)}=-8i\hat{J}_y(\hat{J}_+^2\hat{J}_{+3}+\hat{J}_-^2\hat{J}_{-3}),\\
&Y_3^{(4)}=-2i\hat{J}_y(\hat{J}_{3+}^2\hat{J}_{+3}+\hat{J}_{3-}^2\hat{J}_{-3}), && Y_4^{(4)}=4i\hat{J}_y(\hat{J}_+\hat{J}_{3+}^2+\hat{J}_-\hat{J}_{3-}^2),\\
&Y_5^{(4)}=4i\hat{J}_y(\hat{J}_+\hat{J}_{+3}^2+\hat{J}_-\hat{J}_{-3}^2), && Y_6^{(4)}=-2i\hat{J}_y(\hat{J}_{3+}\hat{J}_{+3}^2+\hat{J}_{3-}\hat{J}_{-3}^2),\\
&Y_7^{(4)}=4i\hat{J}_y(\hat{J}_+\hat{J}_{3+}\hat{J}_{+3}+\hat{J}_-\hat{J}_{3-}\hat{J}_{-3}), && Y_8^{(4)}=16i\hat{J}_y(\hat{J}_+^3+\hat{J}_-^3),\\
&Y_9^{(4)}=-2i\hat{J}_y(\hat{J}_{3+}^3+\hat{J}_{3-}^3), && Y_{10}^{(4)}=-2i\hat{J}_y(\hat{J}_{+3}^3+\hat{J}_{-3}^3).
\end{aligned}$$

This is the point at which we must take into account the special nature of the realization we are using and the fact that there are various relations of the type (2.32) among the generators. This

leads one to the observation that of the 32 invariants given in (4.32) only 9 (actually 8) are functionally independent in the particular problem we are considering. Let us denote these nine invariants as follows:

$$\begin{aligned}
&\hat{J}_{33}=Y_1^{(1)}, \quad \hat{J}=Y_2^{(1)}/2, \quad \hat{J}_y^2=Y_1^{(2)}/4,\\
&X_1=-2Y_2^{(2)}, \quad X_{3/2}=Y_{13}^{(3)}, \quad X_{1/2}=Y_6^{(2)},\\
&X_{-1}=-2Y_3^{(2)}, \quad X_{-(3/2)}=Y_{12}^{(3)}, \quad X_{-(1/2)}=Y_7^{(2)}.
\end{aligned} \tag{4.32}$$

The remaining 23 invariants are expressed in terms of the above operators in the following manner: One first straightforwardly obtains the 19 relations below which give 19 of the 32 invariants as polynomials in the remaining ones

$$\begin{aligned}
&Y_4^{(2)}=4(\hat{J}(\hat{J}+1)-\hat{J}_y^2),\\
&Y_5^{(2)}=4\hat{J}_{33}(\hat{J}+1),\\
&Y_1^{(3)}=-Y_4^{(3)}-X_{-(1/2)}(2\hat{J}-3),\\
&Y_2^{(3)}=Y_5^{(3)}+X_{1/2}(2\hat{J}+5),\\
&Y_3^{(3)}=8\hat{J}_{33}\hat{J}_y^2,\\
&Y_6^{(3)}=-X_{-(1/2)}\hat{J}_{33},\\
&Y_7^{(3)}=2\hat{J}X_{-1},\\
&Y_8^{(3)}=2(\hat{J}+1)X_1,\\
&Y_9^{(3)}=-\hat{J}_{33}X_{1/2},\\
&Y_{10}^{(3)}=-8\hat{J}_{33}(\hat{J}(\hat{J}+1)-\hat{J}_y^2),\\
&Y_1^{(4)}=-4\hat{J}_y^2X_{-(1/2)}+X_{-(1/2)}(2\hat{J}-3)^2+Y_4^{(3)}(2\hat{J}-3),\\
&Y_2^{(4)}=4\hat{J}_y^2X_{1/2}-X_{1/2}(2\hat{J}+5)^2-Y_5^{(3)}(2\hat{J}+5),\\
&Y_3^{(4)}=-(X_{-(1/2)}(2\hat{J}-3)+Y_4^{(3)})\hat{J}_{33},\\
&Y_4^{(4)}=4\hat{J}_y^2X_{-1},\\
&Y_5^{(4)}=-4\hat{J}_y^2X_1,\\
&Y_6^{(4)}=\hat{J}_{33}(Y_5^{(3)}+X_{1/2}(2\hat{J}+5)),\\
&Y_7^{(4)}=-16\hat{J}_y^2\hat{J}_{33},\\
&Y_9^{(4)}=-\tfrac{1}{2}X_{-1}X_{-(1/2)}+X_{-(3/2)}(2\hat{J}-1),\\
&Y_{10}^{(4)}=-X_{3/2}(2\hat{J}+3)+\tfrac{1}{2}X_1X_{1/2}.
\end{aligned} \tag{4.33}$$

Allowing division by diagonal operators further permits us to write $Y_4^{(3)}$, $Y_5^{(3)}$, $Y_{11}^{(3)}$, and $Y_8^{(4)}$ as functions of operators (4.33)

$$\begin{aligned}
Y_4^{(3)}&=\frac{1}{2}X_{1/2}X_{-1}\frac{1}{(\hat{J}_{33}+2)},\\
Y_5^{(3)}&=-4X_{1/2}+\frac{1}{2}X_1X_{-(1/2)}\frac{1}{(\hat{J}_{33}+1)},\\
Y_{11}^{(3)}&=-X_{1/2}X_{-(1/2)}\frac{1}{(\hat{J}_{33}+1)}-16[\hat{J}-\hat{J}^3+\hat{J}_y^2(\hat{J}-2)],\\
Y_8^{(4)}&=-32[\hat{J}_y^2(\hat{J}^2+5\hat{J}-7-\hat{J}_y^2)+3\hat{J}(1-\hat{J}^2)]+2X_{1/2}X_{-(1/2)}\frac{\hat{J}-3}{(\hat{J}_{33}+1)}\\
&\quad+\frac{1}{2}X_{1/2}^2X_{-1}\frac{1}{(\hat{J}_{33}+1)(\hat{J}_{33}+2)}.
\end{aligned} \tag{4.34}$$

We may also remark that $\hat{J}_y^2$ could be eliminated in this fashion

$$\left(\hat{J}_y^2 = \hat{J}^2 - \frac{1}{16} X_1 X_{-1} \frac{1}{(\hat{J}_{33}+1)(\hat{J}_{33}+2)} \right)$$

but since it is a labeling operator we have chosen to keep it in the basis set (4.33).

In principle, a further reduction in the number of independent invariants could be expected from the noncommuting nature of the Γ'-tensor component $\hat{J}_{ij}$ as was explained in Sec. II. However since we shall give below the commutation relations between operators (4.33), there is no use in pursuing this question.

Note that set (4.33) includes the diagonal operators $\hat{H}^{(2)}$, $\hat{J}_{33}$, and $\hat{J}_y^2$; leaving six raising and lowering operators: $X_{\pm(3/2)}, X_{\pm(3/2)}, X_{\pm 1}$. Observe further that

$$X_s^\dagger = X_{-s}, \quad s = \tfrac{3}{2}, \tfrac{1}{2}, 1. \tag{4.35}$$

C. The symmetry algebra: Representation and commuting relations

We now wish to present the action of these operators X_s, $s = \pm(3/2), \pm(1/2), \pm 1$ on the symmetric 3-oscillator states $|j,l,n_3\rangle$ to show that they indeed transform these vectors irreducibly among themselves. This action reads as follows:

$$\hat{J}_{33}|j,l,n_3\rangle = n_3|j,l,n_3\rangle, \tag{4.36}$$

$$\hat{J}|j,l,n_3\rangle = j|j,l,n_3\rangle, \tag{4.37}$$

$$\hat{J}_y^2|j,l,n_3\rangle = l^2|j,l,n_3\rangle, \tag{4.38}$$

$$X_{-1}|j,l,n_3\rangle = 4\sqrt{(j+l)(j-l)(n_3+1)(n_3+2)}\,|j-1,l,n_3+2\rangle, \tag{4.39}$$

$$\begin{aligned} X_{-(3/2)}|j,l,n_3\rangle = &-\sqrt{8(j-l-2)(j-l-1)(j-l)(n_3+1)(n_3+2)(n_3+3)}\,|j-\tfrac{3}{2},l+\tfrac{3}{2},n_3+3\rangle \\ &+\sqrt{8(j+l-2)(j+l-1)(j+l)(n_3+1)(n_3+2)(n_3+3)}\,|j-\tfrac{3}{2},l-\tfrac{3}{2},n_3+3\rangle, \end{aligned} \tag{4.40}$$

$$\begin{aligned} X_{-(1/2)}|j,l,n_3\rangle = &-\sqrt{8(j-l-1)(j-l)(j+l+1)(n_3+1)}\,|j-\tfrac{1}{2},l+\tfrac{3}{2},n_3+1\rangle \\ &+\sqrt{8(j+l-1)(j+l)(j-l+1)(n_3+1)}\,|j-\tfrac{1}{2},l-\tfrac{3}{2},n_3+1\rangle, \end{aligned} \tag{4.41}$$

$$\begin{aligned} X_{1/2}|j,l,n_3\rangle = &\sqrt{8(j-l)(j+l+1)(j+l+2)n_3}\,|j+\tfrac{1}{2},l+\tfrac{3}{2},n_3-1\rangle \\ &-\sqrt{8(j+l)(j-l+1)(j-l+2)n_3}\,|j+\tfrac{1}{2},l-\tfrac{3}{2},n_3-1\rangle, \end{aligned} \tag{4.42}$$

$$\begin{aligned} X_{3/2}|j,l,n_3\rangle = &\sqrt{8(j+l+1)(j+l+2)(j+l+3)n_3(n_3-1)(n_3-2)}\,|j+\tfrac{3}{2},l+\tfrac{3}{2},n_3-3\rangle \\ &-\sqrt{8(j-l+1)(j-l+2)(j-l+3)n_3(n_3-1)(n_3-2)}\,|j+\tfrac{3}{2},l-\tfrac{3}{2},n_3-3\rangle, \end{aligned} \tag{4.43}$$

$$X_1|j,l,n_3\rangle = 4\sqrt{(j+l+1)(j-l+1)n_3(n_3-1)}\,|j+1,l,n_3-2\rangle. \tag{4.44}$$

Notice that in these expressions, formula (4.22) has to be used whenever l becomes negative.

Finally, we give the commutation relations among the eight independent generators of (4.33) [excluding $\hat{J}_y^2$ and taking (4.35) into account]

$$[X_1, X_{3/2}] = 0,$$

$$[X_1, X_{1/2}] = -4X_{3/2},$$

$$[X_1, X_{-(1/2)}] = 2X_1 X_{-(1/2)} \frac{1}{(\hat{J}_{33}+1)} + 8X_{1/2}(\hat{J}_{33}-1),$$

$$[X_1, X_{-(3/2)}] = 6X_{1/2}X_{-1} \frac{(\hat{J}_{33}+1)}{(\hat{J}_{33}+2)} + 12X_{-(1/2)}(4\hat{J}+4\hat{J}\hat{J}_{33}-3\hat{J}_{33}-\hat{J}_{33}^2-4),$$

$$[X_1, X_{-1}] = 16[\hat{J}_{33}(1+2\hat{J}-\hat{J}_{33}(1+2\hat{J})+4(\hat{J}^2-\hat{J}_y^2))] + 32(\hat{J}^2-\hat{J}_y^2), \tag{4.45}$$

$$[X_{3/2}, X_{1/2}] = 3X_1^2,$$

$$[X_{3/2}, X_{-(1/2)}] = -3X_{1/2}^2 - 12X_1(2\hat{J}_{33}(1+\hat{J})-\hat{J}-2\hat{J}_y^2-4),$$

$$[X_{3/2}, X_{-(3/2)}] = X_{1/2}X_{-(1/2)} \frac{(9\hat{J}_{33}^2+9\hat{J}_{33}+6)}{(\hat{J}_{33}+1)} - 96(3\hat{J}^2-4\hat{J}\hat{J}_y^2+5\hat{J}_y^2-3\hat{J}) - 48\hat{J}_{33}(15\hat{J}^2+21\hat{J}_y^2$$
$$+4-12\hat{J}\hat{J}_y^2-3\hat{J}) - 288\hat{J}_{33}^2(\hat{J}_y^2-2\hat{J}_y^2\hat{J}-3\hat{J}-1) - 48\hat{J}_{33}^3(2+3\hat{J}+3\hat{J}^2+3\hat{J}_y^2),$$

$$[X_{1/2}, X_{\cdot(1/2)}] = X_{1/2}X_{-(1/2)} \frac{1}{(\hat{J}_{33}+1)} - \hat{J}_{33}(48\hat{J}^2-80\hat{J}_y^2+48\hat{J}).$$

V. CONCLUSIONS

How the general description of S_N-invariant oscillator states given in Sec. II applies to specific cases should be clear from the discussions in Secs. III and IV. If we wished to work out the case $N=4$ with $H^{(4)}=\Sigma_{i=1}^4 a_i^+ a_i$, for instance, the analysis would proceed as in the last section. The cycles σ_{12}, σ_{123}, and σ_{1234} of length 2, 3, and 4, respectively, can now be used to generate S_4. It remains convenient to pass to relative and CM coordinates and one might again take those of Jacobi. It proves in fact more practical to introduce the set

$$\begin{aligned} \hat{x}_1 &= \tfrac{1}{2}(x_1+x_2-x_3-x_4), \quad \hat{x}_2 = \tfrac{1}{2}(x_2+x_3-x_1-x_4), \\ \hat{x}_3 &= \tfrac{1}{2}(x_1+x_3-x_2-x_4), \quad \hat{x}_4 = \tfrac{1}{2}(x_1+x_2+x_3+x_4), \end{aligned} \tag{5.1}$$

which is such that σ_{12}: $\hat{x}_1 \to \hat{x}_1$, σ_{1234}: $\hat{x}_3 \to -\hat{x}_3$. The Hamiltonian $H^{(4)}$ is U(4) invariant. The operators $H^{(3)}$ and $\hat{J}_{44}$ are taken to be diagonal on the state vectors thus subducing to U(3) representations. Two more S_4-invariant labeling operators are to be found in $\mathcal{U}(\mathrm{U}(3))$. Since S_4 acts by improper three-dimensional rotations on the coordinates $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$, we are led to the following chain of subgroups $\mathrm{U}(4)\supset\mathrm{U}(3)\supset\mathrm{O}(3)\supset\mathrm{O}(2)$.

The case $N=5$ admits of course a similar treatment with the Wigner supermultiplet scheme underlying it: $\mathrm{U}(5)\supset\mathrm{U}(4)\supset\mathrm{U}(2)\times\mathrm{U}(2)$.

The situation obviously becomes more and more complicated as the number of oscillators is increased, and while the formalism applies in general we do not have explicit formulas for arbitrary N. This requires further investigation.

We have been concerned in this paper with the invariance algebra of N-identical oscillators. It could be interesting to examine in the same spirit the spectrum-generating algebras of these systems. Instead of starting from $\mathrm{SU}(N)$, one would base the analysis on $\mathrm{Sp}(2N)$ or even the superalgebra $\mathrm{OSp}(N,2N)$. In this connection, let us remark that the nonlinear algebras of identical oscillators are embedded in the (completion of) universal Heisenberg algebra $\mathcal{U}(\mathcal{H}(N))$ generated by the creation and annihilation operators a_i, a_i^+ $i=1,...,N$. It could prove instructive to study directly the quotient of this algebra under S_N.

Our study raises a number of challenging issues. Noteworthy is the problem of understanding the structure of the nonlinear symmetry algebras and developing their representation theory. This could be done of course in the special case $N=2$ where the symmetries form a cubic SU(2) algebra with three generators. The case $N=3$, has brought a rare example of a dynamical nonlinear algebra of higher dimension. A representation of it could be obtained from the physical problem; this should be of help towards understanding these nonlinear algebras and their representations more abstractly. It is tantalizing that many polynomial algebras of physical interest can be identified as finite W algebras.[14] Whether this is the case in the present problem should be examined.

Finally, we plan to extend the results presented here to systems of the Calogero–Moser type so as to develop the algebraic treatment of these models. We should point out that this has already been done for the special case $N=2$[15] and $N=3$.[16] We hope to report on some of these interesting questions in the near future.

ACKNOWLEDGMENTS

We are grateful to Jean Le Tourneux, Jiri Patera, and Éric Ragoucy for comments and discussions.

This work has been supported in part through funds provided by NSERC (Canada) and FCAR (Québec). One of us (L.L.) holds a FCAR postgraduate scholarship.

[1] A. Polychronakos, ''A classical realization of quantum algebras,'' Mod. Phys. Lett. A **5**, 2325–2334 (1990).
[2] M. Rocek, ''Representation theory of the nonlinear SU(2) algebra,'' Phys. Lett. B **255**, 554–557 (1991).
[3] P. W. Higgs, ''Dynamical symmetries in a spherical geometry I,'' J. Phys. A **12**, 309–323 (1979).
[4] A. S. Zhedanov, ''The 'Higgs Algebra' as a 'quantum' deformation of SU(2),'' Mod. Phys. Lett. A **7**, 507–512 (1992).
[5] C. Quesne, ''Generalized deformed parafermions, nonlinear deformations of so(3) and exactly solvable potentials,'' Phys. Lett. A **193**, 245–250 (1994).
[6] P. Létourneau and L. Vinet, ''Superintegrable systems: polynomial algebra and quasiexactly solvable Hamiltonians,'' Ann. Phys. **243**, 144–168 (1995).
[7] A. J. Dragt, ''Classification of three-particle states according to SU(3),'' J. Math. Phys. **6**, 533–553 (1965).
[8] J.-M. Lévy-Leblond and M. Lévy-Nahas, ''Three-particle nonrelativistic kinematics and phase space,'' J. Math. Phys. **6**, 1571–1575 (1965).
[9] J.-M. Lévy-Leblond, ''Global and democratic methods for classifying N-particle states,'' J. Math. Phys. **7**, 2217–2229 (1966).
[10] P. Kramer and M. Moshinsky, ''Group theory of harmonic oscillators and nuclear structure,'' *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York, 1968), pp. 340–468.
[11] M. Hamermesh, *Group Theory and its Application to Physical Problems*, 2nd ed. (Addison-Wesley, London, 1964).
[12] B. R. Judd, W. Miller, Jr., J. Patera, and P. Winternitz, ''Complete sets of commuting operators and O(3) scalars in the enveloping algebra of SU(3),'' J. Math. Phys. **15**, 1787–1799 (1974).
[13] J. Patera, R. T. Sharp, and P. Winternitz, ''Polynomial irreducible tensors for point groups,'' J. Math. Phys. **19**, 2362–2376 (1978).
[14] J. De Boer, F. Harmsze, and T. Tjin, ''Non-linear finite W-symmetries and applications in elementary systems,'' Phys. Rep. **272**, 139 (1996).
[15] R. Floreanini, L. Lapointe, and L. Vinet, ''The polynomial SU(2) symmetry algebra of the two-body Calogero model,'' Phys. Lett. B **389**, 327–333 (1996).
[16] R. Floreanini, L. Lapointe, and L. Vinet, ''The dynamical algebra of the three-body Calogero model,'' Phys. Lett. B (to appear, 1998).

Rational conformal field theory extensions of $W_{1+\infty}$ in terms of bilocal fields

L. S. Georgiev and I. T. Todorov
Institute for Nuclear Research and Nuclear Energy,
Tsarigradsko Chaussee 72, BG-1784 Sofia, Bulgaria

(Received 29 October 1997; accepted for publication 25 March 1998)

The rational conformal field theory extensions of $W_{1+\infty}$ at $c=1$ are in one-to-one correspondence with one-dimensional integral lattices $L(m)$. Each extension is associated with a pair of oppositely charged ''vertex operators'' of charge square $m \in \mathbf{N}$. Their product defines a bilocal field $V_m(z_1,z_2)$ whose expansion in powers of $z_{12}=z_1-z_2$ gives rise to a series of (neutral) local quasiprimary fields $V^l(z,m)$ (of dimension $l+1$). The associated bilocal exponential of a normalized current generates the $W_{1+\infty}$ algebra spanned by the $V^l(z,1)$ (and the unit operator). The extension of this construction to higher (integer) values of the central charge c is also considered. Applications to a quantum Hall system require computing characters (i.e., chiral partition functions) depending not just on the modular parameter τ, but also on a chemical potential ζ. We compute such a ζ dependence of orbifold characters, thus extending the range of applications of a recent study of affine orbifolds. © *1998 American Institute of Physics.* [S0022-2488(98)01011-1]

I. INTRODUCTION

The unitary positive energy representations of $W_{1+\infty}$ (Ref. 1) correspond to positive integer central charges $c \in \mathbf{N}$. A standard tool in their study[1–4] is the embedding of all irreducible modules for a given c into the Fock space of c pairs of free Weyl fermion fields $\psi_i^*(z)$. It has been used in Ref. 5 to develop a conformal (quantum) field theory (CFT) approach to $W_{1+\infty}$ based on the expansion of the bilocal field

$$V_1(z_1,z_2)=\sum_{i=1}^{c}\left(\psi_i^*(z_1)\psi_i(z_2)-\frac{1}{z_{12}}\right), \quad z_{12}=z_1-z_2 \tag{1.1}$$

into local quasiprimary fields. V_1 is invariant under $\mathrm{SO}(2c)(\supset \mathrm{U}(1)^{\times c})$ gauge transformations in the $2c$-dimensional space of ψ_i and ψ_i^* [it is, in particular, $\mathrm{U}(1)^{\times c}$ neutral]. The index 1 on V indicates the charge square of ψ_i (and ψ_i^*) which coincides with the (total) conformal dimension of V_1: there is a representation $U(\rho)$ of the multiplicative group of positive reals ($\rho>0$) in the state space of our CFT such that

$$U(\rho)V_1(z_1,z_2)U(\rho)^{-1}=\rho V_1(\rho z_1,\rho z_2) \quad (\text{and } U(\rho)|0\rangle=|0\rangle) \tag{1.2}$$

($|0\rangle$ representing the unique zero-energy state, the *vacuum* of the theory). The chiral algebra W_c of c Weyl fermions (and their conjugates) corresponds to the simplest orthogonal lattice, $\mathbf{Z}^c$, of admissible charges: $W_c=\mathscr{A}(\mathbf{Z}^c)$. In general, the abelian compact Lie group $\mathrm{U}(1)^{\times c}$ defines a lattice in $\mathrm{u}(1)^{\otimes c}$ of elements ω satisfying $e^{2\pi i\omega}=1$ (where 1 stands for the group unit). Equipped with a positive definite integral quadratic form, $(\omega|\omega') \in \mathbf{Z}$ $|\omega|^2=(\omega|\omega) \in \mathbf{N}$ for $\omega \neq 0$, L gives rise to a chiral algebra $\mathscr{A}(L)$ whose positive energy (localizable) representations give rise to a rational conformal field theory (RCFT) (see, e.g., Sec. 1 of Ref. 6). Singling out a current $H(z)$ in the Kac–Moody algebra $\hat{\mathrm{u}}(1)^{\otimes c}$ we can view $\mathscr{A}(L)$ as a RCFT extension of the level c $W_{1+\infty}$ algebra which contains H. We define a bilocal field $V_m(z_1,z_2)$, expressed as a bilocal exponential in H—see (2.17) [generalizing (1.1) to any positive integer charge square m] and establish for it

 © 1998 American Institute of Physics

an *operator product expansion* (OPE) of quasiprimary fields $V^l(z,m)$ of dimension $l+1$ ($l=0,1,\ldots$). We prove, however, that only for $m=1$, in which case the fields $V^i(z)=V^l(z,1)$ are expressed in terms of an orthonormal set of u(1) currents J^i,

$$\langle J^i(z_1)J^j(z_2)\rangle=\delta_{ij}z_{12}^{-2},\quad i,j=1,\ldots,c, \tag{1.3}$$

do the V^l (and the unit operator) close an infinite dimensional Lie algebra (which coincides with the level c $W_{1+\infty}$). We work out the details, establishing exact OPE formulas, in Sec. II for the case $c=1$ and indicate the extension to arbitrary $c\in\mathbf{N}$ in Sec. III.

II. RATIONAL EXTENSIONS OF $W_{1+\infty}$ FOR $C=1$. OPE OF QUASIPRIMARY BILOCAL FIELDS

A. Local field basis for $W_{1+\infty}$ in terms of Schur polynomials of a u(1) current and its derivatives

The product of two oppositely charged fields $\phi^{\pm}(z)$ of charge square m can be expressed as a normal ordered exponential of a U(1) current:[7]

$$\phi^+(z_1)\phi^-(z_2)=z_{12}^{-m}:\exp\left\{\int_{z_2}^{z_1}H(z)\mathrm{d}z\right\}:,\quad z_{12}=z_1-z_2. \tag{2.1}$$

(The subtractions involved in the definition of normal ordering can be formally described by dividing the exponential by its vacuum expectation value.) The normalization of the ''Cartan current'' H can be read off the singular terms of the OPE:

$$H(z_1)H(z_2)\sim\frac{m}{z_{12}^2},\quad H(z_1)\phi^{\pm}(z_2)\sim\frac{\pm m}{z_{12}}\phi^{\pm}(z_2). \tag{2.2}$$

For a $c=1$ theory the stress energy tensor $T(z)$ is expressed in terms of H by the Sugawara formula:

$$T(z)=\frac{1}{2m}:H^2(z):\quad\left(:H(z_1)H(z_2):\equiv H(z_1)H(z_2)-\frac{m}{z_{12}^2}\right). \tag{2.3}$$

Its Laurent modes L_n generate the Virasoro algebra:

$$T(z)=\sum_{n\in\mathbf{Z}}L_nz^{-n-2}\left([L_n,L_{n'}]=(n-n')L_{n+n'}+\frac{c}{12}(n^3-n)\delta_{n,-n'},\quad c=1\right). \tag{2.4}$$

A central result in Ref. 5 says that the bilocal field (2.1) is, for $m=1$, a generating functional for a local field basis of the $W_{1+\infty}$ algebra. In fact, it then coincides (up to a c-number term) with (1.1) for central charge $c=1$. We shall reformulate this result using the U(1) current alone (without reference to the free Weyl field ψ) thus extending it to the more general current algebra framework.

Denote the normalized U(1) current $(1/\sqrt{m})H(z)$ by $J^0(z)$. Using the Taylor expansion of the integral in the exponential in (2.1),

$$\int_{z_2}^{z_1}J^0(z)\mathrm{d}z=\sum_{l=0}^{\infty}\frac{z_{12}^{l+1}}{(l+1)!}\partial_2^lJ^0(z_2)\quad\left(\partial_2=\frac{\partial}{\partial z_2};\quad\langle J^0(z_1)J^0(z_2)\rangle_0=z_{12}^{-2}\right) \tag{2.5}$$

we introduce the *Schur polynomials* (in the terminology of Refs. 8 and 9) of J^0 and its derivatives,

$$J^l(z)=l!:S_{l+1}\left(J^0(z),\frac{\partial J^0(z)}{2!},\ldots,\frac{\partial^lJ^0(z)}{(l+1)!}\right):, \tag{2.6}$$

setting

$$V_1(z_1,z_2)=\frac{1}{z_{12}}\left(:\exp\left\{\int_{z_2}^{z_1}J^0(z)\mathrm{dz}\right\}:-1\right)=\sum_{l=0}^{\infty}\frac{z_{12}^l}{l!}J^l(z_2). \tag{2.7}$$

In general, a Schur polynomial of N variables is defined by

$$S_N(t_1,t_2,\ldots,t_k,\ldots)=\sum_{n_1+2n_2+\cdots+kn_k+\cdots=N}\frac{t_1^{n_1}t_2^{n_2}\cdots t_k^{n_k}\cdots}{n_1!n_2!\cdots n_k!\cdots}. \tag{2.8}$$

The series of all Schur polynomials (of an increasing number of variables) is generated by

$$\exp\left(\sum_{k=1}^{\infty}t_kz^k\right)=\sum_{N=0}^{\infty}z^NS_N(t_1,t_2,\ldots,t_k,\ldots). \tag{2.9}$$

$J^l(z)$, as computed from (2.5) and (2.7), satisfy the recurrence relation

$$(l+1)J^l(z)=(J^0+\partial)lJ^{l-1}=\cdots=(J^0+\partial)^lJ^0\left(\equiv(l+1)!S_{l+1}(J^0,\ldots),\partial\equiv\frac{\mathrm{d}}{\mathrm{dz}}\right); \tag{2.10}$$

we have: $2J^1=(J^0)^2+\partial J^0$, $3J^2=(J^0)^3+3J^0\partial J^0+\partial^2J^0$, $4J^3=(J^0)^4+6(J^0)^2\partial J^0+3(\partial J^0)^2+4J^0\partial^2J^0+\partial^3J^0$, etc. (Viewed as quantum operators these expressions should be understood as normal products.) The local fields J^l (which, together with the unit operator span the infinite dimensional Lie algebra $W_{1+\infty}$) are singled for their simple commutation relations (Ref. 5, Lemma 1.2) which are coded in the (singular terms of the) OPE

$$J^k(z_1)J^l(z_2)\simeq\sum_{\nu\geqslant1}\left\{\frac{l!}{(l-\nu)!}J^{k+l-\nu}(z_2)-(-1)^\nu\frac{k!}{(k-\nu)!}J^{k+l-\nu}(z_1)\right\}z_{12}^{-\nu-1}$$
$$+(-1)^kl!k!z_{12}^{-k-l-2}. \tag{2.11}$$

As we shall see shortly, the bilocal field (2.7) can also be expanded in a basis of local *quasiprimary* fields $V^l(z)$ of dimension $l+1$ which are expressed in terms of J^l (Ref. 5, Lemma 1.3) by

$$V^l(z)=\frac{(l!)^2}{(2l)!}\sum_{k=0}^{\infty}(-1)^k\binom{l}{k}\binom{2l-k}{l}\partial^kJ^{l-k}(z). \tag{2.12}$$

A characteristic feature of quasiprimary fields is their orthogonality; indeed, the two-point function of a pair of V^l's is

$$\langle V^l(z_1)V^{l'}(z_2)\rangle_0=\frac{(l!)^2}{(2l)!}z_{12}^{-2l-2}\delta_{l,l'} \tag{2.13}$$

($\langle\cdots\rangle_0$ being a shorthand for the vacuum expectation value $\langle0|\cdots|0\rangle_0$).

According to Ref. 1 each positive energy unitary irreducible representation of $W_{1+\infty}$ for $c=1$ is characterized by a single real number r, the J^0 charge, and the existence of a unique ground state $|r\rangle$ such that

$$(J_0^0-r)|r\rangle=0=(L_0-\tfrac{1}{2}r^2)|r\rangle,\quad J_n^0|r\rangle=0(=L_n|r\rangle)\quad\text{for } n>0. \tag{2.14}$$

Here J_n^0 are the $J^0(z)$ Laurent modes, such that

$$J^0(z)=\sum_{n\in\mathbf{Z}}J_n^0z^{-n-1},\quad[J_n^0,J_{n'}^0]=n\delta_{n,-n'}. \tag{2.15}$$

The *minimal quantum field theory* whose state space includes the irreducible $W_{1+\infty}$ modules $\mathscr{H}$ and $\mathscr{H}_r$ (the *vacuum module* and the charge $r\neq0$ module $\mathscr{H}_r$) is generated by a pair of oppositely charged fields $\phi(z,\pm r)$. Its OPE algebra contains (the $W_{1+\infty}$ currents and) an infinite series of

charged fields $\phi(z,nr)$, $n\in\mathbf{Z}$ (relatively local to J^0—see Ref. 10). For $r^2=m\in\mathbf{N}$ they are also local among themselves Bose or Fermi fields; more precisely, they satisfy

$$z_{12}^{r_1r_2}\phi(z_1,r_1)\phi(z_2,r_2)=z_{21}^{r_1r_2}\phi(z_2,r_2)\phi(z_1,r_1)\quad \text{for } r_{1,2}\in\mathbf{Z}\sqrt{m}\quad (m=1,2,\ldots\text{fixed}). \tag{2.16}$$

In this case we recover the algebra $\mathscr{A}_m$ generated by the fields $\phi^{\pm}(z)=\phi(z,\pm\sqrt{m})$ satisfying (2.1), which provides a RCFT extension of $W_{1+\infty}$. Moreover, as noted in Ref. 5 (Sec. 3) any RCFT extension of $W_{1+\infty}$ for $c=1$ is of this type.

B. Expansion of the dipole field (2.1) into local quasiprimary fields $V^l(z,m)$

We introduce for arbitrary $m(=1,2,\ldots)$ and $c=1$ the counterpart of (1.1), the *bilocal field*

$$V_m(z_1,z_2)=z_{12}^{m-1}\phi^+(z_1)\phi^-(z_2)-z_{12}^{-1}=\frac{1}{z_{12}}\left(:\exp\left\{\int_{z_2}^{z_1}H(z)\mathrm{d}z\right\}:-1\right); \tag{2.17}$$

it is *quasiprimary* of total dimension 1 in the sense that it satisfies

$$[L_n,V_m(z_1,z_2)]=\sum_{i=1}^{2}z_i^n\left(z_i\frac{\partial}{\partial z_i}+\frac{n+1}{2}\right)V_m(z_1,z_2)\quad \text{for } n=0,\pm1. \tag{2.18}$$

(It is easily verified, in particular, that the differential operators on the right-hand side annihilate, for $n=0,\pm1$, the two-point function z_{12}^{-1}.) Equation (2.17) [with the current H satisfying (2.2)] provides a version of the vertex operator construction, thus allowing one to compute the correlation functions of V_m. We have, in particular, $\langle V_m(z_1,z_2)\rangle_0=0$,

$$\begin{aligned}\langle V_m(z_1,z_2)V_m(z_3,z_4)\rangle_0&=\frac{1}{z_{12}z_{34}}\{(1-\eta)^{-m}-1\}\\&=\frac{m}{z_{13}z_{24}}\sum_{n=0}^{\infty}\binom{m+n}{m}\frac{\eta^n}{n+1},\qquad \eta=\frac{z_{12}z_{34}}{z_{13}z_{24}};\end{aligned} \tag{2.19}$$

$$\begin{aligned}&\langle V_m(z_1,z_2)V_m(z_3,z_4)V_m(z_5,z_6)\rangle_0\\&=(z_{12}z_{34}z_{56})^{-m}\left\{2-\left(\frac{z_{13}z_{24}}{z_{23}z_{14}}\right)^m-\left(\frac{z_{35}z_{46}}{z_{36}z_{45}}\right)^m-\left(\frac{z_{15}z_{26}}{z_{16}z_{25}}\right)^m+\left(\frac{z_{13}z_{24}z_{35}z_{46}z_{15}z_{26}}{z_{23}z_{14}z_{36}z_{45}z_{16}z_{25}}\right)^m\right\}.\end{aligned} \tag{2.20}$$

We shall describe the complete OPE of (2.17) into local Bose fields thus extending to arbitrary $m\in\mathbf{N}$ an earlier result about V_1 (cf. Theorem 2.1 of Ref. 5).

Theoorem 2.1: *The bilocal field (2.17) admits an expansion in integrals over local quasiprimary fields V^l (of dimension $l+1$):*

$$V_m(z_1,z_2)=\sum_{l=0}^{\infty}\frac{z_{12}^l}{l!}\int_{z_2}^{z_1}p_l(z;z_1,z_2)V^l(z,m)\mathrm{d}z, \tag{2.21}$$

where p_l is the (normalized) weight function

$$p_l(z;z_1,z_2)=\frac{(2l+1)!}{(l!)^2z_{12}^{2l+1}}(z_1-z)^l(z-z_2)^l\quad \left(\int_{z_2}^{z_1}p_l(z;z_1,z_2)\mathrm{d}z=1\right). \tag{2.22}$$

The two-point functions of V^l (which display their orthogonality),

$$\frac{l+1}{l!}\langle V^l(z_1,m)V^{l'}(z_2,m)\rangle_0=C_{lm}z_{12}^{-2l-2}\delta_{l,l'}, \tag{2.23}$$

involve the same normalization constant as the three-point function

$$(l+1)\langle V_m(z_1,z_2)V^l(z_3,m)\rangle_0 = C_{lm}\frac{z_{12}^l}{(z_{13}z_{23})^{l+1}}. \tag{2.24}$$

The (positive) constants C_{lm} are uniquely determined by the four-point function (2.19). Conversely, $V^l(z,m)$ are computed from $V_m(z_1,z_2)$ by differentiation:

$$V^l(z,m) = \frac{l!}{(2l)!}\lim_{z_1,z_2\to z}\{\partial_1^l(-\partial_2)^l(z_{12}^l V_m(z_1,z_2))\}. \tag{2.25}$$

Furthermore, V^l are again related to the Schur polynomials $S_n(H,\partial H/2!,\ldots,(\partial^{n-1}H)/n!)$ by (the counterpart of) Eq. (2.12).

The *proof* follows, for the most part the argument in Ref. 5 and we shall only sketch it emphasizing the new points.

The weight function (2.22) satisfies the identity

$$\int_{z_2}^{z_1} p_l(z;z_1,z_2)(z-z_3)^{-2l-2}dz = z_{13}^{-l-1}z_{23}^{-l-1}. \tag{2.26}$$

Since, on the other hand, two- and three-point functions of quasiprimary fields are determined (up to normalization constants) by Möbius (conformal) invariance, this implies

$$\langle V_m(z_1,z_2)V^l(z_3,m)\rangle_0 = \frac{z_{12}^l}{l!}\int_{z_2}^{z_1} p_l(z;z_1,z_2)\langle V^l(z,m)V^l(z_3,m)\rangle_0 \mathrm{dz}. \tag{2.27}$$

Orthogonality of V^l [reflected in (2.23)] is a general property of quasiprimary fields of different dimensions. Inserting the expansion (2.21) for $V_m(z_3,z_4)$ into (2.19) and using (2.24) we end up with a triangular set of linear equations for C_{lm}:

$$(1-\eta)^{-m} = 1+\sum_{l=0}^{\infty} C_{lm}\frac{\eta^{l+1}}{(l+1)!}F(l+1,l+1;2l+2;\eta). \tag{2.28}$$

Here we have used the following representation for the Gauss hypergeometric function

$$F(n,n;2n;\eta) = z_{13}^n z_{24}^n\int_{z_4}^{z_3}\frac{p_{n-1}(z;z_3,z_4)\mathrm{dz}}{(z_1-z)^n(z_2-z)^n} \quad \left(\eta = \frac{z_{12}z_{34}}{z_{13}z_{24}}\right). \tag{2.29}$$

We thus have

$$\sum_{l=0}^{\infty} C_{lm}\frac{\eta^{l+1}}{(l+1)!}\sum_{n=0}^{\infty}\binom{l+n}{n}^2\binom{2l+1+n}{n}^{-1}\eta^n = \sum_{l=0}^{\infty}\binom{m+l}{l+1}\eta^{l+1}. \tag{2.30}$$

The first seven C_{lm} are captured by

$$C_{lm} = m^{l+1}+\binom{l+1}{4}\frac{m^{l-1}}{2l-1}+\binom{l+1}{6}\frac{5l^2-7l+6}{8(2l-1)(2l-3)}m^{l-3} \quad \text{for } 0\leqslant l\leqslant 6. \tag{2.31}$$

On the other hand, for $m=1$, we recover the expression $C_{l1}=(l+1)!\binom{2l}{l}^{-1}$ of Ref. 5. The important point is the existence of a unique solution of the (infinite) system (2.30) $C_{lm}>0$.

Finally, Lemma 1.3 of Ref. 5 holds in this case, too, and we can write

$$V^l(z,m) = \frac{(l!)^2}{(2l)!}\sum_{k=0}^{l}(-1)^k\binom{l}{k}\binom{2l-k}{l}(l-k)!\,\partial^k S_{l-k+1}\left(H,\frac{\partial H}{2!},\ldots,\frac{\partial^{l-k}H}{(l-k+1)!}\right), \tag{2.32}$$

where $S_{n+1}(H,\partial H/2!,\ldots,\partial^n H/(n+1)!)=(H+\partial)^n H$ [cf. Eq. (2.10)] and we omit (here and in the following remark) the normal product sign.

Remark 2.1: In spite of the fact that $V^l(z)=V^l(z,1)$ is expressed by the same Schur polynomials as $V^l(z,m)$—only the argument H is substituted by the normalized current J^0 in $V^l(z,1)$—the quasiprimary fields $V^l(z,m)$ do not close under commutation among themselves (and the unit operator) for $m\geqslant 2$. In effect the fields $V^l(z,m)=m^{(l+1/2)}V^l(z)$ for $l=0,1,2$ generate the whole $W_{1+\infty}$ algebra spanned by $V^l(z)$ (and 1). On the other hand, the fields $V^l(z,m)$ and $V^l(z)$ are linearly independent for $l\geqslant 3$, $m\geqslant 2$. Indeed, we have

$$V^3(z,m)=\frac{H^4(z)}{4}+\frac{1}{20}P_1(H,H',H'')=\frac{m^2}{4}J^4(z)+\frac{m}{20}(2JJ''-3J'^2)(z)(\text{for } H=\sqrt{m}J);$$

$$V^4(z,m)=\frac{1}{5}H^5+\frac{1}{7}HP_1,\ \text{etc.} \tag{2.33}$$

Here P_1 is the first of a series of quasiprimary fields that are quadratic forms P_n of H and its derivatives (each term having dimension $2n+2$):

$$P_n(H,H',\ldots,H^{(2n)})=2\sum_{l=0}^{n-1}(-1)^l a_l H^{(l)}H^{(2n-l)}+(-1)^n a_n (H^{(n)})^2,$$

$$a_l=\prod_{k=0}^{l-1}\binom{2n-1+k}{2}\prod_{k=l}^{n-1}\binom{k+2}{2}. \tag{2.34}$$

III. EXTENSION TO ARBITRARY $C\in$ N: DISCUSSION

A. Additive bilocal fields and lattice RCFT extensions

The construction of Sec. II extends to arbitrary positive integer c due to the following simple observation.

Proposition 3.1: (Ref. 5, Lemma 2.2) *Let $V_i(z_1,z_2)$,$1\leqslant i\leqslant c$ be a commuting set of bilocal fields with the properties of $V_1(z_1,z_2)$* (2.7). *Then their sum generates the $W_{1+\infty}$ algebra for central charge c*:

$$\sum_{i=1}^{c}V_i(z_1,z_2)=\sum_{l=0}^{\infty}\frac{z_{12}^l}{l!}J^l(z_2)\Rightarrow[J^k(z_1),J^l(z_2)]=\sum_{\nu\geqslant 1}\left\{(-1)^\nu\binom{l}{\nu}J^{k+l-\nu}(z_2)\right.$$
$$\left.-\binom{k}{\nu}J^{k+l-\nu}(z_1)\}\,\delta^{(\nu)}(z_{12})\right\}-(-1)^l\frac{k!l!c}{(k+l+1)!}\delta^{(k+l+1)}(z_{12}). \tag{3.1}$$

If we use the bosonized expressions

$$V_i(z_1,z_2)=\frac{1}{z_{12}}\left(:\exp\left\{\int_{z_2}^{z_1}J_i^0(z)\mathrm{d}z\right\}:-1\right)\text{with}[J_i^0(z_1),J_j^0(z_2)]=-\delta'(z_{12})\delta_{ij}, \tag{3.2}$$

then $J^l(z)=\Sigma_{i=1}^c J_i^l(z)$ where J_i^l is expressed as a Schur polynomial of $J_i^0(z)$ and its derivatives

$$J_i^l(z)=l!S_{l+1}\left(J_i^0(z),\frac{\partial J_i^0(z)}{2!},\ldots,\frac{\partial^l J_i^0(z)}{(l+1)!}\right). \tag{3.3}$$

Similarly, the basic quasiprimary fields $V^l(z)$ appear as sums of mutually commuting $c=1$ fields $V_i^l(z,1)$.

Let $W^{(c)}$ be the chiral algebra corresponding to the vacuum representation of $W_{1+\infty}$ of central charge c; it is spanned by (normal ordered) polynomials of $V^l(z)$. The set of RCFT extensions is much richer for $c>1$. It includes, in particular, local charged fields' algebras $\mathscr{A}(L)$ associated with a c-dimensional integral Euclidean charge lattice L. In other words, there is a basis $(\mathbf{q}^1,\ldots,\mathbf{q}^c)$ in the space of $\mathrm{U}(1)^{\times c}$ charges such that the (symmetric) positive definite Gram matrix

$(\mathbf{q}^i|\mathbf{q}^j)$ of inner products has integer entries. To each such lattice corresponds a vertex algebra (for a recent review—see Ref. 11) spanned by charged vertex operators $Y(\mathbf{q},z)$ and ''Cartan currents'' $H^{\mathbf{q}}(z)$ satisfying OPE of the type

$$Y(\mathbf{q},z_1)Y(-\mathbf{q},z_2)=\frac{1}{z_{12}^{|\mathbf{q}|^2}}:\exp\left\{\int_{z_2}^{z_1}H^{\mathbf{q}}(z)\mathrm{d}z\right\}:,\quad H^{\mathbf{q}}(z_1)H^{\mathbf{q}'}(z_2)\sim\frac{(\mathbf{q}|\mathbf{q}')}{z_{12}^2},\tag{3.4}$$

$$Y(\mathbf{q},z_1)Y(\mathbf{q}',z_2)\sim\varepsilon(\mathbf{q},\mathbf{q}')z_{12}^{(\mathbf{q}|\mathbf{q}')}Y(\mathbf{q}+\mathbf{q}',z_2)\quad(\text{for }\ \mathbf{q}+\mathbf{q}'\neq 0),\tag{3.5}$$

where ε is a (± 1) valued cocycle that ensures the correct spin-statistics relation between different vertex operators. If the charge lattice is odd, i.e., if there are vectors $\mathbf{q}\in L$ such that $|\mathbf{q}|^2\equiv(\mathbf{q}|\mathbf{q})$ is odd (which is the case of interest in applications to a quantum Hall system—see Refs. 12–14 and references therein) then the resulting chiral algebra $\mathscr{A}(L)$ is a local $\mathbf{Z}_2$ graded superalgebra. In order to identify a $W^{(c)}$ subalgebra we should choose a one-dimensional subspace in the c-dimensional space of currents, thus singling out the current $J^0(z)=\Sigma_{i=1}^c J_i^0(z)$. This is achieved, in particular, by selecting an ''electron charge'' vector $\mathbf{Q}$ in the dual lattice L^* (whose square $|\mathbf{Q}|^2$ gives the quantum Hall filling fraction[12,13]).

The positive energy representations of the RCFT so obtained, in which all vertex operators $Y(\mathbf{q},z)$, $\mathbf{q}\in L$ are single valued, are in one-to-one correspondence with the elements of the (finite) Abelian group L^*/L of order

$$|L^*/L|=\det((\mathbf{q}^i|\mathbf{q}^j)).$$

Let Γ be any finite group of automorphisms of the chiral algebra $\mathscr{A}(L)$ which leaves each element of its $W^{(c)}$ subalgebra invariant. [It is enough to demand the invariance of the first three ''tensor currents'' $V^l(z), l=0,1,2$ which generate the entire $W_{1+\infty}$ algebra.] There is a wealth of such inner automorphism groups in the case when L is a semisimple Lie group lattice (for integer c it may only correspond to a lattice of the A-D-E type, the basis $\mathbf{q}^i$ being identified with a basis of simple roots). Viewing Γ as a *gauge group* we define the *observable algebra* as the subalgebra $\mathscr{A}(L)^\Gamma$ of Γ invariant elements of $\mathscr{A}(L)$. The resulting chiral algebra $\mathscr{A}(L)^\Gamma$ provides another RCFT extension of $W^{(c)}$ (see Ref. 6).

The characters of positive energy irreducible representations of a lattice current algebra span, in general, a finite dimensional representation of the subgroup of the modular group $\mathrm{SL}(2,\mathbf{Z})$ generated by the matrices

$$S=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\quad\text{and}\quad T^2=\begin{pmatrix}1 & 2\\ 0 & 1\end{pmatrix}.\tag{3.6}$$

[If L is an even lattice then they span a representation of the full $\mathrm{SL}(2,\mathbf{Z})$.] Recently, all weak modular invariants [i.e., invariants under the group generated by (3.6) have been classified—see Ref. 15). On the other hand, given a *nonexceptional gauge group* of inner automorphisms Γ [such is, in particular, any finite subgroup of U(n) or SU(n)], it has been demonstrated in Ref. 6 how to construct the specialized characters (that only capture the energy distribution) of the corresponding $\mathscr{A}(L)^\Gamma$ *orbifold modules* and to exhibit their modular transformation properties.

Remark 3.1: Albeit we thus arrive at a rich family of manageable RCFT extensions of $W_{1+\infty}$, the set of such extensions is not exhaustive—even if we admit outer automorphisms in Γ, for which no general construction of orbifold modules is known. RCFT extensions of a different type are obtained through the so-called conformal embedding.[16–21] A simple example of this type is provided by the level 5 su(3)-current algebra $\mathscr{A}_5(\mathrm{su}(3))$ of Virasoro central charge $c_k=8k/(3+k)=5$ for $k=5$, which can be viewed as the *observable subalgebra* of the rank 5 lattice chiral algebra $\mathscr{A}_1(\mathrm{su}(6))$. Indeed, the vacuum space of $\mathscr{A}_1(\mathrm{su}(6))$ splits into a direct sum of two $\mathscr{A}_5(\mathrm{su}(3))$ modules (with the same central charge $c=5$, and hence, the same Virasoro subalgebra): one with su(3) weight $\Lambda_0=(0,0)$ and ground state energy $\Delta_0=0$ and a second one with $\Lambda_1=(2,2)$ and a degenerate 27-dimensional ground state of energy (=conformal weight) $\Delta_1=\frac{1}{24}\{\lambda_1(\lambda_1+3)+\lambda_2(\lambda_2+3)+\lambda_1\lambda_2\}=1$ for $\lambda_1=\lambda_2=2$. Thus $\mathscr{A}_5(\mathrm{su}(3))$ can be viewed as a RCFT extension of either $W^{(5)}$ or the part of $W^{(6)}$ orthogonal to the u(1) current J^0. It cannot be

viewed as the gauge invariant part of the lattice algebra $\mathscr{A}_1(\mathrm{su}(6))$ with respect to a finite gauge group Γ since then the $\mathscr{A}_5(\mathrm{su}(3))$ module of highest weight Λ_1 should have an integer quantum dimension (equal to the dimension of the associated irreducible representation of Γ—see Ref. 21 and references therein) while it is, in fact, irrational:

$$d_q(\Lambda)=\frac{[3]^2[6]}{[2]}=(1+\sqrt{2})^2 \quad \text{for } q^8=-1.$$

Only for $c=1$ we have listed all RCFT extensions of the u(1) current algebra—cf. Ref. 22—and hence all extensions of $W^{(1)}$.

B. Computing character dependence on chemical potential for orbifold models

It is important, as far as applications to the quantum Hall effect are concerned, to compute the character dependence on the chemical and electric potentials combined in a complex variable ζ (see Ref. 14; the ζ dependence carries information about charge distribution and the so-called spectral flow). In the case of lattice orbifolds this requires the following simple (but important) complement of the algorithm presented in Ref. 6.

To set the stage we reformulate briefly some notions and results of Ref. 6 adapted to the present context.

Let L be an integral Euclidean lattice and let $\mathbf{Q}$ be a fixed vector in the dual lattice L^*, the ''electric charge'' [such a pair $(L,\mathbf{Q})$, satisfying some additional requirements, is called a *chiral quantum Hall lattice* in Ref. 12]. We shall assume, for the sake of convenience, that L is *maximal*, i.e., there is no sublattice of L^* that is a proper extension of L. (This assumption does not restrict the generality of the orbifold chiral algebra we shall end up with.) Let G be a connected compact Lie group whose Lie algebra is generated by the zero-components of all (dimension 1) currents in the chiral algebra $\mathscr{A}(L)$. It can hence be assumed acting by unitary operators in each irreducible (positive energy) $\mathscr{A}(L)$ module. G is a rank $l=\dim L$ Lie group: it has a maximal Abelian subgroup $\mathrm{U}(1)^{\times l}$ where $l=\mathrm{rank}\, L$; the center $\mathbf{Z}_G(\subset \mathrm{U}(1)^{\times l})$ of G is nontrivial: it contains a (finite) subgroup isomorphic to $|L^*/L|$. To any $\mathbf{q}\in L^*$ there corresponds a pair of charged vertex operators $Y(\pm\mathbf{q},z)$ [which intertwine the vacuum representation of $\mathscr{A}(L)$ with the charge $\pm\mathbf{q}$ representations] and hence a u(1) current $H^{\mathbf{q}}(z)$ satisfying (3.4) whose zero-component $H_0^{\mathbf{q}}$-belongs to the Lie algebra $\mathscr{G}$ of G. ($\mathscr{G}$ may also involve raising and lowering operators $E^{\pm\alpha}$ for $\underline{\alpha}\in L$, $|\underline{\alpha}|^2=2$ whenever L contains such vectors $\underline{\alpha}$. Let Γ be a (nonexceptional) finite subgroup of G acting by inner automorphisms on $\mathscr{A}(L)$ and leaving the current $H^{\mathbf{Q}}(z)$ invariant. To each inner automorphism $\rho\in \mathrm{Ad}\,\Gamma$ of $\mathscr{A}(L)$ we make correspond a $\beta=\underline{\beta_\rho}\in i\mathscr{G}$ such that

$$\rho(V)=bVb^{-1}, \quad b=e^{2\pi i\beta}, \quad \Gamma_b=\Gamma_\beta, \tag{3.7}$$

where $\Gamma_b(\Gamma_\beta)$ is the centralizer of $b(\beta)$ in Γ. (We identify vectors $\beta\in L$ with corresponding Lie algebra element H_0^β.)

Let $\mathscr{H}_{\underline{\lambda}}, \underline{\lambda}\in L^*/L$ be an irreducible positive energy $\mathscr{A}(L)$ module. Define its character $\chi_{\underline{\lambda}}$

$$\chi_{\underline{\lambda}}(\tau,\underline{\alpha})=\mathrm{tr}_{\mathscr{H}_{\underline{\lambda}}}(q^{L_0-c/24}e^{2\pi i\underline{\alpha}}), \quad q=e^{2\pi i\tau}, \quad \mathrm{Im}\,\tau>0; \quad i\underline{\alpha}\in\mathscr{G}. \tag{3.8}$$

It follows that

$$\chi_{\underline{\lambda}}(\tau,\underline{\alpha})=\frac{1}{[\eta(\tau)]^l}\sum_{\gamma\in L+\underline{\lambda}} q^{(1/2)(\underline{\gamma}|\underline{\gamma})}e^{2\pi i(\underline{\gamma}|\underline{\alpha})} \quad \left(\eta=q^{1/24}\prod_{n=1}^{\infty}(1-q^n),\ \ l=\mathrm{rank}\, L\right). \tag{3.9}$$

The character of a *β-twisted module* $\mathscr{H}_{\underline{\lambda}}^{\beta}$ is [see Eq. (3.18) of Ref. 6]

$$\chi_{\underline{\lambda}}^{\underline{\alpha},\beta}(\tau)=q^{(1/2)|\underline{\beta}|^2}\chi_{\underline{\lambda}}(\tau,\underline{\alpha}-\tau\beta), \quad [\underline{\alpha},\beta]=0. \tag{3.10}$$

Proposition 3:2: Let σ be (the character of) an irreducible representation of the centralizer $\Gamma_{\underline{\beta}}(=\Gamma_b)$ of β in Γ. Then the projected module $\mathscr{H}^{\beta}_{\underline{\lambda},\sigma}$ is only nontrivial if σ and $\underline{\lambda}$ agree on L^*/L and then the character is

$$\chi^{\beta}_{\underline{\lambda},\sigma}(\tau,\zeta)=\frac{1}{|\Gamma_{\underline{\beta}}|}\sum_{h=e^{2\pi i\alpha}\in\Gamma_{\underline{\beta}}} e^{-2\pi i\zeta(\mathbf{Q}|\beta)}\chi^{\zeta\mathbf{Q}+\alpha,\beta}_{\underline{\lambda}}(\tau)\sigma^*(h) \tag{3.11}$$

$|\Gamma_{\underline{\beta}}|$ *standing for the number of elements of* $\Gamma_{\underline{\beta}}$.

Indeed, defining $\mathscr{H}^{\beta}_{\underline{\lambda},\sigma}$ by

$$\mathscr{H}^{\beta}_{\underline{\lambda},\sigma}=P_{\sigma}\mathscr{H}^{\beta}_{\underline{\lambda}} \quad \text{where} \quad P_{\sigma}=\frac{\sigma(1)}{|\Gamma_{\underline{\beta}}|}\sum_{h\in\Gamma_{\underline{\beta}}}\sigma^*(h)h\,(=P_{\sigma}^2) \tag{3.12}$$

one proves by a simple extension of the argument of Sec. 4.1 of Ref. 6 that

$$\chi^{\beta}_{\underline{\lambda},\sigma}(\tau,\zeta)=\mathrm{tr}_{\mathscr{H}^{\beta}_{\underline{\lambda},\sigma}}(q^{L_0-H_0^{\beta}+(1/2)|\beta|^2-c/24}e^{2\pi i\zeta(H_0^{\mathbf{Q}}-(\mathbf{Q}|\beta))}). \tag{3.13}$$

Instead of reproducing the details we shall work out an example of physical interest (cf., Ref. 23).

Example 3.1: Let L be a two-dimensional orthogonal lattice with basis $\mathbf{e}^1,\mathbf{e}^2$ and Gram matrix

$$((\mathbf{e}^i|\mathbf{e}^j))=\begin{pmatrix}(\mathbf{e}^1|\mathbf{e}^1) & (\mathbf{e}^1|\mathbf{e}^2)\\ (\mathbf{e}^2|\mathbf{e}^1) & (\mathbf{e}^2|\mathbf{e}^2)\end{pmatrix}=\begin{pmatrix}m & 0\\ 0 & 1\end{pmatrix},\quad m\in\mathbf{N}, \tag{3.14}$$

and let $\mathbf{Q}=\mathbf{e}_1^*$ (where $\{\mathbf{e}_i^*\}$ stands for the dual basis in L^* satisfying $(\mathbf{e}_i^*|\mathbf{e}^j)=\delta_i^j$). Consider a $\mathbf{Z}_M$ gauge group in $\mathscr{A}(L)$ generated by an automorphism of type (3.7) with

$$\underline{\beta}=\underline{\beta}_M=\frac{1}{M}(\mu\mathbf{e}_1^*+\mathbf{e}_2^*),\quad 0<\mu<M. \tag{3.15}$$

$\mathscr{A}(L)$ has m (untwisted) representations with partition functions

$$\chi_l(\tau,\underline{\alpha})=K_l(\tau,\alpha^1;m)K_0(\tau,\alpha^2;1)\quad \text{for}\ \ \underline{\alpha}=\alpha^i\mathbf{e}_i^*, \tag{3.16}$$

where K_l are the one-dimensional lattice characters (see, e.g., Ref. 22)

$$K_l(\tau,\alpha;m)=\frac{1}{\eta(\tau)}\sum_{n\in\mathbf{Z}} q^{m/2(n+l/m)^2}e^{2\pi i\alpha(n+l/m)} \tag{3.17}$$

l mod m (for m even a canonical choice is $1-m/2\leqslant l\leqslant m/2$).

The lift $b=e^{2\pi iH_0^{\beta}}$ $(H_0^{\beta}=1/M(\mu H_0^{\mathbf{e}_1^*}+H_0^{\mathbf{e}_2^*}))$ of the automorphism group $\mathbf{Z}_M$ to the $\mathscr{A}(L)$ modules $\mathscr{H}_l$ is $\mathbf{Z}_{mM}$. Setting

$$\begin{gathered}\sigma^*(e^{2\pi i\underline{\alpha}})=e^{-2\pi i(\alpha/m)\sigma}\quad (\text{for}\ \ \underline{\alpha}=\alpha(\mu\mathbf{e}_1^*+\mathbf{e}_2^*)),\\ \underline{\beta}=\beta(\mu\mathbf{e}_1^*+\mathbf{e}_2^*);\quad e^{2\pi i(\alpha/m)},\ e^{2\pi i(\beta/m)}\in\mathbf{Z}_{mM},\quad |\Gamma_{\beta}|=mM\end{gathered} \tag{3.18}$$

we can evaluate the finite sum in (3.11) with the result

$$\chi^{\beta}_{l,\sigma}(\tau,\zeta)=\frac{1}{\eta^2(\tau)}\sum_{n_1,n_2\in\mathbf{Z}} \mathrm{e}^{2\pi i(n_1+(l-\mu\beta)/m)\zeta}\mathbf{q}^{(m/2)(n_1+(l-\mu\beta/m)^2+(1/2)(n_2-\beta)}\delta^{\sigma \bmod mM}_{m(\mu n_1+n_2)+\mu l}. \tag{3.19}$$

It follows that $\sigma-\mu l=0$ mod m, in accord with the first statement in Proposition 3.2, and we can take the Kronecker δ into account setting (for M,μ coprime)

$$n_1=\frac{\sigma-\mu l}{m}+r+Mn,\quad n_2=\frac{\mu l-\sigma}{m}(\mu-1)-\mu r+mn',$$
$$r=0,1,\ldots,M-1\quad (n,n'\in\mathbf{Z}). \tag{3.20}$$

Inserting in (3.19) and using once more (3.17) we find

$$\chi_{l,\sigma}^{\nu/M}(\tau,\zeta)=\sum_{r=0}^{M-1} K_{M[\sigma-(\mu-1)l]-\mu\nu+mMr}(\tau,M\zeta;mM^2)K_{M[(l\mu-\sigma)/m](\mu-1)-\nu-\mu Mr}(\tau,0;M^2). \tag{3.21}$$

Remark 3.2: A nonanalytic factor, $e^{-(\pi/m)(\mathrm{Im}\,\zeta)^2/\mathrm{Im}\,\tau}$, is to be assigned to the character in order to ensure its $V:\zeta\to\zeta+\tau$ covariance (see Ref. 14).

Remark 3.3: Given the automorphism ρ there is a ''gauge'' freedom in the choice of b and β in (3.11) [in the example (3.15) it amounts to adding to β an arbitrary vector of L^*] Theorem 4.2 of Ref. 6 guarantees that the resulting orbifold models do not depend on this choice.

To sum up: Quantum field theory realization of $W_{1+\infty}$ are provided by $U(1)^{\times c}$ current algebras which also allow the construction of a large family of RCFT extensions of such algebras. These leave room for additional physical requirements having in mind specific applications, e.g., to quantum Hall systems.

ACKNOWLEDGMENTS

The present paper is an outgrowth of previous work in collaboration with Bojko Bakalov to whom we are indebted, in particular, for an indication of the relevance of Schur's polynomials in the present context. The work has been supported in part by the Bulgarian Foundation for Scientific Research under Contract No. F-404.

[1] V. Kac and A. Radul, Commun. Math. Phys. **157**, 429–457 (1993).
[2] V. Kac and A. Radul, Transformation Groups **1**, 41–70 (1996).
[3] E. Frenkel, V. Kac, A. Radul, and W. Wang, Commun. Math. Phys. **170**, 337–357 (1995).
[4] H. Awata, M. Fukuma, Y. Matsuo, and S. Odake, Prog. Theor. Phys. Suppl. **118**, 343–373 (1995), Report hep-th/9408158.
[5] B. N. Bakalov, L. S. Georgiev, and I. T. Todorov, in *New Trends in Quantum Field Theory*, Proceedings of the 1995 Razlog (Bulgaria) Workshop, edited by A. Ganchev *et al.* (Heron, Sofia, 1996), pp. 147–158 Report hep-th/9512078.
[6] V. G. Kac and I. T. Todorov, Commun. Math. Phys. **190**, 57–111 (1997).
[7] G. F. Dell' Antonio, V. Frishman, and D. Zwanziger, Phys. Rev. D **6**, 988–1007 (1972).
[8] V. Kac and A. K. Raina, *Bombay Lectures on Highest Weight Representations of Infinite Dimensional Lie Algebras*, in Advanced Series in Mathematical Physics Vol. 2 (World Scientific, Singapore, 1987).
[9] B. Bakalov, E. Horozov, and M. Yakimov (unpublished); Report q-alg/9602012.
[10] D. Buchholz, G. Mack, and I. T. Todorov, Nucl. Phys. B (Proc. Suppl.) **5B**, 20 (1988).
[11] V. G. Kac, in *New Trends in Quantum Field Theory*, in Ref. 5, pp. 261–358; *Vertex Algebras for Beginners*, AMS University Lecture Series Vol. 10, (AMS University Press, 1997), p. 141.
[12] J. Fröhlich and M. Thiran, J. Stat. Phys. **76**, 209 (1994); J. Fröhlich, U. M. Studer, and M. Thiran, Report cond-mat/9503113.
[13] A. Cappelli, C. A. Trugenberger, and G. R. Zemba, Nucl. Phys. B **448**, 470 (1995).
[14] A. Cappelli and G. R. Zemba, Nucl. Phys. B: Field Theory Stat. Syst. **FS 490**, 595–632 (1997).
[15] T. Gannon, Report hep-th/9608063.
[16] A. N. Schellekens and N. P. Warner, Phys. Rev. D **34**, 3092 (1986).
[17] P. Bouwknegt and W. Nahm, Phys. Lett. B **184**, 359 (1987).
[18] J. R. C. Arcari, J. F. Gomez, and D. I. Olive, Nucl. Phys. B **285**, 327 (1987).
[19] Ya. S. Stanev, J. Math. Phys. **36**, 2070 (1995).
[20] I. T. Todorov, in *Mathematical Physics Towards the 21st Century*, edited by R. N. Sen and A. Gersten (ben-Gurion University Press, Beer-Sheva, 1994), pp. 160–176.
[21] K.-H. Rehren, Ya. S. Stanev, and I. T. Todorov, Commun. Math. Phys. **174**, 605 (1996).
[22] R. R. Paunov and I. T. Todorov, Phys. Lett. B **196**, 519–526 (1987).
[23] A. Cappelli, L. S. Georgiev, and I. T. Todorov, *A unified CFT description of U=1/2 double layer quantum Hall states* (unpublished).

A histories approach to quantum mechanics

Stanley Gudder
Department of Mathematics and Computer Science, University of Denver, Denver, Colorado 80208

(Received 8 June 1998; accepted for publication 14 July 1998)

A histories approach to quantum mechanics is formulated without the consistency requirement. Plausible, physically motivated axioms for a history structure are presented. In this structure, the consistency requirement is unnecessary because the joint sequential distributions are already probability measures. It is shown that the logic of a history structure is a temporal effect algebra and conversely, any temporal effect algebra is isomorphic to the logic of a history structure. When this general framework is specialized to a Hilbert space context, it reduces to a previously studied formalism. © *1998 American Institute of Physics.*
[S0022-2488(98)04411-9]

I. INTRODUCTION

The consistent histories approach to quantum mechanics has been extensively studied in recent years.[1–8] In this article we shall delete the adjective ''consistent'' and refer to our framework as a histories approach to quantum mechanics. This is because we believe that consistent sets of histories are too restrictive and rarely occur. Even when they do occur, we question their usefulness for describing quantum situations. However, in the absence of the consistency requirement, this approach may provide an important method for formulating and interpreting quantum mechanics. Besides their emphasis on consistency and decoherence, some of these previous works have studied temporal and logical structures in the histories framework.[1,4,5,7,8] Most recently this approach has been extended to include generalized observables given by POV measures and the generalized logics of effect algebras.[7,8] The goal of the present article is to provide a simplification and unification of these previous works without the consistency requirement. Moreover, we argue that consistent sets of histories are unnecessary in a certain sense. Unlike previous works in which the ''logic'' of a history structure is essentially postulated, we shall derive the ''logic'' from more basic principles.

In Sec. II we first present some plausible and physically motivated axioms for a history structure. Although the axioms are completely different, they are motivated by Mackey's quantum logic approach.[9] The general framework and two of the axioms were first introduced by Medina.[10] The existence of sequential joint distributions and distributions of observables follow from the axioms. Conditional probabilities and conditional sequential joint distributions are then defined. Various manifestations of quantum interference are discussed.

The fundamentals of σ-effect algebras and their associated states and observables are reviewed in Sec. III. A temporal effect algebra is a σ-effect algebra that possesses a temporal product that satisfies certain conditions. The basic properties of temporal effect algebras and their automorphism groups are discussed.

Section IV derives the quantum logic associated with a history structure $\mathscr{H}$. Histories that occur with the same probability for every state of the system form an equivalence class that we call a history effect and the set of history effects $\mathscr{E}$ is called the logic of $\mathscr{H}$. It is shown that $\mathscr{E}$ has the algebraic structure of a temporal effect algebra. Conversely, it is shown that any temporal effect algebra is isomorphic to the logic of a history structure. We also prove that states and observables in a history structure $\mathscr{H}$ are represented by corresponding states and observables on the logic of $\mathscr{H}$. Conditional probabilities and quantum stochastic processes on a temporal effect algebra are discussed.

Section V considers Hilbert space history structures. If H is a complex Hilbert space and $\mathscr{E}(H)$ is the set of Hilbert space effects on H, then a natural temporal product can be defined on $\mathscr{E}(H)$ so that $\mathscr{E}(H)$ becomes a temporal effect algebra. We then carry over the concepts developed

 © 1998 American Institute of Physics

in Sec. IV and show that they reduce to a familiar form in $\mathcal{E}(H)$. Decoherence functions are briefly discussed and some observations about them are given.

II. HISTORY STRUCTURES

Denote the σ-algebra of Borel subsets of $\mathbb{R}^n$ by $\mathcal{B}(\mathbb{R}^n)$. A *countable Borel partition* of $\mathbb{R}^n$ is a sequence $A_i \in \mathcal{B}(\mathbb{R}^n)$ such that $A_i \cap A_j = \varnothing$ for $i \neq j$ and $\cup A_i = \mathbb{R}^n$. We denote the set of countable Borel partitions of $\mathbb{R}^n$ by $\mathcal{P}(\mathbb{R}^n)$. If $\mathcal{N} \in \mathcal{P}(\mathbb{R}^n)$, then the σ-algebra generated by $\mathcal{N}$ is denoted by $\mathcal{A}(\mathcal{N})$. Notice that $\mathcal{A}(\mathcal{N})$ consists of all unions of sets in $\mathcal{N}$ and that $\mathcal{A}(\mathcal{N}) \subseteq \mathcal{B}(\mathbb{R}^n)$. It is easy to check that if $\mathcal{N}_i \in \mathcal{P}(\mathbb{R})$, $i=1,...,n$, then

$$\mathcal{M} = \{A_1 \times \cdots \times A_n : \quad A_i \in \mathcal{N}_i, i=1,...,n\} \in \mathcal{P}(\mathbb{R}^n)$$

and we use the notation $\mathcal{A}(\mathcal{M}) = \mathcal{A}(\mathcal{N}_1, \ldots, \mathcal{N}_n)$. We denote the set of probability measures on $(\mathbb{R}^n, \mathcal{A}(\mathcal{N}))$, $\mathcal{N} \in \mathcal{P}(\mathbb{R}^n)$, by $M_1^+(\mathbb{R}^n, \mathcal{A}(\mathcal{N}))$ and use the notation

$$M_1^+ = \cup \{M_1^+(\mathbb{R}^n, \mathcal{A}(\mathcal{N})) : n \in \mathbb{N}, \mathcal{N} \in \mathcal{P}(\mathbb{R}^n)\}.$$

Let $\mathcal{O}$ be a set of observables for a physical system. For $X \in \mathcal{O}$, $A \in \mathcal{B}(\mathbb{R})$, we may view (X,A) as the event for which X has a value in the set A. We do not assume that this event is sharp. It may be unsharp due to an imprecision in a measurement apparatus or in the case of a closed system in which no measurement is performed, the unsharpness may be due to an intrinsic property of the observable X. If $\mathcal{N} \in \mathcal{P}(\mathbb{R})$, we may view $(X, \mathcal{N})$ as a set of possible alternative events that can occur. One interpretation is that $(X, \mathcal{N})$ represents an apparatus for measuring X and the discrete alternatives result from the dial readings of the apparatus. In the absence of a measurement, we may interpret $(X, \mathcal{N})$ as a course-grained set of alternative events.

For $X_i \in \mathcal{O}$, $A_i \in \mathcal{B}(\mathbb{R})$, $i=1,...,n$, we call $h = ((X_1, A_1), \ldots, (X_n, A_n))$ a *history* or *temporal event* and interpret h as an ordered temporal sequence in which X_i has a value in A_i at time τ_i where $\tau_1 < \tau_2 < \cdots < \tau_n$. Alternatively, no specific times may be involved and we regard h as a sequential event in which (X_1, A_1) occurs first, (X_2, A_2) occurs second,..., and (X_n, A_n) occurs last. Notice that the A_i could be singleton sets $A_i = \{a_i\}$ in which case $((X_1, \{a_1\}), \ldots, (X_n, \{a_n\}))$ would be a ''path'' in which X_i has value a_i, $i=1,...,n$. We denote the set of all histories of length n by $\mathcal{H}^n = (\mathcal{O} \times \mathcal{B}(\mathbb{R}))^n$ and the set of all histories by $\mathcal{H} = \cup \{\mathcal{H}^n : n \in \mathbb{N}\}$.

A *history set* is an element of $\hat{\mathcal{H}}^n = (\mathcal{O} \times \mathcal{P}(\mathbb{R}))^n$ for some $n \in \mathbb{N}$ and an $s \in \hat{\mathcal{H}}^n$ has the form $s = ((X_1, \mathcal{N}_1), \ldots, (X_n, \mathcal{N}_n))$ where $X_i \in \mathcal{O}$, $\mathcal{N}_i \in \mathcal{P}(\mathbb{R})$, $i=1,...,n$. We interpret s as a temporal sequence with partitions $\mathcal{N}_i$ in the temporal order $\tau_1 < \tau_2 < \cdots < \tau_n$ or as ordered successive measurements of observables X_i with partitions $\mathcal{N}_i$, $i=1...,n$. We use the notation $\mathcal{A}_s = \mathcal{A}(\mathcal{N}_1, \ldots, \mathcal{N}_n)$ and denote the collection of all history sets by $\hat{\mathcal{H}} = \cup \{\hat{\mathcal{H}}^n : n \in \mathbb{N}\}$. If $h = ((X_1, A_1), \ldots, (X_n, A_n))$ and $s = ((X_1, \mathcal{N}_1), \ldots, (X_n, \mathcal{N}_n))$ with $A_i \in \mathcal{N}_i$, $i=1,...,n$, we write $h \in s$ and say that the history h is an element of the history set s. Our history corresponds to a history filter or homogeneous history and our history set corresponds to a complete set of histories in the literature. We call $(X,A) \in \mathcal{H}^1$ a *one-time event* and $(X, \mathcal{N}) \in \hat{\mathcal{H}}^1$ a *one-time history set*. If $s = ((X_1, \mathcal{N}_1), \ldots, (X_n, \mathcal{N}_n))$, $t = ((Y_1, \mathcal{M}_1), \ldots, (Y_n, \mathcal{M}_n))$ are history sets, we define their *sequential product* by

$$st = ((X_1, \mathcal{N}_1), \ldots, (X_n, \mathcal{N}_n), (Y_1, \mathcal{M}_1), \ldots, (Y_n, \mathcal{M}_n)).$$

Of course, $st \in \hat{\mathcal{H}}$. Notice that $\hat{\mathcal{H}}$ becomes a semigroup under this product and by convention we assume that $\varnothing \in \hat{\mathcal{H}}$ and that $s\varnothing = \varnothing s = s$.

If the system is in a state α and $h = ((X_1, A_1), \ldots, (X_n, A_n)) \in \mathcal{H}$ with $h \in s$, $s \in \hat{\mathcal{H}}$, then the probability that h occurs within s is denoted by $P_{\alpha,s}(A_1 \times \cdots \times A_n)$. We now propose an axiomatic framework for a statistical theory of histories. All of the axioms will concern properties of the probabilities $P_{\alpha,s}$.

Let $\mathcal{S}$ be the set of states and $\mathcal{O}$ the set of observables for a physical system. A *history structure* is a triple $(\mathcal{S}, \mathcal{O}, P)$ where $P : \mathcal{S} \times \hat{\mathcal{H}} \to M_1^+$, $(\alpha, s) \mapsto P_{\alpha,s}$ and $P_{\alpha,s} \in M_1^+(\mathbb{R}^n, \mathcal{A}_s)$ when $s \in \hat{\mathcal{H}}^n$. We postulate that $(\mathcal{S}, \mathcal{O}, P)$ satisfies the following axioms.

(H1) If $s_i \in \hat{\mathscr{H}}^1$ and $P_{\alpha,s_i}(A_i) \leq P_{\alpha,s_{i+1}}(A_{i+1})$ for every $\alpha \in \mathscr{S}$, then there exists an $s \in \hat{\mathscr{H}}^1$ and $B_i \in \mathscr{A}_s$ with $B_i \subseteq B_{i+1}$ such that $P_{\alpha,s_i}(A_i) = P_{\alpha,s}(B_i)$, $i = 1,2,...$, for every $\alpha \in \mathscr{S}$.

(H2) Let $s_1, s_2, t_1, t_2 \in \hat{\mathscr{H}}^1$. If $P_{\alpha,s_1}(A_1) = P_{\alpha,s_2}(A_2)$ and $P_{\alpha,t_1}(B_1) = P_{\alpha,t_2}(B_2)$ for every $\alpha \in \mathscr{S}$, then

$$P_{\alpha,s_1t_1}(A_1 \times B_1) = P_{\alpha,s_2t_2}(A_2 \times B_2)$$

for every $\alpha \in \mathscr{S}$.

(H3) If $s = ((X_1, \mathscr{N}_1), \ldots, (X_n, \mathscr{N}_n)) \in \hat{\mathscr{H}}$ and $A_i \in \mathscr{N}_i$ $i = 1,...,n$, then there exists a $t = (X, \mathscr{N}) \in \hat{\mathscr{H}}^1$ and an $A \in \mathscr{N}$ such that

$$P_{\alpha,t}(A) = P_{\alpha,s}(A_1 \times \cdots \times A_n)$$

for every $\alpha \in \mathscr{S}$.

(H4) If $s \in \hat{\mathscr{H}}$, $t = (X, \mathscr{N}) \in \hat{\mathscr{H}}^1$, $u = (X, \mathscr{M}) \in \hat{\mathscr{H}}^1$, $A \in \mathscr{A}_s$, $B \in \mathscr{A}_t \cap \mathscr{A}_u$, then $P_{\alpha,st}(A \times B) = P_{\alpha,su}(A \times B)$ for every $\alpha \in \mathscr{S}$. Moreover, $P_{\alpha,st}(A \times \mathbb{R}) = P_{\alpha,s}(A)$ for every $\alpha \in \mathscr{S}$.

(H5) If $s = ((X_1, \mathscr{N}_1), \ldots, (X_n, \mathscr{N}_n)) \in \hat{\mathscr{H}}$, $t = ((X_1, \mathscr{M}_1), \ldots, (X_n, \mathscr{M}_n)) \in \hat{\mathscr{H}}$ and $A_i \in \mathscr{N}_i \cap \mathscr{M}_i$, then

$$P_{\alpha,s}(A_1 \times \cdots \times A_n) = P_{\alpha,t}(A_1 \times \cdots \times A_n)$$

for every $\alpha \in \mathscr{S}$.

If $s = ((X_1, \mathscr{N}_1), \ldots, (X_n, \mathscr{N}_n)) \in \hat{\mathscr{H}}$ and $\alpha \in \mathscr{S}$, we call $P_{\alpha,s} \in M_1^+(\mathbb{R}^n, \mathscr{A}_s)$ the *sequential joint distribution of* $X_1, \ldots, X_n$ *with partitions* $\mathscr{N}_1, \ldots, \mathscr{N}_n$ *in the state* α. The main reason for requiring a consistent set of histories in the usual approach is to ensure that $P_{\alpha,s}$ is indeed a probability measure. In the present approach this is not necessary because we are already guaranteed that $P_{\alpha,s}$ is a probability measure. This may appear to be circumventing a problem by decree. However, the point here is that a histories approach with proper statistical properties can be developed without the need for decoherence functions and a consistency requirement.

We now give physical interpretations for the history structure axioms. Axiom (H1) says that an increasing sequence of one-time events can be described by a single observable. In this sense, one-time events in an increasing sequence are compatible. Axiom (H2) says that if pairs of one-time events have the same probabilities of occurrence for every state, then so do the corresponding two-time histories. Our interpretation of (H3) is that any history can be described by a one-time event. This is in accordance with the following statement:[7] ''All meaningful propositions about a physical system specify that the value of some observable lies in some set $A \in \mathscr{B}(\mathbb{R})$.'' The first part of (H4) means that the partition of the last measurement is irrelevant and the second part means that the last measurement does not affect the previous ones. This axiom as well as parts of this general framework were given in Ref. 10. Axiom (H5) says that the probability of a history does not depend on the partitions and hence is unambiguous. However, if s and t are as in (H5) and $A \in \mathscr{A}_s \cap \mathscr{A}_t$, then in general $P_{\alpha,s}(A) \neq P_{\alpha,t}(A)$. This is an unavoidable manifestation of quantum interference and we shall mention other such manifestations in the sequel. (Section V discusses examples in Hilbert space.)

It follows from (H4) by letting $s = \varnothing$ that if $t = (X, \mathscr{N})$, $u = (X, \mathscr{M})$ and $B \in \mathscr{A}(\mathscr{N}) \cap \mathscr{A}(\mathscr{M})$, then $P_{\alpha,t}(B) = P_{\alpha,u}(B)$ for every $\alpha \in \mathscr{S}$. Hence, the probability that X has a value in B is independent of the partition that generates B. The next lemma is stated without proof in Ref. 10.

Lemma 2.1: If $\alpha \in \mathscr{S}$, $X \in \mathscr{O}$, then there exists a unique probability measure $P_{\alpha,X}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $P_{\alpha,X}(A) = P_{\alpha,s}(A)$ for every $s = (X, \mathscr{N})$ and $A \in \mathscr{A}(\mathscr{N})$.

Proof: For $A \in \mathscr{B}(\mathbb{R})$, define $P_{\alpha,X}(A) = P_{\alpha,s}(A)$, where $s = (X, \mathscr{N})$ with $A \in \mathscr{A}(\mathscr{N})$. Our previous observation shows that $P_{\alpha,X}$ is well defined. Then $P_{\alpha,X}(A) \geqslant 0$ for every $A \in \mathscr{B}(\mathbb{R})$ and $P_{\alpha,X}(\mathbb{R}) = 1$. If $A_i \in \mathscr{B}(\mathbb{R})$ are mutually disjoint, $i = 1,2,...$, let $\mathscr{N} \in \mathscr{P}(\mathbb{R})$ contain each A_i, $i = 1,2,...$. If $s = (X, \mathscr{N})$, then

$$P_{\alpha,X}(\cup A_i) = P_{\alpha,s}(\cup A_i) = \sum P_{\alpha,s}(A_i) = \sum P_{\alpha,X}(A_i).$$

Hence, $P_{\alpha,X}$ is a probability measure on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ and it is clear that $P_{\alpha,X}$ is unique. □

We call $P_{\alpha,X}$ the *distribution of X in the state* α. The proof of the following result is similar to that of Lemma 2.1.

Corollary 2.2: If $\alpha\in\mathscr{S}$, $s\in\hat{\mathscr{H}}$, $A\in\mathscr{A}_s$, $X\in\mathscr{O}$, then there exists a unique probability measure $P_{\alpha,s,X}$ such that $P_{\alpha,s,X}(A\times B)=P_{\alpha,s,t}(A\times B)$ for every $t=(X,\mathscr{N})$ and $B\in\mathscr{A}(\mathscr{N})$.

The next lemma follows directly from (H1).

Lemma 2.3: If $P_{\alpha,X_i}(A_i)\leqslant P_{\alpha,X_{i+1}}(A_{i+1})$ for every $\alpha\in\mathscr{S}$, then there exists an $X\in\mathscr{O}$ and $B_i\in\mathscr{B}(\mathbb{R})$ with $B_i\subseteq B_{i+1}$ such that $P_{\alpha,X}(B_i)=P_{\alpha,X_i}(A_i)$ for every $\alpha\in\mathscr{S}$, $i=1,2,\ldots$.

By iterating (H4) we have the following result:[10]

Lemma 2.4: If $\alpha\in\mathscr{S}$, $s\in((X_1,\mathscr{N}_1),\ldots,(X_n,\mathscr{N}))\in\hat{\mathscr{H}}$ and $A\in\mathscr{A}(\mathscr{N}_1)$, then $P_{\alpha,s}(A\times\mathbb{R}^{n-1})=P_{\alpha,X_1}(A)$.

In general, the other marginal distributions are not equal to the one-variable distributions. That is, in general for $k>1$, $s=((X_1,\mathscr{N}_1),\ldots,(X_n,\mathscr{N}_n))$ and $A\in\mathscr{A}(\mathscr{N}_k)$,

$$P_{\alpha,s}(\mathbb{R}^{k-1}\times A\times\mathbb{R}^{n-k})\neq P_{\alpha,X_k}(A).$$

This is another manifestation of quantum interference.

If $s,t\in\hat{\mathscr{H}}$, $A\in\mathscr{A}_s$, $B\in\mathscr{A}_t$ with $P_{\alpha,s}(A)\neq 0$, we defined the *conditional probability of B given A for st in the state* α by

$$P_{\alpha,t|s}(B|A)=\frac{P_{\alpha,st}(A\times B)}{P_{\alpha,s}(A)}.$$

It is clear that if $t\in\hat{\mathscr{H}}^m$, then $P_{\alpha,t|s}(\cdot\,|A)$ is a probability measure on $(\mathbb{R}^m,\mathscr{A}_t)$ which we call a *conditional sequential joint distribution*. The next lemma shows that these distributions have the fundamental properties possessed by sequential joint distributions.

Lemma 2.5: For every $\alpha\in\mathscr{S}$, $s\in\hat{\mathscr{H}}$, $P_{\alpha,\cdot|s}(\cdot\,|A)$ satisfies (H4) and (H5).

Proof: Applying (H4) we have

$$P_{\alpha,tu|s}(B\times C|A)=\frac{P_{\alpha,stu}(A\times B\times C)}{P_{\alpha,s}(A)}=\frac{P_{\alpha,stv}(A\times B\times C)}{P_{\alpha,s}(A)}=P_{\alpha,tv|s}(B\times C|A).$$

Thus, $P_{\alpha,\cdot|s}(\cdot\,|A)$ satisfies the first part of (H4). For the second part of (H4) we have

$$P_{\alpha,tu|s}(B\times\mathbb{R}|A)=\frac{P_{\alpha,stu}(A\times B\times\mathbb{R})}{P_{\alpha,s}(A)}=\frac{P_{\alpha,st}(A\times B)}{P_{\alpha,s}(A)}=P_{\alpha,t|s}(B|A).$$

The proof of (H5) is similar. □

The next result which follows from Corollary 2.2 shows that observables have conditional distributions on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$.

Corollary 2.6: If $\alpha\in\mathscr{S}$, $X\in\mathscr{O}$ and $P_{\alpha,s}(A)\neq 0$, then there exists a unique probability measure $P_{\alpha,X}(\cdot\,|A)$ on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ such that $P_{\alpha,X}(B|A)=P_{\alpha,t|s}(B|A)$ for every $t=(X,\mathscr{N})$ and $B\in\mathscr{A}(\mathscr{N})$.

Another manifestation of quantum interference is the fact that Bayes' formula need not hold,

$$P_{\alpha,t|s}(B|A)P_{\alpha,s}(A)\neq P_{\alpha,s|t}(A|B)P_{\alpha,t}(B).$$

This is because the left-hand side is $P_{\alpha,st}(A\times B)$ and the right-hand side is $P_{\alpha,ts}(B\times A)$ so the order of the measurements or the order of the occurrences is changed. Also the other Bayes' formula need not hold,

$$\sum_{i=1}^{r}P_{\alpha,s}(A_i)P_{\alpha,t|s}(B|A_i)\neq P_{\alpha,t}(B),$$

where $\{A_1,\ldots,A_r\}\in\mathscr{P}(\mathbb{R}^n)$, $s\in\hat{\mathscr{H}}^n$. This is because the left-hand side is

$$\sum_{i=1}^{r} P_{\alpha,st}(A_i\times B)=P_{\alpha,st}(\cup A_i\times B)=P_{\alpha,st}(\mathbb{R}^n\times B)$$

and $P_{\alpha,st}(\mathbb{R}^n\times B)\neq P_{\alpha,t}(B)$ in general. The last inequality comes from the fact that past measurements can affect future ones.

For $h=((X_1,A_1),\ldots,(X_n,A_n))\in\mathscr{H}$ and $\alpha\in\mathscr{S}$, we define $P_\alpha(h)=P_{\alpha,s}(A_1\times\cdots\times A_n)$ for any $s\in\hat{\mathscr{H}}$ with $h\in s$ and (H5) ensures that $P_\alpha(h)$ is well defined. An *automorphism* on $\mathscr{O}$ is a bijection $\psi\colon\mathscr{O}\to\mathscr{O}$ that satisfies

$$P_\alpha((\psi X,A))=P_\alpha((\psi Y,B),(\psi Z,C))$$

and

$$P_\alpha((\psi^{-1}X,A))=P_\alpha((\psi^{-1}Y,B),(\psi^{-1}Z,C))$$

for all $\alpha\in\mathscr{S}$ whenever $P_\alpha((X,A))=P_\alpha((Y,B),(Z,C))$ for all $\alpha\in\mathscr{S}$. Thus, if a two-time history is described by a one-time event, then ψ and ψ^{-1} preserve this relationship. Denoting the set of automorphisms on $\mathscr{O}$ by aut($\mathscr{O}$), it is clear that aut($\mathscr{O}$) is a group under composition. A *dynamical group* on $\mathscr{O}$ is a group homomorphism $\tau\mapsto\psi(\tau)$ from the additive group $\mathbb{R}$ into aut($\mathscr{O}$). We thus have $\psi(\tau_1+\tau_2)=\psi(\tau_1)\circ\psi(\tau_2)$ for all $\tau_1,\tau_2\in\mathbb{R}$ and $\psi(0)=I$, where I is the identity automorphism. We interpret $\psi(\tau)X$ as the evolution of the observable X. If $X_1,\ldots,X_n$ are observables at time $\tau=0$ and $\tau_1<\ldots<\tau_n$, we interpret

$$((\psi(\tau_1)X_1,A_1),\ldots,(\psi(\tau_n)X_n,A_n))$$

as a history in which X_1 has a value in A_1 at time $\tau_1,\ldots,X_n$ has a value in A_n at time τ_n.

III. TEMPORAL EFFECT ALGEBRAS

An *effect algebra*[8,11–16] is an algebraic system $(E,0,1,\oplus)$ where $0,1\in E$ and $\oplus$ is a partial binary operation on E that satisfies

(E1) if $a\oplus b$ is defined, then $b\oplus a$ is defined and $b\oplus a=a\oplus b$;
(E2) if $a\oplus b$ and $(a\oplus b)\oplus c$ are defined, then $b\oplus c$ and $a\oplus(b\oplus c)$ are defined and $a\oplus(b\oplus c)=(a\oplus b)\oplus c$;
(E3) for every $a\in E$, there exists a unique $a'\in E$ such that $a\oplus a'$ is defined and $a\oplus a'=1$;
(E4) if $a\oplus 1$ is defined, then $a=0$.

We call $a\oplus b$ an *orthogonal sum*. It follows from (E2) that existing orthogonal sums do not require parentheses and we write $a\oplus b\oplus c$ for $(a\oplus b)\oplus c$. More generally, we write $a_1\oplus\cdots\oplus a_n$ if this orthogonal sum is defined. We define $a\leqslant b$ if there exists a $c\in E$ such that $a\oplus c=b$. It can be shown[13] that $(E,0,1,\leqslant)$ is a bounded poset and $a\oplus b$ is defined if and only if $a\leqslant b'$. If $a\leqslant b'$, we write $a\perp b$. If $a\leqslant b$, then the unique c that satisfies $a\oplus c=b$ is denoted by $c=b\ominus a$. If the least upper bound $\vee a_i$ exists whenever $a_1\leqslant a_2\leqslant\cdots$, then E is called a *σ-effect algebra*.[8] Suppose $b_i=a_i\oplus\cdots\oplus a_i$ is defined for all $i\in\mathbb{N}$. If $\vee b_i$ exists then we say that $\oplus a_i$ exists and we define $\oplus a_i$ to be $\vee b_i$.

Lemma 3.1: An effect algebra E is a σ-effect algebra if and only if $\oplus a_i$ exists whenever $a_1\oplus\cdots\oplus a_i$ is defined for all $i\in\mathbb{N}$.

Proof: If E is a σ-effect algebra and $b_i=a_1\oplus\cdots\oplus a_i$ is defined for all $i\in\mathbb{N}$, then $b_1\leqslant b_2\leqslant\ldots$, so $\vee b_i$ exists. Conversely, suppose $\oplus a_i$ exists whenever $a_1\oplus\cdots\oplus a_i$ is defined for all $i\in\mathbb{N}$ and let $b_1\leqslant b_2\leqslant\cdots$. Defining $a_1=b_1$, $a_i=b_i\ominus b_{i-1}$, $i=2,3,\ldots$, it easily follows that $a_1\oplus\cdots\oplus a_i=b_i$ exists for all $i\in\mathbb{N}$. Hence, $\vee b_i=\oplus a_i$, exists so E is a σ-effect algebra. □

Let E and F be σ-effect algebras. A map $\phi\colon E\to F$ is *additive* if $a\perp b$ implies $\phi(a)\perp\phi(b)$ and $\phi(a\oplus b)=\phi(a)\oplus(b)$. If ϕ is additive, then ϕ is *monotone* in the sense that $a\leqslant b$ implies $\phi(a)\leqslant\phi(b)$. Indeed, if $a\leqslant b$ then $b=a\oplus c$ for some $c\in E$ so that

$$\phi(a)\leqslant\phi(a)\oplus\phi(c)=\phi(a\oplus c)=\phi(b).$$

An additive map $\phi\colon E\to F$ is *σ-additive* if $a_1\leqslant a_2\leqslant\cdots$ implies that $\phi(\vee a_i)=\vee\phi(a_i)$. The proof of the next result is similar to that of Lemma 3.1.

Lemma 3.2: A map $\phi\colon E\to F$ is σ-additive if and only if $\phi(\oplus a_i)=\oplus\phi(a_i)$ whenever $\oplus a_i$ exists.

A σ-additive map $\phi\colon E\to F$ is a *σ-morphism* if $\phi(1)=1$. It is clear that a σ-morphism satisfies $\phi(a')=\phi(a)'$. If $\phi\colon E\to F$ is a bijective σ-morphism and ϕ^{-1} is a σ-morphism, then ϕ is a *σ-isomorphism*.

An important example of a σ-effect algebra is the unit interval $[0,1]\subseteq\mathbb{R}$. For $a,b\in[0,1]$ we say that $a\oplus b$ is defined if $a+b\leqslant 1$ and in this case $a\oplus b=a+b$. We then have $a'=1-a$, 0 is the zero element and 1 is the one element. A σ-morphism from a σ-effect algebra E into $[0,1]$ is called a *state* on E. We denote the set of states on E by $\mathscr{S}(E)$. We say that $S\subseteq\mathscr{S}(E)$ is *order determining* if $\alpha(a)\leqslant\alpha(b)$ for all $\alpha\in S$ implies that $a\leqslant b$.

Another example of a σ-effect algebra is a measurable space $(\Omega,\mathscr{A})$. For $A,B\in\mathscr{A}$ we say that $A\oplus B$ is defined if $A\cap B=\varnothing$ and in this case $A\oplus B=A\cup B$. We then have A' equal to the complement of A and $0=\varnothing$, $1=\Omega$. A σ-morphism from $(\Omega,\mathscr{A})$ into a σ-effect algebra E is called an *observable* on E with *value space* $(\Omega,\mathscr{A})$. We denote the set of such observables by $\mathscr{O}_{\mathscr{A}}(E)$. Notice that $\mathscr{S}(\Omega,\mathscr{A})$ is precisely the set of probability measures on $(\Omega,\mathscr{A})$. If $X\in\mathscr{O}_{\mathscr{A}}(E)$ and $\alpha\in\mathscr{S}(E)$, then $\alpha\circ X\in\mathscr{S}(\Omega,\mathscr{A})$ is called the *distribution* of X in the state α.

The elements of a σ-effect algebra are called *effects*. If $\alpha\in\mathscr{S}(E)$, then $\alpha(a)$ is the probability that a occurs when the system is in the state α. For $X\in\mathscr{O}_{\mathscr{A}}(E)$, $X(A)$ is the effect that X has a value in A and $\alpha\circ X(A)$ is the probability that $X(A)$ occurs in the state α. We interpret $a\oplus b$ as a statistical sum in the sense that $a\oplus b$ is the effect that satisfies $\alpha(a\oplus b)=\alpha(a)+\alpha(b)$ for every state α.

A *temporal product* on a σ-effect algebra E is a binary operation $(a,b)\mapsto a\circ b$ on E such that $b\mapsto a\circ b$ is σ-additive and $0\circ a=a$, $a\circ 1=a$ for every $a\in E$. Notice that $a\circ 0=0$ for every $a\in E$ because

$$a=a\circ 1=a\circ(0\oplus 1)=a\circ 0\oplus a\circ 1=a\circ 0\oplus a.$$

Hence, by cancellation[13] $a\circ 0=0$. We interpret $a\circ b$ as the effect in which a first occurs and then b occurs. A natural temporal product on the σ-effect algebra $(\Omega,\mathscr{A})$ is given by $A\circ B=A\cap B$. The next lemma shows that there is only one temporal product on $[0,1]$.

Lemma 3.3: The unique temporal product on $[0,1]$ has the form $a\circ b=ab$.

Proof: It is clear that $a\circ b=ab$ is a temporal product on $[0,1]$. Conversely, suppose $a\circ b$ is a temporal product on $[0,1]$. Let $a\in[0,1]$ with $a\neq 0$ and define $\phi_a\colon[0,1]\to[0,1]$ by $\phi_a(b)=(a\circ b)/a$. Then $\phi_a\in\mathscr{S}([0,1])$ and it follows[14] that $\phi_a(b)=b$ for all $b\in[0,1]$. Hence, $a\circ b=ab$. □

In general, a temporal product is not associative. When we write $a_1\circ a_2\circ\cdots\circ a_n$ we mean that the operations are performed from right to left. That is, first $a_{n-1}\circ a_n$ is performed, then $a_{n-2}\circ(a_{n-1}\circ a_n)$ is performed,..., and finally $a_1\circ(a_2\circ\cdots\circ a_n)$ is performed. For example, $a\circ b\circ c$ means $a\circ(b\circ c)$.

Lemma 3.4: Let $\circ$ be a temporal product on E. (i) $a\circ b\leqslant a$ for all $a,b\in E$. (ii) $a\circ b'=a\ominus a\circ b$ for all $a,b\in E$. (iii) If $a\leqslant b$ then $c\circ a\leqslant c\circ b$ for all $c\in E$. (iv) If $a\leqslant b$, then $c\circ(b\ominus a)=c\circ b\ominus c\circ a$ for all $c\in E$.

Proof: (i) For $a,b\in E$ we have

$$a=a\circ 1=a\circ(b\oplus b')=a\circ b\oplus a\circ b'\geqslant a\circ b.$$

(ii) This follows from the proof of (i). (iii) The function $a\mapsto c\circ a$ is additive so it is monotone. (iv) Since $b=a\oplus(b\ominus a)$ we have

$$c\circ a\oplus c\circ(b\ominus a)=c\circ(a\oplus(b\ominus a))=c\circ b.$$

□

A *temporal effect algebra* is a σ-effect algebra that possesses an order determining set of states and a temporal product. For the remainder of this section, E will be a temporal effect algebra with temporal product $\circ$. For $\alpha\in\mathscr{S}(E)$ and $a,b\in E$ with $\alpha(a)\neq 0$, the *conditional probability of b given a in the state α* is defined as $\alpha(b|a)=\alpha(a\circ b)/\alpha(a)$. Notice that $\alpha(\cdot\,|a)\in\mathscr{S}(E)$ and we denote this state by $\alpha|a$. Thus, $(\alpha|a)(b)=\alpha(a\circ b)/\alpha(a)$. By employing the temporal product

$A\circ B=A\cap B$ on $(\Omega,\mathscr{A})$ we see that this definition of conditional probability reduces to the usual concept. Also, it is easy to check that $(\Omega,\mathscr{A})$ and $[0,1]$ are temporal effect algebras.

Temporal effect algebras also give a convenient framework for discussing dynamics. An *automorphism* on E is a σ-isomorphism ϕ: $E\to E$ such that $\phi(a\circ b)=\phi(a)\circ\phi(b)$ for all $a,b\in E$. Notice that the set of automorphisms aut(E) on E forms a group under composition. Indeed, it is clear that if $\phi,\psi\in\text{aut}(E)$, then $\phi\psi\in\text{aut}(E)$. Moreover, if $\phi\in\text{aut}(E)$, then ϕ^{-1} is a σ-isomorphism and

$$\phi[\phi^{-1}(a)\circ\phi^{-1}(b)]=a\circ b$$

so that

$$\phi^{-1}(a)\circ\phi^{-1}(b)=\phi^{-1}(a\circ b).$$

Hence, $\phi^{-1}\in\text{aut}(E)$. If $\phi_1,\ldots,\phi_n\in\text{aut}(E)$, $a_1,\ldots,a_n\in E$, we use the notation $\phi_1\circ a_1\circ\phi_2\circ a_2\circ\cdots\circ\phi_n\circ a_n$ to mean that each ϕ_i operates on the expression to its right. For example,

$$\phi_1\circ a_1\circ\phi_2\circ a_2=\phi_1[a_1\circ(\phi_2(a_2))].$$

A *dynamical group* on E is a one-parameter group of automorphisms on E. That is, a dynamical group on E is a group homomorphism $\tau\mapsto\phi(\tau)$ from the additive group $\mathbb{R}$ into aut(E). We thus have $\phi(\tau_1+\tau_2)=\phi(\tau_2)\circ\phi(\tau_2)$ for all $\tau_1,\tau_2\in\mathbb{R}$ and $\phi(0)=I$ where I is the identity automorphism. It follows that $\phi(-\tau)=\phi(\tau)^{-1}$ for all $\tau\in\mathbb{R}$. If $a\in E$ is an effect at time $\tau=0$, we interpret $\phi(\tau)a$ as the evolution of a at time τ. If $a_1,\ldots,a_n\in E$ correspond to effects at time $\tau=0$ and $\tau_1<\cdots<\tau_n$, we interpret $\phi(\tau_1)a_1\circ\cdots\circ\phi(\tau_n)a_n$ as the effect in which a_1 occurs at time τ_1, a_2 at time $\tau_2,\ldots,a_n$ at time τ_n.

Lemma 3.5: If $\phi(\tau)$ is a dynamical group, then

$$\phi(\tau_1)a_1\circ\cdots\circ\phi(\tau_n)a_n=\phi(\tau_1)\circ a_1\circ\phi(\tau_2-\tau_1)\circ a_2\circ\cdots\circ\phi(\tau_n-\tau_{n-1})\circ a_n.$$

Proof: We proceed by induction on n. The result clearly holds for $n=1$. For $n=2$ we have

$$\begin{aligned}\phi(\tau_1)a_1\circ\phi(\tau_2)a_2&=\phi(\tau_1)a_1\circ\phi(\tau_1)\phi(\tau_2-\tau_1)a_2\\&=\phi(\tau_1)[a_1\circ\phi(\tau_2-\tau_1)a_2]=\phi(\tau_1)\circ a_1\circ\phi(\tau_2-\tau_1)\circ a_2.\end{aligned}$$

Now suppose the result holds for $n\geqslant 1$. Then

$$\begin{aligned}\phi(\tau_1)a_1\circ\cdots\circ\phi(\tau_{n+1})a_{n+1}&=\phi(\tau_1)a_1\circ\phi(\tau_2)[a_2\circ\phi(\tau_3-\tau_2)\circ a_3\circ\cdots\circ\phi(\tau_{n+1}-\tau_n)\circ a_{n+1}]\\&=\phi(\tau_1)\circ a_1\circ\phi(\tau_2-\tau_1)\circ a_2\circ\cdots\circ\phi(\tau_{n+1}-\tau_n)a_{n+1}.\end{aligned}$$

The result follows by induction. □

If $\tau_1<\cdots<\tau_n$, we can interpret Lemma 3.5 as saying that the system evolves until time τ_1 and a_1 occurs, then the system evolves from time τ_1 until time τ_2 and a_2 occurs,..., then the system evolves from time τ_{n-1} until time τ_n and a_n occurs.

If $\phi\in\text{aut}(E)$, we define $\hat{\phi}$: $\mathscr{S}(E)\to\mathscr{S}(E)$ by $\hat{\phi}\alpha(a)=\alpha(\phi^{-1}a)$. Then $\hat{\phi}$ is a bijection on $\mathscr{S}(E)$. Moreover, $(\phi_1\phi_2)\hat{}=\hat{\phi}_1\hat{\phi}_2$ because

$$(\phi_1\phi_2)\hat{}\,\alpha(a)=\alpha((\phi_1\phi_2)^{-1}a)=\alpha(\phi_2^{-1}\phi_1^{-1}a)=\hat{\phi}_2\alpha(\phi_1^{-1}a)=\hat{\phi}_1\hat{\phi}_2\alpha(a).$$

Hence, aut$(E)\hat{}$ is a group and $\phi\mapsto\hat{\phi}$ is a group isomorphism from aut(E) onto aut$(E)\hat{}$. If $\phi(\tau)$ is a dynamical group on E and $\alpha\in\mathscr{S}(E)$ is the state of the system at time $\tau=0$, then we interpret $\hat{\phi}(\tau_0)\alpha$ as the state at time τ_0. If $0<\tau_0<\tau_1<\cdots<\tau_n$, by Lemma 3.5, the probability of the temporal effect $\phi(\tau_1)a_1\circ\cdots\circ\phi(\tau_n)a_n$ in the state $\hat{\phi}(\tau_0)\alpha$ becomes

$$\hat{\phi}(\tau_0)\alpha[\phi(\tau_1)\circ a_1\circ\cdots\circ\phi(\tau_n-\tau_{n-a})\circ a_n]=\alpha[\phi(\tau_1-\tau_0)\circ a_1\circ\cdots\circ\phi(\tau_n-\tau_{n-1})\circ a_n].$$

We can also transfer an automorphism on E into a bijection on $\mathcal{O}(E)=\mathcal{O}_{\mathcal{B}(\mathbb{R})}(E)$. An *observable automorphism* is a bijection $\psi\colon \mathcal{O}(E)\to\mathcal{O}(E)$ such that $(\psi X)(A)=(\psi Y)(B)\circ(\psi Z)(C)$ and $(\psi^{-1}X)(A)=(\psi^{-1}Y)(B)\circ(\psi^{-1}Z)(C)$ whenever $X(A)=Y(B)\circ Z(C)$. It is clear that the set of observable automorphisms $\mathrm{aut}_{\mathcal{O}}(E)$ is a group under composition. If $\phi\in\mathrm{aut}(E)$, define $\tilde{\phi}\colon \mathcal{O}(E)\to\mathcal{O}(E)$ by $(\tilde{\phi}X)(A)=\phi[X(A)]$.

Theorem 3.6: The map $\phi\mapsto\tilde{\phi}$ is a group isomorphism from $\mathrm{aut}(E)$ onto $\mathrm{aut}_{\mathcal{O}}(E)$.

Proof: Letting $\phi\in\mathrm{aut}(E)$, it is straightforward to show that $\tilde{\phi}\colon \mathcal{O}(E)\to\mathcal{O}(E)$ is a bijection. To show that $\tilde{\phi}\in\mathrm{aut}_{\mathcal{O}}(E)$ suppose $X(A)=Y(B)\circ Z(C)$. We then have

$$(\tilde{\phi}X)(A)=\phi[X(A)]=\phi[Y(B)\circ Z(C)]=\phi[Y(B)]\circ\phi[Z(C)]=(\tilde{\phi}Y)(B)\circ(\tilde{\phi}Z)(C).$$

Also, since $(\tilde{\phi})^{-1}=(\phi^{-1})\tilde{}$ we have

$$((\tilde{\phi})^{-1}X)(A)=((\tilde{\phi})^{-1}Y)(B)\circ((\tilde{\phi})^{-1}Z)(C)$$

so $\tilde{\phi}\in\mathrm{aut}_{\mathcal{O}}(E)$. If $\phi_1,\phi_2\in\mathrm{aut}(E)$, then

$$[(\phi_1\phi_2)\tilde{}X](A)=\phi_1\phi_2[X(A)]=\phi_1(\tilde{\phi}_2X)(A)=(\tilde{\phi}_1\tilde{\phi}_2X)(A).$$

Hence, $(\phi_1\phi_2)\tilde{}=\tilde{\phi}_1\tilde{\phi}_2$ so $\phi\mapsto\tilde{\phi}$ is a group homomorphism. To show that $\phi\mapsto\tilde{\phi}$ is injective suppose that $\tilde{\phi}_1=\tilde{\phi}_2$ and $a\in E$. Let $\lambda_1,\lambda_2\in\mathbb{R}$ with $\lambda_1\neq\lambda_2$ and define $X\in\mathcal{O}(E)$ by $X(\lambda_1)=a$, $X(\lambda_2)=a'$ and for $A\in\mathcal{B}(\mathbb{R})$

$$X(A)=\oplus\{X(\lambda_i):\lambda_i\in A,i=1,2\}.$$

We then have

$$\phi_1(a)=\phi_1[X(\lambda_1)]=(\tilde{\phi}_1X)(\lambda_1)=(\tilde{\phi}_2X)(\lambda_1)=\phi_2[X(\lambda_1)]=\phi_2(a).$$

Hence, $\phi_1=\phi_2$ so $\phi\mapsto\tilde{\phi}$ is injective. Finally, to show that $\phi\mapsto\tilde{\phi}$ is surjective, suppose $\psi\in\mathrm{aut}_{\mathcal{O}}(E)$. We have previously shown that for $a\in E$ there exists an $X\in\mathcal{O}(E)$ and an $A\in\mathcal{B}(\mathbb{R})$ such that $X(A)=a$. Define $\phi\colon E\to E$ by $\phi(a)=(\psi X)(A)$. Now ϕ is well defined because if $Y(B)=a$ we have $X(A)=Y(B)\circ Y(\mathbb{R})$ so that

$$(\psi X)(A)=(\psi Y)(B)\circ(\psi Y)(\mathbb{R})=(\psi Y)(B).$$

To show that ϕ is injective, suppose $\phi(a)=\phi(b)$. As before there exists $X,Y\in\mathcal{O}(E),\lambda_1,\lambda_2\in\mathbb{R}$ such that $X(\lambda_1)=a$, $X(\lambda_2)=a'$, $Y(\lambda_1)=b$, $Y(\lambda_2)=b'$. Since

$$X(\{\lambda_1,\lambda_2\})=1=X(\mathbb{R})\circ X(\mathbb{R}),$$

we have

$$(\psi X)(\lambda_1)\oplus(\psi X)(\lambda_2)=(\psi X)(\{\lambda_1,\lambda_2\})=(\psi X)(\mathbb{R})\circ(\psi X)(\mathbb{R})=1,$$

and similarly, $(\psi Y)(\lambda_1)\oplus(\psi Y)(\lambda_2)=1$. Moreover,

$$(\psi X)(\lambda_1)=\phi(a)=\phi(b)=(\psi Y)(\lambda_1)$$

and

$$(\psi X)(\lambda_2)=(\psi X)(\lambda_1)'=(\psi Y)(\lambda_1)'=(\psi Y)(\lambda_2).$$

It follows that $\psi X=\psi Y$ and since ψ is injective, we have $X=Y$. Hence, $a=b$ so ϕ is injective. To show that ϕ is surjective, suppose $b\in E$. Then there exist $Y\in\mathcal{O}(E)$, $\lambda\in\mathbb{R}$ such that $Y(\lambda)=b$. Since ψ is surjective, there exists an $X\in\mathcal{O}(E)$ such that $\psi X=Y$. Letting $a=X(\lambda)$ we have

$$\phi(a)=(\psi X)(\lambda)=Y(\lambda)=b,$$

so ϕ is surjective. Since $X(\mathbb{R})=1$ for $X\in\mathcal{O}(E)$ we have

$$\phi(1)=(\psi X)(\mathbb{R})=1.$$

Assume that $a_i\in E$, $i=1,2,...$, and $\oplus a_i$ exists. Let $\lambda_0,\lambda_1,\ldots\in\mathbb{R}$ be distinct and define $X\in\mathcal{O}(E)$ by $X(\lambda_0)=(\oplus a_i)'$, $X(\lambda_i)=a_i$, $i=1,2,...$, and for $A\in\mathcal{B}(\mathbb{R})$,

$$X(A)=\oplus\{X(\lambda_i):\lambda_i\in A\}.$$

We then have

$$\phi(\oplus a_i)=(\psi X)(\cup\{\lambda_i\})=\oplus(\psi X)(\lambda_i)=\oplus\phi(a_i),$$

and it follows from Lemma 3.2 that ϕ is σ-additive. Define $\phi_1: E\to E$ by $\phi_1(a)=(\psi^{-1}X)(A)$ where $X(A)=a$. As before, ϕ_1 is a well defined σ-morphism and we have

$$\phi\phi_1(a)=(\psi\psi^{-1}X)(A)=X(A)=a.$$

Hence, $\phi\phi_1=I$ and similarly, $\phi_1\phi=I$. It follows that ϕ is a σ-isomorphism. For $a,b\in E$ there exists $X,Y,Z\in\mathcal{O}(E)$, $A,B,C\in\mathcal{B}(\mathbb{R})$ such that $X(A)=a$, $Y(B)=b$, $Z(C)=a\circ b$. Since $Z(C)=X(A)\circ Y(B)$ we have

$$\phi(a\circ b)=\phi(Z(C))=(\psi Z)(C)=(\psi X)(A)\circ(\psi Y)(B)=\phi(a)\circ\phi(b).$$

Hence, $\phi\in\mathrm{aut}(E)$ and for any $X\in\mathcal{O}(E)$, $A\in\mathcal{B}(\mathbb{R})$ we have

$$(\psi X)(A)=\phi[X(A)]=(\tilde{\phi}X)(A).$$

We conclude that $\psi=\tilde{\phi}$ so $\phi\mapsto\tilde{\phi}$ is surjective. □

A *quantum stochastic process* on E is a set of observables X_τ, $\tau\in\mathbb{R}$. If $\phi(\tau)$ is a dynamical group on E and $X\in\mathcal{O}(E)$ then $X_\tau=\phi(\tau)\tilde{}(X)$ is an example of a quantum stochastic process. We then interpret X_τ as the evolution of the observable X.

IV. HISTORY EFFECTS

This section shows that there is a close connection between a history structure and a temporal effect algebra. In this section $(\mathcal{S},\mathcal{O},P)$ will denote a fixed history structure with set of histories $\mathcal{H}$. We have seen that for $h\in\mathcal{H}$, $\alpha\in\mathcal{S}$, the probability $P_\alpha(h)$ is well defined. Moreover, by (H3) there exists a one-time event (X,A) such that

$$P_\alpha(h)=P_{\alpha,X}(A)=P_\alpha((X,A))$$

for every $\alpha\in\mathcal{S}$. For $h_1,h_2\in\mathcal{H}$ we define $h_1\sim h_2$ if $P_\alpha(h_1)=P_\alpha(h_2)$ for every $\alpha\in\mathcal{S}$. In this case h_1 and h_2 are statistically indistinguishable. Clearly, $\sim$ is an equivalence relation on $\mathcal{H}$ and we denote the equivalence class containing h by $[h]$. We call the elements of $\mathcal{E}=\mathcal{H}/\sim=\{[h]:h\in\mathcal{H}\}$ *history effects* and we call $\mathcal{E}$ the *logic* for $(\mathcal{S},\mathcal{O},P)$. Notice that for every $[h]\in\mathcal{E}$ there exists a *one-time effect* $[(X,A)]$ such that $[h]=[(X,A)]$. For $\alpha\in\mathcal{S}$, $[h]\in\mathcal{E}$, we define $P_\alpha([h])=P_\alpha(h)$. We also define $O\in\mathcal{E}$ by $O=[(X,\varnothing)]$ and $1\in\mathcal{E}$ by $1=[(X,\mathbb{R})]$ for any $X\in\mathcal{O}$.

Lemma 4.1: (i) For $a,b\in\mathcal{E}$, if $P_\alpha(a)=P_\alpha(b)$ for every $\alpha\in\mathcal{S}$, then $a=b$. (ii) For $a,b\in\mathcal{E}$, if $P_\alpha(a)+P_\alpha(b)\leqslant 1$ for every $\alpha\in\mathcal{S}$, then there exists a unique $c\in\mathcal{E}$ such that $P_\alpha(c)=P_\alpha(a)+P_\alpha(b)$ for every $\alpha\in\mathcal{S}$.

Proof: (i) Letting $a=[h_1]$, $b=[h_2]$ we have $P_\alpha(h_1)=P_\alpha(h_2)$ for every $\alpha\in\mathcal{S}$. Hence, $h_1\sim h_2$ so $a=b$. (ii) Let $a=[(X,A)]$, $b=[(Y,B)]$. Then for every $\alpha\in\mathcal{S}$ we have

$$P_{\alpha,X}(A)=P_\alpha(a)\leqslant 1-P_\alpha(b)=1-P_{\alpha,Y}(B)=P_{\alpha,Y}(B').$$

By Lemma 2.3, there exists a $Z\in\mathcal{O}$ and C, $D\in\mathcal{B}(\mathbb{R})$ with $C\subseteq D$ such that $P_{\alpha,Z}(C)=P_{\alpha,X}(A)$ and $P_{\alpha,Z}(D)=P_{\alpha,Y}(B')$ for every $\alpha\in\mathcal{S}$. Since $C\cap D'=\varnothing$, letting $c=[(Z,C\cup D')]$ we have

$$P_\alpha(c)=P_{\alpha,Z}(C\cup D')=P_{\alpha,Z}(C)+P_{\alpha,Z}(D')=P_{\alpha,X}(A)+P_{\alpha,Y}(B)=P_\alpha(a)+P_\alpha(b)$$

for every $\alpha\in\mathscr{S}$. The uniqueness of c follows from (i). □

For $a,b\in\mathscr{E}$, we write $a\perp b$ if $P_\alpha(a)+P_\alpha(b)\leqslant 1$ for every $\alpha\in\mathscr{S}$. If $a\perp b$ we define $a\oplus b=c$, where c is the unique element of $\mathscr{E}$ given by Lemma 4.1(ii). We interpret c as the statistical sum of a and b in the sense that c is the unique element of $\mathscr{E}$ such that $P_\alpha(c)=P_\alpha(a)+P_\alpha(b)$ for every $\alpha\in\mathscr{S}$. We define the temporal product of $[(X,A)],[(Y,B)]\in\mathscr{E}$ by

$$[(X,A)]\circ[(Y,B)]=[((X,A),(Y,B))].$$

The temporal product is well defined by (H2). For $a,b\in\mathscr{E}$ we have $a=[(X,A)]$, $b=[(Y,B)]$ for some $X,Y\in\mathscr{O}$ and $A,B\in\mathscr{B}(\mathbb{R})$ and we define $a\circ b=[(X,A)]\circ[(Y,B)]$. The next result shows that the logic of a history structure is a temporal effect algebra.

Theorem 4.2: The algebraic system $(\mathscr{E},0,1,\oplus,\circ)$ is a temporal effect algebra and $\{P_\alpha:\alpha\in\mathscr{S}\}$ is an order determining set of states on $\mathscr{E}$.

Proof: It is clear that if $a\perp b$, then $b\perp a$ and $b\oplus a=a\oplus b$. Now suppose that $a\perp b$ and $a\oplus b\perp c$. Then for every $\alpha\in\mathscr{S}$ we have

$$P_\alpha(a)+P_\alpha(b)+P_\alpha(c)=P_\alpha(a\oplus b)+P_\alpha(c)\leqslant 1$$

so $b\perp c$. Since $P_\alpha(a)+P_\alpha(b\oplus c)\leqslant 1$ for every $\alpha\in\mathscr{S}$, we have $a\perp b\oplus c$. For every $\alpha\in\mathscr{S}$ we have

$$P_\alpha(a\oplus(b\oplus c))=P_\alpha(a)+P_\alpha(b)+P_\alpha(c)=P_\alpha((a\oplus b)\oplus c)$$

so by Lemma 4.1(i), $a\oplus(b\oplus c)=(a\oplus b)\oplus c$. For $a=[(X,A)]\in\mathscr{E}$, let $a'=[(X,A')]$. Then for every $\alpha\in\mathscr{S}$ we have

$$P_\alpha(a)+P_\alpha(a')=P_{\alpha,X}(A)+P_{\alpha,X}(A')=1=P_\alpha(1).$$

Hence, $a\perp a'$ and by Lemma 4.1(i), $a\oplus a'=1$ and a' is unique. If $a\perp 1$, then for every $\alpha\in\mathscr{S}$ we have

$$P_\alpha(a)+1=P_\alpha(a)+P_\alpha(1)\leqslant 1.$$

Hence, $P_\alpha(a)=0$ for every $\alpha\in\mathscr{S}$ so $a=0$. We conclude that $(\mathscr{E},0,1,\oplus)$ is an effect algebra. For $\alpha\in\mathscr{S}$, we have $P_\alpha(1)=1$ and P_α is additive because $a\perp b$ implies that $P_\alpha(a\oplus b)=P_\alpha(a)+P_\alpha(b)$. To show that $\{P_\alpha:\alpha\in\mathscr{S}\}$ is order determining suppose that $P_\alpha(a)\leqslant P_\alpha(b)$ for every $\alpha\in\mathscr{S}$. Then $P_\alpha(a)+P_\alpha(b')\leqslant 1$ for every $\alpha\in\mathscr{S}$ so $a\perp b'$. Since

$$P_\alpha(a)+P_\alpha((a\oplus b')')=P_\alpha(a)+1-P_\alpha(a\oplus b')=P_\alpha(a)+1-P_\alpha(a)-P_\alpha(b')=P_\alpha(b)$$

for every $\alpha\in\mathscr{S}$ we have $a\oplus(a\oplus b')'=b$ so $a\leqslant b$. To show that $\mathscr{E}$ is a σ-effect algebra, suppose that $a_i\in\mathscr{E}$, $i=1,2,...$, with $a_1\leqslant a_2\leqslant\cdots$. Then $a_i=[(X_i,A_i)]$ and since $P_\alpha(a_i)\leqslant P_\alpha(a_{i+1})$ we have $P_{\alpha,X_i}(A_i)\leqslant P_{\alpha,X_{i+1}}(A_{i+1})$, $i=1,2,...$, for every $\alpha\in\mathscr{S}$. Applying Lemma 2.3, there exists an $X\in\mathscr{O}$ and $B_i\in\mathscr{B}(\mathbb{R})$ with $B_i\subseteq B_{i+1}$ such that $P_{\alpha,X}(B_i)=P_{\alpha,X_i}(A_i)$, $i=1,2,...$, for every $\alpha\in\mathscr{S}$. Letting $a=[(X,\cup B_i)]$ we have

$$\lim P_\alpha(a_i)=\lim P_{\alpha,X_i}(A_i)=\lim P_{\alpha,X}(B_i)=P_{\alpha,X}(\cup B_i)=P_\alpha(a)$$

for every $\alpha\in\mathscr{S}$. Since $P_\alpha(a_i)\leqslant P_\alpha(a)$ for every $\alpha\in\mathscr{S}$, we conclude that $a_i\leqslant a$, $i=1,2,...$. If $a_i\leqslant b$, $i=1,2,...$, then $P_\alpha(a_i)\leqslant P_\alpha(b)$ for every $\alpha\in\mathscr{S}$, $i=1,2,...$. Hence, $P_\alpha(a)\leqslant P_\alpha(b)$ for every $\alpha\in\mathscr{S}$ so $a\leqslant b$. Hence, $a=\vee a_i$ so $\mathscr{E}$ is a σ-effect algebra. Also P_α is a state for every $\alpha\in\mathscr{S}$ because we have just shown that if $a_1\leqslant a_2\leqslant\cdots$, then $P_\alpha(\vee a_i)=\lim P_\alpha(a_i)$.

To complete the proof, we must show that $\circ$ is indeed a temporal product. Letting $a=[(X,A)]\in\mathscr{E}$, then for every $\alpha\in\mathscr{S}$ we have by (H4) that

$$P_\alpha(a\circ 1)=P_\alpha([(X,A),(X,\mathbb{R})])=P_{\alpha,s}(A\times\mathbb{R})=P_{\alpha,X}(A)=P_\alpha(a).$$

Hence, $a\circ 1=a$ for every $a\in\mathscr{E}$. Moreover,

$$P_\alpha(0\circ a)=P_\alpha([(X,\varnothing),(X,A)])=P_{\alpha,s}(\varnothing\times A)=P_{\alpha,s}(\varnothing)=0$$

for every $\alpha\in\mathscr{S}$, so $0\circ a=0$. To show that $\circ$ is additive, let $b\perp c$ so that $b\leqslant c'$. Since $P_\alpha(b)\leqslant P_\alpha(c')$ for every $\alpha\in\mathscr{S}$, by Lemma 2.3 there exists an $X\in\mathscr{O}$ and $B, D\in\mathscr{B}(\mathbb{R})$ with $B\subseteq D$ such that $P_{\alpha,X}(B)=P_\alpha(b)$, $P_{\alpha,X}(D)=P_\alpha(c')$ for every $\alpha\in\mathscr{S}$. Then,

$$P_{\alpha,X}(B\cup D')=P_{\alpha,X}(B)+P_{\alpha,X}(D')=P_\alpha(b)+P_\alpha(c)=P_\alpha(b\oplus c)$$

for every $\alpha\in\mathscr{S}$. Hence, $[(X,B\cup D')]=b\oplus c$. For $a=[(Y,A)]\in\mathscr{E}$, by Corollary 2.2 we have

$$\begin{aligned}P_\alpha(a\circ(b\oplus c))&=P_{\alpha,s,X}([((Y,A),(X,B\cup D'))])\\&=P_{\alpha,s,X}([((Y,A),(X,B))])+P_{\alpha,s,X}([((Y,A),(X,D'))])\\&=P_\alpha(a\circ b)+P_\alpha(a\circ c)\end{aligned}$$

for every $\alpha\in\mathscr{S}$. We conclude that $a\circ b\perp a\circ c$ and $a\circ(b\oplus c)=a\circ b\oplus a\circ c$. To show that $\circ$ is σ-additive, suppose that $b_1\leqslant b_2\leqslant\cdots$. By Lemma 3.4(iii) we have that $a\circ b_1\leqslant a\circ b_2\leqslant\cdots$. Applying (H1) there exists $Y\in\mathscr{O}$ and $B_i\in\mathscr{B}(\mathbb{R})$ with $B_i\subseteq B_{i+1}$ such that $P_{\alpha,Y}(B_i)=P_\alpha(b_i)$, $i=1,2,...$, for every $\alpha\in\mathscr{S}$. As in the previous paragraph, we have $[(Y,\cup B_i)]=\vee b_i$. Letting $a=[(X,A)]$ we have by Corollary 2.2 that

$$\begin{aligned}P_\alpha(\vee a\circ b_i)&=\lim P_\alpha(a\circ b_i)=\lim P_{\alpha,s,Y}([((X,A),(Y,B_i))])\\&=P_{\alpha,s,Y}([((X,A),(Y,\cup B_i))])=P_\alpha(a\circ\vee b_i)\end{aligned}$$

for every $\alpha\in\mathscr{S}$. Hence, $\vee a\circ b_i=a\circ\vee b_i$. □

For $X\in\mathscr{O}$, define $\hat{X}:\mathscr{B}(\mathbb{R})\to\mathscr{E}$ by $\hat{X}(A)=[(X,A)]$.

Theorem 4.3: The map $\hat{X}:\mathscr{B}(\mathbb{R})\to\mathscr{E}$ is an observable on $\mathscr{E}$ and the distribution $P_{\alpha,X}$ of X coincides with the distribution $P_\alpha\circ\hat{X}$ of $\hat{X}$ for every $\alpha\in\mathscr{S}$.

Proof: Notice that $\hat{X}(\mathbb{R})=[(X,\mathbb{R})]=1$. If $A, B\in\mathscr{B}(\mathbb{R})$ with $A\cap B=\varnothing$, then

$$P_{\alpha,X}(A)+P_{\alpha,X}(B)=P_{\alpha,X}(A\cup B)\leqslant 1$$

for every $\alpha\in\mathscr{S}$. Hence, $[(X,A)]\perp[(X,B)]$ and

$$P_\alpha([(X,A)])+P_\alpha([(X,B)])=P_\alpha([(X,A\cup B)])$$

for every $\alpha\in\mathscr{S}$ so that $\hat{X}(A\cup B)=\hat{X}(A)\oplus\hat{X}(B)$. Suppose that $A_i\in\mathscr{B}(\mathbb{R})$ with $A_i\cap A_j=\varnothing$, $i\neq j$. Then for $n\in\mathbb{N}$ we have

$$\hat{X}\left(\bigcup_{i=1}^{\infty}A_i\right)\geqslant\hat{X}\left(\bigcup_{i=1}^{n}A_i\right)=\bigoplus_{i=1}^{n}\hat{X}(A_i).$$

Suppose that $a\geqslant\oplus_{i=1}^{n}\hat{X}(A_i)$ for all $n\in\mathbb{N}$. Then

$$P_\alpha(a)\geqslant\sum_{i=1}^{n}P_{\alpha,X}(A_i)$$

for all $n\in\mathbb{N}$ so that

$$P_\alpha(a)\geqslant\sum_{i=1}^{\infty}P_{\alpha,X}(A_i)=P_{\alpha,X}\left(\bigcup_{i=1}^{\infty}A_i\right)=P_\alpha\left(\hat{X}\left(\bigcup_{i=1}^{\infty}A_i\right)\right)$$

for every $\alpha\in\mathscr{S}$. Hence, $a\geqslant\hat{X}(\cup A_i)$ so that

$$\hat{X}(\cup A_i)=\vee\bigoplus_{i=1}^{n}\hat{X}(A_i)=\bigoplus_{i=1}^{\infty}\hat{X}(A_i).$$

It follows from Lemma 3.2 that $\hat{X}\in\mathcal{O}(\mathcal{E})$. Finally, for $\alpha\in\mathcal{S}$, $A\in\mathcal{B}(\mathbb{R})$ we have

$$(P_\alpha\circ\hat{X})(A)=P_\alpha([(X,A)])=P_\alpha((X,A))=P_{\alpha,X}(A). \qquad \square$$

It follows from Theorem 4.2 that any state $\alpha\in\mathcal{S}$ can be represented by a state $P_\alpha\in\mathcal{S}(\mathcal{E})$. Moreover, by Theorem 4.3, any observable $X\in\mathcal{O}$ can be represented by an observable $\hat{X}\in\mathcal{O}(\mathcal{E})$. We can also represent history sets in $\hat{\mathcal{H}}$ by observables on $\mathcal{E}$. If $s=((X_1,\mathcal{N}_1),\ldots,(X_n,\mathcal{N}_n))\in\hat{\mathcal{H}}$ and $A_i\in\mathcal{N}_i$, $i=1,\ldots,n$, define

$$\hat{s}(A_1\times\cdots\times A_n)=[((X_1,A_1),\ldots,(X_n,A_n))]\in\mathcal{E}.$$

Then as in the proof of Theorem 4.3, $\hat{s}$ extends to a unique observable on $\mathcal{E}$ with value space $(\mathbb{R}^n,\mathcal{A}_s)$. Moreover, the joint sequential distribution $P_{\alpha,s}$ of s coincides with the distribution $P_\alpha\circ\hat{s}$ of $\hat{s}$ for every $\alpha\in\mathcal{S}$.

We have shown that the logic of a history structure is a temporal effect algebra. We now prove the converse statement. That is, any temporal effect algebra is σ-isomorphic to the logic of a history structure.

Theorem 4.4: If $(E,0,1,\oplus,\circ)$ is a temporal effect algebra, then there exists a history structure $(\mathcal{S},\mathcal{O},P)$ such that E is σ-isomorphic to the logic of $(\mathcal{S},\mathcal{O},P)$.

Proof: Let $\mathcal{H}$, $\hat{\mathcal{H}}$ be the families of histories and history sets, respectively, for $\mathcal{O}(E)$. That is,

$$\mathcal{H}=\cup\{(\mathcal{O}(E)\times\mathcal{B}(\mathbb{R}))^n:n\in\mathbb{N}\},$$

$$\hat{\mathcal{H}}=\cup\{(\mathcal{O}(E)\times\mathcal{P}(\mathbb{R}))^n:n\in\mathbb{N}\}.$$

For $s=((X_1,\mathcal{N}_1),\ldots,(X_n,\mathcal{N}_n))\in\hat{\mathcal{H}}$, $\alpha\in\mathcal{S}(E)$ and $A_i\in\mathcal{N}_i$, $i=1,\ldots,n$, define

$$P_{\alpha,s}(A_1\times\cdots\times A_n)=\alpha(X_1(A_1)\circ\cdots\circ X_n(A_n)).$$

We now show that $P_{\alpha,s}$ extends uniquely to an element of $M_1^+(\mathbb{R}^n,\mathcal{A}_s)$. Suppose that $\mathcal{N}_i=\{A_i^j:j=1,2,\ldots\}$, $i=1,\ldots,n$. By the σ-additivity of $\circ$ we have

$$\begin{aligned}\bigoplus_{j_n}X_1(A_1^{j_1})\circ\cdots\circ X_n(A_n^{j_n})&=X_1(A_1^{j_1})\circ\cdots\circ X_{n-1}(A_{n-1}^{j_{n-1}})\circ\bigoplus_{j_n}X_n(A_n^{j_n})\\&=X_1(A_1^{j_1})\circ\cdots\circ X_{n-1}(A_{n-1}^{j_{n-1}})\circ X_n\left(\cup_{j_n}A_n^{j_n}\right)=X_1(A_1^{j_1})\circ\cdots\circ X_{n-1}(A_{n-1}^{j_{n-1}}).\end{aligned}$$

Proceeding by induction, we conclude that

$$\bigoplus_{j_1,\ldots,j_n}X_1(A_1^{j_1})\circ\cdots\circ X_n(A_n^{j_n})=1.$$

Hence,

$$\sum_{j_1,\ldots,j_n}P_{\alpha,s}(A_1^{j_1}\times\cdots\times A_n^{j_n})=1.$$

Now any $A\in\mathcal{A}_s$ has the form

$$A=\cup\{A_1^{j_1}\times\cdots\times A_n^{j_n}:j_1\in I_1,\ldots,j_n\in I_n\},$$

where $I_i\subseteq\mathbb{N}$ is an index set depending on A, $i=1,\ldots,n$. We then define $P_{\alpha,s}\in M_1^+(\mathbb{R}^n,\mathcal{A}_s)$ by

$$P_{\alpha,s}(A)=\sum\{P_{\alpha,s}(A_1^{j_1}\times\cdots\times A_n^{j_n}):j_1\in I_1,\ldots,j_n\in I_n\}.$$

We now show that $(\mathscr{S}(E),\mathscr{O}(E),P)$ is a history structure. To verify (H1), suppose that $s_i=(X_i,\mathscr{N}_i)\in\hat{\mathscr{H}}^1$ and $P_{\alpha,s_i}(A_i)\leqslant P_{\alpha,s_{i+1}}(A_{i+1})$ for every $\alpha\in\mathscr{S}(S)$, $i=1,2,\ldots$. Letting $a_i=X_i(A_i)$ we have

$$\alpha(a_i)=\alpha(X_i(A_i))=P_{\alpha,s_i}(A_i)\leqslant P_{\alpha,s_{i+1}}(A_{i+1})=\alpha(X_{i+1}(A_{i+1}))=\alpha(a_{i+1})$$

for every $\alpha\in\mathscr{S}(E)$, $i=1,2,\ldots$. Since $\mathscr{S}(E)$ is order determining, we have $a_i\leqslant a_{i+1}$, $i=1,2,\ldots$. Let $a=\vee a_i$, $b_1=a_1$, $b_i=a_i\ominus a_{i-1}$, $i=2,3,\ldots$, and we have $\oplus b_i=a$. Letting $b_0=a'$, we have $\oplus_{i=0}^{\infty}b_i=1$. Now let $\{\lambda_i:i=0,1,\ldots\}\subseteq\mathbb{R}$, define $X(\lambda_i)=b_i$, $i=0,1,\ldots$, and for $B\in\mathscr{B}(\mathbb{R})$ define

$$X(B)=\bigoplus\{X(\lambda_i):\lambda_i\in B\}.$$

Then $X\in\mathscr{O}(E)$ and for $B_i=\{\lambda_1,\ldots,\lambda_i\}$, $i=1,2,\ldots$, we have $B_i\subseteq B_{i+1}$ and

$$X(B_i)=\bigoplus_{j=1}^{i}X(\lambda_j)=\bigoplus_{j=1}^{i}b=a_i.$$

Letting $\mathscr{N}$ be a partition generated by $\{\lambda_i\}$ $i=1,2,\ldots$, and $s=(X,\mathscr{N})$ we have

$$P_{\alpha,s_i}(A_i)=\alpha(a_i)=\alpha(X(B_i))=P_{\alpha,s}(B_i)$$

for every $\alpha\in\mathscr{S}(E)$. The verifications of (H2)–(H5) are straightforward. Moreover, it is easy to show that the logic of $(\mathscr{S}(E),\mathscr{O}(E),P)$ is σ-isomorphic to E. □

It follows from Theorem 4.4 that a temporal effect algebra E generates a history structure $(\mathscr{S}(E),\mathscr{O}(E),P)$ in a natural way. For $X_i\in\mathscr{O}(E)$, $A_i\in\mathscr{B}(\mathbb{R})$, we call $X_1(A_1)\circ\cdots\circ X_n(A_n)$ a *history effect* and this corresponds to the history $((X_1,A_1),\ldots,(X_n,A_n))$ of the history structure $(\mathscr{S}(E),\mathscr{O}(E),P)$. For $\mathscr{N}_i\in\mathscr{P}(\mathbb{R})$, the observable on E with value space $(\mathbb{R}^n,\mathscr{A}(\mathscr{N}_1,\ldots,\mathscr{N}_n))$ induced by $X_1(A_1)\circ\cdots\circ X_n(A_n)$, $A_i\in\mathscr{N}_i$, $i=1,\ldots,n$, is called a *history observable* and is denoted by $X_s=X_1\circ\cdots\circ X_n$. This observable corresponds to the history set $s=((X_1,\mathscr{N}_1),\ldots,(X_n,\mathscr{N}_n))$ on $(\mathscr{S}(E),\mathscr{O}(E),P)$. If $\alpha\in\mathscr{S}(E)$, then $\alpha(X_1(A_1)\circ\cdots\circ X_n(A_n))$ is the probability of the corresponding history in the state α and the distribution of X_s gives the joint sequential distribution for the history set s in the state α. It is also clear that an observable automorphism $\psi\in\mathrm{aut}_{\mathscr{O}}(E)$ induces an automorphism on the history structure $(\mathscr{S}(E),\mathscr{O}(E),P)$. Moreover, a dynamical group on E induces a dynamical group on $(\mathscr{S}(E),\mathscr{O}(E),P)$.

It is of interest to note that a fixed history set s corresponds to a classical logic in the sense that the range $R(X_s)\subseteq E$ of X_s is a Boolean algebra.[7] In this Boolean algebra, the operations $\vee$ and $\wedge$ are computed within $R(X_s)$ and not relative to the total effect algebra E. (In general E is not even a lattice.) In fact, if $X_s(A),X_s(B)\in R(X_s)$, then $X_s(A)'=X_s(A')$ and

$$X_s(A)\vee X_s(B)=X_s(A\cup B),$$

$$X_s(A)\wedge X_s(B)=X_s(A\cap B).$$

The ''local'' Boolean algebra $R(X_s)$ is contextual in the sense that it depends on s. If t is another history set then there is no relationship between $R(X_s)$ and $R(X_t)$ in general even when s and t both involve the same observables.

V. HILBERT SPACE HISTORIES

Let H be a complex Hilbert space. A linear operator T on H that satisfies $O\leqslant T\leqslant I$ is called a *Hilbert space effect*[17,18] and the set of Hilbert space effects on H is denoted $\mathscr{E}(H)$. For $S,T\in\mathscr{E}(H)$ we write $S\perp T$ if $S+T\in\mathscr{E}(H)$ and in this case we define $S\oplus T=S+T$.

Lemma 5.1: The system $(\mathscr{E}(H),O,I,\oplus)$ is a σ-effect algebra and every state α on $\mathscr{E}(H)$ has the form $\alpha(T)=\mathrm{tr}(WT)$ where W is a unique density operator on H.

Proof: It is easy to show that $(\mathscr{E}(H),O,I,\oplus)$ is an effect algebra. To show that $\mathscr{E}(H)$ is a σ-effect algebra, suppose that $T_i\in\mathscr{E}(H)$ with $T_1\leqslant T_2\leqslant\cdots$. A standard result[19] shows that there exists a $T\in\mathscr{E}(H)$ such that

$$\langle Tx,x\rangle=\lim\langle T_ix,x\rangle$$

for every $x\in H$. That is $\lim T_i=T$ in the weak operator topology. Since $\langle T_ix,x\rangle\leqslant\langle Tx,x\rangle$ for every $x\in H$, we conclude that $T_i\leqslant T$, $i=1,2,...$. Moreover, if $S\in\mathscr{E}(H)$ with $T_i\leqslant S$, $i=1,2,...$, then $\langle T_ix,x\rangle\leqslant\langle Sx,x\rangle$ for every $x\in H$, $i\in\mathbb{N}$. Hence, $\langle Tx,x\rangle\leqslant\langle Sx,x\rangle$ for every $x\in H$ so $T\leqslant S$. Thus, $S=\vee T_i$ so $\mathscr{E}(H)$ is a σ-effect algebra. It is well known that every state on $\mathscr{E}(H)$ has the given form.[17,18] □

Because of Lemma 5.1, we can identify the set $\mathscr{S}(H)$ of density operators on H with the set of states on $\mathscr{E}(H)$ and it is clear that $\mathscr{S}(H)$ is order determining. Moreover, the observables on $\mathscr{E}(H)$ are given by positive operator valued (POV) measures[17,18] on H and we denote the set of observables by $\mathscr{O}(H)$. For $S,T\in\mathscr{E}(H)$ define $S\circ T=S^{1/2}TS^{1/2}$. Since

$$\langle S^{1/2}TS^{1/2}x,x\rangle=\langle TS^{1/2}x,S^{1/2}x\rangle\leqslant\langle S^{1/2}x,S^{1/2}x\rangle=\langle Sx,x\rangle\leqslant\langle x,x\rangle$$

for every $x\in H$, we conclude that $S\circ T\in\mathscr{E}(H)$.

Theorem 5.2: The system $(\mathscr{E}(H),O,I,\oplus,\circ)$ is a temporal effect algebra.

Proof: We only need to show that $\circ$ is a temporal product. It is clear that $O\circ T=O$ and $T\circ I=T$ for all $T\in\mathscr{E}(H)$. If $T_1,T_2\in\mathscr{E}(H)$ with $T_1\perp T_2$, then

$$S\circ(T_1\oplus T_2)=S\circ(T_1+T_2)=S^{1/2}(T_1+T_2)S^{1/2}=S^{1/2}T_1S^{1/2}\oplus S^{1/2}T_2S^{1/2}=S\circ T_1\oplus S\circ T_2.$$

Now suppose that $T_i\in\mathscr{E}(H)$ with $T_1\leqslant T_2\leqslant\cdots$. As in the proof of Lemma 5.1, $\lim T_i=\vee T_i$ in the weak operator topology. Moreover, since $S\circ T_1\leqslant S\circ T_2\leqslant\cdots$, $\lim S\circ T_i=\vee(S\circ T_i)$ in the weak operator topology. Hence, for every $x\in H$, we have

$$\lim\langle S\circ T_ix,x\rangle=\lim\langle T_iS^{1/2}x,S^{1/2}x\rangle=\langle(\vee T_i)S^{1/2}x,S^{1/2}x\rangle=\langle S\circ\vee T_ix,x\rangle.$$

Thus,

$$S\circ\vee T_i=\lim S\circ T_i=\vee(S\circ T_i)$$

so $T\mapsto S\circ T$ is countably additive. □

Since $\mathscr{E}(H)$ is a temporal effect algebra, we can form the corresponding history structure $(\mathscr{S}(H),\mathscr{O}(H),P)$ as in Sec. IV. For $X_i\in\mathscr{O}(H)$, $A_i\in\mathscr{B}(\mathbb{R})$, $i=1,...,n$, we have the history effect

$$h=X_1(A_1)\circ\cdots\circ X_n(A_n)=X_1(A_1)^{1/2}\cdots X_{n-1}(A_{n-1})^{1/2}X_n(A_n)X_{n-1}(A_{n-1})^{1/2}\cdots X_1(A_1)^{1/2}$$

and for a state $W\in\mathscr{S}(H)$ the probability of h becomes

$$P_W(h)=\operatorname{tr}(WX_1(A_1)\circ\cdots\circ X_n(A_n))=\operatorname{tr}(X_n(A_n)^{1/2}\cdots X_1(A_1)^{1/2}WX_1(A_1)^{1/2}\cdots X_n(A_n)^{1/2}).$$

These expressions are given in Ref. 7. For

$$s=((X_1,\mathscr{N}_1),\ldots,(X_n,\mathscr{N}_n))\in\hat{\mathscr{H}}$$

we have the history observable $X_s=X_1\circ\cdots\circ X_n$ with value space $(\mathbb{R}^n,\mathscr{A}_s)$ and for $W\in\mathscr{S}(H)$ the joint sequential distribution of $X_1,\ldots,X_n$ with partitions $\mathscr{N}_1,\ldots,\mathscr{N}_n$ is the probability measure on $(\mathbb{R}^n,\mathscr{A}_s)$ given by $P_{W,s}(X_s(\cdot))=\operatorname{tr}(WX_s(\cdot))$. If $S,T\in\mathscr{E}(H)$ and $W\in\mathscr{S}(H)$ with $\operatorname{tr}(WS)\neq0$, then the conditional probability of T given S is

$$P_W(T|S)=\frac{\operatorname{tr}(WS\circ T)}{\operatorname{tr}(WS)}=\frac{\operatorname{tr}(S^{1/2}WS^{1/2}T)}{\operatorname{tr}(WS)}.$$

Hence, the conditional state $W|S\in\mathscr{S}(H)$ is given by

$$W|S=\frac{S^{1/2}WS^{1/2}}{\mathrm{tr}(WS)}.$$

We now illustrate some manifestations of quantum interference discussed in Sec. II. Let $s=((X_1,\mathcal{N}),(X_2,\mathcal{N}))\in\hat{\mathcal{H}}$ where $\mathcal{N}=\{A_1,A_2\}$, $A_2=A_1'$. The possible alternatives in this history set are the four history effects

$$X_s(A_i\times A_j)=X_1(A_i)\circ X_2(A_j)\quad i,j=1,2.$$

In general

$$\begin{aligned}X_1\circ X_2(\mathbb{R}\times A_1)&=X_s(\mathbb{R}\times A_1)=X_s(A_1\times A_1\cup A_2\times A_1)=X_s(A_1\times A_1)+X_s(A_2\times A_1)\\&=X_1(A_1)\circ X_2(A_1)+X_1(A_2)\circ X_2(A_1)\\&=X_1(A_1)^{1/2}X_2(A_1)X_1(A_1)^{1/2}+X_1(A_2)^{1/2}X_2(A_1)X_1(A_2)^{1/2}\neq X_2(A_1).\end{aligned}$$

Thus, the marginal distribution need not coincide with the one-variable distribution. Of course, this is just an illustration of the nonadditivity in the first argument $(a\oplus b)\circ c\neq a\circ c\oplus b\circ c$.

As another example, let $s=((X_1,\mathcal{N}_1),(X_2,\mathcal{N}_2))$, $t=((X_1,\mathcal{M}_1),(X_2,\mathcal{M}_2))$ and suppose that $A,B\in\mathcal{N}_1$, $A\cup B\in\mathcal{M}_1$, $C\in\mathcal{N}_2\cap\mathcal{M}_2$. Then

$$X_s(A\cup B\times C)=X_s(A\times C\cup B\times C)=X_s(A\times C)\oplus X_s(B\times C)=X_1(A)\circ X_2(C)\oplus X_1(B)\circ X_2(C)$$

$$X_t(A\cup B\times C)=X_1\circ X_2(A\cup B\times C)=X_1(A\cup B)\circ X_2(C)=[X_1(A)\oplus X_1(B)]\circ X_2(C).$$

These two expressions need not coincide. Hence, if $W\in\mathcal{S}(H)$ in general we have

$$P_{W,s}(A\cup B\times C)\neq P_{W,t}(A\cup B\times C).$$

We thus see that refining a partition can change the probability of measurement results. Finally, the fact that Bayes' formula

$$P_W(T|S)P_W(S)=P_W(S|T)P_W(T)$$

does not hold in general follows from the fact that $S\circ T=T\circ S$ need not hold. That is, in general $S^{1/2}TS^{1/2}\neq T^{1/2}ST^{1/2}$.

This framework encompasses the formalism of quantum evolution. Let U: $H\to H$ be a unitary operator and define ϕ: $\mathcal{E}(H)\to\mathcal{E}(H)$ by $\phi(T)=U^*TU$. Then ϕ is an automorphism on $\mathcal{E}(H)$. Indeed, it is straightforward to show that ϕ is a σ-isomorphism and to show that ϕ preserves the temporal product we have

$$\begin{aligned}\phi(S\circ T)&=U^*S^{1/2}TS^{1/2}U=U^*S^{1/2}UU^*TUU^*S^{1/2}U\\&=(U^*SU)^{1/2}U^*TU(U^*SU)^{1/2}=\phi(S)\circ\phi(T).\end{aligned}$$

Let $U(\tau)=e^{-i\tau K/h}$ be a unitary evolution on H where K is the Hamiltonian of the system. It is clear that $U(\tau)$ is a dynamical group on $\mathcal{E}(H)$. If $X\in\mathcal{O}(H)$ is an observable at time zero, then the corresponding observable at time τ is

$$X_\tau(\cdot)=\phi_\tau(X(\cdot))=U(\tau)^*X(\cdot)U(\tau).$$

Thus, X_τ, $\tau\in\mathbb{R}$, is a quantum stochastic process. For times $\tau_1<\tau_2<\cdots<\tau_n$ and $A_1,\ldots,A_n\in\mathcal{B}(\mathbb{R})$ a temporal history effect is given by $h=X_{\tau_1}(A_1)\circ\cdots\circ X_{\tau_n}(A_n)$. To simplify the notation, let $U_i=U(\tau_i)$ and $U_{i,j}=U(\tau_j-\tau_i)$ for $i<j$. We then have

$$\begin{aligned}h&=U_1^*X(A_1)^{1/2}U_1\cdots U_{n-1}^*X(A_{n-1})^{1/2}U_{n-1}U_n^*X(A_n)U_nU_{n-1}^*X(A_{n-1})^{1/2}U_{n-1}\cdots U_1^*X(A_1)^{1/2}U_1\\&=U_1^*X(A_1)^{1/2}U_{2,1}^*\cdots U_{n-1,n-2}^*X(A_{n-1})^{1/2}U_{n,n-1}^*X(A_n)U_{n,n-1}X(A_{n-1})^{1/2}\\&\quad\times U_{n-1,n-2}\cdots U_{2,1}X(A_1)^{1/2}U_1.\end{aligned}$$

If $W\in\mathcal{S}(H)$ is the state at time zero and $\tau_0<\tau_1$ is the initial time under consideration, then the corresponding state at time τ_0 is $W_{\tau_0}=U(\tau_0)^*WU(\tau_0)$. The probability of h for the state W becomes

$$P_{W_{\tau_0}}(h)=\operatorname{tr}(U(\tau_0)^*WU(\tau_0)h)$$
$$=\operatorname{tr}(WU_{1,0}^*X(A_1)^{1/2}U_{2,1}^*\cdots U_{n-1,n-2}^*X(A_{n-1})^{1/2}\cdots U_{2,1}X(A_1)^{1/2}U_{1,0})$$

and this is the form given previously in Ref. 7.

In order to compare our formalism to those given previously, we can reformulate it in the following way. For

$$s=((X_1,\mathcal{N}_1),\ldots,(X_n,\mathcal{N}_n))\in\hat{\mathcal{H}}$$

we use the notation $\mathcal{N}=(\mathcal{N}_1,\ldots,\mathcal{N}_n)$ and for $A_i\in\mathcal{N}_i$, $i=1,...,n$, we write $A=(A_1,\ldots,A_n)\in\mathcal{N}$. For $A\in\mathcal{N}$ we define the *class operator*

$$C_s(A)=X_n(A_n)^{1/2}X_{n-1}(A_{n-1})^{1/2}\cdots X_1(A_1)^{1/2}.$$

Notice that

$$C_s(A)^*C_s(A)=X_1(A_1)\circ\cdots\circ X_n(A_n),$$

which is a history effect. For $W\in\mathcal{S}(H)$, we define the *decoherence function* $D_s^W:\mathcal{N}\times\mathcal{N}\to\mathbb{C}$ by

$$D_s^W(A,B)=\operatorname{tr}(WC_s(A)^*C_s(B)).$$

Then the following properties are easy to check.

(D1) $D_s^W(A,A)=P_W(X_1(A_1)\circ\cdots\circ X_n(A_n))\geqslant 0$;

(D2) $D_s^W(A,B)=D_s^W(B,A)^*$;

(D3) $\Sigma_{A\in\mathcal{N}}D_s^W(A,A)=1$.

We say that s is a *consistent history set relative to* W if $\operatorname{Re}D_s^W(A,B)=0$ for $A\neq B$.

We shall not pursue consistent history sets here because they have been thoroughly studied in the literature. Moreover, the decoherence function D_s^W is a specific Hilbert space construct and its physical significance is not completely clear. For this reason, it is not a natural concept to introduce into our axiomatic framework for a history structure or a temporal effect algebra. Nevertheless, we shall make some observations about D_s^W that do not seem to have been previously considered.

For simplicity, assume that $\mathcal{N}=(\mathcal{N}_1,\ldots,\mathcal{N}_n)$ consists of finite partitions $\mathcal{N}_i$, $i=1,...,n$ (which is frequently done in this formalism) and form the Hilbert space $\ell_2(\mathcal{N})=\{\psi:\mathcal{N}\to\mathbb{C}\}$ with inner product

$$\langle\psi,\phi\rangle=\sum_{A\in\mathcal{N}}\psi(A)\phi(A)^*.$$

Now D_s^W induces a linear operator (which we also denote by D_s^W) on $\ell_2(\mathcal{N})$ given by

$$(D_s^W\psi)(A)=\sum_{B\in\mathcal{N}}D_s^W(A,B)\psi(B).$$

Applying (D2) we have

$$\langle D_s^W\psi,\psi\rangle=\sum_{A,B\in\mathcal{N}} D_s^W(A,B)\psi(B)\psi(A)^*$$

$$=\sum_{A,B\in\mathcal{N}} \operatorname{tr}(WC_s^*(A)C_s(B))\psi(B)\psi(A)^*$$

$$=\operatorname{tr}\left[W\sum_{A\in\mathcal{N}}(\psi(A)C_s(A))^*\sum_{B\in\mathcal{N}}\psi(B)C_s(B)\right]\geqslant 0.$$

Hence, D_s^W is a positive operator on $\ell_2(\mathcal{N})$. It follows from (D3) that the trace of D_s^W is 1. Hence, D_s^W is a density matrix so $D_s^W\in\mathcal{S}(\ell_2(\mathcal{N}))$. It is easy to check that $W\mapsto D_s^W$ is a continuous affine map from $\mathcal{S}(H)$ to $\mathcal{S}(\ell_2(\mathcal{N}))$ so it has a unique continuous linear extension L from the trace class operators on H to the trace class operators on $\ell_2(\mathcal{N})$. The adjoint transformation L^* is a continuous linear transformation from $\mathcal{O}(\ell_2(\mathcal{N}))$ into $\mathcal{O}(H)$. With a little care, these ideas can be generalized to the case in which $\mathcal{N}$ is not finite ($\mathcal{N}$ is always countable). We conclude that each history set s possesses its own Hilbert space $\ell_2(\mathcal{N})$ and observables on $\ell_2(\mathcal{N})$ can be transferred in a natural way to observables on H. In a dual manner, states on H can be transferred to states on $\ell_2(\mathcal{N})$ by employing the decoherence function. This may be a useful physical interpretation for the decoherence function.

[1] T. Dass and Y. Joglekar, J. Math. Phys. **39**, 704 (1998).
[2] M. Gell-Mann and J. Hartle, Phys. Rev. D **47**, 3345 (1993).
[3] R. B. Griffiths, J. Stat. Phys. **36**, 219 (1984); Am. J. Phys. **55**, 11 (1986); Found. Phys. **23**, 1601 (1993).
[4] C. J. Isham, J. Math. Phys. **35**, 2157 (1994).
[5] C. J. Isham and N. Linden, J. Math. Phys. **35**, 5452 (1994); **36**, 5392 (1995).
[6] R. Omnes, J. Stat. Phys. **53**, 893 (1988); **53**, 933 (1988); **53**, 957 (1988); **57**, 357 (1989); Ann. Phys. (N.Y.) **201**, 354 (1990); Rev. Mod. Phys. **64**, 339 (1992).
[7] O. Rudolf, Int. J. Theor. Phys. **35**, 1581 (1996); J. Math. Phys. **37**, 5368 (1996).
[8] S. Pulmannová, Int. J. Theor. Phys. **34**, 189 (1995).
[9] G. W. Mackey, *Mathematical Foundations of Quantum Mechanics* (Benjamin, New York, 1963).
[10] R. Medina, ''Theory of non-Kolmogorovian random variables,'' Rep. Math. Phys. (to be published).
[11] A. Dvurečenskij, Trans. Am. Math. Soc. **147**, 1043 (1995).
[12] A. Dvurečenskij and S. Pulmannová, Int. J. Theor. Phys. **33**, 819 (1994).
[13] D. Foulis and M. K. Bennett, Found. Phys. **24**, 1331 (1994).
[14] S. Gudder, Found. Phys. **26**, 813 (1996); Int. J. Theor. Phys. **35**, 2365 (1996).
[15] F. Kôpka, Tatra Mountain Math. Publ. **1**, 83 (1992).
[16] F. Kôpka and F. Chovanec, Math. Slovaca **44**, 21 (1994).
[17] P. Busch, P. J. Lahti, and P. Mittlestaedt, *The Quantum Theory of Measurements* (Springer, Berlin, 1991).
[18] P. Busch, M. Grabowski, and P. J. Lahti, *Operational Quantum Physics* (Springer, Berlin, 1995).
[19] A. Brown and A. Page, *Elements of Functional Analysis* (Van Nostrand Reinhold, London, 1970).

A number of quasiexactly solvable *N*-body problems

Avinash Khare[a)]
Institute of Physics, Sachivalaya Marg, Bhubaneswar-751005, India

Bhabani Prasad Mandal[b)]
Theory Group, Saha Institute of Nuclear Physics, 1/AF, Bidhannagar Calcutta-700064, India

(Received 26 March 1998; accepted for publication 20 April 1998)

We present several examples of quasiexactly solvable *N*-body problems in one, two, and higher dimensions. We study various aspects of these problems in some detail. In particular, we show that in some of these examples the corresponding polynomials form an orthogonal set, and many of their properties are similar to those of the Bender–Dunne polynomials. We also discuss QES problems where the polynomials do not form an orthogonal set. © *1998 American Institute of Physics.* [S0022-2488(98)04510-1]

I. INTRODUCTION

In last few years, the quasiexactly solvable (QES) problems have attracted a lot of attention.[1] In particular, by now a detailed study has been made of several QES problems in (one-particle) nonrelativistic quantum mechanics in one dimension, and several interesting features have been uncovered. However, very few QES problems in multidimensions or *N*-body problems in one dimension[2,3] have been discussed so far. Further, to the best of our knowledge, no QES *N*-body problem in two and higher dimensions has been discussed so far. The purpose of this paper is to initiate a systematic study of the various aspects of the *N*-body QES problems in multidimensions. In particular, one would first like to discover several *N*-body QES problems in two and higher dimensions. Further, one would like to know if the corresponding polynomials form an orthogonal set or not. Besides, one would like to know if there is an underlying hidden algebra. In this context, it is worth recalling that in the case of the one-body problem in one dimension, it is known that in most QES cases (though not all[4]), there is an underlying hidden algebra $Sl(2)$. Further, the polynomials form an orthogonal set provided the Hamiltonian can be written in terms of the quadratic generators of this Sl(2) algebra,[5] and that in this case one has one[6] or two sets[7] of Bender–Dunne polynomials.

In this paper we discuss in some detail three different *N*-body QES problems namely, Calogero–Marchioro-type *N*-body problem in *D* dimensions[8,9] in Sec. II ($D \geqslant 2$), *N*-body systems exhibiting novel correlations in two dimensions[10] in Sec. III A, and the Calogero–Sutherland-type *N*-body problem in one dimension[11] in Sec. III B. In each case we discuss the QES problem in case the potential is either of the sextic type or is a sum of the oscillator plus Coulomb-type potential. We show that whereas in the former case, the polynomials form an orthogonal set, in the latter case they do not. We discuss the various properties of the orthogonal polynomials in these problems and show that they are similar to those of the Bender–Dunne polynomials in the one-body case. As a special case, we also discuss a self-dual system[12] in which case we can obtain several more QES levels analytically. Further, in this case we obtain a novel relation between the weight functions of the system. Finally, in Sec. IV we summarize our conclusions and point out some of the open problems.

II. CALOGERO–MARCHIORO-TYPE PROBLEM IN *D* DIMENSIONS

The *N*-body Hamiltonian corresponding to the Calogero–Marchioro model in *D* dimensions for arbitrary potential *V* can be written as[9]

[a)]Electronic mail: khare@iopb.stpbh.soft.net
[b)]Electronic mail: bpm@tnp.saha.ernet.in

 © 1998 American Institute of Physics

$$H=-\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2+g\sum_{i<j}^{N}\frac{1}{r_{ij}^2}+G\sum_{i<j,i,j\neq k}\frac{\mathbf{r}_{ki}^2\cdot\mathbf{r}_{kj}^2}{r_{ki}^2 r_{kj}^2}+V\left(\sum_{i<j}^{N}r_{ij}^2\right), \tag{2.1}$$

where $\mathbf{r}_{ij}=(\mathbf{r}_i-\mathbf{r}_j)$ and $\mathbf{r}_i$ is the D-dimensional position vector for the ith particle.

On substituting the ansatz,

$$\psi=\left(\prod_{i<j}\mathbf{r}_{ij}^2\right)^{\Lambda_D/2}\phi(\rho), \tag{2.2}$$

in the Schrödinger equation ($H\psi=\epsilon\psi$) corresponding to the above system (with $\hbar=m=1$), it is easily shown that the equation satisfied by $\phi(\rho)$ is[9]

$$\phi''(\rho)+\frac{2\Gamma_D+1}{\rho}\phi'(\rho)-2V(\rho)\phi(\rho)+E\phi(\rho)=0. \tag{2.3}$$

Here $E=2\epsilon$ while

$$\rho^2=\frac{1}{N}\sum_{i<j}^{N}r_{ij}^2,$$

$$\Gamma_D=\tfrac{1}{2}[D(N-1)-2+\Lambda_D N(N-1)], \tag{2.4}$$

$$\Lambda_D=\sqrt{G}=\tfrac{1}{2}[\sqrt{(D-2)^2+4g}-(D-2)].$$

We now consider two different forms of $V(\rho)$ and obtain QES solutions in each case.

A. Sextic potential and QES problem

Let us first consider $V(\rho)$, as given by

$$V(\rho)=\frac{1}{2}\left[B\rho^2+C\rho^4+H\rho^6+\frac{F}{\rho^2}\right]. \tag{2.5}$$

In the special case when $C=0=H$, an infinite class of exact solutions including the bosonic ground state have already been obtained.[9,13]

On substituting

$$\phi(\rho)=\rho^a\exp(-\alpha\rho^2-\beta\rho^4)\eta(\rho), \tag{2.6}$$

in Eq. (2.3), we obtain

$$\eta''(\rho)+\left(\frac{2a+2\Gamma_D+1}{\rho}-[4\alpha\rho+8\beta\rho^3]\right)\eta'(\rho)$$
$$+(E-4\alpha(a+\Gamma_D+1)+\rho^2[4\alpha^2-B-8\beta(a+\Gamma_D+2)])\eta(\rho)=0, \tag{2.7}$$

where we have chosen, $\alpha=C/4\sqrt{H}$, $\beta=\sqrt{H}/4$, and $F=a^2+2a\Gamma_D$. Thus, a is zero or nonzero depending on if F is zero or nonzero. Finally, on substituting

$$\eta(\rho)=\sum_n\frac{P_n(E)\rho^{2n}}{4^n n!(n+a+\Gamma_D)!}, \tag{2.8}$$

in Eq. (2.7), we obtain the recursion relation satisfied by $P_n(E)$ as

$$P_n(E)+[E-4\alpha(2n-1+a+\Gamma_D)]P_{n-1}(E)+4(n-1)(n-1+a+\Gamma_D)$$
$$\times[4\alpha^2-B-8\beta(2n+a+\Gamma_D-2)]P_{n-2}(E)=0, \tag{2.9}$$

with initial conditions $P_{-1}=0$ and $P_0=1$. Using the well-known theorem,[14,15] ''the *necessary and sufficient* condition for a family of polynomials $\{P_n\}$ (with degree $P_n=n$) to form an orthogonal polynomial system is that $\{P_n\}$ satisfy a three-term recursion relation of the form

$$P_n(E)=(A_nE+B_n)P_{n-1}(E)+C_nP_{n-2}(E), \quad n\geqslant 1, \tag{2.10}$$

where the coefficients A_n, B_n, and C_n are independent of E, $A_n\neq 0$, $C_1=0$, $C_n\neq 0$ for $n\geqslant 1$,'' it then follows that $\{P_n(E)\}$ for this problem forms an orthogonal set of polynomials with respect to some weight function, $\omega(E)$.

Let us write B in the form $B=4\alpha^2-8\beta(2J+a+\Gamma_D)$, where J is any arbitrary number. We shall see that when J is a positive integer, then this represents a QES system. In terms of J, Eq. (2.9) then can be rewritten as

$$\begin{aligned} &P_n(E)+[E-4\alpha(2n-1+a+\Gamma_D)]P_{n-1}(E)-64\beta(n-1)\\ &\quad\times(n-1+a+\Gamma_D)(n-J-1)P_{n-2}(E)=0, \end{aligned} \tag{2.11}$$

from where it is clear that as long as J is a positive integer, this recursion relation will reduce to a two-term recursion relation. Thus, when J is a positive integer, we have a QES system. These recursion relations generate a set of orthogonal polynomials of which the first few are

$$\begin{aligned} P_1&=-E+4\alpha(a+\Gamma_D+1),\\ P_2&=E^2-E[8\alpha(a+\Gamma_D+2)]+16\alpha^2(a+\Gamma_D+1)\\ &\quad\times(a+\Gamma_D+3)-64\beta(a+\Gamma_D+1)(J-1). \end{aligned} \tag{2.12}$$

It is easily seen that when J is a positive integer, exact energy eigenvalues for the first J levels are known. Further, when J is a positive integer, these polynomials exhibit the factorization property, as given by

$$P_{n+J}(E)=P_J(E)Q_n(E), \tag{2.13}$$

where the polynomial set $Q_n(E)$ corresponds to the nonexact spectrum for this problem with $Q_0(E)=1$. For example, for $J=1$, P_{n+1} will be factorized into P_1 and Q_n and the corresponding QES energy level (which in this case is the ground state) is obtained by putting $P_1=0$, i.e.,

$$E_1=4\alpha(a+\Gamma_D+1). \tag{2.14}$$

Similarly for $J=2$, P_{n+2} will be factorized into P_2 and Q_n and the corresponding energy levels are

$$\begin{aligned} E_1&=4\alpha(a+\Gamma_D+2)-4\sqrt{\alpha^2+4\beta(a+\Gamma_D+1)},\\ E_2&=4\alpha(a+\Gamma_D+2)+4\sqrt{\alpha^2+4\beta(a+\Gamma_D+1)}. \end{aligned} \tag{2.15}$$

The quotient polynomials $Q_n(E)$ also form an orthogonal set as they satisfy the recursion relation

$$\begin{aligned} &Q_n(E)+[E-4\alpha\{2(n+J)-1+a+\Gamma_D\}]Q_{n-1}(E)\\ &\quad-64\beta(n+J-1)(n+J-1+a+\Gamma_D)(n-1)Q_{n-2}(E)=0, \end{aligned} \tag{2.16}$$

with $Q_0(E)=1$ and $Q_{-1}(E)=0$.

The square norm of both P_n and Q_n polynomials can be calculated from the recursion relations in Eqs. (2.11) and (2.16), respectively. We obtain

$$\gamma_n^P = 64\beta \prod_{k=1}^{n} k(k+a+\Gamma_D)(J-n),$$
$$\gamma_n^Q = 64\beta \prod_{k=1}^{n} (k+J)(k+J+a+\Gamma_D). \tag{2.17}$$

Note that à la the Bender–Dunne case,[6] the norm for the P_n polynomials vanish for $n \geqslant J$ while it is positive definite for $n<J$. This is an alternative characterization of the QES system. On the other hand, the norm for the Q_n polynomials is always nonvanishing and positive definite.

One can also calculate the weight functions ω_k for the set $P_n(E)$ by using the relations[16]

$$\sum_{k=1}^{J} P_n(E_k)\omega_k = \delta_{n0}, \quad n=0,1,...,J-1. \tag{2.18}$$

For example, for $J=2$ we obtain

$$\omega_1 = \frac{1}{2} + \frac{\alpha}{2\sqrt{\alpha^2+4\beta(a+\Gamma_D+1)}},$$
$$\omega_2 = \frac{1}{2} - \frac{\alpha}{2\sqrt{\alpha^2+4\beta(a+\Gamma_D+1)}}. \tag{2.19}$$

Notice that to the leading order, the weight functions are inversely proportional to the particle number N for a fixed D, while they are inversely proportional to the square root of the space dimension D, when the particle number N remains unchanged. It may also be noted that both the weight functions ω_1 and ω_2 are positive definite. Actually, one can prove on very general grounds that for any J, the weight functions in this case will always be positive. As has been shown by Finkel *et al.*,[15] if the three-term recursion relation is of the form

$$P_{k+1} = (E-b_k)P_k - a_k P_{k-1}, \quad k \geqslant 0, \tag{2.20}$$

with $a_0=0$ and $a_{n+1}=0$, then (i) the weight functions are all positive, (ii) the norm of the polynomials (γ_n^p) is positive for $n<J$, provided b_k is real for $0 \leqslant k \leqslant n$ and $a_k>0$ for $1 \leqslant k \leqslant n$. Using Eq. (2.11) it is then easy to see that for any J, the weight functions and γ_n^p (for $n<J$) will be positive.

One can also calculate the moments $\mu_n = \int dE\ E^n \omega(E)$ of the weight functions. It is easily shown that both the odd and the even moments are nonzero. Further, to the leading order,

$$\mu_n = [4\alpha(a+\Gamma_D+1)]^n + \cdots . \tag{2.21}$$

Thus, to the leading order, the growth rate is proportional to N^{2n} for fixed D and proportional to D^n for fixed N.

Finally, let us discuss the consequences of the anti-isospectral transformation (also termed the duality transformation)[12] in the context of our QES problem. It is easily seen that because of the duality transformation ($x \to ix$), the QES levels of the potential in Eq. (2.5) are related to that of a similar potential (say $\hat{V}$), where C (and hence α) has been replaced by $-C$. In particular, if J levels of the potential in Eq. (2.5) are QES levels with energy eigenvalues and eigenfunctions E_k and ψ_k, respectively ($k=1,2,...,J$), then the corresponding QES eigenvalues and eigenfunctions of $\hat{V}$ are given by

$$\hat{E}_k = -E_{J+1-k}, \quad \hat{\psi}_k(x) = \psi_{J+1-k}(ix). \tag{2.22}$$

For example, for $J=2$, the QES energy eigenvalues of $\hat{V}$ are same as those given by Eq. (2.15), with the precise relationship being $\hat{E}_{1,2} = -E_{2,1}$. The other properties of the dual potential are also similarly related and can be easily worked out.

B. Self-dual potential

Let us now discuss the special case of $C=0$ in Eq. (2.5), in which case we have a self-dual QES potential which has remarkably simple properties. For example, the recursion relation (2.11) now takes the simpler form (note that now $\alpha=C/4\sqrt{H}=0$),

$$P_n(E)+EP_{n-1}(E)-64\beta(n-1)(n-1+a+\Gamma_D)(n-J-1)P_{n-2}(E)=0, \tag{2.23}$$

and similarly the set Q_n satisfy Eq. (2.16) with $\alpha=0$. As a result, both $P_n(E)$ and $Q_n(E)$ are eigenfunctions of parity. First, few polynomials generated by this recursion relation are given by ($s\equiv 1+a+\Gamma_D$)

$$\begin{aligned} P_1&=-E,\\ P_2&=E^2-64\beta s(J-1),\\ P_3&=-E^3+E[64\beta s(J-1)+128\beta(s+1)(J-2)],\\ P_4&=E^4-E^2[64\beta s(J-1)+128\beta(s+1)(J-2)\\ &\quad+192\beta(s+2)(J-3)]+(64\beta)^2 3s(s+2)(J-1)(J-3). \end{aligned} \tag{2.24}$$

From the duality relation (2.22), it follows that for this self-dual system, all the energy eigenvalues must be symmetrically distributed around $E=0$. Further, in this case QES eigenvalues can be analytically calculated very simply for $J=1,2,...,5$. For example, the eigenvalues for $J=3$ and $J=4$ are given as follows: $J=3$,

$$\begin{aligned} E_1&=-8\sqrt{2\beta(2s+1)},\\ E_2&=0,\\ E_3&=8\sqrt{2\beta(2s+1)}, \end{aligned} \tag{2.25}$$

and $J=4$,

$$\begin{aligned} E_1&=-\sqrt{320\beta(s+1)+64\beta\sqrt{16s(s+2)+25}},\\ E_2&=-\sqrt{320\beta(s+1)-64\beta\sqrt{16s(s+2)+25}},\\ E_3&=\sqrt{320\beta(s+1)-64\beta\sqrt{16s(s+2)+25}},\\ E_4&=-\sqrt{320\beta(s+1)+64\beta\sqrt{16s(s+2)+25}}. \end{aligned} \tag{2.26}$$

The square norms of the polynomial sets P_n and $Q_n s$ are again given by Eq. (2.17). Since the polynomials have definite parity, the odd moments vanish. The even moments of the weight functions in the leading order are given by

$$\mu_{2n}\sim[(1+a+\Gamma_D)]^n+\cdots . \tag{2.27}$$

This means μ_{2n} is proportional to N^{2n} for fixed D and proportional to D^n for fixed N. It is worth noting that in the limit $C=0$ (and hence $\alpha=0$), the leading term in the expansion of μ_n in Eq. (2.21) vanishes, and that is why the leading behavior of μ_n as a function of $(1+a+\Gamma)$ when C (and hence α) is nonzero is the same as that of μ_{2n} when C (and hence α)$=0$.

The weight functions in this self-dual QES model are related through a nice relation,

$$\omega_k=\omega_{J+1-k}. \tag{2.28}$$

This can be proved explicitly using relation (2.18) as follows.

We consider n to be odd in Eq. (2.18), i.e.,

$$\sum_k P_{2m+1}(E_k)\omega_k=0, \quad \text{for} \quad k=1,2,...,J. \tag{2.29}$$

Now, since the $P_n(E)$ are of definite parity of E, we can write

$$P_{2m+1}(E_k)=\sum_{n=0}^{m} a_n E_k^{2n+1}, \qquad m=0,1,2,\ldots, \tag{2.30}$$

when a_n are (E-independent) constants and $a_m=-1$.

Putting (2.30) in (2.29), we obtain

$$\sum_{k=1}^{J}\sum_{n=0}^{m} a_n E_k^{2n+1}\omega_k=0. \tag{2.31}$$

This implies that

$$\sum_{k=1}^{J} E_k^{2m+1}\omega_k=0. \tag{2.32}$$

Now in Eq. (2.32), we use the duality relation between the eigenvalues of the self-dual system, i.e., $E_k=-E_{J+1-k}$, to obtain

$$\sum_{k=1}^{J} E_{J+1-k}^{2m+1}\omega_k=0. \tag{2.33}$$

Next, we make a change of variable $k=J+1-j$ in Eq. (2.33) to obtain

$$\sum_{j=1}^{J} E_j^{2m+1}\omega_{J+1-j}=0. \tag{2.34}$$

Subtracting Eq. (2.34) from Eq. (2.32),

$$\sum_{k=1}^{J} E_k^{2m+1}[\omega_k-\omega_{J+1-k}]=0. \tag{2.35}$$

This is true for arbitrary E_k, and hence the result, $\omega_k=\omega_{J+1-k}$.

We now explicitly calculate the weight functions for the case of $J=3$ and $J=4$. We find, for $J=3$,

$$\omega_1=\omega_3=\frac{s}{2(2s+1)}, \qquad \omega_2=\frac{s+1}{2s+2}, \tag{26}$$

and for $J=4$,

$$\omega_1=\omega_4=\frac{1}{4}\left[1-\frac{2s+5}{\sqrt{16s(s+2)+25}}\right], \qquad \omega_2=\omega_3=\frac{1}{4}\left[1+\frac{2s+5}{\sqrt{16s(s+2)+25}}\right]. \tag{2.37}$$

Not surprisingly, the weight function satisfies relation (2.28) and $\Sigma_{k=1}^{J}\ \omega_k=1$.

C. An unusual QES problem

Let us now discuss an unusual QES problem. To that end consider the following potential:

$$V(\rho)=\frac{1}{2}\left[B^2\rho^2-\frac{C}{\rho}+\frac{F}{\rho^2}\right]. \tag{2.38}$$

It is worth noting that for $C=0$ as well as for $B=0$, a class of energy levels can be analytically obtained.[9,13] The special case of $N=2$, i.e., 2-anyons in a magnetic field and experiencing Coulomb potential has been discussed by Myrheim *et al.*[17] We want to show that when both of them are nonzero, it is an example of an unusual QES system, in the sense that for a one-parameter family of potentials (when there is a specific relation between B, C, and F), one always obtains one QES eigenvalue.

On substituting

$$\phi(\rho)=\rho^a\ \exp(-B\rho^2/2)\,\eta(\rho), \tag{2.39}$$

in Eq. (2.3), we obtain

$$\eta''(\rho)+\left[\frac{2a+2\Gamma_D+1}{\rho}-(2B\rho)\right]\eta'(\rho)+\left[E-2(a+\Gamma_D+1)B+\frac{C}{\rho}\right]\eta(\rho)=0, \tag{2.40}$$

where we have chosen $F=a^2+2a\Gamma_D$. Next, we substitute

$$\eta(\rho)=\sum_n\ P_n(E)\rho^n \tag{2.41}$$

in Eq. (2.40) and obtain the recursion relation satisfied by $P_n(E)$ as

$$n(2a+2\Gamma_D+n)P_n(E)+CP_{n-1}(E)+[E-2(n-1+a+\Gamma_D)]P_{n-2}(E)=0, \tag{2.42}$$

with initial conditions $P_{-1}=0$ and $P_0=1$.

From the recursion relation, it follows that the QES levels are obtained in case the coefficient of $P_{n-2}(E)$ vanishes, i.e., the energy of the QES levels is given by $E=2(a+\Gamma_D+n+1)$, where $n=1,2,\ldots$. Further, in that case B, C, and F are not independent. For example, for $n=1$, the relation is $C^2=2(2a+2\Gamma_D+1)B$, while for $n=2$ the relation is $C^2=4(4a+4\Gamma_D+3)B$. The QES level corresponds to an excited (ground) state depending on if $C>(<)0$. On comparing Eq. (2.42) with Eq. (2.10), we conclude that the polynomials in this case do not correspond to orthogonal polynomials.

III. OTHER *N*-BODY PROBLEMS

In this section we briefly discuss two other many-body problems and we will see that in both of these cases we have two QES problems which are very similar to those discussed in the last section.

A. Novel correlations for *N* particles

Recently, a class of exact solutions including the bosonic ground state has been obtained for a many-body Hamiltonian in two dimensions, such that the wave functions have a novel correlation (of the form $X_{ij}=x_iy_j-x_jy_i$) built into them.[10] Here we would like to show that QES states with novel correlations can also be obtained in this case. To that purpose we start from the Hamiltonian,

$$H=-\frac{\hbar^2}{2m}\sum_i \nabla_i^2+\frac{\hbar^2}{2m}\,g_1\sum_{i,j}\ \frac{\mathbf{r}_{ij}^2}{X_{ij}^2}+\frac{\hbar^2}{2m}\,g_2\sum\ \frac{\mathbf{r}_j\cdot\mathbf{r}_k}{X_{ij}\cdot X_{ik}}$$
$$+\frac{\hbar^2}{2m}\left[2\left\{B\sum_i\ r_i^2+C\left(\sum_i\ r_i^2\right)^2+H\left(\sum_i\ r_i^2\right)^3+\frac{F}{\Sigma_i r_i^2}\right\}\right]. \tag{3.1}$$

On substituting the ansatz,

$$\psi=\prod_{i<j}^{N}\ |X_{ij}|^g\phi(\rho), \tag{3.2}$$

in the Schrödinger equation $H\psi=\epsilon\psi$ $(m=\hbar=1)$, one can show that ϕ satisfies the equation

$$\phi''+\frac{2\Delta+1}{\rho}\ \phi'+\left[E-B\rho^2-C\rho^4-H\rho^6-\frac{F}{\rho^2}\right]\phi=0, \tag{3.3}$$

where

$$2\Delta+1=2N-1+2gN(N-1),$$
$$\rho^2=\sum_i\ r_i^2;\quad E=2\epsilon;\quad g_1=g(g-1),\quad g_2=g^2. \tag{3.4}$$

On substituting

$$\phi(\rho)=\rho^a\ \exp(-\alpha\rho^2-\beta\rho^4)\,\eta(\rho), \tag{3.5}$$

in Eq. (3.3), we obtain

$$\eta''(\rho)+\left[\frac{2a+2\Delta+1}{\rho}-(4\alpha\rho+8\beta\rho^3)\right]\eta'(\rho)$$
$$+[E-4\alpha(a+\Delta+1)+\rho^2(4\alpha^2-B-8\beta(a+\Delta+2))]\,\eta(\rho)=0, \tag{3.6}$$

where we have chosen $\alpha=C/4\sqrt{H}$; $\beta=\sqrt{H}/4$; and $F=a^2+2a\Delta$. Now notice that this equation is similar to the corresponding equation in the last section, and hence the rest of the analysis of that section will be considered step by step in this case. Thus, this is another example of a QES N-body problem in two dimensions, where the polynomials satisfy most of the properties of the Bender–Dunne polynomials. Similarly, one can also consider the admixture of the oscillator and Coulomb-type potentials and, as in the last section, again obtain similar QES solutions. In particular, in this case the polynomials do not form an orthogonal set.

B. *N*-body QES problem in one-dimension

Finally, we would like to briefly discuss the Calogero–Sutherland-type N-body QES problems in one dimension.[2,3]

The Hamiltonian corresponding to the N-body QES problem in one-dimension is given by

$$H=-\frac{\hbar^2}{2m}\sum_i\ \nabla_i^2+\sum_{i<j}\ \frac{g}{(x_i-x_j)^2}+B\sum_i\ x_i^2+C\left[\sum\ x_i^2\right]^2 H\left[\sum\ x_i^2\right]^3+\frac{F}{\Sigma x_i^2}. \tag{3.7}$$

We substitute the standard ansatz for the wave function ψ,

$$\psi=Z^{\lambda+1/2}\phi(\rho), \tag{3.8}$$

where $Z=\Pi_{i<j}(x_i-x_j)$, $\rho^2=\Sigma x_i^2$, $\lambda=\frac{1}{2}\sqrt{1+4g}$ in $H\psi=\epsilon\psi$ $(\hbar=m=1)$, and obtain the ϕ equation,

$$\phi''+\frac{2\Delta+1}{\rho}\ \phi'+\left[E-B\rho^2-C\rho^4-H\rho^6-\frac{F}{\rho^2}\right]\phi=0, \tag{3.9}$$

where $E=2\epsilon$ and

$$2\Delta+1=N-1+N(N-1)\lambda. \tag{3.10}$$

This equation is very similar to the Eq. (3.3), and hence the rest of the analysis goes through as in the last section, i.e., the polynomials of the QES problem indeed form an orthogonal set and share most of the properties of the Bender–Dunne polynomials. Similarly, we can show that even in this case the other QES solution corresponding to the admixture of the oscillator and Coulomb potential also exists, and in this case, the polynomials do not form an orthogonal set.

IV. SUMMARY AND OPEN PROBLEMS

In this paper we have obtained QES solutions of a number of N-body problems in one, two, and higher dimensions. We have shown that with a sextic-type potential, in all these cases one has a set of orthogonal polynomials which satisfy almost all the properties, as satisfied by the Bender–Dunne polynomials. However, the hidden algebra is not clear in these cases. In this context, recall that for the case of one particle in one dimension experiencing the sextic potential, the underlying symmetry algebra is $Sl(2)$. In particular, in that case one can express the Hamiltonian in terms of the quadratic (and linear) generators of $Sl(2)$. Further, it will be nice to discover some other QES many-body problems. Finally, it would be most interesting if one can find some application of these QES N-body problems. In this context it is worth noting that one application has already been found by one of the present authors (AK) and Jatkar.[18] In particular, they have considered the model discussed in Sec. III B (with $F=0$) and have shown that the square of the ground state wave function of this model is related to the matrix model corresponding to branched polymers. It will really be interesting if one can find some other application of the N-body QES problems in two and higher dimensions.

[1] A. Ushveridze, *Quasi-Exactly Solvable Models in Quantum Mechanics* (Institute of Physics Publishing, Bristol, 1994).
[2] A. G. Ushveridze, Mod. Phys. Lett. A **6**, 977 (1991).
[3] A. Minzoni, M. Rosenbaum, and A. Turbiner, hep-th/9606092.
[4] D. P. Jatkar, C. Nagaraja Kumar, and A. Khare, Phys. Lett. A **142**, 200 (1989).
[5] A. Khare and B. P. Mandal, Phys. Lett. A **239**, 197 (1998).
[6] C. M. Bender and G. V. Dunne, J. Math. Phys. **37**, 6 (1996).
[7] A. Khare and B. P. Mandal, J. Math. Phys. **39**, 3476 (1998).
[8] F. Calogero and C. Marchioro, J. Math. Phys. **14**, 182 (1973).
[9] A. Khare and K. Ray, Phys. Lett. A **230**, 139 (1997).
[10] M. V. N. Murthy, R. K. Bhaduri, and D. Sen, Phys. Rev. Lett. **76**, 4103 (1996); R. K. Bhaduri, A. Khare, J. Law, M. V. N. Murthy, and D. Sen, J. Phys. A Math. Gen. **30**, 2557 (1997).
[11] F. Calogero, J. Math. Phys. **10**, 2191 (1969); **10**, 2197 (1969); **12**, 419 (1971).
[12] A. Krajewska, A. Ushveridze, and Z. Walczak, Mod. Phys. Lett. A **12**, 1225 (1997).
[13] A. Khare, cond-mat/9712133; J. Math. Phys. (in press).
[14] T. S. Chihara, *An Introduction to Orthogonal Polynomials* (Gordon and Breach, New York, 1975); A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. II.
[15] F. Finkel, A. Gonzalez-Lopez, and M. A. Rodriguez, J. Math. Phys. **37**, 3954 (1996).
[16] A. Krajewska, A. Ushveridze, and Z. Walczak, Mod. Phys. Lett. A **12**, 1131 (1997).
[17] J. Myrheim, E. Halvorsen, and A. Vercin, Phys. Lett. B **278**, 171 (1992).
[18] D. Jatkar and A. Khare, Int. J. Mod. Phys. A **11**, 1357 (1996).

SU(2) quantum kinematics: Rotation-observable versus angular-momentum generalized commutation relations

J. Krause[a)]
Facultad de Física, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile

(Received 21 January 1998; accepted for publication 27 July 1998)

The canonical commutation relations of quantum mechanics are generalized to the case where appropriate dynamical variables are angular-momentum, rotation-angle, and rotation-axis observables. To this end, SU(2) is ''quantized'' on the compact group manifold, according to the standard procedure of non-Abelian quantum kinematics. Quantum-kinematic invariant operators are introduced, and their commutation relations with the rotation variables are found in an explicit manner. The quantum-kinematic invariants yield superselection rules in the form of eigenvalue equations of an isotopic structure (which one should solve in the applications, in order to get multiplets that carry the irreducible representations of the underlying quantum kinematic models). A wide range of applicability of SU(2) quantum kinematics is suggested. © *1998 American Institute of Physics.*
[S0022-2488(98)03111-9]

I. INTRODUCTION

The problem of extending the canonical commutation relations to non-Cartesian dynamical variables is a long-standing question in the general setting of quantum mechanics.[1] Most attempts to solve this problem have been focused on the search of new quantization methods for obtaining generalized minimum-uncertainty states of systems which are primarily described by non-Abelian dynamical variables.[2] However, important as it is, this particular purpose may be too narrow a motivation for such an ambitious endeavor.[3] Indeed, it is quite conceivable that some unknown general constraints should be taken into account for extending the conventional (i.e., Heisenberg) commutation relations to the appropriate ones. Kinematic constraints follow, for instance, from the superselection rules due to the symmetry group that characterizes a system.[4] Likewise, the search for generalized coherent states would be unable by itself to disclose such constraints. Other important motivations for having generalized canonical commutation relations in quantum mechanics appear in the current literature, as well.[5] As a matter of fact, this question has been already so amply discussed in the literature, for so many years, that it seems unnecessary to repeat here the sound physical motivation behind this issue.[6] This paper addresses this fundamental problem for the particular case of rotation configuration variables and angular momenta. Let us here remark that this is a problem in quantum kinematics.[7]

Much progress has been achieved during the last several years on the subject of group-*quantization* and non-Abelian quantum kinematics as used in this paper.[8] Here we shall follow a new group-theoretic approach to quantum kinematics which is quite consistent with the formalism of unitary symmetries used in ordinary quantum mechanics.[9] The new formalism, however, is enhanced by quantizing the parameters of the group (cf. below). In this way, quantum kinematics appears as a self-contained theory which can yield interesting models evolving on the arena of the group manifold. These abstract models can be projected on the configuration states of a physical system (on which the quantized group acts as a symmetry group), and thus a quantal model of the system arises (for details, see Ref. 8). Quantum kinematic theory yields a self-contained formalism of quantum mechanics, for it relinquished the use of ''prequantized'' classical analogs, and everything stems from the assumed symmetry group. Therefore, quantum kinematic models become

a)Electronic mail: jkrause@chopin.fis.puc.cl

 © 1998 American Institute of Physics

interesting in themselves if the quantized group is a physically relevant symmetry group.[10,11]

The analysis of rotational symmetry, and the behavior of physical quantities under rotations, is one of the most common problems in the study of the symmetries of physical systems. In addition, the theory of angular momentum is the prototype of continuous symmetry groups of many types now found useful in the classification of the internal symmetries of elementary particle physics. So it was an important contribution of Wigner[12] to note that SU(2) is the group that enters quantum physics. Of course, the main consequences of this fact are well known and have been studied in the literature for a long time.[13] Nevertheless, some SU(2) quantal features will be considered here under the wide perspective offered by quantum kinematics. In fact, the main point dealt with in this paper is to obtain *generalized Heisenberg commutation relations* for *angular-momenta* and *rotation-angle* operators. (As far as we know, all the fundamental commutation relations presented in this paper are new.) To this end, we use the Euler–Rodrigues parameters[13] for describing the elements of the group, we then introduce the regular representation and we quantize SU(2) on the group manifold. This means that we replace the rotation-angle and the rotation-axis parameters (which are *c*-numbers that label the elements of the group) by a complete set of commuting Hermitian operators, acting as *generalized* position operators of the group manifold, which admit the parameters for spectra. In this approach, the generators of the representation afford the *generalized* momentum operators (i.e., angular momenta) of the model.[4] [The angular momenta and rotation variables mentioned here need not be taken literally, but may be understood in a wide sense. However, it will do just as well if the intuitive example of SO(3), acting in ordinary space, is kept in mind.] As we shall see, according to this formalism one is in a position to introduce a generalized SU(2) Weyl–Heisenberg algebra, and one obtains three quantum-kinematic invariant operators;[14] i.e., superselection rules follow, which reduce the model.

We wish to emphasize that this paper has only an introductory character, since the quantum kinematic theory of SU(2) discussed in the space allotted here is by no means complete. Given the relevance of this group, the endeavor of SU(2) quantum kinematics is so conspicuous, that each of its main features deserves a separate study by itself. Physical applications of SU(2) quantum kinematic theory are many. [Some applications will be given in some forthcoming papers. (For more comments, cf. Sec. IV, below.)] For instance, in the current literature (concerning ordinary rotations in quantum mechanics) the Euler-angle parametrization is predominantly used for describing the elements of SO(3), of which SU(2) is the universal covering group. Thus, in particular, we wish to remark that the introduction of SO(3) quantum kinematics (in terms of fundamental commutation relations for Euler-angle versus angular-momentum operators) follows from the theory presented in this paper, since (of course) the Euler-angles are related with the Euler–Rodrigues parameters in a well-defined way.[13] Strictly speaking, however, this subject belongs to the applications of the present theory, and thus it will be considered elsewhere. Here we are interested exclusively in the SU(2) quantum kinematic theory, for which the discussion in terms of the Euler–Rodrigues variables is the simplest one.[13]

The plan of the paper is the following. Section II deals with group quantization. For quantizing SU(2) we use an embedding approach that describes the group manifold as the surface $\mathcal{S}_3$ of the unit sphere in $\mathcal{E}_4$, in which SU(2) appears as a subgroup of the four-dimensional group of real quaternions.[13] [With this aim, in this paper we adopt the Euler–Rodrigues parameters,[13] which are best suited to this end (cf. Appendix A).] We then deduce the associated quantum kinematic commutation relations within the left regular representation, which is the basic working frame adopted in this article. [For the sake of completeness, both (left and right) regular representations of SU(2) are briefly dealt with in Appendix B.] Section III contains an overview of quantum kinematic invariant operators, which play a central role in this theory. Thus, in Secs. II and III, our discussion will be focused on the most general kinematic features of SU(2)-quantized-group models (although we do not solve for any concrete model in the present paper). Finally, Sec. IV contains our concluding remarks and some perspectives for future work.

II. QUANTUM KINEMATIC COMMUTATION RELATIONS

We begin our work by introducing generalized position operators in the group manifold $\mathcal{S}_3$ of SU(2). Here, $\mathcal{S}_3$ denotes the three-dimensional surface of the unit sphere in four-dimensional Euclidean space $\mathcal{E}_4$, defined by $q^\mu q^\mu = 1$. The four Cartesian variables $q^\mu = (\mathbf{q}, q^4) \in \mathcal{E}_4$, when

TABLE I. Synopsis of the left regular representation formalism on $\mathcal{S}_3$, within the carrier Hilbert space $\mathcal{H}[\mathrm{SU}(2)]$. (For details, see Appendix B.)

$\langle q'\|q\rangle=\mu_0^{-}L(q)\delta^{(3)}(q'-q), \quad \oint_{\mathcal{S}_3} d\mu(q)\|q\rangle\langle q\|=I$	(I.1)
$U(q')\|q\rangle=\|g(q';q)\rangle, \quad U(q)=\oint_{\mathcal{S}_3} d\mu(q')\|g(q;q')\rangle\langle q'\|$	(I.2)
$U(q')U(q)=U[(g(q';q)], \quad U^{\dagger}(q)=U^{-1}(q)=U(\overline{q})$	(I.3)
$U(e+\delta q)=I-\frac{i}{\hbar}\,\delta q^j L_j, \quad U^{\dagger}(q)L_jU(q)=A_{jk}(q)L_k$	(I.4)
$L_j\|q\rangle=i\hbar X_j(q)\|q\rangle, \quad [L_j,L_k]=-2i\hbar\epsilon_{jk\ell}L_{\ell}$	(I.5)

restricted to $q\in\mathcal{S}_3\subset\mathcal{E}_4$, play the role of the Euler–Rodrigues parameters which label the elements of the group in a faithful way.[13] (See Appendix A for some useful details.) We then focus our attention on the regular representation,[15] since this is the paramount structure in the kinematic quantization method. [The regular representation of SU(2) is reviewed in Appendix B.]

Now, given the resolution of the identity stated in Eq. (I.1) (cf. Table I), we define a set of operators Q^{μ} by means of the following spectral integrals:

$$Q^{\mu}=\oint_{\mathcal{S}_3} d\mu(q)\|q\rangle q^{\mu}\langle q\|, \tag{2.1}$$

where $d\mu(q)$ denotes the Hurwitz invariant measure on $\mathcal{S}_3$ defined in Eq. (A17). This is elementary, but fundamental. In general, this means that in Lie group theory one can consider integrals over the group manifold, like (2.1), as a basic geometric definition of quantization. One borrows this definition from standard quantum mechanics. As we will see through the chosen example, in this way one achieves a consistent and fruitful generalized setting for quantum theory. The Q's are generalized position operators on the group manifold $\mathcal{S}_3$, since they satisfy the following demands for such qualification:

$$Q^{\mu\dagger}=Q^{\mu}, \quad [Q^{\mu},Q^{\nu}]=0, \quad Q^{\mu}\|q\rangle=q^{\mu}\|q\rangle. \tag{2.2}$$

In particular, we observe that these operators satisfy the required constraint:

$$Q^{\mu}Q^{\mu}=I. \tag{2.3}$$

So we quantize the group.

In this article we adopt the left-regular representation as the working frame for developing the theory (cf. Appendix B); namely, we here consider exclusively the left unitary group-operators $\mathrm{U}(q)$ defined in Eq. (B7). (See also Table I.)

We next show that the Q's are interesting mathematical objects. We first look for the kinematic law [under SU(2) transformations] obeyed by the position operators of the group. According to their definition, it turns out that the position operators defined in Eq. (2.1) obey the same quaternion kinematic law (A2) that characterizes the parameters of the group SU(2) on $\mathcal{S}_3$:

$$U^{\dagger}(q)Q^{\mu}U(q)=g^{\mu}(q;Q)=\sigma^{\mu}_{\nu\lambda}q^{\nu}Q^{\lambda}. \tag{2.4}$$

$U(q)$ denotes the group operators [cf. Table I, Eq. (I.2)], and the group coefficients $\sigma^{\mu}_{\nu\lambda}$ are those used in Eq. (A2). It must be borne in mind that, in the present theory, Eq. (2.4) is *not* an assumption. As a matter of fact, this analytic identity is a general result of group-theoretic quantum kinematics,[6] whose proof does not demand an explicit realization of the operators involved. One proves Eq. (2.4) in a few lines. Since one adopts the regular representation of the quantized group as the working frame of the formalism, from definition (2.1), using Eqs. (I.2a) and (I.3b) of Table I, and recalling that $d\mu(q)$ is an invariant SU(2) measure on $\mathcal{S}_3$, it follows:

$$U^{\dagger}(q)Q^{\mu}U(q) = \oint_{\mathcal{S}_3} d\mu(q')U^{\dagger}(q)|q'\rangle q'^{\mu}\langle q'|U(q)$$

$$= \oint_{\mathcal{S}_3} d\mu(q')|g(\overline{q};q')\rangle q'^{\mu}\langle g(\overline{q};q')|$$

$$= \oint_{\mathcal{S}_3} d\mu[g(q;q'')]|q''\rangle g^{\mu}(q;q'')\langle q''|$$

$$= \oint_{\mathcal{S}_3} d\mu(q'')|q''\rangle g^{\mu}(q;q'')\langle q''| = g^{\mu}(q;Q), \tag{2.5}$$

which according to (A2) is precisely Eq. (2.4). This is one of the most important results provided by the Lie group quantization technique.

If we now consider the infinitesimal transformation $U^{\dagger}(e+\delta q)Q^{\mu}U(e+\delta q)$, Eq. (I.4) (Table I) and Eq. (2.4) lead us immediately to a set of generalized Heisenberg commutation relations for the position operators on $\mathcal{S}_3$ with the non-Abelian momenta represented by the SU(2) generators. Indeed, a typical group-theoretic calculation yields the desired commutation rules, which read

$$[Q^{\mu},L_j] = i\hbar R_j^{\mu}(Q) = i\hbar\sigma_{j\nu}^{\mu}Q^{\nu} = i\hbar\begin{bmatrix} Q^4 & Q^3 & -Q^2 & -Q^1 \\ -Q^3 & Q^4 & Q^1 & -Q^2 \\ Q^2 & -Q^1 & Q^2 & -Q^3 \end{bmatrix}. \tag{2.6}$$

Here we have written L_j $(j=1,2,3)$ to denote the generators of the left regular representation [see Appendix B, Eqs. (B10) and (B11)]. The entries $R_j^{\mu}(q)$ of the right transport matrix are given in Eq. (A4.1). It is important to note that the σ-coefficients satisfy

$$\sigma_{j\nu}^{\mu}\sigma_{k\lambda}^{\nu} - \sigma_{k\nu}^{\mu}\sigma_{j\lambda}^{\nu} + 2\epsilon_{jk\ell}\sigma_{\ell\lambda}^{\mu} = 0, \tag{2.7}$$

for these imply that the commutation relations (2.6) are consistent with the Lie algebra (B11), in the sense of the Jacobi identity. Furthermore, Eq. (2.6) is also consistent with the constraint $Q^{\mu}Q^{\mu}=I$, because on $\mathcal{S}_3$ one has: $\sigma_{j\nu}^{\mu}q^{\mu}q^{\nu}=0$. (These features are not trivial at all.)

This is the point where the quantum principle begins to emerge from the traditional (i.e., classical) theory of SU(2). Notice that, since $\mu=1,2,3,4$ and $j=1,2,3$, the generalized closed commutation relations shown in Eq. (2.6) do not correspond to a quantized symplectic structure, as do the canonical commutation relations (in the traditional Heisenberg form) of ordinary quantum mechanics. In this sense, considering future applications, let us here remark that it not necessary to redefine Eq. (2.6) in order to get some more conventional 3×3 matrix forms, instead of the 3×4 matrix form obtained here, because *all* four operators Q^{μ} stand on a same footing (see, below) and satisfy the constraint $Q^{\mu}Q^{\mu}=I$. Indeed, as we have discussed in our previous work, non-Abelian quantum kinematics leads us beyond the traditional canonical quantization methods.[8] Much of what follows is a consequence of the generalized commutation relations (2.6).

We now have the kinematic model. Let us proceed to examine the Q's more closely. In general, we define functions of the commuting position operators by means of their spectral integrals:

$$F(Q) = \oint_{\mathcal{S}_3} d\mu(q)|q\rangle F(q)\langle q| \tag{2.8}$$

(see Table I). Then it is worth noting that Eq. (2.6) yields the general commutation rule

$$[F(Q),L_j] = i\hbar R_j^{\mu}(Q)F_{,\mu}(Q) = i\hbar X_j(Q)F(Q), \tag{2.9}$$

where $F_{,\mu}(q)$ is a vector field in $\mathcal{E}_4$, defined by $F_{,\mu}(q)=(\partial/\partial q^{\mu})F(q)$, and where $X_j(q)$ denotes Lie's right vector fields on $\mathcal{S}_3$ [defined in Eq. (A10)]. [Certainly, the fact that the q's are not independent variables on $\mathcal{S}_3$ does not preclude the existence of $F_{,\mu}(q)$ when $q\in\mathcal{S}_3\subset\mathcal{E}_4$. See also the discussion in Appendix A, concerning Eqs. (A11) and (A12).]

Equation (2.9) is a very useful expression. For instance, let us define the rotation axis as follows: $F^j(q)=q^j/\sin\phi=q^j/\sqrt{1-(q^4)^2}\equiv\hat{n}^j$. In quantum kinematics, the rotation axis $\hat{\mathbf{n}}$ becomes a self-adjoint operator $\hat{\mathbf{N}}$: $\hat{\mathbf{n}}\to\hat{\mathbf{N}}$. Using Eq. (2.9), we easily find the generalized commutation relations obeyed by the Cartesian components $\hat{N}^j$ of the rotation-axis-operator with the generators L_j, which read

$$[\hat{N}^j,L_k]=i\hbar[\cot\Phi(\delta^j_k-\hat{N}^j\hat{N}^k)-\epsilon_{jk\not}\hat{N}^{\not}]. \tag{2.10}$$

[In fact, one has $X_k(q)F^j(q)=\cot\phi(\delta^j_k-\hat{n}^j\hat{n}^k)-\epsilon_{jk\not}\hat{n}^{\not}$. If one defines instead $G^j(q)=q^j/\sqrt{q^kq^k}\equiv\hat{n}^j$ (say) one obtains the same result (2.10). Indeed, Eq. (2.10) does not depend on how one defines the unit vector $\hat{\mathbf{n}}$ in terms of the q's. For the operator $\cot\Phi$, see Eq. (2.12), below.] The constraint $\hat{N}^j\hat{N}^j=I$ follows, since the rotation-axis operator $\hat{\mathbf{N}}=(\hat{N}^1,\hat{N}^2,\hat{N}^3)$ is a bounded operator given by

$$\hat{\mathbf{N}}=\mu_0\oint_{\mathcal{S}_3}d\Omega(\phi,\hat{\mathbf{n}})|\phi,\hat{\mathbf{n}}\rangle\hat{\mathbf{n}}\langle\phi,\hat{\mathbf{n}}|, \tag{2.11}$$

so that $\hat{\mathbf{N}}|\phi,\hat{\mathbf{n}}\rangle=\hat{\mathbf{n}}|\phi,\hat{\mathbf{n}}\rangle$ holds. [Concerning the $(\phi,\hat{\mathbf{n}})$ notation, see Appendix B].

In the same manner, we define two bounded operators $\cos\Phi$ and $\sin\Phi$ on $\mathcal{S}_3$:

$$\cos\Phi=\mu_0\oint_{\mathcal{S}_3}d\Omega(\phi,\hat{\mathbf{n}})|\phi,\hat{\mathbf{n}}\rangle\cos\phi\langle\phi,\hat{\mathbf{n}}|, \tag{2.12a}$$

$$\sin\Phi=\mu_0\oint_{\mathcal{S}_3}d\Omega(\phi,\hat{\mathbf{n}})|\phi\cdot\hat{\mathbf{n}}\rangle\sin\phi\langle\phi,\hat{\mathbf{n}}|, \tag{2.12b}$$

which satisfy the rather natural constraint $\cos^2\Phi+\sin^2\Phi=I$ and do not cause difficulties related to periodicity. As the reader can prove, for these operators we obtain the following commutation relations with the L's:

$$[\cos\Phi,L_j]=-i\hbar\hat{N}^j\sin\Phi,\quad[\sin\Phi,L_j]=i\hbar\hat{N}^j\cos\Phi, \tag{2.13}$$

which have a very suggestive form. [See also Eq. (2.17), below.] Similar commutation relations are well known for the (trivial) case of U(1) Abelian quantum kinematics.[1,2] Indeed, from Eq. (2.13) one obtains

$$[\cos\Phi,L]=-i\hbar\sin\Phi,\quad[\sin\Phi,L]=i\hbar\cos\Phi, \tag{2.14}$$

where the Hermitian operator $L=\frac{1}{2}(\hat{\mathbf{N}}\cdot\mathbf{L}+\mathbf{L}\cdot\hat{\mathbf{N}})$ may be interpreted as the U(1) angular-momentum generator in the direction of $\hat{\mathbf{N}}$. [Compare with Ref. 1, and with the references given in Ref. 2.] Equation (2.14) is quite familiar. However, to our best knowledge, Eqs. (2.10) and (2.13) do not figure in the literature, because in order to obtain the SU(2) position operators $(\Phi,\hat{\mathbf{N}})$ one has to quantize the group.

If now one considers the Euler–Rodrigues operators, i.e.,

$$Q^j=\hat{N}^j\sin\Phi,\quad Q^4=\cos\Phi \tag{2.15}$$

(cf. Appendix A), recalling the definition of the σ-coefficients, one can use Eq. (2.6) directly to obtain other basic commutation relations. For example, in this way, the commutators

$$[\hat{N}^j\sin\Phi,L_k]=i\hbar(\delta^j_k\cos\Phi-\epsilon_{jk\not}\hat{N}^{\not}\sin\Phi), \tag{2.16}$$

follow immediately from (2.6). [Of course, these commutators are consistent with Eqs. (2.10) and (2.13).]

Of particular interest are the following commutation relations:

$$[\cos 2\Phi,J_j]=-2i\hbar\hat{N}^j\sin 2\Phi,\quad[\sin 2\Phi,J_j]=2i\hbar\hat{N}^j\cos 2\Phi, \tag{2.17}$$

where we have written $2J_j = L_j$. [The reader will have no difficulty verifying Eq. (2.17), after a short calculation.] These commutation relations are called on to play an important role in the quantum kinematic theory of proper rotations, since the operators J_j are the standard angular-momentum operators in ordinary space, and 2Φ is the SO(3) rotation-angle operator about the axis-operator $\hat{\mathbf{N}}$. [See also Eq. (A9).]

All required generalized commutation relations can be obtained in this fashion, one way or another, from Eqs. (2.6) and (2.9). We have been unable to find such generalized Heisenberg commutation relations in the extensive literature devoted to SU(2).

III. QUANTUM KINEMATIC INVARIANT OPERATORS

We next take advantage of this technique, from a very general point of view. Let us first introduce the following matrix-operator [cf. Eq. (A7)]:

$$A_{jk}(Q) = Q^{\mu} Q_{\mu} \delta_{jk} + 2Q^j Q^k - 2Q^4 \epsilon_{jk\ell} Q^{\ell}, \tag{3.1}$$

which will be referred to as the antiadjoint operator. Since $U^{\dagger}(q)F(Q)U(q) = F[g(q;Q)]$ holds in general, from the discussion presented in Appendix A we conclude that the antiadjoint operator obeys the kinematic law given by the unitary transformation

$$U^{\dagger}(q)\overline{A}_{jk}(Q)U(q) = \overline{A}_{j\ell}(Q)\overline{A}_{\ell k}(q). \tag{3.2}$$

Hence a closed algebra follows immediately:

$$[\overline{A}_{jk}(Q), L_{\ell}] = -2i\hbar \epsilon_{k\ell m}\overline{A}_{jm}(Q). \tag{3.3}$$

Equation (3.3), together with the Lie algebra (B11) and the fact $[\overline{A}_{jk}(Q), \overline{A}_{\ell m}(Q)] = 0$, define the generalized Weyl–Heisenberg algebra of SU(2) quantum kinematics, which is analogous to the Weyl–Heisenberg algebra of the canonical commutation relations of ordinary quantum mechanics. This algebra is interesting, because it produces new invariant operators of the theory, as we shall see presently.

The main reason for introducing the antiadjoint operator of a Lie group, in general, is that it allows the definition of the following operators:[14]

$$\begin{aligned} R_j(Q;L) &= \overline{A}_{jk}(Q)L_k \\ &= (\cos^2\Phi - \sin^2\Phi)L_j + 2\sin^2\Phi \hat{N}^j \hat{N}^k L_k + 2\cos\Phi \sin\Phi \epsilon_{jk\ell}\hat{N}^k L_{\ell}, \end{aligned} \tag{3.4}$$

for which Eq. (3.2) together with (B12) imply the SU(2) invariant property:

$$U^{\dagger}(q)R_j(Q;L)U(q) = R_j(Q;L). \tag{3.5}$$

Thus one has

$$[R_j(Q;L), L_k] = 0, \tag{3.6}$$

for $j,k = 1,2,3$. Moreover, a rather lengthy but straightforward calculation, using the antiadjoint representation, yields $X_k(q)\overline{A}_{jk}(q) = 0$, which means that the invariant operators defined in (3.4) are Hermitian: $R_j^{\dagger}(Q;L) = R_j(Q;L)$, notwithstanding the fact that the Q's and the L's do not commute. Hence, in the kinematic theory of SU(2), besides the familiar Casimir operator $\mathbf{L}^2$, the group acquires three basic Hermitian invariant operators: R_1, R_2, R_3, which stem from the group-quantization procedure.[16]

Furthermore, using well-known properties of the adjoint representation, one proves that the basic invariant operators (3.4) satisfy the right SU(2) Lie algebra:

$$[R_j(Q;L), R_k(Q;L)] = 2i\hbar \epsilon_{jk\ell} R_{\ell}(Q;L) \tag{3.7}$$

TABLE II. Synopsis of the right regular representation formalism on $\mathcal{S}_3$, within the carrier Hilbert space $\mathcal{H}[\mathrm{SU}(2)]$.

$\langle q'\|q\rangle=\mu_0^{-1}R(q)\delta^{3)}(q'-q), \quad \oint_{\mathcal{S}_3} d\mu(q)\|q\rangle\langle q\|=I$	(II.1)
$V(q')\|q\rangle=\|g(q;q')\rangle, \quad V(q)=\oint_{\mathcal{S}_3} d\mu(q')\|g(q';q)\rangle\langle q'\|$	(II.2)
$V(q')V(q)=V[g(q;q')], \quad V^{\dagger}(q)=V^{-1}(q)=V(\overline{q})$	(II.3)
$V(e+\delta q)=I-\frac{i}{\hbar}\delta q^j R_j, \quad V^{\dagger}(q)R_jV(q)=\overline{A}_{jk}(q)R_k$	(II.4)
$R_j\|q\rangle=i\hbar Y_j(q)\|q\rangle, \quad [R_j,R_k]=2i\hbar\epsilon_{jk\not}R_{\not}$	(II.5)

[see Table II, Eq. (II.5)]. For the Casimir operator one has $\mathbf{L}^2=\mathbf{R}^2$. Therefore, we identify the operators $R_j(Q;L)$ as the generators of the right regular representation acting as invariant operators within the left representation of the group.[14] [See also, Eq. (A15).]

To end up this brief study, we consider the following commutation rules:

$$[F(Q),R_j]=i\hbar L_j^{\mu}(Q)F_{,\mu}(Q)=i\hbar Y_j(Q)F(Q), \tag{3.8}$$

where $Y_j(q)$ denotes Lie's left vector fields on $\mathcal{S}_3$ [defined in Eq. (A10)]. These rules yield, for instance,

$$[N^j \sin\Phi,R_k]=i\hbar(\delta_k^j\cos\Phi+\epsilon_{jk\not}N^{\not}\sin\Phi), \tag{3.9}$$

$$[\cos\Phi,R_j]=-i\hbar N^j\sin\Phi, \quad [\sin\Phi,R_j]=i\hbar N_j\cos\Phi. \tag{3.10}$$

Indeed, the kinematic calculus of generalized Heisenberg commutation relations for SU(2) can be applied easily to the kinematic invariant operators.

We here omit the calculations leading to these results. It is important to note, however, that one does not arrive at these results if one does not quantize the group.

IV. CONCLUDING REMARKS AND PERSPECTIVES

In our opinion, the quantum-kinematic operator calculus obtained in this paper bears a great potential value for physics, in general, because it stems from a standard procedure for handling symmetries in quantum theory. In our previous work, a group-quantization program has been developed for the case of noncompact non-Abelian Lie groups (cf. Ref. 8 and references quoted therein). The attained formalism has been applied to some models for which the Schrödinger equation, as well as the corresponding propagation kernel, have been deduced on a strict group-theoretic fashion.[10,11,17,18] It is rather clear, however, that further progress in quantum kinematics requires the consideration of symmetries described by compact Lie groups.

In particular, SU(2) symmetry must be examined upon this new perspective, as we have done in this introductory paper. Quantization of this fundamental symmetry is needed, for instance, for the development of the quantum kinematic theory of the complete Kepler group (recently introduced in the literature[19]), of the Poincaré group and the Galileo group in four-dimensional space–time, and of many other relevant Lie groups which contain SU(2) as a proper subgroup. Work is in progress concerning a companion paper to the present one, devoted to the quantum kinematic theory of the three-dimensional isotropic harmonic oscillator, in which SU(2) quantum kinematics (as introduced here) plays a central role. In fact, the piece of work presented in this paper is ripe for many physical applications.

In light of the previous results, the first point we want to note is that several uncertainty correlations (of the mean-square deviations between rotation observables and angular momentum), which appear in quantum kinematics, are worthy of a detailed study in every application of SU(2) in quantum theory. Maybe there are problems in which new physical predictions can be found by considering SU(2) quantum kinematic models in this sense.

We also have to remark that SU(2) quantum kinematics is quite different from the quantum theory of rotation angles of Barnett and Pegg.[20] As we have shown, the group-quantization approach goes much further into the very roots of the problem set by angle observables in quantum mechanics, than any other attempt considered previously.

We are now in position to use a maximal set of superselection rule operators, as given for instance by $\{\mathbf{R}^2, R_3\}$, in order to obtain full reduction (i.e., maximal diagonalization) of the regular representation. In other words, we can diagonalize the incoherent Hilbert space $\mathcal{H}[\mathrm{SU}(2)]$ into invariant subspaces $\mathcal{H}_{JM}[\mathrm{SU}(2)]$, each carrying an irreducible representation of the group, which corresponds to a SU(2) elementary system.[21] To this end, one must search for a complete orthogonal basis $\{|JMN\rangle\}$ within the left regular representation, associated with the complete set of commuting operators $\{\mathbf{R}^2, R_3, L_3\}$. That is, one looks for vectors satisfying the basic conditions

$$\langle JMN|J'M'N'\rangle = \delta_{JJ'}\delta_{MM'}\delta_{NN'}, \tag{4.1.1}$$

$$\sum_{J=0}^{\infty}\sum_{M=-J}^{J}\sum_{N=-J}^{J}|JMN\rangle\langle JMN| = I, \tag{4.1.2}$$

where the multiplets $|JMN\rangle$ are solutions to the following system of simultaneous eigenvalue equations:

$$\mathbf{R}^2|JMN\rangle = \hbar^2 J(J+2)|JMN\rangle. \tag{4.2.1}$$

$$R_3|JMN\rangle = \hbar M|JMN\rangle, \tag{4.2.2}$$

$$L_3|JMN\rangle = \hbar N|JMN\rangle. \tag{4.2.3}$$

Here: $J = 0,1,2,...$ and $M, N = -J, -J+2, ..., J-2, J$. This problem will be examined in a forthcoming paper.

Here we wish to remark only that, if one considers SU(2) [i.e., the universal covering group of SO(3)] acting as an external symmetry group of a system, then Eqs. (4.2.1) and (4.2.2) define automatically an isotopic angular momentum problem for that system, because both $\mathbf{R}^2$ and R_3 are invariant operators of SU(2). (N is not a good quantum number because L_3 is not an invariant operator.) This result is peculiar to quantum kinematics. It is a very reassuring feature, for it means that quantum kinematics brings into the fore an isotopic structure as a necessary theoretical construct, in a rather natural way indeed.

ACKNOWLEDGMENTS

This paper is devoted to the memory of Professor Lawrence C. Biedenharn, whose wise criticism and encouragement prompted much of the thoughts behind this work. The author acknowledges the referee for many remarks that have improved the presentation of this paper. This research was partially supported by FONDECYT, through Grant No. 1960256.

APPENDIX A: THE REAL-QUATERNION GROUP

It is well known that there is a one-to-one relation between points on the three-dimensional surface $\mathcal{S}_3$ (of the unit sphere in four-dimensional Euclidean space $\mathcal{E}_4$) and elements of SU(2): $\mathcal{S}_3 \ni q \leftrightarrow u(q) \in \mathrm{SU}(2)$, which is preserved by the group law;[13] briefly, the group manifold of SU(2) is the unit spherical surface $\mathcal{S}_3$. Since this is the basic framework adopted in this paper, a compact geometric notation for handling SU(2) on $\mathcal{S}_3$ will be introduced in this appendix, which is a very useful tool for quantum kinematics.

For simplicity and definiteness we label the group elements by means of the Euler–Rodrigues parameters:[13] $\mathbf{q} = \hat{\mathbf{n}} \sin\phi$ and $q^4 = \cos\phi$, with $0 \leqslant \phi \leqslant \pi$ and $\hat{\mathbf{n}}\cdot\hat{\mathbf{n}} = 1$. Thus we write the 2×2 matrices

$$u(q) = u(\mathbf{q}, q^4) = q^4\sigma_4 + iq^k\sigma_k = e^{i\phi\hat{\mathbf{n}}\cdot\sigma} \in \mathrm{SU}(2), \tag{A1}$$

which meaning is clear. One obtains the group law for SU(2) from the so-called quaternionic composition rule,[13] which can be written in a compact manner as follows:

$$q''^{\mu}=g^{\mu}(q';q)=\sigma^{\mu}_{\nu\lambda}q'^{\nu}q^{\lambda}. \tag{A2}$$

One defines the group coefficients $\sigma^{\mu}_{\nu\lambda}$ using well-known properties of the Pauli matrices. Thus one sets: $\sigma^{j}_{k\not4}=-\epsilon_{jk\not4}$, $\sigma^{\mu}_{4\nu}=\sigma^{\mu}_{\nu4}=\delta^{\mu}_{\nu}$, and $\sigma^{4}_{\mu\nu}=\eta_{\mu\nu}$, where the matrix $\eta_{\mu\nu}=\mathrm{diag}(---+)$ denotes the Minkowski metric. Using this notation, the inversion law for the Euler–Rodrigues parameters reads $\overline{q^{\mu}}=\eta_{\mu}=\eta_{\mu\nu}q^{\nu}$, since the identity point is at the ''north pole;'' i.e., $e^{\mu}=\delta^{\mu}_{4}=(0,0,0,1)\in\mathcal{S}_3$.

Useful relations obeyed by the group-coefficients follow from the group property. Especially, we note a remarkable feature of the quaternionic composition law:

$$q''^{\mu}q''^{\mu}=(q'^{\nu}q'^{\nu})(q^{\lambda}q^{\lambda}), \tag{A3}$$

from which the implication $(q';q)\in\mathcal{S}_3\times\mathcal{S}_3\Rightarrow q''\in\mathcal{S}_3$ follows. The Euler–Rodrigues variables are not essential parameters of the group, to be sure, for they do not correspond to independent variables. They are faithful parameters, however, because $q=(\mathbf{q},q^4)=(\hat{\mathbf{n}}\sin\phi,\cos\phi)$ covers the whole group manifold $\mathcal{S}_3$ in a strictly one-to-one way. [So we see that there is indeed a one-to-one relation between points on $\mathcal{S}_3$ and elements of SU(2), which is preserved by the group law.]

By means of the same group-coefficients $\sigma^{\mu}_{\nu\lambda}$ one defines another four-dimensional Lie group in $\mathcal{E}_4$. With this aim, all one needs to modify is the inversion law of the parameters, to read $\overline{q^{\mu}}=\rho^{-2}\eta_{\mu\nu}q^{\nu}$, where $\rho=\sqrt{q^{\mu}q^{\mu}}$, and one identifies the points $\overline{0}=\infty$ and $\overline{\infty}=0$. This group is the real-quaternion group;[13] we denote it by $\Sigma(4)$. The noncompact group $\Sigma(4)$ is isomorphic to $\mathcal{R}_x\otimes\mathrm{SU}(2)$, has the whole Euclidean space $\mathcal{E}_4$ for group manifold, and corresponds to the analytic continuation of SU(2) along the radius $0<p<\infty$. In fact, in this construct, SU(2) appears simply as that subgroup which arises from restricting the action of the quaternion group $\Sigma(4)$ to the locus $\mathcal{S}_3\subset\mathcal{E}_4$.

Since $\Sigma(4)$ is a noncompact (connected and simply connected) Lie group, right and left transport matrices are defined in the usual manner: $R^{\mu}_{\nu}(q)=\lim_{q'\to e}(\partial/\partial q'^{\nu})g^{\mu}(q';q)$ and $L^{\mu}_{\nu}(q)=\lim_{q'\to e}(\partial/\partial q'^{\nu})g^{\mu}(q;q')$, respectively. Expressions for these transport matrices in $\mathcal{E}_4$ are then obtained as follows:

$$R^{\mu}_{\nu}(q)=\sigma^{\mu}_{\nu\lambda}q^{\lambda}=\begin{bmatrix}\delta^{k}_{j}q^4+\epsilon_{jk\not4}q^{\not4} & -q^{j}\\ q^{k} & q^4\end{bmatrix}, \tag{A4.1}$$

$$L^{\mu}_{\nu}(q)=\sigma^{\mu}_{\lambda\nu}q^{\lambda}=\begin{bmatrix}\delta^{k}_{j}q^4-\epsilon_{jk\not4}q^{\not4} & -q^{j}\\ q^{k} & q^4\end{bmatrix}. \tag{A4.2}$$

By the same token, one gets the corresponding inverse transport matrices of $\Sigma(4)$, which are given by $\rho^{-2}\overline{R^{\mu}_{\nu}}(q)$ and $\rho^{-2}\overline{L^{\mu}_{\nu}}(q)$, where

$$\overline{R^{\mu}_{\nu}}(q)=\sigma^{\mu}_{\nu\lambda}\overline{q^{\lambda}}=\begin{bmatrix}\delta^{k}_{j}q^4-\epsilon_{jk\not4}q^{\not4} & q^{j}\\ -q^{k} & q^4\end{bmatrix}, \tag{A5.1}$$

$$\overline{L^{\mu}_{\nu}}(q)=\sigma^{\mu}_{\lambda\nu}\overline{q^{\lambda}}=\begin{bmatrix}\delta^{k}_{j}q^4+\epsilon_{jk\not4}q^{\not4} & q^{j}\\ -q^{k} & q^4\end{bmatrix}. \tag{A5.2}$$

[Here, the superscripts (μ) label the columns, and the subscripts (ν) label the rows.] One checks these matrices against the required property: $R^{\lambda}_{\mu}(q)\overline{R^{\nu}_{\lambda}}(q)=L^{\lambda}_{\mu}(q)\overline{L^{\nu}_{\lambda}}(q)=\rho^2\delta^{\nu}_{\mu}$, for all $q\in\mathcal{E}_4$. Also $R(q)=\det[R^{\mu}_{\nu}(q)]=L(q)=\det[L^{\mu}_{\nu}(q)]=(q^{\mu}q^{\mu})^2=\rho^4$ can be shown rather easily. Moreover, a detailed analysis shows that, if $q\in\mathcal{S}_3$, these matrices transport vectors which are tangent to $\mathcal{S}_3$ into tangent vectors to $\mathcal{S}_3$. Therefore, fixing $q^{\mu}q^{\mu}=1$, Eqs. (A4) and (A5) yield precisely the transport matrices of SU(2) on $\mathcal{S}_3$.

The transport matrices are affine connections on the group manifold. Furthermore, for the noncompact Lie group $\Sigma(4)$ it follows that the mixed transport matrix given by $\rho^{-2}R^{\lambda}_{\nu}(q)L^{\mu}_{\lambda}(q)$ carries the *adjoint representation* of the group.[14] Hence, in the present instance, using Eqs. (A4.1) and (A5.2), for the action of SU(2) on $\mathcal{S}_3$ we get

$$A^{\mu}_{\nu}(q)=\begin{bmatrix} A_{jk}(q) & 0 \\ 0 & 1 \end{bmatrix}. \tag{A6}$$

The 3×3 matrix $A_{jk}(q)$ in Eq. (A6), has entries given by

$$A_{jk}(q)=q^{\mu}q_{\mu}\delta_{jk}+2q^{j}q^{k}+2q^{4}\epsilon_{jk\ell}q^{\ell}, \tag{A7}$$

and carries the adjoint representation of SU(2). In fact, one easily finds

$$u^{\dagger}(q)\sigma_{j}u(q)=A_{jk}(q)\sigma_{k}, \tag{A8}$$

and therefore the A's satisfy the group law $A_{j\ell}(q')A_{\ell k}(q)=A_{jk}[g(q';q)]$. We can further analyze this construct recalling that proper rotations in ordinary space $\mathcal{E}_3$ are characterized by an angle of rotation $\vartheta(0\leqslant\vartheta\leqslant 2\pi)$ about an axis of rotation $\hat{\mathbf{n}}(\hat{\mathbf{n}}\cdot\hat{\mathbf{n}}=1)$, and are represented by the group SO(3) of real orthogonal matrices of the form:

$$R_{jk}(\vartheta,\hat{\mathbf{n}})=\delta_{jk}\cos\vartheta+\hat{n}^{j}\hat{n}^{k}(1-\cos\vartheta)+\epsilon_{jk\ell}\hat{n}^{\ell}\sin\vartheta. \tag{A9}$$

One identifies $2\phi=\vartheta$ and one obtains: $A_{jk}(q)=R_{jk}(\vartheta,\hat{\mathbf{n}})\in \mathrm{SO}(3)$. Also, defining $\mathbf{x}=x^{k}\sigma_{k}$, the unitary transformation $\mathbf{x}'=u^{\dagger}(q)\mathbf{x}u(q)$ yields $x'^{j}=x^{k}R_{jk}(\vartheta,\hat{\mathbf{n}})$. Hence, the double covering of SU(2) onto SO(3) becomes manifest. (These things are, of course, well known.[13])

We next define right and left vector fields for $\Sigma(4)$ in $\mathcal{E}_4$ by means of the transport matrices, as usual: $X_{\mu}(q)=R^{\nu}_{\mu}(q)\partial_{\nu}=\sigma^{\nu}_{\mu\lambda}q^{\lambda}\partial_{\nu}$, and $Y_{\mu}(q)=L^{\nu}_{\mu}(q)\partial_{\nu}=\sigma^{\nu}_{\lambda\mu}q^{\lambda}\partial_{\nu}$. Writing them more explicitly, one sees that these operators correspond to SU(2) right and left generators acting on $\mathcal{S}_3$, which in the Euler–Rodrigues parametrization read:

$$X_{j}(q)=q^{4}\partial_{j}-q^{j}\partial_{4}-\epsilon_{jk\ell}q^{k}\partial_{\ell},\quad Y_{j}(q)=q^{4}\partial_{j}-q^{j}\partial_{4}+\epsilon_{jk\ell}q^{k}\partial_{\ell}, \tag{A10}$$

respectively, as well as to the dilation generator of $\Sigma(4)$ acting along the radius ρ in $\mathcal{E}_4$, which is given by $X_4(q)=Y_4(q)=q^{\mu}\partial_{\mu}=\rho(\partial/\partial\rho)$.

Using (θ,φ,ϕ) as the independent variables of the SU(2) theory [where $\hat{\mathbf{n}}$ is a unit 3-vector corresponding to $\hat{\mathbf{n}}=(\theta,\varphi)$], given any function $\psi(q)$ defined on $\mathcal{S}_3$, the displacement $q\rightarrow(q+dq)\in\mathcal{S}_3$ produces the differential

$$d\psi(q)=\frac{\partial\psi}{\partial q^{j}}\sin\phi d\hat{n}^{j}+\left(\frac{\partial\psi}{\partial q^{j}}\hat{n}^{j}\cos\phi-\frac{\partial\psi}{\partial q^{4}}\sin\phi\right)d\phi, \tag{A11}$$

with $d\hat{n}^{j}=(\partial\hat{n}^{j}/\partial\theta)d\theta+(\partial\hat{n}^{j}/\partial\varphi)d\varphi$. On the other hand, defining $\delta q^{\mu}_{R}=\overline{R^{\mu}_{\nu}}(q)dq^{\nu}$, a straightforward calculation shows that $d\psi(q)=\delta q^{j}_{R}X_{j}(q)\psi(q)$ is given by

$$d\psi(q)=\delta\phi\hat{n}^{j}_{R}\left(\cos\phi\frac{\partial\psi}{\partial q^{j}}-\sin\phi\hat{n}^{j}\frac{\partial\psi}{\partial q^{4}}-\sin\phi\epsilon_{jk\ell}\hat{n}^{k}\frac{\partial\psi}{\partial q^{\ell}}\right). \tag{A12}$$

Hence, consistency demands: $d\phi\hat{n}^{j}=\delta\phi\hat{n}^{j}_{R}$ and $d\hat{n}^{j}=\delta\phi\epsilon_{jk\ell}\hat{n}^{k}\hat{n}^{\ell}_{R}$ [which follow from the transport formula $dq^{\mu}=R^{\mu}_{\nu}(q)\delta q^{\nu}_{R}$], as the reader can prove in a few lines. Furthermore, one also proves that $X_4(q)\psi(q)=0$ follows if $q\in\mathcal{S}_3$. (We left this as an exercise for the reader.) In this fashion we conclude that the action of the four operators [either $X_{\mu}(q)$ or $Y_{\mu}(q)$], on functions $\psi(q)$ defined on the three-dimensional surface $\mathcal{S}_3$, is consistent with the three degrees of freedom of the group SU(2).

As operators in $\mathcal{E}_4$, the generators of $\Sigma(4)$ satisfy the following Lie algebra:

$$[X_{j}(q),X_{k}(q)]=2\epsilon_{jk\ell}X_{\ell}(q),\quad [Y_{j}(q),Y_{k}(q)]=-2\epsilon_{jk\ell}Y_{\ell}(q), \tag{A13}$$

$$[X_j(q),X_4(q)]=0, \quad [Y_j(q),Y_4(q)]=0, \tag{A14}$$

and moreover

$$[X_\mu(q),Y_\nu(q)]=0. \tag{A15}$$

Equation (A13) corresponds to the (right and left) realizations of the SU(2) Lie algebra, whereas Eq. (A14) assures a consistent result (for our purposes here). Equation (A15) is of fundamental importance in quantum kinematics [i.e., see Eq. (3.6)].

Finally, since $\Sigma(4)$ is a noncompact Lie group, an invariant measure in $\mathcal{E}_4$ follows immediately:

$$d\mu(q)=\mu_0\rho^{-4}\, dq^1\, dq^2\, dq^3\, dq^4=\mu_0\,\frac{\sin\,\theta\,\sin^2\,\phi}{\rho}\, d\rho\, d\theta\, d\varphi\, d\phi=\mu_0\,\frac{d\rho}{\rho}\, d\Omega(\phi,\hat{\mathbf{n}}), \tag{A16}$$

where μ_0 denotes an arbitrary normalization constant and $d\Omega(\phi,\hat{\mathbf{n}})$ is the element of solid angle in $\mathcal{E}_4$. The direct-product structure of the quaternion group $\Sigma(4)\sim\mathcal{R}\times\otimes\mathrm{SU}(2)$ means that (if we fix $\rho=1$) we obtain an invariant measure on $\mathcal{S}_3$ as well, which is given by

$$d\mu(q)=\mu_0\,\sin\,\theta\,\sin^2\,\phi\, d\theta\, d\varphi\, d\phi=\mu_0\, d\Omega(\phi,\hat{\mathbf{n}}). \tag{A17}$$

This well-known measure is indeed (right and left) SU(2)-invariant.

APPENDIX B: SU(2) REGULAR REPRESENTATION REVISITED

We here append a brief description of the regular representation of SU(2) as it follows from the embedding scheme $\mathcal{S}_3\subset\mathcal{E}_4$ presented in Appendix A. Henceforth we use a shorthand notation for handling the Euler–Rodrigues parameters $q=(\mathbf{q},q^4)$ in terms of the angle parameters $(\phi,\hat{\mathbf{n}})$. Thus we set: $q=(\hat{\mathbf{n}}\,\sin\,\phi,\cos\,\phi)\equiv(\phi,\hat{\mathbf{n}})\in\mathcal{S}_3$.

As one knows, the regular representation[15] of SU(2) is carried by functions $\psi(q)=\psi(\phi,\hat{\mathbf{n}})$, which have a finite norm with respect to the invariant measure on $\mathcal{S}_3$ defined in Eq. (A17):

$$\langle\psi|\psi\rangle=\mu_0\oint_{\mathcal{S}_3} d\Omega(\phi,\hat{\mathbf{n}})|\psi(\phi,\hat{\mathbf{n}})|^2<\infty. \tag{B1}$$

To proceed, one introduces a Hilbert space $\mathcal{H}[\mathrm{SU}(2)]$ such that $|\psi\rangle\in\mathcal{H}[\mathrm{SU}(2)]\Leftrightarrow\langle\psi|\psi\rangle<\infty$. In order to have a one-to-one mapping between wave functions $\psi(q)$ and vectors $|\psi\rangle$, one considers the associated rigged Hilbert space $\tilde{\mathcal{H}}[\mathrm{SU}(2)]$, which is spanned by continuous basic vectors $|q\rangle$ satisfying the one-to-one correspondence $\mathcal{S}_3\ni q\leftrightarrow|q\rangle\in\tilde{\mathcal{H}}[\mathrm{SU}(2)]$. The basic vectors obey the orthogonal relation:

$$\langle\phi',\hat{\mathbf{n}}'|\phi,\hat{\mathbf{n}}\rangle=\mu_0^{-1}\,\frac{\delta(\phi'-\phi)\,\delta(\theta'-\theta)\,\delta(\varphi'-\varphi)}{\sin^2\,\phi\,\sin\,\theta}, \tag{B2}$$

and yield the following resolution of the identity:

$$\mu_0\oint_{\mathcal{S}_3} d\Omega(\phi,\hat{\mathbf{n}})|\phi,\hat{\mathbf{n}}\rangle\langle\phi,\hat{\mathbf{n}}|=I. \tag{B3}$$

Wave functions are then defined on the group manifold in the usual manner; namely, one writes $\psi(q)=\langle q|\psi\rangle\equiv\psi(\phi,\hat{\mathbf{n}})=\langle\phi,\hat{\mathbf{n}}|\psi\rangle$, whenever $|\psi\rangle\in\mathcal{H}[\mathrm{SU}(2)]$ and $q=(\phi,\hat{\mathbf{n}})\in\mathcal{S}_3$.

Using this notation, the quaternion group law (A2) can be cast in terms of the angle variables $(\phi,\hat{\mathbf{n}})$. In fact, (A2) corresponds to the well-known combination formulas:[13]

$$\hat{\mathbf{n}}''\,\sin\,\phi''=\hat{\mathbf{n}}'\,\sin\,\phi'\,\cos\,\phi+\hat{\mathbf{n}}\,\sin\,\phi\,\cos\,\phi'-\hat{\mathbf{n}}\times\hat{\mathbf{n}}\,\sin\,\phi'\,\sin\,\phi, \tag{B4a}$$

$$\cos\,\phi''=\cos\,\phi'\,\cos\,\phi-\hat{\mathbf{n}}\cdot\hat{\mathbf{n}}\,\sin\,\phi'\,\sin\,\phi. \tag{B4b}$$

At times it is useful to write this law in a compact symbolic way:

$$(\phi'',\hat{\mathbf{n}}'')=(\phi',\hat{\mathbf{n}}')\circ(\phi,\hat{\mathbf{n}}). \tag{B5}$$

For instance, if we consider a point $q=(\phi,\hat{\mathbf{n}})\in\mathcal{S}_3$ which is left-transported by means of $e+\delta q'=(\delta\phi,\hat{\mathbf{n}}')\in\mathcal{S}_3$, Eq. (B4) allows us to express Lie's left-differentials as follows:

$$\delta_L(\hat{\mathbf{n}}\sin\phi)=\delta\phi(\hat{\mathbf{n}}'\cos\phi-\hat{\mathbf{n}}'\times\hat{\mathbf{n}}\sin\phi),\quad \delta_L(\cos\phi)=-\delta\phi\hat{\mathbf{n}}'\cdot\hat{\mathbf{n}}\sin\phi, \tag{B6}$$

where $\hat{\mathbf{n}}''\sin\phi''=\hat{\mathbf{n}}\sin\phi+\delta_L(\hat{\mathbf{n}}\sin\phi)$ and $\cos\phi''=\cos\phi+\delta_L(\cos\phi)$ give us the angle coordinates of the image point, i.e., $(\phi'',\hat{\mathbf{n}})=(\delta\phi,\hat{\mathbf{n}}')\circ(\phi,\hat{\mathbf{n}})$.

A sensible definition of the unitary group-operators $U(q)\equiv U(\phi,\hat{\mathbf{n}})$, of the left regular representation, can then be given in the following manner:

$$U(\phi,\hat{\mathbf{n}})=e^{-(i/\hbar)\phi\hat{\mathbf{n}}\cdot\mathbf{L}}=\mu_0\int_0^{\pi}\mathbf{d}\phi'\ \sin^2\phi'\oint_{\mathcal{S}_2}d\Omega(\hat{\mathbf{n}}')|(\phi,\hat{\mathbf{n}})\circ(\phi',\hat{\mathbf{n}}')\rangle\langle\phi',\hat{\mathbf{n}}'|. \tag{B7}$$

$\mathbf{L}$ denotes the generators of this representation: $\mathbf{L}=(L_1,L_2,L_3)$. In fact, a straightforward calculation yields the unitary representation property obeyed by these operators:

$$U(\phi',\hat{\mathbf{n}}')U(\phi,\hat{\mathbf{n}})=U[(\phi',\hat{\mathbf{n}}')\circ(\phi,\hat{\mathbf{n}})],\quad U^{\dagger}(\phi,\hat{\mathbf{n}})=U^{-1}(\phi,\hat{\mathbf{n}}), \tag{B8}$$

as well as

$$U(\phi',\hat{\mathbf{n}}')|\phi,\hat{\mathbf{n}}\rangle=|(\phi',\hat{\mathbf{n}}')\circ(\phi,\hat{\mathbf{n}})\rangle, \tag{B9}$$

As we see, the operators (B7) realize the left regular representation of SU(2) on $\mathcal{S}_3$. The infinitesimal operators are given by $U(\delta\phi,\hat{\mathbf{n}})=I-(i/\hbar)\delta\phi\hat{\mathbf{n}}\cdot\mathbf{L}$. Therefore, using Eq. (A10), we obtain

$$L_j|\phi,\hat{\mathbf{n}}\rangle=i\hbar X_j(q)|q\rangle=i\hbar(\cos\phi\partial_j-\sin\phi\hat{n}^j\partial_4-\sin\phi\epsilon_{jk\ell}\hat{n}^k\partial_{\ell})|\phi,\hat{\mathbf{n}}\rangle, \tag{B10}$$

for $q=(\phi,\hat{\mathbf{n}})\in\mathcal{S}_3$, as consistency demands. This expression yields the familiar Lie Algebra in terms of the left-regular representation's generators [cf., Eq. (A13)]:

$$[L_j,L_k]=-2i\hbar\epsilon_{jk\ell}L_{\ell}. \tag{B11}$$

[Notice that the generators L_j equal twice the standard physical operators of angular momentum J_j. See, for instance, Eq. (2.17). The reader can change to the conventional scale if he likes.]

Within the left regular representation, the adjoint representation is obtained from $U^{\dagger}(\phi,\hat{\mathbf{n}})U(\delta\phi,\hat{\mathbf{n}}')U(\phi,\hat{\mathbf{n}})$. This yields [see Eq. (A7)]

$$U^{\dagger}(\phi,\hat{\mathbf{n}})L_jU(\phi,\hat{\mathbf{n}})=[(\cos^2\phi-\sin^2\phi)\delta_{jk}+2\sin^2\phi\hat{n}^j\hat{n}^k+2\cos\phi\sin\phi\epsilon_{jk\ell}\hat{n}^{\ell}]L_k \tag{B12}$$

[which leads also to the Lie Algebra (B11)].

To end up this brief revision, let us finally observe that $X_4(q)\psi^*(q)=\langle\psi|X_4(q)|q\rangle=0$ holds for all $|\psi>\in\mathcal{H}[\mathrm{SU}(2)]$ and $q\in\mathcal{S}_3$. So we have

$$q\in\mathcal{S}_3\Rightarrow X_4(q)|q\rangle=0. \tag{B13}$$

In fact, the dilation operator $X_4(q)=Y_4(q)$ is the sole generator of the embedding quaternion group $\Sigma(4)$ not represented in SU(2) quantum kinematics.

Table I summarizes these features. One introduces the right regular representation of SU(2) in a similar manner, defining appropriate group-operators $V(q)$, which act ''from the right'' in the following sense $V(q')|q\rangle=|g(q;q')\rangle$. For the sake of completeness, Table II presents a synoptic setting of the right regular representation. Let us here recall that in the left quantum kinematic formalism, as considered in this article, the right generators R_j presented in Table II play an important role, because they correspond to the basic invariant operators of the theory (cf., Sec. III).

[1] K. Yamada, Phys. Rev. D **25**, 3256 (1982).
[2] M. M. Nieto, Phys. Rev. Lett. **18**, 182 (1967); P. Carruthers and M. M. Nieto, Rev. Mod. Phys. **40**, 411 (1968); H. Bacry, Phys. Rev. A **18**, 617 (1978); see, also, M. M. Nieto, L. M. Simmons, and V. P. Gutschick, Phys. Rev. D **23**, 927 (1981).
[3] Cf., A. Perelomov, *Generalized Coherent States and Their Applications* (Springer, Berlin, 1986). It seems that the motivation behind such studies on generalized minimum uncertainty states has changed after the publication of this timely monograph.
[4] J. Krause, Int. J. Theor. Phys. **33**, 1617 (1994).
[5] Quantum kinematics of non-Abelian variables is also an indispensable device in the formulation of lattice gauge theories which circumvent conventional perturbation theory [K. G. Wilson, Phys. Rev. D **10**, 2445 (1974)] in the search for the appropriate description of quark confinement [J. Kogut and L. Susskind, *ibid.* **11**, 395 (1975); M. Creutz, *ibid.* **15**, 1128 (1977); J. Kogut, Phys. Rep. **67**, 67 (1980)]. See, i.e., M. Kaku, *Quantum Field Theory* (Oxford University Press, Oxford, 1993).
[6] See, e.g., J. Krause, J. Phys. A **18**, 1309 (1985).
[7] H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1931).
[8] J. Krause, Int. J. Theor. Phys. **36**, 847 (1997).
[9] As developed, for instance, in L. Fonda and G. C. Ghirardi, *Symmetry Principles in Quantum Physics* (Dekker, New York, 1970).
[10] J. Krause, Phys. Rev. A **54**, 4691 (1996).
[11] J. Krause, Int. J. Theor. Phys. **37**, 759 (1998).
[12] E. P. Wigner, *Group Theory* (Academic, New York, 1959).
[13] See, for instance, L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison–Wesley, Reading, MA, 1981). If there is a single essential book in the arsenal of a physicist, it is a good book on the theory of angular momentum. The present text (written by two well-known contributors to the field) satisfies all the most exigent criteria to this end. This is the main reference on SU(2) theory used in this paper.
[14] J. Krause, J. Math. Phys. **32**, 348 (1991).
[15] M. A. Naimark and A. I. Stern, *Theory of Group Representations* (Springer, New York, 1982).
[16] Generalized enveloping algebras and coherent states have been considered recently, in terms of the quantum kinematic invariant operators of Lie groups; see J. Krause, Int. J. Theor. Phys. **32**, 1363 (1993). See also, J. Krause, J. Phys. A **26**, 6285 (1993).
[17] J. Krause, J. Math. Phys. **27**, 2922 (1986).
[18] J. Krause, J. Math. Phys. **29**, 393 (1988).
[19] J. Krause, J. Math. Phys. **35**, 5734 (1994).
[20] S. M. Barnett and D. T. Pegg, Phys. Rev. A **41**, 3427 (1990).
[21] It is a well-known property of compact Lie groups that the regular representation contains *all* the irreducible unitary representations. See Ref. 15.

Unified treatment of the Coulomb and harmonic oscillator potentials in *D* dimensions

G. Lévai,[a] B. Kónya, and Z. Papp[b]
Institute of Nuclear Research of the Hungarian Academy of Sciences, P. O. Box 51, H–4001 Debrecen, Hungary

(Received 9 February 1998; accepted for publication 9 June 1998)

Quantum mechanical models and practical calculations often rely on some exactly solvable models like the Coulomb and the harmonic oscillator potentials. The *D* dimensional generalized Coulomb potential contains these potentials as limiting cases, thus it establishes a continuous link between the Coulomb and harmonic oscillator potentials in various dimensions. We present results which are necessary for the utilization of this potential as a model and practical reference problem for quantum mechanical calculations. We define a Hilbert space basis, the generalized Coulomb–Sturmian basis, and calculate the Green's operator on this basis and also present an SU(1,1) algebra associated with it. We formulate the problem for the one-dimensional case, too, and point out that the complications arising due to the singularity of the one-dimensional Coulomb problem can be avoided with the use of the generalized Coulomb potential. © *1998 American Institute of Physics.* [S0022-2488(98)02311-1]

I. INTRODUCTION

Local potentials have been used to model the interactions of the subatomic world ever since the introduction of quantum mechanics. Some of these (like the Coulomb potential) do not differ essentially from the forces observed in nature, while most of them (like the harmonic oscillator, for example) represent approximations of the actual physical situation. The potential shape, defined by the potential type and the parameters in it, is usually chosen in a way that reflects the physical picture our intuition associates with the problem; therefore we can define attractive or repulsive, short-range or long-range potentials, etc.

Some of the potentials used in quantum mechanics are exactly solvable. This means that the energy eigenvalues, the bound-state wave functions and the scattering matrix can be determined in closed analytical form. The range of these potentials has been extended considerably in the recent years by investigations inspired by supersymmetric quantum mechanics (SUSYQM), for example. (See Ref. 1 for recent reviews.) Most of the solvable potentials are one-dimensional problems, which means that they are defined on the $(-\infty,\infty)$ domain (or on some finite interval of it), or they can be reduced to radial problems in higher spatial dimensions, in which case they are confined to the positive real axis. The simplest exactly solvable potentials are those belonging to the so-called shape-invariant class.[2] Many of the most well-known potentials, like the Coulomb, harmonic oscillator, Morse, Pöschl–Teller, etc. potentials, belong to this class. Usually 12 potentials are mentioned to have the property of shape-invariance,[3] although two of them are just alternative forms of other shape-invariant potentials.

A rather more general potential type is that of the Natanzon potentials,[4] the solutions of which can be reduced to a single hypergeometric or confluent hypergeometric function. These potentials are written as complicated six-parameter expressions of the coordinate, and their energy eigenvalues are given by an implicit formula. Although these problems are solvable exactly in principle, only some specific Natanzon potentials have been studied in detail,[5–8] because the formalism is getting more and more complicated with the increasing number of parameters handled. Some of the Natanzon potentials contain certain shape-invariant potentials as special cases: the Ginocchio potential[6] can be considered as a generalization of the Pöschl–Teller potential, while the gener-

[a]Electronic mail: levai@atomki.hu
[b]Electronic mail: pz@indigo.atomki.hu

 © 1998 American Institute of Physics

alized Coulomb potential[8] is a simultaneous extension of the Coulomb and the harmonic oscillator potentials. There are also further potentials beyond the Natanzon class. These include, for example, quasi-exactly solvable (QES) potentials,[9] conditionally exactly solvable (CES) potentials[10] and the SUSYQM partners of Natanzon-class potentials,[11] of which certain CES potentials form a subclass.[12]

The study of exactly solvable potentials represents an interesting field of mathematical physics in itself, but results from this area are also essential for the description of realistic physical problems. More and more interactions can be modeled by taking advantage of the rather flexible potential shapes offered by exactly solvable potentials. The solutions of these can be applied directly, or they can be combined with numerical calculations. In the simplest case analytical calculations can aid numerical studies in areas where numerical techniques might not be safely controlled. This is the case, for example, when bound-state wave functions with arbitrary node numbers are required, for certain singular potentials, or for complex potentials.[13] As the next level of complexity, analytical solutions can supply a basis for numerical calculations.

Virtually all quantum mechanical methods rely in some respect on analytically solvable potentials. Very often their wave function solutions are used as a Hilbert-space basis. More powerful methods can be constructed if we select a basis which allows the exact analytical calculation of the Green's operator of an analytically solvable potential. Some time ago, by utilizing the Coulomb–Sturmian functions as basis functions, the Coulomb Green's operator was calculated analytically and a quantum mechanical approximation method was developed.[14] Recently this method was extended for approaching the really challenging three-body Coulomb problem.[15] We note here that in this work the exact analytical representation of the Coulomb Green's operator on the Coulomb–Sturmian basis was essential. In our recent work[16] we showed that if on some basis representation the Hamiltonian possesses an infinite tridiagonal matrix (J-matrix) form, then its resolvent can be given in terms of continued fractions which can be continued analytically to the whole complex plane. This theorem was exemplified by the Coulomb and harmonic oscillator Green's operators.

The aim of the present work is to demonstrate that various techniques developed for the simple shape-invariant potentials, like the Coulomb and harmonic oscillator potentials, can be generalized for certain Natanzon-class potentials, too, without making the formalism significantly more involved, and that these more general potentials can help our understanding of both practical and theoretical problems. The actual example we consider here is the generalized Coulomb potential,[8] which is the member of the Natanzon confluent potential class.[17] This potential is Coulomb-like asymptotically, while its short-range behavior depends on the parameters: it can be finite or singular as well at the origin. Its shape therefore can approximate various realistic problems, such as nuclear potentials with relatively flat central part, or atomic potentials that incorporate the effect of inner closed shells by a phenomenological repulsive core.

An advantageous feature of the generalized Coulomb problem is that the equivalent of the Coulomb–Sturmian basis can be defined for it, together with all its implications. For example, the method of calculating the matrix elements of the Green's operator as described in Ref. 16 for the Coulomb and the harmonic oscillator potentials can be extended to this case, too, so that the above two problems appear in the present treatment as special cases.

Using a particular differential realization, an SU(1,1) algebra can also be associated with this potential, which reduces to known SU(1,1) algebras in the Coulomb and harmonic oscillator limits.

Another important feature of this potential is that it can be formulated as a one-dimensional problem, but it can avoid the pathologic features of the one-dimensional Coulomb problem, which arise due to the $|x|^{-1}$-like singularity. Since this potential can be chosen to be finite at $x=0$, it offers a convenient basis for calculations which require the use of a one-dimensional Coulomb-like problem.

The arrangement of this paper is the following. In Sec. II we give the basic formalism for the generalized Coulomb problem by recapitulating and extending former results. Section III contains the main results of the paper: *(i)* the discussion of the Coulomb–oscillator connection using the two limiting cases of the potential, *(ii)* the determination of the matrix elements of the Green's operator, *(iii)* the description of an SU(1,1) algebra associated with the problem and *(iv)* discussion on the one-dimensional case, which elucidates the complications related to the singularity of the

one-dimensional Coulomb problem. Finally, we summarize the results and discuss further possibilities in Sec. IV.

II. THE GENERALIZED COULOMB POTENTIAL IN *D* DIMENSIONS

Let us consider the Schrödinger equation in D spatial dimensions with a potential that depends only on the radial variable r. Separating the angular variables from the wave function we obtain the radial Schrödinger equation

$$\hat{H}_0\psi(r)\equiv\left(-\frac{d^2}{dr^2}+\frac{1}{r^2}\left(l+\frac{D-3}{2}\right)\left(l+\frac{D-1}{2}\right)+v(r)\right)\psi(r)=\epsilon\psi(r), \tag{1}$$

where $v(r)\equiv 2m\hbar^{-2}V(r)$ and $\epsilon\equiv 2m\hbar^{-2}E$. The centrifugal term, which depends on the angular momentum l, originates from the kinetic term, i.e., from the D-dimensional Laplace operator after separating the angular variables. The bound-state wave functions solving (1) are normalized as

$$\int_0^\infty |\psi(r)|^2\,dr=1. \tag{2}$$

Here we recapitulate the results obtained for the generalized Coulomb potential in Ref. 8 by adapting it to the D-dimensional case. Changing slightly the notation used in Ref. 8 we define the generalized Coulomb potential as

$$v(r)=-\frac{1}{r^2}\left(l+\frac{D-3}{2}\right)\left(l+\frac{D-1}{2}\right)+\left(\beta-\frac{1}{2}\right)\left(\beta-\frac{3}{2}\right)\frac{C}{4h(r)(h(r)+\theta)}$$
$$-\frac{q}{h(r)+\theta}-\frac{3C}{16(h(r)+\theta)^2}+\frac{5C\theta}{16(h(r)+\theta)^3}, \tag{3}$$

where $h(r)$ is defined in terms of its inverse function

$$r=r(h)=C^{-1/2}\left[\theta\tanh^{-1}\left(\left(\frac{h}{h+\theta}\right)^{1/2}\right)+(h(h+\theta))^{1/2}\right]. \tag{4}$$

The $h(r)$ function maps the $[0,\infty)$ half-axis onto itself and can be approximated with $h(r)\simeq C^{1/2}r$ and $h(r)\simeq Cr^2/(4\theta)$ in the $r\to\infty$ and $r\to 0$ limits, respectively.

Bound states are located at

$$\epsilon_n=-\frac{C}{4}\rho_n^2, \tag{5}$$

where

$$\rho_n=\frac{2}{\theta}\left[\left((n+\beta/2)^2+\frac{q\theta}{C}\right)^{1/2}-(n+\beta/2)\right]. \tag{6}$$

The bound-state wave functions can be written in terms of associated Laguerre polynomials as

$$\psi_n(r)=C^{1/4}\rho_n^{(\beta+1)/2}\left(\frac{\Gamma(n+1)}{\Gamma(n+\beta)(2n+\beta+\rho_n\theta)}\right)^{1/2}$$
$$\times(h(r)+\theta)^{1/4}(h(r))^{(2\beta-1)/4}\exp\left(-\frac{\rho_n}{2}h(r)\right)L_n^{(\beta-1)}(\rho_n h(r)). \tag{7}$$

Potential (3) clearly carries angular momentum dependence: its first term merely compensates the centrifugal term arising from the kinetic term of the Hamiltonian. Its second term also has r^{-2}-like singularity (due to $h^{-1}(r)$), and, as it will be demonstrated later, it cancels the angular

momentum dependent term in the two important limiting cases that recover the D-dimensional Coulomb and the harmonic oscillator potentials. The third term of (3) represents an asymptotically Coulomb-like interaction, while the remaining two terms behave like r^{-2} and r^{-3} for large values of r.

In the remaining part of this section we consider the specific case of $D=3$ and $l=0$. The S matrix of the generalized Coulomb potential can be derived in complete analogy with that of the Coulomb problem for $D=3$. Although this can only be done exactly for $l=0$, the singular term imitating the centrifugal term in Eq. (3) can be defined to be part of the potential. Using the notation of Ref. 18 the equivalents of the general regular and irregular solutions are

$$\varphi^{(G)}(k,r)=C^{-\beta/4}(h(r)+\theta)^{1/4}(h(r))^{(2\beta-1)/4}\exp\left(-\frac{\rho}{2}h(r)\right)\Phi\left(\frac{\beta}{2}+i\nu,\beta;\rho h(r)\right) \tag{8}$$

and

$$f^{(G)}(k,r)=\rho^{-\beta/4}e^{\nu\pi/2}(h(r)+\theta)^{1/4}(h(r))^{(2\beta-1)/4}\exp\left(-\frac{\rho}{2}h(r)\right)\Psi\left(\frac{\beta}{2}+\mathrm{i}\nu,\beta;\rho h(r)\right), \tag{9}$$

where $\Phi(a,c;z)$ and $\Psi(a,c;z)$ are the regular and the irregular confluent hypergeometric functions,[19] and ρ and ν have to be chosen as

$$\rho\equiv\rho(k)=-\frac{2ik}{C^{1/2}},\quad i\nu\equiv i\nu(k)=\frac{\rho\theta}{4}-\frac{q}{C\rho}. \tag{10}$$

The functions (8) and (9) reduce to $\varphi^{(c)}(k,r)$ and $f^{(c)}(k,r)$ of Ref. 18 in the $\theta\to 0$ Coulomb limit. The S matrix for $l=0$ is expressed as

$$S_0(k)=(-1)^{\beta/2+1}\frac{\Gamma(\beta/2+i\nu)}{\Gamma(\beta/2-i\nu)}. \tag{11}$$

The extra phase factor appears because of the r^{-2}-type singular term, which is now defined to be part of the potential. In the Coulomb limit this expression becomes part of the centrifugal term, which is dealt with separately.

The long-range behavior of potential (3) suggests its use in problems associated with the electrostatic field of some charge distribution. The deviation from the Coulomb potential close to the origin can be viewed as replacing the pointlike charge with an extended charged object. The relevant charge density is readily obtained from the potential using

$$\rho(r)=-\frac{\hbar^2}{8\pi me}\Delta v(r). \tag{12}$$

Choosing $D=3$ and neglecting the terms singular at the origin by making the choice $l=0$ and $\beta=1/2$ or $3/2$ we get

$$\rho(r)=-\frac{\hbar^2}{8\pi me}\left[-\frac{2qC}{(h(r)+\theta)^3}+\frac{20qC\theta-9C^2}{8(h(r)+\theta)^4}+\frac{81C^2\theta}{16(h(r)+\theta)^5}-\frac{135C^2\theta^2}{32(h(r)+\theta)^6}\right.$$
$$\left.+\frac{2C^{1/2}}{r}\left(\frac{h(r)}{h(r)+\theta}\right)^{1/2}\left(\frac{q}{(h(r)+\theta)^2}+\frac{3C}{8(h(r)+\theta)^3}-\frac{15C\theta}{16(h(r)+\theta)^4}\right)\right]. \tag{13}$$

In Figs. 1 and 2 we present examples for the actual shape of potential (3) and the corresponding charge distribution (13) for various values of the parameters. It can be seen that this potential is suitable for describing the Coulomb field of extended objects. It is a general feature of potential (3) that for small values of θ a (finite) positive peak appears near the origin, which also manifests itself in a repulsive ''hard core,'' corresponding to a region with positive charge density (see Fig. 2).

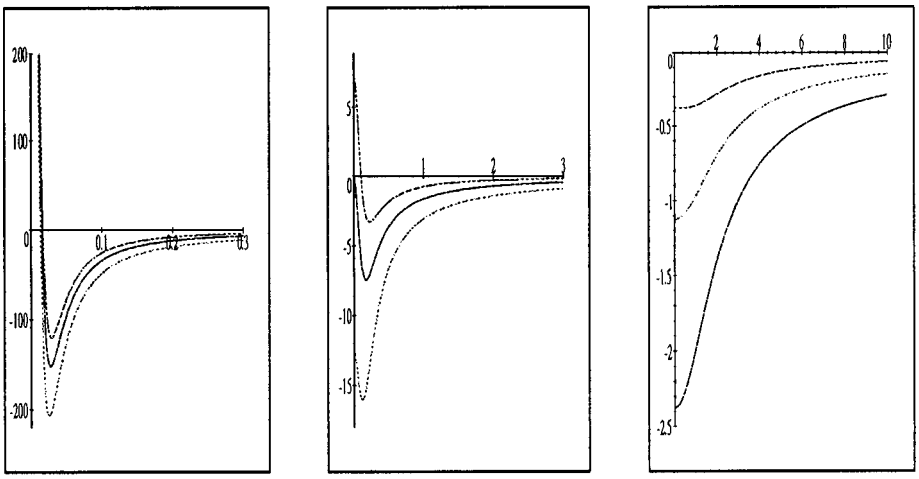

FIG. 1. The generalized Coulomb potential for $q=0.5$, 1.25, 2.5; $\theta=0.01$, 0.1, 1; $C=1$ and $\beta=3/2$. $l=0$ and $D=3$ is also implied. In each panel the largest q corresponds to the lowest curve. Note the different scales of the horizontal (r) and the vertical $(v(r))$ axes.

III. DISCUSSION

A. The Coulomb and harmonic oscillator limits

As described in Ref. 8, the introduction of the generalized Coulomb problem was inspired by a simple method of transforming the Schrödinger equation into the differential equation of the confluent hypergeometric function. The bound-state solutions of these problems contain associated Laguerre polynomials. The three most well-known examples of this type are the Coulomb, the harmonic oscillator and the Morse potentials, which are all members of the shape-invariant potential class.[2] These three potentials are recovered by specific choices of the variable transformation reducing the Schrödinger equation to the differential equation of the associated Laguerre polynomials. This is also reflected by the structure of the function $h(r)$ in Eq. (4): when $h(r)$ is proportional to r and r^2, one obtains the Coulomb problem and the harmonic oscillator potential, respectively. These limits can readily be realized by specific choices of the parameters in Eq. (4): the first one follows from the $\theta\to 0$ limit,[8] while the second one is reached by taking $\theta\to\infty$, while keeping $C/\theta\equiv\tilde{C}$ constant.[20]

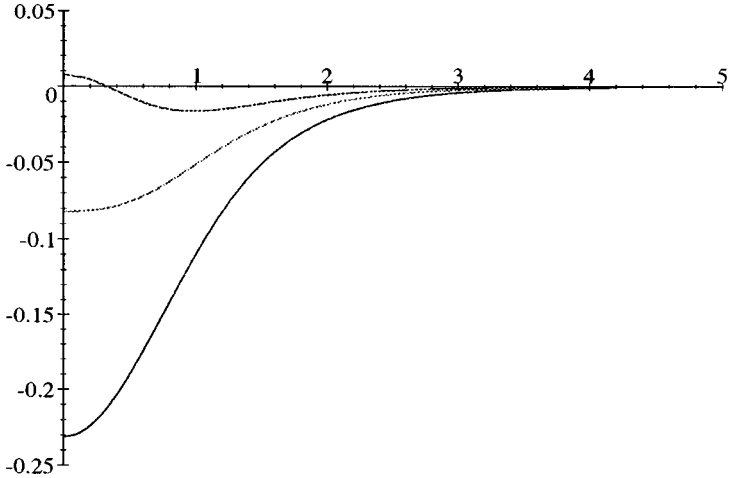

FIG. 2. The charge distributions corresponding to the potentials displayed in Fig. 1 for $\theta=1$ and $q=0.5$, 1.25, 2.5. (The largest q corresponds to the lowest curve.)

Besides taking the $\theta\to 0$ limit, the Coulomb problem in D-dimensions is recovered from Eq. (3) by the $\beta=2l+D-1$ and $C^{-1/2}q=2mZe^2/\hbar^2$ choices: the third term of (3) becomes the Coulomb term, the fifth one vanishes, while the other three all become proportional with r^{-2} and cancel out completely.

In order to reach the oscillator limit one also has to redefine the potential (3) and the energy eigenvalues by adding q/θ to both. (This is equivalent to $\tilde{C}\tilde{D}\theta$ in the notation of Ref. 20.) This choice simply represents resetting the energy scale: $\epsilon=0$ corresponds to $v(r\to\infty)$ for the Coulomb problem, and to $v(r=0)$ for the harmonic oscillator. (Note that the energy eigenvalues also have different signs in the two cases.) Besides $C/\theta=\tilde{C}$, the $\tilde{q}\equiv q/\theta^2$ parameter also has to remain constant in the $\theta\to\infty$ transition here. The potential thus adapted to the harmonic oscillator limit reads

$$\tilde{v}(r)\equiv v(r)+q\theta=-\frac{1}{r^2}\left(l+\frac{D-3}{2}\right)\left(l+\frac{D-1}{2}\right)+\left(\beta-\frac{1}{2}\right)\left(\beta-\frac{3}{2}\right)\frac{\tilde{C}}{4h(r)(1+h(r)/\theta)}$$
$$-\frac{\tilde{q}h(r)}{1+h(r)/\theta}-\frac{3\tilde{C}}{16\theta}\frac{1}{(1+h(r)/\theta)^2}+\frac{5\tilde{C}}{16\theta}\frac{1}{(1+h(r)/\theta)^3}. \tag{14}$$

The harmonic oscillator potential is recovered from (14) by the $\beta=l+D/2$ and $\tilde{C}\tilde{q}=(2m\omega/\hbar)^2$ choice. The two last terms in (14) vanish, the first and the second cancel out, while the third one reproduces the harmonic oscillator potential. The new form of the energy eigenvalues is

$$\tilde{\epsilon}_n\equiv\epsilon_n+q/\theta=\tilde{C}(2n+\beta)\left[\left(\frac{1}{\theta^2}\left(n+\frac{\beta}{2}\right)^2+\frac{\tilde{q}}{\tilde{C}}\right)^{1/2}-\frac{1}{\theta}\left(n+\frac{\beta}{2}\right)\right], \tag{15}$$

which, indeed, reduces to the $\tilde{\epsilon}_n=(2m\omega/\hbar)(2n+l+D/2)$ oscillator spectrum in the $\theta\to\infty$ limit. The wave functions (7) are unchanged, except for the redefinition of the parameters.

Note that the generalized Coulomb problem establishes a link between the Coulomb problem and the harmonic oscillator potential in different spatial dimensions. This is best seen by inspecting the wave functions (7). If the $\beta-1$ parameter of the associated Laguerre polynomial is required to be the same in the two limits, we get an interrelation between the value of the angular momentum and the spatial dimension to be used for the Coulomb and the harmonic oscillator case:

$$l^O+\frac{D^O}{2}=2l^C+D^C-1. \tag{16}$$

Considering $l^O=2l^C$, Eq. (16) implies $D^O=2D^C-2$ which establishes link between the $(D^C,D^O)=(2,2)$, (3,4), (4,6), (5,8), ..., etc. pairs. [Of these the (3,4) pair corresponds to the Kustaanheimo–Stiefel transformation.[21]] This case is called the ''direct map'' between the Coulombic and oscillator solutions in Ref. 22, while with the $l^O=2l^C+\lambda$ choice the ''general map'' can be recovered. These results suggest that the generalized Coulomb potential can be used to formulate a continuous transition between the Coulomb and harmonic oscillator potentials, as opposed to the usual procedure that employs a unique variable and parameter transformation to reach this goal.

B. The generalized Coulomb–Sturmian basis and the matrix elements of the Green's operator

The generalized Coulomb–Sturmian equation, which depends on n as a parameter, has similar structure to the eigenvalue equation in Eq. (1) with potential (3),

$$\hat{X}\phi(\rho,r)\equiv\Bigg[-\frac{d^2}{dr^2}-\frac{3C}{16(h(r)+\theta)^2}+\frac{5C\theta}{16(h(r)+\theta)^3}+\frac{C(\beta-1/2)(\beta-3/2)}{4h(r)(h(r)+\theta)}$$
$$-\left(\frac{\rho^2\theta}{4}+\rho\left(n+\frac{\beta}{2}\right)\right)\frac{C}{h(r)+\theta}+\frac{C}{4}\rho^2\Bigg]\phi(\rho,r)=0, \tag{17}$$

and is solved by the generalized Coulomb–Sturmian (GCS) functions

$$\langle r|n\rangle\equiv\phi_n(\rho,r)=\left(\frac{\Gamma(n+1)}{\Gamma(n+\beta)}\right)^{1/2}(\rho h(r)+\rho\theta)^{1/4}(\rho h(r))^{(2\beta-1)/4}$$
$$\times\exp\left(-\frac{\rho}{2}h(r)\right)L_n^{(\beta-1)}(\rho h(r)). \tag{18}$$

Here ρ is a parameter characterizing the generalized Coulomb–Sturmian basis. The GCS functions, being solutions of a Sturm–Liouville problem, have the property of being orthonormal with respect to the weight function $C^{1/2}(h(r)+\theta)^{-1}$. Introducing the notation $\langle r|\tilde{n}\rangle \equiv \phi_n(\rho,r)C^{1/2}(h(r)+\theta)^{-1}$ the orthogonality and completeness relation of the GCS functions can be expressed as

$$\langle n|\tilde{n}\rangle=\delta_{nn'} \tag{19}$$

and

$$1=\sum_{n=0}^{\infty}|\tilde{n}\rangle\langle n|=\sum_{n=0}^{\infty}|n\rangle\langle\tilde{n}|. \tag{20}$$

Straightforward calculation shows that both the overlap of two GCS functions and the $\langle n'|\hat{H}_0|n\rangle$ matrix element can be expressed as a tridiagonal matrix, therefore the matrix elements of the $\epsilon-\hat{H}_0$ operator also have this feature:

$$\langle n|\epsilon-\hat{H}_0|n'\rangle=\delta_{nn'}\left[\frac{\epsilon}{C^{1/2}\rho}(2n+\beta-\rho\theta)-\frac{C^{1/2}\rho}{4}\left(-\frac{4q}{C\rho}+(2n+\beta)\right)\right]$$
$$-\delta_{nn'+1}(n(n+\beta-1))^{1/2}\left(\frac{\epsilon}{C^{1/2}\rho}+\frac{C^{1/2}\rho}{4}\right)$$
$$-\delta_{nn'-1}((n+1)(n+\beta))^{1/2}\left(\frac{\epsilon}{C^{1/2}\rho}+\frac{C^{1/2}\rho}{4}\right). \tag{21}$$

This means that, similar to the *D*-dimensional Coulomb and harmonic oscillator potential, the matrix elements of the Green's operator can be determined by using continued fractions, as described in Ref. 16. The present results, therefore, extend the applicability of this method to a new potential problem. The formulas presented here reduce to those in Ref. 16 in the appropriate limits discussed in Sec. III A. The role of the Coulomb–Sturm parameter *b* used in Ref. 16 is now played by $C^{1/2}\rho/2$.

C. Algebraic aspects

Here we show that an SU(1,1) algebra can be associated with the generalized Coulomb problem. The generators of this algebra are

$$\hat{J}_3=\frac{h+\theta}{C\rho}\hat{X}+\left(n+\frac{\beta}{2}\right), \quad \hat{J}_1=\hat{J}_3-\frac{\rho}{2}h,$$
$$\hat{J}_2=-\frac{i}{C^{1/2}}(h(h+\theta))^{1/2}\frac{d}{dr}-\frac{i\theta}{4(h+\theta)}, \tag{22}$$

and they satisfy the commutation relations

$$[\hat{J}_1,\hat{J}_2]=-i\hat{J}_3, \quad [\hat{J}_2,\hat{J}_3]=i\hat{J}_1, \quad [\hat{J}_3,\hat{J}_1]=i\hat{J}_2. \tag{23}$$

As can be seen from Eq. (17), $\hat{J}_3$ is diagonal in the basis (18) with eigenvalues $m=n+\beta/2$. The elements of this basis can then be associated with the infinite-dimensional discrete unitary irreducible representation of SU(1,1) called[23] discrete principal series D_j^+, for which the allowed values of m are

$$m=-j,-j+1,-j+2,\ldots, \tag{24}$$

with j being negative. It is natural then to identify j as $j=-\beta/2$. Direct calculations show that the ladder operators connect the neighboring members of this basis:

$$\hat{J}_+\phi_n(\rho,r)\equiv(\hat{J}_1+i\hat{J}_2)\phi_n(\rho,r)=[(n+1)(n+\beta)]^{1/2}\phi_{n+1}(\rho,r), \tag{25}$$

$$\hat{J}_-\phi_n(\rho,r)\equiv(\hat{J}_1-i\hat{J}_2)\phi_n(\rho,r)=[n(n+\beta-1)]^{1/2}\phi_{n-1}(\rho,r). \tag{26}$$

We find that the eigenvalues of the Casimir invariant

$$\hat{C}_2=\hat{J}_3^2-\hat{J}_1^2-\hat{J}_2^2 \tag{27}$$

are $(\beta/2)(\beta/2-1)=j(j+1)$, as expected, and that they set the strength of the fourth term in (17). For $\theta\neq 0$ this is the only singular term and it behaves like γr^{-2} with $\gamma=(\beta-\frac{1}{2})(\beta-\frac{3}{2})=4j(j+1)+\frac{3}{4}$. It is interesting to inspect the allowed values of γ for the different unitary irreducible representations of SU(1,1). For the discrete principal series D_j^+ $\beta>1$ holds, which always secures $-\frac{1}{4}<\gamma$, i.e., the potential has repulsive or ''weakly attractive''[24] r^{-2}-type singularity. For the supplementary series[23] $-\frac{1}{2}<j<0$ holds, which results in $-\frac{1}{4}<\gamma<\frac{3}{4}$. This is exactly the domain where both independent solutions are square integrable at the origin:[24] for $0<\gamma<\frac{3}{4}$ one of these vanishes at $r=0$ and the other one is infinite there, while for $-\frac{1}{4}<\gamma<0$ both solutions vanish at $r=0$. From (18) it is seen that solutions regular and irregular at the origin correspond to $\beta>\frac{1}{2}$ and $\beta<\frac{1}{2}$. This seems to indicate that one regular solution (with $\beta>1$) is associated with D_j^+, while the second square integrable solution, which is either regular or infinite at the origin (depending on β), might be related to the supplementary series, for which $0<\beta<1$ holds. For the sake of completeness we note that for the continuous series[23] C_k^0 and $C_k^{1/2}$, $j=-\frac{1}{2}+ik$ ($k>0$, real) is valid, which results in the strongly singular $\gamma<-\frac{1}{4}$ case. The solutions then oscillate infinitely near the origin and are unbounded from below,[24] which can be interpreted as the falling of the particle into the center of attraction.[25] The situation is similar to that described in Ref. 26 for the $v(x)=\gamma\sin^{-2}x$ potential: the various unitary irreducible representations of the SU(1,1) spectrum generating group there also corresponded to different types of singularities.

We note that we analyzed the singularities of the generalized Coulomb–Sturmian equation (17) and its solutions (18), but similar considerations of the physical potential (3) and its solutions can also be performed taking $D=3$ and $l=0$. The algebraic construction, however, does not apply to this latter problem. This is because the bound-state solutions (7) pick up extra n-dependence through ρ_n, which is not accessible for the ladder operators otherwise changing n as in Eqs. (25) and (26).

The present realization of the SU(1,1) algebra is a special case of that described in Ref. 17 in relation with the Natanzon confluent potentials. Considering the Coulomb and harmonic oscillator limits discussed in Sec. III A and setting the dimension to $D=3$, the generators reduce to the forms presented for the two problems separately in Ref. 27. We note that the spectrum generating

algebra associated this way with the radial harmonic oscillator problem in three dimensions is different from the two-parameter realization of the SU(1,1) algebra discussed in Refs. 28 and 29, because the ladder operators there are linear differential operators and the Hamiltonian is related to the Casimir invariant, while here the ladder operators are second-order differential operators and the Hamiltonian is essentially a linear function of generator $\hat{J}_3$.

Finally, we note that similar limiting cases of other Natanzon-class potentials can also be treated in terms of SU(1,1) algebras. In Ref. 29 an SU(1,1) algebra associated with the Ginocchio potential[5] was analyzed, and it was shown that in two different limiting cases, which both result in the Pöschl–Teller potential, two different algebras can be recovered from the original one: the first being a spectrum generating algebra and the second a potential algebra.

D. The one-dimensional case

The formalism developed previously is valid for $D=1$, too; nevertheless, some particular properties of one-dimensional problems justify a separate treatment of this case. First, the implicit definition of the $h(r)$ function in Eq. (4) has to be extended to negative values of r, which we now denote with x. Using the notation of (4), we can write that $x=r(h)$ for $x\geqslant 0$ and $x=-r(h)$ for $x<0$. The normalization of the wave functions in (7) also has to be modified with a factor of $2^{-1/2}$, accounting for the fact that the integration now runs from $-\infty$ to ∞, but since the potential is symmetric, the bound-state wave functions are either even or odd, therefore $|\psi(x)|^2$ is an even function of x.

For one-dimensional problems it is natural to set l to 0 besides $D=1$, which eliminates the centrifugal term in (3). Furthermore, in order to avoid r^{-2}-like singularities at $x=0$, the second term in (3) also has to be canceled by setting β to either $\frac{1}{2}$ or to $\frac{3}{2}$. Elementary calculations show that the latter choice corresponds to bound-state wave functions that vanish at $x=0$, and essentially represent physical solutions of the problem in higher dimensions as well, while the former choice recovers solutions that do not vanish in general at $x=0$. These two possibilities can naturally be interpreted as odd and even solutions of the one-dimensional potential problem. Furthermore, for $x\geqslant 0$ the two types of wave functions can be rewritten into a common notation (up to a sign) by making use of the relation of associated Laguerre and Hermite polynomials, when the former ones have $\alpha=\frac{1}{2}$ or $\alpha=-\frac{1}{2}$ as parameters:[30]

$$\psi_N^{(D=1)}(x)=C^{1/4}\tilde{\rho}_N^{3/4}2^{-N}\left(\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{N}{2}+1\right)\left(N+\frac{1}{2}+\theta\tilde{\rho}_N\right)\right)^{-1/2}$$
$$\times(h(x)+\theta)^{1/4}\exp\left(-\frac{\tilde{\rho}_N}{2}h(x)\right)H_N((\tilde{\rho}_N h(x))^{1/2}). \tag{28}$$

For $x\leqslant 0$ the bound-state wave functions satisfy

$$\psi_N^{(D=1)}(-x)=\begin{cases}\psi_N^{(D=1)}(x) & \text{for } N=2n;(\beta=\frac{1}{2}),\\ -\psi_N^{(D=1)}(x) & \text{for } N=2n+1;(\beta=\frac{3}{2}).\end{cases} \tag{29}$$

In (28) we defined $\tilde{\rho}_N$ as

$$\tilde{\rho}_N\equiv\frac{1}{\theta}\left[\left((N+1/2)^2+4\frac{q\theta}{C}\right)^{1/2}-\left(N+1/2\right)\right], \tag{30}$$

which reduces to $\rho_{[N/2]}$, where even and odd values of N have to be chosen with $\beta=\frac{1}{2}$ and $\beta=\frac{3}{2}$, respectively, and the integer part of $N/2$ corresponds to n used in ρ_n in Eq. (6).

An interesting aspect of this potential is that it remains finite at $x=0$ $(v(0)=-q/\theta+C/(8\theta^2))$ for any finite value of θ; however, a narrow, finite peak appears in the $\theta\to 0$ limit, which then becomes an attractive $-3/(16r^2)$-like singularity in the Coulomb limit. This is due to the fourth term in (3) and it corresponds to a ''weak'' singularity in the sense that the center of attraction is not strong enough for the particle to become infinitely bound.[25] This finite barrier

arising for small, but finite θ values also introduces the possibility of studying tunneling effects in symmetric potential wells. We also note that besides being finite at $x=0$, potential (3) has continuous derivative there, as can directly be verified.

Based on these features the $D=1$ version of potential (3) can be used to analyze the peculiarities of the one-dimensional Coulomb potential defined as $v^{(c)}=-e^2/|x|$. This problem has been the subject of intensive studies in the past couple of decades, but there is still some controversy in the interpretation of the results (see, e.g., Ref. 31 for a recent review). The unusual features attributed to this singular problem include degenerate eigenvalues[32] interpreted in terms of a hidden O(2) symmetry,[33] an infinitely bound ground state[32] and continuous bound-state spectrum.[34] The last two of these were later found to be based on unacceptable solutions of the Schrödinger equation,[35,36] while the unexpected degeneracy was explained by an impenetrable barrier at $x=0$, which separates the problem into two disjoint, noncommunicating systems with identical energy spectra[36] and makes even the concept of parity obsolete here.[37] Most authors discussing the one-dimensional Coulomb problem agree that the usual techniques of quantum mechanics alone in dealing with potentials are not sufficient in this case. In Ref. 38, for example, self-adjoint extension of the relevant differential operator has been discussed.

The one-dimensional version of potential (3) can be chosen such that it becomes close to nonsingular potentials used in the approximation of the true one-dimensional Coulomb potential. In fact, with appropriate choice of q and θ, any desired Coulomb asymptotics and $v(x=0)$ value can be generated. Figure 1 shows potentials with rounded-off shape near $x=0$ ($\theta=1$, $q=0.5$) and also ones close to the Coulomb potential with a cutoff $-e^2/(|x|+a)$ ($\theta=1$, $q=2.5$). In contrast with these modified Coulomb potentials, all calculations can be performed exactly with (3). In the $\theta\to 0$ limit the generalized Coulomb potential recovers the one-dimensional Coulomb potential supplemented with the $-\frac{3}{16}x^{-2}$ term. This means that the one-dimensional Coulomb potential cannot be reached exactly, nevertheless, reasonable approximations of it can be given, as it will be demonstrated below.

The odd solutions, of course, vanish at $x=0$, while the even solutions have nonzero value there as long as $\theta>0$ holds. In the $\theta\to 0$ limit $\psi^{(D=1)}_{N=2n}(0)$ varies with $\theta^{1/4}$, so the even solutions also tend to zero at $x=0$. This is in accordance with the discussion in Sec. III C on weakly attractive γr^{-2}-type singular potentials on the half line: for $-\frac{1}{4}<\gamma<0$ both independent solutions vanish at the origin, so the wave functions are necessarily zero at $r=0$. If we try to extend the $N=2n$ solutions (28) in the $\theta=0$ Coulomb limit to the $(-\infty,0)$ domain, we find that, due to its $x^{1/4}$-type behavior at the origin, the derivative of an even wave function would not be continuous anymore.

It is also interesting to compare the Coulomb limit of the one-dimensional generalized Coulomb potential with the true one-dimensional Coulomb potential. Switching off gradually the γr^{-2} term we find that the two types of solutions contain confluent hypergeometric functions of the type $\Phi(-n,c;z)$ with $c=1+(1+4\gamma)^{1/2}$ and $c=1-(1+4\gamma)^{1/2}$. Starting with $\gamma=-\frac{3}{16}$ we recover the $c=\frac{3}{2}$ and $\frac{1}{2}$ corresponding to the two $c=\beta$ values assigned to the allowed solutions of the one-dimensional generalized Coulomb potential and its Coulomb limit. Taking the $\gamma\to 0$ limit results in $c=2$ and $c=0$. The first one of these corresponds to the allowed solutions of the one-dimensional Coulomb potential (see, e.g., Ref. 39), while the latter one has to be rejected, because the confluent hypergeometric function $\Phi(a,c;z)$ is not defined for $c=0$. This shows that even solutions which are allowed for the generalized Coulomb potential in one dimension do not have equivalents in the true one-dimensional Coulomb case.

We also note that the true one-dimensional Coulomb potential can be reached in the $\theta\to 0$ limit, setting β to 2 from the beginning. In this case, however, a repulsive $3/(16r^2)$-type singularity appears, which cancels out only for $\theta=0$. This repulsive singularity then separates the problem into two parts (with $x>0$ and $x<0$) from the beginning, so the potential is not a real one-dimensional one. The $\beta=0$ choice, in principle, also gives the same potential; however, the solutions associated with $\beta=0$ are not defined, as we have shown previously. Using $\beta=\varepsilon$ or $2-\varepsilon$ with ε being small we can get as close to the true one-dimensional Coulomb potential as we wish, but the repulsive singularity is then present for all θ values.

The reflection and transmission coefficients can be analyzed using the asymptotic behavior of the general solutions of the one-dimensional problem. These can be chosen to be even and odd functions of x. Making use of Eq. (8) the even and the odd solutions can be defined for $x\geqslant 0$ in

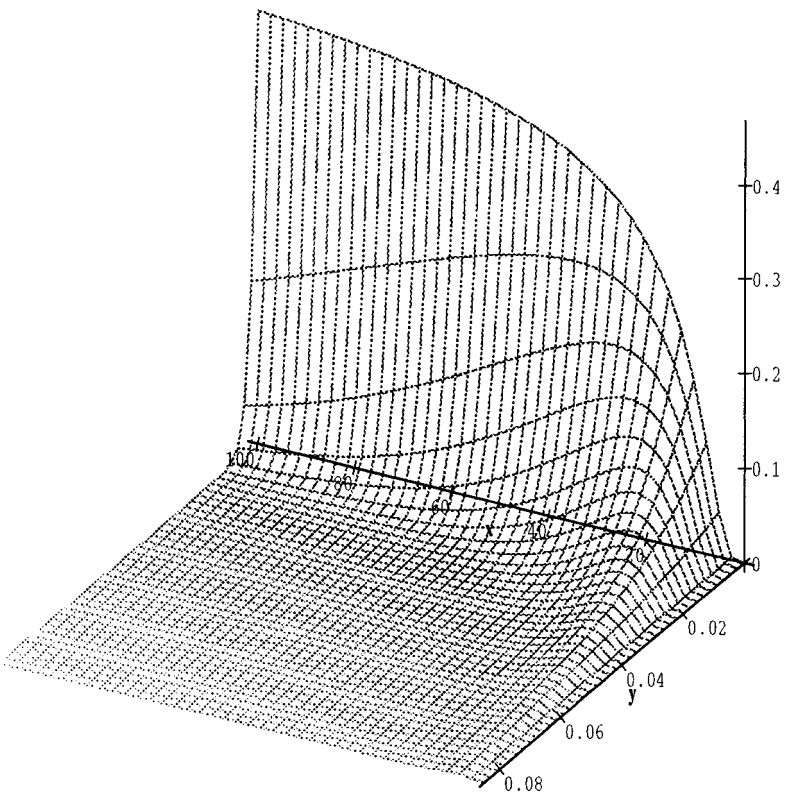

FIG. 3. $|R(k)|^2$ plotted as a function of θ. The notation $x=k$ and $y=\theta$ is used. For technical reasons the plot is made only from $\theta=10^{-5}$ instead of 0.

terms of $\varphi^{(G)}(k,x)$ for $x\geqslant 0$, setting $\beta=\frac{1}{2}$ and $\frac{3}{2}$, respectively; and their extension to $x\leqslant 0$ can be given using a formula similar to Eq. (29). The two solutions are interrelated by Eq. 6.3(3) of Ref. 19. Due to the symmetric nature of the one-dimensional potential $(v(x)=v(-x))$, it is enough to analyze the asymptotic behavior of the solutions for $x\to\infty$: the $x\to-\infty$ case follows naturally. Straightforward calculations show that the reflection coefficient is

$$R(k)=\frac{e^{-i\pi/4}}{2}\left(\frac{\Gamma(\frac{1}{4}+i\nu)}{\Gamma(\frac{1}{4}-i\nu)}-i\frac{\Gamma(\frac{3}{4}+i\nu)}{\Gamma(\frac{3}{4}-i\nu)}\right). \tag{31}$$

Strong reflection is found for potentials having a (finite) barrier in $x=0$ (like those in Fig. 1 with $\theta=0.01$), while more regular shapes (like that in Fig. 1 with $\theta=1$ and $q=2.5$, for example) give weak reflection. The k corresponding to maximal reflection increases as θ gets closer to 0 (i.e., as the barrier at $x=0$ gets higher). $|R(k)|$ is small for very large and very small values of k. Figure 3 presents illustration for $|R(k)|^2$ for various values of θ. Our findings on the reflection coefficient of the one-dimensional generalized Coulomb potential seem to support the existence of the space splitting effect[37] valid for the Coulomb potential on one dimension: in situations close to the true Coulomb case (e.g., $\theta\to 0$) $|R(k)|^2$ is close to 1, which corresponds to a nearly impenetrable barrier at $x=0$. This is also confirmed by our analysis on other approximations of the one-dimensional Coulomb potential, i.e., those with $\beta=\varepsilon$ or $2-\varepsilon$ above.

Besides its mathematical aspects the one-dimensional Coulomb potential has physical relevance, too, in the description of the hydrogen atom in strong magnetic field,[40] for example. In such practical calculations it is reasonable to use a nonsingular Coulomb-like potential instead of the true one-dimensional Coulomb potential: the finite size of the nucleus can be a justification for this. This means that the one-dimensional Coulomb potential might not be sufficient in such calculations: the basis defined with it simply does not contain even-parity states. In practical calculations, therefore, the use of bases like that assigned to the generalized Coulomb potential is necessary.

IV. SUMMARY AND OUTLOOK

We analyzed the generalized Coulomb potential and extended results for the D-dimensional Coulomb and harmonic oscillator potentials to this potential, too, which includes these two simple problems as special cases. The results presented here have both theoretical and practical significance. The theoretical aspects concern the treatment of Coulomb-like problems in one dimension. Due to the singular nature of the one-dimensional Coulomb potential, the usual methods of defining a physically relevant basis are not sufficient in this case. The generalized Coulomb potential avoids these complications and it supplies a basis (including both even and odd functions) which can be used instead of the Coulomb–Sturmian basis, which is otherwise a rather helpful tool in higher dimensions, e.g., the three-body Coulomb problem has been approached this way.[15] We also gave a unification of the SU(1,1) algebras associated with the Coulomb and harmonic oscillator problems separately.

The practical aspects of our work include the introduction of the generalized Coulomb–Sturmian basis, which we used as a basis for analytic representation of the Green's operator. The generalized Coulomb potential (being the member of the rather general Natanzon potential class) offers potential shapes that can approximate various physical problems more accurately than the conventional textbook potentials. In particular, its long-range Coulomb-like behavior makes it a good candidate for describing the Coulomb field of extended (non-point-like) charge distributions. It can be applied either directly, or as a basis for numerical calculations. Its direct applications can be envisaged in describing muonic atoms, where it can model the Coulomb field of the nucleus. It can also be combined with short-range potentials in the potential separable expansion formalism,[14,15] for example. Then the Coulombic asymptotics of the interaction between the strongly interacting charged objects could be taken care of by the Green's operator and only the short-range (nuclear) interaction would have to be approximated.

Similar extensions of other relatively simple (shape-invariant) potentials also seem possible. The example of the generalized Coulomb potential shows that the techniques developed for these well-known potentials can be extended to more general potentials without much technical difficulty. The generalized potentials ''inherit'' many important features of the simple potentials, of which they are extensions, and, at the same time, they can be adapted to more complex physical situations, due to their more flexible shape. There are other examples confirming the expectations that seemingly complicated formulas can be evaluated analytically even for highly nontrivial potentials. In Ref. 41, for example, potentials phase equivalent with the Ginocchio potential[6] have been determined analytically, using the transformations of supersymmetric quantum mechanics.

Based on the present results we believe that the Natanzon-class potentials, which have been analyzed almost exclusively in theoretical studies until now, can be used in practical applications too.

ACKNOWLEDGMENTS

This work has been supported by the OTKA Contracts Nos. F20689 and T17298, and Grant No. JF 345/93 of the US–Hungarian Science and Technology Joint Fund.

[1] G. Lévai, in *Quantum Inversion Theory and Applications*, edited by H. V. von Geramb, Lecture Notes in Physics (Springer, Berlin, 1994), Vol. 427 p. 107; F. Cooper, A. Khare, and U. Sukhatme, Phys. Rep. **251**, 268 (1995); G. Junker, *Supersymmetric Methods in Quantum and Statistical Physics* (Springer, Berlin, 1996).
[2] L. E. Gendenshtein, Zh. Éksp. Teor. Fiz. Pis. Red. **38**, 299 (1983) [English translation: JETP Lett. **38**, 35 (1983)].
[3] G. Lévai, J. Phys. A **22**, 689 (1989).
[4] G. A. Natanzon, Teor. Mat. Fiz. **38**, 146 (1979).
[5] J. N. Ginocchio, Ann. Phys. (N.Y.) **152**, 203 (1984).
[6] J. N. Ginocchio, Ann. Phys. (N.Y.) **159**, 467 (1985).
[7] S. Brajamani and C. A. Singh, J. Phys. A **23**, 3421 (1990); G. Lévai, *ibid.* **24**, 131 (1991); B. W. Williams, *ibid.* **24**, L667 (1991); B. W. Williams and D. Poulios, Eur. J. Phys. **14**, 222 (1993); B. W. Williams, J. L. Rutherford, and G. Lévai, Phys. Lett. A **199**, 7 (1995).
[8] G. Lévai and B. W. Williams, J. Phys. A **26**, 3301 (1993).
[9] A. K. Ushveridze, *Quasi-exactly solvable Models in Quantum Mechanics* (Institute of Physics Publishing, Bristol, 1994).
[10] A. de Souze Dutra, Phys. Rev. A **47**, R2435 (1993); R. Dutt, A. Khare, and Y. P. Varshni, J. Phys. A **28**, L107 (1995).
[11] F. Cooper, J. N. Ginocchio, and A. Khare, Phys. Rev. D **36**, 2458 (1987).
[12] G. Junker and P. Roy, Phys. Lett. A **232**, 155 (1997).
[13] D. Baye, G. Lévai, and J.-M. Sparenberg, Nucl. Phys. A **599**, 435 (1996).

[14] Z. Papp, J. Phys. A **20**, 153 (1987); Phys. Rev. C **38**, 2457 (1988); Phys. Rev. A **46**, 4437 (1992); Comput. Phys. Commun. **70**, 426,435 (1992).
[15] Z. Papp and W. Plessas, Phys. Rev. C **54**, 50 (1996); Z. Papp, *ibid.* **55**, 1080 (1997).
[16] B. Kónya, G. Lévai, and Z. Papp, J. Math. Phys. **38**, 4832 (1997).
[17] P. Cordero and S. Salamó, J. Phys. A **24**, 5299 (1991).
[18] R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).
[19] A. Erdélyi (editor), *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. I.
[20] G. Lévai, in *Proc. 2nd Workshop on Harmonic Oscillators*, NASA Conf. Publications (1994), Vol. 3286, p. 113.
[21] P. Kustaanheimo and E. Stiefel, J. Reine Angew. Math. **218**, 204 (1965).
[22] V. A. Kostelecký, M. M. Nieto, and D. R. Truax, Phys. Rev. D **32**, 2627 (1985).
[23] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[24] W. M. Frank, D. J. Land, and R. M. Spector, Rev. Mod. Phys. **43**, 36 (1971).
[25] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, Oxford, 1977).
[26] G. Lévai, in *Proceedings of the IVth Wigner Symposium*, edited by N. M. Atakishiyev, T. H. Seligman, and K. B. Wolf (World Scientific, Singapore, 1996), p. 88.
[27] I. L. Cooper, J. Phys. A **26**, 1601 (1993).
[28] G. Lévai, J. Phys. A **27**, 3809 (1994).
[29] G. Lévai, in *Proceedings of the XXIth International Colloquium on Group Theoretical Methods in Physics, Vol. 1*, edited by H.-D. Doebner, P. Nattermann, and W. Scherer (World Scientific, Singapore, 1997), p. 461.
[30] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).
[31] A. N. Gordeyev and S. C. Chhajlany, J. Phys. A **30**, 6893 (1997).
[32] R. Loudon, Am. J. Phys. **27**, 649 (1959).
[33] L. S. Davtyan, G. S. Pogosyan, A. N. Sissakian, and V. M. Ter-Antonyan, J. Phys. A **20**, 2765 (1987).
[34] L. K. Haines and D. H. Roberts, Am. J. Phys. **37**, 1145 (1969).
[35] M. Andrews, Am. J. Phys. **34**, 1194 (1966).
[36] M. Andrews, Am. J. Phys. **44**, 1064 (1976).
[37] U. Oseguera and M. de Llano, J. Math. Phys. **34**, 4575 (1993).
[38] F. Gesztesy, J. Phys. A **13**, 867 (1980); W. Fischer, H. Leschke, and P. Müller, J. Math. Phys. **36**, 2313 (1995).
[39] R. G. Newton, J. Phys. A **27**, 4717 (1994).
[40] H. Ruder, G. Wunner, H. Herold, and F. Geyer, *Atoms in Strong Magnetic Fields* (Springer, Berlin, 1994).
[41] G. Lévai, D. Baye, and J.-M. Sparenberg, J. Phys. A **30**, 8257 (1997).

Sigma models coupled with Abelian gauge fields

D. O'Keeffe
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland

D. H. Tchrakian[a)]
Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland and School of Theoretical Physics—DIAS, 10 Burlington Road, Dublin 4, Ireland

Yisong Yang
Department of Applied Mathematics and Physics, Polytechnic University, Brooklyn, New York 11201

(Received 9 February 1998; accepted for publication 8 July 1998)

Two distinct Skyrme terms for the $\mathbb{C}P^1$ model are defined, only one of which preserves the equivalence with the corresponding O(3) model. A gauging prescription in which the composite connection of the $\mathbb{C}P^1$ model is replaced by the U(1) gauge connection is defined. A U(1) gauged $\mathbb{C}P^1$ model and two distinct U(1) gauged $\mathbb{C}P^1$ models with Skyrme terms are defined. The existence of topologically stable vortices in each of the three U(1) gauged models is proved analytically. © *1998 American Institute of Physics.* [S0022-2488(98)02211-7]

I. INTRODUCTION

Abelian gauge theories in $2+1$ dimensions which support topologically stable vortex solutions have proved to be useful models in various areas of theoretical physics; for example, the Abelian Higgs model[1] has been employed in superconductivity and cosmology. More recently, Abelian gauge theories in which the matter field is either $\mathbb{C}P^1$ valued or S^2 (two-sphere) valued, instead of the complex Higgs field of the Abelian Higgs model, have been introduced.[2,3] These are the various two dimensional gauged sigma models, which describe the static vortices of the corresponding $2+1$ dimensional theories. The study of certain Maxwell gauged sigma models in two dimensions is the purpose of this work.

While there are some qualitative and quantitative differences between these U(1) gauged sigma models and the well known Abelian Higgs model, we will not concentrate on this interesting area of comparisons here. We suffice to mention the most obvious of these differences, namely that in the limit of the gauge field decoupling the soliton of the gauged sigma model persists, while the soliton of the Abelian Higgs model collapses.

The two dimensional sigma models under consideration are the $\mathbb{C}P^1$ and the O(3) models. There are *two* distinct U(1) gauging prescriptions for these models.

The *first* gauging prescription[2] involves replacing the composite Abelian connection[4] which is defined in terms of the $\mathbb{C}P^1$ valued field, with the U(1) gauge connection. Using this prescription the ungauged $\mathbb{C}P^1$ models and the O(3) models [obtained from the well known equivalence (4) given below, of the $\mathbb{C}P^1$ valued and the S^2 valued fields] can be gauged. These include both the scale invariant $\mathbb{C}P^1$ model,[4] and the $\mathbb{C}P^1$ model with Skyrme terms. This is the gauging prescription that will be employed in the present work.

The *second* gauging prescription,[3] which applies to the O(3) models only, involves the gauging of an SO(2) subgroup of the global SO(3) group under which these models are invariant. These include both the scale invariant $\mathbb{C}P^1$ model,[4] and the $\mathbb{C}P^1$ model with Skyrme terms.

The most important difference between the two gauging prescriptions is that in the *first* case,[2] the solutions of the resultant gauged models do not saturate the topological lower bound, while in the *second* case,[3] there are gauged models in which the Bogomol'nyi equations support nontrivial

[a)]Electronic mail: tigran.tchrakian@may.ie

 © 1998 American Institute of Physics

solutions. In the case of the $\mathbb{C}P^1$ or O(3) sigma models being gauged according to the *first* prescription, the Bogomol'nyi equations are overdetermined,[5] and as a consequence they have only trivial solitons. We do not give details of the demonstration here since that is given already in Section II of Ref. 5. Inasfar as the existence of self-dual solutions saturating the topological lower bound is concerned, the *second* gauging prescription is mathematically more interesting.

It is nevertheless worthwhile to investigate both the models with and without and Skyrme terms arising from the *first* gauging prescription because they are, in principle, just as relevant to physical applications as the models arising from the second gauging prescription, and in some ways even more interesting physically. Indeed the $\mathbb{C}P^1$ models gauged according to the first[2] prescription, which do not support self-dual solutions, have a distinctive physical quality contrasting with the O(3) model gauged according to the second prescription.[3] This concerns the dual properties of spontaneous symmetry breaking and quantized magnetic flux which are absent in the O(3) models (with or without Skyrme terms) gauged according to the *second* prescription,[3] and which is present in the gauged $\mathbb{C}P^1$ models or their equivalent O(3) models gauged according to the *first* prescription.[2] We can easily see this by noticing that the asymptotic (vacuum) value of the S^2 valued field ϕ^a is $\phi_0^a=(0,0,1)$, which completely preserves the local SO(2) symmetry acting on its first two components, leading to no symmetry breaking. By contrast, the local Abelian U(1) symmetry acts on both of the two components of the $\mathbb{C}P^1$ valued field z^α, with vacuum configuration $z_0^\alpha=(e^{-in\theta},0)$. Hence the symmetry in the latter case is broken and the U(1) gauge field receives a mass proportional to $z^\dagger z=1$. In this sense, the first gauging prescription[2] gives rise to models which have a physically richer particle spectrum. Concerning the arbitrary values of the magnetic flux in the models of the *second* type, this is a consequence of the conditions for the finiteness of energy, yielding topologically significant boundary conditions.

In this paper we employ the *first* gauging prescription, namely that used in Ref. 2, in which the composite connection

$$B_i=iz^\dagger\partial_i z=iz_\alpha^*\partial_i z^\alpha \tag{1}$$

is replaced by a dynamical (independent) U(1) connection. The $\mathbb{C}P^1$ valued field z^α, $\alpha=1,2$ is subject to the constraint $z^\dagger z=z_\alpha^* z^\alpha=1$, and the covariant derivative defined in terms of the composite connection B_i is

$$D_i^{(B)}z=\partial_i z+iB_i z. \tag{2}$$

In Sec. II we introduce the action densities, or in the context of soliton physics the static Hamiltonian densities, and subject them to radial symmetry. In Sec. II A we present the $\mathbb{C}P_1$ models, in their ungauged forms, and then in Sec. II B we give the corresponding gauged models. In Sec. II C we subject the static Hamiltonian densities to radial symmetry. In Sec. III, we give the existence proofs for the vortices in each of these models.

II. THE MODELS

After presenting the ungauged $\mathbb{C}P^1$ models, we will gauge these with a U(1) gauge field described by Maxwell dynamics according to the *first* gauging prescription described above. Then we will impose radial symmetry.

It is equally possible to consider the U(1) gauging with a Chern–Simons (CS) rather than a Maxwell term, as for example in Refs. 6 and 2. There will be no need to do this, however, since the static Hamiltonian density pertaining to the CS-$\mathbb{C}P^1$ model turns out to be identical with the corresponding quantity for the Maxwell–$\mathbb{C}P^1$ model. Hence the corresponding proof given in Sec. III holds for both these systems. It is interesting to note that this coincidence occurs only for the U(1) gauged $\mathbb{C}P^1$ model [or the equivalent O(3) model] subject to the *first* gauging prescription. By contrast, when one employs CS dynamics to gauge the Abelian Higgs model,[7] or to gauge the O(3) model[8,9,5] applying the *second* prescription, one finds that the static Hamiltonian densities of the CS gauged models are quite different from the corresponding quantities of the Maxwell gauged models.[1,3]

A. The ungauged models

In terms of the covariant derivative (2), the action density of the $\mathbb{C}P^1$ model is defined as

$$\mathscr{L}^{(z)} = D_i^{(B)} z^\dagger D_i^{(B)} z = D_i^{(B)} z_\alpha^* D_i^{(B)} z^\alpha. \tag{3}$$

Clearly the system (3) exhibits a local U(1) symmetry given by the composite gauge connection (1).

Since there is a natural isomorphism between the two symmetric coset spaces $\mathbb{C}P^1$ and S^2, the S^2 valued field ϕ^a, $a=1,2,3$, subject to the constraint $|\phi^a|^2=1$ can be expressed in terms of the Pauli matrices σ^a as follows:

$$\phi^a = z^\dagger \sigma^a z = z_\alpha^* (\sigma^a)^\alpha{}_\beta z^\beta. \tag{4}$$

The action density of the O(3) sigma model is defined as

$$\mathscr{L}^{(\phi)} = |\partial_i \phi^a|^2, \tag{5}$$

which can be shown to be equal to the action density (3) of the $\mathbb{C}P^1$ model by substituting (4) in (5). Formally, we have

$$\mathscr{L}^{(\phi)} = 4\mathscr{L}^{(z)}. \tag{6}$$

This equivalence between the densities of the $\mathbb{C}P^1$ model and the O(3) model does not persist when Skyrme (kinetic) terms are introduced.

The Skyrme term of the O(3) sigma model[10] is $|\partial_{[i}\phi^a \partial_{j]}\phi^b|^2$ and it can be expressed in terms of the covariant derivatives (2) and using (4) as follows,

$$|\partial_{[i}\phi^a \partial_{j]}\phi^b|^2 = 2((|\partial_i \phi^a|^2)^2 - |\partial_{(i}\phi^a \partial_{j)}\phi^a|^2) = 2(4|D_i^{(B)} z^\dagger D_i^{(B)} z|^2 - |D_{(i}^{(B)} z^\dagger D_{j)}^{(B)} z|^2, \tag{7}$$

where the $a_{(i}b_{j)}$ denotes the *symmetrization* of the indices i and j, and $a_{[i}b_{j]}$ denotes *antisymmetrization*.

The Skyrme term of the $\mathbb{C}P^1$ sigma model can be chosen in *two* different ways. In the first case, one simply defines the Skyrme term of the $\mathbb{C}P^1$ sigma model to be the Skyrme term (7) of the O(3) model, and so the equivalence of the O(3) model with Skyrme term and $\mathbb{C}P^1$ model with Skyrme term is maintained.

In the second case, a Skyrme term subject to the usual criteria is defined. These criteria are that a kinetic term should feature only the second power of the velocity fields and, in the present case, exhibit the local U(1) gauge invariance characterized by the composite connection B_i in (1). This Skyrme term, denoted, which is simpler and perhaps more natural, is

$$|D_{[i}^{(B)} z^\dagger D_{j]}^{(B)} z|^2 = |D_{[i}^{(B)} z_\alpha^* D_{j]}^{(B)} z^\alpha|^2. \tag{8}$$

The $\mathbb{C}P^1$ model with this Skyrme term is *not* equivalent to the O(3) model with Skyrme term.

The action density of the O(3) model with Skyrme term[10] is

$$\mathscr{L}_{Sk}^{(\phi)} = \kappa_2^4 |\partial_{[i}\phi^a \partial_{j]}\phi^b|^2 + \kappa_2^2 |\partial_i \phi^a|^2 + \lambda_0(1-\phi^3). \tag{9}$$

The action density of the $\mathbb{C}P^1$ model with Skyrme term which is equivalent to the O(3) model with Skyrme term (9) is

$$\mathscr{L}_{Sk1}^{(z)} = 2\kappa_2^4 (4|D_i^{(B)} z^\dagger D_i^{(B)} z|^2 - |D_{(i}^{(B)} z^\dagger D_{j)}^{(B)} z|^2) + \kappa_1^2 D_i^{(B)} z^\dagger D_i^{(B)} z + \lambda_0(1 - z^\dagger \sigma^3 z). \tag{10}$$

The action density of the alternative $\mathbb{C}P^1$ model with Skyrme term which is not equivalent to the O(3) model with Skyrme term (9) is

$$\mathscr{L}_{Sk2}^{(z)} = \kappa_2^4 |D_{[i}^{(B)} z^\dagger D_{j]}^{(B)} z|^2 + \kappa_1^2 D_i^{(B)} z^\dagger D_i^{(B)} z + \lambda_0(1 - z^\dagger \sigma^3 z). \tag{11}$$

The constants κ_1 and κ_2 have the dimensions of length, while the constant λ_0 is dimensionless. Equations (10) and (11) are the two sigma models with Skyrme terms which will be gauged using Maxwell dynamics.

B. U(1) Maxwell gauging prescription

We now subject the $\mathbb{C}P^1$ models with Skyrme terms (10) and (11) to our U(1) Maxwell gauging prescription which consists of replacing the composite connection B_i by the Maxwell U(1) connection A_i, i.e., by replacing the covariant derivatives defined by (2) with

$$D_i^{(A)}z=\partial_i z+iA_i z. \tag{12}$$

We also add the usual Maxwell term F_{ij}^2, where F_{ij} is the curvature of the connection in (12), to the action density. This results in the following two gauged $\mathbb{C}P^1$ models with Skyrme terms

$${}^g\mathscr{L}_{Sk1}^{(z)}=\kappa_3^4F_{ij}^2+2\kappa_2^4(4|D_i^{(A)}z^\dagger D_i^{(A)}z|^2-|D_{(i}^{(A)}z^\dagger D_{j)}^{(A)}z|^2)+\kappa_1^2D_i^{(A)}z^\dagger D_i^{(A)}z+\lambda_0(1-z^\dagger\sigma^3z). \tag{13}$$

$${}^g\mathscr{L}_{Sk2}^{(z)}=\kappa_3^4F_{ij}^2+\kappa_2^4|D_{[i}^{(A)}z^\dagger D_{j]}^{(A)}z|^2+\kappa_1^2D_i^{(A)}z^\dagger D_i^{(A)}z+\lambda_0(1-z^\dagger\sigma^3z). \tag{14}$$

In (13) and (14), κ_3 is a coupling constant which has the dimension of length. If $\kappa_2=0$, then ${}^g\mathscr{L}_{Sk1}^{(z)}={}^g\mathscr{L}_{Sk2}^{(z)}$. We note that ${}^g\mathscr{L}_{Sk1}^{(z)}$ with $\kappa_2=0$, which is the Maxwell–U(1) gauged $\mathbb{C}P^1$ model (without Skyrme term), when subjected to radial symmetry coincides with the Chern–Simons-U(1) gauged $\mathbb{C}P^1$ model studied in Ref. 6 as will be seen in the next section.

C. Radial symmetry

We employ the following radially symmetric ansätze:

$$A_i=\frac{a(r)-n}{r}\varepsilon_{ij}\hat{x}_j, \tag{15}$$

$$\phi^a=(\phi^\alpha,\phi^3)=(\sin f(r)n^\alpha,\cos f(r)); \quad n^\alpha=(\cos n\theta,\sin n\theta), \tag{16}$$

$$z^\alpha=\left(\cos\frac{f(r)}{2}\,e^{-in\theta},\sin\frac{f(r)}{2}\right). \tag{17}$$

Under these ansätze we find that the radially symmetric restrictions of the $\mathbb{C}P^1$ models with Skyrme terms (10) and (11), as well as the U(1) gauged $\mathbb{C}P^1$ models with Skyrme terms (13) and (14), reduce to the following one dimensional Lagrangian densities:

$$L_{Sk}^{(\phi)}=L_{Sk1}^{(z)}=\lambda_2n^2\frac{\sin^2 f}{\rho}f_\rho^2+\lambda_1\left(\rho f_\rho^2+n^2\,\frac{\sin^2 f}{\rho}\right)+\lambda_0 r(1-\cos f\,), \tag{18}$$

$$L_{Sk2}^{(z)}=\lambda_2n^2\frac{\sin^2 2f}{\rho}f_\rho^2+\lambda_1\left(\rho f_\rho^2+n^2\,\frac{\sin^2 f}{\rho}\right)+\lambda_0 r(1-\cos f\,), \tag{19}$$

$$\begin{aligned}{}^gL_{Sk}^{(\phi)}={}^gL_{Sk1}^{(z)}=&\frac{\lambda_3}{\rho}a_\rho^2+2\,\frac{\lambda_2}{\rho}\left[a^2\,\cos^2\frac{f}{2}+(a-n)^2\,\sin^2\frac{f}{2}\right]f_\rho^2\\&+\lambda_1\left(\frac{\rho}{4}f_\rho^2+\frac{1}{\rho}\left[a^2\,\cos^2\frac{f}{2}+(a-n)^2\,\sin^2\frac{f}{2}\right]\right)+\lambda_0\rho(1-\cos f\,),\end{aligned} \tag{20}$$

$${}^gL_{Sk2}^{(z)}=\frac{\lambda_3}{\rho}a_\rho^2+\frac{\lambda_2}{2}n^2\frac{\sin^2 2f}{\rho}f_\rho^2+\lambda_1\left(\frac{\rho}{4}f_\rho^2+\frac{1}{\rho}\left[a^2\,\cos^2\frac{f}{2}+(a-n)^2\,\sin^2\frac{f}{2}\right]\right)+\lambda_0\rho(1-\cos f\,). \tag{21}$$

When $\lambda_2=0$, ${}^gL^{(z)}_{Sk1}={}^gL^{(z)}_{Sk2}$. Let ${}^gL^{(z)}_0$ denote ${}^gL^{(z)}_{Sk1}$ with $\lambda_2=0$. ${}^gL^{(z)}_0$, which is the radially symmetric restriction of the action density of the $\mathbb{C}P^1$ sigma model gauged with Maxwell dynamics, coincides with the corresponding quantity for the $\mathbb{C}P^1$ sigma model gauged with Chern–Simons dynamics of Ref. 6.

$$ {}^gL^{(z)}_0=\frac{\lambda_3}{\rho}a_\rho^2+\lambda_1\left(\frac{\rho}{4}f_\rho^2+\frac{1}{\rho}\left[a^2\,\cos^2\frac{f}{2}+(a-n)^2\,\sin^2\frac{f}{2}\right]\right)+\lambda_0\rho(1-\cos\,f\,). \tag{22}$$

In (18)–(22), we have used a dimensionless radial variable $\rho=\kappa r$ and the dimensionless coupling constants there are defined by $\lambda_1=(\kappa_1/\kappa)^2$, $\lambda_2=(\kappa_2/\kappa)^4$ and $\lambda_3=(\kappa_3/\kappa)^4$.

We note that after the imposition of radial symmetry the Skyrme term in (21) does not exhibit the gauge field function $a(\rho)$.

In the next section we present the proofs of existence for the solitons in the models described above, *analytically*. In connection with this, we emphasize that the existence of these solitons was verified *numerically* for several values of the vorticity n, in Refs. 10, 6, and 2.

III. EXISTENCE OF SOLITONS

A proof of the existence of smooth finite-energy solutions with winding number n of the variational equation of (18) was given in Ref. 11. It is clear that the existence of solutions with winding number n of the variational equation of (19) can be proved in a similar manner to the proof given in Ref. 11 so we do not concern ourselves with it here. We proceed therefore to give the proofs of existence for the gauged Skyrme-Sigma models (20)–(22).

Consider the energy density functionals ${}^gL^{(z)}_0$, ${}^gL^{(z)}_{Sk1}$, ${}^gL^{(z)}_{Sk2}$ and their associated energies. The boundary conditions for solutions of the variational equations with finite energy and with topological charge n given in Refs. 6 and 2 are

$$\lim_{\rho\to 0}\,a(\rho)=n,\quad \lim_{\rho\to\infty}\,a(\rho)=0,\quad \lim_{\rho\to 0}\,f(\rho)=\pi,\quad \lim_{\rho\to\infty}\,f(\rho)=0. \tag{23}$$

Theorem: *The Euler-Lagrange equations of each of the three functionals* ${}^gL^{(z)}_0$, ${}^gL^{(z)}_{Sk1}$ *and* ${}^gL^{(z)}_{Sk2}$ *have finite energy solutions with winding number n satisfying (23).*

To prove the theorem it suffices to find minimizers of each of the energies which satisfy the boundary conditions (23).

To each of the three energies we associate the space of admissible functions:

$$\begin{aligned}\mathscr{A}=\{(a(\rho),f(\rho)):\ & a(\rho)\ \text{and}\ f(\rho)\ \text{are continuous functions on}\ [0,\infty)\ \text{and}\\ & \text{absolutely continuous on every compact subinterval of}\ (0,\infty)\\ & \text{and}\ \lim_{\rho\to 0}\,a(\rho)=n,\quad \lim_{\rho\to 0}\,f(\rho)=\pi,\quad \lim_{\rho\to\infty}\,f(\rho)=0\},\end{aligned}$$

and we consider the optimization problems

$$m\equiv\inf_{(a,f\,)\in\mathscr{A}}\,E(a,f\,), \tag{24}$$

where $E(a,f\,)$ denotes the energy of the respective model which is proportional to

$$\int_0^\infty L\,d\rho,\quad L={}^gL^{(z)}_0,\ {}^gL^{(z)}_{Sk1},\ \text{or}\ {}^gL^{(z)}_{Sk2}.$$

Lemma 1: Let $\{(a_k,f_k)\}$ be a minimizing sequence for any of the three optimization problems (24). Then

$$0\leqslant a_k\leqslant n,\quad 0\leqslant f_k\leqslant\pi,\quad k=1,2,\ldots. \tag{25}$$

Proof: Let $\{(a_k,f_k)\}$ be a minimizing sequence for any of the three optimization problems (24). Define

$$g_k(\rho)=\begin{cases} -f_k(\rho), & \text{if } f_k(\rho)<0,\\ 2\pi-f_k(\rho), & \text{if } f_k(\rho)>\pi,\\ f_k(\rho), & \text{otherwise,}\end{cases}$$

$$b_k(\rho)=\begin{cases} 0, & \text{if } a_k(\rho)<0,\\ n, & \text{if } a_k(\rho)>n,\\ a_k(\rho), & \text{otherwise.}\end{cases}$$

The sequence $\{(b_k,g_k)\}$ lies in $\mathscr{A}$ for any of the three optimization problems, (24). Since $E(b_k,g_k)\leqslant E(a_k,f_k)$ for each of the three energies, $k=1,2,\ldots,$ it follows that $\{(b_k,g_k)\}$ is also a minimizing sequence for any of the three optimization problems (24) which satisfies (25).

Let $\{(a_k,f_k)\}$ be a minimizing sequence for (24) which satisfies (25). Then for any pair of numbers $0<\rho_1<\rho_2$, $\{(a_k,f_k)\}$ is a bounded minimizing sequence in the space $W^{1,2}(\rho_1,\rho_2)$. Hence there exists a pair of functions (a,f) and a subsequence of $\{(a_k,f_k)\}$ [which we also denote by $\{(a_k,f_k)\}$], such that $a_k\to a$ and $f_k\to f$ weakly in $W^{1,2}(\rho_1,\rho_2)$ (as $k\to\infty$) for any ρ_1 and ρ_2 with $0<\rho_1<\rho_2$.

Since there is a compact embedding $W^{1,2}(\rho_1,\rho_2)\to C[\rho_1,\rho_2]$, it follows that $a_k\to a$ and $f_k\to f$ strongly in $C[\rho_1,\rho_2]$ (as $k\to\infty$) for any $0<\rho_1<\rho_2$.

Hence

$$E(a,f)=\lim_{\rho_1\to 0,\rho_2\to\infty}\sup\ \lim_{k\to\infty}\inf\int_{\rho_1}^{\rho_2}\mathscr{E}(a_k,f_k)d\rho\leqslant\lim_{k\to\infty}E(a_k,f_k)=m$$

for any of the three optimization problems (24).

Lemma 2: Let (a,f) be the field configuration pair obtained above in the limit of the sequence $\{(a_k,f_k)\}$. Then (a,f) is a solution of the minimization problem (24).

Proof: It remains to show that (a,f) satisfies the boundary conditions stated in the definition of the admissible space $\mathscr{A}$.

Let $\{(a_k,f_k)\}$ be a minimizing sequence for any of the three optimization problems (24). Since

$$|a_k(\rho)-n|\leqslant\int_0^\rho\left|\frac{da_k}{d\rho'}\right|d\rho'\leqslant\rho\left(\int_0^\infty\frac{1}{\rho'}\left[\frac{da_k}{d\rho'}\right]^2d\rho'\right)^{1/2}\leqslant\rho E^{1/2}(a_k,f_k)\to 0\quad\text{as}\quad\rho\to 0,$$

uniformly for $k=1,2,\ldots,$ for each of the three energies E, so the limit $a(\rho)\to n$ as $\rho\to 0$ follows.

Furthermore, since

$$\begin{aligned}1-\cos f_k&=\frac{f_k^2}{2!}\left(1-\frac{f_k^2}{3\cdot 4}\right)+\frac{f_k^6}{6!}\left(1-\frac{f_k^2}{7\cdot 8}\right)+\cdots\\&\geqslant\frac{f_k^2}{2!}\left(1-\frac{\pi^2}{3\cdot 4}\right)+\frac{f_k^6}{6!}\left(1-\frac{\pi^2}{7\cdot 8}\right)+\cdots>\frac{12-\pi^2}{4!}f_k^2,\end{aligned}$$

then it follows from the respective energy functionals that $\{f_k\}$ satisfies

$$\sup_k\int_0^\infty\rho\left(f_k^2+\left[\frac{df_k}{d\rho}\right]^2\right)d\rho<C$$

for some constant $C>0$. Since $f_k(\rho)\to 0$ as $\rho\to\infty$, then

$$f_k^2(\rho)=-2\int_\rho^\infty f_k\frac{df_k}{d\rho'}\,d\rho'\leqslant\frac{2}{\rho}\left(\int_0^\infty\rho' f_k^2(\rho')d\rho'\right)^{1/2}\left(\int_0^\infty\rho'\left[\frac{df_k}{d\rho'}\right]^2d\rho'\right)^{1/2}\leqslant\frac{2C}{\rho},\quad\rho>0.$$

Hence $f_k(\rho)\to 0$ as $\rho\to\infty$ uniformly for $k=1,2,\ldots$ for each of the energy E. Thus it follows that $f(\rho)\to 0$ as $\rho\to\infty$.

To obtain the boundary condition for f at $\rho=0$, we use the uniform limit $a_k(\rho)\to n\neq 0$ as $\rho\to 0$ to find a number $\delta>0$ so that

$$a_k^2(\rho)>\tfrac{1}{2},\quad k=1,2,\ldots,\quad 0<\rho<\delta.$$

For the energy defined by ${}^gL_{Sk1}^{(z)}$, we have from (23), for all $k=1,2,\ldots$, that

$${}^gL_{Sk1}^{(z)}(a_k,f_k)\geqslant C\frac{1}{\rho}\cos^2\left(\frac{f_k}{2}\right)\left(\frac{df_k}{d\rho}\right)^2,\quad 0<\rho<\delta.$$

Therefore, we see from the estimate

$$\begin{aligned}\left|\sin^2\left(\frac{f_k(\rho)}{2}\right)-1\right|&\leqslant\int_0^\rho\left|\sin\ \text{frac}\ \ f_k 2\cos\frac{f_k}{2}\frac{df_k}{d\rho'}\right|d\rho'\\&\leqslant\rho\left(\int_0^\rho\frac{1}{\rho'}\cos^2\frac{f_k}{2}\left[\frac{df_k}{d\rho'}\right]^2d\rho'\right)^{1/2}\\&\leqslant C\rho E^{1/2}(a_k,f_k),\quad k=1,2,\ldots,\quad 0<\rho<\delta\end{aligned}$$

that $f_k(\rho)\to\pi$ as $\rho\to 0$ uniformly for all k. Hence $f(\rho)\to\pi$ as $\rho\to 0$ as well.

Similarly, we consider the energy defined by ${}^gL_{Sk2}^{(z)}$. Since

$$|\cos\ f_k(\rho)+1|\leqslant\int_0^\rho\left|\sin\ f_k\frac{df_k}{d\rho'}\right|d\rho'\leqslant\rho\left(\int_0^\rho\frac{1}{\rho'}\sin^2\ f_k\left[\frac{df_k}{d\rho'}\right]^2d\rho'\right)^{1/2}\leqslant C\rho E^{1/2}(a_k,f_k)\to 0$$

$$\text{as}\ \ \rho\to 0$$

uniformly for $k=1,2,\ldots$, it follows that $f(\rho)\to\pi$ as $\rho\to 0$.

Using the established fact that $a(\rho)\to 0$ as $\rho\to 0$ and finite energy for the model defined by ${}^gL_0^{(z)}$, we have the convergent integral

$$\int_0^1\frac{1}{\rho}\cos^2\left(\frac{f}{2}\right)d\rho<\infty.$$

Consequently, we obtain

$$\left|\sin^2\left(\frac{f_k(\rho)}{2}\right)-1\right|\leqslant\int_0^\rho\left|\cos\frac{f_k}{2}\frac{df_k}{d\rho'}\right|d\rho'\leqslant C\left(\int_0^\rho\frac{1}{\rho'}\cos^2\frac{f_k}{2}\,d\rho'\right)^{1/2}\sup_k E^{1/2}(a_k,f_k).$$

Applying the Schwarz inequality first and then taking the limit $k\to\infty$, we get

$$\left|\sin^2\left(\frac{f(\rho)}{2}\right)-1\right|\leqslant C_1\left(\int_0^\rho\frac{1}{\rho'}\cos^2\frac{f}{2}\,d\rho'\right)^{1/2}.$$

Letting $\rho\to 0$ we have $f(\rho)\to\pi$ as desired and the proof of Lemma 2 is complete.

It remains to establish the boundary condition $a(\rho)\to 0$ as $\rho\to\infty$. As a preparation, we prove the following.

Lemma 3: Let (a,f) be a minimizer of the energy of any of the three models. Then there hold

$$0<a(\rho)<n,\quad 0<f(\rho)<\pi,\quad \rho>0. \tag{26}$$

Proof: For the model ${}^gL_0^{(z)}$, the Euler–Lagrange equations are

$$\lambda_1\frac{d}{d\rho}\left(\rho\frac{df}{d\rho}\right)=\frac{n}{\rho}\lambda_1(n-2a)\sin\ f+2\lambda_0\rho\ \sin\ f, \tag{27}$$

$$\lambda_3 \frac{d}{d\rho}\left(\frac{1}{\rho}\frac{da}{d\rho}\right)=\frac{\lambda_1}{\rho}\left(a-n\,\sin^2\frac{f}{2}\right). \tag{28}$$

From Lemma 1, we have $0\leqslant a(\rho)\leqslant n$, $0\leqslant f(\rho)\leqslant\pi$. On the other hand, since $f=0,\pi$ are the equilibrium solutions of (27), an easy application of the uniqueness theorem for the solutions of ordinary differential equations shows that $0<f(\rho)<\pi$ for all $\rho>0$.

If there is a $\rho_0>0$ so that $a(\rho_0)=0$, then $a(\rho_0)$ is the minimum of a and $a_\rho=0$ at $\rho=\rho_0$. Inserting this information into (28), we have $a_{\rho\rho}(\rho_0)<0$ which is a contradiction. Similarly, if $a(\rho_0)=n$, then $a_\rho=0$ and $a_{\rho\rho}\leqslant 0$ at $\rho=\rho_0$ which violates again (28).

For a model defined either by ${}^gL^{(z)}_{Sk1}$ or ${}^gL^{(z)}_{Sk2}$, we may use the same argument. Hence the lemma is proved.

Lemma 4: We have the boundary behavior $a(\rho)\to 0$ as $\rho\to\infty$.

Proof: The structure of each of the energy densities implies that there is a sequence $\{\rho_j\}$, $\rho_j\to\infty$ as $j\to\infty$ so that

$$\lim_{j\to\infty} a(\rho_j)=0 \tag{29}$$

because $f(\rho)\to 0$ as $\rho\to\infty$. Suppose otherwise that $a(\rho)\not\to 0$ as $\rho\to\infty$. Then there is some $\varepsilon_0>0$ and a sequence $\{\rho_j'\}$, $\rho_j'\to\infty$ as $j\to\infty$ so that

$$a(\rho_j')\geqslant\varepsilon_0,\quad j=1,2,\ldots. \tag{30}$$

From (26) and (28) it follows that there is a sequence $\{\rho_j^0\}$, $\rho_j^0\to\infty$ as $j\to\infty$ such that $a(\rho_j^0)\geqslant\varepsilon_0$ and ρ_j^0 are local maximum points of $a(\rho)$. Inserting $a_\rho(\rho_j^0)=0$ and $a_{\rho\rho}(\rho_j^0)\leqslant 0$ in the Euler–Lagrange equations of any of the three models, we arrive at a contradiction. For example, for the model defined by ${}^gL^{(z)}_0$, we would get $a_{\rho\rho}(\rho_j^0)>0$ for any sufficiently large j because $f(\rho)\to 0$ as $\rho\to\infty$. This proves the lemma.

In summary, our existence theorem is proved.

Note added. After completing this work it came to our attention that some new[12] Maxwell[13,14] and Chern–Simons[15] gauged O(3) models subject to the *second* gauging prescription which exhibit both symmetry breaking and quantized magnetic flux have been proposed. The vortices of these models, however, are not stabilized by the usual degree of the map, but instead by a partial map which we have not discussed in the present work.

ACKNOWLEDGMENT

This work was carried out under project SC/96/602 of FORBAIRT, and was supported partially by the National Science Foundation under Grant No. DMS-9596041.

[1] A. Jaffe and C. H. Taubes, *Monopoles and Vortices* (Birkhäuser, Basel, 1980).
[2] D. H. Tchrakian and K. Arthur, Phys. Lett. B **352**, 327 (1995).
[3] B. J. Schroers, Phys. Lett. B **356**, 291 (1995).
[4] W. J. Zakrzewski, *Low Dimensional Sigma Models* (Hilger, Bristol, 1989)
[5] K. Arthur, D. H. Tchrakian, and Y. Yang, Phys. Rev. D **54**, 5245 (1996).
[6] B. M. A. G. Piette, D. H. Tchrakian, and W. J. Zakrzewski, Phys. Lett. B **339**, 95 (1994).
[7] J. Hong, Y. Kim, and P. Y. Pac, Phys. Rev. Lett. **64**, 2230 (1990); R. Jackiw and E. Weinberg, *ibid.* **64**, 2234 (1990).
[8] P. K. Ghosh and S. K. Ghosh, Phys. Lett. B **366**, 199 (1996).
[9] K. Kimm, K. Lee, and T. Lee, Phys. Rev. **53**, 4436 (1996).
[10] B. M. A. G. Piette, H. J. W. Müller-Kirsten, D. H. Tchrakian, and W. J. Zakrzewski, Phys. Lett. B **320**, 95 (1994).
[11] K. Arthur, G. Roche, D. H. Tchrakian, and Y. Yang, J. Math. Phys. **37**, 2569 (1996).
[12] B. J. Schroers, Nucl. Phys. B **475**, 440 (1996).
[13] Y. Yang, Phys. Rev. Lett. **80**, 26 (1998).
[14] P. Mukherjee, ''Magnetic vortices in a gauged O(3) sigma model with symmetry breaking self-interaction,'' hep-th/9705196.
[15] P. Mukherjee, Phys. Lett. B **403**, 70 (1997).

Highest weight irreducible representations of the quantum algebra $U_h(A_\infty)$

T. D. Palev[a),b)]
Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627

N. I. Stoilova[a),c)]
Department of Mathematics, University of Queensland, Brisbane Qld 4072, Australia

(Received 2 April 1998; accepted for publication 9 July 1998)

A class of highest-weight irreducible representations of the algebra $U_h(A_\infty)$, the quantum analog of the completion and central extension A_∞ of the Lie algebra gl_∞, is constructed. It is considerably larger than the class of representations known so far. Within each module a basis is introduced and the transformation relations of the basis under the action of the Chevalley generators are explicitly given. The verification of the quantum algebra relations is shown to reduce to a set of nontrivial q-number identities. All representations are restricted in the terminology of S. Levendorskii and Y. Soibelman [Commun. Math. Phys. **140**, 399–414 (1991)].
© *1998 American Institute of Physics.* [S0022-2488(98)02111-2]

I. INTRODUCTION

We construct a class of highest-weight irreducible representations (irreps) of the algebra $U_h(A_\infty)$,[1] the quantum analog of the completion and central extension A_∞ of the Lie algebra gl_∞.[2,3] Our interest in the subject stems from the observation that certain representations of $gl(n)$ (including $n=\infty$)[4] and of A_∞ (see *Example 2* in Ref. 5 and the references therein) are related to a new quantum statistics, the A-statistics. The latter, as it is clear now, is a particular case of the Haldane exclusion statistics,[6] a subject of considerable interest in condensed matter physics (see Sec. 4 in Ref. 7 for more discussions on the subject). It turns out that some of the representations of the deformed algebra $U_h(A_\infty)$ satisfy also the requirements of the Haldane statistics. More precisely, they lead to new solutions for the microscopic statistics of Karabali and Nair[8] directly in the case of infinitely many degrees of freedom. These results will be published elsewhere. We mention them here only in order to justify our motivation for the work we are going to present. One may expect certainly that similar to A_∞[9,10] the representations of $U_h(A_\infty)$ may prove useful also in other branches of physics and mathematics.

The quantum analogs of gl_∞ and A_∞ in the sense of Drinfeld,[11] namely $U_h(gl_\infty)$ and $U_h(A_\infty)$, were worked out by Levendorskii and Soibelman.[1] These authors have constructed a class of highest-weight irreducible representations, writing down explicit expressions for the transformations of the basis under the action of the algebra generators.

The $U_h(A_\infty)$-modules, which we study, are labeled by all possible sequences (see the end of the Introduction for the notation) $\{M\}\equiv\{M_i\}_{i\in\mathbf{Z}}\in\mathbf{C}[h]^\infty$, subject to the conditions:

(a) There exists $m\leqslant n\in\mathbf{Z}$, such that $M_m=M_{m-k}$ and $M_n=M_{n+k}$ for all $k\in\mathbf{N}$;
(b) $M_i-M_j\in\mathbf{Z}_+$ for all $i<j\in\mathbf{Z}$.

Representations, corresponding to two different sequences, $\{M^1\}\neq\{M^2\}$, are inequivalent. The $U_h(A_\infty)$-modules of Levendorskii and Soibelman[1] are labeled by all those sequences $\{M^{(s)}\}$, for which $s=m=n\in\mathbf{Z}$ and $M_i^{(s)}=1$, if $i<s$ and $M_i^{(s)}=0$ for $i\geqslant s$.

The transformation relations of the $U_h(A_\infty)$-modules are given in terms of a basis, which we called a C basis [see Ref. 11]. This basis was initially introduced for a description of the finite-

a)Permanent address: Institute for Nuclear Research and Nuclear Energy, 1784 Sofia, Bulgaria.
b)Electronic mail: tpalev@inrne.bas.bg
c)Electronic mail: stoilova@inrne.bas.bg

 © 1998 American Institute of Physics

signature representations of gl_∞ [12,13] and A_∞ .[14,15] We recall that the *C* basis is different from the known Gel'fand-Zetlin basis (GZ basis)[16] even for $\mathrm{gl}(n)$. The GZ-basis is introduced through a ''left to right'' and ''up to down'' embedding in the chain of subalgebras (see Ref. 13 for the notation)

$$\mathrm{gl}_0(1)\subset\mathrm{gl}_0(2)\subset\cdots\subset\mathrm{gl}_0(n-1)\subset\mathrm{gl}_0(n)\equiv\mathrm{gl}(n), \tag{1}$$

whereas the *C* basis corresponds to a ''central'' embedding

$$\mathrm{gl}_1\subset\mathrm{gl}_2\subset\cdots\subset\mathrm{gl}_{n-1}\subset\mathrm{gl}_n\equiv\mathrm{gl}(n). \tag{2}$$

In the finite-dimensional case the GZ basis and the C basis describe certainly one and the same representation. This is no more the case for gl_∞. The representations obtained through the GZ basis (GZ representations) are not equivalent to the *C* representations. These are different classes of representations. In particular the GZ representations are not highest-weight representations of gl_∞ (with respect to the natural order of the Cartan generators). This is the reason why the GZ representations of gl_∞ are not so interesting in the physical applications.

In the present paper we generalize to the quantum case the C representations. In Ref. 13 we have derived the transformations of the C basis from the results of Gel'fand and Zetlin.[16] This possibility was based on a map φ of gl_n onto $\mathrm{gl}_0(n)$, defined there. Its main property (see *Proposition 3* in Ref. 13) is that φ maps isomorphically the chain (2) onto the chain (1). We were unable to formulate an analog of *Proposition 3* in the quantum case. As a result we were unable to derive the transformation relations [see Eqs. (19)–(22)] from the results of Jimbo.[17] The very idea of Jimbo, however, to replace certain multiples in the transformation relations of the $\mathrm{gl}(n)$ modules by *q* numbers in order to obtain representations in the quantum case turned to work also for the transformation relations of the gl_∞ *modules*. This was certainly only a working hypothesis, because the matrix elements in the GZ basis were too different from those in the C basis. We knew a similar hypothesis worked in the case of the superalgebra $\mathrm{U}_q[\mathrm{gl}(n/m)]$,[18,19] but it did not work in other cases.[20] Therefore our proof of *Proposition 1* [namely of the transformation relations (19)–(22) of the C basis] is based on a direct verification of the Cartan and the Serre relations (3)–(5).

From our results it follows that each A_∞ module with a signature $\{M\}$ can be deformed to a $\mathrm{U}_h(A_\infty)$ module with the same signature. The class of the finite-signature representations of $\mathrm{U}_h(A_\infty)$ is, however, larger, which is due to the fact that the coordinates of the $\mathrm{U}_h(A_\infty)$ signatures take values in $\mathbf{C}[h]$. All representations we obtain are restricted in the sense of Ref. 1, Definition 4.1.

In Sec. II we construct a class of highest-weight irreps of the subalgebra $\mathrm{U}_h(a_\infty)$ of $\mathrm{U}_h(A_\infty)$. Some of these representations, namely the finite-signature representations (*Definition 2*), are extended to representations of $\mathrm{U}_h(A_\infty)$ in Sec. III.

Throughout the paper we use the notation (most of them standard):

N—all positive integers;
$\mathbf{Z}_+$—all non-negative integers;
Z—all integers;
Q—all rational numbers;
C—all complex numbers;
$\mathbf{C}[h]$—the ring of all polynomials in h over **C**;
$\mathbf{C}[[h]]$—the ring of all formal power series in h over **C**;
$X^n=\{(a_1,a_2,\ldots,a_n)|a_i\in X\}$, (including $n=\infty$);
$[a;b]=\{x|a\leqslant x\leqslant b,x\in\mathbf{Z}\}$, $(a;b)=\{x|a<x<b,x\in\mathbf{Z}\}$;
$\Gamma(\{M\})$—the *C*-basis of a module with a signature $\{M\}$;

$$\theta(i)=\begin{cases}1, & \text{for } i\geqslant 0,\\ 0, & \text{for } i<0;\end{cases}$$

$q=e^{h/2}\in\mathbf{C}[[h]]$;
$[x]=(q^x-q^{-x})/(q-q^{-1})\in\mathbf{C}[[h]]$;
$[x,y]_q=xy-qyx$.

II. REPRESENTATIONS OF THE ALGEBRA $U_h(a_\infty)$

The algebra $U_h(a_\infty)$ was defined in Ref. 1. The authors denote it as $U_h(g'(A_\infty))_f$. It is a Hopf algebra, which is a topologically free module over $\mathbf{C}[[h]]$ (complete in h-adic topology), with generators $\{e_i, f_i, h_i, c\}_{i\in\mathbf{Z}}$, and

(1) Cartan relations:

$$[c,\ a]=0, \quad a\in\{h_i, e_i, f_i\}_{i\in\mathbf{Z}}, \tag{3a}$$

$$[h_i,\ h_j]=0, \tag{3b}$$

$$[h_i,\ e_j]-(\delta_{ij}-\delta_{i,j+1})e_j=0, \tag{3c}$$

$$[h_i,\ f_j]+(\delta_{ij}-\delta_{i,j+1})f_j=0, \tag{3d}$$

$$[e_i,\ f_i]-[h_i-h_{i+1}+(\theta(-i)-\theta(-i-1))c]=0, \tag{3e}$$

$$[e_i,\ f_j]=0, \quad i\neq j; \tag{3f}$$

(2) *e*-Serre relations:

$$[e_i,\ e_j]=0, \quad \text{if } |i-j|\neq 1, \tag{4a}$$

$$e_i^2 e_{i+1}-(q+q^{-1})e_i e_{i+1} e_i+e_{i+1}e_i^2=0, \tag{4b}$$

$$e_{i+1}^2 e_i-(q+q^{-1})e_{i+1} e_i e_{i+1}+e_i e_{i+1}^2=0; \tag{4c}$$

(3) *f*-Serre relations:

$$[f_i,\ f_j]=0, \quad \text{if } |i-j|\neq 1. \tag{5a}$$

$$f_i^2 f_{i+1}-(q+q^{-1})f_i f_{i+1} f_i+f_{i+1}f_i^2=0, \tag{5b}$$

$$f_{i+1}^2 f_i-(q+q^{-1})f_{i+1} f_i f_{i+1}+f_i f_{i+1}^2=0. \tag{5c}$$

We do not write the other Hopf algebra maps (Δ,ε,S),[1] since we will not use them. They are certainly also a part of the definition of $U_h(a_\infty)$.

Replacing throughout in the above relations $\{e_i, f_i, h_i, c\}_{i\in\mathbf{Z}}$ with $\{E_i, F_i, H_i, 0\}_{i\in\mathbf{Z}}$, one obtains the definition of $U_h(gl_\infty)$.

In terms of an equivalent set of generating elements $\{\hat{e}_i, \hat{f}_i, h_i, c\}_{i\in\mathbf{Z}}$, with

$$\hat{e}_i=e_i q^{(h_{i+1}-h_i)/2}, \quad \hat{f}_i=f_i q^{(h_i-h_{i+1})/2},$$

one writes the quantum analogs of the Weyl generators $\{e_{ij}\}_{(i,j)\in\mathbf{Z}^2}$:

$$e_{ii}=h_i, \quad e_{i,i+1}=\hat{e}_i, \quad e_{i+1,i}=\hat{f}_i,$$

$$e_{ij}=[\hat{e}_i,[\hat{e}_{i+1},[\ldots,[\hat{e}_{j-2},\hat{e}_{j-1}]_q\ldots]_q]_q]_q, \quad i+1<j,$$

$$e_{ji}=[\hat{f}_i,[\hat{f}_{i+1},[\ldots,[\hat{f}_{j-2},\hat{f}_{j-1}]_q\ldots]_q]_q]_q, \quad i+1<j.$$

The ''commutation relations'' between these generators follow from (3)–(5) and are given in Ref. 1. The relevance of the generators $\{e_{ij}\}_{(i,j)\in\mathbf{Z}^2}$ stems from the observation that the set of ordered monomials (see Ref. 1 for the ordering)

$$c^l \prod_{(i,j)\in\mathbf{Z}^2} e_{ij}^{n_{ij}} \tag{6}$$

with finitely many nonzero exponents $n_{ij}\in\mathbf{Z}_+$, $l\in\mathbf{Z}_+$ forms a (topological) basis in $U_h(a_\infty)$.

Set $H=\oplus_i \mathbf{C}h_i$. Define linear functionals $\varepsilon_i: H\to\mathbf{C}$ by $\varepsilon_i(h_j)=\delta_{ij}$ and set $\alpha_i=\varepsilon_i-\varepsilon_{i+1}$, $Q'_+=\oplus_i \mathbf{Z}_+\alpha_i$. Denote by $\mathrm{U}_h(n_+)$ [respectively $\mathrm{U}_h(n_-)$] the unital subalgebra in $\mathrm{U}_h(a_\infty)$ generated by $\{e_i\}_{i\in\mathbf{Z}}$ (respectively $\{f_i\}_{i\in\mathbf{Z}}$). Then

$$\mathrm{U}_h(n_\pm)=\bigoplus_{\alpha\in Q'_+}\mathrm{U}_h(n_\pm)_{\pm\alpha}, \tag{7}$$

where

$$\mathrm{U}_h(n_\pm)_{\pm\alpha}=\{x\in\mathrm{U}_h(n_\pm)\,|\,[h',x]=\pm\alpha(h')x,\quad \forall h'\in H\} \tag{8}$$

for $\alpha\neq 0$, and $\mathrm{U}_h(n_\pm)_0=\mathbf{C}[[h]]$. Any element $u\in\mathrm{U}_h(a_\infty)$ can be represented as

$$u=\sum_{k=0}^{\infty}h^k\sum_{l=0}^{l(k)}c^l\sum_{\alpha,\beta\in Q'_+}\sum_{\gamma(k,l)\in\mathbf{Z}_+^\infty}\sum_{t=1}^{t(\alpha,\beta)}\mathscr{F}_{\alpha,k,l,t}\prod_{i\in\mathbf{Z}}h_i^{\gamma(k,l)_i}\mathscr{E}_{\beta,k,l,t},\quad \text{finite sums over } \alpha,\beta,\gamma, \tag{9}$$

where $\mathscr{F}_{\alpha,k,l,t}\in\mathrm{U}_h(n_-)_{-\alpha}$, $\mathscr{E}_{\beta,k,l,t}\in\mathrm{U}_h(n_+)_{+\beta}$ and finitely many exponents $\gamma(k,l)_i$ are different from zero. The words ''finite sums over α, β, γ'' have been added in order to indicate that for a fixed k only finitely many summands in (9) are different from zero.

The set $\hat{\mathrm{U}}_h(a_\infty)$, consisting of all $\mathbf{C}[[h]]$-polynomials of the Chevalley generators $\{e_i,f_i,h_i,c\}_{i\in\mathbf{Z}}$, is dense in $\mathrm{U}_h(a_\infty)$ with a basis (6). In particular $\mathscr{F}_{\alpha,k,l,t}$, $\mathscr{E}_{\beta,k,l,t}$ and $\Pi_{i\in\mathbf{Z}}h_i^{\gamma(k,l)_i}$ are in $\hat{\mathrm{U}}_h(a_\infty)$. Then according to (9) any element $u\in\mathrm{U}_h(a_\infty)$ is of the form

$$u=\sum_{k=0}^{\infty}u_kh^k,\quad u_k=\sum_{l=0}^{l(k)}c^l\sum_{\alpha,\beta\in Q'_+}\sum_{\gamma(k,l)\in\mathbf{Z}_+^\infty}\sum_{t=1}^{t(\alpha,\beta)}\mathscr{F}_{\alpha,k,l,t}\prod_{i\in\mathbf{Z}}h_i^{\gamma(k,l)_i}\mathscr{E}_{\beta,k,l,t}\in\hat{\mathrm{U}}_h(a_\infty). \tag{10}$$

We pass to construct a class of highest-weight irreps of $\mathrm{U}_h(a_\infty)$. In the following Propositions 1 and 2 we define the $\mathrm{U}_h(a_\infty)$-modules $V(\{M\};\xi_0,\xi_1)$, each one labeled by ξ_0, $\xi_1\in\mathbf{C}[h]$ and by a sequence $\{M\}\equiv\{M_i\}_{i\in\mathbf{Z}}\in\mathbf{C}[h]^\infty$ such that

$$M_i-M_j\in\mathbf{Z}_+,\quad \forall i<j\in\mathbf{Z}.$$

The basis $\Gamma(\{M\})$ in $V(\{M\};\xi_0,\xi_1)$, called a central basis (C basis), is independent of ξ_0,ξ_1. It is formally the same as the one introduced in Refs. 14 and 15 for a description of representations of a_∞. $\Gamma(\{M\})$ consists of all C patterns

$$|M)\equiv\begin{bmatrix}
.., & M_{1-\theta-k}, & ..., & M_{-1}, & M_0, & M_1, & ..., & M_{k-1},... \\
.., & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & M_{1-\theta-k,2k+\theta-1}, & ..., & M_{-1,2k+\theta-1}, & M_{0,2k+\theta-1}, & M_{1,2k+\theta-1}, & ..., & M_{k-1,2k+\theta-1} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & & & M_{-1,3}, & M_{03}, & M_{13} & & \\
 & & & M_{-1,2}, & M_{02} & & & \\
 & & & & M_{01} & & &
\end{bmatrix}, \tag{11}$$

where $k\in\mathbf{N}$, $\theta=0,1$. Each such pattern is an ordered collection of formal polynomials in h,

$$M_{i,2k+\theta-1}\in\mathbf{C}[h],\quad \forall k\in\mathbf{N},\quad \theta=0,1,\quad i\in[-\theta-k+1;k-1], \tag{12}$$

which satisfy the conditions:

(i) there exists a positive integer $N_{|M)}>1$, depending on $|M)$, such that

$$M_{i,2k+\theta-1}=M_i,\quad \forall 2k+\theta-1\geqslant N_{|M)},\quad \theta=0,1,\quad i\in[1-\theta-k;k-1]; \tag{13a}$$

(ii) for each $k\in\mathbf{N}$, $\theta=0$, 1, and $i\in[1-\theta-k;k-1]$,

$$M_{i+\theta-1,2k+\theta}-M_{i,2k+\theta-1}\in\mathbf{Z}_+,\quad M_{i,2k+\theta-1}-M_{i+\theta,2k+\theta}\in\mathbf{Z}_+. \tag{13b}$$

Denote by $\hat{V}(\{M\};\xi_0,\xi_1)$ the free $\mathbf{C}[[h]]$-module with generators $\Gamma(\{M\})$ and let $V(\{M\};\xi_0,\xi_1)$ be its completion in the h-adic topology. $V(\{M\};\xi_0,\xi_1)$ is a topologically free $\mathbf{C}[[h]]$ module with a (topological) basis $\Gamma(\{M\})$ and $\hat{V}(\{M\};\xi_0,\xi_1)$ is dense in it (in the h-adic topology). $V(\{M\};\xi_0,\xi_1)$ consists of all formal power series in h with coefficients in $\hat{V}(\{M\};\xi_0,\xi_1)$:

$$v=\sum_{i=0}^{\infty} v_i h^i, \quad v_0,v_1,v_2,\ldots\in\hat{V}(\{M\};\xi_0,\xi_1). \tag{14}$$

If a is a $\mathbf{C}[[h]]$-linear map in $\hat{V}(\{M\};\xi_0,\xi_1)$, $a\in\mathrm{End}\,\hat{V}(\{M\};\xi_0,\xi_1)$, extend it to a continuous linear map on $V(\{M\};\xi_0,\xi_1)$ setting

$$av=\sum_{i=0}^{\infty}(av_i)h^i. \tag{15}$$

Therefore the transformation of $V(\{M\};\xi_0,\xi_1)$ under the action of a is completely defined, if a is defined on $\Gamma(\{M\})$.

We proceed to turn $V(\{M\};\xi_0,\xi_1)$ into a $\mathrm{U}_h(a_\infty)$ module. Denote by

$$|M)_{\pm k_1\{j_1,p_1\},\pm k_2\{j_2,p_2\},\ldots}\equiv|M)^{\pm k_1\{j_1,p_1\},\ldots}_{\pm k_2\{j_2,p_2\},\ldots}\equiv|M)^{\pm k_1\{j_1,p_1\},\pm k_2\{j_2,p_2\},\ldots}, \quad k_1,k_2,\ldots=1,2,..., \tag{16}$$

the pattern obtained from the C pattern $|M)$ in (11) after the replacements

$$M_{j_1,p_1}\to M_{j_1,p_1}\pm k_1, \quad M_{j_2,p_2}\to M_{j_2,p_2}\pm k_2,\ldots,$$

correspondingly, and let

$$S(j,l;\nu)=\begin{cases}(-1)^\nu & \text{for } j=l,\\ 1 & \text{for } j<l,\\ -1 & \text{for } j>l\end{cases} \qquad \theta(i)=\begin{cases}1 & \text{for } i\geqslant 0\\ 0 & \text{for } i<0,\end{cases} \quad L_{ij}=M_{ij}-i. \tag{17}$$

Set, moreover,

$$e_i^0=f_i, \quad e_i^1=e_i \quad i\in\mathbf{Z}. \tag{18}$$

Let $\{\rho(e_i),\rho(f_i),\rho(h_i),\rho(c)\}_{i\in\mathbf{Z}}$ be a collection of $\mathbf{C}[[h]]$ endomorphisms of $V(\{M\};\xi_0,\xi_1)$, defined on any C pattern $|M)\in\Gamma(\{M\})$, as follows [see also (46), (47), (50) and (51)]:

$$\rho(e_{-1}^{1-\mu})|M)=([L_{-1,2}-L_{0,1}-\mu][L_{0,1}-L_{0,2}+\mu])^{1/2}|M)_{-(-1)^\mu\{0,1\}}, \quad \mu=0,1, \tag{19}$$

$$\begin{aligned}\rho(e^\mu_{(-1)^\nu i-1})|M)=&-\sum_{j=1-i-\nu}^{i-1}\sum_{l=-i}^{i+\nu-1}S(j,l;\nu)\\ &\times\Bigg(-\frac{\Pi_{k\neq l=-i}^{i+\nu-1}[L_{k,2i+\nu}-L_{j,2i+\nu-1}-(-1)^\nu\mu]\Pi_{k=1-i}^{i+\nu-2}[L_{k,2i+\nu-2}-L_{j,2i+\nu-1}-(-1)^\nu\mu]}{\Pi_{k\neq j=1-i-\nu}^{i-1}[L_{k,2i+\nu-1}-L_{j,2i+\nu-1}][L_{k,2i+\nu-1}-L_{j,2i+\nu-1}+(-1)^{\mu+\nu}]}\\ &\times\frac{\Pi_{k=-i-\nu}^{i}[L_{k,2i+\nu+1}-L_{l,2i+\nu}+(-1)^\nu(1-\mu)]\Pi_{k\neq j=1-i-\nu}^{i-1}[L_{k,2i+\nu-1}-L_{l,2i+\nu}+(-1)^\nu(1-\mu)]}{\Pi_{k\neq l=-i}^{i+\nu-1}[L_{k,2i+\nu}-L_{l,2i+\nu}][L_{k,2i+\nu}-L_{l,2i+\nu}+(-1)^{\mu+\nu}]}\Bigg)^{1/2}\\ &\times|M)^{-(-1)^{\mu+\nu}\{l,2i+\nu\}}_{-(-1)^{\mu+\nu}\{j,2i-1+\nu\}}, \quad i\in\mathbf{N}, \quad \mu,\nu=0,1,\end{aligned} \tag{20}$$

$$\rho(h_i)|M)=\Bigg(\sum_{j=-|i|}^{|i|+\theta(i)-1}M_{j,2|i|+\theta(i)}-\sum_{j=-|i|+1-\theta(i)}^{|i|-1}M_{j,2|i|+\theta(i)-1}+(\xi_1-\xi_0)\,\theta(-i)-\xi_1\Bigg)|M),$$

$$i\in\mathbf{Z}, \tag{21}$$

$$\rho(c)|M)=(\xi_0-\xi_1)|M). \tag{22}$$

Above and throughout $[x]=(q^x-q^{-x})/(q-q^{-1})\in\mathbf{C}[[h]]$. If a pattern from the right-hand side of (20) does not belong to $\Gamma(\{M\})$, i.e., it is not a C pattern, then the corresponding term has to be deleted. (The coefficients in front of all such patterns are undefined; they contain zero multiples in the denominators. Therefore an equivalent statement is that all terms with zeros in the denominators have to be removed.) With this convention all coefficients in front of the C patterns on the rhs of (19)–(22) are well defined as elements from $\mathbf{C}[[h]]$.

Proposition 1: The endomorphisms $\{\rho(e_i),\rho(f_i),\rho(h_i),\rho(c)\}_{i\in\mathbf{Z}}$ *satisfy Eqs. (3)–(5) with* $\rho(e_i)$, $\rho(f_i)$, $\rho(h_i)$, *and* $\rho(c)$ *substituted for* e_i, f_i, h_i, *and* c, *respectively.*

The proof is based on a direct verification of the relations (3)–(5). The most difficult to check is the Cartan relation (3e). In order to show that it holds, one has to prove as an intermediate step that the following identities hold ($k\in\mathbf{N}$):

$$\sum_{s=0}^{1}\sum_{j=1-k}^{k-1}\sum_{l=-k}^{k-1}(-1)^s\frac{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{j,2k-1}+s-1]\Pi_{i=1-k}^{k-2}[L_{i,2k-2}-L_{j,2k-1}+s-1]}{\Pi_{i\neq j=1-k}^{k-1}[L_{i,2k-1}-L_{j,2k-1}+s][L_{i,2k-1}-L_{j,2k-1}+s-1]}$$
$$\times\frac{\Pi_{i=-k}^{k}[L_{i,2k+1}-L_{l,2k}+s]\Pi_{i\neq j=1-k}^{k-1}[L_{i,2k-1}-L_{l,2k}+s]}{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{l,2k}+s][L_{i,2k}-L_{l,2k}+s-1]}$$
$$=\left[\sum_{j=-k+1}^{k-1}L_{j,2k-1}-\sum_{j=-k+1}^{k-2}L_{j,2k-2}-\sum_{j=-k}^{k}L_{j,2k+1}+\sum_{j=-k}^{k-1}L_{j,2k}-1\right], \tag{23a}$$

$$\sum_{s=0}^{1}\sum_{j=-k}^{k-1}\sum_{l=-k}^{k}(-1)^s\frac{\Pi_{i\neq l=-k}^{k}[L_{i,2k+1}-L_{j,2k}-s+1]\Pi_{i=1-k}^{k-1}[L_{i,2k-1}-L_{j,2k}-s+1]}{\Pi_{i\neq j=-k}^{k-1}[L_{i,2k}-L_{j,2k}-s][L_{i,2k}-L_{j,2k}-s+1]}$$
$$\times\frac{\Pi_{i=-k-1}^{k}[L_{i,2k+2}-L_{l,2k+1}-s]\Pi_{i\neq j=-k}^{k-1}[L_{i,2k}-L_{l,2k+1}-s]}{\Pi_{i\neq l=-k}^{k}[L_{i,2k+1}-L_{l,2k+1}-s][L_{i,2k+1}-L_{l,2k+1}-s+1]}$$
$$=\left[\sum_{j=-k-1}^{k}L_{j,2k+2}-\sum_{j=-k}^{k}L_{j,2k+1}-\sum_{j=-k}^{k-1}L_{j,2k}+\sum_{j=-k+1}^{k-1}L_{j,2k-1}-1\right], \tag{23b}$$

$$\sum_{s=0}^{1}\sum_{l=-k}^{k-1}(-1)^s\frac{\Pi_{i=-k}^{k}[L_{i,2k+1}-L_{l,2k}+s]\Pi_{i=1-k;i\neq j,m}^{k-1}[L_{i,2k-1}-L_{l,2k}+s]}{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{l,2k}+s][L_{i,2k}-L_{l,2k}+s-1]}=0, \tag{24a}$$

$$\sum_{s'=0}^{1}\sum_{l=-k}^{k}(-1)^{s'}\frac{\Pi_{i=-k-1}^{k}[L_{i,2k+2}-L_{l,2k+1}-s']\Pi_{i=-k;i\neq j,m}^{k-1}[L_{i,2k}-L_{l,2k+1}-s']}{\Pi_{i\neq l=-k}^{k}[L_{i,2k+1}-L_{l,2k+1}-s'][L_{i,2k+1}-L_{l,2k+1}-s'+1]}=0, \tag{24b}$$

$$\sum_{s'=0}^{1}\sum_{j=1-k}^{k-1}(-1)^{s'}\frac{\Pi_{r=1-k}^{k-2}[L_{r,2k-2}-L_{j,2k-1}-s']\Pi_{r=-k;r\neq l,q}^{k-1}[L_{r,2k}-L_{j,2k-1}-s']}{\Pi_{r\neq j=1-k}^{k-1}[L_{r,2k-1}-L_{j,2k-1}-s'][L_{r,2k-1}-L_{j,2k-1}-s'+1]}=0, \tag{24c}$$

$$\sum_{s=0}^{1}\sum_{j=-k}^{k-1}(-1)^s\frac{\Pi_{i=-k;i\neq l,q}^{k}[L_{i,2k+1}-L_{j,2k}+s]\Pi_{i=1-k}^{k-1}[L_{i,2k-1}-L_{j,2k}+s]}{\Pi_{i\neq j=-k}^{k-1}[L_{i,2k}-L_{j,2k}+s][L_{i,2k}-L_{j,2k}+s-1]}=0. \tag{24d}$$

The identities (24) are relevant also for the proof of the Serre relations (4a) and (5a). Another set of simpler identities, namely,

$$([a-b-1][c-b-1]-[2][a-b]][c-b-1]+[a-b][c-b])\left(\frac{[a-d][c-e-1]}{[d-e-1][c-a-1]}\right.$$

$$\left.+\frac{[c-d-1][a-e]}{[d-e+1][c-a-1]}\right)+\left(\frac{[a-e-1][c-d]}{[d-e-1][c-a+1]}+\frac{[a-d-1][c-e]}{[d-e+1][c-a+1]}\right)([a-b-1]$$

$$\times[c-b-1]-[2][a-b-1][c-b]+[a-b][c-b])=0, \tag{25}$$

$$\frac{[a-1][b-1]-[2][a][b-1]+[a][b]}{[a-b+1]}+\frac{[a-1][b-1]-[2][a-1][b]+[a][b]}{[a-b-1]}=0, \tag{26}$$

$$[a-1]-[2][a]+[a+1]=0. \tag{27}$$

are used in the proof of the Serre relations (4b), (4c), (5b), and (5c).

The proof is lengthy. The main steps of it are outlined in the Appendix. There we prove also the identities (23) and (24), which, we believe, are of independent interest. The approach is similar to the one used in Ref. 19, where a method for proving such q identities was first considered.

Remark: We may have constructed the endomorphisms $\{\rho(e_i),\rho(f_i),\rho(h_i),\rho(c)\}_{i\in\mathbf{Z}}$ from the results on the representations of $\mathrm{U}_h(gl_\infty)$, announced in Ref. 21, and, more precisely, from the endomorphisms $\{\rho(E_i),\rho(F_i),\rho(H_i)\}_{i\in\mathbf{Z}}$ of the $\mathbf{C}[[h]]$ module $V(\{M\})$, which satisfy the Cartan and the Serre relations for $\mathrm{U}_h(gl_\infty)$ [Ref. 21, Eqs. (16)–(18)]. This possibility is based on the observation that the $C[[h]]$ linear map φ, defined on the generators as

$$\begin{gathered}\varphi(E_i)=e_i, \quad \varphi(F_i)=f_i, \quad \varphi(c)=c,\\ \varphi(H_i)=h_i+(\theta(-i)+\alpha)c, \quad \alpha\in\mathbf{C}[h],\end{gathered} \tag{28}$$

and extended by associativity is an (algebra) isomorphism of $\mathrm{U}_h(gl_\infty)\oplus\mathbf{C}[[h]]c$ onto $\mathrm{U}_h(a_\infty)$. Then the endomorphisms $\{\rho(e_i),\rho(f_i),\rho(h_i),\rho(c)\}_{i\in\mathbf{Z}}$ defined according to (28) [with $\alpha(\xi_0-\xi_1)=\xi_1$] as

$$\begin{gathered}\rho(e_i)=\rho(E_i), \quad \rho(f_i)=\rho(F_i), \quad \rho(c)=\xi_0-\xi_1,\\ \rho(h_i)=\rho(H_i)-((\xi_0-\xi_1)\theta(-i)+\xi_1),\end{gathered} \tag{29}$$

lead to the transformation relations (19)–(22). Here we give a direct proof of *Proposition 1*, since the corresponding *Proposition 1* in Ref. 21 was only stated. Its proof would have been based again on the identities (23)–(27).

Consider ρ as a $\mathbf{C}[[h]]$ linear operator from $\mathrm{U}_h(a_\infty)$ into End $V(\{M\};\xi_0,\xi_1)$. So far ρ is defined only on the Chevalley generators $\{e_i,f_i,h_i,c\}_{i\in\mathbf{Z}}$. Extend the domain of its definition on $\hat{\mathrm{U}}_h(a_\infty)$: if ρ has already been defined on $a,b\in\hat{\mathrm{U}}_h(a_\infty)$, then set

$$\rho(\alpha a+\beta b)=\alpha\rho(a)+\beta\rho(b), \quad \rho(ab)=\rho(a)\rho(b), \quad a,b\in\hat{\mathrm{U}}_h(a_\infty), \quad \alpha,\beta\in\mathbf{C}[[h]]. \tag{30}$$

As we know from (10), any element $u\in\mathrm{U}_h(a_\infty)$ can be represented as a sum $u=\Sigma_{i=0}^{\infty}u_ih^i$, $u_i\in\hat{\mathrm{U}}_h(a_\infty)$. Then for an arbitrary $v\in V(\{M\};\xi_0,\xi_1)$, writing it as in (14), we have

$$\left(\sum_{i=0}^{\infty}\rho(u_i)h^i\right)v=\left(\sum_{i=0}^{\infty}\rho(u_i)h^i\right)\left(\sum_{j=0}^{\infty}v_jh^j\right)=\sum_{n=0}^{\infty}\left(\sum_{m=0}^{n}\rho(u_{n-m})v_m\right)h^n\in V(\{M\};\xi_0,\xi_1), \tag{31}$$

since

$$\sum_{m=0}^{n}\rho(u_{n-m})v_m\in\hat{V}(\{M\};\xi_0,\xi_1).$$

Using (31), we extend ρ on $U_h(a_\infty)$:

$$\rho(u)=\sum_{i=0}^{\infty}\rho(u_i)h^i\in \text{End } V(\{M\};\xi_0,\xi_1) \quad \forall u\in \mathrm{U}_h(a_\infty). \tag{32}$$

Hence ρ is a well-defined map from $\mathrm{U}_h(a_\infty)$ into End $V(\{M\};\xi_0,\xi_1)$,

$$\rho:\mathrm{U}_h(a_\infty)\to \text{End } V(\{M\};\xi_0,\xi_1). \tag{33}$$

Proposition 2: The map (33), acting on the C basis according to Eqs. (19)–(22), defines a highest-weight irreducible representation of $\mathrm{U}_h(a_\infty)$ *in* $V(\{M\};\xi_0,\xi_1)$).

Proof: According to *Proposition 1*, (30), and (32), ρ is a $\mathbf{C}[[h]]$ homomorphism of $\mathrm{U}_h(a_\infty)$ in End $V(\{M\};\xi_0,\xi_1)$. It is continuous in the h-adic topology. Indeed, let $u\in \mathrm{U}_h(a_\infty)$. Then any neighborhood $W(\rho(u))$ of $\rho(u)$ contains a basic neighborhood $\rho(u)+h^n$ End $V(\{M\};\xi_0,\xi_1)\subset W(\rho(u))$. Evidently

$$\rho(u+h^n\mathrm{U}_h(a_\infty))\subset\rho(u)+h^n \text{ End } V(\{M\};\xi_0,\xi_1)\subset W(\rho(u)) \tag{34}$$

and therefore ρ is continuous in u for any $u\in \mathrm{U}_h(a_\infty)$. Hence $V(\{M\};\xi_0,\xi_1)$ is a $\mathrm{U}_h(a_\infty)$ module. It is a highest-weight module with respect to the ''Borel'' subalgebra $\mathrm{U}_h(n_+)$. The highest-weight vector $|\hat{M})$, which by definition satisfies the condition $\rho(\mathrm{U}_h(n_+))|\hat{M})=0$ and is an eigenvector of $\rho(H)$, corresponds to the one from (11) with

$$\hat{M}_{i,2k+\theta-1}=M_i, \quad \forall k\in\mathbf{N}, \quad \theta=0,1, \quad i\in[-\theta-k+1;k-1]. \tag{35}$$

The irreducibility of the $\mathrm{U}_h(a_\infty)$ module $V(\{M\};\xi_0,\xi_1)$ follows from the statement that $V(\{M\};\xi_0,\xi_1)$ is an irreducible module also of the non-deformed algebra $\mathrm{U}(A_\infty)$.[15] Let x and y be any two nonzero vectors from $V(\{M\};\xi_0,\xi_1)$. Then there exists a polynomial P of the non-deformed generators such that

$$\rho(P)x=cy, \quad 0\neq c\in\mathbf{C}. \tag{36}$$

Let $\hat{P}$ be the same polynomial, but of the deformed generators. From the non-deformed transformation relations {see Ref. 15, or simply replace the ''quantum'' brackets [ ] with ordinary brackets throughout in Eqs. (19)–(22)} one concludes that $\rho(\hat{P})x=\hat{c}y$, where

$$\hat{c}=c+d, \quad d\in h\mathbf{C}[[h]]. \tag{37}$$

Since $c\neq 0$, then also $\hat{c}\neq 0$. Therefore the representation of $\mathrm{U}_h(a_\infty)$ in $V(\{M\},\xi_0,\xi_1)$ is irreducible.

III. REPRESENTATIONS OF THE ALGEBRA $\mathrm{U}_h(A_\infty)$

In this section we show that some of the $\mathrm{U}_h(a_\infty)$ modules $V(\{M\};\xi_0,\xi_1)$ can be turned into irreducible highest-weight $\mathrm{U}_h(A_\infty)$ modules. First, following Ref. 1, we recall the definition of the quantum algebra $\mathrm{U}_h(A_\infty)$ [denoted by the authors as $\mathrm{U}_h(g'(A_\infty))$] only within the algebra sector (for the other Hopf algebra maps see Ref. 1). $\mathrm{U}_h(A_\infty)$ consists of all elements

$$u=\sum_{k=0}^{\infty}h^k\sum_{l=0}^{l(k)}c^l\sum_{\alpha,\beta\in Q'_+}\sum_{\gamma(k,l)\in\mathbf{Z}_+^\infty}\sum_{t=1}^{t(\alpha,\beta)}\mathscr{F}_{\alpha,k,l,t}\prod_{i\in\mathbf{Z}}h_i^{\gamma(k,l)_i}\mathscr{E}_{\beta,k,l,t}, \quad \text{infinite sums over } \alpha,\beta,\gamma, \tag{38}$$

however certain conditions on the pairs (α,γ) corresponding to the nonzero summands are imposed. In order to state them, set for $\alpha=\Sigma_i m_i\alpha_i\in\mathbf{Q}'_+$, $\gamma=\{\gamma_i\}_{i\in\mathbf{Z}}\in\mathbf{Z}_+^\infty$,

$$S(\alpha)=\{i|m_i\neq 0\}, \quad S(\gamma)=\{i|\gamma_i\neq 0\}, \quad S(\alpha,\gamma)=S(\alpha)\cup S(\gamma). \tag{39}$$

Connecting i and j with a line, if $|i-j|=1$, one can view $S(\alpha,\gamma)$ as a graph. Denote by $\mathscr{T}(\alpha,\gamma)$ the collection of its connected components. For any $u=\Sigma_{k=0}^{\infty}h^k u_k$ as given in (38) consider the series

$$u_k=\sum_{l=0}^{l(k)} c^l \sum_{\alpha,\beta\in Q'_+} \sum_{\gamma(k,l)\in \mathbf{Z}_+^\infty} \sum_{t=1}^{t(\alpha,\beta)} \mathscr{T}_{\alpha,k,l,t}\prod_{i\in\mathbf{Z}} h_i^{\gamma(k,l)_i}\mathscr{E}_{\beta,k,l,t}, \tag{40}$$

and let

$$\mathscr{F}(u,k)=\cup\mathscr{F}(\alpha,\gamma), \tag{41}$$

where the union is taken over all α and γ, which appear in the nonzero summands of (40). For $i\in\mathbf{Z}$ and $k\in\mathbf{Z}_+$ set

$$\text{Int}\,(u,k,i)=\{I\in\mathscr{F}(u,k)|\, i\in I\}. \tag{42}$$

For $r\in\mathbf{N}$ define the series $u(r)$, corresponding to u, by substituting 0 for all h_i ($i\leqslant -r$ or $i>r$) and for all e_i, f_i ($|i|\geqslant r$).

Definition 1:[1] *The series u of the form (38) is said to belong to* $\mathrm{U}_h(A_\infty)$, *provided*

(i) *for any* $k\in\mathbf{Z}_+$ *and any* $i\in\mathbf{Z}$ *the set* Int (u,k,i) *is finite*;

(ii) $u(r)\in\mathrm{U}_h(a_\infty)$ *for all* $r\in\mathbf{N}$.

We turn to construct a class of representations of $\mathrm{U}_h(A_\infty)$. The idea is to show that within certain $V(\{M\};\xi_0,\xi_1)$ the domain of the definition of the operator ρ: $\mathrm{U}_h(a_\infty)\to\text{End}\,V(\{M\};\xi_0,\xi_1)$ [see (33)] can be extended from $\mathrm{U}_h(a_\infty)$ to $\mathrm{U}_h(A_\infty)$, so that ρ is a representation of $\mathrm{U}_h(A_\infty)$ in $V(\{M\};\xi_0,\xi_1)$. The construction is a natural one. We assume that $\rho(u)$ is a continuous $\mathbf{C}[[h]]$ linear operator in $V(\{M\};\xi_0,\xi_1)$ for any $u\in\mathrm{U}_h(A_\infty)$. Then for any $v=\Sigma_{j=0}^\infty v_j h^j\in V(\{M\};\xi_0,\xi_1)$, $v_j\in\hat{V}(\{M\};\xi_0,\xi_1)$

$$\rho(u)v=\sum_{j=0}^{\infty}(\rho(u)v_j)h^j. \tag{43}$$

The above series is well defined, if $\rho(u)v_j\in V(\{M\};\xi_0,\xi_1)$. Since v_j is a (finite) $\mathbf{C}[[h]]$ linear combination of C basis vectors the latter holds, and hence $\rho(u)$ is defined as an operator in $V(\{M\};\xi_0,\xi_1)$, if $\rho(u)|M)\in V(\{M\};\xi_0,\xi_1)$ for any C pattern $|M)$. Thus, the first step is to clarify which are the $\mathrm{U}_h(a_\infty)$ modules $V(\{M\};\xi_0,\xi_1)$, for which $\rho(u)|M)\in V(\{M\};\xi_0,\xi_1)$ holds.

Definition 2: We say that $V(\{M\};\xi_0,\xi_1)$ *is of finite signature or, more precisely, of (m,n) signature and write* $\{M\}=\{M\}_{m,n}$ *if there exist integers* $m\leqslant n\in\mathbf{Z}$, *such that* $M_m=M_{m-k}$ *and* $M_n=M_{n+k}$ *for all* $k\in\mathbf{N}$.

We now proceed to show that each finite signature $\mathrm{U}_h(a_\infty)$ module $V(\{M\}_{m,n};M_m,M_n)$ can be turned into a $\mathrm{U}_h(A_\infty)$ module. To this end we prove first a few preliminary propositions.

Denote by $V(\{M\};\xi_0,\xi_1)_N$, $1<N\in\mathbf{N}$, the subspace of $V(\{M\};\xi_0,\xi_1)$, which is a $\mathbf{C}[[h]]$ linear envelope of all C basis vectors $|M)$, for which

$$M_{i,2k+\theta-1}=M_i \quad \forall 2k+\theta-1\geqslant N, \quad \theta=0,1, \quad i\in[1-\theta-k;k-1]. \tag{44}$$

Note that $V(\{M\};\xi_0,\xi_1)_N$ is a finite-dimensional subspace.

Proposition 3:

$$\rho(e_k)V(\{M\};\xi_0,\xi_1)_N=0, \quad \text{if } k\notin(-\tfrac{1}{2}(N+1);\tfrac{1}{2}(N-2)). \tag{45}$$

Proof: From (20) one obtains

$$\begin{aligned}\rho(e_k)|M)=&-\sum_{j=-k}^{k}\sum_{l=-k-1}^{k} S(j,l;0)\\ &\times\left|\frac{\Pi_{i\neq l=-k-1}^{k}[L_{i,2k+2}-L_{j,2k+1}-1]\Pi_{i=-k}^{k-1}[L_{i,2k}-L_{j,2k+1}-1]}{\Pi_{i\neq j=-k}^{k}[L_{i,2k+1}-L_{j,2k+1}][L_{i,2k+1}-L_{j,2k+1}-1]}\right|^{1/2}\\ &\times\left|\frac{\Pi_{i=-k-1}^{k+1}[L_{i,2k+3}-L_{l,2k+2}]\Pi_{i\neq j=-k}^{k}[L_{i,2k+1}-L_{l,2k+2}]}{\Pi_{i\neq l=-k-1}^{k}[L_{i,2k+2}-L_{l,2k+2}][L_{i,2k+2}-L_{l,2k+2}-1]}\right|^{1/2}|M)^{\{l,2k+2\}}_{\{j,2k+1\}}, \quad k\geqslant 0,\end{aligned} \tag{46}$$

$$\rho(e_{-k})|M)=-\sum_{j=-k+1}^{k-2}\sum_{l=-k+1}^{k-1}S(j,l;1)\times\left|\frac{\Pi_{i\neq l=-k+1}^{k-1}[L_{i,2k-1}-L_{j,2k-2}+1]\Pi_{i=2-k}^{k-2}[L_{i,2k-3}-L_{j,2k-2}+1]}{\Pi_{i\neq j=-k+1}^{k-2}[L_{i,2k-2}-L_{j,2k-2}][L_{i,2k-2}-L_{j,2k-2}+1]}\right|^{1/2}\times\left|\frac{\Pi_{i=-k}^{k-1}[L_{i,2k}-L_{l,2k-1}]\Pi_{i\neq j=-k+1}^{k-2}[L_{i,2k-2}-L_{l,2k-1}]}{\Pi_{i\neq l=-k+1}^{k-1}[L_{i,2k-1}-L_{l,2k-1}][L_{i,2k-1}-L_{l,2k-1}+1]}\right|^{1/2}|M)_{-\{j,2k-2\}}^{-\{l,2k-1\}},\quad k>1. \tag{47}$$

If $|M)\in V(\{M\};\xi_0,\xi_1)_N$ and

$$\begin{aligned}&\text{if } k\geqslant\tfrac{1}{2}(N-2),\quad \text{then } L_{i,2k+3}=L_{i,2k+2}=L_i=M_i-i,\\&\text{if } k\leqslant-\tfrac{1}{2}(N+1),\quad \text{then } L_{i,2k-1}=L_{i,2k}=L_i=M_i-i.\end{aligned} \tag{48}$$

In both cases the rhs of (46) and (47) contain zero multiples (L_l-L_l) and therefore vanish.

Proposition 4:

$$\rho(f_k)V(\{M\}_{m,n};M_m,M_n)_N=0,\quad \text{if } k\notin(\min\{-\tfrac{1}{2}(N+3),m-1\};\max\{\tfrac{1}{2}N,n\}). \tag{49}$$

Proof: Let $|M)\in V(\{M\}_{m,n};M_m,M_n)_N$. The relations that follow from (20) in this case are

$$\rho(f_k)|M)=-\sum_{j=-k}^{k}\sum_{l=-k-1}^{k}S(j,l;0)\left|\frac{\Pi_{i\neq l=-k-1}^{k}[L_{i,2k+2}-L_{j,2k+1}]\Pi_{i=-k}^{k-1}[L_{i,2k}-L_{j,2k+1}]}{\Pi_{i\neq j=-k}^{k}[L_{i,2k+1}-L_{j,2k+1}][L_{i,2k+1}-L_{j,2k+1}+1]}\right|^{1/2}\times\left|\frac{\Pi_{i=-k-1}^{k+1}[L_{i,2k+3}-L_{l,2k+2}+1]\Pi_{i\neq j=-k}^{k}[L_{i,2k+1}-L_{l,2k+2}+1]}{\Pi_{i\neq l=-k-1}^{k}[L_{i,2k+2}-L_{l,2k+2}][L_{i,2k+2}-L_{l,2k+2}+1]}\right|^{1/2}\times|M)_{-\{j,2k+1\}}^{-\{l,2k+2\}},\quad k\in\mathbf{Z}_+, \tag{50}$$

$$\rho(f_{-k})|M)=-\sum_{j=-k+1}^{k-2}\sum_{l=-k+1}^{k-1}S(j,l;1)\times\left|\frac{\Pi_{i\neq l=-k+1}^{k-1}[L_{i,2k-1}-L_{j,2k-2}]\Pi_{i=2-k}^{k-2}[L_{i,2k-3}-L_{j,2k-2}]}{\Pi_{i\neq j=-k+1}^{k-2}[L_{i,2k-2}-L_{j,2k-2}][L_{i,2k-2}-L_{j,2k-2}-1]}\right|^{1/2}\times\left|\frac{\Pi_{i=-k}^{k-1}[L_{i,2k}-L_{l,2k-1}-1]\Pi_{i\neq j=-k+1}^{k-2}[L_{i,2k-2}-L_{l,2k-1}-1]}{\Pi_{i\neq l=-k+1}^{k-1}[L_{i,2k-1}-L_{l,2k-1}][L_{i,2k-1}-L_{l,2k-1}-1]}\right|^{1/2}\times|M)_{\{j,2k-2\}}^{\{l,2k-1\}},\quad k>1. \tag{51}$$

If $k\geqslant N/2$, then $L_{i,2k+3}=L_{i,2k+2}=L_{i,2k+1}=L_{i,2k}=L_i=M_i-i$ and therefore (50) reads

$$\rho(f_k)|M)=-\sum_{j=-k}^{k}\sum_{l=-k-1}^{k}S(j,l;0)\left|\frac{\Pi_{i\neq l=-k-1}^{k}[L_i-L_j]\Pi_{i=-k}^{k-1}[L_i-L_j]}{\Pi_{i\neq j=-k}^{k}[L_i-L_j][L_i-L_j+1]}\right|^{1/2}\times\left|\frac{\Pi_{i=-k-1}^{k+1}[L_i-L_l+1]\Pi_{i\neq j=-k}^{k}[L_i-L_l+1]}{\Pi_{i\neq l=-k-1}^{k}[L_i-L_l][L_i-L_l+1]}\right|^{1/2}|M)_{-\{j,2k+1\}}^{-\{l,2k+2\}}.$$

Above only the term with $j=k$ survives:

$$\rho(f_k)|M\rangle = -\sum_{l=-k-1}^{k} S(k,l;0)\left|\frac{\Pi_{i\neq l=-k-1}^{k}[L_i-L_k]\Pi_{i=-k}^{k-1}[L_i-L_k]}{\Pi_{i=-k}^{k-1}[L_i-L_k][L_i-L_k+1]}\right|^{1/2}$$
$$\times\left|\frac{\Pi_{i=-k-1}^{k+1}[L_i-L_l+1]\Pi_{i=-k}^{k-1}[L_i-L_l+1]}{\Pi_{i\neq l=-k-1}^{k}[L_i-L_l][L_i-L_l+1]}\right|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{l,2k+2\}}$$
$$= -S(k,k;0)$$
$$\times\left|\frac{\Pi_{i=-k-1}^{k-1}[L_i-L_k]\Pi_{i=-k-1}^{k+1}[L_i-L_k+1]\Pi_{i=-k}^{k-1}[L_i-L_k+1]}{\Pi_{i=-k}^{k-1}[L_i-L_k+1]\Pi_{i=-k-1}^{k-1}[L_i-L_k][L_i-L_k+1]}\right|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{k,2k+2\}}$$
$$-\sum_{l=-k-1}^{k-1} S(k,l;0)$$
$$\times\left|\frac{\Pi_{i\neq l=-k-1}^{k}[L_i-L_k]\Pi_{i=-k-1}^{k+1}[L_i-L_l+1]\Pi_{i=-k}^{k-1}[L_i-L_l+1]}{\Pi_{i=-k}^{k-1}[L_i-L_k+1]\Pi_{i\neq l=-k-1}^{k}[L_i-L_l][L_i-L_l+1]}\right|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{l,2k+2\}}.$$

In the last term $l\neq k$. Hence

$$\prod_{i\neq l=-k-1}^{k}[L_i-L_k]=[L_k-L_k]\prod_{i\neq l=-k-1}^{k-1}[L_i-L_k]=0$$

and therefore it vanishes. Then

$$\rho(f_k)|M\rangle = -\left|\frac{\Pi_{i=-k-1}^{k+1}[L_i-L_k+1]}{\Pi_{i=-k-1}^{k-1}[L_i-L_k+1]}\right|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{k,2k+2\}} = -|[L_{k+1}-L_k+1]|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{k,2k+2\}},$$

i.e.,

$$\rho(f_k)|M\rangle = -|[M_{k+1}-M_k]|^{1/2}|M\rangle_{-\{k,2k+1\}}^{-\{k,2k+2\}}. \tag{52}$$

If $k\geqslant n$, then $M_{k+1}=M_k$ and $\rho(f_k)|M\rangle=0$. In a similar way one derives from (51) that $\rho(f_k)|M\rangle=0$ if $k\leqslant\min\{-\frac{1}{2}(N+3),m-1\}$, which proves (49). □

Proposition 5:

$$\rho(h_k)V(\{M\}_{m,n};M_m,M_n)_N=0, \quad \text{if } k\notin(\min\{-\tfrac{1}{2}(N+1),m\};\max\{\tfrac{1}{2}N,n\}). \tag{53}$$

The proof follows easily from (21).

Set

$$r_N=\max\{\tfrac{1}{2}(N+3),1-m,n\}. \tag{54}$$

From the last three propositions one concludes the following.

Corollary 1: If $k\notin(-r_N;r_N)$, then

$$\rho(h_k)V(\{M\}_{m,n};M_m,M_n)_N=\rho(e_k)V(\{M\}_{m,n};M_m,M_n)_N=\rho(f_k)V(\{M\}_{m,n};M_m,M_n)_N=0. \tag{55}$$

Proposition 6: $\rho(\mathrm{U}_h(n_+)_\beta)V(\{M\};\xi_0,\xi_1)_N\subset V(\{M\};\xi_0,\xi_1)_N$ for any $N\in\mathbf{N}$. More precisely,

$$\text{if } S(\beta)\subset\left(-\frac{(N+1)}{2};\frac{(N-2)}{2}\right)\equiv I_N, \text{ then } \rho(U_h(n_+)_{+\beta})V(\{M\};\xi_0,\xi_1)_N\subset V(\{M\};\xi_0,\xi_1)_N. \tag{56}$$

$$\text{If } S(\beta)\not\subset I_N, \text{ then } \rho(\mathrm{U}_h(n_+)_\beta)V(\{M\};\xi_0,\xi_1)_N=0. \tag{57}$$

Proof: From (20) [or directly from (46) and (47)] one concludes that

$$\rho(e_j)|M\rangle \in V(\{M\};\xi_0,\xi_1)_N, \quad \forall j \in \left(-\frac{(N+1)}{2};\frac{(N-2)}{2}\right) \quad \text{and} \quad \forall |M\rangle \in V(\{M\};\xi_0,\xi_1)_N. \tag{58}$$

Hence (56) holds.

Assume $S(\beta) \not\subset I_N$ and let P_β be a monomial of the generators e_i, $i \in S(\beta)$. P_β can be represented as $P_\beta = Q' e_n Q$, where Q depends only on the generators e_j with $j \in S(\beta) \cap I_N$ and $n \notin I_N$. Then (58) yields that $\rho(Q)|M\rangle \in V(\{M\};\xi_0,\xi_1)_N$ and therefore (*Proposition 3*) $\rho(P_\beta)|M\rangle = \rho(Q')\rho(e_n)\rho(Q)|M\rangle = 0$. □

Proposition 7: Let u be any element from $\mathrm{U}_h(A_\infty)$*, represented as in (38). Then*

$$\sum_{k=0}^{\infty} h^k \sum_{l=0}^{l(k)} \rho(c)^l \sum_{\alpha,\beta \in Q'_+} \sum_{\gamma(k,l) \in \mathbf{Z}_+^\infty} \sum_{t=1}^{t(\alpha,\beta)} \rho(\mathscr{T}_{\alpha,k,l,t}) \prod_{i \in \mathbf{Z}} \rho(h_i)^{\gamma(k,l)_i} \rho(\mathscr{E}_{\beta,k,l,t})|M\rangle$$
$$\in V(\{M\}_{m,n};M_m,M_n) \tag{59}$$

for any $|M\rangle$ *from the basis* $\Gamma(\{M\})$ *of* $V(\{M\}_{m,n};M_m,M_n)$*. For a fixed k the number of the nonzero summands in (59) is finite.*

Proof: In order to prove the proposition it suffices to show that

$$\sum_{\alpha,\beta \in Q'_+} \sum_{\gamma(k,l) \in \mathbf{Z}_+^\infty} \sum_{t=1}^{t(\alpha,\beta)} \rho(\mathscr{T}_{\alpha,k,l,t}) \prod_{i \in \mathbf{Z}} \rho(h_i)^{\gamma(k,l)_i} \rho(\mathscr{E}_{\beta,k,l,t})|M\rangle \in \hat{V}(\{M\}_{m,n};M_m,M_n). \tag{60}$$

Assume that $|M\rangle \in V(\{M\}_{m,n};M_m,M_n)_N$ and let β_0 be the weight of $|M\rangle$. Since $\mathscr{E}_{\beta,k,l,t} \in U_h(n_+)_\beta$, $\rho(\mathscr{E}_{\beta,k,l,t})|M\rangle \in V(\{M\}_{m,n};M_m,M_n)_N$ *(Proposition 6).* Each $\rho(\mathscr{E}_{\beta,k,l,t})|M\rangle$ has a weight $\beta+\beta_0$ and the nonzero vectors $\rho(\mathscr{E}_{\beta,k,l,t})|M\rangle$, corresponding to different $\beta \in \mathbf{Q}'_+$, are linearly independent (over $\mathbf{C}[[h]]$). Since $V(\{M\}_{m,n};M_m,M_n)_N$ is a finite-dimensional subspace, $\rho(\mathscr{E}_{\beta,k,l,t})|M\rangle \neq 0$ only for a finite number of $\beta \in \mathbf{Q}'_+$. Hence the sum over β in (60) is finite. Setting $\rho(\mathscr{E}_{\beta,k,l,t})|M\rangle = v_{\beta,k,l,t} \in \hat{V}(\{M\}_{m,n};M_m,M_n)_N$, we obtain for the lhs of (60)

$$\sum_{\alpha,\beta \in Q'_+} \sum_{\gamma(k,l) \in \mathbf{Z}_+^\infty} \sum_{t=1}^{t(\alpha,\beta)} \rho(\mathscr{T}_{\alpha,k,l,t}) \prod_{i \in \mathbf{Z}} \rho(h_i)^{\gamma(k,l)_i} v_{\beta,k,l,t}. \tag{61}$$

We proceed to show that the sum over α and γ in (61) is finite, too. Without loss of generality we assume that every $\mathscr{T}_{\alpha,k,l,t} \Pi_{i \in \mathbf{Z}} h_i^{\gamma(k,l)i}$ is a monomial of $\{f_i,h_i\}_{i \in \mathbf{Z}}$. Consider any term from (61), corresponding to a particular pair (α,γ):

$$\rho(\mathscr{T}_{\alpha,k,l,t}) \prod_{i \in \mathbf{Z}_+} \rho(h_i)^{\gamma(k,l)_i} v_{\beta,k,l,t}. \tag{62}$$

Let $\mathscr{I}(\alpha,\gamma)$ be the connected components of $S(\alpha,\gamma)$:

$$\mathscr{I}(\alpha,\gamma) = \{I_{a_i,b_i} \equiv [a_i;b_i] | i = 1,2,...,n\}, \quad a_i \leqslant b_i \in \mathbf{Z}, \quad |a_i - b_j| > 1, \quad \text{if } i \neq j. \tag{63}$$

In (63) we do not distinguish between a connected component I_{a_i,b_i} (beginning in a_i and ending in b_i) and the corresponding finite integer interval $[a_i;b_i]$. Then

$$\mathscr{T}_{\alpha,k,l,t} \prod_{i \in \mathbf{Z}_+} (h_i)^{\gamma(k,l)_i} = \prod_{i=1}^{n} \mathscr{T}_i, \tag{64}$$

where $\mathscr{T}_i$ is a monomial of f_j,h_j, $j \in [a_i;b_i]$. From (3) and (5) it follows that the multiples $\mathscr{T}_i$ in (64) commute. If $[a_i;b_i] \subset (-\infty;-r_N]$ or $[a_i;b_i] \subset [r_N;\infty)$, then *Corollary 1* yields that $\rho(\mathscr{T}_i)v_{\beta,k,l,t} = 0$. Hence also $\rho(\mathscr{T}_{\alpha,k,l,t}) \Pi_{i \in \mathbf{Z}_+} \rho(h_i)^{\gamma(k,l)_i} v_{\beta,k,l,t} = 0$. Thus, the sum in (60) is over

such pairs (α,γ), for which all connected components of $S(\alpha,\gamma)$, namely the elements from $\mathscr{F}(\alpha,\gamma)$, have nonzero intersection with $(-r_N;r_N)$. There is only a finite number of pairs (α,γ) with this property, for which $\mathrm{Int}(u,k,i)$ is finite.

The conclusion is that the lhs of the series (60) contains a finite number of nonzero summands. Since the generators of $\mathrm{U}_h(A_\infty)$ [see (19)–(22)] transform $\hat{V}(\{M\}_{m,n};M_m,M_n)$ into itself, (60) holds. Hence (59) holds, too. □

Based on *Proposition 7,* we extend the domain of the operator ρ on $\mathrm{U}_h(A_\infty)$, setting for any $u\in \mathrm{U}_h(A_\infty)$ [see (38)]

$$\rho(u)=\sum_{k=0}^{\infty} h^k \sum_{l=0}^{l(k)} \rho(c)^l \sum_{\alpha,\beta\in Q'_+} \sum_{\gamma(k,l)\in \mathbf{Z}_+^\infty} \sum_{t=1}^{t(\alpha,\beta)} \rho(\mathscr{F}_{\alpha,k,l,t}) \prod_{i\in\mathbf{Z}} \rho(h_i)^{\gamma(k,l)_i} \rho(\mathscr{E}_{\beta,k,l,t})$$

$$\in \mathrm{End}\ V(\{M\}_{m,n};M_m,M_n). \tag{65}$$

The map ρ is a homomorphism of $\mathrm{U}_h(A_\infty)$ in $\mathrm{End}\ V(\{M\}_{m,n};M_m,M_n)$. It is continuous in the h-adic topology. Hence ρ defines a representation of $\mathrm{U}_h(A_\infty)$ in $V(\{M\}_{m,n};M_m,M_n)$, which is a highest-weight irreducible representation [since it is a highest weight irrep with respect to the subalgebra $U_h(a_\infty)$].

Let $v=\Sigma_{k\in\mathbf{Z}_+} h^k v_k$, $v_k\in\hat{V}(\{M\}_{m,n};M_m,M_n)$ be an arbitrary element from $V(\{M\}_{m,n};M_m,M_n)$. Since each v_k is a finite linear combination of C vectors, for any $k\in\mathbf{Z}_+$ there exists an integer $N_k>1$ such that $v_k\in V(\{M\}_{m,n};M_m,M_n)_{N_k}$. Then from (55) and (57) one concludes that the following properties hold:

$$(1)\quad \rho(\mathrm{U}_h(n_+)_\beta)v_k=0,\quad \text{if } S(\beta)\not\subset(-r_{N_k};r_{N_k}),$$

$$(2)\quad \rho(\mathrm{U}_h(n_-)_{-\alpha})v_k=0,\quad \text{if } S(\alpha)\subset(-\infty;-r_{N_k}] \text{ or } S(\alpha)\subset[r_{N_k};\infty),$$

$$(3)\quad \rho(h_i)v_k=0,\quad \text{if } |i|\geqslant r_{N_k}. \tag{66}$$

Therefore each finite-signature representation of $\mathrm{U}_h(A_\infty)$ in $V(\{M\}_{m,n};M_m,M_n)$ is a restricted representation (see *Definition 4.1* in Ref. 1).

Let us mention in conclusion that the $\mathrm{U}_h(a_\infty)$ modules $V(\{M\};\xi_0,\xi_1)$, which are not of a finite (m,n) signature and for which $\xi_0\neq M_m$, $\xi_1\neq M_n$, cannot be turned into $\mathrm{U}_h(A_\infty)$ modules. In order to see this consider the transformation of the highest-weight vector $|\hat{M})$ under the action of the operator $I=\Sigma_{i\in\mathbf{Z}}h_i\in\mathrm{U}_h(A_\infty)$. From (21) and (35) one obtains

$$\rho(I)|\hat{M})=\left(\sum_{i=-\infty}^{0}(M_i-\xi_0)+\sum_{i=1}^{\infty}(M_i-\xi_1)\right)|\hat{M}). \tag{67}$$

Only in the finite-signature modules $V(\{M\}_{m,n};M_m,M_n)$ is the rhs of (67) not divergent.

We were dealing with algebras and modules over $\mathbf{C}[[h]]$. The algebra $\mathrm{U}_h(A_\infty)$ is, however, well-defined also for h being a fixed complex number h_c, such that $h_c\notin i\pi\mathbf{Q}$, namely in the case $q=e^{h_c}$ is not a root of 1.[1] In that case the representations we have obtained remain highest-weight irreps of $\mathrm{U}_{h_c}(A_\infty)$ realized in infinite-dimensional complex linear spaces.

ACKNOWLEDGMENTS

We are grateful to Professor Randjbar-Daemi for the kind hospitality at the High Energy Section of ICTP. NIS is grateful to Professor M. D. Gould for the invitation to work in his group at the Department of Mathematics at the University of Queensland. TDP is thankful to Professor S. Okubo for the invitation to carry the research on the Fulbright Project in the Department of Physics and Astronomy, University of Rochester. This work was supported by the Australian Research Council, by the U.S.A. Fulbright Program, Grant No. 21857, and by the Grant Φ-416 of the Bulgarian Foundation for Scientific Research.

APPENDIX: PROOF OF PROPOSITION 1

Here we outline the main steps in the proof of Proposition 1.

(1) The equations

$$[\rho(c),\rho(a)]|M)=0, \quad a\in\{h_i,e_i,f_i\}_{i\in\mathbf{Z}}, \tag{A1}$$

$$[\rho(h_i),\rho(h_j)]|M)=0 \tag{A2}$$

are evident.

(2) The validity of

$$([\rho(h_i),\rho(e_j)]-(\delta_{ij}-\delta_{i,j+1})\rho(e_j))|M)=0, \tag{A3}$$

$$([\rho(h_i),\rho(f_j)]+(\delta_{ij}-\delta_{i,j+1})\rho(f_j))|M)=0 \tag{A4}$$

follows easily from (19)–(22).

(3) We proceed to show that

$$([\rho(e_i),\rho(f_i)]-[\rho(h_i)-\rho(h_{i+1})+(\theta(-i)-\theta(-i-1))\rho(c)])|M)=0, \quad i\in\mathbf{Z}. \tag{A5}$$

(3a) The case in (A5) with $i\in\mathbf{Z}_+$. Applying the transformation relations (20) and (21) and rearranging the terms in an appropriate way, we obtain

$$\begin{aligned}
&([\rho(e_{k-1}),\rho(f_{k-1})]-[\rho(h_{k-1})-\rho(h_k)+(\theta(-k+1)-\theta(-k))\rho(c)])|M)\\
&\quad=\mathscr{A}_{2k}|M)+\sum_{j=-k+1}^{k-1}\sum_{m=-k+1;m\neq j}^{k-1}\mathscr{B}_{2k,j,m}|M)^{\{m,2k-1\}}_{-\{j,2k-1\}}\\
&\qquad+\sum_{l=-k}^{k-1}\sum_{p=-k;p\neq l}^{k-1}\mathscr{C}_{2k-1,p,l}|M)^{\{p,2k\}}_{-\{l,2k\}}, \quad \forall k\in\mathbf{N},
\end{aligned} \tag{A6}$$

where

$$\begin{aligned}
\mathscr{A}_{2k}=&\sum_{s=0}^{1}\sum_{j=1-k}^{k-1}\sum_{l=-k}^{k-1}(-1)^s\frac{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{j,2k-1}+s-1]\Pi_{i=1-k}^{k-2}[L_{i,2k-2}-L_{j,2k-1}+s-1]}{\Pi_{i\neq j=1-k}^{k-1}[L_{i,2k-1}-L_{j,2k-1}+s][L_{i,2k-1}-L_{j,2k-1}+s-1]}\\
&\times\frac{\Pi_{i=-k}^{k}[L_{i,2k+1}-L_{l,2k}+s]\Pi_{i\neq j=1-k}^{k-1}[L_{i,2k-1}-L_{l,2k}+s]}{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{l,2k}+s][L_{i,2k}-L_{l,2k}+s-1]}\\
&-\left[\sum_{j=-k+1}^{k-1}L_{j,2k-1}-\sum_{j=-k+1}^{k-2}L_{j,2k-2}-\sum_{j=-k}^{k}L_{j,2k+1}+\sum_{j=-k}^{k-1}L_{j,2k}-1\right],
\end{aligned} \tag{A7}$$

$$\mathscr{B}_{2k,j,m}=-\sum_{s=0}^{1}\sum_{l=-k}^{k-1}(-1)^s\frac{\Pi_{i=-k}^{k}[L_{i,2k+1}-L_{l,2k}+s]\Pi_{i=1-k;i\neq j,m}^{k-1}[L_{i,2k-1}-L_{l,2k}+s]}{\Pi_{i\neq l=-k}^{k-1}[L_{i,2k}-L_{l,2k}+s][L_{i,2k}-L_{l,2k}+s-1]}, \tag{A8}$$

$$\mathscr{C}_{2k-1,p,l}=\sum_{s'=0}^{1}\sum_{j=1-k}^{k-1}(-1)^{s'}\frac{\Pi_{r=1-k}^{k-2}[L_{r,2k-2}-L_{j,2k-1}-s']\Pi_{r=-k;r\neq l,p}^{k-1}[L_{r,2k}-L_{j,2k-1}-s']}{\Pi_{r\neq j=1-k}^{k-1}[L_{r,2k-1}-L_{j,2k-1}-s'][L_{r,2k-1}-L_{j,2k-1}-s'+1]}. \tag{A9}$$

The rhs of (A6) does not contain all terms, which would appear immediately after applying the transformation relations (20) and (21). The terms, which cancel out in a straightforward way, have already been removed.

We proceed to show that $\mathscr{A}_{2k}=\mathscr{B}_{2k,j,m}=\mathscr{C}_{2k-1,p,l}=0$ or, which is the same, we prove the identities (23a), (24a), and (24c).

(1^0) We begin with $\mathscr{A}_{2k}=0$. Each such coefficient is a formal power series in h, $\mathscr{A}_{2k}=\Sigma_{i=0}^{\infty}a_{ki}h^i$. In order to prove that $\mathscr{A}_{2k}=0$ we have to show that $a_{ki}=0$ for any k and i. To this

end replace h in (A7) by a complex variable $h\in\mathbf{C}$, so that h takes values on any curve $\gamma\subset\mathbf{C}$ and $h\notin i\pi\mathbf{Q}$. Then the complex function $\mathscr{A}_{2k}$ is well defined for any $h\in\gamma$ and $\mathscr{A}_{2k}=\Sigma_{i=0}^{\infty}a_{ki}h^i$. Therefore, $a_{ki}=0$ if the function $\mathscr{A}_{2k}$ of $h\in\gamma$ vanishes. Since $h\in\mathbf{C}$ appears in $\mathscr{A}_{2k}$ only through $q=e^{h/2}$, it suffices to show that the function $\mathscr{A}_{2k}=0$ for q being a number, which is not a root of 1.

Let $q\in\mathbf{C}$ be not a root of unity. Setting

$$\begin{gathered}q^{2L_{i-k,2k-1}}=A_i,\quad i=1,...,2k-1,\\ q^{2L_{i-k-1,2k}}=B_i,\quad i=1,...,2k,\\ q^{2L_{i-k-1,2k+1}}=C_i,\quad i=1,...,2k+1,\\ q^{2L_{i-k,2k-2}}=D_i,\quad i=1,...,2k-2,\\ 2k=n,\end{gathered}\tag{A10}$$

one represents $\mathscr{A}_{2k}\equiv\mathscr{A}_n$ in the form

$$\begin{aligned}\mathscr{A}_n=&\sum_{j=1}^{n-1}\sum_{l=1}^{n}\frac{q\Pi_{i\neq l=1}^{n}(A_j-q^{-2}B_i)\Pi_{i=1}^{n-2}(A_j-q^{-2}D_i)\Pi_{i=1}^{n+1}(B_l-C_i)\Pi_{i\neq j=1}^{n-1}(B_l-A_i)}{A_jB_l\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^{-2}A_i)\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^{-2}B_i)}\\ &-\sum_{j=1}^{n-1}\sum_{l=1}^{n}\frac{q^{-1}\Pi_{i\neq l=1}^{n}(A_j-B_i)\Pi_{i=1}^{n-2}(A_j-D_i)\Pi_{i=1}^{n+1}(B_l-q^2C_i)\Pi_{i\neq j=1}^{n-1}(B_l-q^2A_i)}{A_jB_l\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^2A_i)\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^2B_i)}\\ &-(q-q^{-1})\left(1-q^2\frac{\Pi_{i=1}^{n-2}D_i\Pi_{i=1}^{n+1}C_i}{\Pi_{i=1}^{n-1}A_i\Pi_{i=1}^{n}B_i}\right),\end{aligned}\tag{A11}$$

which can be written also as

$$\begin{aligned}\mathscr{A}_n=&\sum_{j=1}^{n-1}\frac{q\Pi_{i=1}^{n}(A_j-q^{-2}B_i)\Pi_{i=1}^{n-2}(A_j-q^{-2}D_i)}{A_j\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^{-2}A_i)}F_{n,j}\\ &-\sum_{l=1}^{n}\frac{q^{-1}\Pi_{i=1}^{n+1}(B_l-q^2C_i)\Pi_{i=1}^{n-1}(B_l-q^2A_i)}{B_l\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^2B_i)}G_{n,l}-(q-q^{-1})\left(1-q^2\frac{\Pi_{i=1}^{n-2}D_i\Pi_{i=1}^{n+1}C_i}{\Pi_{i=1}^{n-1}A_i\Pi_{i=1}^{n}B_i}\right),\end{aligned}\tag{A12}$$

where

$$F_{n,j}=\sum_{l=1}^{n}\frac{\Pi_{i=1}^{n+1}(B_l-C_i)\Pi_{i\neq j=1}^{n-1}(B_l-A_i)}{(A_j-q^{-2}B_l)B_l\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^{-2}B_i)}\tag{A13}$$

and

$$G_{n,l}=\sum_{j=1}^{n-1}\frac{\Pi_{i\neq l=1}^{n}(A_j-B_i)\Pi_{i=1}^{n-2}(A_j-D_i)}{A_j(B_l-q^2A_j)\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^2A_i)}.\tag{A14}$$

Consider the complex functions

$$f_1(z)=\frac{\Pi_{i=1}^{n+1}(z-C_i)\Pi_{i\neq j=1}^{n-1}(z-A_i)}{(z-q^2A_j)\Pi_{i=1}^{n}(z-B_i)(z-q^{-2}B_i)},\quad f_2(z)=\frac{\Pi_{i\neq l=1}^{n}(z-B_i)\Pi_{i=1}^{n-2}(z-D_i)}{(z-q^{-2}B_l)\Pi_{i=1}^{n-1}(z-A_i)(z-q^2A_i)}.\tag{A15}$$

The function $f_1(z)$ [resp. $f_2(z)$] is analytic in the extended complex plane $\mathbf{C}\cup\infty$ except in its simple poles q^2A_j, $B_1,...,B_n$, $q^{-2}B_1,...,q^{-2}B_n$ (resp. $q^{-2}B_l$, $A_1,...,A_{n-1}$, $q^2A_1,...,q^2A_{n-1}$). Let C be a closed curve whose interior contains all poles of $f_1(z)$ [resp. $f_2(z)$]. The Residue

theorem of complex analysis implies that $\oint_{\mathbf{C}} f_j(z)dz = 2\pi i \Sigma \mathrm{Res}(f_j(z))$, $j=1,2$. On the other hand $f_1(z)$ [resp. $f_2(z)$] has no poles at infinity, Res $f_j(\infty)=0$. Therefore $\oint_{\mathbf{C}} f_j(z)dz=0$, $j=1,2$. Hence

$$\sum_{\text{all residues}} \mathrm{Res}(f_j(z))=0, \quad j=1,2. \tag{A16}$$

From (A16) we obtain

$$F_{n,j}=(q^2-1)\,\frac{\Pi_{i=1}^{n+1}(q^2A_j-C_i)\Pi_{i\neq j=1}^{n-1}(q^2A_j-A_i)}{\Pi_{i=1}^{n}(q^2A_j-B_i)(q^2A_j-q^{-2}B_i)} - \sum_{l=1}^{n} \frac{\Pi_{i=1}^{n+1}(B_l-q^2C_i)\Pi_{i\neq j=1}^{n-1}(B_l-q^2A_i)}{B_l(q^{-2}B_l-q^2A_j)\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^2B_i)}, \tag{A17}$$

$$G_{n,l}=(q^{-2}-1)\,\frac{\Pi_{i\neq l=1}^{n}(B_l-q^2B_i)\Pi_{i=1}^{n-2}(q^{-2}B_l-D_i)}{\Pi_{i=1}^{n-1}(B_l-q^2A_i)(q^{-2}B_l-q^2A_i)} - \sum_{j=1}^{n-1} \frac{\Pi_{i\neq l=1}^{n}(A_j-q^{-2}B_i)\Pi_{i=1}^{n-2}(A_j-q^{-2}D_i)}{A_j(q^2A_j-q^{-2}B_l)\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^{-2}A_i)}. \tag{A18}$$

Inserting (A17) and (A18) into the rhs of (A12) one ends with an expression for A_n, which contains only simple sums:

$$\frac{\mathscr{A}_n}{q-q^{-1}}=\sum_{j=1}^{n-1} \frac{\Pi_{i=1}^{n-2}(A_j-q^{-2}D_i)\Pi_{i=1}^{n+1}(A_j-q^{-2}C_i)}{A_j\Pi_{i\neq j=1}^{n-1}(A_j-A_i)\Pi_{i=1}^{n}(A_j-q^{-4}B_i)} + \sum_{l=1}^{n} \frac{\Pi_{i=1}^{n+1}(B_l-q^2C_i)\Pi_{i=1}^{n-2}(B_l-q^2D_i)}{B_l\Pi_{i\neq l=1}^{n}(B_l-B_i)\Pi_{i=1}^{n-1}(B_l-q^4A_i)} + q^2\,\frac{\Pi_{i=1}^{n-2}D_i\Pi_{i=1}^{n+1}C_i}{\Pi_{i=1}^{n-1}A_i\Pi_{i=1}^{n}B_i}-1. \tag{A19}$$

In order to show that the rhs of (A19) vanishes, consider the function

$$f(z)=\frac{\Pi_{i=1}^{n-2}(z-q^{-2}D_i)\Pi_{i=1}^{n+1}(z-q^{-2}C_i)}{z\Pi_{i=1}^{n-1}(z-A_i)\Pi_{i=1}^{n}(z-q^{-4}B_i)}. \tag{A20}$$

This function is a meromorphic function in the extended complex plane, having a simple pole also at infinity. Then the residue theorem, applied to $f(z)$, yields that the rhs of (A19) vanishes.

For further reference we write down the identity, which we have proved:

$$\sum_{j=1}^{n-1}\sum_{l=1}^{n} \frac{q\Pi_{i\neq l=1}^{n}(A_j-q^{-2}B_i)\Pi_{i=1}^{n-2}(A_j-q^{-2}D_i)\Pi_{i=1}^{n+1}(B_l-C_i)\Pi_{i\neq j=1}^{n-1}(B_l-A_i)}{A_jB_l\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^{-2}A_i)\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^{-2}B_i)} - \sum_{j=1}^{n-1}\sum_{l=1}^{n} \frac{q^{-1}\Pi_{i\neq l=1}^{n}(A_j-B_i)\Pi_{i=1}^{n-2}(A_j-D_i)\Pi_{i=1}^{n+1}(B_l-q^2C_i)\Pi_{i\neq j=1}^{n-1}(B_l-q^2A_i)}{A_jB_l\Pi_{i\neq j=1}^{n-1}(A_j-A_i)(A_j-q^2A_i)\Pi_{i\neq l=1}^{n}(B_l-B_i)(B_l-q^2B_i)} - (q-q^{-1})\left(1-q^2\,\frac{\Pi_{i=1}^{n-2}D_i\Pi_{i=1}^{n+1}C_i}{\Pi_{i=1}^{n-1}A_i\Pi_{i=1}^{n}B_i}\right)=0. \tag{A21}$$

In particular [see (A11)], $\mathscr{A}_n=\mathscr{A}_{2k}=0$.

(2^0) We pass to show that $\mathscr{B}_{2k,j,m}=0$ [see (A8)]. Without loss of generality we may write it as

$$\mathscr{B}_{2k,j,m}=-\sum_{s=0}^{1}\sum_{l=-k}^{k-1}(-1)^s\,\frac{\Pi_{r=-k}^{k}[L_{r,2k+1}-L_{l,2k}+s]\Pi_{r=-k}^{k-4}[L'_{r,2k-1}-L_{l,2k}+s]}{\Pi_{r\neq l=-k}^{k-1}[L_{r,2k}-L_{l,2k}+s][L_{r,2k}-L_{l,2k}+s-1]}. \tag{A22}$$

In order to obtain (A22) we have relabeled the indices of $L_{i,2k-1}$:

$$\{L_{i,2k-1}\}_{i\in\{1-k;k-1\},i\neq j\neq m}=\{L'_{i,2k-1}\}_{i\in\{-k;k-4\}}. \tag{A23}$$

Setting in (A22) $i=l+k+1$ and $t=r+k+1$ and introducing the notation

$$\begin{gathered}L_{t-k-1,2k}=a_t,\quad t=1,...,2k,\\ L_{t-k-1,2k+1}=b_t,\quad t=1,...,2k-1,\\ L_{k,2k+1}=c_{2k-1},\quad L_{k-1,2k+1}=c_{2k-2},\quad L'_{t-k-1,2k-1}=c_t,\quad t=1,...,2k-3,\\ 2k=n,\end{gathered} \tag{A24}$$

we rewrite $\mathscr{B}_{n,j,m}$ in the following form:

$$\mathscr{B}_{n,j,m}=-\sum_{s=0}^{1}\sum_{i=1}^{n}(-1)^s\frac{\Pi_{t=1}^{n-1}[a_i-b_t-s]\Pi_{t=1}^{n-1}[a_i-c_t-s]}{\Pi_{t\neq i=1}^{n}[a_i-a_t-s][a_i-a_t-s+1]}. \tag{A25}$$

In order to prove that $\mathscr{B}_{n,j,m}=0$, it suffices to show that

$$\sum_{s=0}^{1}\sum_{i=1}^{n}(-1)^s\frac{\Pi_{t=1}^{n-1}[a_i-b_t-s]\Pi_{t=1}^{n-1}[a_i-c_t-s]}{\Pi_{t\neq i=1}^{n}[a_i-a_t-s][a_i-a_t-s+1]}=0 \tag{A26}$$

for $q\in\mathbf{C}$ not being a root of 1. Assume this is the case and set $q^{2a_i}=A_i$, $q^{2b_i}=B_i$, $q^{2c_i}=C_i$. Then the lhs of (A26) becomes

$$\frac{\Pi_{t=1}^{n}A_t}{q^{n-1}\Pi_{t=1}^{n-1}B_t^{1/2}C_t^{1/2}}\sum_{i=1}^{n}\frac{1}{A_i}\left(\frac{\Pi_{t=1}^{n-1}(A_i-q^2B_t)(A_i-q^2C_t)}{\Pi_{t\neq i=1}^{n}(A_i-A_t)(A_i-q^2A_t)}-\frac{\Pi_{t=1}^{n-1}(A_i-B_t)(A_i-C_t)}{\Pi_{t\neq i=1}^{n}(A_i-A_t)(A_i-q^{-2}A_t)}\right). \tag{A27}$$

Consider the complex function

$$f(z)=\frac{\Pi_{t=1}^{n-1}(z-q^2B_t)(z-q^2C_t)}{\Pi_{t=1}^{n}(z-A_t)(z-q^2A_t)}. \tag{A28}$$

This function is analytic in the extended complex plane except in its simple poles A_t, q^2A_t, $t=1,2,...,n$. Therefore the sum over all residues of $f(z)$ vanishes. It is a simple calculation to see that

$$\begin{aligned}\sum_{i=1}^{n}\mathrm{Res}(f(A_i))+\sum_{i=1}^{n}\mathrm{Res}(f(q^2A_i))=\frac{1}{1-q^2}\sum_{i=1}^{n}\frac{1}{A_i}\Bigg(&\frac{\Pi_{t=1}^{n-1}(A_i-q^2B_t)(A_i-q^2C_t)}{\Pi_{t\neq i=1}^{n}(A_i-A_t)(A_i-q^2A_t)}\\ &-\frac{\Pi_{t=1}^{n-1}(A_i-B_t)(A_i-C_t)}{\Pi_{t\neq i=1}^{n}(A_i-A_t)(A_i-q^{-2}A_t)}\Bigg)=0.\end{aligned} \tag{A29}$$

Hence (A26) holds and therefore $\mathscr{B}_{n,j,m}\equiv\mathscr{B}_{2k,j,m}=0$ for any k, j and m.

(3^0) The proof of $\mathscr{C}_{2k-1,p,l}=0$ is similar to the previous case. Relabeling the indices of $L_{i,2k}$, one writes (A9) as

$$\begin{aligned}\mathscr{C}_{2k-1,p,l}=&\sum_{s'=0}^{1}\sum_{j=1-k}^{k-1}(-1)^{s'}\\ &\times\frac{\Pi_{r=1-k}^{k-2}[L_{r,2k-2}-L_{j,2k-1}-s']\Pi_{r=1-k}^{k-2}[L'_{r,2k}-L_{j,2k-1}-s']}{\Pi_{r\neq j=1-k}^{k-1}[L_{r,2k-1}-L_{j,2k-1}-s'][L_{r,2k-1}-L_{j,2k-1}-s'+1]}.\end{aligned} \tag{A30}$$

Setting in (A30) $i=j+k$ and $t=r+k$ and introducing the notation

$$a_t=L_{t-k,2k-1}+1,\quad t=1,...,2k-1,$$

$$b_t = L_{t-k,2k-2}, \quad t = 1,...,2k-2,$$

$$c_t = L'_{t-k,2k}, \quad t = 1,...,2k-2, \tag{A31}$$

$$s = 1 - s',$$

$$n = 2k-1,$$

we obtain for $\mathscr{C}_{2k-1,p,l} \equiv \mathscr{C}_{n,p,l}$

$$\mathscr{C}_{n,p,l} = -\sum_{s=0}^{1} \sum_{i=1}^{n} (-1)^s \frac{\Pi_{t=1}^{n-1}[a_i - b_t - s]\Pi_{t=1}^{n-1}[a_i - c_t - s]}{\Pi_{t\neq i=1}^{n}[a_i - a_t - s][a_i - a_t - s + 1]}. \tag{A32}$$

It has already been shown [see (A26)] that the rhs of (A32) vanishes. Hence $\mathscr{C}_{2k-1,p,l} = 0$.

The conclusion so far is that the rhs of (A6) is also equal to zero, i.e., Eq. (A5) holds for any $i \in \mathbf{Z}_+$. It remains to prove (A5) for $i \notin \mathbf{Z}_+$.

(3b) The verification of (A5) corresponding to $i = -1$ is based on Eqs. (19) and (21) and is very simple.

(3c) The proof of (A5), corresponding to $i < -1$, is based essentially on the identities (23b), (24b) and (24d). The latter are proved similarly as (23a), (24a) and (24c).

(4) The verification of last Cartan relation

$$[\rho(e_i), \rho(f_j)]|M) = 0, \quad i \neq j, \tag{A33}$$

as well as of the Serre relations is lengthy but straightforward. It is based on the identities (24)–(27).

[1] S. Levendorskii and Y. Soibelman, Commun. Math. Phys. **140**, 399 (1991).
[2] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, J. Phys. Soc. Jpn. **50**, 3806 (1981).
[3] V. G. Kac and D. H. Peterson, Proc. Natl. Acad. Sci. USA **78**, 3308 (1981).
[4] T. D. Palev, Lie algebraical aspects of the quantum statistics, Thesis, Institute of Nuclear Research and Nuclear Energy, Sofia (1976), Lie algebraic aspects of quantum statistics. Unitary quantization (*A*-quantization) preprint JINR E17-10550 (1977) and hep-th/9705032.
[5] T. D. Palev, Rep. Math. Phys. **31**, 241 (1992).
[6] F. D. M. Haldane, Phys. Rev. Lett. **67**, 937 (1991).
[7] T. D. Palev and N. I. Stoilova, J. Math. Phys. **38**, 2506 (1997); hep-th/9606011.
[8] D. Karabali and V. P. Nair, Nucl. Phys. B **438**, 551 (1995).
[9] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, Publ. RIMS Kyoto Univ. **18**, 1077 (1982).
[10] P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).
[11] V. Drinfeld, Quantum groups. ICM proceedings, Berkeley (1986), p. 798.
[12] T. D. Palev, Funkt. Anal. Prilozh. **24**(1), 82 (1990) (in Russian); Funct. Anal. Appl. **24**, 72 (1990) (English translation).
[13] T. D. Palev, J. Math. Phys. **31**, 579 (1990).
[14] T. D. Palev, Funkt. Anal. Prilozh. **24**(3), 88 (1990) (in Russian); Funct. Anal. Appl. **24**, 250 (1990) (English translation).
[15] T. D. Palev, J. Math. Phys. **31**, 1078 (1990).
[16] I. Gel'fand and M. Zetlin, Dokl. Akad. Nauk SSSR **71**, 825 (1950).
[17] M. Jimbo, Lect. Notes Phys. **246**, 335 (1986).
[18] T. D. Palev and V. N. Tolstoy, Commun. Math. Phys. **141**, 549 (1991).
[19] T. D. Palev, N. I. Stoilova, and J. Van der Jeugt, Commun. Math. Phys. **166**, 367 (1994).
[20] K. A. Nguyen and N. I. Stoilova, J. Math. Phys. **36**, 5979 (1995).
[21] T. D. Palev and N. I. Stoilova, J. Phys. A **30**, L699 (1997).

Symmetries of history quantum theories and decoherence functionals

Oliver Rudolph[a)]
Theoretical Physics Group, Blackett Laboratory, Imperial College of Science, Technology and Medicine, Prince Consort Road, London SW7 2BZ, United Kingdom

(Received 2 May 1998; accepted for publication 8 July 1998)

Recently, Schreckenberg investigated symmetries in the context of *history quantum theories*. In the case that the space of histories is given by the set of projectors on some finite dimensional Hilbert space he obtained, first, a complete characterization of all physical symmetries of history quantum theories — an analog of Wigner's theorem — and, second, a complete mathematical characterization of symmetries of single decoherence functionals. In this paper we extend Schreckenberg's results to the case where the underlying space of histories is given by the set of projectors on some infinite dimensional Hilbert space. © *1998 American Institute of Physics.* [S0022-2488(98)02011-8]

I. INTRODUCTION

History quantum theories are an interesting candidate for quantum theories in which space and time are treated more symmetrically than in standard Hamiltonian quantum mechanics and in which there is no *a priori* external notion of time. They have been introduced by Isham[1] based on ideas of the so called consistent histories approach (see Refs. 2–7). Isham's approach has subsequently been studied by several authors.[8–24] In Isham's approach a history quantum theory is characterized by the space of histories on the one hand, which correspond to the general temporal properties of the quantum system, and by the space of decoherence functionals on the other hand, which correspond to the generalized quantum state of a quantum system in a history theory. To get deeper insight into the structural content of general history quantum theories it is worthwhile to study the mathematical structure of history quantum theories in some detail. In the present paper we will restrict ourselves to the mathematical study of history quantum theories. For a lucid account of the physical motivation for studying history quantum theories the reader is referred to the original works by Isham and co-workers.[1,8–17]

In the present note we study the notion of *symmetry* in history quantum theories. Symmetries in history quantum theories have first been studied by Schreckenberg.[8,9] In these works Schreckenberg considered the situation where the space of histories is given by the set of projection operators on some finite dimensional Hilbert space and, generalizing Wigner's symmetry theorem[25] to history theories, proved that there exists a one-to-one correspondence between physical symmetries of history quantum theories and operators $U\otimes U$ on $\mathfrak{H}\otimes\mathfrak{H}$ where U is a unitary or anti-unitary operator on $\mathfrak{H}$.[8] In Ref. 9 Schreckenberg studied symmetries of decoherence functionals in the finite dimensional case and proved that the symmetries of a given decoherence functional d are in a one-to-one correspondence with operators $U\otimes U$ on $\mathfrak{H}\otimes\mathfrak{H}$ such that $[\mathfrak{X}_d, U\otimes U]=0$, where $\mathfrak{X}_d$ denotes the unique trace class operator associated with d [see Ref. 16 and below, Eq. (1)].

Our aim here is to investigate to what extent Schreckenberg's two main results remain valid when the underlying Hilbert space is infinite dimensional. To this end we use recent results in the representation theory of decoherence functionals in the infinite dimensional case.[21] We shall show in the present note that Schreckenberg's first theorem classifying the physical symmetries of a history quantum theory is valid also in the infinite dimensional case. We shall also investigate to what extent Schreckenberg's second theorem characterizing the symmetries of a decoherence

a)Electronic mail: o.rudolph@ic.ac.uk

 © 1998 American Institute of Physics

functional remains valid and give a complete characterization of symmetries of bounded decoherence functionals in the infinite dimensional case.

This paper is organized as follows: in the next section the necessary mathematical preliminaries are presented. In Sec. III we discuss physical symmetries of history quantum theories and in the fourth section we investigate symmetries of decoherence functionals.

II. MATHEMATICAL PRELIMINARIES

Consider a history quantum theory in which the space of histories is given by the set of projection operators on some in general infinite dimensional Hilbert space. In this case decoherence functionals are defined as follows: let $\mathcal{B}(\mathfrak{H})$ be the space of all bounded operators on a Hilbert space $\mathfrak{H}$ and let $\mathcal{P}(\mathfrak{H})$ be the lattice of projections in $\mathcal{B}(\mathfrak{H})$ which is interpreted as the set of histories. Then a *decoherence functional for* $\mathfrak{H}$ is a complex valued function $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$, defined on all ordered pairs of projections in $\mathcal{P}(\mathfrak{H})$, for which the following conditions are satisfied:

(i) *Hermiticity:* $d(p,q)=d(q,p)^*$ for each p and q in $\mathcal{P}(\mathfrak{H})$. (Here * denotes complex conjugation.)
(ii) *Positivity:* $d(p,p)\geqslant 0$ for each $p\in\mathcal{P}(\mathfrak{H})$.
(iii) *Normalization:* $d(1,1)=1$.
(iv) *Orthoadditivity:* $d(p_1+p_2,q)=d(p_1,q)+d(p_2,q)$ whenever $p_1\in\mathcal{P}(\mathfrak{H})$ and $p_2\in\mathcal{P}(\mathfrak{H})$ are perpendicular and $q\in\mathcal{P}(\mathfrak{H})$ is an arbitrary projection.

Moreover, in the infinite dimensional case, a decoherence functional is said to be *completely additive* if, whenever $\{p_i\}_{i\in I}$ is an infinite collection of pairwise orthogonal projections, $d(\Sigma_{i\in I}p_i,q)=\Sigma_{i\in I}d(p_i,q)$, for each $q\in\mathcal{P}(\mathfrak{H})$. Here the convergence is always absolute. In this note we denote the set of all decoherence functionals on the Hilbert space $\mathfrak{H}$ by $\mathcal{D}(\mathfrak{H})$.

When $\mathfrak{H}$ is finite dimensional and of dimension greater than 2, it follows from Isham, Linden and Schreckenberg[16] that there is a one-to-one correspondence between bounded decoherence functionals d on $\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})$ and trace class operators $\mathfrak{X}$ on $\mathfrak{H}\otimes\mathfrak{H}$ according to the rule

$$d(p,q)=\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}((p\otimes q)\mathfrak{X}), \tag{1}$$

for all $p,q\in\mathcal{P}(\mathfrak{H})$, where

(i) $\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}((p\otimes q)\mathfrak{X})=\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}((q\otimes p)\mathfrak{X}^\dagger)$;
(ii) $\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}((p\otimes p)\mathfrak{X})\geqslant 0$;
(iii) $\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}(\mathfrak{X})=1$.

In contrast to the finite dimensional case where all bounded decoherence functionals admit an *Isham–Linden–Schreckenberg (ILS) representation* as in Eq. (1), in the infinite dimensional case only those decoherence functionals satisfying an additional so-called *tensor boundedness* condition[21] admit an ILS representation.

Denote by $\mathcal{K}(\mathfrak{H})$ the set of compact operators on $\mathfrak{H}$, and by $\mathcal{K}(\mathfrak{H})\bullet\mathcal{K}(\mathfrak{H})$ the algebraic tensor product of $\mathcal{K}(\mathfrak{H})$ with itself. We recall from Ref. 21 that for every bounded decoherence functional there exists a unique bounded bilinear form $\mathcal{D}:\mathcal{B}(\mathfrak{H})\times\mathcal{B}(\mathfrak{H})\to\mathbb{C}$ which extends d and a unique linear functional $\beta:\mathcal{K}(\mathfrak{H})\bullet\mathcal{K}(\mathfrak{H})\to\mathbb{C}$ such that $\beta(x\otimes y)=\mathcal{D}(x,y)$ for all $x,y\in\mathcal{K}(\mathfrak{H})$. A bounded decoherence functional is said to be *tensor bounded* if the associated functional β is bounded on $\mathcal{K}(\mathfrak{H})\bullet\mathcal{K}(\mathfrak{H})$, when $\mathcal{K}(\mathfrak{H})\bullet\mathcal{K}(\mathfrak{H})$ is equipped with the norm induced from $\mathcal{B}(\mathfrak{H}\otimes\mathfrak{H})$. We denote the projection operator from $\mathfrak{H}\otimes\mathfrak{H}$ onto the one dimensional subspace spanned by $\xi\in\mathfrak{H}\otimes\mathfrak{H}$ by p_ξ. A bounded decoherence functional is said to be *tracially bounded* if there exists a constant C such that for every unit vector $\xi\in\mathfrak{H}\otimes_{alg}\mathfrak{H}$ the associated functional β satisfies $|\beta(p_\xi)|\leqslant C$, where $\mathfrak{H}\otimes_{alg}\mathfrak{H}$ denotes the algebraic tensor product of $\mathfrak{H}$ by itself.

In Ref. 21 it has been shown that a bounded decoherence functional is ILS representable if and only if it is tensor bounded (provided the underlying Hilbert space is of dimension greater than two). Moreover, it has been shown that for a tracially bounded decoherence functional d there exists a unique bounded operator $\mathfrak{M}_d$ on $\mathfrak{H}\otimes\mathfrak{H}$ such that $d(p,q)=\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}(p\otimes q\mathfrak{M}_d)$, for all finite rank projections p,q on $\mathfrak{H}$ (provided $\mathfrak{H}$ is of dimension greater than two).

The structure of the space of decoherence functionals has been studied in detail in Ref. 21. We shall need the following result from Ref. 21 in the sequel.

Proposition 2.1: Let $\mathfrak{H}$ be a Hilbert space of dimension greater than two and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a bounded decoherence functional for $\mathfrak{H}$. Then there exist families of trace class operators $\{\mathfrak{X}_i\}_{i\in I}$, and $\{\mathfrak{Y}_i\}_{i\in I}$ on $\mathfrak{H}$, where, for each x and y in $\mathcal{K}(\mathfrak{H})$, $\Sigma_{i\in I}|\mathrm{tr}_{\mathfrak{H}}(x\mathfrak{X}_i)|^2$ and $\Sigma_{i\in I}|\mathrm{tr}_{\mathfrak{H}}(y\mathfrak{Y}_i)|^2$ are convergent, such that, for all $p,q\in\mathcal{P}(\mathfrak{H})$,

$$d(p,q)=\sum_{i\in I}\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}\{p\otimes q(\mathfrak{X}_i\otimes\mathfrak{X}_i^\dagger-\mathfrak{Y}_i\otimes\mathfrak{Y}_i^\dagger)\},\tag{2}$$

where the infinite series is absolutely convergent.

III. SYMMETRIES OF HISTORY QUANTUM THEORIES

Wigner's symmetry theorem in ordinary quantum mechanics is discussed in detail in Refs. 8, 25, and 26. We denote the single-time Hilbert space of standard quantum mechanics by $\mathcal{H}$ and the set of rays in $\mathcal{H}$ by $\mathcal{R}(\mathcal{H})$.

Then, in short, a physical symmetry is an affine bijective map $\xi:\mathcal{R}(\mathcal{H})\to\mathcal{R}(\mathcal{H})$ such that the transition probability between the physical states corresponding to the rays in $\mathcal{H}$ remains unchanged under ξ. Wigner's theorem now states that there is a bijective correspondence between physical symmetries and unitary and antiunitary operators on $\mathcal{H}$.

Considering the histories formulation of standard quantum mechanics[1,16] we see that every physical symmetry ξ in standard quantum mechanics induces an affine orthogonality preserving map $\Xi:\mathcal{R}(\mathfrak{H})\to\mathcal{R}(\mathfrak{H})$ on the rays of the history Hilbert space $\mathfrak{H}$ which in turn induces a map $d\mapsto d^{\Xi}$ on the space of decoherence functionals. Explicitly, if the (anti)unitary operator on $\mathcal{H}$ corresponding to ξ is given by U, then Ξ acts on an n-time history h as $h\mapsto(U\otimes\cdots\otimes U)h(U^\dagger\otimes\cdots\otimes U^\dagger)$, where $U\otimes\cdots\otimes U$ denotes the n-fold tensor product of U by itself. Details can be found in Ref. 9.

The above considerations about symmetries in ordinary quantum mechanics and in the history formulation of standard quantum mechanics motivate the ansatz that in a general history quantum theory a symmetry is still given by an affine one-to-one and onto mapping $\mathcal{Q}:\mathcal{R}(\mathfrak{H})\to\mathcal{R}(\mathfrak{H})$ on the history Hilbert space $\mathfrak{H}$.

The families $\{\mathfrak{X}_i\}_{i\in I}$, and $\{\mathfrak{Y}_i\}_{i\in I}$ on $\mathfrak{H}$ appearing in Eq. (2) in Proposition 2.1 are not unique. For every trace class operator T on $\mathfrak{H}$ there exist two orthonormal families $\{|\psi_k\rangle\}_{k\in K}$ and $\{|\phi_k\rangle\}_{k\in K}$ of vectors in $\mathfrak{H}$ and a family of real positive numbers $\{\lambda_k\}_{k\in K}$ with $\Sigma_{k\in K}\lambda_k<\infty$ such that $T=\Sigma_{k\in K}\lambda_k|\psi_k\rangle\langle\phi_k|$.

Thus, given a representation (2) of a bounded decoherence functional by families of trace class operators $\{\mathfrak{X}_i\}_{i\in I}$ and $\{\mathfrak{Y}_i\}_{i\in I}$, every affine one-to-one map $\mathcal{Q}:\mathcal{R}(\mathfrak{H})\to\mathcal{R}(\mathfrak{H})$ induces a map $\{\mathfrak{X}_i\}_{i\in I}\mapsto\{\mathfrak{X}_i^{\mathcal{Q}}\}_{i\in I}$ and $\{\mathfrak{Y}_i\}_{i\in I}\mapsto\{\mathfrak{Y}_i^{\mathcal{Q}}\}_{i\in I}$. This map, however, depends, in general, obviously on the choice of the representations for the $\mathfrak{X}_i$ and $\mathfrak{Y}_i$. Different choices may lead to different sets of operators $\{\mathfrak{X}_i^{\mathcal{Q}}\}_{i\in I}$ and $\{\mathfrak{Y}_i^{\mathcal{Q}}\}_{i\in I}$. Moreover, the map $\mathcal{Q}$ is, in general, not preserving orthogonality between rays. Nevertheless, it is easy to check that the transformed operators are also compact operators, but *a priori* not necessarily of trace class. The sets of transformed operators may or may not define via Eq. (2) a transformed bounded decoherence functional. Since the representation in the form (2) is not unique, *a priori* we have to take into account the possibilities that the image of d under $\mathcal{Q}$ depends upon the representation (2) chosen and, even worse, that the series defining the image of d under $\mathcal{Q}$ may be not convergent for all $p,q\in\mathcal{P}(\mathfrak{H})$. The well-defined bounded decoherence functionals which can be obtained from a representation (2) of d through the application of $\mathcal{Q}$ will be denoted by $d_s^{\mathcal{Q}}$ where the index s distinguishes the possibly different decoherence functionals arising from different representations of d and of the $\{\mathfrak{X}_i\}_{i\in I}$ and $\{\mathfrak{Y}_i\}_{i\in I}$.

So far, the situation is clearly unsatisfactory from a physical point of view. From a *physical symmetry* $\mathcal{Q}$ of a history theory one would reasonably expect that $\mathcal{Q}$ induces a well-defined map $d\mapsto d^{\mathcal{Q}}$ on the space of decoherence functionals.

This motivates the following definition in the spirit of Ref. 8.

Definition 3.1: Consider a history quantum theory with underlying Hilbert space $\mathfrak{H}$. A physical symmetry *of the history quantum theory is a pair $(\mathcal{Q}_1,\mathcal{Q}_2)$ consisting of an affine bijection*

$\mathcal{Q}_1:\mathcal{R}(\mathfrak{H})\to\mathcal{R}(\mathfrak{H})$ such that $\mathcal{Q}_1$ induces a map $\mathcal{Q}_2:\mathcal{D}(\mathfrak{H})\to\mathcal{D}(\mathfrak{H}), d\mapsto d^{\mathcal{Q}_2}$ on the space of decoherence functionals such that the value of the decoherence functional is preserved, i.e.,

$$d(p,q)=d^{\mathcal{Q}_2}(p^{\mathcal{Q}_1},q^{\mathcal{Q}_1})$$

for all $d\in\mathcal{D}(\mathfrak{H})$ and all $p,q\in\mathcal{R}(\mathfrak{H})$.

The following theorem depends crucially on the choice of the space of decoherence functionals. We assume that the set of all physical decoherence functionals contains as a subset all bounded decoherence functionals.

Theorem 3.2: *Consider a history quantum theory with underlying Hilbert space $\mathfrak{H}$. Then there exists a bijective correspondence between physical symmetries of the history quantum theory and unitary or anti-unitary operators on $\mathfrak{H}$. The physical symmetry $\mathcal{Q}^U=(\mathcal{Q}_1^U,\mathcal{Q}_2^U)$ corresponding to U is explicitly given by*

$$\mathcal{Q}^U:\mathcal{P}(\mathfrak{H})\times\mathcal{D}(\mathfrak{H})\to\mathcal{P}(\mathfrak{H})\times\mathcal{D}(\mathfrak{H}),\mathcal{Q}^U(p,d)=(p^{\mathcal{Q}^U},d^{\mathcal{Q}^U})=(UpU^\dagger,d_U),$$

where d_U is the decoherence functional defined by $d_U(p,q):=d(U^\dagger pU,U^\dagger qU)$, for all $p,q\in\mathcal{P}(\mathfrak{H})$.

Proof: Obviously, $\mathcal{Q}^U$ defines a physical symmetry of the history quantum theory for every unitary or anti-unitary operator U. Conversely, let $(\mathcal{Q}_1,\mathcal{Q}_2)$ be a physical symmetry. We proceed in analogy with Ref. 8. Since the invariance requirement must hold for all physical decoherence functionals, it has particularly to be true for all tensor bounded decoherence functionals whose associated trace class operator [via Eq. (1)] is of finite rank. Consider, for instance, the decoherence functional $d_P(p,q):=\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}(p\otimes qP\otimes P)$ where P is some one dimensional projector on $\mathfrak{H}$. The invariance requirement amounts to $d_P^{\mathcal{Q}_2}(p^{\mathcal{Q}_1},q^{\mathcal{Q}_1})=d_P(p,q)$. Applying Wigner's theorem and using the fact that $\mathcal{Q}_2$ is induced by $\mathcal{Q}_1$ shows that the transformations leaving the values of the decoherence functional invariant are of the form $\mathcal{Q}^U$ for some unitary or anti-unitary operator on $\mathfrak{H}$. Particularly, it follows that $\mathcal{Q}_2$ is, in fact, single valued. □

Notice that the requirement that $\mathcal{Q}_2$ is a map in Definition 3.1 is redundant and actually follows from the requirement that the value of the decoherence functional is invariant. As shown in the proof of Theorem 3.2, the latter requirement already fixes $\mathcal{Q}_1$ to be equal to $\mathcal{Q}^U$ for some unitary or anti-unitary U. However, the relation $\mathcal{Q}_2$ induced by $\mathcal{Q}^U$ is the map $d\mapsto d_U$.

IV. SYMMETRIES OF DECOHERENCE FUNCTIONALS

Definition 4.1: Let $\mathfrak{H}$ be a Hilbert space and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a decoherence functional. A symmetry of *d is a unitary or anti-unitary operator U on $\mathfrak{H}$ such that $d(p,q)=d(UpU^\dagger,UqU^\dagger)$, for all $p,q\in\mathcal{P}(\mathfrak{H})$.*

Proposition 4.2: Let $\mathfrak{H}$ be a Hilbert space and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a decoherence functional. The set $\mathcal{S}[d]$ of symmetries of d carries the structure of a group.

Proof: The unit element of $\mathcal{S}[d]$ is given by the unit operator **1** and the group multiplication is given by multiplication of operators. We show that $U\in\mathcal{S}[d]$ implies $U^\dagger\in\mathcal{S}[d]$. Since U is a symmetry of d, we have $d(p,q)=d(UU^\dagger pUU^\dagger,UU^\dagger qUU^\dagger)=d(U^\dagger pU,U^\dagger qU)$, for all $p,q\in\mathcal{P}(\mathfrak{H})$. This implies $U^\dagger\in\mathcal{S}[d]$. □

Schreckenberg[9] has shown that if $\mathfrak{H}$ is finite dimensional and of dimension greater than 2, then the set $\mathcal{S}[d]$ of all symmetries of a decoherence functional d is given by

$$\mathcal{S}[d]=\{U\in\mathrm{Aut}(\mathfrak{H}):[\mathfrak{X}_d,U\otimes U]=0\},$$

where $\mathrm{Aut}(\mathfrak{H})$ denotes the set of all unitary and anti-unitary operators on $\mathfrak{H}$ and where $\mathfrak{X}_d$ denotes the ILS operator from Eq. (1) associated with d.

We shall need the following lemma.

Lemma 4.3: Let $\mathcal{L}$ be a bounded operator on $\mathfrak{H}\otimes\mathfrak{H}$ such that, for all $\alpha,\beta\in\mathfrak{H}$,

$$\langle\mathcal{L}(\alpha\otimes\beta),\alpha\otimes\beta\rangle=0.$$

Then $\mathcal{L}=0$.

A proof can be found, for instance, in Ref. 21.

Theorem 4.4: *Let $\mathfrak{H}$ be a Hilbert space of dimension greater than 2 and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a tracially bounded, completely additive decoherence functional and denote by $\mathfrak{M}_d$ the bounded operator on $\mathfrak{H}\otimes\mathfrak{H}$ associated with d. The set $\mathcal{S}[d]$ of symmetries of d is given by*

$$\mathcal{S}[d]=\{U\in\mathrm{Aut}(\mathfrak{H}):[\mathfrak{M}_d,U\otimes U]=0\},$$

where $\mathrm{Aut}(\mathfrak{H})$ *denotes the set of all unitary and anti-unitary operators on* $\mathfrak{H}$.

Proof: It is well known that every projection operator P on $\mathfrak{H}$ can be written as a sum over projection operators of rank one, $P=\Sigma_i P_i$. Thus, according to Theorem 4.4 in Ref. 21, whenever p and q are projections in $\mathcal{P}(\mathfrak{H})$ and $\{p_j\}_{j\in J}$ and $\{q_i\}_{i\in I}$ are, respectively, orthogonal families of rank one projections with $p=\Sigma_{j\in J}p_j$ and $q=\Sigma_{i\in I}q_i$, then

$$d(p,q)=\sum_{j\in J}\sum_{i\in I}\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}((p_j\otimes q_i)\mathfrak{M}_d). \tag{3}$$

Let $U\in\{U\in\mathrm{Aut}(\mathfrak{H}):[\mathfrak{M}_d,U\otimes U]=0\}$, then it follows from Eq. (3) that U is a symmetry of d. Conversely, let U by a symmetry of d. Then, by Lemma 4.3, $\mathfrak{M}_d=(U\otimes U)\mathfrak{M}_d(U^\dagger\otimes U^\dagger)$, i.e., $U\in\{U\in\mathrm{Aut}(\mathfrak{H}):[\mathfrak{M}_d,U\otimes U]=0\}$. □

Corollary 4.5: Let $\mathfrak{H}$ be a Hilbert space of dimension greater than 2 and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a tracially bounded decoherence functional and denote by $\mathfrak{M}_d$ the bounded operator on $\mathfrak{H}\otimes\mathfrak{H}$ associated with d. Then the set $\mathcal{S}[d]$ of symmetries of d satisfies

$$\mathcal{S}[d]\subseteq\{U\in\mathrm{Aut}(\mathfrak{H}):[\mathfrak{M}_d,U\otimes U]=0\}.$$

Remarks: (i) In the case of a tensor bounded decoherence functional the requirement of complete additivity can be dropped from the theorem.

(ii) A *weak symmetry* of a decoherence functional d is a unitary or anti-unitary operator $U\in\mathrm{Aut}(\mathfrak{H})$ such that the diagonal values of the decoherence functional (which are interpreted as the physical probabilities) are invariant, i.e.,

$$d(p,p)=d(UpU^\dagger,UpU^\dagger),\forall p\in\mathcal{P}(\mathfrak{H}).$$

For the weak symmetries of a tracially bounded, completely additive decoherence functional d we find, in general, only $\mathfrak{M}_d-(U\otimes U)\mathfrak{M}_d(U^\dagger\otimes U^\dagger)=\Delta_U$, where $\mathrm{tr}_{\mathfrak{H}\otimes\mathfrak{H}}(p\otimes p\Delta_U)=0$ for all $p\in\mathcal{P}(\mathfrak{H})$. A detailed discussion about weak symmetries can be found in Ref. 9 and need not be repeated here.

A bounded decoherence functional (not necessarily tracially bounded), according to Proposition 2.1, can be represented as a series of trace class operators, which, in general, converges only in a weak sense. Thus we obtain the following.

Theorem 4.6: *Let $\mathfrak{H}$ be a Hilbert space of dimension greater than 2 and let $d:\mathcal{P}(\mathfrak{H})\times\mathcal{P}(\mathfrak{H})\to\mathbb{C}$ be a completely additive, bounded decoherence functional. Let $\{\mathfrak{X}_i\}_{i\in I}$ and $\{\mathfrak{Y}_i\}_{i\in I}$ be families of trace class operators forming a representation of d as in Eq. (2). Then, the set $\mathcal{S}[d]$ of symmetries of d is given by*

$$\mathcal{S}[d]=\left\{U\in\mathrm{Aut}(\mathfrak{H}):\sum_{i\in I}\langle[(\mathfrak{X}_i\otimes\mathfrak{X}_i^\dagger-\mathfrak{Y}_i\otimes\mathfrak{Y}_i^\dagger)U^\dagger\otimes U^\dagger,U\otimes U]\xi\otimes\eta,\xi\otimes\eta\rangle=0,\forall\xi,\eta\in\mathfrak{H}\right\}.$$

The proof is analogous to the proof of Theorem 4.4. The representation of d in Proposition 2.1 is not unique. Theorem 4.6 is independent of the choice of the representation in Eq. (2).

ACKNOWLEDGMENTS

The author is a Marie Curie Research Fellow and carries out his research at Imperial College as part of a European Union training project financed by the European Commission under the

Training and Mobility of Researchers (TMR) program. The author would like to thank Professor Christopher J. Isham, Dr. Stephan Schreckenberg, and Professor John D. Maitland Wright for useful discussions and conversations.

[1] C. J. Isham, J. Math. Phys. **35**, 2157 (1994).
[2] R. B. Griffiths, J. Stat. Phys. **36**, 219 (1984).
[3] R. Omnès, *The Interpretation of Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1994).
[4] M. Gell-Mann and J. B. Hartle, in *Proceedings of the 25th International Conference on High Energy Physics, Singapore, August 2–8, 1990*, edited by K. K. Phua and Y. Yamaguchi (World Scientific, Singapore, 1990), p. 1303.
[5] M. Gell-Mann and J. B. Hartle, in *Proceedings of the Third International Symposium on the Foundations of Quantum Mechanics in the Light of New Technology*, edited by S. Kobayashi, H. Ezawa, Y. Murayama, and S. Nomura (Physical Society of Japan, Tokyo, 1990), p. 321.
[6] M. Gell-Mann and J. B. Hartle, in *Complexity, Entropy and the Physics of Information, Santa Fe Institute Studies in the Science of Complexity*, edited by W. Zurek (Addison-Wesley, Reading, MA, 1990), Vol. VIII, p. 425.
[7] M. Gell-Mann and J. B. Hartle, Phys. Rev. D **47**, 3345 (1993).
[8] S. Schreckenberg, J. Math. Phys. **37**, 6086 (1996).
[9] S. Schreckenberg, J. Math. Phys. **38**, 759 (1997).
[10] S. Schreckenberg, J. Math. Phys. **36**, 4735 (1995).
[11] C. J. Isham, *Quantum Logic and Decohering Histories*, Lecture given at the Conference *Theories of Fundamental Interactions*, Maynooth, Eire, 24–26 May 1995; in *Topics in Quantum Field Theory*, edited by D. Tchrakian (World Scientific, Singapore, 1995).
[12] C. J. Isham, Int. J. Theor. Phys. **36**, 785 (1997).
[13] C. J. Isham and N. Linden, J. Math. Phys. **35**, 5452 (1994).
[14] C. J. Isham and N. Linden, J. Math. Phys. **36**, 5392 (1995).
[15] C. J. Isham and N. Linden, Phys. Rev. A **55**, 4030 (1997).
[16] C. J. Isham, N. Linden, and S. Schreckenberg, J. Math. Phys. **35**, 6360 (1994).
[17] C. J. Isham, N. Linden, K. Savvidou, and S. Schreckenberg, J. Math. Phys. **39**, 1818 (1998).
[18] S. Pulmannová, Int. J. Theor. Phys. **34**, 189 (1995).
[19] O. Rudolph, Int. J. Theor. Phys. **35**, 1581 (1996).
[20] O. Rudolph, J. Math. Phys. **37**, 5368 (1996).
[21] O. Rudolph and J. D. M. Wright, J. Math. Phys. **38**, 5643 (1997).
[22] J. D. M. Wright, J. Math. Phys. **36**, 5409 (1995).
[23] J. D. M. Wright, Commun. Math. Phys. **191**, 493 (1998).
[24] J. D. M. Wright, Atti Sem. Mat. Fis. Univ. Modena (to appear).
[25] V. Bargmann, J. Math. Phys. **5**, 147 (1964).
[26] V. S. Varadarajan, *Geometry of Quantum Theory*, 2nd ed. (Springer, Berlin, 1985).

A version of Thirring's approach to the Kolmogorov–Arnold–Moser theorem for quadratic Hamiltonians with degenerate twist

C. Chandre[a] and H. R. Jauslin
Laboratoire de Physique, CNRS, Université de Bourgogne, BP 400, F-21011 Dijon, France

(Received 23 February 1998; accepted for publication 8 July 1998)

We give a proof of the Kolmogorov–Arnold–Moser (KAM) theorem on the existence of invariant tori for weakly perturbed Hamiltonian systems, based on Thirring's approach for Hamiltonians that are quadratic in the action variables. The main point of this approach is that the iteration of canonical transformations on which the proof is based stays within the space of quadratic Hamiltonians. We show that Thirring's proof for nondegenerate Hamiltonians can be adapted to Hamiltonians with degenerate twist. This case, in fact, drastically simplifies Thirring's proof. © *1998 American Institute of Physics.*
[S0022-2488(98)00611-2]

I. INTRODUCTION

Regular Hamiltonian dynamics is characterized by the existence of as many independent conserved quantities as degrees of freedom d. As a consequence, each trajectory is confined to evolve on an invariant torus of dimension d. The Kolmogorov–Arnold–Moser (KAM) technique was developed to prove the stability under small perturbations of a large fraction of these invariant tori. The KAM theorem[1–4] states that if the frequency satisfies a diophantine condition, and the size of the perturbation ε is sufficiently small, then a torus of that frequency will be stable. The proof is based on an iterative algorithm to construct the invariant tori. Each step of the KAM iteration consists of a coordinate transformation that reduces the size of the perturbation from order ε to ε^2.

In this paper, we derive a KAM theorem for a family of Hamiltonians that has been used in Refs. 5–7 within the setup of a renormalization-group approach to the breakup of invariant tori.

We consider the following class of Hamiltonians with d degrees of freedom, that are quadratic in the action variables $\mathbf{A}=(A_1,A_2,\ldots,A_d)\in\mathbb{R}^d$ and described by three scalar functions of the angles $\boldsymbol{\varphi}=(\varphi_1,\varphi_2,\ldots,\varphi_d)\in\mathbb{T}^d$ (the d-dimensional torus parametrized, e.g., by $[0,2\pi]^d$):

$$H(\mathbf{A},\boldsymbol{\varphi})=\frac{1}{2}m(\boldsymbol{\varphi})(\boldsymbol{\Omega}\cdot\mathbf{A})^2+[\boldsymbol{\omega}_0+g(\boldsymbol{\varphi})\boldsymbol{\Omega}]\cdot\mathbf{A}+f(\boldsymbol{\varphi}), \tag{1.1}$$

where $\boldsymbol{\omega}_0\in\mathbb{R}^d$ is the frequency vector of the considered torus, and $\boldsymbol{\Omega}$ is some other constant vector. Without loss of generality, we assume that $\boldsymbol{\Omega}=(\Omega_1,\ldots,\Omega_d)$ is of norm one, i.e., $|\boldsymbol{\Omega}|=\Sigma_{i=1}^d|\Omega_i|=1$. The functions m, g, and f are real analytic on $\mathbb{T}^d$, i.e., they are holomorphic functions on some complex neighborhood of $\mathbb{T}^d$.

We consider $\boldsymbol{\omega}_0$ verifying a diophantine condition

$$|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|^{-1}\leqslant\sigma|\boldsymbol{\nu}|^{\tau},\quad \forall\,\boldsymbol{\nu}=(\nu_1,\ldots,\nu_d)\in\mathbb{Z}^d\backslash\{0\}, \tag{1.2}$$

for some $\tau>d-1$ and $\sigma>0$.

The aim is to prove the existence of a torus with frequency vector $\boldsymbol{\omega}_0$ for Hamiltonian systems described by Eq. (1.1). The method is to find a canonical transformation such that the equations of

[a]Electronic mail: cchandre@jupiter.u-bourgogne.fr

 © 1998 American Institute of Physics

motion expressed in the new coordinates show trivially the existence of this torus. For Hamiltonians of type (1.1), if one takes g and f equal to zero, then the resulting equations of motion are

$$\frac{d\mathbf{A}}{dt} = -\frac{1}{2}\frac{\partial m}{\partial \boldsymbol{\varphi}}(\boldsymbol{\Omega}\cdot\mathbf{A})^2, \tag{1.3}$$

$$\frac{d\boldsymbol{\varphi}}{dt} = m(\boldsymbol{\varphi})(\boldsymbol{\Omega}\cdot\mathbf{A})\boldsymbol{\Omega} + \boldsymbol{\omega}_0. \tag{1.4}$$

Thus there exists a torus with frequency vector $\boldsymbol{\omega}_0$ located at $\mathbf{A}=0$ (even if the resulting Hamiltonian is not globally integrable, i.e., for $\mathbf{A}\neq 0$). This canonical transformation cannot be defined directly, because the formal expressions that appear in the classical perturbation theory do not converge due to the presence of small denominators. The construction is done via an iterative algorithm. We iterate a canonical change of coordinates that maps a Hamiltonian of the form (1.1) with a perturbation (g,f) of order $\mathcal{O}(\varepsilon)$, into a Hamiltonian with (g',f') of order $\mathcal{O}(\varepsilon^2)$. When the iteration converges, the perturbation (g,f) is completely eliminated.

The fact that m does not need to be eliminated to prove the existence of the torus with frequency vector $\boldsymbol{\omega}_0$, allows us to stay with Hamiltonians that are quadratic in the actions at each step of the iteration. This approach has been developed by Thirring.[8,9] In fact, Thirring considered a nondegenerate family of Hamiltonians quadratic in the actions, of the form

$$H(\mathbf{A},\boldsymbol{\varphi}) = \tfrac{1}{2}\,\mathbf{A}\cdot M(\boldsymbol{\varphi})\mathbf{A} + [\boldsymbol{\omega}_0 + \mathbf{g}(\boldsymbol{\varphi})]\cdot\mathbf{A} + f(\boldsymbol{\varphi}), \tag{1.5}$$

where M is a $d\times d$ matrix such that $\det M\neq 0$, and $\mathbf{g}$ a vector. This implies that Thirring's Hamiltonian (1.5) satisfies the ''twist condition''

$$\det\left|\frac{\partial^2 H}{\partial\mathbf{A}\partial\mathbf{A}}\right| = \det M \neq 0. \tag{1.6}$$

The Hamiltonian (1.1) does not satisfy the twist condition for any $\boldsymbol{\varphi}$. The extension of the KAM theorem for degenerate systems was done by Arnold[10] (see also Refs. 11–15). In the case we consider here, the rank of the matrix in (1.6) is one. Thus there is a twist in a particular direction, which is the only relevant one for the considered perturbation.

Condition (1.6) is also called the (standard) nondegeneracy condition. KAM theorems were also proven for Hamiltonians which satisfy the isoenergetic nondegeneracy condition,[16,17] for which the following determinant of order $d+1$ does not vanish

$$\det\begin{vmatrix} \dfrac{\partial^2 H}{\partial\mathbf{A}\partial\mathbf{A}} & \dfrac{\partial H}{\partial\mathbf{A}} \\ \left(\dfrac{\partial H}{\partial\mathbf{A}}\right)^T & 0 \end{vmatrix} \neq 0. \tag{1.7}$$

For $d=2$, the determinant (1.7) is equal to $-m(\boldsymbol{\varphi})[\det(\boldsymbol{\Omega},\boldsymbol{\omega}_0)]^2$. Thus the isoenergetic nondegeneracy condition is not satisfied if $m(\boldsymbol{\varphi})$ has zeroes.

For $d>2$, the model (1.1) is isoenergetically degenerate. The present result is thus not a direct consequence of the ones in Refs. 16 and 17. We notice that in the case where $\boldsymbol{\Omega}$ is parallel to $\boldsymbol{\omega}_0$, the Hamiltonian (1.1) is integrable (because of the existence of d integrals of motion in involution: H and $\boldsymbol{\omega}_0^{\perp}\cdot\boldsymbol{\varphi}$, where $\boldsymbol{\omega}_0^{\perp}$ denotes a vector perpendicular to $\boldsymbol{\omega}_0$, and there are $d-1$ such independant vectors).

In this article, we present a self-contained proof of the KAM theorem for Hamiltonian (1.1) based on Thirring's proof for Hamiltonian (1.5). The advantages are twofold: On one hand, we show that Thirring's proof can be adapted to degenerate twist Hamiltonians, and on the other hand, the resulting proof becomes even simpler than Thirring's. The fact that the iteration stays within the space of Hamiltonians quadratic in the actions is very useful, e.g., for numerical implementations to study the breakup of invariant tori.[5,6]

Before entering into details of the theorem, we give basic definitions and notations: We denote $\mathscr{D}_\rho$ a complex neighborhood of $\mathbb{T}^d$ defined by

$$\mathscr{D}_\rho=\{\boldsymbol{\varphi}\in\mathbb{C}^d \,|\, \|\mathrm{Im}\ \boldsymbol{\varphi}\|\leqslant\rho\}, \tag{1.8}$$

where $\|z\|=\max_i(|z_i|)$, for any $z\in\mathbb{C}^d$. We will consider, in the following calculations, scalar functions defined in $\mathscr{D}_\rho$. More precisely, we define $\mathscr{A}_\rho$ as the set of complex functions $f(\boldsymbol{\varphi})$ defined on $\mathscr{D}_\rho$, analytic in the interior of $\mathscr{D}_\rho$, of period 2π in the variables φ_i, and which have real values when $\boldsymbol{\varphi}\in\mathbb{R}^d$. We define a norm on $\mathscr{A}_\rho$: $\|f\|_\rho=\sup_{\varphi\in\mathscr{D}_\rho}|f(\boldsymbol{\varphi})|$. We define $\langle f\rangle$, the mean value of f by

$$\langle f\rangle=\int_{\mathbb{T}^d}\frac{d^d\boldsymbol{\varphi}}{(2\pi)^d}f(\boldsymbol{\varphi}). \tag{1.9}$$

In the following sections, we will use the notation $\boldsymbol{\partial} f=\partial f/\partial\boldsymbol{\varphi}$ for any function of the angles. In Sec. II, we define the KAM iteration. In Sec. III, we give estimates on the transformed functions. Finally, in Sec. IV, we iterate the transformation, and prove the following KAM theorem for the family of Hamiltonians (1.1):

Theorem 1: *For* $H(\mathbf{A},\boldsymbol{\varphi})=1/2\ m(\boldsymbol{\varphi})(\boldsymbol{\Omega}\cdot\mathbf{A})^2+[\boldsymbol{\omega}_0+g(\boldsymbol{\varphi})\boldsymbol{\Omega}]\cdot\mathbf{A}+f(\boldsymbol{\varphi})$, *suppose that*

(*i*) $|\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}|^{-1}\leqslant\sigma|\boldsymbol{\nu}|^\tau, \forall\,\boldsymbol{\nu}\in\mathbb{Z}^d\backslash\{0\}$, *for some* $\sigma>0$ *and* $\tau>d-1$;

(*ii*) m, g, *and* f *are analytic in* $\mathscr{D}_\rho$, *and* $|\boldsymbol{\Omega}|=1$;

(*iii*) $\max(\|g\|_\rho,\|f\|_\rho)\leqslant\beta_0$, *where*

$$\beta_0=2^{-10}\Gamma^{-6}c^{-3}(h/3)^{3(\tau+d+1)},$$

$$c=2^{3d}\sigma[2e^{-1}(\tau+1)]^{\tau+1},$$

$$\Gamma=\max(1,\|m\|_\rho,|\langle m\rangle|^{-1}),$$

$$h<\inf(2\rho/9,c^{1/(\tau+d+1)}).$$

Then, there exists a canonical transformation, analytic in $\mathscr{D}_{\rho-9h/2}$, *such that the Hamiltonian expressed in the new coordinates is*

$$H^{(\infty)}=\tfrac{1}{2}m^{(\infty)}(\boldsymbol{\varphi}^{(\infty)})(\boldsymbol{\Omega}\cdot\mathbf{A}^{(\infty)})^2+\boldsymbol{\omega}_0\cdot\mathbf{A}^{(\infty)},$$

where $m^{(\infty)}$ *is analytic in* $\mathscr{D}_{\rho-9h/2}$. *As a consequence, the system has an analytic invariant torus of frequency* $\boldsymbol{\omega}_0$.

Remark: If the constant part $\langle m\rangle$ of the quadratic term is zero, the maximal amplitude of the perturbation β_0 given by the theorem becomes zero. Thus in order to have a nontrivial result, we require that $\langle m\rangle$ is nonzero.

II. EXPRESSIONS OF THE GENERATING FUNCTION AND OF THE NEW HAMILTONIAN

We perform a canonical transformation $\mathscr{U}_F:(\boldsymbol{\varphi},\mathbf{A})\mapsto(\boldsymbol{\varphi}',\mathbf{A}')$ defined by a generating function[18,19] linear in the action variables, and characterized by two scalar functions Y, Z, of the angles, and a constant a, of the form

$$F(\mathbf{A}',\boldsymbol{\varphi})=(\mathbf{A}'+a\boldsymbol{\Omega})\cdot\boldsymbol{\varphi}+Y(\boldsymbol{\varphi})\boldsymbol{\Omega}\cdot\mathbf{A}'+Z(\boldsymbol{\varphi}), \tag{2.1}$$

leading to

$$\mathbf{A}=\frac{\partial F}{\partial\boldsymbol{\varphi}}=\mathbf{A}'+(\boldsymbol{\Omega}\cdot\mathbf{A}')\boldsymbol{\partial}Y+a\boldsymbol{\Omega}+\boldsymbol{\partial}Z, \tag{2.2}$$

$$\boldsymbol{\varphi}'=\frac{\partial F}{\partial\mathbf{A}'}=\boldsymbol{\varphi}+Y(\boldsymbol{\varphi})\boldsymbol{\Omega}. \tag{2.3}$$

Inserting Eq. (2.2) into the Hamiltonian (1.1), we obtain the expression of the Hamiltonian in the mixed representation of new action variables and old angle variables

$$\tilde{H}(\mathbf{A}',\boldsymbol{\varphi}) = \tfrac{1}{2}\,\tilde{m}(\boldsymbol{\varphi})(\boldsymbol{\Omega}\cdot\mathbf{A}')^2 + [\boldsymbol{\omega}_0 + \tilde{g}(\boldsymbol{\varphi})\boldsymbol{\Omega}]\cdot\mathbf{A}' + \tilde{f}(\boldsymbol{\varphi}), \tag{2.4}$$

with

$$\tilde{m} = (1+\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y)^2 m, \tag{2.5}$$

$$\tilde{g} = g + \boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Y + mb + \boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y(g+mb), \tag{2.6}$$

$$\tilde{f} = f + \boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Z + \tfrac{1}{2}mb^2 + gb, \tag{2.7}$$

where $b(\boldsymbol{\varphi}) = a\Omega^2 + \boldsymbol{\Omega}\cdot\boldsymbol{\partial} Z$. As the new angles do not depend on the actions, the Hamiltonian (1.1) expressed in the new variables is also quadratic in the actions, and of the same form as (1.1). We notice that this transformation does not change $\boldsymbol{\Omega}$.

We determine the generating function (2.1) such that the new functions $\tilde{g}$, $\tilde{f}$ in $H\circ\mathscr{U}_F$ vanish to the first order in ε. This leads to the conditions

$$\boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Z + f = \text{const}, \tag{2.8}$$

$$\boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Y + g + m(a\Omega^2 + \boldsymbol{\Omega}\cdot\boldsymbol{\partial} Z) = 0. \tag{2.9}$$

The constant a allows us to have $\langle\tilde{g}\rangle = \mathscr{O}(\varepsilon^2)$, in order to keep the frequency $\boldsymbol{\omega}_0$ at the chosen value. We recall that the functions g and f are of order $\mathscr{O}(\varepsilon)$ and m is of order one; as a consequence Y, Z and a are of order $\mathscr{O}(\varepsilon)$. Equations (2.8) and (2.9) are solved by representing them in Fourier space. They define the generating function F as

$$Z(\boldsymbol{\varphi}) = \sum_{\nu\neq 0} \frac{i}{\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}} f_\nu e^{i\nu\cdot\varphi}, \tag{2.10}$$

$$a = -\frac{\langle g\rangle + \langle m\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Z\rangle}{\Omega^2\langle m\rangle}, \tag{2.11}$$

$$Y(\boldsymbol{\varphi}) = \sum_{\nu\neq 0} \frac{i}{\boldsymbol{\omega}_0\cdot\boldsymbol{\nu}} (g_\nu + (m\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Z)_\nu + m_\nu a\Omega^2) e^{i\nu\cdot\varphi}, \tag{2.12}$$

where the scalar functions m, g, f are represented by their Fourier series, e.g.,

$$f(\boldsymbol{\varphi}) = \sum_{\nu\in\mathbb{Z}^d} f_\nu e^{i\nu\cdot\varphi}. \tag{2.13}$$

Thus the scalar functions in $\tilde{H}$ become

$$\tilde{m} = (1+\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y)^2 m, \tag{2.14}$$

$$\tilde{g} = -(\boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Y)(\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y), \tag{2.15}$$

$$\tilde{f} = \tfrac{1}{2}(g - \boldsymbol{\omega}_0\cdot\boldsymbol{\partial} Y)(a\Omega^2 + \boldsymbol{\Omega}\cdot\boldsymbol{\partial} Z). \tag{2.16}$$

The expression of $\tilde{H}$ in the new angles requires the inversion of Eq. (2.3). We denote H' the Hamiltonian expressed in the new variables $H'(\mathbf{A}',\boldsymbol{\varphi}') = \tilde{H}(\mathbf{A}',\boldsymbol{\varphi})$:

$$H'(\mathbf{A}',\boldsymbol{\varphi}') = \tfrac{1}{2}m'(\boldsymbol{\varphi}')(\boldsymbol{\Omega}\cdot\mathbf{A}')^2 + [\boldsymbol{\omega}_0 + g'(\boldsymbol{\varphi}')\boldsymbol{\Omega}]\cdot\mathbf{A}' + f'(\boldsymbol{\varphi}'). \tag{2.17}$$

In the following section, we give estimates on Z, Y and a, in order to derive estimates on m', g' and f'.

III. ESTIMATES ON THE GENERATING FUNCTION AND ON THE NEW HAMILTONIAN

A. Estimate on the generating function

Lemma 1: Let f be an element of $\mathscr{A}_\rho$, and $\boldsymbol{\omega}_0$ satisfy Eq. (1.2), then $Z(\boldsymbol{\varphi})$ given by Eq. (2.10) is analytic on $\mathscr{D}_{\rho-h}$ (for any choice of h with $0<h<\rho<2$), and satisfies the following estimates

$$\|Z\|_{\rho-h}\leqslant \overline{c}h^{-\tau-d}\|f\|_\rho, \tag{3.1}$$

$$\|\boldsymbol{\omega}_0\cdot\boldsymbol{\partial}Z\|_{\rho-h}\leqslant\|f\|_\rho, \tag{3.2}$$

$$\|\boldsymbol{\Omega}\cdot\boldsymbol{\partial}Z\|_{\rho-h}\leqslant ch^{-\tau-d-1}\|f\|_\rho, \tag{3.3}$$

where $\overline{c}=2^{3d}\sigma(2e^{-1}\tau)^\tau$ and $c=2^{3d}\sigma[2e^{-1}(\tau+1)]^{\tau+1}$.

Proof: This lemma is proved, for instance, in Refs. 10 and 20. First, we estimate the Fourier coefficients of f expressed as integrals of f over the torus,

$$f_\nu=\int_{\mathbb{T}^d}\frac{d^d\boldsymbol{\varphi}}{(2\pi)^d}f(\boldsymbol{\varphi})e^{-i\nu\cdot\varphi}. \tag{3.4}$$

By a shift of the angles $\boldsymbol{\varphi}\mapsto\boldsymbol{\varphi}+i\boldsymbol{\eta}$, where $\eta_i=-\rho\nu_i/|\nu_i|$,

$$f_\nu=\int_{\mathbb{T}^d}\frac{d^d\boldsymbol{\varphi}}{(2\pi)^d}f(\boldsymbol{\varphi}+i\boldsymbol{\eta})e^{-i\nu\cdot\varphi}e^{-\rho|\nu|}. \tag{3.5}$$

Then $|f_\nu|\leqslant\|f\|_\rho e^{-\rho|\nu|}$, for all $\boldsymbol{\nu}\in\mathbb{Z}^d$.

To estimate Z given by Eq. (2.10), we use the diophantine property (1.2),

$$|Z(\boldsymbol{\varphi})|\leqslant\sigma\sum_{\nu\neq0}|\boldsymbol{\nu}|^\tau|f_\nu|e^{-\nu\cdot\mathrm{Im}\,\varphi}. \tag{3.6}$$

Then, for all $\boldsymbol{\varphi}\in\mathscr{D}_{\rho-h}$, we have

$$|Z(\boldsymbol{\varphi})|\leqslant\sigma\|f\|_\rho\sum_{\nu\neq0}|\boldsymbol{\nu}|^\tau e^{-h|\nu|}. \tag{3.7}$$

To estimate the sum, we use the following property which is easy to check (see Ref. 2):

$$|\boldsymbol{\nu}|^\tau\leqslant\left(\frac{2\tau}{eh}\right)^\tau e^{|\nu|h/2},\quad\forall\,\tau,h>0. \tag{3.8}$$

From the fact that $1/(1-e^{-x})<2/x$, for all $x\in]0,1[$, we have

$$\sum_{\nu\neq0}e^{-|\nu|h/2}<\left(\frac{8}{h}\right)^d, \tag{3.9}$$

which gives the estimate (3.1). The estimate (3.3) is obtained by the same calculations, and the estimate (3.2) is straightforward from Eq. (2.8). □

Lemma 2: Let m, g, and f be elements of $\mathscr{A}_\rho$, and $\boldsymbol{\omega}_0$ satisfy Eq. (1.2), then $Y(\boldsymbol{\varphi})$ given by Eq. (2.12) is analytic on $\mathscr{D}_{\rho-2h}$. The constant a and the function Y satisfy the following estimates

$$\|Y\|_{\rho-2h}\leqslant\overline{c}h^{-\tau-d}(1+|\langle m\rangle|^{-1}\|m\|_\rho)\times(\|g\|_\rho+ch^{-\tau-d-1}\|f\|_\rho\|m\|_\rho), \tag{3.10}$$

$$\|\boldsymbol{\omega}_0\cdot\boldsymbol{\partial}Y\|_{\rho-2h}\leqslant(1+|\langle m\rangle|^{-1}\|m\|_\rho)\times(\|g\|_\rho+ch^{-\tau-d-1}\|f\|_\rho\|m\|_\rho), \tag{3.11}$$

$$\|\boldsymbol{\Omega}\cdot\boldsymbol{\partial}Y\|_{\rho-2h}\leqslant ch^{-\tau-d-1}(1+|\langle m\rangle|^{-1}\|m\|_\rho)\times(\|g\|_\rho+ch^{-\tau-d-1}\|f\|_\rho\|m\|_\rho), \tag{3.12}$$

$$|a|\Omega^2\leqslant|\langle m\rangle|^{-1}(\|g\|_\rho+ch^{-\tau-d-1}\|f\|_\rho\|m\|_\rho). \tag{3.13}$$

The proof is the same as the one for the estimates on Z.

B. Estimate on the new Hamiltonian

The expression of the scalar functions of the Hamiltonian expressed in the new actions and old angles are explicitly known, and estimates on these functions are easy to obtain from the estimates on the generating function. One has then to invert Eq. (2.3) in order to obtain the estimate on the Hamiltonian expressed in the new angle variables. The Jacobian of this transformation is

$$\det\left|\frac{\partial\varphi_j'}{\partial\varphi_k}\right|=1+\boldsymbol{\Omega}\cdot\boldsymbol{\partial}Y. \tag{3.14}$$

In this section, we denote $V_\rho=\max(\|f\|_\rho,\|g\|_\rho)$ and $\Gamma=\max(1,\|m\|_\rho,|\langle m\rangle|^{-1})$.

From Eq. (2.3) and estimate (3.10), one has the following inequality:

$$|\boldsymbol{\varphi}'-\boldsymbol{\varphi}|\leqslant 2\Gamma^3ch^{-\tau-d}(1+ch^{-\tau-d-1})V_\rho, \tag{3.15}$$

for all $\boldsymbol{\varphi}\in\mathscr{D}_{\rho-2h}$ (we recall that $\Gamma\geqslant 1$). If we assume that V_ρ and h are sufficiently small, one has an estimate on the new angles. More precisely, we assume the following inequalities:

$$\Gamma^3c^2h^{-2(\tau+d+1)}V_\rho\leqslant\tfrac{1}{4}, \tag{3.16}$$

$$ch^{-\tau-d-1}\geqslant 1. \tag{3.17}$$

Then, for all $\boldsymbol{\varphi}\in\mathscr{D}_{\rho-2h}$, we have

$$|\boldsymbol{\varphi}'-\boldsymbol{\varphi}(\boldsymbol{\varphi}')|\leqslant h. \tag{3.18}$$

As a consequence, $\mathscr{D}_{\rho-2h}\subset\Phi(\mathscr{D}_{\rho-3h})$ by the map $\boldsymbol{\varphi}'\mapsto\boldsymbol{\varphi}=\Phi(\boldsymbol{\varphi}')$ given by Eq. (2.3). In order to express the estimates with respect to $\boldsymbol{\varphi}'$, it suffices thus to restrict the width of the strip $\mathscr{D}$ from $\rho-2h$ to $\rho-3h$, e.g., $\|f'\|_{\rho-3h}\leqslant\|\tilde{f}\|_{\rho-2h}$. Then, the estimates on the new perturbation (g',f') are obtained from Eqs. (2.15) and (2.16) and from the estimates on the generating function:

$$\|g'\|_{\rho-3h}\leqslant 2^4\Gamma^6c^3h^{-3(\tau+d+1)}V_\rho^2, \tag{3.19}$$

$$\|f'\|_{\rho-3h}\leqslant 2^3\Gamma^5c^2h^{-2(\tau+d+1)}V_\rho^2. \tag{3.20}$$

Taking into account condition (3.17), we obtain the estimate on $V'_{\rho-3h}=\max(\|g'\|_{\rho-3h},\|f'\|_{\rho-3h})$:

$$V'_{\rho-3h}\leqslant 2^4\Gamma^6c^3h^{-3(\tau+d+1)}V_\rho^2. \tag{3.21}$$

Concerning the quadratic term, we deduce from Eqs. (3.12) and (2.14) that

$$\|m'\|_{\rho-3h}\leqslant\|m\|_\rho(1+4\Gamma^3c^2h^{-2(\tau+d+1)}V_\rho)^2. \tag{3.22}$$

Then using Hypothesis (*iii*) of the theorem, we obtain $\|m'\|_{\rho-3h}\leqslant 2\Gamma$.

The mean value of m' is determined by the integral

$$\langle m'\rangle=\int_{\mathbb{T}^d}\frac{d^d\boldsymbol{\varphi}'}{(2\pi)^d}m'(\boldsymbol{\varphi}'). \tag{3.23}$$

With the change of variable $\boldsymbol{\varphi}' \mapsto \boldsymbol{\varphi}$ given by Eqs. (2.3) and (3.14) and using Eq. (2.14), we rewrite the integral

$$\langle m' \rangle = \int_{\mathbb{T}^d} \frac{d^d\boldsymbol{\varphi}}{(2\pi)^d} |1+\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y| \tilde{m}(\boldsymbol{\varphi}) = \int_{\mathbb{T}^d} \frac{d^d\boldsymbol{\varphi}}{(2\pi)^d} |1+\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y|^3 m(\boldsymbol{\varphi}). \tag{3.24}$$

To estimate $|\langle m' \rangle|^{-1}$, we first estimate the difference $|\langle m' \rangle - \langle m \rangle|$ using Eq. (3.24), and we obtain

$$|\langle m' \rangle - \langle m \rangle| \leqslant \|m\|_\rho (3\|\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y\|_{\rho-2h} + 3\|\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y\|^2_{\rho-2h} + \|\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y\|^3_{\rho-2h}). \tag{3.25}$$

From Eq. (3.12) and condition (3.16), we have the estimate

$$\|\boldsymbol{\Omega}\cdot\boldsymbol{\partial} Y\|_{\rho-2h} \leqslant 4\Gamma^3 c^2 h^{-2(\tau+d+1)} V_\rho \leqslant 1, \tag{3.26}$$

which leads to

$$|\langle m' \rangle - \langle m \rangle| \leqslant 28\Gamma^4 c^2 h^{-2(\tau+d+1)} V_\rho . \tag{3.27}$$

One can easily check that from Hypothesis (*iii*), it follows that

$$|\langle m' \rangle - \langle m \rangle| \leqslant \frac{1}{2\Gamma}. \tag{3.28}$$

Writing that $\langle m' \rangle = \langle m \rangle + \langle m' \rangle - \langle m \rangle$, we have the following estimate on $|\langle m' \rangle|^{-1}$:

$$|\langle m' \rangle|^{-1} \leqslant \frac{|\langle m \rangle|^{-1}}{|1-|\langle m \rangle|^{-1}|\langle m' \rangle - \langle m \rangle||}. \tag{3.29}$$

From Eq. (3.28), we deduce that $|\langle m' \rangle|^{-1} \leqslant 2\Gamma$. Therefore we have proved that $\Gamma' \equiv \max(1, \|m'\|_{\rho-3h}, |\langle m' \rangle|^{-1}) \leqslant 2\Gamma$.

IV. CONVERGENCE OF THE ITERATION: KAM THEOREM

The estimate (3.21) gives a precise meaning to the statement that the perturbation is reduced from V to V^2: we see that the actual reduction is somewhat smaller, since h depends on V_ρ through the condition (3.16) [the actual rate of reduction that takes this into account will be expressed by Eq. (4.11) below]. The counterpart is that the width of the domain of analyticity of the new functions is reduced from ρ to $\rho - 3h$; for this reason, h is called the ''analyticity loss parameter.'' At the n th step of the iteration, we choose h_n as the analyticity loss parameter such that the final domain of analyticity (after an infinite number of iterations) does not shrink to zero. For instance, we choose

$$h_n = h 3^{-n+1}, \quad \text{with} \quad h < \inf(2\rho/9, c^{1/(\tau+d+1)}). \tag{4.1}$$

The second term in the inf is to guarantee also the condition (3.17). We denote $\Gamma_n = \max(1, \|m^{(n)}\|_{\rho_n}, |\langle m^{(n)} \rangle|^{-1})$ and $V_n = \max(\|g^{(n)}\|_{\rho_n}, \|f^{(n)}\|_{\rho_n})$, where $m^{(n)}$, $g^{(n)}$ and $f^{(n)}$ denote the three scalar functions defining the Hamiltonian (1.1) after n iterations, and

$$\rho_n = \rho_{n-1} - 3h_n = \rho - 3\sum_{k=1}^{n} h_k . \tag{4.2}$$

The previous section gave the following estimates

$$V_n \leqslant 2^4 \Gamma^6_{n-1} c^3 h_n^{-3(\tau+d+1)} V^2_{n-1}, \tag{4.3}$$

$$\|m^{(n)}\|_{\rho_n} \leqslant \|m^{(n-1)}\|_{\rho_{n-1}} \times (1 + 4\Gamma^3_{n-1} c^2 h_n^{-2(\tau+d+1)} V_{n-1})^2. \tag{4.4}$$

We now prove that Γ_n is bounded, and that V_n tends to zero faster than geometrically as n goes to infinity.

Denoting

$$V_0=\max(\|g\|_\rho,\|f\|_\rho), \tag{4.5}$$

$$\Gamma=\max(1,\|m\|_\rho,|\langle m\rangle|^{-1}), \tag{4.6}$$

$$\gamma=2^{10}\Gamma^6c^3h^{-3(\tau+d+1)}, \tag{4.7}$$

$$\delta=3^{3(\tau+d+1)}, \tag{4.8}$$

we define the sequence of values

$$\varepsilon_n=\gamma\delta^{n-1}\varepsilon_{n-1}^2=\frac{(\gamma\delta V_0)^{2^n}}{\gamma\delta^{n+1}}. \tag{4.9}$$

We will show by induction that

$$\Gamma_k\leqslant 2\Gamma, \tag{4.10}$$

$$V_k\leqslant\varepsilon_k, \tag{4.11}$$

for all k. If we assume V_0 to be sufficiently small, such that $\gamma\delta V_0<1$, i.e.,

$$2^{10}\Gamma^6c^3h^{-3(\tau+d+1)}3^{3(\tau+d+1)}V_0<1, \tag{4.12}$$

which is the hypothesis (*iii*) of the theorem, then V_n tends to zero faster than $\delta^{-n}=3^{-3n(\tau+d+1)}$ as n goes to infinity. The bounds (4.10) and (4.11) are satisfied for $k=0$, since $\Gamma_0=\Gamma>0$ and $V_0=\varepsilon_0$. Suppose that for all $k<n$, these bounds are satisfied. From Eqs. (4.3) and (4.10), we deduce that

$$V_n\leqslant\gamma\delta^{n-1}V_{n-1}^2. \tag{4.13}$$

Then using Eq. (4.11), we have $V_n\leqslant\varepsilon_n$. Concerning Γ_n, Eq. (4.4) leads to

$$\|m^{(n)}\|_{\rho_n}\leqslant\|m\|_\rho\times\prod_{k=0}^{n-1}(1+2^5\Gamma^3c^2h^{-2(\tau+d+1)}3^{2k(\tau+d+1)}V_k)^2. \tag{4.14}$$

Using Eq. (4.11) and Hypothesis (*iii*), we obtain

$$2^5\Gamma^3c^2h^{-2(\tau+d+1)}3^{2k(\tau+d+1)}V_k\leqslant 3^{-(k+3)(\tau+d+1)}. \tag{4.15}$$

Then we have

$$\|m^{(n)}\|_{\rho_n}\leqslant\Gamma\prod_{k=0}^{\infty}(1+3^{-(k+3)(\tau+d+1)})^2. \tag{4.16}$$

Using the fact that $\Pi(1+x_k)^2\leqslant\exp(2\Sigma x_k)$, we show that the infinite product converges, and that it is smaller than 2. Thus $\|m^{(n)}\|_{\rho_n}\leqslant 2\Gamma$.

For $|\langle m^{(n)}\rangle|^{-1}$, we use estimate (3.29):

$$\frac{|\langle m^{(n)}\rangle|^{-1}}{|\langle m^{(n-1)}\rangle|^{-1}}\leqslant\frac{1}{|1-|\langle m^{(n-1)}\rangle|^{-1}|\langle m^{(n)}\rangle-\langle m^{(n-1)}\rangle||}. \tag{4.17}$$

The difference $|\langle m^{(n)}\rangle-\langle m^{(n-1)}\rangle|$ is evaluated as in the first step (3.27):

$$|\langle m^{(n)}\rangle-\langle m^{(n-1)}\rangle|\leqslant 28\cdot 2^4\Gamma^4c^2h^{-2(\tau+d+1)}\times 3^{2(n-1)(\tau+d+1)}V_{n-1}. \tag{4.18}$$

Using Eq. (4.11), Hypothesis (*iii*), and the fact that Γ and $ch^{-(\tau+d+1)}$ are greater than one,

$$|\langle m^{(n)}\rangle - \langle m^{(n-1)}\rangle| \leqslant 2^{-1}\Gamma^{-1}3^{-(n+2)(\tau+d+1)}. \tag{4.19}$$

Then,

$$|\langle m^{(n)}\rangle|^{-1} \leqslant \Gamma \prod_{k=0}^{\infty} (1-3^{-(k+2)(\tau+d+1)})^{-1}. \tag{4.20}$$

Using the fact that $\Pi(1-x_k)^{-1}=\exp(-\Sigma \ln(1-x_k))$, and that $-\ln(1-x)\leqslant x2\ln 2$, $\forall x\in[0,1/2]$, we show by straightforward calculations that $|\langle m^{(n)}\rangle|^{-1}\leqslant 2\Gamma$.

Thus we have proved (4.10) and (4.11) for all k.

We finally have to verify that with the choices (4.1) and (4.2), the conditions (3.16) and (3.17) are satisfied at each step of the iteration. Equation (3.17) is guaranteed by Eq. (4.1): $ch_n^{-(\tau+d+1)}\geqslant ch^{-(\tau+d+1)}\geqslant 1$. For Eq. (3.16), we first notice that, since $\Gamma\geqslant 1$, $\Gamma_n\leqslant 2\Gamma$, and $ch^{-(\tau+d+1)}>1$, we can write

$$\Gamma_{n-1}^3 c^2 h_n^{-2(\tau+d+1)} < 2^{-7}\gamma\delta^n.$$

Therefore using Eq. (4.11) and Hypothesis (*iii*), we have

$$\Gamma_{n-1}^3 c^2 h_n^{-2(\tau+d+1)} V_{n-1} < 2^{-7}(\gamma\delta V_0)^{2^{n-1}} < \tfrac{1}{4}.$$

We have shown that, for sufficiently small initial perturbation, the Hamiltonian converges under successive iterations of the transformation to a Hamiltonian of the form (1.1) with $g=f=0$. In order to complete the proof, we have still to verify that the composition of the infinitely many canonical transformations is a well-defined finite canonical transformation. It suffices to verify that $|\boldsymbol{\varphi}^{(\infty)}-\boldsymbol{\varphi}|$ and $|\mathbf{A}^{(\infty)}-\mathbf{A}|$ are finite. This is an immediate consequence of the fast convergence of the iteration. From Eq. (3.15), we deduce, e.g., that

$$|\boldsymbol{\varphi}^{(\infty)}-\boldsymbol{\varphi}| \leqslant \sum_{n=1}^{\infty} h_n = 3h/2. \tag{4.21}$$

Analogously, using Eqs. (2.2), (3.3), (3.11), (3.12), and (4.11), for $|\mathbf{A}|\leqslant A_0$, we can bound

$$|\mathbf{A}^{(\infty)}-\mathbf{A}| < \text{const}\times \sum_{n=1}^{\infty} h_n. \tag{4.22}$$

ACKNOWLEDGMENTS

We thank G. Benfatto, L. H. Eliasson, G. Gallavotti, M. Herman, and H. Koch for helpful discussions. Support from EC Contract No. ERBCHRXCT94-0460 for the project ''Stability and universality in classical mechanics'' and from the Conseil Régional de Bourgogne is acknowledged.

[1] A. N. Kolmogorov, Dokl. Akad. Nauk SSSR **98**, 527 (1954) [in *Stochastic Behaviour in Classical and Quantum Hamiltonian Systems*, edited by G. Casati and J. Ford, Lecture Notes in Physics Vol. 93 (Springer, Berlin,1979), p. 51].
[2] V. I. Arnold, Russ. Math. Surv. **18**, 9 (1963).
[3] J. Moser, Nachr. Akad. Wiss. Goett. II, Math.-Phys. Kl. **1**, 1 (1962).
[4] G. Gallavotti, in *Scaling and Self-Similarity in Physics*, edited by J. Fröhlich (Birkhäuser, Boston, 1983).
[5] M. Govin, C. Chandre, and H. R. Jauslin, Phys. Rev. Lett. **79**, 3881 (1997).
[6] C. Chandre, M. Govin, and H. R. Jauslin, Phys. Rev. E **57**, 1536 (1998).
[7] C. Chandre, M. Govin, H. R. Jauslin, and H. Koch, Phys. Rev. E **57**, 6612 (1998).
[8] W. Thirring, *A Course in Mathematical Physics I: Classical Dynamical Systems* (Springer-Verlag, Berlin, 1992), p. 153.
[9] H. R. Jauslin, in *II Granada Lectures in Computational Physics*, edited by P. L. Garrido and J. Marro (World Scientific, Singapore, 1993).
[10] V. I. Arnold, Russ. Math. Surv. **18**, 85 (1963).
[11] Ch.-Q. Cheng and Y.-S. Sun, J. Diff. Eqns. **114**, 288 (1994).
[12] H. Rüssmann, in *Stochastics, Algebra and Analysis in Classical and Quantum Dynamics*, edited by S. Albeverio, Ph. Blanchard, and D. Testard, Mathematics and its Applications, Vol. 59 (Kluwer, Dordrecht, 1990).

[13] G. Gallavotti, Commun. Math. Phys. **164**, 145 (1994).
[14] G. Gallavotti, Rev. Math. Phys. **6**, 343 (1994).
[15] R. D. Paul, archived in mp_arc@math.utexas.edu, #98-9, 1998.
[16] A. Delshams and P. Gutiérrez, J. Diff. Eqns. **128**, 415 (1996).
[17] H. W. Broer and G. B. Huitema, J. Diff. Eqns. **90**, 52 (1991).
[18] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1980).
[19] G. Gallavotti, *The Elements of Mechanics* (Springer-Verlag, New York, 1983).
[20] G. Benettin, L. Galgani, A. Giorgilli, and J. M. Strelcyn, Nuovo Cimento B **79**, 201 (1984).

On periodical trajectories of the billiard systems within an ellipsoid in R^d and generalized Cayley's condition

Vladimir Dragović[a] and Milena Radnović
Mathematical institute SANU, Kneza Mihaila 35, 11000 Beograd, Yugoslavia

(Received 3 April 1998; accepted for publication 20 May 1998)

The analytical condition for a trajectory of the billiard system within an ellipsoid of any dimension is given. An interesting property is obtained—every periodical trajectory with the period less than the dimension of the space is placed in a coordinate hyperplane. © *1998 American Institute of Physics.*
[S0022-2488(98)00711-7]

I. INTRODUCTION

Poncelet's theorem[1,2] is one of most beautiful and most important claims in projective geometry. Suppose we have two conics given in the plane, and a polygon inscribed in one and circumscribed about another conic. Poncelet's theorem states that there exist infinitely many such polygons. Even more, every point of the outer conic is a vertex of some polygonal line with this property, and all these polygons have the same number of sides.

A natural question arises: Is it possible to find an analytical condition for determination if there existed a polygon inscribed in one and circumscribed about another conic? Such a condition was established by Cayley. This result, in a nice geometrical form, can be found in Ref. 2. Griffiths and Harris derived Cayley's theorem using the analytical condition on a point to be of finite order in an elliptic curve group structure.[3] They also gave a generalization of Poncelet's theorem, considering polyhedra inscribed in two quadrics in three-dimensional space.[4]

Poncelet's theorem can be applied in mechanics. Trajectories of the billiard inside an ellipse have an interesting property: every segment of a given trajectory is tangent to the same conic, confocal with the boundary.[5] One can consider the billiard system inside an ellipsoid in the space of any dimension. Any trajectory of that mechanical system in R^d is tangent to $d-1$ quadrics confocal with the outer hypersurface.[6,7] The generalization of Poncelet's theorem (which is made in a different direction than Griffiths and Harris') is valid in the d-dimensional case also: There exists a closed trajectory with $d-1$ given confocal caustics if and only if infinitely many such trajectories exist, and all of them have the same period.[6,7]

The aim of our investigation was to find the corresponding generalization of Cayley's theorem, i.e., the analytical condition on a billiard trajectory within an ellipsoid in R^d to be periodic. Modifying Griffiths and Harris' ideas and applying them to the isospectral curves of Veselov and Moser, we established such a condition in three-dimensional space.[8]

In this paper, we are going to give similar conditions for the ellipsoidal billiard in any dimension. One of the consequences of the obtained result can be restated in a simple manner—there are no periodic billiard trajectories within an ellipsoid in d-dimensional space with nondegenerate caustics, with period less than d. If a trajectory is closed after k bounces, $k<d$, then it is placed in one of the coordinate k-planes and can be examined as in k-dimensional case. Let us note that in the planar case, the obtained condition is not identical to the classical Cayley's formula, although they are similar by their form.

II. ISOSPECTRAL CURVES AND PERIODICAL BILLIARD TRAJECTORIES

Veselov and Moser in Ref. 9 considered the ellipsoidal billiard as a system with discrete time. Let us give a brief account of the algebro-geometric way of its integration, necessary for further exposition. After that, the connection between periodical billiard trajectories and divisors of finite order, established in Ref. 8, will be also demonstrated.

[a]Electronic mail: vladad@mi.sanu.ac.yu

 © 1998 American Institute of Physics

Let an ellipsoid in R^d be given by the equation $(Ax,x)=1$, where A is a diagonal matrix, with different eigenvalues. Let $x_0,x_1,x_2,\ldots,$ be the points of consecutive bounces of the billiard inside that ellipsoid, and $y_k=(x_k-x_{k-1})/|x_k-x_{k-1}|$ be the momenta.

The sequence (x_k,y_k) can be transformed to a sequence which represents a flow of the discrete Heisenberg XYZ model,[9] given by

$$q_{k+1}+q_{k-1}=\nu_k J^{-1}q_k, \tag{1}$$

where ν_k is such that $|q_k|=1$. Namely, using the mapping $\phi:(\underline{x},\underline{y})\to(x',\underline{y'})$ given by $x'_k=J\underline{y}_{k+1}$, $y'_k=-J^{-1}\underline{x}_k$, $J=A^{-1/2}$, and defining the sequence $(\underline{x}_k,\underline{y}_k)$ by $(\underline{x}_0,\underline{y}_0)=(x_0,y_0)$, $(\underline{x}_{k+1},\underline{y}_{k+1})=\phi(\underline{x}_k,\underline{y}_k)$, we obtain $q_k=J^{-1}\underline{x}_k$. The sequence q_k satisfies Eq. (1). Conversely, if q_k is the solution of the Heisenberg system, then the sequence $x_k=(-1)^k Jq_{2k}$ represents a trajectory of a discrete ellipsoidal billiard system.[9]

The isospectral curve, a basic tool of algebro-geometric integration, related to the Heisenberg system is found in[9]

$$\Gamma:\nu^2=\prod_{i=1}^{d-1}(\mu-\mu_i)\prod_{j=1}^{d}(\mu-J_j^2), \tag{2}$$

where $J=\mathrm{diag}(J_1,...,J_d)$, and $\mu_1,\ldots,\mu_{d-1}$ are the roots of the function

$$\phi_\mu(x,Jy)=\sum_{i=1}^{N}\frac{F_i(x,y)}{\mu-J_i^2},\quad F_i=x_i^2+\sum_{j\neq i}\frac{(x\wedge Jy)_{ij}^2}{J_i^2-J_j^2}, \tag{3}$$

with $x=q_{k-1}$, $y=q_k$. The sequence q_k can be uniquely recovered from the divisor sequence $D_{k+1}=D_k+P_\infty-P_0$ on the Jacobian variety of the curve Γ, where P_∞ corresponds to the value $\mu=\infty$, and P_0 to $\mu=0$, $\lambda=(q_k,J^{-1}q_{k+1})^{-1}$.

The connection between periodical billiard trajectories and periodic sequences of the divisors is described in following lemma.

Lemma 1: (Reference 8) If the billiard is periodic with the period n, then the divisor sequence D_k joined to the corresponding Heisenberg XYZ system is also periodic, with period 2n.

By this lemma, we need to investigate when $2n(P_\infty-P_0)=0$ on the Jacobian variety of the spectral curve Γ.

III. POINTS OF FINITE ORDER ON THE JACOBIAN VARIETY OF A HYPERELLIPTIC CURVE

The following discussion is motivated by Eq. (2) of curve Γ and the fact that P_∞ is a branch point of Γ.

Let the curve $\mathscr{C}$ be given by

$$y^2=(x-x_1)\cdots(x-x_{2g+1}),\quad x_i\neq x_j \text{ when } i\neq j.$$

It is a hyperelliptic curve of genus g. Here, we are going to find an analytic characterization of points $P\in\mathscr{C}$ which satisfy $nu(P)=0$, where $u:\mathrm{Div}(\mathscr{C})\to\mathscr{J}(\mathscr{C})$ is the Abel–Jacobi map.

Take E to be the point which corresponds to the value $x=\infty$, and choose $u(E)$ to be the neutral in $\mathscr{J}(\mathscr{C})$. According to the Abel's theorem,[10] $u(P_1)+\cdots+u(P_n)=0$ if and only if there exists a meromorphic function f with zeroes $P_1,\ldots,P_n$ and a pole of order n at the point E. Let $\mathscr{L}(nE)$ be the vector space of meromorphic functions on $\mathscr{C}$ with a unique pole E of order at most n, and $f_1,\ldots,f_k$ a basis of $\mathscr{L}(nE)$. The mapping

$$F:\mathscr{C}\to P^{k-1},\quad X\mapsto[f_1(X),\ldots,f_k(X)]$$

is a projective embedding whose image is a smooth algebraic curve of degree n. Hyperplane sections of this curve are zeroes of functions from $\mathscr{L}(nE)$. Thus, the equality $nu(P)=0$ is equivalent to

$$\operatorname{rank}\begin{bmatrix} f_1(P) & f_2(P) & \cdots & f_k(P) \\ f_1'(P) & f_2'(P) & \cdots & f_k'(P) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ f_1^{(n-1)}(P) & f_2^{(n-1)}(P) & \cdots & f_k^{(n-1)}(P) \end{bmatrix} < k. \tag{4}$$

Lemma 2: For $n \leqslant 2g$, there does not exist a point P on the curve $\mathscr{C}$, such that $nu(P)=0$ and $P \neq E$.

Proof: Let P be a point on $\mathscr{C}$, $P \neq E$, and $x=x_0$ its corresponding value.

Consider the case of n even. Since E is a branch point of a hyperelliptic curve, its Weietstrass gap sequence is $1,3,5,\ldots,2g-1$.[10] Now applying the Riemann–Roch theorem,[10] we obtain $\dim \mathscr{L}(nE)=n/2+1$. Choosing a basis $1,x,\ldots,x^{n/2}$ for $\mathscr{L}(nE)$, and substituting in (4), we come to a contradiction. □

Lemma 3: Let $P(x_0,y_0)$ be a nonbranching point on the curve $\mathscr{C}$. For $n>2g$, equality $nu(P)=0$ is equivalent to

$$\operatorname{rank}\begin{bmatrix} B_{m+1} & B_m & \cdots & B_{g+2} \\ B_{m+2} & B_{m+1} & \cdots & B_{g+3} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ B_{2m-1} & B_{2m-2} & \cdots & B_{m+g} \end{bmatrix} < m-g \quad \textit{when} \quad n=2m$$

and $\sqrt{(x-x_1)\cdots(x-x_{2g+1})}=B_0+B_1(x-x_0)+B_2(x-x_0)^2+B_3(x-x_0)^3+\cdots$. The similar condition can be written for n odd.

Proof: This claim follows from previous results by choosing a basis for $\mathscr{L}(nE)$: $1,x,\ldots,x^m,y,xy,\ldots,x^{m-g-1}y$, $n=2m$, similarly as in Ref. 2. □

IV. ANALYTICAL CHARACTERIZATION OF PERIODICAL BILLIARD TRAJECTORIES

Using the connection between periodical billiard trajectories and divisors of finite order on the spectral curve Γ, we can apply the results of Sec. III whenever Γ is a nonsingular curve.

Let the equation of the boundary be

$$\frac{x_1^2}{J_1^2}+\cdots+\frac{x_d^2}{J_d^2}=1. \tag{5}$$

Every segment of a given billiard trajectory is tangent to the same $d-1$ quadrics, confocal with the outer ellipsoid. Suppose their equations are:

$$\frac{x_1^2}{J_1^2-\alpha_i}+\cdots+\frac{x_d^2}{J_d^2-\alpha_i}=1 \quad (1 \leqslant i \leqslant d-1). \tag{6}$$

Functions F_i defined by (3) are integrals of the billiard. Moreover, they have the same values for all trajectories which share the same caustic hypersufaces.[11] The roots of ϕ_μ are equal to $\alpha_1,\ldots,\alpha_{d-1}$,[12] which can be obtained by direct calculation choosing a line parallel to one of the coordinate axes to be the first segment of the trajectory.

Proposition: The curve Γ is singular if and only if the billiard trajectory is placed in one of coordinate hyperplanes.

Proof: By assumption, $J_1^2,\ldots,J_d^2$ are mutually different. Chasles' theorem[13] states that every line in d-dimensional space is tangent to $d-1$ quadrics from a given confocal family. Thus $\alpha_1,\ldots,\alpha_{d-1}$ are also all different.

Γ is singular if and only if $\alpha_i=J_j^2$ for some i,j, i.e., if some of the caustic quadrics is degenerate. This takes place if the first segment of the trajectory is in one of the coordinate hyperplanes (because of symmetry, then the whole trajectory has to be placed in this hyperplane). □

Summarize the obtained results as follows.

Theorem 1. The condition on a billiard trajectory inside ellipsoid (5) in R^d, with nondegenerate caustics (6), to be perodic with period $n \geqslant d$ is

$$\operatorname{rank}\begin{bmatrix} B_{n+1} & B_n & \cdots & B_{d+1} \\ B_{n+2} & B_{n+1} & \cdots & B_{d+2} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ B_{2n-1} & B_{2n-2} & \cdots & B_{n+d-1} \end{bmatrix} < n-d+1,$$

where $\sqrt{(x-\alpha_1)\cdots(x-\alpha_{d-1})(x-J_1^2)\cdots(x-J_d^2)} = B_0 + B_1 x + B_2 x^2 + \cdots$ *. There are no such trajectories with period less than d.*

Proof: The trajectory is periodic with period n if, by lemma 1, the corresponding divisor sequence on the curve Γ has the period $2n$, i.e., $2n(P_\infty - P_0) = 0$ on $\mathscr{J}(\mathscr{C})$. Curve Γ is hyperelliptic with genus $g = d-1$. Taking $u(P_\infty)$ to be the neutral on $\mathscr{J}(\mathscr{C})$ we get desired results by applying lemmas 2 and 3. □

This is a sort of generalization of Cayley's condition. It is very similar, but even in the case $d=2$, not exactly the same as Cayley's formula. Theorem 1 does not give a complete description of periodical billiard trajectories. For example, it is obvious that there always exists a trajectory with period 2—with both bounces on the same coordinate axis.

Theorem 2: *If the billiard trajectory inside the ellipsoid (5) is periodic with period $n<d$, then all the bounces are placed in one of the coordinate n-dimensional planes.*

Proof: This follows immediately from the proposition, theorem 1, and the fact that the section of the confocal family of quadrics with a coordinate hyperplane is again a confocal family. □

All periodic trajectories of the billiard system either satisfy the condition from Theorem 1 and have at least d bounces, or they are placed in a coordinate k-plane. The analytic condition in the latter case can be obtained by appropriate application of Theorem 1.

ACKNOWLEDGMENT

This research was supported by the Ministry for Science and Technology of the Republic of Serbia, Project No. 04M03.

[1] M. Berger, *Géométrie* (Gauthier-Villars, Paris, 1972).
[2] H. Lebesgue, *Les Coniques* (Gauthier-Villars, Paris, 1942), pp. 115–149.
[3] P. Griffiths and J. Harris, ''On Cayley's explicit solution to Poncelet's porism,'' L'Enseignement Math. **24**, 31–40 (1978).
[4] P. Griffiths and J. Harris, ''A Poncelet theorem in space,'' Comment. Math. Helvetici **52**, 145–160 (1977).
[5] S.-J. Chang and R. Friedberg, ''Elliptical billiards and Poncelet's theorem,'' J. Math. Phys. **29**, 1537–1550 (1988).
[6] S.-J. Chang, B. Crespi, and K.-J. Shi, ''Elliptical billiards and hyperelliptic functions,'' J. Math. Phys. **34**, 2257–2289 (1993).
[7] S.-J. Chang, B. Crespi, and K.-J. Shi, ''Elliptical billiard systems and the full Poncelet's theorem in *n* dimensions,'' J. Math. Phys. **34**, 2242–2256 (1993).
[8] V. Dragović and M. Radnović, ''Conditions of Cayley's type for ellipsoidal billiard,'' J. Math. Phys. **39**, 355–362 (1998).
[9] J. Moser and A. P. Veselov, ''Discrete versions of some classical integrable systems and factorization of matrix polynomials,'' Commun. Math. Phys. **139**, 217–243 (1991).
[10] R. C. Gunning, *Lectures on Riemann Surfaces* (Princeton University Press, Princeton, NJ, 1966).
[11] A. P. Veselov, ''Integrable systems with discrete time and difference operators,'' Funct. Anal. Appl. **22**, 1–13 (1988).
[12] J. Moser, ''Various aspects of integrable hamiltonian systems,'' Newhouse Dynamical Systems, C.I.M.E. Lectures, Bressanone, Italy, 1978.
[13] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics* (Nauka, Moskva, 1989).

Electromagnetic fields in air of traveling-wave currents above the earth

Dionisios Margetis[a)]
Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138-2901

(Received 12 November 1996; accepted for publication 10 July 1998)

The problem of the electromagnetic field created by a thin, straight conductor of infinite length carrying a forward traveling-wave current with a complex propagation constant γ above a homogeneous and isotropic planar earth of wave number k_1 is formulated in terms of contour integrals. In the limit where γ becomes equal to the free-space wave number k_0, the component of the magnetic field in air normal to both conductor and interface is evaluated in closed form in terms of known special functions while the remaining components of the field in air are expressed as series expansions in $\delta=k_0^2(k_1^4-k_0^4)^{-1/2}$ via the application of a contour integration technique. The new analytical formulas involve familiar transcendental functions and are valid at any distance from the source. The analysis sheds light on the intricate nature of low-frequency electromagnetic fields generated by transmission lines in the presence of a conducting or a dielectric half-space. © *1998 American Institute of Physics.* [S0022-2488(98)02811-4]

I. INTRODUCTION

The determination of the propagation modes for the current and evaluation of the ensuing electromagnetic field of a thin, infinitely long straight wire immersed in a stratified medium is a fundamental and fairly old problem in electromagnetic theory. Of particular interest is the case in which the conductor is placed parallel to an isotropic, homogeneous half-space. In the thin-wire approximation, the field is computed with the assumption that the current is axial and concentrated in a line at the center of the cross section of the wire. Early investigations[1,2] employed approximate transmission-line theory and were essentially limited to a relatively dense neighboring medium and to distances from the source that are short compared to the medium's wavelength. In later analyses[3–6] the field components of a conductor carrying an exponential current were expressed in terms of Fourier integrals in the direction normal to the conductor and parallel to the interface, with the tacit or explicit assumption that the field tends to zero when the distance perpendicular to the wire approaches infinity. In these studies, it has not been possible to evaluate all requisite integrals of the so-called Sommerfeld type, even for special values of the propagation constant γ for the current.

Coleman[7] was probably the first to point out that in the limit $a/d\to 0$, where a is the radius of the wire and d is the distance from the interface, the modal equation yields a solution for γ which approaches the wave number k_0 of the ambient medium. Approximate analytical derivations through disparate mathematical methods by Chang and Wait[4] and by King *et al.*[8] with the condition $0<k_0d\ll 1$ and for a relatively dense neighboring medium have corroborated and refined Coleman's[7] findings when $a\ll d$. For the air-earth configuration, deviations from the value $\gamma=k_0$ result then in negligible deviations of the fields in air for all distances of interest at extremely low frequencies. In the case of two or more thin, parallel conductors the condition of a vanishing total current reinforces the reasoning for the assumption $\gamma=k_0$ because of the confinement of the field closer to the source.[9,10] Notably, γ changes drastically when the wire approaches the vicinity of the boundary ($d\sim a$) because of the proximity effect of the earth's surface.[4,7,9] Investigation of the modal equation for a wide range of frequencies is a formidable task.[11,12]

In a recent paper by King and Wu[10] simple approximate formulas for the field were derived under the condition $k_0^2\ll|k_1^2|$ via integration of the approximate lateral-wave formulas for the field

a)Electronic mail: margetis@huhepl.harvard.edu

 © 1998 American Institute of Physics

of a horizontal electric dipole,[13] where k_1 is the wave number of the earth. However, as pointed out in Ref. 10, the violation of the requisite condition $|k_1 r'|>1$ at extremely low frequencies, where r' is the distance from the dipole, introduced an inaccuracy for the axial component of the electric field.

This paper has a twofold purpose. The first is to outline a procedure for obtaining integral representations for all components of the electromagnetic field under the thin-wire approximation in the general case when the current in the x-directed conductor at a fixed frequency ω is of the form $e^{i\gamma x - i\omega t}$, where Re $\gamma>0$, Im $\gamma\geqslant 0$, by invoking the condition that the field be described as a superposition of outgoing waves in the direction normal to both conductor and interface. This in turn leads to a radiation field which is exponentially increasing in the direction normal to the wire when Im $\gamma>0$ and the surrounding medium is a perfect dielectric, such as air, in agreement with the studies on the theory of the microstrip.[14–17] This point is discussed further in Sec. IV.

The second purpose is to evaluate exactly the requisite integrals for the field in air at any distance from the source in the limit $\gamma\to k_0$ by relaxing the condition $k_0^2\ll|k_1^2|$, where k_1 is the wave number of the adjacent medium. This is achieved by applying a contour integration technique. The new integrated formulas involve series expansions in δ^2 for $|\delta|<1$, where δ is expressed as $\delta=\eta_S/k$, η_S being the limit as $\gamma\to k_0$ of the Sommerfeld pole and $k=\sqrt{k_1^2-k_0^2}$, with series coefficients that depend on known special functions. These series are shown to converge uniformly in distance. In certain limiting cases the formulas in question reduce to simplified expressions, such as the well known Carson's series,[1] that have been previously derived by different approximate means. It is emphasized that the proposed treatment is not intended to serve as a substitute for previous simple approximate results; it rather aims at demonstrating rigorously exact solubility of the model in question with the removal of certain restrictions on the physical parameters. Consequently, stringent conditions for the validity of practically appealing simplifications can follow.

Compared with actual current-carrying wires the main idealizations involved here are (1) the current-carrying conductor in air is infinitely long and thin ($a\ll d$); (2) the air-earth interface is planar; (3) the source current has the dependence $e^{i\gamma x}$ with the distance x along the wire. The limit $\gamma\to k_0$ corresponds to the transverse electromagnetic (TEM) mode in air in the absence of the earth. This propagation constant reasonably describes the slow-wave currents in actual multiphase transmission lines above the earth.[9,10] The $e^{-i\omega t}$ time dependence is suppressed throughout the analysis.

II. FORMULATION

The geometry of the problem is shown in Fig. 1. It consists of an infinitely long x-directed conductor lying in the vertical plane $y=0$ in the air (region 0, $z>0$) at height d above the surface ($z=0$) of an isotropic, homogeneous and nonmagnetic earth (region 1, $z<0$). The associated current density is assumed to be

$$\mathbf{J}(\mathbf{r})=e^{i\gamma x}\delta(y)\,\delta(z-d)\hat{\mathbf{x}},\quad \text{Re }\gamma>0,\quad \text{Im }\gamma>0, \tag{1}$$

where $\mathbf{r}=(x,y,z)$. At this point, it is not advisable to employ the usual spatial Fourier transform in y of the field since the requisite integrability condition for $|y|\to\infty$ is not necessarily met. A remedy to this problem is to seek a meaningful integral representation of the form

$$\mathbf{F}_j(x,y,z)=\frac{1}{2\pi}\int_\Gamma d\eta\,\overline{\mathbf{F}}_j(x,\eta,z)e^{i\eta y}, \tag{2}$$

where $\mathbf{F}_j=\mathbf{E}_j$, $\mathbf{B}_j$ $(j=0,1)$ denote the electric and magnetic field, respectively, in each region 0, 1 and Γ is a properly chosen infinite integration path with horizontal asymptotes in the η-plane. Specifically, when a scalar integral coefficient becomes 1, the respective integral yields the Dirac delta function $\delta(y)$. Conventional inversion of the integral formula of Eq. (2) is not generally possible.

Formally, with a dependence of the form

$$\mathbf{F}_j(\mathbf{r})=e^{i\gamma x}\mathbf{f}_j(y,z), \tag{3}$$

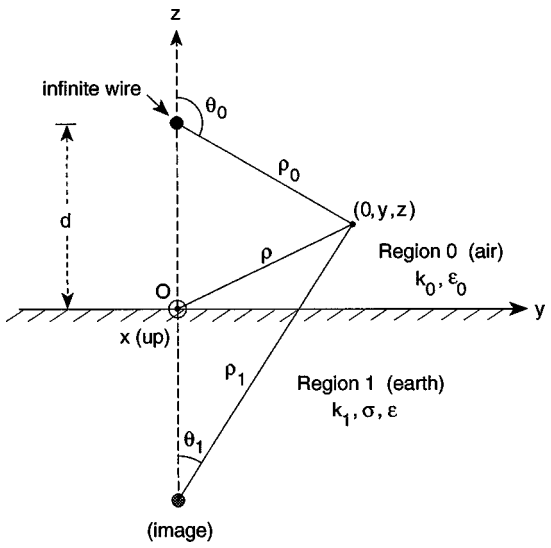

FIG. 1. Cross section of an infinitely thin, infinitely long conductor above the earth.

$\mathbf{E}_j$ and $\mathbf{B}_j$ satisfy Maxwell's equations if the coefficients $\overline{\mathbf{f}}_j=\overline{\mathbf{e}}_j$, $\overline{\mathbf{b}}_j$ $(j=0,1)$ obey the following equations:

$$\overline{e}_{jy}=\frac{i}{k_j^2-\gamma^2}\left[i\eta\gamma\overline{e}_{jx}+\omega\frac{\partial\overline{b}_{jx}}{\partial z}\right], \tag{4}$$

$$\overline{e}_{jz}=\frac{i}{k_j^2-\gamma^2}\left[\gamma\frac{\partial\overline{e}_{jx}}{\partial z}-i\eta\omega\overline{b}_{jx}\right], \tag{5}$$

$$\overline{b}_{jy}=-\frac{1}{k_j^2-\gamma^2}\left[\eta\gamma\overline{b}_{jx}+\frac{ik_j^2}{\omega}\frac{\partial\overline{e}_{jx}}{\partial z}\right], \tag{6}$$

$$\overline{b}_{jz}=\frac{1}{k_j^2-\gamma^2}\left[i\gamma\frac{\partial\overline{b}_{jx}}{\partial z}-\frac{\eta k_j^2}{\omega}\overline{e}_{jx}\right], \tag{7}$$

$$\frac{\partial^2\overline{e}_{jx}}{\partial z^2}+\gamma_j^2\overline{e}_{jx}=-\frac{i\omega\mu_0}{k_j^2}(k_j^2-\gamma^2)\,\delta(z-d), \tag{8}$$

$$\frac{\partial^2\overline{b}_{jx}}{\partial z^2}+\gamma_j^2\overline{b}_{jx}=0, \tag{9}$$

where

$$\gamma_j^2=k_j^2-\gamma^2-\eta^2. \tag{10}$$

The η-Riemann surface for $\overline{\mathbf{e}}_j$, $\overline{\mathbf{b}}_j$ consists of four sheets. For definiteness, the first Riemann sheet $\mathscr{R}_1$ is defined as

$$\forall\,\eta\in\mathscr{R}_1,\quad -\pi/2<\arg\,\gamma_j\leqslant\pi/2. \tag{11}$$

The corresponding branch cuts are parts of hyperbolas, as depicted in Fig. 2. Final expressions for $\overline{\mathbf{e}}_j$, $\overline{\mathbf{b}}_j$ are uniquely determined by imposition of:

(1) Outgoing waves in z; this excludes solutions of the form $e^{-i\gamma_0 z}$ for $z>d$ and $e^{i\gamma_1 z}$ for $z<0$ in Eqs. (8) and (9).

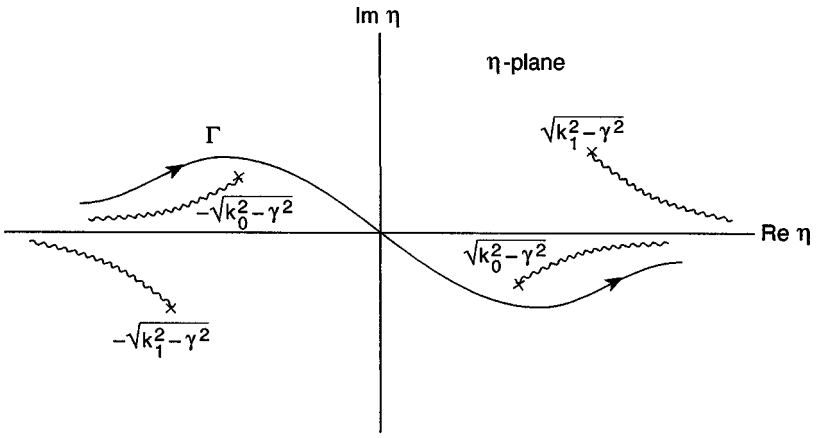

FIG. 2. Integration path Γ in the complex η-plane for the electric and magnetic field of an infinitely long wire in air (wave number k_0) carrying an $e^{i\gamma x}$, Re $\gamma>0$, Im $\gamma>0$, current over the earth (wave number k_1).

(2) Continuity of $\overline{e}_{jx}$, $\overline{b}_{jx}$, $\overline{e}_{jy}$, $\overline{b}_{jy}$ at $z=0$.

Accordingly, for $z>0$ the axial components read

$$\overline{e}_{0x}=-\omega\mu_0\frac{\gamma_1(k_0^2-\gamma^2)+\gamma_0(k_1^2-\gamma^2)}{K(\eta)L(\eta)}e^{i\gamma_0(z+d)}+\frac{i\omega\mu_0}{\gamma_0 k_0^2}(k_0^2-\gamma^2)e^{i\gamma_0 z_>}\sin(\gamma_0 z_<), \tag{12}$$

$$\overline{b}_{0x}=\mu_0\frac{\eta\gamma(k_1^2-k_0^2)}{K(\eta)L(\eta)}e^{i\gamma_0(z+d)}, \tag{13}$$

while for $z<0$

$$\overline{e}_{1x}=-\omega\mu_0\frac{\gamma_1(k_0^2-\gamma^2)+\gamma_0(k_1^2-\gamma^2)}{K(\eta)L(\eta)}e^{i\gamma_0 d}e^{-i\gamma_1 z}, \tag{14}$$

$$\overline{b}_{1x}=\mu_0\frac{\eta\gamma(k_1^2-k_0^2)}{K(\eta)L(\eta)}e^{i\gamma_0 d}e^{-i\gamma_1 z}, \tag{15}$$

where

$$K(\eta)=\gamma_0+\gamma_1, \tag{16a}$$

$$L(\eta)=k_1^2\gamma_0+k_0^2\gamma_1. \tag{16b}$$

In the above, $z_>$ is the larger one of z and d and $z_<$ is the smaller one. Similar formulas for the remaining components are obtained directly from Eqs. (4)–(7). If γ is considered to be real, $\overline{\mathbf{e}}_j$ and $\overline{\mathbf{b}}_j$ designate the two-dimensional spatial Fourier transforms in x,y of the projections in x of the tensor Green's functions for the electric and magnetic field. From Eq. (16b) it is inferred that when $\gamma\neq k_0k_1/\sqrt{k_0^2+k_1^2}$, $L(\eta)$ has four simple zeros in the η-Riemann surface which are the Sommerfeld poles:

$$\eta=\pm\eta_S=\pm\sqrt{\frac{k_0^2k_1^2}{k_0^2+k_1^2}-\gamma^2}. \tag{17}$$

In contrast, $K(\eta)$ is free of zeros in any finite region of this surface. It is noteworthy that the $\overline{b}_{jz}$ are free of the $L(\eta)$ denominator.

An integration path Γ which is symmetric under inversion through the origin is subsequently chosen so that: (1) It lies entirely in $\mathscr{R}_1$. (2) All integrals are absolutely convergent. (3) The positive and negative real axes are asymptotes of the path. Among these requirements, (1) and (2)

suffice to determine the field. The chosen contour is shown in Fig. 2. Integral representations for the field follow from Eq. (2). These satisfy Maxwell's equations in each region 0, 1 along with the prescribed boundary conditions at $z=0$ and are expressed as a Fourier superposition of outgoing waves in z.

In the limit $z\to\infty$, the principal contribution to integration arises from the vicinity of $\eta=0$, yielding a leading term that is exponentially growing in z. Similar asymptotic behavior can be found via the method of steepest descents[18] when $\sqrt{y^2+z^2}\to\infty$ for $z>0$, as outlined in Appendix A.

Evidently, the integral representations may be continued to $\gamma=k_0$. The respective field tends to zero when $\sqrt{y^2+z^2}\to\infty$.

III. THE CASE $\gamma=k_0$

A. Integral representations for the field in air

In the limit $\gamma\to k_0$, the integration path Γ in Eq. (2) can be deformed to coincide with the entire real axis. By use of Eqs. (2)–(7), (12) and (13), and evaluation of the elementary integrals, it is straightforward to arrive at the following expressions for the fields $\mathbf{E}_0$, $\mathbf{B}_0$:

$$E_{0x}=-\frac{\omega\mu_0}{2\pi}\,e^{ik_0x}\int_{-\infty}^{\infty}d\eta e^{i\eta y}e^{-|\eta|(z+d)}\left[\frac{1}{i|\eta|+\sqrt{k^2-\eta^2}}-\frac{k_0^2}{k_0^2\sqrt{k^2-\eta^2}+ik_1^2|\eta|}\right] \tag{18a}$$

$$=\frac{i\omega\mu_0}{2\pi}\,e^{ik_0x}[\Phi_+(y,z;1)-(k_0^2/k_1^2)\Phi_+(y,z;k_0^2/k_1^2)], \tag{18b}$$

$$E_{0y}=\frac{\omega\mu_0k_0}{2\pi}\,e^{ik_0x}\int_{-\infty}^{\infty}d\eta e^{i\eta y}e^{-|\eta|d}\eta\left[\frac{-i\,\sinh(\eta z)}{\eta k_0^2}+\frac{e^{-|\eta|z}}{ik_1^2|\eta|+k_0^2\sqrt{k^2-\eta^2}}\right],\quad 0<z<d \tag{19a}$$

$$=\frac{\mu_0c}{2\pi}\,e^{ik_0x}\left[\left(\frac{y}{\rho_0^2}-\frac{y}{\rho_1^2}\right)+k\,\frac{k_0^2}{k_1^2}\,\Psi_-(y,z;k_0^2/k_1^2)\right], \tag{19b}$$

$$E_{0z}=\frac{\omega\mu_0k_0}{2\pi}\,e^{ik_0x}\int_{-\infty}^{\infty}d\eta e^{i\eta y}e^{-|\eta|d}\left[-\frac{\cosh(\eta z)}{k_0^2}+\frac{\sqrt{k^2-\eta^2}}{ik_1^2|\eta|+k_0^2\sqrt{k^2-\eta^2}}\,e^{-|\eta|z}\right],\quad 0<z<d \tag{20a}$$

$$=\frac{\mu_0c}{2\pi}\,e^{ik_0x}\left[\left(\frac{z-d}{\rho_0^2}-\frac{z+d}{\rho_1^2}\right)-ik\,\frac{k_0^2}{k_1^2}\,U(y,z;k_0^2/k_1^2)\right], \tag{20b}$$

$$B_{0x}=-i\,\frac{\mu_0k_0}{2\pi}\,e^{ik_0x}\int_{-\infty}^{\infty}d\eta e^{i\eta y}e^{-|\eta|(z+d)}\,\mathrm{sgn}(\eta)\left[\frac{1}{\sqrt{k^2-\eta^2}+i|\eta|}-\frac{k_0^2}{ik_1^2|\eta|+k_0^2\sqrt{k^2-\eta^2}}\right] \tag{21a}$$

$$=\frac{\mu_0\omega}{2\pi c}\,e^{ik_0x}[\Phi_-(y,z;1)-(k_0^2/k_1^2)\Phi_-(y,z;k_0^2/k_1^2)], \tag{21b}$$

$$B_{0y}=\frac{\mu_0}{2\pi}\,e^{ik_0x}\int_{-\infty}^{\infty}d\eta e^{i\eta y}e^{-|\eta|d}\left[\cosh(\eta z)+\frac{k_1^2\eta^2-k_0^2(k^2+i|\eta|\sqrt{k^2-\eta^2})}{(i|\eta|+\sqrt{k^2-\eta^2})(ik_1^2|\eta|+k_0^2\sqrt{k^2-\eta^2})}\,e^{-|\eta|z}\right],\quad 0<z<d \tag{22a}$$

$$= -\frac{\mu_0}{2\pi} e^{ik_0x}\left[\left(\frac{z-d}{\rho_0^2} - \frac{z+d}{\rho_1^2}\right) - ikW(y,z;k_0^2/k_1^2)\right], \tag{22b}$$

$$B_{0z} = \frac{\mu_0}{2\pi} e^{ik_0x}\int_{-\infty}^{\infty} d\eta e^{i\eta y} e^{-|\eta|d}\eta\left[\frac{-i\ \sinh(\eta z)}{\eta} + \frac{e^{-|\eta|z}}{i|\eta| + \sqrt{k^2-\eta^2}}\right], \quad 0<z<d \tag{23a}$$

$$= \frac{\mu_0}{2\pi} e^{ik_0x}\left[\left(\frac{y}{\rho_0^2} - \frac{y}{\rho_1^2}\right) + k\Psi_-(y,z;1)\right], \tag{23b}$$

where

$$k = \sqrt{k_1^2 - k_0^2}, \quad \text{Im } k \geqslant 0, \tag{24}$$

$c = (\epsilon_0\mu_0)^{-1/2}$ is the velocity of light, $\text{sgn}(\eta)$ is the sign function

$$\text{sgn}(\eta) = \begin{cases} 1, & \eta > 0 \\ 0, & \eta = 0 \\ -1, & \eta < 0, \end{cases}$$

and, with the introduction of a new variable $s = ik^{-1}|\eta|$,

$$\Phi_\pm(y,z;\beta) = \varphi(\alpha_+;\beta) \pm \varphi(\alpha_-;\beta), \quad \Psi_\pm(y,z;\beta) = \psi(\alpha_+;\beta) \pm \psi(\alpha_-;\beta), \tag{25}$$

$$U(y,z;\beta) = u(\alpha_+;\beta) + u(\alpha_-;\beta) = \frac{1}{\beta}\left(\frac{1}{\alpha_+} + \frac{1}{\alpha_-}\right) - \frac{1}{\beta}\Psi_+(y,z;\beta), \tag{26}$$

$$W(y,z;\beta) = w(\alpha_+;\beta) + w(\alpha_-;\beta) = \Psi_+(y,z;1) - (1+\beta)\Psi_+(y,z;\beta) + \left(\frac{1}{\alpha_+} + \frac{1}{\alpha_-}\right), \tag{27}$$

$$\varphi(\alpha;\beta) = \int_0^{\infty e^{i\tau}} ds e^{-\alpha s}\frac{1}{s + \beta\sqrt{s^2+1}}, \tag{28}$$

$$\psi(\alpha;\beta) = \int_0^{\infty e^{i\tau}} ds e^{-\alpha s}\frac{s}{s + \beta\sqrt{s^2+1}} = -\frac{\partial}{\partial\alpha}\varphi(\alpha;\beta), \tag{29}$$

$$u(\alpha;\beta) = \int_0^{\infty e^{i\tau}} ds e^{-\alpha s}\frac{\sqrt{s^2+1}}{s + \beta\sqrt{s^2+1}} = \frac{1}{\beta}\left[\frac{1}{\alpha} - \psi(\alpha;\beta)\right], \tag{30}$$

$$w(\alpha;\beta) = \int_0^{\infty e^{i\tau}} ds e^{-\alpha s}\frac{s^2 + \beta(1 + s\sqrt{1+s^2})}{(s+\sqrt{1+s^2})(s+\beta\sqrt{1+s^2})} = \psi(\alpha;1) - (1+\beta)\psi(\alpha;\beta) + \frac{1}{\alpha}, \tag{31}$$

$$\alpha_\pm = -ik(z+d\pm iy) = -ik\rho_1 e^{\pm i\theta_1}, \quad -\pi/2 < \theta_1 < \pi/2,$$

$$\rho_0 = \sqrt{(z-d)^2+y^2}, \quad \rho_1 = \sqrt{(z+d)^2+y^2}, \quad \sin\theta_{0,1} = y/\rho_{0,1}, \quad z>0, \tag{32}$$

$$\tau = \frac{\pi}{2} - \arg(k). \tag{33}$$

The integral formulas in the first expressions, Eqs. (18a)–(23a), are invertible.[19] The restriction $0<z<d$ has been removed in the second expressions, Eqs. (18b)–(23b), since the singular term $1/\rho_0^2$ is extracted by direct integration and the ensuing representations may be continued to $z \geqslant d$. In Eq. (29), the interchange of the order of differentiation and integration is legitimate because the integrals are absolutely and uniformly convergent.

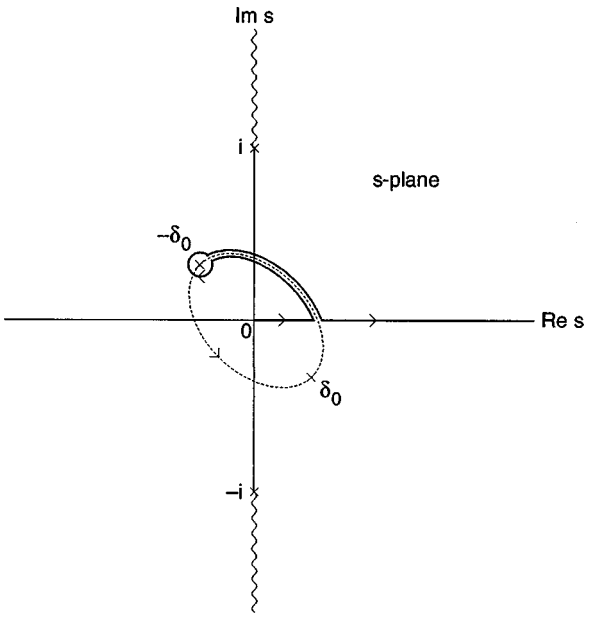

FIG. 3. Branch-cut configuration in the s-plane for the integral of Eq. (34). The modified path serves the analytic continuation of $\varphi(\alpha;\beta)$ along the respective contour enclosing the origin in the β-plane.

The quantities $\Phi_{\pm}$, $\Psi_{\pm}$, U, W introduced above designate solely the corrections to the solution for a perfectly conducting earth ($k_1 \to \infty$). The form of the solution above verifies that for a current supporting a TEM mode in air the x-components of the field satisfy Laplace's equation in y,z in the simply connected region $\{(y,z): z>0\}$ whereas the rest of the components satisfy the Poisson equation in the same region.

An examination of Eqs. (25)–(31) indicates that only the integrations for $\varphi(\alpha;\beta)$, where $\beta = 1$ and k_0^2/k_1^2, need to be carried out.

B. Evaluation of φ, ψ, u, w

For mathematical convenience it is assumed that $\alpha>0$. The extension over the range of complex values of α for which $|\arg \alpha|<\pi$ is attained via analytic continuation. The principal integral reads

$$\varphi(\alpha;\beta)=\int_0^{+\infty} ds e^{-\alpha s} \frac{1}{s+\beta\sqrt{s^2+1}}, \tag{34}$$

where the integration path coincides with the entire positive real axis. The integrand in Eq. (34) has two branch points at $s=\pm i$. A branch of the square root is chosen such that $\sqrt{1+s^2}>0$ for $s>0$. The first Riemann sheet is subsequently defined by taking the branch cuts along the positive and negative imaginary axis, as shown in Fig. 3. A study of the cases $\beta=1$ and $\beta=k_0^2/k_1^2$ essentially evinces the differences introduced by the denominators $K(\eta)$ and $L(\eta)$ of Eqs. (16a) and (16b), respectively. These two possibilities need to be considered separately.

(i) $\beta=1$. In this case, $\varphi(\alpha;\beta)$ is cast in the form

$$\varphi(\alpha;1)=\int_0^{+\infty} ds e^{-\alpha s}(\sqrt{1+s^2}-s)=\frac{1}{\alpha} S_{1,1}(\alpha)-\frac{1}{\alpha^2}. \tag{35}$$

In the above, $S_{\nu,\nu}(\alpha)$ $(\nu=1)$ denotes Lommel's function

$$S_{\nu,\nu}(\alpha)=2^{\nu-1}\pi^{1/2}\Gamma\left(\nu+\frac{1}{2}\right)[\mathbf{H}_\nu(\alpha)-Y_\nu(\alpha)], \tag{36}$$

where $\mathbf{H}_\nu(\alpha)$ and $Y_\nu(\alpha)$ are Struve's function and Neumann's function, respectively, from p. 18 of Ref. 20.

(ii) $\beta=k_0^2/k_1^2$. In this case, it has not been possible to obtain similar expressions in closed form. In particular, in the limit $\beta\to 0$ the integral of Eq. (34) becomes divergent. For $\beta\neq 0$, the denominator of the integrand vanishes to first order at

$$s=\pm\delta, \quad \delta=\beta/(1-\beta^2)^{1/2}=k_0^2/(k_1^4-k_0^4)^{1/2}, \tag{37}$$

which implies the existence of two simple poles in the s-Riemann surface. For definiteness, it is assumed that $\delta>0$ for $0<\beta<1$. Accordingly, the pole $s=-\delta$ alone is present in the selected Riemann sheet.

The analytic continuation of $\varphi(\alpha;\beta)$, qua function of β, along any simple closed curve $\beta=\beta(t)$ enclosing the origin in the β-plane is carried out through deformation of the integration path in the s-plane in order that the pole $s=-\delta(\beta(t))$ does not cross the modified path. In particular, when the initial point is $\beta=\beta_0$ and the contour is described once in the counterclockwise sense in the β-plane, the s-integration picks up the residue at $s=-\delta(\beta_0)\equiv-\delta_0$, as shown in Fig. 3:

$$\varphi(\alpha;\beta_0 e^{i2\pi})-\varphi(\alpha;\beta_0)=-2\pi i\,\mathrm{Res}_{s=-\delta_0}\left[\frac{e^{-\alpha s}}{s+\beta_0\sqrt{s^2+1}}\right]=-2\pi i(1+\delta_0^2)e^{\alpha\delta_0}. \tag{38}$$

This calculation indicates the existence of a logarithmic singularity at $\beta=0$. Accordingly, the task is assigned of expanding the regular part of $\varphi(\alpha;\beta(\delta))$ in powers of δ in the unit disc $\mathscr{D}=\{\delta\in\mathbb{C}:|\delta|<1\}$.

In order to extract the singular contribution, it is desirable to write

$$\frac{1}{s+\beta\sqrt{s^2+1}}=\frac{1+\delta^2}{s+\delta}-\delta\sqrt{1+\delta^2}\,\frac{1}{\sqrt{s^2+1}+\sqrt{1+\delta^2}}, \tag{39}$$

which in turn leads to

$$\varphi(\alpha;\beta)=-(1+\delta^2)e^{\delta\alpha}\mathrm{Ei}(-\delta\alpha)-\delta\sqrt{1+\delta^2}\,\varphi_1(\alpha;\delta), \tag{40}$$

where

$$\varphi_1(\alpha;\delta)=\int_0^\infty ds e^{-\alpha s}\,\frac{1}{\sqrt{s^2+1}+\sqrt{1+\delta^2}}, \tag{41}$$

and $\mathrm{Ei}(-z)$ is the exponential integral defined on p. 267 of Ref. 21 as

$$\mathrm{Ei}(-z)=-\int_z^\infty dt e^{-t}t^{-1}. \tag{42}$$

The integrand in Eq. (41) no longer admits any pole in the first Riemann sheet. By inspection, $\varphi_1(\alpha;\delta)$ is holomorphic in $\mathscr{D}$. It is noted in passing that

$$e^z\mathrm{Ei}(-z)=\sum_{n=0}^{N-1}\frac{(-1)^{n+1}}{n!}z^{-1-n}+\mathrm{O}(|z|^{-N-1}) \quad \text{for } z\to\infty, \quad N=1,2,\ldots,|\arg(z)|<3\pi/2, \tag{43}$$

$$\mathrm{Ei}(-z)=\gamma+\ln z+\sum_{n=1}^{\infty}\frac{(-1)^n z^n}{n!n}, \quad \mathrm{Ei}(-ze^{i2\pi})=\mathrm{Ei}(-z)+2\pi i, \tag{44}$$

where $\gamma=0.5772\ldots$ is Euler's constant.

For the purpose of obtaining a series expansion of $\varphi_1(\alpha;\delta)$ in δ^2, $\varphi_1(\alpha;\delta)$ is expressed as a contour integral, namely,

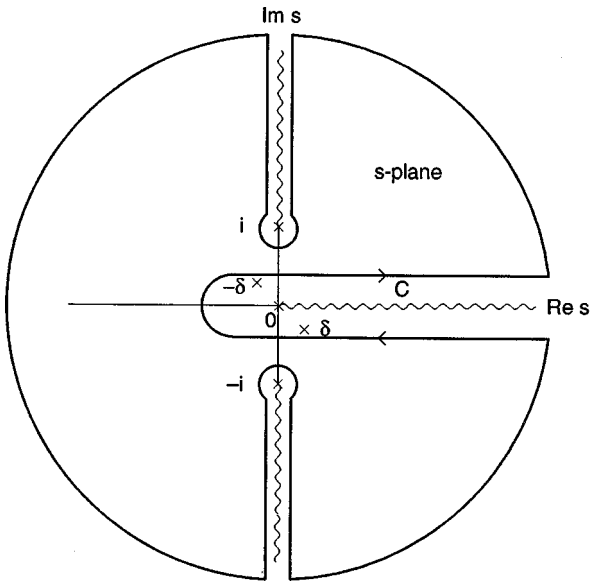

FIG. 4. Branch-cut configuration and contour of integration C in the s-plane for the integral of Eq. (45).

$$\varphi_1(\alpha;\delta)=-\frac{1}{2\pi i}\int_C ds e^{-\alpha s}\mathrm{Ei}(-\alpha s e^{-i\pi})\,\frac{\sqrt{s^2+1}-\sqrt{1+\delta^2}}{s^2-\delta^2}, \tag{45}$$

in view of Eq. (44). A branch of $\mathrm{Ei}(-\alpha s e^{-i\pi})$ is chosen such that $0\leqslant \arg(s)<2\pi$. Contour C is traversed in the clockwise sense and its interior contains the points $s=\pm\delta$, as depicted in Fig. 4. By virtue of

$$\frac{1}{s^2-\delta^2}=\sum_{l=0}^{L-1}\delta^{2l}s^{-2l-2}+\delta^{2L}\,\frac{s^{-2L}}{s^2-\delta^2}, \tag{46}$$

where L is any fixed positive integer, Eq. (45) is rearranged to give

$$\begin{aligned}\varphi_1(\alpha;\delta)=&\sum_{l=0}^{L-1}\delta^{2l}\left(-\frac{1}{2\pi i}\right)\int_C ds e^{-\alpha s}\mathrm{Ei}(-\alpha s e^{-i\pi})s^{-2l-2}\left[\sqrt{1+s^2}-\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}s^{2m}\right]\\&+\sum_{l=0}^{L-1}\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}\delta^{2l}\left(-\frac{1}{2\pi i}\right)\int_C ds e^{-\alpha s}\mathrm{Ei}(-\alpha s e^{-i\pi})s^{2m-2l-2}\\&+\sqrt{1+\delta^2}\,\frac{1}{2\pi i}\int_C ds e^{-\alpha s}\mathrm{Ei}(-\alpha s e^{-i\pi})\,\frac{1}{s^2-\delta^2}\\&-\delta^{2L}\,\frac{1}{2\pi i}\int_C ds e^{-\alpha s}\mathrm{Ei}(-\alpha s e^{-i\pi})s^{-2L}\,\frac{\sqrt{s^2+1}}{s^2-\delta^2}.\end{aligned} \tag{47}$$

In regard to the first sum of the preceding expression, the integration may be performed using the positive real axis since the integrands are all integrable at $s=0,\infty$. The double sum and the third term are zero because the respective integrals can be indented at infinity and no singularity lies within the resulting contour. The fourth term represents the remainder of the summation and is evaluated by folding the contour around the branch cuts in the imaginary axis, as shown in Fig. 4. Accordingly, $\varphi_1(\alpha;\delta)$ reads

$$\varphi_1(\alpha;\delta)=\sum_{l=0}^{L-1}\delta^{2l}q_{2l+2}(\alpha)+R_L(\alpha;\delta), \tag{48}$$

where

$$q_{2l+2}(\alpha)=\int_0^\infty ds e^{-\alpha s}s^{-2l-2}\left[\sqrt{1+s^2}-\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}s^{2m}\right] \tag{49a}$$

$$=(-1)^l\frac{1}{i\pi}\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)-e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2l-2}\sqrt{\xi^2-1}, \tag{49b}$$

and

$$R_L(\alpha;\delta)=(-1)^L\delta^{2L}\frac{1}{i\pi}\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)-e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2L}\frac{\sqrt{\xi^2-1}}{\xi^2+\delta^2}. \tag{50a}$$

When $2L\gg 1$,

$$R_L(\alpha;\delta)=(-1)^L\delta^{2L}\frac{\pi^{-1/2}}{4i}\left\{\frac{1}{1+\delta^2}[e^{i\alpha}\mathrm{Ei}(-i\alpha)-e^{-i\alpha}\mathrm{Ei}(i\alpha)]L^{-3/2}+\mathrm{O}(L^{-5/2})\right\} \tag{50b}$$

for any α, δ ($|\delta|<1$). In the limit $L\to\infty$, $R_L(\alpha;\delta)\to 0$ uniformly in α, δ. Furthermore,

$$\frac{q_{2l+4}(\alpha)}{q_{2l+2}(\alpha)}=-\frac{\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)-e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2l-4}\sqrt{\xi^2-1}}{\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)-e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2l-2}\sqrt{\xi^2-1}}=\mathrm{O}(1) \tag{51a}$$

to all orders in α. Specifically,

$$\frac{q_{2l+4}(\alpha)}{q_{2l+2}(\alpha)}\sim-\frac{\int_1^\infty d\xi\,\xi^{-2l-4}\sqrt{\xi^2-1}}{\int_1^\infty d\xi\,\xi^{-2l-2}\sqrt{\xi^2-1}}=-\frac{2l}{2l+3}\quad(2l\gg 1) \tag{51b}$$

uniformly in α. The preceding expressions are suggestive of the efficiency of the expansion in question when $|\delta|\ll 1$.

The final formula for $\varphi(\alpha;\beta)$ is

$$\varphi(\alpha;\beta(\delta))=-(1+\delta^2)e^{\delta\alpha}\mathrm{Ei}(-\delta\alpha)-\delta\sqrt{1+\delta^2}\sum_{l=0}^{\infty}\delta^{2l}q_{2l+2}(\alpha),\quad\text{for }|\delta|<1, \tag{52}$$

where the $q_{2l+2}(\alpha)$ are given by Eqs. (49a) and (49b) and can be expressed in terms of known special functions for any finite l. An outline of the procedure along with the first six coefficients are provided in Appendix B. A straightforward, yet tedious, alternative derivation of Eq. (52) is carried out in Appendix C by invoking the Mellin transform technique. A strong indication of the validity of Eq. (52) is provided in Tables I and II which display values of the principal quantity $\varphi_1(\alpha;\delta)$ obtained both numerically through the integral of Eq. (41) and from the series of Eq. (48), for selected positive values of δ and α as well as for negative imaginary values of α. As verified analytically, the rate of convergence of the series is essentially independent of α.

Once $\varphi(\alpha;\beta)$ is evaluated, $\psi(\alpha;\beta)$ is obtained by direct differentiation. In particular, for $|\delta|<1$, term-by-term differentiation of the right-hand side of Eq. (52) is justified on grounds of uniform convergence of the resulting series when $|\alpha|>0$. Consequently,

$$\psi(\alpha;1)=-\frac{2}{\alpha^3}-\frac{1}{3}+\frac{1}{3\alpha}S_{2,2}(\alpha), \tag{53}$$

$$\psi(\alpha;\beta(\delta))=\frac{1+\delta^2}{\alpha}+\delta(1+\delta^2)e^{\delta\alpha}\mathrm{Ei}(-\delta\alpha)-\delta\sqrt{1+\delta^2}\sum_{l=0}^{\infty}\delta^{2l}q_{2l+1}(\alpha),\quad\text{for }|\delta|<1, \tag{54}$$

TABLE I. Values for $\varphi_1(\alpha;\delta)$obtained for positive α, δ ($\delta<1$)by numerical integration according to Eq. (41) and by evaluation of series on right-hand side of Eq. (48). The coefficients $q_n(\alpha)$,$n=2$, 4, 6, 8, are evaluated both by numerical integration from Eq. (49a) and by use of the results of Appendix B.

		Series from Eq. (48)			
α	From Eq. (41)	$L=1$	$L=2$	$L=3$	$L=4$
		Case A: $\varphi_1(\alpha;\delta=0.8)$			
10^{-15}	33.481 87	33.654 71	33.441 37	33.495 99	33.476 015
10^{-5}	10.456 17	10.628 98	10.415 68	10.470 29	10.450 32
10^{-2}	3.612 42	3.774 62	3.573 72	3.625 98	3.606 79
1	0.373 44	0.419 04	0.360 50	0.378 28	0.371 36
10	4.3665×10^{-2}	4.9763×10^{-2}	4.1838×10^{-2}	4.4368×10^{-2}	4.3357×10^{-2}
100	4.3846×10^{-3}	$4.999\,75\times10^{-3}$	4.1998×10^{-3}	4.4558×10^{-3}	4.3534×10^{-3}
10^4	4.3848×10^{-5}	5.0000×10^{-5}	4.2000×10^{-5}	4.4560×10^{-5}	4.3536×10^{-5}
		Case B: $\varphi_1(\alpha;\delta=0.1)$			
10^{-15}	33.651 39	33.654 71	33.651 37	33.651 39	33.651 39
10^{-5}	10.625 66	10.628 98	10.625 65	10.625 66	10.625 66
10^{-2}	3.771 49	3.774 62	3.771 48	3.771 49	3.771 49
1	0.418 13	0.419 04	0.418 13	0.418 13	0.418 13
10	4.9640×10^{-2}	4.9763×10^{-2}	4.9639×10^{-2}	4.9640×10^{-2}	4.9640×10^{-2}
100	4.9873×10^{-3}	$4.999\,75\times10^{-3}$	$4.987\,25\times10^{-3}$	4.9873×10^{-3}	4.9873×10^{-3}
10^4	4.9876×10^{-5}	5.0000×10^{-5}	4.9875×10^{-5}	4.9876×10^{-5}	4.9876×10^{-5}

where

$$q_{2l+1}(\alpha)=-\frac{d}{d\alpha}\,q_{2l+2}(\alpha)=\int_0^\infty ds e^{-\alpha s}s^{-2l-1}\left[\sqrt{1+s^2}-\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}s^{2m}\right] \tag{55a}$$

$$=(-1)^{l+1}\,\frac{1}{\pi}\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)+e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2l-1}\sqrt{\xi^2-1}. \tag{55b}$$

The first six coefficients $q_{2l+1}(\alpha)$ of the above series are tabulated in Appendix B. The remainder of this expansion when L terms are summed equals

$$Q_L(\alpha;\delta)=-\frac{\partial}{\partial\alpha}\,R_L(\alpha;\delta)=(-1)^{L+1}\delta^{2L}\,\frac{1}{\pi}\int_1^\infty d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)$$
$$+e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)]\xi^{-2L+1}\,\frac{\sqrt{\xi^2-1}}{\xi^2+\delta^2}, \tag{56}$$

which is $O(\ln\alpha)$ as $\alpha\to0$ and $O(1/\alpha^2)$ as $\alpha\to\infty$, i.e., integrable at $\alpha=0,\infty$. In the limit $L\to\infty$, $Q_L(\alpha;\delta)\to0$ uniformly in $|\delta|<1$, $|\alpha|>0$. The preceding expansion preserves the same interesting features as those implied by Eqs. (51a) and (51b).

Substitution for $\psi(\alpha;\beta)$ into Eqs. (30) and (31) yields

$$u(\alpha;\beta(\delta))=-(1+\delta^2)^{3/2}e^{\delta\alpha}\mathrm{Ei}(-\delta\alpha)-\frac{\delta\sqrt{1+\delta^2}}{\alpha}+(1+\delta^2)\sum_{l=0}^{\infty}\delta^{2l}q_{2l+1}(\alpha), \tag{57}$$

TABLE II. Complex values of $\varphi_1(\alpha;\delta)$ obtained for positive δ and pure imaginary α by numerical integration according to Eq. (41) and by evaluation of series on right-hand side of Eq. (48). These integrals arise, for example, over a very dry earth (conductivity $\sigma\simeq 0$) at $y=0$.

		Series from Eq. (48)		
$i\alpha$	From Eq. (41)	$L=1$	$L=3$	$L=4$
		Case A: $\varphi_1(\alpha;\delta=0.8)$		
10^{-15}	$33.481\,87+i1.570\,80$	$33.654\,71+i1.570\,80$	$33.495\,99+i1.570\,80$	$33.476\,01+i1.570\,80$
10^{-5}	$10.456\,035+i1.570\,64$	$10.628\,87+i1.570\,67$	$10.470\,16+i1.570\,64$	$10.450\,18+i1.570\,64$
10^{-2}	$3.568\,05+i1.506\,88$	$3.736\,64+i1.517\,41$	$3.581\,98+i1.507\,44$	$3.562\,27+i1.506\,66$
1	$0.118\,17+i0.441\,83$	$0.147\,40+i0.496\,82$	$0.122\,12+i0.447\,35$	$0.116\,42+i0.439\,50$
10	9.5813×10^{-7} $+i4.4054\times10^{-2}$	14.4412×10^{-7} $+i5.0269\times10^{-2}$	10.9325×10^{-7} $+i4.4777\times10^{-2}$	8.8476×10^{-7} $+i4.3737\times10^{-2}$
100	$i4.3850\times10^{-3}$	$i5.000\,25\times10^{-3}$	$i4.4562\times10^{-3}$	$i4.3538\times10^{-3}$
10^4	$i4.3848\times10^{-5}$	$i5.0000\times10^{-5}$	$i4.4560\times10^{-5}$	$i4.3536\times10^{-5}$
		Case B: $\varphi_1(\alpha;\delta=0.1)$		
10^{-15}	$33.651\,39+i1.570\,80$	$33.654\,71+i1.570\,80$	$33.651\,39+i1.570\,80$	$33.651\,39+i1.570\,80$
10^{-5}	$10.625\,55+i1.570\,67$	$10.628\,87+i1.570\,67$	$10.625\,55+i1.570\,67$	$10.625\,55+i1.570\,67$
10^{-2}	$3.733\,40+i1.517\,22$	$3.736\,64+i1.517\,41$	$3.733\,40+i1.517\,22$	$3.733\,40+i1.517\,22$
1	$0.146\,80+i0.495\,73$	$0.147\,40+i0.496\,82$	$0.146\,80+i0.495\,73$	$0.146\,80+i0.495\,73$
10	1.4327×10^{-6} $+i5.0144\times10^{-2}$	1.4441×10^{-6} $+i5.0269\times10^{-2}$	1.4327×10^{-6} $+i5.0144\times10^{-2}$	1.4327×10^{-6} $+i5.0144\times10^{-2}$
100	$i4.9878\times10^{-3}$	$i5.000\,25\times10^{-3}$	$i4.9878\times10^{-3}$	$i4.9878\times10^{-3}$
10^4	$i4.9876\times10^{-5}$	$i5.0000\times10^{-5}$	$i4.9876\times10^{-5}$	$i4.9876\times10^{-5}$

$$w(\alpha;\beta(\delta))=-\frac{2}{\alpha^3}-\frac{1}{3}+\frac{1}{3\alpha}S_{2,2}(\alpha)-\frac{\delta[\delta+(1+\delta^2)^{1/2}]}{\alpha}$$
$$-\delta(1+\delta^2)^{1/2}[\delta+(1+\delta^2)^{1/2}]e^{\delta\alpha}\mathrm{Ei}(-\delta\alpha)$$
$$+\delta[\delta+(1+\delta^2)^{1/2}]\sum_{l=0}^{\infty}\delta^{2l}q_{2l+1}(\alpha),\quad \text{for } |\delta|<1. \tag{58}$$

The analytic continuation to complex values of α can be carried out through the appropriate rotation of both contour C and the enclosed branch cut in the right half s-plane. The condition $|\delta|<1$ holds when the surrounding medium is air and the adjacent medium is earth, being dictated on mathematical grounds by the use of a power series representation in δ along with the existence of branch points at $\delta=\pm i$. A physical situation corresponding to these points arises when the ambient medium becomes perfectly conducting and the field is identically zero.

C. The *z*-component of the magnetic field in air

The field is easily evaluated with Eqs. (18b)–(27). Only the z-component of the magnetic field B_{0z} is calculable in simple closed form:

$$B_{0z}=\frac{\mu_0}{2\pi}e^{ik_0x}\left[y\left(\frac{1}{\rho_0^2}-\frac{1}{\rho_1^2}\right)-\frac{4y}{k^2}\frac{y^2-3(z+d)^2}{\rho_1^6}+\frac{1}{3i(z+d-iy)}S_{2,2}(-ky-ik(z+d))\right.$$
$$\left.-\frac{1}{3i(z+d+iy)}S_{2,2}(ky-ik(z+d))\right]. \tag{59}$$

This relatively simple structure is not surprising since it originates from the absence of any pole in the integrand of Eq. (29) when $\beta=1$.

D. Approximate expressions for the field when $k_0^2 \ll |k_1^2|$, $|k_0 k_1^{-1}(k_0\rho_1)| \ll 1$

Clearly, the critical parameters of the problem are $\alpha_\pm = -ik(z+d\pm iy)$ and k_0^2/k_1^2. When $k_0^2 \ll |k_1^2|$, at most only the first term of the associated series expansions needs to be retained essentially regardless of the magnitude of $|\alpha_\pm|$, while the condition $|\delta\alpha_\pm| \sim |k_0 k_1^{-1}(k_0\rho_1)| \ll 1$ enables approximation of the exponential integral $\mathrm{Ei}(-\alpha_\pm\delta)$ according to Eq. (44).

In particular, the exact integrated formulas may be greatly simplified under the condition $|k_1\rho_1|^2 \ll 1$ as, for instance, when the earth is highly conducting and the vertical distance from the image is much shorter than the wavelength in earth. Indeed, then $|\alpha_\pm| \ll 1$ in Eqs. (25)–(27) and small-argument approximations become effective. The resulting approximate expressions are

$$E_{0x} \sim -\frac{i\omega\mu_0}{2\pi} e^{ik_0 x}\left\{ \gamma - \frac{1}{2} - \frac{i\pi}{2} + \frac{2i}{3} k(z+d) + \left[1 + \frac{1}{8} k^2(z+d)^2 - \frac{1}{8} k^2 y^2\right] \ln\left(\frac{1}{2} k\rho_1\right) \right.$$
$$- \frac{1}{32}(5 - 4\gamma + 2\pi i) k^2[(z+d)^2 - y^2] - \frac{1}{4} k^2 (z+d) y \tan^{-1}\left(\frac{y}{z+d}\right)$$
$$\left. - \frac{2k_0^2}{k_1^2}\left[\gamma - \frac{i\pi}{2} + \ln\left(\frac{k_0^2}{k_1^2} k\rho_1\right)\right]\right\}, \tag{60}$$

$$E_{0y} \sim \frac{\mu_0 c}{2\pi} e^{ik_0 x} y\left[\frac{1}{\rho_0^2} - \frac{1 - 2k_0^2/k_1^2}{\rho_1^2}\right], \tag{61}$$

$$E_{0z} \sim \frac{\mu_0 c}{2\pi} e^{ik_0 x}\left[\frac{z-d}{\rho_0^2} - \frac{(1 - 2k_0^2/k_1^2)(z+d)}{\rho_1^2}\right], \tag{62}$$

$$B_{0x} \sim -i \frac{\mu_0 \omega}{2\pi c} e^{ik_0 x}\left\{\left[1 + \frac{1}{8} k^2(z+d)^2 - \frac{1}{8} k^2 y^2\right] \tan^{-1}\left(\frac{y}{z+d}\right) + \frac{2i}{3} ky \right.$$
$$\left. + \frac{1}{4} k^2 (z+d) y \ln\left(\frac{1}{2} k\rho_1\right) - \frac{1}{16}(5 - 4\gamma + 2\pi i) k^2 (z+d) y - \frac{k_0^2}{k_1^2} \tan^{-1}\left(\frac{y}{z+d}\right)\right\}, \tag{63}$$

$$B_{0y} \sim -\frac{\mu_0}{2\pi} e^{ik_0 x}\left\{\frac{z-d}{\rho_0^2} + \frac{1}{4} k^2 (z+d) \ln\left(\frac{1}{2} k\rho_1\right) + \frac{2ik}{3} - \frac{1}{16}(3 - 4\gamma + 2\pi i) k^2 (z+d) \right.$$
$$- \frac{1}{4} k^2 y \tan^{-1}\left(\frac{y}{z+d}\right) + \frac{2i}{15} k^3[(z+d)^2 - y^2] - 2ik \frac{k_0^2}{k_1^2}\left[\ln\left(\frac{2k_1^2}{k_0^2}\right) - 1\right]\Bigg\}, \tag{64}$$

$$B_{0z} \sim \frac{\mu_0}{2\pi} e^{ik_0 x}\left\{ y\left[\frac{1}{\rho_0^2} + \frac{1}{16}(3 - 4\gamma + 2\pi i) k^2 - \frac{1}{4} k^2 \ln\left(\frac{1}{2} k\rho_1\right) - \frac{4}{15} ik^3 (z+d)\right]\right.$$
$$\left. - \frac{1}{4} k^2 (z+d) \tan^{-1}\left(\frac{y}{z+d}\right)\right\}. \tag{65}$$

Equation (60) gives the first few terms of Carson's series[1,3,9] if k is approximated by k_1, with the exception of the correction term proportional to k_0^2/k_1^2. This discrepancy is due to the neglect of the Sommerfeld-pole contribution in Carson's formulation. Equations (60)–(65) agree with the findings of image theory in the static limit $\omega \to 0^+$. In this limit, the exact integrated formulas (18)–(23) reduce to the results of image theory without any restriction on k_0 and k_1. In particular, when the earth has zero conductivity and $\omega \to 0^+$, only the first term in the series expansion of Eq. (54) is nonzero. Only the approximate formula in Eq. (61) and the zeroth-order terms in Eq. (62) for the y- and z-components of the electric field are in accord with the application of the lateral-wave formulas of Ref. 10. No $1/\rho_1^2$ term of the reflected field is involved in the y- and z-components of the magnetic field because the earth was assumed to be nonmagnetic.

Another limiting case involves $|k_1\rho_1| \sim |\alpha_\pm| \gg 1$. Then large-argument approximations apply and the integrated formulas simplify to

$$E_{0x}\sim -\frac{\omega\mu_0}{\pi}\,e^{ik_0x}\left\{\frac{z+d}{k\rho_1^2}-i\frac{(z+d)^2-y^2}{k^2\rho_1^4}-i\frac{k_0^2}{k_1^2}\left[\gamma-\frac{i\pi}{2}+\ln\left(\frac{k_0^2}{k_1^2}\,k\rho_1\right)\right]\right\}, \tag{66}$$

$$E_{0y}\sim \frac{\mu_0 c}{2\pi}\,e^{ik_0x}y\left\{\frac{1}{\rho_0^2}-\frac{1-2k_0^2/k_1^2}{\rho_1^2}\right\}, \tag{67}$$

$$E_{0z}\sim \frac{\mu_0 c}{2\pi}\,e^{ik_0x}\left\{\frac{z-d}{\rho_0^2}-\frac{z+d}{\rho_1^2}+i\frac{k_0^2}{k_1^2}\frac{(z+d)^2-y^2}{k\rho_1^4}+2ik\,\frac{k_0^2}{k_1^2}\left[\gamma-\frac{i\pi}{2}+\ln\left(\frac{k_0^2}{k_1^2}\,k\rho_1\right)\right]\right\}, \tag{68}$$

$$B_{0x}\sim \frac{\mu_0\omega}{\pi c}\,e^{ik_0x}\left\{\frac{y}{k\rho_1^2}-\frac{2iy(z+d)}{k^2\rho_1^4}+i\frac{k_0^2}{k_1^2}\tan^{-1}\left(\frac{y}{z+d}\right)\right\}, \tag{69}$$

$$B_{0y}\sim -\frac{\mu_0}{2\pi}\,e^{ik_0x}\left\{\frac{z-d}{\rho_0^2}-\frac{(1+2k_0^2/k_1^2)(z+d)}{\rho_1^2}+2ik\,\frac{k_0^2}{k_1^2}\left[\gamma-\frac{i\pi}{2}+\ln\left(\frac{k_0^2}{k_1^2}\,k\rho_1\right)\right]\right\}, \tag{70}$$

$$B_{0z}\sim \frac{\mu_0}{2\pi}\,e^{ik_0x}y\left\{\frac{1}{\rho_0^2}-\frac{1}{\rho_1^2}+\frac{4i(z+d)}{k\rho_1^4}\right\}. \tag{71}$$

Only the y-component of the electric field preserves its static character to first order in k_0^2/k_1^2. The correction terms $O(k_0^2/k_1^2)$ become significant over a very dry earth. These either stem from the vicinity of the Sommerfeld pole (logarithm and inverse-tangent terms) or are due to secondary imperfect reflections from the image inside the earth (terms proportional to $1/\rho_1^2$, $1/\rho_1^4$).

IV. CONCLUSIONS AND DISCUSSION

The problem of the field created by an infinitely long and infinitely thin conductor carrying a traveling-wave current above a planar earth is revisited, but with a perspective different from previous analyses. Integral representations for the field in terms of contour integrals are derived in a physically meaningful way. It is shown that, as a direct consequence of this formulation, an exponentially decreasing outgoing current along the conductor (in the positive x direction) causes exponential increase of the radiation field in any direction perpendicular to the line source. This result is in accord with preliminary studies on the theory of the microstrip[14] that seem to have escaped attention in a considerable part of the recent literature.

In the limit where the propagation constant γ for the current becomes equal to the free-space wave number, all integrals for the field in air are, for the first time, evaluated *exactly* in terms of series expansions in the Sommerfeld pole. A prominent feature of the new series representations is the essential independence of their rates of convergence from distance. These series describe corrective contributions for the reflected (secondary) field to take account of the fact that it is not actually the field of an ideal image, and involve coefficients of nontrivial, yet calculable, form in terms of incomplete integrals of cylindrical functions. Finally, the component of the magnetic field normal to both conductor and interface is evaluated in simple closed form in terms of Lommel functions.

It should be borne in mind that the method of solution here crucially depends on the fact that the radiative structure is planar, and therefore is restricted in its applicability. Nevertheless, the present exposition for $\gamma=k_0$ reveals the nature of the field produced by multiphase transmission lines operating at sufficiently low frequencies above the earth, where the total field, trivially obtained via superposition, is spatially confined near the wires.

In the case of a complex propagation constant γ with Im $\gamma>0$, the usual radiation condition in three dimensions[22,23] is not explicitly invoked because the line has infinite length. This point is also discussed in Ref. 14. Heuristically speaking, the field growth in the radial distance $\sqrt{y^2+z^2}$ characterizes an asymptotic region in space where the assumed exponential increase along the *negative x* direction dominates over the free-space radiation decrease inversely proportional with distance. This region extends to infinity, in some sense, when the line becomes infinitely long.

Application of a limiting procedure that, starting with a localized source and the usual boundary condition at infinity, may lead to Eq. (2) and the accompanying assumptions in full mathematical rigor remains an open question for future work.

ACKNOWLEDGMENTS

The author is greatly indebted to Professors Tai Tsun Wu and Ronold W. P. King for their continuous guidance and encouragement. He is also grateful to John Myers, George Fikioris and Margaret Owens for their valuable comments and suggestions. Special thanks are due to Professor James R. Wait and Professor John Fikioris for useful correspondence. This work was supported in part by the Joint Services Electronics Program under Grant No. N00014-89-J-1023 with Harvard University.

APPENDIX A: ON THE ASYMPTOTIC BEHAVIOR OF THE RADIATION FIELD IN AIR

For definiteness, consider $y \geqslant 0$. The phase associated with Eqs. (12) and (13) reads

$$\Phi_{\mp}(\eta) = \sqrt{k_0^2 - \gamma^2 - \eta^2}(z \mp d) + \eta y, \quad z > d. \tag{A1}$$

The corresponding saddle point, $\eta = \eta_{\mp}^{(\mathrm{sp})}$, is

$$\eta_{\mp}^{(\mathrm{sp})} = \sqrt{k_0^2 - \gamma^2}\ \sin\ \theta_{0,1}, \quad 0 \leqslant \theta_{0,1} < \pi/2, \tag{A2}$$

where $-\pi/2 < \arg\sqrt{k_0^2 - \gamma^2} \leqslant 0$, yielding

$$\Phi_{\mp}(\eta^{(\mathrm{sp})}) = \sqrt{k_0^2 - \gamma^2}\rho_{0,1}, \quad \mathrm{Im}\ \Phi_{\mp}(\eta^{(\mathrm{sp})}) < 0, \tag{A3}$$

$$\left.\frac{d^2\Phi_{\mp}}{d\eta^2}\right|_{\eta = \eta^{(\mathrm{sp})}} = -\frac{\rho_1^3}{(z \mp d)^2}\ (k_0^2 - \gamma^2)^{-1/2}. \tag{A4}$$

$\rho_{0,1}$ and $\theta_{0,1}$ are defined by Eqs. (32). For $|\sqrt{k_0^2 - \gamma^2}|\rho_{0,1} \gg 1$ and fixed $\theta_{0,1}$, the steepest descents path is determined by the conditions

$$\mathrm{Re}[\Phi_{\mp}(\eta) - \Phi_{\mp}(\eta^{(\mathrm{sp})})] = 0, \tag{A5}$$

$$\mathrm{Im}[\Phi_{\mp}(\eta) - \Phi_{\mp}(\eta^{(\mathrm{sp})})] > 0. \tag{A6}$$

Once the integration path Γ is properly deformed to pick up the saddle-point contribution, the exponential growth in $\sqrt{y^2 + z^2}$ of the field follows easily from (A3).

Under the usual substitution

$$\eta = \eta(\chi) = \sqrt{k_0^2 - \gamma^2}\ \sin\ \chi, \tag{A7}$$

each $\Phi_{\mp}(\eta)$ is recast in the form

$$\Phi_{\mp}(\eta(\chi)) = \sqrt{k_0^2 - \gamma^2}\rho_{0,1}\ \cos(\theta_{0,1} - \chi). \tag{A8}$$

A branch of $\chi = \chi(\eta)$ can be chosen so that $\chi(0) = 0$. In the χ-plane, the saddle point is $\chi_{\mp}^{(\mathrm{sp})} = \theta_{0,1}$, while the steepest descents path is now described by the relations

$$A\ \cos(\chi_r - \theta_{0,1})\cosh\ \chi_i - D\ \sinh\ \chi_i\ \sin(\chi_r - \theta_{0,1}) = A, \tag{A9}$$

$$A\ \sinh\ \chi_i\ \sin(\chi_r - \theta_{0,1}) + D\ \cos(\chi_r - \theta_{0,1})\cosh\ \chi_i < D, \tag{A10}$$

where $A = \mathrm{Re}\sqrt{k_0^2 - \gamma^2} > 0$, $D = -\mathrm{Im}\sqrt{k_0^2 - \gamma^2} > 0$, and $\chi_r = \mathrm{Re}\ \chi$, $\chi_i = \mathrm{Im}\ \chi$. The path configuration in the χ-plane is depicted in Fig. 5. Note that

$$\tan\ \overline{\theta} = A/D. \tag{A11}$$

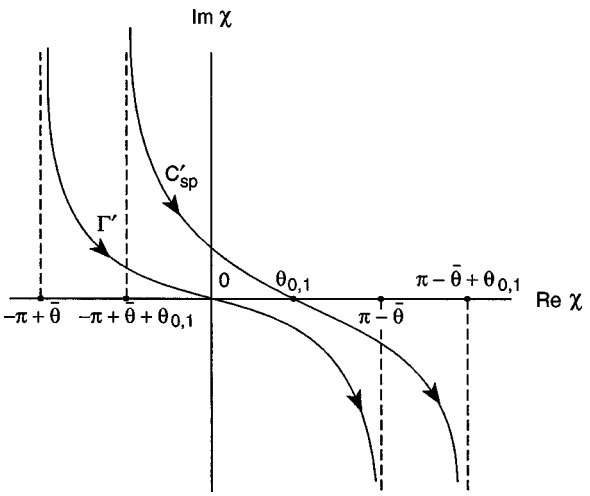

FIG. 5. A sketch of the integration-path configuration in the χ-plane under (A7). Γ' is the image of Γ of Fig. 2 and C'_{sp} is the steepest descents path. For simplicity, the branch points corresponding to $\eta=\pm\sqrt{k_1^2-\gamma^2}$ are not shown here. The dashed lines are vertical asymptotes ($\bar{\theta}=\pi/2-\arg\sqrt{k_0^2-\gamma^2}$).

Possible contribution from a Sommerfeld pole is recognized as exponentially recessive.

APPENDIX B: EVALUATION OF $q_n(\alpha)$

From Eqs. (49a) and (55a), the $q_n(\alpha)$ read

$$q_n(\alpha)=\int_0^{+\infty} ds e^{-\alpha s} s^{-n}\left[\sqrt{1+s^2}-\sum_{m=0}^{[(n-1)/2]}\binom{\frac{1}{2}}{m} s^{2m}\right], \tag{B1}$$

where $\alpha>0$, $n=0,1,\ldots,$ and $[x]$ denotes the integral part of the real number x. These satisfy the following equations:

$$\frac{d}{d\alpha} q_{2l+2}(\alpha)=-q_{2l+1}(\alpha), \tag{B2}$$

$$\frac{d}{d\alpha}\left[q_{2l+1}(\alpha)-\binom{\frac{1}{2}}{l}\ln\alpha\right]=-q_{2l}(\alpha), \tag{B3}$$

where $l\geqslant 0$ and

$$q_0(\alpha)=\int_0^{+\infty} ds e^{-\alpha s}\sqrt{1+s^2}=\frac{1}{\alpha} S_{1,1}(\alpha). \tag{B4}$$

$S_{\nu,\nu}(\alpha)$ is given by Eq. (36).

The application of Lebesgue's dominated convergence theorem to Eq. (49b) when $\alpha\to 0$, $l\geqslant 1$ entails:

$$c_l\equiv\lim_{\alpha\to 0} q_{2l+2}(\alpha)=(-1)^l\int_1^{\infty} d\xi \xi^{-2l-2}\sqrt{\xi^2-1}=\frac{(-1)^l}{2}\frac{\Gamma(3/2)\Gamma(l)}{\Gamma((3/2)+l)}. \tag{B5}$$

For $l\geqslant 1$, the integral of Eq. (55b) is recast in the form

$$q_{2l+1}(\alpha)-\binom{\frac{1}{2}}{l}\ln\alpha=\frac{(-1)^{l-1}}{\pi}\int_1^{\infty} d\xi[e^{i\alpha\xi}\mathrm{Ei}(-i\alpha\xi)+e^{-i\alpha\xi}\mathrm{Ei}(i\alpha\xi)-2\ln\alpha]\xi^{-2l-1}\sqrt{\xi^2-1}. \tag{B6}$$

For $0<\alpha\ll 1$, the preceding integral may be rewritten as

$$\int_1^\infty = \int_1^{O(\alpha^{-\lambda})} + \int_{O(\alpha^{-\lambda})}^\infty,$$

where $0<\lambda<1$. In the limit $\alpha\to 0$, the second term tends to zero because absolute convergence obtains. In regard to the first term, the integrand in brackets is properly expanded according to Eq. (44). This in turn leads to

$$d_l \equiv \lim_{\alpha\to 0}\left\{ q_{2l+1}(\alpha) - \binom{\frac{1}{2}}{l} \ln\,\alpha \right\} = 2\frac{(-1)^{l-1}}{\pi}\int_1^\infty d\xi(\gamma+\ln\,\xi)\xi^{-2l-1}\sqrt{\xi^2-1}$$

$$= 2^{-1}\binom{\frac{1}{2}}{l}\left[2\gamma+\psi(1+l)-\psi\left(-\frac{1}{2}+l\right)\right], \tag{B7}$$

where $l\geqslant 1$ and

$$\psi(z)=\frac{d}{dz}\ln\,\Gamma(z), \quad \psi(1)=-\gamma, \quad \psi(\tfrac{1}{2})=-\gamma-2\ln\,2,$$

$$\psi(z+l)=\frac{1}{z}+\cdots+\frac{1}{z+l-1}+\psi(z).$$

Note that $q_2(\alpha)$ and $q_1(\alpha)-\ln\,\alpha$ exhibit different behavior from that described in Eqs. (B5) and (B7), respectively. By use of Eq. (B1),

$$q_1(\alpha)=\ln\,\alpha+\frac{1}{\alpha}+\gamma+\ln\,2-1+O(\alpha\,\ln\,\alpha) \quad \text{for } \alpha\to 0. \tag{B8}$$

Likewise,

$$q_2(\alpha)=-\ln\,\alpha-\gamma+\ln\,2-1+O(\alpha\,\ln\,\alpha). \tag{B9}$$

Equations (B2)–(B5) and (B7)–(B9) suffice to determine $q_n(\alpha)$ for any $n\geqslant 1$.

In consideration of the following formulas:[20]

$$\frac{d}{dt}[tS_{1,1}(t)]=tS_{0,0}(t), \quad S_{1,1}(t)=1-\frac{d}{dt}S_{0,0}(t), \tag{B10}$$

along with Eqs. (B8) and (B9), Eqs. (B3) and (B2) are properly integrated for $l=0$ to give

$$q_1(\alpha)=\frac{1}{\alpha}+\ln\,\alpha+(\gamma+\ln\,2-1)+\lim_{\delta_1\to 0^+}\int_{\delta_1}^\alpha dt\left\{\frac{1}{t^2}-\frac{1}{t}S_{1,1}(t)\right\}$$

$$=-1+\gamma+\ln(2\alpha)-D(\alpha)+S_{1,1}(\alpha), \tag{B11}$$

$$q_2(\alpha)=-\ln\,\alpha-\gamma+\ln\,2-1-\lim_{\delta_1\to 0^+}\int_{\delta_1}^\alpha dt\left\{\ln\,t+\gamma+\ln\,2-1+S_{1,1}(t)-D(t)-\frac{1}{t}\right\}$$

$$=-\alpha[-1+\gamma+\ln(2\alpha)-D(\alpha)]+S_{0,0}(\alpha)-\alpha S_{1,1}(\alpha). \tag{B12}$$

In the above, use is made of the small-argument approximations for $\mathbf{H}_{0,1}(\alpha)$ and $Y_{0,1}(\alpha)$, and $D(\alpha)$ is defined as

$$D(\alpha)=\int_0^\alpha dtS_{0,0}(t). \tag{B13}$$

This function was initially studied in connection with a problem in acoustic radiation.[24] Despite the fact that[20]

$$\int_0^\alpha dt Y_0(t)=\frac{\pi\alpha}{2}\left[Y_0(\alpha)\mathbf{H}_{-1}(\alpha)+\mathbf{H}_0(\alpha)Y_1(\alpha)\right],\quad \mathbf{H}_{-1}(\alpha)=\frac{2}{\pi}-\mathbf{H}_1(\alpha), \tag{B14}$$

$$\int_0^\alpha dt\mathbf{H}_0(t)=\frac{\alpha^2}{\pi}\,{}_2F_3(1,1;\tfrac{3}{2},\tfrac{3}{2},2;-\alpha^2/4), \tag{B15}$$

the form of Eq. (B13) is more advantageous to employ here. $D(\alpha)$ cannot be expressed in terms of simpler transcendental functions. It is noted in passing that the integral in Eq. (B15) is a special case of the incomplete Lipschitz-Hankel integral of the Struve type.[25]

For $l\geqslant 1$, it is permissible to integrate Eqs. (B2) and (B3) explicitly over the range $[0,\alpha]$ subject to the conditions of Eqs. (B5) and (B7):

$$q_{2l+2}(\alpha)=c_l-\int_0^\alpha dt q_{2l+1}(t),\quad q_{2l+1}(\alpha)=\binom{\frac{1}{2}}{l}\ln\alpha+d_l-\int_0^\alpha dt q_{2l}(t). \tag{B16}$$

Successive integrations by parts according to the recursive scheme

$$\int^t dt t^\nu D(t)=\frac{t^{\nu+1}}{\nu+1}D(t)-\frac{1}{\nu+1}\int^t dt t^{\nu+1}S_{0,0}(t), \tag{B17}$$

$$\int^t dt t^\nu S_{0,0}(t)=t^\nu S_{1,1}(t)-(\nu-1)\int^t dt t^{\nu-1}S_{1,1}(t), \tag{B18}$$

$$\int^t dt t^\nu S_{1,1}(t)=\frac{t^{\nu+1}}{\nu+1}-t^\nu S_{0,0}(t)+\nu\int^t dt t^{\nu-1}S_{0,0}(t) \tag{B19}$$

provide a routine procedure for the evaluation of $q_n(\alpha)$ for any finite $n\geqslant 3$. For all reasonable purposes, the first 13 integrals $q_n(\alpha)$ $(n=0,1,\ldots,12)$ are sufficient. It is readily found that $q_n(\alpha)$ $(n\geqslant 3)$ can be expressed as

$$\begin{aligned}q_n(\alpha)=\frac{(-1)^{n-1}}{(n-1)!}\{&\mathscr{P}^{(1)}_{n-1}(\alpha)[\gamma+\ln(2\alpha)-D(\alpha)]\\&-\mathscr{P}^{(2)}_{n-1}(\alpha)-\mathscr{P}^{(3)}_{n-2}(\alpha)S_{0,0}(\alpha)+\mathscr{P}^{(4)}_{n-1}(\alpha)S_{1,1}(\alpha)\},\end{aligned} \tag{B20}$$

where the $\mathscr{P}^{(j)}_m(\alpha)$ $(j=1,2,3,4)$ denote even or odd mth-degree polynomials in α, $\mathscr{P}^{(j)}_m(-\alpha)=(-1)^m\mathscr{P}^{(j)}_m(\alpha)$. For $n=3,\ldots,12$, the $\mathscr{P}^{(j=1,2,4)}_{n-1}(\alpha)$ and $\mathscr{P}^{(j=3)}_{n-2}(\alpha)$ read

$$\begin{gathered}\mathscr{P}^{(1)}_2(\alpha)=\alpha^2+1,\quad \mathscr{P}^{(1)}_3(\alpha)=\alpha^3+3\alpha,\quad \mathscr{P}^{(1)}_4(\alpha)=\alpha^4+6\alpha^2-3,\\ \mathscr{P}^{(1)}_5(\alpha)=\alpha^5+10\alpha^3-15\alpha,\quad \mathscr{P}^{(1)}_6(\alpha)=\alpha^6+15\alpha^4-45\alpha^2+45,\\ \mathscr{P}^{(1)}_7(\alpha)=\alpha^7+21\alpha^5-105\alpha^3+315\alpha,\\ \mathscr{P}^{(1)}_8(\alpha)=\alpha^8+28\alpha^6-210\alpha^4+1260\alpha^2-1575,\\ \mathscr{P}^{(1)}_9(\alpha)=\alpha^9+36\alpha^7-378\alpha^5+3780\alpha^3-14\,175\alpha,\\ \mathscr{P}^{(1)}_{10}(\alpha)=\alpha^{10}+45\alpha^8-630\alpha^6+9450\alpha^4-70\,875\alpha^2+99\,225,\\ \mathscr{P}^{(1)}_{11}(\alpha)=\alpha^{11}+55\alpha^9-990\alpha^7+20\,790\alpha^5-259\,875\alpha^3+1\,091\,475\alpha;\end{gathered} \tag{B21}$$

$$\mathscr{P}_2^{(2)}(\alpha)=\alpha^2-\frac{1}{2},\quad \mathscr{P}_3^{(2)}(\alpha)=\alpha^3+\frac{3}{2}\alpha,\quad \mathscr{P}_4^{(2)}(\alpha)=\alpha^4+\frac{9}{2}\alpha^2-\frac{3}{4},$$

$$\mathscr{P}_5^{(2)}(\alpha)=\alpha^5+\frac{17}{2}\alpha^3-\frac{75}{4}\alpha,\quad \mathscr{P}_6^{(2)}(\alpha)=\alpha^6+\frac{27}{2}\alpha^4-\frac{225}{4}\alpha^2+\frac{75}{4},$$

$$\mathscr{P}_7^{(2)}(\alpha)=\alpha^7+\frac{39}{2}\alpha^5-\frac{501}{4}\alpha^3+\frac{1785}{4}\alpha,$$

$$\mathscr{P}_8^{(2)}(\alpha)=\alpha^8+\frac{53}{2}\alpha^6-\frac{963}{4}\alpha^4+\frac{3255}{2}\alpha^2-\frac{6195}{8},\tag{B22}$$

$$\mathscr{P}_9^{(2)}(\alpha)=\alpha^9+\frac{69}{2}\alpha^7-\frac{1683}{4}\alpha^5+\frac{8979}{2}\alpha^3-\frac{169\,155}{8}\alpha,$$

$$\mathscr{P}_{10}^{(2)}(\alpha)=\alpha^{10}+\frac{87}{2}\alpha^8-\frac{2745}{4}\alpha^6+\frac{42\,555}{4}\alpha^4-\frac{732\,375}{8}\alpha^2+\frac{424\,305}{8},$$

$$\mathscr{P}_{11}^{(2)}(\alpha)=\alpha^{11}+\frac{107}{2}\alpha^9-\frac{4245}{4}\alpha^7+\frac{90\,525}{4}\alpha^5-\frac{2\,433\,855}{8}\alpha^3+\frac{13\,399\,155}{8}\alpha;$$

$$\mathscr{P}_1^{(3)}(\alpha)=\alpha,\quad \mathscr{P}_2^{(3)}(\alpha)=\alpha^2,\quad \mathscr{P}_3^{(3)}(\alpha)=\alpha^3+3\alpha,\quad \mathscr{P}_4^{(3)}(\alpha)=\alpha^4+7\alpha^2,$$

$$\mathscr{P}_5^{(3)}(\alpha)=\alpha^5+12\alpha^3-45\alpha,\quad \mathscr{P}_6^{(3)}(\alpha)=\alpha^6+18\alpha^4-123\alpha^2,$$

$$\mathscr{P}_7^{(3)}(\alpha)=\alpha^7+25\alpha^5-249\alpha^3+1575\alpha,$$

$$\mathscr{P}_8^{(3)}(\alpha)=\alpha^8+33\alpha^6-441\alpha^4+4959\alpha^2,\tag{B23}$$

$$\mathscr{P}_9^{(3)}(\alpha)=\alpha^9+42\alpha^7-720\alpha^5+11\,790\alpha^3-99\,225\alpha,$$

$$\mathscr{P}_{10}^{(3)}(\alpha)=\alpha^{10}+52\alpha^8-1110\alpha^6+24\,660\alpha^4-354\,195\alpha^2;$$

$$\mathscr{P}_2^{(4)}(\alpha)=\alpha^2,\quad \mathscr{P}_3^{(4)}(\alpha)=\alpha^3+2\alpha,\quad \mathscr{P}_4^{(4)}(\alpha)=\alpha^4+5\alpha^2,$$

$$\mathscr{P}_5^{(4)}(\alpha)=\alpha^5+9\alpha^3-16\alpha,\quad \mathscr{P}_6^{(4)}(\alpha)=\alpha^6+14\alpha^4-51\alpha^2,$$

$$\mathscr{P}_7^{(4)}(\alpha)=\alpha^7+20\alpha^5-117\alpha^3+384\alpha,$$

$$\mathscr{P}_8^{(4)}(\alpha)=\alpha^8+27\alpha^6-229\alpha^4+1497\alpha^2,\tag{B24}$$

$$\mathscr{P}_9^{(4)}(\alpha)=\alpha^9+35\alpha^7-405\alpha^5+4257\alpha^3-18\,432\alpha,$$

$$\mathscr{P}_{10}^{(4)}(\alpha)=\alpha^{10}+44\alpha^8-666\alpha^6+10\,260\alpha^4-85\,095\alpha^2,$$

$$\mathscr{P}_{11}^{(4)}(\alpha)=\alpha^{11}+54\alpha^9-1036\alpha^7+22\,050\alpha^5-290\,925\alpha^3+1\,474\,560\alpha.$$

As explained in Sec. III, the final formulas for $q_n(\alpha)$ may be extended over complex values of α for which $|\arg\alpha|<\pi$. Small-argument approximations for $q_n(\alpha)$ follow from Eqs. (B11)–(B15) and (B20)–(B24) if the series expansions in α for $\mathbf{H}_{0,1}(\alpha)$, $Y_{0,1}(\alpha)$, and ${}_2F_3(1,1;\frac{3}{2},\frac{3}{2},2;-\alpha^2/4)$ are taken into account. Large-argument approximations when $|\alpha|\to\infty$, $|\arg\alpha|<\pi$ can be readily obtained either from Eq. (B1) via Laplace's method[18] or by use of the asymptotic expansions of the related transcendental functions.[20,26]

APPENDIX C: EVALUATION OF $\varphi(\alpha;\beta)$ BY THE MELLIN TRANSFORM TECHNIQUE

In this appendix, Eq. (52) is derived formally via application of the Mellin transform technique.[27,28] For mathematical convenience, let $\beta=i\tilde{\beta}$ and take $\tilde{\beta}>0$. Then,

$$\varphi(\alpha;\beta(\delta))=\int_0^\infty ds\ \frac{e^{-\alpha s}}{s+i\tilde{\beta}\sqrt{1+s^2}}=(1-\tilde{\delta}^2)I_1(\alpha;\tilde{\delta})-i\tilde{\delta}\sqrt{1-\tilde{\delta}^2}I_2(\alpha;\tilde{\delta}), \tag{C1}$$

where

$$\tilde{\delta}=\frac{\tilde{\beta}}{\sqrt{1+\tilde{\beta}^2}}=-i\delta,\quad \delta=\frac{\beta}{\sqrt{1-\beta^2}}, \tag{C2}$$

$$I_1(\alpha;\tilde{\delta})=\int_0^\infty ds\ \frac{se^{-\alpha s}}{s^2+\tilde{\delta}^2}=-\frac{1}{2}\left[e^{i\tilde{\delta}\alpha}\mathrm{Ei}(-i\tilde{\delta}\alpha)+e^{-i\tilde{\delta}\alpha}\mathrm{Ei}(i\tilde{\delta}\alpha)\right], \tag{C3}$$

$$I_2(\alpha;\tilde{\delta})=\int_0^\infty ds\ \frac{\sqrt{1+s^2}}{s^2+\tilde{\delta}^2}\,e^{-\alpha s}. \tag{C4}$$

When $\alpha=\mathrm{O}(1)>0$, Eq. (38) implies that an expansion of $I_2(\alpha;\tilde{\delta})$ in powers of $\tilde{\delta}$ can be obtained quite straightforwardly by invoking its Mellin transform $\overline{I}_2(\alpha;\zeta)$. Avoiding elaboration on mathematical subtleties, such as issues related to convergence of any of the resulting series, consider

$$\overline{I}_2(\alpha;\zeta)=\int_0^\infty d\tilde{\delta}^2(\tilde{\delta}^2)^{-\zeta}I_2(\alpha;\tilde{\delta})=\frac{\pi}{\sin\ \pi\zeta}\int_0^\infty ds e^{-\alpha s}s^{-2\zeta}\sqrt{1+s^2},\quad 0<\mathrm{Re}\ \zeta<\frac{1}{2}, \tag{C5}$$

by interchanging the order of integration. Inversion of the above formula is studied initially according to the formula

$$I_2(\alpha;\tilde{\delta})=\frac{1}{2\pi i}\int_{c'-i\infty}^{c'+i\infty}d\zeta\overline{I}_2(\alpha;\zeta)(\tilde{\delta}^2)^{\zeta-1},\quad 0<c'<\frac{1}{2}. \tag{C6}$$

Of course, $\overline{I}_2(\alpha;\zeta)$ defined by (C5) can be analytically continued into $\mathrm{Re}\ \zeta\leqslant 0$, $\mathrm{Re}\ \zeta\geqslant 1/2$, as shown below. It is noteworthy that

$$\begin{aligned}\overline{I}_2(\alpha;\zeta)&\sim\frac{\pi}{\sin\ \pi\zeta}\int_0^\infty ds e^{-\alpha s}s^{-2\zeta}\\ &=\frac{\pi}{\sin\ \pi\zeta}\,\alpha^{2\zeta-1}\Gamma(1-2\zeta)\\ &=\frac{\pi^2}{\sin\ \pi\zeta\ \sin 2\pi\zeta}\,\frac{\alpha^{2\zeta-1}}{\Gamma(2\zeta)}\\ &\sim-(2\pi)^{3/2}(2\zeta)^{1/2}e^{-2\zeta\ \ln 2\zeta+(2\pm 3i\pi)\zeta+(2\zeta-1)\ln\ \alpha},\\ &\qquad |\zeta|\to\infty,\quad 0<|\arg\ \zeta|<\pi/2,\end{aligned} \tag{C7}$$

where the upper sign holds for $\mathrm{Re}\ \zeta>0$ and the lower sign for $\mathrm{Re}\ \zeta<0$, and use was made of the Stirling formula

$$\Gamma(z)\sim\sqrt{2\pi}e^{-z+(z-1/2)\ln z},\quad |z|\to\infty,\quad |\arg\ z|<\pi.$$

In view of

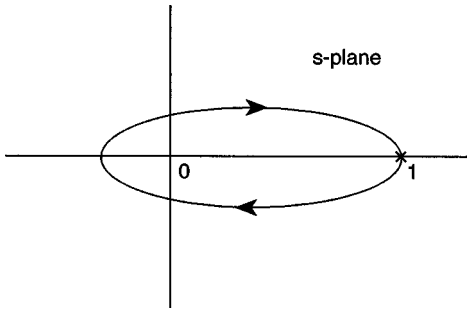

FIG. 6. Contour C_0 that serves the analytic continuation of the integral formula (C5) into the right and left ζ-plane according to (C8).

$$\int_0^1 ds e^{-\alpha s} s^{-2\zeta}\sqrt{1+s^2} = \frac{e^{i2\pi\zeta}}{2i\,\sin 2\pi\zeta}\int_{C_0} ds e^{-\alpha s} s^{-2\zeta}\sqrt{1+s^2}, \tag{C8}$$

where the upper limit $s=1$ is chosen arbitrarily and C_0 is a closed contour passing through 1 and encircling 0 clockwise without enclosing any of the branch points $s=\pm i$ (as shown in Fig. 6), it is not difficult to deduce that $\overline{I}_2(\alpha;\zeta)$ is a meromorphic function of ζ. For Re $\zeta>0$, its poles are located at $\zeta=l$ and $\zeta=l-1/2$, where l is any positive integer. By writing

$$\overline{I}_2(\alpha;\zeta) = \frac{\pi}{\sin\,\pi\zeta}\left\{\int_0^\infty ds\; e^{-\alpha s} s^{-2\zeta}\sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m} s^{2m} + \int_0^\infty ds e^{-\alpha s} s^{-2\zeta} \times\left[\sqrt{1+s^2} - \sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m} s^{2m}\right]\right\}, \tag{C9}$$

the singular behavior of $\overline{I}_2(\alpha;\zeta)(\tilde{\delta}^2)^{\zeta-1}$ in the vicinity of $\zeta=l$ is singled out as

$$\begin{aligned}\overline{I}_2(\alpha;\zeta)(\tilde{\delta}^2)^{\zeta-1} \sim (\tilde{\delta}^2)^{l-1}\Bigg\{&\frac{(-1)^{l+1}}{2(\zeta-l)^2}\sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m)-1}}{(2l-2m-1)!} \\ &+\frac{(-1)^{l+1}}{\zeta-l}\sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m)-1}}{(2l-2m-1)!}\left[\gamma - h_{2l-2m-1} + \ln(\tilde{\delta}\alpha)\right] \\ &+\frac{(-1)^{l+1}}{\zeta-l}\int_0^\infty ds e^{-\alpha s} s^{-2l}\left[\sqrt{1+s^2} - \sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m} s^{2m}\right]\Bigg\} \quad (l=1,2,\ldots),\end{aligned} \tag{C10}$$

where

$$h_n = 1 + 2^{-1} + \cdots + n^{-1}, \quad n \geqslant 1. \tag{C11}$$

In the vicinity of $\zeta=l-1/2$,

$$\overline{I}_2(\alpha;\zeta)(\tilde{\delta}^2)^{\zeta-1} \sim \frac{\pi}{2}\frac{(-1)^l}{\zeta-l+1/2}\,\tilde{\delta}^{2l-3}\sum_{m=0}^{l-1}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m-1)}}{(2l-2m-2)!}. \tag{C12}$$

Equation (C7) suggests that deformation of the original inversion path $\mathscr{C}_1$ so that it finally coincides with $\mathscr{C}_2$, which wraps around the positive real axis (see Fig. 7), is legitimate. Subsequent summation of the residues from all pole contributions there yields

$$I_2(\alpha;\tilde{\delta}) = \frac{1}{2\pi i} \oint_{\mathscr{C}_2} d\zeta \overline{I_2}(\alpha;\zeta)(\overline{\delta^2})^{\zeta-1}$$

$$= \frac{\pi}{2}\mathscr{S}_{21}(\alpha;\tilde{\delta}) + [\gamma + \ln(\tilde{\delta}\alpha)]\mathscr{S}_{22}(\alpha;\tilde{\delta}) - \mathscr{S}_{23}(\alpha;\tilde{\delta}) + \lim_{L\to\infty}\sum_{l=0}^{L}(-1)^l\tilde{\delta}^{2l}q_{2l+2}(\alpha), \tag{C13}$$

where $\tilde{\delta}$ is sufficiently small, $q_{2l+2}(\alpha)$ is given by Eq. (49a), and

$$\mathscr{S}_{21}(\alpha;\tilde{\delta}) \equiv \sum_{l=0}^{\infty}(-1)^l\tilde{\delta}^{2l-1}\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m)}}{(2l-2m)!}$$

$$= \sum_{m=0}^{\infty}\sum_{p=0}^{\infty}(-1)^{p+m}\binom{\frac{1}{2}}{m}\tilde{\delta}^{2(p+m)-1}\frac{\alpha^{2p}}{(2p)!}$$

$$= \frac{1}{\tilde{\delta}}\sqrt{1-\tilde{\delta}^2}\cos\tilde{\delta}\alpha, \tag{C14}$$

$$\mathscr{S}_{22}(\alpha;\tilde{\delta}) \equiv \sum_{l=0}^{\infty}(-1)^l\tilde{\delta}^{2l}\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m)+1}}{(2l-2m+1)!}$$

$$= \sum_{m=0}^{\infty}\sum_{p=0}^{\infty}(-1)^{p+m}\binom{\frac{1}{2}}{m}\tilde{\delta}^{2(p+m)}\frac{\alpha^{2p+1}}{(2p+1)!}$$

$$= \frac{\sqrt{1-\tilde{\delta}^2}}{\tilde{\delta}}\sin\tilde{\delta}\alpha, \tag{C15}$$

$$\mathscr{S}_{23}(\alpha;\tilde{\delta}) \equiv \sum_{l=0}^{\infty}(-1)^l\tilde{\delta}^{2l}\sum_{m=0}^{l}\binom{\frac{1}{2}}{m}\frac{\alpha^{2(l-m)+1}}{(2l-2m+1)!}h_{2l-2m+1}$$

$$= \sum_{p=0}^{\infty}\sum_{m=0}^{\infty}(-1)^{p+m}\binom{\frac{1}{2}}{m}\tilde{\delta}^{2(p+m)}\frac{\alpha^{2p+1}}{(2p+1)!}h_{2p+1}$$

$$= \frac{1}{\tilde{\delta}}\sqrt{1-\tilde{\delta}^2}\sum_{p=0}^{\infty}(-1)^p\frac{(\tilde{\delta}\alpha)^{2p+1}}{(2p+1)!}h_{2p+1}. \tag{C16}$$

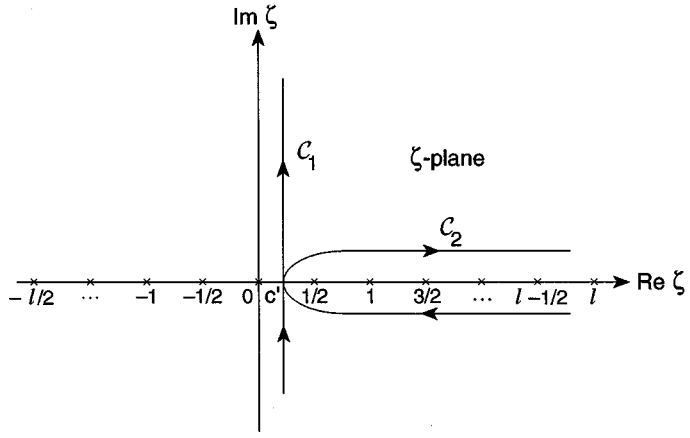

FIG. 7. Inversion paths for the Mellin transform $\overline{I_2}(\alpha;\zeta)$ of (C5). $\mathscr{C}_1$ is the original inversion path and $\mathscr{C}_2$ serves the expansion of $I_2(\alpha;\delta)$ in powers of δ. l is any positive integer.

The last series is easily evaluated by noting that

$$h_m = -\sum_{l=1}^{m} \frac{(-1)^l}{l} \binom{m}{l}. \tag{C17}$$

Therefore

$$\begin{aligned}\sum_{p=0}^{\infty} (-1)^p \frac{z^{2p+1}}{(2p+1)!}\, h_{2p+1} &= \sum_{p=0}^{\infty} \sum_{l=0}^{2p} (-1)^{p+l} \frac{z^{2p+1}}{(l+1)!(l+1)} \frac{1}{(2p-l)!} \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (-1)^{m+n} \frac{z^{2m+2n+1}}{(2n+1)!(2n+1)} \frac{1}{(2m)!} \\ &\quad + \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (-1)^{m+n} \frac{z^{2m+2n+3}}{(2n+2)!(2n+2)} \frac{1}{(2m+1)!} \\ &= z\, \cos z\, F(z) + \sin z\, G(z), \end{aligned} \tag{C18}$$

where

$$F(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n+1)!(2n+1)}, \tag{C19}$$

$$G(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+2}}{(2n+2)!(2n+2)}. \tag{C20}$$

$F(z)$ satisfies

$$\frac{d}{dz}(zF) = \frac{1}{z} \sin z, \quad F(0) = 1, \tag{C21}$$

with the solution

$$F(z) = \frac{1}{z} \int_0^z dt\, \frac{\sin t}{t} = \frac{\pi}{2z} - \frac{1}{2iz} [\mathrm{Ei}(-iz) - \mathrm{Ei}(iz)]. \tag{C22}$$

By inspection of (C20),

$$\begin{aligned} G(z) &= \int_0^z dt\, \frac{1-\cos t}{t} = \int_0^{\infty} \frac{dt}{t} \left(\frac{1}{1+t^2} - \cos t \right) + \int_0^z dt\, \frac{t}{1+t^2} - \int_z^{\infty} \frac{dt}{t(1+t^2)} + \int_z^{\infty} dt\, \frac{\cos t}{t} \\ &= \gamma + \ln z - \tfrac{1}{2} [\mathrm{Ei}(-iz) + \mathrm{Ei}(iz)]. \end{aligned} \tag{C23}$$

Equations (C18), (C22), and (C23), along with (C16), lead to

$$\begin{aligned} \mathscr{S}_{23}(\alpha;\tilde{\delta}) = \tilde{\delta}^{-1} \sqrt{1-\tilde{\delta}^2} \Bigg\{ &\frac{\pi}{2} \cos \tilde{\delta}\alpha + [\gamma + \ln(\tilde{\delta}\alpha)] \sin \tilde{\delta}\alpha \\ &+ \frac{i}{2} [e^{i\tilde{\delta}\alpha} \mathrm{Ei}(-i\tilde{\delta}\alpha) - e^{-i\tilde{\delta}\alpha} \mathrm{Ei}(i\tilde{\delta}\alpha)] \Bigg\}. \end{aligned} \tag{C24}$$

Substitution of (C14), (C15), and (C24) into (C13) yields

$$I_2(\alpha;\tilde{\delta}) = \frac{\sqrt{1-\tilde{\delta}^2}}{2i\tilde{\delta}} [e^{i\tilde{\delta}\alpha} \mathrm{Ei}(-i\tilde{\delta}\alpha) - e^{-i\tilde{\delta}\alpha} \mathrm{Ei}(i\tilde{\delta}\alpha)] + \sum_{l=0}^{\infty} (-1)^l \tilde{\delta}^{2l} q_{2l+2}(\alpha). \tag{C25}$$

By virtue of (C1) and (C3),

$$\varphi(\alpha;\beta)=-(1-\tilde{\delta}^2)e^{i\tilde{\delta}\alpha}\mathrm{Ei}(-i\tilde{\delta}\alpha)-i\tilde{\delta}\sqrt{1-\tilde{\delta}^2}\sum_{l=0}^{\infty}(-1)^l\tilde{\delta}^{2l}q_{2l+2}(\alpha), \tag{C26}$$

which is Eq. (52). This result can be analytically continued to complex values of $\tilde{\delta}=-i\delta$ and α ($|\arg\alpha|<\pi$).

[1] J. R. Carson, Bell Syst. Tech. J. **5**, 539 (1926).
[2] E. Sunde, *Earth Conduction Effects in Transmission Line Systems* (Van Nostrand, New York, 1949).
[3] J. R. Wait, Radio Sci. **7**, 675 (1972); see also J. R. Wait, IEEE Trans. Electromagn. Compat. **38**, 608 (1996).
[4] D. C. Chang and J. R. Wait, IEEE Trans. Commun. **22**, 421 (1974).
[5] E. F. Kuester, D. C. Chang, and S. W. Plate, Electromagnetics **1**, 243 (1981).
[6] M. Wu, R. G. Olsen, and S. W. Plate, J. Electromagn. Waves Appl. **4**, 479 (1990).
[7] B. L. Coleman, Philos. Mag. **41**, 276 (1950).
[8] R. W. P. King, T. T. Wu, and L.-C. Shen, Radio Sci. **9**, 701 (1974).
[9] R. G. Olsen and T. A. Pankaskie, IEEE Trans. Power Appar. Syst. **102**, 769 (1983).
[10] R. W. P. King and T. T. Wu, J. Appl. Phys. **78**, 668 (1995).
[11] M. D'Amore and M. S. Sarto, IEEE Trans. Electromagn. Compat. **38**, 127 (1996).
[12] M. D'Amore and M. S. Sarto, IEEE Trans. Electromagn. Compat. **38**, 139 (1996).
[13] R. W. P. King, M. Owens, and T. T. Wu, *Lateral Electromagnetic Waves: Theory and Applications to Communication, Geophysical Exploration, and Remote Sensing* (Springer-Verlag, New York, 1992).
[14] T. T. Wu, J. Appl. Phys. **28**, 299 (1957).
[15] D. S. Phatak, N. K. Das, and A. P. Defonzo, IEEE Trans. Microwave Theory Tech. **38**, 1719 (1990).
[16] N. K. Das and D. M. Pozar, IEEE Trans. Microwave Theory Tech. **39**, 54 (1991).
[17] J.-Y. Ke, I.-S. Tsai, and C. H. Chen, IEEE Trans. Microwave Theory Tech. **40**, 1970 (1992).
[18] A. Erdélyi, *Asymptotic Expansions* (Dover, New York, 1956).
[19] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain* (American Mathematical Society, Providence, 1987), p. 2.
[20] Bateman Manuscript Project, *Higher Transcendental Functions, Vol. II*, edited by A. Erdélyi (Krieger, Malabar, FL, 1981).
[21] Bateman Manuscript Project, *Higher Transcendental Functions, Vol. I*, edited by A. Erdélyi (Krieger, Malabar, FL, 1981).
[22] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves* (Springer-Verlag, Berlin, 1969), p. 136.
[23] T. T. Wu, Cruft Laboratory, Tech. Report No. 211, Harvard University, Cambridge, MA, 1954.
[24] C. W. Horton, J. Math. Phys. **29**, 31 (1950).
[25] A. R. Miller, J. Franklin Inst. **329**, 563 (1992).
[26] Y. L. Luke, *The Special Functions and their Approximations, Vol. I* (Academic, London, 1969), p. 195.
[27] O. I. Marichev, *Integral Transforms of Higher Transcendental Functions* (Horwood, Chicester, England, 1983).
[28] R. J. Sasiela and J. D. Shelton, J. Math. Phys. **34**, 2572 (1993).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 11 NOVEMBER 1998

Point vortices on a sphere: Stability of relative equilibria

Sergey Pekarsky[a] and Jerrold E. Marsden[b]
Control and Dynamical Systems 107-81, California Institute of Technology, Pasadena, California 91125

(Received 27 January 1998; accepted for publication 10 March 1998)

In this paper we analyze the dynamics of N point vortices moving on a sphere from the point of view of geometric mechanics. The formalism is developed for the general case of N vortices, and the details are worked out for the (integrable) case of three vortices. The system under consideration is SO(3) invariant; the associated momentum map generated by this SO(3) symmetry is equivariant and corresponds to the moment of vorticity. Poisson reduction corresponding to this symmetry is performed; the quotient space is constructed and its Poisson bracket structure and symplectic leaves are found explicitly. The stability of relative equilibria is analyzed by the energy–momentum method. Explicit criteria for stability of different configurations with generic and nongeneric momenta are obtained. In each case a group of transformations is specified, modulo which one has stability in the original (unreduced) phase space. Special attention is given to the distinction between the cases when the relative equilibrium is a nongreat circle equilateral triangle and when the vortices line up on a great circle. © *1998 American Institute of Physics.* [S0022-2488(98)01906-9]

I. INTRODUCTION

The problem of vortex motion has a long and interesting history. It was Helmholtz[1] who introduced the model that is referred to today as point vortices. Several of Helmholtz' contemporaries immediately seized upon and developed the treasures in his paper, such as Kirchhoff[2] and his student Gröbli. An account of some of the history of this problem can be found in Aref, Rott and Thomann[3] and Kidambi and Newton.[4]

We mention a few more historical facts relevant to the present paper. The problem of configurations of vortices that could move without change of shape, namely *relative equilibria* in the language of Poincaré, was analyzed by Thomson,[5] the later Lord Kelvin, and stability aspects of this motion were studied in his later paper, Thomson.[6] The geometric construction was rediscovered, updated and added to by Novikov[7] a century later for the case of equal strength vortices. Synge[8] developed a qualitative classification of all possible motions of three planar vortices and was the first to introduce trilinear coordinates. Aref[9] also treats the case of three vortices of general strength.

The paper by Bogomolov[10] contains the first systematic and thorough derivation of the equations of motion for point vortices on both rotating and nonrotating spheres. A later paper of Bogomolov[11] contains an analysis of the motion of three identical point vortices on the sphere, generalizing the planar result by Novikov.[7] The paper by Kidambi and Newton[4] treats the case of three vortices on a sphere for general vortex strengths, thus generalizing the planar results of Synge[8] and Aref.[9]

The topology of the problem of vortices moving on a sphere is considered by Kirwan,[12] though this paper mainly deals with the symplectic reduction (in the sense of Marsden and Weinstein[13]) of the problem and the study of the topology of the symplectically reduced phase spaces and the number of equilibria, by computing, in the spirit of Smale,[14] some topological invariants, such as Betti numbers.

The dynamics of N vortices on a sphere is a Hamiltonian system (see, e.g., Kidambi and Newton[4] and references therein). This Hamiltonian structure can be obtained using general reduc-

[a]Electronic mail: sergey@cds.caltech.edu
[b]Electronic mail: marsden@cds.caltech.edu

 © 1998 American Institute of Physics

tion techniques starting with the geometrical description of ideal hydrodynamics in terms of diffeomorphism groups; see Marsden and Weinstein[13] and Arnold and Khesin.[15]

In this paper we explicitly carry out Poisson reduction for the 3-vortex problem on a sphere. We calculate the induced Poisson structure on the Poisson reduced space and analyze the associated symplectic stratification. Furthermore, relative equilibria are classified and their stability is determined by the energy–momentum method (see Marsden[16] and references therein). The use of the energy–momentum method for the stability of vortices was studied for certain planar cases by Lewis and Ratiu.[17] As in the basic example of the rigid body, stability in the reduced space can also be studied by considering intersections of the energy levels with the symplectic leaves.

A. The phase space and its Poisson structure

The phase space of the dynamical system of N-vortices moving on the two sphere consists of N copies of a sphere, namely $P=S^2\times\cdots\times S^2$ regarded as being embedded in N copies of three space $\mathbb{R}^{3N}$ as the set defined by $\|\mathbf{x}_n\|=R$, where R is the radius of the sphere and $n=1,\ldots,N$ labels the location of the nth vortex. There are nonzero vortex strengths Γ_n ascribed to each vortex; i.e., to each S^2. The Poisson structure on P is given by

$$\{\cdot,\cdot\}=\sum_{n=1}^{N}\frac{R}{\Gamma_n}\{\cdot,\cdot\}_n, \tag{1}$$

where $\{\ ,\ \}_n$ is the Poisson structure on the nth copy of S^2 corresponding to the natural area symplectic form on S^2; that is, the Poisson structure in each copy of $\mathbb{R}^3$ is the standard Lie–Poisson structure on $\mathbb{R}^3$ considered as $\mathfrak{so}(3)^*$, the dual of the Lie algebra of the rotation group SO(3). (See, for example, Marsden and Ratiu[18] for the basic definitions used here.) The restriction of the Lie–Poisson bracket operation on $\mathbb{R}^3$ to a sphere [which is a symplectic leaf in $\mathfrak{so}(3)^*$] defines an area form. For two functions F,H on the nth copy of $\mathbb{R}^3$, the Lie–Poisson structure is

$$\{F,H\}_n(\mathbf{x}_n)=-\mathbf{x}_n\cdot(\nabla_n F\times\nabla_n H), \tag{2}$$

where $\times$ denotes the vector cross product.

B. The symmetry group and momentum map

Consider the diagonal action of the group SO(3) on P defined by rotations in each $\mathbb{R}^3$. This action is canonical with respect to the Poisson structure (1). The corresponding Lie algebra is naturally identified with $\mathbb{R}^3$ (having the vector product as its Lie bracket operation) and we write $\boldsymbol{\xi}$ for the vector in $\mathbb{R}^3$ corresponding to the matrix $\xi\in\mathfrak{so}(3)$; thus,

$$\mathfrak{so}(3)\simeq(\mathbb{R}^3,\times),\quad \text{i.e.,}\quad [\xi,\eta]=\boldsymbol{\xi}\times\boldsymbol{\eta},\quad \text{for any}\quad \xi,\eta\in\mathfrak{so}(3). \tag{3}$$

The vector field of infinitesimal transformations corresponding to an element ξ in the Lie algebra is given by

$$\xi_P(\mathbf{x}):=\frac{d}{dt}\exp(\xi t)\cdot\mathbf{x}\Big|_{t=0}=(\boldsymbol{\xi}\times\mathbf{x}_1,\ldots,\boldsymbol{\xi}\times\mathbf{x}_N). \tag{4}$$

The space $\mathfrak{so}(3)^*$, which, as mentioned above, is the dual of the Lie algebra $\mathfrak{so}(3)$, is equipped with the natural Lie–Poisson structure given by (2) (after identifying the dual of $\mathbb{R}^3$ with itself using the standard dot product on $\mathbb{R}^3$).

Recall that a *momentum map* $\mathbf{J}:P\to\mathfrak{so}(3)^*\simeq\mathbb{R}^3$ for this action is defined in terms of the Poisson bracket of an arbitrary function F on P by

$$\{F,J(\xi)\}=\xi_P[F],$$

where $J:\mathfrak{so}(3)\to C^\infty(P)$ is related to $\mathbf{J}$ by

$$J(\xi)(z)=\langle\mathbf{J}(z),\xi\rangle,$$

for all $\mathbf{x}\in\mathbf{P}$, $\xi\in\mathfrak{so}(3)$, and where $\langle\cdot,\cdot\rangle$ is the natural paring between the Lie algebra and its dual.

It is readily checked that the momentum map is proportional to the moment of vorticity and is given by

$$\mathbf{J}(\mathbf{x})=-\frac{1}{R}\sum_{n=1}^{N}\Gamma_n\mathbf{x}_n. \tag{5}$$

Denote its components by $\mathbf{J}=(\mathscr{Q},\mathscr{P},\mathscr{S})$. The momentum map is *not surjective* since

$$\max_n|\Gamma_n|-\sum_{m\neq n}|\Gamma_m|<\|\mathbf{J}\|<\sum_n|\Gamma_m|. \tag{6}$$

Denote the range of $\mathbf{J}$ by $\mathscr{R}$.

The momentum map is *equivariant*, that is,

$$\mathrm{Ad}^*_{g^{-1}}(\mathbf{J}(\mathbf{x}))=\mathbf{J}(g(\mathbf{x})), \tag{7}$$

for all $g\in\mathrm{SO}(3)$. Here, the map $\mathrm{Ad}^*_k:\mathfrak{so}(3)^*\to\mathfrak{so}(3)^*$, defined for each $k\in\mathrm{SO}(3)$, denotes the coadjoint action of SO(3) on $\mathfrak{so}(3)^*$. That one can find an *equivariant* momentum map is consistent with general theorems for compact or semisimple groups. In our case, this can be seen directly from (5) as the coadjoint action corresponds simply to rotations in the dual space $\mathfrak{so}(3)^*\simeq\mathbb{R}^3$.

It follows from equivariance of $\mathbf{J}$ or directly, that $\|\mathbf{J}\|^2$ is invariant under the coadjoint action. Hence, smooth functions of $\|\mathbf{J}\|^2$ are also invariant. Thus, if $\mathbf{b}=(b_1,b_2,b_3)\in\mathbb{R}^3$ are coordinates in the dual $\mathfrak{so}(3)^*$, then any smooth function of $\|\mathbf{b}\|^2$ is a Casimir function. Correspondingly, the generic symplectic leaves of $\mathfrak{so}(3)^*$ are spheres defined by the level sets $\|\mathbf{J}\|^2=\mathrm{const}\neq 0$. Note that as SO(3) is compact, the action of it on both P and $\mathfrak{so}(3)^*$ is proper.

C. The Hamiltonian

The Hamiltonian describing the motion of N vortices on the surface of a sphere of radius R is given by (see, e.g., Kidambi and Newton[4])

$$H=\frac{1}{4\pi R^2}\sum_{m<n}\Gamma_m\Gamma_n\ln(l^2_{mn}), \tag{8}$$

where $l^2_{mn}=2(R^2-\mathbf{x}_m\cdot\mathbf{x}_n)$ is the square of the chord distance between two vortices with positions $\mathbf{x}_m$ and $\mathbf{x}_n$. The constraints $\|\mathbf{x}_n\|=R$ are assumed. Notice that the Hamiltonian is invariant with respect to the diagonal action of SO(3) on P described above. Hence, the momentum map $\mathbf{J}$ defined by (5) is constant along the flow of this Hamiltonian.

II. POISSON AND SYMPLECTIC REDUCTION

A. Poisson quotients

In performing Poisson reduction, one normally constructs the quotient space by the symmetry group and calculates its naturally induced Poisson bracket. As is well known, singularities in the quotient space may arise if the group action on the phase space is not free. Strictly speaking, the phase space of the problem is not $S^2\times\cdots\times S^2$ but rather

$$S^2\times\cdots\times S^2\ \Big\backslash\ \sum_k\ \cup_{i_n\neq i_m}\Delta_{i_1\cdots i_k},$$

where $\Delta_{i_1\cdots i_k}=\{\mathbf{x}|$ any two or more of $\mathbf{x}_{i_n}$ coincide$\}$. This is because the self-interaction and collision of vortices (which lead to infinite energy) have been excluded from consideration. This restriction guarantees that the diagonal action of SO(3) on P is free provided $N\geqslant 3$, i.e., there are 3 or more vortices. The action is also proper, as was mentioned above.

Thus, *the quotient* $T=P/\mathrm{SO}(3)$ *is a smooth* $2N-3$ *dimensional Poisson manifold*. In coordinatizing this quotient we shall use the quantities l_{mn}^2, which are functions on P that are invariant with respect to the SO(3) action, with the conditions $l_{mn}^2\neq 0$. In general, there are $1+2(N-2)=2N-3$ independent functions l_{mn}^2, and other invariant functions can be expressed in terms of them.

To describe a configuration of N vortices on a sphere (up to a global rotation), it is sufficient to specify the chord distance between some two vortices and the chord distances from the remaining $N-2$ to those two (to remove the ambiguity of reflection consider, for example, a stereographic projection and choose two vortices such that all the rest lie to one side of the line connecting those two).

When 3 or more vortices are aligned on a great circle, this coordinate system is degenerate, i.e., there are less than $2N-3$ independent functions l_{mn}^2, and so we shall introduce other coordinates in the neighborhood of such points in the quotient space. Specifically, it is easy to see that the differentials of the three square distances associated to three vortices are linearly dependent when the three vortices lie on a great circle. This analysis, obviously, agrees with the dimension of the Poisson quotient. Also, the variables l_{mn}^2 naturally appear in the Hamiltonian for the N-vortex problem on the sphere, which makes the calculation of the reduced Hamiltonian h easy.

It follows from (5) that the square of the momentum map is given by

$$\mathbf{J}^2=:\|\mathbf{J}\|^2=\left(\sum\ \Gamma_n\right)^2-\frac{1}{R^2}\sum_{n<m}\Gamma_n\Gamma_m l_{nm}^2, \tag{9}$$

which, as we mentioned, is invariant under the SO(3) action. Other invariants are given by l_{nm}^2. Denote $\Gamma=\Sigma\Gamma_n$ and define a map $\Phi_\mu: T\to\mathbb{R}$ by

$$\Phi_\mu=(\mu-\Gamma)(\mu+\Gamma)+\frac{1}{R^2}\sum_{n<m}\Gamma_n\Gamma_m l_{nm}^2. \tag{10}$$

Notice that the relation (9) between the variables l_{nm}^2 and $\mathbf{J}$ can be expressed as $\Phi_J(l_{nm^2})=0$.

B. Reduction for the 3-vortex problem

Now we consider the 3-vortex problem and the structure of the corresponding Poisson reduced space in more detail. The phase space of the 3-vortex problem is trivial in a sense that it is diffeomorphic to a product of SO(3) with a ''shape-phase space'' U; that is, $P\cong\mathrm{SO}(3)\times U$, where

$$U=\{(a,\alpha_1,\alpha_2)\,|\,-R<a<R, 0<\alpha_1+\alpha_2<2\pi, \alpha_1<\alpha_2\}\subset\mathbb{R}^3. \tag{11}$$

Here, a can be interpreted as the height of the triangle of the vortices with respect to the sphere and α_n corresponds to an angle opposite the nth vortex. For the computations we will use another atlas which consists of three charts, two of which are nearly identical—they differ only in the orientation and are connected by a $\mathbb{Z}_2$ reflection. That is, for the same chord distances vectors $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$ can form a right-handed or left-handed coordinate system, corresponding to different orientations and thus defining two different configurations, one in each of these two charts.

Denote the coordinates on these charts by $a_1=l_{23}^2, a_2=l_{13}^2, a_3=l_{12}^2$, so that the a_n are the squares of the sides of the triangle inscribed in a circle of radius $r<R$. Thus, all admissible values of a_n can be parameterized by any two angles α_n,α_m. The chart can be given parametrically by an open set $\mathcal{T}\subset\mathbb{R}^3$ defined as the set of triples (a_1,a_2,a_3) given by

$$a_1=2r^2(1-\cos\,\alpha_1),\qquad a_2=2r^2(1-\cos\,\alpha_2),\qquad a_3=2r^2(1-\cos(\alpha_1+\alpha_2)), \tag{12}$$

where $0<\alpha_1+\alpha_2<2\pi$, $0<r<R$. The third chart contains an open neighborhood of the set of great circles and smoothly connects different orientations. Indeed, for great circles $\mathbf{x}_n$ become linearly dependent, and $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$ fail to define either the right- or left-handed coordinate system. The chart can be coordinatized by $V=\mathbf{x}_1\cdot(\mathbf{x}_2\times\mathbf{x}_3)$, i.e., the orientable volume of the parallelepiped formed by the vectors $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$, and any two chord distances a_n,a_m. The sign of V deter-

mines the orientation (by distinguishing between right- and left-handed coordinate systems) and thus specifies one of the above two charts, while $V=0$ corresponds to the great circles. The change of coordinates is checked to be nondegenerate in the open intersections of the charts.

We summarize our results on Poisson reduction for the 3-vortex problem in the following.

Proposition II.1 (Poisson reduction for the 3-vortex problem): The quotient $T=P/\mathrm{SO}(3)$ is a smooth 3 dimensional manifold diffeomorphic to the shape phase space U defined by (11). The natural projection of P to T is a surjective submersion with fibers being the SO(3)-orbits on P.

The manifold T carries the quotient Poisson structure given as follows in the coordinate charts described above. Let f and h be given functions defined on the set $\mathscr{T}\subset T$ and let (a_1,a_2,a_3) lie in the set $\Phi_\mu(a_1,a_2,a_3)=0$. Then

$$\{f,h\}_T(a_1,a_2,a_3)=\frac{4R^3V}{\Gamma_1\Gamma_2\Gamma_3}\nabla\Phi_\mu\cdot(\nabla f\times\nabla h), \tag{13}$$

where V is regarded as a function of **a**; *its sign, which corresponds to different orientations, distinguishes between the two charts $\mathscr{T}$. The Poisson bracket along the set of great circles is given by the following expression:*

$$\{f,h\}_T(V,a_2,a_3)=B_2\left(\frac{\partial f}{\partial a_2}\frac{\partial h}{\partial V}-\frac{\partial f}{\partial V}\frac{\partial h}{\partial a_2}\right)+B_3\left(\frac{\partial f}{\partial a_3}\frac{\partial h}{\partial V}-\frac{\partial f}{\partial V}\frac{\partial h}{\partial a_3}\right). \tag{14}$$

Here

$$B_2=4R\left(\frac{2(a_1+a_2-a_3)R^2-a_1a_2}{\Gamma_1}-\frac{2(a_2+a_3-a_1)R^2-a_2a_3}{\Gamma_3}\right)$$

and

$$B_3=4R\left(\frac{2(a_2+a_3-a_1)R^2-a_2a_3}{\Gamma_2}-\frac{2(a_1+a_3-a_2)R^2-a_1a_3}{\Gamma_1}\right),$$

in which a_1 is regarded as a function of a_2,a_3 (since they are dependent when $V=0$).

Casimir functions on T are generated by Φ_μ; that is, any function of Φ_μ is a Casimir function. The level sets, $\Phi_\mu=0$, determine the symplectic leaves; these leaves are isomorphic to the symplectic-reduced spaces $P_\mu=\mathbf{J}^{-1}(\mu)/\mathrm{SO}(3)_\mu$. The generic leaves are those not containing the great circle equilibria with $\mathbf{J}=\mathbf{0}$ and are open planes that foliate $\mathscr{T}$. For every fixed choice of Γ_n they are parallel to each other and none of them contains the central line $a_1=a_2=a_3$. If $0\in Range\ \mathbf{J}$, then there is a unique nongeneric zero dimensional symplectic leaf that corresponds to a great circle configuration with $\mathbf{J}=0$.

Proof: Define $F=f\circ\pi$, where $\pi:P\to T$ is the projection. The Poisson bracket on P is given by (1) and (2). One computes, in a straightforward way, $\{F,H\}$ using the chain rule to get (13). Then (14) is obtained upon change of coordinates in the chart intersections and setting $V=0$ afterwards; we omit here the required simple but tedious calculations.

The structure of generic symplectic leaves follows from the linearity of the Casimir function (10). □

A Hamiltonian H on P that is invariant under the diagonal action of SO(3) induces a reduced Hamiltonian h on $T=P/\mathrm{SO}(3)$. The corresponding reduced equations on the leaves $\Phi_\mu=0$ in $\mathscr{T}$ are checked to be given by the following (Euler-like) equations (see Kummer[19] and Kirk and Marsden and Silber[20]):

$$\dot{\mathbf{a}}=\frac{4R^3V}{\Gamma_1\Gamma_2\Gamma_3}\nabla h\times\nabla\Phi_\mu, \tag{15}$$

where $\mathbf{a}=(a_1,a_2,a_3)$.

For the Hamiltonian (8) the reduced equations are

$$\dot{a}_i = \frac{V}{\pi R}\Gamma_i\left(\frac{1}{a_j}-\frac{1}{a_k}\right), \tag{16}$$

where (i,j,k) is a cyclic permutation of (1,2,3). Along the set of great circles the equations are represented in a different way, as is the Poisson bracket; in fact, they are given by

$$\dot{V} = \frac{1}{8\pi}\left(2R\left(\frac{a_3-a_1}{a_2}(\Gamma_1+\Gamma_3)+\frac{a_1-a_2}{a_3}(\Gamma_2+\Gamma_1)+\frac{a_2-a_3}{a_1}(\Gamma_3+\Gamma_2)\right)\right.$$
$$\left.-\frac{1}{R}(a_3(\Gamma_1-\Gamma_2)+a_2(\Gamma_3-\Gamma_1)+a_1(\Gamma_2-\Gamma_3))\right), \tag{17}$$

together with $\dot{a}_2=0$ and $\dot{a}_3=0$.

These results reproduce, in the spirit of geometric mechanics, some of the results of Kidambi and Newton.[4] For instance, the second invariant in this reference is interpreted as a linear function of the square of the momentum map, $\|\mathbf{J}\|^2$. They differ only in an overall factor and an additive constant. As it was mentioned above, $\|\mathbf{J}\|^2$ determines the symplectic leaves in $\mathfrak{so}(3)^*$ and naturally leads to conserved quantities.

III. STABILITY OF RELATIVE EQUILIBRIA

A. The energy–momentum method

We will now utilize the energy–momentum method (see Marsden[16] for a summary and references) for the analysis of the stability of *relative equilibria*, i.e., dynamical orbits with initial conditions $\mathbf{x}_e$ such that $\mathbf{x}(t)=\exp(\xi_e t)\mathbf{x}_e$ for some Lie algebra element ξ_e and any time t. As is well known for relative equilibria, the *augmented energy* function $H_{\xi_e}:=H-\langle \mathbf{J}-\mu_e,\xi_e\rangle$ has a critical point at $\mathbf{x}_e$, where $\mu_e=\mathbf{J}(\mathbf{x}_e)$ is the value of the momentum at the relative equilibrium. For notational convenience we will occasionally omit the subscript e.

The orbital stability of a relative equilibrium is equivalent to the stability of the corresponding equilibrium of the reduced system that is induced on the symplectic leaves P_μ of the quotient manifold $P/\mathrm{SO}(3)$. The energy momentum method is designed to enable one to test for orbital stability directly on the unreduced manifold P by constructing a subspace $\mathscr{G}\subset T_{\mathbf{x}_e}P$ which is isomorphic to $T_{\mathbf{x}_e}P_{\mu_e}$. This is done by considering a tangent space to the level set of constant momentum $\mathbf{J}^{-1}(\mu_e)$ and eliminating the neutrally stable directions associated to the isotropy subgroup,

$$\mathrm{SO}(3)_{\mu_e}:=\{g\in \mathrm{SO}(3)\,|\,\mathrm{Ad}^*_g\mu_e=\mu_e\}.$$

The energy–momentum method determines stability by examining definiteness of the second variation of H_{ξ_e} restricted to the subspace $\mathscr{G}$. A detailed description of this method can be found in Simo, Lewis and Marsden.[21]

If one has a definite second variation, then Patrick's theorem (see Patrick[22]) guarantees stability modulo the isotropy subgroup, provided its action on P is proper and the Lie algebra admits an inner product invariant under the adjoint action of the isotropy subgroup. Note that as SO(3) is compact, the assumptions of Patrick's theorem are automatically satisfied for our applications.

As was mentioned above, relative equilibria are critical points of the augmented Hamiltonian H_ξ. For variational calculations, we extend all functions on P to functions on the ambient space $\mathbb{R}^{3N}$, and then restrict variations to the tangent space to P by requiring

$$\delta F(\mathbf{x})\cdot\boldsymbol{\eta}=0,$$

for all $\boldsymbol{\eta}\in T_{\mathbf{x}}P$. For the augmented Hamiltonian corresponding to (8), this results in the following conditions on $\mathbf{x}$:

$$\frac{\Gamma_r}{R}\left(\boldsymbol{\xi}(\mathbf{x})-\frac{1}{2\pi R}\sum_{n\neq r}\Gamma_n\frac{\mathbf{x}_n}{l_{nr}^2}\right)=\kappa_r\frac{\Gamma_r}{R^2}\mathbf{x}_r, \tag{18}$$

where κ_r are constants to be determined.

B. Equidistant relative equilibria

An ***equidistant configuration*** is, by definition, one that satisfies $l_{mn}^2 = l^2$ for all $m \neq n$. Whatever its dynamics, such a configuration is possible only for $N = 2,3,4$ (this follows by geometric arguments similar to those used for the study of regular polytopes in three space); we exclude the simple case $N = 2$ from our considerations.

To verify that an equidistant configuration is a relative equilibrium, one need only check that conditions (18) are satisfied. It is easy to see that

$$\boldsymbol{\xi}(\mathbf{x}) = \frac{1}{2\pi R l^2} \sum_n \Gamma_n \mathbf{x}_n = -\frac{1}{2\pi l^2} \mathbf{J}(\mathbf{x}) \tag{19}$$

solves (18) with $\kappa_r = \Gamma_r / 2\pi l^2$. Notice, that in (19) the vectors $\boldsymbol{\xi}$ and $\mathbf{J}$ have opposite directions. These observations prove the following.

Proposition III.1: Equidistant configurations of relative equilibria satisfying $\mathbf{J}(\mathbf{x}_e) \neq 0$ *are possible only for* $N = 3$ *and* 4 *and are given by equilateral triangles and a tetrahedron, respectively; the associated values of the momentum and the Lie algebra element for these relative equilibria satisfy (19).*

Condition (18) together with $\boldsymbol{\xi} = 0$ defines ***static equilibria***. It follows from (19) that equidistant static equilibria are possible only in the degenerate case of zero momentum. This necessarily implies for $N = 3$ that the vortices lie on a great circle, and for both $N = 3$ and $N = 4$ that all Γ_n are equal, i.e., $\Gamma_n = \Gamma$. Moreover, a tetrahedral configuration with zero momentum $\mathbf{J} = 0$ is necessarily a static equilibrium.

C. Great circle relative equilibria

For $N = 3$ vortices, we have the following classification of ***great circle equilibria*** (see Kidambi and Newton[4]); recall the notations $a_1 = l_{23}^2, a_2 = l_{13}^2, a_3 = l_{12}^2$.

1. Generic momentum, $\mathbf{J}(\mathbf{x}_e) \neq 0$

General relative equilibria correspond to vortices lying on a great circle (and thus satisfying $V = 0$) and also satisfying the following condition:

$$2R\left(\frac{a_3 - a_1}{a_2}(\Gamma_1 + \Gamma_3) + \frac{a_1 - a_2}{a_3}(\Gamma_2 + \Gamma_1) + \frac{a_2 - a_3}{a_1}(\Gamma_3 + \Gamma_2)\right)$$
$$-\frac{1}{R}(a_3(\Gamma_1 - \Gamma_2) + a_2(\Gamma_3 - \Gamma_1) + a_1(\Gamma_2 - \Gamma_3)) = 0, \tag{20}$$

obtained by setting $\dot{V} = 0$ in (17). This implicit formula determines another relation (in addition to $V = 0$), between a_1, a_2 and a_3 for each fixed set of Γ's. This is a nonlinear equation and thus can have multiple solutions.

(a) Isosceles triangular great circle equilibria. A particular family of ***isosceles triangular relative equilibria*** *for arbitrary values of* Γ*'s is given by the following configuration:*

$$a_1 = a_2 = 2R^2, \quad a_3 = 4R^2, \tag{21}$$

or, equivalently, $\alpha_1 = \alpha_2 = \pi/2$, $\alpha_3 = \pi$, as well as configurations obtained from it by cyclic permutations of indices. The whole configuration rotates around the vector

$$\boldsymbol{\xi}(\mathbf{x}) = -\frac{1}{4\pi R^2} \mathbf{J}(\mathbf{x}) \tag{22}$$

and the constants κ_n in (18) are given by

$$\kappa_1=\frac{\Gamma_1}{4\pi R^2}-\frac{\Gamma_2}{8\pi R^2},\quad \kappa_2=\frac{\Gamma_2}{4\pi R^2}-\frac{\Gamma_1}{8\pi R^2},\quad \kappa_3=\frac{\Gamma_3}{4\pi R^2}.$$

(b) Equilateral triangular great circle equilibria. A ***great circle equilateral triangle relative equilibrium*** with $l_{mn}^2=l^2=3R^2$ and ξ given by (19).

Note: When the term *equilateral triangle relative equilibrium* is used, and we do not append ''great circle,'' we will mean that it is a *nongreat circle equilateral triangle relative equilibrium.*

2. Degenerate momentum, $\mathbf{J}(\mathbf{x}_e)=\mathbf{0}$

In this case, the vortices again lie on a great circle, and the whole configuration rotates around the vector

$$\boldsymbol{\xi}(\mathbf{x})=-\frac{1}{2\pi R}\left(\frac{\Gamma_1\mathbf{x}_1}{l_{23}^2}+\frac{\Gamma_2\mathbf{x}_2}{l_{13}^2}+\frac{\Gamma_3\mathbf{x}_3}{l_{12}^2}\right). \tag{23}$$

Remarks.

(1) Another specific family of great circle solutions can be found in case two of the Γ's are equal; for instance, $\Gamma_1=\Gamma_2$. In fact, any isosceles triangle with the corresponding sides of the triangle being also equal, that is, $a_1=a_2$ for $\Gamma_1=\Gamma_2$, solves to (20) and hence is a relative equilibrium for any value of Γ_3.

(2) If we consider the ''inverse'' problem, namely, given a configuration on a great circle find Γ_n satisfying (20) so that this configuration is a relative equilibrium, then condition (20) becomes a linear equation in Γ_n of the form
$$\beta_1\Gamma_1+\beta_2\Gamma_2+\beta_3\Gamma_3=0,$$
where $\beta_n=\beta_n(a_1,a_2,a_3)$ are functions of a great circle configuration. One would expect this to have a two parameter family of solutions.

The structure of the symplectic leaves sheds light on the stability of relative equilibria of the system. In particular, generically, great circle configurations satisfying (20) form a family of one dimensional curves in the Poisson manifold T that intersect symplectic leaves in a point. Similarly, equilateral configurations are isolated points within the symplectic leaves, and stability analysis is done by restricting a proper second variation to the tangent space to these leaves.

D. Geometry of the tangent space of phase space

Following the outline in the beginning of this section, consider a generic *regular* relative equilibrium $\mathbf{x}_e$, that is, its symmetry subgroup is finite, i.e., for each nonzero element $\boldsymbol{\xi}$ of the Lie algebra, the corresponding infinitesimal generator evaluated at $\mathbf{x}_e$, denoted $\boldsymbol{\xi}_P(\mathbf{x}_e)$, is nonzero. Then, the isotropy subgroup of the corresponding nonzero momentum value $\mu_e=\mathbf{J}(\mathbf{x}_e)$ is the group SO(2) of rotations around the vector $\mathbf{J}$. For $\mu_e=0$ the isotropy subgroup is SO(3) itself; (the stability in this is case is simple and will be considered in the end of the section). The isotropy Lie subalgebra is defined by

$$\mathfrak{so}(3)_{\mu_e}=\left\{\boldsymbol{\xi}\in\mathbb{R}^3\,\middle|\,\boldsymbol{\xi}=\varrho\,\mathbf{J}(\mathbf{x}_e)=-\frac{\varrho}{R}\sum_n\Gamma_n\mathbf{x}_{e,n},\ \ \varrho\ \text{ a constant}\right\}. \tag{24}$$

Hence, the tangent space to the $\mathrm{SO}(3)_\mu$ orbit at $\mathbf{x}_e$, which corresponds to the neutrally stable direction, is given by

$$T_{\mathbf{x}_e}(\mathrm{SO}(3)_{\mu_e}\cdot\mathbf{x}_e)=\{\boldsymbol{\xi}\times\mathbf{x}_e\,|\,\boldsymbol{\xi}=\varrho\,\mathbf{J}(\mathbf{x}_e)\}, \tag{25}$$

where again ϱ is a constant. For regular relative equilibria, $\mathrm{Ker}D\mathbf{J}(\mathbf{x}_e)=T_{\mathbf{x}_e}\mathbf{J}^{-1}(\mu_e)$. The derivative of the momentum map $D\mathbf{J}$ as a mapping from TP to $T\mathbb{R}^3$ can be easily computed from (5) to produce

$$D\mathbf{J}(\mathbf{x})\cdot\mathbf{y}=-\frac{1}{R}\sum_n \Gamma_n\mathbf{y}_n,$$

where $\mathbf{y}:=(\mathbf{y}_1,\ldots,\mathbf{y}_N)\in T_{\mathbf{x}}P$ and $\mathbf{y}_n\in T_{\mathbf{x}_n}S^2$ is a tangent vector to the sphere S^2 at the point $\mathbf{x}_n$. Thus, the kernel is determined by the following condition:

$$\mathrm{Ker}D\mathbf{J}(\mathbf{x})=\left\{\mathbf{y}\in T_{\mathbf{x}}P\,\middle|\,\sum_n \Gamma_n\mathbf{y}_n=0\right\}, \tag{26}$$

and is $2N-3$ dimensional.

Using Eqs. (25) and (26) it is easy to see that

$$T_{\mathbf{x}_e}(\mathrm{SO}(3)_\mu\cdot\mathbf{x}_e)\subset\mathrm{Ker}D\mathbf{J}(\mathbf{x}_e).$$

Indeed,

$$\sum_n \Gamma_n\mathbf{y}_n=\sum_n \Gamma_n\varrho\,\mathbf{J}\times\mathbf{x}_n=\varrho\,\mathbf{J}\times\mathbf{J}=0.$$

We proceed to find a subspace $\mathcal{G}\subset\mathrm{Ker}D\mathbf{J}(\mathbf{x}_e)$ that is transversal to the tangent space to the $\mathrm{SO}(3)_\mu$ orbit at $\mathbf{x}_e$. It is done in the following way. Chose two arbitrary vectors $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ such that the plane through them contains no vortices. Then, tangent vectors at each of the vortices,

$$\mathbf{y}_n^{(1)}:=\gamma_n^{(1)}\mathbf{D}^{(1)}\times\mathbf{x}_n,\quad \mathbf{y}_n^{(2)}:=\gamma_n^{(2)}\mathbf{D}^{(2)}\times\mathbf{x}_n \tag{27}$$

span $T_{\mathbf{x}}P$. Notice that (27) guarantees that all $\mathbf{y}_n^{(i)}$ lie in a plane perpendicular to $\mathbf{D}^{(i)}$. Thus, for each $\mathbf{D}^{(i)}$ there are $N-2$ independent zero linear combinations of $\mathbf{y}_n$'s. Also, it follows from (26) that if the coefficients $\gamma_n^{(i)}$ are chosen to satisfy

$$\sum_n \Gamma_n\gamma_n^{(i)}\mathbf{x}_n=\mathbf{D}^{(i)}\quad\text{or}\quad\sum_n \Gamma_n\gamma_n^{(i)}\mathbf{x}_n=0,\quad i=1,2, \tag{28}$$

then the corresponding tangent vectors belong to the Ker$D\mathbf{J}$.

Any of the equalities in (28) has $N-3+1=N-2$ linearly independent solutions for each $\mathbf{D}^{(i)}$, and, hence, a transversal subspace $\mathcal{G}$ is defined by

$$\mathcal{G}=\mathrm{span}\{\mathbf{y}^{(1)}:=(\gamma_n^{(1)}\mathbf{D}^{(1)}\times\mathbf{x}_n),\ \mathbf{y}^{(2)}:=(\gamma_n^{(2)}\mathbf{D}^{(2)}\times\mathbf{x}_n)\}, \tag{29}$$

and $\dim\mathcal{G}=2N-4$. The isotropy subgroup transformations, i.e., rotations around the axis $\mathbf{J}$, is determined by tangent vectors

$$\mathbf{y}_n:=-\frac{1}{R}\,\mathbf{J}\times\mathbf{x}_n$$

and corresponds to an additional one-dimensional neutrally stable subspace in Ker$D\mathbf{J}$.

We note that special choice of $\mathbf{D}^{(i)}$ would result in a diagonal structure of the second variation of H_ξ. We shall see an instance of this below.

E. Definiteness of the second variation

For the calculation of the second variation the Lagrange multiplier method is used. Define the *extended Hamiltonian* $\tilde{H}_\xi$,

$$\tilde{H}_\xi:=H_\xi+\sum_n \lambda_n(\mathbf{x}_n^2-R^2),$$

where $(\mathbf{x}_n^2 - R^2) = 0$ constrains the motion of vortices to the sphere S^2. The Lagrange multipliers λ_n are determined by the condition $\delta \tilde{H}_\xi(\mathbf{x}_e) = 0$ and are given by

$$\lambda_n = -\frac{\kappa_n \Gamma_n}{2R^2},$$

where κ_n are determined from (18). Then the second variation at $\mathbf{x}_e$ is well-defined as a bilinear form on $T_{\mathbf{x}_e} P$. It is given by the following expression:

$$\frac{\partial^2 \tilde{H}_\xi}{\partial x_s^j \partial x_r^i} = \begin{cases} 2\lambda_r \delta^{ij} - \dfrac{\Gamma_r}{\pi R^2} \displaystyle\sum_{n \neq r} \Gamma_n \frac{x_n^i x_n^j}{l_{nr}^4}, & r = s, \\ -\dfrac{\Gamma_r \Gamma_s}{2\pi R^2 l_{rs}^2} \left(\delta^{ij} + 2 \dfrac{x_s^i x_r^j}{l_{rs}^2} \right), & r \neq s. \end{cases} \tag{30}$$

In the case of an equilateral triangle configuration, when $l_{rs}^2 = l^2$, one can choose

$$\mathbf{D}^{(1)} = \mathbf{x}_1 + \mathbf{x}_2 \quad \text{and} \quad \mathbf{D}^{(2)} = \mathbf{x}_2 + \mathbf{x}_3,$$

as a set of vectors defining a basis of the transversal subspace $\mathcal{G}$ according to (27) with the constants $\gamma_n^{(i)}$ that satisfy conditions (28) being given by $\gamma_1^{(1)} = 1/\Gamma_1, \gamma_2^{(1)} = 1/\Gamma_2, \gamma_3^{(1)} = 0$ and $\gamma_1^{(2)} = 0$, $\gamma_2^{(2)} = 1/\Gamma_2, \gamma_3^{(2)} = 1/\Gamma_3$. Then, the restriction of the second variation to it is given by the following expression:

$$\delta^2 \tilde{H}_\xi \big|_{\mathcal{G}} = \frac{V^2}{\pi R^2 l^4} \begin{pmatrix} -\dfrac{\Gamma_3}{\Gamma_1} - \dfrac{\Gamma_3}{\Gamma_2} & 1 \\ 1 & -\dfrac{\Gamma_1}{\Gamma_2} - \dfrac{\Gamma_1}{\Gamma_3} \end{pmatrix}. \tag{31}$$

The second variation is definite provided $\det(\delta^2 \tilde{H}_\xi)$ is positive. Hence, the following.

Theorem III.2 (stability of nongreat circle equilateral triangles): *An equilateral triangle configuration of nongreat circle relative equilibria* $\mathbf{x}_e$ *is stable modulo* SO(2) *rotations around the vector* $\mathbf{J}(\mathbf{x}_e)$ *if*

$$\sum_{n<m} \Gamma_n \Gamma_m > 0, \tag{32}$$

and unstable if

$$\sum_{n<m} \Gamma_n \Gamma_m < 0. \tag{33}$$

This theorem generalizes the known results of Synge[8] for the stability of equilateral relative equilibria of 3 vortices on a plane. Indeed, conditions (32) and (33) are independent of the radius R. Thus, in the limit $R \to \infty$ the spherical stability conditions agree with those for the planar case.

Conjecture: The condition $\Sigma_{n<m} \Gamma_n \Gamma_m = 0$ *corresponds to a (degenerate) Hamiltonian bifurcation.*

Next we analyze stability for the family of great circle relative equilibria given by (21). Choose

$$\mathbf{D}^{(1)} = \mathbf{x}_1 + \mathbf{x}_3 \quad \text{and} \quad \mathbf{D}^{(2)} : (\mathbf{D}^{(2)}, \mathbf{x}_n) = 0, \ \|\mathbf{D}^{(2)}\| = R,$$

as a set of vectors defining a basis of the transversal subspace $\mathcal{G}$ according to (27) with the constants $\gamma_n^{(i)}$ satisfying conditions (28) being given by $\gamma_1^{(1)} = 1/\Gamma_1, \gamma_2^{(1)} = 0$, $\gamma_3^{(1)} = 1/\Gamma_3$ and $\gamma_1^{(2)} = 1/\Gamma_1, \gamma_2^{(2)} = 1/\Gamma_2, \gamma_3^{(2)} = 0$. Then we obtain the following expression for the restriction of the second variation:

$$\delta^2\tilde{H}_\xi|_{\mathscr{S}}=\frac{1}{8\pi}\begin{pmatrix}\dfrac{\Gamma_2}{\Gamma_1} & 0\\ 0 & \left(\dfrac{\Gamma_2}{\Gamma_1}+\dfrac{\Gamma_1}{\Gamma_2}\right)-2\left(1+\dfrac{\Gamma_3}{\Gamma_1}+\dfrac{\Gamma_3}{\Gamma_2}\right)\end{pmatrix}.$$

Stability then follows from a direct analysis of its definiteness; that is, whether or not the two diagonal entries have the same sign or not. In other words, one has stability if the determinant is positive and instability if it is negative. Carrying out this simple calculation gives the following result.

Theorem III.3 (stability of isosceles triangle great circle equilibria): *A great circle configuration of relative equilibrium* $\mathbf{x}_e$ *given by (21) is* stable *if*

$$\Gamma_1^2+\Gamma_2^2>\sum_{n\neq m}\Gamma_n\Gamma_m \tag{34}$$

and unstable *if*

$$\Gamma_1^2+\Gamma_2^2<\sum_{n\neq m}\Gamma_n\Gamma_m. \tag{35}$$

The stability is modulo SO(2) *rotations around* $\mathbf{J}(\mathbf{x}_e)$.

F. Stability of great circle equilateral triangle relative equilibria

The stability analysis of a GCET, a great circle equilateral triangle relative equilibrium is different from the nongreat circle equilateral triangle case. The reason is that the two-dimensional subspace to which the second variation of the augmented Hamiltonian is restricted in the general case fails to be a transversal subspace to the G_μ orbit (rotations around $\mathbf{J}$) within Ker$D\mathbf{J}$ but rather degenerates to a one-dimensional subspace. A complimentary direction transversal to the plane of the triangle has to be taken into account similar to the case of other great circle relative equilibria. Using the notations developed in the section on the geometry of the tangent space, we choose $\mathbf{D}^{(1)}=n\mathbf{x}_1+m\mathbf{x}_2$ and $\mathbf{D}^{(2)}:(\mathbf{D}^{(2)},\mathbf{x}_n)=0$, $\|\mathbf{D}^{(2)}\|=R$ as a set of vectors defining a basis of the transversal subspace $\mathscr{S}$ according to (27) with the constants $\gamma_n^{(i)}$ satisfying conditions (28) being given by $\gamma_1^{(1)}=n/\Gamma_1, \gamma_2^{(1)}=m/\Gamma_2, \gamma_3^{(1)}=0$ and $\gamma_1^{(2)}=1/\Gamma_1, \gamma_2^{(2)}=1/\Gamma_2, \gamma_3^{(2)}=1/\Gamma_3$.

Using this basis, a straightforward computation gives the following expression for the restriction of the second variation:

$$\delta^2\tilde{H}_\xi|_{\mathscr{S}}=\frac{1}{12\pi}\begin{pmatrix}0 & 0\\ 0 & 9-(\Gamma_1+\Gamma_2+\Gamma_3)\left(\dfrac{1}{\Gamma_1}+\dfrac{1}{\Gamma_2}+\dfrac{1}{\Gamma_3}\right)\end{pmatrix}.$$

One concludes from this that *these GCET equilibria are at best, neutrally stable.*

In the paragraphs below, we explore this in a little more detail and identify the source of the zero eigenvector. Compute the gradient of the Casimir function Φ_μ, given by equation (10), which gives the normal direction to the symplectic leaf:

$$\nabla\Phi_\mu=\left(\frac{\partial\Phi_\mu}{\partial\mathbf{x}_n}\right)=-\frac{2}{R^2}\begin{pmatrix}\Gamma_1\Gamma_2\mathbf{x}_2+\Gamma_1\Gamma_3\mathbf{x}_3\\ \Gamma_1\Gamma_2\mathbf{x}_1+\Gamma_2\Gamma_3\mathbf{x}_3\\ \Gamma_1\Gamma_3\mathbf{x}_1+\Gamma_2\Gamma_3\mathbf{x}_2\end{pmatrix}.$$

Evaluate this gradient at the point corresponding to the GCET, and take the gradient in the direction corresponding to the family of equilateral triangles. To determine such a direction, recall that in the coordinates of the trivialization this family is defined by the following curve: $a=a, \alpha_1=\alpha_2=2\pi/3$, where a is the curve parameter. The tangent vector to this curve is (1,0,0) and so in coordinates of the ambient space, the variation of the GCET configuration along the family of equilateral triangles is given by the following expression:

$$w_{\text{gcet}} := \begin{pmatrix} \mathbf{x}_1 \times \mathbf{x}_2 \\ \mathbf{x}_1 \times \mathbf{x}_2 \\ \mathbf{x}_1 \times \mathbf{x}_2 \end{pmatrix} \in T_x P,$$

i.e., the same tangent vector $\mathbf{x}_1 \times \mathbf{x}_2$ is attached at each vortex position.

Intuitively, one can understand this in the following way. Fix a horizontal plane going through the center of the sphere, intersecting it along a great circle. Inscribe an equilateral triangle giving us precisely the GCET configuration. Constrain each vortex to move along a great circle going through its original position and the North Pole. Then, shifting the plane vertically up and down and keeping track on its cross-section with the sphere, defines a family of equilateral triangles. Obviously, the vector of infinitesimal translation at the GCET configuration is given by w_{GCET} above, i.e., at each vortex the vector points strictly vertically.

The gradient $\nabla \Phi_\mu$ evaluated on w_{GCET} at GCET is zero; the volume function V, being the mixed vector product, vanishes at the great circle. This means that such a direction, i.e., the equilateral triangle family of equilibria, is tangential to the leaf at this point. In this sense the GCET is a nonisolated equilibrium within its symplectic leaf. Thus, further analysis of the stability of the GCET equilibrium requires applications of some other, nonstandard techniques.

The preceding considerations are not applicable to a nongreat circle equilateral triangle configuration, for which one shows that in the coordinates given by chord distances l_{nm}, the family of equilateral triangles given by $l_{nm} = l$ for all n, m intersects symplectic leaves, which are planes [see equation (10), Φ_μ is linear], *transversally*.

G. The degenerate case $\mathbf{J}(\mathbf{x}_e) = \mathbf{0}$

Stability in this case is a simple task and can be done by a dimension count. This results in the following theorem.

Theorem III.4 (stability of great circle equilibria with $\mathbf{J} = 0$): *A relative equilibrium with zero vorticity momentum* $\mathbf{J}(\mathbf{x}_e) = 0$, *which necessarily lies on a great circle, is* stable modulo SO(3).

Proof: The isotropy subgroup $\mathrm{SO}(3)_{\mu=0}$ is, in this case, the whole group SO(3) and hence the dimension of $\mathbf{J}^{-1}(0)/\mathrm{SO}(3)_{\mu=0}$ is zero. This implies that

$$\mathrm{Ker} D\mathbf{J}(\mathbf{x}) = T_{\mathbf{x}}(\mathrm{SO}(3)_{\mu=0} \cdot \mathbf{x}).$$

The assumptions of Patrick's theorem are satisfied as SO(3) is compact, and so this proves the theorem. □

H. Stability in the reduced space

One can also study the stability of equidistant configurations of fixed equilibria in the reduced space by analyzing level sets of the integrals of motion. In general, each such integral defines a codimension 1 surface, and trajectories are confined to lie in the intersection of these surfaces. In our case, the flow lines are given by intersecting the 2d energy levels $h = \text{const}$ with the coadjoint orbits which are planes. This is analogous to the rigid body flow on the angular momentum spheres, where the orbits are given by the intersection of the energy ellipsoids $h = \text{const}$ with the coadjoint orbits that are two-spheres (see, e.g., Marsden and Ratiu[18]). Similar to the Energy–Casimir method, this approach, while defining stability conditions, does not specify the transformations in the unreduced space modulo which the stability is understood.

The equidistant fixed equilibria in T are determined by the central line $a_1 = a_2 = a_3 = a$. In the neighborhood of such an equilibrium $a_i = a(1 + \epsilon_i)$, where ϵ_i are small, and the energy levels are given by

$$h = \frac{1}{4\pi R^2}\left(\ln a \sum_{n<m} \Gamma_n \Gamma_m + \Gamma_1 \Gamma_2 \epsilon_3 + \Gamma_1 \Gamma_3 \epsilon_2 + \Gamma_2 \Gamma_3 \epsilon_1 \right)$$

$$- \frac{1}{4\pi R^2} \frac{1}{2} (\Gamma_1 \Gamma_2 \epsilon_3^2 + \Gamma_1 \Gamma_3 \epsilon_2^2 + \Gamma_2 \Gamma_3 \epsilon_1^2) + \ldots . \tag{36}$$

The symplectic leaves are planes; up to a constant they are given by a linear part in (36). Thus, in a small enough neighborhood of an equilibrium, trajectories are determined by the intersections of these planes with the surfaces defined by the quadratic part in (36). Depending on the mutual signs of Γ's these surfaces are either ellipsoids or hyperboloids of one sheet or hyperboloids of two sheets. For instance, if all Γ_n have the same sign, then the quadratic surface is an ellipsoid, and its intersection with any plane is an ellipse. Hence, the fixed point is surrounded by closed planar orbits and is therefore stable. Note that the condition (32) is satisfied. On the contrary, if the signs of Γ_n are different, the quadratic surface is an hyperboloid, and its intersections with a plane are either ellipses or hyperbolas, depending on the position of the plane. This results in either stable or unstable fixed point, respectively, and is determined precisely by the conditions (32) and (33).

IV. CONCLUSIONS

A simple physical system of 3 point vortices on a sphere reveals a surprisingly rich geometrical structure. In this paper we have explicitly constructed the quotient manifold $T=P/\mathrm{SO}(3)$ of the problem and calculated its inherited Poisson bracket.

An analysis of the symplectic structure of the symplectic leaves in this quotient manifold sheds light on the classification of relative equilibria and their stability. By applying the energy–momentum method, we have found explicit criteria for the stability of different configurations of relative equilibria with generic and nongeneric momenta. In each case we have specified a group of transformations modulo which stability in the unreduced space is understood.

In work in progress, we shall explore the link with dual pairs (see Marsden and Weinstein[13] and Weinstein[23]) more thoroughly. Indeed, $P/\mathrm{SO}(3)\overset{\pi}{\leftarrow}P\overset{J}{\rightarrow}\mathscr{R}$ is a full dual pair. This duality is also one way of viewing noncommutative complete integrability of the 3-vortex problem on a sphere.

We also will be exploring the geometric phase (in the sense of Marsden, Montgomery and Ratiu[24]) for the three-vortex problem on a sphere.

ACKNOWLEDGMENTS

We would like to thank Paul Newton for helpful discussions and for insightful remarks on vortex dynamics. We also thank Anthony Blaom, Serge Preston and Tudor Ratiu for their helpful comments and advice on this and related work.

[1] H. Helmholtz, ''On integrals of the hydrodynamical equations which express vortex-motion (trans. p.g. Tait),'' Philos. Mag. **33**(4) (1867).
[2] G. Kirchhoff, *Vorlesungen Über Mathematische Physik*, 3rd ed. (Teubner, Leipzig, 1883).
[3] H. Aref, N. Rott, and H. Thomann, ''Gröbli's solution of the three-vortex problem,'' Annu. Rev. Fluid Mech. **24** (1992).
[4] R. Kidambi and P. Newton, ''Motion of three point vortices on a sphere,'' Physica D **116**, 143–175 (1998).
[5] Sir W. (Lord Kelvin) Thomson, ''On vortex atoms,'' Proc. R. Soc. Edinburgh **6**, 94–105 (1867).
[6] Sir W. (Lord Kelvin) Thomson, ''Floating magnets (illustrating vortex systems),'' Nature (London) **XVIII**, 13–14 (1878).
[7] E. A. Novikov, ''Dynamics and statistics of a system of vortices,'' Sov. Phys. JETP **41**, 937–943 (1975).
[8] J. L. Synge, ''On the motion of three vortices,'' Can. J. Math. **1**, 257–270 (1949).
[9] H. N. Aref, ''Motion of three vortices,'' Phys. Fluids **22**, 393–400 (1979).
[10] V. A. Bogomolov, ''Dynamics of vorticity at a sphere,'' Fluid Dyn. (USSR) **6**, 863–870 (1977).
[11] V. A. Bogomolov, ''Two dimensional fluid dynamics on a sphere,'' Izv. Atmos. Ocean. Phys. **15**, 18–22 (1979).
[12] F. Kirwan, ''The topology of reduced phase spaces of the motion of vortices on a sphere,'' Physica D **30**, 99–123 (1988).
[13] J. Marsden and A. Weinstein, ''Coadjoint orbits, vortices and Clebsch variables for incompressible fluids,'' Physica D **7**, 305–323 (1983).
[14] S. Smale, ''Topology and mechanics,'' Invent. Math. **10**, 305–331 (1970).
[15] V. I. Arnold, and B. A. Khesin, *Topological Methods in Hydrodynamics* (Springer-Verlag, Berlin, 1997).
[16] J. E. Marsden, *Lectures on Mechanics* (Cambridge University Press, 1992).
[17] D. K. Lewis and T. S. Ratiu, ''Rotating n-gon/kn-gon vortex configurations,'' J. Nonlinear Sci. **6**, 385–414 (1996).
[18] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Vol. 17 of *TAM* (Springer-Verlag, Berlin, 1994).
[19] M. Kummer, ''On resonant classical Hamiltonians with *n* frequencies,'' J. Diff. Eqns. **83**, 220–243 (1990).
[20] V. Kirk, J. E. Marsden, and M. Silber, ''Branches of stable three-tori using Hamiltonian methods in Hopf bifurcation on a rhombic lattice,'' Dynam. Stability Systems **11**, 267–302 (1996).
[21] J. C. Simo, D. Lewis, and J. E. Marsden, ''Stability of relative equilibria. part i: the reduced energy–momentum method,'' Arch. Ration. Mech. Anal. **115**, 15–59 (1991).

[22] G. Patrick, ''Relative equilibria in Hamiltonian systems: The dynamic interpretation of nonlinear stability on a reduced phase space,'' J. Geom. Phys. **9**, 111–119 (1992).
[23] A. Weinstein, ''The local structure of Poisson manifolds,'' J. Diff. Geom. **18**, 523–557 (1983).
[24] J. E. Marsden, R. Montgomery, and T. S. Ratiu, ''Reduction, symmetry, and phases in mechanics,'' Mem. AMS **436** (1990).

Symmetries, separability and volume forms

W. Sarlet,[a] R. Sahadevan,[b] and F. Cantrijn[c]
Department of Mathematical Physics and Astronomy, University of Ghent, Krijgslaan 281, B-9000 Ghent, Belgium

(Received 13 April 1998; accepted for publication 20 July 1998)

In order to clarify certain misconceptions in the literature, we discuss the details of the way determining equations for general symmetries of a Lagrangian system differ from determining equations for Noether symmetries, and establish the minimal set of extra conditions which have to be imposed in practical situations for the former to reduce to the latter. We further derive properties by which, in situations where the system is integrable through quadratic integrals in involution, the components of the corresponding Noether symmetries themselves can be used to compute the separation variables, if they exist, for the Hamilton–Jacobi equation. © *1998 American Institute of Physics.* [S0022-2488(98)02711-X]

I. INTRODUCTION

The role and importance of Noether's theorem in the study of Lagrangian systems is widely recognized. It is the vehicle by which, from a physical perspective, the duality is established between symmetries and conservation laws. The mathematical interest of knowing symmetries and/or first integrals of a dynamical system is that they provide means for a reduction of the integration process; knowledge of a sufficient number of such elements may eventually lead to a form of complete integrability of the system.

Many authors have discussed ways of generating, if possible, first integrals also from general symmetries of the Lagrangian (or Hamiltonian) differential equations, i.e., symmetries of non-Noether type. The general mechanisms behind all such constructions are fairly easy to grasp if one makes use of some elementary concepts from the differential geometric description of (Lagrangian) mechanics. Indeed, leaving aside certain slight variations or generalizations, one can roughly capture the full picture in the following scheme. Noether symmetries of a Lagrangian system are symmetries of the Cartan two-form $d\theta$ and the corresponding integral originates from the closedness of the contraction of $d\theta$ with the symmetry generator (see, e.g., Ref. 1). In the case of non-Noether symmetries, there is usually one of two mechanisms at work: if a corresponding first integral has a divergence part, then we are essentially looking at the invariance of the volume form $(d\theta)^n$, where n is the number of degrees of freedom (see, e.g., Ref. 2); alternatively, the invariant object is a type (1,1) tensor field, constructed from $d\theta$ and its Lie derivative under the action of the symmetry generator, and the first integrals then essentially come from the Lax-type equation which expresses this invariance.[3–5]

Unfortunately, there is still a lot of confusion about such issues in the current literature, probably because a number of authors are still not aware or convinced of the benefits and the generality of coordinate free descriptions. To cite one (of many possible) example where this leads to unfruitful efforts Lutzky recently discussed ''generalizations'' of his earlier work on non-Noether symmetries,[6,7] in which a class of velocity-dependent symmetry generators is considered. His new results,[8,9] however, are merely particular cases of the above cited construction of an invariant type (1,1) tensor field, which is valid for all velocity-dependent symmetries.

The need for correcting some inaccurate or misleading statements lies partly at the origin of the present paper. In some of the publications of Lakshmanan and co-workers about integrability aspects of mechanical systems,[10,11] consistent use of Noether's theorem was made to construct the

[a]Electronic mail: willy.sarlet@rug.ac.be
[b]On leave from The Ramanujan Institute for Advanced Study in Mathematics, University of Madras, Chepauk, Chennai–600005, Tamilnadu, India.
[c]Senior Research Associate at the Fund for Scientific Research, Flanders, Belgium.

 © 1998 American Institute of Physics

invariants. In other papers,[12–14] however, the emphasis is much more on a methodology which starts from a systematic construction of (generalized) symmetries of the differential equations (except that in the review paper,[14] a large part deals with Painlevé analysis as well). The adjective generalized refers to the fact that the spirit is much in the line of the Lie approach to differential equations, but nonpoint symmetries are being considered. From the point of view of symmetries of the vector field generating second-order dynamics, however, there is no need to talk about anything generalized here. This is of course not the point of criticism. The point is that the authors believe that they are making use of a wider class of symmetries than just Noether symmetries, as is illustrated by the explicit statement[12,13] that a first integral is being constructed without recourse to Noether's theorem. Yet, the calculations are manifestly not about some volume form or some type (1,1) tensor field. Instead, a certain *ad hoc* ansatz is imposed on the construction of (generalized) Lie symmetries, which apparently has the effect that an invariant of the symmetry generator itself becomes a first integral of the given differential equations. Another, potentially interesting, but in the presentation of the authors rather peculiar story, is about the way separation variables (in the sense of Hamilton-Jacobi theory) can be constructed directly from the components of the symmetry generator. The claim is that such variables follow from a process of ''integration of part of the characteristic equations related to the symmetry vector field.'' However, no justification whatsoever is given concerning reasons why the manipulations which are being carried out should have anything to do with the construction of separation variables.

In Sec. II we will show first of all that the extra anzats which is assumed in Refs. 12 and 13 to find symmetries which lead to a corresponding first integral, is exactly a condition which forces these symmetries to be of Noether type. In Sec. III we will take the opportunity to add new insights into the way Noether symmetries, from a computational point of view, make up a subset of the set of all dynamical symmetries of a class of autonomous Lagrangian systems. Roughly speaking, for symmetries whose leading components are polynomial functions of the velocities of a certain degree, we will prove that the restriction to be a Noether symmetry is equivalent to the requirement that: (i) the coefficients of the terms of nonzero degree in the velocities are totally symmetric; (ii) the symmetry generator preserves energy (a necessary condition for all Noether symmetries of autonomous Lagrangian systems); (iii) the lowest order part (only) is imposed of the determining equations for arbitrary symmetry vector fields.

In Sec. IV we investigate the way the Noether symmetries, corresponding to a set of n quadratic integrals in involution, might be directly related to the construction of separation variables in the sense of Hamilton-Jacobi. For a start, the calculations presented in Refs. 12–14 appear to have little to do with ''integration of characteristic equations'' related to the first-order partial differential operator which comes from the symmetry vector field. Instead, they are merely a set of formal manipulations which, at least in the case of two degrees of freedom[12] and for the examples under consideration, seem to produce the right answers. We will provide an explanation for this feature, which is based on Eisenhart's theorem about the intrinsic characterization of Stäckel systems (see, e.g., Ref. 15), and show that the manipulations under consideration can be given a theoretical backing from a factorization property of the action of the natural volume form on the vector fields corresponding to all quadratic integrals in involution. From this point of view, it would seem that for an extension to more than two degrees of freedom, a similar manipulation of characteristic equations, as attempted in Refs. 13 and 14, has no meaning whatsoever; instead, one is led to compute a determinant from the components of the symmetries and look at a factorization property again. Sec. V contains illustrative examples.

II. COMPUTATIONAL ASPECTS OF NOETHER'S THEOREM

For convenience, we will stick to the framework of autonomous second-order differential equations (as is the case in the papers we wish to clarify).

Let $L(x^i,\dot{x}^i)$, $i=1,\ldots,n$ be the regular Lagrangian of a given second-order dynamical system $\ddot{x}^i=f^i(x,\dot{x})$, with corresponding vector field

$$\Gamma=\dot{x}^i\frac{\partial}{\partial x^i}+f^i(x,\dot{x})\frac{\partial}{\partial \dot{x}^i}, \tag{1}$$

living on the tangent bundle TM of some manifold M. With $\Delta=\dot{x}^i\partial/\partial\dot{x}^i$ representing the Liouville vector field, $E=\Delta(L)-L$ the ''energy function'' associated to L, and $\theta=(\partial L/\partial\dot{x}^i)\,dx^i$ the Poincaré-Cartan one-form, we know that Γ is defined by

$$i_\Gamma d\theta=-dE. \tag{2}$$

If, for simplicity, we restrict ourselves to symmetry vector fields which are time-independent as well, Noether's theorem says: let $Y\in\mathscr{X}(TM)$ be a vector field with the properties

$$\mathscr{L}_Y\theta=df \quad \text{and} \quad Y(E)=0, \tag{3}$$

for some function f on TM, then $F=f-\langle Y,\theta\rangle$ is a first integral, i.e., $\Gamma(F)=0$. The function f is usually referred to as the gauge term in Noether's theorem. Every such Noether symmetry Y is effectively a symmetry of the given differential equations, i.e., we have $[\Gamma,Y]=0$. Recall further that the first condition can equivalently be written as

$$i_Y d\theta=dF, \tag{4}$$

from which in fact it trivially follows that $\Gamma(F)=0$ in view of (2) and $Y(E)=0$. Conversely, if F is a first integral, the relation (4) uniquely defines a vector field Y which will automatically leave the energy function invariant. This way, we obtain a one to one correspondence between Noether symmetries and first integrals (which depends on the Lagrangian). Note in passing that (4) trivially implies that $Y(F)=0$, but needless to say, if for a given symmetry vector field Y of Γ one searches for invariants of Y, there is no reason why these would at the same time be first integrals of the given dynamics as well.

In coordinates, the symmetry generator Y will be of the form

$$Y=\xi^i\frac{\partial}{\partial x^i}+\Gamma(\xi^i)\frac{\partial}{\partial\dot{x}^i}, \tag{5}$$

where the leading components ξ^i are of course functions on TM, i.e., we are talking in general about more than point symmetries.

Let us first link this up with the original version of Noether's theorem, which has to do with invariance of the action integral (up to a gauge term). In that approach, again restricting ourselves to transformations which do not change the time variable, one arrives at the following necessary and sufficient conditions for a Noether symmetry:

$$\xi^i\frac{\partial L}{\partial x^i}+\dot{\xi}^i\frac{\partial L}{\partial\dot{x}^i}=\dot{f}. \tag{6}$$

This is often referred to as the Noether identity. A good reference in this respect is the book by Logan.[16] The right way to look at the conditions (6), from the point of view of writing down determining equations for the construction of Noether symmetries, is that (6) has to be identically satisfied for all values of the independent variables $x^i,\dot{x}^i,\ddot{x}^i$. The terms $\dot{\xi}^i$ and $\dot{f}$ stand for total time derivatives of functions of $x^i,\dot{x}^i$ and hence depend linearly on $\ddot{x}^i$. As a result, Eq. (6) immediately splits and is equivalent with the $n+1$ requirements:

$$\frac{\partial\xi^i}{\partial\dot{x}^j}\frac{\partial L}{\partial\dot{x}^i}=\frac{\partial f}{\partial\dot{x}^j}, \quad j=1,\ldots,n, \tag{7}$$

$$\xi^k\frac{\partial L}{\partial x^k}+\dot{x}^k\frac{\partial\xi^i}{\partial x^k}\frac{\partial L}{\partial\dot{x}^i}=\dot{x}^k\frac{\partial f}{\partial x^k}. \tag{8}$$

The full equivalence between these partial differential equations for the unknown ξ^i, f and the ones which come out of the requirements (3) has been established (also more generally for

time-dependent situations) in Ref. 1. It may be helpful, however, to make a few more observations around this point here. Multiplying Eqs. (7) by f^j, summing over j and adding the result to (8), we see right away that the system [(7) and (8)] is equivalent with

$$\frac{\partial \xi^i}{\partial \dot{x}^j}\frac{\partial L}{\partial \dot{x}^i}=\frac{\partial f}{\partial \dot{x}^j}, \quad j=1,\ldots,n, \tag{9}$$

$$Y(L)=\Gamma(f\ ), \tag{10}$$

with Y as defined in (5). Recalling the identity

$$\mathscr{L}_\Gamma i_Y\theta=i_Y\mathscr{L}_\Gamma\theta+i_{[\Gamma,Y]}\theta$$

(which is valid for arbitrary vector fields Y,Γ and arbitrary θ), we observe that for a vector field of type (5), $[\Gamma,Y]$ is vertical, so that its contraction with the semibasic one-form θ is zero. We thus obtain,

$$\mathscr{L}_\Gamma i_Y\theta=i_Y(i_\Gamma d\theta+di_\Gamma\theta)=i_Y(-dE+d(\Delta L))=Y(L).$$

Putting $F=f-i_Y\theta$, we conclude that (10) is further equivalent to $\Gamma(F)=0$. The point about this remark is the following. Part of the folklore in the literature is that for the explicit calculations involved in the construction of first integrals of Lagrangian systems, one can either follow Darboux' direct approach,[17] i.e., try to solve directly the partial differential equation $\Gamma(F)=0$, or use Noether's method to construct symmetries first, and that these are different methods from the practical point of view. Of course, Noether's theorem expresses that one will lead to the other, but the above remark shows that in fact the calculations are, roughly speaking, the same.

A computation which is different, in principle, but more or less of the same complexity, is the search for solutions of the determining equations for arbitrary symmetries of Γ, i.e., vector fields Y for which $[\Gamma,Y]=0$. Such vector fields will be of the form (5), where the ξ^i have to solve the second-order partial differential equations

$$\Gamma^2(\xi^i)=Y(f^i). \tag{11}$$

If more than just point symmetries are hoped for, an ansatz will be needed about the polynomial dependence of the ξ^i on the velocities, to make such calculations feasible. The set of solutions will always be larger than the solutions of the corresponding search for Noether symmetries with the same ansatz. The construction of general symmetries (11) is the line of approach followed in Refs. 12 and 13, where the authors study Lagrangian systems with 2 and 3 degrees of freedom, for some specific cases of standard Lagrangians of the form

$$L=\frac{1}{2}\sum\ (\dot{x}^i)^2-V(x). \tag{12}$$

The ansatz is that the ξ^i are linear in the velocities and the claim is somehow that with all the symmetries which are obtained ''one can find, without recourse to Noether's theorem, and almost by inspection, a set of involutive first integrals by looking for invariants of the symmetry generator itself.'' Close inspection of their analysis reveals for a start, that in the process of solving Eqs. (11), the authors must have taken for granted certain simplifications, because all the symmetries they obtain will turn out to be Noether symmetries anyway. Moreover, the impression is given that first integrals should follow from ''integrating the equation $Y(F)=0$.'' The fact that the energy integral every time appears to be one of such functions F, points in the direction of Noether symmetries [see the second condition in (3)]. Finally, the rule of thumb by which the other integrals are said to appear is that we ought to expect that our symmetries are such that (in more correct notations):

$$\xi^i=\frac{\partial U}{\partial \dot{x}^i} \quad \text{and} \quad \Gamma(\xi^i)=-\frac{\partial U}{\partial x^i}, \tag{13}$$

for some function U [cf. Eqs. (2.7) and (2.9) in Ref. 12 and (2.6), (2.8), and (2.9) in Ref. 13].

Now, for a Lagrangian of type (12), we have $\theta=\dot{x}^i dx^i$ and hence, with a vector field Y of type (5)

$$i_Y(d\dot{x}^i\wedge dx^i)=\Gamma(\xi^i)dx^i-\xi^i d\dot{x}^i.$$

It follows that the requirements (13) precisely express the condition (4) with $F=-U$, so that the whole idea is bound to be just an application of Noether's theorem, as expected.

In Refs. 11 and 14 no claim is made that the construction of first integrals follows without recourse to Noether's theorem. Symmetry generators are constructed from the general requirement (11), with ξ^i which are of degree 1 or of degree 3 in the velocities. But also here, in the course of the calculations, a number of unspecified simplifications must have been imposed for the ease of finding solutions, because it turns out in the end that again all the symmetries which are found are of Noether type. In the next section, therefore, we will analyze in detail and in some generality what the minimal set of requirements is that one has to impose on general symmetries in order for them to become Noether symmetries.

III. NOETHER VERSUS GENERAL SYMMETRIES

We wish to clarify the situation about conditions which force symmetries to fall into the Noether class from the very practical point of view of actually solving determining equations. For that purpose, we will limit ourselves to purely analytical considerations and a case study for arbitrary n, which covers all the examples of symmetries mentioned in the above cited papers.

The systems under consideration are the Lagrangian ones with a Lagrangian of the form (12). It is known[18,19] that first integrals of such a system, which polynomially depend on the velocities, can be taken to have only terms which are either all of odd or all of even degree in the velocities. We consider the even degree case and assume, to be concrete, that the degree is at most four. The leading coefficients ξ^i of corresponding Noether symmetries will thus contain only terms of odd degree in the velocities and the maximal degree is three. These will then, of course, also be the restrictions we impose on the search for general symmetries. Writing all the indices of undetermined coefficients as bottom indices for easy legibility, we look for functions ξ^i of the following form:

$$\xi^i=a_{ijkl}(x)\dot{x}^j\dot{x}^k\dot{x}^l+b_{ij}(x)\dot{x}^j. \tag{14}$$

Obviously, a_{ijkl} is assumed to be symmetric in its last three indices, but no further symmetry conditions for the coefficients in (14) must hold *a priori* if the search is for general symmetries. The conditions (11) for symmetries, here more specifically of the form

$$\Gamma^2(\xi^i)+Y\left(\frac{\partial V}{\partial x^i}\right)=0, \tag{15}$$

explicitly become

$$\begin{aligned}&\frac{1}{10}a_{i(jkl,mn)}\dot{x}^j\dot{x}^k\dot{x}^l\dot{x}^m\dot{x}^n-a_{i(mkl,j)}\frac{\partial V}{\partial x^j}\,\dot{x}^k\dot{x}^l\dot{x}^m-a_{ij(kl,m)}\frac{\partial V}{\partial x^j}\,\dot{x}^k\dot{x}^l\dot{x}^m-a_{ij(kl}\frac{\partial^2 V}{\partial x^{m)}\partial x^j}\,\dot{x}^k\dot{x}^l\dot{x}^m\\&\quad+a_{jklm}\frac{\partial^2 V}{\partial x^i\partial x^j}\,\dot{x}^k\dot{x}^l\dot{x}^m+\frac{1}{3}b_{i(k,lm)}\dot{x}^k\dot{x}^l\dot{x}^m+6a_{ijkl}\frac{\partial V}{\partial x^j}\,\frac{\partial V}{\partial x^k}\,\dot{x}^l-b_{i(k,l)}\frac{\partial V}{\partial x^k}\,\dot{x}^l-b_{ik,l}\frac{\partial V}{\partial x^k}\,\dot{x}^l\\&\quad-b_{ik}\frac{\partial^2 V}{\partial x^k\partial x^l}\,\dot{x}^l+b_{kl}\frac{\partial^2 V}{\partial x^i\partial x^k}\,\dot{x}^l=0.\end{aligned} \tag{16}$$

As customary, brackets around indices refer to symmetrization. Let us, however, be precise about this to avoid confusion concerning numerical factors: the convention here is that in an expression like $a_{i(mkl,j)}$, where the object is already symmetric in k,l,m by assumption, symmetrization simply means taking a cyclic sum over all indices inside the brackets. The determining equations

for constructing all symmetries with leading components of the form (14) are obtained by putting the coefficients of the terms of degree 5, 3 and 1 in the above expression separately equal to zero.

If the symmetry is of Noether type, the ξ^i will be related to a first integral F via the relations $\xi^i = \partial F/\partial \dot{x}^i$ [strictly speaking, this is minus the F of (4)]. In other words, and in agreement with our analysis in the previous section, Y will be a Noether symmetry if and only if

$$F = \tfrac{1}{4} a_{ijkl} \dot{x}^i \dot{x}^j \dot{x}^k \dot{x}^l + \tfrac{1}{2} b_{ij} \dot{x}^i \dot{x}^j + g(x) \tag{17}$$

is a first integral of Γ for some function g. This implies for a start that we have to take the functions a_{ijkl} and b_{ij} to be symmetric in all indices. Note in passing that these symmetry properties are the integrability conditions for existence of a function f which satisfies (9) and that $\Gamma(F) = 0$, as argued before, is another way of writing (10), which again shows explicitly that we are expressing here the Noether identity (6). Now, $\Gamma(F) = 0$ also gives rise to a relation of degree 5 in the velocities and the terms of order 5, 3 and 1 produce the requirements

$$a_{(ijkl,m)} = 0, \tag{18}$$

$$b_{(ij,k)} = 6 a_{ijkl} \frac{\partial V}{\partial x^l}, \tag{19}$$

$$\frac{\partial g}{\partial x^j} = b_{jk} \frac{\partial V}{\partial x^k}. \tag{20}$$

In summary, functions of the form (14) are the leading components of a Noether symmetry if and only if the a_{ijkl} and b_{ij} are completely symmetric and satisfy (18)–(20) for some function g. The last of the requirements will then give rise to further integrability conditions which read

$$(b_{jk,l} - b_{lk,j}) \frac{\partial V}{\partial x^k} + b_{jk} \frac{\partial^2 V}{\partial x^k \partial x^l} - b_{lk} \frac{\partial^2 V}{\partial x^k \partial x^j} = 0. \tag{21}$$

Clearly, the requirements for Noether symmetries are quite different from those for arbitrary symmetries, if only because in the general case the undetermined functions in (14) need not be totally symmetric. If one wants to pin down a minimal set of extra assumptions which will force symmetries to fall into the subcategory of Noether symmetries, imposing symmetry in all indices is obviously inevitable. Since all Noether symmetries will further preserve the energy function, we wish to investigate now to what extent $Y(E) = 0$ can count for these extra assumptions.

With $E = \frac{1}{2}\Sigma(\dot{x}^i)^2 + V$, we have

$$Y(E) = \dot{x}^i \Gamma(\xi^i) + \xi^i \frac{\partial V}{\partial x^i}$$

and for general ξ^i of the form (14), this becomes

$$Y(E) = \dot{x}^i \left[\frac{1}{4} a_{i(jkl,m)} \dot{x}^j \dot{x}^k \dot{x}^l \dot{x}^m - 3 a_{ijkl} \frac{\partial V}{\partial x^j} \dot{x}^k \dot{x}^l + \frac{1}{2} b_{i(j,k)} \dot{x}^j \dot{x}^k - b_{ij} \frac{\partial V}{\partial x^j} \right]$$
$$+ a_{ijkl} \frac{\partial V}{\partial x^i} \dot{x}^j \dot{x}^k \dot{x}^l + b_{ij} \frac{\partial V}{\partial x^i} \dot{x}^j.$$

Under the assumption of fully symmetric a_{ijkl} and b_{ij}, the terms of order 1 cancel out, whereas the terms of order 5 and 3 can easily be seen to produce exactly (18) and (19). Obviously, we need one further condition to produce the remaining requirement (20) or equivalently (21). For that purpose, we look at the lowest order terms in the general symmetry requirement (16). Originally, they produce the condition:

$$6 a_{ijkl} \frac{\partial V}{\partial x^j} \frac{\partial V}{\partial x^k} - b_{i(l,k)} \frac{\partial V}{\partial x^k} - b_{ij,l} \frac{\partial V}{\partial x^j} - b_{ij} \frac{\partial^2 V}{\partial x^j \partial x^l} + b_{jl} \frac{\partial^2 V}{\partial x^i \partial x^j} = 0. \tag{22}$$

Making use of the symmetry assumption on the a and b functions and condition (19) which already followed from $Y(E)=0$, one readily verifies that (22) becomes identical to (21). The results can be summarized as follows.

Proposition: Consider a vector field Y of the form (5), where the functions $\xi^i(x,\dot{x})$ *have the structure (14), then the following three conditions imply that Y will be a Noether symmetry of the Lagrangian second-order system* Γ *derived from the function L in (12):*

(1) *the coefficients in the expression (14) for* ξ^i *are symmetric in all indices,*
(2) $Y(E)=0$ *(where E is the energy function associated to L),*
(3) *the lowest-order terms (i.e., terms of degree 1) in the nonzero components of the vector field* $[Y,\Gamma]$ *cancel out.*

Remark: Since Noether symmetries are symmetries, the above requirements will guarantee that the terms of degree 5 and 3 in the complicated expression (16) will automatically be zero as well.

Needless to say, the proposition is equally valid for the case of symmetry components ξ^i which are linear in the velocities, with corresponding quadratic first integrals: it suffices to put $a_{ijkl}=0$ in all calculations above. It would not be too difficult to extend it, on the other hand, to higher-degree polynomials and to Lagrangians with a more general metric in the kinetic energy term than the standard Euclidean metric in (12). We finally point to a few other well known facts. The condition (18) in the Noether approach, which comes from the terms of highest degree does not involve the potential and hence expresses that the terms of highest degree in (17) (quartic in our case) will constitute a first integral of the pure geodesic motion produced by the kinetic energy term in L. What the structure of Eqs. (18) further tells us is that we are looking at the components of a Killing tensor: in the present situation, the comma of course refers to taking partial derivates; if the kinetic energy metric would not be Euclidean, covariant derivatives would make their appearance there instead. To name just a few references which deal with the relation between first integrals and Killing tensors, see Refs. 20, 21, 15, and 22. As some of these references indicate, Killing tensors play a key role also in the theory about separation of variables in the Hamilton-Jacobi equation, which we discuss in the next section.

IV. SYMMETRIES AND SEPARABILITY

Separation of variables in the Hamilton-Jacobi equation is a problem with a long history. It is not our intention to even attempt a short survey of the history here. For that we refer to specialists in the field, like Benenti and Kalnins and Miller, who have contributed to the subject with numerous publications of which we cite just a few here.[23,24,15] Probably the best known classical result is Stäckel's theorem about orthogonally separable systems. It is a key result which tells you how separable systems look in the coordinates in which indeed the separation works (generalizations by Benenti cover a similar characterization for the nonorthogonal case). It does not tell you, however, how to construct such coordinates, or in fact whether such coordinates exist (see also Ref. 25). The first fundamental result about existence of separation variables is Eisenhart's theorem,[26] which is in the first place an existence theorem, but also has constructive features which are perhaps underestimated. Eisenhart's theorem essentially gives an intrinsic characterization of the existence of a point transformation which will transform a Hamiltonian system into Stäckel form. Very little is known about the existence of more general canonical transformations which have such effect. A remarkable example of a nonpoint transformation which transforms a Hamiltonian system into separable form is contained in Ref. 27. In Ref. 28, this example is put into a more general framework by the introduction of what the authors call quasipoint transformations. An equivalent result for the same example, by the way, has been obtained through techniques of algebraic geometry. See in that respect Ref. 29 and references therein. In any event, even in the case of point transformations, there is still room for new insights about tests which can tell whether a given system is separable and how one can find the right coordinates explicitly. As a preliminary remark, although Hamilton-Jacobi is of course, strictly speaking, about Hamiltonian systems, we will continue to describe the situation for the corresponding Lagrangian setting. This is certainly justified for systems with a Lagrangian of type (12), for which velocities and momenta are the same.

In Refs. 12–14 it is claimed that a direct calculation on the components of symmetry vector fields can produce separation variables, at least in the case of functions ξ^i which are linear in the velocities. In a program in which one starts the analysis of a given system by constructing symmetries, in particular Noether symmetries and the corresponding first integrals, it is certainly a good idea to investigate whether the symmetries contain by themselves information about the construction of separation variables. Roughly, what the authors do goes as follows. Consider the characteristic equations for the partial differential equation $Y(F)=0$, which read, for the case $n=2$ say,

$$\frac{dx^1}{\xi^1}=\frac{dx^2}{\xi^2}=\frac{d\dot{x}^1}{\Gamma(\xi^1)}=\frac{d\dot{x}^2}{\Gamma(\xi^2)}, \tag{23}$$

and assume that the functions ξ^i are of the form

$$\xi^i=a_{ij}(x)\dot{x}^j, \quad \text{with} \quad a_{ij}=a_{ji}. \tag{24}$$

Then, the first equality can be written in the form

$$a_{2j}\dot{x}^j dx^1-a_{1j}\dot{x}^j dx^2=0.$$

Thinking of $\dot{x}^j$ as standing for dx^j/dt and multiplying by dt, the formal quadratic expression

$$a_{2j}dx^j dx^1-a_{1j}dx^j dx^2=0 \tag{25}$$

is obtained. Solving this, for example, algebraically for dx^2/dx^1 and treating these as first-order differential equations, it so happens that their integration seems to produce the separation variables. The authors call this procedure ''integrating part of the characteristic equations.'' However, when integrating characteristic equations such as (23), the x^i and $\dot{x}^i$ ought to be treated as independent variables, so the formal manipulation we just sketched has nothing to do with that. In any event, the authors do not give any reason why this manipulation should have something to do with the construction of separation variables. The lack of a supporting theory puts the authors on even more shaky ground when they attempt similar manipulations on some three-dimensional systems because there is then more than one symmetry involved and it is in general not clear how one could, even in principle, handle a set of such formal quadratic expressions.

We will develop arguments here to show first that the formal manipulations mentioned above are indeed bound to produce separation variables for $n=2$. These arguments will show at the same time that the route to generalizations to more than two degrees of freedom is quite different from the calculations performed in Refs. 13 and 14. As a preliminary remark, observe that in manipulating the quadratic expression (25), it makes no difference whether one writes dx^i or $\dot{x}^i$, and in fact it will be more appropriate for the developments below to think of it as

$$a_{2j}\dot{x}^j\dot{x}^1-a_{1j}\dot{x}^j\dot{x}^2=0. \tag{26}$$

Since Eisenhart's theorem concerning separability is about Killing tensors, and such objects are related to first integrals which in turn correspond to Noether symmetries, it should not come as a surprise that we will seek an explanation which starts from Eisenhart. We first state the theorem more or less in the form in which it is cited in Ref. 15.

Eisenhart's theorem: *A necessary and sufficient condition that a kinetic energy Lagrangian* $L=\frac{1}{2}g_{ij}(x)\dot{x}^i\dot{x}^j$ *(or better its corresponding Hamiltonian) can be given the Stäckel form is that there exist* $n-1$ *Killing tensors* a^γ_{ij}, $\gamma=1,\ldots,n-1$, *such that:*

(1) *the tensors* $\{g_{ij},a^\gamma_{ij}\}$ *are linearly independent,*
(2) *all roots of the (generalized) eigenvalue problem* $\det(a^\gamma-\lambda^\gamma g)=0$ *are distinct for all* γ *(*a^γ *stands for the matrix with components* a^γ_{ij}*),*
(3) *there exist new coordinates* y^i *such that the vectors* $X_{(k)}$, *with components* $X^j_{(k)}=\partial x^j/\partial y^k$, *are common eigenvectors of all* a^γ, *i.e., we have*

$$(a_{ij}^{\gamma}-\lambda_{(k)}^{\gamma}g_{ij})X_{(k)}^{j}=0, \quad k=1,\ldots,n, \quad \gamma=1,\ldots,n-1. \tag{27}$$

A dual version of this theorem is mentioned by Benenti.[30] If (g^{kl}) denotes the inverse of the matrix of the given metric tensor and a_{γ}^{ij} the contravariant version of the Killing tensors obtained by raising indices in the usual way, then the contravariant version of the third requirement in the theorem says that there are n common orthogonal closed eigenforms $\alpha^{(k)}$:

$$(a_{\gamma}^{ij}-\lambda_{\gamma}^{(k)}g^{ij})\alpha_{j}^{(k)}=0, \quad \text{with} \quad d\alpha^{(k)}=0. \tag{28}$$

Locally then, $\alpha^{(k)}=dy^{k}$ for some functions y^{k}, which define, as in the other version, separation coordinates. It is worth mentioning in this context that Benenti now even has a theorem about the characterization of separable systems in which only one Killing tensor is needed,[23] but this result will not be directly relevant to our further discussion.

Naturally, for the standard metric we are considering in Lagrangians of type (12), the covariant and contravariant versions of the Killing tensors are the same, so if we construct eigenvectors in the algebraic sense, we have the freedom of interpreting them either as vector fields or as one-forms. Strictly speaking, Eisenhart's theorem is about geodesic motions or, in other words, about Lagrangians without potential term. However, it is well known since Levi-Civita that for a Hamiltonian of type $T+V$ to be separable, it is necessary that the kinetic energy part by itself is separable. Moreover, if the full Lagrangian or Hamiltonian has a quadratic first integral, which is the sum of a quadratic part and terms of degree zero, it is the quadratic part only which defines the corresponding symmetry as in (24) and, as said before, this quadratic part determines a Killing tensor for the geodesic motion coming from the kinetic energy term. As a result, for the sake of identifying separation variables and trying to relate them to symmetries, we can, without loss of generality, omit the potential term all together. There are simple criteria then (see again Ref. 23, for example, or even the original theorem of Stäckel) to check, once separation variables have been detected, whether the potential term has a structure which is compatible with the separation.

We now start examining what the implications are of the conditions on the a_{ij}^{γ} in Eisenhart's theorem for the symmetry components ξ^{i} which correspond to these via (24). It is clear that a symmetry vector field Y as in (5) is completely determined by its first term

$$X=\xi^{i}\frac{\partial}{\partial x^{i}}, \tag{29}$$

which looks like a vector field on the base manifold M, whose components, however, are functions on TM. This is called a vector field along the projection $\tau:TM\to M$. So-called semibasic forms on TM likewise can be regarded as differential forms along τ. For an adapted calculus about forms and fields along τ and an application to separability of second-order equations, we refer to Refs. 31–33. We will, however, not really need that sophisticated machinery here. Recall though, that there exists a canonically defined vector field along τ, namely the ''total time derivative operator''

$$\mathbf{T}=\dot{x}^{i}\frac{\partial}{\partial x^{i}} \tag{30}$$

which, in the present context of Lagrangians of type (12), is in fact precisely the vector field along τ determining the Noether symmetry corresponding to the energy integral.

For simplicity, we first discuss the case $n=2$ in some detail, for which there is just one Killing tensor a_{ij} with distinct eigenvalues. It is perhaps instructive to show via Eisenhart's theorem why in such a case separability always works. Using the covariant version, for example, we have two one-dimensional, integrable eigendistributions and their sum is also integrable by dimension. It follows from the Frobenius theorem that there exist coordinates y^{i} such that $\partial/\partial y^{1}$ spans the first distribution and $\partial/\partial y^{2}$ the second. This in turn means that in each of the original eigendistributions, one can find an element such that the two selected eigenvectors commute and hence can simultaneously be straightened out. These elements are of course the eigenvectors with components of the form $\partial x^{j}/\partial y^{k}$ featured in Eisenhart's theorem. Likewise, referring to the contravariant version, one-forms in dimension 2 always have an integrating factor, so that we can always make

the $\alpha^{(k)}$ locally exact, yielding again the desired separation variables. Note that these ideas can in fact be used constructively to find the separation variables for two degrees of freedom systems with a known second quadratic integral (see next section).

Now, if $\alpha^{(k)}=\alpha_i^{(k)}dx^i$, $k=1,2$ are the eigenforms and we consider their formal product to produce a quadratic expression in the dx^i, we know that this expression will become (up to a factor) the formal product dy^1dy^2 when expressed in the right coordinates. Hence if there is any truth in the expectation that the manipulations described about the quadratic expression (25) should indeed lead to the identification of separation variables y^i, there should be a link between (25) and the formal product of the $\alpha^{(k)}$ (to be interpreted properly, for example, as the symmetric tensor product). This is what we shall establish now, taking account of the earlier remark that one can write just as well a quadratic expression in $\dot{x}^i$ instead of dx^i, meaning that we then think of

$$\alpha_i^{(1)}\alpha_j^{(2)}\dot{x}^i\dot{x}^j. \tag{31}$$

Expressing that the $\alpha^{(k)}$ are eigenforms of the matrix a_{ij}, we have

$$a_{ij}\alpha_j^{(k)}=\lambda^{(k)}\alpha_i^{(k)}, \quad k=1,2, \quad i=1,2,$$

or equivalently, since these are functions of the coordinates only,

$$\dot{x}^i a_{ij}\alpha_j^{(k)}=\lambda^{(k)}\alpha_i^{(k)}\dot{x}^i, \quad k=1,2. \tag{32}$$

In the left-hand side, we recognize the contraction of the basic one-form $\alpha^{(k)}$, regarded as one-form along τ, with the vector field X along τ, determined by (29) and (24). The right-hand side contains the contraction of $\alpha^{(k)}$ with $\mathbf{T}$. In other words, the equality (32) has the intrinsic interpretation:

$$\langle X,\alpha^{(k)}\rangle=\lambda^{(k)}\langle \mathbf{T},\alpha^{(k)}\rangle, \quad k=1,2, \tag{33}$$

from which it follows that

$$(\alpha^{(1)}\wedge\alpha^{(2)})(\mathbf{T},X)=(\lambda^{(2)}-\lambda^{(1)})\langle \mathbf{T},\alpha^{(1)}\rangle\langle \mathbf{T},\alpha^{(2)}\rangle. \tag{34}$$

The final point to observe now is that $\alpha^{(1)}\wedge\alpha^{(2)}$, being a volume form (on an open set), is proportional to the standard volume form $dx^1\wedge dx^2$, and

$$(dx^1\wedge dx^2)(\mathbf{T},X)=\xi^2\dot{x}^1-\xi^1\dot{x}^2=a_{2j}\dot{x}^j\dot{x}^1-a_{1j}\dot{x}^j\dot{x}^2. \tag{35}$$

Collecting the results we thus conclude: (35) shows that the quadratic expression (26) (which was obtained through manipulation of characteristic equations in Ref. 12) can be interpreted as coming from the action of a volume form on the symmetry generators $(\mathbf{T},X)$; as such, it is bound to be proportional to the right-hand side of (34) when one thinks of the eigenforms of the matrix (a_{ij}) to construct a volume form; since the eigenvalues $\lambda^{(k)}$ are further assumed to be different in Eisenhart's theorem, this right-hand side is in turn proportional to the expression (31); finally, Eisenhart's theorem guarantees that separation coordinates exist, in which (31) will simply factorize as $\dot{y}^1\dot{y}^2$ (always up to a multiplicative function).

From a practical point of view, once a second quadratic integral has been found (or its corresponding linear Noether symmetry) and the eigenvalues of the matrix (a_{ij}) are different, the above result somehow provides an indirect way of making use, constructively, of Eisenhart's theorem: without bothering about computing eigenforms, we compute a determinant, namely the function $(dx^1\wedge dx^2)(\mathbf{T},X)$ which gives rise to a quadratic form in the $\dot{x}^i$ and one can try to find separation variables by factorizing this expression into the product of two linear expressions. For two degrees of freedom, in fact, this idea can be made stronger because there is a converse result: any such factorization, with factors which are linearly independent, will give rise to separation variables. Indeed, suppose we can write the determinant in question as follows

$$a_{2j}\dot{x}^j\dot{x}^1-a_{1j}\dot{x}^j\dot{x}^2=(\alpha_i^{(1)}\dot{x}^i)(\alpha_j^{(2)}\dot{x}^j).$$

Identification of the coefficients of corresponding terms in both sides then gives:

$$a_{21}=\alpha_1^{(1)}\alpha_1^{(2)}, \quad a_{22}-a_{11}=\alpha_1^{(1)}\alpha_2^{(2)}+\alpha_2^{(1)}\alpha_1^{(2)}, \quad a_{12}=-\alpha_2^{(1)}\alpha_2^{(2)}.$$

Remembering of course that $a_{12}=a_{21}$ it follows that

$$a_{11}\alpha_1^{(1)}+a_{12}\alpha_2^{(1)}=(a_{11}+\alpha_1^{(2)}\alpha_2^{(1)})\alpha_1^{(1)},$$

$$a_{21}\alpha_1^{(1)}+a_{22}\alpha_2^{(1)}=(a_{22}-\alpha_2^{(2)}\alpha_1^{(1)})\alpha_2^{(1)},$$

$$a_{11}\alpha_1^{(2)}+a_{12}\alpha_2^{(2)}=(a_{11}+\alpha_1^{(1)}\alpha_2^{(2)})\alpha_1^{(2)},$$

$$a_{21}\alpha_1^{(2)}+a_{22}\alpha_2^{(2)}=(a_{22}-\alpha_2^{(1)}\alpha_1^{(2)})\alpha_2^{(2)}.$$

This proves that, with

$$\lambda^{(1)}=a_{11}+\alpha_1^{(2)}\alpha_2^{(1)}=a_{22}-\alpha_2^{(2)}\alpha_1^{(1)},$$

$$\lambda^{(2)}=a_{11}+\alpha_1^{(1)}\alpha_2^{(2)}=a_{22}-\alpha_2^{(1)}\alpha_1^{(2)},$$

the $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$ give the components of an eigenform, and we have $\lambda^{(1)}\neq\lambda^{(2)}$ in view of the assumed linear independence. The two eigenforms have an integrating factor by dimension and the resulting new variables will be separation variables by Eisenhart's theorem.

Let us now move on to the case $n=3$. Clearly, if we follow the same line of thought the generalization will not lead to a set of quadratic expressions but rather to one cubic expression, obtained from the computation of a determinant again and having a factorization property as before. A sketch of the reasoning goes as follows. Assume we have obtained two extra quadratic integrals in involution or, equivalently, their corresponding Noether symmetries which are determined by vector fields along τ, say

$$X_\gamma=\xi_\gamma^k\frac{\partial}{\partial x^k}, \quad \text{with} \quad \xi_\gamma^k=a_{\gamma kl}\dot{x}^l, \quad \gamma=1,2. \tag{36}$$

In case of separability, following Eisenhart, there must be closed one-forms $\alpha^{(i)}$, $i=1,2,3$, which are common eigenforms of both matrices a_γ. Denoting the eigenvalues of a_γ, which are assumed to be different, by $\lambda_\gamma^{(k)}$, we thus have, in the same way as (32) was derived and rewritten in the form (33):

$$\langle X_\gamma,\alpha^{(k)}\rangle=\lambda_\gamma^{(k)}\langle\mathbf{T},\alpha^{(k)}\rangle, \quad k=1,2,3, \quad \gamma=1,2. \tag{37}$$

It follows that

$$(\alpha^{(1)}\wedge\alpha^{(2)}\wedge\alpha^{(3)})(\mathbf{T},X_1,X_2)=\rho\langle\mathbf{T},\alpha^{(1)}\rangle\langle\mathbf{T},\alpha^{(2)}\rangle\langle\mathbf{T},\alpha^{(3)}\rangle, \tag{38}$$

with

$$\rho=\lambda_1^{(1)}(\lambda_2^{(2)}-\lambda_2^{(3)})+\lambda_1^{(2)}(\lambda_2^{(3)}-\lambda_2^{(1)})+\lambda_1^{(3)}(\lambda_2^{(1)}-\lambda_2^{(2)}). \tag{39}$$

The function ρ is nonzero because of the first assumption in Eisenhart's theorem. The left-hand side of (38), coming from a volume form, is proportional to the determinant $(dx^1\wedge dx^2\wedge dx^3)\times(\mathbf{T},X_1,X_2)$ which can be computed directly from the symmetry components. The right-hand side of (38) then expresses that this cubic expression in the $\dot{x}^i$ can be factorized as the product of three linear expressions and even in such a way that this product becomes simply $\dot{y}^1\dot{y}^2\dot{y}^3$ in a set of separation variables. For $n>2$ there is no converse in the sense that the existence of a factorization of the determinant by itself does not necessarily imply that each of the factors will be integrable. The results of our analysis can be summarized as follows.

Proposition: Assume that X_1 and X_2 of the form (36) define Noether symmetries of a Lagrangian system (12) with three degrees of freedom. Then, if we are in a situation where corresponding separation variables for the Hamilton-Jacobi equation exist, they can be obtained from factorizing the determinant

$$\begin{vmatrix} \dot{x}^1 & \dot{x}^2 & \dot{x}^3 \\ \xi_1^1 & \xi_1^2 & \xi_1^3 \\ \xi_2^1 & \xi_2^2 & \xi_2^3 \end{vmatrix} \tag{40}$$

into the product of three factors which are linear in the velocities and integrable.

Observe that the potential energy function V is not present in this statement: we repeat that the potential is irrelevant as long as one wants to find out the nature of separation variables if they exist; having found them, one is assured of separability of the kinetic energy part and then it is relatively easy to check whether the form of the potential is such that also the full system will be separable (see next section).

We have proved above a similar result for $n=2$ and it is obvious that it can actually be done for arbitrary n. But beyond $n=3$, it is doubtful that this procedure (as perhaps any other) could be implementable in practice. We will show in the next section, however, that up to three degrees of freedom, there are indeed different ways of making constructive use of this proposition. Our point of view in this respect is that this amounts to making constructive use of the existence theorem by Eisenhart.

On the theoretical side, it looks worthwhile for a future study to go more deeply into the potential role and relevance of volume forms for the characterization of separability. A generalization of the present considerations to nonstandard metrics looks fairly straightforward. A more difficult issue is the question whether anything similar could be done related to nonquadratic integrals, where a transformation to separation variables will not be a point transformation (cf. the already cited example[27]).

Before closing this section, let us make clear why one should not expect any meaningful conclusion from manipulations of quadratic expressions such as (26) when n is greater than 2. Indeed, thinking of the case $n=3$, for example, and the matrix a_{ij} related to one of the symmetry generators, (26) can still be generated by a computation as in (35), and the related computation (34) will still produce a quadratic expression which factorizes in the new variables guaranteed by Eisenhart's theorem. The point then is, however, that $\alpha^{(1)}\wedge\alpha^{(2)}$ will in general contain terms in all $dx^i\wedge dx^j$, or alternatively, $dx^1\wedge dx^2$ will be a combination of all $\alpha^{(i)}\wedge\alpha^{(j)}$ and hence will by itself not factorize.

V. ILLUSTRATIVE EXAMPLES

An example of an integrable and separable system with two degrees of freedom which can be found, e.g., in one of the tables in Ref. 14 is the Lagrangian system (12) with potential (we will write here x,y for the coordinates instead of x^1,x^2)

$$V(x,y)=a(x^2+4y^2)+b(x^4+12x^2y^2+16y^4), \tag{41}$$

for which a second quadratic integral is of the form

$$F=(x\dot{y}-y\dot{x})\dot{x}+2(a+4by^2+2bx^2)x^2y. \tag{42}$$

For the sake of illustrating a number of points discussed in the previous section, we will first of all limit ourselves to quartic potentials ($a=0$) and in fact assume that the class of admissible quartic potentials, in the sense that there exists a first integral with $(x\dot{y}-y\dot{x})\dot{x}$ as quadratic part in the velocities, has as yet to be discovered. Let us say, for example, that we aim at finding potentials of the form

$$V=c_1x^4+c_2x^2y^2+c_3y^4, \tag{43}$$

with c_i to be determined.

When we think of quadratic integrals, the Noether requirements [(18)–(20)] reduce to

$$a_{(ij,k)}=0, \quad \frac{\partial g}{\partial x^j}=a_{jk}\frac{\partial V}{\partial x^k}. \tag{44}$$

Our starting point is the assumption that we have found from the highest-order terms the quadratic part of the above first integral F as a particular solution, or in other words, the symmetry generator (as vector field along τ):

$$X=(x\dot{y}-2y\dot{x})\frac{\partial}{\partial x}+x\dot{x}\frac{\partial}{\partial y}. \tag{45}$$

The potential V plays no role in obtaining this part, which determines a Killing tensor with components

$$A=(a_{ij})=\begin{pmatrix} -2y & x \\ x & 0 \end{pmatrix}. \tag{46}$$

Our objective is not only to find potentials for which a second quadratic integral exists, but also to construct separation coordinates. The determination of the potential can then further be suspended and separation variables, if they exist, should follow from the following procedure. According to Eisenhart's theorem, we construct a set of eigenvectors of the matrix A; they read:

$$(x,y+r) \quad \text{and} \quad (x,y-r) \quad \text{with} \quad r=(x^2+y^2)^{1/2}.$$

If we treat these as components of eigenvector fields, say

$$X_1=x\frac{\partial}{\partial x}+(y+r)\frac{\partial}{\partial y}, \quad X_2=x\frac{\partial}{\partial x}+(y-r)\frac{\partial}{\partial y}, \tag{47}$$

then it is easy to verify that this happens to be a basis of commuting vector fields for the eigendistributions, so that they can simultaneously be straightened out. A coordinate transformation which does this is and therefore defines separation variables reads:

$$u=\ln\frac{x}{\sqrt{r-y}}, \quad v=\ln\sqrt{r-y}.$$

However, any further transformation which does not couple these coordinates will preserve separability. We may, therefore, just as well take

$$u=\sqrt{r+y}, \quad v=\sqrt{r-y}. \tag{48}$$

In agreement with the theoretical discussion of the previous section, we could here also treat the eigenvectors of A as components of eigenforms, and one then readily finds integrating factors which bring these into the form du and dv, with u and v as defined by (48). Finally, the new idea which we have introduced with regard to separation variables, namely as coming from a factorization of the determinant $\dot{x}\xi^2-\dot{y}\xi^1$, can be seen to work as follows. The quadratic expression (35) here becomes:

$$x\dot{x}^2-2y\dot{x}\dot{y}-x\dot{y}^2=\frac{1}{x}(x\dot{x}+(y+r)\dot{y})(x\dot{x}+(y-r)\dot{y}).$$

Again, one can see that the potential is irrelevant in this process of detecting possible separation variables. One way of determining subsequently whether the potential has an appropriate form for separability could go as follows: having found separation variables already, one could first express the given system in these new coordinates and then verify for what values of the parameters the potential satisfies the criteria of Stäckel's theorem. Alternatively, there is an intrinsic character-

ization of the condition which the potential should satisfy, which therefore can be verified directly in the original coordinates. This characterization can be found in Ref. 23, and says that the one-form $K\cdot dV$ should be closed (for all of the Killing tensors K entering Eisenhart's theorem). For the present example, this reduces to the requirement that the functions $a_{ij}\partial V/\partial x^j$ should be the components (locally) of an exact one-form. Not surprisingly, this is exactly the remaining Noether condition in (44), or equivalently, the condition (21). For the Killing tensor A given in (46), we have

$$A\cdot dV=\left(-2y\frac{\partial V}{\partial x}+x\frac{\partial V}{\partial y}\right)dx+x\frac{\partial V}{\partial x}\,dy,$$

and expressing that $d(A\cdot dV)$ should be zero, for a V of the form (43), reduces to the condition

$$(12c_1-c_2)x^3+(8c_2-6c_3)xy^2=0.$$

It follows that we must have $c_2=12c_1$ and $c_3=16c_1$, which is in agreement with the quartic part in the potential function (41).

It is worth mentioning that another, in fact well-known, admissible potential could have been discovered along the lines of this example. Indeed, the quadratic part $(x\dot{y}-y\dot{x})\dot{x}$ of the function F in (42) is the velocity-dependent part of one of the components of the Runge-Lenz vector in the reduced (planar) Kepler problem. One easily verifies that the Kepler potential also satisfies the above requirement $d(A\cdot dV)=0$ and we further conclude that (48) defines separation variables for the Kepler problem as well. A good reference where this example is encountered in the context of linking symmetries to Killing tensors is Ref. 34.

For three degrees of freedom it is not so obvious how one could use in practice the factorization property of the determinant (40) to find separation variables. We will nevertheless illustrate two ways in which one could make use of this result for constructive purposes. We shall write x,y,z for the coordinates x^i now.

Looking once more at the rich collection of examples which is contained in Ref. 14, assume we have a Lagrangian of type (12) which has an angular momentum type first integral, the square of which provides a first quadratic integral, say

$$F_1=(y\dot{z}-z\dot{y})^2. \tag{49}$$

Assume another integral is of the form

$$F_2=(y\dot{x}-x\dot{y})\dot{y}+(z\dot{x}-x\dot{z})\dot{z}+\cdots. \tag{50}$$

As before, we leave the terms of degree zero, both in the Lagrangian and in F_2, out of the discussion for the time being. For the corresponding Noether symmetries, we have

$$\xi_1^1=0,\quad \xi_1^2=-2z(y\dot{z}-z\dot{y}),\quad \xi_1^3=2y(y\dot{z}-z\dot{y}), \tag{51}$$

$$\xi_2^1=y\dot{y}+z\dot{z},\quad \xi_2^2=y\dot{x}-2x\dot{y},\quad \xi_2^3=z\dot{x}-2x\dot{z}. \tag{52}$$

It is obvious from the second row (51) in the determinant (40) that there will be factor $y\dot{z}-z\dot{y}$. The remaining quadratic factor then is

$$(y\dot{y}+z\dot{z})^2+2x(y\dot{y}+z\dot{z})\dot{x}-(y^2+z^2)\dot{x}^2,$$

and it is not too hard to see that this factorizes as

$$(y\dot{y}+z\dot{z}+(x+r)\dot{x})(y\dot{y}+z\dot{z}+(x-r)\dot{x}),$$

with $r=(x^2+y^2+z^2)^{1/2}$. It turns out that these three factors are integrable and identify separation variables u,v,w, defined by

$$u=\sqrt{r+x},\quad v=\sqrt{r-x},\quad w=\arctan\left(\frac{y}{z}\right).$$

However, we should emphasize that our theoretical arguments in Sec. IV do not allow us to draw this conclusion at all. Indeed, it is easy to see that the matrix corresponding to F_1 has a double eigenvalue zero, so that we are not in the situation of Eisenhart's theorem from which our results were derived. The reason is that the essential first integral behind F_1 is the linear expression $y\dot{z}-z\dot{y}$, which corresponds to a Killing vector (not a Killing two-tensor). A rudimentary explanation about the reasons why the above determinant calculation happens to produce the right answers anyway, is that the Killing vectors for a separable Hamiltonian in the end correspond to ignorable coordinates (see Ref. 30). As a result, the system will immediately reduce to a lower dimensional one, and the remaining quadratic expression which we factorized above can then be expected to be in direct correspondence with the quadratic form coming from this lower dimensional system. Note in passing that this situation should not be confused with the well-known property (see, e.g., Ref. 20) that in spaces of constant curvature, every Killing two-tensor is trivial, in the sense of being expressible as a linear combination of symmetrized tensor products of Killing vectors. From the point of view of the present analysis, there is no problem with this form of ''triviality'' as long as the regularity assumptions of Eisenhart's theorem are satisfied. For now, let us again sketch how one can subsequently find appropriate potentials which preserve the separability.

Assume we want to find suitable potentials again in the class of homogeneous functions of degree 4 and, to simplify matters, more particularly of the form

$$V=c_1x^4+c_2y^4+c_3z^4+d_1x^2y^2+d_2x^2z^2+d_3y^2z^2. \tag{53}$$

The one-form $A\cdot dV$ which comes from the Killing tensor related to F_2 reads

$$(yV_y+zV_z)dx+(yV_x-2xV_y)dy+(zV_x-2xV_z)dz.$$

Expressing its closure for a V of type (53), we find restrictions on the undetermined constants which reduce to $c_1=16c_2$, $c_3=c_2$, $d_1=d_2=12c_2$, $d_3=2c_2$. This way, up to a constant factor c_2, the potential is bound to be of the form

$$V=16x^4+(y^2+z^2)^2+12x^2(y^2+z^2). \tag{54}$$

We observe then that this potential is already invariant under the action of the Killing vector $y\partial/\partial z-z\partial/\partial y$ related to F_1. It follows that F_1 will be a quadratic integral for the full system as well.

Let us finally illustrate how one can now further determine the terms in F_2 which are indicated by dots. With the potential (54), the one-form $A\cdot dV$ becomes

$$4(y^2+z^2)(6x^2+y^2+z^2)dx+16x(x^2+y^2+z^2)(ydy+zdz).$$

In agreement with the Noether condition (20), putting this expression equal to dg should produce the remaining function in the quadratic integral. We thus find

$$g=4x(y^2+z^2)^2+8x^3(y^2+z^2). \tag{55}$$

For a third illustration now, assume we start from the following two Killing tensors of the standard metric in dimension 3,

$$F_1=(x\dot{y}-y\dot{x})^2+(x\dot{z}-z\dot{x})^2+(y\dot{z}-z\dot{y})^2+(\dot{x}^2-\dot{z}^2), \tag{56}$$

$$F_2=(y\dot{z}-z\dot{y})^2-(x\dot{y}-y\dot{x})^2-(\dot{x}^2+\dot{z}^2). \tag{57}$$

Also this case is inspired by an example in Ref. 14 and fits as well into the category of systems described in Ref. 35. The idea again is that these should become the leading parts of quadratic integrals of a Lagrangian system of type (12) for some suitable potential. If we first want to verify whether the ultimate system and its integrals will correspond to a case of separability in the

Hamilton-Jacobi sense, then we could try our luck with the factorization property of the determinant (40). Needless to say, we could also try to address directly the conditions of Eisenhart's theorem. The determinant (40) for this case gives rise to a fairly complicated cubic expression in $\dot{x},\dot{y},\dot{z}$ and it is not obvious at first sight how one could think of a way to write it as a product of three linear factors. Nevertheless, we know that there are only a finite number of coordinate systems in $\mathbb{R}^3$ in which the standard metric is separable (the 11 types first derived by Eisenhart[26] and reconstructed also by Benenti[23]). With some experience in the field one can readily rule out a number of these possibilities and then a direct computation on the determinant (40) does become feasible, as we shall now illustrate. One of the more difficult types of coordinates to work with is, for example, the class (IV_1) of Eisenhart.[26] If we denote the new coordinates by u,v,w again, they are defined by relations of the following type

$$x^2=\frac{(\alpha-u)(\alpha-v)(\alpha-w)}{(\alpha-\beta)(\alpha-\gamma)}, \tag{58}$$

$$y^2=\frac{(\beta-u)(\beta-v)(\beta-w)}{(\beta-\alpha)(\beta-\gamma)}, \tag{59}$$

$$z^2=\frac{(\gamma-u)(\gamma-v)(\gamma-w)}{(\gamma-\alpha)(\gamma-\beta)}, \tag{60}$$

for some constants α,β,γ, and the domain of definition is determined by the requirement $\alpha>u>\beta>v>\gamma>w$. It is then a fairly straightforward exercise to transform the cubic polynomial in $\dot{x},\dot{y},\dot{z}$ into a corresponding cubic expression in $\dot{u},\dot{v},\dot{w}$ via the derivatives of (58)–(60). If there indeed exists a transformation of this type which will give rise to separation variables, the new cubic polynomial should reduce to just one term, namely the monomial $\dot{u}\dot{v}\dot{w}$. Doing such a calculation is not as formidable as it may at first sound: first of all, it is typically a symbolic calculation which computer algebra packages can handle; second, to see whether there is any chance for this to work, it suffices to test first just one term, i.e., to check, for example, whether the coefficient of $\dot{u}^3$ can be made to vanish. We have done this for the above situation and found that (58)–(60) does produce a factorization of the determinant (40), provided we choose α,β,γ in such a way that

$$\alpha-\beta=1, \quad \gamma-\beta=-1. \tag{61}$$

As before, one can separately compute what kind of potentials are acceptable and in that process one will find at the same time what extra functions then have to be added to the expressions (56) and (57) in order to obtain quadratic integrals for the full Lagrangian. The procedure works exactly as in the preceding case. We limit ourselves to the observation that it is not possible for this case to find an acceptable potential which is homogeneous of degree four in the coordinates. One does obtain, however, the following class of potentials (with A an arbitrary constant):

$$V=A(x^2+y^2+z^2)+(y^2+2z^2)+(x^2+y^2+z^2)^2. \tag{62}$$

ACKNOWLEDGMENTS

R.S. would like to thank the University of Ghent for a research grant which made this work possible. W.S. and F.C. further thank the Fund for Scientific Research, Flanders, Belgium, for continuing support.

[1] W. Sarlet and F. Cantrijn, ''Generalizations of Noether's theorem in classical mechanics,'' SIAM (Soc. Ind. Appl. Math.) Rev. **23**, 467–494 (1981).
[2] M. Crampin, ''Symmetries and invariants of volume preserving flows,'' Phys. Lett. **79A**, 138–140 (1980).
[3] M. Crampin, ''A note on non-Noether constants of motion,'' Phys. Lett. **95A**, 209–212 (1983).
[4] W. Sarlet and M. Crampin, ''Some recent results on symmetries of Lagrangian systems re-examined,'' J. Phys. A **18**, 2849–2855 (1985).
[5] M. Crampin, G. Marmo, and C. Rubano, ''Conditions for the complete integrability of a dynamical system admitting alternative Lagrangians,'' Phys. Lett. **97A**, 88–90 (1983).

[6] M. Lutzky, ''Origin of non-Noether invariants,'' Phys. Lett. **75A**, 8–10 (1979).
[7] M. Lutzky, ''New classes of conserved quantities associated with non-Noether symmetries,'' J. Phys. A **15**, L87–L91 (1982).
[8] M. Lutzky, ''Remarks on a recent theorem about conserved quantities,'' J. Phys. A **28**, L637–L638 (1996).
[9] M. Lutzky, ''Conserved quantities and velocity-dependent symmetries in Lagrangian mechanics,'' Int. J. Non-Linear Mech. **33**, 393–396 (1997).
[10] R. Sahadevan and M. Lakshmanan, ''Invariance and integrability: Hénon-Heiles and two coupled quartic anharmonic oscillator systems,'' J. Phys. A **19**, L949–L954 (1986).
[11] M. Lakshmanan and R. Sahadevan, ''Generalized Lie symmetries and complete integrability of certain nonlinear Hamiltonian systems with three degrees of freedom,'' J. Math. Phys. **32**, 75–83 (1991).
[12] M. Lakshmanan and M. Senthil Velan, ''Direct integration of generalized Lie symmetries of nonlinear Hamiltonian systems with two degrees of freedom: integrability and separability,'' J. Phys. A **25**, 1259–1272 (1992).
[13] M. Lakshmanan and M. Senthil Velan, ''Direct integration of generalized Lie or dynamical symmetries of three degrees of freedom nonlinear Hamiltonian systems: integrability and separability,'' J. Math. Phys. **33**, 4068–4077 (1992).
[14] M. Lakshmanan and R. Sahadevan, ''Painlevé analysis, Lie symmetries, and integrability of coupled nonlinear oscillators of polynomial type,'' Phys. Rep. **224**, 1–93 (1993).
[15] E. G. Kalnins and W. Miller, ''Killing tensors and variable separation for Hamilton-Jacobi and Helmholtz equations,'' SIAM (Soc. Ind. Appl. Math.) J. Math. Anal. **11**, 1011–1026 (1980).
[16] J. D. Logan, *Invariant Variational Principles* (Academic, New York, 1977).
[17] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge University Press, Cambridge, England, 1964).
[18] G. Thompson, ''Polynomial constants of motion in flat space,'' J. Math. Phys. **25**, 3474–3478 (1984).
[19] J. Hietarinta, ''Direct methods for the search of the second invariant,'' Phys. Rep. **147**, 87–154 (1987).
[20] G. Thompson, ''Killing tensors in spaces of constant curvature,'' J. Math. Phys. **27**, 2693–2699 (1986).
[21] S. Benenti, ''Stäckel systems and Killing tensors,'' Note de Matematica (Suppl.) **9**, 39–58 (1989).
[22] N. M. J. Woodhouse, ''Killing tensors and the separation of the Hamilton-Jacobi equation,'' Commun. Math. Phys. **44**, 39–44 (1975).
[23] S. Benenti, ''Intrinsic characterization of the variable separation in the Hamilton-Jacobi equation,'' J. Math. Phys. **38**, 6578–6602 (1997).
[24] E. G. Kalnins, *Separation of Variables for Riemannian Spaces of Constant Curvature* (Pitman, New York, 1986), Vol. 28.
[25] I. Marshall and S. Wojciechowski, ''When is a Hamiltonian system separable?'' J. Math. Phys. **29**, 1338–1346 (1988).
[26] L. P. Eisenhart, ''Separable systems of Stäckel,'' Ann. Math. **35**, 284–305 (1934).
[27] M. Blaszak and S. Wojciechowski, ''A generalized Henon-Heiles system and related integrable Newton equations,'' J. Math. Phys. **35**, 1693–1709 (1994).
[28] S. Wojciechowski and A. V. Tsiganov, ''Quasi-point separation of variables for the Henon-Heiles system and a system with a quartic potential,'' J. Phys. A **29**, 7769–7778 (1996).
[29] V. Ravoson, L. Gavrilov, and R. Caboz, ''Separability and Lax pairs for Hénon-Heiles system,'' J. Math. Phys. **34**, 2385–2393 (1993).
[30] S. Benenti, ''Separation of variables in the geodesic Hamilton-Jacobi equation,'' in *Symplectic Geometry and Mathematical Physics*, edited by P. Donato, C. Duval, J. Elhadad, and G. M. Tuynman (Progress in Mathematics 99, Birkhäuser, 1991).
[31] E. Martínez, J. F. Cariñena, and W. Sarlet, ''Derivations of differential forms along the tangent bundle projection,'' Diff. Geom. Appl. **2**, 17–43 (1992).
[32] E. Martínez, J. F. Cariñena, and W. Sarlet, ''Derivations of differential forms along the tangent bundle projection II,'' Diff. Geom. Appl. **3**, 1–29 (1993).
[33] E. Martínez, J. F. Cariñena, and W. Sarlet, ''Geometric characterization of separable second-order equations,'' Math. Proc. Camb. Philos. Soc. **113**, 205–224 (1993).
[34] M. Crampin, ''Hidden symmetries and Killing tensors,'' Rep. Math. Phys. **20**, 31–40 (1984).
[35] S. Wojciechowski, ''On a Lax-type representation and separability of the anisotropic harmonic oscillator in a radial quartic potential,'' Lett. Nuovo Cimento **41**, 361–369 (1984).

Relativistic eccentric anomalies

Kostake Teleman[a)]
Faculty of Mathematics, University of Bucharest,
Str. Academeci 14, Bucharest Ro-70109, Romania

(Received 30 March 1998; accepted for publication 17 April 1998)

We use Kepler's eccentric anomaly in order to get explicit equations for the planetary orbits in a solar system. By successively modifying the corresponding Lagrangian, we build a Lagrangian for a relativistic two-body problem, the trajectories of which are conics in Minkowski space. The equations of these trajectories, projected on a three-space, exhibit new anomalies.
© *1998 American Institute of Physics.* [S0022-2488(98)01411-X]

I. INTRODUCTION

The eccentric anomaly α was introduced by Kepler in order to obtain the relationship between the variables r and t appearing in the equations of a planetary motion. Explicitly, one gets [see Ref. 1, formulas (106) and (107) and Ref. 2, p. 83]

$$r=a(1-e\ \cos\ \alpha),\quad t-t_0=\frac{T}{2\pi}\,(\alpha-e\ \sin\ \alpha).$$

In this paper we show briefly the way one can use α in order to get the explicit equations for the planetary orbits. However, our main goal is to obtain a relativistic version of those equations. We solve this problem by successively modifying the Lagrangian of the Newtonian two-body problem, until we get a Poincaré-invariant two-particle Lagrangian, whose trajectories are conics in Minkowski space. The equations of these trajectories are very close to those corresponding to the Newtonian case. However, when we project these on three-space, some new anomalies arise.

We note that in our discussion of both classical and relativistic cases, the eccentric anomaly α and also the Bernoulli–Laplace eccentricity vector A play an essential role (see Ref. 3). This is because the use of α as a parameter on each trajectory leads to linear differential equations and $-A/2E$ gives the centers of the trajectories, when E is the relative energy. Moreover, α plays the role of a relativistic proper time.

We shall also use a modified version of Maupertuis' theorem, which we named **Maupertuis 2000**, since we suppose it is well known and since we cannot give any quotation.

The relativistic two-particle Lagrangians studied in this Note have been obtained in our endeavor to get a link between a similar Lagrangian, introduced in Ref. 4, and the relativistic theory of gravitation.

II. THE NEWTONIAN TWO-BODY PROBLEM

The Lagrangian of the two-body problem in Newtonian mechanics has the form

$$N=\frac{1}{2}\,(m_x\dot{x}^2+m_y\dot{y}^2)+Cm_xm_y\,\frac{1}{r},\quad r=[(x-y)^2]^{1/2}.$$

Hoping to find a realistic relativistic version, we shall first integrate the Euler–Lagrange equations for N, following a nonstandard but possibly already invented method. To this end, we make the notations

$$u=x-y,\quad M=m_x+m_y,\quad v=M^{-1}(m_xx+m_yy),\quad k=CM.$$

Note that v represents the *center of gravity*; we can write

$$N=\frac{M}{2}\,\dot{v}^2+\frac{m_xm_y}{2M}\left(\dot{u}^2+\frac{2k}{r}\right)$$

a)Electronic mail: kostake@math.math.unibuc.ro

 © 1998 American Institute of Physics

and the Euler–Lagrange equations take the form $\ddot{v}=0$, $\ddot{u}=-ku/r^3$.

In order to integrate the second of these equations, we introduce, along each trajectory, s as a new parameter, so that $dt/ds=t'=r$ and then

$$u'=\frac{du}{ds}=r\dot{u},\quad u''=\frac{d^2u}{ds^2}=r^2\ddot{u}+r\dot{r}\dot{u},\quad r'=r\dot{r},\quad r''=r^2\ddot{r}+r\dot{r}^2.$$

The second Euler–Lagrange equation becomes

$$ru''=r'u'-ku,$$

and the quantities

$$E=\frac{\dot{u}^2}{2}-\frac{k}{r}=\frac{u'^2}{2r^2}-\frac{k}{r},\quad L=u\times\dot{u}=\frac{u\times u'}{r}.$$

$$A=u''-2Eu=r\dot{r}\dot{u}+\left(\frac{k}{r}-\dot{u}^2\right)u,\quad L^2=2Er^2+2kr-r'^2=u'^2-r'^2$$

are conserved. Moreover, we have

$$A^2=k^2+2EL^2,\quad r''-2Er=r^2\ddot{r}+r\dot{r}^2-r\dot{u}^2+2k=ru\ddot{u}+2k=k.$$

In the nonparabolic case $E\neq 0$, the equations thus obtained,

$$\text{(I)}\quad u''-2Eu=A,\quad r''-2Er=k,$$

have the general solution

$$u=-\frac{A}{2E}+u_1,\quad r=-\frac{k}{2E}+r_1,\quad u_1''=2Eu_1,\quad r_1''=2Er.$$

If $kE<0$, there are solutions with $u^2=r^2$ which define conic sections with centers $u_0=-A/2E$, foci $u_1=0$, $u_2=-A/E$, and parameters

$$L^2=-2Eb^2,\quad a=-\frac{k\varepsilon}{2E},\quad \varepsilon=\pm 1,$$

$$c=\sqrt{a^2-b^2}=-\frac{|A|}{2E}\,\varepsilon,\quad e=\frac{c}{a}=\frac{|A|}{k}\,\varepsilon.$$

When $E<0$, we have $\varepsilon=1$ and the motion will have the period

$$T=\int_0^\tau r\,ds=\frac{\pi k}{\sqrt{-2E^3}}=\pi\sqrt{\frac{a^3}{2k}},\quad \tau=\frac{2\pi}{\sqrt{-2E}}.$$

The quantity a is *the major semi-axis* of the relative orbit.

As a consequence of the equations above, we have the following.

Theorem: *The Kepler–Newton trajectories are geodesic paths of the hyperplane sections of the light-cone in Minkowski space, in which the time coordinate was replaced by the radial coordinate r.*

An elliptic Kepler–Newton trajectory is defined by the equations

$$u=[c-a\,\cos(hs)]v_1-b\,\sin(hs)v_2,$$

$$t-t_0=as-h^{-1}c\,\sin(hs),\quad r=a-c\,\cos(hs).$$

where v_1, v_2 are constant vectors, a, b, c, h are positive numbers, and

$$v_1 \perp v_2, \quad |v_1| = |v_2| = 1, \quad b^2 = a^2 - c^2, \quad h = 2\pi a T^{-1},$$

$$h^2 = -2E, \quad 2a = -kE^{-1}, \quad 2T^2 = -\pi^2 k^2 E^{-3}.$$

Introducing *the eccentric anomaly* $\alpha = hs$, we obtain

$$u = (c - a\ \cos\ \alpha)v_1 - (b\ \sin\ \alpha)v_2, \quad r = a - c\ \cos\ \alpha,$$

$$t - t_0 = \frac{T}{2\pi}\,(\alpha - e\ \sin\ \alpha), \quad e = \frac{c}{a}.$$

It follows that an increase of α by 2π induces an increase of t by T and

$$\left| t - t_0 - \frac{T\alpha}{2\pi} \right| \leqslant \frac{Te}{2\pi}.$$

III. ALTERNATIVE LAGRANGIANS

The differential equations (I) can also be obtained from the Lagrangian

$$L = \tfrac{1}{2}(\|u'\|^2 - r'^2) + E(\|u\|^2 - r^2) + (\langle A, u\rangle - kr)$$

and have the following consequences:

$$F := ru'' - r'u' + ku = -r'u' + (2Er + k)u + rA = \text{const},$$

$$f := rr'' - r'^2 + kr = -r'^2 + 2Er^2 + 2kr = \text{const},$$

$$\|u'\|^2 - r'^2 - 2E(\|u\|^2 - r^2) - 2(\langle A, u\rangle - kr) = \text{const}.$$

The equations

$$ru'' - r'u' + ku = 0, \quad \ddot{u} + \frac{ku}{r^3} = 0$$

can be obtained as Euler–Lagrange equations of

$$L_1 = \frac{\|u'\|^2}{r} + 2Er, \quad L_2 = \|\dot{u}\|^2 - 2\varepsilon r\dot{s}^2, \quad \varepsilon = \pm 1$$

when we add the constraints

$$\|u\|^2 = r^2, \quad \frac{\|u'\|^2}{r} - 2Er = 2k, \quad r\dot{s} = h, \quad \varepsilon h^2 = k.$$

The trajectories associated with the Lagrangians L_1, L_2 are related to each other according to the following version of Maupertuis' theorem.

The version we give will allow us to get a Lagrangian for a relativistic two-body problem, generalizing the classical two-body problem.

Maupertuis 2000. Suppose we have three Lagrangians such as

$$M_0 = \sum_{i,j=1}^{n} g_{ij}\,\frac{du^i}{ds}\,\frac{du^j}{ds}, \quad M_1 = \sum_{i',j'=n+1}^{n+m} g_{i'j'}\,\frac{du^{i'}}{ds}\,\frac{du^{j'}}{ds}, \quad L = U(M_0 + M_1),$$

where U, $g_{ij} = g_{ji}$ *are functions of* $u^1, \ldots, u^n$ *and* $g_{i'j'} = g_{j'i'}$ *are functions of separate variables* $u^{n+1}, \ldots, u^{n+m}$.

Then the relations

$$\frac{dt}{ds}=U^{-1},\quad M_0^*=U^2M_0,\quad M_1^*=U^2M_1,\quad L^*=M_0^*+fU,\quad f\in\mathbb{R}$$

allow us to associate with each geodesic path of L verifying $L=f$ a trajectory of $L^*=M_0^*+fU$ and a geodesic path of M_1^* verifying

$$M_0^*-fU=-M_1^*.$$

The proof is straightforward.

We give some examples which will guide us towards a relativistic version of the classical two-body problem and which could lead to a proper formulation of the relativistic n-body problem.

Defining

$$r=\|u\|,\quad U=\frac{1}{r},\quad x'=\frac{dx}{ds},\quad \dot{x}=\frac{dx}{dt},$$

we consider the Lagrangians

$$L=\frac{u'^2+v'^2+\lambda t'^2+\lambda_1\tau'^2}{r},\quad L^*=\dot{u}^2+\dot{v}^2+\frac{2k}{r},\quad \lambda\in\mathbb{R},\ \lambda_1\in\mathbb{R},\ k\in\mathbb{R},$$

and it follows that, modulo the constraint $t'=dt/ds=r$, *the trajectories of L verifying*

$$\frac{u'^2+v'^2+\lambda t'^2+\lambda_1\tau'^2}{r}=2k$$

correspond bijectively to the trajectories of L^* verifying

$$L^*-\frac{4k}{r}=\dot{u}^2+\dot{v}^2-\frac{2k}{r}=-(\lambda+\lambda_1C_1^2),$$

where $C_1=\dot{\tau}$ is a conserved quantity.

We make now another choice, by taking

$$U=r,\quad \nu\in\mathbb{R},\quad \nu_1\in\mathbb{R},$$

and replacing τ by σ, t by s, and s by t. We then consider the Lagrangians

$$L_1^*=r^{-1}(u'^2+v'^2)+2Er,\quad L_1=(\dot{u}^2+\dot{v}^2)+\nu r\left(\frac{ds}{dt}\right)^2+\nu_1 r\left(\frac{d\sigma}{dt}\right)^2.$$

According to Maupertuis 2000, the trajectories of L_1^* with

$$r^{-1}(u'^2+v'^2)-2Er=-(\nu+\nu_1C^2)$$

are transformed into geodesic paths of the metric L_1 with $L_1=2E$, when we impose the constraints

$$\frac{ds}{dt}=r^{-1},\quad \frac{d\sigma}{dt}=Cr^{-1},\quad C\in\mathbb{R}.$$

We resume:

The Lagrangians

$$L=r^{-1}(u'^2+v'^2+\lambda t'^2+\lambda_1\tau'^2),\ \lambda\in\mathbb{R},\ \lambda_1\in\mathbb{R},$$

$$L^*=\dot{u}^2+\dot{v}^2+2kr^{-1},\ k\in\mathbb{R},$$

$$L_1=\dot{u}^2+\dot{v}^2+\nu r\dot{s}^2+\nu_1 r\dot{\sigma}^2,\quad \nu\in\mathbb{R},\quad \nu_1\in\mathbb{R},$$

$$L_1^*=r^{-1}(u'^2+v'^2)+2Er,\quad E\in\mathbb{R},$$

give rise to (modulo reparametrizations) equivalent trajectories (u,v) in $\mathbb{R}^3\times\mathbb{R}^3$, *provided we impose the following constraints*:

$$t'=r,\quad \dot{s}=r^{-1},\quad \tau'=C_1 r,\quad \sigma'=C.$$

IV. INVARIANT LAGRANGIANS IN MINKOWSKI SPACE

The previous considerations suggest for us to study the Lagrangians

$$K=\frac{\varepsilon|m_X X'+m_Y Y'|^2+m_X m_Y|X'-Y'|^2}{2G(m_X+m_Y)R},$$

where ε^2, G, m_X, and m_Y are positive numbers and $X=(x,t_x)$, $Y=(y,t_y)$ are trajectories in Minkowski space with a common parameter s, playing the role of a *proper time*. Primes indicate derivatives with respect to s and

$$|X'|^2=x'^2-t_x'^2,\quad |Y'|^2=y'^2-t_y'^2,\quad R^2=(x-y)^2-(t_x-t_y)^2.$$

We shall suppose that $R^2>0$, **i.e.**, $X-Y$ *is spacelike*.

Inspired by classical manipulations, we introduce the four-vectors

$$U=(u,\tau)=X-Y,\quad V=(v,t_V)=m_X^*X+m_Y^*Y,$$

where $m_X^*=M^{-1}m_X$, $m_Y^*=M^{-1}m_Y$, and $M=m_X+m_Y$. Then we have

$$X=V+m_Y^*U,\quad Y=V-m_X^*U,\quad R^2=|U|^2,$$

$$K=\frac{\varepsilon M^2|V'|^2+m_X m_Y|U'|^2}{2GMR}.$$

The Euler–Lagrange equations for the Lagrangian K are

$$(R^{-1}V')'=0,\quad (R^{-1}U')'=-kR^{-2}U,\quad k=GM\,\frac{K}{m_X m_Y}.$$

The first of these equations shows that there is a constant four-vector $W=(w,t_W)$ such that

$$\text{(II)}\quad V'=RW,\quad v'=Rw,\quad t_V'=Rt_W,$$

and the second Euler–Lagrange equation writes

$$\text{(III)}\quad RU''=R'U'-kU.$$

The equation (III) shows that $\langle U,U''\rangle=R'^2-kR$ and the expressions

$$\text{(IV)}\quad E=\frac{|U'|^2}{2R^2}-\frac{k}{R},\quad A=U''-2EU,$$

$$R^{-1}U\wedge U'=R^{-1}(u\times u',\tau' u-\tau u'),\quad U''\wedge U'=A\wedge U'+2EU\wedge U'$$

are constant. In fact,

$$E=-\frac{\varepsilon M^2|W|^2}{2m_X m_Y}$$

so that W and V' are *timelike* when $\varepsilon E>0$.

These relations combined with

$$\langle U,U'\rangle = RR', \quad \langle U,U''\rangle + U'^2 = RR'' + R'^2$$

will imply

$$R''-2ER=k, \quad U'^2=2ER^2+2kR$$

and

$$\frac{|U\wedge U'|^2}{-R^2}=R'^2-2ER^2-2kR=\langle A,U\rangle - kR = R'^2-U'^2=\text{const.}$$

When we introduce *the relative time of the center of gravity*,

$$t=\frac{t_V}{t_W}=\int R\ ds,$$

we get from equations (II) and (III) the astonishing relativistic equation

$$\text{(V)} \quad \frac{d^2U}{dt^2}=-\frac{kU}{R^3}.$$

We will also have, since $t'=R$,

$$R^{-1}U\wedge U' = U\wedge\frac{dU}{dt}=\text{angular } 4-\text{momentum.}$$

The general solution of the differential equations

$$U''=2EU+A, \quad R''=2ER+k$$

writes, for $E<0$ and $\varepsilon<0$,

$$U(s)=-\frac{A}{2E}+U_1\cos(hs)+U_2\sin(hs),$$

$$R(s)=-\frac{k}{2E}+R_1\cos(hs)+R_2\sin(hs),$$

where $h^2=-2E$, U_1 and U_2 are constant four-vectors, and R_1 and R_2 are real numbers.

The image of U in Minkowski space is an ellipse with center $A^*=-A/2E$ and semi-major axis $k^*=-k/2E=(G/M)K/\varepsilon|W|^2$.

The relations established above imply

$$R^{-1}U\wedge U'=\frac{h}{k^*}U_1\wedge U_2, \quad k^*A^*=R_1U_1+R_2U_2,$$

$$kR_1=\langle A,U_1\rangle, \quad kR_2=\langle A,U_2\rangle, \quad \langle U_1,U_2\rangle=R_1R_2,$$

$$|U_1|^2=R_1^2+\frac{k^2-|A|^2}{4E^2}, \quad |U_2|^2=R_2^2+\frac{k^2-|A|^2}{4E^2}, \quad R_1^2+R_2^2=\frac{|A|^2}{4E^2}.$$

We can suppose that $R_1\geqslant 0$ and $R_2=0$. Then

$$kA=4E^2R_1U_1, \quad k^*A^*=R_1U_1, \quad \langle U_1,U_2\rangle=0$$

and

$$|A^*|=R_1, \quad |U_1|^2=\frac{k^2}{4E^2}=k^{*2}, \quad |U_2|^2=\frac{k^2-|A|^2}{4E^2}=k^{*2}-|A^*|^2.$$

As a consequence, we have

$$|A|^2\geqslant 0, \quad |U_1|^2>0,$$

so that U_1 *is a spacelike vector* and $A=(4E^2/k)R_1U_1$ either vanishes (when $R_1=0$, i.e., when $R=-k/2E=k^*$ is constant) or is also spacelike.

The solutions thus obtained generalize the solutions to the Kepler–Newton problem and we obtain those solutions when we select the four-vectors $U_1=(u_1,t_1)$ and $U_2=(u_2,t_2)$ inside $\mathbb{R}^3\times\{0\}$, i.e., when

$$t_1=t_2=0, \quad U_1=(u_1,0), \quad U_2=(u_2,0).$$

Disregarding these conditions, our solutions in Minkowski space can be written in the form $U=(u,\tau)$ with

$$x-y=u=-\frac{a}{2E}+u_1\cos(hs)+u_2\sin(hs),$$

$$A=(a,t_A)=\frac{4E^2}{k}R_1(u_1,t_1), \quad h=\sqrt{-2E},$$

$$t_X-t_Y=\tau=-\frac{t_A}{2E}+t_1\cos(hs)+t_2\sin(hs),$$

and, without restricting the generality, we can add the formulas

$$R=\sqrt{r^2-\tau^2}=-\frac{k}{2E}-\frac{|A|}{2E}\cos(hs), \quad r=\sqrt{u^2}, \quad |A|=-2ER_1,$$

$$V=W\int R\,ds=\left(-\frac{ks}{2E}-\frac{|A|}{2Eh}\sin(hs)\right)W+\text{constant four-vector},$$

where

$$V=(v,t_V)=m_X^*X+m_Y^*Y, \quad W=(w,t_W)=\text{constant four-vector}.$$

When we compare Eq. (V) with the similar equation of Newtonian theory of Gravitation, *we are lead to assimilate the Newtonian time (of the pair (X,Y)) with the variable* $t=\int R\,ds=t_V/t_W$, where t_V is the fourth component of the vector V, which represents the *center of gravity*.

Note that an increase of s by $2\pi/h$ produces an increase of V by

$$\Delta V=-\frac{\pi k}{Eh}W=\frac{2\pi k^*}{h}W.$$

In the frame of center of gravity, $dV/dt=(0,t_W)$; when $E<0$, one obtains for the *period* T of the motion the expression

$$T=\Delta t=-\frac{\pi k}{Eh}=\frac{2\pi k^*}{h}.$$

For $E<0$, the solutions described previously define elliptic trajectories in Minkowski space

$$X=V+m_Y^*U, \quad Y=V-m_X^*U,$$

with centers

$$S_X=-\frac{m_Y^*}{2E}A+V=m_Y^*A^*+V,\quad S_Y=\frac{m_X^*}{2E}A+V=-m_X^*A^*+V$$

moving with constant velocity four-vector

$$\dot{S}_X=\dot{S}_Y=\dot{V}=\frac{dV}{dt}=W=(w,t_W).$$

We have introduced the notations

$$A^*=-\frac{A}{2E}=-\frac{2E}{k}R_1U_1,\quad k^*=-\frac{k}{2E}.$$

Thus we have, for $\lambda=R_1/k^*$: $A^*=\lambda U_1$ and

$$|U_1|^2=k^{*2},\quad |A^*|^2=\lambda^2k^{*2},\quad |U_2|^2=k^{*2}(1-\lambda^2),\quad \langle U_1,U_2\rangle=0.$$

For $\lambda^2\neq 1$, we obtain the following solutions:

$$U(s)=k^*(\lambda+\cos(hs))U_1^*+k^*\sqrt{\varepsilon'(1-\lambda^2)}\,\sin(hs)U_2^*,$$

$$R(s)=k^*(1+\lambda\,\cos(hs)),\quad V(s)=k^*W(s+\lambda h^{-1}\,\sin(hs))+\text{const},$$

where U_1^*, U_2^* are constant four-vectors verifying the conditions:

$$\langle U_1^*,U_2^*\rangle=0,\quad |U_1^*|^2=1,$$

$$|U_2^*|^2=\varepsilon',\quad \varepsilon'^2=1,\quad \varepsilon'(1-\lambda^2)\geqslant 0.$$

We shall suppose, without losing generality, that $\lambda\geqslant 0$.

The formula which defines U implies the identity

$$|U(s)|+|U(s)-2k^*\lambda U_1^*|=2|k^*|;$$

when $U_2\neq 0$, the image of U in Minkowski space is an affine ellipse. Then we interpret the last formula saying that:

When $U_2\neq 0$, the image of U is a Minkowskian ellipse with *focuses* $F_1=0$ and $F_2=2k^*\lambda U_1^*$, *center* $A^*=k^*\lambda U_1^*$, *major axis* $2k^*$, *and Minkowskian eccentricity* $e=\lambda$.

The case $e=\lambda=1$ arrives when the four-vector U_2 is on the light cone; in contrast to properties of conic sections in Euclidean spaces, where eccentricity 1 characterizes parabolas, we see that a Minkowskian ellipse can have $e=1$.

Supposing again that $\lambda^2\neq 1$. Let us now consider the orthogonal projections $u(s)$, u_1^*, u_2^* of the four-vectors

$$U(s)=(u(s),\tau(s)),\quad U_1^*=(u_1^*,t_1^*),\quad U_2^*=(u_2^*,t_2^*)$$

on $\mathbb{R}^3\times\{0\}$. Then we have

$$u(s)=k^*(\lambda+\cos(hs))u_1^*+k^*\sqrt{\varepsilon'(1-\lambda^2)}\,\sin(hs)u_2^*$$

and

$$\|u_1^*\|^2=t_1^{*2}+1,\quad \|u_2^*\|^2=t_2^{*2}+\varepsilon',\quad \langle\langle u_1^*,u_2^*\rangle\rangle=t_1^*t_2^*.$$

The image of u is an ellipse in $\mathbb{R}^3$ with center $a^*=k^*\lambda u_1^*$. When $t_2^*t_1^*\neq 0$, *the foci of this ellipse are not even colinear with the origin; this new anomaly never arises in Newton's theory*, according to which the attractive origin is always in one of the foci.

We get three canonical solutions when we take

$$\varepsilon' = 1, \quad \lambda^2 < 1, \quad U_1^* = (0,1,0,0), \quad U_2^* = (0,0,1,0),$$

$$\varepsilon' = -1, \quad \lambda^2 > 1, \quad U_1^* = (1,0,0,0), \quad U_2^* = (0,0,0,1),$$

$$\lambda = 1, \quad U_1^* = (1,0,0,0), \quad U_2 = (0,0,1,1).$$

All other solutions are obtained from these three particular solutions by applying Lorentz transformations.

The case $\varepsilon' = -1$, $\lambda^2 > 1$, contains the following example:

$$u_\mu(s) = k^*(\lambda + \cos(hs), \ \mu\sqrt{\lambda^2 - 1}\ \sin(hs), 0), \quad \mu \in \mathbb{R},$$

$$\tau_\mu(s) = k^*\sqrt{(\lambda^2 - 1)(\mu^2 + 1)}\ \sin(hs),$$

which is Lorentz equivalent to the solution

$$u_0(s) = k^*(\lambda + \cos(hs), 0, 0), \quad \tau_0(s) = k^*\sqrt{\lambda^2 - 1}\ \sin(hs).$$

This example shows that *an elliptic trajectory in space* $\mathbb{R}^3 \times \{0\}$ *can be Lorentz equivalent to a rectilinear trajectory in three-space*.

The image of the path u_μ is an ellipse in $\mathbb{R}^3$ with center $(k^*\lambda, 0, 0)$. When $\mu^2(\lambda^2 - 1) < 1$, the major axis and the foci are respectively $2|k^*|$ and $(k^*\lambda + c, 0, 0), (k^*\lambda - c, 0, 0)$ with $c^2 = k^{*2}(1 - \mu^2(\lambda^2 - 1))$. However, when $\mu^2(\lambda^2 - 1) > 1$, the major axis and the foci will be $2|k^*|\sqrt{\mu^2(\lambda^2 - 1)}$, respectively, $(k^*\lambda, d, 0), (k^*\lambda, -d, 0)$ with $d^2 = k^{*2}(\mu^2(\lambda^2 - 1) - 1)$.

Notes. This work has been partially inspired by the discussion given at page 11 in Ref. 5.

For the history of the Bernoulli–Laplace eccentricity vector see Ref. 6.

[1] W. R. Hamilton, ''On a general Method in Dynamics,'' Philos. Trans. R. Soc. London **124**, 247–308 (1834), also published in Math. Pap. **2**, 103–162 (1940).
[2] P. I. Bakulin, E. V. Kononowitch, and V. I. Moroz, *A Course in General Astronomy* (in Russian) (Nauka, Moscow, 1977), p. 83.
[3] P. Jordan, *Anschauliche Quantentheorie* (Springer, Berlin, 1936), p. 166.
[4] K. Teleman, ''Conformal Lorentz geometry revisited,'' J. Math. Phys. **37**, 1076–1085 (1996).
[5] I. D. Lawrie, *A Unified Grand Tour of Theoretical Physics* (Institute of Physics, Bristol, 1994).
[6] H. Goldstein, ''More on the prehistory of the Laplace or Runge–Lenz vector,'' Am. J. Phys. **44** , 1123–1124 (1976).

Maxwell fields in anisotropic space–times

Matthias Wollensak,[a)]
Theoretisch-Physikalisches Institut, Friedrich-Schiller-Universität Jena, Max-Wien-Platz 1, D-07743 Jena, Germany

(Received 9 December 1997; accepted for publication 16 June 1998)

Maxwell's equations are rewritten with respect to anholonomic orthonormal frames and applied to the case when the background is described by a diagonal Bianchi-type I universe. It is shown that solutions analogous to plane wave solutions in flat space–time must obey *two* transversality conditions when the space–time is anisotropic. Analytic solutions are presented for propagation along the respective coordinate axes and a special stiff fluid model. Furthermore, by use of a generally applicable ansatz the long- and short-time behavior of waves traveling in the axisymmetric Kasner universe have been investigated. © *1998 American Institute of Physics.* [S0022-2488(98)03010-2]

I. INTRODUCTION

Electromagnetic–gravitational interactions have been studied in numerous investigations and there exists a considerable amount of exact solutions of Einstein–Maxwell systems.[1] In many cases, however, it is sensible and sufficient to treat electromagnetic backreactions on space–time geometry only perturbatively[2,3] or to neglect these effects altogether.[4,5] Hence, provided that the influence of the electromagnetic stress-energy-tensor on the underlying geometry can be assumed to be small enough the situation simplifies considerably. The problem then reduces to the determination of solutions of Maxwell's equations in a given background metric. In the ensuing calculations we adopt this point of view.

If one is interested in effects of electromagnetic-gravitational interactions only, then it is sufficient to compute the Faraday tensor $F_{\mu\nu}$. However, when additional (matter) fields are involved it is even on the classical level necessary to calculate the Maxwell fields A_μ explicitly because the corresponding interaction term is usually written as a bilinear expression of matter and Maxwell fields.[6]

In order to treat Maxwell fields in curved spaces we propose the use of an orthonormal frame instead of the usually employed coordinate frame. This approach puts the treatment of Maxwell fields in a curved background on the same geometric footing as that of Dirac fields[7,8,9] and therefore facilitates in a natural way the coupling of spin-$\frac{1}{2}$ and spin-1-fields. It also has the advantage that the field components with respect to an orthonormal frame represent the physical components.[8,9] Moreover, calculations can be easier performed due to the particularly simple metric tensor.

The paper is organized as follows: In Sec. II we present Maxwell's equations and the Lorenz gauge with respect to orthonormal frames. In Sec. III attention is focused on Maxwell fields in diagonal Bianchi-type I cosmologies. We specialize to the sourcefree case and derive the corresponding differential equations. For solutions analogous to plane waves in Minkowskian space–time we find in addition to the usual transversality condition a second one which has no counterpart in flat space–time. We then develop general expressions for the long- and short-time behavior of those Maxwell fields. In Sec. IV we consider first the special case of waves propagating along coordinate axes. Exact solutions are calculated which confirm our asymptotic expressions. Next we study a universe filled with a perfect stiff fluid. We obtain exact solutions of the corresponding Maxwell system and determine again their asymptotic behavior. In the third model the long- and short-time behavior of Maxwell fields propagating in the axisymmetric Kasner universe are computed.

[a)]Permanent address: Lohengrinstr. 40, 70597 Stuttgart, Germany.

 © 1998 American Institute of Physics

A final remark on notation: Greek indices run from 0 to 3, Latin indices from 1 to 3, and repeated Greek indices are summed over. Throughout this paper we use metric signature $(+---)$.

II. ORTHONORMAL FRAMES AND MAXWELL FIELDS

We give a brief derivation of Maxwell's equations with respect to orthonormal frames which will be used later on. In the following the notion frame will be synonymously used for cobasis, i.e., it stands for a set of (orthonormal) covectors spanning the cotangent space $T_p^*(M)$ of a differentiable manifold M. Our starting point is the frame independent formulation of the Maxwell system,

$$dF=0, \quad \delta F=-J \tag{1}$$

together with Lorenz gauge

$$\delta A=0, \tag{2}$$

where F denotes the electromagnetic field two-form and J, A the four-current, and Maxwell field 1-forms, respectively. Here, δ is the codifferential which is up to a sign identical with $*d*$, where $*$ is the well-known Hodge operator. In general an orthonormal basis is not holonomic so that the commutators of basis vectors e_ν and equivalently the exterior derivative of covectors Θ^μ are not necessarily vanishing any more,[8–10] $[e_\mu,e_\nu]=C^\alpha{}_{\mu\nu}e_\alpha$ and $d\Theta^\mu=-\frac{1}{2}C^\mu{}_{\alpha\beta}\Theta^\alpha\wedge\Theta^\beta$. The commutation (or anholonomic) coefficients $C^\mu{}_{\nu\lambda}$ which take the place of the Christoffel fields are obviously antisymmetric in the second and third indices.

The Maxwell field 1-form, the four-current 1-form and the electromagnetic field 2-form are given by

$$A=A_\mu\Theta^\mu, \quad J=J_\mu\Theta^\mu, \quad F=\tfrac{1}{2}F_{\mu\nu}\Theta^\mu\wedge\Theta^\nu,$$

with

$$F_{\mu\nu}=e_\mu(A_\nu)-e_\nu(A_\mu)-A_\alpha C^\alpha{}_{\mu\nu}.$$

Now since

$$d*F=\frac{1}{3!}\,\epsilon_{\alpha\beta\gamma\delta}\Theta^\alpha\wedge\Theta^\beta\wedge\Theta^\gamma\left\{C^\mu{}_{\mu\nu}F^{\nu\delta}+\frac{1}{2}\,C^\delta{}_{\mu\nu}F^{\mu\nu}+e_\mu(F^{\mu\delta})\right\}$$

and $d*F=*J=(1/3!)\epsilon_{\alpha\beta\gamma\delta}J^\delta\Theta^\alpha\wedge\Theta^\beta\wedge\Theta^\gamma$ one arrives at

$$C^\mu{}_{\mu\nu}F^{\nu\lambda}+\tfrac{1}{2}C^\lambda{}_{\mu\nu}F^{\mu\nu}+e_\mu(F^{\mu\lambda})=J^\lambda. \tag{3}$$

Analogously one finds for the first of Eqs. (1),

$$H_{(\mu\nu\lambda)}=0, \quad \mu<\nu<\lambda, \quad H_{\mu\nu\lambda}=e_\mu(F_{\nu\lambda})+C^\alpha{}_{\mu\nu}F_{\lambda\alpha}, \tag{4}$$

where the brackets denote even permutation of the indices. (3) and (4) represent Maxwell's equations with respect to an *orthonormal* frame.

It remains for us to derive the analogous expression for the Lorenz gauge (2). To this end one calculates the exterior derivative of the 3-form $*A$,

$$d*A=-\Theta^0\wedge\Theta^1\wedge\Theta^2\wedge\Theta^3\{e_\mu(A^\mu)+C^\mu{}_{\mu\nu}A^\nu\},$$

and by use of (2) the Lorenz condition can be read off immediately,

$$e_\mu(A^\mu)+C^\mu{}_{\mu\nu}A^\nu=0. \tag{5}$$

Lorenz gauge is *not* invariant under conformal transformations. Hence, the application of (5) to a term like

$$e_\mu(F^{\mu\lambda})=\eta^{\lambda\kappa}\{e_\mu e^\mu(A_\kappa)+e_\kappa(C^\mu{}_{\mu\nu}A^\nu)+C^\rho{}_{\kappa\mu}e_\rho(A^\mu)+e^\rho(A_\mu C^\mu{}_{\kappa\rho})\}, \tag{6}$$

necessarily spoils conformal invariance of Eq. (3).

III. MAXWELL FIELDS IN BIANCHI-TYPE I UNIVERSES

The general line element of diagonal Bianchi-type I space–times (henceforth abbreviated by BS) is defined by

$$ds^2=dt^2-\sum_{j=1}^{3}\alpha_j^2(t)(dx^j)^2.$$

A natural choice of the orthonormal frame is given by $\Theta^0=dx^0=dt$, $\Theta^j=\alpha_j dx^j$ while the tangent vectors read $e_0=\partial_0=\partial_t$, $e_j=(1/\alpha_j)\partial_j$ satisfying $\Theta^\mu(e_\nu)=\delta^\mu_\nu$. The metric coefficients are given by $\eta_{\mu\nu}=\mathrm{diag}(+1,-1,-1,-1)$. There are three nonzero commutation coefficients

$$C^j{}_{0j}=-\frac{\dot\alpha_j}{\alpha_j}=-C^j{}_{j0}. \tag{7}$$

Overdot means derivation with respect to t ($e_0=\partial_t=\dot{}$). Hence, the field equations (3) take the form

$$e_\mu(F^\mu{}_0)=J_0, \tag{8}$$

$$e_\mu(F^\mu{}_j)=(\mathscr{C}-C_j)F_{0j}+J_j, \tag{9}$$

with $C_j=C^j{}_{0j}$ and $\mathscr{C}=\Sigma_{j=1}^3 C_j=-\partial_t \ln\sqrt{|g|}$, where $|g|=\alpha_1\alpha_2\alpha_3$ denotes the absolute value of $\det g_{\mu\nu}$. On substituting (6) into (8), (9) one gets

$$J_0=e_\mu e^\mu(A_0)-e_0e^\nu(A_\nu)-2\sum_{j=1}^{3}C_j e_j(A_j), \tag{10}$$

$$J_j=e_\mu e^\mu(A_j)-\mathscr{C}\dot A_j+(C_j\mathscr{C}-C_j^2-\dot C_j)A_j-2C_j e_j(A_0). \tag{11}$$

Spatial translation invariance of BS suggests ansatz,

$$A_\mu(t,\mathbf{x})=\int d^3k f_\mu(t,\mathbf{k})e^{i\mathbf{kx}} \tag{12}$$

for the Maxwell fields (bold letters denote 3-vectors). Restricting ourselves from now on to the sourceless case $J_\mu=0$ we find from (10) and (11), resp.,

$$f_0=\frac{i}{p^2}(\dot{\mathbf{p}}\mathbf{f}-\mathbf{p}\dot{\mathbf{f}}) \tag{13}$$

and

$$\partial_t\left(\frac{\sqrt{|g|}}{\alpha_j^2}\partial_t(\alpha_j f_j)\right)+\frac{\sqrt{|g|}}{\alpha_j^2}p^2(\alpha_j f_j)=i\frac{\sqrt{|g|}}{k_j}(\partial_t p_j)^2 f_0, \tag{14}$$

where

$$p_j(t)=\frac{k_j}{\alpha_j(t)} \tag{15}$$

denotes the jth component of the *physical* 3-momentum[11] and $p\equiv|\mathbf{p}|$. Equation (13) is the BS analog of the Minkowskian equation of constraint in the Maxwell theory.

Insertion of ansatz (12) into (5) yields

$$\dot{f}_0 - \mathscr{C} f_0 = \frac{1}{\sqrt{|g|}}\,\partial_t(\sqrt{|g|} f_0) = i\mathbf{pf}, \tag{16}$$

which resembles constraint (13) in that the time-like component f_0 couples to all space-like components f_j.

In what follows we are now seeking solutions of the Maxwell system (13), (14) analogous to plane wave solutions in flat space–time. To this end we consider those fields whose scalaric component vanishes,

$$f_0 \equiv 0. \tag{17}$$

This ansatz is consistent with the Lorenz gauge (16) if the generalized transversality condition

$$\mathbf{pf} = 0 \tag{18}$$

holds. It implies that with increasing t a wave with an arbitrary wave vector will tilt toward that direction with the smallest Hubble constant $H_j = \dot{\alpha}_j / \alpha_j$.

The absence of scalaric photons is established by an additional *second* transversality condition

$$\dot{\mathbf{p}}\mathbf{f} = 0, \tag{19}$$

following from (17), (18) together with constraint (13). (19) enforces orthogonality between the spatial part **f** of the spectral components of the Maxwell fields and the change of momentum **p** with time t. It has no counterpart in flat space. To summarize, ansatz (17) leads to the following (sourcefree) Maxwell system describing transversal fields traveling in a BS background,

$$F_0(t,\mathbf{k}) = 0, \quad \{\partial_t^2 + \Omega_j^2(t)\} F_j(t,\mathbf{k}) = 0, \tag{20}$$

with $F_\mu = |g|^{1/4} f_\mu$ and time-dependent frequency $\Omega_j = \sqrt{\dot{b}_j - b_j^2 + p^2}$, where $b_j = \dot{B}_j / 2B_j$ and $B_j = \alpha_j^2 / \sqrt{|g|}$. The transversality constraints read $\mathbf{pF} = 0 = \dot{\mathbf{p}}\mathbf{F}$. We introduce now a transformation defined by

$$F_j^{\pm}(t,\mathbf{k}) = V_j^{\pm} \exp\left\{ \int_{t_0}^{t} dx [-b_j(x) + ip(x) H_j^{\pm}(x)] \right\}, \quad H_j^{\pm}(x) = \frac{\sqrt{1 + [h_j^{\pm}(x)]^2} \pm 1}{h_j^{\mp}(x)}, \tag{21}$$

where the two functions $F_j^{\pm}$ constitute the fundamental system of solutions of (20). By virtue of (21) Eq. (20) can be cast into a nonlinear first order differential equation for the functions $h_j^{\pm}$,

$$\frac{\partial_x h_j^{\mp}(x)}{\sqrt{1 + [h_j^{\pm}(x)]^2}} \pm h_j^{\mp}(x) \partial_x \ln \beta_j(x) - 2ip(x) = 0, \tag{22}$$

with

$$\beta_j(x) = \frac{\alpha_j^2(x)}{p(x)\sqrt{|g(x)|}}. \tag{23}$$

It turns out that here and in the following one need not care about discontinuities in the interval $(-\infty, 0)$ so that the principal branch of the square root can always be taken.

The form of (21) suggests that $|h_j^{\pm}| \to \infty$ is equivalent to the long-time (or high-frequency) limit. In this case $F_j^{\pm}$ exhibits the typical WKB phase factor because $|H_j^{\pm}|$ tends to unity. Analogously, it will be shown below that $|h_j^{\pm}| \to 0$ corresponds to the short-time limit. For $|h_j^{\pm}| \ll 1$ the square root in (22) can be replaced by unity neglecting terms of order $[h_j^{\pm}]^2$. One obtains a linear first order differential equation with solutions

$$h_j^{\pm}(t) = \beta_j^{\mp 1}(t) \left\{ 2i \int_{t_0}^{t} dx p(x) \beta_j^{\pm 1}(x) + \beta_j^{\pm 1}(t_0) h_j^{\pm}(t_0) \right\}. \tag{24}$$

For $|h_j^{\pm}| \gg 1$ one encounters a nonlinear differential equation of Bernoulli-type where terms of the order $[h_j^{\pm}]^{-2}$ have been discarded. Again, the integration is feasible yielding

$$\frac{1}{h_j^{\pm}(t)} = \pm \int_{t_0}^{t} d\mu_j(x) P(t;x) + \frac{P(t;t_0)}{h_j^{\pm}(t_0)}, \tag{25}$$

with $\mu_j = \ln \beta_j$ and $P(t;x) = \exp\{-2i\int_x^t d\tau p(\tau)\}$. Taking then the expansion of $H_j^{\pm}$ for small $|h_j^{\pm}|$ and inserting the result into (21) one finds

$$F_j^{+}(t,\mathbf{k}) = \tilde{V}_j^{+} \sqrt{\frac{B_j(t_0)}{B_j(t)}} \exp\left\{ i \int_{t_0}^{t} dx p(x) \left[\frac{2}{h_j^{+}(x)} + \lambda_{j,1}^{+}(x) \frac{h_j^{+}(x)}{2} - \lambda_{j,2}^{+}(x) \left(\frac{h_j^{+}(x)}{2} \right)^3 + \cdots \right] \right\}, \tag{26}$$

$$F_j^{-}(t,\mathbf{k}) = \tilde{V}_j^{-} \sqrt{\frac{B_j(t_0)}{B_j(t)}} \exp\left\{ i \int_{t_0}^{t} dx p(x) \left[\frac{h_j^{-}(x)}{2} - \lambda_{j,1}^{-}(x) \left(\frac{h_j^{-}(x)}{2} \right)^3 + \cdots \right] \right\}. \tag{27}$$

Note that the $h_j^{\pm}$ as given by (24) are correct only to the *lowest* order! The higher order corrections are described by the coefficients $\lambda_{j,n}^{\pm}(x)$ which relate the nth order expression ${}^{(n)}h_j^{\pm}$ to the lowest order result $h_j^{\pm}$. For $n=1$ holds, e.g.,

$${}^{(1)}h_j^{\pm}(x) = h_j^{\pm}(x) \left(1 + \frac{1 - \lambda_{j,1}^{\pm}(x)}{4} [h_j^{\pm}(x)]^2 \right). \tag{28}$$

Insertion of (28) into (22) yields after some algebra

$$\lambda_{j,1}^{\pm}(x) = 1 - \frac{\int^x dz I_j^{\pm}(z) \partial_z [h_j^{\pm}(z)]^2}{I_j^{\pm}(x) [h_j^{\pm}(x)]^2}, \tag{29}$$

where

$$I_j^{\pm}(x) = \int^x dy p(y) \beta_j^{\pm 1}(y). \tag{30}$$

For $|h_j^{\pm}| \gg 1$ the procedure can be analogously carried out.

Of particular interest are anisotropically expanding universes with $\alpha_j(t) = t^{q_j}$ subject to constraint $\Sigma_{j=1}^{3} q_j = 1$, $|q_j| < 1$. From (15) and (23) follows $p_j(t) = k_j t^{-q_j}$ and $\beta_j(t) = t^{3q_j - 1}/k_j$. The lowest order expressions for the functions $h^{\pm}(t)$ simplify then to

$$h_j^{+}(t) = \frac{i}{q_j} t p(t), \tag{31}$$

$$h_j^{-}(t) = \frac{i}{p(t)} \sum_{n=1}^{3} \frac{k_n^2}{1 - q_n - q_j} t^{1 - 2q_n}. \tag{32}$$

An additional logarithmic factor appears in (31) when $q_j = 0$ and similarly in (32) if $q_n + q_j = 1$. Note that Eqs. (31), (32) are only sensible in the case $|h_j^{\pm}| \to 0$. This is equivalent to the short-time limit if $p(t)$ diverges slower than t^{-1}. For exponents $|q_j| < 1$ as demanded above this condition is always fulfilled.

IV. EXACT AND ASYMPTOTIC SOLUTIONS

Exact solutions can be easily found if one restricts his attention to fields propagating along the respective coordinate axes. But there exists no ''superposition'' principle which would allow the construction of general results by use of those special results. These solutions are nonetheless interesting as limiting cases of the general situation with arbitrary wave vectors. Below we will present exact solutions for BS with $\alpha_j(t) = t^{q_j}$.

One faces a more complicated situation if one considers fields propagating with an arbitrary wave vector **k** in an axially symmetrical BS. In most cases there are no analytic solutions at disposal. We investigate here two models, an exactly solvable stiff fluid model and a Kasner space–time. Before doing so we note that the transversality conditions (18), (19) together with restriction $\alpha_1(t)=\alpha_2(t)$ lead to

$$f_0(t,\mathbf{k})=f_3(t,\mathbf{k})=0, \quad k_1 f_1(t,\mathbf{k})=-k_2 f_2(t,\mathbf{k}), \tag{33}$$

i.e., linear polarization.

A. Propagation along the coordinate axes

In the ensuing calculation we set without loss of generality $k_1\equiv k$, $k_2=0=k_3$, i.e., we consider propagation along the x^1-axis.[3,4] It follows from (18) that a wave propagating along a special coordinate axis will keep propagating along this axis. But this motion is only stable when the propagation originally started along that axis with the smallest Hubble constant. Otherwise small deviations off the axis will increase with time and cause the wave to tilt toward that axis with the smallest Hubble constant.

The differential equations pertaining to the spatial components $f_j=t^{-1/2}F_j$ read according to (20)

$$\{\partial_t^2+\Omega_j^2(t)\}F_j(t)=0, \tag{34}$$

$$\Omega_j^2(t)=\frac{1-4q_j^2}{4t^2}+k^2t^{-2q_1} \tag{35}$$

$(j\neq 1)$ while the time-like component vanishes. The general solution of (34) may be written as

$$F_j^{\pm}(t)=\sqrt{t}\mathscr{Q}_{\pm\nu_j}(\gamma(t)/2) \quad (j\neq 1), \quad F_0^{\pm}(t)=0=F_1^{\pm}(t), \tag{36}$$

with $\nu_j=q_j/(1-q_1)$ and $\gamma(t)=2kt^{1-q_1}/(1-q_1)$, where $\mathscr{Q}_\nu(z)$ denotes Bessel functions of the first kind J_ν and Y_ν or any linear combination thereof, e.g., Hankel's functions.[12] As is well known J_ν, $J_{-\nu}$ represent a fundamental system of solutions of (34) for noninteger ν while in the case of integer indices a fundamental system is given by J_ν and Y_ν.

We now derive the corresponding expressions in the short- and long-time limit using ansatz (21). Beginning with case $t\to\infty$ one finds upon insertion of $p_j=k_jt^{-q_j}$ and $\beta_j=t^{3q_j-1}/k_j$ into (25)

$$\frac{1}{h_j^{\pm}(t)}\sim\pm\frac{1-q_1-2q_j}{1-q_1}\frac{i}{\gamma(t)}+\mathscr{O}(\gamma^{-2}). \tag{37}$$

On expanding $H_j^{\pm}$ in terms of $1/h_j^{\pm}$ and inserting the result into (21) one arrives at $(j\neq 1)$

$$F_j^{\pm}(t,\mathbf{k})\sim V_j^{\pm}(t_0)t^{q_1/2}\exp\left\{\pm\frac{ik}{1-q_1}t^{1-q_1}\right\} \quad (t\to\infty). \tag{38}$$

(38) exhibits the correct long-time behavior of the exact solutions (36).

The short-time limit coincides with case $|h_j^{\pm}(t)|\to 0$. This has already been shown in Sec. III. One gets from (31), (32),

$$h_j^{+}(t)=\frac{ik}{q_j}t^{1-q_1}, \tag{39}$$

$$h_j^{-}(t)=\frac{ik}{1-q_1-q_j}t^{1-q_1}. \tag{40}$$

The cases $q_j=0$ and $q_j+q_1=1$ have been omitted here because they can be carried out in an entirely analogous way. On calculating the first order corrections

$$\lambda_{j,1}^{+}(0)=\frac{\nu_j}{1+\nu_j},\quad \lambda_{j,1}^{-}(0)=\frac{\nu_j-1}{\nu_j-2},$$

and substituting (39), (40) into (26) and (27), respectively, one obtains in the small-t limit ($j\neq 1$),

$$F_j^{+}(t)=\tilde{V}_j^{+}(t_0)t^{1/2+q_j}\left\{1-\frac{1}{1+\nu_j}\left(\frac{\gamma(t)}{4}\right)^2+\mathcal{O}(\gamma^4)\right\} \tag{41}$$

and

$$F_j^{-}(t)=\tilde{V}_j^{-}(t_0)t^{1/2-q_j}\left\{1-\frac{1}{1-\nu_j}\left(\frac{\gamma(t)}{4}\right)^2+\frac{1}{2}\frac{1}{(1-\nu_j)(2-\nu_j)}\left(\frac{\gamma(t)}{4}\right)^4-\mathcal{O}(\gamma^6)\right\}. \tag{42}$$

These are the short-time expansions up to the order γ^2 and γ^4, respectively, of the exact solutions (36).

B. Stiff-fluid model

We consider exact solutions of Einstein's field equations when the material content is described by the ''perfect-fluid'' stress tensor $T_{\alpha\beta}=(\mu+\mathcal{P})u_\alpha u_\beta+\mathcal{P}g_{\alpha\beta}$, where $\mathcal{P}$ denotes pressure, μ the energy density consisting of the rest mass density plus total internal energy, and u_α the 4-velocity. The equation of state reads $\mathcal{P}=(\zeta-1)\mu$ with ζ a constant. The special case $\zeta=2$ is often referred to as the ''stiff matter'' case. The line element pertaining to it is given by $ds^2=dt^2-\Sigma_{j=1}^{3}(t^{q_j}dx^j)^2$, where $q_1=q_2=q$, $q_3=1-2q$. In the following we are going to study Maxwell fields in a curved space–time described by this model for $q=1/2$, i.e.,

$$q_1=q_2=\tfrac{1}{2},\quad q_3=0. \tag{43}$$

The corresponding differential equation assumes according to (20) the simple form

$$\{\partial_t^2+p^2(t)\}F_j(t,\mathbf{k})=0 \tag{44}$$

with $p^2(t)=k_3^2+\kappa^2/t$, where $\kappa^2=k_1^2+k_2^2$. Here, one must only consider the cases $j=1,2$ because (33) enforces $F_0=F_3=0$. Solutions of (44) are

$$F_j^{+}(t,\mathbf{k})=M_{\eta,1/2}(-2ik_3t), \tag{45}$$

$$F_j^{-}(t,\mathbf{k})=W_{\eta,1/2}(-2ik_3t), \tag{46}$$

where M and W are Whittaker's functions[12] and $\eta=i\kappa^2/2k_3$. The complete set of solutions satisfying constraint (33) reads

$$F_0^{\pm}(t,\mathbf{k})=0=F_3^{\pm}(t,\mathbf{k}),\quad \tilde{F}_1^{\pm}(t,\mathbf{k})=D^{\pm}k_2F_1^{\pm}(t,\mathbf{k}),\quad \tilde{F}_2^{\pm}(t,\mathbf{k})=-D^{\pm}k_1F_2^{\pm}(t,\mathbf{k}), \tag{47}$$

with constant prefactors $D^{\pm}$. We compare these exact results with the corresponding outcome based on our asymptotic expressions. In the long-time limit follows for $k_3\neq 0$,[13]

$$\frac{1}{h_j^{\pm}(t)}=\mp\frac{1-2q_j}{2i|k_3|t}\left\{1+\mathcal{O}\left(\frac{1}{2i|k_3|t};\frac{1}{at}\right)\right\} \tag{48}$$

with $a=k_3^2/\kappa^2$, $\kappa^2=k_1^2+k_2^2$. Expansion of ansatz (21) for large $|h_j^{\pm}|$ eventually gives

$$F_j^{\pm}(t,\mathbf{k})=V_j^{\pm}(t_0,\mathbf{k})t^{\pm i|\eta|}e^{\pm i|k_3|t[1+\mathcal{O}(t^{-2})]}\quad (t\to\infty), \tag{49}$$

where terms of order $(|k_3|t)^{-1}$, $(at)^{-1}$ have been neglected. For the short-time limit we use again (26), (27). After some manipulations one arrives at

$$h_j^+(x)=2ix\sqrt{\frac{\kappa^2}{x}+k_3^2}, \tag{50}$$

$$h_j^-(x)=2i\,\frac{\kappa^2\,\ln(x)+k_3^2}{\sqrt{\frac{\kappa^2}{x}+k_3^2}}, \tag{51}$$

and

$$\lambda_{j,1}^+(x)=\tfrac{1}{2}-\tfrac{1}{6}ax+\mathcal{O}(\kappa^2x)+\mathcal{O}(a^2x^2), \tag{52}$$

$$\lambda_{j,1}^-(x)=\frac{1}{\ln(x)}\left\{1-\frac{2}{\ln(x)}+\frac{2}{[\ln(x)]^2}+\cdots+ax+\mathcal{O}\left(\frac{ax}{\ln(x)}\right)\right\}, \tag{53}$$

where the dots in (53) indicate that there are still terms present that are decaying slower than x for $x\to 0$. But these terms can be omitted here because they do not depend on $a\equiv k_3^2/\kappa^2$ and therefore do not contribute to the ''interaction'' terms. With the interaction the following is meant here: As has been discussed in detail in Sec. IV A one can study propagation along the axes (or within planes in axisymmetric models), i.e., the wave vector has two (or one) vanishing components. One gets exact results depending in the present model on either κ, namely (apart from a constant factor)

$$F_j^+(t,\kappa)=\sqrt{t}J_1(2\kappa\sqrt{t}), \tag{54}$$

$$F_j^-(t,\kappa)=\sqrt{t}\left\{Y_1(2\kappa\sqrt{t})-\frac{2}{\pi}\,[\ln(\kappa)-\psi(1)]J_1(2\kappa\sqrt{t})\right\} \tag{55}$$

(where J_1, Y_1 are the Bessel functions introduced earlier and ψ denotes the Digamma function), or on k_3, namely,

$$F_j^+(t,k_3)=\sin(k_3t), \tag{56}$$

$$F_j^-(t,k_3)=\cos(k_3t). \tag{57}$$

The explicit determination of all *a-independent* terms in (52), (53) can be much easier accomplished by directly using (54), (55). But for an arbitrary wave vector there are additional terms of the form $\kappa^pk_3^q$ which ''couple'' the motions along the x^3-axis ($\kappa=0$) and in the x^1–x^2-plane ($k_3=0$). It is those interaction terms we are after which cannot be obtained from the decoupled results (54)–(57).

Upon insertion of (50), (52) into (26) one finds for small t ($j=1,2$),

$$F_j^+(t,\mathbf{k})=\tilde{V}_j^+(t_0,\mathbf{k})t\{1-\tfrac{1}{2}\kappa^2t-\tfrac{1}{6}k_3^2t^2+\mathcal{O}(\kappa^4t^2)+\mathcal{O}(\kappa^2k_3^2t^3)\}. \tag{58}$$

This is the correct expansion of (45). Here, the lowest order interaction term is proportional to $\kappa^2k_3^2t^3$. For times $t<\min(\kappa^{-2},k_3^{-1})$ one can neglect all interaction terms, but even then (58) is not simply a superposition of (54) and (56).

The short-time limit of F_j^- can be treated in the same way and one gets

$$F_j^-(t,\mathbf{k})=\tilde{V}_j^-(t_0,\mathbf{k})\{1+\mathscr{D}_\kappa^-(t)+\mathscr{D}_{k_3}^-(t)+\mathscr{D}_I^-(t)\} \tag{59}$$

with

$$\mathscr{D}_{\kappa}^{-}(t)=\ln(t)[-\kappa^2 t+\tfrac{1}{2}\kappa^4 t^2+\mathscr{O}(\kappa^6 t^3)]+\kappa^2 t-\tfrac{5}{4}\kappa^4 t^2+\mathscr{O}(\kappa^6 t^3),$$

$$\mathscr{D}_{k_3}^{-}(t)=-\tfrac{1}{2}k_3^2 t^2+\mathscr{O}(k_3^4 t^4), \quad \mathscr{D}_{I}^{-}(t)=\tfrac{1}{6}\kappa^2 k_3^2 t^3[\ln(t)+\mathscr{O}(1)]+\cdots .$$

$\mathscr{D}_{I}^{-}$ denotes the interaction term while $\mathscr{D}_{\kappa}^{-}$, $\mathscr{D}_{k_3}^{-}$ pertain to the special cases of propagation in the x^1–x^2-plane and along the x^3-axis. Since $\mathscr{D}_{\kappa}^{-}$, $\mathscr{D}_{k_3}^{-}$, and $\mathscr{D}_{I}^{-}$ are real, our result (59) is not the small-t expansion of solution (46) but actually belongs to

$$F_j^{-}(t,\mathbf{k})=\tfrac{1}{2}[\Gamma(1-\eta)W_{\eta,1/2}(z)+\Gamma(1+\eta)W_{-\eta,1/2}(-z)]+2\eta[\ln(\kappa)-\psi(1)]M_{\eta,1/2}(z)$$

with $z=-2ik_3 t$ and Γ the gamma and ψ the digamma function. This solution together with (45) represent another possible fundamental system of solutions of Eq. (44).

C. Axially symmetrical Kasner universe

Metrics of the Kasner-type are vacuum solutions of Einstein's field equations.[10] They are special cases of BS with coefficients $\alpha_j=t^{q_j}$ obeying constraints $\Sigma_{j=1}^{3}q_j=1=\Sigma_{j=1}^{3}q_j^2$. The axisymmetric Kasner solution is given by $q_1=2/3=q_2$, $q_3=-1/3$. The differential equations describing the behavior of transversal Maxwell fields in this model read according to (20),

$$\{\partial_t^2+\Omega_j^2(t)\}F_j(t,\mathbf{k})=0 \quad (j=1,2), \quad F_0(t,\mathbf{k})=0=F_3(t,\mathbf{k}) \tag{60}$$

with $\Omega_j^2(t)=p^2(t)+(1-4q_j^2)/4t^2$ and $p(t)=\kappa t^{-2/3}\sqrt{1+at^2}$. There are no nontrivial analytic solutions at disposal but it is not difficult to determine the long- and short-time limits using the results derived in Sec. III. The calculation proceeds along the lines of Sec. IV B. Since

$$\int p(\tau)d\tau=\frac{3}{4}|k_3|\tau^{4/3}\left\{1-\frac{1}{a\tau^2}+\mathscr{O}\left(\frac{1}{a^2\tau^4}\right)\right\},$$

$$\partial_x \ln \beta_j=\frac{2q_j-\frac{1}{3}}{x}-\frac{ax}{1+ax^2},$$

one obtains in the long-time limit with (25)

$$\frac{1}{h_j^{\pm}(t)}=\pm\frac{i\left(\frac{2}{3}-q_j\right)}{|k_3|}t^{-4/3}\left\{1-\frac{2i}{3|k_3|}t^{-4/3}+\mathscr{O}(t^{-2})\right\}\mp\frac{i\kappa^2}{2|k_3|^3}t^{-10/3}+\cdots ,$$

where $k_3\neq 0$.[13] Upon insertion of

$$\int_{t_0}^{t}dx\,\frac{p(x)}{h_j^{\pm}(x)}=\pm i\left(\frac{2}{3}-q_j\right)\left\{\ln\frac{t}{t_0}+\cdots\right\}\pm\frac{i}{4}\frac{\kappa^2}{k_3^2}[t^{-2}-t_0^{-2}]+\cdots$$

into the large-$|h_j^{\pm}|$-expansion of (21) one arrives at the desired long-time limit

$$F_j^{\pm}(t,\mathbf{k})=\epsilon_j(\mathbf{k})V^{\pm}(t_0,\mathbf{k})t^{-1/6}\exp\left\{\pm\frac{3i}{4}|k_3|t^{4/3}\left(1-\frac{\kappa^2}{k_3^2t^2}+\mathscr{O}(t^{-10/3})\right)\right\}, \tag{61}$$

where according to (33) the components ϵ_j of the polarization vectors satisfy $\epsilon_1=k_2$, $\epsilon_2=-k_1$, $\epsilon_3=0$.

Starting point for the determination of the short-time limit are Eqs. (26), (27), where $I_j^{+}(x)=\frac{3}{4}x^{4/3}$, $I_j^{-}(x)=-\frac{3}{2}\kappa^2x^{-2/3}(1+\frac{1}{2}ax^2)$. Equations (31) and (32) lead to

$$h_j^{+}(x)=\frac{3i}{2}\kappa x^{1/3}\sqrt{1+ax^2}, \tag{62}$$

$$h_j^-(x) = -3i\kappa x^{1/3}\frac{1-\frac{1}{2}ax^2}{\sqrt{1+ax^2}}, \tag{63}$$

and via (29) to

$$\lambda_{j,1}^+(x) = \tfrac{2}{3} - \tfrac{1}{3}ax^2 + \mathcal{O}(a^2x^4), \tag{64}$$

$$\lambda_{j,1}^-(x) = -\tfrac{2}{3}\{\ln(x) + \cdots + 5ax^2\ \ln(x) + \mathcal{O}(ax^2)\}. \tag{65}$$

Again, terms of lower order in x than those being kept have been neglected in (65). Because these terms do not contribute to the interaction terms they are not of interest here. From (26) together with (62), (64) follows in the limit $t \to 0 (j = 1,2)$,

$$F_j^+(t,\mathbf{k}) = \epsilon_j(\mathbf{k})\tilde{V}^+(t_0,\mathbf{k})t^{7/6}\{1 + \mathscr{D}_\kappa^+(t) + \mathscr{D}_{k_3}^+(t) + \mathscr{D}_I^+(t)\}, \tag{66}$$

with

$$\mathscr{D}_\kappa^+(t) = -\frac{1}{3}\left(\frac{3\kappa}{2}\right)^2 t^{2/3} + \frac{1}{24}\left(\frac{3\kappa}{2}\right)^4 t^{4/3} + \mathcal{O}(\kappa^6 t^2),$$

$$\mathscr{D}_{k_3}^+(t) = -\frac{1}{6}\left(\frac{3k_3}{4}\right)^2 t^{8/3} + \mathcal{O}(k_3^4 t^{16/3}), \quad \mathscr{D}_I^+(t) = \mathcal{O}(\kappa^2 k_3^2 t^{10/3}). \tag{67}$$

Mode F_j^- given by (27) together with (63) and (65) can be expanded as

$$F_j^-(t,\mathbf{k}) = \epsilon_j(\mathbf{k})\tilde{V}^-(t_0,\mathbf{k})t^{-1/6}\{1 + \mathscr{D}_\kappa^-(t) + \mathscr{D}_{k_3}^-(t) + \mathscr{D}_I^-(t)\} \tag{68}$$

(with $j = 1,2$) where

$$\mathscr{D}_\kappa^-(t) = \left(\frac{3\kappa}{2}\right)^2 t^{2/3} + \frac{3}{4}\left(\frac{3\kappa}{2}\right)^4 t^{4/3} + \mathcal{O}(\kappa^6 t^2) - \ln(t)\left\{\frac{1}{3}\left(\frac{3\kappa}{2}\right)^4 t^{4/3} + \mathcal{O}(\kappa^6 t^2)\right\},$$

$$\mathscr{D}_{k_3}^-(t) = -\frac{1}{2}\left(\frac{3k_3}{4}\right)^2 t^{8/3} + \mathcal{O}(k_3^4 t^{16/3}), \tag{69}$$

$$\mathscr{D}_I^-(t) = -\frac{27}{16}\kappa^2 k_3^2 t^{10/3}[\ln(t) + \mathcal{O}(1)],$$

and as before $\epsilon_1 = k_2$, $\epsilon_2 = -k_1$, $\epsilon_3 = 0$ in (68) and (66) as well. It turns out that $\mathscr{D}_\kappa^\pm$ and $\mathscr{D}_{k_3}^\pm$ can be exactly determined and one finds eventually

$$\mathscr{D}_\kappa^+(t) = \frac{8}{9\kappa^2}\, t^{-2/3} J_2(3\kappa t^{1/3}) - 1,$$

$$\mathscr{D}_{k_3}^+(t) = \sqrt{\frac{2\pi}{3k_3}}\, t^{-2/3} J_{1/2}\left(\frac{3}{4}\, k_3 t^{4/3}\right) - 1,$$

$$\mathscr{D}_\kappa^-(t) = -\frac{9\kappa^2}{4}\, t^{2/3}\left\{\pi Y_2(3\kappa t^{1/3}) - 2\left[\ln\frac{3\kappa}{2} - \psi(1)\right] J_2(3\kappa t^{1/3})\right\} - 1, \tag{70}$$

$$\mathscr{D}_{k_3}^-(t) = \sqrt{\frac{3\pi k_3}{8}}\, t^{2/3} J_{-(1/2)}\left(\frac{3}{4}\, k_3 t^{4/3}\right) - 1.$$

V. CONCLUSION

In this paper we studied the behavior of Maxwell fields in a Bianchi-type I background. Starting with the frame-independent formulation (1), (2) we derived in Sec. II Maxwell's equations and the Lorenz condition with respect to an anholonomic orthonormal frame. In Sec. III we specialized in the sourcefree case and to solutions analogous to plane waves in flat space–time. We found that the spectral components $f_j(t,\mathbf{k})\equiv|g(t)|^{-1/4}F_j(t,\mathbf{k})$ of the spatial Maxwell fields A_j as given in (12) must be perpendicular to the time-dependent 3-momentum $\mathbf{p}$ *and* also to the change $d\mathbf{p}$. As a consequence, the structure of the system of differential equations (20) is in fact complicated since the spatial components F_j are coupled to each other on account of these transversality conditions. However, the coupling becomes trivial in all axisymmetric space–times owing to (33).

We explicitly considered-first the special case of propagation along coordinate axes which has the obvious advantage that the second transversality constraint (19) becomes equivalent to the first one, Eq. (18). This case has also been investigated by Goorjian[3] who treated Maxwell fields with only one nonvanishing component A_j. He obtained a second order partial differential equation of the Euler–Poisson–Darboux-type with solutions which have the structure of decaying standing waves. In contrast, our approach yields decaying traveling waves. This outcome is compatible with the results of Sagnotti and Zwiebach[4] (see also the Appendix) who considered $F_{\mu\nu}$ and the corresponding *E*- and *B*-fields. In the second case we treated a so-called stiff fluid space–time and determined exact general solutions. The corresponding Faraday tensor agrees with the result found earlier by Lotze.[5] Furthermore, we showed that the formalism developed in Sec. III reproduces the correct short- and long-time asymptotics of those exact solutions. In the third model which deals with Maxwell fields in the axially symmetrical Kasner universe we determined the long- and short-time behavior of the two linearly independent solutions.

In all cases we found that the nonvanishing spectral components $f_j^{\pm}$ behave like

$$f_j^{\pm}(t,\mathbf{k})\sim t^{\pm q_j} \tag{71}$$

when t tends to zero. For $t\to\infty$ we showed that the $f_j^{\pm}$ behave like traveling waves with decaying amplitudes, i.e.,

$$|f_j^{\pm}(t,\mathbf{k})|\sim t^{-q_j}. \tag{72}$$

In the coordinate frame $\{dx^{\mu}\}_{\mu}$ one gets instead of (71) $f_j^{\pm}\sim t^{q_j\pm q_j}$ and (72) must be replaced by $|f_j^{\pm}|\sim$ const.[14]

ACKNOWLEDGMENTS

I am grateful to K.-H. Lotze and A. Wipf for stimulating discussions and helpful comments.

APPENDIX: FARADAY TENSOR IN BIANCHI-TYPE I SPACE–TIMES

In order to determine the components of the Faraday tensor in a BS one conveniently starts with field equations (8) and (9). Upon insertion of the ansatz,

$$F_{\mu\nu}(t,\mathbf{x})=\int d^3k f_{\mu\nu}(t,\mathbf{k})e^{i\mathbf{kx}} \tag{A1}$$

into (8) ($J_{\mu}\equiv 0$) one obtains the orthogonality condition $\mathbf{k}\mathbf{\Lambda}=0$. The components of $\mathbf{\Lambda}$ are given by $\Lambda_j=\sqrt{|g|}f_{j0}/\alpha_j$. (A1) together with (9) yields

$$\sum_{l=1}^{3} p_l f_{jl}=-i\,\frac{\alpha_j}{\sqrt{|g|}}\,\dot{\Lambda}_j, \tag{A2}$$

where $p_l=k_l/\alpha_l$. From the second part (4) of Maxwell's equations follows

$$\partial_t(\alpha_m\alpha_n f_{mn})+i\alpha_m\alpha_n(p_m f_{n0}-p_n f_{m0})=0 \tag{A3}$$

and a second orthogonality condition $\mathbf{k\Gamma}=0$ with $\Gamma_j=\sqrt{|g|}\epsilon_{jlm}f_{lm}/\alpha_j$. Note that Λ_j and Γ_j are in flat space–time and just the spatial Fourier transforms of the electric and magnetic field components, respectively. It is useful to introduce a vector $\mathbf{D}^{(\xi)}$ with components

$$D_j^{(\xi)}=\alpha_j^2\Delta_j^{(\xi)},\quad \Delta_j^{(\xi)}=\Gamma_j+i\xi\Lambda_j\quad(\xi=\pm 1), \tag{A4}$$

by use of which Eqs. (A2) and (A3) can be combined to

$$\sqrt{|g|}\dot{\Delta}_j^{(\xi)}+\xi(\mathbf{k}\times\mathbf{D}^{(\xi)})_j=0,\quad \mathbf{k\Delta}^{(\xi)}=0. \tag{A5}$$

If one projects (A5) onto the respective axes of the standard dreibein consisting of $\mathbf{w}_1=(1/k\kappa)\times(k_1k_3,k_2k_3,-\kappa^2)^T$, $\mathbf{w}_2=(1/\kappa)(-k_2,k_1,0)^T$, $\mathbf{w}_3=(1/k)\mathbf{k}$ and then uses $\mathbf{1}\equiv\Sigma_j\mathbf{w}_j\mathbf{w}_j^T$ one recovers Eqs. (1.9)–(1.11) of Ref. 4.

We will now determine $F_{\mu\nu}$ for the stiff-fluid model treated in Sec. IV B. On using (12) and comparing the outcome with (A1) one obtains $f_{0j}=(1/\alpha_j)\partial_t(\alpha_jf_j)$, $f_{mn}=i(p_mf_n-p_nf_m)$, and by use of (47) one gets

$$\begin{aligned}
&f_{12}=-i\,\frac{\kappa^2}{k_2}\,\frac{F_1}{t}=-f_{21},\quad f_{01}=\frac{1}{\sqrt{t}}\,\dot{F}_1=-f_{10},\\
&f_{13}=-ik_3\,\frac{F_1}{\sqrt{t}}=-f_{31},\quad f_{02}=-\frac{1}{\sqrt{t}}\,\frac{k_1}{k_2}\,\dot{F}_1=-f_{20},\\
&f_{23}=i\,\frac{k_1k_3}{k_2}\,\frac{F_1}{\sqrt{t}}=-f_{32},\quad f_{\mu\nu}=0\ \text{ otherwise.}
\end{aligned} \tag{A6}$$

It is then easy to infer from (A4) with the help of (A6),

$$\Delta_1^{(\xi)}\equiv\frac{\sqrt{|g|}}{\alpha_1^2}\,(f_{23}+i\xi f_{10})=i\left(\frac{k_1k_3}{k_2}-\xi\partial_t\right)F_1,$$

$$\Delta_2^{(\xi)}=i\left(k_3+\xi\,\frac{k_1}{k_2}\,\partial_t\right)F_1,\quad \Delta_3^{(\xi)}=-i\,\frac{\kappa^2}{k_2}\,F_1,$$

and hence[5] $\mathbf{w}_1\mathbf{\Delta}^{(\xi)}=i(k\kappa/k_2)F_1$, $\mathbf{w}_2\mathbf{\Delta}^{(\xi)}=i\xi(\kappa/k_2)\dot{F}_1$.

[1] An overview is given, e.g., in D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Deutscher Verlag d. Wiss., Berlin, 1980).
[2] E. M. Lifshitz and I. M. Khalatnikov, Usp. Fiz. Nauk **80**, 391 (1963) [Sov. Phys. Usp. **13**, 745 (1964)].
[3] P. Goorjian, Phys. Rev. D **12**, 2978 (1975).
[4] A. Sagnotti and B. Zwiebach, Phys. Rev. D **24**, 305 (1981).
[5] K.-H. Lotze, Class. Quantum Grav. **7**, 2145 (1990).
[6] An exception to the rule can be found in N. D. Birrell, P. C. W. Davies, and L. H. Ford, J. Phys. A **13**, 961 (1980).
[7] D. R. Brill and J. A. Wheeler, Rev. Mod. Phys. **29**, 465 (1957).
[8] R. U. Sexl and H. K. Urbantke, *Gravitation und Kosmologie*, 3rd ed. (BI-Wiss., Mannheim, 1987).
[9] F. W. Hehl, J. Lemke, and E. W. Mielke, in *Geometry and Theoretical Physics*, edited by J. Debrus and A. C. Hirshfeld (Springer, Berlin, 1991), p. 101.
[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973); N. Straumann, *General Relativity and Relativistic Astrophysics* (Springer, Berlin, 1984).
[11] Since $\mathbf{k}d\mathbf{x}=\mathbf{p}d\mathbf{\Theta}$, transition from the coordinate to the orthonormal frame implies the replacement $\mathbf{k}\to\mathbf{p}$.
[12] M. Abramowitz and I. A. Stegun, *Pocketbook of Mathematical Functions* (Deutsch, Frankfurt, 1984).
[13] The case $k_3=0$ is exactly soluble, see Sec. IV A.
[14] In the case of propagation along a single coordinate axis this is true only for propagation along that axis with the smallest Hubble constant.

Determinant formula for the six-vertex model with reflecting end

Osamu Tsuchiya[a)]
Institute of Physics, University of Tokyo, Komaba, Meguro-ku, Tokyo 153, Japan

(Received 29 April 1998; accepted for publication 13 May 1998)

Using the quantum inverse scattering method for the XXZ model with open boundary conditions, we obtained the determinant formula for the six-vertex model with reflecting end. © *1998 American Institute of Physics.* [S0022-2488(98)01511-4]

I. INTRODUCTION

Determinant representations of correlation functions of one-dimensional quantum integrable models were studied extensively.[1] In particular, Essler *et al.* studied the correlation functions of the XXZ model with periodic boundary conditions.[2,3] In the calculation of scalar products of the XXZ model, the relation between the scalar products of the model and the partition function of the six-vertex with domain wall boundary conditions is essential.[4] Izergin *et al.* used this relation to obtain the determinant formula for the six-vertex model with domain wall boundary conditions.[5]

On the other hand, integrable models with open boundary conditions are interesting. Integrable models with open boundary conditions are related to the B_N type Weyl group[6] and in condenced matter physics they are related to the impurity problems in the Luttinger liquid.[7]

Correlation function of the $1+1$ dimensional delta function interacting Bose gas with open boundary conditions were studied in Ref. 8.

For the XXZ model with open boundary conditions, the algebraic Bethe ansatz of the model can be used to calculate the six-vertex model with a reflecting boundary condition at one boundary and domain wall boundary conditions at the other three boundaries. (In this paper we call a six-vertex model with these types of boundary conditions *six-vertex model with reflecting end*.) In this paper, we obtain the determinant formula for the above six-vertex model using the above relation.

The paper is organized as follows: In Sec. II we give a brief review of the algebraic Bethe ansatz for the open XXZ model. An inhomogeneous six-vertex model with a reflecting end is studied in Sec. III. The determinant formula for the partition function of the six-vertex model with a reflecting end are given in Sec. IV.

II. QUANTUM INVERSE SCATTERING METHOD FOR THE OPEN XXZ MODEL

Let us start with a brief review of the quantum inverse scattering method for the XXZ model with open boundary conditions.[9] The trigonometrical solution of the Yang–Baxter equation is given by

$$R_{00'}(\lambda)=\frac{1}{2}\left\{\left[\frac{\sinh(\lambda-2i\eta)-\sinh(2i\eta)}{\sinh(\lambda)}\right]1_0\otimes 1_{0'}+\sigma_0^x\otimes\sigma_{0'}^x+\sigma_0^y\otimes\sigma_{0'}^y\right.$$
$$\left.+\left[\frac{\sinh(\lambda-2i\eta)+\sinh(2i\eta)}{\sinh(\lambda)}\right]\sigma_0^z\otimes\sigma_{0'}^z\right\}, \tag{1}$$

where $R_{00'}(\mu)$ acts as a linear operator on the tensor product of two auxiliary spaces. The L operators are given by $L_n(\lambda)=\sinh(\lambda)R_{0n}(\lambda)P_{0n}$ where 0 denotes the auxiliary space, n denotes the nth quantum space, and $P_{0n}=(1/2)(1_0\otimes 1_n+\sigma_0^x\otimes\sigma_n^x+\sigma_0^y\otimes\sigma_n^y+\sigma_0^z\otimes\sigma_n^z)$ is a permutation operator. Thus

a)Electronic mail: otutiya@hep1.c.u-tokyo.ac.jp

 © 1998 American Institute of Physics

$$L_n(\lambda)=\tfrac{1}{2}\{[\sinh(\lambda-2i\eta)+\sinh(\lambda)]1_0\otimes 1_n-\sinh(2i\eta)[\sigma_0^x\otimes\sigma_n^x+\sigma_0^y\otimes\sigma_n^y]$$
$$+[\sinh(\lambda-2i\eta)-\sinh(\lambda)]\sigma_0^z\otimes\sigma_n^z\}. \tag{2}$$

As a consequence of the Yang–Baxter equation, the L operators satisfy the fundamental commutation relations

$$R(\lambda-\mu)(L_n(\lambda)\otimes L_n(\mu))=(L_n(\mu)\otimes L_n(\lambda))R(\lambda-\mu). \tag{3}$$

The monodromy matrix of the inhomogeneous XXZ model is constructed from the L operators as

$$T(\lambda,\boldsymbol{\omega})=L_L(\lambda-\omega_L)L_{L-1}(\lambda-\omega_{L-1})\cdots L_1(\lambda-\omega_1). \tag{4}$$

This matrix satisfies fundamental commutation relations similar to (4)

$$R(\lambda-\mu)(T(\lambda,\boldsymbol{\omega})\otimes T(\mu,\boldsymbol{\omega}))=(T(\mu,\boldsymbol{\omega})\otimes T(\lambda,\boldsymbol{\omega}))R(\lambda-\mu). \tag{5}$$

For the model with open boundary conditions, in addition to the Yang–Baxter equation, the reflection equation[9] plays an important role,

$$R_{12}(\lambda-\mu)K_1(\lambda)R_{12}(\lambda+\mu)K_1(\mu)=K_1(\mu)R_{12}(\lambda+\mu)K_1(\lambda)R_{12}(\lambda-\mu), \tag{6}$$

where $K_j(\lambda)$ is the 2×2 matrix which acts to the jth auxiliary space and is related to the reflection of the particle with rapidity λ at the boundaries. In the present paper, we concentrate to the diagonal solution of the reflection equation

$$K(\lambda,\xi)=\frac{1}{\sinh(\xi)}\begin{pmatrix}\sinh(\xi+\lambda) & 0\\ 0 & \sinh(\xi-\lambda)\end{pmatrix}. \tag{7}$$

For the model with open boundary conditions, Sklyanin's monodromy matrix

$$U(\lambda,\boldsymbol{\omega})=T(\lambda,\boldsymbol{\omega})K(\lambda,\xi_-)\grave{T}(\lambda,\boldsymbol{\omega})=\begin{pmatrix}\mathscr{A}(\lambda,\boldsymbol{\omega}) & \mathscr{B}(\lambda,\boldsymbol{\omega})\\ \mathscr{C}(\lambda,\boldsymbol{\omega}) & \mathscr{D}(\lambda,\boldsymbol{\omega})\end{pmatrix}, \tag{8}$$

where

$$\grave{T}(\lambda,\boldsymbol{\omega})=L_1(\lambda+\omega_1)L_2(\lambda+\omega_2)\cdots L_L(\lambda+\omega_L)\propto T^{-1}(-\lambda,\boldsymbol{\omega}) \tag{9}$$

is fundamental, where $\mathscr{A}(\lambda,\boldsymbol{\omega})$, $\mathscr{B}(\lambda,\boldsymbol{\omega})$, $\mathscr{C}(\lambda,\boldsymbol{\omega})$, $\mathscr{D}(\lambda,\boldsymbol{\omega})$ are operators which act on the quantum spaces.

$U(\lambda,\boldsymbol{\omega})$ also satisfies the reflection equation

$$R_{12}(\lambda-\mu)U_1(\lambda,\boldsymbol{\omega})R_{12}(\lambda+\mu)U_1(\mu,\boldsymbol{\omega})=U_1(\mu,\boldsymbol{\omega})R_{12}(\lambda+\mu)U_1(\lambda,\boldsymbol{\omega})R_{12}(\lambda-\mu). \tag{10}$$

The reflection equation is rewritten as the commutation relations of the matrix elements $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$,

$$[\mathscr{A}(\lambda),\mathscr{A}(\mu)]=-[\mathscr{D}(\lambda),\mathscr{D}(\mu)]=\frac{\sinh(2i\eta)}{\sinh(\lambda+\mu-2i\eta)}\{\mathscr{B}(\lambda)\mathscr{C}(\mu)-\mathscr{B}(\mu)\mathscr{C}(\lambda)\}$$
$$=-\frac{\sinh(2i\eta)}{\sinh(\lambda+\mu-2i\eta)}\{\mathscr{C}(\lambda).\mathscr{B}(\mu)-\mathscr{C}(\mu).\mathscr{B}(\lambda)\}$$
$$=\frac{\sinh(\lambda-\mu)\sinh(\lambda+\mu)}{\sinh^2(2i\eta)-\sinh(\lambda+\mu-2i\eta)}[\mathscr{A}(\lambda),\mathscr{D}(\mu)], \tag{11}$$

$$[\mathscr{B}(\lambda),\mathscr{B}(\mu)]=[\mathscr{C}(\lambda),\mathscr{C}(\mu)]=0, \tag{12}$$

$$[\mathscr{B}(\lambda),\mathscr{C}(\mu)]=\frac{\sinh(2i\eta)}{\sinh(\lambda-\mu)\sinh(\lambda+\mu-2i\eta)}$$
$$\times\{\sinh(\lambda+\mu)[\mathscr{D}(\lambda).\mathscr{A}(\mu)-\mathscr{D}(\mu).\mathscr{A}(\lambda)]$$
$$+\sinh(\lambda-\mu)[.\mathscr{A}(\lambda).\mathscr{A}(\mu)-\mathscr{D}(\mu)\mathscr{D}(\lambda)]\}. \tag{13}$$

From the above commutation relations (13) Sklyanin's transfer matrices

$$\tau(\lambda,\boldsymbol{\omega})=\mathrm{tr}_0(K(-\lambda+2i\eta)U(\lambda,\boldsymbol{\omega}))$$
$$=(1/\sinh(\xi_+))[\sinh(\xi_+-\lambda+2i\eta).\mathscr{A}(\lambda,\boldsymbol{\omega})+\sinh(\xi_++\lambda-2i\eta)\mathscr{D}(\lambda,\boldsymbol{\omega})] \tag{14}$$

are commutative for any values of spectral parameter $[\tau(\lambda),\tau(\mu)]$.

Consequently, $\tau(\lambda,\boldsymbol{\omega})$ generate the infinite number of conserved quantities. For example the Hamiltonian is given by

$$\frac{d}{d\lambda}\,\tau(\lambda,0)|_{\lambda=0}=-\frac{1}{\sinh(2i\eta)}\sum_{n=1}^{L-1}\{\sigma_n^x\sigma_{n+1}^x+\sigma_n^y\sigma_{n+1}^y+\cosh(2i\eta)\sigma_n^z\sigma_{n+1}^z\}$$
$$+\coth(\xi_-)\sigma_1^z-\coth(\xi_+)\sigma_L^z+\text{const.} \tag{15}$$

In the algebraic Bethe ansatz, we construct the simultaneous eigenstates of the conserved quantities by applying the nondiagonal elements (creation operator) of the monodromy matrix to the pseudovacuum. We choose as a pseudovacuum the state whose spins are all up

$$|0\rangle=\otimes_{j=1}^{L}|+\rangle_j. \tag{16}$$

This pseudovacuum is the eigenstate of $\mathscr{A}$ and $\mathscr{D}$ and is annihilated by $\mathscr{C}$

$$\mathscr{A}(\lambda,\boldsymbol{\omega})|0\rangle=a(\lambda,\boldsymbol{\omega})|0\rangle,\quad \mathscr{D}(\lambda,\boldsymbol{\omega})|0\rangle=d(\lambda,\boldsymbol{\omega})|0\rangle,\quad \mathscr{C}(\lambda,\boldsymbol{\omega})|0\rangle=0, \tag{17}$$

where

$$a(\lambda,\boldsymbol{\omega})=\sinh(\xi_-+\lambda)\prod_{n=1}^{L}\{\sinh(\lambda-\omega_n-2i\eta)\sinh(\lambda+\omega_n-2i\eta)\},$$
$$d(\lambda)=\sinh(\xi_--\lambda)\prod_{n=1}^{L}\{\sinh(\lambda-\omega_n)\sinh(\lambda+\omega_n)\}. \tag{18}$$

In other words $\mathscr{B}$ has the role of the creation operator and the eigenstates of the Hamiltonian are of the form $\Pi_{n=1}^{N}.\mathscr{B}(\lambda_n,\boldsymbol{\omega})|0\rangle$ if rapidities λ_n satisfy the Bethe ansatz equations.

In the algebraic Bethe ansatz eigenstates and dual states of eigenstates are, respectively, given by

$$\mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle \tag{19}$$

and

$$\langle 0|\mathscr{C}(\mu_1)\cdots\mathscr{C}(\mu_L). \tag{20}$$

III. INHOMOGENEOUS SIX-VERTEX MODEL WITH REFLECTING END

For the scalar products between the eigenstates the following relation holds

$$\langle 0|\mathscr{C}(\mu_1)\cdots\mathscr{C}(\mu_L).\mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle=\langle 0|\mathscr{C}(\mu_1)\cdots\mathscr{C}(\mu_L)|\overline{0}\rangle\langle\overline{0}|.\mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle, \tag{21}$$

where $|\overline{0}\rangle$ is the quantum state with all spins down $\Pi_{n=1}^{L}|-\rangle$. Note that

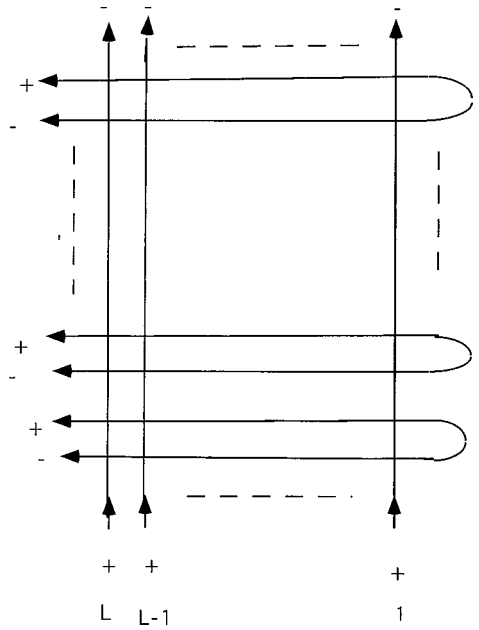

FIG. 1. A six-vertex model with reflecting end: Quantum inverse scattering method formulation. Boundary conditions are all spins: + for the bottom of the lattice, all spins − for the top of the lattice, and + alternate with − for the left of the latlice. The right boundary corresponds to diagonal boundary K matrices.

$$Z_L(\{\lambda\},\{\omega\})=\langle\overline{0}|\mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle, \tag{22}$$

is nothing but the partition function of the inhomogeneous six-vertex model on a $L\times 2L$ lattice with domain wall boundary condition for the lower, upper, and left boundaries and reflecting end for the right boundary (see Fig. 1). $\mathscr{B}(\lambda_n)={}_n^0(+|U(\lambda,\omega)|-)_n^0$ where $|+)_n^0$ and $|-)_n^0$ are the nth auxiliary states with spins + and −, respectively, $[{}_n^0(+|$ and ${}_{n(-|}^0$ are dual states with spins + and −, respectively]. Here we assign the auxiliary space to each $B(\lambda_n)$.

Now we derive the recursion relations to the partition function $Z_L(\{\lambda\},\{\omega\})$.

If the inhomogenuity ω_L is choosen to be $\omega_L=-\lambda_1$ then the partition function of the $L\times 2L$ lattice is related to the partiton function of the $L-1\times 2(L-1)$ lattice.

The L operators are regular, i.e.,

$$L_L(0)=-\sinh(2i\eta)P_{0L}\,. \tag{23}$$

Thus

$$\begin{aligned}\mathscr{B}(\lambda_1)\prod_{j=1}^{L}|+\rangle_j\Big|_{\omega_L=-\lambda_1}&=-\sinh(2i\eta)\,{}_1^0(+|L_L(\lambda_1-\omega_L)\cdots L_{L-1}(\lambda_1+\omega_{L-1})P_{0L}|-)_1^0\prod_{j=1}^{L}|+\rangle_j\\&=-\sinh(2i\eta)\,{}_1^0(+|L_L(\lambda_1-\omega_L)\cdots L_{L-1}(\lambda_1+\omega_{L-1})|+)_1^0|-\rangle_L\prod_{j=1}^{L-1}|+\rangle_j\,.\end{aligned} \tag{24}$$

Noting that $|+)_1|+\rangle_n$ is the eigenstate of $L_n(\lambda_1)$

$$L_n(\lambda_1)|+)_1|+\rangle_n=\sinh(\lambda_1-2i\eta)|+)_1|+\rangle_n\,, \tag{25}$$

$$K(\lambda_1,\xi_-)|+)_1=\frac{\sinh(\xi_-+\lambda_1)}{\sinh(\xi_-)}|+)_1\,, \tag{26}$$

and

$$ {}_1\langle +|L_L(\lambda_n-\omega_L)|-\rangle_L = 2\sinh(\lambda_n-\omega_L)\,{}_1\langle +\|-\rangle_L . \tag{27} $$

We obtain

$$ \begin{aligned} \mathscr{B}(\lambda_1)\prod_{j=1}^{L}|+\rangle_j\Big|_{\omega_L=-\lambda_1} &= -\frac{\sinh(\xi_1+\lambda_1)}{\sinh(\xi_-)}\sinh(2i\eta)\sinh(2\lambda_1) \\ &\quad\times\prod_{j=1}^{L-1}\sinh(\lambda_1-\omega_j-2i\eta)\sinh(\lambda_1+\omega_j-2i\eta) \\ &\quad\times|-\rangle_L\prod_{j=1}^{L-1}|+\rangle_j . \end{aligned} \tag{28} $$

Using

$$ L_L(\lambda_n+\omega_L)|-)_n|-\rangle_L = \sinh(\lambda_n+\omega_L-2i\eta)|-)_n|-\rangle_L , \tag{29} $$

together with Eq. (30) and the locality of the L operators we find the relation

$$ \mathscr{B}(\lambda_n)|-\rangle_L = \sinh(\lambda_n+\omega_L-2i\eta)\sinh(\lambda_n-\omega_L)|-\rangle_L\mathscr{B}_{L-1}(\lambda_n), \tag{30} $$

where

$$ \begin{aligned} \mathscr{B}_{L-1}(\lambda_n) &= {}_n\langle +|L_{L-1}(\lambda_n-\omega_{L-1})\cdots L_1(\lambda_n-\omega_1)K(\lambda_n,\xi_-) \\ &\quad\times L_1(\lambda_n+\omega_1)\cdots L_{L-1}(\lambda_n+\omega_{L-1})|-)_n . \end{aligned} \tag{31} $$

Lemma 1-a: The partition function of the model on the $L\times 2L$ lattice is related to the one on the $L-1\times 2(L-1)$ lattice

$$ \begin{aligned} Z_L(\{\lambda\},\{\omega\})|_{\omega_L=-\lambda_1} &= -\frac{\sinh(\xi_-+\lambda_L)}{\sinh(\xi_-)}\sinh(2i\eta)\sinh(2\lambda_L) \\ &\quad\times\prod_{j=1}^{L-1}\sinh(\lambda_L-\omega_j-2i\eta)\sinh(\lambda_L+\omega_j-2i\eta) \\ &\quad\times\prod_{n=1}^{L-1}\sinh(\lambda_n+\omega_L-2i\eta)\sinh(\lambda_n-\omega_L)Z_{L-1}(\{\lambda\},\{\omega\}). \end{aligned} \tag{32} $$

Here $Z_M(\{\lambda\},\{\omega\})$ is the partition function for the model on the $M\times 2M$ lattice

$$ Z_M(\{\lambda\},\{\omega\}) = \prod_{j=1}^{M}{}_j\langle -|\prod_{n=1}^{M}\mathscr{B}_M(\lambda_n,\omega)\prod_{j=1}^{M}|+\rangle_j , \tag{33} $$

which depend on $\lambda_1\sim\lambda_M$ and $\omega_1\sim\omega_M$.

Lemma 1-b: Folowing the same line of argument as above we have other types of recursion relations

$$ \begin{aligned} Z_L(\{\lambda\},\{\omega\})|_{\omega_L=\lambda_1} &= -\frac{\sinh(\xi_--\lambda_L)}{\sinh(\xi_-)}\sinh(2i\eta)\sinh(2\lambda_L) \\ &\quad\times\prod_{j=1}^{L-1}\sinh(\lambda_L-\omega_j-2i\eta)\sinh(\lambda_L+\omega_j-2i\eta) \\ &\quad\times\prod_{n=1}^{L-1}\sinh(\lambda_n-\omega_L-2i\eta)\sinh(\lambda_n+\omega_L)Z_{L-1}(\{\lambda\},\{\omega\}). \end{aligned} \tag{34} $$

Lemma 2:

$$Z_1(\{\lambda\},\{\omega\}) = {}_1\langle \overline{0}| \mathscr{B}(\lambda)|0\rangle_1 \tag{35}$$

is explicitly calculated as

$$Z_1(\{\lambda\},\{\omega\}) = -\frac{\sinh(2i\eta)}{\sinh(\xi_-)}[\sinh(\xi+\lambda)\sinh(\lambda-\omega)+\sinh(\xi-\lambda)\sinh(\lambda+\omega)]. \tag{36}$$

Lemma 3: $\langle \overline{0}| \mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle$ depend on each λ_n as

$$\langle \overline{0}| \mathscr{B}(\lambda_L)\cdots\mathscr{B}(\lambda_1)|0\rangle = e^{-2L\lambda_n} P_{2L}(e^{2\lambda_n}), \tag{37}$$

where $P_{2L}(x)$ is the polynomial of degree $2L$.

Proof: This can be proved by straightforward calculations.

IV. DETERMINANT FORMULA FOR THE SIX-VERTEX MODEL WITH REFLECTING END

The lemmas in the last section determine the partition function of the model uniquely.

Theorem: The partition function of the six-vertex model with reflecting end is given by

$$\begin{aligned} Z_L(\{\lambda\},\{\omega\}) = & \prod_{j=1}^{L}\prod_{n=1}^{L} \sinh(\lambda_j-\omega_n-2i\eta)\sinh(\lambda_j-\omega_n)\sinh(\lambda_j+\omega_n-2i\eta)\sinh(\lambda_j+\omega_n) \\ & \times\left[\prod_{j>k}\sinh(\lambda_j-\lambda_k)\sinh(\lambda_j+\lambda_k-2i\eta)\prod_{n>m}\sinh(\omega_n-\omega_m)\sinh(\omega_n+\omega_m)\right]^{-1} \\ & \times \det M, \end{aligned} \tag{38}$$

where

$$M_{jn} = \frac{\sinh(2i\eta)}{\sinh(\xi)\sinh(\lambda_j-\omega_n-2i\eta)\sinh(\lambda_j+\omega_n-2i\eta)} \times \left\{\frac{\sinh(\xi_-+\lambda_j)}{\sinh(\lambda_j+\omega_n)}+\frac{\sinh(\xi_--\lambda_j)}{\sinh(\lambda_j-\omega_n)}\right\}. \tag{39}$$

Proof: This is proved along the same line of argument as in Ref. 5.

V. CONCLUSION

In this paper the determinant formula for the partition function of the six-vertex model with reflecting end is given. It is the determinant formula corresponding to the quantum inverse scattering method with open boundary conditions.

In the model with open boundary conditions, as opposed to the model with periodic boundary conditions, one cannot determine the scalar product from the six-vertex model with reflecting end. To calculate the scalar products and correlation functions for the open XXZ model is interesting.

ACKNOWLEDGMENT

I would like to thank to Dr. F. Nakano for comments.

[1] V. E. Korepin, N. M. Bogoliubov, and A. G. Izergin, *Quantum Inverse Scattering Method and Correlation Function* (Cambridge Monographs on Mathematical Physics, 1993).
[2] F. H. L. Essler, H. Frahm, A. G. Izergin, and V. E. Korepin, Commun. Math. Phys. **174**, 191 (1995).
[3] F. H. L. Essler, H. Frahm, A. R. Its, and V. E. Korepin, Nucl. Phys. B: Field Theory Stat. Syst. **446**, 448 (1995).
[4] V. E. Korepin, Commun. Math. Phys. **86**, 391 (1982).
[5] A. G. Izergin, D. A. Coker, and V. E. Korepin, J. Phys. A: Math. Gen. **25**, 4315 (1992).
[6] T. Yamamoto and O. Tsuchiya, J. Phys. A: Math. Gen. **29**, 3977 (1996).
[7] O. Tsuchiya and T. Yamamoto, J. Phys. Soc. Jpn. **66**, 1950 (1997).
[8] T. Kojima, ''Dynamical correlation functions for an impenetrable Bose gas with open boundary conditions,'' preprint, 1997, solv-int/9712007.
[9] E. K. Sklyanin, J. Phys. A: Math. Gen. **21**, 2375 (1988).

Transition to criticality in circle maps at the golden mean

Tony W. Dixon[a] and Brian G. Kenny[b]
Department of Physics, University of Western Australia, Western Australia 6907, Australia

(Received 3 February 1998; accepted for publication 12 August 1998)

The one-dimensional circle map is frequently used to model the quasiperiodic route to chaos. In this paper we study the transition to criticality for circle maps with a golden mean winding number in the limit as the order of the map inflection point $N=1+\epsilon$ approaches unity. The solution to the relevant pair of functional equations consists of the subcritical solution together with a perturbative correction of order ϵ. We derive functional relations that the perturbative correction must satisfy and produce numerical evidence that these equations are indeed satisfied. A similar analysis is performed for the golden mean eigenvalue equations. Power law conjugacy of the subcritical solution to order $N=1+\epsilon$ enables us to obtain a closed form solution for the perturbative correction that satisfies the constraint relations exactly. This is compared with the more interesting critical case associated with a circle map whose inflection point is infinitesimally close to unity. © *1998 American Institute of Physics.* [S0022-2488(98)03611-1]

I. INTRODUCTION

Shenker[1] has argued that the main features of the quasiperiodic route to chaos can be modeled by the sine circle map,

$$\theta_{n+1}=f(\theta_n)=\theta_n+\Omega-\frac{k}{2\pi}\sin(2\pi\theta_n).$$

In particular, he showed that Feigenbaum scaling occurred at the golden mean winding number. For $0\leqslant k<1$ the sine map demonstrates *subcritical* behavior due to the leading linear term for small θ. At $k=1$ (*critical*) the map develops a cubic point of inflection at $\theta=0$ and can demonstrate chaotic behavior.

Associated with this universal scaling at the golden mean is a universal function $g(\theta)$ and an associated universal scaling parameter $\alpha(N)$ that satisfy a pair of equivalent functional equations:

$$g(\theta)=\alpha g[g(\alpha^{-2}\theta)], \tag{1a}$$

$$g(\theta)=\alpha^2 g[\alpha^{-2}g(\alpha^{-1}\theta)], \tag{1b}$$

where we follow the notation of Feigenbaum *et al.*[2] If we set the scale so that $g(0)=1$ and assume a power series expansion of the form

$$g(\theta)=1+g_1\theta^N+g_2\theta^{2N}+\cdots, \quad \theta^N\equiv\theta|\theta|^{N-1}, \tag{2}$$

then one can determine the universal function g and its associated constant α for a critical point of order N.[3,4] Rand *et al.*[5,6] have given an equivalent formulation involving two separate analytic functions on different domains.

[a]Corresponding author. Current address: Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand. Telephone +64 3 479 7779; fax +64 3 479 8427; electronic mail: tdixon@maths.otago.ac.nz
[b]Electronic mail: chaos@cyllene.uwa.edu.au

 © 1998 American Institute of Physics

Linearizing Eqs. (1) about the fixed point $g(\theta)$ (see Ref. 2), one can obtain two equivalent eigenvalue equations for which the largest eigenvalue is the parameter scaling constant $\delta(N)$, and its associated eigenfunction is denoted as $h(\theta)$. In a similar fashion to $g(\theta)$ and α, $h(\theta)$ and δ are also universal, depending only on the order N of the point of inflection.

There is a subcritical solution of Eqs. (1) corresponding to diffeomorphisms of the circle. The equation,

$$g(\theta)=1-\frac{1}{\rho}\,\theta, \tag{3}$$

solves both equations (1a) and (1b) [that is, $N=1$ in the power series expansion (2)] with $\alpha=-1/\rho$; ρ is the golden mean $(\sqrt{5}-1)/2$.

One can generate a family of exact *noncritical* solutions by applying the power conjugacy transformation $g(\theta)\to[g(\theta^N)]^{1/N}$ to the subcritical solution

$$g(\theta)=\left(1-\frac{1}{\rho}\,\theta^N\right)^{1/N}, \quad \theta^N\equiv\theta|\theta|^{N-1}, \tag{4}$$

with associated constant

$$\alpha=-\left(\frac{1}{\rho}\right)^{1/N} \quad \text{or} \quad \alpha^N=-\frac{1}{\rho}. \tag{5}$$

We intend to compare results from these solutions with those obtained for the critical solutions.

Collet *et al.*[7] studied the solution of the period-doubling equation[8,9] in the space of analytic functions of $|x|^{1+\epsilon}$, $\epsilon\geqslant 0$. In a similar fashion, Jonker and Rand[10] used perturbation methods to show that the functional equations associated with the circle map at the golden mean have a solution on the space of analytic functions of $\theta|\theta|^{\epsilon}$, using the formulation of Rand *et al.*[5,6] The existence of a fixed point at $\epsilon=0$ $(N=1)$ greatly simplifies the proof for the golden mean compared to the period-doubling case. We may therefore consider solutions of Eqs. (1) near $N=1$.

This paper is mainly concerned with solutions of Eqs. (1) for small ϵ. We carefully examine the singular transition from subcritical $(\epsilon=0)$ to critical $(\epsilon>0)$ behavior at the golden mean. We derive new functional relations (which characterize this transition) that the first-order corrections to the universal functions g and h must satisfy. The properties of these critical solutions are investigated numerically and compared with those of the class of noncritical functions (4) above.

II. DEVIATION FROM SUBCRITICAL SOLUTION FOR $N=1+\epsilon$

Let us denote an arbitrary solution (noncritical or critical) by $g(N,\theta)$. We want to examine this solution in the case where $N\to 1$ in the infinitesimal limit. If $N=1+\epsilon$, then let us suppose that for small ϵ,

$$g(N,\theta)=g(1+\epsilon,\theta)=g(1,\theta)+\epsilon\mathcal{G}(\theta), \tag{6}$$

where $g(1,\theta)$ is given by (3). If we assume that the normalization is such that $g(N,0)=1$ for all N, then $\mathcal{G}(0)=0$. We do not *a priori* know how α depends on N; numerically we find that the deviation of $\alpha(N)$ from its subcritical value is dependent on ϵ to first order always. If we assume then that

$$\alpha\approx-\frac{1}{\rho}\,(1-c\epsilon), \tag{7a}$$

it follows that

$$\frac{1}{\alpha}\approx-\rho(1+c\epsilon) \tag{7b}$$

to leading order. We expect $c>0$ since $|\alpha|$ decreases as N increases above 1.

For the noncritical functions (4) we know $\alpha=-(1/\rho)^{1/N}$. If $N=1+\epsilon$, for small ϵ,

$$\alpha=-\left(\frac{1}{\rho}\right)^{1/(1+\epsilon)}\approx-\left(\frac{1}{\rho}\right)^{1-\epsilon}\approx-\frac{1}{\rho}\left(1-\epsilon\,\log\frac{1}{\rho}\right).$$

Equating this expression with our first-order approximation (7a) gives

$$c=\log\frac{1}{\rho}. \tag{8}$$

Arneodo and Holschneider[11] have derived the relation

$$\alpha^N(N)=\alpha(1/N), \tag{9}$$

for critical circle maps. We can use this expression to determine the critical value of c as follows. For $N=1+\epsilon$, the LHS of (9) gives

$$\alpha^N(N)\approx-\left(\frac{1}{\rho}\right)^{1+\epsilon}(1-c\epsilon)^{1+\epsilon}\approx-\frac{1}{\rho}\left(1+\epsilon\,\log\frac{1}{\rho}\right)(1-c\epsilon)[1+\epsilon\,\log(1-c\epsilon)]$$
$$\approx-\frac{1}{\rho}\left(1+\epsilon\,\log\frac{1}{\rho}-c\epsilon\right),$$

and the RHS of (9) gives

$$\alpha(1/N)=\alpha(1/(1+\epsilon))\approx\alpha(1-\epsilon)=-\frac{1}{\rho}\,(1+c\epsilon).$$

Equating and rearranging, we find

$$c=\frac{1}{2}\,\log\frac{1}{\rho}. \tag{10}$$

This is the value of the constant c governing the correction to the critical α near $N=1$, a different result to that found for the noncritical α, that is, expression (8).

We note here that by assuming an expansion of the form (2), the following conditions will hold:

$$g(N,1)=1/\alpha,\quad g(N,1/\alpha^2)=1/\alpha^2,$$

$$g'(N,1)=\alpha^{2N-1},\quad g'(N,1/\alpha^2)=\alpha^N.$$

A. Functional equations for $\mathcal{G}(\theta)$

We now proceed to derive constraint equations on $\mathcal{G}(\theta)$ from each of the functional equations (1). In this section, we use the symbol g for the subcritical solution $g(1,\theta)$ and G for the general solutions $g(N,\theta)$.

1. First functional equation

Let $\beta=-1/\rho$, that is, the subcritical value of α at $\epsilon=0$. From (1a) and (7a) we now have

$$G[G(\theta)]=\alpha^{-1}G(\alpha^2\theta), \tag{11a}$$

$$g[g(\theta)]=\beta^{-1}g(\beta^2\theta), \tag{11b}$$

$$\alpha\approx\beta(1-c\epsilon). \tag{11c}$$

To first order in ϵ, the LHS of (11a) is

$$G[G(\theta)]=g[g(\theta)]+\epsilon\mathcal{G}[g(\theta)]+\epsilon\mathcal{G}(\theta)g'[g(\theta)]$$
$$=\beta^{-1}g(\beta^2\theta)+\epsilon\{\mathcal{G}[g(\theta)]+\mathcal{G}(\theta)g'[g(\theta)]\}.$$

As $g(\theta)=1+\beta\theta$, this expression becomes

$$G[G(\theta)]=\beta^{-1}g(\beta^2\theta)+\epsilon[\mathcal{G}(1-\rho^{-1}\theta)-\rho^{-1}\mathcal{G}(\theta)].$$

Also, the RHS of (11a) gives to $O(\epsilon)$

$$\alpha^{-1}G(\alpha^2\theta)=\alpha^{-1}g(\alpha^2\theta)+\alpha^{-1}\epsilon\mathcal{G}(\alpha^2\theta)=\alpha^{-1}g(\alpha^2\theta)+\beta^{-1}\epsilon\mathcal{G}(\beta^2\theta).$$

Equating and rearranging we find

$$\epsilon[\mathcal{G}(1-\rho^{-1}\theta)-\rho^{-1}\mathcal{G}(\theta)+\rho\mathcal{G}(\rho^{-2}\theta)]$$
$$=\alpha^{-1}g(\alpha^2\theta)-\beta^{-1}g(\beta^2\theta)=\alpha^{-1}[1+\alpha^2\beta\theta]-\beta^{-1}[1+\beta^3\theta]=(\alpha^{-1}-\beta^{-1})+\beta\theta(\alpha-\beta)$$
$$=-\rho(1+c\epsilon)+\rho-\rho^{-1}\theta[\rho^{-1}-\rho^{-1}(1-c\epsilon)]=-\epsilon[c\rho+c\rho^{-2}\theta].$$

Therefore, the first-order correction term $\mathcal{G}(\theta)$ must satisfy

$$\mathcal{G}(1-\theta/\rho)-\rho^{-1}\mathcal{G}(\theta)+\rho\mathcal{G}(\rho^{-2}\theta)=-c(\rho+\rho^{-2}\theta). \tag{12}$$

2. Second functional equation

Following the same notation as above, we have, from Eq. (1b),

$$G[\alpha^{-2}G(\theta)]=\alpha^{-2}G(\alpha\theta), \tag{13a}$$

$$g[\beta^{-2}g(\theta)]=\beta^{-2}g(\beta\theta). \tag{13b}$$

To first order in ϵ, (13a) gives, for the LHS and RHS, respectively,

$$G[\alpha^{-2}G(\theta)]=g[\alpha^{-2}g(\theta)]+\epsilon\mathcal{G}[\beta^{-2}g(\theta)]+\epsilon\beta^{-2}\mathcal{G}(\theta)g'[\beta^{-2}g(\theta)],$$

$$\alpha^{-2}G(\alpha\theta)=\alpha^{-2}g(\alpha\theta)+\beta^{-2}\epsilon\mathcal{G}(\beta\theta).$$

Equating these and rearranging, we have

$$-\epsilon\{\mathcal{G}[\beta^{-2}g(\theta)]+\beta^{-2}\mathcal{G}(\theta)g'[\beta^{-2}g(\theta)]-\beta^{-2}\mathcal{G}(\beta\theta)\}=-\alpha^{-2}g(\alpha\theta)+g[\alpha^{-2}g(\theta)]. \tag{14}$$

Again using $g(\theta)=1+\beta\theta$, the quantity in square brackets on the LHS of (14) may be written as

$$\mathcal{G}(\beta^{-2}+\theta\beta^{-1})+\beta^{-1}\mathcal{G}(\theta)-\beta^{-2}\mathcal{G}(\beta\theta)=\mathcal{G}(\rho^2-\rho\theta)-\rho\mathcal{G}(\theta)-\rho^2\mathcal{G}(-\rho^{-1}\theta).$$

Also,

$$\text{RHS of (14)}=-(\alpha^{-2}+\beta\alpha^{-1}\theta)+(1+\beta\alpha^{-2}+\beta^2\alpha^{-2}\theta)$$
$$=-(\alpha^{-2}-1-\beta\alpha^{-2})-\theta(\beta\alpha^{-1}-\beta^2\alpha^{-2})$$
$$=-[-1+\alpha^{-2}(1-\beta)]-\theta[\beta\alpha^{-1}-(\beta/\alpha)^2]$$
$$=-[-1+\rho^2\rho^{-2}(1+2c\epsilon)]-\theta[(1+c\epsilon)-(1+c\epsilon)^2]$$
$$=-2c\epsilon+\theta c\epsilon=-c\epsilon(2-\theta).$$

We have, therefore, a second constraint equation that $\mathcal{G}$ must satisfy:

$$\mathcal{G}(\rho^2-\rho\theta)-\rho\mathcal{G}(\theta)-\rho^2\mathcal{G}(-\rho^{-1}\theta)=c(2-\theta). \tag{15}$$

We anticipate that the equations (12) and (15) are equivalent to the extent that the functional equations (1) are equivalent. The results presented here hold for small $\epsilon<0$ as well.

Jonker and Rand[10] have derived an equation analogous to our first equation (12), using the formulation of Rand *et al.*[5,6] Feigenbaum *et al.*[2] have also studied this transition. They derived constraint equations in what they called the ''standard form,'' where the application of a power conjugacy transformation gives a function g that behaves like $g(\theta)=a+b\theta^{\nu}+\cdots$ at $\theta=0$ and $g(\theta)=c(\theta-1)^{\mu}+\cdots$ at $\theta=1$, with $\mu=\nu=\sqrt{N}$. Our formulation corresponds to the normally encountered situation where g behaves like (2) at $\theta=0$ and is linear at its zero.

3. Evaluation of $\mathcal{G}$ and $\mathcal{G}'$ at special points

Noting again that $\mathcal{G}(0)=0$, substituting $\theta=0$ in Eq. (12) gives $\mathcal{G}(1)=-c\rho$. If we now set $\theta=\rho^2$ in Eq. (12) and use the identity $1-\rho=\rho^2$, then

$$\text{LHS of (12)}=\mathcal{G}(\rho^2)-\rho^{-1}\mathcal{G}(\rho^2)+\rho\mathcal{G}(1)=\rho[\mathcal{G}(1)-\mathcal{G}(\rho^2)],$$

$$\text{RHS of (12)}=-c(\rho+1)=-c/\rho.$$

Therefore, $\mathcal{G}(1)-\mathcal{G}(\rho^2)=-c\rho^{-2}$, or $\mathcal{G}(\rho^2)=c\rho^{-2}+\mathcal{G}(1)=c(\rho^{-2}-\rho)=2c$. Thus

$$\mathcal{G}(\rho^2)=2c, \tag{16a}$$

$$\mathcal{G}(1)=-c\rho. \tag{16b}$$

If we substitute $\theta=0$ in Eq. (15), we again obtain (16a).

It is possible to determine relations involving the derivatives of $\mathcal{G}(\theta)$ at certain points in this way too. However, it is easier to return to the source of these relations, Eq. (6). Differentiating this expression with respect to θ gives

$$g'(1+\epsilon,\theta)-g'(1,\theta)=\epsilon\mathcal{G}'(\theta). \tag{17}$$

At $\theta=1$ we have

$$\begin{aligned}\epsilon\mathcal{G}'(1)&=g'(1+\epsilon,1)-\beta\\&=\alpha^{2N-1}-\beta=-|\alpha|^{1+2\epsilon}+\rho^{-1}=-[\rho^{-1}(1-c\epsilon)]^{1+2\epsilon}+\rho^{-1}\\&=-\rho^{-1}(1-2\epsilon\log\rho^{-1})(1-c\epsilon)+\rho^{-1}=-\rho^{-1}(2\epsilon\log\rho^{-1}-c\epsilon)\\\Rightarrow\mathcal{G}'(1)&=\rho^{-1}[c-2\log\rho^{-1}].\end{aligned} \tag{18}$$

At $\theta=\alpha^{-2}$ (which approaches ρ^2 as $\epsilon\to0$), we have

$$\begin{aligned}\epsilon\mathcal{G}'(\alpha^{-2})&=g'(1+\epsilon,\alpha^{-2})-\beta\\&=\alpha^{N}-\beta=-[\rho^{-1}(1-c\epsilon)]^{1+\epsilon}+\rho^{-1}=-\rho^{-1}(1-\epsilon\log\rho^{-1})(1-c\epsilon)+\rho^{-1}\\&=-\rho^{-1}(\epsilon\log\rho^{-1}-c\epsilon)\\\Rightarrow\mathcal{G}'(\rho^2)&=\rho^{-1}[c-\log\rho^{-1}].\end{aligned} \tag{19}$$

As $g'(N,0)=0$ for all $N>1$, we can see from Eq. (17) that

$$\mathcal{G}'(0)=\frac{1}{\rho\epsilon}\to\infty,\quad\text{as}\quad\epsilon\to0.$$

The correction function $\mathcal{G}$ is thus singular at $\theta=0$. In this sense, the onset of criticality is singular.

B. Exact solution for $\mathcal{G}(\theta)$

Using the exact function $g(N,\theta)$ given by Eq. (4), the $\mathcal{G}(\theta)$ function can be expressed in the form

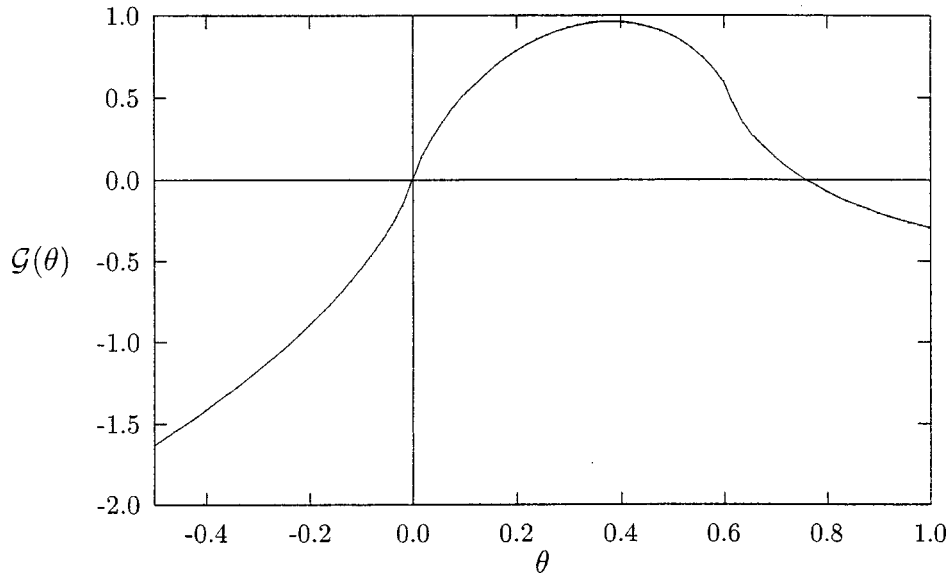

FIG. 1. The exact first-order correction to the subcritical solution for a noncritical function $g(\theta)$ of order N infinitesimally greater than unity.

$$\mathcal{G}(\theta)=\lim_{N\to 1}\frac{\left(1-\frac{1}{\rho}\,\theta^N\right)^{1/N}-\left(1-\frac{1}{\rho}\,\theta\right)}{N-1}=\frac{d}{dN}\,g(N,\theta)\bigg|_{N=1}=\lim_{N\to 1}\frac{d}{dN}\left(1-\frac{1}{\rho}\,\theta^N\right)^{1/N}.$$

Writing

$$F(N,\theta)=\left(1-\frac{1}{\rho}\,\theta^N\right)$$

and

$$\log g(N,\theta)=\frac{1}{N}\log F(N,\theta),$$

we can differentiate $\log g(N,\theta)$ with respect to N, giving

$$\frac{g'(N,\theta)}{g(N,\theta)}=\frac{1}{N}\,\frac{F'(N,\theta)}{F(N,\theta)}-\frac{1}{N^2}\log F(N,\theta)$$

$$\Leftrightarrow g'(N,\theta)=-\frac{1}{\rho N}\left(\frac{d}{dN}\,\theta^N\right)\frac{F(N,\theta)^{1/N}}{F(N,\theta)}-\frac{1}{N^2}\,F(N,\theta)^{1/N}\log F(N,\theta)$$

$$=-\frac{1}{\rho N}\,\theta^N\log\theta\,\frac{F(N,\theta)^{1/N}}{F(N,\theta)}-\frac{1}{N^2}\,F(N,\theta)^{1/N}\log F(N,\theta).$$

Letting $N\to 1$ gives us the solution

$$\mathcal{G}(\theta)=-(\theta/\rho)\log|\theta|-(1-\theta/\rho)\log|1-\theta/\rho|. \tag{20}$$

This solution for $\mathcal{G}(\theta)$ satisfies both of the functional relations (12) and (15) exactly, when c is given by (8). A plot of this noncritical solution is given in Fig. 1. The function values and derivatives at $\theta=\rho^2$ and $\theta=1$ agree with the expected results (16), (18), and (19). One can see from (20) that $\mathcal{G}'(\theta)\approx-(1/\rho)\log|\theta|$ as $\theta\to 0$; that is, the slope of $\mathcal{G}$ is infinite at zero. The logarithmic behavior explains why the singularity is not immediately apparent to the human eye.

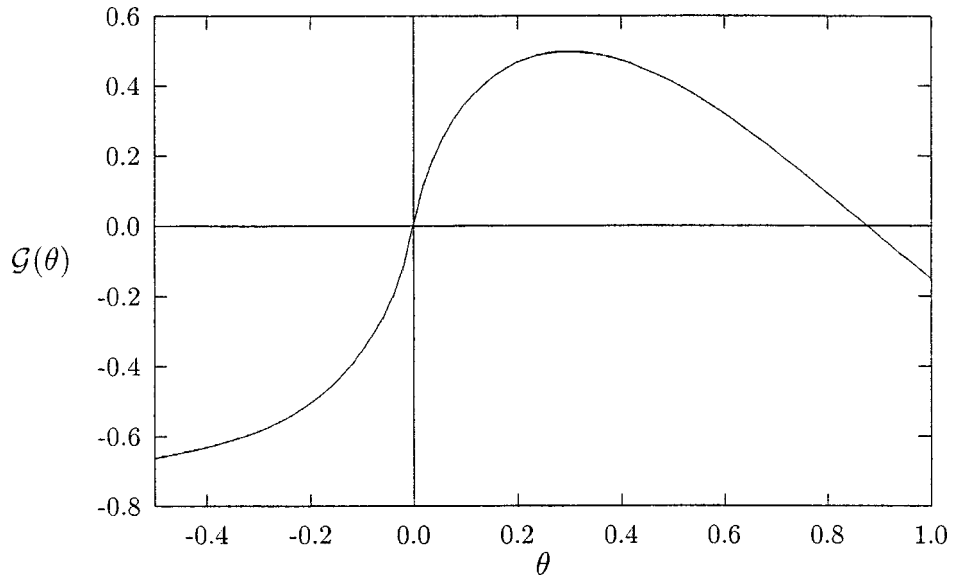

FIG. 2. The first-order correction to the subcritical solution for a critical function $g(\theta)$ with critical point order N infinitesimally greater than unity.

C. Critical $\mathcal{G}(\theta)$

To obtain the critical correction function, we can numerically solve the functional equation (1a) for small ϵ with an expansion of the form (2), and rearrange Eq. (6) to give

$$\mathcal{G}(\theta)=[g(1+\epsilon,\theta)-g(1,\theta)]/\epsilon.$$

A plot of the critical $\mathcal{G}(\theta)$ for $\epsilon=10^{-4}$ is given in Fig. 2, where c is given by (10). Numerical investigations with critical functions of order $N=1+\epsilon$ with ϵ small, have verified that both Eqs. (12) and (15) are satisfied to $O(\epsilon)$. Similarly, the numerical function values are in agreement with predicted results. Comparing the critical and noncritical functions $\mathcal{G}(\theta)$, we see that they are qualitatively similar. The singularity at zero is more apparent in the critical case, though. Also, the critical correction function is regular at $\theta=\rho$, unlike the noncritical plot.

D. Iteration of circle maps

Hu *et al.*[3] and Delbourgo and Kenny[12] have studied the dependence of the critical golden mean scaling constants using the nonanalytic circle map,

$$\theta_{n+1}=f(\theta_n)=\Omega+\theta_n|2\theta_n|^{N-1},\quad -1/2\leqslant\theta<1/2.$$

Reexpressing this map as

$$\theta_{n+1}=f(\theta_n)=\Omega+\theta_n|2\theta_n|^{\epsilon},\quad -1/2\leqslant\theta<1/2, \tag{21}$$

we can find the critical $\mathcal{G}(\theta)$ directly from circle maps.

Shenker[1] showed that one can define a sequence of approximations $\overline{f}_n(\theta)$ to the golden mean universal function, given by

$$\overline{f}_n(\theta)=\alpha^n[f_{K,\Omega}^{F_n}(\alpha^{-n}\theta)-F_{n-1}],$$

where F_n is the nth Fibonacci number. By redefining these universal function approximants in terms of $g_n(\theta)=\alpha\overline{f}_n(\theta)$, and rescaling to get $g_n(0)=1$, we are then able to use Eq. (6) to generate a sequence of $\mathcal{G}_n(\theta)$ that converges to the critical $\mathcal{G}(\theta)$ derived from the functional equations (1) for small ϵ. A plot of $\mathcal{G}(\theta)$ derived from circle maps is exactly the same as Fig. 2; the critical function generated via the circle map (21) is exactly the same as the critical function

generated from the functional equation (1a) using the approach described in the previous section, as one would expect due to the universality arguments developed by Feigenbaum *et al.*[2]

III. DEVIATION FROM SUBCRITICAL EIGENFUNCTION FOR $N=1+\epsilon$

We now focus on the behavior of the linearized function $h(\theta)$ associated with δ at the transition to criticality. This eigenfunction is found by linearizing Eqs. (1) about the fixed point $g(\theta)$:

$$\lambda h[g(\theta)]+h(\theta)g'[g(\theta)]=\frac{\lambda^2}{\alpha}h(\alpha^2\theta), \tag{22a}$$

$$h[\alpha^{-2}g(\theta)]+\frac{\lambda}{\alpha^2}h(\theta)g'[\alpha^{-2}g(\theta)]=\frac{\lambda^2}{\alpha^2}h(\alpha\theta). \tag{22b}$$

The largest eigenvalue $|\lambda|>1$ is δ, the universal constant governing scaling of the parameter space.

Jonker and Rand[10] have shown analytically that for small ϵ, $\delta(N)$ behaves no worse than

$$\delta(1+\epsilon)=\delta(1)+O(\epsilon).$$

Numerically one finds[6]

$$\delta(1+\epsilon)=\delta(1)+O(\epsilon^2).$$

Our numerical studies[13,14] have shown, however, that the following stronger relation appears to hold for all N:

$$\delta(N)=\delta(1/N). \tag{23}$$

Using this conjectured relation, we can illustrate the second-order dependence of δ on ϵ.

We may assume that for small ϵ, $\delta(N)$ deviates from its subcritical value as follows:

$$\delta(1+\epsilon)=-\frac{1}{\rho^2}(1+d\epsilon+e\epsilon^2).$$

Substituting $1/N\approx 1-\epsilon$ into this equation and using the relation (23) above to equate terms gives $d=0$, and therefore $\delta(N)$ depends on ϵ to second order near $N=1$, in agreement with other numerical findings.

We note that recently Briggs *et al.*[4] have shown numerically that, as $N\to\infty$,

$$\delta(N)\to -4.121\,326(\pm 4).$$

For the class of noncritical functions, however,

$$\delta(N)=-\alpha^{2N}=-\frac{1}{\rho^2},$$

which is independent of N or ϵ.

A. Functional equations for $\mathscr{H}(\theta)$

To determine the behavior of $h(\theta)$ as $N\to 1$, we can treat each of the linearized equations (22) in a similar manner to our treatment of Eqs. (1) above. If we assume that

$$h(N,\theta)=h(1,\theta)+\epsilon\mathscr{H}(\theta), \tag{24}$$

where

$$h(1,\theta)=1 \quad \text{and} \quad \mathscr{H}(0)=0,$$

we can show that the following equations are satisfied to $O(\epsilon)$:

$$\rho^2 \mathscr{H}(\theta) + \rho \mathscr{H}\left(1 - \frac{\theta}{\rho}\right) - \mathscr{H}\left(\frac{\theta}{\rho^2}\right) = c + \rho^3 \mathscr{G}'\left(1 - \frac{\theta}{\rho}\right), \tag{25}$$

$$\mathscr{H}(\rho^2 - \rho\theta) + \frac{1}{\rho}\,\mathscr{H}(\theta) - \frac{1}{\rho^2}\,\mathscr{H}\left(-\frac{\theta}{\rho}\right) = 2c + \mathscr{G}'(\rho^2 - \rho\theta). \tag{26}$$

Again, we expect that these relations are equivalent to the extent that the linearized equations (22) are equivalent.

B. Evaluation of $\mathscr{H}$ at special points

Substituting $\theta = 0$ and $\theta = \rho^2$ into Eq. (25), we can show

$$\mathscr{H}(1) = c + \rho^3 \mathscr{G}'(1), \tag{27a}$$

$$\mathscr{H}(\rho^2) = 2c + \mathscr{G}'(\rho^2), \tag{27b}$$

where $\mathscr{G}'(1)$ and $\mathscr{G}'(\rho^2)$ are given by (18) and (19), respectively.

C. Exact solution for $\mathscr{H}(\theta)$

Quite generally, there is a class of solutions to the eigenvalue equations (22) of the form[2]

$$h_n(\theta) = [g(\theta)]^n - \alpha^{n-1}\theta^n g'(\theta), \tag{28}$$

which result from the invariance of the renormalization group functional equations under power conjugacy transformations. These (smooth) coordinate changes generate perturbations that are eigenfunctions of both equations (22). (See Feigenbaum[8,9] or Feigenbaum and Kenway[15] for a treatment of the analogous period-doubling case.) We can use these exact eigenfunctions to derive an exact correction function $\mathscr{H}(\theta)$.

First note that for our noncritical family we have

$$g'(N,\theta) = \alpha^N \theta^{N-1}[1 + \alpha^N \theta^N]^{(1-N)/N} = \alpha^N \theta^{N-1}[g(N,\theta)]^{1-N}. \tag{29}$$

Substituting this expression into (28) gives

$$h_n(N,\theta) = [g(N,\theta)]^n - (\alpha\theta)^{n-1}\alpha^N \theta^N [g(N,\theta)]^{1-N}.$$

If we now choose $n = 1$ (to retain the θ^N dependence of h) and omit the subscript, then

$$h(N,\theta) = g(N,\theta) - \alpha^N \theta^N [g(N,\theta)]^{1-N} = \frac{[g(N,\theta)]^N - \alpha^N \theta^N}{[g(N,\theta)]^{N-1}} = [g(N,\theta)]^{1-N} = \left[1 - \frac{1}{\rho}\,\theta^N\right]^{(1-N)/N}.$$

As we did with the exact $\mathscr{G}(\theta)$ function, we can now proceed to derive an exact solution for $\mathscr{H}(\theta)$, using the equation above for $h(N,\theta)$. Note that $h(N,\theta)$ is singular at $\theta = \rho^{1/N}$. We can express $\mathscr{H}(\theta)$ in the form

$$\mathscr{H}(\theta) = \lim_{N \to 1} \frac{h(N,\theta) - h(1,\theta)}{N-1} = \frac{d}{dN}\, h(N,\theta)\bigg|_{N=1}.$$

Again, defining

$$F(N,\theta) = \left(1 - \frac{1}{\rho}\,\theta^N\right)$$

and

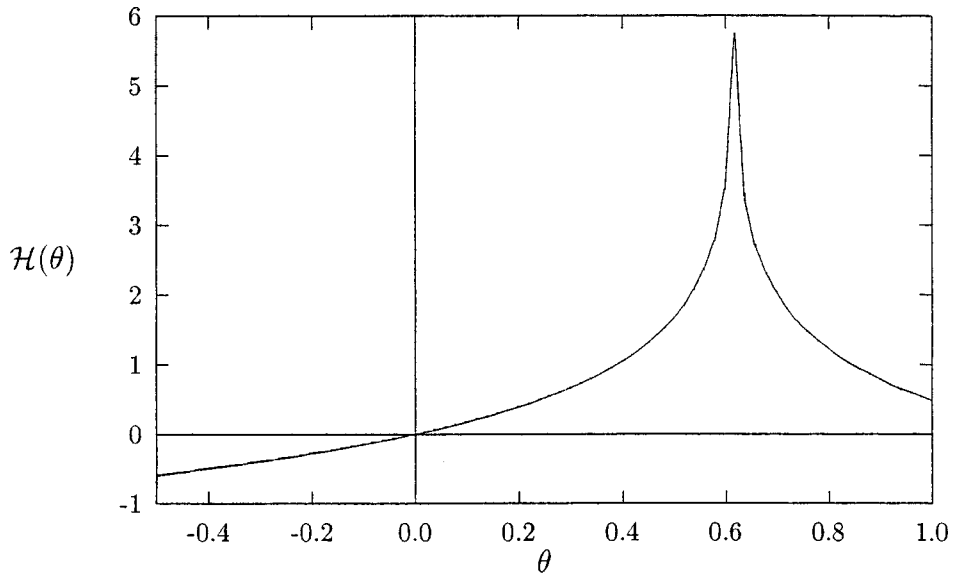

FIG. 3. The exact first-order correction to the subcritical solution for a noncritical $h(\theta)$ of order N infinitesimally greater than unity.

$$\log h(N,\theta)=\frac{1-N}{N}\log F(N,\theta),$$

we can differentiate $\log h(N,\theta)$ with respect to N to get

$$\frac{h'(N,\theta)}{h(N,\theta)}=\frac{1-N}{N}\frac{F'(N,\theta)}{F(N,\theta)}-\frac{1}{N^2}\log F(N,\theta)$$

$$\Leftrightarrow h'(N,\theta)=-\frac{1-N}{\rho N}\left(\frac{d}{dN}\,\theta^N\right)\frac{F(N,\theta)^{(1-N)/N}}{F(N,\theta)}-\frac{1}{N^2}F(N,\theta)^{(1-N)/N}\log F(N,\theta)$$

$$=\frac{N-1}{\rho N}\,\theta^N\log\theta\,\frac{F(N,\theta)^{(1-N)/N}}{F(N,\theta)}-\frac{1}{N^2}F(N,\theta)^{(1-N)/N}\log F(N,\theta)$$

$$=\frac{N-1}{\rho N}\,\theta^N\log\theta\left(1-\frac{1}{\rho}\,\theta^N\right)^{(1-2N)/N}-\frac{1}{N^2}F(N,\theta)^{(1-N)/N}\log F(N,\theta).$$

Letting $N\to 1$ gives

$$\mathcal{H}(\theta)=-\log|1-\theta/\rho|.$$

This exact expression for $\mathcal{H}(\theta)$ satisfies both of the functional equations (25) and (26), when c is given by (8). A plot of this function is given in Fig. 3. Note that the exact $\mathcal{H}(\theta)$ is singular at $\theta=\rho$. The function values at $\theta=1$ and $\theta=\rho^2$ agree with the predicted results (27).

D. Critical $\mathcal{H}(\theta)$

In order to numerically determine the critical $\mathcal{H}$, we first solve for the eigenfunction h for small ϵ using (22a). We then simply rearrange Eq. (24) in terms of $\mathcal{H}$. A plot of the critical $\mathcal{H}(\theta)$ for $\epsilon=10^{-4}$ is given in Fig. 4 when c is given by (10). Numerical investigations verify that the relations (25) and (26) are satisfied by critical functions to $O(\epsilon)$. Similarly, the numerical function values satisfy (27). However, unlike the noncritical case in Fig. 3, the critical $\mathcal{H}$ is not singular. In fact, the functions are quite dissimilar, unlike those for the correction function $\mathcal{G}$.

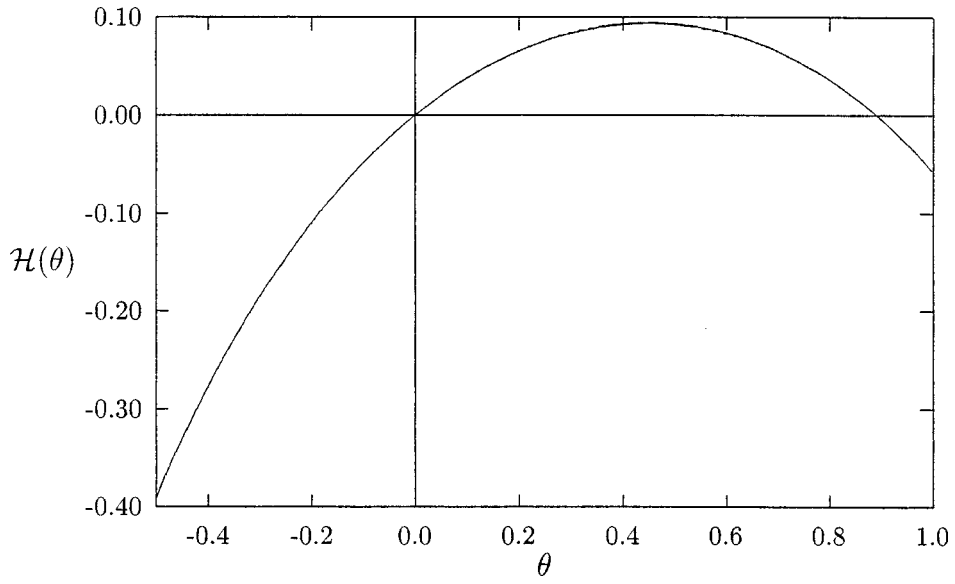

FIG. 4. The first-order correction to the subcritical solution for a critical eigenfunction $h(\theta)$ of order N infinitesimally greater than unity.

IV. CONCLUSION

We have shown that the transition from subcritical to critical dynamics in quasiperiodic systems is a singular phenomenon. We have derived a pair of functional relations, which, together with their respective linearizations, characterize this transition at the golden mean. We have compared the properties of the critical functions with those of the power conjugacy class of the subcritical solution, and in doing so, we have analytically derived the first-order corrections to the subcritical α for both critical and noncritical functions.

An analogous analysis of other quadratic irrationals could be carried out; for example, the silver mean, where three functional equations characterize the transition to chaos. A similar analysis would also describe the behavior near a Siegel boundary.[16,17]

[1] S. J. Shenker, ''Scaling behavior in map of a circle onto itself: Empirical results,'' Physica D **5**, 405–411 (1982).
[2] M. J. Feigenbaum, L. P. Kadanoff, and S. J. Shenker, ''Quasiperiodicity in dissipative systems: A renormalization group analysis,'' Physica D **5**, 370–386 (1982).
[3] B. Hu, A. Valinia, and O. Piro, ''Universal and asymptotic limits of the scaling exponents in circle maps,'' Phys. Lett. A **144**, 7–10 (1990).
[4] K. M. Briggs, T. W. Dixon, and G. Szekeres, ''Analytic solutions of the Cvitanović–Feigenbaum and Feigenbaum–Kadanoff–Shenker equations,'' Int. J. Bifurcation Chaos Appl. Sci. Eng. **8**, 347–357 (1998).
[5] D. Rand, S. Ostlund, J. Sethna, and E. D. Siggia, ''Universal transition from quasiperiodicity to chaos in dissipative systems,'' Phys. Rev. Lett. **49**, 132–135 (1982).
[6] S. Ostlund, D. Rand, J. Sethna, and E. D. Siggia, ''Universal properties of the transition from quasi-periodicity to chaos in dissipative systems,'' Physica D **8**, 303–342 (1983).
[7] P. Collet, J.-P. Eckmann, and O. E. Landford, III, ''Universal properties of maps on an interval,'' Commun. Math. Phys. **76**, 211–254 (1980).
[8] M. J. Feigenbaum, ''Quantitative universality for a class of nonlinear transformations,'' J. Stat. Phys. **19**, 25–52 (1978).
[9] M. J. Feigenbaum, ''The universal metric properties of nonlinear transformations,'' J. Stat. Phys. **21**, 669–702 (1979).
[10] L. Jonker and D. A. Rand, ''Universal properties of maps of the circle with ϵ-singularities,'' Commun. Math. Phys. **90**, 273–292 (1983).
[11] A. Arneodo and M. Holschneider, ''Fractal dimensions and holomorphic conjugacies,'' J. Stat. Phys. **50**, 995–1020 (1988).
[12] R. Delbourgo and B. G. Kenny, ''Relations between universal scaling constants for the circle map near the golden mean,'' J. Math. Phys. **32**, 1045–1051 (1991).
[13] T. W. Dixon, ''Singular limits of universal scaling constants in dynamical systems,'' Ph.D. thesis, University of Western Australia, 1997.
[14] T. W. Dixon, B. G. Kenny, and K. M. Briggs, ''On the universality of singular circle maps,'' Phys. Lett. A **231**, 359–366 (1997).
[15] M. J. Feigenbaum and R. D. Kenway, ''The onset of chaos,'' in *Statistical and Particle Physics*, Proceedings of the 26th Scottish Universities Summer School in Physics, Edinburgh, August 1983, edited by K. C. Bowler and A. J. McKane (SUSSP Publ.), pp. 1–100.

[16] N. S. Manton and M. Nauenberg, ''Universal scaling behaviour for iterated maps in the complex plane,'' Commun. Math. Phys. **89**, 555–570 (1983).
[17] M. Widom, ''Renormalization group analysis of quasi-periodicity in analytic maps,'' Commun. Math. Phys. **92**, 121–136 (1983).

Families of dynamical *r*-matrices and Jacobi inversion problem for nonlinear evolution equations

Yunbo Zeng
Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China and Department of Mathematics, Sophia University, 7-1, Kioicho, Chiyoda-ku, Tokyo 102, Japan

Runliang Lin
Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China

(Received 9 December 1997; accepted for publication 26 May 1998)

Three families of dynamical *r*-matrices are explicitly constructed by means of constrained flows of nonlinear evolution equations (NLEE) associated with polynomial second-order spectral problems. The dynamical classical Yang–Baxter equations for the three linear *r*-matrix algebras are presented with explicit expression for the extra term. The separation variables and the Jacobi inversion problems for the x- and t_n-constrained flow of the NLEEs are found. The factorization of NLEE allows us to obtain the Jacobi inversion problem for the NLEE by combining the above two Jacobi inversion problems. This provides a method of separation of variables to solve the NLEEs. © *1998 American Institute of Physics.* [S0022-2488(98)01611-9]

I. INTRODUCTION

In recent years the study of completely integrable systems admitting a classical *r*-matrix Poisson structure with dynamical *r*-matrix has attracted some attention.[1–7] A family of systems with parabolic separable potentials and parabolic dynamical *r*-matrices has been presented by recurrence relation in Ref. 3. An approach of additive deformation to a quite general family of dynamical *r*-matrices including parabolic and elliptic *r*-matrices is described in Refs. 5 and 6. The stimulus for considering such recurrence relation and additive deformation was inspired by constrained flows of Korteweg–de Vries (KdV) and coupled KdV equations. However, explicit expressions for the family both of integrable systems and of parabolic and elliptic dynamical *r*-matrices have not been carried out. It was pointed out in Refs. 7 and 8 that various constrained flows of soliton equations can be directly used to construct some kinds of dynamical *r*-matrices, since their Lax representations can always be deduced from the adjoint representation of the spectral problem.[9] The first and primary aim of this paper is to construct explicit expressions for three families of integrable systems and their dynamical *r*-matrices directly from the constrained flows of nonlinear evolution equations (NLEEs) associated with two polynomial second-order spectral problems considered in Ref. 10. The first and third families of integrable systems possess elliptic but different dynamical *r*-matrices; the second one has the parabolic dynamical *r*-matrices. We also discuss the dynamical classical Yang–Baxter equations satisfied by the three dynamical *r*-matrix algebras and present a general way to find the explicit expressions for the extra term.[4]

The second aim of this paper is to establish the Jacobi inversion problem for the NLEEs. For the Lax matrices satisfying *r*-matrix algebra with the *r*-matrix of *XXX*, *XXZ*, and *XYZ* type, the problem of determining variables of separation has been essentially solved.[11–14] By using the Lax representation, the separation variables and Jacobi inversion problems for the x- and t_n-constrained flows are found. It was shown in Refs. 15–17 that the soliton equation can always be factorized by x- and t_n-constrained flow. The factorization of the NLEEs together with a combination of the above two Jacobi inversion problems gives rise to the Jacobi inversion problem for the NLEE. The standard Jacobi inversion techniques allow us to find the solution of the NLEE in terms of the Riemann theta function.[18] Therefore, this provides a method of separation of variables to solve the NLEE. The paper is organized as follows.

 © 1998 American Institute of Physics

In Sec. II, for our purposes we reproduce two hierarchies of NLEEs associated with the polynomial second-order spectral problem and its modification, which were considered in Ref. 17, in the framework of zero-curvature representation. In Sec. III we study the first x-constrained flow and t_n-constrained flow and obtain x-finite-dimensional integrable Hamiltonian systems (FDIHS) and t_n-FDIHS. The NLEE is factorized by these two commuting x- and t_n-FDIHSs. In fact we obtain a family of x-FDIHSs depending on the order m of the polynomial. We deduce the Lax representation for the x- and t_n-FDIHSs from the adjoint representation of the spectral problem. Then we construct the classical Poisson structure for the family of x-FDIHSs and obtain explicitly a family of elliptic dynamical r-matrices. Finally we introduce the dynamical classical Yang–Baxter equations which are valid for the dynamical r-matrix algebra. We present a general way to obtain the explicit form for the extra term in the Yang–Baxter equations. In Sec. IV we determine canonically conjugated separation variables and the Jacobi inversion problem for the x- and t_n-FDIHSs by means of their Lax representation. By combining these two Jacobi inversion problems the factorization of the NLEE allows us to find the Jacobi inversion problem for the NLEE. In Secs. V and VI we consider the second x-constrained flow for the first hierarchy of the NLEEs and first x-constrained flow for the second hierarchy of the NLEEs, and present explicit expressions for the second and third families of x-FDIHSs and parabolic and elliptic dynamical r-matrices, respectively. Then we apply the same procedure as in Secs. III and IV to them. In Sec. VII a summary is presented.

II. TWO HIERARCHIES OF NONLINEAR EVOLUTION EQUATIONS

In this section we recall two hierarchies of nonlinear evolution equations (NLEEs) associated with the polynomial second-order spectral problem and its modification. These NLEEs have been described in Ref. 10. For our purpose we rederive these NLEEs in the framework of zero-curvature representation and present some formulas in different form.

A. The first hierarchy of NLEEs

Consider the following polynomial second-order spectral problem,[10]

$$\psi_x = U(u,\lambda)\psi, \quad U(u,\lambda) = \begin{pmatrix} 0 & 1 \\ -\sum_{i=0}^{m} u_i \lambda^i & 0 \end{pmatrix}, \quad \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \tag{2.1}$$

where $u_m = -1$, $u = (u_{m-1},\ldots,u_0)^T$. The adjoint representation of (2.1) is[19]

$$V_x = [U,V] \equiv UV - VU. \tag{2.2}$$

Set

$$V = \sum_{i=0}^{\infty} V_i \lambda^{-i}, \quad V_i = \begin{pmatrix} a_i & b_i \\ c_i & -a_i \end{pmatrix}. \tag{2.3}$$

Equation (2.2) yields

$$a_0 = \cdots = a_m = b_0 = \cdots = b_{m-1} = 0, \quad b_m = 1, \quad b_{m+1} = \tfrac{1}{2} u_{m-1},$$

$$a_{m+1} = -\tfrac{1}{4} u_{m-1,x}, \quad c_0 = 1, \quad c_1 = -\tfrac{1}{2} u_{m-1},\ldots,$$

and, in general,

$$\begin{pmatrix} b_{k+m} \\ \vdots \\ b_{k+1} \end{pmatrix} = L \begin{pmatrix} b_{k+m-1} \\ \vdots \\ b_k \end{pmatrix}, \tag{2.4a}$$

$$a_k = -\frac{1}{2} b_{k,x}, \quad c_k = -\frac{1}{2} b_{k,xx} - \sum_{i=0}^{m} u_i b_{k+i}, \quad k = 1,2,\ldots, \tag{2.4b}$$

where

$$L=\begin{pmatrix} L_{m-1} & L_{m-2} & \cdots & L_1 & L_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$L_0=\tfrac{1}{4}D^2+u_0-\tfrac{1}{2}D^{-1}u_{0,x}, \quad L_i=u_i-\tfrac{1}{2}D^{-1}u_{i,x}, \quad i=1,\ldots,m-1,$$

$$D=\frac{\partial}{\partial x}, \quad DD^{-1}=D^{-1}D=1. \tag{2.5}$$

Take

$$V^{(n)}(u,\lambda)=\sum_{i=0}^{n} V_i\lambda^{n-i}+\begin{pmatrix} 0 & 0 \\ -\sum_{k=1}^{m}\lambda^{m-k}\sum_{i=1}^{k} b_{n+i}u_{m-k+i} & 0 \end{pmatrix}, \tag{2.6}$$

and define

$$\psi_{t_n}=V^{(n)}(u,\lambda)\psi. \tag{2.7}$$

Then the compatibility condition of (2.1) and (2.7) gives rise to the zero-curvature representation[19]

$$U_{t_n}-V_x^{(n)}+[U,\ V^{(n)}]=0, \quad n=1,2,\ldots. \tag{2.8}$$

Equation (2.8) leads to the following hierarchy of NLEEs which can be written as an infinite-dimensional Hamiltonian system,

$$u_{t_n}=\begin{pmatrix} u_{m-1} \\ \vdots \\ u_0 \end{pmatrix}_{t_n}=J\begin{pmatrix} b_{n+m} \\ \vdots \\ b_{n+1} \end{pmatrix}=J\,\frac{\delta H_n}{\delta u}, \quad n=1,2,\ldots, \tag{2.9}$$

where the Hamiltonian H_n and Hamiltonian operator J are given by

$$J=\begin{pmatrix} 0 & 0 & \cdots & 0 & 2D \\ 0 & 0 & \cdots & 2D & J_{m-1} \\ 0 & 0 & \cdots & J_{m-1} & J_{m-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 2D & J_{m-1} & \cdots & J_1 & J_0 \end{pmatrix},$$

$$J_i=-u_{i,x}-2u_iD, \quad i=0,1,\ldots,m-1, \quad H_n=\frac{2}{m-2n-2}\sum_{i=1}^{m} iu_ib_{n+i+1}.$$

Furthermore V satisfies the adjoint representation of (2.7):[19]

$$V_{t_n}=[V^{(n)},\ V], \quad n=1,2,\ldots. \tag{2.10}$$

Under zero boundary condition we have

$$\frac{\delta\lambda}{\delta u}=(\lambda^{m-1}\psi_1^2,\ldots,\psi_1^2)^T, \quad L\,\frac{\delta\lambda}{\delta u}=\lambda\,\frac{\delta\lambda}{\delta u}. \tag{2.11}$$

B. The second hierarchy of NLEEs

Consider the following modified polynomial second-order spectral problem:[10]

$$\psi_x = U(u,\lambda)\psi, \quad U(u,\lambda) = \begin{pmatrix} u_0 & \lambda \\ -\sum_{i=1}^{m} u_i \lambda^{i-1} & -u_0 \end{pmatrix}, \quad \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \tag{2.12}$$

where $u_m = -1$, $u = (u_0,\ldots,u_{m-1})^T$. Equations (2.2) and (2.3) yield

$$a_0 = \cdots = a_{m-2} = b_0 = \cdots = b_{m-3} = 0, \quad b_{m-2} = 1, \quad b_{m-1} = \tfrac{1}{2}u_{m-1},$$

$$a_{m-1} = u_0, \quad c_0 = 1, \quad c_1 = -\tfrac{1}{2}u_{m-1},\ldots,$$

and, in general,

$$\begin{pmatrix} 2a_{k+1} \\ -b_{k+1} \\ \vdots \\ -b_{k+m-1} \end{pmatrix} = L \begin{pmatrix} 2a_k \\ -b_k \\ \vdots \\ -b_{k+m-2} \end{pmatrix}, \tag{2.13a}$$

$$c_{k+1} = a_{k,x} - \sum_{i=1}^{m} u_i b_{k+i-1}, \quad k = 1,2,..., \tag{2.13b}$$

where

$$L = \begin{pmatrix} 0 & -2u_0 + D & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ L_0 & L_1 & L_2 & \ldots & L_{m-2} & L_{m-1} \end{pmatrix},$$

$$L_0 = \tfrac{1}{4}D + \tfrac{1}{2}D^{-1}u_0 D, \quad L_i = \tfrac{1}{2}u_i + \tfrac{1}{2}D^{-1}u_i D, \quad i = 1,...,m-1.$$

Take

$$V^{(n)}(u,\lambda) = \sum_{i=0}^{n} V_i \lambda^{n-i} + \begin{pmatrix} 0 & -b_n \\ \sum_{j=1}^{m-1} \lambda^{j-1} \sum_{i=j+1}^{m} b_{n+i-j-1} u_i & 0 \end{pmatrix}, \tag{2.14}$$

and define

$$\psi_{t_n} = V^{(n)}(u,\lambda)\psi. \tag{2.15}$$

Then the zero-curvature representation (2.8) leads to the following hierarchy of NLEEs which can also be written as an infinite-dimensional Hamiltonian system

$$u_{t_n} = \begin{pmatrix} u_0 \\ \vdots \\ u_{m-1} \end{pmatrix}_{t_n} = J \begin{pmatrix} 2a_n \\ -b_n \\ \vdots \\ -b_{n+m-2} \end{pmatrix} = J \frac{\delta H_n}{\delta u}, \quad n = 1,2,..., \tag{2.16}$$

where the Hamiltonian H_n and Hamiltonian operator J are given by

$$J=\begin{pmatrix} \frac{1}{2}D & 0 & 0 & \cdots & 0 & 0 \\ 0 & J_2 & J_3 & \cdots & J_{m-1} & -2D \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & J_{m-1} & -2D & \cdots & 0 & 0 \\ 0 & -2D & 0 & \cdots & 0 & 0 \end{pmatrix},$$

$$J_i=u_{i,x}+2u_iD, \quad i=0,1,...,m-1, \quad H_n=\frac{2}{m-2n-2}\left[a_{n,x}-\sum_{i=1}^{m} iu_ib_{n+i-1}\right].$$

Also we have

$$\frac{\delta\lambda}{\delta u}=(2\psi_1\psi_2,\psi_1^2,\ldots,\lambda^{m-2}\psi_1^2)^T, \quad L\,\frac{\delta\lambda}{\delta u}=\lambda\,\frac{\delta\lambda}{\delta u}. \tag{2.17}$$

III. THE FIRST FAMILY OF DYNAMICAL *r*-MATRICES

A. The factorization of the NLEEs (2.9)

We first recall the constrained flows and factorization of the NLEEs. The *x*-constrained flow of NLEEs consists of the equations obtained from the spectral problem for N distinct λ_j and of restriction of the variational derivatives for conserved quantities H_k (for any fixed k) and λ_j:[15,16,20]

$$\begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix}_x=U(u,\lambda_j)\begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix}, \quad j=1,...,N, \tag{3.1a}$$

$$\frac{\delta H_k}{\delta u}-\frac{1}{2}\sum_{j=1}^{N}\frac{\delta\lambda_j}{\delta u}=0. \tag{3.1b}$$

The t_n-constrained flow of NLEEs consists of the equations obtained from the evolution equation of the eigenfunction for N distinct λ_j and of the NLEE itself:[15,16]

$$\begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix}_{t_n}=V^{(n)}(u,\lambda_j)\begin{pmatrix}\psi_{1j}\\ \psi_{2j}\end{pmatrix}, \quad j=1,...,N, \tag{3.2a}$$

$$u_{t_n}=J\,\frac{\delta H_n}{\delta u}. \tag{3.2b}$$

Hereafter we denote the inner product in $\mathbf{R}^N$ by $\langle .,. \rangle$ and $\Psi_i=(\psi_{i1},\ldots,\psi_{iN})^T$, $i=1,2$, $\Lambda=\mathrm{diag}\,(\lambda_1,...,\lambda_N)$. By means of (2.11), (3.1) for NLEE (2.9) can be written as

$$\Psi_{1,x}=\Psi_2, \quad \Psi_{2,x}=\Lambda^m\Psi_1-\sum_{i=0}^{m-1}u_i\Lambda^i\Psi_2, \tag{3.3a}$$

$$\begin{pmatrix}b_{m+k}\\ \vdots\\ b_{k+1}\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\langle\Lambda^{m-1}\Psi_1,\Psi_1\rangle\\ \vdots\\ \langle\Psi_1,\Psi_1\rangle\end{pmatrix}. \tag{3.3b}$$

For $k=m$, (3.3b) reads

$$\begin{pmatrix} b_{2m} \\ \vdots \\ b_{m+k} \\ \vdots \\ b_{m+1} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}u_0 + \sum_{i=1}^{m-1} L_i b_{m+i} \\ \vdots \\ \frac{1}{2}u_{m-k} + \sum_{i=1}^{k-1} L_{m-k+i} b_{m+i} \\ \vdots \\ \frac{1}{2}u_{m-1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \langle \Lambda^{m-1}\Psi_1, \Psi_1 \rangle \\ \vdots \\ \langle \Lambda^{k-1}\Psi_1, \Psi_1 \rangle \\ \vdots \\ \langle \Psi_1, \Psi_1 \rangle \end{pmatrix}, \tag{3.4a}$$

which leads to

$$u_{m-k} = \langle \Lambda^{k-1}\Psi_1, \Psi_1 \rangle - \sum_{i=1}^{k-1} L_{m-k+i} \langle \Lambda^{i-1}\Psi_1, \Psi_1 \rangle. \tag{3.4b}$$

Then, using the following kind of equality,

$$\sum_{j=l}^{k} A_{k-j} \sum_{i=l}^{j} B_i C_{j,i} = \sum_{i=l}^{k} B_i \sum_{j=0}^{k-i} A_j C_{k-j,i}, \quad l = 0,1, \tag{3.5}$$

it is found from (3.4) and (3.3a) by a direct computation and induction that

$$u_{m-k} = \sum_{j=1}^{k} (-1)^{j-1} \frac{j+1}{2^j} \sum_{l_1+\cdots+l_j=k-j} \langle \Lambda^{l_1}\Psi_1, \Psi_1 \rangle \cdots \langle \Lambda^{l_j}\Psi_1, \Psi_1 \rangle,$$

$$k = 1,...,m, \tag{3.6}$$

where $l_1 \geqslant 0,...,l_j \geqslant 0$. By substituting (3.6) into (3.3a), the first x-constrained flow of (2.9) can be written in canonical finite-dimensional Hamiltonian system (FDHS)

$$\Psi_{1,x} = \frac{\partial F_0}{\partial \Psi_2}, \quad \Psi_{2,x} = -\frac{\partial F_0}{\partial \Psi_1}, \tag{3.7}$$

with the Hamiltonian F_0 given by

$$F_0 = \frac{1}{2}\langle \Psi_2, \Psi_2 \rangle + \sum_{j=0}^{m} \frac{(-1)^{j-1}}{2^{j+1}} \sum_{l_1+\cdots+l_{j+1}=m-j} \langle \Lambda^{l_1}\Psi_1, \Psi_1 \rangle \cdots \langle \Lambda^{l_{j+1}}\Psi_1, \Psi_1 \rangle.$$

The integrals of motion for (3.7) are

$$F_k = \frac{1}{2}\langle \Lambda^k \Psi_2, \Psi_2 \rangle + \frac{1}{4}\sum_{j=0}^{k-1} [\langle \Lambda^j \Psi_2, \Psi_2 \rangle \langle \Lambda^{k-1-j}\Psi_1, \Psi_1 \rangle - \langle \Lambda^j \Psi_1, \Psi_2 \rangle \langle \Lambda^{k-1-j}\Psi_1, \Psi_2 \rangle]$$

$$+ \sum_{j=0}^{m} \frac{(-1)^{j-1}}{2^{j+1}} \sum_{l_1+\cdots+l_{j+1}=m-j} \langle \Lambda^{l_1}\Psi_1, \Psi_1 \rangle \cdots \langle \Lambda^{l_j}\Psi_1, \Psi_1 \rangle \langle \Lambda^{l_{j+1}+k}\Psi_1, \Psi_1 \rangle,$$

$$k = 0,1,... . \tag{3.8}$$

The FDHS (3.7) was first presented in Ref. 21. However, the Lax representation, r-matrix, and separability for (3.7) have not been studied.

Under (3.6) and (3.7), t_{m+1}-constrained flow (3.2) of NLEE (2.9) can be written as a FDHS,

$$\Psi_{1,t_{m+1}} = \frac{\partial F_1}{\partial \Psi_2}, \quad \Psi_{2,t_{m+1}} = -\frac{\partial F_1}{\partial \Psi_1}, \tag{3.9}$$

and t_{m+2}-constrained flow (3.2) of NLEE (2.9) becomes a FDHS:

$$\Psi_{1,t_{m+2}}=\frac{\partial F_2}{\partial \Psi_2},\quad \Psi_{2,t_{m+2}}=-\frac{\partial F_2}{\partial \Psi_1}. \tag{3.10}$$

We will show in Sec. III C that FDHSs (3.7), (3.9), and (3.10) are a finite-dimensional integrable Hamiltonian system (FDIHS) with the integrals of motion given by (3.8). The construction of the x- and t_n-constrained flow ensures that if (Ψ_1,Ψ_2) satisfies two commuting FDIHSs (3.7) and (3.9) [(3.10)] simultaneously, then u given by (3.6) is a solution of the NLEE (2.9) with $n=m+1(n=m+2)$.

In general, the x-constrained flow (3.1) and t_n-constrained flow (3.2) under (3.1) can be transformed into an x-FDIHS and t_n-FDIHS. Then the NLEE is factorized into these two commuting x-FDIHS and t_n-FDIHS, namely if (Ψ_1,Ψ_2,u) satisfies these two commuting FDIHSs simultaneously, then u is a solution of the NLEE. Therefore, some kind of solution, such as finite-gap solution, for NLEE can be obtained through solving two commuting x- and t_n-FDIHS. We shall find the Jacobi inversion problem for these x- and t_n-FDIHSs later, and combine them to obtain the Jacobi inversion problem for the NLEE, which is solvable in terms of Riemann theta function.

B. Lax representation

By following the method in Ref. 9, the Lax representation for FDHS (3.7) can be deduced from the adjoint representation (2.2). First we have to find the expression of a_i, b_i, c_i under the constraint (3.6) and (2.11), which are denoted by $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{c}_i$. According to (2.4), (2.11), (3.6), and (3.7), it is found by a direct calculation and using (3.5):

$$\tilde{a}_0=\cdots=\tilde{a}_m=\tilde{b}_0=\cdots=\tilde{b}_{m-1}=0,\quad \tilde{b}_m=1,\quad \tilde{c}_0=1,$$

$$\tilde{b}_{m+1+k}=\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_1\rangle,\quad k=0,1,..., \tag{3.11a}$$

$$\tilde{a}_{m+1+k}=-\tfrac{1}{2}\tilde{b}_{m+1+k,x}=-\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_2\rangle,\quad k=0,1,..., \tag{3.11b}$$

$$\tilde{c}_k=-\sum_{i=m-k}^{m}u_i\tilde{b}_{k+i}=\sum_{j=1}^{k}\left(-\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,$$

$$k=1,...,m,$$

$$\tilde{c}_{m+1+k}=-\frac{1}{2}\tilde{b}_{m+1+k,xx}-\sum_{i=0}^{m}u_i\tilde{b}_{m+1+k+i}=-\frac{1}{2}\langle\Lambda^k\Psi_2,\Psi_2\rangle,\quad k=0,1,... . \tag{3.11c}$$

The construction of $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{c}_i$ guarantees that under (3.7)

$$\tilde{V}=\sum_{m=0}^{\infty}\begin{pmatrix}\tilde{a}_i & \tilde{b}_i\\ \tilde{c}_i & -\tilde{a}_i\end{pmatrix}\lambda^{-i}$$

satisfies (2.2). Set

$$M\equiv\lambda^m\tilde{V}=\begin{pmatrix}A(\lambda) & B(\lambda)\\ C(\lambda) & -A(\lambda)\end{pmatrix}. \tag{3.12}$$

We have

$$A(\lambda)=-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j},\quad B(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \tag{3.13a}$$

$$C(\lambda)=\lambda^m+\sum_{k=1}^{m}\lambda^{m-k}\sum_{j=1}^{k}\left(-\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}^2}{\lambda-\lambda_j}. \tag{3.13b}$$

The construction of M ensures the following proposition.

Proposition 1: The Lax representation for the FDHS (3.7) is given by

$$M_x=[\overline{U},\ M], \tag{3.14}$$

where $\overline{U}$ is obtained by inserting (3.6) into U in (2.1).

Similarly, $\tilde{V}$ under (3.9) [(3.10)] satisfies the adjoint representation (2.10). Thus the Lax representation for (3.9) $(n=m+1)$ [(3.10) $(n=m+2)$] is given by

$$M_{t_n}=[\overline{V}^{(n)},\ M], \tag{3.15}$$

where $\overline{V}^{(n)}$ is obtained by inserting (3.6) into $V^{(n)}$ in (2.6).

The equations (3.14) and (3.15) implies that $\mathrm{Tr}\ M^2(\lambda)=A^2(\lambda)+B(\lambda)C(\lambda)$ is the generating function of the integrals of motion for (3.7), (3.9), and (3.10). We have

$$A(\lambda)^2+B(\lambda)C(\lambda)=\lambda^m+\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i}, \tag{3.16}$$

$$F^{(i)}=-\frac{1}{2}\psi_{2i}^2+\frac{1}{2}\left[\lambda_i^m+\sum_{k=1}^{m}\lambda_i^{m-k}\sum_{j=1}^{k}\left(-\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle\right]\psi_{1i}^2$$
$$-\frac{1}{4}\sum_{k\neq i}\frac{(\psi_{1i}\psi_{2k}-\psi_{1k}\psi_{2i})^2}{\lambda_k-\lambda_i},\quad i=1,...,N,$$

where $F^{(i)}$, $i=1,...,N$, are independent integrals of motion for (3.7), (3.9), and (3.10). In fact, we have

$$F_k=\sum_{i=0}^{N}\lambda_i^k F^{(i)},\quad k=0,1,\ldots .$$

C. The first family of dynamical *r*-matrices

In this section we describe the classical Poisson structure associated with the Lax representation for (3.7), (3.9), and (3.10).

With respect to the standard Poisson bracket, it is found by a straightforward but lengthy calculation and using (3.5) that

$$\{A(\lambda),A(\mu)\}=\{B(\lambda),B(\mu)\}=0,\quad \{C(\lambda),C(\mu)\}=2g_1(\lambda,\mu)[A(\lambda)-A([\mu)], \tag{3.17a}$$

$$\{A(\lambda),B(\mu)\}=\frac{1}{\mu-\lambda}[B(\lambda)-B(\mu)],\quad \{B(\lambda),C(\mu)\}=\frac{2}{\mu-\lambda}[A(\lambda)-A(\mu)], \tag{3.17b}$$

$$\{A(\lambda),C(\mu)\}=\frac{1}{\mu-\lambda}[C(\mu)-C(\lambda)]-g_1(\lambda,\mu)B(\lambda), \tag{3.17c}$$

where

$$g_1(\lambda,\mu)=\sum_{k=0}^{m-1}\lambda^k\mu^{m-1-k}+\sum_{l=1}^{m-1}\sum_{k=0}^{m-1-l}\lambda^k\mu^{m-1-l-k}\sum_{j=1}^{l}\left(-\frac{1}{2}\right)^j$$

$$\times(j+1)\sum_{l_1+\cdots+l_j=l-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{3.18}$$

We use the standard notation σ_k, $k=0,1,2,3$, for the Pauli matrices and $\sigma_\pm=\frac{1}{2}(\sigma_1\pm i\sigma_2)$. Denote

$$M_1(\lambda)=M(\lambda)\otimes I,\quad M_2(\mu)=I\otimes M(\mu),$$

where I is the 2×2 unit matrix. Let P be the permutation matrix. Then (3.17) gives rise to the the following proposition.

Proposition 2: The classical Poisson structure for (3.7), (3.9), and (3.10) is given in the following linear r-matrix form:[1]

$$\{M_1(\lambda),M_2(\mu)\}=[r_{12}(\lambda,\mu),M_1(\lambda)]-[r_{21}(\lambda,\mu),M_2(\mu)], \tag{3.19}$$

where the r-matrix is of the form

$$r_{12}(\lambda,\mu)=\frac{1}{\lambda-\mu}P+g_1(\lambda,\mu)S,\quad P=\frac{1}{2}\sum_{k=0}^{3}\sigma_k\otimes\sigma_k,\quad S=\sigma_-\otimes\sigma_-,$$
$$r_{21}(\lambda,\mu)=Pr_{12}(\lambda,\mu)P=r_{12}(\mu,\lambda). \tag{3.20}$$

Proof: The proof is a direct computation.

The family of r-matrices $r_{12}(\lambda,\mu)$ for $m=2,3,...$ is of the dynamical type which depends not only on the spectral parameter but also on the dynamical variables.

According to a remarkable theorem proved in Ref. 1, the involutivity of the integrals of motion

$$\{F^{(i)},F^{(j)}\}=0,\quad i,j=1,...,N, \tag{3.21}$$

is a consequence of (3.19). Therefore (3.7), (3.9), and (3.10) are FDIHSs.[22]

Let us denote

$$M_1(\lambda)=M(\lambda)\otimes I\otimes I,\quad M_2(\mu)=I\otimes M(\mu)\otimes I,\quad M_3(\nu)=I\otimes I\otimes M(\nu),$$

$$P_{12}=P\otimes I,\quad P_{23}=I\otimes P,\quad P_{13}=\frac{1}{2}\sum_{k=0}^{3}\sigma_k\otimes I\otimes\sigma_k,$$

$$S_{12}=S\otimes I,\quad S_{23}=I\otimes S,\quad S_{13}=\sigma_-\otimes I\otimes\sigma_-.$$

It is known that the r-bracket is a Poisson bracket if r-matrix obeys some classical dynamical Yang–Baxter equations. In the following we introduce dynamical Yang–Baxter equations for this family of dynamical r-matrices.

Proposition 3: For the linear r-matrix algebra (3.19) and (3.20), the following dynamical Yang–Baxter equations are valid:

$$[d_{12}(\lambda,\mu),d_{13}(\lambda,\nu)]+[d_{12}(\lambda,\mu),d_{23}(\mu,\nu)]+[d_{32}(\nu,\mu),d_{13}(\lambda,\nu)]+\{M_2(\mu),d_{13}(\lambda,\nu)\}$$
$$-\{M_3(\nu),d_{12}(\lambda,\mu)\}+f(\lambda,\mu,\nu)[S_0,M_2(\mu)-M_3(\nu)], \tag{3.22}$$

plus cyclic permutations, where

$$d_{ij}(\lambda,\mu)=\frac{1}{\lambda-\mu}P_{ij}+g_1(\lambda,\mu)S_{ij},$$

$$d_{ji}(\lambda,\mu)=d_{ij}(\lambda,\mu),\quad i<j,\quad i,j=1,2,3,\quad S_0=\sigma_-\otimes\sigma_-\otimes\sigma_-,$$

$$f(\lambda,\mu,\nu)=-2\sum_{k=0}^{m-2}\sum_{l=0}^{m-2-k}\lambda^k\mu^l\nu^{m-2-l-k}-2\sum_{l=1}^{m-2}\sum_{k=0}^{m-2-l}\sum_{i=0}^{m-2-l-k}\lambda^k\mu^i\nu^{m-2-l-k-i}$$
$$\times\sum_{j=1}^{l}\left(-\frac{1}{2}\right)^j(j+1)^2\sum_{l_1+\cdots+l_j=l-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{3.23}$$

Proof: A direct calculation shows that the validity of (3.22) is due to the following equality

$$2\{A(\mu),g_1(\lambda,\nu)\}+\frac{2}{\lambda-\mu}[g_1(\mu,\nu)-g_1(\lambda,\nu)]-f(\lambda,\mu,\nu)B(\mu)=0, \tag{3.24}$$

where λ,μ are treated as spectral parameters. Notice that if (3.24) holds and $g_1(\lambda,\nu)$ is replaced by $g_1(\lambda,\nu)-\frac{1}{2}\Sigma_{j=1}^{N}\psi_{2j}^2/(\lambda-\lambda_j)$, then (3.24) still holds. Comparing (3.24) with (3.17c), we find that the relation between $g_1(\lambda,\mu)$ and $f(\lambda,\mu,\nu)$ is the same as the relation between $C(\mu)$ and $g_1(\lambda,\mu)$. Thus we may obtain $f(\lambda,\mu,\nu)$ from $-2g_1(\lambda,\mu)$ by replacing m by $m-1$ and μ^k by $\Sigma_{l=0}^{k}\mu^l\nu^{k-l}$. This completes the proof.

The proof provides a general way to obtain $f(\lambda,\mu,\nu)$ from $g(\lambda,\mu)$, namely, $f(\lambda,\mu,\nu)$ can be deduced from $-2g(\lambda,\mu)$ by the same relation as that between $g(\lambda,\mu)$ and $C(\mu)$. The dynamical Yang–Baxter equations (3.22) take the general form introduced in Ref. 4.

IV. THE JACOBI INVERSION PROBLEM FOR NLEE (2.9)

We introduce separation variables w_k,v_k, $k=1,...,N$ for (3.7), (3.9), and (3.10) by zeros of $B(\lambda)$ (Refs. 11–14):

$$B(\lambda)=1+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)}, \tag{4.1a}$$

and

$$v_k=A(w_k)=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{w_k-\lambda_j},\quad k=1,...,N, \tag{4.1b}$$

where

$$K(\lambda)\equiv\prod_{j=1}^{N}(\lambda-\lambda_j)=\sum_{i=0}^{N}\alpha_i\lambda^{N-i},\quad R(\lambda)\equiv\prod_{j=1}^{N}(\lambda-w_j)=\sum_{i=0}^{N}\beta_i\lambda^{N-i},$$

$$\alpha_0=1,\quad \alpha_1=-\sum_{j=1}^{N}\lambda_j,\quad \alpha_2=\sum_{j=1}^{N}\sum_{k=j+1}^{N}\lambda_j\lambda_k,\ldots,\quad \alpha_N=(-1)^N\lambda_1\cdots\lambda_N,$$

$$\beta_0=1,\quad \beta_1=-\sum_{j=1}^{N}w_j,\quad \beta_2=\sum_{j=1}^{N}\sum_{k=j+1}^{N}w_jw_k,\ldots,\quad \beta_N=(-1)^Nw_1\cdots w_N. \tag{4.1c}$$

It follows from (4.1a) that the w_k, $k=1,...,N$, define elliptic coordinates given by the formula

$$\psi_{1j}^2=2\frac{R(\lambda_j)}{K'(\lambda_j)},\quad j=1,...,N, \tag{4.2}$$

where the prime denotes differentiation with respect to λ. Then we obtain from (4.2)

$$\sum_{j=1}^{N}\psi_{2j}d\psi_{1j}=\sum_{j=1}^{N}v_kdw_k,$$

which implies that the coordinates w_k, v_k are canonically conjugated.[22] From (4.1a) we find

$$\langle \Lambda^k \Psi_1, \Psi_1 \rangle = 2\beta_{k+1} + 2\sum_{i=0}^{N} \beta_i \sum_{l=1}^{k+1-i} (-1)^l \prod_{j_1+\cdots+j_l=k+1-i} \alpha_{j_1} \cdots \alpha_{j_l}, \tag{4.3}$$

where $j_i \geqslant 1$, $\alpha_k = \beta_k = 0$ for $k > N$. Combining (3.6) and (4.3) gives rise to the expression of u_k in terms of w_k, which are the symmetric functions of w_k, $k = 1,...,N$. Let

$$A^2(\lambda) + B(\lambda)C(\lambda) = \frac{P(\lambda)}{K(\lambda)}, \quad P(\lambda) = \sum_{i=1}^{N+m+1} P_i \lambda^{i-1}, \tag{4.4}$$

where P_i, $i = 1,...,N$, are also the integrals of motion for (3.7), (3.9), and (3.10). By substituting (3.13) we find

$$P_{N+m+1} = 1, \quad P_{N+m} = \alpha_1, \quad \ldots, \quad P_{N+1} = \alpha_m, \tag{4.5}$$

$$F_0 = -P_N + \alpha_{m+1}, \quad F_1 = -P_{N-1} + \alpha_1 P_N - \alpha_1 \alpha_{m+1} + \alpha_{m+2}, \tag{4.6a}$$

$$F_2 = -P_{N-2} + \alpha_1 P_{N-1} - (\alpha_1^2 - \alpha_2) P_N + \alpha_1^2 \alpha_{m+1} - \alpha_1 \alpha_{m+2} - \alpha_2 \alpha_{m+1} + \alpha_{m+3}, \quad \ldots . \tag{4.6b}$$

In order to write the Hamilton–Jacobi equation from (4.4), we must reinterpret the P_i as integration constants and replace v_k by the partial derivatives $\partial W / \partial w_k$ of the generating function W of canonical transformation.[22] Inserting $\lambda = w_k$, we find from (4.4)

$$v_k = \sqrt{\frac{P(w_k)}{K(w_k)}}, \quad k = 1,...,N, \tag{4.7}$$

which implies that variables in the Hamilton–Jacobi equation are completely separable. Here W can be expressed in the separation form $W(w_1,\ldots,w_N) = \Sigma_{k=1}^{N} W_k(w_k)$. By replacing $v_k = \partial W_k / \partial w_k$ and interpreting P_i as integration constants, (4.4) may be integrated to give the completely separated solution

$$W(w_1,\ldots,w_N) = \sum_{k=1}^{N} \int^{w_k} \sqrt{\frac{P(\lambda)}{K(\lambda)}}\, d\lambda. \tag{4.8}$$

The linearizing coordinates are then

$$Q_i \equiv \frac{\partial W}{\partial P_i} = \frac{1}{2} \sum_{k=1}^{N} \int^{w_k} \frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\, d\lambda, \quad i = 1,...,N. \tag{4.9}$$

This equality provides a map, called the Abel map, from the old coordinates w_k, $k = 1,...,N$, which live on the Riemann surface, to new coordinates Q_k, $k = 1,...,N$, which live on its Jacobi variety. The linear flow induced by (3.7) is then given by [using (4.6a)]

$$Q_i = \gamma_i + \frac{\partial F_0}{\partial P_i} x = \gamma_i - x\,\delta_{i,N}, \quad i = 1,...,N, \tag{4.10}$$

where γ_i are constants.

The linear flow induced by (3.9) is of the form [using (4.6a)]

$$Q_i = \overline{\gamma_i} + \frac{\partial F_1}{\partial P_i} t_{m+1} = \overline{\gamma_i} - [\,\delta_{i,N-1} - \alpha_1 \delta_{i,N}] t_{m+1}, \quad i = 1,...,N. \tag{4.11}$$

The factorization of NLEE (2.9) and combining of Eqs. (4.9), (4.10), and (4.11) give rise to the following proposition.

Proposition 4: The Jacobi inversion problem for FDHS (3.7) is given by

$$\frac{1}{2}\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i-\delta_{i,N}x,\quad i=1,...,N, \tag{4.12}$$

the Jacobi inversion problem for FDHS (3.9) is given by

$$\frac{1}{2}\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i+(\delta_{i,N}\alpha_1-\delta_{i,N-1})t_{m+1},\quad i=1,...,N, \tag{4.13}$$

and the Jacobi inversion problem for NLEE (2.9) with $n=m+1$ is of the form

$$\frac{1}{2}\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i-\delta_{i,N}(x-\alpha_1 t_{m+1})-\delta_{i,N-1}t_{m+1},\quad i=1,...,N. \tag{4.14}$$

The linear flow induced by (3.10) is given by [using (4.6b)]

$$Q_i=\overline{\gamma}_i+\frac{\partial F_2}{\partial P_i}\,t_{m+2}=\overline{\gamma}_i-[\delta_{i,N-2}-\alpha_1\delta_{i,N-1}+(\alpha_1^2-\alpha_2)\delta_{i,N}]t_{m+2},\quad i=1,...,N. \tag{4.15}$$

The factorization of NLEE (2.9) and the combining of Eqs. (4.9), (4.10), and (4.15) lead to the Jacobi inversion problem for FDHS (3.10),

$$\frac{1}{2}\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i-[\delta_{i,N-2}-\alpha_1\delta_{i,N-1}+(\alpha_1^2-\alpha_2)\delta_{i,N}]t_{m+2},\quad i=1,...,N, \tag{4.16}$$

and the Jacobi inversion problem for NLEE (2.9) with $n=m+2$:

$$\frac{1}{2}\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i-[\delta_{i,N-2}-\alpha_1\delta_{i,N-1}]t_{m+2}-[x+(\alpha_1^2-\alpha_2)t_{m+2}]\delta_{i,N},$$
$$i=1,...,N. \tag{4.17}$$

By using standard Jacobi inversion techniques,[18] the solutions u of NLEE (2.9), which are the symmetric functions of w_k, $k=1,...,N$, determined by (3.6) and (4.3), can be given an explicit form in terms of Riemann theta functions. This provides a method of separation of variables to solve NLEE (2.9) for $n=m+1$, $m+2$. The above presented method can be applied to the whole hierarchy (2.9).

V. THE SECOND FAMILY OF DYNAMICAL *r*-MATRICES

We now consider the second x-constrained flow of NLEE (2.9) and its Lax representation, dynamical r-matrix, and Jacobi inversion problem.

A. The second *x*-constrained flow of the NLEE (2.9)

The second x-constrained flow of NLEE (2.9) is obtained from (3.3) for $k=m+1$. Equation (3.3b) with $k=m+1$ reads

$$\begin{pmatrix} b_{2m+1}\\ \vdots\\ b_{m+k}\\ \vdots\\ b_{m+2}\end{pmatrix}=\begin{pmatrix} \Sigma_{i=0}^{m-1}L_i b_{m+i+1}\\ \vdots\\ \frac{1}{2}u_{m-k}+\Sigma_{i=1}^{k-1}L_{m-k+i}b_{m+i}\\ \vdots\\ \frac{1}{2}u_{m-2}+\frac{1}{2}L_{m-1}u_{m-1}\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\langle\Lambda^{m-1}\Psi_1,\Psi_1\rangle\\ \vdots\\ \langle\Lambda^{k-2}\Psi_1,\Psi_1\rangle\\ \vdots\\ \langle\Psi_1,\Psi_1\rangle\end{pmatrix}, \tag{5.1a}$$

which leads to

$$u_{m-k}=\langle\Lambda^{k-2}\Psi_1,\Psi_1\rangle-L_{m-k+1}u_{m-1}-\sum_{i=2}^{k-1}L_{m-k+i}\langle\Lambda^{i-2}\Psi_1,\Psi_1\rangle,\quad k=2,...,m,\tag{5.1b}$$

$$L_0u_{m-1}=\langle\Lambda^{m-1}\Psi_1,\Psi_1\rangle-\sum_{i=1}^{m-1}L_i\langle\Lambda^{i-1}\Psi_1,\Psi_1\rangle.\tag{5.1c}$$

By using (3.5) and the following kind of equality,

$$\sum_{j=2l}^{k}A_{k-j}\sum_{i=l}^{[j/2]}B_iC_{j,2i}=\sum_{i=l}^{[k/2]}B_i\sum_{j=0}^{k-2i}A_jC_{k-j,2i},\quad l=0,1,\tag{5.2}$$

it is found from (5.1b) and (3.3a) by a direct but lengthy computation that

$$u_{m-k}=E_{k,0}u_{m-1}^k+\sum_{i=0}^{k-2}u_{m-1}^i\sum_{j=1}^{[(k-i)/2]}E_{i,j}\sum_{l_1+\cdots+l_j=k-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,$$
$$k=2,...,m,\tag{5.3a}$$

where

$$E_{i,j}=-\left(-\frac{1}{2}\right)^{i+j}\frac{(i+j+1)!}{i!j!}=-(i+j+1)\beta_{i,j},\quad \beta_{i,j}=\left(-\frac{1}{2}\right)^{i+j}\frac{(i+j)!}{i!j!}.\tag{5.3b}$$

Denote

$$q=u_{m-1},\quad p=-\tfrac{1}{8}u_{m-1,x}.\tag{5.4}$$

By substituting (5.3) into (3.3a) and (5.1c), the second x-constrained flow of (2.9) can be written as a canonical FDHS with the extended phase space (Ψ_1,q,Ψ_2,p):

$$\Psi_{1,x}=\frac{\partial F_0}{\partial\Psi_2},\quad \Psi_{2,x}=-\frac{\partial F_0}{\partial\Psi_1},\quad q_x=\frac{\partial F_0}{\partial p},\quad p_x=-\frac{\partial F_0}{\partial q},\tag{5.5}$$

with the Hamiltonian F_0 given by

$$F_0=\frac{1}{2}\langle\Psi_2,\Psi_2\rangle+\beta_{m+2,0}q^{m+2}-4p^2$$
$$+\sum_{i=0}^{m}q^i\sum_{j=1}^{[(m+2-i)/2]}\beta_{i,j}\sum_{l_1+\cdots+l_j=m+2-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle.$$

B. Lax representation

We now derive the Lax representation for the family of FDHS (5.5) from the adjoint representation (2.2) in the same way as for (3.7). According to (2.4), (2.11), (5.1), and (5.5), it is found by a direct calculation

$$\tilde a_0=\cdots=\tilde a_m=\tilde b_0=\cdots=\tilde b_{m-1}=0,\quad \tilde b_m=1,$$
$$\tilde b_{m+1}=\tfrac{1}{2}q,\quad \tilde a_{m+1}=-\tfrac{1}{2}\tilde b_{m+1,x}=2p,\quad \tilde c_0=1\tag{5.6a}$$
$$\tilde b_{m+2+k}=\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_1\rangle,\quad k=0,1,...,$$
$$\tilde a_{m+2+k}=-\tfrac{1}{2}\tilde b_{m+2+k,x}=-\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_2\rangle,\quad k=0,1,...,\tag{5.6b}$$

$$\tilde{c}_k=-\frac{1}{2}\tilde{b}_{k,xx}-\sum_{i=m-k}^{m}u_i\tilde{b}_{k+i}=\beta_{k,0}q^k$$
$$+\sum_{i=0}^{k-2}q^i\sum_{j=1}^{[(k-i)/2]}\beta_{i,j}\sum_{l_1+\cdots+l_j=k-i-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,\quad k=1,\ldots,m+1,$$
$$\tilde{c}_{m+2+k}=-\frac{1}{2}\tilde{b}_{m+2+k,xx}-\sum_{i=0}^{m}u_i\tilde{b}_{m+k+2+i}=-\frac{1}{2}\langle\Lambda^k\Psi_2,\Psi_2\rangle,\quad k=0,1,\ldots. \tag{5.6c}$$

Set

$$M\equiv\lambda^{m+1}\tilde{V}=\begin{pmatrix}A(\lambda) & B(\lambda)\\ C(\lambda) & -A(\lambda)\end{pmatrix}.$$

Then we have

$$A(\lambda)=2p-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j},\quad B(\lambda)=\lambda+\frac{1}{2}q+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \tag{5.7a}$$

$$C(\lambda)=\lambda^{m+1}+\sum_{k=1}^{m+1}\lambda^{m+1-k}\tilde{c}_k-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{2j}^2}{\lambda-\lambda_j}. \tag{5.7b}$$

The construction of M gives rise to the following proposition.

Proposition 5: The Lax representation for the FDHS (5.5) is given by (3.14) with $\bar{U}$ obtained by inserting (5.3) and (5.4) into U in (2.1).

Similarly, the equality

$$A^2(\lambda)+B(\lambda)C(\lambda)=\lambda^{m+2}+F_0+\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i} \tag{5.8}$$

determines $N+1$ independent integrals of motion $F_0,F^{(i)}$, $i=1,\ldots,N$, for FDHS (5.5). We omit the expression for $F^{(i)}$.

C. The second family of dynamical *r*-matrices

By a straightforward but lengthy calculation and using (5.2), it is found that $A(\lambda),B(\lambda),C(\lambda)$ satisfies the same formulas as (3.17) with $g_1(\lambda,\mu)$ replaced by the following $g_2(\lambda,\mu)$:

$$g_2(\lambda,\mu)=-\sum_{l=0}^{m-1}E_{l,0}q^l\sum_{k=0}^{m-1-l}\mu^{m-1-l-k}\lambda^k-\sum_{n=0}^{m-3}\sum_{k=0}^{m-3-n}\lambda^k\mu^{m-3-n-k}\sum_{l=0}^{n}q^l\sum_{j=1}^{[(n+2-l)/2]}E_{l,j}$$
$$\times\sum_{l_1+\cdots+l_j=n+2-l-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{5.9}$$

Then we have the following proposition.

Proposition 6: The classical Poisson structure for FDHS (5.5) is given in the linear r-matrix form (3.19) with

$$r_{12}(\lambda,\mu)=\frac{1}{\lambda-\mu}P+g_2(\lambda,\mu)S,\quad r_{21}(\lambda,\mu)=Pr_{12}(\lambda,\mu)P=r_{12}(\mu,\lambda). \tag{5.10}$$

This gives the second family of dynamical *r*-matrices $r_{12}(\lambda,\mu)$ for $m=2,3,\ldots$. The involutivity of the integrals of motion $F_0,F^{(i)}$, $i=1,\ldots,N$, is a consequence of (3.9) and (5.10). Therefore (5.5) is a FDIHS.

As shown in the proof of Proposition 3, the following proposition is valid.

Proposition 7: For the linear r-matrix algebra (3.19) and (5.10) the dynamical Yang–Baxter equations (3.22) with $d_{i,j}(\lambda,\mu)$ and f given by following formulas are valid:

$$d_{ij}(\lambda,\mu)=\frac{1}{\lambda-\mu}P_{ij}+g_2(\lambda,\mu)S_{ij},\quad d_{ji}(\lambda,\mu)=d_{ij}(\lambda,\mu),\quad i<j,\quad i,j=1,2,3,$$

$$\begin{aligned}f(\lambda,\mu,\nu)=&2\sum_{l=0}^{m-3}\tilde{E}_{l,0}q^l\sum_{k=0}^{m-3-l}\sum_{i=0}^{m-3-l-k}\lambda^k\mu^i\nu^{m-3-l-k-i}\\&+2\sum_{n=0}^{m-5}\sum_{k=0}^{m-5-n}\sum_{i=0}^{m-5-n-k}\lambda^k\mu^i\nu^{m-5-n-k-i}\\&\times\sum_{l=0}^{n}q^l\sum_{j=2}^{[(n+2-l)/2]}\tilde{E}_{l,j}\sum_{l_1+\cdots+l_j=n+2-l-2j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,\end{aligned}$$

where

$$\tilde{E}_{i,j}=(i+j+1)^2\beta_{i,j}.$$

D. The Jacobi inversion problem for FDHS (5.5)

We introduce separation variables w_k,v_k, $k=1,...,N+1$, for (5.5) as

$$B(\lambda)=\lambda+\frac{1}{2}q+\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}=\frac{R(\lambda)}{K(\lambda)},\tag{5.11a}$$

and

$$v_k=A(w_k)=2p-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{w_k-\lambda_j},\quad k=1,...,N+1,\tag{5.11b}$$

where $K(\lambda)$ is given by (4.1c); $R(\lambda)$ and β_k are also given by (4.1c) but with N replaced by $N+1$. It follows from (5.11a) that w_k, $k=1,...,N+1$, define parabolic coordinates given by the formula

$$\psi_{1j}^2=2\frac{R(\lambda_j)}{K'(\lambda_j)},\quad j=1,...,N,\quad q=2\beta_1-2\alpha_1.\tag{5.12}$$

The equation (5.12) leads to

$$\sum_{j=1}^{N}\psi_{2j}d\psi_{1j}+pdq=\sum_{j=1}^{N+1}v_k dw_k.$$

Therefore the coordinates w_k,v_k are canonically conjugated. From (4.1a) we find

$$\langle\Lambda^k\Psi_1,\Psi_1\rangle=2\beta_{k+1}+2\sum_{i=0}^{N+1}\beta_i\sum_{l=1}^{k+1-i}(-1)^l\prod_{j_1+\cdots+j_l=k+1-i}\alpha_{j_1}\cdots\alpha_{j_l},\tag{5.13}$$

where $j_i\geqslant 1$, $\alpha_k=0$, for $k>N$, and $\beta_k=0$ for $k>N+1$. Combining (5.3), (5.12), and (5.13) gives rise to the expression of u_k in terms of w_k, which are the symmetric functions of w_k, $k=1,...,N+1$. Let

$$A^2(\lambda)+B(\lambda)C(\lambda)=\frac{P(\lambda)}{K(\lambda)},\quad P(\lambda)=\sum_{i=1}^{N+m+3}P_i\lambda^{i-1},\tag{5.14}$$

where P_i, $i=1,...,N+1$, are the integrals of motion for FDHS (5.5). By substituting (5.7) into (5.14), one obtains

$$P_{N+m+3}=1, \quad P_{N+m+2}=\alpha_1, \quad \ldots, \quad P_{N+2}=\alpha_{m+1}, \quad P_{N+1}=\alpha_{m+2}-F_0, \quad \ldots . \tag{5.15}$$

In the exact same way as we did in Sec. IV, we can arrive at the following proposition.

Proposition 8: The Jacobi inversion problem for FDHS (5.5) is of the form

$$\frac{1}{2}\sum_{k=1}^{N+1}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}\,d\lambda=\gamma_i-\delta_{i,N+1}x, \quad i=1,...,N+1.$$

VI. THE THIRD FAMILY OF DYNAMICAL *r*-MATRIX

We now apply the same procedure in Sec. III and IV to the NLEEs (2.16).

A. The factorization of the NLEEs (2.16)

The first x-constrained flow (3.1) for $k=m-1$ of NLEE (2.16) reads

$$\Psi_{1,x}=u_0\Psi_1+\Lambda\Psi_2, \quad \Psi_{2,x}=\left(\Lambda^{m-1}-\sum_{i=1}^{m-1}u_i\Lambda^{i-1}\right)\Psi_1-u_0\Psi_2, \tag{6.1a}$$

$$\begin{pmatrix} 2a_{m-1} \\ -b_{m-1} \\ \vdots \\ -b_{2m-3} \end{pmatrix}=\frac{1}{2}\begin{pmatrix} 2\langle\Psi_1,\Psi_2\rangle \\ \langle\Psi_1,\Psi_1\rangle \\ \vdots \\ \langle\Lambda^{m-2}\Psi_1,\Psi_1\rangle \end{pmatrix}, \tag{6.1b}$$

which leads to

$$u_0=\tfrac{1}{2}\langle\Psi_1,\Psi_2\rangle, \tag{6.2a}$$

$$u_{m-k}=-\langle\Lambda^{k-1}\Psi_1,\Psi_1\rangle+\sum_{i=1}^{k-1}L_{m-i}\langle\Lambda^{k-i-1}\Psi_1,\Psi_1\rangle, \quad k=1,...,m-1. \tag{6.2b}$$

Then it is found by a direct but lengthy computation that

$$u_0=\tfrac{1}{2}\langle\Psi_1,\Psi_2\rangle, \tag{6.3a}$$

$$u_{m-k}=-\sum_{j=1}^{k}\frac{j+1}{2^j}\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle, \quad k=1,...,m-1. \tag{6.3b}$$

By substituting (6.3) into (6.1a), the first x-constrained flow of NLEE (2.16) can be written in canonical FDHS

$$\Psi_{1,x}=\frac{\partial F_0}{\partial\Psi_2}, \quad \Psi_{2,x}=-\frac{\partial F_0}{\partial\Psi_1}, \tag{6.4a}$$

with the Hamiltonian F_0 given by

$$F_0=\frac{1}{2}\langle\Lambda\Psi_2,\Psi_2\rangle+\frac{1}{4}\langle\Psi_1,\Psi_2\rangle^2-\sum_{j=1}^{m}\left(\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=m-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{6.4b}$$

Under (6.3) and (6.4), the t_m-constrained flow (3.2) of NLEE (2.16) can be written as a FDHS

$$\Psi_{1,t_m}=\frac{\partial F_1}{\partial\Psi_2}, \quad \Psi_{2,t_m}=-\frac{\partial F_1}{\partial\Psi_1}, \tag{6.5a}$$

with the Hamiltonian F_1 given by

$$F_1=\frac{1}{2}\langle\Lambda^2\Psi_2,\Psi_2\rangle+\frac{1}{2}\langle\Psi_1,\Psi_2\rangle\langle\Lambda\Psi_1,\Psi_2\rangle-\frac{1}{4}\langle\Psi_1,\Psi_1\rangle\langle\Lambda\Psi_2,\Psi_2\rangle$$
$$-\sum_{j=1}^{m}\left(\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=m-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_{j-1}}\Psi_1,\Psi_1\rangle\langle\Lambda^{l_j+1}\Psi_1,\Psi_1\rangle. \quad (6.5b)$$

We will show later that both (6.4) and (6.5) are FDIHS. The construction of the x- and t_m-constrained flow ensures that if (Ψ_1,Ψ_2) satisfies two commuting FDIHSs (6.4) and (6.5) simultaneously, then u given by (6.3) is a solution of the NLEE (2.16) with $n=m$, namely the NLEE (2.16) with $n=m$ is factorized into two commuting x-FDIHS (6.4) and t_m-FDIHS (6.5).

B. Lax representation for (6.4) and (6.5)

It is found by a direct calculation

$$\tilde{a}_0=\cdots=\tilde{a}_{m-2}=\tilde{b}_0=\cdots=\tilde{b}_{m-3}=0,\quad \tilde{b}_{m-2}=1,\quad \tilde{c}_0=1,$$
$$\tilde{b}_{m-1+k}=-\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_1\rangle,\quad k=0,1,...,$$
$$\tilde{a}_{m-1+k}=(u_0-\tfrac{1}{2}D)\tilde{b}_{m-2+k}=\tfrac{1}{2}\langle\Lambda^k\Psi_1,\Psi_2\rangle,\quad k=0,1,...,$$
$$\tilde{c}_k=-\sum_{i=m-k}^{m}u_i\tilde{b}_{k+i-2}=\sum_{j=1}^{k}\left(\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle,\quad k=1,...,m-1,$$
$$\tilde{c}_{m-1+k}=\tilde{a}_{m-2+k,x}-\sum_{i=1}^{m}u_i\tilde{b}_{m+k+i-3}=\frac{1}{2}\langle\Lambda^k\Psi_2,\Psi_2\rangle,\quad k=1,\ldots.$$

Set

$$M\equiv\lambda^{m-1}\tilde{V}=\begin{pmatrix}A(\lambda) & B(\lambda)\\ C(\lambda) & -A(\lambda)\end{pmatrix}, \quad (6.6a)$$

we have

$$A(\lambda)=\frac{1}{2}\lambda\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{\lambda-\lambda_j},\quad B(\lambda)=\lambda-\frac{1}{2}\lambda\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}, \quad (6.6b)$$

$$C(\lambda)=\lambda^{m-1}+\sum_{k=1}^{m-1}\lambda^{m-1-k}\sum_{j=1}^{k}\left(\frac{1}{2}\right)^j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle+\frac{1}{2}\sum_{j=1}^{N}\frac{\lambda_j\psi_{2j}^2}{\lambda-\lambda_j}. \quad (6.6c)$$

Proposition 9: The Lax representation for the FDHS (6.4) is given by (3.14) with $\bar{U}$ obtained by inserting (6.3) into U in (2.12) and M defined by (6.6). The Lax representation for the FDHS (6.5) is given by (3.15) with $\bar{V}^{(m)}$ obtained by inserting (6.3) into $V^{(m)}$ in (2.14).

C. The third family of dynamical *r*-matrices

With respect to the standard Poisson bracket one obtains

$$\{A(\lambda)A(\mu)\}=\{B(\lambda),B(\mu)\}=0,\quad \{C(\lambda),C(\mu)\}=2g_3(\lambda,\mu)[A(\mu)-A(\lambda)],$$
$$\{A(\lambda),B(\mu)\}=\frac{1}{\mu-\lambda}[\lambda B(\mu)-\mu B(\lambda)],\quad \{B(\lambda),C(\mu)\}=\frac{2\lambda}{\mu-\lambda}[A(\mu)-A(\lambda)],$$

$$\{A(\lambda),C(\mu)\}=\frac{\lambda}{\mu-\lambda}\,[C(\lambda)-C(\mu)]+g_3(\lambda,\mu)B(\lambda), \tag{6.7a}$$

where

$$g_3(\lambda,\mu)=\sum_{k=0}^{m-2}\mu^{m-2-k}\lambda^k+\sum_{l=1}^{m-2}\sum_{k=0}^{m-2-l}\lambda^k\mu^{m-2-l-k}\times\sum_{j=1}^{l}\left(\frac{1}{2}\right)^j(j+1)$$

$$\times\sum_{l_1+\cdots+l_j=l-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle. \tag{6.7b}$$

Then (6.7) gives rise to the following proposition.

Proposition 10: The classical Poisson structure for FDHS (6.4) and (6.5) is given by the linear r-matrix form (3.19) with the following r-matrix:

$$r_{12}(\lambda,\mu)=\frac{\mu}{\mu-\lambda}\,P-G-g_3(\lambda,\mu)S,\quad G=\sigma_+\otimes\sigma_-,\quad S=\sigma_-\otimes\sigma_-,$$

$$r_{21}(\lambda,\mu)=Pr_{12}(\lambda,\mu)P=r_{12}(\mu,\lambda). \tag{6.8}$$

The equation (6.8) presents the third family of dynamical r-matrices $r_{12}(\lambda,\mu)$ for $m=3,4,\ldots$. We have

$$A^2(\lambda)+B(\lambda)C(\lambda)=\lambda^m+\lambda\sum_{i=1}^{N}\frac{F^{(i)}}{\lambda-\lambda_i}, \tag{6.9}$$

where

$$F^{(i)}=\frac{1}{2}\,\lambda_i\psi_{2i}^2-\frac{1}{2}\,\lambda_i^{m-1}\psi_{1i}^2+\frac{1}{2}\,\psi_{1i}\psi_{2i}\langle\Psi_1,\Psi_2\rangle-\frac{1}{4}\,\psi_{2i}^2\langle\Psi_1,\Psi_1\rangle$$

$$-\sum_{k=1}^{m-1}\lambda_i^{m-k-1}\sum_{j=1}^{k}\left(\frac{1}{2}\right)^{j+1}\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle\psi_{1i}^2$$

$$+\frac{1}{4}\sum_{k\neq i}\frac{\lambda_k(\psi_{1i}\psi_{2k}-\psi_{1k}\psi_{2i})^2}{\lambda_k-\lambda_i},\quad i=1,\ldots,N.$$

The classical Poisson structure (3.19) and (6.8) guarantees that independent integrals of motion for (6.4) and (6.5) $F^{(i)}$, $i=1,\ldots,N$ are in involution. Therefore (6.4) and (6.5) are FDIHSs.

In the similar way as for Proposition 3, we can prove the following proposition.

Proposition 11: For the linear r-matrix algebra (3.19) and (6.8) the dynamical Yang–Baxter equations (3.22) with the $d_{i,j}$ and f given by following formulas are valid:

$$d_{ij}(\lambda,\mu)=\frac{\mu}{\mu-\lambda}\,P_{ij}-G_{ij}-g_3(\lambda,\mu)S_{ij},\quad d_{ji}(\lambda,\mu)=d_{ij}(\lambda,\mu),\quad j>i,\quad i,j=1,2,3,$$

$$f(\lambda,\mu,\nu)=-2\sum_{k=0}^{m-3}\sum_{l=0}^{m-3-k}\lambda^k\mu^l\nu^{m-3-l-k}-2\sum_{l=1}^{m-3}\sum_{k=0}^{m-3-l}\sum_{i=0}^{m-3-l-k}\lambda^k\mu^i\nu^{m-3-l-k-i}$$

$$\times\sum_{j=1}^{l}\left(\frac{1}{2}\right)^j(j+1)^2\sum_{l_1+\cdots+l_j=l-j}\langle\Lambda^{l_1}\Psi_1,\Psi_1\rangle\cdots\langle\Lambda^{l_j}\Psi_1,\Psi_1\rangle.$$

D. The Jacobi inversion problem for NLEE (2.16)

We introduce canonically conjugated separation variables w_k, v_k, $k=1,...,N$ as

$$B(\lambda)=\lambda\left[1-\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}^2}{\lambda-\lambda_j}\right]=\lambda\frac{R(\lambda)}{K(\lambda)}, \tag{6.10a}$$

and

$$v_k=\frac{A(w_k)}{w_k}=\frac{1}{2}\sum_{j=1}^{N}\frac{\psi_{1j}\psi_{2j}}{w_k-\lambda_j}, \quad k=1,...,N, \tag{6.10b}$$

where $K(\lambda)$, $R(\lambda)$ are given by (4.1c). It follows from (6.10a) that w_k, $k=1,...,N$, define elliptic coordinates given by the formula

$$\psi_{1j}^2=-2\frac{R(\lambda_j)}{K'(\lambda_j)}, \quad j=1,...,N. \tag{6.11}$$

We have from (6.10) that

$$\langle\Lambda^k\Psi_1,\Psi_1\rangle=-2\beta_{k+1}-2\sum_{i=0}^{N}\beta_i\sum_{l=1}^{k+1-i}(-1)^l\prod_{j_1+\cdots+j_l=k+1-i}\alpha_{j_1}\cdots\alpha_{j_l}. \tag{6.12}$$

One obtains from (6.3a), (6.4), and (6.10a) that

$$u_0=\frac{\beta_{N,x}}{2\beta_N}. \tag{6.13}$$

Equations (6.3b), (6.12), and (6.13) imply that the expression of u_k, $k=0,...,m-1$, in terms of w_k are the symmetric functions of w_k, $k=1,...,N$.

Then in the same way as in Sec. IV, we obtain the following proposition.

Proposition 12: The Jacobi inversion problem for FDHS (6.4) is of the form

$$\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}d\lambda=\gamma_i+2\delta_{i,N}x, \quad i=1,...,N, \tag{6.14}$$

the Jacobi inversion problem for FDHS (6.5) is

$$\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}d\lambda=\gamma_i-2\delta_{i,N}\alpha_1 t_m+2\delta_{i,N-1}t_m, \quad i=1,...,N, \tag{6.15}$$

and the Jacobi inversion problem for NLEE (2.16) with $n=m$ is of the form

$$\sum_{k=1}^{N}\int^{w_k}\frac{\lambda^{i-1}}{\sqrt{P(\lambda)K(\lambda)}}d\lambda=\gamma_i+2\delta_{i,N}(x-\alpha_1 t_m)+2\delta_{i,N-1}t_m, \quad i=1,...,N. \tag{6.16}$$

The Jacobi inversion problem (4.14) and (6.16) are almost the same; however, the u of (2.9) and u of (2.16) are in terms of w_k, $k=1,...,N$, in different way.

VII. CONCLUSION

By directly using the constrained flows of soliton equations, we have constructed three families of FDIHSs and their dynamical r-matrices. In this way explicit expressions for the FDIHSs and for the dynamical r-matrices as well as for the dynamical Yang–Baxter equations arise naturally.

Since Lax representation for x- and t_n-constrained flow of soliton equation are naturally deduced from the adjoint representation of the spectral problem, the Jacobi inversion problem for

x- and t_n-constrained flow can always be found by means of the Lax representation. The factorization of soliton equation by x- and t_n-constrained flow allows us to obtain the Jacobi inversion problem for soliton equation by combining the Jacobi inversion problems for x- and t_n-constrained flow. This implies that some kind of solution for soliton equation can be found by the method of separation of variables.

ACKNOWLEDGMENTS

This work was supported by the Chinese Basic Research Project ''Nonlinear Science.'' The author wishes to express his gratitude to the Department of Mathematics, Sophia University, Professor Sunao Ouchi, and Professor Ken-ichi Shinoda for warm hospitality.

[1] O. Babelon and C. M. Viallet, Phys. Lett. B **237**, 411 (1990).
[2] J. Avan and M. Talon, Phys. Lett. B **303**, 33 (1993).
[3] J. C. Eilbeck, V. Z. Enol'skii, V. B. Kuznetsov, and D. V. Leykin, Phys. Lett. A **180**, 208 (1993).
[4] E. K. Sklyanin, St. Petersburg J. Math **6**, 227 (1994).
[5] A. V. Tsiganov, J. Phys. A **27**, 6759 (1994).
[6] P. P. Kulish, S. Rauch-Wojciechowski, and A. V. Tsiganov, J. Math. Phys. **37**, 3463 (1996).
[7] Yunbo Zeng and J. Hietarinta, J. Phys. A **29**, 5241 (1996).
[8] Yunbo Zeng, Phys. Lett. A **216**, 26 (1996).
[9] Yunbo Zeng and Yishen Li, J. Phys. A **26**, L273 (1993).
[10] M. Antonowicz and A. P. Fordy, *Nonlinear Evolution Equations and Dynamical Systems (NEEDS'87)*, edited by J. Leon (World Scientific, Singapore, 1988).
[11] E. K. Sklyanin, J. Sov. Math. **47**, 2473 (1989).
[12] E. K. Sklyanin, Prog. Theor. Phys. Suppl. **118**, 35 (1995).
[13] V. B. Kuznetsov, J. Math. Phys. **33**, 3240 (1992).
[14] J. Harnad and P. Winternitz, Commun. Math. Phys. **172**, 263 (1995).
[15] Yunbo Zeng, Phys. Lett. A **160**, 541 (1991).
[16] Yunbo Zeng, Physica D **73**, 171 (1994).
[17] Yunbo Zeng, J. Phys. A **30**, 3719 (1997).
[18] A. D. Dubrovin, Russian Math. Surveys **36**, 11 (1981).
[19] A. C. Newell, *Solitons in Mathematics and Physics* (SIAM, Philadelphia, 1985).
[20] M. Antonowicz and S. Rauch-Wojciechowski, J. Math. Phys. **33**, 2115 (1992).
[21] Yunbo Zeng and Yishen Li, J. Math. Phys. **31**, 2835 (1990).
[22] V. I. Arnold, *Mathematical Method of Classical Mechanics* (Springer, New York, 1978).

Timelike geodesic motions within the general relativistic gravitational field of the rigidly rotating disk of dust

M. Ansorg[a)]
Friedrich-Schiller-Universität Jena, Theoretisch-Physikalisches Institut, Max-Wien-Platz 1, D-07743 Jena, Germany

(Received 15 June 1998; accepted for publication 3 August 1998)

The general relativistic motion of a test particle near a rigidly rotating disk of dust is investigated. Circular orbits within the plane of the disk (centered on the rotation axis) are special cases of the geodesic motion. One finds that there is always a (stable or unstable) circular orbit for positive angular momentum and a given radius. However, for sufficiently relativistic disks there are regions within the plane of the disk in which a particle with negative angular momentum cannot follow a circular path. If the disk is still more strongly relativistic, then one finds circular orbits with negative energies of arbitrary magnitude. Within the theoretical construction of the Penrose effect, this property can be used to produce arbitrarily high amounts of energy. The study of Hamiltonian mechanics forms another topic of this article. It turns out that the stochastic behavior of the geodesics is related to the position of the region containing all the crossing points of the particle through the plane of the disk. If this region contains points lying inside the disk as well as points outside, the geodesic motion shows highly stochastic behavior. However, if the crossing region is completely inside or outside the disk, the motion proves to be nearly integrable. In these cases the corresponding Hamiltonian system is close to an integrable system of the so-called Liouville class. © *1998 American Institute of Physics.* [S0022-2488(98)03711-6]

I. INTRODUCTION

Einstein's general theory of relativity describes gravitation as a geometric property of the four-dimensional manifold of space and time. Particles moving merely under the influence of gravitation follow the so-called timelike geodesic paths. The investigation of these curves helps in the understanding of the geometrical structure of the manifold in question. Furthermore, accretion processes in which matter is drawn into a central object (e.g., a black hole or a disklike formation) may be approximated by the assumption that the inflowing matter follows the timelike geodesics.

The famous Kerr black hole was the first rotating object for which the relativistic field equations were solved. The timelike geodesic motions near a Kerr black hole have been studied in detail (see, e.g., Ref. 1). For them the corresponding Hamiltonian–Jacobi equation is separable,[2] and the equations of motions are integrable by quadratures. As a consequence, in phase space the trajectories lie within surfaces diffeomorphic to tori.

The general relativistic gravitational field created by a rigidly rotating disk of dust was first studied numerically in 1971 by Bardeen and Wagoner.[3] The global analytic solution of Einstein's field equations for this object was found in 1995 by Neugebauer and Meinel.[4] Their explicit expressions for the metric coefficients allow a direct numerical implementation of the geodesic equations.

In this paper the bounded motions around a rigidly rotating disk of dust are investigated. It turned out that these motions are not integrable by quadratures. In phase space they occupy regions of more complicated structure. However, for a considerable fraction of the set of all bounded timelike geodesic motions, particular properties can be found.

a)Electronic mail: Ansorg@hpfs1.tpi.uni-jena.de

 © 1998 American Institute of Physics

A. The metric

Einstein's field equations for a rigidly rotating disk of dust can be reduced to a single nonlinear complex partial differential equation—the so-called Ernst equation—for which a boundary value problem has to be solved.[5,6] Neugebauer and Meinel succeeded in solving this problem by means of the inverse scattering method which is a technique from the soliton theory.

Writing the metric in Weyl–Papapetrou coordinates (ρ,ζ,φ,t) in the form (a comma in an index represents a partial derivative, a semicolon stands for a covariant derivative)

$$(g_{ij})=\begin{pmatrix} e^{2(k-U)} & 0 & 0 & 0 \\ 0 & e^{2(k-U)} & 0 & 0 \\ 0 & 0 & -a^2e^{2U}+\rho^2e^{-2U} & -ae^{2U} \\ 0 & 0 & -ae^{2U} & -e^{2U} \end{pmatrix} \quad \text{with} \quad g_{ij,\varphi}=0=g_{ij,t},$$

the field equations turn out to be equivalent to the Ernst equation

$$\Re f\left(f_{,\rho\rho}+f_{,\zeta\zeta}+\frac{f_{,\rho}}{\rho}\right)=f_{,\rho}^2+f_{,\zeta}^2$$

for the Ernst potential f which is defined by

$$f=e^{2U}+ib \quad \text{with} \quad b_{,\zeta}=\frac{e^{4U}}{\rho}\,a_{,\rho}, \quad b_{,\rho}=-\frac{e^{4U}}{\rho}\,a_{,\zeta}.$$

The remaining metric function k can be calculated from the functions U and a by quadratures.

Neugebauer and Meinel found the Ernst potential for the rigidly rotating disk of dust in terms of ultraelliptic functions.[4] In their expressions, the Ernst potential depends on a parameter μ, which is related to the angular velocity Ω and the radius ρ_0 of the disk by

$$\mu=2\Omega^2\rho_0^2e^{-2V_0} \quad \text{with} \quad V_0(\mu)=U(\rho=0,\zeta=0;\mu).$$

The parameter μ lies within the interval $(0,\mu_0)$ with $\mu_0=4.629\,66...$. For $\mu\ll1$, one obtains the Newtonian limit of the Maclaurin disk; $\mu\to\mu_0$ and $\rho_0\to0$ yield the extreme Kerr solution. The disk creates an ergosphere [i.e., a region within which the metric function $(-e^{2U})$ is positive] for $\mu>\mu_e\approx1.69$.

In what follows, units are used where the radius ρ_0 as well as the velocity of light c are equal to 1.

B. The Hamiltonian system of the geodesic equations

For axisymmetric stationary space–times the general Hamiltonian system

$$\mathscr{H}=\frac{1}{2}\,g^{ij}p_ip_j \quad \text{with} \quad p_i=g_{ij}\dot{x}^i, \quad \left(=\frac{d}{d\tau}, \quad \tau\text{:proper time}\right)$$

for the timelike geodesic equations

$$\ddot{x}^i+\Gamma^i_{kl}\dot{x}^k\dot{x}^l=0, \quad \dot{x}_k\dot{x}^k=-1$$

can be reduced to a conservative Hamiltonian of two degrees of freedom of the following form:

$$\mathscr{H}=\frac{1}{2}\,e^{2(U-k)}(p_\rho^2+p_\zeta^2)+\frac{1}{2}\left[1-\frac{1}{\rho^2}\,(L^2g_{tt}+2LEg_{\varphi t}+E^2g_{\varphi\varphi})\right]. \tag{1}$$

Here, $L=p_\varphi=\text{const}$ is called the (ζ component of the) angular momentum of the particle and $E=-p_t=\text{const}$ its energy.

The above Hamiltonian is invariant under a simultaneous change of the signs of L and E. Now, these signs are fixed by the condition $\dot{t}>0$ and this condition is then satisfied along the whole trajectory.

The Hamiltonian (1) is arranged such that $\mathscr{H}=0$ holds along the trajectory of the considered test particle. Therefore a restricting condition for the region in which the motion takes place is given by

$$0\leqslant e^{2(U-k)}(p_\rho^2+p_\zeta^2)=\left[\frac{1}{\rho^2}\,(L^2 g_{tt}+2LEg_{\varphi t}+E^2 g_{\varphi\varphi})-1\right]. \tag{2}$$

This region of motion has the following properties:

(1) The region is bounded for $E^2<1$. For $E^2\geqslant 1$ there are unbounded areas allowing an escape of the test particle to infinity. But also, in the case of $E^2\geqslant 1$, there may be additional bounded regions which force the particle to stay inside.
(2) Particles with $L\neq 0$ may not cross the rotation axis.
(3) The momenta p_ρ and p_ζ as well as the velocities $\dot{\rho}$ and $\dot{\zeta}$ vanish at the rim of the considered region. Furthermore, at the boundary of the region the acceleration vector $(\ddot{\rho},\ddot{\zeta})$ is perpendicular to the curve of the boundary and points inwards.
(4) Because of the property $g_{ij}(\rho,-\zeta)=g_{ij}(\rho,\zeta)$, the regions of motion are symmetric with respect to reflection in the plane of the disk.

Illustrative examples for bounded regions of motion can be seen in Figs. 9, 10, 18, and 20.

II. CIRCULAR ORBITS

Motions within the plane of the disk are described by a Hamiltonian with merely one degree of freedom. By the conditions $\dot{\rho}=0=\ddot{\rho}$, circular orbits with centers on the rotation axis are fixed. The following numerical results about circular orbits contain in particular Meinel's and Kleinwächter's analytic results about circular orbits at the rim of the disk.[7] The above conditions $\dot{\rho}=0=\ddot{\rho}$ lead to the equations

$$\begin{aligned} L^2 g_{tt}(\rho,0)+2LEg_{\varphi t}(\rho,0)+E^2 g_{\varphi\varphi}(\rho,0)&=\rho^2,\\ L^2 g_{tt,\rho}(\rho,0)+2LEg_{\varphi t,\rho}(\rho,0)+E^2 g_{\varphi\varphi,\rho}(\rho,0)&=2\rho, \end{aligned} \tag{3}$$

which serve to determine the parameter functions $L_\pm=L_\pm(\rho)$ and $E_\pm=E_\pm(\rho)$. These functions yield for a circular orbit of given radius ρ, the associated pairs of angular momentum and energy. In general, there are two such pairs, denoted by $(L_+(\rho),E_+(\rho))$ and $(L_-(\rho),E_-(\rho))$ where (L_+,E_+) refers to an orbit of positive angular momentum and similarly (L_-,E_-) to an orbit with negative angular momentum. It turns out that the functions $L_+(\rho)$ and $E_+(\rho)$ exist for arbitrary choice of $\rho\geqslant 0$. However, for sufficiently large values of the parameter μ there are no corresponding (L_-,E_-) pairs for ρ values within a certain interval $(\tilde{\rho}_1,\tilde{\rho}_2)$.

In Figs. 1–5, the functions $L_\pm=L_\pm(\rho)$ and $E_\pm=E_\pm(\rho)$ can be seen for different values of the parameter μ. A discussion of these pictures follows.

(1) The pictures for (L_+,E_+) are similar for all values of μ (see Fig. 1). There is a monotonous growth of the functions L_+ and E_+ for $\rho<1$. At $\rho=1$ (here $E_+=1$) both functions turn back and decrease monotonously until they reach another turning point at $\rho=\rho'$. For still greater values of the parameter ρ, the functions L_+ and E_+ grow monotonously again. The turning points are characterized by the conditions $dL_+/d\rho=0=dE_+/d\rho$. As the radius $\rho\to\infty$, E_+ tends to 1 and L_+ to $+\infty$.

Circular orbits with a radius $\rho\in(1,\rho')$ are unstable. Stable circular orbits are those with a radius $\rho\in[0,1)$ or $\rho\in(\rho',\infty)$; the remaining circular orbits with radius 1 or ρ' are marginally unstable.

For circular orbits, the condition

$$\dot{\varphi}=\frac{dE}{dL}\,\dot{t} \tag{4}$$

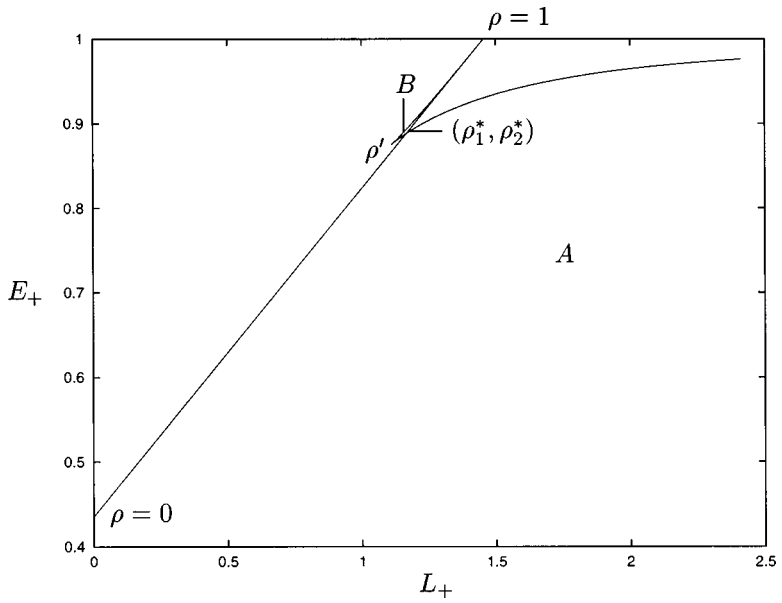

FIG. 1. The functions $L_+(\rho)$ and $E_+(\rho)$ for $\mu = 1.61$.

holds. In all sections of the graphs of the functions L_+ and E_+, the function $E_+ = E_+(L_+)$ grows monotonously. Hence, all circular orbits with positive angular momenta possess positive angular velocity. Furthermore, circular orbits with radii $\rho < 1$ are just those paths along which the dust particles of the disk move. Their four velocity (u^i) is given by $(u^i) = (0,0,\dot{\varphi},\dot{t}) = e^{-V_0}(0,0,\Omega,1)$ (see Ref. 8) from which $dE/dL = \Omega = \text{const}$ follows. Thus, the graph $E_+(L_+)$ is a straight line for $\rho \in [0,1]$.

The region A contains all (L,E) pairs for which no motion of a particle is possible, i.e., either the restricting condition (2) cannot be satisfied or the motion possesses a negative $\dot{t}$. If one chooses for a (not necessarily circular) motion, an (L,E) pair inside the small region B, then there exist two separate compact regions in which bounded motions are possible. At the intersection point of the stable and unstable parts of the (L_+, E_+) graphs these regions degenerate into two

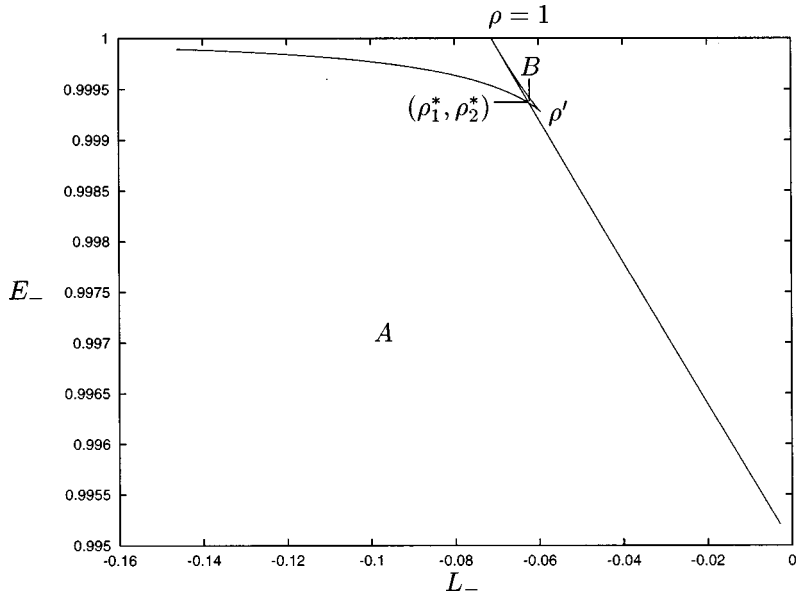

FIG. 2. The functions $L_-(\rho)$ and $E_-(\rho)$ for $\mu = 0.01$.

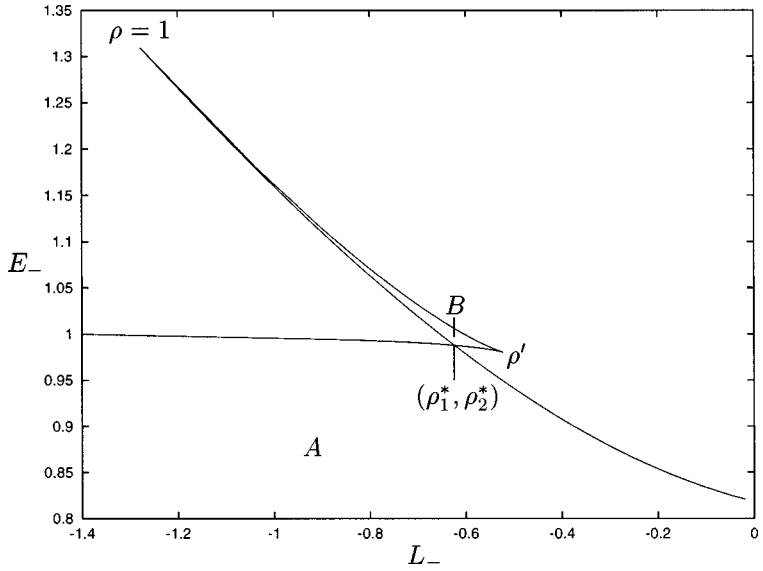

FIG. 3. The functions $L_-(\rho)$ and $E_-(\rho)$ for $\mu = 0.4$.

separate stable circular orbits with radii $\rho_1^* < 1$ and $\rho_2^* > 1$. In Fig. 6, it can be seen how the radii ρ_1^*, ρ_2^*, and ρ' depend on the parameter μ.

(2) In Figs. 2–5, the parameter functions $L_-(\rho)$ and $E_-(\rho)$ are displayed for different values of the parameter μ. For $0 < \mu < 1/2$, the pictures are similar to the graphs of the functions L_+ and E_+. The region B extends now to E_--values greater than 1. Thus, there are stable as well as unstable circular orbits, and furthermore, extended compact regions of motions with energies greater than 1 and corresponding negative angular momenta. From Eq. (4) and the slopes of the functions $E_- = E_-(L_-)$ in Figs. 2 and 3, one concludes that all particles moving along circular orbits with negative angular momenta have negative angular velocity.

(3) As μ approaches the value $\frac{1}{2}$, the region B grows and is unbounded for $\mu \geqslant 1/2$. For radii $\rho \in (\tilde{\rho}_1, \tilde{\rho}_2)$ there are no circular orbits with negative angular momentum. As can be seen in Fig. 4, for sufficiently large values of μ ($\mu > \mu_{\dot{\varphi}} \approx 0.7088$) there are small intervals $[0, \rho_{\dot{\varphi}})$ in which the functions $E_-(L_-)$ grow monotonously. Hence, circular orbits with radii $\rho < \rho_{\dot{\varphi}}$ possess positive

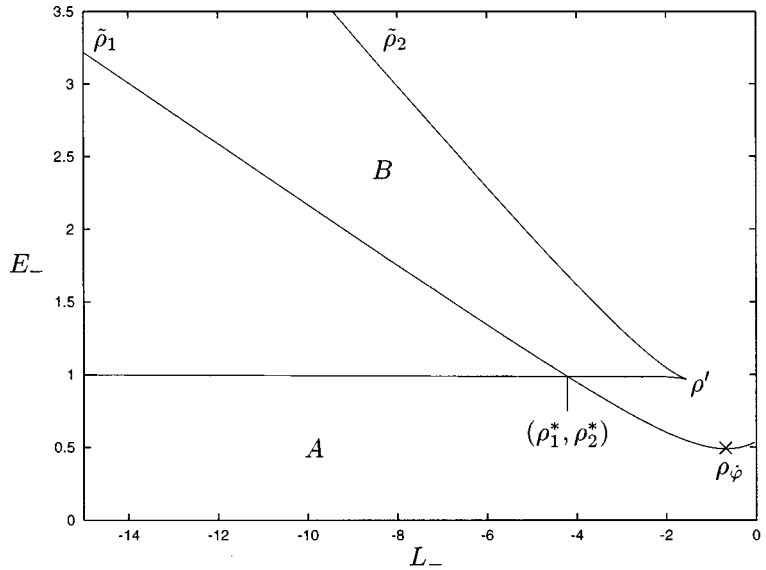

FIG. 4. The functions $L_-(\rho)$ and $E_-(\rho)$ for $\mu = 1.21$.

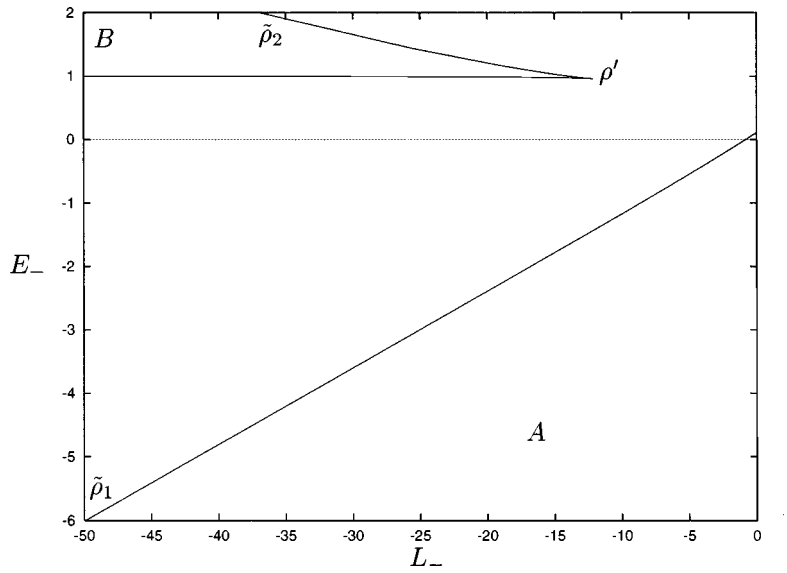

FIG. 5. The functions $L_-(\rho)$ and $E_-(\rho)$ for $\mu=3.5$.

angular velocities in spite of having negative angular momenta. At the radius $\rho_{\dot{\varphi}}$ a particle may remain at rest (in the chosen coordinate system). Such a particle has the smallest energy possible of all motions around the disk.

(4) Finally, in the range $\mu_e<\mu<\mu_0$ (see Fig. 5), there are stable circular orbits with negative energies. As the slope of $E_-(L_-)$ is positive for $\rho<\tilde{\rho}_1$, circular orbits with radii in this range have positive angular velocities. The μ dependencies of the radii $\{\rho_1^*,\rho_2^*,\rho',\tilde{\rho}_1,\tilde{\rho}_2,\rho_{\dot{\varphi}}\}$ as well as those of the radii $\rho_1^{(E)}$ and $\rho_2^{(E)}$, at which the function $E_-=E_-(\rho)$ reaches the value $E_-(\rho_i^{(E)})=1$, are displayed in Fig. 7.

(5) As $\mu\to 0$, the curves $L_\pm(\rho)$ and $E_\pm(\rho)$ tend to the corresponding graphs for the Maclaurin disk (the Newtonian limit of the relativistic disk). Since negative and positive angular momenta are equivalent in Newtonian theory, these graphs possess a reflectional symmetry with respect to the axis $L=0$.

Circular orbits around a Kerr black hole are characterized by the functions (see, e.g., Ref. 1)

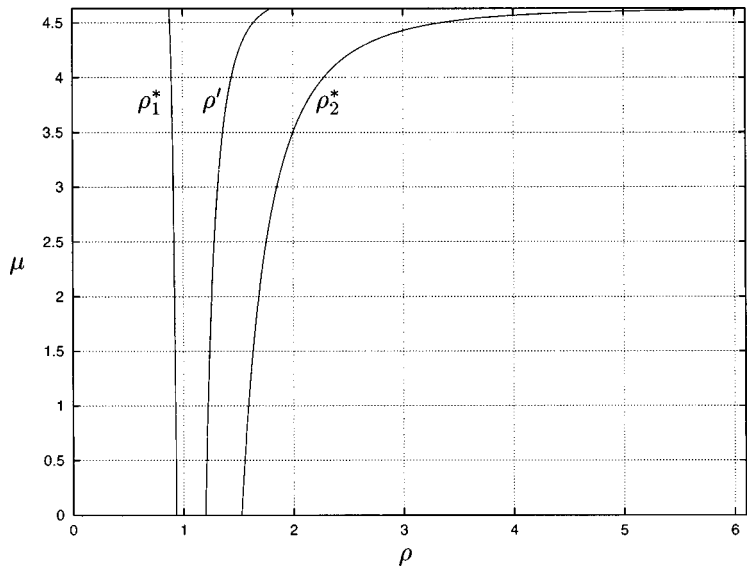

FIG. 6. The μ dependency of the radii ρ_1^*, ρ_2^*, and ρ' for the functions L_+ and E_+.

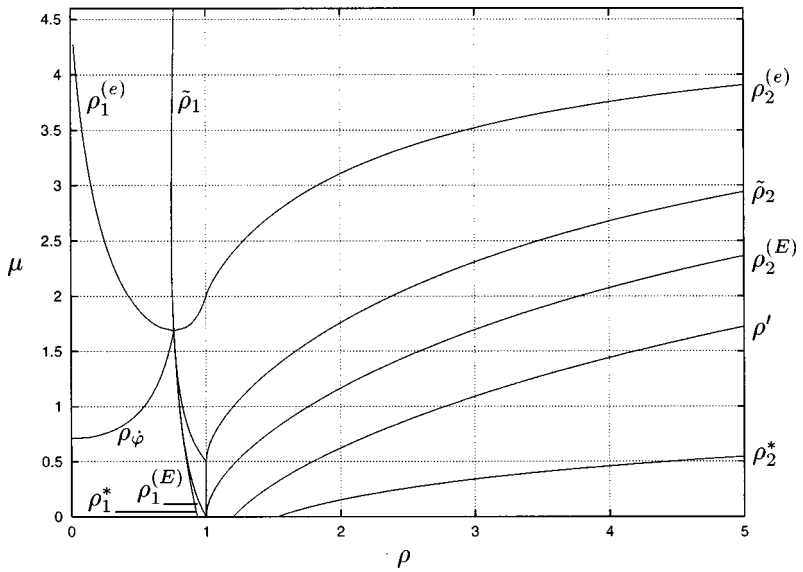

FIG. 7. The μ dependency of the radii $\{\rho_1^*,\rho_2^*,\rho',\tilde{\rho}_1,\tilde{\rho}_2,\rho_{\dot{\varphi}},\rho_1^{(E)},\rho_2^{(E)}\}$ for the functions L_- and E_-. Additionally, the radii $\rho_1^{(e)}$ and $\rho_2^{(e)}$, at which the ergosphere intersects the plane of the disk, are plotted.

$$E_\pm=\frac{r^{3/2}-2mr^{1/2}\pm am^{1/2}}{r^{3/4}(r^{3/2}-3mr^{1/2}\pm 2am^{1/2})^{1/2}},\quad L_\pm=\pm\frac{m^{1/2}[r^2\mp 2a(mr)^{1/2}+a^2]}{r^{3/4}(r^{3/2}-3mr^{1/2}\pm 2am^{1/2})^{1/2}},$$

where $\rho=\sqrt{r^2-2mr+a^2}$, m denotes the mass, and a the specific angular momentum of the black hole. For vanishing denominator in the above expressions (this may happen as ρ reaches the value $\tilde{\rho}$), $L_\pm\to\pm\infty$ and $E_\pm\to+\infty$. Circular orbits with radii less than $\tilde{\rho}$ do not exist. For $\rho>\tilde{\rho}$, the $(L_\pm,E_\pm)$ curves are similar to those of Fig. 5 for $\rho>\tilde{\rho}_2$. Again there is a ρ' which separates unstable (for $\rho<\rho'$) and stable orbits (for $\rho>\rho'$). Furthermore, for circular orbits within the Kerr metric, the signs of angular momentum and velocity always coincide.

Generally one finds that the qualitative behavior of circular motions at sufficiently large radii is similar for the Kerr black hole and the disk. However, circular motions at small radii are quite different.

III. THE PENROSE EFFECT

The Penrose effect is a theoretical construction to produce energy. A test particle (with angular momentum L_0 and energy E_0) is allowed to fall from infinity (hence $E_0>1$) into the ergosphere of the disk. There, it splits into two particles with angular momenta L_1,L_2 and energies E_1,E_2. The total angular momentum and energy are preserved, i.e.,

$$L_0=L_1+L_2,\quad E_0+E_1+E_2.$$

If one of the created particles possesses negative energy (say $E_1<0$), then the other one has an energy $E_2>E_0$. The latter particle escapes to infinity where its higher energy can be utilized.

If a large amount of energy were to be drawn from the disk by this method, the particles remaining within the ergosphere would have a considerable repercussion on the disk. Hence, in the following considerations, when one speaks about arbitrarily high energies, one still means that these energies are negligible compared to the energy of the disk.

A simple way to produce energy would seem to be the following. A particle with an energy E_0 which is slightly larger than 1 falls from infinity into the ergosphere to the point $(\tilde{\rho}_1,0)$ (see Figs. 5 and 7), where it separates into two particles. After the separation, one of the created particles follows the stable circular orbit at $\tilde{\rho}_1$ and has the energy $E_1=E_-(\tilde{\rho}_1)=-\infty$. Hence, the escaping particle possesses an energy $E_2=\infty$. But this does not work since the outgoing particle cannot satisfy the condition (2) at $(\tilde{\rho}_1,0)$, i.e., it never reaches this point. However, it turns out that

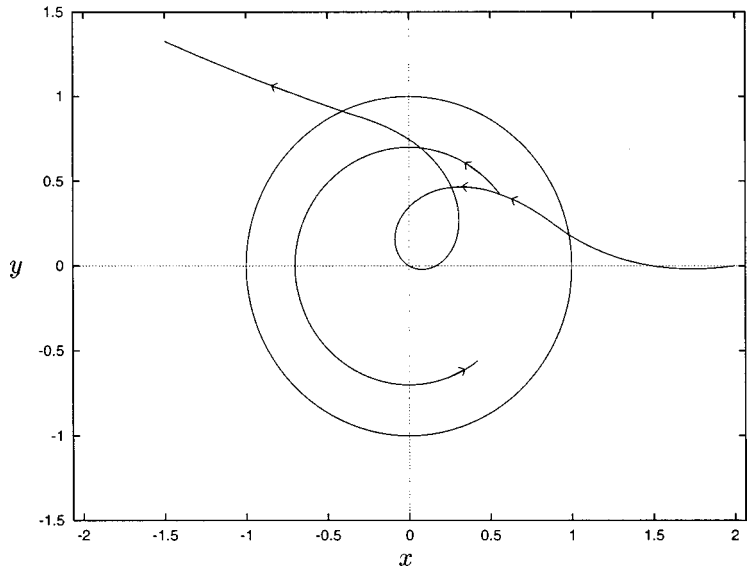

FIG. 8. The Penrose effect. At a value $\mu=2.2$, the parameters $E_0=1.1$ and $\rho_C=0.7$ have been chosen. The energy of the escaping particle is 0.16 greater than the energy of the infalling particle. The curves of the particles can be seen, as the particles move within the plane of the disk (the disk is represented by a circle).

an arbitrary profit of energy can be obtained by giving the infalling particle a high energy $E_0>1$. In the following example, the pair (L_0,E_0) should be chosen above the curve of unstable circular orbits with negative angular momenta (see Fig. 5).

(1) It turns out that it is possible for a particle with an angular momentum L_0 to fall inward toward $(\rho_C,0)$, at which point the corresponding circular orbit has angular momentum $L_-(\rho_C)=L_0$. If the infalling particle remains within the plane of the disk, it suffices to check that (2) holds at $(\rho_C,0)$. In the diagram (E,h), the graph of the function

$$h(E)=\frac{1}{2}\left(1-\frac{1}{\rho_C^2}\left[L_0^2 g_{tt}(\rho_C,0)+2L_0 E g_{\varphi t}(\rho_C,0)+E^2 g_{\varphi\varphi}(\rho_C,0)\right]\right)$$

shows a parabola with negative curvature (since $g_{\varphi\varphi}>0$). Furthermore, $h[E_-(\rho_C)]=0$ since this just represents the first of the conditions (3) for the circular orbit at ρ_C with negative angular momentum. Another conclusion is

$$\frac{dh}{dE}[E_-(\rho_C)]=-\frac{1}{\rho_C^2}\left[L_0 g_{\varphi t}(\rho_C,0)+E_-(\rho_C)g_{\varphi\varphi}(\rho_C,0)\right]=-\dot{t}_C<0,$$

where $\dot{t}_C$ denotes the fourth component of the four velocity of the circular orbit at ρ_C. Thus, h is negative for $E>E_-(\rho_C)$ and hence in particular for $E=E_0$. The infalling particle reaches the point $(\rho_C,0)$.

(2) At the point $(\rho_C,0)$ the particle splits. One of the created particles follows the circular path at ρ_C with negative angular momentum. Hence, it possesses the energy $E_1=E_-(\rho_K)$. The angular momentum of the escaping particle vanishes, whereas its energy $E_2=E_0-E_1$ is greater than E_0 if $E_-(\rho_C)$ is negative. Again, it can be shown that the escaping particle indeed reaches the point $(\rho_C,0)$.

If one goes along the curve of unstable circular orbits (Fig. 5) to higher and higher energies, the corresponding $E_-(\rho_C)$ will be of higher and higher magnitude. Hence, an arbitrary high amount of energy can be obtained if the infalling particle carries a high energy (and a corresponding angular momentum L_0). An example of the described Penrose effect is displayed in Fig. 8.

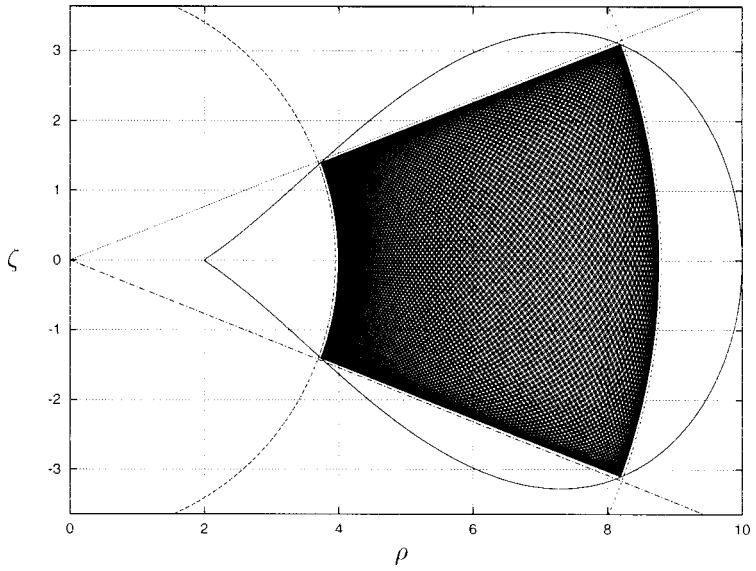

FIG. 9. Example of a geodesic motion in the (ρ,ζ) space with the values $\mu=1$, $L=-1.32$, $E=0.8$, $\rho_1=2$, $\rho_2=10$.

IV. HAMILTONIAN MECHANICS FOR THE GEODESICS

In this section the dynamical properties of bounded geodesics in phase space are investigated. Section IV A treats integrable motions. The regions in phase space occupied by the trajectories are discussed. Furthermore, a class of integrable Hamiltonians—the so-called Liouville class (see, e.g., Ref. 9)—is introduced. In Sec. IV B, this class will turn out to be relevant for the discussion of the nearly integrable geodesics and their properties. Finally, Sec. IV C treats the remaining geodesics which show highly stochastic behavior.

A. Integrable Hamiltonian systems

A famous theorem by Liouville states that in phase space, the bounded trajectories of an integrable Hamiltonian run on geometrical objects diffeomorphic to tori (for a proof see, e.g., Ref. 10). By means of a canonical transformation, it is possible to introduce angle and action variables (ϕ_j, I_j) in which the Hamiltonian merely depends on the action variables I_j. The corresponding Hamiltonian equations

$$\frac{dI_i}{d\tau}=-\frac{\partial \mathscr{H}(I_j)}{\partial \phi_i}=0, \quad \frac{d\phi_i}{d\tau}=\frac{\partial \mathscr{H}(I_j)}{\partial I_i}=\text{const}$$

are easily integrated; one gets $I_j=\text{const}$ and $\phi_j=\omega_j\tau+\phi_{j0}$, where ω_j and ϕ_{j0} are constants.

For an integrable Hamiltonian with two degrees of freedom, one can describe the situation as follows. In the four-dimensional phase space, the action variables $I_j (j=1,2)$ determine a set of surfaces diffeomorphic to two-dimensional tori. Each bounded motion of the considered integrable Hamiltonian runs on one of these surfaces. The angle variables yield a coordinate net on this surface, and the motion within this net proceeds linearly in time. If the ratio ω_1/ω_2 is chosen to be rational, the motion is periodic on the surface; for a ratio $\omega_1/\omega_2=m/n$ $(m,n\in\mathbb{N}$; relatively prime), the path returns to its initial point after a time period of $\tau=2\pi m/\omega_1$. Since the rational numbers are dense in the set of real numbers, the surfaces $I_j=\text{const}$ on which the motion is periodic form a dense subset in the set of all surfaces $I_j=\text{const}$. For the nonperiodic motions with an irrational ratio of the frequencies ω_1 and ω_2, the trajectory fills uniquely the whole surface.

Trajectories in phase space are studied by means of the so-called Poincaré surfaces of section. One chooses a two-dimensional surface in the phase space and labels its two sides (say left and right). Then, the set of crossing points of the trajectory through this surface is investigated. Only intersections are considered for which the trajectory runs through the surface in a particular sense (say from left to right). For integrable Hamiltonians, one gets a closed curve if the ratio of the frequencies is irrational. Otherwise, the Poincaré surfaces of section contain a finite number of

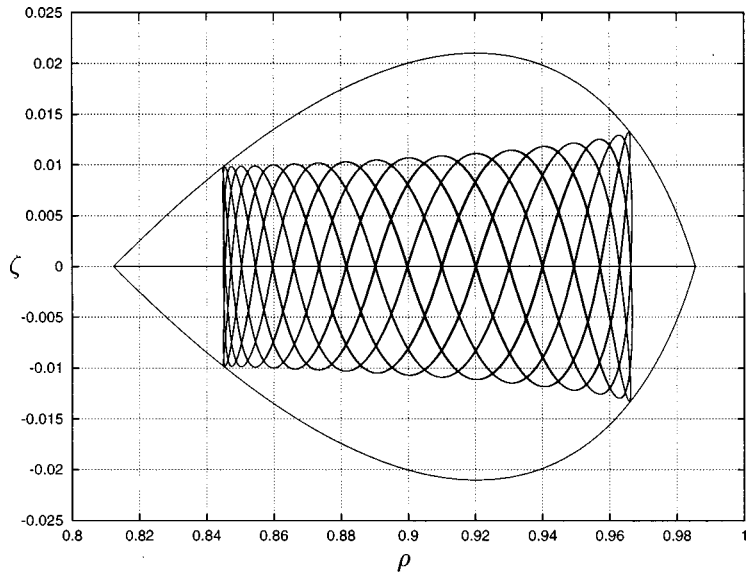

FIG. 10. Example of a geodesic motion in the (ρ,ζ) space with the values $\mu=3$, $L=2.78$, $E=0.77$, $\rho_1=0.81$, $\rho_2=0.986$.

intersection points. For continuously chosen initial conditions of the action variables I_j, the resulting picture consists of closed curves lying inside one another. Lying densely between them, are the finitely many intersection points corresponding to periodic motions.

An important family of integrable Hamiltonian is the so-called Liouville class. In Sec. IV B it will turn out to be the class to which the Hamiltonians for the geodesics are close.

(1) A conformal coordinate transformation $\rho'=\rho'(\rho,\zeta)$ and $\zeta'=\zeta'(\rho,\zeta)$ is defined such that the complex function $z'=z'(\rho,\zeta)=\rho'(\rho,\zeta)+i\zeta'(\rho,\zeta)$ can be expressed only in terms of the variable $z=\rho+i\zeta$, i.e., $z'=z'(z)$. In these coordinates, Hamiltonian systems of the form (1), i.e.,

$$\mathscr{H}=\tfrac{1}{2}e^{\alpha(\rho,\zeta)}(p_\rho^2+p_\zeta^2)+h(\rho,\zeta;L,E), \tag{5}$$

assume the form

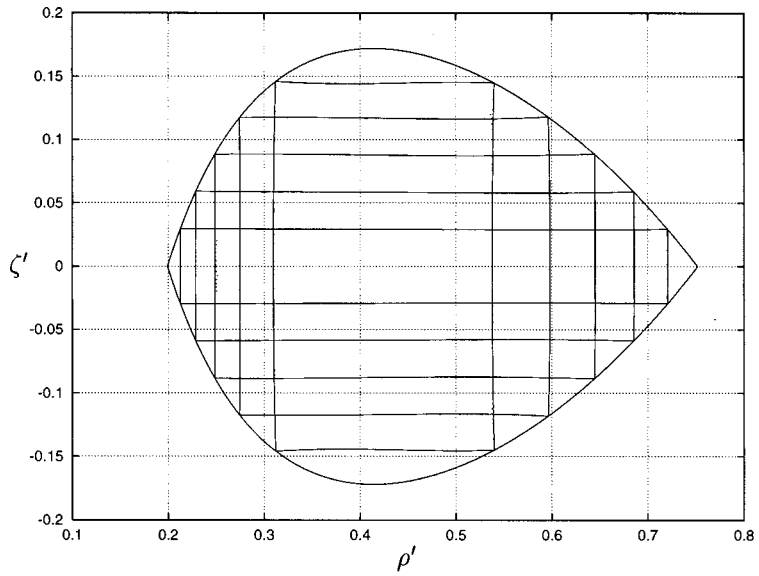

FIG. 11. Boundaries of regions that are filled by Newtonian geodesics in the conformal coordinates (ρ',ζ') introduced by $z'=10z/(10+z^2)$ (the values $\rho_1=0.2$, $\rho_2=0.75$ have been taken).

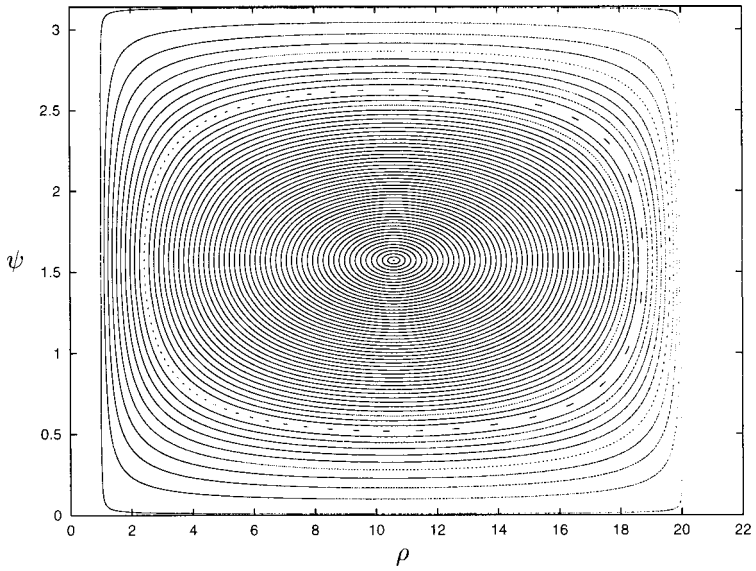

FIG. 12. Poincaré section of surface with $\rho_1 = 1.0001$, $\rho_2 = 20$ (Maclaurin disk).

$$\mathscr{H} = \tfrac{1}{2} e^{\alpha'(\rho',\zeta')}(p_{\rho'}^2 + p_{\zeta'}^2) + h'(\rho',\zeta';L,E)$$

with

$$e^{\alpha'} = e^{\alpha}\left|\frac{dz'}{dz}\right|^2, \quad h'(\rho',\zeta';L,E) = h[\rho(\rho',\zeta'),\zeta(\rho',\zeta');L,E].$$

(2) A Hamiltonian system of the form (5) turns out to be integrable if there are conformal coordinates $z' = z'(z)$ in which the functions α' and h' read

$$e^{\alpha'(\rho',\zeta')} = \frac{1}{R_1(\rho') + Z_1(\zeta')},$$

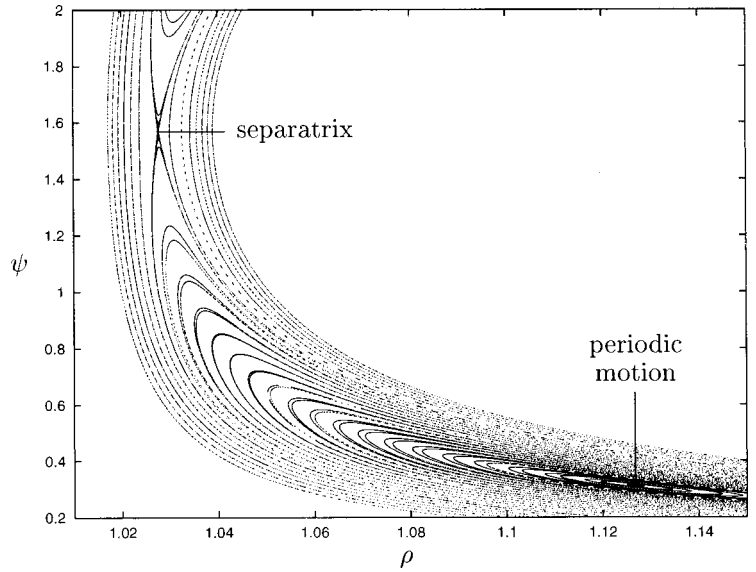

FIG. 13. Magnified extract from Fig. 12.

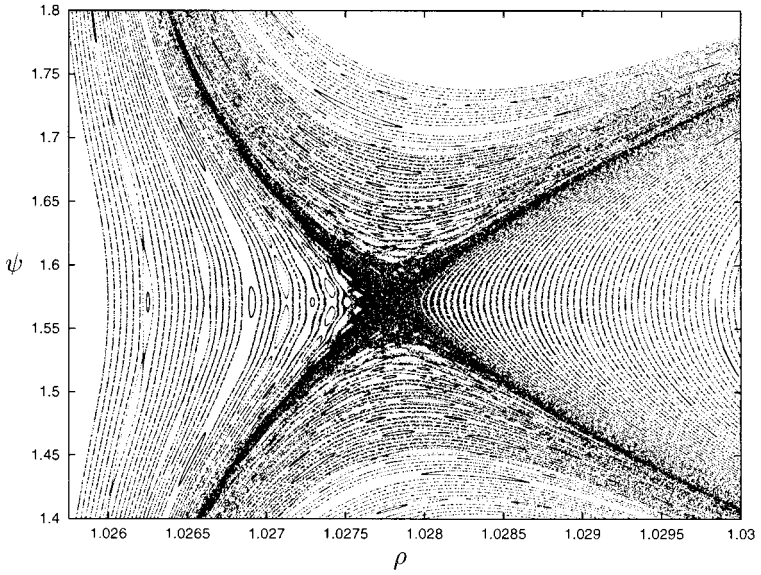

FIG. 14. Still more strongly magnified extract from Fig. 12.

$$h'(\rho',\zeta')=\frac{1}{2}\,e^{\alpha'(\rho',\zeta')}[R_2(\rho')+Z_2(\zeta')]=\frac{1}{2}\cdot\frac{R_2(\rho')+Z_2(\zeta')}{R_1(\rho')+Z_1(\zeta')},$$

where the functions R_i only depend on ρ' and the functions Z_i only on ζ'. Hamiltonians possessing this property belong to the Liouville class of integrable systems. From their first integrals, one concludes

$$\frac{\dot\rho'^2}{\dot\zeta'^2}=\frac{\nu+2\lambda R_1-R_2}{-\nu+2\lambda Z_1-Z_2}=\frac{R_3(\rho';\lambda,\nu)}{Z_3(\zeta';\lambda,\nu)}.$$

If the constants of integration, λ and ν, are fixed by the conditions $R_3(\rho_1';\lambda,\nu)=0=R_3(\rho_2';\lambda,\nu)$, then $\dot\rho'$ vanishes whenever the coordinate ρ' reaches the values ρ_1' or ρ_2'. Simi-

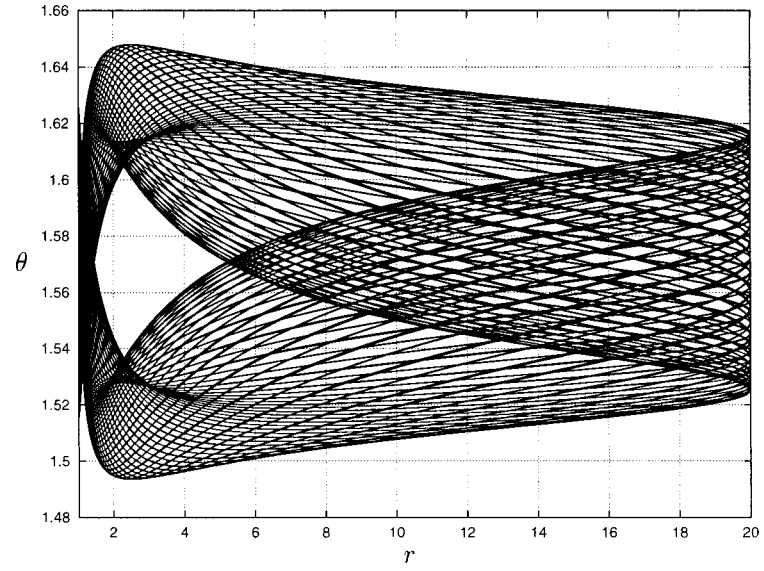

FIG. 15. Trajectory whose intersection points on the Poincaré surface of section fall into the regions of islands displayed in Fig. 13. The curve is presented in spherical polar coordinates ($\rho=r\sin\theta$; $\zeta=r\cos\theta$).

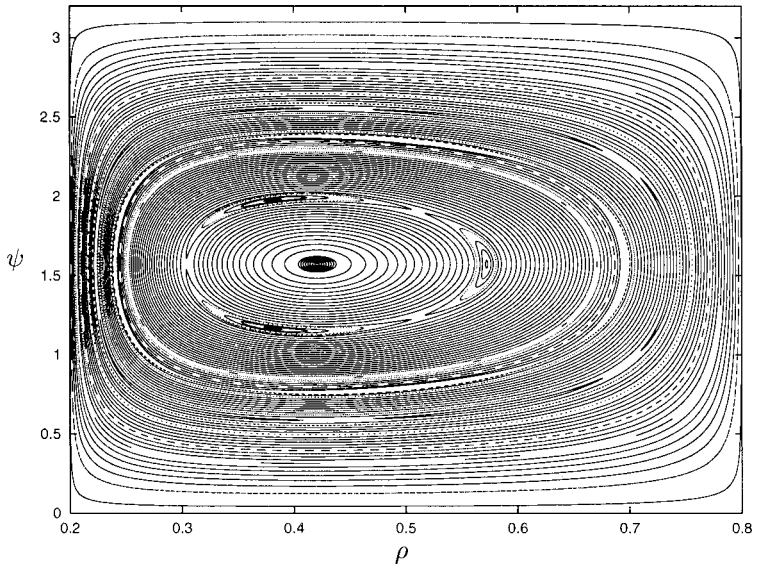

FIG. 16. Poincaré section of surface with $\rho_1=0.2$, $\rho_2=0.8$ (Maclaurin disk).

larly, $\dot{\zeta}'$ vanishes whenever the coordinate ζ' assumes the value of a zero of the function $Z_3(\zeta';\lambda,\nu)$, of which there are at least two for bounded motions. Thus, any bounded nonperiodic motion will gradually fill a region which is spanned by the coordinate axes $\rho'=\text{const}$ and $\zeta'=\text{const}$.

It can be shown that the geodesics within the Kerr metric (the integrability of which was proved by Carter[10]) belong to this class. One gets another important special case by introducing the conformal coordinates $z'=z'(z)=\ln z$, which are closely connected with spherical polar coordinates. Thus, there is a class of integrable Hamiltonians whose (nonperiodic) trajectories in the (ρ,ζ) space densely fill circular segments.

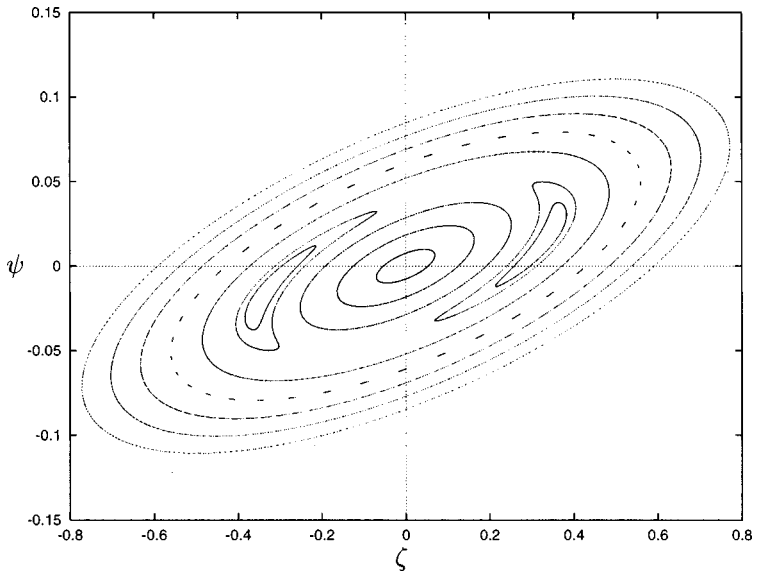

FIG. 17. Poincaré section of surface with $\mu=3.5$, $\rho_1=1.01$, $\rho_2=20$ (relativistic disk). The plane $\rho=\frac{1}{2}(\rho_1+\rho_2)$ and the crossing direction $\dot{\rho}>0$ have been taken.

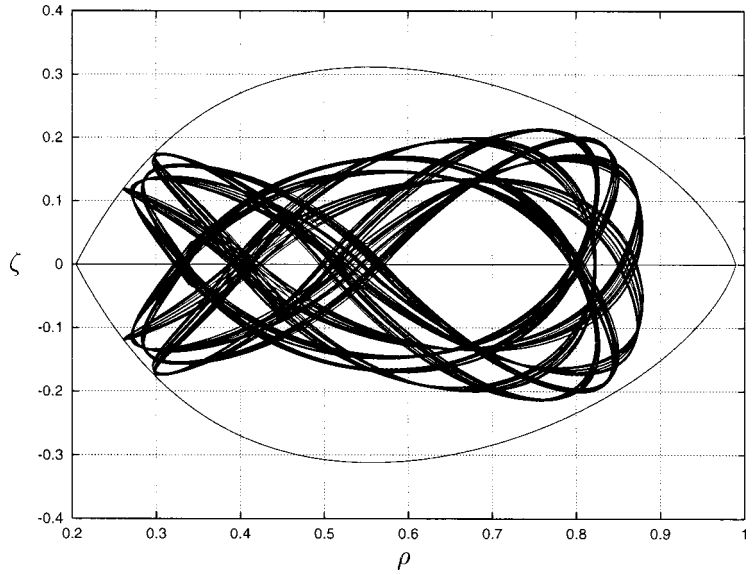

FIG. 18. Example of clearly irregular relativistic geodesic ($\mu=1$, $\rho_1=0.205$, $\rho_2=0.99$).

B. Nearly integrable Hamiltonian systems

Figures 9 and 10 show examples of bounded geodesic motions. In the (ρ,ζ) space, the paths remain within a certain region determined by the formula (2). In Fig. 9, both intersection points (with ρ coordinates ρ_1 and ρ_2) of this region with the disk of the plane are outside the disk, whereas in Fig. 10 both of them are placed inside.

One notices a high regularity of the resulting geodesics. From Fig. 9, one gets the impression that the corresponding Hamiltonian possesses, in spherical polar coordinates, the separation property discussed in Sec. IV A. Also the geodesic shown in Fig. 10 seems to belong to an integrable system of the Liouville class. In Fig. 11, the boundaries of the regions that are gradually filled by the geodesics can be seen in conformal coordinates which are introduced by the transformation $z'=10z/(10+z^2)$.

However, these systems turned out to be clearly nonintegrable. But their deviations from

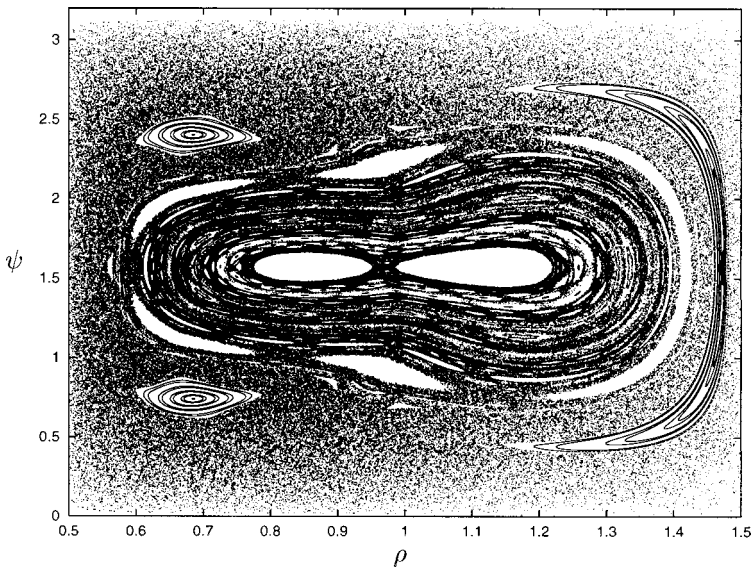

FIG. 19. Poincaré section of surface with $\rho_1=0.5$, $\rho_2=1.5$ (Maclaurin disk).

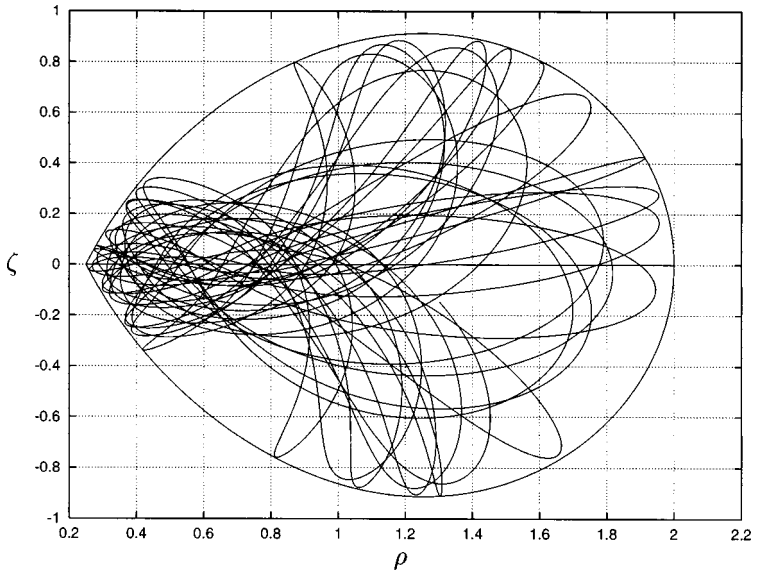

FIG. 20. Example of a geodesic motion in the (ρ,ζ) space with the values $\mu=1$, $L=-0.57$, $E=0.89$, $\rho_1=0.25$, $\rho_2=2$.

integrable systems of the Liouville class are so small that the so-called KAM (Kolmogorov, Arnold, Moser) theorem (see, e.g., Ref. 11) for nearly integrable systems applies.

According to this theorem, the qualitative characteristics of the trajectories of a weakly disturbed integrable Hamiltonian do not change drastically. Most of the curves are still regular, i.e., they are associated with first integrals of motion. Some of the trajectories still lie on tori which now are slightly deformed compared to the integrable case. However, the most regular motions occur on more complicated surfaces. Insight into the topological structure of these surfaces is given by the Poincaré sections of surface. The remaining geodesics show irregular behavior and are not associated with first integrals of motion.

The situation is illustrated by the Poincaré sections of surface displayed in Figs. 12–14. The trajectories are considered in the three-dimensional hypersurface $\mathscr{H}=0$. A coordinate net in this space is given by the coordinates ρ, ζ, and ψ where the angle ψ is defined by $\tan\psi=\dot{\zeta}/\dot{\rho}$. For the Poincaré sections of surface, the intersection plane $\zeta=0$ and the crossing direction $\dot{\zeta}>0$ have

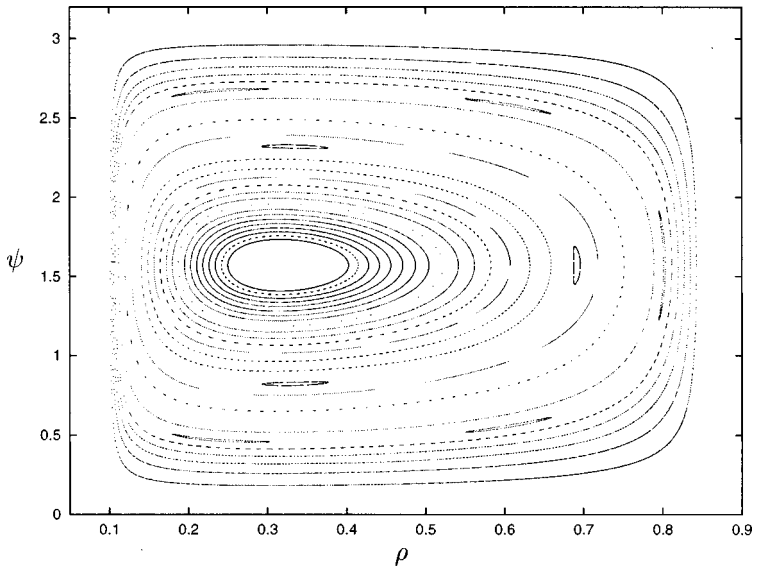

FIG. 21. Poincaré section of surface with $\rho_1=0.1$, $\rho_2=0.85$ (Maclaurin disk).

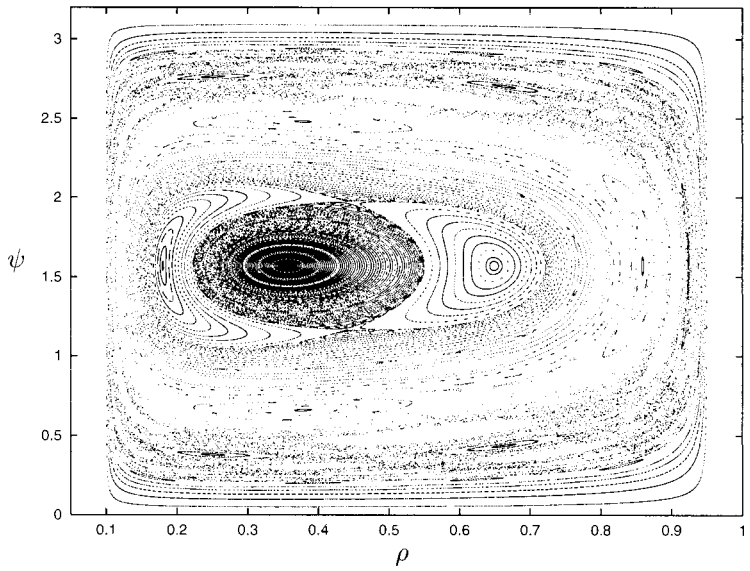

FIG. 22. Poincaré section of surface with $\rho_1 = 0.1$, $\rho_2 = 0.95$ (Maclaurin disk). The distances between the KAM tori have clearly enlarged. The irregular trajectories fill a considerable part of the phase space.

been taken. [The sections of surfaces were created by taking fixed values ρ_1 and ρ_2 and varying the initial conditions $(\rho(0),\zeta(0))$ of the trajectories continuously along the boundary of the region described by formula (2).]

At first glance, one sees as depicted in Fig. 12, what seems to simply be a series of curves lying inside one another, as described in Sec. IV A for integrable motions. But if one enlarges an extract, a more complicated structure becomes evident (Fig. 13). The periodic motions are encircled by islands which lie inside one another. Furthermore, there are so-called separatrices encircling different groups of these islands. In the vicinity of an apex between two of these groups of islands, one recognizes a still more complicated structure (Fig. 14). New groups of islands can be found, and at every new level of magnification, the same qualitative picture emerges. The separatrices represent irregular motions. However, each of the islands belongs to a regular one

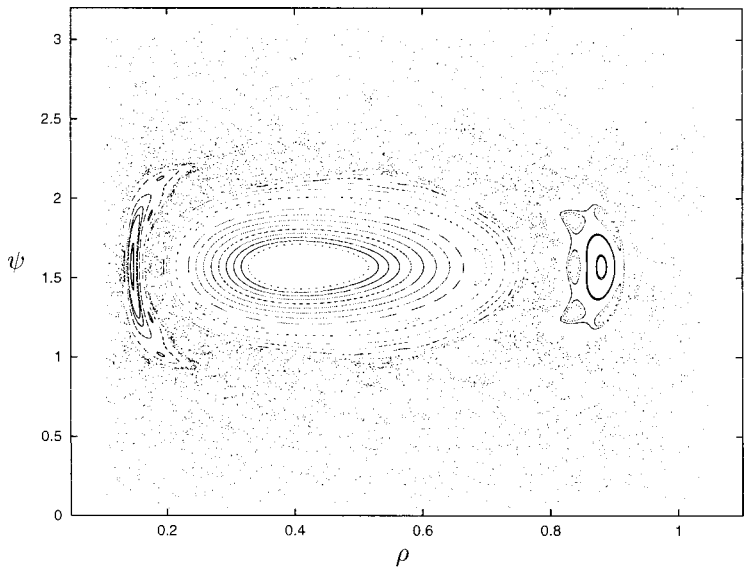

FIG. 23. Poincaré section of surface with $\rho_1 = 0.1$, $\rho_2 = 1.05$ (Maclaurin disk). The characteristical KAM tori have disappeared. Now, there are islands of regular motions within a region of globally stochastic motions.

(with a complicated topological structure of the surface on which the trajectory is trapped). An example for a such a geodesic motion is to be seen in Fig. 15.

Also the geodesics, for which the corresponding region of motion given by formula (2) intersects the plane of the disk at the ρ values $\rho_1<\rho_2<1$, prove to be nearly integrable. As to be seen in Fig. 16, they are not as close to integrable systems as those for the case $1<\rho_1<\rho_2$. But one gets the same qualitative pictures.

The qualitative properties discussed above can be found for the Maclaurin disk (i.e., the Newtonian limit) as well as for the relativistic disk. In general, the relativistic motions show a somewhat stronger deviation from integrable systems, but they still can be regarded as nearly integrable motions. A Poincaré section of surface for the relativistic disk can be seen in Fig. 17. A clearly irregular motion is displayed in Fig. 18.

From these investigations one concludes that the regions in the (ρ,ζ) space that are gradually filled by the nonperiodic geodesics are not spanned by the coordinate axis $\rho'=\text{const}$ and $\zeta'=\text{const}$ of a conformal coordinate system. Rather one finds that the trajectories are trapped by neighboring regions of this form. The distance between these neighboring regions is determined by the deviation of the geodesic in question from an integrable motion.

C. Global stochastic motions

As the departures from an integrable Hamiltonian grow, the distances between the regular curves in phase space become larger and larger. If the disturbance is large enough, then there are no longer any KAM tori, which previously encircled the irregular trajectories. In a global stochastic region, smaller subregions occur, containing groups of islands similar to those described in Sec. IV B. Again, these islands correspond to regular motions.

The remaining bounded geodesic motions, for which the boundary of the region described by formula (2) intersects the plane of the disk at the ρ coordinates $\rho_1<1<\rho_2$, show just this stochastic behavior (as can be seen in Figs. 19 and 20).

The transition from nearly integrable systems to the global stochastic ones is indicated in Figs. 21–23. Here, while keeping the value $\rho_1<1$ fixed, the coordinate ρ_2 is gradually enlarged, which eventually leads to the global stochastic picture in which the KAM tori have disappeared.

Global stochastic motions possess a rather unpredictable behavior. In general, the trajectories are not necessarily trapped into a subregion of the region described by (2) (as was the case for nearly integrable motions). However, if one restricts oneself to motions which lie in a region of islands, one has again nearly integrable motions with particular properties. (For a detailed introduction into the theory of regular and stochastic motions, see, e.g., Ref. 12.)

ACKNOWLEDGMENTS

My warmest thanks go to R. Meinel and G. Neugebauer for their hints and support.

[1] J. M. Bardeen, in *Black Holes*, edited by C. DeWitt and B. S. DeWitt (Gordon and Breach, New York, 1973), p. 219.
[2] B. Carter, Phys. Rev. **174**, 1559 (1970).
[3] J. M. Bardeen and R. V. Wagoner, Astrophys. J. **167**, 359 (1971).
[4] G. Neugebauer and R. Meinel, Phys. Rev. Lett. **75**, 3046 (1995).
[5] F. J. Ernst, Phys. Rev. **167**, 1175 (1968).
[6] D. Kramer and G. Neugebauer, Commun. Math. Phys. **7**, 173 (1968).
[7] R. Meinel and A. Kleinwächter, Einstein Studies (Birkhäuser) **6**, 339 (1995).
[8] G. Neugebauer, A. Kleinwächter, and R. Meinel, Helv. Phys. Acta **69**, 472 (1996).
[9] G. Contopoulos, in *Galactic Dynamics and N-Body Simulations* (Springer, Berlin, 1994), p. 50.
[10] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978, 1989), p. 271.
[11] See Ref. 9, p. 405.
[12] A. J. Lichtenberg and M. A. Liebermann, *Regular and Stochastic Motion* (Springer, New York, 1983), pp. 158–178.

Cauchy horizon end points and differentiability

John K. Beem[a)]
Mathematics Department, University of Missouri–Columbia, Columbia, Missouri 65211

Andrzej Królak[b)]
Max-Planck-Institute for Gravitational Physics, Albert-Einstein-Institute, Schlaatzweg 1, 1447 Potsdam, Germany

(Received 5 March 1998; accepted for publication 15 June 1998)

Cauchy horizons are shown to be differentiable at end points where only a single null generator leaves the horizon. A Cauchy horizon fails to have any null generator end points on a given open subset iff it is differentiable on the open subset and also iff the horizon is (at least) of class C^1 on the open subset. Given the null convergence condition, a compact horizon which is of class C^2 almost everywhere has no end points and is (at least) of class C^1 at all points. © *1998 American Institute of Physics.* [S0022-2488(98)00911-6]

I. INTRODUCTION

Cauchy horizons and black hole event horizons have been extensively studied and used in relativity.[1–12] For general space–times, horizons may fail to be stable under small metric perturbations, however, some sufficiency conditions for various stability questions have been obtained.[13,14] In the present paper, we will consider some differentiability questions for Cauchy horizons.

Let (M,g) be a space–time with a partial Cauchy surface S. The future Cauchy development $D^+(S)$ is the set of points of the space–time where, in theory, one may calculate everything in terms of initial data on S. The future Cauchy horizon $H^+(S)$ is the future boundary of S. We state our results in terms of the future horizon $H^+(S)$, but similar results hold for any past Cauchy horizon $H^-(S)$.

Cauchy horizons are *achronal* (i.e., no two points on the horizon may be joined by a timelike curve) and this implies that Cauchy horizons (locally) satisfy a Lipschitz condition. This, in turn, implies that Cauchy horizons are differentiable almost everywhere. Because they are differentiable except for a set of (three dimensional) measure zero, it seems that they have often been assumed to be smooth except for a set which may be more or less neglected. However, one must remember in the above that (1) differentiable only refers to being differentiable at a single point and (2) sets of measure zero may be quite widely distributed. In fact, Chruściel and Galloway[5] have constructed examples of both a Cauchy horizon and a black hole event horizon where the horizons fail to be differentiable on dense subsets. Thus, they have constructed examples of horizons which contain no open subsets on which they are differentiable at all points. They point out that these examples raise definite questions concerning some major arguments that have been given in the past where smoothness assumptions have been implicitly assumed.[5] In light of these new examples, it is clear that there is a real need for a deeper understanding of the differentiability properties of horizons.

Each point p of a Cauchy horizon $H^+(S)$ lies on at least one null generator.[8] Since partial Cauchy surfaces are edgeless, a null generator continues to lie on the horizon when it is extended in the past. However, null generators may or may not stay on the horizon when they are extended in the future direction. If a null generator leaves the horizon, then there is a last point where it remains on the horizon. This last point is said to be an *end point* of the horizon.

End points where two or more null generators leave the horizon are points where the horizon must fail to be differentiable.[5,7] In addition, Chruściel and Galloway[5] have shown that Cauchy

[a)]Electronic mail: mathjkb@showme.missouri.edu
[b)]On leave from Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, 00-950 Warsaw, Poland. Electronic mail: krolak@aei-potsdam.mpg.de

 © 1998 American Institute of Physics

horizons are differentiable at points which are not end points. Furthermore, Chruściel and Galloway pointed out the need to resolve the differentiability issue for end points where only one generator leaves the horizon. In the present paper, we show that Cauchy horizons are differentiable at these end points. This completes the classification of (pointwise) differentiability for Cauchy horizons in terms of null generators and their end points. However, this result raises the following question. Is it always true that the entire set of end points of a Cauchy horizon will have measure zero? We conjecture that the answer should be affirmative based on known examples. Also, some support for our conjecture is given by our proof that the set of end points with only one generator is in the closure of the set of end points with more than one generator.

Restricting our attention to an open subset W of the Cauchy horizon $H^+(S)$ and assuming that the horizon has no end points on the open set W, we find the horizon must be differentiable at each point of W and, in fact, that the horizon must be at least of class C^1 on W. Conversely, we find differentiability on an open set W implies there are no end points on W. In general, (pointwise) differentiability on an open set yields class C^1, but not necessarily class C^2. We give an example to demonstrate that one may not conclude that either differentiability or lack of end points on an open set W imply class C^2 on W. We also include a very simple example of a Cauchy horizon with an end point where only one null generator leaves the horizon.

Hawking[7] has argued that under certain conditions there should be no compactly generated Cauchy horizons. These are horizons with a compact set K such that all null generators eventually enter and remain in this set K when extended in the past direction. Hawking was interested in establishing arguments against the possibility of having closed timelike curves in physical space–times and, as he mentions, most of his arguments apply to the special case of compact horizons. A key part of Hawking's paper involves flowing the horizon back along null generators to get a contradiction. We use his notation and technique to find sufficient conditions for a compact horizon to have no end points and to be at least of class C^1 at all points. More precisely, we show that if one has the null convergence condition and a compact horizon which is of class C^2 on an open set G with complement of measure zero, then the horizon has no end points and is (at least) of class C^1 at all points.

II. PRELIMINARIES

Let (M,g) be a space–time. Although our results hold for n-dimensional space–times, we will only give the proofs in the four-dimensional case since similar proofs hold in other dimensions. Thus, we will take M to be a smooth, connected, four-dimensional, Hausdorff manifold with a Lorentzian metric, a countable basis, and a time orientation. The Lorentzian metric g will have signature $(-,+,+,+)$. A *partial Cauchy surface* S will be a connected, acausal, edgeless three-dimensional submanifold of (M,g).[2,7,8] The *future Cauchy development* $D^+(S)$ consists of all points $p \in M$ such that each past endless and past directed causal curve from p intersects the set S. The *future Cauchy horizon* is $H^+(S)=\overline{(D^+(S))}-I^-(D^+(S))$. Let p be a point of the Cauchy horizon. It is well known, that there is at least one null generator of $H^+(S)$ containing p. Each null generator is at least part of a null geodesic of M. When a null generator of $H^+(S)$ is extended into the past, it continues to lie on the horizon.[8] However, if a null generator is extended into the future it may have a last point on the horizon which is then said to be an *end point* the horizon. We will define the *multiplicity* of a point p in $H^+(S)$ to be the number of null generators containing p. Points of the horizon which are not end points must have multiplicity one. The multiplicity of an end point may be any positive integer or infinite. We will call the set of end points of multiplicity two or higher the *crease set*.[5]

Consider any fixed point p of the Cauchy horizon $H^+(S)$ and let x^0,x^1,x^2,x^3 be local coordinates defined on an open set about $p=(p^0,p^1,p^2,p^3)$. Let $H^+(S)$ be given near p by an equation of the form

$$x^0=f_H(x^1,x^2,x^3).$$

The horizon $H^+(S)$ is *differentiable* at the point p iff the function f_H is differentiable at the point (p^1,p^2,p^3) using the advanced calculus definition of differentiability.[15] In particular, if $p=(0,0,0,0)$ corresponds to the origin in the given local coordinates and if

$$\Delta x=(x^1,x^2,x^3)$$

represents a small displacement from p in the $x^0=0$ plane, then $H^+(S)$ is differentiable at p iff one has

$$f_H(\Delta x)=f_H(0)+\sum a_i x^i+R_H(\Delta x)=0+\sum a_i x^i+R_H(\Delta x),$$

where the ratio $R_H(\Delta x)/|\Delta x|$ converges to zero as $|\Delta x|$ goes to zero. Here we use

$$|\Delta x|=\sqrt{(x^1)^2+(x^2)^2+(x^3)^2}.$$

Note that even though $|\Delta x|$ represents a coordinate dependent length, the Cauchy surface $H^+(S)$ will be differentiable or not differentiable at p independent of the particular choice of local coordinates. Of course, when $H^+(S)$ is differentiable at p, one has $a_i=\partial f_H/\partial x^i$ evaluated at the origin for $i=1,2,3$. The surface $H^+(S)$ is of class C^r on an open neighborhood of p iff f_H is of class C^r on an open neighborhood of the origin.

If $H^+(S)$ is differentiable at the point p, then there is a well-defined three-dimensional linear subspace N_0 in the tangent space $T_p(M)$ such that N_0 is tangent to the three-dimensional surface $H^+(S)$ at p. In the above notation a basis for N_0 is given by $\{a_i\partial/\partial x^0+\partial/\partial x^i|i=1,2,3\}$. It is clear that if local coordinates x^0,x^1,x^2,x^3 have been chosen as above, then the partials $a_i=\partial f_H/\partial x^i$ at p are determined by N_0 assuming, of course, that $H^+(S)$ is differentiable at p and is given by $x^0=f_H(\Delta x)$ near p. Notice that the given tangent plane N_0 cannot be spacelike since there is at least one null generator containing p. Also, the tangent plane N_0 cannot be timelike since otherwise p would be in chronological past of some points of $H^+(S)$. Thus, N_0 is null [i.e., of signature $(0,+,+)$] and it follows that there is a one-dimensional linear subspace L_0 in N_0 such that L_0 is null. In fact, $N_0=L_0^\perp$ is the orthogonal space to L_0 and is thus uniquely determined by L_0. Given a tangent vector X at p with $X\in N_0$, then X belongs to L_0 iff $g(X,Y)=0$ for all $Y\in N_0$.

At a point p where $H^+(S)$ is differentiable there can only be one null generator through p since two different null generators through p would yield a two-dimensional timelike plane lying in the (three-dimensional) tangent plane to the horizon and this would imply the existence of a timelike plane tangent to $H^+(S)$ at p in contradiction to the above. Thus at points with two or more null generators of $H^+(S)$, the horizon $H^+(S)$ cannot be differentiable.[5] In fact, if p is a point of the horizon having two or more null generators, then all null generators through p must leave the horizon at p when traversed in the future direction. This follows since if $\gamma_1(t)$ and $\gamma_2(t)$ are future directed null generators with $\gamma_1(0)=\gamma_2(0)=p$ and if $\gamma_1(t)$ remains on $H^+(S)$ for some $\gamma_1(t_1)$ with $t_1>0$, one has that all points of $\gamma_2(t)$ with $t<0$ are in the chronological past of $\gamma_1(t_1)$ since one can traverse $\gamma_1(t)$ backward to p, make a corner at p to head in the past direction of the null geodesic $\gamma_2(t)$, and then traverse this null geodesic backward. This would imply the points $\gamma_2(t)$ for $t<0$ are in the set $D^+(S)$ and not on the horizon, in contradiction.

We now give an example of an end point p of multiplicity one.

Example 2.1: Let $M_0=L^3$ be three-dimensional Minkowski space–time with coordinates (t,x,y) and $g=-dt^2+dx^2+dy^2$. Remove the parabola $y=x^2$ from the $t=0$ plane to obtain the space–time M. Let the partial Cauchy surface S be the portion of the $x-y$ plane which is interior to the parabola $y=x^2$ (i.e., $S=\{(0,x,y)|y>x^2\}$). The curvature k of the curve $y=x^2$ has a maximum value of 2 at $x=0$ and is always less than 2 for points on the parabola other than the origin. The osculating circle at the origin to the curve $y=x^2$ is a circle in the $x-y$ plane of radius $1/k=1/2$ and center at $y=1/2$ on the y-axis. The chronological past of the point $(1/2,0,1/2)$ intersected with the $x-y$ plane is the interior of this osculating circle and lies in the set S. It follows that the null geodesic $x=0$, $y=t$ contains a null generator of $H^+(S)$. In particular, the curve $c(t)=(t,0,t)$ for $0<t<\infty$ is a null geodesic which lies on the horizon for $0<t\leqslant 1/2$. Notice that if one intersects the $x-y$ plane with the chronological past of $(t,0,t)$ for $t>1/2$, then one obtains some points outside of the above partial Cauchy surface S. Thus, the null generator $c(t)=(t,0,t)$ must leave the horizon at $t=1/2$ and one finds that $(1/2,0,1/2)$ is an end point of the horizon. However, the end point $p=(1/2,0,1/2)$ has no other null generators since all other null geodesics through

this point intersect the set S. Thus, p is an end point of multiplicity one as desired. In this example one may directly check that $H^+(S)$ is differentiable at the point p. The crease set is a curve which lies above the points on the y-axis for which $1/2<y<\infty$.

In the next example we obtain a horizon which has an open set W where the horizon is of class C^1, but where there are some points in this open set W where the horizon fails to be of class C^2. In particular, there is a certain null generator and the horizon fails to be of class C^2 along this null generator.

Example 2.2: Let $M_0=L^3$. The space–time M will be obtained by removing two half lines and one-fourth of a circle. From M_0 remove the set in the $t=0$ plane given by $y=-1$ and $0\leqslant x$. Then remove the part of the circle $x^2+y^2=1$ for which $x<0$ and $y<0$. Finally, remove the half line in the $t=0$ plane given by $x=-1$ and $0\leqslant y$. The partial Cauchy surface S will be the open set in the $x-y$ plane corresponding to the interior of the convex hull of the deleted set. The null generators corresponding to the quarter circle focus at the point (1,0,0). The null generators corresponding to points on the line $y=-1$ for sufficiently small values of t lie on lines of the form $c(t)=(t,x,t-1)$ where $0\leqslant x$. It follows that for values of x and y near $x=0$ and $y=-1/2$ the height t of the Cauchy horizon is given by $t=f_H(x,y)$ where $f_H(x,y)=1+y$ for $0\leqslant x$ and also $f_H(x,y)=1-(x^2+y^2)^{1/2}$ for $x<0$. It is easily checked that f_H is of class C^1 near $(0,-1/2)$, but that f_H fails to be of class C^2 along the y-axis near $(0,-1/2)$. It follows that there is an open set W in $H^+(S)$ about the point $(1/2,0,-1/2)$ where the horizon is of class C^1, but not of class C^2. In this example there are no end points of multiplicity one. The crease set consists of the closed half line $t=y+1=x+1$ for $0\leqslant y<\infty$ and is a space–like curve.

Both examples 2.1 and 2.2 may be changed to four-dimensional examples using a Cartesian product with a positive definite R^1 corresponding to the z-axis. In particular, to modify example 2.1, one may take $M_0=L^4$, delete the two-dimensional set $\{(0,x,y,z)|y=x^2\}$, and use the three-dimensional partial Cauchy surface $S=\{(0,x,y,z)|y>x^2\}$. This generates a horizon $H^+(S)$ with an end point $p=(1/2,0,1/2,0)$ of multiplicity one. To modify example 2.2, one may take $M_0=L^4$ and delete the set formed by the Cartesian product of the originally deleted set with the z-axis. Of course, in this four-dimensional example, one takes the new partial Cauchy surface to be the Cartesian product of the original Cauchy surface with the z-axis.

III. DIFFERENTIABILITY OF CAUCHY HORIZONS

One may always represent tangent null directions using a normalization based on an auxiliary positive definite metric $g_{\rm pos}$. Given a fixed point p on a null geodesic γ, choose the uniquely defined tangent vector V to γ at p which is future pointing and satisfies $g_{\rm pos}(V,V)=1$. With this normalization, it is clear that the null directions at p form a compact set homeomorphic to S^2. Of course, we are identifying the null direction given by a null vector V as the same as that given by αV for nonzero values of α. Thus, in particular, the null direction represented by V is the same as the null direction represented by $-V$. Note that if one is given any compact subset K of M, the set of null directions attached at points of K forms a compact set.

Let γ_n be a sequence of null generators of $H^+(S)$ and let $p_n\in\gamma_n$ for each n. Assume the sequence $\{p_n\}$ converges to p and let γ be a null generator at p. We will say the sequence γ_n converges to γ if there is a sequence of null vectors $\{V_n\}$ converging to the null vector V such that each V_n is tangent to the corresponding generator γ_n at p_n and V is tangent to γ at p. Notice that in this definition we do not assume that the points $\{p_n\}$ have unique null generators and we do not assume that the point p has a unique null generator. In particular, we allow for the possibility that some (or all) of $\{p_n\}$ and p may be end points.

We now show that if p is a point of multiplicity one, then any point on the horizon sufficiently close to p must have all of its null generators close to the unique null generator containing p.

Lemma 3.1: Let p be a point of multiplicity one on the horizon $H^+(S)$. If $\{p_n\}$ is a sequence of points on the horizon converging to p and if for each n a null generator γ_n containing the point p_n has been chosen, then the sequence γ_n converges to the null generator γ containing p.

Proof: Assume by way of contradiction that the sequence $\{p_n\}$ converges to p, but that the null directions defined by the generators γ_n do not converge to the null direction defined by γ. Using compactness, it follows that there must be a subsequence $\{j\}$ of the sequence $\{n\}$ with the γ_j directions converging to a null direction at p different from that defined by γ. This yields a second null generator at p, in contradiction to the assumption that p was of multiplicity one. □

Clearly, Lemma 3.1 implies that if W is an open subset of the horizon and if each point of W has multiplicity one, then the null generators move in a continuous fashion on W. More precisely, we have the following result.

Lemma 3.2: Let W be an open subset of the future Cauchy horizon $H^+(S)$. If each point of W has multiplicity one, then there is a (nonvanishing) null vector field V tangent to null generators and defined on W such that V is continuous. Furthermore, the vector field V may be taken to be future pointing.

We are now able to show that for open subsets of the horizon, differentiability implies (at least) class C^1. Example 2.2 shows that differentiability on an open set does not, in general, imply class C^2 and hence the following is the best possible result one may obtain without additional assumptions. Recall that if a surface is of class C^r for $r>1$, then it is also of class C^1.

Proposition 3.3: If W is an open subset of $H^+(S)$ and if $H^+(S)$ is differentiable at all points of W, then $H^+(S)$ is of class C^r on W for some $r\geqslant 1$.

Proof: A point where the horizon is differentiable must be a point of multiplicity one. Hence Lemma 3.2 yields the existence of a null vector field V on W which is continuous, tangent to $H^+(S)$ and serves to define the generator directions on W. Let $p\in W$ and, as in Sec. II, introduce local coordinates x^0,x^1,x^2,x^3 near p with $\partial/\partial x^0$ future timelike. Let $H^+(S)$ be given locally by $x^0=f_H(\Delta x)$. The partial derivatives of f_H exist near p since $H^+(S)$ is differentiable near p. Notice that the null vector field V determines the orthogonal space $V^\perp$ which is the null tangent plane to $H^+(S)$ at each point near p. Furthermore, the tangent plane $V^\perp$ determines the first-order partial derivatives of f_H. The continuity of V yields the continuity of $V^\perp$ and hence the continuity of the first-order partial derivatives of f_H. Thus, f_H is at least of class C^1 which is equivalent to $H^+(S)$ being at least of class C^1, as desired. □

Fix a point p on $H^+(S)$ of multiplicity one. Choose normal coordinates x^0,x^1,x^2,x^3 with p corresponding to the origin, with $\partial/\partial x^0$ future timelike at p, and with the future direction of the unique null generator at p corresponding to the direction of $\partial/\partial x^0+\partial/\partial x^1$. Assume also that the natural basis is orthonormal at p. As before, let $H^+(S)$ be given near p by $x^0=f_H(\Delta x)$. If $H^+(S)$ is differentiable at p one has $\partial f_H/\partial x^1=1$ and $\partial f_H/\partial x^2=\partial f_H/\partial x^3=0$ at p. Assuming that f_H is of class C^r for $r\geqslant 1$ near p, define new coordinates z^0,z^1,z^2,z^3 by $z^0=x^0-f_H(\Delta x)$, and $z^i=x^i$ for $i=1,2,3$. Locally $H^+(S)$ corresponds to the set $z^0=0$. Note that z^0,z^1,z^2,z^3 are local coordinates of class C^r in terms of the original x coordinates.

Proposition 3.4: If a Cauchy horizon $H^+(S)$ is differentiable on an open subset W, then the horizon has no end points on W.

Proof: The horizon must be (at least) of class C^1 on W by Proposition 3.3. Fixing $p\in W$, we may use the above local coordinates z^0,z^1,z^2,z^3 which are at least of class C^1 and the horizon is given near p by $z^0=0$. From Lemma 3.2 it follows that there is a continuous future directed null vector field V tangent to $H^+(S)$ near p. Then $V=\Sigma y^i(z^1,z^2,z^3)\partial/\partial z^i$ where $y^i=y^i(z^1,z^2,z^3)$ are continuous functions for $i=1,2,3$. Consider the system

$$dz^i/dt=y^i(z^1,z^2,z^3)$$

for $i=1,2,3$. This is a system which we may regard as being defined on the coordinate plane $z^0=0$ (i.e., on the horizon) and solving the system corresponds to finding integral curves of the vector field V. However, the above system does not necessarily satisfy a Lipschitz condition and thus, in principle, might fail to have unique solutions. Nevertheless, because of the continuity of the functions y^i, one may apply the Cauchy–Peano existence theorem[16] to obtain a class C^1 solution $c(t)$ with $c(0)=p$. Notice that $c(t)$ is a null curve always tangent to $H^+(S)$ because $c'(t)$ is always equal to the vector field V at the corresponding point $c(t)$. Moving along the curve c a little to the past, one obtains a point $r=c(t_1)\in H^+(S)$ with $t_1<0$. Similarly, moving along c a little to the future one obtains $q=c(t_2)\in H^+(S)$ with $t_2>0$. Traversing the curve c backward from q to r one finds it must be a null geodesic lying on the horizon since if it were not, then r would be in the chronological past of q and this would yield a contradiction to the achronality of $H^+(S)$. Since there is a unique null direction tangent to $H^+(S)$ at p, the curve c (at least) restricted to the domain $[t_1,t_2]$ must lie on the generator containing p and it follows that p is not an end point, as desired. □

We note in passing that the proof of Proposition 3.4 shows that, in fact, the system $dz^i/dt = y^i(z^1,z^2,z^3)$ has unique solutions given an initial point $p \in W$ and that each solution $c(t)$ lies on (part of) a null generator.

Fix a point p of multiplicity one on $H^+(S)$ with null direction L_0 tangent to the unique null generator at p. Let Y_0,Y_1,Y_2,Y_3 be an orthonormal basis at p with Y_0 future pointing timelike and with Y_0+Y_1 future pointing null in the direction of L_0. Then Y_2 and Y_3 lie in the orthogonal space to L_0. Let X_0,X_1,X_2,X_3 be a new basis of T_pM given by $X_0=Y_0$, $X_1=Y_0+Y_1$, $X_2=Y_2$, and $X_3=Y_3$. Take normal coordinates x^0,x^1,x^2,x^3 centered at p determined by the basis X_0,X_1,X_2,X_3. Then $\partial/\partial x^0$ is timelike future pointing, $\partial/\partial x^1$ future pointing in the null direction L_0, and both $\partial/\partial x^2$ and $\partial/\partial x^3$ are in the orthogonal space to L_0. Of course, these normal coordinates fail to be orthonormal at p since, in particular, $\partial/\partial x^1$ is null. Notice that the metric tensor g has components (in the x coordinates) which at p have all zero values except for $g_{00}(p)=g_{01}(p)=g_{10}(p)=-1$ and $g_{22}(p)=g_{33}(p)=1$. The x^1 axis is a geodesic which corresponds to a null generator of $H^+(S)$ and (at least) the origin and negative x^1 axis lie on the horizon. The point p is an end point iff points on the positive x^1 axis do not lie on the horizon. In the following we allow p to be either an end point or a non-end-point. Fix a point $r=(0,t_1,0,0)$ on the negative x^1 axis and lying in a convex normal neighborhood of p. Thus, $t_1<0$. Since we have used normal coordinates, we have $p=\exp_r(V_0)$ where $V_0=|t_1|\partial/\partial x^1$ and V_0 is a null vector attached at r. If one considers the exponential map at r restricted to all null vectors at r in some small neighborhood of V_0, then one obtains a smooth null surface which passes through $p=(0,0,0,0)$. Since $\partial/\partial x^1$ at the origin is a null vector tangent to this null surface, it follows that the tangent plane to this null surface at the point p is the orthogonal plane to $\partial/\partial x^1$, which is also the tangent plane to the coordinate plane $x^0=0$ at $p=(0,0,0,0)$. Let $N(V_0)$ be a sufficiently small neighborhood of V_0 in the tangent bundle TM and let $U(r)$ be a (small) three-dimensional neighborhood of r in the coordinate plane $x^1=t_1$. Then points of $U(r)$ are of the form $u=(u^0,t_1,u^2,u^3)$ where the values of of u^0, u^2, and u^3 are all close to zero. For each fixed $u \in U(r)$ one generates a smooth null surface $S(u^0,u^2,u^3)$ defined near the origin by using the exponential map at u and restricting $\exp_u$ to null vectors in the set $N(V_0)$ attached at u. An element $V \in N(V_0)$ attached at $u \in U(r)$ may be represented in local coordinates as $V=V^0\partial/\partial x^0 + \Sigma V^i \partial/\partial x^i$. Thus, using $g(V,V)=0$ to determine the V^0 component in terms of the V^1,V^2,V^3 components, one may parametrize our six-dimensional domain space by u^0,u^2,u^3,V^1,V^2,V^3. Here the values of u^0, u^2, u^3, V^2, and V^3 are close to zero and the value of V^1 is close to $|t_1|$. Note that tangent vectors to this six-dimensional space may be expressed in terms of a basis of the form $\partial/\partial u^0,\partial/\partial u^2,\partial/\partial u^3,\partial/\partial V^1,\partial/\partial V^2,\partial/\partial V^3$. Consider the map

$$E(u,V)=\exp_u(V)$$

taking elements of the domain space to points of M near p. This map may be written as $E(u,V)=(E^0,E^1,E^2,E^3)$ where each component $E^\mu(u^0,u^2,u^3,V^1,V^2,V^3)$ is a real valued function of six variables. The reader will note a slight abuse of notation in the above. For example, we will use u to denote both the point (u^0,t_1,u^2,u^3) in M and also the corresponding coordinates u^0,u^2,u^3 in our domain space. A similar comment holds for our use of V. Notice that if u is chosen to be the point r, then the map $E(r,-)$ takes the null vectors attached at r to the null surface $S(0,0,0)$ which we noted above is tangent to the coordinate plane $x^0=0$ at p. In our domain space the vector V_0 in T_rM corresponds to $(0,0,0|t_1|,0,0)$. Notice that the derivative of the exponential map for $u=r$ (i.e., $(\exp_r)_*$) is nondegenerate at V_0 since r was chosen close to p. Furthermore, $(\exp_r)_*$ takes the null hyperplane at V_0 tangent to the null vectors at r to the null plane tangent to $x^0=0$ at $p=(0,0,0,0)$. Thus, for fixed $u=r$ and $V=V_0$, E_* takes the three-dimensional space with basis $\partial/\partial V^1,\partial/\partial V^2,\partial/\partial V^2$ corresponding to certain vectors tangent to our six-dimensional domain space and attached to this domain space at $V_0=(0,0,0,|t_1|,0,0)$ in a nonsingular fashion to the three-dimensional space with basis $\partial/\partial x^1,\partial/\partial x^2,\partial/\partial x^3$ of vectors tangent to M and attached at p. Thus, using $x^i=E^i(u^0,u^2,u^3,V^1,V^2,V^3)$ one finds that the three by three matrix $[\partial x^i/\partial V^j]$ is nonsingular when evaluated at the point $(0,0,0,|t_1|,0,0)$ of our domain space. Using the implicit function theorem,[17] we may thus solve for the V^1,V^2,V^3 variables in terms of the u^0,u^2,u^3,x^1,x^2,x^3 variables. Hence, one obtains three C^∞ functions V^i

$=V^i(u^0,u^2,u^3,x^1,x^2,x^3)=V^i(u,\Delta x)$. We will let F be the x^0 component of E using the u^0,u^2,u^3,x^1,x^2,x^3 variables. In other words, we define F to be the real valued function

$$F(u^0,u^2,u^3,x^1,x^2,x^3)=E^0(u,V^1(u,\Delta x),V^2(u,\Delta x),V^3(u,\Delta x)).$$

Then, for fixed u, the surface $S(u^0,u^2,u^3)$ is given near the origin by $x^0=F_u(x^1,x^2,x^3)$ where $F_u(\Delta x)=F(u^0,u^2,u^3,x^1,x^2,x^3)$ is a smooth function of six variables. Still holding $u\in U(r)$ fixed, select a point $q=(q^0,q^1,q^2,q^3)$ near the origin on the surface $S(u^0,u^2,u^3)$. Then one has $q^0=F_u(q^1,q^2,q^3)=F_u(\Delta q)$. Now expand F_u about this point (holding the u coordinates fixed) to obtain

$$F_u(x^1,x^2,x^3)=q^0+\sum\ (\partial F_u/\partial x^i)(x^i-q^i)+R(u,q,\Delta(x-q)),$$

where $\Delta(x-q)=(x^1-q^1,x^2-q^2,x^3-q^3)$ and the partials of F_u are evaluated at Δq. Using the smoothness of F, it follows that the remainder term $R(u,q,\Delta(x-q))$ may be bounded on some compact set in the domain space of six variables using the size of the second partial derivatives of F and the magnitude of $|\Delta(x-q)|^2$. Of course the bound will depend on the mixed second partials of F involving the u^0,u^2,u^3 variables as well as the x^1,x^2,x^3 variables.[18] In particular, one may obtain an inequality of the form

$$|R(u,q,\Delta(x-q))|<M|\Delta(x-q)|^2, \tag{3.1}$$

where M is a constant which holds for u near r, q near p, and all sufficiently small $|\Delta(x-q)|$. We will use this bound on the size of the remainder term in the proof of Theorem 3.5 below. We remark in passing that points of any sufficiently small neighborhood $W(p)$ of $p=(0,0,0,0)$ will lie on many of the null surfaces $S(u^0,u^2,u^3)$. In fact, the null cone from each point of a sufficiently small $W(p)$ will intersect the coordinate plane $x^1=t_1$ in a two-dimensional surface and thus through a point of $W(p)$ there will be a two parameter family of null surfaces in the collection $S(u^0,u^1,u^2)$.

The next result answers a question raised by Chruściel and Galloway.[5] They proved that Cauchy horizons fail to be differentiable at end points with more than one null generator and that Cauchy horizons are differentiable at points which are not end points.[5] They mentioned that these results left open the question of the differentiability of Cauchy horizons at end points where there is only null generator (i.e., end points of multiplicity one in our terminology). As Chruściel and Galloway noted, it is of interest to resolve this remaining differentiability question. In Theorem 3.5, we obtain a positive answer to this question. We show that a Cauchy horizon is always differentiable at points of multiplicity one.

Theorem 3.5: *A Cauchy horizon is differentiable at all points of multiplicity one. In particular, a Cauchy horizon is differentiable at an end point where only one null generator leaves the horizon.*

Proof: Let p be a point of multiplicity one. We will use the above-described coordinates x^0,x^1,x^2,x^3, neighborhoods $U(r)$, $N(V_0)$, $W(p)$, null surfaces $S(u^0,u^2,u^3)$, functions $F_u(x^1,x^2,x^3)$ and inequality (3.1).

Assume that the Cauchy horizon $H^+(S)$ is given near p by $x^0=f_H(\Delta x)$. We will show that f_H must be differentiable at the origin and have all of its first-order partials equal to zero at the origin. This will prove that $H^+(S)$ is differentiable at the origin and that it has its tangent plane at the origin tangent to the coordinate plane $x^0=0$. To show the desired properties of f_H are true, it is sufficient to show that $f_H(\Delta x)/|\Delta x|$ converges to zero as $|\Delta x|$ converges to zero. Hence, by way of contradiction, we assume that there is a sequence of points $\{q_k=(q_k^0,q_k^1,q_k^2,q_k^3)\}$ on $H^+(S)$ and converging to p with

$$|f_H(\Delta q_k)|/|\Delta q_k|>c>0$$

for all k. Here

$$\Delta q_k=(q_k^1,q_k^2,q_k^3),\quad |\Delta q_k|=\sqrt{(q_k^1)^2+(q_k^2)^2+(q_k^3)^2}.$$

The surface $S(0,0,0)$ is tangent at the origin to the coordinate plane $x^0=0$ and represents (part of) the null cone from the point $r=(0,t_1,0,0)$. Note that points in the chronological future of r must lie above the Cauchy horizon and that $\partial/\partial x^0$ is future pointing timelike at the origin. It follows that points of the horizon lie on or below the surface $S(0,0,0)$ and hence

$$q_k^0=f_H(\Delta q_k)\leqslant F(0,0,0,q_k^1,q_k^2,q_k^3)$$

for all k. Expanding $F(0,0,0,x^1,x^2,x^3)$ about the origin in the x variables, one has $F(0,0,0,q_k^1,q_k^2,q_k^3)=0+0+R(\Delta q_k)$ where $R(\Delta q_k)/|\Delta q_k|$ converges to zero as k increases. Using $q_k^0=f_H(\Delta q_k)\leqslant F(0,0,0,\Delta q_k)$ and $|f_H(\Delta q_k)|/|\Delta q_k|>c$, one finds that for all large k one must have $f_H(\Delta q_k)<0$. Also, one obtains the inequality $f_H(\Delta q_k)<-c|\Delta q_k|$ for all large k. For each k, choose a null generator γ_k containing q_k. Lemma 3.1 guarantees that the sequence $\{\gamma_k\}$ converges to the unique null generator γ ($=x^1$ axis) containing p. Thus, for sufficiently large k, each γ_k contains points close to r and has tangent directions for these points close to that of γ at r (i.e., close to the direction of the x^1 axis). It follows that for each large k there will be a well-defined point u_k on γ_k with $u_k=(u_k^0,t_1,u_k^2,u_k^3)\in U(r)$ and the sequence u_k will converge to r. Furthermore, we may parametrize γ_k such that $\gamma_k(0)=u_k$ and $\gamma_k(1)=q_k$. Then $V_k=\gamma_k'(0)$ converges to V_0. The null surface $S(u_k^0,u_k^2,u_k^3)$ contains q_k and is given by $x^0=F_k(\Delta x)$ where $F_k(\Delta x)=F(u_k^0,u_k^2,u_k^3,x^1,x^2,x^3)$. We will show that for large k one has $F_k(0,0,0)<0$ and thus that the surface $S(u_k^0,u_k^2,u_k^3)$ cuts the x^0 axis below the origin. To this end, expand F_k about the point q_k in the x variables to obtain

$$F_k(\Delta x)=q_k^0+\sum\ a_{ki}(x^i-q_k^i)+R_k(\Delta(x-q_k)),$$

where $\Delta(x-q_k)=(x^1-q_k^1,x^2-q_k^2,x^3-q_k^3)$, $R_k(\Delta(x-q_k))=R(u_k,q_k,\Delta(x-q_k))=R(u_k^0,u_k^2,u_k^3,q^1,q^2,q^3,\Delta(x-q_k))$ and a_{ki} represents the partial of F_k with respect to x^i evaluated at the point q_k for $i=1,2,3$. Using $x^1=x^2=x^3=0$ and $q_k^0=f_H(\Delta q_k)<-c|\Delta q_k|$, one obtains the following inequality for large k:

$$F_k(0,0,0)<-c|\Delta q_k|+\sum\ a_{ki}(0-q_k^i)+R_k(-q_k^1,-q_k^2,-q_k^3)$$

Recall that as k increases, the tangent direction to γ_k converges to the tangent direction to γ. Thus, the null tangent plane to $x^0=F_k(\Delta x)$ at q_k must converge to the null plane tangent to $x^0=0$ at the origin. It follows that the coefficients a_{ki} converge to zero for $i=1,2,3$ and thus for large k one has an inequality of the form

$$\left|\sum\ a_{ki}(0-q_k^i)\right|<(c/3)|\Delta q_k|.$$

Also, using inequality (3.1) one has $|R_k(-q_k^1,-q_k^3,-q_k^3)|<M|\Delta q_k|^2$ which implies that for large k

$$|R_k(-q_k^1,-q_k^2,-q_k^3)|<(c/3)|\Delta q_k|.$$

Consequently, we find

$$F_k(0,0,0)<-c|\Delta q_k|+(c/3)|\Delta q_k|+(c/3)|\Delta q_k|,$$

which yields $F_k(0,0,0)<0$ for large k since $|\Delta q_k|\neq 0$. Using the fact that the x^0 axis is timelike future directed, one finds that $F_k(0,0,0)$ is in the chronological past of $p=(0,0,0,0)$ for large k. Using the fact that u_k is in the causal past of $F_k(0,0,0)$, one finds that u_k must be in the chronological past of $p=(0,0,0,0)$ for large k. Since p and u_k must lie on $H^+(S)$, this contradicts the achronality of the horizon. We conclude that f_H is differentiable at the origin and all of its partials with respect to x^i are zero at the origin. It follows that $H^+(S)$ is differentiable at the origin, as desired. □

Recall that the *crease* set is the set of points in the horizon having multiplicity at least two and that the horizon fails to be differentiable at any such point. Clearly, Theorem 3.5 is equivalent to the statement *that the crease set consists precisely of those points on the Cauchy horizon where the horizon fails to be differentiable*.

Combining Propositions 3.3 and 3.4 with Theorem 3.5, one obtains the following result.

Proposition 3.6: Let W be an open subset of the Cauchy horizon $H^+(S)$. *Then the following are equivalent:*

(1) $H^+(S)$ *is differentiable on W.*
(2) $H^+(S)$ *is of class* C^r *on W for some* $r \geqslant 1$.
(3) $H^+(S)$ *has no end points on W.*
(4) All points of W have multiplicity one.

Note that the four parts of Proposition 3.6 are logically equivalent for an *open* set *W*, but that, in general, they are not necessarily equivalent for sets which fail to be open. Using the equivalence of parts (1) and (3) of Proposition 3.6, it now follows that near each end point of multiplicity one there must be points where the horizon fails to be differentiable. Hence, each neighborhood of an end point of multiplicity one must contain end points of higher multiplicity. This yields the following corollary.

Corollary 3.7: If p is an end point of multiplicity one on a Cauchy horizon $H^+(S)$, then each neighborhood $W(p)$ of p on $H^+(S)$ contains points where the horizon fails to be differentiable. Hence, the set of end points of multiplicity one is in the closure of the crease set.

We know that the nondifferentiable set (i.e., the crease set) has (three-dimensional) measure zero since the horizon satisfies a Lipschitz condition. On the other hand, since a Cauchy horizon is differentiable at an end point of multiplicity one, the fact that $H^+(S)$ satisfies a Lipschitz condition does not give us any direct information on the measure of the set of end points of multiplicity one. Based on known examples it seems likely that this set should be a relatively small set. Thus, we conjecture that the set of end points of multiplicity one is of measure zero for all Cauchy horizons. Clearly, this is equivalent to the following end point conjecture.

Conjecture: The set of all end points of a Cauchy horizon must have measure zero.

An affirmative or negative answer to the above conjecture would certainly help in the understanding of Cauchy horizons and in future studies of these horizons.

IV. COMPACT HORIZONS

In the previous sections we have not made any curvature assumptions. In this section we will assume $\mathrm{Ric}(V,V) \geqslant 0$ for all null V which is the *null convergence condition*. We will also assume that we have a Cauchy horizon $H^+(S)$ which is compact and of class C^2 on an open subset G such that the complement of G is of three-dimensional measure zero.

Theorem 4.1: *Let (M,g) be a space–time and assume that (M,g) satisfies the null convergence condition (i.e.,* $\mathrm{Ric}(V,V) \geqslant 0$ *for all null V). Let S be a partial Cauchy surface. Assume that the future horizon* $H^+(S)$ *is compact and contains an open set G where it is* C^2 *and is such that the complement of G in* $H^+(S)$ *has three-dimensional measure zero. Then,* $H^+(S)$ *has no end points and hence is differentiable of (at least) class* C^1 *for all points.*

Proof: The compactness of $H^+(S)$ yields the past completeness of all null generators.[8] We use the notation of Hawking[7] to obtain a map $u_t : G \to G$ and[7]

$$d/dt \int_{u_t(G)} dA = 2 \int_{u_t(G)} \rho dA. \tag{4.1}$$

Notice that since $H^+(S)$ is compact we find that $\int_{u_t(G)} dA$ is finite. However, the derivative of $\int_{u_t(G)} dA$ cannot be positive since the set G is mapped into itself. Thus, the left-hand side of Eq. (4.1) is nonpositive. On the other hand, the right-hand side of Eq. (4.1) must be non-negative since $\rho \geqslant 0$. Assume now that $H^+(S)$ has an end point p of a null generator γ. Even if γ does not lie in G, the horizon will be differentiable on the part of γ in the past of p. Choose some q on γ in the past of p and some u on γ in the future of p. Then for some small neighborhood $W(q)$ of q on $H^+(S)$, all null generators though points r of $W(q)$ will have directions close to the direction of

γ at q. Recall that given a compact domain set in the t-axis, geodesics with close initial conditions remain close on the compact domain set. Thus, by choosing $W(q)$ sufficiently small we may get all null generators though points of $W(q)$ come arbitrarily close to u. Thus for sufficiently small $W(q)$ all null generators intersecting $W(q)$ must leave $H^+(S)$ in the future. Since $W(q)$ is open in $H^+(S)$ it must have a nontrivial intersection with the set G. Thus, $u_t(G)$ cannot be all of G for some positive values of t and this yields a negative for some values of t on the left-hand side of Eq. (4.1), in contradiction. Thus, $H^+(S)$ has no end points and by Proposition 3.6 must be (at least) of class C^1 at all points.

ACKNOWLEDGMENTS

The authors wish to thank P. T. Chruściel and G. Galloway for helpful discussions. A.K. wishes to thank Albert Einstein Institute for hospitality where part of this work was done and acknowledge the support of the Polish Science Committee through Grant No. KNB 2 P301 050 07. Furthermore, J.K.B. wishes to thank the Erwin Schrödinger Institute of Mathematical Physics for partial support.

[1] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian Geometry, Second Edition*, Pure and Applied Mathematics Vol. 202 (Dekker, New York, 1996).
[2] J. K. Beem and A. Królak, *Cornelius Lanczos International Centenary Conference Proceedings, Raleigh, NC*, edited by J. D. Brown, M. T. Chu, D. C. Ellison, and R. J. Plemmons, (SIAM, Philadelphia, 1994), p. 515.
[3] R. P. Brady and E. Poisson, Class. Quantum Grav. **9**, 121 (1992).
[4] S. Chandrasekher and J. B. Hartle, Proc. R. Soc. London, Ser. A **384**, 301 (1982).
[5] P. T. Chruściel and G. Galloway, Commun. Math. Phys. **193**, 449 (1998).
[6] S. W. Hawking, Commun. Math. Phys. **25**, 152 (1972).
[7] S. W. Hawking, Phys. Rev. D **46**, 603 (1992).
[8] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space–time* (Cambridge University Press, Cambridge, 1973).
[9] A. Królak, Gen. Relativ. Gravit. **16**, 365 (1984).
[10] A. Królak and W. Rudnicki, Int. J. Theor. Phys. **32**, 137 (1993).
[11] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[12] B. A. O'Neill, *Semi-Riemannian geometry*, Pure and Applied Mathematics Vol. 103 (Academic, New York, 1983).
[13] J. K. Beem, Gen. Relativ. Gravit. **27**, 93 (1995).
[14] P. T. Chruściel and J. Isenberg, Commun. Anal. Geom. (in press).
[15] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. (McGraw–Hill, New York, 1976).
[16] E. A. Coddington and N. Levinson *Theory of Ordinary Differential Equations* (MacGraw–Hill, New York, 1955).
[17] S. Lang, *Real Analysis* (Addison–Wesley, Reading, MA, 1969).
[18] L. J. Corwin and R. H. Szczarba, *Calculus in Vector Spaces*, Pure and Applied Mathematics Vol. 52 (Dekker, New York, 1979).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 11 NOVEMBER 1998

Symmetries of asymptotically flat electrovacuum space–times and radiation

J. Bičák[a] and A. Pravdová[b]
Department of Theoretical Physics, Faculty of Mathematics and Physics, Charles University, V Holešovičkách 2, 180 00 Prague 8, Czech Republic

(Received 24 June 1998; accepted for publication 8 July 1998)

Symmetries compatible with asymptotic flatness and admitting gravitational and electromagnetic radiation are studied by using the Bondi–Sachs–van der Burg formalism. It is shown that in axially symmetric electrovacuum space–times in which at least locally a smooth null infinity in the sense of Penrose exists, the only second allowable symmetry is either the translational symmetry or the boost symmetry. Translationally invariant space–times with, in general, a straight ''cosmic string'' along the axis of symmetry are nonradiative although they can have a nonvanishing news function. The boost-rotation symmetric space–times are radiative. They describe ''uniformly accelerated charged particles'' or black holes which in general may also be rotating—the axial and an additional Killing vector are not assumed to be hypersurface orthogonal. The general functional forms of both gravitational and electromagnetic news functions, and of the mass aspect and total mass of asymptotically flat boost-rotation symmetric space–times at null infinity are obtained. The expressions for the mass are new even in the case of vacuum boost-rotation symmetric space–times with hypersurface orthogonal Killing vectors. In the Appendices some errors appearing in previous works are corrected. © *1998 American Institute of Physics.* [S0022-2488(98)00311-9]

I. INTRODUCTION AND SUMMARY

There is only one class of explicitly known radiative solutions which are asymptotically flat in the sense that null infinity is global, i.e., admits spherical sections, though its generators are not complete: boost-rotation symmetric space–times describing ''uniformly accelerated particles'' of various kinds.[1] A unique role of these solutions is exhibited by a theorem which roughly states that in axially symmetric, asymptotically flat space–times (in the sense that at least a local null infinity exists) the only additional symmetry that does not exclude radiation is the boost symmetry.[2]

The boost-rotation symmetric space–times have been used in various contexts. From a mathematical point of view, these solutions contain the only known space–times in which *arbitrarily strong initial data* with the given symmetry can be chosen on a hyperboloidal hypersurface which lead to the complete, smooth null infinity and regular timelike infinity in future. From a more physical point of view these are—and most probably will long remain—the only exact solutions of Einstein's equations for which one can find such quantities as angular distribution of gravitational radiation emitted by particles (represented by singularities) or by uniformly accelerated black holes.[3] In numerical calculations these space–times have been employed as important test beds in the null cone version of numerical relativity,[4,5] and, most recently, also in the standard approach based on a spacelike initial hypersurface.[6] In the context of quantum gravity the boost-rotation symmetric space–times (as ''generalized C-metric'') have been used to describe production of black-hole pairs in strong background fields (see, e.g., Ref. 7). We refer to the recent review[8] of exact radiative space–times and to the comprehensive analysis of the general structure of the boost-rotation symmetric space–times[1] for more details and references on both the history and recent developments in these issues.

Until now all work on the general properties of the boost-rotation symmetric space–times has

[a]Electronic mail: bicak@mbox.troja.mff.cuni.cz
[b]Electronic mail: pravdova@hp03.troja.mff.cuni.cz

 © 1998 American Institute of Physics

concentrated on the vacuum case with two non-null, hypersurface orthogonal Killing vectors—the rotational (axial) Killing vector and the boost Killing vector. The metric of such space–times, in suitable coordinates, contains just two functions. It can be constructed starting from the appropriately behaved boost-rotation symmetric solution of the flat-space wave equation with sources which is satisfied by one of the functions. The other metric function can then be determined by a quadrature (see Ref. 1). In the present paper we start the systematic study of more general boost-rotation symmetric space–times. We assume that an *electromagnetic field* coupled to gravity may also be present; and we consider *Killing vectors* which need *not* be *hypersurface orthogonal*. Even if only one of these extensions is taken into account, the field equations become fully nonlinear—there is no flat-space wave equation available. The inclusion of an electromagnetic field is also of interest because there exist the analogous features of gravitational fields due to uniformly accelerated masses in general relativity and of electromagnetic fields due to uniformly accelerated charges in special relativity. The study of boost-rotation symmetric fields within the Einstein–Maxwell theory treats both gravitational and electromagnetic fields from a unified point of view.

It is known that the example of a boost-rotation symmetric electrovacuum solution with Killing vectors which are not hypersurface orthogonal exists—this is the charged ''rotating'' C-metric.[9] Here we consider general solutions. Our main result is the theorem which roughly states that even under the presence of electromagnetic field and ''rotating sources,'' the boost symmetry is the only one which can be combined with axial symmetry and radiation exists. However, we shall also obtain other results (e.g., for the news function and total mass) some of which are new even in the vacuum case with hypersurface orthogonal Killing vectors.

In Sec. II we start out from the general form of axially symmetric metric and electromagnetic field in Bondi–Sachs coordinates $\{u,r,\theta,\phi\}$, where $u=\text{const}$ labels null hypersurfaces (in flat space $u=t-r$ is the usual retarded time), r is a luminosity distance along null rays $u=\text{const}$, $\theta=\text{const}$, $\phi=\text{const}$, and θ, ϕ are standard spherical angles. We consider their asymptotic expansions at $r\to\infty$ as they follow from the Einstein–Maxwell equations under the assumption of asymptotic flatness. Since we wish to study space–times in which the axial Killing vector is in general not hypersurface orthogonal we cannot employ the original work of Bondi *et al.*[10] but have to start from its generalization by Sachs[10] or, rather, from van der Burg's[11] explicit treatment of the asymptotic behavior of the coupled Einstein–Maxwell fields in the Bondi–Sachs coordinates. Now van der Burg's work involves complicated equations in which many errors and misprints appear. These do not change basic conclusions of his paper but we need correct forms. Therefore, in Appendix A we first summarize all Einstein–Maxwell equations in the Bondi–Sachs coordinates in general space–time and then give asymptotic forms of their solutions under the assumption of axial symmetry. (All equations were checked by using MAPLE V.) The structure of the field equations (the splitting into twelve main equations and five supplementary conditions), the total quantities (the mass and the charge) and their fluxes given in terms of two gravitational and two electromagnetic news functions are also briefly reviewed in Appendix A. The resulting formulas are used extensively in Sec. II to prove the theorem which states the following:

Suppose that an axially symmetric electrovacuum space–time admits a ''piece'' of null infinity. If this space–time admits an additional Killing vector forming with the axial Killing vector a two-dimensional Lie algebra, then the additional Killing vector either generates a supertranslation or it is the boost Killing vector.

The theorem is proven by first decomposing the additional Killing vector field η^α in the null (Sachs) tetrad and then solving the Killing equations asymptotically in the leading terms. In order to illustrate the form of the Killing equations, the lengthiest among them, $\mathscr{L}_\eta g_{00}=0$, is written down in Appendix B.

In Sec. III the case of the supertranslational Killing field is considered further. By solving Killing equations in the higher orders in r^{-k} and considering also asymptotic solutions of equations $\mathscr{L}_\eta F_{\mu\nu}=0$ (assuming thus that the electromagnetic field shares the same symmetry), we show that the supertranslational Killing field in fact has to be the generator either of translations along the axis of axial symmetry (z-axis) or time translations. The resulting space–times are nonradiative. Somewhat surprisingly perhaps, the news functions of the system need not necessarily be vanishing. This is not because of cylindrical waves, which of course are symmetric under translations along the z-axis. Cylindrical waves are excluded from our analysis if we assume that

the local null infinity exists at $\theta=\pi/2$; there are no cylindrical space–times which admit a regular cross section of null infinity. (We refer to Refs. 12 and 13 for the study of null infinity of cylindrically symmetric space–times and to our forthcoming work in which the Killing equations will be considered in greater detail in further orders in r^{-k} for translational Killing vectors.) Here we consider space–times with a straight cosmic string along the z-axis which, as shown by Bičák and Schmidt,[14] have a nonvanishing news function independent of time. If we assume that the electromagnetic field is regular at the axis of symmetry, only the gravitational news function is nonvanishing. We give the explicit form of the translational Killing vector in this case. This translational Killing vector in the case of the string has not been given in Ref. 2 because of a sign error. The error and misprints appearing in Ref. 2 are corrected in Appendix C. In Appendix D the translations in Bondi's coordinates in a space–time with a straight, nonrotating cosmic string are analyzed in detail. At the end of Sec. III we give explicitly the asymptotic form of both gravitational and electromagnetic fields in the Bondi–Sachs coordinates for stationary space–times without a string.

Section IV is devoted to the case when the additional Killing vector is the boost Killing vector. By analyzing the Killing equations in further orders in r^{-k} and considering also equations $\mathscr{L}_\eta F_{\mu\nu}=0$, we discover that both gravitational and electromagnetic news functions must have the same functional form—they depend on u and θ so that the news function $=f(\sin\theta/u)u^{-2}$, where f is an arbitrary function of its argument. We also derive the general form of the mass aspect and total (Bondi) mass for boost-rotation symmetric space–times. The mass aspect is given by

$$M(u,\theta)=\frac{1}{2\sin\theta}(w^2\mathscr{K}_{,w})_{,w}+\frac{\mathscr{L}}{u^3}, \tag{1}$$

where $w=\sin\theta/u$, $\mathscr{K}(w)$ is an arbitrary function, and $\mathscr{L}(w)=\lambda(w)/w^3$, where $\lambda(w)$ satisfies the simple equation (120) in the main text. $\lambda(w)$ can be determined if the news functions are known. The total Bondi mass is then given by

$$m(u)=\frac{1}{4}\int_0^\pi (w^2\mathscr{K}_{,w})_{,w}\,d\theta+\frac{1}{2}\int_0^\pi \frac{\mathscr{L}w}{u^2}d\theta. \tag{2}$$

The formulas for the mass aspect and total mass are new even in the case of hypersurface orthogonal Killing vectors studied in Ref. 2. Also, the boost Killing vector is here expanded to further orders than in Ref. 2 [see Eq. (122)].

Let us note that in general Bondi's news function has an invariant meaning only if null infinity is topologically $S^2\times\mathrm{R}$. However, as shown in Ref. 14, in axially symmetric space–time relative to a chosen $\partial/\partial\phi$ Killing vector Bondi news and translations have an invariant meaning even locally.

Our conventions for the Riemann and Ricci tensors follow those of Ref. 15 but our signature is -2.

II. AXIALLY SYMMETRIC SPACE–TIMES WITH ANOTHER SYMMETRY

Consider an axially symmetric electrovacuum space–time with circular group orbits; denote the corresponding Killing vector field by $\partial/\partial\phi$. Assume that at least the ''piece of $\mathscr{I}^+$'' exists in the sense of Ref. 16. Then one can introduce the Bondi–Sachs coordinate system $\{u,r,\theta,\phi\}\equiv\{x^0,x^1,x^2,x^3\}$ in which the metric satisfying the Einstein–Maxwell equations has the form[10,11]

$$\begin{aligned}ds^2=&\left(\frac{V}{r}e^{2\beta}-r^2e^{2\gamma}U^2\cosh 2\delta-r^2e^{-2\gamma}W^2\cosh 2\delta-2r^2UW\sinh 2\delta\right)du^2\\&+2e^{2\beta}\,du\,dr+2r^2(e^{2\gamma}U\cosh 2\delta+W\sinh 2\delta)du\,d\theta\\&+2r^2(e^{-2\gamma}W\cosh 2\delta+U\sinh 2\delta)\sin\theta\,du\,d\phi\\&-r^2[\cosh 2\delta(e^{2\gamma}d\theta^2+e^{-2\gamma}\sin^2\theta\,d\phi^2)+2\sinh 2\delta\sin\theta\,d\theta\,d\phi],\end{aligned} \tag{3}$$

where six metric functions U, V, W, β, γ, δ and the Maxwell field $F_{\mu\nu}$ do not depend on ϕ because of axial symmetry.

The complete set of the Einstein–Maxwell equations and the asymptotic expansions of the metric functions and electromagnetic field components at large r are given in Appendix A. We shall need these expansions, in which also the field equations are used, up to the following orders (all ''coefficients'' c, d, M, e, f, ... below being in general functions of u and θ):

$$\begin{aligned}
\gamma &= \frac{c}{r} + O(r^{-3}), \\
\delta &= \frac{d}{r} + O(r^{-3}), \\
\beta &= -\frac{1}{4}(c^2+d^2)\frac{1}{r^2} + O(r^{-4}), \\
U &= -(c_{,\theta} + 2c\cot\theta)\frac{1}{r^2} + O(r^{-3}), \\
W &= -(d_{,\theta} + 2d\cot\theta)\frac{1}{r^2} + O(r^{-3}), \\
V &= r - 2M + O(r^{-1}).
\end{aligned} \tag{4}$$

The leading terms in the electromagnetic field tensor are given by

$$\begin{aligned}
F_{12} &= \frac{e}{r^2} + (2E + ec + fd)\frac{1}{r^3} + O(r^{-4}), \\
F_{13} &= \left(\frac{f}{r^2} + (2F + ed - fc)\frac{1}{r^3} + O(r^{-4})\right)\sin\theta, \\
F_{01} &= -\frac{\epsilon}{r^2} + (e_{,\theta} + e\cot\theta)\frac{1}{r^3} + O(r^{-4}), \\
F_{23} &= \left(-\mu - (f_{,\theta} + f\cot\theta)\frac{1}{r} + O(r^{-2})\right)\sin\theta, \\
F_{02} &= X + (\epsilon_{,\theta} - e_{,u})\frac{1}{r} + O(r^{-2}), \\
F_{03} &= \left(Y - \frac{f_{,u}}{r} + O(r^{-2})\right)\sin\theta.
\end{aligned} \tag{5}$$

The mass aspect $M(u,\theta)$ is connected with the two gravitational news functions $c_{,u}$ and $d_{,u}$, and with the electromagnetic news functions X and Y by the relation

$$M_{,u} = -(c_{,u}^2 + d_{,u}^2) - (X^2+Y^2) + \tfrac{1}{2}(c_{,\theta\theta} + 3c_{,\theta}\cot\theta - 2c)_{,u}. \tag{6}$$

As a consequence of the Einstein–Maxwell equations the electromagnetic functions e, f, ϵ, μ and X, Y are connected by the relations

$$e_{,u} = \tfrac{1}{2}\epsilon_{,\theta} - (cX + dY), \tag{7}$$

$$f_{,u} = -\tfrac{1}{2}\mu_{,\theta} - (-cY + dX), \tag{8}$$

$$\epsilon_{,u} = -X_{,\theta} - X\cot\theta, \tag{9}$$

$$\mu_{,u} = -Y_{,\theta} - Y\cot\theta. \tag{10}$$

The gravitational news functions enter the evolution equations for e and f. The evolution equations for functions E, F are given in Appendix A. There the evolution of the whole Einstein–Maxwell system is summarized and the interpretation of the mass and the charges of the system is recalled. Here let us only notice that both the gravitational news functions $c_{,u}$, $d_{,u}$ and the electromagnetic news functions X, Y are freely specifiable functions.

Now we wish to emphasize that we assume that Eqs. (3)–(10) are valid for all $\phi\in[0,2\pi)$, however, they need *not* be true for all $\theta\in[0,\pi]$. Since we want to admit space–times with only ''local'' $\mathscr{I}^+$, we assume Eqs. (3)–(10) to be satisfied in some open interval of θ, i.e., not necessarily on the whole sphere. In particular, the ''axis of symmetry'' ($\theta=0,\pi$) may contain some nodal singularities and thus need not be regular. The regularity conditions on the axis [which can be obtained by transforming (3) into local Minkowskian coordinates] on functions γ, V, β ... need not be satisfied for any u. If the axis is singular then at least two generators of $\mathscr{I}^+$ would be missing so that $\mathscr{I}^+$ would not be topologically $S^2\times R$ (this is exactly the case which can occur in the boost-rotation symmetric space–times discussed in Sec. IV).

Let us now assume that another Killing vector field η exists which forms together with $\xi=\partial/\partial\phi$ a two-parameter group. In Ref. 2 it is proved (see the Lemma in Sec. 2) that in the case of ξ with circles as integral curves, ξ and η determine an Abelian Lie algebra so that we can assume $[\eta,\xi]=0$. Hence, the components of η^α are independent of ϕ.

Introduce the standard null tetrad field[17] $\{k^\alpha,m^\alpha,t^\alpha,\bar{t}^\alpha\}$, with the overbar denoting the complex conjugation, where $k_\alpha=\partial_\alpha u$, $m^\alpha k_\alpha=1$, $m^\alpha m_\alpha=0$, and the complex vector $t^\alpha=t^\alpha_R+it^\alpha_I$ (the subscripts R and I denoting the real and imaginary parts) obeys $t^\alpha\bar{t}_\alpha=-1$, $t^\alpha t_\alpha=t^\alpha k_\alpha=t^\alpha m_\alpha=0$. A convenient choice of the tetrad reads

$$k_\alpha=[1,0,0,0],\quad m_\alpha=[\tfrac{1}{2}Vr^{-1}e^{2\beta},e^{2\beta},0,0],$$

$$t_\alpha=\tfrac{1}{2}r(\cosh 2\delta)^{-1/2}[(1+\sinh 2\delta)e^{\gamma}U+\cosh 2\delta e^{-\gamma}W+i[(1-\sinh 2\delta)e^{\gamma}U-\cosh 2\delta e^{-\gamma}W], \tag{11}$$

$$0,\quad -(1+\sinh 2\delta+i(1-\sinh 2\delta))e^{\gamma},\quad -(1-i)\cosh 2\delta\sin\theta e^{-\gamma}].$$

This choice indeed implies $g_{\mu\nu}=2k_{(\mu}m_{\nu)}-2\bar{t}_{(\mu}t_{\nu)}$, with $g_{\mu\nu}$ given by (3).

We now decompose the Killing vector field η^α in the null tetrad,

$$\eta^\alpha=Ak^\alpha+Bm^\alpha+Ct^\alpha+\bar{C}\bar{t}^\alpha, \tag{12}$$

or

$$\eta^\alpha=Ak^\alpha+Bm^\alpha+\tilde{f}(t^\alpha_R+t^\alpha_I)+\tilde{g}(t^\alpha_R-t^\alpha_I), \tag{13}$$

where A, B, $C=C_R+iC_I$, $\tilde{f}=C_R-C_I$, $\tilde{g}=C_R+C_I$ are general functions of u, r, θ. The Killing equations

$$\mathscr{L}_\eta g_{\alpha\beta}=0 \tag{14}$$

can then be rewritten in the form

$$\begin{aligned}0=\mathscr{L}_\eta g_{\alpha\beta}=&A\mathscr{L}_k g_{\alpha\beta}+B\mathscr{L}_m g_{\alpha\beta}+2A_{,(\alpha}k_{\beta)}+2B_{,(\alpha}m_{\beta)}\\&+\tilde{f}[(\mathscr{L}_t g_{\alpha\beta})_R+(\mathscr{L}_t g_{\alpha\beta})_I]+2\tilde{f}_{,(\alpha}[t_{R\beta)}+t_{I\beta)}]\\&+\tilde{g}[(\mathscr{L}_t g_{\alpha\beta})_R-(\mathscr{L}_t g_{\alpha\beta})_I]+2\tilde{g}_{,(\alpha}[t_{R\beta)}-t_{I\beta)}].\end{aligned} \tag{15}$$

The easiest is the equation

$$\mathscr{L}_\eta g_{11}=2e^{2\beta}B_{,r}=0, \tag{16}$$

which implies

$$B=B(u,\theta). \tag{17}$$

In the hypersurface orthogonal case, analyzed in Ref. 2, also the equations $\mathscr{L}_\eta g_{03}=\mathscr{L}_\eta g_{13}=\mathscr{L}_\eta g_{23}=0$ can easily be solved. This is not so if $\partial/\partial\phi$ is not hypersurface orthogonal. All Killing equations other than $\mathscr{L}_\eta g_{11}=0$ now become complicated. For illustration, the equation $\mathscr{L}_\eta g_{00}=0$ is written down in Appendix B; the other equations can be found in Ref. 18.

We assume that the coefficients A, $\tilde{f}$, $\tilde{g}$ can be expanded in powers of r^{-k} and solve the Killing equations asymptotically. Denoting by $A^{(k)}$, $f^{(k)}$, $g^{(k)}$ the coefficients of r^{-k} in the expansions, we first notice that $\mathscr{L}_\eta g_{22}=0$, $\mathscr{L}_\eta g_{12}=0$, $\mathscr{L}_\eta g_{13}=0$ imply

$$\begin{aligned} A&=A^{(-1)}r+A^{(0)}+\frac{A^{(1)}}{r}+O(r^{-2}),\\ \tilde{f}&=f^{(-1)}r+f^{(0)}+\frac{f^{(1)}}{r}+O(r^{-2}),\\ \tilde{g}&=g^{(-1)}r+g^{(0)}+\frac{g^{(1)}}{r}+O(r^{-2}), \end{aligned} \tag{18}$$

where $A^{(k)}$, $f^{(k)}$, $g^{(k)}$ are functions of u and θ. The remaining nine Killing equations imply the conditions on functions B, $A^{(k)}$, $f^{(k)}$, $g^{(k)}$. The procedure of their solutions is similar to that in Ref. 2 (cf. Eqs. (20)–(24) therein). However, the equations are now more complicated and there are three additional equations to be satisfied.

In the leading orders in r^{-k} the Killing equations can be written down easily (we omit equations $\mathscr{L}_\eta g_{12}=0$, $\mathscr{L}_\eta g_{13}=0$ since they do not restrict the leading terms in η^α):

$$\mathscr{L}_\eta g_{00}=0 \quad (r^1): \quad A^{(-1)}{}_{,u}=0, \tag{19}$$

$$\mathscr{L}_\eta g_{01}=0 \quad (r^0): \quad B_{,u}+A^{(-1)}=0, \tag{20}$$

$$\mathscr{L}_\eta g_{02}=0 \quad (r^2): \quad f^{(-1)}{}_{,u}=0, \tag{21}$$

$$\mathscr{L}_\eta g_{03}=0 \quad (r^2): \quad g^{(-1)}{}_{,u}=0, \tag{22}$$

$$\mathscr{L}_\eta g_{22}=0 \quad (r^2): \quad f^{(-1)}{}_{,\theta}+A^{(-1)}=0, \tag{23}$$

$$\mathscr{L}_\eta g_{23}=0 \quad (r^2): \quad -g^{(-1)}{}_{,\theta}+g^{(-1)}\cot\theta=0, \tag{24}$$

$$\mathscr{L}_\eta g_{33}=0 \quad (r^2): \quad f^{(-1)}\cot\theta+A^{(-1)}=0. \tag{25}$$

The system of equations (19)–(21), (23), (25) is, at this order, identical to Eqs. (20)–(24) in Ref. 2. The solutions are thus identical, reading

$$\begin{aligned} A^{(-1)}&=k\cos\theta,\\ f^{(-1)}&=-k\sin\theta,\\ B&=-ku\cos\theta+\alpha(\theta), \end{aligned} \tag{26}$$

where $k=\text{const}$ and α is an arbitrary function of θ. Equations (22) and (24) imply

$$g^{(-1)}=h\sin\theta, \tag{27}$$

where $h=\text{const}$.

Regarding Eqs. (11), (13) and (18) we find $\eta^{\phi}=h+O(r^{-1})$. Since the contribution of h to the vector field η^{α} is just $\eta^{\phi}=\text{const}$, which is a constant multiple of the axial Killing vector $\partial/\partial\phi$, we may, without loss of generality, put $h=0$. Therefore, in the lowest order in r^{-1} the general asymptotic form of the Killing vector η turns out to be

$$\eta^{\alpha}=[-ku\cos\theta+\alpha(\theta),\quad kr\cos\theta+O(r^0),\quad -k\sin\theta+O(r^{-1}),\quad O(r^{-1})], \tag{28}$$

where k is a constant, α – an arbitrary function of θ. Thus, in the leading order of the asymptotic expansion the presence of an electromagnetic field satisfying the boundary conditions (5) and the fact that the axial Killing vector need not be hypersurface orthogonal do not change the conclusion obtained in Ref. 2 in the vacuum case with hypersurface orthogonal $\partial/\partial\phi$. When $k=0$, the vector field (28) generates supertranslations. We shall see in Sec. III that the resulting space–times are then nonradiative.

Assuming $k\neq 0$, it is then easy to find a Bondi–Sachs coordinate system with $\alpha=0$ by making a supertranslation $\tilde{u}=u+\hat{\alpha}(\theta)$, $\tilde{r}=r$, $\tilde{\theta}=\theta$, $\tilde{\phi}=\phi$, where $\hat{\alpha}$ satisfies the equation $-\hat{\alpha}_{,\theta}\ k\sin\theta+\hat{\alpha}k\cos\theta=-\alpha$ (cf. Ref. 2). [Notice that no singularity arises for $\hat{\alpha}$ even if $\alpha(0)\neq 0$.] We then obtain $\eta^{\tilde{u}}=-k\tilde{u}\cos\theta$. Hence, we put $\alpha=0$ in Eq. (28) and without loss of generality (buno) we choose $k=1$. Then $B=-u\cos\theta$, $A^{(-1)}=\cos\theta$, $f^{(-1)}=-\sin\theta$ and $g^{(-1)}=0$. The asymptotic form of the Killing vector field η is

$$\eta^{\alpha}=[-u\cos\theta,\quad r\cos\theta+O(r^0),\quad -\sin\theta+O(r^{-1}),\quad O(r^{-1})], \tag{29}$$

which is the *boost Killing vector*. It generates the Lorentz transformations along the axis of axial symmetry.

We have thus proven the following:

Theorem: Suppose that an axially symmetric electrovacuum space–time admits a ''piece'' of $\mathscr{I}^+$ in the sense that the Bondi–Sachs coordinates can be introduced in which the metric takes the form (3)–(4) and the asymptotic form of the electromagnetic field is given by (5). If this space–time admits an additional Killing vector forming with the axial Killing vector a two-dimensional Lie algebra, then the additional Killing vector has asymptotically the form (28). For $k=0$ it generates a supertranslation; for $k\neq 0$ it is the boost Killing field.

On $\mathscr{I}^+$ in the coordinates u, $l=r^{-1}$, θ, and ϕ the boost Killing vector reads (cf. Ref. 2)

$$\eta^{\alpha}_{|\mathscr{I}^+}=[-u\cos\theta,\ \ 0,\ \ -\sin\theta,\ \ 0]. \tag{30}$$

III. THE SUPERTRANSLATIONAL KILLING FIELD

In this section we shall show that if the Killing field (28) for $k=0$ is the supertranslational Killing field, it has, in fact, to be the generator of translations and the resulting space–time is thus nonradiative.

Assuming $k=0$, and considering the Killing equations (14) in the higher orders in r^{-k} than in Eqs. (19)–(25), and taking into account the metric expansions (4), we find the following restrictions on the expansion coefficients of functions A, $\tilde{f}$ and $\tilde{g}$:

$$\mathscr{L}_{\eta}g_{00}=0\quad (r^0):\quad A^{(0)}{}_{,u}=0, \tag{31}$$

$$(r^{-1}):\quad A^{(1)}{}_{,u}-f^{(0)}{}_{,u}(c_{,\theta}+2c\cot\theta)-g^{(0)}{}_{,u}(d_{,\theta}+2d\cot\theta)=0, \tag{32}$$

$$\mathscr{L}_{\eta}g_{01}=0\quad (r^{-1}):\quad \text{satisfied automatically}, \tag{33}$$

$$(r^{-2}):\quad -B(cc_{,u}+dd_{,u})-BM-A^{(1)}+f^{(0)}(c_{,\theta}+2c\cot\theta)+g^{(0)}(d_{,\theta}+2d\cot\theta)=0, \tag{34}$$

$$\mathscr{L}_{\eta}g_{02}=0\quad (r^1):\quad f^{(0)}{}_{,u}=0, \tag{35}$$

$$(r^0):\ (A^{(0)}+\tfrac{1}{2}B)_{,\theta}+f^{(0)}c_{,u}-cf^{(0)}{}_{,u}-f^{(1)}{}_{,u}-2g^{(0)}{}_{,u}\,d=0, \tag{36}$$

$\mathscr{L}_\eta g_{03}=0$ $(r^1):\ g^{(0)}{}_{,u}=0,$ (37)

$$(r^0):\ -g^{(0)}c_{,u}+g^{(0)}{}_{,u}c-g^{(1)}{}_{,u}+2f^{(0)}d_{,u}=0, \tag{38}$$

$\mathscr{L}_\eta g_{12}=0$ $(r^0):\ f^{(0)}+B_{,\theta}=0,$ (39)

$$(r^{-1}):\ f^{(1)}-B(c_{,\theta}+2c\cot\theta)+g^{(0)}d=0, \tag{40}$$

$\mathscr{L}_\eta g_{13}=0$ $(r^0):\ g^{(0)}=0,$ (41)

$$(r^{-1}):\ g^{(1)}-B(d_{,\theta}+2d\cot\theta)-f^{(0)}d=0, \tag{42}$$

$\mathscr{L}_\eta g_{22}=0$ $(r^1):\ f^{(0)}{}_{,\theta}+Bc_{,u}+A^{(0)}-\tfrac{1}{2}B=0,$ (43)

$$(r^0):\ f^{(1)}{}_{,\theta}+B(cc_{,u}+2dd_{,u})-B(c_{,\theta}+2c\cot\theta)_{,\theta}+A^{(1)}+BM+2d(g^{(0)}{}_{,\theta}-g^{(0)}\cot\theta)=0, \tag{44}$$

$\mathscr{L}_\eta g_{23}=0$ $(r^1):\ 2Bd_{,u}+g^{(0)}{}_{,\theta}-g^{(0)}\cot\theta=0,$ (45)

$$(r^0):\ -2A^{(0)}d-4f^{(0)}d\cot\theta+B(d_{,\theta}+2d\cot\theta)_{,\theta}-B(d_{,\theta}+2d\cot\theta)\cot\theta+Bd+g^{(0)}{}_{,\theta}c-g^{(0)}(c_{,\theta}+c\cot\theta)-g^{(1)}{}_{,\theta}+g^{(1)}\cot\theta=0, \tag{46}$$

$\mathscr{L}_\eta g_{33}=0$ $(r^1):\ f^{(0)}\cot\theta-Bc_{,u}+A^{(0)}-\tfrac{1}{2}B=0,$ (47)

$$(r^0):\ \cot\theta[-B(c_{,\theta}+2c\cot\theta)+f^{(1)}-f^{(0)}c]-c_{,\theta}f^{(0)}+c(A^{(0)}-\tfrac{1}{2}B)+A^{(1)}+2Bdd_{,u}+BM=0. \tag{48}$$

First notice that Eqs. (31), (35), (37), (39), (41) and (45) imply

$$A^{(0)}=A^{(0)}(\theta),\quad f^{(0)}=f^{(0)}(\theta)=-B_{,\theta},\quad g^{(0)}=0,\quad d=d(\theta). \tag{49}$$

Using these results in the remaining equations we find

$$A^{(0)}=\tfrac{1}{2}(B_{,\theta\theta}+B_{,\theta}\cot\theta+B), \tag{50}$$

$$c_{,u}=\frac{1}{2B}(B_{,\theta\theta}-B_{,\theta}\cot\theta), \tag{51}$$

$$c=\frac{u}{2B}(B_{,\theta\theta}-B_{,\theta}\cot\theta)+\omega(\theta), \tag{52}$$

$$A^{(1)}=A^{(1)}(\theta), \tag{53}$$

$$f^{(1)}=B(c_{,\theta}+2c\cot\theta), \tag{54}$$

$$g^{(1)}=B(d_{,\theta}+2d\cot\theta)-B_{,\theta}d, \tag{55}$$

where $\omega(\theta)$ is an arbitrary function of θ; ω can be transformed away by a supertranslation $u=\overline{u}+\hat{\alpha}(\theta)$ with $\hat{\alpha}$ satisfying $\omega+(-\hat{\alpha}_{,\theta\theta}+\hat{\alpha}_{,\theta}\cot\theta)/2=0$. Equations (36) and (46) are now satisfied identically. Equations (34), (44) and (48) lead to only one further independent condition:

$$M=-cc_{,u}-B^{-1}[A^{(1)}+B_{,\theta}(c_{,\theta}+2c\cot\theta)]. \tag{56}$$

Important results follow from Eqs. (49) and (51). Equation (49) shows that the news function $d_{,u}=0$. Since $B=B(\theta)$, Eq. (51) implies that the time derivative of the news function $c_{,u}$ must vanish. Therefore, with the supertranslational Killing field, we arrive at the Weyl tensor (see, e.g., Ref. 11)

$$C_{\alpha\beta\gamma\delta}m^{\alpha}t^{\beta}m^{\gamma}t^{\delta}=[(c+id)_{,uu}]\frac{1}{r}+O(r^{-2}), \tag{57}$$

which is nonradiative since the first term proportional to r^{-1} vanishes.

Substituting the expansion of the metric functions (4) into the null tetrad (11) and coefficients A, B, $\tilde{f}$ and $\tilde{g}$ [given by Eqs. (49)–(56)] into Eq. (13), we find the expansion of the supertranslational Killing vector to be

$$\begin{aligned}\eta^{\mu}=\Bigg[&B(\theta),\tfrac{1}{2}(B_{,\theta\theta}+B_{,\theta}\cot\theta)\\&+\Bigg[-B_{,\theta\theta}-2B_{,\theta}B_{,\theta\theta\theta}+2B_{,\theta}{}^{2}B_{,\theta\theta}B^{-1}-2B_{,\theta}{}^{3}\cot\theta B^{-1}+B_{,\theta}{}^{2}\left(3\cot^{2}\theta-\frac{2}{\sin^{2}\theta}\right)\Bigg]\\&\times B^{-1}\frac{u}{4r}+O(r^{-2}),\ \ -B_{,\theta}\frac{1}{r}+B_{,\theta}\frac{c}{r^{2}}+O(r^{-3}),\ \ B_{,\theta}\frac{d}{r^{2}}\sin\theta+O(r^{-3})\Bigg].\end{aligned} \tag{58}$$

Let us now turn to the asymptotic properties of electromagnetic field. We assume $\mathcal{L}_{\eta}F_{\alpha\beta}=0$ where η is now the supertranslational Killing field (58). From the general expression of the Lie derivative of $F_{\mu\nu}$ with respect to general η^{α} decomposed into the null tetrad according to Eq. (13),

$$\begin{aligned}\mathcal{L}_{\eta}F_{\alpha\beta}=&A\mathcal{L}_{k}F_{\alpha\beta}+A_{,\alpha}k^{\gamma}F_{\gamma\beta}+A_{,\beta}k^{\gamma}F_{\alpha\gamma}+B\mathcal{L}_{m}F_{\alpha\beta}+B_{,\alpha}m^{\gamma}F_{\gamma\beta}+B_{,\beta}m^{\gamma}F_{\alpha\gamma}\\&+\tilde{f}[(\mathcal{L}_{l}F_{\alpha\beta})_{R}+(\mathcal{L}_{l}F_{\alpha\beta})_{I}]+\tilde{g}[(\mathcal{L}_{l}F_{\alpha\beta})_{R}-(\mathcal{L}_{l}F_{\alpha\beta})_{I}]\\&+\tilde{f}_{,\alpha}(t_{R}^{\gamma}+t_{I}^{\gamma})F_{\gamma\beta}+\tilde{f}_{,\beta}(t_{R}^{\gamma}+t_{I}^{\gamma})F_{\alpha\gamma}+\tilde{g}_{,\alpha}(t_{R}^{\gamma}-t_{I}^{\gamma})F_{\gamma\beta}+\tilde{g}_{,\beta}(t_{R}^{\gamma}-t_{I}^{\gamma})F_{\alpha\gamma},\end{aligned} \tag{59}$$

and after substituting for A, B, $\tilde{f}$, and $\tilde{g}$ in accordance with Eq. (58), we find that in the first orders in r^{-k} the Lie equations imply

$$\mathcal{L}_{\eta}F_{01}=0\ \ (r^{-2}):\ \ -\epsilon_{,u}B+XB_{,\theta}=0, \tag{60}$$

$$\mathcal{L}_{\eta}F_{02}=0\ \ (r^{0}):\ \ X_{,u}B=0, \tag{61}$$

$$\mathcal{L}_{\eta}F_{03}=0\ \ (r^{0}):\ \ (Y\sin\theta)_{,u}B=0, \tag{62}$$

$$\mathcal{L}_{\eta}F_{12}=0\ \ (r^{-2}):\ \ e_{,u}B+\epsilon B_{,\theta}=0, \tag{63}$$

$$\mathcal{L}_{\eta}F_{13}=0\ \ (r^{-2}):\ \ (f\sin\theta)_{,u}B-\mu\sin\theta B_{,\theta}=0, \tag{64}$$

$$\mathcal{L}_{\eta}F_{23}=0\ \ (r^{0}):\ \ -(\mu\sin\theta)_{,u}B+Y\sin\theta B_{,\theta}=0. \tag{65}$$

From Eq. (61) we get $X=X(\theta)$. Combining Eq. (60) and Maxwell equation (9) we get the equation for X, $(XB\sin\theta)_{,\theta}=0$, the solution being

$$X=\frac{x_{0}}{B\sin\theta},\quad x_{0}=\text{const.} \tag{66}$$

Regarding the last result and integrating Eq. (60), we obtain

$$\epsilon = \frac{x_0 B_{,\theta}}{B^2 \sin\theta} u + \epsilon_1(\theta), \tag{67}$$

where $\epsilon_1(\theta)$ is an arbitrary integration function. Equation (63) then implies

$$e = -\frac{x_0 B_{,\theta}^2}{2B^3 \sin\theta} u^2 - \frac{\epsilon_1 B_{,\theta}}{B} u + e_1(\theta), \tag{68}$$

where $e_1(\theta)$ is again an integration function.

Analogously, using Eqs. (62), (64), (65) and Maxwell equation (10) we find

$$Y = \frac{y_0}{B \sin\theta}, \quad y_0 = \text{const}, \tag{69}$$

$$\mu = \frac{y_0 B_{,\theta}}{B^2 \sin\theta} u + \mu_1(\theta), \tag{70}$$

$$f = \frac{y_0 B_{,\theta}^2}{2B^3 \sin\theta} u^2 + \frac{\mu_1 B_{,\theta}}{B} u + f_1(\theta), \tag{71}$$

where μ_1, f_1 are arbitrary functions of θ.

Comparing Eqs. (68) and (71) with Maxwell equations (7) and (8) we obtain restrictions on ϵ_1 and μ_1:

$$\epsilon_{1,\theta} + \frac{2\epsilon_1 B_{,\theta}}{B} - \frac{2y_0 d}{B \sin\theta} = 0, \tag{72}$$

$$\mu_{1,\theta} + \frac{2\mu_1 B_{,\theta}}{B} + \frac{2x_0 d}{B \sin\theta} = 0. \tag{73}$$

Now we assume that the Killing field η is bounded at the axis, i.e., $B(\theta)$ is bounded for $\theta = 0, \pi$ and the electromagnetic field (5), i.e., X and Y have to be bounded there, too. Therefore constants entering X and Y must vanish:

$$x_0 = y_0 = 0. \tag{74}$$

Taking the time derivative of Eq. (56), regarding Eqs. (49), (51), (53), and comparing with Eq. (6), we arrive at the following equation for function B:

$$\left[\frac{\sin^3\theta}{2B} \left(\frac{B_{,\theta}}{\sin\theta} \right)_{,\theta} \right]_{,\theta} = 0. \tag{75}$$

The general solution can be seen to be

$$B = a \sin\theta \left(\frac{\sin\theta}{\cos\theta + 1} \right)^C + b \sin\theta \left(\frac{\sin\theta}{\cos\theta + 1} \right)^{-C}, \tag{76}$$

where a, b, $C > 0$ are constants. If we assume $C \in [0,1]$ then B is bounded at the axis.

The news function is obtained from (51): $c_{,u} = (C^2 - 1)/(2 \sin^2\theta)$. As described in Ref. 14 and in Appendix D it corresponds to the news function of a string; we exclude $C = 0$. Function B given in Eq. (76) corresponds to translations along the z-axis and t-axis in the space–time of an infinite thin cosmic string described by the deficit angle $2\pi(1 - C)$, $C \in (0,1]$; in the weak-field limit, $C = 1 - 4\mu$, where μ is the mass per unit length of the string. (Notice that $\mu \geqslant 0$ for $C \in (0,1]$.) See Appendix D for details.

In Appendix E, part 1, the explicit form of electromagnetic field for general B, Eq. (76), representing translations in asymptotically flat space–times with a straight string is given.

If $C=1$, there is no string extending to infinity. Equation (76) gives

$$B=(b-a)\cos\theta+(b+a). \tag{77}$$

Then $c_{,u}=0$ and from Eqs. (49) and (50) we get

$$f^{(0)}=(b-a)\sin\theta, \quad A^{(0)}=\tfrac{1}{2}[-(b-a)\cos\theta+b+a]. \tag{78}$$

Consequently, regarding Eqs. (11), (13) and (18), we find the Killing vector field η^α to be asymptotically of the form

$$\eta^\alpha=\left[(b-a)\cos\theta+b+a,\quad -(b-a)\cos\theta+O(r^{-2}),\quad (b-a)\sin\theta\frac{1}{r}+O(r^{-3}),\quad O(r^{-3})\right]. \tag{79}$$

We thus see that the Killing vector field generates translations: with $b-a=0$ this is the time translation, with $a+b=0$ the translation along the z-axis.

Since there is no string extending to infinity, both functions $c=c(\theta)$ and $d=d(\theta)$ are independent of time and both news functions thus vanish; there is no radiation. By employing two transformations from the Bondi–Metzner–Sachs group, functions c and d can be transformed away.

For an illustration let us write down the asymptotic form of both gravitational and electromagnetic fields in the case of the Killing vector representing timelike translations in asymptotically flat space–time without a string; we thus assume $B=\text{const}$ and $c=d=0$. The resulting axially symmetric stationary metric and electromagnetic field then have the asymptotic form

$$\begin{aligned}ds^2=&\left(1-\frac{2M(\theta)}{r}+O(r^{-2})\right)du^2+2(1+O(r^{-4}))du\,dr+2O(r^{-1})du\,d\theta+2O(r^{-1})\sin\theta\,du\,d\phi\\&-r^2[(1+O(r^{-3}))d\theta^2+(1+O(r^{-3}))\sin^2\theta d\phi^2+2O(r^{-3})\sin\theta\,d\theta\,d\phi],\end{aligned}$$

$$F_{01}=-\frac{\epsilon_1}{r^2}+(e_{1,\theta}+e_1\cot\theta)\frac{1}{r^3}+O(r^{-4}),$$

$$F_{02}=O(r^{-2}),$$

$$F_{03}=O(r^{-2}), \tag{80}$$

$$F_{12}=\frac{e_1}{r^2}+O(r^{-3}),$$

$$F_{13}=f_1\sin\theta\frac{1}{r^2}+O(r^{-3}),$$

$$F_{23}=-\mu_1\sin\theta-(f_1\sin\theta)_{,\theta}\frac{1}{r}+O(r^{-2}).$$

Using the consequence of Einstein's equations (A46) and assuming $N_{,u}=0$ (which, as will be proven in our forthcoming publication, follows from further terms in r^{-k} in the Killing equations), we see that $M_{,\theta}=0$, i.e., $A^{(1)}(\theta)=\text{const}$. Constants M, ϵ_1 and μ_1 represent the total mass, electric charge and magnetic charge, respectively.

IV. THE BOOST KILLING VECTOR

In this section we shall find the form of the gravitational news functions $c_{,u}$ and $d_{,u}$ for the case of the boost Killing vector by expanding the Killing equations in further orders in r^{-1}. We shall also obtain the form of the electromagnetic news functions X and Y.

We thus assume the asymptotic form of the Killing vector to be given by Eq. (29). Expanding now the Killing equations (14) in higher orders of r^{-1} we obtain

$$\mathscr{L}_\eta g_{00}=0 \quad (r^0): \quad 2A^{(0)}{}_{,u}+B_{,u}=2A^{(0)}{}_{,u}-\cos\theta=0, \tag{81}$$

$$\mathscr{L}_\eta g_{01}=0 \quad (r^{-1}): \quad \text{satisfied automatically}, \tag{82}$$

$$\mathscr{L}_\eta g_{02}=0 \quad (r^1): \quad f^{(-1)}c_{,u}-f^{(0)}{}_{,u}+A^{(-1)}{}_{,\theta}=-\sin\theta\, c_{,u}-f^{(0)}{}_{,u}-\sin\theta=0, \tag{83}$$

$$\mathscr{L}_\eta g_{03}=0 \quad (r^1): \quad -g^{(0)}{}_{,u}+2f^{(-1)}d_{,u}=-g^{(0)}{}_{,u}-2\sin\theta\, d_{,u}=0, \tag{84}$$

$$\mathscr{L}_\eta g_{12}=0 \quad (r^0): \quad f^{(0)}-f^{(-1)}c+B_{,\theta}=f^{(0)}+\sin\theta\, c+u\sin\theta=0, \tag{85}$$

$$\mathscr{L}_\eta g_{13}=0 \quad (r^0): \quad 2f^{(-1)}d-g^{(0)}=-2\sin\theta\, d-g^{(0)}=0, \tag{86}$$

$$\begin{aligned}\mathscr{L}_\eta g_{22}=0 \quad (r^1): \quad & f^{(0)}{}_{,\theta}+Bc_{,u}+A^{(0)}-\tfrac{1}{2}B \\ & =f^{(0)}{}_{,\theta}-u\cos\theta\, c_{,u}+A^{(0)}+\tfrac{1}{2}u\cos\theta=0,\end{aligned} \tag{87}$$

$$\begin{aligned}\mathscr{L}_\eta g_{23}=0 \quad (r^1): \quad & -g^{(0)}{}_{,\theta}+g^{(0)}\cot\theta-2Bd_{,u}-2df^{(-1)}\cot\theta \\ & =-g^{(0)}{}_{,\theta}+g^{(0)}\cot\theta+2u\cos\theta\, d_{,u}+2d\cos\theta=0,\end{aligned} \tag{88}$$

$$\begin{aligned}\mathscr{L}_\eta g_{33}=0 \quad (r^1): \quad & \tfrac{1}{2}B-A^{(0)}+f^{(-1)}(c_{,\theta}+2c\cot\theta)-f^{(0)}\cot\theta+Bc_{,u} \\ & =-\frac{u}{2}\cos\theta-A^{(0)}-\sin\theta(c_{,\theta}+2c\cot\theta)-f^{(0)}\cot\theta-u\cos\theta\, c_{,u}=0.\end{aligned} \tag{89}$$

From Eqs. (81) and (85) we immediately get

$$A^{(0)}=\frac{u}{2}\cos\theta+\rho(\theta), \tag{90}$$

$$c=-u-\frac{f^{(0)}}{\sin\theta}, \tag{91}$$

where $\rho(\theta)$ is an arbitrary function of θ. Equation (83) is just the u-derivative of Eq. (85). Equation (87) implies

$$c_{,u}=\frac{1}{u\cos\theta}\left(f^{(0)}{}_{,\theta}+A^{(0)}+\frac{u}{2}\cos\theta\right). \tag{92}$$

Using now Eqs. (90)–(92) in Eq. (89) we find that additive function $\rho(\theta)$ in Eq. (90) must vanish, i.e.,

$$A^{(0)}=\frac{u}{2}\cos\theta. \tag{93}$$

Comparing Eqs. (83) and (87) we obtain an equation for $f^{(0)}$,

$$uf^{(0)}{}_{,u}+\tan\theta\, f^{(0)}{}_{,\theta}+2u\sin\theta=0. \tag{94}$$

This equation can be solved by introducing a new variable

$$w=\frac{\sin\theta}{u}. \tag{95}$$

Equation (94) then becomes

$$\partial_u f^{(0)}(u,w)+2uw=0, \tag{96}$$

so that the general solution is

$$f^{(0)}=-u^2w+\mathscr{K}(w)=-u\sin\theta+\mathscr{K}(\sin\theta/u), \tag{97}$$

where $\mathscr{K}(w)$ is an arbitrary function of w. Consequently, Eq. (91) leads to

$$c(u,\theta)=-\frac{\mathscr{K}(w)}{uw}, \tag{98}$$

where w is given by Eq. (95). The news function thus reads

$$c_{,u}(u,\theta)=\frac{\mathscr{K}(w)_{,w}}{u^2}. \tag{99}$$

Before writing down the result for the second news function, let us compare expression (99) with the news function given for the case of the vacuum boost-rotation symmetric space–times with the hypersurface orthogonal Killing vectors in Ref. 2. There the news function is obtained in the form $c_{,u}=F(\tilde{U}/\sin\theta)/\sin^2\theta$, where F is an arbitrary function of $\tilde{U}/\sin\theta$ and the flat-space retarded time $\tilde{U}$ satisfies the equation $\tilde{U}_{,u}u\cot\theta=\tilde{U}\cot\theta-\tilde{U}_{,\theta}$. This equation can be rewritten as the equation $u\tilde{U}_{,u}+\tan\theta\,\tilde{U}_{,\theta}-\tilde{U}=0$, the solution of which is $\tilde{U}=\mathscr{A}(w)u$, $\mathscr{A}$ being function of $w=\sin\theta/u$. We can write $c_{,u}=F(\mathscr{A}(w)/w)/\sin^2\theta=F(\mathscr{A}(w)/w)/(w^2u^2)=\mathscr{K}_{,w}(w)/u^2$, where $\mathscr{K}_{,w}=F(\mathscr{A}(w)/w)/w^2$. Therefore, our result (99) for the general form of the news function is in agreement with Eq. (59) given in Ref. 2.

The second news function, $d_{,u}$, can be found analogously. Equation (84) is the u-derivative of Eq. (86) which gives

$$d=-\frac{g^{(0)}}{2\sin\theta}. \tag{100}$$

Substituting this result and $d_{,u}$ from Eq. (84) into Eq. (88), we obtain the equation for $g^{(0)}$,

$$ug^{(0)}_{,u}+\tan\theta g^{(0)}_{,\theta}=0, \tag{101}$$

which in terms of variables u and w simply yields

$$g^{(0)}=g^{(0)}(w). \tag{102}$$

The second news function is thus given by

$$d_{,u}(u,\theta)=\frac{g^{(0)}(w)_{,w}}{2u^2}, \tag{103}$$

where $g^{(0)}$ is an arbitrary function of $w=\sin\theta/u$.

In order to obtain the mass aspect and the total mass at null infinity we have to expand the Killing equations in higher orders in r^{-1}. Straightforward though rather lengthy calculations lead to the following system of equations (in which $A^{(-1)}$, $A^{(0)}$, B, $f^{(-1)}$, $f^{(0)}$, $g^{(0)}$, c and d are already known but are left unspecified for the sake of compactness):

$$\mathscr{L}_\eta g_{00}=0 \quad (r^{-1}): \quad A^{(1)}{}_{,u}-B_{,u}M+A^{(-1)}(cc_{,u}+dd_{,u}+M)-f^{(-1)}M_{,\theta}$$
$$-(d_{,\theta}+2d\cot\theta)(g^{(0)}{}_{,u}-2f^{(-1)}d_{,u})-(c_{,\theta}+2c\cot\theta)(f^{(0)}{}_{,u}$$
$$-f^{(-1)}c_{,u})=0, \tag{104}$$

$$\mathscr{L}_\eta g_{01}=0 \quad (r^{-2}): \quad (c_{,\theta}+2c\cot\theta)(f^{(0)}-f^{(-1)}c)-f^{(-1)}(dd_{,\theta}+cc_{,\theta})$$
$$-\tfrac{1}{2}B_{,u}(c^2+d^2)-A^{(1)}-BM-B(dd_{,u}+cc_{,u})=0, \tag{105}$$

$$\mathscr{L}_\eta g_{02}=0 \quad (r^{0}): \quad f^{(-1)}(cc_{,u}+2dd_{,u})-f^{(-1)}(c_{,\theta}+2c\cot\theta)_{,\theta}-f^{(-1)}{}_{,\theta}(c_{,\theta}+2c\cot\theta)$$
$$+A^{(0)}{}_{,\theta}+\tfrac{1}{2}B_{,\theta}-f^{(0)}{}_{,u}c+f^{(0)}c_{,u}-f^{(1)}{}_{,u}-2dg^{(0)}{}_{,u}=0, \tag{106}$$

$$\mathscr{L}_\eta g_{03}=0 \quad (r^{0}): \quad -g^{(0)}c_{,u}+g^{(0)}{}_{,u}c-g^{(1)}{}_{,u}+2f^{(0)}d_{,u}+2f^{(-1)}(2dc_{,u}-cd_{,u})$$
$$-f^{(-1)}(d_{,\theta}+2d\cot\theta)_{,\theta}-f^{(-1)}(d_{,\theta}+2d\cot\theta)\cot\theta=0, \tag{107}$$

$$\mathscr{L}_\eta g_{12}=0 \quad (r^{-1}): \quad 2f^{(1)}-2B(c_{,\theta}+2c\cot\theta)-f^{(-1)}(c^2+2d^2)+2dg^{(0)}=0, \tag{108}$$

$$\mathscr{L}_\eta g_{13}=0 \quad (r^{-1}): \quad g^{(1)}-B(d_{,\theta}+2d\cot\theta)-d(f^{(0)}+f^{(-1)}c)=0, \tag{109}$$

$$\mathscr{L}_\eta g_{22}=0 \quad (r^{0}): \quad 2d(g^{(0)}{}_{,\theta}-g^{(0)}\cot\theta)+A^{(-1)}\frac{3d^2}{2}+4f^{(-1)}d^2\cot\theta+A^{(1)}$$
$$+BM+f^{(1)}{}_{,\theta}+B(2dd_{,u}+cc_{,u})-B(c_{,\theta}+2c\cot\theta)_{,\theta}=0, \tag{110}$$

$$\mathscr{L}_\eta g_{23}=0 \quad (r^{0}): \quad -2A^{(0)}d+Bd+4f^{(-1)}d(c_{,\theta}+c\cot\theta)$$
$$-4f^{(0)}d\cot\theta+g^{(0)}{}_{,\theta}c-g^{(0)}(c_{,\theta}+c\cot\theta)-g^{(1)}{}_{,\theta}+g^{(1)}\cot\theta$$
$$+B(d_{,\theta}+2d\cot\theta)_{,\theta}-B(d_{,\theta}+2d\cot\theta)\cot\theta=0, \tag{111}$$

$$\mathscr{L}_\eta g_{33}=0 \quad (r^{0}): \quad A^{(1)}-A^{(0)}c+\tfrac{1}{2}Bc+\tfrac{1}{2}A^{(-1)}(c^2+d^2)+BM-f^{(-1)}d^2\cot\theta$$
$$+2f^{(-1)}d(d_{,\theta}+2d\cot\theta)-B\cot\theta(c_{,\theta}+2c\cot\theta)+2B(cc_{,u}+dd_{,u})$$
$$+f^{(1)}\cot\theta+(c_{,\theta}+2c\cot\theta)(-f^{(0)}+3f^{(-1)}c)$$
$$-c\cot\theta(f^{(0)}+\tfrac{3}{2}f^{(-1)}c)=0. \tag{112}$$

Substituting for c, d and $g^{(0)}$ from Eqs. (98), (100) and (102) into Eqs. (108) and (109) we obtain

$$f^{(1)}=\frac{1}{4\sin\theta}(-2\mathscr{K}^2+g^{(0)2})+u\cot^2\theta(\mathscr{K}_{,w}w+\mathscr{K}), \tag{113}$$

and

$$g^{(1)}=\frac{u}{2}g^{(0)}-\frac{g^{(0)}\mathscr{K}}{\sin\theta}+\frac{u}{2}\cot^2\theta(g^{(0)}{}_{,w}w+g^{(0)}). \tag{114}$$

Using the results for $f^{(-1)}$, $f^{(0)}$, B, ... already obtained, we find that Eqs. (106), (107) and (111) are just identities; Eqs. (110) and (112) also become identities if Eq. (105) is used.

Comparing the remaining Eq. (104) with the u-derivative of Eq. (105), we finally arrive at the equation for the mass aspect M:

$$uM_{,u}+\tan\theta M_{,\theta}+3M-u\tan\theta(c_{,\theta}+2c\cot\theta)_{,u}=0, \tag{115}$$

where $c_{,u}$ is given by Eq. (99). This equation can be solved by the substitution (95) to yield

$$M(u,\theta)=\frac{1}{2u}(\mathcal{K}_{,ww}w+2\mathcal{K}_{,w})+\frac{\mathcal{L}(w)}{u^3}, \tag{116}$$

where $\mathcal{L}(w)$ is an arbitrary function of w. Substituting M into Eq. (105) we obtain function $A^{(1)}$ in the form

$$A^{(1)}=\cos\theta\left(\frac{1}{2}\mathcal{K}_{,ww}w+2\mathcal{K}_{,w}+\frac{\mathcal{K}}{w}\right)-\frac{\cos\theta}{8\sin^2\theta}(4\mathcal{K}^2+g^{(0)2})+\frac{\cos\theta}{u^2}\mathcal{L}(w). \tag{117}$$

The last expression for M can be written as

$$M(u,\theta)=\frac{1}{2\sin\theta}(w^2\mathcal{K}_{,w})_{,w}+\frac{\mathcal{L}}{u^3}. \tag{118}$$

Comparing this form of $M(u,\theta)$ with the consequence of Einstein's equations (6), we find

$$\mathcal{L}(w)=\frac{\lambda(w)}{w^3}, \tag{119}$$

where λ has to satisfy the equation

$$\lambda(w)_{,w}=w^2\left(\mathcal{K}_{,w}^{\,2}+\frac{1}{4}g^{(0)}{}_{,w}^{\,2}+\mathcal{E}^2+\tilde{\mathcal{B}}^2\right)-\frac{1}{2w}(w^3\mathcal{K}_{,ww})_{,w}. \tag{120}$$

Here, $\mathcal{K}$ and $g^{(0)}$ determine the gravitational news functions, $c_{,u}$ and $d_{,u}$, by relations (99) and (103), $\mathcal{E}$ and $\tilde{\mathcal{B}}$ determine the electromagnetic news functions, X and Y, given below by relations (129) and (130). Hence, solving the last equation for λ for *given* $\mathcal{K}$, $g^{(0)}$, $\mathcal{E}$ and $\tilde{\mathcal{B}}$, we find $\mathcal{L}(w)$ and thus the mass aspect $M(u,\theta)$ in the form of Eq. (118). The total mass at $\mathcal{I}^+$ is then given by integrating Eq. (118) over the sphere:

$$m(u)=\frac{1}{2}\int_0^\pi M(u,\theta)\sin\theta\,d\theta=\frac{1}{4}\int_0^\pi(w^2\mathcal{K}_{,w})_{,w}\,d\theta+\frac{1}{2}\int_0^\pi\frac{w\mathcal{L}}{u^2}d\theta. \tag{121}$$

Substituting the expansions of the metric functions, Eq. (4), into the null tetrad, Eq. (11), and coefficients A, B, $\tilde{f}$ and $\tilde{g}$, Eqs. (93), (97), (98), (100), (102), (113), (114), (116), (117), into Eq. (13), we find the expansion of the boost Killing vector to be

$$\begin{aligned}\eta^\mu=\Bigg[&-u\cos\theta,\quad r\cos\theta+u\cos\theta+\cos\theta\left(\mathcal{K}_{,w}+\frac{\mathcal{K}}{w}\right)\frac{1}{r}+O(r^{-2}),\\ &-\sin\theta-u\sin\theta\frac{1}{r}+uc\sin\theta\frac{1}{r^2}+O(r^{-3}),\quad ud\frac{1}{r^2}+O(r^{-3})\Bigg].\end{aligned} \tag{122}$$

It can easily be seen that the boost Killing vector generating Lorentz transformations along the z-axis in Minkowski space, $\eta^\alpha_M=[z,0,t,0]$, in $\{u,\ r,\ \theta,\ \phi\}$ coordinates reads $\eta^\alpha_M=[-u\cos\theta,\ r(1+u/r)\cos\theta,\ -\sin\theta(1+u/r),\ 0]$ so that both vectors differ only in terms proportional to $\mathcal{K}$, c and d.

Let us finally turn to the asymptotic properties of the electromagnetic field which now is assumed to satisfy $\mathcal{L}_\eta F_{\alpha\beta}=0$, Eq. (59), where η is the boost Killing field (122). The first orders of these equations read

$$\mathscr{L}_\eta F_{01}=0 \quad (r^{-2}): \quad \cos\theta \ (\tan\theta \ \ \epsilon_{,\theta}+u\epsilon_{,u}+2\epsilon+u\tan\theta \ \ X)=0, \tag{123}$$

$$\mathscr{L}_\eta F_{02}=0 \quad (r^{0}): \quad -\cos\theta \ (\tan\theta \ \ X_{,\theta}+uX_{,u}+2X)=0, \tag{124}$$

$$\mathscr{L}_\eta F_{03}=0 \quad (r^{0}): \quad -\cos\theta \ [\tan\theta \ \ (Y\sin\theta)_{,\theta}+u(Y\sin\theta)_{,u}+Y\sin\theta]=0, \tag{125}$$

$$\mathscr{L}_\eta F_{12}=0 \quad (r^{-2}): \quad -\cos\theta(\tan\theta \ \ e_{,\theta}+ue_{,u}+2e-u\tan\theta \ \ \epsilon)=0, \tag{126}$$

$$\mathscr{L}_\eta F_{13}=0 \quad (r^{-2}): \quad -\cos\theta \ [\tan\theta \ \ (f\sin\theta)_{,\theta}+u(f\sin\theta)_{,u}+f\sin\theta+u\tan\theta \ \ (\mu\sin\theta)]=0, \tag{127}$$

$$\mathscr{L}_\eta F_{23}=0 \quad (r^{0}): \quad \cos\theta \ [\tan\theta \ \ (\mu\sin\theta)_{,\theta}+u(\mu\sin\theta)_{,u}+\mu\sin\theta+u\tan\theta \ \ (Y\sin\theta)]=0. \tag{128}$$

Using again variable w given by Eq. (95), Eqs. (124) and (125) can easily be solved to yield

$$X(u,\theta)=\frac{\mathscr{E}(w)}{u^2}, \tag{129}$$

$$Y(u,\theta)=\frac{\mathscr{B}(w)}{u\sin\theta}=\frac{\mathscr{B}(w)}{wu^2}=\frac{\tilde{\mathscr{B}}(w)}{u^2}, \tag{130}$$

where functions $\mathscr{E}(w)$ and $\mathscr{B}(w)$ are arbitrary integration functions and $\tilde{\mathscr{B}}(w)=\mathscr{B}(w)/w$. Substituting these results into Eqs. (123) and (128) we find

$$\epsilon=\frac{\mathscr{E}(w)}{u}\cot\theta+\frac{\mathscr{F}(w)}{u^2}, \tag{131}$$

$$\mu=\frac{\mathscr{B}(w)}{\sin\theta}\cot\theta+\frac{\mathscr{C}(w)}{u\sin\theta}=\frac{\tilde{\mathscr{B}}(w)}{u}\cot\theta+\frac{\tilde{\mathscr{C}}(w)}{u^2}, \tag{132}$$

where functions $\mathscr{F}(w)$ and $\mathscr{C}(w)$ are arbitrary integration functions and $\tilde{\mathscr{C}}(w)=\mathscr{C}(w)/w$. Then the solution of Eqs. (126) and (127) reads

$$e=\frac{\mathscr{E}(w)}{2}-\frac{\mathscr{F}(w)}{u}\cot\theta+\frac{\mathscr{G}(w)}{u^2}, \tag{133}$$

$$f=-\frac{\mathscr{B}(w)u}{2\sin\theta}+\frac{\mathscr{C}(w)}{\sin\theta}\cot\theta+\frac{\mathscr{D}(w)}{u\sin\theta}=-\frac{\tilde{\mathscr{B}}(w)}{2}+\frac{\tilde{\mathscr{C}}(w)}{u}\cot\theta+\frac{\tilde{\mathscr{D}}(w)}{u^2}, \tag{134}$$

where $\mathscr{G}(w)$ and $\mathscr{D}(w)$ are again integration functions and $\tilde{\mathscr{D}}(w)=\mathscr{D}(w)/w$. Comparing Eqs. (123) and (128) with Maxwell equations (9) and (10), and using previous results, we get equations for $\mathscr{F}(w)$ and $\mathscr{C}(w)$: $\mathscr{F}_{,w}w+2\mathscr{F}=0$ and $\mathscr{C}_{,w}w+\mathscr{C}=0$. Their solutions are

$$\mathscr{F}=\frac{l}{w^2}, \quad l=\text{const}, \tag{135}$$

$$\mathscr{C}=\frac{p}{w},\quad p=\text{const.} \tag{136}$$

Since $w=\sin\theta/u$, the regularity of the electromagnetic field implies $l=p=0$.

Finally, comparing Maxwell equations (7) and (8) with Eqs. (126) and (127), we get restrictions on functions $\mathscr{D}(w)$ and $\mathscr{G}(w)$ in the form

$$\mathscr{E}_{,w}w-\mathscr{E}+2w^2(\mathscr{G}_{,w}w+2\mathscr{G})+2w\mathscr{E}\mathscr{K}+\mathscr{B}g^{(0)}=0, \tag{137}$$

$$-(\mathscr{B}_{,w}w-2.\mathscr{B})+2w^2(\mathscr{D}_{,w}w+\mathscr{D})-2w.\mathscr{B}\mathscr{K}+w^2g^{(0)}\mathscr{E}=0. \tag{138}$$

ACKNOWLEDGMENTS

We thank Vojtěch Pravda for discussions and help with calculations. We are grateful to Piotr Chruściel for reading the manuscript and making helpful suggestions. We also acknowledge support from Grant Nos. GACR-202/96/0206 of the Czech Republic and GAUK-230/96 of the Charles University, and the hospitality of the Albert Einstein Institute, Potsdam.

APPENDIX A: THE EINSTEIN–MAXWELL EQUATIONS IN THE BONDI–SACHS COORDINATES

In our convention (see the end of Sec. I) the Einstein–Maxwell equations read

$$K_{\mu\nu}\equiv R_{\mu\nu}+8\pi T_{\mu\nu}=0, \tag{A1}$$

where the electromagnetic stress tensor is given by

$$T_{\mu\nu}=\frac{1}{4\pi}\left(F_{\mu\sigma}F_{\nu}{}^{\sigma}-\frac{1}{4}g_{\mu\nu}F_{\sigma\rho}F^{\sigma\rho}\right), \tag{A2}$$

and $T^{\mu}_{\mu}=R=0$; the Maxwell equations are

$$G_{\mu\nu\lambda}\equiv F_{[\mu\nu,\lambda]cykl.}=0, \tag{A3}$$

$$J^{\mu}\equiv F^{\mu\nu}{}_{;\nu}=0. \tag{A4}$$

In Ref. 11 the tensor $E_{\mu\nu}=4\pi T_{\mu\nu}$ is introduced; however, the field equations (A1) are erroneously written without the factor 2 at $E_{\mu\nu}$. This can be ''cured'' by considering $F_{\mu\nu(\text{Burg})}=\sqrt{2}F_{\mu\nu(\text{real})}$. Following Ref. 11, eighteen equations (A1), (A3) and (A4) can be divided into twelve main equations, one trivial equation, and five supplementary conditions. The main equations are

$$\begin{gathered}K_{11}=K_{12}=K_{13}=K_{22}=K_{23}=K_{33}=0,\\ G_{123}=G_{012}=G_{013}=0,\quad J^0=J^2=J^3=0.\end{gathered} \tag{A5}$$

If these are satisfied, then also

$$K_{01}=0. \tag{A6}$$

The only further equations to be satisfied are the supplementary conditions

$$(r^2K_{00})_{,r}=(r^2K_{02})_{,r}=(r^2K_{03})_{,r}=0, \tag{A7}$$

$$G_{023},_r=0, \quad J^1,_r=0,$$

which imply that $r^2(K_{00},K_{02},K_{03})$, G_{023}, and J^1 are functions of u, θ, ϕ only.

Since in Ref. 11 all Einstein equations contain errors due to the factor 2 mentioned above, and Eqs. (5), (6), (7), (14), (15), (17), (18), (19) in Ref. 11 contain additional errors, we write all field equations explicitly here (we checked them by using MAPLE V). Starting from the metric of the form (3) we find that the main field equations are (denoting $\mathrm{ch}=\cosh 2\delta$, $\mathrm{sh}=\sinh 2\delta$)

$$\frac{r}{4}K_{11}=0: \quad \beta,_r=\frac{r}{2}(\gamma,_r^2 \mathrm{ch}^2+\delta,_r^2)$$
$$+\frac{1}{2r}(e^{-2\gamma}\mathrm{ch}\,F_{12}^2+e^{2\gamma}\mathrm{ch}\,F_{13}^2\csc^2\theta-2\,\mathrm{sh}\,F_{12}F_{13}\csc\theta), \tag{A8}$$

$$2r^2K_{12}=0: \quad \{r^4e^{-2\beta}(e^{2\gamma}U,_r\,\mathrm{ch}+W,_r\,\mathrm{sh})\},_r$$
$$=2r^2\left\{\beta,_{r\theta}-\frac{2}{r}\beta,_\theta+2\delta,_r\delta,_\theta-4\gamma,_r\delta,_\theta\,\mathrm{sh}\,\mathrm{ch}-(\gamma,_{r\theta}+2\gamma,_r\cot\theta-2\gamma,_r\gamma,_\theta)\mathrm{ch}^2\right\}$$
$$+2r^2e^{2\gamma}\csc\theta\{-\delta,_{r\phi}-2\delta,_r\gamma,_\phi+2\gamma,_r\delta,_\phi(1+2\,\mathrm{sh}^2)+(\gamma,_{r\phi}+2\gamma,_r\gamma,_\phi)\,\mathrm{sh}\,\mathrm{ch}\}$$
$$-4r^2e^{-2\beta}(F_{01}-UF_{12}-WF_{13}\csc\theta)F_{12}-4(\mathrm{sh}\,F_{12}-e^{2\gamma}\mathrm{ch}\,F_{13}\csc\theta)F_{23}\csc\theta, \tag{A9}$$

$$\frac{2r^2}{\sin\theta}K_{13}=0: \quad \{r^4e^{-2\beta}(e^{-2\gamma}W,_r\,\mathrm{ch}+U,_r\,\mathrm{sh})\},_r$$
$$=2r^2e^{-2\gamma}\{-\delta,_{r\theta}+2\delta,_r\gamma,_\theta-2\delta,_r\cot\theta-(\gamma,_{r\theta}+2\gamma,_r\cot\theta-2\gamma,_r\gamma,_\theta)\mathrm{sh}\,\mathrm{ch}$$
$$-2\gamma,_r\delta,_\theta(1+2\,\mathrm{sh}^2)\}+2r^2\csc\theta\left\{\beta,_{r\phi}-\frac{2}{r}\beta,_\phi+2\delta,_r\delta,_\phi+4\gamma,_r\delta,_\phi\,\mathrm{sh}\,\mathrm{ch}\right.$$
$$\left.+(\gamma,_{r\phi}+2\gamma,_r\gamma,_\phi)\mathrm{ch}^2\right\}-4r^2e^{-2\beta}(F_{01}-UF_{12}-WF_{13}\csc\theta)F_{13}\csc\theta$$
$$-4(e^{-2\gamma}\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)F_{23}\csc\theta, \tag{A10}$$

$$\tfrac{1}{2}e^{2\beta}\{(e^{-2\gamma}K_{22}+e^{2\gamma}\csc^2\theta K_{33})\mathrm{ch}-2\csc\theta K_{23}\,\mathrm{sh}\}=0:$$

$$V,_r=2e^{2\beta}\csc\theta\{(\beta,_{\theta\phi}+\beta,_\theta\beta,_\phi+2\delta,_\theta\delta,_\phi)\mathrm{sh}$$
$$+(\delta,_{\theta\phi}+\delta,_\phi\cot\theta+\delta,_\theta\gamma,_\phi-\gamma,_\theta\delta,_\phi+\delta,_\theta\beta,_\phi+\beta,_\theta\delta,_\phi)\mathrm{ch}\}$$
$$-e^{2(\beta-\gamma)}\{(\beta,_{\theta\theta}+\beta,_\theta^2+\beta,_\theta\cot\theta+2\gamma,_\theta^2+2\delta,_\theta^2-1-\gamma,_{\theta\theta}-3\gamma,_\theta\cot\theta-2\beta,_\theta\gamma,_\theta)\mathrm{ch}$$
$$+(\delta,_{\theta\theta}+3\delta,_\theta\cot\theta+2\beta,_\theta\delta,_\theta-4\gamma,_\theta\delta,_\theta)\mathrm{sh}\}-e^{2(\beta+\gamma)}\{(\beta,_{\phi\phi}+\beta,_\phi^2+2\gamma,_\phi^2+2\delta,_\phi^2+\gamma,_{\phi\phi}$$
$$+2\beta,_\phi\gamma,_\phi)\mathrm{ch}+(\delta,_{\phi\phi}+2\beta,_\phi\delta,_\phi+4\gamma,_\phi\delta,_\phi)\mathrm{sh}\}\csc^2\theta-\frac{r^4}{4}e^{-2\beta}\{(e^{2\gamma}U,_r^2+e^{-2\gamma}W,_r^2)\mathrm{ch}$$
$$+2U,_rW,_r\,\mathrm{sh}\}+\frac{r}{2}(rU,_{r\theta}+rU,_r\cot\theta+4U,_\theta+4U\cot\theta)+\frac{r}{2}\csc\theta(rW,_{r\phi}+4W,_\phi)$$
$$-\frac{1}{r^2}e^{2\beta}\{[r^2e^{-2\beta}(F_{01}-UF_{12}-WF_{13}\csc\theta)]^2+(F_{23}\csc\theta)^2\}, \tag{A11}$$

$$\frac{1}{4r}e^{2\beta}(e^{-2\gamma}K_{22}-e^{2\gamma}K_{33}\csc^2\theta)=0:$$

$$(r\gamma)_{,ur}\,\mathrm{ch}+2r(\gamma_{,u}\delta_{,r}+\delta_{,u}\gamma_{,r})\mathrm{sh}$$
$$=\frac{1}{2}\left(\gamma_{,r}V_{,r}+\gamma_{,rr}V+\frac{1}{r}\gamma_{,r}V\right)\mathrm{ch}+2\gamma_{,r}\delta_{,r}V\,\mathrm{sh}+\frac{r^3}{8}e^{-2\beta}(e^{2\gamma}U_{,r}^2-e^{-2\gamma}W_{,r}^2)$$

$$+\frac{1}{2r}e^{2(\beta-\gamma)}(\beta_{,\theta\theta}+\beta_{,\theta}^2-\beta_{,\theta}\cot\theta)-\frac{1}{2r}e^{2(\beta+\gamma)}(\beta_{,\phi\phi}+\beta_{,\phi}^2)\csc^2\theta$$
$$+\frac{1}{r}e^{2\beta}(\beta_{,\theta}\delta_{,\phi}-\beta_{,\phi}\delta_{,\theta})\csc\theta+\frac{r}{4}e^{2\gamma}\csc\theta\left\{\left(U_{,r\phi}+\frac{2}{r}U_{,\phi}\right)\mathrm{sh}+4\delta_{,r}U_{,\phi}\,\mathrm{ch}\right\}$$
$$-\frac{r}{4}e^{-2\gamma}\left\{\left(W_{,r\theta}-W_{,r}\cot\theta+\frac{2}{r}W_{,\theta}-\frac{2}{r}W\cot\theta\right)\mathrm{sh}+4\delta_{,r}(W_{,\theta}-W\cot\theta)\mathrm{ch}\right\}$$
$$-\tfrac{1}{4}(rU_{,r\theta}+2U_{,\theta}-rU_{,r}\cot\theta-2U\cot\theta+4\gamma_{,\theta}U+4r\gamma_{,r\theta}U+2r\gamma_{,\theta}U_{,r}$$
$$+2r\gamma_{,r}U_{,\theta}+2r\gamma_{,r}U\cot\theta)\mathrm{ch}-r(\delta_{,r}U_{,\theta}+2\gamma_{,r}\delta_{,\theta}U+2\delta_{,r}\gamma_{,\theta}U-\delta_{,r}U\cot\theta)\mathrm{sh}$$
$$+\tfrac{1}{4}\csc\theta(rW_{,r\phi}+2W_{,\phi}-4\gamma_{,\phi}W-4r\gamma_{,r\phi}W-2r\gamma_{,\phi}W_{,r}-2r\gamma_{,r}W_{,\phi})\mathrm{ch}$$
$$+r\csc\theta(\delta_{,r}W_{,\phi}-2\delta_{,r}\gamma_{,\phi}W-2\gamma_{,r}\delta_{,\phi}W)\mathrm{sh}-\frac{1}{2r}\left\{2(e^{-2\gamma}F_{02}F_{12}-e^{2\gamma}F_{03}F_{13}\csc^2\theta)\right.$$
$$\left.-\frac{1}{r}V(e^{-2\gamma}F_{12}^2-e^{2\gamma}F_{13}^2\csc^2\theta)-2(e^{-2\gamma}WF_{12}+e^{2\gamma}UF_{13}\csc\theta)F_{23}\csc\theta\right\},\qquad\text{(A12)}$$

$$\frac{1}{4r}e^{2\beta}\{(e^{-2\gamma}K_{22}+e^{2\gamma}\csc\theta K_{33})\mathrm{sh}-2\csc\theta K_{23}\,\mathrm{ch}\}=0:$$

$$(r\delta)_{,ur}-2r\gamma_{,u}\gamma_{,r}\,\mathrm{sh}\,\mathrm{ch}$$
$$=\frac{1}{2}\left(\delta_{,r}V_{,r}+\delta_{,rr}V+\frac{1}{r}\delta_{,r}V-2\gamma_{,r}^2V\,\mathrm{sh}\,\mathrm{ch}\right)+\frac{r^3}{8}e^{-2\beta}\{(e^{2\gamma}U_{,r}^2+e^{-2\gamma}W_{,r}^2)\mathrm{sh}+2U_{,r}W_{,r}\,\mathrm{ch}\}$$

$$-\frac{1}{2r}e^{2(\beta-\gamma)}(\beta_{,\theta\theta}+\beta_{,\theta}^2-\beta_{,\theta}\cot\theta)\mathrm{sh}-\frac{1}{2r}e^{2(\beta+\gamma)}(\beta_{,\phi\phi}+\beta_{,\phi}^2)\mathrm{sh}\csc^2\theta$$
$$-\frac{1}{r}e^{2\beta}(-\beta_{,\theta\phi}-\beta_{,\theta}\beta_{,\phi}+\beta_{,\phi}\cot\theta+\beta_{,\theta}\gamma_{,\phi}-\beta_{,\phi}\gamma_{,\theta})\mathrm{ch}\csc\theta$$
$$-\frac{r}{2}\left\{2\delta_{,r\theta}U+\frac{2}{r}\delta_{,\theta}U+\delta_{,r}U_{,\theta}+\delta_{,\theta}U_{,r}+\delta_{,r}U\cot\theta-2\gamma_{,r}(U_{,\theta}-U\cot\theta+2\gamma_{,\theta}U)\mathrm{sh}\,\mathrm{ch}\right\}$$
$$-\frac{r}{2}\csc\theta\left\{2\delta_{,r\phi}W+\frac{2}{r}\delta_{,\phi}W+\delta_{,r}W_{,\phi}+\delta_{,\phi}W_{,r}+2\gamma_{,r}(W_{,\phi}-2\gamma_{,\phi}W)\mathrm{sh}\,\mathrm{ch}\right\}$$
$$-\frac{r}{4}e^{-2\gamma}\left\{W_{,r\theta}-W_{,r}\cot\theta+\frac{2}{r}W_{,\theta}-\frac{2}{r}W\cot\theta-4\gamma_{,r}(W_{,\theta}-W\cot\theta)\mathrm{ch}^2\right\}$$
$$-\frac{r}{4}e^{2\gamma}\csc\theta\left(U_{,r\phi}+\frac{2}{r}U_{,\phi}+4\gamma_{,r}U_{,\phi}\,\mathrm{ch}^2\right)-\frac{1}{2r}\left\{2(\mathrm{ch}\,F_{02}-e^{2\gamma}\,\mathrm{sh}\,F_{03}\csc\theta)F_{13}\csc\theta\right.$$

$$-2(e^{-2\gamma}\,\mathrm{sh}\,F_{02}-\mathrm{ch}\,F_{03}\csc\theta)F_{12}+\frac{1}{r}V(e^{-2\gamma}\,\mathrm{sh}\,F_{12}^2+e^{2\gamma}\,\mathrm{sh}\,F_{13}^2\csc^2\theta$$
$$-2\,\mathrm{ch}\,F_{12}F_{13}\csc\theta)+2[U(\mathrm{ch}\,F_{12}-e^{2\gamma}\,\mathrm{sh}\,F_{13}\csc\theta)$$
$$+W(e^{-2\gamma}\,\mathrm{sh}\,F_{12}-\mathrm{ch}\,F_{13}\csc\theta)]F_{23}\csc\theta\Big\}, \tag{A13}$$

$$G_{123}=0:\quad F_{23,r}=F_{13,\theta}-F_{12,\phi}, \tag{A14}$$

$$G_{012}=0:\quad F_{01,\theta}=F_{02,r}-F_{12,u}, \tag{A15}$$

$$G_{013}=0:\quad F_{01,\phi}=F_{03,r}-F_{13,u}, \tag{A16}$$

$$r^2e^{2\beta}J^0=0:\quad \{r^2e^{-2\beta}(F_{01}-UF_{12}-WF_{13}\csc\theta)\}_{,r}$$
$$=-\csc\theta\{\sin\theta(e^{-2\gamma}\,\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)\}_{,\theta}+\csc\theta(\mathrm{sh}\,F_{12}-e^{2\gamma}\,\mathrm{ch}\,F_{13}\csc\theta)_{,\phi}, \tag{A17}$$

$$r^2e^{2\beta}J^2=0:\quad (e^{-2\gamma}\,\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)_{,u}+(e^{-2\gamma}\,\mathrm{ch}\,F_{02}-\mathrm{sh}\,F_{03}\csc\theta)_{,r}$$
$$=\Big\{r^2e^{-2\beta}U(F_{01}-UF_{12}-WF_{13}\csc\theta)+\frac{V}{r}(e^{-2\gamma}\,\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)$$
$$+(e^{-2\gamma}\,\mathrm{ch}\,W+\mathrm{sh}\,U)F_{23}\csc\theta\Big\}_{,r}$$
$$-\csc\theta\Big\{W(e^{-2\gamma}\,\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)+U(\mathrm{sh}\,F_{12}-e^{2\gamma}\,\mathrm{ch}\,F_{13}\csc\theta)$$
$$+\frac{1}{r^2}e^{2\beta}F_{23}\csc\theta\Big\}_{,\phi}, \tag{A18}$$

$$r^2\sin\theta\,e^{2\beta}J^3=0:$$

$$(\mathrm{sh}\,F_{12}-e^{2\gamma}\,\mathrm{ch}\,F_{13}\csc\theta)_{,u}+(\mathrm{sh}\,F_{02}-e^{2\gamma}\,\mathrm{ch}\,F_{03}\csc\theta)_{,r}$$
$$=-\Big\{W(e^{-2\gamma}\,\mathrm{ch}\,F_{12}-\mathrm{sh}\,F_{13}\csc\theta)+U(\mathrm{sh}\,F_{12}-e^{2\gamma}\,\mathrm{ch}\,F_{13}\csc\theta)+\frac{1}{r^2}e^{2\beta}F_{23}\csc\theta\Big\}_{,\theta}.$$
$$-\Big\{r^2e^{-2\beta}W(F_{01}-UF_{12}-WF_{13}\csc\theta)-\frac{V}{r}(\mathrm{sh}\,F_{12}-e^{2\gamma}\,\mathrm{ch}\,F_{13}\csc\theta)$$
$$-(e^{2\gamma}\,\mathrm{ch}\,U+\mathrm{sh}\,W)F_{23}\csc\theta\Big\}_{,r} \tag{A19}$$

Since we assume the axial symmetry, hereafter we put all derivatives $\partial/\partial\phi$ equal to zero.

Now following Ref. 11, we assume functions γ, δ, F_{12} and F_{13} to have the expansions at large r on a hypersurface $u=u_0$ of the form

$$\gamma=\frac{c}{r}+\left(C-\frac{1}{6}c^3-\frac{3}{2}cd^2\right)\frac{1}{r^3}+\frac{D}{r^4}+O(r^{-5}), \tag{A20}$$

$$\delta=\frac{d}{r}+\left(H-\frac{1}{6}d^3+\frac{1}{2}c^2d\right)\frac{1}{r^3}+\frac{K}{r^4}+O(r^{-5}), \tag{A21}$$

$$F_{12}=\frac{e}{r^2}+(2E+ec+fd)\frac{1}{r^3}+O(r^{-4}), \tag{A22}$$

$$F_{13}=\left(\frac{f}{r^2}+(2F+ed-fc)\frac{1}{r^3}+O(r^{-4})\right)\sin\theta, \tag{A23}$$

where c, C, d, ... are prescribed functions of θ on $u=u_0$. This corresponds to the outgoing radiation condition and the absence of logarithmic terms. (We thus do not consider polyhomogeneous null infinity of Ref. 19.) Then twelve main equations determine the other eight functions β, U, V, W, F_{01}, F_{02}, F_{03} and u-derivatives of the four prescribed functions γ, δ, F_{12} and F_{13} on the hypersurface $u=u_0$. Denoting by N, P, M, ϵ, μ, X and Y arbitrary functions of θ on $u=u_0$, we find

$$\beta=-\frac{1}{4}(c^2+d^2)\frac{1}{r^2}+O(r^{-4}), \tag{A24}$$

$$\begin{aligned}U=&-(c_{,\theta}+2c\cot\theta)\frac{1}{r^2}+(2N+3(cc_{,\theta}+dd_{,\theta})+4(c^2+d^2)\cot\theta)\frac{1}{r^3}\\&+\frac{1}{2}(3(C_{,\theta}+2C\cot\theta)-6(cN+dP)-4(2c^2c_{,\theta}+cdd_{,\theta}+c_{,\theta}d^2))\frac{1}{r^4}\\&+\frac{1}{2}(-8c(c^2+d^2)\cot\theta+2(\epsilon e-f\mu))\frac{1}{r^4}+O(r^{-5}),\end{aligned} \tag{A25}$$

$$\begin{aligned}W=&-(d_{,\theta}+2d\cot\theta)\frac{1}{r^2}+(2P+2(c_{,\theta}d-cd_{,\theta}))\frac{1}{r^3}\\&+\frac{1}{2}(3(H_{,\theta}+2H\cot\theta)+(cP-dN)-4(2d^2d_{,\theta}+cdc_{,\theta}+c^2d_{,\theta}))\frac{1}{r^4}\\&+\frac{1}{2}(-8d(c^2+d^2)\cot\theta+2(\mu e+\epsilon f))\frac{1}{r^4}+O(r^{-5}),\end{aligned} \tag{A26}$$

$$\begin{aligned}V=&r-2M-(N_{,\theta}+N\cot\theta-\tfrac{1}{2}(c^2+d^2))\frac{1}{r}\\&-(-(c_{,\theta}+2c\cot\theta)^2-(d_{,\theta}+2d\cot\theta)^2-(\epsilon^2+\mu^2))\frac{1}{r}\\&-\frac{1}{2}(C_{,\theta\theta}+3C_{,\theta}\cot\theta-2C+6N(c_{,\theta}+2c\cot\theta)+6P(d_{,\theta}+2d\cot\theta))\frac{1}{r^2}\\&-\frac{1}{2}(4(2cc_{,\theta}^2+3c_{,\theta}dd_{,\theta}-cd_{,\theta}^2)+8(2c_{,\theta}d^2+3c^2c_{,\theta}+cdd_{,\theta})\cot\theta)\frac{1}{r^2}\\&-\frac{1}{2}(16c(c^2+d^2)\cot^2\theta+2\epsilon(e_{,\theta}+e\cot\theta)-2\mu(f_{,\theta}+f\cot\theta))\frac{1}{r^2}+O(r^{-3}),\end{aligned} \tag{A27}$$

$$F_{01}=-\frac{\epsilon}{r^2}+(e_{,\theta}+e\cot\theta)\frac{1}{r^3}+O(r^{-4}), \tag{A28}$$

$$\mathrm{F}_{23}=\left(-\mu-(f,_\theta+f\cot\theta)\frac{1}{r}+O(r^{-2})\right)\sin\theta, \tag{A29}$$

$$F_{02}=X+(\epsilon,_\theta-e,_u)\frac{1}{r}-\left(\left[E+\frac{1}{2}(ec+fd)\right],_u+\frac{1}{2}(e,_\theta+e\cot\theta),_\theta\right)\frac{1}{r^2}+O(r^{-3}), \tag{A30}$$

$$F_{03}=\left(Y-\frac{f,_u}{r}-\left(\left[F+\frac{1}{2}(ed-fc)\right],_u\right)\frac{1}{r^2}+O(r^{-3})\right)\sin\theta, \tag{A31}$$

and

$$2e,_u=\epsilon,_\theta-2(cX+dY), \tag{A32}$$

$$2f,_u=-\mu,_\theta-2(-cY+dX), \tag{A33}$$

$$4E,_u=-\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)\left[\left(\frac{\partial}{\partial\theta}-\cot\theta\right)e+2(c\epsilon+d\mu)\right], \tag{A34}$$

$$4F,_u=-\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)\left[\left(\frac{\partial}{\partial\theta}-\cot\theta\right)f+2(d\epsilon-c\mu)\right], \tag{A35}$$

$$\begin{aligned}4C,_u=&\,2(c^2-d^2)c,_u+4dcd,_u+2cM+d\left(\frac{\partial}{\partial\theta}+\cot\theta\right)(d,_\theta+2d\cot\theta)\\&-\left(\frac{\partial}{\partial\theta}-\cot\theta\right)N+2(eX-fY),\end{aligned} \tag{A36}$$

$$\begin{aligned}4H,_u=&-2(c^2-d^2)d,_u+4dcc,_u+2dM-c\left(\frac{\partial}{\partial\theta}+\cot\theta\right)(d,_\theta+2d\cot\theta)\\&-\left(\frac{\partial}{\partial\theta}-\cot\theta\right)P+2(eY+fX),\end{aligned} \tag{A37}$$

$$\begin{aligned}4D,_u=&\left(\frac{\partial}{\partial\theta}+3\cot\theta\right)\left[-\left(\frac{\partial}{\partial\theta}-2\cot\theta\right)C+2(cN-dP)\right]-\frac{2}{3}\epsilon\left(\frac{\partial}{\partial\theta}-\cot\theta\right)e\\&+\frac{2}{3}\mu\left(\frac{\partial}{\partial\theta}-\cot\theta\right)f-\frac{4}{3}c(\epsilon^2+\mu^2)+\frac{8}{3}(EX-YF),\end{aligned} \tag{A38}$$

$$\begin{aligned}4K,_u=&\left(\frac{\partial}{\partial\theta}+3\cot\theta\right)\left[-\left(\frac{\partial}{\partial\theta}-2\cot\theta\right)H+2(cP+dN)\right]-\frac{2}{3}\epsilon\left(\frac{\partial}{\partial\theta}-\cot\theta\right)f\\&-\frac{2}{3}\mu\left(\frac{\partial}{\partial\theta}-\cot\theta\right)e-\frac{4}{3}d(\epsilon^2+\mu^2)+\frac{8}{3}(FX+YE).\end{aligned} \tag{A39}$$

In order to simplify the last equations one can, following Ref. 11, introduce ten new quantities

$$\begin{aligned}&c^*=c+id, \quad X^*=X+iY,\\&C^*=C+iH, \quad e^*=e+if,\\&D^*=D+iK, \quad E^*=E+iF,\\&N^*=N+iP, \quad \epsilon^*=\epsilon+i\mu,\\&M^*=M+i\frac{1}{2}\left(\frac{\partial}{\partial\theta}+\cot\theta\right)(d,_\theta+2d\cot\theta),\end{aligned} \tag{A40}$$

$$\mathscr{L}_p=-\left(\frac{\partial}{\partial\theta}-p\cot\theta\right),\quad p=-3,-2\ldots 2,$$

in terms of which Eqs. (A32)–(A39) become

$$2e^*{}_{,u}=-\mathscr{L}_0\overline{\epsilon}^*-2c^*\overline{X}^*, \tag{A41}$$

$$4E^*{}_{,u}=-\mathscr{L}_{-2}(\mathscr{L}_1e^*-2c^*\overline{\epsilon}^*), \tag{A42}$$

$$4C^*{}_{,u}=2c^{*2}\overline{c}^*{}_{,u}+2c^*\overline{M}^*+\mathscr{L}_1N^*+2e^*X^*, \tag{A43}$$

$$4D^*{}_{,u}=-\mathscr{L}_{-3}(\mathscr{L}_2C^*+2c^*N^*)+\tfrac{2}{3}\epsilon^*\mathscr{L}_1e^*-\tfrac{4}{3}c^*\epsilon^*\overline{\epsilon}^*+\tfrac{8}{3}E^*X^*. \tag{A44}$$

The supplementary conditions (A7) have the form

$$M_{,u}=-(c_{,u}^2+d_{,u}^2)-(X^2+Y^2)+\tfrac{1}{2}(c_{,\theta\theta}+3c_{,\theta}\cot\theta-2c)_{,u}, \tag{A45}$$

$$\begin{aligned}3N_{,u}=&-M_{,\theta}-2c\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)c_{,u}-2d\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)d_{,u}\\&-(cc_{,u})_{,\theta}-(dd_{,u})_{,\theta}-2(\epsilon X+\mu Y),\end{aligned} \tag{A46}$$

$$\begin{aligned}3P_{,u}=&\frac{1}{2}\,\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}+\cot\theta\right)(d_{,\theta}+2d\cot\theta)+2c\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)d_{,u}\\&-2d\left(\frac{\partial}{\partial\theta}+2\cot\theta\right)c_{,u}+(cd_{,u})_{,\theta}-(dc_{,u})_{,\theta}-2(\epsilon Y-\mu X),\end{aligned} \tag{A47}$$

$$\epsilon_{,u}=-X_{,\theta}-X\cot\theta, \tag{A48}$$

$$\mu_{,u}=-Y_{,\theta}-Y\cot\theta, \tag{A49}$$

which in terms of quantities (A40) simplify to

$$M^*{}_{,u}=-c^*{}_{,u}\overline{c}^*{}_{,u}-X^*\overline{X}^*+\tfrac{1}{2}\mathscr{L}_{-1}\mathscr{L}_{-2}c^*{}_{,u}, \tag{A50}$$

$$3N^*{}_{,u}=\mathscr{L}_0\overline{M}^*+2c^*\mathscr{L}_{-2}\overline{c}^*{}_{,u}+\mathscr{L}_0(c^*\overline{c}^*{}_{,u})-2\overline{\epsilon}^*X^*, \tag{A51}$$

$$\epsilon^*{}_{,u}=\mathscr{L}_{-1}X^*. \tag{A52}$$

The structure of the field equations is thus as follows. Nine functions γ, δ, F_{12}, F_{13}, M, N, P, ϵ and μ are prescribed on an ''initial'' hypersurface $u=u_0$, and four functions $c_{,u}$, $d_{,u}$, X and Y have to be prescribed for all u. Then the time evolution of gravitational and electromagnetic fields is fully determined. Functions $c_{,u}$ and $d_{,u}$ are the well-known *gravitational news functions*, functions X and Y–the *electromagnetic news functions*. Nonvanishing news functions $d_{,u}$ and Y correspond to a ''rotation'' of a radiating source. In vacuum space–times with hypersurface orthogonal Killing vector $\partial/\partial\phi$ we find $d=0$.

The total mass of the system at a given retarded time u is defined by

$$m(u)=\frac{1}{2}\int_0^\pi M(u,\theta)\sin\theta\,d\theta. \tag{A53}$$

(Notice that the definition of the total mass given in Ref. 11 is different; for example, in the Schwarzschild case it gives $4\pi m_{\text{Schw.}}$.) The time derivative of the total mass is equal to

$$m_{,u} = -\frac{1}{2}\int_0^\pi (c^*{}_{,u}\bar{c}^*{}_{,u} + X^*\bar{X}^*)\sin\theta\, d\theta$$
$$= -\frac{1}{2}\int_0^\pi (c_{,u}^2 + d_{,u}^2 + X^2 + Y^2)\sin\theta\, d\theta \leqslant 0. \tag{A54}$$

Hence, if any of the news functions is nonvanishing, the waves are radiated out and the mass of the system necessarily decreases.

There exist two quantities which are always conserved:

$$\frac{\partial}{\partial u}\int_0^\pi \tfrac{1}{2}\epsilon^* \sin\theta\, d\theta = 0. \tag{A55}$$

These are the ''electric'' and ''magnetic'' charges of the source.

The rate of loss of the electromagnetic energy radiated out from the system is given by

$$\frac{1}{2}\,\frac{\partial}{\partial u}\int_0^\pi X^*\bar{X}^* \sin\theta\, d\theta, \tag{A56}$$

which also implies the loss of mass as seen from Eq. (A54).

APPENDIX B: THE KILLING EQUATION

For an illustration we write down the expression for the Lie derivative component $\mathcal{L}_\eta g_{00} = 0$, where the metric has the form (3) and the vector η is given by Eqs. (11) and (13):

$$\begin{aligned}
\mathcal{L}_\eta g_{00} = & -\frac{1}{2r^3(2\,\mathrm{ch}^2-1)^{3/2}} e^{-3\gamma-2\beta} \\
& \times \Big\{ 2r^2(2\,\mathrm{ch}^2-1)e^{2(\beta+\gamma)}[-\sqrt{2\,\mathrm{ch}^2-1}\,e^{\gamma}(2A_{,u}r + B_{,u}Ve^{2\beta}) \\
& -2e^{2\gamma}(2\tilde{g}_{,u}\,\mathrm{sh}\,\mathrm{ch} + \tilde{f}_{,u})Ur^2 - (2\,\mathrm{ch}^2-1)2r^2\tilde{g}_{,u}W] \\
& +\tilde{g}4r^4e^{2(\beta+\gamma)}[W(2\,\mathrm{ch}^2-1) + 2U\,\mathrm{sh}\,\mathrm{ch}\,e^{2\gamma}][2\delta_{,u}\,\mathrm{sh}\,\mathrm{ch} - (2\,\mathrm{ch}^2-1)\gamma_{,u}] \\
& +\tilde{f}2re^{2\beta}[2r^3W(2\,\mathrm{ch}^2-1)^2(2W\delta_{,\theta}\,\mathrm{sh}\,\mathrm{ch} + (2\,\mathrm{ch}^2-1)(W_{,\theta} + \gamma_{,\theta}W)) \\
& -e^{2(\beta+\gamma)}(2\,\mathrm{ch}^2-1)^2(V_{,\theta} + 2\beta_{,\theta}V) \\
& +4r^3e^{2\gamma}(2\,\mathrm{ch}^2-1)((2\,\mathrm{ch}^2-1)^2UW\delta_{,\theta} + (2\,\mathrm{ch}^2-1)\mathrm{sh}\,\mathrm{ch}(U_{,\theta}W + UW_{,\theta} + 2W\gamma_{,u}) \\
& +W\delta_{,u}) + 2Ur^3e^{4\gamma}((2\,\mathrm{ch}^2-1)^3(U_{,\theta} + U\gamma_{,\theta}) + (2\,\mathrm{ch}^2-1)(2(2\,\mathrm{ch}^2-1)^2-1)\gamma_{,u} \\
& +2U\delta_{,\theta}\,\mathrm{sh}\,\mathrm{ch}(2\mathrm{ch}^2-1)^2 + 2\delta_{,u}\,\mathrm{sh}\,\mathrm{ch})] \\
& +A2r^2e^{\gamma}(2\,\mathrm{ch}^2-1)^{3/2}\Big[2r^2W((2\,\mathrm{ch}^2-1)(-rW\gamma_{,r} + W + rW_{,r}) \\
& +2rW\delta_{,r}\,\mathrm{sh}\,\mathrm{ch}) + 2r^2Ue^{4\gamma}((2\,\mathrm{ch}^2-1)(rU\gamma_{,r} + U + rU_{,r}) + 2rU\delta_{,r}\,\mathrm{sh}\,\mathrm{ch}) \\
& +e^{2(\beta+\gamma)}\Big(-V_{,r} + \frac{V}{r} - 2V\beta_{,r} + 4r\beta_{,u}\Big) \\
& +4r^2e^{2\gamma}((2\,\mathrm{ch}^2-1)rUW\delta_{,r} + \mathrm{sh}\,\mathrm{ch}(rU_{,r}W + 2UW + rUW_{,r}))\Big]
\end{aligned}$$

$$+Be^{2\beta+\gamma}(2\,\mathrm{ch}^2-1)^{3/2}\Big\{e^{2(\beta+\gamma)}\Big(Vr\Big(V_{,r}-\frac{1}{r}V+2V\beta_{,r}\Big)$$

$$-4Vr^2\beta_{,u}-2Ur^2(2\beta_{,\theta}V+V_{,\theta})\Big)$$

$$+2r^4e^{4\gamma}\Big((2\,\mathrm{ch}^2-1)\Big(2U^2r(U_{,\theta}+U\gamma_{,\theta}+\gamma_{,u})-VU\Big(\frac{1}{r}U+U_{,r}+U\gamma_{,r}\Big)\Big)$$

$$+2U^2r\,\mathrm{sh\,ch}\Big(2U\delta_{,\theta}-\frac{V}{r}\delta_{,r}+2\delta_{,u}\Big)\Big)$$

$$+4r^4e^{2\gamma}\Big(UW(2\,\mathrm{ch}^2-1)(-V\delta_{,r}+2Ur\delta_{,\theta}+2r\delta_{,u})$$

$$+\mathrm{sh\,ch}\Big(-V\Big(U_{,r}W+UW_{,r}+\frac{2}{r}UW\Big)+2Ur(U_{,\theta}W+UW_{,\theta})\Big)\Big)$$

$$+r^4\Big((2\,\mathrm{ch}^2-1)2W\Big(-2r\gamma_{,u}W-\frac{V}{r}(W+W_{,r}r-W\gamma_{,r}r)+2Ur(W_{,\theta}-\gamma_{,\theta}W)\Big)$$

$$+4W^2\,\mathrm{sh\,ch}(2Ur\delta_{,\theta}-V\delta_{,r}+2r\delta_{,u})\Big)\Big]\Big\}=0.$$

The Killing equation $\mathscr{L}_\eta g_{00}$ is the lengthiest among all Killing equations $\mathscr{L}_\eta g_{\mu\nu}=0$.

APPENDIX C: CORRECTIONS OF KILLING EQUATIONS GIVEN IN REF. 2

In Ref. 2 a number of misprints appear, and there is also a sign error which does not change the conclusions of the paper but makes wrong Eqs. (53)–(57), including the pathological solution given by Eq. (57) for η^u. Below we give the correct forms of all equations from Ref. 2 in which a misprint appears. First we give the correct forms of erroneous Killing equations from the Appendix in Ref. 2. These are (notice that functions f, g in Ref. 2 are denoted as $\tilde f$, $\tilde g$ in the present paper):

$$\mathscr{L}_\eta g_{01}=e^{2\beta}\beta_{,\theta}(2\tilde f r^{-1}e^{-\gamma}+2BU)+(Be^{2\beta})_{,u}+Ue^{\gamma}(r\tilde f_{,r}-\tilde f-\tilde f r\gamma_{,r})$$
$$+Ur^2e^{2\gamma}BU_{,r}+\Big(A-\frac{1}{2r}BVe^{2\beta}\Big)_{,r}+\frac{1}{r}B_{,r}Ve^{2\beta}, \tag{C1}$$

$$\mathscr{L}_\eta g_{11}=2e^{2\beta}B_{,r}, \tag{C2}$$

$$\mathscr{L}_\eta g_{13}=e^{-\gamma}\sin\theta(\tilde g-\tilde g r\gamma_{,r}-r\tilde g_{,r})+e^{2\beta}B_{,\phi}, \tag{C3}$$

$$\mathscr{L}_\eta g_{22}=-2re^{\gamma}\Big[\tilde f_{,\theta}+rBe^{\gamma}\gamma_{,u}+BUre^{\gamma}\gamma_{,\theta}+Bre^{\gamma}U_{,\theta}+e^{\gamma}\Big(Ae^{-2\beta}-\frac{1}{2r}BV\Big)(1+r\gamma_{,r})\Big]. \tag{C4}$$

In the main text in Ref. 2 the following equations contain misprints and are here corrected:

$$\text{Eq. (34):}\quad B_{,u}+2A^{(0)}{}_{,u}=0, \tag{C5}$$

$$\text{Eq. (46):}\quad f^{(1)}{}_{,\theta}+B_{,\theta\theta}c+A^{(1)}-c(A^{(0)}-\tfrac{1}{2}B)+BM-B(c_{,\theta}+2c\cot\theta)_{,\theta}=0, \tag{C6}$$

$$\text{Eq. (47):}\quad f^{(1)}\cot\theta+B_{,\theta}(c_{,\theta}+c\cot\theta)+c(A^{(0)}-\tfrac{1}{2}B)+A^{(1)}+BM-B\cot\theta(c_{,\theta}+2c\cot\theta), \tag{C7}$$

$$\text{Eq. (48):}\quad [B,_\theta \cot\theta + Bc,_u - (A^{(0)} - \tfrac{1}{2}B)]c = 0. \tag{C8}$$

As with Eq. (48) in Ref. 2 (which contains a sign error), the correct equation (C8) implies that either $c=0$ or the expression in the square brackets has to vanish. The first possibility gives vanishing news function and leads to translations. The second possibility leads to

$$A^{(0)} = \tfrac{1}{2}(B,_{\theta\theta} + B,_\theta \cot\theta + B), \tag{C9}$$

i.e., to Eq. (50) of the present paper. As it is shown below Eq. (50) in the main text, this case corresponds to the translational Killing vector under the presence of a straight cosmic string along the z-axis.

APPENDIX D: TRANSLATIONS IN SPACE–TIMES WITH A STRAIGHT COSMIC STRING

In this Appendix we show that function B in Eq. (76) corresponds to translations along the z-axis and t-axis in the space–time of an infinite thin cosmic string. Let us first recall some results from Ref. 14. The metric outside a straight nonrotating cosmic string along the z-axis can be written in cylindrical coordinates as

$$ds^2 = dt^2 - d\rho^2 - dz^2 - C^2\rho^2\, d\phi^2, \tag{D1}$$

where $\phi \in [0, 2\pi)$ and $C \in (0,1]$ is a constant. [Equation (D1) represents also the asymptotic form of the metric corresponding to a spatially bounded system and the cosmic string along the z-axis.] Introducing spherical flat-space coordinates $\{R, \vartheta, \phi\}$ by $\rho = R\sin\vartheta$, $z = R\cos\vartheta$, $\phi = \phi$, and flat-space retarded time $U = t - R$, we get

$$ds^2 = dU^2 + 2dUdR - R^2(d\vartheta^2 + C^2\sin^2\vartheta\, d\phi^2). \tag{D2}$$

We can now go over to Bondi's coordinates $\{u, r, \theta, \phi\}$—in which the metric has asymptotically Bondi's form—by assuming expansions

$$\begin{aligned} U &= \mathring{\pi}(u,\theta) + O(r^{-1}), \\ R &= q(u,\theta)r + O(r^0), \\ \vartheta &= \mathring{\tau}(u,\theta) + O(r^{-1}). \end{aligned} \tag{D3}$$

Comparing then the resulting metric with the general form of Bondi's metric one obtains the expressions for functions $\mathring{\pi}$, q, $\mathring{\tau}$, ... (see Ref. 14 for details) and the expression for the news function $c,_u$. Here we need only the following results:

$$q(\theta) = \frac{\sin\theta}{C\sin\vartheta}, \qquad \mathring{\tau},_\theta = \pm\frac{1}{q}, \tag{D4}$$

and the news function is

$$c,_u = \frac{C^2 - 1}{2\sin^2\theta}. \tag{D5}$$

Taking the θ-derivative of q, using $\mathring{\tau},_\theta = \pm 1/q$, and excluding the original flat-space ϑ, we obtain the equation

$$q,_\theta^2 - 2qq,_\theta \cot\theta + q^2\left(\cot^2\theta - \frac{C^2}{\sin^2\theta}\right) + 1 = 0, \tag{D6}$$

which implies

$$q_{,\theta}=q\cot\theta\pm\sqrt{\frac{q^2C^2}{\sin^2\theta}-1}. \tag{D7}$$

The solution is

$$q=\frac{\sin\theta}{2C}\left[\left(\chi\frac{\sin\theta}{\cos\theta+1}\right)^C+\left(\chi\frac{\sin\theta}{\cos\theta+1}\right)^{-C}\right],\quad \chi=\text{const.} \tag{D8}$$

Consider now translations along the z-axis and t-axis. In coordinates $\{t,\rho,z,\phi\}$ they have the form $\zeta^\mu_{(z)}=[0,0,a_0,0]$ and $\zeta^\mu_{(t)}=[b_0,0,0,0]$, with a_0 and b_0 constant, so that their general combination is just $\zeta^\mu_{(z)}+\zeta^\mu_{(t)}$.

The asymptotic form of this linear combination in Bondi's coordinates reads

$$\begin{aligned}
\zeta^u_{(z)}+\zeta^u_{(t)}&=b_0q+\frac{a_0}{C}\sin^2\theta\left(\frac{q}{\sin\theta}\right)_{,\theta}+O(r^{-1}),\\
\zeta^r_{(z)}+\zeta^r_{(t)}&=\frac{b_0}{2q}(q_{,\theta}{}^2-q^2+1)\\
&\quad+\frac{a_0}{2Cq^2}[(-q_{,\theta}{}^2+q^2+1)(q\cos\theta-q_{,\theta}\sin\theta)+2q_{,\theta}\sin\theta]+O(r^{-1}),\\
\zeta^\theta_{(z)}+\zeta^\theta_{(t)}&=-\frac{b_0q_{,\theta}}{r}+\frac{a_0}{Cqr}[q_{,\theta}\ (q\cos\theta-q_{,\theta}\sin\theta)-\sin\theta]+O(r^{-2}),\\
\zeta^\phi_{(z)}+\zeta^\phi_{(t)}&=0.
\end{aligned} \tag{D9}$$

Substituting for q from Eq. (D8) we obtain

$$\begin{aligned}
\zeta^u_{(z)}+\zeta^u_{(t)}&=\frac{\sin\theta}{2C}\left(\chi\frac{\sin\theta}{\cos\theta+1}\right)^C(b_0+a_0)+\frac{\sin\theta}{2C}\left(\chi\frac{\sin\theta}{\cos\theta+1}\right)^{-C}(b_0-a_0)\\
&=\sin\theta\left(\frac{\sin\theta}{\cos\theta+1}\right)^C\frac{(b_0+a_0)\chi^C}{2C}+\sin\theta\left(\frac{\sin\theta}{\cos\theta+1}\right)^{-C}\frac{(b_0-a_0)\chi^{-C}}{2C},
\end{aligned} \tag{D10}$$

and similar expressions for $\zeta^r_{(z)}+\zeta^r_{(t)}$ and $\zeta^\theta_{(z)}+\zeta^\theta_{(t)}$; these are somewhat lengthy and are thus not written here explicitly. Comparing now the last results (including those for $\zeta^r_{(z)}+\zeta^r_{(t)}$ and $\zeta^\theta_{(z)}+\zeta^\theta_{(t)}$) with a general asymptotic form of the Killing vector representing a supertranslation, i.e., with Eq. (58) with B given by Eq. (76), we find that this Killing vector is just equal to $\zeta^\mu_{(z)}+\zeta^\mu_{(t)}$ if the constant parameters are related by

$$a=\frac{(b_0+a_0)\chi^C}{2C}, \tag{D11}$$

$$b=\frac{(b_0-a_0)\chi^{-C}}{2C}. \tag{D12}$$

Therefore the supertranslational Killing vector is in fact the translational Killing vector in the asymptotically flat space–time with an infinite cosmic string along the z-axis; $b_0=0$ corresponds to a translation along the z-axis, and $a_0=0$—along the t-axis.

For the special case $C=1$ (no string) the space–time is asymptotically flat, and

$$q=\cos\theta\left(\frac{1}{2\chi}-\frac{\chi}{2}\right)+\frac{1}{2\chi}+\frac{\chi}{2} \tag{D13}$$

corresponds to a standard translation where B is given by Eq. (77).

APPENDIX E: ELECTROMAGNETIC FIELD

1. Translational case

The electromagnetic tensor reads:

$$F_{01}=-\epsilon_1(\theta)\frac{1}{r^2}+[uB^{-2}(-\epsilon_1 BB_{,\theta\theta}+\epsilon_1 B_{,\theta}^2-\epsilon_1 BB_{,\theta}\cot\theta-\epsilon_{1,\theta}BB_{,\theta}) + e_{1,\theta}+e_1\cot\theta]\frac{1}{r^3}+O(r^{-4}), \tag{E1}$$

$$F_{02}=[\epsilon_{1,\theta}+\epsilon_1 B_{,\theta}B^{-1}]\frac{1}{r}+O(r^{-2}), \tag{E2}$$

$$F_{03}=-\mu_1\sin\theta\, B_{,\theta}B^{-1}\frac{1}{r}+O(r^{-2}), \tag{E3}$$

$$F_{12}=[-\epsilon_1 B_{,\theta}B^{-1}u+e_1]\frac{1}{r^2}+O(r^{-3}), \tag{E4}$$

$$F_{13}=[\mu_1\sin\theta\, B_{,\theta}B^{-1}u+f_1\sin\theta]\frac{1}{r^2}+O(r^{-3}), \tag{E5}$$

$$F_{23}=-\mu_1\sin\theta-\{B^{-2}[(\mu_1\sin\theta)_{,\theta}BB_{,\theta} + \mu_1\sin\theta\, BB_{,\theta\theta}-\mu_1\sin\theta\, B_{,\theta}^2]u+(f_1\sin\theta)_{,\theta}\}\frac{1}{r}+O(r^{-2}). \tag{E6}$$

2. Boost case

Assuming the boost-rotation symmetry, the electromagnetic tensor has the form:

$$F_{01}=-\frac{\mathscr{E}}{u}\cot\theta\frac{1}{r^2}+\left[\frac{\cot\theta}{2}(\mathscr{E}_{,w}w+\mathscr{E})+\frac{\cot\theta}{u^2}(\mathscr{G}_{,w}w+\mathscr{G})\right]\frac{1}{r^3}+O(r^{-4}), \tag{E7}$$

$$F_{02}=\frac{\mathscr{E}}{u^2}+\left[\frac{1}{u\sin^2\theta}(\mathscr{E}_{,w}w-\mathscr{E})-\frac{\mathscr{E}_{,w}\sin\theta}{2u^2}+\frac{1}{u^3}(\mathscr{G}_{,w}w+2\mathscr{G})\right]\frac{1}{r}+O(r^{-2}), \tag{E8}$$

$$F_{03}=\frac{\mathscr{B}}{u}+\left[-\frac{1}{2}(\mathscr{B}_{,w}w-\mathscr{B})+\frac{1}{u^2}(\mathscr{D}_{,w}w+\mathscr{D})\right]\frac{1}{r}+O(r^{-2}), \tag{E9}$$

$$F_{12}=\left[\frac{1}{2}\mathscr{E}+\frac{\mathscr{G}}{u^2}\right]\frac{1}{r^2}+O(r^{-3}), \tag{E10}$$

$$F_{13}=\left[-\frac{1}{2}\mathscr{B}u+\frac{\mathscr{D}}{u}\right]\frac{1}{r^2}+O(r^{-3}), \tag{E11}$$

$$F_{23}=-\mathscr{B}\cot\theta+\cos\theta\left(\frac{1}{2}\mathscr{B}_{,w}-\frac{\mathscr{D}_{,w}}{u^2}\right)\frac{1}{r}+O(r^{-2}). \tag{E12}$$

Let us yet illustrate the general asymptotic forms (E8), (E9) for boost-rotation symmetric electromagnetic fields by specific examples. As follows from Eqs. (E8) and (E9), the first nonvanishing terms are given by $F_{02}=\mathscr{E}(w)/u^2$, $F_{03}=\mathscr{B}(w)/u$, with $w=\sin\theta/u$. Denote by e, p, m electric charge, electric dipole and magnetic dipole moments, respectively, and by α^{-1} the magnitude of acceleration. For a uniformly accelerated electric monopole (producing Born's solution), we find $\mathscr{E}(w)=e\alpha^2w/(1+\alpha^2w^2)^{3/2}$, $\mathscr{B}(w)=0$. For a uniformly accelerated electric dipole (see Ref. 20 for the complete field), we get $\mathscr{E}(w)=p\alpha w(2-\alpha^2w^2)/(1+\alpha^2w^2)^{5/2}$, $\mathscr{B}(w)=0$, whereas for a magnetic dipole (see Ref. 21 for the complete field), one finds $\mathscr{E}(w)=0$, $\mathscr{B}(w)=-\alpha mw^2(2-\alpha^2w^2)/(1+\alpha^2w^2)^{5/2}$. We can easily check that $\int(\mathscr{E}\cos\theta/u)d\theta=0$ in the case of electric dipole and $\int\mathscr{B}\cot\theta\,d\theta=0$ in the magnetic case—total charges indeed vanish.

[1] J. Bičák and B. Schmidt, Phys. Rev. D **40**, 1827 (1989).
[2] J. Bičák and B. G. Schmidt, J. Math. Phys. **25**, 600 (1984).
[3] J. Bičák, Proc. R. Soc. London, Ser. A **302**, 201 (1968).
[4] J. Bičák, P. Reilly, and J. Winicour, Gen. Relativ. Gravit. **20**, 171 (1988).
[5] R. Gomźz, P. Papadopoulos, and J. Winicour, J. Math. Phys. **35**, 4184 (1994).
[6] M. Alcubierre, C. Gundlach, and F. Siebel, in Abstracts of Plenary Lectures and Contributed Papers of the 15th International Conference on General Relativity and Gravity, Pune, 1997 (Inter-University Centre for Astronomy and Astrophysics, Pune), p. 83.
[7] S. W. Hawking, G. T. Horowitz, and S. F. Ross, Phys. Rev. D **51**, 4302 (1995).
[8] J. Bičák, in *Relativistic Gravitation and Gravitational Radiation*, Proceedings of the Les Houches School of Physics, edited by J.-A. Marck and J.-P. Lasota (Cambridge University Press, Cambridge, 1995), p. 67.
[9] J. Plebański and M. Demiański, Ann. Phys. (N.Y.) **98**, 98 (1976).
[10] H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. R. Soc. London, Ser. A **269**, 21 (1962); R. K. Sachs, *ibid*. **270**, 103 (1962).
[11] M. G. J. van der Burg, Proc. R. Soc. London, Ser. A **310**, 221 (1969) gives the explicit treatment of the asymptotic behavior of the Einstein–Maxwell fields in the Bondi–Sachs coordinates.
[12] A. Ashtekar, J. Bičák, and B. Schmidt, Phys. Rev. D **55**, 669 (1997).
[13] A. Ashtekar, J. Bičák, and B. Schmidt, Phys. Rev. D **55**, 687 (1997).
[14] J. Bičák and B. Schmidt, Class. Quantum Grav. **6**, 1547 (1989).
[15] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[16] A. Ashtekar and B. Schmidt, J. Math. Phys. **21**, 862 (1980).
[17] We keep the original notation of Sachs [see H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. R. Soc. London, Ser. A **269**, 21 (1962)], which is also used in J. Bičák and B. G. Schmidt, J. Math. Phys. **25**, 600 (1984). Today one rather uses the Newman–Penrose notation $\{l^\alpha,n^\alpha,m^\alpha,\bar{m}^\alpha\}$ for the null tetrad. Notice, however, that usually one chooses $l^r=1$, where r is an affine parameter along an outgoing null geodesic $\{u,\theta,\phi\}=\text{const}$, while in Bondi's formalism r is the luminosity distance, and $k^r\neq1$ in general.
[18] A. Pravdová, Ph.D. dissertation, Department of Theoretical Physics, Charles University, Prague, 1998.
[19] P. T. Chruściel, M. A. H. MacCallum, and D. B. Singleton, Philos. Trans. R. Soc. London, Ser. A **350**, 113 (1995).
[20] J. Bičák and R. Muschall, Wiss. Z. Friedrich-Schiller-Univ. Jena Nat. Wiss. Reihe **39**, 15 (1990).
[21] V. Pravda, Ph.D. dissertation, Department of Theoretical Physics, Charles University, Prague, 1998; see also V. Pravda and A. Pravdová, in Proceedings of The Week of Postgraduate Students, Faculty of Mathematics and Physics, Charles University, Prague, 1998, and gr-qc/9806114, 1998.

Chaos in Robertson–Walker cosmology

Luca Bombelli[a)]
Department of Physics and Astronomy, University of Mississippi, University, Mississippi 38677

Fernando Lombardo[b)]
Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Buenos Aires, Argentina

Mario Castagnino[c)]
Instituto de Astronomía y Física del Espacio, Casilla de Correos 67, Sucursal 28, 1428 Buenos Aires, Argentina

(Received 2 March 1998; accepted for publication 27 April 1998)

Chaos in Robertson–Walker cosmological models where gravity is coupled to one or more scalar fields has been studied by a few authors, mostly using numerical simulations. In this paper we begin a systematic study of the analytical aspect. We consider one conformally coupled scalar field and, using the fact that the model is integrable when the field is massless, we show in detail how homoclinic chaos arises for nonzero masses using a perturbative method. © *1998 American Institute of Physics.* [S0022-2488(98)00511-8]

I. INTRODUCTION

In Ref. 1, Calzetta and El Hasi presented an argument, supported by numerical evidence, for the appearance of chaotic behavior in a spatially closed Robertson–Walker (RW) cosmological model, filled with a conformally coupled massive scalar field. Because cosmological models of this type are practically the simplest possible ones, but give in some sense a rough, smoothed out indication of the average properties of a more realistic cosmology, they have frequently been used as a first testing ground on which to try out approaches to both classical and quantum problems in cosmology. In addition, because of their kinematical simplicity, one would expect the dynamics of these models to be equally simple. It thus may come somewhat as a surprise that chaos should appear at this level, although it is known that it can appear in a nonlinear dynamical system with two degrees of freedom, and one may suspect that it is in fact generic for relativistic systems.

The study of chaos in relativistic cosmology began more than 20 years ago with the study by the Russian school[2] and Misner's group[3] on diagonal Bianchi type IX (mixmaster) vacuum cosmologies. This research is, by itself, very interesting and still active,[4] but the hopes initially placed in these models, essentially the solution of the horizon problem from the mixing effect associated with chaos, could not be sustained, and the Bianchi IX cosmologies have been mostly used as toy models, rather than first approximations from which to extract estimates of physical quantities. On the other hand, although the question of which RW model best approximates our inhomogeneous, anisotropic universe is not yet settled, there is agreement on the relevance of RW models both for classical and quantum cosmology,[5] and physical predictions are often made based on them. Therefore the presence of chaos, in this kind of cosmologies, enlarges the avenues of research in fundamental subjects such as the arrow of time and the issue of irreversibility in the evolution of the universe, or the transition from the quantum regime to the classical one, with the corresponding appearance of decoherence and correlations and related phenomena like particle creation, and has been used in considerations related to physical phenomena in inflationary universes.[6–8]

In their work, Calzetta and El Hasi used mostly a numerical method to investigate the chaotic behavior of the model, but also sketched an analytical, perturbative method by which they considered the model as a perturbation of the integrable one obtained with a vanishing mass for the

[a)]Electronic mail: bombelli@olemiss.edu
[b)]Electronic mail: lombardo@df.uba.ar
[c)]Electronic mail: castagni@iafe.uba.ar

 © 1998 American Institute of Physics

scalar field, and split the Hamiltonian of the system into an integrable part and a coupling term. With a perturbative argument commonly used in the treatment of near-integrable systems, based on the Melnikov criterion for homoclinic chaos and Chirikov's resonance overlap criterion,[9] they argued qualitatively that the Kolmogorov-Arnold-Moser (KAM) tori of the integrable part are destroyed and replaced by stochastic layers under the effect of the perturbation. The resulting strong indications of chaotic behavior were confirmed by the numerical analysis, which allowed them to go beyond perturbation theory.

The generic transition to chaos in a perturbed integrable Hamiltonian system (''soft chaos'') can be seen as taking place in two steps.[9,10] In the first one, a perturbation of the Hamiltonian, with a nonvanishing Fourier component in resonance with one of the tori of the unperturbed system, modifies the dynamics in a neighborhood of this torus to give a new integrable system, which, however, has heteroclinic orbits even if before there were none. These orbits are separatrices, which act as seeds of chaos when in the second step one takes into account the effect of the remaining components of the perturbation on the new integrable dynamics; it is here that one may use the Melnikov method to determine whether in effect the dynamics becomes chaotic. What this method provides is a topological criterion for detecting a chaotic ''tangle'' in the orbits near the separatrices, and is therefore free of the coordinate ambiguities that have arisen, e.g., in the study of the chaotic nature of Bianchi models by other methods; for a quick introduction to the Melnikov method see, e.g., Refs. 11 or 12, and Ref. 13 for a more technical exposition.

The goal of this paper is to continue the work in Ref. 1 by starting to examine in detail the analytical treatment of the model; in future work, we will systematically extend our work to more general models. We will use a different perturbative expansion of the Hamiltonian, and show the onset of chaotic behavior by an explicit computation of the Melnikov integral. Other recent work on chaos in Robertson–Walker models coupled to scalar fields has focused on their application to inflation,[7] again using mostly numerical methods, and an analysis based on the Painlevé analysis of differential equations,[14] which confirmed the nonintegability of the models. The main advantage of our work is the amount of information it can give us regarding the resonances responsible for the onset of chaos and the chaotic region of phase space.

The paper is organized as follows: in Sec. II, we set up the perturbative method we will use, by choosing appropriate variables to describe the cosmological model and separating the corresponding Hamiltonian into an ''unperturbed'' term and a ''perturbation;'' in Sec. III we analyze the first step above for our example, i.e., we discuss the local dynamics in the vicinity of one of the resonant tori, while in Sec. IV we develop the second step, discussing the effect of other perturbation terms on the local dynamics, which becomes chaotic as Melnikov's method shows; Sec. V contains concluding remarks on possible future developments.

II. DYNAMICAL VARIABLES AND HAMILTONIAN: SETUP FOR THE PERTURBATIVE TREATMENT

Let us consider the closed Robertson–Walker (RW) metric for a homogeneous and isotropic universe, written in the form

$$ds^2 = a^2(t)[-dt^2 + d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\varphi^2)], \tag{2.1}$$

where $0 \leqslant \varphi \leqslant 2\pi$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \chi \leqslant \pi$ are the angular coordinates on S^3, and t is the usual ''conformal time'' (see, e.g., Ref. 15). We shall consider only models possessing a cosmic singularity $a=0$, chosen for convenience to occur at $t=0$. We shall also assume that after the big crunch a new cosmological cycle begins, but a changes sign, in such a way that the evolution is smooth, and this process is repeated an infinite number of times. This well-known possibility of extending the evolution beyond the $a=0$ singularity is an artifact of the symmetry possessed by the RW models (we have no grounds for believing that it can be generalized), but it is convenient for us because it allows us to assume that time runs up to $t=+\infty$, and thus to use techniques for detecting chaos, including its very definition, that would otherwise be inapplicable. Any conclusion we draw from the model would of course not have to rely on this extension; the time scale for any chaotic behavior to set in must be shorter than one cycle for its physical consequences to be meaningful.

We start with the Einstein–Hilbert action for the gravitational field,

$$S_{\mathrm{g}}[g]=\frac{1}{16\pi G}\int d^4x\sqrt{-g}R, \tag{2.2}$$

and the action

$$S_{\mathrm{f}}[\Phi,g]=-\frac{1}{2}\int d^4x\sqrt{-g}\left[g^{ab}\nabla_a\Phi\nabla_b\Phi+\left(\mu^2+\frac{1}{6}R\right)\Phi^2\right] \tag{2.3}$$

for a real, conformally coupled scalar field of mass μ. To write down the Hamiltonian, use the fact that the square root of the determinant of the metric (2.1) and its scalar curvature are, respectively,

$$\sqrt{-g}=a^4\sin^2\chi\sin\theta,\quad R=6\left(\frac{\ddot{a}}{a^3}+\frac{1}{a^2}\right), \tag{2.4}$$

where an overdot denotes a t-derivative, and reparametrize the scalar field (which must be homogeneous for consistency) by $\Phi\mapsto\phi:=\sqrt{4\pi G/3}a\Phi$. Then we obtain, up to a constant overall factor,

$$H(a,\phi;\pi,p)=\tfrac{1}{2}[-(\pi^2+a^2)+(p^2+\phi^2)+\mu^2a^2\phi^2], \tag{2.5}$$

where π and p are the momenta conjugate to a and ϕ, respectively. Here, as usual with general relativistic systems, time reparametrization invariance requires that we impose that all solutions satisfy the constraint

$$H=0, \tag{2.6}$$

which is just the (0,0)-component of the Einstein equation.

The system described by the Hamiltonian (2.5) consists of two harmonic oscillators (of which one is ''inverted,'' as is to be expected from any degree of freedom related to the spatial volume element), coupled through a term proportional to μ^2. For $\mu^2=0$, the system is trivially integrable; our main goal is to analyze the effect of the coupling term, at least for small values of μ^2. However, in view of the perturbative treatment to be carried out, the form (2.5) has two disadvantages: the ''perturbation'' term $\mu^2a^2\phi^2$ is not always small, in the sense that for a fixed value of μ^2 it cannot be uniformly bounded for all orbits of the unperturbed system; and the latter system is degenerate, in the sense that the frequencies of both oscillators are constant. We will not attempt to deal directly with the first inconvenience, which at any rate is not really a problem for perturbations of a fixed orbit; we will instead perform a change of variables that will allow us to isolate a different integrable Hamiltonian from H, with enough of the degeneracy removed for us to be able to use the Melnikov method of detecting homoclinic chaos. Regarding the perturbation, it will be sufficient for us to know that it is a well-behaved function on each trajectory.

Following Ref. 1, we first replace the dynamical variables p and ϕ by new variables j and φ, respectively, defined as

$$\phi=\sqrt{\frac{2j}{\omega}}\sin\varphi,\quad p=\sqrt{2\omega j}\cos\varphi, \tag{2.7}$$

where $\omega=\sqrt{1+\mu^2a^2}$ is the instantaneous frequency of the field; thus, this is a first approximation to action-angle variables for the scalar field, which takes into account the coupling term, but not the fact that a is a dynamical variable, and it is a key trick in this treatment of the model. To complete the canonical transformation, we must introduce a new momentum variable conjugate to a, given by

$$P:=\pi-\frac{\mu^2aj}{2\omega^2}\sin 2\varphi, \tag{2.8}$$

and arising from the fact that ω is a-dependent. Now the Hamiltonian in terms of the new variables can be written as a sum $H=\hat{H}_0+\delta\hat{H}$, of an unperturbed Hamiltonian

$$\hat{H}_0(a,\varphi;P,j) = -\tfrac{1}{2}(P^2+a^2) + j\sqrt{1+\mu^2 a^2}, \tag{2.9}$$

which is obviously integrable, since it has the two commuting constants of the motion $\hat{H}_0$ and j, and a perturbation

$$\delta\hat{H}(a,\varphi;P,j) = -\frac{\mu^2 aPj}{2(1+\mu^2 a^2)}\sin 2\varphi - \left[\frac{\mu^2 aj}{4(1+\mu^2 a^2)}\right]^2 (1-\cos 4\varphi). \tag{2.10}$$

However, the system (2.9) is not easy to integrate; since we will need to perform explicit calculations, and in particular to write it in terms of action-angle variables, we find it convenient to simplify it somewhat.

In Ref. 1, this simplification was achieved by noticing that for most of the unperturbed motion, the value of ω is very large, and using the $\omega \gg 1$ approximation to $\hat{H}_0$. In addition, because the treatment was a perturbative one in the parameter μ^2, or more precisely $\mu^2 j$, the second term in the Hamiltonian perturbation (2.10) was dropped, since it contains higher powers of μ^2 than the first one. Notice that, implicit in this way of proceeding is the idea that one considers the μ^2 in $\hat{H}_0$ as a fixed parameter, while in $\delta\hat{H}$, $\mu^2 =: \epsilon$ is rendered temporarily independent of the first one for the sake of the perturbative treatment, and allowed to vary in a neighborhood of $\mu^2 = \epsilon = 0$. (One may be able to set up a perturbation theory for Hamiltonians of the type $H_0(q,p;\epsilon) + \epsilon\delta H(q,p)$, where $H_0(q,p;\epsilon)$ is integrable for all ϵ, but to our knowledge such a theory is not available yet.)

Here we will follow a different path. Since we are going to expand the perturbation in powers of μ^2 and neglect terms higher than μ^4, we will do the same for $\hat{H}_0$, and retain there only the terms up to order μ^2 while we will put in the perturbation the higher order ones. The new unperturbed system obtained in this way will be explicitly integrable. As the perturbation parameter, we will use μ^2, rather than the dimensionless $\mu^2 j$; this is mainly to avoid having to explain already at this stage what values of the variable j we are interested in. At any rate, strictly speaking we will not have to talk about $\epsilon = \mu^2$ being ''small'' (if we did, we might say that $\mu/\mu_{\mathrm{P}} \ll 1$, where μ_{P} is the Planck mass); the perturbative results we will obtain are valid for ϵ in some interval near zero, and such topological statements are ''dimensionless.'' Expanding $\sqrt{1+\mu^2 a^2} = 1 + (1/2)\mu^2 a^2 - (1/4)\mu^4 a^4 + o(\mu^4)$, we obtain a new breakup of the Hamiltonian into

$$H = H_0 + \delta H, \tag{2.11}$$

where the unperturbed part is now

$$H_0(a,\varphi;P,j) = -\tfrac{1}{2}(P^2+a^2) + j\left(1 + \tfrac{1}{2}\mu^2 a^2\right), \tag{2.12}$$

and the perturbation differs from (2.10) by terms proportional to $\mu^4 a^4$ and of higher order in μ^2, namely,

$$\delta H = -\tfrac{1}{2}\mu^2 aPj\,\sin 2\varphi + \mu^4\left[\tfrac{1}{2}a^3 Pj\,\sin 2\varphi - \tfrac{1}{4}a^4 j - \tfrac{1}{16}a^2 j^2(1-\cos 4\varphi)\right] + \mathrm{o}(\mu^4). \tag{2.13}$$

One of the action variables of the new unperturbed system is j itself; to find the other one, we compute the integral $k := (2\pi)^{-1}\int P\,da$, along the trajectory specified by the values of the conserved quantities j and $H_0 = \mathrm{const} := h_0$ (ultimately, we will be interested only in trajectories near $h_0 = 0$, in order to satisfy the constraint $H_0 + \delta H = 0$, but for now we keep the treatment more general), for which

$$P(a) = \pm\sqrt{2(j-h_0) - (1-\mu^2 j)a^2}. \tag{2.14}$$

The integration yields

$$k = \frac{j-h_0}{\sqrt{1-\mu^2 j}}, \tag{2.15}$$

and the angle variables canonically conjugate to k and j, respectively, are given by

$$\theta = \arctan(\sqrt{1-\mu^2 j}\, a/P), \tag{2.16}$$

$$\delta = \varphi - \frac{\mu^2 a P}{4(1-\mu^2 j)}, \tag{2.17}$$

as can be checked by direct computation of the Poisson brackets. Notice that μ^2 here is not the infinitesimal ϵ, but we do want to consider it as being small, so for j not too big $1-\mu^2 j > 0$. With these new variables we can now write down our final expression for the unperturbed Hamiltonian,

$$H_0(k,j) = j - k\sqrt{1-\mu^2 j}. \tag{2.18}$$

Choosing a value for h_0 will then impose a relationship between j and k; in particular, for $h_0 \approx 0$, since j is positive, see (2.12), k must be positive as well. For the perturbation, we invert the transformation (2.15)–(2.17), and substitute in (2.13), which, keeping only the lowest order terms in $\epsilon = \mu^2$, gives

$$\begin{aligned}\delta H = {} & \frac{1}{4}\epsilon kj[\cos(2\theta+2\delta) - \cos(2\theta-2\delta)] + \frac{1}{4}\epsilon^2\Bigg[-\frac{3}{2}k^2 j - \frac{1}{4}kj^2 \\ & + \left(2k^2 j + \frac{1}{4}kj^2\right)\cos 2\theta - \frac{1}{2}k^2 j\,\cos 4\theta - \frac{1}{2}k^2 j\,\cos(2\delta - \pi/2) + \frac{1}{4}kj^2\,\cos 4\delta \\ & + k^2 j\,\cos(2\theta - 2\delta) - k^2 j\,\cos(2\theta + 2\delta) - \frac{1}{8}kj^2\,\cos(2\theta + 4\delta) \\ & - \frac{1}{8}kj^2\,\cos(2\theta - 4\delta) + \frac{\sqrt{5}}{4}k^2 j\,\cos(4\theta + 2\delta + \psi) - \frac{\sqrt{5}}{4}k^2 j\,\cos(4\theta - 2\delta + \psi)\Bigg] \\ & + O(\epsilon^3),\end{aligned} \tag{2.19}$$

where $\psi := -\arcsin(1/\sqrt{5})$, and we have chosen to write the various terms in the form that will be most useful in the following. This completes the setup for the perturbative treatment of the model. The dynamics of the unperturbed Hamiltonian H_0 is trivial, and gives a conditionally periodic motion $\theta = \theta_0 + \omega_k t$, $\delta = \delta_0 + \omega_j t$, with frequencies given by

$$\omega_k = \frac{\partial H_0}{\partial k} = -\sqrt{1-\mu^2 j},$$

$$\omega_j = \frac{\partial H_0}{\partial j} = \frac{\mu^2 k + 2\sqrt{1-\mu^2 j}}{2\sqrt{1-\mu^2 j}}. \tag{2.20}$$

In the next section we will start to take a look at the perturbed dynamics.

III. EFFECT OF THE RESONANT TERMS IN THE PERTURBATION

The effect of the perturbation δH is felt in particular on the rational tori of the unperturbed dynamics, the ones where the motion becomes periodic, because the frequencies (2.20) associated with the two degrees of freedom are related by the resonance condition

$$n_0\omega_k + m_0\omega_j = 0, \tag{3.1}$$

for some integer numbers n_0 and m_0. Using the explicit form of the frequencies in the resonance condition, we obtain a relationship between the action variables k and j, namely, that

$$\frac{2(1-\mu^2 j)}{\mu^2 k + 2\sqrt{1-\mu^2 j}} = \frac{m_0}{n_0} \tag{3.2}$$

must be a rational number. The latter equation can be expressed as

$$\mu^2 k = 2\frac{n_0}{m_0}(1-\mu^2 j) - 2\sqrt{1-\mu^2 j}, \tag{3.3}$$

and, since k must be positive in order for $h_0 \approx 0$, we see that a necessary condition is that n_0 and m_0 have the same sign (we can take them to be positive) with n_0 strictly greater than m_0.

To find out which of the resonant tori of the unperturbed system are affected by δH, we must consider the perturbation (2.19) as a Fourier series with respect to the angular variables θ and δ, of the form

$$\begin{aligned} \delta H(\theta,\delta;k,j) &= \sum_{i=1}^{\infty} \delta H^{(i)}(\theta,\delta;k,j), \\ \delta H^{(i)}(\theta,\delta;k,j) &= \epsilon^i \sum_{n,m} V_{nm}^{(i)}(k,j)\cos(n\theta + m\delta + \psi_{nm}^{(i)}), \end{aligned} \tag{3.4}$$

and check which of the Fourier coefficients $V_{nm}^{(1)}(k,j)$ are nonvanishing. The order ϵ terms in (2.19) are

$$\delta H^{(1)} = \tfrac{1}{4}\,\epsilon kj[\cos(2\theta+2\delta) - \cos(2\theta-2\delta)]. \tag{3.5}$$

Therefore to order μ^2 in the perturbation, the resonant tori which are broken by the perturbation are those for which

$$n_0 = \pm m_0 = 2.$$

However, neither of these pairs satisfies the conditions given below (3.3), which means that these resonant tori are not present in the physically relevant region near $h_0 = 0$.

This forces us to consider the Fourier components of order ϵ^2 in the perturbation (2.19). In order to look at their effect on the dynamics, however, we must somehow consider them as perturbations of a system which includes the order ϵ terms. The dynamical behavior of such a system is qualitatively similar to that of the unperturbed one in the region near $h_0 = 0$, because H_0 has no tori there which resonate with the order ϵ terms, and to set up the system we can use the normal techniques for nonresonant perturbations of integrable systems.

The method (see, e.g., Ref. 16) consists in performing an ϵ-dependent canonical transformation to new coordinates $(\theta',\delta';k',j')$, in which the new Hamiltonian has no perturbation terms of order ϵ, and is obtained by using $k'\theta + j'\delta + \Phi(\theta,\delta;k',j')$ as a generating function, where

$$\Phi(\theta,\delta;k',j') = -\epsilon \sum_{n,m} \frac{V_{nm}^{(1)}(k',j')}{n\omega_k(k',j') + m\omega_j(k',j')} \sin(n\theta + m\delta + \psi_{nm}). \tag{3.6}$$

While the new canonical coordinates cannot be explicitly expressed in terms of the old ones, they differ by a series of powers of ϵ, and for our purposes it will be sufficient to calculate the term of order ϵ. A calculation shows that the new unperturbed Hamiltonian $H_0'(k',j') = H_0(k',j')$, i.e., it is given by the same function (2.18) as in the previous variables, while the perturbation is now of the form

$$\begin{aligned} \delta H' = \epsilon^2 [& V_{00}'^{(2)}(k',j') + V_{20}'^{(2)}(k',j')\cos 2\theta' + V_{02}'^{(2)}(k',j')\cos(2\delta' - \pi/2) + V_{40}'^{(2)}(k',j') \\ & \times \cos 4\theta' + V_{04}'^{(2)}(k',j')\cos 4\delta' + V_{22}'^{(2)}(k',j')\cos(2\theta' + 2\delta') + V_{2-2}'^{(2)}(k',j') \\ & \times \cos(2\theta' - 2\delta') + V_{24}'^{(2)}(k',j')\cos(2\theta' + 4\delta') + V_{2-4}'^{(2)}(k',j')\cos(2\theta' - 4\delta') \\ & + V_{42}'^{(2)}(k',j')\cos(4\theta' + 2\delta' + \psi) + V_{4-2}'^{(2)}(k',j')\cos(4\theta' - 2\delta' + \psi) + V_{44}'^{(2)}(k',j') \\ & \times \cos(4\theta' + 4\delta') + V_{4-4}'^{(2)}(k',j')\cos(4\theta' - 4\delta')] + O(\epsilon^3). \end{aligned} \tag{3.7}$$

Since the unperturbed Hamiltonian is the same as before, the condition for a resonance at $h_0 \approx 0$ is still $n_0 > m_0 > 0$, and the only term in (3.7) satisfying this condition has

$$n_0 = 4, \quad m_0 = 2,$$

in which case

$$\mu^2 k_0 = 2\sqrt{1-\mu^2 j_0}\,(2\sqrt{1-\mu^2 j_0} - 1) \tag{3.8}$$

is positive if $\mu^2 j_0 < 3/4$. We thus choose to study the motion near the resonant torus with action variables (j_0, k_0) characterized by the integer numbers $(n_0, m_0) = (4,2)$, and for which the perturbation coefficient is

$$V'^{(2)}_{42}(k', j') = V^{(2)}_{42}(k', j') = \frac{\sqrt{5}}{16}\, k'^2 j'. \tag{3.9}$$

[It is not necessary for us to write down explicitly all the functions $V'^{(2)}_{mn}(k', j')$ here, but some of them coincide with the $V^{(2)}_{mn}(k', j')$ in (2.19).] As far as this local dynamics is concerned, then, we could have used directly the Hamiltonian $H_0(k, j)$ with perturbation $\delta H^{(2)}(\theta, \delta; k, j)$ as in (2.19), without worrying about the presence of the ϵ term. From now on, we will drop the primes on the variables just introduced.

Since the perturbation with coefficient (3.9) is resonant, it does change the local dynamics qualitatively near the chosen torus. Following the standard procedure for such a case (see, e.g., Ref. 9), we go over to a set of canonical variables $(\gamma, \delta; K, J)$ adapted to the resonant torus, chosen so that one of the momenta will still be a constant of the motion under the resonant term in the perturbation:

$$\begin{gathered} K := \frac{k - k_0}{n_0} = \frac{1}{4}\,(k - k_0), \quad \gamma := n_0\theta + m_0\delta + \psi_0 = 4\theta + 2\delta + \psi_0, \\ J := -\frac{m_0}{n_0}\, k + j = -\frac{1}{2}\, k + j, \end{gathered} \tag{3.10}$$

where ψ_0 is some arbitrary, fixed angle.

If we suppose that $K \ll 1$ is a small increment of the variable k around the resonant value k_0, then we can expand the Hamiltonian, written in the new variables, in powers of K, and study the dynamics generated by the leading terms. We begin with the resonant part of the perturbation,

$$\begin{aligned} \delta H^{(2)} &= \epsilon^2 V^{(2)}_{42}(k, j)\cos(4\theta + 2\delta + \psi) + (\text{terms with different } n \text{ and } m) \\ &= \frac{\sqrt{5}}{16}\, \epsilon^2 k_0^2 j\, \cos(\gamma + \psi - \psi_0) + (\text{terms with different } n \text{ and } m). \end{aligned} \tag{3.11}$$

Here j is to be thought of as $j(K = 0, J)$, with J arbitrary. Since this term in the perturbation depends only on γ and not on δ, J is still a constant of the motion; for simplicity we fix its value at the resonant one, J_0. Then for the unperturbed Hamiltonian we obtain

$$H_0(K, J_0) = H_0(k_0, j_0) + \tfrac{1}{2}\,\Omega K^2 + O(K^3), \tag{3.12}$$

where

$$\Omega := \frac{\partial^2 H_0}{\partial k^2}\, n_0^2 + 2\,\frac{\partial^2 H_0}{\partial j\,\partial k}\, n_0 m_0 + \frac{\partial^2 H_0}{\partial j^2}\, m_0^2 = \frac{8\mu^2}{\sqrt{1-\mu^2 j_0}} + \frac{\mu^4 k_0}{(1-\mu^2 j_0)^{3/2}}. \tag{3.13}$$

So, to the lowest order in K and ϵ, the K dynamics near the resonant torus is generated by

$$H_{\rm loc}(\gamma,K)=H_0(k_0,j_0)+\frac{1}{2}\,\Omega K^2+\frac{\sqrt{5}}{16}\,\epsilon^2 k_0^2 j_0\,\cos(\gamma+\psi-\psi_0), \tag{3.14}$$

which is the Hamiltonian of a well-known system, the nonlinear pendulum.

Let us now find the homoclinic orbits. If we call $h_0=H_0(k_0,j_0)$, as before, and fix some value $H_{\rm loc}=h_{\rm loc}$ for the local Hamiltonian, we can compute from Eq. (3.14)

$$K=\pm\left\{\frac{2}{\Omega}\left[h_{\rm loc}-h_0-\frac{\sqrt{5}}{16}\,\epsilon^2 k_0^2 j_0\,\cos(\gamma+\psi-\psi_0)\right]\right\}^{1/2}.$$

To simplify calculations, we set $\psi_0=\psi+\pi$. Then, for the homoclinic orbit, at the maximum $\gamma=\pi$ of the potential we must have $K=0$, from which we find that $h_{\rm loc}-h_0=(\sqrt{5}/16)\epsilon^2 k_0^2 j_0$, and

$$K=\pm\left\{\frac{\sqrt{5}}{8}\,\Omega^{-1}\epsilon^2 k_0^2 j_0(1+\cos\,\gamma)\right\}^{1/2}=\pm\left(\frac{\sqrt{5}}{4}\,\Omega^{-1}\epsilon^2 k_0^2 j_0\right)^{1/2}\cos\frac{\gamma}{2}. \tag{3.15}$$

Finally, we can compute the evolution of γ by integrating the Hamilton equation $\dot{\gamma}=\partial H_{\rm loc}/\partial K=\Omega K$, and we obtain

$$\gamma(t)=4\,\arctan\,\exp\left\{\left(\frac{\sqrt{5}}{16}\,\epsilon^2\Omega k_0^2 j_0\right)^{1/2}t\right\}-\pi. \tag{3.16}$$

We will use all of these results in the next section.

IV. THE MELNIKOV METHOD

We are now going to study the effect of the other terms of the perturbation

$$\begin{aligned}\delta H'^{(2)}(\gamma,\delta;K,J)&=\epsilon^2\sum_{nm}\,V'^{(2)}_{nm}(k,j)\cos(n\theta+m\delta+\psi_{nm})\\&=\epsilon^2\sum_{nm}\,V'^{(2)}_{nm}(k,j)\cos\left(\frac{n}{4}\,\gamma+\frac{2m-n}{2}\,\delta-\frac{n}{4}\,\psi_0+\psi_{nm}\right),\end{aligned} \tag{4.1}$$

in particular of the terms with $(n,m)\neq(n_0,m_0)$, on the dynamics of the integrable system with Hamiltonian $H_{\rm loc}$, near the homoclinic orbit. We will use the Melnikov method to show that a stochastic layer forms in the vicinity of this destroyed separatrix, which acts as a seed for chaos.

We thus want to show[13] that the Melnikov function of the local K dynamics has a contribution

$$M_{nm}(\psi_{nm}):=\int_{\rm ho}dt\{H_{\rm loc},\delta H'^{(2)}_{nm}\} \tag{4.2}$$

which is an oscillating function of ψ_{nm} and has therefore isolated, transverse zeroes, for some values of (n,m); here the integral in (4.2) is taken over the homoclinic orbit, given by $\gamma(t)$ in (3.16), $\delta(t)=\omega_j t+O(\epsilon^2)$ from $\dot{\delta}=\partial H/\partial j'$, $k=k_0+4K=k_0+O(\epsilon)$, and $j=j_0+O(\epsilon)$.

In the calculation of the Melnikov integral we will keep only the leading order terms in ϵ and K. Then, approximating the Poisson bracket

$$\begin{aligned}\{H_{\rm loc},\delta H'^{(2)}_{nm}\}&\approx\left\{\frac{1}{2}\,\Omega K^2,\epsilon^2 V'^{(2)}_{nm}\,\cos\left(\frac{n}{4}\,\gamma+\frac{2m-n}{2}\,\delta-\frac{n}{4}\,\psi_0+\psi_{nm}\right)\right\}\\&=\frac{1}{4}\,\epsilon^2\Omega K n V'^{(2)}_{nm}(k_0,j_0)\sin\left(\frac{n}{4}\,\gamma+\frac{2m-n}{2}\,\delta-\frac{n}{4}\,\psi_0+\psi_{nm}\right),\end{aligned} \tag{4.3}$$

and using $\dot{\gamma}=\Omega K$, we get

$$M_{nm}(\psi_{nm}) = \int_{-\pi}^{\pi} d\gamma \dot{\gamma}^{-1} \{H_{\rm loc}, \delta H'^{(2)}_{nm}\} = \frac{1}{4}\,\epsilon^2 n V'^{(2)}_{nm}(k_0, j_0)(A_{nm}\,\sin\,\psi'_{nm} + B_{nm}\,\cos\,\psi'_{nm}), \tag{4.4}$$

where we have defined $\psi'_{nm} := \psi_{nm} - (n/4)\psi_0$, and

$$A_{nm} := \int_{-\pi}^{\pi} d\gamma\,\cos\left(\frac{n}{4}\,\gamma + \frac{2m-n}{2}\,\delta(\gamma)\right)$$

$$B_{nm} := \int_{-\pi}^{\pi} d\gamma\,\sin\left(\frac{n}{4}\,\gamma + \frac{2m-n}{2}\,\delta(\gamma)\right), \tag{4.5}$$

and $\delta(\gamma)$ is obtained from (3.16) and $\delta \approx \omega_j t$,

$$\delta(\gamma) = \eta\,\ln\,\tan\frac{\gamma+\pi}{4}, \quad \eta := \omega_j\left(\frac{\sqrt{5}}{16}\,\epsilon^2 \Omega k_0^2 j_0\right)^{-1/2}. \tag{4.6}$$

The function M_{nm} in (4.4) is clearly oscillating whenever $nV'^{(2)}_{nm}A_{nm}$ and/or $nV'^{(2)}_{nm}B_{nm}$ are not zero. We thus have to show that, for some (n,m) in the set $\{(2,0), (2,\pm 2), (2,\pm 4), (4,0), (4,-2), (4,\pm 4)\}$, A_{nm} or B_{nm} do not vanish.

Perhaps surprisingly, the integrals in (4.5) can be calculated analytically, at least for half of the cases of interest, the ones with $n=2$. To do this, notice that A_{nm} and B_{nm} are, respectively, the real and imaginary parts of the complex function

$$Z_{nm} := \int_{-\pi}^{\pi} d\gamma e^{{\rm i}n\gamma/4}\left(\tan\frac{\gamma+\pi}{4}\right)^{{\rm i}(2m-n)\eta/2} = 4e^{-{\rm i}n\pi/4}\int_0^{\pi/2} dx e^{{\rm i}nx}(\tan x)^{{\rm i}(2m-n)\eta/2}, \tag{4.7}$$

where we have defined $x := (1/4)(\gamma+\pi)$. Therefore for $n=2$, we have[17]

$$\begin{aligned} Z_{2m} &= -4{\rm i}\left[\int_0^{\pi/2} dx\,\cos 2x\,\tan^{{\rm i}(m-1)\eta} x + {\rm i}\int_0^{\pi/2} dx\,\sin 2x\,\tan^{{\rm i}(m-1)\eta} x\right] \\ &= -4{\rm i}\left[-\frac{{\rm i}(m-1)\eta\pi}{2}\,\sec\frac{{\rm i}(m-1)\eta\pi}{2} + \frac{{\rm i}(m-1)\eta\pi}{2}\,{\rm cosec}\,\frac{{\rm i}(m-1)\eta\pi}{2}\right] \\ &= -2(m-1)\eta\pi\left[{\rm sech}\,\frac{(m-1)\eta\pi}{2} + {\rm i}\,{\rm cosech}\,\frac{(m-1)\eta\pi}{2}\right], \end{aligned} \tag{4.8}$$

in other words,

$$A_{2m} = -2(m-1)\eta\pi\,{\rm sech}\,\frac{(m-1)\eta\pi}{2}, \quad B_{2m} = -2(m-1)\eta\pi\,{\rm cosech}\,\frac{(m-1)\eta\pi}{2}, \tag{4.9}$$

which do not vanish for any of the values of m listed above.

V. CONCLUSIONS

We have shown analytically that a RW universe, filled with a massive conformally coupled scalar field, has a chaotic behavior, at least for a sufficiently small value of the mass, confirming the numerical results of Ref. 1. This is just the first step in our program to analyze the dynamics of RW models with various types of matter content, in order to relate this to numerical results, and make more precise statements about physical consequences. Among the purely theoretical aspects of this work, on the one hand, some possible generalizations to cosmological models differing by the addition of a cosmological constant and/or other parameters are being currently studied,[18] as well as the possibility of adding a second matter field; on the other hand, we are working on extending the results obtained for the model in this paper to an estimation of the size of the stochastic region in phase space and of the time scale for the manifestation of the chaos we

predict. The latter point is an important one since, as we mentioned in Sec. II, this time scale must be smaller than the lifetime of the universe in our model. Although the simulations in Ref. 1 covered many cycles for the universe, it was stated there that chaos would show up on much smaller time scales, and therefore has, in principle, observable consequences; in Ref. 7, however, the authors argued qualitatively that they expect the time scale to be much larger than the lifetime of the universe, and used this to motivate the use of two different scalar fields in their model. We believe that the issue should be resolved by an actual, quantitative estimate of the time scale.

We will conclude with some comments on possible avenues for future research, concerning other aspects of the relationship between chaos and the physics of the early universe. At a classical level already, chaos opens up ways of explaining the origin of the arrow of time,[19] and in our cosmological setting therefore, of improving our knowledge of cosmological statistical mechanics and thermodynamics. Furthermore, just as nonintegrability and dynamical instability are, at the classical level, the causes of chaos, they are also, at the quantum level, the causes of the instability of quantum states and particles; one therefore expects, for example, classical chaos and semiclassical particle production to be related.[20] At the semiclassical level,[21] one can prove the presence of decoherence and correlations, which cause the transition to the classical regime; stable and unstable scalar field quanta are created during this process. The study of this passage, that leads from particle creation to chaos is, therefore, imperative. In a model very similar to ours,[22] a Lyapunov variable was used to define a growing entropy in the universe; this could be adapted, with minor changes, to the present model, and looks promising since, for example, growth of entropy, particle creation, decoherence and isotropization (a tendency to some kind of equilibrium) are related in semiclassical cosmological models.[23]

The relevance of chaos in the transition to the semiclassical limit of quantum cosmology, for RW models, was studied in Ref. 24. At the complete quantum level, let us recall only that RW models have been used to study the relation between the cosmological and thermodynamical arrows of time;[25] but many other results related to these subjects can be found in the literature. Ultimately, a very interesting goal for future work, and possibly a formidable task even in our simple RW model with chaotic behavior, is to relate all these fundamental features of the universe in a unified theoretical framework.

ACKNOWLEDGMENTS

Part of this work was done when the authors were in the RGGR group at the Université Libre de Bruxelles in Brussels, Belgium, where it was partially supported by the Commission of the European Communities, under Contract No. ECRU002 (DG-III); we are grateful to Edgard Gunzig for making this possible, and for the working atmosphere in the group. F.L. would also like to thank the University of Buenos Aires and the Fundación Antorchas for financial support.

[1] E. Calzetta and C. El Hasi, ''Chaotic Friedmann-Robertson-Walker cosmology,'' Class. Quantum Grav. **10**, 1825–1841 (1993), and gr-qc/9211027; E. Calzetta, ''Homoclinic chaos in relativistic cosmology,'' in *Deterministic Chaos in General Relativity*, edited by D. Hobill, A. Burd, and A. Coley (Plenum, New York, 1994).

[2] V. A. Belinskii, Z. M. Khalatnikov, and E. M. Lifschitz, ''Oscillatory approach to a singular point in the relativistic cosmology,'' Adv. Phys. **19**, 525–573 (1970).

[3] C. W. Misner, ''Mixmaster universe,'' Phys. Rev. Lett. **22**, 1071–1074 (1969).

[4] J. Pullin, ''Time and chaos in general relativity,'' in *Relativity and Gravitation: Classical and Quantum*, edited by J. C. D'Olivo *et al.* (World Scientific, Singapore, 1991); A. Burd and R. Tavakol, ''Invariant Lyapunov exponents and chaos in cosmology,'' Phys. Rev. D **47**, 5336–5341 (1993); S. E. Rugh, ''Chaos in the Einstein equations—Characterization and importance,'' in *Deterministic Chaos in General Relativity*, edited by D. Hobill, A. Burd, and A. Coley (Plenum, New York, 1994); B. K. Berger, D. Garfinkle, and E. Strasser, ''New algorithm for mixmaster dynamics,'' Class. Quantum Grav. **14**, L29–L36 (1997), and gr-qc/9609072.

[5] M. Berry ''Semiclassical mechanics of regular and irregular motion,'' in *Chaotic Behaviour of Deterministic Systems*, edited by G. Iooss, R. H. G. Helleman, and R. Stora (North-Holland, New York, 1983).

[6] E. Calzetta and C. El Hasi, ''Nontrivial dynamics in the early stages of inflation,'' Phys. Rev. D **51**, 2713–2728 (1995), and gr-qc/9408010.

[7] N. J. Cornish and J. J. Levin, ''Chaos, fractals and inflation,'' Phys. Rev. D **53**, 3022–3032 (1996), and astro-ph/9510010.

[8] F. Finelli, G. P. Vacca and G. Venturi, ''Chaotic inflation with a scalar field in nonclassical states,'' gr-qc/9712098.

[9] G. M. Zaslavsky, R. Z. Sagdeev, D. A. Usikov, and A. A. Chernikov, *Weak Chaos and Quasi-Regular Patterns* (Cambridge University Press, Cambridge, England, 1991).

[10] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer, New York, 1990).

[11] A. M. Ozorio de Almeida, *Hamiltonian Systems: Chaos and Quantization* (Cambridge University Press, Cambridge, England, 1988).
[12] L. Bombelli and E. Calzetta, ''Chaos around a black hole,'' Class. Quantum Grav. **9**, 2573–2599 (1992); L. Bombelli, ''Particle motion around perturbed black holes: The onset of chaos,'' in *Deterministic Chaos in General Relativity*, edited by D. Hobill, A. Burd, and A. Coley (Plenum, New York, 1994).
[13] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer, Berlin, 1983).
[14] A. Helmi and H. Vucetich, ''Non-integrability and chaos in classical cosmology,'' Phys. Lett. A **230**, 153–156 (1997), and gr-qc/9705009.
[15] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[16] G. Gallavotti, *The Elements of Mechanics* (Springer, New York, 1983), Sec. 5.10.
[17] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products* (Academic, New York, 1980).
[18] L. Bombelli and K. Rosquist, ''Families of chaotic cosmologies'' in preparation.
[19] E. Calzetta, ''A necessary and sufficient condition for convergence to equilibrium in Kolmogorov systems,'' J. Math. Phys. **32**, 2903–2906 (1991).
[20] W. H. Zurek and J. P. Paz, ''Decoherence, chaos, and the second law,'' Phys. Rev. Lett. **72**, 2508–2511 (1994), and gr-qc/9402006.
[21] M. Castagnino, E. Gunzig, and F. Lombardo, ''Time asymmetry in semiclassical cosmology,'' Gen. Relativ. Gravit. **27**, 257–266 (1995); M. Castagnino and F. Lombardo, ''Decoherence correlations, and unstable quantum states in semiclassical cosmology,'' *ibid.* **28**, 263 (1996), and gr-qc/9404028; F. Lombardo and F. Mazzitelli, ''Einstein-Langevin equations from running coupling constants,'' Phys. Rev. D **55**, 3889–3892 (1997), and gr-qc/9609073.
[22] E. Calzetta, M. Castagnino, and R. Scoccimarro, ''Coarse-graining approach to quantum cosmology,'' Phys. Rev. D **45**, 2806–2813 (1992).
[23] E. Calzetta, ''Particle creation, inflation, and cosmic isotropy,'' Phys. Rev. D **44**, 3043–3051 (1991); E. Calzetta and F. D. Mazzitelli, ''Decoherence and particle creation,'' *ibid.* **42**, 4066–4069 (1990).
[24] E. Calzetta and J. J. González, ''Chaos and semiclassical limit in quantum cosmology,'' Phys. Rev. D **51**, 6821–6828 (1995), and gr-qc/9411045.
[25] S. W. Hawking, ''Arrow of time in cosmology,'' Phys. Rev. D **32**, 2489–2495 (1985); D. N. Page, ''Will entropy decrease if the universe recollapses?,'' *ibid.* **32**, 2496–2499 (1985).

Colliding plane waves in terms of Jacobi functions

Nora Bretón[a)]
Departamento de Física, CINVESTAV-IPN, Apartado Postal 14-740, C.P. 07000, México, D.F., México

Alberto García[b)]
Departamento de Física, CINVESTAV-IPN, Apartado Postal 14-740, C.P. 07000, México, D.F., México and Departamento de Física, Universidad de Santiago de Chile, Avda. Ecuador 3493– Casilla 307.Correo 2, Santiago, Chile

Alfredo Macías[c)]
Departamento de Física, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, C.P. 09340, México, D.F., México

Gustavo Yáñez[d)]
Departamento de Física, CINVESTAV-IPN, Apartado Postal 14-740, C.P. 07000, México, D.F., México

(Received 23 April 1998; accepted for publication 17 June 1998)

We present a general class of noncollinear colliding wave solutions of the Einstein–Maxwell equations given in terms of fourth order polynomials, which in turn can be expressed through Jacobi functions depending on generalized advanced and retarded time coordinates. The solutions are characterized by six free parameters. The parameters can be chosen in such a way as to avoid the generic focusing singularity. © *1998 American Institute of Physics.* [S0022-2488(98)02611-5]

I. INTRODUCTION

Recently, the collision of plane-fronted gravitational waves possibly coupled with electromagnetic waves, has been extensively studied.[1–4] Because gravity is always attractive, it was expected that focusing of the waves would occur and one of the interesting questions is: how much focusing does general relativity predict? Within this framework, strong focusing would appear by the development of space–time curvature singularities. Many solutions have been presented so far, describing collisions of plane-fronted gravitational and electromagnetic waves. And quite a few of them do develop Cauchy horizons. It is important to stress the fact that all of these solutions contain *only second degree* polynomials and that until now no one has gone ahead to *higher degree polynomials* in the sense we are going to explain below.

All known colliding wave solutions are characterized, in null coordinates, by the line element

$$ds^2 = e^{-M} du dv + g_{ab} dx^a dx^b, \quad a,b = 1,2, \tag{1.1}$$

where x^1, x^2 are ignorable coordinates and metric functions depending on (u,v).[5] When written in (η,μ) coordinates,[3] where η is a measure of the time from the instant of collision and μ is a measure of the distance between the plane wavefronts, the metric becomes

$$ds^2 = e^{2\gamma(\eta^2-\mu^2)}\left(\frac{d\eta^2}{P(\eta)} - \frac{d\mu^2}{Q(\mu)}\right) + g_{ab} dx^a dx^b, \tag{1.2}$$

where the polynomials $P(\eta)$, $Q(\mu)$ are at most of the second degree in their respective variables. Generating techniques and other methods have allowed a great deal of generalizations in relation

a)Electronic mail: nora@fis.cinvestav.mx
b)Electronic mail: agarciad@lauca.usach.cl
c)Electronic mail: amac@xanum.uam.mx
d)Electronic mail: gyanez@fis.cinvestav.mx

 © 1998 American Institute of Physics

with the metric components g_{ab} (sometimes called Killing sector).[6–13] The generated solutions preserve the seed's structure of the polynomials $P(\eta)$, $Q(\mu)$, changing in general the form of the metric function γ.

In this paper, we further pursue this and present solutions with $P(\eta)$ and $Q(\mu)$ being polynomials of fourth degree, by means of a generalization of the advanced and retarded coordinates concept. This generalization endows the corresponding space–times with a geometric structure for the time and independently for the spatial coordinate richer than the already known colliding wave solutions. In the limit of polynomials of second degree one recovers the widely used retarded and advanced time coordinates (u,v). The main purpose of this paper is to give the formulation of the most general class presently known of colliding waves with fourth degree polynomials. As usual, it is assumed that in the corresponding space–time, the two waves approach each other from opposite sides in flat Minkowski background; after the collision, a new gravitational field evolves, which satisfies certain continuity conditions. The colliding plane waves posses five symmetries, while the geometry resulting after the collision has two spacelike Killing vectors.

The plan of the paper is as follows: In Sec. II we revisited some generalities on colliding waves concept. In Sec. III the general representation through advanced and retarded time coordinates for fourth degree polynomials is introduced. In Sec. IV the class of colliding waves with fourth degree polynomials is presented. In Sec. V the results are discussed.

II. COLLIDING WAVES

As was stated previously, in this section we review briefly some generalities on electrovacuum colliding waves, introduced by Ernst *et al.*[8]

The set of colliding waves solutions is described by the line element

$$g=2g(u,v)dudv+g_{ab}(u,v)dx^a dx^b, \quad a,b=1,2, \tag{2.1}$$

where $x^1=x$, $x^2=y$ are ignorable coordinates. The domain of the coordinate charts consists of $(x,y)\in\mathbf{R}^2$ and $(u,v)\in\mathbf{R}^2$; it is the union of four continuous regions: $\mathrm{I}:=\{(u,v):0\leqslant u<1,0\leqslant v<1\}$, $\mathrm{II}:=\{(u,v):u\leqslant 0,0\leqslant v<1\}$, $\mathrm{III}:=\{(u,v):0\leqslant u<1,v<0\}$, $\mathrm{IV}:=\{(u,v):u\leqslant 0,v\leqslant 0\}$.

In region IV, a closed subregion of the Minkowski space, it is required that

$$g_{\mu\nu}(u,v)=g_{\mu\nu}(0,0), \quad A_\mu(u,v)=A_\mu(0,0), \tag{2.2}$$

where A_μ is the electromagnetic vector potential. By scaling–shifting transformations the metric can be brought to standard Minkowski metric. In region II, the metric components and the electromagnetic vector potential depend only on v, i.e., $g_{\mu\nu}=g_{\mu\nu}(0,v)$, $A_\mu=A_\mu(0,v)$. In region III these fields are functions of the coordinate u, i.e., $g_{\mu\nu}=g_{\mu\nu}(u,0)$, $A_\mu=A_\mu(u,0)$. In region I, which is occupied by the scattered null fields, the metric components and the electromagnetic field are functions of both u and v coordinates. In this way, the problem is reduced to know the metric (coframe) and the electromagnetic vector function A_μ.

A. Newman–Penrose tetrad

We introduce a complex null frame according to the conventional description:

$$e_\alpha=(\mathbf{l},\mathbf{n},\mathbf{m},\overline{\mathbf{m}}) \tag{2.3}$$

with

$$\mathbf{l}^2=\mathbf{n}^2=\mathbf{m}^2=\overline{\mathbf{m}}^2=\mathbf{l}\cdot\mathbf{m}=\mathbf{n}\cdot\mathbf{m}=\mathbf{l}\cdot\overline{\mathbf{m}}=\mathbf{n}\cdot\overline{\mathbf{m}}=0, \tag{2.4}$$

$$\mathbf{l}\cdot\mathbf{n}=1, \quad \mathbf{m}\cdot\overline{\mathbf{m}}=-1. \tag{2.5}$$

The basis of the corresponding 1-forms, the coframe, is denoted by ϑ^α. If we lower the coframe index, then we use the same letters as with the frame: $\vartheta_\alpha=(\mathbf{l},\mathbf{n},\mathbf{m},\overline{\mathbf{m}})$. Then the coframe reads

$$\vartheta^\alpha=(\mathbf{n},\mathbf{l},-\overline{\mathbf{m}},-\mathbf{m}). \tag{2.6}$$

According to (2.4) and (2.5), we have

$$g=2(\hat{\vartheta^0}\otimes\hat{\vartheta^1}-\hat{\vartheta^2}\otimes\hat{\vartheta^3})=n_{\alpha\beta}\vartheta^\alpha\otimes\vartheta^\beta, \tag{2.7}$$

with

$$n_{\alpha\beta}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&-1\\0&0&-1&0\end{pmatrix}. \tag{2.8}$$

B. Self-dual part of the complex Weyl 2-form

Quite independently from the type of coframe we are using, we can consider some properties of the Weyl 2-form. Let us denote the Weyl curvature 2-form by

$$C_{\alpha\beta}=\tfrac{1}{2}C_{\mu\nu\alpha\beta}\vartheta^\mu\wedge\vartheta^\nu. \tag{2.9}$$

If, in four dimensions, an arbitrary complex 2-form ω is given, then its Hodge dual ${}^\star\omega$ is again a 2-form. Therefore we can build linear combinations of both:

$$\omega^\pm:=\tfrac{1}{2}(\omega\pm i{}^\star\omega). \tag{2.10}$$

Here i is the imaginary unit. We call ω^+ the self-dual and ω^- the antiself-dual part of the 2-form ω. Clearly,

$$\omega=\omega^+ +\omega^-. \tag{2.11}$$

A form ω is called self-dual, if its antiself-dual piece vanishes, i.e., if $\omega=i{}^\star\omega$, it is called antiself-dual, if $\omega=-i{}^\star\omega$.

Let us now assume that the form ω is real, that is, $\omega=\overline{\omega}$, where the bar denotes complex conjugation. Then, according to (2.10), they fulfill the relations:

$$\overline{\omega^+}=\omega^-, \quad \text{and} \quad \overline{\omega^-}=\omega^+. \tag{2.12}$$

Consequently, a real 2-form $\omega=\overline{\omega}$ can be alternatively encoded into its self-dual part ω^+. Incidentally, the same is true for its antiself-dual part.

The Weyl 2-form is a real quantity. Accordingly, we can also take its self-dual part

$$C^+_{\alpha\beta}:=\tfrac{1}{2}(C_{\alpha\beta}+i{}^\star C_{\alpha\beta}). \tag{2.13}$$

Expanding the Weyl 2-form $C^+_{\alpha\beta}$ in terms of the basis $\vartheta^\mu\wedge\vartheta^\nu$ we have

$$C^+_{\alpha\beta}=\tfrac{1}{2}C^+_{\mu\nu\alpha\beta}\vartheta^\mu\wedge\vartheta^\nu, \tag{2.14}$$

which we will assume to be in the standard 2-form self-dual basis [**U,V,W**]. The most general form of our Weyl field components, compatible with colliding wave space–time structure is given by[14]

$$\begin{aligned}
&A_\mu=A_\mu(u,v),\ \ C^+_{abcd}=2\Psi_0U_{ab}U_{cd}+2\Psi_2(U_{ab}V_{cd}+V_{ab}U_{cd}+W_{ab}W_{cd})\\
&\qquad\qquad +2\Psi_4V_{ab}V_{cd},\ \text{region I},\\
&A_\mu=A_\mu(v),\quad C^+_{abcd}=2\Psi_0U_{ab}U_{cd},\quad \text{region II},\\
&A_\mu=A_\mu(u),\quad C^+_{abcd}=2\Psi_4V_{ab}V_{cd},\quad \text{region III},
\end{aligned} \tag{2.15}$$

with

$$W_{ab}=m_a\tilde{m}_b-m_b\tilde{m}_a-k_al_b+k_bl_a,$$

$$V_{ab}=k_a m_b-k_b m_a, \quad U_{ab}=-l_a\tilde{m}_b+l_b\tilde{m}_a \tag{2.16}$$

where m_a, $\tilde{m}_b$, k_a and l_a are null tetrads.

III. GENERALIZED ADVANCED AND RETARDED TIME COORDINATES

We shall consider a metric of the form

$$g=e^{-2\gamma}\left(\frac{dp^2}{P(p)}-\frac{dq^2}{Q(q)}\right)+g_{ab}dx^a dx^b, \tag{3.1}$$

where P,Q are fourth degree polynomials in their respective variable, and γ depending on both (p,q) coordinates.

Since we have real polynomials of fourth degree, when looking for advanced and retarded time variables one has to deal with elliptic integrals, i.e., we shall use the Legendre first kind integrals:

$$\int \frac{dr}{\sqrt{G_4(r)}}=\mu\int \frac{d\phi}{\sqrt{1-k^2\sin^2\phi}}=\mu F(\phi,k), \tag{3.2}$$

where the fourth degree polynomial $G_4(r)$, is represented by the product of four monomials containing the corresponding roots, the coefficient of the higher degree (fourth), denoted by a_4, is assumed to be $a_4=\pm 1$; if this were not the case, i.e., $a_4=\pm\alpha_4$, $\alpha_4>0$, then dividing by α_4 one arrives at the above equation where now μ has to be interpreted as $\mu/\sqrt{\alpha_4}$; in what follows we shall adopt these conventions when needed. The real roots are denoted by r_j, with $j=1,...,4$ and $r_1>r_2>r_3>r_4$, while the complex roots by $s_1\pm it_1$, $s_2\pm it_2$, with $s_1\geqslant s_2$, $t_1>t_2>0$. We also introduce the following useful notation

$$\begin{gathered}
r_{ik}=r_k-r_i, \quad (i,k=1,2,3,4),\\
(r,\beta,\gamma,\delta)=\frac{r-\gamma}{r-\delta}\frac{\beta-\delta}{\beta-\gamma},\\
\tan\theta_1=\frac{r_1-s_1}{t_1}, \quad \tan\theta_2=\frac{r_2-s_1}{t_1},\\
\tan\theta_3=\frac{t_1+t_2}{s_1-s_2}, \quad \tan\theta_4=\frac{t_1-t_2}{s_1-s_2},\\
\tan[(\theta_5/2)^2]=\frac{\cos\theta_3}{\cos\theta_4},\\
\nu=\tan[(\theta_2-\theta_1)/2]\tan[(\theta_2+\theta_1)/2].
\end{gathered} \tag{3.3}$$

The elliptic integral $F(\phi,k)$ is the Legendre integral of the first kind. As is well known, the standard form for writing this function is

$$z=\int^{\phi}\frac{d\phi}{\sqrt{1-k^2\sin^2\phi}}=F(\phi,k), \quad \phi=\text{am } z, \tag{3.4}$$

where am z denotes the function amplitude of z. Replacing ϕ through ω according to

$$\omega=\sin\phi=\sin(\text{am } z):=\text{sn } z, \quad z=\int_0^{\omega}\frac{d\omega}{\sqrt{(1-\omega^2)(1-k^2\omega^2)}}=\tilde{F}(\omega,k), \tag{3.5}$$

where sn z belongs to the Jacobi family of elliptic functions (sn z, cn z, dn z), with well established analytical properties. The values and main properties of these functions can be found in Ref. 15.

Going back to the line element (3.1) and assuming that $P(p)$ and $Q(q)$ are polynomials up to fourth degree on p and q, respectively, then the two dimensional line element [first two terms in (3.1)] can be written in terms of the retarded and advanced time coordinates $\tilde{u}$ and $\tilde{v}$, respectively, as follows

$$e^{-2\gamma(p,q)}\left(\frac{dp^2}{P(p)}-\frac{dq^2}{Q(q)}\right)=e^{-\gamma(\tilde{u},\tilde{v})}d\tilde{u}d\tilde{v}, \tag{3.6}$$

with

$$\left.\begin{matrix}d\tilde{u}\\ d\tilde{v}\end{matrix}\right\}=\frac{dp}{\sqrt{P_4(p)}}\pm\frac{dq}{\sqrt{Q_4(q)}}=\frac{\mu_\phi}{\sqrt{\alpha_4}}\frac{d\phi}{\sqrt{1-k_\phi^2\sin^2\phi}}\pm\frac{\mu_\theta}{\sqrt{\beta_4}}\frac{d\theta}{\sqrt{1-k_\theta^2\sin^2\theta}}, \tag{3.7}$$

where μ_ϕ, μ_θ are constants and α_4, β_4 are the moduli of the coefficients of the higher power in the polynomials P and Q, correspondingly. The variables $\tilde{u}$ and $\tilde{v}$ range the values from $-\infty$ to $+\infty$.

The integration of (3.7) yields

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\frac{\mu_\phi}{\sqrt{\alpha_4}}F(\phi,k_\phi)\pm\frac{\mu_\theta}{\sqrt{\beta_4}}F(\theta,k_\theta), \tag{3.8}$$

where $-\infty<\tilde{u}<\infty$, $-\infty<\tilde{v}<\infty$. Hence, one has relations between the Legendre elliptic integrals and the null coordinates $\tilde{u}$, $\tilde{v}$, i.e.,

$$F(\phi,k_\phi)=\frac{\sqrt{\alpha_4}}{2\mu_\phi}(\tilde{u}+\tilde{v}),\quad F(\theta,k_\theta)=\frac{\sqrt{\beta_4}}{2\mu_\theta}(\tilde{u}-\tilde{v}). \tag{3.9}$$

Consequently, the inversion formulas bring the following functional dependence upon the generalized advanced and retarded time coordinates

$$\phi=\mathrm{am}\left[\frac{\sqrt{\alpha_4}}{2\mu_\phi}(\tilde{u}+\tilde{v})\right],\quad \omega=\sin\phi=\mathrm{sn}\left[\frac{\sqrt{\alpha_4}}{2\mu_\phi}(\tilde{u}+\tilde{v})\right], \tag{3.10}$$

$$\theta=\mathrm{am}\left[\frac{\sqrt{\beta_4}}{2\mu_\theta}(\tilde{u}-\tilde{v})\right],\quad \Omega=\sin\theta=\mathrm{sn}\left[\frac{\sqrt{\beta_4}}{2\mu_\theta}(\tilde{u}-\tilde{v})\right], \tag{3.11}$$

which allow us to give the original p and q coordinates, initially expressed through ϕ and θ, in terms of $\tilde{u}$ and $\tilde{v}$ null coordinates, i.e.,

$$p=p(\tilde{u},\tilde{v}),\quad \text{and}\quad q=q(\tilde{u},\tilde{v}). \tag{3.12}$$

Moreover, one can introduce a new set of variables (u,v)

$$u=\sin\left(\frac{\tilde{u}}{2}\right),\quad \text{and}\quad v=\sin\left(\frac{\tilde{v}}{2}\right) \tag{3.13}$$

where the ranges of the variables are $-\infty<\tilde{u}<\infty$, $-\infty<\tilde{v}<\infty$ and $-1<u<1$, $-1<v<1$. These variables are such that

$$d\tilde{u}=2\frac{du}{U},\quad d\tilde{v}=2\frac{dv}{V}, \tag{3.14}$$

and

$$\frac{\tilde{u}\pm\tilde{v}}{2}=\arcsin u\pm\arcsin v=\arcsin[uV\pm vU], \tag{3.15}$$

where $U=\sqrt{1-u^2}$, $V=\sqrt{1-v^2}$ and the corresponding term in the line element

$$d\tilde{u}\otimes d\tilde{v}=4\frac{du}{U}\otimes\frac{dv}{V}. \tag{3.16}$$

Having in mind colliding wave interpretations of the results, one restricts u and v to the range $0\leqslant u<1$, $0\leqslant v<1$ to be closed to the standard colliding wave approach. Depending on the character, real or complex, of the roots of the fourth degree polynomials, one may encounter, in general, combinations of different possible cases. It is always possible to introduce the retarded and advanced time coordinates u and v. However, only certain solutions satisfy the Ernst[8] requirements for colliding waves. We shall restrict ourselves to metrics which can support a colliding wave interpretation.

It may occur that several mixed cases depend on the degree of the polynomials appearing in the non-Killing sector, for instance,

$$\frac{dp^2}{P_4}-\frac{dq^2}{Q_2},\quad \frac{dp^2}{P_3}-\frac{dq^2}{Q_2}, \tag{3.17}$$

where the subindice indicates the degree of the polynomial we are dealing with; the possibility of interchanging P by Q and p by q and vice versa also has to be taken into account. The case $(dp^2/P_3)-(dq^2/Q_3)$ can be treated in the manner we did with the fourth degree polynomial.

Let us consider the case of $(dp^2/P_4)-(dq^2/Q_2)$. Introducing the null variables $(\tilde{u},\tilde{v})$,

$$\left.\begin{matrix}d\tilde{u}\\ d\tilde{v}\end{matrix}\right\}=\frac{dp}{\sqrt{P_4(p)}}\pm\frac{dq}{\sqrt{Q_2(q)}}, \tag{3.18}$$

in the case when Q_2 can be brought to the form $1-q^2$. They become

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\mu_\phi F(\phi,k_\phi)\pm\arcsin\, q, \tag{3.19}$$

therefore

$$\frac{\tilde{u}-\tilde{v}}{2}=\arcsin\, q, \tag{3.20}$$

$$\frac{\tilde{u}+\tilde{v}}{2}=\mu_\phi F(\phi,k_\phi)=\mu_\phi\int^{\phi}\frac{dx}{\sqrt{1-k_\phi^2\sin^2 x}}=\mu_\phi z,\quad \phi=\mathrm{am}\, z, \tag{3.21}$$

hence

$$q=\sin\left(\frac{\tilde{u}-\tilde{v}}{2}\right),\quad p=p\left(\frac{\tilde{u}+\tilde{v}}{2}\right), \tag{3.22}$$

where the relation of p in terms of $\tilde{u}$, $\tilde{v}$ depend on the roots of the polynomial P_4. The other case with P_3, Q_2 can be treated in a similar way.

For completeness, we also briefly comment on the following cases

(i)

$$\frac{dp^2}{P_{4,3}(p)}-\frac{dq^2}{Q_1(q)}\rightarrow\frac{dp^2}{P_{4,3}(p)}-\beta(dq^{1/2})^2,\quad q>0,\beta>0. \tag{3.23}$$

The corresponding $\tilde{u}$, $\tilde{v}$ are

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\mu_\phi F(\phi,k_\phi)\pm\beta q^{1/2}, \tag{3.24}$$

thus

$$p=p\left(\frac{\tilde{u}+\tilde{v}}{2}\right),\quad q=\left(\frac{\tilde{u}-\tilde{v}}{2\beta}\right)^2, \tag{3.25}$$

(ii)

$$\frac{dp^2}{P_2(p)}-\frac{dq^2}{Q_2(q)}\to\frac{dp^2}{1-p^2}-\frac{dq^2}{1-q^2}. \tag{3.26}$$

The corresponding $\tilde{u}$, $\tilde{v}$ are

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\arcsin\, p\pm\arcsin\, q, \tag{3.27}$$

hence

$$p=\sin\left(\frac{\tilde{u}+\tilde{v}}{2}\right),\quad q=\sin\left(\frac{\tilde{u}-\tilde{v}}{2}\right), \tag{3.28}$$

(iii)

$$\frac{dp^2}{P_{4,3}(p)}-dq^2. \tag{3.29}$$

The corresponding $\tilde{u}$, $\tilde{v}$ are

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\mu_\phi F(\phi,k_\phi)\pm q, \tag{3.30}$$

hence

$$p=p\left(\frac{\tilde{u}+\tilde{v}}{2}\right),\quad q=\frac{\tilde{u}-\tilde{v}}{2}, \tag{3.31}$$

(iv)

$$\frac{dp^2}{P_2(p)}-\frac{dq^2}{Q_1}\to\frac{dp^2}{1-p^2}-\beta(dq^{1/2})^2,\quad \beta>0,q>0. \tag{3.32}$$

The corresponding $\tilde{u}$, $\tilde{v}$ are

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=\arcsin\, p\pm\beta q^{1/2}, \tag{3.33}$$

hence

$$p=\sin\left(\frac{\tilde{u}+\tilde{v}}{2}\right),\quad q=\left(\frac{\tilde{u}-\tilde{v}}{2\beta}\right)^2, \tag{3.34}$$

(v)

$$\frac{dp^2}{P_1(p)}-\frac{dq^2}{Q_1(q)}\to(dp^{1/2})^2-(dq^{1/2})^2,\quad p>0,q>0. \tag{3.35}$$

The corresponding $\tilde{u}$, $\tilde{v}$ are

$$\left.\begin{matrix}\tilde{u}\\ \tilde{v}\end{matrix}\right\}=p^{1/2}\pm q^{1/2}, \tag{3.36}$$

thus

$$p=\left(\frac{\tilde{u}+\tilde{v}}{2}\right)^2,\quad q=\left(\frac{\tilde{u}-\tilde{v}}{2}\right)^2, \tag{3.37}$$

A. Different transformations of $p(\tilde{u},\tilde{v})$ and $q(\tilde{u},\tilde{v})$

In this subsection, we present all possible cases depending upon the roots of the polynomials $P(p)$ and $Q(q)$ and the corresponding transformation.

(1) For P and Q with all real and different roots, i.e.,

$$P=\alpha_4\epsilon_{\pm}[p-p_1][p-p_2][p-p_3][p-p_4],\quad \epsilon_{\pm}=\pm 1,\alpha_4>0, \tag{3.38}$$

$$Q=\beta_4\nu_{\pm}[q-q_1][q-q_2][q-q_3][q-q_4],\quad \nu_{\pm}=\pm 1,\beta_4>0. \tag{3.39}$$

The corresponding transformations read as follows

$P_4(p)$	$\epsilon_{\pm}$	Interval	$p(\tilde{u},\tilde{v})=$	k^2
four	ϵ_+	$p_1\leqslant p$ or $p\leqslant p_4$	$\dfrac{p_1p_{42}-p_2p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{42}-p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	
real	ϵ_+	$p_3\leqslant p\leqslant p_2$	$\dfrac{p_3p_{42}-p_4p_{32}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{42}-p_{32}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	$(\alpha_1,\alpha_2,\alpha_4,\alpha_3)$
roots	ϵ_-	$p_4\leqslant p\leqslant p_3$	$\dfrac{p_4p_{31}+p_1p_{43}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{31}+p_{43}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	
	ϵ_-	$p_2\leqslant p\leqslant p_1$	$\dfrac{p_2p_{31}-p_3p_{21}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{31}-p_{21}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	$(\alpha_3,\alpha_2,\alpha_4,\alpha_1)$

In the transformations given above we must understand that $\tilde{u}=2\arcsin u$ and $\tilde{v}=2\arcsin v$, such that $\tilde{u}\pm\tilde{v}=2\arcsin(uV\pm vU)$ and μ_ϕ has to be understood as $\mu_\phi/\sqrt{\alpha_4}$ when $\alpha_4\neq 1$. For $Q(q)$ the same table is valid making the changes: $p\to q$, $p_i\to q_i$, $\epsilon_{\pm}\to\nu_{\pm}$, $\mu_\phi\to\mu_\theta$ and $\tilde{u}+\tilde{v}\to\tilde{u}-\tilde{v}$, $s_i\to\sigma_i$ and $t_i\to\tau_i$. The same applies in the following cases.

(2) For the case in which P and Q possess two real and different roots, and two complex roots, i.e.,

$$P=\alpha_4\epsilon_{\pm}[p-p_1][p-p_2][p-(s_1+it_1)][p-(s_1-it_1)],\quad \alpha_4>0, \tag{3.40}$$

$$Q=\beta_4\nu_{\pm}[q-q_1][q-q_2][q-(\tilde{\sigma}_1+i\tilde{\tau}_1)][q-(\tilde{\sigma}_1-i\tilde{\tau}_1)],\quad \beta_4>0, \tag{3.41}$$

the transformations are given by

$P_4(p)$	$\epsilon_{\pm}$	Interval	$p(\tilde{u},\tilde{v})=$	k^2
two complex and two real roots	ϵ_+	$-\infty<p<\infty$	$\dfrac{p_1+p_2}{2}-\dfrac{p_1-p_2}{2}\left(\dfrac{\nu_\theta-\mathrm{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]}{1-\nu_\theta\mathrm{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]}\right)$	$\left(\sin\left(\dfrac{\theta_1-\theta_2}{2}\right)\right)^2$

(3) In the case P and Q with four complex roots, respectively, i.e.,

$$P=\alpha_4[p-(s_1+it_1)][p-(s_1-it_1)][p-(s_2+it_2)][p-(s_2-it_2)], \tag{3.42}$$

$$Q=\beta_4[q-(\tilde{\sigma}_1+i\tilde{\tau}_1)][q-(\tilde{\sigma}_1-i\tilde{\tau}_1)][q-(\tilde{\sigma}_2+i\tilde{\tau}_2)][q-(\tilde{\sigma}_2-i\tilde{\tau}_2)], \tag{3.43}$$

the transformations are now

$P_4(p)$	$\epsilon_\pm$	Interval	$p(\tilde{u},\tilde{v})=$	k^2
$s_1>s_2$			$s_1 + t_1\left[\frac{\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\cos((\theta_3/2)+(\theta_4/2))+\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\sin((\theta_3/2)+(\theta_4/2))}{\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\cos((\theta_3/2)+(\theta_4/2))-\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\sin((\theta_3/2)+(\theta_4/2))}\right]$	$\sin^2\theta_5$
	ϵ_+	$-\infty<p<\infty$		
$s_1=s_2$			$s_1 - t_1\left[\frac{\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\cos((\theta_3/2)+(\theta_4/2))+\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\sin((\theta_3/2)+(\theta_4/2))}{\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\cos((\theta_3/2)+(\theta_4/2))-\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})]\sin((\theta_3/2)+(\theta_4/2))}\right]$	
B $t_1>t_2$				

We include the transformations for the case of polynomials of third degree:

$P_3(p)$	$\epsilon_\pm$	Interval	$p(\tilde{u},\tilde{v})=$	k^2
three	ϵ_+	$p_3\leqslant p\leqslant p_2$	$\frac{p_{32}}{p_3}\left(\text{sn}\,\frac{1}{2\mu_\phi}(\tilde{u}+\tilde{v})\right)^2$	$\frac{p_{32}}{p_{31}}$
real		$p_1\leqslant p$	$\frac{p_1-p_2(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{1-(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	
roots	ϵ_-	$p\leqslant p_1$	$\frac{p_1(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2-p_{31}}{(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	$\frac{p_{21}}{p_{31}}$
		$p_2\leqslant p\leqslant p_1$	$\frac{p_2p_{31}-p_3p_{21}(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{31}-p_{21}(\text{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}$	
two	ϵ_+	$p_1\leqslant p$	$p_1-\frac{c_1(1-\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])}{\cos\theta_1(1-\text{cn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])}$	
complex and				$\left(\sin\left[\frac{\theta_1}{2}+\frac{\pi}{4}\right]\right)^2$
one real	ϵ_-	$p\leqslant p_1$		

For the polynomial Q the same tables are valid making the changes: $p\to q$, $p_i\to q_i$, $\epsilon_\pm\to\nu_\pm$, $\mu_\phi\to\mu_\theta$ and $\tilde{u}+\tilde{v}\to\tilde{u}-\tilde{v}$, $s_i\to\sigma_i$ and $t_i\to\tau_i$.

All possible combinations between the transformations corresponding to the different possible ranges of p and q should be taken into account. Moreover, it is straightforward to perform such combinations by using the explicit transformations in the tables given above.

B. Limiting transition

For completeness we include the limiting procedure to obtain from the fourth order polynomials and the third and second degree ones,

$$\begin{aligned} P&=\epsilon_\pm a_4(p-p_1)(p-p_2)(p-p_3)(p-p_4)\\ &=a_4p_1\left(\frac{p}{p_1}-1\right)(p-p_2)(p-p_3)(p-p_4)\\ &=\alpha_4\left(\frac{p}{p_1}-1\right)(p-p_2)(p-p_3)(p-p_4). \end{aligned} \tag{3.44}$$

Taking the limit when $p_1\to\infty$ we recover a third degree polynomial:

$$\lim_{p_1\to\infty} P_4=\epsilon_\pm\alpha_4(p-p_2)(p-p_3)(p-p_4)=\epsilon_\pm\alpha_4P_3. \tag{3.45}$$

Analogously we take the limit on P_3 when $p_2\to\infty$ to get a second order polynomial:

$$\lim_{p_2\to\infty} P_3 = \lim_{p_2\to\infty}\left\{\epsilon_\pm a_4 p_2\left(\frac{p}{p_2}-1\right)(p-p_3)(p-p_4)\right\}$$
$$= \lim_{p_2\to\infty}\left\{\epsilon_\pm \alpha_3\left(\frac{p}{p_2}-1\right)(p-p_3)(p-p_4)\right\}$$
$$= \epsilon_\pm \alpha_3 (p-p_3)(p-p_4) = \epsilon_\pm \alpha_3 P_2 . \tag{3.46}$$

The corresponding procedure can be achieved for Q_4 to obtain a second order polynomial.

C. Concrete example

As a concrete example let us consider, for instance, the case P with four real and different roots and Q also with four real and different roots, in which the coefficients of the higher degree are 1, and the roots of P and Q are denoted by p_i and q_i respectively. For the elliptic integral depending on the p-coordinate, with $P(p)=(p-p_1)(p-p_2)(p-p_3)(p-p_4)$, $(\epsilon_\pm=1)$, one has the relation

$$\int \frac{dx}{\sqrt{P(p)}} = \mu_\phi \int \frac{d\phi}{\sqrt{1-k_\phi^2 \sin^2\phi}} = \mu_\phi F(\phi,k), \tag{3.47}$$

where the explicit relation between p and ϕ reads

$$p = \frac{p_1 p_{42} - p_2 p_{41}\sin^2\phi}{p_{42} - p_{41}\sin^2\phi},$$
$$\sin^2\phi = \frac{p_{42}}{p_{41}}\left(\frac{p-p_1}{p-p_2}\right), \quad \rightarrow \quad \phi = \arcsin\left(\pm\sqrt{\frac{p_{42}}{p_{41}}\left(\frac{p-p_1}{p-p_2}\right)}\right). \tag{3.48}$$

The parameter k_ϕ, $0<k_\phi^2<1$, is given by

$$k^2 = (p_1,p_2,p_4,p_3) = \left(\frac{p_1-p_4}{p_1-p_3}\right)\left(\frac{p_2-p_3}{p_2-p_4}\right), \quad \text{and} \quad \mu = \frac{\mu_\phi}{\sqrt{\alpha_4}} = \frac{2}{\sqrt{p_{31}p_{42}}}. \tag{3.49}$$

It is straightforward to find the analogous expression for the elliptic integral depending on the q-coordinate. The explicit expression of the coordinates p and q through $\tilde{u}$ and $\tilde{v}$ is the following

$$p(\tilde{u},\tilde{v}) = \frac{p_1 p_{42} - p_2 p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}{p_{42} - p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}+\tilde{v})])^2}, \tag{3.50}$$

$$q(\tilde{u},\tilde{v}) = \frac{q_1 q_{42} - q_2 q_{41}(\mathrm{sn}\,[(1/2\mu_\theta)(\tilde{u}-\tilde{v})])^2}{q_{42} - q_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\tilde{u}-\tilde{v})])^2}. \tag{3.51}$$

In terms of the null variables u and v which bring the line element to the form

$$\frac{dp^2}{P_4} - \frac{dq^2}{Q_4} = d\tilde{u}d\tilde{v} = 4\frac{dudv}{UV} \tag{3.52}$$

the p and q variables amount to

$$p(u,v) = \frac{p_1 p_{42} - p_2 p_{41}(\mathrm{sn}\,[(1/\mu_\phi)\,(\arcsin(uV+vU))])^2}{p_{42} - p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,(\arcsin(uV+vU))])^2}, \tag{3.53}$$

$$q(u,v) = \frac{q_1 q_{42} - q_2 q_{41}(\mathrm{sn}\,[(1/2\mu_\theta)\,(\arcsin(uV-vU))])^2}{q_{42} - q_{41}(\mathrm{sn}\,[(1/2\mu_\theta)\,(\arcsin(uV-vU))])^2}. \tag{3.54}$$

On the other hand, the relations between ϕ and θ with p and q are

$$\phi \pm \theta = \arcsin\ \sqrt{\frac{p_{42}}{p_{41}}\left(\frac{p-p_1}{p-p_2}\right)} \pm \arcsin\ \sqrt{\frac{q_{42}}{q_{41}}\left(\frac{q-q_1}{q-q_2}\right)}, \tag{3.55}$$

where ϕ and θ stand correspondingly for the arguments of the elliptical Legendre integrals related with the integration of P and Q. We define the auxiliary non-null variables u' and v' as

$$u' = \sin\left(\frac{\phi+\theta}{2}\right), \quad v' = \sin\left(\frac{\phi-\theta}{2}\right), \tag{3.56}$$

with

$$\left.\begin{matrix}\phi\\ \theta\end{matrix}\right\} = \arcsin\ u' \pm \arcsin\ v' = \arcsin[u'\sqrt{1-v'^2} \pm v'\sqrt{1-u'^2}] = \arcsin[u'V' \pm v'U'], \tag{3.57}$$

where

$$V' = \sqrt{1-v'^2}, \quad \text{and} \quad U' = \sqrt{1-u'^2}. \tag{3.58}$$

Thus, when we substitute (3.57) and (3.58) into (3.50) and (3.51) it is straightforward to find

$$p(u',v') = \frac{p_1 p_{42} - p_2 p_{41}(u'V' + v'U')^2}{p_{42} - p_{41}(u'V' + v'U')^2}, \tag{3.59}$$

$$q(u',v') = \frac{q_1 q_{42} - q_2 q_{41}(u'V' - v'U')^2}{q_{42} - q_{41}(u'V' - v'U')^2}. \tag{3.60}$$

Moreover, from (3.7) and (3.57), the advanced and retarded time coordinates $\tilde{u}$ and $\tilde{v}$ are related with the variables u' and v' through

$$\begin{aligned}\left.\begin{matrix}d\tilde{u}\\ d\tilde{v}\end{matrix}\right\} = &\left[\frac{\mu_\phi}{\sqrt{1-k_\phi^2(u'V'+v'U')^2}} \mp \frac{\mu_\theta}{\sqrt{1-k_\theta^2(u'V'-v'U')^2}}\right]\frac{du'}{U'}\\ &+\left[\frac{\mu_\phi}{\sqrt{1-k_\phi^2(u'V'+v'U')^2}} \mp \frac{\mu_\theta}{\sqrt{1-k_\theta^2(u'V'-v'U')^2}}\right]\frac{dv'}{V'}.\end{aligned} \tag{3.61}$$

According to our general procedure (3.61) can be inverted, yielding $u' = u'(\tilde{u},\tilde{v})$ and $v' = v'(\tilde{u},\tilde{v})$.

In the limit of polynomials of second degree, i.e., P_2, Q_2, one has $k_\phi = k_\theta = 0$, and μ_ϕ, $\mu_\theta \to 1$, hence one recovers from (3.61) the widely used retarded and advanced time coordinates

$$d\tilde{u} = 2\frac{du'}{U'}, \quad d\tilde{v} = 2\frac{dv'}{V'}. \tag{3.62}$$

In the next section we present an explicit class of colliding wave solutions.

IV. A CLASS OF COLLIDING WAVES

Let us consider an explicit solution of the Einstein–Maxwell equations in region I, given by the following coframe in coordinates (x,p,q,y):

$$\begin{aligned}\vartheta^{\hat{0}} &= \frac{1}{H}\sqrt{\frac{\Delta}{Q}}\,dq, \quad \vartheta^{\hat{1}} = \frac{1}{H}\sqrt{\frac{Q}{\Delta}}\,(dx + p^2 dy),\\ \vartheta^{\hat{2}} &= \frac{1}{H}\sqrt{\frac{\Delta}{P}}\,dp, \quad \vartheta^{\hat{3}} = \frac{1}{H}\sqrt{\frac{P}{\Delta}}\,(dx - q^2 dy).\end{aligned} \tag{4.1}$$

Here we have the functions $H=H(p,q)$, $P=P(p)$, $Q=Q(q)$, and $\Delta=\Delta(p,q)$. The coframe is assumed to be orthonormal

$$g=o_{\alpha\beta}\vartheta^{\alpha}\otimes\vartheta^{b}. \tag{4.2}$$

Then the metric explicitly reads

$$g=\frac{1}{H^2}\left\{\frac{Q}{\Delta}\,(dx+p^2dy)^2-\frac{\Delta}{Q}\,dq^2+\frac{\Delta}{P}\,dp^2+\frac{P}{\Delta}\,(dx-q^2dy)^2\right\}, \tag{4.3}$$

with

$$\begin{aligned}
H(p,q)&:=1-\mu pq,\\
P(p)&:=b-g^2+2np-\epsilon p^2+2m\mu p^3+\left(-\frac{\lambda}{3}-\mu^2b-\mu^2e^2\right)p^4,\\
Q(q)&:=-(b+e^2)+2mq-\epsilon q^2+2n\mu q^3+\left(\frac{\lambda}{3}+\mu^2b-\mu^2g^2\right)q^4,\\
\Delta(p,q)&:=p^2+q^2,
\end{aligned} \tag{4.4}$$

while the nonvanishing electromagnetic field components are given by

$$\begin{aligned}
F_{xq}&:=\frac{(p^2-q^2)e-2gpq}{2\Delta^2}, \quad F_{yq}:=p^2F_{xq},\\
F_{xp}&:=\frac{g(q^2-p^2)-2epq}{\Delta^2}, \quad F_{yp}:=-q^2F_{xp}.
\end{aligned} \tag{4.5}$$

In order to express our solutions (4.4) and (4.5) in terms of the advanced and retarded time coordinates $\tilde{u}$, $\tilde{v}$, one has to use the inversion procedure, presented in the previous section, yielding $p=p(u,v)$ and $q=q(u,v)$. The explicit representation requires the use of tables for the sn function.

Our class of solutions, defined in region I, can be extended to the full space–time by introducing the Heaviside step function

$$\Theta(x)=\begin{cases}1, & x\geqslant 0\\ 0, & x<0\end{cases}, \tag{4.6}$$

with $\Theta^2(x)=\Theta(x)$, and performing the following transformation

$$u\to\Theta(u)u, \quad \text{and} \quad v\to\Theta(v)v. \tag{4.7}$$

For the case of real roots of the polynomials, treated extensively in Sec. III B, we obtain for region II

$$p(u,v)=\frac{p_1p_{42}-p_2p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{v}])^2}{p_{42}-p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{v}])^2}, \tag{4.8}$$

$$q(u,v)=\frac{q_1q_{42}-q_2q_{41}(\mathrm{sn}\,[(1/2\mu_\theta)\,\tilde{v}])^2}{q_{42}-q_{41}(\mathrm{sn}\,[(1/2\mu_\theta)\,\tilde{v}])^2}, \tag{4.9}$$

where

$$\mathrm{sn}\,\frac{\tilde{v}}{2\mu}=\mathrm{sn}\left[\frac{1}{\mu}\,\arcsin\,v\right]. \tag{4.10}$$

For region III

$$p(u,v)=\frac{p_1p_{42}-p_2p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{u}])^2}{p_{42}-p_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{u}])^2}, \tag{4.11}$$

$$q(u,v)=\frac{q_1q_{42}-q_2q_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{u}])^2}{q_{42}-q_{41}(\mathrm{sn}\,[(1/2\mu_\phi)\,\tilde{u}])^2}, \tag{4.12}$$

where

$$\mathrm{sn}\,\frac{\tilde{u}}{2\mu}=\mathrm{sn}\left[\frac{1}{\mu}\,\arcsin\,u\right]. \tag{4.13}$$

At this point it is important to mention that not every cylindrically symmetric space–time, even in vacuum, satisfies the requirements of Ernst[8] for being interpreted as colliding waves, only certain classes of cylindrically symmetric solutions can be thought of as solutions generated by collision of waves.

Moreover, until now our analysis has been concerned mostly with metrical aspects, leaving aside the task of having suitable field and matter energy-momentum tensors. As is well known, each participating wave in the head-on collision is assumed to be plane fronted, i.e., it is characterized by a covariantly constant null eigenvector k, i.e., $k_{\mu;\nu}=0$, of the Weyl–Petrov-type N conformal tensor. This condition implies a null energy-momentum tensor for the electromagnetic field, i.e., a radiation field or electrovacuum, let us call it null field for short, into the Einstein field equations, cf. Kramer *et al.*,[16] Sec. 21.5. Therefore, the existence of colliding wave solutions with nonvanishing cosmological λ term is thus forbidden.

Usually one derives colliding wave solutions using the backward procedure, namely, one starts with a cylindrically symmetric solution and an energy-momentum tensor in terms of fields depending on the usual two null variables, say u and v, restricted to a closed interaction region, region I in our approach. Furthermore, by accomplishing on it the transformations (4.7) with Heaviside functions one obtains the space–times prior to the collision. These space–times occur as type N solutions to the Einstein equations in vacuum or coupled with null fields—regions II and III in our approach, where the gravitational and null fields depend solely on a single variable (u or v). At this level, the colliding wave condition allows only for fields which can be matched to the flat Minkowski space–time (region IV) background where the wavefronts propagate. (We appreciate the referee's comment which drew our attention to this characteristic.) In other words, if the cylindrically symmetric space–time is endowed with a cosmological λ term, one has to set it equal to zero in order to connect smoothly the different regions of the collision and in order to avoid meaningless jumps of the cosmological constant across the null lines.

On the other hand, it is worthwhile to mention the results by Chandrasekhar and Xanthopoulos[17] concerning colliding wave solutions with a perfect fluid (stiff matter, fluid pressure=fluid energy density). They found that after the collision takes place, the incoming plane waves supporting radiative fields give rise to solutions with stiff matter perfect fluid tensor. It is an example of the transformation of massless particles describing null trajectories into a perfect fluid in which the stream lines describe timelike trajectories.

In full, in order to interpret our class of cylindrically symmetric solutions (4.4) as a colliding waves class of solutions, it is compulsory to set $\lambda=0$ on it.[16]

V. DISCUSSION

In this paper we present the formulation of the most general class, currently, of colliding waves with fourth degree polynomials determining the non-Killing sector, by means of a generalization of the advanced and retarded coordinates concept. As has been pointed out, after setting the cosmological λ term equal to zero, this class of solutions describes the scattering of two noncollinear polarized gravitation plane waves, having a null electromagnetic field as a source. At the leading edge of each colliding type-N gravitational wave, the curvature tensor exhibits delta and jump discontinuities. The former is interpreted as a gravitational impulsive wave, whereas the latter is attributed to a gravitational shock wave.

As stated above, this class of solutions, defined again in region I, can also be extended to the full space–time by introducing the Heaviside step function Θ.

The electromagnetic field presents delta singularities and jump discontinuities. However, the Bianchi identities hold in a distributional sense, see Ref. 18.

There are no problems on the right-hand side because the delta type singularities of the curvature are multiplied by the smooth distributions $\sqrt{1-\Theta(u)u^2}$ and $\sqrt{1-\Theta(v)v^2}$, respectively.

On the other hand (with respect to the singularities of the solution) we have in region I that the solution is of type *D* in the classification of Petrov. This is so because the only nonvanishing Weyl scalar is Ψ_2, and explicitly, it is given by

$$\Psi_2=\frac{H^3}{2\Delta^3}\,[2(e^2+g^2)(1+\mu pq)(p-iq)^2+q(q^2-3p^2)(a_3-ia_1)-p(p^2-3q^2)(a_1+ia_3)], \tag{5.1}$$

where the parameters $a_1=2n$, $a_3=2m$. The parameters ϵ and b do not appear in the curvature scalar, however the properties of the solution depend on them, as we shall see. The complex structure of Ψ_2 is due to the noncollinear polarization of the incoming waves (nondiagonal metric, see Ref. 19). Since we have the explicit expression for the Weyl scalar, Eq. (5.1), we can study the behavior of the invariants of the space–time and determine if the solution develops a curvature singularity a finite time after the instant of collision. Penrose attributed the development of the singularity in most colliding wave solutions to the self-focusing of the gravitational waves on scattering. In this case the invariants are given by $I=3\Psi_2^2$ and $J=\Psi_2^3$. Then the singularities of Ψ_2 are clearly the same for the invariants. A singularity occurs when $\Delta=p^2+q^2=0$, i.e., when simultaneously $p=0$ and $q=0$. The interesting question is whether there is any case in which the nature of the roots do not permit that $p=0$ and $q=0$ simultaneously. If this is the case, then the solution is free of the focusing singularity. A brief analysis shows that this occurrence can be avoided for certain values of the parameters of the solution.

Let us consider the expressions given in Sec. III C for the case of *P* and *Q*, both with four real roots. In this case (3.53) and (3.54) express p and q in terms of (u,v). Analyzing those expressions we see that $p=0$ and $q=0$ simultaneously if, for instance, $p_{42}=0$, $p_{41}=0$ and $q_{42}=0$, $q_{41}=0$, i.e., when at least three of the four roots coincide: $p_4=p_2=p_1$ and also $q_4=q_2=q_1$. The assumption of the occurrence of three equal roots for *P* and for *Q* leads to algebraic conditions that must be satisfied by the parameters of the solution, *b*, g, n, m, ϵ, *e* and μ. Some of these conditions are incompatible. For instance, in the case considered, the following condition must be valid: $\mu^2(b+e^2)=\mu^2(b-g^2)$. This is only true in the case if $e^2+g^2=0$; i.e., if $e^2+g^2\neq 0$ we do not have that $p=0$ and $q=0$ simultaneously and then the singularity fails to occur. Then we conclude that in some cases the existence of the electromagnetic field permits us to tune the parameters in such a way as to avoid the focusing singularity. A related result was obtained already by Chandrasekhar for a colliding wave solution with a scalar field[20] and by García *et al.* in the framework of metric-affine gauge theories of gravity.[21,22]

ACKNOWLEDGMENTS

We thank Friedrich W. Hehl and José Socorro for useful discussions and literature hints. We also thank the unknown referee for his valuable remarks. This research was supported by CONACyT (México), Grant Nos. 3544-E9311, 26329E, 3692P-E9607, and by the joint German-Mexican project DLR-Conacyt E130-2924 and MXI.6.B0a.6A; and also by FONDECyT (Chile)-1980891.

[1] K. Khan and R. Penrose, Nature (London) **229**, 185 (1971).
[2] P. Szekeres, J. Math. Phys. **13**, 286 (1972).
[3] S. Chandrasekhar and V. Ferrari, Proc. R. Soc. London, Ser. A **396**, 55 (1984).
[4] S. Chandrasekhar and B. C. Xanthopoulos, Proc. R. Soc. London, Ser. A **398**, 223 (1985).
[5] J. B. Griffiths, *Colliding Plane Waves in General Relativity* (Oxford University Press, Oxford, 1991).
[6] F. J. Ernst, in *Gravitational Collapse and Relativity*, edited by H. Sato and T. Nakamura (World Scientific, Singapore, 1986), pp. 141.
[7] F. J. Ernst, A. García, and I. Hauser, J. Math. Phys. **28**, 2155 (1987).
[8] F. J. Ernst, A. García, and I. Hauser, J. Math. Phys. **28**, 2951 (1987).
[9] F. J. Ernst, A. García, and I. Hauser, J. Math. Phys. **29**, 681 (1988).
[10] V. Ferrari and J. Ibañez, Gen. Relativ. Gravit. **19**, 383 (1987).

[11] V. Ferrari and J. Ibañez, Gen. Relativ. Gravit. **19**, 405 (1987).
[12] V. Ferrari and J. Ibañez, Proc. R. Soc. A. **417**, 417 (1988).
[13] V. Ferrari, J. Ibañez, and M. Bruni, Phys. Lett. A **122**, 459 (1987).
[14] A. García, J. Math. Phys. **25**, 1951 (1984).
[15] G. A. Korn and T. M. Korn, *Mathematical Handbook, for Scientists and Engineers. Definitions, Theorems and Formulas for Reference and Review* (MacGraw–Hill, New York, 1968), Chap. 21.
[16] D. Kramer, H. Stephani, M. A. H. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambridge, 1980).
[17] S. Chandrasekhar and B. C. Xanthopoulos, Proc. R. Soc. London, Ser. A **402**, 37 (1985).
[18] A. H. Taub, J. Math. Phys. **21**, 1423 (1980).
[19] A. García, Theor. Math. Phys. **83**, 434 (1990).
[20] S. Chandrasekhar and B. C. Xanthopoulos, Proc. R. Soc. London **414**, 1 (1987).
[21] A. García, C. Lämmerzahl, A. Macías, E. W. Mielke, and J. Socorro, Phys. Rev. D **57**, 3457 (1998).
[22] A. García, F. W. Hehl, A. Macías, and J. Socorro, ''A class of colliding waves in metric affine gravity,'' preprint UAM-I 980223 and CINVESTAV 980346, 1998.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 11 NOVEMBER 1998

A harmonic space approach to quantum gravity of stationary space–times with SO(3) symmetry

Helia Hollmann[a)]
Max-Planck-Institut für Physik, Werner-Heisenberg-Institut, Föhringer Ring 6, 80805 Munich, Germany

(Received 14 April 1997; accepted for publication 6 November 1997)

After dimensional reduction the stationary sector of SO(3) symmetric Einstein's gravity is identified with an $SL(2,\mathbb{R})/SO(2)$ Sigma model coupled to a one-dimensional gravitational remnant. The space of classical solutions consists of the Schwarzschild and Taub–NUT family. A Dirac quantization of this system is performed and the observables—the mass and the charge operator—are shown to be self-adjoint operators with a continuous spectrum ranging from $-\infty$ to ∞. The Hilbert space is constructed explicitly using a harmonic space approach. © *1998 American Institute of Physics.* [S0022-2488(98)03004-7]

I. INTRODUCTION

The quantization of Einstein's gravity is one of the outstanding and most challenging areas in theoretical physics. There have been many attempts at a solution of this problem classified by Isham in his articles about quantum gravity.[1,2] One can start with the classical theory of general relativity and apply some kind of quantization algorithm. General relativity may be considered as a low-energy limit of another theory, for example, superstring theory. One can either try to force some principal ideas of a quantum field theory to be compatible with general relativity or the fundament of quantum gravity is seen to be some radically new perspective which leads to the classical theory of general relativity in certain limiting situations. Attempts in this direction are quantum groups or the noncommutative geometry invented by Connes. Nonperturbative approaches to the problem became of great importance since it is known that quantum gravity is non-renormalizable.

This paper originates in a conservative point of view. Starting from the classical theory of general relativity one achieves a quantization via a canonical scheme, the ADM quantization. For full general relativity this is still an unsolved problem, but freezing all but a few of the infinitely many degrees of freedom of the gravitational field solutions for special sectors of Einstein's vacuum theory are within reach. It is hoped that at least some relevant features of the quantum theory are preserved. Even restricting oneself to a particularly symmetric case, that is spherical symmetry (as it is done here), the space of classical solutions contains a black hole, the Schwarzschild solution, whose special importance is explained now. There are no particles in the sense of ordinary field theory. Even the concept of solitons has to be modified as shown in Refs. 3 and 4. One has to allow singularities in the solutions, leading to the suggestion that black holes are the proper particles of general relativity: coupling a point particle to its own gravitational field turns it into a black hole.[5] Furthermore, classical black holes are stable under local perturbations,[6] and they have remarkable uniqueness properties. This is why they are considered to be the fundamental pieces of a quantum theory of gravitation and the investigation of the spherically symmetric and the SO(3) symmetric sector of Einstein's theory of gravitation is seen in a different light.[3,4,7] As in conventional gauge theory, where one tries to keep manifest Poincaré invariance even in the process of gauge fixing, this sector of Einstein's gravity has a global symmetry group, the $SL(2,\mathbb{R})$, which will be preserved even at a quantum level. The unitary irreducible representations play a key role in the process of quantization. The major advantage of this model is that the quantum theory is explicitly known.

One of the most important developments in field theory during the last two decades was the

[a)]Electronic mail:hollmann@mppmu.mpg.de

 © 1998 American Institute of Physics

discovery of the quantum mechanical instability of black holes due to Hawking radiation. Most of the calculations of black hole radiation involve matter fields quantized in front of a classical background. After the coupling of matter fields this model opens the possibility to take a closer look at Hawking radiation quantum mechanically. However, this is considered to be a project of future research.

This paper mainly splits into a classical and a quantum part. In the beginning the SO(3) symmetric sector of Einstein's gravity is introduced and a Kaluza–Klein-like reduction is performed.[8] This treatment is rather concise, but a more comprehensive and more general exposition can be found in Ref. 8. Gravity reduced from four to three dimensions in its dual representation is equivalent to a $SL(2,\mathbb{R})/SO(2)$ nonlinear coset sigma model coupled to three-dimensional gravity.[9] Using this coset space structure all stationary solutions of Einstein's equations with SO(3) symmetry are constructed in Ref. 10: there are the Schwarzschild solution labeled by a mass parameter and the Taub–NUT family characterized by a charge parameter in addition.[11] The mass parameter m and the charge parameter l labeling the space–time are hidden in the sigma model currents. In the following the problem is treated as a constraint system. The Hamiltonian is the only constraint which survives the process of dimensional reduction. The components of the sigma model currents consequently turn out to be observables in the sense of Dirac and form an $sl(2,\mathbb{R})$ algebra.

During quantization the classical phase space is turned into a Hilbert space on which the operators—the former functions on phase space—act. In general it is a difficult problem to identify the Hilbert space. On the other hand, it is the Hilbert space, which is of fundamental importance to make assertions about the operators to be considered physical. The implementation of the classical $SL(2,\mathbb{R})$ symmetry at a quantum level leads to a solution of this problem. The eigenfunctions of the invariant differential operator on the group span the Hilbert space as is shown by the application of a generalized Plancherel theorem. The measure on the Hilbert space is derived quite naturally from the Haar measure on the group. Once knowing the Hilbert space explicitly one can investigate the self-adjointness and the spectrum of certain differential operators. Both the mass operator and the charge operator are self-adjoint and their spectra are purely continuous. Neither the spectrum of the mass operator nor the spectrum of the charge operator are bounded, that is, negative charges as well as negative masses belong to the spectrum, too.

The most challenging feature of this model is that bases of the Hilbert space are at hand. Therefore the model serves as a kind of test laboratory for several aspects of quantum gravity. However, these topics will be reserved for a forthcoming publication.

This method is not limited to the situation here. It is applicable to various other models occurring in conformal field theory, string theory, quantum cosmology, or supergravity.

II. THE CLASSICAL THEORY

This classical part starts with the application of a Kaluza–Klein reduction[9,12] to a four-dimensional Einstein space M_4 with signature $(+,+,+,-)$ having a one-parameter Abelian isometry group[10] which acts freely and corresponds to time translations; that is there exists a timelike Killing vector field K^M describing the action of the Lie algebra of the isometry group. In a coordinate basis the metric g_{MN} of M_4 decomposes into a metric h_{mn} on the remaining three spatial dimensions, a scalar field τ and the vector field B_m:

$$g_{MN}=\begin{pmatrix} \frac{1}{\tau}\,h_{mn}+\tau B_m B_n & \tau B_n \\ \tau B_m & -\tau \end{pmatrix}.$$

The extra factor τ assures that diffeomorphisms of the special form $t\mapsto t+\Lambda(x^m)$ act as gauge transformations $B_m\mapsto B_m+\partial_m\Lambda$ on the vector field B_m.[3] Capital letters M and N vary from 1 to 4 and small ones from 1 to 3. Plugging this Ansatz for the four-dimensional metric into the Einstein Hilbert action

$$S=-\frac{1}{2}\int d^4\,\sqrt{-g}R \tag{1}$$

leads to the Lagrangian

$$\mathscr{L}^{(4,3)}=-\frac{1}{2}\sqrt{-g}R=\sqrt{h}\left(-\frac{1}{2}{}^{(3)}R-\frac{\tau^2}{8}F^{mn}F_{mn}+\frac{1}{4\tau^2}\partial^m\tau\partial_m\tau\right). \tag{2}$$

Here F_{mn} defined by $F_{mn}=\partial_m B_n-\partial_n B_m$ denotes the field strength of B_m, R and ${}^{(3)}R$ are the scalar curvatures corresponding to the four-dimensional metric g_{MN} and the three dimensional metric h_{mn}, respectively, and $\sqrt{-g}$ and $\sqrt{h}$ are the square roots of the determinants of the metrics as usual.

Under the specified conditions dimensional reduction of four-dimensional pure gravity to three dimensions leads to gravity coupled to an Abelian vector field and a scalar.

The equations of motion derived from (2) can be interpreted as an integrability condition for the vector field B_m:

$$\tau^2F_{mn}=\epsilon_{mnp}\,\partial^p\omega.$$

Elimination of the field strength in terms of the gravito-magnetic potential ω leads to a dual Lagrangian density

$$\mathscr{L}^D=\sqrt{h}\left(\frac{1}{2}{}^{(3)}R-\frac{h^{mn}}{4\tau^2}(\partial_m\tau\partial_n\tau+\partial_m\omega\partial_n\omega)\right),$$

which can be rewritten as

$$\mathscr{L}^D=\sqrt{h}\left(\frac{1}{2}{}^{(3)}R-\frac{h^{mn}}{8}\,\mathrm{Tr}\,(\chi^{-1}\partial_m\chi\chi^{-1}\partial_n\chi)\right), \tag{3}$$

where

$$\chi=\begin{pmatrix}\tau+\dfrac{\omega^2}{\tau} & \dfrac{\omega}{\tau}\\ \dfrac{\omega}{\tau} & \dfrac{1}{\tau}\end{pmatrix},$$

that is four-dimensional reduced gravity in its dual representation is equivalent to an $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$ sigma model coupled to three-dimensional gravity. The matrix χ is an element of the Riemannian symmetric space $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$. The fields contained in χ have a physical interpretation. The norm τ of the Killing vector plays the role of a gravitational potential, and ω is the so-called gravito-magnetic or NUT potential. In the stationary sector the NUT potential can be considered as a kind ''magnetic potential'' in analogy to Maxwell's theory.

For the derivation of the equations of motion one starts from (3). Here $R_{MN}=0$ is equivalent to the set of equations

$${}^{(3)}R_{mn}=\tfrac{1}{2}\,\mathrm{Tr}\,(\chi^{-1}\partial_m\chi\chi^{-1}\partial_n\chi), \tag{4a}$$

$$D^m(\chi^{-1}\partial_m\chi)=0. \tag{4b}$$

Now an additional SO(3) symmetry with S^2 orbits is imposed. Coordinates ρ, ϕ, and θ are introduced such that ρ is constant on the orbits; ϕ and θ are the usual angles.

The three-dimensional metric is then parametrized by

$$h_{mn}=\begin{pmatrix}N^2(\rho) & 0 & \\ & f^2(\rho) & \\ 0 & & f^2(\rho)\,\sin^2\theta\end{pmatrix},\quad m,n=\rho,\theta,\phi. \tag{5}$$

After substitution of this metric into the Lagrangian one obtains

$$\mathscr{L}^{(D)}=N\left[\frac{f'^2}{N^2}+1-\frac{f^2}{4N^2\tau^2}\,(\tau'^2+\omega'^2)\right]. \tag{6}$$

The prime $'$ denotes the derivative with respect to ρ, N is the ''lapse'' function, and ''lapse'' is set in quotation marks because it usually refers to a timelike direction whereas the lapse indicates spacelike propagation.

For SO(3) symmetric gravity $\chi^{-1}\chi'$ is unequal to zero and (4) simplifies to

$${}^{(3)}R_{22}={}^{(3)}R_{33}=\left[\frac{f''}{f}+\left(\frac{f'}{f}\right)^2-\frac{1}{f^2}\right]=0, \tag{7a}$$

$${}^{(3)}R_{11}=-\frac{2f''}{f}=\frac{1}{4}\,\mathrm{Tr}\,(\chi^{-1}\chi')^2, \tag{7b}$$

$$(f^2\chi^{-1}\chi')'=0. \tag{7c}$$

The lapse function refers to a gauge degree of freedom and was set equal to one meanwhile. Dobiasch and Maison[10] calculated the solution of these equations of motion. To show the unique correspondence between the observables and the initial data it is indicated briefly how the structure of the homogeneous space may be used to determine ω and τ classically (see Ref. 10). Using the transformation

$$\rho\mapsto\sigma(\rho)=-\int_\rho^\infty f^{-2}(s)\,ds,$$

(7c) become the geodesic equation in the homogeneous space

$$\frac{d^2\Phi^i}{d\sigma^2}+\Gamma^i_{jk}\frac{d\Phi^j}{d\sigma}\frac{d\Phi^k}{d\sigma}=0.$$

whose invariant line element is given by

$$ds^2=\frac{1}{4}\,\mathrm{Tr}\,(\chi^{-1}d\chi)^2=\frac{1}{2\tau^2}\,(d\tau^2+d\omega^2). \tag{8}$$

The Φ^i are the coordinates ω and τ. This yields the following system of differential equations:

$$\ddot{\tau}-\frac{\dot{\tau}^2}{\tau}+\frac{\dot{\omega}^2}{\tau}=0,\quad \ddot{\omega}-\frac{2\dot{\tau}\dot{\omega}}{\tau}=0. \tag{9}$$

The overdot $\dot{}$ denotes the derivative with respect to σ. The solutions are assumed to be asymptotically flat. From this one derives the boundary conditions $\tau(\sigma=0)=1$ for the gravitational potential τ and $\omega(\sigma=0)=0$ for the gravito-magnetic potential. To better understand the connection between the geodesic coordinate σ and the coordinate ρ one now cites the solutions for $a^2>0$ and $a=0$ from Ref. 10. Here a^2 is defined by $\mathrm{Tr}\,(\chi'\chi^{-1})^2=2a^2/f^4$. For $a^2>0$ the function [Eq. (7a) yields $f^2(\rho)=R^2-a^2$ with $R=\rho-b$] has two simple zeros at $\rho=b\pm a$. This leads to

$$\sigma(\rho)=\frac{1}{a}\,\ln\left(\frac{\rho-b-a}{\rho-b+a}\right).$$

For $\rho=b-a$ the geodesic coordinate σ tends to ∞. At $\sigma=0$ there starts a geodesic in (τ_0,ω_0) and assumes the value (τ_h,ω_h) at $\sigma=\infty$, which determines the position of the horizon of a black hole. The value $a=0$ corresponds to a solution with vanishing mass and charge. For this reason it belongs to flat space. The system of differential equations (9) has the solutions

$$\omega(\sigma)=\cosh\,\beta\,\tanh\,(a\sigma+\beta)-\sinh\,\beta, \tag{10}$$

$$\tau(\sigma)=\frac{\cosh\,\beta}{\cosh\,(a\sigma+\beta)}. \tag{11}$$

To summarize: For each geodesic there is a solution with an asymptotic behavior as demanded, i.e., $(\tau_0,\omega_0)=(1,0)$. It is uniquely determined by its tangent vector

$$(\dot\tau_0,\dot\omega_0)=\left(-a\,\tanh\,\beta,\,\frac{a}{\cosh\,\beta}\right)=(m,l).$$

The parameters m and l range from $-\infty$ to ∞. For the sake of completeness we also give the four-dimensional metric in terms of the one-dimensional data:

$$g_{MN}=\begin{pmatrix} \frac{1}{\tau} & & & & \\ & \frac{f^2}{\tau} & & 0 & \\ & & \frac{f^2}{\tau}\sin^2\,\theta+\frac{f^4}{\tau^3}\cos^2\,\theta\omega'^2 & \frac{f^2}{\tau}\cos\,\theta\omega' & \\ & 0 & & & \\ & & \frac{f^2}{\tau}\cos\,\theta\omega' & -\tau & \end{pmatrix},$$

$M,N=\rho,\theta,\phi,t$. The solution corresponding to $\sin\,\psi=0$ turns out to be the Schwarzschild solution, which is the gravitational field outside a spherically symmetric mass distribution. It is a static solution of Einstein's equations. For $\cos\,\psi=0$ only the μ_1 part remains and it is a Taub–NUT solution with mass parameter equal to zero. For any other value of ψ one obtains a Taub–NUT solution, labeled by a mass and a charge parameter. The geometrical difference between the Schwarzschild and the Taub–NUT solution is that the latter admits an isometry group whose orbits are on three-spheres, whereas in the case of the Schwarzschild solution they are on two-spheres. The ''result'' is the Taub–NUT time being a circle. Another interpretation is given in terms of the difference between staticity and stationarity: In the static case there are two symmetries: a time translation symmetry and a time reflection symmetry. For Taub–NUT the fields being time translation invariant fail to be time reflection invariant: the neighboring orbits of the Killing vector fields twist around each other!

Although it should be clear from the structure of the Lagrangian (3), it is stressed at this point once more that the space of solutions contains more than the Schwarzschild solution. The attempts to quantize spherically symmetric gravity[13–17] can be understood to be the static truncation GL(1) of this SL(2,$\mathbb{R}$)/SO(2) sigma model. The authors do not assume staticity, but the classical solutions turn out to have a hypersurface-orthogonal Killing vector that is timelike in certain regions of space–time.

After this excursion the Lagrangian (3) or (6), respectively, is investigated once more. It is invariant under SL(2,$\mathbb{R}$) transformations $z\mapsto(az+b)/(cz+d)$, $z=\omega+i\tau$, $ad-bc=1$. The corresponding Noether currents J are

$$J=\begin{pmatrix} -J^0 & J^+ \\ J^- & J^0 \end{pmatrix}:=\frac{f^2}{2N}\,\chi^{-1}\chi'=\frac{f^2}{2N}\begin{pmatrix} -\frac{\tau'}{\tau}-\frac{\omega\omega'}{\tau^2} & \frac{\omega'}{\tau^2} \\ \omega'-\frac{2\omega\tau'}{\tau}-\frac{\omega^2\omega'}{\tau^2} & \frac{\omega\omega'}{\tau^2}+\frac{\tau'}{\tau} \end{pmatrix}. \tag{12}$$

As χ is a symmetric matrix only two components of the current matrix are linearly independent. Imposing proper boundary conditions at infinity on the sigma model fields there exists an asymptotic multipole expansion of χ:

$$\chi \sim \sum_{n=0}^{\infty} \rho^{-n} \chi_n(\theta, \phi).$$

A suitable choice of coordinates asymptotically leads to

$$\omega = \frac{l}{\rho} + O\left(\frac{1}{\rho^2}\right) \quad \text{and} \quad \tau = 1 - \frac{2m}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

In particular one is interested in the $1/\rho$ term in the expansion of χ because it contains the mass and the charge parameters m and l, which are the entries of the matrix of global charges

$$Q = \frac{1}{4\pi} \int_{\partial S^2} J_\rho \, d\Sigma^\rho = \chi_0^{-1} \chi_1 .$$

By integration over an infinitely large sphere Q is calculated to be

$$Q = \begin{pmatrix} -2m & l \\ l & 2m \end{pmatrix}.$$

The identification of the mass parameter m and the charge parameter l is obtained by comparison with the standard form of the Schwarzschild and the Taub–NUT solution in the literature.[9] Evaluation of the currents J^0 and J^+ at a classical solution yield m and l, respectively. The J^0 component of the current is related to the mass parameter m, and the J^+ and J^- components lead to the charge parameter l.

So far the space of classical solutions is understood completely. Now one turns to the ADM approach to Einstein's gravity. General relativity is invariant with respect to space–time diffeomorphisms. This local symmetry relates part of the solutions stemming from the same initial conditions. In the Lagrangian formalism this results in the fact that out of ten field equations there are only six independent ones: The reparametrization invariance leads to Noether currents, the so-called Bianchi identities, and in the sequel to the presence of an arbitrary function of time in the general solution of the equations of motion. A solution is mapped to a solution by a gauge transformation. Performing a Hamiltonian formalism the local symmetry transformations yield a system with constraints. This means that there are conditions on the allowed initial data which must be preserved during time evolution. This consistency requirement can lead to secondary and higher-order constraints. The gauge degrees of freedom are hidden in the system of first class constraints (these are those whose Poisson brackets vanish weakly, that is on-shell).

Here a modified Hamiltonian formalism is implemented. The slicing is performed according to the ρ, that is a spacelike, coordinate. How to proceed in this case can be found in Refs. 13 and 18. Due to spherical symmetry the lapse function is the only Lagrange multiplier which survives. In other words, it expresses the invariance under ρ-reparametrization and this leads to the only primary first class constraint of the theory, the Hamiltonian constraint, which generates the gauge transformations

$$H = \frac{1}{4} \pi_f^2 - 1 - \frac{\tau^2}{f^2} (\pi_\tau^2 + \pi_\omega^2).$$

In terms of the fields and their conjugate momenta J reads

$$J = \begin{pmatrix} -\tau\pi_\tau - \omega\pi_\omega & \pi_\omega \\ \tau^2\pi_\omega - 2\omega\tau\pi_\tau - \omega^2\pi_\omega & \tau\pi_\tau + \omega\pi_\omega \end{pmatrix}. \tag{13}$$

By definition an observable is a function on the constraint surface that is gauge invariant. As explained above it has weakly, i.e., on-shell, vanishing brackets with the first class constraints. Observables do not evolve in ''time'' and therefore there is a one-to-one correspondence between the observables and the initial data here: They can be identified with the space of solutions. Here

the classical solutions are parametrized by the mass m and the charge l. Therefore one expects the existence of two observables, which turn out to be ''hidden'' in the current matrix J:

$$\{H,J^0\}=0, \quad \{H,J^+\}=0, \quad \{H,J^-\}=0.$$

On the other hand they generate an $\mathrm{sl}(2,\mathbb{R})$ algebra:

$$\{J^+,J^0\}=-J^+, \quad \{J^+,J^-\}=2J^0, \quad \{J^0,J^-\}=-J^-.$$

The observable J^0 measures the mass, and J^+ and J^- lead to the charge l. Because $\mathrm{Tr}\, J^2$ commutes with all the currents

$$\{H,\mathrm{Tr}\, J^2\}=0, \quad \{\mathrm{Tr}\, J^2,J^+\}=0, \quad \{\mathrm{Tr}\, J^2,J^-\}=0, \quad \{\mathrm{Tr}\, J^2,J^0\}=0,$$

these observables can be ''measured simultaneously'' (simultaneously diagonalized). A ''simultaneous measurement'' of the mass and the charge parameter m and l is not possible.

III. QUANTIZATION

Concerning the canonical approach to quantum gravity there are two main streams to handle the problem:[19] the reduced phase space quantization and the Dirac quantization. From a ''Hamiltonian point of view'' the former deals with the problem by eliminating the first class constraints at an early stage. This amounts to quantizing gauge invariant functions only, i.e., constants of motion. To carry out the reduced phase space quantization one must find a complete set of gauge invariant functions, which is a difficult task in general. Therefore often canonical gauge conditions are imposed to obtain the reduced phase space. After completely fixing the gauge any function of the canonical variables can be viewed as the restriction of a gauge invariant function in that gauge. Hence, once the gauge is fixed one is effectively working with gauge invariant functions. A complete set of independent gauge fixed functions provides a complete set of gauge invariant functions.

Problems of that quantization scheme are that an early elimination of the gauge degrees of freedom may spoil the manifest invariance under important symmetries and in general it destroys locality in space. Furthermore, it might happen that the brackets of the observables are complicated functions and the question arises of how to realize them quantum mechanically. The main advantage is that only the physical degrees of freedom are quantized. Every state in the Hilbert space is a physical one.

In the so-called Dirac quantization one keeps the gauge degrees of freedom. The classical phase space functions become operators acting in the Dirac representation space, which carries nonphysical information. Hence, one has to select a physical subspace of gauge invariant states.

This paper proceeds along the Dirac approach. The $\mathrm{SL}(2,\mathbb{R})$ symmetry of the theory provides the key to quantize this sector of Einstein's gravity. An outline of what happens next is given: The former phase space functions become operators. After this step a naive approach is followed: the Wheeler–DeWitt equation is solved formally. For the investigation of the self-adjointness and the spectra of the physical operators it is necessary to specify the Hilbert space employing group theoretical arguments. Then, to justify the choice of the Hilbert space or to obtain a basis, harmonic analysis is heavily used. That is what is done finally in this quantum part of the paper.

During the process of quantization the functions J^+, J^-, J^0, $\mathrm{Tr}\, J^2$, and H on the phase space become operators on an appropriate Hilbert space. In the Schrödinger representation the fields are turned into multiplication operators and the momenta become differentiation operators $\pi_\Phi \mapsto -i\partial_\Phi$, where ∂_Φ is an abbreviation for ∂/∂_Φ. The operators $\widehat{J^0}$, $\widehat{J^+}$, and $\widehat{J^-}$ turn out to be the differential operators one obtains calculating the Lie derivatives with respect to a suitable basis of the Lie algebra $\mathrm{sl}(2,\mathbb{R})$. This yields an algorithm of how to calculate the Casimir operator $\widehat{\mathrm{Tr}\, J^2}$: one squares the matrix J (12) with entries J^0, J^+, and J^- now in operator form and takes the trace. Therefore, by implementing the symmetry at the quantum level, no ambiguities in the operator ordering arise. The generators of the currents J^0, J^+, J^-, the Casimir $\mathrm{Tr}\, J^2$, and the Hamiltonian H become

$$\widehat{J^0} = -i\tau\partial_\tau - i\omega\partial_\omega,$$

$$\widehat{J^+} = i\partial_\omega,$$

$$\widehat{J^-} = i(\tau^2-\omega^2)\partial_\omega - 2i\omega\tau\partial_\tau,$$

$$\mathrm{Tr}\ J^2 = -\tau^2(\partial_\tau^2+\partial_\omega^2),$$

$$\hat{H} = -\frac{\partial_f^2}{4} + \frac{1}{2f^2}\widehat{\mathrm{Tr}\ J^2} - 1.$$

Applied to a wave function ψ the Hamiltonian defines the so-called Wheeler–DeWitt equation $\hat{H}\psi = 0$. The following commutation relations hold:

$$[\hat{H},\ \widehat{J^0}] = 0, \quad [\hat{H},\ \widehat{J^+}] = 0, \quad [\hat{H}, \widehat{J^-}] = 0,$$

i.e., the current operators are observables. The Casimir operator $\widehat{\mathrm{Tr}\ J^2}$ commutes with the Hamiltonian and the currents

$$[\widehat{\mathrm{Tr}\ J^2},\ \hat{H}] = 0, \quad [\widehat{\mathrm{Tr}\ J^2},\ \widehat{J^0}] = 0, \quad [\widehat{\mathrm{Tr}\ J^2},\ \widehat{J^+}] = 0, \quad [\widehat{\mathrm{Tr}\ J^2}, \widehat{J^{-1}}] = 0,$$

and $\widehat{J^0}$, $\widehat{J^+}$, and $\widehat{J^-}$ form an sl(2,$\mathbb{R}$) algebra:

$$[\widehat{J^+},\ \widehat{J^0}] = -i\widehat{J^+}, \quad [\widehat{J^+},\ \widehat{J^-}] = 2i\widehat{J^0}, \quad [\widehat{J^0},\ \widehat{J^-}] = -i\widehat{J^-}.$$

The Casimir operator $\widehat{\mathrm{Tr}\ J^2}$ is an essential part of the Wheeler–DeWitt equation. As it commutes with the current operators, $\widehat{\mathrm{Tr}\ J^2}$ and any one of the $\hat{J}$'s can be ''measured simultaneously,'' i.e., they can be simultaneously diagonalized. $\widehat{J^0}$, $\widehat{J^+}$, and $\widehat{J^-}$ do not commute with each other. Therefore, even formally a direct ''measurement'' of the mass M and the charge L is not possible.

The interpretation of the current operators as observables of the theory strongly suggests to preserve the SL(2,$\mathbb{R}$) symmetry at the quantum level: $\widehat{J^0}$, $\widehat{J^-}$, $\widehat{J^+}$, and the Casimir operator $\widehat{\mathrm{Tr}\ J^2}$ are forced to become self-adjoint operators during the process of quantization. That is, the quantum mechanical Hilbert space is built from the unitary irreducible representations of the group SL(2,$\mathbb{R}$). Finding a solution of the Wheeler–DeWitt equation therefore basically becomes a group theoretical problem as the equation splits into an f-dependent and an f-independent part, which is the Casimir operator on the group up to a constant term.

There are two particularly useful possibilities to diagonalize the Casimir operator $\widehat{\mathrm{Tr}\ J^2}$. On one hand, one solves the differential equations for $\widehat{\mathrm{Tr}\ J^2}$ and the charge operator $\widehat{J^+}$ simultaneously. The other possibility is to diagonalize the mass operator $\widehat{J^0}$ and the Casimir operator $\widehat{\mathrm{Tr}\ J^2}$.

The former case is investigated first, i.e., the formal solutions of the following system of differential equations,

$$-\tau^2(\partial_\tau^2+\partial_\omega^2)\psi_{\lambda L}(\omega,\tau) = \lambda\psi_{\lambda L}(\omega_r\tau), \tag{14a}$$

$$i\partial_\omega\psi_{\lambda L}(\omega,\tau) = L\psi_{\lambda L}(\omega,\tau), \tag{14b}$$

are obtained. The second equation, (14b), is considered first. The solution is

$$\psi_{\lambda L}(\omega,\tau) = C(\tau)\mathrm{e}^{-iL\omega}.$$

Substitution of this solution into the first equation yields

$$\tau^2\partial_\tau^2 C(\tau) + (\lambda - \tau^2 L^2)C(\tau) = 0.$$

This is a differential equation of Bessel type. If $L^2 \neq 0$, it follows that

$$\psi_{\lambda L}(\omega,\tau) = \psi^1_{\lambda L}(\omega,\tau) + \psi^2_{\lambda L}(\omega,\tau)$$

with

$$\psi^1_{\lambda L}(\omega,\tau) = C_1 \mathrm{e}^{-iL\omega}\sqrt{\tau}\mathrm{I}_k(|L|\tau), \tag{15a}$$

$$\psi^2_{\lambda L}(\omega,\tau) = C_2 \mathrm{e}^{-iL\omega}\sqrt{\tau}\mathrm{K}_k(|L|\tau). \tag{15b}$$

Here I_k and K_k denote the Bessel functions of imaginary argument. For $L^2 = 0$ it follows that

$$C(\tau) = \hat{C}_1 \tau^{1/2+k} + \hat{C}_2 \tau^{1/2-k}, \quad \text{for } k \neq 0,$$

$$C(\tau) = \sqrt{\tau}(\hat{C}_1 + \hat{C}_2 \log \tau, \quad \text{for } k = 0.$$

Now the simultaneous diagonalization of the Casimir operator $\widehat{\mathrm{Tr}\, J^2}$ and the mass operator $\widehat{J^0}$ is carried out. As in the former case the system of differential equations

$$-\tau^2(\partial^2_\tau + \partial^2_\omega)\psi_{\lambda M}(\omega,\tau) = \lambda \psi_{\lambda M}(\omega,\tau), \tag{16a}$$

$$-i(\tau\partial_\tau + \omega\partial_\omega)\psi_{\lambda M}(\omega,\tau) = M\psi_{\lambda M}(\omega,\tau) \tag{16b}$$

has to be solved. The solution of (16b) is

$$\psi_{\lambda M}(\omega,\tau) = C(v)\omega^{iM}, \tag{17}$$

where $v = \tau/\omega$. Alternatively $1/v$ could be used as an independent variable. This seems to be more natural at first sight, but leads to differential equations which are more complicated to handle later on. It turns out that (17) leads to a nice representation in terms of the associated Legendre functions. Substituting (17) into (16a) yields

$$-v^2(v^2+1)C'' + 2mv^3C' - (m(m+1)v^2 - \lambda)C = 0.$$

Here the prime $'$ denotes differentiation with respect to v and m is defined to be $m := iM - 1$. This differential equation is of hypergeometric type and transformed into the hypergeometric differential equation by

$$C(v) = v^{1/2+k}\eta(-v^2),$$

where $\eta(\xi)$ is the solution of

$$\xi(\xi-1)\eta'' + [(\alpha+\beta+1)\xi - \gamma]\eta' + \alpha\beta\eta = 0.$$

The prime $'$ denotes differentiation with respect to ξ, and α, β, and γ are equal to

$$\alpha = \tfrac{1}{4}(1+2k-2m), \quad \beta = -\tfrac{1}{4}(1-2k+2m), \quad \gamma = 1+k.$$

If $\lambda \neq -n^2 + \frac{1}{4}$, $n \in \mathbb{Z}$, $\eta(\xi)$ is calculated to be

$$\eta(\xi) = C_1 F(\alpha,\beta;\gamma;\xi) + C_2 \xi^{1-\gamma} F(\alpha-\gamma+1, \beta-\gamma+1; 2-\gamma; \xi). \tag{18}$$

Equation (18) can be expressed by associated Legendre functions so that one finally obtains

$$\psi_{\lambda M}(\omega,v) = \psi^1_{\lambda M}(\omega,v) + \psi^2_{\lambda M}(\omega,v)$$

with

$$\psi^1_{\lambda M}(\omega,v)=C_1\omega^{iM}\sqrt{v}(1+v^2)^{iM/2-1/4}\mathrm{P}^{-k}_{-iM-1/2}\left(\frac{1}{\sqrt{1+v^2}}\right), \tag{19a}$$

$$\psi^2_{\lambda M}(\omega,\tau)=C_2\omega^{iM}v^{2k+1/2}|v|^{-2k}(1+v^2)^{iM/2-1/4}\mathrm{P}^{k}_{-iM-1/2}\left(\frac{1}{\sqrt{1+v^2}}\right), \tag{19b}$$

where $\lambda\neq\frac{1}{4}-n^2$, $n\in\mathbb{Z}$. For the definition of P see Refs. 20 and 21.

In the coordinates ω and v the differential operators can be written in the following manner:

$$\widehat{J^0}=-i\omega\partial_\omega, \tag{20a}$$

$$\widehat{\mathrm{Tr}\,J^2}=v^2(1+v^2)\partial_v^2+\omega^2v^2\partial_\omega^2-2v\omega^3\partial_\omega\partial_v+2v^3\partial_v. \tag{20b}$$

To solve (20) means seeking solutions of $\widehat{\mathrm{Tr}\,J^2}\psi=\lambda\psi$ of the form $f(\omega)C(v)$.

Of course, the formal solutions of the differential equations do not yield enough structural elements to solve the problem. For a deeper understanding a Hilbert space structure is needed to single out part of the solutions of the differential equations and to show self-adjointness of the physical operators. For this one applies various techniques from functional analysis, the representation theory of SL(2,$\mathbb{R}$), the Plancherel theorem, and its version on the coset space SL(2,$\mathbb{R}$)/SO(2).

The natural choice of the Hilbert space is to take the closure of the functions on the coset space SL(2,$\mathbb{R}$)/SO(2) on which the physical operators can be shown to be essentially self-adjoint. One would also like to have a subset of functions which uniformly approximate all functions on compact subsets of the coset space. This is the content of the theorem by Stone and Weierstrass. It states that, given a compact metric space E, which is constructed below, any subalgebra of continuous functions on that space which contains the unity and separates the points of E is dense in the Banach space of functions E. One immediately recognizes that the space of L^2-functions over the coset space measurable with respect to the Haar measure on the group SL(2,$\mathbb{R}$) is too small: it does not contain the constants. Furthermore, the coset space SL(2,$\mathbb{R}$)/SO(2) is noncompact. This problem is solved by exhausting the coset space by a sequence of compact subsets. The restriction of the solutions on each compact subset of SL(2,$\mathbb{R}$)/SO(2) form (by group theoretical arguments) the required subalgebra of the Stone–Weierstrass theorem. The sequence of compact subsets converges towards SL(2,$\mathbb{R}$)/SO(2).

This paragraph is to motivate the further construction. The Hilbert space $\mathscr{H}$ is defined to be the closure with respect to the scalar product (21) (to follow) of the linear span of the finite norm solutions of the eigenvalue equations corresponding to $\widehat{\mathrm{Tr}\,J^2}$, $\widehat{J^+}$ or $\widehat{\mathrm{Tr}\,J^2}$, $\widehat{J^0}$, respectively. It turns out that the K-Bessel functions (15b) constitute a basis in the case of the diagonalization of $\widehat{\mathrm{Tr}\,J^2}$, $\widehat{J^+}$, and the Hilbert space is the closure of the span of these functions. For the diagonalization of $\widehat{\mathrm{Tr}\,J^2}$, $\widehat{J^0}$ the functions (19a) provide a basis. It can be shown that the Hilbert spaces are isomorphic. The scalar product is defined employing the Haar measure on the coset space dressed with a suitable damping factor, which is chosen such that the constants and enough functions to separate the points on SL(2,$\mathbb{R}$)/SO(2) belong to the Hilbert space. That is, from a technical point of view the Hilbert space theory of the continuation of symmetric operators to their closure is applied.

One realization of the coset space SL(2,$\mathbb{R}$)/SO(2) is the Poincaré upper half-plane $\mathbb{H}$. The details are given in the Appendix.

As motivated above, the Hilbert space $\mathscr{H}(\mathbb{H},d\mu_n)$ is the space of square integrable functions on the upper half-plane, which are measurable with respect to the SL(2,$\mathbb{R}$) invariant measure $d\mu$,

$$d\mu=\frac{d\omega d\tau}{\tau^2},$$

dressed with the damping factor

$$d_n(\omega,\tau)=\frac{\tau^n}{(\omega^2+(1+\tau)^2)^n},\quad \in\mathbb{N}.$$

The Hilbert space measure is defined by $d\mu_n=d_n(\omega,\tau)d\mu$. It is shown below that $\mathscr{H}(\mathbb{H},d\mu_n)$ contains solutions of the differential equations as long as $n>1$. It turns out that they constitute a dense subset of $\mathscr{H}(\mathbb{H},d\mu_n)$. With $d_{n_1}(\omega,\tau)<d_{n_2}(\omega,\tau)$ for $n_1>n_2$ there exists a natural embedding of the L^2 spaces with a smaller n into those with a greater n. In particular $\mathscr{H}(\mathbb{H},d\mu_0)=L^2(\mathbb{H})$. The scalar product $(,)_n$ is defined by

$$(f(\omega,\tau),g(\omega,\tau))_n=\int_{-\infty}^{\infty}d\omega\int_0^{\infty}d\tau\,\frac{d_n(\omega,\tau)}{\tau^2}\,f(\omega,\tau)g^*(\omega,\tau). \tag{21}$$

As to (14) the following lemma holds:

Lemma 1: For $1-4\lambda>0$ *neither* $\psi^1_{\lambda L}$ *nor* $\psi^2_{\lambda L}$ *belong to* $\mathscr{H}(\mathbb{H},d\mu_n)$. *For* $1-4\leqslant 0$ *the* $\psi^2_{\lambda L}$ *are elements of* $\mathscr{H}(\mathbb{H},d\mu_n)$, $n>1$.

It is sufficient to show that

$$(\psi^1_{\lambda L},\psi^1_{\lambda L})_n=\int_{-\infty}^{\infty}d\omega\int_0^{\infty}d\tau\,\frac{d_n(\omega,\tau)}{\tau^2}\,|\psi^1_{\lambda L}(\omega,\tau)|^2$$

and $(\psi^2_{\lambda L},\psi^2_{\lambda L})_n$ are finite. This is essentially proven by splitting the integrals into three parts with respect to τ, namely the asymptotic region near zero, a singularity-free part having ''compact support,'' and the asymptotic region near infinity.

At first the case $1-4\lambda>0$ is studied more closely. Near infinity $|\mathrm{I}_\kappa(|L|\tau)|\sim \mathrm{e}^{\tau}$ holds and therefore the integral does not exist. On the other hand, for small arguments $|\mathrm{K}_k(|L|\tau)|\geqslant k_0\tau^{-k/2}$, which lets the norm of $\psi^2_{\lambda L}$ tend to infinity, too.

For $1-4\lambda=0$ it is clear that I_0 is not bounded and therefore $\psi^1_{\lambda L}$ is not either. Near zero one can approximate $\mathrm{K}_0(x)$ by $\ln 2/x$ and near infinity by $\sqrt{\pi/2x}\mathrm{e}^{-x}$. Splitting the integrals into the three parts mentioned above it can be shown that the norm of $\psi^2_{\lambda L}$ is finite. That the integral over $\psi^2_{\lambda L}$ remains finite for $1-4\lambda<0$, too, is due to the fact that $\mathrm{K}_0(|L|\tau)$ can be used as an upper bound for the function K_k. The easiest way to see this is to use the integral representation for K_k.

To prove the divergence of $(\psi^1_{\lambda L},\psi^1_{\lambda L})$ for $1-4\lambda<0$ the asymptotic expansion of $\mathrm{I}_\nu(z)$ is employed.

Now one answers the question concerning which norm of which part of the solutions $\psi_{\lambda M}$ of the system of differential equations (16) is finite. The parameter κ assumes values on the positive real line $\mathbb{R}_+$ or on the positive imaginary line $i\mathbb{R}_+$.

Lemma 2: For $\lambda\in i\mathbb{R}$ *the functions* $\psi^1_{\lambda M}$ *belong to* $\mathscr{H}(\mathbb{H},d\mu_n)$ *with* $n>1$.

This lemma is proven by rather tricky estimates of the integrals $(\psi^1_{\lambda M},\psi^1_{\lambda M})_n$. The first integral is reduced to one over the damping factor d_n using that by holomorphy there exists an upper bound for the hypergeometric function on the interval $[0,\frac{1}{2}]$. Concerning the second integral, the existence of a lower bound in the neighborhood of 0 is quite useful to prove divergence of the integral.

The Hilbert space is spanned by the solutions of the systems of differential equations (14) or (16), respectively. In particular they form a dense subspace of the Hilbert space. Therefore the symmetry of the operators has to be proven for the solutions of these differential equations only where the operators are diagonal. That is, by construction of the Hilbert space the current operators $\widehat{J^0}$, $\widehat{J^+}$, and the Casimir operator $\widehat{\mathrm{Tr}\,J^2}$ are already symmetric. The self-adjointness of the operators is discussed with the technique of the deficiency indices n_+ and n_-. Given a symmetric operator A, the deficiency index $n_\pm$ is defined to be the dimension of the space $\ker(i\pm A^*)$. The operator is essentially self-adjoint iff the deficiency indices are zero. It has self-adjoint extensions, iff the deficiency indices are equal. Therefore the dimensions n_- of the kernel of $i-\widehat{J^+}^*$ and n_+ of $i+\widehat{J^+}^*$ are calculated next.

Lemma 3: The deficiency indices n_+ *and* n_- *are equal to zero, i.e., the operator* $\widehat{J^+}$ *is already essentially self-adjoint in* $L^2(\mathbb{H},d\mu)$.

The kernel, the dimension n_+ of which is calculated, consists of the functions $\psi_{\lambda L}$ with $\lambda = i$ and $L = -i$ and the ones corresponding to n_- consists of $\psi_{\lambda L}$ with $\lambda = -i$ and $L = i$. The techniques to prove the preceeding Lemma 2 are applied to show the divergence of all the integrals involved.

In an analogous manner one calculates the dimensions n_- of the kernel and $i - \widehat{J^0}$ and n_+ of $i + \widehat{J^0}^*$.

Lemma 4: The deficiency indices n_+ and n_- of the spaces $\ker(i \pm \widehat{J^0}^*)$ *are equal to zero. That is, the operator $\widehat{J^0}$ is already essentially self-adjoint in $L^2(\mathbb{H}, d\mu)$.*

Here $\widehat{\mathrm{Tr}\, J^2}$ is shown to be essentially self-adjoint, too. With the solutions of the differential equations (14) and (16) we have the following.

Lemma 5: The spectrum of the operators $\widehat{J^0}$ and $\widehat{J^+}$ consists of the whole real line.

Lemma 6: The spectrum of the Casimir operator is continuous and consists of a part of the positive real line, i.e., $[\frac{1}{4}, \infty)$.

To summarize: part of the formal simultaneous eigendistributions of the invariant differential operator $\widehat{\mathrm{Tr}\, J^2}$ and the charge operator $\widehat{J^+}$ as well as of $\widehat{\mathrm{Tr}\, J^2}$ and the physical ''mass-operator'' $\widehat{J^0}$ are shown to belong to $\mathscr{H}(\mathbb{H}, d\mu_n)$. They are exactly the functions which are needed to decompose a function on the Poincaré upper half-plane into irreducible parts. This is discussed in the group theoretical context below. The operators $\hat{J}^+$ and $\hat{J}^0$ are shown to be essentially self-adjoint. Finally, no constraints on the spectra of the operators $\widehat{J^+}$ and $\widehat{J^0}$ arise: they consist of the whole real line. Negative charges as well as negative masses belong to the spectrum, too.

Now one switches to the group theoretical point of view. There exists a canonical Hilbert space associated to the group $\mathrm{SL}(2,\mathbb{R})$. The unitary irreducible characters provide a basis of this Hilbert space. There are sufficiently many of them to separate the points of the group which means that a function on the group can be approximated by its characters: one obtains a generalized Fourier transform. This is the reason why the construction of the Hilbert space above makes sense. The property that the characters separate the points of the group reminds one of the assumption of the Stone–Weierstrass theorem mentioned above. The generalized inverse Fourier transform, which is called Plancherel theorem in the mathematical literature, yields an expansion of an arbitrary function on the group into a series or an integral of functions which occur as matrix elements of unitary irreducible representations on the group.[22–25] It contains the characters of the continuous and the discrete series. The formula can then be applied to elements of $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$ with the result that only one part of the continuous series survives. The eigendistributions of the invariant differential operators are shown to appear explicitly in the inverse Fourier transform. That is the deeper reason why the spectra of the mass and the charge operator are purely continuous.

For the classifications of the unitary irreducible representations of $\mathrm{SL}(2,\mathbb{R})$ and the notations used below the author refers to the Appendix.

This approach is similar to the one needed for the investigation of the hydrogen atom. There the coset space $\mathrm{SO}(3)/\mathrm{SO}(2)$, which is isomorphic to the two-sphere S^2, plays the key role. The generalized Fourier transform of $\mathrm{SO}(3)$ represented on S^2 leads to the spherical harmonics. The main difference arises from the fact that the group $\mathrm{SL}(2,\mathbb{R})$ is no longer compact, therefore the unitary irreducible representations are no longer finite dimensional, and apart from a discrete sum there further appears a direct integral in the decomposition formula.

The generalized Fourier transform[22] establishes a topological isomorphism between the C^∞ functions and the space of rapidly decreasing function on the dual of the group $\hat{G}$. ($\hat{G}$ is defined to be the set of all equivalence classes of unitary irreducible representations of the group.)

Theorem 1: *For each function $f \in \mathscr{S}(G)$,*

$$f(g) = \frac{1}{2\pi} \int_0^\infty dt\ \mathrm{Tr}\,[\hat{f}(0, \tfrac{1}{2} + it) V_g^{0,1/2+it}]\, t \tanh \pi t + \frac{1}{2\pi} \int_0^\infty dt\ \mathrm{Tr}\,[\hat{f}(\tfrac{1}{2}, \tfrac{1}{2} + it) V_g^{1/2,1/2+it}]\, t$$

$$\times \coth \pi t + \frac{1}{4\pi} \sum_{\substack{n \in 1/2\mathbb{Z} \\ n \geqslant 1}} (2n-1)\{\mathrm{Tr}\,[\hat{f}(n) U_g^n] + \mathrm{Tr}\,[\hat{f}(-n) U_g^{-n}]\}. \tag{22}$$

In particular

$$f(1)=\frac{1}{2\pi}\int_0^\infty dt(\Theta^{0,1/2+it}(f\,)t\,\tanh\,\pi t+\Theta^{1/2,1/2+it}(f\,)t\,\coth\,\pi t)$$

$$+\frac{1}{4\pi}\sum_{\substack{n\in 1/2\mathbb{Z}\\ n\geqslant 1}}(2n-1)[\Theta^{n}(f\,)+\Theta^{-n}(f\,)]$$

with

$$\Theta^{0,1/2+it}(f\,)=\int_G dg\;f(g^{-1})\Theta^{0,1/2+it}(g).$$

Here $\Theta^{0,1/2+it}$ and $\Theta^{1/2,1/2+it}$ denote the characters of the principal continuous series and Θ^{n} and Θ^{-n} those of the discrete series. Only the discrete and the principal continuous series contribute to the support of the Plancherel measure in $\hat{G}$ in the Plancherel theorem for a general group element g of SL(2,$\mathbb{R}$).

In Einstein's gravity a parametrization of the coset space is given: the upper half-plane. Therefore the formula has to be applied to $\mathbb{H}$. One starts with a function $f\in\mathscr{S}(\mathbb{H})$, which is left and right invariant under the action of the elliptic part K of the group. This leads to the conclusion that the irreducible characters associated with K do not contribute to the Fourier inversion formula for such functions. Without loss of generality g can be taken to be a regular element of the hyperbolic part of the group G'_{hyp}. One calculates

$$\mathrm{Tr}\;\hat{f}(n)U_g^n=\mathrm{Tr}\left(\int_G dl\;f_g(l)U_{l^{-1}}^n\right).$$

Here f_g is both left and right invariant under the action of K, too. The representation $U_{l^{-1}}^n$ is explicitly known. It acts on holomorphic functions s on $\mathbb{H}$ in the following manner:[23]

$$U_{l^{-1}}^n s(z)=\frac{1}{(cz+a)^{2n}}\,s\left(\frac{az+b}{cz+d}\right),\quad l^{-1}=\begin{pmatrix}a & b\\ c & d\end{pmatrix}.$$

For a rotation matrix l the action reads

$$U_{l^{-1}}^n s(z)=\frac{1}{(\sin(\theta/2)z+\cos(\theta/2))^{2n}}\,s\left(\frac{\cos(\theta/2)z-\sin(\theta/2)}{\sin(\theta/2)z+\cos(\theta/2)}\right).$$

The integrand is a holomorphic function on $\mathbb{H}$ and the path of integration is a closed circle around zero. Therefore the integral vanishes and the discrete series does not contribute to the Plancherel formula on $\mathbb{H}$.

The characters of the continuous series enter into the formula in two places.

Due to the symmetry the coth-part denoted by $f_2(g)$ does not contribute to the formula either. By translation all group elements are obtained from unity and

$$\mathrm{Tr}\,[\hat{f}(\tfrac{1}{2},\tfrac{1}{2}+it)V_g^{1/2,1/2+it}]=\Theta^{1/2,1/2+it}(f_g);\tag{23}$$

f_g is two-sided K-invariant leading to $f_g(g)=f_g(-g)$. The character $\Theta^{1/2,1/2+it}$ is[22]

$$\Theta^{1/2,1/2+it}(g)=\begin{cases}\dfrac{\cos\,t\tau}{|\sinh\,\tau|}, & g=g_0a_\tau g_0^{-1}\in G'_{\mathrm{hyp}},\\ -\dfrac{\cos\,t\tau}{|\sinh\,\tau|}, & g=-g_0a_\tau g_0^{-1}\in G'_{\mathrm{hyp}},\\ 0, & g\in G'_{\mathrm{ell}}\cup G_{\mathrm{par}}.\end{cases}$$

It follows that $\Theta^{1/2,1/2+it}(g)=\Theta^{1/2,1/2+it}(-g)$, i.e., the character is point symmetric in τ and (23) vanishes. That is, the coth-part does not contribute to the Plancherel theorem on the upper half-plane, either.

Given a representation space, which is the Poincaré upper half-plane here, it is now possible to prove that the Plancherel theorem on $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$ can be written such that the eigendistributions of the Casimir operator appear as integrand in the formula! This is an application of the known Fourier inversion theorem of zonal functions to this situation.

Theorem 2: *For* $f\in\mathscr{S}(\mathbb{H})$,

$$f(\omega,\tau)=\frac{1}{\pi^2}\int_{L\in\mathbb{R}}dL\int_{t\in\mathbb{R}}dt\ \hat{f}(L,t)e_{L,t}(\omega,\tau)t\ \sinh\ \pi t$$

where

$$e_{L,t}(\omega,\tau)=\exp\ (-iL\omega)\sqrt{\tau}K_{it}(|L|\tau),\quad \text{if } L\neq 0,$$

and

$$\hat{f}(L,t)=\int_{\mathbb{H}}d\omega d\tau\ \frac{f(\omega,\tau)\overline{e_{L,t}(\omega,\tau)}}{\tau^2}.$$

Proof: It is sufficient to show that the tanh part of the Plancherel formula (22), denoted by $f_1(g)$, is equivalent to this lemma. The nuclear spectral theorem[24] yields

$$\begin{aligned}f(\omega,\tau)&=2\pi\int_{L\in\mathbb{R}}dL\int d\lambda(t)\ \hat{f}(L,t)e_{L,t}(\omega,\tau)\\&=2\pi\int d\lambda(t)\int_{\tilde{\omega}\in\mathbb{R}}d\tilde{\omega}\int_{\tilde{\tau}\in\mathbb{R}_+}d\tilde{\tau}\,\frac{f(\tilde{\omega},\tilde{\tau})}{\tilde{\tau}^2}\left(\int_{L\in\mathbb{R}}dL\ \overline{e_{L,t}(\tilde{\omega},\tilde{\tau})}e_{L,t}(\omega,\tau)\right).\end{aligned}\tag{24}$$

The inner integral can be evaluated using the calculus of residues, which leads to an orthogonality relation for the K–Bessel functions:

$$\int_{L\in\mathbb{R}}dL\ \overline{e_{L,t}(\tilde{\omega},\tilde{\tau})}e_{L,t}(\omega,\tau)=\delta(\omega-\tilde{\omega})\,\delta(\tau-\tilde{\tau})\ \frac{\pi}{4\cosh\ t\pi}.\tag{25}$$

Substitution of (25) into (24) yields

$$f(\omega,\tau)=\frac{\pi^2}{2}\int_{t\in\mathbb{R}}d\lambda(t)\int_{\tilde{\omega}\in\mathbb{R}}d\tilde{\omega}\ \frac{f(\tilde{\omega},\tilde{\tau})}{\tilde{\tau}^2}\delta(\omega-\tilde{\omega})\,\delta(\tau-\tilde{\tau})\ \frac{1}{\coth\ \pi t}.\tag{26}$$

On the other hand $f_1(g)=f_{1,g}(1)$. Therefore, and because $\mathrm{SL}(2,\mathbb{R})$ is unimodular, $f_1(g)$ is transformed into

$$f_1(g)=\frac{1}{2\pi}\int_0^\infty dt\Theta^{0,1/2+it}(f_g)t\ \tanh\ \pi t=\frac{1}{2\pi}\int_0^\infty dt\int_G d\tilde{g}\,f_g(\tilde{g}^{-1})\Theta^{0,1/2+it}(\tilde{g})t\ \tanh\ \pi t.$$

The character is symmetric with respect to t (Ref. 22) and the $\mathrm{SL}(2,\mathbb{R})$ acts on the upper half-plane leading to

$$f_1(h)=\frac{1}{4\pi}\int_{-\infty}^{\infty}dt\int_{\mathbb{H}}d\tilde{h}\,f_h(\tilde{h}^{-1})\Theta^{0,1/2+it}(\tilde{h})t\,\tanh\,\pi t$$

$$=\frac{1}{4\pi}\int_{-\infty}^{\infty}dt\int_{\mathbb{H}}d\tilde{h}\,f_h(\tilde{h})\Theta^{0,1/2+it}(\tilde{h}^{-1})t\,\tanh\,\pi t$$

$$=\frac{1}{4\pi}\int_{-\infty}^{\infty}dt\int_{\mathbb{H}}d\tilde{h}\,f(\tilde{h})\Theta_h^{0,1/2+it}(\tilde{h}^{-1})t\,\tanh\,\pi t. \tag{27}$$

The normed character ($\Phi^{0,1/2+it}(e)=1$) is

$$\Theta^{0,1/2+it}(g)=\begin{cases}\cos\,t\tau, & g=g_0a_\tau g_0^{-1}\in G'_{\mathrm{hyp}},\\ 0, & \text{otherwise.}\end{cases}$$

Equations (26) and (27) yield

$$\lambda'(t)=\frac{1}{4\pi}\,\Theta^{0,1/2+it}(e)t\,\tanh\,\pi t\,\frac{2}{\pi^2}\cosh\,\pi t=\frac{t\,\sinh\,\pi t}{2\pi^3}\,\Theta^{0,1/2+it}(e),$$

and therefore

$$\lambda'(t)=\frac{1}{2\pi^3}\,t\,\sinh\,\pi t.$$

This concludes the proof. □

This Fourier inversion formula corresponds to a simultaneous diagonalization of the Casimir operator and the differential operator corresponding to the generator of the Lie algebra, which has a 1 in the upper right corner and 0's elsewhere. The charge L appears explicitly.

On the algebraic level $\widehat{J^0}$ belongs to the generator diag $(1,-1)$. The representation most suitable for the mass M yields another Plancherel inversion formula. Given a $f\in\mathscr{S}(\mathbb{H})$, this function can be written as

$$f(\omega,v)=C\int_{M\in\mathbb{R}}dM\int_{t\in\mathbb{R}}dt\;\hat{f}(M,t)g_{M,t}(\omega,v)d\lambda(M,t),$$

where

$$g_{M,t}(\omega,\tau)=\omega^{iM}\sqrt{v}(1+v^2)^{iM/2-1/4}\mathrm{P}^{-\kappa/2}_{-iM-1/2}\left(\frac{1}{\sqrt{1+v^2}}\right),\quad\text{if } M\neq 0,$$

and

$$\hat{f}(M,t)=\int_{\mathbb{H}}d\omega dv\;\frac{f(\omega,v)\overline{g_{M,t}(\omega,v)}}{v^2\omega}.$$

After the diagonalization of the Casimir operator the Wheeler–DeWitt equation is transformed into

$$\psi''+\left(4-\frac{2\lambda}{f^2}\right)\psi=0. \tag{28}$$

Here λ denotes the eigenvalue of the Casimir operator. Equation (28) can be interpreted as an evolution equation in the coordinate f. In contrast to ρ, f has a physical meaning as $1/f$ is the curvature of the two-spheres. Hence it is also a physical parameter. Equation (28) is a second-order differential equation and therefore there does not exist a positive semi-definite probability density which is invariant under SL(2,$\mathbb{R}$) transformations and conserved during evolution. This

situation is comparable with the one in the case of the Klein–Gordon equation: a conserved current can be associated to the Klein–Gordon equation, but the zero component of the current is not positive definite. Feshbach and Villar[26] interpreted j_0 as a charge density which measures the difference between the numbers of positive and negative charges. In the one-particle case, they showed that the density carries either a positive or a negative sign and that the two degrees of freedom of the second-order differential equation are identified as two possible but equivalent charge states. In Ref. 26 it is outlined that although the nonrelativistic equation only admits charges of one sign, in the relativistic generalization two signs of the charge happen to appear. The two equivalent charge states are basically obtained by transforming the Klein–Gordon equation into a symmetrized Schrödinger equation, i.e., a system of first-order differential equations. Now Eq. (28) is investigated in two steps. At first, the summand $-(2\lambda/f^2)$, is considered to be an f-dependent perturbation and is neglected. This ''free'' equation can easily be solved and two equivalent states are found. The second step consists of the investigation of the ''perturbed'' system. For λ given and finite the solution approaches the free solution in the limit $f\to\infty$. This fixes the asymptotic behavior of the solution. The two linear independent solutions can then again be shown to form two equivalent but independent states.

The ''free'' differential equation $\psi_0''+4\psi_0=0$ has the solution $\psi_0(f)=C_1e^{2if}+C_2e^{-2if}$. It follows that

$$\psi_0^*\partial_f\psi_0-\psi_0\partial_f\psi_0^*=\frac{i}{4}(\chi_0^*\chi_0-\phi_0^*\phi_0)=\frac{i}{4}(|C_1|^2-|C_2|^2)$$

with $\phi_0=C_1e^{2if}$ and $\chi_0=C_2e^{-2if}$. It is not difficult to solve the differential equation (28). The solutions with the proper asymptotic behavior turn out to be the Bessel functions of third kind $H_\nu^{(1)}$ and $H_\nu^{(2)}$:

$$\psi(f)=C_1\sqrt{f}H_\nu^{(1)}(2f)+C_2\sqrt{f}H_\nu^{(2)}(2f),\quad \nu=\tfrac{1}{2}\sqrt{1+8\lambda}.$$

Using that the complex conjugate of $H_\nu^{(1)}$ is $H_\nu^{(2)}$, the density is calculated to be

$$\begin{aligned}\psi\partial_f\psi-\psi\partial_f\psi&=2C_1^2f[H_\nu^{(2)}(2f)\partial_fH_\nu^{(1)}(2f)-H_\nu^{(1)}(2f)\partial_fH_\nu^{(2)}(2f)]\\&\quad+2C_2^2f[H_\nu^{(1)}(2f)\partial_fH_\nu^{(2)}(2f)-H_\nu^{(2)}(2f)\partial_fH_\nu^{(1)}(2f)]\\&=\frac{8i}{\pi}(|C_1|^2-|C_2|^2).\end{aligned}\tag{29}$$

The last equality uses the Wronskian of $H_\nu^{(1)}$ and $H_\nu^{(2)}$. The choice of the equivalent states is not unique as the scalar product is only defined up to phase. With the scalar product given the states are fixed in the evolution of f. Here one finds

$$\phi=\tfrac{1}{2}(\partial_f\psi+i\psi)\quad\text{and}\quad\chi=\tfrac{1}{2}(\partial_f\psi-i\psi).$$

In these new variables the density ρ reads $\rho=2i(\chi\chi^*-\phi\phi^*)$, yet the physical meaning of this discrete symmetry (''charge conjugation'') is not clear.

There is a nice relation between the quantization of the four-dimensional reduced SO(3) symmetric gravity in its dual representation and a WZNW model in the point particle version, which is further reduced to a Liouville theory.[27] There are also some important differences which lead to a deeper understanding of the quantization of the system dealt with in this paper. Performing the reduction to Liouville theory Fülöp ended with a Hamiltonian system with constraints, which correspond to initial values. Choosing these initial values the Hamiltonian of the latter system has exactly the form of the remaining part of the Wheeler–DeWitt equation (28), where the initial values take the place of the eigenvalue of the Casimir operator. The Liouville Hamiltonian is interpreted to be the Hamiltonian of a relativistic particle which moves in a potential. Formally, there are three cases to be distinguished depending on the sign of the constant λ for the initial values. If $\lambda>0$, the particle is repelled by an infinitely high potential barrier, when it travels in the negative f direction. For $\lambda=0$ the particle is free, and for $\lambda<0$ an infinitely deep potential

valley attracts the particle towards the negative f direction. This interpretation is in agreement with the group theoretical facts. $\lambda>0$ corresponds to $\lambda=s(1-s)=\frac{1}{4}+t^2$ and these are the values of the continuous series, which lead to ''scattering states''. $\lambda<0$ means $\lambda=n(1-n)$, $n\in\mathbb{Z}/2$, $n>1$, that is, it is an element of the discrete series of representation theory. Bound states are obtained. Here $\lambda\geqslant 0$ holds, because the Casimir operator can be shown to be the square of the charge and the mass operator. If the Casimir operator on the group would no longer be positive semi-definite, one would expect the discrete series of representations to play a role. The spectra of the physical operators would have a discrete part. However, in the case of stationary SO(3) symmetric gravity this argumentation is certainly not more than a consistency check: here the coordinate f has no intrinsic group theoretical meaning, contrary to the situation which occurs in the Liouville model. It is a ''gravitational remnant.''

Another important difference is that Fülöp considers the momenta of the particle to be fundamental. In particular they become Hermitian operators. Here, as the physical meaning is contained in the currents, there is no reason for the momenta to become Hermitian operators. Moreover, it turns out that starting with Hermitian momenta the currents are non-Hermitian. On the other hand, the conditions that $\widehat{J^0}$, $\widehat{J^+}$, and $\widehat{J^-}$ are Hermitian defines these operators to be Lie derivatives. This also fixes the operator ordering of the Laplacian and therefore of the Hamiltonian, too.

IV. RESULTS AND DISCUSSION

After proper dimensional reduction the SO(3) symmetric sector of Einstein's vacuum theory can be identified with an $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$ sigma model coupled to a gravitational remnant which belongs to a gauge degree of freedom. The gravitative part describes an evolution in the radius of the two-spheres. The ''true'' dynamical degrees of freedom are entirely hidden in the sigma model. In addition to the mass m another parameter of the classical space of solutions—the charge l—shows up. The invariance of the Lagrangian under the group $\mathrm{SL}(2,\mathbb{R})$ offers the key to quantize this part of gravity. The Hilbert space measure is given by the Haar measure of the group. The physical operators are self-adjoint due to the unitarity of the representations of $\mathrm{SL}(2,\mathbb{R})$ used here. The spectra of the mass and the charge operators are purely continuous and consist of the whole real axis. The spectrum of the Casimir operator is $[\frac{1}{4},\infty]$. In a modified Hamiltonian formalism the invariant differential operator on the group—the Casimir operator—appears quite naturally in the Hamiltonian constraint known as the Wheeler–DeWitt equation at the quantum level. After the diagonalization of the Casimir operator the Wheeler–DeWitt equation is reduced to an ordinary differential equation of second order which can be interpreted as Schrödinger equation for two equivalent states. Although this differential equation is of second order a positive definite scalar product on a not uniquely determined subspace can be defined. It is possible to diagonalize the Casimir operator and the mass operator or the charge operator simultaneously. The part of the eigendistributions not increasing more rapidly than a polynomial yields a spectral decomposition of the Laplacian. In the case of the mass operator one finds essentially associated Legendre functions, and in the case of the charge operator there appear K-Bessel functions. The Plancherel formula of $\mathrm{SL}(2,\mathbb{R})$, which allows us to decompose each function on the group into irreducible parts, can be applied to the coset space $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$. In the case of group $\mathrm{SL}(2,\mathbb{R})$ the continuous and the discrete series support the Plancherel measure whereas on the coset space one is left with one part of the continuous series.

The group theoretical methods are not limited to the situation here. They are applicable to various other stationary and static sectors of models in string theory and supergravity as well as to cosmological models. A list of models can be found in Ref. 3.

ACKNOWLEDGMENTS

The author is indebted to P. Breitenlohner and D. Maison for numerous discussions about the subject. She would also like to thank the referee for the encouraging report.

APPENDIX:

1. The isomorphism between SL(2,$\mathbb{R}$)/SO(2) and $\mathbb{H}$

One realization of the coset space SL(2,$\mathbb{R}$)/SO(2) is the Poincaré upper half-plane $\mathbb{H}$. The isomorphism is established now.[28] Each element g of the group SL(2,$\mathbb{R}$) can be represented by a 2×2 matrix with real entries and determinant equal to one:

$$\mathrm{SL}(2,\mathbb{R}) := \left\{ g=\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad ad-bc=1 \right\};$$

$\mathbb{H}$ is defined to be

$$\mathbb{H} := \{ z=\omega+i\tau\in\mathbb{C}, \tau>0 \}.$$

Let the SL(2,$\mathbb{R}$) act on $\mathbb{H}$ in the following manner:

$$g\circ z=\frac{az+b}{cz+d}.$$

It is easy to see that the mapping $g\mapsto g\circ i$ from SL(2,$\mathbb{R}$) into $\mathbb{H}$ induces a bijection SL(2,$\mathbb{R}$)/SO(2)$\rightarrow\mathbb{H}$.

2. The representations of SL(2,$\mathbb{R}$)

The classification of the unitary irreducible representations of SL(2,$\mathbb{R}$) is due to Bargman:[23]

(1) the principal continuous series $V^{j,s}$, $j=0,\ \frac{1}{2}$, $s=\frac{1}{2}+it$, $t\in\mathbb{R}$ (if $j=0$, then $t>0$, and if $j=\frac{1}{2}$, then $t\geqslant 0$),
(2) the limit of the discrete series $U^{1/2}$, $U^{-1/2}$,
(3) the discrete series U^{n}, U^{-n}, $n\in\mathbb{Z}/2$, $n>\frac{1}{2}$,
(4) the complementary series V^{σ}, $\frac{1}{2}<\sigma<1$, and
(5) the trivial representation.

Let λ denote the eigenvalue of the Casimir operator. To the five cases listed above there correspond the following eigenvalues:

(1) $\lambda=s(1-s)$,
(2) $\lambda=\frac{1}{4}$,
(3) $\lambda=n(1-n)$,
(4) $\lambda=\sigma(1-\sigma)$, and
(5) $\lambda=0$.

That is, group theory yields a restriction of the eigenvalues of the Casimir operator and some formal solutions of the differential equations are excluded from the Hilbert space by group theoretical arguments, namely the case $1-4\lambda>0$ mentioned above. To the representations there correspond characters, which are class functions. To evaluate and to apply them it is quite natural to factor the group by the Iwasawa decomposition into an elliptic, a hyperbolic, and a parabolic part, which are denoted by G_{ell}, G_{hyp}, and G_{par}, respectively. Now it is outlined how to proceed.

3. The Iwasawa decomposition of SL(2,$\mathbb{R}$)

Each 2×2 matrix of SL(2,$\mathbb{R}$) can be written as a product of a rotation matrix, the elliptic part of SL(2,$\mathbb{R}$), a diagonal matrix with entries either both positive or both negative, the hyperbolic part of the group, and one matrix with a Jacobi-like upper triangular form, which parametrizes the parabolic part of the group:

$$\mathrm{SL}(2,\mathbb{R})=KAN, \quad g=u_{\theta}a_{\tau}\eta_{\omega},$$

with

$$u_\theta = \begin{pmatrix} \cos\frac{\theta}{2} & \sin\frac{\theta}{2} \\ -\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}, \quad a_\tau = \begin{pmatrix} \tau & 0 \\ 0 & \tau^{-1} \end{pmatrix}, \quad \eta_\omega = \begin{pmatrix} 1 & \omega \\ 0 & 1 \end{pmatrix},$$

where τ is assumed to be positive. Denote $H := A \cup (-1)A$. Then G_{hyp}, G_{ell}, and G_{par} are defined by

$$G_{\text{hyp}} := \bigcup_{g \in G} gHg^{-1}, \quad G_{\text{ell}} := \bigcup_{g \in G} gKg^{-1}, \quad G_{\text{par}} := \bigcup_{g \in G} g(\pm N)g^{-1}.$$

Furthermore, the group $\mathrm{SL}(2,\mathbb{R})$ can be decomposed into disjoint parts of classes of regular elements, i.e., elements with distinct eigenvalues, G''_{ell}, G''_{hyp}, and G''_{par}. G''_{par} is the empty set. G may be represented as the product of G''_{ell}, G''_{hyp}, and G_{par}. The series of the unitary irreducible representations correspond in a one-to-one manner to the factors of the Iwasawa decomposition: the discrete series to the elliptic part, the continuous one to the hyperbolic factor, and the complementary series to the parabolic one.

4. Some relevant spaces

Before dealing with the decomposition of a function on $\mathrm{SL}(2,\mathbb{R})$ into its irreducible parts, the space of rapidly decreasing functions $\mathscr{S}(G), G = \mathrm{SL}(2,\mathbb{R})$ has to be introduced (see Ref. 24):

$$\mathscr{S}(G) := \left\{ f \in C^\infty(G) : \sup_{g \in G} \frac{(\omega^2(1+\tau)^2)^N}{\tau^N} |(D^\alpha f)(g)| < \infty, \quad \forall N \right\}.$$

The tempered distributions $\mathscr{S}'(G)$ are the dual space of the space of the rapidly decreasing functions. They are called slowly increasing. The triplet $\mathscr{S}(G), L^2(G, d\mu_G)$, and $\mathscr{S}'(G)$ is called a Gel'fand triplet. $d\mu_G$ denotes the Haar measure on $\mathrm{SL}(2,\mathbb{R})$. It holds

$$\mathscr{S}(G) \subset L^2(G, d\mu_G) \subset \mathscr{S}'(G).$$

As already mentioned above, $\mathscr{H}(G, d\mu_G) = L^2(G, d\mu_G)$. $\mathscr{S}'(G)$ contains all of the spaces $\mathscr{H}(G, d\mu_n)$. $\mathscr{H}(G, d\mu_G)$ denotes the Hilbert space of functions on G, which are measurable with respect to the Haar measure of G.

[1] C. J. Isham, in *Canonical Gravity*, Proceedings of the 117th WE. Heraeus Seminar in Bad Honnef, edited by J. Ehlers and H. Friedrich (Springer, Berlin, 1994).
[2] C. J. Isham, gr-qc/9510063.
[3] P. Breitenlohner, D. Maison, and G. Gibbons, Commun. Math. Phys. **120**, 295 (1988).
[4] B. Carter, in *Black Holes*, edited by B. De Witt and C. De Witt (Gordon and Breach, New York, 1973).
[5] T. Damour, in *Gravitational Radiation*, edited by N. Deruelle and T. Piran (North Holland, Amsterdam, 1986).
[6] S. Chandrasekhar, *Mathematical Theory of Black Holes* (Oxford U.P., Oxford, 1983).
[7] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge U.P., Cambridge, 1973).
[8] P. Breitenlohner and D. Maison, *Solitons in Kaluza-Klein Theories* (unpublished).
[9] D. Kramer, H. Stephani, E. Herlt, M. MacCallum, and E. Schmutzer, *Exact Solutions of Einstein's Field Equations* (Cambridge U.P., Cambridge, 1980).
[10] P. Dobiasch and D. Maison, Gen. Relativ. Gravit. **14**, 231 (1982).
[11] A. H. Taub, Ann. Math. **53**, 472 (1951).
[12] T. Kaluza, Sitzungsber. K. Preuss. Akad. Wiss. 966 (1921).
[13] M. Cavaglià, V. de Alfaro, and A. T. Filippov, Int. J. Mod. Phys. D **4**, 661 (1995).
[14] M. Cavaglià, V. de Alfaro, and A. T. Filippov, gr-qc/9508062.
[15] H. A. Kastrup and T. Thiemann, Nucl. Phys. B **339**, 211 (1993).
[16] K. Kuchar, Phys. Rev. D **50**, 3961 (1994).
[17] D. Louis-Martinez, J. Gegenberg, and G. Kunstatter, Phys. Lett. B **321**, 193 (1994).
[18] J. Martin, Phys. Rev. D **49**, 5089 (1994).
[19] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton U.P., Princeton, NJ, 1992).
[20] N. N. Lebedew, *Spezielle Funktionen und ihre Anwendung* (B.I.-Wissenschaftsverlag, Mannheim, 1973).
[21] H. Bateman, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. 1.
[22] M. Sugiura, *Unitary Representations and Harmonic Analysis* (North-Holland, New York, 1990).
[23] V. Bargmann, Ann. Math. **48**, 568 (1947).
[24] A. O. Barut and R. Raçzka, *Theory of Group Representations and Applications* (Polish Scientific, Warczawa, 1977).

[25] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications I* (Springer, Berlin, 1985).
[26] H. Feshbach and F. Villar, Rev. Mod. Phys. **21**, 25 (1958).
[27] T. Fülöp, J. Math. Phys. **37**, 1617 (1996).
[28] S. Lang, $SL_2(\mathbb{R})$ (Addison–Wesley, New York, 1975).

The complete Schwarzschild interior and exterior solution in the harmonic coordinate system

Quan-Hui Liu[a)]
Department of Physics, Hunan University, Changsha 410082, People's Republic of China, and Institute of Theoretical Physics, Academia Sinica, P.O. Box 2735, Beijing 100080, People's Republic of China[b)]

(Received 12 January 1998; accepted for publication 23 July 1998)

An exact solution of Schwarzschild interior and exterior solution for a star which has a uniform density and whose center pressure approaches infinite, in the harmonic coordinate system, is given. This solution enables us to demonstrate that in known literatures, a term in the Schwarzschild exterior solution is usually ignored in such a system. Neglect of the term directly leads to contradiction of the continuity of the Schwarzschild interior and exterior solution expressed in the system. © *1998 American Institute of Physics.* [S0022-2488(98)02911-9]

The Harmonic coordinate system (HCS) is an important coordinate system in general relativity. It can be used in studying, e.g., the gravitational fields produced by stable stars,[1–3] black holes,[4,5] gravitational waves,[6] supergravity,[7] quantum gravity,[8] etc. The most fundamental result of HCS is the usual Schwarzschild exterior solution (ES) under HCS.[1] In this paper, we are going to point out that the fundamental ES is not complete indeed if both the Schwarzschild interior solution (IS) and ES are transformed into those in the HCS and then examine their continuity at the boundary.

For simplicity, we will use the rational units in which gravitational constant G and light velocity c are unity.

At first let us examine the general Schwarzschild type solution in the so-called standard coordinate system (SCS)[1] as

$$d\tau^2=E(r)dt^2-G(r)dr^2-r^2(d\theta^2+\sin^2\theta\, d\phi^2). \tag{1}$$

This solution can be expressed in the following coordinates X_μ ($\mu=0,1,2,3$):

$$X_0=t, \tag{2}$$

$$X_1=R(r)\sin\theta\cos\phi, \tag{3}$$

$$X_2=R(r)\sin\theta\sin\phi, \tag{4}$$

$$X_3=R(r)\cos\theta. \tag{5}$$

In these coordinates, the proper time (1) becomes:[1]

$$d\tau^2=E(r)dt^2-\frac{r^2}{R^2}d\mathbf{X}^2-\left(\frac{G}{R^2R_r'^2}-\frac{r^2}{R^4}\right)(\mathbf{X}\cdot d\mathbf{X}), \tag{6}$$

where $R_r'=dR/dr$.

In the new coordinates, the continuity of the solution at the surface of a spherically symmetrical star of radius a is that the metric components and the so-called extrinsic curvature[9] are continuous at $r=a$, respectively. For metric (6) they are

a)Electronic mail: liuqh@itp.ac.cn
b)Mailing address.

 © 1998 American Institute of Physics

$$E_{\rm in}(a)=E_{\rm ex}(a),\quad G_{\rm in}(a)=G_{\rm ex}(a),$$
$$R_{\rm in}(a)=R_{\rm ex}(a),\quad R'_{\rm in}(a)=R'_{\rm ex}(a). \tag{7}$$

where subscripts ''in'' and ''ex'' denote the interior and exterior solution of E, G and R in Eq. (6) respectively.

The coordinates X_μ are to be harmonic if and only if they satisfy the harmonic condition:

$$\Box^2 X_\mu \equiv g^{\lambda\nu}\left[\frac{\partial^2 X_\mu}{\partial x^\nu \partial x^\lambda}-\Gamma^\gamma_{\lambda\nu}\frac{\partial X_\mu}{\partial x^\gamma}\right]=0. \tag{8}$$

It leads to that $R(r)$ must satisfy the differential equation:[1]

$$\frac{d}{dr}\left[r^2\sqrt{\frac{E(r)}{G(r)}}\frac{dR}{dr}\right]-2\sqrt{E(r)G(r)}R=0. \tag{9}$$

Evidently, only after $R(r)$ is solved can we say the metric (6) in HCS is obtained. So, it is necessary to solve ES $R_{\rm ex}(r)$ and IS $R_{\rm in}(r)$, respectively. For ES, $R_{\rm ex}(r)$ is easily available. We have known ES in the SCS for a long time as[1,2,10]

$$E_{\rm ex}(r)=1-\frac{2M}{r}, \tag{10}$$

$$G_{\rm ex}(r)=\frac{1}{1-2M/r}, \tag{11}$$

in which M is a constant of integration. Substituting these relations into Eq. (9), we have

$$\left(1-\frac{2M}{r}\right)R''_{rr}+\frac{2}{r^2}(r-M)R'_r-\frac{2R}{r^2}=0. \tag{12}$$

The general form of the ES $R_{\rm ex}(r)$ is in the interval $0<|r-2M|<\infty$:

$$R_{\rm ex}(r)=C(r-M)+C_1\left[\frac{r-M}{2M}\ln\left(1-\frac{2M}{r}\right)+1\right], \tag{13}$$

where C and C_1 are two arbitrary constants and the constant C can be conveniently chosen to be 1. It is evident in mathematics that the second term cannot be discarded, or say, we cannot put $C_1=0$, when the solution applying to physics objects with $r>2m$. Note that for all spherically symmetrical stable stars, uniform or not, $r\geqslant 9M/4>2M$ always holds.[1] It means that this term probably plays a role for all stars, otherwise the last two equations in continuity equations (7) are not compatible. Surprisingly, in known literature, for example, Refs. 1–5 and 8, C_1 is made zero. In the rest of this paper, an exact solvable star model is given to demonstrate $C_1\neq 0$.

As is well-known, solving IS $R_{\rm in}(r)$ from Eq. (9) is not so simple because the explicit forms of ISs depend on the structure of the stars. For some stars, solving Eq. (9) is mathematically formidable. Fortunately, for a uniform density star whose center pressure approaches infinity, Eq. (9) can be solve analytically. The explicit form of IS for this star will be presented.

First, we need to know the metric components $E_{\rm in}(r), G_{\rm in}(r)$ of IS in the SCS. For an arbitrary uniform density star in the SCS they are in general:[10]

$$E_{\rm in}(r)=\left[A-B\sqrt{1-\frac{r^2}{R_0^2}}\right]^2, \tag{14}$$

$$G_{\rm in}(r)=\frac{1}{1-r^2/R_0^2}, \tag{15}$$

where $R_0^2=3/(8\pi\rho_0)$,[10] A and B are two constants of integration. The above IS can be simplified by considering the internal pressure $p(r)$ of the star, which is[10]

$$8\pi p(r)=\frac{1}{R_0}\left(\frac{3B\xi(r)-A}{A-B\xi(r)}\right), \tag{16}$$

where $\xi(r)=\sqrt{1-r^2/R_0^2}$. Since the pressure equals zero at the surface of the sphere $r=a$, we have

$$3B\xi(a)-A=0, \tag{17}$$

which shows that constants A and B are not independent. And since the star we are interested in has infinite pressure at $r=0$, we have from Eqs. (16) and (17) two relations:

$$A=B, \quad a=\sqrt{\frac{8}{9}}R_0. \tag{18}$$

Substituting these relations into Eqs. (14) and (15) we have the simplified forms of $E_{\text{in}}(r)$ and $G_{\text{in}}(r)$ as

$$E_{\text{in}}(r)=A^2\left[1-\sqrt{1-\frac{8r^2}{9a^2}}\right]^2, \tag{19}$$

$$G_{\text{in}}(r)=\frac{1}{1-8r^2/9a^2}. \tag{20}$$

There are still two arbitrary constants A in IS (19) and M in ES [(10) and (11)]. We can fix them by utilizing the first two equations in (7). And they give:

$$M=\frac{4}{3}\pi\rho_0 a^3, \quad A=\frac{1}{2}, \quad a=\frac{9M}{4}. \tag{21}$$

So far, both the ES [(10) and (11)] and the IS [(19) and (20)] in the SCS for the special uniform density star are complete.

Second, we need to know the IS $R_{\text{in}}(r)$ in the HCS. From Eq. (9) and the above results the equation determining the IS $R_{\text{in}}(r)$ can be established as

$$R''_{rr}+R'_r\left[\frac{2}{r}+\frac{r}{R_0^2}\frac{1}{[1-\xi(r)]\xi(r)}-\frac{r}{R_0^2}\frac{1}{\xi^2(r)}\right]-\frac{2R}{r^2\xi^2(r)}=0. \tag{22}$$

After lengthy and delicate work, we obtain the following general solution of Eq. (22):

$$R_{\text{in}}(r)=C_2 z^{n/2}(1-z)^{m/2}F(\alpha,\beta,\gamma;z)+C_3 z^{n/2}(1-z)^{m/2}z^{1-\gamma}F(\alpha-\gamma+1,\beta-\gamma+1,2-\gamma;z), \tag{23}$$

in which

$$n=\frac{-3+\sqrt{17}}{2}, \tag{24}$$

$$m=1, \tag{25}$$

$$\alpha=\frac{m+n}{2}=\frac{-1+\sqrt{17}}{4}, \tag{26}$$

$$\beta=\frac{m+n+6}{2}=\frac{11+\sqrt{17}}{4}, \tag{27}$$

$$\gamma=\frac{2n+5}{2}=\frac{2+\sqrt{17}}{2}, \tag{28}$$

$$z=\tfrac{1}{2}(1-\xi(r)), \tag{29}$$

$$F(\alpha,\beta,\gamma;z)=1+\sum_{k=1}^{\infty}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\frac{z^k}{k!}, \tag{30}$$

$$(\alpha)_k=\alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1), \tag{31}$$

C_2 *and* C_3 *are two arbitrary constants.* (32)

Thus the general ES $R_{\rm ex}(r)$ and IS $R_{\rm in}(r)$ for the special uniform density star are all obtained. Now we are in a position to discuss the constants C_1, C_2, C_3 in Eqs. (13) and (23). Evidently we must set $C_3=0$ for $z^{n/2+1-\gamma}\sim r^{-1/4(3+\sqrt{17})}\to\infty$ at the vicinity $r=0$. And importantly, we cannot set $C_1=0$ again. Otherwise the mathematical consistency of the last two equations in continuity equations (7) will be violated. In fact the constants C_1 and C_2 are

$$C_1=\frac{(a-M)g'(a)-g(a)}{g(a)f'(a)-g'(a)f(a)}, \tag{33}$$

$$C_2=\frac{(a-M)f'(a)-f(a)}{g(a)f'(a)-g'(a)f(a)}, \tag{34}$$

where

$$f(r)=\frac{r-M}{2M}\ln\left(1-\frac{2M}{r}\right)+1, \tag{35}$$

$$g(r)=z^{n/2}(1-z)^{m/2}F(\alpha,\beta,\gamma;z). \tag{36}$$

Substituting the results (21) into C_1 and C_2, they are simply:

$$C_1=0.060\,498a, \tag{37}$$

$$C_2=0.594\,473a. \tag{38}$$

Now the complete forms of the IS $R_{\rm in}(r)$ and ES $R_{\rm ex}(r)$ in HCS are obtained as

$$R_{\rm ex}(r)=(r-M)+0.060\,498a\left[\frac{r-M}{2M}\ln\left(1-\frac{2M}{r}\right)+1\right], \tag{39}$$

$$R_{\rm in}(r)=0.594\,473az^{n/2}(1-z)^{m/2}F(\alpha,\beta,\gamma;z). \tag{40}$$

In other words, from Eqs. (2)–(6), the complete Schwarzschild interior and exterior solution in the HCS for the star is obtained. As we expected, the constant C_1 has a nonvanishing value. The curve showing $R(r)$ versus r is plotted in Fig. 1, in which we can see the curve at point $r=a$ is smoothly connected indeed.

In summary, we can conclude that the usual ES in HCS is not complete in mathematics, or strictly speaking, is not the correct mathematical solution in HCS corresponding to the ES in the SCS, for a term is neglected as $C_1=0$. The neglected term reflects how the presence of the mass distribution affects the HCS in the surrounding empty space; it is a purely general-relativistic consequence. However, this does not mean that the usual ES are not a correct solution of the Einstein field equation. In fact they are. In light of the principles in general relativity which permits us to use arbitrary coordinate systems, the usual ES in HCS is an exterior solution in a new coordinate system X_μ (2)–(5) with $R(r)=r-M$. But the new coordinate system is not the

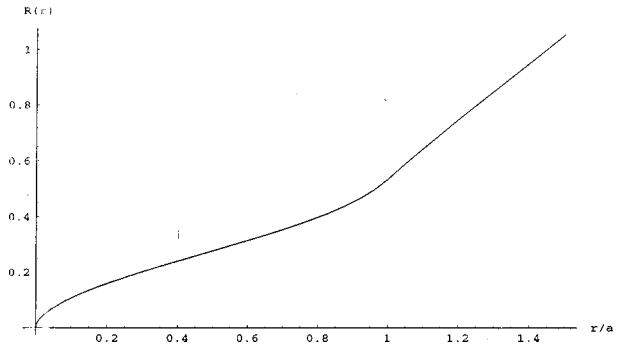

FIG. 1. The curve showing the function relation between $R(r)$ and r in the interval [0,1.5a].

complete HCS at all; and surprisingly, all literature used the HCS in this incomplete sense when an explicit form of HCS was needed.[1–5,8] Finally, we would like to point out that sometimes the neglect of the term may not have a significant influence in physics. Even at the surface of our simplified uniform density star model, only in a percentage point of the results can this term play a role, and the proportion of the second term to the first is approximately 0.04.

ACKNOWLEDGMENTS

I am grateful to Professor Huan-Wu Peng for drawing my attention to the importance of the second term in Eq. (13) and for his encouragement. I also thank Professors Ou-Yang Zhongcan, Yuan-Zhong Zhang, Chao-Guang Huang, Yong-Qiang Yu and Shang-Wu Qian for their helpful discussions. This work was supported by the National Nature Science Foundation of China.

[1] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972), pp. 176–181, 334.
[2] V. Fock, *The Theory of Space, Time and Gravitation* (translated by N. Kemmel) (Pergamon, New York, 1959).
[3] P. Y. Zhou, Sci. Sin. **25**, 628 (1982).
[4] M. Abe, S. Ichinose, and N. Nakanishi, Prog. Theor. Phys. **78**, 1186 (1987); Q. H. Liu, Chin. Phys. Lett. **15**, 313 (1998).
[5] S. G. Turyshev, Can. J. Phys. **74**, 17 (1996), and references therein.
[6] C. M. Will and A. G. Wiseman, Phys. Rev. D **54**, 4813 (1996) and references therein.
[7] A. Galperin, E. Ivanov, and O. Ogievetsky, Ann. Phys. **230**, 201 (1994).
[8] G. W. Gibbons, Branes as BIons, hep-th/9803203, 1998; Rue-Ron Hsu, Bokai Yang, and Chin-Rong Lee, Chin. J. Phys. **34**, 1285 (1996); K. V. Kuchar and C. G. Torre, Phys. Rev. D **44**, 3116 (1991).
[9] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, New York, 1973), p. 553.
[10] R. C. Tolman, *Relativity, Thermodynamics and Cosmology* (Clarendon, Oxford, 1934), pp. 245–247.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 11 NOVEMBER 1998

Applications of topological *-algebras of unbounded operators

F. Bagarello[a)]
Dipartimento di Matematica ed Applicazioni, Facoltà di Ingegneria, Università di Palermo, I-90128 Palermo, Italy

(Received 6 April 1998; accepted for publication 4 June 1998)

In this paper we discuss some physical applications of topological *-algebras of unbounded operators. The first example is a simple system of free bosons. Then we consider other models related to this one. Finally, we discuss the time evolution of two interacting models of matter and bosons. We show that for all these systems it is possible to build up a common mathematical framework where the existence of the thermodynamical limit of the algebraic dynamics can be conveniently analyzed.
© *1998 American Institute of Physics.* [S0022-2488(98)00111-X]

I. INTRODUCTION

Since the first appearance of the algebraic approach to the quantum field theory,[1] many new results have been found and discussed in the literature showing that the structure of C^*-algebras is not rich enough in many relevant physical situations. Difficulties arise already in ordinary quantum mechanics, since the commutation rule $[x,p]=i$ implies that the operators x and p cannot be both bounded as operators on $\mathscr{L}^2(\mathbf{R})$. Obviously, even more involved is the situation for models with infinite degrees of freedom. These physical difficulties have originated much literature about unbounded operators and possible extensions of C^*-algebras: Lassner's quasi *-algebras,[2] partial *-algebras introduced by Antoine and Karwowsky,[3] CQ^*-algebras recently introduced by Trapani and the author,[4] etc. Most of these results are collected, in particular, in two books: the first, essentially mathematically oriented, was written in 1989 by Schmüdgen;[5] the second, which also contains some links with physics, was published one year later by Dubin and Hennings.[6]

These works, however, with few exceptions are mostly concerned with purely mathematical aspects of the problem, while physical applications are missing. In other words, over the years many powerful tools have been proposed, constructed, and refined, but not many relevant physical problems have been solved or even considered using these tools. The present paper tries to make a step to fill this gap. We use topological *-algebras of unbounded operators, of the kind introduced first by Lassner,[2] to show the existence of the algebraic dynamics of several physical models. We begin with an apparently innocuous model of free bosons, where the difficulty of dealing with unbounded operators already appears, and we show how this model can be regularized, using a certain cutoff, and then we prove the existence of the algebraic dynamics of the model when this cutoff is *cleverly* removed. This is the content of the first part of Sec. III. In the second part of the same section we analyze different models which are related to the free bosons by simple transformations. In Sec. IV we first consider an interacting model of single mode bosons interacting with the matter, mimicked by spin observables localized at different lattice sites, and we show again the possibility of regularizing the unbounded Hamiltonian and of finding a limit of the time evolution in a natural topology, similar to that introduced by Lassner.[2] Then we extend our procedure to a generalized version of the same model, where different kinds of bosons are considered. Section V contains the outcome of the paper. We begin, in the following section, with fixing the mathematical structure of our approach. Some of the mathematical details will be analyzed in more detail also in the Appendices.

Before starting a remark is in order: the topological structure of the algebra of the unbounded operators is not always essential. In fact, depending on the approach used for removing the cutoff,

a)Electronic mail: Bagarello@ipamat.math.unipa.it

 © 1998 American Institute of Physics

it can be sufficient to work in an algebra without making reference to any topology. This has been done, for instance, in Refs. 7 and 8, where the limit of several correlation functions of the time evolution of local and delocalized observables has been considered.

II. THE MATHEMATICAL FRAMEWORK

In this paper we will mainly consider a single mode boson oscillator. Here a and $a^\dagger$ are, respectively, the annihilation and the creation operators, satisfying the canonical commutation relation (CCR)

$$[a,a^\dagger]=\mathbb{1}. \tag{2.1}$$

We introduce in a canonical way the Hilbert space of the radiation: we consider a vector Φ_0 which is annihilated by a, and then we build up the linear span of the vectors $\{(a^\dagger)^n\Phi_0, n\in\mathbf{N}\}$, $\mathscr{D}_0$. Taking the completion of this set, we obtain the Fock space of the bosons, $\mathscr{H}$.

Let $N_0\equiv a^\dagger a$. This is a symmetric operator on $\mathscr{D}_0$ which can be uniquely extended to a self-adjoint operator N. Let $D(N^k)$ be the domain of the operator N^k, $k\in\mathbf{N}$, and $\mathscr{D}$ the domain of all the powers of N:

$$\mathscr{D}\equiv D^\infty(N)=\cap_{k\geqslant 0}D(N^k). \tag{2.2}$$

This set is dense in $\mathscr{H}$. Starting from $\mathscr{D}$ we can define, following Lassner,[2] the *-algebra $\mathscr{L}^+(\mathscr{D})$ of all the closable operators defined on $\mathscr{D}$ which, together with their adjoints, map $\mathscr{D}$ into itself. It is clear that all powers of a and $a^\dagger$ belong to this set.

In Ref. 2, the topological structures of both $\mathscr{D}$ and $\mathscr{L}^+(\mathscr{D})$ are discussed in detail; in particular, in $\mathscr{D}$ the topology is defined by the following N-depending seminorms:

$$\phi\in\mathscr{D}\rightarrow\|\phi\|_n\equiv\|N^n\phi\|, \tag{2.3}$$

where n is a natural integer, while $\|\ \|$ is the norm of $\mathscr{H}$. The topology in $\mathscr{L}^+(\mathscr{D})$ is introduced in the following way. We start defining the set $\mathscr{C}$ of all the positive, bounded and continuous functions $f(x)$ on $\mathbf{R}_+$, which are decreasing faster than any inverse power of x. The seminorms on $\mathscr{L}^+(\mathscr{D})$ are labeled by functions in $\mathscr{C}$ and by the integers $\mathbf{N}$. We have

$$X\in\mathscr{L}^+(\mathscr{D})\rightarrow\|X\|^{f,k}\equiv\max\{\|f(N)XN^k\|,\|N^kXf(N)\|\}. \tag{2.4}$$

Here $\|\ \|$ is the usual norm in $B(\mathscr{H})$. We use for this norm the same notation as in Eq. (2.3) since there is no possibility of confusion. Incidentally, we see that a possible way to look at $\mathscr{L}^+(\mathscr{D})$ is as the set of operators such that both $f(N)XN^k$ and $N^kXf(N)$ are bounded for all integers k and for all functions f in $\mathscr{C}$. Moreover, we observe that definition (2.4) easily implies the following invariance property: $\|X\|^{f,k}=\|X^\dagger\|^{f,k}$. We call τ the topology on $\mathscr{L}^+(\mathscr{D})$ defined by the seminorms (2.4). In Ref. 2 it has been proved that $\mathscr{L}^+(\mathscr{D})[\tau]$ is a complete locally convex topological *-algebra.

In the following sections a useful role will be played by some subspaces of $\mathscr{H}$ that we now introduce. We start by defining the one-dimensional spaces $\mathscr{E}_l$ as the set of all the vectors which are proportional to $(a^\dagger)^l\Phi_0$. Then we introduce the $(L+1)$-dimensional space $\mathscr{F}_L\equiv\mathscr{E}_0\oplus\mathscr{E}_1\oplus\cdots\oplus\mathscr{E}_L$, as the direct sum of the first $L+1$ spaces $\mathscr{E}_l$.

These spaces are in one-to-one correspondence with some projection operators. Let $N=\Sigma_{l=0}^\infty l\Pi_l$ be the spectral decomposition of the number operator N. The operators Π_l are projection operators, as well as the operators $Q_L=\Sigma_{l=0}^L\Pi_l$. The following properties are obvious:

$$\Pi_k\Pi_l=\delta_{kl}\Pi_l,\quad \Pi_k^\dagger=\Pi_k, \tag{2.5}$$

and

$$Q_LQ_M=Q_L\quad \text{if } L\leqslant M,\quad Q_L^\dagger=Q_L. \tag{2.6}$$

We see that $\Pi_k:\mathscr{H}\rightarrow\mathscr{E}_k$, while $Q_L:\mathscr{H}\rightarrow\mathscr{F}_L$. From Eqs. (2.5) and (2.6), or alternatively from the definitions of $\mathscr{E}_k$ and $\mathscr{F}_L$, we deduce that the spaces $\mathscr{E}_k$ are mutually orthogonal, while the $\mathscr{F}_L$ are not. Of course, the following inclusions of spaces hold:

$$\mathscr{E}_l \subset \mathscr{F}_L \subset \mathscr{F}_{L+1} \subset \mathscr{D} \subset \mathscr{H}, \tag{2.7}$$

whenever $l \leqslant L$.

In this paper we will not use another possibility which sometimes can be of a certain interest: this consists of considering, instead of $\mathscr{L}^+(\mathscr{D})$, the set $\mathscr{L}(\mathscr{D},\mathscr{D}')$ of the continuous linear maps from $\mathscr{D}$ into its dual $\mathscr{D}'$. Also in this set a topology can be defined,[2] which makes it a complete locally convex space, but the lack of a convenient algebraic structure makes this space of little use to us. In other words, while the problem of finding the time evolution of the product AB when both A and B belong to $\mathscr{L}^+(\mathscr{D})$ is well posed, this is not so in general if A and B are both elements of $\mathscr{L}(\mathscr{D},\mathscr{D}')$. As it is well known, in this case only partial multiplications could be defined, between elements in $\mathscr{L}^+(\mathscr{D})$ and elements of $\mathscr{L}(\mathscr{D},\mathscr{D}')$. A rather complete review on this and related subjects can be found in Ref. 9.

Some final comments concerning the seminorms in (2.4). The first remark is that the two contributions in the definition are exactly of the same form. The estimate of $\|f(N)XN^k\|$ is very similar to the estimate of $\|N^kXf(N)\|$. This is why in the following we will only consider the first of these contributions, identifying from now on $\|X\|^{f,k}$ simply with $\|f(N)XN^k\|$.

With this in mind, we came to the second remark, which is a consequence of the spectral decomposition for N. Using $N=\Sigma_{l=0}^{\infty} l\Pi_l$, we see that the seminorms can be written as follows:

$$X \in \mathscr{L}^+(\mathscr{D}) \to \|X\|^{f,k} = \sum_{l,s=0}^{\infty} f(l)s^k \|\Pi_l X \Pi_s\|. \tag{2.8}$$

We will denote by *physical* the topology τ generated by these seminorms, following, with some freedom, the language usually adopted in the literature.

Since in Sec. IV we will discuss an interacting model with bosons and spins, we introduce here some more relevant spaces and topologies. We refer to Ref. 10 for further details.

Let $\mathbb{Z}^d$ be a d-dimensional infinite lattice, $V \subset \mathbb{Z}^d$, and $|V|$ the number of the points of V. We call $\mathscr{A}_V$ the C^*-algebra generated by the spin operators σ_α^i, $i \in V$ and $\alpha = x,y,z$, and $\mathscr{A}_0$ the norm closure of $\cup_V \mathscr{A}_V$.

We call *relevant* a state ω over $\mathscr{A}_0$ if, denoting by $\mathscr{H}_\omega$ the Hilbert space defined by the Gel'fand–Naimark–Segal (GNS) construction on $\mathscr{A}_0$ and ω, and by Ψ_ω the vector which represents ω in $\mathscr{H}_\omega$ [$\omega(A) = \langle \Psi_\omega, \pi_\omega(A)\Psi_\omega \rangle$, for all $A \in \mathscr{A}_0$], then Ψ_ω belongs to the following set $\mathscr{F}$:

$$\mathscr{F} = \left\{ \Psi \in \mathscr{H}_\omega : \lim_{|V|,\infty} \frac{1}{|V|} \sum_{p \in V} \Pi_\omega(\sigma_3^p)\Psi = \Pi_\omega(\sigma_3^\infty)\Psi, \quad \|\Pi_\omega(\sigma_3^\infty)\| \leqslant 1 \right\}, \tag{2.9}$$

where σ_3^∞ belongs to the center of the algebra $\mathscr{A}_0$ and π_ω is the canonical representation for the spin algebra.[10]

The relevance of this set has been discussed in Ref. 10, where it has been proven, among other things, that all the powers of σ_3^V converge in the $\mathscr{F}$-strong topology (the strong topology ''restricted'' to those vectors which belong to $\mathscr{F}$), as well as the analytic functions of σ_3^V. This reflects the well known fact that σ_3^V does not converge in norm for $|V|$ increasing.[11] The topology is therefore defined by the following seminorms:

$$X \in \mathscr{A}_0 \simeq B(\mathscr{H}_{\text{spin}}) \to \|X\|^\Psi \equiv \|X\Psi\|, \tag{2.10}$$

where $\Psi \in \mathscr{F}$, X is identified with its ''canonical'' representative in $B(\mathscr{H}_{\text{spin}})$, $\mathscr{H}_{\text{spin}}$ being the infinite tensor product of two-dimensional complex spaces $\mathbf{C}_i^2$, $i \in \mathbb{Z}^d$. In this way $\mathscr{A}_0$ is canonically identified with its representative $B(\mathscr{H}_{\text{spin}})$. This identification will be used all throughout this paper in order to simplify the notation.

III. THE FREE BOSONS AND RELATED MODELS

A. The free model

In this section we propose a possible rigorous algebraic treatment of the existence of the time evolution of a single mode of radiation described by the following free Hamiltonian:

$$H=a^{\dagger}a, \tag{3.1}$$

where a and $a^{\dagger}$ are the boson operators introduced in Sec. II. Throughout this paper we will follow the general strategy usually adopted in the literature, whenever a certain mathematical rigor is required: first we regularize the Hamiltonian introducing a certain cutoff; then we obtain the equations of motion which are solved keeping this cutoff fixed, and finally we remove the cutoff. If all these steps can be performed, we define the dynamics for the ''infinite'' model as the limit of these cutoff dynamics.

The first nontrivial problem is the way of implementing this cutoff. As it is obvious from the expression of H, which is nothing but the number operator N, H is unbounded. Nevertheless, if we consider N as acting not on the whole $\mathscr{H}$ but only on some (cleverly chosen) subspaces of the Hilbert space, the unboundedness can be controlled. In fact, it is easily seen that $\|N\varphi\|<c_{\varphi}$ whenever φ is taken in any of the spaces $\mathscr{E}_l$ or $\mathscr{F}_L$, where c_{φ} is a positive constant depending on φ (and therefore on l and L). This fact suggests one consider the following regularization:

$$H_L=Q_LHQ_L\,. \tag{3.2}$$

This choice preserves the Hermitianness of the Hamiltonian and gives the following property: $\forall\varphi\in\mathscr{H}\Rightarrow H_L\varphi=Q_LH\varphi_L$, where $\varphi_L\in\mathscr{F}_L$. Therefore $\forall\varphi\in\mathscr{H}$, $\|H_L\varphi\|<c_{\varphi_L}$. However, this regularization has an unpleasant drawback already evident for the model we will discuss in some detail in Sec. IV, and which is described by the following Hamiltonian,

$$H_V=\frac{1}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+a^{\dagger}a+\gamma(a+a^{\dagger})\sigma_3^V,$$

where $\sigma_3^V=(1/|V|)\Sigma_{i\in V}\sigma_3^i$. In fact, it is clear that, after the regularization procedure (3.2), the following term would appear: $(1/|V|)\Sigma_{i,j\in V}\sigma_3^i\sigma_3^jQ_L$. This contribution is not certainly natural, and it is also technically difficult to be properly considered, since the boson and spin variables appear together already in the free spin Hamiltonian.

All these considerations suggest the use of another regularization, defined starting from the boson operators themselves. We perform the following substitutions:

$$a\to a_L\equiv Q_LaQ_L\,,\quad a^{\dagger}\to a_L^{\dagger}\equiv Q_La^{\dagger}Q_L\,, \tag{3.3}$$

and then we define, for the model (3.1),

$$H_L=a_L^{\dagger}a_L\,. \tag{3.4}$$

This procedure has a natural generalization: if H is a given function χ of certain nonbosonic (bounded) operators $X_1,\ldots,X_n$, as well as of a and $a^{\dagger}$, possibly already depending on some cutoff as the Hamiltonian H_V above, we can consider the *regularized* Hamiltonian H_L as the same function of the same variables $X_1,\ldots,X_n$ and of the regularized boson operators a_L and of $a_L^{\dagger}$:

$$H=\chi(X_1,\ldots,X_n,a,a^{\dagger})\to H_L=\chi(X_1,\ldots,X_n,a_L,a_L^{\dagger}). \tag{3.5}$$

Remark: Using the results of Appendix A it is easy to show that for the free Hamiltonian $H=N$ the above regularizations coincide. Indeed, by Eqs. (A2) and (2.6), we have

$$a_L^{\dagger}a_L=Q_La^{\dagger}Q_LQ_LaQ_L=Q_La^{\dagger}Q_LaQ_L=Q_La^{\dagger}aQ_{L+1}Q_L=Q_LNQ_L\,.$$

This is not true for general models, as the spin-bosons Hamiltonian above explicitly shows.

Once we have fixed the regularization procedure, we come to the main problem of this section, that is to prove the existence of the limit, for L going to infinity, of the following algebraic dynamics:

$$\alpha_L^t(X)=e^{iH_Lt}Xe^{-iH_Lt},$$

where X is an element of $\mathscr{L}^+(\mathscr{D})$. The topology in which this limit will be considered is the physical one defined in the previous section, which makes $\mathscr{L}^+(\mathscr{D})$ a complete topological *-algebra.

Proposition 3.1: The limits of $\alpha_L^t(a^n)$ and $\alpha_L^t((a^\dagger)^n)$ exist in $\mathscr{L}^+(\mathscr{D})[\tau]$ for all natural n.

Proof: We prove the Proposition only for the annihilation operator a. The proof for $a^\dagger$ is essentially the same. We proceed by induction on n. We start therefore by proving the statement for $n=1$.

The first step is to observe that:

$$[H_L,a]_m=(L\Pi_L-Q_{L-1})^m a=(L^m\Pi_L+(-1)^m Q_{L-1})a. \tag{3.6}$$

Again, the proof of these formulas goes on by induction on m, and is a simple consequence of the commutation relations proved in Appendix A. The equality $(L\Pi_L-Q_{L-1})^m=(L^m\Pi_L+(-1)^m Q_{L-1})$ is a consequence of the properties of the projection operators Π_l and Q_L. In particular, we use the equality $\Pi_L Q_{L-1}=0$, which follows from Eq. (2.5) and from the definition of Q_L.

By means of (3.6) we can prove that

$$\alpha_L^t(a)=e^{iH_Lt}ae^{-iH_Lt}=(1-Q_L+e^{itL}\Pi_L+e^{-it}Q_{L-1})a. \tag{3.7}$$

In fact, since H_L is bounded, we have

$$\alpha_L^t(a)=\sum_{m=0}^{\infty}\frac{(it)^m}{m!}[H_L,a]_m=\sum_{m=0}^{\infty}\frac{(it)^m(L^m\Pi_L+(-1)^m Q_{L-1})a}{m!},$$

so that Eq. (3.7) follows.

Now we are ready to prove the τ-Cauchy nature of $\alpha_L^t(a)$. Using the seminorms (2.8) we have

$$Q_{L,M}(t)\equiv\|\alpha_L^t(a)-\alpha_M^t(a)\|^{f,k}=\sum_{l,s=0}^{\infty}f(l)s^k\|\Pi_l(\alpha_L^t(a)-\alpha_M^t(a))\Pi_s\|.$$

We are not making explicit here the dependence of $Q_{L,M}(t)$ from the seminorm, that is from f and k. Using the explicit expression (3.7) for $\alpha_L^t(a)$ we get the following inequality:

$$\begin{aligned}Q_{L,M}(t)\leqslant\sum_{l,s=0}^{\infty}f(l)s^k[&\|\Pi_l(Q_M-Q_L)a\Pi_s\|+|e^{-it}|\|\Pi_l(Q_{L-1}-Q_{M-1})a\Pi_s\|\\&+|e^{itL}|\|\Pi_l\Pi_L a\Pi_s\|+|e^{itM}|\|\Pi_l\Pi_M a\Pi_s\|].\end{aligned}$$

To conclude we use Eq. (2.5) and formula (A4):

$$\|\Pi_l a\Pi_s\|=\delta_{l,s-1}\sqrt{s}, \tag{3.8}$$

which give the following estimate for $Q_{L,M}(t)$:

$$Q_{L,M}(t)\leqslant 2\sum_{s=M+1}^{L}f(s-1)s^{k+1/2}\to 0, \tag{3.9}$$

when L and M both diverge, due to the nature of the functions of the set $\mathscr{C}$. Since $\mathscr{L}^+(\mathscr{D})$ is τ-complete, the above result implies the existence of the limit of $\alpha_L^t(a)$ in $\mathscr{L}^+(\mathscr{D})[\tau]$.

Remark: As announced in Sec. II, we have focused our attention on the seminorm $\|f(N)\cdots N^k\|$. From the above computations, it appears clear how to extend the same estimates to $\|N^k\cdots f(N)\|$: it goes to zero essentially for the same reasons. This implies, due the invariance of the ''complete'' physical topology, given by the seminorms (2.4), with respect to the adjoint, that also $\alpha_L^t(a^\dagger)$ is τ-Cauchy and therefore τ-convergent in $\mathscr{L}^+(\mathscr{D})[\tau]$.

At this point we can prove the second step of the induction. Hence we suppose that $\alpha_L^t(a^{n-1})$ is τ-Cauchy and we also prove that $\alpha_L^t(a^n)$ is τ-Cauchy. The proof goes as follows:

$$Q^n_{L,M}(t)=\|\alpha^t_L(a^n)-\alpha^t_M(a^n)\|^{f,k}\leqslant\|\alpha^t_L(a)((\alpha^t_L(a))^{n-1}-(\alpha^t_M(a))^{n-1})\|^{f,k}$$
$$+\|(\alpha^t_L(a)-\alpha^t_M(a))(\alpha^t_M(a))^{n-1}\|^{f,k}. \tag{3.10}$$

The first contribution can be estimated by the seminorm $\|(\alpha^t_L(a^{n-1})-\alpha^t_M(a^{n-1}))\|^{xf,k}$, which we know by the induction hypothesis to be converging to zero since, if $f(x)\in\mathscr{C}$, then $xf(x)\in\mathscr{C}$ as well. To show this, we call $F_L(t)\equiv 1-Q_L+e^{itL}\Pi_L+e^{-it}Q_{L-1}$, so that $\alpha^t_L(a)=F_L(t)a$. We see that $\|F_L(t)\|\leqslant 3$ and that $F_L(t)$ commutes with Π_l (and Q_L). Therefore the first contribution in (3.10) can be estimated with

$$3\sum_{l,s=0}^{\infty} f(l)s^k\|a\Pi_{l+1}\|\|\Pi_{l+1}((\alpha^t_L(a))^{n-1}-(\alpha^t_M(a))^{n-1})\Pi_s\|,$$

and our claim follows resumming over s and l.

The second contribution in (3.10) can be estimated with similar techniques. Using formula (2.8) for the seminorms, $[F_L(t),\Pi_l]=0$ and formula (A1), we get

$$\begin{aligned}
&\|\Pi_l(\alpha^t_L(a)-\alpha^t_M(a))(F_M(t)a)^{n-1}\Pi_s\|\\
&\quad=\|\Pi_l(\alpha^t_L(a)-\alpha^t_M(a))\Pi_{s-(n-1)}F_M(t)a\Pi_{s-(n-2)}F_M(t)a\cdots\Pi_{s-1}F_M(t)a\Pi_s\|\\
&\quad\leqslant\|\Pi_l(\alpha^t_L(a)-\alpha^t_M(a))\Pi_{s-n+1}\|3^{n-1}\sqrt{(s-n+2)(s-n+3)\ldots(s-2)(s-1)s}\\
&\quad\leqslant(3\sqrt{s})^{n-1}\|\Pi_l(\alpha^t_L(a)-\alpha^t_M(a))\Pi_{s-n+1}\|.
\end{aligned}$$

Here we have also used the estimate $\|a\Pi_l\|\leqslant\sqrt{l}$. After some more easy steps, we finally find

$$\|(\alpha^t_L(a)-\alpha^t_M(a))(\alpha^t_M(a))^{n-1}\|^{f,k}\leqslant 2\sum_{l=M}^{L} f(l)(l+n)^k\sqrt{l+1}(3\sqrt{l+n})^{n-1},$$

which, again, goes to zero for any fixed n when L and M go to infinity, due to the decay properties of the function $f\in\mathscr{C}$. □

B. Variations on the same theme

In this section we briefly discuss two more models which can be easily ''transformed'' into the model (3.1).

The first model is simply a linear perturbation of the original one, and is described by the following Hamiltonian:

$$H=a^\dagger a+\gamma(a+a^\dagger),$$

where γ is real. As it is well known, the above model can be rewritten as a free model in terms of the different boson operators

$$b\equiv a+\gamma.$$

The operator b and its adjoint $b^\dagger$ satisfy the canonical commutation relation, as well as a and $a^\dagger$. Furthermore, in terms of b, we can write $H=b^\dagger b-\gamma^2$. This suggests how to proceed: We introduce the number operator for b, $N_b=b^\dagger b$, and then we consider its spectral decomposition, instead of that of $N_a=a^\dagger a$. Now we can proceed as in the previous section. Every step can be repeated also in this case, so that, again, it is possible to define rigorously the model first introducing, and then removing, the cutoff $b\to b_L=Q_L bQ_L$.

The second model we are going to analyze presents some more interesting aspects. It is a model of two modes of radiation coupled to each other. We assume for it the following Hamiltonian

$$H=a^\dagger a+b^\dagger b+(a^\dagger b+b^\dagger a).$$

Here a and b both satisfy the CCR as well as $[a^\sharp,b^\sharp]=0$, where with $X^\sharp$ we denote X or $X^\dagger$.

Again, the Hamiltonian can be mapped into something very similar to a number operator. We introduce the linear combinations

$$A=\frac{a+b}{\sqrt{2}}, \quad B=\frac{a-b}{\sqrt{2}}. \tag{3.11}$$

With these definitions we get $H=2A^\dagger A$. The commutation relations are

$$[A,A^\dagger]=1, \quad [B,B^\dagger]=1, \quad [A^\sharp,B^\sharp]=0.$$

We can consider now the spectral decomposition of the operator $N_A=A^\dagger A=\Sigma_{l=0}^{\infty} l\Pi_l$, and from the above commutation relations it is clear that Π_l also commutes with B and $B^\dagger$. At this point, with the usual regularization $H\to H_L=2A_L^\dagger A_L$, where $A_L=Q_L A Q_L$ and $Q_L=\Pi_0+\Pi_1+\cdots+\Pi_L$, the limit for $L\to\infty$ of $\alpha_L^t(A^n)$ exists in its natural physical topology τ. Furthermore, it is easily checked that $\alpha_L^t(B^n)=B^n$ for all integers n, so that its limit for L diverging exists as well. Therefore inverting Eqs. (3.11), we also find that $\alpha_L^t(a^n)$ and $\alpha_L^t(b^n)$ do converge in $\mathscr{L}^+(\mathscr{D})[\tau]$, for all integers n.

IV. TWO INTERACTING MODELS

A. Single mode bosons

In this section we will show how to use the framework introduced in Sec. II to discuss the existence of the algebraic dynamics for the model described by the following finite-volume Hamiltonian, already introduced in Sec. III:

$$H_V'=\frac{J}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+a^\dagger a+\gamma(a+a^\dagger)\sigma_3^V. \tag{4.1}$$

Here $\sigma_3^V=(1/|V|)\Sigma_{i\in V}\sigma_3^i$. This is a prototype model of an interaction between the matter, considered as a family of two levels atoms, and conveniently mimicked using spin matrices, with a single boson mode.

We take γ as a ''small'' real parameter, in a sense which will appear clear in the following, and V as a subset of the infinite lattice $\mathbb{Z}^d$. We remind the reader that the Pauli spin matrices satisfy the following commutation rules:

$$[\sigma_\alpha^i,\sigma_\beta^j]=2i\epsilon_{\alpha\beta\gamma}\delta_{i,j}\sigma_\gamma^i. \tag{4.2}$$

A consequence of this commutation relation, together with the definition of σ_3^V, is that

$$[\sigma_3^V,\sigma_\beta^j]=\frac{2i\epsilon_{3\beta\gamma}}{|V|}\sigma_\gamma^i, \tag{4.3}$$

which shows already that, in the limit $|V|\to\infty$, σ_3^V commutes with all the other observables.

It is worth remarking that a volume cutoff appears in H_V'. The necessity of this cutoff in a rigorous treatment of the mean field version of the BCS model for the high temperature superconductivity has been discussed, for instance, in Ref. 11.

Assuming that the perturbation $\gamma(a+a^\dagger)\sigma_3^V$ is small compared with the energy difference between the energy levels of the harmonic oscillator, we can consider ''frozen'' the boson kinetic contribution in H_V'. In this approximation we consider

$$H_V=\frac{J}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+\gamma(a+a^\dagger)\sigma_3^V. \tag{4.4}$$

What we have done is the same kind of approximation which leads to the Hamiltonian for the fractional quantum hall effect (FQHE) starting with an interacting two-dimensional electron gas in the presence of a strong magnetic field. In that system, the Coulomb interaction between the

electrons is considered as a small perturbation of the Hamiltonian of the free model, which consists in a gas of noninteracting electrons localized in the plane $(0;x,y)$ and subjected to a strong constant magnetic field directed along z. The free Hamiltonian of each electron is, in convenient variables, the Hamiltonian of an harmonic oscillator, and the free Hamiltonian of the system is the sum of all these contributions, one for each electron. The assumption on the Coulomb interaction suggests that, without other interactions, all the electrons are in the lowest Landau level (ground state of the oscillators) so that the kinetic part of the complete Hamiltonian (free Hamiltonian plus Coulomb interaction) can be replaced by its mean value in the ground state of the system. The idea of considering only the lowest Landau level in the description of the FQHE is now very common in the literature and was originally proposed by Yoshioka and Fukuyama in Ref. 12.

Before describing our treatment of the dynamics, it is worth remarking on the likeness of this model with the dissipative laser model introduced in Ref. 7. Of course, the present model is quite a simplified version of that one, first for its conservative nature, which allows a Hamiltonian approach, second for the presence of only one mode for the bosons (a condition which will be relaxed in the next section), and finally for the simplified form of the interaction Hamiltonian. As it has been shown in Refs. 7 and 8, a complete analysis of the Alli-Sewell model can be carried out already in a nontopological *-algebra of unbounded operators. In this paper we want to play a similar game with rather different rules, that is we want to introduce a topology related in some way to the model itself.

The way of doing this is suggested by the results contained in Sec. III and in Ref. 10: the algebra $\mathscr{A}$ should reflect the composite nature of the model, and is therefore reasonable to take $\mathscr{A}$ as the tensor product of the bounded spin algebra, $B(\mathscr{H}_{\rm spin})$, and the Lassner boson algebra, $\mathscr{L}^+(\mathscr{D})$: $\mathscr{A}=B(\mathscr{H}_{\rm spin})\otimes\mathscr{L}^+(\mathscr{D})$. The topology on $\mathscr{A}$, τ, is generated by seminorms which deal separately with the spin and the boson variables. Obviously, in view of the results in Sec. III for the free bosons and what is known about the mean field spin models,[10] we take these seminorms essentially as the ones in (2.8) for the boson operators and as the strong ones in (2.10) for the spin observables. We give here the definition of these seminorms which will be discussed in more detail in Appendix B. First of all, following Refs. 10 and 13, we give the definition of what we will call relevant vectors in the present contest. It may be worthwhile to observe that the set below is slightly different from the one introduced in Sec. II:

$$\mathscr{T}=\left\{\Psi\in\mathscr{H}_{\rm spin}:\lim_{|V|,\infty}\frac{1}{|V|}\sum_{p\in V}\sigma_3^p\Psi=\sigma_3^\infty\Psi,\|\sigma_3^\infty\|\leqslant 1\right\},\tag{4.5}$$

σ_3^∞ being an element of the center of the C^*-spin algebra. The usefulness of this set relies on the fact that σ_3^V (together with its powers and its analytical functions) converges strongly only on the vectors of $\mathscr{T}$, not on general vectors of $\mathscr{H}_{\rm spin}$ (and, of course, not in norm).

For future convenience, it is better to introduce the unbounded operator

$$M\equiv N+\mathbb{1}=a^\dagger a+\mathbb{1}=\sum_{l=1}^{\infty} l\Pi_l.\tag{4.6}$$

The only difference between the operators M and N is in the lower value of l in their spectral decompositions. With these considerations we define, for each $X\in\mathscr{L}^+(\mathscr{D})$ and for each $A\in B(\mathscr{H}_{\rm spin})$,

$$\|XA\|^{f,k,\Psi}\equiv\|X\|^{f,k}\|A\|^\Psi=\|f(M)XM^k\|\|A\Psi\|,\tag{4.7}$$

where $f\in\mathscr{C}$, $k\in\mathbf{N}$, $\Psi\in\mathscr{T}$. The final remark is that, again, we are considering only one contribution in the Lassner seminorm; a complete definition should also include $\|M^kXf(M)\|$, see (2.4).

Now we proceed to a complete regularization of the Hamiltonian H_V. The approach we follow is the one already discussed in Sec. III. Therefore we consider the following operator

$$H_{V,L}=J|V|(\sigma_3^V)^2+\gamma(a_L+a_L^\dagger)\sigma_3^V,\tag{4.8}$$

where, as before, $a_L = Q_L a Q_L = a Q_L$, see (A2). Then we define the cutoff algebraic dynamics $\alpha^t_{V,L}$ in a canonical way:

$$\alpha^t_{V,L}(X) = e^{iH_{V,L}t} X e^{-iH_{V,L}t}, \tag{4.9}$$

where $X \in \mathscr{A}$.

For technical convenience it is useful to introduce here a relation between the two different cutoffs. It will be used in the second part of Proposition 5.1, where we will assume that

$$|V| = L^r. \tag{4.10}$$

Here r is an integer bigger than 1. This condition essentially restricts the subspaces $V \subset \mathbb{Z}^d$ which are to be used in the regularization procedure: they are those finite lattices whose number of sites is the rth power of the integers.

We can now state the main result of this section:

Proposition 5.1: The limit of $\alpha^t_{V,L}(a)$ for $|V|$ and L both diverging exists in $\mathscr{A}[\tau]$. In the hypothesis (4.10) the same holds true also for $\alpha^t_{V,L}(\sigma^i_\alpha)$.

Proof: We begin with the proof of the first statement above. Since $[J|V|(\sigma^V_3)^2,\ \gamma(a_L + a_L^\dagger)\sigma^V_3]] = 0$, we can write

$$\begin{aligned}\alpha^t_{V,L}(a) &= e^{i\gamma(a_L + a_L^\dagger)\sigma^V_3 t}(e^{iJ|V|(\sigma^V_3)^2 t} a e^{-i,J|V|(\sigma^V_3)^2 t}) e^{-i\gamma(a_L + a_L^\dagger)\sigma^V_3 t} \\ &= e^{i\gamma(a_L + a_L^\dagger)\sigma^V_3 t} a e^{-i\gamma(a_L + a_L^\dagger)\sigma^V_3 t} = \sum_{n=0}^{\infty} \frac{(i\gamma t \sigma^V_3)^n}{n!} [a_L + a_L^\dagger, a]_n, \end{aligned} \tag{4.11}$$

due to the commutation relation $[a, \sigma^i_\alpha] = 0$. Here $[a_L + a_L^\dagger, a]_n$ is the usual multiple commutator defined as

$$[a_L + a_L^\dagger, a]_0 = a, \quad [a_L + a_L^\dagger, a]_n = [(a_L + a_L^\dagger), [a_L + a_L^\dagger, a]_{n-1}], n \geq 1.$$

We divide the infinite sum in the right-hand side of (4.11) in two contributions:

$$\alpha^t_{V,L}(a) = \sum_{n=0}^{1} \frac{(i\gamma t \sigma^V_3)^n}{n!} [a_L + a_L^\dagger, a]_n + \sum_{n=2}^{\infty} \frac{(i\gamma t \sigma^V_3)^n}{n!} [a_L + a_L^\dagger, a]_n, \tag{4.12}$$

and we show that the last term, $\Sigma_{n=2}^{\infty} \ldots$, converges to zero in τ when L goes to infinity, for any $|V|$.

The proof of this convergence goes as follows:

(a) First we compute, using the formulas in Appendix A, the following multiple commutators:

$$\begin{aligned}&[a_L + a_L^\dagger, a]_0 = a, \quad [a_L + a_L^\dagger, a]_1 = a^2 \Pi_{L+1} + a a^\dagger \Pi_L - Q_L, \\ &[a_L + a_L^\dagger, a]_2 = a^3 \Pi_{L+1} + a^2 a^\dagger (\Pi_L + \Pi_{L+1}) - 2a\Pi_{L+1} - a(a^\dagger)^2 \Pi_{L-1}. \end{aligned} \tag{4.13}$$

(b) Then we observe that $\|f(M)[a_L + a_L^\dagger, a]_2 M^k\| \to 0$ when $L \to \infty$, for any $f \in \mathscr{C}$ and any $k \in \mathbb{N}$. The proof of this convergence is a bit long. Here we only show the convergence to zero of the first contribution in (4.13). We have

$$\|f(M) a^3 \Pi_{L+1} M^k\| = \sum_{l,s=1}^{\infty} f(l) s^k \|\Pi_l a^3 \Pi_{L+1} \Pi_s\| = \sum_{l,s=1}^{\infty} f(l) s^k \delta_{s,L+1} \|\Pi_l a^3 \Pi_s\|.$$

But, since $\|\Pi_l a \Pi_s\| = \sqrt{s}\,\delta_{l,s-1}$, we deduce that, using the idempotence of Π_l, $\Pi_l = \Pi_l^2$ together with formula (A2),

$$\|\Pi_l a^3 \Pi_s\| \leq \|\Pi_l a \Pi_{l+1}\| \|\Pi_{l+1} a \Pi_{l+2}\| \|\Pi_{l+2} a \Pi_s\| \leq \sqrt{(l+1)(l+2)(l+3)}\,\delta_{s,l+3}.$$

Therefore after some algebraic computations,

$$\|f(M)a^3\Pi_{L+1}M^k\| \leqslant f(L-2)(L+1)^k\sqrt{L(L-1)(L-2)},$$

which, of course, goes to zero for $L\to\infty$. Analogous estimates work for the other contributions in $\|f(M)[a_L+a_L^\dagger,a]_2M^k\|$ in (4.13).

(c) In the third step we show how the previous estimate can be extended to the other multiple commutators $[a_L+a_L^\dagger,a]_n$, $n>2$. The estimate goes as follows: first we observe that $[a_L+a_L^\dagger,a]_2$ is essentially the sum of six monomials (we consider two contributions from $-2a\Pi_{L+1}$) in a and $a^\dagger$, each at most raised to the power 3. Since we are interested in very large values of L, it is not very important that sometimes we find Π_L and somewhere else $\Pi_{L\pm1}$. The structure of this commutator is therefore $[a_L+a_L^\dagger,a]_2\simeq 6m_2(a,a^\dagger)\Pi_L$, where m_2 is a monomial of degree $2+1=3$. Analogously we find $[a_L+a_L^\dagger,a]_3\simeq 24m_3(a,a^\dagger)\Pi_L$ where m_3 is of degree 4. The number 24 is the maximum possible number of contributions following from the commutation between $a_L+a_L^\dagger$ and $6m_2(a,a^\dagger)\Pi_L$. This procedure can be extended to each n, and we find that, for $L\gg 1$,

$$[a_L+a_L^\dagger,a]_n\simeq 6\cdot 4^{n-2}m_n(a,a^\dagger)\Pi_L. \tag{4.14}$$

Using this formula in the expression of $\|[a_L+a_L^\dagger,a]_n\|^{f,k}$, together with the $\Pi_L\Pi_s=\Pi_s\delta_{s,L}$ we obtain the following:

$$\|[a_L+a_L^\dagger,a]_n\|^{f,k}\simeq 6L^k4^{n-2}\sum_{l=1}^{\infty}\|\Pi_l m_n(a,a^\dagger)\Pi_L\|.$$

By similar estimates to those used in point (b) above, we see that $\|\Pi_l m_n(a,a^\dagger)\Pi_L\|$ is essentially a polynomial p_n in l and L, times a delta function $\delta_{l,\varphi_{m_n}}$, where φ_{m_n} is an increasing function related to the explicit expression of m_n. Therefore we get

$$\|[a_L+a_L^\dagger,a]_n\|^{f,k}\simeq 6\cdot 4^{n-2}L^k f(\varphi_{m_n}(L))p_n(\varphi_{m_n}(L),L).$$

Due to the nature of f, the right-hand side is bounded by the supremum on n of the same quantity:

$$L^k f(\varphi_{m_n}(L))p_n(\varphi_{m_n}(L),L)\leqslant\sup_n[L^k f(\varphi_{m_n}(L))p_n(\varphi_{m_n}(L),L)]=:F_k(L),$$

where the function $F_k(L)$ goes to zero when $L\to\infty$. We conclude that

$$\|[a_L+a_L^\dagger,a]_n\|^{f,k}\leqslant 6\cdot 4^{n-2}F_k(L). \tag{4.15}$$

(d) At this point we have at hand all the ingredients to prove that the last contribution in (4.12) converges to zero for any V when L diverges. In fact we have, using the definition (4.7) of the seminorms $\|\ \|^{f,k,\Psi}$, the inequality (4.15), and the uniform bound on σ_3^V, $\|(\sigma_3^V)^n\|\leqslant 1$,

$$\left\|\sum_{n=2}^{\infty}\frac{(i\gamma t\sigma_3^V)^n}{n!}[a_L+a_L^\dagger,a]_n\right\|^{f,k,\Psi}\leqslant\sum_{n=2}^{\infty}\frac{|\gamma t|^n}{n!}\|(\sigma_3^V)^n\Psi\|\|[a_L+a_L^\dagger,a]_n\|^{f,k}$$
$$\leqslant\frac{3}{2}\left(\sum_{n=2}^{\infty}\frac{|4\gamma t|^n}{n!}\right)F_k(L)\leqslant\frac{3}{2}e^{|4t\gamma|}F_k(L)\to 0,$$

when $L\to\infty$.

Now, the first statement of the Proposition, that is the τ-convergence of $\alpha_{V,L}^t(a)$, easily follows. In fact, in view of the above result, we can write equality (4.12) as

$$\alpha_{V,L}^t(a)=a+i\gamma\sigma_3^V t(a^2\Pi_{L+1}+aa^\dagger\Pi_L-Q_L)+\tilde{\alpha}_{V,L}^t(a), \tag{4.16}$$

where $\tilde{\alpha}_{V,L}^t(a)\equiv\Sigma_{n=2}^{\infty}[(i\gamma t\sigma_3^V)^n/n!][a_L+a_L^\dagger,a]_n$ is τ-converging to zero with L, and uniformly bounded in $|V|$, so that

$$\|\|\alpha^t_{V,L}(a)-\alpha^t_{V',L'}(a)\|\|^{f,k,\Psi}\leqslant|\gamma t|\|\|\sigma_3^V t(a^2\Pi_{L+1}+aa^\dagger\Pi_L-Q_L)$$
$$-\sigma_3^{V'}t(a^2\Pi_{L'+1}+aa^\dagger\Pi_{L'}-Q_{L'})\|\|^{f,k,\Psi}+O_L,$$

where $O_L\to 0$ when $L\to\infty$.

In this estimate we can handle very easily all the contributions containing any projection operator Π_n, such as $\|\|\sigma_3^V t a^2\Pi_{L+1}\|\|^{f,k,\Psi}$. In fact, they go to zero, and can be seen following essentially the same steps as in point (b) above. What is to be controlled is only the term $\|\|\sigma_3^V Q_L-\sigma_3^{V'}Q_{L'}\|\|^{f,k,\Psi}$. This can be estimated adding and subtracting first the same quantity $\sigma_3^V Q_{L'}$ and observing that, if $L>L'$, then

$$\|Q_L-Q_{L'}\|^{f,k}=\sum_{l=L'+1}^{L}f(l)l^k\to 0,$$

when both L and L' diverge. We have

$$\|\|\sigma_3^V Q_L-\sigma_3^{V'}Q_{L'}\|\|^{f,k,\Psi}\leqslant\|\|\sigma_3^V(Q_L-Q_{L'})\|\|^{f,k,\Psi}+\|\|(\sigma_3^V-\sigma_3^{V'})Q_{L'}\|\|^{f,k,\Psi}$$
$$=\|\sigma_3^V\Psi\|\|Q_L-Q_{L'}\|^{f,k}+\|(\sigma_3^V-\sigma_3^{V'})\Psi\|\|Q_{L'}\|^{f,k}\to O\quad\text{for } L,|V|\to\infty.$$

The existence of $\alpha^t(a)$ finally follows from the completeness of the algebra $\mathscr{A}$.

Remarks: (1) It is clear from the above procedure that the order of the limits ($|V|\to\infty$ or $L\to\infty$) has no importance. (2) It is worthwhile to observe that hypothesis (4.10) has not been used so far. We will use this condition in the second part of the proof.

Let us now discuss the convergence of $\alpha^t_{V,L}(\sigma^i_\alpha)$. Using the commutativity of the two pieces of the Hamiltonian, formula (2.3) of Ref. 13, and the obvious commutation relation $[\sigma_3^V,\sigma_3^i]=0$, we have

$$\begin{aligned}\alpha^t_{V,L}(\sigma^i_\alpha)&=e^{iH_{V,L}t}\sigma^i_\alpha e^{-iH_{V,L}t}\\&=e^{i\gamma(a_L+a_L^\dagger)\sigma_3^V t}(e^{iJ|V|(\sigma_3^V)^2t}\sigma^i_\alpha e^{-iJ|V|(\sigma_3^V)^2t})e^{-i\gamma(a_L+a_L^\dagger)\sigma_3^V t}\\&=\beta^t_{V,L}(\sigma^i_\alpha)\cos^2(S_3^V)-2\epsilon_{3\alpha\beta}\beta^t_{V,L}(\sigma^i_\beta)\sin(S_3^V)\cos(S_3^V)\\&\quad+\sigma_3^i\beta^t_{V,L}(\sigma^i_\alpha)\sigma_3^i\sin^2(S_3^V)+O(|V|^{-1}),\end{aligned}\tag{4.17}$$

where $S_3^V=2Jt\sigma_3^V$, $O(|V|^{-1})$ is norm converging to zero and we have introduced the notation

$$\beta^t_{V,L}(\sigma^i_\alpha)=e^{i\gamma(a_L+a_L^\dagger)\sigma_3^V t}\sigma^i_\alpha e^{-i\gamma(a_L+a_L^\dagger)\sigma_3^V t}.\tag{4.18}$$

The next step consists in proving that

$$\beta^t_{V,L}(\sigma^i_\alpha)=\sigma^i_\alpha+\tilde\beta^t_{V,L},\tag{4.19}$$

where $\tilde\beta^t_{V,L}$ τ-converges to zero when $|V|$ diverges, under the technical assumption (4.10). This result is reasonable, since $[\sigma_3^V,\sigma^i_\alpha]=O(|V|^{-1})$, so that in definition (4.18) we can commute $e^{i\gamma(a_L+a_L^\dagger)\sigma_3^V t}$ with σ^i_α only within an error going to zero when $|V|$ is sent to infinity. We are going to prove (4.19) rigorously. We begin splitting

$$\beta^t_{V,L}(\sigma^i_\alpha)=\sum_{l=0}^{\infty}\frac{(it\gamma(a_L+a_L^\dagger))^l}{l!}[\sigma_3^V,\sigma^i_\alpha]_l=\sigma^i_\alpha+\sum_{l=1}^{\infty}\frac{(it\gamma(a_L+a_L^\dagger))^l}{l!}[\sigma_3^V,\sigma^i_a]_l.\tag{4.20}$$

We will show that the last term, which is exactly what we have called $\tilde\beta^t_{V,L}$, τ-converges to zero, at least under the assumption (4.10). Since

$$\left\|\left|\sum_{l=1}^{\infty}\frac{(it\gamma(a_L+a_L^\dagger))^l}{l!}[\sigma_3^V,\sigma^i_\alpha]_l\right|\right\|^{f,k,\Psi}\leqslant\sum_{l=1}^{\infty}\frac{|t\gamma|^l}{l!}\|(a_L+a_L^\dagger)^l\|^{f,k}\|[\sigma_3^V,\sigma^i_\alpha]_l\Psi\|,$$

we need to estimate $\|(a_L+a_L^\dagger)^l\|^{f,k}$ and $\|[\sigma_3^V,\sigma_\alpha^i]_l\Psi\|$. This last term is easily seen to be bounded by powers of $2/|V|$:

$$\|[\sigma_3^V,\sigma_\alpha^i]_l\Psi\|\leq\left(\frac{2}{|V|}\right)^l. \tag{4.21}$$

As for the estimate of $\|(a_L+a_L^\dagger)^l\|^{f,k}$, the situation is a bit more difficult. In fact, it is possible to prove the following inequality

$$\|(a_L+a_L^\dagger)^l\|^{f,k}\leq c_k((2^k+1)(L+1)^{3/2})^l, \tag{4.22}$$

where $c_k\equiv\Sigma_{l=1}^\infty f(l)l^k$ is certainly finite for any integer k. The proof of the estimate above goes as follows:

$$\|(a_L+a_L^\dagger)^l\|^{f,k}=\|f(M)(a_L+a_L^\dagger)^lM^k\|\leq\|f(M)M^k\|\|M^{-k}(a_L+a_L^\dagger)M^k\|^l.$$

Now, $\|f(M)M^k\|=c_k$ and, using the spectral decomposition for M and formula (A4),

$$\|M^{-k}(a_L+a_L^\dagger)M^k\|=\sum_{l,s=1}^{L}\left(\frac{s}{l}\right)^k\|\Pi_l(a+a^\dagger)\Pi_s\|\leq(2^k+1)L\sqrt{L+1}.$$

Putting it all together, inequality (4.22) follows. Now we can use this inequality and the one in (4.21) to prove our claim on $\tilde{\beta}_{V,L}^t$. Indeed, we have

$$\left\|\left|\sum_{l=1}^{\infty}\frac{(it\gamma(a_L+a_L^\dagger))^l}{l!}[\sigma_3^V,\sigma_a^i]_l\right|\right\|^{f,k,\Psi}\leq c_k\sum_{l=1}^{\infty}\left(\frac{(|2t\gamma|(2^k+1)(L+1)^{3/2})}{|V|}\right)^l\frac{1}{l!}$$
$$=c_k\left(\exp\left\{\frac{(|2t\gamma|(2^k+1)(L+1)^{3/2})}{|V|}\right\}-1\right),$$

which, of course, goes to zero for $L\to\infty$, at least if the hypothesis (4.10) is satisfied. This concludes the proof of Eq. (4.19). Taking into account condition (4.10) we will write

$$\|\|\beta_L^t(\sigma_\alpha^i)-\sigma_\alpha^i\|\|^{f,k,\Psi}=O(L^{-\mu}), \tag{4.23}$$

where μ is a strictly positive number. Inserting (4.23) in (4.18), with $|V|=L^r$, we get

$$\alpha_L^t(\sigma_\alpha^i)=\sigma_\alpha^i\cos^2(S_3^{L^r})-2\epsilon_{3\alpha\beta}\sigma_\beta^i\sin(S_3^{L^r})\cos(S_3^{L^r})$$
$$+\sigma_3^i\sigma_\alpha^i\sigma_3^i\sin^2(S_3^{L^r})+O(L^{-r})+O(L^{-\mu}). \tag{4.24}$$

Here we have written explicitly the two terms going to zero with L, the reason being that while the first goes to zero in the usual norm of the operators, see Ref. 13, the second one goes to zero in the topology τ. Using now the expression above for $\alpha_L^t(\sigma_\alpha^i)$ and the results in Ref. 10, we conclude that the sequence $\alpha_L^t(\sigma_\alpha^i)$ is τ-Cauchy, so that its limit exists in the algebra $\mathscr{A}$. □

Remarks: (a) Our first remark is about condition (4.10): we do believe that this is only a technical requirement, without any physical meaning, and that a refinement of our estimates can be found which does not make use of (4.10). But it is not worth going in this direction.

(b) Even if the approach we are following in this paper is somehow more in line with Refs. 13 and 14, nevertheless it is in Ref. 10 that we find the information we need about what we have called ''$\mathscr{T}$-strong convergence'' of σ_3^V and of its *good*, i.e., analytical, functions. In particular, in Ref. 13 we were forced not to use the strong topology because of the appearance of variables, like $(1/|V|^\gamma)\Sigma_{i\in V}\sigma_3^i$, which, in general, are not uniformly bounded in V for $\gamma<1$. Of course, the algebraic structure for the spin variables of this model could also be constructed using the results in Ref. 13 which, for $\gamma=1$, is nothing but a mean field model, like the present one. This would imply, however, that the strong seminorms (2.10) should be replaced by the Lassner seminorms constructed starting with a certain ''number'' operator which can be introduced also in the spin

algebra.[2,13] Also, the algebra $B(\mathscr{H}_{\rm spin})$ should be replaced by a certain $\mathscr{L}^+(\tilde{\mathscr{D}})$, where $\tilde{\mathscr{D}}$ is the domain of all the powers of this new number operator, and is a subspace of $\mathscr{H}_{\rm spin}$. Since σ_3^V is uniformly bounded, this is not necessary here, so we have used the simpler framework discussed in Ref. 10.

(c) With the same kind of estimates considered above, it is also possible to prove the existence, among the others, of the limits of the following quantities: $\alpha_L^t(a\sigma_\alpha^i)$, $\alpha_L^t(\sigma_{\alpha_1}^{i_1}\cdots\sigma_{\alpha_n}^{i_n})$, $\alpha_L^t(a^n)$. Here n is an arbitrary integer. The proof essentially follows from the explicit expressions (4.16) and (4.24) for the cutoff dynamics and from the algebraic nature of α_L^t.

We can conclude, therefore, that the algebraic dynamics for the model (4.4) can be rigorously defined using the regularization (3.5) and that a natural topological *-algebra considering the model is $\mathscr{A}[\tau]=B(\mathscr{H}_{\rm spin})\otimes\mathscr{L}^+(\mathscr{D})$, $B(\mathscr{H}_{\rm spin})$ taken with its strong topology and $\mathscr{L}^+(\mathscr{D})$ with the physical one.

B. Many mode bosons

The last model we will discuss in this paper is an extension of the interacting model (4.1):

$$H_V'=\frac{J}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+\sum_{l=1}^n a_l^\dagger a_l+\sum_{l=1}^n \gamma_l(a_l+a_l^\dagger)\sigma_3^V. \tag{4.25}$$

In this generalization, n different modes of bosons are introduced, satisfying the CCR

$$[a_l,a_n^\dagger]=\delta_{ln}\mathbb{1},$$

while γ_l are n small real numbers such that, again, the kinetic boson Hamiltonian $\Sigma_{l=1}^n a_l^\dagger a_l$ can be considered frozen in a given level. Again, therefore, we consider the following simplified Hamiltonian

$$H_V=\frac{J}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+\sum_{l=1}^n \gamma_l(a_l+a_l^\dagger)\sigma_3^V. \tag{4.26}$$

The regularization of H_V can be performed in the following way: for each mode we consider the number operator $N_k=a_k^\dagger a_k$, its spectral decomposition $N_k=\Sigma_{l=0}^\infty l\Pi_{k,l}$ and the projection operators $Q_{k,L}\equiv\Sigma_{l=0}^L\Pi_{k,l}$ (also for this model it is convenient to introduce $M_k=N_k+\mathbb{1}$). We also define the Hilbert spaces $\mathscr{H}_k$, one for each mode, and $\mathscr{D}_k\equiv D^\infty(N_k)\subset\mathscr{H}_k$ as in Sec. II. Finally we take $\mathscr{D}\equiv\mathscr{D}_1\otimes\mathscr{D}_2\otimes\cdots\otimes\mathscr{D}_n$ which is dense in the Hilbert space $\mathscr{H}$ of the tensor product of all the $\mathscr{H}_k$. Other useful spaces are $\mathscr{E}_{k,l}=\{(a_k^\dagger)^l\Phi_0\}$ and $\mathscr{F}_{k,L}=\mathscr{E}_{k,0}+\mathscr{E}_{k,1}+\cdots+\mathscr{E}_{k,L}$. Of course, projections corresponding to different modes commute. The algebra is the usual $\mathscr{A}=B(\mathscr{H}_{\rm spin})\otimes\mathscr{L}^+(\mathscr{D})$.

At this point we define the regularized Hamiltonian $H_{V,L}$ as

$$H_{V,L}=\frac{J}{|V|}\sum_{i,j\in V}\sigma_3^i\sigma_3^j+\sum_{l=1}^n \gamma_l(a_{l,L}+a_{l,L}^\dagger)\sigma_3^V, \tag{4.27}$$

where $a_{l,L}=Q_{l,L}a_lQ_{l,L}$.

The existence of the limit of $\alpha_{V,L}^t(a_j)$ is a simple consequence of the results in the first section. In fact, with the same arguments as for the single mode bosons model, we find that

$$\alpha_{V,L}^t(a_j)=e^{i\gamma(a_{j,L}+a_{j,L}^\dagger)\sigma_3^V t}a_j e^{-i\gamma(a_{j,L}+a_{j,L}^\dagger)\sigma_3^V t}.$$

This has exactly the same expression as in Eq. (4.11), so that its convergence in the physical topology can be proven in the same way.

A bit different is the proof of the existence of the limit of $\alpha_{V,L}^t(\sigma_\alpha^i)$. Without going too much into detail, we just observe that the main difference is that in formula (4.18) $\beta_{V,L}^t(\sigma_\alpha^i)$ must be replaced with

$$\varphi^t_{V,L}(\sigma^i_\alpha)=e^{i\Sigma_{l=1}^n \gamma_l(a_{l,L}+a^\dagger_{l,L})\sigma_3^V t}\sigma^i_\alpha e^{-i\Sigma_{l=1}^n \gamma_l(a_{l,L}+a^\dagger_{l,L})\sigma_3^V t}$$
$$=e^{i\gamma_n(a_{n,L}+a^\dagger_{n,L})\sigma_3^V t}(\cdots(e^{i\gamma_1(a_{1,L}+a^\dagger_{1,L})\sigma_3^V t}\sigma^i_\alpha e^{-i\gamma_1(a_{1,L}+a^\dagger_{1,L})\sigma_3^V t})\cdots)e^{-i\gamma_n(a_{n,L}+a^\dagger_{n,L})\sigma_3^V t}.$$

This can still be estimated as in (4.19), $\varphi^t_{V,L}(\sigma^i_\alpha)\simeq\sigma^i_\alpha$, simply by noticing that the contribution τ-converging to zero appears now n times, n being the fixed number of boson modes. We conclude that (4.24) still holds for this multimodes model, and therefore that the limit of $\alpha^t_{V,L}(\sigma^i_\alpha)$ exists in $\mathscr{A}$.

V. OUTCOME AND FUTURE PROJECTS

In this paper some physically relevant models have been analyzed making use of the algebras of unbounded operators of Lassner type. In particular, we have shown that for the free bosons, and for other models related to this one, it is possible to introduce an ''occupation number'' cutoff, whose removal can be performed in the complete topological *-algebra $\mathscr{L}^+(\mathscr{D})$, where $\mathscr{D}$ is the domain of all powers of the number operator N.

Moreover, for an interacting model of spins and bosons, the same kind of regularization allows a rigorous definition of the cutoff Hamiltonian. Again, this cutoff can be removed, together with its other volume cutoff, working in a large topological *-algebra $\mathscr{A}[\tau]=B(\mathscr{H}_{\text{spin}})\otimes\mathscr{L}^+(\mathscr{D})$, where $B(\mathscr{H}_{\text{spin}})$ and $\mathscr{L}^+(\mathscr{D})$ are endowed, respectively, with the strong and with the physical topologies.

For what concerns our future projects, our main goal is to analyze, within the mathematical structure proposed here, some more physically realistic models like the original conservative Dicke model,[15] or its nonconservative generalization proposed by Alli and Sewell.[7]

ACKNOWLEDGMENTS

It is a pleasure to thank Professor G. L. Sewell for many stimulating discussions about the interactions between matter and radiation and for his comments. I also would like to thank Dr. D. Dubin for an interesting discussion concerning topological algebras of unbounded operators, and Dr. C. Trapani for his careful reading of the manuscript and for his comments. This work has been supported by M.U.R.S.T.

APPENDIX A: SOME USEFUL FORMULAS

In this Appendix we will obtain some useful relations concerning the projection operators Π_l and Q_L. We start by proving the following relation:

$$\Pi_{l-1}a=a\Pi_l,\quad l=1,2,3,\ldots. \tag{A1}$$

The proof is the following. First, we observe that $\varphi\in\mathscr{H}$ is such that $\Pi_{l-1}a\varphi=0$ if, and only if, $a\Pi_l\varphi=0$. In fact, if $\Pi_{l-1}a\varphi=0$ then $a\varphi$ must be orthogonal to $\mathscr{E}_{l-1}$ or, that is to say, φ must be orthogonal to $\mathscr{E}_l$, which implies that $\Pi_l\varphi=0$. The converse implication is proved in the same way. This implies that if Ψ is such that $\Pi_{l-1}a\Psi\neq0$ then necessarily also $a\Pi_l\Psi\neq0$. Let Ψ be such a vector. Then $\Pi_l\Psi\in\mathscr{E}_l$ and $a\Pi_l\Psi\in\mathscr{E}_{l-1}$. Analogously $\Pi_{l-1}a\Psi\in\mathscr{E}_{l-1}$. Therefore the two vectors must be necessarily proportional to each other. In other words, a nonzero constant α_l must exist such that $\Pi_{l-1}a\Psi=\alpha_l a\Pi_l\Psi$. Taking the scalar product of both sides of this equality with the vector $(a^\dagger)^{l-1}\Phi_0$ we deduce that $\alpha_l=1$, from which Eq. (A1) follows.

The following relations are now easy consequences of formula (A1):

$$a^\dagger\Pi_{l-1}=\Pi_l a^\dagger,\quad Q_L a=aQ_{L+1},\quad Q_L a^\dagger=a^\dagger Q_{L-1}, \tag{A2}$$

$$[Q_L,a]=\Pi_L a,\quad [Q_L,a^\dagger]=-a^\dagger\Pi_L, \tag{A3}$$

where l and L take integer values.

Let us finally prove the following useful formula:

$$\|\Pi_l a\Pi_s\|=\sqrt{s}\,\delta_{l,s-1}. \tag{A4}$$

This is a consequence of the following equalities:

$$\|\Pi_l a \Pi_s\|^2 = \|a\Pi_{l+1}\Pi_s\|^2 = \delta_{l,s-1}\|a\Pi_s\|^2 = \delta_{l,s-1}\sup_{\|\varphi\|\leqslant 1}|\langle a\Pi_s\varphi, a\Pi_s\varphi\rangle|$$

$$= \delta_{l,s-1}\sup_{\|\varphi\|\leqslant 1}|\langle \Pi_s\varphi, N\Pi_s\varphi\rangle| = s\,\delta_{l,s-1}.$$

APPENDIX B: THE TOPOLOGY τ FOR THE INTERACTING MODELS

Let $\mathscr{H}$ and $\mathscr{H}_{\rm spin}$ be the Hilbert spaces, respectively, of the bosons and the spin. $\mathscr{D}$ is the subset of $\mathscr{H}$ defined as in Sec. II and $\mathscr{A}=B(\mathscr{H}_{\rm spin})\otimes\mathscr{L}^+(\mathscr{D})$ is the relevant *-algebra for the model in Sec. IV.

We start defining, for $X=X_1\otimes X_2\in B(\mathscr{H}_{\rm spin})\otimes B(\mathscr{H})$, and for $\Psi\in\mathscr{T}$ given in (4.5), the following ''strong'' seminorms:

$$\|X\|^{\Psi} = \sup_{\|\chi_1\|\leqslant 1,\|\chi_2\|\leqslant 1,\|\phi_2\|\leqslant 1}|\langle \chi_1\otimes\chi_2, (X_1\otimes X_2)\Psi\otimes\phi_2\rangle| = \|X_1\Psi\|\|X_2\| \tag{B1}$$

(obviously in $\|X_1\Psi\|$ the norm is the one in $\mathscr{H}_{\rm spin}$, while in $\|X_2\|$ is the one in $B(\mathscr{H})$. It is straightforwardly proven that these are really seminorms. The meaning of the above definition is also evident: we are considering the $(\mathscr{T}-)$ strong topology on $B(\mathscr{H}_{\rm spin})$ and the usual norm topology on $B(\mathscr{H})$. At this point, since if $A\in\mathscr{L}^+(\mathscr{D})$ then both $f(M)AM^k$ and $M^kAf(M)$ are bounded operators on $\mathscr{H}$, the following definition appears as the most natural: for any $X\in\mathscr{A}$ we define the following seminorms

$$|||X|||^{f,k,\Psi} \equiv \max\{\|f(M)XM^k\|^{\Psi}, \|M^kXf(M)\|^{\Psi}\}. \tag{B2}$$

Again, it is not difficult to prove that $\{|||\cdots|||^{f,k,\Psi}\}$ is a system of seminorms which define a topology τ and this topology makes of $\mathscr{A}$ a complete topological *-algebra.

[1] R. Haag and D. Kastler, ''An algebraic approach to Quantum Field Theory,'' J. Math. Phys. **5**, 848–861 (1964).
[2] G. Lassner, ''Topological algebras and their applications in Quantum Statistics,'' Wiss. Z. KMU-Leipzig, Math.-Naturwiss. R. **30**, 572–595 (1981); ''Algebras of unbounded operators and quantum dynamics,'' Physica A **124**, 471–480 (1984).
[3] J.-P. Antoine and W. Karwowski, ''Partial *-algebras of closed linear operators in Hilbert space,'' Publ. RIMS Kyoto Univ. **21**, 205–236 (1985).
[4] F. Bagarello and C. Trapani, ''States and Representations of CQ*-Algebras,'' Ann. Inst. Henri Poincaré Phys. Theor. **61**, 103–133 (1994).
[5] K. Schmüdgen, *Unbounded Operator Algebras and Representation Theory* (Akademie-Verlag, Berlin, 1989).
[6] D. A. Dubin and M. A. Hennings, *Quantum Mechanics, Algebras and Distributions* (Longman Group, Harlow, UK, 1990).
[7] G. Alli and G. L. Sewell, ''New methods and structures in the theory of the multi-mode Dicke laser model,'' J. Math. Phys. **36**, 5598–5612 (1995).
[8] F. Bagarello and G. L. Sewell, ''New structures in the theory of the laser model II: Microscopic dynamics and a non-equilibrium entropy principle,'' J. Math. Phys. **39**, 2730–2747 (1998).
[9] C. Trapani, ''Quasi *-algebras of operators and their applications,'' Rev. Math. Phys. **7**, 1303–1332 (1995).
[10] F. Bagarello and G. Morchio, ''Dynamics of mean-field spin models from basic results in abstract differential equations,'' J. Stat. Phys. **66**, 849–866 (1992).
[11] W. Thirring and A. Wehrl, ''On the mathematical structure of the BCS model,'' Commun. Math. Phys. **4**, 303–314 (1967).
[12] D. Yoshioka and H. Fukuyama, ''Charge density wave of two-dimensional electrons in strong magnetic fields,'' J. Phys. Soc. Jpn. **47**, 394–402 (1979).
[13] F. Bagarello and C. Trapani, '''Almost' mean field Ising model: An algebraic approach,'' J. Stat. Phys. **65**, 469–482 (1991).
[14] F. Bagarello and C. Trapani, ''The Heisenberg dynamics of spin systems: A quasi*-algebras approach,'' J. Math. Phys. **37**, 4219–4234 (1996).
[15] R. H. Dicke, Phys. Rev. **93**, 99 (1954).

Calculus on the multivector manifold spin$^+(p,q)$

Stephen Blake
Machine Tool Vision Limited, 53 Lawrence Street, Blackburn, Lancashire, BB2 1QF, England

(Received 13 January 1998; accepted for publication 27 May 1998)

Coordinate-free differential and integral calculus is formulated on the spin groups, regarded as multivector manifolds embedded in Geometric Algebra. The derivative with respect to a point on the manifold is shown to be in the tangent algebra at the inverse point. The curvature of spin(p,q) is calculated and related to the structure constants of its Lie algebra. Volume elements on spin(p,q) are introduced naturally as completely antisymmetrized geometric products of bivectors. They are left invariant like Haar measure and are elements of a multivector-valued Grassman algebra. It is shown that integration in this sense is forbidden for many spin groups. A version of Stokes's theorem is formulated on spin(p,q). © *1998 American Institute of Physics.* [S0022-2488(98)03311-8]

I. INTRODUCTION

This article is concerned with a development of the Geometric Calculus of references Refs. 1–5. Calculus on manifolds is formulated in a coordinate-free way, in the principal reference, Ref. 1, by taking the points of a manifold as vectors in the Geometric Algebra. The notation of Ref. 1 is used in this article. The Geometric Algebra is a graded linear space, whose elements are called multivectors. The grade-0 elements are scalars and the grade-1 elements are vectors. It is reasonable to try to generalize the existing calculus on vector manifolds by formulating differential and integral calculus on manifolds whose points are the higher grade multivectors.

The motivation for this generalization arose from problems involving rotations. In Geometric Algebra, rotations are multivectors R of even grades, called rotors, which satisfy the following constraint:

$$R^\dagger R = 1. \tag{1}$$

Equation (1) picks out a set of points which form a multivector manifold. In applications, one often encounters the need to evaluate the derivative, in the rotor manifold, of a function of a rotation. This led the writer to work on the differential and integral calculus of general multivector manifolds. Guided by the vector manifold theory of Ref. 1, it was necessary to find a projection operator that sends any multivector in the embedding Geometric Algebra into a multivector in the tangent algebra at a point on a multivector manifold. The difficulty is that, on a general multivector manifold, there does not appear to be anything analogous to a pseudoscalar or volume element on a vector manifold. In order to make progress, the scope of the work was confined to the general rotor manifold.

Rotors form the group spin$^+(p,q)$ and Refs. 3 and 2 show that every Lie group can be represented as a spin group spin(p,q). Therefore the restriction to the rotor manifold spin$^+(p,q)$ still encompasses Lie groups as multivector manifolds. Spin(p,q) has the interesting property that the elements of its Lie algebra are bivectors (grade-2 elements of the Geometric Algebra). This fact, and the notion of left and right translation on a group, permit the construction of a projection operator into the tangent algebra of spin$^+(p,q)$ in Sec. III A.

Armed with the projection operator, the differential calculus on the multivector manifold spin$^+(p,q)$ is developed. Two interesting results are obtained. First, the derivative with respect to a point on the multivector manifold is shown, in Sec. III C, to live in the tangent algebra at the inverse point on the manifold. Second, the curvature tensor on the manifold is evaluated in Sec. III G and shown to be related to the structure constants of spin(p,q).

Integration on spin$^+(p,q)$ is formulated in a natural way by defining volume elements as completely antisymmetrized geometric products of bivectors. These volume elements are left

 © 1998 American Institute of Physics

invariant, just like Haar measure Ref. 6 and are shown to form a multivector-valued Grassman algebra in Sec. IV B. The interaction between the multivector grade and the Grassman grade of the volume elements gives rise to the interesting property (Sec. IV C) that not every spin group permits integration. Finally, a version of Stokes's theorem on $\text{spin}^+(p,q)$ is obtained in Sec. IV E and it is used to quickly obtain the curvature formula of Sec. III G.

Lie groups are thoroughly treated as vector manifolds in Ref. 1, so that it is necessary to consider what precisely is gained by the notion of a multivector manifold. In the author's opinion, a multivector can actually be an element of a Lie group because the grading endows it with a rich enough structure. One has no need to construct a representation of the group out of some other mathematical objects which lack the required spectrum of properties. For example, consider the case of spin(3,0). In Chap. 8 of Ref. 1, this group is represented as a vector manifold by the 3-sphere S^3. This is the set of vectors x in four-dimensional (4D) Euclidean space with constraint $x^2=1$. Let x, x', and x'' be three points of the vector manifold which are related by the group multiplication rule. The group multiplication rule on the representative vector manifold is

$$x''=\tfrac{1}{2}(xex'+x'ex)+\tfrac{1}{2}(xex'-x'ex)I, \tag{2}$$

where e is the vector corresponding to the group identity element and I is the unit pseudoscalar of the Geometric Algebra constructed from the 4D Euclidean space. Let rotors R, R', and R'' be elements of spin(3,0). In contrast to Eq. (2), the group multiplication rule is simply $R''=R'R$. This takes the form that is used for the group elements underlying the representation.

Furthermore, a key property of the spin groups is that their Lie algebras are bivectors. Consequently, the tangent algebra of the multivector manifold $\text{spin}(p,q)$ at the identity is distinguished by being the particularly simple space of bivectors. The representation of spin(3,0) as the vector manifold S^3 fails to capture this property because the lack of grading on a vector manifold means that all the tangent spaces must be copies of a linear vector space.

In summary, in order to make progress in physics one needs geometric objects that match the underlying physical reality. The spin groups seem to be an important constituent of this underlying reality. It seems that the grading associated with a multivector makes the notion of a multivector manifold the appropriate geometric object to use when treating the spin groups. For these reasons, the writer thinks that the notion of a multivector manifold might become a fruitful contribution to the program of geometrization of physics.

II. BASIC DEFINITIONS

The Geometric Algebra generated by p vectors of positive norm and q of negative norm is denoted $G_{p,q}$. The grade-r subspace of this algebra is denoted $G^r_{p,q}$.

Orthogonal transformations in $G_{p,q}$ are isometries of the inner product

$$f(x)\cdot f(y)=x\cdot y.$$

The simplest orthogonal transformation in $G_{p,q}$ is the reflection

$$f(x)=-u^{-1}xu,$$

where u is a vector. The most general isometry of the inner product is formed by k reflections

$$f(x)=(-1)^k u_k^{-1}\cdots u_1^{-1}xu_1\cdots u_k.$$

The inverse factors are $u_i^{-1}=u_i/u_i^2$ so nothing is lost by using unit vectors $u_i^2=\pm 1$. A rotation is the result of a pair of reflections so that a general rotation is obtained when k is even. The isometry is now

$$f(x)=U^{\dagger}xU,$$

where the multivector U is

$$U=u_1\cdots u_k$$

with k even and

$$U^\dagger U = u_k \cdots u_1 u_1 \cdots u_k = \pm 1.$$

The group of all products of an even number of unit vector factors is the group spin(p,q). The term *rotor* is used for elements of spin(p,q) satisfying $R^\dagger R = +1$, which has already been encountered as Eq. (1). The subgroup of spin(p,q) just containing the rotors is called the rotor group spin$^+(p,q)$. The reason for specifying rotors to obey Eq. (1) is because rotations are continuously connected to the identity, which is the rotor $R=1$, and $1^\dagger 1 = 1$.

III. DIFFERENTIAL CALCULUS ON THE ROTOR MANIFOLD

Chapter 2 of Ref. 1 shows how to differentiate a function of a multivector variable defined on a linear multivector subspace of $G_{p,q}$. The current section extends this to the case of functions defined on the rotor manifold.

A. The projection operator

The rotor manifold is the set of multivector points in $G_{p,q}$ that belong to spin$^+(p,q)$. It is embedded in $G_{p,q}$. Consider a function $A(R)$, defined on the rotor manifold, that takes values in $G_{p,q}$. For example, suppose that $A(R) = uRv$, where R is a point on the rotor manifold and u and v are arbitrary vectors. The example function uRv sticks out of the tangent algebra of the rotor manifold. In order to verify this, note that the points of the rotor manifold obey Eq. (1). Therefore, a small displacement δR in the rotor manifold obeys

$$(\delta R)^\dagger R + R^\dagger \delta R = 0$$

and a multivector-valued function $A(R)$ sticks out of the tangent algebra of the rotor manifold at point R if it violates the constraint

$$A^\dagger R + R^\dagger A = 0. \tag{3}$$

On substituting the example function $A = uRv$ into the left-hand side of Eq. (3), one obtains the expression $(R^\dagger uR)\cdot v/2$. The example function does indeed stick out of the tangent algebra since the expression is not zero. One needs an operator $P(R,A)$ that will project such a multivector A into the tangent algebra at point R.

The projection operator will be made by using the fact that each multivector defines a map of the geometric algebra into itself. Let L,X,Y be multivectors in $G_{p,q}$ and write $Y = LX$. This defines a left translation by L of the points X of $G_{p,q}$ into points Y. Consider a pair of neighboring points X and $X+\delta X$. Under left translation by L, the small displacement δX at X is mapped to a small displacement $\delta Y = L\,\delta X$ at $Y = LX$. In this way, a multivector-valued function $A(X)$ at X can be transformed, or left translated, into a multivector-valued function taking value $LA(X)$ at $Y = LX$. The next paragraph defines the projection operator by left translating a function to the identity element of the rotor manifold, manipulating it in the tangent algebra at the identity, then left translating it back to its starting point on the rotor manifold.

Let $A(R)$ be a multivector-valued function defined on the rotor manifold. The function is left translated back to the identity as $R^{-1}A(R)$. The Lie algebra of spin(p,q) is the linear space $G^2_{p,q}$ of bivectors. Consequently, taking the grade-2 part of the left-translated function gives a bivector $\langle R^{-1}A(R)\rangle_2$ in the tangent algebra of spin(p,q) at the identity. Left translation of this element of the Lie algebra to the original point R of the rotor manifold gives a multivector $R\langle R^{-1}A(R)\rangle_2$ in the tangent algebra of the rotor manifold at point R. The projection operator is therefore defined by

$$P(R,A) \overset{\text{def}}{=} R\langle R^{-1}A\rangle_2. \tag{4}$$

It is easily verified that the projection obeys the constraint of Eq. (3). The projection of the example function $A = uRv$ into the tangent algebra of the rotor manifold at R is $P(R,A) = R((R^\dagger uR)\wedge v)$.

The projection could also be defined using right translation as

$$P(R,A)=\langle AR^{-1}\rangle_2 R \tag{5}$$

and these alternative definitions turn out to be identical. In order to see this use, the general result $\langle R^\dagger AR\rangle_r=R^\dagger\langle A\rangle_r R$, which is easily obtained by writing A as a sum of blades of anticommuting vectors and using the property that R is an orthogonal transformation. Starting from the right-translated Eq. (5), the left-translated Eq. (4) is obtained as follows:

$$\begin{aligned}P(R,A)&=\langle AR^{-1}\rangle_2 R=\langle AR^\dagger\rangle_2 R=\langle RR^\dagger AR^\dagger\rangle_2 R=\langle R(R^\dagger A)R^\dagger\rangle_2 R\\&=R\langle R^\dagger A\rangle_2 R^\dagger R=R\langle R^\dagger A\rangle_2=R\langle R^{-1}A\rangle_2.\end{aligned}$$

B. Basis and dual basis multivector fields

Let $\{E_i\}$ be a bivector basis for the tangent space $G^2_{p,q}$ of the rotor manifold at the identity. Let $\{E^i\}$ be the dual bivector basis so that $\langle E^iE_j\rangle_0=\delta^i_j$. Expand the bivector part of the projection operator in equation (4) in terms of the basis bivectors. The summation convention is in force from this point onwards.

$$P(R,A)=R\langle R^{-1}A\rangle_2=R\langle\langle R^{-1}A\rangle_2E^i\rangle_0E_i=\langle\langle R^{-1}A\rangle_2E^i\rangle_0RE_i=\langle R^{-1}AE^i\rangle_0RE_i.$$

An application of the cyclic property of the scalar product in the final line gives the following formula for the projection:

$$P(R,A)=\langle AE^iR^{-1}\rangle_0RE_i. \tag{6}$$

The multivector fields RE_i in Eq. (6) are the left invariant fields of the rotor manifold which form a multivector basis for the tangent algebra at R. The form of Eq. (6) also shows that the multivector fields E^iR^{-1} are the dual basis fields. They are also right invariant multivector fields spanning the tangent algebra at the inverse point R^{-1} of the rotor manifold. These multivector fields are a basis and its dual basis because $\langle E^iR^{-1}RE_j\rangle_0=\delta^i_j$.

It is interesting that expansion (6) of a multivector-valued field on the rotor manifold in terms of a set of basis fields at a point should involve the dual basis fields which live at the inverse point on the manifold. On a vector manifold, both the basis and dual basis vector fields span the tangent vector space at the point on the vector manifold. In contrast, on the multivector manifold, the basis multivector fields RE_i span the tangent algebra at point R, while the dual basis multivector fields span the tangent algebra at point R^{-1}, and these tangent algebras are different from one another. This is illustrated in the simple case of spin(2,0). The rotor manifold spin(2,0) is one-dimensional so that a single bivector spans its tangent algebra at the identity. Take $E_1=-E^1=i$ with $i^2=-1$ and consider a point $R=\exp(i\theta/2)$ on the rotor manifold. This point is a rotation by a scalar angle θ. The basis multivector for the tangent algebra at R is

$$RE_1=-\sin(\theta/2)+i\cos(\theta/2),$$

while the basis multivector for the tangent algebra at the inverse point R^{-1} is

$$E^1R^{-1}=-\sin(\theta/2)-i\cos(\theta/2)$$

and it is clear these tangent algebras are different because one cannot write $RE_1=\alpha E^1R^{-1}$ for any scalar α.

C. Differentiation in the rotor manifold

Following Ref. 1, the directed derivative of a multivector-valued function $A(R)$ in the direction δR in the tangent algebra of the rotor manifold at point R is

$$\langle\delta R\partial_R\rangle_0A(R)\overset{\text{def}}{=}A(R+\delta R)-A(R). \tag{7}$$

The meaning of the derivative ∂_R with respect to a point on the rotor manifold can be found by expanding the displacement δR in (7) using (6).

$$\langle \delta R \partial_R \rangle_0 = \langle \langle \delta R E^i R^{-1} \rangle_0 R E_i \partial_R \rangle_0 = \langle \delta R E^i R^{-1} \rangle_0 \langle R E_i \partial_R \rangle_0 = \langle \delta R E^i R^{-1} \langle R E_i \partial_R \rangle_0 \rangle_0 .$$

The final line shows that the derivative with respect to a point is

$$\partial_R = E^i R^{-1} \langle R E_i \partial_R \rangle_0 . \tag{8}$$

The meaning of (8) is that the derivative with respect to a point R on the rotor manifold is a multivector-valued operator in the tangent algebra at the inverse point R^{-1}.

D. Example of differentiation

In applications, it is useful to have a formula for $\partial_R \langle MR \rangle_0$, where M is a constant multivector in $G_{p,q}$.

$$\partial_R \langle MR \rangle_0 = E^i R^\dagger \langle R E_i \dot{\partial}_R \rangle_0 \langle M \dot{R} \rangle_0 = E^i R^\dagger \langle M R E_i \rangle_0 .$$

When one compares the above result with (6), it is clear that

$$\partial_R \langle MR \rangle_0 = P(R^{-1}, M). \tag{9}$$

Similarly,

$$\partial_R \langle R^\dagger M \rangle_0 = \partial_R \langle M^\dagger R \rangle_0 = P(R^{-1}, M^\dagger). \tag{10}$$

As an example of the use of these formulas, the stationary points of the scalar-valued function $\langle v R^\dagger u R \rangle_0$ on spin$^+(p,q)$ are determined. Here, u and v are vectors in $G^1_{p,q}$,

$$\begin{aligned} \partial_R \langle v R^\dagger u R \rangle_0 &= \dot{\partial}_R \langle v \dot{R}^\dagger u R \rangle_0 + \dot{\partial}_R \langle v R^\dagger u \dot{R} \rangle_0 \\ &= \dot{\partial}_R \langle \dot{R}^\dagger u R v \rangle_0 + \dot{\partial}_R \langle v R^\dagger u \dot{R} \rangle_0 \\ &= 2 P(R^\dagger, v R^\dagger u) \\ &= 2 \langle v R^\dagger u R \rangle_2 R^\dagger . \end{aligned} \tag{11}$$

The last line follows from (5). Using (11), the condition for a stationary point is

$$\langle v R^\dagger u R \rangle_2 = v \wedge (R^\dagger u R) = 0. \tag{12}$$

Condition (12) can also be written as an eigenrotor equation

$$v R^\dagger u R = v \cdot (R^\dagger u R) + v \wedge (R^\dagger u R) = \langle v R^\dagger u R \rangle_0 . \tag{13}$$

Multiplying on the right by $R^\dagger$ gives the equivalent eigenrotor equation

$$v R^\dagger u = \langle v R^\dagger u R \rangle_0 R^\dagger \tag{14}$$

with eigenrotors $R^\dagger$ and eigenvalues $\langle v R^\dagger u R \rangle_0$.

E. Lie brackets and commutators

Let A and B be multivector fields in the rotor manifold. Figure 1 shows some integral curves of A and B with parameters α and β, respectively. The multivectors R, S, T, U, and V are points in the rotor manifold that are used to approximate the fields. For example, R and S are separated by a parameter distance α

$$A(R) = \frac{dR}{d\alpha} \approx \frac{S - R}{\alpha}, \tag{15}$$

$$B(R) = \frac{dR}{d\beta} \approx \frac{V - R}{\beta}. \tag{16}$$

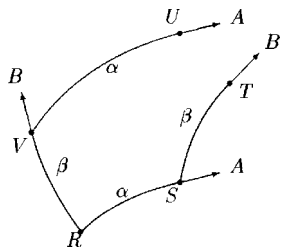

FIG. 1. Gap construction of the Lie bracket.

The conventional *Lie bracket* is extended to the rotor manifold by

$$[B,A] \overset{\text{def}}{=} \langle B\dot{\partial}_R\rangle_0 A(\dot{R}) - \langle A\dot{\partial}_R\rangle_0 B(\dot{R}). \tag{17}$$

Using definition (17) in conjunction with the approximations to the fields, the Lie bracket is shown to be, in essence, the gap in Fig. 1

$$\alpha\beta[B,A] = U - T. \tag{18}$$

Now consider the gap construction to be placed at the identity in the rotor manifold by setting $R=1$. Also, suppose that the multivector fields are no longer arbitrary, but are generators on $\text{spin}^+(p,q)$. This means, for example, that the multivector field $A(V)$ can be obtained by left translation from the field value at the identity as $A(V)=VA(1)$. In other words, when the multivector fields are generators, the points in the gap construction at the identity are not independent since $U=VS$ and $T=SV$. These relations show that the Lie bracket and the commutator are identical for the generators on $\text{spin}^+(p,q)$ at the identity because

$$[B,A] = \frac{1}{\alpha\beta}(VS-SV) = \frac{1}{\alpha\beta}((1+\beta B)(1+\alpha A)-(1+\alpha A)(1+\beta B)) = BA-AB. \tag{19}$$

Reference 1 defines the *commutator product* of a pair of *arbitrary* multivectors by

$$B\times A \overset{\text{def}}{=} \frac{1}{2}(BA-AB)$$

and proves that for A and B *bivectors*,

$$B\times A = \langle BA\rangle_2 .$$

Therefore, the following useful relations hold between bivectors, which are the elements of the tangent algebra $G^2_{p,q}$ at the identity of the rotor manifold $\text{spin}^+(p,q)$:

$$B\times A = \langle BA\rangle_2 = \tfrac{1}{2}(BA-AB) = \tfrac{1}{2}[B,A]. \tag{20}$$

F. Parallel transport on the rotor manifold

The notion of parallel transport on the rotor manifold can be defined with the aid of the projection operator that was introduced in Sec. III A. Let $A(R)$ be a multivector in the tangent algebra of the rotor manifold at R. The result of parallel transport of $A(R)$ to the point $R+\delta R$ on the rotor manifold is the multivector $P(R+\delta R, A(R))$. The change δA under parallel transport by δR is therefore,

$$\delta A = \langle \delta R\dot{\partial}_R\rangle_0 P(\dot{R},A(R)) \overset{\text{def}}{=} P_{\delta R}(R,A(R)) = \langle \delta R\dot{\partial}_R\rangle_0 \dot{R}\langle \dot{R}^{-1}A\rangle_2 = \delta R R^{-1}A - R\langle R^{-1}\delta R R^{-1}A\rangle_2 . \tag{21}$$

In obtaining Eq. (21), the relation $\langle R^{-1}A\rangle_2 = R^{-1}A$ was used since A is in the tangent algebra of the rotor manifold at R and therefore $R^{-1}A$ is the left translation into the tangent algebra $G^2_{p,q}$ at the identity.

Instead of the change δA, which is a multivector that sticks out of the rotor manifold into the embedding algebra $G_{p,q}$, it is more helpful to keep track of the bivector-valued change $\delta(R^{-1}A)$ in the tangent algebra $G^2_{p,q}$ at the identity. This is

$$\delta(R^{-1}A) = -\langle R^{-1}\delta R R^{-1}A\rangle_2 = (R^{-1}A)\times(R^{-1}\delta R). \tag{22}$$

The geodesic equation in $\mathrm{Spin}^+(p,q)$ is obtained by replacing A in Eq. (22) by the tangent multivector $(dR/d\lambda)$ to a path with parameter λ. It is

$$\frac{d}{d\lambda}\left(R^{-1}\frac{dR}{d\lambda}\right) = 0 \tag{23}$$

which has the solution,

$$R(\lambda) = R(0)e^{\lambda B}, \tag{24}$$

where B is a bivector, so that the geodesics are also the one-parameter Abelian subgroups (compare exercise 10.17 of Ref. 7).

The covariant derivative of a multivector-valued function $A(R)$ in the direction δR of the rotor manifold is

$$\langle \delta R\nabla_R\rangle_0 A(R) \overset{\text{def}}{=} P(R,A(R+\delta R)) - P(R,A(R)) = P(R,\langle \delta R\dot{\partial}_R\rangle_0 A(\dot{R})). \tag{25}$$

G. Intrinsic curvature of the rotor manifold

Some results for vector manifolds in Ref. 1 are easily adapted to give the change δA in a multivector A after parallel transport around an infinitesimal parallelogram with sides δR_1 and δR_2 in a general multivector manifold

$$\begin{aligned}\delta A &= P_{\delta R2}(R,P_{\delta R1}(R,A)) - P_{\delta R1}(R,P_{\delta R2}(R,A))\\ &= (\langle \delta R_2\nabla_R\rangle_0\langle \delta R_1\nabla_R\rangle_0 - \langle \delta R_1\nabla_R\rangle_0\langle \delta R_2\nabla_R\rangle_0 - \langle[\delta R_2,\delta R_1]\nabla_R\rangle_0)A.\end{aligned} \tag{26}$$

The case of the rotor manifold can be obtained by replacing the derivatives of the projection operator, $P_{\delta R}(R,A)$, that occur in Eq. (26) with the result in Eq. (21)

$$R^{-1}\delta A = -(R^{-1}A)\times((R^{-1}\delta R_1)\times(R^{-1}\delta R_2)). \tag{27}$$

The details of the calculation of Eq. (27) from the general formula (26) are omitted because the result is most easily obtained by following the derivation of the Riemann curvature tensor in Ref. 8. For this reason, the calculation must be postponed until Stokes's theorem on $\mathrm{spin}^+(p,q)$ is available in Sec. IV E.

The curvature tensor is hidden inside Eq. (27), but it can be made explicit by expressing everything in components with respect to the basis and dual basis multivectors RE_i and E^iR^{-1} of Sec. III B. For example, the multivector A at point R on the rotor manifold is expanded as in Eq. (6),

$$A = \langle AE^iR^{-1}\rangle_0 RE_i = A^iRE_i,$$

where the components of a multivector are distinguished because they carry superscripts. Expanding δA, A, δR_1, and δR_2 in this way in Eq. (27) gives

$$\delta A^l = -E^l\cdot(E_i\times(E_j\times E_k))A^i\delta R_1^j\delta R_2^k$$

and the components B^l_{ijk} of the curvature tensor are

$$B^l_{ijk} = -E^l \cdot (E_i \times (E_j \times E_k)). \tag{28}$$

The structure constants c^k_{ij} of the Lie algebra of spin(p,q) in this basis are obtained from

$$[E_i, E_j] = c^k_{ij} E_k \tag{29}$$

and using Eqs. (20), (29), and (28), the following relation (compare exercise 11.12 of Ref. 7) between the structure constants and the components of the curvature tensor on spin(p,q) is obtained:

$$B^l_{ijk} = -\tfrac{1}{4} c^l_{im} c^m_{jk}. \tag{30}$$

In the standard texts on Lie groups, the Jacobi identity is used to prove the following relation among the structure constants:

$$c^l_{im} c^m_{jk} + c^l_{jm} c^m_{ki} + c^l_{km} c^m_{ij} = 0.$$

The curvature tensor also has the symmetry,

$$B^l_{ijk} + B^l_{jki} + B^l_{kij} = 0$$

and the relation between these two facts is clear from Eq. (30).

IV. INTEGRAL CALCULUS ON THE ROTOR MANIFOLD

Consider Eq. (27) that gives the change in a multivector under parallel transport around a closed contour in the rotor manifold. The comparison of (27) with the analogous equation on a Riemannian manifold Ref. 8 suggests that the bivector $(R^{-1}\delta R_1) \times (R^{-1}\delta R_2)$ is the area of the infinitesimal parallelogram whose edges are the multivector displacements δR_1 and δR_2 on the rotor manifold at R. Equation (20) shows that the key point about this area is that it is a completely antisymmetrized geometric product of bivectors. It seems reasonable to generalize this notion, by defining a k-dimensional volume element on spin$^+(p,q)$ as a completely antisymmetrized geometric product of bivectors, and to formulate integration on spin$^+(p,q)$ in terms of such geometric objects.

A. Directed Haar measure

An infinitesimal k-dimensional volume element $d\mathscr{V}(R)$, at multivector point R on spin$^+(p,q)$, is a completely antisymmetrized geometric product of bivectors

$$d\mathscr{V}(R) = [R^{-1}dR_1, \ldots, R^{-1}dR_k] \overset{\text{def}}{=} \sum_\sigma \ \text{sgn}(\sigma) R^{-1}dR_{\sigma(1)} \cdots R^{-1}dR_{\sigma(k)}. \tag{31}$$

In Eq. (31), σ denotes a permutation of the indices $1,\ldots,k$ and a generalized Lie bracket notation is used to denote a completely antisymmetrized geometric product of bivectors.

If a left translation $T = SR^{-1}$ is made from R to S then a typical dR of the volume element at R is left-translated to $dS = SR^{-1}dR$ and so $S^{-1}dS = R^{-1}dR$. Therefore $d\mathscr{V}(TR) = d\mathscr{V}(R)$ and so the volume elements are a left invariant measure on spin$^+(p,q)$ similar to Haar measure Ref. 6.

B. Grassman algebra of the volume elements

The generators of the Grassman algebra Ref. 9 are the bivectors of $G^2_{p,q}$. A grade-k element of the Grassman algebra is a completely antisymmetrized geometric product of k bivectors $B_1, \ldots, B_k$

$$\mathscr{A} = [B_1, \ldots, B_k] \overset{\text{def}}{=} \sum_\sigma \ \text{sgn}(\sigma) B_{\sigma(1)} \cdots B_{\sigma(k)}. \tag{32}$$

Clearly,

$$[B_{\sigma(1)},\ldots,B_{\sigma(k)}]=\mathrm{sgn}(\sigma)[B_1,\ldots,B_k] \tag{33}$$

so that the brackets can be multiplied by scalars and hence arbitrary elements of the Grassman algebra can be made using linear combinations of the brackets with scalar coefficients.

It is important to note that *grade* now has two meanings. One may refer to the grade of a Grassman element or to its multivector grade. The first is simply the number of bivector generators in the square bracket, the second is the grade of the right-hand side of Eq. (32), considered as a multivector in $G_{p,q}$. The interaction between these two notions of grade is the subject of Sec. IV C.

Elements of the Grassman algebra are multiplied by simply putting the bivector generators together in a geometric product and antisymmetrizing

$$\mathscr{A}\mathscr{B}=[B_1,\ldots,B_k][B_{k+1},\ldots,B_{k+l}]\overset{\text{def}}{=}[B_1,\ldots,B_k,B_{k+1},\ldots,B_{k+l}]. \tag{34}$$

This amounts to just removing the internal square brackets. It is then obvious that multiplication of the Grassman elements is associative. Using Eq. (33), the commutation rule for a pair of elements of Grassman grades k and l is

$$\mathscr{B}\mathscr{A}=(-1)^{kl}\mathscr{A}\mathscr{B}. \tag{35}$$

In Eq. (34), the product of a pair of elements of Grassman grades k and l is of Grassman grade $k+l$. There is also a need for an inner product, written $\mathscr{A}\circ\mathscr{B}$, which reduces the Grassman grade. Using Eq. (32), collect the terms which lead with $B_1,B_2,\ldots,B_k$. The result is a left expansion

$$\mathscr{A}=\sum_{j=1}^{k}(-1)^{j-1}B_j[B_1,\ldots,\check{B}_j,\ldots,B_k] \tag{36}$$

and similarly, a right expansion,

$$\mathscr{A}=\sum_{j=1}^{k}(-1)^{k-j}[B_1,\ldots,\check{B}_j,\ldots,B_k]B_j \tag{37}$$

where the vee means the jth bivector is omitted from the bracket. The rule for constructing the inner product $\mathscr{A}\circ\mathscr{B}$ is to make a right expansion of $\mathscr{A}$ and a left expansion of $\mathscr{B}$. The inner product of Geometric Algebra is then employed to reduce every pair of adjacent bivectors outside the brackets to scalars. The recursive definition is thus,

$$\mathscr{A}\circ\mathscr{B}=[B_1,\ldots,B_k]\circ[B'_1,\ldots,B'_l]$$

$$\overset{\text{def}}{=}\sum_{r=1}^{k}\sum_{s=1}^{l}(-1)^{k-r+s-1}B_r\cdot B'_s[B_1,\ldots,\check{B}_r,\ldots,B_k]\circ[B'_1,\ldots,\check{B}'_s,\ldots,B'_l]. \tag{38}$$

This rule is consistent, because it can be shown that the right-hand side of Eq. (38) preserves the antisymmetry of the elements $\mathscr{A}$ and $\mathscr{B}$. Repeated application of Eq. (38) shows that the Grassman grade of $\mathscr{A}\circ\mathscr{B}$ is $|k-l|$. It can also be shown that the commutation rule for the inner product of elements of Grassman grades k and l with $k\leqslant l$ is

$$\mathscr{B}\circ\mathscr{A}=(-1)^{k(l-k)}\mathscr{A}\circ\mathscr{B}. \tag{39}$$

C. Multivector values of the volume elements

This section studies the interaction between the Grassman grade and the multivector grades of the volume elements that was mentioned in Sec. IV B.

1. Multivector grades of a Grassman element

Consider a Grassman grade-k element $\mathscr{A}$ as a completely antisymmetrized geometric product of bivectors [see Eq. (32) so that it is a multivector in $G_{p,q}$]. Apply the reversion operation of Ref. 1 to this multivector

$$\mathscr{A}^\dagger=[B_1,\ldots,B_k]^\dagger=\sum_\sigma \ \mathrm{sgn}(\sigma)B^\dagger_{\sigma(k)}\cdots B^\dagger_{\sigma(1)}\,.$$

The reverse of a bivector is $B^\dagger=-B$ and each permutation in the sum has also been reversed so that each term gets a sign change $(-1)^{k(k-1)/2}$. Therefore,

$$\mathscr{A}^\dagger=[B_1,\ldots,B_k]^\dagger=(-1)^{k(k+1)/2}[B_1,\ldots,B_k]=(-1)^{k(k+1)/2}\mathscr{A}. \tag{40}$$

Formula (1-1.19) of Ref. 1 can also be used to find the reversion-symmetry of the multivector grade-r part of $\mathscr{A}$

$$\langle\mathscr{A}^\dagger\rangle_r=(-1)^{r(r-1)/2}\langle\mathscr{A}\rangle_r\,. \tag{41}$$

Equations (40) and (41) impose a restriction on the allowable multivector grades of a volume element of Grassman grade-k because

$$\langle\mathscr{A}\rangle_r=0 \ \text{ if } \ (-1)^{r(r-1)/2}\neq(-1)^{k(k+1)/2}. \tag{42}$$

As an example, consider the commutator $[B_1,B_2]$, which, by Eq. (20), is known to be a pure bivector in spite of the fact that it could be of multivector grades $r=0,2,4$ since it is made from geometric products of pairs of bivectors. Setting $k=2$ in Eq. (42) shows that multivector grades $r=0,4$ cannot exist so that the commutator is a pure bivector.

2. Rotor manifolds with nonvanishing volume elements

Let $p+q=n$. The dimension of spin(p,q) is the dimension of its Lie algebra $G^2_{p,q}$ which is $N=n(n-1)/2$. Let $e_1,\ldots,e_n$ be an anticommuting set of basis vectors for $G^1_{p,q}$ with $e_j^2=\pm1$. A bivector basis for the tangent algebra of spin$^+(p,q)$ at the identity is $E_1=e_1e_2,\ldots,E_N=e_{n-1}e_n$. A volume element $d\mathscr{V}$ on spin$^+(p,q)$ is a Grassman grade-N element which, by expanding its bivector factors in terms of the basis bivectors, is shown to be simply proportional to the element $[E_1,\ldots,E_N]$. There are $(n-1)$ copies of e_j for $j=1,\ldots,n$ in each geometric product that forms $[E_1,\ldots,E_N]$. The anticommuting property can be used to bring the vectors together. There is no need to keep track of the sign changes. Each geometric product is of the form $\pm e_1^{n-1}\cdots e_n^{n-1}$. There are two cases to be considered.

If n is odd then each factor e_j^{n-1} contains an even number of vectors and therefore $[E_1,\ldots,E_N]$ must be a pure scalar. Now check that the restriction (42) is compatible with this statement by setting the Grassman grade $k=N$ and the multivector grade $r=0$. The restriction implies that

$$\begin{aligned} d\mathscr{V}&=\langle d\mathscr{V}\rangle_0 \ \text{ for } \ n \ \text{ odd and } \ \frac{n(n-1)(n(n-1)+2)}{8} \ \text{ even},\\ d\mathscr{V}&=0 \ \text{ for } \ n \ \text{ odd and } \ \frac{n(n-1)(n(n-1)+2)}{8} \ \text{ odd}. \end{aligned} \tag{43}$$

If n is even then each factor e_j^{n-1} contains an odd number of vectors and therefore $[E_1,\ldots,E_N]$ must be a pure pseudoscalar. This time, the compatibility of the restriction (42) is checked by setting the Grassman grade $k=N$ and the multivector grade $r=n$. The restriction implies that

$$d\mathscr{V}=\langle d\mathscr{V}\rangle_n \ \text{ for } \ n \ \text{ even and } \ \frac{n(n-1)(n(n-1)-2)}{8} \ \text{ even}, \tag{44}$$

$$d\mathscr{V}=0 \quad \text{for } n \text{ even and } \frac{n(n-1)(n(n-1)-2)}{8} \text{ odd.}$$

Applying these conditions for the low dimensional rotor manifolds, one finds that integration is possible on $\text{spin}^+(p,q)$ for $p+q=2,3,8,9,10,11$, while integration is forbidden for $p+q=4,5,6,7,12,13,14,15$.

D. Example of integration

The volume of spin(3,0),

$$\mathscr{V}=\int d\mathscr{V}=\int [R^{-1}dR_1,R^{-1}dR_2,R^{-1}dR_3]$$

is calculated as an example of integration on a rotor manifold. Introduce basis bivectors E_1,E_2,E_2 for $G^2_{3,0}$ with the properties $E_1^2=E_2^2=E_3^2=-1$ and $E_3E_2=E_1$, $E_1E_3=E_2$, $E_2E_1=E_3$. Coordinates are needed to evaluate the integral. A point on the rotor manifold is $R=\alpha+A$ where α is a scalar and A is a bivector. The components A^1,A^2,A^3 of A in the bivector basis are used as coordinates. The constraint (1) means that

$$\alpha^2=1-(A^1)^2-(A^2)^2-(A^3)^2.$$

The bivector generators turn out to be

$$R^{-1}dR_j=\left(E_j\times A+\alpha E_j+\frac{A^jA}{\alpha}\right)dA^j$$

for $j=1,2,3$ and no summation. A straightforward calculation gives

$$\mathscr{V}=3!\int \frac{dA^1dA^2dA^3}{\alpha}=3!\int \frac{dA^1dA^2dA^3}{\pm\sqrt{1-(A^1)^2-(A^2)^2-(A^3)^2}}.$$

The coordinates range over the unit ball in three dimensions. The integral is therefore over two unit balls because of the sign ambiguity of the square root. The volume elements defined in Sec. IV A are a *directed measure* Ref. 1 and so the integral over the whole of spin(3,0) vanishes because of the sign ambiguity. In accordance with Eq. (43), the volume elements $d\mathscr{V}$ on spin(3,0) are scalar-valued, so a positive result for the volume is obtained by integrating the absolute values of the volume elements

$$V=3!\int \frac{dA^1dA^2dA^3}{|\alpha|}=2\cdot 3!\,\pi^2=12\pi^2.$$

The volumes for SO(3) and SU(2) are stated in Ref. 10 as being $8\pi^2$ and $4\pi^2$, respectively. The different numerical factors are not significant. The volume of spin(3,0) of $12\pi^2$ can also be calculated using Euler angle coordinates, more commonly used for SO(3). When this is done it becomes clear that the different numerical factors are due to a factor of 3! that might be included in the definition of a volume element and a factor of 2^3, caused by the three generators because of the half-angle nature of spin(3,0).

E. Stokes's theorem

Stokes's theorem relates an integral over a k-dimensional region to an integral over its $(k-1)$-dimensional boundary. The integral over the k-dimensional region of $\text{spin}^+(p,q)$ will be

$$\int \mathscr{L}(R,(\partial_R R)\circ[R^{-1}dR_1,\ldots,R^{-1}dR_k])$$

where $\mathscr{L}(\cdot,\cdot)$ is a Grassman-valued function on the rotor manifold (first slot) and is a linear function of a Grassman element (second slot). The second slot is filled with the Grassman inner

product $(\partial_R R)\circ d\mathscr{V}$ where ∂_R is the derivative with respect to a point R on the rotor manifold [Eq. (8)] and $d\mathscr{V}$ is the directed Haar measure [Eq. (31)]. Section III C has shown that ∂_R lives in the tangent algebra at the inverse point R^{-1}. Therefore, it must be right translated, as $\partial_R R$, to the tangent space at the identity of spin$^+(p,q)$, where it is a bivector or generator of the Grassman algebra and it can take part in the Grassman inner product. The derivative ∂_R operates on *all* the R variables in the integral including those in the directed Haar measure.

Stokes's theorem is a straightforward calculation using the definition of the Grassman inner product [Eq. (38)]. The result is

$$\int_{\mathscr{V}} \mathscr{L}(R,(\partial_R R)\circ d\mathscr{V}(R)) = \int_{\partial\mathscr{V}} \mathscr{L}(R,d\mathscr{S}(R)), \tag{45}$$

where $d\mathscr{S}$ is the Grassman grade-$(k-1)$ measure of the boundary of the integration region.

The derivation of formula (27) for the change δA in a multivector A after parallel transport around a closed contour in spin$^+(p,q)$, that was promised in Sec. III G, is now obtained by application of Stokes's theorem. Consider the following integral over a two-dimensional patch of spin$^+(p,q)$ which is converted to an integral around a closed contour by Stokes's theorem:

$$\int (R^{-1}A)\times(\partial_R R\circ[R^{-1}dR_1, R^{-1}dR_2]) = \oint (R^{-1}A)\times(R^{-1}dR). \tag{46}$$

Equation (22) shows that the right-hand side of Eq. (46) is the cumulative change in the bivector that has been left translated to the identity as the loop is traversed

$$\oint d(R^{-1}A) = \int (R^{-1}A)\times(\partial_R R\circ[R^{-1}dR_1, R^{-1}dR_2]). \tag{47}$$

Now evaluate the right-hand side using the definition (38) of the Grassman inner product

$$\oint d(R^{-1}A) = \int \langle dR_1\partial_R\rangle_0((R^{-1}A)\times(R^{-1}dR_2)) - \langle dR_2\partial_R\rangle_0((R^{-1}A)\times(R^{-1}dR_1)).$$

Differentiate by parts, use Eq. (22) for the $R^{-1}A$ terms, use the Jacobi identity for commutator products and the definition (20) of the commutator product, and the result is

$$\oint d(R^{-1}A) = -\int (R^{-1}A)\times(R^{-1}dR_1\times R^{-1}dR_2). \tag{48}$$

This is a bivector change. Multiply by R in order to get the change δA at R on the rotor manifold for an infinitesimal loop. The result,

$$\delta A = -R((R^{-1}A)\times(R^{-1}dR_1\times R^{-1}dR_2))$$

is Eq. (27) of Sec. III G.

[1] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus, A Unified Language for Mathematics and Physics* (Reidel, Dordrecht, 1984).
[2] C. J. L. Doran, ''Geometric algebra and its application to mathematical physics,'' Ph. D. thesis, University of Cambridge, 1994.
[3] C. J. L. Doran, D. Hestenes, F. Sommen, and N. Van Acker, ''Lie groups as spin groups,'' J. Math. Phys. **34**, 3642 (1993).
[4] A. N. Lasenby, C. J. L. Doran, and S. F. Gull, ''Grassman calculus, pseudoclassical mechanics and geometric algebra,'' J. Math. Phys. **34**, 3683 (1993).
[5] D. Hestenes and R. Ziegler, ''Projective geometry with Clifford algebra,'' Acta. Appl. Math. **23**, 25 (1991).
[6] A. P. Balachandran and C. G. Trahern, *Lectures on Group Theory for Physicists* (Bibliopolis, Napoli, 1985).
[7] C. Misner, K. Thorne, and A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[8] L. D. Landau, E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, New York, 1971).
[9] J. F. Cornwell, *Group Theory in Physics* (Academic, New York, 1989), Vol. 3.
[10] J. F. Cornwell, *Group Theory in Physics* (Academic, New York, 1984), Vol. 1.

Foliation of a dynamically homogeneous neutral manifold

Novica Blažić, Neda Bokan,[a] and Zoran Rakić
Faculty of Mathematics, University of Belgrade
Studentski trg 16, p.p. 550, 11000 Belgrade, Yugoslavia

(Received 8 August 1997; accepted for publication 8 July 1998)

It is known that Riemannian and Lorentzian four-dimensional dynamically homogeneous manifolds are two-point homogeneous spaces. This is not true for signature $(--++)$ (neutral or Kleinian signature). In order to better understand their rich structure we study the geometry of nonsymmetric dynamically homogeneous spaces (types II and III): they admit autoparallel distributions and they are locally foliated by totally geodesic, flat, isotropic two-dimensional submanifolds. Moreover we characterize them locally in terms of the existence of an appropriate coordinate system (in the sense of A. G. Walker [Q. J. Math. **1**, 69–79 (1950)]).
© *1998 American Institute of Physics.* [S0022-2488(98)00411-3]

I. INTRODUCTION

The Jacobi operator is important in pseudo-Riemannian geometry as shown in the following examples. A family of free falling particles along a geodesic γ in a Lorentz manifold (M,g), is described by the normal variational Jacobi vector field V along γ. The Jacobi operator plays the role of the tidal force and V satisfies Newton's second law: $V''=R_{V\gamma'}(\gamma')=\mathcal{K}_{\gamma'}V$.[1] The Jacobi operator $\mathcal{K}_{\gamma'}$ is a symmetric operator on the tangent space and from the dynamical point of view the simpliest case to consider is when the algebraic structure of $\mathcal{K}_X$ does not depend on the choice of the unit spacelike or timelike vector X. In other words local dynamics of these spaces is the same in all points and all non-null directions.

Pseudo-Riemannian manifolds with a metric of signature (2,2) have a very interesting geometry. They are also important for physisists. Geometry of these manifolds has been studied, for example, in Refs. 2–4, etc. Spaces with more than one time component have been studied before in Kaluza–Klein theories and a string context (see, Refs. 5–7). $N=2$ string theory is one of the simplest and the most interesting types of string theories. The heterotic version of $N=2$ strings is more intricate than the symmetric one as the dimension for target space is four and its signature (2,2). Its correspondence with self-dual Yang–Mills configuration and a notion of integrability for field theories is studied in Ref. 8. Some other results related to metrics of signature (2,2) one can find, for example, in Refs. 9 and 10, etc.

Let $S^+(p)$ and $S^-(p)$ be the sets of all unit non-null vectors, respectively, at a point $p\in M$. The Jacobi operator $R:Y\mapsto R(Y,X)X$ is a symmetric endomorphism of the tangent space. Its restriction to $X^\perp$ is denoted by $\mathcal{K}_X$. We say that (M,g) is *dynamically homogeneous* or *Osserman manifold* if the Jordan form of $\mathcal{K}_X$ is independent of the choice $X\in S^+(p)\cup S^-(p)$ for arbitrary point $p\in M$.

Now we can state the following question: What is the local and global structure of dynamically homogeneous pseudo-Riemannian manifolds? In a special case of positive definite metric Osserman[11] conjectured that dynamically homogeneous manifolds have to be two-point homogeneous, i.e., rank-one locally symmetric or flat. The conjecture was confirmed by Chi[12] for $\dim M\neq 4k$, $k>1$ (see also Refs. 13, 14 as well as 15 for more details). Also, Lorentzian dynamically homogeneous manifolds have to be of constant sectional curvature (see Ref. 16). But in $(--++)$ signature there are nonsymmetric examples of dynamically homogeneous manifolds ([Ref. 3, examples 8.6–8.8] and Refs. 17 and 18).

If M is a manifold with the diagonalizable Jacobi operator $\mathcal{K}_X$ or if its characteristic polynomial has no triple root then in Ref. 3 it was proved that M is rank-one symmetric space or flat. But

[a]Electronic mail: neda@matf.bg.ac.yu

 © 1998 American Institute of Physics

for $\mathcal{K}_X$ with triple eigenvalue α and the minimal polynomial of the degree 2 or 3 it is not true. These are especially type II and type III dynamically homogeneous manifolds.

We are particularly interested in the neutral signature $(--++)$ and this condition is reduced to the constancy of the minimal polynomial of the Jacobi operator $\mathcal{K}_X$.

The main goal of this paper is to establish some important properties of Osserman manifolds of types II and III. It is known that dynamically homogeneous neutral manifolds have to be self-dual.[19] As we mentioned before self-dual neutral manifolds are broadly studied in Refs. 9 and 10. Here we show that Osserman manifolds of types II and III admit an integrable, autoparallel isotropic two-dimensional distribution in $T_p(M)$, $p \in M$. These distributions have been studied also in the framework of the conformal structure CO(2,2) of four-dimensional manifold M (see Refs. 20 and 2, etc.). We study also some coordinate systems for type II and III dynamically homogeneous manifolds such that the metric has a simpler form. Let us mention that the Walker metric,[21] corresponding to the parallel distributions, belongs to this class and the existence of this type of metric was one of the most important facts in the proof that type II dynamically homogeneous manifolds have to have $\mathcal{K}_X$ with all equal eigenvalues.

II. PRELIMINARIES

Let M be a pseudo-Riemannian manifold of dimension 4 with the metric $\langle \cdot,\cdot \rangle$ of signature (2,2). We also use the notation g for the metric because of convenience reasons. Let $\epsilon_i = -1$, $i=1,2$, and $\epsilon_i = 1$, $i=3,4$. Let e_1, e_2, e_3, e_4 denote an orthonormal basis of T_pM, $\langle e_i, e_j \rangle = \epsilon_i \delta_{ij}$.

ω and Ω are the connection one-form and the curvature two-form, respectively. Then

$$\nabla e_i = \sum \omega_i^s e_s,$$

$$\Omega_j^i = \frac{1}{2} \sum R_{klj}{}^i \theta^k \wedge \theta^l = \frac{1}{2} R_{klj}{}^i \theta^{kl},$$

where θ^i are the dual forms of e_i, i.e., $\theta^i(e_j) = \delta^i_j$ and $\theta^{kl} = \theta^k \wedge \theta^l$. The Cartan structural equations are given as

$$d\theta^i = -\sum \omega_j^i \wedge \theta^j,$$

$$d\omega_j^i = -\sum (\omega_k^i \wedge \omega_j^k + \Omega_j^i) \quad \text{(the second Bianchi identity)}.$$

Let us mention also the symmetry properties of the connection forms as well as the curvature forms

$$\omega_j^i = -\epsilon_i \epsilon_j \omega_i^j, \quad \Omega_j^i = -\epsilon_i \epsilon_j \Omega_i^j.$$

III. TYPE II DYNAMICALLY HOMOGENEOUS MANIFOLDS

In this section we will consider the case when the Jacobi operator $\mathcal{K}_X$ has a characteristic polynomial with triple zero α and the minimal polynomial of degree two. We recall there exists a pseudo-orthonormal basis e_1, e_2, e_3, e_4 such that the matrix of $\mathcal{K}_{e_1}$ has the form

$$\begin{pmatrix} \alpha - \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & \alpha + \frac{1}{2} & 0 \\ 0 & 0 & \beta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} -\alpha + \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & -\alpha - \frac{1}{2} & 0 \\ 0 & 0 & \beta \end{pmatrix}, \tag{3.1}$$

with respect to this basis.

The following one-forms can help us to better understand the structure of these manifolds

$$A=\omega_1^2-\omega_3^4, \quad B=\omega_1^3+\omega_2^4, \quad C=\omega_1^4-\omega_2^3, \quad \tilde{A}=\tfrac{1}{2}(A-B).$$

These one-forms are the linear combinations of the basic forms θ^i, i.e.,

$$A=\sum A_i\theta^i, \quad B=\sum B_i\theta^i, \quad C=\sum C_i\theta^i, \quad \tilde{A}=\sum \tilde{A}_i\theta^i.$$

The coefficients are not independent. More precisely, the following lemma holds.

3.1 Lemma: (Ref. 3) The coefficients $\tilde{A}_i$, $i=1,...,4$ *satisfy the following relations*

$$\tilde{A}_1=\tilde{A}_4=-C_2-C_3 \quad \textit{and} \quad \tilde{A}_2=-\tilde{A}_3=C_1-C_4.$$

We have proved that $\beta=\alpha$ in (3.1) (see [Ref. 3, Corollary 7.7]). Also the nonvanishing generic components of the curvature tensor are the following ones:

$$\begin{gathered} R_{1221}=R_{4334}=\tfrac{1}{2}-\alpha, \quad R_{1331}=R_{4224}=\alpha+\tfrac{1}{2}, \quad R_{1441}=R_{3223}=\beta, \\ R_{2113}=R_{2443}=R_{1234}=-\tfrac{1}{2}, \quad R_{1224}=R_{1334}=R_{1342}=\tfrac{1}{2}. \end{gathered} \tag{3.2}$$

Notice that the scalar curvature is given by $\tau=-12\alpha$.

It is now convenient to consider the components of the curvature tensor in the following isotropic frame

$$F_1=\frac{1}{\sqrt{2}}(e_1-e_4), \quad F_2=\frac{1}{\sqrt{2}}(e_2+e_3), \quad F_3=\frac{1}{\sqrt{2}}(e_2-e_3), \quad F_4=\frac{1}{\sqrt{2}}(e_1+e_4), \tag{3.3}$$

and let ξ^i be the corresponding dual basis and φ_j^i be the corresponding connection forms, so $\nabla F_i=\varphi_i^j F_j$.

Symmetries of the connection forms φ_i^j are the following

$$\begin{gathered} \varphi_1^3+\varphi_2^4=\varphi_3^1+\varphi_4^2=0, \quad \varphi_1^4=\varphi_4^1=0, \\ \varphi_1^2+\varphi_3^4=\varphi_2^1+\varphi_4^3=0, \quad \varphi_2^3=\varphi_3^2=0, \\ \varphi_1^1+\varphi_4^4=\varphi_2^2+\varphi_3^3=0. \end{gathered} \tag{3.4}$$

For type II Osserman manifolds in Ref. 22 it was proved that there exists an isotropic basis F_i, $i=1,...,4$, such that

(i) $\varphi_1^3(F_1)=\varphi_1^3(F_2)=0$.
(ii) $\varphi_1^3(F_3)=-2(\varphi_1^1+\varphi_2^2)(F_2)$ and $\varphi_1^3(F_4)=2(\varphi_1^1+\varphi_2^2)(F_1)$.

Now we are ready to state the theorem.

3.2 Theorem: Let (M,g) be an Osserman manifold of type II. Then, there exists an integrable, autoparallel, isotropic two-dimensional distribution on M.

Proof: Let F_i, $i=1,...,4$ be an already introduced isotropic frame. We will check that the distribution $D=\mathcal{L}(F_1,F_2)$ generated by F_1 and F_2 is autoparallel. We have

$$\nabla_X F_1=\varphi_1^i(X)F_i=\varphi_1^1(X)F_1+\varphi_1^2(X)F_2+\varphi_1^3(X)F_3+\varphi_1^4(X)F_4.$$

Because of symmetries (3.4) we have $\varphi_1^4(X)=0$, and because of (i) it follows $\varphi_1^3(X)=0$ for $X\in D$. This is a consequence of the Osserman condition. Thus $\nabla_X F_1\in D$ for $X\in D$ and similarly $\nabla_X F_2\in D$. Consequently, D is an autoparallel distribution. Since D is autoparallel it is an integrable distribution. □

Remark: Other isotropic two-dimensional distributions $\mathcal{L}(F_i,F_j)$, $\{i,j\}\neq\{1,2\}$ *are not necessarily integrable.*

Remark: The distribution $D=\mathcal{L}(F_1,F_2)$ is parallel if and only if $\varphi_1^3\equiv 0$. In Ref. 3 the possible eigenvalues α and β of the type II manifold were studied. In this context one interesting case was $\beta=4\alpha\neq 0$. It was shown in a few steps that this case is impossible. First, we noticed that $\varphi_1^3\equiv A-B\equiv 0$, consequently D is parallel. Then we introduced the coordinate system where the metric is given by (6.3) for $g_{13}=g_{31}=g_{24}=g_{42}=1$, $g_{14}=g_{41}=g_{23}=g_{32}=0$. The existence of this metric was garanted by Walker.[21] *Studying the Osserman condition in this coordinate system we obtained the contradiction.*

Let us state the following corollary.

3.3 Corollary: An Osserman manifold (M,g) of type II is locally foliated by totally geodesic, flat, isotropic two-dimensional submanifolds.

Proof: First of all notice that the nonzero components of the curvature tensor with respect to the isotropic frame F_i, $i=1,...,4$, are determined by

$$R_{1441}=R_{2332}=R_{1243}=R_{1342}=\alpha \quad \text{and} \quad R_{4334}=2.$$

Since the distribution D is an integrable one, through an arbitrary point $x\in M$ there exists two-dimensional submanifold N defined by D. Since D is autoparallel, N is totally geodesic. The curvature of N is determined by $R_{1221}=0$. □

IV. TYPE III DYNAMICALLY HOMOGENEOUS MANIFOLDS

In this section we will prove the existence of an integrable autoparallel distribution when M is a type III Osserman manifold, i.e., the minimal polynomial has triple zero α.

Similarly to the double zero case we introduce some one-forms

$$A=\omega_1^2+\omega_3^4, \quad B=\omega_2^4-\omega_1^3, \quad C=\omega_2^3+\omega_1^4,$$

$$\tilde{A}=2(A+B), \quad \tilde{B}=2(B-A).$$

Some relations between the forms C, $\tilde{A}$ and $\tilde{B}$ are given by the following lemma.

4.1 Lemma: Let M be a dynamically homogeneous manifold of type III (the minimal polynomial of the Jacobi operator has triple zero). Then, with respect to the pseudoorthonormal basis e_1, e_2, e_3, e_4, we have

(1) $\tilde{A}_1=-\tilde{A}_4=C_2+C_3$, $\tilde{A}_2=-\tilde{A}_3=-(C_1+C_4)$.
(2) $\tilde{B}_1=B_1$, $\tilde{B}_2=B_2$, $\tilde{B}_3=C_1-C_4-B_2$, $\tilde{B}_4=C_3-C_2-B_1$.

Proof: There exists a pseudo-orthonormal frame e_1, e_2, e_3, e_4 (see Ref. 3) such that the nonvanishing generic components of the curvature tensor are determined by

$$R_{1331}=R_{4224}=R_{1441}=R_{3223}=-\alpha, \quad R_{1221}=R_{4334}=\alpha,$$

$$R_{3114}=-R_{3224}=-R_{2114}=-R_{2334}=k,$$

$$R_{1223}=R_{1443}=R_{1332}=-R_{1442}=k.$$

We differentiate the second Bianchi identity to see

$$d\Omega_p^1=\Omega_i^1\wedge\omega_p^i-\omega_i^1\wedge\Omega_p^i,$$

and consequently for $p=2,3,4$ this relation implies

$$\tfrac{1}{2}\tilde{B}(\theta^{13}-\theta^{24})-2B(\theta^{12}-\theta^{34})+C(\theta^{14}+\theta^{23})=0,$$

$$\tfrac{1}{2}\tilde{B}(\theta^{12}-\theta^{34})+2A(\theta^{13}-\theta^{24})-C(\theta^{14}+\theta^{23})=0,$$

$$C(\theta^{12}-\theta^{34}-\theta^{13}+\theta^{24})-\tilde{A}(\theta^{14}+\theta^{23})=0.$$

Comparing coefficients of $\theta^i\wedge\theta^j\wedge\theta^k$ in these three-forms we obtain the system of linear equations, which implies the statement of Lemma. □

We will now change the frame e_i to the isotropic one F_i introduced by (3.3).

As in Sec. III ξ^i is the corresponding dual basis and φ^i_j are the connection one-forms and $\nabla F_i=\varphi^j_i F_j$.

Connection one-forms φ^j_i satisfy the symmetry relations (3.4). Following Ref. 23 we have

$$\widetilde{A}=-4\varphi^1_2,\quad \widetilde{B}=-4\varphi^2_1,\quad C=-\varphi^1_1+\varphi^2_2 \tag{4.1}$$

and

$$\begin{aligned}&\varphi^1_2(F_2)=\varphi^1_2(F_4)=0,\\ &\varphi^1_2(F_1)=2(-\varphi^1_1+\varphi^2_2)(F_2),\\ &\varphi^1_2(F_3)=-2(-\varphi^1_1+\varphi^2_2)(F_4).\end{aligned} \tag{4.2}$$

Now we are able to consider two-dimensional distribution on T_pM. More precisely, one can use (3.3), (4.1) and (4.2) and the same method as in Theorem 3.2 to prove the following theorem.

4.2 Theorem: Let (M,g) be a type III dynamically homogeneous manifold. Then there exists an integrable, autoparallel, isotropic two-dimensional distribution on M.

V. SOME COORDINATE SYSTEM FOR TYPE II AND III DYNAMICALLY HOMOGENEOUS MANIFOLDS

In this section we study some coordinate systems for type II and III dynamically homogeneous manifolds such that the metric has the simplest form as much as possible. We use some results from Sec. III to prove the technical lemma. Then we consider the existence of a coordinate system in Proposition 5.2.

5.1 Lemma: There exist an open set U and functions $A,B\in C^\infty_+(U)$, such that the vector fields $\widetilde{F}_1=AF_1$ and $\widetilde{F}_2=BF_2$ commute, i.e., $[\widetilde{F}_1,\widetilde{F}_2]=0$.

Proof: We use Theorem 3.2 to see that the distribution generated by vectors F_1 and F_2 is integrable and autoparallel. Applying this result we get

$$[F_1,F_2]=\mu_1F_1+\mu_2F_2. \tag{5.1}$$

The straightforward computation gives

$$\begin{aligned}[\widetilde{F}_1,\widetilde{F}_2]=[AF_1,BF_2]&=AB[F_1,F_2]+AF_1(B)F_2-BF_2(A)F_1\\ &=B(A\mu_1-F_2(A))F_1+A(B\mu_2+F_1(B))F_2.\end{aligned}$$

Now the proof follows from the standard fact: for a smooth function μ there exists a smooth solution of the equation

$$\frac{\partial a}{\partial x_1}=\mu_1 a, \tag{5.2}$$

for the appropriate coordinate system (x^1,x^2,x^3,x^4). □

5.2 Proposition: Let (M,g) be a four-dimensional pseudo-Riemannian manifold of signature (2,2) which satisfies the dynamically homogeneous condition. Assume now, that the Jacobi operator is of type II or III. Then there exists a coordinate system $u=(u^1,u^2,u^3,u^4):U\subseteq\mathbb{R}^4\to M$ and an null basis F_1, F_2, F_3, F_4 such that

(i)

$$g=\begin{bmatrix} 0 & 0 & g_{13} & g_{14} \\ 0 & 0 & g_{23} & g_{24} \\ g_{13} & g_{23} & 0 & g_{34} \\ g_{14} & g_{24} & g_{34} & 0 \end{bmatrix}, \quad with \ \det\begin{bmatrix} g_{13} & g_{14} \\ g_{23} & g_{24} \end{bmatrix}=1 \tag{5.3}$$

and

(ii)

$$F_1=\frac{\partial}{\partial u^1}, \quad F_2=\frac{1}{C}\frac{\partial}{\partial u^2}, \quad F_3=CP_i\frac{\partial}{\partial u^i}, \quad F_4=Q_i\frac{\partial}{\partial u^i}, \tag{5.4}$$

where P_i, Q_i, $C\in C^\infty(U)$ *and* C *are nonvanishing on* U.

Proof: We prove Proposition for type II dynamically homogeneous manifolds with all the details. We mention that the proof for type III is the same. Let $\hat{E}_i$ be an arbitrary pseudo-orthonormal basis such that the components of the curvature tensor are given by (3.2) and $\hat{F}_i$ is the corresponding isotropic basis given by (3.3). We already established

$$\hat{\varphi}_1^3(\hat{F}_1)=\hat{\varphi}_1^3(\hat{F}_2)=0, \tag{5.5}$$

$$\hat{\varphi}_1^3(\hat{F}_3)=2(\hat{\varphi}_1^1+\hat{\varphi}_2^2)(\hat{F}_1), \tag{5.6}$$

$$\hat{\varphi}_1^4(\hat{F}_4)=-2(\hat{\varphi}_1^1+\hat{\varphi}_2^2)(\hat{F}_2), \tag{5.7}$$

where $\hat{\varphi}_j^i$ is the corresponding connection form, i.e., $\nabla\hat{F}_i=\hat{\varphi}_i^j\hat{F}_j$.

Now, let $\tilde{F}=(\tilde{F}_1,\tilde{F}_2,\tilde{F}_3,\tilde{F}_4)$ be the following moving frame:

$$\tilde{F}_1=a\hat{F}_1, \quad \tilde{F}_2=b\hat{F}_2, \quad \tilde{F}_3=\frac{1}{b}\hat{F}_3, \quad \tilde{F}_4=\frac{1}{a}\hat{F}_4. \tag{5.8}$$

Notice that $\langle\hat{F}_i,\hat{F}_j\rangle=\langle\tilde{F}_i,\tilde{F}_j\rangle$. The corresponding pseudo-orthonormal frame is

$$\tilde{E}_1=\frac{1}{\sqrt{2}}(\tilde{F}_1+\tilde{F}_4), \quad \tilde{E}_2=\frac{1}{\sqrt{2}}(\tilde{F}_2+\tilde{F}_3), \quad \tilde{E}_3=\frac{1}{\sqrt{2}}(\tilde{F}_2-\tilde{F}_3), \quad \tilde{E}_1=\frac{1}{\sqrt{2}}(\tilde{F}_4-\tilde{F}_1). \tag{5.9}$$

Nonvanishing generic components of the curvature tensor with respect to this frame agree with those given by (3.2) but $\frac{1}{2}$ is replaced by $\frac{1}{2}a^2b^2$. Now we consider Jacobi operator $\mathcal{K}_{\tilde{E}_1}$ and the corresponding frame, i.e., a basis $E_1=\tilde{E}_1,E_2,E_3,E_4$ which satisfies formulas (3.2). Since we assume the Jacobi operator is of type II, then there exists unique two-dimensional invariant subspace generated by $\tilde{E}_2$ and $\tilde{E}_3$, and because of the structure of the curvature tensor follows $\tilde{E}_4=E_4$. There is a hyperbolic rotation which fixes E_1 and E_4 and moves $\tilde{E}_2$ and $\tilde{E}_3$ into E_2 and E_3. Once again, we consider the corresponding null frame defined by (3.3). Clearly,

$$F_1=\tilde{F}_1, \quad F_2=C\tilde{F}_2, \quad F_3=\frac{1}{C}\tilde{F}_3, \quad F_4=\tilde{F}_4, \tag{5.10}$$

where a function C is determined by the above-mentioned hyperbolic rotation.

Since $[\tilde{F}_1,\tilde{F}_2]=0$, Frobenius theorem implies the existence of the coordinate system $\tilde{u}=(\tilde{u}^1,\tilde{u}^2,\tilde{u}^3,\tilde{u}^4)$ in $U\subset M\to\mathbb{R}^4$ such that $\tilde{F}_1=\partial/\partial\tilde{u}^1$ and $\tilde{F}_2=\partial/\partial\tilde{u}^2$.

To complete the proof we will modify slightly this coordinate system to control the determinant condition. Let $u^i=\tilde{u}^i$, $1\leqslant i\leqslant 3$ and

$$\tilde{u}^4=f(u^1,u^2,u^3,u^4).$$

Then the second condition in (5.3) is reduced to

$$\frac{\partial f}{\partial u^4} = \frac{1}{\det(g_{13}g_{24} - g_{14}g_{23})}. \tag{5.11}$$

The solution of (5.11) exists and the proof of Proposition is completed. □

ACKNOWLEDGMENTS

We are very grateful to the referee for suggestions on how to improve and clarify the exposition, particularly the proof of Proposition 5.2. This research was partially supported by SFS, Project No. 04M03.

[1] B. O'Neill, *Semi-Riemannian Geometry with Applications to Relativity* (Academic, New York, 1983).
[2] M. A. Akivis and V. V. Konnov, ''Local aspects in conformal structure theory,'' Russ. Math. Surveys **48**, 1–35 (1993) [Usp. Mat. Nauk **48**, 3–40 (1993)].
[3] N. Blažić, N. Bokan, and Z. Rakić, ''Characterization of 4-dimensional Osserman pseudo-Riemannian manifolds,'' preprint (1995, 1997).
[4] A. Borowiec, M. Ferraris, M. Francaviglia, and I. Volovich, ''Almost complex and almost product Einstein manifolds from a variational principle,'' preprint (1996).
[5] I. Ya. Arefeva and I. V. Volovich, ''Kaluza–Klein theories and the signature of space–time,'' Phys. Lett. B **164**, 287–292 (1985).
[6] I. Bars and C. Kounnas, ''Theories with two times,'' preprint, Chern-TH/97-36 (1997).
[7] A. D. Popov, ''Hyperbolic strings,'' preprint (1990).
[8] H. Oguri and C. Vafa, ''$N=2$ heterotic strings,'' Nucl. Phys. **B367**, 83–104 (1991).
[9] J. Barrett, G. W. Gibbons, M. J. Perry, C. N. Pope, and P. Ruback, ''Kleinian geometry and $N=2$ superstring,'' Int. J. Mod. Phys. A **9**, 1457 (1994).
[10] E. Bergshoeff and E. Sezgin, ''Self-dual supergravity theories in $2+2$ dimensions,'' preprint hep-th/9206101.
[11] R. Osserman, ''Curvature in the eighties,'' Am. Math. Monthly **97**, 731–756 (1990).
[12] Q. S. Chi, ''A curvature characterization of certain locally rank-one symmetric spaces,'' J. Diff. Geom. **28**, 187–202 (1988).
[13] P. Gilkey, ''Manifolds whose curvature operator has constant eigenvalues at the basepoint,'' J. Geom. Anal. **4**, 155–158 (1994).
[14] P. Gilkey, A. Swann, and L. Vanhecke, ''Isoparametric geodesic spheres and a conjecture of Osserman concerning the Jacobi operator,'' Q. J. Math. Oxford. **46**, 299–320 (1995).
[15] N. Blažić, N. Bokan, P. B. Gilkey, and Z. Rakić, ''Pseudo-Riemannian Osserman manifolds,'' Balkan J. Geom. Appl. **2**, 1–12 (1997).
[16] N. Blažić, N. Bokan, and P. Gilkey, ''A Note on Osserman Lorentzian manifolds,'' Bull. London Math. Soc. **29**, 227–230 (1997).
[17] E. Garcia-Rio, M. E. Vázquez-Abal, and R. Vázquez-Lorenzo, ''Nonsymmetric Osserman pseudo-Riemannian manifolds,'' Proc. Am. Math. Soc. (to be published).
[18] Z. Rakić, ''Rank 2 symmetric Osserman spaces,'' Bull. Austral. Math. Soc. **56**, 517–521 (1998).
[19] D. M. Alekseevsky, N. Blažić, N. Bokan, and Z. Rakić, ''Self-dual and pointwise Osserman spaces'' (unpublished).
[20] M. A. Akivis, ''On the real theory of four dimensional conformal structures,'' J. Geom. Phys. **21**, 55–80 (1996).
[21] A. G. Walker, ''Canonical form for a Riemannian space with a parallel field of null planes,'' Q. J. Math. (Oxford) (2) **1**, 69–79 (1950).
[22] N. Blažić, N. Bokan, and Z. Rakić, ''Characterization of type II Osserman manifolds in terms of connection forms,'' preprint (1997).
[23] N. Blažić, N. Bokan, and Z. Rakić, ''The first order PDE system for type III Osserman manifolds,'' Publ. Inst. Math., Nouv. Ser. **62**, 113–119 (1997).

Differential geometry on SU(3) with applications to three state systems

Mark Byrd[a)]
Center for Particle Physics, University of Texas at Austin, Austin, Texas 78712-1081

(Received 22 August 1997; accepted for publication 17 July 1998)

The left and right invariant vector fields are calculated in an ''Euler angle''-type parametrization for the group manifold of SU(3), referred to here as Euler coordinates. The corresponding left and right invariant one-forms are then calculated. This enables the calculation of the invariant volume element or Haar measure. These are then used to describe the density matrix of a pure state and geometric phases for three state systems. © *1998 American Institute of Physics.* [S0022-2488(98)04111-5]

I. INTRODUCTION

In this paper we have two main goals: the first is to construct the invariant vector fields and one-forms on SU(3) in Euler coordinates that are similar to the Euler angle parameters of SU(2); the second is the application of these structures to the description of the geometric phases of three state systems.

The Euler coordinates along with the left and right invariant vector fields and one-forms are constructed here for several reasons. One, the Euler decomposition of a group element is achieved by a method that is very general and could, in principle, be applied to SU(n). Two, this particular parametrization is convenient for many calculations, as will be shown. Three, there will be advantages to having a set of invariant vector fields and corresponding one-forms in the same parametrization of the group.

Color,[1] flavor,[2] and nuclear models[3] use the group SU(3). However, geometric aspects of the group manifold are not emphasized. That is not to say that the differential geometric structures are not relevant to these models; indeed they are! However, as shown in Refs. 4, and 5, the differential geometry of SU(3) is an integral part of the description of three state systems. The Euler coordinates[6] and differential forms provide a complete description of the differential geometry needed to describe the pure state density matrix and geometric phase[7] of three state systems.

II. THE LIE ALGEBRA

The Lie Algebra of a Lie group is a set of left invariant vector fields on the group manifold. These can be constructed by a method that has been applied to SU(2) (see, for example, Ref. 8). The method for decomposing a group element into Euler coordinates and the method for constructing the left invariant vector fields can, in principle, be applied to SU(n). Here we provide the details for SU(3).

A. Euler decomposition

To parametrize the group into the ''Euler angle''-type parametrization, or Euler coordinates, we first consider the Lie algebra of the group. The algebra obeys the following commutation relations:

$$[\lambda_i,\lambda_j]=iC_{ij}^k\lambda_k.$$

The structure constants, C_{ij}^k, can be chosen totally antisymmetric in the three indices. Then the distinction between upper and lower indices is not relevant. (One may also use the Killing metric to lower the index.) In the usual, Gell–Mann basis, the nonzero structure constants are given by

[a)]Electronic mail: mbyrd@physics.utexas.edu

 © 1998 American Institute of Physics

$$C_{123}=2, \quad C_{458}=C_{678}=\sqrt{3},$$

$$C_{147}=C_{246}=C_{257}=C_{345}=C_{516}=C_{637}=1.$$

We can use the Cartan decomposition and split the Lie algebra of SU(3), denoted here $\mathcal{L}(\mathrm{SU}(3))$, into a Lie subalgebra $\mathcal{L}(K)$ and the set of all other SU(3) Lie algebra elements $\mathcal{U}(P)$. The Lie algebra may then be expressed as a (semi-)direct sum,

$$\mathcal{L}(G)=\mathcal{L}(K)\oplus\mathcal{U}(P).$$

To this, there corresponds a decomposition of the group,

$$G=K\cdot P.$$

Here $\mathcal{L}(K)=\{\lambda_1,\lambda_2,\lambda_3,\lambda_8\}$, and $\mathcal{U}(P)=\{\lambda_4,\lambda_5,\lambda_6,\lambda_7\}$, and according to the Cartan decomposition, they obey, for $k_1,\ k_2\in\mathcal{L}(K)$, and $p_1,p_2\in\mathcal{U}(P)$, the following Lie brackets:

$$[k_1,k_2]\in\mathcal{L}(K), \quad [p_1,p_2]\in\mathcal{L}(K),$$

and

$$[k_1,p_1]\in\mathcal{U}(P).$$

Thus K is the U(2) subgroup obtained by exponentiating the corresponding algebra $\{\lambda_1,\lambda_2,\lambda_3\}$ plus λ_8. Since λ_8 commutes with this SU(2) subgroup, an element of K can be written as

$$K=e^{(i\lambda_3\alpha)}e^{(i\lambda_2\beta)}e^{(i\lambda_3\gamma)}e^{(i\lambda_8\phi)},$$

by the Euler angle parametrization of SU(2).

Next, P may be parametrized. Of the four elements of P, we can pick one, say λ_5, and write any element of P as

$$P=Ke^{(i\lambda_5\theta)}K.$$

Dropping the redundancies, we arrive at the following product for SU(3). (This method was adapted from the book by Hermann[9] and first pointed out to me by Biedenharn.[10])

$$D(\alpha,\beta,\gamma,\theta,a,b,c,\phi)=e^{(i\lambda_3\alpha)}e^{(i\lambda_2\beta)}e^{(i\lambda_3\gamma)}e^{(i\lambda_5\theta)}e^{(i\lambda_3\alpha)}e^{(i\lambda_2 b)}e^{(i\lambda_3 c)}e^{(i\lambda_8\phi)}, \tag{1}$$

where D is an arbitrary element of SU(3). This can be written as

$$D(\alpha,\beta,\gamma,\theta,a,b,c,\phi)=D^{(2)}(\alpha,\beta,\gamma)e^{(i\lambda_5\theta)}D^{(2)}(a,b,c)e^{(i\lambda_8\phi)},$$

where the D denotes an arbitrary element of SU(3) and $D^{(2)}$ is an arbitrary element of SU(2) as a subset of SU(3). Since SU(3) is semisimple, it has only one connected component (see, for instance, Ref. 11). Thus, the given parametrization covers the group.

B. A set of invariant vector fields

To find the differential operators corresponding to the Lie algebra elements, one takes derivatives of $D(\alpha,\beta,\gamma,\theta,a,b,c,\phi)$ in (1) with respect to each of its parameters. (For brevity, this will be called D, dropping the explicit dependence on the parameters.) Writing the result of differentiation in the form

$$e^{A}Be^{-A}D,$$

and using the Campbell–Baker–Hausdorff relation to expand this as

$$e^{A}Be^{-A}=B+[A,B]+\tfrac{1}{2}[A,[A,B]]+\cdots,$$

the series can be resummed. Thus, linear combinations of the Lie algebra elements with coefficients given by sines and cosines of the parameters are obtained. The following is an example of such a calculation:

$$\begin{aligned}\frac{\partial D}{\partial \gamma} &= e^{(i\lambda_3\alpha)}e^{(i\lambda_2\beta)}i\lambda_3 e^{(i\lambda_3\gamma)}e^{(i\lambda_5\theta)}e^{(i\lambda_3 a)}e^{(i\lambda_2 b)}e^{(i\lambda_3 c)}e^{(i\lambda_8\phi)}\\ &= e^{(i\lambda_3\alpha)}e^{(i\lambda_2\beta)}(i\lambda_3)e^{(-i\lambda_2\beta)}e^{(-i\lambda_3\alpha)}D\\ &= i(\lambda_1 \cos 2\alpha \sin 2\beta + \lambda_2 \sin 2\alpha \sin \beta + \lambda_3 \cos 2\beta)D.\end{aligned}$$

There are eight such derivatives. These linear combinations may then be expressed as

$$\partial_j = b_j^i \lambda_i , \tag{2}$$

where the parameters have been labeled sequentially as they appear in (1), i.e.,

$$\partial_1 \equiv \frac{\partial}{\partial \alpha}, \quad \partial_2 \equiv \frac{\partial}{\partial \beta}, \quad \partial_3 \equiv \frac{\partial}{\partial \gamma},$$

$$\partial_5 \equiv \frac{\partial}{\partial a}, \quad \partial_6 \equiv \frac{\partial}{\partial b}, \quad \partial_7 \equiv \frac{\partial}{\partial c},$$

$$\partial_4 \equiv \frac{\partial}{\partial \theta}, \quad \partial_8 \equiv \frac{\partial}{\partial \phi}.$$

The matrix b_j^i is then inverted to obtain the representation of the Lie algebra of SU(3) as left invariant vector fields on the group manifold in the Euler parametrization. Note that this calculation relies only on the commutation relations and not on any particular representation of the algebra.

The left invariant vector fields are differential operators on smooth functions D^* on SU(3); they satisfy[12]

$$\Lambda_i D^* = -\lambda^i D^*. \tag{3}$$

Explicit forms of these operators are given in Appendix A. The operators acting from the right are found similarly by taking right derivatives. Note, however, that the right operators obey the commutation relation $[\Lambda_i^r, \Lambda_j^r] = -C_{ij}^k \Lambda_k^r$ since a left translation by a group element is directly related to the right translation of the inverse of this element (see, for example, Refs. 13, 14 or 15).

The calculation of the left invariant vector fields acting on the D matrices was first attempted by Nelson.[16] However, no attempt was made to calculate the right invariant vector fields or the corresponding forms.

As stated above, the right differential operators (the differential operators that correspond to this action when acting *from* the right) are different and are denoted as Λ_i^r. One may find these ''right'' differential operators in two ways. First, one may use the relation

$$D^* \lambda_i = -\Lambda_i^r D^*,$$

and do the same calculation as with the left ones. Second, one may use the fact that

$$\Lambda_i^r = R_i^j \Lambda_j , \tag{4}$$

where the $R_i^j \in \mathrm{SO}(8)$ is an element of the adjoint representation of SU(3). It is therefore a function of the eight parameters above. The matrix R does not cover SO(8), as is clear from the fact that SU(3) has only eight parameters whereas SO(8) has twenty eight.

The right invariant vector fields are also given in Appendix A and the matrix R_i^j has been calculated explicitly in the Euler coordinates as well (see Ref. 17).

III. INVARIANT INTEGRATION

A. Invariant forms

The left invariant forms for the manifold are dual to the left invariant vector fields. This duality between the tangent and cotangent vectors may be used to construct left invariant one-forms from the left invariant vector fields. If the left invariant vector fields are written as

$$\lambda_i = a_i^j \partial_j$$

and the left invariant forms are expressed as

$$\omega^l = b_k^l \, dx^k, \tag{5}$$

then

$$\delta_i^l = \langle \omega^l, \lambda_i \rangle = b_k^l a_i^j \langle dx^k, \partial_j \rangle = b_k^l a_i^j \delta_j^k . \tag{6}$$

Therefore, the matrices b and a are inverse transposes of each other. This is, in fact, the same b given in (2). In this way, we obtain left (and analogously right) invariant forms on the group manifold. Thus, one may now use these to integrate over the whole space or any subspace of the group manifold by the appropriate wedge product of these forms. The explicit left and right invariant one-forms are given in Appendix B.

B. Invariant volume element

The group invariant volume element, or Haar measure, may now be computed, using the invariant forms. To do this, compute the wedge product of the invariant one-forms or, equivalently, take the determinant of the matrix b above. The result is

$$dV = \sin 2\beta \, \sin 2b \, \sin 2\theta \, \sin^2 \theta \, d\alpha \, d\beta \, d\gamma \, d\theta \, da \, db \, dc \, d\phi .$$

This agrees with the result of Holland[18] but is determined only up to normalization. The normalization is determined by having $\int dV = 1$.

With the explicit left and right invariant forms one can check that the left invariant volume element is equal, as expected, to the right invariant volume element by direct computation (see, for instance Ref. 11).

The determination of the ranges of the angles in the Euler coordinates is not a trivial matter. This is related to determining the group invariant volume that would set the normalization of the volume integral. The determination of this group invariant volume has been attempted by many authors (see Ref. 19 and references therein). Taking the products of states in the fundamental irreducible representations and enforcing the orthogonality relations among them, one may infer the minimum ranges for which these orthogonality relations hold. This implies the following ranges of the angles:

$$0 \leqslant \alpha, \gamma, a, c < \pi,$$

$$0 \leqslant \beta, b, \theta \leqslant \frac{\pi}{2}, \quad 0 \leqslant \phi < \sqrt{3}\pi .$$

A further discussion of this is given in Ref. 17.

IV. THREE STATE SYSTEMS

Before discussing the geometric phase, some preliminary remarks are in order. (For a review of geometric phases see Refs. 20, 7, or the review article.[21])

The general pure state density matrices for three-level systems can be represented in the following way:[4,5] Let ψ be a state in a three-dimensional complex Hilbert space $\mathcal{H}^{(3)}$. The density matrix is the matrix ρ described by

$$\rho=\psi\psi^{\dagger}=\tfrac{1}{3}(\mathbb{1}+\sqrt{3}\mathbf{n}\cdot\boldsymbol{\lambda}),$$

$$\psi\in\mathcal{H}^{(3)},\quad(\psi,\psi)=1. \tag{7}$$

Here the dagger denotes the Hermitian conjugate and $\mathbb{1}$ is the 3×3 unit matrix. The dot product of $\mathbf{n}$, and $\boldsymbol{\lambda}$, is the ordinary sum over repeated indices $n^r\lambda_r$. The eight-dimensional unit vector, $\mathbf{n}$, of real functions is restricted to a four-dimensional subspace, as discussed below. The brackets $(\cdot,\cdot)$ denote the inner product on the space $\mathcal{H}^{(3)}$. The $\boldsymbol{\lambda}$ represents the eight Gell–Mann matrices,

$$\lambda_1=\begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix},\quad \lambda_2=\begin{pmatrix}0&-i&0\\i&0&0\\0&0&0\end{pmatrix},\quad \lambda_3=\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix},$$

$$\lambda_4=\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix},\quad \lambda_5=\begin{pmatrix}0&0&-i\\0&0&0\\i&0&0\end{pmatrix},\quad \lambda_6=\begin{pmatrix}0&0&0\\0&0&1\\0&1&0\end{pmatrix},$$

$$\lambda_7=\begin{pmatrix}0&0&0\\0&0&-i\\0&i&0\end{pmatrix},\quad \lambda_8=\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-2\end{pmatrix}.$$

These matrices obey the anticommutation relations,

$$\{\lambda_i,\lambda_j\}=\tfrac{4}{3}\mathbb{1}\delta_{ij}+2d_{ijk}\lambda_k.$$

The nonzero components of the completely symmetric tensor d_{ijk} are

$$d_{118}=d_{228}=d_{338}=-d_{888}=\frac{1}{\sqrt{3}},\quad d_{448}=d_{558}=d_{668}=d_{778}=-\frac{1}{2\sqrt{3}},$$

$$d_{146}=d_{157}=-d_{247}=d_{256}=d_{344}=d_{355}=-d_{366}=-d_{377}=\tfrac{1}{2}.$$

Using the d_{ijk}, the following symmetric product between two elements in the algebra may be defined:

$$(\mathbf{a}\star\mathbf{b})_i=\sqrt{3}d_{ijk}a_jb_k. \tag{8}$$

Now the restrictions on the density matrix:

$$\rho^{\dagger}=\rho=\rho^2\geqslant0,\quad\text{and}\quad \mathrm{Tr}\ \rho=1,$$

are equivalent to the following conditions on n:

$$\mathbf{n}^*=\mathbf{n},\quad \mathbf{n}\cdot\mathbf{n}=1,\quad \mathbf{n}\star\mathbf{n}=\mathbf{n}. \tag{9}$$

A. Pure state density matrices

The density matrix for a pure state, e.g.,

$$\rho_0=\tfrac{1}{3}(\mathbb{1}-\sqrt{3}\lambda_8)=\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}, \tag{10}$$

can be mapped into an arbitrary pure state density matrix, ρ, by an SU(3) transformation:

$$\rho=D\rho_0D^{-1},\quad D\in\mathrm{SU}(3).$$

This mapping can be seen as a projection from SU(3) to SU(3)/SU(2) since the upper left 2×2 block of this matrix is invariant under a U(2) subgroup of SU(3). This subgroup is spanned by the parameters a, b, c, and ϕ. This is clear from the fact that the subalgebra K, defined in Sec. II A, commutes with λ_8. Thus, an arbitrary pure state density matrix for a three state system is an element of the coset space SU(3)/U(2). Moreover, this is expressed in a very natural way in terms of the Euler coordinates of the group manifold coordinates.

It may also be reexpressed in terms of the adjoint representation of the group, viz.,

$$\rho=\tfrac{1}{3}(\mathbb{1}-\sqrt{3}D\lambda_8 D^{-1})=\tfrac{1}{3}(\mathbb{1}-\sqrt{3}R_8^i\lambda_i).$$

This enables the identification of the matrix elements $-R_8^i$ with the vector $\mathbf{n}$ in (7) and one can show by direct computation that this vector satisfies the requirements (9).

B. Geometric phases for three state systems

Using the Euler coordinates for SU(3), in which

$$D=e^{(i\lambda_3\alpha)}e^{(i\lambda_2\beta)}e^{(i\lambda_3\gamma)}e^{(i\lambda_5\theta)}e^{(i\lambda_3 a)}e^{(i\lambda_2 b)}e^{(i\lambda_3 c)}e^{(i\lambda_8\phi)},$$

one can write down explicit coordinates for ψ. This is done by first calculating $\rho=|\psi\rangle\langle\psi|=\psi\psi^\dagger$ by the method above then, and then using $n_i=(\sqrt{3}/2)\psi^\dagger\lambda_i\psi$ to find ψ. The result is

$$\psi=e^{-2i\phi/\sqrt{3}}\begin{pmatrix} e^{i(\alpha+\gamma)}\cos\beta\sin\theta \\ e^{-i(\alpha-\gamma)}\sin\beta\sin\theta \\ \cos\theta \end{pmatrix}.$$

Using this, one can calculate the connection one-form and curvature two-form. The connection one-form for the geometric phase is

$$\mathcal{A}=\psi\, d\psi^\dagger. \tag{11}$$

From this, the connection one-form for the three state system is

$$\begin{aligned}\mathcal{A}&=\frac{2}{\sqrt{3}}\,d\phi+\sin^2\theta[\cos^2\beta(d\alpha+d\gamma)-\sin^2\beta(d\alpha-d\gamma)]\\ &=-\frac{2}{\sqrt{3}}\,d\phi+\sin^2\theta\cos 2\beta d\alpha+\sin^2\theta\, d\gamma,\end{aligned} \tag{12}$$

and the curvature two-form is given by

$$\begin{aligned}F=d\mathcal{A}&=-id\psi^\dagger\wedge d\psi=\sin 2\theta\cos^2\beta\, d\theta\wedge d(\alpha+\gamma)-\sin^2\theta\sin 2\beta\, d\beta\wedge d(\alpha+\gamma)\\ &\quad-\sin 2\theta\sin^2\beta\, d\theta\wedge d(\alpha-\gamma)-\sin^2\theta\sin 2\beta\, d\beta\wedge d(\alpha-\gamma)\\ &=\sin 2\theta\cos 2\beta\, d\theta\wedge d\alpha\\ &\quad-2\sin^2\theta\sin 2\beta\, d\beta\wedge d\alpha+\sin 2\theta\, d\theta\wedge d\gamma.\end{aligned} \tag{13}$$

The geometric phase around a closed loop is given, in general, by

$$\varphi_g=\int\mathcal{A},$$

and specifically for the three state systems by

$$\varphi_g=\int\sin^2\theta[\cos_2\beta(d\alpha+d\gamma)-\sin^2\beta(d\alpha-d\gamma)]=\int[\sin^2\theta\cos 2\beta\, d\alpha+\sin^2\theta\, d\gamma], \tag{14}$$

or by the integral of the curvature over the corresponding surface enclosed by this loop. The pull-back from SU(3)/U(2) to SU(3) of the connection one-form, $\mathscr{A}$, is expressible in terms of linear combinations of the one-forms on the group manifold. Let D be a point in SU(3) as above, and ρ a point in SU(3)/U(2). The push-forward (or derivative mapping) and pull-back maps denoted as ρ^*, and ρ_*, respectively, give the relationship between the vector fields and one-forms on the group manifold and coset space.

V. CONCLUSIONS/COMMENTS

The left and right invariant vector fields on SU(3) in an Euler angle parametrization were displayed explicitly. This enables the calculation of tensor products by direct computation (see, for instance, Ref. 22). In the case of three state systems there is a natural projection map from SU(3) to the coset space SU(3)/U(2) that describes the density matrix of pure states. Having a natural projection map one may use all of the fiber bundle techniques for computing the explicit structures in terms of the Euler coordinates. This enables the calculation of vector fields and forms on the coset space. It also provides a means of calculating geometric phases for the three state systems.

It was stated earlier that the methods for calculating the invariant vector fields and forms for SU(3) are generalizable to SU(n). Though this is true, the left invariant one-forms given here were calculated by hand. This is obviously not an enviable task for an SU(n) group since the dimension of these groups is n^2-1. It is possible to calculate the invariant vector fields with MATHEMATICA.[23] However, this program has some difficulty simplifying products of trigonometric functions of the form found in this calculation. As such programs become more proficient at symbolic manipulations of this sort, the calculations and analyses of such structures for higher-dimensional Lie groups will become more manageable.

ACKNOWLEDGMENTS

I would, first and foremost, like to thank Professor L. C. Biedenharn (now deceased) who, as my advisor, first started me on the construction of these structures. I could not give him too much credit here. I would also like to thank Professor E. C. G. Sudarshan for many helpful comments along with Professor Duane Dicus, whose help and support enabled the completion of this paper. I would like to thank David Cook for deriving the vector fields independently. This provided a check for the results below. This research was supported in part by the U.S. Department of Energy under Contract No. DE-EG013-93ER40757.

APPENDIX A: INVARIANT VECTOR FIELDS

The left invariant vector fields are as follows:

$$\Lambda_1 = i\cos 2\alpha\cot 2\beta\partial_1 + i\sin 2\alpha\partial_2 - i\frac{\cos 2\alpha}{\sin 2\beta}\partial_3, \tag{A1}$$

$$\Lambda_2 = -i\sin 2\alpha\cot 2\beta\partial_1 + i\cos 2\alpha\partial_2 + i\frac{\sin 2\alpha}{\sin 2\beta}\partial_3, \tag{A2}$$

$$\Lambda_3 = i\partial_1, \tag{A3}$$

$$\Lambda_8 = i\sqrt{3}\partial_3 - i\sqrt{3}\partial_5 + i\partial_8, \tag{A4}$$

$$\begin{aligned}\lambda_4 = {} & i\frac{\sin\beta}{\sin 2\beta}\cot\theta\cos(\alpha+\gamma)\partial_1 - i\sin\beta\cot\theta\sin(\alpha+\gamma)\partial_2 \\ & - i\cot 2\beta\sin\beta\cot\theta\cos(\alpha+\gamma)\partial_3 + i\frac{(2-\sin^2\theta)}{\sin 2\theta}\cos\beta\cos(\alpha+\gamma)\partial_3 \\ & + i\cos\beta\sin(\alpha+\gamma)\partial_4 - i2\frac{\cos\beta}{\sin 2\theta}\cos(\alpha+\gamma)\partial_5\end{aligned}$$

$$-i\,\frac{\cot 2b}{\sin\theta}\sin\beta\cos(\alpha-\gamma-2a)\partial_5+i\,\frac{\sin\beta}{\sin\theta}\sin(\alpha-\gamma-2a)\partial_6$$
$$+i\,\frac{\sin\beta}{\sin\theta\sin 2b}\cos(\alpha-\gamma-2a)\partial_7-\frac{\sqrt{3}}{2}\tan\theta\cos\beta\cos(\alpha+\gamma)\Lambda_8, \tag{A5}$$

$$\Lambda_5=-i\,\frac{\sin\beta}{\sin 2\beta}\cos\theta\sin(\alpha+\gamma)\partial_1-\sin\beta\cot\theta\cos(\alpha+\gamma)\partial_2$$
$$+i\cot 2\beta\sin\beta\cot\theta\sin(\alpha+\gamma)\partial_3-i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\cos\beta\sin(\alpha+\gamma)\partial_3$$
$$+i\cos\beta\cos(\alpha+\gamma)\partial_4+i2\,\frac{\cos\beta}{\sin 2\theta}\sin(\alpha+\gamma)\partial_5+i\,\frac{\cot 2b}{\sin\theta}\sin\beta\sin(\alpha-\gamma-2a)\partial_5$$
$$+i\,\frac{\sin\beta}{\sin\theta}\cos(\alpha-\gamma-2a)\partial_6$$
$$-i\,\frac{\sin\beta}{\sin\theta\sin 2b}\sin(\alpha-\gamma-2a)\partial_7+\frac{\sqrt{3}}{2}\tan\theta\cos\beta\sin(\alpha+\gamma)\Lambda_8, \tag{A6}$$

$$\Lambda_6=i\,\frac{\cos\beta}{\sin 2\beta}\cot\theta\cos(\alpha-\gamma)\partial_1+i\cos\beta\cot\theta\sin(\alpha-\gamma)\partial_2$$
$$-i\cot 2\beta\cos\beta\cot\theta\cos(\alpha-\gamma)\partial_3-i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\sin\beta\cos(\alpha-\gamma)\partial_3$$
$$+i\sin\beta\sin(\alpha-\gamma)\partial_4+i2\,\frac{\sin\beta}{\sin 2\theta}\cos(\alpha-\gamma)\partial_5-i\,\frac{\cot 2b}{\sin\theta}\cos\beta\cos(\alpha+\gamma+2a)\partial_5$$
$$-i\,\frac{\cos\beta}{\sin\theta}\sin(\alpha+\gamma+2a)\partial_6+i\,\frac{\cos\beta}{\sin\theta\sin 2b}\cos(\alpha+\gamma+2a)\partial_7$$
$$+\frac{\sqrt{3}}{2}\tan\theta\sin\beta\cos(\alpha-\gamma)\Lambda_8, \tag{A7}$$

$$\Lambda_7=i\,\frac{\cos\beta}{\sin 2\beta}\cot\theta\sin(\alpha-\gamma)\partial_1-i\cos\beta\cot\theta\cos(\alpha-\gamma)\partial_2-i\cot 2\beta\cos\beta\cot\sin(\alpha-\gamma)\partial_3$$
$$-i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\sin\beta\sin(\alpha-\gamma)\partial_3-i\sin\beta\cos(\alpha-\gamma)\partial_4+i2\,\frac{\sin\beta}{\sin 2\theta}\sin(\alpha-\gamma)\partial_5$$
$$-i\,\frac{\cot 2b}{\sin\theta}\cos\beta\sin(\alpha+\gamma+2a)\partial_5+i\,\frac{\cos\beta}{\sin\theta}\cos(\alpha+\gamma+2a)\partial_6$$
$$+i\,\frac{\cos\beta}{\sin\theta\sin 2b}\sin(\alpha+\gamma+2a)\partial_7+\frac{\sqrt{3}}{2}\tan\theta\sin\beta\sin(\alpha-\gamma)\Lambda_8, \tag{A8}$$

where $\eta=\phi/\sqrt{3}$. The right invariant vector fields are given by the following equations:

$$\Lambda_1^r=-i\cos 2c\cot 2b\partial_7-i\sin 2c\partial_6+i\,\frac{\cos 2c}{\sin 2b}\partial_5, \tag{A9}$$

$$\Lambda_2^r=-i\sin 2c\cot 2b\partial_7+i\cos 2c\partial_6+i\,\frac{\sin 2c}{\sin 2b}\partial_5, \tag{A10}$$

$$\Lambda_3^r = i\partial_7, \tag{A11}$$

$$\Lambda_8^r = i\partial_8, \tag{A12}$$

$$\begin{aligned}\Lambda_4^r = &-i\,\frac{\sin b}{\sin 2b}\cot\theta\cos(c+a+3\eta)\partial_7 + i\sin b\cot\theta\sin(c+a+3\eta)\partial_6\\ &+i\cot 2b\sin b\cot\theta\cos(c+a+3\eta)\partial_5 - i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\cos b\cos(c+a+3\eta)\partial_5\\ &-i\cos b\sin(c+a+3\eta)\partial_4 + i2\,\frac{\cos b}{\sin 2\theta}\cos(c+a+3\eta)\partial_3\\ &+i\,\frac{\cot 2\beta}{\sin\theta}\sin b\cos(c-a-2\gamma+3\eta)\partial_3 - i\,\frac{\sin b}{\sin\theta}\sin(c-a-2\gamma+3\eta)\partial_2\\ &-i\,\frac{\sin b}{\sin\theta\sin 2\beta}\cos(c-a-2\gamma+3\eta)\partial_1 - \frac{\sqrt{3}}{2}\tan\theta\cos b\cos(c+a+3\eta)\Lambda_8^r,\end{aligned} \tag{A13}$$

$$\begin{aligned}\Lambda_5^r = &-i\,\frac{\sin b}{\sin 2b}\cot\theta\sin(c+a+3\eta)\partial_7 - i\sin b\cot\theta\cos(c+a+3\eta)\partial_6\\ &+i\cot 2b\sin b\cot\theta\sin(c+a+3\eta)\partial_5 - i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\cos b\sin(c+a+3\eta)\partial_5\\ &+i\cos b\cos(c+a+3\eta)\partial_4 + i2\,\frac{\cos b}{\sin 2\theta}\sin(c+a+3\eta)\partial_3\\ &+i\,\frac{\cot 2\beta}{\sin\theta}\sin b\sin(c-a-2\gamma+3\eta)\partial_3 + i\,\frac{\sin b}{\sin\theta}\cos(c-a-2\gamma+3\eta)\partial_2\\ &-i\,\frac{\sin b}{\sin\theta\sin 2\beta}\sin(c-a-2\gamma+3\eta)\partial_1 - \frac{\sqrt{3}}{2}\tan\theta\cos b\sin(c+a+3\eta)\Lambda_8^r,\end{aligned} \tag{A14}$$

$$\begin{aligned}\Lambda_6^r = &\,i\,\frac{\cos b}{\sin 2b}\cot\theta\cos(c-a-3\eta)\partial_7 + i\cos b\cot\theta\sin(c-a-3\eta)\partial_6\\ &-i\cot 2b\cos b\cot\theta\cos(c-a-3\eta)\partial_5 - \frac{(2-\sin^2\theta)}{\sin 2\theta}\sin b\cos(c-a-3\eta)\partial_5\\ &+i\sin b\sin(c-a-3\eta)\partial_4 + i2\,\frac{\sin b}{\sin 2\theta}\cos(c-a-3\eta)\partial_3\\ &-i\,\frac{\cot 2\beta}{\sin\theta}\cos b\cos(c+a+2\gamma-3\eta)\partial_3 - i\,\frac{\cos b}{\sin\theta}\sin(c+a+2\gamma-3\eta)\partial_2\\ &+i\,\frac{\cos b}{\sin\theta\sin 2\beta}\cos(c+a+2\gamma-3\eta)\partial_1 - \frac{\sqrt{3}}{2}\tan\theta\sin b\cos(c-a-3\eta)\Lambda_8^r,\end{aligned} \tag{A15}$$

$$\begin{aligned}\Lambda_7^r = &-i\,\frac{\cos b}{\sin 2b}\cot\theta\sin(c-a-3\eta)\partial_7 + i\cos b\cot\theta\cos(c-a-3\eta)\partial_6\\ &+i\cot 2b\cos b\cot\theta\sin(c-a-3\eta)\partial_5 + i\,\frac{(2-\sin^2\theta)}{\sin 2\theta}\sin b\sin(c-a-3\eta)\partial_5\\ &+i\sin b\cos(c-a-3\eta)\partial_4 - i2\,\frac{\sin b}{\sin 2\theta}\sin(c-a-3\eta)\partial_3\end{aligned}$$

$$+i\,\frac{\cot 2\beta}{\sin\,\theta}\cos\,b\,\sin(c+a+2\gamma-3\eta)\partial_3-i\,\frac{\cos\,b}{\sin\,\theta}\cos(c+a+2\gamma-3\eta)\partial_2$$

$$-i\,\frac{\cos\,b}{\sin\,\theta\,\sin 2\beta}\sin(c+a+2\gamma-3\eta)\partial_1+\frac{\sqrt{3}}{2}\tan\,\theta\,\sin\,b\,\sin(c-a-3\eta)\Lambda_8^r. \quad \text{(A16)}$$

APPENDIX B: INVARIANT FORMS

The invariant forms on the manifold are given by the following equations:

$$\begin{aligned}\omega^1=&-i\,\sin(2\alpha)d\beta+i\,\cos(2\alpha)\sin(2\beta)d\gamma+i\,\cos(2\alpha)\sin(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))da\\&-i\,\cos(2a+2\gamma)\cos(\theta)\sin(2\alpha)db-i\,\cos(2\alpha)\cos(2\beta)\cos(\theta)\sin(2a+2\gamma)db\\&+i[\cos(2\alpha)\cos(2\beta)\cos(2a+2\gamma)\cos(\theta)\sin(2b)]dc\\&-i\,\cos(\theta)\sin(2\alpha)\sin(2b)\sin(2a+2\gamma)dc+i\,\cos(2\alpha)\cos(2b)\sin(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))dc\\&-i\,\frac{\sqrt{3}}{2}\cos(2\alpha)\sin(2\beta)\sin^2\,\theta d\phi,\end{aligned} \quad \text{(B1)}$$

$$\begin{aligned}\omega^2=&-i\,\cos(2\alpha)d\beta-i\,\sin(2\alpha)\sin(2\beta)d\gamma-i\,\sin(2\alpha)\sin(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))da\\&-i\,\cos(2\alpha)\cos(2a+2\gamma)\cos(\theta)db+i\,\cos(2\beta)\cos(\theta)\sin(2\alpha)\sin(2a+2\gamma)db\\&-i\,\cos(2\beta)\cos(2a+2\gamma)\cos(\theta)\sin(2\alpha)\sin(2b)dc\\&-i\,\cos(2\alpha)\cos(\theta)\sin(2b)\sin(2a+2\gamma)dc-i\,\cos(2b)\sin(2\alpha)\sin(2\beta)\\&\times(1-\tfrac{1}{2}\sin^2(\theta))dc+i\,\frac{\sqrt{3}}{2}\sin(2\alpha)\sin(2\beta)\sin^2(\theta)d\phi,\end{aligned} \quad \text{(B2)}$$

$$\begin{aligned}\omega^3=&-i\,d\alpha-i\,\cos(2\beta)d\gamma-i\,\cos(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))da-i\,\cos(\theta)\sin(2\beta)\sin(2a+2\gamma)db\\&+i\,\cos(2a+2\gamma)\cos(\theta)\sin(2b)\sin(2\beta)dc-i\,\cos(2b)\cos(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))dc\\&+i\,\frac{\sqrt{3}}{2}\cos(2\beta)\,\frac{1}{2}\sin^2(\theta)d\phi,\end{aligned} \quad \text{(B3)}$$

$$\begin{aligned}\omega^4=&-i\,\cos(\beta)\sin(\alpha+\gamma)d\theta+i\tfrac{1}{2}\cos(\beta)\cos(\alpha+\gamma)\sin(2\theta)da\\&+i\,\sin(\beta)\sin(2a-\alpha+\gamma)\sin(\theta)db-i\,\cos(2a-\alpha+\gamma)\sin(2b)\sin(\beta)\sin(\theta)dc\\&+i\tfrac{1}{2}\cos(2b)\cos(\beta)\cos(\alpha+\gamma)\sin(2\theta)dc+i\,\frac{\sqrt{3}}{2}\cos(\beta)\cos(\alpha+\gamma)\sin(2\theta)d\phi,\end{aligned} \quad \text{(B4)}$$

$$\begin{aligned}\omega^5=&-i\,\cos(\beta)\cos(\alpha+\gamma)d\theta-i\tfrac{1}{2}\cos(\beta)\sin(\alpha+\gamma)\sin(2\theta)da\\&-i\,\cos(2a-\alpha+\gamma)\sin(\beta)\sin(\theta)db-i\,\sin(2b)\sin(\beta)\sin(2a-\alpha+\gamma)\sin(\theta)dc\\&-i\tfrac{1}{2}\cos(2b)\cos(\beta)\sin(\alpha+\gamma)\sin(2\theta)dc-i\,\frac{\sqrt{3}}{2}\cos(\beta)\sin(\alpha+\gamma)\sin(2\theta)d\phi,\end{aligned} \quad \text{(B5)}$$

$$\begin{aligned}\omega^6=&-i\,\sin(\beta)\sin(\alpha-\gamma)d\theta+i\,\cos(\beta)\sin(2a+\alpha+\gamma)\sin(\theta)db\\&-i\tfrac{1}{2}\cos(\alpha-\gamma)\sin(\beta)\sin(2\theta)da-i\,\frac{\sqrt{3}}{2}\cos(\alpha-\gamma)\sin(\beta)\sin(2\theta)d\phi\\&-i\,\cos(\beta)\cos(2a+\alpha+\gamma)\sin(2b)\sin(\theta)dc\\&-i\tfrac{1}{2}\cos(2b)\cos(\alpha-\gamma)\sin(\beta)\sin(2\theta)dc,\end{aligned} \quad \text{(B6)}$$

$$\omega^7 = +i\ \cos(\alpha-\gamma)\sin(\beta)d\theta - i\tfrac{1}{2}\sin(\beta)\sin(\alpha-\gamma)\sin(2\theta)da$$
$$-i\ \cos(\beta)\cos(2a+\alpha+\gamma)\sin(\theta)db - i\ \cos(\beta)\sin(2b)\sin(2a+\alpha+\gamma)\sin(\theta)dc$$
$$-i\tfrac{1}{2}\cos(2b)\sin(\beta)\sin(\alpha-\gamma)\sin(2\theta)dc - i\ \frac{\sqrt{3}}{2}\sin(\beta)\sin(\alpha-\gamma)\sin(2\theta)d\phi, \qquad \text{(B7)}$$

$$\omega^8 = i\ \frac{\sqrt{3}}{2}\sin^2(\theta)da + i\ \frac{\sqrt{3}}{2}\cos(2b)\sin^2(\theta)dc - i\left(1-\frac{3}{2}\sin^2(\theta)\right)d\phi. \qquad \text{(B8)}$$

The right invariant forms are given by

$$\omega_r^1 = i\ \cos(2b)\cos(2c)\cos(2a+2\gamma)\cos(\theta)\sin(2\beta)d\alpha - i\ \cos(\theta)\sin(2\beta)\sin(2c)\sin(2a+2\gamma)d\alpha$$
$$+i\ \cos(2\beta)\cos(2c)\sin(2b)(1-\tfrac{1}{2}\sin^2(\theta))d\alpha + i\ \cos(2c)\sin(2b)(1-\tfrac{1}{2}\sin^2(\theta))d\beta$$
$$-i\ \cos(2a+2\gamma)\cos(\theta)\sin(2c)d\gamma - i\ \cos(2b)\cos(2c)\cos(\theta)\sin(2a+2\gamma)d\gamma$$
$$+i\ \cos(2c)\sin(2b)(1-\tfrac{1}{2}\sin^2(\theta))d\theta - i\ \sin(2c)db, \qquad \text{(B9)}$$

$$\omega_r^2 = -i\ \cos(2b)\cos(2a+2\gamma)\cos(\theta)\sin(2\beta)\sin(2c)d\alpha - i\ \cos(2c)\cos(\theta)\sin(2\beta)\sin(2a$$
$$+2\gamma)d\alpha - i\ \cos(2\beta)\sin(2b)\sin(2c)(1-\tfrac{1}{2}\sin^2(\theta))d\alpha$$
$$-i\ \sin(2b)\sin(2c)(1-\tfrac{1}{2}\sin^2(\theta))d\beta - i[\cos(2c)\cos(2a+2\gamma)\cos(\theta)$$
$$-\cos(2b)\cos(\theta)\sin(2c)\sin(2a+2\gamma)]d\gamma - i\ \sin(2b)\sin(2c)(1-\tfrac{1}{2}\sin^2(\theta))d\theta$$
$$+\cos(2c)db, \qquad \text{(B10)}$$

$$\omega_r^3 = i\ \cos(2a+2\gamma)\cos(\theta)\sin(2b)\sin(2\beta)d\alpha - i\ \cos(2b)\cos(2\beta)(1-\tfrac{1}{2}\sin^2(\theta))d\alpha$$
$$-i\ \cos(2b)(1-\tfrac{1}{2}\sin^2(\theta))d\beta - i\ \cos(\theta)\sin(2b)\sin(2a+2\gamma)d\gamma$$
$$-i\ \cos(2b)(1-\tfrac{1}{2}\sin^2(\theta))d\theta + dc, \qquad \text{(B11)}$$

$$\omega_r^4 = -i\ \cos(a-c+2\gamma-3\eta)\sin(b)\sin(2\beta)\sin(\theta)d\alpha + i\tfrac{1}{2}\cos(b)\cos(2\beta)$$
$$\times\cos(a+c+3\eta)\sin(2\theta)d\alpha + i\tfrac{1}{2}\cos(b)\cos(a+c+3\eta)\sin(2\theta)d\beta$$
$$+i\ \sin(b)\sin(\theta)\sin(a-c+2\gamma-3\eta)d\gamma$$
$$+i\ \tfrac{1}{2}\cos(b)\cos(a+c+3\eta)\sin(2\theta)d\theta - \cos(b)\sin(a+c+3\eta)da, \qquad \text{(B12)}$$

$$\omega_r^5 = -i\ \sin(b)\sin(2\beta)\sin(\theta)\sin(a-c+2\gamma-3\eta)d\alpha - i\tfrac{1}{2}\cos(b)\cos(2\beta)$$
$$\times\sin(2\theta)\sin(a+c+3\eta)d\alpha - i\tfrac{1}{2}\cos(b)\sin(2\theta)\sin(a+c+3\eta)d\beta$$
$$-i\ \cos(a-c+2\gamma-3\eta)\sin(b)\sin(\theta)d\gamma - i\tfrac{1}{2}\cos(b)\sin(2\theta)\sin(a+c+3\eta)d\theta$$
$$-i\ \cos(b)\cos(a+c+3\eta)da, \qquad \text{(B13)}$$

$$\omega_r^6 = -i\ \cos(b)\cos(a+c+2\gamma-3\eta)\sin(2\beta)\sin(\theta)d\alpha$$
$$-i\tfrac{1}{2}\cos(2\beta)\cos(a-c+3\eta)\sin(b)\sin(2\theta)d\alpha - i\tfrac{1}{2}\cos(a-c+3\eta)\sin(b)\sin(2\theta)d\beta$$
$$-i\ \cos(b)\sin(\theta)\sin(a+c+2\gamma-3\eta)d\gamma - i\tfrac{1}{2}\cos(a-c+3\eta)\sin(b)\sin(2\theta)d\theta$$
$$+i\ \sin(b)\sin(a-c+3\eta)da, \qquad \text{(B14)}$$

$$\omega_r^7 = -i\ \cos(b)\sin(2\beta)\sin(\theta)\sin(a+c+2\gamma-3\eta)d\alpha + i\tfrac{1}{2}\cos(2\beta)\sin(b)\sin(2\theta)$$
$$\times\sin(a-c+3\eta)d\alpha + i\tfrac{1}{2}\sin(b)\sin(2\theta)\sin(a-c+3\eta)d\beta$$
$$+ i\ \cos(b)\cos(a+c+2\gamma-3\eta)\sin(\theta)d\gamma$$
$$+ i\tfrac{1}{2}\sin(b)\sin(2\theta)\sin(a-c+3\eta)d\theta + i\ \cos(a-c+3\eta)\sin(b)da, \tag{B15}$$

$$\omega_r^8 = i\ \frac{\sqrt{3}}{2}\cos(2\beta)\sin^2(\theta)d\alpha + i\ \frac{\sqrt{3}}{2}\sin^2(\theta)d\beta + i\ \frac{\sqrt{3}}{2}\sin^2(\theta)d\theta - i\ d\phi. \tag{B16}$$

[1] S. Weinberg, *The Quantum Theory of Fields* (Cambridge University Press, Cambridge, 1996), Vol. II Sec. 18.7.
[2] M. Gell-Mann and Y. Ne'eman, *The Eight-Fold Way* (Benjamin, New York, 1964).
[3] J. P. Elliot, ''Collective motion in the nuclear shell model I. Classification schemes for states of mixed configurations,'' Proc. R. Soc. London, Ser. A **245**, 128–145 (1958); and ''Collective motion in the nuclear shell, model II. The introduction of intrinsic wave-functions,'' *ibid.* **245**, 562–581 (1958).
[4] Arvind, K. S. Mallesh, and N. Mukunda, ''A generalized Pancharatnam geometric phase formula fo three-level systems,'' J. Phys. A **30**, 2417–2431 (1997).
[5] G. Khanna, S. Mukhopadhyay, R. Simon, and N. Mukunda, ''Geometric phases for SU(3) representations and three level quantum systems,'' Ann. Phys. (N.Y.) **253**, 55–82 (1997).
[6] There have been other methods of decomposing the group into an Euler angle parametrization. See. Ref. 16 and D. A. Akyeampong and M. A. Rashid, ''On the finite transformations of SU(3),'' J. Math. Phys. **13**, 1218–1224 (1972).
[7] For an excellent review and a collection of reprints, see *Geometric Phases in Physics*, edited by A. Shapere and F. Wilczek (World Scientific, Singapore, 1989).
[8] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics Theory and Application* (Cambridge University Press, Cambridge, 1977), p. 57.
[9] R. Hermann, *Lie Groups for Physicists* (Benjamin, New York, 1966), pp. 30–36.
[10] L. C. Biedenharn (private communication). See also L. C. Biedenharn and Y. Dothan, ''Monopolar harmonics in SU(3) as eigenstates of the Skyrme–Witten model for baryons,'' in *From SU(3) to Gravity* (Cambridge University Press, Cambridge, 1986), pp. 19–34.
[11] T. Bröcker and T. tom Dieck, *Representations of Compact Lie Groups* (Springer-Verlag, New York, 1995) pp. 10, 214.
[12] It is conventional to use D^* for the wave functions in SU(2) (see, for instance, Ref. 8, p. 58). This convention is followed here. The necessary distinction between the two fundamental representations of SU(3) is made when needed in Ref. 17.
[13] See Ref. 8, p. 61.
[14] Y. Choquet-Bruhat and C. DeWitt-Morette, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1982), Part I, p. 164.
[15] C. Isham, *Modern Differential Geometry for Physicists* (World Scientific, Singapore, 1989), p. 72.
[16] T. J. Nelson, ''A set of harmonic functions for the group SU(3) as specialized matrix elements of a general finite transformation,'' J. Math. Phys. **8**, 857–863 (1967).
[17] M. Byrd and E. C. G. Sudarshan, ''SU(3) revisited,'' physics/9803029.
[18] D. F. Holland, ''Finite transformations of SU(3),'' J. Math. Phys. **10**, 531–535 (1969).
[19] M. S. Marinov, ''Invariant volumes of compact groups,'' J. Phys. A **13**, 3357–3366 (1980).
[20] M. Nakahara, *Geometry, Topology and Physics* (IOP Publishing, Philadelphia, 1990), pp. 29, 364.
[21] J. Anandan, J. Christian, and K. Wanelik, ''Resource Letter GPP-1: geometric phases in physics,'' quant-ph/9702011.
[22] Reference 14, p. 138.
[23] A private conversation with D. Cook.

An existence theory for relativistic brachistochrones in stationary space–times

Fabio Giannoni
Dipartimento di Energetica, Universitá de L'Aquila, Italy

Paolo Piccione
Departamento de Matemática, Universidade de São Paulo, São Paulo, Brazil

(Received 16 March 1998; accepted for publication 11 May 1998)

We use a variational principle proven by Giannoni *et al.* [J. Math. Phys. **38**, 6367–6381 (1997)] for relativistic brachistochrones and techniques of global analysis to develop an existence theory for brachistochrones in stationary space–times. We also develop a Ljusternik–Schnirelman theory for brachistochrones, obtaining multiplicity results depending on the topology of the space–time. © *1998 American Institute of Physics.* [S0022-2488(98)01111-6]

I. INTRODUCTION

The classical formulation of the brachistochrone problem is to determine the shape of a frictionless slide between two points in space in such a way that a particle, if released from rest at the first point and driven by a homogeneous gravitational field, would reach the second point in the shortest time. This problem has several generalizations, e.g., the homogeneous gravitational field could be replaced with an arbitrary Newtonian potential, and, instead of releasing the particle from rest, one could prescribe an arbitrary value for the initial speed, leaving the initial direction of the velocity undetermined. Moreover, the brachistochrone problem can also be formulated in the context of general relativity.

The first relativistic versions of the brachistochrone problem appear in Refs. 1 and 2. Perlick (Ref. 3; see also Kamath) has determined the brachistochrone equation in a *regular* stationary Lorentzian manifold, i.e., in a time-independent gravitational field according to general relativity, and the two authors, together with Verderesi, in Ref. 4 have generalized Perlick's result to the case of an arbitrary stationary manifold by reformulating the brachistochrone problem in the context of sub-Riemannian geometry.

If one drops the stationarity assumption for the Lorentzian manifold, the brachistochrone problem can still be formulated in a physically meaningful way (see Ref. 5) and, also in the general case, using the Lagrange multiplier technique one is able to derive a system of differential equations for the brachistochrones and for the Lagrange multipliers. Unfortunately, it does not seem possible to eliminate the Lagrange multipliers from this system without introducing integrals, unless in the stationary case. Thus, it looks as if the brachistochrones in the nonstationary case are not determined by a second-order differential equation, but rather by an integro-differential equation.

The aim of this paper is to use the variational principle for brachistochrones in a stationary space–time proven in Ref. 4 and techniques from critical point theory to develop an existence theory for brachistochrones of fixed energy between an event p and an observer γ.

In order to make the paper self-contained and for the reader's convenience, in Sec. II we will recall the main results of Ref. 4, together with a few other remarks needed in the rest of the paper.

To make our statements precise, we introduce some notation and terminology.

We will denote by $\mathscr{M}$ an $(n+1)$-dimensional smooth manifold, $n \geqslant 1$, and by g a Lorentzian metric on $\mathscr{M}$. More precisely, we assume that $\mathscr{M}$ is a finite-dimensional real C^∞ manifold of dimension greater than two whose topology satisfies the second countability axiom and the Hausdorff separation axiom, and we assume that g is a pseudo-Riemannian C^∞ metric of signature $(+,\ldots,+,-)$ on $\mathscr{M}$. The physically interesting case is, of course, $\dim(\mathscr{M})=4$.

We will also denote by $\langle \cdot,\cdot \rangle$ the bilinear form induced on the tangent space $T_q\mathscr{M}$ by $g(q)$.

Let Y be a smooth timelike Killing vector field on $\mathscr{M}$, i.e., $\langle Y,Y\rangle<0$ everywhere, and the

 © 1998 American Institute of Physics

metric tensor g is invariant by the flow of Y. A Lorentzian manifold endowed with such a vector field is said to be *stationary*.

We will also assume that $\mathscr{M}$ is *causal*, i.e., there exists no closed causal curve in $\mathscr{M}$. We recall that a smooth curve γ is causal if the tangent vector $\dot{\gamma}(s)$ is causal ($\langle \dot{\gamma}(s), \dot{\gamma}(s)\rangle \leqslant 0$) everywhere. Physically, the causality condition on $\mathscr{M}$ means that the possibility that any object, massive or massless, in the universe, traveling at speed less than or equal to the speed of light (in vacuum), may encounter itself at an earlier stage of its life is excluded.

For the main properties of Lorentzian geometry and the basics of general relativity we refer to the classical textbooks, like Refs. 6–8.

Let $\gamma:]a,b[\mapsto \mathscr{M}$ be a *maximal* integral curve of Y and $p \notin \gamma$ be an event of $\mathscr{M}$ (we will also denote by γ the image $\gamma(]a,b[)$ in $\mathscr{M}$). By the causality assumption on $\mathscr{M}$, the curve γ is a *closed embedding* of $]a,b[$ into $\mathscr{M}$, i.e., the relative topology of γ in $\mathscr{M}$ coincides with the topology of $]a,b[$. Moreover, the maximality of γ implies that γ has no *end points* in $\mathscr{M}$, i.e., $\gamma(s)$ is eventually outside of any compact subset of $\mathscr{M}$ for $s \to a,b$.

Let $k \in \mathbb{R}^+$ be a fixed constant. We consider the space $\mathscr{B}^+_{p,\gamma}(k)$ of all piecewise smooth future pointing, timelike curves z from p to γ, parametrized by proper time, and satisfying the conservation law $-\langle \dot{z}, Y\rangle \equiv k > 0$:

$$\mathscr{B}^+_{p,\gamma}(k) = \{z:[0,T_z] \mapsto \mathscr{M} : z \text{ of class } C^1,$$

$$z(0) = p, z(T_z) \in \gamma, \langle \dot{z}, \dot{z}\rangle \equiv -1,$$

$$-\langle \dot{z}, Y\rangle \equiv k > 0 (\text{constant})\}. \tag{1}$$

The curve γ is interpreted as the world line of an observer; the curves z in $\mathscr{B}^+_{p,\gamma}(k)$ represent the world line of massive particles starting at the event p and eventually reaching the observer γ. The condition $\langle \dot{z}, \dot{z}\rangle = -1$ means that z is parametrized by proper time, so that T_z is the *travel time* of z measured on the particle. The quantity $\langle \dot{z}, Y\rangle$ is the component along Y of the four-momentum of the particle z; observe that, since Y is Killing, then it is constant on geodesics. The condition $-\langle \dot{z}, Y\rangle \equiv k$ is the analog of the conservation law of the energy in the Newtonian case.

The positivity of the constant k says that z is future pointing with respect to the time orientation induced on $\mathscr{M}$ by Y. A perfectly similar theory may be developed considering past pointing curves. If one modifies properly the definitions, it is easily checked that all the results proven in this paper work equally well in this case.

Let us consider the open subset $U_k \subset \mathscr{M}$:

$$U_k = \{q \in \mathscr{M} : \langle Y(q), Y(q)\rangle + k^2 > 0\}. \tag{2}$$

Observe that, since Y is Killing, then $\langle Y, Y\rangle$ is constant on the flow lines of Y, and so U_k is invariant by the flow of Y. Observe that, by the wrong way Schwartz's inequality, if $z \in \mathscr{B}^+_{p,\gamma}(k)$ then z is contained in the closure of U_k. Indeed, $-\langle Y, Y\rangle = \langle \dot{z}, \dot{z}\rangle \langle Y, Y\rangle \leqslant \langle \dot{z}, Y\rangle^2$. Moreover, if $-\langle \dot{z}, Y\rangle \equiv k$, then z is entirely contained in U_k if and only if the tangent vector $\dot{z}$ is never a multiple of Y.

We will assume that γ is a maximal integral curve of Y. For all q in $\mathscr{M}$, γ_q will denote the maximal integral line of Y such that $\gamma_q(0) = q$.

Definition I.1: A curve $z \in \mathscr{B}^+_{p,\gamma}(k)$ is said to be a *minimal brachistochrone of energy k* between p and γ if z is a minimal point for the *travel time* functional $F: \mathscr{B}^+_{p,\gamma}(k) \to \mathbb{R}^+$ given by

$$F(z) = T_z.$$

A curve $z \in \mathscr{B}^+_{p,\gamma}(k)$ is a *brachistochrone of energy k* if it minimizes locally its travel time, i.e., if for all $r \in [0,T_z[$ and for $\varepsilon \in]0,T_z - r[$ small enough, the curve $\overline{z}(s) = z(s-r)$, defined on the interval $[0,\varepsilon]$, is a minimal brachistochrone between $z(s)$ and $\gamma_{z(s+\varepsilon)}$.

It is proven in Ref. 4 that the brachistochrones are smooth curves in $\mathscr{M}$ that satisfy the following differential equation:

$$\nabla_{\dot z}\dot z+\frac{2k}{\langle Y,Y\rangle}\nabla_{\dot z}Y+2k^2\frac{\langle\nabla_z Y,Y\rangle}{\langle Y,Y\rangle(k^2+\langle Y,Y\rangle)}\dot z+-2k\frac{\langle\nabla_{\dot z}Y,Y\rangle}{\langle Y,Y\rangle(k^2+\langle Y,Y\rangle)}Y=0, \tag{3}$$

where ∇ is the covariant derivative with respect to the Levi-Civita connection of the Lorentzian metric g.

If Y is a complete vector field on $\mathscr{M}$, then, using the flow of Y, the brachistochrones having fixed energy k are deformed into geodesics with respect to a suitable (sub-)Riemannian structure defined on $\mathscr{M}$.

Our definition of brachistochrones extends the definition of the *τ-brachistochrones* given in Ref. 3 in the case of a regular stationary manifold. In Ref. 3 it is also defined a different brachistochrone problem (in Perlick's terminology: the *t-brachistochrones*), which are curves $z\in\mathscr{B}^+_{p,\gamma}(k)$ minimizing the *arrival time* functional:

$$AT(z)=\gamma^{-1}(z(T_z)),\quad z\in\mathscr{B}^+_{p,\gamma}(k), \tag{4}$$

where γ is assumed to be injective.

The two variational problems are essentially different; as it was shown in Ref. 3, the *t*-brachistochrone problem is also reduced to a Riemannian variational problem consisting of the sum of a length functional and a linear functional in the derivative, representing a sort of *Coriolis* force. In this paper we will only deal with the variational problem of Definition I.1, while the second brachistochrone problem is the subject of a forthcoming paper.

As it is natural to expect, in order to obtain existence results for brachistochrones, one needs to assume the completeness of $\mathscr{M}$ with respect to some Riemannian structure. In the statements of our results we assume the completeness of $\mathscr{M}$ with respect to the Riemannian metric g_R, defined by $g_R[v,v]=g[v,v]$ for v orthogonal to Y, and $g_R[Y,Y]=-g[Y,Y]$ [see formula (7)].

Theorem I.2: *Let $\mathscr{M}$ be a stationary Lorentzian manifold, Y a timelike Killing vector field on $\mathscr{M}$ and $k\in\mathbb{R}^+$ a fixed real constant.*

Suppose that Y is complete, k^2 is a regular value for the function $-\langle Y,Y\rangle$, and assume that Y is bounded from below in U_k, i.e., there exists a positive constant $\nu>0$ such that:

$$-\langle Y(q),Y(q)\rangle\geqslant\nu>0,\quad\forall q\in U_k. \tag{5}$$

Let p be a point in U_k and γ a maximal integral line of Y inside U_k.

Then, if $\overline{U_k}=U_k\cup\partial U_k$ is complete with respect to the Riemannian metric g_R, there exists at least one brachistochrone z of energy k joining p and γ in U_k.

Under the extra assumption that the closed set $\overline{U_k}$ be noncontractible, we also have the following multiplicity result for brachistochrones of fixed energy:

Theorem I.3: *Under the hypotheses of Theorem I.2, if $\overline{U_k}$ is not contractible, then there exists a sequence $\{z_n\}_{n\in\mathbb{N}}$ of brachistochrones of energy k between p and γ such that:*

$$\lim_{n\to\infty}T_{z_n}=+\infty. \tag{6}$$

Theorem I.3 is the analog of Serre's Theorem (see Ref. 9) concerning the multiplicity of geodesics joining two fixed points in a noncontractible Riemannian manifold.

Theorems I.2 and I.3 will be proven in the paper, which is organized as follows. In Sec. II we recall the variational principle for brachistochrones in stationary space–times, relating the brachistochrone curves (with fixed energy) to geodesics in suitable Riemannian and sub-Riemannian structures. In particular, we point out that, using the flow of the vector field Y, the brachistochrones of fixed energy k in $\mathscr{M}$ are deformed into geodesics with respect to the Riemannian metric $\phi_k\cdot g_R$ of $\mathscr{M}$, where $\phi_k=-\langle Y,Y\rangle(k^2+\langle Y,Y\rangle)^{-1}$, hence they are critical points for the Riemannian energy functional E_k associated with this metric.

In Sec. III, using a classical approach to the study of geodesics in Riemannian manifolds with convex boundary, we introduce a *penalized* functional $E_{k,\varepsilon}$, which is a perturbation of E_k. We

prove that, under our hypotheses, the functional $E_{k,\varepsilon}$ satisfies good compactness properties, like the Palais–Smale condition. Such property is essential to obtain existence results for critical points for $E_{k,\varepsilon}$.

In Sec. IV we prove some estimates on the critical points of $E_{k,\varepsilon}$ which do not depend on the parameter ε. These estimates are then used to pass to the limit as $\varepsilon \to 0$, obtaining the proof of Theorem I.2.

Finally, in Sec. V we develop a Ljusternik–Schnirelman theory for the critical points of the penalized functional $E_{k,\varepsilon}$ that, using the estimates proven in Sec. IV, allow us to prove the multiplicity result of Theorem I.3.

We will assume henceforth that the hypotheses of Theorem I.2 are satisfied.

II. THE VARIATIONAL PRINCIPLE FOR BRACHISTOCHRONES

In this section we present our functional framework and we recall the variational principle for brachistochrones and the main results of Ref. 4. The reader will find in Ref. 4 all the details and the proofs of the claims made in this section.

First of all, we define an auxiliary Riemannian structure on $\mathscr{M}$. We denote by Δ the orthogonal distribution to the vector field Y. Since Y is timelike, the wrong-way Schwartz's inequality implies that the restriction of the Lorentzian metric g is positive definite. We consider the following (0,2)-tensor on $\mathscr{M}$:

$$g_R[v,v] = \langle v,v \rangle_{(R)} = \langle v,v \rangle - 2\frac{\langle v, Y(q)\rangle^2}{\langle Y(q), Y(q)\rangle}, \tag{7}$$

where $q \in \mathscr{M}$ and $v \in T_q\mathscr{M}$. The wrong-way Schwartz's inequality implies that $g_R[v,v]$ defines a positive definite, symmetric tensor on $\mathscr{M}$. It is easily checked that the flow of Y preserves g_R, so that Y is Killing with respect to g_R also.

Let $\nabla^{(R)}$ be the covariant derivative with respect to the Levi-Civita connection of g_R. Using standard notation, for a smooth function $\phi: \mathscr{M} \mapsto \mathbb{R}$ we also denote by $\nabla\phi$ ($\nabla_{(R)}\phi$) the *Lorentzian (Riemannian) gradient* of ϕ with respect to g (g_R), which is the vector field defined by $d\phi(q)[\,\cdot\,] = \langle \nabla\phi(q), \cdot \rangle$ ($d\phi(q)[\,\cdot\,] = \langle \nabla_{(R)}\phi(q), \cdot \rangle_{(R)}$), for all $q \in \mathscr{M}$.

Since Y is Killing with respect to g, then for every pair v_1, v_2 of smooth vector fields on $\mathscr{M}$, it is (see Ref. 8):

$$\langle \nabla_{v_1} Y, v_2 \rangle + \langle \nabla_{v_2} Y, v_1 \rangle = 0. \tag{8}$$

The characterization (8) of the Killing property of Y will be used implicitly or explicitly throughout the paper.

We introduce the following functions on U_k:

$$\psi_k(q) = \langle Y(q), Y(q)\rangle + k^2, \quad \text{and} \quad \phi_k(q) = -\frac{\langle Y(q), Y(q)\rangle}{\langle Y(q), Y(q)\rangle + k^2}. \tag{9}$$

Observe that ϕ_k and ψ_k are positive and smooth; moreover, by definition of U_k, they are positive in U_k. Note also that:

$$\partial U_k = \psi_k^{-1}(0), \tag{10}$$

and the condition that k^2 is a regular value for the function $-\langle Y, Y\rangle$ says that the derivative ψ_k' is nonzero on ∂U_k:

$$\psi_k' \neq 0 \quad \text{on} \quad \partial U_k. \tag{11}$$

By the implicit function theorem, (11) implies that ∂U_k is a smooth submanifold of $\mathscr{M}$.

Using the Riemannian metric, we define the space $L^2([0,1], T\mathscr{M})$, consisting of all measurable maps $\zeta: [0,1] \mapsto T\mathscr{M}$ such that $\langle \zeta, \zeta \rangle_{(R)}$ is Lebesgue integrable on $[0,1]$.

The Sobolev space $H^{1,2}([0,1],\mathscr{M})$ is defined as the set of all absolutely continuous curves $z:[0,1]\mapsto\mathscr{M}$ such that $\dot{z}$ is in $L^2([0,1],T\mathscr{M})$. By a classical result of global analysis on manifolds, the spaces $L^2([0,1],T\mathscr{M})$ and $H^{1,2}([0,1],\mathscr{M})$ actually do not depend on the choice of the Riemannian metric on $\mathscr{M}$ (see Ref. 10).

We will also use the spaces $L^p([0,1],T\mathscr{M})$, defined in an obvious way using the Riemannian metric of $\mathscr{M}$, and the Sobolev spaces $H^{m,p}([0,1],\mathscr{M})$, with $m\in\mathbb{N}$ and $p\in[1,+\infty]$. These can be defined as the set of curves of class C^{m-1} in $\mathscr{M}$ possessing a weak mth derivative that belongs to $L^p([0,1],T\mathscr{M})$.

Our functional framework consists of the following spaces:

$$\Omega_{p,\gamma}=\{w\in H^{1,2}([0,1],U_k)\,:\,w(0)=p,w(1)\in\gamma\},\tag{12}$$

$$\Omega_{p,\gamma}(\Delta)=\{w\in\Omega_{p,\gamma}\,:\,\dot{w}(s)\in\Delta\text{ a.e. on}[0,1]\}.\tag{13}$$

Moreover, for $q\in\gamma$, we denote by $\Omega_{p,q}$ the space of all curves $w\in\Omega_{p,\gamma}$ such that $w(1)=q$.

It is well known (see Ref. 10) that $\Omega_{p,\gamma}$ has the structure of a smooth Hilbertian manifold; for $w\in\Omega_{p,\gamma}$ the tangent space $T_w\Omega_{p,\gamma}$ is identified with the Hilbert space:

$$T_w\Omega_{p,\gamma}=\{\zeta:[0,1]\mapsto T\mathscr{M}:\zeta\text{absolutely continuous},$$
$$\zeta(0)=0,\zeta(1)\|Y(w(1)),\quad \nabla_{\dot{w}}\zeta\in L^2([0,1],T\mathscr{M}),\quad \zeta(s)\in T_{w(s)}\mathscr{M}\forall s\},\tag{14}$$

endowed with the Hilbert norm:

$$\|\zeta\|_w=\left(\int_0^1\langle\nabla_{\dot{w}}\zeta,\nabla_{\dot{w}}\zeta\rangle_{(R)}\,ds\right)^{1/2}.$$

Observe that, due to the presence of the boundary ∂U_k, the manifold $\Omega_{p,\gamma}$ (and consequently of $\Omega_{p,\gamma}(\Delta)$) is *not* complete. Namely, if z_n is a Cauchy sequence of curves in $\Omega_{p,\gamma}$, then z_n may converge to a curve z whose image contain points of ∂U_k.

For all $q\in\gamma$, $\Omega_{p,q}$ is a closed submanifold of $\Omega_{p,\gamma}$, and for $w\in\Omega_{p,q}$ the tangent space $T_w\Omega_{p,q}$ is given by

$$T_w\Omega_{p,q}=\{\zeta\in T_w\Omega_{p,\gamma}:\zeta(1)=0\}.\tag{15}$$

In Ref. 4, it is also proven that $\Omega_{p,\gamma}(\Delta)$ is a smooth submanifold of $\Omega_{p,\gamma}$ and that, for $w\in\Omega_{p,\gamma}(\Delta)$, the tangent space $T_w\Omega_{p,\gamma}(\Delta)$ is given by

$$T_w\Omega_{p,\gamma}(\Delta)=\{\zeta\in T_w\Omega_{p,\gamma}:\langle\nabla_{\dot{w}}\zeta,Y\rangle+\langle\dot{w},\nabla_\zeta Y\rangle\equiv 0\}.\tag{16}$$

We can define the following functional E_k on $\Omega_{p,\gamma}(\Delta)$:

$$E_k(w)=\frac{1}{2}\int_0^1\phi_k(w)\cdot\langle\dot{w},\dot{w}\rangle_{(R)}ds.\tag{17}$$

It is easy to see that E_k is smooth on $\Omega_{p,\gamma}$, and so is its restriction to $\Omega_{p,\gamma}(\Delta)$, and, for $\zeta\in T_w\Omega_{p,\gamma}$, its Gateaux derivative $E_k'(w)[\zeta]$ is given by:

$$E_k'(w)[\zeta]=\frac{1}{2}\int_0^1\langle\nabla_{\dot{w}}^{(R)}\phi_k(w),\zeta\rangle_{(R)}\cdot\langle\dot{w},\dot{w}\rangle_{(R)}ds+\int_0^1\phi_k(w)\cdot\langle\nabla_{\dot{w}}^{(R)}\zeta,\dot{w}\rangle_{(R)}ds.\tag{18}$$

We have the following result concerning the critical points of E_k in $\Omega_{p,\gamma}$:

Proposition II.1: Let w be a curve in $\Omega_{p,\gamma}$. *Then,* $E_k'(w)=0$ *if and only if* $w\in\Omega_{p,\gamma}(\Delta)$ *and w is a critical point of the restriction of* E_k *to* $\Omega_{p,\gamma}(\Delta)$. *Such critical points are precisely the geodesics with respect to the Riemannian metric* $\phi_k\cdot g_R$ *on* $\mathscr{M}$ *that join p and* γ *and that are orthogonal to* γ.

Proof: It is an easy observation that, since ϕ_k is constant along the flow lines of Y, then Y is Killing also with respect to the Riemannian metric $\phi_k\cdot g_R$. Since E_k is the energy functional

associated to the Riemannian metric $\phi_k \cdot g_R$, by a well-known result in Riemannian geometry, the set of its critical points in $\Omega_{p,\gamma}$ are precisely the $(\phi_k \cdot g_R)$-geodesics w in $\mathscr{M}$ that arrive orthogonally to γ: $g_R[\dot{w}(1), \dot{\gamma}(w(1))] = 0$. Since γ is the flow line of the Killing vector field Y, then, $\dot{w}$ is orthogonal to Y everywhere, and so $w \in \Omega_{p,\gamma}(\Delta)$. The rest of the proof is easily obtained from Lemma 2.6 of Ref. 4. □

A critical point of E_k in $\Omega_{p,\gamma}$ will be called a *normal* $(\phi_k \cdot g_R)$*-geodesic* between p and γ.

Let's assume now that Y is complete on $\mathscr{M}$, i.e., its integral lines are curves defined on the entire real line. We denote by $\psi: \mathscr{M} \times \mathbb{R} \mapsto \mathscr{M}$ the flow of Y.

We use the flow of Y to define a map from the space $\mathscr{B}^+_{p,\gamma}(k)$ to the space $\Omega_{p,\gamma}(\Delta)$ as follows. First of all, we introduce the space $\tilde{\mathscr{B}}^+_{p,\gamma}(k)$ consisting of all curves in $\mathscr{B}^+_{p,\gamma}(k)$ suitably reparametrized on the interval $[0,1]$:

$$\tilde{\mathscr{B}}^+_{p,\gamma}(k) = \left\{ \sigma:[0,1] \mapsto \mathscr{M} \text{ of class } C^1 \;\middle|\; \sigma(0) = p, \sigma(1) \in \gamma, \exists T_\sigma, \text{ such that} \langle \dot{\sigma}, \dot{\sigma} \rangle \equiv -T_\sigma^2, \frac{\langle \dot{\sigma}, Y(\sigma) \rangle}{T_\sigma} \equiv -k. \right\}. \tag{19}$$

Clearly, the map $R: \mathscr{B}^+_{p,\gamma}(k) \mapsto \tilde{\mathscr{B}}^+_{p,\gamma}(k)$ defined by $R(z) = \sigma$, where $\sigma(s) = z(T_z \cdot s)$, gives a bijection between $\mathscr{B}^+_{p,\gamma}(k)$ and $\tilde{\mathscr{B}}^+_{p,\gamma}(k)$.

The map R is used to identify the spaces $\mathscr{B}^+_{p,\gamma}(k)$ and $\tilde{\mathscr{B}}^+_{p,\gamma}(k)$; a curve $\sigma \in \tilde{\mathscr{B}}^+_{p,\gamma}(k)$ will be called a brachistochrone of energy k if $R^{-1}(\sigma)$ is a brachistochrone of energy k.

We introduce the following map $\mathscr{F}: \tilde{\mathscr{B}}^+_{p,\gamma}(k) \mapsto \Omega_{p,\gamma}(\Delta)$:

$$\mathscr{F}(\sigma)(s) = w_\sigma(s) = \psi(\sigma(s), t_\sigma(s)), \tag{20}$$

where

$$t_\sigma(s) = -\int_0^s \frac{\langle \dot{\sigma}, Y(\sigma) \rangle}{\langle Y(\sigma), Y(\sigma) \rangle} dr = k \cdot T_\sigma \int_0^s \frac{dr}{\langle Y(\sigma), Y(\sigma) \rangle}. \tag{21}$$

The curve w_σ is of class C^1, moreover, it is easy to see that w_σ is in $\Omega_{p,\gamma}(\Delta)$.

Observe that, if for some positive k the curve $\sigma \in \tilde{\mathscr{B}}^+_{p,\gamma}(k)$ has image in the open set U_k of (2), then, since U_k is Y-invariant, also $\mathscr{F}(\sigma)$ has image in U_k.

Our variational principle states that the map $\mathscr{F}$ is a bijection between the set of brachistochrones of fixed energy k joining p with γ and the set of normal $(\phi_k \cdot g_R)$-geodesics.

We can state the variational principle for brachistochrones proven in Ref. 4:

Proposition II.2: Let k be a fixed positive number. If σ is a brachistochrone of energy k between p and γ, then $w_\sigma = \mathscr{F}(\sigma)$ is a normal geodesic joining p and γ in U_k with respect to the metric $\phi_k \cdot g_{\mathrm{sr}}$.

Conversely, if w is a normal geodesic in U_k between p and γ with respect to $\phi_k \cdot g_R$, then there exists a unique brachistochrone σ of energy k between p and γ such that $w = \mathscr{F}(\sigma)$. Moreover, in either case, w is minimal if and only of σ is minimal. □

It is also proven in Ref. 4 that, if $\sigma \in \tilde{\mathscr{B}}^+_{p,\gamma}$ is a brachistochrone of energy k and $w_\sigma = \mathscr{F}(\sigma)$ is the corresponding $(\phi_k \cdot g_R)$-geodesic in $\Omega_{p,\gamma}$, then the quantity $\phi_k(w_\sigma)\langle \dot{w}_\sigma, \dot{w}_\sigma \rangle$ is constant, and

$$2\, E_k(w_\sigma) = \phi_k(w_\sigma)\langle \dot{w}_\sigma, \dot{w}_\sigma \rangle = T_\sigma^2 \equiv -\langle \dot{\sigma}, \dot{\sigma} \rangle. \tag{22}$$

Hence, the proof of Theorem I.2 and Theorem I.3 is reduced to the proof of existence and multiplicity of critical points for the functional E_k in $\Omega_{p,\gamma}$.

III. THE PENALIZED FUNCTIONAL $E_{k,\varepsilon}$ AND THE PALAIS–SMALE CONDITION

The presence of the boundary ∂U_k is the main obstruction for the functional E_k to satisfy good compactness properties on $\Omega_{p,\gamma}$. Moreover, as we have already observed, the manifold $\Omega_{p,\gamma}$ is not complete.

To overcome this problem, we introduce a *penalization* argument, which has been used systematically in the study of geodesics in convex subsets of (semi) Riemannian manifolds.

Recalling the definition (9) of the function ψ_k, for all $\varepsilon \in [0,1]$ we introduce the *penalized* functional $E_{k,\varepsilon}:\Omega_{p,\gamma} \mapsto \mathbb{R}^+$ given by

$$E_{k,\varepsilon}(w) = E_k(w) + \varepsilon \int_0^1 \frac{ds}{\psi_k(w)^2}. \tag{23}$$

Observe that $E_{k,0} = E_k$.

Since ψ_k is smooth, then $E_{k,\varepsilon}$ is a smooth functional on $\Omega_{p,\gamma}$, and its Gateaux derivative is easily computed as

$$E'_{k,\varepsilon}(w)[\zeta] = E'_k(w)[\zeta] - 2\varepsilon \int_0^1 \frac{\langle \nabla_{(R)}\psi_k(w), \zeta\rangle_{(R)}}{\psi_k(w)^3} ds. \tag{24}$$

For all $\varepsilon \in\,]0,1]$, the sublevels of the penalized functional are complete. In order to prove this, we use the following result, known as *Gordon's Lemma*, whose proof can be found, e.g., in Ref. 11 and it is based on the fact that $\psi'_k \neq 0$ on ∂U_k.

We denote by $\mathrm{dist}(\cdot,\cdot)$ the distance function on $\mathscr{M}$ induced by the Riemannian structure g_R.

Lemma III.1: Let $\{w_n\}_n$ be a bounded sequence in $\Omega_{p,\gamma}$ such that:

$$\sup_n \int_0^1 \frac{ds}{\psi_k(w_n)^2} < +\infty.$$

Then, there exists $\rho > 0$ such that $\mathrm{dist}(w_n(s), \partial U_k) \geqslant \rho$ *for all $s \in [0,1]$ and all $n \in \mathbb{N}$.* □

We have the following immediate Corollary:

Corollary III.2: For all $\varepsilon \in\,]0,1]$ and all $c > 0$, the sublevel

$$E^c_{k,\varepsilon} = \{w \in \Omega_{p,\gamma} : E_{k,\varepsilon}(w) \leqslant c\}$$

is a (possibly empty) complete metric subspace of $\Omega_{p,\gamma}$.

Proof: Let w_n be a Cauchy sequence in $\Omega_{p,\gamma}$. Then, w_n is convergent in the $H^{1,2}$-norm to a curve w with image in $\overline{U_k} = U_k \cap \partial U_k$. Since w_n is bounded in $\Omega_{p,\gamma}$ and $E_{k,\varepsilon}(w_n)$ is bounded, from Lemma III.1 it follows that w_n, and hence w, is uniformly far from the boundary ∂U_k. Then, since γ is a maximal integral line of Y, then $w(1) \in \gamma$, so $w \in \Omega_{p,\gamma}$ and w_n converges to w in $\Omega_{p,\gamma}$; moreover, by the continuity of $E_{k,\varepsilon}$, it is $E_{k,\varepsilon}(w) \leqslant c$ and $E^c_{k,\varepsilon}$ is complete. □

We recall that if (X,h) is an Hilbert manifold and $F:X \mapsto \mathbb{R}$ is a C^1-functional on X, then F is said to satisfy the *Palais–Smale* condition at level $c \in \mathbb{R}$ if every sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ satisfying

$$(\mathrm{PS1})_c \quad \lim_{n\to\infty} F(x_n) = c,$$

$$(\mathrm{PS2})_c \quad \lim_{n\to\infty} \|F'(x_n)\| = 0,$$

has a subsequence converging in X. The norm considered in (PS2) is the operator norm in the Hilbert space $T_{x_n}X$.

A sequence x_n in X that satisfies $(\mathrm{PS1})_c$ and $(\mathrm{PS2})_c$ will be called a *Palais–Smale* sequence at level c for the functional F.

Proposition III.3: For all $\varepsilon \in\,]0,1]$, the penalized functional $E_{k,\varepsilon}$ satisfies the Palais–Smale condition at every level $c \in \mathbb{R}$.

Proof: Let $\varepsilon \in]0,1]$ and $c \in \mathbb{R}$ be fixed; let $\{w_n\}_n$ be a Palais–Smale sequence at the level c for $E_{k,\varepsilon}$.

Since $\{w_n\}_n$ is bounded in $\Omega_{p,\gamma}$, from Lemma III.1 it follows that w_n stays uniformly far from the boundary ∂U_k. Moreover, the boundedness in the Hilbert manifold $H^{1,2}([0,1],U_k)$ implies the existence of a subsequence, still denoted by w_n, which is weakly convergent to a curve $w \in H^{1,2}([0,1],U_k)$. In particular, w_n is uniformly convergent to w, and since γ is a maximal integral line of Y, then $w \in \Omega_{p,\gamma}$.

Hence, since w_n and w have image which is uniformly far from ∂U_k, then we can assume that w_n and w have image in a compact subset K of U_k.

Since ϕ_k is uniformly far from zero in U_k, the set K is complete with respect to the Riemannian metric $\phi_k \cdot g_R$.

We now use the fact that $E_{k,\varepsilon}(w_n)$ goes to 0 to prove that the convergence of w_n to w is strong. We argue as follows.

We assume, by the Nash embedding theorem (see Ref. 12), that $\mathscr{M}$ is embedded in some Euclidean space $\mathbb{R}^N$ isometrically with respect to g_R. For all $q \in \mathscr{M}$, we denote by P_q the orthogonal projection of $\mathbb{R}^N$ onto the tangent space $T_q\mathscr{M}$.

If we consider the vector field ζ_n along w_n given by:

$$\zeta_n(s) = P_{w_n(s)}(w_n(s) - w(s) - c_n(s)),$$

where $c_n(s) = s(w_n(1) - w(1))$, then clearly ζ_n is bounded in $T_{w_n}\Omega_{p,\gamma}$ and ζ_n tends weakly to 0 in the Sobolev space $H^{1,2}([0,1],\mathbb{R}^N)$.

Since w_n is uniformly convergent to w, then the vector field:

$$\frac{\nabla_{(R)}\phi_k(w_n)}{\phi_k(w_n)^3}$$

is uniformly convergent. It follows that:

$$\lim_{n\to\infty} \int_0^1 \frac{\langle \nabla_{(R)}\phi_k(w_n), \zeta_n\rangle}{\phi_k(w_n)^3} ds = 0. \tag{25}$$

From (18), (24) and (25) it follows that:

$$\lim_{n\to\infty} E_k'(w_n)[\zeta_n] = \lim_{n\to\infty} \left(\frac{1}{2}\int_0^1 \langle \nabla_{(R)}\phi_k(w_n), \zeta_n\rangle_{(R)} \langle \dot{w}_n, \dot{w}_n\rangle_{(R)} ds + \int_0^1 \phi_k(w_n) \langle \dot{w}_n, \nabla^{(R)}_{\dot{w}_n}\zeta_n\rangle_{(R)} ds \right) = 0. \tag{26}$$

If we consider the first integral on the right-hand side of equality (26), by the weak convergence in $H^{1,2}$ of w_n, we have that the vector field $\nabla_{(R)}\phi_k(w_n)$ is bounded in $L^\infty([0,1],T\mathscr{M})$, ζ_n tends to zero in $L^\infty([0,1],T\mathscr{M})$, and $\langle \dot{w}_n, \dot{w}_n\rangle_{(R)}$ is bounded in $L^1([0,1],\mathbb{R})$. It follows:

$$\lim_{n\to\infty} \int_0^1 \langle \nabla_{(R)}\phi_k(w_n), \zeta_n\rangle_{(R)} \langle \dot{w}_n, \dot{w}_n\rangle_{(R)} ds = 0. \tag{27}$$

Now, we can write:

$$\int_0^1 \phi_k(w_n) \langle \dot{w}_n, \nabla^{(R)}_{\dot{w}_n}\zeta_n\rangle_{(R)} ds = \int_0^1 \phi_k(w_n) \langle \dot{w}_n, (dP_{w_n}[\dot{w}_n])(w_n - w - c_n)\rangle_{(R)} ds + \int_0^1 \phi_k(w_n) \langle \dot{w}_n, P_{w_n}(\dot{w}_n - \bar{w} - \dot{c}_n)\rangle_{(R)} ds, \tag{28}$$

where $dP_{w_n}[\dot{w}_n]$ denotes the linear operator given by the derivative of P_{w_n} in the direction of $\dot{w}_n$.

Once again, the weak convergence of w_n in $\Omega_{p,\gamma}$ implies that the first integral on the right-hand side of (28) is infinitesimal:

$$\lim_{n\to\infty}\int_0^1 \phi_k(w_n)\langle \dot{w}_n, (dP_{w_n}[\dot{w}_n])(w_n - w - c_n)\rangle_{(R)} ds = 0,$$

hence, considering that $\dot{c}_n$ converges to zero uniformly, we have:

$$\lim_{n\to\infty}\int_0^1 \phi_k(w_n)\langle \dot{w}_n, P_{w_n}(\dot{w}_n - \dot{w} - \dot{c}_n)\rangle_{(R)} ds = \lim_{n\to\infty}\int_0^1 \phi_k(w_n)\langle \dot{w}_n, \dot{w}_n - \dot{w}\rangle_{(R)} ds = 0. \quad (29)$$

In (29) we have used the fact that P_{w_n} is symmetric and that $P_{w_n}(\dot{w}_n) = \dot{w}_n$. We also have

$$\lim_{n\to\infty}\int_0^1 \phi_k(w_n)\langle \dot{w}, \dot{w}_n - \dot{w}\rangle_{(R)} ds = 0. \quad (30)$$

To see this, we consider that

$$\lim_{n\to\infty}\left(\int_0^1 \phi_k(w_n)\langle \dot{w}_n - \dot{w}, \dot{w}\rangle_{(R)} ds - \int_0^1 \phi_k(w)\langle \dot{w}_n - \dot{w}, \dot{w}\rangle_{(R)} ds \right) = 0, \quad (31)$$

because $(\phi_k(w_n) - \phi_k(w))$ tends to zero uniformly and $\langle \dot{w}_n - \dot{w}, \dot{w}\rangle_{(R)}$ is bounded in $L^1([0,1], T\mathscr{M})$. Also, by the weak L^2-convergence of $\dot{w}_n$ to $\dot{w}$, it follows immediately that

$$\lim_{n\to\infty}\int_0^1 \phi_k(w)\langle \dot{w}_n - \dot{w}, \dot{w}\rangle_{(R)} ds = 0. \quad (32)$$

Equations (31) and (32) imply immediately (30).

Subtracting (30) from (29), we obtain

$$\lim_{n\to\infty}\int_0^1 \phi_k(w_n)\langle \dot{w}_n - \dot{w}, \dot{w}_n - \dot{w}\rangle_{(R)} ds = 0; \quad (33)$$

finally, since $\inf \phi_k(w_n) > 0$, we get

$$\lim_{n\to\infty}\int_0^1 \langle \dot{w}_n - \dot{w}, \dot{w}_n - \dot{w}\rangle_{(R)} ds = 0, \quad (34)$$

which gives the strong convergence of w_n to w in $\Omega_{p,\gamma}$, concluding the proof. □

Standard techniques in critical point theory are then used to prove the existence of minima for the penalized functionals $E_{k,\varepsilon}$:

Corollary III..4: For all $\varepsilon \in]0,1]$ *the penalized functional* $E_{k,\varepsilon}$ *attains its minimum in* $\Omega_{p,\gamma}$.

Proof: It is a classical argument in critical point theory. Let us fix $\varepsilon \in]0,1]$. Thanks to the Palais–Smale condition and the completeness of the sublevels of the functionals $E_{k,\varepsilon}$, if the infimum i_ε of $E_{k,\varepsilon}$ on $\Omega_{p,\gamma}$ were not a critical value, then it would be possible to find a homotopy between the sublevels $E_{k,\varepsilon}^{i-\eta}$ and $E_{k,\varepsilon}^{i+\eta}$, where $\eta > 0$ is sufficiently small. This is clearly impossible, because, for every $\eta > 0$, $E_{k,\varepsilon}^{i-\eta} = \varnothing$ while $E_{k,\varepsilon}^{i+\eta} \neq \varnothing$. Hence, $E_{k,\varepsilon}$ attains its minimum on $\Omega_{p,\gamma}$. □

IV. *A PRIORI* ESTIMATES FOR THE PENALIZED FUNCTIONAL

The goal of this section is to prove that, given a family w_ε of stationary points for the penalized functionals $E_{k,\varepsilon}$, with $\varepsilon \in]0,1]$, there exists a sequence $\varepsilon_n \downarrow 0$ such that the corresponding sequence w_{ε_n} tends to a curve w which is a critical point for E_k in $\Omega_{p,\gamma}$.

From Corollary III.4, throughout this section we will assume that $\{w_\varepsilon\}$ is a family of critical points for $E_{k,\varepsilon}$ in $\Omega_{p,\gamma}$, with $\varepsilon \in]0,1]$, and that D is a fixed positive constant such that

$$E_{k,\varepsilon}(w_\varepsilon) \leqslant D < +\infty, \forall \varepsilon]0,1]. \tag{35}$$

Since $\varepsilon \mapsto E_{k,\varepsilon}(w)$ is an increasing function for all $w \in \Omega_{p,\gamma}$, then the inequality (35) is satisfied, for instance, if we take w_ε to be a minimum point for $E_{k,\varepsilon}$ and $D = E_{k,1}(w)$, for any fixed $w \in \Omega_{p,\gamma}$.

Observe that, since $\langle Y, Y \rangle_{(R)}$ is bounded away from zero, then (35) implies that the integrals:

$$\int_0^1 \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} ds$$

are bounded. By the completeness of $\overline{U_k}$, it follows that there exists a compact subset $\mathscr{K}$ of $\overline{U_k}$ such that

$$w_\varepsilon(s) \in \mathscr{K}, \quad \forall \varepsilon \in]0,1], \quad s \in [0,1].$$

In the course of this section, it will be useful to have the following formulas for the derivatives of the functions ϕ_k and ψ_k. They are easily computed using the fact that Y is a Killing vector field with respect to g_R:

$$\nabla_{(R)} \psi_k = 2 \nabla^{(R)}_Y Y, \tag{36}$$

$$\nabla_{(R)} \phi_k = -\frac{2k^2}{(k^2 - \langle Y, Y \rangle_{(R)})^2} \nabla^{(R)}_Y Y = -\frac{k^2}{\psi_k^2} \nabla_{(R)} \psi_k . \tag{37}$$

We start by exhibiting the differential equation satisfied by the w_ε's.

Proposition IV.1: For all $\varepsilon \in]0,1]$, the curve w_ε is smooth, and it satisfies the differential equation:

$$\nabla^{(R)}_{\dot{w}_\varepsilon} \dot{w}_\varepsilon + \left\langle \frac{\nabla_{(R)} \phi_k(w_\varepsilon)}{\phi_k(w_\varepsilon)}, \dot{w}_\varepsilon \right\rangle_{(R)} \dot{w}_\varepsilon + 2\varepsilon \frac{\nabla_{(R)} \psi_k(w_\varepsilon)}{\phi_k(w_\varepsilon) \cdot \psi_k(w_\varepsilon)^3} + -\frac{1}{2} \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} \frac{\nabla_{(R)} \phi_k(w_\varepsilon)}{\phi_k(w_\varepsilon)} = 0. \tag{38}$$

Proof: The proof of the smoothness of w_ε is carried out by standard *boot-strap* techniques, and it is omitted.

The differential equation (38) is obtained from the Euler–Lagrange equation for the functional $E_{k,\varepsilon}$. Namely, from (18) and (24), the curve w_ε satisfies:

$$E'_{k,\varepsilon}(w_\varepsilon)[\zeta] = \int_0^1 \left(-2\varepsilon \frac{\langle \nabla_{(R)} \psi_k(w_\varepsilon), \zeta \rangle_{(R)}}{\psi_k(w_\varepsilon)^3} + \phi_k(w_\varepsilon) \langle \nabla^{(R)}_{\dot{w}_\varepsilon} \zeta, \dot{w}_\varepsilon \rangle_{(R)} \right) ds$$
$$+ \frac{1}{2} \int_0^1 \langle \nabla_{(R)} \phi_k(w_\varepsilon), \zeta \rangle_{(R)} \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} ds = 0, \tag{39}$$

for all $\zeta \in T_{w_\varepsilon} \Omega_{p,\gamma}$. If we integrate by parts the term in (39) that contain the covariant derivative $\nabla^{(R)}_{\dot{w}_\varepsilon} \zeta$, considering that $\zeta(0) = \zeta(1) = 0$, we obtain

$$0 = \int_0^1 \langle \zeta, -\phi_k(w_\varepsilon) \nabla^{(R)}_{\dot{w}_\varepsilon} \dot{w}_\varepsilon - \langle \nabla_{(R)} \phi_k(w_\varepsilon), \dot{w}_\varepsilon \rangle_{(R)} \dot{w}_\varepsilon \rangle_{(R)} ds$$
$$+ \int_0^1 \left\langle \zeta, \frac{1}{2} \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} - 2\varepsilon \frac{\nabla_{(R)} \psi_k(w_\varepsilon)}{\psi_k(w_\varepsilon)^3} \right\rangle_{(R)} ds, \tag{40}$$

for all $\zeta \in T_{w_\varepsilon}\Omega_{p,\gamma}$. From the fundamental Lemma of calculus of variations, we have the Euler–Lagrange equations for the functional $E_{k,\varepsilon}$:

$$0 = -\phi_k(w_\varepsilon)\nabla^{(R)}_{\dot{w}_\varepsilon}\dot{w}_\varepsilon - \langle \nabla_{(R)}\phi_k(w_\varepsilon), \dot{w}_\varepsilon\rangle_{(R)}\dot{w}_\varepsilon + \frac{1}{2}\langle \dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)}\nabla_{(R)}\phi_k(w_\varepsilon) - 2\varepsilon\frac{\nabla_{(R)}\psi_k(w_\varepsilon)}{\psi_k(w_\varepsilon)^3}. \tag{41}$$

Equation (38) is obtained by dividing (41) by $-\phi_k(w_\varepsilon)$. □

Remark IV.2: Observe that, using the same argument of the first part of the proof of Proposition IV.1, it is easily seen that the critical points for the functional E_k are smooth curves that satisfy the differential equation obtained from (38) by setting $\varepsilon = 0$.

We now prove that the images of the curves w_ε stay uniformly far from the boundary ∂U_k. The distance from the boundary ∂U_k is measured by the function ψ_k.

Lemma IV.3: There exists $\delta = \delta(D) > 0$ such that $\psi_k(w_\varepsilon(s)) \geqslant \delta$ for all $\varepsilon \in]0,1]$ and all $s \in [0,1]$.

Proof: For all $\varepsilon \in]0,1]$, we consider the smooth function $\rho_\varepsilon(s) = \psi_k(w_\varepsilon(s))$; let $s_\varepsilon \in [0,1]$ be a minimum point for ρ. By contradiction, we assume that

$$\liminf_{\varepsilon\to 0} \rho_\varepsilon(s_\varepsilon) = 0, \tag{42}$$

and thus we can assume that $s_\varepsilon \in]0,1[$. Therefore, we have

$$\rho'_\varepsilon(s_\varepsilon) = \langle \nabla_{(R)}\psi_k(w_\varepsilon(s_\varepsilon)), \dot{w}_\varepsilon(s_\varepsilon)\rangle_{(R)} = 0, \tag{43}$$

and

$$\rho''_\varepsilon(s_\varepsilon) = \langle H^{\psi_k}_{(R)}(w_\varepsilon(s_\varepsilon))\,\dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)} + \langle \nabla_{(R)}\psi_k(w_\varepsilon(s_\varepsilon)), \nabla^{(R)}_{\dot{w}_\varepsilon}\dot{w}_\varepsilon(s_\varepsilon)\rangle_{(R)} \geqslant 0, \tag{44}$$

where $H^{\psi_k}_{(R)}$ denote the Hessian operator of ψ_k with respect to the Riemannian metric g_R.

For the sake of shortness, in the computations that follow we will omit the argument s_ε.

Since $H^{\psi_k}_{(R)}$ is continuous in $\mathscr{M}$ and w_ε has image in a compact subset of $\overline{U}_k$, then there exists a positive constant ν_0 such that

$$|\langle H^{\psi_k}_{(R)}(w_\varepsilon)\dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)}| \leqslant \nu_0 \cdot \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)}, \quad \forall \varepsilon \in]0,1]. \tag{45}$$

Using the differential equation (38), the formula (36) and the fact that $\langle \nabla_{(R)}\psi_k(w_\varepsilon), \dot{w}_\varepsilon\rangle_{(R)} = 0$, from (44) and (45) we obtain

$$\begin{aligned}
\rho''_\varepsilon(s_\varepsilon) &= \frac{1}{2}\frac{\langle \dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)}}{\phi_k(w_\varepsilon)} \cdot \langle \nabla_{(R)}\phi_k(w_\varepsilon), \nabla_{(R)}\psi_k(w_\varepsilon)\rangle_{(R)} \\
&\quad - \frac{2\varepsilon}{\phi_k(w_\varepsilon)\cdot\psi_k(w_\varepsilon)^3}\langle \nabla_{(R)}\psi_k(w_\varepsilon), \nabla_{(R)}\psi_k(w_\varepsilon)\rangle_{(R)} + \langle H^{\psi_k}_{(R)}(w_\varepsilon)\dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)} \\
&\leqslant \left(\nu_0 - \frac{k^2\langle \nabla_{(R)}\psi_k(w_\varepsilon), \nabla_{(R)}\psi_k(w_\varepsilon)\rangle_{(R)}}{2\phi_k(w_\varepsilon)\cdot\psi_k(w_\varepsilon)^2} \right) \cdot \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon\rangle_{(R)} \\
&\quad - \frac{2\varepsilon}{\phi_k(w_\varepsilon)\cdot\psi_k(w_\varepsilon)^3}\langle \nabla_{(R)}\psi_k(w_\varepsilon), \nabla_{(R)}\psi_k(w_\varepsilon)\rangle_{(R)}.
\end{aligned} \tag{46}$$

Since $\nabla_{(R)}\psi_k$ does not vanish near ∂U_k and the image of the w_ε's lie in a compact subset of $\overline{U}_k$, then

$$\langle \nabla_{(R)}\psi_k(w_\varepsilon), \nabla_{(R)}\psi_k(w_\varepsilon)\rangle_{(R)} \geqslant \mu_0 > 0$$

for some suitable constant μ_0; moreover, the function $(\phi_k \cdot \psi_k^2)$ is infinitesimal near the boundary ∂U_k.

Then, if (42) holds, then it would be possible to find $\varepsilon > 0$ small enough so that:

$$\left(\nu_0 - \frac{k^2 \langle \nabla_{(R)} \psi_k(w_\varepsilon), \nabla_{(R)} \psi_k(w_\varepsilon) \rangle_{(R)}}{2\, \phi_k(w_\varepsilon) \cdot \psi_k(w_\varepsilon)^2} \right) \cdot \langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} \leqslant 0. \tag{47}$$

Since

$$-\frac{2\,\varepsilon}{\phi_k(w_\varepsilon) \cdot \psi_k(w_\varepsilon)^3} \langle \nabla_{(R)} \psi_k(w_\varepsilon), \nabla_{(R)} \psi_k(w_\varepsilon) \rangle_{(R)} < 0, \tag{48}$$

from (46), (47), and (48) it would follow:

$$\rho_\varepsilon''(s_\varepsilon) < 0, \tag{49}$$

which contradicts (44) and concludes the proof. □

From Lemma IV.3, it follows that the function $\psi_k(w_\varepsilon)$ is uniformly bounded away from zero.

We are now ready to prove the main result of the section:

Proposition IV.4: There exists a sequence $\{\varepsilon_n\}$ *tending to zero as* $n \to \infty$ *and a smooth curve* $w \in \Omega_{p,\gamma}$ *such that* w_{ε_n} *tends to* w in $C^2([0,1], \mathscr{M})$. *Moreover, such a curve* w *is a critical point for the functional* E_k in $\Omega_{p,\gamma}$.

Proof: From (35), we have that the family $\{w_\varepsilon\}$ is bounded in $\Omega_{p,\gamma}$ and, from Lemma IV.3, the images of the curves w_ε stay uniformly far from the boundary ∂U_k.

We consider Eq. (38) satisfied by the w_ε's. Since $\psi_k(w_\varepsilon)$ is bounded away from zero, and w_ε is uniformly bounded, then the term:

$$\frac{\nabla_{(R)} \psi_k(w_\varepsilon)}{\phi_k(w_\varepsilon) \psi_k(w_\varepsilon)^3} \tag{50}$$

is bounded in $L^\infty([0,1], T\mathscr{M})$.

Moreover, since $\dot{w}_\varepsilon$ is bounded in $L^2([0,1], T\mathscr{M})$, then the term

$$\left\langle \frac{\nabla_{(R)} \phi_k(w_\varepsilon)}{\phi_k(w_\varepsilon)}, \dot{w}_\varepsilon \right\rangle_{(R)} \dot{w}_\varepsilon \tag{51}$$

is bounded in $L^1([0,1], T\mathscr{M})$.

Also the term:

$$\langle \dot{w}_\varepsilon, \dot{w}_\varepsilon \rangle_{(R)} \frac{\nabla_{(R)} \phi_k(w_\varepsilon)}{\phi_k(w_\varepsilon)} \tag{52}$$

is bounded in $L^1([0,1], T\mathscr{M})$.

Then, from (38), we have that $\nabla^{(R)}_{\dot{w}_\varepsilon} \dot{w}_\varepsilon$ is bounded in $L^1([0,1], T\mathscr{M})$. This implies that $\dot{w}_\varepsilon$ is bounded in $L^\infty([0,1], T\mathscr{M})$.

Repeating the same argument, we have now that the terms (50)–(52) are bounded in $L^\infty([0,1], T\mathscr{M})$, and this implies that $\nabla^{(R)}_{\dot{w}_\varepsilon} \dot{w}_\varepsilon$ is bounded in $L^\infty([0,1], T\mathscr{M})$, so w_ε is bounded in $H^{1,\infty}([0,1], \mathscr{M})$.

Proceeding similarly, we obtain that w_ε is bounded in $H^{2,2}([0,1], \mathscr{M})$. Then, up to subsequences, w_n is convergent in $C^1([0,1], \mathscr{M})$. Hence, by (41), the convergence is in $C^2([0,1], \mathscr{M})$.

Since the convergence of w_{ε_n} to w is C^2, we can pass to the limit as $\varepsilon \to 0$ in the differential equation (38), obtaining that w satisfies

$$\nabla^{(R)}_{\dot{w}}\dot{w}+\left\langle \frac{\nabla_{(R)}\phi_k(w)}{\phi_k(w)},\dot{w}\right\rangle_{(R)}\dot{w}-\frac{1}{2}\langle \dot{w},\dot{w}\rangle_{(R)}\nabla_{(R)}\phi_k(w)=0. \tag{53}$$

Such equation is the Euler–Lagrange equation for the functional E_k on $\Omega_{p,\gamma}$ (see Remark IV.2), thus w is a critical point for E_k and we are done. □

Observe that, as a consequence of Proposition IV.4, we have proved Theorem I.2.

V. LJUSTERNIK–SCHNIRELMAN THEORY AND MULTIPLICITY OF BRACHISTOCHRONES

The goal of this section is to give a proof of Theorem I.3 by means of the Ljusternik–Schnirelman theory for Palais–Smale functionals. Observe that, by Proposition II.2 and formula (22), if $\sigma\in\tilde{\mathscr{B}}^+_{p,\gamma}$ is a brachistochrone of energy k and w is the corresponding $\phi_k\cdot g_R$ normal geodesic in $\Omega_{p,\gamma}$, then the travel time T_σ of σ and the value of the functional $E_k(w)$ is

$$T_\sigma=\sqrt{2\,E_k(w)}. \tag{54}$$

We recall the following definition:

Definition V.1: If X is a topological space and B any subset of X, the Ljusternik–Schnirelman category $\mathrm{cat}_X(B)$ of B in X is the minimal number (possibly infinite) of closed, contractible subsets of X that cover B. We denote by $\mathrm{cat}(X)=\mathrm{cat}_X(X)$.

Clearly, the Ljusternik–Schnirelman category is increasing with respect to the inclusion, i.e., if A and B are subsets of X, then

$$A\subseteq B\ \Rightarrow\ \mathrm{cat}_X(A)\leqslant\mathrm{cat}_X(B). \tag{55}$$

Moreover, the Ljusternik–Schnirelman category is a homotopical invariant, i.e., if A and B are homotopical subsets of X, then $\mathrm{cat}_X(A)=\mathrm{cat}_X(B)$.

It is easy to see that, since Y is complete and γ is contractible in $\mathscr{M}$, then, for any fixed point $q\in\gamma$, the spaces $\Omega_{p,q}$ and $\Omega_{p,\gamma}$ have the same homotopy type. Namely, given any curve $z\in\Omega_{p,\gamma}$, one can consider the curve $\tilde{z}\in\Omega_{p,q}$ given by

$$\tilde{z}(s)=\psi(z(s),\phi_z(s)),$$

where

$$\phi_z(s)=(\gamma^{-1}(q)-\gamma^{-1}(z(1)))\cdot s.$$

Observe that the map γ^{-1} is well defined and continuous on $\gamma(\mathbb{R})$ because γ is injective, due to the causality of $\mathscr{M}$. Moreover, the evaluation map $z\mapsto z(1)$ is continuous in $\Omega_{p,\gamma}$, and so the map $z\mapsto\phi_z\in C^1([0,1],\mathbb{R})$ is continuous in $\Omega_{p,\gamma}$.

The map $\mathscr{H}:\Omega_{p,\gamma}\times[0,1]\mapsto\Omega_{p,\gamma}$ given by

$$\mathscr{H}(z,\sigma)(s)=\psi(z(s),\phi_z(s)\cdot\sigma)$$

is clearly continuous; moreover it satisfies $\mathscr{H}(z,0)=z$, $\mathscr{H}(z,\sigma)=z$ for all $z\in\Omega_{p,q}$ and all $\sigma\in[0,1]$, and $\mathscr{H}(z,1)=\tilde{z}$, and so it is a *strong deformation retract* of $\Omega_{p,\gamma}$ onto $\Omega_{p,q}$. In particular, the map $\mathscr{H}(\cdot,1)$ is a homotopy equivalence between $\Omega_{p,\gamma}$ and $\Omega_{p,q}$ and the two spaces have the same homotopy type.

A well known result by Fadell and Husseini (see Ref. 13) states that, if $\overline{U_k}$ is not contractible, then the category of the space $\Omega_{p,q}$, and so also the category of $\Omega_{p,\gamma}$, is infinite:

$$\mathrm{cat}(\Omega_{p,\gamma})=\mathrm{cat}(\Omega_{p,q})=+\infty;$$

moreover, there exists a sequence K_n of compact subsets of $\Omega_{p,\gamma}$ such that

$$\lim_{n\to\infty}\mathrm{cat}_{\Omega_{p,\gamma}}(K_n)=+\infty. \tag{56}$$

For all $n \in \mathbb{N}$, we denote by Γ_n the collection of all compact subsets of $\Omega_{p,\gamma}$ having category strictly larger than to n:

$$\Gamma_n = \{B \text{ compact subset of } \Omega_{p,\gamma} : \mathrm{cat}_{\Omega_{p,\gamma}}(B) \geqslant n+1\}. \tag{57}$$

Observe that, by (56), $\Gamma_n \neq \varnothing$ for all $n \in \mathbb{N}$.

For all $c \in \mathbb{R}$ and $\varepsilon \in [0,1]$, we denote by $E^c_{k,\varepsilon}$ the closed c-sublevel of the functional $E_{k,\varepsilon}$ in $\Omega_{p,\gamma}$:

$$E^c_{k,\varepsilon} = \{w \in \Omega_{p,\gamma} : E_{k,\varepsilon}(w) \leqslant c\};$$

clearly

$$E^c_{k,\varepsilon} \subseteq E^c_k, \quad \forall c \in \mathbb{R}, \quad \forall \varepsilon \in [0,1].$$

A well known argument in critical point theory shows that, if $F: X \mapsto \mathbb{R}$ is a C^1-functional defined on the Banach manifold X, satisfying the Palais–Smale condition at the level $c' \in \mathbb{R}$ and such that the sublevel $F^{c'}$ is complete, then, the category $\mathrm{cat}_X(F^{c'})$ is finite.

In the following preliminary Lemma we show that, for all $c \in \mathbb{R}$, the category $\mathrm{cat}_{\Omega_{p,\gamma}}(E^c_{k,\varepsilon})$ is bounded *uniformly* with respect to ε.

Lemma V.2: For all $c \in \mathbb{R}$ there exists $N = N(c) \in \mathbb{N}$ such that

$$\mathrm{cat}_{\Omega_{p,\gamma}}(E^c_{k,\varepsilon}) \leqslant N(c), \tag{58}$$

for all $\varepsilon \in [0,1]$.

Proof: We will show that, for all $c \in \mathbb{R}$, there exists $c' \in \mathbb{R}$ and, for all $\varepsilon \in [0,1]$, a homotopy between the sublevel $E^c_{k,\varepsilon}$ and a subset of the sublevel $F^{c'}$, where F is a suitable smooth functional satisfying the properties described above.

By the completeness of $\overline{U_k}$, for all $c \in \mathbb{R}$ there exists a compact subset $K_c \subset \overline{U_k}$ containing the images of all the curves in $E^c_{k,\varepsilon}$, for all $\varepsilon \in [0,1]$. Indeed, as we have already observed at the beginning of Sec. IV, all such curves w have bounded energy:

$$\int_0^1 \langle \dot{w}, \dot{w} \rangle_{(R)} ds \leqslant d_0 < +\infty.$$

For $\delta > 0$, we denote by V_δ the closed subset of U_k given by

$$V_\delta = \{x \in U_k : \psi_k(x) \geqslant \delta\}. \tag{59}$$

Let now $\delta > 0$ be small enough and $H: [0,1] \times (\overline{U_k} \cap K_c) \mapsto K_c$ be a map of class C^1 satisfying:

(1) $H(0,x) = x$ for all $x \in \overline{U_k} \cap K_c$;
(2) $H(\sigma, V_\delta) \subset V_\delta$ for all $\sigma \in [0,1]$;
(3) $H(1, \overline{U_k} \cap K_c) \subset V_\delta$.

Such a map is built using the flow of the vector field $\nabla_{(R)} \psi_k$, which is bounded away from 0 in the compact set $K_c \cap \partial U_k$.

Finally, we consider the map $\mathscr{H}: [0,1] \times E^c_k \mapsto \Omega_{p,\gamma}$ defined by

$$\mathscr{H}(\sigma, w)(s) = H(\sigma \cdot s, w(s)), \quad \forall \sigma, s \in [0,1]. \tag{60}$$

Clearly, $\mathscr{H}$ is continuous; moreover, by the construction of H, the map $\mathscr{H}(0,\cdot)$ is the identity on E^c_k. By construction, the curves in $\mathscr{H}(1, E^c_k)$ have image in the set V_δ, and so they stay uniformly far from ∂U_k.

Since the curves in E^c_k have image in a fixed compact set, it is easy to see by a direct computation that the map H has bounded partial derivatives, and so the map $\mathscr{H}(1,\cdot)$ is Lipschitz continuous on E^c_k. It follows that there exists a real number c' such that

$$\sup_{w \in \mathscr{H}(1,E_k^c)} E_k(w) \leqslant c'.$$

We denote by $A_{k,\delta,c'}$ the set of curves in $E_k^{c'}$ having image in V_δ:

$$A_{k,\delta,c'} = \{w \in E_k^{c'} : w(s) \in V_\delta, \forall s\};$$

we have just proven that

$$\mathscr{H}(1,E_k^c) \subseteq A_{k,\delta,c'}.$$

Let us consider the functional $F:\Omega_{p,\gamma} \mapsto \mathbb{R}$ given by

$$F(w) = E_k(w) + \int_0^1 \frac{ds}{G(\psi_k(w))^2}, \tag{61}$$

where $G:[0,+\infty[\mapsto \mathbb{R}^+$ is a smooth function satisfying

$$G(0) = 0,\ G'(0) > 0, \quad \text{and } G(r) \equiv 0 \quad \text{if } x \in \left[\frac{\delta}{2}, +\infty\right[.$$

Now, the same proofs of Corollary III.2 and of Proposition III.3 can be repeated *verbatim* to show that F is a smooth functional on $\Omega_{p,\gamma}$ that satisfies the Palais–Smale condition, and whose sublevels are complete. It follows that, for all $c' \in \mathbb{R}$, there exists a natural number $N(c')$ such that:

$$\mathrm{cat}_{\Omega_{p,\gamma}}(F^{c'}) = N(c') < +\infty. \tag{62}$$

Since the function G vanishes identically on the images of the curves in $A_{k,\delta,c'}$, then $A_{k,\delta,c'} \subseteq F^{c'}$, and, in particular,

$$\mathrm{cat}_{\Omega_{p,\gamma}}(A_{k,\delta,c'}) \leqslant \mathrm{cat}_{\Omega_{p,\gamma}}(F^{c'}) = N(c'). \tag{63}$$

Thus, by the homotopical invariance and the monotonicity of the Ljusternik–Schnirelman category, it is:

$$\mathrm{cat}_{\Omega_{p,\gamma}}(E_{k,\varepsilon}^c) \leqslant \mathrm{cat}_{\Omega_{p,\gamma}}(E_k^c) = \mathrm{cat}_{\Omega_{p,\gamma}}(\mathscr{H}(1,E_k^c)) \leqslant \mathrm{cat}_{\Omega_{p,\gamma}}(A_{k,\delta,c'}), \tag{64}$$

which concludes the proof. □

A well known *minimax* argument in critical point theory shows that the numbers:

$$c_m^\varepsilon = \inf_{B \in \Gamma_m} \sup_{x \in B} E_{k,\varepsilon}(x) \tag{65}$$

are critical values for the functional $E_{k,\varepsilon}$ for all $m \in \mathbb{N}$ and all $\varepsilon \in]0,1]$. Observe that, by the definition (57) of Γ_n, each c_m^ε is well defined and finite. Also, since $\varepsilon \mapsto E_{k,\varepsilon}(w)$ is increasing on $[0,1]$ for all $w \in \Omega_{p,\gamma}$, it follows easily that

$$0 \leqslant c_m^\varepsilon \leqslant c_m^1, \quad \forall \varepsilon \in [0,1], \quad \forall m \in \mathbb{N}. \tag{66}$$

Since Γ_n is nonempty for all n, from (65) it follows that $E_{k,\varepsilon}$ has arbitrarily large critical values for all $\varepsilon \in]0,1]$; in order to prove Theorem I.3, we need to prove that the same claim is true for the functional $E_k = E_{k,0}$.

Lemma V.3: In the notations of Lemma V.2, for all $c \in \mathbb{R}$, if $m \geqslant N(c)$ then

$$c_m^\varepsilon \geqslant c, \quad \forall \varepsilon \in]0,1]. \tag{67}$$

Proof: Let $\varepsilon \in]0,1]$ and $c \in \mathbb{R}$ be fixed. By contradiction, suppose that there exists $m \geqslant N(c)$ and $B \in \Gamma_m$ such that:

$$\sup_{x \in B} E_{k,\varepsilon}(x) \leqslant c.$$

Then, it would be:

$$B \subseteq E^c_{k,\varepsilon} \subseteq E^c_k,$$

and, by the monotonicity of the Ljusternik–Schnirelman category, it would be

$$\mathrm{cat}_{\Omega_{p,\gamma}}(B) \leqslant \mathrm{cat}_{\Omega_{p,\gamma}}(E^c_k) \leqslant N(c). \tag{68}$$

The inequality (68) contradicts the fact that $B \in \Gamma_m$ and concludes the proof. □

We are ready for the proof of Theorem I.3:

Proof of Theorem I.3 Let $\{\varepsilon_n\}_n \subset]0,1]$ be any decreasing sequence converging to 0 and let E_{k,ε_n} be the corresponding sequence of functionals on $\Omega_{p,\gamma}$.

For all $m \in \mathbb{N}$, we define:

$$\overline{c}_m = \lim_{n \to \infty} \inf c^{\varepsilon_n}_m. \tag{69}$$

By (66), the $\overline{c}_m$ are well defined (finite), and they form a nondecreasing sequence.

By Proposition IV.4, for all $m \in \mathbb{N}$, $\overline{c}_m$ is a critical value for E_k.

Finally, by Lemma V.3, the sequence $\overline{c}_m$ is unbounded, hence E_k has arbitrarily large critical values. The conclusion follows immediately from Proposition II.2 and formula (54). □

ACKNOWLEDGMENTS

F. G. partially supported by MURST 40% and 60%. P. P. was partially supported by CNPq, Processo No. 301410/95-0.

[1] F. Goldstein and C. M. Bender, ''Relativistic Brachistochrone,'' J. Math. Phys. **27**, 507–511 (1985).
[2] G. Kamath, ''The brachistochrone in almost flat space,'' J. Math. Phys. **29**, 2268–2272 (1988).
[3] V. Perlick, ''The brachistochrone problem in a stationary space–time,'' J. Math. Phys. **32**, 3148–3157 (1991).
[4] F. Giannoni, P. Piccione, and J. A. Verderesi, ''An approach to the relativistic brachistochrone problem by sub–Riemannian geometry,'' J. Math. Phys. **38**, 6367–6381 (1997).
[5] V. Perlick and P. Piccione (unpublished).
[6] J. K. Beem, P. E. Ehrlich, and K. L. Easly, *Global Lorentzian Geometry* (Dekker, New York, 1996).
[7] S. W. Hawking and G. F. Ellis, *The Large Scale Structure of Space–Time* (Cambridge University Press, London, 1973).
[8] B. O'Neill, *Semi–Riemannian Geometry with Applications to Relativity* (Academic, New York, 1983).
[9] J. P. Serre, ''Homologie singuliere des espaces fibres,'' Ann. Math. **54**, 425–505 (1951).
[10] R. Palais, *Foundations of Global Nonlinear Analysis* (Benjamin, New York, 1968).
[11] V. Benci, D. Fortunato, and F. Giannoni, ''On the existence of geodesics in static Lorentz manifolds with nonsmooth boundary,'' Ann. Sci. Norm. Sup. Serie IV **XIX**, 255–289 (1992).
[12] J. Nash, ''The embedding problem for riemannian manifolds,'' Ann. Math. **63**, 20–63 (1956).
[13] E. Fadell and S. Husseini, ''Category of loop spaces of open subsets in Euclidean spaces,'' Nonlinear Analysis: Theory, Methods and Applications **17**, 1153–1161 (1991).

Modified real structures for quantum groups and applications

G. Grensing[a)]
Institut für Theoretische Physik und Astrophysik, Universität Kiel, D-24118 Kiel, Germany

(Received 28 October 1997; accepted for publication 28 July 1998)

A modification of the standard $*$-structure on a quasitriangular Hopf algebra is introduced by means of which, in particular, the quantum deformation $U_q(su_2)$ can be endowed with a real structure for q a pure phase. As opposed to the standard definition, its modified version not only works for the quantum group regarded as an algebra, but as a quasitriangular Hopf algebra if q is of modulus one. Applications to two relevant examples are given, which are known to require quantum symmetries: the anisotropic Heisenberg chain and the Landau theory on the torus, thus solving some unitarity problems that have always resisted a satisfactory solution. © *1998 American Institute of Physics.* [S0022-2488(98)03511-7]

I. INTRODUCTION

More than half a century was needed to realize that some quantum mechanical models do not admit conventional symmetries in the form of Lie groups, but instead require symmetries which have come to be known as quantum groups. Their underlying algebraic structure was revealed by Drinfeld;[1] they may all be subsumed under the class of quasitriangular Hopf algebras. The deformations of the classical Lie algebras, the q-deformed quantum enveloping algebras, are main examples of these constructs (see, e.g., Ref. 2).

The case of interest in many applications, both in mathematics and physics, is the one for which the deformation parameter q is a pure phase, or even a root of unity. On the mathematical side, this occurs for invariants of three-manifolds.[3] On the physical side, what we have in mind are examples such as the anisotropic spin chain,[4] living in one dimension, and the Landau theory of two-dimensional electrons,[5,6] which underlies the quantum Hall effect.

Despite its relevance, the case q a pure phase, or even a root of unity, is also the most difficult to handle. One reason is that no real structure that extends to the quantum enveloping algebra as a whole is known, because the standard definition[7] is not compatible with the coalgebra structure. Since a real structure dictates what is to be called a unitary or Hermitian operator, one thus gets confronted with severe unitarity problems. As for the Heisenberg spin chain, unitarity seems to be violated for $|q|=1$ because the Hamiltonian is not Hermitian; for the Landau theory of two-dimensional electrons, the quantum group which acts on the lowest Landau level wave functions carries the wrong real structure $U_q(su_{1,1})$, if one relies on the results given in the literature.

It is the purpose of the present paper to contribute to the solution of these problems. Related ideas have appeared in Refs. 8 and 9, but our approach is rather different; the treatment given in Ref. 8, which employs category-theoretic ideas, seems to be closest to ours.

The paper may roughly be divided into two parts: the first is concerned with the mathematical aspects of the problem. In Sec. II the standard definition of a real structure on a Hopf algebra is modified and extended for the quasitriangular case to the universal $\mathcal{R}$-matrix. The applications to the basic example $U_q(sl_2)$ are given in Sec. III, and it is shown that the number of its real structures is simply doubled; in particular, the case $|q|=1$, which has always played a singular role, is now included for the ''compact'' real form. Furthermore, the modified $*$-structure is applied to representations of the real form $U_q(su_2)$ with q a pure phase, which are now representations of the quasitriangular Hopf algebra, and not of the algebra only.

The second part is concerned with some physical aspects of the problem. In Sec. IV we deal with the applications to the XXZ-spin chain which, for open boundary conditions, admits a global

a)Electronic mail: g.grensing@theo-physik.uni-kiel.de

 © 1998 American Institute of Physics

$U_q(su_2)$ symmetry. The Hamilton operator is not hermitian in the naive sense, but the appropriate definition of hermiticity, derived from the modified real structure, will guarantee the eigenvalues to be real. We end this Section with a short discussion of closed boundary conditions, which are considerably more difficult to treat.[4] Recent work[10] has given hints that in this case a two-parameter deformation of gl_2 gets involved. We take another route to demonstrate this on invoking a generalization of the Wigner–Jordan transformation, which will turn out to be intimately connected with the notion of anyons.[11] The resulting two-parameter deformation $U_{q,t}(gl_2)$ is then shown to constitute the symmetry group of this system; but what we have not been able to resolve is the unitarity problem for closed boundary conditions. In Sec. V we deal with the Landau theory on the torus[6] with twisted boundary conditions, and it is shown that there is a residual symmetry, resulting from a discrete subgroup of the magnetic translation group, which gives rise to a unitary and irreducible representation of the finite-dimensional restriction of the quantum group $U_q(su_2)$ at a primitive root of unity. Section VI ends with some concluding remarks.

II. STANDARD AND MODIFIED REAL STRUCTURES ON QUASITRIANGULAR HOPF ALGEBRAS

To begin with, we briefly review the standard definition of a $*$-structure on a Hopf algebra. Let $\mathcal{A}$ be an associative algebra with unit over the field $k=\mathbb{C}$; then a $*$-structure is a conjugate-linear antimorphism $*:\mathcal{A}\to\mathcal{A}$ with $*^2=id_{\mathcal{A}}$. If, in addition, $\mathcal{A}$ is a bialgebra, the compatibility requirements for the comultiplication Δ and the count ε are usually taken to be

$$*\otimes *\circ\Delta=\Delta\circ *, \quad *\circ\varepsilon=\varepsilon\circ *, \tag{1}$$

where $*\circ\varepsilon(a)=\overline{\varepsilon(a)}$ with $a\in\mathcal{A}$ is the complex conjugate. Furthermore, if $\mathcal{A}$ is also a Hopf algebra with antipode σ, which we always assume to be invertible, there is no need to impose $*\circ\sigma=\sigma^{-1}\circ *$ as a further axiom, since this property follows from the conditions stated above.

There is another option to define the $*$-structure of a bialgebra, as we now explain. In order to motivate this modification, let us show the defining properties of a $*$-structure on an algebra $\mathcal{A}$ in the form of commutative diagrams. The $*$-involution is chosen so as to make the following diagrams commute:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{*} & \mathcal{A} \\ {\scriptstyle \mu_{op}}\uparrow & & \uparrow{\scriptstyle \mu} \\ \mathcal{A}\otimes\mathcal{A} & \xrightarrow[*\otimes *]{} & \mathcal{A}\otimes\mathcal{A} \end{array} \qquad \begin{array}{ccc} k & \xrightarrow{*} & k \\ {\scriptstyle \iota}\downarrow & & \downarrow{\scriptstyle \iota} \\ \mathcal{A} & \xrightarrow[*]{} & \mathcal{A} \end{array}$$

where ι is the identity, μ the multiplication, and μ_{op} its opposite.

For the standard choice of a $*$-structure on a bialgebra, the comultiplication and the count are $*$-morphisms of $\mathcal{A}$ as a coalgebra. A rather obvious modification is to define a $*$-structure for a coalgebra such that it is dual to the one for an algebra, i.e., in the form of commuting diagrams,

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{*} & \mathcal{A} \\ {\scriptstyle \Delta}\downarrow & & \downarrow{\scriptstyle \Delta^{op}} \\ \mathcal{A}\otimes\mathcal{A} & \xrightarrow[*\otimes *]{} & \mathcal{A}\otimes\mathcal{A} \end{array} \qquad \begin{array}{ccc} k & \xrightarrow{*} & k \\ {\scriptstyle \epsilon}\uparrow & & \uparrow{\scriptstyle \epsilon} \\ \mathcal{A} & \xrightarrow[*]{} & \mathcal{A} \end{array}$$

with $\Delta^{op}=\tau\circ\Delta$ and τ the transposition in $\mathcal{A}\otimes\mathcal{A}$. We thus see that the defining properties for the standard case are replaced by

$$*\otimes *\circ\Delta=\Delta^{op}\circ *, \quad *\circ\varepsilon=\varepsilon\circ *. \tag{2}$$

This is the case which we call modified in the sequel. That such modified $*$-structures indeed exist will be shown in Sec. III on the basic example $U_q(sl_2)$. Furthermore, if $\mathcal{A}$ is also a Hopf algebra, the antipode obeys $*\circ\sigma=\sigma\circ *$ for the modified $*$- structure, the proof being rather similar to the corresponding statement for the standard case.

We now assume that $\mathcal{A}$ is also endowed with a quasitriangular structure; what one needs is the behavior of the universal $\mathcal{R}$-matrix under the tensor product map $*\otimes *$. We find the two possi-

bilities: $*\otimes *(\mathcal{R})=\check{\mathcal{R}}$ and $*\otimes *(\mathcal{R})=\mathcal{R}^{-1}$ with $\check{\mathcal{R}}=\tau\mathcal{R}$ for the standard case, whereas $*\otimes *(\mathcal{R})=\check{\mathcal{R}}^{-1}$ and $*\otimes *(\mathcal{R})=\mathcal{R}$ for the modified case. It is straightforward to prove that all these possibilities are indeed compatible with the quasitriangular structure.

Finally, let us briefly discuss tensor product representations of quasitriangular Hopf algebras with standard and modified real structures.[12] If $\rho_V:\mathcal{A}\to End(V)$ is a $*$-representation of $\mathcal{A}$ on a finite-dimensional vector space V over $\mathbb{C}$ with a nondegenerate Hermitian form, then the Hermitian adjoint of the representation $\rho_{V,W}(a)=\rho_V\otimes\rho_W\circ\Delta(a)$ of $\mathcal{A}$ on $V\otimes W$ reads as $\rho^{\dagger}_{V,W}(a)=\rho_V\otimes\rho_W\circ *\otimes *\circ\Delta(a)$. This can be rewritten in terms of the adjoint $*(a)$ of $a\in A$. For the standard $*$-structure, the result is

$$\rho^{\dagger}_{V,W}(a)=\rho_{V,W}(*(a)), \tag{3}$$

and so the tensor product representation is again a $*$-representation; for the modified $*$-structure, we find

$$\rho^{\dagger}_{V,W}(a)=R_{V,W}\rho_{V,W}(*(a))R^{-1}_{V,W}. \tag{4}$$

In Sec. IV to come, we will have an occasion to discuss this result.

III. REAL STRUCTURES OF $U_q(sl_2)$ AND $*$-REPRESENTATIONS

Let us apply the above results to the quantum enveloping algebra $U_q(sl_2)$ with generators X_+, X_-, and H and relations

$$q^{H/2}X_{\pm}q^{-H/2}=q^{\pm 1}X_{\pm}, \quad [X_+,X_-]=\frac{q^H-q^{-H}}{q-q^{-1}}. \tag{5}$$

For the sake of brevity, the additional properties needed to endow this deformation with a quasitriangular Hopf structure are omitted.[2]

For the discussion of the $*$-structures of $U_q(sl_2)$, we rely on the known fact that real forms exist for q real, a pure phase, and purely imaginary; since in the latter case no classical limit exists, we confine ourselves to the first two possibilities. Furthermore, it is known that there are three nonequivalent real forms for the standard $*$-structure.[7] The modified $*$-structure adds three additional real forms, which just cover the alternative possibility, i.e., if $q\in\mathbb{R}$ is realized in the standard case (i), then the modified case (ii) realizes $|q|=1$, and vice versa. For the compact case $U_q(su_2)$ with $*(H)=H$ and $*(X_{\pm})=X_{\mp}$ the result is

$$\text{(i)}\quad q\in\mathbb{R}:*\otimes *(\mathcal{R})=\check{\mathcal{R}}, \quad \text{(ii)}\quad |q|=1:*\otimes *(\mathcal{R})=\check{\mathcal{R}}^{-1}. \tag{6}$$

For the two noncompact cases, we find for $U_q(su_{1,1})$ with $*(H)=H$ and $*(X_{\pm})=-X_{\mp}$:

$$\text{(i)}\quad q\in\mathbb{R}:*\otimes *(\mathcal{R})=\check{\mathcal{R}}, \quad \text{(ii)}\quad |q|=1:*\otimes *(\mathcal{R})=\check{\mathcal{R}}^{-1}, \tag{7}$$

and for $U_q(sl_2,\mathbb{R})$ with $*(H)=-H$ and $*(X_{\pm})=-X_{\pm}$]:

$$\text{(i)}\quad |q|=1:*\otimes *(\mathcal{R})=\mathcal{R}^{-1}, \quad \text{(ii)}\quad q\in\mathbb{R}:*\otimes *(\mathcal{R})=\mathcal{R}. \tag{8}$$

The proof (cf. also Ref. 2), which may be greatly simplified by using basic elements of q-analysis, is rather straightforward. As a rule, $K=e^H$ obeys $*(K)=K$ for the standard real structures, and $*(K)=K^{-1}$ for the modified real structures.

We thus see that all four possibilities for $*\otimes *(\mathcal{R})$, having been encountered in Sec. II, really do appear. Furthermore, we are now entitled to discuss $*$-representations of the ''compact'' real form $U_q(su_2)$ for q a pure phase, which are compatible with the structure of $U_q(su_2)$ as a Hopf algebra, and not as an algebra only.

So let us give a brief review of the known representation theory of the quantum deformation $U_q(su_2)$,[13] where we restrict ourselves to the case with $q=e^{2\pi i/p}$ a primitive root of unity. Since

the representations with dimension $n+1>p$ are reducible but not completely reducible,[2] we choose $n+1\leqslant p$. The irreducible highest weight representations take the form

$$X_+ w_r = \epsilon[n-r+1][r]w_{r-1}, \quad X_- w_r = w_{r+1}, \quad K w_r = \epsilon q^{n-2r} w_r, \tag{9}$$

with $0\leqslant r\leqslant n$, $\epsilon=\pm 1$, and $[x]=(q^x-q^{-x})/(q-q^{-1})$; this result holds as well for $q\in\mathbb{R}$ and $n\geqslant 0$. For q a root of unity, the generators obey the additional restrictions $K^{2p}-1=0$ and $X_\pm^p=0$. If q is an odd root of unity and $\epsilon=+1$, we even have the stronger restriction $K^p=1$; but, as will be seen, we shall have a need for the weaker version.

Since these restrictions are obeyed on all irreducible representations, one can as well impose them on the algebra level as additional relations. This is achieved on passing from K to the square root $L=K^{1/2}$ and, with the obvious changes in the defining properties of $U_q(sl_2)$, it is straightforward to prove that the two-sided ideal I, generated by the elements $L^{4p}-1$ and $X_\pm^p$, is a Hopf ideal. Hence, $U_q(sl_2)'=U_q(sl_2)/I$ is again a Hopf algebra. This finite-dimensional restriction[2] is even quasitriangular with a universal $\mathcal{R}$-matrix,

$$\mathcal{R}=\mathcal{R}_L \sum_{m=0}^{p-1} r_m (X_+ L^{+1}\otimes L^{-1} X_-)^m, \tag{10}$$

where $\mathcal{R}_L=(1/4p)\,\Sigma_{k,l=0}^{4p-1} q^{-(1/2)kl} L^k\otimes L^l$ and $r_m=q^{(1/2)m(m-1)}(q-q^{-1})^m/[m]!$; furthermore, we have $*\otimes *(\mathcal{R})=\check{\mathcal{R}}^{-1}$. For q an even root of unity, this is a known result;[3] the proof for q an odd root of unity is rather similar.

A nondegenerate Hermitian form can be constructed on the $n+1$-dimensional vector space, with respect to which the representations are $*$-representations. For $q\in\mathbb{R}$, they read

$$X_\pm e^j{}_m = (\epsilon[j\mp m][j\pm m+1])^{1/2} e^j{}_{m\pm 1}, \quad K e^j{}_m = \epsilon q^{2m} e^j{}_m, \tag{11}$$

with $n=2j$ and $+j\geqslant m\geqslant -j$; they are well-known for $\epsilon=+1$.[14] For q an even root of unity and $\epsilon=+1$, it is easy to prove that the irreducible representations with $n+1\leqslant p$ are unitary and take the same form; this result is related to Jones' work on subfactors.[15]

If q is an odd root of unity, no general statement can be made. Of special importance are those representations which are on the verge of getting reducible and indecomposable, i.e., those with $n+1=p$. For $\epsilon=-1$, they are unitary and can be shown to read

$$X_\pm e^j{}_m = [m\pm\tfrac{1}{2}] e^j_{m\pm 1}, \quad K e^j{}_m = -q^{2m} e^j{}_m. \tag{12}$$

This result, which may also formally be obtained from (11) on choosing $q=e^{2\pi i/p}$ with p odd and $\epsilon=-1$, will be referred to in Sec. V below.

IV. SPIN CHAINS AND UNITARITY

Let us apply the above results to the anisotropic XXZ-spin chain, which is described by the Hamiltonian,[16,4,17]

$$\hat{H}=-\frac{J}{4}\sum_{i=1}^{N-1}\{2(\sigma_i^+\sigma_{i+1}^- + \sigma_i^-\sigma_{i+1}^+) + \gamma\sigma_i^3\sigma_{i+1}^3 + \delta(\sigma_i^3-\sigma_{i+1}^3) + \varepsilon\}. \tag{13}$$

Here, the σ_i are local spin-$\frac{1}{2}$ operators, living on the sites $i=1,\ldots,N-1$ of a one-dimensional finite lattice with boundary conditions which, for the present, we assume to be open. The local symmetry is the quantum algebra $U_q(su_2)$ since its fundamental representation coincides with the undeformed version. The constants appearing in (13) are given in terms of the deformation parameter q as $\gamma=(q+q^{-1})/2$, $\delta=-(q-q^{-1})/2$, and $\epsilon=-\delta$. They are designed so as to render the Hamiltonian invariant under a global $U_q(su_2)$, the generators of which are determined by the $(N-1)$-fold coproduct. There are two routes to prove the global invariance, both of which will be made use of in the subsequent development.

The first proof relies on the observation that one can read the Hamiltonian as[4] $\hat{H}=(J/2)\,\Sigma_{i=1}^{N-1}E_i$, where the operators $E_i=(\overset{i-1}{\otimes}1)\otimes E\otimes(\overset{N-i-1}{\otimes}1)$ with

$$E=-\tfrac{1}{2}(2(\sigma^+\otimes\sigma^-+\sigma^-\otimes\sigma^+)+\gamma\sigma^3\otimes\sigma^3+\delta(\sigma^3\otimes 1-1\otimes\sigma^3)+\epsilon 1\otimes 1) \tag{14}$$

form a Temperley–Lieb–Jones (TLJ) algebra;[18] our conventions are fixed by $E_i^2=2\gamma E_i$. Hence, the operators[15] $S_i=qE_i-1$ obey the braid group relations; furthermore, the global generators commute with $\hat{H}$, since they commute with each single TLJ operator.

The second proof relies on the observation that the Hamiltonian is built from the coproduct of the Casimir operator $C=J^+J^-+[J^3][J^3-1]$, where the J's are now understood to denote the generators of $U_q(su_2)$ in an arbitrary representation. Indeed, in the fundamental representation, we have $\Delta^\natural(C)=-E+2\gamma 1\otimes 1$. On introducing $\Delta^\natural_{i,i+1}(C)=(\overset{i-1}{\otimes}1)\otimes\Delta^\natural(C)\otimes(\overset{N-i-1}{\otimes}1)$, the Hamiltonian can be expressed as a sum of such terms. So the proof is complete if we can show that

$$\Delta_{i,i+1}(C)\Delta^{N-1}(a)=\Delta^{N-1}(a)\Delta_{i,i+1}(C), \tag{15}$$

for all $i=1,\ldots,N-1$ and a in $U_q(su_2)$. But this property follows by induction, using coassociativity and the fact that the Casimir operator is central.

However, what has always been something of a mystery is that, for q a pure phase, the Hamiltonian is not self-adjoint because the boundary term in (13) (the one proportional to ε) is purely imaginary; so one would not expect the eigenvalues to be real.

This intriguing contradiction is resolved by means of the modified real structure of $U_q(su_2)$. Again, we make use of the fact that $\hat{H}$ may be written as a sum of Casimir operators, where it is essential that the prefactor γ of the additional term is a real number and thus does not affect the conclusions to follow.

For the discussion of the hermiticity properties of $\hat{H}$, we may thus restrict ourselves to the investigation of $\Delta(C)$, and for this purpose we go back to Sec. II. The naive Hermitian conjugation of

$$\Delta(C)=\Delta(J^+)\Delta(J^-)+[\Delta(J^3)][\Delta(J^3)-1], \tag{16}$$

with $*(J^\pm)=J^\mp$ amounts to Eq. (1) with $a=C$, i.e., $*\otimes *\cdot\Delta(C)=\Delta(*(C))=\Delta(C)$ since C is self-adjoint; but for $|q|=1$ this is the wrong result. The reason is, of course, that Eq. (1) only holds for q real. For q a pure phase, we must use instead Eq. (2), i.e., $*\otimes *\cdot\Delta(C)=\Delta^{op}(C)$, and this is indeed the correct result.

Turning to representations, we know that for q an even root of unity, the representations are $*$-representations with a positive definite Hermitian form. In particular, we can thus resort to [see (4)] $\Delta^\natural(C)^\dagger=R^\natural\Delta^\natural(C)R^{\natural\,-1}$, where $R^\natural$ denotes the representation of $\mathcal{R}$ in the tensor product of two fundamental representations. As can easily be shown now, although $\Delta^\natural(C)$ is only ''almost'' Hermitian, it nevertheless has real eigenvalues.

In summary, the spin chain Hamiltonian with open boundary conditions is not Hermitian in the naive sense, but the generalized hermiticity property (4) at least guarantees its eigenvalues to be real. This property holds without qualification only if q is an even root of unity.

We want to conclude this section with some comments on the spin chain with closed boundary conditions. These amount to the replacement $N\to N+1$ and, for the term with $i=N$ in (13), one defines $\sigma^\pm_{N+1}=e^{\pm i\varphi}\sigma^\pm_1$ and $\sigma^3_{N+1}=\sigma^3_1$, where the angle φ describes the twisting. It is advantageous to remove the exceptional role of σ_{N+1} by means of the substitution $\sigma_i^\pm\to t^{\pm i}\sigma_i^\pm$ and $\sigma_i^3\to\sigma_i^3$ with $t^N=e^{i\varphi}$, which does not change the local algebra; the boundary conditions then become periodic in the usual sense. The Hamiltonian now takes the form as above, with N replaced by $N+1$, and

$$E(q,t)=-\tfrac{1}{2}(2(t^{-1}\sigma^+\otimes\sigma^-+t\sigma^-\otimes\sigma^+)+\gamma\sigma^3\otimes\sigma^3+\delta(\sigma^3\otimes 1-1\otimes\sigma^3)+\epsilon 1\otimes 1). \tag{17}$$

Remarkably, the operators $E_i(q,t)$ still obey a TLJ algebra for t arbitrary.[19]

In order to motivate what follows, let us consider the following generalization of the Jordan–Wigner transformation:

$$S_i^- = e^{i\pi\kappa\sum_{j=1}^{N}\theta(i-j)c_j^\dagger c_j} c_i, \quad S_i^3 = c_i^\dagger c_i - \frac{1}{2}, \tag{18}$$

where the c_i are fermionic operators with $\{c_i, c_j^\dagger\} = \delta_{ij}$ and θ is the ''lattice angle'' in one dimension with $\theta(j)$ equal to 0 for $j \leqslant 0$ and 1 for $j > 0$; the original case[20] is recovered for $\kappa = 1$. This generalization leads to anyons,[11,21] usually considered in two dimensions only, but this concept works in one dimension as well. Indeed, the operators $S_i^\pm$ and S_i^3 obey relations such as $S_i^+ S_j^- + e^{-i\pi\kappa} S_j^- S_i^+ = 0$ for $i \neq j$, being characteristic of anyons. This relation guides us to introduce the deformed commutator,

$$[S_i^+, S_j^-]_t = t^{-1} S_i^+ S_j^- - t S_j^- S_i^+ = t^{-2S_i^3} [2S_i^3]_q \delta_{ij}, \tag{19}$$

with $t^2 = -e^{-i\pi\kappa}$, where the last equality is easily verified; the other relations remain undeformed. Hence, the anyonization of one-dimensional fermions on a lattice leads to a two-parameter deformation of sl_2, q being a measure of the anisotropy, and t is related to the boundary conditions.

We thus have obtained a two-parameter deformation $U_{q,t}(sl_2)$,[22] but what is still missing is the global invariance of the Hamiltonian. This can be achieved on supplementing the generators $S^\pm$ and S^3 by a central element Z. Then the coproduct must be modified as follows:

$$\Delta(S^\pm) = S^\pm \otimes q^{+S^3} t^{-(S^3 \pm Z)} + q^{-S^3} t^{-(S^3 \mp Z)} \otimes S^\pm, \quad \Delta(S^3) = S^3 \otimes 1 + 1 \otimes S^3, \tag{20}$$

and for the count we choose $\epsilon(Z) = 0$; a Hopf algebra is obtained on defining the antipode as

$$\sigma(S^3) = -S^3, \quad \sigma(S^\pm) = -(qt)^{\pm 1} S^\pm t^{2S^3}, \quad \sigma(Z) = -Z. \tag{21}$$

This is a deformation $U_{q,t}(gl_2)$ of the general linear group in two dimensions (Ref. 10); see also Ref. 23). As can now be shown by explicit computation, the generalized TLJ operators $E_i(q,t)$ commute with the generators,

$$J^\pm = \sum_{i=1}^{N+1} \left(\overset{i-1}{\otimes} q^{-1/2\,\sigma^3} t^{-(1/2)(\sigma^3 \mp 1)} \right) \otimes \sigma^\pm \otimes \left(\overset{N-i+1}{\otimes} q^{+(1/2)\sigma^3} t^{-(1/2)(\sigma^3 \pm 1)} \right), \tag{22}$$

$$J^3 = \sum_{i=1}^{N+1} (\overset{i-1}{\otimes} 1) \otimes \frac{1}{2}\sigma^3 \otimes (\overset{N-i+1}{\otimes} 1),$$

of the global $U_{q,t}(gl_2)$, which thus constitutes the quantum symmetry of the spin chain with closed boundary conditions.[24]

V. LANDAU THEORY AND QUANTUM SYMMETRY

Another respectable problem where quantum symmetries make their appearance is the quantum Hall effect.[25] Recently, the deformation $U_q(su_2)$ has been shown to be the spectrum generating algebra for odd filling fractions $\nu = 1/p$.[26] Here, we are concerned with the underlying Landau theory of two-dimensional electrons, confined to a rectangular sample of extensions L_x and L_y, and subjected to a strong constant magnetic field B in the third direction. Hence, boundary conditions will again play a crucial role.

Thus, the point of departure is the Hamilton operator,

$$\hat{H} = \frac{1}{2m}(\pi_x{}^2 + \pi_y{}^2), \tag{23}$$

with $\pi_i = p_i + (e/c)A_i$ and $i = x,y$. The magnetic translations[27] $T(\alpha,\beta) = \exp((1/2)(\alpha Y - \beta X))$, where $X = x - (1/\omega B)\pi_y$ and $Y = y + (1/\omega B)\pi_x$, commute with the Hamiltonian and yield a Heisenberg–Weyl group; here $l = \sqrt{\hbar c/eB}$ denotes the magnetic length and $\omega = eB/mc$ the cyclotron frequency. On wave functions they act as

$$T(\alpha,\beta)\psi(x,y) = \exp\left(-\frac{i}{l^2}\left(\frac{1}{2}\alpha\beta + \beta x \right) \right) \psi(x+\alpha, y+\beta), \tag{24}$$

and choosing the Landau gauge[28] $A_x = -By, A_y = 0$, we impose twisted boundary conditions, which are compatible with the symmetries of the system, by requiring

$$T(L_x,0)\psi_0(x,y) = e^{2\pi i\mu}\psi_0(x,y), \quad T(0,L_y)\psi_0(x,y) = e^{2\pi i\nu}\psi_0(x,y), \tag{25}$$

with $\mu, \nu \in [0,1]$. The ground state solutions can be expressed in terms of θ-functions with characteristics[29] as

$$\psi_{o;m}(x,y) = \langle x,y|m\rangle = c_0 \exp\left(-\frac{1}{2l^2}y^2 \right) \theta\begin{bmatrix} \dfrac{m+\mu}{N} \\ \nu \end{bmatrix} (N\overline{v}|N\tau), \tag{26}$$

with $v = (x+iy)/L_x$, and $\tau = iL_y/L_x$ the modular parameter of the torus; furthermore, $N = L_xL_y/2\pi l^2$ is the degree of degeneracy, and m takes values $(N-1)/2 \geqslant m \geqslant -(N-1)/2$. These ground state wave functions are orthonormal for $c_0 = (1/\pi l^2 L_x^2)^{1/4}$, where the integration extends over the fundamental domain of the torus only.

There remains a discrete subgroup of the Weyl group, which arises as follows. The ''latticized'' magnetic translations $T = T(L_x/N,0)$ and $S = T(0,L_y/N)$ obey the relation $TS = qST$, where $q = e^{2\pi i(1/N)}$. They act on the ground state wave functions as

$$S|m\rangle = q^{m+\mu}|m\rangle, \quad T|m\rangle = q^{\nu}|m-1\rangle, \tag{27}$$

with $S^N = e^{2\pi i\mu}$ and $T^N = e^{2\pi i\nu}$ in this representation. As a consequence, the discrete subgroup, which is generated by S and T, subject to the above relation, acts unitarily and irreducibly on the N-dimensional Hilbert space, being built from the lowest Landau level wave functions on the torus.

On passing from the discrete subgroup to the corresponding group algebra, it is legitimate to take linear combinations of products of its generating elements. In particular, it is easy to guess the form of the operators,

$$J^{\pm} = \frac{q^{\mp 1/2}S - q^{\pm 1/2}S^{-1}}{q - q^{-1}} T^{\mp 1}, \quad K = -S^2, \tag{28}$$

being designed so as to act on the basis as

$$J^{\pm}|m\rangle = \left[m \pm \frac{1}{2} \right] |m \pm 1\rangle, \quad K|m\rangle = -q^{2m}|m\rangle, \tag{29}$$

for $\mu = \nu = 0$.[30] This is recognized as the irreducible and unitary representation (12) of Sec. III, and so the operators (28) are the generators of the finite-dimensional restriction $U_q(su_2)'$; furthermore, the degeneracy number N must necessarily be an odd[31] integer. Remarkably, they belong to the class of representations with quantum dimension zero, which are regarded as ''unphysical'' in another context.[32]

The final result, Eq. (29), looks similar to what is reported in Refs. 6,33, but differs in at least one essential aspect, since the action of the generators J^+ and J^- does not contain a relative minus sign. This makes a big difference because we then had the wrong real structure $U_q(su_{1,1})$, and it is known that this ''noncompact'' real form does not admit any finite-dimensional irreducible and unitary representations.

To collect what we have shown, the (modular invariant) wave functions of the lowest Landau level on the torus constitute the representation space of an irreducible and unitary representation of the finite-dimensional quantum group $U_q(su_2)'$ for q an odd root of unity, the generators being built from the discrete subgroup of the Weyl group. To put it differently,[34] theta functions are intimately related to representations of quantum groups at the roots of unity.

VI. CONCLUSION

Having provided for a real structure that makes sense of the quantum deformation $U_q(su_2)$ as a quasitriangular Hopf algebra for q a pure phase, nevertheless, there remains some work to be done.

On the mathematical side, we have left open the application to quantum enveloping algebras of the form $U_q(\mathcal{G})$ with $\mathcal{G}$ an element of the Cartan list, as well as their affine generalizations; the standard case is solved in Ref. 35. We also have not touched upon the implications for the Faddeev–Reshetikhin–Takhtadjian construction.

On the physical side, as for the applications to quantum spin chains, for closed boundary conditions we have proven the invariance under a two-parameter deformation, but unitarity remains an issue. As for the Landau theory, we have not gone very far since what one aims at is the transition to the many-particle problem. The simplest kind of approach would be to pass to the N_e-fold tensor product. Since we know now that the finite-dimensional restriction $U_q(su_2)'$ for q an odd root of unity acts irreducibly and unitarily on the Hilbert space of the lowest Landau level with quantum dimension zero, the braiding of the electrons is then determined by the $\mathcal{R}$-matrix. A more ambitious program would be to take the Laughlin ground state[36] as the vacuum, and to perform second quantization over this new ground state. But the Laughlin ground state is not appropriate since it is not even a singlet under the quantum symmetry; the trivial representation $\varepsilon: U_q(su_2) \to \mathbb{C}$ is not realized because $\varepsilon(K)=1$ is violated. What one needs instead is the analog of the Laughlin ground state on the torus. This was given soon after the advent of Laughlin's theory in an important paper by Haldane and Rezayi.[37] But we have reasons[26] to believe that this degenerate ground state is restricted to the sector with total charge zero, and so is hardly related to the quantum Hall effect.

[1] V. G. Drinfeld, in *Proceedings of the International Congress of Mathematics, Berkeley 1986*, edited by A. M. Gleason (American Mathematical Society, Providence, RI, 1987).

[2] V. Chari and A. Pressley, *A Guide to Quantum Groups* (Cambridge University Press, Cambridge, 1994); S. Majid, *Foundations of Quantum Group Theory* (Cambridge University Press, Cambridge, 1995); C. Kassel, *Quantum Groups* (Springer-Verlag, New York, 1995).

[3] N. Yu. Reshetikhin and V. G. Turaev, Invent. Math. **103**, 547 (1991).

[4] V. Pasquier and H. Saleur, Nucl. Phys. B **330**, 523 (1990).

[5] N. B. Wiegmann and A. V. Zabrodin, Nucl. Phys. B **422**, 495 (1994).

[6] I. I. Kogan, Int. J. Mod. Phys. A **9**, 3887 (1994).

[7] T. Masuda, K. Mimachi, Y. Nakagami, M. Noumi, Y. Saburi, and K. Ueno, Lett. Math. Phys. **19**, 187 (1990).

[8] V. V. Lyubashenko, in *Quantum Groups*, edited by P. P. Kulish (Springer-Verlag, Berlin, 1992).

[9] G. Mack and V. Schomerus, Nucl. Phys. B **370**, 185 (1992).

[10] M. R. Monteiro, I. Roditi, L. M. C. S. Rodrigues, and S. Sciuto, Mod. Phys. Lett. A **5**, 419 (1995).

[11] F. Wilczek, in *Fractional Statistics and Anyon Superconductivity* (World Scientific, Singapore, 1990).

[12] Note that we stick to the conventional definition of the *- structure on tensor products; in particular, on $\mathcal{A}\otimes\mathcal{A}$ we have $*\otimes*(a\otimes a')=*(a)\otimes*(a')$ for $a,a'\in\mathcal{A}$. In Ref. 9 the order of the factors on the right-hand side is reversed.

[13] P. Roche and D. Arnaudon, Lett. Math. Phys. **17**, 295 (1989); G. Keller, *ibid.* **21**, 273 (1991).

[14] A. N. Kirillov and N. Yu. Reshetikhin, in *Infinite Dimensional Lie Algebras and Groups*, edited by V. G. Kac (World Scientific, Singapore, 1989).

[15] V. Jones, Ann. Math. **126**, 335 (1987).

[16] F. C. Alcaraz, M. N. Barber, M. T. Batchelor, R. I. Baxter, and G. R. W. Quispel, J. Phys. A **20**, 6397 (1987).

[17] L. Mecinescu and R. I. Nepomechie, in *Quantum Groups*, edited by T. Curtright, D. Fairlie, and C. Zachos (World Scientific, Singapore, 1990).

[18] H. Temperley and E. Lieb, Proc. R. Soc. London **322**, 251 (1971).

[19] One could guess that the additional parameter t gives rise to the two-variable knot invariant, but in the skein relations it drops out; cf. also N. Yu. Reshetikhin, Lett. Math. Phys. **20**, 331 (1990).

[20] P. Jordan and E. P. Wigner, Z. Phys. **47**, 631 (1928).

[21] M. Lüscher, Nucl. Phys. B **326**, 557 (1989); E. Fradkin, Phys. Rev. Lett. **63**, 322 (1989).

[22] This approach is rather different from the one given by A. Lerda and S. Sciuto, Nucl. Phys. B **401**, 613 (1993), where it is essential that two Schwinger fermions are needed to connect the anyonic operators with the generators of the one-parameter deformation $U_q(sl_2)$. The approach given here can also be generalized to two dimensions on replacing

the step function θ by the two-dimensional lattice angle function, as given from Chern–Simons theory.
[23] N. Yu. Reshetikhin, Lett. Math. Phys. **20**, 331 (1990); A. Sudbery, J. Phys. A **23**, L697 (1990); A. Schirrmacher, J. Wess, and B. Zumino, Z. Phys. C **49**, 317 (1991); C. Burdik and L. Hlavaty, *ibid.* **24**, L165 (1991).
[24] What we have not been able to show is that the Hamiltonian has only real eigenvalues because we do not know of a real structure for $|q|=1$ and $|t|=1$, which is compatible with the coalgebra structure; it exists for $|q|=1$ and t real, but this is not the case of interest.
[25] R. Prange and S. Girvin, in *The Quantum Hall Effect* (Springer-Verlag, New York, 1990); M. Stone, in *Quantum Hall Effect* (World Scientific, Singapore, 1992).
[26] G. Grensing, ''Fractional quantum Hall effect and quantum symmetry,'' preprint university of Kiel, August 1997.
[27] E. Brown, Phys. Rev. A **133**, 1038 (1964); J. Zak, *ibid.* **134**, 1602,1607 (1964).
[28] From Kogan's approach (Ref. 6) one might infer that it is essential to work in the symmetric gauge; but on the torus one must stay in the unsymmetric gauge.
[29] D. Mumford, *Tata Lectures on Theta I*, Progress in Mathematics 28 (Birkhäuser, Boston, 1983).
[30] For nonzero values of the phases μ and ν, the representations generally belong to the cyclic series. However, there is one special case for which, again, a highest weight representation is obtained. Namely, if the ground state wave functions are required to be modular invariant, one can show that this requirement is met if the phases take the values $\mu=-\nu=\frac{1}{2}$; hence, there are no spin structures. It is straightforward to prove that the resulting representation is equivalent to (29).
[31] For q an even root of unity, the representations with $n+1=p$ cannot be brought into the form (12); the best that can be achieved is $X_{\pm}e^j_m=[j\mp m]e^j_{m\pm1}$ and $Ke^j_m=q^{2m}e^j_m$ in the notation of Sec. III. Also note the crucial role of the choice $\epsilon=-1$ in (12).
[32] L. Alvarez-Gaumé, C. Goméz, and G. Sierra, Nucl. Phys. B **330**, 347 (1990); C. Goméz and G. Sierra, *ibid.* **352**, 791 (1991).
[33] H. T. Sato, Mod. Phys. Lett. A **9**, 1819 (1994).
[34] G. Grensing, Phys. Lett. B **419**, 258 (1998).
[35] E. Twietmeyer, Lett. Math. Phys. **24**, 49 (1992).
[36] R. B. Laughlin, Phys. Rev. Lett. **50**, 1395 (1983).
[37] F. D. M. Haldane and E. H. Rezayi, Phys. Rev. B **31**, 2529 (1985); G. Cristofano, G. Maiella, R. Musto, and F. Nicodemi, Mod. Phys. Lett. A **6**, 1779 (1991); Nucl. Phys. B (Proc. Suppl.) **33**C, 119 (1993).

Hamiltonian structure and Darboux theorem for families of generalized Lotka–Volterra systems

Benito Hernández-Bermejo and Víctor Fairén[a)]
Departamento de Física Fundamental, Universidad Nacional de Educación a Distancia, Senda del Rey S/N, 28040 Madrid, Spain

(Received 3 November 1997; accepted for publication 16 July 1998)

This work is devoted to the establishment of a Poisson structure for a format of equations known as generalized Lotka–Volterra systems. These equations, which include the classical Lotka–Volterra systems as a particular case, have been deeply studied in the literature. They have been shown to constitute a whole hierarchy of systems, the characterization of which is made in the context of simple algebra. Our main result is to show that this algebraic structure is completely translatable into the Poisson domain. Important Poisson structures features, such as the symplectic foliation and the Darboux canonical representation, rise as a result of rather simple matrix manipulations. © *1998 American Institute of Physics.* [S0022-2488(98)02411-6]

I. INTRODUCTION

Poisson structures[1,2] (sometimes named *generalized Hamiltonian structures* in the literature) are ubiquitous in all fields of mathematical physics, from finite-dimensional dynamical systems[3–7] to field theories:[8,9] Fluid dynamics,[10,11] magnetohydrodynamics,[11,12] plasmas,[13–15] continuous media,[15] condensed matter,[16] etc. Reformulating a given problem in terms of a Poisson structure provides fruitful insight into the behavior of the system, which may take the form of perturbative solutions,[17] nonlinear stability analysis through the energy-Casimir algorithm[7,18] or the energy-momentum method,[19] bifurcation properties and characterization of chaotic dynamics,[20] integrability results,[21] application of reduction of order procedures[2,22] or explicit determination of new solutions.[14,23]

In the present work, we shall restrict ourselves to finite-dimensional Poisson structures. In terms of local coordinates, a Poisson system defined on an n-dimensional manifold takes the following form:

$$\dot{x}_i = \sum_{j=1}^{n} J_{ij}\partial_j H, \quad i=1,...,n. \tag{1}$$

The smooth, real-valued function $H(x)$ in (1) is a constant of motion of the system, which plays the role of Hamiltonian, and the $J_{ij}(x)$ are also smooth and real-valued, being the entries of a $n \times n$ skew-symmetric structure matrix $\mathscr{J}$ which verifies the Jacobi equations:

$$\sum_{l=1}^{n} (J_{li}\partial_l J_{jk} + J_{lj}\partial_l J_{ki} + J_{lk}\partial_l J_{ij}) = 0. \tag{2}$$

Here ∂_l means $\partial/\partial x_l$ and indices i,j,k run from 1 to n. The flow (1) can then be expressed as $\dot{x}_i = [x_i, H]$, in terms of the Poisson bracket defined by

$$[F,G] = \sum_{i,j=1}^{n} \frac{\partial F}{\partial x_i} J_{ij} \frac{\partial G}{\partial x_j},$$

a)Author to whom correspondence should be addressed. Electronic mail: vfairen@uned.es

 © 1998 American Institute of Physics

where F and G are smooth real-valued functions defined on the Poisson manifold. Consequently, Poisson structures generalize classical Hamiltonian systems, for which $\mathscr{J}$ is the well-known symplectic matrix. In particular, the classical restriction to even-dimensional manifolds is not present in Poisson systems. However, Poisson dynamics preserves the Hamiltonian character of the motion. This is proven by Darboux theorem,[2] which states that there exist local coordinates in the neighborhood of every point of the Poisson manifold, such that the equations of motion take essentially the classical Hamiltonian form. The practical construction of Darboux coordinates is, however, a complicated task in general, which has been carried out only for a limited sample of systems.[2,24,25]

An important question is that of characterizing a given vector field not in form (1), as an actual Poisson system. In the finite-dimensional case the problem amounts to giving a procedure for decomposing (whenever possible) a smooth function $f(x):\Omega\subset R^n\to R^n$, where Ω is open, as $f(x)=\mathscr{J}(x)\cdot\nabla H(x)$, where $\mathscr{J}$ is a solution of the nonlinear PDE (2) and $H(x)$ is a real-valued function. This is a nontrivial problem to which important efforts have been devoted in past years in a variety of approaches.[3–8,25–28] The question is well understood in the simplest cases—two and three dimensions—and the existence and determination of at least one Poisson structure is ensured if a first integral is known for the system.[25,28] In higher dimensions the situation is by no means so clear, and comparable results are still lacking.

The main exception to this absence of results in n-dimensional systems is, to our knowledge, given by the Lotka–Volterra equations (LV from now on). They were introduced by Lotka[29] and Volterra[30] in chemical and biological contexts, respectively, and Volterra himself was already aware of the (classical) Hamiltonian nature of some LV systems. However, the main and more systematic advances are due to Kerner,[31] who developed a biological analog of classical statistical mechanics for predator-prey systems. More recently, the attention has turned to the search of Poisson structures for more general LV models, since Poisson structures generalize classical Hamiltonian systems while retaining the Hamiltonian nature of the dynamics—in this sense, the Poisson structure of many two and three-dimensional LV systems has been established,[3,25,27,28] and all three-dimensional biHamiltonian LV systems have been classified by Plank.[5] Also, in the domain of n-dimensional LV flows, a first tentative classification of Poisson structures has been carried out.[6]

Both the relevance of LV equations and the importance of finding their Hamiltonian or Poisson representations, transcends the fact that LV models are appropriate in describing many problems in Biology, Chemistry or Physics. Cairó and Feix,[32] for example, refer to a fairly long list of systems modeled by LV equations—their ubiquity has even prompted Peschel and Mende[33] to head their book on the issue with the title: *Do we live in a Volterra World?*

Actually, the significance of LV equations goes beyond a strict modeling context, because they have been shown to be canonical representatives of infinite families, or classes, of very general flows,[33,34] to which, following Brenig, we shall refer as Generalized Lotka–Volterra (GLV) systems. There is a whole formalism associated to the GLV equations.[33–36] As we shall see later, the most relevant feature of this formalism is that of permitting the analysis and interpretation of certain properties of the vector field in purely algebraic terms.

Our purpose in this article is to demonstrate how these algebraic properties are of fundamental importance in understanding the Poisson structure of LV and GLV models. We shall investigate Poisson structures for GLV families of systems. They include as particular elements all LV models which have been the object of interest in the literature in relation to Poisson structures (in works of Volterra, Kerner and Plank). We shall show that the algebraic GLV matrix properties can be translated into the Poisson context and acquire a new significance. The reverse is also true, thus defining a close connection between GLV algebraic properties and the Poisson structure of the system.

Implementing the structure of the GLV formalism on its Poisson counterpart has very interesting consequences. First of all, we are able to include in our scheme systems which are more general than the LV ones. Second, we can take into account larger sets of LV flows than those studied by previous approaches. For example, we are neither limited to LV systems of even dimension, nor cases with a unique fixed point. These are two common restrictions often imposed in the literature,[6,37] which we obviate at once. Third, we are able to capture important phase-space features in terms of simple properties of constant matrices. Finally, our approach leads to an

algorithmic reduction to the Darboux form for the equations. And last but not least, our construction is always global.

II. OVERVIEW OF THE GLV FORMALISM

We proceed now to briefly summarize the main features of the GLV formalism. We refer the reader interested in a more detailed exposition to the original references.[33–36]

Definition 2.1: A GLV system is a set of ordinary differential equations which is defined in the real positive orthant and complies to the form:

$$\dot{x}_i = x_i\left(\lambda_i + \sum_{j=1}^{m} A_{ij}\prod_{k=1}^{n} x_k^{B_{jk}}\right), \quad i=1...n, \tag{3}$$

where n and m are positive integers, $m \geqslant n$, and A, B and λ are $n\times m$, $m\times n$ and $n\times 1$ real matrices, respectively.

The m nonlinear terms of the right-hand side of (3) are usually known as *quasimonomials*. Sometimes we shall group all the coefficients of A and λ in a single, composite matrix, $M=(\lambda|A)$. We shall assume that matrix B is of maximal rank. This is a standard case to which every GLV system can be reduced.[36] Notice also that the well-known LV equations

$$\dot{x}_i = x_i\left(\lambda_i + \sum_{j=1}^{n} A_{ij}x_j\right), \quad i=1...n$$

are a particular case of (3) where $m=n$ and B is the $n\times n$ identity matrix.

System (3) is form-invariant under quasimonomial transformations (or QMTs from now on):

$$x_i = \prod_{k=1}^{n} y_k^{C_{ik}}, \quad i=1,...,n, \quad |C|\neq 0. \tag{4}$$

Under (4), matrices B, A, λ and M change to $B'=B\cdot C$, $A'=C^{-1}\cdot A$, $\lambda'=C^{-1}\cdot\lambda$ and $M'=C^{-1}\cdot M$, respectively, but the GLV format is preserved. Obviously, all GLV systems which can be connected through QMTs share the value of the product $B\cdot M$. These families of systems are, in fact, classes of equivalence, the product $B\cdot M$ being a class invariant. The QMTs are diffeomorphisms defined in the positive orthant, and are orientation-preserving iff $|C|>0$. Consequently, QMTs preserve the topology of the phase portrait modulo an inversion.

Definition 2.2: A GLV class of equivalence for which rank $(M)=r$, and whose members are n-dimensional and have m quasimonomials, is denoted as an (r,n,m)-class.

The kind of manipulations in which we shall be interested later will transform a GLV system into another one belonging to the same or, eventually, to a different class. However, these manipulations will affect neither r nor m, but may change n. We shall always operate, however, in the range $r\leqslant n\leqslant m$. Obviously, a QMT does not modify anyone of these three indexes.

If $m=n$, we can perform a QMT of matrix $C=B^{-1}$. The result is another flow for which $B'=I_n$, that is, a LV system. Such a system can be taken as the canonical representative of the GLV class of equivalence.

In the complementary case $m>n$, there is no LV system inside the class of equivalence. However, the reduction to the LV form is possible if we perform an embedding, just by adding new variables to system (3).

Definition 2.3: We define a p-embedding as the result of adding to a GLV system p new variables in the following way:

$$\dot{x}_i = 0, \quad x_i(0)=\alpha_i>0, \quad i=n+1,...,n+p, \quad 1\leqslant p\leqslant m-n.$$

Let A, B, and λ be the matrices of the original GLV system. The p-embedded system is also GLV, and its characteristic matrices are:

$$\tilde{B}=(B|B^{*}_{m\times p}), \quad \tilde{\lambda}=\begin{pmatrix}\lambda\\ O_{p\times 1}\end{pmatrix}, \quad \tilde{A}=\begin{pmatrix}A\cdot E\\ O_{p\times m}\end{pmatrix}, \tag{5}$$

where

$$E=\mathrm{diag}(e_1,\ldots,e_m),\quad e_j=\left(\prod_{k=n+1}^{n+p}\alpha_k^{\tilde{B}_{jk}}\right)^{-1},\quad j=1,\ldots,m. \tag{6}$$

In (5), O denotes a submatrix of null entries, while $B^*_{m\times p}$ has arbitrary real entries appropriately chosen for $\tilde{B}$ to be of maximal rank. The subscripts such as $m\times p$ indicate the size of the corresponding submatrix (B^* in this case); we shall maintain this notation henceforth.

Notice how the previous operation transforms a GLV system from an (r,n,m)-class, with $m>n$, into a GLV system belonging to an $(r,n+p,m)$-class, with $1\leqslant p\leqslant m-n$. The embedded system is topologically equivalent to the original one in the manifold $x_i=\alpha_i$, $i=n+1,\ldots,n+p$.

For fixed p, infinitely many different $(r,n+p,m)$-classes can be reached by means of an embedding, depending on the entries of matrix E. Also, in the particular case $p=m-n$ the target system belongs to an (r,m,m)-class, in which a LV representative can be reached by a QMT of matrix $C=\tilde{B}^{-1}$ (since $\mathrm{rank}(\tilde{B})=m$). Thus, the recasting of the original GLV flow into the LV form is always possible. For this m-dimensional LV system, however, $\mathrm{rank}(\tilde{M}_{\mathrm{LV}})<m$, and is not maximum. Whenever this happens in a GLV system, it indicates the existence of quasimonomial constants of motion, as the following proposition shows:

Proposition 2.4: In a GLV system (3), $\mathrm{rank}(M)=r<n$ if and only if there exist $(n-r)$ functionally independent constants of the motion which are time-independent and have quasimonomial form.

Proof: We can assume, without loss of generality, that the r first rows of M are the linearly independent ones. Then, there exist real constants γ_{ki}, with $i=1,\ldots,r$, and $k=r+1,\ldots,n$, such that:

$$M_{kl}=\sum_{i=1}^{r}\gamma_{ki}M_{il},\quad \forall l=1,\ldots,m+1.$$

From (3), we arrive at:

$$\frac{\dot{x}_k}{x_k}=\sum_{i=1}^{r}\gamma_{ki}\frac{\dot{x}_i}{x_i}.$$

After a simple integration this leads to the set of $(n-r)$ constants of motion:

$$x_k^{-1}\prod_{i=1}^{r}x_i^{\gamma_{ki}}=c_k,$$

where the c_k are real constants given by the initial conditions. The functional independence holds immediately from simple evaluation of the Jacobian. The proof in the opposite sense is straightforward after this. Q.E.D.

Therefore, a p-embedding in the $m>n$ case introduces p quasimonomial constants of motion, which are obviously form-invariant under QMTs. This invariance implies that the quasimonomial constants of motion can always be decoupled from a GLV system by means of an appropriate QMT. When this is done, what we are doing is to reverse the p-embedding procedure, actually. The first step to show this is the following result:

Proposition 2.5: Let A^*, B^* and λ^* be the matrices of a GLV system belonging to an $(r,n+p,m)$-class, where $r\leqslant n$ and $0<p\leqslant m-n$. Then there exists a quasimonomial transformation that leads to an $(n+p)$-dimensional GLV system of matrices

$$\lambda=\begin{pmatrix}\overline{\lambda}\\ O_{p\times 1}\end{pmatrix},\quad A=\begin{pmatrix}\overline{A}\\ O_{p\times m}\end{pmatrix}. \tag{7}$$

Proof: We shall omit it, since it is based on simple matrix algebra properties.

In (7), we have decoupled the final p components of the vector field: Let $x_1,\ldots,x_{n+p}$ be the variables of the system of matrices (7), and let $x_{n+i}(0)=\alpha_{n+i}>0$, $i=1,\ldots,p$, be the initial condi-

tions of the decoupled variables. Let us also write $B=(\overline{B}|\overline{B}'_{m\times p})$ for the matrix of exponents of this system. Then, when we restrict the dynamics to an n-dimensional flow, the result is another GLV system from an (r,n,m)-class, which is characterized by three matrices $\hat{B}$, $\hat{\lambda}$ and $\hat{A}$, given by:

$$\hat{B}=\overline{B}, \quad \hat{\lambda}=\overline{\lambda}, \quad \hat{A}=\overline{A}\cdot E^{-1},$$

where again

$$E=\mathrm{diag}(e_1,\ldots,e_m), \quad e_j=\left(\prod_{k=n+1}^{n+p}\alpha_k^{B_{jk}}\right)^{-1}, \quad j=1,\ldots,m. \tag{8}$$

Definition 2.6: The previous operation transforming a GLV system in an $(r,n+p,m)$-class, with $r\leqslant n$ and $0<p\leqslant m-n$, into a GLV system belonging to an (r,n,m)-class, is called a p-decoupling.

Even for fixed n, there are again infinite possible target classes, due to the arbitrariness in the initial conditions represented by E. In any case, there is obviously a conservation of the topological properties of the flow in the process. It is also clear that, in the especial case in which we choose $n=r$, we have the maximum reduction possible by means of this method; otherwise the decoupling is partial. In either case, the simplification is possible because we are, in fact, restricting the dynamics to the level surfaces of quasimonomial constants of motion.

To summarize, we have a multilevel structure of (r,n,m)-classes of equivalence, with n ranging in the interval $r\leqslant n\leqslant m$. We can transform freely every GLV system inside this scheme by means of the QMTs and the two basic—and opposite—operations: p-embeddings and p-decouplings, which proceed by the introduction of quasimonomial first integrals, or by the restriction of the system dynamics to their level surfaces, respectively.

III. GLV FAMILIES OF POISSON SYSTEMS

We start by characterizing the systems of interest. In what follows, the superscript T denotes the transpose of a matrix.

Theorem 3.1: Let us consider a GLV system of the form (3) such that

$$\lambda=K\cdot L, \quad A=K\cdot B^T\cdot D, \tag{9}$$

with K, L and D matrices of real entries, where K is $n\times n$ and skew-symmetric; L is $n\times 1$; and D is $m\times m$, diagonal and of maximal rank. Then the system has a constant of motion of the form:

$$H=\sum_{i=1}^{m} D_{ii}\prod_{k=1}^{n} x_k^{B_{ik}}+\sum_{j=1}^{n} L_j\,\ln(x_j). \tag{10}$$

Moreover, the system is Poisson with Hamiltonian H.

Proof: The GLV flow complies with the format $\dot{x}=\mathscr{J}\cdot\nabla H$, where the Hamiltonian is smooth in the positive orthant and given by H in (10), while $\mathscr{J}$ is the smooth matrix

$$\mathscr{J}=X\cdot K\cdot X, \quad X=\mathrm{diag}(x_1,\ldots,x_n). \tag{11}$$

That H is the Hamiltonian implies that it is a constant of motion. Q.E.D.

Notice that the first part of the Hamiltonian is associated to the m quasimonomials of the GLV vector field (in fact, it is a linear combination of them), while the logarithmic terms are closely connected to the linear contributions. The observation that a matrix of the form $X\cdot K\cdot X$ is a structure matrix iff $K^T=-K$ is due to Plank.[6] From now on, we shall denote the systems described by Theorem 3.1 as the GLV-Poisson (GLVP).

Proposition 3.2: The Poisson structure of GLVP systems is form-invariant under a QMT. After a QMT of matrix C, the characteristic matrices of the transformed Poisson structure are:

$$K'=C^{-1}\cdot K\cdot (C^{-1})^T, \quad L'=C^T\cdot L, \quad D'=D.$$

In particular, both the Hamiltonian and the structure matrix are form-invariant under QMTs.

Proof: The simplest proof is the algebraic one. After a QMT we have:

$$\lambda' = C^{-1}\cdot K\cdot L = C^{-1}\cdot K\cdot (C^{-1})^T\cdot C^T\cdot L = K'\cdot L'$$

$$A' = C^{-1}\cdot K\cdot B^T\cdot D = K'\cdot (B')^T\cdot D'$$

and $D' = D$. Then, from Theorem 3.1 the new system is also GLVP, and its structure matrix and Hamiltonian are, respectively, $\mathscr{T}' = Y\cdot K'\cdot Y$, and

$$H' = \sum_{i=1}^{m} D'_{ii}\prod_{k=1}^{n} y_k^{B'_{ik}} + \sum_{j=1}^{n} L'_j \ln(y_j).$$

This demonstrates the result. Q.E.D.

Corollary 3.3: The Poisson bracket of a GLVP system is form-invariant under QMTs.

There is an important degree of freedom in the Poisson structure: Let $N\in \mathrm{Ker}\{K\}$. Then, the GLVP system we obtain does not change if we replace L by $L+N$, since from (9) we have $\lambda = K\cdot L = K\cdot (L+N)$. This is, in fact, a source of ambiguity in the Hamiltonian itself, because the flow is unaltered if we add to H an extra term of the form:

$$\phi_N = \sum_{j=1}^{n} N_j \ln(x_j), \quad N\in \mathrm{Ker}\{K\}. \tag{12}$$

This degree of freedom is precisely the one associated to the well-known Casimir functions:

Proposition 3.4: Let $r = \mathrm{rank}(K)$ in a given GLVP structure. There is a complete set of $n-r$ functionally independent Casimir functions of the form (12).

Proof: Evidently, $\mathrm{rank}(\mathscr{T}) = \mathrm{rank}(K) = \text{constant}$ everywhere in the positive orthant. Then there are exactly $n-\mathrm{rank}(K)$ functionally independent Casimirs. If $N\in \mathrm{Ker}\{K\}$ we have $\mathscr{T}\cdot\nabla\phi_N = 0$, and all such ϕ_N are Casimirs. But $\dim(\mathrm{Ker}\{K\}) = n-r$: Then we can get a maximal set of independent Casimirs by choosing $n-r$ linearly independent vectors of $\mathrm{Ker}\{K\}$, i.e., a basis of $\mathrm{Ker}\{K\}$. The functional independence can be readily verified in this case. Q.E.D.

Moreover, under a QMT of matrix C, ϕ_N is form-invariant and changes to $\phi_{N'}$, where $N' = C^T\cdot N$ and $N'\in \mathrm{Ker}\{K'\}$. Thus $\phi_{N'}$ is a Casimir of the transformed system. Let $\{N^{(1)},\ldots,N^{(n-r)}\}$ be a basis of $\mathrm{Ker}\{K\}$, where $r = \mathrm{rank}(K)$ as before. Then $\{\phi_{N^{(1)}},\ldots,\phi_{N^{(n-r)}}\}$ is a complete set of Casimirs of the initial system. After a QMT of matrix C, the Casimirs are transformed into a new family of Casimirs $\{\phi_{N'^{(1)}},\ldots,\phi_{N'^{(n-r)}}\}$, where $N'^{(i)} = C^T\cdot N^{(i)}$ for all i. In fact, the new set $\{N'^{(1)},\ldots,N'^{(n-r)}\}$ is also a basis of $\mathrm{Ker}\{K'\}$. Therefore, every QMT carries a complete set of independent Casimirs of the form (12) into its counterpart for the target system. In other words, the symplectic foliation of a GLVP system is a class property.

We can equivalently express the Casimir functions (12) in quasimonomial form as:

$$\phi_N = \prod_{j=1}^{n} x_j^{N_j}, \quad N\in \mathrm{Ker}\{K\}.$$

However, we have seen in Proposition 2.4 that the quasimonomial first integrals arise in a purely GLV context, independently of the existence of a Poisson structure of the system. We saw that they are associated to a degeneracy in the rank of M. Clearly, there must be a close relationship between quasimonomial Casimirs and quasimonomial first integrals in general. We shall now demonstrate that both sets do coincide:

Theorem 3.5: Every quasimonomial constant of motion of a GLVP system is a Casimir function.

Proof: Notice that a quasimonomial constant of motion can be expressed as:

$$\prod_{j=1}^{n} x_j^{N_j} = \text{constant}, \quad N\in \mathrm{Ker}\{M^T\}.$$

Therefore, our statement will be automatically demonstrated if we show that $\mathrm{Ker}\{M^T\} = \mathrm{Ker}\{K\}$.

First, we demonstrate that $\mathrm{rank}(A) = \mathrm{rank}(M)$ for GLVP systems, where $M = (\lambda | A)$. From (9), we can write symbolically $M = K \cdot (L | B^T \cdot D)$. Expressed in this form, it is immediate to see that $\mathrm{rank}(A) = \mathrm{rank}(M)$ by simple inspection. Consequently, $\mathrm{Ker}\{M^T\} = \mathrm{Ker}\{A^T\}$ for GLVP systems, and the theorem will be demonstrated if $\mathrm{Ker}\{A^T\} = \mathrm{Ker}\{K\}$.

It is evident that $\mathrm{Ker}\{K\} \subset \mathrm{Ker}\{A^T\}$. To show that $\mathrm{Ker}\{A^T\} \subset \mathrm{Ker}\{K\}$, note that $\mathrm{rank}(D \cdot B) = n$, and therefore $\mathrm{Ker}(D \cdot B) = \{0\}$. The result is then straightforward, and the theorem is demonstrated. Q.E.D.

Corollary 3.6: In every GLVP system, $\mathrm{rank}(M) = \mathrm{rank}(A) = \mathrm{rank}(\mathscr{J}) = \mathrm{rank}(K)$.

Theorem 3.5 will be of fundamental importance in what follows. It summarizes very well the interplay between algebraic and Poisson properties, which is present in GLVP systems. The level surfaces of the Casimir functions yield the global structure of the Poisson system—they constitute the symplectic foliation of the phase-space. We are now able to reconstruct this important feature in terms of a very simple and purely algebraic scheme, which is summarized in the proof of Proposition 2.4. More concisely, we can say that the symplectic foliation of the system is condensed in the rank properties of the GLV matrix M. Conversely, purely algebraic properties of the system in the GLV context now assume a completely new role. This is the case of the quasimonomial first integrals, which now become Casimir functions. Moreover, this parallelism is not only valid for a single system, but it is a class property. Obviously, some logical consequences arise from this interplay: For example, $\mathrm{rank}(M)$ may take any value between 0 and n in a general GLV system. However, from Corollary 3.6 this must be an even number for GLVP systems, an evident restriction if the level surfaces of the quasimonomial first integrals are to be a symplectic foliation.

In Sec. II, we elaborated in some detail on the techniques for the manipulation of the quasimonomial constants of motion in a pure GLV context. These procedures involved inter-class transformations which might, if desired, eliminate all such first integrals. These ideas acquire completely new implications in the light of the last results: Such manipulations give now the key for restricting the dynamics to the symplectic leaves or, in the opposite sense, to embed the system in Poisson structures of higher dimensionality. We shall devote the next section to give a systematic treatment of these issues.

IV. TRANSFORMATIONS ON THE SYMPLECTIC FOLIATION

According to Sec. II, three basic procedures are those which allow the manipulation of GLV systems: QMTs, embeddings, and decouplings. In Sec. III, we have seen how QMTs preserve the GLVP structure. That is, we have demonstrated that the GLVP structure is a class property. In this section, our aim will be to show that the inter-class operations do also preserve in an appropriate way the same GLVP character. This will complete the Poisson description of these systems and reinforce the parallelism between algebraic and Poisson properties.

We shall first establish the result for the embeddings:

Proposition 4.1: If a GLVP system which belongs to an (r,n,m)-class, with $m > n$, is subjected to a p-embedding, with $1 \leqslant p \leqslant m - n$, then the resulting system is also GLVP.

Proof: Relations $A = K \cdot B^T \cdot D$ and $\lambda = K \cdot L$ in the original system, imply that $\tilde{A} = \tilde{K} \cdot \tilde{B}^T \cdot \tilde{D}$ and $\tilde{\lambda} = \tilde{K} \cdot \tilde{L}$ in the embedded vector field, where:

$$\tilde{K} = \begin{pmatrix} K & O_{n\times p} \\ O_{p\times n} & O_{p\times p} \end{pmatrix}, \quad \tilde{L} = \begin{pmatrix} L \\ L^*_{p\times 1} \end{pmatrix}, \quad \tilde{D} = D \cdot E.$$

Here, L^* is composed of arbitrary entries, E is given by (6), and $\tilde{A}$, $\tilde{B}$, and $\tilde{\lambda}$ are those in (5). With the help of Theorem 3.1, this proves the result. Q.E.D.

Corollary 4.2: Under the same hypotheses of Proposition 4.1, all the members of the $(r, n+p, m)$-class to which the expanded system belongs, are GLVP systems.

Corollary 4.3: Given a GLVP flow belonging to an (r,n,m)-class, every m-dimensional LV representative of the system is also GLVP.

We shall demonstrate now that we have an analogous situation in the case of the decouplings. For this we need a preliminary result:

Proposition 4.4: Given a GLVP system of matrices M^* and K^*, which belongs to an $(r,n+p,m)$-class, with $r\leqslant n$ and $0<p\leqslant m-n$, then there exists at least one QMT such that the transformed flow has matrices of the form:

$$M=\begin{pmatrix} \overline{M} \\ O_{p\times(m+1)} \end{pmatrix}, \quad K=\begin{pmatrix} \overline{K} & O_{n\times p} \\ O_{p\times n} & O_{p\times p} \end{pmatrix}. \tag{13}$$

Proof: Let C be the matrix of the QMT. From the transformation rule $K=C^{-1}\cdot K^*\cdot (C^{-1})^T$ and the skew-symmetry of K, it is clear that there exists an invertible C which recasts K^* in the desired way. The form of M is a consequence of the form of K, since from (9) we have $\lambda=K\cdot L$ and $A=K\cdot B^T\cdot D$. Q.E.D.

This leads to the main result for reductions:

Proposition 4.5: If a GLVP system belonging to an $(r,n+p,m)$-class, with $r\leqslant n$ and $0<p\leqslant m-n$, is subjected to a p-decoupling, then the resulting system is also GLVP.

Proof: From Proposition 4.4, we can first transform the GLVP system into another one in the same $(r,n+p,m)$-class, with matrices like those in (13). We shall denote the rest of matrices of the latter as A, B, λ, L and D (A and λ being given by Eq. (7)). We shall also, for convenience, express B and L as composed of submatrices in the usual form $B=(\overline{B}|\underline{B}'_{m\times p})$, and $L=(\overline{L}|\underline{L}'_{p\times 1})^T$. As we know from Sec. II, the matrices of the restricted system are $\hat{B}=B$, $\hat{\lambda}=\overline{\lambda}$, and $\hat{A}=A\cdot E^{-1}$, where E is given by (8). It is not difficult to check that the relations $A=K\cdot B^T\cdot D$ and $\lambda=K\cdot L$, imply that $\hat{A}=\hat{K}\cdot\hat{B}^T\cdot\hat{D}$ and $\hat{\lambda}=\hat{K}\cdot\hat{L}$ in the reduced flow, where:

$$\hat{K}=\overline{K}, \quad \hat{L}=\overline{L}, \quad \hat{D}=D\cdot E^{-1}.$$

This proves the result. Q.E.D.

Therefore, not only the GLV format itself, but also the proper GLVP structure is preserved when a reduction is carried out. This is consistent with the fact that quasimonomial constants of motion are Casimirs of the Poisson structure. We can then state:

Corollary 4.6: If a GLV system of an (r,n,m)-class is GLVP, then all the (r,k,m)-classes, for $k=r,\ldots,m$, which can be reached by means of successive embeddings and/or decouplings, are composed of GLVP systems.

Consequently, all the transformations that we have defined within and among the classes preserve both the algebraic GLV properties of the equations and the Poisson format. A decoupling amounts to restricting the system totally or partially to the level surfaces of the Casimirs. An embedding adds new Casimirs to the system, thus increasing its dimensionality. The dynamics on the symplectic leaves remains, on the other hand, always intact. Knowing how to increase the dimension by means of embeddings is important, because it connects every class with the LV format—and vice versa in the case of decouplings. In particular, this tells us how an LV system can be simplified.

In the case in which we perform a maximal decoupling, we obviously arrive at an (r,r,m)-class of GLVP systems, with r even: The members of this class are symplectic systems, since all the Casimirs have been removed. The previous results not only allow, however, the mere transformation of the vector field into symplectic form: They can also be used to reach the full reduction to the canonical forms of Darboux or Hamilton. This is the issue of the next section.

V. REDUCTION TO DARBOUX CANONICAL FORM

We detail here three ways for constructing the Darboux canonical form. We shall first address the more general approach, and then comment on two more specific possibilities.

A. General method

Proposition 5.1: Given a GLVP system belonging to an (r,n,m)-class, there exists a quasimonomial transformation such that for the transformed system:

$$K'=\mathscr{S}(r,n-r)=\mathrm{diag}(S_1,S_2,\ldots,S_{r/2},\overbrace{0,\ldots,0}^{(n-r)}), \tag{14}$$

where

$$S_i = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \forall i = 1,...,r/2.$$

Proof: From Corollary 3.6, $\mathrm{rank}(K) = r$ for the original system. After a QMT of matrix C, we have from Proposition 3.2 that $K' = C^{-1} \cdot K \cdot (C^{-1})^T$, which is a congruence transformation over K. But, since K is skew-symmetric and of rank r, it is congruent[38] to a matrix of the form (14). Then, the QMT exists. Q.E.D.

We can now state the following result:

Theorem 5.2: Every GLVP system belonging to an (r,n,m)-class can be globally reduced to Darboux canonical form inside the positive orthant.

Proof: The proof is constructive. We shall assume that the system has been already transformed in such a way that its matrix K complies to format (14). We can then introduce the following transformation, which is a global orientation-preserving diffeomorphism inside the positive orthant:

$$y_i = \ln(x_i), \quad i = 1,...,n. \tag{15}$$

We now take into account the equation for the transformation of the structure matrix under general diffeomorphisms, $y_i = y_i(x)$:

$$(\mathscr{J}')_{ij}(y) = \sum_{k,l=1}^{n} \frac{\partial y_i}{\partial x_k} \mathscr{J}_{kl}(x) \frac{\partial y_j}{\partial x_l}. \tag{16}$$

The result, upon applying (15), is:

$$(\mathscr{J}')_{ij} = (\mathscr{S}(r,n-r))_{ij}, \quad \forall i,j = 1,...,n.$$

This transforms the original GLVP into a non-GLVP system that conforms, however, to Darboux form: There are obviously $r/2$ pairs of canonically conjugate variables, and $(n-r)$ trivial Casimirs. The transformation of the Hamiltonian is straightforward. This proves the result. Q.E.D.

B. Decoupling method

Notice that in the case $r < n$, i.e., when the GLVP system is not symplectic, we can make use of the reduction procedure of the previous section, instead of applying Theorem 5.2 from the very beginning. The result would be another GLVP system, now symplectic, which belongs to an (r,r,m)-class. Making use of Theorem 5.2 on the reduced flow, would lead to a purely Hamiltonian system, since

$$\mathscr{J}' = \mathscr{S}(r,0),$$

which is the classical symplectic matrix of dimension r.

C. Linear transformation method

Assume that we first subject the initial system of an (r,n,m)-class and matrix K, to transformation (15). From (16), we find:

$$\mathscr{J}' = K.$$

The resulting system $\dot{y} = K \cdot \nabla H'(y)$ is not GLVP, though it is Poisson. A linear change of variables $w = C \cdot y$, where C is an invertible $n \times n$ matrix, leads to another Poisson system of constant structure matrix,

$$\dot{w} = (C \cdot K \cdot C^T) \cdot \nabla H''(w).$$

Finally, there exists a C such that $C \cdot K \cdot C^T = \mathscr{S}(r,n-r)$, and Darboux form is achieved. This procedure has already been applied in the literature to certain symplectic LV systems, of even-dimension and with a single fixed point.[37]

VI. EXAMPLE: 3D LOTKA–VOLTERRA EQUATIONS

A. Poisson structure

As an illustration of the previous results, we shall look upon the 3D LV Poisson structure first characterized by Nutku.[27] The flow is given by the equations:

$$\dot{x}_1 = x_1(\rho + cx_2 + x_3),$$

$$\dot{x}_2 = x_2(\mu + x_1 + ax_3),$$

$$\dot{x}_3 = x_3(\nu + bx_1 + x_2).$$

As Nutku has pointed out, this is a Poisson system if

$$abc = -1, \quad \nu = \mu b - \rho ab.$$

In this case, the structure matrix and the Hamiltonian are, respectively:

$$\mathscr{J} = \begin{pmatrix} 0 & cx_1x_2 & bcx_1x_3 \\ -cx_1x_2 & 0 & -x_2x_3 \\ -bcx_1x_3 & x_2x_3 & 0 \end{pmatrix},$$

obeying to form (11), and

$$H = abx_1 + x_2 - ax_3 + \nu \ln(x_2) - \mu \ln(x_3),$$

which is of the form (10). The system is thus GLVP with characteristic matrices:

$$B = I_{3\times 3}, \quad M = \begin{pmatrix} \rho & 0 & c & 1 \\ \mu & 1 & 0 & a \\ \nu & b & 1 & 0 \end{pmatrix},$$

$$K = \begin{pmatrix} 0 & c & bc \\ -c & 0 & -1 \\ -bc & 1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} ab & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -a \end{pmatrix}, \quad L = \begin{pmatrix} 0 \\ \nu \\ -\mu \end{pmatrix}.$$

Notice that $\mathrm{rank}(M) = \mathrm{rank}(A) = \mathrm{rank}(K) = \mathrm{rank}(\mathscr{J}) = 2$ inside the positive orthant of R^3. There is then one independent Casimir function. By noting that, in M, $\mathrm{row}(3) = (1/c) \times \mathrm{row}(1) + b \times \mathrm{row}(2)$, we immediately find the quasimonomial first integral:

$$x_1^{ab} x_2^{-b} x_3 = \text{constant},$$

which is also a Casimir of the Poisson structure from Theorem 3.5. This way of recovering Casimir functions of the system is certainly more economic than solving the PDE $\mathscr{J} \cdot \nabla \phi = 0$, which is the usual approach.

B. Darboux form: General method

Let us now subject the system to a QMT of matrix:

$$C = \begin{pmatrix} c & 0 & 0 \\ 0 & 1 & 0 \\ 1 & b & -1 \end{pmatrix}.$$

We arrive to a new GLVP of matrices:

$$B'=\begin{pmatrix} c & 0 & 0 \\ 0 & 1 & 0 \\ 1 & b & -1 \end{pmatrix}, \quad M'=\begin{pmatrix} \rho/c & 0 & 1 & 1/c \\ \mu & 1 & 0 & a \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$K'=\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad D'=\begin{pmatrix} ab & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -a \end{pmatrix}, \quad L'=\begin{pmatrix} -\mu \\ \rho/c \\ \mu \end{pmatrix}.$$

The Casimir function has been decoupled, and now is just x_3'. A change of variables $y_i=\ln(x_i')$, $i=1,2,3$ yields Darboux's form, with:

$$\mathscr{J}(y)=\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{17}$$

and

$$H(y)=abe^{cy_1}+e^{y_2}-ae^{y_1+by_2-y_3}-\mu y_1+(\rho/c)y_2+\mu y_3. \tag{18}$$

C. Darboux form: Decoupling method

Although the previous one is the shortest way to achieve the transformation into Darboux form, it may be sometimes more convenient to proceed in a two-step alternative: The first step is the transformation of the system into a symplectic flow. This might be more appropriate in systems of higher dimensions, in which an initial reduction of the dimensionality of the problem may produce the most manageable system. We shall briefly display it for the sake of illustration.

We can first make a QMT of matrix:

$$C_1=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/c & b & -1 \end{pmatrix}.$$

If we then decouple the third variable, assuming for simplicity that its initial condition is $x_3'(0)=1$, the reduced GLVP system is given by:

$$\hat{B}'=\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1/c & b \end{pmatrix}, \quad \hat{M}'=\begin{pmatrix} \rho & 0 & c & 1 \\ \mu & 1 & 0 & a \end{pmatrix},$$

$$\hat{K}'=\begin{pmatrix} 0 & c \\ -c & 0 \end{pmatrix}, \quad \hat{D}'=\begin{pmatrix} ab & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -a \end{pmatrix}, \quad \hat{L}'=\begin{pmatrix} -\mu/c \\ \nu-\mu b \end{pmatrix}.$$

We now perform a second QMT, this time acting on the reduced (2,2,3)-class:

$$C_2=\begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix}.$$

In the resulting flow, we have:

$$\hat{K}''=S(2,0)=\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

After the final change of variables $y_i=\ln(x_i'')$, $i=1,2$, we arrive at a Hamiltonian system in which:

$$H(y)=abe^{cy_1}+e^{y_2}-ae^{y_1+by_2}-\mu y_1+(\rho/c)y_2.$$

Notice that this Hamiltonian can be obtained from (18) with $y_3=0$. This is due to the initial condition we have assumed for x_3' in the 1-decoupling. We obviously retrieve, up to trivial differences in form, the Darboux system (17–18).

VII. FINAL REMARKS

We have seen that there is a close parallelism between the Poisson structure of GLVP flows and the algebraic properties of GLV equations. The deep and unexpected interplay between both aspects of the systems results in an economy in their description. It also establishes an operational framework for their manipulation and simplification. This is, to our knowledge, a novel approach to the treatment of finite-dimensional Poisson structures.

We end this work by giving an evaluation in relation to what has been done previously in the literature. Unfortunately, the only way for doing this is by particularizing the comparison to LV models, for which earlier results are available. We shall only consider previous approaches which are valid for n-dimensional LV models, for arbitrary n. We can then say that most Hamiltonian and Poisson LV systems treated in the literature have a matrix A which is of maximal rank, i.e., they have a single fixed point.[6,30,37] Also, they are often restricted to even dimensionality,[30,37] which usually entails that the system is symplectic (these are, of course, two requirements which are implicit in the classical Hamiltonian studies). None of these restrictions are present, as we have seen, in our models. On the other hand, our treatment joins previous works in what concerns certain requirements on matrix A: For a GLVP Lotka–Volterra system, we have that $A=K\cdot D$, where K is skew-symmetric. This implies, as it can be readily seen, that $D_{ii}A_{ij}=-D_{jj}A_{ji}$, which is exactly the same kind of generalized skew-symmetry which can be found in the works of Kerner[37] and some cases from Plank,[6] for example. Therefore, the scope of our treatment does not differ, in this sense, to that of previous ones. Notice also how our Hamiltonian (10) reduces, in the case of LV systems, to a generalization of the classical Volterra's constant of the motion:

$$H_V=\sum_{i=1}^{n}\beta_i(x_i-p_i\ln(x_i)),$$

where p_i are the coordinates of the (unique) fixed point of Volterra's systems.

Another interesting issue which we would like to comment on here concerns the use of an arbitrary Hamiltonian, while retaining the form (11) for the structure matrix. This leads, of course, to the generation of a wide range of Poisson systems. This procedure can be found, for example, in Plank's work.[6] We may mention that many of our previous results still hold in this Hamiltonian-independent situation. This is the case, for instance, in the reduction of the system to the Darboux form. The reason is that the criterion to decide whether a system complies to Darboux format or not, relies on the form of the structure matrix, exclusively. Consequently, the manipulations to which the system is to be subjected concern the recasting of $\mathscr{T}$ in the desired form, H being irrelevant for that case—which is the situation in Sec. V.

ACKNOWLEDGMENTS

This work has been supported by the DGICYT of Spain (Grant No. PB94-0390) and by the E.U. (Esprit WG 24490). B. H. acknowledges a doctoral fellowship from Comunidad de Madrid.

[1] A. Lichnerowicz, J. Diff. Geom. **12**, 253 (1977); A. Weinstein, *ibid.* **18**, 523 (1983).
[2] P. J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed. (Springer, New York, 1993).
[3] H. Gümral and Y. Nutku, J. Math. Phys. **34**, 5691 (1993); F. Haas and J. Goedert, Phys. Lett. A **199**, 173 (1995).
[4] S. A. Hojman, J. Phys. A **24**, L249 (1991); **29**, 667 (1996).
[5] M. Plank, Nonlinearity **9**, 887 (1996).
[6] M. Plank, J. Math. Phys. **36**, 3520 (1995).
[7] J. Goedert, F. Haas, D. Hua, M. R. Feix, and L. Cairó, J. Phys. A **27**, 6495 (1994).
[8] W. Pauli, Nuovo Cimento **10**, 648 (1953).
[9] J. Gibbons, D. D. Holm, and B. Kupershmidt, Phys. Lett. **90A**, 281 (1982); D. D. Holm and B. A. Kupershmidt, *ibid.* **91**, 425 (1982); D. D. Holm and B. A. Kupershmidt, *ibid.* **93**, 177 (1983); J. E. Marsden, R. Montgomery, P. J. Morrison, and W. B. Thompson, Ann. Phys. **169**, 29 (1986); D. Chinea, J. C. Marrero, and M. de León, J. Phys. A **29**, 6313 (1996).

[10] D. D. Holm, Physica D **17**, 1 (1985); D. Lewis, J. Marsden, R. Montgomery, and T. Ratiu, *ibid.* **18**, 391 (1986).
[11] P. J. Morrison and J. M. Greene, Phys. Rev. Lett. **45**, 790 (1980).
[12] D. D. Holm, Phys. Lett. **114A**, 137 (1986).
[13] J. E. Marsden and A. Weinstein, Physica D **4**, 394 (1982).
[14] R. D. Hazeltine, D. D. Holm, and P. J. Morrison, J. Plasma Phys. **34**, 103 (1985).
[15] D. D. Holm and B. A. Kupershmidt, Physica D **6**, 347 (1983).
[16] I. E. Dzyaloshinskii and G. E. Volovick, Ann. Phys. **125**, 67 (1980).
[17] R. G. Littlejohn, J. Math. Phys. **20**, 2445 (1979); **23**, 742 (1982); J. R. Cary and R. G. Littlejohn, Ann. Phys. **151**, 1 (1983).
[18] H. D. I. Abarbanel, D. D. Holm, J. E. Marsden, and T. Ratiu, Phys. Rev. Lett. **52**, 2352 (1984); D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, Phys. Rep. **123**, 1 (1985); D. Lewis, J. Marsden, and T. Ratiu, J. Math. Phys. **28**, 2508 (1987); D. D. Holm and K. B. Wolf, Physica D **51**, 189 (1991).
[19] J. C. Simo, T. A. Posbergh, and J. E. Marsden, Phys. Rep. **193**, 279 (1990); J. C. Simo, D. Lewis, and J. E. Marsden, Arch. Ration. Mech. Anal. **115**, 15 (1991); J. C. Simo, T. A. Posbergh, and J. E. Marsden, *ibid.* **115**, 61 (1991).
[20] D. David, D. D. Holm, and M. V. Tratnik, Phys. Lett. A **137**, 355 (1989); **138**, 29 (1989); Phys. Rep. **187**, 281 (1990).
[21] F. Magri, J. Math. Phys. **19**, 1156 (1978); P. J. Olver, Phys. Lett. A **148**, 177 (1990).
[22] D. David and D. D. Holm, J. Nonlinear Sci. **2**, 241 (1992).
[23] P. J. Morrison and R. D. Hazeltine, Phys. Fluids **27**, 886 (1984).
[24] R. G. Littlejohn, AIP Conf. Proc. **88**, 47 (1982).
[25] L. Cairó and M. R. Feix, J. Phys. A **25**, L1287 (1992).
[26] M. Razavy and F. J. Kennedy, Can. J. Phys. **52**, 1532 (1974); C. A. Lucey and E. T. Newman, J. Math. Phys. **29**, 2430 (1988); Y. Nutku, J. Phys. A **23**, L1145 (1990).
[27] Y. Nutku, Phys. Lett. A **145**, 27 (1990).
[28] B. Hernández-Bermejo and V. Fairén, Phys. Lett. A **234**, 35 (1997).
[29] A. J. Lotka, *Elements of Mathematical Biology* (Dover, New York, 1956).
[30] V. Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie* (Gauthier Villars, Paris, 1931).
[31] E. H. Kerner, Bull. Math. Biophys. **26**, 151 (1964); E. H. Kerner, in *Advances in Chemical Physics*, edited by I. Prigogine and S. A. Rice (Wiley, New York, 1971), Vol. 19; N. S. Goel, S. C. Maitra, and E. W. Montroll, Rev. Mod. Phys. **43**, 231 (1971); E. H. Kerner, *Gibbs Ensemble: Biological Ensemble* (Gordon and Breach, New York, 1972).
[32] L. Cairó and M. R. Feix, J. Math. Phys. **33**, 2440 (1992).
[33] M. Peschel and W. Mende, *The Predator-Prey Model. Do we live in a Volterra world?* (Springer, Vienna, 1986).
[34] L. Brenig, Phys. Lett. A **133**, 378 (1988); L. Brenig and A. Goriely, Phys. Rev. A **40**, 4119 (1989); A. Br'uno, *Local Methods in the Theory of Differential Equations* (Springer, New York, 1989); J. L. Gouzé, *Transformation of Polynomial Differential Systems in the Positive Orthant* (Rapport INRIA No. 1308, Sophia-Antipolis, 06561 Valbonne Cedex, France, 1990).
[35] A. Goriely and L. Brenig, Phys. Lett. A **145**, 245 (1990); A. Goriely, J. Math. Phys. **33**, 2728 (1992); B. Hernández-Bermejo and V. Fairén, Phys. Lett. A **206**, 31 (1995); V. Fairén and B. Hernández-Bermejo, J. Phys. Chem. **100**, 19023 (1996); B. Hernández-Bermejo and V. Fairén, Math. Biosci. **140**, 1 (1997).
[36] L. Brenig and A. Goriely, in *Computer Algebra and Differential Equations*, edited by E. Tournier (Cambridge University Press, Cambridge, England, 1994).
[37] E. H. Kerner, Phys. Lett. A **151**, 401 (1990); J. Math. Phys. **38**, 1218 (1997).
[38] F. Ayres Jr., *Schaum's Outline of Matrices* (McGraw–Hill, New York, 1962).

Conserved operators of the generalized elliptic Ruijsenaars models

Yasushi Komori[a] and Kazuhiro Hikami[b]
Department of Physics, Graduate School of Science, University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo 113, Japan

(Received 2 July 1997; accepted for publication 1 June 1998)

We construct the integrable families of the generalized elliptic Ruijsenaars models by use of the Yang–Baxter equation and the reflection equation. The higher order conserved operators of these models are explicitly presented. © *1998 American Institute of Physics.* [S0022-2488(98)00811-1]

I. INTRODUCTION

Since a relativistic analog of the Calogero–Sutherland–Moser (CSM) model was first introduced by Ruijsenaars,[1] this model has been studied extensively in both classical and quantum cases.[2–6] In his studies Ruijsenaars defined the ''relativistic'' Hamiltonian $\mathscr{H}_A$ as

$$\mathscr{H}_A=\sum_{j=1}^{N}\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sqrt{\frac{\vartheta_1(z_{jk}-\mu)}{\vartheta_1(z_{jk})}}\right)\exp\left(\beta\,\frac{\partial}{\partial z_j}\right)\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sqrt{\frac{\vartheta_1(z_{kj}-\mu)}{\vartheta_1(z_{kj})}}\right)$$
$$+\sum_{j=1}^{N}\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sqrt{\frac{\vartheta_1(z_{kj}-\mu)}{\vartheta_1(z_{kj})}}\right)\exp\left(-\beta\,\frac{\partial}{\partial z_j}\right)\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sqrt{\frac{\vartheta_1(z_{jk}-\mu)}{\vartheta_1(z_{jk})}}\right). \tag{1.1}$$

Here z_{jk} denotes z_j-z_k, and β and μ are arbitrary parameters. The function $\vartheta_r(z)$ is the Jacobi theta function (see the Appendix for detail). One sees that the Hamiltonian $\mathscr{H}_A$ is invariant under permutations of variables $z_j\leftrightarrow z_k$, and is called the *A*-type Ruijsenaars model. As a generalization of the *A*-type model, also considered is a *D*-type analog of the Ruijsenaars model,[6–8] i.e., a model invariant not only under permutations $z_j\leftrightarrow z_k$ but sign flips of variables $z_j\leftrightarrow -z_j$. The *D*-type Hamiltonian $\mathscr{H}_D$ reads as

$$\mathscr{H}_D=\sum_{j=1}^{N}\Psi_j(z)^{1/2}\exp\left(2\beta\,\frac{\partial}{\partial z_j}\right)\Psi_j(-z)^{1/2}$$
$$+\sum_{j=1}^{N}\Psi_j(-z)^{1/2}\exp\left(-2\beta\,\frac{\partial}{\partial z_j}\right)\Psi_j(z)^{1/2}+\Psi_0(z), \tag{1.2}$$

where functions $\Psi_j(z)$ and $\Psi_0(z)$ are

$$\Psi_j(z)=\left(\prod_{\substack{k=1\\k\neq j}}^{N}\frac{\vartheta_1(z_{jk}-\mu)}{\vartheta_1(z_{jk})}\,\frac{\vartheta_1(z_j+z_k-\mu)}{\vartheta_1(z_j+z_k)}\right)\left(\prod_{r=0}^{3}\frac{\vartheta_{r+1}(z_j-\nu_r)}{\vartheta_{r+1}(z_j)}\,\frac{\vartheta_{r+1}(z_j+\beta-\overline{\nu_r})}{\vartheta_{r+1}(z_j+\beta)}\right), \tag{1.3a}$$

[a]Electronic mail: komori@monet.phys.s.u-tokyo.ac.jp
[b]Electronic mail: hikami@phys.s.u-tokyo.ac.jp

 © 1998 American Institute of Physics

$$\Psi_0(z)=-\sum_{p=0}^{3}\left(\frac{\pi}{\vartheta_1'(0)}\right)^2\frac{2}{\vartheta_1(-\mu)\vartheta_1(-2\beta-\mu)}\left(\prod_{r=0}^{3}\vartheta_{r+1}(-\beta-\nu_{\pi_p r})\vartheta_{r+1}(-\overline{\nu}_{\pi_p r})\right)$$
$$\times\left(\prod_{j=1}^{N}\frac{\vartheta_{p+1}(z_j-\beta-\mu)}{\vartheta_{p+1}(z_j-\beta)}\frac{\vartheta_{p+1}(-z_j-\beta-\mu)}{\vartheta_{p+1}(-z_j-\beta)}\right). \tag{1.3b}$$

Here π_p for $p=0,1,2,3$ denote permutations; $\pi_0=\mathbb{1}$, $\pi_1=(01)(23)$, $\pi_2=(02)(13)$, and $\pi_3=(03)(12)$. Parameters β, μ, ν_r and $\overline{\nu}_r$ $(r=0,1,2,3)$ are arbitrary constants. We note that the *A*-type model and the *D*-type model contain two and ten parameters, respectively. Mathematically the trigonometric limits of these operators are related to the Macdonald operators (or, for the case of *D* type, the Macdonald–Koornwinder operator), and their eigenfunctions are much investigated in *q*-orthogonal polynomial theory.[9–11] There the affine Hecke algebra plays an important role.[12–16]

In this paper, we define two families of commutative difference operators in terms of solutions of both the Yang–Baxter equation (YBE) and the reflection equation (RE). Here we regard these solutions as operators acting on function spaces. While these elliptic difference operators were constructed in the form of an infinite sum or product of trigonometric solution,[17,18] our method deals only with finite products since the solutions of the YBE and the RE are elliptic. We prove that the difference operators possess the Weyl group invariance. This fact follows from the quasi-classical properties of the YBE and the RE. Owing to these symmetries, we present the explicit forms of all the conserved operators for both *A*-type and *D*-type models.

This paper is organized as follows. In Sec. II, we briefly review the operator-valued solutions of the YBE and the RE. We also mention characteristic properties used in later discussions. In Sec. III, we define a family of commutative difference operators in terms of the solution of the YBE. These operators are supposed to act on symmetric space in coordinates. When we assign a proper figure to each operator, the symmetry of the difference operators becomes manifest. We see that the higher conserved operators are explicitly calculated and reduce to the elliptic Macdonald operators. We show in Sec. IV that a similar discussion can be applied to the *D*-type model. As in the previous sections, we construct a set of difference operators using solutions of the YBE and the RE. The symmetry is also clarified by a pictorial analysis. Though the calculation is not straightforward and needs several steps in contrast to the *A*-type, we obtain the explicit forms of the *D*-type operators. Section V is devoted to the conclusion.

II. YANG–BAXTER EQUATION AND THE REFLECTION EQUATION

We shall review operator-valued solutions[19–23] of the YBE and the RE.[25,26] The YBE and the RE are, respectively, defined by

$$R^{12}(\xi_{12})R^{13}(\xi_{13})R^{23}(\xi_{23})=R^{23}(\xi_{23})R^{13}(\xi_{13})R^{12}(\xi_{12}), \tag{2.1}$$

$$R^{12}(\xi_{12})\overset{1}{K}(\xi_1)R^{21}(\xi_1+\xi_2)\overset{2}{K}(\xi_2)=\overset{2}{K}(\xi_2)R^{12}(\xi_1+\xi_2)\overset{1}{K}(\xi_1)R^{21}(\xi_{12}). \tag{2.2}$$

Here ξ_j is called the spectral parameter, and ξ_{jk} denotes $\xi_{jk}=\xi_j-\xi_k$. The operator $R^{jk}(\xi)$ acts nontrivially on the jth and the kth variables, and $\overset{j}{K}(\xi)$ acts as $K(\xi)$ only on the jth variable.

We regard solutions of both the YBE and RE as operators acting on the functional space of z_j. We use operators $\hat{s}_{jk}$ and $\hat{t}_j$ as a permutation and a sign flip of variables, respectively;

$$z_j\hat{s}_{jk}=\hat{s}_{jk}z_k, \quad z_j\hat{t}_j=-\hat{t}_jz_j. \tag{2.3}$$

As shown in Refs. 4, 23, and 24, we have the following theorems:

Theorem 2.1: *The elliptic R-operator, defined by*

$$R^{jk}(\xi)=\sigma_\mu(z_{jk})-\sigma_\xi(z_{jk})\cdot\hat{s}_{jk}, \tag{2.4}$$

with arbitrary constant μ, satisfies the YBE (2.1). This R-operator possesses the unitarity,

$$R^{12}(\xi)R^{21}(-\xi)=c(\xi)\cdot\mathbb{1}, \tag{2.5}$$

where function $c(\xi)$ *is given by* $c(\xi)=\wp(\mu)-\wp(\xi)$.

Theorem 2.2: *With the elliptic R-operator (2.4), the elliptic K-operator defined by*

$$K(\xi)=\sum_{r=0}^{3} g_r\sigma^r_{2\xi}(z)-\sum_{r=0}^{3} g_r\sigma^r_{2\nu}(z)\cdot\hat{t}, \tag{2.6}$$

satisfies the RE (2.2). Here parameters ν, g_r $(r=0,1,2,3)$ *are arbitrary. This K-operator (2.6) has the unitarity,*

$$K(\xi)K(-\xi)=d(\xi)\cdot\mathbb{1}, \tag{2.7}$$

where function $d(\xi)$ *is given by*

$$d(\xi)=\mathbf{g}^{\mathrm{T}}\mathbf{S}\mathbf{\Lambda}(\xi)\mathbf{S}\mathbf{g},$$

$$\mathbf{S}=\frac{1}{2}\begin{pmatrix}1&1&1&1\\1&1&-1&-1\\1&-1&1&-1\\1&-1&-1&1\end{pmatrix},\quad \mathbf{g}=\begin{pmatrix}g_0\\g_1\\g_2\\g_3\end{pmatrix},$$

$$\mathbf{\Lambda}(\xi)=\mathrm{diag}_{r=0,1,2,3}(\wp(\nu+\omega_r)-\wp(\xi+\omega_r)).$$

We remark that the *R*- and *K*-operators satisfy the quasiclassical conditions as follows;

$$R^{12}(\mu)=\sigma_\mu(z_{12})(1-\hat{s}_{12}), \tag{2.8a}$$

$$R^{12}(-\mu)=(1+\hat{s}_{12})\sigma_\mu(z_{12}), \tag{2.8b}$$

$$K(\nu)=\left(\sum_{r=0}^{3} g_r\sigma^r_{2\nu}(z)\right)(1-\hat{t}), \tag{2.8c}$$

$$K(-\nu)=(1+\hat{t})\left(\sum_{r=0}^{3} g_r\sigma^r_{-2\nu}(z)\right). \tag{2.8d}$$

It should also be noted that (2.5) and (2.7) include identities

$$R^{12}(\mu)R^{21}(-\mu)=0, \tag{2.9a}$$

$$K(\nu)K(-\nu)=0. \tag{2.9b}$$

III. ELLIPTIC RUIJSENAARS MODEL OF TYPE-*A*

We shall study the elliptic Ruijsenaars model $\mathscr{H}_A$ (1.1) of type-*A*. We begin with a family of difference operators called the elliptic Macdonald operators,

$$\mathscr{M}_n=\sum_{\substack{J\subset\{1,\ldots,N\}\\|J|=n}}\left(\prod_{\substack{j\in J\\k\in J^c}}\sigma_\mu(z_{jk})\right)\prod_{j\in J}T_j(\beta), \tag{3.1}$$

where $T_j(\beta)$ is a shift operator and is defined by $T_j(\beta)=\exp(-\beta\partial/\partial z_j)$. Throughout this section, we denote $T_j(\beta)$ as T_j. The Macdonald operators are supposed to act on a symmetric space in coordinates $z_1,z_2,\ldots,z_N$, i.e., an invariant space under the action of the Weyl group $W_A\simeq S_N$. Thus hereafter we restrict our discussions to such a space. These operators are connected with the Hamiltonian (1.1) of the elliptic Ruijsenaars model of type-*A* by a gauge transformation.[6]

In our previous paper,[6] we proved that the Ruijsenaars model, or the elliptic Macdonald operator $\mathscr{M}_n$ (3.1), is integrable, and that the operators $\mathscr{M}_n$ are related to the elliptic R-operator (2.4) as follows.

(1) the difference operators $\mathscr{D}_n(\xi)=\mathscr{D}_n(\xi_1,\ldots,\xi_N)$ defined by

$$\mathscr{D}_n(\xi)=\mathscr{R}^N_{N-n+1}(\xi)\prod_{m=N-n+1}^{N} T_m, \tag{3.2}$$

commute each other for arbitrary ξ,

$$[\mathscr{D}_n(\xi),\mathscr{D}_m(\xi)]=0. \tag{3.3}$$

Here the operator $\mathscr{R}^n_m(\xi)$ corresponds to a fused R-operator[4] given by

$$\mathscr{R}^n_m(\xi)=\prod_{j=m}^{n}\left(\prod_{k=1}^{\overset{\frown}{m-1}} R^{jk}(\xi_{jk})\right). \tag{3.4}$$

(2) The difference operator $\mathscr{D}_n(\xi)$ reduces to the elliptic Macdonald operator with a specific choice of the spectral parameters,

$$\mathscr{M}_n=\mathscr{D}_n(\xi)|_{\xi_{jk}=(k-j)\mu}, \tag{3.5}$$

where the spectral parameters are set to be $\xi_{jk}=(k-j)\mu$.

Though we have proved the equality (3.5) after a lengthy computation,[6] we shall take another simple method. To clarify the structure of our difference operators, we assign a figure for each operator. The R-operator and the shift operator T can be represented pictorially as

$R^{jk}(\xi)$ =

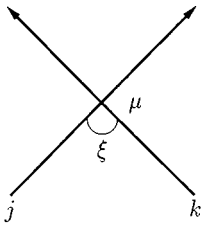

$T_j(\beta)$ = j

(3.6)

Using these figures, the difference operators $\mathscr{D}_n(\xi)$ are depicted as follows:

$\mathscr{D}_n(\xi)$ =

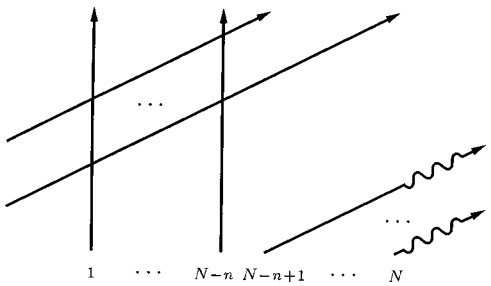

(3.7)

Note that both ends of the figures are supposed to be connected.

Theorem 3.1: *The difference operators $\mathscr{D}_n(\xi)$ preserve the W_A-invariant space when the spectral parameters are set to be $\xi_{jk}=(k-j)\mu$.*

Proof: It is easy to see that the elliptic R-operator (2.4) commutes with T_jT_k;

$$R^{jk}(\xi)T_jT_k=T_jT_kR^{jk}(\xi), \tag{3.8}$$

since the R-operator (2.4) depends only on the difference of variables. We depict some relations as follows:

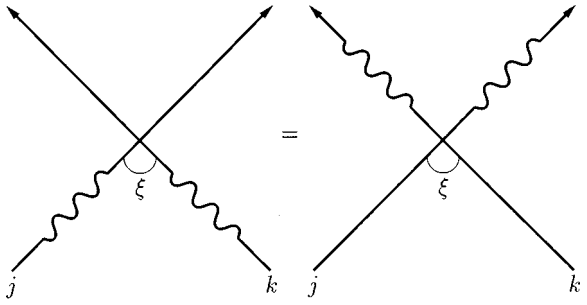

$$R^{jk}(\xi)\,T_j\,T_k \quad = \quad T_j\,T_k\,R^{jk}(\xi) \tag{3.8$'$}$$

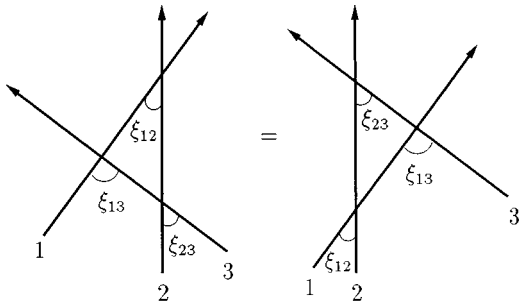

$$R^{12}(\xi_{12})R^{13}(\xi_{13})R^{23}(\xi_{23}) \;=\; R^{23}(\xi_{23})R^{13}(\xi_{13})R^{12}(\xi_{12}) \tag{2.1$'$}$$

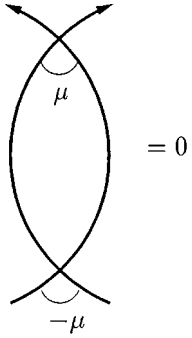

$$R^{jk}(\mu)R^{kj}(-\mu) = 0 \tag{2.9a$'$}$$

We also note that, by the quasiclassical condition (2.8), the operator $R^{jk}(\mu)$ vanishes when it acts on symmetric space in z_j, z_k;

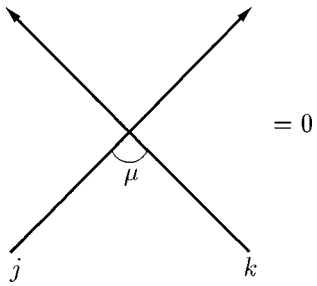

(3.9)

To prove the statement of Theorem 3.1, we only need to check it for the generators of W_A, i.e.,

$$R_{ii+1}(\xi_{ii+1})\mathscr{D}_n(\xi)\big|_{\xi_{jk}=(k-j)\mu}=\sigma_\mu(z_{ii+1})(1-\hat{s}_{ii+1})\mathscr{D}_n(\xi)\big|_{\xi_{jk}=(k-j)\mu}=0. \tag{3.10}$$

This equation can be shown as follows.

(1) $1\leqslant i\leqslant N-n-1$: By using the YBE (2.1′), one finds that the left-hand side of (3.11) is equivalent to the right-hand side of (3.11), which vanishes when it acts on the W_A-invariant space due to (3.9).

(2) $i=N-n$: (3.12) vanishes immediately due to (2.9a′).

(3) $N-n+1\leqslant i\leqslant N-1$: By use of the YBE (2.1′) and the relation (3.8′), one finds that the left-hand side of (3.13) is equivalent to the right-hand side of (3.13), which vanishes if it acts on the W_A-invariant space due to (3.9),

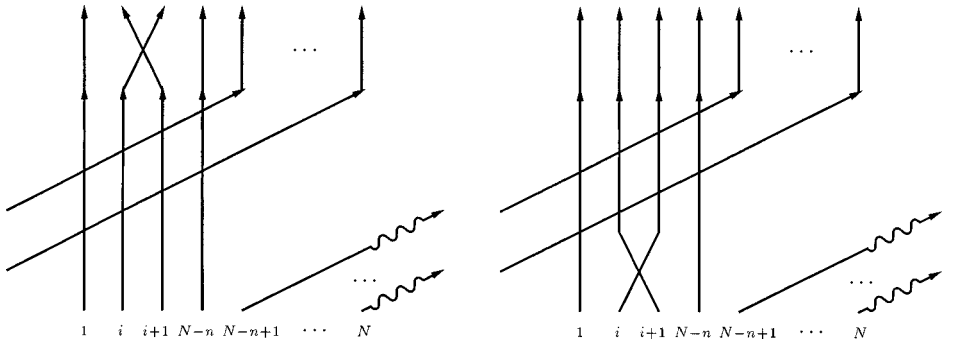

(3.11)

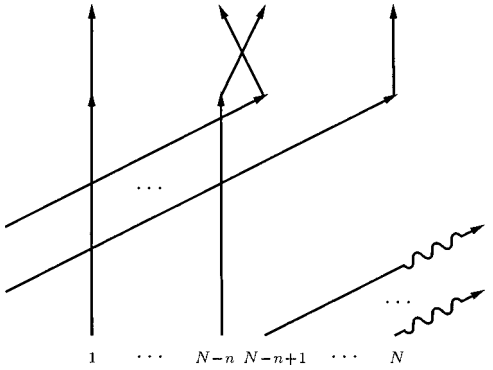

(3.12)

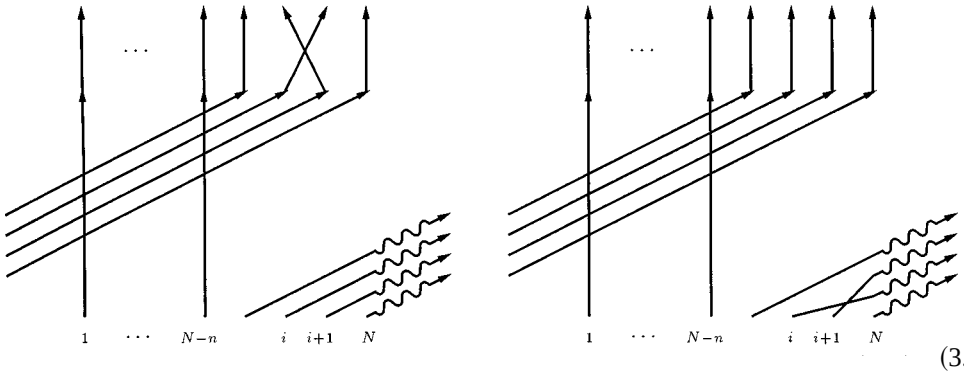

(3.13)

□

Now it is straightforward to calculate the explicit forms of the difference operators $\mathscr{D}_n(\xi)$ with $\xi_{jk}=(k-j)\mu$.

Theorem 3.2: *We have*

$$\mathscr{D}_n(\xi)\big|_{\xi_{jk}=(k-j)\mu}=\mathscr{M}_n, \tag{3.14}$$

where operators $\mathscr{M}_n$ are the elliptic Macdonald operators (3.1).

Proof: When we expand the operators $\mathscr{D}_n(\xi)$ by substituting the elliptic R-operator (2.4) into (3.2), we see that the term $\Pi_{j=N-n+1}^{N} T_j$ is computed simply as

$$\prod_{j=N-n+1}^{N}\left(\prod_{k=1}^{N-n}\sigma_\mu(z_{jk})\right)\prod_{j=N-n+1}^{N} T_j. \tag{3.15}$$

Due to the W_A-invariance of the operators $\mathscr{D}_n(\xi)$ with $\xi_{jk}=(k-j)\mu$, we obtain the result by symmetrizing (3.15). □

IV. ELLIPTIC RUIJSENAARS MODEL OF TYPE-*D*

We introduced the *D*-type Ruijsenaars operators $\mathscr{W}_n$,[6] which form a commutative family of difference operators. The explicit form of the lowest order operator is given by

$$\begin{aligned}\mathscr{W}_1=&\sum_{j=1}^{N}\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sigma_\mu(z_{jk})\sigma_\mu(z_j+z_k)\right)\left(\sum_{r=0}^{3} g_r\sigma_{2\nu}^{r}(z_j)\right)\\&\times\left(\sum_{r=0}^{3}\overline{g}_r\sigma_{2\nu}^{r}(z_j+\beta)T_j(2\beta)-\sum_{r=0}^{3}\overline{g}_r\sigma^{r}_{-2(N-1)\mu-2\nu}(z_j+\beta)\right)\\&+\sum_{j=1}^{N}\left(\prod_{\substack{k=1\\k\neq j}}^{N}\sigma_\mu(z_{kj})\sigma_\mu(-z_j-z_k)\right)\left(\sum_{r=0}^{3} g_r\sigma_{2\nu}^{r}(-z_j)\right)\\&\times\left(\sum_{r=0}^{3}\overline{g}_r\sigma_{2\nu}^{r}(-z_j+\beta)T_j(-2\beta)-\sum_{r=0}^{3}\overline{g}_r\sigma^{r}_{-2(N-1)\mu-2\nu}(-z_j+\beta)\right).\end{aligned} \tag{4.1}$$

The operators $\mathscr{W}_n$ are supposed to act on a symmetric space under permutations and sign flips, i.e., the Weyl group $W_B\simeq S_N\ltimes\{\mathbb{Z}_2\}^N$. We impose such restriction on the space in what follows.

In this section, we shall give the explicit forms of the higher order operators $\mathscr{W}_n$ by a similar technique discussed for the *A*-type operator. To define the operators $\mathscr{W}_n$, we introduce a set of integrable difference operators $\mathscr{Y}_n(\xi)$ for $\xi=\{\xi_1,\ldots,\xi_N\}$ as[6]

$$\mathscr{Y}_n(\xi)=\mathscr{R}^{N}_{N-n+1}(\xi)\mathscr{K}^{N}_{N-n+1}(\xi)\overline{\mathscr{R}}^{N}_{N-n+1}(\xi)\overline{\mathscr{K}}^{N}_{N-n+1}(\xi), \tag{4.2}$$

where $\mathscr{R}^{n}_{m}(\xi)$ is the same as (3.4), and the others are given by

$$\overline{\mathscr{R}}^{n}_{m}(\xi)=\prod_{j=m}^{n}\left(\prod_{k=1}^{m-1}R^{kj}(\xi_k+\xi_j)\right), \tag{4.3}$$

$$\mathscr{K}^{n}_{m}(\xi)=\prod_{j=m}^{n}{}^{j}K(\xi_j)\left(\prod_{k=j+1}^{n}R^{kj}(\xi_k+\xi_j)\right), \tag{4.4}$$

$$\overline{\mathscr{K}}^{n}_{m}(\xi)=\prod_{j=m}^{n}\left(\prod_{k=m}^{j-1}R^{kj}(\xi_k+\xi_j)\right)\overline{{}^{j}K}(\xi_j). \tag{4.5}$$

In this expression, we have used a ''conjugate'' boundary operator $\overline{K}(\xi)$ as

$$\overline{K}(\xi)=T(-\beta)\cdot\hat{t}\cdot K(\xi)\cdot\hat{t}\cdot T(\beta), \tag{4.6}$$

where parameters in the elliptic K-operator (2.6) are replaced as $g_r \to \overline{g_r}$ and $\underline{\nu} \to \overline{\nu}$. As the boundary K-operator (2.6) satisfies the RE (2.2), the conjugate boundary operator $\underline{K}(\xi)$ defined in (4.6) is proved to satisfy the ''conjugate'' RE;

$$\underline{\overset{2}{K}}(\xi_2)R^{21}(\xi_1+\xi_2)\underline{\overset{1}{K}}(\xi_1)R^{12}(\xi_{12}) = R^{21}(\xi_{12})\underline{\overset{1}{K}}(\xi_1)R^{12}(\xi_1+\xi_2)\underline{\overset{2}{K}}(\xi_2). \tag{4.7}$$

It should be noted that (3.4) and (4.3)–(4.5) are closely related to the fusion procedure and they also satisfy the YBE and the RE.

The operators K and $\underline{K}$ are, respectively, depicted as follows:

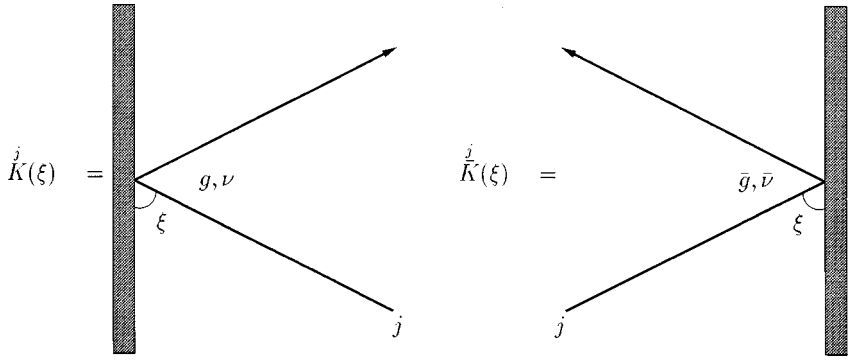

(4.8)

Using the above interpretations, the difference operators $\mathscr{Y}_n(\xi)$ defined in (4.2) can be depicted as

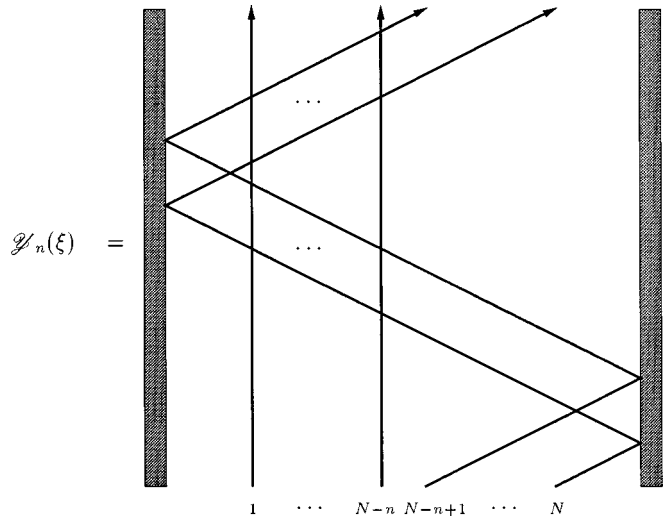

(4.9)

It should be noted that the difference operators $\mathscr{Y}_n(\xi)$ are proved[6] to constitute a commutative family for arbitrary spectral parameters, by applying the YBE and the RE.

In terms of the integrable operators $\mathscr{Y}_n(\xi)$, we define[6] the D-type Ruijsenaars operators $\mathscr{W}_n$ by fixing the spectral parameters as

$$\mathscr{W}_n = \mathscr{Y}_n(\xi)|_{\xi_j = -\nu-(j-1)\mu}. \tag{4.10}$$

The meaning of this definition is clarified by the following theorem.

Theorem 4.1: *The difference operators $\mathscr{Y}_n(\xi)$ preserve the W_B-invariant space when each spectral parameter is set to be $\xi_j = -\nu-(j-1)\mu$.*

Proof: We apply the similar discussion in Theorem 3.1 to prove this theorem. Note that the RE (2.2), the conjugate RE (4.7), and the unitarity condition (2.9b) can be graphically interpreted as follows:

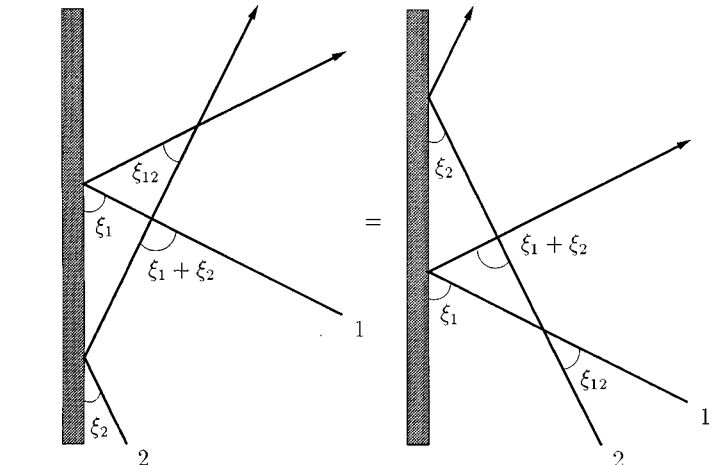

$$R^{12}(\xi_{12})\overset{1}{K}(\xi_1)R^{21}(\xi_1+\xi_2)\overset{2}{K}(\xi_2) = \overset{2}{K}(\xi_2)R^{12}(\xi_1+\xi_2)\overset{1}{K}(\xi_1)R^{21}(\xi_{12}) \tag{2.2′}$$

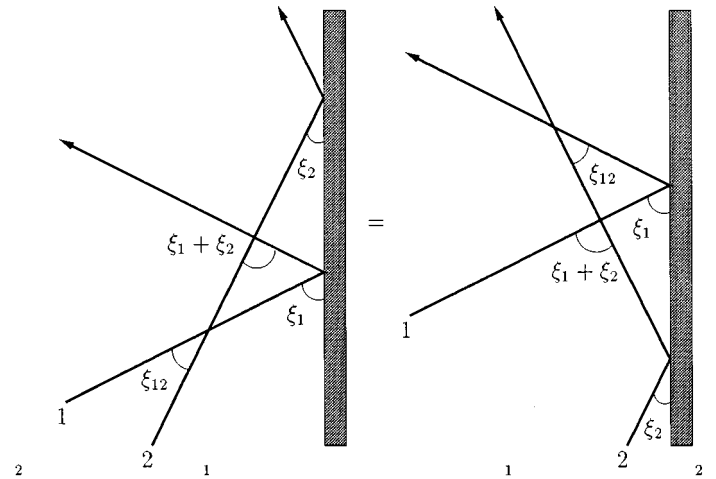

$$\overset{2}{K}(\xi_2)R^{21}(\xi_1+\xi_2)\overset{1}{K}(\xi_1)R^{12}(\xi_{12}) = R^{21}(\xi_{12})\overset{1}{K}(\xi_1)R^{12}(\xi_1+\xi_2)\overset{2}{K}(\xi_2) \tag{4.7′}$$

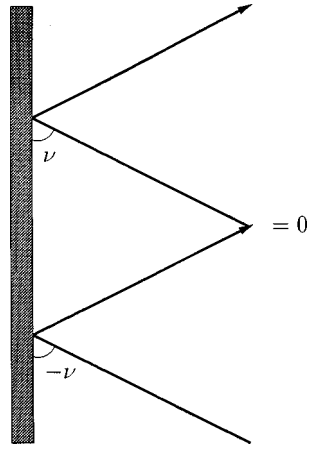

(2.9b′)

We also depict the quasiclassical condition of the boundary operator $K(\nu)$ acting on a invariant space in a sign flip $z_j \leftrightarrow -z_j$;

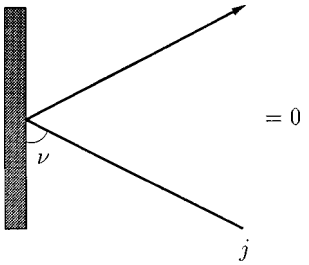

(4.11)

In the following, we fix the spectral parameters ξ as $\xi_j=-\nu-(j-1)\mu$. By noting that the Weyl group W_B is generated by $\hat{t}_1$ and $\hat{s}_{ii+1}$, it is sufficient to show that the following equalities hold:

$$R_{ii+1}(\xi_{ii+1})\mathscr{Y}_n(\xi)=\sigma_\mu(z_{ii+1})(1-\hat{s}_{ii+1})\mathscr{Y}_n(\xi)=0, \tag{4.12}$$

$$\overset{1}{K}(-\xi_1)\mathscr{Y}_n(\xi)=\left(\sum_{r=0}^{3} g_r\sigma^r_{2\nu}(z)\right)(1-\hat{t}_1)\mathscr{Y}_n(\xi)=0. \tag{4.13}$$

These equations are proved pictorially. Equation (4.12) is proved as

(1) $1\leqslant i\leqslant N-n-1$: By using the YBE (2.1′), one finds that the left-hand side of (4.14) is equivalent to the right-hand side of (4.14), which vanishes when it acts on the W_B-invariant space due to (3.9).

(2) $i=N-n$: Equation (4.15) vanishes immediately due to (2.9a′).

(3) $N-n+1\leqslant i\leqslant N-1$: By use of the YBE (2.1′), the RE (2.2′), and the conjugate RE (4.7′), one finds that the left-hand side of (4.16) is equivalent to the right-hand side of (4.16), which vanishes if it acts on the W_B-invariant space due to (3.9).

Equation (4.13) is proved in the same way as

(1) $n\neq N$: By use of the YBE (2.1′) and the RE (2.2′), one finds that the left-hand side of (4.17) is equivalent to the right-hand side of (4.17), which vanishes if it acts on the W_B-invariant space due to (4.11).

(2) $n=N$: Equation (4.18) vanishes immediately due to (2.9b′).

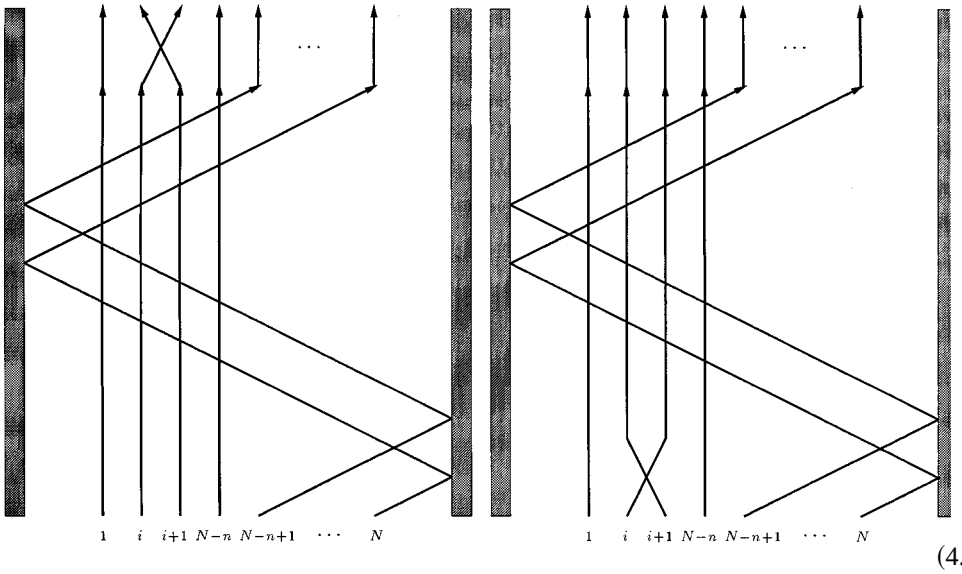

(4.14)

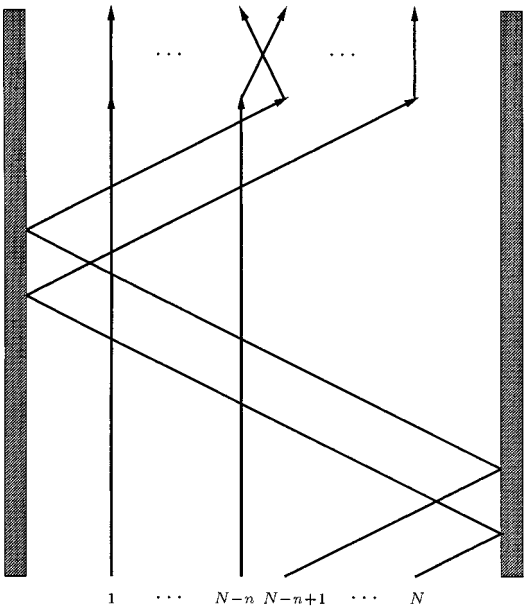

(4.15)

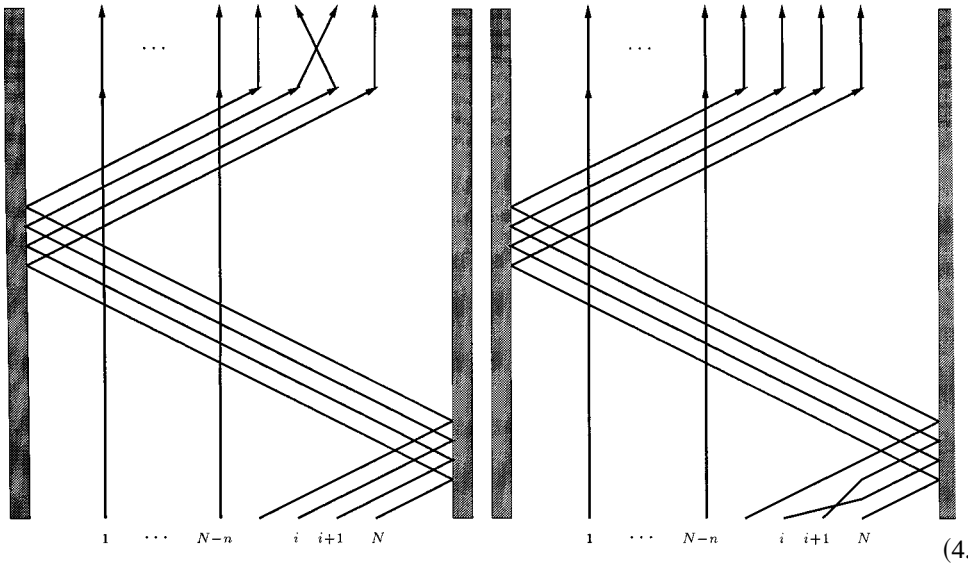

(4.16)

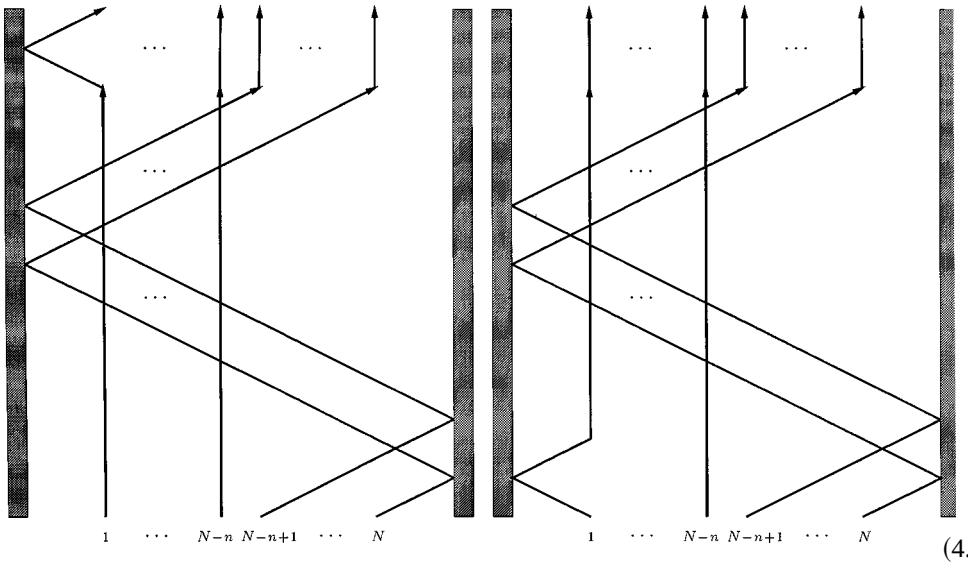

(4.17)

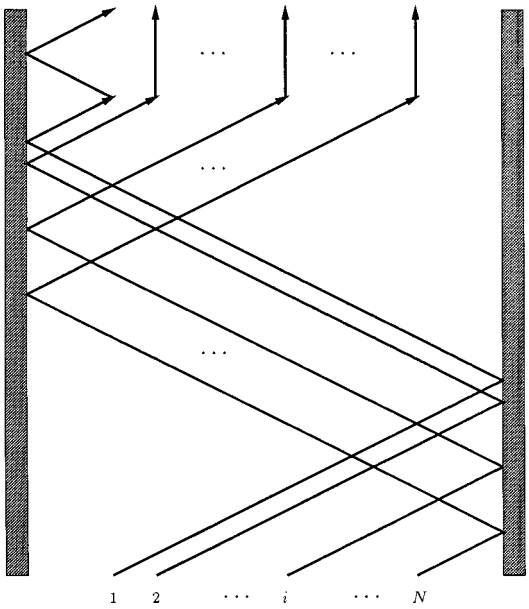

(4.18)

□

Based on Theorem 4.1 we now compute the explicit forms of the conserved operators $\mathscr{W}_n = \mathscr{Y}_n(\xi)|_{\xi_j=-\nu-(j-1)\mu}$. Explicit forms of these operators are given by the following theorem.

Theorem 4.2: *The Ruijsenaars operators $\mathscr{W}_n$ (4.10) are explicitly given by*

$$\mathscr{W}_n = \sum_{\substack{J\subset\{1,\ldots,N\},|J|\leqslant n \\ \epsilon_j=\pm 1, j\in J}} U_{J^c,n-|J|} V_{\epsilon J;J^c} T_{\epsilon J}, \quad \text{for } 1\leqslant n\leqslant N. \tag{4.19}$$

Here the shift operator $T_{\epsilon J}$ and functions $U_{K,m}$ and $V_{\epsilon J;K}$ are, respectively, defined as

$$T_{\epsilon J} = \prod_{j\in J} T_j(2\epsilon_j\beta), \tag{4.20a}$$

$$V_{\epsilon J;K} = \prod_{j\in J} H(\epsilon_j z_j)\overline{H}(\epsilon_j z_j+\beta) \prod_{\substack{j,j'\in J \\ j<j'}} \sigma_\mu(\epsilon_j z_j+\epsilon_{j'} z_{j'})\sigma_\mu(\epsilon_j z_j+\epsilon_{j'} z_{j'}+2\beta)$$

$$\times \prod_{\substack{j\in J \\ k\in K}} \sigma_\mu(\epsilon_j z_j+z_k)\sigma_\mu(\epsilon_j z_j-z_k), \tag{4.20b}$$

$$U_{K,m} = (-1)^m \sum_{\substack{J\subset K,|J|=m \\ \epsilon_j=\pm 1, j\in J}} w_{m,|K|-m}(-\epsilon_{j_1} z_{j_1},\ldots,-\epsilon_{j_m} z_{j_m}) \prod_{j\in J} H(\epsilon_j z_j)$$

$$\times \prod_{\substack{j,j'\in J \\ j<j'}} \sigma_\mu(\epsilon_j z_j+\epsilon_{j'} z_{j'}) \prod_{\substack{j\in J \\ k\in K\setminus J}} \sigma_\mu(\epsilon_j z_j+z_k)\sigma_\mu(\epsilon_j z_j-z_k), \tag{4.20c}$$

where

$$G(\xi,z) = \sum_{r=0}^{3} g_r \sigma_{2\xi}^r(z), \quad H(z) = G(\nu,z), \tag{4.21a}$$

$$\overline{G}(\xi,z) = \sum_{r=0}^{3} \overline{g}_r \sigma_{2\xi}^r(z), \quad \overline{H}(z) = \overline{G}(\overline{\nu},z). \tag{4.21b}$$

The function $w_{p,q}(z_1,\ldots,z_p)$ is a symmetric function of p variables, and recursively defined as

$$
\begin{aligned}
w_{p+1,q}(z_1,\ldots,z_{p+1}) = w_{p,q}(z_1,\ldots,z_p)\Bigg(&\Bigg(\prod_{k=1}^{p}\sigma_\mu(z_{kp+1})\Bigg)\overline{G}(-\nu-(p+q)\mu,-z_{p+1}+\beta)\\
&-\sum_{j=1}^{p}\sigma_{-2\nu-(p+2q)\mu}(z_{jp+1})\Bigg(\prod_{\substack{k=1\\k\neq j}}^{p}\sigma_\mu(z_{kj})\Bigg)\overline{G}(-\nu-(p+q)\mu,-z_j+\beta)\Bigg)\\
&-\sum_{j=1}^{p}w_{p-1,q}(z_1,\ldots,z_{j-1},z_{j+1},\ldots,z_p)\Bigg(\prod_{\substack{k=1\\k\neq j}}^{p}\sigma_\mu(z_{kj})\Bigg)\\
&\times\Bigg(\prod_{\substack{k=1\\k\neq j}}^{p}\sigma_\mu(z_k+z_j-2\beta)\Bigg)\overline{H}(-z_j+\beta)\\
&\times\overline{H}(z_j-\beta)\sigma_{-2\nu-(p+2q)\mu}(-z_j-z_{p+1}+2\beta),
\end{aligned}
\tag{4.22}
$$

where the first two functions are given by

$$
w_{1,q}(z_1)=\overline{G}(-\nu-q\mu,-z_1+\beta),
\tag{4.23a}
$$

$$
\begin{aligned}
w_{2,q}(z_1,z_2)=&\overline{G}(-\nu-q\mu,-z_1+\beta)\overline{G}(-\nu-q\mu,-z_2+\beta)\sigma_\mu(z_1+z_2-2\beta)\\
&-\overline{d}(-\nu-q\mu)\sigma_{-2\nu-(2q+1)\mu}(-z_1-z_2+2\beta).
\end{aligned}
\tag{4.23b}
$$

The function $w_{p,q}(z_1,\ldots,z_p)$ is indeed a symmetric function as will be shown in Lemma 4.3. We have not found a simple expression for a symmetric function $w_{p,q}$ in the case of $p\geqslant 3$. We remark that the function $w_{p,q}(z_1,\ldots,z_p)$ can be rewritten as

$$
\begin{aligned}
w_{p,q}(z_1,\ldots,z_p)=&\prod_{k=1}^{p}\overline{G}(-\nu-q\mu,-z_k+\beta)\prod_{1\leqslant k<k'\leqslant p}\sigma_\mu(z_k+z_{k'}-2\beta)\\
&+\overline{d}(-\nu-q\mu)f_{p,q}(z_1,\ldots,z_p),
\end{aligned}
\tag{4.24}
$$

for some symmetric function $f_{p,q}(z_1,\ldots,z_p)$. By using this expression, one sees, in the case of $\nu+\overline{\nu}=0$, that the highest conserved operator $\mathscr{W}_N$ takes a simple form since $\overline{d}(\overline{\nu})=0$ and that it reduces to the operator proposed in Ref. 7 (see also Ref. 6).

To prove Theorem 4.2, we use the following two lemmas.

Lemma 4.3: The zeroth power term of the shift operator in $\overline{\mathscr{K}}_m^n$ (4.5) *is given by function $w_{n-m+1,m-1}(z_m,\ldots,z_n)$, which satisfies* (4.22) *and* (4.23).

Proof: We show this lemma by mathematical induction for n. For $n=m$ and $n=m+1$ cases, Eq. (4.23) can be shown by a direct calculation. We check the $n+1$ case. Since the operator $\overline{\mathscr{K}}_m^{n+1}$ can be rewritten by use of $\overline{\mathscr{K}}_m^n$ as

$$
\overline{\mathscr{K}}_m^{n+1}=\overline{\mathscr{K}}_m^n\Bigg(\prod_{k=m}^{n}\sigma_\mu(z_{kn+1})-\sum_{l=m}^{n}\sigma_{\xi_m+\xi_{n+1}}(z_{ln+1})\prod_{\substack{k=m\\k\neq l}}^{n}\sigma_\mu(z_{kl})\hat{s}_{ln+1}\Bigg)\overline{K}^{n+1}(\xi_{n+1}),
\tag{4.25}
$$

one finds that the zeroth power term and the $T_j(-2\beta)$ term in $\overline{\mathscr{K}}_m^n$ only contribute to the zeroth power term in $\overline{\mathscr{K}}_m^{n+1}$. As the $T_n(-2\beta)$ term in $\overline{\mathscr{K}}_m^n$ is easily given by

$$
w_{n-m,m}(z_m,\ldots,z_{n-1})\prod_{k=m}^{n-1}\sigma_\mu(z_{kn})(-\overline{H}(-z_n+\beta))T_n(-2\beta)\hat{t}_n,
\tag{4.26}
$$

the $T_j(-2\beta)$ terms in $\bar{\mathscr{K}}_m^n$ are given by symmetrizing (4.26). Substituting these expressions into (4.25) and gathering the zeroth power terms, we obtain the recursive equations for functions $w_{p,q}$ (4.22).

The fact that the function $w_{p,q}$ is symmetric is trivial by construction. □

Lemma 4.4: On the space both W_A-invariant in $z_m,\ldots,z_n$ and W_B-invariant in $z_1,\ldots,z_{m-1}$, the operator $\mathscr{R}_m^n.\mathscr{K}_m^n.\bar{\mathscr{R}}_m^n$ is explicitly given by

$$\mathscr{R}_m^n.\mathscr{K}_m^n.\bar{\mathscr{R}}_m^n=(-1)^{n-m+1}\sum_{\substack{J\subset\{1,\ldots,n\},|J|=n-m+1\\ \epsilon_j=\pm 1,j\in J}}\prod_{j\in J}H(\epsilon_j z_j)\prod_{\substack{j,j'\in J\\ j<j'}}\sigma_\mu(\epsilon_j z_j+\epsilon_{j'}z_{j'})$$
$$\times\prod_{\substack{j\in J\\ k\in\{1,\ldots,n\}\setminus J}}\sigma_\mu(\epsilon_j z_j+z_k)\sigma_\mu(\epsilon_j z_j-z_k)\left(\prod_{j\in J,\epsilon_j=1}\hat{t}_j\right)\hat{s}_{\{m,\ldots,n\},J}. \tag{4.27}$$

Proof: First we note that this operator maps the above space to W_B-invariant space in $z_1,\ldots,z_n$. Hence it is sufficient to calculate the $\Pi_{j=m}^n\hat{t}_j$ term, which is simply computed as

$$\left(\prod_{j=m}^n(-H(z_j))\right)\left(\prod_{m\leqslant j<j'\leqslant n}\sigma_\mu(z_j+z_{j'})\right)\left(\prod_{j=m}^n\prod_{k=1}^{m-1}\sigma_\mu(z_{jk})\sigma_\mu(z_j+z_k)\right)\prod_{j=m}^n\hat{t}_j. \tag{4.28}$$

By the W_B-invariance, we obtain the result. □

Now we are in position to show Theorem 4.2.

Proof: We shall calculate terms $\Pi_{j=q}^N T_j(2\beta)$ for $N-n+1\leqslant q\leqslant N$ in operators $\mathscr{W}_n$, which appear only in the following form:

$$\mathscr{R}_{N-n+1}^{q-1}.\mathscr{K}_{N-n+1}^{q-1}.\bar{\mathscr{R}}_{N-n+1}^{q-1}\left(\prod_{j=q}^N\prod_{k=N-n+1}^{q-1}\sigma_\mu(z_{jk})\right)\bar{\mathscr{K}}_{N-n+1}^{q-1}$$
$$\times\left(\prod_{j=q}^N H(z_j)\bar{H}(z_j+\beta)\right)\left(\prod_{q\leqslant j<j'\leqslant N}\sigma_\mu(z_j+z_{j'})\sigma_\mu(z_j+z_{j'}+2\beta)\right)$$
$$\times\left(\prod_{j=q}^N\prod_{k=1}^{N-n}\sigma_\mu(z_{jk})\sigma_\mu(z_j+z_k)\right)\left(\prod_{j=q}^N\prod_{k=N-n+1}^{q-1}\sigma_\mu(z_j+z_k)\right)\prod_{j=q}^N T_j(2\beta). \tag{4.29}$$

By Lemma 4.3 and 4.4, we can pick up the zeroth power terms in $\mathscr{R}_{N-n+1}^{q-1}.\mathscr{K}_{N-n+1}^{q-1}.\bar{\mathscr{R}}_{N-n+1}^{q-1}$ and $\bar{\mathscr{K}}_{N-n+1}^{q-1}$. Then we have the explicit forms

$$U_{\{1,\ldots,q-1\},q-N+n-1}\left(\prod_{j=q}^N H(z_j)\bar{H}(z_j+\beta)\right)\left(\prod_{q\leqslant j<j'\leqslant N}\sigma_\mu(z_j+z_{j'})\sigma_\mu(z_j+z_{j'}+2\beta)\right)$$
$$\times\left(\prod_{j=q}^N\prod_{k=1}^{q-1}\sigma_\mu(z_{jk})\sigma_\mu(z_j+z_k)\right)\prod_{j=q}^N T_j(2\beta). \tag{4.30}$$

The W_B-invariance of operator $\mathscr{W}_n$ implies Theorem 4.2. □

V. CONCLUSION

We have constructed two sets of commutative difference operators associated with the Weyl group of type-*A* and type-*B*. By use of a pictorial interpretation, we have shown that they possess the Weyl group invariance, and that the higher conserved operators can be constructed in terms of the operator-valued solutions of the YBE and the RE. We have also computed the explicit forms of all the conserved operators. It has been shown that the conserved operators of type-*A* coincide with the elliptic Macdonald operators, which are gauge equivalent with the Hamiltonians of the Ruijsenaars models. We note that our *D*-type Hamiltonian (4.1) is the most general integrable

Hamiltonian, which contains ten arbitrary parameters. One sees[4] that, in a certain limit, it reduces to the so-called Macdonald–Koornwinder operator i.e., the multivariable Askey–Wilson operator. Contrary to the case of the Askey–Wilson polynomial, a unity is not an eigenfunction of our operator $\mathscr{W}_1$ (4.1) except the case, $\overline{\nu}+\nu=-(N-1)\mu$. It is a challenging problem to determine how we diagonalize the elliptic difference operators $\mathscr{M}_n$ (3.1) and $\mathscr{W}_n$ (4.10) by use of the ''symmetric'' and ''nonsymmetric'' elliptic functions.

ACKNOWLEDGMENTS

The authors would like to thank Miki Wadati for kind interest in this work. One of the authors (Y.K.) is a Research Fellow of the Japan Society for the Promotion of Science.

APPENDIX: FUNDAMENTAL FUNCTIONS AND IDENTITIES

We shall establish notations and useful identities on the theta functions[27,28] used in this paper. The Jacobi theta functions are defined for $\Im\tau>0$ as

$$\vartheta_1(z)=-\mathrm{i}\sum_{n\in\mathbb{Z}}\exp\left(\mathrm{i}\pi\left(n+\frac{1}{2}\right)^2\tau+2\pi\mathrm{i}\left(n+\frac{1}{2}\right)z+\mathrm{i}\pi n\right), \tag{A1a}$$

$$\vartheta_2(z)=\sum_{n\in\mathbb{Z}}\exp\left(\mathrm{i}\pi\left(n+\frac{1}{2}\right)^2\tau+2\pi\mathrm{i}\left(n+\frac{1}{2}\right)z\right), \tag{A1b}$$

$$\vartheta_3(z)=\sum_{n\in\mathbb{Z}}\exp(1\pi n^2\tau+2\pi\mathrm{i}nz), \tag{A1c}$$

$$\vartheta_0(z)=\sum_{n\in\mathbb{Z}}\exp(\mathrm{i}\pi n^2\tau+2\pi\mathrm{i}nz+\mathrm{i}\pi n), \tag{A1d}$$

and $\vartheta_4(z)=\vartheta_0(z)$. These functions have a kind of periodic behavior under the translation generated by 1 and τ. Note that $\vartheta_1(z)$ is odd while the other three are even.

Let us define functions $\sigma_\mu^r(z)$ for $r=0,1,2,3$ in terms of the Jacobi theta functions by

$$\sigma_\mu^r(z)=\frac{\vartheta_{r+1}(z-\mu)\vartheta_1'(0)}{\vartheta_{r+1}(z)\vartheta_1(-\mu)}, \tag{A2}$$

where $\vartheta_1'(z)$ denotes a derivative of $\vartheta_1(z)$ with respect to z. In this paper we often refer to $\sigma_\mu^0(z)$ as $\sigma_\mu(z)$ for brevity. We remark that the elliptic functions $\sigma_\mu^r(z)$ satisfy[27,28]

$$\sigma_{-\mu}^r(-z)=-\sigma_\mu^r(z), \tag{A3}$$

$$\sigma_\mu^r(z)\sigma_\mu^r(-z)=\wp(\mu)-\wp(z+\omega_r), \tag{A4}$$

$$\sum_{i=1}^{q}\left(\prod_{\substack{k=1\\k\neq i}}^{q}\sigma_{\mu_k}(z_{ki})\right)\sigma_{\Sigma_{j=1}^q\mu_j}(z_i)=\prod_{k=1}^{q}\sigma_{\mu_k}(z_k), \tag{A5}$$

where $\wp(z)$ is Weierstrass's $\wp$-function with periods 1 and τ.

[1] S. N. M. Ruijsenaars, Commun. Math. Phys. **110**, 191 (1987).
[2] F. W. Nijhoff, V. B. Kuznetsov, E. K. Sklyanin, and O. Ragnisco, J. Phys. A **29**, L333 (1996).
[3] V. B. Kuznetsov, F. W. Nijhoff, and E. K. Sklyanin, Commun. Math. Phys. **189**, 855 (1997).
[4] K. Hikami and Y. Komori, Mod. Phys. Lett. A **11**, 2861 (1996); **12**, 751 (1997).
[5] K. Hasegawa, Commun. Math. Phys. **187**, 289 (1997).
[6] Y. Komori and K. Hikami, J. Phys. A **30**, 4341 (1997).
[7] J. F. van Diejen, J. Math. Phys. **35**, 2983 (1994).
[8] J. F. van Diejen, J. Math. Phys. **36**, 1299 (1995).
[9] I. G. Macdonald, Sémin. Bourbaki **797**, 1 (1994–95).
[10] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed. (Oxford University Press, Oxford, 1995).

[11] T. H. Koornwinder, Contemp. Math. **138**, 189 (1992).
[12] I. V. Cherednik, Int. Math. Res. Notices **9**, 171 (1992).
[13] M. Noumi, RIMS Kôkyûroku **919**, 44 (1995), in Japanese.
[14] A. N. Kirillov and M. Noumi, internal report, 1996, q-alg/9605004.
[15] K. Hikami, J. Phys. A **29**, L281 (1996).
[16] K. Hikami and M. Wadati (unpublished).
[17] I. V. Cherednik, Commun. Math. Phys. **169**, 441 (1995).
[18] I. V. Cherednik, Int. Math. Res. Notices **1**, 43 (1995).
[19] M. Gaudin, J. Phys. (France) **49**, 1857 (1988).
[20] Y. Shibukawa and K. Ueno, Lett. Math. Phys. **25**, 239 (1992).
[21] K. Hikami, Phys. Lett. A **205**, 167 (1995).
[22] K. Hikami, J. Phys. A **29**, 2135 (1996).
[23] Y. Komori and K. Hikami, Nucl. Phys. B **494**, 687 (1997); K. Hikami and Y. Komori, Mod. Phys. Lett. A **11**, 751 (1997).
[24] V. M. Buchstaber, G. Felder, and A. P. Veselov, Duke Math. J. **76**, 885 (1994).
[25] I. Cherednik, Theor. Math. Phys. **61**, 35 (1984).
[26] E. K. Sklyanin, J. Phys. A **21**, 2375 (1988).
[27] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed. (Cambridge University Press, Cambridge, 1927).
[28] A. Hurwitz and R. Courant, *Vorlesungen über Allgemeine Funktionentheorie und Elliptische Funktionen* (Springer, Berlin, 1964).

Discrete spectral triples and their symmetries

Mario Paschke[a] and Andrzej Sitarz[b]
Institut für Physik, Johannes-Gutenberg Universität, 55130 Mainz, Germany

(Received 28 February 1997; accepted for publication 24 September 1997)

We classify zero-dimensional spectral triples over complex and real algebras and provide some general statements about their differential structure. We investigate also whether such spectral triples admit a symmetry arising from the Hopf algebra structure of the finite algebra. We discuss examples of commutative algebras and group algebras. © *1998 American Institute of Physics.* [S0022-2488(98)01603-X]

I. INTRODUCTION

Recently Connes[1] has proposed a definition of *spectral triples*, which is supposed to extend the notion of the Riemannian geometry to a noncommutative setup. This could be a natural geometrical framework for theories of fundamental interactions and gravity at a small distance scale and could possibly open a way towards quantum gravity.

One of the promising applications to fundamental particle physics is an attempt to explain the particle contents and interaction pattern in the Standard Model. A finite algebra $M_3(\mathbb{C})\oplus\mathbb{H}\oplus\mathbb{C}$ plays a significant role in the spectral triple in question. In this paper we shall not discuss this particular physical application. (For a more detailed discussion of the Standard Model spectral triple see Ref. 2.)

We derive here the general structure of zero-dimensional spectral triples and of the Dirac operator for complex and real semisimple algebras. Moreover, assuming that such algebras are Hopf algebras, we discuss possible symmetry structures. As examples we present $\mathbb{C}^n$ commutative algebra with discrete group structure as well as group algebras, particularly $\mathbb{C}S_3$ for the permutation group S_3.

II. COMPLEX SPECTRAL TRIPLES

Let $\mathcal{A}_F$ be a finite-dimensional semi-simple algebra over the field of complex numbers $\mathbb{C}$. (C^* algebras are semisimple, thus we consider only them. Of course, nonsemisimple algebras (Grassman variables, for instance) may also appear in physical models. Whether one can successfully construct a corresponding theory for them remains an interesting problem.) Such an algebra is isomorphic to $M_{n_1}(\mathbb{C})\oplus\cdots\oplus M_{n_k}(\mathbb{C})$ for some nonzero integers $n_1,\ldots,n_k$ (see Ref. 3, pp. 29–36).

A *spectral triple* for a finite algebra means a finite Hilbert space $\mathcal{H}$, a faithful representation $\pi:\mathcal{A}\to B(\mathcal{H})$ over the field $\mathbb{C}$, a reality structure J, such that J is an antilinear isometry of $\mathcal{H}$, $J^2=1$, and $\pi^0=J\circ\pi^*\circ J$ is a representation of the opposite algebra $\mathcal{A}^0$, which commutes with π:

$$[\pi^0(a),\pi(b)]=0 \quad \forall a,b\in\mathcal{A}.$$

We assume that there exists a grading $\gamma\in B(\mathcal{H})$, such that $\gamma^*=\gamma$, $\gamma^2=1$, $J\gamma=\gamma J$, $\gamma\pi(a)=\pi(a)\gamma$, and γ is equal to $\pi(c)$ for some Hochschild cocycle $c\in Z_0(\mathcal{A},\mathcal{A}\otimes\mathcal{A}^0)$.

We begin with some simple observations.

Observation 1: $\mathcal{H}$ is a bimodule over $\mathcal{A}$. If $\xi\in\mathcal{H}$, then we have $a\xi=\pi(a)\xi$ and $\xi a=\pi^0(a)\xi$.

We can now use the structure of the finite algebra $\mathcal{A}$ to learn more about the Hilbert space of the spectral triple.

Lemma 1: Let P_i be the element of $\mathcal{A}$ with identity matrix in the ith entry and zeroes elsewhere. Let H_{ij} be $P_i\mathcal{H}P_j$. Then $\mathcal{H}$ is a direct sum of all H_{ij},

[a]Electronic mail: paschke@dipmza.physik.uni-mainz.de
[b]Present address: IF UJ, Reymonta 4, 30-059 Kraków, Poland. Electronic mail: sitarz@higgs.physik.uni-mainz.de

 © 1998 American Institute of Physics

$$\mathscr{H}=\oplus_{i,j} H_{ij},$$

and if for every $a\in\mathscr{A}$ *we denote* $P_i a = a P_i$ as a_i *we have*

$$(a\xi)_{ij}=a_i\xi_{ij},\quad (\xi a)_{ij}=\xi_{ij}a_j.$$

The proof is a simple calculation and we skip it.

Lemma 2: The Hilbert subspace H_{ij} *has the following form:*

$$H_{ij}=\mathbb{C}^{n_i}\otimes\mathbb{C}^{r_{ij}}\otimes\mathbb{C}^{n_j},$$

where r_{ij} *are some integers,* $r_{ij}\geqslant 0$. *We take* $H_{ij}=0$ *iff* $r_{ij}=0$. *An element* a *acts on* H_{ij} *from the left as* $a_i\otimes 1\otimes 1$ *and from the right as* $1\otimes 1\otimes a_j^T$, *where* ^{T} *denotes the matrix transposition.*

From the previous lemma we know that H_{ij} is the representation space for $M_{n_i}(\mathbb{C})$ and for the opposite algebra of $M_{n_j}(\mathbb{C})$. Because these representations commute, the structure of H_{ij} must have the above form.

Lemma 3: The grading operator γ *is* ± 1 *when restricted to the subspace* H_{ij}.

Proof: Since γ commutes with both the elements of $\mathscr{A}$ and J, and is self-adjoint, we conclude that it also commutes with the right multiplication on $\mathscr{H}$:

$$a(\gamma\xi)b=\gamma(a\xi b).$$

Hence, $\gamma H_{ij}\subset H_{ij}$. When restricted to H_{ij}, γ is an isometry, which commutes with the left multiplication by matrices from $M_{n_i}(\mathbb{C})$ and right multiplication by matrices from $M_{n_j}(\mathbb{C})$. Therefore it must be of the form

$$\gamma_{ij}: H_{ij}\to H_{ij},\quad \gamma=1_{n_i}\otimes\Gamma_{ij}\otimes 1_{n_j},$$

where Γ_{ij} is an arbitrary self-adjoint matrix, such that $\Gamma_{ij}^2=1$ acting on $\mathbb{C}^{r_{ij}}$. However, if we take into account that γ is an image of a Hochschild cocycle, i.e., there exists an element c of $\mathscr{A}\otimes\mathscr{A}^0$ such that $\pi(c)=\gamma$, we see that Γ_{ij} must be proportional to the identity matrix, so it leaves only the possibility for $+1$ or -1. ■

Lemma 4: The reality J *maps* H_{ij} *onto* H_{ji}, *therefore* r_{ij} *must be equal to* r_{ji}.

Proof: Let us take $\xi\in H_{ij}$. There exists $\nu\in\mathscr{H}$ such that $\xi=P_i\nu P_j$. Next, we calculate $J\xi$, writing all the expressions carefully:

$$J\xi=JP_i\nu P_j=J\pi(P_i)(J\pi(P_j)J)\nu=(JP_iJ)P_j(J\nu)=P_j(J\nu)P_i\in H_{ji}.$$

Of course, since $J^2=\mathrm{id}$ we must have $r_{ij}=r_{ji}$ so that the dimensions of both Hilbert spaces are equal. ■

Observation 2: If we denote by γ_{ij} *the sign of the grading on* H_{ij}, *then* $\gamma_{ij}=\gamma_{ji}$.

This is a trivial consequence of the commutation relation $[J,\gamma]=0$.

We can now introduce $q_{ij}=r_{ij}\gamma_{ij}$, which, later on, could be identified as the intersection form for our spectral triple. So far, we know that q_{ij} is a symmetric matrix with integer entries.

Lemma 5: There exists a basis of H_{ij}, H_{ji} *of the form* $H_{ij}\ni\nu=\nu_i\otimes\nu_{ij}\otimes\nu_j$ *such that*

$$J\nu=\nu_j\otimes\nu_{ji}\otimes\nu_i\in H_{ji}. \tag{1}$$

Proof: We begin with the $i\neq j$ case. First of all, let us fix an orthonormal basis of the spaces H_{ij} and H_{ji} of the form $e_1\otimes e_2\otimes e_3$ and define $\tilde{J}$ as an antilinear isometry operator, which exchanges the first and the third element of the tensor product:

$$\tilde{J}(e_1\otimes e_2\otimes e_3)=e\otimes e_2\otimes e_1.$$

Then $\tilde{J}$ satisfies all the axioms for the spectral triple, i.e., $\tilde{J}^2=1$ and $a^0=\tilde{J}a^\dagger\tilde{J}$. Next, observe that $J\circ\tilde{J}$ is a linear invertible map, such that

$$J\circ\tilde{J}a = aJ\circ\tilde{J}$$

and

$$J\circ\tilde{J}a^0 = a^0 J\circ\tilde{J}.$$

We prove only the first of the two:

$$J\circ\tilde{J}a = J\circ\tilde{J}a\tilde{J}^2 = J(a^\dagger)^0\tilde{J} = J^2 aJ\circ\tilde{J} = aJ\circ\tilde{J}.$$

Therefore $J\circ\tilde{J}$ commutes both with the left and the right multiplication, so it has to be of the form $1\otimes\tilde{l}\otimes 1$, where $\tilde{l}$ is a linear isometry $l:\mathbb{C}^{r_{ij}}\to\mathbb{C}^{r_{ji}}$. Thus $J=(1\otimes\tilde{l}^{-1}\otimes 1)\circ\tilde{J}$ and the only possible nontrivial part of J comes from $\tilde{l}$.

In the last step we have to consider the cases $j=i$ and $j\neq i$ separately. The latter one is simpler, we just chose any orthonormal basis e_s of $C^{r_{ij}}$ (same as in the definition of $\tilde{J}$) and then $\tilde{l}^{-1}e_s$ constitutes a basis of $C^{r_{ji}}$. Take as ν_{ij} just the vectors of the chosen basis and ν_{ji} their image under $\tilde{l}^{-1}$. Then the property (1) follows immediately.

A more subtle situation occurs in the $i=j$ case. Here, however, we can use another simple argument. Take $\nu\in C^{r_{ii}}$. Then either $\tilde{l}^{-1}\nu$ is proportional to ν or is linearly independent from it. In the latter case it is sufficient to consider ν_{ii} to be $\nu+\tilde{l}^{-1}\nu$ and $i(\nu-\tilde{l}^{-1}\nu)$. In the first case $\tilde{l}^{-1}\nu=e^{i\phi}\nu$ and then $e^{i\phi/2}\nu$ will do as ν_{ii}. Since our space is finite dimensional and l is an isometry we can carry on with this procedure until we find a complete basis of H_{ii} satisfying our requirement. ■

A. The Dirac operator

Up to this point we have constructed the basic geometry of a finite spectral triple and in the next step we shall introduce the Dirac operator, which gives the differential and metric structures.

The Dirac operator D of a finite spectral triple is a self-adjoint linear operator on $\mathscr{H}$, which commutes with J: $DJ=JD$, anticommutes with γ: $D\gamma=-\gamma D$, and satisfies the following relation:

$$[[D,\pi(a)],\ \pi^0(b)]=0 \quad \forall a,b\in\mathscr{A}. \tag{2}$$

Observation 3: Let $D_{ij,kl}=P_{ij}DP_{kl}$, where P_{ij} is the projection operator on H_{ij}: ($P_{ij}\nu=P_i\nu P_j$). Then the following is true:

(1) *$D_{ij,kl}:H_{kl}\to H_{ij}$ is a linear map, which is zero unless $\gamma_{ij}\gamma_{kl}<0$ (i.e., they are of opposite sign).*
(2) *$D^\dagger_{ij,kl}=D_{kl,ij}$. This follows easily from the requirement $D=D^\dagger$.*
(3) *$D_{ij,kl}=JD_{ji,lk}J$. This follows from the relation $[D,\ J]=0$ and $J^2=\mathrm{id}$. If we choose the basis of the spaces H_{ij} so that the operator J has a preferred form $J\nu=\pm\nu$, the above condition can be simplified to $D_{ij,kl}=D^*_{ji,lk}$, where $*$ means the complex conjugation of matrix elements of $D_{ji,lk}$ in the preferred basis.*

The strongest restriction for the Dirac operator follows from the commutation property (2).

Lemma 6: The Dirac operator component $D_{ij,kl}$ vanishes unless $i=k$ or $j=l$. In either of these cases D must commute with the corresponding action of the $\mathscr{A}$ or $\mathscr{A}^0$, respectively.

Proof: Let us rewrite the condition (2) for $\nu\in H_{kl}$:

$$D_{ij,kl}(a_k\nu b_l)-a_iD_{ij,kl}(\nu b_l)-(D_{ij,kl}(a_k\nu))b_j+a_i(D_{ij,kl}\nu)b_j=0.$$

Choosing $b=P_j$, $j\neq l$, we get that $D_{ij,kl}=0$ unless $i=k$ and, for $i=k$, $D_{ij,il}a_i=a_iD_{ij,il}$ for every $a\in\mathscr{A}$.

Similarly, if we choose $a=P_i$, $i\neq k$, we get that $D_{ij,kl}=0$ unless $j=l$ and additionally $D_{ij,kj}(\nu b_j)=(D_{ij,kj}\nu)b_j$.

Of course, the case $i=k$ and $j=l$ is excluded by the requirement that D acts only between spaces of different grading. ■

B. Differential algebra

The Dirac operator D provides us with both the metric and the differential properties on the finite space corresponding to the algebra $\mathcal{A}$. We shall now concentrate on some general properties of the differential structure.

Observation 4: The representation π of $\mathcal{A}$ extends to the universal differential algebra $\Omega_u\mathcal{A}$, by

$$\pi(a_0 da_1 d_2 \cdots da_n) = \pi(a_0)[D,\pi(a_1)]\cdots[D,\pi(a_n)].$$

The differential calculus is usually constructed as follows: first, $\Omega^1(\mathcal{A})$ is the quotient of $\Omega^1_u\mathcal{A}$ by the kernel of π. Therefore the bimodule $\Omega^1(\mathcal{A})$ is isomorphic to $\pi(\Omega^1_u(\mathcal{A}))$. Higher-order forms are obtained by taking the quotient of the universal forms of a given order by the ideal $\Omega^n_u\mathcal{A}\cap(\ker\pi\cup d\ \ker\pi)$.

Let us state some general results.

Lemma 7: The first-order differential calculus is inner; there exists a one-form

$$\xi=\sum_{i\neq j} P_i dP_j, \tag{3}$$

such that for every $a\in\mathcal{A}$ one has $da=[\xi,\ a]$. The Dirac operator $D=\pi(\xi)+J\pi(\xi)J$.

Proof: Of course, from the definition we have $\pi(da)=[D,\ \pi(a)]$, so one has to show

$$\sum_{i\neq j}[\pi(P_i)[D,\ \pi(P_j)],\ \pi(a)]=[D,\ \pi(a)].$$

Let us compute $\xi_{ij,kl}=P_{ij}\pi(\xi)P_{kl}$. For any $h\in H_{kl}$ we have

$$(\pi(\xi)h)_{ij}=\left(\sum_{s\neq t} P_s[D,\ P_t]h\right)_{ij}=\sum_{s\neq t}(P_sDP_th)_{ij}=(P_iDP_kh)_{ij}=\delta_{jl}D_{ij,kj}h.$$

Note that

$$\xi=\sum_{i\neq j} P_i dP_j=-\sum_i P_i dP_i.$$

The relation $D=\pi(\xi)+J\pi(\xi)J$ follows immediately from the properties of J and D, additionally $[J\xi J,\ a]=0$ for every element $a\in\mathcal{A}$, so that the property $da=[D,\ a]=[\xi,\ a]$ holds.■

Observation 5: $\pi(\xi)$ is a self-adjoint operator, thus the first-order calculus has a natural involution map such that $d\circ *=-(*\circ d)$ if we define $\xi^*=\xi$.

The one-form ξ has some more interesting properties, which make the task of determining the differential structure easier. Before we discuss the construction of $\Omega^2(\mathcal{A})$ let us make another useful remark.

Observation 6: The bimodule $\Omega^1(\mathcal{A})$ has no center. If $a\omega=\omega a$ for every element $a\in\mathcal{A}$, then $\omega=0$.

Proof: First, let us observe that for every idempotent $e=e^2$, we have $edee=0$. This follows from differentiating $0=e^2-e$ and multiplying it from the left and from the right by e. Therefore $P_idP_iP_i=P_i\xi P_i=0$.

Every element of $\Omega^1(\mathcal{A})$ could be written as a finite sum of the type $\omega=\Sigma_\alpha b_\alpha\xi c_\alpha$. Suppose that it commutes with every $a\in\mathcal{A}$, in particular with P_i. Then if we multiply both sides by P_i, since $P_i^2=P_i$ we get $P_i\omega P_i=P_i\omega$. However, $P_i\omega P_i=0$ as P_i commute with elements of the algebra and $P_i\xi P_i=0$. So $P_i\omega=0$ for every i, and therefore $\omega=0$. ■

Lemma 8: $d\xi=\xi\xi-\Sigma P_i\xi\xi P_i$.

Proof:

$$d\xi=-\sum_i dP_i dP_i=-\sum_i[\xi,\ P_i][\xi,\ P_i]=\xi\xi+\sum_i P_i\xi\xi P_i.$$

■

The above equality holds only within the differential algebra, which means that all products of one-forms in the line above must be calculated using the product in $\Omega^2(\mathcal{A})$. We have not determined it yet (in a general situation it is a difficult technical calculation), however, using the above relations we will be able to say something about it.

Lemma 9: Let $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})\ni\Xi=\Sigma_i P_i\xi\otimes_{\mathcal{A}}\xi P_i$. *Let* $\mathcal{J}_1=\ker\pi$ [*where* π *on* $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})$ *is defined in the usual way*: $\pi(\omega_1\otimes_{\mathcal{A}}\omega_2)=\pi(\omega_1)\pi(\omega_2)$], $\mathcal{J}_2$ *be the subbimodule generated by commutators* $[a,\Xi]$ *for all* $a\in\mathcal{A}$, *and finally let* $\mathcal{J}_3$ *be the subbimodule generated by all elements of the form* $\Sigma_i a_i(\xi\otimes_{\mathcal{A}}\xi-\Xi)b_i$ *for all* a_i, b_i *such that* $\Sigma_i a_i\xi b_i=0$. *Then* $\mathcal{J}=\mathcal{J}_1\oplus\mathcal{J}_2\oplus\mathcal{J}_3$ *is a subbimodule of* $\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})$ *and* $\Omega^2(\mathcal{A})=\Omega^1(\mathcal{A})\otimes_{\mathcal{A}}\Omega^1(\mathcal{A})/\mathcal{J}$.

Proof: We know that every element of $\Omega^1(\mathcal{A})$ can be written as a finite sum of $\omega=a\xi b$. One easily finds $d\omega$,

$$d\omega=\xi\omega-\omega\xi+a(\tilde{\Xi}-\xi\xi)b, \tag{4}$$

and, in a special case, when ω is of the form $da=\xi a-a\xi$, we obtain

$$d^2a=\tilde{\Xi}a-a\tilde{\Xi}, \tag{5}$$

where $\tilde{\Xi}$ denotes an element in $\Omega^2(\mathcal{A})$ which is the image of Ξ. Now, we clearly see that division by $\mathcal{J}_2$ guarantees that d is well defined ($d0=0$) and $d^2\equiv 0$ if and only if we quotient out $\mathcal{J}_3$. Note that in order to obtain a differential algebra we do not need to include $\mathcal{J}_1$ as the quotient; however, we proceed so in order to be in agreement with the *standard* definition of the differential algebra for spectral triples. ■

Before we start considering an interesting simplification, let us say a bit more about the interpretation of $\xi\xi$ and Ξ. The image of the latter (under π) is an operator which links pairs of subspaces of $\mathcal{H}$, which have the same first index i. Clearly, it means that the Dirac operator ''connects'' $H_{ik}\to H_{pk}\to H_{ik}$, which is always the case, as ξ is self-adjoint (for every map $H_{ik}\to H_{pk}$ there exists a conjugate map $H_{pk}\to H_{ik}$). Now, $\pi(\Xi-\xi\xi)$ is a set of maps between subspaces of $\mathcal{H}$ with different first index, so it does not vanish if and only if ξ connects $H_{ij}\to H_{kj}\to H_{pj}$ for some $p\neq i$ and the composition of these maps does not vanish.

An interesting situation occurs when $\pi(\xi\otimes_{\mathcal{A}}\xi)=\pi(\Xi)$, so that $\mathcal{J}_2\subset\mathcal{J}_1$.

1. Inner calculi for spectral triples

Lemma 10: When $\pi(\xi\otimes_{\mathcal{A}}\xi)=\pi(\Xi)$, *then* $\tilde{\Xi}=\xi\xi$ *and the calculus remains inner in* $\Omega^2(\mathcal{A})$:

$$d\omega=\xi\omega+\omega\xi, \tag{6}$$

for any one-form ω. *The subbimodule* $\mathcal{J}$ *contains the kernel of* π *and all commutators* $[a,\ \xi\otimes_{\mathcal{A}}\xi]$.

This happens, for instance, if $\pi(\xi)\pi(\xi)$ is in $\pi(\mathcal{A})$; additionally, if it commutes with $\pi(\mathcal{A})$, then $\mathcal{J}=\ker\pi$ and the structure of $\Omega^2(\mathcal{A})$ simplifies considerably.

The next lemma makes these observations more precise:

Lemma 11: If $\pi(\xi)\pi(\xi)$ *commutes with* $\pi(Z(\mathcal{A}))$, *where* $Z(\mathcal{A})$ *is the center of* $\mathcal{A}$, *then* $\pi(\xi)\pi(\xi)=\pi(\Xi)$. *The converse is also true.*

Proof: Of course, if $\pi(\xi)\pi(\xi)$ commutes with $\pi(Z(\mathcal{A}))$ it commutes with every $\pi(P_i)$. Using $\Sigma_i P_i\equiv 1$ we get the required identity. For the converse, let us assume that there exists i such that $\pi(P_i)$ does not commute with $\pi(\xi)\pi(\xi)$ and let the commutator be ρ. Then it is easy to verify that $\pi(\xi)\pi(\xi)-\pi(\Xi)=\rho$. ■

It is quite interesting to characterize the space in which $\pi(\xi)\pi(\xi)$ takes value for this particular situation. We already know that it commutes with $\pi(\mathcal{A}^0)$. Since additionally it has to commute with the center of $\pi(\mathcal{A})$, the only possibility is that it is diagonal (i.e., when restricted to H_{ij} maps it to itself), and using the decomposition into tensor products presented in Sec. II, we find that it is $T_{ij}\otimes\mathrm{id}$ on each of the subspaces H_{ij}.

If the calculus [at least up to $\Omega^2(\mathcal{A})$] is inner, we can immediately make an interesting observation about the gauge potentials.

Observation 7: If H is a self-adjoint one-form and $F(H)=dH+HH$ is its curvature form, then for the inner calculus, as described above, $F(-2\xi)=0$.

This follows from a simple calculation. Of course, we have presented here only the *formal* solution of the $F(H)=0$ equation in $\Omega^2(\mathcal{A})$. It does not guarantee that this solution is unique; it could happen (for instance, if we take a tensor product of the spectral triple in question with a triple corresponding to a manifold) that there are far more solutions or even the whole subbimodule $\Omega^2(\mathcal{A})$ for the discrete part vanishes.

2. Commutative discrete spectral triples

A lot more can be said about the commutative case $\mathcal{A}=\mathbb{C}^n$. The generators of the algebra could be identified with the earlier introduced projection operators P_i. The structure of $\Omega^1(\mathcal{A})$ is completely determined by the elements $P_i dP_j$, $i\neq j$ (using $\Sigma_i dP_i=0$). In our representation we have

$$\pi(P_i dP_j)h=P_i DP_j h-P_i\delta_{ij}Dh,$$

which for the kl-component becomes

$$(\pi(P_i dP_j)h)_{kl}=\delta_{ik}D_{kljp}h_{jp}-\delta_{ik}\delta_{ij}D_{klps}h_{ps}=\cdots,$$

which, after taking into account the properties of D, becomes

$$\cdots=\delta_{ik}D_{iljl}h_{jl}-\delta_{ik}\delta_{ij}D_{klpl}h_{pl}.$$

Since we are interested in the case $i\neq j$ we obtain that for $P_i dP_j\neq 0$ it is sufficient that there exists l such that $D_{iljl}\neq 0$.

Let us now investigate $\Omega^2(\mathcal{A})$, which would be composed out of the bimodule $\Omega_u^2(\mathcal{A})$ after we quotient out the subbimodule generated by $\ker\pi\cup d\ker\pi$.

First, we show the representation action of the basis of $\Omega_u^2(\mathcal{A})$, $e_{kij}=P_k dP_i dP_j$, $k\neq i$ and $i\neq j$ (again, to see that this is the basis we use the identity $\Sigma_i dP_i=0$).

Observation 8: For i,j,k such that $k\neq i$ and $i\neq j$, $\pi(e_{kij})$ is a collection of maps $H_{jl}\mapsto H_{kl}$, each of the form

$$\pi(e_{kij})=D_{klil}D_{iljl}.$$

Next we shall show the structure of $d\ker\pi$.

Let us take a one-form ρ in $\ker\pi$. Of course, it is sufficient to take a generator: $\pi(P_i dP_j)=0$ for some $i\neq j$.

Let us calculate $\pi(dP_i dP_j)$. After little calculation we get the following result:

$$\pi(dP_i dP_j)h=-P_i D^2 P_j h \quad \text{if } P_i dP_j\in\ker\pi,\quad i\neq j. \tag{7}$$

Rewriting all that in terms of the basis of Ω_u^2 we obtain that for each $i\neq j$ if $\pi(P_i dP_j)=0$, we have the following generator added to the *junk*:

$$\sum_{i\neq s\neq j} P_i dP_s dP_j.$$

An easy way to visualize the construction of the differential structure coming from the commutative spectral triple is by the way of graphs. Imagine we have n vertices, and each vertex is *split* into n points. The nonzero entries of the Dirac operator shall correspond to arrows linking two points of two different vertices or within the same vertex (of course, due to the reality structure there exists a link between both).

If we look at the differential forms, the *fine* structure of the vertices becomes not so important.

Now, as for the one-forms, we know that $P_i dP_j$, $i\neq j$, exists if there is at least one arrow connecting vertex j with i. For a two-form $P_k dP_i dP_j$ there must exist links $j\to i\to k$. However, if there is no direct link $j\to k$, then the sum of all forms $\Sigma P_k dP_i dP_j$ over $i\neq k$ and $i\neq j$, vanishes. This bears a resemblance to a general structure of differential geometry on graphs.[4]

3. The metric

On the algebra $\mathscr{A}$ we have a natural trace coming from the representation on the Hilbert space $\mathscr{H}$. Since our representation extends naturally to the differential algebra, such a trace gives a *generic* scalar product on the space of one-forms. In particular, for the generators of the commutative algebra (which is the easiest example) we have

$$(P_i dP_j, P_k dP_l) = \mathrm{Tr}\,(\pi^*(P_i dP_j)\pi(P_k dP_l)) = \delta_{ik}\delta_{jl}\sum_p \mathrm{Tr}\,(D_{lpkp}D_{kplp}).$$

The last term of the expression is a self-adjoint map from H_{lp} into itself, therefore the scalar product is positive. All generating one-forms $P_i dP_j$ are orthogonal to each other and the norm of each is fixed by the elements of the Dirac operator. Of course, should the norm be zero then it is obvious that $\forall p\ D_{ipjp}=0$ and hence the one-form $P_i dP_j$ vanishes.

The above choice is not unique, and one may consider its deformation of the following type:

$$(\omega,\rho)z = \mathrm{Tr}\,(Z\pi(\omega)\pi(\rho)^\dagger),$$

where Z is an operator, for which it is sufficient to assume that Z is self-adjoint, and to keep the gauge invariance we need $\pi(U)Z\pi(U)^\dagger = Z$ for any unitary U from the algebra.

The reason for introducing such scaling may follow from the requirement that the trace has some additional symmetry (for instance, we might fix it to recover the Haar measure on the algebra, when it has a Hopf algebra structure).

Of course, changing the trace would have significant physical consequences (for details of the construction and application to the Standard Model, see the forthcoming Ref. 2).

C. Finite spectral triples with real algebras

So far we have discussed finite spectral triples with algebras over $\mathbb{C}$. Since the possible high-energy physics application of noncommutative geometry uses rather a real algebra, we shall now briefly discuss the above construction assuming that $\mathscr{A}$ is over the field of real numbers.

Lemma 12 (See Ref. 3): The simple finite algebras over $\mathbb{R}$ *are of the type* $M_n(F)$*, where* F *could be* $\mathbb{R}$, $\mathbb{C}$, *or* $\mathbb{H}$.

Therefore every semi-simple algebra over $\mathbb{R}$ could be decomposed as

$$\mathscr{A} = \bigoplus_i M_{n_i}(F_i), \tag{8}$$

where each F_i could be the field of real, complex, or quaternionic numbers.

We shall now investigate irreducible representations of $\mathscr{A}$ on a *complex* Hilbert space. Consider $M_n(\mathbb{R})$ and $M_n(\mathbb{H})$ as real algebras. Then each has only one irreducible representation on $\mathbb{C}$ (for quaternions the conjugate representation is equivalent to the fundamental one). As for $M_n(\mathbb{C})$ we have two inequivalent irreducible representations:

$$\pi(a) = a, \quad \overline{\pi}(a) = \overline{a},$$

where $\overline{a}$ means the involution of matrix elements.

How does the description of our spectral triple change? Apparently for $M_n(\mathbb{R})$ and $M_n(\mathbb{H})$ there is no change at all, whereas for $M_n(\mathbb{C})$ we have to take into account the existence of these two inequivalent representations:

Lemma 13: If $F_i=\mathbb{C}$*, then the Hilbert subspace* H_{ij} *decomposes into* H_{ij} *and* $H_{\bar{\imath}j}$*, with the representation of* $M_{n_i}(F_i)$ *by* m *and* $\overline{m}$*, respectively.*

Of course, an analogous decomposition takes place if one takes into account the right action of the algebra on $\mathscr{H}$.

Until now, there has not been much change apart from the additional splitting of the Hilbert space. All other results remain unchanged, in particular, γ cannot be different for H_{ij} and $H_{\bar{\imath}j}$.

If we look at the restrictions for the Dirac operator, which come from (2) we see that the constraints on allowed values of D also remain unchanged; however, one should add some more,

which arise from the fact that one cannot mix the fundamental and the conjugated representation. To see this explicitly one should replace in the proof of Lemma 6 $b=P_j$ by $b=zP_j$ for any complex number z.

Generally speaking, the derivation of the constraint, which we had for the complex case, could be repeated separately for each of the real situations. Therefore, we may summarize the result by stating that D does not mix two inequivalent representations of the same algebra.

D. The intersection form

The intersection form in K-theory is a map $K_*(\mathcal{A})\times K_*(\mathcal{A})\to\mathbb{Z}$ and its invertibility (Poincaré duality) is fundamental for a characterization of homotopy types of spaces, which possess the structure of a smooth manifold.

For finite complex spectral triples the intersection form, evaluated on generators of the K_0-group, e,f, is[1]

$$\langle e,f\rangle=\langle D,e\otimes f^o\rangle,$$

and, generally, the pairing $\langle D,E\rangle$ for a projector E means

$$\langle D,E\rangle=\dim\,\mathrm{coker}_{E\mathcal{H}_R}(ED^+E)-\dim\,\ker(ED^+E),$$

where $\mathcal{H}_L=\frac{1}{2}(1+\gamma).\mathcal{H}$, $\mathcal{H}_R=\frac{1}{2}(1-\gamma).\mathcal{H}$, and $D^+=\frac{1}{4}(1-\gamma)D(1+\gamma)$.

Since for matrix algebras the group K_1 is trivial (any element can be deformed to 1), the only nontrivial part is K_0. For $M_n(\mathbb{C})$ all projectors are equivalent [within $M_\infty(M_n(\mathbb{C}))$] to a diagonal matrix with 1 at the first diagonal entry and zeroes elsewhere, which we shall call e. Of course, for a direct sum of N simple algebras the corresponding K-theory will have N independent generators, each coming from the simple component of the sum.

Let us calculate the $\langle e_i,e_j\rangle$ entry. First, we have to determine the dimensions of Hilbert subspaces appearing in the above formulas:

$$\dim\,e_ie_j^0.\mathcal{H}_L=\begin{cases}r_{ij}, & \text{if } \gamma_{ij}=-1,\\ 0, & \text{if } \gamma_{ij}=1,\end{cases}\tag{9}$$

$$\dim\,e_ie_j^0.\mathcal{H}_R=\begin{cases}r_{ij}, & \text{if } \gamma_{ij}=1,\\ 0, & \text{if } \gamma_{ij}=-1.\end{cases}\tag{10}$$

Now, observe out of $e_ie_j^0.\mathcal{H}_L$ and $e_ie_j^0.\mathcal{H}_R$ one must always be empty, so the $e_ie_j^0D^+e_ie_j^0$ operator, which acts between them, has either domain or target space empty. Thus, the index is independent of D and reads

$$\langle e_i,e_j\rangle=\gamma_{ij}r_{ij}=q_{ij},$$

so we recover the matrix q as our intersection form.

For real algebras the calculation is slightly more complicated. First, one has to take into account the form of projectors for quaternions, which results in the doubling of the corresponding entries in the intersection form. Moreover, for real algebras the group K_1 is nontrivial, as $K_1(M_n(\mathbb{R}))=\mathbb{Z}_2$; however, this still does not change the intersection form if one does not take into account torsion in K-theory.

III. HOPF ALGEBRA SYMMETRIES

Spectral triples are expected to be an extension of the notion of Riemannian manifolds to noncommutative geometry. Similarly as in the standard differential geometry we may attempt to explore symmetries of such manifolds. We expect, however, that the correct idea of a symmetry should be extended from a group to a Hopf algebra (for a good introduction to Hopf algebras and quantum groups we refer the reader to Ref. 5). Another potentially interesting point is the speculation[1,6,7] that the finite algebra of the Standard Model originates from the q-deformation of the Spin structure and may possess some Hopf algebra symmetry (for details see also Ref. 8).

We shall try to answer the question of whether spectral triples admit (and if so, to what extent) a symmetry understood as a Hopf algebra symmetry. Of course, the natural thing is to use the Haar measure as a trace on the algebra and to extend it to the whole representation space. Another possibility is to demand that differential structures of the spectral triple are also symmetric, for instance, that they are left-covariant or bicovariant. Furthermore, the Hilbert space may be investigated for the existence of a comodule structure. Last but not least, we may turn our attention to the tensor product of representations (this is allowed only if a coproduct on the algebra exists), which physically has the meaning of constructing composite states. In this paper we concentrate our efforts on the first two aspects of symmetries, leaving the others for future investigations.

A. Discrete group structure on $\mathbb{C}^n$ algebra

One of the simplest examples of real spectral triples are those based on a finite-dimensional commutative algebra $\mathbb{C}^n$ understood as function algebras on a discrete n-point space. The general construction scheme was presented by Connes[1] and falls into the general classification of complex spectral triples presented in the previous section.

Suppose that we assume the discrete space to have the structure of a finite group and demand that the first-order differential calculus generated by D is bicovariant.

First, let us recall the basics of the bicovariant differential structure for finite groups (details in Ref. 9). The calculus is generated by left-covariant forms:

$$\chi^g=\sum_h e_{gh}de_g,$$

where e_g is a delta-Kronecker function: 1 at g and 0, elsewhere. The right coaction on them is

$$\Delta_R\chi^g=\sum_h \chi^{hgh^{-1}}\otimes e_h.$$

Using the spectral triple approach, we may calculate the χ^g forms as operators on our Hilbert space. It appears that

$$(\chi^g h)_{ij}=D_{ij,(ig^{-1})l}h_{(ig^{-1})l}.$$

Because of the properties of D we see that it is only possible that $j=l$ and therefore finally

$$(\chi^g h)_{ij}=D_{ij,(ig^{-1})j}h_{(ig^{-1})j},$$

so it is a collection of operators from $H_{(ig^{-1})j}\to H_{ij}$. Of course, if $\gamma_{ij}\gamma_{(ig^{-1})j}=1$ they all vanish because D acts only between spaces of different chirality.

What we have obtained so far is the representation of left-invariant forms as operators on our Hilbert space. We see that all χ^g which do not vanish are linearly independent (since they act between different Hilbert subspaces).

If we demand that the calculus is bicovariant, we immediately see that if for any g, χ^g vanishes, then for every h, $\chi^{hgh^{-1}}$ must vanish as well. Of course, for commutative groups such a condition is void; however, for non-Abelian ones it leads to certain restrictions on possible forms of D, which we make precise in the following lemma:

Lemma 14: If the spectral-triple calculus on $\mathbb{C}^n$ with a structure of the discrete group G is bicovariant, then

$$\text{if } \chi^g\neq 0, \quad \text{then } \forall i\in G\,\exists j\in G: D_{ij,(ig)j}\neq 0,$$

$$\text{if } \chi^g=0, \quad \text{then } \forall i\in G\,\forall j\in G: D_{ij,(ig)j}=0.$$

The proof is an immediate consequence of previous observations.

1. Function algebra on S_3

Consider as an example the six-dimensional space S_3. We shall check what restrictions on the spectral triple and the Dirac operator are set by the requirements that the corresponding calculus is bicovariant.

First, let us take the universal calculus. It appears that there exists a spectral triple which gives such a calculus. The simplest example is given by the following intersection form,

$$\begin{pmatrix} -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 \end{pmatrix},$$

and is 36-dimensional! It is quite easy to see that this must be the case—for each χ^g not to vanish, the Dirac operator must link (for every i) some $H_{(ig^{-1})j}$ with H_{ij}. For the intersection matrix it means that for every row there always exists an entry which has an opposite sign to an entry in the same column of another row. It is easy to verify that such a matrix is nondegenerate so that the Poincaré duality axiom is fulfilled. Of course, to get the universal calculus, all possible Dirac operator elements must not vanish.

A more interesting case is when we require that the calculus is the lowest-dimensional bicovariant. For S_3 such a calculus is generated by χ^{ab} and χ^{ba}, where a and b denote the generators of S_3.

If we order the elements of S_3 as $a,b,c,1,ab,ba$, the example of a spectral triple which gives the desired calculus is defined by

$$\begin{pmatrix} 1 & 1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}.$$

Again, all the possible elements of the Dirac operator (components which act between spaces of different chirality) must not vanish. Of course if we slightly change the intersection form (introduce, for instance, other nonzero entries in the intersection matrix) the bicovariance (two-dimensional) shall also be preserved provided that the additional allowed Dirac operator components vanish!

For physical interpretation this would mean that the symmetry requires some of the fermion particles in the model to be massless, whereas in the spectral triple presented above all mass matrices must be nonzero.

Noncommutative finite algebras cannot carry such a simple group structure and therefore we have to look for non-Abelian Hopf algebras. This itself is an interesting topic and very little is known about the general classification of such (semisimple) objects. The easiest examples come from group algebras and as we try to make our examples comprehensible, we shall use the simplest one of them, the group algebra $\mathbb{C}S_3$.

B. Group algebra S_3

We shall begin here with some generalities on Hopf and bicovariant differential structures on group algebras and then we shall present the interesting example of S_3.

1. Generalities on group algebras

The group algebra $\mathbb{C}G$ has a natural structure of a Hopf algebra with the following coproduct, counit, and antipode:

$$\Delta g = g \otimes g, \quad \varepsilon(g) = 1, \quad Sg = g^{-1}.$$

The adjoint coaction is trivial:

$$\mathrm{ad}(g) = g \otimes 1.$$

The Haar invariant measure μ is just $\mu(g) = 0$ for g different from the neutral element of the group.

Finally let us state some observations on the differential calculi.

Observation 9: The differential calculi on group algebras are always inner, i.e., there exists a one-form χ such that $dg = [g, \chi]$.

It is sufficient to take

$$\chi = -\frac{1}{N_G}\left(\sum_h h^{-1} dh\right).$$

Then

$$\begin{aligned}[g, \chi] &= -\frac{1}{N_G}\sum_h (gh^{-1}dh - h^{-1}dhg) \\ &= -\frac{1}{N_G}\sum_h (gh^{-1}dh - h^{-1}d(hg) - dg) \\ &= \frac{1}{N_G}\sum_h dg = dg.\end{aligned}$$

Observation 10: Since $\mathbb{C}G$ is cocommutative, every left-covariant calculus is bicovariant. There exists a 1:1 correspondence between bicovariant calculi and representations of the group G.

The first remark is obvious, so let us concentrate on the latter. Suppose we have a representation of G on some vector space V. We may define Ω_1 as a right-module $\mathbb{C}G \otimes V$, with the multiplication from the right by elements of $\mathscr{A}$ only on the first component of the tensor product. Then the following rule (it is sufficient to define it only for the generators) makes Ω_1 a bimodule:

$$h(g \otimes \nu) = (hg) \otimes (\pi(h)\nu).$$

Of course, Ω_1 is a bicovariant bimodule, with all the elements of type $1 \otimes \nu$ being left and right invariant. We already know that every differential calculus is inner and, moreover, it is easy to check that the one-form χ is left and right invariant. Therefore, to construct the calculus, it is sufficient to choose a $\chi = 1 \otimes \nu \in 1 \otimes V$ and define $dg = [g, \chi]$. It remains to show that $\mathrm{lm}\, d$ generates Ω_1. This is equivalent to verifying whether $\{\nu - \pi(h)\nu\}$, $h \in G$, generates the whole space V. Note that the representation π could be linearly extended on the whole group algebra $\mathbb{C}G$. Then the latter requirement tells us that the kernel of the counit acting on ν spans the whole space V.

The proof of the converse (every bicovariant calculus provides us with a representation of the group) is obvious.

Observation 11: The central part of $\Omega^1(\mathscr{A})$ consists of all one-forms of the type $\omega = \Sigma_g g \otimes \nu_g$, where $\nu_g \in V$ are vectors satisfying for every $g, h \in G$:

$$\pi(h)\nu_g = \nu_{hgh^{-1}}. \tag{11}$$

Of course, this condition may be void for some representation of the group G (in particular for a nontrivial one-dimensional representation).

Proof: Every one-form can be written as $\omega=\Sigma_g g\otimes\nu_g$ for some $\nu_g\in V$. If we require that $h\omega=\omega h$ for every $h\in G$, then it is easy to verify that (11) follows. ■

As an example we shall consider the smallest noncommutative group algebra $\mathbb{C}S_3$.

2. The group algebra $\mathbb{C}S_3$

The group S_3 has two generators a,b with the following multiplication rules:

$$a^2=e, b^2=e, aba=bab.$$

It is convenient to name the element $aba=c$, $c^2=e$ and use the rules for multiplication between a,b,c:

$$ab=bc,\quad ac=ba,\quad bc=ca.$$

As an algebra $\mathbb{C}S_3$ is isomorphic to the algebra $M_2(C)\oplus\mathbb{C}\oplus\mathbb{C}$. We shall present the form of this isomorphism i for the generators a,b:

$$i(a)=\begin{pmatrix} 1 & 0 & & \\ 0 & -1 & & \\ & & 1 & \\ & & & -1 \end{pmatrix},\quad i(b)=\begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} & & \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & & \\ & & 1 & \\ & & & -1 \end{pmatrix}.$$

The Haar measure on the Hopf algebra is defined as a linear map which satisfies

$$1\mu(f)=(\mathrm{id}\otimes\mu)\Delta f=(\mu\otimes\mathrm{id})\Delta f.$$

On any group algebra there exists only one such measure, which is $\mu(g)=0$, $g\neq 0$ and $\mu(1)=1$ (normalized). For our four-dimensional representation $[M\in M_2(C), p,q\in C]$ it is given by

$$\mu\begin{pmatrix} M & & \\ & p & \\ & & q \end{pmatrix}=\frac{1}{3}\,\mathrm{Tr}\,M+\frac{1}{6}\,p+\frac{1}{6}\,q.$$

Before we start constructing spectral triples let us have a look at the covariant differential calculi, constructed according to the general scheme. Therefore, first we have to give the representations of the group S_3:

dimension	$\pi(a)$	$\pi(b)$
1^*	1	1
1	-1	-1
2	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$

where 1^* is the trivial representation. Every representation of higher dimension is a direct sum of them.

(i) One-dimensional calculus: For the one-dimensional calculus we take the representation 1 (the trivial one gives $d\equiv 0$).

There is only one generating form χ and the bimodule rules follow from the one-dimensional representation of S_3. Since there is only one such representation $a=b=-1$, we obtain

$$a\chi = -\chi a, \quad b\chi = -\chi b,$$

and the external derivative is

$$da = a\chi, \quad db = b\chi, \quad dc = c\chi.$$

Notice that

$$d(ab) = 0, \quad d(ba) = 0.$$

Such a calculus is extendible to a higher-order calculus with $d^2 = 0$ provided that we set $d\chi = -\chi\chi$.

This calculus has an interesting splitting property:

Observation 12: Let $A = M_2(C)$ and $B = C \oplus C$, $\mathbb{C}S_3 = A \oplus B$. Then the split short exact sequence of algebras,

$$0 \to A \to \mathbb{C}S_3 \to B \to 0,$$

extends to a split short exact sequence of differential modules,

$$0 \to \Omega(A) \to \Omega(\mathbb{C}S_3) \to \Omega(B) \to 0,$$

where $\Omega(\mathbb{C}S_3)$ is the one-dimensional bicovariant calculus, and $\Omega(A)$ and $\Omega(B)$ are its restriction to the corresponding subalgebras.

This tells us immediately that no spectral triple data could generate this calculus. Indeed, since the Dirac operator D can act only between spaces of different left (or right) representation of the algebra, such a splitting of differential bimodules is not possible.

(ii) Two-dimensional calculus: As we have pointed out there exists only one nontrivial two-dimensional representation of S_3 (here nontrivial means that there exists a vector ν such that (ker ε) ν spans the whole representation space, which, for instance, does not hold for $1 \oplus 1$). Various possibilities for the choice of ν would correspond to differential calculi related by automorphism of the group algebra.

If we choose the above presented matrix form of a and b and $\nu = (1, -\sqrt{3})$ we arrive at the following relations:

$$\chi^1 a = -a\chi^1, \quad \chi^2 b = -b\chi^2,$$

$$\chi^1 b = b(\chi^1 - \chi^2), \quad \chi^2 a = a(\chi^2 - \chi^1),$$

where $\chi^1 = ada$ and $\chi^2 = bdb$.

The same relations expressed in terms of dg become

$$(da)b = c(da) - a(db), \quad (db)a = c(db) - b(da).$$

Then $dc = (da)ba + a(da)b + ab(da) = 0$.

All other two-dimensional differential calculi could be obtained through an automorphism of the algebra.

Note that for this calculus we may construct a one-form in the center of $\Omega^1(\mathcal{A})$. For instance, the vector $a\nu_a + b\nu_b - c(\nu_a + \nu_b)$, where ν_a and ν_b are eigenvectors (to eigenvalue 1) of $\pi(a)$ and $\pi(b)$, respectively, gives such a form. Using χ^a and χ^b we can write it as

$$(2a - b - c)\chi^2 + (2b - a - c)\chi^1,$$

so, again this rules out the compatibility between bicovariance and spectral-triple calculus for $\mathbb{C}S_3$.

Our result extends also for higher-dimensional calculi, as all higher-order representations of S_3 contain the two-dimensional part (and, if not, they must be composed out of 1 and 1*) and therefore one may repeat the arguments.

Thus, we may state a simple no-go lemma:

Lemma 15: For $\mathbb{C}S_3$ the differential structure from spectral triple construction cannot carry a bicovariance symmetry.

Knowing that we may now concentrate on spectral triples and the symmetry of the measure.

C. Spectral triples for $\mathbb{C}S_3$

Following the general construction scheme presented earlier we shall now turn our attention to an example of low-dimensional spectral triples for the algebra $\mathbb{C}S_3$. We will be interested in the nontrivial case, which admits a nonzero Dirac operator.

It is easy to verify that the lowest-dimensional spectral triple is given by the intersection form:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The Hilbert space has the dimension 7 (it is odd, as it contains a vector in the subspace H_{33} such that $J\nu=\nu$!).

Instead of the Dirac operator it is more convenient to use the one-form ξ. Since the Dirac operator D can only have nonzero entries $D_{12,32}$ and $D_{23,33}$ (and, respectively, their Hermitian conjugates and J-conjugates) we have the following ξ (where the Hilbert spaces on which ξ acts are ordered as follows: $H_{12},H_{32},H_{23},H_{33}$):

$$\pi(\xi)=\begin{pmatrix} 0 & 0 & x & 0 & 0 \\ 0 & 0 & y & 0 & 0 \\ x^* & y^* & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & z \\ 0 & 0 & 0 & z^* & 0 \end{pmatrix}.$$

Observation 13: The operator $\pi(\xi)\pi(\xi)$ is in $\pi(\mathcal{A})$ provided that $xx^*+yy^*=zz^*$. In any case it commutes with the image of the center of the algebra and therefore it follows from Lemma 11 that $\pi(\xi)^2=\pi(\Xi)$ and the calculus is inner:

$$\pi(\xi)^2=\begin{pmatrix} xx^* & xy^* & 0 & 0 & 0 \\ yx^* & yy^* & 0 & 0 & 0 \\ 0 & 0 & xx^*+yy^* & 0 & 0 \\ 0 & 0 & 0 & zz^* & 0 \\ 0 & 0 & 0 & 0 & zz^* \end{pmatrix}.$$

We shall finish the investigation by writing the *measure operator*, which gives the Haar measure on $\mathbb{C}S_3$.

Observation 14: If $Z=\frac{1}{3}P_1+\frac{1}{18}P_2+\frac{1}{12}P_3$, where P_i denote the corresponding projections in the algebra, then, for every $a\in\mathcal{A}$,

$$\mathrm{Tr}_Z\pi(a)=\mathrm{Tr}\,(\pi(Z)\pi(a))$$

is the Haar measure on $\mathbb{C}S_3$.

One may now use this deformed scalar product on forms to calculate the Yang–Mills action. Note that $[D,\pi(Z)]\neq 0$ and the results would differ from the ones obtained using the *generic* scalar product. For physical models this would result in different mass relations between the gauge bosons in the theory based upon such a triple.

IV. CONCLUSIONS

The world of noncommutative geometry is much bigger than that of classical geometry. In the classical differential geometry, once given a smooth manifold we already know its differential

bundle. Given a Lie group we have a natural notion of the Lie group action on differential structures. All this seems to be lost when we enter the noncommutative world. For a single algebra we have many choices of differential structures and there is no universal rule to tell us which one to choose. Symmetries, or at least group symmetries, are in many cases no longer present or drastically reduced.

The notion of spectral triples is an attempt to bring some order into our understanding of noncommutative differential geometry, by specifying what structures must appear in the models to make them a real *geometry*.

What we have tried to investigate in this paper are the restrictions, which are introduced by this order to the realms of zero-dimensional geometry. We have tried even more to answer the question whether the proposed structures admit and restrict the possible zero-dimensional noncommutative symmetries.

We have classified finite spectral triples and the allowed Dirac operators. This, for instance, enables us to state that the Standard Model Dirac operator is not the most general one, and in fact one may consider an identical spectral triple with some additional components of the Dirac operator, which would link left-handed antileptons with right-handed quarks. Of course, the assumption that such a component exists might have some deep consequences for the model, as it could potentially lead to the breaking of the SU(3) strong symmetry. Another consequence (mentioned originally by D. Testard, see Ref. 10) is that Poincaré duality enforces that there are no right-handed neutrinos. This is correct provided that we add a pair of them (particle and antiparticle). However, it is easy to verify that adding only one right-handed Majorana particle (we doubt whether the name *neutrino* would be justified) still does preserve the Poincaré duality. These and other observations concerning the standard model will be the topic of our forthcoming paper.[2]

The symmetries of spectral triples are a far more complicated problem. We have seen that only in some cases there exists an extension (on the differential level at least) of the Hopf algebra structure into the spectral triple theory. Whether this is a general pattern is difficult to say; however, we believe that symmetries can also be realized in the zero-dimensional case. The examples we have analyzed do not give a conclusive answer; what they suggest, however, is that the symmetry restrictions would be much stronger and of different type than in classical geometry.

ACKNOWLEDGMENTS

We would like thank Florian Scheck for valuable discussions and J. M. Warzecha for remarks on the manuscript.

While preparing a final version of this paper we learned that similar results were obtained by T. Krajewski,[11] whom we would also like to thank for comments and remarks.

[1] A. Connes, ''Gravity coupled with matter and the foundation of noncommutative geometry,'' Commun. Math. Phys. **182**, 155 (1996).
[2] M. Paschke, F. Scheck, and A. Sitarz, ''Can (noncommutative) geometry accommodate leptoquarks?,'' hep-th/9709009.
[3] *Encyclopaedia of Mathematical Sciences, Vol. 18, Algebra II* (Springer-Verlag, Berlin, 1991).
[4] A. Sitarz, ''Noncommutative Differential Geometry and Connections on Simplicial Manifolds,'' Acta Phys. Pol. B **28**, 1039 (1997).
[5] S. Majid, *Foundations of Quantum Group Theory* (Cambridge U.P., Cambridge, 1995).
[6] A. Connes and A. Chamseddine, ''The Spectral Action Principle,'' Commun. Math. Phys. **186**, 731 (1997).
[7] A. Connes, ''Noncommutative Geometry and Reality,'' J. Math. Phys. **36**, 6194 (1995).
[8] R. Coquereaux, ''On the finite dimensional quantum group $M_3+(M_{2|1}(\Lambda^2))_0$,'' hep-th/9610114.
[9] K. Bresser, A. Dimakis, F. Mueller-Hoissen, and A. Sitarz, ''Noncommutative Geometry of Finite Groups,'' J. Phys. A **29**, 2705 (1996).
[10] C. P. Martin, J. M. Gracia-Bondia, and J. C. Varilly, ''The Standard Model as a noncommutative geometry: the low energy regime,'' hep-th/9605001, to appear in Physics Reports (1997).
[11] T. Krajewski, ''Classification of finite spectral triples,'' preprint CPT-96/P.3409, hep-th/9701081.

The trace formulas yield the inverse metric formula

Ronaldo Rodrigues Silva[a)]
Centro Brasileiro de Pesquisas Físicas, Rua Dr. Xavier Sigaud 150, 22290-180, Rio de Janeiro, RJ, Brazil

(Received 12 May 1998; accepted for publication 8 July 1998)

It is a well-known fact that the first and last nontrivial coefficients of the characteristic polynomial of a linear operator are, respectively, its trace and its determinant. This work shows how to compute recursively all the coefficients as polynomial functions in the traces of the successive powers of the operator. With the aid of Cayley–Hamilton's theorem the trace formulas provide a rational formula for the resolvent kernel and an operator-valued null identity for each finite dimension of the underlying vector space. The four-dimensional resolvent formula allows an algebraic solution of the inverse metric problem in general relativity. © *1998 American Institute of Physics.* [S0022-2488(98)01311-5]

I. INTRODUCTION

The eigenvalue problem arises in a variety of different branches of mathematical physics. For instance, it is well-known that quantum systems reach stationary states that are the eigenvectors of a suitable linear operator defined throughout a complex vector space representing the physical quantum states.

In the infinite dimensional case real and complex analytical methods have been developed to compute the eigenvalues of a linear operator. We mention, for instance, global variational methods, which consist of investigating stationary levels of a suitable energy functional, and local perturbative methods by means of complex analytic continuation.

On the other hand, if the underlying vector space is finite dimensional the eigenvalue problem becomes an algebraically well-posed problem by means of the characteristic polynomial. It is exactly this feature that justifies the algebraic approach in this work. Following this general device, the algebraic environment is set by the complex number field. After working out a few tools, the most important of these being Newton's identities, and setting very fundamental concepts and basic results in complex linear algebra, we succeed in achieving the mathematical core of this work: a recursive algorithm to compute the coefficients of the characteristic polynomial as algebraic functions in the traces of the successive powers of the linear operator.

With the aid of Cayley–Hamilton's theorem the trace formulas already obtained provide an operator-valued null identity for each finite dimension of the underlying vector space. As a by-product it sheds light on the algebraic structure of the associative algebra of complex linear operators. The computational skill just attained is enough to yield an operator-valued polynomial with rational coefficients for the finite-dimensional resolvent kernel, which improves a known result by revealing its rational dependence with respect to the spectral variable, as well as with respect to the linear operator.

In the context of general relativity, the four-dimensional characteristic formula endows a polynomial expression for the volume scalar density. Furthermore, the four-dimensional resolvent formula yields a tensor-valued third-degree polynomial for the inverse metric, thus avoiding the computational and formal drawbacks of the Neumann series.

II. NEWTON'S IDENTITIES

The fundamental theorem of algebra[1] together with Euclid's division algorithm imply algebraic closure[2] of the complex number field; this basic result is encompassed without proof by

[a)]Electronic mail: ronaldor@cat.cbpf.br

 © 1998 American Institute of Physics

Theorem 0: *If* $p(z)=z^n+D_1z^{n-1}+\cdots+D_{n-1}z+D_n$ *is a polynomial with complex coefficients* D_k *then there exist complex numbers* $\lambda_1,\lambda_2,\ldots,\lambda_n$, *called roots of p, such that* $p(z)=(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n)$.

As a consequence of the identity principle, the following relations between coefficients and roots hold:

$$\begin{aligned}
-D_1&=\lambda_1+\lambda_2+\cdots+\lambda_n,\\
+D_2&=\lambda_1\lambda_2+\cdots+\lambda_{n-1}\lambda_n,\\
&\vdots,\\
(-)^kD_k&=\sum\ k\text{-products of }\lambda\text{'s},\\
&\vdots,\\
(-)^nD_n&=\lambda_1\lambda_2\cdots\lambda_n.
\end{aligned}\tag{1}$$

The right-hand side of (1) defines the elementary symmetric functions[3] in the variables $\lambda_1,\lambda_2,\ldots,\lambda_n$. Next to them, the most important symmetric functions are the sums of like powers:

$$\begin{aligned}
T_1&=\lambda_1+\lambda_2+\cdots+\lambda_n,\\
T_2&=\lambda_1^2+\lambda_2^2+\cdots+\lambda_n^2,\\
&\vdots,\\
T_k&=\lambda_1^k+\lambda_2^k+\cdots+\lambda_n^k,\\
&\vdots
\end{aligned}\tag{2}$$

D_k and T_k, besides being symmetric, are homogeneous functions of degree k in the variables $\lambda_1,\lambda_2,\ldots,\lambda_n$. We shall derive a set of recursive relations connecting D's and T's by which the λ's are eliminated from (1) and (2).

To pursue this goal we need two lemmas. The first one is an improved version of the remainder theorem.

Proposition 1: If $p(z)=z^n+D_1z^{n-1}+\cdots+D_{n-1}z+D_n$ *is a polynomial with complex coefficients* D_k *and* λ *is a complex number then*

$$\frac{p(z)-p(\lambda)}{z-\lambda}=z^{n-1}+(\lambda+D_1)z^{n-2}+(\lambda^2+D_1\lambda+D_2)z^{n-3}+\cdots+(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1}).$$

Proof: By induction on the degree. For $n=1$, $p(z)-p(\lambda)=(z+D_1)-(\lambda+D_1)=z-\lambda$. *For generic n,*

$$\begin{aligned}
p(z)-p(\lambda)&=(z^n+D_1z^{n-1}+\cdots+D_{n-1}z+D_n)-(\lambda^n+D_1\lambda^{n-1}+\cdots+D_{n-1}\lambda+D_n)\\
&=z(z^{n-1}+D_1z^{n-2}+\cdots+D_{n-1})-\lambda(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1})\\
&=z((z^{n-1}+D_1z^{n-2}+\cdots+D_{n-1})-(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1}))\\
&\quad+(z-\lambda)(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1}).
\end{aligned}$$

From the inductive hypothesis the last expression reads

$$
\begin{aligned}
&z(z-\lambda)(z^{n-2}+(\lambda+D_1)z^{n-3}+\cdots+(\lambda^{n-2}+D_1\lambda^{n-3}+\cdots+D_{n-2}))\\
&\quad+(z-\lambda)(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1})\\
&\quad=(z-\lambda)(z^{n-1}+(\lambda+D_1)z^{n-2}+\cdots+(\lambda^{n-2}+D_1\lambda^{n-3}+\cdots+D_{n-2})z\\
&\qquad+(\lambda^{n-1}+D_1\lambda^{n-2}+\cdots+D_{n-1})).
\end{aligned}
$$

The second lemma is a contribution of differential calculus to algebra.

Proposition 2: If $\lambda_1,\lambda_2,\ldots,\lambda_n$ are the complex roots of the nth-degree polynomial p and p' is its derivative polynomial, then

$$
p'(z)=\sum_{k=1}^{n}\frac{p(z)}{z-\lambda_k}.
$$

Proof: From $p(z)=d_0(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n)$, it follows

$$
\begin{aligned}
p'(z)&=d_0(z-\lambda_2)(z-\lambda_3)\cdots(z-\lambda_n)+d_0(z-\lambda_1)(z-\lambda_3)\cdots(z-\lambda_n)\\
&\quad+\cdots+d_0(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_{n-1})\\
&=\frac{p(z)}{z-\lambda_1}+\frac{p(z)}{z-\lambda_2}+\cdots+\frac{p(z)}{z-\lambda_n}.
\end{aligned}
$$

Thus, we are able to get

Theorem 1: *(Newton's formulas[4]) If $\lambda_1,\lambda_2,\ldots,\lambda_n$ are the complex roots of the polynomial $p(z)=z^n+D_1z^{n-1}+D_2z^{n-2}+\cdots+D_{n-1}z+D_n$ with complex coefficients D_k and $T_k=\lambda_1^k+\lambda_2^k+\cdots+\lambda_n^k$ then*

$$
T_k+D_1T_{k-1}+\cdots+D_{k-1}T_1+kD_k=0,\quad k=1,2,\ldots,n.
$$

Proof: Propositions 1 and 2 yield

$$
\begin{aligned}
&nz^{n-1}+(n-1)D_1z^{n-2}+(n-2)D_2z^{n-3}+\cdots+D_{n-1}\\
&\quad=\sum_{k=1}^{n}\{z^{n-1}+(\lambda_k+D_1)z^{n-2}+(\lambda_k^2+D_1\lambda_k+D_2)z^{n-3}\\
&\qquad+\cdots+(\lambda_k^{n-1}+D_1\lambda_k^{n-2}+\cdots+D_{n-2}\lambda_k+D_{n-1})\}\\
&\quad=nz^{n-1}+(T_1+nD_1)z^{n-2}+(T_2+D_1T_1+nD_2)z^{n-3}\\
&\qquad+\cdots+(T_{n-1}+D_1T_{n-2}+\cdots+D_{n-2}T_1+nD_{n-1}).
\end{aligned}
$$

Equating coefficients we obtain

$$
\begin{aligned}
(n-1)D_1&=T_1+nD_1,\\
(n-2)D_2&=T_2+D_1T_1+nD_2,\\
&\vdots,\\
D_{n-1}&=T_{n-1}+D_1T_{n-2}+\cdots+D_{n-2}T_1+nD_{n-1},
\end{aligned}
$$

from which follow the $(n-1)$ first relations to be shown. Since each λ_k is a root of p it verifies $0=\lambda_k^n+D_1\lambda_k^{n-1}+\cdots+D_{n-1}\lambda_k+D_n$. The addition of these relations yields

$$
0=\sum_{k=1}^{n}\{\lambda_k^n+D_1\lambda_k^{n-1}+\cdots+D_{n-1}\lambda_k+D_n\}=T_n+D_1T_{n-1}+\cdots+D_{n-1}T_1+nD_n,
$$

which is the remaining formula to be proved.

III. THE TRACE FORMULAS

The characteristic polynomial[5] of a complex linear operator $\mathbf{T}$ in a finite dimensional complex vector space $\mathbb{V}$ is defined by $p(z)=\det(z\mathbf{I}-\mathbf{T})$, where $\mathbf{I}$ is the identity operator in $\mathbb{V}$. The degree of p as a polynomial in the complex variable z equals the dimension of $\mathbb{V}$ as a complex vector space; its roots are called the characteristic values of $\mathbf{T}$ and the set of characteristic values is called the spectrum of $\mathbf{T}$.

These definitions and properties lead to

Proposition 3: The trace of a complex linear operator $\mathbf{T}$ *equals the sum of its characteristic values counted with multiplicities.*

Proof: $\det(z\mathbf{I}-\mathbf{T})=(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n)$. *By the matrix representation* T_j^i *of* $\mathbf{T}$ *we cast the left-hand side as* $z^n-(T_1^1+T_2^2+\cdots+T_n^n)z^{n-1}+$ *terms of order* $\leqslant n-2$. *In the right-hand side we have* $z^n-(\lambda_1+\lambda_2+\cdots+\lambda_n)z^{n-1}+$ *terms of order* $\leqslant n-2$. *Equating coefficients we obtain*

$$T_1^1+T_2^2+\cdots+T_n^n=\lambda_1+\lambda_2+\cdots+\lambda_n.$$

The roots of unity in the complex field yield the following factorization lemma in the associative algebra of complex linear operators.

Proposition 4: If $\mathbf{T}$ *and* $\mathbf{W}$ *are commuting complex linear operators and* $\theta=e^{i2\pi/k}=\cos(2\pi/k)+i\sin(2\pi/k)$ *with* k *a positive integer, then*

$$\mathbf{T}^k-\mathbf{W}^k=(\mathbf{T}-\mathbf{W})(\mathbf{T}-\theta\mathbf{W})(\mathbf{T}-\theta^2\mathbf{W})\cdots(\mathbf{T}-\theta^{k-1}\mathbf{W}).$$

Proof: As $\mathbf{T}$ *and* $\mathbf{W}$ *commute we collect terms as*

$$\begin{aligned}&(\mathbf{T}-\mathbf{W})(\mathbf{T}-\theta\mathbf{W})(\mathbf{T}-\theta^2\mathbf{W})\cdots(\mathbf{T}-\theta^{k-1}\mathbf{W})\\&\quad=\mathbf{T}^k+d_1\mathbf{W}\mathbf{T}^{k-1}+d_2\mathbf{W}^2\mathbf{T}^{k-2}+\cdots+d_{k-1}\mathbf{W}^{k-1}\mathbf{T}+d_k\mathbf{W}^k,\end{aligned}$$

where the complex numbers $d_1,d_2,\ldots,d_k$ *are the coefficients of the polynomial* $(z-1)(z-\theta)\times(z-\theta^2)\cdots(z-\theta^{k-1})=z^k-1$, *since* $(\theta^j)^k=((e^{i2\pi/k})^j)^k=1$ *for each* $j=0,1,\ldots,k-1$. *Therefore* $d_1=d_2=\cdots=d_{k-1}=0$ *and* $d_k=-1$.

From the multiplicative and homogeneity properties of determinants we get

Proposition 5: If $\lambda_1,\lambda_2,\ldots,\lambda_n$ *are the characteristic values of the complex linear operator* $\mathbf{T}$ *then* $\lambda_1^k,\lambda_2^k,\ldots,\lambda_n^k$ *are the characteristic values of the linear operator* $\mathbf{T}^k$.

Proof: For the particular case where $\mathbf{W}=w\mathbf{I}$, *proposition 4 yields*

$$\begin{aligned}\det(\mathbf{T}^k-w^k\mathbf{I})&=\det(\mathbf{T}-w\mathbf{I})\det(\mathbf{T}-\theta w\mathbf{I})\cdots\det(\mathbf{T}-\theta^{k-1}w\mathbf{I})\\&=(\lambda_1-w)\cdots(\lambda_n-w)(\lambda_1-\theta w)\cdots(\lambda_n-\theta w)\cdots(\lambda_1-\theta^{k-1}w)\cdots(\lambda_n-\theta^{k-1}w)\\&=(\lambda_1-w)(\lambda_1-\theta w)\cdots(\lambda_1-\theta^{k-1}w)\cdots(\lambda_n-w)(\lambda_n-\theta w)\cdots(\lambda_n\theta^{k-1}w)\\&=(\lambda_1{}^k-w^k)\cdots(\lambda_n{}^k-w^k).\end{aligned}$$

Thus $\det(z\mathbf{I}-\mathbf{T}^k)=(z-\lambda_1{}^k)\cdots(z-\lambda_n{}^k)$.

Collecting the former results we are ready to achieve the *recursive formulas* enclosed by

Theorem 2: *If* $p(z)=z^n+D_1z^{n-1}+\cdots+D_{n-1}z+D_n$ *is the characteristic polynomial of the complex linear operator* $\mathbf{T}$ *and* $T_k=\mathrm{trace}\,(\mathbf{T}^k)$ *then*

$$T_k+D_1T_{k-1}+\cdots+D_{k-1}T_1+kD_k=0,\quad k=1,2,\ldots,n.$$

Proof: It is a straightforward consequence of theorem 1 together with propositions 3 and 5.

The key content of theorem 2 shall be emphasized by the following *trace formulas statement: The coefficients of the characteristic polynomial of a linear operator can be recursively computed as polynomial functions in the traces of its successive powers.*

The *trace formulas* $D_k=D_k(T_1,T_2,\ldots,T_{k-1},T_k)$ for the coefficients of the characteristic polynomial are listed below for k up to 8.

$$
\begin{aligned}
D_1 &= -T_1,\\
D_2 &= \tfrac{1}{2}T_1^2-\tfrac{1}{2}T_2,\\
D_3 &= -\tfrac{1}{6}T_1^3+\tfrac{1}{2}T_1T_2-\tfrac{1}{3}T_3,\\
D_4 &= \tfrac{1}{24}T_1^4-\tfrac{1}{4}T_1^2T_2+\tfrac{1}{3}T_1T_3+\tfrac{1}{8}T_2^2-\tfrac{1}{4}T_4,\\
D_5 &= -\tfrac{1}{120}T_1^5+\tfrac{1}{12}T_1^3T_2-\tfrac{1}{6}T_1^2T_3-\tfrac{1}{8}T_1T_2^2+\tfrac{1}{4}T_1T_4+\tfrac{1}{6}T_2T_3-\tfrac{1}{5}T_5,
\end{aligned}
\tag{3}
$$

$$
D_6=\tfrac{1}{720}T_1^6-\tfrac{1}{48}T_1^4T_2+\tfrac{1}{18}T_1^3T_3+\tfrac{1}{16}T_1^2T_2^2-\tfrac{1}{8}T_1^2T_4-\tfrac{1}{6}T_1T_2T_3+\tfrac{1}{5}T_1T_5-\tfrac{1}{48}T_2^3+\tfrac{1}{8}T_2T_4+\tfrac{1}{18}T_3^2-\tfrac{1}{6}T_6,
$$

$$
\begin{aligned}
D_7=&-\tfrac{1}{5040}T_1^7+\tfrac{1}{240}T_1^5T_2-\tfrac{1}{72}T_1^4T_3-\tfrac{1}{48}T_1^3T_2^2+\tfrac{1}{24}T_1^3T_4+\tfrac{1}{12}T_1^2T_2T_3-\tfrac{1}{10}T_1^2T_5+\tfrac{1}{48}T_1T_2^3-\tfrac{1}{8}T_1T_2T_4\\
&-\tfrac{1}{18}T_1T_3^2+\tfrac{1}{6}T_1T_6-\tfrac{1}{24}T_2^2T_3+\tfrac{1}{10}T_2T_5+\tfrac{1}{12}T_3T_4-\tfrac{1}{7}T_7,
\end{aligned}
$$

$$
\begin{aligned}
D_8=&\tfrac{1}{40\,320}T_1^8-\tfrac{1}{1440}T_1^6T_2+\tfrac{1}{360}T_1^5T_3+\tfrac{1}{192}T_1^4T_2^2-\tfrac{1}{96}T_1^4T_4-\tfrac{1}{36}T_1^3T_2T_3+\tfrac{1}{30}T_1^3T_5-\tfrac{1}{96}T_1^2T_2^3+\tfrac{1}{16}T_1^2T_2T_4\\
&+\tfrac{1}{36}T_1^2T_3^2-\tfrac{1}{12}T_1^2T_6+\tfrac{1}{24}T_1T_2^2T_3-\tfrac{1}{10}T_1T_2T_5-\tfrac{1}{12}T_1T_3T_4+\tfrac{1}{7}T_1T_7+\tfrac{1}{384}T_2^4-\tfrac{1}{32}T_2^2T_4-\tfrac{1}{36}T_2T_3^2\\
&+\tfrac{1}{12}T_2T_6+\tfrac{1}{15}T_3T_5+\tfrac{1}{32}T_4^2-\tfrac{1}{8}T_8.
\end{aligned}
$$

The *characteristic formulas* for dimensions up to four are written down:

$$
\begin{aligned}
\det(z\mathbf{I}-\mathbf{T}) &= z-T_1,\\
\det(z\mathbf{I}-\mathbf{T}) &= z^2-T_1z+\tfrac{1}{2}(T_1^2-T_2),\\
\det(z\mathbf{I}-\mathbf{T}) &= z^3-T_1z^2+\tfrac{1}{2}(T_1^2-T_2)z-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3),\\
\det(z\mathbf{I}-\mathbf{T}) &= z^4-T_1z^3+\tfrac{1}{2}(T_1^2-T_2)z^2-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3)z\\
&\quad+\tfrac{1}{24}(T_1^4-6T_1^2T_2+8T_1T_3+3T_2^2-6T_4).
\end{aligned}
\tag{4}
$$

IV. NULL IDENTITIES

The polynomial $p(z)=d_0z^n+d_1z^{n-1}+\cdots+d_{n-1}z+d_n$ with complex coefficients d_k is said to annihilate the complex linear operator $\mathbf{T}$ if $p(\mathbf{T})=d_0\mathbf{T}^n+d_1\mathbf{T}^{n-1}+\cdots+d_{n-1}\mathbf{T}+d_n\mathbf{I}=\mathbf{0}$, the identically null operator.

The knowledge of the principal ideal[6] of polynomials that annihilate a linear operator $\mathbf{T}$ is essential to attaining computational skill in the associative algebra generated by $\mathbf{T}$. To ensure this goal we state without proof one of the fundamental results in linear algebra.

Theorem 3: *(Cayley–Hamilton[7]) The characteristic polynomial of a linear operator annihilates it.*

Joining Cayley–Hamilton's theorem with the trace formulas statement we conclude that *for each finite dimension of the underlying vector space there is a fundamental null identity in the associative algebra of linear operators*. For instance, in dimensions up to four, the characteristic formulas (4) yield the following *null identities*:

$$
\begin{aligned}
&\mathbf{T}-T_1\mathbf{I}=\mathbf{0},\\
&\mathbf{T}^2-T_1\mathbf{T}+\tfrac{1}{2}(T_1^2-T_2)\mathbf{I}=\mathbf{0},\\
&\mathbf{T}^3-T_1\mathbf{T}^2+\tfrac{1}{2}(T_1^2-T_2)\mathbf{T}-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3)\mathbf{I}=\mathbf{0},\\
&\mathbf{T}^4-T_1\mathbf{T}^3+\tfrac{1}{2}(T_1^2-T_2)\mathbf{T}^2-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3)\mathbf{T}+\tfrac{1}{24}(T_1^4-6T_1^2T_2+8T_1T_3+3T_2^2-6T_4)\mathbf{I}=\mathbf{0}.
\end{aligned}
$$

V. THE FINITE-DIMENSIONAL RESOLVENT KERNEL

The resolvent of a complex linear operator $\mathbf{T}$ is the operator-valued function $\mathbf{R}$ defined by $\mathbf{R}(z)=(z\mathbf{I}-\mathbf{T})^{-1}$. It is a well-known fact that $\mathbf{R}$ is an operator-valued analytic function outside the spectrum of $\mathbf{T}$. We shall refine such a result showing that in the finite dimensional case $\mathbf{R}$ is an operator-valued rational function completely tied by a finite set of complex-valued rational functions, whose coefficients are exactly the trace formulas.

In this vein we need an improved version of the remainder theorem in the associative algebra of complex linear operators, whose content is the extension of proposition 1 to operator-valued polynomials.

Proposition 6: If $p(w)=w^n+D_1w^{n-1}+\cdots+D_{n-1}w+D_n$ *is a polynomial with complex coefficients* D_k *and* $\mathbf{T}$ *and* $\mathbf{W}$ *are commuting complex linear operators then*

$$p(\mathbf{W})-p(\mathbf{T})=(\mathbf{W}-\mathbf{T})[\mathbf{T}^{n-1}+(\mathbf{W}+D_1\mathbf{I})\mathbf{T}^{n-2}+(\mathbf{W}^2+D_1\mathbf{W}+D_2\mathbf{I})\mathbf{T}^{n-3}+\cdots+(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}+\cdots+D_{n-1}\mathbf{I})\mathbf{I}].$$

Proof: By induction on the degree. For $n=1$, $p(\mathbf{W})-p(\mathbf{T})=(\mathbf{W}+D_1\mathbf{I})-(\mathbf{T}+D_1\mathbf{I})=\mathbf{W}-\mathbf{T}$. *For generic n,*

$$\begin{aligned}p(\mathbf{W})-p(\mathbf{T})&=\mathbf{W}(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}+\cdots+D_{n-1}\mathbf{I})+D_n\mathbf{I}\\&\quad-\mathbf{T}(\mathbf{T}^{n-1}+D_1\mathbf{T}^{n-2}+\cdots+D_{n-1}\mathbf{I})-D_n\mathbf{I}\\&=(\mathbf{W}-\mathbf{T})(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}+\cdots+D_{n-1}\mathbf{I})+\mathbf{T}[(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}\\&\quad+\cdots+D_{n-1}\mathbf{I})-(\mathbf{T}^{n-1}+D_1\mathbf{T}^{n-2}+\cdots+D_{n-1}\mathbf{I})].\end{aligned}$$

From the inductive hypothesis the last expression reads

$$\begin{aligned}&(\mathbf{W}-\mathbf{T})(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}+\cdots+D_{n-1}\mathbf{I})+\mathbf{T}(\mathbf{W}-\mathbf{T})[\mathbf{T}^{n-2}+(\mathbf{W}+D_1\mathbf{I})\mathbf{T}^{n-3}\\&\quad+\cdots+(\mathbf{W}^{n-2}+D_1\mathbf{W}^{n-3}+\cdots+D_{n-2}\mathbf{I})\mathbf{I}]\\&=(\mathbf{W}-\mathbf{T})[\mathbf{T}^{n-1}+(\mathbf{W}+D_1\mathbf{I})\mathbf{T}^{n-2}\\&\quad+\cdots+(\mathbf{W}^{n-2}+D_1\mathbf{W}^{n-3}+\cdots+D_{n-2}\mathbf{I})\mathbf{T}\\&\quad+(\mathbf{W}^{n-1}+D_1\mathbf{W}^{n-2}+\cdots+D_{n-1}\mathbf{I})\mathbf{I}].\end{aligned}$$

We are ready to achieve a *rational formula* for the resolvent kernel.

Theorem 4: *If* $p(w)=w^n+D_1w^{n-1}+\cdots+D_{n-1}w+D_n$ *is the characteristic polynomial of the complex linear operator* $\mathbf{T}$ *and z is any complex number not belonging to the spectrum of* $\mathbf{T}$ *then*

$$(z\mathbf{I}-\mathbf{T})^{-1}=\frac{1}{p(z)}\,\mathbf{T}^{n-1}+\frac{z+D_1}{p(z)}\,\mathbf{T}^{n-2}+\frac{z^2+D_1z+D_2}{p(z)}\,\mathbf{T}^{n-3}+\cdots+\frac{z^{n-1}+D_1z^{n-2}+\cdots+D_{n-1}}{p(z)}\,\mathbf{I}.$$

Proof: It is enough to consider Cayley–Hamilton's theorem for the linear operator $\mathbf{T}$ *and set* $\mathbf{W}=z\mathbf{I}$ *in proposition 6; notice that in such case* $\mathbf{W}^k+D_1\mathbf{W}^{k-1}+\cdots+D_k\mathbf{I}=(z^k+D_1z^{k-1}+\cdots+D_k)\mathbf{I}$ *for each* $k=1,2,...,n$.

The joint content of theorem 4 and the trace formulas statement shall be pointed out as follows: *For each finite dimension of the underlying vector space there is a fundamental rational formula for the resolvent kernel of a linear operator*. For instance, in dimensions up to four, the trace formulas (3) provide the following *resolvent formulas*:

$$
\begin{aligned}
(z\mathbf{I}-\mathbf{T})^{-1} &= (z-T_1)^{-1}\mathbf{I},\\
(z\mathbf{I}-\mathbf{T})^{-1} &= (z^2-T_1z+\tfrac{1}{2}(T_1^2-T_2))^{-1}[\mathbf{T}+(z-T_1)\mathbf{I}],\\
(z\mathbf{I}-\mathbf{T})^{-1} &= (z^3-T_1z^2+\tfrac{1}{2}(T_1^2-T_2)z-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3))^{-1}[\mathbf{T}^2+(z-T_1)\mathbf{T}\\
&\quad +(z^2-T_1z+\tfrac{1}{2}(T_1^2-T_2))\mathbf{I}],\\
(z\mathbf{I}-\mathbf{T})^{-1} &= (z^4-T_1z^3+\tfrac{1}{2}(T_1^2-T_2)z^2-\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3)z+\tfrac{1}{24}(T_1^4-6T_1^2T_2+8T_1T_3+3T_2^2\\
&\quad -6T_4))^{-1}[\mathbf{T}^3+(z-T_1)\mathbf{T}^2+(z^2-T_1z+\tfrac{1}{2}(T_1^2-T_2))\mathbf{T}+(z^3-T_1z^2+\tfrac{1}{2}(T_1^2-T_2)z\\
&\quad -\tfrac{1}{6}(T_1^3-3T_1T_2+2T_3))\mathbf{I}].
\end{aligned}
\tag{5}
$$

VI. APPLICATIONS TO GENERAL RELATIVITY

In general relativity the fundamental physical object is an effective geometry, mathematically represented by a second-rank covariant tensor field $\mathbf{g}$, defined throughout a suitable four-dimensional manifold. In a local coordinate system $\mathbf{x}=(x^\alpha)$ the metric tensor can be written as $\mathbf{g}=g_{\mu\nu}dx^\mu\otimes dx^\nu$.

Investigations on general relativity are frequently carried out under the assumption that there exists some background geometry $\mathring{\mathbf{g}}=\mathring{g}_{\mu\nu}dx^\mu\otimes dx^\nu$. The metric properties to be assumed on $\mathring{\mathbf{g}}$ vary depending on the gravitational scenario. In order to perform calculations it suffices to set $\mathring{\mathbf{g}}$ a Ricci-flat metric. However, to interpret the results as physically meaningful it is generally agreed that one should require $\mathring{\mathbf{g}}$ to be a flat metric. Even this case is sometimes thought to be too broad, as some authors claim to set harmonic coordinates[8] or even Cartesian coordinates.[9] We do not take into account here any suplementary conditions on the background geometry $\mathring{\mathbf{g}}$.

The effective and background geometries are related by a tensor field $\mathbf{h}=h_{\mu\nu}dx^\mu\otimes dx^\nu$ by a connecting equation, the generally accepted form of which is $\mathbf{g}=\mathring{\mathbf{g}}+\mathbf{h}$. The scalar density $\sqrt{-g}d^4x$ associated with $\mathbf{g}$ requires us to compute $\det(\mathbf{g})=\det(\mathring{\mathbf{g}}+\mathbf{h})$. This can be achieved by means of

Proposition 7: If $\mathbf{g}=\mathring{\mathbf{g}}+\mathbf{h}$, $\mathbf{H}=\mathring{\mathbf{g}}^{-1}\mathbf{h}$, $H_k=\mathrm{trace}\,(\mathbf{H}^k)$ *and the underlying manifold is four-dimensional then*

$$
\begin{aligned}
\frac{\det(\mathbf{g})}{\det(\mathring{\mathbf{g}})} &= 1+H_1+\frac{1}{2}(H_1^2-H_2)+\frac{1}{6}(H_1^3-3H_1H_2+2H_3)\\
&\quad +\frac{1}{24}(H_1^4-6H_1^2H_2+8H_1H_3+3H_2^2-6H_4).
\end{aligned}
$$

Proof: From $\mathbf{g}=\mathring{\mathbf{g}}+\mathbf{h}=\mathring{\mathbf{g}}(\mathbf{I}+\mathbf{H})$ *it follows that* $\det(\mathbf{g})=\det(\mathring{\mathbf{g}})\det(\mathbf{I}+\mathbf{H})$. *Now it is enough to consider the four-dimensional characteristic formula in (4) with* $z=1$ *and* $\mathbf{T}=-\mathbf{H}$; *notice that* $\mathbf{T}^k=(-)^k\mathbf{H}^k$ *implies* $T_k=(-)^kH_k$.

The Levi-Civita connection associated with $\mathbf{g}$ requires us to compute $\mathbf{g}^{-1}=(\mathring{\mathbf{g}}+\mathbf{h})^{-1}=(\mathbf{I}+\mathbf{H})^{-1}\mathring{\mathbf{g}}^{-1}$. The known explicit form is given by the Neumann series[10]

$$
(\mathbf{I}+\mathbf{H})^{-1}=\mathbf{I}-\mathbf{H}+\mathbf{H}^2-\mathbf{H}^3+\cdots .
$$

In local coordinates it reads

$$
g^{\mu\nu}=(\delta^\mu_\beta-\mathring{g}^{\mu\alpha}h_{\alpha\beta}+\mathring{g}^{\mu\alpha}h_{\alpha\lambda}\mathring{g}^{\lambda\epsilon}h_{\epsilon\beta}-\mathring{g}^{\mu\alpha}h_{\alpha\lambda}\mathring{g}^{\lambda\epsilon}h_{\epsilon\rho}\mathring{g}^{\rho\sigma}h_{\sigma\beta}+\cdots)\mathring{g}^{\beta\nu},
$$

where $g^{\mu\nu}$ and $\mathring{g}^{\mu\nu}$ are well-defined by $g^{\mu\alpha}g_{\alpha\nu}=\delta^\mu_\nu=\mathring{g}^{\mu\beta}\mathring{g}_{\beta\nu}$.

Besides convergence requirements on the above series, we stress that such expression leads to technical difficulties when developing the Lagrangian variational formalism. These problems were already dealt with in the literature, the proposed solution being to modify the above form of the connecting equation.[11] We show how to overcome such drawbacks by means of

Proposition 8: If $\mathbf{g}=\mathring{\mathbf{g}}+\mathbf{h}$, $\mathbf{H}=\mathring{\mathbf{g}}^{-1}\mathbf{h}$, $H_k=\mathrm{trace}\,(\mathbf{H}^k)$ *and the underlying manifold is four-dimensional then*

$$\mathbf{g}^{-1}=(1+H_1+\tfrac{1}{2}(H_1^2-H_2)+\tfrac{1}{6}(H_1^3-3H_1H_2+2H_3)+\tfrac{1}{24}(H_1^4-6H_1^2H_2+8H_1H_3+3H_2^2-6H_4))^{-1}[-\mathbf{H}^3+(1+H_1)\mathbf{H}^2-(1+H_1+\tfrac{1}{2}(H_1^2-H_2))\mathbf{H}+(1+H_1+\tfrac{1}{2}(H_1^2-H_2)+\tfrac{1}{6}(H_1^3-3H_1H_2+2H_3))\mathbf{I}]\mathring{\mathbf{g}}^{-1}.$$

Proof: From $\mathbf{g}=\mathring{\mathbf{g}}+\mathbf{h}=\mathring{\mathbf{g}}(\mathbf{I}+\mathbf{H})$ *it follows that* $\mathbf{g}^{-1}=(\mathbf{I}+\mathbf{H})^{-1}\mathring{\mathbf{g}}^{-1}$. *Now it is enough to consider the four-dimensional resolvent formula in (5) with* $z=1$ *and* $\mathbf{T}=-\mathbf{H}$; *notice that* $\mathbf{T}^k=(-)^k\mathbf{H}^k$ *implies* $T_k=(-)^kH_k$.

ACKNOWLEDGMENTS

The author thanks Dr. Renato Klippert Barcellos for having collaborated with him on this work, for valuable discussions, and for the implementation of the computer algebra algorithm. The author owes special thanks to Dr. Nelson Pinto Neto and Dr. Vitorio Alberto De Lorenci for encouragement, and for having drawn his attention to the relevance of the subject of the present work. The author is also indebted to Carla da Fonseca-Bernetti for helpful corrections of the manuscript. This work was financially supported by CNPq.

[1] J. V. Uspensky, *Theory of Equations* (McGraw–Hill, New York, 1948).
[2] S. Lang, *Algebra* (Addison–Wesley, London, 1971), revised printing, p 119.
[3] B. L. van der Warden, *Algebra*, 5th ed. (Springer, New York, 1991), Vol. I, p. 99.
[4] See Ref. 1, p. 260.
[5] S. Mac Lane and G. Birkhoff, *Algebra*, 4th printing (MacMillan London, 1970), p 310; see also Ref. 2, p. 399.
[6] See Ref. 5, p. 152, Ref. 2, p. 58.
[7] See Ref. 5, p. 316; Ref. 2, p. 400.
[8] V. Fock, *The Theory of Space, Time and Gravitation*, 2nd revised ed. (Pergamon, Oxford, 1966), p. 319.
[9] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, The Classical Theory of Fields, Vol. 2, 4th revised English ed. (Butterworth, Oxford, 1975), p. 330.
[10] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Functional Analysis Vol. I (Academic, New York, 1972), p. 191.
[11] L. P. Grishchuk, A. N. Petrov, and A. D. Popova, Commun. Math. Phys. **94**, 379 (1984).

Branching rules for restriction of the Weil representations of Sp($n,\mathbb{R}$) to its maximal parabolic subgroup CM(n)

D. J. Rowe
Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

J. Repka
Department of Mathematics, University of Toronto, Toronto, Ontario M5S 3G3, Canada

(Received 3 March 1998; accepted for publication 8 July 1998)

The symplectic group $\mathrm{Sp}(n,\mathbb{R})$ is the group of linear canonical transformations of a real $2n$-dimensional phase space and $\mathrm{CM}(n)\subset\mathrm{Sp}(n,\mathbb{R})$ is a maximal parabolic subgroup. The symplectic groups are the fundamental dynamical groups of classical and quantal Hamiltonian mechanics. In particular, $\mathrm{Sp}(3,\mathbb{R})$ is the dynamical group of the spherical harmonic oscillator and its Weil (harmonic series) representations are important for the microscopic (shell model) description of the collective motions of many-particle systems. The subgroup $\mathrm{CM}(3)\subset\mathrm{Sp}(3,\mathbb{R})$ also appears in the microscopic theory of nuclear collective motion as the dynamical group of a hydrodynamic model of quadrupole vibrations and rotations of a nucleus. Thus, the $\mathrm{Sp}(3,\mathbb{R})\rightarrow\mathrm{CM}(3)$ branching rules are needed in finding the embedding of the hydrodynamic collective model in the microscopic shell model. Some new developments are made in the vector-coherent-state theory of induced representations. © *1998 American Institute of Physics.* [S0022-2488(98)01211-0]

I. INTRODUCTION

The symplectic group $\mathrm{Sp}(n,\mathbb{R})$ is the group of linear transformations of a real $2n$-dimensional phase space, $\mathbb{R}^{2n}$, that leave the antisymmetric biform

$$(\tilde{p},\tilde{x})\begin{pmatrix}0 & I_n\\ -I_n & 0\end{pmatrix}\begin{pmatrix}p\\ x\end{pmatrix}=(\tilde{p}x-\tilde{x}p),\quad x\in\mathbb{R}^n,\quad p\in\mathbb{R}^n, \tag{1}$$

invariant; $\tilde{x}$ denotes the transpose (row vector) of x (a column vector). Thus, $\mathrm{Sp}(n,\mathbb{R})$ is the group of $2n\times 2n$ real matrices

$$\mathrm{Sp}(n,\mathbb{R})=\left\{\begin{pmatrix}A & B\\ C & D\end{pmatrix}\middle|\begin{pmatrix}\tilde{A} & \tilde{C}\\ \tilde{B} & \tilde{D}\end{pmatrix}\begin{pmatrix}0 & I_n\\ -I_n & 0\end{pmatrix}\begin{pmatrix}A & B\\ C & D\end{pmatrix}=\begin{pmatrix}0 & I_n\\ -I_n & 0\end{pmatrix}\right\}. \tag{2}$$

The symplectic groups are the fundamental dynamical groups of classical and quantal Hamiltonian mechanics. In quantum mechanics, one is interested, for example, in their unitary representations on many-particle Hilbert spaces. The latter are spaces of square integrable wave functions $\mathscr{L}^2(\mathbb{R}^{Nn})$ for N-particles in n dimensions. The corresponding representations are known as Weil or Segal–Shale–Weil representations.[1] They are also known as harmonic oscillator or harmonic series representations.

It is known (cf. Gelbart[2]) that the space $\mathscr{L}^2(\mathbb{R}^{Nn})$, with $N=2n$, carries all the holomorphic discrete series representations of $\mathrm{Sp}(n,\mathbb{R})$. For other N, some of the discrete series representations are missing and, for $N<2n$, other irreps (irreducible representations) which are limits of discrete series appear. The harmonic series comprise all representations with lowest weights; for N even these representations are true representations of $\mathrm{Sp}(n,\mathbb{R})$ while for N odd they are projective representations, i.e., true representations of the double covering metaplectic group $\mathrm{Mp}(n,\mathbb{R})$. [For the remainder of this paper we shall mean by a representation of $\mathrm{Sp}(n,\mathbb{R})$ either a true or a projective representation according as N is even or odd.]

$\mathrm{CM}(n)$ (a mnemonic for ''collective motion in n dimensions'') is the maximal parabolic subgroup of matrices of the form

 © 1998 American Institute of Physics

$$\mathrm{CM}(n)=\left\{\begin{pmatrix} A & B \\ 0 & D \end{pmatrix}\in \mathrm{Sp}(n,\mathbb{R})\right\}\subset \mathrm{Sp}(n,\mathbb{R}). \tag{3}$$

It is generated by the matrices

$$\begin{aligned} M_1(g)&=\begin{pmatrix} g & 0 \\ 0 & \tilde{g}^{-1} \end{pmatrix}, \quad g\in \mathrm{GL}(n,\mathbb{R}),\\ M_2(B)&=\begin{pmatrix} I_n & B \\ 0 & I_n \end{pmatrix}, \quad B=\tilde{B}\in \mathbb{R}^{n\times n}, \end{aligned} \tag{4}$$

where $\mathbb{R}^{n\times n}$ is the set of $n\times n$ real matrices. Thus, $\mathrm{CM}(n)$ is isomorphic to the semidirect product group $[\mathbb{R}^{(1/2)n(n+1)}]\mathrm{GL}(n,\mathbb{R})$ of which the normal subgroup $\mathbb{R}^{(1/2)n(n+1)}$ is the additive Abelian group of real $n\times n$ symmetric matrices. The isomorphism is given by the map

$$M_2(B)M_1(g)\rightarrow(B,g), \tag{5}$$

from which one infers the group product

$$(B,g)\circ(B',g')=(B+gB'\tilde{g},gg'). \tag{6}$$

To generate the group $\mathrm{Sp}(n,\mathbb{R})$, one needs, in addition to the generators of $\mathrm{CM}(n)$, the matrix

$$J=\begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}. \tag{7}$$

The group CM(3) appears, for example, in the microscopic theory of nuclear collective motion as the dynamical group of a hydrodynamic model of quadrupole vibrations and rotations of a nucleus.[3] The shape and orientation of a nucleus in this model are characterized by a 3×3 real symmetric matrix Q whose entries are the quadrupole moments

$$Q_{ij}=\int x_i x_j \rho(x)\mathrm{d}v, \tag{8}$$

of the nuclear matter density distribution ρ, where (x_1,x_2,x_3) are Cartesian coordinates of a point $x\in\mathbb{R}^3$. The momentum observables of the model are infinitesimal generators $\{S_{ij}\}$ of the general linear transformations

$$Q\rightarrow\tilde{g}Qg, \quad g\in\mathrm{GL}(3,\mathbb{R}). \tag{9}$$

Together, the set $\{Q_{ij},S_{ij}\}$ comprise an algebra of observables for the CM(3) model. Note that the algebra $\mathbb{R}^6$ spanned by the quadrupole moments $\{Q_{ij}\}$ is both an Abelian Lie algebra and an Abelian Lie group with product given by addition. Moreover, an arbitrary element $B=\Sigma_{ij}\,B_{ij}Q_{ij}\in\mathbb{R}^6$ transforms under $\mathrm{GL}(3,\mathbb{R})$ according to

$$g{:}B\rightarrow\sum_{ij}\,B_{ij}(\tilde{g}Qg)_{ij}=\sum_{ij}\,(gB\tilde{g})_{ij}Q_{ij}, \quad g\in\mathrm{GL}(3,\mathbb{R}), \tag{10}$$

consistent with Eq. (6). The representations of CM(3) that have a microscopic interpretation for an N-nucleon nucleus are the ones carried by subspaces of $\mathscr{S}^2(\mathbb{R}^{3N})$.[4] We construct these representations in the following. Moreover, we determine which $\mathrm{CM}(n)$ irreps are contained within the harmonic series of $\mathrm{Sp}(3,\mathbb{R})$ irreps.

In addition to its CM(3) subgroup, the symplectic group $\mathrm{Sp}(3,\mathbb{R})$ contains the unitary group U(3) as a subgroup. This is significant because U(3) is the symmetry group of the spherical harmonic oscillator; a fact which makes it particularly easy to construct the irreps of $\mathrm{Sp}(3,\mathbb{R})$ in a basis which reduces the subgroup chain

$$\mathrm{Sp}(3,\mathbb{R})\supset \mathrm{U}(3)\supset \mathrm{SO}(3). \tag{11}$$

For this reason, microscopic collective model calculations[5] have been carried out in such a basis. However, to interpret the results of such calculations, one would like to express the wave functions in a basis that reduces the subgroup chain

$$\mathrm{Sp}(3,\mathbb{R})\supset \mathrm{CM}(3)\supset \mathrm{SO}(3). \tag{12}$$

The $\mathrm{Sp}(n,\mathbb{R})\to\mathrm{U}(n)$ branching rules have been long known for the holomorphic discrete series representations. They were derived for all the harmonic series by Rowe, Wybourne, and Butler.[6] The $\mathrm{Sp}(n,\mathbb{R})\to\mathrm{CM}(n)$ branching rules were given for the holomorphic discrete series by Rosensteel and Rowe.[7] We give them herein for the whole harmonic series.

II. THE WEIL REPRESENTATIONS

The Weil (harmonic series) representations of $\mathrm{Sp}(n,\mathbb{R})$ on $\mathscr{L}^2(\mathbb{R}^{Nn})$, with $N=1,2,3,...$, are defined [1,2] by

$$\begin{aligned} M_1(e^a)&\to W(M_1(e^a))=\exp\left[\frac{i}{2}\sum_{ij} a_{ij}\hat{S}_{ij}\right],\\ M_2(B)&\to W(M_2(B))=\exp\left[-\frac{i}{2}\sum_{ij} B_{ij}\hat{Q}_{ij}\right],\\ J&\to W(J)=\exp\left[-\frac{i\pi}{2}\hat{H}\right] \end{aligned} \tag{13}$$

where

$$\begin{aligned} [\hat{S}_{ij}\psi](X)&=\sum_{s=1}^{N}(X_{si}P_{sj}+P_{sj}X_{si})\psi(X),\\ [\hat{Q}_{ij}\psi](X)&=\sum_{s=1}^{N}X_{si}X_{sj}\psi(X),\\ [\hat{H}\psi](X)&=\frac{1}{2}\sum_{s=1}^{N}(X_{si}^2+P_{si}^2)\psi(X); \end{aligned} \tag{14}$$

$X=\{X_{si}\}$ is an element of $\mathbb{R}^{Nn}$ and $P_{si}=-i\partial/\partial X_{si}$. Equivalently, one can show that

$$\begin{aligned} [W(M_1(g))\psi](X)&=[R(g)\psi](X)=[\det g]^{N/2}\psi(Xg),\\ [W(M_2(B))\psi](X)&=\exp\left[-\frac{i}{2}\,\mathrm{Tr}(XB\tilde{X})\right]\psi(X),\\ [W(J)\psi](X)&=(-i)^{Nn/2}\phi(X), \end{aligned} \tag{15}$$

where ϕ is the Fourier transform of ψ. As already noted, the Weil representations are projective, for N odd, i.e., they are true representations of the double covering metaplectic group $\mathrm{Mp}(n,\mathbb{R})$.

A Weil representation of $\mathrm{Sp}(n,\mathbb{R})$ can be combined with the corresponding representation of $\mathrm{O}(N)$

$$\omega\to\rho(\omega),\quad [\rho(\omega)\psi](X)=\psi(\omega^{-1}X),\quad \omega\in\mathrm{O}(N), \tag{16}$$

to give a representation T of the direct product group $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$ for which

$$T(\omega,\alpha)\psi=\rho(\omega)W(\alpha)\psi,\quad \omega\in\mathrm{O}(N),\quad \alpha\in\mathrm{Sp}(n,\mathbb{R}). \tag{17}$$

The decomposition of T is then given by the following theorem.

Theorem: (Kashiwara and Vergne)[1] Let ρ^σ denote an irreducible $O(N)$ representation of highest weight σ and let $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ denote the sum of the $O(N)$-invariant subspaces of $\mathscr{L}^2(\mathbb{R}^{Nn})$ on which the action of $O(N)$ is equivalent to ρ^σ. Then, $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ is the carrier space for an irreducible representation T^σ (projective when N is odd) of $O(N)\times Sp(n,\mathbb{R})$. Moreover, $T=\Sigma_\sigma T^\sigma$ is the direct sum of such irreps.

III. GENERIC REPRESENTATIONS OF THE MAXIMAL PARABOLIC SUBGROUP

First we consider the case where $N>n$. Let S denote the surface of all points in $\mathbb{R}^{Nn}$ for which the matrix of quadrupole moments $Q(X)=\tilde{X}X$ is equal to the $n\times n$ identity matrix, i.e.,

$$S=\{\xi\in\mathbb{R}^{Nn}|\tilde{\xi}\xi=I_n\}. \tag{18}$$

This surface is invariant under both the left action of $O(N)$ and the right action of $O(n)$. Moreover, the measure on S inherited from the Euclidean measure on $\mathbb{R}^{Nn}$ is also invariant under these $O(N)$ and $O(n)$ actions. It follows that the space $U=\mathscr{L}^2(S)$ of square integrable functions on S, relative to this $O(n)$-invariant measure, carries a unitary representation ρ of $O(n)$ with

$$[\rho(\Omega)\varphi](\xi)=\varphi(\xi\Omega),\quad \Omega\in O(n),\quad \varphi\in U. \tag{19}$$

Furthermore, an irreducible right $O(n)$-invariant subspace $\mathscr{U}^v\subset U$ is the carrier space for both an irreducible $O(n)$ subrepresentation ρ^v of ρ and a unitary irrep $\mathscr{R}^v$ of the semidirect product group $[\mathbb{R}^{(1/2)n(n+1)}]O(n)$ with

$$[\mathscr{R}^v(B,\Omega)\varphi](\xi)=e^{-(i/2)\mathrm{Tr}[B]}\varphi(\xi\Omega),\quad (B,\Omega)\in[\mathbb{R}^{(1/2)n(n+1)}]O(n),\quad \varphi\in\mathscr{U}^v. \tag{20}$$

The objective is now to induce a unitary irrep of $CM(n)$, on a subspace of $\mathscr{L}^2(\mathbb{R}^{Nn})$, from the irrep $\mathscr{R}^v$ of $[\mathbb{R}^{(1/2)n(n+1)}]O(n)\subset CM(n)$ on $\mathscr{U}^v$. This would be straightforward if $\mathscr{U}^v$ were a closed subspace of $\mathscr{L}^2(\mathbb{R}^{Nn})$, which it is not. Instead, we make use of the fact that the functionals on $\mathscr{U}^v$ are naturally extended to functionals on the dense subspace $\mathbb{H}_D$ of Schwartz functions[8] in $\mathscr{L}^2(\mathbb{R}^{Nn})$. This enables us to map $\mathscr{U}^v$ into the dual space $\mathbb{H}_D^*$.

An element $\langle\varphi_U|$ of the dual space U^* of functionals on U can be seen as an element $\langle\varphi|$ of the dual $\mathbb{H}_D^*$ of the dense subspace $\mathbb{H}_D$ of Schwartz functions in $\mathscr{L}^2(\mathbb{R}^{Nn})$ as follows. Let $\psi_{|S}$ denote the restriction of the function $\psi\in\mathbb{H}_D$ to the subset $S\subset\mathbb{R}^{Nn}$. The function $\psi_{|S}$ is an element of $\mathscr{L}^2(S)$ (because it is continuous and therefore bounded on the compact set S). Therefore, given a linear functional $\langle\varphi_U|\in U^*$, we can define $\langle\varphi|\in\mathbb{H}_D^*$ by

$$\langle\varphi|\psi\rangle=\langle\varphi_U|\psi_{|S}\rangle. \tag{21}$$

Combined with the standard map $U\to U^*;\varphi\mapsto\langle\varphi_U|$, defined by the Hermitian inner product of $U=\mathscr{L}^2(S)$, we then have a map $U\to\mathbb{H}_D^*$ defined by the sequence $U\to U^*\to\mathbb{H}_D^*;\varphi\mapsto\langle\varphi_U|\mapsto\langle\varphi|$.

The $U\to\mathbb{H}_D^*$ map can be constructed explicitly as follows. Recall that there is a natural map from $\mathbb{R}^{Nn}$ to $\mathbb{H}_D^*$ in which an element $X\in\mathbb{R}^{Nn}$ maps to the functional $\langle X|$ for which

$$\langle X|\psi\rangle=\psi(X),\quad \psi\in\mathbb{H}_D. \tag{22}$$

Thus, for $\varphi\in U$, we define the functional on $\mathbb{H}_D$,

$$\langle\varphi|=\int_S d\xi\ \varphi^*(\xi)\langle\xi|, \tag{23}$$

where $d\xi$ is the $O(N)$-invariant measure on S.

Let $\{\langle vK|\}$ in $\mathbb{H}_D^*$ denote the images under the above map of a set of orthonormal basis functions $\{\varphi_K^v\}$ for $\mathscr{U}^v$. We then have a map $\Pi^v:\mathscr{L}^2(\mathbb{R}^{Nn})\to\mathscr{U}^v$ defined on the dense subspace $\mathbb{H}_D\subset\mathscr{L}^2(\mathbb{R}^{Nn})$ by

$$\Pi^{\upsilon}=\sum_{K}\ \varphi_{K}^{\upsilon}\langle \upsilon K|. \tag{24}$$

Claim: Π^{υ} is an $[\mathbb{R}^{(1/2)n(n+1)}]\mathrm{O}(n)$-intertwining operator for which

$$\Pi^{\upsilon}W(M_2(B)M_1(\Omega))=\mathscr{R}^{\upsilon}(B,\Omega)\Pi^{\upsilon},\quad (B,\Omega)\in[\mathbb{R}^{(1/2)n(n+1)}]\mathrm{O}(n). \tag{25}$$

Proof: The group $[\mathbb{R}^{(1/2)n(n+1)}]\mathrm{O}(n)$ is a subgroup of $\mathrm{Sp}(n,\mathbb{R})$. Thus, it has a (reducible) representation on $\mathbb{H}_D$ given by the restriction of the Weil representation W of $\mathrm{Sp}(n,\mathbb{R})$. For this representation, we have, for any $|\psi\rangle\in\mathbb{H}_D$

$$\sum_{K}\ \varphi_{K}^{\upsilon}\langle \upsilon K|W(M_2(B))|\psi\rangle=\sum_{K}\ \varphi_{K}^{\upsilon}\int_{S}d\xi\ \varphi_{K}^{\upsilon}(\xi)^{*}\langle\xi|W(M_2(B))|\psi\rangle$$

$$=e^{-(i/2)\mathrm{Tr}[B]}\sum_{K}\ \varphi_{K}^{\upsilon}\langle \upsilon K|\psi\rangle, \tag{26}$$

implying that

$$\Pi^{\upsilon}W(M_2(B))=e^{-(i/2)\mathrm{Tr}[B]}\Pi^{\upsilon}. \tag{27}$$

We also have

$$\Pi^{\upsilon}W(M_1(\Omega))=\rho^{\upsilon}(\Omega)\Pi^{\upsilon}. \tag{28}$$

Therefore

$$\Pi^{\upsilon}W(M_2(B)M_1(\Omega))=e^{-(i/2)\mathrm{Tr}[B]}\rho^{\upsilon}(\Omega)\Pi^{\upsilon} \tag{29}$$

and the claim follows.

Claim: The VCS (vector-coherent-state) wave functions [9,10]

$$\mathscr{F}^{\upsilon}=\{\psi:\mathrm{GL}(n,\mathbb{R})\to\mathscr{C}^{\upsilon}\,|\,\psi(g)=\Pi^{\upsilon}W(M_1(g))|\psi\rangle;\ \ g\in\mathrm{GL}(n,\mathbb{R}),\ \ |\psi\rangle\in\mathbb{H}_D\} \tag{30}$$

satisfy the identity

$$\psi(\Omega g)=\rho^{\upsilon}(\Omega)\psi(g),\quad \Omega\in\mathrm{O}(n),\quad g\in\mathrm{GL}(n,\mathbb{R}), \tag{31}$$

and carry a unitary representation, $\overline{\Gamma}^{\upsilon}$, of $\mathrm{CM}(n)$ defined by

$$[\overline{\Gamma}^{\upsilon}(B,g)\psi](g')=e^{-(i/2)\mathrm{Tr}[g'B\tilde{g}']}\psi(g'g). \tag{32}$$

This representation is irreducible.

Proof: The identity (31) follows directly from Eq. (28). Under restriction of the Weil representation to $\mathrm{CM}(n)$, an element $(B,g)\in\mathrm{CM}(n)$ maps

$$|\psi\rangle\to W(M_2(B)M_1(g))|\psi\rangle. \tag{33}$$

The corresponding map $\psi\to\overline{\Gamma}^{\upsilon}(B,g)\psi$ of the VCS wave function ψ for $|\psi\rangle$ is therefore given by

$$[\overline{\Gamma}^{\upsilon}(B,g)\psi](g')=\Pi^{\upsilon}W(M_1(g'))W(M_2(B)M_1(g))|\psi\rangle$$

$$=\Pi^{\upsilon}W(M_2(g'B\tilde{g}')M_1(g'g))|\psi\rangle$$

$$=e^{-(i/2)\mathrm{Tr}[g'B\tilde{g}']}\psi(g'g). \tag{34}$$

Thus, $\overline{\Gamma}^{\upsilon}$ is equivalent to a subrepresentation of the unitary Weil representation restricted to $\mathrm{CM}(n)$. The representation $\overline{\Gamma}^{\upsilon}$ is irreducible by Mackey's theory of induced representations.[11]

Note, that there is a close similarity, but also important differences, between the representations of $\mathrm{CM}(n)$ obtained in this way and those constructed[4] by Mackey induction.[11] In Mackey's

construction, one starts with the set of square integrable $\mathrm{GL}(n,\mathbb{R})\to\mathscr{C}^v$ functions that satisfy Eq. (31). One then defines a representation on this space in accord with Eq. (32). An advantage of Mackey's approach is that the Hilbert space of the representation is equipped with an inner product from the start. In contrast, more work is required to infer the inner product in the VCS method. The advantage of the VCS construction is that it tells us which representations of $\mathrm{CM}(n)$, equivalent to Mackey representations, exist within a given reducible representation on some completely different space (in our case $\mathscr{L}^2(\mathbb{R}^{Nn})$). Moreover, having found such a representation, we can make use of Mackey's theory to infer that it is irreducible.

IV. BRANCHING RULES

Determination of the $\mathrm{CM}(n)$ irreps contained in the restriction to $\mathrm{CM}(n)$ of an irreducible harmonic series (Weil) representation of $\mathrm{Sp}(n,\mathbb{R})$ is simple in a VCS representation.

Observation: The surface S, defined by Eq. (18) for $N>n$, is diffeomorphic to the Stiefel manifold $\mathrm{O}(N)/\mathrm{O}(N-n)$.

Proof: The $\mathrm{O}(N)$ action on $\mathbb{R}^{Nn}$, in which an element $\omega\in\mathrm{O}(N)$ sends X to ωX, leaves S invariant. Furthermore, an $N\times n$ matrix is in S if and only if its columns are orthonormal; such a matrix is of the form $\omega\dot{I}$, where $\dot{I}$ is the matrix with entries $\dot{I}_{si}=\delta_{si}$ and $\omega\in\mathrm{O}(N)$. Thus, S is an $\mathrm{O}(N)$ orbit in $\mathbb{R}^{Nn}$ containing $\dot{I}$. The stability subgroup at $\dot{I}$ is $\mathrm{O}(N-n)$.

By the Peter–Weyl theorem, an orthonormal basis for $U=\mathscr{L}^2(\mathrm{O}(N)/\mathrm{O}(N-n))$ is given by the $\mathrm{O}(N)/\mathrm{O}(N-n)\to\mathbb{C}$ functions $\{\phi^\sigma_{\nu r}\}$, where

$$\phi^\sigma_{\nu r}(\xi)=\sqrt{d_\sigma}\;D^\sigma_{\nu,r}(\xi)^*,\quad \xi\in\mathrm{O}(N)/\mathrm{O}(N-n); \tag{35}$$

$D^\sigma_{\nu,r}$ is a matrix coefficient function for an $\mathrm{O}(N)$ irrep of highest weight σ and dimension d_σ; ν indexes an orthonormal basis for the carrier space V^σ of this irrep and r indexes an orthonormal basis for the $\mathrm{O}(N-n)$-invariant subspace of V^σ. For each σ and r, the functions $\{\phi^\sigma_{\nu r}\}$ are a basis for a representation ρ^σ of $\mathrm{O}(N)$ defined by

$$[\rho^\sigma(\omega)\phi^\sigma_{\nu r}](\xi)=\phi^\sigma_{\nu r}(\omega^{-1}\xi),\quad \omega\in\mathrm{O}(N). \tag{36}$$

Let $\{\langle\nu\sigma r|\}$ in $\mathbb{H}_D^*$ denote the images under the $U\to\mathbb{H}_D^*$ map (defined in Sec. III) of the orthonormal basis functions $\{\phi^\sigma_{\nu r}\}$. And let U^σ denote the subspace of U spanned by the functions $\{\phi^\sigma_{\nu r}\}$ of a given σ. Then, for each σ, we have a map $\Pi^\sigma:\mathscr{L}^2(\mathbb{R}^{Nn})\to U^\sigma$ defined on the dense subspace $\mathbb{H}_D\subset\mathscr{L}^2(\mathbb{R}^{Nn})$ by

$$\Pi^\sigma=\sum_{\nu r}\phi^\sigma_{\nu r}\langle\nu\sigma r|. \tag{37}$$

Claim: The VCS wave functions

$$F^\sigma=\{\Psi:\mathrm{GL}(n,\mathbb{R})\to U^\sigma\,|\,\Psi(g)=\Pi^\sigma W(M_1(g))|\psi\rangle;\ \ g\in\mathrm{GL}(n,\mathbb{R}),\ \ |\psi\rangle\in\mathbb{H}_D\} \tag{38}$$

carry a representation Γ^σ of $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$ defined by

$$[\Gamma^\sigma(\omega,\alpha)\Psi](g)=\Pi^\sigma W(M_1(g))T(\omega,\alpha)|\psi\rangle,\quad (\omega,\alpha)\in\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R}), \tag{39}$$

where T is the representation of $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$ defined by Eq. (17). This representation is equivalent to the irreducible subrepresentation T^σ of the representation T.

Proof: Let $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ be the sum of $\mathrm{O}(N)$-invariant subspaces of $\mathscr{L}^2(\mathbb{R}^{Nn})$ on which the action of $\mathrm{O}(N)$ is equivalent to ρ^σ; as in the Kashiwara–Vergne theorem. Then, by the Kashiwara–Vergne theorem, $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ is a Hilbert space for an irreducible subrepresentation T^σ of the representation T of $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$. Moreover, the dense subspace $\mathbb{H}^\sigma_D=\mathbb{H}_D\cap\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ is invariant under T^σ. To prove the claim, we now show that F^σ is isomorphic to $\mathbb{H}^\sigma_D=\mathbb{H}_D\cap\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$.

First observe that a vector-valued function $\Psi\in F^\sigma$ is uniquely characterized by the set of $\mathrm{GL}(n,\mathbb{R})\to\mathbb{C}$ functions $\{\Psi^\sigma_{\nu r}\}$ defined by

$$\Psi^{\sigma}_{\nu r}(g)=\langle \nu\sigma r|W(M_1(g))|\psi\rangle, \quad g\in \mathrm{GL}(n,\mathbb{R}), \tag{40}$$

for some $|\psi\rangle\in\mathbb{H}_D$. These functions remain the same if $|\psi\rangle$ is replaced by its projection onto $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$. So we can assume that $|\psi\rangle\in\mathbb{H}_D\cap\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$.

Any continuous function $\psi\in\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ is uniquely defined by its restriction to the dense set of matrices $X\in\mathbb{R}^{Nn}$ that have maximal rank, i.e., rank equal to n. For such an X, performing a Gram–Schmidt orthonormalization on the columns amounts to finding $g\in\mathrm{GL}(n,\mathbb{R})$ so that $Xg\in S$ (after all, being in S is equivalent to having orthonormal columns). This implies that any continuous function $\psi\in\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ is uniquely defined by the restrictions to S of its $\mathrm{GL}(n,\mathbb{R})$-translates. In turn, this implies that ψ is determined by the overlaps $\{\langle \nu\sigma r|W(M_1(g))|\psi\rangle; g\in\mathrm{GL}(n,\mathbb{R})\}$. Thus, F^σ is isomorphic to $\mathbb{H}^\sigma_D=\mathbb{H}_D\cap\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ and the claim is proved.

Irreps of $\mathrm{Sp}(n,\mathbb{R})$ are constructed on subspaces of F^σ as follows. First observe that the operator Π^σ satisfies the identity

$$\Pi^\sigma T(\omega,\alpha)=\rho^\sigma(\omega)\Pi^\sigma W(\alpha), \quad \omega\in\mathrm{O}(N), \quad \alpha\in\mathrm{Sp}(n,\mathbb{R}). \tag{41}$$

It follows that the action of the $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$ representation Γ^σ is given by

$$[\Gamma^\sigma(\omega,\alpha)\Psi](g)=\rho^\sigma(\omega)\Pi^\sigma W(M_1(g)\alpha)|\psi\rangle, \quad (\omega,\alpha)\in\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R}). \tag{42}$$

Now F^σ is a direct sum $F^\sigma=\Sigma_\nu F^\sigma_\nu$, where

$$F^\sigma_\nu=\{\Psi:\mathrm{GL}(n,\mathbb{R})\to U^\sigma_\nu|\Psi(g)=\Pi^\sigma_\nu W(M_1(g))|\psi\rangle;\ g\in\mathrm{GL}(n,\mathbb{R}),\ |\psi\rangle\in\mathbb{H}_D\}, \tag{43}$$

U^σ_ν is the subspace of U^σ spanned by the basis wave functions $\{\phi^\sigma_{\nu r}\}$ with both σ and ν held fixed, and

$$\Pi^\sigma_\nu=\sum_r \phi^\sigma_{\nu r}\langle \nu\sigma r|. \tag{44}$$

From the Kashiwara–Vergne theorem, we know that each subspace F^σ_ν is invariant and irreducible under the restriction of Γ^σ to $\mathrm{Sp}(n,\mathbb{R})$. Thus, F^σ_ν is the carrier space for an $\mathrm{Sp}(n,\mathbb{R})$ irrep Γ^σ_ν for which

$$[\Gamma^\sigma_\nu(\alpha)\Psi](g)=\Pi^\sigma_\nu W(M_1(g)\alpha)|\psi\rangle, \quad \alpha\in\mathrm{Sp}(n,\mathbb{R}), \quad \Psi(g)=\Pi^\sigma_\nu W(M_1(g))|\psi\rangle. \tag{45}$$

We now obtain the $\mathrm{Sp}(n,\mathbb{R})\to\mathrm{CM}(n)$ branching rules by expanding the $\mathrm{Sp}(n,\mathbb{R})$ intertwining operator Π^σ_ν as a sum of intertwining operators for $\mathrm{CM}(n)$ irreps. Recall that the label r of the U basis wave functions $\{\phi^\sigma_{\nu r}\}$ indexes a basis for the $\mathrm{O}(N-n)$-invariant subspace of the carrier space V^σ for an $\mathrm{O}(N)$ irrep equivalent to ρ^σ. Let us choose a basis for this subspace of V^σ that reduces the subgroup chain

$$\begin{array}{ccccc}\mathrm{O}(N) & \supset & \mathrm{O}(n)\times\mathrm{O}(N-n),\\ \sigma & & \delta\ \upsilon K & & 0\end{array} \tag{46}$$

where K labels a basis for an $\mathrm{O}(n)$ irrep υ, 0 labels an identity representation of $\mathrm{O}(N-n)$ and δ is a multiplicity index. The index r is then replaced by the set of indices $(\delta\upsilon K)$. It follows that Π^σ_ν can be expanded

$$\Pi^\sigma_\nu=\sum_{\delta\upsilon K}\Pi^\sigma_{\nu\delta\upsilon K}. \tag{47}$$

Moreover, we can identify the $\mathrm{CM}(n)$ intertwining operator Π^υ with any one of the $\Pi^\sigma_{\nu\delta\upsilon K}$ operators of the same υ.

Thus, $\mathrm{Sp}(n,\mathbb{R})\to\mathrm{CM}(n)$ branching rules for the $N\geqslant n$ harmonic series irreps are given by the following claim:

Theorem 1(A): If $N>n$ and an $O(n)$ irrep ρ^v occurs κ times in combination with the identity irrep of $O(N-n)$ in the restriction of the irrep ρ^σ of $O(N)$ to its $O(n)\times O(N-n)$ subgroup, then the irrep Γ^v of $CM(n)$ occurs κ times in the restriction of the $N>n$ harmonic series irrep Γ^σ_ν of $Sp(n,\mathbb{R})$ to $CM(n)$.

An equivalent, and more useful, form of this branching rule is given by equating it with a corresponding $U(n)\to O(n)$ branching rule. Let $U(n)$ be the maximal compact subgroup of $Sp(n,\mathbb{R})$. On restriction to this subgroup, a harmonic series irrep of $Sp(n,\mathbb{R})$ restricts to an infinite sum of $U(n)$ irreps each of which is labeled by a highest weight. Let σ denote the $U(n)$ highest weight having lowest value in this sum. With σ interpreted in this way, the $Sp(n,\mathbb{R})\to CM(n)$ branching rules are given for the $N>n$ harmonic series irreps by the following claim.

Theorem 1(B): If an $O(n)$ irrep ρ^v occurs κ times in the restriction of the irrep of highest weight σ of $U(n)$ to its $O(n)$ subgroup, then the irrep Γ^v of $CM(n)$ occurs κ times in the restriction of an $N>n$ harmonic series irrep Γ^σ_ν of $Sp(n,\mathbb{R})$ to $CM(n)$.

Proof: The claim is proved by comparing the known $U(n)\to O(n)$ and $O(N)\to O(n)\times O(N-n)$ branching rules both of which are given by King.[12]

V. BRANCHING RULES FOR THE REMAINING HARMONIC SERIES REPRESENTATIONS

Now we consider the harmonic representations for which $N\leqslant n$. When $N\leqslant n$, we define the surface S in $\mathbb{R}^{Nn}$ to be the subset of matrices

$$S=\{\xi\in\mathbb{R}^{Nn}\,|\,\tilde{\xi}\xi=I^n_N\}, \tag{48}$$

where I^n_N is the $n\times n$ diagonal rank-N matrix with elements

$$(I^n_N)_{ij}=\delta_{ij}\times\begin{cases}1 & \text{if } i\leqslant N\\ 0 & \text{if } i>N.\end{cases} \tag{49}$$

S is generated by the right action of the subgroup $O(N)\subset O(n)$ on the matrix $\dot{I}\in\mathbb{R}^{Nn}$, where $\dot{I}$ is the matrix with entries $I_{si}=\delta_{si}$ and an element $\Omega\in O(n)$ acts on $\mathbb{R}^{Nn}$ in the usual way by sending $X\in\mathbb{R}^{Nn}$ to $X\tilde{\Omega}$. The stability subgroup of this $O(N)$ action at $\dot{I}$ is the identity subgroup. Thus, S is diffeomorphic to $O(N)$.

The square integrable functions $U=\mathscr{L}^2(O(N))$ on $O(N)\sim S$, relative to the $O(N)$-invariant measure, carry a right representation ρ_R of $O(N)$ with

$$[\rho_R(\omega)\phi](\omega')=\phi(\omega'\omega). \tag{50}$$

It follows that an irreducible right $O(N)$-invariant subspace $\mathscr{U}^\sigma\subset U$ is the carrier space for both an irrep of $O(N)$ and a unitary irrep $\mathscr{R}^\sigma$ of $[\mathbb{R}^{(1/2)n(n+1)}]O(N)$ with

$$[\mathscr{R}^\sigma(B,\omega)\phi](\omega')=e^{-(i/2)\mathrm{Tr}[BI^n_N]}\phi(\omega'\omega),\quad (B,\omega)\in[\mathbb{R}^{(1/2)n(n+1)}]O(N),\quad \phi\in\mathscr{U}^\sigma. \tag{51}$$

We now induce a representation of $CM(n)$ from the representation $\mathscr{R}^\sigma$.

By the Peter–Weyl theorem, an orthonormal basis for $U=\mathscr{L}^2(O(N))$ is given by the functions $\{\phi^\sigma_{\mu\nu}\}$, where

$$\phi^\sigma_{\mu\nu}(\omega)=\sqrt{d_\sigma}\;D^\sigma_{\mu,\nu}(\omega),\quad \omega\in O(N), \tag{52}$$

and $D^\sigma_{\mu,\nu}$ is a matrix coefficient function for an $O(N)$ irrep ρ^σ of dimension d_σ. Let U^σ_μ denote the subspace of U^σ spanned by the subset of functions $\{\phi^\sigma_{\mu\nu}\}$ with σ and μ fixed. The $U^\sigma_\mu\subset U^\sigma$ subspaces are invariant and irreducible under the right $O(N)$ action ρ_R; they carry equivalent $O(N)$ irreps given by

$$\rho^\sigma_R(\omega)\phi^\sigma_{\mu\nu}=\sum_{\nu'}\phi^\sigma_{\mu\nu'}D^\sigma_{\nu',\nu}(\omega),\quad \omega\in O(N). \tag{53}$$

One sees that the irreducible right $O(N)$-invariant subspace $\mathscr{U}^\sigma\subset U$ can be identified with any one of the U^σ_μ of the same σ.

Let $\{\langle\mu\sigma\nu|\}$ in $\mathbb{H}_D^*$ denote the images under the $U\to\mathbb{H}_D^*$ map (defined in Sec. III) of the orthonormal basis functions $\{\phi^\sigma_{\mu\nu}\}$ and let

$$\Pi^\sigma_\mu=\sum_\nu \phi^\sigma_{\mu\nu}\langle\mu\sigma\nu|. \tag{54}$$

Claim: The VCS wave functions

$$F^\sigma_\mu=\{\Psi:\mathrm{GL}(n,\mathbb{R})\to U^\sigma_\mu|\Psi(g)=\Pi^\sigma_\mu W(M_1(g))|\psi\rangle;\ \ g\in\mathrm{GL}(n,\mathbb{R}),\ \ |\psi\rangle\in\mathbb{H}_D\} \tag{55}$$

satisfy the identity

$$\Psi(\omega g)=\rho^\sigma_R(\omega)\Psi(g),\quad \omega\in\mathrm{O}(N)\subset\mathrm{O}(n),\quad g\in\mathrm{GL}(n,\mathbb{R}), \tag{56}$$

and carry an irreducible representation, $\overline{\Gamma}^\sigma_\mu$, of $\mathrm{CM}(n)$, for which

$$[\overline{\Gamma}^\sigma_\mu(B,g)\Psi](g')=e^{-(i/2)\mathrm{Tr}[g'B\tilde{g}'I^n_N]}\Psi(g'g). \tag{57}$$

Proof: Equation (56) follows from the identity

$$\Pi^\sigma_\mu W(M_1(\omega))=\rho^\sigma_R(\omega)\Pi^\sigma_\mu,\quad \omega\in\mathrm{O}(N)\subset\mathrm{O}(n). \tag{58}$$

Under restriction of the Weil representation to $\mathrm{CM}(n)$, an element $(B,g)\in\mathrm{CM}(n)$ maps $|\psi\rangle\to W(M_2(B)M_1(g))|\psi\rangle$. The corresponding map $\Psi\to\overline{\Gamma}^\sigma_\mu(B,g)\Psi$ of the VCS wave function Ψ for $|\psi\rangle$ is given therefore by

$$\begin{aligned}[\overline{\Gamma}^\sigma_\mu(B,g)\Psi](g')&=\Pi^\sigma_\mu W(M_1(g')M_2(B)M_1(g))|\psi\rangle\\&=\Pi^\sigma_\mu W(M_2(g'B\tilde{g}')M_1(g'g))|\psi\rangle\\&=e^{-(i/2)\mathrm{Tr}[g'B\tilde{g}'I^n_N]}\Psi(g'g).\end{aligned} \tag{59}$$

This representation $\overline{\Gamma}^\sigma_\mu$ is the representation of $\mathrm{CM}(n)$ induced from the representation $\mathscr{R}^\sigma$. It is known, from Mackey's theory,[11] to be irreducible.

We now show that F^σ_μ also carries an $\mathrm{Sp}(n,\mathbb{R})$ irrep.

Theorem 2: Let σ label an irrep of $\mathrm{O}(N)$. The Kashiwara–Vergne theorem associates to σ a harmonic series irrep of $\mathrm{Sp}(n,\mathbb{R})$. On restriction to its maximal parabolic subgroup, $\mathrm{CM}(n)$, this irrep of $\mathrm{Sp}(n,\mathbb{R})$ remains irreducible when $N<n$. The $\mathrm{CM}(n)$ representation obtained by the restriction is the representation induced from the irrep $\mathscr{R}^\sigma$ of the $[\mathbb{R}^{(1/2)n(n+1)}]\mathrm{O}(N)\subset\mathrm{CM}(n)$ subgroup.

Proof: Let $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$ denote the sum of $\mathrm{O}(N)$-invariant subspaces of $\mathscr{L}^2(\mathbb{R}^{Nn})$ on which the action of $\mathrm{O}(N)$ is equivalent to ρ^σ. Then, as shown in Sec. IV, $F^\sigma=\Sigma_\mu F^\sigma_\mu$ is isomorphic to a dense subspace of $\mathscr{L}^2_\sigma(\mathbb{R}^{Nn})$. It follows from the Kashiwara–Vergne theorem that F^σ carries a unitary irrep Γ^σ of $\mathrm{O}(N)\times\mathrm{Sp}(n,\mathbb{R})$ for which

$$\begin{aligned}\Gamma^\sigma(\omega,\alpha)\Psi](g)&=\sum_\mu \Pi^\sigma_\mu W(M_1(g))T(\omega,\alpha)|\psi\rangle,\\&=\sum_\mu \Pi^\sigma_\mu T(\omega,M_1(g)\alpha)|\psi\rangle.\end{aligned} \tag{60}$$

The restriction of this irrep, Γ^σ, to $\mathrm{Sp}(n,\mathbb{R})$ is reducible but, again by the Kashiwara–Vergne theorem, each subspace $F^\sigma_\mu\subset F^\sigma$ is invariant and irreducible.

VI. COMMENTS ON VCS INDUCTION

In deriving the above branching rules, we have relied on the VCS inducing construction and introduced some new techniques. Since these techniques are more generally useful, we conclude with some brief comments on VCS theory in the context of general inducing constructions.

The essential ingredients of all inducing constructions are as follows. Let G be a group and $H \subset G$ a subgroup. Let ρ be a representation of H on a vector space U and let $\mathscr{F}$ be a G-invariant space of $G \to U$ (vector-valued) functions for which the G action is expressed

$$G \times \mathscr{F} \to \mathscr{F}; \quad g \times \psi \mapsto \psi_g, \quad \text{where} \quad \psi_g(g') = \psi(g'g). \tag{61}$$

The space $\mathscr{F}$ carries a (generally reducible) representation of G. A subrepresentation carried by the G-invariant subset of functions

$$\overline{\mathscr{F}} = \{\psi \in \mathscr{F} \,|\, \psi(hg) = \rho(h)\psi(g), \ \forall h \in H\}, \tag{62}$$

is then said to be induced from the representation ρ of H. Different choices of $\mathscr{F}$ give rise to different inducing constructions. In Mackey's theory,[11] for example, $\mathscr{F}$ is usually set equal to $\mathscr{L}^2(G)$, relative to a G-invariant or quasi-invariant measure.

The VCS construction starts with a representation T of G on some space (e.g., a Hilbert space) $\mathbb{H}$ and seeks an intertwining operator Π which maps $\mathbb{H}$ to the carrier space U of a representation ρ of a subgroup $H \subset G$, i.e., Π satisfies

$$\Pi T(h) = \rho(h)\Pi, \quad h \in H. \tag{63}$$

This ensures that the VCS wave functions

$$\overline{\mathscr{F}} = \{\psi \,|\, \psi(g) = \Pi T(g)\Psi; \Psi \in \mathbb{H}, g \in G\} \tag{64}$$

satisfy the identity $\psi(hg) = \rho(h)\psi(g)$, for all $h \in H$, and carry a representation Γ for which

$$[\Gamma(g)\psi](\alpha) = \Pi T(\alpha)T(g)\Psi = \psi(\alpha g), \quad g \in G. \tag{65}$$

The VCS representation Γ is seen to be isomorphic to a subrepresentation of T.

The standard (i.e., Mackey) construction has the advantage that the Hilbert space $\mathscr{L}^2(G)$ is automatically equipped with an inner product, whereas the inner product for a VCS representation has to be inferred from that of the Hilbert space $\mathbb{H}$ in which it is embedded. This can be an advantage, e.g., for representations for which the $\mathscr{L}^2(G)$ inner product diverges, as for nondiscrete series representations of $\mathrm{Sp}(n, \mathbb{R})$. Moreover, even when defined, the $\mathscr{L}^2(G)$ integrals can be much harder to evaluate than the corresponding expressions of the VCS construction. For present purposes, the special property of the VCS induction that we have exploited is that it gives the representations contained within a (generally reducible) representation T on an arbitrary space $\mathbb{H}$ [not necessarily $\mathscr{L}^2(G)$ or even a Hilbert space].

In the special case that $\mathbb{H}$ is set equal to $\mathscr{L}^2(G)$ and T is the regular representation, VCS induction may or may not be equivalent to that of Mackey, depending on the choice of Π. For some choices of Π it may be possible to induce irreducible VCS representation of G whereas the standard induced representation is reducible. The space $\mathbb{H}$ might also be a space of holomorphic sections of a G/H bundle or a Hilbert space of many-particle wave functions, as in this paper. There are many possibilities. Thus, VCS theory provides tools for finding the subrepresentations of a group within a particular space of interest. The practicality of the VCS construction relies on the existence of systematic ways of choosing the representation ρ to achieve some desired purpose and for determining the corresponding space $\overline{\mathscr{F}}$ of wave functions. Moreover, by means of the so-called K-matrix theory,[9,10] there are systematic ways of determining the inner product for $\overline{\mathscr{F}}$ inherited from $\mathbb{H}$.

In the first applications of VCS theory, the carrier space U for the representation ρ of H was taken to be a subspace of $\mathbb{H}$ and Π was an $\mathbb{H} \to U$ projection operator. The projection operator was constructed explicitly by choosing an orthonormal basis $\{\varphi_\alpha\}$ for U and identifying the wave functions of this basis with corresponding vectors $\{|\alpha\rangle\}$ in $\mathbb{H}$ (in which U is embedded as a subspace). The $\mathbb{H} \to U$ projection operator Π is then defined

$$\Pi = \sum_\alpha \varphi_\alpha \langle \alpha|. \tag{66}$$

The only problem with Π defined in this way is that it limits the construction to induction from a representation ρ of the subgroup H that occurs in a direct sum decomposition of $T|_H$, the restriction of T to H. However, as shown explicitly in this paper, this limitation can be relaxed and the VCS construction also applied with representations that occur in a direct integral decomposition of $T|_H$. We have shown that there is a natural embedding of the dual U^* of U in the space of functionals $\mathbb{H}_D^*$ on a dense subspace $\mathbb{H}_D$ of $\mathbb{H}$. Thus, if $\{\varphi_\alpha\}$ is an orthonormal basis for U we obtain a sequence of maps $U \to U^* \to \mathbb{H}_D^*$ in which the basis functions $\{\varphi_\alpha\}$ map to functionals $\{\langle\alpha|\}$ in $\mathbb{H}_D^*$. We then again have an intertwining operator $\Pi = \Sigma_\alpha \varphi_\alpha \langle\alpha|$. The only difference is that now Π is only defined on the dense subspace $\mathbb{H}_D$ but that is sufficient.

ACKNOWLEDGMENTS

The authors wish to thank G. T. Rosensteel and J. Harnad for helpful comments.

[1] M. Kashiwara and M. Vergne, Invent. Math. **44**, 1 (1978).
[2] S. Gelbart, Invent. Math. **19**, 49 (1973).
[3] L. Weaver, L. C. Biedenharn, and R. Y. Cusson, Ann. Phys. (N.Y.) **77**, 250 (1973).
[4] G. Rosensteel and D. J. Rowe, Ann. Phys. (N.Y.) **96**, 1 (1976); **123**, 36 (1979).
[5] G. Rosensteel and D. J. Rowe, Phys. Rev. Lett. **38**, 10 (1977); Ann. Phys. (N.Y.) **126**, 343 (1980); P. Park, J. Carvalho, M. Vassanji, D. J. Rowe, and G. Rosensteel, Nucl. Phys. A **414**, 93 (1984); D. J. Rowe, Rep. Prog. Phys. **48**, 1419 (1985).
[6] D. J. Rowe, B. G. Wybourne, and P. H. Butler, J. Phys. A **18**, 939 (1985).
[7] G. R. Rosensteel and D. J. Rowe, Phys. Rev. Lett. **46**, 1119 (1981); D. J. Rowe, Prog. Theor. Phys. Suppl. **74,75**, 306 (1983).
[8] The Schwartz functions are a subset of smooth functions whose Fourier transforms are normalizable (Hilbert space) functions; cf. G. G. Emch, *Mathematical and Conceptual Foundations of 20th-Century Physics* (North-Holland, Amsterdam, 1984).
[9] D. J. Rowe, J. Math. Phys. **25**, 2662 (1984); D. J. Rowe, G. Rosensteel, and R. Carr, J. Phys. A **17**, L399 (1984); D. J. Rowe, R. Le Blanc, and K. T. Hecht, J. Math. Phys. **29**, 287 (1988).
[10] D. J. Rowe and J. Repka, J. Math. Phys. **32**, 2614 (1991).
[11] G. W. Mackey, Ann. Math. **55**, 101 (1952); *Induced Representation of Groups and Quantum Mechanics* (Benjamin, New York, 1968); *Unitary Group Representations in Physics, Probability and Number Theory* (Benjamin, Reading, MA, 1978).
[12] R. C. King, J. Phys. A **8**, 429 (1975).

Algebraic expressions for some SO(N) Racah coefficients

Dong Ruan[a)]
Department of Physics, Tsinghua University, Beijing 100084, People's Republic of China

Hong-zhou Sun
Department of Physics, Tsinghua University, Beijing 100084, People's Republic of China and Institute of Theoretical Physics, Academia Sinica, Beijing 100080, People's Republic of China

(Received 20 May 1998; accepted for publication 24 June 1998)

In this paper the algebraic expressions for SO(N) Racah coefficients involving the coupling of three irreducible representations $(1\,\dot{0})$, $(m_1\,m_2\,m_3\,\dot{0})$ and $(1\,\dot{0})$ are given by using the building-up process. Their properties are discussed in detail. © *1998 American Institute of Physics.* [S0022-2488(98)01811-8]

I. INTRODUCTION

Racah coefficients[1] (or the highly symmetrical $6j$ symbols[2]) for SO(N) are very useful in calculating energy levels and transition rates in atomic, molecular and nuclear structure theory. They appear in the calculation of the matrix element of any coupled tensor operator when the Wigner–Eckart theorem is applied. In fact, Racah coefficients are the transformation coefficients between different coupling schemes of three angular momenta in quantum mechanics, where the underlying group is SO(3). This concept and the cases of the coupling of more than three angular momenta (or $3nj$ symbols) have been discussed widely[3–6] and been generalized to various groups such as point groups,[7,8] symmetry groups,[9] classical Lie groups SU(N),[10–12] SO(N),[13–16] Sp($2N$),[17] and exceptional Lie group G_2.[14,18]

There are many approaches to the Racah coefficients. Usually, a basic and direct approach is to use the Biedenharn–Elliott sum rule (including the Racah backcoupling formula and the orthonormality relations of Racah coefficients).[3–6] The other methods are the use of the bootstrap procedure,[15,19,20] the generating functions method,[17] the building-up process using the known reduction factors (reduced Wigner factors),[13,21,22] the diagrammatic method[23,24] and so on.

Besides the well-known SO(3), the algebraic expressions for some special Racah coefficients of SO(N) were discussed by Judd *et al.*[14,15] and Ališauskas.[16] Judd *et al.* considered the cases of single-parameter irreducible representations for SO(N) (so only one parameter w in $6j$ symbol), and of all six irreducible representations having the type $(1^a\,0^{l-a})$ for SO($2l+1$). Ališauskas considered the semistretched case of the symmetry irreducible representations $(l\,\dot{0})$ for SO(N) ($N>5$) and gave the polynomial-type expressions for $6j$ symbol, namely, they are expressed by the sums over a series of $6j$ symbols of SU(2).

Recently, Sun *et al.*[25,26] obtained the algebraic expressions for SO(N)$\supset$SO($N-1$) reduction factors for $(m_1\,m_2\,m_3\,\dot{0})\times(1\,\dot{0})$ by use of the irreducible tensor basis method.[27,28] Therefore, it allows us, using the building-up process, to derive algebraic expressions for all SO(N) Racah coefficients for $(1\,\dot{0})\times(m_1\,m_2\,m_3\,\dot{0})\times(1\,\dot{0})$.

This paper is arranged as follows. In Sec. II, the definitions of Racah coefficient and the corresponding $6j$ symbol for an arbitrary group and the building-up equation we will utilize are reviewed simply. In Sec. III, the algebraic expressions for SO(N) Racah coefficients are calculated and given in convenient tabular forms. A simple discussion is given in the final section.

II. DEFINITIONS AND BUILDING-UP EQUATION

Let Γ label the irreducible representations of group G and Λ label those of the subgroup G'. Then $|\Gamma\,\Lambda\,\lambda\rangle$ are the basis vectors carrying the irreducible representations Γ, where λ, as the

a)Electronic mail: gllong@mail.tsinghua.edu.cn

 © 1998 American Institute of Physics

components of Λ, distinguishes the individual basis vectors. When we consider the coupling of products of the basis states of three irreducible representations Γ_1, Γ_2 and Γ_3 for the definitive resultant coupled states $|\Gamma\,\Lambda\,\lambda\rangle$, there are two coupling sequences: one is first coupling $|\Gamma_1\,\Lambda_1\,\lambda_1\rangle$ with $|\Gamma_2\,\Lambda_2\,\lambda_2\rangle$ to $|\Gamma_{12}\,\Lambda_{12}\,\lambda_{12}\rangle$, then coupling $|\Gamma_{12}\,\Lambda_{12}\,\lambda_{12}\rangle$ with $|\Gamma_3\,\Lambda_3\,\lambda_3\rangle$ to $|\Gamma\,\Lambda\,\lambda\rangle$; another is first coupling $|\Gamma_2\,\Lambda_2\,\lambda_2\rangle$ with $|\Gamma_3\,\Lambda_3\,\lambda_3\rangle$ to $|\Gamma_{23}\,\Lambda_{23}\,\lambda_{23}\rangle$, then coupling $|\Gamma_1\,\Lambda_1\,\lambda_1\rangle$ with $|\Gamma_{23}\,\Lambda_{23}\,\lambda_{23}\rangle$ to $|\Gamma\,\Lambda\,\lambda\rangle$. (Note that here and afterwards, i.e., in Eqs. (1) and (4) and in the definitions of Racah coefficient (2), $6j$ symbol (3) and reduction factor (5), we have dropped the additional labels because in our case they are multiplicity-free.) The two resultant coupled states $|(\Gamma_1\Gamma_2)\,\Gamma_{12},\Gamma_3;\Gamma\,\Lambda\,\lambda\rangle$ and $|\Gamma_1,(\Gamma_2\,\Gamma_3)\,\Gamma_{23};\Gamma\,\Lambda\,\lambda\rangle$ must be related by a unitary transformation matrix $\mathrm{U}(\Gamma)$, such that

$$|(\Gamma_1\Gamma_2)\Gamma_{12},\Gamma_3;\Gamma\Lambda\lambda\rangle=\sum_{\Gamma_{23}}\mathrm{U}_{\Gamma_{12}\Gamma_{23}}(\Gamma)|\Gamma_1,(\Gamma_2\Gamma_3)\Gamma_{23};\Gamma\Lambda\lambda\rangle \tag{1}$$

where matrix element $\mathrm{U}_{\Gamma_{12}\,\Gamma_{23}}(\Gamma)$, which is independent of Λ and λ,

$$\mathrm{U}_{\Gamma_{12}\,\Gamma_{23}}(\Gamma)\equiv W\begin{pmatrix}\Gamma_1 & \Gamma_2 & \Gamma_{12}\\ \Gamma_3 & \Gamma & \Gamma_{23}\end{pmatrix}, \tag{2}$$

is called a normalized Racah coefficient of group G. Its corresponding $6j$ symbol, which may display better symmetries, is[15]

$$\begin{Bmatrix}\Gamma_1 & \Gamma_2 & \Gamma_{12}\\ \Gamma_3 & \Gamma & \Gamma_{23}\end{Bmatrix}=[\mathrm{Dim}(\Gamma_{12})\,\mathrm{Dim}(\Gamma_{23})]^{-1/2}W\begin{pmatrix}\Gamma_1 & \Gamma_2 & \Gamma_{12}\\ \Gamma_3 & \Gamma & \Gamma_{23}\end{pmatrix}, \tag{3}$$

where $\mathrm{Dim}(\Gamma_i)$ is the dimension of the irreducible representation Γ_i. Equation (3) is analogous to the definition of Jahn[29] for SO(3).

By standard recoupling theory, the ket in Eq. (1) may be expressed in terms of sums of products of coupling coefficients. Then, multiplying this equation on both sides by $\langle\,\Gamma_1\,\Lambda_1,(\Gamma_2\,\Gamma_3)\,\Gamma_{23}\Lambda_{23}|\Gamma\,\Lambda\,\rangle$, and using the Racah factorization lemma together with the unitarity properties of the coupling coefficients, we can obtain the following very useful relationship[13,22,28]

$$\begin{aligned}W\begin{pmatrix}\Gamma_1 & \Gamma_2 & \Gamma_{12}\\ \Gamma_3 & \Gamma & \Gamma_{23}\end{pmatrix}\begin{bmatrix}\Gamma_1 & \Gamma_{23}\\ \Lambda_1 & \Lambda_{23}\end{bmatrix}\Gamma\Lambda\Big]=&\sum_{\Lambda_2,\,\Lambda_{12},\,\Lambda_3}W\begin{pmatrix}\Lambda_1 & \Lambda_2 & \Lambda_{12}\\ \Lambda_3 & \Lambda & \Lambda_{23}\end{pmatrix}\begin{bmatrix}\Gamma_1 & \Gamma_2\\ \Lambda_1 & \Lambda_2\end{bmatrix}\Gamma_{12}\Lambda_{12}\Big]\\ &\times\begin{bmatrix}\Gamma_3 & \Gamma_{12}\\ \Lambda_3 & \Lambda_{12}\end{bmatrix}\Gamma\Lambda\Big]\begin{bmatrix}\Gamma_2 & \Gamma_3\\ \Lambda_2 & \Lambda_3\end{bmatrix}\Gamma_{23}\Lambda_{23}\Big]\end{aligned} \tag{4}$$

where

$$W\begin{pmatrix}\Gamma_1 & \Gamma_2 & \Gamma_{12}\\ \Gamma_3 & \Gamma & \Gamma_{23}\end{pmatrix}\quad\text{and}\quad W\begin{pmatrix}\Lambda_1 & \Lambda_2 & \Lambda_{12}\\ \Lambda_3 & \Lambda & \Lambda_{23}\end{pmatrix}$$

are the normalized Racah coefficients of G and of its subgroup G', respectively, and the remaining factors, like

$$\begin{bmatrix}\Gamma_1 & \Gamma_{23}\\ \Lambda_1 & \Lambda_{23}\end{bmatrix}\Gamma\Lambda\Big], \tag{5}$$

stand for the $G\supset G'$ reduction factors. Equation (4) is the so-called building-up equation, which relates the Racah coefficients of G with those of its subgroup G' by the $G\supset G'$ reduction factors. It is a straightforward generalization of the relationships between SO(3) Racah coefficients and the Clebsch–Gordan coefficients of SO(3) [i.e., the $\mathrm{SO}(3)\supset\mathrm{SO}(2)$ reduction factors].[3]

TABLE I. SO(3) Racah coefficients $W\begin{pmatrix} 1 & j & j_{12} \\ 1 & j' & j_{23} \end{pmatrix}$

(a) $j'=j+1$

$j_{12}\backslash j_{23}$	j	$j+1$
j	$\frac{1}{j+1}$	$\frac{\sqrt{j(j+2)}}{j+1}$
$j+1$	$\frac{\sqrt{j(j+2)}}{j+1}$	$-\frac{1}{j+1}$

(b) $j'=j-1$

$j_{12}\backslash j_{23}$	$j-1$	j
$j-1$	$\frac{1}{j}$	$\frac{\sqrt{(j-1)(j+1)}}{j}$
j	$\frac{\sqrt{(j-1)(j+1)}}{j}$	$-\frac{1}{j}$

(c) $j'=j$

$j_{12}\backslash j_{23}$	$j-1$	j	$j+1$
$j-1$	$\frac{1}{j(2j+1)}$	$-\frac{\sqrt{2j-1}}{j\sqrt{2j+1}}$	$\frac{\sqrt{(2j-1)(2j+3)}}{2j+1}$
j	$-\frac{\sqrt{2j-1}}{j\sqrt{2j+1}}$	$\frac{j^2+j-1}{j(j+1)}$	$\frac{\sqrt{(2j-1)(2j+3)}}{(j+1)(2j+1)}$
$j+1$	$\frac{\sqrt{(2j-1)(2j+3)}}{2j+1}$	$\frac{\sqrt{(2j-1)(2j+3)}}{(j+1)(2j+1)}$	$\frac{1}{(j+1)(2j+1)}$

III. CALCULATIONS OF RACAH COEFFICIENTS

We have observed that Eq. (4) is a kind of recursion relationship. This equation makes it possible that the Racah coefficients, to within an overall phase convention, of groups of higher ranks may be obtained step by step from those of their subgroups through a suitable group chain provided that the corresponding reduction factors have been known.

In order to obtain SO(N) Racah coefficients, we choose a canonical group chain $\mathrm{SO}(N)\supset\mathrm{SO}(N-1)\supset\cdots\supset\mathrm{SO}(3)\supset\mathrm{SO}(2)$. Due to the trivial SO(2), whose Racah coefficients only take either 1 when the ordinary weight addition rules are satisfied or 0 otherwise, we first need to derive the algebraic expressions for Racah coefficients of the lowest but nontrivial subgroup, SO(3). Fortunately, as mentioned in Sec. I, the case of SO(3) has been treated very well in the quantum theory of angular momentum.[3,5,6] The algebraic expressions for SO(3) Racah coefficients for $1\times j\times 1$ considered in this paper, cast into the corresponding unitary matrices, are given in Table I. It is worth reminding readers to note the differences between the phase conventions of SO(3) Racah coefficients in Table I and the usual ones of the Racah coefficients of the angular momenta[3] because the phase conventions of the coupling coefficients of SO(3) we used[25] are different from those of the coupling coefficients of the angular momenta.[3] The convenient phase conventions of SO(3) Racah coefficients in Table I have been adopted in order to be consistent with the phase conventions of SO(N) ($N>3$) Racah coefficients calculated later.

Then, using the SO(3) Racah coefficients (Table I) and the $\mathrm{SO}(4)\supset\mathrm{SO}(3)$ reduction factors for $(m_{14}m_{24})\times(1\ 0)$,[25] we can obtain all SO(4) Racah coefficients for $(1\ 0)\times(m_{14}\ m_{24})\times(1\ 0)$ by the standard technique, that is, in the building-up equation (4) Λ_{23} and Λ taking their highest weights respectively while Λ_2, Λ_{12} and Λ_3 are running their range of allowed values. This method may go further to SO(5), SO(6), ... We have calculated up to SO(7), and obtained SO(7) Racah coefficients, however, here the procedure is no longer repeated for economy of

TABLE II. All schemes of coupling Γ_{12} or Γ_{23} with 1 to Γ.

Γ	Γ_{12} or Γ_{23}
$(m_1-2\ m_2\ m_3\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})$
$(m_1\ m_2-2\ m_3\ \dot{0})$	$(m_1\ m_2-1\ m_3\ \dot{0})$
$(m_1\ m_2\ m_3-2\ \dot{0})$	$(m_1\ m_2\ m_3-1\ \dot{0})$
$(m_1+2\ m_2\ m_3\ \dot{0})$	$(m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1\ m_2+2\ m_3\ \dot{0})$	$(m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1\ m_2\ m_3+2\ \dot{0})$	$(m_1\ m_2\ m_3+1\ \dot{0})$
$(m_1\ m_2\ m_3\ 2\ \dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1-1\ m_2-1\ m_3\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2-1\ m_3\ \dot{0})$
$(m_1-1\ m_2\ m_3-1\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2\ m_3-1\ \dot{0})$
$(m_1\ m_2-1\ m_3-1\ \dot{0})$	$(m_1\ m_2-1\ m_3\ \dot{0})\ (m_1\ m_2\ m_3-1\ \dot{0})$
$(m_1+1\ m_2+1\ m_3\ \dot{0})$	$(m_1\ m_2+1\ m_3\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1+1\ m_2\ m_3+1\ \dot{0})$	$(m_1\ m_2\ m_3+1\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1\ m_2+1\ m_3+1\ \dot{0})$	$(m_1\ m_2\ m_3+1\ \dot{0})\ (m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1+1\ m_2\ m_3\ 1\dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1\ m_2+1\ m_3\ 1\dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})\ (m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1\ m_2\ m_3+11\ \dot{0})$	$(m_1\ m_2\ m_3\ 1\dot{0})\ (m_1\ m_2\ m_3+1\ \dot{0})$
$(m_1-1\ m_2\ m_3\ 1\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1\ m_2-1\ m_3\ 1\ \dot{0})$	$(m_1\ m_2-1\ m_3\ \dot{0})\ (m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1\ m_2\ m_3-1\ 1\ \dot{0})$	$(m_1\ m_2\ m_3-1\ \dot{0})\ (m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1-1\ m_2+1\ m_3\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1-1m_2m_3+1\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2\ m_3+1\ \dot{0})$
$(m_1\ m_2-1\ m_3+1\ \dot{0})$	$(m_1\ m_2-1\ m_3\ \dot{0})\ (m_1\ m_2\ m_3+1\ \dot{0})$
$(m_1+1\ m_2-1\ m_3\ \dot{0})$	$(m_1\ m_2-1\ m_3\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1+1\ m_2\ m_3-1\ \dot{0})$	$(m_1\ m_2\ m_3-1\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})$
$(m_1\ m_2+1\ m_3-1\ \dot{0})$	$(m_1\ m_2\ m_3-1\ \dot{0})\ (m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1\ m_2\ m_3\ \dot{0})$	$(m_1-1\ m_2\ m_3\ \dot{0})\ (m_1+1\ m_2\ m_3\ \dot{0})\ (m_1\ m_2-1\ m_3\ \dot{0})$
	$(m_1\ m_2+1\ m_3\ \dot{0})\ (m_1\ m_2\ m_3-1\ \dot{0})\ (m_1\ m_2\ m_3+1\ \dot{0})$
	$(m_1\ m_2\ m_3\ 1\ \dot{0})$

pages. From the obtained results of SO(4), SO(5), SO(6) and SO(7), we have found that there exists a very simple relation depending on rank N $(=4,\ldots,7)$ in the corresponding Racah coefficients—the factors involving $m_{iN}+m_{jN}$ contain rank N whereas the factors involving $m_{iN}-m_{jN}$ do not contain rank N. This result is not surprising because a similar property occurred in the $\mathrm{SO}(N)\supset\mathrm{SO}(N-1)$ reduction factors.[25] Therefore, we can extend easily the SO(7) Racah coefficients for $(1\ 0\ 0)\times(m_{17}\ m_{27}\ m_{37})\times(1\ 0\ 0)$ to the case of SO(N) of the arbitrary rank N for $(1\ \dot{0})\times(m_1\ m_2\ m_3\ \dot{0})\times(1\ \dot{0})$, where the three entries m_1, m_2 and m_3 in the irreducible representation $(m_1\ m_2\ m_3\ \dot{0})$ of SO(N) must be integer and obey the betweenness relation $m_1\geqslant m_2\geqslant m_3\geqslant 0$. (Here for simplicity, the subscript N of m_{iN} has been omitted, i.e., $m_{iN}\equiv m_i$, and a dot over a number implies that this number is repeated as often as necessary.) Thus using Eq. (4) and the $\mathrm{SO}(N)\supset\mathrm{SO}(N-1)$ reduction factors,[25,26] we can gain by mathematical induction all SO(N) Racah coefficients of the type

$$W\begin{pmatrix}(1\ \dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & \Gamma_{12}\\ (1\,\dot{0}) & \Gamma & \Gamma_{23}\end{pmatrix}.$$

Correspondingly, in Eq. (4), we have to choose $\Gamma_1=\Gamma_3=(1\ \dot{0})$ and $\Gamma_2=(m_1m_2\ m_3\dot{0})$. Then Γ_{12} and Γ_{23} may be different for different Γ's. They can be gained by the method of weight decomposition or Shur function method,[22,28] and are listed in Table II.

Hence eighty-six Racah coefficients in all needed calculating. Furthermore, they may be put into twenty-six unitary matrices, i.e., seven 1×1 matrices, eighteen 2×2 matrices and one 7×7 matrix. For the first seven cases in Table II, the normalized Racah coefficients are

$$W\begin{pmatrix}(1\,\dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & (m_1\pm1\,m_2\,m_3\,\dot{0})\\ (1\,\dot{0}) & (m_1\pm2\,m_2\,m_3\,\dot{0}) & (m_1\pm1\,m_2\,m_3\,\dot{0})\end{pmatrix}=1,$$

$$W\begin{pmatrix}(1\,\dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & (m_1\,m_2\pm1\,m_3\,\dot{0})\\ (1\,\dot{0}) & (m_1\,m_2\pm2\,m_3\,\dot{0}) & (m_1\,m_2\pm1\,m_3\,\dot{0})\end{pmatrix}=1,$$

$$W\begin{pmatrix}(1\,\dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & (m_1\,m_2\,m_3\pm1\,\dot{0})\\ (1\,\dot{0}) & (m_1\,m_2\,m_3\pm2\,\dot{0}) & (m_1\,m_2\,m_3\pm1\,\dot{0})\end{pmatrix}=1,$$

$$W\begin{pmatrix}(1\,\dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & (m_1\,m_2\,m_3\,1\,\dot{0})\\ (1\,\dot{0}) & (m_1\,m_2\,m_3\,2\,\dot{0}) & (m_1\,m_2\,m_3\,1\,\dot{0})\end{pmatrix}=1\,, \tag{6}$$

owing to the fact that for a fixed Γ there is only one coupling scheme (i.e., only one Γ_{12} and one Γ_{23}). The other SO(N) Racah coefficients for $\Gamma=(m_1\pm1\,m_2\pm1\,m_3\,\dot{0})$, $\Gamma=(m_1\pm1\,m_2\,m_3\pm1\,\dot{0})$ and $\Gamma=(m_1\,m_2\pm1\,m_3\pm1\,\dot{0})$ are given in Table III(a)–III(c), respectively; those for $\Gamma=(m_1+1\,m_2\,m_3\,1\,\dot{0})$, $\Gamma=(m_1\,m_2+1\,m_3\,1\,\dot{0})$ and $\Gamma=(m_1\,m_2\,m_3+1\,1\,\dot{0})$ are given in Table III(d)–III(f), respectively; those for $\Gamma=(m_1\pm1\,m_2\mp1\,m_3\,\dot{0})$, $\Gamma=(m_1\pm1\,m_2\,m_3\mp1\,\dot{0})$ and $\Gamma=(m_1\,m_2\pm1\,m_3\mp1\,\dot{0})$ are given in Table III(g)–III(i), respectively; those for $\Gamma=(m_1-1\,m_2\,m_3\,1\,\dot{0})$, $\Gamma=(m_1\,m_2-1\,m_3\,1\,\dot{0})$ and $\Gamma=(m_1\,m_2\,m_3-1\,1\,\dot{0})$ are given in Table III(j)–III(l), respectively, and, finally, those for $\Gamma=(m_1\,m_2\,m_3\,\dot{0})$ are given in Table III(m).

However, the $6j$ symbol forms of SO(N) Racah coefficients for $(1\,\dot{0})\times(m_1\ m_2\ m_3\,\dot{0})\times(1\,\dot{0})$ are not given here because they may be obtained easily by Eq. (3) together with the dimension formula of the irreducible representation $(m_1\ m_2\ m_3\,\dot{0})$ of SO(N),[26,28]

$$\begin{aligned}\mathrm{Dim}(m_1\,m_2\,m_3\,\dot{0})=\ & \frac{(2m_1+N-2)(2m_2+N-4)(2m_3+N-6)}{(N-2)(N-3)(N-4)^2(N-5)^2(N-6)^3}\\ &\times(m_1-m_2+1)(m_1+m_2+N-3)(m_1-m_3+2)\\ &\times(m_1+m_3+N-4)(m_2-m_3+1)(m_2+m_3+N-5)\\ &\times\begin{pmatrix}m_1+N-5\\ m_1+2\end{pmatrix}\begin{pmatrix}m_2+N-6\\ m_2+1\end{pmatrix}\begin{pmatrix}m_3+N-7\\ m_3\end{pmatrix},\end{aligned} \tag{7}$$

where the binomal symbol $\binom{n}{k}=n!/k!\,(n-k)!$ has been used.

We have observed the following properties for the SO(N) Racah coefficients.

(1) Like the Racah coefficients of SO(3), SO(N) Racah coefficients depend on the irreducible representations of SO(N) only, and do not depend on those of SO($N-1$).

(2) There exist two types of Racah coefficients. The first type of Racah coefficients are rank N-independent [see Table III(a)–III(f)] because they are no trace contractions for the irreducible representations in the coupling, and the second type of Racah coefficients are rank N-dependent [see Table III(g)–III(m)].[30] This property of SO(N) Racah coefficients is little different from that of SU(N) Racah coefficients, in which all SU(N) Racah coefficients are rank N-independent.[10,11,31]

From the obtained SO(N) Racah coefficients (Table III), we can gain the following simple rules: compared with $\Gamma_2=(m_1\ m_2\ m_3\,\dot{0})$, if any two entries in a Γ (see Table II), say m_i and m_j $(i\neq j)$, change by the same interval, i.e., $+1$ or -1, simultaneously, then all Racah coefficients in the corresponding unitary matrix U(Γ) only contain the factors involving m_i-m_j, and hence

TABLE III. SO(N) Racah coefficients $W\begin{pmatrix}(1\,\dot{0}) & (m_1\,m_2\,m_3\,\dot{0}) & \Gamma_{12}\\ (1\,\dot{0}) & \Gamma & \Gamma_{23}\end{pmatrix}$.

(a) $\Gamma=(m_1\pm1\ m_2\pm1\ m_3\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\pm1\ m_2\ m_3\,\dot{0})$	$(m_1\ m_2\pm1\ m_3\,\dot{0})$
$(m_1\pm1\ m_2\ m_3\,\dot{0})$	$\mp\frac{1}{m_1-m_2+1}$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)}}{m_1-m_2+1}$
$(m_1\ m_2\pm1\ m_3\,\dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)}}{m_1-m_2+1}$	$\pm\frac{1}{m_1-m_2+1}$

(b) $\Gamma=(m_1\pm1\ m_2\ m_3\pm1\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\pm1\ m_2\ m_3\,\dot{0})$	$(m_1\ m_2\ m_3\pm1\,\dot{0})$
$(m_1\pm1\ m_2\ m_3\,\dot{0})$	$\mp\frac{1}{m_1-m_3+2}$	$\frac{\sqrt{(m_1-m_3+1)(m_1-m_3+}}{m_1-m_3+2}$
$(m_1\ m_2\ m_3\pm1\,\dot{0})$	$\frac{\sqrt{(m_1-m_3+1)(m_1-m_3+3)}}{m_1-m_3+2}$	$\pm\frac{1}{m_1-m_3+2}$

(c) $\Gamma=(m_1\ m_2\pm1\ m_3\pm1\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\ m_2\pm1\ m_3\,\dot{0})$	$(m_1\ m_2\ m_3\pm1\,\dot{0})$
$(m_1\ m_2\pm1\ m_3\,\dot{0})$	$\mp\frac{1}{m_2-m_3+1}$	$\frac{\sqrt{(m_2-m_3)(m_2-m_3+2)}}{m_2-m_3+1}$
$(m_1\ m_2\ m_3\pm1\,\dot{0})$	$\frac{\sqrt{(m_2-m_3)(m_2-m_3+2)}}{m_2-m_3+1}$	$\pm\frac{1}{m_2-m_3+1}$

(d) $\Gamma=(m_1+1\ m_2\ m_3\,1\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$(m_1+1\ m_2\ m_3\,\dot{0})$
$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$\frac{1}{m_1+3}$	$\frac{\sqrt{(m_1+2)(m_1+4)}}{m_1+3}$
$(m_1+1\ m_2\ m_3\,\dot{0})$	$\frac{\sqrt{(m_1+2)(m_1+4)}}{m_1+3}$	$-\frac{1}{m_1+3}$

(e) $\Gamma=(m_1\ m_2+1\ m_3\,1\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$(m_1\ m_2+1\ m_3\,\dot{0})$
$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$\frac{1}{m_2+2}$	$\frac{\sqrt{(m_2+1)(m_2+3)}}{m_2+2}$
$(m_1\ m_2+1\ m_3\,\dot{0})$	$\frac{\sqrt{(m_2+1)(m_2+3)}}{m_2+2}$	$-\frac{1}{m_2+2}$

(f) $\Gamma=(m_1\ m_2\ m_3+1\,1\,\dot{0})$

$\Gamma_{12}\setminus\Gamma_{23}$	$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$(m_1\ m_2\ m_3+1\,\dot{0})$
$(m_1\ m_2\ m_3\ 1\,\dot{0})$	$\frac{1}{m_3+1}$	$\frac{\sqrt{m_3(m_3+2)}}{m_3+1}$
$(m_1\ m_2\ m_3+1\,\dot{0})$	$\frac{\sqrt{m_3(m_3+2)}}{m_3+1}$	$-\frac{1}{m_3+1}$

(g) $\Gamma = (m_1 \pm 1\, m_2 \mp 1\, m_3\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1 \pm 1\ m_2\ m_3\ \dot{0})$	$(m_1\ m_2 \mp 1\ m_3\ \dot{0})$
$(m_1 \pm 1\ m_2\ m_3\ \dot{0})$	$\mp\frac{1}{m_1+m_2-3+N}$	$\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)}}{m_1+m_2-3+N}$
$(m_1\ m_2 \mp 1\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)}}{m_1+m_2-3+N}$	$\pm\frac{1}{m_1+m_2-3+N}$

(h) $\Gamma = (m_1 \pm 1\, m_2\, m_3 \mp 1\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1 \pm 1\ m_2\ m_3\ \dot{0})$	$(m_1\ m_2\ m_3 \mp 1\ \dot{0})$
$(m_1 \pm 1\ m_2\ m_3\ \dot{0})$	$\mp\frac{1}{m_1+m_3-4+N}$	$\frac{\sqrt{(m_1+m_3-5+N)(m_1+m_3-3+N)}}{m_1+m_3-4+N}$
$(m_1\ m_2\ m_3 \mp 1\ \dot{0})$	$\frac{\sqrt{(m_1+m_3-5+N)(m_1+m_3-3+N)}}{m_1+m_3-4+N}$	$\pm\frac{1}{m_1+m_3-4+N}$

(i) $\Gamma = (m_1\, m_2 \pm 1\, m_3 \mp 1\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2 \pm 1\ m_3\ \dot{0})$	$(m_1\ m_2\ m_3 \mp 1\ \dot{0})$
$(m_1\ m_2 \pm 1\ m_3\ \dot{0})$	$\mp\frac{1}{m_2+m_3-5+N}$	$\frac{\sqrt{(m_2+m_3-6+N)(m_2+m_3-4+N)}}{m_2+m_3-5+N}$
$(m_1\ m_2\ m_3 \mp 1\ 1\ \dot{0})$	$\frac{\sqrt{(m_2+m_3-6+N)(m_2+m_3-4+N)}}{m_2+m_3-5+N}$	$\pm\frac{1}{m_2+m_3-5+N}$

(j) $\Gamma = (m_1 - 1\, m_2\, m_3\, 1\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1 - 1\ m_2\ m_3\ \dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1 - 1\ m_2\ m_3\ \dot{0})$	$\frac{1}{m_1-5+N}$	$\frac{\sqrt{(m_1-6+N)(m_1-4+N)}}{m_1-5+N}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{\sqrt{(m_1-6+N)(m_1-4+N)}}{m_1-5+N}$	$-\frac{1}{m_1-5+N}$

(k) $\Gamma = (m_1\, m_2 - 1\, m_3\, 1\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2 - 1\ m_3\ \dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1\ m_2 - 1\ m_3\ \dot{0})$	$\frac{1}{m_2-6+N}$	$\frac{\sqrt{(m_2-7+N)(m_2-5+N)}}{m_2-6+N}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{\sqrt{(m_2-7+N)(m_2-5+N)}}{m_2-6+N}$	$-\frac{1}{m_2-6+N}$

(l) $\Gamma = (m_1\, m_2\, m_3 - 1\, 1\, \dot{0})$

$\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2\ m_3 - 1\ \dot{0})$	$(m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1\ m_2\ m_3 - 1\ \dot{0})$	$\frac{1}{m_3-7+N}$	$\frac{\sqrt{(m_3-8+N)(m_3-6+N)}}{m_3-7+N}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{\sqrt{(m_3-8+N)(m_3-6+N)}}{m_3-7+N}$	$-\frac{1}{m_3-7+N}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$

(1) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1 - 1\; m_2\; m_3\; \dot{0})$
$(m_1 - 1\; m_2\; m_3\; \dot{0})$	$\frac{[(N-2)(m_1+1)(m_1-4+N)-(N-4)m_2(m_2-4+N)](m_1+2)(m_1-5+N)-[(N-4)(m_1+2)(m_1-5+N)-(N-6)(m_2+1)(m_2-5+N)]m_3(m_3-6+N)}{(m_1-5+N)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}$
$(m_1 + 1\; m_2\; m_3\; \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+1)(m_1-m_3+3)(m_1+m_3-5+N)(m_1+m_3-3+N)(2m_1-4+N)(2m_1+N)(m_1+2)(m_1-4+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_1-2+N)\sqrt{(m_1+3)(m_1-5+N)}}$
$(m_1\; m_2 - 1\; m_3\; \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1-m_3+1)(m_1+m_3-5+N)(m_2-m_3)(m_2+m_3-6+N)(m_2+1)(m_1+2)(2m_1-4+N)(2m_2-6+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1-5+N)(m_2-6+N)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1\; m_2 + 1\; m_3\; \dot{0})$	$-\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+1)(m_1+m_3-5+N)(m_2-m_3+2)(m_2+m_3-4+N)(m_2-5+N)(m_1+2)(2m_1-4+N)(2m_2-2+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1-5+N)(m_2+2)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1\; m_2\; m_3 - 1\; \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1-m_3+3)(m_2-m_3+2)(m_2+m_3-6+N)(2m_1-4+N)(2m_3-8+N)(m_1+2)m_3}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1-5+N)(m_3-7+N)}}$
$(m_1\; m_2\; m_3 + 1\; \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1+m_2-4+N)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-4+N)(2m_1-4+N)(2m_3-4+N)(m_1+2)(m_3-6+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1-5+N)(m_3+1)}}$
$(m_1\; m_2\; m_3\; 1\; \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(2m_1-4+N)(m_2-5+N)(m_2+1)(m_3-6+N)m_3(m_1-m_2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1+m_3-5+N)}}{(m_1-5+N)\sqrt{(m_1+3)(2m_1-2+N)(m_2+2)(m_2-6+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(2) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1+1\; m_2\; m_3\; \dot{0})$
$(m_1-1\; m_2\; m_3\; \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+1)(m_1-m_3+3)(m_1+m_3-5+N)(m_1+m_3-3+N)(2m_1-4+N)(2m_1+N)(m_1+2)(m_1-4+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_1-2+N)\sqrt{(m_1+3)(m_1-5+N)}}$
$(m_1+1\; m_2\; m_3\; \dot{0})$	$\frac{[(N-2)(m_1+2)(m_1-3+N)-(N-4)m_2(m_2-4+N)](m_1+3)(m_1-4+N)-[(N-4)(m_1+3)(m_1-4+N)-(N-6)(m_2+1)(m_2-5+N)]m_3(m_3-6+N)}{(m_1+3)(2m_1-2+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}$
$(m_1\; m_2-1\; m_3\; \dot{0})$	$\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-6+N)(m_1-4+N)(m_2+1)(2m_1+N)(2m_2-6+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1+3)(m_2-6+N)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1\; m_2+1\; m_3\; \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1-m_3+3)(m_1+m_3-3+N)(m_2-m_3+2)(m_2+m_3-4+N)(2m_1+N)(2m_2-2+N)(m_1-4+N)(m_2-5+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_2-4+N)(m_1+3)(m_2+2)}}$
$(m_1\; m_2\; m_3-1\; \dot{0})$	$\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-2+N)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3+2)(m_2+m_3-6+N)(2m_1+N)(2m_3-8+N)(m_1-4+N)m_3}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1+3)(m_3-7+N)}}$
$(m_1\; m_2\; m_3+1\; \dot{0})$	$\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-2+N)(m_1-m_3+1)(m_1-m_3+3)(m_2-m_3)(m_2+m_3-4+N)(2m_1+N)(2m_3-4+N)(m_1-4+N)(m_3-6+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1+3)(m_3+1)}}$
$(m_1\; m_2\; m_3\; 1\; \dot{0})$	$\frac{\sqrt{(N-6)(m_1+2)(2m_1+N)(m_2-5+N)(m_2+1)(m_3-6+N)m_3(m_1-m_2+2)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-3+N)}}{(m_1+3)\sqrt{(m_1-5+N)(2m_1-2+N)(m_2+2)(m_2-6+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(3) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\, m_2 - 1\, m_3\, \dot{0})$
$(m_1 - 1\, m_2\, m_3\, \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1-m_3+1)(m_1+m_3-5+N)(m_2-m_3)(m_2+m_3-6+N)(m_2+1)(m_1+2)(2m_1-4+N)(2m_2-6+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1-5+N)(m_2-6+N)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1 + 1\, m_2\, m_3\, \dot{0})$	$\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-6+N)(m_1-4+N)(m_2+1)(2m_1+N)(2m_2-6+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1+3)(m_2-6+N)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1\, m_2 - 1\, m_3\, \dot{0})$	$\frac{[(N-4)(m_1+1)(m_1-3+N)-(N-2)m_2(m_2-5+N)](m_2+1)(m_2-6+N)-[(N-6)(m_1+2)(m_1-4+N)-(N-4)(m_2+1)(m_2-6+N)]m_3(m_3-6+N)}{(m_2-6+N)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}$
$(m_1\, m_2 + 1\, m_3\, \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1+m_2-4+N)(m_1+m_2-2+N)(m_2-m_3)(m_2-m_3+2)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_2-5+N)(m_2+1)(2m_2-6+N)(2m_2-2+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_2-4+N)\sqrt{(m_2+2)(m_2-6+N)}}$
$(m_1\, m_2\, m_3 - 1\, \dot{0})$	$-\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-4+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2-m_3)(m_2-m_3+2)(2m_2-6+N)(2m_3-8+N)(m_2+1)m_3}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_2-4+N)(2m_3-6+N)(m_2-6+N)(m_3-7+N)}}$
$(m_1\, m_2\, m_3 + 1\, \dot{0})$	$-\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2+m_3-6+N)(m_2+m_3-4+N)(2m_2-6+N)(2m_3-4+N)(m_2+1)(m_3-6+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_2-4+N)(2m_3-6+N)(m_2-6+N)(m_3+1)}}$
$(m_1\, m_2\, m_3\, 1\, \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2-5+N)(2m_2-6+N)(m_3-6+N)m_3(m_1-m_2+2)(m_1+m_2-4+N)(m_2-m_3)(m_2+m_3-6+N)}}{(m_2-6+N)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(2m_2-4+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(4) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2+1\ m_3\ \dot{0})$
$(m_1-1\ m_2\ m_3\ \dot{0})$	$-\frac{\sqrt{(m_1+m_2-4+N)(m_1+m_2-2+N)(m_1-m_3+1)(m_1+m_3-5+N)(m_2-m_3+2)(m_2+m_3-4+N)(m_2-5+N)(m_1+2)(2m_1-4+N)(2m_2-2+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(m_1-5+N)(m_2+2)(2m_1-2+N)(2m_2-4+N)}}$
$(m_1+1\ m_2\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1-m_3+3)(m_1+m_3-3+N)(m_2-m_3+2)(m_2+m_3-4+N)(2m_1+N)(2m_2-2+N)(m_1-4+N)(m_2-5+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)\sqrt{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_2-4+N)(m_1+3)(m_2+2)}}$
$(m_1\ m_2-1\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1-m_2+2)(m_1+m_2-4+N)(m_1+m_2-2+N)(m_2-m_3)(m_2-m_3+2)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_2-5+N)(m_2+1)(2m_2-6+N)(2m_2-2+N)}}{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_2-4+N)\sqrt{(m_2+2)(m_2-6+N)}}$
$(m_1\ m_2+1\ m_3\ \dot{0})$	$\frac{[(N-4)(m_1+1)(m_1-3+N)-(N-2)(m_2+1)(m_2-4+N)](m_2+2)(m_2-5+N)-[(N-6)(m_1+2)(m_1-4+N)-(N-4)(m_2+2)(m_2-5+N)]m_3(m_3-6+N)}{(m_2+2)(2m_2-4+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}$
$(m_1\ m_2\ m_3-1\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_2-5+N)m_3(2m_2-2+N)(2m_3-8+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2+2)(m_3-7+N)(2m_2-4+N)(2m_3-6+N)}}$
$(m_1\ m_2\ m_3+1\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1+m_2-2+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2-m_3)(m_2-m_3+2)(m_2-5+N)(m_3-6+N)(2m_2-2+N)(2m_3-4+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2+2)(m_3+1)(2m_2-4+N)(2m_3-6+N)}}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(2m_2-2+N)(m_3-6+N)m_3(m_1-m_2)(m_1+m_2-2+N)(m_2-m_3+2)(m_2+m_3-4+N)}}{(m_2+2)\sqrt{(m_1+3)(m_1-5+N)(m_2-6+N)(2m_2-4+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(5) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2\ m_3 - 1\ \dot{0})$
$(m_1 - 1\ m_2\ m_3\ \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1-m_3+3)(m_2-m_3+2)(m_2+m_3-6+N)(2m_1-4+N)(2m_3-8+N)(m_1+2)m_3}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1-5+N)(m_3-7+N)}}$
$(m_1 + 1\ m_2\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-2+N)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3+2)(m_2+m_3-6+N)(2m_1+N)(2m_3-8+N)(m_1-4+N)m_3}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1+3)(m_3-7+N)}}$
$(m_1\ m_2 - 1\ m_3\ \dot{0})$	$-\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-4+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2-m_3)(m_2-m_3+2)(2m_2-6+N)(2m_3-8+N)(m_2+1)m_3}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_2-4+N)(2m_3-6+N)(m_2-6+N)(m_3-7+N)}}$
$(m_1\ m_2 + 1\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_2-5+N)m_3(2m_2-2+N)(2m_3-8+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2+2)(m_3-7+N)(2m_2-4+N)(2m_3-6+N)}}$
$(m_1\ m_2\ m_3 - 1\ \dot{0})$	$\frac{-[(N-4)(m_1+1)(m_1-3+N)-(N-2)(m_3-1)(m_3-6+N)]m_3(m_3-7+N)+[(N-6)(m_1+2)(m_1-4+N)-(N-4)m_3(m_3-7+N)](m_2+1)(m_2-5+N)}{(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}$
$(m_1\ m_2\ m_3 + 1\ \dot{0})$	$\frac{\sqrt{(m_1-m_3+1)(m_1-m_3+3)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3)(m_2-m_3+2)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_3-6+N)m_3(2m_3-8+N)(2m_3-4+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_3-6+N)\sqrt{(m_3+1)(m_3-7+N)}}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(m_2-5+N)(m_3-6+N)(2m_3-8+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2-m_3+2)(m_2+m_3-6+N)}}{(m_3-7+N)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(6) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2\ m_3 + 1\ \dot{0})$
$(m_1 - 1\ m_2\ m_3\ \dot{0})$	$-\frac{\sqrt{(m_1-m_2)(m_1+m_2-4+N)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-4+N)(2m_1-4+N)(2m_3-4+N)(m_1+2)(m_3-6+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1-5+N)(m_3+1)}}$
$(m_1 + 1\ m_2\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-2+N)(m_1-m_3+1)(m_1-m_3+3)(m_2-m_3)(m_2+m_3-4+N)(2m_1+N)(2m_3-4+N)(m_1-4+N)(m_3-6+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_1-2+N)(2m_3-6+N)(m_1+3)(m_3+1)}}$
$(m_1\ m_2 - 1\ m_3\ \dot{0})$	$-\frac{\sqrt{(m_1-m_2+2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2+m_3-6+N)(m_2+m_3-4+N)(2m_2-6+N)(2m_3-4+N)(m_2+1)(m_3-6+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(2m_2-4+N)(2m_3-6+N)(m_2-6+N)(m_3+1)}}$
$(m_1\ m_2 + 1\ m_3\ \dot{0})$	$\frac{\sqrt{(m_1-m_2)(m_1+m_2-2+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2-m_3)(m_2-m_3+2)(m_2-5+N)(m_3-6+N)(2m_2-2+N)(2m_3-4+N)}}{(m_2-m_3+1)(m_2+m_3-5+N)\sqrt{(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2+2)(m_3+1)(2m_2-4+N)(2m_3-6+N)}}$
$(m_1\ m_2\ m_3 - 1\ \dot{0})$	$\frac{\sqrt{(m_1-m_3+1)(m_1-m_3+3)(m_1+m_3-5+N)(m_1+m_3-3+N)(m_2-m_3)(m_2-m_3+2)(m_2+m_3-6+N)(m_2+m_3-4+N)(m_3-6+N)m_3(2m_3-8+N)(2m_3-4+N)}}{(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)(2m_3-6+N)\sqrt{(m_3+1)(m_3-7+N)}}$
$(m_1\ m_2\ m_3 + 1\ \dot{0})$	$\frac{-[(N-4)(m_1+1)(m_1-3+N)-(N-2)m_3(m_3-5+N)](m_3+1)(m_3-6+N)+[(N-6)(m_1+2)(m_1-4+N)-(N-4)(m_3+1)(m_3-6+N)](m_2+1)(m_2-5+N)}{(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(m_2-5+N)m_3(2m_3-4+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-4+N)}}{(m_3+1)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$

(m) $\Gamma = (m_1\, m_2\, m_3\, \dot{0})$ (*Continued.*)

(7) $\Gamma_{12} \setminus \Gamma_{23}$	$(m_1\ m_2\ m_3\ 1\ \dot{0})$
$(m_1 - 1\ m_2\ m_3\ \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(2m_1-4+N)(m_2-5+N)(m_2+1)(m_3-6+N)m_3(m_1-m_2)(m_1+m_2-4+N)(m_1-m_3+1)(m_1+m_3-5+N)}}{(m_1-5+N)\sqrt{(m_1+3)(2m_1-2+N)(m_2+2)(m_2-6+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}}$
$(m_1 + 1\ m_2\ m_3\ \dot{0})$	$\frac{\sqrt{(N-6)(m_1+2)(2m_1+N)(m_2-5+N)(m_2+1)(m_3-6+N)m_3(m_1-m_2+2)(m_1+m_2-2+N)(m_1-m_3+3)(m_1+m_3-3+N)}}{(m_1+3)\sqrt{(m_1-5+N)(2m_1-2+N)(m_2+2)(m_2-6+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_1-m_3+2)(m_1+m_3-4+N)}}$
$(m_1\ m_2 - 1\ m_3\ \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2-5+N)(2m_2-6+N)(m_3-6+N)m_3(m_1-m_2+2)(m_1+m_2-4+N)(m_2-m_3)(m_2+m_3-6+N)}}{(m_2-6+N)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(2m_2-4+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$
$(m_1\ m_2 + 1\ m_3\ \dot{0})$	$\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(2m_2-2+N)(m_3-6+N)m_3(m_1-m_2)(m_1+m_2-2+N)(m_2-m_3+2)(m_2+m_3-4+N)}}{(m_2+2)\sqrt{(m_1+3)(m_1-5+N)(m_2-6+N)(2m_2-4+N)(m_3+1)(m_3-7+N)(m_1-m_2+1)(m_1+m_2-3+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$
$(m_1\ m_2\ m_3 - 1\ \dot{0})$	$-\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(m_2-5+N)(m_3-6+N)(2m_3-8+N)(m_1-m_3+3)(m_1+m_3-5+N)(m_2-m_3+2)(m_2+m_3-6+N)}}{(m_3-7+N)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)(m_3+1)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$
$(m_1\ m_2\ m_3 + 1\ \dot{0})$	$\frac{\sqrt{(N-6)(m_1-4+N)(m_1+2)(m_2+1)(m_2-5+N)m_3(2m_3-4+N)(m_1-m_3+1)(m_1+m_3-3+N)(m_2-m_3)(m_2+m_3-4+N)}}{(m_3+1)\sqrt{(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)(m_3-7+N)(2m_3-6+N)(m_1-m_3+2)(m_1+m_3-4+N)(m_2-m_3+1)(m_2+m_3-5+N)}}$
$(m_1\ m_2\ m_3\ 1\ \dot{0})$	$\frac{[m_1(m_1-2+N)+(2N-9)]m_2(m_2-4+N)m_3(m_3-6+N)-(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)+(N-6)[(m_1+2)(m_1-5+N)+m_1]m_3(m_3-6+N)}{(m_1+3)(m_1-5+N)(m_2+2)(m_2-6+N)(m_3+1)(m_3-7+N)}$

belong to the first type. If the two entries change by the opposite interval, i.e., one $+1$ and another -1, then the corresponding Racah coefficients only contain the factors involving m_i+m_j, and hence belong to the second type. Furthermore, all of them are independent of the remaining entries [see Table III(a)–III(l)].

However, compared with Γ_2, if Γ does not change, i.e., $\Gamma=\Gamma_2$, then all Racah coefficients in the corresponding unitary matrix U(Γ) include both the factors involving m_i-m_j and the factors involving m_i+m_j, and hence belong to the second type [see Table III(m)].

(3) Each unitary matrix U(Γ) is symmetric and its matrix elements, SO(N) Racah coefficients, are real. They satisfy the orthonormality relations

$$\sum_{\Gamma_{12}} W\begin{pmatrix} \Gamma_1 & \Gamma_2 & \Gamma_{12} \\ \Gamma_3 & \Gamma & \Gamma_{23} \end{pmatrix} W\begin{pmatrix} \Gamma_1 & \Gamma_2 & \Gamma_{12} \\ \Gamma_3 & \Gamma & \overline{\Gamma}_{23} \end{pmatrix} = \delta(\Gamma_{23},\overline{\Gamma}_{23}). \tag{8}$$

The orthonormality of Racah coefficients in Table III have been checked by software ''MATHEMATICA.'' It is worth reminding readers that when $N=7$, in Table III(d)–III(f), III(j)–III(l) and III(m) the values of the type ($\bullet\bullet\bullet$ 1 $\dot{0}$) taken by Γ_{12} and Γ_{23} should be replaced by ($\bullet\bullet\bullet$).

(4) Similar to the symmetries of SO(3) Racah coefficients in angular momentum theory under the substitution $l\rightarrow -l-1$, SO(N) ($N>3$) Racah coefficients hold the kind of so-called weight-space symmetries[32] as well. For the case in this paper, any number of the independent substitutions

$$m_i\rightarrow -m_i-N+2i, \qquad i=1,2,3, \tag{9}$$

are allowed. They leave the algebraic expressions for SO(N) Racah coefficients untouched, but merely change the relative phases among them and the appearance of the irreducible representations occurring in $W(\cdots)$. When combined with the substitutions (9), the irreducible representations occurring in $W(\cdots)$ are transformed as follows:

$$(m_1\, m_2\, m_3\, \dot{0})\rightarrow(m_1\, m_2\, m_3\, \dot{0}),$$

$$(m_1\, m_2\, m_3 1\, \dot{0})\rightarrow(m_1\, m_2\, m_3 1\, \dot{0}),$$

$$(m_1\pm 2\, m_2\, m_3\, \dot{0})\rightarrow(m_1\mp 2\, m_2\, m_3\, \dot{0}),$$

$$(m_1\, m_2\pm 2\, m_3\, \dot{0})\rightarrow(m_1\, m_2\mp 2\, m_3\, \dot{0}),$$

$$(m_1\, m_2\, m_3\pm 2\, \dot{0})\rightarrow(m_1\, m_2\, m_3\mp 2\, \dot{0}),$$

$$(m_1\pm 1\, m_2\, m_3\, \dot{0})\rightarrow(m_1\mp 1\, m_2\, m_3\, \dot{0}),$$

$$(m_1\, m_2\pm 1\, m_3\, \dot{0})\rightarrow(m_1\, m_2\mp 1\, m_3\, \dot{0}),$$

$$(m_1\, m_2\, m_3\pm 1\, \dot{0})\rightarrow(m_1\, m_2\, m_3\mp 1\, \dot{0}),$$

$$(m_1\pm 1\, m_2\, m_3 1\, \dot{0})\rightarrow(m_1\mp 1\, m_2\, m_3 1\, \dot{0}),$$

$$(m_1\, m_2\pm 1\, m_3 1\, \dot{0})\rightarrow(m_1\, m_2\mp 1\, m_3 1\, \dot{0}),$$

$$(m_1\, m_2\, m_3\pm 1\, 1\, \dot{0})\rightarrow(m_1\, m_2\, m_3\mp 1\, 1\, \dot{0}),$$

$$(m_1\pm 1\, m_2\pm 1\, m_3\, \dot{0})\rightarrow(m_1\mp 1\, m_2\mp 1\, m_3\, \dot{0}),$$

$$(m_1 \pm 1\ m_2\ m_3 \pm 1\ \dot{0}) \rightarrow (m_1 \mp 1\ m_2\ m_3 \mp 1\ \dot{0}),$$

$$(m_1\ m_2 \pm 1\ m_3 \pm 1\ \dot{0}) \rightarrow (m_1\ m_2 \mp 1\ m_3 \mp 1\ \dot{0}),$$

$$(m_1 \mp 1\ m_2 \pm 1\ m_3\ \dot{0}) \rightarrow (m_1 \pm 1\ m_2 \mp 1\ m_3\ \dot{0}),$$

$$(m_1 \mp 1\ m_2\ m_3 \pm 1\ \dot{0}) \rightarrow (m_1 \pm 1\ m_2\ m_3 \mp 1\ \dot{0}),$$

$$(m_1\ m_2 \mp 1\ m_3 \pm 1\ \dot{0}) \rightarrow (m_1\ m_2 \pm 1\ m_3 \mp 1\ \dot{0}).$$

IV. CONCLUSIONS

In this paper, the explicit algebraic expressions for a class of multiplicity-free SO(N) Racah coefficients for $(1\ \dot{0}) \times (m_1\ m_2\ m_3\ \dot{0}) \times (1\ \dot{0})$ are given. These results take on the stronger regularity, therefore, as the base of mathematical induction, they may be extended to the more general case of SO(N) for $(1\ \dot{0}) \times (m_1 m_2 \ldots m_{[N/2]}) \times (1\ \dot{0})$. They, and SO($N$) Racah coefficients of the spinor type for $(\frac{1}{2} \pm \dot{\frac{1}{2}}) \times (m_1 m_2 \ldots m_{[N/2]}) \times (\frac{1}{2} \pm \dot{\frac{1}{2}})$, are being studied.

ACKNOWLEDGMENTS

The authors would like to thank Professor Q. Z. Han and Professor G. L. Long for helpful discussions.

[1] R. Racah, Phys. Rev. **62**, 438 (1942); *ibid.* **63**, 367 (1943).
[2] E. P. Wigner, ''On the matrices which reduce the Kronecker products of representations of simply reducible groups,'' in *Quantum Theory of Angular Momentum*, edited by L. C. Biedenharn and J. D. Louck (Academic, New York, 1965).
[3] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1957).
[4] B. R. Judd, *Operator Techniques in Atomic Spectroscopy* (McGraw–Hill, New York, 1963).
[5] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison–Wesley, Reading, MA, 1981).
[6] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonsk, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).
[7] P. H. Butler, *Point Group Symmetry Applications* (Plenum, New York, 1981).
[8] J. Raynal and R. Conte, J. Math. Phys. **26**, 2413, 2441 (1985).
[9] R. W. Haase and D. Dirl, J. Math. Phys. **27**, 900 (1986).
[10] H. Z. Sun and Q. Z. Han, Sci. Sin. **24**, 914 (1981).
[11] J. Q. Chen, P. N. Wang, Z. M. Lu, and X. B. Wu, *Tables of the CG, Racah and Subduction Coefficients of $SU(n)$ Groups* (World Scientific, Singapore, 1987)
[12] C. Quesne, J. Math. Phys. **31**, 1600 (1990).
[13] K. T. Hecht, Nucl. Phys. A **102**, 11 (1967).
[14] B. R. Judd, G. M. S. Lister, and M. C. M. O'Brien, J. Phys. A **19**, 2473 (1986).
[15] B. R. Judd, R. C. Leavitt, and G. M. S. Lister, J. Phys. A **23**, 385 (1990).
[16] S. Ališauskas, J. Phys. A **20**, 35 (1987).
[17] B. R. Judd and G. M. S. Lister, J. Phys. A **20**, 3159 (1987).
[18] B. R. Judd, J. Math. Phys. **27**, 2616 (1986).
[19] R. P. Bickerstaff and B. G. Wybourne, J. Phys. G **7**, 275 (1981).
[20] B. G. Searle and P. H. Butler, J. Phys. A **21**, 3041 (1988).
[21] J. D. Vergados, Nucl. Phys. A **111**, 681 (1968).
[22] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[23] E. El-Baz, J. Lafoucrière, and B. Castel, *Traitement Graphique de l'Algébre des Moments Angulaires* (Masson & Cie, Paris, 1969).
[24] Y. Bordarier, Contribution à l'emploi de Méthodes Graphiques en spectroscopie Atomique, Thèse, Faculté des Sciences d'Orsay, Université de Paris, 1970.
[25] H. Z. Sun and D. Ruan, J. Math. Phys. **39**, 630 (1998).
[26] H. Z. Sun, Q. Z. Han, M. Zhang, and D. Ruan, Commun. Theor. Phys. **30**, 1080 (1998).
[27] H. Z. Sun, Acta Phys. Sin. **20**, 483 (1964).
[28] Q. Z. Han and H. Z. Sun, *Group Theory* (Peking U.P., Beijing, 1987).
[29] H. A. Jahn, Proc. R. Soc. London, Ser. A **205**, 192 (1951).
[30] D. Ruan and H. Z. Sun, to appear in Commun. Theor. Phys. (1998).
[31] G. L. Long, J. Phys. A **28**, 64 (1995); G. L. Long, J. Y. Zhang, and W. L. Zhang, to appear in J. Phys. A (1998).
[32] A. P. Jucys and A. A. Bandzaitis, *Theory of Angular Momentum in Quantum Mechanics* (Mokslas, Vilnius, 1977) (in Russian).

Erratum: "All stationary vacuum metrics with shearing geodesic eigenrays" [J. Math. Phys. 13, 1695 (1972)]

J. Kóta and Z. Perjés
KFKI Research Institute for Particle and Nuclear Physics, Budapest, 114, P.O. Box 49, H-1525 Hungary

(Received 14 May 1998; accepted for publication 15 May 1998)

[S0022-2488(98)01711-3]

Three vacuum metrics have been obtained in the paper. In the third metric, given in Eq. (62), the y factors were misplaced. Here is how the metric should read:

$$ds^2=-(f^0y^2/f)(r^{1-\gamma^0}dx^2+r^{1+\gamma^0}dy^2)+2dr[dt-\gamma^0x^2ydy]+f[dt-\gamma^0x^2ydy]^2.$$

 © 1998 American Institute of Physics

Erratum: "On quantization of Z_2-graded algebras" [J. Math. Phys. 38, 476–483 (1997)]

A. I. Mudrov[a)]
Institute of Physics, St. Petersburg State University, St. Petersburg 198904, Russia

(Received 12 January 1998; accepted for publication 12 January 1998)

[S0022-2488(98)01707-1]

There is an inaccuracy on p. 479. The linear mapping $U(\tilde{\mathbf{V}}_1^*)\to U^*(\tilde{\mathbf{V}}_1)$ in the diagram (right downward arrow) should be replaced by $U'(\tilde{\mathbf{V}}_1^*)U(\tilde{\mathbf{V}}_{n-1}^*)\to[U(\tilde{\mathbf{V}}_{n-1})U'(\tilde{\mathbf{V}}_1)]^*$, the prime denoting the ideal spanned by the monomials of positive degrees.

[a)]Electronic mail: aimudrov@DG2062.spb.edu

 © 1998 American Institute of Physics

Wigner distribution function for finite systems

Natig M. Atakishiyev,[a),b)] Sergey M. Chumakov, and Kurt Bernardo Wolf
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas—Cuernavaca, Universidad Nacional Autónoma de México, Apartado Postal 48-3, 62251 Cuernavaca, Morelos, México

(Received 9 May 1997; accepted for publication 11 August 1998)

We construct a Wigner distribution function for finite data sets. It is based on a finite optical system; a linear wave guide where the finite number of discrete sensors is equal to the number of modes which the guide can carry. The dynamical group for this model is SU(2) and the wave functions are sets of $N=2l+1$ data points. The Wigner distribution function assigns classical *c*-numbers to the operators of position, momentum, and wave guide mode. © *1998 American Institute of Physics.* [S0022-2488(98)00312-0]

I. CONTINUOUS AND FINITE SIGNAL ANALYSIS

The phase-space representation of data sets can be analyzed with the aid of the Wigner distribution function. This was introduced originally in quantum mechanics, where the ''signal'' is a wave function $\psi(q)$, and its Wigner function (for $\hbar=1$) is defined as[1]

$$W_\psi(q,p)=\frac{1}{2\pi}\int_{\mathfrak{R}} dx[\,\psi(q-\tfrac{1}{2}x)]^* e^{-ixp}\psi(q+\tfrac{1}{2}x). \tag{1.1}$$

When (q,p) are understood as canonically conjugate coordinates of a phase-space plane, the value of $W_\psi(q,p)$ mirrors closely the intuitive objects in the model. They can be ''particles'' in quantum mechanics, with position q and canonically conjugate momentum p; in quantum optics these objects may be the coherent states of the radiation field; in monochromatic paraxial wave optics they are often beams with Gaussian position and inclination distributions. The Heisenberg uncertainty relation is built into the Wigner picture by the Fourier transform between the position and momentum representations.

The importance of the Wigner distribution function on finite data sets can be explained in terms of Lohmann's rendering of the Wigner function in music,[2] whereas the graphs of the acoustic signal or the frequency spectrum of a performance are meaningless to visual inspection, the Wigner function will exhibit peaks at positions that are the notes in a pentagram. In effect, the Wigner function is a musical score—*partitura* of the data set $\{\psi(q)\}$, $q\in\mathfrak{R}$, which also contains the information of its Fourier transform function $\{\tilde{\psi}(p)\}$, $p\in\mathfrak{R}$. Lohmann also presents optical devices to produce essentially a photograph of the Wigner function of a line segment of data on a slide.[3]

In this paper we analyze specifically the phase-space representation of finite data sets. We are interested in the features specific to finite systems. They include the discreteness of measurement and operator spectra, and the fact that these operators are realizations of a compact Lie algebra of *difference* operators. Physicists using Lie-theoretical methods are much more familiar with algebras of *differential* operators. We should stress that the domain of difference operators are functions on the complex plane; in the same way as the difference equation $\Gamma(x+1)=x\Gamma(x)$ generalizes the factorial product of the nonnegative integers to any $x\in\mathfrak{R}$, difference operators will generally relate three neighboring points of x separated by unity. Moreover, finite difference equations have a recognizedly richer structure of solutions than their limit differential equations.[4] When the interval and density of data points approach infinity, the results obtained will match with those of the standard formalism for continuous signals.

[a)]Electronic mails: natig@, sergey@, and bwolf@ce.ifisicam.unam.mx
[b)]Instituto de Matemáticas, UNAM, Mexico.

 © 1998 American Institute of Physics

We consider $2l+1$ complex data values

$$\mathbf{f}=\{f_m | m=-l,-l+1,...,l\}, \quad f_m \in \mathcal{C}, \tag{1.2a}$$

where $2l+1$ is a non-negative integer. These can be thought of as complex light amplitude values measured on an array of $2l+1$ points on a screen, equidistant by $\lambda>0$ and centered at the origin, i.e.,

$$f_m=f(q_m), \quad q_m=m\lambda, \quad m=-l,-l+1,...,l. \tag{1.2b}$$

In ordinary finite Fourier analysis one asks for the exponential Fourier transform to provide the wave number content $\tilde{\mathbf{f}}$ of the signal $\mathbf{f}$ in the $2l+1=N$-dimensional orthonormal basis $N^{-1/2}\exp(2\pi nm/N)$, $n=1,...,N$, consisting of Brillouin waves in a periodic lattice. This basis is adequate for a model of a finite system if the system is homogeneous under (dihedral) rotations of the lattice, so the points q_{-l} and q_l are first neighbors. When the two points q_{-l} and q_l are poles apart of an array where the middle portion carries the most significant part of the information, a different basis and transform are called for. Moreover, since the spectrum of the finite second-difference matrix is not equally spaced, the finite Fourier-exponential transform[5] is a poor basis for Lie-theoretical classification and evolution treatment of finite data in a system.

In Sec. II we present the finite wave guide model.[6] This is a physical system which embodies well the mathematical developments we present below; it is a *finite oscillator*. As will become clear at the end of the paper, the wave guide model serves as a working definition of *finite optics*. We start with Newton's equation for the classical harmonic oscillator, and show that both the usual infinite-spectrum quantum mechanical oscillator and the present finite-spectrum oscillator satisfy that equation; the former with the Heisenberg–Weyl algebra, the latter with the su(2) compact algebra. We stress that our approach uses the group of 2×2 unimodular unitary matrices (twofold cover of the rotation group in three dimensions), rather than the usual Heisenberg–Weyl group, as an arena for the Wigner function. In this way, we are assured of the existence of a *position* operator whose spectrum, $m=-l,-l+1,...,l$, is discrete and finite.

Section III defines the Wigner operator in the group ring,[7] and the Wigner distribution function as a bilinear form of the data values, which depend on the classical position, momentum, and energy variables. Section IV examines the SU(2) covariance of the Wigner function and addresses its computation by using group theoretical properties. It is clarified that for each fixed number of sensors $N=2l+1$, the essential information can be plotted and inspected visually on the surface of a two-sphere. Coherent SU(2) states are examined in Sec. V, where we also plot other interesting data sets, such as Schrödinger-cat states with their concomitant interference phenomenon. The concluding Sec. VI recapitulates our construction and indicates other finite systems with various dynamical laws, which can be analyzed by the same mathematical tools and physical concepts.

II. FINITE WAVE GUIDE MODEL AND THE SU(2) GROUP

Finite data sets and their parallel processing by optical means will be based on a planar multimodal wave guide, such as would be part of photonic devices fabricated by doping a strip on a transparent substratum.[6] The *finite* wave guide model has a refractive index whose profile is parabolic (i.e., of the form $n(q)=n_0-\nu q^2+\cdots$ in a neighborhood of its axis); it acts as a harmonic oscillator on the input wave field (signal) produced by a linear array of N coherent light sources; the output is received by the same number of wave field sensors. We stress that our wave guide model is capable of carrying only a *finite* number of modes, and corresponds to a quantum system with finite number of bound states equidistant in energy (see Fig. 1).

A. Newton equation for the finite oscillator

A unit mass in the classical one-dimensional harmonic oscillator potential $V(q)=\frac{1}{2}\omega^2 q^2$ obeys Newton's equation $\ddot{q}=-\omega^2 q$ (the dots indicate time derivatives); the same equation applies in geometric optics to the transverse ray coordinate within a wave guide. As in the quantization (or *wavization*) of geometric to paraxial wave (Fourier) optics,[8] we assume that there exists a *position* operator Q, and time derivatives of the classical observables are replaced by Lie brackets (commutators) of the corresponding operators with a Hamiltonian evolution operator H (times i). For the systems at hand, Newton's equation thereby becomes the *Lie–Newton equation*

$$[H,\,[H,\,Q]]=\omega^2 Q. \tag{2.1a}$$

By definition, the commutator of the Hamiltonian and position operators yields the *momentum* operator

$$P=i[H,\,Q], \tag{2.1b}$$

so the Lie–Newton equation (2.1a) becomes

$$[H,\,P]=i\omega^2 Q. \tag{2.1c}$$

The three operators Q, P, and H satisfy two commutation relations, (2.1b) and (2.1c), which embody the *Hamilton* equations of oscillator/wave guide systems; their geometry and dynamics, respectively. Notice that the third commutator, $[P, Q]$, is so far unspecified; for the three operators to close into a Lie algebra, their Jacobi identity requires that $[H,\,[P,\,Q]]=0$. This implies that $[P, Q]$ is constant under evolution by H, and has the form

$$[P,\,Q]=i(\sigma H+C), \tag{2.2}$$

where $\sigma\in\{+1,0,-1\}$, and C is any operator that commutes with the original three, i.e., is in the center of the Lie algebra. [We placed the i in (2.2) to use self-adjoint operators below.] The values of the coefficient σ determine the *dynamical* Lie algebra to be su(2), iso(2) or su(1,1), respectively, and C may be a function of the Casimir invariant plus/or the generator of a central extension, plus/or a constant. [If a nonlinear function of H is placed in (2.2), other algebraic structures may arise; we will not examine them here.]

In the familiar formulation of quantum phase space with operators Q and P, one sets $\sigma=0$ and $C=\hbar\mathbf{1}$; this defines the Heisenberg–Weyl–Lie algebra, and Eqs. (2.1) are satisfied when $H=\frac{1}{2}(P^2+\omega^2Q^2)$, in accordance with the classical Hamiltonian formulation.[9] In Fourier polychromatic optics, the natural constant $\hbar$ is replaced by the reduced wavelength $\lambda/2\pi$, $\lambda\in\Re-\{0\}$.[7] In Ref. 10 we examined the case when in (2.1a) and (2.1c) the oscillator frequency ω is zero, and in (2.2) $\sigma=-1$, $C=0$; we have then a Euclidean dynamical Lie algebra, iso(2). In this article we shall examine the su(2) structure that fits the finite wave guide model.

B. Difference operators for finite systems

In finite systems, the position coordinate ranges over a finite (discrete) set of integer (or half-integer) values, $q_m=m$, $m=-l,-l+1,...,l$. This set of values we here interpret as the *spectrum* of the position operator Q of the model, in a $(2l+1)$-dimensional vector space $\Re^{2l+1}$ of signals (1.2).

Denoting by $q\in\Re$ the continuous coordinate whose integer (or half-integer) values are the normalized sensor positions m, and using the right- and left-shift operators $e^{\pm\partial_q}f(q)=f(q\pm1)$, we recall from Refs. 6 and 11 the *difference* operators

$$J_1=Q=q, \tag{2.3a}$$

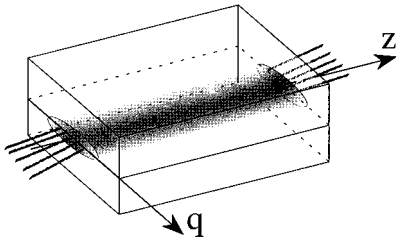

FIG. 1. Finite wave guide model. A shallow, planar wave guide doped into a transparent substrate, capable of confining only $2l+1$ transverse modes, transmits in parallel at most that number of signals to the same number of sensors.

$$J_2 = -P = -i\tfrac{1}{2}[\alpha_l(-q)e^{\partial_q} - \alpha_l(q)e^{-\partial_q}], \quad \alpha_l(q) = \sqrt{(l+q)(l-q+1)}, \tag{2.3b}$$

$$J_3 = H - l - \tfrac{1}{2} = -\tfrac{1}{2}[\alpha_l(q)e^{-\partial_q} + \alpha_l(-q)e^{\partial_q}]. \tag{2.3c}$$

These operators satisfy the Lie–Hamilton commutation relations (2.1) and (2.2) with $\sigma = +1$ and $C = -(l+\frac{1}{2})\mathbf{1}$, as generators of a Lie algebra su(2),

$$[J_1, J_2] = iJ_3, \quad [J_2, J_3] = iJ_1, \quad [J_3, J_1] = iJ_2. \tag{2.4}$$

Since the Casimir operator in this realization is $J^2 = J_1^2 + J_2^2 + J_3^2 = l(l+1)\mathbf{1}$, it is equivalent to the SU(2) unitary irreducible representation familiar from quantum angular momentum theory, but with J_1 identified as the position operator.

In this way the angular momentum algebra su(2) is woven into the foundation of finite systems, i.e., systems with a finite number of independent states, and in particular of the multimodal planar wave guide model. The spectra of positions, momenta, and energies are intrinsically discrete, finite, and equally-spaced. When we change the basis (2.3), so that the Hamiltonian $H = J_3 + l + \frac{1}{2}$ be diagonal, we find the eigenvalue equation

$$H(q)\phi_n(q,2l) = (n+\tfrac{1}{2})\phi_n(q,2l), \quad n = 0,1,...,2l, \tag{2.5}$$

that governs the wave field normal modes in the finite wave guide. The eigenfunctions $\phi_n(q,2l)$ can be written in terms of Kravchuk polynomials,[12] which satisfy a discrete orthogonality relation with a binomial weight function. Multiplying the polynomials with the square root of the latter defines the Kravchuk *functions*,[6,13] which are continuously defined in the interval $-l-1 \leqslant q \leqslant l+1$ and satisfy the discrete orthogonality relation $\Sigma_{m=-1}^{l} \phi_n'(m,2l)\phi_{n'}(m,2l) = \delta_{n,n'}$. In Ref. 13 we showed that when $l \to \infty$ and the spacing between points vanishes as l^{-1}, they become the usual quantum harmonic oscillator wave functions.

C. Polar parametrization of SU(2)

Exponentiation of the Lie algebra of operators (2.3) yields the elements of the SU(2) group, whose generic element can be parametrized by a three-vector $\mathbf{y}$, whose unit direction vector $\mathbf{v} = \mathbf{y}/|\mathbf{y}|$ is the rotation axis, and whose length $\eta = |\mathbf{y}|$ is the rotation angle around that axis, namely,

$$g[\mathbf{y}] = g[\eta\mathbf{v}] = \exp(-i\mathbf{y}\cdot\mathbf{J}) = \exp[-i\eta(v_1J_1 + v_2J_2 + v_3J_3)]. \tag{2.6}$$

$$\mathbf{y} = \eta\mathbf{v}(\theta,\phi), \quad \mathbf{v} = \begin{pmatrix} \sin\theta\sin\phi \\ \sin\theta\cos\phi \\ \cos\theta \end{pmatrix}, \quad \begin{cases} 0 \leqslant \eta < 4\pi, \\ 0 \leqslant \theta \leqslant \pi, \\ 0 \leqslant \phi < 2\pi. \end{cases} \tag{2.7}$$

This is the *polar* parametrization of the group, indicated hereafter by the square brackets; it provides an $\mathfrak{R}^3$ coordinate system where the unit element is at the origin, $g[\mathbf{0}]$, for all directions (θ,ϕ) and the SU(2) ''minus unit,'' also in the center (commutant) of the group, is the element $g[2\pi\mathbf{v}(\theta,\phi)]$, also for all directions. The vanishing of the Haar measure at $\eta = 0,2\pi$ in Eq. (2.10) below, indicates that each of these element counts for a single point of the manifold. In polar coordinates, the nth power is simply expressed as $(g[\mathbf{y}])^n = g[n\mathbf{y}]$; this defines fractional powers and yields the inverse element for $n = -1$. The angle η is counted modulo 4π and labels the conjugation classes of the group; the group elements within each class are distinguished by the axis $\mathbf{v}(\theta,\phi) \in \mathcal{S}_2$ on the two-sphere. (Relations with the more commonly used Euler angle parametrization $g(\phi,\theta,\psi)$ (Refs. 14, 15), indicated by round brackets, will appear below.)

The three generators $\mathbf{J} = (J_1, J_2, J_3)$ transform under the action of the SU(2) group as the components a row vector under the *adjoint* representation $\mathbf{R}$ by 3×3 orthogonal matrices

$$g[\mathbf{z}]\mathbf{J}(g[\mathbf{z}])^{-1} = \mathbf{J}\mathbf{R}[\mathbf{z}]. \tag{2.8}$$

The generic operator in the su(2) algebra has the form $\mathbf{y}\cdot\mathbf{J} = \Sigma_{i=1}^{3} y_i J_i$; therefore, from (2.6) and writing $\mathbf{y} = \eta\mathbf{v}(\theta,\phi)$, it follows that in polar parametrization,

$$g[\mathbf{z}]g[\mathbf{y}](g[\mathbf{z}])^{-1}=g[\mathbf{R}[\mathbf{z}]\mathbf{y}]=g[\,\eta\mathbf{R}[\mathbf{z}]\mathbf{v}]. \tag{2.9}$$

In these coordinates, the invariant measure [for which $\int_{SU(2)} dg\ f(g)=\int_{SU(2)} dg\ f(gg_0)=\int_{SU(2)} dg\ f(g^{-1})$] is[14,15]

$$dg[\mathbf{y}]=\gamma(\eta)d\mathbf{y}=\tfrac{1}{2}\sin^2\tfrac{1}{2}\eta d\,\eta d\mathbf{v},\quad d\mathbf{v}(\vartheta,\varphi)=\sin\vartheta d\vartheta d\varphi, \tag{2.10a}$$

where $d\mathbf{y}=dy_1\,dy_2\,dy_3$ is the Cartesian measure in $\mathfrak{R}^3$, the weight function is

$$\gamma(\eta)=\frac{\frac{1}{2}\sin^2\frac{1}{2}\eta}{\eta^2}=\tfrac{1}{8}\,\mathrm{sinc}\,\tfrac{1}{2}\eta, \tag{2.10b}$$

and $\mathrm{sinc}\,x=x^{-1}\sin x$ is the *sinus cardinalis* function. Finally, we recall that the SU(2) manifold is the three-sphere $\mathcal{S}_3$ and $\mathrm{vol\,SU(2)}=\int_{SU(2)}dg=2\pi^2=\mathrm{vol}\,\mathcal{S}_3$.

On the su(2) algebra, the Hamiltonian generates evolution through $g[z,\mathbf{k}]=\exp(-izH)$, i.e., a rotation by the angle z around the unit vector $\mathbf{k}$ along the three-axis. In the finite wave guide model defined by (2.3), it represents evolution of the $\mathfrak{R}^{2l+1}$ wave fields along the optical z-axis of the wave guide (times a common phase $\exp[-iz(l+\frac{1}{2})]$). Other useful optical elements are also contained in this su(2) model; a thin *wedge* (prism) of distinct refracting index in the wave guide, of small angle ω, multiplies the signal along the sensor array by $e^{-i\omega m}$; this is generated by Q and is a rotation by ω around the one-axis of the group. A tilted plate of such material *translates* signal positions at the center of the guide; it is a rotation around the two-axis of the group generated by P.

III. WIGNER OPERATOR AND DISTRIBUTION FUNCTION

In this section we define the Wigner operator on a Lie group[7] and examine the SU(2) Wigner function associated with the finite wave guide model.

A. Wigner operator and its covariance

Consider a d-parameter Lie group $\mathcal{G}$ with generators $\mathbf{J}=\{J_i\}_{i=1}^d$, whose elements in polar parametrization are $g[\mathbf{y}]=\exp(-i\mathbf{y}\cdot\mathbf{J})$, where $\mathbf{y}\cdot\mathbf{J}:=\Sigma_{j=1}^D y_jJ_j$. We define the *Wigner operator* as the family of formal operators, function of $\mathbf{x}\in\mathfrak{R}^d$, given by

$$\mathcal{W}(\mathbf{x})=\int_{\mathcal{G}}dg[\mathbf{y}]\exp[i\mathbf{y}\cdot(\mathbf{x}-\mathbf{J})]=\int_{\mathcal{G}}dg[\mathbf{y}]\exp(i\mathbf{x}\cdot\mathbf{y})g[\mathbf{y}]. \tag{3.1}$$

In the case of the unitary group SU(2), $d=3$ and the dot product is the rotation-invariant scalar product between three-vectors, $\mathbf{x}\cdot\mathbf{y}=\Sigma_{i=1}^3 x_iy_i$. The direct integral of group elements is well defined when the operator acts on functions over a homogeneous space under the group; the SU(2)-invariant measure is (2.10).

Note first that the Wigner operator at the coordinate origin, $\mathcal{W}(\mathbf{0})=\mathcal{I}$, is the unit element in the SU(2) group ring of formal operators $\mathcal{A}=\int_{\mathcal{G}} dg\ A(g)g$ (i.e., $\mathcal{I}\mathcal{A}=\mathcal{A}\mathcal{I}=\mathcal{A}$). Next, observe that the range of the polar parameters $\mathbf{y}$ is the compact SU(2) manifold, while the space $\mathbf{x}$ in the argument of the Wigner operator is $\mathfrak{R}^3$. Moreover, the function $\exp(i\mathbf{x}\cdot\mathbf{y})$ in (3.1) is not periodic over the group; for general $\mathbf{x}$ it has different values at $g[\,\eta\mathbf{v}]$ and at $g[(\eta+4\pi)\mathbf{v}]$, which represent the same element $g[\mathbf{y}]$. We must therefore fix the definition of the Wigner operator (3.1) by imposing an otherwise natural condition for operators in Lie group rings, self-adjointness. We consider the group elements $g[\mathbf{y}]$ in (3.1) acting as unitary operators on an appropriate Hilbert space, so their operator adjoint is their inverse, $(g[\mathbf{y}])^\dagger=g[-\mathbf{y}]$. Then, we obtain the adjoint of $\mathcal{W}(\mathbf{x})$ through complex conjugation of the exponential function; changing integration variables by $\mathbf{y}\mapsto-\mathbf{y}$, we reproduce (3.1) only if $\int_{SU(2)} dg[-\mathbf{y}]\cdots=\int_{SU(2)} dg[\mathbf{y}]\cdots$. The range of $\mathbf{y}=\eta\mathbf{v}(\theta,\phi)$ will be the same as that of its inverse $-\mathbf{y}$ if we agree that η, naturally ranging in $[0,4\pi)$ in accordance with (2.6)–(2.7) and counted modulo 4π, will be integrated over $(-2\pi,2\pi]$. Only in this way the Wigner operator will be self-adjoint (in the same Hilbert space): $[\mathcal{W}(\mathbf{x})]^\dagger=\mathcal{W}(\mathbf{x})$ for all $\mathbf{x}\in\mathfrak{R}^3$.

The Wigner operator (3.1) satisfies the important property of SU(2)-*covariance*,

$$g[\mathbf{y}]\mathcal{W}(\mathbf{x})(g[\mathbf{y}])^{-1}=\mathcal{W}(\mathbf{R}[\mathbf{y}]\mathbf{x}). \tag{3.2a}$$

This is a consequence of the bilateral invariance of $dg[\mathbf{y}]$ and the property (2.9) of the adjoint representation; it follows from rewriting the left-hand side of (3.2a) as

$$\int_{\mathrm{SU}(2)} dg[\mathbf{y}]\exp(i\mathbf{y}'\cdot\mathbf{x})g[\mathbf{R}\mathbf{y}']=\int_{\mathrm{SU}(2)} dg[\mathbf{y}'']\exp(i\mathbf{y}''\cdot\mathbf{R}\mathbf{x})g[\mathbf{y}''], \tag{3.2b}$$

where $\mathbf{R}=\mathbf{R}[\mathbf{y}]$ and $\mathbf{y}''=\mathbf{R}\mathbf{y}'$, to obtain the right-hand side of the equation.

B. Wigner matrix

The action of the Wigner operator (3.1) on column vectors $\mathbf{f}\in\mathfrak{R}^{2l+1}$ [whose vector components are the observed wave field values (1.2)], is

$$\mathcal{W}(\mathbf{x})\mathbf{f}=\int_{\mathrm{SU}(2)} dg[\mathbf{y}]\exp(i\mathbf{x}\cdot\mathbf{y})\mathbf{D}^l[\mathbf{y}]\mathbf{f}=\mathbf{W}^l(\mathbf{x})\mathbf{f}, \tag{3.3}$$

where $\mathbf{D}^l[\mathbf{y}]$ are the spin-l unitary irreducible representation matrices of SU(2) in polar coordinates, well known as the *Wigner D-matrices* in angular momentum theory.[14,15] We call $\mathbf{W}^l(\mathbf{x})$ the SU(2) *Wigner W-matrices*; they are the *matrix representation* of the SU(2) Wigner operator on $\mathfrak{R}^{2l+1}$.

The Wigner *D*- and *W*-matrices are essentially Fourier conjugates of each other; the integral in Eq. (3.3) can be inverted because it has the transform kernel $\exp(i\mathbf{x}\cdot\mathbf{y})$. Thus, the transform pair is

$$\mathbf{W}^l(\mathbf{x})=\int_{\mathrm{SU}(2)} \gamma(|\mathbf{y}|)d\mathbf{y}\ \exp(i\mathbf{x}\cdot\mathbf{y})\mathbf{D}^l[\mathbf{y}], \tag{3.4a}$$

$$\mathbf{D}^l[\mathbf{y}]=\frac{1}{(2\pi^3)\gamma(|\mathbf{y}|)}\int_{\mathfrak{R}^3} d\mathbf{x}\ \exp(-i\mathbf{x}\cdot\mathbf{y})\mathbf{W}^l(\mathbf{x}). \tag{3.4b}$$

From (3.4) we see that since the Wigner *D*-matrices are unitary, $(\mathbf{D}^l[\mathbf{y}])^\dagger=\mathbf{D}^l[-\mathbf{y}]$, and since we have chosen the integration range to be invariant under inversions, it follows that $\mathbf{W}^l(\mathbf{x})^\dagger=\mathbf{W}^l(\mathbf{x})$, i.e., the Wigner *W*-matrices are self-adjoint. Therefore, their eigenvalues, determinant and trace, are real. Also from (3.4), the known integral and a special value of the *D*-matrices imply

$$\mathbf{W}^l(\mathbf{0})=\int_{\mathrm{SU}(2)} dg[\mathbf{y}]\mathbf{D}^l[\mathbf{y}]=2\pi^2\delta_{l,0}, \tag{3.5a}$$

$$\int_{\mathfrak{R}^3} d\mathbf{x}\mathbf{W}^l(\mathbf{x})=(2\pi)^3\gamma(\mathbf{0})\mathbf{D}^l[\mathbf{0}]=\pi^3\mathbf{1}. \tag{3.5b}$$

The unitarity of the *D*-matrices in (3.4a) further determines the integral

$$\int_{\mathfrak{R}^3} d\mathbf{x}[\mathbf{W}^l(\mathbf{x})]^\dagger\mathbf{W}^l(\mathbf{x})=(2\pi)^3\int_{\mathrm{SU}(2)}[\gamma(|\mathbf{y}|)]^2d\mathbf{y}\mathbf{1}=\pi^4 s_4\mathbf{1}, \tag{3.6}$$

where $s_4=\int_0^{2\pi}\mathrm{sinc}^4\ z\ dz=1.0467...$ is a constant. Finally, notice that (3.4) is a transform pair between functions W of $\mathbf{x}\in\mathfrak{R}^3$ and functions D of $g[\mathbf{y}]\in\mathrm{SU}(2)$ with support inside a sphere of radius 2π in $\mathfrak{R}^3$. Payley–Wiener theorems should yield analiticity properties of the Wigner *W*-matrices that will be explored elsewhere.

C. Wigner function and its covariance

We define the *Wigner distribution function* of the finite signals $\mathbf{f}=\{f_m\}_{m=-l}^l$ and $\mathbf{g}=\{g_m\}_{m=-l}^l$ as the sesquilinear form

$$W^l(\mathbf{f},\mathbf{g}|\mathbf{x}) = \mathbf{f}^\dagger \mathbf{W}^l(\mathbf{x})\mathbf{g} = \sum_{m,m'=-l}^{l} [f_m]^* W^l_{m,m'}(\mathbf{x}) g_{m'}. \tag{3.7}$$

The Wigner matrix elements are, from (2.7), (2.10), and (3.4a),

$$W^l_{m,m'}(\chi\mathbf{u}) = \int_{-2\pi}^{2\pi} \tfrac{1}{2}\sin^2\tfrac{1}{2}\eta d\eta \int_{\mathcal{S}_2} d\mathbf{v}\, \exp(i\chi\eta\mathbf{u}\cdot\mathbf{v}) D^l_{m,m'}[\eta\mathbf{v}]. \tag{3.8}$$

In Dirac's notation one can usefully write $W^l_{m,m'}(\mathbf{x}) = \langle l,m|\mathcal{W}(\mathbf{x})|l,m'\rangle$. When $\mathbf{f}=\mathbf{g}$, we indicate the Wigner function simply by $W^l(\mathbf{f}|\mathbf{x})$; this is the expectation value of the Wigner operator in the state $\mathbf{f}$.

From the SU(2)-covariance of the Wigner operator (3.2) follows the SU(2)-covariance of the Wigner matrix and of the Wigner function,

$$\mathbf{D}^l[\mathbf{y}]\mathbf{W}^l(\mathbf{x})\mathbf{D}^l[-\mathbf{y}] = \mathbf{W}^l(\mathbf{R}[\mathbf{y}]\mathbf{x}), \tag{3.9}$$

$$W^l(\mathbf{D}^l[-\mathbf{y}]\mathbf{f},\mathbf{D}^l[-\mathbf{y}]\mathbf{g}|\mathbf{x}) = W^l(\mathbf{f},\mathbf{g}|\mathbf{R}[\mathbf{y}]\mathbf{x}). \tag{3.10}$$

One consequence is that a signal vector $\boldsymbol{\alpha}^\mu = \{\alpha^\mu_m\}^l_{m=-l}$ which is an eigenvector of any operator $J_a = \mathbf{a}\cdot\mathbf{J}$ in the su(2) algebra and with the eigenvalue μ, will have a Wigner function that will be invariant under rotations of $\mathbf{x}$ around the axis $\mathbf{a}$. This is so because rotation by $\exp(-i\omega J_a)$ will multiply its eigenvector $\boldsymbol{\alpha}^\mu$ by $\exp(-i\omega\mu)$; but in (3.3)–(3.7) we see that the Wigner function is insensitive to common signal phase factors, $W^l(e^{i\phi}\mathbf{f}|\mathbf{x}) = W^l(\mathbf{f}|\mathbf{x})$. Such is the eigenbasis of *normal modes* of the finite wave guide, which are the eigenfunctions of the Hamiltonian H and given by the Kravchuk functions (2.5).

D. Meta-phase space

The definition, properties, and use of the common Wigner distribution function (1.1) were briefly discussed in the Introduction. We propose (3.7) as the proper Wigner function for finite signals $\mathbf{f}$ in a planar multimode wave guide. Equations (2.3)–(2.6) establish a correspondence between the (c-number) arguments of $W^l(\mathbf{f}|\mathbf{x})$ and the physical observables of a wave field in a shallow waveguide, namely,

$$\mathbf{x} = \chi\boldsymbol{\mu}(\vartheta,\varphi) = \begin{pmatrix} q \\ -p \\ E-l-\frac{1}{2} \end{pmatrix} = \begin{pmatrix} \text{position} \\ -\text{momentum} \\ \text{energy}-l-\frac{1}{2} \end{pmatrix}. \tag{3.11}$$

Because the first two coordinates are of ordinary phase space, let us call the three-dimensional space of $\mathbf{x}\in\mathfrak{R}^3$ the *metaphase space* of the finite wave guide model. When l is large and E is small, in a neighborhood of the South pole we can approximate the sphere $|\mathbf{x}|^2 = q^2+p^2+(e-l-\frac{1}{2})^2 \approx l^2$, by the osculating paraboloid $E \approx (p^2+q^2)/2l$, which corresponds to the classical energy in an oscillator. We should thus expect the values of the Wigner function to peak in $\chi = |\mathbf{x}|$ between l and $l+1$. This will be borne out below.

IV. WIGNER MATRIX AND PROJECTIONS

With the aim of computing the Wigner function of finite signals, we now find expressions for the D- and W-matrix elements and some of their main properties. To understand the results in three-dimensional space, we examine spherical slices and projections.

A. D^l matrix elements in polar coordinates

For analytic and computational work it is convenient to use both polar and Euler angle coordinate systems in SU(2). Their relation is

$$g[\mathbf{y}] = g(\phi,\theta,0)g(0,0,\eta)[g(\phi,\theta,0)]^{-1}, \quad \mathbf{y} = \eta\mathbf{v}(\theta,\phi), \tag{4.1}$$

where (θ,ϕ) determine the direction of the unit vector **v**. This also holds for any matrix representation. The *D*-matrix elements are commonly written in terms of Euler angles as

$$D^l_{m,m'}(\phi,\theta,\psi)=e^{im\phi}d^l_{m,m'}(\theta)e^{im'\psi}, \tag{4.2}$$

where ϕ and ψ are rotations around the three-axis, $e^{im\phi}=\langle lm|e^{i\phi J_3}|lm\rangle$, and the *little-d*'s represent rotations around the one-axis, $d^l_{m,m'}(\theta)=\langle lm|e^{i\theta J_1}|lm'\rangle$, noting that we follow the convention of Vilenkin and Klymik[14] in rotating the second Euler angle around the one-axis; this is a form preferred for SO(N) systematics and is slightly different from that of Biedenharn and Louck in Ref. 15, Sec. 3.6, who rotate around the two-axis. The *D*-matrix elements in polar parametrization are thus bilinear in the *d*-functions,

$$\begin{aligned} D^l_{m,m'}[\eta\mathbf{v}(\theta,\phi)]&=\sum_{n=-l}^{l} D^l_{m,n}(\phi,\theta,0)e^{-in\eta}[D^l_{m',n}(\phi,\theta,0)]^* \\ &=e^{i(m'-m)\phi}\sum_{n=-l}^{l} d^l_{m,n}(\theta)e^{-in\eta}[d^l_{m',n}(\theta)]^*. \end{aligned} \tag{4.3}$$

B. W^l matrix elements on the two-sphere

Any point $\mathbf{x}=\chi\mathbf{u}(\vartheta,\varphi)\in\mathfrak{R}^3$ [cf. Eq. (2.7)] can be obtained by rotating $\chi\mathbf{k}=(0,0,\chi)$ from the North pole to $(\vartheta,\varphi)\in\mathcal{S}_2$, through

$$\mathbf{u}(\vartheta,\varphi)=\mathbf{R}(\varphi,\vartheta,0)\mathbf{k}. \tag{4.4}$$

Because of SU(2)-covariance (3.2a)–(3.9), we can write the Wigner W^l-matrix over the sphere in terms of its values at the North pole, where it is diagonal, $W^l_{m,m'}(\chi\mathbf{k})=\delta_{m,m'}W^l_m(\chi)$. Thus we have

$$W^l_{m,m'}(\chi\mathbf{u}(\vartheta,\varphi))=e^{-i(m-m')\varphi}\sum_{n=-l}^{l} d^l_{m,n}(\vartheta)W^l_n(\chi)[d^l_{m',n}(\vartheta)]^*, \tag{4.5}$$

which separates the angular dependence from the radial coordinate $\chi=|\mathbf{x}|\in\mathfrak{R}^+$, and

$$\begin{aligned} W^l_m(\chi)&=2\pi\sum_{n=-l}^{l}\int_{-1}^{1} d\,\cos\,\theta|d^l_{m,n}(\theta)|^2\int_{-2\pi}^{2\pi}\frac{1}{2}\sin^2\frac{1}{2}\,\psi d\psi\,\exp(i\psi[\chi\,\cos\,\theta-n]) \\ &=(-1)^{2l}\frac{\pi}{4}\sum_{n=-l}^{l}\int_{-1}^{1} ds|d^l_{m,n}(\arccos s)|^2 \\ &\quad\times\sin(2\pi\chi s)\left[\frac{-1}{\chi s-n+1}+\frac{2}{\chi s-n}+\frac{-1}{\chi s-n-1}\right], \end{aligned} \tag{4.6}$$

where in the last expression we have used $s:=\cos\,\theta$. Note that the integral we performed in ψ is the second difference in n of $\mathrm{sinc}\,2\pi(\chi s-n)$, and that the poles within the brackets cancel the zeros of the sine function. The eigenvalues of the self-adjoint Wigner matrix $\mathbf{W}^l(\chi\mathbf{u})$ are $\{W^l_m(\chi)\}^l_{m=-l}$. The Wigner matrix elements $W^l_{m,m'}(\mathbf{x})$ merit a deeper group-theoretical analysis; this will be done elsewhere, but it suffices to point out here that by virtue of their covariance properties, they are the *transform kernel* between functions of the continuous space coordinates $\mathbf{x}\in\mathfrak{R}^3$ and functions of the discrete space of irreducible representation indices (l,m,m'). They intertwine the angular momentum operators in their well-known differential form with their form (2.3) as difference operators.

C. Radial projection

The *radial marginal distribution* of the SU(2)-Wigner function $W^l(\mathbf{f},\mathbf{g}|\mathbf{x})$, $\mathbf{x}=\chi\mathbf{u}$, is its projection on the radius χ by integration over the spherical coordinates of $\mathbf{u}(\vartheta,\varphi)$,

$$M^l_{\text{radial}}(\mathbf{f},\mathbf{g}|\chi)=\int_{\mathcal{S}_2} d\mathbf{u}\,W^l(\mathbf{f},\mathbf{g}|\chi\mathbf{u}). \tag{4.7}$$

When we evaluate the integral (4.7) over $\mathbf{u}\in\mathcal{S}_2$, the exponential factor $\exp(i\chi\eta\mathbf{u}\cdot\mathbf{v})$ yields 4π sinc $\chi\eta$. Next, the $\mathcal{S}_2$-integral over the group parameters $\mathbf{v}(\theta,\phi)$ is computed by using

$$\int_{\mathcal{S}_2} d\mathbf{v}D^l_{m,m'}[\eta\mathbf{v}]=\frac{4\pi}{2l+1}\,\delta_{m,m'}\,\frac{\sin(l+\frac{1}{2})\eta}{\sin\frac{1}{2}\eta}. \tag{4.8}$$

Substituting, we obtain the radial projection of the Wigner function in the form

$$M^l_{\text{radial}}(\mathbf{f},\mathbf{g}|\chi)=(\mathbf{f},\mathbf{g})R^l(\chi), \tag{4.9}$$

where $(\mathbf{f},\mathbf{g})=\Sigma^l_{m=-l} f^*_m g_m$ is the inner product of the two signals (or the squared norm $|\mathbf{f}|^2$ of one signal when $\mathbf{f}=\mathbf{g}$). All radial dependence is contained in the factor

$$\begin{aligned}R^l(\chi)&=\frac{8\pi^2}{2l+1}\int_{-2\pi}^{2\pi} d\eta\,\sin\frac{1}{2}\,\eta\,\text{sinc}\,\chi\eta\,\sin\left(l+\frac{1}{2}\right)\eta\\&=\frac{4\pi^2}{(2l+1)\chi}\,[\text{Si}\,2\pi(\chi+l)+\text{Si}\,2\pi(\chi-l)-\text{Si}\,2\pi(\chi+l+1)-\text{Si}\,2\pi(\chi-l-1)],\end{aligned} \tag{4.10}$$

given in terms of the sine integral functions $\text{Si}\,x=\int_0^x \text{sinc}\,y\,dy$ (Ref. 16, Eq. 3.7632). Finally, the integral of the Wigner function over $\mathbf{x}\in\mathfrak{R}^3$ is proportional to the signal inner product (or squared norm),

$$M^l(\mathbf{f},\mathbf{g})=\int_{\mathfrak{R}^3} d\mathbf{x}W^l(\mathbf{f},\mathbf{g}|\mathbf{x})=\pi^3(\mathbf{f},\mathbf{g}). \tag{4.11}$$

In Fig. 2 we plot $R^l(\chi)$, the radial factor (4.10); taking into account (4.11) it shows that the Wigner function is concentrated in a spherical shell between radii $\chi=l$ and $l+1$. It is not difficult to show that the function $\chi R^l(\chi)$ has extremal points at every integer and half-integer $\chi\geqslant 0$, except for $\chi=l$ and $\chi=l+1$; the largest maximum of $\chi R^l(\chi)$ is at $\chi=l+\frac{1}{2}$, and $R^0(0)=2(2\pi)^3$. We emphasize that in $R^l(\chi)$ or $\chi^2R^l(\chi)$ (with the radial measure $\chi^2d\chi$), the position of this maximum shifts right or left (respectively) by small amounts but stays well within the interval $(l,l+1)$.

The number of sensors in the wave guide is fixed, so the Wigner function corresponds to a single value of l. In searching for convenient ways to plot this function of $\mathbf{x}\in\mathfrak{R}^3$, we have chosen to show its level curves on a *spherical slice* of $\mathfrak{R}^3$, for the standard value $\chi=l+\frac{1}{2}$. We have checked that similar plots at various radii near $l+\frac{1}{2}$ produce very similar figures, only with smaller maxima and curves somewhat more rounded. (Another graphing strategy for the Wigner function would be to project it by integration over χ with the radial measure $\chi^2d\chi$, or with any special or convenient measure; we leave this possibility open for future work.)

V. COHERENT STATES, ELEMENTARY STATES, CAT STATES

In this section we present the Wigner function of some signals of interest suggested by our group-theoretical treatment; they are concepts taken from the quantum theory of angular momentum. We direct our attention to coherent states and spherical harmonics.

A. Vacuum coherent state

The *vacuum* coherent state is the lowest-energy state (signal) of the system. For SU(2) and in the eigenbasis of J_3 it is the $(2l+1)$-dimensional column vector indicated

$$\boldsymbol{\alpha}(-\mathbf{k})=(0,\ldots,0,1)^{\mathsf{T}}, \tag{5.1}$$

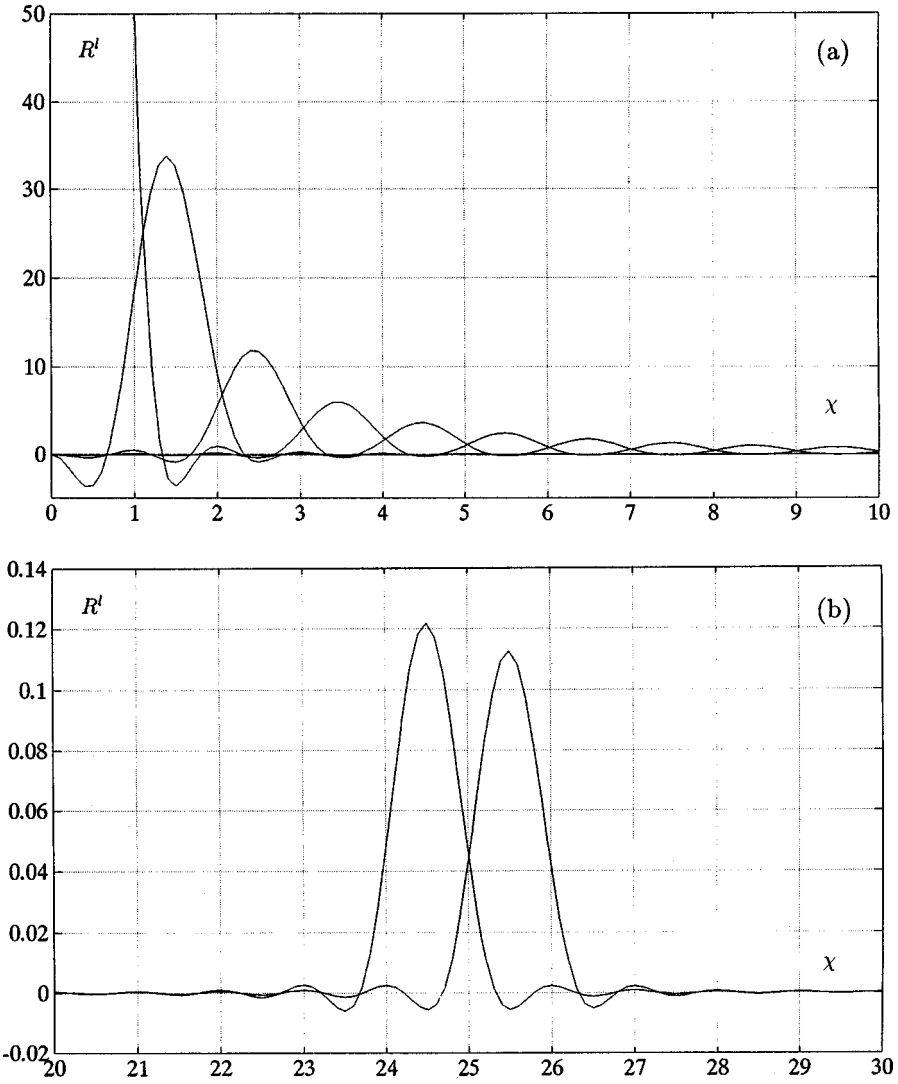

FIG. 2. Radial factor $R^l(\chi)$ in Eq. (4.10). (a) For $l=0,1,2,...,10$ in $0\leqslant\chi\leqslant 10$; the $l=0$ curve has its maximum value $2(2\pi)^3=496.10...$ at $\chi=0$. (b) Amplification of $l=24$ and $l=25$ in $20\leqslant\chi\leqslant 30$.

where T is the transpose and $-\mathbf{k}$ is a unit vector pointing to the South pole of the sphere.

In the basis where the position operator $Q=J_1$ is diagonal, the same state (5.1) represents the $(2l+1)$-point signal

$$\tilde{\boldsymbol{\alpha}}(-\mathbf{k})=\exp(-i\tfrac{1}{2}\pi J_2)\,\boldsymbol{\alpha}(-\mathbf{k})=\mathbf{D}^l(-\tfrac{1}{2}\pi,-\tfrac{1}{2}\pi,0)\,\boldsymbol{\alpha}(-\mathbf{k}), \tag{5.2a}$$

$$\tilde{\alpha}_m(-\mathbf{k})=e^{-i\pi m/2}d^l_{m,-l}(-\tfrac{1}{2}\pi)=\frac{e^{-i\pi m/2}}{2^l}\sqrt{\binom{2l}{l-m}}=e^{-i\pi m/2}\phi_0(m,2l). \tag{5.2b}$$

Here we recognize the Kravchuk function of order zero, $\phi_0(m,2l)$ in (2.5).[13] At the integers m, this function is the square root of the binomial distribution. In the limit when $l\to\infty$, $\phi_0(q,2l)$ becomes the Gaussian function of lowest oscillator state of quantum mechanics. In Fig. 3, the Wigner function (3.7) of the signal (5.2) is plotted on the sphere of radius $|\mathbf{x}|=l+\frac{1}{2}$,

$$W^l_{\text{sphere}}(\boldsymbol{\alpha}(-\mathbf{k}))|\,\vartheta,\varphi)=W^l_{-l,-l}(\chi\mathbf{u}(\vartheta,\varphi))\big|_{\chi=l+1/2}=\sum_{m=-l}^{l}W^l_m[d^l_{-l,m}(\vartheta)]^2, \tag{5.3}$$

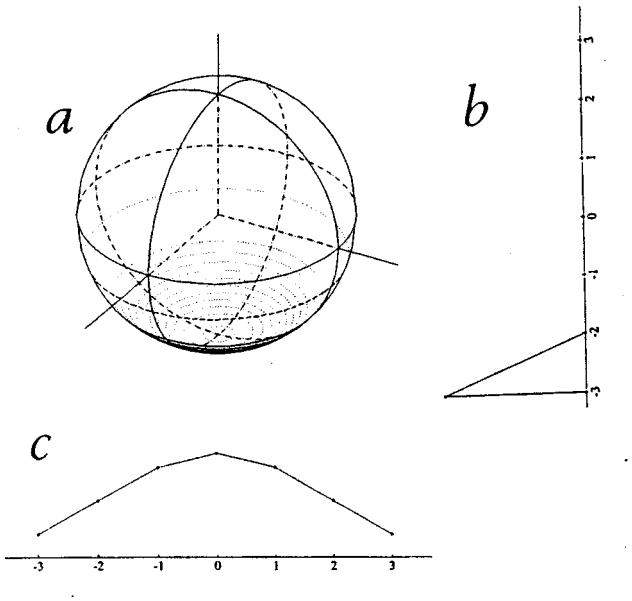

FIG. 3. Faces of the vacuum coherent state $\boldsymbol{\alpha}(-\mathbf{k})$ for $l=3$ and 7 ''sensor'' points: (a) Wigner function (5.3). Positive level contours are indicated with unbroken lines for function values 0.35, 0.7,...; zero and negative levels are indicated with broken lines. (b) Energy (mode) content. (c) The signal in the ''position'' ($Q=J_1$) basis of sensors; it is a Kravchuk function of order zero which is also the root of the binomial distribution.

where $W_m^l = W_m^l(l+\frac{1}{2})$ are constants obtained from (4.6) for the chosen value of the radius χ. Figure 3 shows the signal, its mode (energy) content, and the level curves of the Wigner function on the sphere; it shows the *partitura* of the vacuum coherent state, the ''fundamental note'' of the system. Evolution along the z-axis of the wave guide is generated by the Hamiltonian H in (2.3c), and appears as a rotation of the Wigner function around the three-axis of the sphere. Under this transformation, the vacuum state is simply multiplied by the phase $e^{-iz/2}$; hence its Wigner function is invariant.

B. SU(2) coherent states

The Wigner function of the vacuum coherent state can be rotated to any new position on the sphere (ϑ,φ) by means of SU(2)-optical transformations of the signal, i.e., by free propagation along a waveguide, with refracting wedges and tilted plates. One thus generalizes the concept of coherent states to $\boldsymbol{\alpha}(\mathbf{a})$ for any and all directions $\mathbf{a}(\vartheta,\varphi)\in\mathcal{S}_2$. If

$$\mathbf{a}=e^{i\vartheta J_1}e^{i\varphi J_3}\mathbf{k}, \quad \text{then} \quad \alpha_m(\mathbf{a}(\vartheta,\varphi))=D^l_{m,-l}(-\varphi,-\vartheta,0)=e^{im\varphi}d^l_{-l,m}(\vartheta) \tag{5.4}$$

are the components of the coherent state along **a**. In Figs. 4 we show the *antivacuum coherent* state $\boldsymbol{\alpha}(\mathbf{k})$, which is the highest energy mode in the finite planar wave guide. This is obtained from the vacuum state $\boldsymbol{\alpha}(-\mathbf{k})$ by a rotation of the sphere around the one-axis; such rotation can be physically produced with a refracting wedge that impresses a phase $e^{i\pi m}=(-1)^m$ on the input signal values; the output signal for continuous q is the Kravchuk function $\phi_{2l}(q,2l)$. For integer values m of q, this signal has the envelope of the vacuum coherent state $\phi_0(m,2l)$, but with values alternating in sign.

C. Elementary states

When the unit vector **a** is along the one-axis **i**, the coherent state $\boldsymbol{\alpha}(\mathbf{i})$ is the signal (1,0,...,0). When this signal propagates along the wave guide, the vector **a** will move around the equator; the signal will undergo the fractional Fourier–Kravchuk transformation cycle.[6] We call *elementary*

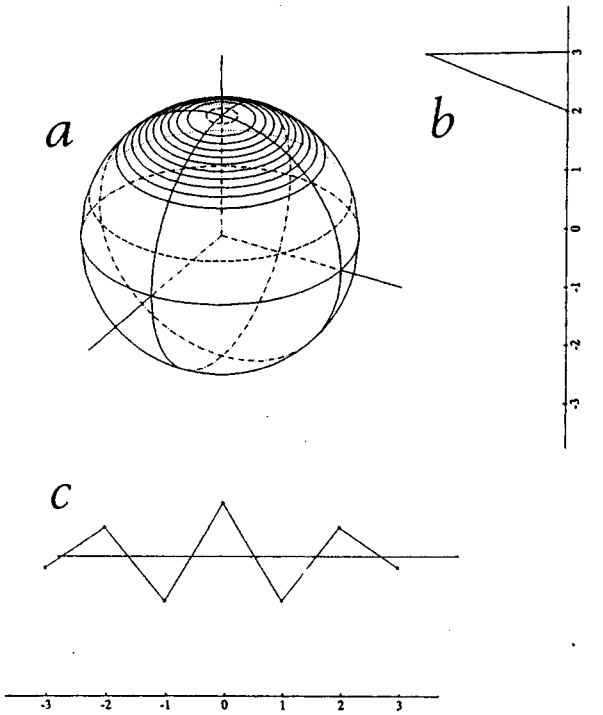

FIG. 4. Antivacuum coherent state $\boldsymbol{\alpha}(\mathbf{k})$ for $l=3$, with the parameters of the previous figure. (a) Wigner function. (b) Energy (mode) content. (c) Signal at the sensors ($Q=J_1$ basis) as in the previous figure but with alternating signs.

states (in an array of $2l+1$ sensors) to those signals with single 1 at some position ν ($l-\nu$ integer) and zeros on the rest. They are the eigenfunctions of J_1 with eigenvalues $\nu\in\{l,l-1,...,-l\}$ and will be indicated by $\boldsymbol{\alpha}^{\nu}(\mathbf{i})$.

The Wigner function (3.7) of elementary states can be then written in polar coordinates $\mathbf{u}_1(\vartheta_1,\varphi_1)$, referred to the one-axis [so $\mathbf{u}_1(0,\varphi_1)=\mathbf{i}$],

$$W^l_{\text{sphere}}(\boldsymbol{\alpha}^{\nu}(\mathbf{i})|\vartheta_1,\varphi_1)=W^l_{\nu,\nu}(\chi\mathbf{u}_1(\vartheta_1,\varphi_1))|_{\chi=l+1/2}=\sum_{m=-l}^{l} W^l_m[d^l_{\nu,m}(\vartheta_1)]^2. \tag{5.5}$$

[cf. Eq. (5.3).] The Wigner functions of elementary signals are independent of φ_1, and thus invariant under rotations of the sphere around the one-axis. Conversely, when the Wigner function of a signal $\mathbf{f}$ is invariant under rotations around some axis $\mathbf{a}$, $\mathbf{f}$ can be multiplied only by an overall phase when acted upon by $\exp(i\omega\mathbf{a}\cdot\mathbf{J})$; therefore $\mathbf{f}$ will be an eigenfunction of $\mathbf{a}\cdot\mathbf{J}$. From the plot of the Wigner function one can determine readily whether or not an output signal $\mathbf{f}=\{f_m\}^l_{m=-l}$ derives from the input of an elementary signal (0,...,0,1,0,...,0) through an SU(2)-optical system.

D. Schrödinger-cat signals and their smile function

The sum of two signals is a *Schrödinger-cat signal*. The original Schrödinger ''paradox'' of coherently superposing one dead and one live cat states applies fruitfully to signal analysis.

Consider two signals $\mathbf{f}$ and $\mathbf{g}$; the Wigner function of their linear combination $c_f\mathbf{f}+c_g\mathbf{g}$ is

$$W(c_f\mathbf{f}+c_g\mathbf{g}|\mathbf{x})=|c_f|^2W(\mathbf{f}|\mathbf{x})+|c_g|^2W(\mathbf{g}|\mathbf{x})+2\,\text{Re}\,c_f^*c_gW(\mathbf{f},\mathbf{g}|\mathbf{x}). \tag{5.6}$$

The Wigner functions of the two signals thus not only add, but *interfere*. The cross term $S(\mathbf{f},\mathbf{g}|\mathbf{x})=2\,\text{Re}\,\mathbf{f}^{\dagger}\mathbf{W}(\mathbf{x})\mathbf{g}$ has been called the *smile* function of the Schrödinger-cat state; it is a most prominent feature of the Wigner function of cat states that carries the holographic information of the object signal $\mathbf{g}$ with the reference signal $\mathbf{f}$.[17]

In Fig. 5 we plot Schrödinger cat states composed of two normal modes of the wave guide, for $l=3$ and 7 sensors. If we place the one-axis along the three-axes of the figures, the Schrödinger

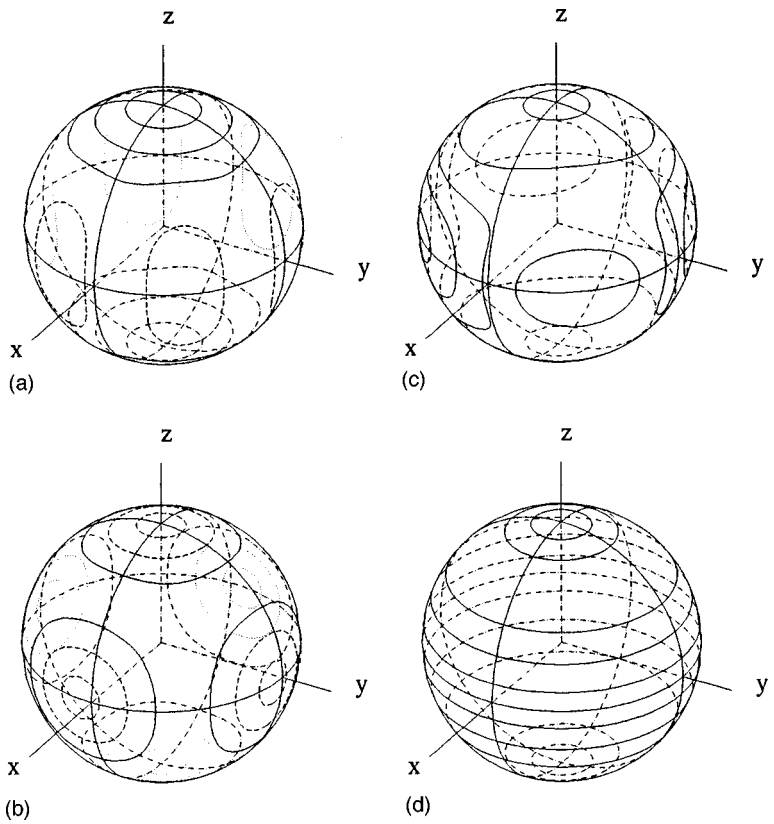

FIG. 5. Schrödinger-cat states of two normal modes or elementary seven-point signals ($l=3$). (a) ($1/\sqrt{2}$, 0, 0, 0, 0, 0, $1/\sqrt{2}$), (b) (0, $1/\sqrt{2}$, 0, 0, 0, $1/\sqrt{2}$, 0), (c) (0, 0, $1/\sqrt{2}$, 0, $1/\sqrt{2}$, 0, 0), and (d) (0,0,0,1,0,0,0).

cats are ''bilocalized'' states and the last is an elementary state. Compare Fig. 5(a) with Figs. 3 and 4, noticing the six oscillations (*teeth*) of the smile across the equator; Fig. 5(b) exhibits four, Fig. 5(c) shows two, and Fig. 5(d) has axial symmetry.

VI. CONCLUDING REMARKS

We represented finite data sets by their Wigner function in an optical model of planar multimodal wave guides. Position, momentum, and energy (mode number) are *c*-number coordinates which are associated to the generators of su(2) algebra; the latter are self-adjoint difference operators with discrete, equally-spaced, and finite spectra. The *partitura* of a signal is drawn on a sphere. The vacuum coherent state is analogous to the ground harmonic oscillator wave function; from it, all other coherent states can be produced by SU(2)-optical elements, wave guide, wedge, and slab. These transformations are the counterpart of rotation and translation of classical phase space.

The applicability of our construction extends to other mechanical and optical systems, also based on finite data sets but with different dynamics. For instance, the quantum *point rotor* (a mass point constrained to move on $\mathcal{S}_2$) has symmetry group SO(3), which is covered twice by SU(2). This system is isomorphic to the mechanical vibrating sphere and, by stereographic projection, to the two-dimensional Maxwell fisheye wave-optical medium.[18] In both of these cases, wave functions are on the sphere and generally contain *all* spherical harmonics,

$$\Psi(\beta,\gamma)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\Psi_m^l Y_m^l(\beta,\gamma). \tag{6.1}$$

The data set is now infinite and given by the tower of coefficients $\mathbf{\Psi}=\{\mathbf{\Psi}^l\}_{l=0}^{\infty}$, $\mathbf{\Psi}^l=\{\Psi_m^l\}_{m=-l}^{l}$.

The SU(2) Wigner function (3.7) was built for a single value of l corresponding to a fixed number of sensors of the model, and the radial factor (4.10) indicated that it is concentrated in a spherical shell of mean radius $l+\frac{1}{2}$. If now the data vector is (6.1), the Wigner operator (3.1) will act irreducibly on each subvector $\mathbf{\Psi}^l$, and the Wigner function of $\mathbf{\Psi}$ will be the sum of the Wigner functions in the irreducible subspaces, i.e.,

$$W(\mathbf{\Psi}|\mathbf{x})=\sum_{l=0}^{\infty} W^l(\mathbf{\Psi}^l|\mathbf{x})=\sum_{l=0}^{\infty} \mathbf{\Psi}^{l\dagger}\mathbf{W}^l(\mathbf{x})\mathbf{\Psi}^l. \tag{6.2}$$

Such a function must be plotted in $\mathfrak{R}^3$-space $\mathbf{x}$ and will show the quasiprobability distribution of the angular momentum vector $\mathbf{J}$ in the mechanical model, and of the beam parameters in the Maxwell fisheye. As drawn in elementary texts, the extreme spherical harmonic (coherent state) $Y_l^l(\beta,\gamma)$ has a minimal uncertainty cone around the three-axis, $Y_{l-1}^l(\beta,\gamma)$ has a larger cone, etc., until $Y_0^l(\beta,\gamma)$ opens the cone into the equator. Thus, while the wave function on the sphere $Y_l^l(\beta,\gamma)$ have their square maximum on the equator $\beta=\frac{1}{2}\pi$ (where the classical particle is), the Wigner function is concentrated in the North polar cap around $\mathbf{x}=(l+\frac{1}{2})\mathbf{k}$ (where the angular momentum is). On the other hand, strongly localized rotor wave functions, containing a wide spread of l's, will have Wigner functions (6.2) widely spread in $|\mathbf{x}|$.

The dynamics of the rotor models (vibrating sphere and Maxwell fisheye) are different from the wave guide, however; their Hamiltonian is proportional to the Casimir operator of so(3)=su(2), rather than to one of its generators. Since under the corresponding ''time'' evolution τ, each component $\mathbf{\Psi}^l$ is multiplied by the phase $\exp[i\tau l(l+1)]$, it follows that the Wigner function (6.2) is invariant under rotor dynamics.

Nonlinear processes on discrete, finite systems, can also be analyzed with the SU(2) Wigner function (3.7) when the Hamiltonian belongs to the su(2) *enveloping* algebra. This is the case of the finite analog of the optical Kerr medium studied in Ref. 19, whose Hamiltonian is $\sim J_3+\alpha J_3^2$. It also applies for Hamiltonians contrieved to have the form $\frac{1}{2}J_2^2+V(J_1)$, to mimic the classical form of mechanics $\frac{1}{2}p^2+V(q)$, but restricted to a finite number of bound states. Natural applications for SU(2)-Wigner functions will also include spin and pseudo-spin systems, as they appear, e.g., in quantum optics, when a collection of two-level atoms interact with the radiation field under small-volume approximation, as in the Dicke model.[20,21] Time evolution in these cases deforms the shape of the initial coherent state on the sphere; in the Kerr medium, the SU(2) Wigner function gives a transparent picture of the nature of fractional-period resonances.[19] Covariance does not hold beyond SU(2), so we can expect purely quantum effects (including squeezing) to show up in the Wigner function on the sphere.

ACKNOWLEDGMENTS

We are very grateful to the referee for carefully reading the original manuscript and his critical remarks which have helped us greatly to improve the presentation of our results. We thank Guillermo Krötzsch for his kind help with the figures. This work was performed under the support of Project DGAPA-UNAM 106595 *Optica Matemática.*

[1] E. Wigner, ''On the quantum correction for thermodynamic equilibrium,'' Phys. Rev. **40**, 749–759 (1932); M. Hillery, R. F. O'Connell, M. O. Scully, and E. P. Wigner, ''Distribution functions in physics: Fundamentals,'' Phys. Rep. **106**, 121–167 (1984); H.-W. Lee, ''Theory and application of the quantum phase-space distribution functions,'' *ibid.* **259**, 147–211 (1995).

[2] A. Lohmann, ''The Wigner function and its optical production,'' Opt. Commun. **42**, 32–37 (1980).

[3] H. O. Bartelt, K.-H. Brenner, and H. Lohmann, ''The Wigner distribution function and its optical production,'' Opt. Commun. **32**, 32–38 (1980); H. Bartelt and K.-H. Brenner, ''The Wigner distribution function: An alternate signal representation in optics,'' Isr. J. Tech. **18**, 260–262 (1980); K.-H. Brenner and H. Lohmann, ''Wigner distribution function display of complex 1D signals,'' Opt. Commun. **42**, 310–314 (1982).

[4] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge University Press, Cambridge, 1990); A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of a Discrete Variable* (Springer, Berlin, 1991).

[5] K. B. Wolf, *Integral Transforms in Science and Engineering* (Plenum, New York, 1979), Chap. 1–3.

[6] N. M. Atakishiyev and K. B. Wolf, ''Fractional Fourier–Kravchuk transform,'' J. Opt. Soc. Am. A **14**, 1467–1477 (1997).

[7] K. B. Wolf, ''Wigner distribution function for paraxial polychromatic optics,'' Opt. Commun. **132**, 343–352 (1996).

[8] J. W. Goodman, *Introduction to Fourier Optics* (McGraw–Hill, New York, 1968).

[9] J. R. Klauder and E. C. G. Sudarshan, *Fundamentals of Quantum Optics* (Benjamin, Reading, 1968).

[10] L. M. Nieto, N. M. Atakishiyev, S. M. Chumakov, and K. B. Wolf, ''Wigner distribution function for Euclidean systems,'' J. Phys. A **31**, 3875–3895 (1998).
[11] N. M. Atakishiyev and S. K. Suslov, ''Difference analogs of the harmonic oscillator,'' Theor. Math. Phys. **85**, 1055–1062 (1991).
[12] M. Krawtchouk, ''Sur une généralization des polinômes d'Hermite,'' C. R. Acad. Sci., Paris **189**, 620–622 (1929); A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. 2.
[13] N. M. Atakishiyev and K. B. Wolf, ''Approximation on a finite set of points through Kravchuk functions,'' Rev. Mex. Fis. **40**, 366–377 (1994).
[14] N. Ja. Vilenkin and A. U. Klimyk, *Representation of Lie Group and Special Functions* (Kluwer Academic, Dordrecht, 1991), Vol. 1.
[15] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics*, in the Encyclopedia of Mathematics and its Applications Series, edited by G.-C. Rota (Addison–Wesley, Reading, 1981).
[16] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed. (Academic, New York, 1965).
[17] K. B. Wolf and A. L. Rivera, ''Holographic information in the Wigner function,'' Opt. Commun. **144**, 36–42 (1997).
[18] A. Frank, F. Leyvraz, and K. B. Wolf, ''Hidden symmetry and potential group of the Maxwell fish-eye,'' J. Math. Phys. **31**, 2757–2768 (1990).
[19] S. M. Chumakov, A. Frank, and K. B. Wolf, ''Finite Kerr medium: Schrödinger cats and Wigner functions on the sphere'' (preprint No. 65, IIMAS-UNAM, October 1997) (submitted).
[20] R. H. Dicke, ''Coherence in spontaneous radiation processes,'' Phys. Rev. **93**, 99–110 (1954); D. J. Wineland, J. J. Bollinger, W. M. Itano, and D. J. Heinzen, ''Squeezed atomic states and projection noise in spectroscopy,'' Phys. Rev. A **50**, 67–88 (1994).
[21] M. Kozierowski, A. A. Mamedov, and S. M. Chumakov, ''Spontaneous emission by a system of *N* two-level atoms in terms of the SU(2)-group representations,'' Phys. Rev. A **42**, 1762–1766 (1990); S. M. Chumakov and M. Kozierowski, ''Collective emission in a lossless cavity: Analytical approach,'' Quantum Semiclassic. Opt. **8**, 775–803 (1996).

Resolution of some mathematical problems arising in the relativistic treatment of the *S* states of three-electron systems

D. Matthew Feldmann, Paul J. Pelzl, and Frederick W. King
Department of Chemistry, University of Wisconsin—Eau Claire, Eau Claire, Wisconsin 54702

(Received 16 April 1998; accepted for publication 27 August 1998)

Some of the mathematical difficulties that arise in the evaluation of the Breit–Pauli relativistic energy corrections for the *S* states of three-electron systems are resolved. Evaluation of the expectation value of the Breit–Pauli Hamiltonian using explicitly correlated wave functions leads to sets of integrals that diverge individually. By appropriately combining these integrals, and using some judicious series expansions, all the integration problems are resolved in terms of well-known auxiliary functions. © *1998 American Institute of Physics.* [S0022-2488(98)02512-2]

I. INTRODUCTION

There has been considerable progress over the past ten years on the calculation of high-precision properties of the lithium atom and other members of the Li isoelectronic series (for a recent summary of this progress see Refs. 1 and 2). A large part of this effort has been achieved using Hylleraas (or configuration interaction (CI)-Hylleraas) type wave functions.

One important property that has received limited theoretical attention is the high-precision calculation of the relativistic correction to the energy, ΔE_{REL}. This is a key contribution in determining a precise theoretical estimate of the first ionization potential. At present, the only calculations of the relativistic correction to the ground state energy for the lithium atom are due to Chung,[3] who used the CI technique, the very recent paper of Yan and Drake using the Hylleraas technique,[4] and the work of King *et al.*,[5] who also employed the Hylleraas method. Chung's calculation required a significant core correction, determined by first computing the relativistic correction for Li^+, and then comparing with precise calculations for Li^+ using Hylleraas-type wave functions.[6,7] This core correction was then applied directly to the Li calculation. A very effective approach to obtaining high-precision estimates of ΔE_{REL} is to make use of large-scale Hylleraas-type expansions. Progress in this area has been hampered by the mathematical difficulties which arise for the operators

$$H_{\mathrm{mass}}=-\frac{\alpha^2}{8}\sum_{i=1}^{3}\nabla_i^4, \tag{1a}$$

$$H_{\mathrm{oo}}=\frac{\alpha^2}{2}\sum_{i=1}^{3}\sum_{j>i}^{3}\frac{1}{r_{ij}}\left(\nabla_i\cdot\nabla_j+\frac{\mathbf{r}_{ij}\cdot(\mathbf{r}_{ij}\cdot\nabla_i)\nabla_j}{r_{ij}^2}\right), \tag{1b}$$

where α is the fine structure constant and r_{ij} is the interelectronic separation. For the three-electron problem, evaluation of the matrix elements of H_{mass} or H_{oo} [in the form displayed in Eqs. (1a) and (1b)] using a general Hylleraas expansion can be shown to involve the following integrals:

$$I(i,j,k,l,m,n,\alpha,\beta,\gamma)\equiv\int r_1^i r_2^j r_3^k r_{23}^l r_{31}^m r_{12}^n e^{-\alpha r_1-\beta r_2-\gamma r_3}\,d\mathbf{r}_1\,d\mathbf{r}_2\,d\mathbf{r}_3, \tag{2}$$

where $i,j,k\geqslant -2$, $l\geqslant -2$, $m\geqslant -2$, and $n\geqslant -1$;

 © 1998 American Institute of Physics

$$I_1(i,j,k,l,m,\alpha,\beta,\gamma) \equiv \int r_1^i r_2^j r_3^k \left(\frac{r_1^2 - r_2^2}{r_{12}^3} \right) r_{23}^l r_{31}^m e^{-\alpha r_1 - \beta r_2 - \gamma r_3} \, d\mathbf{r}_1 \, d\mathbf{r}_2 \, d\mathbf{r}_3, \tag{3}$$

$$I_2(i,j,k,l,m,\alpha,\beta,\gamma) \equiv \int r_1^i r_2^j r_3^k \left(\frac{r_{31}^2 - r_{23}^2}{r_{12}^3} \right) r_{23}^l r_{31}^m e^{-\alpha r_1 - \beta r_2 - \gamma r_3} \, d\mathbf{r}_1 \, d\mathbf{r}_2 \, d\mathbf{r}_3, \tag{4}$$

and

$$I_3(i,j,k,l,m,\alpha,\beta,\gamma) \equiv \int r_1^i r_2^j r_3^k \left(\frac{r_1^2 - r_2^2}{r_{12}^2} \right) r_{23}^l r_{31}^m e^{-\alpha r_1 - \beta r_2 - \gamma r_3} \, d\mathbf{r}_1 \, d\mathbf{r}_2 \, d\mathbf{r}_3, \tag{5}$$

where $i,j,k \geqslant -2$, $l \geqslant -1$, and $m \geqslant 1$. The combination of terms displayed in Eqs. (3) and (4) is essential, if a finite value of the expectation value of the Breit–Pauli Hamiltonian is to be obtained using a first-order perturbation theory approach. For Eq. (2) and some related generalizations, several evaluation approaches exist in the literature for the case where l, m, and n are $\geqslant -1$,[8–18] and for the more difficult cases involving $l = -2$ (and $m = -2$).[19–24] Some of the published techniques could be applied to evaluate I_3.

The purpose of this paper is to present an efficient method for the evaluation of the integrals I_1, I_2, and I_3. The integrals are evaluated for the cases $i,j,k \geqslant -2$, $l \geqslant -1$, and $m \geqslant -1$. The obvious reduction of the I_1 and I_2 integrals into a difference of I integrals is not possible, because the separate I integrals both diverge. This is discussed further in Appendix A. While I_3 could be calculated by taking a difference of I integrals, a computationally superior approach is discussed in Sec. IV. The solution method is to reduce each of the integrals to a series of well-known auxiliary functions.

II. EVALUATION OF I_1

To evaluate I_1, the following expansions are employed:[25–27]

$$r_{23}^l = \sum_{w_2=0}^{\infty} R_{l w_2}(r_2, r_3) P_{w_2}(\cos \theta_{23}), \tag{6}$$

$$r_{31}^m = \sum_{w_3=0}^{\infty} R_{m w_3}(r_1, r_3) P_{w_3}(\cos \theta_{31}), \tag{7}$$

and

$$\frac{1}{r_{12}^3} = \frac{1}{r_{12>}^2 - r_{12<}^2} \sum_{w_1=0}^{\infty} (2w_1 + 1) \frac{r_{12<}^{w_1}}{r_{12>}^{w_1+1}} P_{w_1}(\cos \theta_{12}). \tag{8}$$

Equations (6) and (7) are the Sack expansions for the interelectron coordinates.[25] Equation (8) is discussed in Appendix B. Here $r_{12<}$ denotes the lesser of (r_1, r_2), $r_{12>}$ represents the greater of (r_1, r_2), and $P_n(x)$ designates a Legendre polynomial throughout this work.

Substituting Eqs. (6), (7), and (8) into Eq. (3) gives

$$I_1 = \sum_{w_1=0}^{\infty} \sum_{w_2=0}^{\infty} \sum_{w_3=0}^{\infty} (2w_1 + 1) I_R(w_1, w_2, w_3) I_\Omega(w_1, w_2, w_3), \tag{9}$$

where I_R is the integral over the radial coordinates,

$$I_R(w_1, w_2, w_3) = \int r_1^{i+2} r_2^{j+2} r_3^{k+2} r_{12<}^{w_1} r_{12>}^{-w_1-1} \left(\frac{r_1^2 - r_2^2}{r_{12>}^2 - r_{12<}^2} \right) R_{l w_2}(r_2, r_3) \times R_{m w_3}(r_1, r_3) e^{-\alpha r_1 - \beta r_2 - \gamma r_3} \, dr_1 \, dr_2 \, dr_3, \tag{10}$$

and I_Ω is the integral over the angular variables,

$$I_\Omega(w_1,w_2,w_3)=\int P_{w_1}(\cos\,\theta_{12})P_{w_2}(\cos\,\theta_{23})P_{w_3}(\cos\,\theta_{31})d\Omega_1\,d\Omega_2\,d\Omega_3. \tag{11}$$

The functional dependence of I_R on $\{i,j,k,l,m,\alpha,\beta,\gamma\}$ is suppressed to simplify the notation. Expanding the Legendre polynomials in terms of spherical harmonics, the angular integral simplifies to

$$I_\Omega(w_1,w_2,w_3)=\frac{64\pi^3}{(2w_1+1)^2}\,\delta_{w_1w_2}\delta_{w_2w_3}, \tag{12}$$

where δ_{ij} denotes a Kronecker delta. Equation (9) now becomes

$$I_1=64\pi^3\sum_{w_1=0}^{\infty}\frac{I_R(w_1)}{(2w_1+1)}, \tag{13}$$

where $I_R(w_1)\equiv I_R(w_1,w_1,w_1)$.

To evaluate the radial integral the formulas for the Sack radial functions are employed:[25]

$$R_{lw_1}(r_2,r_3)=\sum_{p=0}^{\infty}a_{w_1lp}r_{23<}^{w_1+2p}r_{23>}^{l-w_1-2p} \tag{14}$$

and

$$R_{mw_1}(r_1,r_3)=\sum_{q=0}^{\infty}a_{w_1mq}r_{31<}^{w_1+2q}r_{31>}^{m-w_1-2q}, \tag{15}$$

where

$$a_{tuv}=\frac{(-u/2)_t(t-u/2)_v(-1/2-u/2)_v}{(1/2)_t v!(t+3/2)_v} \tag{16}$$

and $(z)_n$ denotes a Pochhammer symbol. Inserting Eqs. (14) and (15) into Eq. (10) gives

$$\begin{aligned}I_R(w_1)=\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}a_{w_1lp}a_{w_1mq}\int r_1^{i+2}r_2^{j+2}r_3^{k+2}r_{12<}^{w_1}r_{12>}^{-w_1-1}\left(\frac{r_1^2-r_2^2}{r_{12>}^2-r_{12<}^2}\right)\\ \times r_{23<}^{w_1+2p}r_{23>}^{l-w_1-2p}r_{31<}^{w_1+2q}r_{31>}^{m-w_1-2q}e^{-\alpha r_1-\beta r_2-\gamma r_3}\,dr_1\,dr_2\,dr_3.\end{aligned} \tag{17}$$

If the integration in Eq. (17) is broken up into the six regions $r_i\leqslant r_j\leqslant r_k$, then I_1 can be written as

$$\begin{aligned}I_1(i,j,k,l,m,\alpha,\beta,\gamma)=64\pi^3\sum_{w=0}^{\infty}\frac{1}{(2w+1)}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}a_{wlp}a_{wmq}\\ \times\{W(k+2+2p+2q+2w,j+2+l-2p,i+m+1-2q-2w,\gamma,\beta,\alpha)\\ +W(j+2+2w+2p,k+2+2q+l-2p,i+m+1-2q-2w,\beta,\gamma,\alpha)\\ +W(j+2+2w+2p,i+1+2q,k+m+l+2-2w-2p-2q,\beta,\alpha,\gamma)\\ -W(i+2+2w+2q,j+1+2p,k+m+l+2-2w-2p-2q,\alpha,\beta,\gamma)\\ -W(i+2+2w+2q,k+2+m-2q+2p,j+1+l-2w-2p,\alpha,\gamma,\beta)\\ -W(k+2+2p+2q+2w,i+2+m-2q,j+1+l-2w-2p,\gamma,\alpha,\beta)\},\end{aligned} \tag{18}$$

where

$$W(L,M,N,a,b,c)=\int_0^\infty x^L e^{-ax}\,dx\int_x^\infty y^M e^{-by}\,dy\int_y^\infty z^N e^{-cz}\,dz. \tag{19}$$

The W integrals have been well studied in the literature,[8,9,13,16,28–30] and efficient algorithms exist for the evaluation of these auxiliary functions.

We now examine the summation limits of the w, p, and q sums of Eq. (18). Because of the following property of the Pochhammer symbol,

$$(-k)_l=0, \quad l>k \text{ for positive integer } k, \tag{20}$$

the p summation terminates at $(l+1)/2$ for odd l, and $(l/2)-w$ for even l. Similarly, the q summation terminates at $(m+1)/2$ for odd m, and $(m/2)-w$ for even m. These conditions follow from the definition of a_{tuv} given in Eq. (16). This leads to the following termination conditions for the w sum:

$$\frac{m}{2}, \quad m \text{ even and } n \text{ odd},$$

$$\frac{n}{2}, \quad n \text{ even and } m \text{ odd},$$

$$\min\left\{\frac{m}{2},\frac{n}{2}\right\}, \quad m \text{ and } n \text{ both even}.$$

For m and n both odd, the w sum is nonterminating.

In addition to the individual constraints on i, j, k, l, and m mentioned in the Introduction, it is also necessary that

$$i+j+k+l+m\geqslant -7. \tag{21}$$

This follows directly from a known constraint on the arguments of the W integrals, namely

$$L+M+N\geqslant -2. \tag{22}$$

Yan and Drake[23] give a reduction formula [see their Eq. (128)] for a radial integral which is related to I_1. This radial integral is part of a nested sum, which for the general case would involve a double infinite sum.

III. EVALUATION OF I_2

For this integral, the coordinate system of choice is a three-dimensional analog of the one used by Hylleraas,[31] where the directions of the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ are given relative to $\mathbf{r}_3$. This method has been previously employed to consider other three-electron integration problems.[10] In this coordinate system the volume element becomes

$$d\mathbf{r}_1\,d\mathbf{r}_2\,d\mathbf{r}_3=r_1^2r_2^2r_3^2\,dr_1\,dr_2\,dr_3\sin\theta_3\,d\theta_3\,d\phi_3\sin\theta_{23}\,d\theta_{23}\,d\phi_{23}\sin\theta_{31}\,d\theta_{31}\,d\phi_{31}. \tag{23}$$

If $r_{31}^2-r_{23}^2$ is written as

$$r_{31}^2-r_{23}^2=r_1^2-r_2^2-2r_1r_3\cos\theta_{31}+2r_2r_3\cos\theta_{23}, \tag{24}$$

then I_2 may be expressed as

$$I_2(i,j,k,l,m,\alpha,\beta,\gamma)=I_1(i,j,k,l,m,\alpha,\beta,\gamma)-2J_1(i,j,k,l,m,\alpha,\beta,\gamma), \tag{25}$$

where

$$J_1(i,j,k,l,m,\alpha,\beta,\gamma)=\int r_1^i r_2^j r_3^{k+1}\left(\frac{r_1\cos\theta_{31}-r_2\cos\theta_{23}}{r_{12}^3}\right) r_{23}^l r_{31}^m e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3. \tag{26}$$

J_1 is simplified by substituting Eqs. (6), (7), (8), (14), and (15) into Eq. (26) and integrating over θ_3 and ϕ_3 to obtain

$$J_1=4\pi\sum_{w_1=0}^{\infty}(2w_1+1)\sum_{w_2=0}^{\infty}\sum_{w_3=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}a_{w_2lp}a_{w_3mq}\int r_1^{i+2}r_2^{j+2}r_3^{k+3}r_{23<}^{w_2+2p}r_{23>}^{l-w_2-2p}$$
$$\times r_{31<}^{w_3+2q}r_{31>}^{m-w_3-2q}r_{12<}^{w_1}r_{12>}^{-w_1-1}\left(\frac{r_1\Phi_1-r_2\Phi_2}{r_{12>}^2-r_{12<}^2}\right)e^{-\alpha r_1-\beta r_2-\gamma r_3}\, dr_1\, dr_2\, dr_3, \tag{27}$$

where

$$\Phi_1=\int_0^{2\pi}d\phi_{23}\int_0^{2\pi}d\phi_{31}\int_0^{\pi}d\theta_{31}\sin\theta_{31}\int_0^{\pi}d\theta_{23}\sin\theta_{23}\cos\theta_{31}$$
$$\times P_{w_1}(\cos\theta_{12})P_{w_2}(\cos\theta_{23})P_{w_3}(\cos\theta_{31}) \tag{28}$$

and

$$\Phi_2=\int_0^{2\pi}d\phi_{23}\int_0^{2\pi}d\phi_{31}\int_0^{\pi}d\theta_{31}\sin\theta_{31}\int_0^{\pi}d\theta_{23}\sin\theta_{23}\cos\theta_{23}$$
$$\times P_{w_1}(\cos\theta_{12})P_{w_2}(\cos\theta_{23})P_{w_3}(\cos\theta_{31}). \tag{29}$$

Evaluation of the integrals Φ_1 and Φ_2 leads to

$$\Phi_1=\frac{16\pi^2\delta_{w_1w_2}}{2w_1+1}\left\{\frac{w_2}{4w_2^2-1}\,\delta_{w_2,w_3+1}+\frac{w_3}{4w_3^2-1}\,\delta_{w_3,w_2+1}\right\} \tag{30}$$

and

$$\Phi_2=\frac{16\pi^2\delta_{w_1w_3}}{2w_1+1}\left\{\frac{w_2}{4w_2^2-1}\,\delta_{w_2,w_3+1}+\frac{w_3}{4w_3^2-1}\,\delta_{w_3,w_2+1}\right\}. \tag{31}$$

The solution of these integrals is discussed in Appendix C. Substituting Eqs. (30) and (31) into Eq. (27) leads to

$$J_1=64\pi^3\sum_{w_2=0}^{\infty}\sum_{w_3=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}a_{w_2lp}a_{w_3mq}\int r_1^{i+2}r_2^{j+2}r_3^{k+3}r_{23<}^{w_2+2p}r_{23>}^{l-w_2-2p}r_{31<}^{w_3+2q}$$
$$\times r_{31>}^{m-w_3-2q}e^{-\alpha r_1-\beta r_2-\gamma r_3}\left(\frac{r_1r_{12<}^{w_2}r_{12>}^{-w_2-1}-r_2r_{12<}^{w_3}r_{12>}^{-w_3-1}}{r_{12>}^2-r_{12<}^2}\right)$$
$$\times\left[\frac{w_2}{4w_2^2-1}\,\delta_{w_2,w_3+1}+\frac{w_3}{4w_3^2-1}\,\delta_{w_3,w_2+1}\right]dr_1\, dr_2\, dr_3. \tag{32}$$

Using the fact that

$$a_{(t+1)uv}=\left(\frac{2t+3}{2t+1}\right)\left(\frac{2t+2v-u}{2t+2v+3}\right)a_{tuv} \tag{33}$$

allows I_2 to be simplified to

$$
\begin{aligned}
I_2(i,j,k,l,m,\alpha,\beta,\gamma) &= I_1(i,j,k,l,m,\alpha,\beta,\gamma) \\
&+ 128\pi^3 \sum_{w=0}^{\infty} \frac{(w+1)}{(2w+1)^2} \sum_{q=0}^{\infty} \sum_{p=0}^{\infty} a_{wmq} a_{wlp} \left[\left(\frac{2w+2p-l}{2w+2p+3} \right) \right. \\
&\times \{ W(i+2+2w+2q, j+1+2p, k+m+l+2-2w-2p-2q, \alpha, \beta, \gamma) \\
&+ W(i+2+2w+2q, k+m+4+2p-2q, j+l-1-2w-2p, \alpha, \gamma, \beta) \\
&+ W(k+4+2w+2p+2q, i+m+2-2q, j+l-1-2w-2p, \gamma, \alpha, \beta) \} \\
&- \left(\frac{2w+2q-m}{2w+2q+3} \right) \{ W(j+2+2w+2p, i+1+2q, k+m \\
&+ l+2-2w-2p-2q, \beta, \alpha, \gamma) \\
&+ W(j+2+2w+2p, k+4+l-2p+2q, i+m-1-2w-2q, \beta, \gamma, \alpha) \\
&\left. + W(k+4+2w+2p+2q, j+2+l-2p, i+m-1-2w-2q, \gamma, \beta, \alpha) \} \right].
\end{aligned} \tag{34}
$$

Equation (34) represents the solution to I_2. The summation limits of the w, p, and q sums are the same as those for the I_1 integral. The constraint given in Eq. (21) also applies for Eq. (34).

VI. EVALUATION OF I_3

The generating function for the Chebyshev polynomials of the first kind is

$$
\frac{1-r^2}{1-2rx+r^2} = 1 + 2 \sum_{w_1=1}^{\infty} r^{w_1} T_{w_1}(x). \tag{35}
$$

Letting $r = r_{12<}/r_{12>}$ and $x = \cos\theta_{12}$ in Eq. (35) yields

$$
\frac{1}{r_{12}^2} = \frac{1}{r_{12>}^2 - r_{12<}^2} \left[1 + 2 \sum_{w_1=1}^{\infty} \left(\frac{r_{12<}}{r_{12>}} \right)^{w_1} T_{w_1}(\cos\theta_{12}) \right]. \tag{36}
$$

Substituting Eqs. (6), (7), and (36) into Eq. (5) leads to the result

$$
I_3 = F(i,j,k,l,m,\alpha,\beta,\gamma) + 2 \sum_{w_1=1}^{\infty} \sum_{w_2=0}^{\infty} \sum_{w_3=0}^{\infty} I_\Omega(w_1,w_2,w_3) I_R(w_1,w_2,w_3), \tag{37}
$$

where

$$
\begin{aligned}
F(i,j,k,l,m,\alpha,\beta,\gamma) = \sum_{w_2=0}^{\infty} \sum_{w_3=0}^{\infty} \int r_1^i r_2^j r_3^k \left(\frac{r_1^2 - r_2^2}{r_{12>}^2 - r_{12<}^2} \right) R_{lw_2}(r_2,r_3) R_{mw_3}(r_3,r_1) \\
\times P_{w_2}(\cos\theta_{23}) P_{w_3}(\cos\theta_{31}) d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3,
\end{aligned} \tag{38}
$$

$$
I_\Omega = \int T_{w_1}(\cos\theta_{12}) P_{w_2}(\cos\theta_{23}) P_{w_3}(\cos\theta_{31}) d\Omega_1\, d\Omega_2\, d\Omega_3 \tag{39}
$$

and

$$
I_R = \int r_1^{i+2} r_2^{j+2} r_3^{k+2} r_{12<}^{w_1} r_{12>}^{-w_1} \left(\frac{r_1^2 - r_2^2}{r_{12>}^2 - r_{12<}^2} \right) R_{lw_2}(r_2,r_3) R_{mw_3}(r_3,r_1) dr_1\, dr_2\, dr_3. \tag{40}
$$

Employing Eqs. (14) and (15), and expanding the Legendre polynomials in terms of spherical harmonics, Eq. (38) may be easily evaluated to yield

$$
\begin{aligned}
F(i,j,k,l,m,\alpha,\beta,\gamma) = 64\pi^3 \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} a_{0lp} a_{0mq} \\
&\times \{ W(k+2+2p+2q, j+2+l-2p, i+m+2-2q, \gamma, \beta, \alpha) \\
&+ W(j+2+2p, k+2+l-2p+2q, i+m+2-2q, \beta, \gamma, \alpha) \\
&+ W(j+2+2p, i+2+2q, k+m+l+2-2p-2q, \beta, \alpha, \gamma) \\
&- W(i+2+2q, j+2+2p, k+m+l+2-2p-2q, \alpha, \beta, \gamma) \\
&- W(i+2+2q, k+m+2+2p-2q, j+l+2-2p, \alpha, \gamma, \beta) \\
&- W(k+2+2p+2q, i+m+2-2q, j+l+2-2p, \gamma, \alpha, \beta) \}. \qquad (41)
\end{aligned}
$$

Due to the property of the Pochhammer symbol given in Eq. (20), the p summation in Eq. (41) will terminate at $l/2$ for even l, and $(l+1)/2$ for odd l. Similarly, the q summation will terminate at $m/2$ for even m, and $(m+1)/2$ for odd m.

The integral I_Ω can be evaluated to give

$$
I_\Omega = \frac{32\pi^3}{2w_2+1}\, \delta_{w_2 w_3} \times \begin{cases} \dfrac{\sqrt{\pi}\,\Gamma(w_1+1)}{2\Gamma(w_1+\frac{3}{2})}, & w_1 = w_2, \\ \dfrac{-w_1}{(w_1-w_2)(w_1+w_2+1)} \dfrac{\Gamma\left(\dfrac{w_1-w_2-1}{2}\right)\Gamma\left(\dfrac{w_1+w_2}{2}\right)}{\Gamma\left(\dfrac{w_1-w_2}{2}\right)\Gamma\left(\dfrac{w_1+w_2+1}{2}\right)}, & w_1 \geqslant w_2+2, \quad w_1+w_2 \text{ even}, \\ 0, & \text{elsewhere}, \end{cases} \qquad (42)
$$

where $\Gamma(n)$ denotes the Gamma function. The evaluation of this integral is discussed in Appendix D. Since I_Ω is independent of $\{i,j,k,l,m,\alpha,\beta,\gamma\}$, it can be stored in table form to increase the computational speed of the integral evaluation. Equation (37) may now be written

$$
I_3 = F(i,j,k,l,m,\alpha,\beta,\gamma) + 2 \sum_{w_1=1}^{\infty} \sum_{w_2=0}^{w_1} I_\Omega(w_1,w_2) I_R(w_1,w_2), \qquad (43)
$$

where

$$
I_\Omega(w_1,w_2) \equiv I_\Omega(w_1,w_2,w_2) \qquad (44)
$$

and

$$
I_R(w_1,w_2) \equiv I_R(w_1,w_2,w_2). \qquad (45)
$$

Inserting Eqs. (14) and (15) into Eq. (43), then breaking up the region of integration as $r_i \leqslant r_j \leqslant r_k$, allows us to write Eq. (43) as

$$
\begin{aligned}
&I_3(i,j,k,l,m,\alpha,\beta,\gamma)\\
&\quad = F(i,j,k,l,m,\alpha,\beta,\gamma) + 2 \sum_{\substack{w_1=1 \\ (w_1+w_2 \text{ even})}}^{\infty} \sum_{w_2=0}^{w_1} I_\Omega(w_1,w_2) \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} a_{w_2 lp} a_{w_2 mq}\\
&\qquad \times \{ W(j+2+w_1+w_2+2p, i+2-w_1+w_2+2q, k+2+l+m-2w_2-2p-2q, \beta,\alpha,\gamma)\\
&\qquad + W(j+2+w_1+w_2+2p, k+2+l-2p+2q, i+2-w_1+m-w_2-2q, \beta,\gamma,\alpha)\\
&\qquad + W(k+2+2w_2+2p+2q, j+2+l+w_1-2p-w_2, i+2+m-w_1-w_2-2q, \gamma,\beta,\alpha)\\
&\qquad - W(k+2+2w_2+2p+2q, i+2+w_1+m-w_2-2q, j+2-w_1+l-w_2-2p, \gamma,\alpha,\beta)\\
&\qquad - W(i+2+w_1+w_2+2q, j+2-w_1+w_2+2p, k+2+l+m-2w_2-2p-2q, \alpha,\beta,\gamma)\\
&\qquad - W(i+2+w_1+w_2+2q, k+2+m-2q+2p, j+2-w_1-w_2+l-2p, \alpha,\gamma,\beta) \}. \quad (46)
\end{aligned}
$$

Due to properties of the Pochhammer symbol, the p summation in Eq. (46) will terminate at $l/2$ for even l, and $(l+1)/2$ for odd l. Similarly, the q summation will terminate at $m/2$ for even m, and $(m+1)/2$ for odd m. Also, due to the Pochhammer symbols in Eq. (46) and the restriction on w_2 given in Eq. (42), the w_2 sum has the following termination conditions:

$$
\begin{aligned}
&\min\left\{ w_1, \frac{l}{2} \right\}, \quad l \text{ even and } m \text{ odd},\\
&\min\left\{ w_1, \frac{m}{2} \right\}, \quad m \text{ even and } l \text{ odd},\\
&\min\left\{ w_1, \frac{l}{2}, \frac{m}{2} \right\}, \quad l \text{ and } m \text{ both even},\\
&w_1, \quad l \text{ and } m \text{ both odd}.
\end{aligned}
$$

The w_1 sum is always nonterminating.

In addition to the individual constraints on i, j, k, l, and m mentioned in the Introduction, Eq. (46) requires that

$$
i+j+k+l+m \geqslant -8, \quad (47)
$$

which follows from Eq. (22).

V. NUMERICAL EVALUATION

For even l or m, the evaluation of I_1 and I_2 reduces to a finite sum of terms. For l and m both odd, it is useful to examine the asymptotic behavior of the series. I_1 is considered first.

The asymptotic behavior of a single W integral appearing in Eq. (18) is[32]

$$
W \sim \frac{1}{w^2}. \quad (48)
$$

However, if the W integrals are taken in the appropriate pairs, it can be shown that the set of six W integrals in Eq. (18) behaves like

$$
\text{set of six } W \sim \frac{1}{w^3}, \quad (49)
$$

due to the difference in signs. The product of the a_{tuv} coefficients in Eq. (17) exhibits the asymptotic behavior[32]

TABLE I. Values of $I_1(i,j,k,l,m,n,\alpha,\beta,\gamma)$.

i	j	k	l	m	α	β	γ	$I_1(i,j,k,l,m,n,\alpha,\beta,\gamma)$
0	0	0	−1	1	2.7	2.9	0.65	8.947 959 925 047 512 835 052 492
2	−1	1	3	−1	2.7	2.9	0.65	$3.066\,249\,710\,874\,801\,140\,638\,611\times10^{4}$
1	2	3	1	3	2.7	2.9	0.65	$-7.127\,529\,951\,937\,274\,842\,240\,935\times10^{7}$
0	2	0	2	1	2.7	2.9	0.65	$-1.155\,955\,548\,991\,711\,453\,287\,647\times10^{4}$
−1	3	2	0	4	2.7	2.9	0.65	$-4.898\,965\,561\,162\,249\,945\,560\,093\times10^{7}$
1	2	−1	3	4	2.7	2.9	0.65	$-5.595\,918\,656\,196\,395\,007\,062\,751\times10^{6}$

$$\frac{(-m/2)_w(-l/2)_w}{[(\frac{1}{2})_w]^2}\sim\frac{1}{w^{(l+m+2)/2}}. \tag{50}$$

Combining Eqs. (49) and (50) with the factor $1/(2w+1)$ from Eq. (18) leads to

$$I_1\sim\sum_w \frac{1}{w^{(l+m+10)/2}}. \tag{51}$$

The worst case that arises in the evaluation of the matrix elements of H_{oo} is $l=-1$ and $m=1$, where

$$I_1\sim\sum_w \frac{1}{w^5}, \tag{52}$$

which is suitable for direct numerical evaluation.

The asymptotic form for I_2 can be determined in a similar manner to that described for I_1. The result is

$$I_2\sim\sum_w \frac{1}{w^{(l+m+10)/2}}. \tag{53}$$

The worst case asymptotic behavior for I_2 is the same as that of I_1.

While the series representations of I_1 and I_2 are suitable for direct summation, such a process is rather time consuming for the most slowly converging cases. An improved approach is to apply the techniques discussed in Ref. 18. By first converting these monotonic series into equivalent alternating series, then applying convergence accelerators to the transformed series, a high-precision evaluation scheme is produced which is far more rapid for the most slowly converging integrals.

The sign differences on the coefficients of the W integrals in Eqs. (18) and (34) might be suggestive of possible numerical precision problems. However, both equations have been tested and found to be numerically stable. Tables I and II present some representative values for the I_1 and I_2 integrals, respectively.

A useful check for I_1 and I_2 can be derived by considering the integral

TABLE II. Values of $I_2(i,j,k,l,m,n,\alpha,\beta,\gamma)$.

i	j	k	l	m	α	β	γ	$I_2(i,j,k,l,m,n,\alpha,\beta,\gamma)$
0	2	0	−1	−1	2.7	2.9	0.65	−2.943 608 978 543 536 288 422 283
1	4	3	−1	3	2.7	2.9	0.65	$2.755\,380\,752\,419\,424\,787\,112\,100\times10^{6}$
2	1	1	1	−1	2.7	2.9	0.65	$-1.874\,264\,201\,450\,095\,466\,617\,095\times10^{2}$
1	3	2	0	2	2.7	2.9	0.65	$6.929\,145\,321\,870\,892\,125\,904\,854\times10^{3}$
0	4	0	2	−1	2.7	2.9	0.65	$-2.953\,164\,176\,902\,131\,531\,016\,592\times10^{3}$
3	2	−1	3	2	2.7	2.9	0.65	$2.411\,321\,913\,805\,718\,749\,043\,867\times10^{5}$

$$I_4(i,j,k,l,m,\alpha,\beta,\gamma)=\int r_1^i r_2^j r_3^k \frac{(r_1^2-r_2^2)(r_{31}^2-r_{23}^2)}{r_{12}^3} r_{23}^l r_{31}^m e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3. \tag{54}$$

Equation (54) can be evaluated using either I_1 or I_2, resulting in the following relationship:

$$\begin{aligned} &I_1(i,j,k,l,m+2,\alpha,\beta,\gamma)-I_1(i,j,k,l+2,m,\alpha,\beta,\gamma)\\ &\quad=I_2(i+2,j,k,l,m,\alpha,\beta,\gamma)-I_2(i,j+2,k,l,m,\alpha,\beta,\gamma). \end{aligned} \tag{55}$$

Hence, Eqs. (18) and (34) may checked against each other. Also, if $l=m=0$ in Eq. (26), $J_1=0$, and therefore

$$I_1(i,j,k,0,0,\alpha,\beta,\gamma)=I_2(i,j,k,0,0,\alpha,\beta,\gamma). \tag{56}$$

The asymptotic behavior of I_3 is the most difficult to estimate, due to the nested w_2 sum in Eq. (46). Let us define K to be the infinite series portion of the I_3 integral. The technique used to analyze the asymptotic behavior of K will be to divide the w_1, w_2 plane into three regions:

$$w_2=0,$$

$$0<w_2<w_1,$$

$$w_2=w_1.$$

Making these simplifications, it is possible to arrive at upper and lower bound estimates on the convergence rate for the w_1 summation.

First we examine the region where $w_2=0$. From Eq. (42), it can easily be shown that

$$I_\Omega\sim\frac{1}{w_1^2}. \tag{57}$$

If taken in the appropriate pairs, the sum of the W integrals in Eq. (46) behaves like

$$\text{set of six } W\sim\frac{1}{w_1^2}. \tag{58}$$

For the case $w_2=0$, the a_{tuv} coefficients in Eq. (46) lead to

$$\frac{(-m/2)_0(-l/2)_0}{[(\frac{1}{2})_0]^2}\sim 1. \tag{59}$$

Combining Eqs. (57), (58), and (59) gives the result

$$K\sim\sum_{w_1}\frac{1}{w_1^4}. \tag{60}$$

The asymptotic behavior of K in the other two regions may be obtained by similar methods. The results are

$$K\sim\sum_{w_1}\frac{1}{w_1^4} \quad\text{to}\quad \sum_{w_1}\frac{1}{w_1^{(l+m+14)/2}} \tag{61}$$

and

$$K\sim\sum_{w_1}\frac{1}{w_1^{(l+m+11)/2}}, \tag{62}$$

TABLE III. Values of $I_3(i,j,k,l,m,\alpha,\beta,\gamma)$.

i	j	k	l	m	α	β	γ	$I_3(i,j,k,l,m,\alpha,\beta,\gamma)$
1	−1	0	−1	1	2.7	2.9	0.65	$1.356\,087\,352\,189\,092\,17\times10^{2}$
1	2	2	1	1	2.7	2.9	0.65	$-6.371\,117\,492\,583\,756\times10^{4}$
3	1	2	3	1	2.7	2.9	0.65	$4.870\,134\,178\,451\,231\,6\times10^{7}$
2	2	3	−1	3	2.7	2.9	0.65	$6.471\,807\,181\,649\,945\times10^{5}$
4	1	2	3	3	2.7	2.9	0.65	$3.203\,917\,521\,846\,703\,1\times10^{10}$
3	1	−1	1	5	2.7	2.9	0.65	$5.849\,805\,492\,834\,822\,33\times10^{6}$

respectively. The second of the two results given in Eq. (61) was obtained using $w_2=w_1/2$. If instead we employ $w_2\approx w_1$ (with $w_2<w_1$), the same asymptotic behavior given in Eq. (62) would be obtained.

The convergence of the entire sum will be determined by the most slowly converging portion of that sum, which was found at $w_2=0$. So the entire sum behaves asymptotically like

$$K\sim\sum_{w_1}\frac{1}{w_1^4}. \tag{63}$$

It should be clear from this result that a large number of terms are needed to obtain an accurate result for K. Rather than directly summing the series, a convergence acceleration technique was applied to the infinite series portion of Eq. (46). The acceleration method employed was the Richardson extrapolation,[33] which has been used previously in the evaluation of other three-electron integration problems.[22] The Richardson extrapolation is given by

$$S_0=\sum_{k=0}^{N}\frac{S_{n+k}(n+k)^N(-1)^{k+N}}{k!(N-k)!}, \tag{64}$$

where S_0 is an approximation to the total sum in Eq. (46). A well-known difficulty associated with the application of Eq. (64) is that it is subject to numerical precision loss for higher values of N.

Careful examination of Eq. (46) shows it to be composed of two monotonically decreasing series, one for w_1 even and one for w_1 odd. Because the individual terms of the two series differ significantly in magnitude, a direct application of the Richardson extrapolation produces quite poor results. However, if Eq. (64) is applied individually to these two series, then a significant improvement in convergence is obtained. For both of these series, choosing $n=1$, the optimal value of N appears to be at about $N=23$. At this value, approximately 17 digits of precision may be obtained using 30 digit arithmetic. As was found for I_1 and I_2, the sign differences on the coefficients of the W integrals in Eq. (46) do not appear to be problematic. Some representative values of I_3 are presented in Table III.

For the case l and m both odd, evaluation of Eq. (5) via the approach developed in Sec. IV requires about half the computational time of previously published methods. For a number of test cases investigated, several more significant digits were obtained using the present approach. The increased speed of evaluation is directly tied to the much simpler formula that results in the present work, in comparison with what is obtained by breaking the integral into two separate parts. For l and m not both odd, evaluation methods may be employed which are both computationally faster and produce higher-precision results.[19]

VI. DISCUSSION

In this work, some key integrals are solved which are needed for the evaluation of ΔE_{REL} (the non-fine-structure contributions) for the ^2S states of three-electron systems. The basic approach employed in this work is to treat each of the integrand factors $(r_1^2-r_2^2)r_{12}^{-3}$, $(r_{31}^2-r_{23}^2)r_{12}^{-3}$, and $(r_1^2-r_2^2)r_{12}^{-2}$ as single expansions. For the I_1 and I_2 integrals investigated, the option of splitting these factors into separate parts does not exist, as each of the separate integrals diverge. A recursive scheme for an integral related to I_1 is developed by Yan and Drake.[23]

With the evaluation of I_1, I_2, and I_3, the difficulties associated with the evaluation of the matrix elements of the Breit–Pauli relativistic operators now lie elsewhere. Integrals of the form shown in Eq. (2) with $l=-2$ also arise in the evaluation of H_{oo} and H_{mass}, and these have known computational difficulties.[19] While several methods exist for their evaluation,[19–24] they are relatively slow for some choices of the arguments and are subject to precision loss. Development of better methods for handling integrals of this form would be useful.

ACKNOWLEDGMENTS

Support from the National Science Foundation (Grant Nos. PHY-9300863 and PHY-9600926) is greatly appreciated. The initial stages of this project were partially supported by the Donors of the Petroleum Research Fund, Administered by the American Chemical Society.

APPENDIX A: DIVERGENCE OF *I* INTEGRAL WHEN $n=-3$

This appendix illustrates that Eq. (2) diverges for at least one of l, m and $n=-3$. Consequently, Eqs. (3) and (4) cannot be broken into separate integrals of the form of Eq. (2).

Let

$$I(i,j,k,l,m,-3,\alpha,\beta,\gamma)\equiv\int r_1^i r_2^j r_3^k r_{23}^l r_{31}^m r_{12}^{-3} e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3. \tag{A1}$$

Because the integrand is everywhere positive,

$$I\geqslant\int_{(r_3-\varepsilon)\geqslant r_p} r_1^i r_2^j r_3^k r_{23}^l r_{31}^m r_{12}^{-3} e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3, \tag{A2}$$

where $(r_3-\varepsilon)>0$ and $p=1,2$.

Furthermore,

$$I\geqslant \min_{\substack{\mathbb{R}^9\\(r_3-\varepsilon)\geqslant r_p}}\{r_{23}^l\}\ \min_{\substack{\mathbb{R}^9\\(r_3-\varepsilon)\geqslant r_p}}\{r_{31}^m\}\int_{(r_3-\varepsilon)\geqslant r_p} r_1^i r_2^j r_3^k r_{12}^{-3} e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3. \tag{A3}$$

Due to the restrictions placed on the region of integration,

$$\min_{\substack{\mathbb{R}^9\\(r_3-\varepsilon)\geqslant r_p}}\{r_{23}^l\}\ \min_{\substack{\mathbb{R}^9\\(r_3-\varepsilon)\geqslant r_l}}\{r_{31}^m\}=\varepsilon^{l+m}. \tag{A4}$$

Substituting Eq. (A4) into Eq. (A3) yields,

$$I\geqslant\varepsilon^{l+m}\times\int_{(r_3-\varepsilon)\geqslant r_p} r_1^i r_2^j r_3^k r_{12}^{-3} e^{-\alpha r_1-\beta r_2-\gamma r_3}\, d\mathbf{r}_1\, d\mathbf{r}_2\, d\mathbf{r}_3. \tag{A5}$$

Integration over $\mathbf{r}_3$ produces a finite sum of positive terms involving integrals of the form

$$\int r_1^s r_2^t r_{12}^{-3} e^{-ar_1-br_2}\, d\mathbf{r}_1\, d\mathbf{r}_2. \tag{A6}$$

Choosing perimetric coordinates[6] it can easily be shown that Eq. (A6) diverges irrespective of the values of s, t, a, and b. Therefore Eq. (A5) must be divergent, and so Eq. (A1) must also diverge.

The extension to N-electron systems is obvious.

APPENDIX B: EXPANSION FOR r_{12}^{-3}

The generating function for the Legendre polynomials is[34]

$$\frac{1}{\sqrt{1-2rx+r^2}}=\sum_{n=0}^{\infty} r^n P_n(x). \tag{B1}$$

If the operator $\hat{O}\equiv 1+2r\ \partial/\partial r$ is applied to both sides of Eq. (B1)[26] one obtains

$$\frac{1-r^2}{(1-2rx+r^2)^{3/2}}=\sum_{n=0}^{\infty} (2n+1)r^n P_n(x). \tag{B2}$$

Making the substitutions $r=r_{12<}/r_{12>}$, $x=\cos\theta_{12}$, and $n=w_1$ in Eq. (B2) results in Eq. (8).

APPENDIX C: SOLUTION OF ANGULAR INTEGRALS FOR I_2

This appendix discusses the evaluation of the angular integrals Φ_1 and Φ_2 which arise in Eq. (27). If the addition theorem for Legendre polynomials,[26]

$$P_{w_1}(\cos\ \theta_{12})=P_{w_1}(\cos\ \theta_{23})P_{w_1}(\cos\ \theta_{31}) + 2\sum_{n=1}^{w_1} \frac{(w_1-n)!}{(w_1+n)!} P_{w_1}^n(\cos\ \theta_{23})P_{w_1}^n(\cos\ \theta_{31})\cos(n(\phi_{23}-\phi_{31})), \tag{C1}$$

is applied to $P_{w_1}(\cos\theta_{12})$ in Eq. (28), and the integration over ϕ_{23} and ϕ_{31} is performed, then the finite sum in Eq. (C1) vanishes. Thus Φ_1 simplifies to

$$\Phi_1=4\pi^2\int_{-1}^{1} xP_{w_1}(x)P_{w_3}(x)dx\int_{-1}^{1} P_{w_1}(x)P_{w_2}(x)dx. \tag{C2}$$

Using the standard recurrence relationship

$$(2v+1)xP_v(x)=(v+1)P_{v+1}(x)+vP_{v-1}(x), \tag{C3}$$

and the orthogonality property of the Legendre polynomials, allows Eq. (C2) to be simplified to Eq. (30). The result for Φ_2 can be obtained by symmetry [switch $1\leftrightarrow 2$ and $w_2\leftrightarrow w_3$ in Eq. (28), thereby obtaining Eq. (31)].

APPENDIX D: SOLUTION OF ANGULAR INTEGRAL FOR I_3

This appendix discusses the evaluation of the angular integral I_Ω, Eq. (39), which occurs in I_3. We choose our coordinate system such that

$$d\Omega_1\ d\Omega_2\ d\Omega_3=\sin\ \theta_{12}\ d\theta_{12}\ d\phi_{12}\ \sin\ \theta_{31}\ d\theta_{31}\ d\phi_{31}\ \sin\ \theta_1\ d\theta_1\ d\phi_1. \tag{D1}$$

Expanding $P_{w_2}(\cos\theta_{23})$ by the addition theorem for Legendre polynomials, then integrating over θ_1, ϕ_1, ϕ_{12}, ϕ_{31}, and θ_{31}, yields

$$I_\Omega(w_1,w_2,w_3)=\frac{32\pi^3}{2w_2+1}\ \delta_{w_2w_3}\int_0^{\pi} T_{w_1}(\cos\ \theta_{12})P_{w_2}(\cos\ \theta_{12})\sin\ \theta_{12}\ d\theta_{12}. \tag{D2}$$

Expanding $P_{w_2}(\cos\theta_{12})$ in a Fourier sine series,[34] and using the identity

$$T_{w_1}(\cos\ \theta_{12})=\cos(w_1\theta_{12}), \tag{D3}$$

produces

$$I_\Omega = \frac{64\pi^{5/2}}{2w_2+1}\,\delta_{w_2 w_3}\sum_{k=0}^{\infty}\frac{(\frac{1}{2})_k\Gamma(w_2+k+1)}{k!\,\Gamma(w_2+k+\frac{3}{2})}\int_0^{\pi}\cos(w_1\theta_{12})\sin[(w_2+2k+1)\theta_{12}]\sin\,\theta_{12}\,d\theta_{12}. \tag{D4}$$

Evaluation of this integral, followed by some straightforward simplification, leads to Eq. (42).

[1] F. W. King, J. Mol. Struct. (Theochem) **400**, 7 (1997).
[2] F. W. King, Adv. At. Mol. Opt. Phys. **40**, 57 (1999).
[3] K. T. Chung, Phys. Rev. A **44**, 5421 (1991).
[4] Z.-C. Yan and G. W. F. Drake, Phys. Rev. Lett. **81**, 774 (1998).
[5] F. W. King, D. G. Ballegeer, D. J. Larson, P. J. Pelzl, S. A. Nelson, T. J. Prosa, and B. M. Hinaus, Phys. Rev. A **58** (1998) (in press).
[6] C. L. Pekeris, Phys. Rev. **112**, 1649 (1958).
[7] C. L. Pekeris, Phys. Rev. **126**, 143 (1962).
[8] H. M. James and A. S. Coolidge, Phys. Rev. **49**, 688 (1936).
[9] Y. Öhrn and J. Nordling, J. Chem. Phys. **39**, 1864 (1963).
[10] L. Szász, J. Chem. Phys. **35**, 1072 (1961).
[11] J. F. Perkins, J. Chem. Phys. **48**, 1985 (1968).
[12] Y. K. Ho and B. A. P. Page, J. Comput. Phys. **17**, 122 (1975).
[13] A. Berk, A. K. Bhatia, B. R. Junker, and A. Temkin, Phys. Rev. A **34**, 4591 (1986).
[14] D. M. Fromm and R. N. Hill, Phys. Rev. A **36**, 1013 (1987).
[15] E. Remiddi, Phys. Rev. A **44**, 5492 (1991).
[16] G. W. F. Drake and Z.-C. Yan, Phys. Rev. A **52**, 3681 (1995).
[17] F. E. Harris, Phys. Rev. A **55**, 1820 (1997).
[18] P. J. Pelzl and F. W. King, Phys. Rev. E **57**, 7268 (1998).
[19] F. W. King, Phys. Rev. A **44**, 7108 (1991).
[20] F. W. King, K. J. Dykema, and A. Lund, Phys. Rev. A **46**, 5406 (1992).
[21] A. Lüchow and H. Kleindienst, Int. J. Quantum Chem. **45**, 445 (1993).
[22] I. Porras and F. W. King, Phys. Rev. A **49**, 1637 (1994).
[23] Z.-C. Yan and G. W. F. Drake, J. Phys. B **30**, 4723 (1997). For an application to the fine-structure terms of the $1s^2 2p\ ^2P$ state of the lithium atom, see Z.-C. Yan and G. W. F. Drake, Phys. Rev. Lett. **79**, 1646 (1997).
[24] F. W. King, Int. J. Quantum Chem. (submitted for publication).
[25] R. A. Sack, J. Math. Phys. **5**, 245 (1964).
[26] H. Hochstadt, *The Functions of Mathematical Physics* (Dover, New York, 1986).
[27] S. Chapman, Q. J. Math. **185**, 16 (1916).
[28] V. McKoy, J. Chem. Phys. **42**, 2959 (1965).
[29] F. Byron, Jr. and C. J. Joachain, Phys. Rev. **146**, 1 (1966).
[30] A. M. Frolov and V. H. Smith, Jr., Int. J. Quantum Chem. **63**, 269 (1997).
[31] E. A. Hylleraas, Z. Phys. **48**, 469 (1928).
[32] S. Larsson, Phys. Rev. **169**, 49 (1968).
[33] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw–Hill, New York, 1978), p. 375.
[34] M. Abramowitz and I. A. Stegun (editors), *Handbook of Mathematical Functions* (Dover, New York, 1970).

Local $U(2,2)$ symmetry in relativistic quantum mechanics

Felix Finster
Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138

(Received 15 May 1998; accepted for publication 9 June 1998)

Local gauge freedom in relativistic quantum mechanics is derived from a measurement principle for space and time. For the Dirac equation, one obtains local $U(2,2)$ gauge transformations acting on the spinor index of the wave functions. This local $U(2,2)$ symmetry allows a unified description of electrodynamics and general relativity as a classical gauge theory. © *1998 American Institute of Physics.* [S0022-2488(98)03012-6]

I. CONNECTION BETWEEN LOCAL GAUGE FREEDOM AND POSITION MEASUREMENTS

In Ref. 1, it was suggested to link the physical gauge principle with quantum mechanical measurements of the position variable. In the present paper, we will extend this concept to relativistic quantum mechanics and apply it to the Dirac equation. For Dirac spinors, we obtain local $U(2,2)$ gauge freedom. Our main result is that this $U(2,2)$ symmetry allows a natural description of both electrodynamics and general relativity as a classical gauge theory. This is shown by deriving a $U(2,2)$ spin connection from the Dirac operator and analyzing the geometry of this connection. Although we develop the subject from a particular point of view, this paper can be used as an introduction to the Dirac theory in curved space–time.

In contrast to Ref. 1, where the point of interest is the measure theoretic derivation of local gauge transformations, here we will concentrate on the differential geometry of the Dirac operator. In order to keep measure theory out of this paper, we will use a bra/ket notation in position space. This allows us to explain the basic ideas and results of Ref. 1 in a simple, nontechnical way, which will be sufficient for the purpose of this paper. Nevertheless, one should keep in mind that the bra/ket symbols and the δ normalizations are only a formal notation; the mathematical justification for this formalism is given in Ref. 1.

We begin with the example of a scalar particle in nonrelativistic quantum mechanics. The particle is described by a wave function $\Psi(\mathbf{x})$, which is a vector of the Hilbert space $H = L^2(\mathbb{R}^3)$. The physical observables correspond to self-adjoint operators on H. The position operators $\mathbf{X}$, for example, are given as multiplication operators with the coordinate functions, X^i: $\Psi(\mathbf{x}) \to x^i \Psi(\mathbf{x})$, $i = 1,...,3$. Our definition of the Hilbert space as a space of functions in the position variable was only a matter of convenience; e.g., we could just as well have introduced H as functions in momentum space. Therefore it seems reasonable to forget about the fact that H is a function space and consider it merely as an abstract Hilbert space. After this generalization, which is often implicitly assumed in quantum mechanics, we must construct the representation of H as wave functions. For this purpose, we choose an ''eigenvector basis'' $|\mathbf{x}\rangle$ of the position operators,

$$X^i|\mathbf{x}\rangle = x^i|\mathbf{x}\rangle, \quad \langle \mathbf{x}|\mathbf{y}\rangle = \delta^3(\mathbf{x}-\mathbf{y}), \tag{1}$$

and define the wave function of a vector $\Psi \in H$ by $\Psi(\mathbf{x}) = \langle \mathbf{x}|\Psi\rangle$.

This ''position representation'' of the Hilbert space is quite elementary and is currently used in physics. We point out that it is not unique, because the ''eigenvectors'' $|\mathbf{x}\rangle$ are only determined up to a phase. Namely, we can transform $|\mathbf{x}\rangle$ according to

$$|\mathbf{x}\rangle \to e^{ie\Lambda(\mathbf{x})}|\mathbf{x}\rangle, \tag{2}$$

with a real function $\Lambda(\mathbf{x})$. This corresponds to a local phase transformation,

 © 1998 American Institute of Physics

$$\Psi(\mathbf{x}) \rightarrow e^{-ie\Lambda(\mathbf{x})} \Psi(\mathbf{x}), \tag{3}$$

of the wave functions. The arbitrariness of the local phase of the wave functions can also be understood directly from the fact that the wave function itself is not observable, only its absolute square $|\Psi|^2$ has a physical interpretation as probability density. It is important for the following that the local phase transformations (2), (3) can be interpreted as $U(1)$ gauge transformations. To see this, we consider the canonical momentum operator $\boldsymbol{\pi} = -i\nabla - e\mathbf{A}$ with magnetic vector potential $\mathbf{A}$. Under the transformation (2), the canonical momentum behaves like

$$\boldsymbol{\pi} \rightarrow e^{-ie\Lambda(\mathbf{x})} \boldsymbol{\pi} e^{ie\Lambda(\mathbf{x})} = -i\nabla - e(\mathbf{A} - \nabla\Lambda),$$

which corresponds to the gauge transformation $\mathbf{A} \rightarrow \mathbf{A} - \nabla\Lambda$ of the vector potential.

In this way, we have explained the local $U(1)$ gauge symmetry of the magnetic field from the fact that space is a quantum mechanical observable. Unfortunately, this natural explanation of local $U(1)$ gauge freedom only works for scalar particles. In the general case with spin, the wave functions have several components, $\Psi \in H = L^2(\mathbb{R}^3)^m$ (i.e., $m = 2s+1$ for particles with spin s). We again consider H as an abstract Hilbert space. In order to construct the representation of a vector $\Psi \in H$ as a function, we choose an ''orthonormal basis'' $|\mathbf{x}\alpha\rangle$, $\mathbf{x} \in \mathbb{R}^3$, $\alpha = 1,...,m$ of the position operators,

$$X^i |\mathbf{x}\alpha\rangle = x^i |\mathbf{x}\alpha\rangle, \quad \langle \mathbf{x}\alpha | \mathbf{y}\beta \rangle = \delta_{\alpha\beta} \delta^3(\mathbf{x}-\mathbf{y}), \tag{4}$$

and define the wave function by $\Psi^\alpha(\mathbf{x}) = \langle \mathbf{x}\alpha | \Psi \rangle$. The basis $|\mathbf{x}\alpha\rangle$ is only unique up to local unitary transformations,

$$|\mathbf{x}\alpha\rangle \rightarrow \sum_{\beta=1}^{m} (U^{-1})^\alpha_\beta(\mathbf{x}) |\mathbf{x}\beta\rangle, \quad \text{with} \quad U(\mathbf{x}) \in U(m), \tag{5}$$

as is verified by substituting into (4). Thus the wave functions can be transformed according to

$$\Psi(\mathbf{x}) \rightarrow U(\mathbf{x}) \Psi(\mathbf{x}). \tag{6}$$

It would be nice if we could again identify the local $U(m)$ transformations (5), (6) with physical gauge transformations. Then the local gauge principle would no longer be an *a priori* principle in physics. It would be a consequence of a quantum mechanical ''measurement principle,'' namely the description of space with observables X^i on an abstract Hilbert space. Notice that the local $U(m)$ transformations (6) are a generalization of the phase transformations (3). Our idea is that the additional degrees of freedom of the larger gauge group in (6) might make it possible to describe additional interactions (like gravitation or the weak and strong forces). Notice that, as a great advantage of our procedure, the local gauge group could no longer be chosen arbitrarily; it would be determined by the configuration of the spinors. Thus, we could hope to get into the position to explain the gauge groups observed in physics. At the moment, however, it is not clear if the interpretation of (5), (6) as gauge transformations really makes physical sense. We will in the following simply assume it as a postulate and want to study its further consequences.

In view of our later generalization to curved space (and curved space–time), it is convenient to introduce the spectral measure $dE_\mathbf{x}$ of the position operators: We form the ''projectors'' $E_\mathbf{x}$ on the ''eigenspaces'' of $\mathbf{X}$,

$$E_\mathbf{x} = \sum_{\alpha=1}^{m} |\mathbf{x}\alpha\rangle\langle\mathbf{x}\alpha|. \tag{7}$$

The $E_\mathbf{x}$ do not depend on the choice of $|\mathbf{x}\alpha\rangle$; we can characterize them by the operator relations,

$$X^i E_\mathbf{x} = x^i E_\mathbf{x}, \quad E_\mathbf{x} E_\mathbf{y} = \delta^3(\mathbf{x}-\mathbf{y}) E_\mathbf{x}. \tag{8}$$

The δ^3 normalization in (8) leads to difficulties as soon as curved coordinate systems are considered. To avoid these problems, it is useful to combine $E_{\mathbf{x}}$ with the integration measure by introducing the spectral measure $dE_{\mathbf{x}} := E_{\mathbf{x}}\, d\mathbf{x}$. Integrating over the spectral measure yields an operator on the Hilbert space. For example, we have

$$X^i = \int_{\mathbb{R}^3} x^i \, dE_{\mathbf{x}}, \quad \chi_V = \int_V dE_x, \tag{9}$$

where χ_V is the multiplication operator with the characteristic function [i.e., $(\chi_V \Psi)(\mathbf{x})$ is equal to $\Psi(\mathbf{x})$ if $\mathbf{x} \in V$ and vanishes otherwise]. These relations can be verified directly with the help of (4) and (7). Actually, the spectral measure is completely characterized by (9). In the mathematical paper,[1] spectral measures on a manifold are used as the starting point. In this more general approach, the number of components of the wave functions $m = m(\mathbf{x})$ may vary in space. We call it the *spin dimension*.

II. GENERALIZATION TO DIRAC SPINORS, $U(2,2)$ GAUGE SYMMETRY

If our explanation of local gauge freedom shall have general significance, it must be possible to extend it to the relativistic setting. We start the analysis with Dirac wave functions in Minkowski space. [We mention for completeness that *Minkowski space* is a four-dimensional real vector space with a scalar product of signature $(+---)$. A coordinate system where this scalar product coincides with the Minkowski metric η_{ij} is called *reference frame*. We will mainly work in a fixed reference frame and can then identify Minkowski space with $\mathbb{R}^4$.] Since we do not yet want to introduce Dirac matrices and the Dirac operator, we consider the wave functions $\Psi, \Phi, \ldots$, just as four-component functions. We define a scalar product on the spinors,

$$\prec \Psi, \Phi \succ (x) = \sum_{\alpha=1}^{4} s_\alpha (\Psi^\alpha(x))^* \Phi^\alpha(x), \quad \text{with } s_1 = s_2 = 1, \quad s_3 = s_4 = -1, \tag{10}$$

which is called *spin scalar product*. It is indefinite of signature (2,2). Later, after the introduction of the Dirac matrices in Dirac representation, the spin scalar product will coincide with $\overline{\Psi}\Phi$, where $\overline{\Psi} = \Psi^* \gamma^0$ is the adjoint spinor. Our definition without referring to Dirac matrices will clarify the relation between coordinate and gauge transformations in Sec. III.

The basic question is how we want to describe space–time. At the moment, position and time are merely parameters of the wave functions. This is the usual description in relativistic quantum mechanics. In the previous section, however, it was essential for the explanation of local gauge transformations that space corresponds to some operators on the wave functions. In order to generalize this ''measurement principle'' to the relativistic setting, we introduce ''observables'' for space and time as multiplication operators with the coordinate functions,

$$X^i : \Psi(x) \to x^i \Psi(x), \quad i = 0, \ldots, 3. \tag{11}$$

We point out that these operators, especially the time operator X^0, are commonly not used in relativistic quantum mechanics. Nevertheless, our description of Minkowski space makes sense, as we will see in the following.

The next step is the introduction of a scalar product on the wave functions. Usually one considers the positive product,

$$(\Psi | \Phi) = \int_{\mathbb{R}^3} \Psi^*(t, \mathbf{x}) \Phi(t, \mathbf{x}) \, d\mathbf{x}, \tag{12}$$

where the spinors are integrated over a space-like hypersurface $t = \text{const}$ (we avoid the notation $\Psi^* \Phi = \overline{\Psi} \gamma^0 \Phi$ with the adjoint spinor because we have no Dirac matrices at the moment). The integrand of $(\Psi | \Psi)$ can be interpreted as the probability density of the particle. For solutions of the Dirac equation, current conservation implies that (12) is independent of t. As an apparent problem of this scalar product, time measurements do not make sense, because expectations $(\Psi | X^0 | \Psi)$

$=t(\Psi|\Psi)$ depend on the choice of the hypersurface. The only way out is to introduce a different scalar product where the spinors are also integrated over the time variable, namely,

$$\langle \Psi | \Phi \rangle = \int_{\mathbb{R}^4} <\Psi|\Phi> d^4x. \tag{13}$$

In contrast to (12), the scalar product (13) has no immediate physical interpretation. Nevertheless, it causes no problem to consider (13) as the fundamental scalar product on the spinors, for the following reason: The scalar product does not enter in the physical equations (i.e., the Dirac equation and the classical field equations), we only need it for the probabilistic interpretation of the wave functions. For this final interpretation, we can just introduce (12) by an additional mathematical construction [see Eq. (27)].

We choose (13) as the fundamental scalar product on the wave functions. We denote the corresponding function space by $(H, \langle \cdot | \cdot \rangle)$ and consider it together with the position/time operators (11) as an abstract scalar product space. In order to represent the vectors of H as wave functions, we choose an ''eigenvector basis'' $|x\alpha\rangle$, $x \in \mathbb{R}^4$, $\alpha = 1,...,4$ of the position/time operators,

$$X^i |x\alpha\rangle = x^i |x\alpha\rangle, \quad \langle x\alpha | y\beta \rangle = s_\alpha \delta_{\alpha\beta} \delta^4(x-y), \tag{14}$$

and define for $\Psi \in H$ the corresponding wave function by $\Psi^\alpha(x) = \langle x\alpha | \Psi \rangle$. This is a relativistic analog of (4). The additional factors $s_\alpha = \pm 1$ occur as a consequence of the indefinite spin scalar product. The arbitrariness of the choice of $|x\alpha\rangle$ leads to the transformations

$$|x\alpha\rangle \rightarrow \sum_{\beta=1}^{4} (U^{-1})^\alpha_\beta(x) |x\beta\rangle$$

$$\Psi(x) \rightarrow U(x)\Psi(x), \quad \text{with } U(x) \in U(2,2), \tag{15}$$

which we interpret as local $U(2,2)$ gauge transformations. According to this interpretation, we call $|x\alpha\rangle$ a *gauge*. Notice that the gauge group is noncompact, which is directly related to the indefiniteness of the spin scalar product. Similar to the nonrelativistic case, the spectral measure is given by

$$dE_x = \sum_{\beta=1}^{4} s_\beta |x\beta\rangle\langle x\beta| d^4x.$$

We say that the spin dimension is (2,2).

We remark that the indefiniteness of the scalar product $\langle \cdot | \cdot \rangle$ leads to mathematical problems in the rigorous derivation of gauge transformations. A first, but not fully convincing attempt toward a satisfying mathematical formulation was made in Ref. 2 (which also contains a preliminary version of Ref. 1).

III. THE FREE DIRAC OPERATOR, LORENTZIAN TRANSFORMATIONS

Having introduced the Dirac wave functions in Minkowski space, we can now define the free Dirac operator: We fix the reference frame, choose a special gauge, and introduce the partial differential operator $i\partial\!\!\!/ := i\gamma^j \partial_j$ on the wave functions, where γ^j denote the usual (4×4)-matrices in the Dirac representation acting on the spinor index,

$$\gamma^0 = \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \quad i = 1,2,3$$

($\sigma^1, \sigma^2, \sigma^3$ are the Pauli matrices). Now we forget about the fact that the free Dirac operator was defined in a special gauge and a special reference frame and consider it as an operator on H. We denote this operator by G and write the free Dirac equation in the coordinate and gauge invariant form $(G-m)\Psi = 0$.

The Dirac operator is Hermitian (with respect to the scalar product $\langle \cdot | \cdot \rangle$). This is most conveniently verified in the original gauge and reference frame where $G = i\partial\!\!\!/$: For simplicity, we rewrite the factors s_α in (10) with the matrix γ^0 and obtain

$$\langle G\Psi | \Phi \rangle = \int_{\mathbb{R}^4} <G\Psi | \Phi>_{|x} \, d^4x = \int_{\mathbb{R}^4} (i\gamma^j \, \partial_j \Psi)(x)^* \gamma^0 \Phi(x) d^4x = -i \int_{\mathbb{R}^4} (\partial_j \Psi)^* (\gamma^j)^* \gamma^0 \Phi \; d^4x$$

$$= i \int_{\mathbb{R}^4} \Psi^* \gamma^0 (\gamma^0 (\gamma^j)^* \gamma^0) \partial_j \Phi = i \int_{\mathbb{R}^4} \Psi^* \gamma^0 \gamma^j \, \partial_j \Phi = \int_{\mathbb{R}^4} <\Psi | G\Phi>_{|x} d^4x = \langle \Psi | G\Phi \rangle .$$

Note that the γ-matrices are not Hermitian [more precisely, $(\gamma^0)^* = \gamma^0$ and $(\gamma^j)^* = -\gamma^j$ for $j = 1,2,3$], but they are self-adjoint with respect to the spin scalar product, $<\gamma^j \Psi | \Phi> = <\Psi | \gamma^j \Phi>$.

We point out that G is as operator on H a coordinate- and gauge-independent mathematical object (although its definition might depend on the special gauge and reference frame chosen at the beginning). Both coordinate and gauge transformations are merely passive transformations and lead to equivalent representations of G. Especially, transformations of the space–time coordinates and of the spinors are independent of each other. This is a major difference to the usual description of the Dirac equation. Therefore we will now discuss coordinate and gauge transformations and explain what ''Lorentzian invariance'' of the free Dirac operator exactly means in our setting.

We first look at the Dirac operator in a general gauge, but still in the original reference frame. The gauge transformation (15) leads to

$$i\partial\!\!\!/ \rightarrow G = U i\partial\!\!\!/ U^{-1} = iG^j \frac{\partial}{\partial x^j} + B,$$

with

$$G^j(x) = U(x) \gamma^j U(x)^{-1}, \quad B(x) = U(x) \gamma^j (\partial_j U(x)^{-1}).$$

As explained before, all these operators are equivalent; of course, it is most convenient to work in the original gauge with $G = i\partial\!\!\!/$.

Next we look at a Lorentzian transformation. The transformation of the space–time variables is described by a coordinate transformation $x^j \rightarrow \Lambda^j_k x^k$, where Λ is an isometry of Minkowski space. The Dirac operator transforms like

$$i\partial\!\!\!/ \rightarrow G = i\gamma^j \Lambda^k_j \partial_k . \tag{16}$$

Thus, in the new reference frame, the Dirac operator no longer has the original form. This might seem to contradict Lorentzian invariance, but we must keep in mind that all operators related to each other by gauge transformations are equivalent. Thus we must look for a gauge transformation such that the Dirac operator in the new reference frame again coincides with $i\partial\!\!\!/$. As is shown in standard textbooks on relativistic quantum mechanics (see, e.g., Ref. 3), there is a (4×4)-matrix $U(\Lambda)$ with

$$U(\Lambda) \gamma^j \Lambda^k_j U(\Lambda)^{-1} = \gamma^k .$$

It is important for our purpose that this matrix is unitary with respect to the spin scalar product, $U(\Lambda) \in U(2,2)$. Therefore we can perform a gauge transformation with $U(x) \equiv U(\Lambda)$. In the new gauge we again have $G = i\partial\!\!\!/$, which shows Lorentzian invariance of G. To summarize, we describe a Lorentzian transformation in two steps. The first step is a coordinate; the second a gauge transformation. If both steps are performed at once, the transformation of the space–time coordinates and of the spinors are related in the usual way. Notice, however, that the second step is only a matter of convenience; we could just as well work in the new reference frame in any other gauge. This splitting of the space–time and spinor transformation becomes possible because the conditions $G^j = \gamma^j$ give a link between coordinate and gauge transformations.

We remark that it is not trivial that a spinorial equation can be described with passive coordinate and gauge transformations. As a counterexample, we consider the Weyl equation: Chiral fermions are usually described by two-component spinors χ satisfying the equation

$$i\sigma^j \frac{\partial}{\partial x^j} \chi = 0, \tag{17}$$

with Pauli matrices $\sigma^0 = \mathbb{1}, \boldsymbol{\sigma}$. This equation is Lorentzian invariant in the sense that there is a one-to-one correspondence between the solutions of (17) in different reference frames (see, e.g., Ref. 3). We assume that the Weyl operator $W := i\sigma^j \partial_j$ were given as a coordinate- and gauge-independent object (we will not specify the scalar product on the wave functions). After a Lorentzian transformation $x^j \to \Lambda^j_k x^k$, the Weyl operator has the form $W = i\sigma^k \Lambda^j_k \partial_j$. If W was Lorentzian invariant, it should be possible to perform a gauge transformation such that $W = i\sigma^j \partial_j$. Thus there should be a matrix $U(\Lambda)$ with

$$U(\Lambda) \sigma^k \Lambda^j_k U(\Lambda)^{-1} = \sigma^j.$$

Taking the trace on both sides, however, yields the equation $\Lambda^j_0 = \delta^j_0$, which is only satisfied for $\Lambda = \mathbb{1}$. We conclude that the Weyl operator cannot be described as a Lorentz invariant operator in our setting (this is not a serious problem because the Weyl equation can always be obtained as the left- or right-handed component of the massless Dirac equation).

IV. GENERAL DEFINITION OF THE DIRAC OPERATOR

In the previous section, we introduced the free Dirac operator as an operator on the wave functions in Minkowski space. We point out that this definition involved some global assumptions on space–time. First of all, the reference frames gave global coordinate systems. Furthermore, we chose a special gauge at the beginning and described Lorentzian transformations using global gauge transformations (i.e., transformations independent of x). According to the principle of general relativity, however, global conditions do not make sense. In this section, we replace them by corresponding local conditions, which will yield the general definition of the Dirac operator.

We start with the observation that the Minkowski metric can be derived from the Dirac matrices by $\eta^{ij}\mathbb{1} = \frac{1}{2}\{G^i, G^j\}$. This allows us to forget about the Minkowski metric and view Minkowski space simply as a four-dimensional vector space. For the generalization to curved coordinate systems, we replace this vector space by a four-dimensional (smooth) manifold M. Notice that M has no Lorentzian structure; we will later recover the Lorentzian metric from the Dirac operator.

Next we must generalize the scalar product space $(H, \langle \cdot | \cdot \rangle)$ and the position/time operators to the setting of a space–time manifold. This can be most easily done with the help of spectral measures: We consider an indefinite scalar product space $(H, \langle \cdot | \cdot \rangle)$ and introduce a spectral measure $(dE_x)_{x \in M}$ on M of spin dimension (2,2). For any chart (x^i, U), we define corresponding position/time operators by

$$X^i = \int_U x^i \, dE_x. \tag{18}$$

We choose an ''eigenvector basis'' $|x\alpha\rangle$, $x \in U$, $\alpha = 1,...,4$ of the X^i,

$$X^i |x\alpha\rangle = x^i |x\alpha\rangle, \quad \langle x\alpha | y\beta \rangle = s_\alpha \delta_{\alpha\beta} \delta^4(x-y), \tag{19}$$

and represent a vector $\Psi \in H$ as the wave function $\Psi^\alpha(x) = \langle x\alpha | \Psi \rangle$, $x \in U$. This construction generalizes (9), (14) to curved coordinate systems. The spectral measure in (18) allows us to extend the usual changes of charts to the level of a calculus of operators.

After these preparations, we can introduce the Dirac operator G. First of all, it shall be a Hermitian operator on H. In a representation of H as wave functions, we assume it to be a partial differential operator of first order, i.e.,

$$G = iG^j \frac{\partial}{\partial x^j} + B, \tag{20}$$

with (4×4)-matrices $G^j(x)$, $B(x)$. Since G is (as an operator on H) a coordinate- and gauge-independent object, the matrices $G^j(x)$ and $B(x)$ have a well-defined behavior under coordinate and gauge transformations. As an additional condition, we demand that the matrices G^j shall coincide locally with the usual Dirac matrices:

Definition IV.1: A partial differential operator G on H is called a **Dirac operator**, *if, for any* $p \in M$, *there is a chart* (x^i, U) *around p and a gauge* $|x\alpha\rangle$ *such that G has the form (20) with*

$$G^j(p) = \gamma^j. \tag{21}$$

This definition can also be understood in more physical terms: Let us assume that an observer (e.g., a spacecraft passing through the space–time point p) chooses a coordinate system and gauge satisfying (21). Since the Dirac matrices at p coincide with those of the free Dirac operator, the observer at p has the impression that space–time locally resembles Minkowski space. Thus, we expect that the Dirac operator gives a Lorentz metric g_{jk}, which coincides with the Minkowski metric at p, $g_{jk}(p) = \eta_{jk}$. We can really derive this Lorentz metric from G; it is most conveniently given by

$$g^{jk}(x)\mathbb{1} = \tfrac{1}{2}\{G^j(x), G^k(x)\}. \tag{22}$$

In this way, our definition of the Dirac operator incorporates the principle of general relativity; the coordinate and gauge satisfying (21) gives a local reference frame. If the partial derivatives $\partial_j g_{kl}(p)$ of the metric also vanish (which can be arranged by choosing a so-called ''normal coordinate system''), the reference frame is a ''local system of inertia,'' where the observer feels no gravitational force at p. It might happen that $\partial_j G^k(p) \neq 0$, although $\partial_j g_{kl}(p)$ vanishes. This corresponds to a ''spin-mixing force'' which is related to our $U(2,2)$ symmetry. The matrix $B(x)$ can be used for the description of additional interactions, especially of electromagnetism $G = i\partial\!\!\!/ + eA\!\!\!/$. Note that $B(x)$ was not at all specified in our definition of the Dirac operator. This will turn out to be too general for physical applications. We postpone the discussion and specification of $B(x)$ to Sec. VI.

Unfortunately, the equation (19) for the ''eigenvector basis'' is too simple and must be modified. This is only a minor technical point, but, nevertheless, we give it in detail: The δ^4-distribution in (19) is useful in combination with the integration measure d^4x, because integrations can be easily carried out using the relation $\int f(x)\,\delta^4(x)\,d^4x = f(0)$. On a manifold, however, one works with the invariant measure $d\mu := \sqrt{|g|}\,d^4x$ instead of d^4x. In order to compensate the factor $\sqrt{|g|}$ of the integration measure in the integral over the δ-distribution, we must view the combination $|g(y)|^{-1/2}\delta^4(x-y)$ as the ''invariant δ-distribution'' at a point y of the manifold. Thus, we are led to replace (19) by

$$X^i|x\alpha\rangle = x^i|x\alpha\rangle, \quad \langle x\alpha|y\beta\rangle = s_\alpha \delta_{\alpha\beta} \frac{1}{\sqrt{|g|}}\,\delta^4(x-y). \tag{23}$$

This definition of a gauge is much better than (19) because $|x\alpha\rangle$ now behaves under coordinate transformations $x \to x'(x)$ simply like $|x\alpha\rangle \to |x'\alpha\rangle$. We again introduce the wave functions by $\Psi^\alpha(x) = \langle x\alpha|\Psi\rangle$ and obtain a representation of G as the differential operator (20). Notice that this indirect definition of a gauge by first introducing (19) and then replacing it by (23) was necessary because we had no invariant measure at the beginning. We must check that this procedure is consistent: The transition from (19) to (23) is described by the transformation $|x\alpha\rangle \to |g(x)|^{-1/4}|x\alpha\rangle$. Consequently, the Dirac operator transforms as $G \to |g(x)|^{-1/4} G |g(x)|^{1/4}$, and thus $G^j \to G^j$, $B \to B + (i/4)G^j(\partial_j \log|g|)$. We conclude that the definition of the Dirac operator and the Lorentz metric (22) are independent of whether (19) or (23) are used.

We will in the following always work with (23) as the definition of a *gauge*. The freedom in the choice of the gauge again describes local $U(2,2)$ transformations of the wave functions

$$\Psi \to U(x)\Psi(x), \quad \text{with } U(x) \in U(2,2). \tag{24}$$

We define the spin scalar product by (10). According to (23), the scalar product is obtained by integrating over the spin scalar product,

$$\langle \Psi | \Phi \rangle = \int_M < \Psi | \Phi > d\mu. \tag{25}$$

The Dirac equation is again introduced as the operator equation

$$(G - m)\Psi = 0. \tag{26}$$

On solutions of the Dirac equation, we finally define a generalization of the scalar product (12): We choose a space-like hypersurface $\mathcal{H}$ with normal vector field ν and set

$$(\Psi | \Phi)_{\mathcal{H}} = \int_{\mathcal{H}} < \Psi | G^j | \Phi > \nu_j \, d\mu_{\mathcal{H}}, \tag{27}$$

where $d\mu_{\mathcal{H}}$ is the measure on $\mathcal{H}$ induced by the Lorentzian metric.

V. CONSTRUCTION OF THE SPIN DERIVATIVE

We just introduced the Dirac operator on a manifold by combining the principle of general relativity with our $U(2,2)$ gauge symmetry. Although this definition was very natural, it is not clear what it precisely means: The Dirac operator is characterized by the matrices $G^j(x)$ and $B(x)$ in (20). We saw that $G^j(x)$ induces a Lorentzian structure and that $B(x)$ is composed of potentials that might be identified with usual gauge potentials. The precise physical interpretation of the degrees of freedom of $G^j(x)$, $B(x)$ is not clear, however. Furthermore, we expect in analogy to the discussion of Lorentzian transformations in Sec. III that coordinate and gauge transformations can be linked by imposing some conditions on the Dirac matrices. This would imply that the freedom in choosing the coordinate system is related to the $U(2,2)$ gauge symmetry. But this relation is also very vague at the moment.

In order to clarify the situation, we proceed in this section with some differential geometry. We will try to avoid abstract formalisms and prefer calculations in explicit coordinate systems and gauges. This is less elegant than a coordinate-free formulation, but it leads to a more elementary and direct approach. Our constructions are helpful for the physical interpretation, as will be explained in Sec. VI. In nontechnical terms, our aim is to rewrite the Dirac operator in a form which gives a better geometrical understanding of the matrices $G^j(x)$, $B(x)$.

Before starting the mathematical analysis, we point out that we consider the Dirac operator (apart from the wave functions of the physical particles) as the only *a priori* given object on the manifold. This means that all additional structure (like the metric, curvature, classical potentials, and field tensors) must be constructed from G. It might seem unusual to take only the Dirac operator as the starting point of a classical gauge theory. The motivation for this very restrictive procedure is the following: A physical theory is more convincing if it is developed from few objects, which are given from the very beginning and are considered as the ''fundamental'' objects of the theory. In relativistic quantum mechanics, the Dirac operator is needed in any case for the formulation of the Dirac equation. The objects of classical field theory, however, either enter directly in the Dirac operator or can be expressed in terms of derivatives of potentials occurring in G. We conclude that it is both desirable and might be possible to consider them as derived objects, which are only constructed from the Dirac operator for a convenient formulation of the classical interactions.

The Lorentzian metric (22) induces the Levi-Civita connection ∇, which gives a parallel transport of vector and tensor fields along geodesics. The geometry of space–time can be described as usual with the Riemannian curvature tensor R^i_{jkl}, and we could thus formulate general relativity by writing down Einstein's field equations. This procedure only tells about part of the geometry, however. We did not at all use the matrix $B(x)$ and did not take into account the $U(2,2)$ gauge symmetry. If we want to understand all the degrees of freedom of the Dirac operator geometrically, it is a better idea to study the parallel transport of spinors (instead of tensor fields). In the infinitesimal version, this parallel transport is given by a $U(2,2)$ gauge covariant derivative

D on the wave functions, which we call a *spin derivative*. It is characterized by the condition that the derivative $D_j \Psi$ of a wave function shall again behave under gauge transformations according to (24). The basic question is if the Dirac operator induces a spin derivative.

For the construction of the spin derivative, we will analyze gauge transformations explicitly. In a special gauge and coordinate system, D shall have the representation

$$D_j = \frac{\partial}{\partial x^j} - iC_j(x), \tag{28}$$

with suitable (4×4)-matrices $C_j(x)$. Under a gauge transformation (24), it transforms like

$$D_j \to U D_j U^{-1} = \partial_j - iUC_j U^{-1} + U(\partial_j U^{-1}). \tag{29}$$

Thus, our aim is to find matrices $C_j(x)$ satisfying the transformation rule,

$$C_j \to UC_j U^{-1} + iU(\partial_j U^{-1}). \tag{30}$$

These matrices must be formed out of $G^j(x)$, $B(x)$, which behave under gauge transformations like

$$G^j \to UG^j U^{-1}, \quad B \to UBU^{-1} + iUG^j(\partial_j U^{-1}). \tag{31}$$

It is far from being obvious how an explicit formula for the matrices C_j should look like. But we can already say something about its general structure: since only first derivatives of U occur in (30), we will only use G^j, B, and first derivatives of the Dirac matrices $\partial_k G^l$ for the construction. As a consequence, we need not care about second derivatives of U; they will not enter in our calculations. Constant gauge transformations (i.e., transformations U with $\partial_j U = 0$) are also irrelevant, because they only describe a common unitary transformation$\to U$. U^{-1} of all matrices. The point of interest is the first-order term $iU(\partial_j U^{-1})$ in (30).

We introduce some notation: In analogy to the bilinear and pseudoscalar covariants of the Dirac theory, we define the matrices

$$\sigma^{jk}(x) = \frac{i}{2}\,[G^j, G^k], \quad \rho(x) = \frac{i}{4!}\,\epsilon_{jklm} G^j G^k G^l G^m.$$

In a local reference frame (21), they coincide with the usual matrices σ^{jk}, γ^5 in the Dirac representation (we use the notation ρ instead of γ^5, because the ''tensor index'' [5] might be confusing on a manifold). A (4×4)-matrix is called *even* or *odd* if it commutes, resp. anticommutes, with ρ. The matrices

$$G^j, \quad \rho G^j, \quad \mathbb{1}, \quad i\rho, \quad \sigma^{jk}, \tag{32}$$

form a basis of the 16-dimensional (real) vector space of self-adjoint matrices (with respect to $<\cdot|\cdot>$). G^j, ρG^j are odd; $\mathbb{1}$, $i\rho$, and σ^{jk} are even.

Since we must expect the matrices C_j to be complicated expressions in G^j and B, it is useful to first consider special gauges where these expressions *should have* a simple form. For this purpose, we study the term $\nabla_k G^j = \partial_k G^j + \Gamma^j_{kl} G^l$. Under coordinate transformations, it behaves like a tensor. Under gauge transformations, however, first derivatives of U occur,

$$\begin{aligned} \nabla_k G^j \to \nabla_k (UG^j U^{-1}) &= U(\nabla_k G^j) U^{-1} + (\partial_k U) G^j U^{-1} + UG^j (\partial_k U^{-1}) \\ &= U(\nabla_k G^j) U^{-1} - [U(\partial_k U^{-1}), UG^j U^{-1}]. \end{aligned} \tag{33}$$

We can use the second summand in (33) for a partial gauge fixing.

Lemma V.1: For any space–time point $p \in M$ there is a gauge such that

$$\nabla_k G^j(p) = 0. \tag{34}$$

Proof: Notice that the matrix $\partial_j \rho$ is odd, because

$$0 = \partial_j \mathbb{1} = \partial_j(\rho\rho) = (\partial_j\rho)\rho + \rho(\partial_j\rho). \tag{35}$$

We start with an arbitrary gauge and construct the desired gauge with two subsequent gauge transformations.

(1) According to (35), the matrix $i\rho(\partial_j\rho)$ is self-adjoint. We can thus perform a gauge transformation U with $U(p)=\mathbb{1}$, $\partial_j U(p) = 1/2\rho(\partial_j\rho)$. The matrix $\partial_j\rho(p)$ vanishes in the new gauge, since

$$\partial_j\rho_{|p} \to \partial_j(U\rho U^{-1})_{|p} = \partial_j\rho_{|p} + \tfrac{1}{2}[\rho(\partial_j\rho),\rho]_{|p} = \partial_j\rho_{|p} - \rho^2(\partial_j\rho)_{|p} = 0.$$

By differentiating the relation $\{\rho, G^j\}=0$, we conclude that the matrix $\nabla_k G^j_{|p}$ is odd. We can thus represent it in the form

$$\nabla_k G^j_{|p} = \Lambda^j_{km} G^m_{|p} + \Theta^j_{km}\rho G^m, \tag{36}$$

with suitable coefficients Λ^j_{km}, Θ^j_{km}.

This representation can be further simplified: According to Ricci's Lemma, we have at p,

$$0 = 2\nabla_n g^{jk} = \{\nabla_n G^j, G^k\} + \{G^j, \nabla_n G^k\} = 2\Lambda^j_{nm} g^{mk} - \Theta^j_{nm} 2i\rho\sigma^{mk} + 2\Lambda^k_{nm} g^{mj} - \Theta^k_{nm} 2i\rho\sigma^{mj}, \tag{37}$$

and thus

$$\Lambda^j_{nm} g^{mk}_{|p} = -\Lambda^k_{nm} g^{mj}_{|p}. \tag{38}$$

In the case $j=k\neq m$, (37) yields that $\Theta^j_{nm}=0$. For $j\neq k$, we obtain $\Theta^j_{nj}\sigma^{jk} + \Theta^k_{nk}\sigma^{kj} = 0$ and thus $\Theta^j_{nj} = \Theta^k_{nk}$ (j and k denote fixed indices; no summation is performed). We conclude that there are coefficients Θ_k with

$$\Theta^j_{km} = \Theta_k \delta^j_m. \tag{39}$$

(2) We perform a gauge transformation U with $U(p)=\mathbb{1}$ and

$$\partial_k U = -\frac{1}{2}\,\Theta_k\rho - \frac{i}{4}\,\Lambda^m_{kn} g^{nl}\sigma_{ml}.$$

Using the representation (36) together with (38),(39), the matrix $\nabla_k G^j$ transforms like

$$\nabla_k G^j \to \nabla_k G^j + [\partial_k U, G^j] = \Lambda^j_{km} G^m + \Theta_k \rho G^j - \Theta_k \rho G^j - \frac{i}{4}\,\Lambda^m_{kn} g^{nl}[\sigma_{ml}, G^j]$$

$$= \Lambda^j_{km} G^m - \frac{i}{4}\,\Lambda^m_{kn} g^{nl} 2i(G_m \delta^j_l - G_l \delta^j_m)$$

$$= \Lambda^j_{km} G^m + \tfrac{1}{2}\Lambda^m_{kn} g^{nj} G_m - \tfrac{1}{2}\Lambda^j_{km} G^m = 0. \qquad \blacksquare$$

In general, the condition $\nabla_k G^j$ can only be satisfied ''locally'' in one point $p \in M$. We call a gauge satisfying (34) a *normal gauge* around p.

Now we look to which extent the gauge is fixed by Lemma V.1. According to (33), a transformation between normal gauges must satisfy the condition

$$[U(\partial_j U^{-1}), U G^j U^{-1}]_{|p} = 0.$$

As a consequence, the matrix $iU(\partial_j U^{-1})_{|p}$ must be a multiple of the identity, as is verified in the basis (32) using the commutation relations between the Dirac matrices. Since constant gauge transformations and higher than first-order derivatives of U are irrelevant, we can assume that $U(x)$ itself is a multiple of the identity. In other words, we can restrict ourselves to $U(1)$ gauge transformations. This is very helpful because in this special case there is a simple expression showing the transformation law (30), namely,

$$\tfrac{1}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)\mathbb{1} \to \tfrac{1}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)\mathbb{1} + \tfrac{1}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j G^k i U(\partial_j U^{-1})\mathbb{1}) = \tfrac{1}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)\mathbb{1} + iU(\partial_j U^{-1}).$$

We can identify this expression with C_j and use (28) as the definition for a spin derivative.

Definition V.2: We define the **canonical spin derivative** *D by the condition that it has in normal gauges around p the form*

$$D_j(p) = \frac{\partial}{\partial x^j} - \frac{i}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)_{|p}\mathbb{1}. \tag{40}$$

In general gauges, the canonical spin derivative can be written as

$$D_j = \frac{\partial}{\partial x^j} - iE_j - \frac{i}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)\mathbb{1}, \tag{41}$$

with additional matrices $E_j(x)$. These matrices are characterized by the condition that $E_j(p)$ vanishes in normal gauges around p and behaves according to (30) under $SU(2,2)$ gauge transformations. They are given by the formula

$$E_j = \frac{i}{2}\,\rho(\partial_j \rho) - \frac{i}{16}\,\mathrm{Tr}(G^m \nabla_j G^n) G_m G_n + \frac{i}{8}\,\mathrm{Tr}(\rho G_j \nabla_m G^m)\rho, \tag{42}$$

as can be verified by a straightforward calculation. Unfortunately, this expression is rather complicated. We will in the following always avoid working with E_j explicitly.

The canonical spin derivative has some nice properties. We first collect them in a theorem and discuss them afterward.

Theorem V.3: *The canonical spin derivative satisfies the equation*

$$[D_k, G^j] + \Gamma^j_{kl} G^l = 0. \tag{43}$$

It is compatible with the spin scalar product,

$$\partial_j \langle \Psi | \Phi \rangle = \langle D_j \Psi | \Phi \rangle + \langle \Psi | D_j \Phi \rangle. \tag{44}$$

The operator $iG^j D_j$ is a well-defined Hermitian operator on H. Furthermore,

$$-i\nabla_j \langle G^j \Psi | \Phi \rangle = \langle iG^j D_j \Psi | \Phi \rangle - \langle \Psi | iG^j D_j \Phi \rangle. \tag{45}$$

Proof: The left side of (43) behaves under gauge transformations according to the adjoint representation. $i \to U \cdot U^{-1}$ of the gauge group. Thus it suffices to check (43) in a normal gauge, where

$$[D_k, G^j] + \Gamma^j_{kl} G^l = \nabla_k G^j - \frac{i}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B)[\mathbb{1}, G^j] = 0.$$

For the canonical spin derivative, the matrix C_j in (28) is self-adjoint, as can be verified with (41), (42). This immediately implies (44).

According to its behavior under coordinate and gauge transformations, we can view $iG^j D_j$ as an operator on H. Relation (44) yields the equation

$$\nabla_j \langle G^j \Psi | \Phi \rangle = \langle D_j G^j \Psi | \Phi \rangle + \langle G^j \Psi | D_j \Phi \rangle + \Gamma^j_{jk} \langle G^k \Psi | \Phi \rangle.$$

Applying (43) and the self-adjointness of the Dirac matrices, we obtain (45). If we integrate this equation with respect to the invariant measure $d\mu = \sqrt{|g|}\,d^4x$, the left side vanishes with Gauss' theorem. On the right side, we substitute (25) and conclude that the operator $iG^j D_j$ is Hermitian. ∎

Generally speaking, this theorem shows that the connections ∇, D, the Dirac matrices G^j, and the scalar products $g_{jk}, \langle \cdot | \cdot \rangle$ can be consistently used in combination with each other, if all

expressions are written in a coordinate and gauge invariant form (by appropriately using ∇_j, D_j, and Γ^j_{kl}). We can leave out the derivatives of G^j and may use the product rule inside the spin scalar product. Relation (44) formally corresponds to Ricci's Lemma in Riemannian geometry if we replace the metric by the spin scalar product and ∇ by D.

Since the operators G and iG^jD_j are both Hermitian and coincide up to zeroth-order contributions, we can represent the Dirac operator in the form

$$G = iG^jD_j + H, \tag{46}$$

with a self-adjoint matrix $H(x)$. Under gauge transformations, H simply behaves according to the adjoint representation. This is an advantage over the formula (20), where the transformation of B involves derivatives of U.

Equation (45) implies a general version of current conservation: Assume that Ψ and Φ are eigenvectors of G with real eigenvalue m. Then the vector field $<G^j\Psi|\Phi>$ is divergence-free,

$$-i\nabla_j <G^j\Psi|\Phi> = <(G-H)\Psi|\Phi> - <\Psi|(G-H)\Phi> = 0.$$

It is a natural generalization of the electromagnetic current $\overline{\Psi}\gamma^j\Psi$ of the Dirac theory. According to Gauss' law, the current conservation implies that the scalar product (27) is independent of the choice of the hypersurface $\mathcal{H}$. The integrand of the scalar product $(\Psi|\Psi)$ has the interpretation as the probability density of the particle.

The Christoffel symbol in (43) indicates that the spin derivative in some way includes the Levi-Civita connection ∇. The following bundle construction makes this relation more precise: We denote the sections of a vector bundle N over M by $\Gamma(N)$ [e.g., $\Gamma(TM)$ are the vector fields on M]. The Levi-Civita connection is a connection on the tangent bundle TM,

$$\nabla : T_pM \times \Gamma(TM) \to T_pM : (X,Y) \to \nabla_X Y.$$

We can view the wave functions as sections of a vector bundle SM with fiber $(\mathbb{C}^4, <\cdot|\cdot>)$. In order to describe the $U(2,2)$ gauge symmetry, we make this *spin bundle* to a principal bundle with structure group $U(2,2)$, which acts on the fiber in the fundamental representation. $\to U$. . The spin derivative is a connection on the spin bundle,

$$D : T_pM \times \Gamma(SM) \to S_pM : (X,\Psi) \to D_X\Psi.$$

We define L_pM as the vector space of self-adjoint transformations of S_pM. In a special gauge, L_pM can be represented as (4×4)-matrices; (32) gives an explicit basis. The gauge transformations act on the matrices according to the adjoint representation. $\to U_{|p} . U_{|p}^{-1}$. We denote the corresponding principal bundle by LM. The spin derivative induces a connection on LM by

$$D : T_pM \times \Gamma(LM) \to L_pM : (X,A) \to [D_X, A]. \tag{47}$$

We can view the tangent bundle as a subspace of LM, because there are canonical embeddings and projections,

$$\iota_p : T_pM \hookrightarrow L_pM : X \to X^i G_i,$$

$$\pi_p : L_pM \to T_pM : A \to \frac{1}{4}\,\mathrm{Tr}(G^jA)\,\frac{\partial}{\partial x^j},$$

with $\pi\iota = id$. By restricting the connection (47) to the tangent bundle, we can form a connection $\tilde{\nabla}$ on TM,

$$\tilde{\nabla} : T_pM \times S(TM) \to T_pM : (X,Y) \to \pi[D_X, \iota Y].$$

Equation (43) states that this connection coincides with the Levi-Civita connection, since

$$[D_X, \iota Y] = X^j[D_j, G^kY_k] = X^j((\partial_j Y_k)G^k - \Gamma^k_{jl}Y_kG^l) = X^j(\nabla_j Y_k)G^k = \iota\,\nabla_X Y, \tag{48}$$

and thus

$$\tilde{\nabla}_X Y = \pi[D_X, \iota Y] = \pi \iota\, \nabla_X Y = \nabla_X Y.$$

In this way, the Levi-Civita connection is recovered as a suitable restriction of the spin connection. The geometry of the spin bundle generalizes the Lorentzian geometry of the manifold.

It remains to construct the geometrical invariants of the spin connection. On an elementary level, one can look for expressions in D_k, G^j, which transform both like a tensor and according to a local representation of the gauge group. This leads to the definition of the 2-forms,

$$T_{jk} = \frac{i}{2}\,([D_j, G_k] - [D_k, G_j]),$$

$$R_{jk} = \frac{i}{2}\,[D_j, D_k],$$

which are called *torsion* and *curvature*.

Theorem V.4: *The canonical spin connection is torsion-free. Curvature has the form*

$$R_{jk} = \tfrac{1}{8} R_{mnjk} \sigma^{mn} + \tfrac{1}{2}(\partial_j a_k - \partial_k a_j)\mathbb{1}, \tag{49}$$

with the Riemannian curvature tensor R_{mnjk} and the potential

$$a_j = \tfrac{1}{4}\,\mathrm{Re}\,\mathrm{Tr}(G_j B). \tag{50}$$

Proof: Relation (43) yields

$$[D_j, G_k] = [D_j, g_{kl} G^l] = (\partial_j g_{kl}) G^l - g_{kl} \Gamma^l_{jm} G^m = \Gamma^m_{jk} G_m,$$

and thus, using that the Levi-Civita connection is torsion-free,

$$T_{jk} = \frac{i}{2}\,(\Gamma^m_{jk} - \Gamma^m_{kj}) G_m = 0.$$

We iterate formula (48) and express the Riemannian curvature tensor in terms of the curvature of the spin connection,

$$\begin{aligned} G_i R^i_{jkl} Z^l = \iota(\nabla_j \nabla_k Z - \nabla_k \nabla_j Z) &= [D_j, [D_k, \iota Z]] - [D_k, [D_j, \iota Z]] \\ &= [[D_j, D_k], \iota Z] = -2i[R_{jk}, \iota Z]. \end{aligned}$$

This equation determines curvature up to a multiple of the identity matrix,

$$R_{jk}(x) = \tfrac{1}{8} R_{mnjk} \sigma^{mn} + \lambda_{jk}\mathbb{1}.$$

Thus, it remains to calculate the trace of curvature,

$$\tfrac{1}{4}\,\mathrm{Tr}(R_{jk}) = \tfrac{1}{8}\mathrm{Tr}(\partial_j C_k - \partial_k C_j) = \tfrac{1}{2}(\partial_j a_k - \partial_k a_j),$$

where we used (41) and the fact that the matrices E_j are trace-free. ∎

We remark that the canonical spin derivative is not the only spin derivative, which can be constructed from the Dirac operator. A general spin derivative $\overline{D}$ differs from the canonical spin derivative by matrices $F_j = D_j - D_j$, which transform according to the adjoint representation of the gauge group. It is no loss of generality to work with the canonical spin derivative because the matrices F_j can also be dealt as additional potentials. Because of the useful properties of Theorem V.4, it is most convenient to use the canonical spin derivative.

VI. INTERPRETATION, THE PHYSICAL DIRAC OPERATOR

For the physical interpretation of the previous constructions, we return to the discussion of the moving observer on page 9. We saw that the coordinate system and gauge satisfying (21) give a local reference frame for the observer. The gravitational field can be compensated locally by choosing a normal coordinate system. According to Lemma V.1, we can further compensate the dynamics of the Dirac matrices: in a normal gauge even the derivatives of the Dirac matrices vanish, $\partial_k G^j(p) = \nabla_k G^j(p) = 0$. This means that the above-mentioned ''spin-mixing force'' can also be compensated locally by choosing an appropriate coordinate system and gauge. We cannot do even better because second-order derivatives of the metric and the Dirac matrices involve the Riemannian curvature tensor. We call a coordinate system and gauge satisfying the conditions $G^j(p) = \gamma^j$, $\partial_k G^j(p) = 0$ a *normal reference frame* around p.

We come to the discussion of the matrix $B(x)$, which was not at all specified in Definition IV.1. The strongest local condition is to impose that $B(p)$ vanishes in a suitable normal reference frame around p. This condition turns out to be of physical interest, and we use it as the definition for the physical Dirac operator:

Definition VI.1: A Dirac operator G is called a **physical Dirac operator** *if for any $p \in M$ there is a normal reference frame around p such that $B(p) = 0$.*

In other words, the physical Dirac operator is characterized by the condition that it locally coincides with the free Dirac operator, which means more precisely that there is a coordinate system and a gauge with $G^j(p) = \gamma^j$, $\partial_k G^j(p) = 0$, and $B(p) = 0$.

We again consider the physical Dirac operator as the only *a priori* given object on the manifold and construct the metric and the canonical spin derivative from G. In a normal reference frame with $G = iG^j \partial_j$, the canonical spin derivative (40) coincides with the partial derivatives, $D_j = \partial_j$. Thus, the physical Dirac operator has the representation

$$G = iG^j D_j, \tag{51}$$

which is also valid in a general gauge and coordinate system. In comparison to the representation (20), the degrees of freedom of the matrix $B(x)$ now occur in the canonical spin derivative. This has the advantage that they can be immediately identified with the $U(2,2)$ gauge potentials in (41). The last summand in (41) describes the $U(1)$ phase transformations of electrodynamics. We define the electromagnetic potential by

$$A_j = \frac{1}{4e} \text{ Re Tr}(G_j B),$$

and rewrite the canonical spin derivative in the more familiar form

$$D_j = \frac{\partial}{\partial x^j} - iE_j - ieA_j.$$

The matrices E_j take into account the gravitational field. According to Theorem V.4, the curvature of the spin connection is formed of the Riemannian curvature tensor and the electromagnetic field tensor $F_{jk} = \partial_j A_k - \partial_k A_j$. We can write down the classical action in terms of these tensor fields. The classical variational principle yields a unified description of general relativity and electromagnetism as a $U(2,2)$ gauge symmetry.

The definition of the physical Dirac operator can also be understood in more technical terms with the canonical spin derivative: The Dirac operator of Definition IV.1 has the representation (46) with a self-adjoint matrix $H(x)$. According to the construction of the spin derivative, the matrices $G^j H$ are trace-free. Thus $H(x)$ is characterized by 12 degrees of freedom, which correspond to additional potentials (more precisely, one scalar, one pseudoscalar, four pseudovector, and six bilinear potentials). These potentials are not gauge potentials. They do not seem to occur in nature. Therefore it is a reasonable physical condition to assume that H vanishes. This leads to the representation (51) of the physical Dirac operator.

Our definition of the Dirac operator introduces the gauge potentials and in this way replaces the usual minimal coupling procedure. We explain for clarity why simple minimal coupling does

not make sense in the context of our $U(2,2)$ symmetry: Assume that we had (instead of constructing the spin derivative from the Dirac operator) introduced a $U(2,2)$ gauge covariant derivative D of the form (28) with gauge potentials C_j. According to minimal coupling, we must replace the partial derivatives in the free Dirac operator by gauge covariant derivatives, which gives the operator $i\gamma^j D_j$. This operator does not make sense, however. First of all, it is not Hermitian (with respect to $<\cdot|\cdot>$). Furthermore, the current $<\gamma^j\Psi|\Psi>$ is, in general, not divergence-free, so that the probabilistic interpretation of the Dirac wave function breaks down. The basic reason why minimal coupling does not work is that the $U(2,2)$ gauge potentials C_j do, in general, not commute with the Dirac matrices γ^j. In order to bypass these problems, we must replace the γ-matrices by dynamical Dirac matrices $G^j(x)$. The two replacements $\partial_j \rightarrow D_j$ and $\gamma^j \rightarrow G^j$ must be coordinated in such a way that the Dirac operator iG^jD_j is Hermitian and allows a reasonable definition of a conserved current. This is accomplished by our definition of the Dirac operator and the subsequent construction of the spin derivative.

Our description of gravitation differs considerably from the usual formulation of general relativity as a gauge theory (see, e.g., Ref. 4). As one of the major differences, we avoid the principal bundle of orthonormal frames (with the Lorentz group as structure group); instead, we are working with the spin bundle and local $U(2,2)$ transformations. This replacement is possible because the subgroup of $U(2,2)$ generated by the bilinear covariants σ^{ij} is locally isomorphic to the Lorentz group. In contrast to the connection on the bundle of orthonormal frames, the canonical spin connection is torsion-free. Remember that the $U(2,2)$ spin bundle arose as a consequence of our measurement principle for space and time. Thus, as an advantage of our procedure, the usual bundle constructions for the definition of gauge fields and of spinors in curved space–time are no longer necessary.

VII. CONCLUSION

We saw that the adaptation of the ideas in Ref. 1 to the relativistic context yields a local $U(2,2)$ gauge symmetry of the Dirac equation. In order to describe the physical interactions with this gauge symmetry, it is necessary to consider the Dirac operator as the basic object on the manifold. The gauge potentials are implicitly contained in the Dirac operator. By constructing the spin derivative, they are recovered as describing the electromagnetic and gravitational field. In this way, we conclude that the local $U(2,2)$ symmetry in relativistic quantum mechanics makes physical sense. The concept of measurability of space–time gives a fundamental explanation for this gauge symmetry.

Our description has the advantage that both the Dirac theory and classical field theory are developed from few *a priori* given objects: The fermionic particles correspond to vectors of the indefinite scalar product space $(H,\langle\cdot|\cdot\rangle)$. Space–time is described by the spectral measure dE_x on the manifold M. The Dirac operator gives the gauge potentials and determines the interaction between the fermions and the gauge fields. This description is conceptually simple. It is the starting point for further constructions which finally lead to the ''Principle of the Fermionic Projector'' as introduced in Ref. 5.

ACKNOWLEDGMENT

Supported by the Deutsche Forschungsgemeinschaft, Bonn.

[1] F. Finster, ''Derivation of local gauge freedom from a measurement principle,'' preprint funct-an/9701002, 1997.
[2] F. Finster, Diplomarbeit mathematik, Universität Heidelberg, 1992.
[3] J. Bjorken and S. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964).
[4] M. Göckeler and T. Schücker, *Differential Geometry, Gauge Theories and Gravity* (Cambridge University Press, Cambridge, 1987).
[5] F. Finster, ''Derivation of field equations from the principle of the fermionic projector,'' preprint gr-qc/9606040, 1996.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

Stochastically positive structures on Weyl algebras. The case of quasi-free states

R. Gielerak,[a] L. Jakóbczyk,[b] and R. Olkiewicz[c]
Institute of Theoretical Physics, University of Wroclaw, Wroclaw, Poland

(Received 20 May 1997; accepted for publication 21 August 1998)

We consider quasi-free stochastically positive ground and thermal states on Weyl algebras in the imaginary time formulation. In particular, we obtain a new derivation of a general form of thermal quasi-free state and give conditions when such a state is stochastically positive, i.e., when it defines a periodic stochastic process with respect to imaginary time, a so-called thermal process. Then we show that the thermal process completely determines modular structure canonically associated with the quasi-free thermal state on Weyl algebra. We discuss a variety of examples connected with free quantum field theories on globally hyperbolic stationary space–times and models of quantum statistical mechanics. © *1998 American Institute of Physics.* [S0022-2488(98)01712-5]

I. INTRODUCTION

It is well known that the idea of analytic continuation to the imaginary time (Euclidean) variables is very fruitful for studying several quantum systems, especially Wightman quantum field theories on a flat space–time. From the general properties of Wightman distributions follows the possibility of analytic continuation to the Euclidean region, which gives an alternative, purely Euclidean description in terms of so-called Schwinger functions.[1–4] In the context of bosonic fields, we obtain commutative objects, which are much easier to analyze.[5] A particulary nice situation occurs in the case of free scalar fields. As was noticed by Nelson[6] (see also Ref. 7), the corresponding Euclidean structure is given by a Gaussian Markov generalized random field. This observation and the general reconstruction theorem,[8] gave powerful input for the development of the constructive quantum field theory on a flat space–time (Refs. 9 and 10 and references therein).

One of the main objectives of the present paper is to develop systematically the rigorous imaginary time formalism in the context of equilibrium states given by the Kubo–Martin–Schwinger (KMS) condition. As an interesting application of general results, we consider imaginary time approach to quantum field theories at nonzero temperature on globally hyperbolic stationary space–times.[11,12] Some aspects of imaginary time formalism in this context were discussed in the literature at various levels of mathematical rigor (see, e.g., Refs. 13–16), but, to our knowledge, there is still the lack of a systematic and mathematically rigorous approach to these problems. In the context of quantum statistical mechanics, the imaginary time approach developed in Refs. 17–27 also appears to be very useful. In particular, for a class of quantum systems with so-called stochastic positivity, which are described by some stochastic processes, methods of classical statistical mechanics can be applied to study several questions like the existence and properties of the limiting thermal states, phase transitions, etc.[20,21,24–27]

In the present paper a general conceptual framework for stochastically positive structures on Weyl algebras is introduced (Sec. II). As a starting point for further development, we analyze in detail the case of quasi-free states (Sec. III). Although basic structural results characterizing quasi-free thermal states were obtained a long time ago,[28–30] we present a new (in our opinion, simpler) proof of the theorem giving the general form of a quasi-free state satisfying the KMS condition (see Theorems 3.1 and 3.7). Section III D contains complete characterizations of stochastically positive quasi-free KMS states and corresponding periodic stochastic processes. The

[a]Electronic mail: gielerak@ift.uni.wroc.pl
[b]Electronic mail: ljak@ift.uni.wroc.pl
[c]Electronic mail: rolek@ift.uni.wroc.pl

 © 1998 American Institute of Physics

important problem of the equivalence of the modular structure given by a stochastic process (Theorem 3.3) with the canonical modular structure associated with a quasi-free KMS state is solved in full generality in Sec. III E. This result shows that in the case of quasi-free stochastically positive states, all relevant information about KMS structure is contained in the commutative sector given by a thermal stochastic process. Similar results are true also for the ground state case (Sec. III G). Let us emphasize that it gives new possibilities for the description of KMS structures in the case when interaction is present. Having described quasi-free systems in terms of stochastic processes, one may perturb them with multiplicativelike functionals, thereby creating some new non-Gaussian thermal process. Furthermore, given such a process, we can reproduce its KMS structure. Some results in this direction were obtained in Ref. 24 for gentle perturbations of the free Bose gas in the noncritical region of densities and in Ref. 25 for the critical case. Essentially equivalent (via the Feynman–Kac formula) types of perturbations which can be studied using this method are central perturbations of a quasi-free state, by which we mean the perturbations of the form ''$H_\omega + V$,'' where H_ω is the generator of the free evolution in the thermal representation, and V is an operator affiliated with the von Neumann algebra $\pi_\omega(\mathfrak{U})''$, for an appropriate Abelian subalgebra $\mathfrak{U}$ of the Weyl algebra. That kind of perturbation will be discussed in the second part of the present work. Section IV contains several examples to which our general arguments can be applied. In the particular case of the scalar quantum field theory on a globally hyperbolic stationary space–time, we give a general result on the existence of Markov thermal process, which determines the whole modular structure of the theory. Similar arguments are also valid for ground state structure. In the case of a static space–time, we obtain much more explicit description of the arising process and we are able to discuss its continuity properties. In particular, in the case of the Rindler wedge (Example 4.1.2), we show stochastic positivity of the KMS structure arising in the context of the Bisognano–Wichmann theorem.[31] These results can be used to develop a rigorous constructive approach to nonlinear quantum field theories by adopting the tools worked out in constructive quantum field theory on a flat space–time. A detailed study of the perturbed Euclidean quantum fields on stationary globally hyperbolic space–times will be presented in our forthcoming publication. In this section we discuss the models of quantum statistical mechanics, including the model of nonrelativistic Bose matter and infinite harmonic crystal, and describe the corresponding stochastic processes.

II. WEYL ALGEBRA. VACUUM AND TEMPERATURE GREEN FUNCTIONS

A. Weyl algebra

Let $\mathcal{D}$ be a real vector space with a locally convex topology τ. If σ is a τ-continuous symplectic form on $\mathcal{D}$ (i.e., σ is a bilinear, antisymmetric and nondegenerate mapping from $\mathcal{D}\times\mathcal{D}$ into $\mathbb{R}$), then $((\mathcal{D},\tau),\sigma)$ is called a vector symplectic space. Let W_f be the real function on $\mathcal{D}$ defined by

$$W_f(g)=\begin{cases}1, & \text{if } f=g,\\ 0, & \text{if } f\neq g.\end{cases}$$

With the product

$$W_f W_g = e^{-i\sigma(f,g)/2} W_{f+g}$$

and involution

$$W_f^* = W_{-f},$$

the complex algebra generated by W_f, $f\in\mathcal{D}$ becomes a*-algebra $\mathfrak{W}_0(\mathcal{D},\sigma)$. We define $\mathfrak{W}(\mathcal{D},\sigma)$ [the Weyl algebra over $((\mathcal{D},\sigma))$] in the following way. Let

$$\left\| \sum_{k=1}^{N} z_k W_{f_k} \right\|_1 = \sum_{k=1}^{N} |z_k|.$$

Here $\|\cdot\|_1$ is a $*$-norm on $\mathfrak{W}_0(\mathcal{D},\sigma)$ and let the completion $\overline{\mathfrak{W}(\mathcal{D},\sigma)}^{\|\cdot\|_1}$ be a $*$-Banach algebra with unit. Let $\mathcal{F}$ be the set of states on $\overline{\mathfrak{W}(\mathcal{D},\sigma)}^{\|\cdot\|_1}$. Then we define a C^*-algebra norm as follows:[32]

$$\|W\| = \sup_{\rho \in \mathcal{F}} \sqrt{\rho(W^*W)},$$

then $\mathfrak{W}(\mathcal{D},\sigma)$ is the completion of $\overline{\mathfrak{W}_0(\mathcal{D},\sigma)}^{\|\cdot\|_1}$ with respect to this norm.

A state ω on $\mathfrak{W}(\mathcal{D},\sigma)$ is called regular if it is τ-continuous, i.e., the function

$$f \to \omega(W_f)$$

is τ-continuous.

Remark: Usually, the weaker form of continuity is assumed, namely the map

$$\mathbb{R} \ni t \to \omega(W_{tf})$$

should be continuous, for every $f \in \mathcal{D}$.

The set of all regular states on $\mathfrak{W}(\mathcal{D},\sigma)$ will be denoted by $E(\mathfrak{W})$. For a given $\omega \in E(\mathfrak{W})$, $(\mathfrak{h}_\omega, \pi_\omega, \Omega_\omega)$ will be the corresponding Gel'fand–Nalmark–Segal (GNS) representation.

Let $\{T_t\}_{t \in \mathbb{R}}$ be a one-parameter group of τ-continuous symplectic mappings from $\mathcal{D}$ onto $\mathcal{D}$. In the following, $(\mathcal{D},\sigma,\{T_t\})$ will be called a linear dynamical system. If we put

$$\alpha_t(W_f) = W_{T_t f},$$

we obtain a one-parameter group of *-automorphisms of the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$. For a given linear dynamical system $(\mathcal{D},\sigma,\{T_t\})$, let $E^\alpha(\mathfrak{W})$ be the set of all regular α_t-invariant states on $\mathfrak{W}(\mathcal{D},\sigma)$. If $\omega \in E^\alpha(\mathfrak{W})$, then on the GNS space $\mathfrak{h}_\omega$, the evolution α_t is represented by a unitary group $U_\omega(t)$ with a self-adjoint generator H_ω.

Definition 2.1:

(1) $\omega \in E^\alpha(\mathfrak{W})$ is *a ground state* iff $H_\omega \geqslant 0$.

(2) Let $0 < \beta < \infty$ be given. A faithful state $\omega \in E^\alpha(\mathfrak{W})$ is *an α_t-KMS state at the inverse temperature β* iff in the corresponding GNS representation the unitary group e^{itH_ω} defines a weakly continuous group of *-automorphisms of $\pi_\omega(\mathfrak{W}(\mathcal{D},\sigma))''$ such that

$$\langle e^{-(\beta/2)H_\omega} \pi_\omega(W_f)\Omega_\omega, e^{-(\beta/2)H_\omega} \pi_\omega(W_g)\Omega_\omega \rangle = \langle \pi_\omega(W_g)^*\Omega_\omega, \pi_\omega(W_f)^*\Omega_\omega \rangle$$

for all $f,g \in \mathcal{D}$. In this case Ω_ω is cyclic and separating for $\pi_\omega(\mathfrak{W}(\mathcal{D},\sigma))''$ and the modular operator associated with Ω_ω coincides with $e^{-(\beta/2)H_\omega}$.

B. Imaginary time Green's functions

For any $\omega \in E^\alpha(\mathfrak{W})$ define *multi-time Green's functions*

$$\begin{aligned} G(f_1,\ldots,f_n;t_1,\ldots,t_n) &= \omega(\alpha_{t_1}(W_{f_1}) \cdots \alpha_{t_n}(W_{f_n})) \\ &= \langle \Omega_\omega, U_\omega(t_1)\pi_\omega(W_{f_1})U_\omega(-t_1) \\ &\quad \cdots U_\omega(t_n)\pi_\omega(W_{f_n})U_\omega(-t_n)\Omega_\omega \rangle \end{aligned}$$

for $t_1,\ldots,t_n \in \mathbb{R}$. If $\omega_\infty \in E^\alpha(\mathfrak{W})$ is a ground state, then it can be shown that the corresponding Green's functions $G_\infty(f_1,\ldots,f_n;t_1,\ldots,t_n)$ can be analytically continued to the functions denoted by $G_\infty(f_1,\ldots,f_n;z_1,\ldots,z_n)$, holomorphic in the region

$$\mathfrak{T}_n^\infty = \{(z_1,\ldots,z_n) \in \mathbb{C}^n : -\infty < \mathrm{Im}\ z_1 < \mathrm{Im}\ z_2 < \ldots < \mathrm{Im}\ z_n < \infty\}$$

and continuous on the boundary of $\overline{\mathfrak{T}_n^{\infty}}$. Similarly, if $\omega_\beta \in E^\alpha(\mathfrak{W})$ is a strongly β-KMS state, then by the theorem of Araki,[33] the corresponding Green's functions $G_\beta(f_1,\ldots,f_n;t_1,\ldots,t_n)$ can be analytically continued to the functions denoted by $G_\beta(f_1,\ldots,f_n;z_1,\ldots,z_n)$, holomorphic in the region

$$\mathfrak{T}_n^\beta=\{(z_1,\ldots,z_n)\in\mathbb{C}^n: -\beta/2<\mathrm{Im}\ z_1<\cdots<\mathrm{Im}\ z_n<\beta/2\}$$

and continuous on the boundary of $\overline{\mathfrak{T}_n^\beta}$. Let

$$\mathfrak{E}_n^\infty=\{(s_1,\ldots,s_n)\in\mathbb{R}^n: s_1\leqslant\cdots\leqslant s_n\}$$

and

$$\mathfrak{E}_n^\beta=\{(s_1,\ldots,s_n): -\beta/2\leqslant s_1\leqslant\cdots\leqslant s_n\leqslant\beta/2\}.$$

Definition 2.2:

(1) The restrictions of $G_\infty(f_1,\ldots,f_n;z_1,\ldots,z_n)$ to the purely imaginary points

$$(is_1,\ldots,is_n),\quad (s_1,\ldots,s_n)\in\mathfrak{E}_n^\infty$$

will be called *imaginary time* (*or Euclidean*) Green's functions corresponding to the ground state ω_∞, and will be denoted by $G_\infty^E(f_1,\ldots,f_n;s_1,\ldots,s_n)$.

(2) Similarly, the restrictions of $G_\beta(f_1,\ldots,f_n;z_1,\ldots,z_n)$ to the points

$$(is_1,\ldots,is_n),\quad (s_1,\ldots,s_n)\in\mathfrak{E}_n^\beta$$

will be called *imaginary time* (*or Euclidean*) Green's functions corresponding to the KMS state ω_β, and will be denoted by $G_\beta^E(f_1,\ldots,f_n;s_1,\ldots,s_n)$.

Remark: By linearity and continuity, we extend $G_{\infty(\beta)}^E(f_1,\ldots,f_n;s_1,\ldots,s_n)$ to the whole Weyl algebra and obtain

$$G_{\infty(\beta)}^E(W_1,\ldots,W_n;s_1,\ldots,s_n),\quad W_1,\ldots,W_n\in\mathfrak{W}(\mathcal{D},\sigma).$$

The properties of imaginary time Green's functions corresponding to the KMS state ω_β were studied in Ref. 23. Similarly we can obtain the properties of ground state imaginary time Green's functions. In the formulation of these properties, the following notation will be used: for $W_1,\ldots,W_k\in\mathfrak{W}(\mathcal{D},\sigma)$,

$$\underline{W}^k=(W_1,\ldots,W_k)\quad\text{and}\quad \underline{W}^{k*}=(W_k^*,\ldots,W_1^*)$$

for $f_1,\ldots,f_k\in\mathcal{D}$, $\underline{f}^k=(f_1,\ldots,f_k)$. Similarly

$$\underline{s}^k=(s_1,\ldots,s_k)\quad\text{and}\quad \underline{s}^{k*}=(-s_k,\ldots,-s_1).$$

Moreover,

$$G_{\infty(\beta)}^E(\underline{f}^k;\underline{s}^k)=G_{\infty(\beta)}^E(f_1,\ldots,f_k;s_1,\ldots,s_k),$$

$$G_{\infty(\beta)}^E(\underline{W}^k;\underline{s}^k)=G_{\infty(\beta)}^E(W_1,\ldots,W_k;s_1,\ldots,s_k),$$

and

$$\underline{W}^{k-1}(i)=(W_1,\ldots,W_{i-1},W_iW_{i+1},W_{i+2},\ldots,W_k),$$

$$\underline{W}^{k-1}([i])=(W_1,\ldots,W_{i-1},W_{i+1},\ldots,W_k),$$

$$\underline{s}^{k-1}([i])=(s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_k).$$

For the ground state $\omega_\infty \in E^\alpha(\mathfrak{W})$ the set of imaginary time Green's functions $\mathbb{G}^E_\infty = \{G^E_\infty(\underline{W}^k;\underline{s}^k)\}$ has the following properties.

(EG1)$_\infty$: (1) For each $\underline{W}^k$ the map

$$\underline{s}^k \to G^E_\infty(\underline{W}^k;\underline{s}^k)$$

is continuous.

(2) For a fixed $\underline{s}^k \in \mathfrak{E}^\infty_k$, the map

$$\underline{W}^k \to G^E_\infty(\underline{W}^k;\underline{s}^k)$$

is multilinear and bounded, i.e.,

$$|G^E_\infty(\underline{W}^k;\underline{s}^k)| \leqslant \prod_{i=1}^{k} \|W_i\|.$$

Moreover, the map

$$\underline{f}^k \to G^E_\infty(\underline{f}^k;\underline{s}^k)$$

is τ-continuous.

(3) For any $s \in [0,\infty)$

$$G^E_\infty(\underline{W}^k;\underline{s}^k + s) = G^E_\infty(\underline{W}^k;\underline{s}^k).$$

(4) For any $\underline{W}^k$ and $\underline{s}^k \in \mathfrak{E}^\infty_k$ such that $s_i = s_{i+1}$, $1 \leqslant i \leqslant k-1$,

$$G^E_\infty(\underline{W}^k;\underline{s}^k) = G^E_\infty(\underline{W}^{k-1}(i),\underline{s}^{k-1}([i+1])).$$

(5) For any $\underline{W}^k$ such that $W_i = \mathbb{I}$: $1 \leqslant i \leqslant k$,

$$G^E_\infty(\underline{W}^k;\underline{s}^k) = G^E_\infty(\underline{W}^{k-1}([i]);\underline{s}^{k-1}([i])).$$

(6)

$$G^E_\infty(\mathbb{I};0) = 1.$$

(EG2)$_\infty$ (OS-positivity): For every terminating sequences

$$(\underline{W}^0,\underline{W}^1,\ldots,\underline{W}^k,\ldots), \quad (\underline{s}^0,\underline{s}^1,\ldots,\underline{s}^k,\ldots), \quad \underline{s}^k \in \mathfrak{E}^{\infty,+}_k = \{\underline{s}^k \in \mathfrak{E}^\infty_k : s_1 \geqslant 0\}.$$

we have

$$\sum_{k,l} G^E_\infty(\underline{W}^{k*},\underline{W}^l;\underline{s}^{k*},\underline{s}^l) \geqslant 0.$$

(EG3)$_\infty$: For every terminating sequences

$$(\underline{W}^0,\underline{W}^1,\ldots,\underline{W}^k,\ldots), \quad (\underline{s}^0,\underline{s}^1,\ldots,\underline{s}^k,\ldots), \quad \underline{s}^k \in \mathfrak{E}^{\infty,+}_k = \{\underline{s}^k \in \mathfrak{E}^\infty_k : s_1 \geqslant 0\}$$

and every $W \in \mathfrak{W}(\mathcal{D},\sigma)$ we have

$$\sum_{k,l} G^E_\infty(\underline{W}^{k*},W^*,W,\underline{W}^l\underline{s}^{k*},0,0\underline{s}^l) \leqslant \|W\|^2 \sum_{k,l} G^E_\infty(\underline{W}^{k*},\underline{W}^l;\underline{s}^{k*},\underline{s}^l).$$

Similarly, for the β-KMS state ω_β the set $\mathbb{G}^E_\beta = \{G^E_\beta(\underline{W}^k;\underline{s}^k)\}$ of imaginary time Green's functions has the following properties.

(EG1)$_\beta$: (1) For each $\underline{W}^k$ the map

$$\underline{s}^k \to G^E_\beta(\underline{W}^k;\underline{s}^k)$$

is continuous.

(2) For a fixed $\underline{s}^k \in \mathfrak{E}^\beta_k$ the map

$$\underline{W}^k \to G^E_\beta(\underline{W}^k;\underline{s}^k)$$

is multilinear, and bounded

$$|G^E_\beta(\underline{W}^k;\underline{s}^k)| \leqslant \prod_{i=1}^{k} \|W_i\|.$$

Moreover, the map

$$\underline{f}^k \to G^E_\beta(\underline{f}^k;\underline{s}^k)$$

is τ-continuous,

(3) For any $\underline{s}^k \in \mathfrak{E}^\beta_k$ and $s \in [-\beta/2,\beta/2]$ such that $s_k + s \leqslant \beta/2$, the imaginary time Green's functions are locally shift invariant, i.e.,

$$G^E_\beta(\underline{W}^k;\underline{s}^k + s) = G^E_\beta(\underline{W}^k;\underline{s}^k).$$

(4) For any $\underline{W}^k$ and $\underline{s}^k \in \mathfrak{E}^\beta_k$ such that $s_i = s_{i+1}$, $1 \leqslant i \leqslant k-1$,

$$G^E_\beta(\underline{W}^k;\underline{s}^k) = G^E_\beta(\underline{W}^{k-1}(i);\underline{s}^{k-1}([i+1])).$$

(5) For any $\underline{W}^k$ such that $W_i = 1$: $1 \leqslant i \leqslant k$,

$$G^E_\beta(\underline{W}^k;\underline{s}^k) = G^E_\beta(\underline{W}^{k-1}([i]);\underline{s}^{k-1}([i])).$$

(6)

$$G^E_\beta(\mathbb{I};0) = 1.$$

(EG2)$_\beta$ [Osterwalder–Schrader (OS-positivity)]: For every terminating sequences

$$(\underline{W}^0,\underline{W}^1,\ldots,\underline{W}^k,\ldots), \quad (\underline{s}^0,\underline{s}^1,\ldots,\underline{s}^k,\ldots), \quad \underline{s}^k \in \mathfrak{E}^{\beta,+}_k = \{\underline{s}^k \in \mathfrak{E}^\beta_k : s_1 \geqslant 0\}$$

we have

$$\sum_{k,l} G^E_\beta(\underline{W}^{k*},\underline{W}^l;\underline{s}^{k*},\underline{s}^l) \geqslant 0.$$

(EG3)$_\beta$: For every terminating sequences

$$(\underline{W}^0,\underline{W}^1,\ldots,\underline{W}^k,\ldots), \quad (\underline{s}^0,\underline{s}^1,\ldots,\underline{s}^k,\ldots), \quad \underline{s}^k \in \mathfrak{E}^{\beta,+}_k = \{\underline{s}^k \in \mathfrak{E}^\beta_k : s_1 \geqslant 0\}$$

and every $W \in \mathfrak{W}(\mathcal{D},\sigma)$ we have

$$\sum_{k,l} G^E_\beta(\underline{W}^{k*},W^*,W,\underline{W}^l;\underline{s}^{k*},0,0,\underline{s}^l) \leqslant \|W\|^2 \sum_{k,l} G^E_\beta(\underline{W}^{k*},\underline{W}^l;\underline{s}^{k*},\underline{s}^l).$$

(EG4)$_\beta$ (weak form of KMS condition): Let us define for $0 \leqslant s_1 \leqslant \cdots \leqslant s_n \leqslant \beta$

$$\hat{G}^E_\beta(W_0,\ldots,W_n;s_1,\ldots,s_n) := G^E_\beta(W_0,\ldots,W_n;-\beta/2,s_1-\beta/2,\ldots,s_n-\beta/2).$$

Then for each n

$$\hat{G}^E_\beta(W_0,\ldots,W_n;s_1,\ldots,s_n) = \hat{G}^E_\beta(W_n,W_0,W_1,\ldots,W_{n-1};\beta-s_n,\beta-s_n+s_1,\ldots,\beta-s_n+s_{n-1}).$$

On the other hand, starting with imaginary time Green's functions and proceeding similarly as in Ref. 23 (see also Ref. 18), we can obtain the following reconstruction theorems. Thus for ground state case we have the following.

Theorem 2.1: *Let $\mathbb{G}^E_\infty$ be an abstract set of imaginary time Green's functions on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ with properties* $(\mathrm{EG1})_\infty - (\mathrm{EG3})_\infty$. *Then there exist*

(1) *the Hilbert space $\mathfrak{H}_\infty$,*
(2) *the vector $\Omega_\infty \in \mathfrak{H}_\infty$,*
(3) *the*-representation π_∞ of $\mathfrak{W}(\mathcal{D},\sigma)$ in $\mathbb{B}(\mathfrak{H}_\infty)$, and*
(4) *the weakly continuous one-parameter group of unitary operators $U_\infty(t)=e^{itH_\infty}$ with $H_\infty \geqslant 0$*

such that

$$G^E_\infty(f^k;\underline{s}^k)=\langle \Omega_\infty, e^{-s_1 H_\infty}\pi_\infty(W_{f_1})\cdots e^{-(s_k-s_{k-1})H_\infty}\pi_\infty(W_{f_k})\Omega_\infty\rangle.$$

Moreover, the vector Ω_∞ is cyclic for the algebra $\mathfrak{M}^\infty_0$ generated by

$$e^{it_1 H_\infty}\pi_\infty(W_{f_1})e^{-it_1 H_\infty}\cdots e^{it_n H_\infty}\pi_\infty(W_{f_n})e^{-it_n H_\infty}, \quad t_1,\ldots,t_n\in\mathbb{R},$$

and the linear space spanned by vectors

$$e^{-s_1 H_\infty}\pi_\infty(W_{f_1})\cdots e^{-(s_k-s_{k-1})H_\infty}\pi_\infty(W_{f_k})\Omega_\infty$$

is dense in $\mathfrak{H}_\infty$. Additionally, $\alpha^\infty_t(W)=e^{itH_\infty}We^{-itH_\infty}$ is σ-weakly continuous group of automorphisms of $\mathfrak{M}^\infty=(\mathfrak{M}^\infty_0)''$.

And similarly, for temperature case, we have the following.

Theorem 2.2: *Let $\mathbb{G}^E_\beta$ be an abstract set of imaginary time Green's functions on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ with properties* $(\mathrm{EG1})_\beta - (\mathrm{EG4})_\beta$. *Then there exist*

(1) *the Hilbert space $\mathfrak{H}_\beta$,*
(2) *the vector $\Omega_\beta \in \mathfrak{H}_\beta$,*
(3) *the *-representation π_β of $\mathfrak{W}(\mathcal{D},\sigma)$ in $\mathbb{B}(\mathfrak{H}_\beta)$, and*
(4) *the weakly continuous one-parameter group of unitary operators $U_\beta(t)=e^{itH_\beta}$*

such that if $-\beta/2\leqslant s_1\leqslant s_2\leqslant\cdots\leqslant s_j\leqslant 0\leqslant s_{j+1}\leqslant\cdots\leqslant s_k\leqslant\beta/2$, then

$$G^E_\beta(f^k;\underline{s}^k)=\langle e^{s_j H_\beta}\pi_\beta(W_{f_j})^* e^{-(s_j-s_{j-1})H_\beta}\pi_\beta(W_{f_{j-1}})^*\cdots e^{-(s_2-s_1)H_\beta}\pi_\beta(W_{f_1})^*\Omega_\beta,$$
$$e^{-s_{j+1}H_\beta}\pi_\beta(W_{f_{j+1}})e^{-(s_{j+2}-s_{j+1})H_\beta}\pi_\beta(W_{f_{j+2}})\cdots e^{-(s_k-s_{k-1})H_\beta}\pi_\beta(W_{f_k})\Omega_\beta.$$

Moreover, the vector Ω_β is cyclic for the algebra $\mathfrak{M}^\beta_0$ generated by operators of the form

$$e^{it_1 H_\beta}\pi_\beta(W_{f_1})e^{-it_1 H_\beta}\cdots e^{it_n H_\beta}\pi_\beta(W_{f_n})e^{-it_n H_\beta}, \quad t_1,\ldots,t_n\in\mathbb{R},$$

and the state $\omega(M)=\langle\Omega_\beta, M\Omega_\beta\rangle$ is a β-KMS state on $\mathfrak{M}^\beta=(\mathfrak{M}^\beta_0)''$ with respect to the unitary group e^{itH_β}.

In the following, the $W^*-\beta$-KMS system constructed in Theorem 2.2 will be denoted as

$$\mathfrak{C}=(\mathfrak{H}_\beta,\Omega_\beta,\pi_\beta,e^{itH_\beta},\mathfrak{M}^\beta).$$

C. Stochastic positivity

For a general KMS state ω on an abstract C^*-algebra $\mathfrak{A}$, Klein and Landau[18] discussed the problem of construction of a stochastic process corresponding to ω. As they showed, such a process can be constructed using some Abelian sub-C^*-algebra $\mathfrak{B}$ of $\mathfrak{A}$ on which imaginary time Green's functions corresponding to ω are positive in some special sense. In such a case, the process has values in the spectrum of the Abelian algebra $\mathfrak{B}$. To study the existence of a stochastic process in the case of Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$, we consider Abelian subalgebras of $\mathfrak{W}(\mathcal{D},\sigma)$ defined in terms of so-called *A*belian splitting of the symplectic space $(\mathcal{D},\sigma)$.

Definition 2.3: A pair $(\mathcal{D}_+,\mathcal{D}_-)$ of linear subspaces of the symplectic space $(\mathcal{D},\sigma)$ is called an Abelian splitting if $\mathcal{D}=\mathcal{D}_+ + \mathcal{D}_-$ and $\sigma(\mathcal{D}_\pm,\mathcal{D}_\pm)=0$. For the Abelian splitting $(\mathcal{D}_+,\mathcal{D}_-)$, let $\mathfrak{W}_+$ and $\mathfrak{W}_-$ be the Abelian subalgebras of the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ generated by W_f with $f\in\mathcal{D}_+$ and $\mathcal{D}_-$, respectively.

Now the desired positivity condition is the following:

Definition 2.4:

(1) The set of imaginary time Green's functions $\mathbb{G}^E_\beta$ on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ with the given Abelian splitting is $\mathfrak{W}_\pm$-stochastically positive if for every positive element $W_1,\ldots,W_n \in \mathfrak{W}_\pm$ and every $(s_1,\ldots,s_n)\in\mathfrak{E}^\beta_n$

$$G^E_\beta(W_1,\ldots,W_n;s_1,\ldots,s_n)\geqslant 0.$$

(2) In the case of ground state system $\mathbb{G}^E_\infty$ the definition is analogous.

In the following discussion we will use the notions of thermal and ground state processes with values in some Hausdorff topological space $\mathbb{V}$.

Definition 2.5:

(1) Let $\mathbb{V}$ be a Hausdorff space. A stochastic process ξ_t taking values in $\mathbb{V}$ will be called a *thermal process* (with the inverse temperature β) if

TP1$_\beta$: ξ_t is stochastically continuous and faithful, i.e., for any $f\in C_b(\mathbb{V})$, $f>0$, we have $\mathbb{E}f(\xi_t)\neq 0$.

TP2$_\beta$: ξ_t is periodic with period β, i.e., for all $t\in\mathbb{R}$, $\xi_t=\xi_{t+\beta}$ in law (if K_β is the circle with length β parametrized as $[-\beta/2,\beta/2]$ with end points glued, then ξ_t can be indexed by $t\in K_\beta$).

TP3$_\beta$: ξ_t is OS-positive on K_β, i.e., for every $F\in C_b(\mathbb{V}^n)$ and $t_1,\ldots,t_n\in[0,\beta/2]$,

$$\mathbb{E}\overline{F}(\xi_{-t_1},\ldots,\xi_{-t_n})F(\xi_{t_1},\ldots,\xi_{t_n})\geqslant 0.$$

TP4$_\beta$: ξ_t is shift invariant on K_β, i.e., $\xi_{t+\tau}=\xi_t$ for all $t,\tau\in K_\beta$, and $t+\tau$ is defined modulo β.

TP5$_\beta$: ξ_t is reflection invariant, i.e., for every $F\in C_b(\mathbb{V}^n)$ and $t_1,\ldots,t_n\in K_\beta$,

$$\mathbb{E}F(\xi_{t_1},\ldots,\xi_{t_n})=\mathbb{E}F(\xi_{-t_1},\ldots,\xi_{-t_n}).$$

(2) Similarly, the process ξ_t is called a *ground state process* if it is stochastically continuous and faithful, shift and reflection invariant on $\mathbb{R}$ and OS-positive on $\mathbb{R}$.

Let ξ_t be the thermal process with values in $\mathcal{D}^*_+$, where $\mathcal{D}^*_+$ is an algebraic dual to $\mathcal{D}_+$. Define

$$G^\xi(\underline{f}^k;\underline{s}^k)=\mathbb{E}\left(\prod_{l=1}^k e^{i\langle\xi_{s_l},f_l\rangle}\right),$$

where $\underline{f}^k\in\mathcal{D}^k_+$ and $\underline{s}^k\in K^k_\beta$. It can be shown that $\mathbb{G}^\xi=\{G^\xi(\underline{f}^k;\underline{s}^k)\}$ satisfy conditions $(\mathrm{EG1})_\beta-(\mathrm{EG4})_\beta$ for the Abelian algebra $\mathfrak{W}_+$. Thus from Theorem 2.2, there exists the unique (up to the unitary equivalence) $W^*-\beta$-KMS system

$$\mathfrak{C}^\xi=(\mathfrak{H}^\xi,\Omega^\xi,\pi^\xi,e^{itH^\xi},\mathfrak{M}^\xi)$$

such that for every $\underline{f}^k\in\mathcal{D}^k_+$ and $-\beta/2\leqslant s_1\leqslant s_2\leqslant\cdots\leqslant s_j\leqslant 0\leqslant s_{j+1}\leqslant\cdots\leqslant s_k\leqslant\beta/2$,

$$G^E_\xi(\underline{f}^k;\underline{s}^k)=\langle e^{s_jH^\xi}\pi^\xi(W_{f_j})^* e^{-(s_j-s_{j-1})H^\xi}\pi^\xi(W_{f_{j-1}})^*\cdots e^{-(s_2s_1)H^\xi}\pi^\xi(W_{f_1})^*\Omega^\xi,$$
$$e^{-s_{j+1}H^\xi}\pi^\xi(W_{f_{j+1}})e^{-(s_{j+2}-s_{j+1})H^\xi}\pi^\xi(W_{f_{j+2}})\cdots e^{-(s_k-s_{k-1})H^\xi}\pi^\xi(W_{f_k})\Omega^\xi\rangle.$$

Now we show the converse:

Theorem 2.3:

(1) *Let $(\mathcal{D}_+,\mathcal{D}_-)$ be the Abelian splitting of the symplectic space $(\mathcal{D},\sigma)$. If $\mathbb{G}^E_\beta$ is the set of $\mathfrak{W}_+$-stochastically positive imaginary time Green's functions on $\mathfrak{W}(\mathcal{D},\sigma)$, there exists the unique (up to the stochastic equivalence) thermal process ξ^β_t with values in $\mathcal{D}^*_+$ such that, for every $f_1,\ldots,f_k\in\mathcal{D}_+$ and $\underline{s}^k\in K^k_\beta$,*

$$G^E_\beta(\underline{f}^k;\underline{s}^k)=\mathbb{E}\left(\prod_{l=1}^{k} e^{i\langle \xi^\beta_{s_1}, f_l\rangle}\right). \tag{2.1}$$

(2) *Similarly, if* $\mathbb{G}^E_\infty$ *is the set of* $\mathfrak{W}_+$*-stochastically positive ground state imaginary time Green's functions, there exists the unique (up to the stochastic equivalence) ground state process* ξ^∞_t *with values in* $\mathcal{D}^*_+$ *such that, for every* $f_1,\ldots,f_k\in\mathcal{D}_+$ *and* $-\infty<s_1\leqslant\ldots\leqslant s_k<\infty$,

$$G^E_\infty(\underline{f}^k;\underline{s}^k)=\mathbb{E}\left(\prod_{l=1}^{k} e^{i\langle \xi^\infty_{s_l}, f_l\rangle}\right). \tag{2.2}$$

Proof: Let $\mathbb{C}^+$ be the $W^*-\beta$-KMS system generated by the set $\mathbb{G}^+$ of imaginary time Green's functions $G^E(\underline{f}^k;\underline{s}^k)$ restricted to $\underline{f}^k\in\mathcal{D}^k_+$. Then, using the result of Ref. 18, we conclude that there exists the thermal process ξ^{KL}_t with values in the spectrum of the von Neumann algebra $(\pi_+(\mathfrak{W}_+))''$ (where π_+ denotes the representation defined by the set $\mathbb{G}^+$) and such that

$$G^E_\beta(\underline{f}^k;\underline{s}^k)=\int_{Q_\beta}\prod_{l=1}^{k}\widehat{\pi_+(W_{f_l})}(q(s_l))\,d\mu^{\xi^{KL}}(q),$$

where Q_β is the path space of the process ξ^{KL} and $\widehat{\pi_+(W_f)}$ is the Gel'fand transformation of $\pi_+(W_f)$.

For fixed $\underline{f}^k$ and $\underline{s}^k$ define the map

$$\underline{\alpha}\to\Gamma_{\underline{f}^k,\underline{s}^k}(\underline{\alpha}),$$

where $\underline{\alpha}=(\alpha_1,\ldots,\alpha_k)\in\mathbb{R}^k$, by

$$\Gamma_{\underline{f}^k,\underline{s}^k}(\underline{\alpha})=G^E(\alpha_1 f_1,\ldots,\alpha_k f_k;s_1,\ldots,s_k).$$

Here Γ is positive definite on $\mathbb{R}^k$ since for every $c_l\in\mathbb{C}$, $\underline{\alpha}^l\in\mathbb{R}^k$, $l=1,\ldots,M$,

$$\sum_{l,m=1}^{M} c_l\overline{c_m}\Gamma_{\underline{f}^k,\underline{s}^k}(\underline{\alpha}^l-\underline{\alpha}^m)=\int_{Q_\beta}\left|\sum_{l=1}^{M}\prod_{i=1}^{k}\widehat{\pi_+(W_{\alpha'_i f_i})}(q(s_i))\right|^2 d\mu^{\xi^{KL}}(q)\geqslant 0.$$

By the Bochner theorem, there exists Borel probability measure $d\nu^{\underline{f}^k,\underline{s}^k}$ on $\mathbb{R}^k$ such that

$$\Gamma_{\underline{f}^k,\underline{s}^k}(\underline{\alpha})=\int e^{i\langle\underline{\alpha},\underline{\gamma}\rangle}d\nu^{\underline{f}^k,\underline{s}^k}.$$

The system of finite-dimensional measures $d\nu^{\underline{f}^k,\underline{s}^k}$ forms a compatible system of measures on the cylindrical sets of $\mathcal{D}^{*K_\beta}_+$ as it follows from $(\mathrm{EG1})_\beta$ (5). Thus, by the theorem of Kolmogorov, there exists unique, up to the stochastic equivalence stochastic process, ξ^β_t with values in $\mathcal{D}^*_+$ such that (2.1) is satisfied. Similar arguments can be used in the ground state case.

Remarks:

(1) It is worthwhile to stress that the process corresponding to stochastically positive imaginary time Green's functions on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ can be realized as a process with values not in an abstract spectrum of Abelian algebra, but in the dual space to $\mathcal{D}_+$. In particular situations, the space of values of the process ξ^β_t can be localized in a much smaller subspace of $\mathcal{D}^*_+$ (see, i.e., Sec. IV and Refs. 24 and 25).

(2) In order to reconstruct all relevant informations about KMS structure (or ground state structure) from the stochastic process, it is important to take the Abelian subalgebra in such a way that the W^*-systems $\mathfrak{C}$ and $\mathfrak{C}^\xi$ (or analogous in the ground state case) are unitarily equivalent. It is interesting that in the quasi-free case discussed in the next section, Abelian subalgebras defined by some natural splitting always have this property.

III. QUASI-FREE STOCHASTICALLY POSITIVE STATES

A. Quasi-free states

To discuss the properties of quasi-free stochastically positive systems, assume that $\mathcal{D}$ has a structure of complex pre-Hilbert space with a scalar product $\langle\cdot,\cdot\rangle$ and $\sigma(f,g)=\mathrm{Im}\langle f,g\rangle$. Moreover, assume that T_t is a group of unitary operators with respect to $\langle\cdot,\cdot\rangle$ which leaves $\mathcal{D}$ invariant and has a self-adjoint generator $\mathbf{h}$. Let $\mathbb{B}(f,g)$ be a positive, sesqulinear form on $\mathcal{D}$ which is T_t-invariant and such that

$$|\sigma(f,g)|^2 \leqslant \mathbb{B}(f,f)\mathbb{B}(g,q).$$

Then ω defined by

$$\omega(W_f)=e^{-(1/4)\mathbb{B}(f,f)} \tag{3.1}$$

and extended by linearity and continuity to the whole $\mathfrak{W}(\mathcal{D},\sigma)$ is a so-called *gauge invariant quasi-free state*[34] on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$.

For quasi-free states, two-point Green's functions

$$G^{(2)}(f,g:t)=\omega(W_f W_{T_t g})$$

can be calculated, and have the form

$$G^{(2)}(f,g;t)=e^{-(1/4)\mathbb{B}(f,f)-(1/4)\mathbb{B}(g,g)}e^{-(1/2)F(f,g;t)}, \tag{3.2}$$

where

$$F(f,g;t)=\mathrm{Re}\ \mathbb{B}(f,T_t,g)+i\sigma(f,T_t g).$$

B. KMS quasi-free states

Now we consider a quasi-free state ω which is α_t-KMS state at the inverse temperature β for α_t defined by $T_t=e^{it\mathbf{h}}$. Then for every $f,g\in\mathcal{D}$ there exists a function $\Phi(f,g;z)$, analytic in the strip $0<\mathrm{Im}\, z<\beta$ and continuous on the boundary, such that

$$\Phi(f,g;t+i0)=F(f,g;t)$$

and

$$\Phi(f,g;t+i\beta)=F(g,f;-t).$$

Then

$$G^{(2)}(f,g;t)=e^{-(1/4)(\mathbb{B}(f,f)+\mathbb{B}(g,g))}e^{-(1/2)\Phi(f,g;t+i0)}$$

and

$$G^{(2)}(g,f;-t)=e^{-(1/4)(\mathbb{B}(f,f)+\mathbb{B}(g,g))}e^{-(1/2)\Phi(f,g;t+i\beta)}.$$

To find the explicit form of $\mathbb{B}(f,g)$ let us first consider the case when $\mathcal{D}$ is a Hilbert space and $\mathbb{B}(f,g)$ is bounded on $\mathcal{D}$, hence defined by some bounded operator B such that $B\geqslant 1$.

Theorem 3.1: *Suppose that ω defined by (3.1) is an α_t-*KMS *state at the inverse temperature β. Then*

(1) *there exists $\epsilon>0$ such that $\mathbf{h}\geqslant\epsilon$, and*

(2)

$$B=\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Proof: Since ω is α_t-KMS state, it is α_t-invariant. Thus B commutes with $T_t=e^{it\mathbf{h}}$, hence B commutes with spectral projectors $E(\lambda)$ of $\mathbf{h}$. First we show that 0 is not an eigenvalue of $\mathbf{h}$. Suppose that there exists $f_0\in\mathcal{D}$ such that $\mathbf{h}f_0=0$. Then for every $t\in\mathbb{R}$, $T_tf_0=f_0$. Because

$$F(f,f_0;t)=\mathrm{Re}\ \mathbb{B}(f,T_tf_0)+i\mathrm{Im}\langle f,T_tf_0\rangle=F(f,f_0;0),$$

$\Phi(f,f_0,z)$ is identically constant. Hence

$$F(f,f_0;0)=F(f_0,f;0).$$

On the other hand, $F(f,f_0;0)=\overline{F(f_0,f;0)}$, which implies $\sigma(f,f_0)=0$. But σ is nondegenerate, so $f_0=0$. Because $\Phi(f,f;z)$ is the analytic continuation of $F(f,f;t)$ and $F(f,f;-t)$, so[22]

$$\mathcal{F}(\mathbb{T}_F)(-p)=e^{-\beta p}\mathcal{F}(\mathbb{T}_F)(p),$$

where $\mathbb{T}_F$ is a distribution from $\mathcal{S}'$ given by

$$\langle\mathbb{T}_F,\varphi\rangle=\int_{-\infty}^{\infty}F(f,f,t)\varphi(t)\ dt,\quad \varphi\in\mathcal{S},$$

and $\mathcal{F}$ is the Fourier transform in $\mathcal{S}'$. Because

$$\begin{aligned}F(f,f;t)&=\mathrm{Re}\langle f,Be^{it\mathbf{h}}f\rangle+i\mathrm{Im}\langle f,e^{it\mathbf{h}}f\rangle\\&=\mathrm{Re}\int_{\mathbb{R}\setminus\{0\}}e^{it\lambda}d\langle Bf,E(\lambda)f\rangle+i\mathrm{Im}\int_{\mathbb{R}\setminus\{0\}}e^{it\lambda}d\langle f,E(\lambda)f\rangle\\&=\frac{1}{2}\int_{\mathbb{R}\setminus\{0\}}e^{it\lambda}\ d\langle(B+\mathbb{I})f,E(\lambda)f\rangle+\frac{1}{2}\int_{\mathbb{R}\setminus\{0\}}e^{it\lambda}\ d\langle(B-\mathbb{I})f,E(\lambda)f\rangle\end{aligned}$$

for any $\varphi\in\mathcal{S}$ we have

$$\begin{aligned}\langle\mathcal{F}(\mathbb{T}_F),\varphi_{\mathrm{inv}}\rangle&=\frac{1}{2}\int_{-\infty}^{\infty}dt\int_{\mathbb{R}\setminus\{0\}}d\langle(B+\mathbb{I})f,E(\lambda)f\rangle e^{it\lambda}\mathcal{F}(\sigma_{\mathrm{inv}})(t)\\&\quad+\frac{1}{2}\int_{-\infty}^{\infty}dt\int_{\mathbb{R}\setminus\{0\}}d\langle(B-\mathbb{I})f,E(\lambda)f\rangle e^{-it\lambda}\mathcal{F}(\varphi_{\mathrm{inv}})(t)\\&=\frac{\sqrt{2\pi}}{2}\int_{\mathbb{R}\setminus\{0\}}d\langle(B+\mathbb{I})f,E(\lambda)f\rangle\varphi(-\lambda)+\frac{\sqrt{2\pi}}{2}\int_{\mathbb{R}\setminus\{0\}}d\langle(B-\mathbb{I})f,E(\lambda)f\rangle\varphi(\lambda).\end{aligned}$$

On the other hand, it is equal to

$$\begin{aligned}\langle e^{-\beta\cdot}\mathcal{F}(\mathbb{T}_F),\varphi\rangle&=\frac{\sqrt{2\pi}}{2}\int_{\mathbb{R}\setminus\{0\}}d\langle(B+\mathbb{I})f,E(\lambda)f\rangle e^{-\beta\lambda}\varphi(\lambda)\\&\quad+\frac{\sqrt{2\pi}}{2}\int_{\mathbb{R}\setminus\{0\}}d\langle(B-\mathbb{I})f,E(\lambda)f\rangle e^{\beta\lambda}\varphi(-\lambda).\end{aligned}$$

Because Ker $\mathbf{h}=0$,

$$\mathbf{h}=\int_{-\infty}^{0^-}\lambda\ dE(\lambda)+\int_{0^+}^{\infty}\lambda\ dE(\lambda).$$

Assume that $E((-\infty,0))\neq 0$ and take $f\neq 0$, $f\in\tilde{\mathcal{D}}=E((-\infty,0))\mathcal{D}$. Suppose also that $\varphi\geqslant 0$ and supp $\varphi=[0,\infty)$. Then

$$\int_{-\infty}^{0} d\langle (B+\mathbb{I})f, E(\lambda)f\rangle \varphi(-\lambda) = \int_{-\infty}^{0} d\langle (B-\mathbb{I})f, E(\lambda)f\rangle e^{\beta\lambda}\varphi(-\lambda),$$

thus

$$\langle f, (B+\mathbb{I})\varphi(\mathbf{h}^{-1})f\rangle = \langle f, (B-\mathbb{I})e^{-\beta\mathbf{h}^{-}}\varphi(\mathbf{h}^{-})f\rangle,$$

where

$$\mathbf{h}^{-} = \int_{-\infty}^{0} (-\lambda)\, dE(\lambda).$$

Because $B:\tilde{\mathcal{D}}\to\tilde{\mathcal{D}}$, so by polarization we obtain that

$$(B+\mathbb{I})\varphi(\mathbf{h}^{-}) = (B-\mathbb{I})e^{-\beta\mathbf{h}^{-}}\varphi(\mathbf{h}^{-}).$$

But φ is arbitrary, so

$$B|_{\tilde{\mathcal{D}}} = \frac{\mathbb{I}+e^{-\beta\mathbf{h}^{-}}}{e^{-\beta\mathbf{h}^{-}}-\mathbb{I}} < -\mathbb{I}.$$

Thus we get the contradiction. Hence $\mathbf{h}>0$ and using similar arguments we show that

$$B = \frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Suppose now that for every $n\in\mathbb{N}$ the projector $E((0,1/n))$ is nonzero. Take $f_n\in E((0,1/n))\mathcal{D}$ such that $\|f_n\|=1$. Then

$$\|B^{1/2}f_n\|^2 = \int_0^{1/n} a(\lambda)\, d\rho_n(\lambda),$$

where $a(\lambda)=(1+e^{-\beta\lambda})^{-1}(1-e^{-\beta\lambda})$ and $d\rho_n(\lambda)=d\langle f_n, E(\lambda)f_n\rangle$. But for every $n\in\mathbb{N}$ the measure $d\rho_n$ is normed to 1 on $(0,1/n)$, so

$$\|B^{1/2}f_n\|^2 \geqslant a\left(\frac{1}{n}\right)\to\infty.$$

It contradicts the assumption that B is bounded.

C. Analytic continuation

Starting with the functions $\Phi(f,g;z)$ we can define a two-point imaginary time Green's function in terms of

$$S(f,g;s) = \Phi(f,g;is) \tag{3.3}$$

for $s\in[0,\beta]$. Then

$$G^{E,2}(f,g;s) = e^{-(1/4)(\mathbb{B}(f,f)+\mathbb{B}(g,g))}e^{-(1/2)S(f,g;s)}. \tag{3.4}$$

Proposition 3.1: $S(f,g;s)$ defined by (3.3) satisfies the following.
(1) For every $f,g\in\mathcal{D}$ and $s\in[0,\beta]$

$$S(f,g;s) = S(g,f;\beta-s).$$

(2) For every $f\in\mathcal{D}$ the mapping $s\to S(f,f;s)$ is OS-positive, i.e., for every sequence $\{s_k\}$, $s_k\in[0,\beta/2]$; $\{c_k\}$, $c_k\in\mathbb{C}$.

$$\sum_{k,l} \overline{c_k} c_l S(f,f;s_k+s_l) \geqslant 0.$$

More generally, for every terminating sequence $\{f_k\}$, $f_k \in \mathcal{D}$; $\{s_k\}$, $s_k \in [0,\beta/2]$, and $\{c_k\}$, $c_k \in \mathbb{C}$,

$$\sum_{k,l} \overline{c_k} c_l S(f_k,f_l;s_k+s_l) \geqslant 0.$$

Proof: The function $S(f,g;s)$ can be represented by the following integral formula,[22]

$$S(f,g;s) = \int_{-\infty}^{\infty} [\mathfrak{B}(\rho,s)F(f,g;\rho) + \mathfrak{B}(\rho,\beta-s)F(g,f;-\rho)]\, d\rho, \tag{3.5}$$

with the kernel

$$\mathfrak{B}(\rho,s) = \frac{1}{2\beta} \sin \frac{\pi s}{\beta} \left(\cosh \frac{\pi \rho}{\beta} - \cos \frac{\pi s}{\beta} \right)^{-1}.$$

From this formula, we see that $S(f,g;s) = S(g,f;\beta-s)$. On the other hand, the function $t \to F(f,f;t)$ is positive definite since

$$F(f,g;t) = F_+(f,g;t) + F_-(f,g;t),$$

where

$$F_+(f,g;t) = \tfrac{1}{2}\langle f, (B+\mathbb{I})e^{it\mathbf{h}} g\rangle, \quad F_-(f,g;t) = \tfrac{1}{2}\langle g, (B-\mathbb{I})e^{-it\mathbf{h}} f\rangle,$$

and the functions $F_+(f,f;t), F_-(f,f;t)$ are obviously positive definite. Thus $S(f,f;s)$ is OS-positive as a Laplace transform of positive measure. By polarization

$$\sum_{k,l} \overline{c_k} c_l F_{\pm}(f_k,f_l;t_k-t_l) \geqslant 0$$

and

$$F_{\pm}(f,g;t) = \int_{-\infty}^{\infty} e^{it\lambda}\, d\mu_{f,g}^{\pm}(\lambda)$$

for complex-valued measure $\mu_{f,g}^{\pm}$. Thus

$$\sum_{k,l} \overline{c_k} c_l F(f_k,f_l;t_k-t_l) \geqslant 0$$

and

$$F(f,q;t) = \int_{-\infty}^{\infty} e^{it\lambda}\, d\mu_{f,g}(\lambda)$$

for $\mu_{f,g} = \mu_{f,g}^{+} + \mu_{f,g}^{-}$.

Hence the matrix of measures (μ_{f_k,f_l}) is positive definite, and the matrix

$$(S(f_k,f_l;s_k+s_l)) = \left(\int_{-\infty}^{\infty} e^{-s(s_k+s_l)\lambda}\, d\mu_{f_k,f_l}(\lambda) \right)$$

is also positive definite.

Corollary 3.1:

$$G^{E,2}(f,g;s)$$

satisfies

(1) $$G^{E,2}(f,g;s)=G^{E,2}(g,f;\beta-s), \quad s\in[0,\beta];$$

(2) $$\sum_{k,l}\overline{c}_k c_l G^{E,2}(-f_k,f_l,s_k+s_l)\geqslant 0, \quad s_k\in[0,\beta/2].$$

Remark: Multi-time Green's functions corresponding to the quasi-free KMS state can be obtained as follows. By a direct computation we show that

$$G(f_1,\ldots,f_n;t_1,\ldots,t_n)=\omega(W_{T_{t_1}f_1}\cdots W_{T_{t_n}f_n})=\prod_{l=1}^{n} e^{[(n-2)/4]\mathbb{B}(f_l,f_l)}\prod_{(j_n,k_n)} G^{(2)}(f_{j_n},f_{k_n};t_{k_n}-t_{j_n}),$$

where the second product is taken over all pairs of different indices (j_n,k_n) such that $j_n,\ k_n \in\{1,\ldots,n\},\ t_{j_n}<t_{k_n}$. Analytic continuation of the right-hand side of this expression gives after restriction to imaginary times the formula

$$G^{E}(f_1,\ldots,f_n;s_1,\ldots,s_n)=\prod_{l=1}^{n} e^{[(n-2)/4]\mathbb{B}(f_l,f_l)}\prod_{(j_n,k_n)} G^{E,2}(f_{j_n},f_{k_n};s_{k_n}-s_{j_n}),$$

which can be rewritten in the following way:

$$G^{E}(f_1,\ldots,f_n;s_1,\ldots,s_n)=\prod_{k=1}^{n} e^{-(1/4)\mathbb{B}(f_k,f_k)}\prod_{(j_k,l_k)} e^{-(1/2)S(f_{j_k},f_{l_k};s_{l_k}-s_{j_k})}.$$

It can be checked that $G^{E}(f_1,\ldots,f_n;s_1,\ldots,s_n)$ satisfy all conditions of Theorem 2.1.

D. Stochastic positivity and thermal process

Now we can pass to the question of stochastic positivity of quasi-free imaginary time Green's functions. This property can be naturally formulated in terms of the functions $S(f,g;s)$.

Let C be an abstract complex conjugation on $\mathcal{D}$, i.e., C is antiunitary and $C^2=1$. The mapping C naturally defines the Abelian splitting $(\mathcal{D}_+,\mathcal{D}_-)$ of $\mathcal{D}$:

$$\mathcal{D}_+=\{f\in\mathcal{D}:Cf=f\}, \quad \mathcal{D}_-=\{f\in\mathcal{D}:Cf=-f\}.$$

Lemma 3.1: The functions $S(f,g;s)$ restricted to $\mathcal{D}_+$ satisfy

$$S(f,g;s)=S(g,f,;s), \quad s\in[0,\beta], \tag{3.6}$$

if $\mathbf{h}$ is C-real, i.e., C leaves the domain $D(\mathbf{h})$ invariant and

$$C\mathbf{h}f=\mathbf{h}Cf$$

for every $f\in D(\mathbf{h})$.

Proof: By analytic continuation, the property (3.6) is equivalent to

$$F(f,g;t)=F(g,f;t)$$

for $f,g\in\mathcal{D}_+$ and $t\in\mathbb{R}$. So we have to show that

$$\mathrm{Re}\langle f,(T_t-T_{-t})Bg\rangle=0 \tag{3.7}$$

and

$$\mathrm{Im}\langle f,(T_t+T_{-t})g\rangle=0 \tag{3.8}$$

for all $f,g\in\mathcal{D}_+$. Because $\mathbf{h}$ is C-real, C commutes with spectral projectors of $\mathbf{h}$ and

$$CT_t = T_{-t}C,$$

thus

$$T_t + T_{-t} : \mathcal{D}_+ \to \mathcal{D}_+ ,$$

$$T_t - T_{-t} : \mathcal{D}_+ \to \mathcal{D}_- .$$

Hence (3.8) is satisfied. B as a real function of $\mathbf{h}$ is C-real, too, and so B: $\mathcal{D}_+ \to \mathcal{D}_+$. But

$$\mathrm{Re}\langle f_+ , g_- \rangle = 0$$

for all $f_+ \in \mathcal{D}_+$, $g_- \in \mathcal{D}_-$, so (3.7) follows.

Combining this result with that of Proposition 3.1 we get

$$S(f,g;s) = S(f,g;\beta - s), \quad s \in [0,\beta].$$

This allows us to extend the function $S(f,g;s)$ (for the fixed $f,g \in \mathcal{D}_+$) to the periodic function of s with the period β, defined for all $s \in \mathbb{R}$. The extended functions will also be denoted by $S(f,g;s)$.

Proposition 3.2: For every fixed $f,g \in \mathcal{D}_+$ there exists a finite, real-valued measure $\nu_{f,g}$ on $[0,\infty)$ such that

$$S(f,g;s) = \int_0^\infty (e^{-sp} + e^{-(\beta - s)p})\, d\nu_{f,g} . \tag{3.9}$$

Proof: For a fixed $f \in \mathcal{D}_+$, the function $S(f,f,s)$ is OS-positive, thus by the Widder theorem there exists a finite measure $\tilde{\nu}_{f,f}$ on $\mathbb{R}$ such that

$$S(f,f;s) = \int_{-\infty}^\infty e^{-sp}\, d\tilde{\nu}_{f,f}$$

for $s \in [0,\beta]$. Since $S(f,f;s) = S(f,f;\beta - s)$,

$$d\tilde{\nu}_{f,f}(-p) = e^{-\beta p} d\tilde{\nu}_{f,f}(p).$$

Thus there exists a unique finite measure $\nu_{f,f}$ with support on $[0,\infty)$ such that

$$S(f,f;s) = \int_0^\infty (e^{-sp} + e^{-(\beta - s)p})\, d\nu_{f,f} .$$

By polarization, this gives the integral representation of $S(f,g;s)$ in terms of real-valued measure $\nu_{f,g}$.

Proposition 3.3: $S(f,g;s)$ is positive definite, i.e.,

$$\sum_{k,l=1}^n \overline{c_k} c_l S(f_k, f_l; s_k - s_l) \geqslant 0, \tag{3.10}$$

for all terminating sequences $f_1,\ldots,f_n \in \mathcal{D}_+$; $s_1,\ldots,s_n \in \mathbb{R}$ and $c_1,\ldots,c_n \in \mathbb{C}$.

Proof: Since

$$e^{-sp} + e^{-(\beta - s)p} = \sum_{n \in \mathbb{Z}} c_n(p) e^{i2\pi n s/\beta} \tag{3.11}$$

for all $s \in [0,\beta]$, where

$$c_n(p) = ((p\beta)^2 + (2\pi n)^2)^{-1} (2\beta p (1 - e^{-\beta p})) \geqslant 0,$$

the right-hand side of (3.11) defines a periodic function for all $s \in \mathbb{R}$.[35] Combining this with (3.9) we get

$$S(f,f;s)=\sum_{n\in\mathbb{Z}} e^{i2\pi ns/\beta}\int_0^\infty c_n(p)\,d\nu_{f,f}(p),$$

so for a fixed $f\in\mathcal{D}_+$, $S(f,f;s)$ is a real-valued positive-definite function as a function of s. Thus, it is a Fourier transform of a positive measure $m_{f,f}$ on $\mathbb{R}$. Using the polarization formula for $S(f,g;s)$ we obtain that $S(f,g;s)$ is a Fourier transform of real-valued measure $m_{f,g}$ with finite variation. Let

$$m=(m_{f_k,f_l})$$

be an $n\times n$ matrix of such measures. We show that m is positive definite, i.e., for every Borel subset $E\subset\mathbb{R}$, the matrix $(m_{f_k,f_l}(E))$ is positive definite. Let $\varphi\in\mathcal{S}$ be non-negative. For $g_k=r_kf_k$, $r_k\in\mathbb{R}$, we have

$$\sum_{k,l} r_k r_l m_{f_k,f_l}(\varphi)=\sum_{k,l} m_{g_k,g_l}(\varphi)=m_{(\Sigma_k g_k),(\Sigma_k g_k)}(\varphi)\geqslant 0.$$

It implies that the matrix

$$(S(f_k,f_l;s_k-s_l))=\left(\int_{-\infty}^{\infty} e^{i(s_k-s_l)p}\,dm_{f_k,f_l}(p)\right) \tag{3.12}$$

is also positive definite. To show this, notice that the integral on the right-hand side of (3.12) can be approximated by the sum

$$\sum_{j=0}^{N} e^{i(s_k-s_l)p_j}m_{f_k,f_l}(E_j).$$

Because $(e^{i(s_k-s_l)p_j})$ and $(m_{f_k,f_l}(E_j))$ are positive-definite matrices, from the Schur lemma the matrix

$$(e^{i(s_k-s_l)p_j}m_{f_k,f_l}(E_j))$$

is positive definite for each j, so the matrix (3.12) is positive definite. Thus $S(f_k,f_l;s_k-f_l)$ is a real-valued positive-definite matrix in real vector space. It follows that it is also positive definite as a matrix in complex vector space.

Theorem 3.2: *$S(f,g;s)$ defines an operator-valued covariance function $R_\beta(s)$ of a periodic Gaussian OS-positive stochastic process indexed by $\mathcal{D}_+$. Thus $R_\beta(s)$ is an operator-valued positive-definite function on $\mathcal{D}_+$ which is periodic and OS-positive. Moreover,*

$$S(f,g;s)=\langle f,R_\beta(s)g\rangle.$$

Proof: From the positive definiteness of $S(f,g;s)$ we have

$$|S(f,g;s)|^2\leqslant S(f,f;0)S(g,g;0)\leqslant\|B\|^2\|f\|^2\|g\|^2.$$

Since $S(f,g;s)$ is bilinear and symmetric, there exists a bounded and positive operator $R_\beta(s)$ on $\mathcal{D}_+$ such that

$$S(f,g;s)=\langle f,R_\beta(s)g\rangle.$$

Moreover, the function

$$s\to R_\beta(s)$$

is positive definite, OS-positive and weakly periodic.

Corollary 3.2: Let ξ_s^β be the Gaussian process indexed by $\mathcal{D}_+$ defined by

$$\mathbb{E}(\langle \xi_s^\beta, f\rangle)=0$$

and

$$\mathbb{E}(\langle \xi_{s_1}^\beta, f_1\rangle\langle \xi_{s_2}^\beta, f_2\rangle)=\tfrac{1}{2}\langle f_1, R_\beta(s_2-s_1)f_2\rangle.$$

Then, for $f_1, f_2 \in \mathcal{D}_+$,

$$G^{E,2}(f_1,f_2;s_2-s_1)=\mathbb{E}(e^{i\langle \xi_{s_1}^\beta, f_1\rangle}e^{i\langle \xi_{s_2}^\beta, f_2\rangle}).$$

Similarly, for $f_1,\ldots,f_n \in \mathcal{D}_+$,

$$G^E(f_1,\ldots,f_n;s_1,\ldots,s_n)=\mathbb{E}(e^{i\Sigma_{k=1}^n\langle \xi_{s_k}^\beta, f_k\rangle}).$$

Proposition 3.4: If $f \in \mathcal{D}_+$ is such that $m(f)=\int_0^\infty p\, d\nu_{f,f}(p)<\infty$, then the coordinate process $\xi_s^f := \langle \xi_s^\beta, f\rangle$ has a version (denoted by the same symbol) with Hölder continuous paths. More precisely, for any $0<\gamma<\frac{1}{2}$ there exists an integrable random variable $d(f,\gamma)$ such that

$$|\xi_s^f-\xi_{s'}^f|\leqslant d(f,\gamma)|s-s'|^\gamma$$

with probability one.

Proof: Since

$$|S(f,f;h)-S(f,f;0)|=\left|\int_0^\infty ((e^{-hp}-1)+(e^{-(\beta-h)p}-e^{-\beta p}))\,d\nu_{f,f}(p)\right|$$

$$\leqslant 2|h|\int_0^\infty p\,d\nu_{f,f}(p)=2m(f)|h|,$$

it follows that

$$|\mathbb{E}e^{z|\xi^f_{s+h}-\xi^f_s|}|\leqslant \mathbb{E}e^{|z|(\xi^f_{s+h}-\xi^f_s)}+\mathbb{E}e^{-|z|(\xi^f_{s+h}-\xi^f_s)}=2e^{(|z|^2/2)\mathbb{E}(\xi^f_{s+h}-\xi^f_s)^2}$$

$$\leqslant 2e^{|z||S(f,f;0)-S(f,f;h)|}\leqslant 2e^{2|z|h|m(f)}.$$

On the other hand, by the Cauchy integral formula,

$$|\xi^f_{s+h}-\xi^f_s|^n=\frac{n!}{2\pi i}\int_{C_r}\frac{e^{z|\xi^f_{s+h}-\xi^f_s|}}{z^{n+1}}\,dz.$$

Thus

$$\mathbb{E}|\xi^f_{s+h}-\xi^f_s|^n\leqslant 2n!\,\frac{e^{2r^2|h|m(f)}}{r^n}.$$

Now taking $r=|h|^{-1/2}$ we obtain

$$\mathbb{E}|\xi^f_{s+h}-\xi^f_s|^n\leqslant c_n|h|^{n/2}.$$

The assertion follows by the application of the Kolmogorov continuity test.

Definition 3.1: A periodic stochastic process ξ_s indexed by $\mathcal{D}_+$ has the *two-sided Markov property on the circle* K_β iff for all $r,s \in K_\beta$,

$$E_{[s,r]}E_{[r,s]}=E_{\{r,s\}}E_{[r,s]},$$

where for $I \subset K_\beta$, E_I denotes conditional expectation with respect to the σ-algebra generated by $\{\xi_s, s \in I\}$, and $\{r,s\}=\{s,r\}$ is the set consisting of the two elements r,s.

Proposition 3.5: ξ_s^β has a version with the two-sided Markov property on the circle K_β.

Proof: The covariance operator $R_\beta(s)$ has the form

$$R_\beta(s)=\frac{e^{-s\mathbf{h}}+e^{-(\beta-s)\mathbf{h}}}{1-e^{-\beta\mathbf{h}}}.$$

Let us introduce the new scalar product in $\mathcal{D}_+$ given by

$$\langle f,g\rangle_\beta := \langle f,(1-e^{-\beta\mathbf{h}})^{-1}g\rangle.$$

The norms $\|\cdot\|$ and $\|\cdot\|_\beta$ are obviously equivalent. Let $\tilde{\xi}_s$ be the Gaussian process indexed by $(\mathcal{D}_+,\langle\cdot,\cdot\rangle_\beta)$ with zero mean and covariance operator

$$\tilde{R}_\beta(s)=e^{-s\mathbf{h}}+e^{-(\beta-s)\mathbf{h}}.$$

Here $\tilde{\xi}_s$ is stochastically equivalent to ξ_s^β and by the result of Ref. 35 it satisfies the two-sided Markov property on the circle K_β.

E. KMS structure generated by thermal process

Let ξ_s^β be a Gaussian process constructed above and let (Q,Σ,μ) be its underlying probability space. Since the process is stationary, $u(t)$ defined by

$$u(t)(e^{i\langle\xi^\beta_{s_1},f_1\rangle}\cdots e^{i\langle\xi^\beta_{s_n},f_n\rangle})=e^{i\langle\xi^\beta_{s_1+t},f_1\rangle}\cdots e^{i\langle\xi^\beta_{s_n+t},f_n\rangle}$$

extends to a one-parameter group of unitary operators on $L^2(Q,\Sigma,\mu)$. By periodicity, $u(\beta)=\mathbb{I}$. Since the process is symmetric, Θ defined by

$$\Theta(e^{i\langle\xi^\beta_{s_1},f_1\rangle}\cdots e^{i\langle\xi^\beta_{s_n},f_n\rangle})=e^{i\langle\xi^\beta_{-s_1},f_1\rangle}\cdots e^{i\langle\xi^\beta_{-s_n},f_n\rangle}$$

extends to a unitary operator on $L^2(Q,\Sigma,\mu)$ such that $\Theta^2=\mathbb{I}$. Finally, since the process is OS-positive,

$$\langle\Theta F,F\rangle_{L^2}\geqslant 0$$

for all $F\in L^2(Q,\Sigma_{[0,\beta/2]},\mu)$, where for $S\subset\mathbb{R},\Sigma_S$ denotes the σ-algebra generated by $\{\xi_s^\beta : s \in S\}$.

Theorem 3.3: *Let ξ_s^β be a Gaussian, periodic (with period β), OS-positive stochastic process indexed by $\mathcal{D}_+$. Then there exist a Hilbert space $\mathfrak{H}_\xi$ with a unit vector Ω_ξ, a weakly continuous one-parameter group of unitary operators $U_\xi(t)=e^{itH_\xi}$, and a von Neumann algebra $\mathfrak{M}_\xi$ of operators acting on $\mathfrak{H}_\xi$ such that Ω_ξ is cyclic and separating for $\mathfrak{M}_\xi$ and $\alpha_t^\xi(M)=e^{itH_\xi}Me^{-itH_\xi}$ is the modular automorphisms group associated with Ω_ξ.*

Proof: On the space $L^2(Q,\Sigma_{[0,\beta/2]},\mu)$ define a sesquilinear form by

$$\langle F,G\rangle=\langle\Theta F,G\rangle_{L^2}.$$

By OS-positivity, it is positive semi-definite. Let

$$\mathcal{N}=\{F\in L^2(Q,\Sigma_{[0,\beta/2]},\mu):\langle F,F\rangle=0\}.$$

Then

$$\mathfrak{D}=L^2(Q,\Sigma_{[0,\beta/2]},\mu)/\mathcal{N}$$

is a pre-Hilbert space with respect to the inner product

$$\langle[F],[G]\rangle=\langle F,G\rangle,$$

where $[F]$ denotes the class containing F. Here $\mathfrak{H}_\xi$ is defined as a Hilbert space completion of $\mathcal{D}$ and $\Omega_\xi=[1]$. Let $\mathcal{D}_t$ be the linear space generated by vectors $[F]$ for $F\in L^2(Q,\Sigma_{[0,\beta/2-t]},\mu)$, $t\in[0,\beta/2]$. For every $t\in[0,\beta/2]$ we can define the linear operator $p(t)$ with domain $\mathfrak{D}_t$ by

$$p(t)[F]=[u(t)F],$$

and we can show that $(p(t),\mathfrak{D}_t)$ form a symmetric local semigroup.[36] Hence there exists a unique self-adjoint operator H_ξ on $\mathfrak{H}_\xi$ such that $p(t)=e^{-tH_\xi}$. Here $U_\xi(t)$ is defined by $U_\xi(t)=e^{itH_\xi}$. Now let $F_0\in L^\infty(Q,\Sigma_0,\mu)$. Then

$$\pi_0(F_0)[F]=[F_0F]$$

defines a bounded operator on $\mathfrak{H}_\xi$ and

$$\mathfrak{M}_0=\{\pi_0(F_0):F_0\in L^\infty(Q,\Sigma_0,\mu)\}$$

is a von Neumann algebra of operators on $\mathfrak{H}_\xi$. Let $\mathfrak{M}_\xi$ be the von Neumann algebra generated by elements

$$e^{it_1H_\xi}B_1e^{-it_1H_\xi}\cdots e^{it_nH_\xi}B_ne^{-it_nH_\xi}$$

with $t_j\in\mathbb{R}$, $B_j\in\mathfrak{M}_0$. Then Ω_ξ is cyclic and separating for $\mathfrak{M}_\xi$.[23] Using the properties of the process ξ_s^β we can now define the modular conjugation and modular group corresponding to Ω_ξ. Let

$$\Theta_{\beta/4}=u(\beta/4)\Theta u(-\beta/4)$$

and

$$J_\xi[F]=[\overline{\Theta_{\beta/4}F}].$$

Then

$$\langle J_\xi[F],J_\xi[G]\rangle=\langle\Theta\overline{\Theta_{\beta/4}F},\overline{\Theta_{\beta/4}G}\rangle_{L^2}=\langle G,\Theta F\rangle_{L^2}=\langle[G],[F]\rangle$$

since $\Theta_{\beta/4}$ commutes with Θ by periodicity. Hence J_ξ can be extended to an antiunitary operator on $\mathfrak{H}_\xi$ such that $J_\xi^2=\mathbb{I}$. Computing the action of J_ξ on $\mathfrak{M}_\xi\Omega_\xi$, $M\in\mathfrak{M}_\xi$, we can show that

$$J_\xi M^*\Omega_\xi=e^{(-\beta/2)H_\xi}M\Omega_\xi.$$

Now since $\mathfrak{M}_\xi\Omega_\xi$ is a core for $e^{(-\beta/2)H_\xi}$,[18] J_ξ defined above is the modular conjugation operator and the corresponding modular operator Δ_ξ can be identified with $e^{(-\beta/2)H_\xi}$. For more details see Ref. 18.

Let $(\mathfrak{H}_\omega,\pi_\omega,\Omega_\omega)$ be the GNS representation defined by quasi-free KMS state ω. Then Ω_ω is cyclic and separating for $\pi_\omega(\mathfrak{W}(\mathcal{D},\sigma))''$. Let J_ω and Δ_ω be the corresponding modular conjugation and modular operator. We are going to show that in the case of the quasi-free state the modular structure constructed from the process ξ_s^β is unitarily equivalent to the canonical modular structure defined by KMS state ω. Thus all relevant information about KMS structure is contained in the (commutative) stochastic process ξ_s^β. To obtain this result we need the following property of the dynamics $e^{it\mathbf{h}}$.

Theorem 3.4: *Let* $\mathbf{h}$ *be C-real on* $\mathcal{D}$ *and let* $E(\{0\})=0$, *where* $dE(\lambda)$ *is the spectral measure of* $\mathbf{h}$. *Then the family*

$$\left\{\sum_{k=1}^n e^{it_k\mathbf{h}}f_k:n\in\mathbb{N},t_k\in\mathbb{R},f_k\in\mathcal{D}_+\right\}$$

is dense in $\mathcal{D}$.

Proof: Because C commutes with spectral projectors of $\mathbf{h}$ and 0 is not an eigenvalue of $\mathbf{h}$, so

$$\mathbf{h}=\mathbf{h}_1\oplus\mathbf{h}_2, \quad \mathcal{D}=\mathcal{D}_1\oplus\mathcal{D}_2 \quad \text{and} \quad \mathcal{D}_+=\mathcal{D}_{+,1}\oplus\mathcal{D}_{+,2},$$

where

$$\mathcal{D}_1=E((0,\infty))\mathcal{D}, \quad \mathcal{D}_2=E((-\infty,0))\mathcal{D}, \quad \mathcal{D}_{+,1}=\mathcal{D}_+\cap\mathcal{D}_1, \quad \mathcal{D}_{+,2}=\mathcal{D}_+\cap\mathcal{D}_2.$$

Hence it is enough to consider a positive operator $\mathbf{h}$ such that $\operatorname{Ker}\mathbf{h}=\{0\}$. Because $\mathcal{D}$ is separable, $\mathbf{h}$ can be written as

$$\mathbf{h}=\oplus_{l=1}^{\infty}\mathbf{h}_l, \quad \mathcal{D}=\bigoplus_{l=1}^{\infty}\mathcal{D}_l,$$

where each $\mathbf{h}_l$ is a positive operator with a simple spectrum. Moreover, each generating vector g_l for a Hilbert space $\mathcal{D}_l$ can be taken from $\mathcal{D}_+$ and normalized. Indeed, let us take arbitrary $g_1\in\mathcal{D}_+$ with $\|g_1\|=1$ and define

$$\mathcal{D}_1=\overline{\operatorname{Lin}_{\mathbb{C}}\{E(\sigma)g_1\}}, \quad \sigma\in\mathcal{B}(0,\infty).$$

Because C commutes with all $E(\sigma)$, so $C:\mathcal{D}_1\to\mathcal{D}_1$ and hence $C:\mathcal{D}_1^{\perp}\to\mathcal{D}_1^{\perp}$. In the second step we take $g_2\in\mathcal{D}_1^{\perp}\cap\mathcal{D}_+$, $\|g_2\|=1$ and define

$$\mathcal{D}_2=\overline{\operatorname{Lin}_{\mathbb{C}}\{E(\sigma)g_2\}}.$$

It is clear that $\mathcal{D}_1\perp\mathcal{D}_2$. The total decomposition follows from the Kuratowski–Zorn lemma. Because every $\mathcal{D}_l$ reduces $\mathbf{h}$, $\mathbf{h}=\bigoplus_{l=1}^{\infty}\mathbf{h}_l$ where $\mathbf{h}_l:D(\mathbf{h})\cap\mathcal{D}_l\to\mathcal{D}_l$. Moreover, each $\mathbf{h}_l$ is $C|_{\mathcal{D}_l}$-real.

Let us assume that we have proved that for every $n\in\mathbb{N}$ the set

$$\left\{\sum_{k=1}^{m} e^{it_k\mathbf{h}_k}f_n^k : m\in\mathbb{N}, t_k\in\mathbb{R}, f_n^k\in\mathcal{D}_{+,n}\right\}$$

is dense in $\mathcal{D}_n$. Then for any $\epsilon>0$ and for any $f\in\mathcal{D}$ we can find $n_0\in\mathbb{N}$ such that

$$\left\|\bigoplus_{n=1}^{n_0} f_n-f\right\|<\epsilon/3$$

and for these $f_n\in\mathcal{D}_n$ there are $f_n^k\in\mathcal{D}_{+,n}$ such that

$$\left\|f_n-\sum_{k=1}^{m_n} e^{it_k(n)\mathbf{h}_n}f_n^k\right\|\leqslant\epsilon/2^n.$$

It follows that

$$\left\|f-\bigoplus_{n=1}^{n_0}\sum_{k=1}^{m_n} e^{it_k(n)\mathbf{h}_n}f_n^k\right\|\leqslant\left\|f-\bigoplus_{n=1}^{n_0}f_n\right\|+\left\|\bigoplus_{n=1}^{n_0}f_n-\bigoplus_{n=1}^{n_0}\sum_{k=1}^{m_n} e^{it_k(n)\mathbf{h}_n}f_n^k\right\|$$

$$\leqslant\epsilon/3+\left[\sum_{n=1}^{n_0}\left\|f_n-\sum_{k=1}^{m_n} e^{it_k(n)\mathbf{h}_n}f_n^k\right\|^2\right]^{1/2}\leqslant\epsilon/3+\left[\sum_{n=1}^{n_0}(\epsilon^2/4^n)\right]^{1/2}<\epsilon.$$

Because every f_n^k is an element of $\mathcal{D}_+$ and we can write

$$e^{it_k(n)\mathbf{h}_n}f_n^k=e^{it_k(n)\mathbf{h}}f_n^k,$$

we obtain

$$\left\| f-\sum_{n=1}^{n_0}\sum_{k=1}^{m_n} e^{it_k(n)\mathbf{h}} f_n^k \right\| < \epsilon.$$

Hence to finish the proof we need only the following Lemma.

Lemma 3.2: Suppose that $\mathbf{h}$ *is C-real and positive on* $\mathcal{D}$. *If* Ker $\mathbf{h}=\{0\}$ *and* $\mathbf{h}$ *is simple with a generating vector* $g\in\mathcal{D}_+$, $\|g\|=1$, *then the family*

$$\left\{ \sum_{k=1}^{n} e^{it_k\mathbf{h}} f_k : n\in\mathbb{N}, t_k\in\mathbb{R}, f_k\in\mathcal{D}_+ \right\}$$

is dense in $\mathcal{D}$.

Proof: By the spectral theorem, $\mathbf{h}$ is unitarily isomorphic to the multiplication operator $\hat{x}$ in $L^2(\mathbb{R}_+,\mathcal{B},\rho)$, where $\rho(\sigma)=\langle g,E(\sigma)g\rangle$. This means that the unitary operator U given by

$$Ua=\int_0^\infty a(\lambda)\, dE(\lambda)g$$

maps $D(\hat{x})$ onto $D(\mathbf{h})$ and $U^*\mathbf{h}U=\hat{x}$. Because C commutes with $E(\sigma)$ and $Cg=g$, we also have that $U(L_r^2)=\mathcal{D}_+$ where

$$L_r^2=\{a\in L^2(\mathbb{R}_+,\mathcal{B},\rho): \overline{a}=a\}.$$

Thus for $a_k\in L_r^2$

$$U\left(\sum_{k=1}^{n} e^{it_k\hat{x}} a_k\right)=\sum_{k=1}^{n} e^{it_kU\hat{x}U^*}(Ua_k)=\sum_{k=1}^{n} e^{it_k\mathbf{h}} f_k,$$

where $f_k=Ua_k$. So it is enough to prove the statement for $\hat{x}$.

Let K be an arbitrary compact set $K\subset(0,\infty)$. Let us consider the family of continuous, real-valued functions on K given by

$$\left\{ a: a=\sum_{k=1}^{n} \sin(t_k\hat{x})a_k, n\in\mathbb{N}, a_k\in C_r(K) \right\}.$$

It is clear that this family is an algebra which separates points in K and does not vanish identically on some point in K. By Stone–Weierstrass theorem this family is dense in $C_r(K)$. Now let $f\in C(K)$. Then $f=f_1+if_2, f_1,f_2\in C_r(K)$. At first we find a finite sum $\Sigma_{k=1}^n \sin(t_k\hat{x})a_k, a_k\in C_r(K)$, such that

$$\left\| f_2-\sum_{k=1}^{n} \sin(t_k\hat{x})a_k \right\|_{\sup} < \epsilon.$$

Next put $t_0=0$, and $a_0=f_1-\Sigma_{k=1}^n \cos(t_k\hat{x})a_k$. Then

$$\left\| f-\sum_{k=0}^{n} e^{it_k\hat{x}} a_k \right\|_{\sup} = \left\| f_2-\sum_{k=1}^{n} \sin(t_k\hat{x})a_k \right\|_{\sup} < \epsilon.$$

Because ρ is a probability measure on $\mathbb{R}_+$, we have also that

$$\left\| f-\sum_{k=0}^{n} e^{it_k\hat{x}} a_k \right\|_{L^2} < \epsilon.$$

But $\mathbb{R}_+$ is locally compact and ρ is a Borel measure, so for every $\varphi\in L^2(\mathbb{R}_+,\mathcal{B},\rho)$ there is a compact set $K\subset\mathbb{R}_+$ and a function $f\in C(K)$ such that $\|\varphi-f\|_{L^2}<\epsilon$. Thus the proof of Lemma 3.2 is finished.

Theorem 3.5: *For quasi-free KMS state ω defined on the Weyl algebra $\mathfrak{W}(\mathcal{D},\sigma)$ the canonical modular structure $(\pi_\omega(\mathfrak{W}(\mathcal{D},\sigma))'',\Delta_\omega,J_\omega)$ is unitarily equivalent to the modular structure $(\mathfrak{M}_\xi,\Delta_\xi,J_\xi)$ constructed from the stochastic process ξ_s^β.*

Proof: Let $w(f)$ denote the following elements of $L^\infty(Q,\Sigma_0,\mu)$:

$$w(f)=e^{i\langle\xi_0^\beta,f\rangle}.$$

On a dense set of vectors in $\mathfrak{H}_\xi$ generated by elements

$$e^{it_1H_\xi}\pi_0(w(f_1))e^{-it_1H_\xi}\cdots e^{it_nH_\xi}\pi_0(w(f_n))e^{-t_nH_\xi}\Omega_\xi,\quad t_1,\ldots,t_n\in\mathbb{R},f_1,\ldots,f_n\in\mathcal{D}_+,$$

we define a map V by

$$\begin{aligned}&V(e^{it_1H_\xi}\pi_0(w(f_1))e^{-it_1H_\xi}\cdots e^{it_nH_\xi}\pi_0(w(f_n))e^{-it_nH_\xi}\Omega_\xi)\\&\quad=e^{it_1H_\omega}\pi_\omega(W_{f_1})e^{-it_1H_\omega}\cdots e^{it_nH_\omega}\pi_\omega(W_{f_n})e^{-it_nH_\omega}\Omega_\omega.\end{aligned}$$

Since the functions

$$\phi_\xi(t_1,\ldots,t_n)=\langle\Omega_\xi,e^{it_1H_\xi}\pi_0(w(f_1))e^{-it_1H_\xi}\cdots e^{it_nH_\xi}\pi_0(w(f_n))e^{-it_nH_\xi}\Omega_\xi\rangle$$

and

$$\phi_\omega(t_1,\ldots,t_n)=\langle\Omega_\omega,e^{it_1H_\omega}\pi_\omega(W_{f_1})e^{-it_1H_\omega}\cdots e^{it_nH_\omega}\pi_\omega(W_{f_n})e^{-it_nH_\omega}\Omega_\omega\rangle$$

can be analytically continued to $\mathfrak{T}_n^\beta$ and coincide on imaginary times, they are equal for all $t_1,\ldots,t_n\in\mathbb{R}$. Thus V is an isometry with dense domain in $\mathfrak{H}_\xi$. But

$$e^{it_1H_\omega}\pi_\omega(W_{f_1})e^{-it_1H_\omega}\cdots e^{it_nH_\omega}\pi_\omega(W_{f_n})e^{-it_nH_\omega}=\pi_\omega(W_{\Sigma_{k=1}^n e^{it_k\mathbf{h}}f_k})\prod_{l\leqslant k<l\leqslant n}e^{-(i/2)\sigma(f_k,e^{i(t_l-t_k)\mathbf{h}}f_l)},$$

hence the range of V is dense in $\mathfrak{H}_\omega$ by Theorem 3.4 and continuity of $f\to\omega(W_f)$ with respect to the Hilbert space norm on $\mathcal{D}$. Moreover, $V\Omega_\xi=\Omega_\omega$. By direct computation we show that

$$V\pi_0(w(f))V^*=\pi_\omega(W_f),\quad f\in\mathcal{D}_+;\quad Ve^{itH_\xi}V^*=e^{itH_\omega}$$

and

$$V\left(\prod_{k=1}^m e^{it_kH_\xi}\pi_0(w(f_k))e^{-it_kH_\xi}\right)V^*=\prod_{k=1}^m e^{it_kH_\omega}\pi_\omega(W_{f_k})e^{i-it_kH_\omega}.$$

Thus

$$V\mathfrak{M}_\xi V^*=\pi_\omega(\mathfrak{W}(\mathcal{D},\sigma))''$$

and

$$VJ_\xi V^*=J_\omega,\quad V\Delta_\xi V^*=\Delta_\omega.$$

F. General case

If the operator **h** is only bounded from below, that is

$$\mathbf{h}\geqslant\mu\mathbb{I},\quad \mu<0,$$

then we can still obtain a quasi-free KMS state on the whole algebra $\mathfrak{W}(\mathcal{D},\sigma)$ if we replace **h** by $\tilde{\mathbf{h}}=\mathbf{h}-(\mu-1)\mathbb{I}$, for example. Then the operator B will be equal to

$$B=\frac{\mathbb{I}+ze^{-\beta\mathbf{h}}}{\mathbb{I}-ze^{-\beta\mathbf{h}}},$$

where $z=e^{\beta(\mu-1)}$ and B will be bounded. But at the same time the dynamics α_t have to be replaced by $\tilde{\alpha}_t$ given by

$$\tilde{\alpha}_t(W_f)=W_{e^{it\tilde{\mathbf{h}}}f}=W_{z^{-it/\beta}T_tf}.$$

In some cases it is necessary to consider the operator $\mathbf{h}$ itself. Then we have to restrict $\mathfrak{W}(\mathcal{D},\sigma)$ to a Weyl algebra over a suitable subspace of the Hilbert space $\mathcal{D}$.

Theorem 3.6: *Suppose that* $\mathbf{h}\geqslant 0$ *and* Ker $\mathbf{h}=0$. *Let* $B=(\mathbb{I}+e^{-\beta\mathbf{h}})^{-1}(\mathbb{I}+e^{-\beta\mathbf{h}})$ *and*

$$\mathbb{B}(f,g)=\langle B^{1/2}f,B^{1/2}g\rangle \quad with \quad D(\mathbb{B})=D(B^{1/2}).$$

Let us define

$$\omega(W_f)=e^{-(1/4)\mathbb{B}(f,f)} \quad \text{for} \quad f\in D(\mathbb{B})$$

and extend it onto the Weyl algebra $\mathfrak{W}(D(\mathbb{B}),\sigma)$. *Then* ω *is an* α_t*-KMS state at the inverse temperature* β *for the dynamics* α_t *given by* $T_t=e^{it\mathbf{h}}$.

Proof: First let us check that α_t maps $\mathfrak{W}(D(\mathbb{B}),\sigma)$ into itself. But since T_t commutes with spectral projectors of $\mathbf{h}$,

$$T_t:D(B^{1/2})\to D(B^{1/2}),$$

hence α_t leaves $\mathfrak{W}(D(\mathbb{B}),\sigma)$ invariant. Now consider two-point Green's function $G^{(2)}(f,g;t)$ given by (3.2) for $f,g\in D(\mathbb{B})$. It is clear that this function is continuous and bounded. Hence to show the existence of an analytic function interpolating between $G^{(2)}(f,g;t)$ and $G^{(2)}(g,f;-t)$ in the strip $\mathbb{R}\times i[0,\beta]$ it is enough to check that

$$\mathcal{F}(\mathbb{T}_{F(f,g;\cdot)})=e^{-\beta p}\mathcal{F}(\mathbb{T}_{F(g,f;\cdot)}).$$

For $f=g$ it can be done by direct calculations similarly as in the proof of Theorem 3.1. The general result follows from the polarization formula. It is clear that such an analytic function exists for any

$$a_n=\sum_{k=1}^{n}c_kW_{f_k},\quad b_n=\sum_{k=1}^{n}d_kW_{g_k}$$

in $\mathfrak{W}(D(\mathbb{B}),\sigma)$. Now for arbitrary $a,b\in\mathfrak{W}(D(\mathbb{B}),\sigma)$

$$G^{(2)}(a,b;t)=\lim_{n\to\infty}G^{(2)}(a_n,b_n;t)$$

and this function is continuous and bounded, so

$$\mathbb{T}_{G^{(2)}(a,b;\cdot)}=\lim_{n\to+\infty}\mathbb{T}_{G^{(2)}(a_n,b_n;\cdot)}$$

in $\mathcal{S}'$. But since $\mathcal{F}$ is a homeomorphism of $\mathcal{S}'$

$$\mathcal{F}(\mathbb{T}_{G^{(2)}(a,b;\cdot)})(-p)=e^{-p\beta}\mathcal{F}(\mathbb{T}_{G^{(2)}(b,a;\cdot)})(p).$$

Remark: In the above case the state ω is not $\mathcal{D}$-continuous in the Hilbert space topology.

Conversely, we show that among quasi-free gauge-invariant states this is the only possibility.

Theorem 3.7: *Suppose that* $\mathbb{B}$ *is a positive, closed sesquilinear form on* $\mathcal{D}$ *with domain* $D(\mathbb{B})$. *Let* $\mathbf{h}\geqslant 0$ *and* Ker $\mathbf{h}=0$. *If* $e^{it\mathbf{h}}$: $D(\mathbb{B})\to D(\mathbb{B})$ *and a state* ω *determined by* $\mathbb{B}$ *by formula* (4.1) *is an* α_t*-KMS state at inverse temperature* β, *defined on the Weyl algebra* $\mathcal{W}(D(\mathbb{B}),\sigma)$, *then*

$$\mathbb{B}(f,g)=\langle B^{1/2}f,B^{1/2}g\rangle \quad \text{where } B=\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Proof: Let $f\in D(\mathbb{B})$. Then

$$\int_0^\infty d\langle (B+\mathbb{I})f,E(\lambda)f\rangle\varphi(-\lambda)+\int_0^\infty d\langle (B-\mathbb{I})f,E(\lambda)f\rangle\varphi(\lambda)$$
$$=\int_0^\infty d\langle (B+\mathbb{I})f,E(\lambda)f\rangle e^{-\beta\lambda}\varphi(\lambda)+\int_0^\infty d\langle (B-\mathbb{I})f,E(\lambda)f\rangle e^{\beta\lambda}\varphi(-\lambda),$$

where $E(\lambda)$ are spectral projectors of $\mathbf{h}$. Suppose that $\varphi>0$ for every $x\in\mathbb{R}$. Then

$$\langle (B-\mathbb{I})f,\varphi(\mathbf{h})f\rangle=\langle (B+\mathbb{I})f,e^{-\beta\mathbf{h}}\varphi(\mathbf{h})f\rangle.$$

By polarization we get

$$\varphi(\mathbf{h})(B-\mathbb{I})f=\varphi(\mathbf{h})e^{-\beta\mathbf{h}}(B+\mathbb{I})f.$$

But $\varphi(\mathbf{h})$ is injective, hence

$$(B-\mathbb{I})f=e^{-\beta\mathbf{h}}(B+\mathbb{I})f.$$

So

$$(\mathbb{I}-e^{-\beta\mathbf{h}})Bf=(\mathbb{I}=e^{-\beta\mathbf{h}})f.$$

It implies that $f\in D((\mathbb{I}-e^{-\beta\mathbf{h}})^{-1}(\mathbb{I}+e^{-\beta\mathbf{h}}))$ and

$$\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}f=Bf.$$

Thus

$$B\subset\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Because both of them are self-adjoint, $D(B)=D((\mathbb{I}+e^{-\beta\mathbf{h}})^{-1}(\mathbb{I}-e^{-\beta\mathbf{h}}))$ and

$$B=\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Remark: The results of Secs. III C and D can also be extended to this case. In particular, the functions $S(f,g;s)$ restricted to $D(\mathbb{B})_+=\{f\in D(\mathbb{B}):Cf=f\}$ define periodic Gaussian OS-positive stochastic process indexed by $D(\mathbb{B})_+$, where $D(\mathbb{B})_+$ is the real Hilbert space with respect to the inner product

$$\langle f,g\rangle_B=\langle B^{1/2}f,B^{1/2}g\rangle.$$

Similarly, since $e^{it\mathbf{h}}$ is unitary with respect to the inner product $\langle\cdot,\cdot\rangle_B$, Theorem 3.5 can be applied to this case. Moreover, the mapping

$$f\to\omega(W_f)$$

is continuous with respect to the norm $\|\cdot\|_B=\langle\cdot,\cdot\rangle_B^{1/2}$. Thus also in this case the process ξ_s^β determines the modular structure associated with the state ω. Similarly, one can show that ξ_s^β has a version with two-sided Markov property on K_β.

G. Ground state process

Now consider a quasi-free state ω_0 which is a ground state with respect to the evolution α_t defined by $T_t=e^{it\mathbf{h}}$. Then, for every $f,g\in\mathcal{D}$ there exists a function $\Phi_0(f,g;z)$, analytic and bounded for Im $z>0$, continuous on Im $z\geqslant 0$ and such that

$$\Phi_0(f,g;t)=F(f,g;t)$$

for all $t\in\mathbb{R}$. We know that for the ground state case, supp $\mathcal{F}(\mathbb{T}_F)\subseteqq[0,\infty)$. Similar arguments as used in the proof of Theorem 3.1 lead to the conclusion that for ω_0,

$$F(f,g;t)=\langle f,e^{it\mathbf{h}}g\rangle$$

and $\mathbf{h}\geqslant 0$. By analytic continuation to the imaginary times we obtain the functions

$$S_0(f,g;s)=\Phi_0(f,g;is) \tag{3.13}$$

defined for $s\geqslant 0$. Two-point imaginary time Green's functions are now given by

$$G_0^{E,2}(f,g;s)=e^{-(1/4)(\mathbb{B}(f,f)+\mathbb{B}(g,g)-(1/2)S_0(f,g;s)}.$$

Proposition 3.6: $S_0(f,g;s)$ defined by (3.13) satisfies the following.

(1) For every $f\in\mathcal{D}$ and $s\in[0,\infty)$, $S(f,f;s)$ is OS-positive.
(2) More generally, for every terminating sequences $f_k\in\mathcal{D}$, $s_k\in[0,\infty)$, $c_k\in\mathbb{C}$,

$$\sum_{k,l}\overline{c_k}c_lS_0(f_k,f_l;s_k+s_l)\geqslant 0.$$

Proof: Since the function

$$t\to F(f,f;t)=\langle f,e^{it\mathbf{h}}f\rangle$$

is positive definite, $S_0(f,f;s)$ is OS-positive as a Laplace transform of positive measure. The general positivity condition follows by polarization.

Similarly as in the nonzero temperature case, the existence of complex conjugation C defining the Abelian splitting $(\mathcal{D}_+,\mathcal{D}_-)$ of $\mathcal{D}$, commuting with $\mathbf{h}$ is necessary for stochastic positivity of $S_0(f,g;s)$ restricted to $\mathcal{D}_+$. Adopting the arguments from Sec. III C to the ground state case we can also prove the following Proposition.

Proposition 3.7: Let $\mathbf{h}$ be C-real. Then functions $S_0(f,g;s)$ restricted to $\mathcal{D}_+$ satisfy the following.

(1) $S_0(f,g;s)=S_0(g,f;s)$, $s\in[0,\infty)$.

(2) $S_0(f,g;s)$ can be extended to the function of $s\in\mathbb{R}$ (denoted by the same symbol) such that for all terminating sequences $f_1,\ldots,f_n\in\mathcal{D}_+$; $s_1,\ldots,s_n\in\mathbb{R}$ and $c_1,\ldots,c_n\in\mathbb{C}$,

$$\sum_{k,l=1}^{n}\overline{c_k}c_lS_0(f_k,f_l;s_k-s_l)\geqslant 0.$$

(3) $S_0(f,g;s)$ defines an operator-valued covariance function $R_\infty(s)$ of Gaussian OS-positive stochastic process indexed by $\mathcal{D}_+$.

Let ξ_t^∞ be the Gaussian process indexed by $\mathcal{D}_+$ defined by

$$\mathbb{E}(\langle\xi_t^\infty,f\rangle)=0;\quad \mathbb{E}(\langle\xi_{t_1}^\infty,f_1\rangle\langle\xi_{t_2}^\infty,f_2\rangle)=\tfrac{1}{2}\langle f_1,R_\infty(t_2-t_1)f_2\rangle.$$

If, in addition, Ker $\mathbf{h}=\{0\}$, then as in the KMS state case, the process ξ_t^∞ corresponding to the ground state completely determines the ground state structure.

Theorem 3.8: *Let ξ_t^∞ be a Gaussian, OS-positive stochastic process indexed by $\mathcal{D}_+\times\mathbb{R}$. Then there exists a Hilbert space $\mathfrak{H}_\xi^\infty$ with a unit vector Ω_ξ^∞, a weakly continuous one-parameter group of unitary operators $U_\xi^\infty(t)=e^{itH_\xi^\infty}$ with $H_\xi^\infty\geqslant 0$, Ker $H_\xi^\infty=\{0\}$ and a von Neumann algebra $\mathfrak{M}_\xi^\infty$ of*

operators acting on $\mathfrak{H}_\xi^\infty$ *such that* Ω_ξ^∞ *is cyclic for* $\mathfrak{M}_\xi^\infty$. *Moreover, the canonical ground state structure* $(\pi_{\omega_0}(\mathfrak{W}(\mathcal{D},\sigma))'', e^{itH_{\omega_0}}, \Omega_{\omega_0})$ *reconstructed from the state* ω_0 *is unitarily equivalent to the ground state structure* $(\mathfrak{M}_\xi^\infty, e^{itH_\xi^\infty}, \Omega_\xi^\infty)$.

Remark: In the case $\mathbf{h}>0$, there is the following relation between covariance R_β and R_∞. If $\beta\to\infty$, then

$$R_\beta(s)\to R_\infty(s)$$

weakly. On the other hand,

$$R_\beta(s)=\sum_{n\in\mathbb{Z}} R_\infty(s+n\beta),$$

where the series on the right-hand side is weakly convergent.

IV. EXAMPLES

A. Ground states and KMS states for quantum fields on a stationary space–time

Let $(\mathcal{M},g)$ be a stationary space–time, i.e., $(\mathcal{M},g)$ is space and time orientable with a global timelike Killing vector field X. Thus, $(\mathcal{M},g)$ can always be realized as $(\mathbb{R}\times\mathcal{C},g)$ where $(\mathcal{C},\hat{g})$ is a Riemannian three-manifold and

$$g=\begin{pmatrix} a^2-b^ib_i & -b_i \\ -b_i & -\hat{g}_{ij}\end{pmatrix}$$

with a scalar field a (laps field) and a vector field b (shift field) satisfying

$$a>0,\quad a^2-\hat{g}(b,b)>0,$$

and with the Killing vector field

$$X:=\frac{\partial}{\partial t}=aN(\mathcal{C})+b,$$

where $N(\mathcal{C})$ is a unit future-pointing normal vector field to $\mathcal{C}$. If $(\mathcal{M},g)$ is globally hyperbolic, then $\mathcal{C}$ can be chosen to be a global Cauchy surface. On $(\mathcal{M},g)$ we consider the covariant Klein–Gordon equation

$$(g^{\mu\nu}\nabla_\mu\nabla_\nu+m^2+V)\varphi=0.$$

Given some Cauchy surface $\mathcal{C}$, let

$$D(\mathcal{C})=C_0^\infty(\mathcal{C})+C_0^\infty(\mathcal{C})$$

be the space of real smooth Cauchy data of compact support. Then, by the Leray theorem,[37] the Cauchy data $\Phi\in D(\mathcal{C})$ given by

$$\Phi=\begin{pmatrix} f \\ p\end{pmatrix}$$

define a unique solution φ of the Klein–Gordon equation having compact support on every other Cauchy surface and such that

$$\varphi|c=f,\quad N(\mathcal{C})\varphi|c=p.$$

Thus, we may view time evolution as a one-parameter group

$$T_t: D(\mathcal{C})\to D(\mathcal{C}).$$

Moreover, T_t preserves the symplectic form

$$\hat{\sigma}(\Phi_1,\Phi_2)=\int_{\mathcal{C}}(f_1p_2-p_1f_2)\ d\eta(\mathcal{C}),$$

where $\eta(\mathcal{C})$ is the Riemannian volume element on $(\mathcal{C},\hat{g})$ and

$$\frac{d}{dt}\,T_t\Phi|_{t=0}=-\hat{h}\Phi$$

with $\hat{h}=-\mathbf{gA}$ and

$$\mathbf{A}=\begin{pmatrix}-(\partial^i a)\partial_i+a(m^2-\Delta(\mathcal{C})+V) & -(\nabla_i b^i+b^i\partial_i)\\ b^i\partial_i & a\end{pmatrix},\quad \mathbf{g}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix},$$

where $\Delta(\mathcal{C})$ is the Laplace–Beltrami operator on $(\mathcal{C},\hat{g})$ and ∇_i is the covariant derivative on $(\mathcal{C},\hat{g})$.

In order to apply the results of Sec. III we need a Hilbert space $(\mathcal{D},\langle,\rangle_{\mathcal{D}})$ containing $D(\mathcal{C})$, such that $\hat{\sigma}(\cdot,\cdot)=\mathrm{Im}\langle\cdot,\cdot\rangle_{\mathcal{D}}$ and $T_t=e^{it\mathbf{h}}$, where $\mathbf{h}$ is self-adjoint on $\mathcal{D}$. The result of Kay[38] shows that such *a one-particle Hilbert space structure* exists under the following (mass gap) assumptions:

(1) V is stationary and $\inf V(x)+m^2>0$,
(2) $\inf a>0$ on $\mathcal{C}$, and
(3) $\inf(a-b^ib_i/a)>0$ on $\mathcal{C}$.

The above assumptions also imply that the generator $\mathbf{h}$ has a bounded inverse.

Proposition 4.1: Let $(\mathcal{D},e^{it\mathbf{h}})$ be the one-particle Hilbert space structure corresponding to the Klein–Gordon equation on a globally hyperbolic stationary space–time $(\mathcal{M},g)$. There exists a complex conjugation C on $\mathcal{D}$ such that $\mathbf{h}$ is C-real.

Proof: To show the existence of complex conjugation C we quote basic steps of the Kay construction. Since

$$\langle\Phi,\mathbf{A}\Phi\rangle_{L^2+L^2}>\epsilon\|\Phi\|^2_{L^2+L^2}$$

for some $\epsilon>0$, we define the $\mathbf{A}$-norm

$$\|\Phi\|^2_{\mathbf{A}}=\langle\Phi,\mathbf{A}\Phi\rangle_{L^2+L^2}$$

and $D_{\mathbf{A}}$ as the completion of $D(\mathcal{C})$ in this norm. Then, it can be shown that $\hat{h}$ is essentially skew-adjoint and $(\overline{\hat{h}})^{-1}$ exists and is bounded. Since $\hat{\sigma}$ is continuous in $\mathbf{A}$-norm, it can be extended to $\hat{\sigma}'$ on $D_{\mathbf{A}}$ and

$$\hat{\sigma}'(\Phi,\Psi)=\langle\Phi,(\overline{\hat{h}})^{-1}\Psi\rangle_{\mathbf{A}}.$$

Let $D^{\mathbb{C}}_{\mathbf{A}}$ be the natural complexification of the real Hilbert space $D_{\mathbf{A}}$. We change the inner product in $D^{\mathbb{C}}_{\mathbf{A}}$ by modifying the complex structure on $D^{\mathbb{C}}_{\mathbf{A}}$. To do this we introduce a unitary operator I satisfying $I^2=-1$ and $\hat{\sigma}'(\Phi,I\Phi)>0$. Since $i\overline{\hat{h}}$ is self-adjoint on $D^{\mathbb{C}}_{\mathbf{A}}$ we can define

$$I=i(P_+-P_-)=|\overline{\hat{h}}|^{-1}\overline{\hat{h}},$$

where P_+,P_- are projectors onto positive and negative parts of its spectrum. Now we define the new inner product by

$$\langle\Phi,\Psi\rangle_{\mathcal{D}}=\hat{\sigma}'(\Phi,I\Psi)+i\hat{\sigma}'(\Phi,\Psi).$$

Let $\mathcal{D}$ be the Hilbert space completion of $D_{\mathbf{A}}^{\mathbb{C}}$ with respect to $\langle,\rangle_{\mathcal{D}}$. From the construction of $\mathcal{D}$ we see that T_t is the unitary group with respect to the inner product $\langle,\rangle$, and the generator $\mathbf{h}$ of it is strictly positive. The complex conjugation C in $\mathcal{D}$ is defined as follows. Let C_0 be the complex conjugation in $D_{\mathbf{A}}^{\mathbb{C}}$ with respect to the complex structure given by multiplication by i. Then

$$C = C_0(P_+ - P_-)$$

is the complex conjugation in $\mathcal{D}$ such that $\mathbf{h}$ is C-real.

Applying the results of Sec. III we have the following.

Corollary 4.1:

(1) Let ω be the KMS state at the inverse temperature β, defined on $\mathfrak{W}(\mathcal{D},\sigma)$ as in Theorem 3.1, and let ξ_t^β be the corresponding Gaussian Markov thermal process indexed by $\mathcal{D}_+$ (where $\mathcal{D}_+$ is defined by the complex conjugation C). The modular structure defined by ω is unitarily equivalent to the modular structure reconstructed from ξ_t^β as in Theorem 3.5.

(2) Similarly, let ω_0 be the ground state on $\mathfrak{W}(\mathcal{D},\sigma)$ and let ξ_t^∞ be the corresponding ground state process. The ground state structure defined by ω_0 is unitarily equivalent to the ground state structure reconstructed from the process ξ_t^∞ as in Theorem 3.8.

In the case of *a static* space–time, we can realize $(\mathcal{M},g)$ as above, but with the vector field $b=0$. Then the matrix $\mathbf{A}$ becomes diagonal

$$\mathbf{A} = \begin{pmatrix} A & 0 \\ 0 & a \end{pmatrix}$$

with

$$A = -(\partial^i a)\partial_i + a(m^2 - \Delta(\mathcal{C}) + V.$$

In this case the one-particle structure can be described in a more explicit way. As was shown by Kay,[38] $\mathcal{D}$ can be identified with $L^2(\mathcal{C},d\eta)$ and the space $D(\mathcal{C})$ of Cauchy data is mapped into $L^2(\mathcal{C},d\eta)$ by

$$(f,p) \to Yf + IY^{*-1}p,$$

where

$$Y = (\overline{a}^{1/2}\overline{A}\overline{a}^{1/2})^{-1/4}\overline{a}^{1/2}$$

and $\overline{B}$ means closure of operator B. The one-particle Hamiltonian $\mathbf{h}$ is now given by

$$\mathbf{h} = (\overline{a}^{1/2}\overline{A}\overline{a}^{1/2})^{1/2},$$

and the complex conjugation C is the natural one in $L^2(\mathcal{C},d\eta)$. Then $\mathcal{D}_+$ is defined as a Hilbert space completion in $L^2(\mathcal{C},d\eta)$ of the linear space generated by vectors

$$Yf, \quad f \in C^\infty_{0,\mathbb{R}}(\mathcal{C}),$$

and, similarly, $\mathcal{D}_-$ is the completion in $L^2(\mathcal{C},d\eta)$ of linear space generated by

$$Y^{*-1}p, \quad p \in C^\infty_{0,\mathbb{R}}(\mathcal{C}).$$

From the general results of Sec. III it follows that there exist thermal (respectively, ground state) processes indexed by $\mathcal{D}_+$ and $\mathcal{D}_-$. The former we call *field thermal* (respectively, *ground state*) *process* and denote by ξ_t^β (respectively ξ_t^∞) and the latter we call *momentum thermal* (respectively *ground state*) *process* and denote by π_t^β (respectively π_t^∞). It is enough to consider the covariances of these processes for the indexes of the form Yf or $Y^{*-1}p$ with $f,p \in C_0^\infty(\mathcal{C})$. Thus for thermal processes we obtain

$$\mathbb{E}(\langle \xi_t^\beta, f\rangle\langle \xi_s^\beta, g\rangle) = \int_{\mathcal{C}\times\mathcal{C}} (Yf)(x) R^\beta(t,s;x,y)(Yg)(y)\, d\eta(x)\, d\eta(y),$$

$$\mathbb{E}(\langle \pi_t^\beta, f\rangle\langle \pi_s^\beta, g\rangle) = \int_{\mathcal{C}\times\mathcal{C}} (Y^{*-1}f)(x) R^\beta(t,s;x,y)(Y^{*-1}g)(y)\, d\eta(x)\, d\eta(y),$$

where

$$R^\beta(t,s;x,y) = \frac{e^{-|t-s|\mathbf{h}} + e^{-(\beta-|t-s|)\mathbf{h}}}{1-e^{-\beta\mathbf{h}}}(x,y)$$

for $|t-s| \leqslant \beta$. Similarly, for the ground state processes,

$$\mathbb{E}(\langle \xi_t^\infty, f\rangle\langle \xi_s^\infty, g\rangle) = \int_{\mathcal{C}\times\mathcal{C}} (Yf)(x) R^\infty(t,s;x,y)(Yg)(y)\, d\eta(x)\, d\eta(y),$$

$$\mathbb{E}(\langle \pi_i^\infty, f\rangle\langle \pi_s^\infty, g\rangle) = \int_{\mathcal{C}\times\mathcal{C}} (Y^{*-1}f)(x) R^\infty(t,s;x,y)(Y^{*-1}g)(y)\, d\eta(x)\, d\eta(y),$$

where

$$R^\infty(t,s;x,y) = e^{-|t-s|\mathbf{h}}(x,y).$$

Let $dE_{\mathbf{h}}(\lambda)$ be the spectral resolution of the operator $\mathbf{h}$ in the Hilbert space $L^2(\mathcal{C}, d\eta)$. Then

$$\mathbb{E}(\langle \xi_l^\beta, f\rangle\langle \xi_0^\beta, g\rangle) = \int_0^\infty \frac{e^{-t\lambda} + e^{-(\beta-t)\lambda}}{\lambda(1-e^{-\beta\lambda})}\, d\langle \overline{a}^{1/2} f, E_{\mathbf{h}}(\lambda)\overline{a}^{1/2} g\rangle.$$

Therefore, if f is such that $\overline{a}^{1/2} f \in L^2(\mathcal{C}, d\eta)$, then

$$m(f) = \int_0^\infty \frac{e^{-t\lambda} + e^{-(\beta-t)\lambda}}{1-e^{-\beta\lambda}}\, d\langle \overline{a}^{1/2} f, E_{\mathbf{h}}(\lambda)\overline{a}^{1/2} f\rangle < \text{const} \|\overline{a} f\|^2_{L^2(\mathcal{C}, d\eta)},$$

and we can apply Proposition 3.6 to conclude that for every such f there exists a version of the coordinate process $\langle \xi_t^\beta, f\rangle$ with Hölder continuous paths. Similar arguments work also for the ground state (field) process and momentum processes. Thus we obtain the following.

Corollary 4.2: Let $(\mathcal{M}, g)$ be a static, globally hyperbolic space–time with the laps field a. Then

(1) For any $f \in C^\infty_{0,\mathbb{R}}(\mathcal{C})$ such that $\overline{a}^{1/2} f \in L^2(\mathcal{C}, d\eta)$ and $\beta \in (0,\infty]$ there exists a version of the coordinate process $\langle \xi_t^\beta, f\rangle$ such that for every $\gamma \in (0,1/2)$ there is an integrable random variable $d(f,\gamma)$ such that

$$|\langle \xi_t^\beta, f\rangle - \langle f_s^\beta, f\rangle| \leqslant d(f,\gamma)|t-s|^\gamma$$

with probability one.

(2) For any $f \in C^\infty_{0,\mathbb{R}}(\mathcal{C})$ such that $\overline{a}^{1/2} f \in H_{1/2}(\mathcal{C}, d\eta)$ where

$$H_{1/2}(\mathcal{C}, d\eta) = \text{metric completion of } C^\infty_{0,\mathbb{R}}(\mathcal{C}) \text{ with respect to (Sobolev-like) norm,}$$

$$\|f\|^2_{1/2} = \int_{\mathcal{C}\times\mathcal{C}} (\mathbf{h}^{1/2} f)(x)(\mathbf{h}^{1/2} f)(y)\, d\eta(x)\, d\eta(y),$$

the coordinate process $\langle \pi_t^\beta, f\rangle$ for $\beta \in (0,\infty]$ has a version with Hölder continuous paths similarly as in (1).

Example 4.1.1: Let $\mathcal{M}_d$ be a flat d-dimensional Minkowski space–time with the metric tensor

$$(g_{\mu\nu})=\begin{pmatrix}1 & 0\\ 0 & -\mathbb{I}_{d-1}\end{pmatrix}.$$

Taking

$$\mathcal{C}=\{(x^0,\mathbf{x})\in\mathcal{M}_d : x^0=0\}$$

and $V=0$, we obtain the following covariances of thermal and ground state (field) processes:

$$\mathbb{E}(\langle\xi_t^\beta,f\rangle\langle\xi_s^\beta,g\rangle)=\int_{\mathbb{R}^{d-1}}\frac{d\mathbf{p}}{\sqrt{\mathbf{p}^2+m^2}}\,\hat f(\mathbf{p})\hat g(\mathbf{p})\,\frac{e^{-|t-s|\sqrt{\mathbf{p}^2+m^2}}+e^{-(\beta-|t-s|)\sqrt{\mathbf{p}^2+m^2}}}{1-e^{-\beta\sqrt{\mathbf{p}^2+m^2}}}$$

for $|t-s|\leqslant\beta$ and

$$\mathbb{E}(\langle\xi_t^\infty,f\rangle\langle\xi^{\infty}{}_s,g\rangle)=\int_{\mathbb{R}^{d-1}}\frac{d\mathbf{p}}{\sqrt{\mathbf{p}^2+m^2}}\,\hat f(\mathbf{p})\hat g(\mathbf{p})e^{-|t-s|\sqrt{\mathbf{p}^2+m^2}},$$

where $f,g\in C_0^\infty(\mathbb{R}^{d-1})$. And similarly for momentum processes:

$$\mathbb{E}(\langle\pi_t^\beta,f\rangle\langle\pi_s^\beta,g\rangle)=\int_{\mathbb{R}^{d-1}}d\mathbf{p}\sqrt{\mathbf{p}^2+m^2}\hat f(\mathbf{p})\hat g(\mathbf{p})\,\frac{e^{-|t-s|\sqrt{\mathbf{p}^2+m^2}}+e^{-(\beta-|t-s|)\sqrt{\mathbf{p}^2+m^2}}}{1-e^{-\beta\sqrt{\mathbf{p}^2+m^2}}},$$

$$\mathbb{E}(\langle\pi_t^\infty,f\rangle\langle\pi_s^\infty,g\rangle)=\int_{\mathbb{R}^{d-1}}d\mathbf{p}\sqrt{\mathbf{p}^2+m^2}\hat f(\mathbf{p})\hat g(\mathbf{p})e^{-|t-s|\sqrt{\mathbf{p}^2+m^2}}.$$

The law of the process ξ_t^β is given by the Gaussian random field μ_0^β indexed by $\mathcal{S}(K_\beta\times\mathbb{R}^{d-1})$ and defined by

$$\mathbb{E}^{\mu_0^\beta}(\langle\varphi,f\rangle)=0\quad\text{and}\quad\mathbb{E}^{\mu_0^\beta}(\langle\varphi,f\rangle\langle\varphi,g\rangle)=\langle f,(-\Delta_d^\beta+m^2)^{-1}g\rangle,$$

where $-\Delta_d^\beta$ denotes the d-dimensional Laplace operator with periodic boundary conditions in the time direction [$-\Delta_d^\infty$ is defined as the Friedrichs extension of $-\Delta$ in $L^2(\mathbb{R}^d)$]. The law of the momentum process π_t^β is given by the Gaussian random field ν_0^β indexed by $\mathcal{S}(K_\beta\times\mathbb{R}^{d-1})$ and defined by

$$\mathbb{E}^{\nu_0^\beta}(\langle\psi,f\rangle)=0\quad\text{and}\quad\mathbb{E}^{\nu_0^\beta}(\langle\psi,f\rangle\langle\psi,g\rangle)=\langle f,(-\Delta_{d-1}+m^2)(-\Delta_d^\beta+m^2)^{-1}g\rangle.$$

In the present case we are able to analyze the properties of continuity of the processes in more detail.

Proposition 4.2: Let $\rho\in L^2(\mathbb{R}^{d-1})\cap C(\mathbb{R}^{d-1})$, $\alpha>(d-2)/2$. Then the field process ξ_t^β has a version realized in the space

$$\bigcap_{\alpha>(d-2)/2}H^\rho_{-\alpha}(\mathbb{R}^{d-1}),$$

where

$$H^\rho_{-\alpha}(\mathbb{R}^{d-1})=\text{metric completion of } C_0^\infty(\mathbb{R}^{d-1}) \text{ in the norm,}$$

$$\|f\|^\rho_{-\alpha}=\left[\int\frac{|\hat f(p)|^2}{(p^2+m^2)^\alpha}\,\rho^2(p)dp\right]^{1/2},$$

and such that for any $\alpha>(d-1)/2$:

(1) If $\beta<\infty$, there exists a Borel measurable function $\Theta_\beta:\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1})\to\mathbb{R}_+$ such that for all $t,s\in K_\beta$ such that $|t-s|<1$

$$\|\xi_t^\beta - \xi_s^\beta\|_{-\alpha}^\rho \leqslant \Theta_\beta\left(\frac{|t-s|}{\ln(1/|t-s|)}\right)^{1/2}.$$

(2) If $\beta=\infty$, then for any $n\in\mathbb{N}$ there exists a Borel measurable function $\Theta_n:\mathcal{S}'(\mathbb{R}^d)\to\mathbb{R}_+$ such that for any $t,s\in[-n,n]$ such that $|t-s|<e^{-1}$

$$\|\xi_t^\infty - \xi_s^\infty\|_{-\alpha}^\rho \leqslant \Theta_n\left(\frac{|t-s|}{\ln(1/|t-s|)}\right)^{1/2}.$$

In the case of momentum process π_t^∞, let ρ be as above. Then the process π_t^∞ is realized in the space

$$\bigcap_{\alpha>d/2} H_{-\alpha}^\rho(\mathbb{R}^{d-1}).$$

Moreover, for any $\alpha>(d+1)/2$, π_t^β has a version with Hölder continuous [in $H_{-\alpha}^\rho(\mathbb{R}^{d-1})$] paths and obeys inequalities similar to (1) and (2).

Sketch of the proof: Let

$$C_0^\beta(x,y)=(-\Delta_d^\beta+m^2)^{-1}(x,y)$$

be the covariance of the Gaussian random field μ_0^β. Let $\{\chi_\epsilon\}$ be the net of functions satisfying: $\chi_\epsilon\in C_0^\infty(\mathbb{R}^{d-1})$, $\chi_\epsilon>0$, $\int_{\mathbb{R}^{d-1}}\chi_\epsilon(x)dx=1$, i.e., $\chi_\epsilon\to\delta$ in $\mathcal{D}'(\mathbb{R}^{d-1})$. Define

$$C_{0,\epsilon}^\beta = C_0^\beta * (\chi_\epsilon\otimes\chi_\epsilon)$$

and let $\mu_{0,\epsilon}^\beta$ be the corresponding Gaussian measure on $(\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1}),\mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra of sets in $\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1})$. For every $t>0$ and $\mu_{0,\epsilon}^\beta$-almost every $\varphi\in\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1})$ we can define the map

$$x\to X_t^\epsilon(x)=\varphi(t,x).$$

Moreover, it can be shown that $X_t^\epsilon(x)$ takes values in the space $H_{-\alpha}^\rho(\mathbb{R}^{d-1})$.[39] Suitable modification of X_t^ϵ gives rise to $L^2(\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1}),\mu_0^\beta|_{H_{-\alpha}^\rho(\mathbb{R}^{d-1})})$ stochastic process and its L^2 limit is easily seen to be the process ξ_t^β. Thus, there exists a version of ξ_t^β (denoted by the same symbol) which is $L^2(\mathcal{S}'(K_\beta\times\mathbb{R}^{d-1}),\mu_0^\beta|_{H_{-\alpha}^\rho(\mathbb{R}^{d-1})})$ stochastic process and which takes values in $H_{-\alpha}^\rho(\mathbb{R}^{d-1})$ for all $\alpha>(d-2)/2$. Notice that for $\alpha>(d-1)/2$

$$\sup_{|t-s|\leqslant r}(\mathbb{E}(\|\xi_t^\beta-\xi_s^\beta\|_{-\alpha}^\rho)^2)^{1/2}\leqslant \text{const}|t-s|^{1/2}$$

and

$$\lim_{t\to s}\mathbb{E}(\|\xi_t^\beta-\xi_s^\beta\|_{-\alpha}^\rho)^2=0.$$

Thus we can apply the continuity criterion due to Preston and Garsia[40,41] as formulated in Theorem 5.1 in Ref. 42. This leads to the Hölder continuity of ξ_t^β in $H_{-\alpha}^\rho(\mathbb{R}^{d-1})$ for all $\alpha>(d-1)/2$ and ρ as above. Similarly, we can prove the analogous properties of the momentum processes.

Remark: For more refined continuity properties and additional references to other continuity properties of the process ξ_t^∞ we refer to Ref. 42. Other aspects of the process ξ_t^β are discussed in Refs. 43 and 35. It is interesting to note the essential difference between Markov properties of the field and momentum processes. For example, if $\beta=\infty$, then the law of ξ_t^∞ fulfills the so-called sharp Markov property,[6] while the law of π_t^∞ satisfies another kind of Markov property discussed in Ref. 44.

Example 4.1.2:

Let

$$W_{\mathrm{R}}=\{(T,X)\in\mathcal{M}_2: T^2-X^2<0,\ X>0\}$$

be the right wedge in two-dimensional Minkowski space–time. In the hyperbolic coordinates (τ,x) defined by

$$T=e^x \sinh\,\tau,\quad X=e^x\cosh\,\tau,$$

the action of the Lorentz boost

$$\Lambda(t)=\begin{pmatrix}\cosh\,t & \sinh\,t\\ \sinh\,t & \cosh\,t\end{pmatrix}$$

becomes

$$\Lambda(t)(\tau,x)=(\tau+t,x).$$

Let

$$\mathcal{C}=\{(\tau,x):\tau=0\}=\mathbb{R}.$$

Then, on

$$\hat{D}_{\mathcal{C}}=\{(f,\hat{p}): f\in C_0^\infty(\mathcal{C}),\hat{p}=e^x p, p\in C_0^\infty(\mathcal{C})\},$$

the symplectic form

$$\hat{\sigma}((f,\hat{p}),(f',\tilde{p}'))=\int_{\mathcal{C}}(f\hat{p}'-\hat{p}f')\,dx$$

is invariant under the induced action of

$$\hat{\Lambda}(t)=e^{-t\mathcal{L}},\quad \text{where}\quad \mathcal{L}=\begin{pmatrix}0 & -1\\ A & 0\end{pmatrix}\quad\text{and}\quad A=-\frac{\partial^2}{\partial x^2}+e^{2x}m^2.$$

Although in this case the mass gap conditions are not fulfilled, a one-particle Hilbert space can also be constructed.[45] $\mathcal{D}$ can be identified with $L^2(\mathbb{R},dx)$ and $\mathbf{h}=A^{1/2}$. The space $\hat{D}(\mathcal{C})$ of Cauchy data is mapped in $L^2(\mathbb{R},dx)$ by

$$(f,\hat{p})\to A^{-1/4}f+iA^{1/4}\hat{p}.$$

As was shown in Sec. III F we can define the thermal state (for any $\beta>0$) on the Weyl algebra over $D(B^{1/2})\subset L^2(\mathbb{R},dx)$, where

$$B=\frac{\mathbb{I}+e^{-\beta\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}.$$

Using the natural complex conjugation on $L^2(\mathbb{R},dx)$ we obtain thermal processes indexed by $D(B^{1/2})_{\pm}$. On a suitable domain in $C_0^\infty(\mathbb{R})$ we can compute the covariance of the field process, denoted now by l_τ^β, for the indexes of the form $A^{-1/4}f$, $f\in C_0^\infty(\mathbb{R})$, and we obtain

$$\mathbb{E}(\langle l_\tau^\beta,f\rangle\langle l_{\tau'}^\beta,g\rangle)=\left\langle f,\ \frac{e^{-|\tau-\tau'|A^{1/2}}+e^{-(\beta-|\tau-\tau'|)A^{1/2}}}{A^{1/2}(\mathbb{I}-e^{-\beta A^{1/2}})}\,g\right\rangle_{L^2(\mathbb{R})}.$$

Let φ be a free scalar massive quantum field on $\mathcal{M}_2$. It is well known that for f real, the field operator $\varphi(f)$ is essentially self-adjoint on D^φ defined by

$$D^\varphi=\mathrm{lh}\{\varphi(f_1)\cdots\varphi(f_n)\Omega_0: f_1,\ldots,f_n\in C_0^\infty(\mathcal{M}_2), n\in\mathbb{N}\},$$

where Ω_0 is the Fock vacuum. For any open $G\subset\mathcal{M}_2$, let $\mathfrak{N}_G$ be the von Neumann algebra generated by

$$e(f)=e^{\overline{i\varphi(f)}} \quad \text{with supp } f\subset G.$$

If $G\subset\mathcal{M}_2$ is such that its causal complement is nonempty, then by the Reeh–Schlieder theorem, Ω_0 is cyclic and separating for $\mathfrak{N}_G$. In particular, we can put $G=W_{\mathrm{R}}$ and, using the Bisognano–Wichmann theorem,[31] we conclude that $(\mathfrak{N}_{W_{\mathrm{R}}},\Omega_0,\Lambda(\tau))$ forms a W^*-KMS system at $\beta=2\pi$. On the other hand, for any $f\in C_0^\infty(\mathbb{R})$, $t\in\mathbb{R}$, the operator $\varphi(\delta_t\otimes f)$ is also essentially self-adjoint on D^φ. Thus we can define the Abelian von Neumann algebra $\mathfrak{U}_R$ generated by

$$e^{\overline{i\varphi(\delta_l\otimes f)}}, \quad \text{with } f=\overline{f} \text{ supp } f\subset\{(x,0):x>0\}.$$

By the direct computation we also obtain

$$\langle\Omega_0,e^{i\varphi(\delta_0\otimes f)}\Lambda(i\tau)e^{i\varphi(\delta_0\otimes g)}\Omega_0\rangle=\mathbb{E}(e^{i\langle l_0^{2\pi},f\rangle}e^{i\langle l_\tau^{2\pi},g\rangle})$$

for $\tau\in[0,2\pi]$, supp f,supp $g\subset\{(x,0):x>0\}$. Moreover, it is easy to show that the modular structure reconstructed from the thermal (boost) process $l_\tau^{2\pi}$ is unitarily equivalent to $(\mathfrak{N}_{W_{\mathrm{R}}},\Omega_0,\Lambda(\tau))$.

Remark: Similar construction of the thermal process can be done in the case of the exterior Schwarzschild right wedge region in the Kruskal space–time $\mathcal{M}\simeq\mathbb{R}^2\times\Omega$ where Ω is some $d-2$-dimensional Riemann manifold with the metric dl_Ω^2. The metric on $\mathcal{M}$ is given by

$$ds^2=32M^3r^{-1}e^{-r/(2M)}(dT^2-dX^2)-r^2dl_\Omega^2,$$

where M is the mass of the black hole and r is the Schwarzschild radius defined by

$$T^2-X^2=(1-r/(2M))e^{r/(2M)}.$$

The exterior Schwarzschild right wedge is defined as

$$\mathcal{R}=\{(T,X,\zeta):X>|T|,r>2M,\zeta\in\Omega\}.$$

The corresponding thermal process has the covariance defined as above, but in terms of the operator A given by

$$A=-\frac{\partial^2}{\partial x^2}+(1-2M/r)(2M/r^3-\Delta_\Omega/r^2+m^2),$$

where Δ_Ω is the Laplace operator on Ω. Detailed analysis of this process together with the construction corresponding to the de Sitter case will be discussed in a separate paper.

B. Nonrelativistic Bose matter

Let $\mathcal{D}=L^2(\mathbb{R}^d)$ and $T_t=e^{it\mathbf{h}}$ where $\mathbf{h}\geqslant 0$, Ker $\mathbf{h}=\{0\}$, commutes with the natural complex conjugation on $L^2(\mathbb{R}^d)$. Then there exist thermal (and ground state) processes indexed by a suitable subspaces of $L^2_{\mathbb{R}}(\mathbb{R}^d)$. In particular, then the covariance operator of the thermal process ξ_τ^β is given by

$$R^\beta(\tau)=\frac{e^{-|\tau|\mathbf{h}}+e^{-(\beta-|\tau|)\mathbf{h}}}{\mathbb{I}-e^{-\beta\mathbf{h}}}$$

for $|\tau|\leqslant\beta$. As in the previous examples the modular structure defined by the canonical KMS state at the inverse temperature β coincides with the modular structure obtained from the thermal process ξ_τ^β. If additionally inf $\sigma(\mathbf{h})>0$, it is easy to see that for any $f\in L^2_{\mathbb{R}}(\mathbb{R}^d)$ the assumption of Proposition 3.4 is satisfied, hence each coordinate process $\langle\xi^\beta,f\rangle$ has a version with Hölder continuous paths.

The case when the operator **h** (called a kinetic energy) is of the form $\mathbf{h}=\mathcal{E}(-i\nabla)$, where $\mathcal{E}$ is some locally bounded, measurable function satisfying $\mathcal{E}(p)\geqslant 0$, is of particular interest to physics. The choice $\mathcal{E}(p)=p^2+\mu,\ \mu>0$, corresponds to the standard kinetic energy,[24] whereas $\mathcal{E}(p)=\sqrt{p^2+m^2}$ corresponds to the semirelativistic kinetic energy. For a given $\mathcal{E}\in\mathcal{B}_0(\mathbb{R}^d)$ where

$$\mathcal{B}_0(\mathbb{R}^d)=\{\text{the set of locally bounded, measurable functions on } \mathbb{R}^d \text{ satisfying } \inf \mathcal{E}(p)=\epsilon>0\},$$

we denote

$$\hat{C}^\beta(p)=\frac{1+e^{-\beta\mathcal{E}(p)}}{1-e^{-\beta\mathcal{E}(p)}},\quad \hat{S}^\beta(\tau,p)=\frac{e^{-|\tau|\mathcal{E}(p)}+e^{-(\beta-|\tau|)\mathcal{E}(p)}}{1-e^{-\beta\mathcal{E}(p)}}.$$

Since

$$\hat{C}^\beta(p)=1+\frac{2e^{-\beta\mathcal{E}(p)}}{1-e^{-\beta\mathcal{E}(p)}},$$

the corresponding expression in x space has a δ singularity at $x=0$. If we assume that $\mathcal{E}\in C(\mathbb{R}^d)$ and

$$\mathcal{E}(p)\sim|p|^\eta \quad \text{for} \quad |p|\to\infty,\quad \eta>0,$$

we obtain

$$C^\beta(x)=\delta(x)+\mathcal{R}^\beta(x),$$

where $\mathcal{R}^\beta$ is continuous and

$$\mathcal{R}^\beta\in\bigcap_{p\geqslant 1}L^p(\mathbb{R}^d).$$

The kernel $\hat{S}^\beta(\tau,p)$ has similar properties.

Remark: The technical difficulties arise only in the case when the set

$$\mathcal{N}_\mathcal{E}=\{p\in\mathcal{R}^d:\mathcal{E}(p)=0\}$$

is nonempty. In that case, we have to restrict the index space of the process ξ_τ^β. For example, if $\mathcal{E}$ is continuous and

$$\mathcal{E}(p)\sim|p|^\nu \quad \text{as} \quad |p|\to+0,$$

we can take as the index space $C^\infty_{\eta^*}(\mathbb{R}^d)$ defined as follows:

$$C^\infty_{\eta^*}(\mathbb{R}^d)=\left\{f\in\mathcal{S}(\mathbb{R}^d):\ \frac{\partial^{|\mathbf{i}|}}{\partial x_1^{i_1}\cdots\partial x_n^{i_n}}f\Big|_{x=0}=0,\quad \text{for all } \left|\mathbf{i}\right|\leqslant\eta^*\right\},$$

where $\eta^*\geqslant d-1-\nu$.

Using similar arguments as in the proof of Proposition 4.2, we obtain the following:

Proposition 4.3:

(1) Let $\mathcal{E}\in\mathcal{B}_0(\mathbb{R}^d)$. Then for any $\alpha>d/2$, there exists a version of the process ξ_τ^β with values in the space

$$H^\rho_{-\alpha}(\mathbb{R}^d)=\left\{f\in\mathcal{S}'(\mathbb{R}^d):\int_{\mathbb{R}^d}\frac{|\hat{f}(p)|^2}{(p^2+1)^\alpha}\,\rho^2(p)dp<\infty\right\}.$$

If, in addition, $\mathcal{E}(p)\sim|p|^\eta$ for $|p|\to\infty$, then for any $\alpha>(d+\eta)/2$ there exists a version of ξ_τ^β such that with probability one

$$\|\xi_\tau^\beta - \xi_{\tau'}^\beta\|_{-\alpha}^2 \leqslant \Theta_\beta \left(\frac{|\tau-\tau'|}{\ln(1/|\tau-\tau'|)} \right)^{1/2}.$$

for sufficiently small $|\tau-\tau'|$ and some integrable random variable Θ_β.

(2) Let $\mathcal{E}$ be such that $\mathcal{E}(p)\sim|p|^\eta$ as $|p|\to+0$ and $\mathcal{E}(p)\sim|p|^\nu$ as $|p|\to\infty$. Let

$$H^\rho_{-\alpha,\eta}(\mathbb{R}^d) = \Big\{ \text{metric completion of } C_0^\infty(\mathbb{R}^d) \text{ with the norm } \|f\|^{\rho 2}_{-\alpha,\eta}$$
$$= \int_{\mathbb{R}^d} \frac{|p|^\eta |f(p)|^2}{(p^2+1)^\alpha}\, \rho^2(p)dp \Big\}.$$

Then for any $\alpha>(d+\eta)/2$ there exists a version of the process ξ_τ^β with values in $H^\rho_{-\alpha,\eta}(\mathbb{R}^d)$ and for any $\alpha>(d+\eta+\nu)/2$ there exists a version of ξ_τ^β with Hölder continuous paths as in (1).

Remark: The case of critical (standard) Bose gas defined by the thermal state ω_{cr} on the Weyl algebra over $\mathcal{S}(\mathbb{R}^d)$ given by

$$\omega_{\mathrm{cr}}(W_f) = \exp\, c|\hat f(0)|^2) \exp\left[-\frac{1}{4} \int_{\mathbb{R}^d} \frac{1+e^{-\beta p^2}}{1-e^{-\beta p^2}} |\hat f(p)|^2\, dp \right],$$

where $c>0$ is a constant and $d\geqslant 3$, was discussed in Ref. 25. Here we quote only some continuity results.

Let $d\geqslant 3$ and $\xi_\tau^{\beta,\mathrm{cr}}$ be the thermal process indexed by $\mathcal{S}(\mathbb{R}^d)$ with the covariance given by the density

$$S^{\beta,\mathrm{cr}}(\tau,p) = c\,\delta(p) + \frac{e^{-|\tau|p^2} + e^{-(\beta-|g t|)p^2}}{1-e^{-\beta p^2}}.$$

For any $\alpha>d/2$ there exists a version of $\xi_\tau^{\beta,cr}$ with values in $H_{-\alpha}(\mathbb{R}^d)$ and for any $\alpha>(d+1)/2$ there exists a version of $\xi_\tau^{\beta,cr}$ with Hölder continuous paths.

C. Quantum lattice models

Let $\mathbb{L}$ be a countable set with elements denoted by $\mathbf{l}$. Assume that for any $\mathbf{l}\in\mathbb{L}$ there is a separable Hilbert space $\mathcal{D}_\mathbf{l}$ with a unitary group $T_{\mathbf{l},t}=e^{it\mathbf{h}_\mathbf{l}}$, where $\mathbf{h}_\mathbf{l}$ is self-adjoint and $\sigma(\mathbf{h}_\mathbf{l})\subset[\epsilon,\infty)$, $\epsilon>0$. Assume also that there is a complex conjugation $C_\mathbf{l}$ on $\mathcal{D}_\mathbf{l}$ such that $\mathbf{h}_\mathbf{l}$ is $C_\mathbf{l}$-real. Let us define

$$\mathcal{D}=\bigoplus_{\mathbf{l}\in\mathbb{L}} \mathcal{D}_\mathbf{l}, \quad T_t=\bigoplus_{\mathbf{l}\in\mathbb{L}} T_{\mathbf{l},t}, \quad C=\bigoplus_{\mathbf{l}\in\mathbb{L}} C_\mathbf{l}.$$

On $\mathcal{D}$ we define a symplectic form

$$\sigma(\varphi,\psi)=\sum_{\mathbf{l}\in\mathbb{L}} \mathrm{Im}\langle \varphi_\mathbf{l}, \psi_\mathbf{l}\rangle_{\mathcal{D}_\mathbf{l}},$$

where $\varphi=(\varphi_\mathbf{l})_\mathbf{l}\in\mathbb{L}\in\mathcal{D}$. Consider a state ω defined on the Weyl algebra $\mathfrak{M}(\mathcal{D},\sigma)$ by

$$\omega(W_\varphi)=\prod_{\mathbf{l}\in\mathbb{L}} \exp\left[-\frac{1}{4} \left\langle \varphi_\mathbf{l}, \coth\frac{\beta}{2}\mathbf{h}_\mathbf{l}\varphi_\mathbf{l} \right\rangle_{\mathcal{D}_l} \right]. \tag{4.1}$$

The state (4.1) is an α_t-KMS state (for α_t defined by T_t) at the inverse temperature β. The corresponding thermal process indexed by $\mathcal{D}_+=\bigoplus \mathbf{l}\in\mathbb{L}\mathcal{D}_{+,l}$ will be denoted by $\xi_\tau^{\beta,\mathbb{L}}$. The process $\xi_\tau^{\beta,\mathbb{L}}$ is a direct sum of the site processes $\xi_\tau^{\beta,\mathbf{l}}$, indexed by $\mathcal{D}_{+,\mathbf{l}}$, and its covariance is given by

$$\mathbb{E}\langle \xi_\tau^{\beta,\mathbb{L}},\varphi\rangle\langle \xi_\tau^{\beta,\mathbb{L}},\psi\rangle=\sum_{\mathbf{l}\in\mathbb{L}} \mathbb{E}\langle \xi_\tau^{\beta,\mathbf{l}},\varphi_\mathbf{l}\rangle\langle \xi_{\tau'}^{\beta,\mathbf{l}},\psi_\mathbf{l}\rangle. \tag{4.2}$$

From Proposition 3.4 and Eq. (4.2) follows that for $\varphi=(\varphi_l)_{l\in\mathbb{L}}$ such that

$$m(\varphi)=\sum_{\mathbf{l}\in\mathbb{L}} m_{\mathbf{l}}(\varphi_{\mathbf{l}})\leqslant\infty,$$

where $m_{\mathbf{l}}(\varphi_{\mathbf{l}})$ is the moment of the measure corresponding to the covariance of $\xi^{\beta,\mathbf{l}}$ (see Proposition 3.4), there exists a Hölder continuous version of the coordinate process $\langle\xi_\tau^{\beta,\mathbb{L}},\varphi\rangle = \sum_{\mathbf{l}\in\mathbb{L}}\langle\xi_\tau^{\beta,\mathbf{l}},\varphi_{\mathbf{l}}\rangle$.

Let A be a self-adjoint, non-negative operator on $\mathcal{D}$ such that $\oplus_{\mathbf{l}\in\mathbb{L}}\mathbf{h}_{\mathbf{l}}+A$ is essentially self-adjoint on $D(\oplus_{\mathbf{l}\in\mathbb{L}}\mathbf{h}_{\mathbf{l}})$. Define $\mathbf{h}_A=\overline{\oplus_{\mathbf{l}\in\mathbb{L}}\mathbf{h}_{\mathbf{l}}+A}$. The state

$$\omega_A(W_\varphi)=\exp\left[-\frac{1}{4}\left\langle\varphi,\frac{\mathbb{I}+e^{-\beta\mathbf{h}_A}}{\mathbb{I}-e^{-\beta\mathbf{h}_A}}\,\varphi\right\rangle\right]$$

is a KMS state with respect to the evolution defined by $T_t^A=e^{it\mathbf{h}_A}$. If $\mathbf{h}_A$ is C-real, there exists a thermal process $\xi_\tau^{\beta,A}$ indexed by $\mathcal{D}_+=\oplus_{\mathbf{l}\in\mathbb{L}}\mathcal{D}_{+,\mathbf{l}}$, with the covariance

$$\mathbb{E}\langle\xi_\tau^{\beta,A},\varphi\rangle\langle\xi_{\tau'}^{\beta,A},\psi\rangle=\left\langle\varphi,\frac{e^{-|\tau-\tau'|\mathbf{h}_A}+e^{-(\beta-|\tau-\tau'|)\mathbf{h}_A}}{\mathbb{I}-e^{-\beta\mathbf{h}_A}}\,\psi\right\rangle.$$

The law of this process is given by a centered Gaussian random field μ_A^β with covariance

$$\mathbb{E}^{\mu_A^\beta}\langle\Phi,\varphi\otimes f\rangle\langle\Phi,\psi\otimes g\rangle=\left\langle\varphi\otimes f,\left[\left(-\frac{d^2}{d\tau^2}\right)_{\text{per}}+\mathbf{h}_A\right]^{-1}\psi\otimes g\right\rangle,$$

where $(-d^2/d\tau^2)_{\text{per}}$ is a periodized version of the operator $-d^2/d\tau^2$. A similar conclusion is true for the ground state process.

Harmonic crystal model. Let $\mathbb{L}=\mathbb{Z}^d$ be a lattice consisting of points in $\mathbb{R}^d$ with all coordinates integer. To each $\mathbf{l}\in\mathbb{Z}^d$ we associate a copy of the complex plane $\mathbb{C}$ with a natural complex conjugation $z\to\overline{z}$ and a copy of an unitary group $u_1(t)=\exp(i/2)t$. By the direct sum construction we obtain the Hilbert space $\mathcal{D}=l^2(\mathbb{Z}^d)$ with the complex conjugation $C(z_{\mathbf{l}})_{\mathbf{l}\in\mathbb{Z}^d}=(\overline{z_{\mathbf{l}}})_{\mathbf{l}\in\mathbb{Z}^d}$ and the unitary group $T_t=\oplus_{\mathbf{l}\in\mathbb{Z}^d}u_l(t)$. Defining the KMS state by (4.1) and proceeding as above, we obtain a thermal process $\xi_\tau^{\beta,\mathbb{Z}^d}$ which is a direct sum of the periodized Ornstein–Uhlenbeck processes $\xi_\tau^{\beta,\mathbf{l}}$ with covariances

$$\mathbb{E}\xi_\tau^{\beta,\mathbf{l}}\xi_0^{\beta,\mathbf{l}'}=\frac{1}{2}\,\frac{e^{-|\tau|/2}+e^{-(\beta-|\tau|)/2}}{1-e^{-\beta/2}}\,\delta_{\mathbf{l}\mathbf{l}'},$$

where $|\tau|\leqslant\beta$

Now let A be a self-adjoint and non-negative operator on $l^2(\mathbb{Z}^d)$ with real matrix coefficients. Since A is C-real, $\mathbf{h}_A=1/2\mathbb{I}+A$ is also C-real, and we obtain a thermal process $\xi_\tau^{\beta,A}$ indexed by $l^2_{\mathbb{R}}(\mathbb{Z}^d)$. Its covariance is given by

$$\mathbb{E}\langle\xi_\tau^{\beta,A},(z_{\mathbf{l}})\rangle\langle\xi_0^{\beta,A},(z'_{\mathbf{l}})\rangle=\left\langle(z_{\mathbf{l}}),\frac{e^{-|\tau|\mathbf{h}_A}+e^{-(\beta-|r|)\mathbf{h}_A}}{\mathbb{I}-e^{-\beta\mathbf{h}_A}}\,(z'_{\mathbf{l}})\right\rangle_{l^2(\mathbb{Z}^d)}. \tag{4.3}$$

In the Gibbs state approach discussed in Refs. 46–49, the process $\xi_\tau^{\beta,A}$ can be obtained as a perturbation of the process $\xi_\tau^{\beta,\mathbb{Z}^d}$. The perturbation is defined as follows. Let

$$d\mu_0^\beta=\bigotimes_{\mathbf{l}\in\mathbb{Z}^d} d\mu_0^{\beta,\mathbf{l}},$$

where $d\mu_0^{\beta,\mathbf{l}}$ is the measure on trajectories of the periodic Ornstein–Uhlenbeck process $\xi_\tau^{\beta,\mathbf{l}}$. Then

$$d\mu^{\beta,A}((\omega_{\mathbf{l}})_{\mathbf{l}\in\mathbb{Z}^d}) = \lim_{\Lambda\uparrow\mathbb{Z}^d} \frac{1}{Z_A} \exp\left[-\int_{-\beta/2}^{\beta/2} \sum_{\mathbf{l},\mathbf{k}\in\Lambda} A_{\mathbf{lk}}\omega_{\mathbf{l}}(\tau)\omega_{\mathbf{k}}(\tau)\, d\tau \right] \cdot d\mu_0^{\beta}((\omega_{\mathbf{l}})_{\mathbf{l}\in\mathbb{Z}^d}),$$

$$Z_{\Lambda} = \int d\mu_0^{\beta}((\omega_{\mathbf{l}})_{\mathbf{l}\in\mathbb{Z}^d}) \exp\left[-\int_{-\beta/2}^{\beta/2} \sum_{\mathbf{l},\mathbf{k}\in\Lambda} A_{\mathbf{lk}}\omega_{\mathbf{l}}(\tau)\omega_{\mathbf{k}}(\tau) d\tau \right],$$

where Λ is a finite subset of $\mathbb{Z}^d$ and the limit is taken in the sense of Dobrushin–Landford–Rueke (DLR) equations.[50,51] Clearly it defines a measure on trajectories of the process $\xi_{\tau}^{\beta,A}$. In fact, the set of limiting Gibbs measures can contain also some nontranslationary invariant solutions,[52] which differ from those chosen above only by nonzero means. This is connected with the nontrivial kernel of the operator $d^2/d\tau^2+\mathbf{h}_A$. Using this and general results of Sec. III E we obtain the following:

Proposition 4.4: The modular structure of the KMS state corresponding to the harmonic crystal model with the Fock space Hamiltonian $d\Gamma(\mathbf{h}_A)$ formally given by

$$d\Gamma(\mathbf{h}_A) = '' \sum_{\mathbf{l}\in\mathbb{Z}^d} \left(-\frac{1}{2}\frac{d^2}{dx_{\mathbf{l}}^2} + \frac{1}{2} x_{\mathbf{l}}^2 \right) + \sum_{\mathbf{l},\mathbf{k}\in\mathbb{Z}^d} A_{\mathbf{lk}} x_{\mathbf{l}} x_{\mathbf{k}} '',$$

where $(A_{\mathbf{lk}}) = A \geqslant 0$, is determined by the thermal process $\xi_{\tau}^{\beta,A}$ with the covariance (4.3).

Remark: The result that the modular structure of the harmonic crystal model is stochastically determined seems to be new. For other aspects of this model, we refer to Ref. 53. Gibbsian perturbations of the Abelian sector were discussed recently in Refs. 48 and 49.

ACKNOWLEDGMENTS

A part of this work has been done during the visits of the first (RG) and the third (RO) authors in BiBoS Research Center under the financial support of CIPA-CT92-4016. This work was also supported by KBN Grant No. 2PO3B 122 11.

[1] K. Osterwalder and R. Schrader, ''Axioms for Euclidean Green Functions. I.,'' Commun. Math. Phys. **31**, 83–112 (1973).
[2] K. Osterwalder and R. Schrader, ''Axioms for Euclidean Green Functions. II.,'' Commun. Math. Phys. **42**, 281–305 (1975).
[3] V. Glaser, ''On the equivalence of the Euclidean and Wightman formulation of field theory,'' Commun. Math. Phys. **37**, 257–272 (1974).
[4] Yu. M. Zinoviev, ''Equivalence of Euclidean and Wightman field theory,'' Commun. Math. Phys. **174**, 1 (1995/96).
[5] H. J. Borchers and J. Yngvason, ''Necessary and sufficient conditions for integral representations of Wightman functionals at Schwinger points,'' Commun. Math. Phys. **47**, 197–213 (1976).
[6] E. Nelson, ''The free Markov field,'' J. Funct. Anal. **12**, 211–217 (1973).
[7] K. Symanzik, ''Euclidean quantum field theory,'' *Local Quantum Theory*, edited by R. Jost (Academic, New York, 1969), pp. 152–225.
[8] E. Nelson, ''The construction of quantum fields from Markov fields,'' J. Funct. Anal. **12**, 97–112 (1973).
[9] B. Simon, *The $P(\phi)_2$ Euclidean (quantum) field theory* (Princeton U. P., Princeton, NJ, 1974).
[10] J. Glimm and A. Jaffe, *Quantum physics. A functional integral point of view* (Springer-Verlag, New York, 1987).
[11] N. D. Birrell and P. C. W. Davies, *Quantum fields in curved space* (Cambridge U. P. Cambridge, 1982).
[12] S. A. Fulling, *Aspects of quantum field theory in curved space-time, London Math. Soc. Student Texts, Vol. 17* (Cambridge U. P., Cambridge, 1989).
[13] S. A. Fulling and S. N. M. Ruijsenaars, ''Temperature, periodicity and horizons,'' Phys. Rep. **152**, 135–176 (1987).
[14] R. M. Wald, ''Euclidean approach to quantum field theory in curved space-time,'' Commun. Math. Phys. **70**, 221 (1979).
[15] G. W. Gibbons and S. W. Hawking (eds.), *Euclidean quantum gravity* (World Scientific, Singapore, 1993).
[16] S. W. Hawking, ''Path integrals in application to quantum gravity,'' *General relativity*, edited by S. W. Hawking and W. Israel, An Einstein centenary survey (Cambridge U. P., Cambridge, 1979).
[17] S. Albeverio and R. Hoegh-Krohn, ''Homogeneous random fields and statistical mechanics,'' J. Funct. Anal. **19**, 242–272 (1975).
[18] A. Klein and L. J. Landau, ''Stochastic processes associated with KMS states,'' J. Funct. Anal. **42**, 368–428 (1981).
[19] J. Fröhlich, ''The reconstruction of quantum fields from Euclidean Green's functions at arbitrary temperature in models of a self-interacting Bose fields in two space-times dimensions,'' Helv. Phys. Acta **48**, 355–363 (1975).
[20] W. Driessler, L. Landau, and J. F. Perez, ''Estimates of critical lengths and critical temperatures for classical and quantum lattice systems,'' J. Stat. Phys. **20**, 123–162 (1979).
[21] L. A. Pastur and B. A. Khoruzhenko, ''Phase transitions in quantum models of rotators and ferroelectrics,'' Theor. Math. Phys. **73**, 111–124 (1987).
[22] R. Gielerak, L. Jakóbczyk, and R. Olkiewicz, ''Reconstruction of KMS structure from Euclidean Green functions,'' J. Math. Phys. **35**, 3726 (1994).

[23] R. Gielerak, L. Jakóbczyk, and R. Olkiewicz, ''W*-KMS structure from multitime Euclidean Green functions,'' J. Math. Phys. **35**, 6291 (1994).
[24] R. Gielerak and R. Olkiewicz, ''Gentle perturbations of the free Bose gas. I.,'' J. Stat. Phys. **80**, 875 (1995).
[25] R. Gielerak and R. Olkiewicz, ''Gentle perturbations of the free Bose gas. II. The critical regime,'' J. Math. Phys. **37**, 1268–1286 (1996).
[26] R. Gielerak and R. Olkiewicz, ''Gentle perturbations of free Bose gas. III. The states and cyclicity of the thermal vacuum,'' (to appear).
[27] R. Gielerak and R. Olkiewicz, ''Stochastic approach to many Bosons physics,'' in *The interplay between stochastic and deterministic behaviour*, edited by P. Garbaczewski, M. Wolf, and A. Weron, Proc. of the XXXI Karpacz Winter School in Theoretical Physics. Lecture Notes in Physics (Springer-Verlag, New York, 1995), Vol. 1457, pp. 435–444.
[28] H. Araki and E. J. Woods, ''Representations of the canonical commutation relations describing a nonrelativistic infinite free bose gas,'' J. Math. Phys. **4**, 637–662 (1963).
[29] F. Rocca, M. Sirugue, and D. Testard, ''On a class of equilibrium states under the Kubo-Martin-Schwinger condition. II. Bosons,'' Commun. Math. Phys. **19**, 119–141 (1970).
[30] O. Bratteli and D. Robinson, *Operator algebras and quantum statistical mechanics. II* (Springer-Verlag, New York, 1981).
[31] J. J. Bisognano and E. H. Wichmann, ''On the duality condition for a Hermitian scalar field,'' J. Math. Phys. **16**, 985–1007 (1975).
[32] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, ''The smallest C^*-algebra for canonical commutation relations,'' Commun. Math. Phys. **32**, 231–242 (1973).
[33] H. Araki, ''Multi time analyticity of a quantum statistical state satisfying the KMS boundary condition,'' Publ. RIMS **4**, 361 (1968).
[34] J. Manuceau and A. Verbeure, ''Quasi-free states of the C.C.R.-algebra and Bogoliubov transformations,'' Commun. Math. Phys. **9**, 293–302 (1968).
[35] A. Klein and L. J. Landau, ''Periodic Gaussian Osterwalder-Schrader positive processes and the two-sided Markov property on the circle,'' Pacific J. Math. **94**, 341–367 (1981).
[36] A. Klein and L. J. Landau, ''Construction of a unique self-adjoint generator for a symmetric local semigroup,'' J. Funct. Anal. **44**, 121–137 (1981).
[37] J. Leray, *Hyperbolic Differential Equations* (The Institute for Advanced Study, Princeton, NJ, 1953).
[38] B. S. Kay, ''Linear spin-zero quantum fields in external gravitational and scalar fields. I. A one particle structure for the stationary case,'' Commun. Math. Phys. **62**, 55–70 (1978).
[39] S. Albeverio, R. Gielerak, and F. Russo, ''On the path Holder continuity in models of Euclidean quantum field theory,'' preprint Inst. Galilei, Univ. Paris Nord, 1996.
[40] C. Preston, ''Continuity properties of some Gaussian processes,'' Ann. Math. Stat. **43**, 285–292 (1972).
[41] A. M. Garsia, ''On the smoothness of functions satisfying certain integral inequalities,'' in *Proceedings of the Symposium on functional analysis at Monterey*, edited by C. Wilde (Academic, New York, 1970), pp. 127–162.
[42] M. Röckner, ''Traces of harmonic functions and a new path space for the free quantum field,'' J. Funct. Anal. **79**, 211–249 (1988).
[43] R. Hoeg-Krohn, ''Relativistic quantum statistical mechanics in two-dimensional space-time,'' Commun. Math. Phys. **38**, 195–224 (1974).
[44] S. Kotani, ''On a Mankov property for stationary Gaussian processes with a multidimensional parameter,'' Lect. Notes Math. **330**, 239–250 (1974).
[45] B. S. Kay, ''The double-wedge algebra for quantum fields on Schwarzschild and Minkowski spacetimes,'' Commun. Math. Phys. **100**, 57–81 (1985).
[46] S. A. Globa and Ju. G. Kondratiev, ''Gibbs measures for quantum lattice systems,'' Selecta Math. Sov. **9**, 297 (1990).
[47] Ju. G. Kondratiev, ''Phase transitions in quantum models of ferroelectrics,'' BiBoS preprint No 487, 1991.
[48] Y. M. Park and H. J. Yoo, ''A characterisation of Gibbs sates of lattice Boson systems,'' J. Stat. Phys. **75**, 215–239 (1994).
[49] Y. M. Park and H. J. Yoo, ''Uniqueness and clustering properties of Gibbs sates for classical and quantum unbounded spin systems,'' J. Stat. Phys. **80**, 223–271 (1995).
[50] B. Simon, *The statistical mechanics of lattice gases* (Princeton U. P., Princeton, NJ, 1993), Vol. I.
[51] R. B. Israel, *Convexity in theory of lattice gases* (Princeton U. P., Princeton, 1979).
[52] R. L. Dobrushin, ''Gaussian random fields: Gibbsian point of view,'' in *Multicomponent random systems*, edited by R. L. Dobrushin and Ya. G. Sinai (Mov, Moscow, 1987), pp. 119–153.
[53] A. F. Verbeure and E. J. Verboven, ''States of infinitely many oscillators,'' Physica **37**, 23 (1967).

Gibbs states for AF algebras

Valentin Ya. Golodets[a] and Sergey V. Neshveyev
Institute for Low Temperature Physics & Engineering,
Lenin Ave 47, Kharkov 310164, Ukraine

(Received 13 February 1998; accepted for publication 17 June 1998)

We consider a special class of C^*-systems containing asymptotically Abelian binary shifts and shifts of Temperley–Lieb algebras. We study Gibbs states for these systems corresponding to potentials with finite range interaction, and obtain the same results as the well-known Araki's results for a one-dimensional quantum lattice. In particular, it is proved that a Gibbs state in the infinite volume is a translation invariant KMS state having the exponential uniform clustering property. Entropic properties of the Gibbs states are also discussed. This allows us, in particular, to construct new examples of quantum K-systems. © *1998 American Institute of Physics.* [S0022-2488(98)01112-8]

I. INTRODUCTION

In Ref. 1, Araki studied a one-dimensional infinite quantum spin lattice system with a finite range interaction. He showed that the Gibbs state ϕ of the system in the infinite volume is a limit of local Gibbs states, is a factor state of the UHF algebra A corresponding to the quantum spin lattice system, is invariant under time and lattice translation, satisfies the KMS boundary condition, and has the exponential uniform clustering property.

Developing the transfer matrix technique in a fashion analogous to Ref. 2, Araki introduced an auxiliary operator $\mathcal{L}$ acting on A, and found an eigenstate ν for $\mathcal{L}$. It turned out that the Gibbs state ϕ is a perturbation of the state ν. To realize this approach, Araki developed the Tomonaga–Shwinger–Dyson perturbation theory to the case of UHF algebras (see Refs. 3 and 4 for a more detailed exposition and applications).

It is a natural problem to extend Araki's theory to a more general class of AF algebras. We would like to include in this class such important algebras as Temperley–Lieb algebras and asymptotically Abelian binary shifts. So we will consider quadruples $(A,\{A_{[n,m]}\},\tau,\gamma)$, where A is a unital C^*-algebra, $A_{[n,m]}$, $n,m\in\mathbb{Z}$, $n\leqslant m$, is a finite-dimensional subalgebra of A such that $A_{[n,m]}\subset A_{[n',m']}$ for $n'\leqslant n\leqslant m\leqslant m'$, $\cup_n A_{[-n,n]}$ is dense in A, τ is a faithful trace on A, and γ is a τ-preserving automorphism of A such that $\gamma(A_{[n,m]})=A_{[n+1,m+1]}$. Moreover, we suppose that the trace τ has some multiplicativity property, and the local algebras satisfy Popa's commuting square condition (see Sec. II). Nevertheless, there are difficulties in this way. For example, we cannot define an analog of Araki's operator $\mathcal{L}$ on the whole algebra A, since, in general, the algebra cannot be ''cut'' into two commuting pieces as in Ref. 1. Fortunately, it is possible to define such an operator on the subalgebra $A_{[0,\infty)}$ as in Ruelle's paper.[2] Then, basing on Araki's perturbation technique we can prove the existence of an eigenstate ν on $A_{[0,\infty)}$. The Gibbs state ϕ on A can be restored from this state ν. Again, as in Ref. 1, the Gibbs state is the limit of local Gibbs states, invariant under γ, satisfies the KMS boundary condition and has the exponential uniform clustering property. These results are proved in Sec. III.

The constructed Gibbs state is a σ_t-KMS state ($\beta=1$), where σ_t is the time evolution. For a one-dimensional quantum lattice it is a unique σ_t-KMS state (see Refs. 5 and 4). The same result holds for binary shifts, or for the C^*-tensor product of a binary shift and a quantum spin lattice system. To prove this, one can apply the same arguments as in Ref. 5, or Ref. 4 (see Remark III.13), but, in general, this problem apparently requires new arguments.

The present research has been undertaken to construct new examples of quantum Kolmogorov systems in the sense of Narnhofer and Thirring.[6] Indeed, let ϕ be a Gibbs state, π_ϕ the GNS

[a]Telephone: 38(0572) 30-85-85; Fax (0572) 322370; Electronic mail: golodets@ilt.kharkov.ua

 © 1998 American Institute of Physics

representation of A corresponding to ϕ. Then (M, ϕ, γ), where $M=\pi_\phi(A)''$, is a quantum K-system. We deduce this result in Sec. IV from the proved clustering property and from the sufficient condition for the K-property in our paper.[7]

In Sec. IV we also compute the mean entropy of a Gibbs state for certain C^*-dynamical systems. In particular, for a binary shift and for a Temperley–Lieb algebra of index $\leqslant 4$. It is a very interesting problem to investigate when this mean entropy is equal to the dynamical entropy $h_\phi(\gamma)$. But this problem has not been solved, even for a quantum spin lattice system[8] (see also Remark IV.7 below).

It is important to note that Hiai and Petz[9] considered quantum statistical thermodynamics in AF C^*-systems. They studied, in particular, such problems as the Gibbs conditions, the variational principle for states, their relations with the KMS conditions, and presented a lot of interesting examples. But Araki's approach to Gibbs states was not considered.

II. NOTATIONS AND EXAMPLES

We consider quadruples $(A,\{A_{[n,m]}\}_{n\leqslant m},\tau,\gamma)$, $n,m\in\mathbb{Z}$, where A is a unital C^*-algebra, $A_{[n,m]}$ is a finite-dimensional C^*-subalgebra of A such that $A_{[n',m']}\subset A_{[n'',m'']}$ for $n''\leqslant n'\leqslant m'\leqslant m''$, and $\cup_n A_{[-n,n]}$ is dense in A, τ a faithful trace of A, γ a τ-preserving automorphism of A such that $\gamma(A_{[n,m]})=A_{[n+1,m+1]}$. For any subset Λ of $\mathbb{Z}$, we denote by A_Λ the C^*-subalgebra of A generated by $A_{[k,n]}$, $[k,n]\subset\Lambda$, and write A_n instead of $A_{[n,n]}$.

Throughout the paper we suppose that the following conditions are satisfied.

Assumptions II.1:

(i) There exists $n_0\geqslant 0$ such that $A_{(-\infty,0]}$ and $A_{[n_0+1,\infty)}$ commute.

(ii) $\tau(xy)=\tau(x)\tau(y)$ for $x\in A_{(-\infty,0]}$, $y\in A_{[n_0+1,\infty)}$.

(iii) There exists a τ-preserving conditional expectation $\mathcal{E}:A_{[0,\infty)}\to A_{[1,\infty)}$, and $\mathcal{E}(A_{[0,n]})=A_{[1,n]}$ for any $n\geqslant 1$. Equivalently,

$$\begin{array}{ccc} A_{[0,n]} & \subset & A_{[0,n+1]} \\ \cup & & \cup \\ A_{[1,n]} & \subset & A_{[1,n+1]} \end{array}$$

is a commuting square[10,11] for any n.

Besides a one-dimensional quantum spin lattice system we have the following standard examples.

Example II.2. Automorphism θ_λ:[11]

Let $\lambda\in\{(4\cos^2(\pi/n))^{-1}\}_{n>3}\cup(0,\frac{1}{4}]$, A the C^*-algebra generated by projections e_n, $n\in\mathbb{Z}$, satisfying $e_ie_j=e_je_i$ for $|i-j|\geqslant 2$, and $e_ie_{i\pm1}e_i=\lambda e_i$, $A_{[n,m]}=C^*(e_n,e_{n+1},\ldots,e_m)$, τ the λ-Markov trace on A, i.e., $\tau(we_n)=\lambda\tau(w)$ for any $w\in A_{(-\infty,n-1]}$, $\gamma(e_n)=e_{n+1}$.

Example II.3. Canonical shift on the tower of relative commutants:[12]

Let M be a II_1 factor, $N\subset M$ its subfactor of finite index, $N\subset M\subset^{e_1}M_1\subset^{e_2}M_2\subset\cdots$ the Jones tower, $\cdots\subset^{e_{-2}}M_{-2}\subset^{e_{-1}}M_{-1}=N\subset^{e_0}M_0=M$ a tunnel, $A_{[n,m]}=M'_{2n}\cap M_{2m}$, $A=\overline{\cup_n A_{[-n,n]}}$, τ the restriction of the trace on $\cup_n M_n$. For any $n\in\mathbb{N}$, the triple $M_{-n}\subset M_n\subset M_{3n}$ is a basic construction in a canonical way. In particular, a (not necessarily τ-preserving) antiautomorphism $\gamma_n(x)=J_nx^*J_n$ of $M'_{-n}\cap M_{3n}$ is defined, where J_n is the canonical conjugation on $L^2(M_n)$. Then $\gamma(x)=\gamma_{n+1}(\gamma_n(x))$ for $x\in M'_{-n}\cap M_{3n}$.

Example II.4. Asymptotically Abelian binary shift:[13,14]

Let X be a nonempty finite subset of $\mathbb{N}$, $a(x)$ the characteristic function of X, A the C^*-algebra with unit generated by symmetries s_n, $n\in\mathbb{Z}$, satisfying $s_is_j=(-1)^{a(|i-j|)}s_js_i$, $A_{[n,m]}=C^*(s_n,s_{n+1},\ldots,s_m)$, $\tau(w)=0$ for any nonempty word w in s_n's, $\gamma(s_n)=s_{n+1}$.

It is worth noting that the C^*-tensor product of systems satisfying Assumptions II.1 also satisfies these assumptions.

By an interaction potential we mean a mapping $X\mapsto\Phi(X)\in A_X$ defined on finite subsets X of $\mathbb{Z}$ such that $\Phi(X)=\Phi(X)^*$, $\Phi(X+n)=\gamma^n(\Phi(X))$. We suppose that $\Phi(X)=0$ if X is not within an interval of length $r>0$. Following Araki, we denote by $U(\Lambda)$ the Hamiltonian for a finite interval Λ,

$$U(\Lambda)=\sum_{X\subset\Lambda}\Phi(X).$$

We write $U(a,b)$ instead of $U([a,b])$.

For a finite interval Λ, the local Gibbs state is

$$\phi_\Lambda(Q)=\frac{\tau(Qe^{-U(\Lambda)})}{\tau(e^{-U(\Lambda)})}.$$

III. EXISTENCE AND CLUSTERING PROPERTIES OF GIBBS STATES

Our main results in this section are as follows.

Theorem III.1:

(i) The limit $P(\Phi)=\lim_{b-a\to\infty}[1/(b-a)]\log\tau(e^{-U(a,b)})$ *exists and is finite.*

(ii) For any $Q\in A$, *there exists the limit* $\phi(Q)=\lim_{\substack{a\to-\infty\\ b\to+\infty}}\phi_{[a,b]}(Q)$. *Moreover, there exist* $C,q>0$ *such that*

$$|\phi_{[-n+a,b+n]}(Q)-\phi(Q)|\leqslant Ce^{-qn}\|Q\|,\quad for\quad Q\in A_{[a,b]},\quad n\in\mathbb{N}.$$

Theorem III.2: *The state* ϕ *is uniformly exponentially clustering, i.e., there exist* $C,q>0$, *such that*

$$|\phi(Q_1Q_2)-\phi(Q_1)\phi(Q_2)|\leqslant Ce^{-qn}\|Q_1\|\|Q_2\|,$$

for $Q_1\in A_{(-\infty,-n)}$, $Q_2\in A_{[n,\infty)}$, $n\in\mathbb{N}$.

We shall call the state ϕ the Gibbs state corresponding to the interaction potential $\Phi(X)$.

The proofs of Theorems follow closely the work by Araki.[1] We introduce an auxiliary operator $\mathcal{L}$ and found an eigenstate ν for this operator. Then the Gibbs state can be restored from ν. The main difference of our work from Araki's one is that the operator $\mathcal{L}$ cannot be defined on the whole algebra A, but only on $A_{[0,\infty)}$. So ν is a state on $A_{[0,\infty)}$, and the perturbation argument by Araki cannot be applied. Our arguments are closer to Ruelle's work[2] (see also Ref. 15).

Following Araki, for $Q\in A_{[0,\infty)}$, we define

$$\|Q\|_l=\inf_{Q_l\in A_{[0,l]}}\|Q-Q_l\|,$$

$$|||Q|||_{N,x}=\|Q\|+\sum_{n=N}^{\infty}\|Q\|_n x^n\quad(x>1,\ N\in\mathbb{N}),$$

$$A(x)=\{Q\in A_{[0,\infty)}|\,|||Q|||_{1,x}<\infty\}.$$

Then $A(x)$ is a Banach space with respect to the norm $|||\ |||_{1,x}$. Every $|||\ |||_{N,x}$ is an equivalent norm on $A(x)$. For $Q>0$, we also define

$$\alpha_l(Q)=\inf_{Q_l\in A_{[0,l]},Q_l>0}\|Q-Q_l\|\|Q_l^{-1}\|,$$

$$\alpha(Q)=\|Q\|\|Q^{-1}\|=\frac{\text{l.u.b. Spec }Q}{\text{g.l.b. Spec }Q}.$$

Lemma III.3: For $Q>0$, *we have the following:*

(i) $\alpha_l(Q)\leqslant[\alpha(Q)-1]/[\alpha(Q)+1]<1$;

(ii) there exists $Q_l\in A_{[0,l]}$, $Q_l>0$, *such that* $\alpha_l(Q)=\|Q-Q_l\|\|Q_l^{-1}\|$, *then*

$$(1-\alpha_l(Q))\alpha(Q)\leqslant\|Q\|\|Q_l^{-1}\|\leqslant(1+\alpha_l(Q))\alpha(Q);$$

(iii) if $\|Q\|_l\|Q^{-1}\|<1$, *then* $\alpha_l(Q)\leqslant\|Q\|_l\|Q^{-1}\|/(1-\|Q\|_l\|Q^{-1}\|)$;
(iv) $\|Q\|_l\|Q^{-1}\|\leqslant\alpha_l(Q)/[1-\alpha_l(Q)]$.

Proof: See Ref. 1, Lemma 3.9: It is worth only to note that if $\|Q\|_l\|Q^{-1}\|<1$, then, for $Q_l=Q_l^*\in A_{[0,l]}$, $\|Q-Q_l\|=\|Q\|_l$, we have $Q_l>0$. Indeed,

$$\|\|Q\|1-Q_l\|\leqslant\|\|Q\|1-Q\|+\|Q\|_l=\|Q\|-\|Q^{-1}\|^{-1}+\|Q\|_l<\|Q\|.$$

□

Let us also define

$$\Phi=\sum_{I\subset[0,r]}\frac{\Phi(I)}{n(I)}\quad\text{and}\quad H(I)=\sum_{n:[n,n+r]\subset I}\gamma^n(\Phi),$$

where $n(I)$ is the number of translates $I+a$ of I that are still in $[0,r]$. $H(a,b)=H([a,b])$ and $U(a,b)$ differ only near the two ends. More precisely, there exist $\Delta^-\in A_{[0,r]}$ and $\Delta^+\in A_{[-r,0]}$ such that, for $b-a>2r$, we have

$$U(a,b)=H(a,b)+\gamma^a(\Delta^-)+\gamma^b(\Delta^+).$$

For any subset $I\subset\mathbb{Z}$, the sequence $\mathrm{Ad}\exp(itH(I\cap[-n,n]))$ pointwise converges to a strongly continuous one-parameter automorphism group σ_t^I on A, and every $Q\in\cup_nA_{[-n,n]}$ is a σ_t^I-analytic element (see Ref. 1, Theorem 4.2, or Ref. 4, Theorem 6.2.4). We write σ_t instead of $\sigma_t^{\mathbb{Z}}$. For $Q\in\cup_nA_{[-n,n]}$, $\lambda\in\mathbb{R}$, let us also consider the perturbed dynamics $(\sigma_t^I)^Q$ (see Ref. 4, Section 5.4.1). These dynamics are related via a σ_t^I-cocycle Γ_t^Q. Following Araki, an analytic continuation of this cocycle at the point $t=-i\lambda$ is denoted by $E(\lambda Q;\lambda H(I))$,

$$E(\lambda Q;\lambda H(I))=1+\sum_{n=1}^{\infty}\int_0^1ds_1\int_0^{s_1}ds_2\cdots\int_0^{s_{n-1}}ds_n\;\sigma^I_{-i\lambda s_n}(\lambda Q)\cdots\sigma^I_{-i\lambda s_1}(\lambda Q),$$

(see Ref. 1, Definition 5.1, or Ref. 4, p. 149). If I is finite, then

$$E(\lambda Q;\lambda H(I))=e^{\lambda Q+\lambda H(I)}e^{-\lambda H(I)}.$$

Lemma III.4: There exist constants $q>0$ *and* $C_n>0$, $n\in\mathbb{N}$, *such that*

$$\|E(Q;\lambda H(I))\|\leqslant e^{C_n\|Q\|}$$

and

$$\|E(Q;\lambda H(I))-E(Q;\lambda H(I\cap[-N,n+N]))\|\leqslant C_n\frac{q^{[N/(r+n_0)]}}{(1+[N/(r+n_0)])!}e^{C_n\|Q\|},$$

for $Q\in A_{[0,n]}$, $n\geqslant r$, $N>0$, $|\lambda|\leqslant1$, $I\subset\mathbb{Z}$, *where* n_0 *is defined in Assumptions II.1.*

Proof: See Ref. 1, Theorem 4.2 and Lemma 5.2. We would only mention that if $[\gamma^{j_m}(\Phi),[\ldots,[\gamma^{j_1}(\Phi),Q]\cdots]]\neq\varnothing$ then $[j_k,j_k+r]$ has a nonempty intersection with

$$[-n_0,n+n_0]\cup\bigcup_{1\leqslant l<k}[j_l-n_0,j_l+r+n_0],$$

for $k=1,\ldots,m$. But $[j_k,j_k+r]\cap[a-n_0,b+n_0]\neq\varnothing$ iff $[j_k,j_k+r+n_0]\cap[a,b+n_0]\neq\varnothing$. So the estimates by Araki hold with the replacement of r by $r+n_0$. □

We define the transfer operator $\mathcal{L}:A_{[0,\infty)}\to A_{[0,\infty)}$ as follows: $\mathcal{L}(Q)=\gamma^{-1}\mathcal{E}(K^*QK)$, where $K=E(-\frac{1}{2}\Phi;-\frac{1}{2}H(1,\infty))$ and $\mathcal{E}:A_{[0,\infty)}\to A_{[1,\infty)}$ is the τ-preserving conditional expectation.

Lemma III.5:

(i) $\mathcal{L}$ *is a faithful completely positive mapping. Moreover, if* $Q\geqslant0$ *is invertible,* $\mathcal{L}(Q)$ *is invertible too.*

(ii) $\mathcal{L}^n(Q)=\gamma^{-n}\mathcal{E}_n(K_n^*QK_ne^{-H(0,n-n_0-1)})$, *where* $\mathcal{E}_n:A_{[0,\infty)}\to A_{[n,\infty)}$ *is the* τ*-preserving conditional expectation,*

$$K_n=\begin{cases} E\left(-\frac{1}{2}\sum_{j=0}^{n-1}\gamma^j(\Phi);-\frac{1}{2}H(n,\infty)\right), & n\leqslant r+n_0,\\ E(-\frac{1}{2}\gamma^{n-1}(\Psi);-\frac{1}{2}[H(0,n-n_0-1)+H(n,\infty)]), & n>r+n_0; \end{cases}$$

$$\Psi=\sum_{j=1-r-n_0}^{0}\gamma^j(\Phi).$$

Proof: We follow the proof of Lemma 6.3 in Ref. 1.

(i) is trivial, since $\mathcal{E}$ is faithful and K is invertible ($K^{-1}=E(\frac{1}{2}\Phi;\frac{1}{2}H(1,\infty))^*$).

(ii) Define $K_n'=K\gamma(K)\cdots\gamma^{n-1}(K)$. Then $\mathcal{L}^n(Q)=\gamma^{-n}\mathcal{E}_n(K_n'^*QK_n')$. Indeed, for $n=1$ it is true by definition. By induction,

$$\begin{aligned}\mathcal{L}^n(Q)&=\gamma^{-n+1}\mathcal{E}_{n-1}(K_{n-1}'^*\mathcal{L}(Q)K_{n-1}')\\&=\gamma^{-n+1}\mathcal{E}_{n-1}\gamma^{-1}\mathcal{E}((K\gamma(K_{n-1}'))^*QK\gamma(K_{n-1}'))\\&=\gamma^{-n}\mathcal{E}_n(K_n'^*QK_n'),\end{aligned}$$

since $\mathcal{E}_n|_{A_{[1,\infty)}}=\gamma\mathcal{E}_{n-1}\gamma^{-1}$ and $\mathcal{E}_n\mathcal{E}=\mathcal{E}_n$.

Using the identity $E(Q_1;Q_2+R)E(Q_2;R)=E(Q_1+Q_2;R)$, we obtain

$$K_n'=E\left(-\frac{1}{2}\sum_{j=0}^{n-1}\gamma^j(\Phi);-\frac{1}{2}H(n,\infty)\right)$$

$$\begin{aligned}(\text{for } n>r+n_0)&=E\left(-\frac{1}{2}\sum_{j=n-r-n_0}^{n-1}\gamma^j(\Phi);-\frac{1}{2}\left[\sum_{j=0}^{n-r-n_0-1}\gamma^j(\Phi)+H(n,\infty)\right]\right)\\&\quad\times E\left(-\frac{1}{2}\sum_{j=0}^{n-r-n_0-1}\gamma^j(\Phi);-\frac{1}{2}H(n,\infty)\right)\\&=E\left(-\frac{1}{2}\gamma^{n-1}(\Psi);-\frac{1}{2}[H(0,n-n_0-1)+H(n,\infty)]\right)\\&\quad\times E\left(-\frac{1}{2}H(0,n-n_0-1);-\frac{1}{2}H(n,\infty)\right)\\&=E\left(-\frac{1}{2}\gamma^{n-1}(\Psi);-\frac{1}{2}[H(0,n-n_0-1)+H(n,\infty)]\right)\\&\quad\times e^{-(1/2)H(0,n-n_0-1)},\end{aligned}$$

since $H(0,n-n_0-1)$ commute with $A_{[n,\infty)}$. It remains to note that $\mathcal{E}_n(xy)=\mathcal{E}_n(yx)$ for any $x\in A_{[0,n-n_0-1]}$ and $y\in A_{[0,\infty)}$ by Assumption II.1 (i). □

Lemma III.6: There exist constants $C,q>0$ such that, for $Q>0$,

(i) $\alpha(\mathcal{L}^n(Q))\leqslant C\alpha(Q)$;

(ii) $\alpha_l(\mathcal{L}^n(Q))\leqslant C(\alpha(Q)\delta_{l-r}+\alpha_{n+l}(Q))$, for $l\geqslant r$, where $\delta_{l-r}=q^{[(l-r+1)/(n_0+r)]}/(1+[(l-r+1)/(n_0+r)])!$ (see Lemma III.4);

(iii) if $|||Q|||_{M,x}\|Q^{-1}\|\leqslant a$, then $\alpha(\mathcal{L}^n(Q))\leqslant C$ for any $n\geqslant N(a,M,x)$ for a constant $N(a,M,x)$, depending on $a>0$, $x>1$, $M\in\mathbb{N}$.

Proof: See Ref. 1, Lemma 6.4. We give a proof to demonstrate that all the Assumptions II.1 are applied.

(i) Let K_n be as in Lemma III.5. By Lemma III.4, there exists a constant C such that $\|K_n\|$, $\|K_n^{-1}\|\leqslant C$.

Let $p_n=\tau(e^{-H(0,n-n_0-1)})$. Then, by Assumption II.1(ii), $\mathcal{E}_n(e^{-H(0,n-n_0-1)})=p_n1$, so that

$$\|(K_n^*QK_n)^{-1}\|^{-1}\leqslant\frac{1}{p_n}\,\mathcal{L}^n(Q)\leqslant\|K_n^*QK_n\|,$$

hence $\alpha(\mathcal{L}^n(Q))\leqslant C^4\alpha(Q)$.

(ii) Let $Q_{n+l}\in A_{[0,n+l]}$, $Q_{n+l}>0$, $\alpha_{n+l}(Q)=\|Q-Q_{n+l}\|\,\|Q_{n+l}^{-1}\|$ [Lemma III.3(ii)],

$$K_{(n,l)}=E(-\tfrac{1}{2}\gamma^{n-1}(\Psi);-\tfrac{1}{2}[H(0,n-n_0-1)+H(n,n+l)]),$$

$Q_l'=\gamma^{-n}\mathcal{E}_n(K_{(n,l)}^*Q_{n+l}K_{(n,l)}e^{-H(0,n-n_0-1)})$. By Assumption II.1(iii), $Q_l'\in A_{[0,l]}$. Let us compute $\|\mathcal{L}^n(Q)-Q_l'\|\,\|(Q_l')^{-1}\|$.

By Lemma III.4,

$$\|K_{(n,l)}\|,\|K_{(n,l)}^{-1}\|\leqslant C\quad\text{and}\quad\|K_n-K_{(n,l)}\|\leqslant C\delta_{l-r}.$$

Hence,

$$\|K_n^*QK_n-K_{(n,l)}^*QK_{(n,l)}\|\leqslant 2C^2\delta_{l-r}\|Q\|\quad\text{and}\quad\|K_{(n,l)}^*(Q-Q_{n+l})K_{(n,l)}\|\leqslant C^2\|Q-Q_{n+l}\|.$$

Being a completely positive unital mapping, $A_{[0,\infty)}\ni x\mapsto(1/p_n)\mathcal{E}_n(xe^{-H(0,n-n_0-1)})$ has norm one, so that

$$\|\mathcal{L}^n(Q)-Q_l'\|\leqslant p_nC^2(2\,\delta_{l-r}\|Q\|+\|Q-Q_{n+l}\|).\tag{1}$$

On the other hand, $K_{(n,l)}^*Q_{n+l}K_{(n,l)}\geqslant\|(K_{(n,l)}^*Q_{n+l}K_{(n,l)})^{-1}\|^{-1}\geqslant C^{-2}\|Q_{n+l}^{-1}\|^{-1}$. Hence

$$Q_l'\geqslant\frac{p_n}{C^2}\,\|Q_{n+l}^{-1}\|^{-1}.\tag{2}$$

Using (1) and (2), we obtain

$$\alpha_l(\mathcal{L}^n(Q))\leqslant\|\mathcal{L}^n(Q)-Q_l'\|\,\|(Q_l')^{-1}\|\leqslant C^4(2\,\delta_{l-r}\|Q\|\,\|Q_{n+l}^{-1}\|+\alpha_{n+l}(Q)).$$

Since $\|Q\|\,\|Q_{n+l}^{-1}\|\leqslant 2\alpha(Q)$ by Lemma III.3(i), (ii), the proof of (ii) is complete.

(iii) Let $Q_k\in A_{[0,k]}$, $Q_k>0$, $\|Q-Q_k\|\,\|Q_k^{-1}\|=\alpha_k(Q)$,

$$K_{n,l}=E(-\tfrac{1}{2}\gamma^{n-1}(\Psi);-\tfrac{1}{2}[H(n-l,n-n_0-1)+H(n,\infty)]),$$

$Q'=\gamma^{-n}\mathcal{E}_n(K_{n,l}^*Q_kK_{n,l}e^{-H(0,n-n_0-1)})$. Then, as in the proof of (ii), we conclude that, for a constant C depending only on the potential, we have $\|K_{n,l}\|$, $\|K_{n,l}^{-1}\|\leqslant C$ and

$$\|\mathcal{L}^n(Q)-Q'\|\,\|(Q')^{-1}\|\leqslant C(\alpha(Q)\,\delta_{l-r-n_0}+\alpha_k(Q)).\tag{3}$$

We denote the right-hand part of (3) by Δ. Then

$$\mathcal{L}^n(Q)\leqslant Q'+\|\mathcal{L}^n(Q)-Q'\|\leqslant Q'+\|\mathcal{L}^n(Q)-Q'\|\,\|(Q')^{-1}\|Q'\leqslant(1+\Delta)Q',$$

$$\mathcal{L}^n(Q)\geqslant Q'-\|\mathcal{L}^n(Q)-Q'\|\geqslant Q'-\|\mathcal{L}^n(Q)-Q'\|\,\|(Q')^{-1}\|Q'\geqslant(1-\Delta)Q',$$

so that

$$\alpha(\mathcal{L}^n(Q))\leqslant\frac{1+\Delta}{1-\Delta}\,\alpha(Q'),\quad\text{whenever }\Delta<1.\tag{4}$$

Since $|||Q|||_{M,x}\|Q^{-1}\|\leqslant a$, we have $\alpha(Q)\leqslant a$ and $\|Q\|_k\|Q^{-1}\|\leqslant ax^{-k}$ for $k\geqslant M$, hence [Lemma III.3(iii)] $\alpha_k(Q)\leqslant ax^{-k}/(1-ax^{-k})$. We see that we can choose $l\geqslant r+n_0$ and k such that $\Delta\leqslant\frac{1}{2}$ independently of Q. Let $N(a,M,x)=k+l+n_0+1$. Then, for $n\geqslant N$, $K_{n,l}$ and Q_k commute. If $p=\tau(Q_ke^{-H(0,n-n_0-1)})$, the mapping $y\mapsto(1/p)\mathcal{E}_n(Q_k^{1/2}yQ_k^{1/2}e^{-H(0,n-n_0-1)})$ is completely positive and unital, so that

$$\frac{1}{C^2}\leqslant\|(K_{n,l}^*K_{n,l})^{-1}\|^{-1}\leqslant\frac{1}{p}\,\mathcal{E}_n(K_{n,l}^*K_{n,l}Q_ke^{-H(0,n-n_0-1)})\leqslant\|K_{n,l}^*K_{n,l}\|\leqslant C^2,$$

hence $\alpha(Q')\leqslant C^4$. By virtue of this inequality, the inequality (4) and the choice of N, we obtain $\alpha(\mathcal{L}^n(Q))\leqslant 3C^4$. □

The operator $\psi\mapsto\psi(\mathcal{L}(1))^{-1}\psi\mathcal{L}$ is defined on the state space of $A_{[0,\infty)}$. By the Schauder theorem, there exists a state ν of $A_{[0,\infty)}$ such that $\nu\mathcal{L}=\lambda\nu$ with $\lambda=\nu(\mathcal{L}(1))>0$.

Let $L=\lambda^{-1}\mathcal{L}$.

Corollary III.7: The sequence $\{L^n\}_{n=1}^{\infty}$ is bounded.

Proof: First let $Q>0$ and $\alpha(Q)\leqslant 2$. By Lemma III.6(i), $\alpha(L^n(Q))=\alpha(\mathcal{L}^n(Q))\leqslant 2C$ for a constant C independent of Q. Then

$$\|L^n(Q)\|=\alpha(L^n(Q))\|L^n(Q)^{-1}\|^{-1}\leqslant 2C\nu(L^n(Q))=2C\nu(Q)\leqslant 2C\|Q\|.$$

For arbitrary $Q\geqslant 0$, let $Q'=Q+\|Q\|1$. Then $\alpha(Q')\leqslant 2$. Hence

$$\|L^n(Q)\|\leqslant\|L^n(Q')\|+\|Q\|\,\|L^n(1)\|\leqslant 6C\|Q\|.$$

So for any $Q\in A_{[0,\infty)}$, we have $\|L^n(Q)\|\leqslant 24C\|Q\|$. □

Proposition III.8: There exists an element $h\in\cap_{x>1}A(x)$ such that $h>0$, $\nu(h)=1$, $Lh=h$. Then

$$|||L^n(Q)-\nu(Q)h|||_{1,x}\leqslant C_xe^{-q_xn}|||Q|||_{1,x},\quad \textit{for any } Q\in A(x)$$

for constants C_x, $q_x>0$ depending on x.

Proof: See Lemmas 6.5, 7.5, 7.6 in Ref. 1.

Step 1. Existence of h.

Let C be chosen as in the formulation of Lemma III.6, and δ_l be as defined there. Let X be the norm closure of the convex hull of $\{L^n(1)\}_{n=1}^{\infty}$. The set X is L-invariant. We prove that X is compact.

Let $Y=\{Q\in A_{[0,\infty)}|\ \|Q\|\leqslant C,\ \|Q\|_l\leqslant C\delta_{l-r}/(1-C\delta_{l-r}),\ l\geqslant r\}$. The set Y is compact, since the compact ball of radius C in $A_{[0,l]}$ is a $2C\delta_{l-r}/(1-C\delta_{l-r})$-net for Y. X is a subset of Y. Indeed, it is enough to show that $L^n(1)\in Y$ for any n. We have

$$\|L^n(1)\|=\alpha(L^n(1))\|L^n(1)^{-1}\|^{-1}\leqslant\alpha(L^n(1))\nu(L^n(1))=\alpha(L^n(1))\leqslant C$$

and

$$\|L^n(1)\|_l\leqslant\frac{\alpha_l(L^N(1))}{1-\alpha_l(L^n(1))}\,\|L^n(1)^{-1}\|^{-1}\leqslant\frac{\alpha_l(L^n(1))}{1-\alpha_l(L^n(1))}\leqslant\frac{C\delta_{l-r}}{1-C\delta_{l-r}},$$

by Lemma III.6(i), (ii) and Lemma III.3(iv).

By the Schauder theorem, there exists $h\in X$ such that $Lh=h$. We have $h\geqslant 0$, $\nu(h)=1$, and $h\in\cap_{x>1}A(x)$, since $Y\subset\cap_{x>1}A(x)$. Since

$$L^n(1)\geqslant\frac{\|L^n(1)\|}{\alpha(L^n(1))}\geqslant\frac{\nu(L^n(1))}{C}=\frac{1}{C},$$

we have $h\geqslant C^{-1}$.

Step 2. Continuity of $L|_{A(x)}$.

Consider $Q=Q^*$, $|||Q|||_{1,x}\leqslant 1$. Let $Q'=2+Q$. Then $1\leqslant Q'\leqslant 3$ and $\|Q'\|_l=\|Q\|_l$ for any l. By Lemma III.6(i) we obtain $\alpha(L(Q'))\leqslant 3C$, and, as in Step 1, we have

$$\|L(Q')\|_l\leqslant\frac{\alpha_l(L(Q'))}{1-\alpha_l(L(Q'))}\,\nu(Q')\leqslant\frac{3}{2}\,(3C+1)\alpha_l(L(Q')),\tag{5}$$

where the latter inequality follows from Lemma III.3(i). By Lemma III.6(ii), for $l\geqslant r$,

$$\alpha_l(L(Q'))\leqslant C(3\delta_{l-r}+\alpha_{l+1}(Q')),$$

and by Lemma III.3(iii),

$$\alpha_{l+1}(Q')\leqslant\frac{\|Q'\|_{l+1}\|(Q')^{-1}\|}{1-\|Q'\|_{l+1}\|(Q')^{-1}\|}\leqslant\frac{\|Q'\|_{l+1}}{1-\|Q'\|_{l+1}}\leqslant\frac{\|Q'\|_{l+1}}{1-x^{-1}},$$

so that

$$\alpha_l(L(Q'))\leqslant C\left(3\delta_{l-r}+\frac{\|Q'\|_{l+1}}{1-x^{-1}}\right). \tag{6}$$

Using (5), (6), and $\Sigma_l\|Q'\|_l x^l\leqslant 1$, we obtain

$$\begin{aligned}|||L(Q')|||_{r,x}&=\|L(Q')\|+\sum_{l=r}^{\infty}\|L(Q')\|_l x^l\\&\leqslant 3\|L(1)\|+\frac{3}{2}(3C+1)\sum_{l=r}^{\infty}C\left(3\delta_{l-r}+\frac{\|Q'\|_{l+1}}{1-x^{-1}}\right)x^l\\&\leqslant 3\|L(1)\|+\frac{9C(3C+1)}{2}\sum_{l=r}^{\infty}\delta_{l-r}x^l+\frac{3C(3C+1)}{2(x-1)}.\end{aligned}$$

In particular, $|||L(1)|||_{r,x}<\infty$. Since $|||L(Q)|||_{r,x}\leqslant 2|||L(1)|||_{r,x}+|||L(Q')|||_{r,x}$, the continuity of $L|_{A(x)}$ is proved.

Step 3. Convergence proof.

Let C be chosen as in the formulation of Lemma III.6. Then we take

$$a\geqslant 4C,$$

and choose $M\in\mathbb{N}$ such that

$$2Cax^{-M}\leqslant\frac{1}{4},\quad Ca\delta_{l-r}\leqslant\frac{1}{4},\quad\text{for}\quad l\geqslant M,\quad 4C\sum_{l=M}^{\infty}\delta_{l-r}x^l\leqslant\frac{1}{4}.$$

Finally, we fix $N\in\mathbb{N}$ such that

$$8Cx^{-N}\leqslant\tfrac{1}{4}\quad\text{and}\quad N\geqslant N(a,M,x),$$

where $N(a,M,x)$ is as in the formulation of Lemma III.6(iii).

We define an auxiliary operator Λ on $A(x)$,

$$\Lambda(Q)=L^N(Q)-\frac{\nu(Q)}{2C}.$$

We state that

$$\text{if}\ \ Q>0,|||Q|||_{M,x}\|Q^{-1}\|\leqslant a,\quad\text{then}\ \ \Lambda(Q)>0,\quad |||\Lambda(Q)|||_{M,x}\|\Lambda(Q)^{-1}\|\leqslant a. \tag{7}$$

Suppose the statement is proved. Let $q_x=-(1/N)\log(1-1/2C)>0$. Then $\nu(\Lambda(Q))=(1-1/2C)\nu(Q)=e^{-q_xN}\nu(Q)$. In particular, for $\nu(Q)=0$, we have $\Lambda^n(Q)=L^{Nn}(Q)$.

Consider $Q=Q^*\in A(x)$, $|||Q|||_{M,x}\leqslant 1$, $\nu(Q)=0$. Define $Q_1=2+Q$, $Q_2=2$. Then $1\leqslant Q_1\leqslant 3$, $|||Q_1|||_{M,x}\|Q_1^{-1}\|\leqslant|||Q_1|||_{M,x}\leqslant 3\leqslant a$. By (7),

$$|||\Lambda^n(Q_1)|||_{M,x}\leqslant a\|\Lambda^n(Q_1)^{-1}\|^{-1}\leqslant a\,\nu(\Lambda^n(Q_1))=ae^{-q_xNn}\nu(Q_1)=2ae^{-q_xNn},$$

analogously $|||\Lambda^n(2)|||_{M,x}\leqslant 2ae^{-q_xNn}$, so that

$$|||L^{Nn}(Q)|||_{M,x}=|||\Lambda^n(Q)|||_{M,x}=|||\Lambda^n(Q_1)-\Lambda^n(Q_2)|||_{M,x}\leqslant 4ae^{-q_xNn}.$$

For arbitrary $k\in\mathbb{N}$ we obtain

$$|||L^k(Q)|||_{M,x}\leqslant|||L|||_{M,x}^{k-[k/N]N}4ae^{-q_x[k/N]N}\leqslant 4ae^{q_xN}(1+|||L|||_{M,x})^Ne^{-q_xk},$$

where $|||L|||_{M,x}$ is the norm of the operator L on the Banach space $(A(x),|||\ |||_{M,x})$ that is finite by Step 2. Let $\tilde{C}_x=4ae^{q_xN}(1+|||L|||_{M,x})^N$. Then, for $Q=Q^*\in A(x)$, we have

$$\begin{aligned}|||L^k(Q)-\nu(Q)h|||_{M,x}&=|||L^k(Q-\nu(Q)h)|||_{M,x}\\&\leqslant\tilde{C}_xe^{-q_xk}|||Q-\nu(Q)h|||_{M,x}\\&\leqslant\tilde{C}_x(1+|||h|||_{M,x})e^{-q_xk}|||Q|||_{M,x},\end{aligned}$$

so that, for arbitrary $Q\in A(x)$,

$$|||L^k(Q)-\nu(Q)h|||_{M,x}\leqslant 2\tilde{C}_x(1+|||h|||_{M,x})e^{-q_xk}|||Q|||_{M,x}.$$

It remains to prove (7). We have $\alpha(L^N(Q))\leqslant C$ by Lemma III.6(iii) and the choice of N. Hence

$$L^N(Q)\geqslant\frac{\|L^N(Q)\|}{\alpha(L^N(Q))}\geqslant\frac{\|L^N(Q)\|}{C}\geqslant\frac{\nu(Q)}{C}.$$

In particular, $\Lambda(Q)\geqslant\nu(Q)/2C>0$ and $[\nu(Q)/C]\|L^N(Q)^{-1}\|\leqslant 1$. Further,

$$\begin{aligned}\alpha(\Lambda(Q))&=\frac{\|L^N(Q)\|-\dfrac{\nu(Q)}{2C}}{\|L^N(Q)^{-1}\|^{-1}-\dfrac{\nu(Q)}{2C}}=\frac{\alpha(L^N(Q))-\dfrac{\nu(Q)}{2C}\|L^N(Q)^{-1}\|}{1-\dfrac{\nu(Q)}{2C}\|L^N(Q)^{-1}\|}\\&\leqslant\frac{\alpha(L^N(Q))}{1-\dfrac{1}{2}}=2\alpha(L^N(Q))\leqslant 2C\leqslant\frac{a}{2}\end{aligned}\tag{8}$$

and

$$\begin{aligned}\|\Lambda(Q)\|_l\|\Lambda(Q)^{-1}\|&=\|L^N(Q)\|_l\|\Lambda(Q)^{-1}\|\\&=\|L^N(Q)\|_l\left(\|L^N(Q)^{-1}\|^{-1}-\frac{\nu(Q)}{2C}\right)^{-1}\\&\leqslant 2\|L^N(Q)\|_l\|L^N(Q)^{-1}\|\leqslant\frac{2\alpha_l(L^N(Q))}{1-\alpha_l(L^N(Q))},\end{aligned}\tag{9}$$

where we have used Lemma III.3(iv) in the latter inequality. By Lemma III.6(ii),

$$\alpha_l(L^N(Q))\leqslant C(\alpha(Q)\delta_{l-r}+\alpha_{l+N}(Q)).\tag{10}$$

Since $\|Q\|_{l+N}\|Q^{-1}\|\leqslant ax^{-l-N}$ and $\alpha_{l+N}(Q)\leqslant\|Q\|_{l+N}\|Q^{-1}\|/(1-\|Q\|_{l+N}\|Q^{-1}\|)$ by Lemma III.3(iii), $C\alpha_{l+N}(Q)\leqslant C[ax^{-M}/(1-ax^{-M})]\leqslant\frac{1}{4}$. We also have $C\alpha(Q)\delta_{l-r}\leqslant Ca\,\delta_{l-r}\leqslant\frac{1}{4}$ for $l\geqslant M$, so that $\alpha_l(L^N(Q))\leqslant\frac{1}{2}$ for any $l\geqslant M$. Together with (9), this gives

$$\|\Lambda(Q)\|_l\|\Lambda(Q)^{-1}\|\leqslant 4\alpha_l(L^N(Q)).$$

Now it follows from (10) again that

$$\|\Lambda(Q)\|_l\|\Lambda(Q)^{-1}\|\leqslant 4C(\alpha(Q)\delta_{l-r}+\alpha_{l+N}(Q))$$
$$\leqslant 4Ca\,\delta_{l-r}+4C\,\frac{\|Q\|_{l+N}\|Q^{-1}\|}{1-ax^{-M}}\leqslant 4Ca\,\delta_{l-r}+8C\|Q\|_{l+N}\|Q^{-1}\|. \quad (11)$$

Finally, using (8), (11), and the choice of M and N, we obtain

$$|||\Lambda(Q)|||_{M,x}\|\Lambda(Q)^{-1}\|=\alpha(\Lambda(Q))+\sum_{l=M}^{\infty}\|\Lambda(Q)\|_l\|\Lambda(Q)^{-1}\|x^l$$
$$\leqslant\frac{a}{2}+4Ca\sum_{l=M}^{\infty}\delta_{l-r}x^l+8C\sum_{l=M}^{\infty}\|Q\|_{l+N}\|Q^{-1}\|x^l$$
$$\leqslant\frac{a}{2}+\frac{a}{4}+8Cax^{-N}\leqslant a.$$

□

Corollary III.9: For any $x>1$, there exist C_x, $q_x>0$, such that

$$|||L^{N+n}(Q)-\nu(Q)h|||_{1,x}\leqslant C_xe^{-q_xn}\|Q\|, \quad \forall Q\in A_{[0,N]}, \quad \forall n,N.$$

Proof: It is enough to prove that there exists a constant C_x independent of N such that $|||L^N(Q)|||_{1,x}\leqslant C_x\|Q\|$, $\forall Q\in A_{[0,N]}$. As in the proof of Proposition III.8, Step 2, it suffices to consider $Q\in A_{[0,N]}$, $1\leqslant Q\leqslant 3$. By the same arguments as given there, we have $\alpha(L^N(Q))\leqslant 3C$ and, for $l\geqslant r$,

$$\|L^N(Q)\|_l\leqslant\frac{3C(3C+1)}{2}\,(3\,\delta_{l-r}+\alpha_{l+N}(Q)).$$

But $\alpha_{l+N}(Q)=0$ for any $Q\in A_{[0,N]}$, $Q>0$, and we obtain

$$|||L^N(Q)|||_{1,x}\leqslant rx^{r-1}\|L^N(Q)\|+\sum_{l=r}^{\infty}\|L^N(Q)\|_lx^l\leqslant 3rx^{r-1}\|L^N\|+\frac{9C(3C+1)}{2}\sum_{l=r}^{\infty}\delta_{l-r}x^l.$$

From Corollary III.7, we conclude that the latter expression is bounded by a constant $\widetilde{C}_x$ independent of Q and N. □

Lemma III.10: The sequence $\{\nu\gamma^n\}_{n=1}^{\infty}$ converges to a state ϕ on $A_{[0,\infty)}$ in norm. Moreover, there exist C, $q>0$ such that $\|\phi-\nu\gamma^n\|\leqslant Ce^{-qn}$.

Proof: For n and N, $N+n_0<n$, let

$$K_{n-N,n}=E(-\tfrac{1}{2}\gamma^{n-N-1}(\Psi);-\tfrac{1}{2}[H(0,n-N-n_0-1)+H(n-N,n-n_0-1)]).$$

According to Lemma III.4, Corollary III.7, and Proposition III.8, there exist C, $q>0$ such that

$$\|K_{n-N,n}K_{n-N}^{-1}-1\|\leqslant Ce^{-qN},$$
$$\|L^n\|\leqslant C,$$
$$\|L^n(1)-h\|\leqslant Ce^{-qn},$$

where K_{n-N}'s are defined in Lemma III.5.

Since $\gamma^n(Q)$ and $K_{n-N,n}$ commute for any $Q\in A_{[0,\infty)}$, we have

$$L^{n-N}((K_{n-N,n}K_{n-N}^{-1})*\gamma^n(Q)K_{n-N,n}K_{n-N}^{-1})=\gamma^N(Q)L^{n-N}((K_{n-N,n}K_{n-N}^{-1})*K_{n-N,n}K_{n-N}^{-1}).$$

So

$$\|L^{n-N}(\gamma^n(Q))-\gamma^N(Q)L^{n-N}(1)\|\leqslant C\|\gamma^n(Q)-(K_{n-N,n}K_{n-N}^{-1})^*\gamma^n(Q)K_{n-N,n}K_{n-N}^{-1}\|$$
$$+C\|(K_{n-N,n}K_{n-N}^{-1})^*K_{n-N,n}K_{n-N}^{-1}-1\|\ \|Q\|\leqslant C'e^{-qN}\|Q\|,$$

where $C'=2C^2(2+C)$ (note that $\|K_{n-N,n}K_{n-N}^{-1}\|\leqslant 1+C$). Then

$$\|L^{n-N}(\gamma^n(Q))-\gamma^N(Q)h\|\leqslant C'e^{-qN}\|Q\|+\|L^{n-N}(1)-h\|\ \|Q\|\leqslant(C'e^{-qN}+Ce^{-q(n-N)})\|Q\|,$$

hence $|\nu(\gamma^n(Q))-\nu(\gamma^N(Q)h)|\leqslant(C'e^{-qN}+Ce^{-q(n-N)})\|Q\|$, and

$$\|\nu\gamma^n-\nu\gamma^m\|\leqslant 2C'e^{-qN}+Ce^{-q(n-N)}+Ce^{-q(m-N)},$$

for n, $m>N+n_0$. We see that the sequence $\{\nu\gamma^n\}_n$ converges. Denoting its limit by ϕ and letting $m\to\infty$ in the latter inequality, we obtain

$$\|\nu\gamma^n-\phi\|\leqslant 2C'e^{-qN}+Ce^{-q(n-N)}.$$

Taking $N=[\frac{1}{2}n]$, we obtain the desired. □

The state ϕ of $A_{[0,\infty)}$ is obviously γ-invariant. Hence there exists a unique γ-invariant state of A that, being restricted to $A_{[0,\infty)}$, coincides with ϕ. We denote this state by the same letter ϕ, and our aim is to prove that ϕ is the required Gibbs state.

First, we shall prove that ϕ satisfies the conclusion of Theorem III.2.

Proposition III.11: There exist C, $q>0$, such that

$$|\nu(Q_1Q_2)-\nu(Q_1)\nu(Q_2)|\leqslant Ce^{-qn}\|Q_1\|\ \|Q_2\|,$$

for $Q_1\in A_{[0,N]}$ and $Q_2\in A_{[N+n,\infty)}$ for any N, n.

Proof: Replacing Q_1 by $Q_1-\nu(Q_1)1$ we can suppose that $\nu(Q_1)=0$.

By the same arguments as in Lemma III.10, we can find C, q such that

$$\|L^{N+m}(Q_1Q_2)-\gamma^{-N-m}(Q_2)L^{N+m}(Q_1)\|\leqslant Ce^{-q(n-m)}\|Q_1\|\ \|Q_2\|,$$

for $m<n-n_0$. On the other hand, $\|L^{N+m}(Q_1)\|\leqslant Ce^{-qm}\|Q_1\|$ by Corollary III.9. Hence

$$\|L^{N+m}(Q_1Q_2)\|\leqslant C(e^{-qm}+e^{-q(n-m)})\|Q_1\|\ \|Q_2\|,$$

so that

$$|\nu(Q_1Q_2)|\leqslant C(e^{-qm}+e^{-q(n-m)})\|Q_1\|\ \|Q_2\|.$$

Taking $m=[\frac{1}{2}n]$ we obtain the required. □

Now we can prove that ϕ is uniformly exponentially clustering in the sense of Theorem III.2:

If C, q are chosen according to Proposition III.11 then, for $m>N>n$, $Q_1\in A_{[-N,-n]}$, $Q_2\in A_{[n,N]}$, we have

$$|\nu(\gamma^m(Q_1Q_2)-\nu(\gamma^m(Q_1))\nu(\gamma^m(Q_2))|\leqslant Ce^{-2qn}\|Q_1\|\ \|Q_2\|.$$

Letting $m\to\infty$, we obtain

$$|\phi(Q_1Q_2)-\phi(Q_1)\phi(Q_2)|\leqslant Ce^{-2qn}\|Q_1\|\ \|Q_2\|.$$

Lemma III.12: Let $\overline{Z}_n=\tau(e^{-H(0,n)-\gamma^n(\Delta^+)})$ and $\overline{\phi}_n(Q)=\overline{Z}_n^{-1}\tau(Qe^{-H(0,n)-\gamma^n(\Delta^+)})$ for $Q\in A_{[0,\infty)}$. Then

(i) $\lim_{n\to\infty}(1/n)\ \log\overline{Z}_n=\log\lambda$;

(ii) there exist C, $q>0$ such that

$$|\nu(Q)-\overline{\phi}_{n+N}(Q)|\leqslant Ce^{-qn}\|Q\|,\quad \forall Q\in A_{[0,N]},\quad \forall N,n.$$

Proof: See Lemma 8.1 in Ref. 1.

Let $Q\in A_{[0,N]}$. Proving (ii) we can suppose that $Q>0$ and $\alpha(Q)\leqslant 2$ as in Corollary III.7. For $n>m+r$, let

$$K_{N+m,N+n}=E(-\tfrac{1}{2}\gamma^{N+m-1}(\Psi);-\tfrac{1}{2}[H(0,N+m-n_0-1)+H(N+m,N+n)+\gamma^{N+n}(\Delta^+)]).$$

Then

$$e^{-(1/2)H(0,N+n)-(1/2)\gamma^{N+n}(\Delta^+)}=K_{N+m,N+n}e^{-(1/2)H(0,N+m-n_0-1)-(1/2)H(N+m,N+n)-(1/2)\gamma^{N+n}(\Delta^+)},$$

so that

$$\begin{aligned}\overline{\phi}_{N+n}(Q)&=\overline{Z}_{N+n}^{-1}\tau(K_{N+m,N+n}^*QK_{N+m,N+n}e^{-H(0,N+m-n_0-1)-H(N+m,N+n)-\gamma^{N+n}(\Delta^+)})\\&=\overline{Z}_{N+n}^{-1}\tau(\gamma^{-N-m}\mathcal{E}_{N+m}(K_{N+m,N+n}^*QK_{N+m,N+n}e^{-H(0,N+m-n_0-1)})e^{-H(0,n-m)-\gamma^{n-m}(\Delta^+)})\\&=\overline{Z}_{N+n}^{-1}\overline{Z}_{n-m}\overline{\phi}_{n-m}(\gamma^{-N-m}\mathcal{E}_{N+m}(K_{N+m,N+n}^*QK_{N+m,N+n}e^{-H(0,N+m-n_0-1)})).\end{aligned}$$

By Corollary III.9 and a version of Lemma III.4, there exist C, $q>0$ such that

$$\|K_{N+m,N+n}\|,\ \|K_{N+m}\|,\ \|K_{N+m}^{-1}\|\leqslant C,$$

$$\|L^{N+m}(Q)-\nu(Q)h\|\leqslant Ce^{-qm}\|Q\|,$$

$$\|K_{N+m}-K_{N+m,N+n}\|\leqslant Ce^{-q(n-m)}.$$

Let $p_{N+m}=\tau(e^{-H(0,N+m-n_0-1)})$. Then $x\mapsto p_{N+m}^{-1}\mathcal{E}_{N+m}(xe^{-H(0,N+m-n_0-1)})$, being a unital completely positive mapping, has norm one. From this observation, the choice of C, and Lemma III.5, we infer

$$\begin{aligned}&\|\gamma^{-N-m}\mathcal{E}_{N+m}(K_{N+m,N+n}^*QK_{N+m,N+n}e^{-H(0,N+m-n_0-1)})-\mathcal{L}^{N+m}(Q)\|\\&\quad\leqslant p_{N+m}\|K_{N+m,N+n}^*QK_{N+m,N+n}-K_{N+m}^*QK_{N+m}\|\leqslant p_{N+m}2C^2e^{-q(n-m)}\|Q\|,\end{aligned}$$

and $\mathcal{L}^{N+m}(Q)\geqslant p_{N+m}\|K_{N+m}^{-1}\|^{-2}\|Q^{-1}\|^{-1}$, so that $\|\mathcal{L}^{N+m}(Q)^{-1}\|\leqslant p_{N+m}^{-1}C^2\|Q^{-1}\|$.

Hence

$$\begin{aligned}&\|\lambda^{-N-m}\gamma^{-N-m}\mathcal{E}_{N+m}(K_{N+m,N+n}^*QK_{N+m,N+n}e^{-H(0,N+m-n_0-1)})-L^{N+m}(Q)\|\|L^{N+m}(Q)^{-1}\|\\&\quad\leqslant 2C^4e^{-q(n-m)}\alpha(Q)\leqslant 4C^4e^{-q(n-m)}.\end{aligned}$$

Applying $\overline{\phi}_{n-m}$, we obtain

$$\begin{aligned}&|\lambda^{-N-m}\overline{Z}_{N+n}\overline{Z}_{n-m}^{-1}\overline{\phi}_{N+n}(Q)-\overline{\phi}_{n-m}(L^{N+m}(Q))|\\&\quad\leqslant 4C^4e^{-q(n-m)}\|L^{N+m}(Q)^{-1}\|^{-1}\leqslant 4C^4e^{-q(n-m)}\nu(L^{N+m}(Q))\leqslant 4C^4e^{-q(n-m)}\|Q\|.\end{aligned}$$

Hence

$$|\lambda^{-N-m}\overline{Z}_{N+n}\overline{Z}_{n-m}^{-1}\overline{\phi}_{N+n}(Q)-\nu(Q)\overline{\phi}_{n-m}(h)|\leqslant(4C^4e^{-q(n-m)}+Ce^{-qm})\|Q\|. \tag{12}$$

In particular,

$$|\lambda^{-N-m}\overline{Z}_{N+n}\overline{Z}_{n-m}^{-1}-\overline{\phi}_{n-m}(h)|\leqslant 4C^4e^{-q(n-m)}+Ce^{-qm}. \tag{13}$$

Having fixed $n-m$ sufficiently large, we see from $\|\underline{h}^{-1}\|^{-1}\leqslant\overline{\phi}_{n-m}(h)\leqslant\|h\|$ and the latter inequality that the sequence $\{-(N+n)\log\lambda+\log Z_{N+n}\}_n$ is bounded. In particular, $\lim_{n\to\infty}n^{-1}\log Z_n=\log\underline{\lambda}$. So we have proved (i).

Multiplying (13) by $\phi_{N+n}(Q)$ we also deduce from (12) and (13) that

$$|\overline{\phi}_{N+n}(Q)-\nu(Q)|\leqslant 2(4C^4e^{-q(n-m)}+Ce^{-qm})\|Q\|\|h^{-1}\|.$$

Taking $m=[\frac{1}{2}n]$, we obtain (ii). □

Proof of Theorem III.1: Let $Q\in A_{[a,b]}$,

$$\begin{aligned}\phi_{[-n+a,b+n]}(Q)&=\frac{\tau(Qe^{-U(-n+a,b+n)})}{\tau(e^{-U(-n+a,b+n)})}\\&=\frac{\tau(\gamma^{n-a}(Q)e^{-U(0,b-a+2n)})}{\tau(e^{-U(0,b-a+2n)})}\\&=\frac{\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\gamma^{n-a}(Q)\Delta_{b-a+2n})}{\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\Delta_{b-a+2n})},\end{aligned}$$

where $\overline{\phi}_{b-a+2n}$ is introduced in Lemma III.12,

$$\Delta_{b-a+2n}=E(-\tfrac{1}{2}\Delta^-;-\tfrac{1}{2}[H(0,b-a+2n)+\gamma^{b-a+2n}(\Delta^+)]).$$

Let C, q be chosen as in Lemma III.10, Proposition III.11 and Lemma III.12. By a version of Lemma III.4, we can also suppose that

$$\|\Delta_m\|,\|\Delta_m^{-1}\|\leqslant C \quad\text{and}\quad \|\Delta_m-\Delta_k\|\leqslant Ce^{-q\,\min\{m,k\}}.$$

We have the following inequalities ($m<n-n_0$):

$$|\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\gamma^{n-a}(Q)\Delta_{b-a+2n})-\overline{\phi}_{b-a+2n}(\gamma^{n-a}(Q)\Delta^*_m\Delta_m)|\leqslant 2C^2e^{-qm}\|Q\|, \quad (14)$$

since $\gamma^{n-a}(Q)$ and Δ_m commute;

$$|\overline{\phi}_{b-a+2n}(\gamma^{n-a}(Q)\Delta^*_m\Delta_m)-\nu(\gamma^{n-a}(Q)\Delta^*_m\Delta_m)|\leqslant C^3e^{-qn}\|Q\|, \quad (15)$$

by Lemma III.12,

$$|\nu(\gamma^{n-a}(Q)\Delta^*_m\Delta_m)-\nu(\gamma^{n-a}(Q))\nu(\Delta^*_m\Delta_m)|\leqslant C^3e^{-q(n-m)}\|Q\|, \quad (16)$$

by Proposition III.11,

$$|\nu(\gamma^{n-a}(Q))\nu(\Delta^*_m\Delta_m)-\phi(Q)\nu(\Delta^*_m\Delta_m)|\leqslant C^3e^{-qn}\|Q\|, \quad (17)$$

by Lemma III.10,

$$|\phi(Q)\nu(\Delta^*_m\Delta_m)-\phi(Q)\overline{\phi}_{b-a+2n}(\Delta^*_m\Delta_m)|\leqslant C^3e^{-q(b-a+2n-m)}\|Q\|\leqslant C^3e^{-qn}\|Q\|, \quad (18)$$

by Lemma III.12. Finally,

$$|\phi(Q)\overline{\phi}_{b-a+2n}(\Delta^*_m\Delta_m)-\phi(Q)\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\Delta_{b-a+2n})|\leqslant 2C^2e^{-qm}\|Q\|, \quad (19)$$

by the property of Δ_m's. Summing the inequalities (14)–(19), we obtain

$$\begin{aligned}&|\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\gamma^{n-a}(Q)\Delta_{b-a+2n})-\phi(Q)\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\Delta_{b-a+2n})|\\&\quad\leqslant\tilde{C}(e^{-qn}+e^{-qm}+e^{-q(n-m)}),\end{aligned}$$

for a constant $\tilde{C}$, depending only on the potential.

Since

$$|\overline{\phi}_{b-a+2n}(\Delta^*_{b-a+2n}\Delta_{b-a+2n})|\geqslant\|(\Delta^*_{b-a+2n}\Delta_{b-a+2n})^{-1}\|^{-1}\geqslant\frac{1}{C^2},$$

we have

$$|\phi_{[-n+a,b+n]}(Q)-\phi(Q)|\leqslant\tilde{C}C^2(e^{-qn}+e^{-qm}+e^{-q(n-m)}).$$

Taking $m=[\frac{1}{2}n]$, we obtain (ii). Since

$$\tau(e^{-U(a,b)})=\overline{Z}_{b-a}\overline{\phi}_{b-a}(\Delta^*_{b-a}\Delta_{b-a}),$$

$1/C^2\leqslant\overline{\phi}_{b-a}(\Delta^*_{b-a}\Delta_{b-a})\leqslant C^2$ and $(1/n)\log\overline{Z}_n\to\log\lambda$, we have (i). □

Remark III.13: The Gibbs state ϕ is a σ_t-KMS state ($\beta=1$). It is known that for a one-dimensional quantum lattice this is a unique σ_t-KMS state (see Ref. 5, or Ref. 4). The same result holds for binary shifts (Example II.4). Indeed, by Refs. 16 and 17, there exist $p,k\in\mathbb{N}$ such that $A_{[0,k+pn]}$ is a full matrix algebra for any $n\in\mathbb{N}$. Then one can apply Ref. 5.

IV. ENTROPIC PROPERTIES OF GIBBS STATES

In the sequel, ϕ denotes a Gibbs state on $(A,\{A_{\{n,m\}}\}_{n\leqslant m},\tau,\gamma)$ corresponding to an interaction, $\Phi(X)$.

Theorem IV.1: *Let $M=\pi_\phi(A)''$. Then (M,ϕ,γ) is an entropic K-system.*

Proof: Let M_0 be the W^*-subalgebra of M generated by $\pi_\phi(A_{(-\infty,0]})$. Using Theorem III.2, one concludes that $\cap_n\gamma^n(M_0)=\mathbb{C}1$. Hence (M,ϕ,γ) is an entropic K system by Ref. 7, Theorem 3.1. □

Corollary IV.2: If $\pi_\phi(A)\neq\mathbb{C}1$, then $h_\phi(\gamma)>0$. □

Corollary IV.3: For all the examples of Sec. II we have $0<h_\phi(\gamma)<\infty$.

Proof: Let d_n be the tracial dimension of $A_{[1,n]}$, i.e., the dimension of a maximal Abelian subalgebra in $A_{[1,n]}$. Then

$$h_\phi(\gamma)\leqslant\liminf_{n\to\infty}\frac{S(\phi|_{A_{[1,n]}})}{n}\leqslant\sup_n\frac{\log d_n}{n}.$$

So to prove the finiteness it suffices to show that $\sup_n(1/n)\log d_n<\infty$.

It is obvious for binary shifts (Example II.4), since $\dim A_{[1,n]}=2^n$. For the automorphism θ_λ (Example II.2) and for the canonical shift on the tower of relative commutants (Example II.3) we note that the tracial dimension of $M'\cap M_n$ does not exceed $[M_n:M]=[M:N]^n$.[18]

To prove that the entropy is nonzero, we have to show that each of the algebras under consideration has no ideals of codimension one. The algebra of a binary shift is simple as well as the algebra of relative commutants corresponding to a finite depth inclusion. In general, an algebra of higher relative commutants can be nonsimple. Nevertheless, it always contains a Temperley–Lieb algebra that has no ideals of codimension one. □

Next, we shall discuss mean entropy. First we need a form of the Gibbs condition (called the Gibbs condition in the strong sense in Ref. 9).

Proposition IV.4: For a finite interval $\Lambda=[k,l]\subset\mathbb{Z}$, let $W(\Lambda)$ denote the surface energy, i.e.,

$$W(\Lambda)=\sum_{\substack{X:\\ X\cap\Lambda\neq\varnothing\\ X\cap\Lambda^c\neq\varnothing}}\Phi(X)+\sum_{\substack{X\subset\Lambda^c:\\ X\cap([k-n_0,k-1]\cup[l+1,l+n_0])\neq\varnothing}}\Phi(X).$$

Then, for the perturbed state $\phi^{W(\Lambda)}$ (see, for example, Ref. 4, Section 5.4.1), we have

$$\phi^{W(\Lambda)}(ab)=\phi_\Lambda(a)\phi^{W(\Lambda)}(b),\quad\forall a\in A_\Lambda,\quad\forall b\in A_{(-\infty,k-n_0)\cup(l+n_0,\infty)}.$$

Proof: Let $K=E(\frac{1}{2}W(\Lambda);-\frac{1}{2}U(\mathbb{Z}))$,

$$K_n=E(\tfrac{1}{2}W(\Lambda);-\tfrac{1}{2}U(-n,n))=e^{(1/2)W(\Lambda)-(1/2)U(-n,n)}e^{(1/2)U(-n,n)}.$$

Then K and K_n are invertible, K_n converges to K in norm, and

$$\phi^{W(\Lambda)}=\frac{\phi(K^*\cdot K)}{\phi(K^*K)},\quad \phi^{W(\Lambda)}_{[-n,n]}=\frac{\phi_{[-n,n]}(K_n^*\cdot K_n)}{\phi_{[-n,n]}(K_n^*K_n)}.$$

Since $\phi_{[-n,n]}$ pointwise converges to ϕ, we see that $\phi^{W(\Lambda)}_{[-n,n]}$ pointwise converges to $\phi^{W(\Lambda)}$. For n sufficiently large, we have

$$\begin{aligned}\phi^{W(\Lambda)}_{[-n,n]}(ab)&=\frac{\tau(abe^{-U(-n,n)+W(\Lambda)})}{\tau(e^{-U(-n,n)+W(\Lambda)})}\\&=\frac{\tau(abe^{-U(\Lambda)}e^{-U([-n,k-n_0-1]\cup[l+n_0+1,n])})}{\tau(e^{-U(-n,n)+W(\Lambda)})}\\&=\frac{\tau(ae^{-U(\Lambda)})\tau(be^{-U([-n,k-n_0-1]\cup[l+n_0+1,n])})}{\tau(e^{-U(-n,n)+W(\Lambda)})}\\&=\frac{\tau(ae^{-U(\Lambda)})}{\tau(e^{-U(\Lambda)})}\cdot\frac{\tau(be^{-U(\Lambda)}e^{-U([-n,k-n_0-1]\cup[l+n_0+1,n])})}{\tau(e^{-U(-n,n)+W(\Lambda)})}\\&=\phi_\Lambda(a)\phi^{W(\Lambda)}_{[-n,n]}(b).\end{aligned}$$

□

With the help of the Gibbs condition the mean entropy is computed in a standard way (see Refs. 4 and 9). Namely, we have the following.

Corollary IV.5: Let P_n be the density operator for $\tau|_{A_{[1,n]}}$. Then

$$\lim_{n\to\infty}\frac{1}{n}\,(\phi(\log P_n)+S(\phi|_{A_{[1,n]}}))=\phi(\Phi)+P(\Phi)=\phi(E_\Phi)+P(\Phi),$$

where $E_\Phi=\Sigma_{X\ni 0}\Phi(X)/|X|$. In particular, the mean entropy $s(\phi)=\lim_n(1/n)S(\phi|_{A_{[1,n]}})$ exists iff the limit $\lim_n(1/n)\phi(\log P_n)$ exists.

Proof: Let Q_n be the density operator for $\phi|_{A_{[1,n]}}$. The operator $[e^{-U(1,n)}/\tau(e^{-U(1,n)})]P_n$ is the density operator for $\phi_{[1,n]}|_{A_{[1,n]}}=\phi^{W(1,n)}|_{A_{[1,n]}}$. Then by Ref. 4, p. 278,

$$0\leqslant \mathrm{Tr}_{A_{[1,n]}}\left(Q_n\left(\log Q_n-\log\left(\frac{e^{-U(1,n)}}{\tau(e^{-U(1,n)})}\,P_n\right)\right)\right)\leqslant S(\phi^{W(1,n)},\phi)\leqslant 2\|W(1,n)\|,$$

so that

$$0\leqslant -S(\phi|_{A_{[1,n]}})-\phi(\log P_n)+\phi(U(1,n))+\log\,\tau(e^{-U(1,n)})\leqslant 2\|W(1,n)\|.$$

Since $(1/n)\phi(U(1,n))\to\phi(\Phi)=\phi(E_\Phi)$, $(1/n)\log\tau(e^{-U(1,n)})\to P(\Phi)$, and the sequence $\{W(1,n)\}_n$ is bounded, we obtain the desired. □

Corollary IV.6: The mean entropy exists for each of the following cases:

(1) the automorphism θ_λ, $\lambda\geqslant\frac{1}{4}$,

$$s(\phi)=\phi(E_\Phi)+P(\Phi)-\tfrac{1}{2}\log\lambda;$$

(2) the canonical shift on the tower of relative commutants corresponding to a subfactor of finite depth,

$$s(\phi)=\phi(E_\Phi)+P(\Phi)+\log[M:N];$$

(3) a binary shift,

$$s(\phi)=\phi(E_\Phi)+P(\Phi)+\tfrac{1}{2}\log 2.$$

Proof: For all these systems the limit $\lim_n(1/n)\log P_n$ exists in norm. The result is well known for the systems in (1) and (2) (see, e.g., Ref. 9, Examples 1.3, 1.4). Concerning binary shifts the sequence $\{A_{[1,n]}\}_n$ is periodic.[17] Hence, the limit $\lim_n -(1/n)\log P_n$ exists, is scalar, and equals the entropy $h_\tau(\gamma)$ of γ with respect to τ.[19] This entropy is equal to $\frac{1}{2}\log 2$.[19] □

Remark IV.7: It is not known whether the mean entropy equals the dynamical entropy, even for a quantum spin lattice system. But such an equality is possible. This is the case when an interaction potential lies in the diagonal (then, by the general theory,[8] the dynamical entropy coincides with entropy of the restriction of the automorphism to the diagonal). A bit less trivial example can be obtained as follows.

Suppose an interaction potential on a one-dimensional quantum lattice has the property $\sigma_t(A_0)\subset A_{[-n,n]}$ for some $n\in\mathbb{N}$. Then it is easy to see that $h_\phi(\gamma)=s(\phi)$ (in fact, the state ϕ is n-Markov, i.e., it would be Markov if we consider A as a quantum lattice system with the one-site algebra $A_{[0,n-1]}$,[20] and we can refer to Ref. 21; see also Ref. 7). One can construct potentials with such a property, but that do not lie in the diagonal. For example, we can choose self-adjoint a_0, a_1, $b\in \mathrm{Mat}_d(\mathbb{C})$, $d\geqslant 3$, such that a_0 commutes with a_1 and b, but a_1 and b do not commute. Define $\Phi(\{0\})=b$, $\Phi(\{0,1\})=a_0\otimes a_1$, $\Phi(X)=0$ for $|X|\geqslant 3$. Then $\sigma_t(A_0)\subset A_{[-1,1]}$, but the centralizer A^σ does not contain any diagonal of A_0.

[1] H. Araki, ''Gibbs states of a one dimensional quantum lattice,'' Commun. Math. Phys. **14**, 120–157 (1969).
[2] D. Ruelle, ''Statistical mechanics of a one-dimensional lattice gas,'' Commun. Math. Phys. **9**, 267–278 (1968).
[3] H. Araki, ''Expansionals in Banach algebras,'' Ann. Sci. Ecole Norm. Sup. Ser. 4 **6**, 67–84 (1973).
[4] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II* (Springer-Verlag, New York, 1987).
[5] H. Araki, ''On uniqueness of KMS states of one-dimensional quantum lattice systems,'' Commun. Math. Phys. **44**, 1–7 (1975).
[6] H. Narnhofer and W. Thirring, ''Quantum K-systems,'' Commun. Math. Phys. **125**, 564–577 (1989).
[7] V. Ya. Golodets and S. V. Neshveyev, ''Non-Bernoullian quantum K-systems,'' Commun. Math. Phys. **195**, 213–232 (1998).
[8] A. Connes, H. Narnhofer, and W. Thirring, ''Dynamical entropy of C^*-algebras and von Neumann algebras,'' Commun. Math. Phys. **112**, 691–719 (1987).
[9] F. Hiai and D. Petz, ''Quantum mechanics in AF C^*-systems,'' Rev. Math. Phys. **8**, 819–859 (1996).
[10] S. Popa, ''Classification of subfactors: the reduction to commuting squares,'' Invent. Math. **101**, 19–43 (1990).
[11] M. Pimsner and S. Popa, ''Entropy and index for subfactors,'' Ann. Sci. Ecole Norm. Sup. Ser. 4 **19**, 57–106 (1986).
[12] A. Ocneanu, ''Quantized groups, string algebras and Galois theory for algebras,'' *Operator Algebras and Applications*, London Math. Soc. Lect. Note Ser. 136, edited by D. E. Evans and M. Takesaki, Vol. 2, 1988, pp. 119–172.
[13] R. T. Powers, ''An index theory for semigroups of *-endomorphisms of $B(H)$ and type II_1-factors,'' Can. J. Math. **40**, 86–114 (1988).
[14] V. Ya. Golodets and E. Størmer, ''Entropy of C^*-dynamical systems defined by bit-streams,'' Ergod. Th. & Dynam. Sys. **18**, 1–16 (1998).
[15] R. Bowen, ''Equilibrium states and the ergodic theory of Anosov diffeomorphisms,'' *Lecture Notes in Mathematics 470* (Springer-Verlag, Berlin, 1975).
[16] R. T. Powers and G. L. Price, ''Binary shifts on the hyperfinite II_1-factor,'' Contemp. Math. **145**, 453–464 (1993).
[17] M. Enomoto, M. Nagisa, Y. Watatani, and H. Yoshida, ''Relative commutant algebras of Powers' binary shifts on the hyperfinite II_1-factor,'' Math. Scand. **68**, 115–130 (1991).
[18] V. Jones, ''Index for subfactors,'' Invent. Math. **72**, 1–25 (1983).
[19] M. Choda, ''Entropy for *-endomorphisms and relative entropy for subalgebras,'' J. Operator Theory **25**, 125–140 (1991).
[20] S. V. Neshveyev, ''Quantum Markov K-systems,'' Mat. Fiz., Anal. Geom. **5**, 87–94 (1998) (in Russian).
[21] D. Petz, ''Entropy of Markov states,'' Math. Pure Appl. **14**, 33–42 (1994).

Matrix elements for a generalized spiked harmonic oscillator

Richard L. Hall, Nasser Saad,[a] and Attila B. von Keviczky
Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montréal, Québec H3G 1M8, Canada

(Received 12 June 1998; accepted for publication 8 July 1998)

Closed form expressions for the singular-potential integrals $\langle m|x^{-\alpha}|n\rangle$ are obtained with respect to the Gol'dman and Krivchenkov eigenfunctions for the singular potential Bx^2+A/x^2, $B>0$, $A\geqslant 0$. The formulas obtained are generalizations of those found earlier by use of the odd solutions of the Schrödinger equation with the harmonic oscillator potential [Aguilera-Navarro *et al.*, J. Math. Phys. **31**, 99 (1990)]. © *1998 American Institute of Physics.* [S0022-2488(98)00812-3]

I. INTRODUCTION

In 1990 Aguilera-Navarro *et al.*[1] employed a perturbative scheme to provide a variational analysis for the lowest eigenvalue of the Schrödinger operator,

$$H=-\frac{d^2}{dx^2}+x^2+\frac{\lambda}{x^\alpha},\quad 0\leqslant x<\infty,$$

where α is a positive constant. Writing $H\equiv H_0+\lambda V$ with H_0 standing for the harmonic oscillator Hamiltonian and $V=x^{-\alpha}$, Aguilera-Navarro *et al.*[1] used the basis set

$$\psi_n(x)=A_n e^{-x^2/2}H_{2n+1}(x),\quad A_n^{-2}=4^n(2n+1)\Gamma(1/2),\quad n=0,1,2,..., \tag{1}$$

constructed from the normalized solutions of $H_0\psi=E\psi$ to evaluate the matrix elements of H. They found that

$$H_{m+1,n+1}\equiv(3+4n)\,\delta_{m,n}+\lambda\langle m|x^{-\alpha}|n\rangle,\quad m,n=0,1,2,...,N-1, \tag{2}$$

where

$$\langle m|x^{-\alpha}|n\rangle=(-1)^{n+m}\,\frac{\sqrt{(2n+1)!(2m+1)!}}{2^{n+m}n!m!}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\,\frac{\Gamma(k+(3-\alpha)/2)\Gamma(n-k+\alpha/2)}{\Gamma(k+3/2)\Gamma(-k+\alpha/2)}. \tag{3}$$

Our aim in this article is to extend these results to treat the more general spiked harmonic oscillator Hamiltonian,

$$H=-\frac{d^2}{dx^2}+Bx^2+\frac{A}{x^2}+\frac{\lambda}{x^\alpha},\quad B>0,\quad A\geqslant 0. \tag{4}$$

The particular case of $A=0$ and $B=1$ allows us, of course, to recover the results of Aguilera-Navarro *et al.*[1] as a special case. The existence of all the exact eigenfunctions for a problem with a singular potential is the principal motivation for the present work: we expect that such a set of functions will be more effective for the analysis of other singular problems than are the Hermite functions generated, as they are, by the nonsingular harmonic-oscillator potential x^2.

The article is organized as follows. In Sec. II we provide an orthonormal set of functions that we use to compute the matrix elements of the Hamiltonian (4). In Sec. III we prove that this set of

[a]Present address: Department of Mathematics, Faculty of Science, Notre Dame University, Beirut, Lebanon.

 © 1998 American Institute of Physics

functions is complete in the sense that it is a basis for $L^2(0,\infty)$. We then compute the matrix elements using this basis in Sec. IV. In Sec. V the comparison with the result of Aguilera-Navarro *et al.*[1] are presented for the special case $B=1$ and $A=0$, where we shall point out some errors in their value for the matrix element $\langle \psi_3|x^{-\alpha}|\psi_3\rangle$ as quoted in the appendix of Ref. 1.

II. AN ORTHONORMAL BASIS

Gol'dman and Krivchenkov[2] have provided a clear description of the exact solutions of the following one dimensional Schrödinger equation (in units $\hbar=2m=1$),

$$-\psi''+V_0\left(\frac{a}{x}-\frac{x}{a}\right)^2\psi=E_n\psi, \quad x\in[0,\infty), \tag{5}$$

with ψ satisfying the Dirichlet boundary condition $\psi(0)=0$. They showed that the energy spectrum in terms of the parameters V_0 and a is given by

$$E_n=\frac{4}{a}\sqrt{V_0}\left\{n+\frac{1}{2}+\frac{1}{4}\left(\sqrt{1+4V_0a^2}-2a\sqrt{V_0}\right)\right\}. \tag{6}$$

To simplify notation we introduce the parameters $B=V_0a^{-2}$ and $A=V_0a^2$, and obtain thereby an exact solution to the Schrödinger equation with the singular potential

$$V(x)=Bx^2+\frac{A}{x^2}, \quad B>0, \quad A\geqslant 0, \tag{7}$$

where the energy spectrum is now given in terms of the parameters A and B by

$$E_n=\sqrt{B}(4n+2+\sqrt{1+4A}), \quad n=0,1,2,\ldots. \tag{8}$$

The wave functions have the form

$$\langle x|n\rangle\equiv\psi_n(x)=C_nx^{(1/2)(1+\sqrt{1+4A})}e^{-(1/2)\sqrt{B}x^2}\,{}_1F_1(-n,1+\tfrac{1}{2}\sqrt{1+4A};\sqrt{B}x^2), \tag{9}$$

where $n=0,1,2,\ldots$, and ${}_1F_1$ is the confluent hypergeometric function[3]

$${}_1F_1(a,b;z)=\sum_k\frac{(a)_kz^k}{(b)_kk!}.$$

The Pochhammer symbols $(a)_k$ are defined as

$$(a)_k=a(a+1)(a+2)\cdots(a+k-1)=\frac{\Gamma(a+k)}{\Gamma(a)}, \quad k=1,2,\ldots, \tag{10}$$

where $\Gamma(a)$ is the gamma function. Note that we have corrected the misprint of Ref. 4 for the power of x in the wave functions (9). The constant C_n is determined from the normalization condition

$$\int_0^\infty\psi_n^2(x)dx=1,$$

which requires use of the identity[4]

$$\int_0^\infty e^{-\lambda x}x^{\nu-1}[{}_1F_1(-n,\gamma;kx)]^2dx$$

$$=\frac{n!\Gamma(\nu)}{k^\nu\gamma(\gamma+1)\cdots(\gamma+n-1)}$$

$$\times\left\{1+\sum_{s=0}^{n-1}\frac{n(n-1)\cdots(n-s)(\gamma-\nu-s-1)(\gamma-\nu-s)\cdots(\gamma-\nu+s)}{[(s+1)!]^2\gamma(\gamma+1)\cdots(\gamma+s)}\right\}, \quad (11)$$

thereby yielding

$$C_n^2=\frac{2B^{1/2+1/4\sqrt{1+4A}}\Gamma(n+1+\frac{1}{2}\sqrt{1+4A})}{n![\Gamma(1+\frac{1}{2}\sqrt{1+4A})]^2}. \quad (12)$$

In order to prove that the $\psi_n(x)$, $n=0,1,2,...$, defined by (9) are orthonormal, we have to demonstrate

$$\int_0^\infty \psi_n(x)\psi_m(x)dx=0 \quad (n\neq m). \quad (13)$$

We know, however, that $\psi_n(x)$ and $\psi_m(x)$ satisfy the Schrödinger equations,

$$\psi_n''+\left(E_n-Bx^2-\frac{A}{x^2}\right)\psi_n=0, \quad E_n=\sqrt{B}(4n+2+\sqrt{1+4A}), \quad (14)$$

$$\psi_m''+\left(E_m-Bx^2-\frac{A}{x^2}\right)\psi_m=0, \quad E_m=\sqrt{B}(4m+2+\sqrt{1+4A}). \quad (15)$$

Multiplying (14) by ψ_m and (15) by ψ_n and then subtracting the resulting equations, we obtain

$$\frac{d}{dx}(\psi_m\psi_n'-\psi_n\psi_m')+\sqrt{B}(n-m)\psi_n\psi_m=0.$$

After integrating this equation over $[0,\infty)$ and using $\psi_n(0)=\psi_m(0)=0$ and $\psi_n(\infty)=\psi_m(\infty)=0$, we get

$$(n-m)\int_0^\infty \psi_n(x)\psi_m(x)dx=0,$$

which proves (13). Thus we obtain the following identity:

$$\int_0^\infty e^{-\lambda x^2}x^{2\gamma-1}\,{}_1F_1(-n,\gamma;\lambda x^2){}_1F_1(-m,\gamma;\lambda x^2)dx=\begin{cases}0, & \text{if } n\neq m,\\ \dfrac{1}{2}\dfrac{n!\Gamma(\gamma)}{\lambda^\gamma(\gamma)_n}, & \text{if } n=m,\end{cases}$$

wherein the confluent hypergeometric functions ${}_1F_1$ are defined as follows:[3]

$${}_1F_1(-n,\gamma;r)\equiv-\frac{1}{2\pi i}\frac{\Gamma(n+1)\Gamma(\gamma)}{\Gamma(n+\gamma)}\oint_{C'}e^{tr}(-t)^{-n-1}(1-t)^{\gamma-n-1}dt$$

$$=\frac{\Gamma(\gamma)}{\Gamma(\gamma+n)}r^{1-\gamma}e^rD^n(r^{\gamma+n-1}e^{-r})=\frac{\Gamma(\gamma)}{\Gamma(\gamma+n)}r^{1-\gamma}(D-1)^n(r^{\gamma+n-1}), \quad (16)$$

for any simply closed rectifiable contour C' starting at 1 and enclosing the straight line segment from 0 to 1 in the complex plane, as illustrated in Fig. 1. With $\gamma=1+\frac{1}{2}\sqrt{1+4A}$ and $\lambda=\sqrt{B}$, the set of $L^2(0,\infty)$-functions,

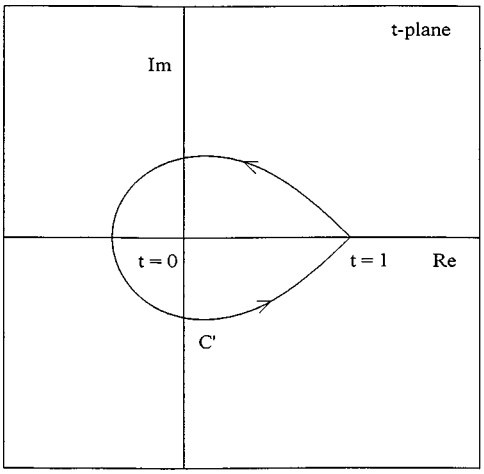

FIG. 1. The contour C' in the t-plane starting at 1 and enclosing the line segment from 0 to 1.

$$\psi_n(x)=C_n x^{\gamma-1/2}e^{-(1/2)\lambda x^2}\,{}_1F_1(-n,\gamma;\lambda x^2),\quad n=0,1,2,..., \tag{17}$$

constitutes an orthonormal system of the Hilbert space $L^2(0,\infty)$.

III. PROOF OF COMPLETENESS

For the orthonormal functions $\{\psi_n\}$ to qualify as a basis for $L^2(0,\infty)$, we must demonstrate the density of the linear manifold generated by these functions in the topology induced by the norm determined by the inner product $\langle\cdot|\cdot\rangle$. This is equivalent to showing that if $\langle\psi_n|f\rangle=0$ for all $n=0,1,2,...$, then $f=0$ a.e. on $(0,\infty)$. To this end we note that out of the fourth expression of (16) follows:

$${}_1F_1(-n,\gamma;\lambda x^2)=\sum_{k=0}^{n}\binom{n}{k}\frac{(-\lambda)^{n-k}\Gamma(\gamma)}{\Gamma(\gamma+n-k)}\,x^{2(n-k)}. \tag{18}$$

Hence the basis representation of the functions (vectors),

$$\{{}_1F_1(-n,\gamma;\lambda x^2),\ {}_1F_1(-(n-1),\gamma;\lambda x^2),\ldots,\ {}_1F_1(-1,\gamma;\lambda x^2),\ {}_1F_1(-0,\gamma;\lambda x^2)\},$$

in terms of the basis

$$\{x^{2n},x^{2(n-1)},\ldots,x^2,1\},$$

is achieved by a lower triangular $(n+1)\times(n+1)$ matrix, whose diagonal entries are $(-\lambda)^{n-k}\Gamma(\gamma)/\Gamma(\gamma+n-k)$ for $k=0,1,2,...,n$—i.e., this matrix is invertible provided $\lambda\neq 0$. Thus each x^{2n} is a unique linear combination of the $n+1$ functions ${}_1F_1(-(n-k),\gamma;\lambda x^2)$ for $k=0,1,2,...,n$, which conclusion carries over to the $2n$-th degree Taylor polynomial,

$$e_n\left(-\frac{\mu x^2}{2}\right)=\sum_{k=0}^{n}\frac{1}{k!}\left(-\frac{\mu x^2}{2}\right)^k,$$

of $e^{-\mu x^2/2}$ about the point 0, where μ is an arbitrary parameter.

Let f be an $L^2(0,\infty)$-function orthogonal to each of the ψ_n, which is equivalent to saying

$$\left\langle e_n\left(-\frac{\mu\cdot^2}{2}\right)\middle|f\right\rangle=\int_0^\infty x^{\gamma-1/2}e_n\left(-\frac{\mu x^2}{2}\right)f(x)\,dx=0, \tag{19}$$

for all $n=0,1,2,...$. Here we note that

$$x^{\gamma-1/2}e^{-\lambda x^2/4}f(x),$$

is an $L^1(0,\infty)$-function whose absolute value majorizes $x^{\gamma-1/2}e^{-\lambda x^2/4}e^{|\mu|x^2/4}f(x)$ for $|\mu|\leqslant\lambda/4$ and consequently also $x^{\gamma-1/2}e^{-\lambda x^2/4}e_n(-\mu x^2/4)f(x)$. Because $x^{\gamma-1/2}e^{-\lambda x^2/4}e_n(-\mu x^2/4)f(x)$ converges to $x^{\gamma-1/2}e^{-\lambda x^2/2}f(x)$ a.e. on $(0,\infty)$ as $n\to\infty$, we conclude by means of the Lebesgue dominated convergence theorem[5] that we may replace $e_n(-\mu x^2/2)$ by $e^{-\mu x^2/2}$ in Eq. (19) for all complex numbers μ such that $|\mu|\leqslant\lambda/4$, which, after setting $x=\sqrt{2t}$, yields the Laplace-Transform expression

$$\mathscr{L}\{F\}(z)=\int_0^\infty e^{-zt}(\sqrt{2t})^{\gamma-3/2}f(\sqrt{2t})\,dt=0, \quad |z-\lambda|\leqslant\frac{\lambda}{4}. \tag{20}$$

However, the Laplace transform of the measurable Laplace-transformable function $F(t)=e^{-zt}(\sqrt{2t})^{\gamma-3/2}f(\sqrt{2t})$ defines a holomorphic function of variable z in the right half plane $\Re(z)>0$ vanishing in the disc $|z-\lambda|\leqslant\lambda/4$. By uniqueness of the analytic function[6] the Laplace transform of the function must vanish in the right half plane, specifically $\mathscr{L}\{F\}(s)=0$ for all s on the interval $(0,\infty)$. Further, the Laplace transform determines $F(t)$ uniquely[7] a.e. in t on $(0,\infty)$, hence $F(t)=0$ a.e. in t or f is the zero $L^2(0,\infty)$-function. Consequently, $\{\psi_n:\ n=0,1,2,...\}$ is an orthonormal basis of $L^2(0,\infty)$.

IV. THE MATRIX ELEMENTS $\langle m|x^{-\alpha}|n\rangle$

Let us now split the Hamiltonian (5) into an H_0 part,

$$H_0=-\frac{d^2}{dx^2}+Bx^2+\frac{A}{x^2}, \quad x\geqslant 0, \tag{21}$$

and a perturbation,

$$H_I=\lambda/x^\alpha.$$

The eigenstates of H_0 are now given by (9) and their unperturbed energy is given by (8). All we need to do is to evaluate the matrix elements $\langle m|x^{-\alpha}|n\rangle$ using the basis (9), namely,

$$\langle m|x^{-\alpha}|n\rangle=C_nC_m\int_0^\infty e^{-\sqrt{B}x^2}x^{-\alpha+1+\sqrt{1+4A}}\ {}_1F_1\left(-n,1+\frac{1}{2}\sqrt{1+4A};\sqrt{B}x^2\right)$$
$$\times\ {}_1F_1\left(-m,1+\frac{1}{2}\sqrt{1+4A};\sqrt{B}x^2\right)dx, \quad \alpha<2+\sqrt{1+4A}\,. \tag{22}$$

This is equivalent to

$$\langle m|x^{-\alpha}|n\rangle=\frac{C_nC_m}{2}\,B^{-(1/2)(-\alpha+2+\sqrt{1+4A})}\times I, \tag{23}$$

where

$$I=\int_0^\infty e^{-r}r^{\gamma-s}\ {}_1F_1(-n,\gamma;r)\ {}_1F_1(-m,\gamma;r)\,dr, \tag{24}$$

with $r=\sqrt{B}x^2$, $\gamma=1+\frac{1}{2}\sqrt{1+4A}$, and $s=1+\alpha/2$.

From the Fubini–Tonneli theorem[5] combined with the Leibniz formula for differentiating the product of two functions, as well as exponential shift from the third expression to the fourth in Eq. (16), with n replaced by m, we find that I is given by the expression

$$I=(-1)^n\frac{n![\Gamma(\gamma)]^2}{\Gamma(n+\gamma)\Gamma(m+\gamma)}(2\pi i)^{-1}\oint_{C'}t^{-n-1}(1-t)^{\gamma-n-1}\int_0^\infty e^{-(1-t)r}r^{1-s}$$
$$\times\left[\sum_{k=0}^{m}(-1)^k\binom{m}{k}(\gamma+m-1)(\gamma+m-2)\cdots(\gamma+k)r^{\gamma+m-1-(m-k)}\right]dr\,dt.$$

Further, owing to the fact that the simply closed rectifiable contour C' lies to the left of the complex number 1, Fig. 1, we have

$$\int_0^\infty e^{-(1-t)r}r^{\gamma-s+k}\,dr=\Gamma(\gamma-s+k+1)(1-t)^{-\gamma+s-k-1},$$

and our expression for I thereby reduces to

$$I=(-1)^n\frac{n![\Gamma(\gamma)]^2}{\Gamma(n+\gamma)\Gamma(m+\gamma)}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\frac{\Gamma(m+\gamma)\Gamma(\gamma-s+k+1)}{\Gamma(k+\gamma)}(2\pi i)^{-1}$$
$$\times\oint_{C'}t^{-n-1}(1-t)^{s+n-k-2}\,dt. \tag{25}$$

Since the contour C' has 0 in its inside, Fig. 1, and the integrand has a weak singularity at 1 [in consequence of $\Re(s+n-k-2)>-1$], the Cauchy integral formula lets us write the contour integral multiplied by $(2\pi i)^{-1}$ as the n-th derivatives of $(1-t)^{s+n-k-2}$ evaluated at $t=0$. Utilizing thereafter the Pochhammer symbol in the Gamma function format, Eq. (10), we arrive at

$$I=\frac{[\Gamma(\gamma)]^2}{\Gamma(n+\gamma)}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\frac{\Gamma(\gamma-s+k+1)\Gamma(s+n-k-1)}{\Gamma(\gamma+k)\Gamma(s-k-1)}. \tag{26}$$

Therefore, the matrix elements are given by

$$\langle m|x^{-\alpha}|n\rangle=(-1)^{n+m}\frac{C_nC_m}{2}B^{-(1/4)(-\alpha+2+\sqrt{1+4A})}\frac{[\Gamma(1+\frac{1}{2}\sqrt{1+4A})]^2}{\Gamma(n+1+\frac{1}{2}\sqrt{1+4A})}$$
$$\times\sum_{k=0}^{m}(-1)^k\binom{m}{k}\frac{\Gamma\left(k+1+\frac{1}{2}\sqrt{1+4A}-\frac{\alpha}{2}\right)\Gamma(\alpha/2+n-k)}{\Gamma(k+1+\frac{1}{2}\sqrt{1+4A})\Gamma(\alpha/2-k)},\alpha<2+\sqrt{1+4A}, \tag{27}$$

with normalization coefficients C_n given in Eq. (12). In case $\alpha/2-k$ is a negative integer; then[3] $1/\Gamma(\alpha/2-k)=0$ for such k and the terms involving these k's shall not appear in the summation of Eq. (27). Further, by expressing the confluent hypergeometric functions ${}_1F_1(-n,\gamma;r)$ and ${}_1F_1(-m,\gamma;r)$ by means of the fourth formula in Eq. (19) and substituting these into Eq. (18), we immediately see that the sum appearing in Eq. (27) is a polynomial of degree $m+n$ in α.

With Eq. (27) we have therefore computed the matrix elements of the operator $x^{-\alpha}$ in the complete basis given by the Gol'dman and Krivchenkov eigenfunctions (9). Concomitant to our result, the matrix elements,

$$\langle 0|x^{-\alpha}|n\rangle=(-1)^nB^{\alpha/4}\sqrt{\frac{\Gamma(1+\frac{1}{2}\sqrt{1+4A})}{n!\Gamma(n+1+\frac{1}{2}\sqrt{1+4A})}}\frac{\Gamma(1+\frac{1}{2}\sqrt{1+4A}-\alpha/2)\Gamma(\alpha/2+n)}{\Gamma(1+\frac{1}{2}\sqrt{1+4A})\Gamma(\alpha/2)}, \tag{28}$$

are of special interest.

V. EXPLICIT FORMS OF THE MATRIX ELEMENTS

In terms of the parameter $\gamma=1+\frac{1}{2}\sqrt{1+4A}$, the explicit forms of the first ten matrix elements of $x^{-\alpha}$ with $\alpha<2\gamma$ are

$$x_{00}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{\Gamma(\gamma)},$$

$$x_{01}^{-\alpha}=-\frac{B^{\alpha/4}}{2}\frac{\Gamma(-\alpha/2+\gamma)}{\sqrt{\gamma}\Gamma(\gamma)}\,\alpha,$$

$$x_{02}^{-\alpha}=\frac{B^{\alpha/4}}{2^2}\frac{\Gamma(-\alpha/2+\gamma)}{\sqrt{2!\,\gamma(\gamma+1)}\Gamma(\gamma)}\,\alpha(\alpha+2),$$

$$x_{03}^{-\alpha}=-\frac{B^{\alpha/4}}{2^3}\frac{\Gamma(-\alpha/2+\gamma)}{\sqrt{3!\,\gamma(\gamma+1)(\gamma+2)}\Gamma(\gamma)}\,\alpha(\alpha^2+6\alpha+8),$$

$$x_{11}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{4\Gamma(\gamma+1)}\,(\alpha^2-2\alpha+4\gamma),$$

$$x_{12}^{-\alpha}=-B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{8\sqrt{2}\sqrt{\gamma+1}\Gamma(\gamma+1)}\,\alpha(\alpha^2-2\alpha+8\gamma),$$

$$x_{13}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{16\sqrt{3!(\gamma+2)(\gamma+1)}\Gamma(\gamma+1)}\,\alpha(\alpha^3+(12\gamma-4)\alpha+24\gamma),$$

$$x_{22}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{32\Gamma(\gamma+2)}\,(\alpha^4-4\alpha^3-16(1+2\gamma)\alpha+4(3+4\gamma)\alpha^2+32\gamma(1+\gamma)),$$

$$x_{23}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{32\sqrt{3!2!(\gamma+2)}\Gamma(\gamma+2)}\,\alpha(\alpha^4-4\alpha^3+4(5+6\gamma)\alpha^2-16(2+3\gamma)\alpha+96\gamma(1+\gamma)),$$

$$x_{33}^{-\alpha}=B^{\alpha/4}\frac{\Gamma(-\alpha/2+\gamma)}{384\Gamma(\gamma+3)}\,(\alpha^6-8\alpha^5+(72+36\gamma)\alpha^4-(208+144\gamma)\alpha^3$$
$$+(272+720\gamma+288\gamma^2)\alpha^2-(192+1152\gamma+576\gamma^2)\alpha+384\gamma(1+\gamma)(2+\gamma)).$$

These matrix elements can be compared for the special case $(A,B)=(0,1)$ with the matrix elements computed by the simple harmonic oscillator representation supplemented by the Dirichlet boundary condition.[1] We have found an error in the value of the matrix element $x_{33}^{-\alpha}$ as given by Ref. 1; this error is confirmed by a re-computation according to the matrix element expression given by them. Indeed, the matrix element $x_{33}^{-\alpha}$ should read as

$$x_{33}^{-\alpha}=\frac{\Gamma((3-\alpha)/2)}{7!\,\Gamma(3/2)}\,(\alpha^6-6\alpha^5+106\alpha^4-384\alpha^3+2080\alpha^2-3408\alpha+5040),$$

instead of

$$x_{33}^{-\alpha}=\frac{\Gamma((3-\alpha)/2)}{7!\,\Gamma(3/2)}\,(\alpha^6-6\alpha^5+106\alpha^4-454\alpha^3+1660\alpha^2-3968\alpha+5040),$$

as quoted by Aguilera-Navarro *et al.*[1]

In this work we have proved that the Gol'dman–Krivchenkov wave functions constitute an orthonormal basis for the Hilbert space $L^2(0,\infty)$. Using this orthonormal basis, we are able to construct the matrix elements of $x^{-\alpha}$: the general result [Eq. (27)] is convenient for use in any practical application which involves such singular potential terms. It is also interesting that, with minor changes,[8] essentially involving only the value of the coefficient A, the same formulas apply

immediately to the corresponding problems with nonzero angular momentum and in arbitrary spatial dimension $N \geqslant 2$. A detailed variational analysis of the spiked harmonic Hamiltonian operator based on these matrix elements is presently in progress.

ACKNOWLEDGMENT

Partial financial support of this work under Grant No. GP3438 from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

[1] V. C. Aguilera-Navarro, G. A. Estévez, and R. Guardiola, J. Math. Phys. **31**, 99–104 (1990).
[2] I. I. Gol'dman and D. V. Krivchenkov, *Problems in Quantum Mechanics* (Pergamon, London, 1961).
[3] L. J. Slater, *Confluent Hypergeometric Functions* (Cambridge University Press, Cambridge, 1960); F. W. Schäfke, *Einführung in die Theorie der Speziellen Funktionen der Mathematischen Physik* (Springer-Verlag, Berlin, 1963), Satz 1, p. 162.
[4] L. D. Landau and M. E. Lifshitz, *Quantum Mechanics: Non-relativistic Theory* (Pergamon, Oxford, 1977).
[5] W. Rudin, *Real and Complex Analysis*, 3rd ed. (McGraw-Hill, New York, 1987), Theorem 1.3.4, p. 26; the Fubini–Tonelli theorem is discussed on pp. 164–166.
[6] D. V. Widder, *The Laplace Transform* (Princeton University Press, Princeton, 1972), Corollary 9.3b, p. 80; G. Doetsch, *Handbuch der Laplace Transformation* (Band I) (Basel, Birkhäuser, 1971), Satz 4, p. 74.
[7] H. Behnke and F. Sommer, *Theorie der Analytischen Funktionen einer Komplexen Veränderlichen* (Springer-Verlag, Berlin, 1976), Satz 28, p. 138.
[8] R. L. Hall and N. Saad, J. Chem. Phys. **109**, 2983–2986 (1998).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

On infravacua and the localization of sectors

Walter Kunhardt[a)]
Institut für Theoretische Physik der Universität Göttingen, Bunsenstraße 9, 37073 Göttingen, Germany

(Received 24 June 1998; accepted for publication 27 August 1998)

A certain class of superselection sectors of the free massless scalar field in three space dimensions is considered. It is shown that these sectors, which cannot be localized with respect to the vacuum, acquire a much better localization, namely in spacelike cones, when viewed in front of suitable ''infravacuum'' backgrounds. These background states coincide, essentially, with a class of states introduced by Kraus, Polley, and Reents as models for clouds of infrared radiation. © *1998 American Institute of Physics.* [S0022-2488(98)02412-8]

I. INTRODUCTION

In the analysis of superselection sectors, the localizability properties of charges are crucial for defining notions as charge composition and conjugation, statistics or a (global) gauge group. In a classical work,[1] Doplicher, Haag, and Roberts carried through such a program for charges which fulfill what is now called the DHR criterion, i.e., which are compactly localized. Now this criterion is very restrictive, and Buchholz and Fredenhagen[2] established that sectors of theories in three space dimensions without massless particles in general only comply with the so-called BF criterion, i.e., they are localized in spacelike cones. Still, these authors could extend the analysis of Ref. 1 to charges with such a weaker localization behavior.

The situation is more difficult for theories with massless particles. Typically, these theories possess sectors whose localization is too poor for the DHR framework to be applicable. Motivated by what is expected to happen in quantum electrodynamics (QED), it has been proposed by Buchholz in Ref. 3 to improve the localization by viewing the charges in front of some suitable background field instead of the vacuum. In QED, such background fields should correspond to clouds of infrared radiation. An appropriate mathematical description of such infrared clouds has been introduced by Kraus, Polley, and Reents (KPR).[4]

Here, we want to verify this mechanism in a simpler model, namely in the theory of the free massless scalar field in three space dimensions.[5] More precisely, we will consider a certain class of (non-Lorentz invariant) sectors described by automorphisms of the observable algebra and analyze their localization properties in terms of the BF criterion. In particular, we will show that the sectors under consideration do satisfy this criterion with respect to a KPR-like background but do not satisfy it with respect to the vacuum. (Calling the background fields ''KPR-like'' should indicate that they are very similar, yet not identical, to those of Ref. 4.)

As to the consistency of such an approach, it should be kept in mind that, in a theory whose charges are compactly localized, the superselection structure can be described without any difference with respect to the vacuum as well as with respect to so-called infravacua, the latter being generalizations of the KPR-like background states considered here. As has been shown in Ref. 6, when viewed in front of such an infravacuum, the charges remain compactly localized and have the same fusion structure and statistics as in front of the vacuum. Moreover, positivity of the energy in a sector does not depend on the background chosen, nor do the masses of massive particles possibly contained in such a theory.

The above-mentioned class of sectors of the free massless field has been studied recently by Buchholz *et al.*[5] with the purpose of modeling charges of electromagnetic type. In the following, we will stick very closely to the notations introduced there, but we should emphasize that our point of view is slightly different from that adopted in Ref. 5: Buchholz *et al.* achieved a better

[a)]Electronic mail: kunhardt@theorie.physik.uni-goettingen.de

 © 1998 American Institute of Physics

localization of the sectors by restricting them to a (non-Lorentz invariant) subnet $\mathcal{A}_0 \subset \mathcal{A}$ of the observable net. The sectors then even became localized in the DHR sense, which permitted them to carry through a DHR-like analysis, even though the net $\mathcal{A}_0$ does not fulfill Haag duality. Here, in contrast, the subnet $\mathcal{A}_0$ will play no role, and the localization obtained will be in a weaker sense.

We end this Introduction by recalling the definition of the model under consideration. The observable algebra of the free massless scalar field is defined in its vacuum representation. More precisely, let $\mathcal{K} := L^2(\mathbf{R}^3, d^3k)$ be the Hilbert space of momentum space wave functions, $\omega(\vec{k}) := |\vec{k}|$ the one-particle energy and $U(t,\vec{x}) = e^{i(\omega(\hat{k})t - \vec{k}\vec{x})}$ the usual representation of the space–time translations. The vacuum Hilbert space of our model will be the bosonic Fock space $\mathcal{H}$ over $\mathcal{K}$; the induced unitary representation of the space–time translations will still be denoted by $U(t,\vec{x})$ without any risk of confusion. For any $v \in \mathcal{K}$, $W(v) \in \mathcal{B}(\mathcal{H})$ will denote the corresponding Weyl operator. The normalization is chosen such that the Weyl relations read $W(u)W(v) = e^{-(i/2)\mathrm{Im}\langle u,v\rangle} W(u+v)$. For any real linear subspace $\mathcal{L} \subset \mathcal{K}$, $\mathcal{W}(\mathcal{L})$ denotes the C*-subalgebra of $\mathcal{B}(\mathcal{H})$ generated by the operators $W(f)$, $f \in \mathcal{L}$. The net of observables now is given as

$$\mathcal{O} \mapsto \mathcal{A}(\mathcal{O}) := \mathcal{W}(\mathcal{L}(\mathcal{O}))'',$$

where $\mathcal{O} \mapsto \mathcal{L}(\mathcal{O})$ is the isotonous, local and covariant net of symplectic subspaces in $\mathcal{K}$ (indexed by the set of open double cones in Minkowski space) defined as follows: If $\mathcal{O} := (\{0\} \times O)''$ is the causal completion of an open ball $O \subset \mathbf{R}^3$ at time $t = 0$, one has

$$\mathcal{L}(\mathcal{O}) := \omega^{-1/2} \widehat{\mathcal{D}_{\mathbf{R}}(O)} + i\omega^{+1/2} \widehat{\mathcal{D}_{\mathbf{R}}(O)},$$

where $\mathcal{D}_{\mathbf{R}}(O)$ is the set of all real-valued smooth functions with support in O and $\hat{}$ denotes the Fourier transform. For other double cones $\mathcal{O}$, the space $\mathcal{L}(\mathcal{O})$ is defined by translation covariance and additivity. The symplectic form σ on $\mathcal{L} := \cup_{\mathcal{O}} \mathcal{L}(\mathcal{O}) \subset \mathcal{K}$ reads

$$\sigma(f_1, f_2) := -\mathrm{Im}\langle f_1, f_2\rangle,$$

and locality for the net $\mathcal{L}(\cdot)$ just means $\sigma(\mathcal{L}(\mathcal{O}_1), \mathcal{L}(\mathcal{O}_2)) = 0$ whenever $\mathcal{O}_1$ and $\mathcal{O}_2$ are spacelike to each other. As usual, we also associate symplectic subspaces of $\mathcal{L}$ (resp. C*-subalgebras of $\mathcal{B}(\mathcal{H})$) to *un*bounded regions in $\mathbf{R}^{1+3}$ by additivity (resp. additivity and norm closure) and simply denote by $\mathcal{A}$ the quasilocal algebra $\mathcal{A}(\mathbf{R}^{1+3})$.

The charges under consideration are given in terms of net automorphisms $\gamma \in \mathrm{Aut}\,\mathcal{A}$ which are labeled uniquely by elements of the (additive) Abelian group

$$\mathcal{L}_\Gamma := \omega^{-1/2} \widehat{\mathcal{D}_{\mathbf{R}}(\mathbf{R}^3)} + i\omega^{-3/2} \widehat{\mathcal{D}_{\mathbf{R}}(\mathbf{R}^3)}.$$

Any element $\gamma \in \mathcal{L}_\Gamma$ gives rise to a linear form $l_\gamma : \mathcal{L} \to \mathbf{C}$,

$$l_\gamma(f) := -\mathrm{Im} \int d^3k\, \overline{\gamma(\vec{k})} f(\vec{k}),$$

and hence to an automorphism, again denoted by γ, of $\mathcal{A}$ by

$$\gamma(W(f)) := e^{i l_\gamma(f)} W(f).$$

As explained in Ref. 5, γ is indeed a well-defined automorphism of the quasi-local algebra $\mathcal{A}$ since, by Huygens' principle, it turns out to be locally normal; as a consequence, it can be extended by weak continuity from the local Weyl algebras $\mathcal{W}(\mathcal{L}(\mathcal{O}))$ to the local von Neumann algebras $\mathcal{A}(\mathcal{O})$.

There will be no risk of confusion in viewing the real vector space $\mathcal{L}_\Gamma$ as an Abelian subgroup of $\mathrm{Aut}\,\mathcal{A}$. In particular, a sum $\gamma_1 + \gamma_2$ in $\mathcal{L}_\Gamma$ corresponds to the composition $\gamma_1 \circ \gamma_2$ in $\mathrm{Aut}\,\mathcal{A}$. Moreover, γ_1 and γ_2 define the same sector of $\mathcal{A}$, i.e., they are unitarily equivalent in $\mathcal{B}(\mathcal{H})$, iff $\gamma_1 - \gamma_2 \in \mathcal{L}_\Gamma \cap \mathcal{K}$. In this case, the Weyl operator $W(\gamma_1 - \gamma_2)$ is well defined and implements the unitary equivalence $\gamma_1 \cong \gamma_2$ on $\mathcal{A}$.

Any $\gamma \in \mathscr{S}_\Gamma$ can be written uniquely in the form $\gamma = \omega^{-1/2}\hat{\sigma} + i\omega^{-3/2}\hat{\rho}$ with functions $\sigma, \rho \in \mathscr{D}_{\mathbf{R}}(\mathbf{R}^3)$. Since $\hat{\sigma}$ and $\hat{\rho}$ are analytic, it is obvious that γ is square integrable, i.e., $\gamma \in \mathscr{K}$, iff $\hat{\rho}(0) = 0$. As a consequence, the sectors considered are labeled by a single real parameter

$$q_\gamma := \hat{\rho}(0) = \int d^3x \rho(\vec{x})$$

which is interpreted as the charge of the sector $[\gamma]$. In particular, this shows that the sectors are transportable; as a matter of fact, they even have positive energy.[5]

II. BAD LOCALIZATION OF THE SECTORS IN FRONT OF THE VACUUM

It has been shown in Ref. 5 that the automorphisms $\gamma \in \mathscr{S}_\Gamma$ do not satisfy the DHR localization criterion. Here, we want to strengthen this result and show with closely related methods that they do not even satisfy the BF criterion, that is, that they are not localizable in spacelike cones. To this end, it is sufficient to prove the following

Proposition 2.1: Let $C \subset \mathbf{R}^3$ be an open convex cone having 0 as its apex and denote with $\mathscr{C} := (\{0\} \times C)''$ its causal completion. Then, for any $\gamma \in \mathscr{S}_\Gamma$,

$$\gamma|_{\mathscr{A}(\mathscr{C})} \cong \mathrm{id}|_{\mathscr{A}(\mathscr{C})} \quad \text{iff} \quad q_\gamma = 0.$$

The ''if'' part of this proposition is trivial, and before proving the ''only if'' part, we recall some facts about the dilation covariance of the model. The dilation group $\mathbf{R}_{>0}$ acts unitarily on $\mathscr{K}$ and leaves the space $\mathscr{S}$ invariant. More precisely, $f \in \mathscr{S}(\mathscr{O})$ is mapped onto $f_\lambda \in \mathscr{S}(\lambda\mathscr{O})$, where $f_\lambda(\vec{k}) := \lambda^{3/2} f(\lambda\vec{k})$. Writing $f = \omega^{-1/2}\hat{h} + i\omega^{+1/2}\hat{g}$, it is verified by a straightforward computation that this entails for the linear form l_γ, $\gamma \in \mathscr{S}_\Gamma$:

$$l_\gamma(f_\lambda) = \int \frac{d^3k}{\omega^2} \overline{\hat{\rho}(\vec{k}/\lambda)}\hat{h}(\vec{k}) - \frac{1}{\lambda}\int d^3k \overline{\hat{\sigma}(\vec{k}/\lambda)}\hat{g}(\vec{k}).$$

In the limit $\lambda \to \infty$, $\vec{k} \mapsto \overline{\hat{\rho}(\vec{k}/\lambda)}\hat{h}(\vec{k})$ converges to $\hat{\rho}(0)\hat{h}$ in the space of test functions, and since $2\pi^2/r$ is the Fourier transform of $1/\omega^2$ in the sense of distributions, one obtains

$$\lim_{\lambda\to\infty} l_\gamma(f_\lambda) = q_\gamma \kappa_f \quad \text{with} \quad \kappa_f := 2\pi^2 \int \frac{d^3x}{|\vec{x}|} h(\vec{x}).$$

This allows us to prove the following

Lemma 2.2: Let $f \in \mathscr{S}(\mathscr{O}')$, where $\mathscr{O} \subset \mathbf{R}^{1+3}$ is a neighborhood of 0. For any $\gamma \in \mathscr{S}_\Gamma$, one then has

$$\text{w-}\lim_{\lambda\to\infty} \gamma(W(f_\lambda)) = e^{iq_\gamma\kappa_f} e^{-(1/4)\|f\|^2}\mathbf{1}.$$

Proof: Since the dilations act geometrically, it follows by locality from the special form of the localization region of f that $\lim_{\lambda\to\infty} \sigma(f_\lambda, f') = 0$ for any $f' \in \mathscr{S}$. Hence, $(W(f_\lambda))_{\lambda>0}$ is a central sequence of unitaries in $\mathscr{W}(\mathscr{S})$ whose set of weak limits is, by the irreducibility of the vacuum representation, a (nonempty) subset of $\mathbf{C1}$. On the other hand, unitarity of the dilations permits us to evaluate this limit in the vacuum state: $\omega_0(W(f_\lambda)) \xrightarrow{\lambda\to\infty} e^{-(1/4)\|f\|^2}$. But this means that $W(f_\lambda)$ has $e^{-(1/4)\|f\|^2}\mathbf{1}$ as its *unique* weak limit for $\lambda \to \infty$, establishing thus the assertion for $\gamma = 0$. For arbitrary $\gamma \in \mathscr{S}_\Gamma$, it now follows easily in view of the discussion in the preceding paragraph. ∎

Physically, the sequence $(W(f_\lambda))_{\lambda\to\infty}$ is interpreted as a measurement of the asymptotic behavior (in the spatial directions determined by the smearing function h) of the ''Coulomb potential'' of the ''charge density'' ρ. In QED, one expects that operators measuring the

asymptotic electric flux distribution play a similar role.[3] In the present case, the leading $1/r$ behavior of the Coulomb potential is isotropic in all sectors $[\gamma]$. This fact, reflected by the factorizing of $\lim l_\gamma(f_\lambda)$ as seen above, is relevant in the following.

Proof of Proposition 2.1: Let $\gamma \in \mathscr{G}_\Gamma$ with $q_\gamma \neq 0$. Choose a nonvanishing, non-negative test function $h \in \mathscr{D}_{\mathbf{R}}(C)$. Letting $f := \omega^{-1/2}\hat{h}$, this implies $\kappa_f \neq 0$ and $f \in \mathscr{S}(\mathscr{O} \cap \mathscr{O}')$ for some neighborhood $\mathscr{O} \subset \mathbf{R}^{1+3}$ of 0. Since $e^{iq_\gamma \kappa_f} \neq 1$ can always be achieved by a mere rescaling of h, Lemma 2.2 shows that the weak limits (as $\lambda \to \infty$) of $W(f_\lambda)$ and $\gamma(W(f_\lambda))$ are different scalar multiples of the unit operator. But since $W(f_\lambda) \in \mathscr{A}(\mathscr{O})$ for all $\lambda > 0$, this implies $\gamma|_{\mathscr{A}(\mathscr{O})} \neq \mathrm{id}|_{\mathscr{A}(\mathscr{O})}$. ■

III. INFRAVACUUM BACKGROUND STATES

In this section we introduce a class of background states in front of which the automorphisms γ will be shown (in Sec. IV) to have better localization properties. Apart from two modifications necessitated by the present model, these background states are of the same type as those introduced by Kraus, Polley, and Reents[4] as a model for infrared clouds in QED or, more generally, in any theory containing massless particles.

A. Preliminaries on quasifree states

First, we recall that a *quasifree state* on $\mathscr{A}$ is a locally normal state ω_T which is, on the Weyl operators $W(f) \in \mathscr{A}$, $f \in \mathscr{S}$, of the form

$$\omega_T(W(f)) = e^{-(1/4)\|Tf\|^2}.$$

Here, $T: D_T \to \mathscr{K}$ is a real linear, symplectic (i.e., fulfilling $\mathrm{Im}\langle Tv, Tw\rangle = \mathrm{Im}\langle v, w\rangle, v, w \in D_T$) operator defined on a dense, real linear subspace D_T which contains $\mathscr{S}$. In the case at hand, we will have in addition $\overline{T\mathscr{S}} = \mathscr{K}$, which entails that ω_T is a pure state. Its Gel'fand–Naimark–Segal (GNS) representation π_T acts irreducibly on the vacuum Hilbert space $\mathscr{H}$ as $\pi_T(W(f)) = W(Tf)$, $f \in \mathscr{S}$.

Next, we describe the *real linear* operator T in terms of a pair of *complex linear* operators T_1, T_2 defined on complex linear subspaces D_{T_j}, $j = 1,2$.

Lemma 3.1: Let $\Gamma: \mathscr{K} \to \mathscr{K}$ *be an antiunitary involution. Then, the formulas*

$$T := T_2 \frac{1+\Gamma}{2} + T_1 \frac{1-\Gamma}{2},$$

$$D_T := \left\{ v \in \mathscr{K} \,\middle|\, \frac{1+\Gamma}{2} v \in D_{T_2},\ \frac{1-\Gamma}{2} v \in D_{T_1} \right\}$$

establish a bijection between

(i) *densely defined,* Γ*-invariant* **R***-linear operators* $T: D_T \to \mathscr{K}$ *and*
(ii) *densely defined,* Γ*-invariant* **C***-linear operators* $T_j: D_{T_j} \to \mathscr{K}$, $j = 1,2$

(*where* $T: D_T \to \mathscr{K}$ *being* Γ*-invariant means* $\Gamma D_T = D_T$ *and* $[\Gamma, T] = 0$ *on* D_T). *Moreover,* T *is symplectic iff* $\langle T_1 u_1, T_2 u_2\rangle = \langle u_1, u_2\rangle$ *for all* $u_j \in D_{T_j}$.

Since all assertions can be checked by simple calculations, we omit the formal proof of this Lemma and merely point out that the converse formulas expressing T_1 and T_2 in terms of T read

$$D_{T_2} = \left\{ v \in \mathscr{K} \,\middle|\, \frac{1+\Gamma}{2} \mathbf{C} v \subset D_T \right\}, \quad T_2 = T \frac{1+\Gamma}{2} - iT \frac{1+\Gamma}{2}\, i,$$

$$D_{T_1} = \left\{ v \in \mathscr{K} \,\middle|\, \frac{1-\Gamma}{2} \mathbf{C} v \subset D_T \right\}, \quad T_1 = T \frac{1-\Gamma}{2} + iT \frac{1-\Gamma}{2}\, i.$$

Remark: The involution Γ induces the notion of real and imaginary parts of vectors $v \in \mathscr{K}$: $\mathrm{Re}\, v = [(1+\Gamma)/2]v$, $\mathrm{Im}\, v = [(1-\Gamma)/2i]v$. Then, T_2 acts on the real and T_1 on the imaginary parts:

$$\mathrm{Re}\ Tv = T_2\ \mathrm{Re}\ v, \quad \mathrm{Im}\ Tv = T_1\ \mathrm{Im}\ v, \quad v \in D_T.$$

From now on, we will fix Γ to be pointwise complex conjugation in position space. In terms of momentum space wave functions $v \in \mathscr{K}$, this means

$$(\Gamma v)(\vec{k}) := \overline{v(-\vec{k})}.$$

(For the sake of completeness, we point out that Kraus *et al.* used pointwise conjugation in *momentum space* for defining their background states in Ref. 4. In their case as well as in ours, the choice of the involution Γ is dictated by the set of sectors under consideration.)

B. Quasifree states with positive energy

Before describing in detail the operators T_1 and T_2, we introduce some notation: for any $\epsilon > 0$, let $P_\epsilon : \mathscr{K} \to \mathscr{K}$ be the projector onto the subspace $P_\epsilon \mathscr{K} = \{v \in \mathscr{K} \mid v(\vec{k}) = 0 \text{ if } |\vec{k}| < \epsilon\}$ and denote by

$$D_0 := \bigcup_{\epsilon > 0} P_\epsilon \mathscr{K}$$

the dense subspace of functions vanishing in some neighborhood of $\vec{k} = 0$. Note that $[P_\epsilon, \Gamma] = 0$ and $\Gamma D_0 = D_0$. The subspace D_0 will serve as a provisional domain for T_1 and T_2.

Now we follow Ref. 4 and choose

(i) a sequence $(\epsilon_i)_{i \in \mathbf{N}}$ in $\mathbf{R}_{>0}$ satisfying $\epsilon_{i+1} < \epsilon_i$ and $\epsilon_i \xrightarrow{i \to \infty} 0$. This sequence induces a decomposition of momentum space into concentric spherical shells. The projections onto the associated spectral subspaces of $\mathscr{K}$ will be denoted by $P_i := P_{\epsilon_{i+1}} - P_{\epsilon_i}$. For notational convenience, we also put $P_0 := P_{\epsilon_1}$.

(ii) a sequence $(Q_i)_{i \in \mathbf{N}}$ of orthogonal projections in $\mathscr{K}$ with finite rank $\mathrm{rk} Q_i$ satisfying $Q_i \Gamma = \Gamma Q_i$, $Q_i P_i = Q_i$.

(iii) a sequence $(b_i)_{i \in \mathbf{N}}$ in $]0,1[$ satisfying $b_i \xrightarrow{i \to \infty} 0$ and $\Sigma_i (\epsilon_i / b_i^2) \mathrm{rk} Q_i < \infty$. If, e.g., the ϵ_i decrease exponentially and $\mathrm{rk} Q_i$ is polynomially bounded, this can be satisfied by $b_i \propto i^{-\alpha}$, $\alpha > 0$.

With these data, define $\mathbf{C}$-linear operators T_1 and T_2 on the subspace D_0 by

$$T_1 := 1 + \text{s-lim}_{n \to \infty} \sum_{i=1}^{n} (b_i - 1) Q_i, \quad T_2 := 1 + \text{s-lim}_{n \to \infty} \sum_{i=1}^{n} \left(\frac{1}{b_i} - 1 \right) Q_i.$$

Since, on every $v \in D_0$, the number of terms which contribute on the right-hand side is finite, these operators are well defined and map D_0 into itself. Moreover, the relations

$$T_1 P_i = ((1 - Q_i) + b_i Q_i) P_i, \quad T_2 P_i = ((1 - Q_i) + (1/b_i) Q_i) P_i$$

show that the subspace $P_i \mathscr{K}$ decomposes into a subspace $(1 - Q_i) P_i \mathscr{K}$ where both T_1 and T_2 act trivially and an orthogonal subspace $Q_i P_i \mathscr{K} = Q_i \mathscr{K}$ where they act as multiplications with the scalars b_i and $1/b_i$, respectively. As a consequence, T_1 and T_2 are inverses of each other. Because of $\lim_{i \to \infty} b_i = 0$, T_1 is bounded ($\|T_1\| = 1$), whereas T_2 is not. Also, it is clear that T_1 and T_2 are Γ-invariant and symmetric. In particular, it follows that $\langle T_1 u_1, T_2 u_2 \rangle = \langle u_1, T_1 T_2 u_2 \rangle = \langle u_1, u_2 \rangle$ for any $u_1, u_2 \in D_0$. We are thus in the situation of Lemma 3.1 and obtain an unbounded symplectic operator

$$T : D_0 \to \mathscr{K}, \quad T = T_2 \frac{1 + \Gamma}{2} + T_1 \frac{1 - \Gamma}{2}.$$

In the next step, T has to be extended to a larger domain $D_T \supset \mathscr{L}$. To this end, we analyze its singular behavior for $|\vec{k}| \to 0$ by comparing it with powers of (a regularized version ω_r of) the one-particle Hamiltonian ω. Setting

$$\omega_r := \omega(1-P_0) + \epsilon_1 P_0 = \begin{cases} \omega & \text{on } (1-P_0)\mathscr{K}, \\ \epsilon_1 \mathbf{1} & \text{on } P_0\mathscr{K}, \end{cases}$$

and noting that $\omega_r^{1/2} D_0 \subset D_0$, we obtain the following Lemma:

Lemma 3.2: $T_2 \omega_r^{1/2}$ is bounded.

Proof: Making use of $\|\omega_r P_i\| = \epsilon_i$ for $i \in \mathbf{N}$, one obtains for $v \in D_0$

$$\|(T_2-1)\omega_r^{1/2} v\|^2 = \left\| \sum_i \left(\frac{1}{b_i}-1\right) Q_i \omega_r^{1/2} v \right\|^2 = \sum_i \left(\frac{1}{b_i}-1\right)^2 \langle \omega_r^{1/2} v, Q_i \omega_r^{1/2} v \rangle$$

$$\leqslant \sum_i \left(\frac{1}{b_i}-1\right)^2 \mathrm{rk} Q_i \langle \omega_r^{1/2} v, P_i \omega_r^{1/2} v \rangle \leqslant \sum_i \left(\frac{1}{b_i}-1\right)^2 \mathrm{rk} Q_i \epsilon_i \|v\|^2.$$

From the conditions imposed on the b_i, it follows that $\Sigma_i (1/b_i - 1)^2 \,\mathrm{rk} Q_i \epsilon_i$ is finite. Thus $(T_2 - 1)\omega_r^{1/2}$ is bounded, hence also $T_2 \omega_r^{1/2}$. ∎

We now can extend T_1 by continuity to all of $\mathscr{K} =: D_{T_1}$ and T_2 by the formula

$$T_2 v := T_2 \omega_r^{1/2} \omega_r^{-1/2} v, \quad v \in \omega_r^{1/2} \mathscr{K}$$

to the dense subspace $\omega_r^{1/2}\mathscr{K} =: D_{T_2}$. (Strictly speaking, the symbol $T_2 \omega_r^{1/2}$ on the right-hand side stands for the continuous extension to $\mathscr{K}$ of the operator considered in the previous Lemma.) Note that T_1 and T_2 still are Γ-invariant. The relevant properties are collected in the following.

Lemma 3.3:

(1) $D_T := \{v \in \mathscr{K} \mid [(1+\Gamma)/2]v \in \omega_r^{1/2}\mathscr{K}\}$ *is a real linear dense subspace of* $\mathscr{K}$.
(2) $T = T_2(1+\Gamma)/2 + T_1(1-\Gamma)/2$ *is well defined on* D_T.
(3) $T: D_T \to \mathscr{K}$ *is a symplectic operator.*
(4) $\mathscr{L} \subset D_T$, *and* $T\mathscr{L}$ *is dense in* $\mathscr{K}$.

Proof: Part 1 is obvious, since $D_0 \subset D_T$; part 2 has been shown in the previous paragraph. For 3, we have to show that $\langle T_1 u_1, T_2 u_2 \rangle = \langle u_1, u_2 \rangle$ remains true for all $u_1 \in D_{T_1}$ and $u_2 \in D_{T_2}$. First, assume $u_1 \in D_0$. Since D_0 is dense in $\mathscr{K}$ and invariant under $\omega_r^{1/2}$, there exists a sequence $u_2^{(n)} \in D_0$, $n \in \mathbf{N}$, such that $\omega_r^{-1/2} u_2 = \lim \omega_r^{-1/2} u_2^{(n)}$, implying $u_2 = \lim u_2^{(n)}$. Using the boundedness of $T_2 \omega_r^{1/2}$, we can compute

$$\langle T_1 u_1, T_2 u_2 \rangle = \langle T_1 u_1, T_2 \omega_r^{1/2} \omega_r^{-1/2} u_2 \rangle = \left\langle T_1 u_1, T_2 \omega_r^{1/2} \lim_{n\to\infty} \omega_r^{-1/2} u_2^{(n)} \right\rangle$$

$$= \lim_{n\to\infty} \langle T_1 u_1, T_2 \omega_r^{1/2} \omega_r^{-1/2} u_2^{(n)} \rangle = \lim_{n\to\infty} \langle u_1, u_2^{(n)} \rangle = \langle u_1, u_2 \rangle.$$

Since T_1 is bounded, the restriction on u_1 can now be dropped by continuity, thus yielding the assertion. Finally, $\mathscr{L} \subset D_T$ is obvious, and the remaining part of 4 is equivalent, in terms of T_1 and T_2, to

$$\frac{1+\Gamma}{2} T\mathscr{L} = T_2 \frac{1+\Gamma}{2}\,\mathscr{L} = T_2 \omega^{-1/2} \widehat{\mathscr{D}_{\mathbf{R}}} \text{ is dense in } \frac{1+\Gamma}{2}\,\mathscr{K},$$

$$\frac{1-\Gamma}{2i} T\mathscr{L} = T_1 \frac{1-\Gamma}{2i}\,\mathscr{L} = T_1 \omega^{1/2} \widehat{\mathscr{D}_{\mathbf{R}}} \text{ is dense in } \frac{1-\Gamma}{2i}\,\mathscr{K}.$$

By $\mathbf{C}$-linearity, this in turn is equivalent to $T_2 \omega^{-1/2} \widehat{\mathscr{D}_{\mathbf{C}}} = T_2 \omega_r^{1/2} \omega_r^{-1/2} \omega^{-1/2} \widehat{\mathscr{D}_{\mathbf{C}}}$ and $T_1 \omega^{1/2} \widehat{\mathscr{D}_{\mathbf{C}}}$ both being dense in $\mathscr{K}$. But this is implied by the fact that, on the one hand, both operators $T_2 \omega_r^{1/2}$ and

T_1 are bounded and have dense images (since they are invertible on the dense, invariant subspace D_0) and that, on the other hand, the subspaces $\omega_r^{-1/2}\omega^{-1/2}\widehat{\mathscr{D}_{\mathbf{C}}}$ and $\omega^{1/2}\widehat{\mathscr{D}_{\mathbf{C}}}$ are dense in $\mathscr{K}$ (by the spectral calculus of ω). ■

With the above preparations, we can define a state $\omega_T : \mathscr{A} \to \mathbf{C}$ and analyze its main properties.

Proposition 3.4: The quasifree state ω_T, defined on $\mathscr{W}(\mathscr{L})$ by

$$\omega_T(W(f)) = e^{-(1/4)\|Tf\|^2}, \quad f \in \mathscr{L},$$

extends to a unique locally normal state ω_T over the quasilocal algebra $\mathscr{A}$. This state is pure and has positive energy.

Proof: The difficult part of this proof is to obtain local normality of ω_T on the net $\mathscr{O} \mapsto \mathscr{W}(\mathscr{L}(\mathscr{O}))$ of Weyl algebras. To this end, recall that T is (on $\mathscr{L}$) the strong limit of symplectic operators T_n such that $T_n - 1$ have finite rank. As a consequence, the associated quasifree states ω_{T_n} are vector states in the vacuum representation and converge weakly to ω_T on $\mathscr{W}(\mathscr{L})$. Now since the Fredenhagen–Hertel compactness condition $C_\sharp$[7,8] is known to be fulfilled in the present model, we can conclude that ω_T is locally normal if the sequence $(\omega_{T_n})_{n \in \mathbf{N}}$ is bounded with respect to some exponential energy norm $\|\cdot\|_\beta$, $\beta > 0$ defined by $\|\omega\|_\beta^2 := \omega(e^{2\beta H})$. But this follows from $\Sigma_i(\epsilon_i/b_i^2)\mathrm{rk}Q_i < \infty$, as F. Hars has shown in Ref. 9, adapting ideas from Ref. 4. (Although our involution Γ differs from that of Refs. 4 and 9, the arguments leading to this conclusion are still valid.) Hence, ω_T is locally normal on $\mathscr{W}(\mathscr{L})$ and thus extends uniquely to a locally normal state on $\mathscr{A}$. Since it is a weak limit of states in the vacuum representation with positive energy, the arguments of Buchholz and Doplicher[10] can be applied to show that ω_T has positive energy, too. Finally, the relation $\overline{T\mathscr{L}} = \mathscr{K}$, established in Lemma 3.3, implies that ω_T is pure, as has been noted at the very beginning of this section. ■

Remark: The inequality $\Sigma_i(\epsilon_i/b_i^2)\mathrm{rk}Q_i < \infty$, which played a crucial role in the previous proof, has a direct physical interpretation. Indeed, performing the limit $\omega_{T_n} \to \omega_T$ corresponds to the excitation of more and more low-energy ''photon'' modes in comparison to the vacuum, namely those singled out by the projections Q_i, $i = 1, \ldots, n$, which appear in T. Since $1/b_i$ measures the amplitude of these modes, each of them carries an energy of about ϵ_i/b_i^2. Hence the modes in the energy interval $[\epsilon_{i+1}, \epsilon_i]$ contribute with (at most) $(\epsilon_i/b_i^2)\mathrm{rk}Q_i$ to the mean energy of the state ω_T, and the above inequality thus means that ω_T describes an infrared cloud with finite total energy. We conjecture that these arguments can be sharpened in order to prove that the transition energy[11] between the sectors π_0 and π_T vanishes. In terms of Ref. 6, the properties of π_T could then be summarized by saying that it is an ''infravacuum representation,'' and we will indeed use this terminology in the sequel.

C. KPR-like quasifree states

We reach our goal of improving the localization of the automorphisms γ by considering a special class of infravacuum representations. The main idea, due to Ref. 4, is to control the angular momentum carried by the low-energy modes. It may be formalized as follows.

Definition: A state ω_T over $\mathscr{A}$ based on the sequences ϵ_i, Q_i, b_i as described above is called a KPR-like state [and π_T (resp. T) a KPR-like representation (resp. symplectic operator)] if the following two additional conditions are fulfilled:

(1) $(\ln(\epsilon_i/\epsilon_{i+1}))_{i \in \mathbf{N}}$ is polynomially bounded, and $\Sigma_i b_i^2 \ln(\epsilon_i/\epsilon_{i+1}) < \infty$.

(2) With respect to the tensor product structure of the subspace $P_i\mathscr{K}$, $P_i\mathscr{K} \cong L^2([\epsilon_{i+1}, \epsilon_i], \omega^2 d\omega) \otimes L^2(S^2)$, the projections Q_i read

$$Q_i = \frac{|\xi_i\rangle\langle\xi_i|}{\langle\xi_i|\xi_i\rangle} \otimes \tilde{Q}_i \quad \text{with} \quad \tilde{Q}_i := \sum_{0 < l \leqslant i} \sum_{m=-l}^{l} |Y_{lm}\rangle\langle Y_{lm}|;$$

here the vector $\xi_i \in L^2([\epsilon_{i+1}, \epsilon_i], \omega^2 d\omega)$ is given by $\xi_i(\omega) = \omega^{-3/2}$ and $Y_{lm} \in L^2(S^2)$ are the spherical harmonics.

This definition has been formulated so as to imply the regularity property of the bounded operator T_1 formulated in the next Lemma. It is only through this result that the two additional

properties of KPR-like infravacua enter the analysis of Sec. IV. It is apparent from the ensuing proof that the above definition may be generalized in several respects. However, we refrain from discussing these possibilities here.

In contrast, we draw the reader's attention to the following crucial difference between our KPR-like states and the ''true'' KPR states as defined in Ref. 4: In our case, the projection $\tilde{Q}_i$ contains no summand $|Y_{00}\rangle\langle Y_{00}|$. In physical terms, this means that the infrared cloud does not contain any spherically symmetric low-energy modes. Such a restriction is necessary, since it is precisely by such modes or, equivalently, by the isotropic long-range behavior of the ''Coulomb potential,'' that the sectors $[\gamma]$ differ from each other. Too strong an $l=0$ contribution to the infrared cloud would therefore render the sectors indistinguishable in front of that background. (Indeed, if one had $0\leqslant l\leqslant i$ in the definition of $\tilde{Q}_i$, one would obtain, instead of Lemma 3.6 below, that $\pi_T\circ\gamma\cong\pi_T$ for all $\gamma\in\mathscr{S}_\Gamma$.) This seemingly artificial restriction on the background states mimics the situation in QED, where the Coulomb field $\vec{\mathscr{E}}(\vec{k})\sim i\vec{k}/\omega^2$ cannot be compensated by transverse photons.

Lemma 3.5: Let the sequences ϵ_i, Q_i, b_i be such that ω_T is a KPR-like state. Let $u\in\mathscr{K}$ have, in a neighborhood of $\vec{k}=0$, the form $u(\vec{k})=\eta(\vec{k}/|\vec{k}|)$ with some $\eta\in C^\infty(S^2)\subset L^2(S^2)$. Then the sequence $(T_1\omega^{-3/2}P_{\epsilon_n}u)_{n\in\mathbf{N}}$ converges iff $\eta\perp Y_{00}$.

Proof: Without any restriction, one may assume $u=c\otimes\eta$ with $c(\omega)=1$ if $\omega<\epsilon_1$. Let $\eta\perp Y_{00}$. For $0<m<n$, one computes

$$T_1\omega^{-3/2}P_{\epsilon_n}u-T_1\omega^{-3/2}P_{\epsilon_m}u=T_1\omega^{-3/2}\sum_{i=m}^{n-1}P_i(c\otimes\eta)=\sum_{i=m}^{n-1}T_1P_i(\xi_i\otimes\eta)$$

$$=\sum_{i=m}^{n-1}((1-Q_i)+b_iQ_i)(\xi_i\otimes\eta)=\sum_{i=m}^{n-1}\xi_i\otimes((1-\tilde{Q}_i)\eta+b_i\tilde{Q}_i\eta).$$

Now $\eta\in C^\infty(S^2)$ entails that $\|(1-\tilde{Q}_i)\eta\|^2=\Sigma_{l>i}\Sigma_m|\langle Y_{lm},\eta\rangle|^2$, $i\in\mathbf{N}$, is a sequence of rapid decrease (since $\eta\in D(\vec{L}^{2N})$ for any N, $\vec{L}$ denoting the angular momentum operator). Thus, using $\|\xi_i\|^2=\int_{\epsilon_{i+1}}^{\epsilon_i}\omega^2d\omega 1/\omega^3=\ln(\epsilon_i/\epsilon_{i+1})$ and $\|b_i\tilde{Q}_i\eta\|^2\leqslant b_i^2\|\eta\|^2$, one obtains

$$\|T_1\omega^{-3/2}P_{\epsilon_n}u-T_1\omega^{-3/2}P_{\epsilon_m}u\|^2\leqslant\sum_{i=m}^{n-1}\|\xi_i\|^2(\|(1-\tilde{Q}_i)\eta\|^2+b_i^2\|\tilde{Q}_i\eta\|^2)\leqslant\sum_{i=m}^{n-1}\ln\frac{\epsilon_i}{\epsilon_{i+1}}\left(\frac{c_N}{i^N}+b_i^2\right).$$

With suitably chosen N, the right-hand side vanishes as $m,n\to\infty$ due to the conditions imposed on ϵ_i and b_i. Hence $(T_1\omega^{-3/2}P_{\epsilon_n}u)_{n\in\mathbf{N}}$ is a Cauchy sequence. Conversely, assume $\langle Y_{00},\eta\rangle\neq0$. With $\eta=\langle Y_{00},\eta\rangle Y_{00}+\eta_1$, $(T_1\omega^{-3/2}P_{\epsilon_n}(c\otimes\eta_1))_{n\in\mathbf{N}}$ is convergent, hence $(T_1\omega^{-3/2}P_{\epsilon_n}u)_{n\in\mathbf{N}}$ is divergent because $(T_1\omega^{-3/2}P_{\epsilon_n}(c\otimes Y_{00}))_{n\in\mathbf{N}}=(\omega^{-3/2}P_{\epsilon_n}(c\otimes Y_{00}))_{n\in\mathbf{N}}$ is. ■

We end this section with a result which shows that the KPR-like infravacua do not affect the superselection structure of the present model. As the previous Lemma, it makes essential use of the fact that $Tf=f$ for all rotation invariant elements $f\in D_T$.

Lemma 3.6: Let π_T be a KPR-like infravacuum representation. Then, for any $\gamma_1,\gamma_2\in\mathscr{S}_\Gamma$, one has

$$\pi_0\circ\gamma_1\cong\pi_0\circ\gamma_2\quad\text{iff}\quad\pi_T\circ\gamma_1\cong\pi_T\circ\gamma_2.$$

Proof: Let $\pi_0\circ\gamma_1\cong\pi_0\circ\gamma_2$. Then $\gamma:=\gamma_1-\gamma_2\in\mathscr{S}_\Gamma$ has charge $q_\gamma=0$, as noted in the Introduction, which does not only yield $\gamma\in\mathscr{K}$, but even $\gamma\in D_T$. Hence, the unitary $W(T\gamma)$ is well-defined and intertwines the representations $\pi_T\circ\gamma_1$ and $\pi_T\circ\gamma_2$. Conversely, assume $\pi_0\circ\gamma_1\ncong\pi_0\circ\gamma_2$, i.e., $q_{\gamma_1}\neq q_{\gamma_2}$. For any *rotation invariant* test function $h\in\mathscr{D}_{\mathbf{R}}(\mathbf{R}^3\backslash\{0\})$, one has $\omega^{-1/2}\hat{h}=:f\in\mathscr{S}(\mathscr{O}')$ for some open neighborhood $\mathscr{O}\subset\mathbf{R}^{1+3}$ of 0. Since $Tf_\lambda=f_\lambda$, Lemma 2.2 implies

$$\pi_T\circ\gamma_j(W(f_\lambda))=\gamma_j(W(f_\lambda))\xrightarrow{\lambda\to\infty}e^{iq_{\gamma_j}\kappa_f}\,e^{-(1/4)\|f\|^2}\mathbf{1}.$$

As it is always possible to obtain $e^{iq_{\gamma_1}\kappa_f} \neq e^{iq_{\gamma_2}\kappa_f}$ by a rescaling of h, it follows that $\pi_T \circ \gamma_1 \not\cong \pi_T \circ \gamma_2$. ■

IV. BETTER LOCALIZATION OF THE SECTORS IN FRONT OF KPR-LIKE INFRAVACUA

The main aim of this section is to prove the following result which establishes some (non-Lorentz invariant) version of BF localization. In the sequel, we will denote by $\mathscr{C} = (\{t\} \times C)''$ an ''upright'' spacelike cone whose basis is the open convex cone $C \subset \mathbf{R}^3$ at time t. Note that the set of upright spacelike cones is translation invariant and that an arbitrary spacelike cone can be obtained from an upright one by a Lorentz transformation.

Proposition 4.1: Let π_T be a KPR-like infravacuum representation, and let $\gamma \in \mathscr{S}_\Gamma$. Then one has for any upright spacelike cone $\mathscr{C}$:

$$\pi_T \circ \gamma|_{\mathscr{A}(\mathscr{C}')} \cong \pi_T|_{\mathscr{A}(\mathscr{C}')}.$$

To prove this assertion, we will first deal with a special case in which the relevant computations can be carried out quite explicitly. Eventually, the formal proof will consist in reducing the general case to the special one.

The case discussed first amounts to the following two assumptions:

(i) $\mathscr{C} = (\{0\} \times C)''$ and the apex of C is the origin $0 \in \mathbf{R}^3$;

(ii) $\gamma \in \mathscr{S}_\Gamma$ has the special form $\gamma = i\omega^{-3/2}\hat{\rho}$, where $\rho \in \mathscr{D}_{\mathbf{R}}(\mathbf{R}^3)$ satisfies $\rho = -\Delta\Phi$ with a rotation invariant function $\Phi \in C^\infty_{\mathbf{R}}(\mathbf{R}^3)$ obeying, for some $0 < r_1 < r_2 < \infty$,

$$\Phi(\vec{x}) = \begin{cases} 0 & \text{if } |\vec{x}| < r_1, \\ \dfrac{q_\gamma}{4\pi|\vec{x}|} & \text{if } |\vec{x}| > r_2. \end{cases}$$

To proceed, we note that the cone $C \subset \mathbf{R}^3$ determines, by projection onto the unit sphere S^2, a subset of S^2 which we denote by C, too. Now we choose a function $\chi^C \in C^\infty_{\mathbf{R}}(S^2)$ with the properties

$$(i)\ \ \chi^C|_{S^2 \setminus C} = 1 \quad \text{and} \quad (ii)\ \ \langle Y_{00}, \chi^C \rangle = 0,$$

and denote by $\Phi^C \in C^\infty_{\mathbf{R}}(\mathbf{R}^3)$ the product

$$\Phi^C(\vec{x}) := (\Phi \cdot \chi^C)(\vec{x}) := \Phi(\vec{x})\chi^C(\vec{x}/|\vec{x}|).$$

This function will now be used to construct a unitary intertwiner from π_T to $\pi_T \circ \gamma$ on the C*-algebra $\mathscr{W}(\mathscr{A}(\mathscr{C}'))$. [Here and in the sequel, we use the notation $\Psi \cdot \eta$ for the pointwise product of a rotation invariant function Ψ and the function $\vec{u} \mapsto \eta(\vec{u}/|\vec{u}|)$, where $\eta \in C^\infty(S^2)$. For definiteness, we let $(\Psi \cdot \eta)(0) := 0$.]

For this purpose, we calculate (using spherical coordinates)

$$-\Delta\Phi^C = \rho \cdot \chi^C + \frac{\Phi}{r^2} \cdot \vec{L}^2 \chi^C.$$

This function is square integrable, hence its Fourier transform $u^C := \widehat{-\Delta\Phi^C}$ lies in $\mathscr{K}$, and

$$v_n^C := i\omega^{-3/2} P_{\epsilon_n} u^C, \quad n \in \mathbf{N},$$

is a well-defined sequence in D_0 which approximates the linear form l_γ on $\mathscr{C}'$ in the following sense:

Lemma 4.2: For any $f \in \mathscr{A}(\mathscr{C}')$, one has $l_\gamma(f) = -\lim_{n\to\infty} \mathrm{Im}\langle v_n^C, f \rangle$.

Proof: Write $f = \omega^{-1/2}\hat{h} + i\omega^{+1/2}\hat{g}$ with $h, g \in \mathscr{D}_{\mathbf{R}}(\mathbf{R}^3 \setminus \overline{C})$ and consider

$$-\mathrm{Im}\langle v_n^C, f\rangle = -\mathrm{Im}\langle i\omega^{-3/2}P_{\epsilon_n}u^C, \omega^{-1/2}\hat{h} + i\omega^{+1/2}\hat{g}\rangle = \int_{|\vec{k}|>\epsilon_n} d^3k\,\omega^{-2}\overline{u^C(\vec{k})}\hat{h}(\vec{k}).$$

Since $u^C \in L^\infty_{\mathrm{loc}}(\mathbf{R}^3)$ (cf. Lemma 4.3), whence $\widehat{\Phi^C} = \omega^{-2}u^C \in L^1_{\mathrm{loc}}(\mathbf{R}^3)$, it follows that this sequence converges for $n\to\infty$ to

$$\int_{\mathbf{R}^3} d^3k\,\overline{\widehat{\Phi^C}(\vec{k})}\hat{h}(\vec{k}) = \Phi^C(h) = \Phi(h) = \int_{\mathbf{R}^3} d^3k\,\overline{\hat{\Phi}(\vec{k})}\hat{h}(\vec{k}).$$

In the previous line, we have viewed Φ^C and Φ as distributions and made use of the fact that they coincide on supp h. The proof is now completed by a straightforward computation showing that the last expression equals $l_\gamma(f)$. ■

Whereas property (*i*) of χ^C was essential for the previous Lemma, the following one will show how property (*ii*) determines the behavior of u^C in a neighborhood of $\vec{k}=0$.

Lemma 4.3: There exists a smooth function $\eta \in C^\infty(S^2)$ *with* $\langle Y_{00}, \eta\rangle = 0$ *and an analytic function* $R:\mathbf{R}^3\to\mathbf{C}$ *with* $R(0)=0$ *such that*

$$u^C(\vec{k}) = \eta(\vec{k}/|\vec{k}|) + R(\vec{k}) \quad for \;\; \vec{k}\neq 0.$$

Proof: Let $\mathscr{S}_{00}$ denote the set of all rotation-invariant test functions. Since $\langle Y_{00}, \vec{L}^2\chi^C\rangle = 0$, there exists a unique distribution F_1 on $\mathbf{R}^3$ which is homogeneous of degree -3 and which coincides on $\mathbf{R}^3\backslash\{0\}$ with $(q_\gamma/4\pi)(1/r^3)\cdot\vec{L}^2\chi^C$. By Theorems 7.1.16 and 18 of Ref. 12 it follows that its Fourier transform $\widehat{F_1}$ is homogeneous of degree 0 and restricts on $\mathbf{R}^3\backslash\{0\}$ to a smooth function, i.e., $\hat{F}_1(\vec{k}) = \eta(\vec{k}/|\vec{k}|)$ for $\vec{k}\neq 0$ with some $\eta\in C^\infty(S^2)$. Moreover, since $F_1|_{\mathscr{S}_{00}} = 0$ and $\mathscr{S}_{00}$ is stable under Fourier transformations, it follows that $\langle Y_{00}, \eta\rangle = 0$. Now consider the distribution $F_2 := -\Delta\Phi^C - F_1$. For $r\neq 0$, it is given by $F_2 = \rho\cdot\chi^C + (\Phi - q_\gamma/4\pi r)1/r^2\cdot\vec{L}^2\chi^C$ and thus has compact support. Hence, its Fourier transform is an analytic function R: $\hat{F}_2(\vec{k}) = R(\vec{k})$, $\vec{k}\in\mathbf{R}^3$. As χ^C was assumed to fulfill $\langle Y_{00}, \chi^C\rangle = 0$, it follows in particular that $R(0) = \int d^3x(\rho\cdot\chi^C)(\vec{x}) = 0$. To sum up, we have $-\Delta\Phi^C = F_1 + F_2$ (in the sense of distributions) and thus $-\widehat{\Delta\Phi^C} = \hat{F}_1 + \hat{F}_2$. Since all three terms of this last equation are smooth on $\mathbf{R}^3\backslash\{0\}$, this implies the identity $u^C(\vec{k}) = -\widehat{\Delta\Phi^C}(\vec{k}) = \eta(\vec{k}/|\vec{k}|) + R(\vec{k})$ for all $\vec{k}\neq 0$. ■

The knowledge of u^C at $\vec{k}=0$ now permits us to establish the connection with the KPR-like infravacuum representations described in Sec. III.

Lemma 4.4: Let T be a KPR-like symplectic operator. Then we have the following.

(1) *The limit* $v_T^C := \lim_{n\to\infty} Tv_n^C$ *exists in* $\mathscr{K}$.
(2) *The unitary* $W(v_T^C)$ *satisfies*

$$\mathrm{Ad}\; W(v_T^C)\circ\pi_T = \pi_T\circ\gamma \;\; on \;\; \mathscr{A}(\mathscr{C}').$$

Proof: Since $\Delta\Phi^C$ is real-valued, one has $\Gamma u^C = u^C$, hence $\Gamma v_n^C = -v_n^C$. Thus, $Tv_n^C = T_1v_n^C = iT_1\omega^{-3/2}P_{\epsilon_n}u^C = iT_1\omega^{-3/2}P_{\epsilon_n}(u_1^C + u_2^C)$ with $u_1^C, u_2^C \in \mathscr{K}$ defined by $u_1^C := (1-P_0)(1\cdot\eta)$ and $u_2^C := u^C - u_1^C = P_0(1\cdot\eta) + R$, where η and R are as in the previous Lemma. In particular, $R(0) = 0$ yields $u_2^C \in D_{\omega^{-3/2}}$ which implies $T_1\omega^{-3/2}P_{\epsilon_n}u_2^C = T_1P_{\epsilon_n}\omega^{-3/2}u_2^C \xrightarrow{n\to\infty} T_1\omega^{-3/2}u_2^C$ by the boundedness of T_1. On the other hand, it follows from $\langle Y_{00}, \eta\rangle = 0$ by Lemma 3.5 that the sequence $(T_1\omega^{-3/2}P_{\epsilon_n}u_1^C)_{n\in\mathbf{N}}$ is convergent, which completes the proof of Part 1. Part 2 is a straightforward computation: Let $f\in\mathscr{S}(\mathscr{C}')$; then, by Lemma 4.2, $\mathrm{Im}\langle v_T^C, Tf\rangle = \lim_{n\to\infty}\mathrm{Im}\langle Tv_n^C, Tf\rangle = \lim_{n\to\infty}\mathrm{Im}\langle v_n^C, f\rangle = -l_\gamma(f)$, which implies

$$\mathrm{Ad}\; W(v_T^C)(\pi_T(W(f))) = W(v_T^C)W(Tf)W(v_T^C)^* = e^{-i\,\mathrm{Im}\langle v_T^C, Tf\rangle}W(Tf)$$
$$= e^{il_\gamma(f)}W(Tf) = \pi_T\circ\gamma(W(f)).$$

This establishes the stated equivalence on $\mathscr{W}(\mathscr{L}(\mathscr{C}'))$ and, hence, by local normality (of both π_T and γ) also on $\mathscr{A}(\mathscr{C}')$. ■

With these preparations, we are ready for the following proof.

Proof of Proposition 4.1: By standard arguments using transportability of the charges, $\gamma_x \cong \gamma$, and translation covariance of the representation π_T, it can always be assumed that the apex of $\mathscr{C}$ is $0 \in \mathbf{R}^{1+3}$. To remove the assumption on the special form of γ as well, we note that, for any $\gamma \in \mathscr{S}_\Gamma$, there exists some equivalent $\gamma_0 \in \mathscr{S}_\Gamma$ with the special form considered. Such a γ_0 automatically satisfies $\gamma_0 - \gamma \in D_T$ and thus provides a unitary $W(T(\gamma_0 - \gamma))$ performing the equivalence $\pi_T \circ \gamma \cong \pi_T \circ \gamma_0$ on all of $\mathscr{A}$. Taking into account all the above, this proves Proposition 4.1. ■

V. CONCLUSIONS

The present work has shown in a concrete example that choosing a background different from the vacuum can improve the localizability properties of superselection sectors in theories with massless particles. Typically, such backgrounds correspond to clouds of infrared radiation which exist in great variety in any such theory, but it appears that *suitable* background states have to be chosen carefully in order to match with the sectors under consideration. To illustrate this point, we recall that this led us, in particular, to choose complex conjugation in *position* space as the involution Γ. If we had, on the other hand, chosen complex conjugation $\hat{\Gamma}$ in *momentum* space instead of Γ, our sectors would only have been localizable in *upright spacelike double cones*, i.e., in regions of the form $\mathscr{C} \cup (a - \mathscr{C})$, where $\mathscr{C}$ is an upright spacelike cone and $a \in \mathbf{R}^{1+3}$, but not in spacelike cones any more. (The former statement can be verified with the method of Sec. IV, the latter can be reduced to an application of Lemma 2.2.)

Finally, we remark that we were unable, with our methods, to establish the *full* BF localization criterion for the sectors $[\gamma]$, i.e., $\pi_T \circ \gamma|_{\mathscr{A}(\mathscr{C}')} \cong \pi_T|_{\mathscr{A}(\mathscr{C}')}$ even for spacelike cones $\mathscr{C}$ which do not contain an upright one. Apart from the obvious fact that the KPR-like states break Lorentz covariance explicitly, there seem to be other indications that such a result might indeed not be true. However, we do not pursue this point further, since the localization properties obtained here should be sufficient for carrying through a DHR-like analysis (along the lines of Ref. 2) in front of the infravacuum background. If this is indeed possible, it has to be studied in a subsequent step under which conditions (to be imposed on the infravacuum) the superselection structure thus obtained will be independent from the particular background.

ACKNOWLEDGMENTS

I would like to thank Professor D. Buchholz for numerous helpful discussions and stimulating interest in this work. Partial financial support from the Deutsche Forschungsgemeinschaft (''Graduiertenkolleg Theoretische Elementarteilchenphysik'' at Hamburg University) is also gratefully acknowledged.

[1] S. Doplicher, R. Haag, and J. E. Roberts, Commun. Math. Phys. **23**, 199 (1971); **35**, 49 (1974).
[2] D. Buchholz and K. Fredenhagen, Commun. Math. Phys. **84**, 1 (1982).
[3] D. Buchholz, Commun. Math. Phys. **85**, 49 (1982).
[4] K. Kraus, L. Polley, and G. Reents, Ann. Inst. Henri Poincaré, Sect. A **26**, 109 (1977).
[5] D. Buchholz, S. Doplicher, G. Morchio, J. E. Roberts, and F. Strocchi, in *Operator Algebras and Quantum Field Theory*, edited by S. Doplicher, R. Longo, J. E. Roberts, and L. Szido (International Press, Cambridge, 1997).
[6] W. Kunhardt, J. Math Phys. **39**, 3589 (1998).
[7] K. Fredenhagen and J. Hertel, unpublished manuscript (1979).
[8] D. Buchholz and M. Porrmann, Ann. Inst. Henri Poincaré Phys. Theor. **52**, 237 (1990).
[9] F. Hars, ''Infravakua in der Quantenelektrodynamik,'' Diploma thesis, Hamburg University, 1993.
[10] D. Buchholz and S. Doplicher, Ann. Inst. Henri Poincaré Phys. Theor. **40**, 175 (1984).
[11] R. Wanzenberg, ''Energie und Präparierbarkeit von Zuständen in der lokalen Quantenfeldtheorie,'' Diploma thesis, Hamburg University, 1987.
[12] L. Hörmander, *The Analysis of Linear Partial Differential Operators I* (Springer, Berlin, 1990).

Coexistent observables and effects in a convexity approach

Pekka Lahti[a)]
Department of Physics, University of Turku, FIN-20014 Turku, Finland

Sylvia Pulmannova[b)]
Mathematical Institute, Slovak Academy of Sciences, 81473 Bratislava, Slovakia

Kari Ylinen[c)]
Department of Mathematics, University of Turku, FIN-20014 Turku, Finland

(Received 20 November 1997; accepted for publication 17 July 1998)

We characterize functionally coexistent observables in terms of biobservables and joint observables. Coexistent sets of effects are characterized in terms of projective systems of simple observables. © *1998 American Institute of Physics.* [S0022-2488(98)00712-9]

I. INTRODUCTION

In an operational formulation of quantum mechanics the notions of *states* (equivalence classes of preparation procedures) and *observables* (equivalence classes of measurements or registration procedures) occur in a statistical duality: each pair (s,o) of a state s and an observable o determines a probability distribution $p(s,o)$ of the possible measurement outcomes in the given experiment.[1,2] This leads one to express the notion of an observable as an assignment $s\mapsto p(s,o)$. Should it be possible to mix preparation procedures into new preparation procedures, the set of states acquires the structure of a convex set. In such a case it is natural to require that the assignment $s\mapsto p(s,o)$ preserves this convex structure.

Given two observables $s\mapsto p(s,o_1)$ and $s\mapsto p(s,o_2)$, the question arises under which conditions it is possible to collect the relevant experimental data under a single observable $s\mapsto p(s,o)$. This leads to the notion of the *coexistence* of observables, and it is an important question in quantum mechanics to characterize such (pairs of or collections of) observables. In Sec. I of this paper we shall define the notion of an observable as an affine function from an abstract convex set to the set of probability measures on a measurable space, and we introduce the notion of the coexistence of such observables. Theorem 2.1.5 then characterizes an important class of functionally coexistent observables in terms of so-called biobservables and joint observables. In the remaining part of Sec. I we illustrate this result in the framework of a total effect algebra (Sec. II B) as well as in the Hilbert space formulation of quantum mechanics (Sec. II C).

In Sec. III we regard an *effect* as a $[0,1]$-valued affine function on an abstract convex set of states. Adopting the usual notion of the coexistence of effects in this setting, we then characterize coexistent sets of effects in terms of projective systems of simple observables.

II. COEXISTENT OBSERVABLES

A. Observables as affine functions

Let S be a convex set, that is, a nonempty set with a convex structure:[3]

(CS1) for any finite set of positive numbers $w_1,\ldots,w_n$ satisfying $\Sigma w_i=1$ and any $s_1,\ldots,s_n\in S$ there is a unique element $\langle w_1,\ldots,w_n;s_1,\ldots,s_n\rangle\in S$;
(CS2) $\langle w_1,\ldots,w_n;s,\ldots,s\rangle=s$.

[a)]Electronic mail: pekka.lahti@utu.fi
[b)]Electronic mail: pulmann@mau.savba.sk
[c)]Electronic mail: kari.edvin.ylinen@utu.fi

 © 1998 American Institute of Physics

We assume that the set of (normalized) states of a physical system forms a convex set and we interpret $\langle w_1,\ldots,w_n;s_1,\ldots,s_n\rangle$ as a mixture of the states $s_1,\ldots,s_n$ with the weights $w_1,\ldots,w_n$, respectively.

Let $(\Omega,\mathcal{B})$ be a measurable space, where $\mathcal{B}$ is a σ-algebra of subsets of a nonempty set Ω, and let $M_1^+(\Omega,\mathcal{B})$ denote the convex set of probability measures on $(\Omega,\mathcal{B})$. Let $E:S\to M_1^+(\Omega,\mathcal{B})$ be an affine function, that is, a function $s\mapsto E(s)$ which has the property

$$E(\langle w,1-w;s_1,s_2\rangle)=wE(s_1)+(1-w)E(s_2),$$

for all $s_1,s_2\in S$, and for any $0\leqslant w\leqslant 1$. Let $A_1^+(S)$ denote the set of affine functions $f:S\to[0,1]$. Then we get a map $\tilde{E}:\mathcal{B}\to A_1^+(S)$ by defining

$$\tilde{E}(X)(s):=E(s)(X),$$

for all $s\in S,X\in\mathcal{B}$. Equipping $A_1^+(S)$ with the topology τ of pointwise convergence, the induced product topology of $A_1^+(S)\subset[0,1]^S$, we get

$$\tilde{E}\ \text{ is } \sigma\text{-additive with respect to } \tau,$$

$$\tilde{E}(\Omega)=1_S,$$

where 1_S is the constant one function.

Theorem 2.1.1: The map $E\mapsto\tilde{E}$ is an affine bijection from the convex set of affine functions $E:S\to M_1^+(\Omega,\mathcal{B})$ onto the convex set of σ-additive functions $F:\mathcal{B}\to A_1^+(S)$ (with respect to τ) satisfying $F(\Omega)=1_S$.

Proof: The injectivity is obvious. Let now $F:\mathcal{B}\to A_1^+(S)$ be σ-additive with respect to τ and satisfy $F(\Omega)=1_S$. Define $E(s)(X):=F(X)(s)$ for all $s\in S,X\in\mathcal{B}$. Then $S\ni s\mapsto E(s)\in M_1^+(\Omega,\mathcal{B})$ is affine, and $F=\tilde{E}$.

Definition 2.1.2: (*a*) An affine map $E:S\to M_1^+(\Omega,\mathcal{B})$ is called an *observable*. (*b*) Two observables $E_i:S\to M_1^+(\Omega_i,\mathcal{B}_i)$, $i=1,2$, are called *coexistent*, if there is a measurable space $(\Omega,\mathcal{B})$ with an observable $E:S\to M_1^+(\Omega,\mathcal{B})$ such that

$$\tilde{E}_1(\mathcal{B}_1)\cup\tilde{E}_2(\mathcal{B}_2)\subseteq\tilde{E}(\mathcal{B}).$$

(*c*) Two observables $E_i:S\to M_1^+(\Omega_i,\mathcal{B}_i)$, $i=1,2$, are *functionally coexistent* if there is a measurable space $(\Omega,\mathcal{B})$ with an observable $E:S\to M_1^+(\Omega,\mathcal{B})$ such that for some measurable functions $f_i:\Omega\to\Omega_i$,

$$E_i(s)(X)=E(s)(f_i^{-1}(X)),$$

for all $s\in S,X\in\mathcal{B}_i$, $i=1,2$.

We observe that the condition $E_i(s)(X)=E(s)(f_i^{-1}(X))$, $s\in S,X\in\mathcal{B}_i$, is equivalent to the condition $\tilde{E}_i(X)=\tilde{E}(f_i^{-1}(X))$, $X\in\mathcal{B}_i$. Obviously, functionally coexistent observables are coexistent.

For any two measurable spaces $(\Omega_i,\mathcal{B}_i)$, $i=1,2$, we let $BM_1^+(\Omega_1\times\Omega_2,\mathcal{B}_1\times\mathcal{B}_2)$ denote the set of probability bimeasures on $(\Omega_1\times\Omega_2,\mathcal{B}_1\times\mathcal{B}_2)$, where $\mathcal{B}_1\times\mathcal{B}_2=\{(X,Y)|X\in\mathcal{B}_1,Y\in\mathcal{B}_2\}$. By definition, if $\Phi\in BM_1^+(\Omega_1\times\Omega_2,\mathcal{B}_1\times\mathcal{B}_2)$, then $\Phi(\Omega_1\times\Omega_2)=1$, and for each $X\in\mathcal{B}_1$ the partial function $Y\mapsto\Phi(X,Y)$ is a $[0,1]$-valued measure on $(\Omega_2,\mathcal{B}_2)$ and for each $Y\in\mathcal{B}_2$ the function $X\mapsto\Phi(X,Y)$ is a $[0,1]$-valued measure on $(\Omega_1,\mathcal{B}_1)$. Let $B:S\to BM_1^+(\Omega_1\times\Omega_2,\mathcal{B}_1\times\mathcal{B}_2)$ be an affine function, and define $\tilde{B}:\mathcal{B}_1\times\mathcal{B}_2\to A_1^+(S)$ by putting

$$\tilde{B}(X,Y)(s):=B(s)(X,Y),$$

for all $s\in S,X\in\mathcal{B}_1,Y\in\mathcal{B}_2$. Analogously with Theorem 2.1.1 we get the following.

Theorem 2.1.3: The map $B \mapsto \tilde{B}$ is an affine bijection from the convex set of affine functions $B:S \to BM_1^+(\Omega_1 \times \Omega_2, \mathcal{B}_1 \times \mathcal{B}_2)$ onto the convex set of separately σ-additive (with respect to τ) functions $C:\mathcal{B}_1 \times \mathcal{B}_2 \to A_1^+(S)$ satisfying $C(\Omega_1, \Omega_2) = 1_S$.

Definition 2.1.4: An affine map $B:S \to BM_1^+(\Omega_1 \times \Omega_2, \mathcal{B}_1 \times \mathcal{B}_2)$ is called a *biobservable*.

The question arises under which conditions a biobservable determines an observable.

Assume now that Ω is a Hausdorff space, and let $\mathcal{B}(\Omega)$ be its Borel σ-algebra, that is, the σ-algebra generated by the open subsets of Ω. We call $(\Omega, \mathcal{B}(\Omega))$ a Hausdorff Borel space. Following Ref. 4 we call a probability measure $\mu:\mathcal{B}(\Omega) \to [0,1]$ a *Radon probability measure* if it satisfies the regularity condition:

$$\mu(X) = \sup\{\mu(C) \mid C \subseteq X, C \text{ compact}\}$$

for each $X \in \mathcal{B}(\Omega)$. We let $RM_1^+(\mathcal{B}(\Omega))$ denote the set of Radon probability measures on $(\Omega, \mathcal{B}(\Omega))$, and we call an affine map $E:S \to RM_1^+(\mathcal{B}(\Omega))$ a *Radon observable*. Similarly, if $(\Omega_1, \mathcal{B}(\Omega_1))$ and $(\Omega_2, \mathcal{B}(\Omega_2))$ are two Hausdorff Borel spaces, we say that a probability bimeasure $\Phi:\mathcal{B}(\Omega_1) \times \mathcal{B}(\Omega_2) \to [0,1]$ is a Radon probability bimeasure if

$$\Phi(X,Y) = \sup\{\Phi(C,D) \mid C \subseteq X, D \subseteq Y, C, D \text{ compact}\}$$

for all Borel sets $X \in \mathcal{B}(\Omega_1)$ and $Y \in \mathcal{B}(\Omega_2)$. We let $RBM_1^+(\mathcal{B}(\Omega_1) \times \mathcal{B}(\Omega_2))$ denote the set of Radon probability bimeasures on $(\Omega_1 \times \Omega_2, \mathcal{B}(\Omega_1) \times \mathcal{B}(\Omega_2))$, and we call an affine map $B:S \to RBM_1^+(\mathcal{B}(\Omega_1) \times \mathcal{B}(\Omega_2))$ a *Radon biobservable*.

Let $(\Omega_1, \mathcal{B}(\Omega_1))$ and $(\Omega_2, \mathcal{B}(\Omega_2))$ be two Hausdorff Borel spaces, and let $\mathcal{B}(\Omega_1 \times \Omega_2)$ denote the Borel σ-algebra of $\Omega_1 \times \Omega_2$. Let $E_1:S \to RM_1^+(\mathcal{B}(\Omega_1))$ and $E_2:S \to RM_1^+(\mathcal{B}(\Omega_2))$ be two Radon observables. We say that E_1 and E_2 have a *Radon joint observable* if there is a Radon observable $E:S \to RM_1^+(\mathcal{B}(\Omega_1 \times \Omega_2))$ such that $E_1(s)(X) = E(s)(X \times \Omega_2)$ and $E_2(s)(Y) = E(s)(\Omega_1 \times Y)$ for all $X \in \mathcal{B}(\Omega_1), Y \in \mathcal{B}(\Omega_2)$, and for all $s \in S$. Similarly, we say that E_1 and E_2 have a Radon biobservable if there is a Radon biobservable $B:S \to RBM_1^+(\mathcal{B}(\Omega_1) \times \mathcal{B}(\Omega_2))$ such that $\tilde{E}_1(X) = \tilde{B}(X, \Omega_2)$ and $\tilde{E}_2(Y) = \tilde{B}(\Omega_1, Y)$ for all $X \in \mathcal{B}(\Omega_1), Y \in \mathcal{B}(\Omega_2)$. Similarly, we may specify, for instance, the notion of functional coexistence of Definition 2.1.2 to the notion of *functional R-coexistence* assuming that all the observables involved are Radon observables.

Theorem 2.1.5: Let $(\Omega_1, \mathcal{B}(\Omega_1))$ and $(\Omega_2, \mathcal{B}(\Omega_2))$ be two Hausdorff Borel spaces. For any two Radon observables $E_1:S \to RM_1^+(\mathcal{B}(\Omega_1))$ and $E_2:S \to RM_1^+(\mathcal{B}(\Omega_2))$ the following three statements are equivalent:

(a) E_1 and E_2 have a Radon biobservable;
(b) E_1 and E_2 have a Radon joint observable;
(c) E_1 and E_2 are functionally R-coexistent.

Proof: $(a) \Rightarrow (b)$ Let B be a Radon biobservable associated with E_1 and E_2. Then, for any $s \in S$, the Radon probability bimeasure $B(s)$ determines a unique Radon probability measure $\mu(B,s)$ on $(\Omega_1 \times \Omega_2, \mathcal{B}(\Omega_1 \times \Omega_2))$ such that for all $X \in \mathcal{B}(\Omega_1)$, $Y \in \mathcal{B}(\Omega_2)$,

$$\mu(B,s)(X \times Y) = B(s)(X,Y)$$

(see Theorem 1.10, p. 24, of Ref. 4). Putting $E(s)(Z) = \mu(B,s)(Z)$ for all $s \in S, Z \in \mathcal{B}(\Omega_1 \times \Omega_2)$, one defines a Radon observable $E:S \to RM_1^+(\mathcal{B}(\Omega_1 \times \Omega_2))$ with the property

$$\tilde{E}(X \times \Omega_2) = \tilde{B}(X, \Omega_2) = \tilde{E}_1(X),$$

$$\tilde{E}(\Omega_1 \times Y) = \tilde{B}(\Omega_1, Y) = \tilde{E}_2(Y),$$

for all $X \in \mathcal{B}_1, Y \in \mathcal{B}_2$.

$(b) \Rightarrow (c)$ Let now F be a Radon joint observable of E_1 and E_2, and let π_1 and π_2 be the coordinate projections $\Omega_1 \times \Omega_2 \to \Omega_1$, $\Omega_1 \times \Omega_2 \to \Omega_2$, respectively. Then $\tilde{E}_1(X) = \tilde{F}(\pi_1^{-1}(X))$ and $\tilde{E}_2(Y) = \tilde{F}(\pi_2^{-1}(Y))$, showing that E_1 and E_2 are functionally R-coexistent.

$(c)\Rightarrow(a)$ If $\tilde{E}_1(X)=\tilde{E}(f_1^{-1}(X))$ and $\tilde{E}_2(Y)=\tilde{E}(f_2^{-1}(Y))$, for some Radon observable $E:S\to M_1^+(\mathcal{B}(\Omega))$ and some Borel functions $f_i:\Omega\to\Omega_i$, $i=1,2$, then $\tilde{B}(X,Y)(s):=\tilde{E}(f_1^{-1}(X)\cap f_2^{-1}(Y))(s)$, $X\in\mathcal{B}(\Omega_1), Y\in\mathcal{B}(\Omega_2), s\in S$, defines a biobservable B with the desired properties.

Remark 2.1.6: In physical applications, the underlying spaces Ω,Ω_1,Ω_2, of observables $E:S\to M_1^+(\Omega,\mathcal{B})$ and biobservables $B:S\to BM_1^+(\Omega_1\times\Omega_2,\mathcal{B}_1\times\mathcal{B}_2)$ are often equipped with locally compact metrizable and separable topologies. The measures on the Borel σ-algebras of such spaces are necessarily Radon measures.[5] In such cases one may, therefore, just refer to observables, joint observables, and biobservables without any further qualification of their being Radon, e.g., in Theorem 2.1.5. In particular, this is the case for Ω being the real Borel space $(\mathbf{R},\mathcal{B}(\mathbf{R}))$, or its n-fold Cartesian product.

Remark 2.1.7: The framework of Theorem 2.1.5 is very general. In interesting cases this framework is further specified not only with respect to the measurable spaces $(\Omega,\mathcal{B})$ involved, cf. above, but also giving some further structure on the set of states, like σ-additivity (with respect to an appropriate topology). In such cases it is natural to restrict also the notion of observable further to those affine functions $E:S\to M_1^+(\Omega,\mathcal{B})$ which preserve this additional structure. In the next section we shall come to this question.

B. Observables on effect algebras

Definition 2.2.1: An *effect algebra* is a set L with two distinguished elements 0_L and 1_L, and with a partially defined binary operation $L\times L\ni(a,b)\mapsto a\oplus b\in L$ such that for any $a,b,c\in L$ the following conditions are satisfied, see, e.g., Ref. 6:

(EA1) $a\oplus b$ is defined, if and only if $b\oplus a$ is defined, and then $a\oplus b=b\oplus a$ (commutativity);
(EA2) $b\oplus c$, $a\oplus(b\oplus c)$ are defined, if and only if $a\oplus b$, $(a\oplus b)\oplus c$ are defined, and then $a\oplus(b\oplus c)=(a\oplus b)\oplus c$ (associativity);
(EA3) for every $a\in L$, there is a unique $b\in L$ such that $a\oplus b=1_L$ (orthocomplementation);
(EA4) $1_L\oplus a$ is defined, if and only if $a=0_L$ (zero-one law).

If the assumptions of (EA2) are satisfied, we write $a\oplus b\oplus c$ for the element $a\oplus(b\oplus c)=(a\oplus b)\oplus c$. We say that $a,b\in L$ are $\oplus$-*orthogonal* if $a\oplus b$ is defined. Occasionally we then write $a\perp b$. The element b in (EA3) for which $a\oplus b=1$ is called the *orthocomplement* of a, and we write $b=a'$.

A partial order $\leqslant$ can be defined on L as follows: $a\leqslant b$, if there is a $c\in L$ such that $c\perp a$ and $a\oplus c=b$. The element c is uniquely defined, since (EA3) enables one to prove the cancellation law: if $a\oplus c=b\oplus c$, then $a=b$. In this ordering, 0_L is the smallest and 1_L is the greatest element in L.

We say that a finite sequence $a_1,a_2,\ldots,a_n$ of elements of L is $\oplus$-*orthogonal*, if the element $a_1\oplus a_2\oplus\ldots\oplus a_n$ exists in L, where the latter element is defined recursively using (EA1) and (EA2). An arbitrary system $G=\{a_i\}_{i\in I}$ of (not necessarily different) elements of L is $\oplus$-orthogonal if every finite subsystem $\{a_i\}_{i\in F}$, $F\subset I$, of G is $\oplus$-orthogonal. A $\oplus$-orthogonal system G has $\oplus$-sum, if the least upper bound $\oplus G:=\vee_F(\oplus_{i\in F}a_i)$ over all finite $\oplus$-sums exists. An effect algebra L is called *σ-complete (complete)*, if $\oplus$-sums exist for any countable (arbitrary) system $\{a_i\}_{i\in I}$ of $\oplus$-orthogonal elements of L.

Example 2.2.2: Let $[0,1]$ be the unit interval of the real numbers. For any $a,b\in[0,1]$, define $a\oplus b:=a+b$, whenever $a+b\leqslant 1$. Then $([0,1],0,1,\oplus)$ is a complete effect algebra.

Example 2.2.3: Let $\mathcal{B}$ be a σ-algebra of a nonempty set Ω. $(\mathcal{B},\varnothing,\Omega,\oplus)$ is a σ-complete effect algebra, with the $\oplus$-operation defined as the disjoint set theoretical union: $X\oplus Y:=X\cup Y$, whenever $X\cap Y=\varnothing$.

Example 2.2.4: Let $\mathcal{F}$ be a set of functions from a nonempty set S to the unit interval $[0,1]$ such that (F1a) $1_S\in\mathcal{F}$, (F1b) if $f\in\mathcal{F}$, then also $1_S-f\in\mathcal{F}$, and (F2) if $f,g,h\in\mathcal{F}$, $f+g\in\mathcal{F}$, $f+g+h\in\mathcal{F}$, then also $g+h\in\mathcal{F}$. Defining $f\oplus g=f+g$, whenever $f+g\in\mathcal{F}$, one obtains that $(\mathcal{F},0_S,1_S,\oplus)$ is an effect algebra. If, in addition, the following condition is satisfied,

$$\text{(F3)}\quad f,\ g\in\mathcal{F},\quad f+g\leqslant 1_S\Rightarrow f+g\in\mathcal{F}$$

(with $f+g\leqslant 1_S$ referring to the pointwise ordering), then the partial order on $\mathcal{F}$, derived from the operation $\oplus$, agrees with the pointwise ordering of functions.[7] The above condition may also be strengthened to a σ-additive version:

(F3σ) if $(f_i)_{i\geqslant 1}$ is any countable sequence of functions in $\mathcal{F}$ such that $\Sigma_{i\geqslant 1} f_i \leqslant 1_S$, then $\Sigma_{i\geqslant 1} f_i$ belongs to $\mathcal{F}$.

Here, again, the convergence is the pointwise convergence. If this condition is also satisfied, then $(\mathcal{F},0_S,1_S,\oplus)$ is a σ-complete effect algebra.

Definition 2.2.5: (*a*) Let L and K be two effect algebras. A function $\Phi: L\to K$ is an *effect morphism,* if for all $a,b\in L$,

$$a\perp b \Rightarrow \Phi(a)\perp\Phi(b), \text{ and then } \Phi(a\oplus b)=\Phi(a)\oplus\Phi(b).$$

An effect morphism Φ is *unital* if $\Phi(1_L)=1_K$. An effect morphism is *σ-additive (completely additive)* if it preserves all the existing countable (arbitrary) $\oplus$-sums.

(*b*) A unital effect morphism s from an effect algebra L to the effect algebra $[0,1]$ is called a *state* on L. We let $S(L)$ denote the set of all states on L. A state $s\in S(L)$ is σ-additive (completely additive), if $s(\oplus_{i\in I} a_i)=\Sigma_{i\in I} s(a_i)$ whenever $\oplus_{i\in I} a_i$ exists, where I is a countable (arbitrary) index set.

(*c*) A set $S\subseteq S(L)$ of states on an effect algebra L is *separating,* if for any $a,b\in L$, if $a\neq b$, then $s(a)\neq s(b)$ for some $s\in S$. S is *ordering,* if the condition $s(a)\leqslant s(b)$, $s\in S$, implies $a\leqslant b$ for any $a,b\in L$.

Clearly, an ordering set of states is separating. We also note that the set of all states $S(L)$ on L is a convex set (as a subset of $[0,1]^L\subset \mathbf{R}^L$): if $s_1,s_2\in S(L)$ and $0\leqslant\lambda\leqslant 1$, then $\lambda s_1+(1-\lambda)s_2\in S(L)$ [where $\lambda s_1+(1-\lambda)s_2$ is defined pointwise]. Moreover, $S(L)$ is σ-convex with respect to the topology of pointwise convergence.

Let (L,S) be an effect algebra L with a nonempty set S of states on L. For every $a\in L$, define the function $\underline{a}: S\to[0,1]$, $\underline{a}(s):=s(a)$. Then $a\mapsto\underline{a}$ is a unital morphism from L to the effect algebra of functions from S to $[0,1]$ (that can be considered as an effect algebra in the sense of Example 2.2.4). Let $\mathcal{F}_L(S):=\{\underline{a}: a\in L\}$. Then $L\ni a\mapsto\underline{a}\in\mathcal{F}_L(S)$ is a surjective mapping. If the set S of states on L is separating, this mapping is also injective, so that L and $\mathcal{F}_L(S)$ are, in fact, isomorphic sets. In that case we can transfer the effect algebra structure from L to $\mathcal{F}_L(S)$ by defining $\underline{a}\perp\underline{b}$ whenever $a\perp b$, and putting then $\underline{a}\oplus\underline{b}=\underline{a\oplus b}=\underline{a}+\underline{b}$. However, the ordering on $\mathcal{F}_L(S)$ induced by $\oplus$ need not agree with the pointwise ordering of functions. In fact, these two orderings coincide if and only if S is an ordering set of states on L. Indeed, S is ordering if and only if $a\perp b\Leftrightarrow\underline{a}+\underline{b}\leqslant 1_S$. Moreover, the effect algebra L is σ-complete and the states in the ordering set S are σ-additive, if and only if $\mathcal{F}_L(S)$ satisfies (F3σ). We also note that if S is a convex set of states on L, then $\mathcal{F}_L(S)$ is a subset of the set $A_1^+(S)$ of affine functions $S\to[0,1]$.

Let L be a σ-complete effect algebra, and let $(\Omega,\mathcal{B})$ be a measurable space. Let $A:\mathcal{B}\to L$ be a unital σ-additive morphism. If s is a σ-additive state on L, then $s\circ A$ is a probability measure on $(\Omega,\mathcal{B})$. Therefore, if S is a convex set of σ-additive states on L, then $S\ni s\mapsto s\circ A\in M_1^+(\Omega,\mathcal{B})$ is an observable. The question arises under which conditions an observable $E: S\to M_1^+(\Omega,\mathcal{B})$ comes in this way from a σ-additive morphism $\mathcal{B}\to L$. For a given observable E, we let $\mathrm{ran}(\tilde{E}):=\{\tilde{E}(X)\,|\,X\in\mathcal{B}\}$ denote the range of $\tilde{E}$. It is a subset of $A_1^+(S)$.

Definition 2.2.6: Let (L,S) be an effect algebra L with a convex separating set S of σ-additive states.

(a) The pair (L,S) is said to be *total with respect to a measurable space* $(\Omega,\mathcal{B})$ if, for each observable $E: S\to M_1^+(\Omega,\mathcal{B})$, $\mathrm{ran}\tilde{E}\subseteq\mathcal{F}_L(S)$.
(b) The pair (L,S) is said to be *total* if $\mathcal{F}_L(S)=A_1^+(S)$.

Proposition 2.2.7: Let (L,S) be an effect algebra with a convex separating set of σ-additive states.

(i) If (L,S) is total, then S is ordering.
(ii) If (L,S) is total, then it is total with respect to every measurable space $(\Omega,\mathcal{B})$.

Proof: (i): Since S is separating, L is isomorphic to $\mathcal{F}_L(S)$. The totality of (L,S) implies that $\mathcal{F}_L(S)=A_1^+(S)$. But then S is ordering on L. The other statement is immediate.

Let (L,S) be a total effect algebra with a convex separating set of σ-additive states. If $E:S\to M_1^+(\Omega,\mathcal{B})$ is an observable, we get a map $\hat{E}:\mathcal{B}\to L$ by requiring

$$\underline{\hat{E}(X)}=\tilde{E}(X),$$

for all $X\in\mathcal{B}$, that is, by putting

$$s(\hat{E}(X))=\tilde{E}(X)(s)=E(s)(X),$$

for all $s\in S, X\in\mathcal{B}$. Clearly, $\hat{E}$ is a unital σ-additive morphism $\mathcal{B}\to L$. Indeed, $\underline{\hat{E}(\Omega)}=1_S$, so that $\hat{E}(\Omega)=1_L$, and, if $(X_i)_{i\geqslant 1}$ is a disjoint sequence, then, for any $s\in S$, $s(\hat{E}(\cup X_i))=\tilde{E}(\cup X_i)(s)=\Sigma\tilde{E}(X_i)(s)=\Sigma s(\hat{E}(X_i))=s(\oplus\hat{E}(X_i))$, showing that $\hat{E}(\oplus_{i\geqslant 1}X_i)=\oplus_{i\geqslant 1}\hat{E}(X_i)$.

Corollary 2.2.8: Let (L,S) be a total effect algebra with a convex separating set of σ-additive states. For any measurable space $(\Omega,\mathcal{B})$, the map $E\mapsto\hat{E}$ is an affine bijection from the convex set of observables $E:S\to M_1^+(\Omega,\mathcal{B})$ onto the convex set of unital σ-additive morphisms $A:\mathcal{B}\to L$.

For a total effect algebra L with a convex separating set S of σ-additive states the notion of an observable as an affine function $E:S\to M_1^+(\Omega,\mathcal{B})$ has two alternative equivalent formulations, a σ-additive function $F:\mathcal{B}\to A_1^+(S)$, with the property $F(\Omega)=1_S$, and a unital σ-additive morphism $A:\mathcal{B}\to L$, and we have $s(A(X))=F(X)(s)=E(s)(X)$ for all $s\in S, X\in\mathcal{B}$. The notions of coexistence and functional coexistence of any two observables $E_1:S\to M_1^+(\Omega_1,\mathcal{B}_1)$ and $E_2:S\to M_1^+(\Omega_2,\mathcal{B}_2)$ may then also be given in terms of the morphisms $\hat{E}_1:\mathcal{B}_1\to L$ and $\hat{E}_2:\mathcal{B}_2\to L$. In particular, Theorem 2.1.5 may then be read as a characterization of functional coexistence of such morphisms.

C. Observables in the Hilbert space framework

Let $\mathcal{H}$ be a complex separable Hilbert space, and let $\mathcal{T}(\mathcal{H})_1^+$ denote the set of positive trace one operators on $\mathcal{H}$. It is a convex set as a subset of the real vector space $\mathcal{T}_s(\mathcal{H})$ of the self-adjoint trace class operators on $\mathcal{H}$. Let $(\Omega,\mathcal{B})$ be a measurable space, and consider an observable $E:\mathcal{T}(H)_1^+\to M_1^+(\Omega,\mathcal{B})$. Since $\mathcal{T}(\mathcal{H})_1^+$ is a base for the positive cone $\mathcal{T}(\mathcal{H})^+$ of the vector space $\mathcal{T}_s(\mathcal{H})$, any $f\in A_1^+(\mathcal{T}(\mathcal{H})_1^+)$ extends uniquely to a positive linear functional on $\mathcal{T}_s(\mathcal{H})$. Such a functional is known to be trace-norm continuous,[8] so that, by the duality of the trace class operators and the bounded operators on $\mathcal{H}$, for each $f\in A_1^+(\mathcal{T}(\mathcal{H})_1^+)$ there is a unique positive operator $A\geqslant O$, bounded by the unit operator, $A\leqslant I$, such that $f(T)=\mathrm{tr}[TA]$ for all $T\in\mathcal{T}(\mathcal{H})_1^+$. This shows that for each $X\in\mathcal{B}$, there is a unique $O\leqslant A(X)\leqslant I$ such that $E(T)(X)=\tilde{E}(X)(T)=\mathrm{tr}[TA(X)]$ for all $T\in\mathcal{T}(\mathcal{H})_1^+$. Moreover, if $(X_i)\subseteq\mathcal{B}$ is a disjoint sequence, then, for each $T\in\mathcal{T}(\mathcal{H})_1^+$, $\mathrm{tr}[TA(\cup X_i)]=\tilde{E}(\cup X_i)(T)=\Sigma\tilde{E}(X_i)(T)=\Sigma\ \mathrm{tr}[TA(X_i)]$, showing that the operator-valued function $X\mapsto A(X)$ is σ-additive in the weak operator topology of $\mathcal{L}(\mathcal{H})$, the set of bounded operators on $\mathcal{H}$. Therefore, an observable $E:\mathcal{T}(\mathcal{H})_1^+\to M_1^+(\Omega,\mathcal{B})$ determines a unique normalized $(A(\Omega)=I)$ positive operator-valued measure $\mathcal{B}\ni X\mapsto A(X)\in\mathcal{L}(\mathcal{H})$. Clearly, any such measure $X\mapsto A(X)$ determines an observable $T\mapsto E(T)$ via the formula $E(T)(X):=\mathrm{tr}[TA(X)]$.

Let $\mathcal{E}(\mathcal{H}):=\{A\in\mathcal{L}(\mathcal{H})|O\leqslant A\leqslant I\}$. We equip $\mathcal{E}(\mathcal{H})$ with the $\oplus$-operation defined as $A\oplus B=A+B$, whenever $A+B\leqslant I$. Clearly, the order defined by $\oplus$ on $\mathcal{E}(\mathcal{H})$ coincides with the usual ordering of $\mathcal{E}(\mathcal{H})$. Since every upward directed family of self-adjoint operators which is bounded from above by a self-adjoint operator has a supremum, $\mathcal{E}(\mathcal{H})$ is a complete effect algebra. However, it is not a lattice, unless $\mathcal{H}$ is one-dimensional.

The set $\mathcal{E}(\mathcal{H})$ is also a convex set [as a subset of $\mathcal{L}(\mathcal{H})$], and the set of states on $\mathcal{E}(\mathcal{H})$ in the sense of Definition 2.2.5 equals now the set of affine functions on $\mathcal{E}(\mathcal{H})$, that is, $S(\mathcal{E}(\mathcal{H}))=A_1^+(\mathcal{E}(\mathcal{H}))$. As is well known, the subset of σ-additive states on $\mathcal{E}(\mathcal{H})$ is in one-to-one correspondence with the set $\mathcal{T}(\mathcal{H})_1^+$ of positive trace one operators on $\mathcal{H}$. Indeed, if $s:\mathcal{E}(\mathcal{H})\to[0,1]$ is a unital σ-additive morphism, then there is a unique $T\in\mathcal{T}(\mathcal{H})_1^+$ such that for all $A\in\mathcal{E}(\mathcal{H})$, $s(A)=\mathrm{tr}[TA]$, and conversely, any $T\in\mathcal{T}(\mathcal{H})_1^+$ defines, by this formula, a σ-additive state on $\mathcal{E}(\mathcal{H})$, see, e.g., Refs. 1 and 8. The set $\mathcal{T}(\mathcal{H})_1^+$ is an ordering set of states on $\mathcal{E}(\mathcal{H})$. Moreover, the

pair $(\mathcal{E}(\mathcal{H}),\mathcal{T}(\mathcal{H})_1^+)$ is total. Indeed, as already pointed out above, for each $f\in A_1^+(\mathcal{T}(\mathcal{H})_1^+)$ there is a unique $A\in\mathcal{E}(\mathcal{H})$ such that $f(T)=\mathrm{tr}[TA]$ for all $T\in\mathcal{T}(\mathcal{H})_1^+$, showing that $\mathcal{F}_{\mathcal{E}(\mathcal{H})}(\mathcal{T}(\mathcal{H})_1^+)=A_1^+(\mathcal{T}(\mathcal{H})_1^+)$. This means, in particular, that any observable $E:\mathcal{T}(\mathcal{H})_1^+\to M_1^+(\Omega,\mathcal{B})$ corresponds to a unital σ-additive morphism $A:\mathcal{B}\to\mathcal{E}(\mathcal{H})$. Such a morphism is a normalized positive operator-valued measure.

We conclude that Theorem 2.1.5 serves, in this context, to characterize the functional coexistence of observables represented as normalized positive operator-valued measures. Such a characterization was obtained also in Ref. 9 using different methods and results from Ref. 10.

III. COEXISTENT EFFECTS

Consider, again, an abstract convex set S and let $A_1^+(S)$ be the set of affine functions on S. We call these functions *effects*. If $E:S\to M_1^+(\Omega,\mathcal{B})$ is an observable, then the range $\mathrm{ran}(\tilde{E})$ of $\tilde{E}$ is a subset of $A_1^+(S)$.

Definition 3.1: A set of effects $\mathcal{A}\subseteq A_1^+(S)$ is called *coexistent* if there is a measurable space $(\Omega,\mathcal{B})$ with an observable $E:S\to M_1^+(\Omega,\mathcal{B})$ such that $\mathcal{A}\subset\mathrm{ran}(\tilde{E})$.

The next theorem characterizes coexistent sets of effects. For that we need to introduce the following notions.

Definition 3.2: (*a*) Let $(A,\leqslant)$ be a directed set and Ω_α a finite set for each $\alpha\in A$. Whenever $\alpha,\beta\in A$ and $\alpha\leqslant\beta$, let there be given a mapping $g_{\alpha\beta}:\Omega_\beta\to\Omega_\alpha$, and denote $\mathcal{G}=(g_{\alpha,\beta})_{\alpha\leqslant\beta}$. The pair $((\Omega_\alpha)_{\alpha\in A},\mathcal{G})$ is called a *projective system of finite sets* if the following conditions hold:

(*i*) $g_{\alpha\alpha}$ is the identity map on Ω_α for all $\alpha\in A$;
(*ii*) $g_{\alpha\beta}\circ g_{\beta\gamma}=g_{\alpha\gamma}$ whenever $\alpha\leqslant\beta\leqslant\gamma$.

(*b*) We say that $((\Omega_\alpha,P_\alpha)_{\alpha\in A},\mathcal{G})$ is a *projective system of simple probability spaces* if $((\Omega_\alpha)_{\alpha\in A},\mathcal{G})$ is a projective system of finite sets and, for each $\alpha\in A$, $P_\alpha:2^{\Omega_\alpha}\to[0,1]$ is a probability measure such that the following compatibility condition holds:

(*iii*) $P_\alpha(X)=P_\beta(g_{\alpha\beta}^{-1}(X))$, for all $X\subseteq 2^{\Omega_\alpha}$, whenever $\alpha\leqslant\beta$.

(*c*) We say that $((\Omega_\alpha,E_\alpha)_{\alpha\in A},\mathcal{G})$ is a *projective system of simple observables* if $((\Omega_\alpha)_{\alpha\in A},\mathcal{G})$ is a projective system of finite sets and, for each $\alpha\in A$, $E_\alpha:S\to M_1^+(\Omega_\alpha,2^{\Omega_\alpha})$ is an observable such that the following compatibility condition holds:

(*iv*) $E_\alpha(s)(X)=E_\beta(s)(g_{\alpha\beta}^{-1}(X))$, for all $X\subseteq 2^{\Omega_\alpha}$, $s\in S$, whenever $\alpha\leqslant\beta$.

Theorem 3.3: For $\mathcal{A}\subseteq A_1^+(S)$ the following conditions are equivalent:

(*i*) there is a compact Hausdorff space Ω with a Radon observable $E:S\to M_1^+(\mathcal{B}(\Omega))$ such that $\mathcal{A}\subseteq\mathrm{ran}(\tilde{E})$;
(*ii*) there is a measurable space $(\Omega,\mathcal{B})$ with an observable $E:S\to M_1^+(\Omega,\mathcal{B})$ such that $\mathcal{A}\subseteq\mathrm{ran}(\tilde{E})$;
(*iii*) there is a projective system $((\Omega_\alpha,E_\alpha)_{\alpha\in A},\mathcal{G})$ of simple observables such that for each $a\in\mathcal{A}$ there is an $\alpha\in A$ such that $a\in\mathrm{ran}(\tilde{E}_\alpha)$.

Proof: The implication $(i)\Rightarrow(ii)$ is trivial. Assume (*ii*), and let $E:S\to M_1^+(\Omega,\mathcal{B})$ be an observable such that $\mathcal{A}\subset\mathrm{ran}(\tilde{E})$. Let $\mathcal{R}$ be the set of all finite disjoint measurable partitions of Ω. For each $R=\{X_i\}_{i\leqslant n}\in\mathcal{R}$, the function $R\ni X_i\mapsto\tilde{E}(X_i)=:\tilde{E}^R(\{X_i\})$ corresponds uniquely (Theorem 2.1.1) to a simple observable $S\ni s\mapsto E^R(s)\in M_1^+(R,2^R)$, where R is the finite set $\{X_1,\ldots,X_n\}$. Let $R_1,R_2\in\mathcal{R}$, and assume that R_1 is a refinement of R_2, that is, for every $X_i\in R_1$, there is (exactly one) $Y_j\in R_2$, such that $X_i\subseteq Y_j$. We then write $R_2\leqslant R_1$. For $R_2\leqslant R_1$, the function $f_{R_2R_1}:R_1\to R_2, X_i\mapsto Y_j$ (with $X_i\subseteq Y_j$) is then uniquely defined. If $R_2=R_1$, this function is the identity. If $R_2\leqslant R_1$ and $R_3\leqslant R_2$, then also $R_3\leqslant R_1$, and $f_{R_3R_1}=f_{R_3R_2}\circ f_{R_2R_1}$. Finally, for any two partitions $R_1,R_2\in\mathcal{R}$, the partition $R=\{X_i\cap Y_j: X_i\in R_1, X_j\in R_2\}$ belongs to $\mathcal{R}$, and it is a common refinement of R_1 and R_2, that is, $R_1\leqslant R$, $R_2\leqslant R$. Let $\mathcal{G}:=\{f_{RS}|R,S\in\mathcal{R},R\leqslant S\}$. It is now immediate to check that $((R,E^R)_{R\in\mathcal{R}},\mathcal{G})$ is a projective system of simple observables. By con-

struction it also covers $\mathcal{A}$, that is, $\mathcal{A}\subseteq\cup_{R\in\mathcal{R}}\mathrm{ran}(\tilde{E}^R)$. This shows that $(ii)\Rightarrow(iii)$. Finally, assume (iii). Using Proposition 4 in Ref. 11, p. 120, we see that the projective limit $\Omega=\lim_{\leftarrow}(\Omega_\alpha,g_{\alpha\beta})$ is a compact Hausdorff space, and the projective system $((\Omega_\alpha)_{\alpha\in A},\mathcal{G})$ satisfies Bochner's sequential maximality condition. Let $\pi_\alpha:\Omega\to\Omega_\alpha$ be the natural coordinate projection, $\alpha\in A$. Bochner's generalization of the Kolmogorov consistency theorem (see, e.g., Theorem 5 in Ref. 11, p. 121, or Ref. 12, p. 120) yields for each $s\in S$ a Radon probability measure $P(s):\mathcal{B}(\Omega)\to[0,1]$ such that $P(s)(\pi_\alpha^{-1}(X))=E_\alpha(s)(X)$. Since the projective limit is unique and the function $s\mapsto P(s)$ is affine coordinatewise, that is, for each $\alpha\in A$ the map $s\mapsto E_\alpha(s)$ is affine, we conclude that $P:S\to M_1^+(\mathcal{B}(\Omega))$ is an observable satisfying (i). This completes the proof.

Remark 3.4: Let (L,S) be a total effect algebra L with a convex separating set S of σ-additive states. Since now $L\simeq\mathcal{F}_L(S)=A_1^+(S)$, the above theorem allows one to characterize coexistent subsets of L. As another application, taking $S=\mathcal{T}(\mathcal{H})_1^+$, one obtains a characterization of the coexistent subsets of the Hilbert space effect algebra $\mathcal{E}(\mathcal{H})$. In Ref. 9 an alternative characterization was obtained.

ACKNOWLEDGMENT

We are indebted to an anonymous referee of Ref. 9 for valuable comments which, although dealing with a special case, influenced some ingredients in the proofs of Theorems 2.1.5 and 3.3.

[1] G. Ludwig, *Foundations of Quantum Mechanics I* (Springer, Berlin, 1983).
[2] P. Busch, M. Grabowski, and P. Lahti, *Operational Quantum Physics*, **LNP m 31** (Springer, 1995, corrected printing 1997).
[3] S. P. Gudder, *Stochastic Methods in Quantum Mechanics* (North-Holland, Amsterdam, 1979).
[4] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups* (Springer, Berlin, 1984).
[5] P. Halmos, *Measure Theory* (Springer, Berlin, 1974).
[6] D. J. Foulis and M. K. Bennett, ''Effect algebras and unsharp quantum logics,'' Found. Phys. **24**, 1331–1352 (1994).
[7] S. Pulmannova, ''A remark to orthomodular partial algebras,'' *Demonstratio Mathematica* **27**, 687–699 (1994).
[8] E. B. Davies, *The Quantum Theory of Open Systems* (Academic, New York, 1976).
[9] P. Lahti and S. Pulmannova, ''Coexistent observables and effects in quantum mechanics,'' Rep. Math. Phys. **39**, 339–351 (1997).
[10] K. Ylinen, ''Positive operator bimeasures and a noncommutative generalization,'' *Studia Mathematica* **118**, 157–168 (1996).
[11] M. M. Rao, *Foundations of Stochastic Analysis* (Academic, New York, 1981).
[12] S. Bochner, *Harmonic Analysis and the Theory of Probability* (Univ. of California, Berkeley and Los Angeles, 1955).

Strict quantization of coadjoint orbits

N. P. Landsman[a)]
Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

(Received 8 June 1998; accepted for publication 3 August 1998)

A strict quantization of a compact symplectic manifold S on a subset $I\subseteq\mathbb{R}$, containing 0 as an accumulation point, is defined as a continuous field of C^*-algebras $\{\mathfrak{A}_\hbar\}_{\hbar\in I}$, with $\mathfrak{A}_0=C_0(S)$, and a set of continuous cross sections $\{\mathcal{Q}(f)\}_{f\in C^\infty(S)}$ for which $\mathcal{Q}_0(f)=f$. Here $\mathcal{Q}_\hbar(f^*)=\mathcal{Q}_\hbar(f)^*$ for all $\hbar\in I$, whereas for $\hbar\to 0$ one requires that $i[\mathcal{Q}_\hbar(f),\mathcal{Q}_\hbar(g)]/\hbar\to\mathcal{Q}_\hbar(\{f,g\})$ in norm. We discuss general conditions which guarantee that a (deformation) quantization in a more physical sense leads to one in the above sense. Using ideas of Berezin, Lieb, Simon, and others, we construct a strict quantization of an arbitrary integral coadjoint orbit $\mathcal{O}_\lambda$ of a compact connected Lie group G, associated to a highest weight λ. Here $I=0\cup 1/\mathbb{N}$, so that $\hbar=1/k$, $k\in\mathbb{N}$, and $\mathfrak{A}_{1/k}$ is defined as the C^*-algebra of all matrices on the finite-dimensional Hilbert space $\mathcal{V}_{k\lambda}$ carrying the irreducible representation $U_{k\lambda}(G)$ with highest weight $k\lambda$. The quantization maps $\mathcal{Q}_{1/k}(f)$ are constructed from coherent states in $\mathcal{V}_{k\lambda}$, and have the special feature of being positive maps. © *1998 American Institute of Physics.* [S0022-2488(98)00512-X]

I. INTRODUCTION

The aim of this paper is to combine a number of what appear to the author to be good ideas in mathematical physics, whose interplay has so far not sufficiently been studied. First, there now exists a satisfying C^*-algebraic definition of quantization, which enables one to link Poisson and symplectic geometry[1,2] with noncommutative geometry.[3] In particular, the geometric theory of classical mechanics and reduction is thereby related to the C^*-algebraic formulation of quantum mechanics and induction.[4] The main mathematical idea of this definition goes back to Rieffel,[5] who showed how deformation quantization makes sense in an operator-algebraic context. The physical postulates, of course, may be traced back to Dirac. In this definition, like in conventional (''formal'') deformation quantization, it is crucial that one studies the quantum theory for a family of values of Planck's constant $\hbar$.[6]

Second, one has the well-known connection between coadjoint orbits of certain Lie groups, unitary representations, and geometric quantization.[7] This connection works particularly well for either compact or exponential nilpotent Lie groups. Here one keeps $\hbar$ fixed.

Our third source of inspiration is the work of Lieb,[8] Simon,[9] and others on the classical limit of quantum spin systems and their generalizations to arbitrary compact Lie groups. This work is closely related to Perelomov's coherent states,[10,11] as well as to Berezin's approach to quantization and the classical limit.[6]

Historical comments and extensive references concerning these ideas may be found in Ref. 4.

Using the second and third groups of ideas at both a conceptual and a technical level, we will construct a C^*-algebraic quantization of an arbitrary integral coadjoint orbit of a compact connected Lie group. A different approach to this problem has recently been considered in Ref. 12; see also Ref. 13.

Section II contains a general C^*-algebraic definition of quantization. A quantization satisfying this definition is called ''strict.'' We present criteria on a given quantization which guarantee that the postulates in the definition are met. It is on this basis that we will actually construct the quantizations in this paper. Section III describes our approach to coherent states and Berezin quantization, and develops conditions under which the Berezin quantization constructed from a

a)Electronic mail: npl@wins.uva.nl

 © 1998 American Institute of Physics

family of coherent states is strict. This is done in terms of a so-called pure state quantization, which, in a heuristic sense, is dual to a C^*-algebraic quantization. The material in Secs. II and III is model independent, and should be relevant to quantization theory in general.

Section IV is a brief review of Perelomov coherent states, coadjoint orbits of compact Lie groups, the momentum map, and the connection between these concepts. The material in this section is not new, but is worth summarizing in preparation for our main results. Section V contains the two principal theorems of this paper. The first states that Perelomov's coherent states for compact connected Lie groups define pure state quantizations of particular coadjoint orbits. The second states that such pure state quantizations lead to strict quantizations of the coadjoint orbits in question.

II. ON QUANTIZATION

The central notion in C^*-algebraic quantization theory is that of a continuous field of C^*-algebras.[14] For our purposes the following reformulation is useful.[15]

*Definition 1: A continuous field of C^*-algebras $(\mathfrak{C},\{\mathfrak{A}_x,\varphi_x\}_{x\in X})$ over a locally compact Hausdorff space X consists of a C^*-algebra $\mathfrak{C}$, a collection of C^*-algebras $\{\mathfrak{A}_x\}_{x\in X}$, and a set $\{\varphi_x:\mathfrak{C}\to\mathfrak{A}_x\}_{x\in X}$ of surjective *-homomorphisms, such that for all $A\in\mathfrak{C}$:*

(1) *the function $x\to\|\varphi_x(A)\|$ is in $C_0(X)$;*

(2) *one has $\|A\|=\sup_{x\in X}\|\varphi_x(A)\|$;*

(3) *for any $f\in C_0(X)$ there is an element $fA\in\mathfrak{C}$ for which $\varphi_x(fA)=f(x)\varphi_x(A)$ for all $x\in X$.*

Here $C_0(X)$ is the space of continuous functions on X which vanish at infinity. Note that it is sufficient that each φ_x has dense image, since the image of a *-homomorphism between C^*-algebras is automatically closed. The continuous cross sections of the field in the sense of Ref. 14 consist of those elements $\{A_x\}_{x\in X}$ of $\Pi_{x\in X}\mathfrak{A}_x$ for which there is a $A\in\mathfrak{C}$ such that $A_x=\varphi_x(A)$ for all $x\in X$.

The following definition of quantization, which is a slight reformulation of a definition in Ref. 16, seems to combine the best of the previous definitions in this direction in Refs. 5 and 17.

Definition 2: Let $I\subseteq\mathbb{R}$ contain 0 as an accumulation point. A strict quantization of a Poisson manifold P on I consists of

(1) *a continuous field of C^*-algebras $(\mathfrak{C},\{\mathfrak{A}_\hbar,\varphi_\hbar\}_{\hbar\in I})$ over I, with $\mathfrak{A}_0=C_0(P)$;*

(2) *a dense *-subalgebra $\tilde{\mathfrak{A}}_0$ of $C_0(P)$ on which the Poisson bracket is defined, and which is closed under taking Poisson brackets (so that $\tilde{\mathfrak{A}}_0$ is a complex Poisson algebra);*

(3) *a linear map $\mathcal{Q}:\tilde{\mathfrak{A}}_0\to\mathfrak{C}$ which (with $\mathcal{Q}_\hbar(f)\equiv\varphi_\hbar(\mathcal{Q}(f))$) for all $f\in\tilde{\mathfrak{A}}_0$ and $\hbar\in I$ satisfies*

$$\mathcal{Q}_0(f)=f, \tag{1}$$

$$\mathcal{Q}_\hbar(f^*)=\mathcal{Q}_\hbar(f)^*, \tag{2}$$

and for all $f,g\in\tilde{\mathfrak{A}}_0$ satisfies Dirac's condition

$$\lim_{\hbar\to 0}\left\|\frac{i}{\hbar}[\mathcal{Q}_\hbar(f),\mathcal{Q}_\hbar(g)]-\mathcal{Q}_\hbar(\{f,g\})\right\|=0. \tag{3}$$

Elements of I are interpreted as possible values of Planck's constant $\hbar$, and $\mathfrak{A}_\hbar$ is the quantum algebra of observables of the theory at the given value of $\hbar\neq 0$. For real-valued f, the operator $\mathcal{Q}_\hbar(f)$ is the quantum observable associated with the classical observable f. This interpretation is possible because of condition (2) in Definition 2.

In the examples of the present paper, the Poisson manifold P will be a compact symplectic manifold S, and we will choose $\tilde{\mathfrak{A}}_0=C^\infty(S)$. It will be a special feature that $\mathcal{Q}$ is defined on all of $\mathfrak{A}_0$; condition (3) of course makes sense on $\tilde{\mathfrak{A}}_0$ only.

The connection between Definition 2 and the more physically and historically oriented definition of quantization proposed in Refs. 17 and 4 is established in Proposition 1 below. This

requires two lemmas, the first of which, as a bonus, entails the equivalence between the definition of a continuous field given above, and the one given by Dixmier[14] (restricted to the case that the base space is locally compact).

Lemma 1: The C-algebra $\mathfrak{C}$ of (sections of) a continuous field is locally uniformly closed. That is, if $A \in \Pi_x \mathfrak{A}_x$ is such that for every $y \in X$ and every $\epsilon > 0$ there exists a $B^y \in \mathfrak{C}$ and a neighborhood $\mathcal{N}^y$ of y in which $\|A_x - B^y_x\| < \epsilon$ for all $x \in \mathcal{N}^y$, and also $\lim_{x\to\infty}\|A_x\| = 0$, then $A \in \mathfrak{C}$.*

Alternatively, if the function $x \to \|A_x - C_x\|$ lies in $C_0(X)$ for each $C \in \mathfrak{C}$, then $A \in \mathfrak{C}$.

In the situation of the first part, there is a compact set $K \subseteq X$ for which $\|A_x\| < \epsilon$ outside K, as well as a finite subcover $\{\mathcal{N}^{x_1}, \ldots \mathcal{N}^{x_n}\}$ of K. Taking a partition of unity $\{u_i\}$ on K subordinate to this subcover, the operator $B = \Sigma_i u_i B^{x_i}$ lies in $\mathfrak{C}$ because of Definition 1.3, and satisfies $\sup_{x\in X}\|A_x - B_x\| < \epsilon$. Hence $A \in \mathfrak{C}$ by Definition 1.2 and the completeness of $\mathfrak{C}$.

Given any $A \in \Pi_x \mathfrak{A}_x$ and $y \in X$, because φ_y is surjective there is a $B^y \in \mathfrak{C}$ so that $A_y = B^y_y$. The assumption in the second part of the lemma then implies that the conditions in the first part are satisfied, so that $A \in \mathfrak{C}$. ■

The next lemma adapts Propositions 10.2.3 and 10.3.2 in Ref. 14, which relate to Dixmier's own definition of a continuous field, to Definition 1.

Lemma 2: Suppose one has a family $\{\mathfrak{A}_x\}_{x\in X}$ of C-algebras indexed by a locally compact Hausdorff space X, as well as a subset $\tilde{\mathfrak{C}} \subseteq \Pi_x \mathfrak{A}_x$ which satisfies the following conditions:*

(1) *the set $\{A_x \mid A \in \tilde{\mathfrak{C}}\}$ is dense in $\mathfrak{A}_x$ for each $x \in X$;*
(2) *the function $x \to \|A_x\|$ is in $C_0(X)$ for each $A \in \tilde{\mathfrak{C}}$;*
(3) *the set $\tilde{\mathfrak{C}}$ is closed under pointwise scalar multiplication, addition, adjointing, and operator multiplication.*

There exists a unique continuous field of C-algebras $(\mathfrak{C}, \{\mathfrak{A}_x, \varphi_x\}_{x\in X})$ whose collection of continuous cross sections contains $\tilde{\mathfrak{C}}$. First, as a set $\mathfrak{C}$ consists of all $A \in \Pi_x \mathfrak{A}_x$ for which the function $x \to \|A_x - C_x\|$ lies in $C_0(X)$ for each $C \in \tilde{\mathfrak{C}}$. This set is regarded as a C*-algebra under the pointwise operations listed in item (3) above, and the norm defined in Definition 1.2. Second, $\varphi_x(A) = A_x$ is the evaluation map.*

We first show that if $A \in \Pi_x \mathfrak{A}_x$ is such that for every $x_0 \in X$ and every $\epsilon > 0$ there exists a $B \in \mathfrak{C}$ and a neighborhood $\mathcal{N}$ of x_0 such that $\|A_x - B_x\| < \epsilon$ for all $x \in \mathcal{N}$, and also $\lim_{x\to\infty}\|A_x\| = 0$, then $A \in \mathfrak{C}$. Indeed, take $C \in \tilde{\mathfrak{C}}$ arbitrary, and define the functions $f_{AC}: x \to \|A_x - C_x\|$ and $f_{BC}: x \to \|B_x - C_x\|$. Using the inequality

$$|(\|X\| - \|Y\|)| \leqslant \|X - Y\|, \tag{4}$$

one finds $|f_{AC}(x) - f_{BC}(x)| < \epsilon$ for all $x \in \mathcal{N}$. By assumption, f_{BC} is continuous, so that $|f_{BC}(x) - f_{BC}(x_0)| < \epsilon$ for all x in some neighborhood $\mathcal{N}'$ of x_0. Combining the two inequalities yields $|f_{AC}(x) - f_{AC}(x_0)| < 3\epsilon$ for all $x \in \mathcal{N} \cap \mathcal{N}'$. Hence f_{AC} is continuous at x_0, which was arbitrary, so that $A \in \mathfrak{C}$ by definition of $\mathfrak{C}$.

Using this property, it is easily shown that $\mathfrak{C}$ is a C*-algebra, and that condition 3 in Definition 1 is satisfied. It is clear from Definition 1.1 and the definition of $\mathfrak{C}$ that $\mathfrak{C}$ is maximal. On the other hand, according to the second part of Lemma 1, $\mathfrak{C}$ is minimal, so that it is unique.

Adding an assumption satisfied by the examples in this paper, we are now in a position to relate Definitions 1 and 2.

Proposition 1: Suppose one has a Poisson manifold P and a family $\{\mathfrak{A}_\hbar\}_{\hbar\in I}$ of C-algebras, with $\tilde{\mathfrak{A}}_0 \subset \mathfrak{A}_0 = C_0(P)$ as in Definition 2, and a collection of linear maps $\{\mathcal{Q}_\hbar : \tilde{\mathfrak{A}}_0 \to \mathfrak{A}_\hbar\}_{\hbar\in I}$ satisfying (1), (2), (3), as well as*

$$\lim_{\hbar\to 0} \| \tfrac{1}{2}(\mathcal{Q}_\hbar(f)\mathcal{Q}_\hbar(g) + \mathcal{Q}_\hbar(g)\mathcal{Q}_\hbar(f)) - \mathcal{Q}_\hbar(fg)\| = 0; \tag{5}$$

$$\lim_{\hbar\to 0} \|\mathcal{Q}_\hbar(f)\| = \|f\|_\infty \tag{6}$$

for all $f,g\in\tilde{\mathfrak{A}}_0$, *and finally the completeness condition that the collection* $\{\mathcal{Q}_\hbar(f)\mid f\in\tilde{\mathfrak{A}}_0\}$ *be dense in* $\mathfrak{A}_\hbar$ *for each* $\hbar\in I$. *Furthermore, assume that* $\hbar$ *assumes discrete values.*

There exists a unique continuous field of C-algebras* $(\mathfrak{C},\{\mathfrak{A}_\hbar,\varphi_\hbar\}_{\hbar\in I})$ *whose collection of continuous cross sections* $\{\varphi_\hbar(A)\}_{\hbar\in I}$, $A\in\mathfrak{C}$, *contains all maps* $\{\mathcal{Q}_\hbar(f)\}_{\hbar\in I}$, $f\in\tilde{\mathfrak{A}}_0$.

One defines $\tilde{\mathfrak{C}}\subset\Pi_\hbar\mathfrak{A}_\hbar$ as the complex linear span of all expressions of the form $\hbar\to\mathcal{Q}_\hbar(f_1)\ldots\mathcal{Q}_\hbar(f_n)$, where $f_i\in\tilde{\mathfrak{A}}_0$.

Given that I is discrete, the continuity of each function $F_{f_1,\ldots,f_n}:\hbar\to\|\mathcal{Q}_\hbar(f_1)\ldots\mathcal{Q}_\hbar(f_n)\|$ away from 0 is trivial. Now note that (5) and (3) imply that

$$\lim_{\hbar\to0}\|\mathcal{Q}_\hbar(f)\mathcal{Q}_\hbar(g)-\mathcal{Q}_\hbar(fg)\|=0 \tag{7}$$

for all $f,g\in\tilde{\mathfrak{A}}_0$. Hence $\lim_{\hbar\to0}\|\mathcal{Q}_\hbar(f_1)\mathcal{Q}_\hbar(f_2\cdots f_n)-\mathcal{Q}_\hbar(f_1\cdots f_n)\|=0$, and by induction $\lim_{\hbar\to0}\|\mathcal{Q}_\hbar(f_1)\cdots\mathcal{Q}_\hbar(f_n)-\mathcal{Q}_\hbar(f_1\cdots f_n)\|=0$. Equation (4) then yields $\lim_{\hbar\to0}\|\mathcal{Q}_\hbar(f_1)\cdots\mathcal{Q}_\hbar(f_n)\|-\|\mathcal{Q}_\hbar(f_1\cdots f_n)\|=0$, so that finally $\lim_{\hbar\to0}\|\mathcal{Q}_\hbar(f_1)\cdots\mathcal{Q}_\hbar(f_n)\|=\|f_1\cdots f_n\|_\infty$ by (6). Because of (1), this shows that $F_{f_1,\ldots,f_n}$ is continuous at $\hbar=0$.

Hence one is in the situation of Lemma 2, and the claim follows. ■

Although it is irrelevant for the present paper, we note that Proposition 1 holds equally well (with a different proof) if the set I is not discrete, provided that all $\mathfrak{A}_\hbar$ are identical for $\hbar\neq0$, and in addition the function $\hbar\to\mathcal{Q}_\hbar(f)$ is continuous for each $f\in\tilde{\mathfrak{A}}_0$.[4]

III. QUANTIZATION OF PURE STATES AND BEREZIN QUANTIZATION

A Hilbert space $\mathcal{H}$ is a symplectic manifold, with symplectic form $\omega(\Phi,\Omega)=2\,\mathrm{Im}\,(\Phi,\Omega)$ [here the inner product $(\,,\,)$ on $\mathcal{H}$ is linear in the second entry].[1,2] This form is invariant under the standard action $\exp(i\alpha):\Psi\to\exp(i\alpha)\Psi$ of $U(1)$ on $\mathcal{H}$, so that the quotient $\mathcal{H}^*/U(1)$ is a Poisson manifold (here $\mathcal{H}^*=\mathcal{H}\backslash\{0\}$). The symplectic leaves[2] of $\mathcal{H}^*/U(1)$ are the spaces $S_r=\mathcal{H}_r/U(1)$, where $\mathcal{H}_r=\{\Psi\in\mathcal{H}\mid(\Psi,\Psi)=r^2\}$. In particular, the projective space $\mathbb{P}\mathcal{H}$ may be identified with S_1, and is therefore a symplectic manifold, with symplectic form $\omega_{\mathbb{P}\mathcal{H}}$.

In addition, $\mathbb{P}\mathcal{H}$ is equipped with a transition probability $p:\mathbb{P}\mathcal{H}\times\mathbb{P}\mathcal{H}\to[0,1]$, given by $p(\psi,\varphi)=|(\Psi,\Phi)|^2$; here Ψ and Φ are arbitrary lifts of ψ and φ to unit vectors in $\mathcal{H}$. Equipped with these transition probabilities and with the Poisson bracket defined by $\omega_{\mathbb{P}\mathcal{H}}$, the manifold $\mathbb{P}\mathcal{H}$ is the pure state space of a quantum system without superselection rules. See Ref. 4 and references therein.

The pure state space of a classical system is a symplectic manifold (S,ω_S), supporting the Liouville measure μ_L. Such a classical pure state space may be seen as carrying the ''classical'' transition probability p_0, defined by $p_0(\rho,\sigma)=\delta_{\rho\sigma}$.

Definition 3: Let $I\subseteq\mathbb{R}$ *be as in Definition II, and put* $I_0=I\backslash\{0\}$. *A pure state quantization of a symplectic manifold* (S,ω) *consists of a collection of Hilbert spaces* $\{\mathcal{H}_\hbar\}_{\hbar\in I_0}$ *and a collection of smooth injections* $\{q_\hbar:S\to\mathbb{P}\mathcal{H}_\hbar\}_{\hbar\in I_0}$, *for which the following requirements are satisfied.*

(1) *There exists a positive function* $c:I_0\to\mathbb{R}\backslash\{0\}$ *such that for all* $\hbar\in I_0$ *and all* $\psi\in\mathbb{P}\mathcal{H}_\hbar$ *one has*

$$c(\hbar)\int_S d\mu_L(\sigma)\,p(q_\hbar(\sigma),\psi)=1. \tag{8}$$

(2) *For all fixed* $f\in C_c(S)$ *and* $\rho\in S$ *the function*

$$\hbar\to\int_S d\mu_L(\sigma)\,p(q_\hbar(\rho),q_\hbar(\sigma))f(\sigma)$$

is continuous on I_0 *and satisfies*

$$\lim_{\hbar\to0}c(\hbar)\int_S d\mu_L(\sigma)\,p(q_\hbar(\rho),q_\hbar(\sigma))f(\sigma)=f(\rho). \tag{9}$$

(3) *For each* $\hbar \in I_0$, *the map* $q_\hbar$ *is a symplectomorphism, that is,*

$$q_\hbar^* \omega_{\mathbb{P}\mathcal{H}} = \omega_S . \tag{10}$$

Since $f \in C_c(S)$, the continuity requirement on I_0 in this condition is equivalent to the continuity of the function $\hbar \to p(q_\hbar(\rho), q_\hbar(\sigma))$ for fixed ρ and σ. Using Urysohn's lemma, it is not difficult to show[4] that (9) and (8) imply

$$\lim_{\hbar \to 0} p(q_\hbar(\rho), q_\hbar(\sigma)) = \delta_{\rho\sigma} . \tag{11}$$

That is, in quantizing pure states the quantum-mechanical transition probabilities should converge to the classical ones for $\hbar \to 0$.

More generally, one may replace the Liouville measure μ_L in Definition 3 by a family of measures $\mu_\hbar$, and substitute an appropriate limiting condition for (10); see Ref. 4. This generality is not needed for the present paper. The author is indebted to E. Hawkins for this remark.

A pure state quantization naturally leads to the quantization of observables.

Definition 4: Let $\{\mathcal{H}_\hbar, q_\hbar\}_{\hbar \in I_0}$ *be a pure state quantization of a symplectic manifold* S. *The Berezin quantization of a function* $f \in L^\infty(S)$ *is the family of operators* $\{\mathcal{Q}_\hbar^B(f)\}_{\hbar \in I_0}$, *where* $\mathcal{Q}_\hbar^B(f) \in \mathfrak{B}(\mathcal{H}_\hbar)$ *is defined by the weak integral*

$$\mathcal{Q}_\hbar^B(f) = c(\hbar) \int_S d\mu_L(\sigma) f(\sigma) [q_\hbar(\sigma)]. \tag{12}$$

Here $[q_\hbar(\sigma)]$ *is the projection onto the one-dimensional subspace in* $\mathcal{H}_\hbar$ *whose image in* $\mathbb{P}\mathcal{H}_\hbar$ *is* $q_\hbar(\sigma)$. *In case that* $f \in L^1(S) \cap L^\infty(S)$, *the integral is an ordinary Lebesgue integral (with values in a Banach space).*

A number of properties of $\mathcal{Q}_\hbar^B$ are immediately evident (cf. Ref. 9): $\mathcal{Q}_\hbar^B$ is positive (that is, $f \geqslant 0$ almost everywhere on S implies $\mathcal{Q}_\hbar^B(f) \geqslant 0$ in $\mathfrak{B}(\mathcal{H}_\hbar)$), and if f is real-valued then $\mathcal{Q}_\hbar^B(f)$ is self-adjoint. Moreover, $\mathcal{Q}_\hbar^B(f)$ is bounded, with

$$\| \mathcal{Q}_\hbar^B(f) \| \leqslant \| f \|_\infty . \tag{13}$$

To prove the last property, let A be a bounded symmetric operator such that $|(\Psi, A\Psi)| \leqslant c\|\Psi\|^2$ for some $c > 0$, and all Ψ. One then replaces Ψ by $\Psi \pm A\Psi/c$, and subtracts the two inequalities thus obtained. This implies the inequality $\|A\Psi\| \leqslant c\|\Psi\|$, showing that A is bounded with norm $\leqslant c$. This argument with (8) implies (13).

A Berezin quantization is not necessarily strict when restricted to, say, $C_c^\infty(S)$; even (10) does not imply (3). At the present abstract level, all that can be inferred is the following.

Proposition 2: Let $f \in C_0(S)$, *and assume that* I *is discrete. Then*

$$\lim_{\hbar \to 0} \| \mathcal{Q}_\hbar^B(f) \| = \| f \|_\infty . \tag{14}$$

We initially assume that $f \in L^1(S) \cap C_0(S)$, and at the end extend the result to $f \in C_0(S)$ using the continuity of $\mathcal{Q}_\hbar^B$. Equation (13) implies

$$\limsup_{\hbar \to 0} \| \mathcal{Q}_\hbar^B(f) \| \leqslant \| f \|_\infty . \tag{15}$$

On the other hand, for $f \in C_0(S)$ we can find $\rho \in S$ for which $\|f\|_\infty = |f(\rho)|$. Now for any unit vector $\Psi \in \mathcal{H}_\hbar$, Eq. (12) implies that

$$(\Psi, \mathcal{Q}_\hbar^B(f)\Psi) = c(\hbar) \int_S d\mu_L(\sigma)\, p(q_\hbar(\sigma), \psi) f(\sigma), \tag{16}$$

where ψ is the projection of Ψ to $\mathbb{P}\mathcal{H}_\hbar$. We now use the obvious inequality $\| \mathcal{Q}_\hbar^B(f) \| \geqslant |(\Psi, \mathcal{Q}_\hbar^B(f)\Psi)|$, take Ψ to be a lift of $q_\hbar(\rho)$, and use (9) to find

$$\liminf_{\hbar\to 0}\|\mathcal{Q}^B_\hbar(f)\|\geqslant\|f\|_\infty. \tag{17}$$

Combining this with (15), Eq. (14) follows. ■

IV. COADJOINT ORBITS AND PERELOMOV'S COHERENT STATES

Let G be a Lie group, with Lie algebra $\mathfrak{g}$. The dual $\mathfrak{g}^*$ of $\mathfrak{g}$ is a Poisson manifold under the so-called Lie-Poisson bracket[2,4]

$$\{f,g\}(\theta)=-\theta([df_\theta,dg_\theta]); \tag{18}$$

here the differential df_θ of $f\in C^\infty(\mathfrak{g}^*)$ at $\theta\in\mathfrak{g}^*$, which is a linear map from $T_\theta\mathfrak{g}^*\simeq\mathfrak{g}^*$ to $\mathbb{R}$, is identified with an element of $\mathfrak{g}\simeq\mathfrak{g}^{**}$, so that the right-hand side of (18) is the Lie bracket in $\mathfrak{g}$.

The symplectic leaves of $\mathfrak{g}^*$ with respect to the Lie–Poisson structure are the coadjoint orbits of G.[2,4] This endows each coadjoint orbit $\mathcal{O}$ with the so-called Lie–Kirillov symplectic structure. It is clear that $\theta\in\mathfrak{g}^*$ satisfies $\theta([X,Y])=0$ for all $X,Y\in\mathfrak{g}_\theta$, where $\mathfrak{g}_\theta$ is the Lie algebra of the stabilizer G_θ of θ under the coadjoint action. In other words, $\theta:\mathfrak{g}_\theta\to\mathbb{R}$ is a Lie algebra homomorphism.

Definition 5: A coadjoint orbit $\mathcal{O}\in\mathfrak{g}^*$ is called integral if for some (hence all) $\theta\in\mathcal{O}$ the functional $\theta\restriction\mathfrak{g}_\theta$ exponentiates to a character of G_θ.

In other words, θ is integral iff there is a character U_θ of G_θ such that $\theta=idU_\theta$ on $\mathfrak{g}_\theta$. If this holds for one $\theta\in\mathcal{O}$ it holds for all, since one has $U_{\mathrm{Co}(x)\theta}=U_\theta\circ\mathrm{Ad}(x^{-1})$. Here and in what follows, Co and Ad stand for the coadjoint and the adjoint action of G on $\mathfrak{g}^*$ and on $\mathfrak{g}$, respectively.

In the remainder of this paper, G is a compact connected Lie group. We assume familiarity with the standard Cartan–Weyl description of the unitary irreducible representations of G.[18] We fix a maximal torus $T\subset G$. The coadjoint orbits of T in the dual $\mathfrak{t}^*$ of the Lie algebra $\mathfrak{t}$ of T are points. The integral coadjoint orbits of T form the weight lattice $\Lambda\subset\mathfrak{t}^*$. The Weyl group $W=N(T)/T$ [where $N(T)$ is the normalizer of T] acts on T by conjugation. The derivative of this action is a W-action on $\mathfrak{t}$, whose dual action on $\mathfrak{t}^*$ is the projection of the coadjoint action of $N(T)$ to $N(T)/T$. This action maps the weight lattice Λ into itself.

A functional $\lambda\in\mathfrak{t}^*$ is regular when $w\lambda=\lambda$ for $w\in W$ implies $w=e$ (and singular otherwise); this defines the sets $\mathfrak{t}^*_r$ and $\Lambda_r=\mathfrak{t}^*_r\cap\Lambda$ of regular elements and regular weights in $\mathfrak{t}^*$, respectively. A Weyl chamber is a connected component C of $\mathfrak{t}^*_r$, and thereby forms an open convex cone in $\mathfrak{t}^*$. Singular weights clearly lie on the boundary of some Weyl chamber. One singles out an arbitrary Weyl chamber C_d, and declares a weight dominant if it lies in the closure $\overline{C_d}$. The point is now that each W-orbit intersects a given closed Weyl chamber $\overline{C}$ in exactly one point.

Combining the Cartan–Weyl theory with that of Kostant,[7] one obtains a number of parametrizations of the unitary dual of G (i.e., the set of unitary irreducible representations of G modulo unitary equivalence).

Proposition 3: There exist bijective correspondences between the unitary dual $\hat{G}$ of G, the set of W-orbits in Λ, the set $\Lambda_{\mathrm{d}}=\Lambda\cap\overline{C}_{\mathrm{d}}$ of dominant weights, and the set of integral coadjoint orbits in $\mathfrak{g}^*$. The latter set is isomorphic to $\mathfrak{t}^*/W$. In other words, one has

$$\hat{G}\simeq\Lambda/W\simeq\Lambda_d\simeq(\mathfrak{g}^*/W)_{\mathrm{integral}}. \tag{19}$$

Though exceedingly well-known, we recall the explicit form of the bijection $\hat{G}\simeq\Lambda_d$ in Proposition 3, as it will play an important role in the proof of our main result. We use the standard Cartan–Weyl basis $\{H_j\}_{j=1,\ldots,r}\cup\{E_\alpha,E_{-\alpha}\}_{\alpha\in\Delta^+}$ of $\mathfrak{g}_{\mathbb{C}}$, where r is the rank of G, and Δ^+ is the set of positive roots (relative to a choice of C_d). The vectors H_i lie in $\mathfrak{t}$, and the $E_{\pm\alpha}$ are eigenvectors of each H_i under restriction from $\mathfrak{g}$ to $\mathfrak{t}$ of the adjoint representation of $\mathfrak{g}$ on $\mathfrak{g}_{\mathbb{C}}$.

A Hilbert space $\mathcal{V}_\lambda$ carrying an irreducible representation $U_\lambda(G)$ corresponding to a dominant weight $\lambda\in\Lambda_d$ has a highest weight vector Ψ_λ of norm 1, unique up to a phase, such that

$$dU_\lambda(H_j)\Psi_\lambda=-i\lambda(H_j)\Psi_\lambda \tag{20}$$

for $j=1,\ldots,r$, whereas for all $\alpha\in\Delta^+$ one has

$$dU_\lambda(E_\alpha)\Psi_\lambda=0. \tag{21}$$

Since $dU_\lambda(E_\alpha)^*=-dU_\lambda(E_{-\alpha})$, it follows that

$$(\Psi_\lambda,dU_\lambda(E_\alpha)\Psi_\lambda)=0 \tag{22}$$

for all $\alpha\in\Delta$.

The association of a coadjoint orbit in $\mathfrak{g}^*$ with a W-orbit in $\mathfrak{t}^*$ is as follows. We write $\mathfrak{t}_{\mathbb{C}}^{\perp}$ for the span of all $E_{\pm\alpha}$, and $\mathfrak{t}^{\perp}=\mathfrak{t}_{\mathbb{C}}^{\perp}\cap\mathfrak{g}$. Given a weight $\lambda\in\mathfrak{t}^*$, putting $\theta(\lambda)=0$ on $\mathfrak{t}^{\perp}$ and $\theta(\lambda)=\lambda$ on $\mathfrak{t}$, one obtains an extension $\theta(\lambda)\in\mathfrak{g}^*$ of λ. Thus the coadjoint orbit $\mathcal{O}_\lambda=\mathcal{O}_{\theta(\lambda)}$ associated to λ is the coadjoint orbit through $\theta(\lambda)$; it is obvious from the definition of the W-action on $\mathfrak{t}^*$ that all points of W-orbit of λ are mapped into $\mathcal{O}_\lambda$. We will use the label λ to denote a dominant weight in $\Lambda_d\subset\mathfrak{t}^*$, as well as the corresponding element $\theta(\lambda)$ in the coadjoint orbit $\mathcal{O}_\lambda\subset\mathfrak{g}^*$.

The correspondence $\Lambda_d\simeq\mathfrak{g}^*/G$ is most easily described in terms of the momentum map.[1,2] The representation $U_\lambda(G)$ on $\mathcal{V}_\lambda$ quotients to a G-action $\tilde{U}_\lambda$ on the projective space $\mathbb{P}\mathcal{V}_\lambda$. Since U_λ is unitary, this quotient action may be computed from the action of $U_\lambda(G)$ on unit vectors in $\mathcal{V}_\lambda$. As explained at the beginning of Sec. III, $\mathbb{P}\mathcal{V}_\lambda$ is a symplectic manifold. It is clear from the definition of the symplectic structure on $\mathbb{P}\mathcal{V}_\lambda$ and the unitarity of U_λ that the reduced G-action $\tilde{U}_\lambda$ preserves the symplectic form $\omega_{\mathbb{P}\mathcal{V}_\lambda}$. More strongly, this action admits an equivariant momentum map $J:\mathbb{P}\mathcal{V}_\lambda\to\mathfrak{g}^*$, given by

$$J_X(\psi)=i(\Psi,dU_\lambda(X)\Psi). \tag{23}$$

As always, the unit vector $\Psi\in\mathcal{V}_\lambda$ is an arbitrary lift of $\psi\in\mathbb{P}\mathcal{V}_\lambda$.

Proposition 4: The coadjoint orbit $\mathcal{O}_\lambda$ corresponding to an irreducible representation U_λ with highest weight vector Ψ_λ contains $J\psi_\lambda$. In fact, the restriction of J to $\mathbb{P}U_\lambda(G)\Psi_\lambda$ is a symplectomorphism onto $\mathcal{O}_\lambda$.

Here it is understood that $\mathbb{P}U_\lambda(G)\Psi_\lambda$ inherits the usual symplectic structure of $\mathbb{P}\mathcal{V}_\lambda$, and that $\mathcal{O}_\lambda$ is endowed with the Lie–Kirillov symplectic form. For convenience, as well as for later reference, we include a proof of this fundamental result (see also Ref. 1 or 4).

Equations (23), (20), (21), and (22) imply that $\langle J(\psi_\lambda),X\rangle$ equals $\lambda(X)$ for $X\in\mathfrak{t}$ and equals 0 for $X\in\mathfrak{t}^{\perp}$. Hence $J(\psi_\lambda)$ is precisely the element $\theta(\lambda)\in\mathfrak{g}^*$ discussed after (22), proving the first claim.

By (23), the stability group $G_{J(\psi_\lambda)}$ of $J(\psi_\lambda)$ consists of those $x\in G$ for which

$$(U_\lambda(x)\Psi_\lambda,dU_\lambda(Y)U_\lambda(x)\Psi_\lambda)=(\Psi_\lambda,dU_\lambda(Y)\Psi_\lambda) \tag{24}$$

for all $Y\in\mathfrak{g}$. Since U_λ is irreducible this implies that Ψ_λ and $U_\lambda(x)\Psi_\lambda$ define the same element of $\mathbb{P}\mathcal{H}_\lambda$, proving that $G_{J(\psi_\lambda)}\subseteq G_{\psi_\lambda}$. The opposite inclusion is trivial from the equivariance of J. ∎

The unit vectors in $\mathcal{V}_\lambda$ of the form $U_\lambda(x)\Psi_\lambda$, where $x\in G$, are Perelomov's coherent states[10,11], which are parametrized by G. It should be clear by now that we are not interested in these states themselves, but in their projections to $\mathbb{P}\mathcal{V}_\lambda$. It follows from Proposition 3 that the ensuing family is parametrized by the coadjoint orbit $\mathcal{O}_\lambda$. Remarkably, the coadjoint orbits of maximal dimension among all coadjoint orbits $\mathfrak{g}$, which are the ones for which $G_\lambda=T$, are precisely the ones labeled by a regular weight λ. The ones of smaller dimension, for which G_λ properly contains T, are labeled by a singular weight.

V. BEREZIN QUANTIZATION OF COADJOINT ORBITS

Proposition 4 shows how to construct the coadjoint orbit $\mathcal{O}_\lambda$ corresponding to a given unitary irreducible representation U_λ by Proposition 3. The converse passage is accomplished by the Borel–Weil–Bott theory, which may be seen as a special case of geometric quantization.[7] In what follows we assume the correspondence in question to be given, and instead are interested in constructing an entire family of representations of G that are in some sense associated to a given

coadjoint orbit. The fundamental idea of rescaling the label of an irreducible representation by multiplying with $1/\hbar$ (which accordingly has to be quantized in the compact case) is due to Berezin (1975); see also Perelomov (1986). In a more intuitive setting, this rescaling was explicit in the early years of quantum mechanics, and seems to comprise one of the faces of Bohr's correspondence principle. The idea was further developed in the context of the classical limit of quantum mechanics by Lieb[8] and Simon.[9]

Theorem 1: *Let G be a compact connected Lie group, and $\mathcal{O}_\lambda$ an integral coadjoint orbit (cf. 5), corresponding to a highest weight $\lambda \in \Lambda_d$. For $\hbar = 1/k$, $k \in \mathbb{N}$, define $\mathcal{H}_\hbar = \mathcal{V}_{\lambda/\hbar}$, i.e., the carrier space of the irreducible representation $U_{\lambda/\hbar}(G)$ with highest weight $\lambda/\hbar = k\lambda$.*

Let $\tau: \mathcal{H}_\hbar \to \mathbb{P}\mathcal{H}_\hbar$ be the canonical projection. The map $q_\hbar : \mathcal{O}_\lambda \to \mathbb{P}\mathcal{H}_\hbar$, given by

$$q_\hbar(\mathrm{Co}(x)\lambda) = \tau(U_{\lambda/\hbar}(x)\Psi_{\lambda/\hbar}), \tag{25}$$

is well defined and injective, and provides a pure state quantization of $\mathcal{O}_\lambda$ (equipped with minus the Lie–Kirillov symplectic structure) on $I = 1/\mathbb{N} \cup 0$.

One should note here that $k\lambda \in \overline{C_d}$ when $\lambda \in \overline{C_d}$, since Weyl chambers are convex cones. The map $q_\hbar$ is well-defined and injective by the equation $G_{k\lambda} = G_\lambda$ plus the argument on stability groups used in the proof of Proposition 4. In fact, if we define $J_\hbar : \mathbb{P}\mathcal{H}_\hbar \to \mathfrak{g}^*$ by (23) with the right-hand side divided by k, it follows from (4) that $J_\hbar$ takes values in $\mathcal{O}_\lambda$ and is a left-inverse of $q_\hbar$.

We start from the fact that the Haar measure on G (with total mass 1) pushes forward to the Liouville measure derived from the Lie symplectic structure under the canonical projection $G \to \mathcal{O}_\lambda \simeq G/G_\lambda$. Using the invariance of the Haar measure and the unitarity of $U_{k\lambda}$, we then have

$$\int_{\mathcal{O}_\lambda} d\mu_L(\sigma)\, p(q_\hbar(\rho), q_\hbar(\sigma)) f(\sigma) = \int_G dx\, |(\Psi_{k\lambda}, U_{k\lambda}(x)\Psi_{k\lambda})|^2 f_\lambda(yx) \tag{26}$$

for all $f \in C(\mathcal{O}_\lambda)$, where $f_\lambda = \tau^*_{G \to G/G_\lambda} f \in C(G)$ is a right-G_λ-invariant function, and y is such that $\tau_{G \to G/G_\lambda}(y) = \rho \in \mathcal{O}_\lambda$. Choosing $f = 1$, the orthogonality relations for compact groups then imply (8) with

$$c(\hbar) = d_{\lambda/\hbar}, \tag{27}$$

i.e., the dimension of $\mathcal{H}_\hbar = \mathcal{V}_{\lambda/\hbar}$.

Equation (10) follows from Proposition 4. To prove (9) we use a result of Gilmore.[19]

Lemma 3: Let λ_i be dominant weights with highest weight representations and vectors U_{λ_i} and Ψ_{λ_i}, respectively $(i = 1,2)$. Then for each $x \in G$,

$$(\Psi_{\lambda_1}, U_{\lambda_1}(x)\Psi_{\lambda_1}) \cdot (\Psi_{\lambda_2}, U_{\lambda_2}(x)\Psi_{\lambda_2}) = (\Psi_{\lambda_1+\lambda_2}, U_{\lambda_1+\lambda_2}(x)\Psi_{\lambda_1+\lambda_2}). \tag{28}$$

This is immediate from (20), (21), and the connectedness of G. ∎

This lemma implies that

$$(\Psi_{k\lambda}, U_{k\lambda}(x)\Psi_{k\lambda}) = (\Psi_\lambda, U_\lambda(x)\Psi_\lambda)^k. \tag{29}$$

Using (26), we can write the left-hand side of (9) as

$$\lim_{\hbar \to 0} \int_{\mathcal{O}_\lambda} d\mu_L(\sigma)\, p(\rho_\hbar, \sigma_\hbar) f(\sigma) = \lim_{k \to \infty} \int_G d\mu_k(x) f_\lambda(yx),$$

where μ_k is a probability measure on G defined by

$$d\mu_k(x) = d_{k\lambda}\, dx |(\Psi_\lambda, U_\lambda(x)\Psi_\lambda)|^{2k}. \tag{30}$$

It is obvious that each μ_k is right-G_λ-invariant. It follows from (21), (22), and the fact that the exponential map is surjective for compact Lie groups, that $|(\Psi_\lambda, U_\lambda(x)\Psi_\lambda)|$, which is evidently $\leqslant 1$, equals 1 iff $x \in G_\lambda$. Hence for large k the support of μ_k is increasingly concentrated on G_λ. This suggests that

$$\lim_{k\to\infty} \mu_k(f) = \int_{G_\lambda} dh\, f(h) \tag{31}$$

for all $f \in C(G)$, where dh is the normalized Haar measure on G_λ. This is indeed the case, as proved by Duffield[20] (related results may be found in Refs. 8 and 9). For the right-G_λ-invariant function $f_\lambda \in C(G)^{G_\lambda}$ one therefore obtains

$$\lim_{k\to\infty} \mu_k(f_\lambda) = f_\lambda(e).$$

This proves (9), which finishes the proof of Theorem 1. ■

The Berezin quantization $\mathcal{Q}^B_\hbar$ associated with the pure state quantization in Theorem 1 is defined on $\tilde{\mathfrak{A}}_0 = C^\infty(\mathcal{O}_\lambda)$. By (12), one has

$$\mathcal{Q}^B_{1/k}(f) = d_{k\lambda} \int_G dx\, f_\lambda(x)[U_{k\lambda}(x)\Psi_{k\lambda}]; \tag{32}$$

this is an element of $\mathfrak{A}_{1/k} = \mathfrak{M}_{d_{k\lambda}}(\mathbb{C})$ (the C^*-algebra of complex $d_{k\lambda} \times d_{k\lambda}$ matrices). Recall that $[\Phi]$ is the projection onto $\mathbb{C}\Phi$.

The most important property of $\mathcal{Q}^B_\hbar$ is its G-equivariance. For $x \in G$ we write

$$\alpha^{1/k}_x(A) = U_{k\lambda}(x) A U_{k\lambda}(x)^*, \tag{33}$$

where $A \in \mathfrak{A}_{1/k}$, and for $f \in \mathfrak{A}_0 = C(\mathcal{O}_\lambda)$ we put

$$\alpha^0_x(f) = \mathrm{Co}(x^{-1})^* f. \tag{34}$$

Proposition 5: For all $k \in \mathbb{N}$, $x \in G$, and $f \in L^\infty(\mathcal{O}_\lambda)$ one has

$$\mathcal{Q}^B_{1/k}(\alpha^0_x(f)) = \alpha^{1/k}_x(\mathcal{Q}^B_{1/k}(f)). \tag{35}$$

This is immediate from (32), (34), the fact that $(\mathrm{Co}(y)^* f)_\lambda = L^*_y f_\lambda$, the right-invariance of the Haar measure, and (33). ■

We are now ready for the main result of this paper.

Theorem 2: *The Berezin quantization (32), defined on the space $C^\infty(\mathcal{O}_\lambda)$, is strict.*

We show that the assumptions in Proposition 1 are met. Equation (6) follows from Theorem 1 and Proposition 2. The completeness condition is an easy consequence of Schur's lemma and the irreducibility of $U_{k\lambda}$.

We will now prove the remaining conditions (5) and (3), using the notation of the proof of Theorem 1. We pick a unit vector Φ_k in each $\mathcal{V}_{k\lambda}$, and use the invariance of the Haar measure and the orthogonality relations for compact groups to write

$$(\Phi_k, (\mathcal{Q}^B_{1/k}(f)\mathcal{Q}^B_{1/k}(g) - \mathcal{Q}^B_{1/k}(fg))\Phi_k) = d_{k\lambda} \int_G dx\, f_\lambda(x)(\Phi_k, U_{k\lambda}(x)\Psi_{k\lambda}) I_k(x), \tag{36}$$

where

$$I_k(x) = d_{k\lambda} \int_G dy\, (\Psi_{k\lambda}, U_{k\lambda}(y)\Psi_{k\lambda}) F^x_\lambda(y), \tag{37}$$

in which we have abbreviated

$$F^x_\lambda(y)=(U_{k\lambda}(xy)\Psi_{k\lambda},\Phi_k)[g_\lambda(xy)-g_\lambda(x)]. \tag{38}$$

In the notation used after (26), the function F^x_λ on G corresponds to a function F^x on $\mathcal{O}_\lambda$.

Using (29), we can write $(\Psi_{k\lambda},U_{k\lambda}(y)\Psi_{k\lambda})=\exp(-kS_\lambda(y))$, where $S_\lambda(y)=-\log(\Psi_\lambda,U_\lambda(y)\Psi_\lambda)$ (in view of the exponentiation, the choice of the branch cut of the logarithm is irrelevant). The function S_λ is right-G_λ-invariant; we denote the corresponding function on G/G_λ by S. We identify G/G_λ with $\mathcal{O}_\lambda$, so that the coset $[G_\lambda]\in G/G_\lambda$ is identified with $\lambda\in\mathcal{O}_\lambda$.

Putting $S^+_\lambda(y)=-\log|(\Psi_\lambda,U_\lambda(y)\Psi_\lambda)|$, the absolute value of $\exp(-kS)$ is $\exp(-kS^+)$. As in the argument preceding (31) we see that S^+ takes values in $[0,\infty]$ and assumes its unique absolute minimum 0 at λ. Since F^x_λ in (37) is bounded, a standard argument implies that to $O(\exp(-k))$ we may replace the integration over G/G_λ by one over any neighborhood of λ.

We identify $T_\lambda\mathcal{O}_\lambda$ with $\mathfrak{g}/\mathfrak{g}_\lambda$, and use complex coordinates $\{z_\alpha,\overline{z_\alpha}\}_{\alpha\in\Delta^+_\lambda}$, where Δ^+_λ consists of those positive roots for which $(\lambda,\alpha)\neq 0$. By the definition of a highest weight, this implies that $(\lambda,\alpha)>0$ for all $\alpha\in\Delta^+_\lambda$. The coordinates $(z_\alpha,\overline{z_\alpha})$ correspond to the point in $\mathcal{O}_\lambda$ given by

$$\mathrm{Co}\left[\exp\left(i\sum_{\alpha\in\Delta^+_\lambda}(z_\alpha E_\alpha-\overline{z}_\alpha E_{-\alpha})\right)\right]\lambda.$$

A simple computation, using the Cartan–Weyl form of the commutation relations in $\mathfrak{g}$ (Ref. 18) and (20)–(22), leads to

$$S(z_\alpha,\overline{z_\alpha})=\sum_{\alpha\in\Delta^+_\lambda}(\lambda,\alpha)z_\alpha\overline{z_\alpha}+O(|z|^3). \tag{39}$$

Hence to $O(\exp(-k))$ we may approximate $I_k(x)$ by

$$d_{k\lambda}\int_{\mathfrak{g}/\mathfrak{g}_\lambda}\left(\prod_{\alpha\in\Delta^+_\lambda}\frac{dz_\alpha d\overline{z_\alpha}}{2\pi}\right)J(z_\alpha,\overline{z_\alpha})\exp\left(-k\sum_{\alpha\in\Delta^+_\lambda}(\lambda,\alpha)z_\alpha\overline{z_\alpha}\right)F^x(z_\alpha,\overline{z_\alpha}),$$

where J is a Jacobian, and F^x_λ has been extended to $\mathfrak{g}/\mathfrak{g}_\lambda$ by, say, the exponential map. If we omit the factor $[\cdots]$ in (38), the integral (37) can be done, using the orthogonality relations for compact groups. On the other hand, we can compute the above integral to lowest order in the steepest descent approximation; this avoids the need to compute $J(0)$. Comparing the results computes the prefactor in the steepest descent approximation as unity. As a by-product we obtain the asymptotic expression for $k\to\infty$,

$$d_{k\lambda}\sim\frac{\Pi_{\alpha\in\Delta^+_\lambda}(\lambda,\alpha)}{J(0)}k^{1/2\dim(\mathcal{O}_\lambda)}, \tag{40}$$

where $\dim(\mathcal{O}_\lambda)$ is given by

$$\dim(\mathcal{O}_\lambda)=\dim(\mathfrak{g})-\dim(\mathfrak{t})-2\,\mathrm{Card}\{\alpha\in\Delta^+|(\lambda,\alpha)=0\}. \tag{41}$$

Comparison with the Weyl dimension formula[18] then yields $J(0)=\Pi_{\alpha\in\Delta^+}(\alpha,\delta)$, where $\delta=\frac{1}{2}\Sigma_{\alpha\in\Delta^+}\alpha$. Equation (41) follows from the decomposition

$$\mathfrak{g}_\lambda=\mathfrak{t}\bigoplus\oplus_{\alpha\in\Delta^+|(\lambda,\alpha)=0}\mathfrak{g}^{\mathbb{R}}_\alpha, \tag{42}$$

where $\mathfrak{g}^{\mathbb{R}}_\alpha=\mathfrak{g}\cap(\mathbb{C}E_\alpha\oplus\mathbb{C}E_{-\alpha})$. This decomposition is easily derived from the proof of Proposition 4, the Cartan–Weyl commutation relations, and the fact that $\lambda(E_\alpha)=0$ for all $\alpha\in\Delta$.

Thus the steepest descent approximation to the above integral, and therefore to (37), reads

$$I_k(x)=\sum_{l=0}^{N}\frac{1}{l!}D^l(JF^x)(0)+O(k^{-N-1}), \tag{43}$$

where, abbreviating $\partial_\alpha=\partial/\partial z_\alpha$ and $\overline{\partial}_\alpha=\overline{\partial}/\overline{\partial z}_\alpha$, we have put

$$D=\sum_{\alpha\in\Delta_\lambda^+}\frac{1}{(\lambda,\alpha)}\partial_\alpha\overline{\partial}_\alpha. \tag{44}$$

Substituting this expansion in (36) we see that

$$(\Phi_k,(\mathcal{Q}^B_{1/k}(f)\mathcal{Q}^B_{1/k}(g)-\mathcal{Q}^B_{1/k}(fg))\Phi_k)=O(1/k). \tag{45}$$

To analyze the remainder of $O(1/k)$ we note that the l'th term in the expansion leads to an x-integrand in (36) of the form

$$(U_{k\lambda}(x^{-1})\Phi_k,\Psi_{k\lambda})\overline{(U_{k\lambda}(x^{-1})\Phi_k,\Psi^{(l_1)})}f\partial^{l_2}\overline{\partial}^{l_3}g_\lambda(x),$$

where $l_i\leqslant l$ and the $\Psi^{(l_1)}$ are given by the action of products of $dU_{k\lambda}(E_\alpha)$ and $dU_{k\lambda}(E_{-\alpha})$ on $\Psi_{k\lambda}$. The important point is now that the orthogonality relations (applied to the x-integration) then imply that the $O(k^{-N-1})$ term is bounded by $C\|\Phi_k\|^2/k^{N+1}$ for some constant C. Hence (5) follows.

To prove (3) we need the $l=1$ term in (43). We substitute (44), and perform some partial integrations in the remaining x-integral (using the invariance of the Haar measure). We abbreviate $A=(\Phi_k,U_{k\lambda}(x)\Psi_{k\lambda})$; then (21) implies that $\partial_\alpha A$ and $\overline{\partial}_\alpha\overline{A}$ vanish at $z_\alpha=\overline{z}_\alpha=0$. Terms of the form $\partial_\alpha\overline{\partial}_\alpha A$ (or $\overline{A}$) drop out in the commutator, as do contributions from J (whose first derivatives at 0 already vanish identically). What remains is

$$(\Phi_k,ik[\mathcal{Q}^B_{1/k}(f),\mathcal{Q}^B_{1/k}(g)]-\mathcal{Q}^B_{1/k}(\{f,g\})\Phi_k)=O(1/k), \tag{46}$$

where, in the realization of f,g as G_λ-invariant functions f_λ,g_λ on G,

$$\{f_\lambda,g_\lambda\}=-i\sum_{\alpha\in\Delta_\lambda}\frac{1}{(\lambda,\alpha)}\xi^L_\alpha f_\lambda\,\xi^L_{-\alpha}g_\lambda. \tag{47}$$

Here the left-invariant vector fields $\xi^L_{\pm\alpha}$ on G are defined by first expressing $E_{\pm\alpha}\in\mathfrak{g}_{\mathbb{C}}$ in terms of elements of $\mathfrak{g}$, and then using the usual definition $\xi^L_X f(y)=df(y\exp(tX)/dt(t=0)$. Also, $\Delta_\lambda=\Delta_\lambda^+\cup\Delta_\lambda^-$, i.e., the set of all roots α for which $(\lambda,\alpha)\neq 0$.

To finish the proof, we remark that (47) is precisely the Lie–Kirillov Poisson bracket on $\mathcal{O}_\lambda$; this may be verified at the point $\lambda\in\mathcal{O}_\lambda$ (or $e\in G$) by direct computation from (18), from which the general statement follows by the G-invariance of the Poisson structure.

It is manifest that the right-hand side of (47) is left-G-invariant if f_λ and g_λ are. To prove the right-G_λ-invariance of (47), it is sufficient to establish that it is right-invariant under the derived action of the Lie algebra $\mathfrak{g}_\lambda$, for G_λ is connected. Recall that $[E_\alpha,E_\beta]=N_{\alpha,\beta}E_{\alpha+\beta}$, where $\beta\neq-\alpha$. We now need the identity $N_{-\alpha-\beta,\beta}=-N_{\alpha,\beta}$ (where $\beta\neq\pm\alpha$), which follows from the $\mathrm{Ad}(\mathfrak{g})$-invariance of the inner product on $\mathfrak{g}_{\mathbb{C}}$, combined with the normalization of the E_α. The right-invariance of (47) under $\mathfrak{g}_\lambda$ follows by combining this identity with (42) and the Cartan–Weyl commutation relations.

The higher-order terms in (46) are dealt with as in the above proof of (5). This proves (3), finishing the proof of Theorem 2. ∎

Finally, we remark that the results in this section have an obvious yet somewhat cumbersome generalization: If the orbit $\mathcal{O}_\lambda$ is not integral, but such that $\mathcal{O}_{\lambda/c}$ is integral for some $c\in\mathbb{R}\setminus\{0\}$, we can construct a strict quantization for the values $\hbar=c/k$, $k\in\mathbb{N}$.

ACKNOWLEDGMENT

Supported by a fellowship from the Royal Netherlands Academy of Arts and Science (KNAW).

[1] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge University Press, Cambridge, 1984).
[2] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry* (Springer, New York, 1994).
[3] A. Connes, *Noncommutative Geometry* (Academic, San Diego, 1994).
[4] N. P. Landsman, *Mathematical Topics Between Classical and Quantum Mechanics* (Springer, New York, 1998).
[5] M. A. Rieffel, ''Deformation quantization of Heisenberg manifolds,'' Commun. Math. Phys. **122**, 531–562 (1989).
[6] F. A. Berezin, ''General concept of quantization,'' Commun. Math. Phys. **40**, 153–174 (1975).
[7] B. Kostant, ''Quantization and unitary representations,'' Lect. Notes Math. **170**, 87–208 (1970).
[8] E. H. Lieb, ''The classical limit of quantum spin systems,'' Commun. Math. Phys. **62**, 327–340 (1973).
[9] B. Simon, ''The classical limit of quantum partition functions'' Commun. Math. Phys. **71**, 247–276 (1980).
[10] A. M. Perelomov, ''Coherent states for arbitrary Lie groups,'' Commun. Math. Phys. **26**, 222–236 (1972).
[11] A. Perelomov, *Generalized Coherent States and their Applications* (Springer, Berlin, 1986).
[12] E. Hawkins, ''Quantization of equivariant vector bundles,'' e-print q-alg/9708030.
[13] D. Barmoshe and M. S. Marinov, ''Realization of compact Lie algebras in Káhler manifolds,'' J. Phys. A **27**, 6287–6298 (1994).
[14] J. Dixmier, *C*-Algebras* (North-Holland, Amsterdam, 1977).
[15] E. Kirchberg and S. Wassermann, ''Operations on continuous bundles of C^*-algebras,'' Math. Ann. **303**, 677–697 (1995).
[16] M. A. Rieffel, ''Quantization and operator algebras,'' in *Meeting with the Platypus,* Proceedings of the XIIth International Congress of Mathematical Physics, Brisbane, 1997, edited by A. J. Bracken, D. De Wit, M. Gould, and P. Pearce (International Press, Singapore, 1998).
[17] N. P. Landsman, ''Strict deformation quantization of a particle in external gravitational and Yang-Mills fields,'' J. Geom. Phys. **12**, 93–132 (1993).
[18] B. Simon, *Representations of Finite and Compact Groups* (American Mathematical Society, Providence, 1996).
[19] R. Gilmore, ''The classical limit of quantum nonspin systems,'' J. Math. Phys. **20**, 891–893 (1979).
[20] N. G. Duffield, ''Classical and thermodynamic limits for generalized quantum spin systems,'' Commun. Math. Phys. **127**, 27–39 (1990).

Deformations of massless Gupta–Bleuler triplets in 3+2 De Sitter space

Marc Lesimple
Dipartimento di Matematica, Università degli Studi di Padova, via Belzoni 7, 35131 Padova, Italy

(Received 27 May 1998; accepted for publication 27 July 1998)

We study deformations of massless indecomposable representations (Gupta–Bleuler triplets) for the De Sitter group SO(3,2), showing that there is no deformation of such representations toward a kind of massive Gupta–Bleuler triplets which would have corresponded to an analog of a Higgs–Kibble mechanism in anti De Sitter space. © *1998 American Institute of Physics.* [S0022-2488(98)03411-2]

I. INTRODUCTION

Passage from massless field theory to interacting massive field theory involves deformations of massless gauge theories, and so brings forth the stability problem (in the sense of deformations) for a given massless gauge theory.

In a $3+2$ De Sitter space massless fields are described by indecomposable representations with invariant scalar product, which reduce (for some minimal gauge fixing) to representations of the form $\rho=(\pi_3,V_3)\to(\pi_2,V_2)\to(\pi_1,V_g)$, with kernel (π_1,V_g) of type $D(s+2,s-1)$ or $D(2-s,s)$, in the discrete series of the De Sitter group [denoting representations of the De Sitter algebra so(3,2) of the De Sitter group, associated with elementary particles—with lower energy E_0 and spin s—by $D(E_0,s)$].

In particular, the occurrence of the identity representation [as a submodule of (π_2,V_2)] in one of the Gupta–Bleuler triplets of electrodynamics presumes the existence of Higgs–Kibble phenomena[1] to describe interacting massive spin 1 fields.

A possible Higgs–Kibble mechanism describing massive fields in $3+2$ De Sitter space would depend on deformations of massless gauge theories; which one aspect is deformations of massless Gupta–Bleuler triplets.

Spontaneously broken symmetries in the Lagrangian of a massless field would give the occurrence of massless Goldstone fields,[2] now belonging to the kernel V_g^λ of a deformed Gupta–Bleuler triplet ρ^λ containing an extension, say $(\pi_2^\lambda,V_2^\lambda)\to(\pi_1^\lambda,V_g^\lambda)$; where submodules $(\pi_1^\lambda,V_g^\lambda)$, $(\pi_2^\lambda,V_2^\lambda)$, describe, respectively, longitudinal and transverse modes of a massive field.

In other words Higgs–Kibble phenomena invoke stability of massless gauge theories under deformations inside massless gauge theories. The existence (or nonexistence) of a kernel in ρ^λ could be compared with the conservation (or nonconservation) of the sources of interacting massive spin 2 fields in Minkowski space.[3]

In Sec. II we recall the structure of massless representations. We give a characterization of the first cohomological spaces corresponding to massless Gupta–Bleuler triplets.

In Sec. III, considering the problem of stability, we study deformations of a massless Gupta–Bleuler triplet ρ. We shall see that no indecomposable representations, nonequivalent to ρ, can be obtained by deformations.

II. MASSLESS GUPTA–BLEULER TRIPLETS

Presentation of the unitary irreducible representations involved.

We consider De Sitter space as a covering of the hyperboloid $(y^0)^2-(y^1)^2-(y^2)^2-(y^3)^2+(y^5)^2=R^{-1}$ in $(\mathbb{R}^5,g)$, R being the curvature and g the metric $g_{\alpha\beta}=(1,-1,-1,-1,1)$.

$\mathbf{G}$ will denote the connected component $\mathrm{SO}_0(3,2,)$, $\tilde{\mathbf{G}}$ its universal covering, $\mathbf{g}$ the complexified of its Lie algebra $\mathbf{g}_0$, and U($\mathbf{g}$) the universal enveloping algebra of $\mathbf{g}$. Let $\{L_{\alpha\beta}\}_{\alpha,\beta=0,1,2,3,5}$ be a basis of $\mathbf{g}_0$,[4] that satisfies $L_{\alpha\beta}=-L_{\beta\alpha}$, $[L_{\alpha\beta},L_{\gamma\delta}]=g_{\alpha\beta}L_{\beta\gamma}+g_{\beta\gamma}L_{\alpha\delta}-g_{\alpha\gamma}L_{\beta\delta}-g_{\beta\delta}L_{\alpha\gamma}$. We keep the notations of Ref. 5.

 © 1998 American Institute of Physics

Let $(H,H',E_{\pm\alpha},E_{\pm\beta},E_{\pm\gamma},E_{\pm\delta})$ be defined by

$$H=L_{21}+L_{50},\quad H'=L_{21}-L_{50},$$

$$E_{\pm\alpha}=L_{25}+L_{10}\pm i(L_{51}+L_{20}),\quad E_{\pm\beta}=L_{25}-L_{10}\pm i(L_{51}-L_{20}),$$

$$E_{\pm\gamma}=L_{35}\pm iL_{30},\quad E_{\pm\delta}=\pm L_{23}+iL_{31}.$$

We consider the four subalgebras of $\mathbf{g}$, of type A_1, $\mathbf{L}$ (respectively, $\mathbf{L}'$) generated by $(E_{\pm\alpha},H)$ [respectively $(E_{\pm\beta},H')$], $\mathbf{L}_1$ (respectively, $\mathbf{L}_2$) by $(H_1,-E_{\pm\delta})$ with $H_1=-i/2(H+H')$ [respectively $(H_2,-E_{\pm\gamma})$] with $H_2=-i/2(H-H')$.

Let us rewrite the commutation relations for $\mathbf{L}\oplus\mathbf{L}'$,

$$[H,E_{\pm\alpha}]=\pm 2iE_{\pm\alpha},\quad [H,E_{\pm\beta}]=0,$$

$$[H',E_{\pm\beta}]=\pm 2iE_{\pm\beta},\quad [H',E_{\pm\alpha}]=0,$$

$$[E_\alpha,E_{-\alpha}]=-4iH,\quad [E_\beta,E_{-\beta}]=-4iH',\quad [H,H']=0.$$

Let $\mathbf{K}=\mathrm{SO}(3)\times\mathrm{SO}(2)$ be the maximal compact subgroup of $\mathbf{G}$, $\tilde{\mathbf{K}}=\mathrm{SU}(2)\times\mathbb{R}$ its universal covering, $\mathbf{k}=\mathrm{so}(2)\oplus\mathrm{so}(3)$ the complexification of its Lie algebra. The center of the enveloping algebra of, respectively, $\mathbf{L},\mathbf{L}',\mathbf{L}_1,\mathbf{L}_2$ is generated by the following elements:

$$Q_L=H^2+H+E_\alpha E_{-\alpha},\quad Q_{L'}=H'^2+H'+E_{-\beta}E_\beta,$$

$$Q_{L_1}=-\frac{1}{4}(H^2+H'^2)-\frac{1}{2}HH'-\frac{i}{2}(H+H')+E_{-\delta}E_\delta,$$

$$Q_{L_2}=-\frac{1}{4}(H^2+H'^2)+\frac{1}{2}HH'-\frac{i}{2}(H-H')+E_{-\gamma}E_\gamma.$$

Given a unitary irreducible representation of $\tilde{\mathbf{G}}$ in a Hilbert space $\mathscr{H}$, $(U,\mathscr{H})$ of type $D(E_0,s)$ with extremal weight (E_0,s).

The space V of $\mathbf{k}$-finite vectors of $\pi=DU$ (the differential of U) decomposes as the vectorial sum of weight subspaces $V_{\sigma\delta}$, of weight (σ,δ), $V=\oplus_{\substack{\sigma=E_0,E_0+1,\dots\\ \delta=-s,\dots,+s}} V_{\sigma\delta}$.[5,6] There exists an orthonormal basis $(v_{\sigma\delta}^{nn'})n,n'\in\mathbb{N}$ of $V_{\sigma\delta}$ for which the action of $\mathbf{g}$ is given by, noting $l=-(\sigma+\delta)/2$ and $l'=-(\sigma-\delta)/2$,

$$\begin{aligned}
&\pi(H)v_{\sigma\delta}^{nn'}=2i(-l+n)v_{\sigma\delta}^{nn'},\\
&\pi(H')v_{\sigma\delta}^{nn'}=-2i(-l+n')v_{\sigma\delta}^{nn'},\\
&\pi(E_\alpha)v_{\sigma\delta}^{nn'}=2i\alpha_{n+1}^{\sigma\delta}v_{\sigma\delta}^{n+1,n'},\\
&\pi(E_{-\alpha})v_{\sigma\delta}^{nn'}=2i\alpha_n^{\sigma\delta}v_{\sigma\delta}^{n-1,n'}\quad\text{where}\quad \alpha_{n+1}^{\sigma\delta}=[\lambda_n^{\sigma\delta}(\lambda_{n+1}^{\sigma\delta})-l(l+1)]^{1/2},\\
&\pi(E_\beta)v_{\sigma\delta}^{nn'}=2i\alpha_{n'}^{\sigma\delta}v_{\sigma\delta}^{n,n'-1},\quad(\text{with}\ \lambda_n^{\sigma\delta}=-l+n);\\
&\pi(E_{-\beta})v_{\sigma\delta}^{nn'}=2i\alpha_{n'+1}^{\sigma\delta}v_{\sigma\delta}^{n,n'+1}\quad\text{where}\quad \alpha_{n'+1}^{\sigma\delta}=[\lambda_{n'}^{\sigma\delta}(\lambda_{n'+1}^{\sigma\delta})-l'(l'+1)]^{1/2}\\
&\qquad(\text{with}\ \lambda_{n'}^{\sigma\delta}=-l'+n').
\end{aligned}\tag{2.1}$$

$$\pi(E_\delta)v^{nn'}_{\sigma\delta} = -i[(\lambda^{\sigma\delta}_n - l)(\lambda^{\sigma\delta}_{n'} + l')]^{1/2} a_{\sigma\delta} v^{n,n'-1}_{\sigma+1,\delta} - i[(\lambda^{\sigma\delta}_n + l + 1)(\lambda^{\sigma\delta}_{n'} - l' - 1)]^{1/2} b_{\sigma\delta} v^{n+1,n'}_{\sigma-1,\delta}$$
$$- i[(\lambda^{\sigma\delta}_n + l + 1)(\lambda^{\sigma\delta}_{n'} + l')]^{1/2} c_{\sigma\delta} v^{n+1,n'-1}_{\sigma,\delta-1}$$
$$- i[(\lambda^{\sigma\delta}_n - l)(\lambda^{\sigma\delta}_{n'} - l' - 1)]^{1/2} d_{\sigma\delta} v^{n,n'}_{\sigma,\delta+1}, \tag{2.2}$$

where $a_{\sigma\delta}, b_{\sigma\delta}, c_{\sigma\delta}, d_{\sigma\delta} \in \mathbb{C}$, given for $a_{\sigma\delta}$ by

$$a_{\sigma\delta} = \left[\frac{(\sigma+s)(\sigma-s-1)(\sigma-E_0+1)(\sigma+E_0-2)}{(\sigma+s)(\sigma+\delta-1)(\sigma-\delta)(\sigma-\delta-1)} \right]^{1/2}.$$

And similar relations for $\pi(E_{-\delta})$, $\pi(E_{\pm\gamma})$ (see Ref. 5).

We shall need to write the decomposition of V corresponding to the restriction $U_{|\tilde{K}}$, $V = \oplus_{(j,\epsilon)} V_{j\epsilon}$, where $V_{j\epsilon}$ is the $2j+1$ dimensional weight space corresponding to the weight (j,ϵ).

We denote by $(\phi^m_{j\epsilon})_{m=-j,\ldots,j}$ an othonormal basis of $V_{j\epsilon}$.[5,6]

Let us recall,

$$\pi(H_1)\phi^m_{j\epsilon} = m\phi^m_{j\epsilon}, \quad \pi(H_2)\phi^m_{j\epsilon} = \epsilon\phi^m_{j\epsilon} \tag{2.3}$$

and $\pi(Q_{L_1})\phi^m_{j\epsilon} = j(j+1)\phi^m_{j\epsilon}$.

Let us now introduce a few notations. A Gupta–Bleuler triplet is given by an indecomposable unitary representation of $\tilde{\mathbf{G}}$ $(U, \mathscr{H}')$ in a Hilbert space $\mathscr{H}'$ defined in the following way.

We note V' the space of $\mathbf{k}$-finite vectors of $\rho = DU$ and V, V_1 the submodules occurring in the decomposition of V', $V_1 \subset V \subset V'$,

$$(\rho, V') = (\pi_3, V_3) \to (\pi_2, V_2) \to (\pi_1, V_1)$$

with $V_2 \simeq V/V_1$ and $V_3 \simeq V'/V$; the $\mathbf{g}$-simple submodules (π_3, V_3) and (π_1, V_1) being conjugated.[7]

Massless Gupta–Bleuler triplets [that is with (π_1, V_1) of type $D(s+2, s-1)$ or $D(2-s, s)$] appear at limit of the discrete series $E_0 = s+1$ ($s \geqslant 1$, integer or half-integer).

$\phi^m_{ij\epsilon}$ [$(\phi^m_{ij\epsilon})$ is the basis of V_i as defined in (2.3)] can be realized as homogeneous and symmetric tensors (s integer) or tensors-spinors (s half-integer) on De Sitter space, solutions of gauge-fixed wave equations.[4] For some minimal gauge fixing the representation ρ reduces to

$$D(s+2, s-1) \to D(s+1, s) \to D(s+2, s-1) \quad \text{or} \tag{2.4}$$

$$D(2-s,s) \nearrow D(s+1,s) \searrow \quad D(2-s,s). \quad D(2-s,s) \searrow D(1-s,s-1) \nearrow \tag{2.5}$$

Given (π_1, V_1) and (π_3, V_3) [respectively, (π_2, V_2)] of type $D(s+2, s-1)$ [respectively, $D(s+1, s)$]. We shall characterize extensions of the form $\rho = \pi_3 \to \pi_2 \to \pi_1$, which remain Harish–Chandra modules; consequently it is enough to restrict to relative cohomology.

Let $Z^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$ [respectively, $B^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$] be the space of cocycles (respectively, coboundaries) vanishing on $\mathbf{h}$, the Cartan subalgebra of $\mathbf{g}$ [$\mathscr{L}(V_2, V_1)$ is the space of continuous linear mapping from V_2 into V_1].

We note $H^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$ the coset $Z^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}} / B^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$.[8] Similarly $H^1(\mathbf{g}, \mathscr{L}(V_3, V))^{\mathbf{h}}$ denotes the relative first cohomological space of $\mathbf{g}$-mod. $\mathbf{h}$, with coefficients in $\mathscr{L}(V_3, V)$ with $V = V_1 \oplus V_2$.

Characterization of $H^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$ and $H^1(\mathbf{g}, \mathscr{L}(V_3, V))^{\mathbf{h}}$ have been carried on only so far as to establish results relevant for Sec. III. We denote $\{v^{nn'}_{i\sigma\delta}\}$ and $\{\phi^m_{ij\epsilon}\}$ the relative basis of V_i, $i = 1,2,3$ as defined above (2.3). First we show that $H^1(\mathbf{g}, \mathscr{L}(V_2, V_1))^{\mathbf{h}}$ is generated by cocycles ξ satisfying,

$$\xi_{E_\beta} = \xi_{E_{-\alpha}} = \xi_{E_{-\gamma}} = 0 \quad \xi_{|L_1} = 0,$$

$$\xi_{E_{-\beta}} v^{00}_{2s+1,s} = v^{00}_{1s+2,s-1}.$$

We have $\pi_2(H_2)v^{nn'}_{2\sigma\delta}=(\sigma+n+n')v^{nn'}_{2\sigma\delta}$ and $\pi_2(H_1)v^{nn'}_{2\sigma\delta}=(\sigma+n+n')v^{nn'}_{2\sigma\delta}$. From commutations relations for $E_{-\beta}$ and H_1,H_2, one has $[H_1,\xi_{E_{-\beta}}]=-\xi_{E_{-\beta}}$ and $[H_2,\xi_{E_{-\beta}}]=\xi_{E_{-\beta}}$. Therefore one deduces that

$$\pi_1(H_2)\xi_{E_{-\beta}}v^{00}_{2s+1,s}=(s+2)\xi_{E_{-\beta}}v^{00}_{2s+1,s} \quad \text{and} \quad \pi_1(H_1)\xi_{E_{-\beta}}v^{00}_{2s+1,s}=(s-1)\xi_{E_{-\beta}}v^{00}_{2s+1,s}.$$

From commutation relations for $E_{-\beta}$ and H,H', we deduce that

$$\pi_1(H)\xi_{E_{-\beta}}v^{00}_{2s+1,s}=2i\left(\frac{2s+1}{2}\right)\xi_{E_{-\beta}}v^{00}_{2s+1,s} \quad \text{and} \quad \pi_1(H')\xi_{E_{-\beta}}v^{00}_{2s+1,s}=-2i\left(\frac{3}{2}\right)\xi_{E_{-\beta}}v^{00}_{2s+1,s}.$$

Since $((2s+1)/2,\frac{3}{2})$ is a minimal weight for π_1, we have that $\xi_{E_{-\beta}}v^{00}_{2s+1,s}\in V^{00}_{1s+2,s-1}$. Similarly we have $\xi_{E_\delta}V^m_{2j\epsilon}\subset\oplus_{j<\epsilon}V^{m+1}_{1j\epsilon}$ [with eventually $V^{m+1}_{1j\epsilon}=0$ if (j,ϵ) is not a weight of π_1]. Consider $V^{L_1}_{ij\epsilon}=\pi_i(U(\mathbf{L}_1))\phi^m_{ij\epsilon}$ with $\phi^m_{ij\epsilon}\neq 0$, $i=1,2$; then $\xi_{|_{\mathbf{L}_1}}\in Z^1(\mathbf{L}_1,\mathscr{L}(\oplus_{j<\epsilon}V^{L_1}_{2j\epsilon},\oplus_{j<\epsilon}V^{L_1}_{1j\epsilon}))^{\mathbf{h}_1}$ ($\mathbf{h}_1$ being the Cartan subalgebra of $\mathbf{L}_1$). From Whitehead's lemma, there exist $B_1\in\mathscr{L}(V_2,V_1)$ such that $\xi_{|_{\mathbf{L}_1}}=dB_1$ (d is the coboundary operator).

From commutations relations of $E_{-\gamma}$ with H_1 and H_2, one deduces that $\xi_{E_{-\gamma}}V^m_{2j\epsilon}\subset\oplus_{j<\epsilon-1}V^m_{1j\epsilon-1}$, therefore there exist $B_2\in\mathscr{L}(V_2,V_1)$ defined by $B_2\phi^m_{2j\epsilon}=\Sigma_{j'<\epsilon}b_{j'\epsilon}\phi^m_{1j'\epsilon}$, such that $\xi_{E_{-\gamma}}=[E_{-\gamma},B_2]$. dB_2 vanishes on $\mathbf{h}$ and $\mathbf{L}_1$, thus we can restrict ourselves to cocycles vanishing on $\mathbf{L}_1$ and $E_{-\gamma}$.

Now coboundaries vanishing on $\mathbf{h}$, $\mathbf{L}_1$, and $E_{-\gamma}$ are of the form dB with B diagonal in the basis $\{\phi^m_{ij\epsilon}\}_{i=1,2}$, so $dB(E_{-\beta})v^{00}_{2s+1,s}=0$; consequently $\xi_{E_{-\beta}}v^{00}_{2s+1,s}\neq 0$.

Let us now determine the elements of $H^1(\mathbf{g},\mathscr{L}(V_3,V_1\oplus V_2))^{\mathbf{h}}$ corresponding to the extension $\rho=\pi_3\rightarrow(\pi_2\rightarrow\pi_1)$.

ρ can be written in the following form:

$$\rho=\begin{pmatrix}\pi_1 & \xi & \zeta\\ 0 & \pi_2 & \eta\\ 0 & 0 & \pi_3\end{pmatrix} \quad \text{with} \quad \eta\in Z^1(\mathbf{g},\mathscr{L}(V_3,V_2))^{\mathbf{h}}, \quad \zeta\in Z^1(\mathbf{g},\mathscr{L}(V_3,V_1))^{\mathbf{h}}$$

(Ref. 7). First, ζ defines an element of $H^1(\mathbf{g},\mathscr{L}(V_3,V_1))^{\mathbf{h}}$, but (π_3,V_3) and (π_1,V_1) being equivalent, every restriction of ζ to alternatively $\mathbf{L}$, $\mathbf{L}'$, $\mathbf{L}_1$ and $\mathbf{L}_2$ corresponds to an extension of type $D_k^+\rightarrow D_k^+$ (k integer or half integer) which is inessential.[8]

Thus, up to a coboundary, $\zeta=0$. η defines an element of $H^1(\mathbf{g},\mathscr{L}(V_3,V_2))^{\mathbf{h}}$. We have, in an analogous manner as previously, $\eta_{E_\beta}v^{00}_{3s+2,s-1}=v^{00}_{2s+1,s}$ and $\eta_{E_{-\gamma}}=\eta_{E_{-\beta}}=0$ as well as $\eta_{|_{\mathbf{L}_1}}=0$.

We can summarize,

Proposition 1: All nontrivial extensions $\pi_3\rightarrow\pi_2\rightarrow\pi_1$ with (π_1,V_1), of type $D(s+2,s-1)$ and (π_2,V_2) of type $D(s+1,s)$, are equivalent to an extension of the following form:

$$\rho=\begin{pmatrix}\pi_1 & \xi & 0\\ 0 & \pi_2 & \eta\\ 0 & 0 & \pi_3\end{pmatrix} \quad \textit{with} \quad \xi_X\in\mathscr{L}(V_2,V_1), \quad \eta_X\in\mathscr{L}(V_3,V_2) \quad \textit{for every } X \textit{ in } \mathbf{g}.$$

ξ, η satisfying

$$\xi_{E_{-\beta}}v^{00}_{2s+1,s}=v^{00}_{1s+2,s-1}, \quad \xi_{E_\beta}=\xi_{E_{-\alpha}}=\xi_{E_{-\gamma}}=\xi_{|_{\mathbf{L}_1}}=0,$$

and

$$\eta_{E_\beta}v^{00}_{3s+2,s-1}=v^{00}_{2s+1,s}, \quad \eta_{E_{-\beta}}=\eta_{E_{-\gamma}}=\eta_{|_{\mathbf{L}_1}}=0.$$

$v^{00}_{1s+2,s-1}$, $v^{00}_{2s+1,s}$, and $v^{00}_{3s+2,s-1}$ are cyclic vectors for, respectively, π_1, $\pi_2\rightarrow\pi_1$, and ρ.

For instance a realization of the Gupta–Bleuler triplet (2.4) for $s=1$, corresponding to electrodynamics is given in Ref. 9, with $\xi_{E_{-\beta}} v^{00}_{2s+1,s} = v^{00}_{1s+2,s-1}$ and $\eta_{E_\beta} v^{00}_{3s+2,s-1} = v^{00}_{2s+1,s}$, others terms being null.

Remark: We had previously developed with an irreducible representation π_2 [of type $D(s+1,s)$] but similar arguments extend to the case of a quasisimple (Ref. 10, p. 257) representation. In particular that includes the second triplet (2.5).

III. DEFORMATIONS OF MASSLESS GUPTA-BLEULER TRIPLETS

Let (ρ, V') be a massless Gupta–Bleuler triplet, it can be written in the form

$$\rho = \begin{pmatrix} \pi_1 & \xi & 0 \\ 0 & \pi_2 & \eta \\ 0 & 0 & \pi_3 \end{pmatrix}$$

with $\xi_X \in \mathscr{L}(V_2, V_1)$, $\eta_X \in \mathscr{L}(V_3, V_2)$ for every X in $\mathbf{g}$ $(V' = V_1 \oplus V_2 \oplus V_3,\ V = V_1 \oplus V_2)$. (π_1, V_1), (π_3, V_3) being conjugated and (π_2, V_2) quasisimple. $V^\perp$ will designate the orthogonal of V with respect to the $\tilde{\mathbf{G}}$-invariant inner product in V'.

We now turn to deform the representation (ρ, V'). We denote by $Z^1(\mathbf{g}, \mathscr{L}(V'))$ [respectively, $B^1(\mathbf{g}, \mathscr{L}(V'))$] the space of cocycles (respectively, coboundaries), and by $H^1(\mathbf{g}, \mathscr{L}(V'))$ the first-order cohomological space of $\mathbf{g}$, with coefficients in $\mathscr{L}(V')$ [$\mathscr{L}(V')$ is the space of continuous linear mappings]. We restrict to deformations ρ^λ ($\lambda \in \mathbb{C}$) of ρ which remain Gupta–Bleuler triplets (with components in the discrete or supplementary series).

Looking first to the infinitesimal parts, we shall determine corresponding elements of $H^1(\mathbf{g}, \mathscr{L}(V'))$. Next we shall show that massless gauge theories are stable, as far as deformations to gauge theories that are of concern:

Proposition 2: Let us take $\rho = \pi_3 \to \pi_2 \to \pi_1$, with π_1 of type $D(s+2,s-1)$ or $D(2-s,s)$. Given a deformation $(\rho^\lambda, V'^\lambda)$ of (ρ, V'), integrable to a deformation U^λ of U. Suppose that $(\rho^\lambda, V'^\lambda)$ admits a simple $\mathbf{g}$-submodule V_1^λ of V'^λ with $V_1^{\lambda\perp} \neq V_1^\lambda$, such that restriction θ_1^λ of ρ^λ to V_1^λ is a Verma module and there are no $\mathbf{g}$-invariant complement of V_1^λ in V'^λ.

Then the representation ρ^λ is equivalent to the representation ρ.

In other words, there are no deformations of ρ with a kernel.

Proof: We assume that π_2 is irreducible of type $D(s+1,s)$ and ρ defined by proposition 1 but the same arguments could be repeated with a quasisimple representation π_2.

Let V'^λ be the subspace of $\mathbf{k}$-finite vectors of ρ^λ. Up to equivalence we have $V'^\lambda = V'$ for every λ.[11] Since $V_1^{\lambda\perp} \neq V_1^\lambda$, ρ^λ is a Gupta–Bleuler triplet; there is a submodule V^λ of V', $V_1^\lambda \subset V^\lambda \subset V'$ such that V^λ / V_1^λ is a simple Verma module and: the quotient representation $(\rho^\lambda, V'/V^\lambda)$ is conjugated to $(\rho^\lambda, V_1^\lambda)$.[7]

Actually V_1^λ, V^λ / V_1^λ and V'/V^λ are deformations of the irreducible representations (π_1, V_1), (π_2, V_2) and (π_3, V_3), respectively. So let us denote the representations obtained, restricting ρ^λ to successively V_1^λ, V^λ, and V' modulo the corresponding invariant subspaces, by $(\pi_1^\lambda, V_1^\lambda)$, $(\pi_2^\lambda, V_2^\lambda)$, and $(\pi_3^\lambda, V_3^\lambda)$. Thus ρ^λ is of the form

$$\rho^\lambda = \begin{pmatrix} \pi_1^\lambda & u_{12}^\lambda & u_{13}^\lambda \\ 0 & \pi_2^\lambda & u_{23}^\lambda \\ 0 & 0 & \pi_3^\lambda \end{pmatrix} \quad \text{with} \quad V' = V_1^\lambda \oplus V_2^\lambda \oplus V_3^\lambda .$$

Since we restrict ourselves to the ρ^λ which remain Harish–Chandra modules, we have

$$u_{12}^\lambda \in Z^1(\mathbf{g}, \mathscr{L}(V_2^\lambda, V_1^\lambda))^{\mathbf{h}}, \quad u_{13}^\lambda \in Z^1(\mathbf{g}, \mathscr{L}(V_3^\lambda, V_1^\lambda))^{\mathbf{h}},$$

and $u_{23}^\lambda \in Z^1(\mathbf{g}, \mathscr{L}(V_3^\lambda, V_2^\lambda))^{\mathbf{h}}$. As a submodule of (ρ^λ, V'), $(\rho^\lambda{}_{|V^\lambda}, V^\lambda)$ is a deformation of $(\rho_{|V}, V)$; as well $(\rho^\lambda{}_{|V_1^\lambda}, V_1^\lambda)$ is a deformation of $(\rho_{|V_1}, V_1)$.

Therefore up to an isomorphism (which will not be unitary in general) we can suppose that (ρ^λ, V') decomposes as $V'=V\oplus V_3$ with $V=V_1\oplus V_2$ (that is $V_1^\lambda \simeq V_1$, $V_2^\lambda \simeq V_2$ and $V_3^\lambda \simeq V_3$ as subspaces of V'). Let us write $\rho^\lambda(X)$, $X\in\mathbf{g}$ as

$$\rho^\lambda(X)=\rho(X)+\lambda C(X)+\sum_{n\geqslant 2}\lambda^n C_n(X) \quad \text{with } C(X),C_n(X)\in\mathscr{L}(V').$$

ρ^λ is integrable to a deformation U^λ of U as a representation of $\tilde{\mathbf{G}}$ in $\mathscr{H}'$ (Ref. 11, proposition 9).

We shall characterize the elements C in $Z^1(\mathbf{g},\mathscr{L}(V'))$ corresponding to such deformations. First, we show that we can restrict ourselves to cocycles C vanishing on $\mathbf{L}\oplus\mathbf{L}'$ [in other words we show that C vanishes on C_3, the commutant of so(3) in $U(\mathbf{g})$, Ref. 6],

$$\rho^\lambda=\pi_3^\lambda\to\pi_2^\lambda\to\pi_1^\lambda \quad \text{where } \pi_i^\lambda \text{ are of type } D(E_0^\lambda,s^\lambda).$$

Let $\mathbf{G}_{L\oplus L'}$ be a subgroup of $\mathbf{G}$, connected, closed, associated to $\mathbf{L}\oplus\mathbf{L}'$ ($\mathbf{G}_{L\oplus L'}\simeq \mathrm{SO}(2,2)$). $\mathbf{G}_{L\oplus L'}$ has a maximal compact subgroup $\mathbf{K}_{L\oplus L'}$, thus considering the restriction $C_{|_{\mathbf{L}\oplus\mathbf{L}'}}$ of C to $\mathbf{L}\oplus\mathbf{L}'$, $C_{|_{\mathbf{L}\oplus\mathbf{L}'}}$ defines an element of $H^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V'))$ integrable to an element of $H^1(\mathbf{G}_{L\oplus L'},\mathscr{L}(\mathscr{H}'))$ (Ref. 8, proposition 7.13). But, ρ being integrable to a representation U of $\tilde{\mathbf{G}}$ and $\mathbf{G}$, $H^1(\mathbf{G}_{L\oplus L'},\mathscr{L}(\mathscr{H}'))=H^1(\mathbf{G}_{L\oplus L'},\mathscr{L}(\mathscr{H}'))^{\mathbf{K}_{L\oplus L'}}$. So $C_{|_{\mathbf{L}\oplus\mathbf{L}'}}$ belongs—up to a coboundary—to $Z^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V'))^{\mathbf{h}}$. Consequently we can restrict to cocycles C vanishing on $\mathbf{h}$.

We have

$$V'=\bigoplus_{\substack{(\sigma,\delta)\\ i=1,2,3}} V_{i\sigma\delta}$$

[$V_{i\sigma\delta}$ being the null space if (σ,δ) is not a weight of π_i] with $V_{i\sigma\delta}=\pi_i(U(\mathbf{L}\oplus\mathbf{L}'))v_{i\sigma\delta}^{00}$. Let us consider the following Verma $\mathbf{L}\oplus\mathbf{L}'$-modules:

$$F_{i\sigma\delta}=\rho(U(\mathbf{L}\oplus\mathbf{L}'))v_{i\sigma\delta}^{00}, \quad i=1,2,3.$$

One has $F_{1\sigma\delta}\subset V_1$, $F_{2\sigma\delta}\subset V_1\oplus V_2$, and $F_{3\sigma\delta}\subset V_1\oplus V_2\oplus V_3$. We can show that $H^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(F_{i\sigma\delta}))^{\mathbf{h}}=\{0\}$, but every extension appearing in $\rho_{|_{\mathbf{L}\oplus\mathbf{L}'}}$ being trivial,[8] to simplify, we can assume that $F_{i\sigma\delta}=V_{i\sigma\delta}$, $i=1,2,3$. (ρ^λ,V') admits a decomposition relative to $\mathbf{L}\oplus\mathbf{L}'$ as

$$V'=\bigoplus_{\substack{i=1,2,3\\ (\sigma(\lambda),\delta(\lambda))}} V_{i\sigma(\lambda)\delta(\lambda)},$$

where $V_{i\sigma(\lambda)\delta(\lambda)}=\rho^\lambda((U(\mathbf{L}\oplus\mathbf{L}'))v_{i\sigma\delta}^{00}$ (denoting $(\sigma(\lambda),\delta(\lambda))$ the weights of $\rho^\lambda{}_{|_{\mathbf{L}\oplus\mathbf{L}'}}$). Considering

$$\left(\rho^\lambda{}_{|_{\mathbf{L}\oplus\mathbf{L}'}}, \bigoplus_{i'=1,\ldots,i} V_{i'\sigma(\lambda)\delta(\lambda)}\right)$$

as a deformation of

$$\left(\rho_{|_{\mathbf{L}\oplus\mathbf{L}'}}, \bigoplus_{i'=1,\ldots,i} V_{i'\sigma\delta}\right),$$

we can suppose—up to an isomorphism—that $V_{i\sigma(\lambda)\delta(\lambda)}=V_{i\sigma\delta}$ $(i=1,2,3)$ for every λ.

Let us note $V'_{\sigma\delta}=\oplus_{i=1,2,3}V_{i\sigma\delta}$. Now $C_{|_{\mathbf{L}\oplus\mathbf{L}'}}$ defines an element of

$$H^1\left(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}\left(\bigoplus_{i'=1,\ldots,i} V_{i'\sigma\delta}\right)\right)^{\mathbf{h}}$$

and every extension of type $D_k^+\to D_{k'}^+$ $(k,k'\geqslant\frac{1}{2})$ for $sl(2)$ are inessential[8] ($\mathbf{L}\oplus\mathbf{L}'$—submodules occurring in $(\rho_{|_{\mathbf{L}\oplus\mathbf{L}'}},V_{i\sigma\delta})$ $i=1,2,3$, are of type $D_{\frac{1}{2}(\sigma+\delta)}\oplus D_{\frac{1}{2}(\sigma-\delta)}$).

One deduces that

$$H^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V_{i\sigma\delta},V_{i'\sigma\delta}))^{\mathbf{h}}=\{0\},\quad i,i'=1,2,3;$$

thus $H^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V'_{\sigma\delta}))^{\mathbf{h}}=\{0\}$ for every weight (σ,δ).[12]

V' being the vectorial sum of $\{V'_{\sigma\delta}\}_{(\sigma,\delta)}$, there exist $T\in\mathscr{L}(V')\cap\mathscr{L}(V'_{\sigma\delta})$, for every (σ,δ), such that $C(X)=dT(X)$ for every $X\in\mathbf{L}\oplus\mathbf{L}'$. That is to say, $H^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V'))^{\mathbf{h}}=\{0\}$. Then we can restrict ourselves to cocycles C vanishing on $\mathbf{L}\oplus\mathbf{L}'$, and

$$\rho^\lambda|_{\mathbf{L}\oplus\mathbf{L}'}=\rho|_{\mathbf{L}\oplus\mathbf{L}'}+\lambda^2C_2|_{\mathbf{L}\oplus\mathbf{L}'}+\cdots,\quad\text{with } C_2|_{\mathbf{L}\oplus\mathbf{L}'}\in Z^1(\mathbf{L}\oplus\mathbf{L}',\mathscr{L}(V'))^{\mathbf{h}}.$$

Thus up to equivalence, we can suppose that $C_2|_{\mathbf{L}\oplus\mathbf{L}'}=0$; repeating for $C_3,\ldots$, one has—up to equivalence—$\rho^\lambda|_{\mathbf{L}\oplus\mathbf{L}'}=\rho|_{\mathbf{L}\oplus\mathbf{L}'}$. Therefore ρ^λ is a massless Gupta–Bleuler triplet. In order to show that ρ^λ is isomorphic to ρ, it is enough to prove that $\rho^\lambda(Q_{L_1})=\rho(Q_{L_1})$.

Let us consider the decomposition of V', relative to $\tilde{\mathbf{K}}$,

$$V'=\bigoplus_{\substack{(j,\epsilon)\\ i=1,\ldots,3}} V_{ij\epsilon},$$

where as previously $V_{ij\epsilon}$ denote the $2j+1$ dimensional weight subspace (with basis $\{\phi^m_{ij\epsilon}\}_m$) associated to the weight (j,ϵ) for π_i. Let us note $V^{L_1}_{ij\epsilon}=\pi_i(U(\mathbf{L}_1))\phi^m_{ij\epsilon}$ with $\phi^m_{ij\epsilon}\neq0$ and

$$V'^{L_1}_{ij\epsilon}=\bigoplus_{i'=1,\ldots,i} V^{L_1}_{i'j\epsilon}.$$

Then $(\pi_i|_{\mathbf{L}_1},V^{L_1}_{ij\epsilon})$ is a $2j+1$ dimensional representation of type $D(j)$ [the spectrum of H_1 is $j-n$, $0\leqslant n\leqslant 2j$ and the Casimir value is $j(j+1)$]. Moreover $C(X)V^{L_1}_{ij\epsilon}\subset\oplus_{j'<\epsilon}V'^{L_1}_{ij'\epsilon}$ $X\in\mathbf{L}_1$, $i=1,2,3$, since $[\rho(H_2),C(X)]=0$ if X belongs to $\mathbf{L}_1$. So $C|_{\mathbf{L}_1}$ is an element of $Z^1(\mathbf{L}_1,\mathscr{L}(\oplus_{\substack{j<\epsilon\\ i=1,2,3}}V^{L_1}_{ij\epsilon}))^{\mathbf{h}_1}$ and a coboundary by Whitehead's lemma; actually

$$C|_{\mathbf{L}_1}\in\bigoplus_{\substack{j<\epsilon\\ j'<\epsilon\\ i\geqslant i'}} B^1(\mathbf{L}_1,\mathscr{L}(V^{L_1}_{ij\epsilon},V^{L_1}_{i'j'\epsilon}))^{\mathbf{h}_1}.$$

Finally, $C|_{\mathbf{L}_1}=dT_1$, with $T_1\in\mathscr{L}(V')\cap\mathscr{L}(V'^{L_1}_{ij\epsilon})$, $i=1,2,3$.

Indeed $\mathbf{L}_1$-submodules $V^{L_1}_{ij\epsilon}$, $V^{L_1}_{ij'\epsilon}$ are nonequivalent irreducible representations if $j\neq j'$ [of type $D(j)$, $D(j')$] thus corresponding components of the coboundary dT_1 vanish on Q_{L_1}.[8] Consequently, $D(j)$, $D(j')$ $j\neq j'$, having different Casimir values,

$$T_1(V'^{L_1}_{ij\epsilon})\subset V'^{L_1}_{ij\epsilon},\quad i=1,2,3.$$

In other terms $dT_1(Q_{L_1})=0$ on V'. Then $\rho^\lambda|_{\mathbf{L}_1}=\rho|_{\mathbf{L}_1}+\lambda dT_1+\lambda^2C_2|_{\mathbf{L}_1}+\cdots$. Up to an equivalence of type $(Id+\lambda T_1)$, we can assume that $\rho^\lambda|_{\mathbf{L}_1}=\rho|_{\mathbf{L}_1}+\lambda^2C_2|_{\mathbf{L}_1}+\cdots$. Therefore $C_2\in Z^1(\mathbf{L}_1,\mathscr{L}(V'))^{\mathbf{h}_1}=B^1(\mathbf{L}_1,\mathscr{L}(V'))^{\mathbf{h}_1}$ by Whitehead's lemma. Up to an equivalence of type $(Id+\lambda^2T_2)$, $T_2\in\mathscr{L}(V')\cap\mathscr{L}(V'^{L_1}_{ij\epsilon})$, we have, $\rho^\lambda|_{\mathbf{L}_1}=\rho|_{\mathbf{L}_1}+\lambda^3C_3|_{\mathbf{L}_1}+\cdots$. Now we can repeat the same arguments for $C_3,\ldots$. In this way we construct a formal equivalence, A^λ, of $\rho^\lambda|_{\mathbf{L}_1}$ with $\rho|_{\mathbf{L}_1}$ (actually the equivalence is not formal[11] but we do not need it here) with $A^\lambda\in GL(V')\cap GL(V'^{L_1}_{ij\epsilon})$. Thus $\rho^\lambda(Q_{L_1})=A^{\lambda^{-1}}\rho(Q_{L_1})A^\lambda$, restricting to $V'^{L_1}_{ij\epsilon}$ we obtain $\rho^\lambda(Q_{L_1})|_{V'^{L_1}_{ij\epsilon}}=j(j+1)Id_{V'^{L_1}_{ij\epsilon}}$, that is to say $\rho^\lambda(Q_{L_1})=\rho(Q_{L_1})$ on V'.

Conclusion, $\rho^\lambda|_{V_1}\simeq\rho|_{V_1}$ of type $D(s+2,s-1)$ with $v^{00}_{1s+2,s-1}$ as cyclic vector; the quotient representations of ρ^λ on alternatively V/V_1, V'/V is of type $D(s+1,s)$ and $D(s+2,s-1)$. So $\rho^\lambda\simeq\rho$ as a representation of $\mathbf{g}$.

ACKNOWLEDGMENTS

I would like to thank Professor Moshé Flato for enlightening discussions and constant encouragement and Professor Tullio Valent for his hospitality at the University of Padova.

[1] M. Flato and C. Fronsdal, ''Spontaneously generated field theories, zero-center modules, colored singletons and the virtue of $N=6$ supergravity,'' in *Essays on Supersymmetry*, edited by C. Fronsdal (Reidel, Dordrecht, 1986), pp. 123–162.
[2] E. S. Abers and B. W. Lee, ''Gauge theories,'' Phys. Rep., Phys. Lett. **9**, 1–141 (1973).
[3] C. Fronsdal, ''Smooth massless limit of field theories,'' Nucl. Phys. B **167**, 237–248 (1980).
[4] E. Angelopoulos, M. Flato, C. Fronsdal, and D. Sternheimer, ''Massless particles, conformal group and de Sitter universe,'' Phys. Rev. D **23**, 1278–1289 (1981).
[5] N. T. Evans, ''Discrete series for the universal covering group of the $3+2$ de Sitter group,'' J. Math. Phys. **8**, 170–184 (1967).
[6] E. Angelopoulos, ''$SO_0(3,2)$: Linear and unitary irreducible representations,'' in *Quantum Theory Groups, Fields and Particles*, edited by A. O. Barut (Reidel, Dordrecht, 1983), pp. 101–148.
[7] H. Araki, ''Indecomposable representations with invariant inner product. A theory of Gupta-Bleuler triplet,'' Commun. Math. Phys. **97**, 149–159 (1985).
[8] G. Pinczon and J. Simon, ''Extensions of representations and cohomology,'' Rep. Math. Phys. **16**, 49–77 (1979).
[9] B. Binegar, C. Fronsdal, and W. Heidenreich, ''de Sitter QED,'' Ann. Phys. (N.Y.) **149**, 254–272 (1983).
[10] G. Warner, *Harmonic Analysis on Semi-simple Lie Groups* (Springer, Berlin, 1972).
[11] M. Lesimple and G. Pinczon, ''Deformations of Lie group and Lie algebra representations,'' J. Math. Phys. **34**, 4251–4272 (1993).
[12] A. Guichardet, *Cohomologie des Groupes Topologiques et des Algèbres de Lie* (Cedic/Nathan, Paris, 1980).

Solitary excitations of a two-dimensional electron gas

A. Nerses and E. E. Kunhardt
Department of Physics, Stevens Institute of Technology, Hoboken, New Jersey 07030

(Received 16 March 1998; accepted for publication 12 June 1998)

The nonlinear collective excitations of a two-dimensional electron gas (2DEG) formed at the interface of a heterostructure are presented. A matrix formulation of the coupled particle dynamics–electromagnetic field equations permits the extraction of the equation of evolution for these excitations. The stationary solutions of the equation are presented. A new class of solitary excitations is shown to form part of the nonlinear mode spectrum of excitations of the 2DEG in the low wave-vector plasmon–polariton regime. © *1998 American Institute of Physics.*
[S0022-2488(98)04010-9]

I. INTRODUCTION

Interest in the study of the linear plasmon spectrum of low-dimensional electron gas semiconductor systems has grown considerably in the last decade due to technological advances in the fabrication of submicron, quantum structures. This interest has been driven both by the challenge to fabricate and probe devices based on these low-dimensional structures, and by the search for novel behavior and/or modified behavior from bulk semiconductor systems. The dispersion relation for the plasmons,[1–11] as well as the coupling of these modes to phonons and photons,[8,12–20] in these low-dimensional systems has been studied extensively both experimentally and theoretically, using linearized hydrodynamic models,[10,11,20,21] and quantum many-body formulations.[1,3,18] The results from these models yield similar behavior in the low ω, k regime, where $k<k_F$, the Fermi wave vector.[10] Thus far in the literature, there has been no treatment of the nonlinear behavior of these excitations. Furthermore, the nonlinear regime has not been experimentally accessible due to the weak coupling schemes used to excite the plasmons, including prism and grating couplers.

In this paper, we present a theoretical investigation of the nonlinear excitations in a two-dimensional electron gas (2DEG), e.g., formed at the interface of an AlGaAs/GaAs heterostructure. We find a new class of solitary excitations to form part of the nonlinear mode spectrum in the low wave-vector regime, where the plasmon–polaritons have a strong photon content.

Our objective is to obtain an equation of evolution in space for the excitation amplitudes since this description is suitable for the analysis of experiments in which the generation and detection of the excitations occur at two different spatial locations. For this purpose, we present a novel approach which involves: (1) a matrix formulation of coupled particle dynamics–electromagnetic field equations, (2) transformation of the coupled system of equations into the frequency domain in order to isolate the plasmon spectrum, and (3) transformation back into the time domain, thereby obtaining an equation of evolution in space–time for the amplitude of the excitations. This approach permits the analysis of nonlinear excitations containing a wide spectrum of frequencies (such as pulse excitations). This program is carried out in detail in Sec. III, and the resulting equation of evolution presented in Sec. IV. From the equation of evolution, we find a new class of solitary excitations as part of the nonlinear mode spectrum of the 2DEG. The matrix formulation and the properties of the excitations are presented in Sec. II.

II. MODEL FOR DESCRIBING ELECTRON DYNAMICS

The system under consideration is a 2DEG, e.g., which forms at the interface of an AlGaAs/GaAs heterostructure. The interface in which electron motion is free will be considered as the $y-z$ plane, while the x-axis will denote the direction perpendicular to this plane, where electron motion is confined. For simplicity, the finite width of the potential well at the interface is neglected and the wave function for the 2D system of electrons is assumed to be exactly confined in the $y-z$

 © 1998 American Institute of Physics

plane of the interface, with no extension outside of the plane. This model provides a good description for the nonlinear excitations at low temperatures, for which only the lower subband is occupied.

The dynamics of the 2DEG is described via a hydrodynamic model, with the electrons possessing an isotropic mass and the ions are assumed to form a compensating uniform background. The variables that comprise the hydrodynamic model are: the 2D electron density $n(\mathbf{r},t)$ and the current density $\mathbf{J}(\mathbf{r},t)$. The electron density consists of two contributions, the time-independent background n_0 and time-dependent $n_1(\mathbf{r},t)$; that is,

$$n(\mathbf{r},t)=n_0+n_1(\mathbf{r},t), \tag{1}$$

where $n_1(\mathbf{r},t)<n_0$. For charges confined to a single layer, the plane $x=0$, the three-dimensional number density and current density become singular and can be expressed as follows:

$$n(\mathbf{r},t)=n(z,t)\,\delta(x) \tag{2}$$

and

$$\mathbf{J}(\mathbf{r},t)=J_z(z,t)\,\delta(x)\tilde{\mathbf{z}}, \tag{3}$$

where $\mathbf{z}$ is a unit vector in the propagation direction (in the plane of the 2DEG).

These variables satisfy the continuity equation and the Euler equation of motion,[20] namely:

$$\partial_t(-en(\mathbf{r},t))+\boldsymbol{\nabla}\cdot\mathbf{J}(\mathbf{r},t)=0, \tag{4}$$

$$\partial_t\mathbf{J}(\mathbf{r},t)-\frac{1}{en_0}\,\boldsymbol{\nabla}\cdot(\mathbf{J}(\mathbf{r},t)\mathbf{J}(\mathbf{r},t))=\frac{e^2}{m}\,n(\mathbf{r},t)\mathbf{E}(\mathbf{r},t)+e\beta^2\boldsymbol{\nabla}n(\mathbf{r},t)-\frac{1}{\tau}\,\mathbf{J}(\mathbf{r},t), \tag{5}$$

where β is the spatial dispersion parameter with units of speed, and τ is the momentum relaxation time. The terms on the right-hand side of Eq. (5) contribute to the acceleration of the electrons due to density gradients and an applied electric field, and to the decay of the excitations.

The nonlinear effects arise from the $\boldsymbol{\nabla}\cdot(\mathbf{J}(\mathbf{r},\mathbf{t})\mathbf{J}(\mathbf{r},\mathbf{t}))$ and $n(\mathbf{r},t)\mathbf{E}(\mathbf{r},t)$ terms in Eq. (5). The $\boldsymbol{\nabla}\cdot(\mathbf{J}(\mathbf{r},\mathbf{t})\mathbf{J}(\mathbf{r},\mathbf{t}))$ term is due to the tendency of elements of the 2DEG with different average current density to compress or expand, whereas the $n(\mathbf{r},t)\mathbf{E}(\mathbf{r},t)$ term is due to the dependence on the local density of the force driving a change in the average current density. Thus far in the literature where the hydrodynamic model has been applied to one-dimensional (1D) and 2D systems,[2,10,20] Eq. (5) is linearized for the purpose of obtaining the linear plasmon spectrum.

The field variables, $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$, obey Maxwell's equations:

$$\begin{aligned}\boldsymbol{\nabla}\times\mathbf{E}(\mathbf{r},t)&=-\partial_t\mathbf{B}(\mathbf{r},t),\\ \boldsymbol{\nabla}\times\mathbf{B}(\mathbf{r},t)&=\epsilon\partial_t\mathbf{E}(\mathbf{r},t)+\mathbf{J}(\mathbf{r},t),\\ \boldsymbol{\nabla}\cdot\epsilon\mathbf{E}(\mathbf{r},t)&=\rho,\qquad \boldsymbol{\nabla}\cdot\mathbf{B}(\mathbf{r},t)=0,\end{aligned} \tag{6}$$

where ρ is the charge density, $\rho=-en_1(\mathbf{r},t)$. For our purposes, Maxwell's equations are expressed through the vector potential $\mathbf{A}(\mathbf{r},t)$ in the gauge where the scalar potential vanishes. In this gauge, the vector potential satisfies the following wave equation:

$$[\boldsymbol{\nabla}^2-\mu_0\epsilon\partial_t^2]\mathbf{A}(\mathbf{r},t)=-\mu_0\mathbf{J}(\mathbf{r},t)\tilde{\mathbf{z}}-\frac{e}{i\omega\epsilon}\,\partial_z n(\mathbf{r},t)\tilde{\mathbf{z}} \tag{7}$$

and is related to the electric and magnetic fields as follows:

$$\mathbf{E}(\mathbf{r},t)=-\partial_t\mathbf{A}(\mathbf{r},t),\qquad \mathbf{B}(\mathbf{r},t)=\boldsymbol{\nabla}\times\mathbf{A}(\mathbf{r},t). \tag{8}$$

To simplify the system of equations, ϵ, the background dielectric constant, will be used to characterize both the AlGaAs and GaAs surrounding the 2DEG.

The system of Eqs. (4)–(8) forms the basis for the investigations of the nonlinear excitations. This coupled system of equations will be further simplified by neglecting variations in the y-direction and assuming that β and τ are velocity independent. For the present, we choose to emphasize the characteristics of the nonlinear excitations by avoiding mathematical complexities due to the inclusion of microscopic phenomena (such as exchange and correlation effects, and wave–particle interactions) that can obscure the nature of these excitations. Furthermore, this simplification of the model contains sufficient information to describe the nonlinear plasmon excitations in the $k<k_F$ regime. For the future, the formulation presented below can be extended to include more complete descriptions of the electron gas.

III. MATRIX FORMULATION OF THE SYSTEM OF EQUATIONS

To obtain the equation of evolution in space for the amplitude of the nonlinear excitations, Eqs. (4), (5), and (7) are cast in a matrix form. To this end, the gradient operator is decomposed as

$$\boldsymbol{\nabla}=\boldsymbol{\nabla}_t+\partial_z\tilde{\mathbf{z}}, \tag{9}$$

where $\boldsymbol{\nabla}_t$ is the transverse part of the gradient operator, with respect to the propagation direction, $\tilde{\mathbf{z}}$, and the equations are transformed into the frequency domain to facilitate the extraction of the plasmon excitations. With these observations, Eqs. (4), (5), and (7) can be expressed as:

$$-i\partial_z^2 J_z(z,\omega)+\omega e\partial_z n_1(z,\omega)=0, \tag{10}$$

$$-\frac{\beta^2}{i\omega}\partial_z^2 J_z(z,\omega)\tilde{\mathbf{z}}+i\omega\left(1+\frac{i}{\omega\tau}\right)J_z(z,\omega)\tilde{\mathbf{z}}+e^2\frac{n_0}{m}i\omega\mathbf{A}(0,z,\omega)=\mathrm{NL}\tilde{\mathbf{z}}, \tag{11}$$

$$(\partial_z^2+\partial_x^2+\omega^2\mu_0\epsilon)A_z(x,z,\omega)\tilde{\mathbf{z}}+\mu_0\delta(x)J_z(z,\omega)\tilde{\mathbf{z}}+\frac{1}{\omega^2\epsilon}\delta(x)\partial_z^2 J_z(z,\omega)\tilde{\mathbf{z}}=0, \tag{12}$$

$$-\partial_z^2\partial_x A_z(x,z,\omega)\tilde{\mathbf{x}}+\omega^2\mu_0\epsilon\partial_z A_x(x,z,\omega)\tilde{\mathbf{x}}+\partial_z^2(\partial_z A_x(x,z,\omega))\tilde{\mathbf{x}}=0, \tag{13}$$

where NL represents the nonlinear term

$$\begin{aligned}\mathrm{NL}=&-\frac{2}{en_0}\int\int d\omega_1 d\omega_2\delta(\omega-\omega_1-\omega_2)J_z(z,\omega_1)\partial_z J_z(z,\omega_2)\\&+\frac{e}{m}\int\int d\omega_1 d\omega_2\delta(\omega-\omega_1-\omega_2)\frac{\omega_1}{\omega_2}A_z(0,z,\omega_1)\partial_z J_z(z,\omega_2)\end{aligned} \tag{14}$$

and ω represents the transformation variable in time; that is, $G(z,\omega)=\int dt e^{i\omega t}G(z,t)$ with $G(z,t)$ corresponding to any of the independent variables in the model. Note that the wave equation (7) has been written in component form to arrive at Eqs. (12) and (13).

Equations (10)–(13) can be recast in matrix form as a nonlinear equation of evolution along the **z**-direction:

$$\Gamma\partial_z^2\psi(\mathbf{r},\omega)+\mathbf{L}(\boldsymbol{\nabla}_t,\omega)\psi(\mathbf{r},\omega)=\mathrm{NL}[\psi(\mathbf{r},\omega)], \tag{15}$$

where

$$\mathbf{L}(\boldsymbol{\nabla}_t,\omega)=\begin{bmatrix}\frac{1}{c}\{(\partial_x^2+\omega^2\mu_0\epsilon)\tilde{\mathbf{z}}\tilde{\mathbf{z}}-\partial_x\tilde{\mathbf{z}}\tilde{\mathbf{x}}\}+\omega^2\mu_0\epsilon\tilde{\mathbf{x}}\tilde{\mathbf{x}} & \omega\mu_0\delta(x)\tilde{\mathbf{z}}\tilde{\mathbf{z}}\\ \omega\mu_0\delta(x)\tilde{\mathbf{z}}\tilde{\mathbf{z}} & \delta(x)\omega^2 c\frac{\mu_0 m}{e^2 n_0}\left[\left(1+\frac{i}{\tau\omega}\right)\tilde{\mathbf{z}}\tilde{\mathbf{z}}-\frac{e^2\beta^2}{\kappa_z^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}\right]\end{bmatrix}, \tag{16}$$

$$\Gamma=\begin{bmatrix} -(\partial_x\tilde{\mathbf{xz}}-\tilde{\mathbf{xx}}) & 0 \\ 0 & \delta(x)c\,\dfrac{\mu_0 m\beta^2}{e^2 n_0}\left(\tilde{\mathbf{zz}}-\dfrac{e\omega}{\kappa_z^2}\,\tilde{\mathbf{xz}}\right) \end{bmatrix}, \tag{17}$$

$$\psi(\mathbf{r},\omega)=\begin{bmatrix} \omega(A_z(x,z,\omega)\tilde{\mathbf{z}}+\partial_z A_x(x,z,\omega)\tilde{\mathbf{x}}) \\ \dfrac{1}{c}\,(J_z(z,\omega)\tilde{\mathbf{z}}+i\partial_z n(z,\omega)\tilde{\mathbf{x}}) \end{bmatrix}, \tag{18}$$

and

$$\mathrm{NL}=-\frac{i\omega\mu_0 m}{e^2 n_0}\,\delta(x)\tilde{\mathbf{z}}\left[\int\int d\omega_1 d\omega_2\,\delta(\omega-\omega_1-\omega_2)\begin{Bmatrix} 0 \\ \dfrac{2}{en_0}\,J_{z\omega_1}-\dfrac{e}{m}\,\dfrac{\omega_1}{\omega_2}\,A_{z\omega_1} \end{Bmatrix}\partial_z J_{z\omega_2}\right], \tag{19}$$

where **1** is the unit dyad; **L** is a linear operator that depends only on frequency and the component of the gradient operator, $\boldsymbol{\nabla}_t$, transverse to the direction of propagation of the plasmon excitations (i.e., the z-direction).

Operator **L** is decomposed into a lossless, Hermitian operator, $L_0=L_{\tau=\infty}$, and a smaller perturbative term L_{coll}, such that $L=L_0+L_{\mathrm{coll}}$ and $L_0\gg\mathbf{L}_{\mathrm{coll}}$:

$$L_0(\boldsymbol{\nabla}_t,\omega)=\begin{bmatrix} \dfrac{1}{c}\,\{(\partial_x^2+\omega^2\mu_0\epsilon)\tilde{\mathbf{zz}}-\partial_x\tilde{\mathbf{zx}}\}+\omega^2\mu_0\epsilon\tilde{\mathbf{xx}} & \omega\mu_0\delta(x)\tilde{\mathbf{zz}} \\ \omega\mu_0\delta(x)\tilde{\mathbf{zz}} & \delta(x)\omega^2 c\,\dfrac{\mu_0 m}{e^2 n_0}\left[\tilde{\mathbf{zz}}-\dfrac{e^2\beta^2}{\kappa_z^2}\,\tilde{\mathbf{xx}}\right] \end{bmatrix} \tag{20}$$

and L_{coll}:

$$\mathbf{L}_{\mathrm{coll}}=\begin{bmatrix} 0 & 0 \\ 0 & \delta(x)c\,\dfrac{\mu_0 m}{e^2 n_0}\,\dfrac{i\omega}{\tau}\,\tilde{\mathbf{zz}} \end{bmatrix}. \tag{21}$$

Solutions to (15) are sought in the form of an expansion in terms of the eigenfunctions of the linear, lossless operator, $\mathbf{L}_0$. That is, $\Psi(\mathbf{r},\omega)$ lies in the eigenspace of $\mathbf{L}_0$ and thus can be represented as

$$\Psi(\mathbf{r},\omega)=\sum_\alpha\,\Phi_\alpha(\mathbf{x},\omega)a_\alpha(z,\omega), \tag{22}$$

where $\boldsymbol{\Phi}_\alpha(\mathbf{x},\omega)$ are the eigenfunctions of $\mathbf{L}_0$ defined as

$$\mathbf{L}_0(-i\omega,\partial_x)\Phi_\alpha(\mathbf{x},\omega)=\kappa_{z\alpha}^2(\omega)\Gamma\boldsymbol{\Phi}_\alpha(\mathbf{x},\omega). \tag{23}$$

The eigenfunctions $\Phi_\alpha(\mathbf{x},\omega)$ possess the orthonormality property

$$(\Phi_\alpha(\mathbf{x},\omega),\Gamma\Phi_\beta(\mathbf{x},\omega))=2\,\delta_{\alpha\beta} \tag{24}$$

in the plane perpendicular to **z** and have the form

$$\Phi_\alpha(\mathbf{x},\omega)=\begin{bmatrix} \omega(\varphi^{A_z}\tilde{\mathbf{z}}+\varphi^{\partial_z A_x}\tilde{\mathbf{x}})_\alpha \\ \dfrac{1}{c}\,(\varphi^{J_z}\tilde{\mathbf{z}}+i\varphi^{\partial_z n}\tilde{\mathbf{x}})_\alpha \end{bmatrix}, \tag{25}$$

where, e.g., $\varphi^{J_z}(x,\omega)$ denotes the contribution to the eigenfunctions due to the current density within the subband. In general, this formulation permits the extraction of the equation of evolution for any mode of interest.

IV. SOLUTION OF EIGENVALUE EQUATION FOR $\mathbf{\Phi}_\alpha(\mathbf{x},\omega)$

Equation (23) is solved by obtaining expressions for $\mathbf{\Phi}_\alpha(\mathbf{x},\omega)$ in the regions $x>0$ and $x<0$, respectively, and satisfying boundary conditions at the interace obeyed by the transverse and longitudinal components of the vector potential. From the continuity of the tangential component of the vector potential at $x=0$:

$$\varphi^{A_z}{}_\alpha(0^+,\omega)=\varphi^{A_z}{}_\alpha(0^-,\omega) \tag{26}$$

and the orthonormality relation, Eq. (24), the eigenfunctions are found to have the form:

$$\Phi_\alpha(\mathbf{x},\omega)=\begin{bmatrix} -\dfrac{\omega}{\sqrt{\omega_p}}\left(1-\dfrac{\beta^2\kappa_{z\alpha}^2}{\omega^2}\right)e^{-|\kappa_{x\alpha}|x}\left(\tilde{\mathbf{z}}\pm\dfrac{\kappa_{z\alpha}^2}{|\kappa_x|}\,\tilde{\mathbf{x}}\right) \\ \dfrac{en_0}{mc\sqrt{\omega_p}}\left(e\tilde{\mathbf{z}}-i\,\dfrac{\kappa_{z\alpha}^2}{\omega}\,\tilde{\mathbf{x}}\right) \end{bmatrix}, \tag{27}$$

where

$$|\kappa_{x\alpha}|^2=\kappa_{z\alpha}^2-\omega^2\mu_0\epsilon, \tag{28}$$

$$\omega_p=\frac{n_0e^2}{2m\epsilon c}. \tag{29}$$

c is the speed of light in the medium and the top (bottom) sign is for solutions $x>0$ ($x<0$).

The boundary condition on the transverse (to z) potential is obtained by integrating the eigenvalue equation, Eq. (23), over the coordinate x, transverse to the direction of propagation. From this boundary condition, the following expression for the transverse wave number is obtained:

$$|\kappa_x|=-\frac{\omega_A}{2\beta}+\frac{\omega_A}{2\beta}\sqrt{1+4\frac{\omega^2}{\omega_A^2}\left(1-\frac{\beta^2}{c^2}\right)}, \tag{30}$$

where

$$\omega_A=\frac{n_0e^2}{2\epsilon m\beta}. \tag{31}$$

For our system, only one subband is occupied and hence there exists only one allowed plasmon–polariton mode of propagation, i.e., α takes on only one value such that $\mathbf{\Phi}_\alpha(x,\omega)\equiv\Phi(x,\omega)$.

From Eqs. (28)–(31), the well-known dispersion relation for linear 2D electron plasmon[21] is obtained, namely,

$$\omega^2=\beta^2\kappa_z^2+\omega_pc\sqrt{\kappa_z^2-\frac{\omega^2}{c^2}}. \tag{32}$$

At very low frequencies, $\omega\ll\omega_p$, the transverse wave number $|\kappa_x|\to0$ and the dispersion relation tends to $\omega\approx c\kappa_z$; hence, the mode structure contained in Eq. (27) becomes electromagneticlike in that the longitudinal component of the wave vector $A_z\to0$. At the other extreme, where $\omega\gg\omega_p/(\beta/c)$, the excitations are nondispersive and characterized by $A_x\to0$, traveling at the thermal speed, that is $\omega\approx\beta\kappa_z$. In the region between these two extremes, the excitations are mixed and dispersive, due to the non-negligible contributions of both the transverse and longitudinal field components to the mode structure.

V. EQUATION OF EVOLUTION FOR THE EXCITATION AMPLITUDE

The equation of evolution for the plasmon excitation amplitude, $a_\alpha(z,\omega)$, may be obtained from Eq. (15) using Eqs. (22), (25), and the orthonormality relation (24). Inserting Eq. (22) into (15) yields:

$$\sum_\alpha (L-L_0)a_\alpha(z,\omega)\Phi_\alpha(x,\omega)+\sum_\alpha L_0 a_\alpha(z,\omega)\mathbf{\Phi}_\alpha(\mathbf{x},\boldsymbol{\omega})+\sum_\alpha \Gamma\partial_z^2 a_\alpha(z,\omega)\mathbf{\Phi}_\alpha(\mathbf{x},\boldsymbol{\omega})=\mathrm{NL}. \tag{33}$$

Again, since only one propagating mode exists, the summation can be dropped and $a_\alpha(\mathbf{z},\boldsymbol{\omega}) \equiv a(z,\omega)$ just as $\mathbf{\Phi}_\alpha(x,\omega)\equiv\Phi(x,\omega)$. Using the eigenfunctions and their orthogonality properties, the equation for $a(z,\omega)$ is obtained as follows. Taking the inner with $\mathbf{\Phi}(x,\omega)$ and using the eigenvalue equation, the above equation becomes

$$\begin{aligned}\{\mathbf{\Phi}(x,\omega),(L-L_0)\mathbf{\Phi}(x,\omega)\}a(z,\boldsymbol{\omega})+\{\mathbf{\Phi}(x,\omega),\Gamma\mathbf{\Phi}(x,\omega)\}\kappa_z^2(\omega)a(z,\boldsymbol{\omega})\\ +\{\mathbf{\Phi}(x,\omega),\Gamma\mathbf{\Phi}(x,\omega)\}\partial_z^2 a(z,\omega)=\{\mathbf{\Phi}(x,\boldsymbol{\omega}),\mathrm{NL}\}.\end{aligned} \tag{34}$$

Applying the orthogonality condition and evaluating the inner products as follows:

$$\{\mathbf{\Phi}(x,\omega),(L-L_0)\mathbf{\Phi}(x,\omega)\}=\frac{2}{c^2\tau}\,i\boldsymbol{\omega} \tag{35}$$

and

$$\{\Phi,\mathrm{NL}\}=\int_{-\infty}^{\infty}[(\varphi^{\partial_z A_x{}^*}\tilde{\mathbf{x}}+\varphi^{A_z{}^*}\tilde{\mathbf{z}})\quad(\varphi^{J_z^*}\tilde{\mathbf{z}}-i\varphi^{\partial_z n^*}\tilde{\mathbf{x}})]\cdot\begin{bmatrix}0\\ I\delta(x)\tilde{\mathbf{z}}\end{bmatrix}dx, \tag{36}$$

where

$$I=-\frac{i\omega\mu_0}{c\sqrt{\omega_p}}\int\int d\omega_1 d\omega_2\,\delta(\omega-\omega_1-\omega_2)\left\{\frac{2}{en_0}J_z(\boldsymbol{\omega}_1)-\frac{e\omega_1}{m\omega_2}A_z(\boldsymbol{\omega}_1)\right\}\partial_z J_z(\boldsymbol{\omega}_2). \tag{37}$$

Equation (34) becomes

$$\begin{aligned}\partial_z^2 a(z,\omega)=-\kappa_z^2(\omega)a(z,\omega)-\frac{i\omega}{\tau c^2}a(z,\boldsymbol{\omega})+\frac{i\omega e^3 n_0\mu_0}{m^2 c\omega_p\sqrt{\omega_p}}\int\int d\omega_1 d\omega_2\,\delta(\omega-\omega_1-\omega_2)\\ \times\left\{2+\frac{\omega_1}{\omega_2}\left(1-\frac{\beta^2\kappa_z^2(\boldsymbol{\omega}_1)}{\boldsymbol{\omega}_1^2}\right)\right\}a(\boldsymbol{\omega}_1)\partial_z a(\boldsymbol{\omega}_2).\end{aligned} \tag{38}$$

VI. EQUATION OF EVOLUTION IN THE REGIME $\omega<\omega_p$

To obtain the nonlinear equation of evolution in space–time, Eq. (38) is inverse transformed to the time domain. In this paper, we shall focus on the nonlinear excitations with frequency bandwidth $\omega<\omega_p$, for which the plasmon–polariton has a strong photon content and the spatial dispersion parameter, β, plays a negligible role. The role of spatial dispersion in the $\omega>\omega_p$ regime will be discussed in a later publication. Thus, in carrying out the inverse transformation, an approximate expansion of $\kappa_z(\boldsymbol{\omega})$ can be used in Eq. (38), valid for $\omega<\omega_p$. In this regime, Eq. (28) becomes

$$|\kappa_x|\approx\frac{\omega^2}{\omega_A\beta}\left(1-\frac{\beta^2}{c^2}\right)=\frac{\omega^2}{\omega_p c}\left(1-\frac{\beta^2}{c^2}\right), \tag{39}$$

such that

$$\kappa_z(\omega_2)\cong\frac{\omega_2}{c}\left(1+\frac{\omega_2^2}{2\omega_p^2}\left(1-\frac{\beta^2}{c^2}\right)^2\right). \tag{40}$$

Using the above approximation, Eq. (38) yields

$$\begin{aligned}\partial_z^2 a(z,\omega)=&\frac{(-i\omega)^2}{c^2}\,a(z,\omega)-\frac{(-i\omega)^4}{\omega_p^2c^2}\,K^2a(z,\omega)+\frac{(-i\boldsymbol{\omega})}{\tau c^2}\,a(\mathbf{z},\boldsymbol{\omega})\\&-\frac{ei\omega}{2mc\sqrt{\omega_p}}\int\int d\omega_1 d\omega_2\,\delta(\omega-\omega_1-\omega_2)\Big\{-2(-i\boldsymbol{\omega}_2)-(-i\boldsymbol{\omega}_1)\mathbf{K}\\&-\frac{\beta^2K^2}{\omega_p^2c^2}(-i\omega_1)^3+\frac{K^2}{f_p^2}(-i\omega_2)^3+\frac{K^2}{2\boldsymbol{\omega}_p^2}(-i\boldsymbol{\omega}_1)(-i\boldsymbol{\omega}_2)^2-\frac{\beta^2K^2}{2\omega_p^2c^2}(-i\omega_1)(-i\omega_2)^2\\&+\frac{\beta^2K^4}{2\omega_p^4c^2}(-i\boldsymbol{\omega}_1)^3(-i\boldsymbol{\omega}_2)^2\Big\}a(\boldsymbol{\omega}_1)a(\boldsymbol{\omega}_2),\end{aligned} \tag{41}$$

where

$$K=\left(1-\frac{\beta^2}{c^2}\right).$$

As a result of the transformation, $-i\omega\rightarrow\partial_t$ (for excitations with spectral distribution $\boldsymbol{\omega}<\boldsymbol{\omega}_p$). After introducing the following change of variables:

$$t\rightarrow f_p t \quad \text{and} \quad z\rightarrow z\,\frac{f_p}{c},$$

$$a(z,t)\rightarrow\frac{1}{a'}\,a(z,t) \quad \text{where}\ \frac{1}{a'}=\frac{e}{mc\sqrt{\boldsymbol{\omega}_p}}, \tag{42}$$

we obtain the nonlinear equation of evolution in space for the amplitude of the excitations:

$$\begin{aligned}\partial_z^2 a(z,t)=&\partial_t^2 a(z,t)-K^2\partial_t^4 a(z,t)+\frac{2}{f_p\tau}\,\partial_t a(z,t)+\frac{1}{2}\,\partial_t\Big\{-(2+K)a(z,t)\partial_t a(z,t)\\&+K^3a(z,t)\partial_t^3a(z,t)+\frac{K^3}{2}\,\partial_t a(z,t)\partial_t^2a(z,t)+\frac{K^4\beta^2}{2c^2}\,\partial_t^2a(z,t)\partial_t^3a(z,t)\Big\}.\end{aligned} \tag{43}$$

Equation (43) is a nonlinear, dispersive equation which is second order in space, implying solutions that propagate in the $+z$ and $-z$ directions.

Using the reductive perturbative method,[22] the evolution equation (43) can be reduced to a more tractable nonlinear equation describing a wave traveling in one direction in the weak amplitude limit. Introducing a small parameter, $\boldsymbol{\epsilon}\ll 1$, and performing the following change of variables:

$$Z=\boldsymbol{\epsilon}^3 z, \qquad T=\boldsymbol{\epsilon}^{1/2}(z-t). \tag{44}$$

The amplitude $a(z,t)$ is expressed in a perturbation series:

$$a=\boldsymbol{\epsilon}a_1+\boldsymbol{\epsilon}^2a_2+\cdots, \tag{45}$$

where $\boldsymbol{\epsilon}^{1/2}$ and $\boldsymbol{\epsilon}^3$ are chosen in order to balance the dispersive and nonlinear terms. Substituting Eqs. (44) and (45) into (43) and integrating once with respect to time, we find to lowest order in $\boldsymbol{\epsilon}$:

$$\partial_Z \hat{a}(Z,T) + K^2 \partial_T^3 \hat{a}(Z,T) - \left(1 + \frac{K}{2}\right) \hat{a}(Z,T) \partial_T \hat{a}(Z,T) \exp\left(-\frac{z}{\omega_p \tau}\right) = 0, \tag{46}$$

where $a_1(z,t) \equiv \hat{a}(z,t)\exp(-z/\omega_p\tau)$. This reduces to the Korteweg de Vries (KdV) equation in the limit $\tau \to \infty$. The KdV equation admits solitary wave solutions, which are localized solutions belonging to the class of traveling waves.[23]

To next higher order in ϵ, Eq. (43) reduces to

$$\partial_Z \hat{z}(Z,T) = K^2 \partial_T^3 \hat{a}(Z,T) - \left\{ \tfrac{1}{2}(2+K) \hat{a}(Z,T) \partial_T \hat{a}(Z,T) + \frac{K^3}{2} \hat{a}(Z,T) \partial_T^3 \hat{a}(Z,T) \right.$$
$$\left. + \frac{K^3}{4} \partial_T \hat{a}(Z,T) \partial_T^2 \hat{a}(Z,T) + \frac{K^4 \beta^2}{4c^2} \partial_T^2 \hat{a}(Z,T) \partial_T^3 \hat{a}(Z,T) \right\} \exp\left(-\frac{z}{\omega_p \tau}\right). \tag{47}$$

We will now consider the damping factor $\exp(-z/\omega_p\tau)$ in the limit $\tau \to \infty$. The effects of damping of the nonlinear terms will be left for a numerical investigation of Eq. (47) to be done at a later date. Equation (47) can be rewritten in the general form shown below:

$$\partial_z \hat{a}(z,t) = v(\hat{a}) \partial_t \hat{a}(z,t) + \chi(\hat{a}) \partial_t^3 \hat{a}(z,t), \tag{48}$$

where

$$\begin{aligned} v(\hat{a}) &= v_0 + \alpha \hat{a}(z,t) + \eta \partial_t^2 \hat{a}(z,t), \\ \chi(\hat{a}) &= \gamma + \delta \hat{a}(z,t) + \xi \partial_t^2 \hat{a}(z,t). \end{aligned} \tag{49}$$

This is a generalized dispersive equation with amplitude dependent coefficients. It contains the Airy equation[23] in the linear regime, where $\hat{a}(z,t) \to 0$ in $v(\hat{a})$ and $\chi(\hat{a})$, and, as shown above, the KdV equation to lowest order in the nonlinear regime.

VII. STATIONARY SOLUTIONS OF THE EQUATION OF EVOLUTION IN THE PLASMON–POLARITON REGIME

The elementary excitations of the system are found by seeking the stationary solutions to Eq. (47) which possess a permanent profile of the form

$$a(Z,T) = f(\theta), \tag{50}$$

where

$$\theta = T - \frac{Z}{v}. \tag{51}$$

θ represents the normalized position in a coordinate system moving at a normalized velocity v for which the wave appears stationary. Substituting this functional relation into Eq. (47) yields the following equation in θ:

$$d_\theta a(\theta) = -u^3 K^2 d_\theta^3 a(\theta) + \tfrac{1}{4} u (2+K) d_\theta a^2(\theta) - \tfrac{1}{2} K^3 u^3 d_\theta \{a(\theta) d_\theta^2 a(\theta)\}$$
$$+ \tfrac{1}{8} K^3 u^3 d_\theta \{d_\theta a(\theta)\}^2 - \frac{\beta^2}{8c^2} u^5 K^4 d_\theta \{d_\theta^2 a(\theta)\}^2, \tag{52}$$

where $d_\theta \equiv d/d\theta$. Integrating once with respect to θ (and setting integration constants to zero since we are looking for localized solutions and derivatives that tend to zero as $\theta \to \pm\infty$):

$$a(\theta) = -u^3 K^2 d_\theta^2 a(\theta) + \tfrac{1}{4} u (2+K) a^2(\theta) - \tfrac{1}{2} u^3 K^3 a(\theta) d_\theta^2 a(\theta)$$
$$+ \tfrac{1}{8} u^3 K^3 \{d_\theta a(\theta)\}^2 - \frac{1}{8} \frac{\beta^2}{c^2} u^5 K^4 \{d_\theta^2 a(\theta)\}^2. \tag{53}$$

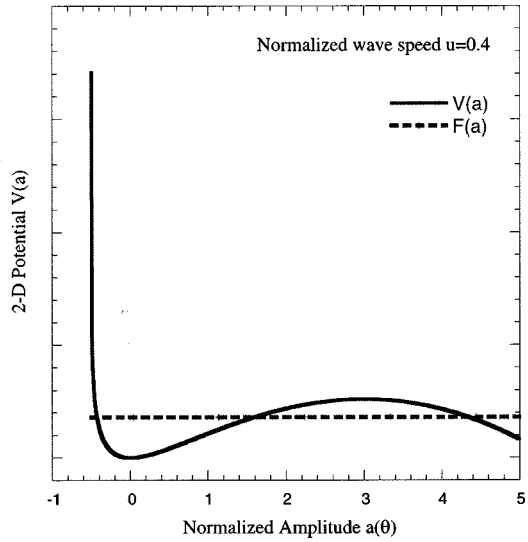

FIG. 1. Potential as a function of normalized amplitude. As the initial amplitude approaches the point of unstable equilibrium, the period of the cnoidal wave$\rightarrow\infty$, corresponding to solitary excitations. The bound, cnoidal excitations exist only for the initial excitation amplitude falling within the range $-0.883<a<3.33$.

Equation (53) can be expressed as:

$$K^2u^3d_\theta^2a(\theta)=-\frac{dV(a)}{da}+\frac{u^3K^3(d_\theta a(\theta))^2-u^5K^4\,\dfrac{\beta^2}{c^2}\,(d_\theta^2a(\theta))^2}{8(1+\frac{1}{2}Ka(\theta))}, \tag{54}$$

where

$$\frac{dV(a)}{da}=\frac{a(\theta)-\frac{1}{4}u(2+K)a^2(\theta)}{1+\frac{1}{2}Ka(\theta)}. \tag{55}$$

We present both a phase plane analysis [i.e., plots of $d_\theta a(\theta)$ vs $a(\theta)$], and a numerical analysis to understand the nature of the solutions of Eq. (54). In general, equations that take the form of a particle in a potential well

$$\frac{1}{2}\,u^3d_\theta^2a(\theta)=-\frac{dV(a)}{da} \tag{56}$$

exhibit solitary behavior near the extrema of $a(\theta)$.[23] At this limit, the second term on the right-hand side of Eq. (54) can be neglected to obtain, approximately, the regions of θ where bound solutions of $a(\theta)$ exist. Neglecting this second term, Eq. (54) takes the form of Eq. (56), with

$$V(a)=\left(\frac{2}{K}\right)^2\left\{\left(1+u\,\frac{2+K}{2K}\right)\left(\frac{K}{2}\,a-\ln\left(1+\frac{1}{2}\,Ka\right)\right)-u\,\frac{K(2+K)}{16}\,a^2\right\}. \tag{57}$$

This potential is plotted as a function of normalized amplitude for a normalized speed of $u=0.6$, in Fig. 1. Integrating Eq. (56) once yields

$$\tfrac{1}{2}u^3(d_\theta a(\theta))^2=V(a_0)-V(a)\equiv F(a), \tag{58}$$

where $V(a_0)$ is a constant of integration and a_0 is the initial excitation amplitude.

Real solutions of Eq. (58) exist for $F(a)>0$ and bound solutions exist for $F(a)$ positive between two of its simple zeros. It has been shown[23] that there exist periodic solutions $F(a)$ if

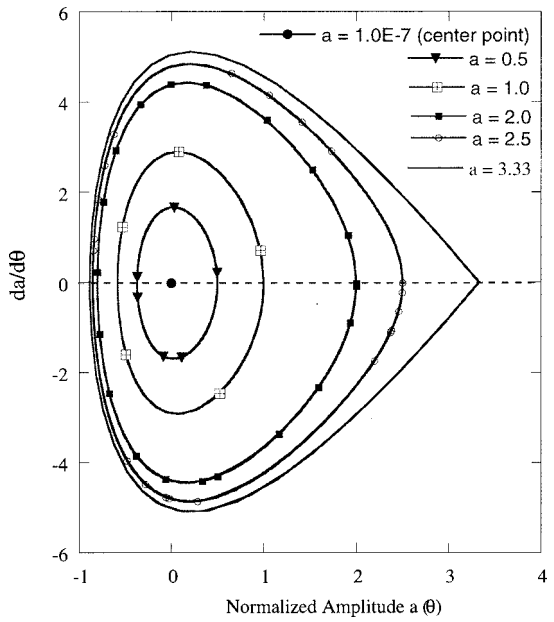

FIG. 2. Phase plane curves $da/d\theta$ vs $a(\theta)$ varying from the linear wave limit (at the center of the curves) to the solitary excitation represented by the largest phase curve, corresponding to an initial excitation amplitude $a_0=3.33$.

$(d_\theta a(\theta))^2$ is positive between two simple zeros of $(d_\theta a(\theta))^2$, and there exist solitary-wave solutions if $(d_\theta a(\theta))^2$ is positive between a simple zero and a double zero [i.e., a zero at a local maximum of minimum of $(d_\theta a(\theta))^2$].

Phase plane curves of Eq. (54) are shown in Fig. 2. From numerical solution of Eq. (54), bound solutions are found for initial amplitudes falling within the range $-0.883 \leqslant a_0 \leqslant 3.33$. The curves in Fig. 2 are plotted with values of the parameter a_0 chosen within this range, and at a fixed value of the normalized speed u. The phase curve for a linear wave reduces to a point at the center of the family of curves. The curves around this point cross normal to the a-axis and correspond to the periodic, nonlinear cnoidal waves. The outermost curve, where parameter $a_0=3.33$, does not exhibit this property and represents a solitary wave. This value of a_0 closely agrees with the initial amplitude corresponding to the maximum bounding potential determined from Fig. 1 [i.e., $F(a)$ is positive between a simple zero and a double zero, which occurs at the point of unstable equilibrium $a \approx 3.33$]. At this limit, the dispersion balances the effects of the nonlinearity, leading to a pulselike wave, a solitary wave, which propagates with constant velocity and profile.

The wave profile of the full equation (55), including all terms, is determined by numerical solution. Figure 3 is a plot of the normalized amplitude versus θ for two different initial excitation amplitudes, a_0. The linear excitations shown in Fig. 3 are obtained for $a_0=1.0E-7$ and, as discussed earlier, are represented by the point in the phase plane curves of Fig. 2. The sinusoidal behavior with θ is characteristic of linear waves. In contrast, the wave profile obtained for $a_0=3.3$ shows the characteristic steepening of nonlinear excitations. These bound, periodic excitations are the cnoidal waves, for which phase plane curves were plotted in Fig. 2. One wavelength (as shown in Fig. 3) corresponds to traveling once around the curve of Fig. 2. The solitary excitations correspond to the limit of infinite period, where $a(\theta)$ tends to a single pulse with $a_0=3.33$. This limit is represented by the outermost phase curve of Fig. 2, where the curve is traveled just once for the entire range $-\infty < \theta < \infty$.

VIII. CONCLUDING REMARKS

In this paper, we have introduced a novel matrix formulation of coupled particle-field equations leading to a nonlinear equation of evolution for the plasmon–polaritons of a 2DEG. The stationary solutions of the evolution equation were shown to yield solitary excitations. The procedure developed here can incorporate more sophisticated effects that may play an important role

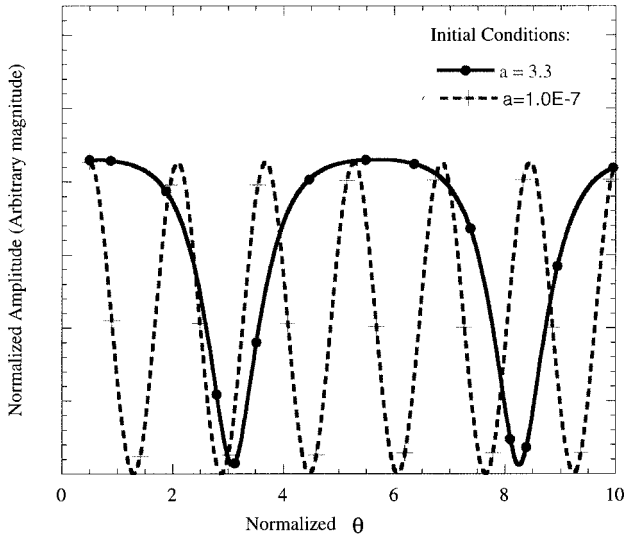

FIG. 3. Normalized amplitude $a(\theta)$ vs θ. The period of oscillation for the normalized amplitude grows as the initial amplitude approaches the point of unstable equilibrium, found from Figs. 1 and 2 to be $a \approx 3.33$.

in the evolution of these solitary excitations. Whether or not these solitary excitations are ''true'' solitons, as well as the stability and general mathematical properties of these excitations, remain to be investigated.

ACKNOWLEDGMENT

This work is supported by ARL/BMDO.

[1] F. Stern, Phys. Rev. Lett. **18**, 546 (1967).
[2] A. Eguiluz and A. A. Maradudin, Surf. Sci. **73**, 437 (1978).
[3] S. Das Sarma and J. J. Quinn, Phys. Rev. B **25**, 7603 (1982).
[4] Q. Li and S. Das Sarma, Phys. Rev. B **40**, 5860 (1983).
[5] T. Demel and D. Heitmann, Phys. Rev. Lett. **66**, 2657 (1991).
[6] S. Das Sarma and Wu-yan Lai, Phys. Rev. B **32**, 1401 (1985).
[7] A. Puri and W. L. Schaich, Phys. Rev. B **31**, 974 (1985).
[8] W. Xiaoguang, F. M. Peeters, and J. T. Devreese, Phys. Rev. B **32**, 6982 (1985).
[9] J. Wang and J. P. Leburton, Phys. Rev. B **41**, 7846 (1990).
[10] B. S. Mendoza and W. L. Schaich, Phys. Rev. B **43**, 6590 (1991).
[11] B. S. Mendoza and W. L. Schaich, Phys. Rev. B **43**, 9275 (1991).
[12] P. J. Price, Ann. Phys. (N.Y.) **133**, 217 (1981).
[13] B. K. Ridley, J. Phys. C **15**, 5899 (1982).
[14] L. Wendler and R. Pechstedt, Phys. Status Solidi B **138**, 197 (1986).
[15] W. Cai, M. C. Marchetti, and M. Lax, Phys. Rev. B **34**, 8573 (1986).
[16] R. Jalabert and S. Das Sarma, Phys. Rev. B **40**, 9723 (1989).
[17] L. Wendler, R. Haupt, and R. Pechstedt, Phys. Rev. B **52**, 9031 (1995).
[18] E. H. Hwang and S. Das Sarma, Phys. Rev. B **52**, 8668 (1995).
[19] P. A. Knipp and T. L. Reinecke, Phys. Rev. B **52**, 5923 (1995).
[20] A. D. Boardman, *Electromagnetic Surface Modes* (Academic, New York, 1982).
[21] A. L. Fetter, Ann. Phys. (N.Y.) **81**, 367 (1973).
[22] T. Taniuti and C. C. Wei, J. Phys. Soc. Jpn. **24**, 941 (1968).
[23] E. Infeld and G. Rowlands, *Nonlinear Waves, Solitons and Chaos* (Cambridge University Press, New York, 1990).

Transition probabilities between quasifree states

H. Scutaru[a)]
Department of Theoretical Physics, Institute of Atomic Physics, Bucharest-Magurele, POB MG-6, Romania

(Received 16 April 1998; accepted for publication 27 August 1998)

We obtain a general formula for the transition probabilities between any state of the C^* algebra of the canonical commutation relations (CCR-algebra) and a squeezed quasifree state (Theorem III.1). Applications of this formula are made for the case of multimode thermal squeezed states of quantum optics using a general canonical decomposition of the correlation matrix valid for any quasifree state. In the particular case of a one-mode CCR-algebra we show that the transition probability between two quasifree squeezed states is a decreasing function of the geodesic distance between the points of the upper half-plane representing these states. In the special case of the purification map it is shown that the transition probability between the state of the enlarged system and the product state of real and fictitious subsystems can be a measure for the entanglement. © *1998 American Institute of Physics.* [S0022-2488(98)02212-9]

I. INTRODUCTION

The notion of the quasifree state has appeared and was developed in the framework of the C^*-algebra approach to the canonical commutation relations (CCR).[1–11] The quasifree states are the natural ground states for the Hamiltonians which are at most quadratic in the bosonic creation and annihilation operators. The essential property of the quasifree states is the fact that all their correlations are expressible in terms of the one- and two-point functions.

The advantage of the C^*-algebra approach to the theory of coherent and squeezed states (thermal or not) follows from the fact that the results are obtained in a representation independent form. This representation independence allows us to elude the hard noncommutative calculation. However, a price must be paid for this. A simple result in usual quantum mechanics is obtained, in this case, with the help of slightly difficult results concerning the transition probabilities between the states of C^*-algebras.[12]

We shall illustrate this assertion by giving a representation-independent formula for the transition probability between a squeezed state and an arbitrary state of the C^*-algebra of CCRs (Theorem III.1).

Many interesting states of the radiation field in quantum optics are quasifree states. The factorization of the correlations in coherent states was discovered by Glauber as the essential property of these states of coherent light. The squeezed states are also quasifree states. The most general case is that of multimodal thermal displaced and squeezed states.

The main result of the paper, given in Sec. III, is the formula for the transition probabilities between any state of the C^*-algebra of the canonical commutation relations and any squeezed quasifree pure state (Theorem III.1). This result is obtained using the fact that any quasifree squeezed pure state can be obtained from a coherent state by a Bogoliubov automorphism and the fact that for any coherent state there is an element of the CCR-algebra which is a minimal orthoprojection.[4] Then a general result of P. M. Alberti and V. Heinemann[12] is applicable.

A new approach to the theory of multimode squeezed thermal states is described in Sec. IV, and is based on the combination of the Williamson theorem,[13] which gives the most general structure of a positive definite matrix with a theorem of Balian, De Dominicis, and Itzykson,[14] which gives the most general structure of a symplectic matrix. In this way a general natural parametrization of the correlation matrix of a quasifree state is obtained. This parametrization

[a)]Electronic mail: scutaru@theor1.ifa.ro

 © 1998 American Institute of Physics

points out the squeezing and the orthogonal symplectic transformations. The orthogonal transformations mix the different modes.

The geometric interpretation, given in Sec. V, is valid only for the one-mode case and shows that the transition probability between two pure squeezed states decreases with the increase of the geodesic distance between the corresponding points of the Poincaré upper half plane.

In Sec. VI the relation between the Wigner function of a quasifree state of a system and that of a subsystem is given in most general form. A first application of this relation is given in Sec. VII where the correlation matrix of a pure quasifree state of an enlarged system (which reduces to the mixed state of the orginal system) is obtained in an elementary way.[4]

In Sec. VIII it is shown that the transition probability between the above pure state obtained for the enlarged system and the mixed state obtained as the direct product of the state of the original subsystem with the state of the fictitious subsystem can be a measure of the entanglement of these two subsystems reflected in the pure state of the total system.

Finally, in Sec. IX, it is shown how the correlations between the original subsystem and the fictitious one are created by the Bogoliubov transformations which map the Fock state into the above pure state obtained by the purification procedure.

II. QUASIFREE STATES

The phase space considered in the present paper is a finite-dimensional symplectic space (E,σ). This is a real vector space endowed with a real, bilinear, antisymmetric form $\sigma(.,.)$ which gives the symplectic structure on E. Then E is of even real dimension $2n$ and there exist in E symplectic bases of vectors $\{e_j,f_j\}_{j=1,\ldots,n}$, i.e., reference systems such that $\sigma(e_j,e_k)=\sigma(f_j,f_k)=0$ and $\sigma(e_j,f_k)=-\sigma(f_k,e_j)=\delta_{jk}$, $j,k=1,\ldots,n$. The coordinates (ξ^j,η^j) of a vector $u\in E$ in a symplectic basis $[u=\Sigma_{j=1}^{n}(\xi^j e_j+\eta^j f_j)]$ are called symplectic coordinates. The measure $dm(u)=\Pi_{j=1}^{n}d\xi^j d\eta^j$ is the same for all symplectic coordinates and is called the Liouville measure on (E,σ).[1] There is a one-to-one correspondence between the symplectic bases and the linear operators J on E defined by $Je_k=-f_k$ and $Jf_k=e_k$, $k=1,\ldots,n$. The essential properties of these operators are $\sigma(Ju,u)\geqslant 0$, $\sigma(Ju,v)+\sigma(u,Jv)=0$ $(u,v\in E$ and $J^2=-I$, I denotes the identity operator on $E)$. Such operators are called complex structures.[1] In the following we shall use the matricial notations with $u\in E$ as column vectors. Then $\sigma(u,v)=u^T Jv$ and the scalar product is given by $\sigma(Ju,v)=u^T v, u,v\in E$. A linear operator S on E is called a symplectic operator if $S^T JS=J$. When S is a symplectic operator, then S^T and S^{-1} are also symplectic operators. The group of all symplectic operators $\mathrm{Sp}(E,\sigma)$ is called the symplectic group of (E,σ). The Lie algebra of $\mathrm{Sp}(E,\sigma)$ is denoted by $\mathrm{sp}(E,\sigma)$ and its elements are operators R on E with the property $(JR)^T=JR$. If J and K are two complex structures, there exists a symplectic transformation S such that $J=S^{-1}KS$.

Definition II.1:[1,3] *The C*-algebra* $\mathbf{A}$ *of the canonical commutation relations (CCR) is obtained by the completion of the *-algebra* $\mathrm{Span}\{\delta_u : u\in E\}$, *the elements of which satisfy the Weyl relations:*

$$\delta_u\delta_v=\exp(-i\sigma(u,v)/2)\,\delta_{u+v},\quad \delta_u^*=\delta_{-u},u,v\in E. \tag{2.1}$$

A state on $\mathbf{A}$ is a positive linear functional ω normalized by $\omega(\delta_0)=1$.[1–11] The complex valued function on the phase space $\omega(\delta_u)$ is called the characteristic or the generating function of the state ω. A continuous complex-valued function f on E, with $f(0)=1$, is the characteristic function of a state on $\mathbf{A}$ if and only if

$$\sum_{j,k=1}^{N}\overline{a}_j a_k\exp(i\sigma(u_j,u_k)/2)f(u_k-u_j)\geqslant 0 \tag{2.2}$$

for all $a_k\in\mathbf{C}$, $u_k\in E$, $k=1,\ldots,N$, and $N\in\mathbf{N}$.

For any $v\in E$ an automorphism τ_v of $\mathbf{A}$ can be defined by

$$\tau_v(x)=\delta_v^* x\,\delta_v. \tag{2.3}$$

A Bogoliubov transformation of $\mathbf{A}$ is a *-automorphism of $\mathbf{A}$ defined for any $S\in\mathrm{Sp}(E,\sigma)$ by

$$\alpha_S(\delta_u)=\delta_{Su},\ u\in E. \tag{2.4}$$

The quasifree automorphisms of **A** are those of the form $\tau_v\circ\alpha_S$ where $v\in E$ and $S\in \mathrm{Sp}(E,\sigma)$.

Let A be a symmetric ($A^T=A$) and positive ($u^TAu\geqslant 0$ for all $u\in E$) matrix.

Definition II.2:[1–11] *A quasifree state* $\omega_{(A,v)}$ *on* **A** *is a state whose characteristic function is given by the formula*

$$\omega_{(A,v)}(\delta_u)=\exp(-u^TAu/4+iu^Tv) \tag{2.5}$$

with the restriction

$$-JAJ\geqslant A^{-1}. \tag{2.6}$$

Evidently, we have

$$\omega_{(A,v)}\circ\alpha_S=\omega_{(S^TAS,S^Tv)} \tag{2.7}$$

and $\omega_{(A,v)}=\omega_{(A,0)}\circ\tau_v$. The state $\omega_{(A,v)}$ is pure iff $\omega_{(A,0)}$ is pure and this is the case iff $-(JA)^2=I$ [this is equivalent with $JA\in \mathrm{Sp}(E,\sigma)$] and, in particular, when $A=I$, i.e., when the state is a coherent state. The pure squeezed states are the quasifree states of the form $\omega_{(I,v)}\circ\alpha_S=\omega_{(S^TS,S^Tv)}$.

III. TRANSITION PROBABILITIES

The transition probability $P(\omega_1;\omega_2)$ between two arbitrary states ω_1 and ω_2 on **A** can be defined in many ways.[15–17,12]

Theorem III.1: *If the state* ω_1 *is a squeezed state, the transition probability to an arbitrary state* ω_2 *can be computed by the following representation-independent formula:*

$$P(\omega_1;\omega_2)=(2\pi)^{-n}\int_E \omega_1(\delta_{-u})\omega_2(\delta_u)\,dm(u), \tag{3.1}$$

where $dm(u)$ *is the Liouville measure defined above.*

A proof of this result can be obtained by combining the construction of a projection as in Lemma 3 from Ref. 4 with the assertion (vi) of Sec. II from Ref. 12.

Lemma III.1: For any coherent state $\omega_{(I,v)}$ *we can define an element* $p_v\in\mathbf{A}$ *such that it is a projector (i.e.,* $p_v^2=p_v$ *and* $p_v^*=p_v$*)* and has the property

$$p_vxp_v=\omega_{(I,v)}(x)p_v \tag{3.2}$$

for any $x\in\mathbf{A}$.

Proof: We take

$$p_v=(2\pi)^{-n}\int_E \omega_{(I,v)}(\delta_{-u})\delta_u\,dm(u). \tag{3.3}$$

The fact that p_v is a self-adjoint element of **A** is evident. The fact that it is an idempotent element of **A** results from

$$p_v^2=(2\pi)^{-2n}\int_E\int_E \omega_{(I,v)}(\delta_{-u})\omega(\delta_{-w})\delta_u\delta_w\,dm(u)\,dm(w)$$

$$=(2\pi)^{-2n}\int_E \omega_{(I,z)}(\delta_{-z})I(z)\,dm(z), \tag{3.4}$$

where $I(z)=\int_E[\omega_{(I,v)}(\delta_{-u})\omega_{(I,v)}(\delta_{-z+u})/\omega_{(I,v)}(\delta_{-z})]\exp(-(i/2)\sigma(u,z))\,dm(u)$. Thus it is sufficient to prove that $I(z)=(2\pi)^n$ and this follows by direct computations. In this way we have associated to any coherent state $\omega_{(I,v)}$ a projector from **A**. In order to prove the last assertion it is sufficient to prove that

$$\int_E \omega_{(I,v)}(\delta_{-w})\omega_{(I,v)}(\delta_{-y+w+u})\exp\left\{-\frac{1}{2}i(\sigma(w,u)+\sigma(w,y))\right\}dm(w)$$

$$=(2\pi)^n\omega_{(I,v)}(\delta_u)\omega_{(I,v)}(\delta_{-y})\exp\frac{1}{2}i\sigma(u,y). \tag{3.5}$$

As it results from Refs. 1–3, the property (3.5) is valid for any $x\in\mathbf{A}$ and (3.2) follows.

Then we can apply the following result of P. M. Alberti and V. Heinemann given in Sec. II of Ref. 12 at the point (vi):

Lemma III.2: Let A a C-algebra and $p\in A$ be a minimal orthoprojection, i.e., $pAp=\mathbf{C}p$. Let $pxp=\nu(x)p$, with $\nu(x)\in\mathbf{C}$. Then the map $\nu:x\to\nu(x)\in\mathbf{C}$ defines a state $\nu\in S(A)$ (the set of states on A) with $\nu(p)=1$ and we have*

$$P(\omega;\nu)=\omega(p) \tag{3.6}$$

for any state $\omega\in S(A)$.

Proof of the Theorem: We must verify the fact that

$$\omega_{(I,v)}(p_v)=1. \tag{3.7}$$

We have

$$\omega_{(I,v)}(p_v)=(2\pi)^{-n}\int_E\exp\left(-\frac{1}{2}w^Tw\right)dm(x)=1. \tag{3.8}$$

Thus we have proved that for any state ω and coherent state $\omega_{(I,v)}$ the transition probability is given by

$$P(\omega;\omega_{(I,v)})=\omega(p_v)=(2\pi)^{-n}\int_E\omega_{(I,v)}(\delta_{-v})\omega(\delta_v)\,dm(v). \tag{3.9}$$

Corrolary III.1: The general formula for the transition probability between any quasifree state and a pure quasifree state is given by

$$P(\omega_{(A,v)},\omega_{(B,w)})=\left(\det\left(\frac{A+B}{2}\right)\right)^{-1/2}\exp(-(w-v)^T(A+B)^{-1}(w-v)). \tag{3.10}$$

Proof: It is well known[15–17,12] that for any automorphism α of **A** we have

$$P(\omega_1\circ\alpha;\omega_2\circ\alpha)=P(\omega_1;\omega_2). \tag{3.11}$$

When the state ω is a quasifree state ω_A we have

$$P(\omega_A;\omega_{(I,v)})=\det\left(\frac{I+A}{2}\right)^{-1/2}\exp(-v^T(I+A)^{-1}v). \tag{3.12}$$

The transition probability between two squeezed states can be also given if we take into account the fact that any squeezed state is obtained by a Bogoliubov automorophism from a coherent state and the invariance of the transition probabilities with respect to automorphisms.

Remark: When both states are squeezed pure states we have

$$P(\omega_{(A,v)};\omega_{(B,w)})=\det\left(\frac{I+S^TS}{2}\right)^{-1/2}\exp(-(w-v)^T(A+B)^{-1}(w-v)), \tag{3.13}$$

where S is the symplectic matrix such that $A=S^TBS$.

IV. APPLICATIONS IN QUANTUM OPTICS

Theorem IV.1: *The most general form of a correlation matrix A is given by*

$$A=O'^T\begin{pmatrix}M & 0\\ 0 & M^{-1}\end{pmatrix}O^T\begin{pmatrix}D & 0\\ 0 & D\end{pmatrix}O\begin{pmatrix}M & 0\\ 0 & M^{-1}\end{pmatrix}O', \tag{4.1}$$

where O and O' are symplectic and orthogonal ($O^TO=I$) operators and M is a diagonal $n\times n$ matrix.

Proof: The proof results by combining two well-known results. Because the correlation matrix A is positive definite it follows[13] that there exists $S\in \mathrm{Sp}(E,\sigma)$ such that

$$A=S^T\begin{pmatrix}D & 0\\ 0 & D\end{pmatrix}S, \tag{4.2}$$

where D is a diagonal $n\times n$ matrix.

The most general real symplectic transformation $S\in \mathrm{Sp}(E,\sigma)$ has[14] the following structure:

$$S=O\begin{pmatrix}M & 0\\ 0 & M^{-1}\end{pmatrix}O', \tag{4.3}$$

where O and O' are symplectic and orthogonal ($O^TO=I$) operators and M is a diagonal $n\times n$ matrix.

Examples: Various particular kinds of such matrices are obtained taking O, O', D, or M to be equal or proportional to the corresponding identity operator. A pure squeezed state is obtained when $D=I$. If this condition is not satisfied, the state is a mixed state called thermal squeezed state.[18] When $M=I$ there is no squeezing and the correspondig states are pure coherent states or thermal coherent states. All these states have correlations between the different modes produced by the orthogonal symplectic operators O and O'. Combining Eqs. (3.2) and (3.12) we obtain that

$$P(\omega_A\circ\alpha_S;\omega_{(I,0)})=P(\omega_A;\omega_{(I,0)}) \tag{4.4}$$

iff $S^TAS=A$ or iff $S^TS=I$, i.e., iff S is orthogonal.

The first situation appears, for example, in the case of a parametric amplifier[19] and this property explains the fact that the nonentanglement of states is preserved by such a transformation. Hence the parametric amplifier is a transition probability preserving device. The second situation appears if there is no squeezing, i.e., $M=I$. As it was shown in Ref. 7 for any quasifree state ω_A we have

$$\omega_A=(\omega_{d_1}\otimes\omega_{d_2}\otimes\ldots\otimes\omega_{d_n})\circ\alpha_S, \tag{4.5}$$

where $d_1,\ldots,d_n$ are the diagonal elements of the matrix D, and where ω_{d_i}, $i=1,\ldots,n$, are the one-mode quasifree states defined on the phase spaces E_i generated by the vectors $\{e_i,f_i\}$ (on which the operator A acts by $Ae_i=d_ie_i$ and $Af_i=d_if_i$) by the formula

$$\omega_{d_i}(\delta_{\xi^ie_i+\eta^if_i})=\exp(-d_i((\xi^i)^2+(\eta^i)^2)/4) \tag{4.6}$$

with $d_i\geqslant 1$ as it follows from Eq. (2.6). It is clear that

$$P(\omega_{d_1}\otimes\omega_{d_2}\otimes\cdots\otimes\omega_{d_n};\omega_{(I,0)})=\prod_{i=1}^{n}P(\omega_{d_i};\omega_{(I,0)}). \tag{4.7}$$

Combining Eqs. (4.5) and (4.7) we obtain that the equality

$$P(\omega_A;\omega_I)=\prod_i^n P(\omega_{d_i};\omega_I) \tag{4.8}$$

is valid only for nonsqueezed states (i.e., when $M=I$).

For pure squeezed states ($D=I$) we still have such a decomposition of the transition probability as a product of transition probabilities for the individual modes. Moreover, there is another class of quasifree states defined by the condition $O=I$ in Eq. (4.3), and for which such a decomposition is valid. The formula which describes these two situations is

$$P(\omega_A;\omega_I)=\prod_i^n \frac{2}{\sqrt{(m_i^2+d_i)(m_i^{-2}+d_i)}}. \tag{4.9}$$

The thermal squeezed states considered in Ref. 18 are obtained when $M=mI$ and $D=dI$. In this case the correlation matrix A takes the following form:

$$A=O'^T\begin{pmatrix} m^2dI & 0 \\ 0 & m^{-2}dI \end{pmatrix}O' \tag{4.10}$$

and

$$P(\omega_A;\omega_I)=\left(\frac{2}{\sqrt{(m^2+d)(m^{-2}+d)}}\right)^n. \tag{4.11}$$

We shall take into account the fact that the most general form of an orthogonal symplectic matrix is[14,13]

$$O'=\begin{pmatrix} X & Y \\ -Y & X \end{pmatrix}, \tag{4.12}$$

where X and Y are $n\times n$ matrices which satisfy the conditions $X^TX+Y^TY=I$ and $X^TY=Y^TX$.

When $Y=0$ the corresponding quasifree state is nonentangled and remains nonentangled in all symplectic frames obtained by such Bogoliubov transformations (i.e., orthogonal with $Y=0$).[18] The ideal (lossless) beam splitter affects a transformation of such kind and the nonentanglement is preserved only when all modes are with the same temperature and equally squeezed.[19] It follows from the above discussion that the ideal beam splitter is a transition probability preserving device.

V. A GEOMETRIC INTERPRETATION

The squeezed states are in a one-to-one correspondence with the elements of $\mathbf{R}=\mathrm{sp}(E,\sigma)\cap\mathrm{Sp}(E,\sigma)$. This correspondence is covariant with respect to the adjoint action of the symplectic group $\mathrm{Sp}(E,\sigma)$ on $\mathbf{R}$ and the action of this group on the squeezed states given by the corresponding Bogoliubov authomorphisms. Because the symplectic group acts transitively on $\mathbf{R}$ it follows that this is a coadjoint orbit of $\mathrm{Sp}(E,\sigma)$. In fact, this is the Hermitian symmetric space $\mathrm{Sp}(E,\sigma)/U(n)$, where $U(n)$ is the subgroup of $\mathrm{Sp}(E,\sigma)$ the elements of which are also orthogonal. Hence the squeezed states are labeled by the elements of this Hermitian symmetric space. In the one-mode case ($n=1$) the Hermitian symmetric space $\mathrm{Sp}(2,R)/U(1)$ is the Poincaré upper half-plane $\mathbf{H}$.[20] The group $\mathrm{Sp}(2,R)$ acts on $\mathbf{H}$ in the usual way:

$$\gamma: z\rightarrow \frac{az+b}{cz+d}, \tag{5.1}$$

where $z=x+\sqrt{-1}y$ and where

$$\gamma=\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{5.2}$$

with $\det(\gamma)=ad-bc=1$. The function

$$u(z,z')=\frac{|z-z'|^2}{4yy'} \tag{5.3}$$

is invariant to the action of the group $\mathrm{Sp}(2,R)$:

$$u(\gamma z,\gamma z')=u(z,z'). \tag{5.4}$$

The Poincaré metrics on **H** is also invariant and is given by

$$ds^2=\frac{dx^2+dy^2}{y^2}=\frac{dz\overline{dz}}{y^2}. \tag{5.5}$$

The corresponding distance function $\rho(z,z')$ is equal with the length of the geodesic between z and z'.

Proposition V.1: For any pair of squeezed states ω_z and $\omega_{z'}$,

$$P(\omega_z;\omega_{z'})=\frac{1}{ch(s/2)}, \tag{5.6}$$

where s is the geodesic distance between z and z'.

Proof: For any two points $z,z'\in\mathbf{H}$ there exists $\gamma\in\mathrm{Sp}(2,R)$ such that $\gamma z=\sqrt{-1}$ and $\gamma z'=\sqrt{-1}y_0$ for a real $y_0\geqslant 1$. Evidently $\rho(z,z')=\rho(\sqrt{-1},\sqrt{-1}y_0)$. The geodesic between $\sqrt{-1}$ and $\sqrt{-1}y_0$ is the vertical straight line connecting these two points.[20] Then

$$\rho(\sqrt{-1},\sqrt{-1}y_0)=\int_1^t\frac{dy}{y}=\ln(t)=s \tag{5.7}$$

and $1+u(\sqrt{-1},\sqrt{-1}t)=ch^2(s/2)$. From $\gamma\sqrt{-1}=(a\sqrt{-1}+b)/(c\sqrt{-1}+d)=x+\sqrt{-1}y$ we obtain $x=(bd+ac)/(c^2+d^2)$ and $y=(ad-bc)/(c^2+d^2)$ and

$$1+u(z,\sqrt{-1})=\frac{a^2+b^2+c^2+d^2+2}{4}. \tag{5.8}$$

The correspondence between the points of the Poincaré half-plane **H** and the squeezed states is given by

$$\sqrt{-1}\leftrightarrow J=\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{5.9}$$

and $z\leftrightarrow J\gamma$. Then

$$P(\omega_z,\omega_{\sqrt{-1}})=\det\left(\frac{1+\gamma^T\gamma}{2}\right)^{-1/2}=\frac{1}{ch(s/2)}. \tag{5.10}$$

Evidently, this result is valid for any pair of squeezed states.

VI. THE REDUCTION OF THE WIGNER FUNCTIONS TO THE SUBSYSTEMS

In the following we shall use the fact that the Wigner functions of the states on the algebra of the bosonic commutation relations give a complete description of these states.

Definition IV.1: The Wigner function of the state ω is defined as the Fourier symplectic transform of the corresponding characteristic function $\omega(\delta_u)$:

$$W_\omega(u)=(2\pi)^{-2n}\int_E \exp[i\sigma(u,v)]\omega(\delta_v)\,dm(v). \tag{6.1}$$

Proposition VI.1: The Wigner function of a quasifree state is given by

$$W_\omega(u)=\pi^{-n}(\det G)^{1/2}\exp(-u^TGu), \tag{6.2}$$

where $G=-JA^{-1}J$ *is a* $2n\times 2n$ *real, symmetric and positive definite matrix which satisfies a restriction which is equivalent with the restriction (2.6):*

$$-(JG)^2\leqslant I. \tag{6.3}$$

Remarks: We have also $JG\in \mathrm{sp}(E,\sigma)$. If ω is a pure quasifree state, then $G=A$ and $\det G=\det A=1$ [because G and A belong to $\mathrm{Sp}(E,\sigma)$].

Corrolary VI.1: The most general form of the matrix G which defines the Wigner function of a quasifree state is the following:

$$G=O'^T\begin{pmatrix} M & 0\\ 0 & M^{-1}\end{pmatrix}O^T\begin{pmatrix} D^{-1} & 0\\ 0 & D^{-1}\end{pmatrix}O\begin{pmatrix} M & 0\\ 0 & M^{-1}\end{pmatrix}O', \tag{6.4}$$

where O *and* O' *are symplectic and orthogonal* $(O^TO=I)$ *matrices and* M *is a diagonal* $n\times n$ *matrix.*

Proof: Since $S^TJS=(S^{-1})^TJS^{-1}=SJS^T=J$ for any symplectic matrix S it follows that

$$G=-JA^{-1}J=-JS^{-1}\begin{pmatrix} D & 0\\ 0 & D\end{pmatrix}(S^T)^{-1}J=S^T\begin{pmatrix} D^{-1} & 0\\ 0 & D^{-1}\end{pmatrix}S. \tag{6.5}$$

The most general real symplectic transformation $S\in \mathrm{Sp}(E,\sigma)$ has[17] the following structure

$$S=O\begin{pmatrix} M & 0\\ 0 & M^{-1}\end{pmatrix}O', \tag{6.6}$$

where O and O' are symplectic and orthogonal $(O^TO=I)$ operators and M is a diagonal $n\times n$ matrix.

As a consequence we have that $\det G=(\det D)^{-2}$.

Let us suppose that the state of the quantum system $\mathcal{S}$ is described by the Wigner function $W(u)$ defined on the phase space (E,ω) which can be considered as being composed from two subsystems $\mathcal{S}_1$ and $\mathcal{S}_2$ with the phase spaces (E_1,ω_1) and (E_2,ω_2), respectively, where $E=E_1\oplus E_2$ and ω_1 and ω_2 are the restrictions of the state ω to E_1 and to E_2, respectively. In this case the Wigner function $W_\omega(u)=\pi^{-n}(\det G)^{1/2}\exp(-u^TGu)$ can be written as

$$W(x,y)=\pi^{-n}\left(\det\begin{pmatrix} A & B\\ B^T & C\end{pmatrix}\right)^{1/2}\exp-(x^Ty^T)\begin{pmatrix} A & B\\ B^T & C\end{pmatrix}\begin{pmatrix} x\\ y\end{pmatrix}. \tag{6.7}$$

Theorem VI.1: *If* $\det C\neq 0$, *then the Wigner function of the subsystem* $\mathcal{S}_1$ *is given by*

$$W_1(x)=\pi^{-n/2}(\det(A-BC^{-1}B^T))^{1/2}\exp[-x^T(A-BC^{-1}B^T)x], \tag{6.8}$$

and if $\det A\neq 0$, *the Wigner function of the subsystem* $\mathcal{S}_2$ *is given by*

$$W_2(y)=\pi^{-n/2}(\det(C-B^TA^{-1}B))^{1/2}\exp[-y^T(C-B^TA^{-1}B)y]. \tag{6.9}$$

Proof: The Wigner function $W_1(x)$ of the subsystem $\mathcal{S}_1$ is defined by the integration on the coordinates of the subsystem $\mathcal{S}_2$:

$$W_1(x)=\int_{E_2} W(x,y)\,dm(y). \tag{6.10}$$

If $\det C \neq 0$, then

$$W_1(x) = \pi^{-n/2}(\det(A - BC^{-1}B^T))^{1/2}\exp[-x^T(A - BC^{-1}B^T)x]. \tag{6.11}$$

We have used Schur's formula

$$\det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \det C \det(A - BC^{-1}B^T). \tag{6.12}$$

Analogously, if $\det A \neq 0$, then

$$W_2(y) = \pi^{-n/2}(\det(C - B^TA^{-1}B))^{1/2}\exp[-y^T(C - B^TA^{-1}B)y]. \tag{6.13}$$

Also we have used the fact that

$$\det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \det A \det(C - B^TA^{-1}B). \tag{6.14}$$

VII. THE PURIFICATION OF A MIXED STATE

Theorem VII.1: *For any quasifree mixed state ω_A on the C*-algebra of commutation relations defined on the phase space E there exists a pure quasifree state $\tilde{\omega}_A$ on the C*-algebra of commutation relations defined on the phase space $E \oplus E$ endowed with the symplectic structure $\begin{pmatrix} J & 0 \\ 0 & -J \end{pmatrix}$. This pure quasifree state $\tilde{\omega}_A$ is defined by the correlation matrix $\tilde{A}$ (which is a $4n \times 4n$ real, symmetric and positive definite matrix)*

$$\tilde{A} = \begin{pmatrix} A & A\sqrt{I + (JA)^{-2}} \\ A\sqrt{I + (JA)^{-2}} & A \end{pmatrix}. \tag{7.1}$$

Proof: This pure quasifree state $\tilde{\omega}_A$ is defined by the correlation matrix $\tilde{A}$ (which is a $4n \times 4n$ real, symmetric and positive definite matrix)

$$\tilde{A} = \begin{pmatrix} U & V \\ V & U \end{pmatrix}, \tag{7.2}$$

which satisfies the restriction

$$-\left(\begin{pmatrix} J & 0 \\ 0 & -J \end{pmatrix}\begin{pmatrix} U & V \\ V & U \end{pmatrix}\right)^2 = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \tag{7.3}$$

i.e.,

$$(JU)^2 - (JV)^2 = -I, JUJV = JVJU. \tag{7.4}$$

The map which associates a pure quasifree state $\tilde{\omega}_A$ to any mixed quasifree state, ω_A is called the purification map. Because $\tilde{\omega}_A$ is a pure quasifree state the corresponding Wigner function is defined by the same matrix:

$$\tilde{G} = \begin{pmatrix} U & V \\ V & U \end{pmatrix}. \tag{7.5}$$

We remark that the original system can be considered as the reduction of the extended one. Then one must have the relation

$$\int_E W_{\tilde{\omega}_A}((u, \tilde{u}))\, dm(\tilde{u}) = W_{\omega_A}(u). \tag{7.6}$$

Hence we can use the result of the preceeding section, i.e., we have

$$G_1 = U - VU^{-1}V = G = -JA^{-1}J. \tag{7.7}$$

From the Schur formula it follows that

$$\det \tilde{G} = \det U \det(U - VU^{-1}V) = \det A \det A^{-1} = 1. \tag{7.8}$$

From the equations (7.3) and (7.6) we have $JGJU = JU - (JV)^2(JU)^{-1}$. Hence $JGJU = (JU)^2 - (JV)^2 = -I$, from which it follows that $U = A$ and $(JV)^2 = I + (JA)^2$. It is easy to prove that $V = A\sqrt{I + (JA)^{-2}}$ is a solution of the last equation.

Remarks: The map defined by (7.6) which acts on the Wigner functions is the inverse map of the purification map. From the formula

$$W_{\tilde{\omega}_A}((u,\tilde{u})) = \pi^{-2n} \exp\{-u^T A u - \tilde{u}^T A \tilde{u} - V^T \tilde{u} - \tilde{u}^T V u)\} \tag{7.9}$$

it follows that the correlations between the variables u of the real system and the variables $\tilde{u}$ of the fictitious one are given by the terms $(A\sqrt{I + (JA)^{-2}}u)^T\tilde{u} + \tilde{u}^T(A\sqrt{I + JA})^{-2}u)$ which appear at the exponent.

In the general case

$$A = S^T \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} S \tag{7.10}$$

and $JS^T = S^{-1}J$ and $J\left(\begin{smallmatrix} D & 0 \\ 0 & D \end{smallmatrix}\right) = \left(\begin{smallmatrix} D & 0 \\ 0 & D \end{smallmatrix}\right)J$. From these relations it follows that

$$\begin{aligned} (JA)^{-2} &= A^{-1}JA^{-1}J \\ &= S^{-1}\begin{pmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix}(S^T)^{-1}JS^{-1}\begin{pmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix}(S^T)^{-1}J \\ &= -S^{-1}\begin{pmatrix} D^{-2} & 0 \\ 0 & D^{-2} \end{pmatrix}S. \end{aligned}$$

Hence

$$\begin{aligned} I + (JA)^{-2} &= I - S^{-1}\begin{pmatrix} D^{-2} & 0 \\ 0 & D^{-2} \end{pmatrix}S \\ &= S^{-1}\begin{pmatrix} I - D^{-2} & 0 \\ 0 & I - D^{-2} \end{pmatrix}S \end{aligned}$$

and

$$A\sqrt{I + (JA)^{-2}} = S^T \begin{pmatrix} \sqrt{D^2 - I} & 0 \\ 0 & \sqrt{D^2 - I} \end{pmatrix} S. \tag{7.11}$$

VIII. A MEASURE OF THE ENTANGLEMENT

A characteristic feature of quantum mechanics is the presence of the correlations between the subsystem of a quantum systems described by a pure state. In this situation the states of the subsystems are mixed. Any quantum system can be considered to be a subsystem of a larger quantum system which is in a pure state. The enlarged system can contain real or fictitious additional subsystems. The appearance of the quantum correlations between these auxiliary subsystems and the initial one is the price which must be paid for the purity of the state of the whole system.

For any mixed quasifree state on an *n*-mode bosonic system there exists a standard procedure which defines a compound *2n*-mode bosonic system in a pure quasifree state such that the restriction of this state to the initial subsystem is the initial mixed quasifree state. This procedure is called the purification operation.

The corresponding compound system has two subsystems which are identical with the initial one and which are strongly correlated. These correlations are explicitly defined by the purification operation.

In Sec. III we have obtained the following general formula for the transition probability between any quasifree state and a pure quasifree state:

$$P(\omega_A,\omega_B)=\left(\det\left(\frac{A+B}{2}\right)\right)^{-1/2}. \tag{8.1}$$

Because the state obtained by purification is a pure state we can apply this formula in the case in which the mixed quasifree state is that given by the direct product of the mixed states of the real and fictitious systems. The last one is a mixed quasifree state with the correlation matrix given by

$$\begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}. \tag{8.2}$$

Hence we must find

$$\det\begin{pmatrix} A & \frac{A}{2}\sqrt{I+(JA)^{-2}} \\ \frac{A}{2}\sqrt{I+(JA)^{-2}} & A \end{pmatrix}. \tag{8.3}$$

One obtains

$$P(\omega_{A\oplus A},\tilde{\omega}_A)=(\det(\tfrac{3}{4}D^2+\tfrac{1}{4}I))^{-1}. \tag{8.4}$$

This transition probability can be considered as a measure of the correlation between the real and fictitious subsystems. One can remark that this transition probability is a decreasing function of the temperature of each mode, i.e., the correlation is increasing with the temperature.

IX. CORRELATIONS THAT ARE GIVEN BY THE BOGOLIUBOV AUTOMORPHISMS

Another useful result of the theory of bosonic commutation relations that can be applied to the pure state which appears from the purification operation is the fact that all quasifree states can be obtained from the Fock state by a Bogoliubov transformation. In other words, the pure quasifree states are squeezed states in a generalized meaning. Hence the strong correlations between the real modes and the fictitious ones are produced by the action of the Bogoliubov automorphisms which are of two kinds: squeezing transformations of each mode and transformations which mix different modes.

In the following we shall disscuss the case of pure quasifree states in detail. If ω_A is a pure quasifree state, then $-(JA)^2=I$. From the Williamson theorem[13] for any positive definite symmetric matrix A there exists an element $S\in \mathrm{Sp}(E,\sigma)$ and a diagonal matrix with positive entries D and with the property $DJ=JD$ and such that $A=S^TDS$. Then from the restriction $-(JA)^2=I$ one obtains $D^2=I$, i.e., $D=I$. Hence for any pure quasifree state ω_A the correlation matrix A is of the form S^TS with $S\in \mathrm{Sp}(E,\sigma)$, i.e., $\omega_A=\omega_I\circ\alpha_S$, where ω_I is called the Fock state. In other words any pure quasifree state can be obtained from the Fock state by a Bogoliubov transformation. Because the Fock state is a state without correlations between different modes it follows that in the state with the correlation matrix ω_A the correlations are produced under the action of the Bogoliubov automorphism α_S, $S\in \mathrm{Sp}(E,\sigma)$. We can apply these considerations to the quasifree pure state $\tilde{\omega}_A$ which arises by purification of the quasifree state ω_A. It follows that $\tilde{\omega}_A$ can be obtained

from the Fock state by the Bogoliubov automorphism α_S. As we have seen in the preceding section, the correlation matrix of the state obtained by purification is given by

$$\begin{pmatrix} S^T & 0 \\ 0 & S^T \end{pmatrix} \begin{pmatrix} \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} & \begin{pmatrix} \sqrt{D^2-I} & 0 \\ 0 & \sqrt{D^2-I} \end{pmatrix} \\ \begin{pmatrix} \sqrt{D^2-I} & 0 \\ 0 & \sqrt{D^2-I} \end{pmatrix} & \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} \end{pmatrix} \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}. \tag{9.1}$$

Hence the most general pure quasifree state is obtained from the mixed quasifree state described by the following correlation matrix:

$$\mathbf{D} = \begin{pmatrix} \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} & \begin{pmatrix} \sqrt{D^2-I} & 0 \\ 0 & \sqrt{D^2-I} \end{pmatrix} \\ \begin{pmatrix} \sqrt{D^2-I} & 0 \\ 0 & \sqrt{D^2-I} \end{pmatrix} & \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} \end{pmatrix} \tag{9.2}$$

using a specific Bogoliubov transform which does not couple the real and the fictious systems:

$$\begin{pmatrix} S^T & 0 \\ 0 & S^T \end{pmatrix} \begin{pmatrix} J & 0 \\ 0 & -J \end{pmatrix} \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix} = \begin{pmatrix} J & 0 \\ 0 & -J \end{pmatrix}. \tag{9.3}$$

In the enlarged phase space there exists a Bogoliubov transform $\mathbf{S}$ such that $\mathbf{S}^T\mathbf{S}=\mathbf{D}$. If we denote by $\mathbf{J}=\left(\begin{smallmatrix} J & 0 \\ 0 & J \end{smallmatrix}\right)$, then $\mathbf{S}^T\mathbf{J}\mathbf{S}=\mathbf{J}$. The solution is

$$\mathbf{S} = \begin{pmatrix} \begin{pmatrix} \sqrt{\frac{D+I}{2}} & 0 \\ 0 & \sqrt{\frac{D+I}{2}} \end{pmatrix} & \begin{pmatrix} \sqrt{\frac{D-I}{2}} & 0 \\ 0 & \sqrt{\frac{D-I}{2}} \end{pmatrix} \\ \begin{pmatrix} \sqrt{\frac{D-I}{2}} & 0 \\ 0 & \sqrt{\frac{D-I}{2}} \end{pmatrix} & \begin{pmatrix} \sqrt{\frac{D+I}{2}} & 0 \\ 0 & \sqrt{\frac{D+I}{2}} \end{pmatrix} \end{pmatrix}. \tag{9.4}$$

If we denote by $\mathcal{D}$ the matrix

$$\begin{pmatrix} \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} \end{pmatrix}, \tag{9.5}$$

then the transition probability is given by

$$P(\omega_A, \omega_B) = \left(\det\left(\frac{\mathbf{I} + \mathbf{S}\mathcal{D}\mathbf{S}^T}{2} \right) \right)^{-1/2}, \tag{9.6}$$

and we reobtain the result (8.4).

X. CONCLUSIONS

The quasifree states are among the most simple and interesting, from the physical point of view, because they are states on which the quantum correlations (the entanglement) between systems can be relatively easily described. We have applied to the finite-dimensional systems the

ideas and the methods developed for the infinite-dimensional systems by D. Kastler, J. Manuceau, A. Van Daele, A. Verbeure, M. Fannes, and A. S. Holevo. For such systems, which are interesting from the physical point of view, the theory was developed (with rare exceptions) in an (apparently) independent way which is based on noncommutative calculations. In comparison, the methods developed for infinite-dimensional systems are simpler and more adequate in the finite-dimensional case.

The fundamental object which contains all the information about a quasifree state is its correlation matrix. We have shown in the present paper how we can extract this information by elementary means. For this purpose we have used results obtained many years ago. Of great importance is a result of R. Balian, C. De Dominicis, and C. Itzykson concerning the structure of covariance matrices. According to this result all covariance matrices can be obtained from the covariance matrix of a thermal state by orthogonal symplectic transformations and squeezing symplectic transformations. This structure theorem is of fundamental importance and explains all results on the structure of correlations matrices obtained until now. All operations of restriction to subsystems and of extension from a subsystem to the whole system can also be described in terms of correlation matrices. The transition probability from an arbitrary state to a squeezed coherent state can be obtained as the value of the arbitrary state on an element (a projector) of the CCR-algebra which is associated in a natural way to any squeezed coherent state. In the present paper we have shown only that the methods developed for squeezed states can be applied also to the squeezed coherent states. We have constructed explicitly the projector associated to any squeezed coherent state. The transition probability is an immediate result.

The ultimate goal of this research is the desciption of the entanglement in the case of quasifree states. We have shown that the most entangled quasifree states are those obtained by the purification of quasifree states. In the case of pure states all corelations are obtained by Bogoliubov transformations. For the states obtained by purifications we have determined explicitly the Bogoliubov transformation by which these states can be obtained from the Fock state. We have also determined explicitly the transition probability from the state obtained by purification to the state of the system obtained by direct product of the original system to the twin fictitious one. This transition probability is a decreasing function of the temperatures of the original system. This means that a state obtained by the purification is most correlated when the original state is most mixed. Hence we have obtained that the probability transition can be considered as a measure of the entanglement.

[1] D. Kastler, Commun. Math. Phys. **1**, 14 (1965).
[2] J. Manuceau, Ann. Inst. Henri Poincaré, Sect. A **8**, 139 (1968).
[3] J. Manuceau and A. Verbeure, Commun. Math. Phys. **9**, 293 (1968).
[4] A. Van Daele and A. Verbeure, Commun. Math. Phys. **20**, 268 (1971).
[5] M. Fannes, Commun. Math. Phys. **51**, 55 (1976).
[6] A. S. Holevo, Probl. Peredachi Inf. **6**, 44 (1970).
[7] A. S. Holevo, Teor. Mat. Fiz. **6**, 3 (1971).
[8] A. S. Holevo, Teor. Mat. Fiz. **13**, 184 (1979).
[9] A. S. Holevo, IEEE Trans. Inf. Theory **IT-21**, 533 (1972).
[10] H. Scutaru, Phys. Lett. A **141**, 223 (1989).
[11] H. Scutaru, Phys. Lett. A **167**, 326 (1992).
[12] P. M. Alberti and V. Heinemann, J. Math. Phys. **30**, 2083 (1989).
[13] G. B. Folland, *Harmonic analyses in phase space* (Princeton U.P., Princeton, NJ, 1989), p. 177, Proposition 4.22.
[14] R. Balian, C. De Dominicis, and C. Itzykson, Nucl. Phys. **67**, 609 (1965).
[15] D. J. C. Bures, Trans. Am. Math. Soc. **135**, 199 (1969).
[16] A. Uhlmann, Rep. Math. Phys. **9**, 273 (1976).
[17] H. Araki and G. A. Raggio, Lett. Math. Phys. **6**, 237 (1982).
[18] H. Ezawa, A. Mann, K. Nakamura, and M. Revzen, Ann. Phys. (N.Y.) **209**, 216 (1991).
[19] H. Huang and G. S. Agarwal, Phys. Rev. A **49**, 52 (1994).
[20] S. Lang, $SL_2(\mathbf{R})$ (Addison–Wesley, Reading, MA, 1975), Chap. XIV, Sec. 1.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

Symmetry adapted states for Hubbard clusters

O. Tjernberg
European Synchrotron Radiation Facility, BP 220, 38 043 Grenoble Cedex, France

(Received 27 May 1998; accepted for publication 11 August 1998)

The application of spin, pseudo spin and space group symmetry to find symmetry adapted states for the square planar Hubbard model is discussed. An approach based on pseudo spin configurations and the application of Young tableaux to the permutation group is presented. The method is illustrated for the case of a 4*4 cluster and the complete classification of the states for this cluster is given. It is shown that the linear dimension of the largest matrix block is reduced by three orders of magnitude by application of the above symmetries. © *1998 American Institute of Physics.* [S0022-2488(98)00612-4]

I. INTRODUCTION

In the last decade there has been a strong interest in highly correlated electron systems. This interest was at least partly generated by the discovery of high temperature superconductivity (HTC) in cuprates.[1] In the attempts to unravel the mechanism behind HTC attention has often been focused on the Hubbard model[2] or related models such as the t-J model.[3] These models are considered to be the simplest possible models that might describe physical systems such as the cuprates or other transition metal oxides. Even though the Hubbard model is a very simplified model which only takes intra atomic repulsion into account, it has turned out to be very difficult to treat analytically and an analytical solution has only been obtained in the case of one dimension.[4] In view of these difficulties, numerical calculations on clusters have become an important tool in the study of models for strongly correlated systems.[5] Two approaches to numerical calculations on clusters are in wide use. One is the Quantum Monte Carlo (QMC)[6] technique and the other is the method of exact diagonalization, often implemented by use of a Lanczos algorithm.[7] The (QMC) technique is suitable for larger clusters but the so called sign problem induces large errors at a low temperature. As a result it is difficult to obtain information about ground state properties or low lying excitations from QMC simulations. The exact diagonalization technique is well suited to provide this kind of information and it is also capable of providing dynamical properties. The problem is, of course, that the dimensionality of the Hilbert space involved in the calculations become very large for larger clusters. In the case of a 4*4 cluster at half filling the number of bases states exceeds $6*10^8$ and it becomes necessary to reduce the Hamiltonian matrix to block diagonal form. For larger clusters it is therefore crucial to exploit the available symmetry as much as possible. In previous reports only certain of the available symmetries have been exploited such as the space group symmetry[8] or the spin and translational symmetries.[9] In this paper the application of the spin, pseudo spin and space group symmetries to exact diagonalization studies of clusters is discussed and as an example the classification of all the states of the half filled 4*4 cluster is given.

II. SYMMETRIES

If the Hamiltonian is invariant under spin rotation, i.e., it has a SU(2) symmetry characterized by the usual spin operators,

$$S_z=\frac{1}{2}\sum_i (c^\dagger_{i\uparrow}c_{i\uparrow}-c^\dagger_{i\downarrow}c_{i\downarrow}), \tag{2.1}$$

$$S_+=\sum_i c^\dagger_{i\uparrow}c_{i\downarrow}, \tag{2.2}$$

 © 1998 American Institute of Physics

$$S_- = \sum_i c_{i\downarrow}^\dagger c_{i\uparrow}, \tag{2.3}$$

$$S^2 = \frac{1}{2}(S_+ S_- + S_- S_+) + S_z^2, \tag{2.4}$$

then S^2 and S_z are good quantum numbers which can be used to classify the states of the Hamiltonian. In the standard notation used here, $c_{i\uparrow}^\dagger$ and $c_{i\downarrow}$ represent creation and destruction operators, respectively. The index i represents the lattice site and the arrow indicates the m_z value of the created alt. destroyed electron. In the case of the Hubbard model on a bipartite lattice there is a second SU(2) symmetry available.[10,11] This symmetry, commonly called pseudo spin symmetry, is generated by the following operators:

$$J_z = \frac{1}{2}\sum_i (c_{i\uparrow}^\dagger c_{i\uparrow} + c_{i\downarrow}^\dagger c_{i\downarrow} - 1), \tag{2.5}$$

$$J_+ = \sum_i (-1)^\tau c_{i\uparrow}^\dagger c_{i\downarrow}^\dagger, \tag{2.6}$$

$$J_- = \sum_i (-1)^\tau c_{i\downarrow} c_{i\uparrow}, \tag{2.7}$$

$$J^2 = \frac{1}{2}(J_+ J_- + J_- J_+) + J_z^2. \tag{2.8}$$

In the above equations τ is 1 or 2 depending on which sub-lattice i belongs to. J_z is seen to be the charge operator while J_+ and J_- flip the pseudo spin, i.e., change a positive ion to a negative and vice versa. It can be directly verified that J_z and J^2 commute with the Hubbard Hamiltonian given by[2]

$$H = -t\sum_{i,\delta,\sigma} c_{i,\sigma}^\dagger c_{i+\delta,\sigma} + U\sum_i (c_{i\uparrow}^\dagger c_{i\uparrow} c_{i\downarrow}^\dagger c_{i\downarrow}), \tag{2.9}$$

where t is the hopping energy, U the on site Coulomb repulsion and $\sigma = \uparrow, \downarrow$. The following commutation relations can also be directly verified:

$$[J_z, J_\pm] = J_\pm, \tag{2.10}$$

$$[J_+, J_-] = 2J_z. \tag{2.11}$$

The states of the Hubbard model can therefore be simultaneously characterized by their spin and pseudo spin quantum numbers.

This does, however, not exhaust the available symmetry in most cases since, in general, the cluster under study possesses some space symmetry. In the following, attention will be given to the two dimensional square lattice with periodic boundary conditions but the discussion can equally well be applied to lattices of different shapes and dimensions. The space group of a square planar lattice with periodic boundary conditions has previously been studied in connection to the Hubbard model by Fano *et al.*[8] The derivation of the irreducible representations in the square-planar case is given in Ref. 8 but it should be noted that the irreducible representations of any symmorphic space group can be written down directly with aid of the following theorem.[12]

Let $\vec{k}$ be an allowed $\vec{k}$-vector of the Brillouin zone and $G_0(\vec{k})$ the subgroup of the point group G whose elements conserve $\vec{k}$ to within a reciprocal lattice vector. Let R_i be g/h left coset representatives for the decomposition of G with respect to $G_0(\vec{k})$, where g and h are the order of G and $G_0(\vec{k})$, respectively. Then the elements of the irreducible representations of the space group are given by

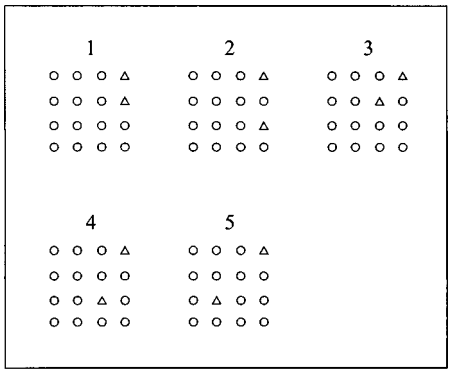

FIG. 1. All possible inequivalent pseudo spin configurations for 2 pseudo spins. ○ represents a spin while △ represents an ion or pseudo spin.

$$\Gamma^{\vec{k}p}(\vec{T},R)_{kt,jr}=\begin{cases} e^{-\{i(R_k\vec{k})\circ\vec{T}\}}\Gamma^{p}_{G_0(\vec{k})}(R_k^{-1}RR_j)_{t,r}, & \text{if } (R_k^{-1}RR_j)\in G_0(\vec{k}), \\ 0, \quad \text{otherwise.} \end{cases} \tag{2.12}$$

In the expression above, R denotes an element of the point group, $\vec{T}$ denotes a translation element of the space group and p denotes an irreducible representation of $G_0(\vec{k})$. The point group in the current case is C_{4v} and the space group of the example of a 4*4 lattice is given in the Appendix in Table II. The above theorem comes from the fact that the space group can be written as the semidirect product of the translation group and the point group. Since the translation group is an Abelian invariant group the method of induced representations, in form of the above theorem, generates the whole set of irreducible representations.

It is worth mentioning that the bipartite requirement for pseudospin symmetry excludes the possibility of including both pseudo spin and translational symmetry for square planar lattices with an odd number of sites, e.g., 3*3.

III. APPLICATION

In order to apply the above discussed symmetries, all possible permutations of spin and pseudo spin are first divided into sets of pseudo spin configurations which do not mix under the operations of the space group. In the example of a 4*4 lattice at half filling, set representatives of all the configurations with 2 pseudo spins are shown in Fig. 1. The representative for a given set of equivalent configurations is picked at random. A circle (○) indicates that the particular site is singly occupied (spin 1/2) and a triangle (△) indicates that the site is unoccupied or doubly occupied (pseudo spin 1/2). By applying all the space group elements to the shown configurations, one can show that none of the configurations mix and that all possible configurations are generated.

The space group symmetry elements which leave each set representative unchanged are listed in Table I, where each operation is written as a product of a translation and an operation of the point group. T_x^i and T_y^i signify an i step translation in the x and y directions, respectively, C_4^n a rotation through $n(\pi/2)$ and σ a reflection in a mirror plane.

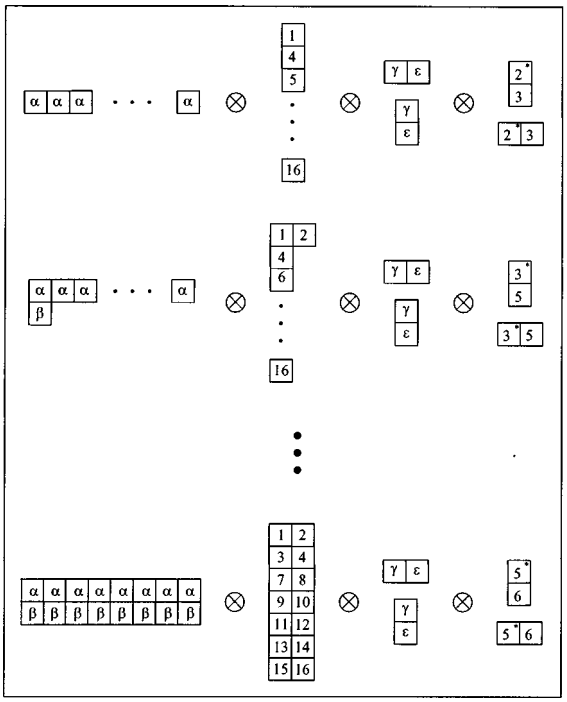

FIG. 2. A few examples of possible states with 2 pseudo spins and 14 spins. α and β denote spin up and spin down while γ and ε denote pseudo spin up and pseudo spin down. For the orbital diagrams the numbers signify that the corresponding orbital is occupied. The two possible pseudo spin configurations have been written out explicitly. In the orbital tableaux for the pseudo spin part, the * signifies double occupancy while no * signifies zero occupancy.

To find the states generated by a given pseudo spin configuration and having definite values of spin and pseudo spin, eigenstates of the spin and pseudo spin operators need first to be constructed. This is accomplished with the aid of the Young tableaux representation of the permutation group.[13,14] In Fig. 2 a few examples of the many possible states for 14 spins and 2 pseudo spins are shown. The highest spin state is given in each case while $J_z=0$ for all states at half filling. Since the total wavefunction must be antisymmetric with respect to any odd permutation only spin and orbital states represented by conjugate Young tableaux are possible as indicated in Fig. 2. In the construction of the states, the usual rules of regular application have been applied.[14] Once states of a given spin and pseudo spin have been obtained, only the orbital parts need to be kept since the Hamiltonian does not depend on spin. In order to proceed and calculate the states also belonging to a specific irreducible representation of the space group, it is necessary to find the characters of the representations generated by the states corresponding to a particular set of orbital Young diagram and pseudo spin configuration. At this point, the fact that the space group is isomorph with a subgroup of the permutation group of all the orbitals is used. By finding the permutation equivalent to the space group element the character generated by the set of pseudo spin configurations with a given spin and pseudo spin can thus be taken from the corresponding permutation group. In the present case this has been done by calculating the character of a given permutation corresponding to an allowed Young diagram or irreducible representation. For small permutation groups, the characters can be taken directly from tables. The characters needed for larger groups can be calculated directly, by e.g., regular applications or directly from the Frobenius formula.[14]

As an example we look at configuration one for two ions in Fig. 1. From Table I, it is seen that the configuration has three symmetry elements, excluding the identity, which belong to $\mathfrak{C}4$, $\mathfrak{C}8$ and $\mathfrak{C}9$ of the space group. Looking at Fig. 1, one finds that the space group operations correspond to the following permutations:

$$T_y^2\sigma_h \rightarrow 2^7 \text{ on spins and 2 on pseudospins,} \tag{3.1}$$

$$^1T_x^3T_y^2C_2 \rightarrow 2^7 \text{ on spins and } 2 \text{ on pseudospins,} \tag{3.2}$$

$$T_x^3\sigma_h' \rightarrow 2^4,1^6 \text{ on spins and } 1^2 \text{ on pseudospins.} \tag{3.3}$$

In the notation above, the permutation x^y denotes y cycles of length x. There are, however, still many spin and pseudo spin states corresponding to configuration two. If spin singlets and pseudo spin singlets, represented by Young diagrams of the type shown in the bottom line of Fig. 2, are chosen as an example then the generated representation has the following characters: 13728 0 0 280 0 0 0 140 104 0 0 0 0 0 0 0 0 0 0 0. The characters are calculated in the following way. The dimension of the representation is given by $g/h*f^{[\lambda]}$ where g is the order of the space group, h is the order of the group which leaves the configuration unchanged and $f^{[\lambda]}$ is the dimension of the irreducible representation corresponding to Young diagram $[\lambda]$ (including both spin and pseudospin). Since the dimension of the irreducible representation corresponding to the single occupation coordinate diagram in the bottom line of Fig. 2 is 429, each equivalent pseudo spin configuration is found to be 429 dimensional due to the isomorphism. As a result, the set of all equivalent pseudo spin configurations of type two generate a $128/4*429=13728$ dimensional representation. For $\mathfrak{C}4$ there are 8 equivalent pseudo spin configurations which remain unaffected by $T_y^2\sigma_h$. All of these equivalent configurations have the same permutational behavior given by expression 3.1. The characters of the irreducible representations in the bottom line of Fig. 2 for the permutation cycles 2^7 and 2 are 35 and 1, respectively. The current representation thus generates $8*35=280$ as a character for $\mathfrak{C}4$. The characters for $\mathfrak{C}8$ and $\mathfrak{C}9$ are found in a similar manner. In some cases the permutations corresponding to a particular space group operation are not the same for all equivalent configurations and care must be taken to calculate the contribution from each configuration separately. With the aid of Table II the found representation can now be decomposed into irreducible representations of the space group.

The same method can be applied to all allowed combinations of configurations, spins and pseudo spins. The result of the complete classification is given in the Appendix in Table IV. As seen from the table, the linear dimension of the largest matrix block which needs to be diagonalized has been reduced by approximately three orders of magnitude from the original 601080390 to 831398. This is one order of magnitude larger reduction than what can be achieved with application of only the space group. Assuming that $U>0$ and $t>0$, it is known[15] that the ground state is a spin and pseudo spin singlet with $k=0$ and the ground state can thus be found by diagonalizing a matrix with a linear dimension no larger than 107 358.

Due to the formidable number of states available it is not feasible to do exact diagonalization on Hubbard clusters much larger than 4*4. It is, however, possible to do an approximate treatment by including only states with a limited number of pseudo spins. This approximation is easily applied in the present context and should be applicable for $U \gg t$.

In order to obtain explicit expressions for the states found during classification, the standard projection operators of the space group are applied to the spin symmetry adapted states represented by the Young tableaux. The projection operators are given by

$$P_{ik}^{\chi_i} = c \sum_R \Gamma_{ik}^{\chi_i}(R)*R, \tag{3.4}$$

where c is a normalization constant, R an operation of the space group and $\Gamma^{\chi_i}(R)$ the matrix representation of operation R in the irreducible representation χ_i as given by (2.12). Applied to an arbitrary Young tableaux representing a given spin, pseudo spin and pseudo spin configuration, this will result in a completely symmetry adapted state. In the application of the projection operators, it is convenient to make use of the fact that the space group operations can be replaced by the equivalent permutation operations. In the case of pseudo spin configurations which possess some symmetry elements, the application of the projection operators can be written as[16]

$$\psi_{ik}^{\chi_i}(S,J) = c \sum_{R_a,Q,t} \Gamma_{ik}^{\chi_i}(R_aQ)*\Gamma_{tt_0}^{[S,J]}(Q^p)R_a\phi^{[S,J]}(t). \tag{3.5}$$

In the above expression, Q are the space group elements which leave the pseudo spin configuration unchanged, R_a are the coset representatives for the division of the space group into left cosets of Q and Q^p is the permutation corresponding to Q. The state $\phi^{[S,J]}(t)$ represents a Young tableaux of spin S and pseudo spin J.

IV. SUMMARY

A technique for finding the spin, pseudo spin and space group symmetry adapted states of Hubbard clusters has been demonstrated. The technique has been applied to classify the states of square planar clusters of size 4*4. The results presented show that the application reduces the dimensions of the involved matrices considerably. In the 4*4 case the reduction is approximately three orders of magnitude. These reductions are of great value especially when the complete or a large part of the energy spectrum is needed, e.g., in calculations of thermodynamic quantities. It should be noted that the method presented in this paper can equally well be applied to any Hamiltonian with all or part of the presented symmetry and with arbitrary filling.

APPENDIX A: TABLES FOR THE 4*4 CLUSTER

TABLE I. All symmetry elements, identity excluded, which leave the different pseudo spin configurations in Fig. 1 unchanged. The labeling of configurations correspond to that of Fig. 1.

Config.	Space group operations																			
1	$T_y^2\sigma_h$	$T_x^3T_y^2C_2$	$T_x^3\sigma_h'$																	
2	$T_y\sigma_h$	T_y^2	$T_y^3\sigma_h$	$T_x^3\sigma_h'$	$T_x^3T_yC_2$	$T_x^3T_y^2\sigma_h'$	$T_x^3T_y^3C_2$													
3	σ_v	$T_x^2T_y^2C_2$	$T_x^2T_y^2\sigma_v'$																	
4	$T_y^3\sigma_h$	$T_x^2T_yC_2$	$T_x^2T_y^2\sigma_h'$																	
5	σ_v	$T_y^3C_4^3$	$T_y^3\sigma_h$	$T_xT_yC_2$	$T_xT_y\sigma_v'$	$T_xT_y^2C_4$	$T_xT_y^2\sigma_h'$	$T_x^2T_yC_4^3$	$T_x^2T_y\sigma_h$	$T_x^2T_y^2$	$T_x^2T_y^2\sigma_v$	$T_x^3C_4$	$T_x^3\sigma_h'$	$T_x^3T_y^3C_2$	$T_x^3T_y\sigma_v'$					

TABLE II. The character table for the space group of a 4*4 square lattice with periodic boundary conditions. For each class the cycles of the corresponding permutation is also given. The superscript on each irreducible representation represents the values of the corresponding k-vector k_x, k_y. The notation for the operations is the same as in Table I with T_x^i and T_y^i signifying an i step translation in the x and y directions, respectively, C_4^n a rotation through $n(\pi/2)$, σ a reflection in a mirror plane and E the identity.

	$\mathfrak{C}_1$	$\mathfrak{C}_2$	$\mathfrak{C}_3$	$\mathfrak{C}_4$	$\mathfrak{C}_5$	$\mathfrak{C}_6$	$\mathfrak{C}_7$	$\mathfrak{C}_8$	$\mathfrak{C}_9$	$\mathfrak{C}_{10}$	$\mathfrak{C}_{11}$	$\mathfrak{C}_{12}$	$\mathfrak{C}_{13}$	$\mathfrak{C}_{14}$	$\mathfrak{C}_{15}$	$\mathfrak{C}_{16}$	$\mathfrak{C}_{17}$	$\mathfrak{C}_{18}$	$\mathfrak{C}_{19}$	$\mathfrak{C}_{20}$
	1^{16}	4^4	2^8	2^8	2^61^4	4^4	$4^32^12^1$	2^8	2^41^8	4^4	8^2	2^8	2^8	4^4	4^4	2^61^4	4^4	4^4	2^8	2^8
$\chi_1^{0,0}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\chi_2^{0,0}$	1	1	1	−1	−1	1	1	1	−1	−1	−1	1	−1	−1	1	1	−1	1	−1	1
$\chi_3^{0,0}$	1	−1	1	1	−1	1	−1	1	1	1	−1	1	1	−1	1	1	1	1	1	1
$\chi_4^{0,0}$	1	−1	1	−1	1	1	−1	1	−1	−1	1	1	−1	1	1	1	−1	1	−1	1
$\chi_5^{0,0}$	2	0	−2	0	0	2	0	−2	0	0	0	2	0	0	2	−2	0	2	0	2
$\chi_6^{0,1\pi/2}$	4	0	0	2	0	2	0	0	2	0	0	0	−2	0	0	0	0	−2	−2	−4
$\chi_7^{0,1\pi/2}$	4	0	0	−2	0	2	0	0	−2	0	0	0	2	0	0	0	0	−2	2	−4
$\chi_8^{0,1\pi}$	2	0	2	2	0	0	0	0	0	0	0	2	2	0	−2	−2	−2	0	0	2
$\chi_9^{0,1\pi}$	2	0	2	−2	0	0	0	0	0	0	0	2	−2	0	−2	−2	2	0	0	2
$\chi_{10}^{0,1\pi}$	2	0	−2	0	0	0	0	0	−2	2	0	2	0	0	−2	2	0	0	−2	2
$\chi_{11}^{0,1\pi}$	2	0	−2	0	0	0	0	0	2	−2	0	2	0	0	−2	2	0	0	2	2
$\chi_{12}^{1\pi/2,1\pi/2}$	4	0	0	0	2	0	0	0	0	0	0	−4	0	−2	0	0	0	0	0	4
$\chi_{13}^{1\pi/2,1\pi/2}$	4	0	0	0	−2	0	0	0	0	0	0	−4	0	2	0	0	0	0	0	4
$\chi_{14}^{1\pi/2,1\pi}$	4	0	0	2	0	−2	0	0	−2	0	0	0	−2	0	0	0	0	2	2	−4
$\chi_{15}^{1\pi/2,1\pi}$	4	0	0	−2	0	−2	0	0	2	0	0	0	2	0	0	0	0	2	−2	−4
$\chi_{16}^{1\pi,1\pi}$	1	1	1	1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	−1	−1	1
$\chi_{17}^{1\pi,1\pi}$	1	1	1	−1	−1	−1	−1	−1	1	1	1	1	−1	−1	1	1	−1	−1	1	1
$\chi_{18}^{1\pi,1\pi}$	1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	−1	1	1	1	−1	−1	1
$\chi_{19}^{1\pi,1\pi}$	1	−1	1	−1	1	−1	1	−1	1	1	−1	1	−1	1	1	1	−1	−1	1	1
$\chi_{20}^{1\pi,1\pi}$	2	0	−2	0	0	−2	0	2	0	0	0	2	0	0	2	−2	0	−2	0	2

TABLE III. All symmetry operations grouped into classes. The superscript i on some operations can take on all allowed values.

$\mathfrak{C}_1$	E							
$\mathfrak{C}_2$	C_4^i	$T_y^2C_4^i$	$T_xT_yC_4^i$	$T_xT_y^3C_4^i$	$T_x^2C_4^i$	$T_x^2T_y^2C_4^i$	$T_x^3T_yC_4^i$	$T_x^3T_y^3C_4^i$
$\mathfrak{C}_3$	C_2	$T_y^2C_2$	$T_x^2C_2$	$T_x^2T_y^2C_2$				
$\mathfrak{C}_4$	σ_h	σ_h'	$T_y^2\sigma_h$	$T_x^2\sigma_h'$				
$\mathfrak{C}_5$	σ_v	σ_v'	$T_xT_y\sigma_v'$	$T_xT_y^3\sigma_v$	$T_x^2T_y^2\sigma_v$	$T_x^2T_y^2\sigma_v'$	$T_x^3T_y\sigma_v$	$T_x^3T_y^3\sigma_v'$
$\mathfrak{C}_6$	T_y	T_y^3	T_x	T_x^3				
$\mathfrak{C}_7$	$T_yC_4^i$	$T_y^3C_4^i$	$T_xC_4^i$	$T_xT_y^2C_4^i$	$T_x^2T_yC_4^i$	$T_x^2T_y^3C_4^i$	$T_x^3C_4^i$	$T_x^3T_y^2C_4^i$
$\mathfrak{C}_8$	T_yC_2	$T_y^3C_2$	T_xC_2	$T_xT_y^2C_2$	$T_x^2T_yC_2$	$T_x^2T_y^3C_2$	$T_x^3C_2$	$T_x^3T_y^2C_2$
$\mathfrak{C}_9$	$T_y\sigma_h$	$T_y^3\sigma_h$	$T_x\sigma_h'$	$T_x^3\sigma_h'$				
$\mathfrak{C}_{10}$	$T_y\sigma_h'$	$T_y^3\sigma_h'$	$T_x\sigma_h$	$T_xT_y^2\sigma_h$	$T_x^2T_y\sigma_h'$	$T_x^2T_y^3\sigma_h'$	$T_x^3\sigma_h$	$T_x^3T_y^2\sigma_h$
$\mathfrak{C}_{11}$	$T_y\sigma_v^i$	$T_y^3\sigma_v^i$	$T_x\sigma_v^i$	$T_xT_y^2\sigma_v^i$	$T_x^2T_y\sigma_v^i$	$T_x^2T_y^3\sigma_v^i$	$T_x^3\sigma_v^i$	$T_x^3T_y^2\sigma_v^i$
$\mathfrak{C}_{12}$	T_y^2	T_x^2						
$\mathfrak{C}_{13}$	$T_y^2\sigma_h'$	$T_x^2\sigma_h$	$T_x^2T_y^2\sigma_h$	$T_x^2T_y^2\sigma_h'$				
$\mathfrak{C}_{14}$	$T_y^2\sigma_v$	$T_y^2\sigma_v'$	$T_xT_y\sigma_v$	$T_xT_y^3\sigma_v'$	$T_x^2\sigma_v$	$T_x^2\sigma_v'$	$T_x^3T_y\sigma_v'$	$T_x^3T_y^3\sigma_v$
$\mathfrak{C}_{15}$	T_xT_y	$T_xT_y^3$	$T_x^3T_y$	$T_x^3T_y^3$				
$\mathfrak{C}_{16}$	$T_xT_yC_2$	$T_xT_y^3C_2$	$T_x^3T_yC_2$	$T_x^3T_y^3C_2$				
$\mathfrak{C}_{17}$	$T_xT_y\sigma_h$	$T_xT_y\sigma_h'$	$T_xT_y^3\sigma_h$	$T_xT_y^3\sigma_h'$	$T_x^3T_y\sigma_h$	$T_x^3T_y\sigma_h'$	$T_x^3T_y^3\sigma_h$	$T_x^3T_y^3\sigma_h'$
$\mathfrak{C}_{18}$	$T_xT_y^2$	$T_x^2T_y$	$T_x^2T_y^3$	$T_x^3T_y^2$				
$\mathfrak{C}_{19}$	$T_xT_y^2\sigma_h'$	$T_x^2T_y\sigma_h$	$T_x^2T_y^3\sigma_h$	$T_x^3T_y^2\sigma_h'$				
$\mathfrak{C}_{20}$	$T_x^2T_y^2$							

TABLE IV. All states of the 4*4 periodic Hubbard cluster classified according to spin, pseudospin and space group representation.

2S+1	2J+1	χ_1	χ_2	χ_3	χ_4	χ_5	χ_6	χ_7	χ_8	χ_9	χ_{10}
1	1	56490	54272	56388	54356	107358	216906	217084	110652	108456	107530
1	3	105729	105969	105666	106032	214281	428322	425400	211635	213399	212883
1	5	76978	74996	76897	75083	150129	302240	300796	151893	150717	149491
1	7	28327	28011	28247	28095	57138	114604	112952	56262	56846	56398
1	9	6411	5895	6379	5915	11898	24300	23832	12270	12022	11686
1	11	696	646	681	663	1435	2920	2708	1327	1399	1345
1	13	68	33	65	37	71	177	150	98	80	61
1	15					3	7	2		2	1
1	17	1									
3	1	105729	105969	105666	106032	214281	428322	425400	211635	213399	212883
3	3	209986	206294	209770	206504	413862	831398	827392	416058	414594	412066
3	5	144185	142255	143963	142489	287538	577024	571844	286224	287100	285182
3	7	53494	51476	53347	51617	104067	210152	207412	104823	104319	102841
3	9	10680	10090	10595	10173	20937	42464	41116	20685	20853	20347
3	11	1270	1013	1241	1042	2160	4574	4246	2253	2191	2024
3	13	65	37	61	44	111	253	188	99	107	85
3	15	5		3	1		5	2	3	1	
5	1	76978	74996	76897	75083	150129	302240	300796	151893	150717	149491
5	3	144185	142255	143963	142489	287538	577024	571844	286224	287100	285182
5	5	96810	93826	96578	94040	189462	381890	377640	190398	189774	187554
5	7	32489	31169	32328	31330	63875	129070	126248	63497	63749	62625
5	9	6020	5323	5945	5401	11079	22855	21866	11268	11142	10661
5	11	527	401	507	422	939	2007	1752	909	929	833
5	13	33	8	24	13	22	66	46	31	25	18
7	1	28327	28011	28247	28095	57138	114604	112952	56262	56846	56398
7	3	53494	51476	53347	51617	104067	210152	207412	104823	104319	102841
7	5	32489	31169	32328	31330	63875	129070	126248	63497	63749	62625
7	7	9885	8928	9780	9033	18473	37898	36512	18707	18551	17883
7	9	1380	1140	1346	1177	2532	5309	4834	2487	2517	2332
7	11	101	46	85	59	110	275	222	131	117	98
9	1	6411	5895	6379	5915	11898	24300	23832	12270	12022	11686
9	3	10680	10090	10595	10173	20937	42464	41116	20685	20853	20347
9	5	6020	5323	5945	5401	11079	22855	21866	11268	11142	10661
9	7	1380	1140	1346	1177	2532	5309	4834	2487	2517	2332
9	9	153	78	130	95	183	436	364	207	191	165
11	1	696	646	681	663	1435	2920	2708	1327	1399	1345
11	3	1270	1013	1241	1042	2160	4574	4246	2253	2191	2024
11	5	527	401	507	422	939	2007	1752	909	929	833
11	7	101	46	85	59	110	275	222	131	117	98
13	1	68	33	65	37	71	177	150	98	80	61
13	3	65	37	61	44	111	253	188	99	107	85
13	5	33	8	24	13	22	66	46	31	25	18
15	1					3	7	2		2	1
15	3	5		3	1		5	2	3	1	
17	1	1									

TABLE IV. (*Continued*.)

2S+1	2J+1	χ_{11}	χ_{12}	χ_{13}	χ_{14}	χ_{15}	χ_{16}	χ_{17}	χ_{18}	χ_{19}	χ_{20}
1	1	109548	217078	216912	217084	216906	54364	54184	54264	54266	109554
1	3	214041	426924	426798	425400	428322	106032	107430	105969	107493	212517
1	5	152111	301602	301434	300796	302240	75083	75721	74996	75802	151305
1	7	57466	113864	113692	112952	114604	28091	28835	28015	28923	56554
1	9	12402	24088	24044	23832	24300	5919	6127	5891	6151	12146
1	11	1485	2830	2798	2708	2920	663	753	646	768	1363
1	13	106	167	160	150	177	37	47	33	50	89
1	15	4	5	4	2	7		2		3	1
1	17										
3	1	214041	426924	426798	425400	428322	106032	107430	105969	107493	212517
3	3	417536	829602	829188	827392	831398	206510	208300	206288	208504	415326
3	5	289486	574668	574200	571844	577024	142483	144845	142261	145079	286662
3	7	106085	208926	208638	207412	210152	51617	52843	51476	52990	104571
3	9	21527	41874	41706	41116	42464	10173	10763	10090	10848	20769
3	11	2414	4438	4382	4246	4574	1043	1178	1012	1205	2222
3	13	142	227	214	188	253	43	70	38	76	103
3	15	5	5	2	2	5	1	1		3	2
5	1	152111	301602	301434	300796	302240	75083	75721	74996	75802	151305
5	3	289486	574668	574200	571844	577024	142483	144845	142261	145079	286662
5	5	192428	379982	379548	377640	381890	94046	95948	93820	96168	190086
5	7	65195	127820	127498	126248	129070	31330	32580	31169	32741	63623
5	9	11776	22437	22284	21866	22855	5401	5819	5323	5894	11205
5	11	1068	1901	1858	1752	2007	421	528	402	550	919
5	13	44	62	50	46	66	14	17	7	24	28
7	1	57466	113864	113692	112952	114604	28091	28835	28015	28923	56554
7	3	106085	208926	208638	207412	210152	51617	52843	51476	52990	104571
7	5	65195	127820	127498	126248	129070	31330	32580	31169	32741	63623
7	7	19425	37308	37102	36512	37898	9034	9623	8927	9724	18629
7	9	2777	5109	5034	4834	5309	1176	1377	1141	1415	2502
7	11	165	263	234	222	275	59	71	46	87	124
9	1	12402	24088	24044	23832	24300	5919	6127	5891	6151	12146
9	3	21527	41874	41706	41116	42464	10173	10763	10090	10848	20769
9	5	11776	22437	22284	21866	22855	5401	5819	5323	5894	11205
9	7	2777	5109	5034	4834	5309	1176	1377	1141	1415	2502
9	9	253	418	382	364	436	96	113	77	132	199
11	1	1485	2830	2798	2708	2920	663	753	646	768	1363
11	3	2414	4438	4382	4246	4574	1043	1178	1012	1205	2222
11	5	1068	1901	1858	1752	2007	421	528	402	550	919
11	7	165	263	234	222	275	59	71	46	87	124
13	1	106	167	160	150	177	37	47	33	50	89
13	3	142	227	214	188	253	43	70	38	76	103
13	5	44	62	50	46	66	14	17	7	24	28
15	1	4	5	4	2	7		2		3	1
15	3	5	5	2	2	5	1	1		3	2
17	1										

[1] J. G. Bednorz and K. A. Müller, Z. Phys. B **64**, 189 (1986).
[2] J. Hubbard, Proc. R. Soc. London, Ser. A **276**, 238 (1963).
[3] P. W. Anderson, Science **235**, 1196 (1987).
[4] E. H. Lieb and F. Y. Wu, Phys. Rev. Lett. **20**, 1445 (1968).
[5] E. Dagotto, Rev. Mod. Phys. **66**, 763 (1994), and references therein.
[6] K. Binder and D. W. Heermann, *Monte Carlo Simulations in Statistical Physics*, Vol. 80 of Springer Series in Solid-State Sciences (Springer-Verlag, Berlin, 1992).
[7] C. Lanczos, J. Res. Natl. Bur. Stand. **45**, 255 (1950).
[8] G. Fano, F. Ortolani, and F. Semeria, Phys. Rev. B **46**, 1048 (1992).
[9] C. Noce and M. Cuoco, Phys. Rev. B **54**, 13047 (1996).
[10] C. N. Yang, Phys. Rev. Lett. **63**, 2144 (1989).
[11] S. C. Zhang, Phys. Rev. Lett. **65**, 120 (1990).
[12] J. F. Cornwell, *Group Theory in Physics, Techniques of Physics*, 5th ed. (Academic, London, 1993).
[13] A. Young, Proc. London Math. Soc. **28**, 255 (1928).
[14] N. Hamermesh, *Group Theory and its Application to Physical Problems* (Dover, New York, 1989).
[15] E. H. Lieb, Phys. Rev. Lett. **62**, 1201 (1989).
[16] I. G. Kaplan, *Symmetry of Many-Electron Systems* (Academic, New York, 1975).

The *q*-harmonic oscillator in a lattice model

Hans van Leeuwen[a] and Hans Maassen[b]
Department of Mathematics KUN, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

(Received 21 January 1998; accepted for publication 25 May 1998)

We give an explicit proof of the pair partitions formula for the moments of the *q*-harmonic oscillator, and of the claim made by Parisi that the *q*-deformed lattice Laplacian on the *d*-dimensional lattice tends to the *q*-harmonic oscillator in distribution for $d\to\infty$. © *1998 American Institute of Physics.*
[S0022-2488(98)00412-5]

I. INTRODUCTION

In 1994, Parisi published his paper ''*D*-dimensional arrays of Josephson junctions, spin glasses and *q*-deformed harmonic oscillators'' (Ref. 1). It describes a lattice model (to be called the ''Parisi model'' here) that shows a connection with the *q*-harmonic oscillator. We prove some of the claims and conjectures made by Parisi after setting the stage for them in some detail.

We shall work within the framework of noncommutative probability theory or ''quantum probability'' (see, for example, Refs. 2–6), to be described briefly in Sec. II. A fine example of noncommutative probability is free probability theory, as discovered by Voiculescu.

Here we are concerned with an interpolation between the classical Gaussian distribution and the ''free Gaussian'' distribution: Wigner's semicircle law. This is done by *q* deforming the quantum mechanical harmonic oscillator. In Sec. III we show how the *q*-deformed Fock space, as introduced by Bożejko and Speicher in Refs. 7 and 8, can be constructed in an algebraic way, starting from the *q*-deformed commutation relations. It is shown that the moments of the ground state distribution of the *q*-harmonic oscillator can be calculated as a sum over pair partitions interpolating nicely between the well-known moment formulas of the Gauss and Wigner distributions.

In Sec. IV we prepare for the Parisi model by introducing the *d*-dimensional lattice, $\mathcal{T}(N,d)$, and its boundary and coboundary operators. The Parisi model itself is introduced in Sec. V. Section VI shows how we can use pair partitions to describe the walks on $\mathcal{T}(N,d)$ that turn out to be relevant for the Parisi model. Using the possibility of describing walks in terms of partitions, we show in Sec. VII that the position operator of a *q*-harmonic oscillator has the same distribution as the normalized *q*-deformed lattice Laplacian in the Parisi model if we let $d\to\infty$.

In Sec. VIII we show that, in fact, sequences of independent *q*-Gaussians are present in the Parisi model, although on the grounds of Ref. 9 we hesitate to call them ''*q*-independent.''

II. NONCOMMUTATIVE PROBABILITY AND INDEPENDENCE

In this section we shall give a brief outline of noncommutative probability theory and the free probabilistic example.

We shall first describe the transition from classical to noncommutative probability theory.[3,4] This transition follows much the same recipe as the search for ''noncommutative versions'' of mathematical objects as initiated by Connes.[10] A classical probability space will be, following the axiomatic approach outlined by Kolmogorov, a triple $(\Omega,\Sigma,\mathbb{P})$, where Ω is a state space, Σ is a σ-algebra of events, and $\mathbb{P}$ is a probability measure on Σ. The transition of classical probability to noncommutative probability theory is preceded by a replacement of functions by multiplication operators: on $(\Omega,\Sigma,\mathbb{P})$ we can construct a commutative algebra of functions together with a state that contains the same information as $(\Omega,\Sigma,\mathbb{P})$ itself. Problems defined on the space can then be translated to algebraic problems and worked out in the commutative algebra of functions. Then we

[a]Electronic mail: H.van.Leeuwen@abp.nl
[b]Electronic mail: maassen@sci.kun.nl

 © 1998 American Institute of Physics

drop the commutativity requirement on the algebra. The noncommutative algebra can no longer be identified with a space, but it still represents some kind of probability theory.

Definition 1: A general probability space is a unital von Neumann algebra, $\mathcal{A}$, together with a normal state φ.

The events in a general probability space are the projections $p=p^*=p^2$ in $\mathcal{A}$, and the probability that p occurs is $\varphi(p)$. A self-adjoint element $X\in\mathcal{A}$ will be called a random variable and $\varphi(X)$ its expectation. The distribution, μ_X, of a random variable $X\in\mathcal{A}$ will be a linear functional on the ring of complex polynomials in one variable, $\mu_X:\mathbb{C}[x]\to\mathbb{C}$, that satisfies $\mu_X(f)=\varphi(f(X))$ for every $f\in\mathbb{C}[x]$. Such a linear functional uniquely determines a probability measure with compact support on the real line, which is also denoted by μ_X.

As an illustration we shall reconstruct the classical probability space from a commutative general probability space $(\mathcal{A},\varphi)$. First, if $(\Omega,\Sigma,\mathbb{P})$ is a classical probability space, then the associated general probability space $(\mathcal{A},\varphi)$ is $\mathcal{A}:=L^\infty(\Omega,\Sigma,\mathbb{P})$ and $\varphi(f):=\int f d\mathbb{P}$, the expectation of f with respect to $\mathbb{P}$. Now, we can *re*construct the events in Σ by considering all projections $p\in\mathcal{A}$, which are of the form $p=1_S$ for some $S\in\Sigma$. Up to equivalence we can then reconstruct the concrete realization of Ω we had.

In this paper we choose to extend the classical definition of independence in the following way.

Definition 2 (independence): Random variables $X_1,\ldots,X_m$ in a general probability space $(\mathcal{A},\varphi)$ are called independent if for all $n\leqslant m$ and polynomials $f_1,\ldots,f_n$, we have

$$\varphi(f_1(X_{k(1)})\cdots f_n(X_{k(n)}))=\varphi(f_1(X_{k(1)}))\cdots\varphi(f_n(X_{k(n)})), \tag{1}$$

provided $k\in\{1,\ldots,m\}^n$ has all components $k(i)$, $1\leqslant i\leqslant n$, different from each other.

A completely noncommutative notion paralleling the idea of independence is the concept of free independence of random variables.

Definition 3 (free independence): Random variables $X_1,\ldots,X_m$ in a general probability space $(\mathcal{A},\varphi)$ are called freely independent if for all $n\in\mathbb{N}$ and polynomials $f_1,\ldots,f_n$ we have

$$\forall i\leqslant n:\varphi(f_i(X_{k(i)}))=0\Rightarrow\varphi(f_1(X_{k(1)})\cdots f_n(X_{k(n)}))=0,$$

provided $k\in\{1,\ldots,m\}^n$ satisfies $k(1)\neq k(2)\neq k(3)\neq\cdots\neq k(n)$.

Proposition 4: Free independence implies independence.

Proof: Suppose $X_1,\ldots,X_m$ are freely independent. Note that (1) is valid for $n=1$. Suppose that (1) is valid for all polynomials $f_1,\ldots,f_{n-1}$ and different indices $k(1),\ldots,k(n-1)$. Now let $f_1,\ldots,f_n$ be given polynomials and $k(1),\ldots,k(n)$ be different indices. By g_j we denote the difference between f_j and the constant $\varphi(f_j(X_{k(j)}))$ so that $\varphi(g_j(X_{k(j)}))=0$. Then free independence implies

$$\varphi(g_1(X_{k(1)})\cdots g_n(X_{k(n)}))=0.$$

Writing out this product, we find

$$\varphi(f_1(X_{k(1)})\cdots f_n(X_{k(n)}))=-\sum_{\gamma\subsetneq\{1,\ldots,n\}}(-1)^{n-\#\gamma}\left(\prod_{i\notin\gamma}\varphi(f_i(X_{k(i)}))\right)\varphi\left(\prod_{j\in\gamma}f_j(X_{k(j)})\right).$$

Since γ contains less than n elements, the expectation of $\Pi_{j\in\gamma}f_j(X_{k(j)})$ factorizes by the induction hypothesis. So all the terms in the sum are equal apart from their sign. Since the sum of these signs is 1, the result follows. □

III. *q*-HARMONIC OSCILLATORS AND THE *q*-DEFORMED FOCK SPACE

In this section we show how the q-deformed Fock space, as introduced in the papers of Bożejko and Speicher,[7,8] is found naturally, starting from the commutation relation. We show how the moments of the q-harmonic oscillator can be expressed in terms of pair partitions.

We start from the operators $\mathbf{1}$ and $a_1,a_2,\ldots,a_m$, generating a *-algebra $\mathcal{A}_q$ and satisfying

$$a_ia_j^*-qa_j^*a_i=\delta_{i,j}\mathbf{1},\quad q\in(-1,1). \tag{2}$$

On $\mathcal{A}_q$ we introduce a linear functional φ_q that we require to satisfy

$$\varphi_q(\mathbf{1})=1, \quad \varphi_q(a_i^* a_{\mathbf{k}}^{\epsilon})=0, \quad \varphi_q(a_{\mathbf{k}}^{\epsilon} a_i)=0, \tag{3}$$

for $i,j\in\{1,...,m\}$, $\mathbf{k}\in\{1,...,m\}^n$, $n\in\mathbb{N}$, and $\epsilon\in\{1,*\}^n$, i.e., a sequence with length n of 1's and $*$'s. So, by $a_{\mathbf{k}}^{\epsilon}$ we mean the ordered product:

$$a_{\mathbf{k}}^{\epsilon}=\prod_{i=1}^{n} \lambda_i, \quad \text{where } \lambda_i=\begin{cases} a_{\mathbf{k}(i)}, & \text{if } \epsilon(i)=1, \\ a_{\mathbf{k}(i)}^*, & \text{if } \epsilon(i)=*. \end{cases}$$

Next, we show the connection between pair partitions and $\varphi_q(a_{\mathbf{k}}^{\epsilon})$. Let $S=\{1,...,n\}$, for some $n\in\mathbb{N}$. A sequence,

$$\Pi=\{(\alpha_1,\beta_1),(\alpha_2,\beta_2),\ldots,(\alpha_{n/2},\beta_{n/2})\}, \quad \text{for } n \text{ even},$$

such that $\cup_{i=1}^{n/2}\{\alpha_i,\beta_i\}=S$, will be called a *pair partition* of S. A pair partition will be ordered in the sense that $\alpha_i<\beta_i$ for all $i\in\{1,...,n/2\}$ and $\alpha_1<\cdots<\alpha_{n/2}$. The collection of all such pair partitions of the set S will be denoted by $\mathcal{P}_2(S)$ or $\mathcal{P}_2(n)$. For n odd this is just the empty set. A crossing in Π is a subset of Π with two elements, $\{(\alpha_i,\beta_i),(\alpha_j,\beta_j)\}$, which satisfies either $\alpha_i<\alpha_j<\beta_i<\beta_j$ or $\alpha_j<\alpha_i<\beta_j<\beta_i$. The set of all crossings of a partition Π can be conveniently labeled by

$$c(\Pi):=\{(i,j)\,|\,1\leqslant i<j\leqslant n/2,\{(\alpha_i,\beta_i),(\alpha_j,\beta_j)\} \text{ is a crossing}\}.$$

Lemma 5: For all $n\in\mathbb{N}$, $n\geqslant 1$, $\epsilon\in\{1,*\}^n$, and $\mathbf{h}\in\{1,...,m\}^n$, we have

$$\varphi_q(a_{\mathbf{h}}^{\epsilon})=\sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\Pi)} \prod_{(i,j)\in\Pi} \varphi_q(a_{\mathbf{h}(i)}^{\epsilon(i)} a_{\mathbf{h}(j)}^{\epsilon(j)}). \tag{4}$$

Note that in this notation (i,j) can stand for an element of Π or for a pair of elements. The context will make clear, which is meant.

A pair partition $\Pi\in\mathcal{P}_2(n)$ is compatible with ϵ, denoted $\Pi\sim\epsilon$, if pairs $(\alpha_i,\beta_i)\in\Pi$ are such that $\epsilon(\alpha_i)=1$ and $\epsilon(\beta_i)=*$ for all $i\in\{1,...,n/2\}$. Compatibility of Π with $\mathbf{h}$ means that for each pair $(\alpha_i,\beta_i)\in\Pi$ we have $\mathbf{h}(\alpha_i)=\mathbf{h}(\beta_i)$. Using this notion of compatibility we can rewrite relation (4) as follows:

$$\varphi_q(a_{\mathbf{h}}^{\epsilon})=\sum_{\substack{\Pi\in\mathcal{P}_2(n)\\ \Pi\sim\epsilon,\mathbf{h}}} q^{\#c(\Pi)}. \tag{5}$$

We shall now prove Lemma 5.

Proof: Suppose that $\epsilon\in\{1,*\}^n$ can be written as $(\sigma,1,*,\varrho)$, where $\sigma\in\{1,*\}^k$ and $\varrho\in\{1,*\}^l$, with $k+l+2=n$. Suppose, furthermore, that $\mathbf{i}\in\{1,...,m\}^k$ and $\mathbf{j}\in\{1,...,m\}^l$ with $(\mathbf{h}(1),\ldots,\mathbf{h}(n))=(\mathbf{i}(1),\ldots,\mathbf{i}(k),\mathbf{h}(k+1),\mathbf{h}(k+2),\mathbf{j}(1),\ldots,\mathbf{j}(l))$. Then the following holds:

$$\varphi_q(a_{\mathbf{h}}^{\epsilon})=\varphi_q(a_{\mathbf{i}}^{\sigma} a_{\mathbf{h}(k+1)} a_{\mathbf{h}(k+2)}^* a_{\mathbf{j}}^{\varrho})=\delta_{\mathbf{h}(k+1),\mathbf{h}(k+2)}\varphi_q(a_{\mathbf{i}}^{\sigma} a_{\mathbf{j}}^{\varrho})+q\varphi_q(a_{\mathbf{j}}^{\sigma} a_{\mathbf{h}(k+2)}^* a_{\mathbf{h}(k+1)} a_{\mathbf{j}}^{\varrho}).$$

This relation, together with (3), determines the left-hand side of (4), and therefore φ_q, completely.

We shall refer to the right-hand side of (5) as $F(a_{\mathbf{h}}^{\epsilon})$ putting $F(\mathbf{1})=1$, and shall show that F satisfies the same recursion relation. Let $\Pi\in\mathcal{P}_2(n)$. Then, for every pair $(i,j)\in\Pi$ we have $\varphi_q(a_{\mathbf{h}(i)}^{\epsilon(i)} a_{\mathbf{h}(j)}^{\epsilon(j)})\neq 0$ if and only if $\epsilon(i)=1$ and $\epsilon(j)=*$ and $\mathbf{h}(i)=\mathbf{h}(j)$. There are now two possibilities: $(k+1,k+2)\in\Pi$ or $(k+1,k+2)\notin\Pi$. In the first case $\Pi':=\Pi\backslash\{(k+1,k+2)\}$ is a partition of $\{1,...,k,k+3,...,n\}$ with $\#c(\Pi)=\#c(\Pi')$. In the second case there must be $i<k+1$ and $j>k+2$ such that $(i,k+2)$ and $(k+1,j)$ are crossing pairs in Π. It is then possible to remove the crossing and construct

$$\Pi''=(\Pi\backslash\{(i,k+2),(k+1,j)\})\cup\{(i,k+1),(k+2,j)\},$$

for which $\#c(\Pi'')=\#c(\Pi)-1$. Now we have $\Pi''\sim(\sigma,*,1,\varrho),\mathbf{h}''$, where $\mathbf{h}''=(\mathbf{i},\mathbf{h}(k+2),\mathbf{h}(k+1),\mathbf{j})$. Note that conversely every pair partition of n, compatible with $\boldsymbol{\epsilon}$ and $\mathbf{h}$ but not containing $(k+1,k+2)$, can be found in $\mathcal{P}_2(n)$ as an element with a crossing between the two pairs that contain $k+1$ and $k+2$. We find

$$
\begin{aligned}
F(a^{\boldsymbol{\epsilon}}_{\mathbf{h}}) &:= \sum_{\substack{\Pi\in\mathcal{P}_2(n)\\ \Pi\sim\boldsymbol{\epsilon},\mathbf{h}}} q^{\#c(\Pi)} \\
&= \sum_{\substack{\Pi\in\mathcal{P}_2(n)\\ \Pi\sim(\sigma,1,*,\varrho)\mathbf{h}\\ (k+1,k+2)\in\Pi}} q^{\#c(\Pi)} + \sum_{\substack{\Pi\in\mathcal{P}_2(n)\\ \Pi\sim(\sigma,1,*,\varrho)h\\ (k+1,k+2)\notin\Pi}} q^{\#c(\Pi)} \\
&= \delta_{\mathbf{h}(k+1),\mathbf{h}(k+2)} \sum_{\substack{\Pi'\in\mathcal{P}_2(n-2)\\ \Pi'\sim(\sigma,\varrho),(\mathbf{i},\mathbf{j})}} q^{\#c(\Pi')} + \sum_{\substack{\Pi''\in\mathcal{P}_2(n)\\ \Pi''\sim(\sigma,*,1,\varrho),\mathbf{h}''}} q^{\#c(\Pi'')+1} \\
&= \delta_{\mathbf{h}(k+1),\mathbf{h}(k+2)} F(a^{\sigma}_{\mathbf{i}}a^{\varrho}_{\mathbf{j}}) + qF(a^{\sigma}_{\mathbf{i}}a^{*}_{\mathbf{h}(k+2)}a_{\mathbf{h}(k+1)}a^{\varrho}_{\mathbf{j}}).
\end{aligned}
$$

Relation (5) now follows because the right-hand side $F(a^{\boldsymbol{\epsilon}}_{\mathbf{h}})$ satisfies the same recursion relation as $\varphi(a^{\boldsymbol{\epsilon}}_{\mathbf{h}})$ with the boundary conditions (3). □

Since the position operator of a harmonic oscillator is usually represented by an operator of the form

$$X_i := a_i + a_i^*,$$

we define X_i to be the position of a q-harmonic oscillator and show that its moments under φ_q can be calculated as a sum over partitions. The operators X_i, $i\in\{1,...,m\}$, generate the *-algebra $\mathfrak{B}_q\subset\mathcal{A}_q$. We shall refer to $\mathfrak{B}_q$ as the q-harmonic oscillator algebra. Let ρ_q denote the restriction of φ_q to $\mathfrak{B}_q$.

Theorem 6: For all $\mathbf{h}\in\{1,...,m\}^n$, $n\in\mathbb{N}$, we have

$$\rho_q(X_{\mathbf{h}}) = \sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\Pi)} \prod_{(l,m)\in\Pi} \delta_{\mathbf{h}(l),\mathbf{h}(m)}.$$

Proof: This follows from Lemma 5 by summation over all $\boldsymbol{\epsilon}\in\{1,*\}^n$. □

Corollary 7: The linear functional ρ_q satisfies the following.

(1) For all $j\in\{1,...,m\}$ and $n\in\mathbb{N}$, we have

$$\rho_q(X_j^n) = \sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\Pi)}.$$

(2) For $\mathbf{i}\in\{1,...,m\}^n$ and a cyclic permutation $\tau\in S_n$, we have

$$\rho_q(X_{\mathbf{i}(\tau(1))}\cdots X_{\mathbf{i}(\tau(n))}) = \rho_q(X_{\mathbf{i}(1)}\cdots X_{\mathbf{i}(n)}).$$

Proof of 1: Put $\mathbf{h}(1)=\cdots=\mathbf{h}(n)=j$ in Theorem 6. □

Before proving (2) we first give some considerations concerning pair partitions.

The usual way to visualize a partition, $\Pi\in\mathcal{P}_2(S)$, would be to draw the elements of S on a straight line and connect every two points belonging to the same pair in Π with an arc above the

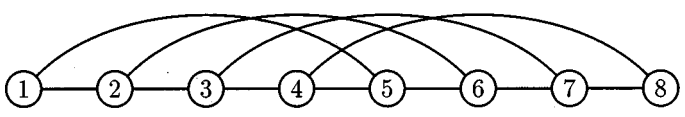

FIG. 1. The line representation of the partition $\Pi=\{(1,5),(2,6),(3,7),(4,8)\}$.

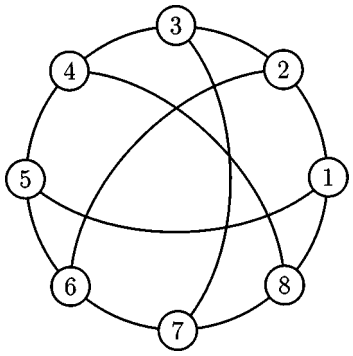

FIG. 2. A circle representation of the partition $\Pi=\{(1,5),(2,6),(3,7),(4,8)\}$.

line in such a way that two different lines cross at most once and no three lines intersect in one point. As an example, in Fig. 1 we draw the partition $\Pi=\{(1,5),(2,6),(3,7),(4,8)\}$. We will refer to this method of visualization as the line representation of a pair partition.

Another way to visualize a pair partition is its circle representation. This consists in drawing the points of S on a circle and connecting them by lines inside the circle, subject to the same restrictions. An example is given in Fig. 2. We note that this can be done in more than one way.

To make the circle representation of pair partitions explicit, we regard the circle as the unit circle in $\mathbb{C}$, and we make the map

$$\mathfrak{T}:S\to\mathbb{C}:s\mapsto e^{2i(s-1)\pi/n}.$$

This map converts every pair $(\alpha,\beta)\in\Pi$ to a pair (f,g) on the unit circle in $\mathbb{C}$.

Definition 8: Two pairs (f,g) and (f',g') of different points on the unit circle in $\mathbb{C}$ are said to be separated if and only if the straight line from f to g crosses the straight line from f' to g' inside the unit circle.

(See Fig. 3.) It is obvious that two pairs (α_i,β_i) and (α_j,β_j) in a partition Π are crossing if and only if the pairs $(\mathfrak{T}(\alpha_i),\mathfrak{T}(\beta_i))$ and $(\mathfrak{T}(\alpha_j),\mathfrak{T}(\beta_j))$ are separated. Let $\tau\in S_n$ be a cyclic permutation and $\Pi\in\mathcal{P}_2(n)$ any pair partition. By $\tau(\Pi)$ we denote the rotated partition:

$$\tau(\Pi):=\{(\tau(\alpha_1),\tau(\beta_1)),\ldots,(\tau(\alpha_{n/2}),\tau(\beta_{n/2}))\}.$$

Since separated pairs on the unit circle remain separated under rotation, we have

$$\#c(\Pi)=\#c(\tau(\Pi)). \tag{6}$$

We now turn to the proof of the second part of Corollary 7.

Proof of (2): The essence of the proof that ρ_q has this cyclic property is the fact that the circle representation of a partition $\Pi\in\mathcal{P}_2(n)$ also has this cyclic property. This is shown in Fig. 4.

More formally, for $\mathbf{i}\in\{1,\ldots,m\}^n$ and $\tau\in S_n$ cyclic, we have the following:

FIG. 3. Two separated and two nonseparated pairs.

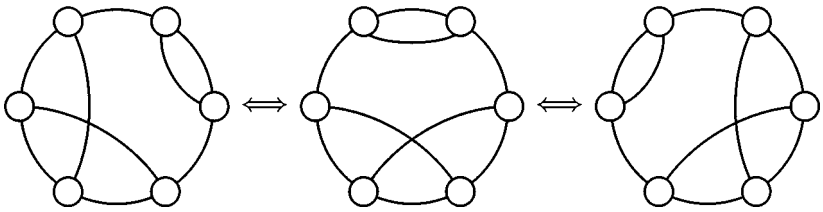

FIG. 4. The linear functional ρ_q satisfies a cyclic property.

$$\rho_q(X_{\mathbf{i}(\tau(1))}\cdots X_{\mathbf{i}(\tau(n))}) = \sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\Pi)} \prod_{(l,m)\in\Pi} \delta_{\mathbf{i}(\tau(l)),\mathbf{i}(\tau(m))}$$

$$= \sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\Pi)} \prod_{(l,m)\in\tau(\Pi)} \delta_{\mathbf{i}(l),\mathbf{i}(m)}$$

$$= \sum_{\Pi\in\mathcal{P}_2(n)} q^{\#c(\tau^{-1}(\Pi))} \prod_{(l,m)\in\Pi} \delta_{\mathbf{i}(l),\mathbf{i}(m)} = \rho_q(X_{\mathbf{i}(1)}\cdots X_{\mathbf{i}(n)}).$$

We shall represent the pair $(\mathcal{A}_q,\varphi_q)$ on a Hilbert space $\mathcal{K}$ with inner product $\langle\cdot,\cdot\rangle_q$ in which a unit vector Ψ, from now on referred to as vacuum vector, is singled out. To make this representation explicit, let $\mathfrak{h}$ be a Hilbert space with $\dim\mathfrak{h}=m$ and orthonormal basis $\{e_1,\ldots,e_m\}$. Because of (3) we require that for all $i\in\{1,\ldots,m\}$:

$$a_i\Psi = 0.$$

If now we put for all $\mathbf{k}\in\{1,\ldots,m\}^n$:

$$a^*_{\mathbf{k}(1)}a^*_{\mathbf{k}(2)}\cdots a^*_{\mathbf{k}(n)}\Psi =: e_{\mathbf{k}(1)}\otimes\cdots\otimes e_{\mathbf{k}(n)},$$

then we see that for $\mathcal{K}$ we can take the full Fock space over $\mathfrak{h}$, denoted $\mathcal{F}(\mathfrak{h})$:

$$\mathcal{F}(\mathfrak{h}) := \bigoplus_{i=1}^{\infty} \mathfrak{h}^{\otimes i}, \quad \mathfrak{h}^{\otimes 0} := \mathbb{C},$$

with vacuum vector $\Psi = 1\oplus 0\oplus 0\oplus\cdots$. We define a bilinear form $\langle\cdot,\cdot\rangle_q$ on $\mathcal{K}$ as follows:

$$\langle e_{\mathbf{k}(1)}\otimes\cdots\otimes e_{\mathbf{k}(n)}, e_{\mathbf{l}(1)}\otimes\cdots\otimes e_{\mathbf{l}(n')}\rangle_q = \langle a^*_{\mathbf{k}(1)}\cdots a^*_{\mathbf{k}(n)}\Psi, a^*_{\mathbf{l}(1)}\cdots a^*_{\mathbf{l}(n')}\Psi\rangle_q$$

$$:= \varphi_q(a_{\mathbf{k}(n)}\cdots a_{\mathbf{k}(1)}a^*_{\mathbf{l}(1)}\cdots a^*_{\mathbf{l}(n')})$$

$$= \sum_{\Pi\in\mathcal{P}_2(n+n')} q^{\#c(\Pi)} \prod_{(p,q)\in\Pi} \varphi_q(a^{\epsilon(p)}_{\mathbf{h}(p)}a^{\epsilon(q)}_{\mathbf{h}(q)}),$$

for $l\in\{1,\ldots,m\}^{n'}$, $n'\in\mathbb{N}$, $\mathbf{h}=(\mathbf{k}(n),\ldots,\mathbf{k}(1),\mathbf{l}(1),\ldots,l(n))\in\{1,\ldots,m\}^{n+n'}$, and $\epsilon\in\{1,*\}^{n+n'}$, such that $\epsilon(1)=\cdots=\epsilon(n)$ and $\epsilon(n+1)=\cdots=\epsilon(n+n')=*$. Bożejko and Speicher in Ref. 7 show positivity of the bilinear form $\langle\cdot,\cdot\rangle_q$ by proving positive definiteness of the function $S_n\to\mathbb{C}$: $\sigma\mapsto q^{\#i(\sigma)}$. Here the set $i(\sigma)$ is the set of inversions of the permutation $\sigma\in S_n$:

$$i(\sigma) := \{(l,m)\,|\,1\leq l<m\leq n, \sigma(l)>\sigma(m)\}.$$

Because of its positivity, the bilinear form $\langle\cdot,\cdot\rangle_q$ can be regarded as an inner product on $\mathcal{K}$, so we conclude that the linear functional φ_q is, in fact, a state. Therefore the second part of Corollary 7 implies that ρ_q is a trace state.

Since $\varphi_q(a^{\epsilon(p)}_{\mathbf{h}(p)}a^{\epsilon(q)}_{\mathbf{h}(q)})$ only yields something different from 0 for $\epsilon(p)=1$ and $\epsilon(q)=*$, we know that $\langle\cdot,\cdot\rangle_q\neq 0$ only in case $n=n'$, and that the partitions $\Pi\in\mathcal{P}_2(2n)$ contributing to this

inner product have to be compatible with ϵ. This only happens for partitions where, for all $i \in \{1,...,n\}$, we have $1 \leqslant \alpha_i \leqslant n$ and $n+1 \leqslant \beta_i \leqslant 2n$. There are exactly $n!$ of these partitions in $\mathcal{P}_2(2n)$, to be labeled by permutations σ of $\{1,...,n\}$. We find

$$\varphi_q(a_{\mathbf{h}}^{\epsilon}) = \delta_{n,n'} \sum_{\substack{\Pi \in \mathcal{P}_2(2n) \\ \Pi \sim \epsilon}} q^{\#c(\Pi)} \prod_{(p,q)\in\Pi} \delta_{\mathbf{h}(p),\mathbf{h}(q)} = \delta_{n,n'} \sum_{\sigma \in S_n} q^{\#i(\sigma)} \delta_{\mathbf{h}(1),\mathbf{h}(\sigma(1))} \cdots \delta_{\mathbf{h}(n),\mathbf{h}(\sigma(n))}.$$

Repeated use of the commutation relations (2) and the fact that $a_i \Psi = 0$ for all $i \in \{1,...,m\}$ yields, for all $\mathbf{h} \in \{1,...,m\}^n$, the following action of a_i:

$$a_i e_{\mathbf{h}(1)} \otimes \cdots \otimes e_{\mathbf{h}(n)} := \sum_{j=1}^{n} q^{j-1} \delta_{i,\mathbf{h}(j)} e_{\mathbf{h}(1)} \otimes \cdots \check{e}_{\mathbf{h}(j)} \cdots \otimes e_{\mathbf{h}(n)},$$

where by $e_{\mathbf{h}(1)} \otimes \cdots \check{e}_{\mathbf{h}(j)} \cdots \otimes e_{\mathbf{h}(n)}$ we mean the tensor product $e_{\mathbf{h}(1)} \otimes \cdots \otimes e_{\mathbf{h}(j-1)} \otimes e_{\mathbf{h}(j+1)} \otimes \ldots \otimes e_{\mathbf{h}(n)}$.

IV. THE DISCRETE LATTICE $\mathcal{T}(N,d)$

For d, $N \in \mathbb{N}$ we will define the Parisi model on a discrete d-dimensional lattice:

$$\mathcal{T}(N,d) := \{-N,\ldots,0,...,N\}^d.$$

Define the set of h-dimensional, $h \in \mathbb{N}$, elementary facets (i.e., points, edges, planes, etc.) in $\mathcal{T}(N,d)$ as follows:

$$\mathcal{X}_h := \{(v,\gamma) \in \mathcal{T}(N,d) \times \mathcal{P}_h(\{1,...,d\}) | j \in \gamma \Rightarrow v(j) \neq N\},$$

where $\mathcal{P}_h(\{1,...,d\})$ is the collection of all subsets of $\{1,...,d\}$ that contain exactly h elements. In this notation (v,γ) stands for the elementary facet, which has $v \in \mathcal{T}(N,d)$ as its lowest vertex, and whose spatial orientation is defined by the set of spatial directions $\gamma \subset \{1,2,...,d\}$. Then $\mathcal{X}_0$ stands for the set $\mathcal{T}(N,d)$ of vertices itself, $\mathcal{X}_1$ is the set of edges in $\mathcal{T}(N,d)$, and $\mathcal{X}_2$ is the set of two-dimensional planes in $\mathcal{T}(N,d)$, from now on referred to as plaquettes. By $\mathcal{G}_h$ we shall mean the set of all functions from $\mathcal{X}_h$ to $\mathbb{Z}$:

$$\mathcal{G}_h := \mathbb{Z}^{\mathcal{X}_h},$$

and by Ω_h the set of all functions from $\mathcal{X}_h$ to $\mathbb{C}$:

$$\Omega_h := \mathbb{C}^{\mathcal{X}_h}.$$

Then Ω_h is the set of ''forms'' on $\mathcal{X}_h$. They are a discrete analog of the differential forms of Cartan. The set Ω of all differential forms will be denoted by $\Omega := \oplus_{h=0}^{\infty} \Omega_h$. Similarly, we define $\mathcal{G} := \oplus_{h=0}^{\infty} \mathcal{G}_h$. Note that $\#(\mathcal{X}_h) = \dim \mathcal{G}_h = \dim \Omega_h = \binom{d}{h}(2N)^h(2N+1)^{d-h}$. An element of $\mathcal{G}_1$ will be called a curve [on $\mathcal{T}(N,d)$] and an element of $\mathcal{G}_2$ will be called a surface [on $\mathcal{T}(N,d)$]. The mapping,

$$L: \mathcal{G}_1 \to \mathbb{N}: l \mapsto \sum_{j \in \mathcal{X}_1} |l(j)|,$$

associates to a curve its length, and the mapping,

$$A: \mathcal{G}_2 \to \mathbb{N}: k \mapsto \sum_{j \in \mathcal{X}_2} |k(j)|,$$

sends a surface to its area. By a *walk* of length $n \in \mathbb{N}$ on $\mathcal{T}(N,d)$ we shall mean a series of consecutively neighboring points $(\alpha_1,\ldots,\alpha_n)$ in $\mathcal{T}(N,d)$ that trace out a curve $l \in \mathcal{G}_1$. This curve l assigns to every edge $x \in \mathcal{X}_1$, the number of times the walk $(\alpha_1,\ldots,\alpha_n)$ runs through x in positive direction minus the number of times this walk runs through x in the negative direction. So

a curve is thought of as just any configuration of edges equipped with a direction and a multiplicity. A surface is conceived of analogously. A walk will be called *closed* if $\alpha_1=\alpha_n$.

We define the integral of an h-form $\omega\in\Omega_h$ over some element $k\in\mathcal{G}_h$ as follows.

Definition 9:

$$\int_k \omega := \langle k,\omega\rangle = \sum_{x\in\mathcal{X}_h} k(x)\omega(x).$$

We define the boundary and coboundary operators $\partial_h:\mathcal{G}_h\to\mathcal{G}_{h-1}$ and $\delta_h:\Omega_h\to\Omega_{h+1}$ as follows:

$$(\partial_h g)(v,\gamma) := \sum_{j\notin\gamma} \epsilon(j,\gamma)\big(g(v-e_j,\gamma\cup\{j\})-g(v,\gamma\cup\{j\})\big) \quad h\geqslant 1,$$

$$(\delta_h f\ )(v,\gamma) := \sum_{j\in\gamma} \epsilon(j,\gamma\backslash\{j\})\big(f(v+e_j,\gamma\backslash\{j\})-f(v,\gamma\backslash\{j\})\big), \quad h\geqslant 0,$$

where $\epsilon(j,\varrho):=(-1)^{\#\{i\in\varrho|i<j\}}$ and e_j, $j\in\{1,...,d\}$, denotes the unit vector in the jth direction in $\mathcal{T}(N,d)$. We define the operator δ on Ω as δ_h for a differential form in Ω_h, so $\delta_h=\delta|_{\Omega_h}$. The operator δ can be considered the discrete version of the derivative operator on h forms. The operator ∂ on $\mathcal{G}$, defined as $\partial_h=\partial|_{\mathcal{G}_h}$, will be referred to as the boundary operator, since it yields the $(h-1)$-dimensional boundary of an h-dimensional object for $h\geqslant 1$. Stokes' theorem is the statement that δ and ∂ are each other's adjoints.

Theorem 10 (Stokes): For $\omega\in\Omega_h$ and $k\in\mathcal{G}_{h+1}$, we have

$$\int_k \delta\omega = \int_{\partial k}\omega.$$

Proof: By $1_{(v,\gamma)}$ we denote the characteristic function of $(v,\gamma)\in\mathcal{X}_1$. Then to prove this theorem it suffices to show that $\langle\partial 1_{(v,\gamma)},\omega\rangle=\langle 1_{(v,\gamma)},\delta\omega\rangle$ for every $(v,\gamma)\in\mathcal{X}_{h+1}$ and $\omega\in\Omega_h$:

$$\begin{aligned}\langle\partial 1_{(v,\gamma)},\omega\rangle &= \sum_{(w,\varrho)\in\mathcal{X}_h}(\partial 1_{(v,\gamma)})(w,\varrho)\omega(w,\varrho)\\ &= \sum_{(w,\varrho)\in\mathcal{X}_h}\sum_{j\notin\varrho}\epsilon(j,\varrho)\big(1_{(v,\gamma)}(w-e_j,\varrho\cup\{j\})-1_{(v,\gamma)}(w,\varrho\cup\{j\})\big)\omega(w,\varrho)\\ &= \sum_{j\in\gamma}\epsilon(j,\gamma\backslash\{j\})\big(\omega(v+e_j,\gamma\backslash\{j\})-\omega(v,\gamma\backslash\{j\})\big)=(\delta\omega)(v,\gamma)=\langle 1_{(v,\gamma)},\delta\omega\rangle. \quad \square\end{aligned}$$

Theorem 11: $\delta^2=0$.

Proof: For $\omega\in\Omega_h$ and $i<j$ we have, for all $(v,\gamma)\in\mathcal{X}_{h+2}$:

$$(\delta_{h+1}\delta_h\omega)(v,\gamma)=\sum_{i\in\gamma}\epsilon(i,\gamma)\big(\delta_h\omega(v+e_i,\gamma\backslash\{i\})-\delta_h\omega(v,\gamma\backslash\{i\})\big)=\sum_{\substack{i,j\in\gamma\\ i\neq j}}F(i,j)G(i,j),$$

where

$$F(i,j)=\epsilon(i,\gamma)\epsilon(j,\gamma\backslash\{i\})$$

and

$$G(i,j)=\omega(v+e_i+e_j,\gamma\backslash\{i,j\})-\omega(v+e_i,\gamma\backslash\{i,j\})-\omega(v+e_j,\gamma\backslash\{i,j\})+\omega(v,\gamma\backslash\{i,j\}).$$

Note that G is symmetric and F is antisymmetric:

$$F(i,j)F(j,i)=\epsilon(i,\gamma)\epsilon(j,\gamma\backslash\{i\})\epsilon(j,\gamma)\epsilon(i,\gamma\backslash\{j\})$$
$$=(-1)^{\#\{k\in\gamma|k<i\}}(-1)^{\#\{k\in\gamma\backslash\{i\}|k<j\}}(-1)^{\#\{k\in\gamma|k<j\}}(-1)^{\#\{k\in\gamma\backslash\{j\}|k<i\}}=-1.$$

Therefore $\delta_{h+1}\delta_h\omega=\Sigma_{i,j\in\gamma}F(i,j)G(i,j)=0$ for all $\omega\in\Omega_h$. □

Because $\partial=\delta^*$, viewed as $(\#\mathcal{X}_h)\times(\#\mathcal{X}_{h+1})$ matrices, Corollary 12 immediately follows.

Corollary 12: $\partial^2=0$.

Definition 13: An element $p\in\mathcal{G}_h$ is called closed if $\partial_h(p)=0$.

We shall call a closed curve a *loop*. The boundary of a surface k is always a loop since $\partial^2 k=0$. We shall say that $\partial_2 k$ *spans* k. With every loop l we can associate a class of surfaces $\Gamma_l:=\{k\in\mathcal{G}_2|\partial_2 k=l\}$, the class of surfaces spanned by l. A surface $p\in\Gamma_l$ will be called *minimal*, with respect to l, if its area $A(p)$ is minimal in $A(\Gamma_l)$.

Definition 14: By the area $A(l)$ of a loop we shall mean the area of a surface that is minimal with respect to l.

A closed walk traces out a loop. By the area of a closed walk $(\alpha_1,\ldots,\alpha_n=\alpha_1)$ we shall mean the area of the loop traced out by this walk, and denote it by $A((\alpha_1,\ldots,\alpha_n))$.

V. THE PARISI MODEL

In the description of the Parisi model we try to stay as close as possible to the notation used in Ref. 1.

Consider $\mathcal{T}(N,d)$ and put on every plaquette a magnetic field, i.e., we define a 2-form $B\in\Omega_2$ with strength $\mathcal{B}\in[0,\pi]$, the sign of which will be chosen independently for every pair of spatial directions:

$$B(v,\{i,j\})=S_{\{i,j\}}\mathcal{B},$$

where $S_{\{i,j\}}$ is a random variable depending on the pair $\{i,j\}$. The random variable $S_{\{i,j\}}$ is a coin toss, i.e., the value of $S_{\{i,j\}}$ is chosen from $\{-1,1\}$ with distribution $\{1/2,1/2\}$. It is obvious that, for any of the $2^{d(d-1)/2}$ choices for S, the constant field B on $\mathcal{T}(N,d)$ is divergence-free: $\delta B=0$. We put $q:=\cos\mathcal{B}$; then $q\in[-1,1]$.

By $\Phi(k)$ we denote the magnetic flux through some surface $k\in\mathcal{G}_2$. This flux is simply the sum of the fluxes through the plaquettes in k, which, by Definition 9, equals

$$\Phi(k)=\sum_{x\in\mathcal{X}_2}k(x)B(x)=\int_k B.$$

We define the magnetic flux enclosed by a loop $l\in\mathcal{G}_1$ as the flux $\Phi(k)$ for some k spanned by l, and denote it by $\Phi(l)$. Since the field B is divergence-free there must exist some $A\in\Omega_1$ for which $\delta_1 A=B$. Indeed, if, for $k\in\{1,\ldots,d\}$ we choose any $\mathcal{C}_0,\ldots,\mathcal{C}_d\in\mathbb{R}$, and put

$$A(w,\{k\})=\mathcal{C}_0+\sum_{j=1}^{k-1}(\mathcal{C}_j+w(j)S_{\{j,k\}}\mathcal{B});$$

then the requirement $\delta_1 A=B$ is fulfilled, as can be checked easily by calculating $(\delta_1 A)$ at $(v,\{i,j\})$ explicitly. By Stokes' theorem we find, for $\Phi(l)$,

$$\Phi(l)=\int_k\delta A=\int_l A=\sum_{x\in\mathcal{X}_1}l(x)A(x),$$

for all $k\in\Gamma_l$. So $\Phi(l)$ is well defined and does not depend on the choice of k.

Let the magnetic field B on $\mathcal{T}(N,d)$ induce a deformation of the nearest-neighbor $(2N+1)^d\times(2N+1)^d$-interaction matrix or lattice Laplacian, leading to the deformed lattice Laplacian Δ_q. This matrix $(\Delta_q)_{v,w\in\mathbb{N}^d}$ is defined by

$$|(\Delta_q)_{v,w}|=\begin{cases}1, & \text{if } |v-w|=1, \quad \text{for } v,w\in\mathcal{X}_0,\\ 0, & \text{otherwise,}\end{cases}$$

where the phases are determined by the field

$$(\Delta_q)_{v,v+e_k} := e^{iA(v,\{k\})},$$

$$(\Delta_q)_{v+e_k,v} := e^{-iA(v,\{k\})},$$

for all $(v,\{k\}) \in \mathcal{X}_1$. From this definition it is clear that $(\Delta_q)_{v,w} = \overline{(\Delta_q)_{w,v}}$. Now let

$$W(l) := e^{i\Phi(l)} = \prod_{x \in \mathcal{X}_1} e^{il(x)A(x)}$$

be the product of the Δ_q's along a loop l. $W(l)$ is known as the ''Wilson loop,'' although $W(l)$ itself is not a loop but a complex number assigned to the loop l. For a closed walk l' that traces out a loop $l \in \mathcal{G}_1$ we have that the product of the Δ_q's along l' equals the product of the Δ_q's along l, hence $W(l') = W(l)$.

We define an operator $\hat{X}$ as follows:

$$\hat{X} = \frac{1}{\sqrt{2d}} \Delta_q, \tag{7}$$

where the overcaret symbolizes the dependence on N and d. Then $\hat{X}$ is an element in the algebra $\mathcal{A}$ of matrix-valued functions,

$$\{-1,+1\}^{\binom{d}{2}} \to M_{(2N+1)^d}, \tag{8}$$

where $\{-1,+1\}^{\binom{d}{2}}$ is the space of outcomes of the coin tosses and $M_{(2N+1)^d}$ denotes the $(2N+1)^d \times (2N+1)^d$ matrices with complex entries. On $M_{(2N+1)^d}$ we have a normalized trace $\mathrm{tr} = (1/(2N+1)^d)\,\mathrm{Tr}$, where Tr is the standard trace on $M_{(2N+1)^d}$. Since we have here a trace that satisfies $\mathrm{tr}(\mathbf{1}) = 1$, its expectation can serve as a generalized probability measure (a state) $\hat{\varphi}$ on $\mathcal{A}$, and it is therefore possible to calculate generalized expectations of elements of $\mathcal{A}$:

$$\hat{\varphi}(\hat{X}^n) := \hat{\mathbb{E}}(\mathrm{tr}(\hat{X}^n)) = \frac{1}{(2d)^{n/2}(2N+1)^d}\, \hat{\mathbb{E}}(\mathrm{Tr}(\Delta_q^n)), \quad \text{for all } n \in \mathbb{N}.$$

Here $\hat{\mathbb{E}}$ yields the expectation value with respect to the $d(d-1)/2$ coin tosses. The standardization factor $1/\sqrt{2d}$ in (7) ensures that $\hat{X}$ has variance 1 in the limit $N \to \infty$. It remains to show that indeed $\hat{\varphi}(\hat{X}^n)$ can be interpreted as a sum over walks on $\mathcal{T}(N,d)$:

$$\hat{\varphi}(\hat{X}^n) = \frac{1}{(2N+1)^d}\, \hat{\mathbb{E}}(\mathrm{Tr}(\hat{X}^n)) = \frac{1}{(2N+1)^d}\, \hat{\mathbb{E}}\left(\sum_{i_1 \in \mathcal{T}(N,d)} \cdots \sum_{i_n \in \mathcal{T}(N,d)} \hat{X}_{i_1,i_2} \hat{X}_{i_2,i_3} \cdots \hat{X}_{i_n,i_1} \right). \tag{9}$$

The product $\hat{X}_{i_1,i_2} \cdots \hat{X}_{i_n,i_1}$ on the right-hand side of this equation yields something different from 0 if and only if

$$|i_1 - i_2| = |i_2 - i_3| = \cdots = |i_{n-1} - i_n| = |i_n - i_1| = 1;$$

$(i_1,i_2,i_3,\ldots,i_n,i_1)$ is a closed walk on $\mathcal{T}(N,d)$ that starts from i_1 and returns to i_1 and hence describes a loop in $\mathcal{G}_1$. If the walk crosses some point in $\mathcal{T}(N,d)$ more than once, then there is more than one walk yielding the same loop. Therefore the sum over walks in (9) cannot be reduced to a sum over loops easily. However, we shall see that in the limit $d \to \infty$ only a narrow class of walks survive.

VI. PAIR PARTITIONS, WALKS, AND LOOPS

In this section we shall describe the connection between pair partitions, as described in the Introduction, closed walks on $\mathcal{T}(N,d)$, and loops in $\mathcal{G}_1$.

Suppose we have some $\Pi \in P_2(S)$ consisting of n pairs $(\alpha_1,\beta_1),\ldots,(\alpha_n,\beta_n)$ for $S=\{1,\ldots,2n\}$. To every pair (α_1,β_1), $i\in\{1,\ldots,n\}$, we assign the unit vector in the ith direction, e_i. Now for every element $s\in S$, starting with 1, we make a step in the lattice in the direction assigned to the pair in Π to which this number belongs. If s is the first element of such a pair, then the step will be taken in the positive direction. If s is the second element of such a pair, then the step will be taken in the negative direction. If we choose the origin as a starting point for our walks, then we always get a loop on $\{0,1\}^n$, the corners of the unit cube in n dimensions.

For fixed Π, define a mapping $\sigma_\Pi : S\to\{1,\ldots,n\}$ that indicates to what pair $s\in S$ belongs: $\sigma_\Pi(s)=i$ if $s\in(\alpha_i,\beta_i)$. Furthermore, define a mapping $\vartheta_\Pi : S\to\{-1,+1\}$ that indicates whether it is the first or the second element of this pair: $\vartheta_\Pi(s)=1$ if $s=\alpha_i$ and $\vartheta_\Pi(s)=-1$ if $s=\beta_i$. To keep track of our walk we define a mapping $v:S\to\mathcal{X}_0$ as follows:

$$v(s) := \sum_{j=1}^{s} \vartheta_\Pi(j) e_{\sigma_\Pi(j)}, \quad \text{with } v(0) := 0.$$

To fix the starting point and direction of every edge in the walk, we need the following mapping:

$$w: S\to\mathcal{X}_0 : i \mapsto \min(v(i-1),v(i)),$$

where $\min(v(i-1),v(i))$ is the component-wise minimum of $v(i-1)$ and $v(i)$. Now, let $\gamma_1:\mathcal{P}_2(S)\to\mathcal{G}_1$ associate to Π the loop in $\mathcal{G}_1$ traced out by the walk $(0,v(1),v(2),\ldots,v(2n-1),0)$:

$$\gamma_1(\Pi) := \sum_{j=1}^{2n} \vartheta_\Pi(j) 1_{(w(j),\{\sigma_\Pi(j)\})}.$$

Then one easily checks that $\gamma_1(\Pi)$ is indeed a loop, i.e., $\partial_1\gamma_1(\Pi)=0$.

Apart from γ_1 we shall also need an injective mapping $\gamma_2:\mathcal{P}_2(S)\to\mathcal{G}_2$ that maps a partition to a surface spanned by $\gamma_1(\Pi)$. To this end let us regard the pair partition in the circle representation as a planar graph with a closed outer edge. For every planar graph we can construct a dual by regarding every sector inside the outer edge of the graph as a vertex in the dual graph. The vertices in the dual graph are then connected by edges when the corresponding sectors have an edge in common.

Definition 15: The dual graph of a pair partition is the dual of the planar graph generated by its circle representation.

Figure 5 shows the construction of the dual graph of the partition $\Pi=\{(1,5),(2,6),(3,7),(4,8)\}$. To construct γ_2, consider a planar graph, H, given by a circle representation of a pair partition $\Pi\in\mathcal{P}_2(S)$. The disk sectors in this graph represent elements of $\mathcal{X}_0$, as is clear from the construction of the circle representation. Namely, we assign to sectors in

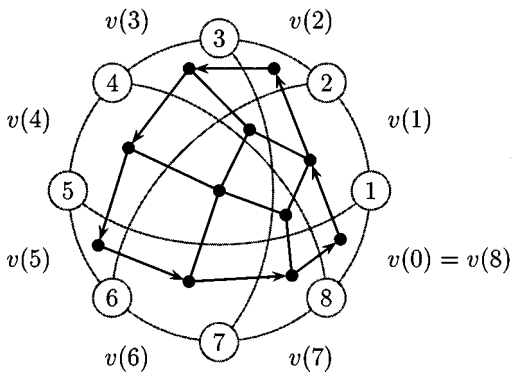

FIG. 5. The dual graph of the pair partition $\Pi=\{(1,5),(2,6),(3,7),(4,8)\}$. The orientation of the walk that traces out $\gamma_1(\Pi)$ is indicated by arrows.

H a vertex in space as follows: denote by A_0 the sector in H that has an edge in common with the circle segment between 1 and $2n$. Then to every sector A in H we associate a vertex $y=(y(1),\ldots,y(n))$ according to the rule.

$$y(i)=\begin{cases}0, & \text{if } A \text{ and } A_0 \text{ lie on the same side of the connecting line with index } i,\\ 1, & \text{otherwise.}\end{cases}$$

As a consequence to A_0 is associated $0\in\{0,1\}^n$. This procedure is illustrated in Fig. 6, which shows a picture of the partition $\Pi=\{(1,5),(2,4),(3,6)\}$ with the corner points in every sector as an element of $\{0,1\}^3$. Since two sectors in the graph that have an edge in common can be associated to points in $\mathcal{X}_0$ that differ only one step in the direction corresponding to the shared edge, we can connect two such points by an edge in the dual graph, H', of Π, corresponding to an edge in $\mathcal{X}_1$. This way the sectors in H' point at plaquettes in $\mathcal{X}_2$, since every crossing of connecting lines in H has four sectors in H that share that point. Since any two lines cross at most once, the number of plaquettes enclosed by H' is equal to $\#c(\Pi)$, the number of crossings of Π. In fact, H' contains a plaquette, based upon the minimum of its four corner points, u_{ij}, for every pair $(i,j)\in c(\Pi)$. We can now define

$$\gamma_2(\Pi):=\sum_{(i,j)\in c(\Pi)} 1_{(u_{ij},\{i,j\})}.$$

Note that every plaquette in $\gamma_2(\Pi)$ receives positive orientation in this definition.

Theorem 16: *For every* $\Pi\in\mathcal{P}_2(2n)$ *we have* $\partial_2\gamma_2(\Pi)=\gamma_1(\Pi)$.

The content of this theorem becomes apparent if one realizes that the outer edge of the dual graph of the pair partition Π represents the walk $(v(0),v(1),\ldots,v(2n))$, as shown in Fig. 5. It is immediately clear that noncrossing partitions $\Pi\in\mathcal{P}_2(2n)$ yield $\gamma_2(\Pi)=0$, so we have $\partial_2\gamma_2(\Pi)=0$. This agrees with $\gamma_1(\Pi)=0$: every step that is taken is retraced later. This, however, will be a special case of the proof we shall give.

Proof: Let H be the planar graph given by the circle representation of Π. Fix an edge $(v,\{i\})$ in the dual graph, H', of Π. This edge crosses a segment, κ, of the ith connecting line in H. There are now three possibilities.

(1) The segment κ is the ith connecting line itself, i.e., the ith connecting line crosses no other connecting lines in H.

(2) The segment κ connects the edge of the circle to a crossing with the jth connecting line in H.

(3) The segment k connects 2 crossings, say with the j_1th and the j_2th connecting line.

In the first case the connecting line in the circle representation corresponding to e_i splits the circle into two separate parts connected by one edge, ζ, in H'. The ith connecting line therefore crosses no other connecting lines. This means that there exist no $k\in\{1,\ldots,n\}$ such that $(i,k)\in c(\Pi)$; therefore $\gamma_2(\Pi)(v,\{i\})=0$. Since the walk that traces out $\gamma_1(\Pi)$ is closed, it has to visit ζ twice; once in the positive direction and once in the negative direction, so $\gamma_1(\Pi)(\zeta)$ also vanishes.

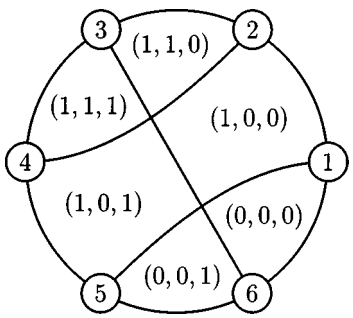

FIG. 6. The corner points of the surface generated by the partition $\Pi=\{(1,5),(2,4),(3,6)\}$.

In the second case the edge $(v,\{i\})$ is in the boundary of the plaquette $(u_{ij},\{i,j\})$, so

$$(\partial_2\gamma_2(\Pi))(v,\{i\})=\epsilon(i,\{j\})(1_{(u_{ij}+e_j,\{i\})}(v,\{i\})-1_{(u_{ij},\{i\})}(v,\{i\})). \tag{10}$$

First suppose $i<j$; then the crossing pairs (α_i,β_i) and (α_j,β_j) in Π satisfy $\alpha_i<\alpha_j<\beta_i<\beta_j$. Now if $v=u_{ij}$, meaning that v and u_{ij} lie on the same side of the ith and jth connecting line as 0, then the walk that traces out $\gamma_1(\Pi)$ reaches v before crossing the jth connecting line. Therefore the step this walk takes from v onward corresponds to α_i, so $\gamma_1(\Pi)(v,\{i\})=1$. From (10) we see that also $(\partial_2\gamma_2(\Pi))(v,\{i\})=1$. If, on the other hand, $v=u_{ij}+e_j$, then v and u_{ij} lie on the same side of the ith connecting line, but not on the same side of the jth connecting line. This means that the walk that traces out $\gamma_1(\Pi)$ has to cross the jth connecting line before it can reach v. But this means that a step in the ith direction has also been taken. Therefore we know that the step in the ith direction corresponds to β_i, so $\gamma_1(\Pi)(v,\{i\})=-1$. From (10) we see that also $(\partial_2\gamma_2(\Pi))(v,\{i\})=-1$. Now suppose $i>j$, then the crossing pairs (α_i,β_i) and (α_j,β_j) in Π satisfy $\alpha_j<\alpha_i<\beta_j<\beta_i$. The same type of argument as used for the case $i<j$ yields

$$v=u_{ij}\Rightarrow(\partial_2\gamma_2(\Pi))(v,\{i\})=\gamma_1(\Pi)(v,\{i\})=-1,$$

$$v=u_{ij}+e_j\Rightarrow(\partial_2\gamma_2(\Pi))(v,\{i\})=\gamma_1(\Pi)(v,\{i\})=1.$$

In the third case $(v,\{i\})$ is in the boundary of two plaquettes; $p_1=1_{(u_{ij_1},\{i,j_1\})}$ and $p_2=1_{(u_{ij_2},\{i,j_2\})}$. Since we are free to choose $j_1<j_2$, we can distinguish three cases:

(a) $i<j_1<j_2$,
(b) $j_1<i_1<j_2$, and
(c) $j_1<j_2<i$.

To prove case (a) we note that in this case we have $u_{ij_2}=u_{ij_1}+e_{j_1}$ and $v=u_{ij_2}$. Now calculate

$$(\partial_2 p_1)(v,\{i\})=\epsilon(j_1,\{i\})(1_{(u_{ij_1}+e_{j_1},\{i\})}(v,\{i\})-1_{(u_{ij_1},\{i\})}(v,\{i\}))=-1,$$

$$(\partial_2 p_2)(v,\{i\})=\epsilon(j_2,\{i\})(1_{(u_{ij_2}+e_{j_2},\{i\})}(v,\{i\})-1_{(u_{ij_2},\{i\})}(v,\{i\}))=1.$$

Therefore we have that $(\partial_2(p_1+p_2))(v,\{i\})=(\partial_2\gamma_2(\Pi))(v,\{i\})=0$. Since $(v,\{i\})$ is not an outer edge of H', we also have that $\gamma_1(\Pi)(v,\{i\})=0$. The reader can now easily verify cases (b) and (c), since the method of proof for these cases is the same as for case (a).

Finally, there are the edges in $\mathcal{X}_1$ that are not part of the dual graph of Π. We have $\gamma_2(\Pi)(u_1)=0$ for every $u_1\in\mathcal{X}_2$ not corresponding to a plaquette in H'. For every such u_1, we have $\partial_2(\gamma_2(\Pi)(u_1))=0$. Furthermore, we have $\gamma_1(\Pi)(u_2)=0$ for every u_2 not in the walk $\{(w(1),\{\sigma_\Pi(1)\}),\ldots,(w(2n),\{\sigma_\Pi(2n)\})\}$, so on the edges that are not in the dual graph of Π we have that $\partial_2\gamma_2(\Pi)=\gamma_1(\Pi)=0$, since every $u_2\in\{(w(1),\{\sigma_\Pi(1)\}),\ldots,(w(2n),\{\sigma_\Pi(2n)\})\}$ is represented in the dual graph of Π. □

Theorem 17: For every $\Pi\in\mathcal{P}_2(2n)$ the surface $\gamma_2(\Pi)$ is minimal in $\Gamma_{\gamma_1(\Pi)}$.

Proof: In case $n=1$ the loop $\gamma_1(\Pi)$ cannot enclose a plaquette. It follows from the definition of $\gamma_2(\Pi)$ that in this case $\gamma_2(\Pi)=0$, so the theorem holds for $n=1$. In the following we assume $n\geqslant 2$ and $i<j$ for $i,j\in\{1,\ldots,n\}$.

On the unit vectors $\{e_1,\ldots,e_n\}$ we introduce the following projection operator:

$$Q_{ij}e_k:=1_{\{i,j\}}(k)e_k.$$

Then the element $\gamma_2(\Pi)$ will be minimal in $\Gamma_{\gamma_1(\Pi)}$ if for every pair $i,j\in\{1,\ldots,n\}$ the projection of the closed walk that traces out $\gamma_1(\Pi)$ is a closed walk around $(0,\{i,j\})\in\mathcal{X}_2$, provided that $(i,j)\in c(\Pi)$. We now introduce

$$\Upsilon_{ij}:=\bigcup_{s=1}^{2n}\{Q_{ij}v(s)\},$$

the set of elements in $\mathcal{X}_0$ the projection to the $\{i,j\}$ plane of the walk that traces out $\gamma_1(\Pi)$ visits, and note that the proof is finished if we show $(i,j)\in c(\Pi)\Rightarrow \#\Upsilon_{ij}=4$.

Suppose $(i,j)\in c(\Pi)$, then (α_i,β_i) and (α_j,β_j) are crossing pairs in Π, so $\alpha_i<\alpha_j<\beta_i<\beta_j$. This means that $\Upsilon_{ij}=\{0,e_i,e_i+e_j,e_j\}$, so $\#\Upsilon_{ij}=4$. □

Corollary 18: For all $\Pi\in\mathcal{P}_2(2n)$ we have that $A(\gamma_2(\Pi))=A(\gamma_1(\Pi))=\#c(\Pi)$.

VII. THE PARISI MODEL AND THE *q*-HARMONIC OSCILLATOR

To show the connection between the *q*-harmonic oscillator and the Parisi model, we prove that, for $d\to\infty$ and $N\to\infty$, the moments of $\hat{X}$ converge to the moments of $X=a+a^*$. In short, $\hat{X}$ converges in distribution to X.

Theorem 19: *Let φ_q be the vacuum state on the q-harmonic oscillator algebra generated by $X=a+a^*$. Then for all $n\in\mathbb{N}$ the following holds:*

$$\varphi_q(X^n)=\lim_{d\to\infty}\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n).$$

Corollary 7 states how the moments of the *q*-harmonic oscillator can be calculated as a sum over pair partitions. Therefore, here it suffices to show that

$$\lim_{d\to\infty}\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n)=\sum_{\Pi\in\mathcal{P}_2(n)}q^{\#c(\Pi)}.$$

Proof: From Eq. (9) we know that $\hat{\varphi}(\hat{X}^n)$ can be interpreted as a sum over closed walks in $\mathcal{T}(N,d)$. The theorem then is trivial for n odd since an odd number has no pair partitions and a closed walk cannot return to its starting point in an odd number of steps. So we may assume $n=2r$, $r\in\mathbb{N}$, in the following.

We define a sublattice $\mathcal{T}'(N,d)$ of $\mathcal{T}(N,d)$ as follows:

$$\mathcal{T}'(N,d):=\{-N+(r+1),\ldots,N-(r+1)\}^d.$$

The set of all walks that start from some $x\in\mathcal{T}'(N,d)$ and have length n will be defined as

$$\mathcal{W}(x,n):=\{(x,j_1,\ldots,j_{n-1},x)|j_1,\ldots,j_{n-1}\in\mathcal{T}(N,d),\quad |x-j_1|=|j_1-j_2|=\cdots=|j_{n-1}-x|=1\}.$$

Note that we have

$$\#\mathcal{W}(x,n)=2^r d(d-1)\cdots(d-r+1).$$

To every walk $w\in\mathcal{W}(x,n)$ we can assign some surface $s\in\Gamma_w$, the set of minimal surfaces s that have the closed walk w as a boundary. We can then choose a set $\mathcal{S}(x,n)$ that contains exactly one surface $s\in\Gamma_w$ for every $w\in\mathcal{W}(x,n)$, i.e., $\#\mathcal{S}(x,n)=\#\mathcal{W}(x,n)$. Every surface $s\in\mathcal{S}(x,n)$ has to be minimal with respect to its corresponding walk $w\in\mathcal{W}(x,n)$ in order for the surface to have the corner points of the plaquettes in this surface in $\mathcal{T}(N,d)$.

With the use of these definitions we can rewrite the sum over i_1 in Eq. (9) as follows:

$$\hat{\varphi}(\hat{X}^n)=\frac{1}{(2N+1)^d}\sum_{i_1\in\mathcal{T}'(N,d)}\hat{\mathbb{E}}\left(\sum_{i_2\in\mathcal{T}(N,d)}\cdots\sum_{i_n\in\mathcal{T}(N,d)}\hat{X}_{i_1,i_2}\hat{X}_{i_2,i_3}\cdots\hat{X}_{i_n,i_1}\right)$$
$$+\frac{1}{(2N+1)^d}\sum_{\substack{i_1\in\mathcal{T}(N,d)\\ i_1\notin\mathcal{T}'(N,d)}}\hat{\mathbb{E}}\left(\sum_{i_2\in\mathcal{T}(N,d)}\cdots\sum_{i_n\in\mathcal{T}(N,d)}\hat{X}_{i_1,i_2}\hat{X}_{i_2,i_3}\cdots\hat{X}_{i_n,i_1}\right),\qquad(11)$$

and calculate

$$\hat{\mathbb{E}}(\hat{X}_{i_1,i_2}\cdots\hat{X}_{i_n,i_1}) = \frac{1}{(2d)^{n/2}} \sum_{w\in\mathcal{W}(x,n)} \hat{\mathbb{E}}(W(w)) = \frac{1}{(2d)^{n/2}} \sum_{w\in\mathcal{W}(x,n)} \hat{\mathbb{E}}(e^{i\Sigma_{p\in\mathcal{X}_1} w(p)A(p)})$$
$$= \frac{1}{(2d)^{n/2}} \sum_{w\in\mathcal{W}(x,n)} \hat{\mathbb{E}}(e^{i\int_w A}) = \frac{1}{(2d)^{n/2}} \sum_{s\in\mathcal{S}(x,n)} \hat{\mathbb{E}}(e^{i\int_s B}),$$

where we used Stokes' theorem. A walk that starts in $\mathcal{T}'(N,d)$ can never reach the boundary of $\mathcal{T}(N,d)$, so we can identify every walk $w\in\mathcal{W}(x,n)$ with a walk $w'\in\mathcal{W}(y,n)$, for $x,y\in\mathcal{T}'(N,d)$, via

$$w=(x,j_2,\ldots,j_n,x)\mapsto(y,j_2-(x-y),\ldots,j_n-(x-y),y)=w'.$$

From the fact that B is constant, it follows that the magnetic flux through $s\in\mathcal{S}(x,n)$, corresponding to w, is equal to the magnetic flux through $s'\in\mathcal{S}(y,n)$, corresponding to w', for $x,y\in\mathcal{T}'(N,d)$. This implies that we have

$$\sum_{s\in\mathcal{S}(x,n)} \hat{\mathbb{E}}(e^{i\int_s B}) = \sum_{s\in\mathcal{S}(y,n)} \hat{\mathbb{E}}(e^{i\int_s B}).$$

We can therefore perform the first sum over i_1 in (11), yielding

$$\frac{(2(N-(r+1))+1)^d}{(2N+1)^d}\,\hat{\mathbb{E}}\left(\sum_{i_2\in\mathcal{T}(N,d)}\cdots\sum_{i_n\in\mathcal{T}(N,d)}\hat{X}_{0,i_2}\hat{X}_{i_2,i_3}\cdots\hat{X}_{i_n,0}\right).$$

Since for every point in $\mathcal{T}(N,d)\backslash\mathcal{T}'(N,d)$ we have, at most, as many loops as for points in $\mathcal{T}'(N,d)$ we can estimate an upper bound for the second sum over i_1 in (11):

$$\frac{(2(N-(r+1))+1)^d-(2N+1)^d}{(2N+1)^d}\,\#\mathcal{W}(0,n).$$

This term tends to 0 for $N\to\infty$, so we find

$$\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n)=\hat{\mathbb{E}}\left(\sum_{i_2\in\mathbb{Z}^d}\cdots\sum_{i_n\in\mathbb{Z}^d}\hat{X}_{0,i_2}\hat{X}_{i_2,i_3}\cdots\hat{X}_{i_n,0}\right).$$

We now turn to the limit $d\to\infty$. There are exactly $\#\mathcal{P}_2(n)\binom{d}{r}2^r$ walks that start in 0, have length n, and go through r different spatial directions. The number of closed walks that go through less than r, say r', spatial directions is less than $(r')^n\binom{d}{r'}$, so the contribution of these walks vanishes in the limit $d\to\infty$ due to the standardization of $\hat{X}$. The closed walks that go through r spatial directions take a step in each direction exactly twice, and therefore correspond to a pair partition $\Pi\in\mathcal{P}_2(2r)$. The expectation value of the magnetic flux through the area of the loop $\gamma_1(\Pi)$ such a walk traces out, is given by

$$\hat{\mathbb{E}}(W(\gamma_2(\Pi)))=\hat{\mathbb{E}}\left(\prod_{x\in\mathcal{X}_2'}e^{i\gamma_2(\Pi)(x)B(x)}\right)=\prod_{x\in\mathcal{X}_2'}\hat{\mathbb{E}}(e^{i\gamma_2(\Pi)(x)B(x)})=(\cos\mathcal{B})^{A(\gamma_2(\Pi))}=q^{A(\gamma_2(\Pi))},$$

where $\mathcal{X}_2':=\{(v,\gamma)\in\mathbb{Z}^d\times\mathfrak{P}_2(\{1,\ldots,d\})\}$ with the same conventions we have for $\mathcal{X}_2$. In the above calculation the expectation of the product over $\mathcal{X}_2$ is interpreted as a product of expectations. To justify this we note that $\gamma_2(\Pi)$ has, at most, one plaquette for every pair of spatial directions and that the sign of the magnetic field B for a pair of spatial directions is independent of that for every other pair of directions. This implies that the fluxes through different plaquettes in $\gamma_2(\Pi)$ are independent. It is now possible to write $\hat{\varphi}(\hat{X}^{2r})$ as a sum over partitions in the limit $d,N\to\infty$:

$$\lim_{d\to\infty}\lim_{N\to\infty}\hat{\varphi}(\hat{X}^{2r})=\lim_{d\to\infty}\frac{1}{(2d)^r}\sum_{i_2\in\mathbb{Z}^d}\cdots\sum_{i_{2r}\in\mathbb{Z}^d}\hat{\mathbb{E}}((\Delta_q)_{0,i_2}\cdots(\Delta_q)_{i_{2r},0})$$

$$=\sum_{\Pi\in\mathcal{P}_2(2r)}q^{A(\gamma_1(\Pi))}=\sum_{\Pi\in\mathcal{P}_2(2r)}q^{\#c(\Pi)},$$

where we used Corollary 18. □

VIII. NONCOMMUTATIVE INDEPENDENCE IN THE PARISI MODEL

Until now we concerned ourselves with only one q-Gaussian random variable. In order to illustrate the concept of q independence in this model, we shall have to define more, say $\mathcal{N}\in\mathbb{N}$, q-Gaussian random variables. The Parisi model allows for such an extension in a straightforward way.

If d is divisible by $\mathcal{N}$, the lattice $\mathcal{T}(N,d)$ can be decomposed as a product of $\mathcal{N}$ sublattices as follows:

$$\mathcal{T}(N,d)=\mathcal{M}_1\times\mathcal{M}_2\times\cdots\times\mathcal{M}_{\mathcal{N}},$$

where $\dim\mathcal{M}_j=d/\mathcal{N}$ for $j\in\{1,...,\mathcal{N}\}$. We can now use the q-deformed lattice Laplacian Δ_q defined in Sec. V to define $\mathcal{N}$ q-Gaussian random variables $\hat{X}_1,\ldots,\hat{X}_{\mathcal{N}}\in\mathcal{A}$ as follows:

$$(\hat{X}_i)_{u,w}=\begin{cases}\sqrt{\dfrac{\mathcal{N}}{2d}}\,(\Delta_q)_{v,w}, & \text{if } v-w\in\mathcal{M}_i,\\ 0, & \text{otherwise.}\end{cases}$$

In this way $\hat{X}_i, i\in\{1,...,\mathcal{N}\}$, is the standardized deformed lattice Laplacian on M_i. From the previous section we can deduce that the operators $\hat{X}_1,\ldots,\hat{X}_{\mathcal{N}}$ defined in this way converge, in distribution, to the q-Gaussian random variables $X_i=a_i+a_i^*$ for $d\to\infty$ and $N\to\infty$. Now fix $\mathcal{N}=2$ and let $\hat{X}=\hat{X}_1$ and $\hat{Y}=\hat{X}_2$.

Theorem 20 (independence): *For $n,m\in\mathbb{N}$ we have that*

$$\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n\hat{Y}^m)=\lim_{N\to\infty}\hat{\varphi}(\hat{Y}^m\hat{X}^n)=\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n)\hat{\varphi}(\hat{Y}^m).$$

Proof: Write out $\hat{\varphi}(\hat{X}^n\hat{Y}^m)$ to find

$$\hat{\varphi}(\hat{X}^n\hat{Y}^m)=\frac{1}{(2N+1)^d}\,\hat{\mathbb{E}}\left(\sum_{\substack{i_1,\ldots,i_n\\ \in\mathcal{T}(N,d)}}\sum_{\substack{j_1,\ldots,j_m\\ \in\mathcal{T}(N,d)}}\hat{X}_{i_1,i_2}\cdots\hat{X}_{i_n,j_1}\hat{Y}_{j_1,j_2}\cdots\hat{Y}_{j_m,i_1}\right).$$

Now, obviously the nonvanishing terms all satisfy

$$(i_1-i_2)+\cdots+(i_n-j_1)+(j_1-j_2)+\cdots+(j_m-i_1)=0.$$

The first n differences all lie in $\mathcal{M}_1$ and the last m differences all lie in $\mathcal{M}_2$ therefore,

$$(i_1-i_2)+\cdots+(i_n-j_1)=(j_1-j_2)+\cdots+(j_m-i_1)=0,$$

and it follows that $i_1=j_1$. With reference to the proof of Theorem 19, we can choose $i_1=j_1=0$, so we find

$$\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n\hat{Y}^m)=\hat{\mathbb{E}}\left(\sum_{\substack{i_2,\ldots,i_n\\ \in\mathbb{Z}^d}}\sum_{\substack{j_2,\ldots,j_m\\ \in\mathbb{Z}^d}}\hat{X}_{0,i_2}\cdots\hat{X}_{i_n,0}\hat{Y}_{0,j_2}\cdots\hat{Y}_{j_m,0}\right)$$

$$=\hat{\mathbb{E}}\left(\sum_{\substack{i_2,\ldots,i_n\\ \in\mathbb{Z}^d}}\hat{X}_{0,i_2}\cdots\hat{X}_{i_n,0}\right)\hat{\mathbb{E}}\left(\sum_{\substack{j_2,\ldots,j_m\\ \in\mathbb{Z}^d}}\hat{Y}_{0,j_2}\cdots\hat{Y}_{j_m,0}\right)=\lim_{N\to\infty}\hat{\varphi}(\hat{X}^n)\hat{\varphi}(\hat{Y}^m).$$

Since $\hat{\varphi}$ is a trace state, we have that $\hat{\varphi}(\hat{X}^n\hat{Y}^m)=\hat{\varphi}(\hat{Y}^m\hat{X}^n)$. □
The same type of proof shows that, for general $\mathcal{N}$, we have

$$\hat{\varphi}(\hat{X}_{k(1)}^{n_1}\cdots\hat{X}_{k(\mathcal{N})}^{n_\mathcal{N}})=\hat{\varphi}(\hat{X}_{k(1)}^{n_1})\cdots\hat{\varphi}(\hat{X}_{k(\mathcal{N})}^{n_\mathcal{N}}),$$

provided the values $k(1),\ldots,k(\mathcal{N})$ are all different.

We conclude that in the sense of Definition 2, the $\hat{X}_i$ tend to independent random variables in the limit $N\to\infty$, and to independent q-Gaussians in the limit $N\to\infty$, $d\to\infty$.

[1] G. Parisi, ''d-dimensional arrays of Josephson junctions, spin glasses and q-deformed harmonic oscillators,'' J. Phys. A **27**, 7555–7568 (1994).
[2] B. Kümmerer, ''Survey on a theory of non-commutative stationary Markov processes,'' *Quantum Probability and Applications III*, Lecture Notes in Mathematics, edited by L. Accardi and W. von Waldenfels (Springer-Verlag, Berlin, 1987), Vol. 1303, pp. 154–182.
[3] B. Kümmerer and H. Maassen, ''Elements of quantum probability,'' to appear in *Quantum Probability Communications X*, edited by R. Hadron and M. Lindsay (World Scientific, Singapore, 1998), in press.
[4] H. Maassen, ''Theoretical concepts in quantum probability; quantum Markov processes,'' *Fractals, Quasicrystals, Chaos, Knots and Algebraic Quantum Mechanics*, Nato ASI Series C: Mathematical and Physical Sciences, edited by A. Amann, L. Cederbaum, and W. Gans (Kluwer Academic, Dordrecht, 1987), Vol. 235, pp. 287–302.
[5] P. A. Meyer, ''Quantum probability for probabilists,'' *Lecture Notes in Mathematics* (Springer-Verlag, Berlin, 1993), Vol. 1538.
[6] K. R. Parthasarathy, *An Introduction to Quantum Stochastic Calculus*, Monographs in Mathematics (Birkhäuser, Basel, 1992), Vol. 85.
[7] M. Bożejko and R. Speicher, ''An example of a generalized Brownian motion,'' Commun. Math. Phys. **137**, 519–531 (1991).
[8] M. Bożejko and R. Speicher, ''An example of a generalized Brownian motion II,'' *Quantum probability and related topics VII*, 1992, pp. 67–77.
[9] H. van Leeuwen and H. Maassen, ''An obstruction for q-deformation of the convolution product,'' J. Phys. A **29**, 4741–4748 (1996).
[10] A. Connes, *Non-Commutative Geometry* (Academic, New York, 1990).

q-phase-coherent states and their squeezing properties

Yaping Yang[a)]
CCAST (World Laboratory), Beijing, People's Republic of China
and Department of Physics, Tongji University,
Shanghai 200092, People's Republic of China

Zhixin Lin and Shuangyuan Xie
Department of Physics, Tongji University, Shanghai 200092, People's Republic of China

Weiguo Feng and Xiang Wu
CCAST (World Laboratory), Beijing, People's Republic of China
and Department of Physics, Tongji University,
Shanghai 200092, People's Republic of China

(Received 27 February 1998; accepted for publication 17 August 1998)

Some kinds of *q*-phase-coherent states of a *q*-harmonic osscillator in a finite-dimensional Hilbert space are constructed. Some properties of these states are discussed. Second-order squeezing properties of these states with respect to the phase quadrature operators are studied. The number-phase squeezing and number-phase uncertainty relations are also studied in detail for a two-state system. Some new number-phase minimum uncertainty states are found. © *1998 American Institute of Physics.* [S0022-2488(98)01412-1]

I. INTRODUCTION

The concept of amplitude and phase of an oscillator plays an important part in mechanics and optics and was introduced into quantum mechanics and electrodynamics at an early stage.[1,2] In quantum theory, the amplitude of the electromagnetic field is proportional to the square root of the photon-number operator, and the phase is described by the exponential phase operators[3–6] which were established by Susskind and Glogower (SG). Unfortunately, the SG phase operators are not unitary, hence the Hermitian phase operators cannot be constructed.

In the late 1980s, Pegg and Barnett (PB)[7–9] suggested a formalism in which all the operators are defined in a finite but arbitrarily large state space. The Hermitian phase operator is then defined as the projection operator on the particular phase state multiplied by the corresponding value of the phase. All the mean values of observables depend parametrically on the dimension of the Hilbert space. When the infinite limit of the dimension is taken, the physical results correspond to the results obtainable in ordinary quantum mechanics. The PB phase-operator formalism has been applied in a variety of problems in quantum mechanics.[10,11] In particular, the coherent states (CSs) of the harmonic osillator (HO) in the finite-dimensional number-state or phase-state space were defined, and their squeezing properties of these CSs were analyzed in detail.[12–14]

In the past few years, *q*-deformed HO[15–17] was introduced in research of the representations of quantum algebra. Futhermore, a polar decomposition phase operator h in context of *q*-deformed oscillator was introduced first by Chaichian and Ellinas,[18] and *q*-deformed SG phase operator and *q*-coherent states were also constructed,[19–21] and the the properties of macro-deformed CSs were discussed.[22,23] But the *q*-deformed SG phase operator is also not unitary, therefore, Nelson and his co-workers have adopted the PB formalism to study the quantum phase problem in the deformed case.[19,22] *q*-Deformed operators in the finite-dimensional Hilbert space were defined,[24,25] and another kind of *q*-coherent states of the *q*-HO was established.[25]

The aim of this paper is to construct *q*-phase-coherent states (*q*-PCSs) based on the *q*-deformed operators of the phase quanta defined in Ref. 25. We shall also pay attention to phase fluctuation and amplitude and phase squeezing properties of *q*-PCSs in a finite-dimensional basis. We will also determine whether *q*-PCSs are minimum uncertainty states (MUSs)[4,5] or intelligent

[a)]Mailing address; Electronic mail: ypyangk@online.sh.cn

 © 1998 American Institute of Physics

states[26,27] with respect to the number and phase operators. In Sec. II we present a short review on the q-deformed operators of the phase quanta and discuss their algebraic properties. In Sec. III we define the q-PCSs in the finite-dimensional space and study some properties of these states. In Sec. IV we analyze second-order squeezing properties with respect to the q-PCSs in an arbitrary finite-dimensional phase-state space. Certainly, we also recognize that there are many cases of interest where the state space is genuinely finite. In Sec. V we investigate in detail number-phase squeezing and number-phase uncertainty relations for the simplest in a two-state system.

II. *q*-DEFORMED OPERATORS OF PHASE QUANTA

Let us consider a finite but arbitrarily large state space Ψ of $(s+1)$ dimensions of a q-deformed HO. The number states $|n\rangle \in \Psi$ are orthonormal:

$$\langle n|m\rangle = \delta_{nm}, \quad \sum_{n=0}^{s} |n\rangle\langle n| = 1. \tag{1}$$

The q-annihilation, creation and number operators of q-harmonic oscillator in the finite-dimensional Hilbert space may be defined through the following projection operators:

$$\hat{a}^{+} = \sum_{n=1}^{s} \sqrt{[n]}|n\rangle\langle n-1|, \quad \hat{a} = \sum_{n=1}^{s} \sqrt{[n]}|n-1\rangle\langle n|, \quad \hat{N} = \sum_{n=1}^{s} |n\rangle\langle n|, \tag{2}$$

where $[x]=(q^{x}-q^{-x})/(q-q^{-1})$, and x is a number or an operator. If s goes to infinity, those operators $\hat{a}^{+}$, $\hat{a}$, $\hat{N}$ will be ordinary q-deformed creation, annihilation, and number operator in infinite-dimensional space:

$$[\hat{N},\hat{a}^{+}] = \hat{a}^{+}, \quad [\hat{N},\hat{a}] = -\hat{a}, \quad [\hat{N}] = \hat{a}^{+}\hat{a},$$

$$\hat{a}\hat{a}^{+} - q\hat{a}^{+}\hat{a} = q^{-\hat{N}} - [s+1]|s\rangle\langle s|, \quad \hat{a}\hat{a}^{+} - q^{-1}\hat{a}^{+}\hat{a} = q^{\hat{N}} - [s+1]|s\rangle\langle s|. \tag{3}$$

Let us define a unitary exponential phase operator from the following equation:

$$\hat{a} = e^{i\Phi_\theta} R(\hat{N}). \tag{4}$$

$R(\hat{N})$ represents the amplitude part of $\hat{a}$. Using the similar method in Refs. 18, 24, and 25, we can obtain

$$\hat{a} = e^{i\Phi_\theta}\sqrt{[\hat{N}]}, \quad e^{i\Phi_\theta} = \sum_{n=1}^{s} |n-1\rangle\langle n| + e^{i\psi}|s\rangle\langle 0|. \tag{5}$$

Here ψ is an arbitrary real number. The eigenstate of the unitary exponential phase operator $e^{i\Phi_\theta}$ with eigenvalue $e^{i\theta_m}$ is

$$|\theta_m\rangle = (s+1)^{-1/2} \sum_{n=0}^{s} e^{i\theta_m n}|n\rangle. \tag{6}$$

The phase quanta θ_m is equal to $\theta_0 + (2m\pi)/(s+1)$, and $\theta_0 = \psi/(s+1)$. The phase states $|\theta_m\rangle (m=0,1,...,s)$ are orthonormal

$$\langle \theta_m|\theta_n\rangle = \delta_{m,n}, \quad \sum_{m=0}^{s} |\theta_m\rangle\langle\theta_m| = 1. \tag{7}$$

And the Hermitian phase operator can be expressed as

$$\Phi_\theta=\sum_{m=0}^{s}\theta_m|\theta_m\rangle\langle\theta_m|=\theta_0+\frac{s\pi}{s+1}+\frac{2\pi}{s+1}\sum_{m\neq n}^{s}\frac{\exp(-i(m-n)\theta_0)}{\exp(-i2\pi(m-n)/(s+1))-1}|n\rangle\langle m|, \quad (8)$$

and

$$e^{\pm i\Phi_\theta}=\sum_{m=0}^{s}e^{\pm i\theta_m}|\theta_m\rangle\langle\theta_m|,$$

$$\cos\Phi_\theta=\sum_{m=0}^{s}\cos\theta_m|\theta_m\rangle\langle\theta_m|, \quad \sin\Phi_\theta=\sum_{m=0}^{s}\sin\theta_m|\theta_m\rangle\langle\theta_m|. \quad (9)$$

The q-deformed creation and annihilation operators of phase quanta can be also defined as[25]

$$\hat{\phi}^+=\sqrt{[\Phi_\theta]}e^{(i2\pi\dot{N})/(s+1)}, \quad \hat{\phi}=e^{-(i2\pi\dot{N})/(s+1)}\sqrt{[\Phi_\theta]}. \quad (10)$$

In phase-state space, they can be rewritten as

$$\begin{aligned}\hat{\phi}^+&=\sum_{m=1}^{s}\sqrt{[\theta_m]}|\theta_m\rangle\langle\theta_{m-1}|+\sqrt{[\theta_0]}|\theta_0\rangle\langle\theta_s|,\\ \hat{\phi}&=\sum_{m=1}^{s}\sqrt{[\theta_m]}|\theta_{m-1}\rangle\langle\theta_m|+\sqrt{[\theta_0]}|\theta_s\rangle\langle\theta_0|.\end{aligned} \quad (11)$$

For simplicity, we take $\theta_0=0$ in the following discussion. From the above formulas we obtain

$$\begin{aligned}\hat{\phi}\hat{\phi}^\dagger-q^{2\pi/(s+1)}\hat{\phi}^\dagger\hat{\phi}&=\left[\frac{2\pi}{s+1}\right]q^{-\hat{\Phi}_\theta}-[2\pi]\hat{\mathcal{M}}_{ss},\\ \hat{\phi}\hat{\phi}^\dagger-q^{-2\pi/(s+1)}\hat{\phi}^\dagger\hat{\phi}&=\left[\frac{2\pi}{s+1}\right]q^{\hat{\Phi}_\theta}-[2\pi]\hat{\mathcal{M}}_{ss},\\ [\hat{\Phi}_\theta,\hat{\phi}^{\dagger k}]&=\frac{2\pi}{s+1}k\hat{\phi}^k, \quad 1\leqslant k\leqslant s,\\ [\hat{\Phi}_\theta,\hat{\phi}^k]&=-\frac{2\pi}{s+1}k\hat{\phi}^{\dagger k}, \quad 1\leqslant k\leqslant s,\end{aligned} \quad (12)$$

$$\begin{aligned}[\hat{\phi}^\dagger,\hat{\mathcal{M}}_{ij}]&=\sqrt{[\theta_{i+1}]}\mathcal{H}(s-i)\hat{\mathcal{M}}_{(i+1)j}-\sqrt{[\theta_j]}\hat{\mathcal{M}}_{i(j-1)},\\ [\hat{\phi},\hat{\mathcal{M}}_{ij}]&=\sqrt{[\theta_i]}\hat{\mathcal{M}}_{(i-1)j}-\sqrt{[\theta_{j+1}]}\mathcal{H}(s-j)\hat{\mathcal{M}}_{i(j+1)},\\ [\hat{\Phi}_\theta,\hat{\mathcal{M}}_{ij}]&=(\theta_i-\theta_j)\hat{\mathcal{M}}_{ij}, \quad [\hat{\mathcal{M}}_{ij},\hat{\mathcal{M}}_{kl}]=\delta_{jk}\hat{\mathcal{M}}_{il}-\delta_{il}\hat{\mathcal{M}}_{kj},\end{aligned}$$

where $\hat{\mathcal{M}}_{ij}$ is a projection operator in the phase-state space,[14] $\hat{\mathcal{M}}_{ij}=|\theta_i\rangle\langle\theta_j|$, $(i,j=0,1,...,s)$. The $\mathcal{H}(x)$ is the step function defined as $\mathcal{H}(x)\{^{1\ x>0}_{0\ x\leqslant 0}$. In this section, we can see that, in the phase-state basis, the creation and annihilation operators of the phase quanta are deformed and the number operator is not deformed. When $q\to 1$, these equations can be simplified as the corresponding equations of the normal HO.[14]

III. *q*-PHASE-COHERENT STATES OF *q*-HARMONIC OSCILLATOR

We now give a definition of the q-phase coherent states (q-PCSs) of the finite-dimensional q-HO as follows:

$$|\alpha,s\rangle = C(\alpha,s) e_{q^{2\pi/(s+1)}}^{\alpha \hat{\phi}^\dagger} |0\rangle_p = C(\alpha,s) \sum_{m=0}^{s} \frac{\tilde{\alpha}^m}{\sqrt{[\theta_m]!}} |\theta_m\rangle, \tag{13}$$

where we have used Eq. (11), and introduced the following symbols:

$$[\theta_0]! = 1, \quad [\theta_m]! = \prod_{i=1}^{m} [\theta_i], \quad m = 1,2,...,s,$$

and $e_{q^{2\pi/(s+1)}}^{x}$ is a q-exponential function

$$e_{q^{2\pi/(s+1)}}^{x} = \sum_{m=0}^{\infty} \frac{x^m}{[m]_{q^{2\pi/(s+1)}}!},$$

$$[m]_{q^{2\pi/(s+1)}} = \frac{q^{(2\pi m)/(s+1)} - q^{-(2\pi m)/(s+1)}}{q^{2\pi/(s+1)} - q^{-2\pi/(s+1)}} = \frac{[\theta_m]}{[2\pi/s+1]},$$

and $\tilde{\alpha} = [2\pi/(s+1)]\,\alpha$, α is an arbitrary complex number, and the normalization constant $C(\alpha,s)$ is given by

$$C(\alpha,s) = \left(\sum_{m=0}^{s} \frac{(\tilde{\alpha}\tilde{\alpha}^*)^m}{[\theta_m]!} \right)^{-1/2} = (e_{q,p,s}^{|\tilde{\alpha}|^2})^{-1/2}, \tag{14}$$

where $\tilde{\alpha}^*$ denotes the complex conjugation of $\tilde{\alpha}$ and $e_{q,p,s}^{x}$ is defined as

$$e_{q,p,s}^{x} = \sum_{m=0}^{s} \frac{x^m}{[\theta_m]!}.$$

Obviously, when $s \to \infty$, $e_{q,p,s}^{x} \to e_{q,p}^{x} = \Sigma_{m=0}^{\infty} x^m/[\theta_m]!$, and the q-PCSs become the q-PCSs of the q-HO in an infinite-dimensional Hilbert space.[18,19] When $q \to 1$, $e_{q,p,s}^{x} \to \Sigma_{m=0}^{s}(x^m)/(\theta_m!)$, and the q-PCSs become the PCSs of normal HO in a finite-dimensional Hilbert space.[14]

The q-annihilation operator of the phase quanta $\hat{\phi}$ acts on the q-PCSs $|\alpha,s\rangle$ in the following form:

$$\hat{\phi}|\alpha,s\rangle = \tilde{\alpha}\left\{ |\alpha,s\rangle - (e_{q,p,s}^{|\tilde{\alpha}|^2})^{-1/2} \frac{\tilde{\alpha}^s}{\sqrt{[\theta_s]!}} |\theta_s\rangle \right\}, \tag{15}$$

which means that the q-PCSs are not eigenstates of the q-annihilation operators of the phase quanta for any finite value of s. The second term on the right-hand side of the above formula originates from the finiteness of the dimension. When s goes to infinity, the second term vanishes, and $|\alpha,s\rangle$ is changed to an eigenstate of annihilation operator $\hat{\phi}$.

The q-PCSs are nonorthogonal to each other

$$\langle \alpha,s|\alpha',s\rangle = (e_{q,p,s}^{|\tilde{\alpha}|^2} e_{q,p,s}^{|\tilde{\alpha}'|^2})^{-1/2} e_{q,p,s}^{\tilde{\alpha}^*\tilde{\alpha}'}, \tag{16}$$

but the q-PCSs are complete

$$\int |\alpha,s\rangle\langle\alpha,s| d\mu(\alpha) = \sum_{m=0}^{s} |\theta_m\rangle\langle\theta_m| = 1, \tag{17}$$

where

$$d\mu(\alpha) = \frac{1}{2\pi} e_{q,p}^{-|\tilde{\alpha}|^2} e_{q,p,s}^{|\tilde{\alpha}|^2} d_{q^{2\pi/(s+1)}} |\tilde{\alpha}|^2 d\theta.$$

Note that the integral over θ is a normal integration, but the integral over $|\tilde{\alpha}|^2$ is a q-integration. The q-derivative is defined as

$$\frac{d}{d_{q^{2\pi/(s+1)}}} f(x) = \frac{f(q^{2\pi/(s+1)}x) - f(q^{-2\pi/(s+1)}x)}{(q^{2\pi/(s+1)} - q^{-2\pi/(s+1)})x}.$$

The q-integration can be derived from the definition of the q-derivative.

By the two above properties, the q-PCSs are overcomplete. An arbitrary q-PCS $|\alpha',s\rangle$ can be expressed as

$$|\alpha',s\rangle = \int (e^{|\tilde{\alpha}|^2}_{q,p,s} e^{|\tilde{\alpha}'|^2}_{q,p,s})^{-1/2} e^{\tilde{\alpha}^*\tilde{\alpha}'}_{q,p,s} |\alpha,s\rangle d\mu(\alpha). \tag{18}$$

The phase-state representation of the q-PCSs $|\alpha,s\rangle$, given in Eq. (13), implies that the phase-probability distribution is

$$P(\theta_m,s) = |\langle \theta_m|\alpha,s\rangle|^2 = \frac{|\tilde{\alpha}|^{2m}}{[\theta_m]!} (e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}, \tag{19}$$

with the normalization condition

$$\sum_{m=0}^{s} P(\theta_m,s) = 1. \tag{20}$$

On account of the completeness relation of the q-PCSs, we can expand the phase states $|\theta_m\rangle$ in terms of the q-PCSs$|\alpha,s\rangle$ as

$$|\theta_m\rangle = \frac{1}{2\pi} \int e^{-|\tilde{\alpha}|^2}_{q,p} (e^{|\tilde{\alpha}|^2}_{q,p,s})^{1/2} \frac{(\tilde{\alpha}^*)^m}{\sqrt{[\theta_m]!}} |\alpha,s\rangle d_{q^{2\pi/(s+1)}} |\tilde{\alpha}|^2 d\theta. \tag{21}$$

After substituting Eq. (6) into Eq. (13), we get

$$|\alpha,s\rangle = (s+1)^{-1/2} (e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1/2} \sum_{n,m=0}^{s} \frac{\tilde{\alpha}^m}{\sqrt{[\theta_m]!}} \exp(in\theta_m)|n\rangle. \tag{22}$$

So the probability distribution of the q-PCSs in the finite-dimensional number-state space is

$$P(n,s) = (s+1)^{-1} (e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1} \left| \sum_{m=0}^{s} \frac{\tilde{\alpha}^m}{\sqrt{[\theta_m]!}} \exp(in\theta_m) \right|^2, \tag{23}$$

with the normalization condition

$$\sum_{n=0}^{s} P(n,s) = 1. \tag{24}$$

Similarly, the number states $|n\rangle$ can also be expanded in terms of the q-PCSs $|\alpha,s\rangle$

$$|n\rangle = \frac{1}{2\pi(s+1)^{1/2}} \sum_{m=0}^{s} \int e^{-|\tilde{\alpha}|^2}_{q,p} (e^{|\tilde{\alpha}|^2}_{q,p,s})^{1/2} \exp(-in\theta_m) \frac{(\tilde{\alpha}^*)^m}{\sqrt{[\theta_m]!}} |\alpha,s\rangle d_{q^{2\pi/(s+1)}} |\tilde{\alpha}|^2 d\theta. \tag{25}$$

IV. SECOND-ORDER SQUEEZING PROPERTIES

In this section, we investigate the second-order squeezing properties of the q-PCSs mentioned in Sec. III. First, we introduce a pair of Hermitian phase quadrature operators $\hat{X}_\phi$ and $\hat{Y}^{12}_\phi$ corresponding to the q-creation and annihilation operators of the phase quanta under consideration as

$$\hat{X}_\phi=\tfrac{1}{2}(\hat{\phi}+\hat{\phi}^\dagger),\quad \hat{Y}_\phi=\frac{1}{2i}(\hat{\phi}-\hat{\phi}^\dagger), \tag{26}$$

which satisfy the following commutation relation:

$$[\hat{X}_\phi,\hat{Y}_\phi]=\frac{i}{2}[\hat{\phi},\hat{\phi}^\dagger], \tag{27}$$

and the uncertainty relation

$$\langle(\Delta\hat{X}_\phi)^2\rangle\cdot\langle(\Delta\hat{Y}_\phi)^2\rangle\geqslant\tfrac{1}{4}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|^2, \tag{28}$$

where the variances of the operators are defined as $\langle(\Delta\hat{X}_\phi)^2\rangle=\langle\hat{X}_\phi^2\rangle-(\langle\hat{X}_\phi\rangle)^2$, and $\langle(\Delta\hat{Y}_\phi)^2\rangle=\langle\hat{Y}_\phi^2\rangle-(\langle\hat{Y}_\phi\rangle)^2$. We shall say that the variances of the operators $\hat{X}_\phi$ and $\hat{Y}_\phi$ are squeezed if

$$\langle(\Delta\hat{X}_\phi)^2\rangle<\tfrac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|\ \text{ or }\ \langle(\Delta\hat{Y}_\phi)^2\rangle<\tfrac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|.$$

To measure the degree of squeezing with respect to the pair of operators, we introduce two parameters S_X and S_Y

$$S_X=\frac{\langle(\Delta\hat{X}_\phi)^2\rangle-\frac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|}{\frac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|},\quad S_Y=\frac{\langle(\Delta\hat{Y}_\phi)^2\rangle-\frac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|}{\frac{1}{2}|\langle[\hat{X}_\phi,\hat{Y}_\phi]\rangle|}. \tag{29}$$

The squeezing condition now takes the simple form

$$S_X<0\ \text{ or }\ S_Y<0. \tag{30}$$

For the q-PCSs defined in Eq. (13), we have some expectation values:

$$\begin{aligned}
&\langle\alpha,s|\hat{\phi}|\alpha,s\rangle=\tilde{\alpha}(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-1},\quad \langle\alpha,s|\hat{\phi}^\dagger|\alpha,s\rangle=\tilde{\alpha}^*(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-1},\\
&\langle\alpha,s|\hat{\phi}^2|\alpha,s\rangle=\tilde{\alpha}^2(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-2},\quad \langle\alpha,s|\hat{\phi}^{\dagger2}|\alpha,s\rangle=\tilde{\alpha}^{*2}(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-2},\\
&\langle\alpha,s|\hat{\phi}^\dagger\hat{\phi}|\alpha,s\rangle=|\tilde{\alpha}|^2(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-1},\\
&\langle\alpha,s|\hat{\phi}\hat{\phi}^\dagger|\alpha,s\rangle=(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}\left(q^{2\pi/(s+1)}|\tilde{\alpha}|^2e^{|\tilde{\alpha}|^2}_{q,p,s-2}+\left[\frac{2\pi}{s+1}\right]e^{|\tilde{\alpha}|^2q^{-2\pi/(s+1)}}_{q,p,s-1}\right).
\end{aligned} \tag{31}$$

So we obtain the variances of the operators $\hat{X}_\phi$ and $\hat{Y}_\phi$

$$\begin{aligned}
\langle\alpha,s|(\Delta\hat{X}_\phi)^2|\alpha,s\rangle=\frac{1}{4}(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}\Bigg\{&(\tilde{\alpha}^2+\tilde{\alpha}^{*2}+q^{2\pi/(s+1)}|\tilde{\alpha}|^2)e^{|\tilde{\alpha}|^2}_{q,p,s-2}+\left[\frac{2\pi}{s+1}\right]e^{|\tilde{\alpha}|^2q^{-2\pi/(s+1)}}_{q,p,s-1}\\
&+|\tilde{\alpha}|^2e^{|\tilde{\alpha}|^2}_{q,p,s-1}\Bigg\}-\{(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-1}\ \mathrm{Re}(\tilde{\alpha})\}^2,
\end{aligned} \tag{32}$$

$$\begin{aligned}
\langle\alpha,s|(\Delta\hat{Y}_\phi)^2|\alpha,s\rangle=-\frac{1}{4}(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}\Bigg\{&(\tilde{\alpha}^2+\tilde{\alpha}^{*2}-q^{2\pi/(s+1)}|\tilde{\alpha}|^2)e^{|\tilde{\alpha}|^2}_{q,p,s-2}-\left[\frac{2\pi}{s+1}\right]e^{|\tilde{\alpha}|^2q^{-2\pi/(s+1)}}_{q,p,s-1}\\
&-|\tilde{\alpha}|^2e^{|\tilde{\alpha}|^2}_{q,p,s-1}\Bigg\}-\{(e^{|\tilde{\alpha}|^2}_{q,p,s})^{-1}e^{|\tilde{\alpha}|^2}_{q,p,s-1}\ \mathrm{Im}(\tilde{\alpha})\}^2,
\end{aligned} \tag{33}$$

and the expectation value of their commutator $[\hat{X}_\phi,\hat{Y}_\phi]$,

$$\langle\alpha,s|[\hat{X}_\phi,\hat{Y}_\phi]|\alpha,s\rangle=\frac{i}{2}\,(e_{q,p,s}^{|\tilde{\alpha}|^2})^{-1}\left\{\left[\frac{2\pi}{s+1}\right]e_{q,p,s-1}^{|\tilde{\alpha}|^2q^{-2\pi/(s+1)}}+q^{2\pi/(s+1)}|\tilde{\alpha}|^2e_{q,p,s-2}^{|\tilde{\alpha}|^2}-|\tilde{\alpha}|^2e_{q,p,s-1}^{|\tilde{\alpha}|^2}\right\}. \tag{34}$$

Eqs. (32)–(34) show that the q-PCSs $|\alpha,s\rangle$ are not the MUSs. However, in the limit $s\to\infty$, we have

$$\langle\alpha,s|(\Delta\hat{X}_\phi)^2|\alpha,s\rangle=\langle\alpha,s|(\Delta\hat{Y}_\phi)^2|\alpha,s\rangle=\tfrac{1}{2}|\langle\alpha,s|[\hat{X}_\phi,\hat{Y}_\phi]|\alpha,s\rangle|=0 \tag{35}$$

and

$$S_X=S_Y=0. \tag{36}$$

The equality in the uncertainty relation (28) holds, i.e., the q-PCSs are the MUSs.

We investigate below the second-order squeezing for α real.

When $[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2\,e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}>0$, we obtain the two parameters:

$$\begin{aligned}S_X&=\frac{2\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}+2\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}-4\tilde{\alpha}^2(e_{q,p,s}^{\tilde{\alpha}^2})^{-1}(e_{q,p,s-1}^{\tilde{\alpha}^2})^2}{[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}},\\ S_Y&=\frac{2\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}-2\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}}{[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}}\geqslant 0,\end{aligned} \tag{37}$$

Thus, in this case, there does not exist any squeezing of the $\hat{Y}_\phi$ component for an arbitrary value that s can take. However, it can be deduced from the above formula that squeezing of the $\hat{X}_\phi$ component may be obtained if

$$\begin{aligned}&\left[\frac{2\pi}{s+1}\right]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}>0,\\ &2\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}+2\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}-4\tilde{\alpha}^2(e_{q,p,s}^{\tilde{\alpha}^2})^{-1}(e_{q,p,s-1}^{\tilde{\alpha}^2})^2<0.\end{aligned} \tag{38}$$

When $[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}<0$, the representations of the two parameters become

$$\begin{aligned}S_X&=\frac{2\tilde{\alpha}^2(1+q^{2\pi/(s+1)})e_{q,p,s-2}^{\tilde{\alpha}^2}+2[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-4\tilde{\alpha}^2(e_{q,p,s}^{\tilde{\alpha}^2})^{-1}(e_{q,p,s-1}^{\tilde{\alpha}^2})^2}{\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}-[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}},\\ S_Y&=\frac{2\tilde{\alpha}^2(q^{2\pi/(s+1)}-1)e_{q,p,s-2}^{\tilde{\alpha}^2}-2[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}}{\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}-[2\pi/(s+1)]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}}.\end{aligned} \tag{39}$$

Obviously, squeezing of the $\hat{X}_\phi$ component is obtained when

$$\begin{aligned}&\left[\frac{2\pi}{s+1}\right]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-\tilde{\alpha}^2e_{q,p,s-1}^{\tilde{\alpha}^2}+q^{2\pi/(s+1)}\tilde{\alpha}^2e_{q,p,s-2}^{\tilde{\alpha}^2}<0,\\ &2\tilde{\alpha}^2(1+q^{2\pi/(s+1)})e_{q,p,s-2}^{\tilde{\alpha}^2}+2\left[\frac{2\pi}{s+1}\right]e_{q,p,s-1}^{\tilde{\alpha}^2q^{-2\pi/(s+1)}}-4\tilde{\alpha}^2(e_{q,p,s}^{\tilde{\alpha}^2})^{-1}(e_{q,p,s-1}^{\tilde{\alpha}^2})^2<0.\end{aligned} \tag{40}$$

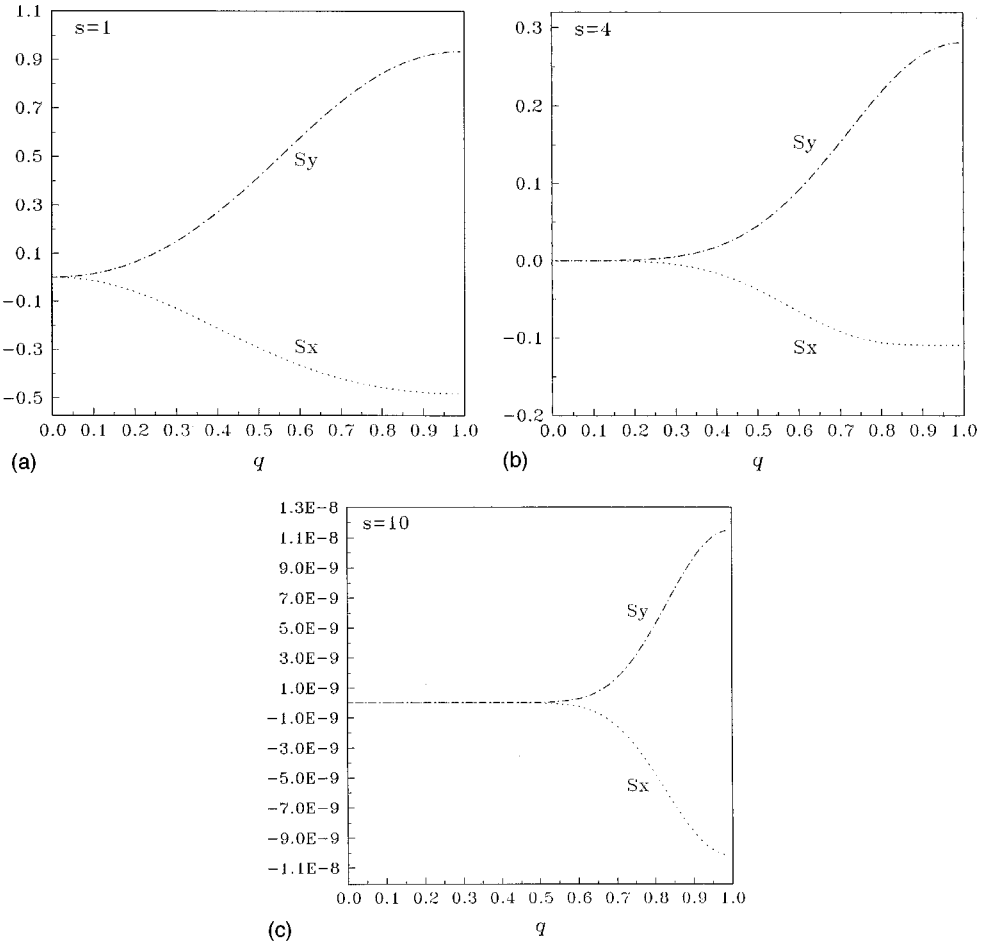

FIG. 1. Parameters S_X and S_Y [see Eqs. (37) and (39)] for the q-PCSs with $\alpha=1$ are plotted as a function of q.

There seems to exist a squeezing of the $\hat{Y}_\phi$ component in some term. However, after quantities of numerical analyses, one can understand that S_Y is always greater than zero, whatever q, s and α are (where $0<q<1$, α is a real number, s is a positive integer), so that there is not any squeezing of the $\hat{Y}_\phi$ component.

In Figs. 1 and 2 we plot the parameters S_X and S_Y for the q-PCSs $|\alpha,s\rangle$ with $\alpha=1,4$ in $(s+1)$-dimensional state space for various values of s as a function of q. From these figures it is clearly seen that S_Y is always greater that zero and in certain cases S_X is less than zero.

From the above analyses, we can conclude that when α is a real number, there does not exist any squeezing of the $\hat{Y}_\phi$ component, but squeezing of the $\hat{X}_\phi$ component can be obtained. A similar discussion shows that when α is a pure imaginary number, no $\hat{X}_\phi$ squeezing appears, while there may exist $\hat{Y}_\phi$ squeezing.

V. NUMBER-PHASE SQUEEZING AND ITS UNCERTAINTY RELATIONS FOR A TWO-STATE SYSTEM

It is instructive to see how the ideas advanced in this paper can be illustrated with the simplest finite-dimensional Hilbert space, namely that of a two-state system (i.e., the $s=1$ case), which is of some importance in quantum optics. We take $\theta_0=0$, then $\theta_1=\pi$. The relations between number-state basis and phase-state basis are

$$|0\rangle=\frac{1}{\sqrt{2}}\,(|0\rangle_p+|\theta_1\rangle),\quad |1\rangle=\frac{1}{\sqrt{2}}\,(|0\rangle_p-|\theta_1\rangle), \tag{41}$$

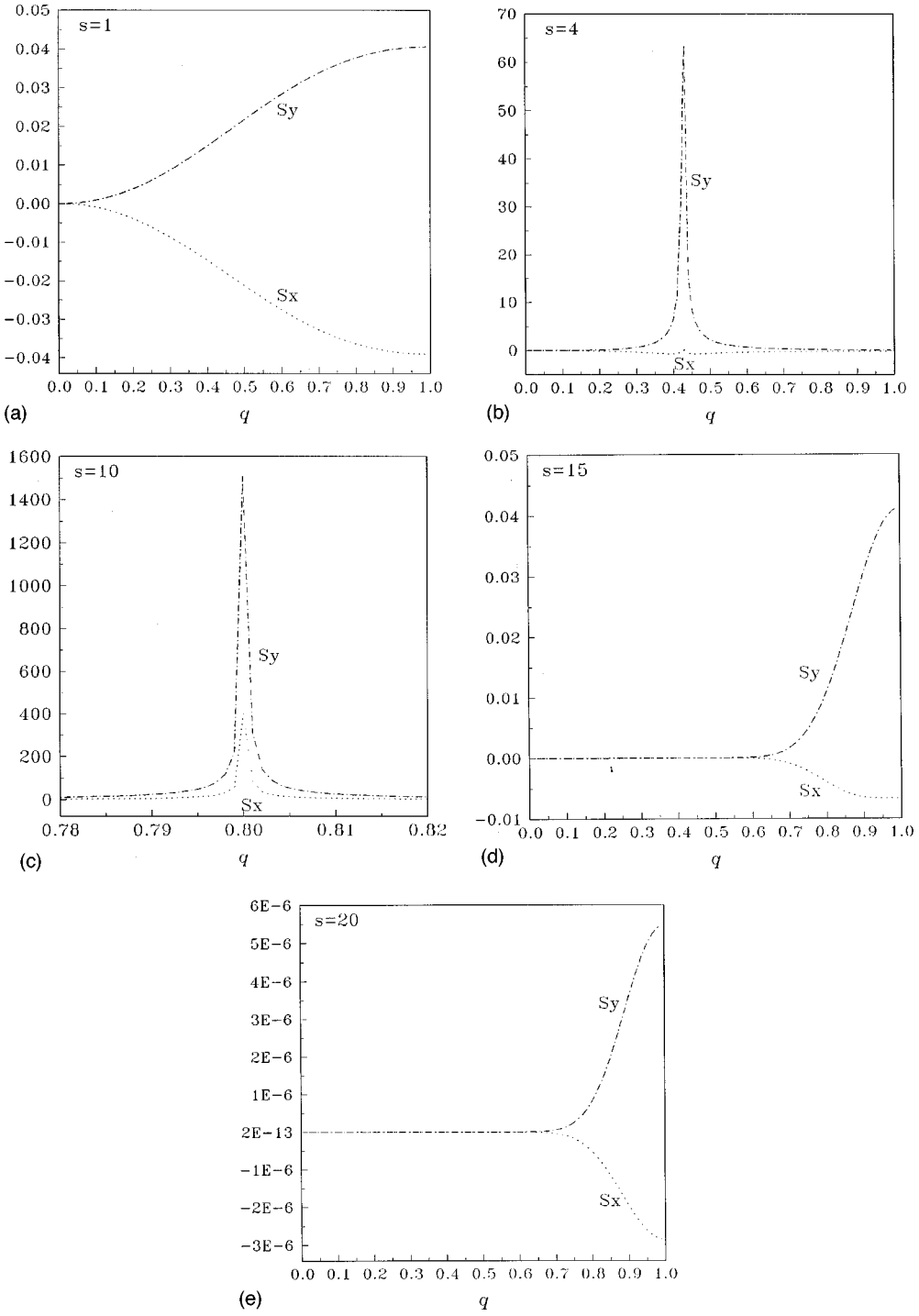

FIG. 2. Parameters S_X and S_Y [see Eqs. (37) and (39)] for the q-PCSs with $\alpha=4$ are plotted as a function of q.

$$|0\rangle_p=\frac{1}{\sqrt{2}}\,(|0\rangle+|1\rangle),\quad |\theta_1\rangle=\frac{1}{\sqrt{2}}\,(|0\rangle-|1\rangle).$$

The number operator and phase operator can be written in the two bases, respectively

$$\hat{N}=|1\rangle\langle 1|=\tfrac{1}{2}(|0\rangle_{pp}\langle 0|+|\theta_1\rangle\langle\theta_1|-|0\rangle_p\langle\theta_1|-|\theta_1\rangle_p\langle 0|),\tag{42}$$

$$\hat{\Phi}_\theta = \pi|\theta_1\rangle\langle\theta_1| = \frac{\pi}{2}(|0\rangle\langle 0| + |1\rangle\langle 1| - |0\rangle\langle 1| - |1\rangle\langle 0|).$$

Using Eq. (9), we get

$$\cos\hat{\Phi}_\theta = |0\rangle_{pp}\langle 0| - |\theta_1\rangle\langle\theta_1|. \tag{43}$$

Then we have the following commutation relations:

$$\begin{aligned}[\hat{N},\hat{\Phi}_\theta] &= \frac{\pi}{2}(|\theta_1\rangle_p\langle 0| - |0\rangle_p\langle\theta_1|),\\ [\hat{N},\cos\hat{\Phi}_\theta] &= |0\rangle_p\langle\theta_1| - |\theta_1\rangle_p\langle 0|.\end{aligned} \tag{44}$$

When $s=1$, the q-PCSs can be expressed as

$$\begin{aligned}|\alpha,1\rangle &= \left(1+\frac{|\tilde{\alpha}|^2}{[\pi]}\right)^{-1/2}\left(|0\rangle_p + \frac{\tilde{\alpha}}{\sqrt{[\pi]}}|\theta_1\rangle\right)\\ &= A|0\rangle_p + \exp(i\phi)B|\theta_1\rangle\\ &= \frac{1}{\sqrt{2}}(A+\exp(i\phi)B)|0\rangle + \frac{1}{\sqrt{2}}(A-\exp(i\phi)B)|1\rangle,\end{aligned} \tag{45}$$

where $\tilde{\alpha}=|\tilde{\alpha}|e^{i\phi}$, $A\equiv(1+|\tilde{\alpha}|^2/[\pi])^{1/2}$, $B\equiv|\tilde{\alpha}|/\sqrt{[\pi]}(1+|\tilde{\alpha}|^2/[\pi])^{-1/2}$, and they satisfy $A^2+B^2=1$.

A. $\hat{\Phi}_\theta - \hat{N}$ squeezing

The variances of the phase and number operator and the expectation value of their commutater in the q-PCSs (45) are

$$\begin{aligned}\langle(\Delta\hat{N})^2\rangle &= \tfrac{1}{4} - A^2B^2\cos^2\phi,\\ \langle(\Delta\hat{\Phi}_\theta)^2\rangle &= \pi^2A^2B^2,\\ \langle[\hat{N},\hat{\Phi}_\theta]\rangle &= -i\pi AB\sin\phi,\end{aligned} \tag{46}$$

so that the corresponding parameters are

$$\begin{aligned}S_\Phi &= \frac{2\pi AB - |\sin\phi|}{|\sin\phi|},\\ S_N &= \frac{1-4A^2B^2\cos^2\phi - 2\pi AB|\sin\phi|}{2\pi AB|\sin\phi|}.\end{aligned} \tag{47}$$

In what follows we will study the q-PCSs (45) in which the operators $\hat{N}$ and $\hat{\Phi}_\theta$ are subfluctuant, that is, in which squeezing of the variances of these operators can be observed.

1. Squeezing in $\hat{\Phi}_\theta$

Let us suppose $\theta_0-\phi=n\pi+\pi/2$ (n is a integer). In this case the q-PCSs (45) take the form

$$|\alpha,1\rangle = A|0\rangle_p \pm iB|\theta_1\rangle. \tag{48}$$

It is easy to check that these are the intelligent states with respect to the operators $\hat{\Phi}_\theta$ and $\hat{N}$ with the Heisenberg relation

$$\langle(\Delta\hat{N})^2\rangle\langle(\Delta\hat{\Phi}_\theta)^2\rangle=\frac{1}{4}\,|\langle[\hat{N},\hat{\Phi}_\theta]\rangle|^2=\frac{\pi^2}{4}\,A^2B^2. \tag{49}$$

The squeezing parameters describing the degree of squeezing of the variances of operators $\hat{\Phi}_\theta$ and $\hat{N}$ in the states (48) are

$$S_\Phi=2\pi AB-1,\quad S_N=\frac{1}{2\pi AB}-1. \tag{50}$$

From above it follows that when $2\pi AB<1$, the phase operator $\hat{\Phi}_\theta$ is subfluctuant (i.e., $S_\Phi<0$) and the number operator $\hat{N}$ is superfluctuant ($S_N>0$) in these states (48). Moreover, when $2\pi AB>1$, $\hat{\Phi}_\theta$ is superfluctuant and $\hat{N}$ is subfluctuant. In particular, if $2\pi AB\to 0$ (i.e., $\alpha\to 0$), $S_\Phi\to -1$ while $S_N\to +\infty$. Simultaneously the Heisenberg relation (49) reaches its minimum, which implies that the q-PCSs under consideration are MUSs. From Eq. (48) we find that the MUSs are

$$\lim_{\alpha\to 0}|\alpha,1\rangle=|0\rangle_p\,. \tag{51}$$

The phase operator $\hat{\Phi}_\theta$ is ultimately subfluctuant (100% squeezed) in these states.

2. Squeezing in $\hat{N}$

If $AB=1$ (i.e., $|\alpha|=1/\sqrt{[}\pi]$), the q-PCSs (45) take the form

$$|\alpha,1\rangle=\frac{1}{\sqrt{2}}\,(|0\rangle_p+\exp(i\phi)|\theta_1\rangle). \tag{52}$$

These are also the intelligent states with respect to the operators $\hat{\Phi}_\theta$ and $\hat{N}$ with the Heisenberg relation

$$\langle(\Delta\hat{N})^2\rangle\langle(\Delta\hat{\Phi}_\theta)^2\rangle=\frac{1}{4}\,|\langle[\hat{N},\hat{\Phi}_\theta]\rangle|^2=\frac{\pi^2}{16}\sin^2\,\phi. \tag{53}$$

In this case, the squeezing parameters take the form

$$S_\Phi=\frac{\pi-|\sin\,\phi|}{|\sin\,\phi|}>0,\quad S_N=\frac{|\sin\,\phi|-\pi}{\pi}<0, \tag{54}$$

which imply that the phase operator is superfluctuant (i.e., $S_\Phi>0$), while the number operator is subfluctuant (i.e., $S_N<0$). In particular, the q-PCSs under consideration become MUSs when $\phi=n\pi$. At this time, $S_N\to -1$, and $S_\Phi\to +\infty$. From Eq. (52) we can obtain the following MUSs:

$$|\alpha,1\rangle=\begin{cases}\dfrac{1}{\sqrt{2}}\,(|0\rangle_p+|\theta_1\rangle), & (\phi=0)\\[2ex] \dfrac{1}{\sqrt{2}}\,(|0\rangle_p-|\theta_1\rangle), & (\phi=\pi)\end{cases}. \tag{55}$$

B. cos $\hat{\Phi}_\theta$–$\hat{N}$ squeezing

The variance of the operator $\cos\hat{\Phi}_\theta$ and the expectation value of the commutator $[\cos\hat{\Phi}_\theta,\hat{N}]$ in the q-PCSs (45) are

$$\langle(\Delta\ \cos\hat{\Phi}_\theta)^2\rangle = 4A^2B^2,$$

$$\langle[\hat{N},\cos\hat{\Phi}_\theta]\rangle = 2iAB\ \sin\ \phi. \tag{56}$$

The corresponding squeezing parameters are

$$S_{\cos\Phi} = \frac{4AB - |\sin\ \phi|}{|\sin\ \phi|}, \quad S_N = \frac{1 - 4A^2B^2\ \cos^2\ \phi - 4AB|\sin\ \phi|}{4AB|\sin\ \phi|}. \tag{57}$$

If $\theta_0 - \phi = n\pi + \pi/2$ (n is a integer), the q-PCSs take the form of Eq. (48), and are also the intelligent states with respect to the operators $\cos\hat{\Phi}_\theta$ and $\hat{N}$ with the Heisenberg relation

$$\langle(\Delta\hat{N})^2\rangle\langle(\Delta\ \cos\hat{\Phi}_\theta)^2\rangle = \tfrac{1}{4}|\langle[\hat{N},\cos\hat{\Phi}_\theta]\rangle|^2 = A^2B^2. \tag{58}$$

In this case, the squeezing parameters become

$$S_{\cos\Phi} = 4AB - 1,$$

$$S_N = \frac{1}{4AB} - 1. \tag{59}$$

From above it can be seen that when $4AB < 1$, $\cos\hat{\Phi}_\theta$ is superfluctuant ($S_\Phi < 0$) while $\hat{N}$ is subfluctuant ($S_N > 0$), and when $4AB > 1$, $\hat{\Phi}_\theta$ is subfluctuant while $\hat{N}$ is superfluctuant. In particular, if $AB \to 0$ (i.e., $\alpha \to 0$), $S_{\cos\Phi} \to -1$ and $S_N \to +\infty$, which means that the operator $\cos\hat{\Phi}_\theta$ is subfluctuant (with 100% degree of squeezing) in these states, and that the q-PCSs under consideration are MUSs as in Eq. (51).

If we take $2AB = 1$, we can get the similar squeezing properties of $\hat{N}$ as in Sec. V A.

VI. CONCLUSIONS

In this paper we have constructed the q-PCSs of the q-HO in the finite-dimensional Hilbert space and discussed some properties of them, for instance, overcompleteness, phase (or number)-distribution, and uncertainty relation. Those results are of interest for working on the q-deformed oscillator problem, and they may be used to investigate the phase properties of a q-deformed single-mode electromagnetic field. We have also studied the second-order squeezing properties of the phase quadrature operators with respect to the q-PCSs. In particular, we have investigated in detail the number-phase squeezing properties and the number-phase uncertainty relations for a two-state system and found some new number-phase MUSs.

ACKNOWLEDGMENT

The authors would like to thank the referee for helpful suggestions.

[1] F. London, Z. Phys. **37**, 915 (1926).
[2] P. A. M. Dirac, Proc. R. Soc. London, Ser. A **114**, 243 (1927).
[3] L. Susskind and J. Glogower, Physics (Long Island City, NY) **1**, 49 (1964).
[4] P. Carruthers and M. M. Nieto, Phys. Rev. Lett. **14**, 387 (1965).
[5] P. Carruthers and M. M. Nieto, Mod. Phys. Rev. **40**, 411 (1968).
[6] Y. P. Yang, W. G. Feng, and X. Wu, Mod. Phys. Lett. A **12**, 2141 (1997).
[7] D. T. Pegg and S. M. Barnett, Europhys. Lett. **6**, 483 (1988).
[8] S. M. Barnett and D. T. Pegg, J. Mod. Opt. **36**, 7 (1989).
[9] D. T. Pegg and S. M. Barnett, Phys. Rev. A **39**, 1665 (1989).
[10] J. A. Vaccaro and D. T. Pegg, J. Mod. Opt. **37**, 17 (1990).
[11] S. M. Barnett and D. T. Pegg, Phys. Rev. A **42**, 6713 (1990).
[12] V. Bužek, A. D. Wilson-Gordon, P. L. Knight, and W. K. Lai, Phys. Rev. A **45**, 8079 (1992).
[13] L. M. Kuang, F. B. Wang, and Y. G. Zhou, J. Mod. Opt. **41**, 1307 (1994).
[14] L. M. Kuang and X. Chen, Phys. Lett. A **186**, 8 (1994).
[15] L. C. Biedenharn, J. Phys. A **22**, L873 (1989).
[16] A. J. Macfarlane, J. Phys. A **22**, 4581 (1989).

[17] C. P. Sun, J. F. Lu, and M. L. Ge, J. Phys. A **23**, L1199 (1990).
[18] M. Chaichian and D. Ellinas, J. Phys. A **23**, L291 (1990).
[19] S. H. Chiu, R. W. Gray, and C. A. Nelson, Phys. Lett. A **164**, 237 (1992).
[20] Z. R. Yu, Phys. Lett. A **175**, 391 (1993).
[21] Y. P. Yang, W. G. Feng, and X. Wu, J. Math. Phys. **38**, 6315 (1997).
[22] C. A. Nelson and M. H. Fields, Phys. Rev. A **51**, 2410 (1995).
[23] W. C. Kimler IV and C. A. Nelson, J. Phys. A **30**, 949 (1997).
[24] D. Ellinas, Phys. Rev. A **45**, 3358 (1992).
[25] Y. P. Yang and Z. R. Yu, Mod. Phys. Lett. A **10**, 347 (1995).
[26] C. Aragone, G. Guerri, S. Salamó, and J. L. Tani, J. Phys. A **7**, L149 (1974).
[27] C. Aragone, E. Ghalbaud, and S. Salamó, J. Math. Phys. **17**, 1963 (1976).

On separable Schrödinger–Maxwell equations

Renat Zhdanov[a] and Maxim Lutfullin
Institute of Mathematics, 3 Tereshchenkivska Street, 252004 Kiev, Ukraine

(Received 8 May 1998; accepted for publication 11 August 1998)

We obtain the most general time-dependent potential $V(t,x_1,x_2)$ enabling separating of variables in the (1+2)-dimensional Schrödinger equation. With the use of this result the four classes of separable Schrödinger–Maxwell equations are constructed. © *1998 American Institute of Physics.* [S0022-2488(98)01212-2]

I. INTRODUCTION

In the Refs. 1 and 2 we have suggested an efficient direct approach to a problem of variable separation in the (1+2)-dimensional Schrödinger equation (SE) having a time-independent potential

$$i\psi_t+\psi_{x_1x_1}+\psi_{x_2x_2}=V(x_1,x_2)\psi. \tag{1}$$

With the help of this approach we have classified potentials $V(x_1,x_2)$ providing separability of Eq. (1) and, furthermore, obtained the full list of coordinate systems such that the corresponding Schrödinger equations are separable. It occurs that this approach can be extended in order to take into consideration the Schrödinger equation with time-dependent potential, namely

$$i\psi_t+\psi_{x_1x_1}+\psi_{x_2x_2}=V(t,x_1,x_2)\psi. \tag{2}$$

This class of SEs has numerous physical applications, in particular, they can be applied to describe the multiphoton ionization problem in the one electron atom.[3] Furthermore, the Keldysh type models in two space dimensions can be reduced to the particular form (2).[4]

In the present paper we obtain the complete description of potentials $V(t,x_1,x_2)$ such that SE is solvable by the method of separation of variables. Furthermore, we give the forms of the most general coordinate systems providing separability of the corresponding Schrödinger equations. As an application of thus obtained results we consider a problem of solving the Schrödinger–Maxwell equations by the method of separation of variables. This yields four classes of the Schrödinger–Maxwell equations solvable by the method of separation of variables.

II. SEPARATION OF VARIABLES IN EQ. (2)

SE (2) is solvable in the sense of Ref. 2 via the separation Ansatz

$$\psi=Q(t,x_1,x_2)\varphi_0(t)\varphi_1(\omega_1(t,x_1,x_2))\varphi_2(\omega_2(t,x_1,x_2)), \tag{3}$$

if and only if the functions Q,ω_1,ω_2 satisfy the following system of nonlinear partial differential equations:

$$\omega_{1x_1}\omega_{2x_1}+\omega_{1x_2}\omega_{2x_2}=0, \tag{4}$$

$$B_{1a}(\omega_1)(\omega_{1x_1}^2+\omega_{1x_2}^2)+B_{2a}(\omega_2)(\omega_{2x_1}^2+\omega_{2x_2}^2)+R_a(t)=0, \tag{5}$$

$$2(\omega_{ax_1}Q_{x_1}+\omega_{ax_2}Q_{x_2})+Q(i\omega_{at}+\omega_{ax_1x_1}+\omega_{ax_2x_2})=0, \tag{6}$$

[a]Electronic mail: rzhdanov@apmat.freenet.kiev.ua

 © 1998 American Institute of Physics

$$(B_{01}(\omega_1)(\omega_{1x_1}^2+\omega_{1x_2}^2)+B_{02}(\omega_2)(\omega_{2x_1}^2+\omega_{2x_2}^2))Q$$
$$+iQ_t+Q_{x_1x_1}+Q_{x_2x_2}+R_0(t)Q-V(t,x_1,x_2)Q=0, \tag{7}$$

where B_{0a}, B_{1a}, B_{2a}, R_0, R_a are smooth functions of the given arguments and $a=1,2$.

The general solution of the system of Eqs. (4)–(6) has been constructed in Ref. 2. It splits into the four inequivalent classes given below

$$\text{I.}\quad \omega_1=A(t)x_1+W_1(t),\quad \omega_2=B(t)x_2+W_2(t),$$
$$Q(t,x_1,x_2)=\exp\left\{-\frac{i}{4}\left(\frac{A'}{A}x_1^2+\frac{B'}{B}x_2^2\right)-\frac{i}{2}\left(\frac{W_1'}{A}x_1+\frac{W_2'}{B}x_2\right)\right\};$$
$$\text{II.}\quad x_1=W(t)e^{\omega_1}\sin\omega_2+W_1(t),\quad x_2=W(t)e^{\omega_1}\cos\omega_2+W_2(t),$$
$$Q(t,x_1,x_2)=\exp\{iR(t,x_1,x_2)\};$$
$$\text{III.}\quad x_1=\tfrac{1}{2}W(t)(\omega_1^2-\omega_2^2)+W_1(t),\quad x_2=W(t)\omega_1\omega_2+W_2(t), \tag{8}$$
$$Q(t,x_1,x_2)=\exp\{iR(t,x_1,x_2)\};$$
$$\text{IV.}\quad x_1=W(t)\cosh\omega_1\cos\omega_2+W_1(t),\quad x_2=W(t)\sinh\omega_1\sin\omega_2+W_2(t),$$
$$Q(t,x_1,x_2)=\exp\{iR(t,x_1,x_2)\}.$$

Here A, B, W, W_1, W_2 are arbitrary smooth functions and the function $R=R(t,x_1,x_2)$ is given by the formula

$$R(t,x_1,x_2)=\frac{W'}{4W}((x_1-W_1)^2+(x_2-W_2)^2)+\frac{1}{2}(W_1'x_1+W_2'x_2).$$

Inserting the above formulas into Eq. (7) yields the four inequivalent classes of potentials $V(t,x_1,x_2)$ providing separability of SE (2)

$$\text{I.}\quad V(t,x_1,x_2)=F_0(t)+A^2F_1(\omega_1)+B^2F_2(\omega_2)+x_1^2\left[\frac{A''}{4A}-\frac{(A')^2}{2A^2}\right]$$
$$+x_2^2\left[\frac{B''}{4B}-\frac{(B')^2}{2B^2}\right]-x_1\left(\frac{A'W_1'}{A^2}-\frac{W_1''}{2A}\right)-x_2\left(\frac{B'W_2'}{B^2}-\frac{W_2''}{2B}\right), \tag{9}$$
$$\text{II.}\quad V(t,x_1,x_2)=F_0(t)+e^{-2\omega_1}W^{-2}[F_1(\omega_1)+F_2(\omega_2)]$$
$$-\frac{W''}{4W}(x_1^2+x_2^2)+x_1\left(\frac{W_1W''}{2W}-\frac{W_1''}{2}\right)+x_2\left(\frac{W_2W''}{2W}-\frac{W_2''}{2}\right), \tag{10}$$
$$\text{III.}\quad V(t,x_1,x_2)=F_0(t)+(\omega_1^2+\omega_2^2)^{-1}W^{-2}[F_1(\omega_1)+F_2(\omega_2)]$$
$$-\frac{W''}{4W}(x_1^2+x_2^2)+x_1\left(\frac{W_1W''}{2W}-\frac{W_1''}{2}\right)+x_2\left(\frac{W_2W''}{2W}-\frac{W_2''}{2}\right), \tag{11}$$
$$\text{IV.}\quad V(t,x_1,x_2)=F_0(t)+(\sinh^2\omega_1+\sin^2\omega_2)^{-1}W^{-2}[F_1(\omega_1)+F_2(\omega_2)]$$
$$-\frac{W''}{4W}(x_1^2+x_2^2)+x_1\left(\frac{W_1W''}{2W}-\frac{W_1''}{2}\right)+x_2\left(\frac{W_2W''}{2W}-\frac{W_2''}{2}\right), \tag{12}$$

where the functions ω_1,ω_2 are given by the corresponding formulas I–IV from Eq. (8) and F_0, F_1, F_2 are arbitrary functions.

Formulas (9)–(12) give the explicit forms of the most general potentials $V=V(t,x_1,x_2)$ such that the corresponding SE is solvable by the separation of variables. However, the potentials obtained can be substantially simplified if we take into account that the class of partial differential equations (2) is invariant with respect to the following coordinate transformations:

$$1.\ \tilde{t}=t, \quad \tilde{x}_1=x_1+a(t), \quad \tilde{x}_2=x_2+b(t), \tag{13}$$

$$\tilde{\psi}(\tilde{t},\tilde{x}_1,\tilde{x}_2)=\psi(t,x_1,x_2)\exp\left\{\frac{i}{2}\left[a'(t)x_1+b'(t)x_2\right]\right\}; \tag{14}$$

$$2.\ \tilde{t}=T(t), \quad \tilde{x}_1=c(t)x_1, \quad \tilde{x}_2=c(t)x_2, \tag{15}$$

$$\tilde{\psi}(\tilde{t},\tilde{x}_1,\tilde{x}_2)=\psi(t,x_1,x_2)\exp\left\{\frac{ic'(t)}{4c(t)}\left(x_1^2+x_2^2\right)\right\}; \tag{16}$$

$$3.\ \tilde{t}=t, \quad \tilde{x}_1=x_1, \quad \tilde{x}_2=x_2, \quad \tilde{\psi}(\tilde{t},\tilde{x}_1,\tilde{x}_2)=\psi(t,x_1,x_2)\exp\{id(t)\}.$$

Here $a(t)$, $b(t)$, $c(t)$, $d(t)$ are arbitrary smooth functions and the function $T(t)$ is determined by the equation $(dT/dt)=c^2(t)$. Using the above enumerated coordinate transformations we can choose $W_1=0$, $W_2=0$, $B=1$, $W=1$ and get the following expressions for V:

$$1.\ V(t,x_1,x_2)=A^2F_1(\omega_1)+F_2(\omega_2)+x_1^2\left[\frac{A''}{4A}-\frac{(A')^2}{2A^2}\right], \tag{17}$$

$$2.\ V(t,x_1,x_2)=e^{-2\omega_1}[F_1(\omega_1)+F_2(\omega_2)], \tag{18}$$

$$3.\ V(t,x_1,x_2)=(\omega_1^2+\omega_2^2)^{-1}[F_1(\omega_1)+F_2(\omega_2)], \tag{19}$$

$$4.\ V(t,x_1,x_2)=(\sinh^2\omega_1+\sin^2\omega_2)^{-1}[F_1(\omega_1)+F_2(\omega_2)], \tag{20}$$

where the functions F_1,F_2 are arbitrary, the functions ω_1,ω_2 are given by the corresponding relations below

$$1.\ x_1=\frac{1}{A}\omega_1, \quad x_2=\omega_2;$$

$$2.\ x_1=e^{\omega_1}\sin\omega_2, \quad x_2=e^{\omega_1}\cos\omega_2;$$

$$3.\ x_1=\tfrac{1}{2}(\omega_1^2-\omega_2^2), \quad x_2=\omega_1\omega_2;$$

$$4.\ x_1=\cosh\omega_1\cos\omega_2, \quad x_2=\sinh\omega_1\sin\omega_2,$$

and $A=A(t)$ is a solution of the nonlinear ordinary differential equation $(4A)^{-1}A''-(2A)^{-2}(A')^2+C_1A+C_2=0$ with arbitrary constant C_1,C_2.

Note that the Schrödinger equations with potentials (9)–(12) and (17)–(20), being equivalent from the point of view of the standard theory of partial differential equations, are inequivalent within the context of quantum mechanics. The reason for this situation is that the transformations (13) and (15) change the quantity $|\psi(t,x_1,x_2)|$ which is regarded in quantum mechanics as the probability density. This means, in particular, that from the point of view of quantum mechanics time dependence of the potential $V(t,x_1,x_2)$ is essential and cannot be neglected. Nevertheless, in a sequel we restrict our considerations to the simplest case when potentials are of the forms (17)–(20).

III. SEPARATION OF VARIABLES IN THE SCHRÖDINGER–MAXWELL SYSTEM

Now we apply the above results to separate variables in the Schrödinger–Maxwell system

$$(p_0 - eA_0)\psi - (\vec{p} - e\vec{A})^2\psi = 0,$$

$$\Box A_0 - \frac{\partial}{\partial t}(A_{0t} + \operatorname{div}\vec{A}) = 0,$$

$$\Box\vec{A} + \overrightarrow{\operatorname{grad}}(A_{0t} + \operatorname{div}\vec{A}) = \vec{0}.$$

Here $p_0 = i\partial/\partial t$, $p_a = i\partial/\partial x_a$, $a = 1,2,3$, $e = \text{const}$, $\Box = \partial^2/\partial t^2 - \Delta$ is the d'Alembert operator.

To this end we consider the expressions (17)–(20) as the Ansätze for the vector potential of a special form

$$A = (A_0(t,\vec{x}) \equiv e^{-1}V(t,x_1,x_2),\vec{0}). \tag{21}$$

If we insert the above expression into the vacuum Maxwell equations without currents then the three equations for the function V

$$V_{x_1x_1} + V_{x_2x_2} = 0, \quad V_{tx_a} = 0, \quad a = 1,2,$$

are obtained. Hence we get restrictions on the choice of the functions F_1, F_2 in the form of ordinary differential equations. After solving these we arrive at the following expressions for the functions $V(t,x_1,x_2)$:

1. $V = C_1(x_1^2 - x_2^2) + C_2x_1 + C_3x_2$, (22)

2. $V = C_1 \ln(x_1^2 + x_2^2) + (x_1^2 + x_2^2)^{-2}[C_2(x_1^2 - x_2^2) + C_3x_1x_2]$, (23)

3. $V = C_1x_1 + \dfrac{1}{\sqrt{x_1^2 + x_2^2}}\left[C_2\sqrt{\sqrt{x_1^2 + x_2^2} + x_1} + C_3\sqrt{\sqrt{x_1^2 + x_2^2} - x_1}\right]$, (24)

4. $V = \dfrac{f^2}{f^4 + x_2^2}\left[(C_2 + C_1 \operatorname{arcsinh} f)f\sqrt{1 + f^2} + \left(C_3 + C_1 \arcsin\dfrac{x_2}{f}\right)\dfrac{x_2\sqrt{f^2 - x_2^2}}{f^2}\right]$, (25)

where

$$f^2 = \tfrac{1}{2}[x_1^2 + x_2^2 - 1 + \sqrt{(x_1^2 + x_2^2)^2 - 2(x_1^2 - x_2^2) + 1}]$$

and C_1, C_2, C_3 are integration constants.

SE having the potential (22) separates in Cartesian coordinates into two harmonic oscillator one-dimensional stationary Schrödinger equations and, consequently, is exactly solvable. Given the potentials (23) and (25) with $C_1 = 0$, SE (2) separates into the Hill and Mathieu equations. Hence we conclude that its solutions are represented as products of the degenerated hypergeometric and Mathieu functions.[5] At last, SE having the potential (24) separates in parabolic coordinates into two quasi-exactly solvable one-dimensional stationary Schrödinger equations and, consequently, is quasi-exactly solvable. This means that the finite part of the spectrum of the operator $-\Delta + V(x_1,x_2)$ can be computed by purely algebraic methods (for more details about quasi-exactly solvable models of quantum mechanics, see Ref. 6).

IV. CONCLUDING REMARKS

The variable separation technique used in the present paper can be directly applied to the Schrödinger equations with electro-magnetic fields of more general configurations than Eq. (21). Namely it is possible to classify vector potentials of the electro-magnetic field of the form

$$A=(A_0(t,x_1,x_2),\quad A_1(t,x_1,x_2),\quad A_2(t,x_1,x_2),\quad 0).$$

such that the corresponding SEs are solvable by the method of separation of variables. Furthermore, it is also possible to modify this technique in order to take into consideration the Schrödinger equation in $(1+3)$-dimensions. Both problems are under investigation now and will be a topic of our future publications.

ACKNOWLEDGMENT

One of the authors (R.Zh.) is supported by the Ukrainian DFFD Foundation (project 1.4/356).

[1] R. Z. Zhdanov, I. V. Revenko, and W. I. Fushchych, J. Math. Phys. **36**, 5506 (1995).
[2] R. Z. Zhdanov, J. Math. Phys. **38**, 1197 (1997).
[3] J. H. Eberly, J. Javanainen, and K. Rzazewski, Phys. Rep. **204**, 331 (1991).
[4] L. V. Keldysh, Sov. Phys. JETP **20**, 1307 (1965).
[5] E. Kamke, *Differentialgleichungen Lösungmethoden und Lösungen* (Akademische Verlagsgesellschaft, Leipzig, 1976).
[6] A. G. Ushveridze, *Quasi-exactly solvable models in quantum mechanics* (IOP Publ., Bristol, 1993).

Snapback repellers as a cause of chaotic vibration of the wave equation with a van der Pol boundary condition and energy injection at the middle of the span

Dedicated to Professor John E. Lagnese on the occasion of his 60th birthday

Goong Chen[a)]
Department of Mathematics, Texas A&M University, College Station, Texas 77843

Sze-Bi Hsu[b)]
Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

Jianxin Zhou[c)]
Department of Mathematics, Texas A&M University, College Station, Texas 77843

(Received 29 May 1998; accepted for publication 8 July 1998)

A wave equation on a one-dimensional interval I has a van der Pol type nonlinear boundary condition at the right end. At the left end, the boundary condition is fixed. At exactly the midpoint of the interval I, energy is injected into the system through a pair of transmission conditions in the feedback form of anti-damping. We wish to study chaotic wave propagation in the system. A cause of chaos by snapback repellers has been identified. These snapback repellers are repelling fixed points possessing homoclinic orbits of the non-invertible map in 2D corresponding to wave reflections and transmissions at, respectively, the boundary and the middle-of-the-span points. Existing literature [F. R. Marotto, J. Math. Anal. Appl. **63**, 199–223 (1978)] on snapback repellers contains an error. We clarify the error and give a refined theorem that snapback repellers imply chaos. Numerical simulations of chaotic vibration are also illustrated. © *1998 American Institute of Physics.* [S0022-2488(98)02012-X]

I. INTRODUCTION

Earlier, in a series of papers (Refs. 1–4), we studied chaotic vibrations of the wave equation due to a nonlinear self-excitation boundary condition of a van der Pol type: For the wave equation

$$w_{tt}(x,t)-w_{xx}(x,t)=0, \quad 0<x<1, \quad t>0, \tag{1}$$

let the right-end boundary condition be self-exciting:

$$w_x(1,t)=\alpha w_t(1,t)-\beta w_t^3(1,t), \quad \alpha,\beta>0, \quad t>0; \tag{2}$$

and let the left-end boundary condition be either fixed or free:

$$w(0,t)=0 \ \ (\text{fixed end}) \quad \text{or} \quad w_x(0,t)=0 \ \ (\text{free end}), \tag{3}$$

then the study in Refs. 2 and 4 shows that for initial conditions of generic type, the gradient (w_x,w_t) and the Riemann invariants (w_x+w_t,w_x-w_t) of (1) will be asymptotically periodic with one (for $0<\alpha\leqslant 1$) or two frequencies (for $\alpha>1$). Thus the wave equation does not have chaotic vibrations. However, instead of (3), if energy is injected into the system at the left end in the form of the boundary condition

[a)]Electronic mail: gchen@math.tamu.edu
[b)]Electronic mail: sbhsu@am.nthu.edu.tw
[c)]Electronic mail: jzhou@math.tamu.edu

 © 1998 American Institute of Physics

$$w_x(0,t) = -\eta w_t(0,t), \quad \eta > 0, \quad \eta \neq 1. \tag{4}$$

Then for a certain parameter range of η, the injected energy will excite (asymptotically) periodic vibrations into chaos (Refs. 3 and 4). The proofs in Refs. 1 and 4 are based upon the reflection relations of the equation (1) at the two endpoints $x=0$ and $x=1$, which is perhaps the most natural Poincaré section of the initial-boundary value problem stated above.

Existing study on chaos in differential equations is concerned mostly with unpredictable behavior originating from nonlinearities appearing either in the governing equations or/and boundary conditions. Very few works have treated chaotic nonlinear systems containing disturbances or excitations from in-span or interior *pointwise sources*. So let us consider the following scenario. Let (1)–(3) stay the same as above and, in addition, assume that energy is injected into the system at an *in-span point*, say $x=a$, where $0<a<1$. This is a *pointwise excitation*. Can this also rouse periodic vibrations to chaos?

A volcanic eruption on earth may be idealized as a pointwise disturbance or excitation to the physical system governing global weather, because the scale of the volcano is negligibly small compared with that of the earth. The eruption injects ashes, chemicals, and energy, among other things, into the atmospheric currents and the stratosphere and can easily throw the global weather patterns off balance and into chaos, provided that the eruption is long and forceful. Our study undertaken here bears the resemblance of the kind of pointwise excitation occurring in Mother Nature and, thus, signifying certain mathematical and physical relevance of both models meriting investigation. However, we must also point out the dissimilarity: the volcanic eruption injects energy into the weather system as *exogenous forcing*, whereas the energy injection at an in-span point considered in this paper is in the *endogenous feedback form of anti-damping.*

We now describe the details of the mathematical model we wish to treat. Consider the wave equation

$$w_{tt}(x,t) - w_{xx}(x,t) = 0, \quad -1<x<0, \quad 0<x<1, \quad t>0. \tag{5}$$

The spatial interval is chosen to be $(-1,1)$ just for convenience. The wave speed plays a very minor role in the subsequent mathematical analysis, so we just set it to be equal to 1 in (5). Again, let the right-end boundary condition be self-exciting:

$$w_x(1,t) = \alpha w_t(1,t) - \beta w_t^3(1,t), \quad 0<\alpha \leqslant 1, \quad \beta>0, \quad t>0. \tag{6}$$

Note here that we require $0<\alpha \leqslant 1$ in order to avoid hysteresis and nonuniqueness of solutions.[2,4] At the left end, assume that the boundary condition be fixed:

$$w(-1,t) = 0, \quad t>0. \tag{7}$$

[If this is replaced by the free end boundary condition $w_x(-1,t)=0, \quad t>0$, then the mathematical analysis remains qualitatively the same.] At *exactly the middle of the span, $x=0$, we consider two types of transmission conditions:*

$$\text{(Type I)} \begin{cases} w_t(0+,t) - w_t(0-,t) = -\eta w_x(0+,t), \\ w_x(0-,t) = w_x(0+,t), \end{cases} \quad t>0, \quad \eta>0, \quad \eta \neq 2, \tag{8}$$

or

$$\text{(Type II)} \begin{cases} w(0-,t) = w(0+,t) \\ w_x(0+,t) - w_x(0-,t) = -\eta w_t(0+,t), \end{cases} \quad t>0, \quad \eta>0, \quad \eta \neq 2. \tag{8'}$$

Also prescribed are two initial conditions

$$w(x,0) = w_0(x), \quad w_t(x,0) = w_1(x), \quad -1<x<1, \tag{9}$$

where the initial state (w_0, w_1) lies in appropriate function spaces. The energy, $E(t)$, of the overall system (5)–(9), at time t, is defined to be

$$E(t)=\frac{1}{2}\int_{-1}^{1}[w_x(x,t)^2+w_t(x,t)^2]\,dx.$$

We now examine the two sets of transmission conditions (8) and (8′). Consider the time rate of change of energy:

$$\begin{aligned}\frac{d}{dt}E(t)&=\int_{-1}^{0}(w_xw_{xt}+w_tw_{tt})\,dx+\int_{0}^{1}(w_xw_{xt}+w_tw_{tt})\,dx\\&=\ldots[\text{integrating by parts and utilizing}(5)\text{–}(8')]\\&=\alpha w_t(1,t)^2-\beta w_t(1,t)^4+\begin{cases}\eta w_x(0+,t)^2 & \text{for (8)},\\ \eta w_t(0+,t)^2 & \text{for (8)}'.\end{cases}\end{aligned}\tag{10}\tag{10′}$$

We see that if $\eta<0$, then the transmission conditions (8) or (8)′ would have contributed *loss* of energy to the system. Indeed, for $\eta<0$, (8) and (8)′ model (the *only*) two primary *feedback damping* designs we know of today in structural dynamics (Ref. 5, pp. 50–51). But here in (8) or (8)′, we require $\eta>0$. Therefore the in-span conditions (8) or (8)′ contribute *energy increase* to the system. Therefore, physically, (8) or (8)′ correspond to *feedback anti-damping* devices which *inject energy* into the system.

Note that the first two terms $\alpha w_t(1,t)^2-\beta w_t(1,t)^4$, in (10) or (10)′, signify as in Refs. 2–4 the *self-regulating* (or self-exciting) effect of the boundary condition (6) because we have

$$\alpha w_t(1,t)^2-\beta w_t(1,t)^4\begin{cases}\geqslant 0 & |w_t(1,t)|\leqslant(\alpha/\beta)^{1/2},\\ \text{if} & \\ <0 & |w_t(1,t)|>(\alpha/\beta)^{1/2},\end{cases}\tag{11}\tag{11′}$$

i.e., it causes energy to *rise* if the velocity magnitude $|w_t(1,t)|$ *is small*, and to *fall* if $|w_t(1,t)|$ *is large*, just like what the damping terms do in the van der Pol ordinary differential equation $m\ddot{x}+(-\alpha\dot{x}+\beta\dot{x}^3)+kx=0$. Incidentally, the transmission conditions (81) can be incorporated into the governing equation (5) by rewriting it as

$$w_{tt}(x,t)-w_{xx}(x,t)-\eta w_t(0+,t)\,\delta(x)=0,\quad t>0,\tag{12}$$

where $\delta(x)$ is the Dirac delta distribution concentrated at $x=0$. Therefore, (8)′ does indeed correspond to a pointwise disturbance or excitation. Even though (8) also corresponds to a pointwise excitation, we do not as yet know of any similar way to incorporate (8) into the governing equation (5) through the adding of some delta functions.

The advantage of choosing the exact middle-of-the-span point $x_0=0\pm$ in (8) or (8)′ is that it makes the *method of characteristics* easily applicable to our model problem for the purpose of mathematical analysis. If, instead, we replace $x_0=0\pm$ by $x_0=a\pm$ therein, for some arbitrary a, $-1<a<1$, then the problem obviously is much more generally posed. Unfortunately, this generality (of x_0) also renders the problem highly intractable in mathematical technicality. As we will see below, the approach of the method of characteristics adopted by us is *not robust* with respect to the choice of energy injection point $x_0=0\pm$ in the sense that a slight perturbation, say, changing $x_0=0\pm$ to $x_0=\varepsilon\pm$ for some small $\varepsilon\neq 0$, would immediately fail all the mathematical analysis based on this approach. So the question is, how good or sound is the main conclusion of the paper that chaotic vibrations exist when $x_0\neq 0\pm$? This may be responded to, in a nonrigorous way, as follows. From the work in Ref. 6, we know that, in the linear case, corresponding to the choice $x_0=\varepsilon\pm$ with ε being irrational, there are many *aperiodic* solutions and, thus, the general system (with $x_0=\varepsilon\pm$) should be ''even more chaotic'' than the special case $x_0=0\pm$ when the nonlinearity (6) is present. However, a rigorous proof of this heuristic claim seems to be far out of reach for the time being; more efforts are required in order to be able to treat the general case.

We are now in a position to apply the method of characteristics to treat (5)–(9) as follows. Define, by folding the interval $(-1,0)$ onto the interval $(0,1)$, the following:

$$\left.\begin{aligned} w^{(1)}(x,t) &= -w(-x,t) \\ w^{(2)}(x,t) &= w(x,t) \end{aligned}\right\} \quad \text{for } x\in[0,1], \quad t\geqslant 0, \tag{13}$$

$$u_i=\tfrac{1}{2}(w_x^{(i)}+w_t^{(i)}), \quad v_i=\tfrac{1}{2}(w_x^{(i)}-w_t^{(i)}), \quad i=1,2. \tag{14}$$

Then the wave equation (5) is converted into a first-order hyperbolic system

$$\frac{\partial}{\partial t}\begin{bmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \frac{\partial}{\partial x}\begin{bmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{bmatrix}, \quad 0<x<1, \quad t>0. \tag{15}$$

The reflection relation at $x=1$, according to (6), (7), and (14), is

$$\begin{bmatrix} u_1(1,t) \\ u_2(1,t) \end{bmatrix} = \mathcal{R}_{1,\alpha\beta}\left(\begin{bmatrix} v_1(1,t) \\ v_2(1,t) \end{bmatrix}\right) \equiv \begin{bmatrix} v_1(1,t) \\ F_{\alpha,\beta}(v_2(1,t)) \end{bmatrix}, \tag{16}$$

where for each given $v\in\mathbb{R}$ the function $u=F_{\alpha,\beta}(v)$ is well defined through the following implicit cubic equation,[2]

$$\beta(u-v)^3+(1-\alpha)(u-v)+2v=0. \tag{17}$$

Thus (16) constitutes the right-end boundary condition for (15). In what follows, we often write $\mathcal{R}_{1,\alpha\beta}$ simply as $\mathcal{R}_1$, in case no ambiguities should occur.

Remark I.1: In (17), for given fixed α, $0<\alpha\leqslant 1$, there exists a unique $u\in\mathbb{R}$ satisfying (17) for each given $v\in\mathbb{R}$. Contrarily, if $\alpha\notin(0,1]$, i.e., $1-\alpha\notin[0,1)$, for each given $v\in\mathbb{R}$, then there may exist one, two, or three real solution(s) $u\in\mathbb{R}$, and thus $F_{\alpha,\beta}$ is no longer a well-defined function. See more in (31)–(33) below. □

The reflection (i.e., transmission) relation at $x=0$, according to (8), (8′), and (14), is, respectively,

$$\begin{bmatrix} v_1(0,t) \\ v_2(0,t) \end{bmatrix} = \mathcal{R}_{0,\eta}\left(\begin{bmatrix} u_1(0,t) \\ u_2(0,t) \end{bmatrix}\right) \equiv \begin{bmatrix} \dfrac{\eta}{2-\eta} & \dfrac{2}{2-\eta} \\ \dfrac{2}{2-\eta} & \dfrac{\eta}{2-\eta} \end{bmatrix} \begin{bmatrix} u_1(0,t) \\ u_2(0,t) \end{bmatrix} \quad \text{for (8)}; \tag{18}$$

$$\begin{bmatrix} v_1(0,t) \\ v_2(0,t) \end{bmatrix} = -\mathcal{R}_{0,\eta}\left(\begin{bmatrix} u_1(0,t) \\ u_2(0,t) \end{bmatrix}\right), \quad \text{for (8′)}. \tag{18′}$$

The above [(18) or (18′)) constitutes the left-end boundary condition for (15). From now on, we often abbreviate $\mathcal{R}_{0,\eta}$ as $\mathcal{R}_0$. By abuse of notation, we will make no distinction between $\mathcal{R}_{0,\eta}$ and the matrix on the rhs of (18).

The original initial conditions (9) now lead to

$$\begin{bmatrix} u_1(x,0) \\ u_2(x,0) \\ v_1(x,0) \\ v_2(x,0) \end{bmatrix} = \begin{bmatrix} u_{1,0}(x) \\ u_{2,0}(x) \\ v_{1,0}(x) \\ v_{2,0}(x) \end{bmatrix}, \quad 0<x<1, \tag{19}$$

for some functions $u_{i,0}(x), v_{i,0}(x), x\in(0,1), i=1,2$, according to (14). In summary, (15)–(19) constitute the complete set of a well-posed initial-boundary value problem. This system has a unique solution (u_1,u_2,v_1,v_2): for $t=2k+\tau$, $k=0,1,2,\ldots,0\leqslant\tau<2$ and $0\leqslant x\leqslant 1$,

$$\begin{bmatrix} u_1(x,t) \\ u_2(x,t) \end{bmatrix} = \begin{cases} (\mathcal{R}_1 \circ \mathcal{R}_0)^k \left(\begin{bmatrix} u_{1,0}(x+\tau) \\ u_{2,0}(x+\tau) \end{bmatrix} \right), & \tau \leqslant 1-x, \\ \mathcal{R}_0^{-1} \circ (\mathcal{R}_0 \circ \mathcal{R}_1)^{k+1} \left(\begin{bmatrix} v_{1,0}(2-x-\tau) \\ v_{2,0}(2-x-\tau) \end{bmatrix} \right), & 1-x < \tau \leqslant 2-x, \\ (\mathcal{R}_1 \circ \mathcal{R}_0)^{k+1} \left(\begin{bmatrix} u_{1,0}(\tau+x-2) \\ u_{2,0}(\tau+x-2) \end{bmatrix} \right), & 2-x < \tau \leqslant 2; \end{cases}$$

$$\begin{bmatrix} v_1(x,t) \\ v_2(x,t) \end{bmatrix} = \begin{cases} (\mathcal{R}_0 \circ \mathcal{R}_1)^k \left(\begin{bmatrix} v_{1,0}(x-\tau) \\ v_{2,0}(x,\tau) \end{bmatrix} \right), & \tau \leqslant x, \\ \mathcal{R}_0 \circ (\mathcal{R}_0 \circ \mathcal{R}_1)^k \left(\begin{bmatrix} u_{1,0}(\tau-x) \\ u_{2,0}(\tau-x) \end{bmatrix} \right), & x < \tau \leqslant 1+x, \\ (\mathcal{R}_0 \circ \mathcal{R}_1)^{k+1} \left(\begin{bmatrix} v_{1,0}(2+x-\tau) \\ v_{2,0}(2+x-\tau) \end{bmatrix} \right), & 1+x < \tau \leqslant 2, \end{cases} \tag{20}$$

which has the same form as Ref. 2, (13) and (14). In the above, all the $\mathcal{R}_0$'s may be replaced by $-\mathcal{R}_0$'s if (18′) [or, equivalently, (8′))] takes place in lieu of (18).

Naturally, from the explicit representation (20) of the solution, the system (15)–(19) manifests chaotic behavior if and only if (the iterates of) the composite map(s) $\mathcal{R}_0\mathcal{R}_1$ and/or $\mathcal{R}_1\mathcal{R}_0$ is/are chaotic. Actually, $\mathcal{R}_0\mathcal{R}_1$ and $\mathcal{R}_1\mathcal{R}_0$ have identical dynamical behavior. This is seen in the following.

Proposition I.1: For each given α, $0<\alpha\leqslant 1$, $\beta>0$, and $\eta>0$, $\eta\neq 2$, the maps $\mathcal{R}_{0,\eta}\mathcal{R}_{1,\alpha\beta}$ and $\mathcal{R}_{1,\alpha\beta}\mathcal{R}_{0,\eta}$ are topologically conjugate.

Proof: First, we note that $\mathcal{R}_{0,\eta}$ is an invertible 2×2 matrix. The rest is then obvious from the following commutative diagram:

$$\begin{array}{ccc} \mathbb{R}^2 & \xrightarrow{\mathcal{R}_{0,\eta}\mathcal{R}_{1,\alpha\beta}} & \mathbb{R}^2 \\ {\scriptstyle \mathcal{R}_{0,\eta}^{-1}}\downarrow & & \downarrow{\scriptstyle \mathcal{R}_{0,\eta}^{-1}} \\ \mathbb{R}^2 & \xrightarrow{\mathcal{R}_{1,\alpha\beta}\mathcal{R}_{0,\eta}} & \mathbb{R}^2 \end{array}$$

□

Therefore, investigation of the periodic and chaotic behavior of only $\mathcal{R}_0\mathcal{R}_1$ suffices, because it implies that of $\mathcal{R}_1\mathcal{R}_0$, and vice versa.

If the transmission conditions (8′) take effect rather than (8), then we need to use (18′) and, consequently, (20), but with all $\mathcal{R}_0$'s therein substituted by $-\mathcal{R}_0$'s. This means that we must investigate the chaotic behavior of $-\mathcal{R}_0\mathcal{R}_1$ rather than $\mathcal{R}_0\mathcal{R}_1$. From the mathematical analysis point of view, we have found that the treatment of $-\mathcal{R}_0\mathcal{R}_1$ is qualitatively the same as that of $\mathcal{R}_0\mathcal{R}_1$. Henceforth, we will therefore only consider (8), (18), and the ensuing $\mathcal{R}_0\mathcal{R}_1$.

The main objective of this paper is to study the occurrence of chaos for the wave equation system as described. Recall from our earlier studies in Refs. 2–4 that the parameter β plays the *role of scaling*. Therefore, we may just fix β to be a positive constant, say $\beta=1$. By varying α and η increasingly from zero, either separately or jointly, we have observed through computer simulations at least the following three routes/sources of chaos for the map $\mathcal{R}_{0,\eta}R_{1,\alpha\beta}$:

A. Period doubling

For positive α and η close to zero, the map $\mathcal{R}_{0,\eta}\mathcal{R}_{1,\alpha\beta}$ has a stable period-4 orbit in two dimensions (2D). (It does not have any period-2 orbit, according to Proposition III.5 in Sec. III below.) Let us, say, fix $\alpha=1/2$ (and $\beta=1$), for example. By increasing η, at $\eta\approx 0.6725$, we observe that this period-4 orbit loses stability, and a new stable period-8 orbit supersedes it. The ratios of successive differences of parametrical values η where period doublings occur have been verified to tend to Feigenbaum's constant. So the numerical evidence in support of period doubling is strong and beyond doubt. Nevertheless, at this moment, the authors are only partially successful in establishing a period-doubling bifurcation theorem in 2D and in carrying out a computer-aided verification of the theorem. We hope to be able to defer a complete presentation of this to a sequel.

B. Snapback repellers

These are repelling fixed points or periodic points in whose neighborhoods a homoclinic orbit originates. For example (see Example V.1 in Sec. V), the origin in $\mathbb{R}^2$, as a repelling fixed point, is a snapback repeller for a restricted range of α and η values. So far, this is the only case for which we have sufficient understanding; its discussion will be the focus of the study in this paper (see more in Secs. IV and V).

C. Nondiffeomorphic horseshoes

The Smale horseshoe is a powerful technique for proving chaos of multidimensional maps. However, the powerful Smale–Birkhoff homoclinic theorem (Ref. 7, pp. 482–483) requires that the map be a *diffeomorphism*; so it must at least be *invertible*. Ours is not the case here because $\mathcal{R}_{1,\alpha\beta}$ is not invertible for any α, $0<\alpha\leqslant 1$, $\beta>0$; see Remark I.1 and (31)–(33) below. The lack of invertibility of $\mathcal{R}_1$ and, consequently, of $\mathcal{R}_0\mathcal{R}_1$ is related to the prevalent *irreversible* behavior of time-dependent nonlinear PDEs; see Sec. II. Numerical evidence strongly suggests that the map $\mathcal{R}_0\mathcal{R}_1$ has many periodic points of *saddle node* type in whose neighborhoods homoclinic orbits originate and, thus, we speculate that they cause chaos. (For lack of a better term, we call this a ''non-diffeomorphic horseshoe'' for the time being.) See some details in Example V.2. Notwithstanding, we must concede that before a rigorous proof is given, this remains just a speculation. For general non-invertible maps in two- or higher-dimensional spaces, as pointed out by Mira *et al.*[8] there are *not many optional* theoretical methods available to rigorously prove the occurrence of chaos. More dedicated cultivation of this area is very desirable.

The organization of the paper proceeds as follows. In Sec. II, we discuss the time irreversibility of our PDE system. In Sec. III, elementary properties of the map, including fixed points, stability, and invariant domains, are studied. In Sec. IV, we study chaos caused by snapback repellers. We point out an error in an earlier work by Marotto[9] and give a refined proof. In Sec. V, we present examples and illustrations of chaotic vibrations.

II. IRREVERSIBILITY FOR NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS: THE LACK OF DIFFEOMORPHISM TO FORM A SMALE HORSESHOE

Recall from Ref. 2 that for $0<\alpha\leqslant 1$, $\beta>0$, the implicit relation (17) determines a unique function $u=F(v)=F_{\alpha,\beta}(v)$, where

$$u=F(v)=v+\left[-\frac{v}{\beta}+\sqrt{\frac{1}{27}\left(\frac{1-\alpha}{\beta}\right)^3+\frac{v^2}{\beta^2}}\right]^{1/3}+\left[-\frac{v}{\beta}-\sqrt{\frac{1}{27}\left(\frac{1-\alpha}{\beta}\right)^3+\frac{v^2}{\beta^2}}\right]^{1/3}, \tag{21}$$

by Cardan's formula. The function F is odd, with two critical points—one maximum and one minimum—at, respectively $-v_c^*$ and v_c^*, where $v_c^*=[(2-\alpha)/3]\sqrt{(1+\alpha)/3\beta}$. We also know that F satisfies the following properties:

$$\text{(i)}\quad F'(0)=-\frac{1+\alpha}{1-\alpha}, \tag{22}$$

$$\text{(ii)}\quad 0<F'(v)<1 \quad \text{for } |v|>v_c^*; \tag{23}$$

$$\text{(iii)}\quad F \text{ has three intercepts at } v=-v_I,\ 0,\ v_I, \quad \text{where } v_I\equiv\sqrt{\frac{1+\alpha}{\beta}}. \tag{24}$$

Also, recall from Ref. 3 that for η, $0<\eta<1$, the map $u=G\circ F(v)=G_\eta\circ F_{\alpha,\beta}(v)$, where $G_\eta\equiv(1+\eta)/(1-\eta)$, has exactly three fixed points

$$v=G\circ F(v), \quad \text{for } v=-v^*,\ 0,\ v^*; \quad v^*\equiv\frac{1+\eta}{2\eta}\sqrt{\frac{1+\alpha\eta}{\beta\eta}}. \tag{25}$$

Now, let us consider the question whether the system (15)–(19) is *time reversible*. We make a change of variable $t\mapsto -t$ and consider $t\geqslant 0$; we obtain the time-reversed system

$$\frac{\partial}{\partial t}\begin{bmatrix}\tilde{u}_1\\ \tilde{u}_2\\ \tilde{v}_1\\ \tilde{v}_2\end{bmatrix}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix}\tilde{u}_1\\ \tilde{u}_2\\ \tilde{v}_1\\ \tilde{v}_2\end{bmatrix},\quad 0<x<1,\quad t>0, \tag{26}$$

with boundary conditions

$$\text{at } x=0\begin{bmatrix}\tilde{v}_1(0,t)\\ \tilde{v}_2(0,t)\end{bmatrix}=\tilde{\mathcal{R}}_0\left(\begin{bmatrix}\tilde{u}_1(0,t)\\ \tilde{u}_2(0,t)\end{bmatrix}\right)=\begin{bmatrix}-\frac{\eta}{2+\eta}&\frac{2}{2+\eta}\\ \frac{2}{2+\eta}&-\frac{\eta}{2+\eta}\end{bmatrix}\begin{bmatrix}\tilde{u}_1(0,t)\\ \tilde{u}_2(0,t)\end{bmatrix},\quad t>0,\quad (\eta>0), \tag{27}$$

$$\text{at } x=1\begin{bmatrix}u_1(1,t)\\ u_2(1,t)\end{bmatrix}=\tilde{\mathcal{R}}_1\left(\begin{bmatrix}\tilde{v}_1(1,t)\\ \tilde{v}_2(1,t)\end{bmatrix}\right)=\begin{bmatrix}\tilde{v}_1(1,t)\\ \tilde{F}_{\alpha,\beta}(\tilde{v}_2(1,t)\end{bmatrix},\quad t>0, \tag{28}$$

and certain initial conditions

$$\tilde{u}_i(x,0)=\tilde{u}_{i,0}(x),\quad \tilde{v}_i(x,0)=\tilde{v}_{i,0}(x),\quad i=1,2,\quad x\in[0,1], \tag{29}$$

where, in (28), the relation $\tilde{u}=\tilde{F}_{\alpha,\beta}(\tilde{v})$ is defined through the following cubic equation:

$$\beta(\tilde{u}-\tilde{v})^3-(1+\alpha)(\tilde{u}-\tilde{v})-2\tilde{v}=0,\quad 0<\alpha\leqslant 1,\quad \beta>0. \tag{30}$$

Note that $\tilde{u}$ and $\tilde{v}$ are restricted to be reals in (30). The relation $\tilde{F}_{\alpha,\beta}$ is *not* a function, as Cardan's formula for cubic algebraic equations gives the solutions of (30) as follows: let

$$D(\tilde{v})\equiv-\frac{1}{27}\frac{(1+\alpha)^3}{\beta^3}+\frac{\tilde{v}^2}{\beta^2}$$

be the discriminant for (30). Then
(i) if $D(\tilde{v})>0$, then $\tilde{u}$ is uniquely determined by $\tilde{v}$:

$$\tilde{u}=\tilde{F}_{\alpha,\beta}(\tilde{v})=\tilde{v}+\left[-\frac{\tilde{v}}{\beta}+\sqrt{D(\tilde{v})}\right]^{1/3}+\left[-\frac{\tilde{v}}{\beta}-\sqrt{D(\tilde{v})}\right]^{1/3}, \tag{31}$$

(ii) If $D(\tilde{v})<0$, then (30) yields three real solutions:

$$\tilde{u}_k=2S\ \cos\frac{\phi+2k\pi}{3};\quad S\equiv\sqrt{\frac{1}{3}\frac{1+\alpha}{\beta}},\quad \phi\equiv\cos^{-1}\left[\frac{3\sqrt{3}\tilde{v}}{(1+\alpha)}\sqrt{\frac{\beta}{1+\alpha}}\right],\quad k=1,2,3. \tag{32}$$

(iii) If $D(\tilde{v})=0$, then (30) yields two distinct real solutions:

$$\tilde{u}_1=2\sqrt[3]{\frac{\tilde{v}}{\beta}},\quad \tilde{u}_2=\sqrt[3]{-\frac{\tilde{v}}{\beta}}\ \text{(with multiplicity 2)}. \tag{33}$$

The multiplicity of solutions given in (32) and (33) spells trouble for the system (26)–(29). We have the following.

Theorem II.1 [Lack of global irreversibility of the system (15)–(19)]: *Let (26)–(29) be the time-reversed system of (15)–(19). Assume that the initial conditions* $\tilde{u}_{i,0}(x),\tilde{v}_{i,0}(x)$, $i=1,2$, *in (29) are continuous functions on [0,1] such that not all of these functions are identically zero. Then there exists an* $\varepsilon>0$ *such that if*

$$0 \leqslant \inf_{x\in[0,1]} |\tilde{u}_{i,0}(x)| < \varepsilon, \quad \text{or} \quad 0 \leqslant \inf_{x\in[0,1]} |\tilde{v}_{i,0}(x)| < \varepsilon, \quad \text{for some } i = 1 \text{ or } 2,$$

then the solution of the system (26)–(29) is not unique for $t>T$, for some $T>0$.

Proof: If $\tilde{v}_{2,0}(x) \not\equiv 0$, and if there exists some $\xi \in (0,1)$ such that

$$|\tilde{v}_{2,0}(\xi)| \leqslant \frac{1+\alpha}{3\sqrt{3}} \sqrt{\frac{1+\alpha}{\beta}}, \tag{34}$$

then $D(\tilde{v}_{2,0}(\xi)) \leqslant 0$. By (32) or (33), after the reflection at $x=1$ takes place at time $t=1-\xi$ through the relation (28), we see that the $\tilde{u}_2$-component lacks uniqueness. Therefore the solution of the system (26)–(29) loses uniqueness after $T=1-\xi$.

If $\tilde{v}_{2,0}(x) \equiv 0$ or if (34) does not hold for any $\xi \in (0,1)$, then there exists some $\xi \in (0,1)$ such that one of the following,

$$0 < |\tilde{u}_{1,0}(\xi)| < \varepsilon, \quad 0 < |\tilde{u}_{2,0}(\xi)| < \varepsilon, \quad 0 < |\tilde{v}_{1,0}(\xi)| < \varepsilon, \tag{35}$$

is true. We note that at $x=0$ the reflection matrix $\tilde{\mathcal{R}}_0$ has two eigenvalues: -1 and $(2-\eta)/(2+\eta)$, both with magnitudes not larger than 1. Thus $\tilde{\mathcal{R}}_0$ is nonexpansive. By choosing ε sufficiently small and tracing reflections along characteristics, by using the nonexpansiveness of $\tilde{\mathcal{R}}_0$ plus some detailed arguments (which are omitted), we obtain from (20) that at some $t_0>0$, we will gain

$$|\tilde{v}_2(\tilde{\xi}, t_0)| \leqslant \frac{1+\alpha}{3\sqrt{3}} \sqrt{\frac{1+\alpha}{\beta}}, \quad \text{for some } \tilde{\xi} \in (0,1).$$

Therefore again (32) or (33) applies, and, by (28), the solution loses uniqueness. □

The above irreversible behavior has all but ruled out the direct applicability of Smale's horseshoe to our problem.

III. ELEMENTARY PROPERTIES OF THE MAP $\mathcal{R}_0\mathcal{R}_1$

In this section, we perform an elementary stability analysis of fixed points of $\mathcal{R}_0\mathcal{R}_1$ and determine some invariant regions.

Proposition III.1: Let $0<\alpha\leqslant 1$, $\beta \geqslant 1$, and $\eta>0, \eta\neq 2$. Then the map $\mathcal{R}_0\mathcal{R}_1$ has exactly three fixed points:

$$(0,0), (v_1^*, v_2^*) \equiv \left(\frac{1}{2\eta} \sqrt{\frac{1+\alpha\eta}{\beta\eta}}, \frac{1+\eta}{2\eta} \sqrt{\frac{1+\alpha\eta}{\beta\eta}} \right), \quad (-v_1^*, -v_2^*). \tag{36}$$

Proof: We determine the fixed points by solving

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathcal{R}_0\mathcal{R}_1 \left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right) = \begin{bmatrix} \dfrac{\eta}{2-\eta} v_1 + \dfrac{2}{2-\eta} F(v_2) \\ \dfrac{2}{2-\eta} v_1 + \dfrac{\eta}{2-\eta} F(v_2) \end{bmatrix}. \tag{37}$$

From the first component equation in (37), we obtain

$$v_1 = \frac{F(v_2)}{1-\eta}, \quad \eta \neq 1, \tag{38a}$$

$$F(v_2) = 0, \quad \eta = 1. \tag{38b}$$

For $\eta \neq 1$, substituting $(38a)_1$ into the second component equation of (37), we obtain

$$v_2=\frac{1+\eta}{1-\eta}F(v_2). \tag{39}$$

However, in the notation of Ref. 3, (2.1), (39) says exactly that $v_2=G\circ F(v_2)$, i.e., v_2 is a fixed point of the 1D map $G\circ F$, which according to (25) and (39) gives three values of v_2:

$$v_2=0,\quad \frac{1+\eta}{2\eta}\sqrt{\frac{1+\alpha\eta}{\beta\eta}},\quad -\frac{1+\eta}{2\eta}\sqrt{\frac{1+\alpha\eta}{\beta\eta}}. \tag{40}$$

Substituting these values first into (39) to obtain $F(v_2)$ and next into (38a), we obtain (36).

For $\eta=1$, (38b) says that v_2 is a v-axis intercept of F. By (24), there are three intercepts:

$$v_2=0,\quad \sqrt{\frac{1+\alpha}{\beta}},\quad -\sqrt{\frac{1+\alpha}{\beta}}. \tag{41}$$

Using these values and (38b) in the second component equation of (37), we obtain $v_1=v_2/2$. Thus we again have (36) with η being set to 1 therein. □

For any differentiable map $\mathcal{F}$: $\mathbb{R}^N\to\mathbb{R}^N$, we let $D\mathcal{F}(x)$ denote the Jacobian matrix of $\mathcal{F}$ at x. For our map $\mathcal{R}_0\mathcal{R}_1$, we have

$$D(\mathcal{R}_0\mathcal{R}_1)(v_1,v_2)=\begin{bmatrix}\frac{\eta}{2-\eta} & \frac{2}{2-\eta}F'(v_2)\\ \frac{2}{2-\eta} & \frac{\eta}{2-\eta}F'(v_2)\end{bmatrix}. \tag{42}$$

The stability analysis of fixed points and periodic points is well known to be important in the understanding of the dynamics of the map $\mathcal{R}_0\mathcal{R}_1$. We perform such an analysis for the three fixed points given in Proposition III.1.

Proposition III.2: Let $0<\alpha\leqslant 1,\beta>0$, and $0<\eta<2$. The fixed point (0,0) is a repelling fixed point of $\mathcal{R}_0\mathcal{R}_1$.

Proof: First, note from Ref. 2, Sec. III, (P. 3), that $F'(0)=-(1+\alpha)/(1-\alpha)$. From (42), we get

$$J\equiv D(\mathcal{R}_0\mathcal{R}_1)(0,0)=\begin{bmatrix}\frac{\eta}{2-\eta} & -\frac{2}{2-\eta}\frac{1+\alpha}{1-\alpha}\\ \frac{2}{2-\eta} & -\frac{\eta}{2-\eta}\frac{1+\alpha}{1-\alpha}\end{bmatrix}. \tag{43}$$

The eigenvalues of J are computed from

$$0=\det(\lambda I_2-J)\quad (\det\equiv\text{determinant})=\lambda^2-\left(1-\frac{1+\alpha}{1-\alpha}\right)\frac{\eta}{2-\eta}\lambda-\frac{1+\alpha}{1-\alpha}\left[\left(\frac{\eta}{2-\eta}\right)^2-\left(\frac{2}{2-\eta}\right)^2\right],$$

which are

$$\lambda_1,\lambda_2=\frac{1}{2}\left[-\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\pm\sqrt{\left(\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\right)^2-\frac{4(1+\alpha)}{1-\alpha}\frac{2+\eta}{2-\eta}}\right], \tag{44}$$

$$(\text{``}+\text{''}\quad\text{for}\ \lambda_1,\text{``}-\text{''}\text{for}\lambda_2).$$

If the discriminant Δ is negative,

$$\Delta\equiv\left(\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\right)^2-\frac{4(1+\alpha)}{1-\alpha}\frac{2+\eta}{2-\eta}<0, \tag{45}$$

which happens if

$$\eta^2-4(1-\alpha^2)<0, \tag{46}$$

then λ_1 and λ_2 are complex conjugates of each other and, for $0<\eta<2$,

$$|\lambda_1|^2=|\lambda_2|^2=\lambda_1\lambda_2=-\frac{1+\alpha}{1-\alpha}\left[\left(\frac{\eta}{2-\eta}\right)^2-\left(\frac{2}{2-\eta}\right)^2\right]=\frac{1+\alpha}{1-\alpha}\frac{2+\eta}{2-\eta}>1.$$

Therefore the origin is a repelling *spiral point*.

On the other hand, if $\Delta\geqslant 0$, then $\lambda_2<\lambda_1<0$. We now show that $\lambda_1<-1$, or, from (44),

$$-\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}+\sqrt{\left(\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\right)^2-\frac{4(1+\alpha)}{1-\alpha}\frac{2+\eta}{2-\eta}}<-2,$$

or

$$\sqrt{\left(\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\right)^2-\frac{4(1+\alpha)}{1-\alpha}\frac{2+\eta}{2-\eta}}<-2+\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}. \tag{47}$$

Note that the rhs of the inequality above is non-negative because

$$\Delta\geqslant 0\Rightarrow\alpha^2+\frac{\eta^2}{2^2}\geqslant 1\Rightarrow\alpha+\frac{\eta}{2}\geqslant\left(\alpha^2+\frac{\eta^2}{2^2}\right)^{1/2}\geqslant 1 \quad \text{for } \alpha>0, \quad \eta>0$$

$$\Rightarrow 2\alpha+\eta-2\geqslant 0\Rightarrow -2+\frac{2\alpha}{1-\alpha}\frac{\eta}{2-\eta}\geqslant 0 \quad \text{for } 0<\alpha<1, \quad 0<\eta<2.$$

Squaring both sides of (47), we see that $\lambda_1<-1$ if and only if

$$-\frac{4(1+\alpha)}{1-\alpha}\cdot\frac{2}{2-\eta}<8, \quad \text{for } 0<\alpha<1, \quad 0<\eta<2,$$

which is always valid. Therefore $\lambda_2<\lambda_1<-1$. □

Corollary III.1: Let $0<\alpha\leqslant 1, \beta>0$, *and* $0<\eta<2$. *Then*

(i) if $\eta^2-4(1-\alpha^2)<0$, *then* (0,0) *is a repelling spiral point of* $\mathcal{R}_0\mathcal{R}_1$.

(ii) if $\eta^2-4(1-\alpha^2)\geqslant 0$, *then* (0,0) *is a repelling (star) nodal point of* $\mathcal{R}_0\mathcal{R}_1$. □

Proposition III.3: Let $0<\alpha\leqslant 1, \beta>1$, *and* $0<\eta<2$. *Then the fixed points* (v_1^*, v_2^*) *and* $(-v_1^*, -v_2^*)$ *are unstable saddle nodes of* $\mathcal{R}_0\mathcal{R}_1$.

Proof: We need only consider the stability of (v_1^*, v_2^*); that of $(-v_1^*, -v_2^*)$ follows from a symmetry argument.

The Jacobian matrix at (v_1^*, v_2^*), by (42), is

$$J\equiv\begin{bmatrix}\frac{\eta}{2-\eta} & \frac{2}{2-\eta}F'(v_2^*)\\ \frac{2}{2-\eta} & \frac{\eta}{2-\eta}F'(v_2^*)\end{bmatrix}.$$

The eigenvalues λ_1 and λ_2 of J satisfy

$$\begin{aligned}0&=\det(\lambda I_2-B)=\lambda^2-\frac{\eta}{2-\eta}(1+F'(v_2^*))-\frac{2+\eta}{2-\eta}F'(v_2^*),\\ \lambda_1,\lambda_2&=\frac{1}{2}\left[\frac{\eta}{2-\eta}(1+F'(v_2^*))\pm\sqrt{\left(\frac{\eta}{2-\eta}\right)^2(1+F'(v_2^*))^2+4\cdot\frac{2+\eta}{2-\eta}F'(v_2^*)}\right],\\ &(\text{``}+\text{''} \text{ for } \lambda_1, \quad \text{``}-\text{''} \text{ for } \lambda_2).\end{aligned} \tag{48}$$

From (22), (23), and Ref. 3, Lemma 2.4, we have

$$0<F'(v_2^*)<1, \quad \frac{1+\eta}{1-\eta}F'(v_2^*)>1. \tag{49}$$

Therefore we have $\lambda_1>0$, $\lambda_2>0$. We want to show further that

$$\lambda_1>1, \quad -1<\lambda_2<0. \tag{50}$$

From (48), we see that $\lambda_1>1$ holds if and only if

$$\sqrt{\left(\frac{\eta}{2-\eta}\right)^2(1+F'(v_2^*))^2+4\cdot\frac{2+\eta}{2-\eta}F'(v^*)}>2-\frac{\eta}{2-\eta}(1+F'(v_2^*)). \tag{51}$$

If $2-[\eta/(2-\eta)](1+F'(v_2^*))\leqslant 0$, then (51) automatically holds. So we consider $2-[\eta/(2-\eta)](1+F'(v_2^*))>0$. Squaring both sides of (51) and simplifying yields

$$\lambda_1>0\Leftrightarrow 4\cdot\frac{2+\eta}{2-\eta}F'(v_2^*)>4-\frac{4\eta}{2-\eta}(1+F'(v_2^*)),$$

$$\Leftrightarrow F'(v_2^*)>\frac{1-\eta}{1+\eta},$$

which holds by (49).

Next, we show that $0>\lambda_2>-1$. From (48),

$$\lambda_2>-1\Leftrightarrow 2+\frac{\eta}{2-\eta}(1+F'(v_2^*))>\sqrt{\left(\frac{\eta}{2-\eta}\right)^2(1+F'(v_2^*))^2+4\cdot\frac{2+\eta}{2-\eta}F'(v_2^*)}. \tag{52}$$

Squaring both sides of (52) and simplifying yields

$$\frac{4\eta}{2-\eta}(1+F'(v_2^*))+4>4\cdot\frac{2+\eta}{2-\eta}F'(v_2^*),$$

$$F'(v_2^*)<1.$$

The above again is true by (23). Therefore $\lambda_2>-1$. □

We now determine an elementary invariant rectangle, for positive η which is close to 0.

Proposition III.4: Assume $0<\alpha\leqslant 1,\beta>0$, *and* $0<\eta<1$. *Let* $G=G_\eta, F=F_{\alpha,\beta}$, v^*, v_I *and* v_c^* *be defined as in* (22)–(25). *Assume that* $\eta_{\alpha,\beta}^*$ *satisfies*

$$G_{\eta_{\alpha,\beta}^*}F_{\alpha,\beta}(-v_c^*)\leqslant v^*. \tag{53}$$

Then for η, $0<\eta<\eta_{\alpha,\beta}^*$,

$$D=\{(v_1,v_2)\in\mathbb{R}^2 \,|\, |v_1|\leqslant v_1^*, |v_2|\leqslant v_2^*\} \tag{54}$$

is invariant under $\mathcal{R}_0\mathcal{R}_1$, *i.e.,* $\mathcal{R}_0\mathcal{R}_1(D)\subset D$, *where*

$$v_1^*\equiv\frac{v^*}{1+\eta}, \quad v_2^*\equiv v^*. \tag{55}$$

Proof: Let $(v_1,v_2)\in D$ and

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathcal{R}_0\mathcal{R}_1\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right) = \begin{bmatrix} \frac{\eta}{2-\eta}v_1 + \frac{2}{2-\eta}F(v_2) \\ \frac{2}{2-\eta}v_1 + \frac{\eta}{2-\eta}F(v_2) \end{bmatrix}. \tag{56}$$

We want to show that $(u_1,u_2)\in D$.

By (25) and (53), using Ref. 3, Lemma 2.5, we first note that $\mathcal{I}\equiv[-v^*,v^*]$ is an invariant interval of $G\circ F$. So we have

$$-v_2^* \leqslant \frac{1+\eta}{1-\eta}F(v_2) \leqslant v_2^*, \quad \text{for all } v_2, |v_2| \leqslant v_2^* = v^*. \tag{57}$$

Thus, for $v_1, |v_1| \leqslant v_1^*$,

$$u_1 = \frac{\eta}{2-\eta}v_1 + \frac{2}{2-\eta}F(v_2) \leqslant \frac{\eta}{2-\eta}v_1^* + \frac{2}{2-\eta}F(v_2).$$

However,

$$\frac{\eta}{2-\eta}v_1^* + \frac{2}{2-\eta}F(v_2) \leqslant v_1^* \Leftrightarrow \frac{2}{2-\eta}F(v_2) \leqslant \frac{2(1-\eta)}{2-\eta}v_1^* \Leftrightarrow \frac{1+\eta}{1-\eta}F(v_2) \leqslant v_2^*. \tag{58}$$

By (57) we know that the last inequality in (58) holds. Therefore

$$u_2 \leqslant v_1^*. \tag{59}$$

Next,

$$u_1 = \frac{\eta}{2-\eta}v_1 + \frac{2}{2-\eta}F(v_2) \geqslant -\frac{\eta}{2-\eta}v_1^* + \frac{2}{2-\eta}F(v_2).$$

We have

$$\begin{aligned} -\frac{\eta}{2-\eta}v_1^* + \frac{2}{2-\eta}F(v_2) \geqslant -v_1^* &\Leftrightarrow \frac{2}{2-\eta}F(v_2) \\ &\geqslant \frac{(-2)(1-\eta)}{2-\eta}v_1^* \Leftrightarrow \frac{1+\eta}{1-\eta}F(v_2) \\ &\geqslant -(1+\eta)v_1^* \\ &= -v_2^* \Leftrightarrow \frac{1+\eta}{1-\eta}F(v_2) \geqslant -v_2^*. \end{aligned} \tag{60}$$

By (57), we again know that the last inequality in (60) holds. Hence $u_1 \geqslant -v_1^*$, which, incorporated with (59), gives $|u_1| \leqslant v_1^*$.

The proof that u_2 satisfies $|u_2| \leqslant v_2^*$ is similar and therefore omitted. □

Example III.1:

By Ref. 3, (2.20), and (25), we can determine $\eta_{\alpha,\beta}^*$ from the equality

$$G_{\eta_{\alpha,\beta}^*}F_{\alpha,\beta}(-v_c^*) = \frac{1+\eta_{\alpha,\beta}^*}{1-\eta_{\alpha,\beta}^*}\frac{1+\alpha}{3}\sqrt{\frac{1+\alpha}{3\beta}} = v^* = \frac{1+\eta_{\alpha,\beta}^*}{2\eta_{\alpha,\beta}^*}\sqrt{\frac{1+\alpha\eta_{\alpha,\beta}^*}{\beta\eta_{\alpha,\beta}^*}}.$$

Let $\alpha = 1/2$, $\beta = 1$, for example; we have

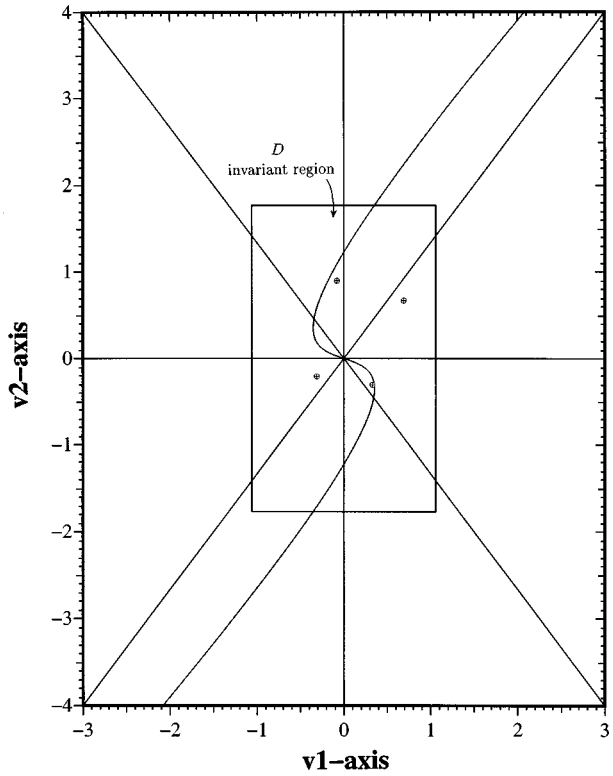

FIG. 1. An attracting period-4 orbit for Example III.1, where $\alpha=1/2$, $\beta=1$, and $\eta=2/3$ are used. The square is an invariant region guaranteed by Proposition III.4.

$$v_c^*\approx 0.3536,\ v^*\approx 1.7678 \quad \text{and} \quad \eta^*_{\alpha,\beta}=2/3\approx 0.6667,$$

$$v_1^*\approx 1.0607, \quad v_2^*=v^*\approx 1.7678.$$

For these values of α, β, and $\eta^*_{\alpha,\beta}$, we use the computer to plot the orbit $(\mathcal{R}_0\mathcal{R}_1)^k P_0$, with $P_0=(0.005,0.005)$ and $10^4\leqslant k\leqslant 10^5$. What we obtain is just an attracting period-4 orbit as shown in Fig. 1. There is another attracting period-4 orbit which is symmetric with respect to the origin of the one shown in Fig. 1 but not plotted there. □

The following can be verified in a straightforward way; its proof is omitted.

Proposition III.5: For given $\alpha, 0<\alpha\leqslant 1$, $\beta>0$, and $\eta>0, \eta\neq 2$, *the map* $\mathcal{R}_0\mathcal{R}_1$ *does not have any prime period-2 points.* □

The following theorem gives two unbounded invariant domains whereupon all trajectories grow unbounded.

Theorem III.1: Let $0<\alpha\leqslant 1$, $\beta>0$, *and* $0<\eta<1$. *Then the following two sets,*

$$U_1=\{(v_1,v_2)\in\mathbb{R}^2 | v_1\geqslant v_1^*, v_2\geqslant v_2^*\},$$

$$U_2=-U_1=\{(v_1,v_2)\in\mathbb{R}^2 | (-v_1,-v_2)\in U_1\},$$

are invariant under $\mathcal{R}_0\mathcal{R}_1$. *Furthermore, for each point* $(v_1,v_2)\in U_1\backslash\{(v_1^*,v_2^*)\}$ *we have*

$$\lim_{n\to\infty}|(\mathcal{R}_0\mathcal{R}_1)^n(v_1,v_2)|_2=\infty, \tag{61}$$

where $||_2$ is the Euclidean norm in $\mathbb{R}^2$. Same for U_2.

Proof: Let $(v_1,v_2)\in U_1$ and $(u_1,u_2)\in\mathcal{R}_0\mathcal{R}_1(v_1,v_2)$. Then

$$
\begin{aligned}
u_1 &= \frac{\eta}{2-\eta} v_1 + \frac{2}{2-\eta} F(v_2) \\
&\geqslant \frac{\eta}{2-\eta} v_1^* + \frac{2}{2-\eta} F(v_2^*) \\
&= \frac{\eta}{2-\eta} v_1^* + \frac{2}{2-\eta} \cdot \frac{1-\eta}{1+\eta} v_2^* \quad \text{[by (25),(55)]} \\
&= \left[\frac{\eta}{2-\eta} + \frac{2}{2-\eta} \cdot \frac{1-\eta}{1+\eta} \cdot (1+\eta) \right] v_1^* = v_1^*, \quad \text{[by (55)];}
\end{aligned}
$$

$$
\begin{aligned}
u_2 &= \frac{2}{2-\eta} v_1 + \frac{\eta}{2-\eta} F(v_2) \\
&\geqslant \frac{2}{2-\eta} v_1^* + \frac{\eta}{2-\eta} F(v_2^*) \\
&= \left[\frac{2}{2-\eta} \cdot \frac{1}{1+\eta} + \frac{\eta}{2-\eta} \frac{1-\eta}{1+\eta} \right] v_2^* = v_2^* \quad \text{[by (25),(55)].}
\end{aligned}
$$

Therefore U_1 is invariant under $\mathcal{R}_0\mathcal{R}_1$. The invariance of U_2 follows immediately from the oddness of the map $\mathcal{R}_0\mathcal{R}_1$.

For any initial state $(v_1^{(0)}, v_2^{(0)}) \in U_1$, let $(v_1^{(n)}, v_2^{(n)}) = (\mathcal{R}_0\mathcal{R}_1)^n (v_1^{(0)}, v_2^{(0)})$ be the nth iterate. Define the following Liapunov function on U_1:

$$
W(v_1, v_2) = \delta v_1 + v_2, \quad \text{for some } \delta > 0.
$$

Then

$$
\begin{aligned}
\Delta_1 W &\equiv W(v_1^{(1)}, v_2^{(1)}) - W(v_1^{(0)}, v_2^{(0)}) \\
&= (\delta v_1^{(1)} + v_2^{(1)}) - (\delta v_1^{(0)} + v_2^{(0)}) \\
&= \delta \left[\frac{\eta}{2-\eta} v_1^{(0)} + \frac{2}{2-\eta} F(v_2^{(0)}) \right] \\
&\quad + \left[\frac{2}{2-\eta} v_1^{(0)} + \frac{\eta}{2-\eta} F(v_2^{(0)}) \right] - (\delta v_1^{(0)} + v_2^{(0)}) \\
&= \left(\frac{\delta\eta + 2}{2-\eta} - \delta \right) v_1 + \left(\frac{2\delta + \eta}{2-\eta} F(v_2) - v_2 \right). \qquad (62)
\end{aligned}
$$

Now choose $\delta = 1/(1-\eta)$. Then $(\delta\eta + 2)/(2-\eta) - \delta = 0$, and

$$
\frac{2\delta + \eta}{2-\eta} = \frac{1+\eta}{1-\eta}.
$$

Continuing from (62), we get

$$
\Delta_1 W = \frac{1+\eta}{1-\eta} F(v_2^{(0)}) - v_2^{(0)} \geqslant 0, \quad \text{for } v_2 \geqslant v_2^*. \qquad (63)
$$

Note that equality in (63) holds when and only when $v_2 = v_2^*$. However, if $v_1^{(0)} > v_1^*$, $v_2^{(0)} = v_2^*$, then

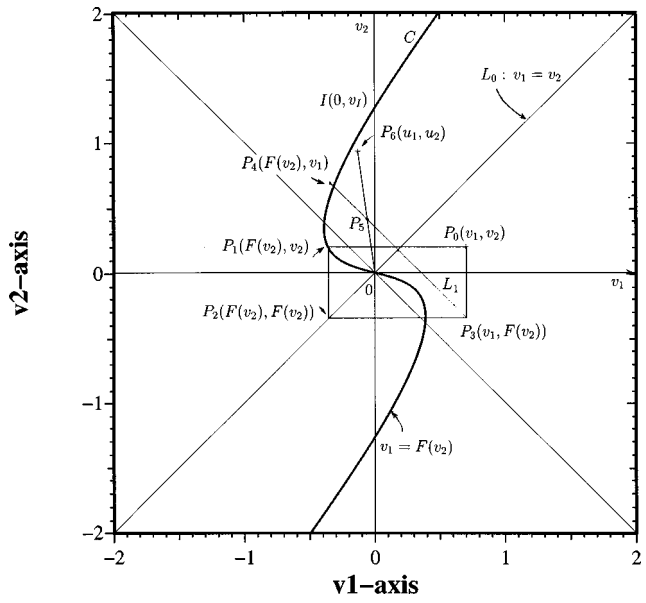

FIG. 2. Graphical construction of $(u_1,u_2)=\mathcal{R}_0\mathcal{R}_1(v_1,v_2)$.

$$v_2^{(1)}=\frac{2}{2-\eta}v_1^{(0)}+\frac{\eta}{2-\eta}F(v_2^*)>\frac{\eta}{2-\eta}v_1^*+\frac{\eta}{2-\eta}\frac{1-\eta}{1+\eta}v_2^*$$
$$=\left(\frac{2}{2-\eta}\frac{1}{1+\eta}+\frac{\eta}{2-\eta}\frac{1-\eta}{1+\eta}\right)v_2^*=v_2^*.$$

Therefore $\Delta_1 W>0$ for $(v_1^{(0)},v_2^{(0)})\in U_1\setminus\{(v_1^*,v_2^*)\}$. A little more elaborate argument further shows that $W(v_1^{(n)},v_2^{(n)})\to\infty$ as $n\to\infty$. Details are omitted. We therefore have (61). □

Remark III.1: By Theorem III.1, clearly chaos will not occur in U_1 and U_2 because trajectories originating or passing therein grow unbounded. This information is rather useful in eliminating nonchaotic cases.

Then, will chaos occur in the bounded invariant rectangle D in (53) for η in the range $0<\eta<\eta^*_{\alpha,\beta}$? Through numerical experiments, we have found that the answer is also negative; the map $\mathcal{R}_0\mathcal{R}_1$ displays only periodic behavior with periods 2^n, i.e., for the parameter range $\eta, 0<\eta<\eta^*_{\alpha,\beta}$, all orbits are asymptotically periodic; see Example III.1. What we have found empirically is that strange attractors (invariant regions) for chaotic $\mathcal{R}_0\mathcal{R}_1$ consist mainly of points in D, plus a small portion of points lying *outside* $D\cup U_1\cup U_2$ (see Figs. 3 and 14, for example).

Also, even though in Proposition III.4 and Theorem III.1 the range of η is restricted to $0<\eta<1$ (rather than $\eta>0$, $\eta\neq 2$), numerical experiments indicate that once η approaches $1-$, then trajectories become unbounded. Thus, it appears that no chaotic cases are lost if we restrict η, $0<\eta<1$.

Graphical methods, such as Lienard's (Ref. 10, pp. 31–33), have been found effective and useful in studying systems of nonlinear differential equations in 2D. To conclude this section, we describe a graphical method for the map $\mathcal{R}_0\mathcal{R}_1$. For any given point (v_1,v_2), we rewrite the relation (56) as

$$\begin{bmatrix}u_1\\u_2\end{bmatrix}=\frac{2+\eta}{2-\eta}\left\{\frac{\eta}{2+\eta}\begin{bmatrix}v_1\\F(v_2)\end{bmatrix}+\frac{2}{2+\eta}\begin{bmatrix}F(v_2)\\v_1\end{bmatrix}\right\}. \tag{64}$$

Note that in (64), the terms inside the curly parentheses are a convex combination of two points $(v_1,F(v_2))$ and $(F(v_2),v_1)$. Thus such a sum of terms corresponds to a point on the line segment with endpoints $(v_1,F(v_2))$ and $(v_2,F(v_1))$. Using (64), we show the graphical construction in Fig. 2 where C is the curve representing the function $v_1=F_{\alpha,\beta}(v_2)$, I is the point $(0,v_I)$, L_0 is the straight line $v_1=v_2$, P_0 is the point (v_1,v_2), P_1 is the point $(F(v_2),v_2)$, P_2 is the point

$(F(v_2),F(v_2))$, P_3 is the point $(v_1,F(v_2))$, P_4 is the point $(F(v_2),v_1)$, L_1 is the line segment joining P_3 with P_4, P_5 is a point on L_1 satisfying $P_5P_3{:}P_5P_4=\eta{:}2$, and P_6 is the image point (u_1,u_2) satisfying $OP_6{:}OP_5=(2+\eta)/(2-\eta)$.

IV. SNAPBACK REPELLERS AS A CAUSE OF CHAOS

Difficulties associated with proving chaos for noninvertible maps in two- or higher-dimensional spaces have been described itemwise in Secs. I A, B and C. As promised, we now focus our study on Sec. I B, the snapback repellers.

Snapback repellers for (noninvertible) one-dimensional maps are repelling fixed points with homoclinic orbits. The proof that they cause chaos may be found in Devaney (Ref. 11, Theorem 1.16.5, p. 124). We have found that that theorem can be properly generalized to N-dimensional noninvertible maps, and that for our 2D model problems under study here, such snapback repellers actually exist and, thus, they cause chaos. However, we are alerted by Ref. 9, p. 63, that a reference citation for snapback repellers by Marotto[9] already exists in the literature. It turns out that, however, by comparing Marotto's proof with ours, we are surprised that there is an error in Ref. 9. The error is not grievous; at his writing Marotto might think that it was just a simple fact (by changing to an equivalent norm to make things work) not worth mentioning at all. We nevertheless feel that the situation is confusing at least, leading us to clarify it below. We will then present a somewhat refined theorem in the spirit of Ref. 11, Theorem 1.16.5, *loc. cit.* (Even though the authors seem to have independently rediscovered such snapback repellers here, the credit of priority clearly goes to Marotto.[9])

First, recall the definition of a snapback repeller based on Marotto (Ref. 9, Def. 2.3, p. 203).

Definition IV.1: Let $\mathcal{F}{:}\mathbb{R}^N\to\mathbb{R}^N$ be C^1. Let Z be a fixed point of $\mathcal{F}$ such that all of the eigenvalues of $D\mathcal{F}(Z)$ have absolute values larger than 1. We say that Z is a snapback repeller if there exists a point X_0 in $W^u_{\rm loc}(Z)$, the local unstable set of Z, and some integer M, such that $\mathcal{F}^M(X_0)=Z$ and $\det D\mathcal{F}^M(X_0)\neq 0$. □

Now, let us restate the main theorem in Ref. 9.

Theorem IV.1: *[Marotto (Ref. 9, Theorem 3.1, pp. 204–205)] Let $\mathcal{F}{:}\mathbb{R}^N\to\mathbb{R}^N$ be C^1. If $\mathcal{F}$ possesses a snapback repeller, then $\mathcal{F}$ is chaotic. That is, there exists*

(i) a positive integer n such that for each integer $p\geqslant n$, $\mathcal{F}$ has a point of period p;

(ii) a ''scrambled set'' of $\mathcal{F}$, i.e., an uncountable set S containing no periodic points of $\mathcal{F}$ such that:

(a) $\mathcal{F}(S)\subset S$,

(b) for every $X,Y\in S$ with $X\neq Y$,

$$\overline{\lim_{k\to\infty}}|\mathcal{F}^k(X)-\mathcal{F}^k(Y)|>0,$$

(c) for every $X\in S$ and any periodic point Y of $\mathcal{F}$,

$$\overline{\lim_{k\to\infty}}|\mathcal{F}^k(X)-\mathcal{F}^k(Y)|>0;$$

(iii) an uncountable subset S_0 of S such that for every $X,Y\in S_0$,

$$\overline{\lim_{k\to\infty}}|\mathcal{F}^k(X)-\mathcal{F}^k(Y)|=0.$$

□

The statements in Theorem IV.1 are all correct without question. However, an error was made in the proof: in Ref. 9, p. 202, line 34, Marotto first states that in $\mathbb{R}^N$ that the usual Euclidean norm is to be used. Let us write it as

$$|x|_2=\left(\sum_{i=1}^N x_i^2\right)^{1/2},\quad \text{for } x=(x_1,\ldots,x_N)\in\mathbb{R}^N,$$

where the subscript 2 denotes the ℓ_2-norm. Then in Ref. 9, p. 203, lines 15–17, he writes:

''If all eigenvalues of $DF(Z)$ are greater than 1 in norm, then F displays the following local behavior at Z. For some $s>1$ and $r>0$:

$$\|F(X)-F(Y)\|>s\|X-Y\| \quad \text{for all } X,Y\in B_r(Z).'' \tag{65}$$

($B_r(Z)$ is the ball with radius r centered at Z.)

There, Marotto obviously was using the following property:

''Let A be an $N\times N$ real constant matrix such that all of its eigenvalues are larger than 1 in absolute value. Then A satisfies

$$|Ax|\geqslant\mu|x|, \quad \text{for some } \mu>1, \quad \text{for all } x\in\mathbb{R}^N.'' \tag{66}$$

Since in (65), the norm of Marotto's choice for the underlying space is the ℓ_2-norm, he must be interpreted to mean that in (66), there is a $\mu>1$ such that

$$|Ax|_2\geqslant\mu|x|_2, \quad \text{for all } x\in\mathbb{R}^N. \tag{67}$$

But it is well known that (67) is false. The following is a simple counterexample.

Example IV.1:Let ε and δ be positive. The matrix

$$A=\begin{bmatrix} 1+\epsilon+\dfrac{\delta}{2} & \dfrac{1}{2\delta} \\ \dfrac{\delta^3}{2} & 1+\epsilon+\dfrac{\delta}{2} \end{bmatrix}$$

has two eigenvalues $1+\epsilon$ and $1+\epsilon+\delta/2$, with respective corresponding eigenvectors

$$\begin{bmatrix} 1 \\ -\delta^2 \end{bmatrix}, \begin{bmatrix} 1 \\ \delta^2 \end{bmatrix}.$$

Choose a unit vector

$$x=\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Then

$$Ax=\frac{1}{\sqrt{2}}\begin{bmatrix} 1+\epsilon+\dfrac{\delta}{2}-\dfrac{1}{2\delta} \\ \dfrac{\delta^3}{2}-1-\epsilon-\dfrac{\delta}{2} \end{bmatrix}.$$

Take $\delta=1/2$. Then

$$|Ax|_2=\frac{1}{\sqrt{2}}\left[\left(\epsilon+\frac{1}{4}\right)^2+\left(1+\epsilon+\frac{3}{16}\right)^2\right]^{1/2}=\frac{1}{\sqrt{2}}\left[\frac{377}{256}+\epsilon\left(\frac{27}{8}+2\epsilon\right)\right]^{1/2}<0.9<1,$$

if ϵ is small. □

All norms in $\mathbb{R}^N$ or $\mathbb{C}^N$ are equivalent. What we need is a certain norm in $\mathbb{R}^N$ or $\mathbb{C}^N$ making (66) valid. We show, step by step, how to do this in the following.

Lemma IV 1: Let A be an invertible $N\times N$ real (resp. complex) constant matrix and let $|\ |$ be a norm in $\mathbb{R}^N$ (resp. $\mathbb{C}^N$). Then there exists a $\mu>0$ such that $|Ax|\geqslant\mu|x|$ holds for all $x\in\mathbb{R}^N$ (resp. $\mathbb{C}^N$) if and only if $\|A^{-1}\|\leqslant\mu^{-1}$, where $\|A^{-1}\|$ is the operator norm of A^{-1} corresponding to the norm $|\ |$ of the underlying $\mathbb{R}^N$ (resp. $\mathbb{C}^N$).

Proof: For any $x\in\mathbb{R}^N$ or $\mathbb{C}^N$, $x\neq 0$, we have

$$\mu^{-1}=\|A^{-1}\|=\sup_{y\neq 0}\frac{|A^{-1}y|}{|y|}\geqslant\frac{|A^{-1}w|}{|w|}=\frac{|x|}{|Ax|}\quad(x=A^{-1}w).$$

Therefore $|Ax|\geqslant\mu|x|$ for all $x\in\mathbb{R}^N$ or $\mathbb{C}^N$, $x\neq 0$, if and only if $\|A^{-1}\|\leqslant\mu^{-1}$. When $x=0$, the proof is trivial. □

Lemma IV.2: Let $D=\mathrm{diag}(\lambda_1,\ldots,\lambda_N)$ *be an* $N\times N$ *diagonal matrix with diagonal entries* $\lambda_1,\ldots,\lambda_N\in\mathbb{C}^N$. *Let* $\mathbb{C}^N$ *be the N-dimensional complex Euclidean space with* ℓ^2*-norm*

$$|z|_2=\left(\sum_{j=1}^{N}|z_j|^2\right)^{1/2},\quad z=(z_1,\ldots,z_N)\in\mathbb{C}^N.$$

Then $\|D\|_2$, *the operator norm of D corresponding to* $|\ |_2$, *satisfies*

$$\|D\|_2=\max_{1\leqslant j\leqslant N}|\lambda_j|.$$

Proof: Just expand any vector $z\in\mathbb{C}^N$ in terms of the basis eigenvectors $e_1,e_2,\ldots,e_n$, where e_j is the unit vector pointing in the jth axial direction. □

Lemma IV.3: Let J be an $m\times m$ *Jordan matrix with eigenvalue* λ *of Riesz index m:*

$$J=\begin{bmatrix}\lambda & 1 & 0 & \ldots & 0\\ 0 & \lambda & 1 & & \vdots\\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1\\ 0 & \ldots & 0 & & \lambda\end{bmatrix}.\tag{68}$$

Then for any given ε, *J has a similarity matrix* $\tilde{J}=D^{-1}JD$, *where D is an invertible diagonal matrix, such that*

$$\tilde{J}=\begin{bmatrix}\lambda & \varepsilon & & & 0\\ & \lambda & \varepsilon & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon\\ 0 & & & & \lambda.\end{bmatrix}.\tag{69}$$

Proof: This can be found in Devaney (Ref. 11, Proposition 2.1.12, p. 168) or Franklin (Ref. 12, Exercise 13, p. 174). Its proof can be easily illustrated through the following 3×3 similarity matrix:

$$D^{-1}JD=\begin{bmatrix}a^{-1} & 0 & 0\\ 0 & a_2^{-1} & 0\\ 0 & 0 & a_3^{-1}\end{bmatrix}\begin{bmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{bmatrix}\begin{bmatrix}a_1 & 0 & 0\\ 0 & a_2 & 0\\ 0 & 0 & a_3\end{bmatrix}=\begin{bmatrix}\lambda & a_1^{-1}a_2 & 0\\ 0 & \lambda & a_2^{-1}a_3\\ 0 & 0 & \lambda\end{bmatrix}.$$

Therefore, if we choose a_1,a_2, and a_3, the diagonal entries of D, to satisfy $a_1^{-1}a_2=a_2^{-1}a_3=\varepsilon$, we obtain the desired form (69). □

Theorem IV.2: Let A be an $N\times N$ *(real or complex) constant matrix such that all of its eigenvalues have absolute values larger than* μ. *Then there exists a norm* $|\cdot|$ *in* $\mathbb{C}^N$ *such that the associated operator-norm of* A^{-1}, $\|A^{-1}\|$, *satisfies* $\|A^{-1}\|\leqslant\mu^{-1}$. *Consequently, A satisfies* $|Ax|\geqslant\mu|x|$ *for all* $x\in\mathbb{R}^N$ *or* $\mathbb{C}^N$.

Proof: We first perform a similarity transformation to obtain $A^{-1}=PJP^{-1}$, where J is a Jordan canonical form of A^{-1}; J consists of submatrices J_k diagonally, for $k=1,2,\ldots,l$, where each J_k has the form (68). Applying Lemma IV.3, we may now assume that each J_k has the form

(69). Therefore, J is an ε-perturbation of a diagonal matrix; call that diagonal matrix D. The diagonal entries of D consist of λ_j^{-1}, the multiplicative inverses of the eigenvalues λ_j of A. By Lemma IV.2 and the assumption that $|\lambda_j|>\mu$ for all λ_j, we have

$$\|D\|_2=\max_j|\lambda_j^{-1}|<\mu^{-1}.$$

Since J is an ε-perturbation of D, for ε sufficiently small we obtain

$$\|J\|_2\leqslant\|D\|_2+\varepsilon'\leqslant\mu^{-1}, \tag{70}$$

because $\varepsilon'>0$ depends only on ε and can thus be made as small as we wish.

In $\mathbb{C}^N$, we now define a P-norm by

$$|z|_P=|P^{-1}z|_2.$$

Then, for a $z\in\mathbb{C}^N$,

$$|A^{-1}z|_P=|P^{-1}A^{-1}z|_2=|JP^{-1}z|_2\leqslant\|J\|_2|P^{-1}z|_2\leqslant\mu^{-1}|z|_P, \text{ by (70).}$$

Therefore

$$\|A^{-1}\|_P=\sup\frac{|A^{-1}z|_P}{|z|_P}\leqslant\mu^{-1}.$$

By Lemma IV.1, we have

$$|Ax|_P\geqslant\mu|x|_P, \quad \text{for all } x\in\mathbb{R}^N \quad \text{or } \mathbb{C}^N.$$

By incorporating Theorem IV.2 into its proof, Theorem IV.1 is now true. However, in what follows, let us provide a somewhat refined version of Theorem IV.1. Let

$$\mathcal{F}:\mathbb{R}^N\to\mathbb{R}^N, \quad \mathcal{F}(x)=(f_1(x),f_2(x),\ldots,f_N(x)), \quad x\in\mathbb{R}^N,$$

be a C^1 map. Then for $x^{(1)},x^{(2)}\in\mathbb{R}^N$, by applying the mean value theorem to the scalar-valued functions $f_1,\ldots,f_N$, we have

$$\mathcal{F}(x^{(1)})-\mathcal{F}(x^{(2)})=(\nabla f_1(y^1)\cdot(x^{(1)}-x^{(2)}),\nabla f_2(y^2)\cdot(x^{(1)}-x^{(2)}),\ldots,\nabla f_N(y^N)\cdot(x^{(1)}-x^{(2)})), \tag{71}$$

for some $y^1,\ldots,y^N\in\mathbb{R}^N$, where each y^j lies on the open line segment $L=\{y\in\mathbb{R}^N|y=\alpha x^{(1)}+(1-\alpha)x^{(2)},\alpha\in(0,1)\}$. We write (71) as

$$\mathcal{F}(x^{(1)})-\mathcal{F}(x^{(2)})=\mathcal{D}F\cdot[x^{(1)}-x^{(2)}], \tag{72}$$

where $\mathcal{DF}=\mathcal{DF}(Y)$, with $Y=(y^1,\ldots,y^n)\in\mathbb{R}^{N\times N}$, is an $N\times N$ matrix whose ith row vector is $\nabla f_i(y^i)$, where $\nabla f_i(y^i)$ makes (71) satisfied.

Theorem IV.3: Let $\mathcal{F}:\mathbb{R}^N\to\mathbb{R}^N$ be C^1, and p be a snapback repeller of $\mathcal{F}$. Then for each neighborhood U of p, there is an integer $n>0$ such that $\mathcal{F}^n$ has a hyperbolic invariant subset in U on which $\mathcal{F}^n$ is topologically conjugate to the shift map on the binary symbol space Σ_2.

Proof: The argument goes the same way as that in Ref. 11, Theorem 1.16.5, p. 124. Since p is a snapback repeller, all of the eigenvalues of $D\mathcal{F}(p)$ have absolute values larger than 1. By Theorem IV.2, we can define a norm $|\cdot|$ on $\mathbb{R}^N$ such that $\|[D\mathcal{F}(p)]^{-1}\|\leqslant\mu^{-1}$, for some $\mu>1$. By continuity, we can find an open neighborhood W of p in U and some $\tilde{\mu},1<\tilde{\mu}<\mu$, such that

$$\|[\mathcal{DF}(Y)]^{-1}\|\leqslant\tilde{\mu}^{-1}, \quad \text{for all } Y=(y^1,\ldots,y^N)\in W^N. \tag{73}$$

By the assumption of a snapback repeller, we can take a point $q\in W$ such that $\mathcal{F}^n(q)=p$ and $\det(D(\mathcal{F}^n)(q))\neq0$. Since the orbit of q contains only a finite number of points that do not lie in

W, and since $\det(D\mathcal{F}^n(q))\neq 0$, by continuity we can find a neighborhood V of q in W such that $\mathcal{DF}^n(Y)$ is invertible for all $Y=(y^1,\ldots,y^N)\in V^N$. Therefore there exists some $\varepsilon>0$ such that

$$\|[\mathcal{DF}^n(Y)]^{-1}\|<\varepsilon^{-1}, \quad \text{for all } Y=(y^1,\ldots,y^N)\in V^N. \tag{74}$$

From the invertibility of $\mathcal{DF}^n(Y)$, we see that $\mathcal{F}^n$ maps V diffeomorphically onto an open set $\mathcal{F}^n(V)$ containing p in its interior.

We now choose j so that $(\tilde{\mu}^j\varepsilon)^{-1}<1$. We choose V sufficiently small so that $\mathcal{F}^{n+i}(V)\subset W$, but $\mathcal{F}^{n+i}(V)\cap V=\varnothing$, for $i=1,2,\ldots,j$. Since $V\subset W$, from (72), (73), and Lemma IV.1, for $k>j$ and for any two points $x^1,x^2\in V$, we have

$$\begin{aligned}
|\mathcal{F}^{n+k}(x^1)-\mathcal{F}^{n+k}(x^2)| &= |\mathcal{F}^{n+k}(x^1)-\mathcal{F}^{n+k}(x^2)| \\
&= |\mathcal{DF}(Y^1)\cdot(\mathcal{F}^{n+k-1}(x^1)-\mathcal{F}^{n+k-1}(x^2))| \\
&\geqslant \tilde{\mu}|\mathcal{F}^{n+k-1}(x^1)-\mathcal{F}^{n+k-1}(x^2)| \quad (\text{where } Y^1\in W^N) \\
&= \tilde{\mu}|\mathcal{DF}(Y^2)\cdot(\mathcal{F}^{n+k-2}(x^1)-\mathcal{F}^{n+k-2}(x^2))| \quad (\text{where } Y^2\in W^N) \\
&\geqslant \tilde{\mu}^2|\mathcal{F}^{n+k-2}(x^1)-\mathcal{F}^{n+k-2}(x^2)| \\
&\geqslant \cdots \\
&\geqslant \tilde{\mu}^k|\mathcal{F}^n(x^1)-\mathcal{F}^n(x^2)| = \tilde{\mu}^k|\mathcal{DF}^n(Y^k)\cdot(x^1-x^2)| \quad (\text{where } Y^k\in V^N) \\
&\geqslant \tilde{\mu}^k\varepsilon|x^1-x^2| \geqslant \tilde{\mu}^{k-j}|x^1-x^2|.
\end{aligned}$$

Therefore $\mathcal{F}^{n+k}$ expands V and for k sufficiently large, $\mathcal{F}^{n+k}(V)$ covers V. We have a diffeomorphism $\mathcal{F}^{n+k}:V\to\mathcal{F}^{n+k}(V)$ such that $p\in\mathcal{F}^{n+k}(V)$ and $V\subset\mathcal{F}^{n+k}(V)$.

Let us choose a sufficiently small open subset V' containing $(\mathcal{F}^{n+k})^{-1}(V)$. By choosing V sufficiently small and k large, we can make V' satisfy

(i) $V\cap V'=\varnothing$;
(ii) $\mathcal{F}^{n+k}(V)\supset V\cup V'$;
(iii) $\mathcal{F}^{n+k}(V')\supset V$;
(iv) $\mathcal{F}^{n+k}(V')\subset W$;
(v) $\mathcal{F}^{n+k}$ is a strict expansion on V'.

Therefore the map $\mathcal{F}^{n+k}$ has a shift sequence

$$V'\to V\to V'\cup V,$$

where each member (except the leader) of the sequence is covered by the image of the predecessor. Therefore $\mathcal{F}^{n+k}(V'\cup V)\supset V'\cup V$, and $V'\cup V$ has an invariant subset

$$\Lambda\equiv\{x\in V'\cup V|(\mathcal{F}^{n+k})^j(x)\in V'\cup V, j=0,1,2,\ldots\}.$$

For any $x\in\Lambda$ we define its binary (itinerary) symbol as

$$s(x)=(s_0s_1s_2\ldots s_j\ldots), \quad s_j=0 \text{ if } (\mathcal{F}^{n+k})^jx\in V', \quad s_j=1 \text{ if } (\mathcal{F}^{n+k})^jx\in V.$$

Then it is a standard procedure to show that $\mathcal{F}^{n+k}$ is topologically conjugate to the shift automorphism on Σ_2, the space of all binary symbols. The proof is complete. □

Note that all consequences (i), (ii), and (iii) in Marotto's Theorem IV.1 can be deduced from the symbolic dynamics implied by Theorem IV.3.

Remark IV.1: In verifying the snapback repeller assumption in order to be able to apply Theorem IV.3, one needs to check the following conditions according to Definition IV.1:

(i) p is a fixed point of $\mathcal{F}$ such that all eigenvalues of $D\mathcal{F}(p)$ have absolute values larger than 1; (75)
(ii) there exists a $q\in W^u_{\text{loc}}(p)$ such that $\mathcal{F}^M(q)=p$ for some positive integer M; (76)
(iii) $\det D\mathcal{F}^M(q)\neq 0$. (77)

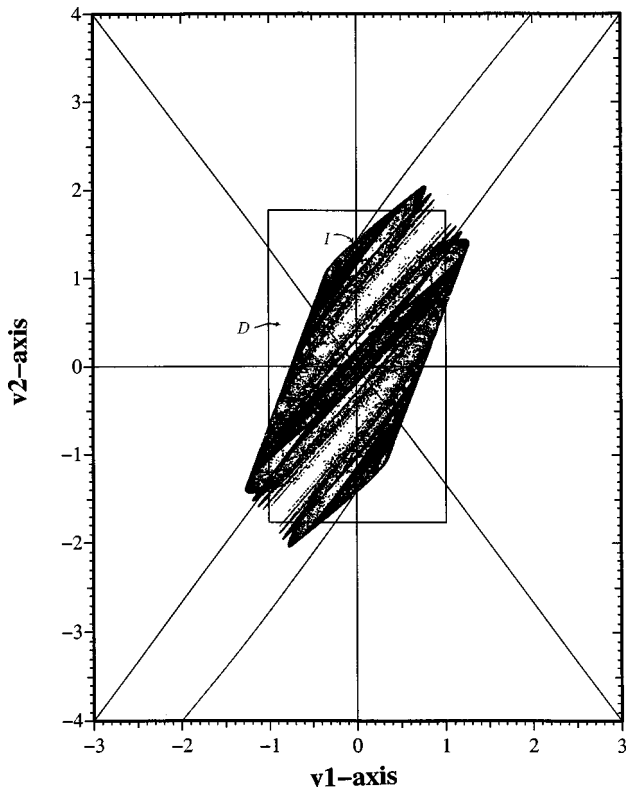

FIG. 3. The strange attractor for Example V.1, where $\alpha=0.9980$, $\beta=1$, and $\eta=0.7570$ are used. D represents the region defined by (54). The origin lies inside the ''cloud'' and, thus, is a snapback repeller.

In most cases, the verification work must be done through the aid of a computer, which automatically involves numerical errors. The verification of (i) is easiest, and the computer result can be trusted with total confidence. To verify (ii), one needs to do computer-aided search of *preimages* of p and then check whether such preimages can wind up in $W^u_{\text{loc}}(p)$. This kind of reverse search can easily accumulate large numerical errors. Therefore, the value of M, even if computable in principle, would *not be entirely trustworthy*. Nevertheless, using the *continuity property of* $\mathcal{F}$, one can still conclude whether such an M (and, consequently, a homoclinic orbit) *exists or not*.

The verification of (iii) is the *most difficult* in general, because as we just pointed out in the above paragraph, it is very difficult to pin down an accurate value of M. Even if M is firmly determined, checking whether $\det D\mathcal{F}^M(q)\neq 0$ is still generally *impossible* because the evaluation of $D\mathcal{F}^M(q)$ usually involves computer roundoff errors. On the one hand, using the holomorphic property of the map $\mathcal{F}$ with respect to the parameter(s), one may claim that there exist at least some parameter values such that (4.13) holds for some snapback homoclinic orbit. On the other hand, we can expect, ''almost beyond reasonable doubt,'' that even if (77) fails, i.e., $\det D\mathcal{F}^M(q)=0$, there should still be chaos. The situation $\det D\mathcal{F}^M(q)=0$ signifies *degeneracy* and leads us to believe that what we have is a *degenerate (snapback) homoclinic orbit*. According to our knowledge of 1D maps, degenerate homoclinic orbits actually causes perhaps the *strongest chaos*.[11] This leads us to pose a conjecture below:

Conjecture: Let $\mathcal{F}:\mathbb{R}^N\to\mathbb{R}^N$ be C^1 satisfying (75) and (76). Then there exists a hyperbolic invariant subset of $W^u_{\text{loc}}(p)$ on which $\mathcal{F}^n$ on Λ is topologically conjugate to the shift map on Σ_2. □

V. NUMERICAL EXAMPLES

We provide two examples in this section. Throughout, we fix $\beta=1$.

Example V.1: The origin as a snapback repeller of $\mathcal{R}_0\mathcal{R}_1$. For α, $0<\alpha\leqslant 1$, we know that the origin (0,0) is a repelling fixed point by Proposition III.2.

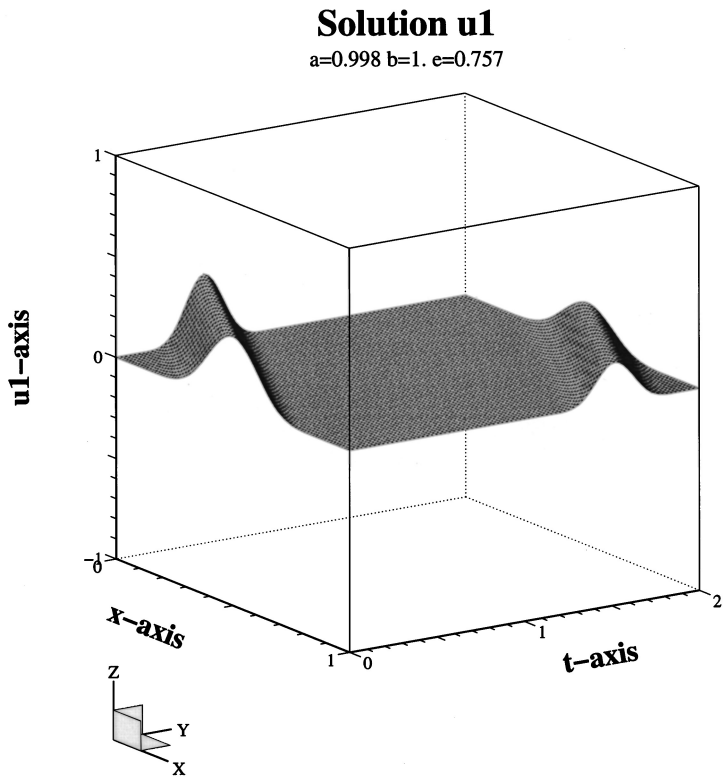

FIG. 4. The spatiotemporal profile of u_1 for initial time duration $[0,\ 2]$.

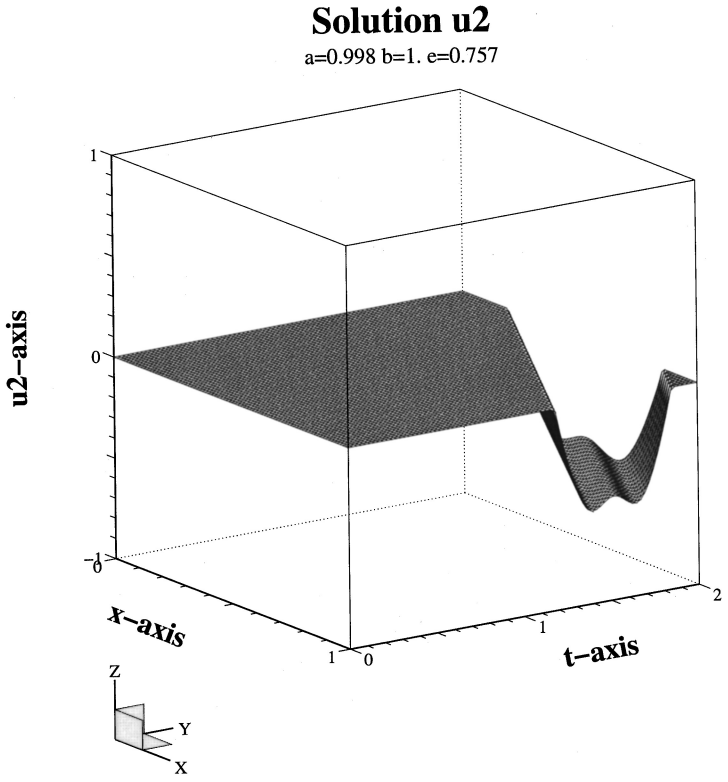

FIG. 5. The spatiotemporal profile of u_2 for initial time duration $[0,\ 2]$.

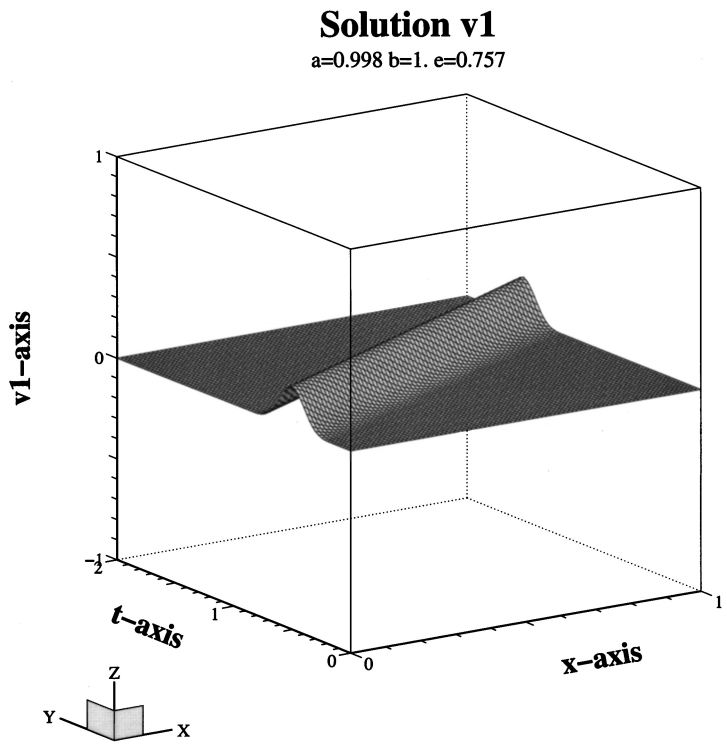

FIG. 6. The spatiotemporal profile of v_1 for initial time duration $[0,\ 2]$.

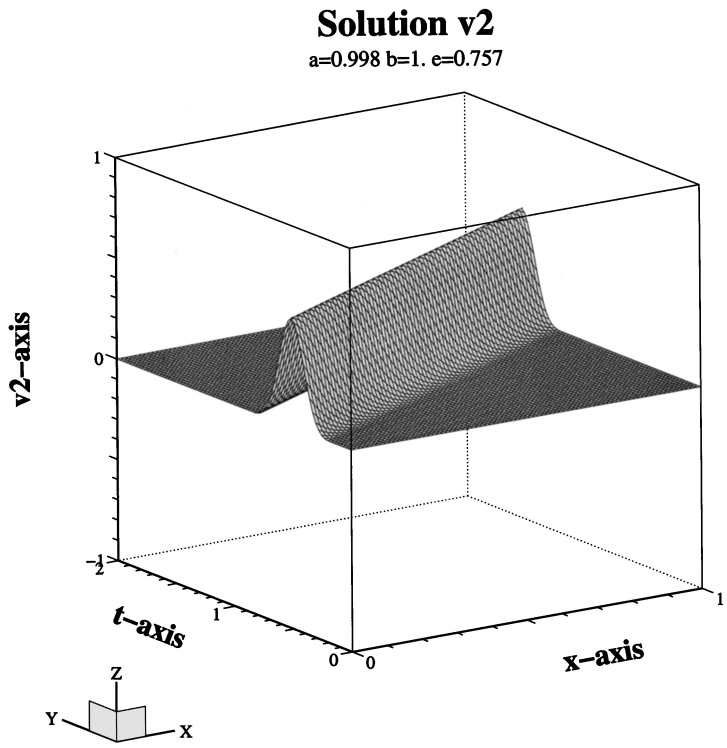

FIG. 7. The spatiotemporal profile of v_2 for initial time duration $[0,\ 2]$.

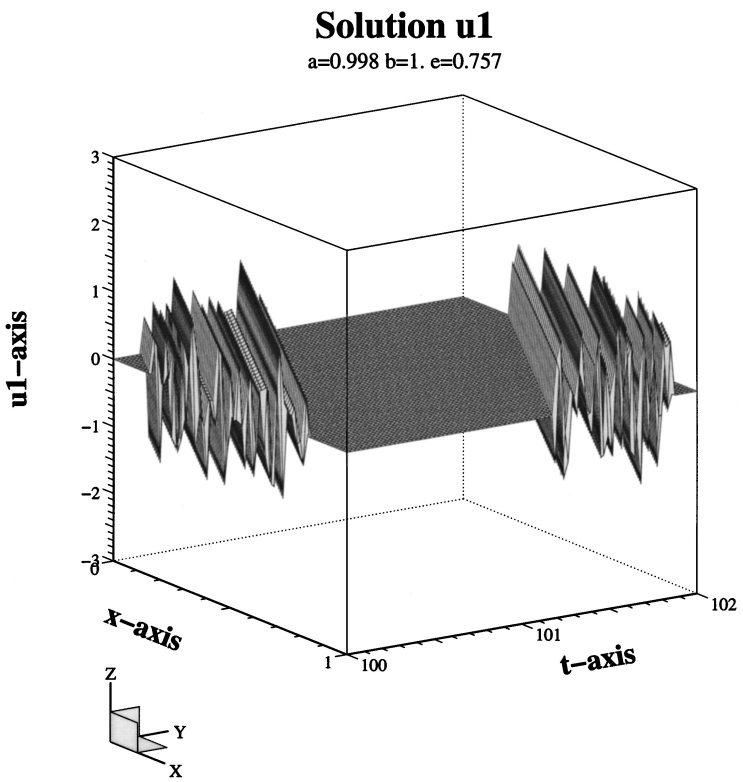

FIG. 8. The spatiotemporal profile of u_1 for time duration [100, 102].

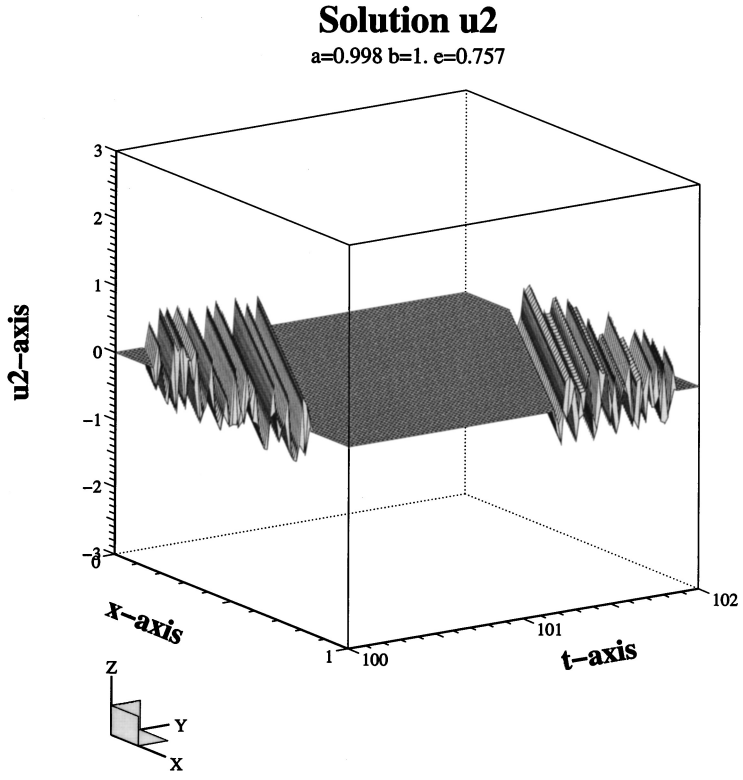

FIG. 9. The spatiotemporal profile of u_2 for time duration [100, 102].

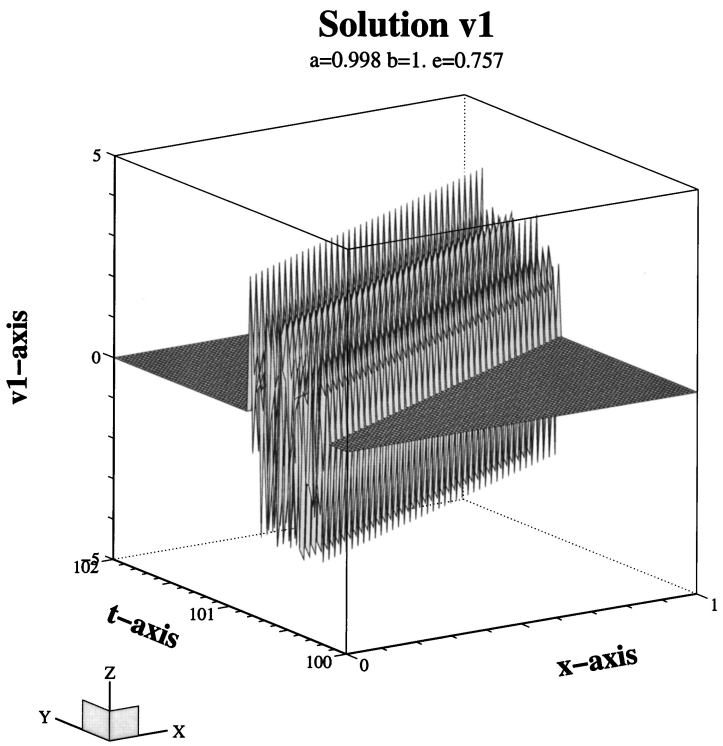

FIG. 10. The spatiotemporal profile of v_1 for time duration [100, 102].

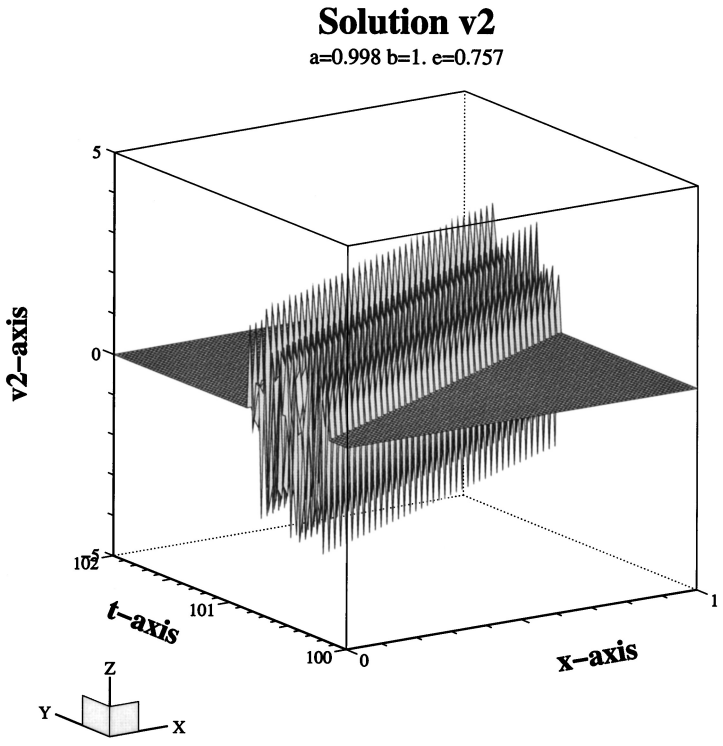

FIG. 11. The spatiotemporal profile of v_2 for time duration [100, 102].

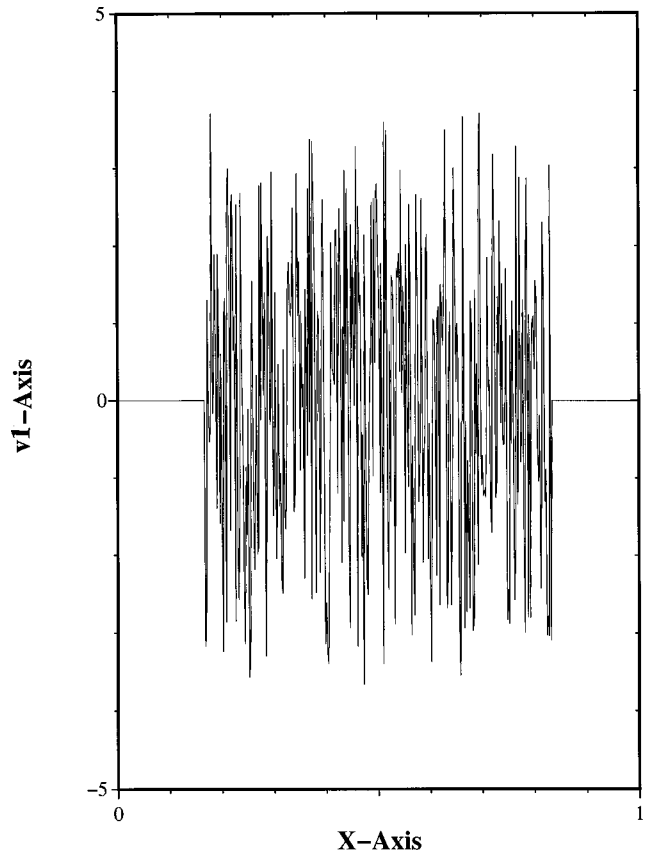

FIG. 12. The snapshot of v_1 at time $t=101$.

Through many numerical experiments, we have found that the origin will not be a snapback repeller unless α is close to 1.

We choose $\alpha=0.9980$, $\beta=1$, and $\eta=0.7570$. To see whether there is a strange attractor, we plot the iterates

$$(\mathcal{R}_0\mathcal{R}_1)^k P_0, \quad P_0=(0.005,0.005), \quad 10^4\leqslant k\leqslant 10^5.$$

What we obtain is a ''cloud,'' the strange attractor, as shown in Fig. 3. Note that the point $I=(0,v_I)$ (cf. Fig. 2) is mapped into the origin:

$$(\mathcal{R}_0\mathcal{R}_1)I=(0,0),$$

and in Fig. 3, I is ''hidden in the cloud.'' Therefore, by Remark IV.1, the continuity of $\mathcal{R}_0\mathcal{R}_1$ implies that the origin is a snapback repeller, i.e., (76) is satisfied. The verification of (77) is difficult, however, as noted in Remark IV.1. We have made no attempts to verify it.

To see chaotic vibration, we consider the system (15)–(18), with the initial data

$$u_{1,0}(x)=\frac{1}{12}\cdot\begin{cases}(x-x_1)^3/h^3, & x_1\leqslant x\leqslant x_2,\\ 1+\dfrac{3(x-x_2)}{h}+\dfrac{3(x-x_2)^2}{h^2}-\dfrac{3(x-x_2)^3}{h^3}, & x_2\leqslant x\leqslant x_3,\\ 1-\dfrac{3(x-x_4)}{h}+\dfrac{3(x-x_4)^2}{h^2}+\dfrac{3(x-x_4)^3}{h^3}, & x_3\leqslant x\leqslant x_4,\\ (x_5-x)^3/h^3, & x_4\leqslant x\leqslant x_5\\ 0, & \text{elsewhere,}\end{cases} \tag{78}$$

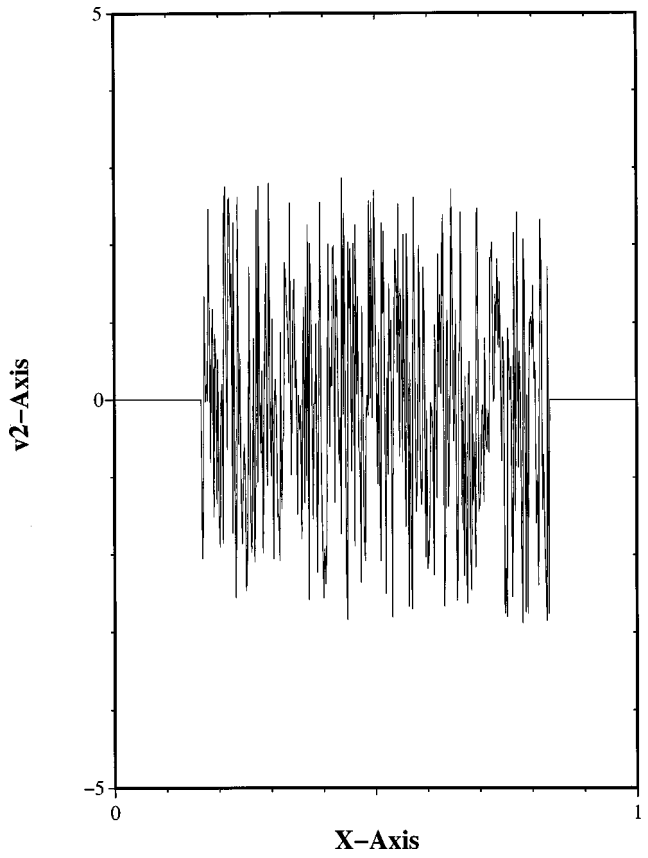

FIG. 13. The snapshot of v_2 at time $t=101$.

$$h=1/6, \quad x_j=j/6, \quad \text{for } j=1,2,3,4,5;$$
$$u_{2,0}(x)=v_{1,0}(x)=v_{2,0}(x)\equiv 0, \quad 0\leqslant x\leqslant 1. \tag{79}$$

Note that the function given in (78) is a C^2-continuous smoothest spline of degree 3. The initial data (78) and (79) satisfy the boundary conditions (16) and (18). It is not difficult to show (by mimicking the proof in Ref. 3, Theorem 6.1 that the system has a unique C^2-solution (u_1,u_2,v_1,v_2) on $(x,t)\in[0,1]\times[0,\infty)$.

The spatiotemporal profiles of u_1,u_2,v_1,v_2 are displayed, respectively, in Figs. 4, 5, 6, and 7 for the first two time units, i.e., $t\in[0,2]$.

For time $t\in[100,102]$, the spatiotemporal profiles of u_1,u_2,v_1,v_2 are displayed, respectively, in Figs. 8, 9, 10, and 11.

The snapshots at $t=101$ of v_1 and v_2 are displayed, respectively, in Figs. 12 and 13. We do not need to display u_1 and u_2 at $t=101$ because u_1 and u_2 are identically zero.

These profiles are computed by using the explicit representation formulas (20). Note that because of the *nondispersive* effects of wave propagation, all components u_1,u_2,v_1, and v_2 of the system can display a ''totally serene'' (i.e., zero or no disturbance) zone right next to the chaotic spatiotemporal region. □

Example V.2: A period-4 point as a snapback repeller of $(\mathcal{R}_0\mathcal{R}_1)^4$. Let us choose $\alpha=0.6$, $\beta=1$, and $\eta=0.8200$.

The map $\mathcal{R}_0\mathcal{R}_1$ has a strange attractor as shown in Fig. 10, suggesting that $\mathcal{R}_0\mathcal{R}_1$ is chaotic. Note in Fig. 14 that neither the point I nor the origin lies inside the strange attractor ''cloud;'' in this case, we can easily rule out the origin as a snapback repeller.

The map $\mathcal{R}_0\mathcal{R}_1$ has many period-4 orbits. So let us compute the fixed points of $(\mathcal{R}_0\mathcal{R}_1)^4$ by Newton's method. We have obtained three sets of such orbits (80), (82), and (84) and other relevant data as given below:

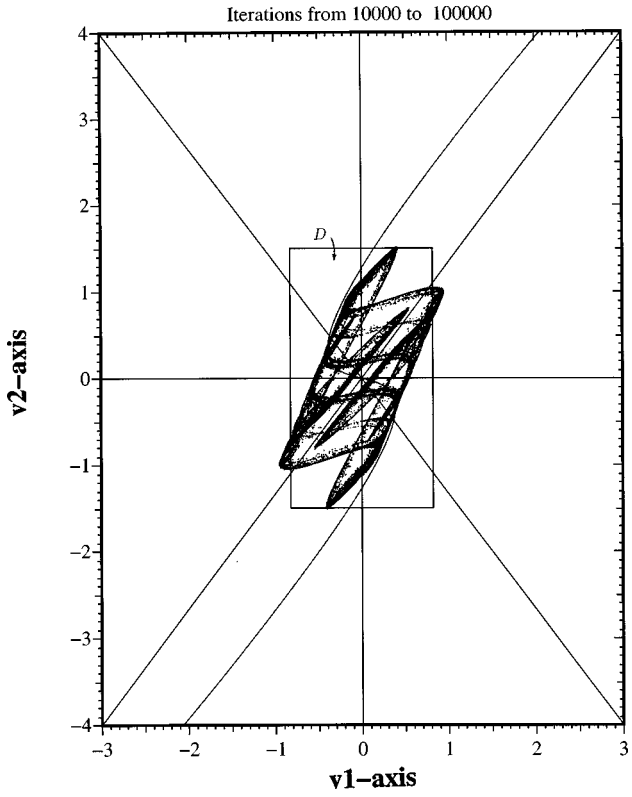

FIG. 14. The strange attractor for Example V.2, where $\alpha=0.6$, $\beta=1$, and $\eta=0.82$ are used. D is the rectangle defined by (54). Note that the origin is not inside the ''cloud'' of the strange attractor.

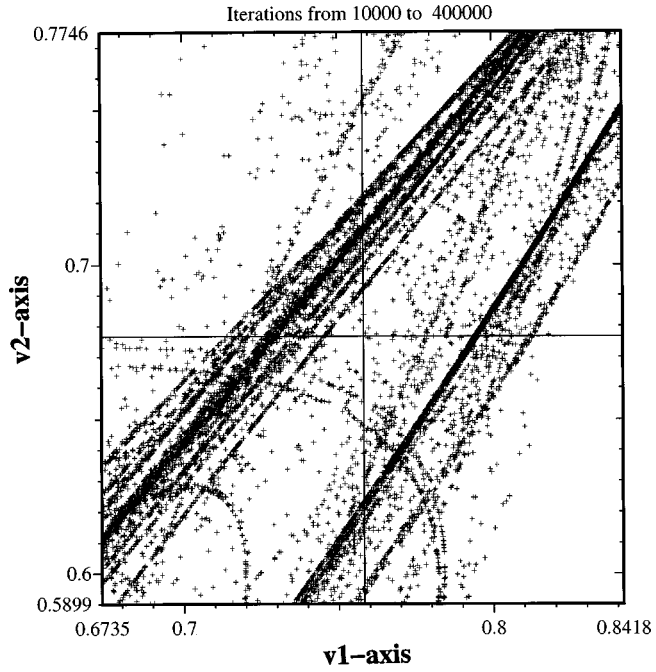

FIG. 15. This is a zoom-in of the vicinity of the period-4 point $P_1^{(1)}=(0.7582,\ 0.6762)$, cf. (80), from Fig. 14. Note that $P_1^{(1)}$is highlighted as the intersection of the horizontal and vertical lines near the center of the figure. Chaos in the immediate vicinity of $P_1^{(1)}$ is not strong.

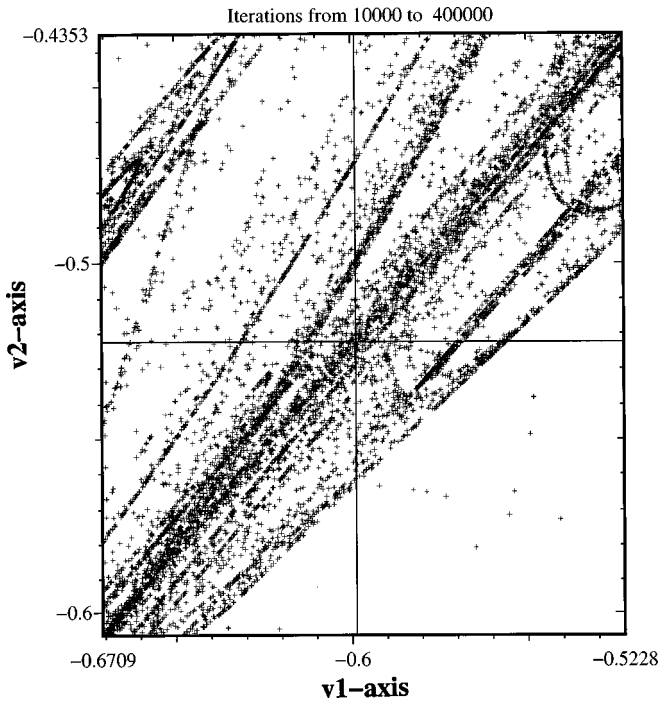

FIG. 16. This is a zoom-in of the vicinity of the period-4 point $P_2^{(1)}=(-0.5988,\ -0.5229)$, cf. (81), from Fig. 14. Note that $P_2^{(1)}$ is highlighted as the intersection of the horizontal and vertical lines near the center of the figure. Chaos is strong in the immediate vicinity of $P_2^{(1)}$.

$$\text{(i)}\quad \begin{cases} P_1^{(1)}=(0.7582,\quad 0.6762), \quad P_1^{(2)}=(2.1507,\quad 1.0699), \\ P_1^{(3)}=(-0.1942,\quad -7.6573), \quad P_1^{(4)}=(0.2591,\quad -0.1675); \end{cases} \tag{80}$$

$$D((\mathcal{R}_0\mathcal{R}_1)^4)(P_1^{(1)})=\begin{bmatrix} -4.4737 & -1.9625 \\ -8.2024 & -5.7953 \end{bmatrix},\ \text{with eigenvalues } -9.2007 \text{ and } -1.0683. \tag{81}$$

$$\text{(ii)}\quad \begin{cases} P_2^{(1)}=(-0.5988,\ -0.5229), \quad P_2^{(2)}=(0.1968,\ -0.7636), \\ P_2^{(3)}=(0.5988,\ 0.5229), \quad P_2^{(4)}=(-0.1968,\ 0.7636); \end{cases} \tag{82}$$

$$D((\mathcal{R}_0\mathcal{R}_1)^4(P_2^{(1)}))=\begin{bmatrix} 3.9208 & 1.2102 \\ 4.4494 & 1.4967 \end{bmatrix},\quad \text{with eigenvalues } 0.0908 \text{ and } 5.3267. \tag{83}$$

$$\text{(iii)}\quad \begin{cases} P_3^{(1)}=(-0.4492,\ -0.0324), \quad P_3^{(2)}=(-0.1076,\ -0.6774), \\ P_3^{(3)}=(0.4492,\ 0.0324), \quad P_3^{(4)}=(0.1076,\ 0.6774). \end{cases} \tag{84}$$

$$D((\mathcal{R}_0\mathcal{R}_1)^4(P_3^{(1)}))=\begin{bmatrix} -6.1730 & 44.6753 \\ -13.7347 & 90.5629 \end{bmatrix},\quad \text{with eigenvalues } 0.6515,\ 83.7384. \tag{85}$$

Note that $P_i^{(j)}$ satisfies

$$(\mathcal{R}_0\mathcal{R}_1)^k(P_i^{(j)})=P_i^{((k+j)\bmod 4)},\quad i=1,2,3.$$

Let us now zoom in on Fig. 10 about these period-4 points $P_1^{(1)}, P_2^{(1)}$, and $P_3^{(1)}$ (see Figs. 15, 16, and 17, respectively.)

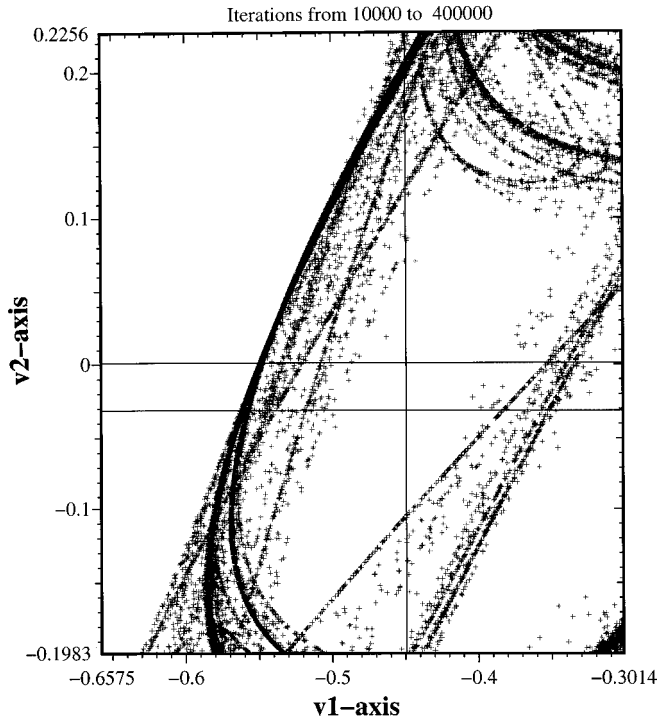

FIG. 17. This is a zoom-in of the vicinity of the period-4 point $P_3^{(1)}=(-0.4492,\ -0.0324)$, cf. (84), from Fig. 14. This point is not inside the ''cloud'' of the strange attractor; it does not have a homoclinic orbit. There is no chaos in the close vicinity of $P_3^{(1)}$.

From (81) and Fig. 15, by the continuity argument as given in Remark IV.1, it is quite clear that with $\mathcal{F}=(\mathcal{R}_0\mathcal{R}_1)^4$, the point $P_1^{(1)}$ satisfies (75) and (76). Thus, $P_1^{(1)}$ ''should be'' a snapback repeller. However, we observe in Fig. 15 that *chaos is not strong in the immediate vicinity of* $P_1^{(1)}$.

From (83) and Fig. 16, we see that $P_2^{(1)}$ is a saddle node fixed point $(\mathcal{R}_0\mathcal{R}_1)^4$ with a homoclinic orbit. *Chaos is very strong in the immediate vicinity of* $P_2^{(1)}$.

From (85), we note that $P_3^{(1)}$ is again a saddle node fixed point of $(\mathcal{R}_0\mathcal{R}_1)^4$. However, $P_3^{(1)}$ is not on the strange attractor, as can be seen from Fig. 17, suggesting that $P_3^{(1)}$ does not have a homoclinic orbit. *There is no chaos at all in the immediate vicinity of* $P_3^{(1)}$.

This example suggests the following:

(1) Snapback repellers imply chaos, but its strength may not be strong when other sources of chaos are present.
(2) A saddle node fixed point having a (forward) homoclinic orbit seems to cause strong chaos, even without the diffeomorphism assumption.

Point (2) above appears particularly important. We hope to be able to report some results in the future. □

ACKNOWLEDGMENTS

G.C. and J.Z. were supported in part by NSF Grant DMS No. 96-10076, Texas Advanced Research Program Grant No. 010366-046, and Texas A&M University Grant No. 96-36. S.B.H. was supported in part by a grant from the National Council of Science of the Republic of China.

[1] G. Chen, H. B. Hsu, and J. Zhou, ''Linear superposition of chaotic and orderly vibrations on two serially connected strings with a van der Pol joint,'' Int. J. Bifurcation Chaos Appl. Sci. Eng. **6**, 1509–1527 (1996).
[2] G. Chen, H. B. Hsu, and J. Zhou, ''Chaotic vibration of the one-dimensional wave equation due to a self-excitation boundary condition, Part I: controlled hysteresis,'' Transactions Amer. Math. Soc. (to appear).
[3] G. Chen, H. B. Hsu, and J. Zhou, ''Chaotic vibration of the one-dimensional wave equation due to a self-excitation

boundary condition, Part II: energy injection, period doubling and homoclinic orbit,'' Int. J. Bifurcation Chaos Appl. Sci. Eng. **8**, 423–445 (1998).
[4] G. Chen, H. B. Hsu, and J. Zhou, ''Chaotic vibration of the one-dimensional wave equation due to a self-excitation boundary condition, Part III: natural hysteresis memory effects,'' Int. J. Bifurcation Chaos Appl. Sci. Eng. **8**, 447–470 (1998).
[5] G. Chen and J. Zhou, *Vibration and Damping in Distributed Systems, Vol. I, Analysis, Estimation, Attenuation and Design* (CRC, Boca Raton, FL, 1994).
[6] K. S. Liu, F. L. Huang, and G. Chen, ''Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **49**, 1694–1707 (1989).
[7] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer-Verlag, New York, 1990).
[8] C. Mira, L. Gardini, A. Barugola, and J. C. Cathala, *Chaotic Dynamics in Two-Dimensional Noninvertible Maps* (World Scientific, Singapore, 1996).
[9] F. R. Marotto, ''Snap-back repellers imply chaos in $\mathbb{R}^n$,'' J. Math. Anal. Appl. **63**, 199–223 (1978).
[10] J. J. Stoker, *Nonlinear Vibrations* (Wiley, New York, 1950).
[11] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, 2nd ed. (Addison–Wesley, New York, 1989).
[12] J. Franklin, *Matrix Theory* (Prentice–Hall, Englewood Cliffs, NJ, 1968).

Information gain within nonextensive thermostatistics

Lisa Borland,[a] Angel R. Plastino,[b] and Constantino Tsallis[c]
Centro Brasileiro de Pesquisas Fisicas, Rua Dr. Xavier Sigaud 150, 22290-180 Rio de Janeiro, RJ, Brazil

(Received 8 April 1998; accepted for publication 17 June 1998)

We discuss the information theoretical foundations of the Kullback information gain, recently generalized within a nonextensive thermostatistical formalism. General properties are studied and, in particular, a consistent test for measuring the degree of correlation between random variables is proposed. In addition, minimum entropy distributions are discussed and the H-theorem is proved within the generalized context. © *1998 American Institute of Physics.* [S0022-2488(98)02712-1]

I. INTRODUCTION

There is a currently growing body of evidence implying that a nonextensive thermostatistical description is needed in order to adequately understand a vast class of physical phenomena, ranging from long-range interactions to self-organized criticality. A generalized thermostatistical formalism recently introduced by one of us[1] may be an interesting candidate for dealing with those kinds of physical scenarios. The generalized formalism has already been applied to astrophysical self-gravitating systems,[2] cosmology,[3] the solar neutrino problem,[4] anomalous diffusion of correlated[5–7] and Lévy[8] types, pure-electron plasma turbulence phenomena,[9] low-[10] and high-dimensional[11] dissipative nonlinear dynamical systems, anomalous Hamiltonian systems,[12] linear response theory,[13] electron–phonon anomalous relaxation,[14] variational and perturbation methods,[15] simulated annealing and other optimization techniques,[16] among many others.[1]

The generalized formalism is based upon the nonlogarithmic entropy[1]

$$S_q = k \sum_j p_j^q \frac{(1-p_j^{1-q})}{1-q}, \tag{1}$$

where q is a real number and the p_j are the usual probabilities of events j; k is a positive constant, from now on taken to be unity without loss of generality. The standard Shannon entropy[17] $S = -\Sigma_j p_j \ln p_j$ is recovered for $q \to 1$,[1] but S_q has some important properties which differ from the Shannon one in the general $q \neq 1$ case. Most noteworthy is the fact that the generalized entropy is *nonextensive*, in the sense that $S_q(A+B) = S_q(A) + S_q(B) + (1-q) S_q(A) S_q(B)$, where A and B are two *independent* systems, i.e., with probabilities satisfying $p_{ij}(A+B) = p_i(A) p_j(B)$. This entropy has been shown to lay the foundation to a generalized statistical mechanics, playing the same role the standard logarithmic entropy does within the standard information theory approach pioneered by Jaynes.[18] General theoretical properties of the nonlogarithmic entropy have been studied in Refs. 19–21. Also, connections have been made with the dynamical thermostatting approach to statistical ensembles.[22]

An important quantity closely related to the standard entropy is the Kullback information gain,[23] which measures the mean change of information associated with two probability distributions p and p'. This measure shows many desirable properties. First, it is *invariant* under arbitrary coordinate transformations, allowing the formulation of physical theories in a covariant fashion. Second, it provides a *minimum* entropy formalism for statistical inference, automatically incorporating a not necessarily uniform prior.[24] Third, the Kullback entropy is of utmost importance in,

[a]Electronic mail: lisa@cat.cbpf.br
[b]Permanent address: Facultad de Ciencias Astronómicas y Geofísicas, Universidad Nacional de La Plata, C. C. 727, 1900 La Plata, Argentina. Electronic mail: plastino@fcaglp.fcaglp.unlp.edu.ar
[c]Electronic mail: tsallis@cat.cbpf.br

 © 1998 American Institute of Physics

for example, correctly formulating the H-theorem for the Fokker–Planck equation based on the relative entropy between *any* two solutions.[25,26] In addition, this entropy also measures the degree of statistical dependence of two random variables.[26]

Recently, one of us proposed a generalization of the Kullback measure[27] within the framework of the generalized thermostatistics. In this paper, we discuss some of the main properties of that measure. In Sec. II we derive the generalized Kullback information gain and in Sec. III we study some general theorems and inequalities. Also, a consistent test for measuring the degree of dependence between random variables (introduced in Ref. 27) is worked out in detail. Section IV presents the minimum Kullback entropy distributions and Sec. V proves the H-theorem for the Fokker–Planck and Liouville equations, based on the generalized Kullback measure. Finally, some conclusions are addressed in Sec. VI.

II. DERIVATION OF THE GENERALIZED KULLBACK INFORMATION GAIN

Let us review the derivation of the Kullback information gain measure in the standard case.[23,26] Assume that a particular set of measurements gives the probabilities p_j of obtaining the outcomes denoted by the index j. Then the Shannon entropy of the system can be written as

$$S=-\sum_j p_j \ln p_j, \tag{2}$$

where we can interpret the quantity $\sigma_j \equiv -\ln p_j$ as the information content of the outcome j. Let us denote the first set of measurements by p'_j and assume that a new set of measurements is made which yields a new set of values p_j for the probabilities of the outcomes j. The change of information $\Delta\sigma_j$ is thus equal to $\Delta\sigma_j=-(\ln p'_j-\ln p_j)$, and the mean change of information is obtained by averaging over the new distribution defined by the p_j to obtain

$$K(p,p')\equiv\sum_j p_j\Delta\sigma_j=\sum_j p_j \ln\frac{p_j}{p'_j}. \tag{3}$$

This is the standard derivation and definition of the Kullback–Leibler information gain (also known as *mutual information, cross entropy*, and *relative entropy*).

We now show that the generalized Kullback–Leibler measure follows naturally from the standard derivation of the Kullback entropy, if only the nonextensive formalism[1] is employed instead of the standard one. Within that setting, the entropy of Eq. (1) is used instead of the Shannon entropy, and all expectation values are replaced with the so-called *q-expectation values* of the form $\langle\omega\rangle_q\equiv\Sigma_j p_j^q\omega_j$, where ω is a state variable of the system.

It follows then that Eq. (1) can be interpreted as the q-average of the generalized information content given by $\sigma_j^q\equiv-(p_j^{1-q}-1)/(1-q)$. We denote the probabilities of the first set of measurements by p'_j and assume that we then obtain a new set of measurements with the probabilities p_j. The change in information is given by $\Delta\sigma_j^q=[1/(1-q)]((1-p_j'^{1-q})-(1-p_j^{1-q}))$. Taking the q-average of $\Delta\sigma_j$ yields

$$K_q(p,p')=\sum_j \frac{p_j^q}{1-q}(p_j^{1-q}-p_j'^{1-q}), \tag{4}$$

which is the generalized Kullback–Leibler information gain. This can alternatively be written as

$$K_q(p,p')=\sum_j p_j \ln_q\frac{p_j}{p'_j}, \tag{5}$$

where we make use of the definition of the q-logarithm function $\ln_q$ (Ref. 28):

$$\ln_q(p)\equiv\frac{p^{1-q}-1}{1-q}. \tag{6}$$

Hence, $\ln_1(p)$ recovers the standard $\ln(p)$. In continuous variables, the measure takes on the form

$$K_q(p,p')=\int dx\,\frac{p(x)^q}{1-q}\,(p(x)^{1-q}-p'(x)^{1-q})=\int dx\;p(x)\ln_q\left(\frac{p(x)}{p'(x)}\right), \tag{7}$$

which coincides exactly with the measure proposed in Eq. (5) of Ref. 27.

III. GENERAL PROPERTIES AND THEOREMS

A. Positivity and concavity properties

As in the standard $q=1$ case, the generalized Kullback information measure has a definite sign (depending on q). Here we provide an alternative proof to the discussion presented in Ref. 27. For $q>0$ we obtain

$$K_q(p,p')=\sum_j\;\frac{p_j}{1-q}\left(1-\left(\frac{p'_j}{p_j}\right)^{1-q}\right) \tag{8}$$

$$\geqslant\frac{1}{1-q}\left(1-\left(\sum_j\;p_j\,\frac{p'_j}{p_j}\right)^{1-q}\right) \tag{9}$$

$$=\frac{1}{1-q}\left(1-\left(\sum_j\;p'_j\right)^{1-q}\right) \tag{10}$$

$$=\frac{1}{1-q}\,(1-1)=0, \tag{11}$$

where we have used Jensens inequality: If $\Sigma_j\alpha_j=1$ and $f''(x)>0$, then

$$\sum_j\;\alpha_j f(x_j)\geqslant f\left(\sum_j\;\alpha_j x_j\right). \tag{12}$$

Here, $f(x)=[1/(1-q)](1-x^{1-q})$ with $x=p'_j/p_j$ and $\alpha_j=p_j$. Conversely, for $q<0$, $f''(x)<0$ and the sign of the inequality reverses. Therefore, we have shown that

$$K_q\geqslant 0,\quad q>0, \tag{13}$$

$$K_q=0,\quad q=0, \tag{14}$$

$$K_q\leqslant 0,\quad q<0. \tag{15}$$

The equalities (for $q\neq 0$) hold if and only if $p_j=p'_j$ for all j. Along the same lines, we obtain for $q>0$

$$\left(\sum_i\;a_i\right)\left\{\frac{[(\Sigma_i a_i)/(\Sigma_j b_j)]^{q-1}-1}{q-1}\right\}\leqslant\sum_i\;a_i\left\{\frac{[a_i/b_i]^{q-1}-1}{q-1}\right\}, \tag{16}$$

where Jensens inequality was employed using

$$f(x)=x\left(\frac{x^{q-1}-1}{q-1}\right), \tag{17}$$

$$\alpha_i=b_i\Big/\left(\sum_j\;b_j\right), \tag{18}$$

$$x_i=a_i/b_i\,. \tag{19}$$

For $q=1$, this expression reduces to the well-known log-sum inequality.[26] A direct consequence of Eq. (16) is that the measure $K_q(p,p')$ is a convex function of its both arguments. This can be seen if we consider the probability distributions p and p' to be equal mixtures of two other distributions, $p=\lambda p_j^1+(1-\lambda)p_j^2$ and $p'=\lambda p_j'^1+(1-\lambda)p_j'^2$. We then apply the log-sum inequality Eq. (16) to each such term, by identifying $a_1=\lambda p_j^1$ and $a_2=(1-\lambda)p_j^2$ and similarly $b_1=\lambda p_j'^1$ and $b_2=(1-\lambda)p_j'^2$. Finally, we obtain after summing over j

$$K_q(\lambda p^1+(1-\lambda)p^2,\lambda p'^1+(1-\lambda)p'^2)\leqslant\lambda K(p^1,p'^1)+(1-\lambda)K(p^2,p'^2) \tag{20}$$

for $q>0$. For $q<0$, $K_q(p,p')$ is instead a concave function of its arguments, and the sign of the inequality reverses.

B. Relative entropy

Let us now inspect the special case when $p=p(x,y)$ and $p'=p(x)p(y)$, corresponding respectively to the *joint* and *marginal* probability distributions of a system consisting of two random variables x and y. In this case, the Kullback information measure is often referred to as the *relative entropy* between x and y, and we shall refer to it as $I(x,y)$. Notice that $I(x,y)$ is a special case of $K(p,p')$ for a particular choice of p and p'. In the $q=1$ case, it is well known that[26]

$$I(x,y)=S(x)+S(y)-S(x,y) \tag{21}$$

$$=S(x)-S(x|y) \tag{22}$$

$$=S(y)-S(y|x), \tag{23}$$

where $S(x)$ and $S(y)$ are the Shannon entropies of the marginal distributions and $S(x,y)$ is the Shannon entropy of the joint probability distribution. $I(x,y)$ vanishes if and only if x and y are independent variables. The conditional entropy of type $S(x|y)$ is however defined as $S(x|y)=-\Sigma_{x,y}p(x,y)\ln p(x|y)=-E_{p(x,y)}[\ln p(x|y)]$, where E_p denotes the expectation value with respect to the distribution p. Notice that all terms in the condition of Eq. (21) can equally be expressed in terms of E_p, namely

$$I(x,y)=-E_{p(x,y)}[\ln p(x)+\ln p(y)-\ln p(x,y)] \tag{24}$$

$$=-E_{p(x,y)}[\ln p(x)-\ln (p(x|y)] \tag{25}$$

$$=-E_{p(x,y)}[\ln p(y)-\ln (p(y|x)], \tag{26}$$

because $p(x,y)=p(y|x)p(x)=p(x|y)p(y)$ and $\Sigma_y p(y|x)=\Sigma_x p(x|y)=1$.

The idea now is to see how these properties generalize in the $q\neq 1$ case. We write

$$I_q(x,y)=-\sum_{x,y}\frac{p(x,y)^q}{1-q}\left((p(x)p(y))^{1-q}-p(x,y)^{1-q}\right) \tag{27}$$

$$=\sum_{x,y}\frac{p(x,y)^q}{1-q}\left((1-(p(x)p(y))^{1-q})-(1-p(x,y)^{1-q})\right). \tag{28}$$

Applying the relation

$$1-ab=(1-a)+(1-b)-(1-a)(1-b) \tag{29}$$

to rewrite the first term in the brackets on the rhs, we immediately obtain

$$I_q(x,y)=\sum_{x,y}\frac{p(x,y)^q}{1-q}\left((1-p(x)^{1-q})+(1-p(y)^{1-q})\right.$$
$$\left.-(1-p(x)^{1-q})(1-p(y)^{1-q})-(1-p(x,y)^{1-q})\right). \tag{30}$$

Using the definition (6) of the $\ln_q$ function, we see that this yields

$$I_q(x,y) = -E^q_{p(x,y)}[\ln_q p(x) + \ln_q p(y) + (1-q)\ln_q p(x)\ln_q p(y) - \ln_q p(x,y)], \tag{31}$$

where we have introduced the notation E^q_p to denote the q-expectation value over the distribution p. One immediately verifies that Eq. (31) is a very natural generalization of the property equation (21) or (24), where the expection values of $\ln p(x)$ over $p(x,y)$ are replaced by the q-expectation values over $p(x,y)$, and the sum of $\ln p(x)$ and $\ln p(y)$ are replaced by the *pseudo-sum* of the nonextensive formalism, namely $\ln_q p(x) + \ln_q p(y) + (1-q)\ln_q p(x) \ln_q p(y)$.

We now return to Eq. (28) and rewrite $p(x,y)$ in the second term within the brackets on the rhs as $p(x,y) = p(x)p(y|x)$. Applying the expansion of Eq. (29) to that term as well, we see that our result of Eq. (31) can equally be written as

$$\begin{aligned} I_q(x,y) = -E^q_{p(x,y)}[&\ln_q p(x) - \ln_q p(x|y) \\ &-(1-q)(\ln_q p(y)\ln_q p(x|y) - \ln_q p(x)\ln_q p(y))] \end{aligned} \tag{32}$$

$$\begin{aligned} = -E^q_{p(x,y)}[&\ln_q p(y) - \ln_q p(y|x) \\ &-(1-q)(\ln_q p(x)\ln_q p(y|x) - \ln_q p(x)\ln_q p(y))]. \end{aligned} \tag{33}$$

These expressions are again very natural generalizations of the properties of Eqs. (25) and (26) for the standard $q=1$ case.

C. Upper and lower bounds

We are now interested in defining bounds on the measure $I_q(x,y)$. If we wish to use it as a criterion to determine the degree of correlation between the two variables x and y, then it is desirable to know its value corresponding to the *minimum* and *maximum* correlation between the variables. Let us first discuss the $q>0$ case. As follows from Eq. (13), we know that $I_q \geqslant 0$, and we see [for example, from Eq. (32)] that $I_q = 0$ occurs when $p(x|y) = p(x)$ which corresponds to the case when x is totally independent of y. Alternatively, we see that if x is maximally correlated with y such that there is a one-to-one mapping between the two variables, then there is no uncertainty of x if one knows y and consequently the conditional entropy

$$S_q(x|y) \equiv \sum_x p^q(x|y)\ln_q p(x|y) \tag{34}$$

must vanish. This case should yield the upper bound on I_q because one expects the mutual information between x and y to be at its height if the two variables are completely correlated. Indeed we see that this is true because the following two terms of I_q of Eq. (32), namely $-E^q_{p(x,y)} \ln_q p(x|y)$ and $-E^q_{p(x,y)} \ln_q p(y) \ln_q p(x|y)$ must vanish in this case, whereas they constitute positive quantities which are subtracted, respectively, from the extensive and nonextensive parts of I_q in all other cases. To see that they do in fact vanish, we consider first the term $E^q_{p(x,y)} \ln_q p(x|y)$ which satisfies

$$0 \leqslant -E^q_{p(x,y)}[\ln_q p(x|y)] = -\sum_{x,y} p^q(y)p^q(x|y)\ln_q p(x|y) \leqslant S_q(x|y) = 0. \tag{35}$$

In a similar way, we see that

$$\begin{aligned} -E^q_{p(x,y)}[\ln_q p(y) \ln_q p(x|y)] &= -\sum_y p(y)^q \ln_q p(y) \sum_x p(x|y)^q \ln_q p(x|y) \\ &= \sum_y p(y)^q \ln_q p(y) S_q(x|y) = 0. \end{aligned} \tag{36}$$

We then see, from Eq (32), that the upper bound $I^{\max}_q$ is given by

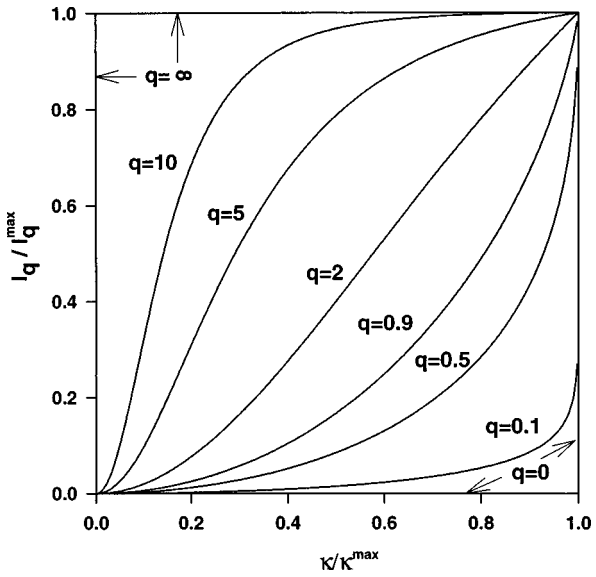

FIG. 1. The normalized index $R \equiv I_q / I_q^{\max}$ is shown as a function of $\kappa/\kappa^{\max}$ for different values of q. The ratio $\kappa/\kappa^{\max}$ is a measure of the degree of correlation of two binary random variables, whose joint probability distribution is given by $p(1,1)=p^2+\kappa$, $p(1,2)=p(2,1)=p(1-p)-\kappa$, and $p(2,2)=(1-p)^2+\kappa$. The curves correspond to typical values of q and $p=\frac{1}{2}$. For this special case one can express R analytically as a function of q (Ref. 40): $R(q)=[(1+4\kappa)^q+(1-4\kappa)^q-2]/[(1-2^{(1-q)})[(1+4\kappa)^q+(1-4\kappa)^q]]$.

$$I_q^{\max}(x,y) = -E^q_{p(x,y)}[\ln_q p(x) + (1-q)\ln_q p(x) \ln_q p(y)]. \tag{37}$$

D. Criterion for statistical testing

Equation (37) together with Eq. (31) enables the definition of the ratio

$$R \equiv I_q / I_q^{\max}, \in [0,1], \tag{38}$$

which can be used as a *criterion for measuring the degree of dependence between random variables* (time series and others). In Fig. 1 we plot the ratio R as a function of the correlation between the variables x and y, for typical values of $q>0$. This ratio varies from 0 (corresponding to zero correlation, i.e., independence between x and y) to 1 (corresponding to maximum correlation, i.e., a one-to-one mapping between x and y). Note that there is an *optimal* q^{opt} (see Fig. 2) for a given degree of correlation, in the sense that the gradient of the index R is most sensitive and therefore most capable of detecting small variations of correlation among the variables. In Fig. 3 we show how the curve R versus q varies for different degrees of correlation.

The criterion R of Eq. (38) can hopefully be quite useful since it is (i) *consistent*, in the sense that it monotonically varies with the degree of dependence of the random variables; (ii) *versatile*, in the sense that it includes a parameter, the entropic index q, which can be tuned, for each given specific problem, to an optimal value q^{opt} for which the sensitivity of the criterion is maximal; (iii) *optimal*, in the sense that its two extreme values precisely correspond to full independence and full dependence.

For $q<0$ all signs reverse but we have the additional problem that the q-generalized entropies diverge whenever the probabilities p_j of any of the microstates j approach 0. Indeed, for negative q, certainty implies that the entropy diverges.[1] Hence, the conditional entropy $S(x|y)\to\infty$ in the case of complete correlation between the variables x and y. Therefore, the difference between the index I_q and $I_q^{\max}$ as defined in Eq. (37) can become arbitrarily large.

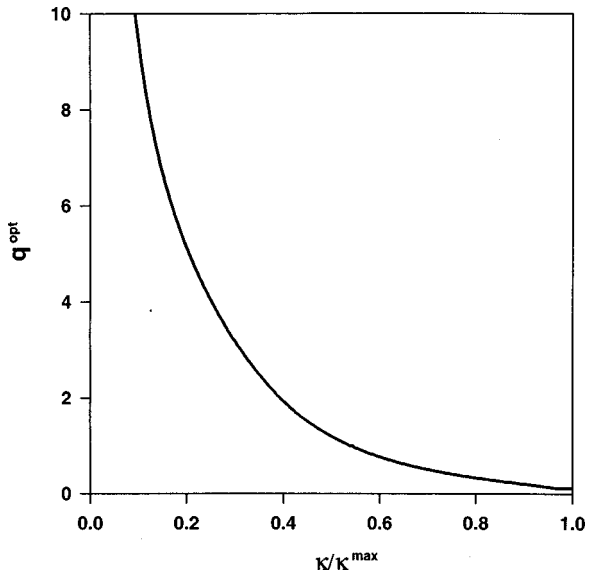

FIG. 2. As can be seen in Fig. 1, for a given value of $\kappa/\kappa^{\max}$, there is an optimal value of q, in the sense that the index R is most sensitive to detect variations on the degree of statistical correlation. Here we plot, for the same example presented in Fig. 1, the optimal value q^{opt} as a function of $\kappa/\kappa^{\max}$.

E. Shift information

Another interesting property of the generalized Kullback information measure is the q-invariance of its relationship to the Fisher information. From now on we treat the measure in its continuous form [Eq. (7)]. As in the $q=1$ case, we consider a monoparametric family of distributions characterized by the parameter θ. One obtains

$$K_q(p(\theta),p(\theta+\Delta))=\frac{1}{q-1}\int dx\; p(x,\theta)\left\{\left(\frac{p(x,\theta)}{p(x,\theta+\Delta)}\right)^{q-1}-1\right\}. \tag{39}$$

A Taylor expansion in the variable Δ yields, up to second order,

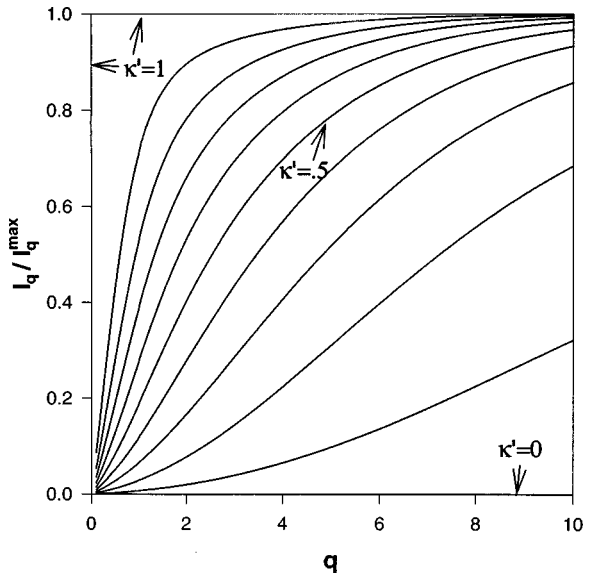

FIG. 3. The index $R=I_q/I_q^{\max}$ as a function of q for different values of $\kappa'=\kappa/\kappa^{\max}$ between 0 and 1, plotted for the same example as in Figs. 1 and 2. The curves for $\kappa'=0$, 0.5 and 1 are indicated, and the intermediate curves correspond to 0.1 increments of κ'.

$$K_q(p(x,\theta),p(x,\theta+\Delta)) \approx \frac{1}{2}\, q\Delta^2 \int \frac{1}{p}\left(\frac{\partial p}{\partial \theta}\right)^2 dx. \tag{40}$$

Up to the factor q, this coincides with the standard result for $q=1$.[29] Note that it is the standard ($q=1$) Fisher information that appears here, and not the q-generalized Fisher information that was introduced in Ref. 30.

IV. MINIMUM RELATIVE INFORMATION DISTRIBUTIONS

Now we will briefly discuss the minimum relative entropy distributions consistent with a set of M given constraints.[24] Let's assume that the only information we know about a probability distribution p_j, $j=1,...,N$, are the q-generalized expectation values of M quantities A_r, $r=1,...M$,

$$\langle A_r \rangle_q = \sum_j A_{rj} p_j^q. \tag{41}$$

Following the q-generalized maximum entropy prescription, one would adopt the probability distribution that *maximizes* the entropy S_q, under the constraints imposed by normalization and the set (41) of mean values. However, such a procedure has a strong, and sometimes undesirable, bias towards the uniform distribution.[23] In many applications, we may prefer a nonuniform prior probability distribution $\tilde{p}_j$ that would be our default choice in the case that no further information was available. A natural way of incorporating an arbitrary prior distribution $\tilde{p}_j$ is to find the probability distribution that *minimizes* the generalized Kullback information measure $K_q(p,\tilde{p})$, under the relevant constraints. Introducing appropriate Lagrange multipliers, the concomitant variational principle reads

$$\delta\left\{ K_q + \alpha \sum_{j=1}^{N} p_j + \sum_{r=1}^{M} \lambda_r \left(\sum_{j=1}^{N} A_{rj} p_j^q \right) \right\} = 0. \tag{42}$$

The Kullback minimum information principle[23] is recovered in the limit $q \to 1$. A straightforward calculation yields

$$p_j = Z_q^{-1} \left[\tilde{p}_j^{1-q} - (1-q) \sum_{r=1}^{M} \lambda_r A_{rj} \right]^{1/(1-q)}, \tag{43}$$

where the partition function Z_q is given by

$$Z_q = \sum_j \left[\tilde{p}_j^{1-q} - (1-q) \sum_{r=1}^{M} \lambda_r A_{rj} \right]^{1/(1-q)}. \tag{44}$$

It is easy to verify that the Lagrange multipliers λ_r, $r=1,...,M$, the corresponding mean values $\langle A_r \rangle_q$, and the relative entropy K_q satisfy

$$K_q = -\lambda_J - \sum_{r=1}^{M} \lambda_r \langle A_r \rangle_q, \tag{45}$$

where the Jaynes parameter λ_J is given by

$$\lambda_J \equiv \frac{Z_q^{1-q} - 1}{1-q} = \ln_q Z_q. \tag{46}$$

After some algebra, it is also possible to obtain Jaynes' well-known thermodynamic relations relating the entropy, the relevant mean values, and the corresponding Lagrange multipliers, i.e.,

$$\frac{\partial \lambda_J}{\partial \lambda_r} = -\langle A_r \rangle_q \quad (r=1,...M) \tag{47}$$

and

$$\frac{\partial K_q}{\partial \langle A_r \rangle_q} = -\lambda_r \quad (r=1,...M). \tag{48}$$

Here, the q-invariance becomes explicit. These thermodynamic relations were recently proved[31,32] to hold true for, essentially, *any* entropic measure. They are a universal feature of the maximum entropy variational procedure. However, the present case of the q-generalized relative information measure K_q shows the interesting property that the partition function Z_q can be explicitly written as a function of the Lagrange multipliers λ_r. This is not a universal characteristic of the maximum entropy prescription. Up to now, there are only a few known entropic functionals that share this property.[33]

V. H-THEOREM

The D-dimensional Fokker–Planck (FP) equation reads[25]

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x} (V_i(\mathbf{x})p) + \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} (G_{ij}(\mathbf{x})p), \tag{49}$$

where $p(\mathbf{x},t)$, $\mathbf{x} \in R^D$, is a time-dependent probability distribution; $V_i(\mathbf{x})$, $i=1,...,D$, stand for the components of the drift vector $\mathbf{V}$; and $G_{ij}(\mathbf{x})$, $i,j=1,...,N$, are the components of the diffusion tensor $\mathbf{G}$. We will consider the behavior of the relative information between two arbitrary time-dependent solutions (p_1 and p_2) to the FP equation, i.e.,

$$K_q = \frac{1}{1-q} \int d\mathbf{x}\, p_1(\mathbf{x},t) \left[1 - \left(\frac{p_2(\mathbf{x},t)}{p_1(\mathbf{x},t)} \right)^{1-q} \right]. \tag{50}$$

It is worthwhile to discuss first in some detail the particular, but important, *deterministic* case of a vanishing diffusion tensor.

A. Liouville equation

The D-dimensional Fokker–Planck equation with a vanishing diffusion tensor reduces to the Liouville equation. Although the Liouville equation is usually associated to Hamiltonian systems, it encompasses more general situations as well. The general Liouville equation

$$\frac{\partial p}{\partial t} + \nabla \cdot (\mathbf{V}p) = 0 \tag{51}$$

describes the time evolution of an ensemble of identical dynamical systems (but with different initial conditions), each one of them evolving according to the set of coupled ordinary differential equations:

$$\frac{dx_i}{dt} = V_i(\mathbf{x}) \quad (i=1,...,D). \tag{52}$$

Hamiltonian systems belong to the special family of dynamical systems with divergenceless phase-space flows, i.e.,

$$\nabla \cdot \mathbf{V} = 0. \tag{53}$$

Liouville himself[34] introduced his celebrated equation in the form (51), within the general context of arbitrary dynamical systems.

The behavior of the Shannon entropy $S(p(\mathbf{x},t))$, where $p(\mathbf{x},t)$ stands for a solution to the Liouville equation, has been discussed by many authors.[35–37] Here the time derivative of S is given by the mean value of the divergence of the flow,

$$\frac{dS}{dt}=\langle\nabla\cdot\mathbf{V}\rangle. \tag{54}$$

This implies that the entropy is time-independent for divergenceless flows. However, the entropy increases when the flow in phase space has a positive mean divergence. The above result can be easily generalized for the case of the entropy S_q of Eq. (1). One obtains[5]

$$\frac{dS_q}{dt}=\langle\nabla\cdot\mathbf{V}\rangle_q. \tag{55}$$

It is interesting to realize that the q-generalized mean values allow us to write the well-known equation (54) in a q invariant way.

In spite of its importance within the statistical approach to dynamical systems, Eq. (54) [or its q-generalization (55)] constitutes a somewhat nonoptimal result. It means that the entropy measure associated with the *fine-grained* phase space distribution describing a *deterministic* system may grow with time. Intuitively one expects entropy to increase when there is a loss of information associated with some kind of coarse graining, or with an incomplete knowledge of the equations of motion (such as when stochastic forces are present). A measure that appropriately takes into account these objections is provided by the relative information $K(p_1,p_2)$ [or its q-generalization $K_q(p_1,p_2)$], where $p_1(\mathbf{x},t)$ and $p_2(\mathbf{x},t)$ are two *arbitrary* time-dependent solutions to the Liouville equation. The time derivative of K_q is given by

$$\begin{aligned}\frac{dK_q}{dt}&=\frac{1}{q-1}\int d\mathbf{x}\left\{q\left(\frac{\partial p_1}{\partial t}\right)\left(\frac{p_2}{p_1}\right)^{1-q}+(1-q)\left(\frac{\partial p_2}{\partial t}\right)\left(\frac{p_2}{p_1}\right)^{-q}\right\}\\&=q\int d\mathbf{x}\left[p_2\left(\frac{p_2}{p_1}\right)^{-q-1}-p_1\left(\frac{p_2}{p_1}\right)^{-q}\right]\sum_i V_i\,\frac{\partial}{\partial x_i}\left(\frac{p_2}{p_1}\right)=0.\end{aligned} \tag{56}$$

We can see that $K_q(p_1,p_2)$ is *a constant of motion for the Liouville equation*. Given two arbitrary initial distributions $p_1(\mathbf{x},t_0)$ and $p_2(\mathbf{x},t_0)$, the associated mean change of information $K_q(p_1,p_2)$ is not modified if we let p_1 and p_2 to evolve according to the Liouville equation (i.e., with a *fully deterministic dynamics*).

Our results are related to the analysis in Ref. 38 of the behavior of the generalized q-entropy within Liouville dynamics. There, the functional

$$Q=\int d\mathbf{x}\;p(\mathbf{x},t)\left(\frac{p(\mathbf{x},t)}{\gamma(\mathbf{x})}\right)^{q-1} \tag{57}$$

is introduced, where $p(\mathbf{x},t)$ is a solution of the Liouville equation. The function $\gamma(\mathbf{x})$ is interpreted in terms of a volume element $d\Omega=\gamma(\mathbf{x})d\mathbf{x}$. The motivation for introducing the function $\gamma(\mathbf{x})$ and the functional Q is to work within a covariant formalism. The functional Q is interpreted as a covariant form of the entropy S_q. It was shown[38] that the functional Q is time invariant whenever the covariant divergence

$$\tilde{V}(\mathbf{x})=\frac{1}{\gamma}\,\nabla\cdot(\gamma\mathbf{V}) \tag{58}$$

vanishes. However, as already mentioned, the relation $\tilde{V}=0$ implies that $\gamma(\mathbf{x})$ constitutes a stationary solution to the Liouville equation (51). Besides, it is clear that the functional Q, up to unessential multiplicative and additive constants, is just the relative information $K_q(p,\gamma)$. Hence, that result can be reformulated by stating that the relative information between an arbitrary time-dependent solution $p(\mathbf{x},t)$ *and a stationary solution* $\gamma(\mathbf{x})$ of the Liouville equation is constant in

time. In the present work we have proved that the result of Ref. 38 is only a particular case of a more general property: the relative entropy between *two arbitrary solutions* is time independent.

B. Fokker–Planck equation

Now we will consider the case of the full Fokker–Planck equation [Eq. (49)] with a nonvanishing diffusion tensor. It is clear from the last section that the drift term in the Fokker–Planck equation makes no contribution to the time derivative of K_q. After two successive integrations by parts, the time derivative of K_q can be cast as

$$\frac{dK_q}{dt}=q\int d\mathbf{x}\sum_{ij} G_{ij}\left\{p_2\,\frac{\partial}{\partial x_i}\left[\left(\frac{p_2}{p_1}\right)^{-1-q}\frac{\partial}{\partial x_j}\left(\frac{p_2}{p_1}\right)\right]-p_1\,\frac{\partial}{\partial x_i}\left[\left(\frac{p_2}{p_1}\right)^{-q}\frac{\partial}{\partial x_j}\left(\frac{p_2}{p_1}\right)\right]\right\}, \quad (59)$$

which can be simplified to obtain

$$\frac{dK_q}{dt}=-q\int d\mathbf{x}\; p_2\left(\frac{p_2}{p_1}\right)^{-2-q}\sum_{ij} G_{ij}\left[\frac{\partial}{\partial x_i}\left(\frac{p_2}{p_1}\right)\right]\left[\frac{\partial}{\partial x_j}\left(\frac{p_2}{p_1}\right)\right]\leqslant 0, \quad (60)$$

where we have used the fact that G_{ij} is a positive definite tensor. In the $q\to 1$ limit, this expression reduces to the standard H-theorem for the Fokker–Planck equation.[25]

VI. CONCLUSIONS

The q-generalized Kullback information gain has been shown to satisfy suitably generalized versions of some basic theorems and properties fulfilled by the standard $(q=1)$ Kullback one. As already proposed by one of us,[27] the generalized relative entropy can be used as an efficient criterion for statistical independence of, for example, time series data. In the present paper we improved upon that idea by proposing a normalized version of that criterion. Hopefully, it may be quite useful since it is (i) *consistent*, in the sense that it monotonically varies with the degree of dependence of the random variables; (ii) *versatile*, in the sense that it includes a parameter, the entropic index q, which can be tuned, for each given specific problem, to an optimal value q^{opt} for which the sensitivity of the criterion is maximal; (iii) *optimal*, in the sense that its two extreme values precisely correspond to full independence and full dependence.

The q-maximum entropy distributions have already been shown to appropriately describe some experimental situations where the standard theory fails.[9] It is therefore natural to believe that the generalized Kullback minimum entropy solutions, which allow the incorporation of a nonuniform prior, may enlarge the range of successful applications of the q-nonextensive formalism. The H-theorem for a general Fokker–Planck (linear) equation was also reformulated using the generalized information gain. We remark that there is an appreciable advantage to studying the relative change in entropy of two arbitrary solutions. Indeed, in the case of the Liouville equation (purely deterministic dynamics), the cross entropy remains constant in time, whereas for the full Fokker–Planck equation (both deterministic and random components in the dynamics), the cross entropy monotonically decreases with time. Our treatment generalizes and unifies previous results already existing in the literature.[26,35–39] Based on the present considerations, we believe that the generalized Kullback measure may find useful future applications in both statistical physics and information theory.

ACKNOWLEDGMENTS

We are grateful to G. Mato and H. Wio for useful remarks, and are indebted to G. Mato for kindly providing to us the analytic expression of $R(q)$ appearing in the caption of Fig. 1. ARP gratefully acknowledges partial support from Argentina's National Research Council (CONICET). Partial support by PRONEX and CNPq (Brazilian Agencies) is also acknowledged. LB acknowledges support from the U.S. National Science Foundation. This material is based upon work supported by the National Science Foundation under Grant No. INT-9703649.

[1] C. Tsallis, J. Stat. Phys. **52**, 479 (1988); E. M. F. Curado and C. Tsallis, J. Phys. A **24**, L69 (1991); Corrigenda: **24**, 3187 (1991) and **25**, 1019 (1992); C. Tsallis, Physica A **221**, 277 (1995); C. Tsallis, Fractals **6**, 539 (1995), and references therein; an updated bibliography can be found in http://tsallis.cat.cbpf.br/biblio.htm

[2] A. R. Plastino and A. Plastino, Phys. Lett. A **174**, 384 (1993); A. R. Plastino and A. Plastino, Phys. Lett. A **193**, 251 (1994); J. J. Aly, in *N-Body Problems and Gravitational Dynamics*, Proceedings of the Meeting held at Aussois, France (21–25 March 1993), edited by F. Combes and E. Athanassoula (Publications de l'Observatoire de Paris, Paris, 1993).
[3] V. H. Hamity and D. E. Barraco, Phys. Rev. Lett. **76**, 4664 (1996); D. F. Torres, H. Vucetich, and A. Plastino, Phys. Rev. Lett. **79**, 1588 (1997).
[4] G. Kaniadakis, A. Lavagno, and P. Quarati, Phys. Lett. B **369**, 308 (1996); P. Quarati, A. Carbone, G. Gervino, G. Kaniadakis, A. Lavagno, and E. Miraldi, Nucl. Phys. A **621**, 345c (1996).
[5] A. R. Plastino and A. Plastino, Physica A **222**, 347 (1995).
[6] C. Tsallis and D. J. Bukman, Phys. Rev. E **54**, R2197 (1996).
[7] L. Borland, Phys. Rev. E **57**, 6634 (1998).
[8] P. A. Alemany and D. H. Zanette, Phys. Rev. E **49**, 956 (1994); D. H. Zanette and P. A. Alemany, Phys. Rev. Lett. **75**, 366 (1995); M. O. Caceres and C. E. Budde, *ibid.* **77**, 2589 (1996); D. H. Zannette and P. A. Alemany, *ibid.* **77**, 2590 (1996); C. Tsallis, S. V. F. Levy, A. M. C. Souza, and R. Maynard, Phys. Rev. Lett. **75**, 3589 (1995); Erratum: **77**, 5442 (1996).
[9] B. M. R. Boghosian, Phys. Rev. E **53**, 4754 (1995); C. Anteneodo and C. Tsallis, J. Mol. Liq. **71**, 255 (1997).
[10] C. Tsallis, A. R. Plastino, and W.-M. Zheng, Chaos Solitons Fractals **8**, 885 (1997); M. L. Lyra and C. Tsallis, Phys. Rev. Lett. **80**, 53 (1998).
[11] F. A. Tamarit, S. A. Cannas, and C. Tsallis, Eur. Phys. J. B **1**, 545 (1998); A. R. R. Papa and C. Tsallis, Phys. Rev. E **57**, 3923 (1998).
[12] C. Anteneodo and C. Tsallis, Phys. Rev. Lett. **80**, 5313 (1998).
[13] A. K. Rajagopal, Phys. Rev. Lett. **76**, 3469 (1996).
[14] I. Koponen, Phys. Rev. E **55**, 7759 (1997).
[15] E. K. Lenzi, L. C. Malacarne, and R. S. Mendes, Phys. Rev. Lett. **80**, 218 (1998).
[16] T. J. P. Penna, Phys. Rev. E **51**, R1 (1995); Comput. Phys. **9**, 341 (1995); D. A. Stariolo and C. Tsallis, *Annual Reviews of Computational Physics, Vol. II*, edited by D. Stauffer (World Scientific, Singapore, 1995), p. 343; K. C. Mundim and C. Tsallis, Int. J. Quantum Chem. **58**, 373 (1996); J. Schulte, Phys. Rev. E **53**, 1348 (1996); I. Andricionaei and J. E. Straub, Phys. Rev. E **53**, R3055 (1996); P. Serra, A. F. Stanton, and S. Kais, Phys. Rev. E **55**, 1162 (1997); M. A. Moret, P. M. Bisch, and F. M. C. Vieira, Phys. Rev. E **57**, R2535 (1998).
[17] C. E. Shannon, Bell Syst. Tech. J. **27**, 379, 623 (1948).
[18] E. T. Jaynes, in *Statistical Physics*, edited by W. K. Ford (Benjamin, New York, 1963); A. Katz, *Statistical Mechanics* (Freeman, San Francisco, 1967).
[19] G. A. Raggio, J. Math. Phys. **36**, 4785 (1995).
[20] A. M. Meson and F. Vericat, J. Math. Phys. **37**, 4480 (1996).
[21] R. J. V. dos Santos, J. Math. Phys. **38**, 4104 (1997).
[22] A. R. Plastino and C. Anteneodo, Ann. Phys. **255**, 250 (1997).
[23] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
[24] A. R. Plastino, H. Miller, and A. Plastino, Phys. Rev. E **56**, 3927 (1997).
[25] H. Risken, *The Fokker-Planck Equation* (Springer, New York, 1989).
[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[27] C. Tsallis, Phys. Rev. E **58**, 1442 (1998).
[28] C. Tsallis, Quimica Nova **17**, 468 (1994); E. P. Borges, J. Phys. A **31**, 5281 (1998).
[29] G. V. Vstovsky, Phys. Rev. E **51**, 975 (1995).
[30] A. Plastino, A. R. Plastino, and H. G. Miller, Physica A **235**, 577 (1997); F. Pennini and A. Plastino, Physica A **247**, 559 (1997).
[31] A. Plastino and A. R. Plastino, Phys. Lett. A **226**, 257 (1997).
[32] R. S. Mendes, Physica A **242**, 299 (1997).
[33] E. M. F. Curado, communicated at the International Workshop on the Foundations of Statistical Mechanics and Thermodynamics, Natal, Brazil, October 1997.
[34] J. Liouville, J. Math. Pure Appl. **3**, 342 (1838).
[35] W.-H. Steeb, Physica A **95**, 181 (1979).
[36] L. Andrey, Phys. Lett. A **111**, 45 (1985).
[37] L. Andrey, Phys. Lett. A **114**, 183 (1986).
[38] J. D. Ramshaw, Phys. Lett. A **175**, 171 (1993).
[39] A. M. Mariz, Phys. Lett. A **165**, 409 (1992).
[40] G. Mato, private communication.

The quantum canonical ensemble

Dorje C. Brody[a)]
DAMTP, Silver Street, Cambridge CB3 9EW United Kingdom
and Churchill College, Cambridge CB3 0DS United Kingdom

Lane P. Hughston[b)]
Merrill Lynch International, 25 Ropemaker Street, London EC2Y 9LY United Kingdom
and King's College London, The Strand, London WC2R 2LS United Kingdom

(Received 22 June 1998; accepted for publication 19 August 1998)

The phase space Γ of quantum mechanics can be viewed as the complex projective space CP^n endowed with a Kählerian structure given by the Fubini-Study metric and an associated symplectic form. We can then interpret the Schrödinger equation as generating a Hamiltonian dynamics on Γ. Based upon the geometric structure of the quantum phase space we introduce the corresponding natural microcanonical and canonical ensembles. The resulting density matrix for the canonical Γ-ensemble differs from the density matrix of the conventional approach. As an illustration, the results are applied to the case of a spin one-half particle in a heat bath with an applied magnetic field. © *1998 American Institute of Physics.* [S0022-2488(98)00212-6]

I. QUANTUM PHASE SPACE

With a view to certain geometrical generalizations we reformulate the conventional quantum theory in the following manner. Consider a finite dimensional complex Hilbert space $\mathcal{H}_{\mathbf{C}}^{n+1}$ with complex elements Z^α. We identify the state vector Z^α with its multiples λZ^α $(\lambda \in \mathbf{C}-\{0\})$ to obtain the complex projective space CP^n. Here we use Greek indices $(\alpha=0,1,...,n)$ for vectors in $\mathcal{H}_{\mathbf{C}}^{n+1}$, and regard CP^n as the true ''state space'' of quantum mechanics. The complex conjugate of Z^α is written $\overline{Z}_\alpha$, with the Hermitian inner product $Z^\alpha \overline{Z}_\alpha$. We let the state vector Z^α represent homogeneous coordinates for CP^n. Then, the complex conjugate of a point P^α in CP^n corresponds to the hyperplane $\overline{P}_\alpha Z^\alpha = 0$. The points on this plane are the states orthogonal to the original state P^α. Thus, CP^n is equipped with a complex conjugation operation that maps points to hyperplanes of codimension one, and vice versa.

Distinct points X^α and Y^α are joined by a complex projective line $L^{\alpha\beta}=X^{[\alpha}Y^{\beta]}$, representing complex superpositions of the original two states. The quantum mechanical transition probability between two states X^α and Y^α is the cross ratio $\kappa = X^\alpha \overline{Y}_\alpha Y^\beta \overline{X}_\beta / X^\gamma \overline{X}_\gamma Y^\delta \overline{Y}_\delta$. If the system is in the state Y^α and a measurement is made with the projection operator $X^\alpha \overline{X}_\beta / X^\gamma \overline{X}_\gamma$, then κ is the probability that the result is the eigenvalue unity. The angle θ defined by $\kappa = \cos^2(\theta/2)$ can be interpreted as the distance between the states X^α and Y^α. Then if we set $\theta = ds$, $X^\alpha = Z^\alpha$, and $Y^\alpha = Z^\alpha + dZ^\alpha$, retaining terms to second order, we obtain the unitarily invariant Fubini-Study metric[1–3] on the state space, given by

$$ds^2 = 8(Z^\alpha \overline{Z}_\alpha)^{-2} Z^{[\alpha} dZ^{\beta]} \overline{Z}_{[\alpha} d\overline{Z}_{\beta]}. \tag{1}$$

Now suppose we write H^α_β for the Hamiltonian operator, assumed Hermitian. Then, if we set $\hbar = 1$, the Schrödinger equation reads $dZ^\alpha = iH^\alpha_\beta Z^\beta dt$. However, the projective Schrödinger equation

$$Z^{[\alpha} dZ^{\beta]} = iZ^{[\alpha} H^{\beta]}_\gamma Z^\gamma dt, \tag{2}$$

a)Electronic address: d.brody@damtp.cam.ac.uk
b)Electronic address: lane@ml.com

 © 1998 American Institute of Physics

which eliminates the superfluous freedom associated with the scale of Z^α, is defined directly on the state space.

An alternative way of viewing this structure is to regard CP^n as a real manifold Γ of dimension $2n$, equipped with a Riemannian metric g_{ab} and a compatible symplectic structure Ω_{ab}, satisfying $g^{ab}\Omega_{ac}\Omega_{bd}=g_{cd}$ and $\nabla_a\Omega_{bc}=0$, where ∇_a is the metric connection. We use Roman indices ($a=1,2,...,2n$) for tensorial operations in the tangent space of Γ. The Schrödinger equation then takes the form of a Hamiltonian flow $\xi^a=2\Omega^{ab}\nabla_b H$ on Γ. Here the generating function $H(x)$ is given at each point x in Γ by the expectation $H^\alpha_\beta\Pi^\beta_\alpha(x)$ of the Hamiltonian operator, where $\Pi^\beta_\alpha(x)=Z^\beta\overline{Z}_\alpha/Z^\gamma\overline{Z}_\gamma$ is the projection operator for the point $x\in\Gamma$ represented by the state vector $Z^\alpha(x)$. Thus, in quantum mechanics the dynamical trajectories are given by a symplectic flow on the quantum phase space Γ, generated by the Hamiltonian function $H(x)$. The flow is necessarily a Killing field, satisfying

$$\nabla_{(a}\xi_{b)}=0, \tag{3}$$

where $\xi_a=g_{ab}\xi^b$. In other words, the isometries of the Fubini-Study metric on Γ can be lifted to $\mathcal{H}^{n+1}_{\mathbf{C}}$ to yield unitary transformations. Conversely, given the Killing field we recover the observable function $H(x)$ on Γ, up to an additive constant, by use of the relation

$$\Omega^{ab}\nabla_a\xi_b=2(n+1)(H-\overline{H}), \tag{4}$$

where $\overline{H}=H^\alpha_\alpha/(n+1)$ is the average of the eigenvalues of H^α_β.

II. QUANTUM MICROCANONICAL POSTULATES

In this paper we use the metrical symplectic geometry of the quantum phase space as the basis for a re-examination of the traditional hypotheses of quantum statistical mechanics. To this end, we observe that the quantum phase space Γ admits a natural foliation by the $(2n-1)$-dimensional hypersurfaces $\mathcal{E}_E$ determined by level values $H(x)=E$ of the Hamiltonian function. The number of quantum mechanical microscopic configurations (pure states) with expected energy in the small range E to $E+\Delta E$ is then given by $\Omega(E)\Delta E$, where the state density $\Omega(E)$ for energy E is

$$\Omega(E)=\int_{\mathcal{E}_E}\frac{\nabla_a H d\sigma^a}{(\nabla_b H\nabla^b H)}. \tag{5}$$

Here $d\sigma^a=g^{ab}\epsilon_{bc\cdots d}dx^c\cdots dx^d$ is the natural vector-valued $(2n-1)$-form on Γ. In the case of an isolated quantum mechanical system with energy in the given small range, we can adopt the notion of the microcanonical ensemble, and identify the entropy by use of the Boltzmann relation $S(E)=\ln(\Omega(E)\Delta E)$. Here, we implicitly assume what might be called the *quantum microcanonical postulate*, which asserts that for an isolated system in equilibrium all states on a given energy surface in the quantum phase space are equally probable. As a consequence, the temperature of such a system is given by $\beta=dS(E)/dE$, where $\beta=1/kT$. Thus for an isolated quantum system with expected energy E the equilibrium configuration is given by a uniform distribution on the energy surface $\mathcal{E}_E$, with entropy $S(E)$ and temperature $T(E)$ as given above. The corresponding probability density function on Γ, which we call the microcanonical Γ-distribution, is

$$\mu_E(x)=\frac{\delta(H(x)-E)}{\Omega(E)}. \tag{6}$$

It is a straightforward exercise to verify that the state density $\Omega(E)=\int_\Gamma\delta(H(x)-E)dV$, which appears here as a normalization factor, is consistent with expression (5). Here dV is the volume element on Γ.

A general measurable function $F(x)$ on Γ represents a nonlinear observable.[1,5] We interpret $F(x)$ as the conditional expectation $\langle F\rangle_x$ of the observable F in the pure state x. The unconditional expectation of F in the microcanonical Γ-ensemble is then given by $\langle F\rangle_E=\int_\Gamma F(x)\mu_E(x)dV$. In the case of a conventional linear observable we have $F(x)=F^\beta_\alpha\Pi^\alpha_\beta(x)$. Then the unconditional expectation in the state $\mu_E(x)$ is $\langle F\rangle_E=F^\beta_\alpha\mu^\alpha_\beta(E)$, where the quantum microcanonical density matrix $\mu^\alpha_\beta(E)$, parameterized by E, is

$$\mu_\beta^\alpha(E)=\int_\Gamma \Pi_\beta^\alpha(x)\mu_E(x)dV. \tag{7}$$

Providing we only consider linear observables, the state is fully characterized by $\mu_\beta^\alpha(E)$. Now suppose $W(E)$ denotes the total phase space volume for states such that $H(x)\leqslant E$. Then the density matrix $\mu_\beta^\alpha(E)$ can be calculated explicitly by use of the formula

$$\mu_\beta^\alpha(E)=\left(\frac{\partial W(E)}{\partial \overline{H}}\right)^{-1}\frac{\partial W(E)}{\partial H_\alpha^\beta}. \tag{8}$$

III. CANONICAL Γ-ENSEMBLE

Next, we motivate a construction for the corresponding canonical ensemble, representing the situation where a quantum mechanical system is in contact with a heat bath at a fixed temperature. The results we obtain are related to those that follow from the conventional density matrix for the canonical state, though differ in certain essential respects. First we note that the projection operator $\Pi_\beta^\alpha(y)$, for a given state y, can itself be thought of as the density matrix corresponding to a mass distribution $\delta_y(x)$ on Γ, concentrated at the point y. Therefore, we can write $\Pi_\beta^\alpha(y)=\int_\Gamma \Pi_\beta^\alpha(x)\delta_y(x)dV$. Then the conventional density matrix of quantum statistical mechanics, which we denote r_β^α, can be derived from a point mass distribution with density

$$r(x)=Z(\beta)^{-1}\sum_k e^{-\beta E_k}\delta_{y_k}(x) \tag{9}$$

on Γ, concentrated at the energy eigenstates y_k, with E_k being the corresponding eigenvalues and $Z(\beta)=\Sigma_k e^{-\beta E_k}$ the associated partition function. Hence, for the conventional density matrix we have

$$r_\beta^\alpha=Z(\beta)^{-1}\sum_k e^{-\beta E_k}\Pi_\beta^\alpha(y_k). \tag{10}$$

It should be evident that, despite its general usefulness and wide acceptance as a fundamental basis for quantum statistical mechanics, the conventional density matrix r_β^α is in some respects unnatural in the absence of some mechanism forcing the system to energy eigenstates, since its phase space distribution is characterized by $\delta_{y_k}(x)$. Thus we are led to consider an alternative expression, suggested by the geometry of Γ, in which we do not assume a concentration on eigenstates.

Consider two systems I and II and let them make contact to form a combined system I+II. We suppose that the systems only interact weakly, so the support of the resulting state on the quantum phase space Γ_{1+2} of the combined system is concentrated on the product $\Gamma_{1\cdot 2}\subset\Gamma_{1+2}$ of the phase spaces Γ_1 and Γ_2 for the systems I and II. We have in mind the situation where system I represents a heat bath with a given inverse temperature β, while II represents a small system immersed in the bath. We wish to calculate the probability density $p(E_2)$ that system II lies on the energy surface $\mathcal{E}_{E_2}$ in Γ_2, conditional on the total energy of the combined system I+II lying in the small range E to $E+\Delta E$. Now, according to the quantum microcanonical hypothesis, the equilibrium distribution over the energy surface $\mathcal{E}_E$ in $\Gamma_{1\cdot 2}$ is uniform. Therefore, conditional on a given value E for the combined system I+II, the joint probability density function $p_E(E_1,E_2)$ for system I to lie on $\mathcal{E}_{E_1}$ in Γ_1 and for system II to lie on $\mathcal{E}_{E_2}$ in Γ_2 is

$$p_E(E_1,E_2)=\frac{\delta(E-E_1-E_2)\Omega_1(E_1)\Omega_2(E_2)}{\Omega_{1\cdot 2}(E)}. \tag{11}$$

Here, $\Omega_1(E_1)$ and $\Omega_2(E_2)$ are the state densities for the energy surfaces $\mathcal{E}_{E_1}$ in Γ_1 and $\mathcal{E}_{E_2}$ in Γ_2, and the state density $\Omega_{1\cdot 2}(E)$ for the energy surface $\mathcal{E}_E$ in $\Gamma_{1\cdot 2}$ is

$$\Omega_{1\cdot 2}(E)=\int_{-\infty}^{\infty}\Omega_1(E-\epsilon)\Omega_2(\epsilon)d\epsilon. \tag{12}$$

Our goal is to calculate the conditional probability density for system II to have energy E_2. This is given by $p(E_2)=\int_{-\infty}^{\infty}p_E(E_1,E_2)dE_1$, which implies

$$p(E_2)=\frac{\Omega_1(E-E_2)\Omega_2(E_2)}{\Omega_{1\cdot 2}(E)}. \tag{13}$$

Since system I, the heat bath, is in equilibrium, by the microcanonical hypothesis it has entropy $S_1(E_1)=\ln(\Omega_1(E_1)\Delta E)$, and hence temperature $\beta(E_1)=\partial S_1/\partial E_1$. If $E_2\ll E$, then to first order in E_2 we have $\Omega_1(E-E_2)=\Omega_1(E)e^{-\beta E_2}$, where $\beta(E_1)$ is the inverse temperature of the heat bath. As a consequence we deduce that the probability density $p(E_2)$ for system II to lie on the energy surface $\mathcal{E}_{E_2}$ in Γ_2 is given by $p(E_2)=\Omega_2(E_2)e^{-\beta E_2}/\int_{-\infty}^{\infty}\Omega_2(\epsilon)e^{-\beta\epsilon}d\epsilon$. More succinctly, now we drop the subscript 2, and regard the conditioning as implicit in the specification of the parameter β. Then for the probability density of the energy distribution on Γ we have

$$p(\epsilon)=\frac{\Omega(\epsilon)\exp(-\beta\epsilon)}{Z(\beta)}, \tag{14}$$

where $\Omega(\epsilon)$ is the state density on Γ per unit of energy and $Z(\beta)=\int_{-\infty}^{\infty}\Omega(\epsilon)\exp(-\beta\epsilon)d\epsilon$ is the partition function. By the *quantum canonical postulate* we mean the assumption that a small quantum system in equilibrium with a heat bath will be in a state characterized by a distribution over the energy surfaces of Γ with density (14), having a uniform distribution on each such surface. The corresponding probability density function $\rho_\beta(x)$ on Γ is

$$\rho_\beta(x)=\frac{\exp(-\beta H(x))}{Z(\beta)}, \tag{15}$$

which we shall call the *canonical* Γ*-distribution*. Alternatively, we can write $\rho_\beta(x)=\int_\epsilon p(\epsilon)\mu_\epsilon(x)d\epsilon$, expressing $\rho_\beta(x)$ as a weighted average of microcanonical ensembles, where $p(\epsilon)$ is given as in (14). The associated density matrix for the canonical Γ-distribution is therefore

$$\rho_\beta^\alpha(T)=\int_\Gamma \Pi_\beta^\alpha(x)\rho_\beta(x)dV. \tag{16}$$

IV. TWO-STATE SYSTEM

To illustrate how the consequences of the canonical Γ-distribution (15) differ from those of the conventional treatment in terms of the density matrix r_β^α, we study the example of a spin one-half particle in a heat bath. First we review briefly the geometry of the Schrödinger dynamics for a spin one-half particle. The quantum phase space in this case is CP^1, which can be viewed as a sphere S^2. The poles of the sphere correspond to the energy eigenstates, with eigenvalues $E_0=-h$ and $E_1=h$, and the Schrödinger trajectories are generated by a rigid rotation of the sphere about the axis through the poles, the two stationary points. This rotation gives rise to a Killing field on the sphere, where the angular velocity is $2h$, the energy difference between the poles. Thus the Schrödinger evolution generates a latitudinal circle. Now we introduce a complex Hilbert space with coordinates Z^α $(\alpha=0,1)$ which we regard as homogeneous coordinates for CP^1. The complex conjugate of Z^α is the ''plane'' $\overline{Z}_\alpha$ in CP^1, which in this dimension is a point. The point corresponding to $\overline{Z}_\alpha$ is $\overline{Z}^\alpha=\epsilon^{\alpha\beta}\overline{Z}_\beta$, where $\epsilon^{\alpha\beta}$ is the natural symplectic form on the two-dimensional Hilbert space. The formalism in this case is equivalent to the algebra of two-component spinors. By use of the spinor identity $2X^{[\alpha}Y^{\beta]}=\epsilon^{\alpha\beta}X_\gamma Y^\gamma$, where $X^\alpha\epsilon_{\alpha\beta}=X_\beta$, we obtain $ds^2=4Z_\alpha dZ^\alpha Z_\beta d\overline{Z}^\beta/(Z_\gamma\overline{Z}^\gamma)^2$ for the Fubini-Study metric, and $Z_\gamma dZ^\gamma=iH_{\alpha\beta}Z^\alpha Z^\beta dt$ for the projective Schrödinger equation. The Hamiltonian has the representation

$$H_{\alpha\beta}=2h\,\frac{P_{(\alpha}\overline{P}_{\beta)}}{\overline{P}_{\gamma}P^{\gamma}}, \tag{17}$$

where P^{α} and $\overline{P}^{\alpha}$ are the stationary points, satisfying $H^{\alpha}_{\beta}\overline{P}^{\beta}=hP^{\alpha}$ and $H^{\alpha}_{\beta}\overline{P}^{\beta}=-h\overline{P}^{\alpha}$. It follows that the velocity of the trajectory through Γ is

$$\frac{ds}{dt}=2h\,\sin\theta, \tag{18}$$

a special case of the Anandan–Aharonov relation.[4] Here, θ is the distance from Z^{α} to P^{α}, given by the angular coordinate measured down from the north pole. The transition probability from Z^{α} to the north pole P^{α} is $(1+\cos\theta)/2$, and for the evolutionary trajectory we obtain

$$Z^{\alpha}=\cos(\theta/2)e^{i(ht+\phi)}P^{\alpha}+\sin(\theta/2)e^{-i(ht+\phi)}\overline{P}^{\alpha}, \tag{19}$$

where θ and ϕ are the initial coordinates on the sphere for the state at $t=0$. The vector Z^{α} is normalized by setting $P^{\alpha}\overline{P}_{\alpha}=-P_{\alpha}\overline{P}^{\alpha}=1$. The expectation E of the energy for this state is $E=h\,\cos\theta$, and its variance is $h^2\,\sin^2\theta$.

With this in mind we now examine the situation where the spin one-half particle is immersed in a heat bath. The conventional density matrix $r^{\alpha}_{\beta}(T)$ for the system in this case can be expressed as a superposition of projection operators onto energy eigenstates, with Boltzmann weights,

$$r^{\alpha}_{\beta}=\frac{-e^{-\beta E_0}\overline{P}^{\alpha}P_{\beta}+e^{-\beta E_1}P^{\alpha}\overline{P}_{\beta}}{e^{-\beta E_0}+e^{-\beta E_1}}. \tag{20}$$

This can also be written $r^{\alpha}_{\beta}=Z^{-1}\,\exp(-\beta H^{\alpha}_{\beta})$. The fact that r^{α}_{β} has trace unity follows from the normalization condition. For the expected energy $H^{\alpha}_{\beta}r^{\beta}_{\alpha}$ we then obtain $E=-h\,\tanh(\beta h)$.

Next we apply the canonical Γ-distribution to this problem. For this ensemble the probability distribution of the energy in the spin one-half case is

$$p(E)dE=-2\pi Z(\beta)^{-1}\,\sin\theta\,\exp(-\beta h\,\cos\theta)d\theta, \tag{21}$$

where $E=h\,\cos\theta$. The normalization condition on the density function $p(E)$ implies $Z(\beta)=4\pi(\beta h)^{-1}\,\sinh(\beta h)$, from which the expected energy $E=-\partial\,\ln Z/\partial\beta$ can be deduced. The result is $E=kT-\mu B\,\coth(\mu B/kT)$. Here we have reinstated Boltzmann's constant k, the particle's magnetic moment μ, and the external magnetic field strength B, with $h=\mu B$. The behavior of the energy for the cases considered above is sketched in Fig. 1. It is clear that the increase in magnetization, when the temperature decreases, is slower in the case of the canonical Γ-distribution than for the conventional ensemble. This is because the latitudinal circles closer to the equator have larger weights than those closer to the poles. In Fig. 2 we plot the relationship between the specific heat and the temperature.

We note that in the case of the canonical Γ-ensemble the heat capacity for this system is nonvanishing at zero temperature. For many bulk substances, on the other hand, the heat capacity vanishes as zero temperature is approached. Therefore, it would be interesting to enquire if a single electron has a different behavior. To that end, we note that in the case of a general two state system, with energy eigenvalues E_0 and E_1, the density matrix of the canonical Γ-distribution, in the basis of energy eigenstates, is

$$\rho^{\alpha}_{\beta}=-\left[\frac{1}{\beta(E_0-E_1)}+\frac{1}{1-e^{-\beta(E_1-E_0)}}\right]\overline{P}^{\alpha}P_{\beta}+\left[\frac{1}{\beta(E_1-E_0)}+\frac{1}{1-e^{-\beta(E_0-E_1)}}\right]P^{\alpha}\overline{P}_{\beta}. \tag{22}$$

This expression can be contrasted with the conventional result for such a system, given in formula (20). In fact it is not difficult to obtain analogous expressions in higher dimensions, some of which may characterize other systems that can be implemented in experiments to determine whether the Γ-distribution introduced here provides a better account of phenomena in certain situations.

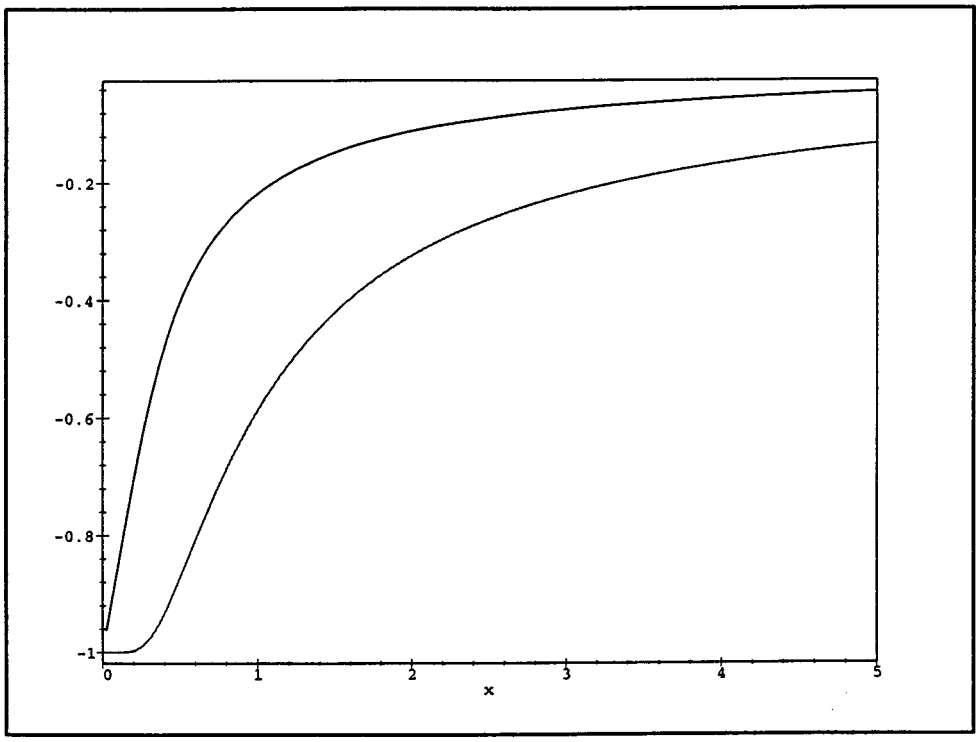

FIG. 1. The energy ratio $E/\mu_e B$ for an electron in a heat bath with temperature T and an applied magnetic field strength B of one Tesla. The x axis is the temperature in degrees Kelvin. Results for the canonical Γ-ensemble (top) and the conventional ensemble (bottom) are shown.

V. DISCUSSION

The foregoing constructions are based on the quantum microcanonical postulate, which for an isolated system in equilibrium implies a uniform distribution over any energy surface of the quantum phase space. We therefore enquire whether this postulate can be derived from the basic principles of quantum mechanics. This would follow, e.g., if quantum dynamics were ergodic on the energy surfaces $\mathcal{E}_E$. In the example of the spin one-half particle, ergodicity is indeed guaranteed by the periodicity of the Schrödinger evolution and the dimensionality of the energy surface.

In higher dimensions, the generic energy surface $\mathcal{E}_E$ is parameterized by $(n-1)$ angular parameters and n phase variables. For each fixed set of angular variables we obtain an n-torus T^n in $\mathcal{E}_E$, and by varying the angular parameters the energy surface is foliated by an $(n-1)$-dimensional family of such tori. If we assume that the Hamiltonian is nondegenerate, and that the ratios E_i/E_j $(i\neq j)$ of the energy eigenvalues are irrational, then the Schrödinger evolution is nonperiodic on each T^n. As a result, ergodicity is guaranteed on the toroidal energy subsurfaces. On the other hand, unitary evolution does not change the angular parameters on $\mathcal{E}_E$. It follows that the energy surfaces are not fully ergodic with respect to the Schrödinger equation, and that we cannot deduce the microcanonical postulate directly from the principles of quantum mechanics. It would be interesting to see if this postulate could be derived as a consequence of a suitable generalization of ordinary quantum mechanics, e.g., nonlinear[1–3,5] or stochastic[6] quantum dynamics.

To summarize, we have formulated quantum mechanical analogs of the microcanonical and canonical ensembles by consideration of the geometry of the quantum phase space. In particular, a key distinction in our proposal is that phase space weightings are fully taken into account in the construction of equilibrium ensembles, whereas in the conventional approach such weights arise only in the event of energy eigenvalue degeneracies. We are thus able to give a coherent meaning to the quantum microcanonical ensemble, and derive an expression for the corresponding density matrix, a construction that is missing in the conventional theory. It may be that for bulk substances there is a further component to the dynamics, perhaps due to internal interactions at low tempera-

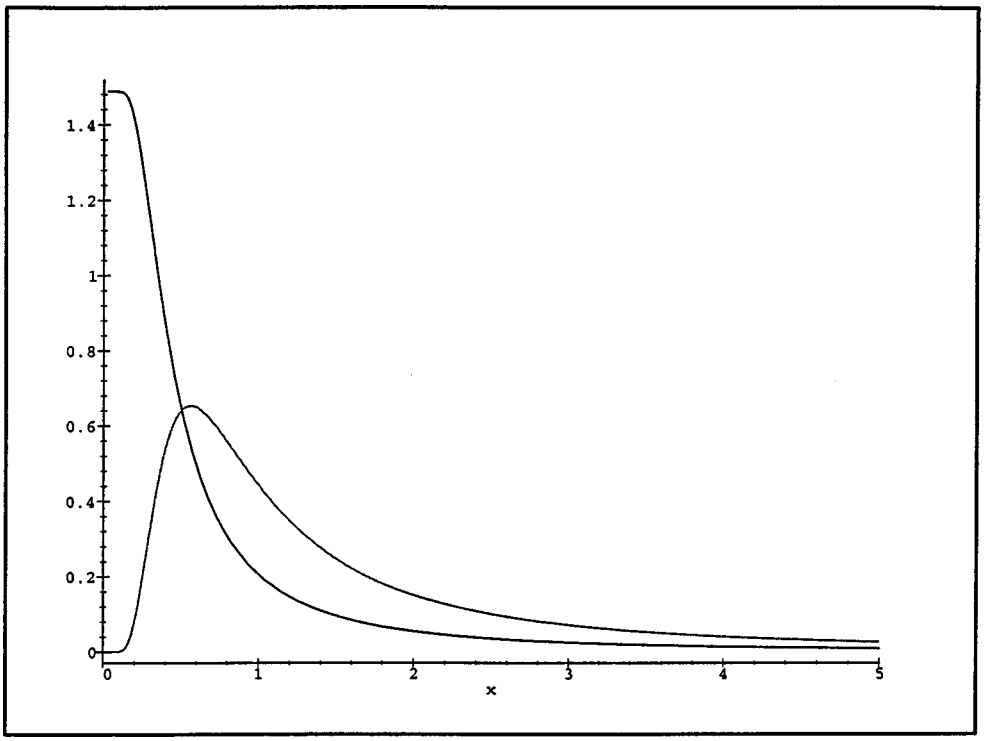

FIG. 2. The heat capacity ratio $C/\mu_e B$ for the system represented in Fig. 1, plotted against the temperature. Results for the canonical Γ-ensemble (top left) and the conventional ensemble (bottom left) are shown.

ture, that forces the quantum phase space to contract down to the lattice of energy eigenstates, with appropriate multiplicities. In that case we recover the conventional theory. It would be interesting to determine, by suitable measurements on quantum mechanical systems, whether the ensembles introduced here are indeed appropriate.

ACKNOWLEDGMENT

D.C.B. is grateful to PPARC for financial support.

[1] T. W. B. Kibble, Commun. Math. Phys. **65**, 189 (1979).
[2] G. W. Gibbons, J. Geom. Phys. **8**, 147 (1992); L. P. Hughston, in *Twistor Theory*, edited by S. Huggett (Marcel Dekker, New York, 1995).
[3] A. Ashtekar and T. A. Schilling, in *On Einstein's Path*, edited by A. Harvey (Springer, Berlin, 1998).
[4] J. Anandan and Y. Aharonov, Phys. Rev. Lett. **65**, 1697 (1990).
[5] S. Weinberg, Phys. Rev. Lett. **62**, 485 (1989); Ann. Phys. N.Y. **194**, 336 (1989).
[6] G. C. Ghirardi, A. Rimini, and T. Weber, Phys. Rev. D **34**, 470 (1989); N. Gisin, Helv. Phys. Acta **62**, 363 (1989); I. C. Percival, Proc. R. Soc. London **447**, 189 (1994); L. P. Hughston, *ibid.* **452**, 953 (1996).

Dirichlet forms and Dirichlet operators for infinite particle systems: Essential self-adjointness

Veni Choi[a)]
Institute for Mathematical Sciences, Yonsei University, Seoul 120-749, Korea

Yong Moon Park[b)]
Department of Mathematics and Institute for Mathematical Sciences, Yonsei University, Seoul 120-749, Korea

Hyun Jae Yoo[c)]
Institute for Mathematical Sciences, Yonsei University, Seoul 120-749, Korea

(Received 15 April 1998; accepted for publication 12 June 1998)

We study the Dirichlet forms and the associated Dirichlet operators for Gibbs measures on infinite particle configuration space. The Dirichlet forms are defined to be ''gradient-type'' forms by introducing a measurable field of rigged Hilbert spaces on the configuration space. Under mild conditions on the interaction including singular potentials, we show that the pre-Dirichlet operator is symmetric and that the closure of the pre-Dirichlet form satisfies the Markovian property. When the interaction is three times differentiable and decreasing sub-exponentially, we show that the Dirichlet operator is essentially self-adjoint on a domain consisting of bounded smooth local functions. © *1998 American Institute of Physics.* [S0022-2488(98)02912-0]

I. INTRODUCTION

In this paper we study the Dirichlet forms and the associated Dirichlet operators for Gibbs measures on the configuration space of interacting particles. More precisely, we consider the infinite particle systems on a Euclidean space $\mathbb{R}^d$ interacting via superstable and lower regular two-body interactions in the sense of Ruelle.[1,2] This ensures[2] the existence of a Gibbs measure μ on the locally finite (particle) configuration space Ω. We define a gradient type pre-Dirichlet form on a domain of bounded smooth cylinder functions from $L^2(\Omega,\mu)$. We show that the associated Dirichlet operator is symmetric and the closure of the form satisfies the Markovian property.[3,4] Furthermore, if the interaction is three times differentiable and decreasing faster than any polynomial, we show that the Dirichlet operator is essentially self-adjoint on a space of bounded smooth local functions. See Sec. II for the details.

The study of this paper is closely related to the following first order stochastic differential equations:

$$\begin{cases} dx_k(t) = -\dfrac{1}{2}\beta \sum\limits_{j\neq k} \operatorname{grad}\Phi(x_k - x_j)dt + dW_k(t), \\ x_k(0) = z_k, \quad k \in S, \end{cases} \tag{1.1}$$

where S is a countable set, $\omega := \{x_k(t)\}_{k\in S}$ is a particle configuration and $\{W_k(t)\}_{k\in S}$ is the independent standard Brownian motion on $\mathbb{R}^d$. In a formal level, any stochastic solution to the system (1.1) defines a diffusion process under which any Gibbs measure with interacton Φ at inverse temperature β is stationary. The study of diffusions associated to the system (1.1) has been initiated by Lang[5,6] for smooth compact support potential Φ using a finite dimensional approximation. This approach had been further developed by Fritz,[7] still in the case of smooth compact

a)Electronic mail: veni@math.yonsei.ac.kr
b)Electronic mail: ympark@bubble.yonsei.ac.kr
c)Present address: Department of Mathematics, Kyungpook National University, Taegu 702-701, Korea; electronic mail: yoohj@phya.yonsei.ac.kr

 © 1998 American Institute of Physics

support potentials and under the restriction $d \leqslant 4$. On the other hand, Osada and Yoshida, in their papers, Ref. 8 and Ref. 9, respectively, have used a Dirichlet form approach to construct the diffusion process corresponding to (1.1). Under mild conditions on the interaction Φ, including singular potentials such as Lenard-Jones type potentials, they have shown that the corresponding canonical Dirichlet form is quasi-regular in the sense defined in Ref. 4 and the existence of a diffusion process is provided by a general theory.[3,4] In order to set the problem in a systematic way, Albeverio, Kondratiev, and Röckner have extensively developed the analytic and geometric theories for the configuration spaces.[10–13] The main idea is a ''lifting procedure'' from underlying one-particle space to the configuration space. From this approach, one considerably benefits in investigating the Dirichlet form theory. In Refs. 12 and 13, the diffusion process is constructed and also the ergodicity of the process has been established (see also Ref. 14).

In Ref. 7, the essential self-adjointness of the Dirichlet operator associated to the Dirichlet form has been studied. This property is used in the proof of the Gaussian central limit theorem for the measures of the density fluctuation fields in Ref. 15. The essential self-adjointness of the Liouville operator corresponding to an infinite Newtonian system has been solved for one-dimensional hard cores system with singular (and of finite range) two-body interactions in Ref. 16. The essential self-adjointness of the Dirichlet operator is also deeply related to the equivalence of the corresponding strong and weak Sobolev spaces.[13,17] Our purpose in this paper is to reformulate the gradient-type Dirichlet forms in terms of a measurable field of rigged Hilbert spaces and then to show the essential self-adjointness of the Dirichlet operators for Gibbs measures of infinite range interactions.

Let us briefly describe the contents of this paper. Let Ω be the space of locally finite particle configurations on $\mathbb{R}^d$, Φ a two-body interaction, and μ a Gibbs measure for the interaction Φ. We will impose Φ to be superstable and lower regular in the sense of Ruelle so that Gibbs measures do exist.[1,2] The main body of this paper consists of (i) developing a measurable field of rigged Hilbert spaces, a gradient operator for functions on Ω and defining a pre-Dirichlet form on $L^2(\Omega,\mu)$ using those ingredients, (ii) showing that the form is closable, and finally (iii) proving the essential self-adjointness of the corresponding Dirichlet operator on a space of smooth local functions. By using the diffentiability concept of functions on Ω developed in Ref. 6, we will define the ''gradient'' ∇f of a function f. ∇f will be understood as an element of a measurable field of Hilbert spaces, so our Dirichlet form shall be of the form

$$\mathcal{E}(u,v) := \frac{1}{2} \int \langle \nabla u(\omega), \nabla v(\omega) \rangle_{\omega,0} d\mu(\omega), \qquad u,v \in C^2_{b,loc}(\Omega), \tag{1.2}$$

where for each $\omega \in \Omega$, $\langle \cdot,\cdot \rangle_{\omega,0}$ denotes the inner product in a Hilbert space $\mathcal{H}_{\omega,0}$ and we will define further Hilbert spaces $\mathcal{H}_{\omega,+}$ and $\mathcal{H}_{\omega,-}$ so that

$$\mathcal{H}_{\omega,+} \subset \mathcal{H}_{\omega,0} \subset \mathcal{H}_{\omega,-} \tag{1.3}$$

shall be a rigged Hilbert space. $C^2_{b,loc}(\Omega)$ denotes the space of bounded, local, twice-differentiable functions on Ω with bounded derivatives. The form (1.2) is essentially equivalent to that defined in Refs. 9 and 13, respectively. One can see the concept of measurable field of Hilbert spaces, i.e., a ''tangent space'' in Refs. 10–13 and also in Ref. 17. The introduction of the rigged Hilbert space (1.3) is typical in this paper and shall be very much useful in showing the essential self-adjointness of Dirichlet operator. This idea has been used for lattice systems.[18–20] Next, we define the Dirichlet operator H_μ:

$$\begin{aligned} H_\mu v(\omega) &:= -\tfrac{1}{2}\Delta v(\omega) - \tfrac{1}{2}\langle b(\omega), \nabla v(\omega) \rangle_{\omega,0}, \qquad v \in C^2_{b,loc}(\Omega), \\ b(\omega) &= \left(-\beta \sum_{y \in \omega \setminus x} \operatorname{grad} \Phi(x-y) \right)_{x \in \omega}, \end{aligned} \tag{1.4}$$

so that

$$\mathcal{E}(u,v) = (u, H_\mu v)_{L^2(\Omega,\mu)}. \tag{1.5}$$

Being a form defined through a positive symmetric operator, it is closable and it is not hard to show the Markovian property of the form under appropriate conditions on Φ. A similar procedure has been done also in Refs. 8–9 and 13. For the essential self-adjointness of the Dirichlet operator, we will use the general parabolic criterion for the essential self-adjointness developed in Ref. 21 and use uniform estimates on partial dynamics for which we have modified the method developed in Ref. 7 to deal with infinite range interactions.

We organize this paper as follows: In Sec. II, we give necessary definitions, review some properties of Gibbs measures, and give the main results of this paper. In Sec. III, we prove the closability and Markovian property of the forms. Section IV is devoted to showing the essential self-adjointness. In the Appendix, we provide crucial, but technical lemmas used in the contents.

In this paper, we improve some part of the doctoral dissertation of one of the authors (Ref. 22) and the results have been announced at the conference at Rome in February, 1997.

II. NOTATIONS, PRELIMINARIES, AND MAIN RESULTS

In this section, we give basic notations, necessary preliminaries, and main results. Related results can be found also in Refs. 8–9 and 12–13. We begin with a configuration space and definition of Gibbs measures.

A. Configuration space and Gibbs measures

Let $\mathcal{N}$ be the space of non-negative integer-valued Radon measures on $\mathbb{R}^d$ equipped with the vague topology. It is well known[23] that by this topology $\mathcal{N}$ becomes a Polish space, i.e., a complete separable metric space. An element ω of $\mathcal{N}$ is of the form

$$\omega=\sum_{i\geqslant 1}\delta_{x_i},$$

where $(x_i)_{i\geqslant 1}$ is a finite or infinite sequence and δ_x is the Dirac measure at $x\in\mathbb{R}^d$. Let us define by $\mathcal{C}$ the class of all bounded Borel subsets of $\mathbb{R}^d$. For each $\Delta\in\mathcal{C}$, let $\mathcal{N}_\Delta$ be the totality of $\mathbb{Z}_+$-valued Radon measures on Δ, and define a mapping $p_\Delta:\mathcal{N}\to\mathcal{N}_\Delta$ by

$$p_\Delta(\omega)=\sum_{x_i\in\Delta}\delta_{x_i},\quad \text{for } \omega=\sum_{i\geqslant 1}\delta_{x_i}\in\mathcal{N}. \tag{2.1}$$

Frequently we will use the notation ω_Δ for $p_\Delta(\omega)$. We note that for each $\Delta\in\mathcal{C}$, there is a bijection between $\mathcal{N}_\Delta$ and the topological direct sum $\Sigma_{m=0}^{\infty}\Delta^m$ given by

$$\mathcal{N}_\Delta\ni\sum_{i=1}^{m}\delta_{x_i}\leftrightarrow\{x_1,\ldots,x_m\}\in\Delta^m. \tag{2.2}$$

For each $\Delta\in\mathcal{C}$, let $N(\Delta)$ be the counting variable:

$$N(\Delta)(\omega):=\omega(\Delta)=\operatorname{card}(\omega\cap\Delta)\equiv|\omega\cap\Delta|. \tag{2.3}$$

Let $\mathcal{F}_\Delta$ be the σ-algebra on $\mathcal{N}$ generated by $N(\Lambda)$, $\Lambda\subset\Delta$, i.e., $\mathcal{F}_\Delta$ is generated by the sets of the form

$$\{\omega\in\mathcal{N}:N(\Lambda)(\omega)=m\},\qquad \Lambda\subset\Delta,\Lambda\in\mathcal{C}\quad\text{and}\qquad m=0,1,2,\ldots.$$

We notice that the Borel σ-algebra on $\mathcal{N}$ coincides with the σ-algebra generated by $\cup_{\Delta\in\mathcal{C}}\mathcal{F}_\Delta$, which we will donote by $\mathcal{F}$.[2,23]

Let Ω be a subset of $\mathcal{N}$ defined by

$$\Omega:=\{\omega\in\mathcal{N}:\omega(\{x\})=0\quad\text{or}\quad 1,\forall\, x\in\mathbb{R}^d,\quad\text{and}\quad \omega(\Delta)<\infty\,\forall\,\Delta\in\mathcal{C}\}. \tag{2.4}$$

By abuse of words we will regard $\omega \in \Omega$ as a measure and also as a subset of $\mathbb{R}^d$ consisting of the support of ω. We will use also the same notations $\mathcal{F}$ and $\mathcal{F}_\Delta$, $\Delta \in \mathcal{C}$, for the σ-algebras on Ω whose elements are the intersections of those in $\mathcal{F}$ and $\mathcal{F}_\Delta$, $\Delta \in \mathcal{C}$, with Ω. We write Ω_Λ for $p_\Lambda \Omega$, $\Lambda \in \mathcal{C}$.

Let us now introduce the notion of interactions and Gibbs measures. For $x=(x^1,\ldots,x^d) \in \mathbb{R}^d$ and $r=(r^1,\ldots,r^d) \in \mathbb{Z}^d$ we will use the following norms:

$$|x| := ((x^1)^2 + \cdots + (x^d)^2)^{1/2} \qquad \text{and} \qquad |r| := \max_{1 \le k \le d} |r^k|. \tag{2.5}$$

For each $r=(r^1,\ldots,r^d) \in \mathbb{Z}^d$, define a unit cube $Q(r)$ by

$$Q(r) := \{x=(x^1,\ldots,x^d) \in \mathbb{R}^d : r^i - \tfrac{1}{2} \le x^i < r^i + \tfrac{1}{2}, i=1,\ldots,d\}.$$

We say that a Lebesgue measurable function Φ is a pair potential when it satisfies

$$\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\},$$

$$\Phi(-x) = \Phi(x), \qquad \forall x \in \mathbb{R}^d.$$

We shall assume that the interaction between particles is given by a pair potential, i.e., the energy of m particles $\omega=\{x_1,\ldots,x_m\} \subset \mathbb{R}^d$ is

$$U(\omega) \equiv U(x_1,x_2,\ldots,x_m) = \begin{cases} \dfrac{1}{2} \displaystyle\sum_{i \neq j} \Phi(x_i - x_j), & m=2,3,\ldots, \\ 0, & m=0,1. \end{cases} \tag{2.6}$$

For any $\omega=\{x_1,\ldots,x_m\}$ and $\eta=\{y_1,\ldots,y_n\}$, we write $\omega\eta=\{x_1,\ldots,x_m,y_1,\ldots,y_n\}$ and the interaction energy $W(\omega,\eta)$ between ω and η is defined to satisfy

$$\begin{gathered} U(\omega\eta) = U(\omega) + U(\eta) + W(\omega,\eta), \\ W(\omega,\eta) = \infty, \quad \text{if } U(\omega\eta) = \infty. \end{gathered} \tag{2.7}$$

We will impose the superstability and lower regularity to the pair potential in the sense of Ruelle.[1,2] For $\omega=\{x_i\}_{i \ge 1}$, $n(\omega:r)$ denotes the number of points x_i of ω contained in $Q(r)$, i.e., $n(\omega:r) = \text{card}(\omega \cap Q(r))$. A pair potential Φ is said to be *superstable* if the following holds:

(SS) There exist $A>0$, $B \ge 0$ such that if $\mathcal{R}$ is a finite subset of $\mathbb{Z}^d$ and $\omega=\{x_1,x_2,\ldots,x_m\} \subset \cup_{r \in \mathcal{R}} Q(r)$, then

$$U(\omega) \ge \sum_{r \in \mathcal{R}} [An(\omega:r)^2 - Bn(\omega:r)]. \tag{2.8}$$

A pair potential Φ is said to be *lower regular* if the following holds:

(LR) There exists a decreasing positive function $\Psi : \mathbb{Z}_+ \to \mathbb{R}_+$ such that $\Sigma_{r \in \mathbb{Z}^d} \Psi(|r|) < \infty$ and if $\mathcal{R}$ and $\mathcal{S}$ are finite subsets of $\mathbb{Z}^d$ such that $\omega=\{x_1,x_2,\ldots,x_m\} \subset \cup_{r \in \mathcal{R}} Q(r)$, $\eta=\{y_1,y_2,\ldots,y_n\} \subset \cup_{s \in \mathcal{S}} Q(s)$, then

$$W(\omega,\eta) \ge -\sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \Psi(|s-r|) \left\{ \frac{1}{2} n(\omega:r)^2 + \frac{1}{2} n(\eta:s)^2 \right\}. \tag{2.9}$$

In order to define Gibbs measures we need a reference measure, which is the Poisson measure on $(\Omega,\mathcal{F})$. Let $z>0$ be an arbitrary but fixed number (the chemical potential) and π the Poisson measure on $(\Omega,\mathcal{F})$ with intensity dx, the Lebesgue measure on $\mathbb{R}^d$, and parameter z. For each $\Lambda \in \mathcal{C}$, let π_Λ be the restriction of π to the sub-σ-algebra $\mathcal{F}_\Lambda$. Obviously, we may regard π_Λ as a measure on Ω_Λ, i.e., for any $\mathcal{F}_\Lambda$-measurable function f on Ω,

$$\int_\Omega f(\omega)\pi(d\omega)\equiv\int_{\Omega_\Lambda} f(\omega_\Lambda)\pi_\Lambda(d\omega_\Lambda)=\sum_{n=0}^{\infty} e^{-z|\Lambda|}\frac{z^n}{n!}\int_{\Lambda^n} f(x_1,\ldots,x_n)dx_1\cdots dx_n, \quad (2.10)$$

where $|\Lambda|$ is the volume of Λ. For $N=1,2,\ldots$, let

$$\mathcal{S}_N=\left\{\omega\in\Omega:\forall l\in\mathbb{N},\sum_{\substack{r\in\mathbb{Z}^d\\ |r|\leqslant l}} n(\omega:r)^2\leqslant N^2(2l+1)^d\right\},$$
$$\mathcal{S}=\cup_{N=0}^{\infty}\mathcal{S}_N. \quad (2.11)$$

The following definitions were given by Ruelle[2].

Definition II.1: *A probability measure μ on Ω is said to be regular if there exist $\overline{\gamma}>0$ and $\overline{\delta}>0$ such that*

$$\mu_\Lambda(d\omega_\Lambda)\leqslant\exp\left\{\sum_{r\in\mathbb{Z}^d}(-\overline{\gamma}n(\omega_\Lambda:r)^2+\overline{\delta}n(\omega_\Lambda:r))\right\}\pi_\Lambda(d\omega_\Lambda),$$

where $\Lambda\in\mathcal{C}$ and μ_Λ is the restriction of μ to $\mathcal{F}_\Lambda$.

Definition II.2: *A probability measure μ on Ω is said to be tempered if $\mu(\mathcal{S})=1$.*

Definition II.3: *A probability measure μ on $(\Omega,\mathcal{F})$ is called a Gibbs measure with potential Φ, activity z, and inverse temperature β if it is tempered and satisfies the following Gibbs condition: for any Borel function $f\in L^1(\Omega,\mu)$ and $\Lambda\in\mathcal{C}$,*

$$\int_\Omega f(\omega)\mu(d\omega)=\int_\Omega\mu(d\eta)\frac{1}{Z_\Lambda(\eta)}\int_{\Omega_\Lambda} f(\omega_\Lambda\cup\eta_{\Lambda^c})\exp\{-\beta[U(\omega_\Lambda)+W(\omega_\Lambda,\eta_{\Lambda^c})]\}\pi_\Lambda(d\omega_\Lambda),$$

where $Z_\Lambda(\eta)=\int_{\Omega_\Lambda}\exp\{-\beta[U(\omega_\Lambda)+W(\omega_\Lambda,\eta_{\Lambda^c})]\}\pi_\Lambda(d\omega_\Lambda)$ is the partition function on Λ with boundary condition $\eta\in\Omega$.

For a superstable and lower regular pair potential Φ the existence of a Gibbs measure was shown by Ruelle.[2]

Theorem II.4: *Suppose that Φ is superstable and lower regular. Then, the set $\mathcal{G}^\Phi(z,\beta)$ of Gibbs measures is nonempty, convex, compact in the topology of local uniform convergence, and Choquet simplex. Furthermore, each $\mu\in\mathcal{G}^\Phi(z,\beta)$ is regular.*

For later use, we introduce the concept of a system of density distributions and a correlation function.[1,2]

Definition II.5: *A family (σ_Δ^m) where $m\in\mathbb{Z}_+$ and Δ runs over the bounded Lebesgue measurable subsets of $\mathbb{R}^d$ is called a system of density distributions if it satisfies the following conditions:*

(a) σ_Δ^m is a positive Lebesgue integrable function on Δ^m and

$$\sigma_\Delta^m(x_{i_1},\ldots,x_{i_m})=\sigma_\Delta^m(x_1,\ldots,x_m),$$

for every permutation $(1,\ldots,m)\to(i_1,\ldots,i_m)$.

(b) (Normalization)

$$\sum_{m=0}^{\infty}\frac{1}{m!}\int\sigma_\Delta^m(x_1,\ldots,x_m)dx_1\cdots dx_m=1.$$

(c) (Compatibility) If $\Delta\subset\Delta'$, then

$$\sigma_\Delta^m(x_1,\ldots,x_m)=\sum_{n=0}^{\infty}\frac{1}{n!}\int_{(\Delta'\setminus\Delta)^n}\sigma_{\Delta'}^{m+n}(x_1,\ldots,x_{m+n})dx_{m+1}\cdots dx_{m+n}.$$

The correlation function is defined by

$$\rho(x_1,\ldots,x_m)=\sum_{n=0}^{\infty}\frac{1}{n!}\int_{\Delta^n}\sigma_\Delta^{m+n}(x_1,\ldots,x_{m+n})dx_{m+1}\cdots dx_{m+n},$$

for $x_1,\ldots,x_m\in\Delta$.

Remark II.6: For any system of density distributions (σ_Δ^m)*, one can associate a probability measure* μ *and conversely, any probability measure* μ *determines* (σ_Δ^m) *completely. They are related as follows: for any* $\mathcal{F}_\Delta$*-measurable function f on* Ω*,*

$$\int_\Omega f(\omega)\mu(d\omega)=\sum_{m=0}^{\infty}\frac{1}{m!}\int_{\Delta^m}f(x_1,\ldots,x_m)\sigma_\Delta^m(x_1,\ldots,x_m)dx_1\cdots dx_m.$$

The following proposition was given by Ruelle (Ref. 2, Propositon 5.2, Corollary 5.3).

Proposition II.7: Let the interaction U be associated with a pair potential Φ *which is superstable, lower regular, and a Borel function. Assume that*

$$\int_{\mathbb{R}^d}|1-e^{-\beta\Phi(x)}|dx<\infty.$$

If a probability measure μ *on* $(\Omega,\mathcal{F})$ *is tempered and satisfies the Gibbs condition, then* μ *is associated with a system of density distributions* (σ_Δ^m) *such that the following inequality holds:*

$$\sigma_\Delta^m(x_1,\ldots,x_m)\leqslant\exp\left\{\sum_{r\in\mathbb{Z}^d}(-\overline{\gamma}n(x_1,\ldots,x_m:r)^2+\overline{\delta}n(x_1,\ldots,x_m:r))\right\},$$

for some $\overline{\gamma}>0$ *and* $\overline{\delta}$ *real and the correlation function defined in Definition II.5 satisfies the Mayer equations,*

$$\rho(\omega\eta)=z^m e^{-\beta[U(\omega)+W(\omega,\eta)]}\sum_{p=0}^{\infty}\frac{1}{p!}\int K(\omega,\zeta)\rho(\eta\zeta)dv_1\cdots dv_p, \tag{2.12}$$

where

$$\omega=(x_1,\ldots,x_m),\qquad \eta=(y_1,\ldots,y_n),\qquad \zeta=(v_1,\ldots,v_p),$$

$$K((x_1,\ldots,x_m),\qquad (v_1,\ldots,v_p))=\prod_{j=1}^{p}K((x_1,\ldots,x_m),v_j),$$

$$K((x_1,\ldots,x_m),v)=\exp\left\{-\beta\sum_{i=1}^{m}\Phi(v-x_i)\right\}-1.$$

Moreover, there exists a constant ξ *such that the correlation function also satisfies*

$$\rho(x_1,\ldots,x_m)\leqslant\xi^m,\qquad x_1,\ldots,x_m\in\mathbb{R}^d,\qquad m\geqslant 0. \tag{2.13}$$

B. Differential operators

In this subsection, we recall the notion of the differentiability introduced by Lang[6] and then define the gradient operator on smooth local functions defined on the configuration space Ω. Our definition of the gradient is essentially same as that introduced in Ref. 12. But there is a minor difference, because we introduce a rigging structure in the field of Hilbert spaces.

Following Lang,[6] let us define

$$\begin{aligned}\Omega^*&=\{(x,\omega)\in\mathbb{R}^d\times\Omega:x\in\omega\},\\ \Omega^{**}&=\{(x,y,\omega)\in\mathbb{R}^d\times\mathbb{R}^d\times\Omega:x\in\omega,y\in\omega\backslash x\}.\end{aligned} \tag{2.14}$$

Ω^* and Ω^{**} are equipped with the topologies induced by those of $\mathbb{R}^d\times\Omega$ and $\mathbb{R}^d\times\mathbb{R}^d\times\Omega$, respectively. For $\omega=\{x_i\}_{i\geqslant 1}$ and $x\in\omega$, $\omega\backslash x$ means the set $\{x_i\in\omega: x_i\neq x\}$ and for $y\in\mathbb{R}^d$ and $\Lambda\in\mathcal{C}$, $\omega\cup y$ means $\omega\cup\{y\}$. For a function $u:\Omega\to\mathbb{R}$ and $(x,\omega)\in\Omega^*$, define a function $u_{\omega\backslash x}$ on $\mathbb{R}^d$ by

$$u_{\omega\backslash x}(p)=u((\omega\backslash x)\cup p),\qquad p\in\mathbb{R}^d. \tag{2.15}$$

Also, for $(x,y,\omega)\in\Omega^{**}$, a function $u_{\omega\backslash x\backslash y}$ on $\mathbb{R}^d\times\mathbb{R}^d$ is defined by

$$u_{\omega\backslash x\backslash y}(p,q)=u((\omega\backslash x\backslash y)\cup p\cup q),\qquad p,q\in\mathbb{R}^d. \tag{2.16}$$

Definition II.8: A function $u:\Omega\to\mathbb{R}$ is said to be continuously differentiable if the following conditions are satisfied:

(a) For each $(x,\omega)\in\Omega^*$, the function $u_{\omega\backslash x}:\mathbb{R}^d\to\mathbb{R}$ is differentiable at x. We write

$$D_x u(\omega):=\left(\frac{\partial u_{\omega\backslash x}}{\partial p^1}(x),\ldots,\frac{\partial u_{\omega\backslash x}}{\partial p^d}(x)\right). \tag{2.17}$$

(b) The mapping $(x,\omega)\mapsto D_x u(\omega)$ is continuous on Ω^*.

Definition II.9: A function $u:\Omega\to\mathbb{R}$ is said to be twice continuously differentiable if the following conditions are satisfied:

(a) For each $(x,\omega)\in\Omega^*$ and $(x,y,\omega)\in\Omega^{**}$, $u_{\omega\backslash x}$ and $u_{\omega\backslash x\backslash y}$ are twice continuously differentiable at x and (x,y)*, respectively. We write*

$$\begin{aligned} D^2_{xx}u(\omega)&:=\left(\frac{\partial^2 u_{\omega\backslash x}}{\partial p^l\,\partial p^m}(x)\right)_{1\leqslant l,m\leqslant d},\\ D^2_{xy}u(\omega)&:=\left(\frac{\partial^2 u_{\omega\backslash x\backslash y}}{\partial p^l\,\partial q^m}(x,y)\right)_{1\leqslant l,m\leqslant d}. \end{aligned} \tag{2.18}$$

(b) The mappings $(x,\omega)\mapsto D^2_{xx}u(\omega)$ and $(x,y,\omega)\mapsto D^2_{xy}u(\omega)$ are continuous on Ω^* and Ω^{**}, respectively.

Let us introduce the spaces of functions on Ω. Put

$$\begin{aligned} \mathcal{F}C^\infty_b(\Omega)=\{u\in L^2(\Omega,\mu):\ &\exists l_1,\ldots,l_n\in C^\infty_0(\mathbb{R}^d)\ \text{ and }\ f\in C^\infty_b(\mathbb{R}^n),\\ &\text{such that }\ u(\omega)=f(\langle l_1,\omega\rangle,\ldots,\langle l_n,\omega\rangle)\ \text{ for }\ \omega\in\Omega, \end{aligned} \tag{2.19}$$

where $C^\infty_0(\mathbb{R}^k)$ and $C^\infty_b(\mathbb{R}^k)$ are the totality of infinitely many continuously differentiable functions on $\mathbb{R}^k$ such that partial derivatives of all order have compact supports and are bounded, respectively, and $\langle\cdot,\cdot\rangle$ is the dualization between $C_0(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d)'$, i.e., $\langle l,\omega\rangle=\Sigma_{i\geqslant 1}l(x_i)$ for $\omega=\Sigma_{i\geqslant 1}\delta_{x_i}$. For $k=1,2$, we define

$$\begin{aligned} C^k(\Omega)&=\{u:\Omega\to\mathbb{R}:\quad u\ \text{ is }\ k\text{-times continuously differentiable}\},\\ C^k_{loc}(\Omega)&=\{u\in C^k(\Omega):\quad u\ \text{ is }\ \mathcal{F}_\Lambda\text{-measurable for some }\ \Lambda\in\mathcal{C}\},\\ C^1_{b,loc}(\Omega)&=\{u\in C^1_{loc}(\Omega):\quad u\ \text{ is bounded and }\ \sup_{(x,\omega)\in\Omega^*}|D_x u(\omega)|<\infty\},\\ C^2_{b,loc}(\Omega)&=\{u\in C^2_{loc}(\Omega):\quad u\ \text{ is bounded, }\ \sup_{(x,\omega)\in\Omega^*}|D_x u(\omega)|<\infty, \end{aligned} \tag{2.20}$$

$$\sup_{(x,\omega)\in\Omega^*} |D^2_{xx}u(\omega)|<\infty \quad \text{and} \quad \sup_{(x,y,\omega)\in\Omega^{**}} |D^2_{xy}u(\omega)|<\infty\}.$$

Analogously, we can define the spaces $C^k(\Omega), C^k_{loc}(\Omega)$, etc., for $k\geqslant 3$.

In order to define Dirichlet forms via the so called square field type operator on $L^2(\Omega,\mu)$, where μ is a Gibbs measure, we introduce some Hilbert space structure, a ''tangent space,'' and the notion of the gradient operator. Let $\nu>d$ be an arbitrary but fixed number. For each $\omega\in\Omega$ and $\varepsilon\in\{-,0,+\}$, we define real Hilbert spaces as follows:

$$\mathcal{H}_{\omega,\varepsilon}=\{h:\omega\to\mathbb{R}^d:|h|_{\omega,\varepsilon}<\infty\}. \tag{2.21}$$

Here, the norms are induced from the following inner products: for $\omega\in\Omega$, $\varepsilon\in\{-,0,+\}$ and $h_1, h_2:\omega\to\mathbb{R}^d$:

$$\langle h_1,h_2\rangle_{\omega,\varepsilon}=\sum_{x\in\omega}(1+|x|)^{\varepsilon\nu}\langle h_1(x),h_2(x)\rangle_{\mathbb{R}^d}. \tag{2.22}$$

Notice that for each $\omega\in\Omega$, the inclusions,

$$\mathcal{H}_{\omega,+}\subset\mathcal{H}_{\omega,0}\subset\mathcal{H}_{\omega,-}, \tag{2.23}$$

hold. The embeddings are everywhere dense and $\mathcal{H}_{\omega,0}$ becomes a rigged Hilbert space by $\mathcal{H}_{\omega,+}$ and $\mathcal{H}_{\omega,-}$. For $\varepsilon=+,0,-$, denote $T_\varepsilon(\Omega)=\cup_{\omega\in\Omega}\mathcal{H}_{\omega,\varepsilon}$ and let $\Gamma(T_\varepsilon(\Omega))$ be the space of functions (sections) $f:\Omega\to T_\varepsilon(\Omega)$ such that $f(\omega)\in\mathcal{H}_{\omega,\varepsilon}$ for each $\omega\in\Omega$ and the mapping $\omega\mapsto|f(\omega)|_{\omega,\varepsilon}$ is measurable. We define the gradient operator ∇ as follows:

$$\begin{gathered} D(\nabla)=C^1_{b,loc}(\Omega),\\ \nabla:D(\nabla)\to\Gamma(T_+(\Omega)),\\ u\mapsto\nabla u,\\ \nabla u(\omega)=(D_x u(\omega))_{x\in\omega}, \quad \text{for } \omega\in\Omega. \end{gathered} \tag{2.24}$$

Since u is $\mathcal{F}_\Lambda$ -measurable for some $\Lambda\in\mathcal{C}$, $\nabla u(\omega)$ has at most finitely many nonzero components for each $\omega\in\Omega$ and so $\nabla u(\omega)$ is well-defined. It may be worthy noting that the gradient operator introduced above is essentially the same as that introduced by Albeverio, Kondratiev, and Röckner.[12]

The Laplace operator Δ is defined as follows:

$$\begin{gathered} D(\Delta)=C^2_{b,loc}(\Omega),\\ \Delta:D(\Delta)\to L^2(\Omega,\mu),\\ u\mapsto\Delta u,\\ (\Delta u)(\omega)=\sum_{x\in\omega}\sum_{k=1}^{d}\frac{\partial^2 u_{\omega\setminus x}}{\partial p^k\partial p^k}(x)\\ \equiv\sum_{x\in\omega}\Delta_x u(\omega). \end{gathered} \tag{2.25}$$

In Sec. III, we will show that for any regular Gibbs measure μ and $u\in C^1_{b,loc}(\Omega)$ [resp., $u\in C^2_{b,loc}(\Omega)$], the function $\omega\mapsto|\nabla u(\omega)|_{\omega,+}$ [resp., $\omega\mapsto\Delta u(\omega)$] is an element of $L^2(\Omega,\mu)$.

C. Main results

We begin with a definition of a pre-Dirichlet form. In the sequel, we assume that the hypotheses of Proposition II.7 are satisfied and fix a Gibbs measure μ. We note that μ is regular in the sence of Definition II.1 and the corresponding system of density functions satisfies the boundedness given in the statement of Proposition II.7. Combining these facts we see that the condition (A.2) of Ref. 8 is satisfied and by the same method given in Lemma 2.4 of Ref. 8 we see that $\mathcal{F}C_b^\infty(\Omega)$ and $C_{b,loc}^2(\Omega)$ are dense in $L^2(\Omega,\mu)$. We define a pre-Dirichlet form as follows:

$$D(\mathcal{E}) := C_{b,loc}^2(\Omega),$$

$$\mathcal{E}(u,v) := \frac{1}{2}\int_\Omega \langle \nabla u(\omega), \quad \nabla v(\omega)\rangle_{\omega,0}\mu(d\omega) \tag{2.26}$$

$$= \frac{1}{2}\int_\Omega \sum_{x\in\omega} \langle D_x u(\omega), \quad D_x v(\omega)\rangle_{\mathbb{R}^d}\mu(d\omega).$$

From the regularity of μ (Definition II.1), it is not hard to see that the form is well-defined. In order to show that the pre-Dirichlet form defined in (2.26) is closable and its closure is a Dirichlet form, we will impose further conditions to the interaction function $\Phi(x)$.

Assumption II.10: The pair potential $\Phi(x)$ satisfies the following conditions:

(a) Φ is superstable and lower regular;

(b) $\int_{\mathbb{R}^d}|1-e^{-\beta\Phi(x)}|dx<\infty$;

(c) $\int_{\mathbb{R}^d}|\text{grad }\Phi(x)|^4 e^{-\beta\Phi(x)}\,dx<\infty$;

(d) there exists $d_1>2, c>0$ such that $|\text{grad }\Phi(x)|\leqslant c(1+|x|)^{-\nu}$ for $|x|>d_1$, where $\nu>d$ is the constant given in the definition of Hilbert spaces $\mathcal{H}_{\omega,\varepsilon}$ in (2.21)–(2.22).

Let us define the pre-Dirichlet operator as follows:

$$D(H_\mu) := C_{b,loc}^2(\Omega),$$

$$(H_\mu v)(\omega) := -\tfrac{1}{2}\Delta v(\omega) - \tfrac{1}{2}\langle b(\omega), \nabla v(\omega)\rangle_{\omega,0}, \quad v\in C_{b,loc}^2(\Omega), \tag{2.27}$$

where

$$b(\omega) = \left(-\beta \sum_{y\in\omega\setminus x} (\text{grad}\,\Phi)(x-y)\right)_{x\in\omega}.$$

The following is one of the main results of this paper.

Theorem II.11: *Suppose that the hypotheses of Assumption II.10 are satisfied. Then the pre-Dirichlet operator H_μ in (2.27) is well defined, symmetric, and the following relation:*

$$\mathcal{E}(u,v) = (u, H_\mu v)_{L^2(\Omega,\mu)}, \qquad u,v\in C_{b,loc}^2(\Omega),$$

holds. Furthermore, the pre-Dirichlet form $(\mathcal{E}, C_{b,loc}^2(\Omega))$ is closable and the closure, which we write as $(\mathcal{E}, D(\mathcal{E}))$ again, is a Dirichlet form.

Remark II.12: Since $\mathcal{F}C_b^\infty(\Omega)\subset C_{b,loc}^2(\Omega)$, the results in Theorem II.11 hold for $\mathcal{F}C_b^\infty(\Omega)$ instead of $C_{b,loc}^2(\Omega)$.

The proof of the above theorem will be given in the next section.

Having established a Dirichlet form one is attempted to study whether there exists a diffusion process which is associated with the Dirichlet form $(\mathcal{E}, D(\mathcal{E}))$. In Ref. 4, it was proved that if a Dirichlet form is quasi-regular, there exists a special standard process which is associated with the given Dirichlet form. See Ref. 4 for details. We do not want to go deep into this direction. Instead, we mention that the Dirichlet forms constructed in this paper and in Refs. 8,9 and also in Ref. 13

are all the same. The quasi-regularity of the Dirichlet form has been proved in Refs. 8,9 and Ref. 13. Here, we just state the result. For a technical reason we will regard the form $(\mathcal{E},D(\mathcal{E}))$ as a form on $L^2(\tilde{\Omega},\mu)$, where $\tilde{\Omega}$ is the closure of Ω. We adjoint an extra point $\square$ (the cemetry) to $\tilde{\Omega}$ and write $\tilde{\Omega}_\square := \tilde{\Omega}\cup\{\square\}$. $\tilde{\Omega}_\square$ has a Borel σ-algebra $\mathcal{F}_\square : \mathcal{F}\cup\{F\cup\{\square\}:F\in\mathcal{F}\}$ and any function $u:\tilde{\Omega}\to\mathbb{R}$ is considered as a function on $\tilde{\Omega}_\square$, putting $u(\square)=0$.

Theorem II.13: *(Ref. 8, Theorem 1, Ref. 9, Theorem 3, Ref. 13, Theorem 5.2) Suppose that the hypotheses in Assumption II.10 are satisfied. Then, the Dirichlet form $(\mathcal{E}, D(\mathcal{E}))$ on $L^2(\tilde{\Omega},\mu)$ constructed through Theorem II.11 is quasi-regular, and hence there exists a Markov process $\mathbb{M}:=(X,\mathcal{A},(X_t)_{t\geqslant 0},(P_z)_{z\in\tilde{\Omega}_\square})$ such that $\mathbb{M}$ is associated with $(\mathcal{E}, D(\mathcal{E}))$, i.e.,*

$$(T^t u)(z)=\int u(X_t)dP_z, \qquad \mu-a.a., \qquad z\in\tilde{\Omega}_\square,$$

where $\{T^t\}_{t\geqslant 0}$ is the strongly continuous semigroup associated with $(\mathcal{E},D(\mathcal{E}))$.

Remark II.14: It is obvious that $(\mathcal{E},D(\mathcal{E}))$ has a local property, i.e., $\mathcal{E}(u,v)=0$ for all $u,v\in D(\mathcal{E})$ with $supp[u]\cap supp[v]=0$. Therefore the Markov process in Theorem II.13 is a diffusion process (Refs. 3–4).

From now on we consider the essential self-adjointness of the Dirichlet operator. We will impose smoothness to the interaction function Φ. We begin with introducing an allowed configuration space.[7]

For $\omega=\{x_i\}\in\Omega, m\in\mathbb{R}^d$, and $l>0$, define

$$H(\omega,m,l):=\sum_{k:|x_k-m|\leqslant l}\left\{1+2B+\sum_{\substack{j\neq k\\ j:|x_j-m|\leqslant l}}\Phi(x_k-x_j)\right\}, \tag{2.28}$$

where $B\geqslant 0$ is the constant given in the superstability assumption (2.8) of Φ. This is a weighted sum of potential energy and particle number in the ball $B_l(m)$ with center m and radius l. We note that it is non-negative.

For $t\in\mathbb{R}$, let

$$g(t)=[1+\log(1+t)]^{1/d} \tag{2.29}$$

and

$$\overline{H}(\omega)=\sup_{m\in\mathbb{Z}^d}\sup_{r\in\mathbb{Z}}\{[rg(|m|)]^{-d}H(\omega,m,rg(|m|))+1\}, \tag{2.30}$$

where $\mathbb{N}$ is the set of positive integers.

The set of allowed configurations is defined to be

$$\overline{\Omega}=\{\omega\in\Omega:\overline{H}(\omega)<\infty\}. \tag{2.31}$$

By a superstable estimate (regularity of Gibbs measures in Definition II.1), it is not hard to show that any Gibbs measure is supported on $\overline{\Omega}$.[7] We now introduce the space of functions defined on $\overline{\Omega}$. For $h>0$, let

$$\overline{\Omega}_h=\{\omega\in\overline{\Omega}:\overline{H}(\omega)\leqslant h\},$$

and for a measurable function $u:\overline{\Omega}\to\mathbb{R}$, let

$$\|u\|_h=\sup\{|u(\omega)|:\omega\in\Omega_h\}.$$

We introduce some useful function spaces.[7]

Definition II.15:

$$C(\overline{\Omega})=\{u:\overline{\Omega}\to\mathbb{R}:\quad \text{the restriction of } u \text{ to each nonempty } \overline{\Omega}_h \text{ is uniformly continuous and bounded}\},$$

$$C_b(\overline{\Omega})=\{u\in C(\overline{\Omega}):\quad u \text{ is bounded }\},$$

$$C_p(\overline{\Omega})=\{u\in C(\overline{\Omega}):\quad \|u\|_h \text{ is bounded by a polynomial of } h\},$$

$$C_1(\overline{\Omega})=\{u\in C(\overline{\Omega}):\quad \|u\|_h\leqslant\exp(c_u h^p) \text{ with } p<1\}.$$

Recall that for $\omega=\{x_i\}_{i\geqslant 1}\in\overline{\Omega}$, $D_{x_i}u$ and $D^2_{x_ix_j}u$ denote the vector in $\mathbb{R}^d$ and $d\times d$ matrix of the first and second partial derivatives of u with respect to the coordinates of x_i and x_j, respectively. For the notational convenience, we write

$$D_i\equiv D_{x_i},$$

$$D_iD_j\equiv D^2_{x_ix_j}, \text{ etc.}$$

Definition II;16: Let $D^q_e(\overline{\Omega})$, $q=1,2,3$, *denote the space of functions* $u\in C_1(\overline{\Omega})$ *possessing derivatives of order less than or equal to* q; *each of these derivatives belongs to* $C_1(\overline{\Omega})$. *Moreover, there exists an* $\varepsilon\equiv\varepsilon_u>0$ *such that for any* $h>0$, *we have*

$$\begin{aligned}
&\||D_ku(\omega)|\exp(\varepsilon|x_k|)\|_h\leqslant c_u(\varepsilon,h), \quad \text{for } q=1,2,3,\\
&\||D_jD_ku(\omega)|\exp(\varepsilon|x_k|+\varepsilon|x_j|)\|_h\leqslant c_u(\varepsilon,h), \quad \text{for } q=2,3,\\
&\||D_iD_jD_ku(\omega)|\exp(\varepsilon|x_k|+\varepsilon|x_j|+\varepsilon|x_i|)\|_h\leqslant c_u(\varepsilon,h), \quad \text{for } q=3.
\end{aligned}\tag{2.32}$$

Also, for $q=1,2,3$, *we define the spaces*

$$D^q_{e1}(\overline{\Omega})=\{u\in D^q_e(\overline{\Omega}):\quad \exists p<1 \text{ such that the first } q \text{ inequalities from (2.32) hold with } c_u(\varepsilon,h)=\exp[c_u(\varepsilon)(1+h)^p]\},$$

$$D^q_0(\overline{\Omega})=\{u\in D^q_{e1}(\overline{\Omega}):\quad u \text{ is } \mathcal{F}_\Lambda\text{-measurable for some } \Lambda\in\mathcal{C}\}.$$

We notice that for $q=1,2,3, C^q_{b,loc}(\overline{\Omega})\subset D^q_0(\overline{\Omega})$.

One of the main purposes of this paper is to show the essential self-adjointness of the Dirichlet operator on a proper domain. For the purpose we need to impose a decreasing rate of the interaction according to the interaction distance. Let $\alpha:\mathbb{R}_+\to\mathbb{R}$ be any monotonic increasing and concave function such that

(i) $\alpha(0)\geqslant 1$ and $\alpha(\lambda)\to\infty$ as $\lambda\to\infty$.

(ii) $\alpha'(\lambda)\leqslant[1/(1+\lambda)]\,\alpha(\lambda)$ for $\lambda\geqslant 0$, and there exists a constant $c>0$ such that $\alpha''(\lambda)\geqslant -c[1/(1+\lambda)]$.

We notice that $\alpha(\lambda)$ might be a very slowly increasing function and there are so many such functions. For example, let $l(\lambda):=\log(1+\lambda)$, $\lambda\geqslant 0$. Then, for any $n\geqslant 1$, the function $\alpha(\lambda):=1+\overbrace{l\circ\cdots\circ l}^{n-\text{times}}(\lambda)$ satisfies the above conditions.

Theorem II.17: *Suppose that a potential* Φ *is superstable and lower regular. Furthermore, we assume that* $\Phi\in C^3_b(\mathbb{R}^d)$ *and there exist a constant* $c_0>0$ *and a function* α *that satisfies the conditions* (i) *and* (ii) *above and for all* $x\in\mathbb{R}^d, i,j,k=1,\ldots,d$,

$$\left|\frac{\partial\Phi}{\partial x^i}(x)\right|+\left|\frac{\partial^2\Phi}{\partial x^j\partial x^i}(x)\right|+\left|\frac{\partial^3\Phi}{\partial x^k\partial x^j\partial x^i}(x)\right|\leqslant\exp[-c_0\log(1+|x|^2)\alpha(1+|x|^2)].$$

Then the pre-Dirichlet operator H_μ defined in (2.27) *is essentially self-adjoint on the domain* $D_0^2(\overline{\Omega})$.

The proof of the above Theorem will be given in Sec. IV. The main idea for the proof is to use a general parabolic criterion of essential self-adjointness[21] and a uniform estimates on partial dynamics.[7]

Before closing this subsection, it may worth giving some comments on Theorem II.17.

Remark II.18: (a) We believe that the result in Theorem II.17 holds under the condition that the derivatives of Φ up to third order are bounded by $(1+|x|)^{-2\nu}$ for $\nu>d$. But the method we used in the proof of Theorem II.17 in this paper may not cover the above case.

(b) Since in the proof of the theorem we have used the method of stochastic differential equations,[7] we were not able to handle singular interactions.

(c) We conjecture that $\mathcal{FC}_b^\infty(\Omega)$ is a core for H_μ. Without the proof, Fritz claimed that for $\Phi\in C_0^3(\mathbb{R}^3)$, H_μ is essentially self-adjoint on $\mathcal{FC}_b^\infty(\Omega)$. See Remark 10 of Ref. 7.

III. DIRICHLET FORMS, DIRICHLET OPERATORS, AND THE ASSOCIATED PROCESSES

In this section, we will give the proof of Theorem II.11. We first collect some estimates, which will be used in the proof. Recall that for any $\omega\in\Omega$ and $r\in\mathbb{Z}^d$, $n(\omega:r)=|\omega\cap Q(r)|$.

Lemma III.1: Let μ be a regular Gibbs measure. Then, for any $p\in\mathbb{N}$ and $r\in\mathbb{Z}^d$,

$$\int_\Omega n(\omega:r)^p\mu(d\omega)\leqslant c(p)<\infty.$$

Proof: Note that $n(\omega:r)$ is $\mathcal{F}_{Q(r)}$-measurable. Hence for $Q(r)\subset\Lambda$,

$$\int_\Omega n(\omega:r)^p\mu(d\omega)=\int_{\Omega_\Lambda} n(\omega_\Lambda:r)^p\mu_\Lambda(d\omega_\Lambda),$$

and so the result follows from the regularity of μ (Definition II.1). ■

Lemma III.2: Let ρ be the correlation function defined in Definition II.5. Then, there exists a constant $b\geqslant 0$ such that for any $\{x_1,\ldots,x_m\}\subset\Delta\in\mathcal{C}$, the bound

$$\rho(x_1,\cdots,x_m)\leqslant z\xi^{m-1}e^{b|\Delta|}\exp\left\{-\beta\sum_{k=2}^m\Phi(x_1-x_k)\right\}$$

holds.

Proof: We prove only for $m=2$. The other cases can be done similarly. By (2.12) of Proposition II.7, we have

$$\begin{aligned}\rho(x_1,x_2)&\equiv\rho((x_1),(x_2))\\&=z\exp\{-\beta\Phi(x_1-x_2)\}\sum_{p=0}^\infty\frac{1}{p!}\int_{\Delta^p}K((x_1),(v_1,\ldots,v_p))\\&\quad\times\rho((x_2)(v_1,\ldots,v_p))\,dv_1\cdots dv_p.\end{aligned}$$

By the lower regularity of Φ, we may assume that $\Phi\geqslant -b_1$ for some $b_1\geqslant 0$. Thus we have

$$|K((x_1),(v_1,\ldots,v_p))|=\left|\prod_{j=1}^p[\exp\{-\beta\Phi(v_j-x_1)\}-1]\right|\leqslant(e^{\beta b_1}+1)^p,\tag{3.1}$$

and by (2.13),

$$\rho((x_2),(v_1,\ldots,v_p))\leqslant \xi^{p+1}. \tag{3.2}$$

Combining (3.1) and (3.2), we obtain the result. ■

Proposition III.3: For any $v\in C^2_{b,loc}(\Omega)$, *there exist constants* K_1 *and* K_2 *such that the bounds*

$$\|\Delta v\|_{L^2(\mu)}\leqslant K_1,$$

$$\int_\Omega |\nabla v(\omega)|^4_{\omega,+}\mu(d\omega)\leqslant K_2,$$

hold.

Proof: For a given $v\in C^2_{b,loc}$, let v be $\mathcal{F}_\Lambda$-measurable for some $\Lambda\in\mathcal{C}$. Then

$$\begin{aligned}\|\Delta v\|^2_{L^2(\mu)}&=\int |\Delta v(\omega)|^2\mu(d\omega)\\ &\leqslant\int\left(\sum_{x\in\omega_\Lambda}\left|\Delta_x v(\omega)\right|\right)^2\mu(d\omega)\\ &\leqslant c(v)\int\left(\sum_{i\in\mathbb{Z}^d\cap\Lambda} n(\omega{:}i)\right)^2\mu(d\omega).\end{aligned}$$

By Lemma III.1, it follows that $\|\Delta v\|_{L^2(\mu)}\leqslant K_1$ for some constant K_1. Since

$$|\nabla v(\omega)|^4_{\omega,+}=\left(\sum_{x\in\omega_\Lambda}(1+|x|)^\nu|D_x v(\omega)|^2\right)^2\leqslant c(\Lambda,v)\left(\sum_{i\in\mathbb{Z}^d\cap\Lambda} n(\omega{:}i)\right)^2,$$

the second bound also follows from Lemma III.1. ■

Now we are in a position to prove Theorem II.11.

Proof of Theorem II.11: First, we will show that the pre-Dirichlet operator H_μ is well-defined as an operator on a dense domain $C^2_{b,loc}(\Omega)$ of $L^2(\Omega,\mu)\equiv L^2(\mu)$. For $v\in C^2_{b,loc}(\Omega)$, by the definition (2.27) of H_μ and Schwarz inequality,

$$\|H_\mu v\|_{L^2(\mu)}\leqslant\frac{1}{2}\|\Delta v\|_{L^2(\mu)}+\frac{1}{2}\left(\int_\Omega|b(\omega)|^4_{\omega,-}\,\mu(d\omega)\right)^{1/4}\left(\int_\Omega|\nabla v(\omega)|^4_{\omega,+}\mu(d\omega)\right)^{1/4}.$$

By Proposition III.3, it is enough to show that

$$\int |b(\omega)|^4_{\omega,-}\mu(d\omega)<\infty.$$

We may assume that $v\in C^2_{b,loc}(\Omega)$ is $\mathcal{F}_\Lambda$-measurable for some $\Lambda\in\mathcal{C}$. A direct computation yields

$$\begin{aligned}\int |b(\omega)|^4_{\omega,-}\mu(d\omega)&\leqslant 8\beta^4\Bigg[\int_\Omega\left\{\sum_{x\in\omega}(1+|x|)^{-\nu}\left(\sum_{\substack{y\in\omega\setminus x\\ |x-y|\geqslant d_1}}|\operatorname{grad}\Phi(x-y)|\right)^2\right\}^2\mu(d\omega)\\ &\quad+\int_\Omega\left\{\sum_{x\in\omega}(1+|x|)^{-\nu}\left(\sum_{\substack{y\in\omega/x\\ |x-y|<d_1}}|\operatorname{grad}\Phi(x-y)|\right)^2\right\}^2\mu(d\omega)\Bigg]\\ &\equiv 8\beta^4[I_1+I_2].\end{aligned}$$

Here, $d_1>0$ is the constant given in Assumption II.10(d). First, we consider the term I_1. By Assumption II.10(d), we have

$$I_1 \leqslant c^4 \int_\Omega \left\{ \sum_{x \in \omega} (1+|x|)^{-\nu} \left(\sum_{\substack{y \in \omega \setminus x \\ |x-y| \geqslant d_1}} (1+|x-y|)^{-\nu} \right)^2 \right\}^2 \mu(d\omega)$$

$$\leqslant c_1 \int_\Omega \left\{ \sum_{i \in \mathbb{Z}^d} (1+|i|)^{-\nu} n(\omega:i) \left(\sum_{\substack{j \in \mathbb{Z}^d \\ |i-j| \geqslant 1}} (1+|i-j|)^{-\nu} n(\omega:j) \right)^2 \right\}^2 \mu(d\omega).$$

Using the Schwarz inequality, the fact $\nu > d$, and Lemma III.1, we obtain $I_1 < \infty$.

Next, let us consider the term I_2,

$$I_2 \leqslant c_2 \int_\Omega \left\{ \sum_{i \in \mathbb{Z}^d} (1+|i|)^{-\nu} \sum_{x \in \omega \cap Q(i)} \left(\sum_{\substack{y \in \omega \setminus x \\ |x-y| < d_1}} |\operatorname{grad} \Phi(x-y)| \right)^2 \right\}^2 \mu(d\omega)$$

$$\leqslant c_3 \int_\Omega \left\{ \sum_{i \in \mathbb{Z}^d} (1+|i|)^{-\nu} \sum_{x \in \omega_{\Delta(i)}} \left(\sum_{\substack{y \in \omega_{\Delta(i)} \\ y \neq x}} |\operatorname{grad} \Phi(x-y)| \right)^2 \right\}^2 \mu(d\omega),$$

where, for each $i \in \mathbb{Z}^d$, $\Delta(i)$ is the finite union of unit cubes such that $\Delta(i) \supset B_{d_1+1}(i)$. We use Schwarz inequality and the translation invariance of interaction and Gibbs measures to see that

$$I_2 \leqslant c_4 \int_\Omega \left\{ \sum_{x \in \omega_\Delta} \left(\sum_{\substack{y \neq \epsilon x_\Delta \\ by \in \omega_\Delta}} |\operatorname{grad} \Phi(x-y)| \right)^2 \right\}^2 \mu(d\omega).$$

Here, Δ is a finite union of unit cubes containing $B_{d_1+1}(0)$. Now, we note that the integrand in the above inequality is an $\mathcal{F}_\Delta$-measurable local function, so we can apply the equality in Remark II.6. Hence,

$$I_2 \leqslant c_4 \sum_{m=2}^{\infty} \frac{1}{m!} \int_{\Delta^m} \left\{ m \left(\sum_{k=2}^{m} |\operatorname{grad} \Phi(x_1 - x_k)| \right)^2 \right\}^2 \sigma_\Delta^m(x_1,\ldots,x_m) dx_1 \cdots dx_m$$

$$\leqslant c_4 \sum_{m=2}^{\infty} \frac{m^2}{m!} (m-1)^3 \int_{\Delta^m} \sum_{k=2}^{m} |\operatorname{grad} \Phi(x_1 - x_k)|^4 \sigma_\Delta^m(x_1,\ldots,x_m) dx_1 \cdots dx_m$$

$$= c_4 \sum_{m=2}^{\infty} \frac{m^2}{m!} (m-1)^4 \int_{\Delta^m} |\operatorname{grad} \Phi(x_1 - x_2)|^4 \sigma_\Delta^m(x_1,\ldots,x_m) dx_1 \cdots dx_m$$

$$= c_4 \sum_{m=2}^{5} \frac{m(m-1)^3}{(m-2)!} \int_{\Delta^m} |\operatorname{grad} \Phi(x_1 - x_2)|^4 \sigma_\Delta^m(x_1,\ldots,x_m)\, dx_1 \cdots dx_m$$

$$+ c_4 \sum_{m=6}^{\infty} \frac{m(m-1)^3}{(m-2)!} \int_{\Delta^m} |\operatorname{grad} \Phi(x_1 - x_2)|^4 \sigma_\Delta^m(x_1,\ldots,x_m) dx_1 \cdots dx_m$$

$$\equiv c_4 (I_{2,1} + I_{2,2}).$$

Consider the first term $I_{2,1}$:

$$I_{2,1}\leqslant c_5\sum_{m=2}^{5}\frac{1}{(m-2)!}\int_{\Delta^m}|\mathrm{grad}\,\Phi(x_1-x_2)|^4 s_\Delta^m(x_1,\ldots,x_m)\,dx_1\cdots dx_m$$

$$\leqslant c_5\int_{\Delta^2}|\mathrm{grad}\,\Phi(x_1-x_2)|^4\left\{\sum_{m=2}^{\infty}\frac{1}{(m-2)!}\int_{\Delta^{m-2}}\sigma_\Delta^m(x_1,\ldots,x_m)\,dx_3\cdots dx_m\right\}dx_1\,dx_2$$

$$= c_5\int_{\Delta^2}|\mathrm{grad}\,\Phi(x_1-x_2)|^4\rho(x_1,x_2)dx_1\,dx_2$$

$$\leqslant c_6\int_{\Delta^2}|\mathrm{grad}\,\Phi(x_1-x_2)|^4 e^{-\beta\Phi(x_1-x_2)}dx_1\,dx_2\leqslant c_6|\Delta|\int_{\mathbb{R}^d}|\mathrm{grad}\,\Phi(x)|^4 e^{-\beta\Phi(x)}\,dx.$$

We have used Lemma III.2 in the third inequality. By Assumption II.10(c), $I_{2,1}$ is finite. To prove the finiteness of $I_{2,2}$, note that $m(m-1)^3/(m-2)!\leqslant c_7[1/(m-6)!]$ for some constant $c_7>0$. So, we have

$$I_{2,2}\leqslant c_7\int_{\Delta^6}|\mathrm{grad}\,\Phi(x_1-x_2)|^4$$
$$\times\left\{\sum_{m=6}^{\infty}\frac{1}{(m-6)!}\int_{\Delta^{m-6}}\sigma_\Delta^m(x_1,\ldots,x_6,x_7,\ldots,x_m)\,dx_7\cdots dx_m\right\}dx_1\cdots dx_6.$$

The finiteness of $I_{2,2}$ comes from Lemma III.2 and Assumption II.10(c). Thus, the proof of the fact that $H_\mu v\in L^2(\mu)$ for $v\in C_{b,loc}^2(\Omega)$ is complete.

Now we prove the relation $\mathcal{E}(u,v)=(u,H_\mu v)_{L^2(\mu)}$ for $u,v\in C_{b,loc}^2(\Omega)$. We may assume that both u and v are $\mathcal{F}_\Lambda$-measurable for some $\Lambda\in\mathcal{C}$. By the definiton of $\mathcal{E}$, the Gibbs condition of μ, and the definition (2.10) of the Poisson measure, we see that

$$\mathcal{E}(u,v)=\frac{1}{2}\int\langle\nabla u(\omega),\nabla v(\omega)\rangle_{\omega,0}\mu(d\omega)$$

$$=\frac{1}{2}\int\mu(d\eta)\frac{1}{Z_\Lambda(\eta)}\int_{\Omega_\Lambda}\sum_{x\in\omega_\Lambda}\langle D_x u(\omega_\Lambda),D_x v(\omega_\Lambda)\rangle_{\mathbb{R}^d}$$
$$\times\exp[-\beta(U(\omega_\Lambda)+W(\omega_\Lambda,\eta_{\Lambda^c}))]\pi_\Lambda(d\omega_\Lambda)$$

$$=\frac{1}{2}\int\mu(d\eta)\frac{1}{Z_\Lambda(\eta)}\sum_{n=0}^{\infty}e^{-z|\Lambda|}\frac{z^n}{n!}$$
$$\times\int_{\Lambda^n}\sum_{i=1}^{n}\langle D_{x_i}u(\{x_1,\ldots,x_n\}),D_{x_i}v(\{x_1,\ldots,x_n\})\rangle_{\mathbb{R}^d}$$
$$\times\exp\{-\beta[U(x_1,\ldots,x_n)+W(\{x_1,\ldots,x_n\},\eta_{\Lambda^c})]\}\,dx_1\cdots dx_n.$$

Applying an integration by parts formula and repeating the above procedure conversely, we obtain the desired equality. By this relation, it follows immediately that H_μ is a positive-definite symmetric operator and so it is closable.

To prove the Markovian property of $\mathcal{E}$, it suffices[4] to consider functions $u\in C_{b,loc}^2(\Omega)$. For each $\varepsilon>0$, choose $\phi_\varepsilon\in C^\infty(\mathbb{R})$ such that $\phi_\varepsilon(t)=t$ for $0\leqslant t\leqslant 1$, $-\varepsilon\leqslant\phi_\varepsilon(t)\leqslant 1+\varepsilon$ for $t\in\mathbb{R}$ and $0\leqslant\phi_\varepsilon(t)-\phi_\varepsilon(s)\leqslant t-s$ for $s<t$. Then for $u\in C_{b,loc}^2(\Omega)$, $\phi_\varepsilon\circ u\in C_{b,loc}^2(\Omega)$ and by a chain rule,

$$\langle\nabla(\phi_\varepsilon\circ u)(\omega),\nabla(\phi_\varepsilon\circ u)(\omega)\rangle_{\omega,0}=(\phi_\varepsilon'(u(\omega)))^2\langle\nabla u(\omega),\nabla u(\omega)\rangle_{\omega,0}\leqslant\langle\nabla u(\omega),\nabla u(\omega)\rangle_{\omega,0}.$$

So, we obtain $\mathcal{E}(\phi_\varepsilon\circ u,\phi_\varepsilon\circ u)\leqslant\mathcal{E}(u,u)$ for $u\in C_{b,loc}^2(\Omega)$. Thus the proof of Theorem II.11 is complete. ■

IV. ESSENTIAL SELF-ADJOINTNESS OF THE DIRICHLET OPERATOR

In this section we show the essential self-adjointness of the Dirichlet operator on a suitable space of local functions, i.e., we prove Theorem II.17. The main ingredient of the proof is the general parabolic criterion developed in Ref. 21. This is an approximate argument. The dynamics of partial systems shall play the role of an approximation.[7] Thus uniform estimates for partial systems are needed. It was extensively investigated by Dobrushin and Fritz.[7,24,25] We will modify it in order to include the case of infinite range interactions. Throughout this section we assume that the interaction Φ satisfies the hypotheses of Theorem II.17. We fix a corresponding Gibbs measure μ, which is a probability measure on unlabeled particle configurations. In order to discuss the dynamics of a given set of particles, it is convenient to label each particle by a given index set. We begin by introducing a partial dynamics.

Let $C(\mathbb{R}_+,\mathbb{R}^d)$ denote the space of continuous functions of $[0,\infty)$ into $\mathbb{R}^d$ with the usual topology corresponding to uniform convergence on compact sets. For a countable index set S, let $\mathcal{W}=[C(\mathbb{R}_+,\mathbb{R}^d)]^S$ with the product topology, $\mathcal{A}$ the associated Borel σ-algebra, and $\mathcal{A}_t$ the smallest σ-algebra for which the projections $w_k(s), k\in S, s\leqslant t$, of $w\in\mathcal{W}$ are measurable. Let $\{W_k\}_{k\in S}$ be a family of independent standard Wiener processes having values in $\mathbb{R}^d$. We assume that we are given a probability measure P on $\mathcal{A}$ such that the Wiener processes $\{W_k\}_{k\in S}$ are realized as the random element $w=\{W_k\}_{k\in S}$ of $\mathcal{W}$. Henceforth, we denote by E^W the expectation with respect to the measure P.

Let us consider the following system of stochastic differential equations: For $\omega=\{x_i\}_{i\geqslant 1}$,

$$\begin{aligned} dx_k(t)&=\tfrac{1}{2}b_k(\omega)\,dt+dW_k(t),\\ \omega(0)&=z=\{z_k\}_{k\in S}\in\overline{\Omega}, \end{aligned} \tag{4.1}$$

where

$$b_k(\omega)=-\beta\sum_{j\neq k}(\operatorname{grad}\Phi)(x_k-x_j).$$

In the case that Φ is compact supported and nonsingular and $d\leqslant 4$, Fritz proved[7] the existence of a unique tempered solution in the sense that the solution $\omega(t,z,W)\equiv\omega(t)$ satisfies $\overline{H}(\omega(t))<\infty$ a.s. for any finite time $t>0$ whenever $\overline{H}(z)<\infty$. Regarding to the essential self-adjointness of the Dirichlet operator H_μ on a dense subset of $L^2(\mu)$, Fritz gave an approximation criterion and claimed that $\mathcal{F}C_b^\infty(\Omega)$ would be a core for H_μ.[7] This approximation uses a uniform *a priori* bound for the partial dynamics crucially. (See also an argument by Spohn.[15]) We will use also an approximation argument (a general parabolic criterion for essential self-adjointness) developed in Ref. 21. We will use the Fritz' method, but to accommodate for infinite range interactions, we need a proper modification. A partial dynamics is defined as follows: For each $\rho\geqslant 1$, define a cut-off function by

$$f_\rho(x)=f_0\left(\frac{|x|}{\rho}\right),\quad \text{for } x\in\mathbb{R}^d, \tag{4.2}$$

where $f_0:\mathbb{R}\to[0,1]$ is a function such that $f_0\in C^3(\mathbb{R})$ and

$$f_0(x)=\begin{cases}1, & \text{for } |x|\leqslant 1,\\ 0, & \text{for } |x|\geqslant 2.\end{cases} \tag{4.3}$$

For $\omega=\{x_i\}_{i\in S}\in\overline{\Omega}$, we define

$$H_k(\omega):=\sum_{j\neq k}\Phi(x_k-x_j), \tag{4.4}$$

$$b_{k,\rho}(\omega):=e^{\beta H_k(\omega)}D_k[f_\rho^2(x_k)e^{-\beta H_k(\omega)}]=f_\rho^2(x_k)b_k(\omega)+2f_\rho(x_k)(\operatorname{grad}f_\rho)(x_k), \tag{4.5}$$

$$\sigma_{k,\rho}(\omega) := f_\rho(x_k). \tag{4.6}$$

Note that $b_{k,\rho}(\omega)=b_k(\omega)$, $\sigma_{k,\rho}(\omega)=1$ for $|x_k|\leqslant\rho$ and $b_{k,\rho}=\sigma_{k,\rho}=0$ for $|x_k|\geqslant 2\rho$. Comparing with the system of (4.1) the partial dynamics is governed by the following stochastic differential equations: Let $\omega=\{x_k\}_{k\in S}$,

$$\begin{gathered} dx_k(t,\rho)=\tfrac{1}{2}b_{k,\rho}(\omega)dt+\sigma_{k,\rho}(\omega)dW_k(t), \qquad k\in S, \\ \omega(0)=z=\{z_k\}_{k\in S}\in\overline{\Omega}. \end{gathered} \tag{4.7}$$

We notice that by the above system (4.7) of stochastic differential equations the particles outside the ball of radius 2ρ centered at 0 are frozen. Note that for each $\omega\in\overline{\Omega}\subset\mathcal{S}$, $b_{k,\rho}(\omega)$ is finite.[2] So, (4.7) is a finite system. Since $\Phi\in C_b^3(\mathbb{R}^d)$, a unique strong solution $\omega=\omega(t,z,\rho)=(x_k(t,z,\rho))_{k\in S}$ of (4.7) exists[5,6,26,27] and $\omega\in\overline{\Omega}_h$ for some $h>0$.[7] If we put $P_\rho^t\phi(z)=E^W[\phi(\omega(t,z,\rho))]$ for a function $\phi\in L^2(\overline{\Omega},\mu)$, then for each $\rho>1$, $\{P_\rho^t\}_{t\geqslant 0}$ is a strongly continuous semigroup of self-adjoint contractions in $L^2(\overline{\Omega},\mu)$. Denote by $H_{\mu,\rho}$ its infinitesimal generator.

For $u\in D_0^2(\overline{\Omega})$, $H_{\mu,\rho}$ has the following explicit form:

$$(H_{\mu,\rho}u)(\omega)=-\tfrac{1}{2}\Delta_\rho u(\omega)-\tfrac{1}{2}\langle b_\rho(\omega), \quad \nabla u(\omega)\rangle_{\omega,0}, \tag{4.8}$$

where

$$\begin{gathered} \Delta_\rho u(\omega)=\sum_{k\in S} f_\rho^2(x_k)\Delta_{x_k}u(\omega), \\ b_\rho(\omega)=(b_{k,\rho}(\omega))_{k\in S}. \end{gathered} \tag{4.9}$$

It is not hard to see that (the closure of) $H_{\mu,\rho}$ is the infinitesimal generator of the following (pre-)Dirichlet form: For each $\rho>1$,

$$\begin{gathered} D(\mathcal{E}_\rho)=D_0^2(\overline{\Omega}), \\ \mathcal{E}_\rho(u,v)=\frac{1}{2}\int_{\overline{\Omega}}\sum_{x\in\omega}\langle D_x u(\omega), \quad D_x v(\omega)\rangle_{\mathbb{R}^d} f_\rho^2(x)\mu(d\omega), \qquad u,v\in D(\mathcal{E}_\rho). \end{gathered} \tag{4.10}$$

We have also the relation

$$\mathcal{E}_\rho(u,v)=(u,H_{\mu,\rho}v)_{L^2(\overline{\Omega},\mu)}, \qquad u,v\in D_0^2(\overline{\Omega}). \tag{4.11}$$

The following is a general parabolic criterion for essential self-adjointness properly expressed for our use (see Ref. 21, Chapter 5, Theorem 1.10).

Proposition IV.1: For $\rho\geqslant 1$, consider the Cauchy problem:

$$\frac{d}{dt}u_\rho(t)+H_{\mu,\rho}u_\rho(t)=0, \qquad t\in[0,1],$$

$$u_\rho(0)=\phi\in D_0^2(\overline{\Omega}),$$

where $H_{\mu,\rho}$ is an infinitesimal generator of the semigroup P_ρ^t induced from the solution of the stochastic differential equation (4.7) and $\phi\in D_0^2(\overline{\Omega})$. Suppose that there exist strong solutions,

$$u_\rho:[0,1]\to L^2(\overline{\Omega},\mu),$$

such that

(a) $t \to u_\rho(t)$ *is differentiable in* $L^2(\bar{\Omega}, \mu)$,

(b) $\forall \rho > 1$ *and* $\forall t \in [0,1]$, $u_\rho(t) \in D_0^2(\bar{\Omega})$,

(c) $\int_0^1 \| (H_\mu - H_{\mu,\rho}) u_\rho(t) \|_{L^2(\bar{\Omega},\mu)} dt \to 0$ *as* $\rho \to \infty$.

Then, H_μ *is essentially self-adjoint on* $D_0^2(\bar{\Omega})$.

In order to apply the above result, we need a uniform bound estimation for finite systems. For this, we use the method developed in Ref. 7. For $\rho \geqslant 1$, let $\omega(t,z,\rho) = (x_k(t,z,\rho))_{k \in S}$ be the strong solution of the stochastic differential equation (4.7). Let $\bar{D}_n^{(r)}$, $n \in S$, $r = 1,2,\ldots,d$, denote the differential operator with respect to $z_n^{(r)}$ ($\equiv r^{\text{th}}$ component of $z_n = (z_n^{(1)}, \ldots, z_n^{(d)})$) and $\bar{D}_n = (\bar{D}_n^{(1)}, \ldots, \bar{D}_n^{(d)})$. Since the system of stochastic differential equations (4.7) is a finite system with smooth coefficients, the following processes:

$$u_k := \bar{D}_n^{(r)} x_k(t,z,\rho), \qquad v_k := \bar{D}_m^{(s)} u_k = \bar{D}_m^{(s)} \bar{D}_n^{(r)} x_k(t,z,\rho), \text{etc.},$$

are well-defined.[28] Therefore for any $\phi \in D_{e1}^1(\bar{\Omega})$ we may have $\bar{D}_n P_\rho^t \phi(z)$, etc., where $P_\rho^t \phi(z) = E^W[\phi(\omega(t,z,\rho))]$. We need a uniform bound on the estimation $\bar{D}_n P_\rho^t \phi(z)$. It will be bounded by an estimation of the following quantities that have been defined in Ref. 7. For $\rho > 1$, define

$$\begin{aligned} Z_p(t,z) &:= t + \int_0^t \bar{H}(\omega(s,z,\rho))^{p/2} ds, \\ V_p(t,z) &:= 1 + Z_p(t,z) + \bar{H}(\omega(t,z,\rho))^{p/2}. \end{aligned} \tag{4.12}$$

For $\omega = (x_i)_{i \geqslant 1}$, we let

$$\begin{aligned} q_k(\omega) &= \bar{H}(\omega)^{1/2} (1 + \log(1 + |x_k|^2))^{1/2}, \\ \dot{Z}_p &= \frac{\partial Z_p}{\partial t}. \end{aligned} \tag{4.13}$$

For the notational brevity, let us write

$$F(t,z;p) := V_p(t,z) + \frac{Z_1(t,z)^2}{h(Z_1(t,z))^{1/2}}, \tag{4.14}$$

where $h(\lambda) := \log(e + \alpha(\lambda))$ and $\alpha(\lambda)$ is the function introduced in Theorem II.17. The bounds in the following proposition corresponds to those in Proposition 4 and Proposition 6 of Ref. 7.

Proposition IV.2: Suppose that the hypotheses of Theorem II.17 are satisfied. Then, we can find a $p \in (1,2)$, $K > 0$, *such that the bounds,*

$$|\bar{D}_n P_\rho^t \phi(z)| \leqslant (1 + |z_n|^2)^{-\nu} \{E^W[\exp[KF(t,z;p)]]\}^{1/2}$$

$$|\bar{D}_m \bar{D}_n P_\rho^t \phi(z)| \leqslant (1 + |z_m|^2)^{-\nu} (1 + |z_n|^2)^{-\nu} \{E^W[\exp[KF(t,z;p)]]\}^{1/2},$$

hold uniformly in $\rho \geqslant 1$.

The proof of the above proposition will be given in the Appendix.

Lemma IV.3: For any $p \in (1,2)$, $K > 0$, *and* $0 \leqslant t \leqslant T$,

$$\int_{\bar{\Omega}} E^W[\exp[KF(t,z;p)]] \mu(dz) \leqslant C_p(T,K) < \infty,$$

where C_p *does not depend on* ρ *and it is a nondecreasing function with respect to its variables* T *and* K.

The proof of this lemma will be given in the Appendix.

We are now ready to give the proof of Theorem II.17.

Proof of Theorem II.17: It is enough to show that all the conditions of Proposition IV.1 are satisfied. So, let $\phi \in D_0^2(\overline{\Omega})$ and $z=\{z_k\}_{k\in S}\in \overline{\Omega}$ be given. Let $\omega(t,z,\rho)=(x_k(t,z,\rho))_{k\in S}$ be the strong solution of the system of stochastic differential equations (4.7). Define $u_\rho(t):=P_\rho^t\phi(z)$ $=E^W[\phi(\omega(t,z,\rho))]$, $t\geqslant 0$. Since the system of stochastic equations (4.7) is a finite system with smooth coefficients, $u_\rho(t)$ solves the Cauchy problem stated in Proposition IV.1. Furthermore, it can be checked that the function $t\mapsto u_\rho(t)$ is differentiable in $L^2(\overline{\Omega},\mu)$.[28] (See also an argument used in Ref. 18, Lemma 5.) In the Appendix, we will show the invariance of $D_0^2(\overline{\Omega})$ under P_ρ^t, i.e., $P_\rho^t D_0^2(\overline{\Omega})\subset D_0^2(\overline{\Omega})$ (Lemma A.4). So, it remains to prove the condition (c) in Proposition IV.1. A direct computation yields

$$\begin{aligned}
&\int_0^1 \|(H_\mu - H_{\mu,\rho})u_\rho(t)\|_{L^2(\overline{\Omega},\mu)}dt \\
&\quad\leqslant \frac{1}{2}\Bigg[\int_0^1 \|\Delta u_\rho(t)-\Delta_\rho u_\rho(t)\|_{L^2(\overline{\Omega},\mu)}dt + \int_0^1\Bigg(\int_{\overline{\Omega}} |b(z)-b_\rho(z)|_{z,-}^4\mu(dz)\Bigg)^{1/4} \\
&\qquad\times\Bigg(\int_{\overline{\Omega}} |\nabla u_\rho(t)(z)|_{z,+}^4\mu(dz)\Bigg)^{1/4}dt\Bigg] \\
&\quad\equiv \frac{1}{2}\Bigg[\int_0^1 I_1(t)dt + \int_0^1 I_{2,1}(t)^{1/4}I_{2,2}(t)^{1/4}dt\Bigg],
\end{aligned} \tag{4.15}$$

where

$$\begin{aligned}
I_1(t)&=\|\Delta u_\rho(t)-\Delta_\rho u_\rho(t)\|_{L^2(\overline{\Omega},\mu)}, \\
I_{2,1}(t)&=\int_{\overline{\Omega}} |b(z)-b_\rho(z)|_{z,-}^4\mu(dz), \\
I_{2,2}(t)&=\int_{\overline{\Omega}} |\nabla u_\rho(t)(z)|_{z,+}^4\mu(dz).
\end{aligned} \tag{4.16}$$

Let us consider the first term $I_1(t)$. Note that $0\leqslant f_\rho\leqslant 1$ and $|\Delta_{z_k}u(z)|\leqslant c|\overline{D}_k\overline{D}_k u(z)|$ for some constant $c>0$. Since $P_\rho^t\phi\in D_0^2(\overline{\Omega})$ and $D_0^2(\overline{\Omega})\subset D_{e1}^2(\overline{\Omega})$, by the second inequality of Proposition IV.2 and Hölder's inequality, we obtain from (4.16) that

$$\begin{aligned}
I_1(t)^2&=\int_{\overline{\Omega}}\Bigg\{\sum_{|z_k|>\rho}|1-f_\rho^2(z_k)||\Delta_{z_k}P_\rho^t\phi(z)|\Bigg\}^2\mu(dz) \\
&\leqslant \int_{\overline{\Omega}}\Bigg\{\sum_{|z_k|>\rho}|\Delta_{z_k}P_\rho^t\phi(z)|\Bigg\}^2\mu(dz) \\
&\leqslant c^2\int_{\overline{\Omega}}\Bigg\{\sum_{|z_k|>\rho}(1+|z_k|^2)^{-2\nu}(E^W[\exp[KF(t,z;p)]])^{1/2}\Bigg\}^2\mu(dz) \\
&\leqslant c^2\Bigg(\int_{\overline{\Omega}}E^W[\exp[2KF(t,z;p)]]\,\mu(dz)\Bigg)^{1/2}\Bigg(\int_{\overline{\Omega}}\Bigg\{\sum_{|z_k|>\rho}(1+|z_k|^2)^{-2\nu}\Bigg\}^4\mu(dz)\Bigg)^{1/2}.
\end{aligned}$$

By Lemma IV.3, the first integration of the last expression is not greater than $C_p(1,2K)<\infty$ for $0\leqslant t\leqslant 1$. We use the regularity of Gibbs measures (Definition II.1) to show the finiteness of the second integral as follows:

$$\int_{\bar{\Omega}}\left\{\sum_{|z_k|>\rho}(1+|z_k|^2)^{-2\nu}\right\}^4\mu(dz)$$

$$\leqslant c_1\int_{\bar{\Omega}}\left\{\sum_{\substack{|i|\geqslant\rho-1\\ i\in\mathbb{Z}^d}}(1+|i|^2)^{-2\nu}n(z{:}i)\right\}^4\mu(dz)$$

$$\leqslant c_1\int_{\bar{\Omega}}\left(\sum_{\substack{|i|\geqslant\rho-1\\ i\in\mathbb{Z}^d}}(1+|i|^2)^{-\frac{4}{3}\nu}\right)^3\left(\sum_{\substack{|i|\geqslant\rho-1\\ ie\mathbb{Z}^d}}(1+|i|^2)^{-4\nu}n(z{:}i)^4\right)\mu(dz)$$

$$\leqslant c_2\sum_{\substack{|i|\geqslant\rho-1\\ i\in\mathbb{Z}^d}}(1+|i|^2)^{-4},$$

where c_1 and c_2 are appropriate constants. Therefore,

$$\int_0^1 I_1(t)dt\leqslant c_3\left(\sum_{\substack{|i|\geqslant\rho-1\\ i\in\mathbb{Z}^d}}(1+|i|^2)^{-4\nu}\right)^{1/4}\to 0,\quad \text{as } \rho\to\infty. \tag{4.17}$$

We next consider $I_{2,1}(t)$. Recall the definitions of $b(z)$ and $b_\rho(z)$ in (2.27) and (4.9), respectively. It follows from (4.16) that

$$I_{2,1}(t)=\int_{\bar{\Omega}}\left\{\sum_{\substack{z_k\in z\\ |z_k|>\rho}}(1+|z_k|)^{-\nu}|b_k(z)-b_{k,\rho}(z)|^2\right\}^2\mu(dz).$$

By the hypotheses of Theorem II.17, we see that

$$I_{2,1}(t)\leqslant c_4\int_{\bar{\Omega}}\left\{\sum_{\substack{z_k\in z\\ |z_k|>\rho}}(1+|z_k|)^{-\nu}|b_k(z)|^2\right\}^2\mu(dz)$$

$$\leqslant c_5\int_{\bar{\Omega}}\left\{\sum_{\substack{z_k\in z\\ |z_k|>\rho}}(1+|z_k|)^{-\nu}\times\left(\sum_{z_j\neq z_k}\exp[-c_0\log(1+|z_k-z_j|^2)\alpha(1+|z_k-z_j|^2)]\right)^2\right\}^2\mu(dz).$$

As in the case of $I_1(t)$, we replace the summations w.r.t. z_k and z_j by summations w.r.t. $k,j\in\mathbb{Z}^d$. Then

$$I_{2,1}(t)\leqslant c_6\int_{\bar{\Omega}}\sum_{\substack{k\in\mathbb{Z}^d\\|k|\geqslant\rho-1}}(1+|k|)^{-\nu}n(z{:}k)^2\mu(dz)$$
$$\times\int_{\bar{\Omega}}\left[\sum_{\substack{k\in\mathbb{Z}^d\\|k|\geqslant\rho-1}}(1+|k|)^{-\nu}\left\{\sum_{j\in\mathbb{Z}^d}\exp\left[-\frac{2}{3}c_0\log(1+|k-j|^2)\mathring{a}(1+|k-j|^2)\right]\right\}\right]^3$$
$$\times\left\{\sum_{j\in\mathbb{Z}^d}n(z{:}j)^4\exp[-2c_0\log(1+|k-j|^2)\alpha(1+|k-j|^2)]\right\}\Bigg]\mu(dz)$$
$$\leqslant c_7\sum_{\substack{k\in\mathbb{Z}^d\\|k|>\rho-1}}(1+|k|)^{-\nu}. \tag{4.18}$$

For $I_{2,2}(t)$, applying the first inequality of Proposition IV.2 and using the Schwarz inequality, we see that

$$I_{2,2}(t)=\int_{\bar{\Omega}}\left\{\sum_{z_k\in z}(1+|z_k|)^{\nu}\left|\overline{D}_kP_\rho^t\phi(z)\right|_{\mathbb{R}^d}^2\right\}^2\mu(dz)$$
$$\leqslant c_8\int_{\bar{\Omega}}\left\{\sum_{z_k\in z}(1+|z_k|)^{\nu}(1+|z_k|^2)^{-2\nu}E^W[\exp[KF(t,z;p)]]\right\}^2\mu(dz)$$
$$\leqslant c_8\left(\int_{\bar{\Omega}}E^W[\exp[4KF(t,z;p)]]\mu(dz)\right)^{1/2}\left(\int_{\bar{\Omega}}\left(\sum_{k\in\mathbb{Z}^d}(1+|k|^2)^{-2\nu+\nu/2}\right)\right.$$
$$\left.\times\left(\sum_{k\in\mathbb{Z}^d}(1+|k|^2)^{-2\nu+\nu/2}n(z{:}k)^4\right)\mu(dz)\right)^{1/2}<\infty. \tag{4.19}$$

By (4.18) and (4.19),

$$\int_0^1 I_{2,1}(t)^{1/4}I_{2,2}(t)^{1/4}\,dt\leqslant c_{10}\left(\sum_{\substack{k\in\mathbb{Z}^d\\|k|>\rho-1}}(1+|k|)^{-\nu}\right)^{1/4}\to 0,\qquad \text{as } \rho\to\infty. \tag{4.20}$$

By (4.15), (4.17), and (4.20) we obtain

$$\int_0^1\|(H_\mu-H_{\mu,\rho})u_\rho(t)\|_{L^2(\bar{\Omega},\mu)}dt\to 0,\qquad \text{as } \rho\to\infty.$$

The proof of Theorem II.17 is complete. ■

APPENDIX: PROOFS OF PROPOSITION IV.2 AND LEMMA IV.3

In this appendix, we provide with the proofs of Proposition IV.2 and Lemma IV.3. The main ideas for the proof of Proposition IV.2 were given in the proof of Proposition 4 and Proposition 6 of Ref. 7. In order to include also the infinite range (sub-exponentially decreasing) interactions, we need proper modifications.

Let $\alpha:\mathbb{R}^+\to\mathbb{R}$ be the function given in Theorem II.17. Define

$$h(\lambda):=\log(e+\alpha(\lambda)),\qquad \lambda\geqslant 0. \tag{A1}$$

Then, h is monotonic increasing and concave. Moreover,

(i) $h(0)\geqslant 1$ and $h(\lambda)\to\infty$ as $\lambda\to\infty$.

(ii) $h'(\lambda)\leqslant 1/(1+\lambda)\leqslant[1/(1+\lambda)]\,h(\lambda),\lambda\geqslant 0$, and there exists a constant $c>0$ such that $h''(\lambda)\geqslant -c[1/(1+\lambda)]$. (A2)

For $x\in\mathbb{R}^d$ and $\lambda\geqslant 0$, define

$$\theta(x,\lambda) := \exp\left[-\frac{\lambda}{h(\lambda)^{1/2}}[1+\lambda^2+\log(1+|x|^2)h(\sqrt{1+|x|^2})]^{1/2}\right](1+|x|^2)^{-\nu}. \tag{A3}$$

The above is a proper modification of the corresponding θ-function in (5.1) of Ref. 7. The followings are crucial lemmas for the estimate of uniform bounds for partial dynamics. Let us denote by $\theta'(x,\lambda)$ the derivative w.r.t. a λ-variable and grad $\theta(x,\lambda), \Delta\theta(x,\lambda)$, the gradient and Laplacian w.r.t. the x-variable, respectively.

Lemma A.1: There exists positive constants c_1, c_2, c_3, and c_4 such that the following holds:

(a) $-\theta'(x,\lambda) \geqslant c_1(1+\log(1+|x|^2))^{1/2}\theta(x,\lambda)$;

(b) $|\text{grad}\,\theta(x,\lambda)| + |\Delta\theta(x,\lambda)| \leqslant c_2\theta(x,\lambda)$;

(c) $\theta(x,\lambda) \leqslant c_3\,\theta(y,\lambda)\exp[c_4\log(1+|x-y|^2)h(1+|x-y|^2)]$.

Proof: (a) Let $G(x,\lambda) := 1+\lambda^2+\log(1+|x|^2)h(\sqrt{1+|x|^2})$. Since $G'(x,\lambda) \geqslant 0$ by (A2) (ii),

$$\begin{aligned} -\theta'(x,\lambda) &\geqslant \left(h(\lambda)^{-1/2} - \tfrac{1}{2}\lambda h(\lambda)^{-3/2}h'(\lambda)\right) G(x,\lambda)^{1/2}\theta(x,\lambda) \\ &\geqslant \tfrac{1}{2}h(\lambda)^{-1/2}G(x,\lambda)^{1/2}\theta(x,\lambda) \\ &\geqslant \tfrac{1}{2}h(G(x,\lambda))^{-1/2}G(x,\lambda)^{1/2}\theta(x,\lambda). \end{aligned}$$

Note that by (A2) (ii), $h(y)^{-1}y$ is nondecreasing and $h(y) \leqslant c(1+\log(1+y))$ for $y \geqslant 0$ and some constant $c>0$. Thus, it is easy to check that

$$h(G(x,\lambda))^{-1}G(x,\lambda) \geqslant h(G(x,0))^{-1}G(x,0) \geqslant c'(1+\log(1+|x|^2)),$$

for some constant $c'>0$. Thus there exists a $c_1>0$ such that

$$-\theta'(x,\lambda) \geqslant c_1(1+\log(1+|x|^2))^{1/2}\theta(x,\lambda).$$

(b) By a direct calculation and the fact that $h(\sqrt{1+|x|^2}) \leqslant c''\log(e+|x|^2)$ for some constant $c''>0$, we obtain the desired result.

(c) Without loss of generality we may assume $|y| \geqslant |x|$. By a fundamental theorem of calculus,

$$\begin{aligned} &\left| \frac{\lambda}{h(\lambda)^{1/2}}G(y,\lambda)^{1/2} - \frac{\lambda}{h(\lambda)^{1/2}}G(x,\lambda)^{1/2}\right| \\ &\quad = \frac{\lambda}{h(\lambda)^{1/2}}\int_{|x|}^{|y|}\frac{d}{du}(1+\lambda^2+\log(1+u^2)h(\sqrt{1+u^2}))^{1/2}du \\ &\quad \leqslant \int_{|x|}^{|y|}\frac{d}{du}(\log(1+u^2)h(\sqrt{1+u^2}))du \\ &\quad = \log(1+|y|^2)h(\sqrt{1+|y|^2}) - \log(1+|x|^2)h(\sqrt{1+|x|^2}). \end{aligned} \tag{A4}$$

Notice that the function $0 \leqslant u \mapsto \log(1+u^2)h(\sqrt{1+u^2})$ is increasing and concave in the region $u \geqslant u_0$ for some constant $u_0>0$. Thus, the last term of (A4) is bounded by

$$\log(1+(u_0+|y|-|x|)^2)h(\sqrt{1+(u_0+|y|-|x|)^2}).$$

Since h and the logarithmic function are concave, we obtain the desired bound. The remaining polynomial part is trivially bounded by the desired term. ■

Lemma A.2: For any $\varepsilon>0$, there exist positive constants c_5 and $c_6(\varepsilon)$ such that the bound

$$e^{-\varepsilon|x|} \leqslant \theta(x,\lambda)\exp\left[c_5\frac{\lambda}{h(\lambda)^{1/2}}(1+\lambda) + c_6(\varepsilon)\right]$$

holds.

Proof: We see that

$$\theta(x,\lambda)^{-1}e^{-\varepsilon|x|}=\exp\left[\frac{\lambda}{h(\lambda)^{1/2}}[1+\lambda^2+\log(1+|x|^2)h(\sqrt{1+|x|^2})]^{1/2}-\varepsilon|x|\right](1+|x|^2)^{\nu}.$$

We devide the x-λ region into two subregions: $\lambda\leqslant|x|^{1/4}$ and $\lambda\geqslant|x|^{1/4}$.

Consider the case $\lambda\leqslant|x|^{1/4}$. Since the function $\lambda\mapsto\lambda/h(\lambda)^{1/2}$ is increasing and $h(\lambda)\geqslant 1$,

$$\begin{aligned}\theta(x,\lambda)^{-1}e^{-\varepsilon|x|}&\leqslant\exp[|x|^{1/4}[1+|x|^{1/2}+\log(1+|x|^2)h(\sqrt{1+|x|^2})]^{1/2}-\varepsilon|x|](1+|x|^2)^{\nu}\\&\leqslant c_6(\varepsilon),\qquad\text{for some } c_6(\varepsilon)>0,\qquad\text{for all } x\in\mathbb{R}^d.\end{aligned}$$

On the other hand, for $\lambda\geqslant|x|^{1/4}$, since $|x|\mapsto\log(1+|x|^2)h(\sqrt{1+|x|^2})$ is increasing,

$$\begin{aligned}\theta(x,\lambda)^{-1}e^{-\varepsilon|x|}&\leqslant\exp\left[\frac{\lambda}{h(\lambda)^{1/2}}[1+\lambda^2+\log(1+\lambda^8)h(\sqrt{1+\lambda^8})]^{1/2}\right](1+\lambda^8)^{\nu}\\&\leqslant\exp\left[c_5\frac{\lambda}{h(\lambda)^{1/2}}(1+\lambda)\right],\end{aligned}$$

for some constant $c_5>0$. ■

Recall the definitions of $b_{k,\rho}(\omega)$ and $\sigma_{k,\rho}(\omega)$ in (4.5) and (4.6). Let $b_{kj}(\omega,\rho)=D_jb_{k,\rho}(\omega)$, $\sigma_{kj}(\omega,\rho)=D_j\sigma_{k,\rho}(\omega)$, $b_{kji}(\omega,\rho)=D_iD_jb_{k,\rho}(\omega)$, and $\sigma_{kji}(\omega,\rho)=D_iD_j\sigma_{k,\rho}(\omega)$. σ_{kji} is a matrix, while b_{kji} is a vector of $d\times d$ matrices, namely $b_{kji}^{(r)}=D_iD_j\langle b_{k,\rho},e_r\rangle$ for $r=1,2,\ldots,d$. Recall the definition of $q_k(\omega)$ in (4.13). We need some estimates.

Lemma A.3: Suppose that the hypotheses of Theorem II.17 are satisfied. Then for any $\gamma>0$ and $i,j,k\in S$, we have a uniform constant $K_1>0$ such that the following bounds hold:

(a) $\Sigma_{j\in S}|b_{kj}(\omega,\rho)|\exp[\gamma\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)]\leqslant K_1q_k(\omega)$;

(b) $\Sigma_{k\in S}|b_{kj}(\omega,\rho)|\exp[\gamma\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)]\leqslant K_1q_j(\omega)$;

(c) $\Sigma_{j,i}|b_{kji}(\omega,\rho)|\exp[\gamma\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)+\gamma\log(1+|x_k-x_i|^2)h(1+|x_k-x_i|^2)]\leqslant K_1q_k(\omega)$;

(d) $\Sigma_k|b_{kji}(\omega,\rho)|\exp[\gamma\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)+\gamma\log(1+|x_k-x_i|^2)\,h(1+|x_k-x_i|^2)]\leqslant K_1q_i(\omega)^{1/2}q_j(\omega)^{1/2}$;

(e) $\Sigma_{j\in S}|\sigma_{kj}(\omega,\rho)|\leqslant K_1$;

(f) $\Sigma_{k\in S}|\sigma_{kj}(\omega,\rho)|\leqslant K_1$;

(g) $\Sigma_{j,i}|\sigma_{kji}(\omega,\rho)|\leqslant K_1$;

(h) $\Sigma_k|\sigma_{kji}(\omega,\rho)|\leqslant K_1$.

Proof: We prove only (a). The bounds (e)–(h) are direct consequences of the definition of $\sigma_k(\omega,\rho)$ in (4.6). The bounds (b)–(d) can be proved by the method used for (a). Recall the definition of $n(\omega:j)$. We have for any $j,k\in\mathbb{Z}^d$,

$$\begin{aligned}n(\omega:j)&=g(|j|)^{-d/2}n(\omega:j)g(|j|)^{d/2}\\&\leqslant[g(|j|)^{-d}n(\omega:j)^2+1]^{1/2}(1+\log(1+|j|))^{1/2}\\&\leqslant c_7\overline{H}(\omega)^{1/2}(1+\log(1+|x_k|))^{1/2}(1+\log(1+|x_k-j|))^{1/2}\\&=c_7q_k(\omega)(1+\log(1+|x_k-j|))^{1/2}.\end{aligned}\tag{A5}$$

From the definition of $b_{k,\rho}(\omega)$, the assumption for Φ in Theorem II.17, and the inequality in (A5), we get

$$\sum_{j\in S}|b_{kj}(\omega,\rho)|\exp[\gamma\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)]$$

$$\leqslant c_8\sum_{j\in S}\exp[\log(1+|x_k-x_j|^2)(-c_0\alpha(1+|x_k-x_j|^2)+\gamma h(1+|x_k-x_j|^2))]$$

$$\leqslant c_9\sum_{j\in\mathbb{Z}^d}n(\omega:j)\exp[\log(1+|x_k-j|^2)(-c_0\alpha(1+|x_k-j|^2)+\gamma h(1+|x_k-j|^2))]$$

$$\leqslant c_{10}q_k(\omega)\sum_{j\in\mathbb{Z}^d}(1+\log(1+|x_k-j|^2))^{1/2}\exp[\log(1+|x_k-j|^2)$$
$$\times(-c_0\alpha(1+|x_k-j|^2)+\gamma h(1+|x_k-j|^2))]\leqslant c_{11}q_k(\omega).$$

The last inequality was obtained by noting $h(\lambda)=\log(e+\alpha(\lambda))$. ■

We are now prepared to prove Proposition IV.2.

Proof of Proposition IV.2: First, let us prove the first bound in the proposition by using the method similar to that employed in the proof of Proposition 4 of Ref. 7. By a chain rule, we have

$$\bar{D}_n^{(r)}\phi(\omega)=\sum_{k\in S}\langle D_k\phi(\omega),u_k\rangle_{\mathbb{R}^d},$$

where $\omega=\omega(t,z,\rho)\equiv(x_k(t,z,\rho))_{k\in S}$ is the solution of (4.7) and $u_k\equiv\bar{D}_n^{(r)}x_k(t,z,\rho)$. Since $D_{e1}^2(\bar{\Omega})\subset D_{e1}^1(\bar{\Omega})$, there exists an $\varepsilon>0$ and $p_0\in(0,1)$ such that

$$|D_k\phi(\omega)|\leqslant e^{-\varepsilon|x_k|}\exp[c_{\phi,\varepsilon}(1+\bar{H}(\omega))^{p_0}].$$

Since $1+\bar{H}(\omega)^{p/2}=\dot{Z}_p$, $p>0$, we thus have a $p_1\in(1,2)$ such that

$$|\bar{D}_n^{(r)}\phi(\omega)|\leqslant\exp[c_{\phi,\varepsilon}\dot{Z}_{p_1}]\sum_{k\in S}e^{-\varepsilon|x_k|}|u_k|\leqslant\exp[c_{\phi,\varepsilon}\dot{Z}_{p_1}]\left(\sum_{k\in S}e^{-\varepsilon|x_k|}\right)^{1/2}\left(\sum_{k\in S}e^{-\varepsilon|x_k|}|u_k|^2\right)^{1/2}. \tag{A6}$$

On the other hand,

$$\sum_{k\in S}e^{-\varepsilon|x_k|}\leqslant\sum_{k\in\mathbb{Z}^d}n(\omega:k)e^{-\varepsilon|k|}\leqslant c_{12}\sum_{k\in\mathbb{Z}^d}\bar{H}(\omega)^{1/2}e^{-\varepsilon|k|}(1+\log(1+|k|))^{1/2}\leqslant c_{13}\dot{Z}_1. \tag{A7}$$

From (A6), (A7), and Lemma A.2, we obtain

$$|\bar{D}_n^{(r)}\phi(\omega)|\leqslant\exp\left[K_2\dot{Z}_{p_1}+K_2\frac{Z_1}{h(KZ_1)^{1/2}}(1+Z_1)+K_2\right]R_1(t,KZ_1)^{1/2}, \tag{A8}$$

where

$$R_1(t,\lambda)=\sum_{k\in S}\theta(x_k(t,z,\rho),\lambda)|u_k(t,z,\rho)|^2. \tag{A9}$$

The constant K_2 depends on ϕ and ε and K is arbitrary. We want to show that

$$E^W[R_1(t,KZ_1(t,z))]\leqslant(1+|z_n|^2)^{-\nu} \tag{A10}$$

with a large $K>0$. Then, (A8), (A10), and the definition of $F(t,z;p)$ in (4.14) imply the first bound in the proposition. We have to calculate the stochastic differential of R_1 when dx_k is given by (4.7) while,

$$du_k=\sum_{j\in S} b_{kj}u_j dt+\sum_{j\in S}\langle\sigma_{kj},u_j\rangle dW_k(t). \tag{A11}$$

As in Ref. 7, if an $\mathcal{A}_t$-adapted process has a stochastic differential $dI=a\,dt+\Sigma b_k\,dW_k$, then $d_dI:=a\,dt$ denotes the deterministic part of dI. For R_1, by Itô's formula, we have

$$d_dR_1(t,KZ_1)=-\dot{Z}_1I_0dt+(I_1+I_2+I_3+I_4)dt, \tag{A12}$$

where

$$I_n=\sum_{k\in S} I_{nk}, \qquad \text{for } n=0,1,2,3,4,$$

$$I_{0k}=-K\theta_k'|u_k|^2,$$

$$I_{1k}=\left(\langle \text{grad}\,\theta_k,b_{k,\rho}\rangle+\tfrac{1}{2}\Delta\theta_k\sigma_{k,\rho}^2\right)|u_k|^2,$$

$$I_{2k}=2\theta_k\sum_{j\in S}\langle u_k,b_{kj}u_j\rangle,$$

$$I_{3k}=d\,\theta_k\left(\sum_{j\in S}\langle\sigma_{kj},u_j\rangle\right)^2,$$

$$I_{4k}=\langle\text{grad}\,\theta_k,u_k\rangle\sigma_{k,\rho}\sum_{j\in S}\langle\sigma_{kj},u_j\rangle.$$

Note that $I_0=-K(\partial R_1/\partial\lambda)$, thus $I_n\leqslant\frac{1}{4}\dot{Z}_1I_0$, $n=1,2,3,4$, implies (A10) as $R_1(0,0)=(1+|z_n|^2)^{-\nu}$. First, by using Lemma A.1 (b) and the assumption in Theorem II.17, we see that

$$|I_1|\leqslant c_{14}\left(\sum_{k\in S}\sum_{j\in S}\exp[-c_0\log(1+|x_k-x_j|^2)\alpha(1+|x_k-x_j|^2)]\theta_k|u_k|^2+\sum_{k\in S}\theta_k|u_k|^2\right).$$

By a similar method as in the proof of Lemma A.3 (a), we have the bound

$$|I_1|\leqslant c_{15}\sum_{k\in S} q_k(\omega)\theta_k|u_k|^2.$$

Consider the term I_2. By Lemma A.1 (c),

$$\begin{aligned}|I_2|&\leqslant\sum_{k\in S}2\theta_k\sum_{j\in S}|u_k||b_{kj}||u_j|\\&\leqslant\sum_{k\in S}\sum_{j\in S}\theta_k|b_{kj}||u_k|^2+\sum_{k\in S}\sum_{j\in S}\theta_k|b_{kj}||u_j|^2\\&\leqslant\sum_{k\in S}\sum_{j\in S}\theta_k|b_{kj}||u_k|^2+c_3\sum_{k\in S}\sum_{j\in S}\theta_j|b_{kj}||u_j|^2\exp[c_4\log(1+|x_k-x_j|^2)h(1+|x_k-x_j|^2)].\end{aligned}$$

By Lemma A.3 (b) and the same technique used for I_1, we have also

$$|I_2|\leqslant c_{16}\sum_{k\in S} q_k(\omega)\theta_k|u_k|^2.$$

Similarly, we obtain the same inequalities for I_3 and I_4. Thus, there exists a constant $K_0>0$ such that

$$|I_1|+|I_2|+|I_3|+|I_4|\leqslant K_0\sum_{k\in S} q_k(\omega)\theta_k|u_k|^2.$$

On the other hand, by Lemma $A.1$ (a),

$$\begin{aligned}\dot{Z}_1 I_0 &= \dot{Z}_1\left(-K\frac{\partial R_1}{\partial\lambda}\right)\\ &= (1+\overline{H}(\omega)^{1/2})(-K)\sum_{k\in S}\theta_k'|u_k|^2\\ &\geqslant Kc_1\overline{H}(\omega)^{1/2}\sum_{k\in S}(1+\log(1+|x_k|^2))^{1/2}\theta_k|u_k|^2 = Kc_1\sum_{k\in S} q_k(\omega)\theta_k|u_k|^2.\end{aligned}$$

So, in order to have the inequality (A10), it suffices to take $Kc_1\geqslant K_0$. This proves the first inequality.

In order to prove the second inequality, one may use the method employed in the proof of Proposition 6 of Ref. 7 together with necessary bounds in Lemma A.6 and the method used in the above. Since the proof is not so hard to carry out, we leave the details to the reader. ■

We now prove Lemma IV.3.

Proof of Lemma IV.3: We have to prove

$$\int_{\Omega} E^W\left[\exp\left[K\left(V_p(t,z)+\frac{Z_1(t,z)^2}{h(Z_1(t,z))^{1/2}}\right)\right]\right]\mu(dz)<\infty,\qquad \forall K>0,\qquad 1<p<2. \tag{A13}$$

By using Schwarz inequality, it suffices to show that

$$\int_{\Omega} E^W[\exp[K'V_p(t,z)]]\mu(dz)<\infty,\qquad \forall K'>0,\qquad p<1 \tag{A14a}$$

and

$$\int_{\Omega} E^W\left[\exp\left[K'\frac{Z_1(t,z)^2}{h(Z_1(t,z))^{1/2}}\right]\right]\mu(dz)<\infty,\qquad \forall K'>0. \tag{A14b}$$

We will prove only the second bound [(A14b)]. The first one can be proved in a similar manner. We notice that the function $0\leqslant\lambda\mapsto\lambda^2/h(\lambda)^{1/2}$ is convex, so by Jensen's inequality and the invariance of μ w.r.t. P_ρ^t, it reduces to show that for any $K'>0$,

$$\int_{\Omega}\exp\left[K'\frac{(1+\overline{H}(z)^{1/2})^2}{h(1+\overline{H}(z)^{1/2})^{1/2}}\right]\mu(dz)<\infty. \tag{A15}$$

Since $h(u)\to\infty$ as $u\to\infty$, it is sufficient to show that

$$\int_{\Omega}\exp(\lambda\overline{H}(\omega))\mu(d\omega)<\infty, \tag{A16}$$

for sufficiently small $\lambda>0$. A direct computation shows that for any $m\in\mathbb{Z}^d$ and $l>0$,

$$H(\omega,m,l)\leqslant c\sum_{i\in B_l(m)\cap\mathbb{Z}^d}(n(\omega{:}i)^2+1),$$

for some constant $c>0$ independent of m and l. Thus, regularity of μ implies that there exists λ_0 such that for $\lambda\leqslant\lambda_0$ the bounds

$$\int_{\bar{\Omega}} \exp(\lambda H(\omega,m,l))\mu(d\omega) \leqslant \exp(Kl^d) \tag{A17}$$

hold for some constant $K>0$ independent of m and l. The bound (A16) follows from the bounds in (A17) and the argument similar to that used in the proof of Lemma 2.1 of Ref. 16. See also the proof of Proposition 1 in Ref. 24. For the detailed derivation of (A16) from (A17), we refer to Ref. 16. ■

Finally, we provide with the proof of condition (b) in Proposition IV.1, that is, we show the invariance of $D_0^2(\bar{\Omega})$ under P_ρ^t.

Lemma A.4: For any $\rho \geqslant 1$, $t \geqslant 0$, $P_\rho^t(D_0^2(\bar{\Omega})) \subset D_0^2(\bar{\Omega})$.

Proof: Recall $u_\rho(t) := P_\rho^t \phi(z) = E^W[\phi(\omega(t,z,\rho))]$ for a given $\phi \in D_0^2(\bar{\Omega})$. For a fixed $T>0$, consider $t \in [0,T]$. By (4.7), it is clear that $u_\rho(t)$ is also a local function. Thus, it remains only to prove the proper bounds of the derivatives of $u_\rho(t)$ in the sense as defined in Definition II.16.

First, we consider the first order derivative of $u_\rho(t)$. Recall the notation $V_p(t,z) = 1 + \bar{H}^{p/2}(\omega(t,z,\rho)) + t + \int_0^t \bar{H}^{p/2}(\omega(s,z,\rho))ds$ in (4.12). We assert that if we can find constants (that may depend on ρ and T) $\sigma>0$, K, and $p \in (1,2)$, such that the bounds

$$|\bar{D}_n P_\rho^t \phi(z)| \leqslant \exp(-\sigma|z_n|) E^W[\exp(KV_p(t,z))], n \in S, \tag{A18}$$

hold (cf. Proposition 4 of Ref. 7), then the lemma is proved. In fact, we observe that the particle numbers inside the ball $B_{2\rho}(0)$ are conserved because the particles outside the ball are frozen in (4.7). By the superstability and decay property of the interaction Φ, one can check that there exists a positive constant $c(\rho)$ such that the bound

$$\bar{H}(\omega(t,z,\rho)) \leqslant c(\rho)\bar{H}(z), \qquad 0 \leqslant t \leqslant T, \tag{A19}$$

holds. Thus, it follows from (4.19) that

$$E^W[\exp(KV_p(t,z))] \leqslant \exp[K'(1+\bar{H}(z)^{p/2})], \tag{A20}$$

for some constant $K'>0$. Inserting this bound into (A18) we have proven our assertion.

Let us now prove (A18). We emphasize that we can not apply Proposition IV.2 directly. First, if $z_n \notin B_{2\rho}(0)$, then $x_n(t,z,\rho) = z_n$. Thus, since $\phi \in D_0^2(\bar{\Omega})$, we can find constants $\varepsilon>0$, $c_\phi(\varepsilon)$, and $p \in (1,2)$ such that the bounds

$$\begin{aligned} |\bar{D}_n P_\rho^t(z)| &= |E^W[\bar{D}_n \phi(\omega(t,z,\rho))]| \\ &\leqslant e^{-\varepsilon|z_n|} E^W[\exp\{c_\phi(\varepsilon)(1+\bar{H}(\omega(t,z,\rho))^{p/2})\}], \quad z_n \notin B_{2\rho}(0) \end{aligned} \tag{A21}$$

hold.

Next, consider the case in which $z_n \in B_{2\rho}(0)$. Since the particles outside the ball are frozen, we have that

$$u_k := \bar{D}_n^{(r)} x_k(t,z,\rho) = 0, \qquad \text{for } z_n \in B_{2\rho}(0) \text{ and } z_k \notin B_{2\rho}(0).$$

Thus, for $z_n \in B_{2\rho}(0)$, $\bar{D}_n P_\rho^t \phi(z) = E^W[\Sigma_{k \in S: \omega_k \in B_{2\rho}(0)} \langle D_k \phi(\omega), u_k \rangle_{\mathbb{R}^d}]$ and the sum in (A11) is actually a finite sum:

$$du_k = \sum_{\substack{j \in S: \\ z_j \in B_{2\rho}(0)}} b_{kj} u_j \, dt + \sum_{\substack{j \in S: \\ z_j \in B_{2\rho}(0)}} \langle \sigma_{kj}, u_j \rangle dW_k(t), \qquad z_k \in B_{2\rho}(0). \tag{A22}$$

Therefore, for a fixed $\rho>0$, we may effectively consider the system as a finite range interaction system with interaction range $2\rho>0$. Now, the exact method used in the proof of Proposition 4 of Ref. 7 gives the bound in (A18).

For the second order derivatives of $u_\rho(t)$, we use a similar argument as above and the method used in the proof of Proposition 6 of Ref. 7. The proof of the lemma is completed. ■

ACKNOWLEDGMENTS

This work was supported in part by Basic Science Research Institute program, Korean Ministry of Education (BSRI 98-1421), and KOSEF (95-0701-04-01-3). H. J. Yoo wishes to acknowledge the financial support of Korea Research Foundation made in the program year of 1997.

[1] D. Ruelle, *Statistical Mechanics. Rigorous Results* (Benjamin, New York, 1969).
[2] D. Ruelle, ''Superstable interactions in classical statistical mechanics,'' Commun. Math. Phys. **18**, 127–159 (1970).
[3] M. Fukusima, *Dirichlet Forms and Markov Processes* (North-Holland, Amsterdam, 1980).
[4] Z. Ma and M. Röckner, *Introduction to the Theory of (Non-Symmetric) Dirichlet Forms* (Springer-Verlag, Berlin, 1992).
[5] R. Lang, ''Unendlich-dimensionale Wiener processe mit wechselwirkung. I. Existenz,'' Z. Wahrsch. Verw. Gebiete **38**, 55–72 (1977).
[6] R. Lang, ''Unendlich-dimensionale Wiener processe mit Wechselwirkung. II. Die reversiblen Maβe sind kanonische Gibbs-Maβe,'' Z. Wahrsch. Verw. Gebiete **39**, 277–299 (1977).
[7] J. Fritz, ''Gradient dynamics of infinite point systems,'' Ann. Prob. **15**, 478–514 (1987).
[8] H. Osada, ''Dirichlet form approach to infinite-dimensional Wiener processes with singular interactions,'' Commun. Math. Phys. **176**, 117–132 (1996).
[9] M. W. Yoshida, ''Construction of infinite dimensional interacting diffusion processes through Dirichlet form,'' Prob. Theor. Relat. Fields **106**, 265–297 (1996).
[10] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Differential geometry of Poisson spaces,'' C. R. Acad. Sci. Paris **323**, 1129–1134 (1996).
[11] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Canonical Dirichlet operator and distorted Brownian motion on Poisson spaces,'' C. R. Acad. Sci. Paris **323**, 1179–1184 (1996).
[12] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Analysis and geometry on configuration spaces,'' SFB-343 preprint 97-050, Bielefeld, 1997.
[13] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Analysis and geometry on configuration spaces: Gibbsian case,'' SFB-343 preprint, Bielefeld, 1997.
[14] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Ergodicity of L^2-semigroups and extremality of Gibbs states,'' J. Funct. Anal. **144**, 394–423 (1997).
[15] H. Spohn, ''Equilibrium fluctuations for interacting Brownian particles,'' Commun. Math. Phys. **103**, 1–33 (1986).
[16] C. Marchioro, A. Pellegrinotti, and M. Pulvirenti, ''Self-adjointness of the Liouville operator for infinite classical systems,'' Commun. Math. Phys. **58**, 113–129 (1978).
[17] A. Eberle, ''Weak Sobolev spaces and Markov uniqueness of operators,'' C. R. Acad. Sci. Paris **320**, 1249–1254 (1995).
[18] S. Albeverio, Yu. G. Kondratiev, and M. Röckner, ''Dirichlet operators via stochastic analysis,'' J. Funct. Anal. **128**, 102–138 (1993).
[19] H. Y. Lim, Y. M. Park, and H. J. Yoo, ''Dirichlet forms, Dirichlet operators, and log-Sobolev inequalities for Gibbs measures of classical unbounded spin systems,'' J. Korean Math. Soc. **34**, 731–770 (1997).
[20] H. Y. Lim, Y. M. Park, and H. J. Yoo, ''Dirichlet forms and Dirichlet operators for Gibbs measures of quantum unbounded spin systems: essential self-adjointness and log-Sobolve inequality,'' J. Stat. Phys. **90**, 949–1002 (1998).
[21] Yu. M. Berezansky and Yu. G. Kondratiev, *Spectral Methods in Infinite Dimensional Analysis* (Nankova Dumka, Kiev, 1988); English translation (Kluwer, Boston, 1993).
[22] V. Choi, ''Dirichlet forms for Gibbs measures of classical particle systems,'' Dissertation, Yonsei University, 1997.
[23] O. Kallenberg, *Random Measures* (Akademie-Verlag, Berlin, 1983).
[24] R. L. Dobrushin and J. Fritz, ''Non-equilibrium dynamics of one-dimensional infinite particle systems with a hard-core interaction,'' Commun. Math. Phys. **55**, 275–292 (1977).
[25] J. Fritz and R. L. Dobrushin, ''Non-equilibrium dynamics of two-dimensional infinite particle systems with a singular interaction,'' Commun. Math. Phys. **57**, 67–81 (1977).
[26] G. Leha and G. Ritter, ''On diffusion processes and their semi groups in Hilbert spaces with an application to interacting stochastic systems,'' Ann. Prob. **12**, 1077–1112 (1984).
[27] G. Leha and G. Ritter, ''One solution to stochastic differential equations with discontinuous drift in Hilbert space,'' Math. Ann. **270**, 109–123 (1985).
[28] Yu. L. Daleckii and S. V. Fomin, *Measures and Differential Equations in Infinite Dimensional Spaces* (Nauka, Moscow, 1983).

Chaos in the Gyldén problem

Florin Diacu[a)]
Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia V8W 3P4, Canada

Dan Şelaru
Institute for Gravitation and Space Sciences, Laboratory for Gravitation, Str. Mendeleev 21-25, Bucharest, Romania

(Received 8 October 1997; accepted for publication 18 August 1998)

We consider the Gyldén problem—a perturbation of the Kepler problem via an explicit function of time. For certain general classes of planar periodic perturbations, after proving a Poincaré–Melnikov-type criterion, we find a manifold of orbits in which the dynamics is given by the shift automorphism on the set of bi-infinite sequences with infinitely many symbols. We achieve the main result by computing the Melnikov integral explicitly. © *1998 American Institute of Physics.* [S0022-2488(98)01912-4]

I. INTRODUCTION

In a note[1] published in 1884, the Swedish astronomer Johan August Hugo Gyldén (the same who in 1889 would compete for King Oscar's Prize—awarded to Poincaré—and then fuel the scandal that followed due to the essential mistake found in the prized paper[2]) proposed a model aimed to describe the secular acceleration of the moon's longitude. He considered a perturbed Kepler problem in which the perturbation μ is an explicit function of time. In its most general form, this problem is given by the Hamiltonian function

$$H(p_1,p_2,q_1,q_2,t)=(p_1^2+p_2^2)/2-(1-\epsilon\mu(t))(q_1^2+q_2^2)^{-1/2}. \tag{1}$$

Later on, Gyldén's model was used to tackle different physical phenomena: a central-force problem with variable mass, the motion of a test particle around a star whose radiation pressure is periodic, a two-body problem in a Dirac-type universe (i.e., one in which the gravitational constant is replaced by a parameter that varies in time), etc. More details on the physical significance of this model can be found in Refs. 3 and 4.

From the mathematical point of view, approximate solutions for slow variations of the perturbation function μ have been obtained by Hadjidemetriou,[5] Verhulst,[6] and Cucu-Dumitrescu and Şelaru.[7] The case of a periodic perturbation function has been tackled by Saslaw (see Ref. 4) and Şelaru, Cucu-Dumitrescu, and Mioc.[8,9] Also, Şelaru and Mioc have applied KAM theory to this problem.[10]

In this paper we consider two classes of planar perturbations: the one for which μ is a periodic even function and the one for which $\mu(t)=A\cos(2\pi t/T)+B\sin(2\pi t/T)$, where A or B is nonzero, and $T>0$ is the period. We show that in both cases the system exhibits chaos in the sense of symbolic dynamics. More precisely, we prove that the dynamics of a manifold of solutions is similar to that of a shift automorphism on the set of bi-infinite sequences with infinitely many symbols. To achieve this goal we first consider the unperturbed problem via McGehee transformations, determine its global flow, and find the Poincaré map near the homoclinic orbit to the degenerate periodic orbit at infinity. Then, using a result of McGehee,[11] we show that this Poincaré map possesses analytic stable and unstable manifolds. Pursuing the Poincaré-Melnikov method, we further prove a Melnikov-type criterion,[12] which tells under what conditions planar periodic perturbations make the stable and unstable manifolds of the periodic orbit at infinity for the Poincaré map split and intersect transversally. Then, using this criterion, we show that the Melnikov integral has zeroes for each of the above-described class of periodic perturbation func-

a)Electronic mail: diacu@math.uvic.ca

 © 1998 American Institute of Physics

tions; consequently, the stable and unstable manifolds of the periodic orbit at infinity intersect transversely, giving rise to chaos in the neighborhood of the homoclinic orbit.

The above scenario has been previously used by Moser[13] to prove the existence of chaos in the Sitnikov problem and by Xia[14] to put into evidence the same phenomenon in the elliptic restricted three-body problem. In general, this method is hard to apply due to the difficulty of evaluating the Melnikov integral. However, for the above-described classes of Gyldén systems, we manage to overcome the technical challenge. Moreover, we are able to compute the Melnikov integral explicitly.

II. THE EQUATIONS OF MOTION

Consider a unit point mass moving around a center in a Newtonian gravitational field perturbed by a periodic function of time. The equations of motion are described by the Hamiltonian (1), or in explicit form by the system

$$\begin{aligned} \dot{q}_1 &= p_1, \\ \dot{q}_2 &= p_2, \\ \dot{p}_1 &= -(1-\epsilon\mu(t))q_1(q_1^2+q_2^2)^{-3/2}, \\ \dot{p}_2 &= -1(1-\epsilon\mu(t))q_2(q_1^2+q_2^2)^{-3/2}, \end{aligned} \tag{2}$$

for which (q_1,q_2) and (p_1,p_2) denote the position and the momentum of the moving particle, μ is the periodic perturbation function (of prime period $T>0$), and $\epsilon \geqslant 0$ is a parameter. If $\epsilon=0$, the equations (2) define the *unperturbed system*, whereas if ϵ is positive and small, the equations (2) are called the *perturbed system*. The latter expresses the planar periodic Gyldén problem and the former describes the classical Kepler problem; in both cases the center body and the moving particle have unit mass.

Under the inverse of the transformation

$$(0,\infty)\times S^1\times\mathbb{R}^2\to(\mathbb{R}^2\backslash\{(0,0)\})\times\mathbb{R}^2,$$

$$(r,\theta,p_r,p_\theta)\to(q_1,q_2,p_1,p_2)$$

given by

$$\begin{aligned} q_1 &= r\cos\theta, \\ q_2 &= r\sin\theta, \\ p_1 &= p_r\cos\theta-(p_\theta/r)\sin\theta, \\ p_2 &= p_r\sin\theta+(p_\theta/r)\cos\theta, \end{aligned} \tag{3}$$

which is a real analytic diffeomorphism, the equations (2) become

$$\begin{aligned} \dot{r} &= p_r, \\ \dot{\theta} &= r^{-2}p_\theta, \\ \dot{p}_r &= r^{-3}p_\theta^2-r^{-2}(1-\epsilon\mu(t)), \\ \dot{p}_\theta &= 0. \end{aligned} \tag{4}$$

Notice that the equations (4) have the Hamiltonian

$$\overline{H}(p_r,p_\theta,r,\theta)=p_r^2/2+p_\theta^2/(2r^2)-(1-\epsilon\mu(t))/r. \tag{5}$$

Also, in system (4) the last equation can be dropped.

We will further use *McGehee transformations of the first kind*, given by the inverse of the function

$$(0,\infty)\times S^1\times\mathbb{R}\to(0,\infty)\times S^1\times\mathbb{R},$$

$$(x,\theta,y)\to(r,\theta,p_r),$$

where

$$\begin{aligned} r&=1/x^2,\\ y&=p_r. \end{aligned} \tag{6}$$

(There are two kinds of McGehee transformations: the first ones, which we use here, have been defined in Ref. 11 to address the concept of a ''periodic orbit at infinity;'' the second ones were considered in Ref. 15 to blow up the triple-collision singularity of the rectilinear three-body problem.) The real analytic diffeomorphism (6) transforms the equations (4) into

$$\begin{aligned} \dot{x}&=-x^3y/2,\\ \dot{\theta}&=kx^4,\\ \dot{y}&=-x^4(1-\epsilon\mu(t))+k^2x^6. \end{aligned} \tag{7}$$

Note that the McGehee transformations destroy the Hamiltonian structure of the equations of motion. In exchange they will allow us to understand the dynamics at infinity by bringing infinity to the origin of the system of coordinates. In fact, since the equations (7) make sense for $x=0$, the phase space $(0,\infty)\times S^1\times\mathbb{R}$ can be extended to $[0,\infty)\times S^1\times\mathbb{R}$.

The second equation in (7) can be solved independently on the other two, so from now on we will focus on the nonautonomous system

$$\begin{aligned} \dot{x}&=-x^3y/2,\\ \dot{y}&=-x^4(1-\epsilon\mu(t))+k^2x^6, \end{aligned} \tag{8}$$

in the phase plane $[0,\infty)\times\mathbb{R}$.

III. THE KEPLER PROBLEM IN MCGEHEE COORDINATES

Let us first tackle the particular case of the unperturbed (autonomous) system, i.e., the Kepler problem in McGehee coordinates given by the equations (8) with $\epsilon=0$:

$$\begin{aligned} \dot{x}&=-x^3y/2,\\ \dot{y}&=-x^4+k^2x^6, \end{aligned} \tag{9}$$

with the energy relation

$$y^2/2+k^2x^4/2-x^2=h. \tag{10}$$

The time-rescaling transformation $ds=x^3dt$, which is also a real analytic diffeomorphism, changes system (9) into

$$\begin{aligned} x'&=-y/2,\\ y'&=-x+k^2x^3, \end{aligned} \tag{11}$$

where the prime denotes differentiation with respect to the new time variable s. The equations (11) have two equilibria: $(0,0)$ and $(1/k^2,0)$. For the origin, the eigenvalues of the attached linear

system are $\pm\sqrt{2}/2$, so (0,0) is a saddle; for the other equilibrium, the attached linear system has the eigenvalues $\pm i\sqrt{2}/2$, so no direct conclusion can be drawn about the nonlinear system. However, from the energy relation (10), we can tell that $(1/k^2,0)$ is a center.

For the original system (9), the origin is a degenerate equilibrium, since every solution $(0,y)$ is an equilibrium. Thus in the phase-space picture, the periodic orbits correspond to $h<0$, the heteroclinic ones to $h>0$, and the homoclinic orbit has the energy constant $h=0$. Let us tackle each case.

For $h<0$ and $y\geqslant 0$, the periodic orbits are described by the equation

$$\dot{x}=-(1/2)x^3\sqrt{-k^2x^4+2x^2+2h}, \tag{12}$$

which if integrated yields

$$t=\frac{\sqrt{-k^2x^4+2x^2+2h}}{2hx^2}+\frac{1}{2\sqrt{2}h\sqrt{-h}}\arcsin\frac{x^2+2h}{x^2\sqrt{1+2hk^2}} \tag{13}$$

for the admissible values of x. It looks hopeless to try to express x as an explicit function of t.

For $h=0$ and $y\geqslant 0$ the homoclinic orbit is described by the equation

$$\dot{x}=-(1/2)x^4\sqrt{2-k^2x^2},$$

which if integrated leads to

$$t=(1+k^2x^2)\sqrt{2-k^2x^2}/(3x^3),$$

for $x\in(0,\sqrt{2}/|k|]$. With the substitution $u=x^2$, we obtain the cubic equation

$$u^3+(3/2)k^2u^2-(9t^2+k^6)/2=0,$$

which by Cardano's formula yields

$$u(t)=(1/2)[(3t+\sqrt{9t^2+k^6})^{2/3}+(3t-\sqrt{9t^2+k^6})^{2/3}-k^2].$$

Thus the homoclinic orbit is given by

$$\begin{aligned} x(t)&=\xi(t,k)=\sqrt{2}[(3t+\sqrt{9t^2+k^6})^{2/3}+(3t-\sqrt{9t^2+k^6})^{2/3}-k^2]^{-1,2},\\ y(t)&=\eta(t,k)=\begin{cases}(2x^2(t)-k^2x^4(t))^{1/2}, & t\geqslant 0,\\ -(2x^2(t)-k^2x^4(t))^{1/2}, & t<0,\end{cases}\end{aligned} \tag{14}$$

for $k\neq 0$. The homoclinic orbit intersects the Ox axis at $x=\sqrt{2}/|k|$. Notice that for $k=0$ the motion takes place on a line and that the solution leads to a collision and/or an ejection.

For $h>0$ and $y\geqslant 0$, the heteroclinic orbits are also described by the equation (12). In this case, however, integrating (12) leads to

$$t=\frac{\sqrt{-k^2x^4+2x^2+2h}}{2hx^2}+\frac{1}{2\sqrt{2}h^{3/2}}\log\frac{2\sqrt{2h}\sqrt{-k^2x^4+2x^2+2h}+2x+4h}{x}, \tag{15}$$

for which it also looks hopeless to write x as an explicit function of t.

Up to now we have understood the qualitative behavior of the unperturbed system (the Kepler problem in McGehee coordinates), but have only a very limited understanding of its analytic aspect. Since it looks hopeless to solve the unperturbed system (8) or even to understand its global flow, the natural questions that arise concern the existence of periodic, oscillatory, and dense orbits for the equations (8). A classical expansion theorem in perturbation theory (see, e.g., Ref. 16, Chap. 9) indicates that if periodic orbits of system (8) exist, they can be found only in the region within the homoclinic loop to (0,0), with $O(\epsilon)$ error of the unperturbed orbit. However, this result offers little insight into the problem. Therefore we now proceed with developing a stronger method, along the lines of Poincaré and Melnikov.

IV. THE MELNIKOV FUNCTION

Let us assume from now on that μ has also the property

$$\int_0^T \mu(t)dt = 0.$$

(In fact this condition does not restrain the generality since we can introduce in the Newtonian term of the interaction the temporal average of the perturbation.)

Also notice that since μ is analytic, bounded, and T-periodic, it has the Fourier representation,

$$\mu(t) = \sum_{n=1}^{\infty} [C_n \cos(2\pi nt/T) + S_n \sin(2\pi nt/T)],$$

where the coefficients C_n and S_n are real for all $n \geqslant 1$.

Let us further rewrite the equations (8) as an autonomous system:

$$\frac{dx}{d\tau} = -\frac{x^3 y}{2},$$

$$\frac{dy}{d\tau} = -x^4 + k^2 x^6 + \epsilon x^4 \mu(t), \tag{16}$$

$$\frac{dt}{d\tau} = 1.$$

Consider further a global section $\Sigma^{\tau=\tau_0}$ and denote by $P_{\tau=\tau_0}(x,y)$ the first return of the Poincaré function defined on $\Sigma^{\tau=\tau_0}$. Notice that the origin O is a degenerate fixed point for $P_{\tau=\tau_0}$.

For degenerate situations of this type, McGehee has obtained a result that shows the existence of stable and unstable manifolds for the Poincaré map [see Ref. 11, p. 82, system (7.3) and Proposition (11)]. With the help of this result we can now prove the following:

Lemma 1: The stable and unstable manifold $W^s(\Lambda)$ *and* $W^u(\Lambda)$ *of the periodic orbit* $\Lambda = \{x=0, y=0, \tau \in [0,T)\}$ *are analytic manifolds for* $x>0$ *and for a fixed value of the angular momentum k.*

Proof: McGehee's result cannot be directly applied to system (16), so we will first consider the change of variables

$$\overline{x} = x,$$

$$\overline{t} = t,$$

$$\overline{y} = y + \epsilon x^4 \sum_{n=1}^{\infty} \frac{T}{2\pi n}\left(C_n \sin\frac{2\pi nt}{T} - S_n \cos\frac{2\pi nt}{T}\right),$$

which fulfills the condition $D(x,y,t)/D(\overline{x},\overline{y},\overline{t}) = 1$, and which transforms system (16) into

$$\frac{d\overline{x}}{d\tau} = -\overline{x}^3\left(\frac{\overline{y}}{2} + g_1(\overline{x},\overline{y},\overline{t})\right),$$

$$\frac{d\overline{y}}{d\tau} = -\overline{x}^3(\overline{x} + g_2(\overline{x},\overline{y},\overline{t})),$$

$$\frac{d\overline{t}}{d\tau} = 1,$$

where the functions g_1, g_2 are periodic in $\overline{t}$, C^∞, and such that $|g_i(\overline{x},\overline{y},\overline{t})| \leqslant O(|(x,y)|^2)$, $i=1,2$.

McGehee's result can now be applied to the above equations and Lemma 1 follows. This completes the proof.

Recall that in the unperturbed case the stable and unstable manifolds of Λ are identical and form the homoclinic loop (14). We will see that in the perturbed case the two manifolds split and intersect transversally. The first step in formulating the necessary conditions of transverse intersections is to find suitable estimates of the stable and unstable manifolds of the perturbed system. This is done in the following:

Lemma 2: The solutions $(x^{s,u}(\tau,\tau_0), y^{s,u}(\tau,\tau_0))$ *that remain in the manifolds* $W^s(\Lambda)$ *and* $W^u(\Lambda)$, *and for which* $(x^{s,u}(0,\tau_0), y^{s,u}(0,\tau_0))$ *belong to the Poincaré section at time* $\tau=\tau_0$, *admit, for* ϵ *sufficiently small, the estimates*

$$x^s(\tau,\tau_0)=\xi(\tau-\tau_0,k)+\epsilon x_1^s(\tau,\tau_0)+O(\epsilon^2), \quad \tau\in[\tau_0,\infty),$$

$$y^s(\tau,\tau_0)=\eta(\tau-\tau_0,k)+\epsilon y_1^s(\tau,\tau_0)+O(\epsilon^2), \quad \tau\in[\tau_0,\infty),$$

$$x^u(\tau,\tau_0)=\xi(\tau-\tau_0,k)+\epsilon x_1^u(\tau,\tau_0)+O(\epsilon^2), \quad \tau\in(\infty,\tau_0],$$

$$y^u(\tau,\tau_0)=\eta(\tau-\tau_0,k)+\epsilon y_1^u(\tau,\tau_0)+O(\epsilon^2), \quad \tau\in(\infty,\tau_0],$$

uniformly in the indicated time intervals. The functions $x_1^{s,u}(\tau,\tau_0)$, $y_1^{s,u}(\tau,\tau_0)$ *are determined by the first variational equation of system* (8) *along the unperturbed homoclinic orbit* (14).

The proof of Lemma 2, based on Gronwall's lemma and on the fact that the perturbed and unperturbed manifolds are C^∞-close, can be found in Ref. 17, 18, or 19. The necessary condition for the proposed estimates to be valid is that the period T of the function μ is neither too big nor too small.

The vector

$$\mathbf{d}(\tau_0)=(x^s(\tau_0,\tau_0)-x^u(\tau_0,\tau_0), y^s(\tau_0,\tau_0)-y^u(\tau_0,\tau_0))$$

measures the splitting of the manifolds $W^s(\Lambda)$ and $W^u(\Lambda)$ in the Poincaré section $\Sigma^{\tau=\tau_0}$. Denoting by H_0 the Hamiltonian (1) for $\epsilon=0$, notice that the projection of $\mathbf{d}$ onto the vector

$$\mathbf{H}_{0N}=\left(\frac{\partial H_0}{\partial x}(\xi(0,k),\eta(0,k)), \frac{\partial H_0}{\partial y}(\xi(0,k),\eta(0,k))\right),$$

normal to $H_0(x,y)$ at the point $(\xi(0,k),\eta(0,k))$, approximates the separation of the two manifolds $W^s(\Lambda)$ and $W^u(\Lambda)$ (see Ref. 14). Denoting by $\langle,\rangle$ the scalar product, in the first approximation we have

$$\left\langle \mathbf{d}(\tau_0), \frac{\mathbf{H}_{0N}}{|\mathbf{H}_{0N}|}\right\rangle = \epsilon \frac{\langle \mathbf{H}_{0N}, (x_1^s(\tau_0,\tau_0)-x_1^u(\tau_0,\tau_0), y_1^s(\tau_0,\tau_0)-y_1^u(\tau_0,\tau_0))\rangle}{|\mathbf{H}_{0N}|}+O(\epsilon^2),$$

which, using Lemma 2 (see also Ref. 14), leads to

$$\left\langle \mathbf{d}(\tau_0), \frac{\mathbf{H}_{0N}}{|\mathbf{H}_{0N}|}\right\rangle = \epsilon \frac{H_0(x^s(\tau_0,\tau_0),y^s(\tau_0,\tau_0))-H_0(x^u(\tau_0,\tau_0),y^u(\tau_0,\tau_0))}{|\mathbf{H}_{0N}|}+O(\epsilon^2)$$

$$=\frac{\epsilon}{|\mathbf{H}_{0N}|}\int_{-\infty}^{+\infty}\frac{dH_0}{d\tau}\,d\tau+O(\epsilon^2)=\frac{\epsilon}{|\mathbf{H}_{0N}|}M(\tau_0)+O(\epsilon^2),$$

where $M(\tau_0)$ is the Melnikov function (see Ref. 12, 17, or 19):

$$M(\tau_0)=\frac{1}{\epsilon}\int_{-\infty}^{+\infty}\frac{dH_0(\xi(\tau-\tau_0,k),\eta(\tau-\tau_0,k))}{d\tau}\,d\tau$$

$$=\int_{-\infty}^{+\infty}\xi^4(\tau-\tau_0,k)\,\eta(\tau-\tau_0,k)\,\mu(\tau)d\tau$$

$$=\int_{-\infty}^{+\infty}\xi^4(\tau,k)\,\eta(\tau,k)\,\mu(\tau+\tau_0)d\tau. \tag{17}$$

Substituting the Fourier series representation of μ into the Melnikov function (17), we obtain

$$M(\tau_0)=\sum_{n=1}^{\infty}\left[\left(C_n\cos\frac{2\pi n\tau_0}{T}+S_n\sin\frac{2\pi n\tau_0}{T}\right)\int_{-\infty}^{+\infty}\xi^4(\tau,k)\,\eta(\tau,k)\cos\frac{2\pi n\tau}{T}\,d\tau\right.$$

$$\left.+\left(-C_n\sin\frac{2\pi n\tau_0}{T}+S_n\cos\frac{2\pi n\tau_0}{T}\right)\int_{-\infty}^{+\infty}\xi^4(\tau,k)\,\eta(\tau,k)\sin\frac{2\pi n\tau}{T}\,d\tau\right]$$

$$=\sum_{n=1}^{\infty}\left(-C_n\sin\frac{2\pi n\tau_0}{T}+S_n\cos\frac{2\pi n\tau_0}{T}\right)\int_{-\infty}^{+\infty}\xi^4(\tau,k)\,\eta(\tau,k)\sin\frac{2\pi n\tau}{T}\,d\tau,$$

where the first of the integrals cancels because the function under the integral is odd. We further have

$$M'(\tau_0)=\frac{dM}{d\tau}(\tau_0)=-\sum_{n=1}^{\infty}\frac{2\pi n}{T}\left(C_n\cos\frac{2\pi n\tau_0}{T}+S_n\sin\frac{2\pi n\tau_0}{T}\right)I_n,$$

where $I_n>0$ stands for

$$I_n=\int_{-\infty}^{+\infty}\xi^4(\tau,k)\,\eta(\tau,k)\sin\frac{2\pi n\tau}{T}\,d\tau.$$

The actual form of I_n will not be involved in our future endeavors. However, we compute I_n in the Appendix at the end of our paper to show that $I_n>0$.

Taking now into account the form of the vector $\mathbf{H}_{0N}$ and the equations (14), a straightforward computation leads to

$$|\mathbf{H}_{0N}|=\sqrt{(2k^2\xi^3(0,k)-2\xi(0,k))^2+\eta^2(0,k)}=2\sqrt{2}/|k|,$$

so we have

$$\left\langle\mathbf{d}(\tau_0),\frac{\mathbf{H}_{0N}}{|\mathbf{H}_{0N}|}\right\rangle=\epsilon\frac{|k|M(\tau_0)}{2\sqrt{2}}+O(\epsilon^2).$$

This formula will allow us to reach our conclusions.

V. THE MAIN RESULTS

From the above formula we can see that the necessary conditions for the transverse intersections of $W^s(\Lambda)$ and $W^u(\Lambda)$ are

(i) ϵ is sufficiently small.
(ii) μ is T-periodic, analytic, and bounded.
(iii) There exists $\tau_0\in[0,T)$ such that $M(\tau_0)=0$, the derivative M' is defined and differentiable at τ_0, and $M'(\tau_0)\neq 0$.

We have thus proved the following criterion:

Theorem 3: *For the equations (8), which define the Gyldén problem, for μ T-periodic, analytic, and bounded, for a sufficiently large value of the angular momentum k, and for a sufficiently small $\epsilon>0$, if there exists a $\tau_0\in[0,T)$ such that $M(\tau_0)=0$ and $M'(\tau_0)\neq 0$, then the manifolds $W^s(\Lambda)$ and $W^u(\Lambda)$ of the Poincaré map* (16) *intersect transversally.*

From Theorem 3 we can now draw the following conclusion:

Corollary 4: For the equations (8)*, which define the Gyldén problem, for μ even, T-periodic, analytic, and bounded [i.e., $\mu(t)=\Sigma_{n=1}^{\infty}C_n\cos(2\pi nt/T)$], for a sufficiently large value of the angular momentum k, and for a sufficiently small $\epsilon>0$, the manifolds $W^s(\Lambda)$ and $W^u(\Lambda)$ of the Poincaré map* (16) *intersect transversally. This happens generically, except possibly for functions μ that belong to a certain negligible set included in the set of all functions μ with the above-mentioned properties.*

Proof: Observe that the Melnikov function becomes in this case

$$M(\tau_0)=-\sum_{n=1}^{\infty} C_nI_n \sin\frac{2\pi n\tau_0}{T},$$

and its derivative has the form

$$M'(\tau_0)=-\sum_{n=1}^{\infty} \frac{2\pi nC_nI_n}{T}\cos\frac{2\pi n\tau_0}{T} I_n.$$

Notice that $M(\tau_{0i})=0$, $i=1,2$, at least for the values $\tau_{01}=0$ and $\tau_{02}=T/2$, whereas for the derivative of the Melnikov function at these points we obtain

$$M'(\tau_{01})=-\sum_{n=1}^{\infty} \frac{2\pi nC_n}{T} I_n, \quad M'(\tau_{02})=\sum_{n=1}^{\infty} (-1)^{n+1}\frac{2\pi nC_n}{T} I_n.$$

Clearly, $M'(\tau_{0i})\neq 0$ for most of the coefficients $C_1,C_2,\ldots$. More precisely, $M'(\tau_{0i})\neq 0$ in the space of coefficients, $\{(C_1,C_2,\ldots)|C_i\in\mathbb{R}\}$, except possibly for a negligible subset. This completes the proof.

Another consequence we can draw from Theorem 3 is the following:

Corollary 5: For the equations (8), which define the Gyldén problem, for functions μ that are T-periodic, analytic, bounded and have a single harmonic [i.e., $\mu(t)=A\cos(2\pi t/T)+B\sin(2\pi t/T)$, with $\sqrt{A^2+B^2}\neq 0$], for a sufficiently large value of the angular momentum k, and for a sufficiently small $\epsilon>0$, the manifolds $W^s(\Lambda)$ and $W^u(\Lambda)$ of the Poincaré map (16) *intersect transversally.*

Proof: The Melnikov function becomes in this case

$$M(\tau_0)=\left(-A\sin\frac{2\pi\tau_0}{T}+B\cos\frac{2\pi n\tau_0}{T}\right)I_1,$$

and its derivative takes the form

$$M'(\tau_0)=-\frac{2\pi}{T}\left(A\cos\frac{2\pi\tau_0}{T}+B\sin\frac{2\pi n\tau_0}{T}\right)I_1.$$

We can consider $B(A^2+B^2)^{-1/2}=\sin\alpha$ and $A(A^2+B^2)^{-1/2}=\cos\alpha$, and obtain the expressions

$$M(\tau_0)=I_1\sqrt{A^2+B^2}\sin(\alpha-2\pi\tau_0/T),$$

$$M'(\tau_0)=-I_1\sqrt{A^2+B^2}(2\pi/T)\cos(\alpha-2\pi\tau_0/T)$$

for the Melnikov function and its derivative. The equation $M(\tau_0)=0$ admits the only solutions $\tau_{01}=\alpha T/2\pi$ and $\tau_{02}=(\alpha-\pi)T/2\pi$, for which

$$M'(\tau_{01})=-I_1\sqrt{A^2+B^2}(2\pi/T)\neq 0, \quad M'(\tau_{02})=I_1\sqrt{A^2+B^2}(2\pi/T)\neq 0.$$

By Theorem 3 the conclusion of Corollary 5 follows. This completes the proof.

We will close our endeavors with a symbolic-dynamics description of the above results, which will allow us a better understanding of the dynamics of the Gyldén problem. For this, consider a general Poincaré map P for a transversal homoclinic orbit Λ and let $\mathcal{R}$ be a small rectangle having two of its sides on the stable and unstable manifolds of Λ, respectively, at a point at which the two manifolds intersect transversally. For a point q in $\mathcal{R}$, let n be the smallest positive integer for which $P^n(q)$ belongs to $\mathcal{R}$, and denote by $\mathcal{Q}$ the set of all those q for which such an n (which depends on q) exists. In Ref. 13, Moser shows that $\mathcal{Q}$ is nonempty and proves that a shift automorphism on infinitely many symbols can be embedded in the set $\mathcal{Q}$. The precise formulation of this result is as follows:

Theorem 6: *Let $\mathcal{T}(q)=P^n(q)$, for all $q\in\mathcal{Q}$, and let $\mathcal{B}$ denote the set of bi-infinite sequences on infinitely many symbols. Then there is an invariant set $I\subset\mathcal{Q}$ for the map $\mathcal{T}$, homeomorphic to $\mathcal{B}$, such that the map $\mathcal{T}$ is topologically conjugate to the shift automorphism on $\mathcal{B}$.*

This result has two main consequences. First it proves the existence of periodic orbits of any period, of dense orbits, as well as that of oscillatory ones, or of any other orbits that can be put into correspondence with a symbolic sequence. Second, it shows that real analytic first integrals do not exist. All of these characterize chaotic motion in the Gyldén problem and put into the evidence some of the complicated dynamics in the phase space of the equations (8).

Notice that due to the low dimension of system (8), transition-tori phenomena (like the Arnold diffusion[20–23]) cannot occur. However, such subtle dynamical aspects may very well characterize higher-dimensional systems of the same type, i.e., n-body Gyldén problems with $n\geqslant 3$.

APPENDIX

In this Appendix we compute I_n. For this define

$$J_\omega=\int_{-\infty}^{\infty} x^4 y\,\sin\,\omega\tau\,d\tau$$

along the unperturbed orbit (16). From the equations (11), integrating J_ω by parts, we have

$$J_\omega=-\int_{-\infty}^{\infty}\frac{d}{d\tau}(x^2)\sin\,\omega\tau\,d\tau=\omega\int_{-\infty}^{\infty}\frac{\cos\,\omega\tau}{r}\,d\tau.$$

Using the angular momentum integral and the expression of r as a function of the true anomaly v for unperturbed parabolic motions, given by $r=1/(1+\cos v)$, we obtain

$$J_\omega=\frac{\omega k}{2}\int_{-\pi}^{\pi}(1+\tan^2(v/2))\cos\,\omega\tau\,dv.$$

Denoting $\tan(v/2)=\alpha$ and taking into account the fact that parabolic motion implies that $\tau=(k^3/4)(\alpha+\alpha^3/3)$, we obtain

$$J_\omega=\omega k^3\int_{-\infty}^{\infty}\cos[(\omega k^3/4)(s+s^3/3)]$$

or

$$J_\omega=\frac{2\pi\omega k^3}{3}\left[\tilde{I}\left(-\frac{1}{3},\frac{\omega k^3}{6}\right)-\tilde{I}\left(\frac{1}{3},\frac{\omega k^3}{6}\right)\right],$$

where $\tilde{I}(\nu,z)$ is the modified Bessel function of the first kind.[24] However,

$$\tilde{I}(-\nu,z)-\tilde{I}(\nu,z)=\frac{2\sin\,\pi\nu}{\pi}K_\nu(z),$$

where $K_\nu(z)$ is the modified Bessel function of the second kind (see Ref. 25, p. 78). So

$$J_\omega = \frac{4}{3}\,\omega k^3 \sin\frac{\pi}{3}\,K_{1/3}\left(\frac{\omega k^3}{6}\right). \tag{18}$$

However, $K_\nu(z)$ tends exponentially to 0 when $z\to\infty$ through positive values, and since $I_n = J_{2\pi n/T}$, with $J_{2\pi n/T}$ given by (18), the positivity of I_n follows.

ACKNOWLEDGMENT

Florin Diacu was supported in part by Grant No. OGP0122045 of the NSERC of Canada.

[1] H. Gyldén, ''Die Bahnbewegung in einem Systeme von zwei Körpern in dem Falle dass die Massen Veränderungen unterworfen sind,'' Astron. Nachr. **109**, 1–6 (1884).
[2] F. Diacu and P. Holmes, *Celestial Encounters—The Origins of Chaos and Stability* (Princeton U.P., Princeton, NJ, 1996).
[3] A. Deprit, ''The secular accelerations in Gyldén's problem,'' Celest. Mech. **31**, 1–22 (1983).
[4] W. C. Saslaw, ''Motion around a source whose luminosity changes,'' Astrophys. J. **226**, 240–252 (1978).
[5] J. D. Hadjidemetriou, ''Two-body problem with variable mass: a new approach,'' Icarus **2**, 440–448 (1963).
[6] F. P. Verhulst, ''Asymptotic expansion in the perturbed two-body problem with application to systems with variable mass,'' Celest. Mech. **11**, 95–129 (1975).
[7] C. Cucu-Dumitrescu and D. Şelaru, ''Gyldén model: the slowly changing equivalent gravitational parameter,'' Astrophys. Space Science (to appear).
[8] D. Şelaru, C. Cucu-Dumitrescu, and V. Mioc, ''On a two-body problem with periodically changing equivalent gravitational parameter,'' Astron. Nachr. **313**, 257–263 (1993).
[9] D. Şelaru, C. Cucu-Dumitrescu, and V. Mioc, ''Periodic motions around pulsating stars,'' Astrophys. Space Sci. **202**, 11–19 (1993).
[10] D. Şelaru and V. Mioc, ''Le problème de Gyldén du point de vue de la théorie KAM,'' Compt. Rendus t. 325, Série IIb, 487–490 (1997).
[11] R. McGehee, ''A stable manifold theorem for degenerate fixed points with applications to celestial mechanics,'' J. Diff. Eqns. **14**, 70–88 (1973).
[12] V. K. Melnikov, ''On the stability of the center for time periodic perturbations,'' Trans. Moscow Math. Soc. **12**, 1–56 (1963).
[13] J. Moser, *Stable and Random Motions in Dynamical Systems*, Annals of Mathematics Studies, Vol. 77 (Princeton U. Press, Princeton, NJ, 1973).
[14] Z. Xia, ''Melnikov method and transversal homoclinic points in the restricted three-body problem,'' J. Diff. Eqns. **96**, 170–184 (1992).
[15] R. McGehee, ''Triple collision in the collinear three-body problem,'' Invent. Math. **27**, 191–227 (1974).
[16] F. P. Verhulst, *Nonlinear Differential Equations and Dynamical Systems* (Springer Verlag, Berlin, 1990).
[17] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer Verlag, New York, 1983).
[18] P. Holmes, ''Averaging and chaotic motions in forced oscillations,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **38**, 65–80 (1980).
[19] J. A. Sanders, ''Melnikov's method and averaging,'' Celest. Mech. **28**, 171–181 (1982).
[20] V. I. Arnold, ''Instability of dynamical systems with several degrees of freedom,'' Sov. Math. Dokl. **5**, 581–585 (1963).
[21] R. Moeckel, ''Transition tori in the five-body problem,'' J. Diff. Eqns. **129**, 290–314 (1996).
[22] Z. Xia, ''Arnold diffusion in the elliptic restricted three-body problem,'' J. Dynamics and Differential Equations **5**, 219–240 (1993).
[23] Z. Xia, ''Arnold diffusion and oscillatory solutions in the planar three-body problem,'' J. Diff. Eqns. **110**, 289–321 (1994).
[24] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington DC, 1968).
[25] G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge U.P., New York, 1995).

A generalized Hirota equation in 2+1 dimensions

Attilio Maccari
Technical Institute ''G. Cardano,'' Piazza della Resistenza 1, 00015 Monterotondo (Rome), Italy

(Received 2 March 1998; accepted for publication 10 May 1998)

Using an asymptotically exact reduction method based on Fourier expansion and spatiotemporal rescaling, a new integrable and nonlinear partial differential equation (PDE) in $2+1$ dimensions is obtained starting from the Kadomtsev–Petviashvili equation. We apply the reduction technique to the Lax pair of the Kadomtsev–Petviashvili equation and demonstrate the integrability property of the new equation, because we obtain the corresponding Lax pair. The new equation reduces to the Hirota equation in the $1+1$-dimensional limit. © *1998 American Institute of Physics.* [S0022-2488(98)03911-5]

I. INTRODUCTION

It is well known that certain ''universal'' nonlinear evolution PDEs can be obtained, by a limiting procedure involving the wave modulation induced by weak nonlinear effects, from very large classes of nonlinear evolution equations in $1+1$ and $2+1$ dimensions, with a dispersive linear part. The most interesting result is that only a very limited number of model equations are necessary. This reduction technique is the correct one to evince weakly nonlinear effects, and then these model equations are usually of applicative relevance (in plasma physics, nonlinear optics, hydrodynamics). The model equations are integrable, since it is sufficient that the very large class from which they are obtainable contain just one integrable equation, because it is clear from this method that the property of integrability is inherited through this limiting technique.[1–5] The reduction method can be also applied to the weakly nonlinear ordinary differential equations.[6]

The first step of the reduction method (in $2+1$ dimensions) is to use a moving frame of reference via the transformation

$$\xi=\epsilon^{p_1}(x-V_1t), \quad \eta=\epsilon^{p_2}(y-V_2t), \quad \tau=\epsilon^q t, \tag{1.1}$$

where p_1, $p_2>0$, $q>0$, and $V_1=V_1(K_1,K_2)$, $V_2=V_2(K_1,K_2)$ are the components of the group velocity $\mathbf{V}(\mathbf{K})\equiv(V_1(K_1,K_2),V_2(K_1,K_2))$ of the linearized equation, i.e., of the equation obtained after neglecting all the nonlinear terms, and ϵ is the expansion parameter, supposed to be sufficiently small. Fourier modes describe the linear evolution, as they have a constant amplitude and a well-defined group velocity. Nonlinear terms induce a modulation of the amplitude of the Fourier mode. The modulation is best described in terms of the rescaled variables ξ, η, τ, that account for the need to look on larger space and time scales, to obtain a non-negligible contribution from the nonlinear term.

In preceding papers, this method has been applied to certain integrable equations in $1+1$ and $2+1$ dimensions and new integrable nonlinear PDEs have been derived.[7–9] Moreover, the reduction technique is able to derive equations of applicative relevance.[10] In this paper we consider the Kadomtsev–Petviashvili (KP) equation[11] (integrable by means of the spectral transform[12]):

$$U_{xt}(x,y,t)+U_{xxxx}(x,y,t)+sU_{yy}(x,y,t)-3(U^2(x,y,t))_{xx}=0, \tag{1.2}$$

where $U=U(x,y,t)$ is real, $s=\pm1$, and the subscripts denote differentiation. In the following we consider the case $s=1$, because the other possibility has been previously studied.[9] If $p_1\neq p_2$, a new model equation can be derived,

$$i\Psi_\tau+\Psi_{\xi\eta}+i\Psi_{\xi\xi\xi}+\Psi\Phi-i|\Psi|^2\Psi_\xi=0, \tag{1.3a}$$

$$3\Phi_\xi+(|\Psi|^2)_\eta=0, \tag{1.3b}$$

 © 1998 American Institute of Physics

with $\Psi=\Psi(\xi,\eta,\tau)$ complex and $\Phi=\Phi(\xi,\eta,\tau)$ real. If we take $\xi=\eta$, we recover the integrable equation in $1+1$ dimensions obtained by Hirota.[13] The Lax pair relative to Eq. (1.3) will be deduced from the Lax pair of the KP equation.

II. MODEL NONLINEAR PDEs OBTAINED FROM THE KP EQUATION

The linear dispersive part of the KP equation (1.2) is simply described in terms of Fourier modes, with a group velocity $\mathbf{V}(\mathbf{K})=(V_1(K_1,K_2),V_2(K_1,K_2))$,

$$V_1(K_1,K_2)=-3K_1^2-\frac{K_2^2}{K_1^2},\quad V_2(K_1,K_2)=\frac{2K_2}{K_1}. \tag{2.1}$$

We introduce a formal asymptotic Fourier expansion,

$$U(x,y,t)=\sum_{n=-\infty}^{n=+\infty}\epsilon^{\gamma_n}\psi_n(\xi,\eta,\tau;\epsilon)\exp\{in(K_1x+K_2y-\omega t)\}, \tag{2.2}$$

with $\gamma_n=|n|$, for $n\neq 0$, $\gamma_0=r$ a rational number to be determined later on, and $\psi_{-n}(\xi,\eta,\tau;\epsilon)=\psi_n^*(\xi,\eta,\tau;\epsilon)$, because $U(x,y,t)$ is real.

The functions $\psi_n(\xi,\eta,\tau,\epsilon)$'s depend on the parameter ϵ, and we suppose that the limit of the ψ_n's for $\epsilon\to 0$ exists and is finite and, moreover, they can be expanded in power series of ϵ, i.e.,

$$\psi_n(\xi,\eta,\tau;\epsilon)=\sum_{i=0}^{\infty}\epsilon^i\psi_n^{(i)}(\xi,\eta,\tau). \tag{2.3}$$

In the following, for simplicity, we use the abbreviations $\psi_n^{(0)}=\psi_n$ for $n\neq 0,1$ and $\psi_1^{(0)}=\Psi$, $\psi_0^{(0)}=\Psi_0$.

The final goal is to obtain the evolution equation for the modulation amplitude $\Psi=\Psi(\xi,\eta,\tau)$ and to understand how it is modified by choosing different wave numbers. In particular, an interesting result shall be derived, if the special condition $K_2=\sqrt{3}K_1^2$ is verified. We substitute the expansion (2.2) in the KP equation (1.2) and consider the different equations obtained from the coefficients of the Fourier modes. To simplify the treatment, it is more convenient to separate the contributions from the linear and nonlinear parts: $\epsilon^{\gamma_n}[D_n\psi_n]=\epsilon^2F_n$, where D_n is a linear differential operator acting on $\psi_n(\xi,\eta,\tau)$ and F_n is the contribution of the nonlinear part. The operator D_n is

$$\begin{aligned}D_n=&(-in\omega+\epsilon^q\partial_\tau-V_1(K_1,K_2)\epsilon^{p_1}\partial_\xi-V_2(K_1,K_2)\epsilon^{p_2}\partial_\eta)(\epsilon^{p_1}\partial_\xi+inK_1)\\&+(inK_1+\epsilon^{p_1}\partial_\xi)^4+(inK_2+\epsilon^{p_2}\partial_\eta)^2.\end{aligned} \tag{2.4}$$

It is easily seen that the first F_n have the following explicit form:

$$F_0=6(|\Psi|^2)_{\xi\xi}\epsilon^{2p_1}+o(\epsilon^{2p_1+2},\epsilon^{2p_1+2r-2}), \tag{2.5a}$$

$$F_2=-12K_1^2\Psi^2+o(\epsilon^{p_1}), \tag{2.5b}$$

$$F_1=-6K_1^2\Psi_0\Psi\epsilon^r-6K_1^2\psi_2\Psi^*\epsilon+O(\epsilon^{p_1+r-1},\epsilon^{p_1+1}). \tag{2.5c}$$

After taking $q=2$, $p_1=p_2=1$, $r=2$ for the proper balance of terms, we obtain at the lowest order for $n=0$ and $n=2$,

$$V_1\Psi_{0,\xi\xi}+V_2\Psi_{0,\xi\eta}-\Psi_{0,\eta\eta}+6(|\Psi|^2)_{\xi\xi}=0, \tag{2.6a}$$

$$\psi_2^{(0)}=\frac{3K_1^2}{K_2^2-6K_1^4}\Psi^2, \tag{2.6b}$$

and, for $n=1$,

$$iK_1\Psi_\tau-(6K_1^2+V_1)\Psi_{\xi\xi}-V_2\Psi_{\xi\eta}+\Psi_{\eta\eta}+6K_1^2(\Psi_0\Psi+\psi_2\Psi^*)=0. \tag{2.6c}$$

Substituting (2.6b) into (2.6c) and after a trivial rescaling we arrive at the model equation of a nonlinear Schrodinger type:

$$i\Psi_\tau+L_1\Psi+\Psi\Psi_0=0, \tag{2.7a}$$

$$L_2\Psi_0=2L_1(|\Psi|^2), \tag{2.7b}$$

with the linear differential operators L_1 and L_2 given by

$$L_1=\frac{1}{4}(1-\lambda^2)\frac{\partial^2}{\partial\xi^2}+\lambda\frac{\partial}{\partial\xi}\frac{\partial}{\partial\eta}-\frac{\partial^2}{\partial\eta^2}, \tag{2.7c}$$

$$L_2=-\frac{1}{4}(1+\lambda^2)\frac{\partial^2}{\partial\xi^2}+\lambda\frac{\partial}{\partial\xi}\frac{\partial}{\partial\eta}-\frac{\partial^2}{\partial\eta^2}, \tag{2.7d}$$

with $\lambda=K_2/\sqrt{3}K_1^2$. By means of a trivial transformation in the spatial variables, the λ can be eliminated, thereby reducing Eq. (2.7) to the integrable Davey–Stewartson equation.[14,15]

Now, we consider the particular case $K_2=\sqrt{3}K_1^2$ and note that if $p_2>p_1$, a different result is obtained for $n=0$ and $r=2$,

$$\Psi_0=-\frac{1}{K_1^2}|\Psi|^2, \tag{2.8a}$$

and then the second term and the nonlinear term in Eq. (2.6c) vanish. We conclude that in this situation the preceding choice for the coefficients p_1, p_2, q is not able to capture the contribution of nonlinear terms.

We reconsider Eq. (2.6a) and see that if $p_2>p_1$ the most important terms lead to the equation

$$\epsilon^{2p_1+2}V_1\Psi_{0,\xi\xi}+6\epsilon^{2+2p_1}(|\Psi|^2)_{\xi\xi})+\epsilon^{2p_1+3}\Phi_{\xi\xi}+\epsilon^{p_1+p_2+2}V_2\Psi_{0,\xi\eta}=0 \tag{2.8b}$$

where $\Phi=\psi_0^{(1)}$ [see (2.3)]. With the choice $p_2=p_1+1$, we obtain

$$\sqrt{3}K_1^3\Phi_\xi-(|\Psi|^2)_\eta=0, \tag{2.9a}$$

while Ψ_0 is always given by Eq. (2.8a). In a similar way, we obtain, for $n=2$,

$$\psi_2^{(0)}=\frac{1}{K_1^2}\psi^2,\quad \psi_2^{(1)}=-\frac{2i}{K_1^3}\Psi\Psi_\xi,\quad p_1=1,\quad p_2=2. \tag{2.9b}$$

Taking $n=1$ and with the choice $q=3$, a new model nonlinear PDE is obtained:

$$i\Psi_\tau-2\sqrt{3}\Psi_{\xi\eta}+4i\Psi_{\xi\xi\xi}+6K_1\Psi\Phi-\frac{12i}{K_1^2}|\Psi|^2\Psi=0. \tag{2.10}$$

After a trivial rescaling, we arrive at the nonlinear model equation (1.4). It is expected that it is integrable, taking into account that it has been obtained from an equation integrable by the spectral transform.

III. THE LAX PAIR FOR THE MODEL NONLINEAR PDE

In order to demonstrate the integrability of the new equation (1.3) by the spectral transform, we apply the reduction method to the Lax pair of the KP equation. The Lax operators are

$$L=\frac{1}{\sqrt{3}}\frac{\partial}{\partial y}+\frac{\partial^2}{\partial x^2}-U(x,y,t),\quad L\varphi(x,y,t)=0, \tag{3.1}$$

$$A=4\frac{\partial^3}{\partial y^3}-6U(x,y,t)\frac{\partial}{\partial x}-3U_x(x,y,t)+\sqrt{3}W(x,y,t), \tag{3.2}$$

with $W_x(x,y,t)=U_y(x,y,t)$ and

$$\varphi_t(x,y,t)+A\varphi(x,y,t)=0. \tag{3.3}$$

We introduce an asymptotic Fourier expansion,

$$\varphi(x,y,t)=\sum_{n=-\infty}^{n=+\infty}\epsilon^{\gamma_n}\varphi_n(\xi,\eta,\tau;\epsilon)\exp\left(i\frac{n}{2}z+i(\lambda_1 x+\lambda_2 y+\lambda_3 t)\right), \tag{3.4}$$

where n is odd, $z=K_1x+K_2y-\omega t$, $\gamma_{n+2}=1+\gamma_n$, $\gamma_n=\gamma_{-n}$ for $n>0$, the $\varphi_n(\xi,\eta,\tau;\epsilon)$'s are functions that parametrically depend on ϵ and remain finite if $\epsilon\to 0$ and $\lambda_1,\lambda_2,\lambda_3$ are real constants to be determined later on. Substituting the expression for $\varphi(x,y,t)$ in Eq. (3.1), the coefficients of the Fourier modes generate a series of relations. Obviously, each relation must be valid for a given order of approximation in ϵ. In particular, for the fundamental harmonics $\varphi_{\pm}(x,y,t)=\varphi_{\pm 1}(x,y,t)$, considering terms $O(\epsilon^0)$ in (3.1) and (3.3), the constants $\lambda_1,\lambda_2,\lambda_3$ can be fixed:

$$\lambda_1=-\frac{K_1}{2},\quad \lambda_2=0,\quad \lambda_3=-K_1^3. \tag{3.5}$$

The new spectral problem is obtained by means of the successive order ϵ for the equation (3.1),

$$K_1(i-1)\varphi_{+,\xi}=\Psi\varphi_-, \tag{3.6a}$$

$$K_1(i+1)\varphi_{-,\xi}=-\Psi^*\varphi_+. \tag{3.6b}$$

The spectral problem can be obtained again when we consider the temporal evolution equation (3.3) at the same order ϵ. Only if we take into account the next orders of approximation of (3.3), can the temporal evolution be determined. However, new quantities, the corrections $\tilde{\varphi}_{\pm}(\xi,\eta,\tau)$ of order ϵ to the fundamental harmonics $\varphi_{\pm}(\xi,\eta,\tau)$, appear. These unknown quantities can be eliminated in (3.3), taking advantage of the relation obtained from (3.1), considering terms of order ϵ^2 and ϵ^3. This elimination is possible only because Eqs. (3.1) and (3.3) are identical at the orders ϵ and ϵ^2. The Lax pair of the new integrable equation (1.3) is then

$$L\hat{\varphi}=0,\quad \hat{\varphi}_\tau+A\hat{\varphi}=0, \tag{3.7a}$$

with

$$L=-\frac{1}{2\sqrt{3}}\begin{pmatrix}\partial_\xi & (1+i)\Psi\\ (1-i)\Psi^* & \partial_\xi\end{pmatrix},\quad \hat{\varphi}=\begin{pmatrix}\varphi_+\\ \varphi_-\end{pmatrix} \tag{3.7b}$$

and

$$A=\begin{pmatrix}a_1\Phi+a_2(|\Psi|^2)_\xi+a_3\Psi^*\Psi_\xi & a_4\Psi_{\xi\xi}+a_5|\Psi|^2\Psi+a_6\Psi_\eta\\ a_4^*\Psi^*_{\xi\xi}+a_5^*|\Psi|^2\Psi^*+a_6^*\Psi^*_\eta & a_1^*\Phi+a_2(|\Psi|^2)_\xi+a_3\Psi\Psi^*_\xi\end{pmatrix}, \tag{3.7c}$$

$$a_1=\frac{i}{2},\quad a_2=-\frac{1}{6},\quad a_3=\frac{1}{3},\quad a_4=\frac{i+1}{2\sqrt{3}},\quad a_5=-\frac{i+1}{6\sqrt{3}},\quad a_6=\frac{1-i}{2\sqrt{3}}. \tag{3.7d}$$

IV. CONCLUSION

A new integrable nonlinear evolution PDE in $2+1$ dimensions has been derived from the KP equation, by means of a reduction method based on Fourier decomposition and space–time rescalings. We have applied the reduction method to the Lax pair of the original equation and have

demonstrated the integrability property of the new equation, by exhibiting the corresponding Lax. The reduction method is then able to identify integrable nonlinear evolution PDEs.

[1] F. Calogero and W. Eckhaus, ''Necessary conditions for integrability of nonlinear PDEs,'' Inverse Probl. **3**, L27–L32 (1987).
[2] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescalings, model PDEs and their integrability. I,'' Inverse Probl. **3**, 229–262 (1987).
[3] F. Calogero and W. Eckhaus, ''Nonlinear evolution equations, rescalings, model PDEs and their integrability. II,'' Inverse Probl. **4**, 11–33 (1988).
[4] F. Calogero and A. Maccari, ''Equations of nonlinear Schrödinger type in $1+1$ and $2+1$ dimensions obtained from integrable PDEs,'' *Inverse Problems: An Interdisciplinary Study*, Proceedings of the Meeting on Inverse Problems, Montpellier, 1986, Advances in Electronics and Electron Physics, edited by P. C. Sabatier (Academic, New York, 1988), Vol. 19, pp. 462–480.
[5] F. Calogero, ''Why are certain nonlinear PDEs both widely applicable and integrable,'' in *What is integrability (For Nonlinear PDEs)*, edited by V. E. Zakharov (Springer-Verlag, Berlin, 1991).
[6] A. Maccari, ''Dissipative bidimensional systems and resonant excitations,'' Int. J. Non-Linear Mech. **33**, 713–726 (1998).
[7] A. Maccari, ''The Boussinesq equation as a source of equations integrable by the spectral transform,'' Nonlinearity **10**, 849–856 (1997).
[8] A. Maccari, ''Universal and integrable nonlinear evolution systems of equations in $2+1$ dimensions,'' J. Math. Phys. **38**, 4151–4164 (1997).
[9] A. Maccari, ''The Kadomtsev–Petviashvili equation as a source of integrable model equations,'' J. Math. Phys. **37**, 6207–6212 (1996).
[10] A. Maccari, ''A model equation for non-resonant interacting ion acoustic plasma waves,'' J. Plasma Phys. (in press).
[11] B. B. Kadomtsev and V. I. Petviashvili, ''On the stability of solitary waves in weakly dispersive media,'' Sov. Phys. Dokl. **15**, 539–541 (1970).
[12] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[13] R. Hirota, ''Exact envelope-soliton solutions of a nonlinear wave equation,'' J. Math. Phys. **14**, 805–809 (1973).
[14] A. Davey and K. Stewartson, ''On three-dimensional packets of surface waves,'' Proc. R. Soc. London, Ser. A **338**, 101–110 (1974).
[15] D. Anker and N. C. Freeman, ''On the soliton solutions of the Davey–Stewartson equation for long waves,'' Proc. R. Soc. London, Ser. A **360**, 529–540 (1978).

Bifurcations of new eigenvalues for the Benjamin–Ono equation

Dmitry E. Pelinovsky and Catherine Sulem
Department of Mathematics, University of Toronto, Toronto, Ontario M5S3G3, Canada

(Received 20 April 1998; accepted for publication 14 August 1998)

A criterion for the emergence of new eigenvalues is found for the linear scattering problem associated with the Benjamin–Ono (BO) equation. This bifurcation occurs due to perturbations of nongeneric potentials which include the soliton solutions of the BO equation. The asymptotic approximation of an exponentially small new eigenvalue is derived. The method is based on the expansion of a localized function through a complete set of unperturbed eigenfunctions. Explicit expressions are obtained for the soliton potentials. © *1998 American Institute of Physics.* [S0022-2488(98)00112-1]

I. INTRODUCTION

This paper studies the problem of soliton generation in the integro-differential Benjamin–Ono (BO) equation,

$$u_t+2uu_x+Hu_{xx}=0, \tag{1.1}$$

where

$$Hu=\frac{1}{\pi}\,\text{p.v.}\int_{-\infty}^{\infty}\frac{u(y)dy}{y-x}$$

and p.v. stands for the principal value of the integral. The evolution equation (1.1) describes the propagation of long internal waves in a deep fluid (see Ref. 1 for review and references). The algebraic and spectral properties of the BO equation were intensely studied in development of the bilinear and inverse scattering transform methods.[2,3] The latter method allows one to relate the number and parameters of solitons generated by a smooth localized initial perturbation $u(x)=u(x,0)$ to the number and location of eigenvalues of the discrete spectrum for the linear spectral problem associated with the potential $u(x)$.

Generations of new solitons (and associated eigenvalues) were studied for several $(1+1)$-dimensional evolution equations in the limit of small and, alternatively, large number of solitons.[4] For instance, the Korteweg–de Vries (KdV) equation associated with the Schrödinger operator possesses *no threshold* for the generation of solitons by a small initial disturbance[4] while the modified KdV (mKdV) equation *does* possess a certain threshold on the area of the potential,

$$\mathcal{M}[u]=\int_{-\infty}^{\infty}u\,dx. \tag{1.2}$$

This threshold is related to spectral properties of the Dirac operator.[5] When the potential has a large amplitude, the problem under consideration has a uniform quasiclassical solution (see Ref. 4 for the KdV equation) which was derived also within the rigorous analysis of the dispersionless limit of the evolution equations.[6]

Surprisely enough, the problem of soliton generation for the BO equation (1.1) has not been finally solved since some independent results have appeared to be contradictory. Suppose that the initial perturbation can be renormalized in dimensionless units, $u=U_0U(x/L_0)$, where U_0 is a characteristic amplitude and L_0 is a characteristic length. Then, the number of solitons essentially

 © 1998 American Institute of Physics

depends on the sole parameter (the so-called Ursell parameter), $\sigma = U_0 L_0$. In the limit $\sigma \gg 1$, the potential generates a large number m of solitons, an approximation for m was found by Matsuno[7] and confirmed by Miloh *et al.*,[8]

$$m = \frac{1}{2\pi} \int_{u(x) \geqslant 0} u(x) dx. \tag{1.3}$$

Some early numerical and analytical data on the generation of solitons by an initial disturbance of an algebraic profile, $u = 2a/(1+x^2)$ were reported in Refs. 9 and 10. An important question, whether a small initial perturbation, i.e., that in the limit $\sigma \ll 1$ ($a \ll 1$), can support propagation of at least a single soliton, remained open however. This problem was first addressed by Pelinovsky and Stepanyants[11] who considered linear approximation in the pulse propagation. The linear part of the BO equation (1.1) has a self-similar solution preserving the area (1.2),

$$u = \frac{1}{t^{1/2}} U\left(\frac{x}{t^{1/2}} \right),$$

i.e., the characteristic amplitude and length of an initial pulse evolves as $U_0 \sim t^{-1/2}$ and $L_0 \sim t^{1/2}$, so that the Ursell parameter σ is effectively constant in time. These arguments imply that an initially small wave disturbance with $\sigma \ll 1$ remains effectively linear for all times t and hence does not support propagation of a soliton since the latter realizes a balance of nonlinear and dispersive effects for the finite value of σ.

Another way to predict the existence of a threshold for the soliton generation (see Ref. 8) is to use the fact that the area (1.2) is invariant for soliton solutions of Eq. (1.1). A single soliton is

$$u = u_s(x - vt) = \frac{2v}{1 + v^2 (x - vt - x_0)^2}, \tag{1.4}$$

where x_0, v are arbitrary parameters with the constraint $v > 0$. Using Eqs. (1.2) and (1.4), one concludes that $\mathcal{M}_s = \mathcal{M}[u_s] = 2\pi$, and thus the area does not depend on the amplitude of the soliton. For the mKdV equation, this property is related to the existence of threshold on the soliton generation, where perturbations with $\mathcal{M}[u] \leqslant \frac{1}{2}\mathcal{M}_s$ do not support a soliton.[5] Direct numerical simulations of the BO equation did not display formation of a soliton for an initial perturbation with $\mathcal{M} < \mathcal{M}_s$ (see Fig. 2 in Ref. 8). However, contrary to these preliminary predictions, we show that a soliton can still be generated in the BO equation (1.1) by a small initial perturbation without a threshold.

The problem at the center of the analysis is the linear spectral problem associated with the potential $u(x)$,

$$i\phi_x^+ + k(\phi^+ - \phi^-) = -u(x)\phi^+, \tag{1.5}$$

where k is a spectral parameter and $\phi^{\pm}(x)$ are the limit values, when x tends to the real axis, of analytical functions in upper/lower half-planes of x. The pioneer analysis of direct and inverse scattering problems for Eq. (1.5) was developed by Fokas and Ablowitz,[12] who showed that the continuous spectrum is located for real positive values of k while the discrete spectrum consists of isolated nondegenerate eigenvalues for real negative values of k. Various aspects of the spectral theory including in particular the complicated asymptotic behavior of the Jost functions (the eigenfunctions associated with the continuous spectrum) were addressed by many authors.[13–18]

Based on the ''improved'' eigenfunctions (i.e., those free of secular divergences at $k \to 0^+$), Coifman and Wickerhauser[15] proved that the scattering problem (1.5) has no bound states in a neighborhood of the origin, if $u(x) \to O(|x|^{-1-\mu})$, $\mu > 0$ (Theorem 7.1 of Ref. 15). Although this result seems to agree with qualitative arguments mentioned above, direct analysis of the problem (1.5) for the algebraic potential $u(x) = 2a/(1+x^2)$ did reveal the existence of a single negative eigenvalue for $0 < a < 1$.[10] The eigenvalue approaches zero very fast as $a \to 0^+$ and this fact makes the detection of the eigenvalue difficult for a small.

The reason for apparent discrepancy between the analytical and numerical results is caused by a very special structure of the Jost functions in the limit $k \to 0^+$. Theorem 7.1 of Ref. 15 is valid for *generic* potentials satisfying $n_0 \neq 0$, where

$$n_0 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} u(x) n(x) dx. \tag{1.6}$$

Here $n(x)$ is the limiting Jost function (see Eq. (2.13) below) which is logarithmically divergent as $k \to 0^+$. However, if the potential is *nongeneric*, i.e., it satisfies $n_0 = 0$, then the limiting Jost function is bounded in the limit $k \to 0^+$ and properties of the scattering problem are modified.[16,17] The zero potential $u(x) = 0$ as well as soliton solutions are particular examples of nongeneric potentials.

Recently, Kaup *et al.*[18] proved the completeness relation for the problem (1.5) and calculated variations of the scattering data under the action of a perturbation of the potential $u(x)$. These variational formulas are derived for generic potentials when the number of bound states is invariant under small perturbations of a potential.

In this paper we study the perturbation of nongeneric potentials where the number of bound states may change depending on the type of the perturbation. More precisely, we consider a potential $u(x)$ in the form $u^\epsilon = u(x) + \epsilon \Delta u(x)$, where $\epsilon \ll 1$ and $u(x)$ satisfies the constraint $n_0 = 0$. In the particular case $u(x) = 0$, this problem reduces to the problem of soliton generation by a small initial perturbation. We derive a criterion for a new eigenvalue to emerge from the edge of the continuous spectrum at $k = 0$. In particular, new eigenvalues may always appear due to perturbations of the zero background and the soliton solutions. We find that the asymptotic behavior of the new eigenvalue is *exponentially small* in terms of the parameter of the perturbation ϵ and obtain the corresponding bound state and the perturbed Jost functions of the continuous spectrum.

This result is to be compared to the classical spectral problem for the Schrödinger operator $-\Delta \psi + \epsilon V(x) \psi = \lambda \psi$, with $\epsilon \ll 1$.[19,20] As above, the problem is to find the condition on the potential so that there exists an eigenvalue λ that bifurcates from the edge of the continuous spectrum, and to compute, as $\epsilon \to 0$, its asymptotic value and the corresponding eigenfunction. In one space dimension, the bifurcation occurs when $\mathcal{M} = \int_{-\infty}^{\infty} V(x) dx < 0$ and one finds the eigenvalue $\lambda \sim -\epsilon^2 \mathcal{M}^2/4$.[19] The situation is more complex in two dimensions even with a radially symmetric potential. The criterion for the bifurcation to occur is $\mathcal{M} = \int_0^{\infty} V(r) r dr < 0$ and the eigenvalue is $\lambda \sim -\lambda_0 \exp(2/\epsilon \mathcal{M})$, where λ_0 can be calculated explicitly in terms of integrals of the potential.[21] Note that the behavior of the new eigenvalue for the BO spectral problem is very similar to the latter case.

For the Schrödinger spectral problem, the analysis of the bifurcation can be performed by various methods such as a direct two-scale asymptotic expansion technique,[19] the analysis based on the positive definiteness of the self-adjoint operators,[20] the analysis involving the Green's function representation,[21] or the study of analytical behavior of transmission coefficients in the scattering problem or, their generalizations given by the Evans' function.[22]

The BO spectral problem has the form of an integro-differential equation which make direct applications of the above methods difficult. In this article, we develop a new approach based on the completeness of Jost functions of the unperturbed problem.

The plan of this article is as follows. In Sec. II we recall the main results on the spectral properties of Eq. (1.5). The regular perturbation theory is described in Sec. III. The bifurcation of a new negative eigenvalue from the origin cannot be obtained through a regular perturbation method and we develop a revised method in Sec. IV to analyze this bifurcation. We derive the leading order terms for the new eigenvalue and for the associated bound state at short and large distances, as well as for the variation of the continuous spectrum data. Explicit expressions are given in Sec. V for the soliton potentials. Section VI is devoted to discussions. Appendix A gives useful relations for eigenfunctions of the BO scattering problem and Appendix B presents new relations for the roots of the Laguerre polynomials which appear in the course of our analysis.

II. SPECTRAL PROPERTIES OF THE BO ASSOCIATED PROBLEM

In this section, we recall some basic results on the direct scattering problem (1.5) that will be used in the analysis. We assume that the potential $u=u(x)$ is real and has a sufficient decay to ensure the convergence of the area integral (1.2), i.e., $u\sim O(|x|^{-2-\mu})$ as $x\to\infty$, $\mu>0$. The details and proofs of these results can be found in Refs. 12 and 17.

A. Spectrum and scattering data

For all $k>0$, there are two solutions N and $\overline{N}$ (the Jost functions) to Eq. (1.5), defined by their boundary conditions as $x\to\infty$,

$$\overline{N}(x,k)\to 1, \quad N(x,k)\to e^{ikx}. \tag{2.1}$$

Using the projection operators

$$P^{\pm}(v)=\pm\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{v(y)dy}{y-(x\pm i0)}=\frac{1}{2}\,(v\mp iHv), \tag{2.2}$$

it is proved that $\overline{N}$ and N satisfy the integro-differential equations

$$i\overline{N}_x+k\overline{N}=-P^{+}(u\overline{N})+k, \tag{2.3a}$$

$$iN_x+kN=-P^{+}(uN). \tag{2.3b}$$

Equation (2.1) has also a discrete number of bound states $\{\Phi_j(x)\}_{j=1}^{m}$, where m is finite, associated to discrete eigenvalues $k_j<0$. They satisfy

$$i\Phi_{jx}+k_j\Phi_j=-P^{+}(u\Phi_j), \tag{2.3c}$$

with the boundary condition

$$\Phi_j(x)\to\frac{1}{x} \text{ as } x\to\infty.$$

We denote ϕ^{+} the set of eigenfunctions $\phi^{+}=\{\overline{N}(x,k);N(x,k);\{\Phi_j(x)\}_{j=1}^{m}\}$. Alternatively, one can consider the set of eigenfunctions $\{M(x,k);\overline{M}(x,k);\{\Phi_j(x)\}_{j=1}^{m}\}$ which satisfy the same equations (2.3) but with the left-side boundary conditions

$$M(x,k)\to 1, \quad \overline{M}(x,k)\to e^{ikx}, \quad \Phi_j(x)\to\frac{1}{x}, \tag{2.4}$$

as $x\to-\infty$. These two sets are related by the scattering equations

$$M(x,k)-\overline{N}(x,k)=\beta(k)N(x,k), \quad \overline{M}(x,k)=\Gamma(k)N(x,k), \quad k>0. \tag{2.5}$$

In the case of a real potential $u(x)$, the scattering coefficients are given by[17]

$$\Gamma(k)=\exp\left[\frac{1}{2\pi i}\int_0^k\frac{|\beta(k')|^2}{k'}\,dk'\right] \tag{2.6a}$$

and

$$\beta(k)=i\int_{-\infty}^{\infty}u(x)N^{*}(x,k)dx. \tag{2.6b}$$

The scattering equation (2.5) relates the functions $M(x,k)$ and $\overline{N}(x,k)$ which are analytic in $\mathrm{Im}(k)>0$ and $\mathrm{Im}(k)<0$, respectively. It has the form of a Riemann–Hilbert problem.[3] In Ref. 12, Fokas and Ablowitz derived a solution of this Riemann–Hilbert problem for the function $\overline{N}(x,k)$ at $\mathrm{Im}(k)\leqslant 0$, in terms of $N(x,k)$ for real k,

$$\overline{N}(x,k)=1-i\sum_{j=1}^{m}\frac{\Phi_j(x)}{k-k_j}+\frac{1}{2\pi i}\int_0^\infty \frac{\beta(k')N(x,k')dk'}{k'-(k-i0)}. \tag{2.7}$$

Here simple poles correspond to the discrete spectrum located at m negative eigenvalues $k_j<0$ while the integral represents the contribution from the continuous spectrum located for $k>0$. The scattering data form a closed system when Eq. (2.7) is coupled with the relation,

$$\frac{\partial N(x,k)}{\partial k}=ixN(x,k)+\frac{\beta^*(k)}{2\pi ik}\overline{N}(x,k) \tag{2.8a}$$

and the limiting behavior of $\overline{N}(x,k)$ as $k\to k_j$ is to be used,

$$\lim_{k\to k_j}\left[\overline{N}(x,k)+\frac{i\Phi_j(x)}{k-k_j}\right]=(x+\gamma_j)\Phi_j(x), \tag{2.8b}$$

where γ_j are the phase constants. The spectral data $\{k_j,\gamma_j\}_{j=1}^m$ can be expressed through the eigenfunctions $\Phi_j(x)$ according to the expressions,

$$k_j=\frac{1}{2\pi i}\int_{-\infty}^{\infty}u(x)\Phi_j(x)dx \tag{2.9a}$$

and

$$\gamma_j=\frac{1}{2\pi k_j}\int_{-\infty}^{\infty}x|\Phi_j(x)|^2dx-\frac{i}{2k_j}. \tag{2.9b}$$

Equation (2.9a) follows from Eq. (2.3c) (see Refs. 12 and 17), while Eq. (2.9b) can be obtained from Eq. (A4a) of Appendix A.

It was recently found for other spectral problems of inverse scattering[23] that the coalescence of two simple poles may occur and, in this case, the proper solution to the Riemann–Hilbert problem must include also multiple poles corresponding to degenerate eigenvalues. However, this coalescence is not possible for the BO spectral problem (1.5) associated with the real potentials. Indeed, suppose that $k_1=k_2+\epsilon$, where $\epsilon\ll 1$ and expand a solution to Eqs. (2.7) and (2.8) into an asymptotic series,[23]

$$\begin{cases}\Phi_1=-\dfrac{1}{\epsilon}\,\Phi^{(-1)}(x)+\Phi_1^{(0)}(x)+O(\epsilon),\\ \Phi_2=\dfrac{1}{\epsilon}\,\Phi^{(-1)}(x)+\Phi_2^{(0)}(x)+O(\epsilon),\end{cases} \tag{2.10a}$$

$$\begin{cases}\gamma_1=\dfrac{i}{\epsilon}+\gamma^{(0)}+O(\epsilon),\\ \gamma^2=-\dfrac{i}{\epsilon}+\gamma^{(0)}+O(\epsilon).\end{cases} \tag{2.10b}$$

This expansion provides a double pole in the limit $\epsilon\to 0$ for the function $\overline{N}(x,k)$ and satisfy the limiting relations (2.8b). However, it contradicts the balance prescribed by Eq. (2.9b), the latter being valid for real potential $u(x)$. Therefore, multiple poles are not present in the representation (2.7) and the eigenvalues $k=k_j$ for $j=\overline{1,m}$ are all simple.

The spectral data $\{\beta(k)\},\{k_j,\gamma_j\}_{j=1}^m$ characterize the BO equation as a completely integrable system[18] and they can be calculated from the Jost functions $N(x,k)$ and the bound states $\Phi_j(x)$ by means of Eqs. (2.6) and (2.9). The functions $N(x,k)$ and $\Phi_j(x)$ satisfy a self-consistent system of integral and algebraic equations[17] which is a crucial step in inverse scattering. The potential $u(x,t)$ can be recovered for $t\geqslant 0$ by means of the formula,

$$u=\frac{1}{2\pi i}\int_0^\infty \beta(k')N(x,k')dk'+i\sum_{j=1}^m \Phi_j(x)+\text{c.c.} \tag{2.11}$$

if the time evolution of the spectral data is taken into account.[12,17] Inverse scattering results will not be used for the problem under consideration. Instead, we mention the well-known fact[12] that the continuous spectrum of the scattering problem (1.5) corresponds to the dispersive wave packets (radiation) induced by the initial potential $u(x)$ while the bound states of the discrete spectrum correspond to solitons (1.4) with the parameters $v=-2k_j$ and $x_0=-\text{Re}(\gamma_j)$. These parameters induced by the initial potential are constant in the time, while the solitons become separated due to different values of their velocities.

B. Asymptotic behavior of the Jost functions

The asymptotic behavior of the Jost functions and of the scattering data as $k\to 0^+$ can be studied by means of their Green's function representation.[16,17] For generic potentials, the Jost functions vanish as $k\to 0^+$ according to the scaling approximation,

$$N(x,k)\to\frac{n(x)}{1+n_0(\gamma+\ln(ik))}+O\left(\frac{k}{\ln k}\right), \tag{2.12}$$

where $\gamma\approx 0.577$ is the Euler's constant, n_0 is defined by Eq. (1.6) and $n(x)$ satisfies

$$in_x=-P^+(un). \tag{2.13a}$$

Equivalently, $n(x)$ is the unique solution to the integral equation

$$n(x)=1-i\int_x^\infty u(y)n(y)dy-\frac{1}{2\pi}\int_{-\infty}^\infty u(y)n(y)\ln(x-y-i0)dy. \tag{2.13b}$$

We notice that the representation (2.12) describes an improved approximation for the asymptotic behavior of the Jost function compared to the conventional expansion,[13]

$$N(x,k)\to\frac{n(x)}{n_0\ln k}+O\left(\frac{1}{(\ln k)^2}\right). \tag{2.14a}$$

Using Eqs. (2.6) and (2.14a) we find the uniform asymptotic representation for the scattering coefficients as $k\to 0^+$,

$$\beta(k)\to\frac{2\pi i}{\ln k}+O\left(\frac{1}{(\ln k)^2}\right),\quad \Gamma(k)=1+\frac{2\pi i}{\ln k}+O\left(\frac{1}{(\ln k)^2}\right). \tag{2.14b}$$

It follows from Eq. (2.13b) that the limiting function $n(x)$ is logarithmically secular and violates the boundary condition (2.1)

$$n(x)\to\begin{cases}1-n_0\ln x & \text{as } x\to\infty\\ 1-n_0\,\text{Ln}\,x & \text{as } x\to\text{-}\infty,\end{cases} \tag{2.15}$$

where Ln x is the principal value of the logarithmic function. Thus, the limiting point $k=0$ does not belong to the continuous spectrum for generic potentials, i.e., $N(x,k)\to 0$ as $k\to 0^+$.

On the contrary, the condition $n_0=0$ defines a class of nongeneric potentials which include the zero background $u(x)=0$ and soliton solutions. In this case, the limiting function $n(x)$ is not secular [see Eq. (2.15)] and the limiting point $k=0$ *does* belong the continuous spectrum, i.e.,

$$N(x,k) \to n(x) + k n_1(x) + O(k^2 \ln k), \tag{2.16a}$$

and

$$\beta(k) \to k\beta_1 + O(k^2 \ln k)$$

$$\Gamma(k) \to 1 + \frac{|\beta_1|^2}{4\pi i} k^2 + O(k^3 \ln k) \tag{2.16b}$$

as $k \to 0^+$. In order to prove validity of the expansions (2.16), we use the relations (2.6b) and (2.8a) in the limit $k \to 0^+$. At leading order, $N(x,k) \sim n(x)$ as $k \to 0^+$.[17] As a result, we find from Eq. (2.8a) the following representation for $n_1(x)$:

$$n_1(x) = i\left(x - \frac{\beta_1^*}{2\pi} \right) n(x), \tag{2.17a}$$

and from Eq. (2.6b) the explicit formula for β_1,

$$\beta_1 = \text{p.v.} \int_{-\infty}^{\infty} x u(x) n^*(x) dx. \tag{2.17b}$$

Any perturbation to the nongeneric potential changes the asymptotic behavior as $k \to 0^+$ and may lead to the occurrence of a new bound state for $k < 0$. This bifurcation is studied in Sec. IV.

C. Number of bound states

For both generic and nongeneric potentials, the total area (1.2) is related to the number of bound states m by

$$\mathcal{M}[u] = 2\pi m - \frac{1}{2\pi} \int_0^{\infty} \frac{|\beta(k)|^2}{k} \, dk. \tag{2.18}$$

To obtain this formula (see Ref. 17), one writes Eq. (2.7) at $k = 0$, multiplies it by u and integrates over $\mathbf{R}$. The integral $\int_{-\infty}^{\infty} u(x) N(x,0) dx = 0$ for both generic and nongeneric potentials. The other terms are evaluated using Eqs. (2.6b) and (2.9a).

Another way to see that

$$\frac{1}{2\pi} \left(\mathcal{M}[u] + \frac{1}{2\pi} \int_0^{\infty} \frac{|\beta(k)|^2}{k} dk \right)$$

is an integer is to take the limit $k \to \infty$, $x \to -\infty$ in the second equation (2.5) and to use Eqs. (2.4), (2.6a) as well as the asymptotic representation of $N(x,k)$ for $k \to \infty$,[17]

$$N(x,k) \sim \exp\left[ikx - i \int_x^{\infty} u(x) dx \right] \quad \text{for real } x. \tag{2.19}$$

In agreement with the distribution (1.3) for large m, the positive regions of the potential $u(x)$ give birth to bound states (solitons) while the negative regions are sources of continuous spectrum (radiation).

III. REGULAR PERTURBATION THEORY

Suppose that the scattering problem (1.5) for $u = u(x)$ has m bound states $\Phi_j(x)$ for $k = k_j < 0$, $j = 1,m$ and let $N(x,k)$ be the Jost functions for $k > 0$. In Ref. 18, Kaup *et al.* proved the orthogonality conditions

$$\int_{-\infty}^{\infty} N^*(x,k') N(x,k) dx = 2\pi \delta(k - k'), \tag{3.1a}$$

$$\int_{-\infty}^{\infty} N^*(x,k)\Phi_j(x)dx=0, \tag{3.1b}$$

$$\int_{-\infty}^{\infty} \Phi_l^*(x)\Phi_j(x)dx=-2\pi k_j\delta_{jl} \tag{3.1c}$$

as well as the completeness relation,

$$\int_0^{\infty} N^*(y,k)N(x,k)dk-\sum_{j=1}^{m} \frac{\Phi_j^*(y)\Phi_j(x)}{k_j}=\frac{-i}{y-(x+i0)}. \tag{3.2}$$

Any function $\Phi^+(x)$ such that $\Phi^+\sim O(|x|^{-1-\mu})$ as $x\to\infty$, $\mu\geqslant 0$, can be expanded over the basis $\{N(x,k);\{\Phi_j(x)\}_{j=1}^m\}$ as follows

$$\Phi^+(x)=\int_0^{\infty} \frac{\alpha(k)N(x,k)}{k-\kappa}\,dk+\sum_{j=1}^{m} \frac{\alpha_j\Phi_j(x)}{k_j-\kappa}, \tag{3.3}$$

where $\alpha(k)$, α_j are coefficients of the expansion and the arbitrary parameter κ is introduced for convenience in the following analysis and will be specified in the next subsection. The coefficients $\alpha(k)$ and α_j can be expressed through the function $\Phi^+(x)$ using the orthogonality relations (3.1).

A classical problem of perturbation theory is to study variations of the spectral data subject to a small variation of the potential. We assume that the perturbed potential is expressed as $u^\epsilon(x)=u(x)+\epsilon\Delta u(x)$, where $\epsilon\ll 1$, and the functions $u(x)$ and $\Delta u(x)$ decay at infinity as $u(x),\Delta u(x)\sim O(|x|^{-2-\mu})$ with $\mu\geqslant 0$. Here and in subsequent sections, we study the variations of the spectral data for perturbed generic and nongeneric potentials.

A. Variations of data of the discrete spectrum

Suppose that $\Phi^+(x)$ solves Eq. (2.3c) with $u^\epsilon=u(x)+\epsilon\Delta u(x)$ and κ is the corresponding eigenvalue. The linear eigenvalue problem (2.3c) can be reduced to a set of integral equations for the coefficients $\alpha(k)$ and α_j in the form

$$\alpha(k)=\frac{\epsilon}{2\pi}\left[\int_0^{\infty} \frac{K(k,k')\alpha(k')}{k'-\kappa}\,dk'+\sum_{j=1}^{m} \frac{K_j(k)\alpha_j}{k_j-\kappa}\right], \tag{3.4a}$$

$$\alpha_j=-\frac{\epsilon}{2\pi k_j}\left[\int_0^{\infty} \frac{K_j^*(k')\alpha(k')}{k'-\kappa}\,dk'+\sum_{l=1}^{m} \frac{K_{jl}\alpha_l}{k_l-\kappa}\right], \tag{3.4b}$$

where

$$K(k,k')=\int_{-\infty}^{\infty}\Delta u(x)N^*(x,k)N(x,k')dx,$$

$$K_j(k)=\int_{-\infty}^{\infty}\Delta u(x)N^*(x,k)\Phi_j(x)dx,$$

$$K_{jl}=\int_{-\infty}^{\infty}\Delta u(x)\Phi_j^*(x)\Phi_l(x)dx.$$

Consider a particular eigenvalue $\kappa=k_j^\epsilon<0$ and the corresponding bound state $\Phi^+=\Phi_j^\epsilon(x)$. The analysis of Eq. (3.4b) shows that the asymptotic balance is satisfied by the following Taylor expansion:

$$k_j^\epsilon=k_j+\epsilon\Delta k_j+\epsilon^2\Delta_2 k_j+O(\epsilon^3) \tag{3.5a}$$

and

$$\alpha_j=(k_j-k_j^\epsilon)(1+\epsilon\alpha_j^{(1)}+O(\epsilon^2)), \tag{3.5b}$$

$$\alpha_l=-\frac{\epsilon K_{lj}}{2\pi k_l}+O(\epsilon^2), \quad l\neq j, \tag{3.5c}$$

$$\alpha(k)=\frac{\epsilon K_j(k)}{2\pi}+O(\epsilon^2). \tag{3.5d}$$

The correction Δk_j is obtained from Eq. (3.4b) as

$$\Delta k_j=\frac{1}{2\pi k_j}K_{jj}. \tag{3.6}$$

We then proceed with the first-order correction term in $\Phi_j^\epsilon(x)$ which is given through Eqs. (3.3) and (3.4) as $\Phi_j^\epsilon=\Phi_j+\epsilon\Delta\Phi_j+O(\epsilon^2)$, where

$$\Delta\Phi_j=\alpha_j^{(1)}\Phi_j+\frac{1}{2\pi}\int_0^\infty\frac{K_j(k)N(x,k)}{k-k_j}\,dk-\sum_{l\neq j}\frac{K_{lj}\Phi_l(x)}{2\pi k_l(k_l-k_j)}. \tag{3.7}$$

The constant $\alpha_j^{(1)}$ can be found from the boundary condition (2.3) by removing the $O(x^{-1})$ term as $x\to\infty$. Using the boundary conditions (A3) of Appendix A this leads to the expression,

$$\alpha_j^{(1)}=\sum_{l\neq j}\frac{K_{lj}}{2\pi k_l(k_l-k_j)}-\frac{K_j(0)}{2\pi i k_j}-\frac{1}{4\pi^2}\int_0^\infty\frac{\beta^*(k)K_j(k)}{k(k-k_j)}\,dk. \tag{3.8a}$$

This formula is equivalent to that given in Ref. 18,

$$\alpha_j^{(1)}=-\frac{1}{2\pi i k_j}\int_{-\infty}^\infty\Delta u(x)(x+\gamma_j)|\Phi_j(x)|^2dx. \tag{3.8b}$$

In order to prove this, we rewrite Eq. (3.8b) as

$$\alpha_j^{(1)}=-\frac{1}{2\pi i k_j}\left(\int_{-\infty}^\infty P^+(\Delta u\Phi_j)dx+\int_{-\infty}^\infty(x+\gamma_j)\Phi_j^*P^+(\Delta u\Phi_j)dx\right). \tag{3.9}$$

Using the completeness formula (3.2) we represent $P^+(\Delta u\Phi_j)$ as

$$P^+(\Delta u\Phi_j)=\frac{1}{2\pi}\int_0^\infty K_j(k)N(x,k)dk-\sum_{l=1}^m\frac{K_{lj}\Phi_l(x)}{2\pi k_l}. \tag{3.10}$$

Substituting Eq. (3.10) into Eq. (3.9) and using Eqs. (A2) and (A4) of Appendix A we recover Eq. (3.8a).

Applying the same method, we expand the relations (2.9) into the asymptotic series for k_j^ϵ [see Eq. (3.5a)] and for γ_j^ϵ, $\gamma_j^\epsilon=\gamma_j+\epsilon\Delta\gamma_j+O(\epsilon^2)$. The first-order perturbation terms are thus expressed as

$$\Delta k_j=\frac{1}{2\pi i}\int_{-\infty}^\infty(\Delta u\Phi_j+u\Delta\Phi_j)dx, \tag{3.11a}$$

$$\Delta\gamma_j=\frac{1}{2\pi k_j}\int_{-\infty}^\infty(x+\gamma_j)(\Phi_j\Delta\Phi_j^*+\Phi_j^*\Delta\Phi_j)dx. \tag{3.11b}$$

Substitution of Eqs. (3.7) and (3.10) into (3.11a) reproduces the result (3.6), while substitution of Eqs. (3.7), (3.8a), and (A4) into Eq. (3.11b) leads to the expression,

$$\Delta\gamma_j = -\frac{1}{2\pi k_j^2}\int_{-\infty}^{\infty}\Delta u(x)(x+\gamma_j)|\Phi_j(x)|^2dx - \frac{i}{2\pi k_j}\sum_{l\neq j}\frac{K_{lj}-K_{jl}}{(k_l-k_j)^2} + \frac{i}{4\pi^2 k_j}\int_0^{\infty}\frac{\beta^*(k)K_j(k)-\beta(k)K_j^*(k)}{(k-k_j)^2}\,dk. \tag{3.12}$$

The formulas (3.6)–(3.8) and (3.12) coincide with the variational derivatives found in Ref. 18 with a different approach. In addition, we find from Eq. (3.4b) the second-order correction to the eigenvalue,

$$\Delta_2 k_j = \frac{1}{2\pi k_j}\left(\int_0^{\infty}\frac{|K_j(k)|^2}{k-k_j}\,dk - \sum_{l\neq j}\frac{|K_{jl}|^2}{2\pi k_l(k_l-k_j)}\right). \tag{3.13}$$

These results are valid both for generic and nongeneric potentials, i.e., those with the asymptotic behavior (2.14) and (2.16), respectively. The variations of data of the discrete spectrum for the BO spectral problem are similar to standard form of perturbation theory in quantum mechanics.[19]

B. Variations of data of the continuous spectrum

The perturbation theory for the Jost function can be constructed with the help of a modified integral representation,

$$N^{\epsilon}(x,\kappa) = N(x,\kappa) + \int_0^{\infty}\frac{\alpha(k,\kappa)N(x,k)}{k-(\kappa-i0)}\,dk + \sum_{j=1}^{m}\frac{\alpha_j(\kappa)\Phi_j(x)}{k_j-\kappa}, \tag{3.14}$$

where $N^{\epsilon}(x,\kappa)$ solves Eq. (2.3b) for $u^{\epsilon}=u(x)+\epsilon\Delta u(x)$ and $k=\kappa$. According to the boundary conditions (2.1), (2.4) and the scattering problem (2.5), Eq. (3.14) reduces in the limit $x\to-\infty$ to the following relation:

$$\frac{1}{\Gamma^{\epsilon}(k)} - \frac{1}{\Gamma(k)} = -\frac{2\pi i\alpha(k,k)}{\Gamma(k)}. \tag{3.15}$$

The coefficients $\alpha(k,\kappa)$ and $\alpha_j(\kappa)$ satisfy a system of integral equations modified compared to Eqs. (3.4),

$$\alpha(k,\kappa) = \frac{\epsilon}{2\pi}\left[K(k,\kappa) + \int_0^{\infty}\frac{K(k,k')\alpha(k',\kappa)}{k'-(\kappa-i0)}\,dk' + \sum_{j=1}^{m}\frac{K_j(k)\alpha_j(\kappa)}{k_j-\kappa}\right], \tag{3.16a}$$

$$\alpha_j(\kappa) = -\frac{\epsilon}{2\pi k_j}\left[K_j^*(\kappa) + \int_0^{\infty}\frac{K_j^*(k')\alpha(k',\kappa)}{k'-(\kappa-i0)}\,dk' + \sum_{l=1}^{m}\frac{K_{jl}\alpha_l(\kappa)}{k_l-\kappa}\right]. \tag{3.16b}$$

As $\epsilon\to 0$, these integral equations can be solved by a Taylor expansion,

$$\alpha(k,\kappa) = \frac{\epsilon}{2\pi}K(k,\kappa) + O(\epsilon^2), \tag{3.17a}$$

$$\alpha_j(\kappa) = -\frac{\epsilon}{2\pi k_j}K_j^*(\kappa) + O(\epsilon^2). \tag{3.17b}$$

Using this explicit asymptotic behavior, the spectral coefficient can also be expanded in the Taylor series, $\Gamma^{\epsilon}(k)=\Gamma(k)+\epsilon\Delta\Gamma(k)+O(\epsilon^2)$, where

$$\frac{\Delta\Gamma(k)}{\Gamma(k)} = iK(k,k). \tag{3.18}$$

Since $\lim_{k\to\infty}K(k,k)=\int_{-\infty}^{\infty}\Delta u(x)dx\equiv\Delta\mathcal{M}$, it follows from Eqs. (2.6a) and (3.18) that

$$\Delta \mathcal{M} = -\frac{1}{2\pi} \Delta \int_0^\infty \frac{|\beta(k)|^2}{k}\, dk. \tag{3.19}$$

This relation is nothing but the variation of Eq. (2.18) which proves that the number m of solitons is invariant under a small change of the potential, i.e., $\Delta m = 0$. However, this result as well as formulas (3.17)–(3.19) of the regular perturbation theory are valid only for generic potentials. When the potential is nongeneric, i.e., it possesses the asymptotic representation (2.16), the regular perturbation theory fails in the limit $k \to 0^+$. Indeed, using Eqs. (2.16) we find that

$$\Delta \Gamma(0) = i \int_{-\infty}^{\infty} \Delta u(x) |n(x)|^2 dx \neq 0. \tag{3.20}$$

This result contradicts the asymptotic behavior (2.14b), which provides $\Delta\Gamma(k) \to 0$ as $k \to 0^+$. For comparison, the relation (3.18) for generic potentials exhibits the limit behavior, $\Delta\Gamma(k) \sim O((\ln k)^{-2})$ as $k \to 0^+$ which is in agreement with Eq. (2.14b). The reason for failure of the regular perturbation theory is related to the fact that the integral in Eq. (3.14) becomes singular as $\kappa \to 0^+$. Thus, the perturbation theory must be revised for perturbations of the nongeneric potentials. This revision is closely related to analysis of bifurcation of a new eigenvalue in a neighborhood of $k = 0$.

IV. REVISED PERTURBATION THEORY

Since multiple degenerate eigenvalues are not permitted in the BO spectral problem, as shown in Sec. II A, the only possible bifurcation may be a formation of a new bound state from the delocalized limiting eigenfunction of the continuous spectrum. In this section we show that this bifurcation can happen due to perturbations of the nongeneric potentials. Using the integral representation (3.3) and (3.4) we find the asymptotic approximations for the new eigenvalue and the corresponding eigenfunction.

A. Asymptotic series for a new eigenvalue

Suppose that $\kappa = k_{m+1} = -p(\epsilon)$, where k_{m+1} is a new eigenvalue detaching from the continuum as $\epsilon \to 0$ and $p(\epsilon) > 0$. The integral in Eq. (3.4a) is singular as $\kappa \to 0^-$ unless $K(k,0) = 0$. The latter condition is always true for the generic potentials since $N(x,k)$ vanishes in the limit $k \to 0^+$. However, for nongeneric potentials, since $N(x,0) \sim n(x)$ [see Eq. (2.16a)], $K(k,0)$ is generally nonzero. Thus, the singularity in Eq. (3.4a) is of resonance pole type and accounts for a nontrivial solution of integral equations which corresponds to a new bound state $\Phi^+ = \Phi_{m+1}(x)$.

In order to construct an asymptotic solution to Eqs. (3.4) at $\kappa = -p(\epsilon)$ we evaluate the singular contribution explicitly,

$$\alpha(k) = -\frac{\epsilon}{2\pi} \ln\left(\frac{p}{1+p}\right) K(k,0)\alpha(0) + \frac{\epsilon}{2\pi}\left[\int_0^1 \frac{K(k,k')\alpha(k') - K(k,0)\alpha(0)}{k'+p}\, dk' + \int_1^\infty \frac{K(k,k')\alpha(k')}{k'+p}\, dk' + \sum_{j=1}^m \frac{K_j(k)\alpha_j}{k_j+p}\right]. \tag{4.1}$$

[The other equation (3.4b) can be transformed in a similar manner.] We notice that the integrals in Eq. (4.1) are now free of singular terms in the limit $p(\epsilon) \to 0^+$. [Still a Taylor expansion with respect to $p(\epsilon)$ would lead to new (weak) singular terms in Eq. (4.1). This modification is required only for calculation of exponentially small corrections of the asymptotic series (4.2) and will be neglected henceforth.] The asymptotic balance of the singular term occurs for $\epsilon \ln p \sim O(1)$ as $\epsilon \to 0$. Thus, the new eigenvalue is exponentially small in terms of ϵ. Under the balance above, the asymptotic series solving Eq. (4.1) can be presented in the form,

$$\ln p = \frac{1}{\epsilon} p_{-1} + \sum_{l=0}^{\infty} \epsilon^l p_l + O(\epsilon^{-1} p(\epsilon)), \tag{4.2a}$$

$$\alpha(k)=\alpha_0(k)+\sum_{l=1}^{\infty}\epsilon^l\alpha_l(k)+O(p(\epsilon)), \tag{4.2b}$$

$$\alpha_j=\alpha_j^{(0)}+\sum_{l=1}^{\infty}\epsilon^l\alpha_j^{(l)}+O(p(\epsilon)). \tag{4.2c}$$

We require $p_{-1}<0$ to ensure the eigenvalue $p=p(\epsilon)$ to be small in the limit $\epsilon\to 0$. All terms of the Taylor series for $\alpha(k)$ and α_j can be found recursively in an explicit form from Eq. (4.1) by neglecting the exponentially small residue terms, i.e., the terms containing the eigenvalue p. We confine our analysis only to the leading-order and first-order approximations in these asymptotic series. The leading order term is

$$\alpha_0(k)=-\frac{p_{-1}}{2\pi}K(k,0)\alpha_0(0). \tag{4.3}$$

The coefficient p_{-1} follows from this equation in the limit $k\to 0$,

$$p_{-1}=-\frac{2\pi}{K(0,0)}. \tag{4.4}$$

Since the restriction $p_{-1}<0$ was assumed in Eq. (4.2a), the following criterion,

$$K(0,0)=\int_{-\infty}^{\infty}\Delta u(x)|n(x)|^2dx>0, \tag{4.5}$$

should be satisfied for the bifurcation of the new eigenvalue $k=k_{m+1}=-p(\epsilon)$ to occur.

In order to find the pre-exponential factor p_0 in the asymptotic expansion for $p=p(\epsilon)$, we derive from Eqs. (3.4b) and (4.1) the following expressions for $\alpha_j^{(0)}$ and $\alpha_1(k)$:

$$\alpha_j^{(0)}=-\frac{K_j^*(0)}{k_jK(0,0)}\alpha_0(0) \tag{4.6a}$$

and

$$\begin{aligned}\alpha_1(k)=&\frac{K(k,0)}{K(0,0)}\alpha_1(0)-\frac{p_0}{2\pi}K(k,0)\alpha_0(0)\\&+\frac{\alpha_0(0)}{2\pi K(0,0)}\left[\int_0^1\frac{K(k,k')K(k',0)-K(k,0)K(0,0)}{k'}dk'\right.\\&\left.+\int_1^{\infty}\frac{K(k,k')K(k',0)}{k'}dk'-\sum_{j=1}^{m}\frac{K_j(k)K_j^*(0)}{k_j^2}\right].\end{aligned} \tag{4.6b}$$

Then, the coefficient p_0 follows from Eq. (4.6b) in the limit $k\to 0$,

$$\begin{aligned}p_0=&\frac{1}{[K(0,0)]^2}\int\int_{-\infty}^{\infty}\Delta u(x)\Delta u(y)n^*(x)n(y)dxdy\left[\int_0^1\frac{N^*(y,k)N(x,k)-n^*(y)n(x)}{k}dk\right.\\&\left.+\int_1^{\infty}\frac{N^*(y,k)N(x,k)}{k}dk-\sum_{j=1}^{n}\frac{\Phi_j^*(y)\Phi_j(x)}{k_j^2}\right].\end{aligned} \tag{4.7}$$

This formula can be simplified if the nongeneric potential reduces to a pure m-soliton solution (see examples in Sec. V). We formulate the main result of our analysis in the following statement.

Proposition 1: Suppose that the potential $u(x)$ satisfies the constraint $n_0=0$ and the perturbation $\Delta u(x)$ satisfies the criterion $K(0,0)>0$. Then, the potential $u^{\epsilon}=u(x)+\epsilon\Delta u(x)$ supports an eigenvalue $k=-p(\epsilon)$ in a neighborhood of $k=0$ for $\epsilon>0$, where

$$p(\epsilon)=c\ \exp(\epsilon^{-1}p_{-1})[1+O(\epsilon)] \tag{4.8}$$

and $c=e^{p_0}$. *The coefficients* p_{-1}, p_0 *are given by Eqs. (4.4) and (4.7). No eigenvalue exists in a neighbourhood of* $k=0$ *for* $\epsilon<0$.

We notice that if $K(0,0)=0$, the asymptotic approximation of $\ln p$ has generally the $O(\epsilon^{-2})$ term and the asymptotic expressions (4.2)–(4.8) must be reconsidered. This special situation occurs if the perturbation $\Delta u(x)$ still conserves the constraint $n_0=0$ for the limiting Jost function $n(x)$ extended to the first-order of the perturbation theory. Alternatively, if $K(0,0)<0$, the bifurcation analyzed above occurs for $\epsilon<0$.

B. The bound state corresponding to the new eigenvalue

Proposition 2: The eigenfunction $\Phi_{m+1}(x)$ *corresponding to the eigenvalue* $k=k_{m+1}=-p(\epsilon)$ *has the asymptotic form*

$$\Phi_{m+1}(x)=\frac{\alpha_0(0)}{K(0,0)}\int_0^\infty \frac{K(k,0)N(x,k)}{k+p(\epsilon)}\,dk+O(\epsilon). \tag{4.9}$$

The spectral data γ_{m+1} *defined in Eq. (2.9b) is given by*

$$\gamma_{m+1}=-\frac{\int_{-\infty}^{\infty}x\Delta u(x)|n(x)|^2dx}{\int_{-\infty}^{\infty}\Delta u(x)|n(x)|^2dx}+\frac{\beta_1+\beta_1^*}{4\pi}(1+\ln p)+\frac{i}{2p}, \tag{4.10}$$

where β_1 *is defined in Eq. (2.17b).*

The eigenfunction $\Phi_{m+1}(x)$ follows from Eqs. (3.3) and (4.2b). At the leading order, this produces the expression (4.9), where the integral has a singular contribution as $\epsilon\to 0$.

The behavior of this eigenfunction is different along different scales of the x variations. In the outer region, i.e., in the limit $x\to\infty$, the function $\Phi_{m+1}(x)$ is localized. Using the boundary conditions (A3) and the asymptotic representation (2.16), we find the limiting behavior,

$$\Phi_{m+1}(x)\big|_{\text{outer}}\to-\frac{\alpha_0(0)}{x}\left[\frac{1}{ip}+\frac{\beta_1^*\ln p}{2\pi}\right]+O(\epsilon), \tag{4.11a}$$

where β_1 is given by Eq. (2.17b). The second term in the brackets is exponentially small compared to the first one, and, therefore, we define $\alpha_0(0)$ as $\alpha_0(0)=-ip$ according to the boundary condition $\Phi_{m+1}(x)\to x^{-1}$ as $x\to\infty$.

In the inner region, where $x\sim O(1)$, the amplitude of the eigenfunction $\Phi_{m+1}(x)$ is exponentially small in terms of ϵ. Using the integral representation (4.9) we find the inner asymptotic expansion for $\Phi_{m+1}(x)$ as follows:

$$\Phi_{m+1}(x)\big|_{\text{inner}}\to ip(\ln p)n(x)+O(p). \tag{4.11b}$$

In order to compute the spectral data γ_{m+1} we use Eqs. (2.9b), (4.9), and (A5) from Appendix A. In addition, we find from Eqs. (2.16a) and (2.17a) that

$$\lim_{k\to 0}\left(\frac{\partial K(k,0)}{\partial k}\right)=-i\int_{-\infty}^{\infty}x\Delta u(x)|n(x)|^2dx-\frac{\beta_1}{2\pi i}K(0,0),$$

where β_1 is given by Eq. (2.17b). The result of these computations is expressed in Eq. (4.10).

C. Revised variations of data of the continuous spectrum

Here we revise the regular perturbation theory for Jost functions. The regular asymptotic expansion is described in Sec. III C, where it was shown that the contradiction (3.20) appears in the limit $k\to 0^+$ for perturbations of nongeneric potentials. In order to calculate this limit correctly, we notice that the integral in Eq. (3.16a) is singular in the limit $\kappa\to 0^+$ if $K(k,0)\neq 0$ and $\alpha(0,\kappa)\neq 0$. Therefore, we cannot use the Taylor expansion (3.17) to solve Eqs. (3.16). Instead, we evaluate the second term of the r.h.s. of Eq. (3.16a) in the form,

$$\int_0^\infty \frac{K(k,k')\alpha(k',\kappa)}{k'-(\kappa-i0)}\,dk' = -\pi i K(k,\kappa)\alpha(\kappa,\kappa) + K(k,\kappa)\alpha(\kappa,\kappa)\ln\left(\frac{1-\kappa}{\kappa}\right) + \text{p.v.}\int_0^1 \frac{K(k,k')\alpha(k',\kappa)-K(k,\kappa)\alpha(\kappa,\kappa)}{k'-\kappa}\,dk' + \int_1^\infty \frac{K(k,k')\alpha(k',\kappa)}{k'-\kappa}\,dk'.$$

Now we impose a scaling transformation,

$$\alpha(k,\kappa) = \frac{1}{\ln\kappa - \ln p + \pi i}\,\beta(k,\kappa), \tag{4.12a}$$

where $\ln p$ is defined by solving Eq. (4.1) and $\beta(k,\kappa)$ satisfies the integral equation,

$$\beta(k,\kappa) = \frac{\epsilon}{2\pi}K(k,\kappa)[-\ln p + (\ln\kappa + \pi i)(1-\beta(\kappa,\kappa))] + \frac{\epsilon}{2\pi}\left[\text{p.v.}\int_0^1 \frac{K(k,k')\beta(k',\kappa)-K(k,\kappa)\beta(\kappa,\kappa)}{k'-\kappa}\,dk' + \int_1^\infty \frac{K(k,k')\beta(k',\kappa)}{k'-\kappa}\,dk' + \sum_{j=1}^m \frac{K_j(k)\beta_j(\kappa)}{k_j-\kappa}\right]. \tag{4.13}$$

Here $\beta_j(\kappa)$ are defined through $\alpha_j(\kappa)$ by the same transformation as Eq. (4.12a). In the asymptotic limit $\kappa \sim p(\epsilon) \ll 1$ we use the approximation $K(k,\kappa) = K(k,0) + O(\kappa)$ and $\beta(\kappa,\kappa) = \beta(0,\kappa) + O(\kappa)$ and compare the integral equation (4.1) with Eq. (4.13). Neglecting the terms of $O(p(\epsilon))$ and $O(\kappa)$ in the integral equations, we derive the following simple result:

$$\beta(k,\kappa) = \frac{\alpha(k)}{\alpha(0)} + O(\kappa, p(\epsilon)), \tag{4.12b}$$

where $\alpha(k)$ is represented by the asymptotic series (4.2b). The remainder term of $O(\kappa)$ follows from the analysis of Eq. (4.13) in the limit $\epsilon \to 0$. Using Eqs. (3.14) and (4.12) we verify that the Jost function $N^\epsilon(x,k)$ has the correct asymptotic behavior as $k \to 0^+$,

$$N^\epsilon(x,k) = \frac{-\epsilon^{-1}p_{-1}}{\ln k - \ln p + \pi i}\,n(x) + O\left(\frac{k}{\ln k}\right). \tag{4.14}$$

This formula agrees with the representation (2.12). Furthermore, comparing the denominators of Eqs. (2.12) and (4.14), we find the relation between the asymptotic expansion of the parameters n_0^ϵ [see Eq. (1.6)] and $p(\epsilon)$ for the new eigenvalue,

$$\frac{1}{n_0^\epsilon} = -\ln p(\epsilon) + \frac{i\pi}{2} - \gamma + O(p(\epsilon)). \tag{4.15}$$

At the leading order, this formula implies that

$$n_0^\epsilon = \frac{\epsilon}{2\pi}K(0,0) + O(\epsilon^2). \tag{4.16}$$

Thus, the criterion for bifurcation of a new eigenvalue ($K(0,0) > 0$) can also be formulated as $n_0^\epsilon > 0$. Furthermore, we use Eqs. (3.15) and (4.12) to find the spectral data $\Gamma^\epsilon(k)$ as $k \sim p(\epsilon) \ll 1$. Since $\Gamma(k) = 1 + O(k^2)$ for nongeneric potentials [see Eq. (2.16b)], we have

Proposition 3: When $k \sim p(\epsilon) \ll 1$, the spectral data $\Gamma^\epsilon(k)$ are given by

$$\Gamma^{\epsilon}(k)=1+\frac{2\pi i}{\ln k-\ln p+\pi i}+O\left(\frac{k}{\ln k},\epsilon\right). \tag{4.17}$$

It is important to notice that the perturbation to the nongeneric potential produces an $O(1)$ effect which occurs however only for $k\sim p(\epsilon)\ll 1$. In the limit $k\ll p(\epsilon)$ the asymptotic solutions (4.14), (4.17) reproduce the behavior of scattering data for the generic potentials [see Eqs. (2.14)]. In the opposite limit $k\gg p(\epsilon)$ the asymptotic solution (4.12) matches with the Taylor expansion (3.17a), while the Jost function (4.14) approaches to the limiting function $n(x)$. Thus, the revised perturbation theory provides a uniform asymptotic solution describing a structural transformation of asymptotic behavior of the Jost function and the scattering coefficients as $k\to 0^+$ induced due to a perturbation of the nongeneric potential.

Finally, we notice that the uniform asymptotic solution (4.12) prescribes a ''hidden'' pole at $k=p(\epsilon)$. This resonant pole occurs due to strong excitation of radiation in the BO equation (1.1) which compensates the unbalance of the area integral (1.2) induced due to formation of a new soliton. Indeed, if the small perturbation to the potential leads to a new soliton (new eigenvalue), then the following relation must be satisfied for Eq. (2.18) at the leading order,

$$\lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_0^{\infty}\frac{|\beta^{\epsilon}(k)|^2-|\beta(k)|^2}{k}\,dk=2\pi. \tag{4.18}$$

We check this limit by using the asymptotic representation for $\beta^{\epsilon}(k)$ following from Eqs. (2.6b), (2.12), and (4.15),

$$\beta^{\epsilon}(k)=\frac{2\pi i}{\ln k-\ln p-\pi i}+O\left(\frac{k}{\ln k},\epsilon\right). \tag{4.19}$$

The dominant pole-type term is the leading-order term in the difference between $\beta^{\epsilon}(k)$ and $\beta(k)$, so that we find

$$\lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_0^{\infty}\frac{|\beta^{\epsilon}(k)|^2-|\beta(k)|^2}{k}\,dk=2\pi\int_0^{\infty}\frac{dk}{k((\ln k-\ln p)^2+\pi^2)}=2\pi.$$

V. EXPLICIT EXAMPLES

In this section we apply our general results to the important class of the pure m-soliton potential $u(x)$. In particular, these explicit formulas solve the problem of soliton generation by a small initial disturbance, i.e., when $u(x)=0$. Explicit expressions are also computed for an algebraic potential $u(x)=2a/(1+x^2)$.

A. Asymptotic approximations for *m*-soliton potentials

For m-soliton potentials, the scattering data are $\beta(k)=0$ and $\Gamma(k)=1$ for all k, and the Jost function $N(x,k)$ has the simple explicit form,

$$N(x,k)=n(x)e^{ikx},\quad n(x)=\exp\left[-i\int_x^{\infty}u_0(x)dx\right]. \tag{5.1}$$

Thus, the asymptotic expression (2.19) for large k is valid for any k in the case of m-soliton potentials and provides a function which is analytic in $\mathrm{Im}(x)\geqslant 0$. To prove the latter statement, we rewrite the m-soliton solution through its pole expansions,[2]

$$u(x)=\sum_{j=1}^{m}\left[\frac{i}{x-x_j}-\frac{i}{x-x_j^*}\right], \tag{5.2}$$

where x_j are complex poles of the solutions with $\mathrm{Im}(x_j)<0$. Then, we find the explicit formula for $n(x)$,

$$n(x)=\prod_{j=1}^{m}\frac{x-x_j^*}{x-x_j} \tag{5.3}$$

which is analytical for $\mathrm{Im}(x)\geqslant 0$. It follows from Eqs. (1.6) and (5.1) that the constraint $n_0=0$ is satisfied identically for any m-soliton solution. As a result, we find from Eq. (4.5) that the variation of m-soliton solutions may generate an additional eigenvalue, if

$$\Delta\mathcal{M}=\int_{-\infty}^{\infty}\Delta u(x)dx>0, \tag{5.4}$$

where $\Delta\mathcal{M}$ is the variation of the area integral (1.2). [Note that the area integral for the m-soliton solution is $\mathcal{M}=2\pi m$.] The asymptotic approximation of the eigenvalue $k=-p(\epsilon)$ follows from Eqs. (4.4), (4.7), and (4.8). Using the fact that $|n(x)|^2=1$ [see Eq. (5.1)] and the table integral (see p. 432 in Ref. 24),

$$\int_0^1\frac{e^{ik(x-y)}-1}{k}\,dk+\int_1^{\infty}\frac{e^{ik(x-y)}}{k}\,dk=-\gamma-\ln|x-y|+\frac{i\pi}{2}\,\mathrm{sign}(x-y) \tag{5.5}$$

we find the parameters p_{-1} and p_0 explicitly

$$p_{-1}=-\frac{2\pi}{\Delta\mathcal{M}} \tag{5.6a}$$

and

$$p_0=-\gamma-\frac{\int\int_{-\infty}^{\infty}\Delta u(x)\Delta u(y)\ln|x-y|dxdy}{(\Delta\mathcal{M})^2}-\sum_{j=1}^{m}\frac{\left|\int_{-\infty}^{\infty}\Delta u(x)n^*(x)\Phi_j(x)dx\right|^2}{k_j^2(\Delta\mathcal{M})^2}. \tag{5.6b}$$

The first integral term in Eq. (5.6b) represents the second-order correction to the new eigenvalue from the continuous spectrum while the second integral term is the correction induced due to the presence of other bound states. We notice that the latter contribution is additive which is a consequence of the Darboux transformation of linear eigenvalue problems with bound states.[25] In the particular case $m=0$, when all other bound states are absent, the expression (5.6) give an approximation for a single eigenvalue supported by a small perturbation to the zero background $u(x)=0$ with a positive area integral (1.2).

B. Algebraic potential

The single-humped algebraic potential $u(x)=2a/(1+x^2)$ was considered as an initial condition for numerical simulations of the BO equation.[8,9] In the case $a=1$, this potential is exactly the soliton solution (1.4) for $t=0$, $x_0=0$, and $v=1$ [Due to the scaling invariance of the BO equation the parameter v can be always scaled by unity.] Kodama *et al.*[10] proved that this potential generates an m-soliton solution for $a=a_m=m$ with eigenvalues k_j given by zeros of the Laguerre polynomials $L_m(-2k)$ (see Appendix B). They also proved that, if $a_m<a\leqslant a_{m+1}$, there are $m+1$ bound states. Since the exact solutions of Eq. (1.5) for the algebraic potential are not known, the eigenvalues were calculated numerically. Here we find the explicit analytical approximations of the new small eigenvalue which emerges for $a=a_m+\epsilon$, where $0<\epsilon\ll 1$.

Since the function $n(x)$ is analytical for $\mathrm{Im}(x)\geqslant 0$ we calculate $P^+[un]$ for the algebraic potential $u=2a/(1+x^2)$ as follows:

$$P^+[un]=un+\frac{ian(i)}{x-i}.$$

In addition, we find from Eq. (1.6) that

$$n_0=\frac{1}{2\pi}\int_{-\infty}^{\infty}u(x)n(x)dx=an(i).$$

Then, the limiting function $n(x)$ satisfies a simple differential equation following from Eq. (2.13a),

$$in_x+un=-\frac{in_0}{x-i}.$$

This equation has an explicit solution

$$n(x)=-n_0\left(\frac{x-i}{x+i}\right)^a(I_{a+1}(x)-\alpha), \tag{5.7}$$

where

$$I_{a+1}(x)=\int\frac{(x+i)^a}{(x-i)^{a+1}}\,dx$$

and the constant α is introduced to cancel the nonanalyticity of $n(x)$ at the branching point $x=i$. Using the recurrent relation,

$$I_{a+1}(x)=I_a(x)-\frac{1}{a}\left(\frac{x+i}{x-i}\right)^a \tag{5.8}$$

the function $n(x)$ can be represented for $m\leqslant a<(m+1)$ as

$$n(x)=n_0\left[\sum_{k=0}^{m}\frac{1}{a-k}\left(\frac{x-i}{x+i}\right)^k-\left(\frac{x-i}{x+i}\right)^a(I_{a-m}(x)-I_{a-m}(i))\right]. \tag{5.9}$$

The integral $I_a(x)$ for $0\leqslant a<1$ can be evaluated in two exceptional cases: for $a=1/2$ (see Ref. 16), when $I_{1/2}=\ln(x+\sqrt{1+x^2})$, and for $a=0$, when $I_0=\ln(x+i)$. Using the last formula and comparing the asymptotical behavior of $n(x)$ as $x\to\infty$ given by Eqs. (2.15) and (5.9), we find the asymptotic representation for n_0 as $a=m+\epsilon$,

$$\frac{1}{n_0}=\frac{1}{\epsilon}+\ln 2+\frac{i\pi}{2}+\sum_{k=1}^{m}\frac{1}{k}+O(\epsilon). \tag{5.10}$$

This exact computation of n_0 enables us to find from Eq. (4.15) the asymptotic approximation for a new eigenvalue $k_{m+1}=-p(\epsilon)$ supported by a perturbed algebraic potential,

$$p(\epsilon)=\tfrac{1}{2}\exp(-\epsilon^{-1}-\gamma-q_m+O(\epsilon))\quad q_m=\sum_{k=1}^{m}\frac{1}{k}. \tag{5.11}$$

In Appendix B we derive the formula (5.11) with the help of Eqs. (5.6). This (equivalent) technique leads to a new relation for zeros of Laguerre polynomials.

VI. CONCLUSION

We have shown that small variations of the nongeneric potentials support a new bound state in the BO scattering problem if the generic perturbation leads to a positive value of the governing parameter n_0. For the soliton potentials, the latter condition is equivalent to a positive contribution to the area integral (1.2) induced by the perturbation. The self-consistent asymptotic expansions are constructed and studied not only for the new eigenvalue and an associated bound state but also for the perturbed Jost functions of the BO scattering problem.

Our results on bifurcation of new eigenvalues in the spectral problem (1.5) solve the classical problem of soliton generation by a small initial perturbation within the BO equation. The expo-

nential behavior of the new eigenvalue explains why our results do not fit into the classification of Ref. 11. Although the BO equation (1.1) realizes a marginal case in the criterion on the soliton generation given in Ref. 11, the generation still takes place but is accompanied by a strong radiation. The energy of an initial pulse measured by the integral $P=\int_{-\infty}^{\infty}u^2dx$ all goes to the linear dispersive waves except for an exponentially small addition to the algebraic soliton (1.4). As an indication of this smallness, the new soliton was not identified in the numerical simulations of Ref. 8.

It is interesting to compare the no-threshold soliton generation in the BO equation (1.1) with that in the KdV equation, $u_t+6uu_x+u_{xxx}=0$. In the KdV equation,[4] a small initial pulse with the mass invariant $M=\epsilon\Delta M$, where $\Delta M=\int_{-\infty}^{\infty}u(x,0)dx$, also leads to generation of a single soliton with the mass $M_s=2\epsilon\Delta M$ and emission of linear radiative waves with the mass $M_{\rm rad}=M-M_{\rm sol}=-\epsilon\Delta M$, i.e., all integral quantities remain small in this evolution process. In the BO equation, an initial pulse with the mass $M=\epsilon\Delta M$ generates a soliton with finite mass $M_{\rm sol}=2\pi$ and strong radiation with finite mass $M_{\rm rad}=M-M_{\rm sol}=-2\pi+\epsilon\Delta M$ [see formulas (2.18) and (4.18)].

Similar bifurcations of new eigenvalues (solitons) may also occur when the perturbation of the potential is given not by an initial condition but by a perturbation to the underlying equation itself. The perturbed BO equation was considered recently by Matsuno and Kaup[26,27] who studied variations of the soliton solutions induced by the perturbation. The formation of new solitons has not been discussed yet.

Finally, since the BO equation is believed to be a pivot to multidimensional evolution equations such as the KP1 and DS2 equations (see Ref. 3), another interesting problem is to construct the bifurcation theory for new eigenvalues (solitons) in the multidimensional evolution equation integrable by the inverse scattering method.

ACKNOWLEDGMENTS

D.P. thanks Yu. A. Stepanyants for bringing his attention to the problem and comments on the manuscript. We have also benefited from useful discussions with D. Kaup, Y. Kodama, T. Lakoba, E. Pelinovsky, and I.M. Sigal. D.P. acknowledges support from a NATO fellowship provided by NSERC and C.S. acknowledges support from NSERC Operating Grant No. OGP0046179.

APPENDIX A: USEFUL RELATIONS FOR PERTURBATION THEORY

Taking the limits (2.8b) into Eq. (2.7) we derive a system of algebraic equations for $\Phi_j(x)$, $j=\overline{1,m}$,

$$(x+\gamma_j)\Phi_j(x)=1-i\sum_{l\neq j}\frac{\Phi_l(x)}{k_j-k_l}+\frac{1}{2\pi i}\int_0^{\infty}\frac{\beta(k)N(x,k)}{k-k_j}\,dk. \tag{A1}$$

Furthermore, using Eqs. (2.3b), (2.3c), (2.6b), (2.9a) as well as the formula

$$\lim_{x\to+\infty}\ \text{p.v.}\ \frac{e^{\pm ikx}}{k}=\pm i\pi\delta(k),$$

we find the following relations:

$$\int_{-\infty}^{\infty}\Phi_j dx=-\pi i, \tag{A2a}$$

$$\int_{-\infty}^{\infty}N(x,k)dx=2\pi\delta(k)-\frac{i\beta^*(k)}{2k}. \tag{A2b}$$

Equivalently, these formulas also follow from the boundary conditions for $N(x,k)$ and $\Phi_j(x)$ as $x\to\infty$ [cf. Eq. (2.1)],

$$N(x,k) \to e^{ikx} + \frac{\beta^*(k)}{2\pi kx} + O(x^{-2}), \tag{A3a}$$

$$\Phi_j(x) \to \frac{1}{x} + O(x^{-2}). \tag{A3b}$$

We derive from Eqs. (A1), (A2), and orthogonality conditions (3.1) the following important relations:

$$\int_{-\infty}^{\infty} (x+\gamma_j) |\Phi_j(x)|^2 dx = \pi i, \tag{A4a}$$

$$\int_{-\infty}^{\infty} (x+\gamma_j) \Phi_l^*(x) \Phi_j(x) dx = \frac{\pi i (k_j + k_l)}{k_j - k_l}, \quad j \neq l, \tag{A4b}$$

$$\int_{-\infty}^{\infty} (x+\gamma_j) N^*(x,k) \Phi_j(x) dx = 2\pi \delta(k) + \frac{i\beta(k)(k_j+k)}{2k(k_j-k)}. \tag{A4c}$$

In addition, we use Eqs. (2.7), (2.8a), (3.1a), and (A2b) in order to compute an equivalent formula for the Jost functions $N(x,k)$. As a result of straightforward calculations, we derive the relation,

$$\int_{-\infty}^{\infty} x N^*(x,k') N(x,k) dx = -2\pi i \frac{d\delta}{dk}(k-k') + \frac{\beta^*(k)}{2k} \delta(k') + \frac{\beta^*(k')}{2k'} \delta(k) + \frac{i\beta(k')\beta^*(k)(k+k')}{4\pi k k'(k-k')}. \tag{A5}$$

APPENDIX B: DERIVATION OF THE ASYMPTOTIC FORMULA (5.11)

Here we check the asymptotic result (5.11) by virtue of formulas (5.6). First, the potential $u(x) = 2a/(1+x^2)$ can be decomposed for $a = m + \epsilon$ as

$$u(x) = \frac{2m}{1+x^2} \quad \text{and} \quad \Delta u(x) = \frac{2}{1+x^2}. \tag{B1}$$

Then, we find from Eqs. (5.4) and (5.6a) that $\Delta M = 2\pi$ and $p_{-1} = -1$ in agreement with Eq. (5.11). Then, the first integral term in Eq. (5.6b) can be evaluated explicitly[24] as

$$\int \int_{-\infty}^{\infty} \Delta u(x) \Delta u(y) \ln|x-y| dx dy = -8 \int_0^{\pi} t \ln(\sin t) dt = 4\pi^2 \ln 2 \tag{B2}$$

which gives the factor $\frac{1}{2}$ in Eq. (5.11).

Next, in order to evaluate the contribution from the other bound states, we find from Eq. (5.3) the limiting eigenfunction $n(x)$ for the algebraic potential $u(x)$ given by Eq. (B1),

$$n(x) = \frac{(x-i)^m}{(x+i)^m}. \tag{B3}$$

Using Eq. (2.9a), we have

$$k_j = \frac{1}{2\pi i} \int_{-\infty}^{\infty} u(x) \Phi_j(x) dx = -im \Phi_j(i).$$

Similarly to Eq. (5.7), the bound states $\Phi_j(x)$ can also be expressed through an integral representation (see Eq. (7.23) in Ref. 10),

$$\Phi_j(x) = -ik_j\left(\frac{x-i}{x+i}\right)^m e^{ik_j x}\int_{-\infty}^{x}\frac{(y+i)^m}{(y-i)^{m+1}}\,e^{-ik_j y}dy.$$

This integral can be computed through the finite sum,

$$\Phi_j(x) = ik_j\sum_{k=1}^{m}\sum_{l=1}^{k}\frac{(l-1)!m!}{(k!)^2(m-k)!}\,\frac{(2i)^l(2k_j)^{k-l}}{(x+i)^m(x-i)^{l-m}} - ik_j\left(\frac{x-i}{x+i}\right)^m e^{ik_j x}\int_{-\infty}^{x}\frac{e^{ik_j y}}{y-i}\,dy\;L_m(-2k_j), \tag{B4}$$

where $L_m(x)$ is the Laguerre polynomial[24]

$$L_m(x) = \sum_{k=0}^{m}\frac{m!}{(k!)^2(m-k)!}\,(-x)^k. \tag{B5}$$

The eigenvalues k_j satisfy $L_m(-2k_j)=0$. Therefore, the integrals $\int_{-\infty}^{\infty}\Delta u(x)n^*(x)\Phi_j(x)dx$ in Eq. (5.6b) can be evaluated from (B1), (B3), and (B4). As a result, we find from Eqs. (5.6) and (5.11) the following representation for q_m:

$$q_m = \sum_{j=1}^{m}\left(\sum_{k=1}^{m}\sum_{l=1}^{k}\frac{(-1)^l(l-1)!m!}{(k!)^2(m-k)!}\,(2k_j)^{k-l}\right)^2. \tag{B6}$$

According to Eq. (5.11), the sum q_m is equal to $q_m=\Sigma_{k=1}^{m}(1/k)$ which gives a new relation for zeros of the Laguerre polynomials. Indeed, we have checked the first terms following from Eqs. (B5) and (B6): $q_1=1$, $q_2=3/2$, $q_3=11/6$. These results agree with Eq. (5.11).

[1] R. Grimshaw, ''Internal solitary waves,'' in *Advances in Coastal and Ocean Engineering*, edited by P. L.-F. Liu (World Scientific, Singapore, 1996), Vol. 3, pp. 1–30.
[2] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[3] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991).
[4] V. I. Karpman, *Nonlinear Waves in Dispersive Media* (Pergamon, New York, 1975).
[5] M. Desaix, D. Anderson, M. Lisak, and M. L. Quiroga-Teixeiro, ''Variationally obtained approximate eigenvalues of the Zakharov–Shabat scattering problem for real potentials,'' Phys. Lett. A **212**, 332–338 (1996).
[6] P. D. Lax and C. D. Levermore, ''The small dispersion limit of the Korteweg–de Vries equation I, II, III,'' Commun. Pure Appl. Math. **37**, 253–290 (1983); **37**, 571–593 (1983); **37**, 809–830 (1983).
[7] Y. Matsuno, ''Asymptotic properties of the Benjamin–Ono equation,'' J. Phys. Soc. Jpn. **51**, 667–674 (1982).
[8] T. Miloh, M. Prestin, L. Shtilman, and M. P. Tulin, ''A note on the numerical and *N*-soliton solutions of the Benjamin–Ono evolution equation,'' Wave Motion **17**, 1–10 (1993).
[9] J. D. Meiss and N. R. Pereira, ''Internal solitons,'' Phys. Fluids **21**, 700–702 (1978).
[10] Y. Kodama, M. J. Ablowitz, and J. Satsuma, ''Direct and inverse scattering problems of the nonlinear intermediate long wave equation,'' J. Math. Phys. **23**, 564–576 (1982).
[11] E. N. Pelinovsky and Yu. A. Stepanyants, ''Linear approximation in pulse propagation problems in nonlinear media,'' Radiophysics and Quantum Electronics **21**, 1186–1188 (1979).
[12] A. S. Fokas and M. J. Ablowitz, ''The inverse scattering transform for the Benjamin–Ono equation: A pivot to multidimensional problems,'' Stud. Appl. Math. **68**, 1–10 (1983).
[13] P. M. Santini, M. J. Ablowitz, and A. S. Fokas, ''On the limit from the intermediate long wave equation to the Benjamin–Ono equation,'' J. Math. Phys. **25**, 892–899 (1984).
[14] R. L. Anderson and E. Taflin, ''The Benjamin–Ono equation: Recursivity of linearization maps: Lax pairs,'' Lett. Math. Phys. **9**, 299–311 (1985).
[15] R. R. Coifman and M. V. Wickerhauser, ''The scattering transform for the Benjamin–Ono equation,'' Inverse Probl. **6**, 825–861 (1990).
[16] G. Scoufis and C. M. Cosgrove, ''On the initial value problem for the modified Benjamin–Ono equation,'' J. Math. Phys. **36**, 5753–5759 (1995).
[17] D. J. Kaup and Y. Matsuno, ''On the inverse scattering transform for the Benjamin–Ono equation,'' Stud. Appl. Math., to appear.
[18] D. J. Kaup, T. I. Lakoba, and Y. Matsuno, ''Complete integrability of the Benjamin–Ono equation by means of action-angle variables,'' Phys. Lett. A **238**, 123–133 (1998).
[19] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, New York, 1985).
[20] M. Klaus and B. Simon, ''Coupling constant thresholds in nonrelativistic quantum mechanics. I. Short-range two-body case,'' Ann. Phys. (N.Y.) **130**, 251–281 (1980).
[21] S. M. Apenko, ''Weakly bound states in $2+\epsilon$ dimensions,'' J. Phys. A **31**, 1553–1562 (1998).
[22] R. L. Pego and M. I. Weinstein, ''Eigenvalues and instabilities of solitary waves,'' Philos. Trans. R. Soc. London, Ser. A **340**, 47–94 (1992).

[23] M. J. Ablowitz and J. Villarroel, ''Solutions to the time-dependent Schrödinger and the Kadomtsev–Petviashvili equations,'' Phys. Rev. Lett. **78**, 570–573 (1997).
[24] I. S. Gradshtein and I. M. Ryzhik, *Tables of Integrals, Sums, Series and Products* (Academic, New York, 1980).
[25] V. B. Matveev and M. A. Salle, *Darboux Transformation and Solitons* (Springer, Berlin, 1992).
[26] Y. Matsuno, ''Multisoliton perturbation theory for the Benjamin–Ono equation and its application to real physical systems,'' Phys. Rev. E **51**, 1471–1483 (1995).
[27] Y. Matsuno and D. J. Kaup, ''Initial value problem of the linearized Benjamin–Ono equation and its applications,'' J. Math. Phys. **38**, 5198–5224 (1997).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

Timelike infinity and asymptotic symmetry

Uchida Gen
Department of Physics, The University of Tokyo, Tokyo 113, Japan

Tetsuya Shiromizu
Department of Physics, The University of Tokyo, Tokyo 113, Japan
and Research Center for the Early Universe (RESCEU), The University of Tokyo, Tokyo 113, Japan

(Received 9 September 1997; accepted for publication 20 August 1998)

By extending Ashtekar and Romano's definition of spacelike infinity to the timelike direction, a new definition of asymptotic flatness at timelike infinity for an isolated system with a source is proposed. The treatment provides unit spacelike three-hyperboloid timelike infinity and avoids the introduction of the troublesome differentiability conditions which were necessary in the previous works on asymptotically flat space–times at timelike infinity. Asymptotic flatness is characterized by the falloff rate of the energy-momentum tensor at timelike infinity, which makes it easier to understand physically what space–times are investigated. The notion of the order of the asymptotic flatness is naturally introduced from the rate. The definition gives a systematized picture of hierarchy in the asymptotic structure, which was not clear in the previous works. It is found that if the energy-momentum tensor falls off at a rate faster than $\sim t^{-2}$, the space–time is asymptotically flat and asymptotically stationary in the sense that the Lie derivative of the metric with respect to ∂_t falls off at the rate $\sim t^{-2}$. It also admits an asymptotic symmetry group similar to the Poincaré group. If the energy-momentum tensor falls off at a rate faster than $\sim t^{-3}$, the four-momentum of a space–time may be defined. On the other hand, angular momentum is defined only for space–times in which the energy-momentum tensor falls off at a rate faster than $\sim t^{-4}$. © *1998 American Institute of Physics.* [S0022-2488(98)01812-X]

I. INTRODUCTION

If gravitational collapse takes place in an asymptotically flat space–time, we naively expect that the energy of fields and particles in the exterior region would be either radiated away to infinity or fall into the black hole and that the space–time would approach a vacuum space–time. Because the lost energy cannot be restored, the dynamical processes would eventually die away and the space–time would settle down to a stationary state. Thus, we expect that the final state of gravitational collapse be described by a Kerr black hole, since the uniqueness theorem states that the Kerr space–time is the unique asymptotically flat, stationary, vacuum solution.[1] This is the reason why the Kerr solution is used to analyze the black hole candidates observed in the universe. In this paper, to confirm the naive picture above, we shall closely examine the asymptotic structure of the very late time regime of a space–time, i.e., timelike infinity, and see that an ''asymptotic'' vacuum state actually induces an ''asymptotic'' stationary state. This approach may be more reasonable than the assumption of the exact stationarity in the sense that *exact* vacuum or *exact* stationary state would not be achieved in a real gravitational collapse.

A solution of the Einstein equation is normally given in terms of the metric components with respect to some specific coordinates. However, it is not easy from the component expression to grasp what physical system the solution actually represents. One way to overcome the difficulty is to introduce a certain physical index that characterizes the system. For example, if the energy of the system at a given time is defined, we can picture the evolution of the system to some extent from the change in its value. Moreover, if the multipole moments of a space–time are defined, we may obtain a precise Newtonian analog by constructing a Newtonian potential having the same moments. The analog could serve as a powerful guide for the physical interpretation of the solution. Because the gravitational field itself acts as its own source, we expect multipole moments

 © 1998 American Institute of Physics

can be defined only in space–times with gravitational fields satisfying a suitable fall-off condition. It is shown that *stationary* space–times are sufficiently asymptotically flat for the multipole moments to be well defined.[2] However, there are no arguments on more general space–times. In this paper, as a first step towards defining multipole moments of general asymptotically flat space–times, we explore timelike infinity of a space–time to seek what falloff condition the gravitational field must satisfy so that four-momentum or angular-momentum may be defined.

There are some previous works on timelike infinity. A definition of timelike infinity with *everywhere-smooth* unphysical metric is proposed by Cutler in Ref. 3. The high differentiability condition results in demanding that the space–time have vanishing final Bondi mass at timelike infinity. This means the definition applies only for space–times with no bounded sources such as black holes and stars. On the other hand, Porill's definition of the asymptotic flatness at timelike infinity is applicable to space–times with bounded sources.[4] However, as a result of the ''tieing'' of the late time regime with the null regime, the treatment compresses the late time regime, which has a finite volume, down to a point. As a result, we cannot demand smoothness or even differentiability of the unphysical metric at the point. Instead, we need to formulate a notion of direction-dependent ''differentiability'' and demand that the metric possess this awkward property. The treatment makes it difficult to understand what kinds of space–times are investigated.

Recently, Ashtekar and Romano formulated an elegant treatment of spacelike infinity in a coordinate-independent manner.[5] In their work, the corresponding ''awkward'' differentiability at spacelike infinity is avoided by regarding the late time regime of a space–time not as a point but as a three-manifold. This is achieved by abandoning the idea of ''tieing'' the late time regime and the null regime together. As a result, there is no necessity to use a conformally completed unphysical metric for the description of the gravitational field. Instead, a normal vector field and an intrinsic metric on the timelike slices foliating the unphysical space–time are defined, and they provide the three-manifold timelike infinity.

In the present paper, we modify Ashtekar and Romano's treatment of spacelike infinity to the timelike direction and investigate the asymptotic structure of space–times around timelike infinity. Our investigation is more or less a direct application of their work but differs in a number of important points, which enable us to discuss (1) the asymptotic vacuum state and the asymptotic stationary state of a space–time to explore the final product of gravitational collapse, and (2) the order of asymptotic flatness to explore how flat the space–time must be in order to define physical indices. First, we take a different view on how the falloff condition on the energy-momentum tensor should be imposed. We impose conditions on the *physical* components of the physical energy-momentum tensor, while they impose conditions on the *unphysical* components. This change gives a picture of hierarchy in the asymptotic structure at timelike infinity and helps us to understand how the structure is systematized, which is a point that was obscure in the previous treatments. Second, a notion of asymptotic Killing fields is introduced. The formulation of the order of asymptotic symmetries is possible only by the introduction of such a notion. This is also crucial in examining and specifying how fast a space–time becomes stationary. Also, special attention is paid to a bounded source that persists even in the very late time regime of a space–time, in contrast with the spacelike direction.

The rest of the paper is organized as follows. In Sec. II, we construct timelike infinity with certain desirable features for Minkowski space–time. Following the completion of Minkowski space–time, we propose a definition of asymptotic flatness at future timelike infinity for a general space–time with a source in Sec. III. In Sec. IV, we examine the fundamental asymptotic structure of space–times that possess the defined asymptotic flatness. In Sec. V, we propose an asymptotic symmetry group which preserves the studied structure. A notion of asymptotic Killing fields is introduced in Sec. VI, and its relation to the falloff of the energy-momentum tensor is clarified. In Sec. VII, we examine the asymptotic behavior of the gravitational fields. In Sec. VIII, we define conserved quantities associated with the asymptotic symmetry. The Appendix gives some examples of asymptotically flat space–times with timelike infinity defined in this paper and the flat Friedmann space–time which is not asymptotically flat as a counterexample. Throughout the paper, we follow the notation of Wald.[6]

II. COMPLETION OF THE MINKOWSKI SPACE–TIME

In this section, we complete Minkowski space–time $(\hat{\mathcal{M}}, \hat{\eta}_{ab})$ in such a way that future timelike infinity is represented not as a point but as a three-manifold. The construction motivates the general definition of asymptotic flatness in the next section. Hereafter, we only consider *future* timelike infinity and thus omit the word future. All the discussion presented here is also applicable to past timelike infinity with minor changes.

In static spherical coordinates (t,r,θ,ϕ), the metric of Minkowski space–time takes the form

$$\hat{\eta}_{ab} = -(dt)_a(dt)_b + (dr)_a(dr)_b + r^2(d\sigma)_{ab}, \tag{2.1}$$

where $(d\sigma)_{ab} := (d\theta)_a(d\theta)_b + \sin^2\theta(d\phi)_a(d\phi)_b$ is the unit two-sphere metric. Introduce the standard hyperbolic coordinates (ρ,χ,θ,ϕ) in the *interior* of the future light cone at the origin: $t = \rho\cosh\chi$ and $r = \rho\sinh\chi$, where $\chi\in[0,\infty)$ and $\rho\in(0,\infty)$. In this chart, timelike infinity is specified by the limit surface $\rho\to\infty$ and the metric takes the form

$$\hat{\eta}_{ab} = -(d\rho)_a(d\rho)_b + \rho^2 h_{ab}, \tag{2.2}$$

where $h_{ab} := (d\chi)_a(d\chi)_b + \sinh^2\chi(d\sigma)_{ab}$ is the unit spacelike three-hyperboloid metric. If a new coordinate $\Omega := \rho^{-1}$ is introduced and $\hat{\mathcal{M}}$ is extended to include the points on which $\Omega = 0$, timelike infinity is specified by the surface $\Omega = 0$ and the metric is found to be singular there:

$$\hat{\eta}_{ab} = -\Omega^{-4}(d\Omega)_a(d\Omega)_b + \Omega^{-2}h_{ab}. \tag{2.3}$$

To complete the space–time conformally, we take a conformal factor of Ω^4 and obtain a nonsingular unphysical metric:

$$\eta_{ab} := \Omega^4\hat{\eta}_{ab} = -(d\Omega)_a(d\Omega)_b + \Omega^2 h_{ab}. \tag{2.4}$$

With respect to the metric, timelike infinity $\Omega = 0$ is a point with no volume. This is because the components of the metric normal to and those tangent to the Ω-const surfaces are rescaled by the same power of Ω although they diverge with different powers. In other words, if we rescale them according to the power of their divergence, we obtain timelike infinity that is not compressed to a point. Thus, define

$$n^a := \Omega^{-4}\hat{\eta}^{ab}(d\Omega)_b = -(\partial_\Omega)^a \tag{2.5}$$

to the direction normal to the Ω-const surfaces and

$$q_{ab} := \Omega^2(\hat{\eta}_{ab} + \Omega^{-4}F^{-1}(d\Omega)_a(d\Omega)_b) = h_{ab}, \quad \text{where } F := -\pounds_n\Omega = 1 \tag{2.6}$$

to the direction tangent to the Ω-const surfaces. The information on the gravitational fields is completely given by the pair (n^a, q_{ab}), and we do not need the conformally rescaled unphysical metric η_{ab}. With respect to this pair, timelike infinity is a three-surface whose intrinsic metric is h_{ab} and whose normal is ∂_Ω. That is, timelike infinity is a unit spacelike three-hyperboloid.

Note here that the completion presented here does not conserve the causal structure of the physical space–time, in contrast with the conformal completion. Hence, it is inappropriate, for example, to discuss the relation of timelike infinity with null infinity using the method. However, the method provides an adequate basis to investigate asymptotic structure of a spacetime at timelike infinity.

III. DEFINITIONS

In this section, we introduce a definition of asymptotic flatness at timelike infinity to order n and define some tensor fields that are useful for the coming discussion.

We have demonstrated the completion of Minkowski space–time in the previous section. It is important to note that the way which the completion is accomplished crucially depends on the structure of Minkowski space–time at timelike infinity. This motivates us to define an asymptotic flat space–time as a space–time which can be completed ''asymptotically'' in a similar fashion to

Minkowski space–time. More precisely, a space–time is said to be asymptotically flat if the space–time can be extended to enclose the points of infinity and if it is possible to construct the pair (n^a, q_{ab}) that has a smooth limit to the infinity.

Definition: A physical space–time $(\hat{\mathcal{M}}, \hat{g}_{ab})$ is said to possess an *asymptote at future timelike infinity* $\check{\imath}^+$ *to order n* (AFTI-*n*) for a non-negative integer *n*, if there exists a manifold $\mathcal{M}$ with boundary $\mathcal{H}$, a smooth function Ω defined on $\mathcal{M}$, and an imbedding Ψ of an open subset $\hat{\mathcal{F}}$ in $\hat{\mathcal{M}}$ to $\mathcal{M}-\mathcal{H}$ satisfying the following conditions:

(1) $\check{\imath}^+ := \partial\mathcal{F} \cap (\mathcal{M} - \Psi(\hat{\mathcal{M}}))$ is not empty and $\check{\imath}^+ \in I^+(\mathcal{F})$ where $\mathcal{F} := \Psi(\hat{\mathcal{F}})$;
(2) $\Omega \doteq 0$ and $\nabla_a \Omega \neq 0$, where $\doteq$ denotes the equality evaluated on $\check{\imath}^+$;
(3) $n^a := \Omega^{-4} \Psi^* \hat{g}^{ab} \nabla_b \Omega$ and $q_{ab} := \Omega^2 (\Psi^* \hat{g}_{ab} + \Omega^{-4} F^{-1} \nabla_a \Omega \nabla_b \Omega)$ admit smooth limits to $\check{\imath}^+$ with q_{ab} having signature$(+++)$ on $\check{\imath}^+$, where $F := -\pounds_n \Omega$; and
(4) $\lim_{\to \check{\imath}^+} \Omega^{-(2+n)} \hat{T}_{\hat{\mu}\hat{\nu}} = 0$ where $\hat{T}_{\hat{\mu}\hat{\nu}} := \Psi^*[(\hat{e}_\mu)^a (\hat{e}_\nu)^b \hat{T}_{ab}]$ in which $\hat{T}_{ab}$ is the physical energy-momentum tensor and $\{(\hat{e}_\mu)^a\}_{\mu=0,1,2,3}$ is a tetrad of $(\hat{\mathcal{M}}, \hat{g}_{ab})$.

Let us now discuss the meaning of the conditions. First of all, note that the asymptotic flatness is defined for an open *subset* $\hat{\mathcal{F}}$ of the space–time manifold. This is because isolated bodies may be present at the late time regime of a space–time and thus the *whole* space–time is not expected to be asymptotically flat. If there are isolated bodies at the late time regime, the region excluding the bodies should be chosen as $\hat{\mathcal{F}}$. The former part of the first condition ensures that the infinity $\check{\imath}^+$ consists of points that are included by extending the physical manifold. Such $\check{\imath}^+$ resides not in the region $\Psi(\mathcal{M})$, where the physical manifold is mapped, but on the boundary of the region of interest, $\partial\mathcal{F}$. The latter part assures that $\check{\imath}^+$ exists in the *future* of the physical space–time so that $\check{\imath}^+$ represents *future* timelike infinity. The second condition assures that $\check{\imath}^+$ is infinitely far away. The third demands that the asymptotic space–time can be completed asymptotically in a similar fashion to the case of Minkowski space–time and that the infinity $\check{\imath}^+$ is *timelike*. This means that Ω asymptotically plays a role of t^{-1} in Minkowski space–time. The fourth condition specifies how the physical components of the physical energy-momentum tensor should decay. This condition is different from the corresponding one in the definition given by Ashtekar and Romano.[5] If we literally translated their prescription for spacelike infinity to our timelike infinity, we would demand that the energy-momentum tensor fall off faster than Ω^1, $\lim_{\to \check{\imath}^+} \Omega^{-1} \Psi^* \hat{T}_{ab} = 0$, and consider such terms as $q_a^m q_b^n \Omega^{-1} \Psi^* \hat{R}_{mn}$ vanish on $\check{\imath}^+$. Although it is not explicitly stated in Ref. 5, this means that the energy-momentum tensor is evaluated by the components with respect to an *unphysical* coordinate basis, $(e_\mu)^a (e_\nu)^b \Psi^* \hat{T}_{ab}$ where $\{(e_\mu)^a\}_{\mu=0,1,2,3}$ is a tetrad of the unphysical space–time. Here, we take an alternative view. We evaluate the *physical* energy-momentum tensor by its components with respect to a *physical* coordinate basis, $\hat{T}_{\hat{\mu}\hat{\nu}} = \Psi^*[(\hat{e}_\mu)^a (\hat{e}_\nu)^b \hat{T}_{ab}]$. In this view, it is found that the richness of asymptotic structures that a space–time possesses depends on how fast the energy-momentum tensor falls off in the space–time. Thus, we need to specify the rate of the falloff when defining asymptotic flatness. To evaluate $\hat{T}_{\hat{\mu}\hat{\nu}}$ hereafter, we use a tetrad consisting of a unit normal vector to Ω-const surfaces, $\hat{n}^a$, and the triad on Ω-const surfaces, $\{(\hat{e}_I)^a\}_{I=1,2,3}$. They satisfy

$$\hat{g}^{ab} = -\hat{n}^a \hat{n}^b + (\hat{e}_I)^a (\hat{e}_I)^b, \tag{3.1}$$

$$\Psi^* \hat{n}^a = \frac{n^a}{\sqrt{-\Psi^* \hat{g}_{ab} n^a n^b}} = \Omega^2 F^{-1/2} n^a, \tag{3.2}$$

$$\Psi^* (\hat{e}_I)^a = \Omega (e_I)^a, \tag{3.3}$$

where $\{(e_I)^a\}_{I=1,2,3}$ is the smooth triad of q_{ab}. Equation (3.3) will be obtained below. Since all the equations appearing in the following discussion are those on $\mathcal{M}$, unless it may cause ambiguity, we omit hereafter Ψ^* in front of the tensors defined on $\hat{\mathcal{M}}$ for brevity.

It is meaningful to note here that a scalar field ϕ in the Schwarzschild space–time decays asymptotically at the rate $\sim t^{-2}$ or faster, according to the Price's theorem.[7] Thus, the Schwarzschild space–time with such a scalar field is asymptotically flat to order 1.

Now consider the Schwarzschild space–time. We can take the region $r>2m$ for $\mathcal{F}$. Then, all the conditions above are satisfied and $\check{i}^+$ is represented by a unit spacelike three-hyperboloid with a point removed from it. (See the Appendix for the details.) The removed point corresponds to the future of the region $r\leqslant 2m$, or the region excluded from $\hat{\mathcal{M}}$. Recalling that a unit spacelike three-hyperboloid has a topology of $S^2\times\mathbb{R}^+$, we see that the topology of $\check{i}^+$ is now $S^2\times(\mathbb{R}^+ -\{p\})$ where $\mathbb{R}^+$ denotes $[0,\infty)$. This result may be regarded as a common feature of the space–times with sources and motivates the following definition.

Definition: A space–time $(\hat{\mathcal{M}},\hat{g}_{ab})$ is said to be *asymptotically flat at future timelike infinity with a source to order n* (AFFTIS-*n*) if $(\hat{\mathcal{M}},\hat{g}_{ab})$ possesses an asymptote at timelike infinity to order n and $\check{i}^+$ has topology of $S^2\times(\mathbb{R}^+-\{p\})$.

For the convenience of the following discussion, we define some useful tensors and show the relations among them. From the definition, the physical metric $\hat{g}_{ab}$ and its inverse $\hat{g}^{ab}$ are given by

$$\hat{g}_{ab}=-\Omega^{-4}F^{-1}\nabla_a\Omega\nabla_b\Omega+\Omega^{-2}q_{ab}\,, \tag{3.4}$$

$$\hat{g}^{ab}=-\Omega^4F^{-1}n^an^b+\Omega^2q^{ab}, \tag{3.5}$$

respectively, with functions and tensors that are smooth in $\mathcal{F}$ and on $\check{i}^+$. From Eqs. (3.5) and (3.1), we obtain Eq. (3.3). A smooth projection operator on the Ω-const surfaces can be defined from smooth q^{ab} and q_{ab}:

$$q^a{}_b:=q^{ac}q_{cb}=\delta^a{}_b+F^{-1}n^a\nabla_b\Omega. \tag{3.6}$$

The operator is the same as the one defined from $\hat{q}^{ab}$ and $\hat{q}_{ab}$. Thus, derivative operators D_a and $\hat{D}_a$ that are associated with q_{ab} and $\hat{q}_{ab}$, respectively, do not differ. Hence,

$${}^3\hat{R}_{abc}{}^d={}^3R_{abc}{}^d \quad\text{and}\quad {}^3\hat{R}_{ab}={}^3R_{ab}\,, \tag{3.7}$$

where $(\hat{D}_a\hat{D}_b-\hat{D}_b\hat{D}_a)\omega_c=:{}^3\hat{R}_{abc}{}^d\omega_d$, ${}^3\hat{R}_{ab}:={}^3\hat{R}_{abc}{}^c$, $(D_aD_b-D_bD_a)\omega_c=:{}^3R_{abc}{}^d\omega_d$ and ${}^3R_{ab}:={}^3R_{acb}{}^c$. On the other hand, the physical extrinsic curvature of the Ω-const surfaces does not admit a smooth limit to $\check{i}^+$:

$$\hat{K}_{ab}:=\tfrac{1}{2}\,\pounds_{\hat{n}}\hat{q}_{ab}=\Omega^{-1}F^{1/2}q_{ab}+\tfrac{1}{2}\,F^{-1/2}\pounds_nq_{ab}\,. \tag{3.8}$$

Thus, define a smooth tensor

$$K_{ab}:=\Omega\hat{K}_{ab}=F^{1/2}q_{ab}+\tfrac{1}{2}\,\Omega F^{-1/2}\pounds_nq_{ab}\,. \tag{3.9}$$

Hereafter, we raise and lower the indices of tensors without hats on Ω-const surfaces by the smooth metric q_{ab} and the inverse q^{ab}. For example, $K^a_b:=q^{ac}K_{cb}$.

IV. THE ZEROTH-ORDER ASYMPTOTIC STRUCTURE

In this section, we examine the asymptotic structure of an AFTI-0 space–time using the Einstein equation. Consider the momentum constraint equation expressed in terms of smooth tensors and $\hat{T}_{\hat{\mu}\hat{\nu}}$:

$$\Omega^{-2}(\hat{e}_I)^a\hat{n}^b\hat{T}_{ab}=[D_cK^c_a-D_aK](e_I)^a. \tag{4.1}$$

In AFTI-0 space–times, the left-hand side vanishes on $\check{i}^+$ because of the falloff condition on the energy-momentum tensor. Therefore, we get

$$D_cK_a{}^c-D_aK\stackrel{=}{}0. \tag{4.2}$$

Substituting Eq. (3.9) into this equation, we find

$$F\stackrel{=}{}\text{const}. \tag{4.3}$$

By virtue of the conformal freedom of the function F, the above constant can be set to unity without any loss of generality. To see this, complete the space–time with a function $\Omega' := \alpha\Omega$ instead of Ω, where α is a smooth function that is constant on $\check{\imath}^+$. Then, the function $F' := -\pounds_{n'}\Omega'$ in this completion is given by $F' \doteq \alpha^{-2}F$. Thus, there is always a function Ω' that induces $F' \doteq 1$ for any given function Ω. We call a completion with such a function a *unit hyperboloid completion* for the reason that will be clarified below. Because there is no loss of generality, hereafter we consider only unit hyperboloid completions. In the completion,

$$F = 1 + \sum_{l=1}^{\infty} {}^{(l)}F \cdot \Omega^l, \tag{4.4}$$

where ${}^{(n)}F$ are defined by

$${}^{(1)}F := \lim_{\to \check{\imath}^+} \Omega^{-1}(F-1) \quad \text{and} \quad {}^{(n)}F := \lim_{\to \check{\imath}^+} \Omega^{-n}\left(F - 1 - \sum_{l=1}^{n-1} {}^{(l)}F \right) \quad \text{for } n \geqslant 2; \tag{4.5}$$

and

$$K_{ab} = q_{ab} + \tfrac{1}{2}\Omega({}^{(1)}F q_{ab} + \pounds_n q_{ab}) + O(\Omega^2). \tag{4.6}$$

Now consider the space–space components of the Ricci tensor:

$$\Omega^{-2}(\hat{e}_I)^a(\hat{e}_J)^b \hat{R}_{ab} = [{}^3R_{ab} + KK_{ab} - F^{1/2}K_{ab} - 2K_{ac}K^c_b + \Omega F^{-1/2}\pounds_n K_{ab} + F^{-1/2}D_aD_bF^{1/2}] \times (e_I)^a(e_J)^b. \tag{4.7}$$

In AFTI-0 space–times, the left-hand side vanishes on $\check{\imath}^+$ by virtue of the falloff condition on the energy-momentum tensor. Substituting Eqs. (4.4) and (4.6) into Eq. (4.7), we obtain

$${}^3R_{ab} \doteq -2q_{ab}. \tag{4.8}$$

Here, recall that $\check{\imath}^+$ is a submanifold of $\mathcal{M}$ and thus there is an identity imbedding Π of $\check{\imath}^+$ in $\mathcal{M}$. Using the map Π, a tensor field A_{ab} defined on $\mathcal{M}$ induces a tensor field $\mathbf{A}_{ab} := \Pi^+ A_{ab}$ on $\check{\imath}^+$. (Hereafter, we denote the tensor fields on the manifold $\check{\imath}^+$ in boldface.) It is important to note here that if a tensor field is tangential to $\check{\imath}^+$, or $q_a{}^m q_b{}^n A_{mn} \doteq A_{ab}$, the induced tensor field contains the same information as the original. Hence, the investigation of an equation on $\check{\imath}^+$ with tensor fields tangential to $\check{\imath}^+$ can be done by considering the induced equation on $\check{\imath}^+$. Thus, let us consider the following equation induced by Eq. (4.8) on $\check{\imath}^+$:

$${}^3\mathbf{R}_{ab} \doteq -2\mathbf{q}_{ab}. \tag{4.9}$$

The equation tells us that the unphysical three-space–time $(\check{\imath}^+, \mathbf{q}_{ab})$ has negative constant curvature:

$${}^3\mathbf{R}_{abcd} \doteq -2\mathbf{q}_{a[c}\mathbf{q}_{d]b}. \tag{4.10}$$

In other words, timelike infinity $(\check{\imath}^+, \mathbf{q}_{ab})$ is isometric to a unit spacelike hyperboloid in Minkowski space–time and $\mathbf{q}_{ab} \doteq \mathbf{h}_{ab}$. Note here that it is isometric to a *unit* hyperboloid because the *unit* hyperboloid completion induces $F \doteq 1$. This is where the name comes from.

The falloff condition on the time–time component of the energy-momentum tensor gives nothing new.

Now that we have explored everything that can be derived from the falloff condition which AFTI-0 space–times satisfy, the asymptotic structure of AFTI-0 space–time can be summarized as follows.

Definition: The function F which is unity on $\check{\imath}^+$, i.e., $F \doteq 1$ and the manifold $\check{\imath}^+$ with normalized negative-constant curvature, i.e., $\mathbf{q}_{ab} \doteq \mathbf{h}_{ab}$ are called the *zeroth-order asymptotic structure* of an AFTI-0 space–time.

The zeroth-order asymptotic structure corresponds to the universal structure of AM space–times in Ref. 5. However, they differ in that the former does not contain the equivalence class of Ω constructed by regarding two completions Ω and Ω' equivalent if and only if they satisfy

$$\lim_{\to \check{i}^+} \frac{\Omega' - \Omega}{\Omega^2} = 0 \tag{4.11}$$

while the latter does. The reason is that the equivalence class does not arise on its own from the definition of an AFTI-0 space–time. This observation is also supported by the discussion of asymptotic symmetries in Sec. V.

Recalling that the difference between an AFTI-0 space–time and an AFFTIS-0 space–time is the topology of $\check{i}^+$, we define the asymptotic structure of an AFFTIS-0 space–time as follows.

Definition: The function F which is unity on $\check{i}^+$, i.e., $F \stackrel{\check{i}^+}{=} 1$, and the manifold $\check{i}^+$ with normalized negative-constant curvature, i.e., $\mathbf{q}_{ab} \stackrel{\check{i}^+}{=} \mathbf{h}_{ab}$ and topology of $S^2 \times (\mathbb{R}^+ - \{p\})$, are called the *zeroth-order asymptotic structure* of an AFFTIS-0 space–time.

Let us consider the form of the physical metric $\hat{g}_{ab}$ in an AFTI-0 space–time to understand the zeroth-order asymptotic structure more. From the fact that an AFTI-0 space–time possesses the zeroth-order asymptotic structure, we see that the expansions of the function F and the tensor field q_{ab} are given by

$$F = 1 + O(\Omega),$$

$$q_{ab} = h_{ab} + f (d\Omega)_a (d\Omega)_b + f^I (d\Omega)_{(a} (e_I)_{b)} + O(\Omega)$$

in an AFTI-0 space–time where f and f^I are smooth functions. With the expansion above, the definitions of F and q_{ab} give

$$\hat{g}_{ab} = -\Omega^{-4}[1 + O(\Omega)](d\Omega)_a (d\Omega)_b + \Omega^{-2}[h_{ab} + f (d\Omega)_a (d\Omega)_b + f^I (d\Omega)_{(a} (e_I)_{b)} + O(\Omega)]. \tag{4.12}$$

Introducing a conformal time $\eta := \ln \Omega$, we obtain

$$\hat{g}_{ab} = {}^{(0)}\hat{g}_{ab} + \Omega\, {}^{(1)}\hat{g}_{ab} + \cdots, \tag{4.13}$$

where

$${}^{(0)}\hat{g}_{ab} := (e^{-\eta})^2[-(d\eta)_a (d\eta)_b + h_{ab}] \tag{4.14}$$

and $e^{2\eta} \cdot {}^{(0)}\hat{g}_{ab}, e^{2\eta} \cdot {}^{(1)}\hat{g}_{ab}, \ldots$ are tensor fields that do not depend on Ω. If we transform the coordinates, it can be seen that ${}^{(0)}\hat{g}_{ab}$ is a metric of Minkowski space–time. Note that ${}^{(1)}\hat{g}_{ab}$ is arbitrary in an AFTI-0 space–time. That is, Eq. (4.13) tells us that an AFTI-0 space–time is asymptotically Minkowskian but how it approaches an asymptotically Minkowski space–time is not specified in any way.

V. ASYMPTOTIC SYMMETRIES

In this section, we define asymptotic symmetries by considering the transformations that preserve the zeroth-order asymptotic structure of an AFTI-0 and an AFFTIS-0 space–time and analyze their structures. Basically, this section is a review of the corresponding work of Ref. 5 but differs in the following points. First, the definition of the asymptotic symmetries is different due to the difference in the asymptotic structure they are to preserve. Second, one of the subgroups of the asymptotic symmetry group is examined in detail for the explicit evaluation of four-momentum in Sec. VIII. Third, the structure of the asymptotic symmetries differ due to the presence of a bounded source at the late time regime in an AFFTIS space–time.

In Minkowski space–time, the symmetry group, i.e., the Poincaré group, is a group of diffeomorphisms $\{\phi\}$ that preserve the structure of the space–time $\hat{\eta}_{ab}$: $\phi^* \hat{\eta}_{ab} = \hat{\eta}_{ab}$. Because the

structure that is universal to AFTI-0 or AFFTIS-0 space–times is the zeroth-order asymptotic structure, we define the asymptotic symmetries of such space–times to be transformations that preserve the asymptotic structures.

First we examine the infinitesimal asymptotic symmetries. As they are *local* transformations, the structure they preserve is that of an AFTI-0 space–time.

Definition: The infinitesimal Ti symmetry group $\mathcal{L}_{\mathcal{G}}$ is a group of equivalence classes of a C^∞ vector field ξ^a that satisfies the following conditions:

(1) ξ^a induces a smooth function $\boldsymbol{\alpha}$ given by

$$\boldsymbol{\alpha} := \lim_{\to \check{\imath}^+} \frac{\pounds_\xi \Omega}{\Omega^2} \tag{5.1}$$

and

(2) ξ^a induces a Killing vector of $(\check{\imath}^+, \mathbf{h}_{ab})$,

$$\pounds_{\boldsymbol{\xi}} \mathbf{h}_{ab} \doteq 0. \tag{5.2}$$

Two vector fields ξ_1^a and ξ_2^a are regarded as equivalent if and only if $(\boldsymbol{\alpha}_1, \boldsymbol{\xi}_1^a) \doteq (\boldsymbol{\alpha}_2, \boldsymbol{\xi}_2^a)$. [Ti stands for timelike infinity and rhymes with Scri($\mathcal{I}^\pm$) as Ashtekar's Spi(i^0) does.]

The first condition demands that $\Psi_\xi^*(\Omega)$ differs from Ω by terms of $O(\Omega^2)$ so that $\Psi_\xi^*(\Omega)$ also induces a unit hyperboloid completion and thus ξ^a preserves the zeroth-order asymptotic structure $F \doteq 1$, where Ψ denotes an infinitesimal map generated by ξ. The second condition ensures that ξ^a preserves the zeroth-order asymptotic structure $\mathbf{q}_{ab} \doteq \mathbf{h}_{ab}$. The equivalence class is defined because the vector fields that belong to the same class cannot be distinguished as a generator of symmetries.

Let us investigate the structure of the infinitesimal Ti group. Because its element is a C^∞ vector field, $\mathcal{L}_{\mathcal{G}}$ has the structure of a Lie algebra. The Lie bracket on the pairs is given by

$$[(\boldsymbol{\alpha}, \boldsymbol{\xi}^a), (\boldsymbol{\beta}, \boldsymbol{\eta}^a)] \hat{=} (\pounds_{\boldsymbol{\xi}} \boldsymbol{\beta} - \pounds_{\boldsymbol{\eta}} \boldsymbol{\alpha}, [\boldsymbol{\xi}, \boldsymbol{\eta}]^a). \tag{5.3}$$

The introduction of the following subgroup of $\mathcal{L}_{\mathcal{G}}$ helps the investigation of the structure on $\mathcal{L}_{\mathcal{G}}$.

Definition: The *infinitesimal Ti supertranslation group* $\mathcal{L}_{\mathcal{S}}$ is a subgroup of $\mathcal{L}_{\mathcal{G}}$ whose element $(\boldsymbol{\alpha}, \boldsymbol{\xi})$ satisfies

$$\boldsymbol{\xi}^a \doteq 0. \tag{5.4}$$

Because the subgroup contains the following infinitesimal Ti *translation* group, we call it infinitesimal Ti *supertranslation* group.

Definition: The *infinitesimal Ti translation group* $\mathcal{L}_{\mathcal{T}}$ is a subgroup of $\mathcal{L}_{\mathcal{S}}$ whose element $(\boldsymbol{\alpha}, 0)$ satisfies

$$\mathbf{D}_a \mathbf{D}_b \boldsymbol{\alpha} - \boldsymbol{\alpha} \mathbf{h}_{ab} \hat{=} 0, \tag{5.5}$$

where $\mathbf{D}_a$ is the derivative operator associated with the intrinsic metric $\mathbf{h}_{ab}$ on $\check{\imath}^+$. (The group does not depend on the function Ω used in the completion of a space–time because ξ^a vanishes on $\check{\imath}^+$ and thus α does not depend on Ω.)

It is important to note here that there are *four* independent functions that satisfy Eq. (5.5):

$$\boldsymbol{\alpha}_t \hat{=} \cosh \chi, \quad \boldsymbol{\alpha}_x \hat{=} \sinh \chi \sin \theta \cos \phi, \quad \boldsymbol{\alpha}_y \hat{=} \sinh \chi \sin \theta \sin \phi, \quad \text{and} \quad \boldsymbol{\alpha}_z \hat{=} \sinh \chi \cos \theta, \tag{5.6}$$

where (χ, θ, ϕ) is the standard hyperbolic coordinate on $\check{\imath}^+$. It can be shown that if $\check{\imath}^+$ is timelike infinity of n-dimensional space–time and thus $\mathbf{h}_{ab}$ in Eq. (5.5) is the $(n-1)$-hyperboloid metric, there are n independent functions satisfying the equation. The number corresponds to the number of possible ''directions'' of translation in the space–time. When Minkowski space–time is completed, the translational Killing fields correspond to the elements of $\mathcal{L}_{\mathcal{T}}$. It will be shown that the group is closely related with the asymptotic *translation* Killing fields introduced in Sec. VI and the four-momentum of the AFFTIS space–time introduced in Sec. VIII.

The structure of $\mathcal{L}_{\mathcal{G}}$ can be examined by noting that $\mathcal{L}_{\mathcal{S}}$ is a Lie ideal of $\mathcal{L}_{\mathcal{G}}$.[5] Hence, the quotient $\mathcal{L}_{\mathcal{G}}/\mathcal{L}_{\mathcal{S}}$ has the structure of a Lie algebra consisting of cosets $(\{\boldsymbol{\alpha}\},\boldsymbol{\xi}^a)$ where $\{\boldsymbol{\alpha}\}$ is a set of all the functions smooth on $\check{\imath}^+$. Thus, each element of the quotient is characterized by $\boldsymbol{\xi}^a$. In other words, $\mathcal{L}_{\mathcal{G}}/\mathcal{L}_{\mathcal{S}}$ is isomorphic to the group of the Killing fields of $(\check{\imath}^+,\mathbf{h}_{ab})$. Recalling that such Killing fields are in the Lie algebra of the Lorentz group $\mathcal{L}_L$, we obtain

$$\mathcal{L}_{\mathcal{G}}/\mathcal{L}_{\mathcal{S}}\cong\mathcal{L}_L \quad \text{and thus} \quad \mathcal{L}_{\mathcal{G}}\cong\mathcal{L}_{\mathcal{S}}\textcircled{S}\mathcal{L}_L, \tag{5.7}$$

where Ⓢ denotes the semi-direct sum here.

Next, let us consider the asymptotic symmetry group. Following the symmetry group of Minkowski space–time, we define the asymptotic symmetry group as a group of diffeomorphisms. Because AFTI-0 space–times possess only the zeroth-order asymptotic structure in common, the diffeomorphisms can be only defined on $\check{\imath}^+$. However, $\check{\imath}^+$ of an AFTI-0 space–time is not specified enough to consider diffeomorphisms generated on itself. Hence, we consider $\check{\imath}^+$ of an AFFTIS-0 space–time whose topology is restricted and demand that $\check{\imath}^+$ be geodesically complete so that diffeomorphisms could be generated on it.

Definition: The *Ti group* $\mathcal{G}$ is a group of diffeomorphisms generated by elements of $\mathcal{L}_{\mathcal{G}}$ on $\check{\imath}^+$ that is geodesically complete except for the geodesics through a point and that has topology of $S^2\times(\mathbb{R}^+-\{p\})$.

Recall that a three-manifold with negative-constant curvature with topology of $S^2\times(\mathbb{R}^+-\{p\})$ is a spacelike three-hyperboloid with a removed point. Hence, it is only when the removed point is chosen as the origin of the Lorentz transformations that the elements of $\mathcal{L}_L$ do generate diffeomorphisms. Otherwise, the integral curves of the elements of $\mathcal{L}_L$ would be incomplete due to the point. Hence, the Ti group $\mathcal{G}$ is a semi-direct product of the infinite-dimensional additive group of smooth functions on the unit spacelike hyperboloid with the Lorentz group around the removed point, L_{RP}:

$$\mathcal{G}\cong\mathcal{S}\textcircled{S}L_{\mathrm{RP}}. \tag{5.8}$$

The above additive group is *infinite* dimensional in contrast with the *four*-dimensional additive subgroup of the Poincaré groups, the translation group. This is because, as for transformations of the zeroth-order asymptotic structures, translations are the changes of the function Ω used in completions and the changes are possible in ''infinitely different ways.'' It is interesting that this structure is very similar to that of the Bond–Metzner–Sachs (BMS) group defined at null infinity which is a semi-direct product of the infinite-dimensional additive group of *conformally weighted functions on the two-sphere* and the Lorentz group.[8–11]

However, it is important to note here that the additive groups $\mathcal{S}$ in Eq. (5.8) consists of all the possible finite supertranslation. This is against our intuition that there are no spacelike translations when there exists a source, or a removed point on $\check{\imath}^+$, that breaks the symmetries. The reason traces back to the fact that the Ti group is defined only to preserve the structure of $\check{\imath}^+$ but the structure of $\check{\imath}^+$ is preserved as long as the function α is smooth, regardless of the detailed nature of α, which characterizes supertranslations. In the next section, we find that a translation whose associated function is α transforms the points in the vicinity of $\check{\imath}^+$ in $\mathcal{F}$ by the generator $\Omega D_a\alpha$. Hence, to take the actions of the supertranslations into the symmetry group more appropriately, we need to define the group $\mathcal{G}'$ to consist of diffeomorphisms in *the vicinity* of $\check{\imath}^+$. Now, the generators of the group must be vanishing in the vicinity of the removed point on $\check{\imath}^+$, and thus $\mathcal{G}'$ is a semi-direct product of the infinite-dimensional additive group $\mathcal{S}_{\mathrm{RP}}$ of smooth functions whose gradient vanishes in the vicinity of the removed point and the Lorentz group around the removed point:

$$\mathcal{G}'\cong\mathcal{S}_{\mathrm{RP}}\textcircled{S}L_{\mathrm{RP}}. \tag{5.9}$$

This modified Ti group $\mathcal{G}'$ does not contain ''spacelike supertranslation'' because, as we shall see in the next section, the gradients of the functions α associated with spacelike translations do not vanish in the vicinity of the removed point.

VI. ASYMPTOTIC KILLING FIELDS

In this section, we introduce the notion of asymptotic Killing fields. The introduction helps us to understand that the asymptotic symmetries are closely related with the hierarchy found in the asymptotic structure and gives a basis for discussing how fast a space–time becomes stationary.

If a vector field $\hat{\xi}^a$ is a Killing vector field, the Lie derivative of the metric with respect to $\hat{\xi}^a$ vanishes. This fact motivates us to call the vector field $\hat{\xi}^a$ an *asymptotic* Killing field if $\pounds_{\hat{\xi}}\hat{g}_{ab}$ *asymptotically* vanishes to the limit $\Omega\to 0$.

Definition: A space–time $(\hat{\mathcal{M}},\hat{g}_{ab})$ is said to admit an *asymptotic Killing field* $\hat{\xi}^a$ to order n if

$$\lim_{\to \check{\imath}^+} \Omega^{-n}(\pounds_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}=0, \tag{6.1}$$

where $(\pounds_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}:=\Psi^*((\hat{e}_\mu)^a(\hat{e}_\nu)^b\pounds_{\hat{\xi}}\hat{g}_{ab})$.

The local asymptotic symmetries defined above are related to the infinitesimal Ti group defined in the previous section as follows.

Proposition 1: In an AFTI-0 space–time, a vector field $\hat{\xi}^a$ belongs to the infinitesimal Ti group if and only if $\hat{\xi}^a$ is an asymptotic Killing field to order 0.

Proof of If: The components of the Lie derivative of the metric with respect to $\hat{\xi}^a$ are given by

$$\begin{aligned}
\hat{n}^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=4\Omega^{-1}\pounds_{\hat{\xi}}\Omega-2F^{-1}\pounds_{[\hat{\xi},n]}\Omega-F^{-1}\pounds_{\hat{\xi}}F,\\
(\hat{e}_I)^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=[\Omega^{-1}F^{-1/2}D_m(\pounds_{\hat{\xi}}\Omega)-\Omega q_{ma}F^{-1/2}\pounds_{\hat{\xi}}n^a](e_I)^m,\\
(\hat{e}_I)^a(\hat{e}_J)^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=[-2\Omega^{-1}(\pounds_{\hat{\xi}}\Omega)q_{mn}+q_m{}^aq_n{}^b\pounds_{\hat{\xi}}q_{ab}](e_I)^m(e_J)^n
\end{aligned} \tag{6.2}$$

in terms of smooth functions and tensors. Because $\hat{\xi}^a$ is an asymptotic Killing field to order 0, all the components must be of $O(\Omega^1)$ by the definition. Hence, from the time–time component of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ in the above equation, we find that $\alpha:=\pounds_{\hat{\xi}}\Omega/\Omega^2$ must be a smooth function. Hence, $\boldsymbol{\alpha}:=\lim_{\to \check{\imath}^+}\alpha$ is a smooth function on $\check{\imath}^+$ and the above identities become

$$\begin{aligned}
\hat{n}^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=0+\Omega\mathcal{L}^{(1)}_{\hat{\xi}}F+O(\Omega^2),\\
(\hat{e}_I)^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=[0+\Omega(D_m\alpha-q_{ma}\mathcal{L}_{\hat{\xi}}n^a)+O(\Omega^2)](e_I)^m,\\
(\hat{e}_I)^a(\hat{e}_J)^b\pounds_{\hat{\xi}}\hat{g}_{ab}&=[q_m{}^aq_n{}^b\mathcal{L}_{\hat{\xi}}q_{ab}-2\Omega\alpha q_{mn}](e_I)^m(e_J)^n
\end{aligned} \tag{6.3}$$

in AFTI-0 space–times where Eq. (4.4) is used in the evaluation. Next, from the space–space component, we find

$$q_m{}^aq_n{}^b\pounds_{\hat{\xi}}q_{ab}\doteq 0, \tag{6.4}$$

must be satisfied so that all the components of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ are of $O(\Omega^1)$. The equation above induces

$$\pounds_{\hat{\boldsymbol{\xi}}}\mathbf{h}_{ab}\doteq 0 \tag{6.5}$$

on $\check{\imath}^+$. The equation tells us that $\hat{\xi}^a$ induces a Killing field of $(\check{\imath}^+,\mathbf{h}_{ab})$.

To summarize, an asymptotic Killing field to order 0 generates a smooth function $\boldsymbol{\alpha}$ on $\check{\imath}^+$ and induces a Killing field of $(\check{\imath}^+,\mathbf{h}_{ab})$. The vector field belongs to the infinitesimal Ti group. □

Proof of Only If: If $\hat{\xi}^a$ belongs to the infinitesimal Ti group, $\alpha:=\pounds_{\hat{\xi}}\Omega/\Omega^2$ is a smooth function from the definition. Thus, in AFTI space–times Eqs. (6.3) are satisfied. Further, $\hat{\xi}^a$ induces a Killing field of $(\check{\imath}^+,\mathbf{h}_{ab})$. Hence, the first term in the space–space component vanishes resulting in $(\pounds_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}=O(\Omega^1)$. That is, $\hat{\xi}^a$ is an asymptotic Killing field to order 0. □

Equations (6.3) show that it was appropriate to define the zeroth-order asymptotic structure such that it does not contain the equivalence class of Ω defined by Eq. (4.11). As we have seen above, if the components of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ are of order $O(\Omega)$, then the vector field ξ^a preserves the zeroth-order asymptotic structure. However, the vector field ξ^a preserves the equivalence class of Ω, only when $D_m\alpha-q_{ma}\pounds_{\hat{\xi}}n^a\doteq 0$ were satisfied in addition, or the time–space components of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ were of $O(\Omega^2)$, as it can be seen from Eqs. (6.3). (For the details on the preservation of the

class, see Ref. 5.) The fact that the time–space components of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ must be $O(\Omega^2)$ instead of $O(\Omega^1)$ to preserve the class indicates that the equivalence class of Ω is of *an order higher* asymptotic structure.

Next, we investigate the asymptotic translational Killing fields. This is to see how fast an AFTI-0 space–time becomes stationary.

Proposition 2: In an AFTI-0 space–time,

$$\begin{aligned}
\hat{\xi}_t{}^a &= \Omega[\sinh\chi (e_1)^a] + \Omega^2 F^{-1/2}\cosh\chi n^a,\\
\hat{\xi}_x{}^a &= \Omega[\cosh\chi \sin\theta\cos\phi (e_1)^a + \cos\theta\cos\phi (e_2)^a - \sin\phi (e_3)^a] + \Omega^2 F^{-1/2}\sinh\chi n^a,\\
\hat{\xi}_y{}^a &= \Omega[\cosh\chi \sin\theta\sin\phi (e_1)^a + \cos\theta\sin\phi (e_2)^a + \cos\phi (e_3)^a] + \Omega^2 F^{-1/2}\sinh\chi n^a,\\
\hat{\xi}_z{}^a &= \Omega[\cosh\chi \cos\theta (e_1)^a - \sin\theta (e_2)^a] + \Omega^2 F^{-1/2}\sinh\chi n^a
\end{aligned} \tag{6.6}$$

are admitted as asymptotic Killing fields to order one, where $e_1=\partial_\chi$, $e_2=(\sinh^{-1}\chi)\partial_\theta$ and $e_3=(\sinh^{-1}\chi\sin^{-1}\theta)\partial_\phi$. $\{\hat{\xi}_{x^\mu}\}_{\mu=0,1,2,3}(\{x^\mu\}=\{t,x,y,z\})$ are four *independent asymptotic translational Killing fields* in the sense that, with respect to the physical metric $\hat{g}_{ab}$, their norms are constant and they are orthogonal to each other:

$$\hat{g}_{ab}(\hat{\xi}_{x^\mu})^a(\hat{\xi}_{x^\nu})^b = \mp\delta_{\mu,\nu},$$

where $-$ is for $\mu=0$ and $+$ is for $\mu=1,2,3$.

Proof: The space components of the above asymptotic translational Killing vector fields are of $O(\Omega)$ and the time components are of $O(\Omega^2)$. Thus, the expansion of the vector field $\hat{\xi}^a$ is given by

$$\hat{\xi}^a = \sum_{l=1}^{\infty}\Omega^{l^{(l)}}\xi^I(e_I)^a + \sum_{l=2}^{\infty}\Omega^{l^{(l)}}\xi^0(\partial_\Omega)^a, \tag{6.7}$$

where $\{{}^{(n)}\xi^\mu\}$ is a set of smooth functions independent of Ω. Hence,

$$\alpha := \pounds_{\hat{\xi}}\Omega/\Omega^2 = \sum_{l=2}^{\infty}{}^{(l)}\xi^0\Omega^{l-2} \tag{6.8}$$

is a smooth function. Thus, the components of $\pounds_{\hat{\xi}}\hat{g}_{ab}$ are given by

$$\begin{aligned}
\hat{n}^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab} &= 0 + O(\Omega^2),\\
(\hat{e}_I)^a\hat{n}^b\pounds_{\hat{\xi}}\hat{g}_{ab} &= [\Omega(D_m\alpha - {}^{(1)}\xi^J(e_J)_m) + O(\Omega^2)](e_I)^m,\\
(\hat{e}_I)^a(\hat{e}_J)^b\pounds_{\hat{\xi}}\hat{g}_{ab} &= [\Omega(q_m{}^a q_n{}^b\pounds^{(1)}\xi^J(e_J)q_{ab} - 2\alpha q_{mn}) + O(\Omega^2)](e_I)^m(e_J)^n.
\end{aligned} \tag{6.9}$$

From the equations above, we see that $(\pounds_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}} = O(\Omega^2)$ if the space components of the $O(\Omega^1)$ terms of the vector field $\hat{\xi}^a$ satisfy

$$D_m\alpha - {}^{(1)}\xi^J(e_J)_m \doteq 0 \quad \text{and} \quad q_m{}^a q_n{}^b\pounds^{(1)}\xi^J(e_J)q_{ab} - 2\alpha q_{mn} \doteq 0. \tag{6.10}$$

These equations are tangential to $\check{\imath}^+$ and thus induce

$${}^{(1)}\boldsymbol{\xi}^a \doteq \mathbf{D}^a\boldsymbol{\alpha} \quad \text{and} \quad \pounds^{(1)}\boldsymbol{\xi}\mathbf{h}_{ab} - 2\boldsymbol{\alpha}\mathbf{h}_{ab} \doteq 0 \tag{6.11}$$

on $\check{\imath}^+$ without any loss of information where ${}^{(1)}\boldsymbol{\xi}^a$ and $\boldsymbol{\alpha}$ denote the vector field and the function induced by ${}^{(1)}\xi^J(e_J)^a$ and $\lim_{\to\check{\imath}^+}\alpha := {}^{(2)}\xi^0$ on $\check{\imath}^+$, respectively. Thus, the conditions can be restated as

$$\mathbf{D}_a\mathbf{D}_b\boldsymbol{\alpha}-\boldsymbol{\alpha}\mathbf{h}_{ab}\doteq 0, \tag{6.12}$$

$$^{(1)}\boldsymbol{\xi}^a\doteq\mathbf{D}^a\boldsymbol{\alpha}. \tag{6.13}$$

In other words, $(\pounds_{\hat{\xi}}\hat{g})_{\hat{\mu}\hat{\nu}}=O(\Omega^2)$ is satisfied if the leading term of the time component of $\hat{\xi}$, $\boldsymbol{\alpha}=^{(2)}\boldsymbol{\xi}^0$, satisfies Eq. (6.12) and if those of the space components $^{(1)}\boldsymbol{\xi}^a$ are given by Eq. (6.13). Note here that Eq. (6.12) is the same as Eq. (5.5) used in defining the infinitesimal Ti *translation* and that $\{\boldsymbol{\alpha}_{x^\mu}\}$ in Eq. (5.6) are the four independent smooth functions that satisfy the equation. Then, from Eq. (6.13), we see that the corresponding independent space components are

$$\begin{aligned}
^{(1)}\boldsymbol{\xi}_t{}^a&\doteq\sinh\chi(\mathbf{e}_1)^a,\\
^{(1)}\boldsymbol{\xi}_x{}^a&\doteq\cosh\chi\,\sin\theta\,\cos\phi(\mathbf{e}_1)^a+\cos\theta\,\cos\phi(\mathbf{e}_2)^a-\sin\phi(\mathbf{e}_3)^a,\\
^{(1)}\boldsymbol{\xi}_y{}^a&\doteq\cosh\chi\,\sin\theta\,\sin\phi(\mathbf{e}_1)+\cos\theta\,\sin\phi(\mathbf{e}_2)^a+\cos\phi(\mathbf{e}_3)^a\\
^{(1)}\boldsymbol{\xi}_z{}^a&\doteq\cosh\chi\,\cos\theta(\mathbf{e}_1)^a-\sin\theta(\mathbf{e}_2)^a.
\end{aligned} \tag{6.14}$$

The space components of the vector fields in Eqs. (6.6) do take this form and thus are asymptotic Killing fields to order one in Ω. □

As a corollary of Proposition 2, we obtain the relation between the infinitesimal Ti translational group $\mathcal{L}_{\mathcal{T}}$ and the above asymptotic translational Killing fields.

Corollary: The asymptotic translational Killing fields to order one, Eqs. (6.6), belong to the infinitesimal Ti translational group $\mathcal{L}_{\mathcal{T}}$.

Proof: Admitted as asymptotic Killing fields to order one, the asymptotic translational Killing fields belong to the infinitesimal Ti group according to Proposition 1. In addition, the fields vanish on $\check{\imath}^+$ and generate smooth functions satisfying Eq. (5.5). □

Let us consider the finite translations in AFFTIS-0 space–times. From Eqs. (6.7) and (6.13), we see that the generators of translations vanish on $\check{\imath}^+$ but transform the points in vicinity of $\check{\imath}^+$ in the direction of $\Omega D^a\alpha$, or $^{(1)}\xi^a$. Although the generator $^{(1)}\xi_t{}^a$ of a timelike translation vanishes in the vicinity of the removed point $\chi=0$ on $\check{\imath}^+$, the generators $^{(1)}\xi_x{}^a$, $^{(1)}\xi_y{}^a$, and $^{(1)}\xi_z{}^a$ of spacelike translations do not. Hence, there exist only timelike translation symmetry $\hat{\xi}_t$ and no spacelike translation symmetries in AFFTIS-0 space–times. In other words, the existence of source in an AFFTIS-0 space–time breaks the spacelike translation symmetries but leaves the timelike symmetries.

From the above discussion, it can be said that if the energy-momentum tensor of a space–time falls off at a rate faster than Ω^2, the space–time becomes asymptotically stationary at the rate of Ω^2.

We conclude the section with a remark on the relation of the asymptotic symmetries with the hierarchy in asymptotic structure. We have shown that asymptotic translational Killing fields are admitted to order 1 in AFTI-0 space–times. However, we can see that the fields are not admitted to order 2 in general. For the vector fields to be admitted to order 2, the coefficients of the $O(\Omega^2)$ terms in Eqs. (6.9) need to vanish. These terms contain the higher-order asymptotic structure, such as $^{(1)}F$. For example, the $O(\Omega^2)$ term of the time–time component is $\Omega^2[D^{a(1)}FD_a\alpha+^{(1)}F\alpha-2n^a\nabla_a\alpha]$. These higher-order asymptotic quantities are *arbitrary* in AFTI-0 space–times. Hence, in general, the $O(\Omega^2)$ terms in Eqs. (6.9) do not vanish and thus asymptotic translational Killing fields to order two are not admitted in general AFTI-0 space–times. The asymptotic translational Killing fields to order 2 are the asymptotic symmetries that belong to the zeroth-order asymptotic structure.

VII. GRAVITATIONAL FIELDS

In this section, we investigate the gravitational fields and find hierarchy in their asymptotic behaviors.

Because the Ricci tensor of $\hat{g}_{ab}$ is reserved for the definition of AFTI-n space–times, we examine the Weyl tensor $\hat{C}_{abcd}$ to investigate the gravitational fields. For convenience of the discussion, decompose the Weyl tensor into the electric part $\hat{E}_{ab}:=\hat{C}_{ambn}\hat{n}^m\hat{n}^n$ and the magnetic

part $\hat{B}_{ab}:=*\hat{C}_{ambn}\hat{n}^m\hat{n}^n$, where $*\hat{C}_{ambn}$ denotes the dual of the two-form $\hat{C}_{[am]bn}$. In terms of quantities associated with the Ω-const surfaces and the physical components of the energy-momentum tensor, $\hat{E}_{ab}$ and $\hat{B}_{ab}$ are given by

$$\hat{E}_{ab}=\hat{K}_{ar}\hat{K}^r{}_b-\pounds_{\hat{n}}\hat{K}_{ab}+\hat{D}_{(a}\hat{a}_{b)}+\hat{a}_a\hat{a}_b+\tfrac{1}{2}(\hat{q}^r_a\hat{q}^s_b-\hat{q}_{ab}\hat{n}^r\hat{n}^s)\hat{L}_{rs},$$
$$\hat{B}_{ab}=\hat{\epsilon}_{ra}{}^s\hat{D}^r\hat{K}_{bs}+\tfrac{1}{2}\hat{\epsilon}_{ab}{}^r\hat{n}^s\hat{L}_{rs}, \tag{7.1}$$

where $\hat{L}_{ab}:=\hat{R}_{ab}-\frac{1}{6}\hat{R}\hat{g}_{ab}$ and $\hat{a}_a:=\hat{q}_{ar}\hat{n}^s\hat{\nabla}_s\hat{n}^r$.

In AFTI-0 space–times, we find that $\hat{E}_{ab}$ and $\hat{B}_{ab}$ vanish on $\check{\imath}^+$:

$$\hat{E}_{ab}=\sum_{l=1}^{\infty}\Omega^{l(l)}E_{ab} \quad \text{and} \quad \hat{B}_{ab}=\sum_{l=1}^{\infty}\Omega^{l(l)}B_{ab}, \tag{7.2}$$

where ${}^{(n)}E_{ab}$ and ${}^{(n)}B_{ab}$ are defined by

$${}^{(1)}E_{ab}:=\lim_{\to\check{\imath}^+}\Omega^{-1}\hat{E}_{ab}, \quad {}^{(n)}E_{ab}:=\lim_{\to\check{\imath}^+}\Omega^{-n}\left(\hat{E}_{ab}-\sum_{l=1}^{n-1}{}^{(l)}E_{ab}\right) \quad \text{for } n\geqslant 2, \tag{7.3}$$

$${}^{(1)}B_{ab}:=\lim_{\to\check{\imath}^+}\Omega^{-1}\hat{B}_{ab}, \quad {}^{(n)}B_{ab}:=\lim_{\to\check{\imath}+}\Omega^{-n}\left(\hat{B}_{ab}-\sum_{l=1}^{n-1}{}^{(l)}B_{ab}\right) \quad \text{for } n\geqslant 2. \tag{7.4}$$

Note here that, although they depend on the choice of Ω, ${}^{(n)}E_{ab}$ and ${}^{(n)}B_{ab}$ are defined, regardless of whether their lower-order terms vanish. This is in contrast with the treatment of Ref. 5 where the subsequent order terms are defined only when their lower-order terms vanish. Their leading terms are given by

$${}^{(1)}E_{ab}=\tfrac{1}{2}(D_aD_b-h_{ab})^{(1)}F+\tfrac{1}{2}(e_I)_a(e_J)_b({}^{(3)}\hat{L}_{\hat{I}\hat{J}}+\delta_{I,J}{}^{(3)}\hat{L}_{\hat{0}\hat{0}}), \tag{7.5}$$

$${}^{(1)}B_{ab}=\tfrac{1}{2}\epsilon_{ra}{}^sD^{r(1)}K_{bs}+\tfrac{1}{2}\epsilon_{ab}{}^r(e_I)_r{}^{(3)}\hat{L}_{\hat{0}\hat{I}}, \tag{7.6}$$

where ${}^{(1)}K_{ab}:=\lim_{\to\check{\imath}^+}\Omega^{-1}(K_{ab}-h_{ab})$, $\hat{L}_{\hat{0}\hat{0}}:=\Psi*[\hat{L}_{ab}\hat{n}^a\hat{n}^b]$, $\hat{L}_{\hat{0}\hat{I}}:=\Psi*[\hat{L}_{ab}\hat{n}^{(a}(\hat{e}_I)^{b)}]$, $\hat{L}_{\hat{I}\hat{J}}:=\Psi*[\hat{L}_{ab}(\hat{e}_I)^a(\hat{e}_J)^b]$, and ${}^{(n)}\hat{L}_{\hat{\mu}\hat{\nu}}$ denotes the coefficient of the $O(\Omega^n)$ term of $\hat{L}_{\hat{\mu}\hat{\nu}}$. ${}^{(1)}E_{ab}$ and ${}^{(1)}B_{ab}$ depend on ${}^{(3)}\hat{L}_{\hat{\mu}\hat{\nu}}$ that consists of the higher-order asymptotic structure such as ${}^{(1)}F$ and ${}^{(1)}K_{ab}$. Because these quantities are arbitrary in AFTI-0 space–times, we cannot conclude anything more than that $\hat{E}_{ab}$ and $\hat{B}_{ab}$ vanish on $\check{\imath}^+$.

Let us examine the gravitational fields in AFTI-1 space–times. The falloff condition on the energy-momentum tensor demands ${}^{(3)}\hat{L}_{\hat{\mu}\hat{\nu}}$ vanish on $\check{\imath}^+$ and we obtain

$${}^{(1)}E_{ab}\stackrel{\check{}}{=}\tfrac{1}{2}(D_aD_b-h_{ab})^{(1)}F \quad \text{and} \quad {}^{(1)}B_{ab}\stackrel{\check{}}{=}\tfrac{1}{2}\epsilon_{ra}{}^sD^{r(1)}K_{bs}. \tag{7.7}$$

The equation on ${}^{(1)}E_{ab}$ induces

$${}^{(1)}\mathbf{E}_{ab}\stackrel{\check{}}{=}\tfrac{1}{2}(\mathbf{D}_a\mathbf{D}_b-\mathbf{h}_{ab})^{(1)}\mathbf{F} \tag{7.8}$$

on $\check{\imath}^+$. Using Eq. (4.10), it can be shown that ${}^{(1)}\mathbf{E}_{ab}$ obeys the field equation

$$\mathbf{D}_{[a}{}^{(1)}\mathbf{E}_{b]c}\stackrel{\check{}}{=}0. \tag{7.9}$$

Contracting over a and b and recalling that ${}^{(1)}\mathbf{E}_{ab}$ is traceless by definition, we find that ${}^{(1)}\mathbf{E}_{ab}$ is divergence-free:

$$\mathbf{D}_a{}^{(1)}\mathbf{E}^{ab}\stackrel{\check{}}{=}0. \tag{7.10}$$

It can be shown that ${}^{(1)}\mathbf{B}_{ab}$ satisfies the same field equation. Contracting the Bianchi identity $\nabla_{[a}\hat{R}_{bc]}{}^{de}=0$ with $\hat{n}^a\hat{\epsilon}^{bc}{}_m\hat{\epsilon}_{den}$, we obtain

$$\hat{\epsilon}^{ac}{}_{m}\hat{D}_a\hat{B}_{cn}-\hat{n}^a\hat{\nabla}_a\hat{E}_{mn}+\hat{\epsilon}^{ac}{}_{m}\hat{\epsilon}^{de}{}_{n}\hat{K}_{ad}\hat{E}_{ce}-2\hat{\epsilon}^{bc}{}_{(m}\hat{B}_{n)c}\hat{a}_b + \tfrac{1}{2}(\hat{n}^{[a}\hat{q}^{c]e}\hat{q}_{mn}-\hat{n}^{[a}\hat{q}^{c]}_n\hat{q}^e_m)\hat{\nabla}_a\hat{L}_{ce}=0 \tag{7.11}$$

from which we find, in AFTI-n space–times,

$$D_{[a}{}^{(1)}B_{b]c}=O(\Omega)+O(\Omega^n), \tag{7.12}$$

where we have used the equations $n^r\nabla_r(\hat{e}_I)_a=-(e_I)_a+O(\Omega)$, $\Omega\cdot n^r\nabla_r\hat{L}_{\hat{I}\hat{J}}=O(\Omega^{n+3})$, $\hat{L}_{\hat{\mu}\hat{\nu}}=O(\Omega^{n+3})$ and $D_a\hat{L}_{\hat{0}\hat{I}}=O(\Omega^{n+3})$ to evaluate the derivative of $\hat{L}_{ab}$:

$$\begin{aligned}(\hat{n}^{[a}\hat{q}^{c]e}\hat{q}_{mn}-\hat{n}^{[a}\hat{q}^{c]}_n\hat{q}^e_m)\hat{\nabla}_a\hat{L}_{ce}&=\Omega^{-1}[q_{mn}(\Omega n^a\nabla_a\hat{L}_{\hat{I}\hat{J}}-2\hat{L}_{\hat{I}\hat{I}}+2\hat{L}_{\hat{0}\hat{0}}-(e_I)^eD_e\hat{L}_{\hat{0}\hat{I}})\\&\quad+(e_I)_m(e_J)_n(\Omega n^a\nabla_a\hat{L}_{\hat{I}\hat{J}}-\hat{L}_{\hat{I}\hat{J}})+(e_I)_mD_n\hat{L}_{\hat{0}\hat{I}}+O(\Omega^{n+4})]\\&=O(\Omega^{n+2}).\end{aligned}$$

Hence, in AFTI-1 spacetimes, the equation $D_{[a}{}^{(1)}B_{b]c}\stackrel{=}{=}0$ is satisfied and induces

$$\mathbf{D}_{[a}{}^{(1)}\mathbf{B}_{b]c}\stackrel{=}{=}0 \tag{7.13}$$

on $\check{\imath}^+$. Thus ${}^{(1)}\mathbf{B}_{ab}$ is also divergence-free:

$$\mathbf{D}_a{}^{(1)}\mathbf{B}^{ab}\stackrel{=}{=}0. \tag{7.14}$$

Here, it is important to note that the explicit form of the function ${}^{(1)}\mathbf{F}$ can be obtained. Because ${}^{(1)}\mathbf{E}_{ab}$ is traceless by definition, the contraction of Eq. (7.8) gives a differential equation for the function ${}^{(1)}\mathbf{F}$**:**

$$\mathbf{D}^a\mathbf{D}_a{}^{(1)}\mathbf{F}-3{}^{(1)}\mathbf{F}\stackrel{=}{=}0. \tag{7.15}$$

We have obtained the nontrivial equation because the Einstein equation is used in a nontrivial way to obtain Eq. (7.8). That is, the terms that are not traceless, $\frac{1}{2}(e_I)_a(e_J)_b({}^{(3)}\hat{L}_{\hat{I}\hat{J}}+\delta_{I,J}{}^{(3)}\hat{L}_{\hat{0}\hat{0}})$, are discarded to obtain the equation. In contrast, we do not obtain any nontrivial equations from noting that ${}^{(1)}B_{ab}$ is traceless because the discarded terms to obtain the corresponding equation (7.7) are traceless. The general solution of Eq. (7.15) takes the form

$${}^{(1)}\mathbf{F}(\chi,\theta,\phi)=\sum_{\substack{l\geqslant 0\\|m|\leqslant l}}\mathbf{a}_{lm}{}^{(1)}\mathbf{F}_{lm}(\chi,\theta,\phi), \tag{7.16}$$

where

$${}^{(1)}\mathbf{F}_{lm}(\boldsymbol{\chi},\boldsymbol{\theta},\boldsymbol{\phi}):=\frac{1}{\sqrt{\sinh\boldsymbol{\chi}}}\mathbf{P}^{l+1/2}_{3/2}(\cosh\boldsymbol{\chi})\mathbf{Y}_{lm}(\boldsymbol{\theta},\boldsymbol{\phi}). \tag{7.17}$$

Here, $P^\mu_\nu(z)$ denote the associated Legendre functions defined by

$$P^\mu_\nu(z):=\frac{1}{\Gamma(1-\mu)}\left(\frac{z+1}{z-1}\right)^{\mu/2}F\left(-\nu,\nu+1,1-\mu;\frac{1-z}{2}\right)\quad\text{for } z>1, \tag{7.18}$$

where $F(\alpha,\beta,\gamma;z)$ are the hypergeometric functions. Here Y_{lm} stands for the spherical harmonics with their coefficients chosen as

$$\mathbf{Y}_{\diagup m}(\theta,\phi)=\begin{cases}\sqrt{2\pi}\mathbf{P}^m_l(\cos\theta)\cdot\cos m\phi, & m\geqslant 0,\\ \sqrt{2\pi}\mathbf{P}^m_l(\cos\theta)\cdot\sin m\phi, & m<0,\end{cases} \tag{7.19}$$

so that their physical meanings can be conveniently read in the next section. Explicitly, we obtain

$$^{(1)}\mathbf{F}_{00}=2\left(2\sinh\boldsymbol{\chi}+\frac{1}{\sinh\boldsymbol{\chi}}\right), \tag{7.20}$$

$$^{(1)}\mathbf{F}_{11}=2\left(2\cosh\boldsymbol{\chi}-\frac{\cosh\boldsymbol{\chi}}{\sinh^2\boldsymbol{\chi}}\right)\sin\boldsymbol{\theta}\cos\phi. \tag{7.21}$$

When the Schwarzschild spacetimes are completed, the function F indeed consists of the above functions. See the Appendix for the explicit calculations.

To define an angular momentum, it is important to see the conditions for $^{(2)}\mathbf{B}_{ab}$ to be divergence-free. This can be done using the Bianchi identity $\nabla_{[a}\hat{R}_{bc]}{}^{de}=0$ contracted with $\hat{\boldsymbol{\epsilon}}^{abc}\hat{n}_d\hat{q}^s{}_e$,

$$\hat{D}_r\hat{B}^{rs}=\tfrac{1}{2}\hat{\boldsymbol{\epsilon}}_s{}^{rp}\hat{n}^q\nabla_r\hat{L}_{pq}. \tag{7.22}$$

Keeping in mind that $^{(1)}\mathbf{B}_{ab}$ is divergence-free, we find that the above equation induces

$$\mathbf{D}_r{}^{(2)}\mathbf{B}^{rs}\stackrel{\scriptscriptstyle\sim}{=}{}^{(4)}\mathbf{L}_{IJ}\boldsymbol{\epsilon}^{srp}(\mathbf{e}_I)_r(\mathbf{e}_J)_p. \tag{7.23}$$

Hence, $^{(2)}\mathbf{B}_{ab}$ is not divergence-free in general AFTI-1 space–times. However, in AFTI-2 space–times $^{(4)}\mathbf{L}_{IJ}$ vanishes on $\check{\imath}^+$ and $^{(2)}\mathbf{B}_{ab}$ is divergence-free:

$$\mathbf{D}_r{}^{(2)}\mathbf{B}^{rs}\stackrel{\scriptscriptstyle\sim}{=}0. \tag{7.24}$$

Note here that $^{(1)}B_{ab}$ does not need to vanish on $\check{\imath}^+$ for $^{(2)}\mathbf{B}^{ab}$ to be divergence-free. This point is unclear in, Ref. 5.

We conclude the section with a remark on the relation of the asymptotic behavior of gravitational fields with the hierarchy in the asymptotic structure. The gravitational fields of each order, $^{(n)}E_{ab}$ and $^{(n)}B_{ab}$, are constructed from the asymptotic structure of the corresponding order. Hence, if the energy-momentum tensor falls off at a low rate and the space–time lacks the higher-order asymptotic structure, the asymptotic behaviors of the higher-order gravitational fields are arbitrary. In an AFTI-0 space–time, $^{(0)}E_{ab}$ and $^{(0)}B_{ab}$ vanish on $\check{\imath}^+$ and $^{(l)}E_{ab}$ and $^{(l)}B_{ab}$ take arbitrary forms for $l\geqslant 1$. In an AFTI-1 space–time, $^{(1)}E_{ab}$ and $^{(1)}B_{ab}$ induces divergence-free tensor fields on $\check{\imath}^+$ that satisfy the field equations (7.9) and (7.13), respectively, and $^{(1)}E_{ab}$ takes the form given by Eqs. (7.16) and (7.7). $^{(l)}E_{ab}$ and $^{(l)}B_{ab}$ take arbitrary forms for $l\geqslant 2$. In an AFTI-2 space–time, $^{(2)}B_{ab}$ induce a divergence-free tensor field on $\check{\imath}^+$.

VIII. CONSERVED QUANTITIES

In this section, we define two conserved quantities, namely four-momentum of ATI-1 space–times and angular momentum of ATI-2 space–times, which are associated with the asymptotic symmetries that the spacetimes possess.

As Minkowski space–time admits infinitesimal *translation* symmetries, it possesses a conserved four-momentum. The four-momentum is defined as a linear mapping from the four-dimensional vector space of infinitesimal translations to real numbers. Because ATI-n space–times possess infinitesimal Ti translations, we expect that the four-momentum of ATI-n space–times can be also defined as a linear mapping of infinitesimal Ti translations to real numbers.

The four-momentum representing the state of the space–time should be defined with information from the gravitational fields. However, ATI-0 space–times possess no asymptotic gravitational fields, as we have seen in the previous section. This means that the four-momentum cannot be defined in ATI-0 space–times. Now, consider ATI-1 space–times. Here, we have gravitational fields $^{(1)}\mathbf{E}_{ab}$ and $^{(1)}\mathbf{B}_{ab}$. On the other hand, infinitesimal Ti, translations may be represented by $\mathbf{D}^a\boldsymbol{\alpha}_{\mathcal{T}}$, where $\boldsymbol{\alpha}_{\mathcal{T}}$ is a smooth function generated by the elements of $\mathcal{L}_{\mathcal{T}}$. The vector field is a conformal Killing field of $(\check{\imath}^+,\mathbf{h}_{ab})$ by definition:

$$\mathbf{D}_{(a}\mathbf{D}_{b)}\boldsymbol{\alpha}_{\mathcal{T}}\stackrel{\scriptscriptstyle\sim}{=}2\boldsymbol{\alpha}_{\mathcal{T}}\mathbf{h}_{ab}. \tag{8.1}$$

With these $^{(1)}\mathbf{E}_{ab}$ and $\mathbf{D}^a\boldsymbol{\alpha}_{\mathcal{T}}$, the four-momentum of an ATI-1 space–time can be defined as follows.

Definition: Four-momentum $\mathbf{P}$ of an ATI-1 space–time with a region $\mathcal{A}$ on $\check{\imath}^+$ that satisfies

(1) $\mathcal{A}$ is bounded by cross sections $\mathcal{S}_a$ and $\mathcal{S}_b$ that do not intersect with each other, i.e., $\partial\mathcal{A}=\mathcal{S}_a\cap\mathcal{S}_b$ and $\mathcal{S}_a\cap\mathcal{S}_b=\varnothing$,
(2) q_{ab} and n^a have smooth limit to $\mathcal{A}\subset\check{\imath}^+$ and functions $\boldsymbol{\alpha}_{\mathcal{T}}$ generated by elements of $\mathcal{L}_{\mathcal{T}}$ are smooth on $\mathcal{A}$, is a linear mapping of $\boldsymbol{\alpha}_{\mathcal{T}}$ to real numbers defined by

$$\mathbf{P}:\boldsymbol{\alpha}_{\mathcal{T}}\mapsto\mathbb{R} \overset{\text{def}}{\Leftrightarrow} \mathbf{P}(\boldsymbol{\alpha}_{\mathcal{T}}):=-\frac{1}{8\pi}\int_{\mathcal{S}}\boldsymbol{\epsilon}_{ab}{}^{m(1)}\mathbf{E}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}, \tag{8.2}$$

where $\mathcal{S}$ is a cross section of the region $\mathcal{A}$.

Note that, in general, ATI-1 space–times do not possess a region $\mathcal{A}$ satisfying the above properties; q_{ab} and n^a are demanded to have smooth limits to $\check{\imath}^+\cap\mathcal{F}$ in ATI-1 space–times but the region $\check{\imath}^+\cap\mathcal{F}$ does not need to satisfy the above properties (1) and (2). However, we need such a region on $\check{\imath}^+$ to define a ''conserved quantity'' as we shall see below.

The four-momentum can be considered as a ''conserved quantity'' because it is invariant under a change of the cross section $\mathcal{S}$ chosen for the evaluation. This is because the two-form integrand is closed in $\mathcal{A}$: the three-form $\mathbf{D}_{[c}\boldsymbol{\epsilon}_{ab]}{}^{m(1)}\mathbf{E}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}$ is proportional to

$$\boldsymbol{\epsilon}^{cab}\mathbf{D}_c(\boldsymbol{\epsilon}_{ab}{}^{m(1)}\mathbf{E}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}})\doteq(\mathbf{D}^{m(1)}\mathbf{E}_{mn})\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}+{}^{(1)}\mathbf{E}_{mn}(\tfrac{1}{2}\boldsymbol{\alpha}_{\mathcal{T}}\mathbf{h}^{mn})\doteq 0, \tag{8.3}$$

where we used, in the first line, the fact that $\mathbf{D}^a\boldsymbol{\alpha}_{\mathcal{T}}$ is a conformal Killing field and, in the second line, the facts that ${}^{(1)}\mathbf{E}_{ab}$ is divergence-free and traceless. In other words, the four-momentum is ''conserved'' in $\mathcal{A}$ because it consists of a smooth conformal Killing field and a smooth divergence-free, traceless gravitational field.

Recall that there are *four* independent functions of $\boldsymbol{\alpha}_{\mathcal{T}}$ as shown in Eq. (5.6). Thus, there are *four* independent components of *four* momentum $\mathbf{P}$ as we expect.

Now let us evaluate the four-momentum defined above and see where the information is contained. First, we evaluate the $\boldsymbol{\alpha}_t$ component. For simplicity, choose the $\chi=$const surface in $\mathcal{A}$ as $\mathcal{S}$. We obtain

$$\mathbf{P}(\boldsymbol{\alpha}_t)=\frac{1}{16\pi}\int_{\chi=\text{const}}\left(\frac{\partial^2}{\partial\boldsymbol{\chi}^2}{}^{(1)}\mathbf{F}-{}^{(1)}\mathbf{F}\right)\sinh\boldsymbol{\chi}\sqrt{\mathbf{h}}\,d\boldsymbol{\theta}d\boldsymbol{\phi}, \tag{8.4}$$

where $\mathbf{h}:=\det(\mathbf{h}_{ij})=\sinh^2\boldsymbol{\chi}\sin\boldsymbol{\theta}$. Substituting the general solution of ${}^{(1)}\mathbf{F}$, Eq. (7.16), and using the orthogonality of spherical harmonics

$$\int\mathbf{Y}_{lm}\,d(\cos\boldsymbol{\theta})d\boldsymbol{\phi}=4\sqrt{2}\pi^{3/2}\delta_{l,0}\delta_{m,0}, \tag{8.5}$$

we find that all the terms but ${}^{(1)}\mathbf{F}_{00}$ of ${}^{(1)}\mathbf{F}$ vanish and obtain

$$\mathbf{P}(\boldsymbol{\alpha}_t)=\mathbf{a}_{00}. \tag{8.6}$$

The spatial momentum can be evaluated likewise as

$$\mathbf{P}(\boldsymbol{\alpha}_x)=\mathbf{a}_{11},\quad \mathbf{P}(\boldsymbol{\alpha}_y)=\mathbf{a}_{1-1},\quad \mathbf{P}(\boldsymbol{\alpha}_z)=\mathbf{a}_{10}. \tag{8.7}$$

Here, we find that the $l=0$, 1 modes of ${}^{(1)}\mathbf{F}$ contain the information on the four-momentum of an AFTI-0 space–time. Since the function ${}^{(1)}\mathbf{F}$ is of the first-order asymptotic structure, the coefficients of ${}^{(1)}\mathbf{F}$ and thus the four momentum $\mathbf{P}$ is also of the first-order asymptotic structure. If we complete the Kerr space–time and the Schwarzschild space–time, we obtain the expected answers for the four-momentum **P:** $\mathbf{a}_{00}$ equals the mass parameter m. See Eqs. (A3) and (A9) for the explicit calculations. Moreover, $\mathbf{a}_{11}$ equals $-mv$ in the leading order of v for the Schwarzschild space–time boosted by a constant velocity v along the x axis. See Eq. (A11).

Because ${}^{(1)}\mathbf{B}_{ab}$ is also divergence-free and traceless and $\boldsymbol{\xi}\in\mathcal{L}_L$ is a conformal Killing field of $(\check{\imath}^+,\mathbf{h}_{ab})$, we obtain a new ''conserved quantity'' if we replace ${}^{(1)}\mathbf{E}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}$ in Eq. (8.2) with one

of ${}^{(1)}\mathbf{B}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}$, ${}^{(1)}\mathbf{E}_{mn}\boldsymbol{\xi}^n$, and ${}^{(1)}\mathbf{B}_{mn}\boldsymbol{\xi}^n$. However, it is found that they all give exact two-form integrands. Hence, the ''conserved quantities'' identically vanish.[5] Because there are no more gravitational fields in an AFFTIS-1 space–time, four-momentum $\mathbf{P}$ can be considered as the sole ''conserved quantity'' of AFFTIS-1 space–times.

Next, consider AFFTIS-2 space–times. There exists a gravitational field ${}^{(2)}\mathbf{B}_{ab}$ that is divergence-free and traceless in such space–times. Hence, we yield another ''conserved quantity'' by replacing ${}^{(1)}\mathbf{E}_{mn}\mathbf{D}^n\boldsymbol{\alpha}_{\mathcal{T}}$ in Eq. (8.2) with ${}^{(2)}\mathbf{B}_{mn}\boldsymbol{\xi}^n$. This quantity can be regarded as angular momentum if a gauge condition ${}^{(1)}\mathbf{B}_{ab}\doteq 0$ is imposed. This is to reduce the Ti group to the Poincaré group so that the angular momentum is defined around an origin in the *four*-dimensional affine space of the infinitesimal Ti translations.[5] Note here that it is clarified that the gauge condition is needed for the quantity to be defined around the four-dimensional origin, not for the quantity to be ''conserved.'' This point was left unclear in Ref. 5.

We conclude the section with a remark on how the conserved quantities are associated with the hierarchy in the asymptotic structure. To define the conserved quantities, we needed vector fields that respect the asymptotic symmetries and gravitational fields that are traceless and divergence-free. However, the asymptotically flat space–times that satisfy only the weakest falloff condition on the energy-momentum tensor do not possess sufficient asymptotic structures that the space–times lack for the conserved quantities to be well defined. As a result, the four-momentum is defined for an AFFTIS-1 space–time and angular momentum is defined for an AFFTIS-2 space–time.

IX. SUMMARY AND DISCUSSIONS

In this paper, we have proposed a new definition of asymptotic flatness at timelike infinity. Instead of rescaling the metric of a space–time conformally as in the previous works,[3,4] we rescaled three-metrics and normal vector fields to the time slices with different powers to provide the information on the gravitational fields. As a result, the causal structure of the physical space–time is not manifest in the unphysical space–time and the timelike infinity arises not as a point but as a *three-manifold*. The three-manifold timelike infinity enables us to characterize the asymptotically flat space–time under investigation with a falloff condition on the physical components of the physical energy-momentum tensor. Hence, it is possible to discuss the properties of such a space–time with *the energy-momentum tensor falling off at a certain rate* where the rate is specified with the function Ω that reads asymptotically the inverse of t, the observer's proper time. This makes a sharp contrast with the treatment using the conformal completion in which the space–time is characterized with the awkward differentiability conditions on the unphysical metric and thus such an argument is not possible. Hereafter in this section, we refer to the physical components of the physical energy-momentum tensor simply by *the energy-momentum tensor* for brevity.

We found the following hierarchy in the asymptotic structure of an asymptotically space–time. If the energy-momentum tensor falls off faster than the rate Ω^2, the space–time possesses an intrinsic structure called the zeroth-order asymptotic structure. The structure consists of a function F that is unity on timelike infinity $\check{\imath}^+$, and the manifold $\check{\imath}^+$ with a unit spacelike three-hyperboloid metric $\mathbf{h}_{ab}$. The presence of the structure means that the space–time is asymptotically Minkowskian. An asymptotic symmetry group, called the Ti group, can be defined to preserve the structure. The group is found to be similar to the Poincaré group but differs in that the four-dimensional translation group is replaced with an infinite-dimensional additive group. Further, by introducing the notion of an asymptotic Killing field, it is found that such a space–time becomes asymptotically stationary at the rate Ω^2 and in general does not become stationary at a faster rate. Although there exists an asymptotically symmetry group, the space–time is just asymptotically Minkowskian and thus does not possess gravitational fields with sufficient properties to have the four-momentum defined. If the energy-momentum tensor falls off faster than the rate Ω^3, there is enough higher-order asymptotic structure for the four-momentum to be defined. The form of one of the asymptotic structure ${}^{(1)}F$ is determined in such space–times. A conserved quantity associated with angular momentum can be defined only for space–times with even more asymptotic structure: the energy-momentum tensor needs to fall off faster than the rate Ω^4. For the quantity to be regarded as angular-momentum in the sense that the angular-momentum should be defined

around the *four*-dimensional origin, the leading order of one of the gravitational fields, ${}^{(1)}B_{ab}$, must vanish.

ACKNOWLEDGMENTS

We would like to thank Katsuhiko Sato, Akio Hosoya, and Yasushi Suto for their encouragement. We especially thank Akio Hosoya for his careful reading of the manuscript. We also thank Keita Ikumi for his fruitful discussion with us.

APPENDIX: SOME AFFTIS SPACE–TIMES AND SOME THAT ARE NOT

We give the explicit construction of the completion of the Schwarzschild space–time and show that it is an AFFTIS space–time. Then, it is shown that the nature of the charge and the angular momentum of the Kerr–Newmann space–time appears in the second-order asymptotic structure. The function $F:=-\pounds_n\Omega$ which contains the information on four-momentum is explicitly calculated for a boosted Schwarzschild space–time. At the end, an example of space–times which do not satisfy the conditions of AFFTIS-n space–times is given.

In static spherical coordinates, the metric of Schwarzschild space–time takes the form

$$\hat{g}_{ab}=-(1-2m/r)(1-2m/r)(dt)_a(dt)_b+(1-2m/r)^{-1}(dr)_a(dr)_b+r^2(d\sigma)_{ab}. \tag{A1}$$

Introducing new coordinates $t=\Omega^{-1}\cosh\chi$ and $r=\Omega^{-1}\sinh\chi$, we obtain smooth n^a, F and q_{ab}:

$$\begin{aligned} n^a&=-\frac{1+4m\Omega\sinh\chi-4m^2\Omega^2}{1-2m\Omega/\sinh\chi}(\partial_\Omega)^a-4m\frac{\cosh\chi-m\Omega/\tanh\chi}{1-2m\Omega/\sinh\chi}(\partial_\chi)^a\\ &=-(\partial_\Omega)^a-\left[\Omega m^{(1)}F_{00}+\Omega^2\frac{4m^2}{\tanh^2\chi}+O(\Omega^3)\right](\partial_\Omega)^a\\ &\quad-\left[4m\cosh\chi+\Omega\frac{4m^2}{\tanh\chi}+O(\Omega^2)\right](\partial_\chi)^a, \end{aligned} \tag{A2}$$

$$F=\frac{1+4m\Omega\sinh\chi-4m^2\Omega^2}{1-2m\Omega/\sinh\chi}=1+\Omega m^{(1)}F_{00}+\Omega^2\frac{4m^2}{\tanh^2\chi}+O(\Omega^3), \tag{A3}$$

and

$$\begin{aligned} q_{ab}&=16m^2\frac{\cosh^2\chi-2m\Omega\cosh\chi/\tanh\chi+m^2\Omega^2/\tanh^2\chi}{1+2m\Omega(2\sinh\chi-1/\sinh\chi)-12m^2\Omega^2+8m^3\Omega^3/\sinh\chi}(d\Omega)_a(d\Omega)_b\\ &\quad-4m\frac{\cosh\chi-m\Omega/\tanh\chi}{1-2m\Omega/\sinh\chi}[(d\Omega)_a(d\chi)_b+(d\chi)_a(d\Omega)_b]\\ &\quad+\frac{1+4m\Omega\sinh\chi-4m^2\Omega^2}{1-2m\Omega/\sinh\chi}(d\chi)_a(d\chi)_b+\sinh^2\chi(d\sigma)_{ab}\\ &=h_{ab}+\left[16m^2\cosh^2\chi-\Omega 64\frac{m^3\cosh\chi\sinh^2\chi}{\tanh\chi}+O(\Omega^2)\right](d\Omega)_a(d\Omega)_b\\ &\quad-\left[4m\cosh\chi+\Omega\frac{4m^2}{\tanh\chi}+O(\Omega^2)\right](d\chi)_{(a}(d\Omega)_{b)}+\left[\Omega 2m^{(1)}F_{00}+\Omega^2\frac{4m^2}{\tanh\chi}+O(\Omega^3)\right]\\ &\quad\times(d\chi)_a(d\chi)_b. \end{aligned} \tag{A4}$$

Together with the fact that it is a vacuum solution of the Einstein equation, the above calculations show that the Schwarzschild space–time is an AFTI-∞ space–time, where we denote a space–time that is an AFTI-n space–time for arbitrary n by an ATI-∞ space–time. Now take the outside of the event horizon $r>2m$ in the original coordinates as $\hat{\mathcal{F}}$. Because the region $r\leqslant 2m$ corresponds to

the point $\chi=0$ on the $\Omega=0$ surface, the choice provides timelike infinity $\check{\imath}^+$ with the point missing. Thus, topology of $\check{\imath}^+$ is $S^2\times(\mathbb{R}-\{p\})$ and the Schwarzschild space–time satisfies the condition of an AFFTIS-n space–time to arbitrary n. It is important to note here that the Oppenheimer–Snyder space–time, which describes dust collapsing to form a Schwarzschild black hole, can be completed to possess $\check{\imath}^+$ with the same structure as the above completed Schwarzschild space–time does. The space–time has the Schwarzschild metric outside the dust and the surface of the dust collapses along $r=(r_{ini}/2)(1+\cos\ \eta)$ where r and r_{ini} are the Schwarzschild radial coordinate and the initial dust radius at time $t=0$, respectively; and η is related to the Schwarzschild time coordinate t by

$$t=2m\ \ln\left|\frac{\sqrt{r_{ini}/2m-1}+\tan\ \eta/2}{\sqrt{r_{ini}/2m-1}-\tan\ \eta/2}\right|+2m\sqrt{r_{ini}/2m-1}\,(\eta+(\eta+\sin\ \eta)r_{ini}/4m). \tag{A5}$$

If we choose the region $r>r_{ini}$ as region $\hat{\mathcal{F}}$, the metric of the region is that of the Schwarzschild space–time and thus $\hat{\mathcal{F}}$ satisfies all the conditions of AFFTIS-n for arbitrary n and can be completed to have exactly the same n^a, F, and q_{ab} in $\mathcal{F}$ as those of the Schwarzschild space–time. Since the region $r\leqslant r_{ini}$ corresponds to the point $\chi=0$ on $\check{\imath}^+$ in this completion also, the global structure of $\check{\imath}^+$ is also the same as that of the above completed Schwarzschild space–time.

In static spherical coordinates, the metric of Kerr–Newmann space–time is given by

$$\begin{aligned}\hat{g}_{ab}=&-\frac{\Delta-a^2\ \sin^2\ \theta}{\Sigma}\,(dt)_a(dt)_b-2\,\frac{a\ \sin^2\ \theta(r^2+a^2-\Delta)}{\Sigma}\,(dt)_{(a}(d\phi)_{b)}+\frac{\Sigma}{\Delta}\,(dr)_a(dr)_b\\&+\Sigma(d\theta)_a(d\theta)_b+\frac{(r^2+a^2)^2-\Delta a^2\ \sin^2\ \theta}{\Sigma}\,\sin^2\ \theta(d\phi)_a(d\phi)_b,\end{aligned} \tag{A6}$$

where $\Sigma=r^2+a^2\ \cos^2\ \theta$ and $\Delta=r^2+a^2+e^2-2mr$. Introducing $\eta:=\ln\Omega$, we clearly see that the nature of the charge and the angular momentum of the space–time appear in the second-order asymptotic structure and do not alter the lower-order asymptotic structure:

$$\begin{aligned}n^a&=e^{-\eta}[{}^{(0)}n^a+\Omega^{(1)}n^a+\Omega^{2\ (2)}n^a+O(\Omega^3)],\\F&={}^{(0)}F+\Omega^{(1)}F+\Omega^{2\ (2)}F+O(\Omega^3),\\q_{ab}&={}^{(0)}q_{ab}+\Omega^{(1)}q_{ab}+\Omega^{2\ (2)}q_{ab}+O(\Omega^3),\end{aligned} \tag{A7}$$

where

$$\begin{aligned}&{}^{(0)}n^a=-(\partial_\eta)^a,\quad {}^{(1)}n^a=m^{(1)}F_{00}(\partial_\eta)^a-4m\ \cosh\ \chi(\partial_\chi)^a,\\&{}^{(2)}n^a=\left(\frac{4m^2-e^2}{\tanh\ \chi}-e^2-a^2\ \sin^2\ \theta\right)(\partial_\eta)^a-\frac{4m^2-2e^2-a^2\ \sin^2\ \theta}{\tanh\ \chi}\,(\partial_\chi)^a+2ma\,\frac{\cosh\ \chi}{\sinh^3\ \chi}\,(\partial_\phi)^a;\end{aligned} \tag{A8}$$

$$\begin{aligned}&{}^{(0)}F=1,\quad {}^{(1)}F=m^{(1)}F_{00},\\&{}^{(2)}F=\frac{4m^2-e^2}{\tanh\ \chi}-e^2-a^2\ \sin^2\ \theta;\end{aligned} \tag{A9}$$

$$\begin{aligned}&{}^{(0)}q_{ab}=h_{ab},\quad {}^{(1)}q_{ab}=-2(4m\ \cosh\ \chi)(d\eta)_{(a}(d\chi)_{b)}+2m^{(1)}F_{00}(d\chi)_a(d\chi)_b,\\&{}^{(2)}q_{ab}=16m^2\ \cosh^2\ \chi(d\eta)_a(d\eta)_b-2\,\frac{4m^2-2e^2-a^2\ \sin\ \theta}{\tanh\ \chi}\,(d\eta)_{(a}(d\chi)_{b)}\\&\quad+4ma\ \sin^2\ \theta\left(\frac{(d\eta)_{(a}(d\phi)_{b)}}{\tanh\ \chi}-(d\chi)_{(a}(d\phi)_{b)}\right)+\left(\frac{4m^2-e^2-a^2\ \sin\ \theta}{\tanh^2\ \chi}-e^2\right)(d\chi)_a(d\chi)_b\\&\quad+a^2\ \cos^2\ \theta(d\theta)_a(d\theta)_b+a^2\ \sin^2\ \theta(d\phi)_a(d\phi)_b.\end{aligned} \tag{A10}$$

In the boosted Schwarzschild space–time with velocity of v along the x axis, or to be precise in the Schwarzschild space–time that is observed by an observer with velocity of v along the x axis,

$$F=1+\Omega\left\{2m\left(2\sinh\chi+\frac{1}{\sinh\chi}\right)-2mv\left(2\cosh\chi-\frac{\cosh\chi}{\sinh^2\chi}\right)\sin\theta\cos\phi\right\}+O(\Omega^2)$$
$$=1+\Omega\{m^{(1)}F_{00}-mv^{(1)}F_{11}+O(v^2)\}+O(\Omega^2). \tag{A11}$$

Let us consider the flat Friedmann–Robertson–Walker space–time, which does not qualify as an AFTI space–time as one expects. The metric of the space–time filled with perfect fluid is

$$\hat{g}_{ab}=-(dt)_a(dt)_b+a^2(t)[(dr)_a(dr)_b+r^2(d\sigma)_{ab}],$$

where $a(t)=a_0(t/t_0)^{2/3(1+w)}$ if the energy density ρ and the pressure p is related by $p=w\rho$. In order to get smooth n^a and q_{ab}, we need to choose the new coordinates (Ω,χ) as

$$t=a_0\frac{1+3w}{1+w}\Omega^{-1}(\cosh\chi)^{(3+3w)/(1+3w)},$$

$$r=t_0^{2/(3+3w)}\Omega^{-(1+3w)/(3+3w)}(\sinh\chi).$$

They give us

$$n^a=\frac{1}{a_0^2}\left(\frac{3+3w}{1+3w}\right)^2(\cosh\chi)^{-4/(1+3w)}(\partial_\Omega)^a, \tag{A12}$$

$$q_{ab}=a_0^2(\cosh\chi)^{4/(1+3w)}[(d\chi)_a(d\chi)_b+\sinh^2\chi(d\sigma)_{ab}]. \tag{A13}$$

However, with this choice of Ω, the time–time component of the energy-momentum tensor decays as Ω^2:

$$\hat{n}^a\hat{n}^b\hat{T}_{ab}=9\rho_0\left(\frac{t}{t_0}\right)^{-2}\cosh^2\chi+O(\Omega)=\Omega^2\rho_0\left(\frac{3t_0(1+w)}{a_0(1+3w)}\right)^2(\cosh\chi)^{-4/(1+3w)}+O(\Omega^3),$$

from which we see that the space–time does not satisfy the third condition of AFTI-n space–times for non-negative n and thus the space–time is not an AFTI-n space–time. Hence, the space–time does not need to possess the properties of an AFTI-n space–time: the function

$$F\doteq\frac{1}{a_0^2}\left(\frac{3+3w}{1+3w}\right)^2(\cosh\chi)^{-4/(1+3w)} \tag{A14}$$

is not unity on $\check{\imath}^+$, and the space–time does not admit an asymptotic stationary Killing field.

[1] M. Heusler, *Black Hole Uniqueness Theorem* (Cambridge Univeristy Press, Cambridge, 1996) and references therein.
[2] R. O. Hansen, J. Math. Phys. **15**, 46 (1974).
[3] C. Cutler, Class. Quantum Grav. **6**, 1075 (1989).
[4] J. Porrill, Proc. R. Soc. London, Ser. A **381**, 323 (1982).
[5] A. Ashtekar and J. D. Romano, Class. Quantum Grav. **9**, 1069 (1992).
[6] R. M. Wald, *General Relativity* (Chicago University Press, Chicago, 1984).
[7] R. H. Price, Phys. Rev. D **5**, 2419 (1972); **5**, 2439 (1972); C. Gundlach, R. Price, and J. Pullen, *ibid.* **49**, 883 (1994); E. S. C. Ching, P. T. Leung, W. N. Suen, and K. Young, *ibid.* **52**, 2118 (1995); W. Krivan, P. Lagura, and P. Papadopoulos, *ibid.* **54**, 4728 (1996).
[8] H. Bondi, A. W. K. Metzner, and M. J. G. Van der Berg, Proc. R. Soc. London, Ser. A **269**, 21 (1962).
[9] R. K. Sachs, Proc. R. Soc. London, Ser. A **270**, 103 (1962); Phys. Rev. **128**, 2851 (1962).
[10] R. Penrose, Phys. Rev. Lett. **10**, 66 (1963); Proc. R. Soc. London, Ser. A **284**, 159 (1965).
[11] A. Ashtekar and R. O. Hansen, J. Math. Phys. **19**, 1542 (1978).

Simplicial quantum gravity in the elongated phase

Gabriele Gionti[a)]
Scuola Internazionale Superiore di Studi Avanzati, Via Beirut 2-4, 34013 Trieste, Italy

(Received 19 November 1997; accepted for publication 8 May 1998)

We discuss the elongated phase of four-dimensional simplicial quantum gravity by exploiting recent analytical results. In particular using Walkup's theorem we prove that the dominating configurations in the elongated phase are treelike structures called ''stacked spheres.'' Such configurations can be mapped into branched polymers and baby universes arguments are used in order to analyze the critical behavior of theory in the weak coupling regime. © *1998 American Institute of Physics.* [S0022-2488(98)03011-4]

I. INTRODUCTION

The key concept of simplicial quantum gravity is to work with a discrete version of space–time. To this purpose it is necessary to consider a simplicial complex Σ by gluing four-dimensional simplexes through their three-dimensional faces in such a way that the link of every p-dimensional simplex σ, $p=0,1,2$, $link(\sigma,\Sigma)$, is homeomorphic to a $(4-p-1)$ sphere. Analytically this last condition implies that the following Dehn–Sommerville constraints should hold:

$$\sum_{i=0}^{4}(-1)^i N_i(T)=\chi(T), \tag{1.1}$$

$$\sum_{i=2k-1}^{4}(-1)^i\frac{(i+1)!}{(i-2k+2)!(2k-1)!}N_i(T)=0 \tag{1.2}$$

with $1\leqslant k\leqslant 2$ and $N_i(T)$ the number of i-dimensional simplexes of the simplicial manifold T and where $\chi(T)$ is the Euler–Poincaré characteristic. Equation (1.1) is just the Euler–Poincaré equation, while (1.2) is a consequence of the fact that the link of every $2k$-simplex is an odd dimensional sphere, and hence has Euler number zero.

Two simplicial manifolds T and T' are said to be Picewise Linear (PL)-equivalent if there exist simplicial subdivisions of them, $\tilde{T}$ and $\tilde{T}'$, such that $\tilde{T}$ and $\tilde{T}'$ are simplicially isomorphic. A simplicial isomorphism[1] is a one to one map ϕ between the vertex of the simplicial manifolds such that $(\phi(v_i),\phi(v_j))$ is an edge of T' if and only if (v_i,v_j) is an edge of T and so on for every simplex of any dimension. A simplicial isomorphism is an equivalence relation among simplicial manifolds. An equivalence class of simplicial manifolds with respect to the isomormorphism above is a Picewise Linear (PL)-manifold.

A dynamical triangulation T_a is a simplicial manifold, that is a representative of a PL manifold, generated by gluing, along their adjacent faces, $N_4(T_a)$ equilateral simplexes, say σ^4, with fixed edge-length a. A dynamical triangulation gives a metric structure on a PL manifold (see, e.g., Ref. 2, p. 20).

The numbers $N_i(T)$ $i=0,...,4$ can be seen as the component of a five-dimensional vector f,

$$f\equiv(N_0(T),N_1(T),N_2(T),N_3(T),N_4(T)) \tag{1.3}$$

called the f-vector of the triangulation T. Dehn–Sommerville equations fix three components of f, leaving two undetermined. In simplicial gravity $N_2(T_a)$ and $N_4(T_a)$ are, usually, the two undetermined components. Their values are subject to some limitations as will be explained below.

a)Electronic mail: gionti@sissa.it

 © 1998 American Institute of Physics

Every two-dimensional simplex, in four dimension, is called bone (hinges in Regge calculus). Let us label the bones by an index α and denote by $q(\alpha)$ the number of four simplexes incident on it; the average number of simplexes incident on a bone is

$$b \equiv \frac{1}{N_2} \sum_{\alpha} q(\alpha) = 10 \left(\frac{N_4(T_a)}{N_2(T_a)} \right) \tag{1.4}$$

since each simplex is incident on 10 different bones. b ranges between two kinematical bounds which follows from Dehn–Sommerville and a theorem by Walkup.[3] This latter theorem is also relevant in classifying the ''elongated phase'' of four-dimensional simplicial quantum gravity as we shall see below.

Theorem: *If T is a triangulation of a closed, connected four-dimensional manifold then*

$$N_1(T) \geqslant 5N_0(T) - \tfrac{15}{2}\chi(T). \tag{1.5}$$

Moreover, equality holds if and only if $T \in H^4(1-\frac{1}{2}\chi(T))$, where the class of triangulations $H^4(n)$ is defined inductively according to: (a) The boundary complex of any abstract five-simplex $(Bd\sigma)$ is a member of $H^4(0)$. (b) If K is in $H^4(0)$ and σ is a four-simplex of K, then K' is in $H^4(0)$, where K' is any complex obtained from K by deleting σ and adding the joinder of the boundary complex $Bd\sigma$ and a new vertex distinct from the vertices of K. (c) If K is in $H^d(n)$, then K' is in $H^d(n+1)$ if there exist two four-simplexes σ_1 and σ_2 with no common vertices and a dimension preserving simplicial map ϕ from $K-\sigma_1-\sigma_2$ onto K' which identifies $Bd\sigma_1$ with $Bd\sigma_2$ but otherwise is one to one.

In other words $H^4(0)$ is built up by gluing together five-dimensional simplexes through their four-dimensional faces and considering only the boundary of this resulting complex. $H^4(n)$ differs from $H^4(0)$ by the fact that it has n handles. This way of constructing a triangulation of a four-sphere has a natural connection with the definition of a baby universe. A baby universe[4,5] is associated with a triangulation in which we can distinguish two pieces. A piece that contains the majority of the simplices of the triangulation—called the ''mother''—and a small part called the ''baby.'' In the ''minbus'' (minimum neck baby universes) the two parts are glued together along the boundary of a four-dimensional simplex (in four dimension) that is the ''neck'' of the baby universe. Thus the ''stacked spheres'' can be considered as a network of minbus in which the mother disappears and the baby universes have a minimal volume, that is the boundary of a five simplices minus the simplices of the necks through which they are glued to the others. We will exploit this parallel in Sec. VI to give an estimate of the number of distinct stacked spheres.

As established in Ref. 2, the Walkup theorem fixes a bound on the values of b, in fact formula (1.5) together with the Dehn–Sommerville equations leads to $b \geqslant 4 - 10(\chi/N_2)$.
From the Dehn–Sommerville equations and the fact that $N_0 \geqslant 5$, it follows that $b \leqslant 5 + (10\chi - 50)/N_2$. Thus in the limit of large N, one obtains $4 \leqslant b \leqslant 5$.

As in Regge[6] calculus the curvature of a dynamical triangulation is concentrated on the bones and depends on the number $q(B)$ of incident four-dimensional simplexes. More precisely, the curvature $K(B)$ at a bone B is

$$K(B) = \frac{48\sqrt{15}}{a^2} \left[\frac{2\pi - q(B)\cos^{-1}\frac{1}{4}}{q(B)} \right]. \tag{1.6}$$

The Regge version of Einstein–Hilbert action with cosmological constant for a dynamical triangulation T_a[5] is

$$S(T_a) = k_4 N_4(T_a) - k_2 N_2(T_a), \tag{1.7}$$

where k_4 is a constant depending linearly on the inverse of the gravitational constant and on the cosmological constant, whereas k_2 is proportional to the inverse of the gravitational constant; these constants depend also on the edge-length a; typically to simplify calculations one sets $a = 1$.

The partition function for four-dimensional dynamical triangulations is defined as

$$Z(k_2,k_4)=\sum_{\mathrm{top}(M)}\sum_{N_4(T_a)}\sum_{N_2(T_a)} W_{\mathrm{top}}(N_2,N_4)e^{-k_4N_4+k_2N_2}, \tag{1.8}$$

where $W_{\mathrm{top}}(N_4,N_2)$ is the number of distinct á la Tutte dynamical triangulations with a fixed number of four simplices N_4 and of bones N_2, and a fixed topology. Since in three and in four dimensions there is not yet a mathematical status for the sum over topologies we restrict ourselves to the simply connected topologies:

$$Z_{\mathrm{sc}}(k_2,k_4)=\sum_{N_4(T_a)}\sum_{N_2(T_a)} W_{\mathrm{sc}}(N_2,N_4)e^{-k_4N_4+k_2N_2}. \tag{1.9}$$

Equation (1.9) has the structure of a grand canonical partition function:

$$Z_{\mathrm{GC}}=\sum_N\sum_{\sigma_N} e^{-\beta H(\sigma_N)}z^N \tag{1.10}$$

with the number of simplexes that is playing the role of the number of particles. Following this analogy we refer to $W(N_2,N_4)$ as the microcanonical partition function and to

$$Z(k_2,N_4)=\sum_{N_2} W(N_2,N_4)e^{k_2N_2} \tag{1.11}$$

as the canonical partition function.

II. CANONICAL PARTITION FUNCTION

Recently it has been shown analytically[2] that the dynamical triangulations in four dimensions are characterized by two phases: a strong coupling phase in the region $k_2^{\mathrm{inf}}=\log 9/8<k_2<k_2^{\mathrm{crit}}$ and a weak coupling phase for $k_2>k_2^{\mathrm{crit}}$. k_2^{crit} is the value of k_2 for which in the infinite volume limit the theory has a phase transition from the strong to the weak phase (for a detailed analysis see Ref. 2). The transition between these two phases is characterized by the fact that the subdominant asymptotics of the number of distinct triangulations passes from an exponential to polynomial behavior (cf. Ref. 2, p. 106). Presently the precise value of k_2^{crit} has not been established yet, it is just known that it is close but distinct from $k_2^{\mathrm{max}}=\log 4$, recent numerical simulations suggest that $k_2^{\mathrm{crit}}=1.24$.

In the strong coupling phase the leading term of the asymptotic expansion of the canonical partition function is (cf. Ref. 2)

$$Z(N_4,k_2)=c_4\left(\frac{A(k_2)+2}{3A(k_2)}\right)^{-4}N_4^{-5}\exp[-m(\eta^*)N^{1/n_H}]\exp\left[\left[10\log\frac{A(k_2)+2}{3}\right]N_4\right], \tag{2.1}$$

where for notational convenience we have set

$$A(k_2)\equiv[\tfrac{27}{2}e^{k_2}+1+\sqrt{(\tfrac{27}{2}e^{k_2}+1)^2-1}]^{1/3}+[\tfrac{27}{2}e^{k_2}+1-\sqrt{(\tfrac{27}{2}e^{k_2}+1)^2-1}]^{1/3}-1 \tag{2.2}$$

and

$$\eta^*(k_2)=\frac{1}{3}\left(1-\frac{1}{A(k_2)}\right). \tag{2.3}$$

The explicit form of $m(\eta^*(k_2))$ and n_H, the Hausdorff dimension, are at present unknown.

In the weak phase $k_2>k_2^{\mathrm{crit}}$ the number of distinct dynamical triangulations with equal curvature assignment have a power law behavior in N_4. The phase is characterized by two distinct asymptotic regimes of the leading term of the canonical partition function: the critical coupling regime and the weak coupling regime. In the critical coupling regime $k_2^{\mathrm{crit}}<k_2<k_2^{\mathrm{max}}$ we have

$$Z(N_4,k_2)\asymp c_4\left(\frac{A(k_2)+2}{3A(k_2)}\right)^{-4} N_4^{\tau(\eta^*)-5}\ \exp\left[\left[10\log\frac{A(k_2)+2}{3}\right]N_4\right], \tag{2.4}$$

in which $\tau(\eta^*)$ is not explicitly known.

Instead in the weak coupling regime $k_2>k_2^{\max}$,

$$Z(N,k_2)\asymp \frac{c_4 e}{\sqrt{2\pi}}\frac{\eta_{\max}^{-1/2}(1-2\eta_{\max})^{-4}}{\sqrt{(1-3\eta_{\max})(1-2\eta_{\max})}}(\hat{N}+1)^{\tau-11/2}.\ \frac{e^{(\hat{N}+1)f(\eta_{\max},k_2)}}{k_2^{\sup}-k_2} \tag{2.5}$$

in which $\eta^{\max}=\frac{1}{4}$ for the sphere S^4 and

$$f(\eta,k_2)\equiv-\eta\log\eta+(1-2\eta)\log(1-2\eta)-(1-3\eta)\log(1-3\eta)+k_2\eta. \tag{2.6}$$

From this form of the canonical partition function it follows that in the case of the sphere S^4 and in the infinite volume limit the average value of b, see Eq. (1.4), is a decreasing function of k_2 in the critical coupling regime and it is constantly equal to 4 in the weak coupling regime, that is to say

$$\begin{gathered}\lim_{N_4\mapsto\infty}\langle b\rangle_{N_4}=\frac{1}{\eta(k_2)},\quad k_2^{\mathrm{crit}}\leqslant k_2\leqslant k_2^{\max},\\ \lim_{N_4\to\infty}\langle b\rangle_{N_4}=4,\quad k_2\geqslant k_2^{\max}.\end{gathered} \tag{2.7}$$

If we look at the average curvature we have

$$\lim_{N\to\infty}\frac{1}{N}\left\langle\sum_B K(B)\mathrm{vol}_4(B)\right\rangle=\pi a^2\sqrt{3}\left(10\eta^*(k_2)-\frac{5}{\pi}\cos^{-1}\frac{1}{4}\right),\quad k_2^{\mathrm{crit}}<k_2\leqslant k_2^{\max} \tag{2.8}$$

and

$$\lim_{N\to\infty}\frac{1}{N}\left\langle\sum_B K(B)\mathrm{vol}_4(B)\right\rangle=\pi a^2\sqrt{3}\left(\frac{5}{2}-\frac{5}{\pi}\cos^{-1}\frac{1}{4}\right),\quad k_2>k_2^{\max}, \tag{2.9}$$

in which $\mathrm{vol}_4(B)$ is the volume of the four simplexes incident on the bone B, and it is $a^2(\sqrt{3}/2)q(B)$. As we have already said the value of the average curvature is saturated.

This result was already known in numerical simulations[7] and, as we will see in detail in Sec. III, it is due to the prevalence of stacked spheres in the weak coupling regime.

III. ELONGATED PHASE

We have seen that in the case of triangulations of the sphere S^4 for $k_2\mapsto\log 4$, in the infinite volume limit, $\langle b\rangle_{N_4},\mapsto 4$. Walkup's theorem[3] implies that the minimum value of b is reached on triangulations K of the sphere S^4 that belong to $H^4(0)$ (stacked spheres). Then for $k_2>\log 4$ one has $\langle b\rangle_{N_4}=4$, that means that in this region of k_2 the statistical ensemble of quantum gravity is strongly dominated by stacked spheres.

As explicitly shown in Ref. 3, p. 83, the elements of $H^4(0)$ can be put in correspondence with a tree structure. Let us recall that a d-dimensional simplicial complex T, $d\geqslant 1$, is called a simple d-tree if it is the closure of its d-simplexes $\sigma_1,\ldots,\sigma_t$ and these d-simplexes can be ordered in such a way that

$$\mathrm{Cl}\ \sigma_j\cap\left\{\bigcup_{i=1}^{j-1}\mathrm{Cl}\ \sigma_i\right\}=\mathrm{Cl}\ \tau_j \tag{3.1}$$

for some $(d-1)$-face τ_j of σ_j, $j\geqslant 2$, and where the τ_j are all distinct. This ordering of the simplexes of T induces a natural ordering of its vertices in $v_1,\ldots,v_{t+d}$, where v_{i+d} is the vertex

of σ_i not in Cl τ_i. Note that the interior part of T contains the simplexes σ_i and faces τ_i. The boundary of T, Bd T, consists of the boundary of the σ_i minus the τ_i, and is topologically equivalent to S^{d-1}.

It can be shown,[3] and it is very easy to check, that any element of $K \in H^d(0)$ is the boundary of a simple $(d+1)$-tree T, and that K determines uniquely the simple $(d+1)$-tree T for $d \geqslant 2$.

Note that any stacked sphere $K \in H^4(0)$ can be mapped into a tree graph (see also Ref. 8). This mapping is defined in the following way: Let us consider the (unique) simple five-simple tree T associated with K, every five-simplex is mapped to a vertex and every four-dimensional face in common with two five simplexes is mapped to an edge which has end points at the two vertices which represent the two five simplexes. Since the map between K and T is one to one, we have a map from a stacked-sphere into a tree-graph whose number of links at every vertex can be at most 6 (since a five simplex has *six* faces). This map, from the stacked spheres $H^4(0)$ to the simple tree-graphs, is not one to one. The mathematical reason is that the previous construction maps every T to a tree-graph by an application that is the dual map restricted to the five- and four-dimensional simplexes of T in its domain and whose image is the tree graphs that are the one-dimensional skeleton of the dual complex. It is well known that the dual map is a one to one correspondence only if we take into account the simplexes of T of any dimension. It follows that our map is such that a simple tree-graph may correspond to many stacked spheres. To illustrate better this point we shall use a picture closer to the physical intuition. Let us consider a stacked sphere K and its simple tree T. Consider on T one of its ordered simplexes σ_j and let τ_j be the face that it shares with one of the simplexes $\sigma_1,\ldots,\sigma_t$ introduced previously. Repeating the same arguments used in the calculations of the inequivalent triangulations for baby universes,[4] we can cut T in τ_j into two parts, such that each part has two copies of τ_j as a part of its boundary. Since τ_j is a four-dimensional simplex it is easy to admit that we can glue these two different copies in $5 \cdot 4 \cdot 3$ ways to rebuild again a simple five tree T'. In general all the possible gluings will generate distinct triangulations and consequently distinct stacked spheres (if the two parts are highly symmetric triangulations some of the 60 ways of joining will not be distinct but, since for large N_4 asymmetrical triangulations will dominate, the number of cases in which this will happen will be trascurable). The corresponding tree graph associated with these configurations T' will always be the same since this operation has not modified the one skeleton of the tree T.

In statistical mechanics simple-tree structures correspond to ''branched-polymers.'' So we have that for $k_2 \geqslant k_2^{\max}$ the dominant configurations for the triangulations of S^4 are branched polymers. This fact was already observed in numerical simulations in four dimensions. In Ref. 9 a network of baby universes of minimum neck (minbu) and of minimum size (blips) without a mother universe was obtained. These have been interpreted as branched polymerlike structures.

IV. STACKED SPHERES AND BRANCHED POLYMERS

In this section we will establish a parallel between the standard mean field theory of branched polymers and the stacked spheres, in the sense that we will use the counting techniques of branched polymers to give an upper and lower estimation of the number of inequivalent stacked spheres. As we have stressed in Sec. III there exists a map between the stacked spheres and tree graphs and this map is *not* one to one in the sense that there are more stacked spheres than tree graphs. This means that the number of inequivalent stacked spheres with a fixed number N_5 of five-simplexes is bounded below by the corresponding number of tree graphs. Now we will study the statistical behavior of the tree graphs using the measure of simplicial quantum gravity and restricting it to the stacked spheres. In this analysis we will follow[10] in a way adapted to our case. The reader may refer to it for the details. A more mathematical analysis on the enumeration of inequivalent tree graphs is contained in Ref. 11.

First of all we notice that the boundary of a five-simplex has six four-simplexes and every edge of a tree graph corresponds to a cancellation of two four-simplexes in the corresponding boundary of the stacked sphere. Since in a tree graph with N_5 vertices there are N_5-1 edges, we have that the boundary of a stacked sphere, whose corresponding tree graph has N_5 points, is made by a number N_4 of four-simplexes,

$$N_4 = 4N_5 + 2. \tag{4.1}$$

From the condition for stacked spheres (1.5), we have that the Einstein–Hilbert action, for dynamical triangulations, restricted to the stacked spheres is

$$S=N_4(k_4-\tfrac{5}{2}k_2)-5k_2. \tag{4.2}$$

It follows that the Gibbs factor for the ensemble of tree graphs (branched polymers) corresponding to stacked spheres is

$$\exp(-(4N_5+2)(k_4-\tfrac{5}{2}k_2)+5k_2). \tag{4.3}$$

Following the same notation of Ref. 10, let $r^6(N_5)$ be the number of inequivalent rooted tree graphs with N_5 vertices and with at most six incident edges on each vertex, with one rooted vertex with one incident edge. $\xi^6(N_5)$ is the number of inequivalent tree graphs with N_5 vertices. By simple geometrical considerations, or from the asymptotic formula[10,11] for $r^6(N_5)$ and $\xi^6(N_5)$, we get the asymptotic formula

$$\xi^6(N_5)=\frac{1}{N_5}\,r^6(N_5+1). \tag{4.4}$$

Now let $R^6(k_4,k_2)$ be the partition function for rooted tree graphs with statistical weight (4.3), and $Z^6(k_4,k_2)$ be the partition function for unrooted tree graphs. We have

$$R^6(k_4,k_2)\equiv\sum_{N_5=2}^{\infty} e^{-(4(N_5-1)+2)(k_4-\frac{5}{2}k_2)+5k_2}r^6(N_5), \tag{4.5}$$

$$Z^6(k_4,k_2)\equiv\sum_{N_5=1}^{\infty} e^{-(4N_5+2)(k_4-\frac{5}{2}k_2)+5k_2}\xi^6(N_5). \tag{4.6}$$

Let us define

$$R^{*6}(\Delta k_4)\equiv(e^{-2(k_4-5)}R^6(k_4,k_2))=\sum_{N_5=2}^{\infty} e^{-4(N_5-1)\Delta k_4}r^6(N_5), \tag{4.7}$$

$$Z^{*6}(\Delta k_4)\equiv(e^{-2(k_4-5)}Z^6(k_4,k_2))=\sum_{N_5=1}^{\infty} e^{-4N_5\Delta k_4}\xi^6(N_5), \tag{4.8}$$

where $\Delta k_4\equiv k_4-\frac{5}{2}k_2$. It is easy to see, from (4.4), that

$$R^{*6}(\Delta k_4)=-\frac{1}{4}\,\frac{d}{d\Delta k_4}\,Z^{*6}(\Delta k_4). \tag{4.9}$$

From simplicial quantum gravity it is well known that the susceptibility $\chi(k_4,k_2)$ is given by

$$\chi(k_4,k_2)\asymp\frac{\partial^2}{\partial k_4^2}\,Z(k_4,k_2). \tag{4.10}$$

By analogous consideration on the graph it follows that

$$\chi(\Delta k_4)\asymp\frac{d^2}{d(\Delta k_4)^2}\asymp\frac{d}{d(\Delta k_4)}\,R^{*6}(\Delta k_4). \tag{4.11}$$

Now we will study the critical behavior of this system in order to obtain some information about the critical behavior of stacked spheres.

As easily verified the following identity, which is true for rooted tree graphs

$$r^6(N_5)=\sum_{\gamma=0}^{5}\ \frac{1}{\gamma!}\sum_{n_1,\ldots,n_\gamma;\Sigma_{i=1}^{\gamma}n_i=n-\gamma-2}\ \prod_{i=1}^{\gamma} r^6(n_i), \tag{4.12}$$

implies

$$R^{*6}(\Delta k_4)=e^{-4\Delta k_4}\left(\sum_{\gamma=0}^{5}\ \frac{1}{\gamma!}\,R^{*6}(\Delta k_4)\right). \tag{4.13}$$

Differentiating the previous equation we get a differential equation for $R^{*6}(\Delta k_4)$,

$$\frac{d}{d\Delta k_4}\,R^{*6}(\Delta k_4)=-\frac{1}{4}\,R^{*6}(\Delta k_4)\left[1+\frac{e^{-4\Delta k_4}}{5!}\,(R^{*6}(\Delta k_4))^5-R^{*6}(\Delta k_4)\right]^{-1}. \tag{4.14}$$

So $R^{*6}(\Delta k_4)$ shows a singularity when

$$R^{*6}((\Delta k_4)^c)=1+\frac{e^{-4\Delta k_4}}{5!}\,(R^{*6}((\Delta k_4))^c)^5. \tag{4.15}$$

This equation has only one real solution and since $(d/d\Delta k_4)R^{*6}|_{(\Delta k_4)^c}$ diverges, the inverse function is zero

$$\left.\frac{d(\Delta k_4)(R^{*6})}{dR^{*6}}\right|_{(\Delta k_4)^c}=0. \tag{4.16}$$

By the implicit function theorem $(\Delta k_4)(R^{*6})$ is analytic function at $R^{*6}((\Delta k_4)^c)$, which implies

$$\Delta k_4(R^{*6})-\Delta k_4^c=\frac{1}{2}\left.\frac{d^2(\Delta k_4)(R^{*6})}{d(R^{*6})^2}\right|_{R^{*6}((\Delta k_4)^c)}(R^{*6}(\Delta k_4)-R^{*6}(\Delta k_4{}^c))^2$$
$$+o((R^{*6}(\Delta k_4)-R^{*6}(\Delta k_4^c))^2). \tag{4.17}$$

So near $(\Delta k_4)^c$ we can write

$$R^{*6}(\Delta k_4)\asymp R^{*6}((\Delta k_4)^c)+C((\Delta k_4)-(\Delta k_4)^c)^{1/2}. \tag{4.18}$$

From (4.11) we get

$$\chi^6(\Delta k_4)\asymp((\Delta k_4)-(\Delta k_4)^c)^{-1/2}. \tag{4.19}$$

This means that the critical exponent of the susceptibility for the system of these tree graphs is $\gamma=\frac{1}{2}$.

The partition function that has been studied is a lower estimate of the partition function of the stacked spheres. Consider, now, a stacked sphere K and a face τ_j through which two five-simplexes are glued together (a link on the corresponding tree graph). We can glue a two four face of a stacked sphere in $5\cdot4\cdot3$ different ways, for large number of simplexes this will generate distinct configurations of stacked spheres whose associated tree-graph is always the same. Repeating the same argument for every j, $j=1,\ldots,N_5-1$ (that is to say for every link of the dual tree graph) we obtain a factor, $(5\cdot4\cdot3)^{N_5-1}$, that multiplied by $r^6(N_5)$ and $\xi^6(N_5)$, gives an upper bound on the number of, respectively, rooted and unrooted inequivalent stacked spheres. In other words

$$Z_{\text{tree}}(k_4,k_2)\leqslant Z_{\text{ss}}(k_4,k_2)\leqslant\tilde{Z}_{\text{tree}}(k_4,k_2) \tag{4.20}$$

where $Z_{\text{tree}}(k_4,k_2)$ is the partition function for tree graphs studied above, $Z_{\text{ss}}(k_4,k_2)$ is the partition function for stacked spheres, and $\tilde{Z}_{\text{tree}}(k_4,k_2)$ is the partition function for tree graphs with the additional weight defined above.

A similar analysis as for $Z_{\text{tree}}(k_4,k_2)$ shows that the critical line of $\tilde{Z}_{\text{tree}}(k_4,k_2)$ is, of course, a straight line parallel and above with respect to the k_4 axis to $Z_{\text{tree}}(k_4,k_2)$ one, and the susceptibility exponent is again $\gamma=\frac{1}{2}$.

Obviously the estimates (4.20) are true for the canonical partition functions too,

$$Z_{\text{tree}}(N_4,k_2)\leqslant Z_{\text{ss}}(N_4,k_2)\leqslant \tilde{Z}_{\text{tree}}(N_4,k_2), \tag{4.21}$$

and the previous calculations show that

$$Z_{\text{tree}}(N_4,k_2)\asymp N_4^{-5/2}e^{N_4(k_4-(5/2)k_2-t_4)} \tag{4.22}$$

and

$$\tilde{Z}_{\text{tree}}(N_4,k_2)\asymp N_4^{-5/2}e^{-N_4(k_4-(5/2)k_2+t_4-(1/4)\ \log 60)}, \tag{4.23}$$

where t_4 is a constant that may be calculated by Eq. (4.15). More easily from the table in Ref. 11 we get that $t_4\approx\frac{1}{4}\log 0.34$.

The circumstance that in the weak coupling region the partition function of quantum gravity, as found in an analytically way in Ref. 2, is strongly dominated by stacked spheres, allows us to write

$$Z_{\text{ss}}\asymp N_4^{\gamma_s-3}e^{k_4^c(k_2)}, \tag{4.24}$$

where γ_s is the susceptibility exponent.[12]

Thus the critical line of the stacked spheres is a straight line parallel and between the critical lines of the systems of the two branched polymers. This implies

$$k_4-\tfrac{5}{2}k_2+t_4\leqslant k_4^c(k_2)\leqslant k_4-\tfrac{5}{2}k_2+t_4-\tfrac{1}{4}\log 60. \tag{4.25}$$

Moreover from Eqs. (4.22) and (4.23) the one loop green functions[5] of the two model of branched polymers have, respectively, near their critical lines, the asymptotic behavior

$$G_{\text{tree}}(\Delta k_4)\asymp \text{cost}_1+(\Delta k_4-\Delta k_4^c)^{1/2},\quad \tilde{G}_{\text{tree}}(\Delta\tilde{k}_4)\asymp \text{cost}_2+(\Delta\tilde{k}_4-\Delta\tilde{k}_4^c)^{1/2}. \tag{4.26}$$

Equation (4.21) together with Eqs. (4.21) and (4.24) prove that the one loop function of the stacked spheres $G_{\text{ss}}(k_4,k_2)$ near the critical line has the asymptotic behavior

$$G_{\text{ss}}(k_4,k_2)\asymp \text{cost}_3+(k_4-k_4^c(k_2))^{1-\gamma_s}, \tag{4.27}$$

with $\gamma_s<1$ [the value $\gamma_s=1$ is not allowed because in this case the one loop green function of stacked spheres at the critical line will have a behavior like $\Sigma_{N_4=5}^{\infty}\,1/N$ that is divergent and then incompatible with the upper bound given by the second equation of (4.26)].

Motivated by physical considerations, we can use a well-known argument[5,13] in favor of the fact that the susceptibility exponent of the stacked spheres is $\gamma_s=\frac{1}{2}$. More precisely we will show that a model of proliferating baby universes, with the measure of quantum gravity restricted to stacked spheres, can be put in correspondence with the statistical system of stacked spheres.

Let us consider four-dimensional triangulations that are (boundary of the) stacked spheres in which there can be loops made by two three-dimensional simplexes. This is possible whenever the stacked spheres are pinched on a three simplex creating a bottle neck loop of two three-simplexes. These loops could be either the loops of a Green function or the minimum bottle neck of a baby universes. This class of triangulations, following the notation in literature, is called $\mathbf{T}_2$. The other class of triangulations is the stacked spheres in which the minimum loop length can be made by the boundary of a four-simplex that are five three-simplexes. We call this last class $\mathbf{T}_5$. In the two-dimensional theory, the introduction in $\mathbf{T}_2$ of two link loops (the two-dimensional analog of two three-simplex loop) corresponds in the matrix model ϕ^3 to consider Feynman diagrams with self-energy (cf. Refs. 5, 13). Since the minima loops of $\mathbf{T}_2$ and $\mathbf{T}_5$, for which they differ, are of

the order of lattice spacing we will expect that the two classes of triangulations, as a statistical mechanics system, coincide in the scaling limit, that is to say they belong to the same universality class.

Let us consider the minimum neck one loop function $G(\Delta k_4)$ in $\mathbf{T}_2$. In every triangulation of $\mathbf{T}_2$ we can cut out the maximal size baby universe of minimum neck and close the two three-simplex loop. We will obtain again a triangulation that belongs to $\mathbf{T}_2$. In this way we will obtain all the triangulations of $\mathbf{T}_2$ from the triangulations of the stacked spheres $\mathbf{T}_5$ considering that for each three-simplex either leave them in their actual form or we can open the triangulation to create a two three-simplex loop and gluing on it a whole one loop universe $G(\Delta k_4)$. We note that in the triangulations of $\mathbf{T}_5$ $\bar{N}_3=5/2\bar{N}_4$ (Dehn–Sommerville). Calling the one loop function of $\mathbf{T}_5$ $\bar{G}(\overline{\Delta k_4})$, the above considerations lead to the identity

$$G(\Delta k_4)=\sum_{\mathbf{T}\in\mathbf{T}_5} e^{-\bar{N}_4\Delta k_4}(1+G(\Delta k_4))^{\bar{N}_3}=\sum_{\mathbf{T}\in\mathbf{T}_5} e^{-\bar{N}_4\overline{\Delta k_4}}=\bar{G}(\overline{\Delta k_4}), \tag{4.28}$$

in which $\Delta k_4=k_4-5/2k_2$ comes out from restricting the action of quantum gravity to stacked spheres (4.2) and where we have defined

$$\overline{\Delta k_4}=\Delta k_4-\tfrac{5}{2}\log(1+G(\Delta k_4)). \tag{4.29}$$

By (4.28) we can also write Eq. (4.29) as

$$\Delta k_4=\overline{\Delta k_4}+\tfrac{5}{2}\log(1+\bar{G}(\overline{\Delta k_4})). \tag{4.30}$$

By universality and the estimates (4.21) it follows that near the critical point we have $G(\Delta k_4)\asymp \mathrm{cost}+(\Delta k_4-\Delta k_4^c)^{1-\gamma_s}$ with $\gamma_s<1$.

The susceptibility functions of $\mathbf{T}_2$ and $\mathbf{T}_5$ by (4.28) are

$$\chi(\Delta k_4)\asymp -\frac{d}{d\Delta k_4}G(\Delta k_4),\quad \bar{\chi}(\overline{\Delta k_4})\asymp -\frac{d}{d\overline{\Delta k_4}}\bar{G}(\overline{\Delta k_4}). \tag{4.31}$$

From (4.30) we have

$$\frac{d(\Delta k_4)}{d(\overline{\Delta k_4})}=1-\frac{5}{2}\frac{\bar{\chi}(\overline{\Delta k_4})}{(1+\bar{G}(\overline{\Delta k_4})}. \tag{4.32}$$

If we calculate di derivative with respect to Δk_4 of the one loop function $G(\Delta k_4)$ and use the previous equation we get

$$\chi(\Delta k_4)=\frac{\bar{\chi}(\overline{\Delta k_4})}{1-\frac{5}{2}\frac{\bar{\chi}(\overline{\Delta k_4})}{(1+\bar{G}(\overline{\Delta k_4})}}. \tag{4.33}$$

Now it is clear that $\chi(\Delta k_4)\mapsto+\infty$ for $\Delta k_4\mapsto(\Delta k_4)^c$ and with the same critical exponent $\bar{\chi}(\overline{\Delta k_4})\mapsto+\infty$ for $\overline{\Delta k_4}\mapsto(\overline{\Delta k_4})^c$. From Eq. (4.33) when $\chi(\Delta k_4)\mapsto+\infty$ we have that $\bar{\chi}(\overline{\Delta k_4})\mapsto\frac{2}{5}(1+\bar{G}(\overline{\Delta k_4}(\Delta k_4^c)))<+\infty$, then the system $\mathbf{T}_5$ is above his critical line, i.e., $\overline{\Delta k_4}(\Delta k_4^c)>\overline{\Delta k_4}^c$. These facts imply that at $\overline{\Delta k_4}(\Delta k_4^c)d(\Delta k_4)/d(\overline{\Delta k_4})=0$ and around it $\bar{\chi}(\overline{\Delta k_4})/(1+G(\overline{\Delta k_4}))$ is a decreasing monotonic function of Eq. (4.32) because $\chi(\Delta k_4)\mapsto+\infty$, we can expand Eqs. (4.30) and (4.33) around $\overline{\Delta k_4}(\Delta k_4^c)$

$$\Delta k_4-\Delta k_4^c=\mathrm{cost}(\overline{\Delta k_4}-\overline{\Delta k_4}(\Delta k_4^c))^2. \tag{4.34}$$

$$\chi(\Delta k_4)\asymp\frac{1}{\overline{\Delta k_4}-\overline{\Delta k_4}(\Delta k_4^c)}\asymp\frac{1}{\sqrt{\Delta k_4-\Delta k_4^c}}. \tag{4.35}$$

The second asymptotic equality of the last equation implies

$$\gamma_s = \tfrac{1}{2}. \tag{4.36}$$

The dominance of stacked spheres in the weak phase allows us to fix the parameter τ in the partition function of quantum gravity in the weak coupling regime:

$$\tau - \tfrac{11}{2} = -\tfrac{5}{2} \Rightarrow \tau = 3. \tag{4.37}$$

ACKNOWLEDGMENTS

It is a pleasure to thank J. Ambjørn, M. Carfora, M. Caselle, A. D'Adda, and D. Gabrielli for interesting discussions. In particular G. Esposito and C. Stornaiolo for careful reading of the manuscript.

[1] W. J. Tutte, ''A census of planar triangulations,'' Can. J. Math. **14**, 21–38 (1962).
[2] J. Ambjørn, M. Carfora, and A. Marzuoli, ''The geometry of dynamical triangulations,'' Lect. Notes Phys. **50**, (1997).
[3] D. Walkup, ''The lower bound conjecture for 3- and 4-manifolds,'' Acta Math. **125**, 75–107 (1970).
[4] S. Jain and S. D. Mathur, ''World-sheet geometry and baby universes in 2D quantum gravity,'' Phys. Lett. B **286**, 239–246 (1992).
[5] J. Ambjørn, *Quantization of Geometry*, in Les Houches 1994 Session LXII (North-Holland, Amsterdam, 1996), pp. 77–193 (hep-lat/9412006).
[6] T. Regge, ''General relativity without coordinates,'' Nuovo Cimento **19**, 558–571 (1961).
[7] B. V. de Bakker, *Simplicial Quantum Gravity*, Ph.D. Thesis (hep-lat/9508006).
[8] D. Gabrielli, ''Polymeric phase of simplicial quantum gravity,'' Phys. Lett. B **421**, 79–85 (1998) (hep-lat/9710055).
[9] J. Ambjørn, J. Jurkiewicz, and Y. Watabiki, ''Dynamical triangulations, a gateway to quantum gravity,'' J. Math. Phys. **36**, 6299–6339 (1995).
[10] J. Frohlich, *Mathematical Aspects of Physics of Disordered Systems*, in Les Houches 1984 Session XLIII (North-Holland, Amsterdam, 1986), pp. 727–893.
[11] R. Otter, ''The numbers of trees,'' Ann. Math. **49**, 583–599 (1948).
[12] J. Ambjørn and J. Jurkiewiez, ''Scaling in four dimensional quantum gravity,'' Nucl. Phys. B **451**, 643–676 (1995).
[13] J. Ambjørn, B. Durhuus, and T. Jonsson, *Quantum Geometry* (Cambridge University Press, Cambridge, 1997)

Is there a general area theorem for black holes?

Domenico Giulini[a)]
Institut für theoretische Physik, Universität Zürich,
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

(Received 15 July 1998; accepted for publication 3 August 1998)

The general validity of the area law for black holes is still an open problem. We first show in detail how to complete the usually incompletely stated textbook proofs under the assumption of piecewise C^2-smoothness for the surface of the black hole. Then we prove that a black hole surface necessarily contains points where it is not C^1 (called ''cusps'') at any time before caustics of the horizon generators show up, like, e.g., in merging processes. This implies that caustics never disappear in the past and that black holes without initial cusps will never develop such. Hence black holes which will undergo any nontrivial processes anywhere in the future will always show cusps. Although this does not yet imply a strict incompatibility with piecewise C^2 structures, it indicates that the latter are likely to be physically unnatural. We conclude by calling for a purely measure theoretic proof of the area theorem. © *1998 American Institute of Physics.* [S0022-2488(98)01612-0]

I. INTRODUCTION

It seems to be widely accepted as fact that the surface area of a black hole cannot decrease with time. However, the proofs offered in standard textbooks, like Refs. 1–3, are basically content with the remark that this law follows from the nonconvergence of the generators of the future event horizon. It would indeed follow from this remark and some elementary differential geometric considerations if the horizon were a sufficiently smooth submanifold. Mathematically there is absolutely no reason why this should be true in general,[4] which means that extra assumptions must be invoked (implicitly) in the textbooks arguments. However, not much precise information seems to exist in the literature concerning these extra assumptions. Perhaps the clearest statement is given in Ref. 4, where the authors mention that in the textbook proofs of the area theorem ''something close to C^2 differentiability 'almost everywhere' of the event horizon seems to have been assumed.'' The textbooks themselves do not mention any such condition. Below we show explicitly how to complete the textbook argumentation under the assumption of piecewise C^2-smoothness. But this clearly does not imply its necessity.

General considerations only prove the horizon to be locally Lipschitz continuous (denoted by C^{1-}).[1] Mathematically this implies (pointwise) differentiability almost everywhere (with respect to the Lebesgue measures defined by the charts, see Ref. 5 Theorem 3.1.6), but it still allows the points of nondifferentiability to be densely distributed.[4] Hence horizons exist which are nowhere C^1. Given these mathematical facts, it is of interest to learn what *physical* conditions imply a breakdown of C^1 differentiability. We prove that a black hole whose surface is C^1 at one time can never merge with other black holes and, more generally, never encounter new null generators for its future event horizon. In other words, an initial C^1-condition basically rules out any interesting physical process to happen in the future. Hence the only dynamically interesting holes for which the area law is actually proven are those whose surfaces are piecewise C^2 but not C^1. Presumably this class does not contain many, if any, physically realistic members. Cusps on the surface of colliding black holes can be clearly seen in numerical studies,[6] but no analytic proof of their general existence seems to have been given so far.

Given the widely believed connection of the area law with thermodynamic properties of black holes on one side, and the widely expressed hope that this connection may be of heuristic value in understanding certain aspects of quantum gravity on the other, it seems important to know the most general conditions under which the area law is valid.

a)Electronic mail: giulini@physik.unizh.ch

 © 1998 American Institute of Physics

II. NOTATION, FACTS, AND ASSUMPTIONS

We assume space–time (M,g) to be strongly asymptotically predictable (in the sense of Ref. 3) and globally hyperbolic. (It would be sufficient to restrict to a globally hyperbolic portion, as in Thm. 12.2.6 of Ref. 3.) $\mathcal{I}^+$ (scri-plus) denotes future null infinity, $J^-(\mathcal{I}^+)$ its causal past and $B:=M-J^-(\mathcal{I}^+)$ the black hole region. Its boundary, $\partial B=:H$, is the future event horizon. H is a closed, imbedded, achronal three-dimensional C^{1-}-submanifold of M (Proposition 6.3.1 in Ref. 1). H is generated by null geodesics without future endpoints. Past endpoints occur only where null geodesics—necessarily coming from $J^-(\mathcal{I}^+)$—join onto H. Such points are called ''caustics.'' Only at a caustic can a point of the horizon be intersected by more than one generator, and *all* generators that intersect a caustic enter the horizon at this point. Once a null geodesic has joined onto H it will never encounter a caustic again (i.e., not intersect another generator) and never leave H. See Box 34.1 in Ref. 2 for a lucid discussion and partial proofs of these statements. Hence there are two different processes through which the area of a black hole may increase: First, new generators can join the horizon and, second, the already existing generators can mutually diverge.

Caustic points where n (possibly infinite) new generators join in are said to be of multiplicity n. In Ref. 7 it is proven that H is not differentiable at p iff p is a caustic of multiplicity $n\geqslant 2$, and that caustics of multiplicity 1 are contained in the closure of those of higher multiplicity. In particular, H cannot be of class C^1 at any caustic point. Points where H is not C^1 will be called ''cusps.'' By definition, being C^1 at p implies that H is differentiable in a whole neighborhood U of p. Conversely, it was shown that differentiability in some open neighborhood U of p implies that H is C^1 in U (Ref. 7, Prop. 3.3), so that the set of points where H is C^1 is open. It follows that being C^1 at p is in fact equivalent to being differentiable in some neighborhood of p. Hence the set of cusps is the closure of the set of points where H is nondifferentiable (caustics of multiplicity $\geqslant 2$) and hence also the closure of the set of all caustics.

Let Σ be a suitably smooth (usually C^2) Cauchy surface, then $\mathcal{B}:=B\cap\Sigma$ is called the black-hole region at time Σ and $\mathcal{H}:=H\cap\Sigma=\partial\mathcal{B}$ the (future-event-) horizon at time Σ. A connected component $\mathcal{B}_i$ of $\mathcal{B}$ is called a black hole at time Σ. Its surface is $\mathcal{H}_i=\partial B_i$, which is a two-dimensional, imbedded C^{1-}-submanifold of Σ. We have seen that in general $\mathcal{H}$ may contain all kinds of singularities which would render standard differential geometric methods inapplicable. Adding the hypothesis of piecewise C^2-smoothness circumvents this problem.

By $\exp{:}TM\to M$ we denote the exponential map. Recall that $\exp_p(v):=\gamma(1)$, where γ is the unique geodesic with initial conditions $\gamma(0)=p\in M$ and $\dot{\gamma}(0)=v\in T_p(M)$. For each p it is well defined for v in some open neighborhood of $0\in T_p(M)$. One has $\gamma(t)=\exp_p(tv)$. We shall assume the Lorentzian metric g of M to be C^2, hence the connection (i.e., the Christoffel symbols) is C^1 and, therefore, the map exp is also C^1. The last assertion is e.g., proven in Ref. 8.

III. LOCAL FORMULATION OF THE AREA LAW

We consider two C^2 Cauchy surfaces with Σ' to the future of Σ. The corresponding blackhole regions and surfaces are denoted as above, with a prime distinguishing those on Σ'. We make the *assumption* that $\mathcal{H}$ is piecewise C^2, i.e., each connected component $\mathcal{H}_i$ of $\mathcal{H}$ is the union of open subsets $\mathcal{H}_i^k$ which are C^2 submanifolds of M and whose two-dimensional measure exhaust that of $\mathcal{H}_i$: $\mu(\mathcal{H}_i-\cup_k\mathcal{H}_i^k)=0$, where μ is the measure on $\mathcal{H}$ induced from the metric g.

For each point $p\in\mathcal{H}_i^k$ there is a unique future- and outward-pointing null direction perpendicular to $\mathcal{H}_i^k$, which we generate by some future directed $l(p)\in T_p(M)$. We can choose a C^1-field $p\mapsto l(p)$ of such vectors over $\mathcal{H}_i^k$. The geodesics $\gamma_p: t\mapsto\gamma_p(t):=\exp_p(tl(p))$ are generators of H without future endpoint. Therefore, each γ_p cuts Σ' in a unique point $p'\in\mathcal{H}'$ at a unique parameter value $t=\tau(p)$. By appropriately choosing the affine parametrizations of γ_p as p varies over $\mathcal{H}_i^k$ we can arrange the map τ to be also C^1. Hence $p\mapsto m(p):=\tau(p)l(p)$ is a null vector field of class C^1 over $\mathcal{H}_i^k$. We can now define the map

$$\Phi_i^k:\mathcal{H}_i^k\to\mathcal{H}',\quad p\mapsto\Phi_i^k(p):=\exp_p(m(p)),\tag{1}$$

which satisfies the following.

Lemma 1: Φ_i^k is (i) C^1, (ii) injective, (iii) not measure-decreasing.

(i) follows from the fact that the functions m and exp are C^1. Injectivity must hold, since otherwise some of the generators of H through $\mathcal{H}_i^k$ would cross in the future. By being not measure-decreasing we mean the following: Let μ and μ' be the measures on $\mathcal{H}$ and $\mathcal{H}'$ induced by the space–time metric g. Then $\mu[U] \leqslant \mu'[\Phi_i^k(U)]$ for *each* measurable $U \subset \mathcal{H}_i^k$. Assuming the weak energy condition, this is a consequence of the *nowhere* negative divergence for the future geodesic congruence $p \mapsto \gamma_p$ (Lemma 9.2.2 in Ref. 1), as we will now show.

Proof of (iii): Set $H_i^k := \cup_{p,t} \exp_p(tl(p))$, $\forall p \in \mathcal{H}_i^k$ and $\forall t \in R_+$, which is a C^1-submanifold of M (the future of $\mathcal{H}_i^k$ in H). Let l be the unique (up to a constant scale) future directed null geodesic (i.e., $\boldsymbol{\nabla}_l l = 0$) vector field on H_i^k parallel to the generators. Then $0 \leqslant \nabla_\mu l^\mu = \pi_\nu^\mu \nabla_\mu l^\nu$, where π denotes the map given by the g-orthogonal projection $T(M)|_{H_i^k} \to T(H_i^k)$, followed by the quotient map $T(H_i^k) \to T(H_i^k)/\mathrm{span}\{l\}$. Note that tangent spaces of C^1-cross sections of H_i^k at the point p are naturally identified with $T_p(H_i^k)/\mathrm{span}\{l(p)\}$. Since $\pi_\nu^\mu l^\nu = 0$, we also have $\pi_\nu^\mu \nabla_\mu k^\nu \geqslant 0$ for $k = \lambda l$ and any C^1-function λ: $H_i^k \to \mathbf{R}_+$. Hence this inequality is valid for any future pointing C^1-vector-field k on H_i^k parallel to the generators. Given that, let $t \mapsto \phi_t$ be the flow of k and $A(t) := \mu_t[\phi_t(U)] := \int_{\phi_t(U)} d\mu_t$, then $\dot{A}(t) = \int_{\phi_t(U)} \pi_\nu^\mu(t) \nabla_\mu k^\nu(t) d\mu_t \geqslant 0$, where $\pi(t)$ projects onto $T(\phi_t(\mathcal{H}_i^k))$, $k(t) = d/(dt')|_{t'=t} \phi_{t'}$ and $\mu_t =$ measure on $\phi_t(\mathcal{H}_i^k)$. Now choose k such that $\phi_{t=1} = \Phi_i^k$, then $\mu'[\Phi_i^k(U)] - \mu[U] = \int_0^1 dt \dot{A}(t) \geqslant 0$. □

Part (iii) of Lemma 1 is the local version of the area law. By turning it into a global statement about areas one usually abandons some of its information. The most trivial global implication is that the total sum of areas cannot decrease. A more refined version is as follows: Recall that black holes cannot bifurcate in the future (Proposition 9.2.5 of Ref. 1). Hence all surface elements $\mathcal{H}_i^k$ of the ith black hole at time Σ are mapped via Φ_i^k into the surface $\mathcal{H}'_i$ of a *single* black hole at time Σ'. We call $\mathcal{H}'_i$ (i.e., the connected component of $\Sigma' \cap H$ into which $\mathcal{H}_i$ is mapped) the development of $\mathcal{H}_i$ at time Σ'. Lemma 1 now implies that its area cannot be less than that of $\mathcal{H}_i$. The nonbifurcation result implies that if the number N' of black holes at time Σ' is bigger than the number N at time Σ, then there is an intermediate formation of $K \geqslant N' - N$ new black holes. That these black holes are ''new,'' i.e., not present at time Σ, means that all generators of H which intersect $\mathcal{H}'_1 \cup \cdots \cup \mathcal{H}'_K$ must have past endpoints somewhere between Σ and Σ'. A black hole at time Σ' which is smaller than any black hole at time Σ must also be new in this sense. Hence one way to express an area law would be as follows:

Assertion (Area Law): Consider two Cauchy surfaces, Σ and Σ', with Σ' to the future of Σ. Then the area of the development $\mathcal{H}'_i$ of any $\mathcal{H}_i$ cannot be less than that of $\mathcal{H}_i$. In particular, black holes at time Σ' whose area is smaller than that of any black at time Σ must have been formed in the meantime.

Presently we do not have a proof that this statement is true in general. But since it is a statement about measures, we suggest that it should be possible to give a proof without invoking fiducial (and probably irrelevant) differentiability assumptions.

IV. CONSEQUENCES

The previous discussion allows one to show that black holes whose surface is C^1 at one instant cannot undergo any nontrivial change in the future, like merging processes or any other process involving the incorporation of new generators. This we shall now explain in more detail. Let $\mathcal{H}_1$ be the surface a black hole at time Σ. We assume $\mathcal{H}_1$ to be a compact, two-dimensional C^1-submanifold. As before, Σ' is to the future of Σ and $\mathcal{H}' = H \cap \Sigma'$. We can construct a map $\Phi_1 : \mathcal{H}_1 \to \mathcal{H}'$, just analogous to the construction of Φ_i^k above, but now defined on *all* of $\mathcal{H}_1$. The C^1-condition on $\mathcal{H}_1$ now implies that Φ_1 is C^0. Φ_1 is also injective for the same reason as given for Φ_i^k. Since $\mathcal{H}_1$ is connected, its image under Φ_1 is also connected. Let $\mathcal{H}'_1 \subset \mathcal{H}'$ be the connected part containing the image of Φ_1. We show

Lemma 2: $\Phi_1 : \mathcal{H}_1 \to \mathcal{H}'_1$ is a homeomorphism.

Proof: Φ_1 is a closed map, because if $U \subset \mathcal{H}_1$ is closed $\Rightarrow U$ is compact (since $\mathcal{H}_1$ is compact) $\Rightarrow U' := \Phi_1(U)$ is compact (since Φ_1 is continuous) $\Rightarrow U'$ is closed (since $\mathcal{H}'_1$ is Hausdorff). From this it follows that Φ_1 is a homeomorphism onto its image. But Φ_1 is also open; this follows directly from Brouwer's theorem on the invariance-of-domain, which states that any continuous injective map from an open $X \subset R^n$ into R^n is open (Proposition 7.4 in Ref. 9). This clearly

generalizes to manifolds. Hence the image $\Phi_1(\mathcal{H}_1)\subset\mathcal{H}_1'$ is open, closed and connected, and hence all of $\mathcal{H}_1'$. □

Surjectivity of Φ_1 implies that all generators of H which intersect $\mathcal{H}_1'$ also intersect $\mathcal{H}_1$. Hence nowhere in its future will $\mathcal{H}_1$ be joined by new generators. Similar to the definition of H_i^k above, let $H_1\subset H$ denote the future of $\mathcal{H}_1$ in H; then it follows that H_1 is free of caustics and therefore C^1. Hence we have

Proposition 1: Let the surface of a black hole at time Σ be without cusps. Then this black hole will never encounter cusps to the future of Σ, in particular, it will not merge with other black holes.

Another equivalent formulation, emphasizing that cusps will not die out in the past, is as follows:

Proposition 2: At no time to the past of a cusp on H will the surface of a black hole be without cusps.

It is sometimes suggested that caustics just exist for some finite time interval during which the actual processes take place, like collision and coalescence of black holes or the infall of matter through the horizon (see e.g., Ref. 2, 34.5). Proposition 2 shows that this is not quite the right picture.

[1] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time* (Cambridge University Press, Cambridge, 1973).
[2] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[3] R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago and London, 1984).
[4] P. T. Chruściel and G. J. Galloway, ''Horizons non-differentiable on a dense set,'' Commun. Math. Phys. **193**, 449–470 (1998).
[5] H. Federer, *Geometric measure theory* (Springer Verlag, New York, 1969).
[6] J. Libson, J. Massó, E. Seidel, W.-M. Suen, and P. Walker, ''Event horizons in numerical relativity: Methods and tests,'' Phys. Rev. D **53**, 4335–4350 (1996).
[7] J. K. Beem and A. Królak, ''Cauchy horizon endpoints and differentiability,'' gr-qc/9709046.
[8] S. Lang, *Differential Manifolds* (Springer-Verlag, Berlin, 1985).
[9] A. Dold, *Lectures on algebraic topology* (Springer Verlag, Berlin, 1972).

Time-independent bounds on solutions with shocks to the relativistic Euler equations on conformally flat space–times with curvature singularities

Jeffrey M. Groah
Department of Mathematics and Computing, La Sierra University, Riverside, California 92515

(Received 15 October 1997; accepted for publication 7 November 1997)

We consider the initial value problem of the relativistic Euler equations when the underlying space–time is not flat but conformally flat, and demonstrate existence of solutions with shocks, with time-independent bounds, in some cases from big bang till big crunch. Our theorem requires that the space–time metric satisfy certain bounds, but these do not constrain the curvature of space–time, and hence our theorem may guarantee uniform bounds on solutions until the formation of curvature singularities. © *1998 American Institute of Physics.*
[S0022-2488(98)04003-1]

I. INTRODUCTION

We study the global evolution of shock dynamics on space–times given a fixed conformally flat metric. We analyze the fluid motion using Glimm's method with operator splitting, a numerical method that accounts for the effects of the fluid in the first step, using Glimm's method, and the effects of the geometric sources in the second iteration of the method. Glimm's method with operator splitting is the only method of analysis that has been successfully shown to yield global estimates for the shock dynamics when sources are present.[1] In Ref. 2, we considered conformally flat space–times with trace-free energy-momentum tensor, and found a special compatibility between the gravitational sources and the fluid which allowed us to obtain solutions with uniform total variation bounds independent of the strength of the gravitational potential, and hence from big bang till big crunch. Here, we do not assume that the energy-momentum tensor is trace-free, and we find it necessary to assume certain bounds on the gravitational sources in order to obtain uniform bounds on the fluid solutions. Since these restrictions do not constrain the space–time curvature, they allow us to establish existence of uniformly bounded solutions on space–times having both a big bang and a big crunch singularity.

We study here the initial value problem for solutions with shocks to the relativistic Euler equation on a 4-dimensional space–time on which we assume a fixed background metric which is conformally flat, $g=e^{2\varphi}\eta$, where η is the Minkowski metric diag$(-1,1,1,1)$, where fluid motion occurs in one spatial coordinate direction and along the time coordinate, and with the equation of state for isothermal flow, or the extreme relativistic limit of free particles, $p=\sigma^2\rho$. Here, ρ is the fluid pressure, σ is the constant sound speed, and ρ is the mass-energy density of the fluid as measured in units of mass in a reference frame moving with the fluid particles. Let the space–time manifold M be given coordinates (x^0,x^1,x^2,x^3) and assume, without loss of generality, that $x^0=ct$ is the time coordinate, and $x^1=x$ is the spatial coordinate along which fluid moves.

We will show that, under these conditions, the relativistic Euler equations can be written in the form

$$\frac{1}{c}\frac{\partial}{\partial t}\left\{\overline{\rho}\left((\sigma^2+c^2)\,\frac{c^2}{c^2-v^2}-\sigma^2\right)\right\}+\frac{\partial}{\partial x}\left\{\overline{\rho}(\sigma^2+c^2)\,\frac{cv}{c^2-v^2}\right\} \tag{1}$$

$$=-\overline{\rho}\,\frac{1}{c}\,\frac{\partial\varphi}{\partial t}\,\{\sigma^2(n-1)-c^2\}, \tag{2}$$

 © 1998 American Institute of Physics

$$\frac{1}{c}\frac{\partial}{\partial t}\left\{\overline{\rho}(\sigma^2+c^2)\,\frac{cv}{c^2-v^2}\right\}+\frac{\partial}{\partial x}\left\{\overline{\rho}\left((\sigma^2+c^2)\,\frac{v^2}{c^2-v^2}+\sigma^2\right)\right\} \tag{3}$$

$$=\overline{\rho}\,\frac{\partial\varphi}{\partial x}\{\sigma^2(n-1)-c^2\}, \tag{4}$$

where $\overline{\rho}=\rho e_{n\varphi}$, c is the speed of light, and v is the particle velocity. Our main theorem establishes the existence of uniformly bounded weak solutions to (4) on conformally flat space–times when the gravitational potential φ satisfies certain bounds.

Theorem I.1: *Suppose a 4-dimensional space–time manifold is given a coordinate system* (x^0,x^1,x^2,x^3), *a conformally flat metric* $g=e^{2\varphi}\eta$ *of signature* $(-1,1,1,1)$ *where* φ *is a function of the space–time coordinates* x^0 *and* x^1, *and arbitrary initial data* $\rho_0(x^1)$, $v_0(x^1)$ *satisfying*

$$\mathrm{TV}(\ln(\rho_0))<\infty \tag{5}$$

and

$$\mathrm{TV}\left\{\ln\left(\frac{v_0(\cdot)+c}{v_0(\cdot)-c}\right)\right\}<\infty, \tag{6}$$

where $\mathrm{TV}(f)$ *denotes the total variation of the function* f. *Suppose that* $|\varphi_{,0}|+|\varphi_{,1}|\in L^1(R\times R^+)$, *and*

$$\int_0^\infty \mathrm{TV}_x(|\varphi_{,0}|+\varphi_{,1}|)dt\leqslant\infty. \tag{7}$$

Then there exists a bounded weak solution $\rho(x,t),v(x,t)$ *of system (4) satisfying:*

$$\mathrm{TV}\ \ln(\rho)\leqslant\mathrm{TV}\ \ln(\rho_0)+C \tag{8}$$

and

$$\mathrm{TV}\left\{\ln\left(\frac{v(\cdot,t)+c}{v(\cdot,t)-c}\right)\right\}\leqslant\mathrm{TV}\left\{\ln\left(\frac{v_0(\cdot)+c}{v_0(\cdot)-c}\right)\right\}+C, \tag{9}$$

where C *is a constant depending only on the metric* g. *Moreover, the solution is a limit of approximate solutions* $(\rho_{\Delta x},v_{\Delta x})$ *generated by a Glimm type operator splitting scheme, and converges a.e., and in* L^1_{loc} *at each time, uniformly on bounded subsets of* (x,t)*-space.*

Theorem I.1 generalizes the results in Ref. 2 to space–times without trace-free energy-momentum tensors, but on this more general class of spaces, the metric must satisfy the bound (7). The approximation scheme used here is necessarily different due to the different form of the lower order gravitational source terms, and is similar to an operator splitting scheme used by Luskin and Temple in Ref. 3 (see also Ref. 4). The operator splitting scheme consists of two parts which we call the Conservation Law Step and the ODE Step. Our proof of the Conservation Law Step bounds in Lemma IV.9 follows the flat Minkowski space case proved by Smoller and Temple in Ref. 5, and is based on the geometry of the shock-curves is the plane of Riemann invariants, first used by Nishida[6] in his solution of the classical case. Our assumption of the bounds (7) is necessary for our proof of the ODE Step bounds, and amounts to assuming that the gravitational source term adds a finite amount of wave strength. Assuming the spatial total variation of the source is bounded implies the source adds a finite amount of wave strength in each time interval, while assuming that the spatial total variation of the source is integrable implies that the total contribution to wave strength over all time will be integrable. The assumption that the gravitational source term is $L^1(R\times R^+)$ assures that the source diminishes in space and time, and is used in our proof of convergence.

The bounds on $\varphi_{,0}$ and $\varphi_{,1}$ are not restrictions on space–time curvature since curvature depends not only on these quantities but also on the second derivatives of the metric. Instead, the bounds (7) are restrictions on the divergence of the coordinate vector fields $\partial/\partial x^0$ and $\partial/\partial x^1$.

When the Einstein equation $G=\kappa T$ holds for a perfect fluid with the equation of state $p=\sigma^2\rho$, curvature is singular if and only if the mass-energy-momentum tensor is singular. Here, however, we do not require that the Einstein equation be satisfied, and as long as the integrability conditions (7) hold, solutions with uniform bounds to (4) may be found, even in the presence of curvature singularities such as occur in a big bang or a black hole. This theorem applies in the presence of these singularities but only when the integrability conditions (7) are satisfied.

For classical fluids, the equation of state $p=\sigma^2\rho$ is considered theoretically as the limiting case of a polytropic gas, $p=\sigma^2\rho^\gamma$, as $\gamma\to 1^+$. In relativistic physics, the case $p=\sigma^2\rho$ is the equation of state for a gas in which radiation energy dominates, or the extreme relativistic limit of free particles (see Ref. 7 p. 12 or Ref. 8 p. 51). It arises in connection with the big bang, massless neutrinos, ultrarelativistic electron-positron pairs, Neutron stars and the early stages of stellar formation. See also Ref. 5.

The organization of this paper is as follows. In Sec. II, the equations we study are derived. The necessity of coordinate vector fields in the construction of a grid and the Courant–Friedrichs–Lewy condition is described in Sec. III, and then the approximate solutions are defined in Sec. III A, and the operator splitting scheme described in detail. Wave strengths and the potential are then defined. In Sec. IV, the geometry of the shock curves is discussed, followed by the solution of the Riemann problem, and then bounds are demonstrated to exist on the growth of the total wave strength on the conservation law step of the operator splitting scheme. In Sec. V, bounds are demonstrated to exist on the growth of the total wave strength on the ODE step of the operator splitting scheme. Finally, in Sec. VI, we show that a subsequence of the approximate solutions converges to a solution.

II. DERIVATION OF THE EQUATIONS

In this section the equations are put into the form needed for the operator splitting scheme. Then we discuss a restriction on the metric which arises in connection with construction of a grid on which to carry out a Glimm-type approximation scheme. The approximate solutions are next defined and the operator splitting scheme described in detail. Wave strengths and the potential are then defined.

One spatial dimensional fluid motion requires that the initial data be constructed appropriately, and that the space–time manifold have sufficient symmetry so that the curvature of space–time will not cause fluid to flow in directions other than those permitted. The assumption that fluid moves along the x^0 and x^1 coordinates directly affects our choice of metric. The symmetry of fluid motion requires that the space–time manifold have sufficient symmetry; otherwise, gravitation will cause fluid motion in the x^2 and x^3 coordinate directions. The following lemma gives sufficient conditions on the metric so that fluid will move only in the x^0 and x^1 coordinate directions.

Lemma II.1: Suppose the space–time metric has the block structure,

$$\begin{pmatrix} \tilde{g} & 0 \\ 0 & \tilde{g} \end{pmatrix}, \tag{10}$$

where $\hat{g}$ is a 2×2 block. Given the equation of state $p(\rho)$, when $g_{ij}, i,j\leqslant 1$ and ρ are functions of x^0 and x^1 alone, and $u^2\equiv u^3\equiv 0$, then the Eqs. (4) $j\geqslant 2$ are identically satisfied, and hence the Eqs. (4) $j\leqslant 1$ may be solved independently. Moreover, the components of acceleration in the x^2 and x^3 coordinates directions, as well as the pressure gradient in these directions, is zero.

Proof: That the Eqs. (4) $j\leqslant 1$ may be solved independently of the Eqs. (4) $j\geqslant 2$ may be readily checked. The particle acceleration is given by $\nabla_U U$. That this acceleration should vanish in the coordinate directions $\partial/\partial x^k$, $k=2,3$ is expressed by $0=g(\nabla_U U,\partial/\partial x^k)$, $k=2,3$ or in coordinates by

$$u^i u^j_{,i} g_{jk} + u^i u^j \Gamma^l_{ij} g_{lk} = 0. \tag{11}$$

To analyze the second term, let i, $j\leqslant 1$, $k\geqslant 2$. Since the metric is block diagonal,

$$\Gamma^l_{ij} g_{lk} = \frac{g^{lm}}{2}(-g_{ij,m}+g_{mi,j}+g_{jm,i})g_{lk} = \frac{-1}{2} g_{ij,k}. \tag{12}$$

Since the metric components $g_{ij}, i,j \leqslant 1$ are functions of x^0 and x^1 alone,

$$\Gamma^l_{ij} g_{lk} = 0. \tag{13}$$

This implies that if the solution satisfies the equation

$$u^i u^j_{,i} g_{jk} = 0, \tag{14}$$

then the fluid is not accelerated in the x^2 and x^3 coordinate directions. Since $k \geqslant 2$, and $u^k \equiv 0$, $k \geqslant 2$, (14) is satisfied. Thus the acceleration in the x^2 and x^3 coordinate directions is zero.

We now analyze the pressure gradient. To this end, consider the relativistic Euler equations, which can be written in the form

$$0 = \mathrm{div}(T) = U(p+\rho c^2)U + (p+\rho c^2)\nabla_U U + (p+\rho c^2)\mathrm{div}(U)U + \mathrm{grad}(p).$$

Since U is a unit vector (and since the metric is parallel), $0 = U(g(U,U)) = 2g(\nabla_U U, U)$, so $\nabla_U U \perp U$. From this we deduce the two equations

(1) $c^2 U(\rho) = -(p+\rho c^2)\mathrm{div}(U)$ (energy equation);

(2) $(p+\rho c^2)\nabla_U U = -\mathrm{grad}_\perp p$ (force equation),

where $\mathrm{grad}_\perp(p)$ is the component of $\mathrm{grad}(p)$ orthogonal to U, and is known as the *spatial pressure gradient* (see Ref. 9 p. 339.) If $g(\nabla_U U, \partial/\partial x^k) = 0$ for $k = 2,3$, the force equation implies $0 = g(\mathrm{grad}_\perp p, \partial/\partial x^k)$, $k \geqslant 2$. Now, $\mathrm{grad}(p) = -U(p)U + \mathrm{grad}_\perp p$, so it must be that $0 = p_{,k}$, $k \geqslant 2$. Given the equation of state $p(\rho)$, we deduce that since $\rho_{,k} = 0$ when $k \geqslant 2$, $0 = g(\mathrm{grad}_\perp p, \partial/\partial x^k)$, for $k \geqslant 2$. □

For conformally flat space–times, the condition that the functions $g_{ij}, i,j \leqslant 1$ are functions of x^0 and x^1 alone implies φ is a function of x^0 and x^1 alone.

Let T be the energy-momentum tensor for a fluid defined on M. Conservation of mass-energy and momentum is expressed by the equation

$$\mathrm{div}\ T = 0. \tag{15}$$

We will use the following proposition in the first step in rewriting the Euler equations.

Proposition II.0.1: Suppose T is a symmetric (2,0) tensor, $\overline{T} = e^{(n+2)\varphi} T$ *where n is the dimension of the manifold, and we have two metrics g and* $\overline{g}$*, which are conformally related,* $\overline{g} = e^{2\varphi} g$*. Then*

$$e^{(n+2)\varphi}\ \mathrm{div}_{\overline{g}} T = \mathrm{div}_g \overline{T} - e^{(n+2)\varphi}\ \mathrm{grad}_g\ \varphi \mathrm{tr}_g(T), \tag{16}$$

and therefore $\mathrm{div}_{\overline{g}} T = 0$ *if and only if*

$$\mathrm{div}_g\ \overline{T} = e^{(n+2)\varphi}\ \mathrm{grad}_g(\varphi)\mathrm{tr}_g(T). \tag{17}$$

Moreover, T is a weak solution of $\mathrm{div}_{\overline{g}} T = 0$ *if and only if it is a weak solution to (17).*

Proof: When $\overline{g} = e^{2\varphi} g$, the connection $\overline{\nabla}$ with respect to the metric $\overline{g}$ is related to the connection ∇ with respect to the metric g by the equation

$$\overline{\nabla}_X Y = \nabla_X Y + X(\varphi)Y + Y(\varphi)X - g(X,Y)\mathrm{grad}_g(\varphi), \tag{18}$$

(see Ref. 10). Therefore,

$$\overline{\Gamma}^j_{ik} = \Gamma^j_{ik} + \varphi_{,k}\delta^j_i + \varphi_{,i}\delta^j_k - \varphi_{,l} g_{ik} g^{jl}. \tag{19}$$

In addition, we use the following identity:

$$\overline{\Gamma}^i_{ik} = \tfrac{1}{2}\ln(-\overline{g})_{,k}, \tag{20}$$

(see Refs. 11, 12) whereby

$$\overline{\Gamma}^i_{ik} = \tfrac{1}{2}\ln(-e^{2n\varphi}g)_{,k} = \tfrac{1}{2}\{\ln(-g)+2n\varphi\}_{,k} = \Gamma^i_{ik} + n\varphi_{,k}. \tag{21}$$

Using (19) and (20) to compute the covariant divergence of T with respect to the metric $\overline{g}$ we obtain

$$\begin{aligned}\operatorname{div}_{\overline{g}} T^j &= T^{ij}{}_{,i} + \overline{\Gamma}^i_{ik}T^{kj} + \widetilde{\Gamma}^j_{ik}T^{ik}\\ &= T^{ij}{}_{,i} + (\Gamma^i_{ik} + n\varphi_{,k})T^{kj} + (\Gamma^j_{ik} + \varphi_{,k}\delta^j_i + \varphi_{,i}\delta^j_k - \varphi_{,l}g^{lj}g_{ik})T^{ik}\\ &= \operatorname{div}_g T^j + (n+2)\varphi_{,k}T^{kj} - \varphi_{,k}g^{kj}\ \operatorname{tr}_g(T),\end{aligned}$$

where by $\operatorname{div}_{\overline{g}}T^j$ we mean the jth component of $\operatorname{div}_{\overline{g}}T$. Setting $\overline{T} = e^{(n+2)\varphi}T$ and using the last equation we find

$$\begin{aligned}\operatorname{div}_g \overline{T}^j = \overline{T}^{ij}{}_{,i} &= (n+2)\varphi_{,i}e^{(n+2)\varphi}T^{ij} + e^{(n+2)\varphi}T^{ij}{}_{;i}\\ &= e^{(n+2)\varphi}\{(n+2)\varphi_{,i}T^{ij} + \operatorname{div}_g T^j\}\\ &= e^{(n+2)\varphi}\{g^{ij}\varphi_{,i}\ \operatorname{tr}_g(T) + \operatorname{div}_{\overline{g}}T^j\}.\end{aligned}$$

The same equation written invariantly is

$$\operatorname{div}_g \overline{T} = e^{(n+2)\varphi}\{\operatorname{grad}_g(\varphi)\operatorname{tr}_g(T) + \operatorname{div}_{\overline{g}}T\}, \tag{22}$$

verifying (16). From (22) it follows directly that

$$\operatorname{div}_g \overline{T} = e^{(n+2)\varphi}\ \operatorname{grad}(\varphi)\operatorname{tr}(T) \tag{23}$$

if and only if $\operatorname{div}_{\overline{g}}T = 0$.

By equation (22),

$$\operatorname{div}_g T = e^{-(n+2)\varphi}\{\operatorname{div}_g \overline{T} - e^{(n+2)\varphi}\ \operatorname{grad}_g\ \varphi\ \operatorname{tr}_g(T)\}. \tag{24}$$

Therefore, weak solutions to $0 = \operatorname{div}_{\overline{g}}T$ are weak solutions to

$$0 = e^{-(n+2)\varphi}\{\operatorname{div}_g \overline{T} - e^{(n+2)\varphi}\ \operatorname{grad}_g\ \varphi\ \operatorname{tr}_g(T)\}. \tag{25}$$

Equation (25) differs from (17) by a factor of $e^{-(n+2)\varphi}$. Weak solutions must satisfy

$$0 = \int_{\infty}^{\infty}\int_{t_0}^{T} e^{-(n+2)\varphi}\{\operatorname{div}_g \overline{T} - e^{(n+2)\varphi}\ \operatorname{grad}_g\ \varphi\ \operatorname{tr}_g(T)\}\phi\, dx^0\, dx^1, \tag{26}$$

but the factor $e^{-(n+2)\varphi}$ may be combined with the function ϕ in Eq. (26) because $e^{-(n+2)\varphi}$ is nonzero and smooth, yielding

$$0 = \int_{\infty}^{\infty}\int_{t_0}^{T} \{\operatorname{div}_g \overline{T} - e^{(n+2)\varphi}\ \operatorname{grad}_g\ \varphi\ \operatorname{tr}_g(T)\}\psi\, dx^0\, dx^1, \tag{27}$$

where $\psi = e^{-(n+2)\varphi}\phi$. Therefore, T is a weak solution of $\operatorname{div}_{\overline{g}}T = 0$ if and only if it is a weak solution to (17). □

There are several alternate forms of Eq. (17). From the following facts:

(1) $\operatorname{tr}_{\overline{g}}(\underline{T}) = \overline{g}_{ik}T^{ik} = e^{2\varphi}g_{ik}T^{ik} = e^{2\varphi}\operatorname{tr}_g(T)$,

(2) $\operatorname{tr}_g(T) = \operatorname{tr}_g(e^{(n+2)\varphi}T) = e^{(n+2)\varphi}\operatorname{tr}_g(T)$.

The following are equivalent to $\operatorname{div}_g T = 0$:

$$\operatorname{div}_g(\overline{T}) = e^{(n+2)\varphi}\ \operatorname{grad}_g(\varphi)\operatorname{tr}_g(T) = e^{n\varphi}\ \operatorname{grad}_g(\varphi)\operatorname{tr}_{\overline{g}}(T) = \operatorname{grad}_g(\varphi)\operatorname{tr}_g(\overline{T}) = e^{-2\varphi}\ \operatorname{grad}_g(\varphi)\operatorname{tr}_{\overline{g}}(\overline{T}).$$

For the Minkowski metric case presented in Ref. 5, the Christoffel symbols are zero. In addition, $\varphi \equiv 0$, so the source term is zero, leaving a pure conservation law. Note that from Eq.

(17) we see that it is not necessary to assume that $\varphi \equiv 0$ in order to obtain a pure conservation law. The assumption that $\text{grad}(\varphi)\text{tr}(T)=0$ will accomplish the same result. This assumption implies either $\text{grad}(\varphi)=0$ on an open set, in which case the metric in question is the Minkowski metric, or $\text{tr}(T)=0$, the case investigated in Ref. 2.

The energy-momentum tensor for a perfect fluid is given by

$$T^{ij}=(p+\rho c^2)u^i u^j+pg^{ij}, \tag{28}$$

where ρ is the mass-energy density, p is the fluid pressure, u^i, $i=0,1,2,3$ are the components of the 4-velocity vector field U of fluid particles, and g is the space–time metric. If T is the energy-momentum tensor for a perfect fluid, then (15) is called the *relativistic Euler equation.*

Perfect fluids with trace free energy-momentum tensors are precisely those with the equation of state

$$p=\frac{c^2}{(n-1)}\,\rho, \tag{29}$$

since

$$0=\text{tr}(T)=g_{ik}T^{ik}=g_{ik}\{(p+\rho c^2)u^i u^k+pg^{ik}\}=-(p+\rho c^2)+np=(n-1)p-\rho c^2.$$

Here we assume only that $p=\sigma^2\rho$, where σ is the constant sound speed, $0<\sigma<c$.

With the metric $g=e^{2\varphi}\eta$, (17) takes the form

$$\text{div}_\eta\,\overline{T}=e^{n\varphi}\,\text{grad}_\eta(\varphi)\text{tr}_g(T). \tag{30}$$

Substituting the energy-momentum tensor for a perfect fluid we obtain,

$$\begin{aligned}\{e^{(n+2)\varphi}((p+\rho c^2)u^i u^j+pe^{-2\varphi}\eta^{ij})\}_{,i}&=e^{n\varphi}\varphi_{,i}\eta^{ij}g_{lk}\{(p+\rho c^2)u^l u^k+pg^{lk})\}\\&=e^{n\varphi}\varphi_{,i}\eta^{ij}\{-(p+\rho c^2)+pn\}\\&=e^{n\varphi}\varphi_{,i}\eta^{ij}\{p(n-1)-\rho c^2\}.\end{aligned}$$

With the equation of state $p=\sigma^2\rho$, we obtain the system of equations

$$\begin{aligned}&\{e^{n\varphi}\rho((\sigma^2+c^2)(u^0)^2-\sigma^2e^{-2\varphi})\}_{,0}+\{e^{n\varphi}(\rho(\sigma^2+c^2)u^0u^1\}_{,1}\\&\quad=-e^{n\varphi}\rho\varphi_{,0}\{\sigma^2(n-1)-c^2\}\{e^{n\varphi}\rho((\sigma^2+c^2)u^0u^1)\}_{,0}\\&\qquad+\{e^{n\varphi}\rho((\sigma^2+c^2)(u^1)^2+\sigma^2e^{-2\varphi})\}_{,1}\\&\quad=e^{n\varphi}\rho\varphi_{,1}\{\sigma^2(n-1)-c^2\}.\end{aligned} \tag{31}$$

Since $-1=g(U,U)$ and $u^1=\dfrac{v}{c}\,u^0$, it follows that

$$u^1=\frac{v^2e^{-2\varphi}}{c^2-v^2}. \tag{32}$$

Substituting (32) into the system (31), and setting $\overline{\rho}=\rho e^{n\varphi}$, we have

$$\left\{\overline{\rho}\left((\sigma^2+c^2)\,\frac{c^2}{c^2-v^2}-\sigma^2\right)\right\}_{,0}+\left\{\overline{\rho}(\sigma^2+c^2)\,\frac{cv}{c^2-v^2}\right\}_{,1}=-\overline{\rho}\varphi_{,0}\{\sigma^2(n-1)-c^2\}, \tag{33}$$

$$\left\{\overline{\rho}(\sigma^2+c^2)\,\frac{cv}{c^2-v^2}\right\}_{,0}+\left\{\overline{\rho}\left((\sigma^2+c^2)\,\frac{v^2}{c^2-v^2}+\sigma^2\right)\right\}_{,1}=\overline{\rho}\varphi_{,1}\{\sigma^2(n-1)-c^2\}, \tag{34}$$

from which we obtain (4).

In the classical theory of shock waves, a conservation law is a PDE of the form

$$U_t + F(U)_x = 0, \tag{35}$$

where $U \in R^n$, $n \in N$ and $(x,t) \in R \times R^+$ (Ref. 1, p. 239). However, due to the existence of a gravitational source term, it will be seen that the relativistic Euler equations for one-dimensional fluid flow with conformally flat metrics can be written as a system of two equations of the form

$$U_t + F(U)_x = G(x,t,U). \tag{36}$$

This type of equation is known as a system of conservation laws with a lower order source term.

Set

$$U^0 \overset{\text{def}}{=} \left\{ \overline{\rho}(\sigma^2 + c^2) \, \frac{c^2}{c^2 - v^2} - \sigma^2 \right\},$$

$$U^1 \overset{\text{def}}{=} \left\{ \overline{\rho}(\sigma^2 + c^2) \, \frac{cv}{c^2 - v^2} \right\},$$

$$U^2 \overset{\text{def}}{=} \left\{ \overline{\rho}(\sigma^2 + c^2) \, \frac{v^2}{c^2 - v^2} + \sigma^2 \right\},$$

as well as $U \overset{\text{def}}{=} (U^0, U^1)^T$, $F(U) \overset{\text{def}}{=} (U^1, U^2)^T$, and

$$G(x^0, x^1, U) \overset{\text{def}}{=} - \begin{pmatrix} -e^{(n+2)\varphi} \overline{\rho} \varphi_{,0} \{\sigma^2 (n-1) - c^2\} \\ e^{(n+2)\varphi} \overline{\rho} \varphi_{,1} \{\sigma^2 (n-1) - c^2\} \end{pmatrix}. \tag{37}$$

The system (31) can be put into the form (36), if we can show that U^2 is a function of U^0 and U^1, and if we can show that G is a function of U and the coordinates x^0 and x^1 alone.

First, define $\Phi(\rho, v) = (U^0, U^1)$, where for simplicity, we write ρ instead of $\overline{\rho}$. If Φ is one-to-one, then we may set $\Phi^{-1}(U^0, U^1) = (\rho, v)$, where we can write $\Phi_1^{-1}(U) = \rho$ and $\Phi_2^{-1}(U) = v$, and

$$f(U) \overset{\text{def}}{=} U^2(\Phi_1^{-1}(U), \Phi_2^{-1}(U)) = U^2(\rho, v). \tag{38}$$

From this we may set $(U^1, U^2)^T = (U^1, f(U))^T$ which shows that F is indeed a function of U. We must still show that Φ is invertible, which is accomplished via the following lemma.

Let J be the subset of the (ρ, v) plane corresponding to positive densities $\rho > 0$ and timelike velocities $|v| < c$.

Lemma II.2: The function $\Phi(\rho, v) = (U^0, U^1)$ *satisfies the following on the region* J:

(1) *the Jacobian determinant of* Φ *is continuous,*

(2) *the Jacobian determinant of* Φ *does not vanish,*

(3) Φ *is one-to-one.*

Proof: The Jacobian determinant of Φ is

$$|J\Phi| = \frac{c(p + \rho c^2)}{-e^{4\varphi}(c^2 - v^2)^2} \{c^4 - \sigma^2 v^2\} \tag{39}$$

which is continuous on J.

If $|J\Phi| = 0$, then $c^4 = \sigma^2 v^2 < c^2 v^2$ which in turn implies that $c < |v|$. Thus, v is outside of the assumed range and $|J\Phi| \neq 0$ in the region of interest.

To show that Φ is one-to-one, let $(\rho_1,v_1)\neq(\rho_2,v_2)$ be two points such that $\Phi(\rho_1,v_1)=\Phi(\rho_2,v_2)$. First, note that if $\rho_1=\rho_2$, then since $U^0(\rho_1,v_1)=U^0(\rho_2,v_2)$ we have that $c^2-v_1^2=c^2-v_2^2$. Substituting this into the equation $U^1(\rho_1,v_1)=U^1(\rho_2,v_2)$ we find that $v_1=v_2$. Similarly, if $v_1=v_2$, then $\rho_1=\rho_2$.

Since $\Phi(\rho_1,v_1)=\Phi(\rho_2,v_2)$,

$$\rho_1\left(\frac{(c^2+\sigma^2)cv_1e^{-2\varphi}}{c^2-v_1^2}\right)=\rho_2\left(\frac{(c^2+\sigma^2)cv_2e^{-2\varphi}}{c^2-v_2^2}\right), \tag{40}$$

$$\rho_1e^{-2\varphi}\left(\frac{(c^2+\sigma^2)c^2}{c^2-v_1^2}-\sigma^2\right)=\rho_2e^{-2\varphi}\left(\frac{(c^2+\sigma^2)c^2}{c^2-v_2^2}-\sigma^2\right). \tag{41}$$

Dividing (41) by (40), we have

$$\frac{c}{v_1}-\frac{\sigma^2(c^2-v_1^2)}{(c^2+\sigma^2)cv_1}=\frac{c}{v_2}-\frac{\sigma^2(c^2-v_2^2)}{(c^2+\sigma^2)cv_2}. \tag{42}$$

After some simplification, (42) becomes

$$c^2(c^2+\sigma^2)(v_2-v_1)=\sigma^2(v_2-v_1)\{c^2-v_1v_2\}. \tag{43}$$

Assuming that $v_1\neq v_2$, we may simplify (43) further as

$$\frac{c^4}{\sigma^2}=-v_1v_2. \tag{44}$$

Taking the absolute value of both sides of (44), we have

$$|v_1v_2|\geqslant c^2, \tag{45}$$

which is not possible for timelike particles. Thus $v_1=v_2$, which implies that $\rho_1=\rho_2$ as we have seen. We conclude that Φ is one-to-one. □

III. THE APPROXIMATE SOLUTIONS

In order to implement a Glimm-type scheme it is necessary to construct a grid, for which, in order to be well-defined, it is necessary that the grid lines be constructed from coordinate lines. To see this, suppose that two arbitrary linearly independent vector fields X and Y are chosen, in place of $\partial/\partial x^0$ and $\partial/\partial x^1$, with which we are to construct a grid. Following the vector field X a distance Δx from some point (x_0,y_0), and then following the vector field Y a distance Δy, we end up at some point (x_1,y_1). Conversely, if we follow the vector field Y a distance Δy from the point (x_0,y_0), and then following the vector field X a distance Δx we end up at some point (x_2,y_2). In general $(x_1,y_1)\neq(x_2,y_2)$ making it impossible to form a closed grid, unless X and Y are *coordinate* vector fields (see Ref. 13 pp. 218, 219).

The establishment of a Courant–Friedrichs–Lewy (CFL) condition is complicated by the fact that in some coordinate systems on space–time manifolds, the speed of light is not bounded. For lightlike particles,

$$\frac{dx^1}{dx^0}=\left(\frac{-g_{01}\pm\sqrt{-\hat{g}}}{g_{11}}\right), \tag{46}$$

where $g_{11}\neq 0$. If the supremum in absolute value of (46) exists,

$$C=\sup_{p\in M}\left|\frac{-g_{01}(p)\pm\sqrt{-\hat{g}(p)}}{g_{11}(p)}\right|, \tag{47}$$

then we can set $\Delta x^1/\Delta x^0>2C$. In the present case, $dx^1/dx^0=\pm 1$.

As examples of the CFL condition, we consider conformally flat spaces, the Schwarzschild metric, and a Robertson–Walker space. For a conformally flat space, $C=1$ so a single grid can be used clear till the space crashes to a point (or from a point). For the Robertson–Walker metric

$$ds^2=-(dx^0)^2+f^2(x^0)((dx^1)^2+(dx^2)^2+(dx^3)^2), \tag{48}$$

we have

$$C=\sup_{p\in M} 1/f \tag{49}$$

so, as long as f is bounded away from zero, one grid can be used for as long as time makes sense. This means that big bangs and big crunches must be avoided in order for the existence theorem to go through, though this is only partly true. In the limit, existence is guaranteed from big bang to big crunch though uniform bounds are not guaranteed by this method. This should not be surprising since density is expected to become unbounded in these extreme situations anyway. Note however, that uniform bounds have been demonstrated to exist from big bang to big crunch on conformally flat space–times with trace-free energy-momentum tensors in Ref. 2.

For the Schwarzschild metric,

$$ds^2=-\left(1-\frac{2M}{r}\right)dt^2+\left(1-\frac{2M}{r}\right)^{-1}dr^2+r^2\,d\Omega^2, \tag{50}$$

where $d\Omega^2=d\theta^2+\sin^2\theta\,d\varphi^2$, we have

$$C=\sup_{p\in M}|1-2M/r|. \tag{51}$$

In the region $r\geqslant\delta>0$, the sup exists and a CFL condition can be imposed on a grid.

A. The approximate solutions

An operator splitting technique is used to construct the approximate solutions for the initial value problem

$$U_t+F(U)_x=G(x,t,U), \tag{52}$$

$$U(x,0)=U_0(x). \tag{53}$$

The approximates are constructed in two steps which will be called the *Conservation law step*, and the *ODE step*. On the conservation law step the solution at time t_i is discretized (made piecewise constant) by Glimm's method (see, e.g., Refs. 14, 1), and the resulting Riemann problems are solved with respect to the equations

$$U_t+F(U)_x=0 \tag{54}$$

on the discretized data, and for a time step. Then, on the ODE step, the ODE's $U_t=G(x,t,U)$ are solved with initial values being the solutions to the Riemann problems. The approximates at time $t\in[t_i,t_{i+1}]$ are defined to be the value of the solution to the ODE after $t-i\Delta t$ units of time. More precisely, let Δx and Δt be chosen so as to satisfy the CFL condition, and set $x_i=i\Delta x$ and $t_i=t_0+i\Delta t$. We may, without loss of generality, set $t_0=0$. Select a sequence of random points $\theta=(\theta_1,\theta_2,\cdots),\theta_i\in(0,1)$, and denote the solution to the conservation law step on the ith time step by ${}^R_iU$ and the solution to the ODE step on the ith time step by ${}^{\mathrm{ODE}}_iU$. Then ${}^R_iU$ solves the homogeneous equation (54) on $R\times[t_i,t_{i+1}]$ with initial data being ${}^{\mathrm{ODE}}_{i-1}U(x,t_i)$. Let ${}^{\mathrm{Temp}}U(t,U_0)$ solve the ODE

$$U_t=G(x,t,U),$$

$$U(0)=U_0.$$

Then ${}^{\mathrm{ODE}}_{i}U$ is defined by ${}^{\mathrm{ODE}}_{i}U(x,t)={}^{\mathrm{Temp}}U(t-t_i,{}^{R}_{i}U(x,t)),\ t\in[t_i,t_{i+1}]$.

B. The potential

We next define I-curves, and a potential on the I-curves. As in Ref. 3, we define *vertex points* as

$$(x_i,t_j+\tfrac{1}{2}\Delta t),$$

$$(x_i+\theta_i\Delta x_i,t_j),$$

where $j\geqslant 0$ and $i\in Z$. An I-curve is a piecewise linear curve in xt-space which connects a vertex point on t_j to a vertex point on $t_j\pm\frac{1}{2}\Delta t$ with line segments, and where I-curves are functions of x. I-curves are given a partial ordering by defining that larger curves lie toward larger time. We call I_2 an *immediate successor* to I_1 if these two I-curves connect the same vertices except one, and if $I_2>I_1$.

Let γ be a wave. We define its strength by $|\gamma|=|\ln(\rho_r)-\ln(\rho_l)|$. Given this definition of wave strength we define the potential for the conservation law step to be

$$L(I)=\sum_{\gamma\ \mathrm{crosses}\ I}|\gamma|. \tag{55}$$

IV. CONSERVATION LAW STEP BOUNDS

In this section we consider the solutions on the conservation law step, and bound the potential. We first find the eigenvalues and Riemann invariants, and we determine when the conservation laws are strictly hyperbolic and genuinely nonlinear. Next we prove that the Lax entropy inequalities hold, and parameterize the shock curves. The geometry of the shock curves is then described. Finally, the Riemann problems are solved and it is shown that the potential is nonincreasing on the conservation law step.

By demonstrating that the entropy inequalities hold, we show that the approximate solutions we construct are physically relevant, and any shock waves in our approximate solutions are not entropy violating rarefaction shocks. The geometry of the shock curves as described in Lemma IV.6 is necessary, first to demonstrate existence of solutions to Riemann problems, and then to show that interacting waves produce waves of equal or lesser total variation. In this way, the approximate solutions on the conservation law step will be total variation diminishing. Note, however, that the gravitational source term can increase the total variation of the approximate solutions.

On the conservation law step, we set the source term to zero and solve Riemann problems on the source-free equations, isolating the analysis of the sources in the next section. Therefore, our analysis in this section will closely follow the Minkowski metric case, which can be found in Ref. 5. We note that our proof that the Lax entropy inequalities hold is ours.

A. Eigenvalues and Riemann invariants

For the conservation law step we set the source term to zero and consider only the equations

$$U_{,0}+F(U)_{,1}=0. \tag{56}$$

For what follows, set $A=U^0$, $B=U^1$, and $C=U^2$. We also write, for simplicity, ρ in place of $\overline{\rho}$.

The wave speeds are the eigenvalues λ_i of the Jacobian matrix $dF=\partial F/\partial U$. Let R_i denote the ith eigenvector of dF. We will calculate the wave speeds and the Riemann invariants using the following well-known theorem due to Lax.[15]

Theorem IV.1 (Lax): *Assume that the system (56) is strictly hyperbolic, and for some fixed state U_l, the solutions U and s of the Rankine–Hugoniot relation $s(U-U_l)=F(U)-F(U_l)$ can be described (in a neighborhood of U_l) by two families of smooth curves $U=\mathbf{S}_i(\epsilon)$, $\mathbf{S}_i(0)=U_l$, with corresponding speeds $s_i(\epsilon), i\in\{1,2\}$. Moreover, as $\epsilon\to 0$, we have $s_i(\epsilon)\to\lambda_i(U_l)$ and $U_i'(\epsilon)\to R_i$. Here, the parameter ϵ can be taken to be (Euclidean) arc length along the shock curve $\mathbf{S}_i$ in U-space.*

First, we use this theorem to calculate the eigenvalues and eigenvectors for the equation of state $p(\rho)$. The Rankine–Hugoniot jump conditions imply that if the state (ρ_r, v_r) lies on a shock curve starting at (ρ_l, v_l), then

$$[B]^2=[A][C]. \tag{57}$$

Since (57) defines v implicitly as a function of ρ, we may differentiate (57) with respect to ρ and divide by $[B]$ to obtain

$$2B'=\frac{[A]}{[B]}C'+\frac{[C]}{[B]}A', \tag{58}$$

where the prime denotes differentiation with respect to ρ. Using L'Hospital's rule,

$$\lim_{\rho\to\rho_l}\frac{[A]}{[B]}=\left.\frac{A'}{B'}\right|_{\rho=\rho_l}. \tag{59}$$

Thus, at $\rho=\rho_l$, (57) can be written as

$$(B')^2=A'C'. \tag{60}$$

We solve (60) for v' obtaining

$$\frac{v'}{c^2-v^2}=\pm\frac{\sqrt{p'}}{p+\rho c^2}. \tag{61}$$

We use (61) to find the Riemann invariants. As defined in Ref. 1 p. 321, a k-Riemann invariant is a smooth function w such that $\langle R_k, \nabla w\rangle=0$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product on R^n. We note that Riemann invariants are constant along integral curves of eigenvector fields of dF. By Theorem IV.1, we ultimately obtain the Riemann invariants

$$r=\frac{1}{2}\ln\left|\frac{c+v}{c-v}\right| c\int_1^\rho\frac{\sqrt{p'(\epsilon)}}{p(\epsilon)+\epsilon c^2}\,d\epsilon, \tag{62}$$

$$s=\frac{1}{2}\ln\left|\frac{c+v}{c-v}\right|+c\int_1^\rho\frac{\sqrt{p'(\epsilon)}}{p(\epsilon)+\epsilon c^2}\,d\epsilon. \tag{63}$$

We now calculate the eigenvalues. Again using Theorem IV.1 and L'Hospital's rule we have

$$\lambda_i=\lim_{\rho\to\rho_l}\frac{[B]}{[A]}=\left.\frac{B'}{A'}\right|_{\rho=\rho_l}. \tag{64}$$

B'/A' is a quantity involving $\rho, v, p'=dp/d\rho$, and $v'=dv/d\rho$. Substituting (61) into (64) and simplifying we obtain

$$\lambda_1=\frac{cv-\sqrt{p'}c}{c^2-\sqrt{p'}v}, \tag{65}$$

$$\lambda_2=\frac{cv+\sqrt{p'}c}{c^2+\sqrt{p'}v}. \tag{66}$$

We are now in a position to prove strict hyperbolicity and genuine nonlinearity.

Theorem IV.2: *System (56) is strictly hyperbolic at* (ρ,v) *if and only if*

$$\sqrt{p'(\rho)}<c. \tag{67}$$

Moreover, assuming (67), system (4) is genuinely nonlinear at each (ρ,v) *if and only if*

$$p''(\rho)\neq -2\frac{(c^2-p')p'}{(p+\rho c^2)}. \tag{68}$$

Proof: We first show that $\lambda_1<\lambda_2$ if and only if $\sqrt{p'(\rho)}<c$ and hence (56) is strictly hyperbolic. To this end, a calculation can show that

$$\lambda_1-\lambda_2=-\frac{2c\sqrt{p'}(c^2-v^2)}{c^4-p'v^2}. \tag{69}$$

The right-hand side of Eq. (69) is negative for all $|v|<c$ if and only if p' is bounded by the square of the speed of light, c^2. We have shown that $\lambda_1<\lambda_2$, which demonstrates strict hyperbolicity, if and only if $\sqrt{p'(\rho)}<c$.

We are now in a position to verify that r (respectively, s) is indeed a 1- (respectively, 2-) Riemann invariant. To determine under what conditions the system (4) will be genuinely nonlinear, we must calculate the right eigenvectors of dF. We do this using the Riemann invariants and then check that $\langle \text{grad}\,\lambda_i,R_i\rangle\neq 0$. The eigenvectors may be found using the equations

$$R_1=\left(-\frac{\partial r}{\partial v},-\frac{\partial r}{\partial \rho}\right)^T,$$

$$R_2=\left(-\frac{\partial s}{\partial v},-\frac{\partial s}{\partial \rho}\right)^T=\left(\frac{\partial r}{\partial v},-\frac{\partial r}{\partial \rho}\right)^T,$$

from which we obtain

$$R_1=\left(\frac{-c}{(c^2-v^2)},\frac{\sqrt{p'}}{p+\rho c^2}\right)^T,$$

$$R_2=\left(\frac{c}{(c^2-v^2)},\frac{\sqrt{p'}}{p+\rho c^2}\right)^T.$$

A straightforward computation shows that

$$\text{grad}\,\lambda_i=\frac{1}{\{c^2\mp\sqrt{p'}v\}^2}\left(\pm\frac{p''c(c^2-v^2)}{2\sqrt{p'}},c(p'-c^2)\right)^T.$$

The gradient here is with respect to the Euclidean inner product on R^2. It can now be shown that

$$\langle \text{grad}\,\lambda_i,R_i\rangle=\frac{c^2}{\{c^2\mp\sqrt{p'}v^2\}^2}\left(\frac{p''}{2\sqrt{p'}}+\frac{\sqrt{p'}(c^2-p')}{(p+\rho c^2)}\right). \tag{70}$$

From this it follows that the system (4) is genuinely nonlinear if and only if (68) holds. □

B. The shock curves and the entropy inequalities

Simplifying the Eq. (57) we obtain

$$\frac{(p_r-p_l)(\rho_r-\rho_l)}{(p_r+\rho_r c^2)(p_l+\rho_l c^2)}=\frac{(v_r-v_l)^2}{(c^2-v_r^2)(c^2-v_l^2)}. \tag{71}$$

With the equation of state $p=\sigma^2\rho$, (71) becomes

$$\frac{(\rho_r-\rho_l)^2}{\rho_r\rho_l}=\frac{(\sigma^2+c^2)^2(v_r-v_l)^2}{\sigma^2(c^2-v_r^2)(c^2-v_l^2)}. \tag{72}$$

The right-hand side of (72) we define to be

$$2\beta \overset{\text{def}}{=} \frac{(\sigma^2+c^2)^2(v_r-v_l)^2}{\sigma^2(c^2-v_r^2)(c^2-v_l^2)}. \tag{73}$$

Setting

$$z \overset{\text{def}}{=} \frac{\rho_r}{\rho_l}, \tag{74}$$

and solving for z in (72) yields

$$z = 1+\beta\left\{1 \pm \sqrt{1+\frac{2}{\beta}}\right\} \overset{\text{def}}{=} f_{\pm}(\beta). \tag{75}$$

For simplicity of the equations that follow, set

$$K \overset{\text{def}}{=} \frac{2\sigma^2c^2}{(\sigma^2+c^2)^2}. \tag{76}$$

We now address the Lax entropy inequalities, which are shown to hold all along the shock curves. The following lemma appears in Ref. 5 p. 81 in their treatment of the 2-dimensional space–time Minkowski metric case, though the proof given here is different.

Lemma IV.3: The shock speed s_i is monotone all alone the shock curve $\mathbf{S}_i$ of system (4), $i=1,2$, and moreover, the following inequalities (Lax shock conditions) hold at each state $U_r \neq U_l$ on the shock curve $\mathbf{S}_i$:

$$\lambda_i(\rho_r,v_r) < s_i < \lambda_i(\rho_l,v_l). \tag{77}$$

Proof: Recall that (57) defines v as a function of ρ. Therefore, the equation $s=[B]/[A]$ defines the shock speeds s as a function of ρ. We show that $ds/d\rho \neq 0$. To this end, note that since

$$s^2 = \frac{[C]}{[A]}, \tag{78}$$

we have

$$2s\,\frac{ds}{d\rho} = \frac{1}{[A]}\{C'-s^2A'\}. \tag{79}$$

Note that $ds/d\rho$ equals zero if and only if

$$s^2 = \frac{C'}{A'} = \lim_{\rho_l \to \rho_r} \frac{[C]}{[A]}. \tag{80}$$

Since

$$\lambda_i(\rho_r,v_r) = \frac{B'}{A'}, \tag{81}$$

Eqs. (80) and (81) together imply that the only way that $s'=0$ is if $s=\lambda_i(\rho_r,v_r)$. We solve

$$\lambda_i = s_j = \frac{[B]}{[A]} \tag{82}$$

for v_r, where the numerator and denominator of the right-hand side of (82) is divided through by ρ_l and ρ_r/ρ_l is replaced by $1+\beta\{1\mp\sqrt{1+(2/\beta)}\}$. β is then replaced by

$$\frac{(\sigma^2+c^2)^2(v_r-v_l)^2}{2\sigma^2(c^2-v_r^2)(c^2-v_l^2)}. \tag{83}$$

After this tedious computation, the only solutions are $v_r=v_l$, in which case there is no shock, and

$$v_r=\frac{c^4}{\sigma^2 v_l}. \tag{84}$$

Since $\sigma<c$ and $|v_l|<c$, we have $|v_r|>c$. Note that this contradiction implies that either $\lambda_i(\rho_{r,l},v_{r,l})<s_j$ or $\lambda_i(\rho_{r,l},v_{r,l})>s_j$. We have shown that the shock speed s_i is monotone all along the shock curves $\mathbf{S}_i$, since $s'\neq 0$ on these curves. There is now a simple procedure to verify that the shock inequalities hold. Since we have shown that $s_i\neq\lambda_j$, we need only test one point in each of the six relevant intervals. This was done and it was found that

$$\begin{gathered}\lambda_1(r)<s_1<\lambda_1(l)\lambda_2(r)<s_2<\lambda_2(l),\\ s_1<\lambda_2(r),\quad \lambda_1(l)<s_2.\end{gathered} \tag{85}$$

□

C. Geometry of the shock curves

The goal of this section is to prove that the i-shock curves at different points are translates of each other in Riemann invariant coordinates on state space, and that 1-shock and 2-shock curves are reflections of each other about a particular line in the state space under Riemann invariant coordinates.

All of the lemma's in this section are the same as in Ref. 5 pp. 84–87,90. (Note that some of the pages were printed out of order.) All of the proofs are essentially identical. However, these lemma's are important to the existence theorem so, they will be included here.

Lemma IV.4: Let $r\equiv r(\rho,v)$, $s\equiv s(\rho,v)$, $\Delta r\equiv r(\rho,v)-r(\rho_l,v_l)$ *and* $\Delta s\equiv s(\rho,v)-s(\rho_l,v_l)$, *where we let* $(\rho,v)\equiv(\rho_l,v_l)$. *Then the 1-shock curve* $\mathbf{S}_1$ *for system (4) based at* (r_l,s_l) *is given by the following parametrization with respect to* $\beta,0\leqslant\beta<\infty$:

$$\Delta r=-\frac{1}{2}\ln\{f_+(2K\beta)\}-\sqrt{\frac{K}{2}}\ln\{f_+(\beta)\}, \tag{86}$$

$$\Delta s=-\frac{1}{2}\ln\{f_+(2K\beta)\}+\sqrt{\frac{K}{2}}\ln\{f_+(\beta)\}; \tag{87}$$

and the 2-shock curve $\mathbf{S}_2$ *based at* (r_l,s_l) *is given for* $0\leqslant\beta<\infty$, *by*

$$\Delta r=-\frac{1}{2}\ln\{f_+(2K\beta)\}-\sqrt{\frac{K}{2}}\ln\{f_-(\beta)\}, \tag{88}$$

$$\Delta s=-\frac{1}{2}\ln\{f_+(2K\beta)\}+\sqrt{\frac{K}{2}}\ln\{f_-(\beta)\}. \tag{89}$$

Lemma IV.5: The 2-shock curve based at an arbitrary point (r_l,s_l) *is the reflection in the* rs*-plane of the 1-shock curve based at the same point, where the axis of reflection is the line passing through* (r_l,s_l), *parallel to the liner* $r=s$.

Lemma IV.6: The shock curves $\mathbf{S}_i$ *define convex curves in the* rs*-plane and moreover,*

$$0\leqslant\frac{ds}{dr}<\frac{\sqrt{2K}-1}{-\sqrt{2K}-1}<1 \tag{90}$$

all along the 1-shock curve $\mathbf{S}_1$, *and*

$$0 \leqslant \frac{dr}{ds} < \frac{\sqrt{2K}-1}{-\sqrt{2K}-1} < 1 \tag{91}$$

all along the 2-shock curve $\mathbf{S}_2$.

D. The Riemann problem

The Riemann problem solutions we generate are not Riemann problem solutions to the full relativistic Euler equations. Instead, they are Riemann problem solutions to the Euler equations with the source set to zero. The geometry of the shock curves in the plane of Riemann invariants as described in Lemma's IV.5, and IV.6 is special to the equation of state $p=\sigma^2\rho$. Given a different equation of state, the geometry of the shock curves could be much more complex. This remarkable geometry was first noticed by Nishida in Ref. 6 in his treatment of the classical case. It was also used by Smoller and Temple in Ref. 5 in their treatment of the special relativistic case. We take advantage of the special geometry in this case to demonstrate the existence of solutions to each Riemann problem, and to obtain an inequality which will be important in proving the bounds (8), (9) in Theorem I.1.

Theorem IV.7: *There exists a solution of the Riemann problem for system (4) in the case of an equation of state of the form* $p=\sigma^2\rho$, $\sigma^2<c^2$, *so long as density is positive and velocities are for timelike particles. Moreover, the solution is given by a 1-wave followed by a 2-wave, also has positive density, and particle velocities are for timelike particles. The solution is unique in the class of rarefaction waves and admissible shock waves.*

Briefly, the proof of Theorem IV.7 is as follows. Lemmas IV.5 and IV.6 show that the wave curves at a point (r_l,s_l) form a coordinate system. Thus for each point (r_r,s_r) there is a unique 1-wave curve and a unique 2-wave curve such that we may pass from (r_l,s_l) to (r_r,s_r) along the path consisting of the 1-wave curve followed by the 2-wave curve. The solution to the Riemann problem (U_l,U_r) will be these two waves.

The wave interaction estimates for solutions to the Riemann problems on the conservation law step can be obtained exactly as in Ref. 5 pp. 92–93 by which we have the following lemma.

Lemma IV.8: Let U_l, U_m, *and* U_r *denote three arbitrary states with positive density and timelike velocities. Let* $\alpha_1,\alpha_2,\beta_1,\beta_2$, and γ_1,γ_2 *denote the waves that solve the Riemann problems*

$$[U_l,U_m],[U_m,U_r],[U_l,U_r], \tag{92}$$

respectively. Then

$$|\gamma_1|+|\gamma_2| \leqslant |\alpha_1|+|\alpha_2|+|\beta_1|+|\beta_2|. \tag{93}$$

The idea of the proof is quite simple. The strengths of the waves going in for all possible cases of wave interactions are compared to the strengths of the waves going out using the geometry of the wave curves in the plane of Riemann invariants. The inequalities (90) and (91) are particularly helpful in demonstrating the inequality (93) when shock waves are involved in the interaction. Note that since rarefaction waves are straight lines in the plane of Riemann invariants, equality holds in (93) when only rarefaction waves are involved.

Lemma IV.8 establishes the following.

Lemma IV.9: Recall that $L(j)$ *is the total strength of the waves in the approximate solution at time step* j, *and let* $\hat{L}(j)$ *denote the total strength of the waves in the approximate solutions after the conservation law step (but before the ODE step) at time step* j. *Then* $\hat{L}(j+1) \leqslant L(j)$.

V. ODE STEP BOUNDS

In the construction of the approximate solutions we first solve conservation laws with the source term set to zero. We then modify these solutions by a system of ODE's which adds the effects of the source term. The potential function may increase through this process, so bounds on the potential function must be demonstrated to exist. In this section we find bounds on the

solutions needed in order to maintain bounds on the potential. It will not be possible to demonstrate the existence of these bounds for all metrics. In general, black holes and big bangs can cause the potential to become unbounded in finite time.

The wave strengths are measured in terms of variations in $\log \rho$, and therefore, the total wave strength is the total variation in $\log \rho$. The source term may add to the total variation in $\log \rho$ by increasing wave strengths and adding new waves. For our proof of convergence to go through, we must show that the total variation in $\log \rho$ will remain bounded. The following theorem shows that if the total variation of the gravitational source term is of bounded variation for each time t, then it cannot increase the total wave strength by more than a finite amount on each time step. Furthermore, if the total variation of the gravitational source term is an integrable function of time, then total wave strength contributed by the gravitational source term is integrable, and hence the total variation in $\log \rho$ will remain uniformly bounded. If, however, the gravitational source term does not depend on time, then the sources will continue to feed wave strength into the solution, and hence we can at best guarantee a linear growth bound on the total variation of the solution.

Theorem V.1: *If*

$$\int_0^\infty \mathrm{TV}_x(\varphi,i)(t)dt<\infty \tag{94}$$

for $i\in\{0,1\}$, *then* $L(j)\leqslant L(0)+K$, *where* K *is a constant depending only on the metric. If* $\varphi_{,i}$, $i\in\{0,1\}$ *is a function of bounded variation in* x *alone, then* $L(j)\leqslant L(0)+Kt$.

Proof: Let ${}^R\rho^j$ denote the solution of the Conservation Law Step on the time interval $[t_j,t_{j+1}]$ at time t_{j+1}^-, since this will be the data for the ODE's at time t_j. Denote by ρ^j the solution of the ODE's on $[t_j,t_{j+1}]$, i.e., after the ODE's have operated on the solutions for time $t-t_j$. We will write the ODE's in the form $\rho_t/\rho=\sigma$, where σ is the source term. In this case, for a fixed x,

$$\ln(\rho^j)=\ln({}^R\rho^j)+\int_{t_j}^t \sigma\, dt. \tag{95}$$

Therefore,

$$\mathrm{TV}_x(\ln(\rho^j))\leqslant \mathrm{TV}_x(\ln({}^R\rho^j))+\mathrm{TV}_x\left(\int_{t_j}^t \sigma\, dt\right) \tag{96}$$

$$\leqslant \mathrm{TV}_x(\ln({}^R\rho^j))+\int_{t_j}^t \mathrm{TV}_x(\sigma)(t)dt. \tag{97}$$

By the Mean Value Theorem for Integrals, (97) leads to

$$\mathrm{TV}_x(\ln(\rho^j))\leqslant \mathrm{TV}_x(\ln({}^R\rho^j))+\mathrm{TV}_x(\sigma)(\xi_j)\Delta t, \tag{98}$$

where $\xi_j\in[t_j,t]$. By induction, (98) implies

$$\mathrm{TV}_x(\ln(\rho^{j+1}))\leqslant \mathrm{TV}_x(\ln({}^R\rho^0))+\sum_{k=0}^{j} \mathrm{TV}_x(\sigma)(\xi_k)\Delta t, \tag{99}$$

and therefore

$$\mathrm{TV}_x(\ln(\rho^j))\leqslant \mathrm{TV}_x(\ln({}^R\rho^0))+K\int_0^{t_{j+1}} \mathrm{TV}_x(\sigma)(t)dt. \tag{100}$$

Equation (100) shows that if

$$\int_0^\infty \mathrm{TV}_x(\sigma)(t)dt \tag{101}$$

converges, then the total variation of $\log(\rho)$ is uniformly bounded. But, if the source σ is a function of x alone, then the total variation will grow at most linearly.

We must now examine the source term of the relativistic Euler equations. On the ODE step we solve the ODE's

$$U_t = G(x,t,U). \tag{102}$$

Now, $U_t = U_\rho \rho_t + U_v v_t$, which may be solved for ρ_t and v_t, where we are using the convention that ρ here means $\rho e^{n\varphi}$ for simplicity. We may write the ODE's for the Euler equations in the form

$$U^0_{,\rho}\rho_{,0} + U^0_{,v} v_{,0} = G^0, \tag{103}$$

$$U^1_{,\rho}\rho_{,0} + U^1_{,v} v_{,0} = G^1. \tag{104}$$

Set

$$D = \begin{vmatrix} U^0_{,\rho} & U^0_{,v} \\ U^1_{,\rho} & U^1_{,v} \end{vmatrix}, \tag{105}$$

$$D_\rho = \begin{vmatrix} G^0_{,\rho} & U^0_{,v} \\ G^1_{,\rho} & U^1_{,v} \end{vmatrix}, \tag{106}$$

$$D_v = \begin{vmatrix} U^0_{,\rho} & G^0_{,v} \\ U^1_{,\rho} & G^1_{,v} \end{vmatrix}.$$

Then by Cramer's rule, $\rho_{,0} = D_\rho / D$ and $v_{,0} = D_v / D$. Using the form for U^0, U^1, G^0, and G^1 in (4), we obtain

$$\frac{\rho_{,0}}{\rho} = -\frac{(\sigma^2(n-1)-c^2)[(c^2+v^2)\varphi_{,0} + 2cv\varphi_{,1}]}{c^4 - \sigma^2 v^2}, \tag{107}$$

$$v_{,0} = \frac{(\sigma^2(n-1)-c^2)(c^2-v^2)}{c(\sigma^2+c^2)(c^4-\sigma^2 v^2)}\,((c^4+\sigma^2 v^2)\varphi_{,1} + (\sigma^2+c^2)cv\varphi_{,0}). \tag{108}$$

Set

$$\sigma_0(v,\varphi) = \frac{(\sigma^2(n-1)-c^2)[(c^2+v^2)\varphi_{,0} + 2cv\varphi_{,1}]}{c^4 - \sigma^2 v^2}, \tag{109}$$

and

$$\sigma_1(v,\varphi) = \frac{(\sigma^2(n-1)-c^2)(c^2-v^2)}{c(\sigma^2+c^2)(c^4-\sigma^2 v^2)}\,((c^4+\sigma^2 v^2)\varphi_{,1} + (\sigma^2+c^2)cv\varphi_{,0}), \tag{110}$$

and let $S(t)$ be the set of points at which the approximate solution is discontinuous at time t. Note that σ_0 is in the form $\alpha(v)\varphi_{,0} + \beta(v)\varphi_{,1}$. Then,

$$\mathrm{TV}_x(\sigma_0(v,\varphi)) = \int_{R-S(t)} \left(\alpha'(v) v_x \varphi_{,0} + \beta'(v) v_x \varphi_{,1} + \alpha(v)\,\frac{\partial \varphi_{,0}}{\partial x} + \beta(v)\,\frac{\partial \varphi_{,1}}{\partial x} \right) dx$$

$$+ \sum_{S(t)} |[\sigma(v,\varphi)]|. \tag{111}$$

Noting that $\alpha(v)$, $\alpha'(v)$, $\beta(v)$, and $\beta'(v)$ are bounded, and since $\mathrm{TV}_x(\varphi,_i)$, $i \in \{0,1\}$ is bounded, $\varphi_{,i}$, $i \in \{0,1\}$ is bounded as well, we deduce that

$$\mathrm{TV}_x(\sigma_0(v,\varphi)) \leqslant C_1 \mathrm{TV}_x(v) + C_2 \mathrm{TV}_x(\varphi_{,0}) + C_3 \mathrm{TV}_x(\varphi_{,1}). \tag{112}$$

Therefore, substituting (112) into (94) implies

$$\mathrm{TV}_x(\log(\rho^j)) \leqslant \mathrm{TV}_x(\log({}^R\rho^j)) \tag{113}$$

$$+ \int_{t_j}^{t} (C_1 \mathrm{TV}_x(v) + C_2 \mathrm{TV}_x(\varphi_{,0}) + C_3 \mathrm{TV}_x(\varphi_{,1}))dt. \tag{114}$$

Similarly,

$$\mathrm{TV}_x\left(\log\left(\frac{c+v}{c-v}\right)\right)^j \leqslant \mathrm{TV}_x^R\left(\log\left(\frac{c+v}{c-v}\right)\right)^j \tag{115}$$

$$+ C_4 \int_0^{\infty} (\mathrm{TV}_x(\varphi_{,0}) + \mathrm{TV}_x(\varphi_{,1}))dt. \tag{116}$$

It follows that if the bounds (94) hold, then the total variation on $\log((c+v)/(c-v))$ is uniformly bounded. In addition, since

$$\frac{\partial}{\partial x}\left(\log\left(\frac{c+v}{c-v}\right)\right) = \frac{v_x}{c^2-v^2}, \tag{117}$$

it follows that

$$\mathrm{TV}_x\left(\log\left(\frac{c+{}^Rv^j}{c-{}^Rv^j}\right)\right) = C\ \mathrm{TV}_x({}^Rv^j) \tag{118}$$

since (116) implies that velocity must be bounded away from the speed of light. Substituting (118) into (114), shows that as long as (94) holds, then we can achieve the bounds (7) sought in Theorem I.1. □

Note that a uniform total variation bound on $\log \rho$ was sufficient to demonstrate a uniform total variation bound on the solution U in the Minkowski metric case in Ref. 5 since the v and ρ are related through the geometry of the wave curves in the plane of Riemann invariants. Here, however, we must bound both the variations of $\log \rho$ and $\log((c+v)/(c-v))$ since on the ODE step the geometry of the wave curves is changed by the action of the ODE. Placing bounds on the variations of $\log \rho$ and $\log((c+v)/(c-v))$ is equivalent to bounding the arc length of the wave curves.

Also, we have loosely used the phrase ''bound the gravitational source term'' since more than gravitation is involved. In particular, $\varphi_{,0}$ is proportional to the divergence of the time vector field, and $\varphi_{,1}$ is proportional to the divergence of the coordinate vector field in the direction of fluid motion. To see this,

$$\mathrm{div}\left(\frac{\partial}{\partial x^0}\right) = \Gamma^i_{i0} \tag{119}$$

$$= \tfrac{1}{2}\log(-g)_{,0} \tag{120}$$

$$= n\varphi_{,0},$$

where n is the dimension of the space–time manifold. Similarly, $\mathrm{div}(\partial/\partial x^0) = n\varphi_{,1}$. These divergences depend on the gravitational field, but they also depend on the coordinate system we choose. Recall that we must choose coordinate vector fields in order to construct a grid on which to implement this scheme. With spherically symmetric coordinates on Minkowski space–time, the divergence of the radial vector field becomes unbounded at the origin. In this case, the ''source term'' becomes unbounded, even though there is *no* gravitation in Minkowski space. In this way it can be seen that coordinates can contribute equally to increases in total variation with gravitation.

VI. CONVERGENCE

In this section we show that a subsequence of approximate solutions, as described in Sec. III A, converges to a solution of the relativistic Euler equations. The analysis is similar to that in Ref. 3 with some modifications.

A function U is a weak solution to (36) if for each $\phi \in C_0^\infty(R\times R^+)$

$$0=\int_{-\infty}^{\infty}\int_0^{\infty}\{U\phi_t+F(U)\phi_x+G(x,t,U)\phi\}dt\ dx+\int_{-\infty}^{\infty}U(x,0^+)\phi(x,0)dx. \tag{121}$$

Define the error function for the approximate solutions as

$$\epsilon(\theta,\Delta x,\phi)\overset{\text{def}}{=}\int_{-\infty}^{\infty}\int_0^{\infty}\{U_{\Delta x}\phi_t+F(U_{\Delta z})\phi_x+G(x,t,U_{\Delta x})\phi\}dt\ dx + \int_{-\infty}^{\infty}U_{\Delta x}(x,0^+)\phi(x,0)dx.$$

It will be shown that there exists a sequence $\{\Delta x_n\}$ such that $\Delta x_n\to 0$ and $\epsilon(\theta,\Delta x_n,\phi)\to 0$ as $n\to\infty$ for all $\phi\in C_0^\infty(R\times(0,\infty))$.

It follows from the definition of the approximate solutions that

$$U_{\Delta x}(x,t)={}^RU_{\Delta x}(x,t)+\int_{t_j}^{t}G(x,\xi,\ {}^{\text{Temp}}U_{\Delta x}(\xi-t_j,\ {}^RU_{\Delta x}(x,t)))d\xi. \tag{122}$$

Multiplying (122) by ϕ_t and integrating gives

$$\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}U_{\Delta x}\phi_t\ dt\ dx=\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}{}^RU_{\Delta x}\phi_t\ dt\ dx + \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}\int_{t_j}^{t}G(x,\xi,\ {}^{\text{Temp}}U_{\Delta x}(\xi-t_j,\ {}^RU_{\Delta x}))d\xi\ \phi_t\ dt\ dx. \tag{123}$$

Since ${}^RU_{\Delta x}$ solves the Riemann problems for the homogeneous equation (54), we have

$$0=\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}{}^RU_{\Delta x}\phi_t\ dt\ dx+\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}F({}^RU_{\Delta x})\phi_x\ dt\ dx + \int_{-\infty}^{\infty}\{{}^RU_{\Delta x}(x,t_j^+)\phi(x,t_j)-{}^RU_{\Delta x}(x,t_{j+1}^-)\phi(x,t_{j+1})\}dx. \tag{124}$$

Using integration by parts, substituting (122), and using the fact that

$$\frac{d}{dt}\int_a^t f(\xi,t)d\xi=f(t,t)+\int_a^t\frac{\partial}{\partial t}f(\xi,t)d\xi, \tag{125}$$

we have

$$\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}\left\{\int_{t_j}^{t}G(x,\xi,\ {}^{\text{Temp}}U_{\Delta x}(\xi-t_j,\ {}^RU_{\Delta x}(x,t)))d\xi\right\}\phi_t\ dt\ dx$$
$$=\int_{-\infty}^{\infty}\left\{\int_{t_j}^{t_{j+1}}G(x,\xi,\ {}^{\text{Temp}}U_{\Delta x}(\xi-t_j,\ {}^RU_{\Delta x}((x,t_{j+1}^-)))d\xi\right\}\phi(x,t_{j+1})dx$$
$$-\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}\left\{\frac{\partial}{\partial t}\int_{t_j}^{t}G(x,\xi,\ {}^{\text{Temp}}U_{\Delta x}(\xi-t_j,\ {}^RU_{\Delta x}(x,t)))d\xi\right\}\phi\ dt\ dx$$

$$= \int_{-\infty}^{\infty} \{U_{\Delta x}(x,t_{j+1}^-) - {}^R U_{\Delta x}(x,t_{j+1}^-)\}\phi(x,t_{j+1})dx$$

$$- \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}} G(x,t,U_{\Delta x}(x,t))\phi \; dt \; dx$$

$$- \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}} \left\{ \int_{t_j}^{t} \frac{\partial}{\partial t} G(x,\xi, {}^{\text{Temp}}U_{\Delta x}(\xi - t_j, {}^R U_{\Delta x}(x,t)))d\xi \right\} \phi(x,t) dt \; dx. \quad (126)$$

Substituting (124) and (126) into (123) we have

$$\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}} U_{\Delta x}\phi_t \; dt \; dx = -\int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}} \{F({}^R U_{\Delta x})\phi_x + G(x,t,U_{\Delta x})\phi\} dt \; dx$$

$$+ \int_{-\infty}^{\infty} \{U_{\Delta x}(x,t_{j+1}^-)\phi(x,t_{j+1}) - U_{\Delta x}(x,t_j^+)\phi(x,t_j)\}dx$$

$$- \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}\int_{t_j}^{t} \frac{\partial}{\partial t} G(x,\xi, {}^{\text{Temp}}U_{\Delta x}(\xi - t_j, {}^R U_{\Delta x}(x,t)))d\xi \phi(x,t) dt \; dx.$$

Thus after summing over j we obtain

$$\epsilon(\theta,\Delta x,\phi) = \sum_{j=0}^{\infty} \epsilon_j(\theta,\Delta x,\phi) + \sum_{j=1}^{\infty} E_j(\theta,\Delta x,\phi), \quad (127)$$

where we define

$$\epsilon_j(\theta,\Delta x,\phi) \overset{\text{def}}{=} \int_{-\infty}^{\infty} \{U_{\Delta x}(x,t_j^-) - U_{\Delta x}(x,t_j^+)\}\phi(x,t_j)dx \quad (128)$$

and

$$E_{j+1}(\theta,\Delta x,\phi) \overset{\text{def}}{=} \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}} \{F(U_{\Delta x}) - F({}^R U_{\Delta x})\}\phi \; dt \; dx$$

$$- \int_{-\infty}^{\infty}\int_{t_j}^{t_{j+1}}\int_{t_j}^{t} \frac{\partial}{\partial t} G(x,\xi, {}^{\text{Temp}}U(\xi - t_j, {}^R U_{\Delta x}(x,t)))d\xi \; \phi \; dt \; dx,$$

where $U_{\Delta x}(x,t_0^-) \overset{\text{def}}{=} 0$. We may write

$$\epsilon_j(\theta,\Delta x,\phi) = \sum_i \int_{x_{i-1}}^{x_i} \{U_{\Delta x}(x,t_j^-) - U_{\Delta x}(x_i + \theta_j \Delta x, t_j^-)\}\phi(x,t_j)dx, \quad (129)$$

which shows that coordinatewise we have

$$|\epsilon_j(\theta,\Delta x,\phi)| \leq C\left(\sum_{i,l} |\gamma_{i,j}^l|\right)\Delta x |\phi|_\infty \leq C \; \text{TV} \; \Delta x |\phi|_\infty, \quad (130)$$

so

$$\left|\sum_{j=1}^{\infty} \epsilon_j(\theta,\Delta x,\phi)\right| \leq C \; \text{TV} T \frac{\Delta x}{\Delta t} |\phi|_\infty = C|\phi|_\infty, \quad (131)$$

where the support of ϕ is in $R \times [0,T)$. We also have that

$$\left| \int_0^1 \epsilon_j(\theta,\Delta x,\phi)\,d\theta_j \right| \leqslant C\left(\sum_{i,l} |\gamma_{i,l}^j| \right) \Delta x^2 |\phi_x|_\infty \leqslant C\ \mathrm{TV}\Delta x^2 |\phi_x|_\infty = C\Delta x^2 |\phi_x|_\infty .$$

Therefore, for $0<j_1<j_2$,

$$\left| \int \epsilon_{j_1}(\theta,\Delta x,\phi)\epsilon_{j_2}(\theta,\Delta x,\phi)\,d\theta \right| = \left| \int \left\{ \int_0^1 \epsilon_{j_2}(\theta,\Delta x,\phi)\,d\theta_{j_2} \right\} \epsilon_{j_1}(\theta,\Delta x,\phi)\,d\hat{\theta}_{j_2} \right|$$
$$\leqslant C\Delta x^3 |\phi|_\infty |\phi_x|_\infty ,$$

where $\hat{\theta}_{j_2}$ denotes the sequence obtained from θ by deleting the j_2th element. By (130) we see that

$$\int \epsilon_j(\theta,\Delta x,\phi)^2\,d\theta \leqslant C\Delta x^2 |\phi|_\infty^2. \tag{132}$$

Thus

$$\int \left(\sum \epsilon_j(\theta,\Delta x,\phi) \right)^2 d\theta = 2\sum_{j_1<j_2} \int \epsilon_{j_1}\epsilon_{j_2}\,d\theta + \sum_j \int \epsilon_j^2\,d\theta$$
$$\leqslant \frac{T^2}{\Delta t}\, C\ \mathrm{TV}_0^2 \Delta x^3 |\phi|_\infty |\phi_x|_\infty$$
$$+ \frac{T^2}{\Delta t}\, C\ \mathrm{TV}_0^2 \Delta x^2 |\phi|_\infty^2 = C\Delta x(|\phi|_\infty^2 + |\phi_x|_\infty^2).$$

We deduce that for each $\phi \in C_0^\infty(R\times(0,\infty))$ there exists a sequence $\Delta x_m \to 0$ and a set $A \subset \Pi_{i=1}^\infty [0,1]$ with measure one such that if $\theta \in A$, then

$$\sum_j \epsilon_j(\theta,\Delta x_m,\phi) \to 0, \tag{133}$$

as $\Delta x_m \to 0$. Now, let $\{\phi_n\}_{n\in N}$ be a countable dense set in $C_0^\infty(R\times(0,\infty))$ with respect to the norm $|\cdot|_\infty$. After refining Δx_m by a diagonal argument and after taking the appropriate countable intersection of sets A_n of measure one, we can conclude that there exists a sequence $\Delta x_m \to 0$ and a set $A \subset \Pi_{i=1}^\infty [0,1]$ with measure one such that if $\theta \in A$ and $n \in N$, then

$$\sum_j \epsilon_j(\theta,\Delta x_m,\phi_n) \to 0, \tag{134}$$

as $m \to \infty$. It follows from (131) that if $\theta \in A$ and $\phi \in C_0^\infty(R\times(0,\infty))$, then

$$\sum_j \epsilon_j(\theta,\Delta x_m,\phi) \to 0. \tag{135}$$

It remains to show that the second term on the right-hand side of Eq. (127) goes to zero as $\Delta x \to 0$. Note that by (122), the Mean Value Theorem for Integrals,

$$|U_{\Delta x} - {}^R U_{\Delta x}| = \left| \int_{t_j}^t G \right| = G(x,\xi_j)\Delta t. \tag{136}$$

Therefore,

$$\int_{-\infty}^\infty \int_{t_j}^{t_{j+1}} |U_{\Delta x} - {}^R U_{\Delta x}|\,|\phi| = \Delta t^2 |\phi|_\infty \int_{-\infty}^\infty G(x,\xi_j)\,dx. \tag{137}$$

Since the source term is assumed to be $L^1(R\times R^+)$,

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}|U_{\Delta x}-{}^{R}U_{\Delta x}||\phi|_{\infty}\leqslant\Delta t^{2}|\phi|_{\infty}\sum_{j=0}^{\infty}\int_{-\infty}^{\infty}G(x,\xi_{j})dx \tag{138}$$

$$\leqslant C\Delta t^{2}|\phi|_{\infty}. \tag{139}$$

These estimates may be interpreted coordinatewise. We want the same bounds on

$$|F(U_{\Delta x})-F({}^{R}U_{\Delta x})|. \tag{140}$$

Since $F(U)=(U^{1},U^{2})$, we have the estimate (139) on the first coordinate of $F(U)$. The Jacobian $dF(U)$ is

$$\frac{-c^{2}(\sigma^{2}-v^{2})}{(c^{4}-\sigma^{2}v^{2})}, \tag{141}$$

which is bounded. We make this calculation by setting $F(U)=\Psi(\Phi^{-1}(U))$, where $\Phi(U)$ is defined in Lemma II.2, and $\Psi(\rho,v)$ is defined by $F(U)=\Psi(\Phi^{-1}(U))$. By the Mean Value Theorem,

$$|F(U_{\Delta x})-F({}^{R}U_{\Delta x})|\leqslant C|U_{\Delta x}-{}^{R}U_{\Delta x}|, \tag{142}$$

from which we conclude

$$\left|\int_{-\infty}^{\infty}\int_{t_{j}}^{t_{j+1}}\{F(U_{\Delta x})-F({}^{R}U_{\Delta x})\}\phi\ dt\ dx\right|\leqslant C\Delta t^{2}|\phi|_{\infty}. \tag{143}$$

In addition, for $x\in[x_{i-1},x_{i}]$,

$$\left|\int_{t_{j}}^{t_{j+1}}\int_{t_{j}}^{t}\frac{\partial}{\partial t}\,G(x,\xi,U(\xi-t_{j},\,{}^{R}U_{\Delta x}(x,t)))d\xi\ \phi\ dt\right| \tag{144}$$

$$=\left|\int_{t_{j}}^{t_{j+1}}\int_{t_{j}}^{t}G_{U}U_{\mathrm{Temp}_{U}}{}^{R}U_{\Delta x,t}\phi\ d\xi\ dt\right| \tag{145}$$

$$\leqslant C\Delta t|\phi|_{\infty}\int_{t_{j}}^{t_{j+1}}|{}^{R}U_{\Delta x,t}|dt$$

$$\leqslant C\Delta t|\phi|_{\infty}(|\gamma_{i-1,j}^{2}|+|\gamma_{i,j}^{1}|). \tag{146}$$

Hence,

$$\left|\int_{-\infty}^{\infty}\int_{t_{j}}^{t_{j+1}}\int_{t_{j}}^{t}\frac{\partial}{\partial t}\,G(x,t,{}^{\mathrm{Temp}}U(\xi-t_{j},{}^{R}U_{\Delta x}(x,t)))\phi\ d\xi\ dx\ dt\right|$$

$$\leqslant C\Delta t\Delta x|\phi|_{\infty}\left(\sum_{i,l}\ |\gamma_{i,j}^{l}|\right)$$

$$\leqslant C(\mathrm{TV})\Delta t\Delta x|\phi|_{\infty}$$

$$\leqslant C\Delta t\Delta x|\phi|_{\infty}. \tag{147}$$

It follows from (143) and (147) that

$$\left|\sum_{j=1}^{\infty}E_{j}(\theta,\Delta x,\phi)\right|\leqslant C\cdot T\Delta t|\phi|_{\infty}+C\cdot T\Delta x|\phi|_{\infty}. \tag{148}$$

Thus, from (135) and (148) we can conclude that if $\theta\in A$ and $\phi\in C_{0}^{\infty}(R\times R^{+})$, then

$$\epsilon(\theta,\Delta x_m,\phi)\to 0 \tag{149}$$

as $m\to\infty$.

Now, choose $T>0$ and let $\{t_i\}_{i\in N}\subset(0,T)$ be countable and dense and let $\theta\in A$ be fixed. Since

$$\mathrm{TV}_{x\in R}U_{\Delta x}(\cdot,t)\leqslant C \tag{150}$$

for $t\in(0,\infty)$, Helly's theorem (Ref. 16 p. 372) implies that there exists a function $U(\cdot,t)$: $\{t_i\}\to L^1(R)$ such that (after further refining $\{\Delta x_m\}$) for $i\in N$, $U_{\Delta x_m}(\cdot,t_i)\to U(\cdot,t_i)$, in $L^1(R)$ as $m\to\infty$.

To proceed, we must prove that the approximates are L_1 Lipschitz in time.

Lemma VI.1: There exists a constant C, such that

$$\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_2)-U_{\Delta x}(x,t_1)|dx\leqslant C\{|t_2-t_1|+\Delta t\}. \tag{151}$$

Proof: Suppose that

$$0\leqslant t_{m-1}\leqslant t_1<\cdots<t_M<t_2\leqslant t_{M+1}. \tag{152}$$

Then

$$\begin{aligned}\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_2)-U_{\Delta x}(x,t_1)|dx\leqslant{}&\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_m^-)-U_{\Delta x}(x,t_1)|dx+\sum_{j=m}^{M-1}\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_{j+1}^-)\\&-U_{\Delta x}(x,t_j^+)|dx+\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_2)-U_{\Delta x}(x,t_M^+)|dx\\&+\sum_{j=m}^{M}\int_{-\infty}^{\infty}|U_{\Delta x}(x,t_j^+)-U_{\Delta x}(x,t_j^-)|dx.\end{aligned} \tag{153}$$

For $s,\ t\in[t_j,t_{j+1}]$ we have that

$$\begin{aligned}\int_{-\infty}^{\infty}|U_{\Delta x}(x,s)-U_{\Delta x}(x,t)|dx\leqslant{}&\int_{-\infty}^{\infty}|U_{\Delta x}(x,s)-{}^RU_{\Delta x}(x,s)|dx+\int_{-\infty}^{\infty}|{}^RU_{\Delta x}(x,s)-{}^RU_{\Delta x}(x,t)|dx\\&+\int_{-\infty}^{\infty}|{}^RU_{\Delta x}(x,t)-U_{\Delta x}(x,t)|dx.\end{aligned} \tag{154}$$

But

$$\int_{x_{i-1}}^{x_i}|{}^RU_{\Delta x}(x,s)-{}^RU_{\Delta x}(x,t)|dx\leqslant C\Delta x(|\gamma_{i-1,j}^2|+|\gamma_{i-1,j}^1|). \tag{155}$$

Thus

$$\int_{-\infty}^{\infty}|{}^RU_{\Delta x}(x,s)-{}^RU_{\Delta x}(x,t)|dx\leqslant C\Delta x\hat{L}(j). \tag{156}$$

Moreover, by (122) it follows that

$$\max_x|U_{\Delta x}(x,s)-{}^RU_{\Delta x}(x,s)|\leqslant C|s-t_j|, \tag{157}$$

for $s\in[t_j,t_{j+1}]$, so by (154)

$$\int_{-\infty}^{\infty}|U_{\Delta x}(x,s)-U_{\Delta x}(x,t)|dx\leqslant C\Delta xL(j)+C\Delta t. \tag{158}$$

In addition, by (157)

$$\int_{x_{i-1}}^{x_i} |U_{\Delta x}(x,t_j^+) - U_{\Delta x}(x,t_j^-)|dx = \int_{x_{i-1}}^{x_i} |U_{\Delta x}(x+\theta_j \Delta x, t_j^-) - U_{\Delta x}(x,t_j^-)|dx$$

$$\leqslant C\Delta x(|\gamma^2_{i-1,j-1}| + |\gamma^1_{i,j-1}|) + C\Delta x \Delta t,$$

so

$$\int_{-\infty}^{\infty} |U_{\Delta x}(x,t_j^+) - U_{\Delta x}(x,t_j^-)|dx \leqslant C\Delta x Ł(j-1) + C\Delta t. \tag{159}$$

Now, putting (158) and (159) into (153) and summing yields

$$\int_{-\infty}^{\infty} |U_{\Delta x}(x,t_2) - U_{\Delta x}(x,t_1)|dx \leqslant C\Delta x(M-m+1)\mathrm{TV}_0 + C\Delta t(M-m+1) \leqslant C\{|t_2-t_1| + \Delta t\}.$$

Now, it follows from Lemma VI.1 that $U_{\Delta x_m}(\cdot,t)$ converges to $U(\cdot,t)$ in $L^1(R)$ for each $t \in (0,T)$ and

$$\|U(\cdot,t_2) - U(\cdot,t_1)\|_{L^1(R)} \leqslant C|t_2-t_1|, \tag{160}$$

for $t_1,\ t_2 \in (0,T)$. Hence,

$$\int_{-\infty}^{\infty}\int_0^{\infty} \{U_{\Delta x_m}\phi_t + F(U_{\Delta x_m})\phi_x + G(x,t,U_{\Delta x_m})\phi\}dt\ dx$$

$$\to \int_{-\infty}^{\infty}\int_0^{\infty} \{U\phi_t + F(U)\phi_x + G(x,t,U)\phi\}dx\ dt$$

as $m \to \infty$.

Since $\epsilon(\theta, \Delta x_m, \phi) \to 0$, U is a weak solution to (36). Furthermore, since $U_{\Delta x}(0) \to U_0$ in $L^1(R)$, (160) implies that $\lim_{t\to 0} U(\cdot,t) = U_0$.

ACKNOWLEDGMENT

The author was supported in part by a research grant from La Sierra University.

[1] J. Smoller, *Shock Waves and Reaction Diffusion Equations* (Springer-Verlag, New York, 1983).
[2] J. Groah, ''Shock-wave solutions of the general relativistic Euler equations'' (unpublished).
[3] M. Luskin and J. B. Temple, ''The existence of a global weak solution to the nonlinear waterhammer problem,'' Commun. Pure Appl. Math. **35**, 697–735 (1982).
[4] D. Marchesin and P. J. Paes-Leme, ''Shocks in gas pipelines,'' SIAM (Soc. Ind. Appl. Math.) J. Sci. Stat. Comput. **4**, 105–16 (1983).
[5] J. Smoller and J. B. Temple, ''Global solutions of the relativistic Euler equations,'' Commun. Math. Phys. **156**, 67–99 (1993).
[6] T. Nishida, ''Global solution for an initial boundary value problem of a quasilinear hyperbolic system,'' Proc. Jpn. Acad. **44**, 642–646 (1968).
[7] A. M. Anile, *Relativistic Fluids and Magneto-Fluids* (Cambridge University Press, New York, 1989).
[8] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972).
[9] B. O'Neill, *Semi-Riemannian Geometry* (Academic, London, 1983).
[10] S. Goldberg, *Curvature and Homology* (Dover, New York, 1982).
[11] D. Laugwitz, *Differential Geometry* (Academic, New York, 1965).
[12] M. Carmeli, *Classical Fields* (Wiley, New York, 1982).
[13] M. Spivak, *Differential Geometry* (Publish or Perish, Inc., Houston, Texas, 1979).
[14] J. Glimm, ''Solutions in the large for nonlinear hyperbolic systems of equations,'' Commun. Pure Appl. Math. **18**, 697–715 (1965).
[15] P. D. Lax, ''Hyperbolic systems of conservation laws, II,'' Commun. Pure Appl. Math. **10**, 537–566 (1957).
[16] Kolmogorov-Fomin, *Introductory Real Analysis* (Dover, Mineola, 1970).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

Rest frame system for asymptotically flat space–times

Osvaldo M. Moreschi[a] and Sergio Dain[b]
FaMAF, Ciudad Universitaria, Universidad Nacional de Córdoba, (5000) Córdoba, Argentina

(Received 7 July 1997; accepted for publication 8 July 1998)

The notion of center of mass for an isolated system has been previously encoded in the definition of the so-called nice sections. In this article we present a generalization of the proof of existence of solutions to the linearized equation for nice sections, and formalize a local existence proof of nice sections relaxing the radiation condition. We report on the differentiable and non-self-crossing properties of this family of solutions. We also give a proof of the global existence of nice sections. © *1998 American Institute of Physics.* [S0022-2488(98)00912-8]

I. INTRODUCTION

There is an increasing interest in obtaining results from general relativity that could be applied to real systems. In particular, the possibility of detecting gravitational waves in the near future motivates the research for the description of systems containing very compact objects. In modeling these systems as isolated ones, it is possible to distinguish at least two very different approaches for tackling this problem. In one of them the description is attempted by introducing data, with some criteria, for the numeric solution of the field equation. In the other the description is attempted by trying to reduce the infinite degrees of freedom of the fields to some finite set, analogous to what is done in the particlelike description in Newtonian or in post-Newtonian dynamics. If this second approach has any chance of success, it will be because it is possible to ascribe, in a geometric invariant way, the finite degrees of freedom to the physical system one wants to describe.

It is important to emphasize that the present indication of the existence of gravitational radiation comes from the usage of the balance equation in the description of the binary pulsar PSR 1913+16. The applicability of the present derivations of the balance equation to the binary pulsar system is still a matter of discussion. However, even in the case of having a formally complete derivation of this equation valid at an instant of time, one is still faced with the issue of its validity for long periods of time, or for systems involving strong gravitational fields and possible high velocities. The nature of the problem can be seen by noting that the balance equation relates time derivatives of the energy of the systems with time derivatives of the quadrupole moments of the system. Although the total energy of the system could be related with the unambiguously defined Bondi momentum; the notion of quadrupole moment has meaning only after one has settled down the issue of supertranslation freedom. Our construction of the center of mass sections at future null infinity through the definition of nice sections provides a procedure to give a precise meaning to concepts such as angular momentum and multipole moments.

Therefore, from the study of gravitating isolated systems in the vicinity of the asymptotic region, i.e., future null infinity, one concludes that it is essential[1] to introduce a family of sections of scri, embodying the notion of generalized center of mass. This is the analog of having a frame of reference with respect to which one can study the asymptotic fields. With this aim in Ref. 2 the ''supercenter of mass'' at future null infinity was introduced; which was defined in terms of the so-called *nice sections*. It was indicated in this reference that the notion of center of mass can be attained by requiring an asymptotic nice section for the retarded time $u \to -\infty$ to satisfy that the angular momentum coincides with the intrinsic angular momentum. Then, if the subsequent nice sections are constructed by infinitesimal time translations which are parallel to the Bondi momen-

[a]Member of CONICET; Electronic mail: moreschi@fis.uncor.edu
[b]Fellowship holder from CONICOR; Electronic mail: dain@fis.uncor.edu

 © 1998 American Institute of Physics

tum, one obtains a one parameter family of nice sections at future null infinity embodying the notion of center of mass.

In this work we study some mathematical properties of nice sections. In Sec. II we review the nice section equation and some of its basic properties. We present a more general proof of the existence of solutions to the linearized equation for nice sections, and formalize a local existence proof of nice sections in Sec. III. The differentiable properties of nice sections are studied in Sec. IV. We report on the desired non-self-crossing property of this family of solutions in Sec. V. In Sec. VI we present a proof of global existence of nice sections. Finally, several complementary results are presented in three Appendices.

II. THE NICE SECTION EQUATION

We will start by recalling the nice section equation. Given a section S of future null infinity, one can always find a Bondi coordinate (see, e.g., Ref. 3) system $(u,\zeta,\bar{\zeta})$ such that S is given by the equation $u=0$. The supermomentum at S is defined[2] by

$$P_{lm}(S)\equiv -\frac{1}{\sqrt{4\pi}}\int_S Y_{lm}(\zeta,\bar{\zeta})\Psi dS^2, \tag{1}$$

where

$$\Psi\equiv\Psi_2+\sigma\dot{\bar{\sigma}}+\eth^2\bar{\sigma}, \tag{2}$$

with Ψ_2 and σ being the leading order asymptotic behavior of the second Weyl tensor component and the Bondi shear, respectively, and where we are using the Geroch–Held-Penrose (GHP) notation[4] for the edth operator of the unit sphere. It is probably worth noting that in a Bondi system[3] the quantity Ψ is a real function on future null infinity (see, e.g., Eqs. (3.35) and (5.20) in Ref. 5).

Given another section $\tilde{S}$ of scri, one can find another Bondi system $(\tilde{u},\tilde{\zeta},\bar{\tilde{\zeta}})$ such that $\tilde{S}$ is given by $\tilde{u}=0$. One can further demand for that Bondi system to be aligned with the total (Bondi) momentum at $\tilde{S}$, i.e., that the $l=1$ components of the momentum are zero:

$$\tilde{P}_{l=1,m}(\tilde{S})=0.$$

The relation between the new and the original Bondi system is given by the Bondi–Metzner–Sach (BMS) transformation

$$\tilde{u}=K(\zeta,\bar{\zeta})(u-\gamma(\zeta,\bar{\zeta})), \tag{3}$$

$$\bar{\zeta}=\frac{a\zeta+b}{c\bar{\zeta}+d} \tag{4}$$

with

$$ad-bc=1,\quad K=\frac{(1+\zeta\bar{\zeta})}{(a\zeta+b)(\bar{a}\bar{\zeta}+\bar{b})+(c\zeta+d)(\bar{c}\bar{\zeta}+\bar{d})},$$

where a, b, c and d are complex constants. Note that $\tilde{S}$ can also be determined by $u=\gamma(\zeta,\bar{\zeta})$.

The section $\tilde{S}$ is said to be of the nice[2] type if

$$\tilde{P}_{lm}(\tilde{S})=0\quad\text{for}\quad l\neq 0; \tag{5}$$

that is all the ''spacelike'' components of the supermomentum, when calculated with respect to the adapted Bondi system, are zero; and the only nonvanishing one, namely $\tilde{P}_{00}(\tilde{S})$, coincide with the total Bondi mass at $\tilde{S}$.

This is the original version of the nice section equation; however its information is hidden in the simplicity of its form. Since the nice section can be determined by the supertranslation $\gamma(\zeta,\bar{\zeta})$, mentioned above, it is more useful to restate the nice section equation in terms of an equation for $\gamma(\zeta,\bar{\zeta})$.

The supermomentum at $\tilde{S}$ can also be expressed in terms of the original Bondi coordinate system:[2]

$$\tilde{P}_{lm}(\tilde{S}) \equiv -\frac{1}{\sqrt{4\pi}} \int_{\tilde{S}} Y_{lm}(\tilde{\zeta},\bar{\tilde{\zeta}})\tilde{\Psi}\, d\tilde{S}^2$$

$$= -\frac{1}{\sqrt{4\pi}} K_{lm}^{l'm'} \int_{\tilde{S}=(u=\gamma)} Y_{l'm'}(\zeta,\bar{\zeta})[\Psi(u=\gamma) - \eth^2\bar{\eth}^2\gamma] dS^2, \tag{6}$$

where the matrix $K_{lm}^{l'm'}$ is the transformation matrix of the generators of supertranslations, that is:

$$\tilde{p}_{lm} = K_{lm}^{l'm'} p_{l'm'}, \tag{7a}$$

with

$$p_{lm} = Y_{lm}(\zeta,\bar{\zeta})\frac{\partial}{\partial u}. \tag{7b}$$

The fourth-order edth operator appearing in Eq. (6) comes from the transformation properties of the quantities entering into the definition of Ψ; in particular, the term $\eth^2\bar{\sigma}$ transforms under the law:

$$\tilde{\eth}^2\tilde{\bar{\sigma}} = \frac{1}{K^3}(\eth^2\bar{\sigma} - \eth^2\eth^2\gamma), \tag{8}$$

and similarly, after a straightforward but long calculation, one can prove that Ψ transforms as

$$\tilde{\Psi} = \frac{1}{K^3}(\Psi - \eth^2\bar{\eth}^2\gamma), \tag{9}$$

where the transformation of Bondi systems under BMS transformations, appearing in the appendix of Ref. 3, have been used in deriving the last two relations. The nice section equation can be understood as a condition for $\gamma(\zeta,\bar{\zeta})$ and K.

It is important to note that there exists different choices for Ψ in the literature; because by adding any function with $l \geqslant 2$ (e.g., $\eth^2\bar{\sigma}$) one obtains the same expression for the Bondi momentum. But the supermomentum (2) is the only one that has the simple transformation (9). We mention also (but we do not make use of this) that the time derivative of (2) have also a very simple form:

$$\dot{\Psi} = \dot{\sigma}\dot{\bar{\sigma}},$$

the other definitions of supermomentum have a flux that is not positive definite.

Although in Eq. (6) one has an expression that involves the original Bondi coordinate system $(u,\zeta,\bar{\zeta})$, the difficulty with it is the appearance of the infinite transformation matrix $K_{lm}^{l'm'}$; therefore it is convenient to seek a simpler expression. It was indicated in Ref. 2 that Eq. (5) can be expressed by

$$\eth^2\bar{\eth}^2\gamma = \Psi(u=\gamma,\zeta,\bar{\zeta}) + K(\gamma;\zeta,\bar{\zeta})^3 M(\gamma), \tag{10}$$

where the function on the sphere K is the conformal factor of the BMS transformation that aligns the $\tilde{p}_{00}$ with the Bondi momentum at $u=\gamma(\zeta,\bar{\zeta})$; which justifies the notation indicating the γ dependence on it; and finally the total mass M is also calculated at $\tilde{S}$. Equation (10) is our main

equation that is studied throughout the rest of this article. The validity of Eq. (10) can be deduced either from the discussion in Ref. 2, using the properties of the matrix $K_{lm}^{l'm'}$; or it can be deduced from Eqs. (5) and (9); since the nice section equation (5) can be understood as requiring that $\tilde{\Psi}$ be a constant when evaluated at the section $u=\gamma(\zeta,\bar{\zeta})$; and this constant is precisely the negative of the Bondi mass M evaluated on this section.

Clearly $K(\gamma;\zeta,\bar{\zeta})$ and $M(\gamma)$ are determined by the value of the total momentum at $\tilde{S}$, which can explicitly be given by

$$P^{\underline{a}}(\tilde{S})=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\Psi(u=\gamma,\zeta,\bar{\zeta})dS^2, \tag{11}$$

where $\underline{a}=0,1,2,3$, and the vector $l^{\underline{a}}$ is given in terms of the first spherical harmonics by

$$l^{\underline{a}}=\sqrt{4\pi}\left(Y_{00},-\frac{1}{\sqrt{6}}(Y_{11}-Y_{1-1}),\frac{i}{\sqrt{6}}(Y_{11}+Y_{1-1}),\frac{1}{\sqrt{3}}Y_{10}\right). \tag{12}$$

Note that one can also express the total momentum at $\tilde{S}$ by

$$\begin{aligned}P^{\underline{a}}(\tilde{S})&=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\left(\int_0^{\gamma}\dot{\Psi}(u',\zeta,\bar{\zeta})du'+\Psi(u=0,\zeta,\bar{\zeta})\right)dS^2\\&=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\int_0^{\gamma}\dot{\Psi}(u',\zeta,\bar{\zeta})du'\;dS^2+P^{\underline{a}}(S),\end{aligned} \tag{13}$$

where a dot means u derivative; this shows a more explicit dependence of $P^{\underline{a}}(\tilde{S})$ on $\gamma(\zeta,\bar{\zeta})$ and the momentum flux. Then, the conformal factor K can be given by

$$K(\gamma;\zeta,\bar{\zeta})=\frac{M(\gamma)}{P^{\underline{a}}(\tilde{S})l^{\underline{b}}(\zeta,\bar{\zeta})\eta_{\underline{a}\underline{b}}}, \tag{14}$$

where the Lorentzian metric $\eta_{\underline{a}\underline{b}}$ was defined in Ref. 3 and

$$M(\gamma)=\sqrt{P^{\underline{a}}(\tilde{S})P^{\underline{b}}(\tilde{S})\eta_{\underline{a}\underline{b}}}. \tag{15}$$

Then, it is observed that Eq. (10) is an elliptic, nonlinear, integro-differential equation for $\gamma(\zeta,\bar{\zeta})$. Let us mention some properties of this equation. Calling D the differential operator appearing on the left-hand side of Eq. (10), we have that:

$$D\equiv\eth^2\bar{\eth}^2=\tfrac{1}{4}\Delta^2-\tfrac{1}{2}\Delta, \tag{16}$$

where Δ is the Laplacian of the unit sphere. Note that when the operator D is applied to the spherical harmonics it gives[3]

$$DY_{lm}=\frac{(l-1)l(l+1)(l+2)}{4}Y_{lm}. \tag{17}$$

This means that the left-hand side of (10) has an expansion in spherical harmonics with $l\geqslant 2$; and in order to be a consistent equation the right-hand side must also have this property for an arbitrary Ψ (see Appendix 1 for a proof of this property).

In the stationary case, $\Psi=\Psi(\zeta,\bar{\zeta})$, the equation becomes linear and always has solutions.[2] Let γ_0 be the *a* solution of the stationary case, then if $x(\zeta,\bar{\zeta})$ is such that $Dx=0$ we have that γ_0+x is also a solution. Then, in the stationary case there exists a four-parameter family of solution (the real solutions of $Dx=0$ are linear combinations of the spherical harmonics with $l\leqslant 1$, i.e., they are determined by four real numbers). To illustrate this situation let us consider for example the Schwarzschild space–time. In the usual coordinates and null tetrad the equation

simplifies considerably since $\Psi=-M$ and $K=1$; where M is the Schwarzschild mass. Therefore, in this case, the right-hand side of Eq. (10) vanishes, and so $\gamma_0=0$ is a nice section. This means that the usual coordinate system is a ''global rest frame'' system; and one still has the freedom of the translation x's. If instead we had begun with an arbitrary system in the Schwarzschild space–time, then we would have $\Psi\neq -M$ and we would need a nontrivial γ_0 to solve the nice section equation (10). This γ_0 is the supertranslation we have to make in order to come back to ''global rest frame.'' In order to gain the notion of center of mass from this family of solutions, one impose on the ''first'' section at $u=-\infty$ the condition that the angular momentum coincide with the intrinsic angular momentum; this eliminates a three-dimensional freedom in the choice of the translation x's. Then, if from the first nice section we only allow for infinitesimal translations that are parallel to the total momentum, we generate from $u=-\infty$ the ''center of mass system'' with the solutions of the nice section equation.

What happens in general, with a radiative asymptotically flat space–time? Let us first put a conjecture:

Conjecture: Under mild hypothesis for Ψ, there exists solutions γ of (10) of the form $\gamma=x+y$ for each choice of translation x ($Dx=0$), and with y a supertranslation (i.e., with spherical harmonics $l\geqslant 2$ expansion). The solution is C^2 in the sphere and if x_1 and x_2 are time constant translations, with x_1 in the future of x_2; that is,

$$x_1\geqslant x_2$$

then:

$$\gamma_1\geqslant\gamma_2 (*)$$

where γ_1 is the solution corresponding to x_1 and γ_2 that corresponding to x_2.

If this conjecture were true, and by using the selection of the ''initial'' nice section as it was explained for the Schwarzschild case, we would have that for any asymptotically flat space–time there exists one globally defined ''rest frame'' at null infinity, that have the appropriate properties to be interpreted as the ''center of mass'' system of the space–time. We note that the last property (*) requires that the sections defining the center of mass system must be a one-parameter family, without self-crossing behavior. We also remark that the differentiability property means that they are smooth sections of scri.

Of course this conjecture is very difficult to prove, because Eq. (10) is a complicated nonlinear equation, but here we can give an argument[2] (not a proof) that can help one to understand why one might think the conjecture is true.

The idea is to construct a perturbation series beginning with an arbitrary translation x. Let,

$$\gamma_0=x,\quad \gamma_n=x+\sum_{k=1}^{n} y_k,$$

where y_n is the solution of the linear equation:

$$\eth^2\bar{\eth}^2 y_n=\Psi(u=\gamma_{n-1},\zeta,\overline{\zeta})+K(\gamma_{n-1};\zeta,\overline{\zeta})^3 M(\gamma_{n-1}).$$

If the series converges we have for each translation x a solution γ.

In the following part of this paper we prove the following results that indicate to the validity of the conjecture.

(1) ''Local'' version of the conjecture: Given a space–time which satisfies, using geometric units, $|\partial\Psi/\partial u|\leqslant\sqrt{27/4}$(**) and assuming there exists a solution γ_0, then there exists locally for each infinitesimal translation a solution γ, which is C^2 and satisfies (*).

This is done in Secs. III, IV, and V. Let us note that the condition (**) means that the space–time does not admit arbitrary amount of radiation; however, for realistic astrophysical systems it turns out that (**) is always true (see Ref. 2).

(2) If Ψ is near the stationary case, then the there exist a global four-parameter family of solution.

This result is proved in Sec. VI.

III. LOCAL EXISTENCE OF SOLUTIONS

The result of the local existence of solutions will be constructed around the implicit function theorem; therefore appropriate Banach spaces must be introduced. We begin by introducing some notation.

For any pair of regular functions on the spheres a and b let the inner product $\langle a|b\rangle$ be defined by

$$\langle a|b\rangle = \int \overline{a}b \; dS^2, \tag{18}$$

and associated to this let the norm of a function a be given by

$$\|a\|^2 = \langle a|a\rangle, \tag{19}$$

that is, $\|\ \|$ is the standard L^2 norm on the unit sphere S^2.

We will next introduce three sets of real regular functions on the sphere. Let us start with the set:

$$X = \{x(\zeta,\overline{\zeta}) \in L^2(S^2)/Dx = 0\}, \tag{20}$$

note that the real functions $x(\zeta,\overline{\zeta})$ are determined by four real numbers.

Let Z be the set of real functions on the sphere with finite norm and residing in the orthogonal complement of X; equivalently,

$$Z = X^{\perp} = \{z(\zeta,\overline{\zeta}) \in L^2(S^2)/\langle x|z\rangle = 0 \,\forall x \in X\}. \tag{21}$$

Note that if γ is in L^2, then it can be decomposed as a sum of two terms; one in X and the other in Z; that is $\gamma = x + z$, $x \in X$, $z \in Z$. Furthermore, it is convenient to define the projector operator T, such that $T(\gamma) = T(x+z) = x$; that is T projects to the subspace X and therefore satisfies $T^2 = T$.

It is clear from the definitions that X and Z are Hilbert spaces.

Let us now define the following inner product. For any four time differentiable functions y_1 and y_2, in Z, let $\langle|\rangle_D$ be defined by

$$\langle y_1|y_2\rangle_D \equiv \int D\overline{y_1} \; Dy_2 \; dS^2, \tag{22}$$

the fact that this is really an inner product in Z follows from the existence of an inverse of the operator D; as it is deduced from Eq. (17). It is natural to define the norm

$$\|y\|_D^2 \equiv \langle y|y\rangle_D . \tag{23}$$

As it was stated in Ref. 2 it is not difficult to see that

$$\|y\| \leqslant \|y\|_D . \tag{24}$$

Let us define now the space of real functions on the sphere Y by

$$Y = \{y(\zeta,\overline{\zeta}) \in Z/\|y\|_D < \infty \text{ completed with respect to this norm}\}. \tag{25}$$

If one expresses the functions x and y in terms of spherical harmonics one has

$$x = \sum_{\substack{l=0 \\ |m|\leqslant l}}^{1} x^{lm} Y_{lm}(\zeta,\overline{\zeta}) \tag{26}$$

and

$$y=\sum_{\substack{l=2\\|m|\leqslant l}}^{\infty} y^{lm}Y_{lm}(\zeta,\zeta), \tag{27}$$

for some constant coefficients x^{lm} and y^{lm}.

Note that the space Y is also a Hilbert space.

Given regular radiation data at future null infinity, let the real function $f(x,y)$ on the sphere be defined in terms of functions $x\in X$ and $y\in Y$ by the expression:

$$f(x,y)=Dy-(\Psi(u=x+y,\zeta,\bar{\zeta})+K(x+y;\zeta,\bar{\zeta})^3M(x+y)); \tag{28}$$

where the radiation data enter into the determination of Ψ, K and M at $u=\gamma=x+y$; and it is easily seen that the condition $f(x,y)=0$ coincides with the nice section equation. It is also a property of any $f(x,y)$, so constructed, that it lives in the orthogonal complement space of X; in other words, for any f and any x' in X one has $\langle f,x'\rangle=0$; see Appendix A for a proof of this assertion. This means that f is expressible in terms of spherical harmonics Y_{lm} with $l\geqslant 2$; that is,

$$f=\sum_{\substack{l=2\\|m|\leqslant l}}^{\infty} f^{lm}Y_{lm}(\zeta,\zeta). \tag{29}$$

It is worth mentioning that the decomposition of the supertranslation γ as the sum $\gamma=x+y$ is a natural one due to the fact that $Dx=0$ for any x in X.

Since Ψ, K and M are completely determined by the free data at future null infinity, one can see that for smooth data, any f will have finite norm and so any f, so defined, belongs to Z.

Let us now consider the difference

$$\begin{aligned} &f(x,y+\delta y)-f(x,y)\\ &\quad=D\,\delta y-\left(\int_{x+y}^{x+y+\delta y}\dot{\Psi}(u',\zeta,\bar{\zeta})du'+K(x+y+\delta y;\zeta,\bar{\zeta})^3M(x+y+\delta y)\right.\\ &\qquad\left.-K(x+y;\zeta,\bar{\zeta})^3M(x+y)\right). \end{aligned}$$

From the definition (28) one can prove that

$$f(x,y+\delta y)-f(x,y)=f_y(x,y)\,\delta y+O(\delta y^2) \tag{30}$$

with

$$f_y(x,y)\,\delta y=D\,\delta y-\left[\dot{\Psi}(u=x+y,\zeta,\bar{\zeta})\,\delta y+K^3\left(\frac{4P_{\underline{a}}}{M}-3Kl_{\underline{a}}(\zeta,\bar{\zeta})\right)\delta P^{\underline{a}}\right], \tag{31}$$

where $P_{\underline{a}}$, M, and K are evaluated at $u=x+y$, and $\delta P^{\underline{a}}$ is given by

$$\delta P^{\underline{a}}=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\dot{\Psi}(u=x+y,\zeta,\bar{\zeta})\,\delta y\;dS^2, \tag{32}$$

therefore by definition, $f_y(x,y)$ is the Fréchet derivative[6] of the map f in terms of the argument y.

Let us observe that for smooth data $f_y(x,y)$ is continuous in x, and

$$\delta z=f_y(x,y)\,\delta y \tag{33}$$

is a linear continuous map from Y onto Z. The *onto* property is deduced from the fact that the first term in Eq. (31) is an operator providing an onto map.

Let us assume that

$$f(x_0,y_0)=0; \tag{34}$$

that is, the pair (x_0,y_0) defines a nice section through the relation $\gamma_0=x_0+y_0$, then

$$P^{\underline{a}}(x_0,y_0)=M(x_0,y_0)\,\delta^{\underline{a}}_0, \tag{35}$$

which implies that

$$K(x_0,y_0)=1. \tag{36}$$

So the last term in Eq. (31) is given by

$$K^3\left(\frac{4P_{\underline{a}}}{M}-3Kl_{\underline{a}}(\zeta,\overline{\zeta})\right)\delta P^{\underline{a}}=(4\delta^0_{\underline{a}}-3l_{\underline{a}}(\zeta,\overline{\zeta}))\,\delta P^{\underline{a}}, \tag{37}$$

which is clearly an expression involving only spherical harmonics Y_{lm} with $l\leqslant 1$.

Therefore one has that (see also Appendix A)

$$f_y(x_0,y_0)\,\delta y=D\,\delta y-[F\,\delta y-T(F\,\delta y)], \tag{38}$$

with $F=F(\zeta,\overline{\zeta})\equiv\dot{\Psi}(u=x_0+y_0,\zeta,\overline{\zeta})$. As it was done in Ref. 2, we will assume that the modulus of the flux is bounded by the quantity λ, i.e.,

$$|\dot{\Psi}(u,\zeta,\overline{\zeta})|\leqslant\lambda; \tag{39}$$

and therefore $|F|\leqslant\lambda$.

Then the following result holds:

Lemma 3.1: If $(\lambda^2\frac{4}{27})<1$ then the map $\delta z=f_y(x_0,y_0)\,\delta y$ from Y onto Z is invertible.

Proof: We start with the expression

$$D\,\delta y-[F\,\delta y-T(F\,\delta y)]=\delta z. \tag{40}$$

Note that from Eq. (17) one can see that there exists an operator D^{-1}, acting on elements of Z, such that on spherical harmonicas, with $l\geqslant 2$, it gives

$$D^{-1}Y_{lm}=\frac{4}{(l-1)l(l+1)(l+2)}\,Y_{lm}, \tag{41}$$

then one deduces that $z\in Z\Rightarrow D^{-1}z\in Z$. Let us now define the operator $A:Y\to Y$, such that $\forall\,\delta z\in Z$,

$$A\,\delta y\equiv D^{-1}(\delta z+F\,\delta y-T(F\,\delta y)). \tag{42}$$

Then Eq. (40) is equivalent to the equation:

$$A\,\delta y=\delta y. \tag{43}$$

We will next show that if $(\lambda^2\frac{4}{27})<1$ then A is a contracting map,[6] and therefore Eq. (43) has a unique solution.

It is shown in Appendix B that for any function $\Gamma(\zeta,\overline{\zeta})\in Y$ one has the following inequality:

$$\|\Gamma\|^2\leqslant\tfrac{4}{27}\|\Gamma\|^2_D. \tag{44}$$

This clearly improves the inequality (24). It then follows that

$$\begin{aligned}\|A\,\delta y-A\,\delta y'\|^2_D&=\|D^{-1}F(\delta y-\delta y')-D^{-1}T(F(\delta y-\delta y'))\|^2_D\\&=\|F(\delta y-\delta y')-T(F(\delta y-\delta y'))\|^2\\&\leqslant\|F(\delta y-\delta y')\|^2\leqslant\lambda^2\|(\delta y-\delta y')\|^2\leqslant\lambda^2\tfrac{4}{27}\|(\delta y-\delta y')\|^2_D,\end{aligned} \tag{45}$$

where we have used the fact that for any $\gamma \in L^2(S^2)$ one has $\|\gamma - T\gamma\| \leqslant \|\gamma\|$. Consequently A defines a contracting map, which implies that (43) has a unique solution. Then $f_y(x_0, y_0)$ provides with a one-to-one map and onto from Z to Y. ■

Therefore $f_y(x_0, y_0)$ is an homeomorphism of Y onto Z.

This result along with the implicit function theorem[6] allows us to prove the following:

Theorem 3.1: Let $\gamma_0 = x_0 + y_0$ be a nice section, i.e., $f(x_0, y_0) = 0$, and the smooth data $\dot{\Psi}$ satisfying $|\dot{\Psi}(u, \zeta, \bar{\zeta})| \leqslant \lambda$, with $\lambda < \sqrt{\frac{27}{4}}$, then there exist a local four-parameter family of nice sections near γ_0.

Proof: Lemma 3.1 allows us to use the implicit function theorem[6] for the Banach spaces X, Y and Z, and the map $z = f(x, y)$; which tell us that there is a unique continuous mapping $g: U \to Y$ defined in a neighborhood U of x_0 such that $y_0 = g(x_0)$ and

$$Dg(x) = \Psi(u = x + g(x), \zeta, \bar{\zeta}) + K(x + g(x); \zeta, \bar{\zeta})^3 M(x + g(x)). \tag{46}$$

Since there is a four-parameter family of solutions to the equation $D\,x = 0$; one deduces that given a nice section and data satisfying $\lambda < \sqrt{\frac{27}{4}}$, there exists a four-parameter family of them in a neighborhood of the original one. ■

This result generalizes the calculations of Ref. 2 in several ways; notably by pushing the permissible upper value of λ from 1 to $\sqrt{\frac{27}{4}}$, and also by allowing a general translation x.

IV. DIFFERENTIABLE PROPERTIES OF THE LOCAL SOLUTIONS

In order to study the differentiable properties of nice sections we need to introduce further notation. Let us express the Sobolev norm H_4 by[7]

$$\|y\|_{H_4} = \sum_{i=0}^{4} \|\nabla^{(i)} y\|, \tag{47}$$

where the multi-index i is used as

$$|\nabla^{(i)} y|^2 = \nabla^{a_i} \nabla^{a_2} \cdots \nabla^{a_i} y \nabla_{a_1} \nabla_{a_2} \cdots \nabla_{a_i} y. \tag{48}$$

Since the operator D is strongly elliptic[8] of fourth degree on the compact manifold S^2, then the following inequalities hold:[9,10]

$$\|Dy\| \leqslant c_1 \|y\|_{H^4}, \tag{49}$$

$$\|y\|_{H^4} \leqslant c_2 (\|Dy\| + \|y\|), \tag{50}$$

for some positive constants c_1 and c_2.

We will use these relations to prove the next result.

Lemma 4.1: There exist constants C and C', independent of y such that $\forall y \in Y$ one has that:

$$C\|y\|_{H^4} \leqslant \|y\|_D \leqslant C' \|y\|_{H^4}. \tag{51}$$

That is the norms are equivalent.

Proof: The second inequality is just relation (49). The first inequality follows from (50) and the fact that for $y \in Y$ one has:

$$\|y\|_{H^4} \leqslant c(\|Dy\| + \|y\|) = c(\|y\|_D + \|y\|) \leqslant c\left(d\|y\|_D + \sqrt{\frac{4}{27}}\, \|y\|_D \right) = c\left(1 + \sqrt{\frac{4}{27}} \right) \|y\|_D .$$

■

It then follows:

Theorem 4.1: The local four-parameter family of nice sections is two times differentiable functions on the sphere.

Proof: Using the last Lemma and the Sobolev imbedding theorem for compact manifolds (see, e.g., Ref. 7) one deduces that the solutions $g(x)$, mentioned in the proof of the theorem of the last section, are in $C^2(S^2)$. ■

V. NON SELF-CROSSING PROPERTY OF THE LOCAL SOLUTIONS

Let $\gamma(\tau)$ be a one-parameter family of nice sections for $\tau\in[0,t]$. Then since $\gamma(x,y)$ is linearly expressed in terms of x and y by $\gamma(\tau)=x(\tau)+y(x(\tau))$, it is clear that the Fréchet derivative of $f(\gamma(x,y))$, with respect to γ, $f_\gamma(\gamma(x,y))$, is given by the analog expression (31) when one substitutes δy by $\delta\gamma$, that is,

$$f_\gamma(\gamma(x,y))\,\delta\gamma=D\,\delta\gamma-\left[\dot{\Psi}(u=\gamma,\zeta,\bar{\zeta})\,\delta\gamma+K^3\left(\frac{4P_{\underline{a}}}{M}-3Kl_{\underline{a}}(\zeta,\bar{\zeta})\right)\delta P^{\underline{a}}(\delta\gamma)\right], \tag{52}$$

where $P_{\underline{a}}$, M, and K are evaluated at $u=\gamma$, and $\delta P^{\underline{a}}(\delta\gamma)$ is given by

$$\delta P^{\underline{a}}(\delta\gamma)=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\dot{\Psi}(u=\gamma,\zeta,\bar{\zeta})\,\delta\gamma\,dS^2. \tag{53}$$

Recalling that the nice section equation is equivalent to $f(x,y)=0$, it is deduced from the properties of the equation that its Fréchet derivative with respect to γ must also vanish. But it is observed that the equation $f_\gamma(\gamma)\,\delta\gamma=0$ coincides with the linearization of the nice section equation; therefore, if one takes $\delta\gamma=(d\gamma/d\tau)d\tau$ and wants to study the properties of $d\gamma/d\tau$, one needs to analyze the linearized nice section equation.

So, let us consider now the solutions of the linear nice section equation of the form

$$\delta\gamma=\delta x+\Gamma, \tag{54}$$

with $\delta x\in X$ and $\Gamma\in Y$. From the arguments of Ref. 2, and their extension to general translations, presented above, it is deduced that a solution can be constructed by the sequence $\delta\gamma_n=\delta x+\Gamma_n$ defined in the following form:

$$\delta\gamma_0=\delta x, \tag{55}$$

$$\delta\Gamma_1=A_0\,\delta\gamma_0, \tag{56}$$

and in general, for $n>1$,

$$\delta\Gamma_n=A_0(\delta x+\delta\Gamma_{n-1}) \tag{57}$$

or

$$\delta\gamma_n=\delta x+A_0\,\delta\gamma_{n-1}, \tag{58}$$

where A_0 is the trivial extension of the operator defined in (42) to act on $X\oplus Y$, with $\delta z=0$. Let us note that if $\delta x=0$, then $\delta\gamma_1=0$ which implies that $\delta\gamma_n=0\,\forall n$, or equivalently $\Gamma=0$.

It is interesting to observe that this solution for the homogeneous linear equation can also be expressed by

$$\delta\gamma=\sum_{i=0}^{\infty}A_0^i\,\delta x; \tag{59}$$

in particular, one can see that

$$\Gamma_n=\sum_{i=1}^{n}A_0^i\,\delta x. \tag{60}$$

The proof that (59) gives the solution of the homogeneous linear equation can be given in the following way. First let us check that if the series (59) converges then it is a solution, since

$$A_0\delta\gamma = A_0\left(\sum_{i=0}^{\infty} A_0^i \delta x\right) = \sum_{i=1}^{\infty} A_0^i \delta x = \Gamma, \tag{61}$$

which is the representation of the homogeneous linear equation in terms of the operator A_0. Now let us prove that the series (59) converges. Note that

$$\|\delta\gamma_n - \delta\gamma_{n-1}\|_D = \|\Gamma_n - \Gamma_{n-1}\|_D = \|A_0^n \delta x\|_D; \tag{62}$$

then, since for $n \geqslant 1$, $A_0^n \delta x \in Y$, one has from (45) that

$$\|A_0 y\|_D \leqslant \lambda \sqrt{\tfrac{4}{27}} \|y\|_D; \tag{63}$$

therefore

$$\|\delta\gamma_n - \delta\gamma_{n-1}\|_D \leqslant (\lambda \sqrt{\tfrac{4}{27}})^{n-1} \|A_0 \delta x\|_D, \tag{64}$$

which implies that the sequence converges if $(\lambda \sqrt{\frac{4}{27}}) < 1$.

Using the notation of the last section, and denoting $\dot{\Psi}_0 = \dot{\Psi}(u = x_0 + g(x_0), \zeta, \overline{\zeta})$, one can prove the following result:

Theorem 5.1: Let $\gamma_0 = x_0 + g(x_0)$ be a nice section and let δx_0 be a positive constant such that $x = x_0 + \delta x_0 \in U$. There exist $\lambda_* \in (0, \sqrt{\frac{27}{4}})$ such that for data satisfying $|\dot{\Psi}_0| < \lambda_*$ the solution $\delta\gamma(\delta x_0)$ of the linear nice section equation is also positive.

This implies, as is mentioned below, that the solutions of the nice section equation generated by an aligned time translation do not cross among them.

Proof: Using the fact that for any $\gamma \in L^2(S^2)$, $\|\gamma - T\gamma\| \leqslant \|\gamma\|$, one deduces that if $\delta\gamma = \delta x + \Gamma$ is a solution of the linear nice section equation, then

$$\|\Gamma\|_D^2 \leqslant \|\dot{\Psi}_0(\delta x + \Gamma)\|^2 \leqslant \lambda^2 \|\delta x + \Gamma\|^2 \leqslant \lambda^2 \|\Gamma\|^2 + \lambda^2 \|\delta x\|^2 \leqslant \lambda^2 (\tfrac{4}{27} \|\Gamma\|_D^2 + \|\delta x\|^2), \tag{65}$$

therefore

$$\|\Gamma\|_D^2 \leqslant \frac{\lambda^2}{1 - \frac{4}{27}\lambda^2} \|\delta x\|^2. \tag{66}$$

From Appendix B, one has

$$|\Gamma(\zeta, \overline{\zeta})|^2 \leqslant \frac{1}{4\pi} \frac{4}{27} \|\Gamma\|_D^2; \tag{67}$$

therefore, it is deduced that

$$\sup|\Gamma|^2 < \frac{4}{27} \frac{\lambda^2}{1 - \frac{4}{27}\lambda^2} (\delta x_0)^2. \tag{68}$$

Then, the inequality

$$\sup|\Gamma|^2 < (\delta x_0)^2, \tag{69}$$

will be satisfied if

$$\lambda < \lambda_* = \frac{1}{\sqrt{2}} \sqrt{\frac{27}{4}} \cong 1.8371 \qquad \blacksquare \tag{70}$$

The constant λ_* appearing in the result depends on the technique one uses for the proof. In order to illustrate this, in Appendix C we present three other alternative proofs, using different techniques, which impose three different bounds on the data, which do not improve the last one.

It is not yet clear to us whether the above value of λ_* is the best possible one.

Theorem 5.1 proves that $\delta\gamma>0$ for an aligned future time translation $d\tau$, which implies that $d\gamma/d\tau>0$, and therefore that the nice section family does not show self-crossing since $\gamma(\tau+d\tau)$ is to the future of $\gamma(\tau)$.

VI. GLOBAL EXISTENCE OF NICE SECTIONS NEAR THE STATIONARY CASE

The results appearing in the previous sections deal only with local properties of nice sections; in this section the issue of global existence will be tackled. In Ref. 2 it was proved that there exists a four-parameter family of nice sections in a stationary space–time. This suggests that when radiation is ''small'' one could probably find an analogous result. In what follows we will apply once more the implicit function theorem to prove this assertion.

Let $\mathcal{C}^0(\mathcal{I}^+)$ represent the vector space of all bounded continuous functions on scri plus, with the norm $\|F\|_{\mathcal{C}^0(\mathcal{I}^+)}=\sup|F(u,\zeta,\bar{\zeta})|$. The vector space $\mathcal{C}^0(\mathcal{I}^+)$ defines a Banach space.

Let us denote with $\mathcal{C}^0_{-\infty}(\mathcal{I}^+)$ the set of functions $F(u,\zeta,\bar{\zeta})$ at scri plus, which belong to the set $\mathcal{C}^0(\mathcal{I}^+)$ and such that

$$\int_{-\infty}^{u} F(u',\zeta,\bar{\zeta})du' \in C^0(I^+), \tag{71}$$

where $C^0(\mathcal{I}^+)$ represents the set of bounded continuous functions on scri plus. The set $\mathcal{C}^0_{-\infty}(\mathcal{I}^+)$ is a vector subspace of $\mathcal{C}^0(\mathcal{I}^+)$, and therefore is a Banach space.

Let X, Y and Z represent the spaces defined in Sec. III, and let us denote with Ξ the space whose elements are of the form $\xi=(x,F)$, with $x\in X$, $F\in\mathcal{C}^0_{-\infty}(\mathcal{I}^+)$ and with norm $\|\xi\|=\|x\|_x+\|F\|_{\mathcal{C}^0(\mathcal{I}^+)}$. The space Ξ is a Banach space. Consider now the map from (Ξ,Y) into Z defined by

$$\Phi(\xi,y)=Dy-\left(\int_{-\infty}^{x+y}F(u',\zeta,\bar{\zeta})du'+K(x+y,F;\zeta,\bar{\zeta})^3M(x+y,F)\right), \tag{72}$$

which coincide with the map $f(x,y)$ defined in Sec. III, but where now it is taken into account explicitly the dependence on the flux F. The Fréchet derivative of this map with respect to the variable y is

$$\Phi_y(\xi,y)\,\delta y=D\,\delta y-\left[F(u=\gamma=x+y,\zeta,\bar{\zeta})\,\delta y+K^3\left(\frac{4P_{\underline{a}}}{M}-3Kl_{\underline{a}}(\zeta,\bar{\zeta})\right)\delta P^{\underline{a}}\right], \tag{73}$$

where $P_{\underline{a}}$, M, and K are evaluated at $u=\gamma=x+y$; in particular one has

$$P^{\underline{a}}(\gamma)=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})\int_{-\infty}^{\gamma}F(u',\zeta,\bar{\zeta})du'\;dS^2+P^{\underline{a}}(u=-\infty) \tag{74}$$

and $\delta P^{\underline{a}}$ is given by

$$\delta P^{\underline{a}}=-\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})F(u=\gamma=x+y,\zeta,\bar{\zeta})\,\delta y\;dS^2. \tag{75}$$

Let us denote with ξ_0 the nonradiation case, that is, $\xi_0=(x_0,F=0)$; we have shown in Ref. 2 that there exists a relation $y_0(x_0)$ such that $\Phi(\xi_0,y_0)=0$, for any translation x_0; in other words, in the absence of radiation $(F=0)$, the map $\gamma_0(x_0)=x_0+y_0(x_0)$ defines a four-parameter family of nice sections globally on scri. It is observed that $\delta P^{\underline{a}}(F=0)=0$; therefore

$$\Phi_y(\xi(x,F=0),y)\,\delta y=D\,\delta y, \tag{76}$$

and it is clear that the map $\delta z=\Phi_y(\xi(x,F=0),y)\,\delta y$ from Y onto Z is invertible; that is, it is an homeomorphism of Y onto Z.

The implicit function theorem implies that there exists a unique continuous mapping G defined in a neighborhood $U_0\subset\mathcal{C}^0_{-\infty}(\mathcal{I}^+)$ of $F=0$, $G:U_0\to Y$, parametrized by x_0, such that $G(\xi(x_0,F=0))=y_0$ and $\Phi(\xi(x_0,F),y)=0$ for $F\in U_0$.

Let us call *small radiation data* the one that satisfies $F\in U_0$ in the last result. Then, the continuity properties contained in the implicit function theorem and the fact that x lives in a four-dimensional vector space, imply the following result:

Theorem 6.1: For *small radiation data* there exists a global four-parameter family of nice sections at scri.

VII. FINAL COMMENTS

Summarizing, in this article several results concerning local and global properties of the center of mass of isolated systems in general relativity have been presented; in terms of the so-called nice sections. In Sec. VI the global existence of nice sections has been established for *small radiation data*. In Sec. III there appears a formal proof of existence of nice sections with less restrictive assumptions on the strength of the radiation data that had appeared in the literature before.[2]

Although the global result of Sec. VI is very important from a formal perspective of the whole picture based on nice sections, the local result of Sec. III might be more important from the operational point of view, since if one wants to make a concrete calculation based on this construction it is natural to start by taking an initial nice section and then calculating the dynamics of the system on a local family around the initial one. The relaxation of the strength of the radiation warrants a wider set of applicability of this formalism.

Differentiable properties of nice sections have been established in Sec. IV. And in Sec. V the desired non-self-crossing property of nice sections was proved.

It is important to emphasize that this last property is what one expects from making the analogy with the case of a set of particles in linearized gravity; since in this case any sensible notion of center of mass, defined in the interior of the space–time, would follow a timelike curve whose future null cones would define the corresponding one-parameter family of center of mass at future null infinity. This latter family clearly could not show self-crossing since this would imply a faster than light velocity for the trajectory of the center of mass in the interior of the spacetime.

We conclude then that the present approach for the notion of center of mass of isolated gravitating systems, based on the definition of nice sections, not only has the necessary mathematical properties, but it also has the appropriate physical properties that make possible its interpretation.

ACKNOWLEDGMENTS

Most of this work was developed while one of the authors (O.M.) was visiting the Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, Potsdam, Germany. This author is very grateful to all the people at the Institute and in particular Dr. Helmut Friedrich, for their kind hospitality. It is a pleasure to thank Dr. H. Friedrich, Dr. O. Reula, Dr. B. Schmidt, and Dr. E. T. Newman for illuminating discussions on our work. This research was supported by the ''Consejo de Investigaciones Científicas y Tecnológicas de la Provincia de Córdoba'' (CONICOR), the ''Consejo Nacional de Investigaciones Científicas y Técnicas'' (CONICET) and the ''Deutsche Forschungsgemeinschaft'' (DFG).

APPENDIX A

For any regular real function on the sphere f let us define

$$f^{\underline{a}}=\frac{1}{4\pi}\int l^{\underline{a}}(\zeta,\bar{\zeta})f(\zeta,\bar{\zeta})dS^2,$$

and let us also define

$$\hat{f}\equiv f^{\underline{a}}l_{\underline{a}}=f^0-\hat{f}^{(3)},$$

with $\hat{f}^{(3)} \equiv (f^1 l^1 + f^2 l^2 + f^3 l^3)$. Then it is easy to prove by direct calculation

Lemma A1.1:

$$T(f\ ) = f^0 - 3\hat{f}^{(3)}.$$

The following result is also useful.

Lemma A1.2: For any timelike vector U^a one has

$$\frac{1}{4\pi} \int \frac{l^{\underline{a}}}{(U^{\underline{b}} l_{\underline{b}})^3}\, dS^2 = \frac{U^{\underline{a}}}{(U^{\underline{b}} U_{\underline{b}})^2}.$$

Proof: First let us prove that the left-hand side transforms as a four-vector under Lorentz transformations. This can easily be seen from the fact that under Lorentz transformations:

$$l^{\underline{a}} \to K l^{\underline{a}}, \quad dS^2 \to K^2\, dS^2.$$

Therefore

$$\frac{1}{4\pi} \int \frac{l^{\underline{a}}}{(U^{\underline{b}} l_{\underline{b}})^3}\, dS^2 = \frac{V^{\underline{a}}}{(U^{\underline{b}} U_{\underline{b}})^2}$$

for some four-vector $V^{\underline{a}}$

Since U^a is a timelike vector, one can choose a frame for which

$$U^{\underline{a}} = (U^0, 0, 0, 0),$$

from which it is immediately deduced that

$$V^{\underline{a}} = (U^0, 0, 0, 0) = U^{\underline{a}}.$$

■

Then it follows:

Corollary A1.3: Let U^a be a timelike vector, then:

$$T\left(\frac{1}{(U^{\underline{a}} l_{\underline{a}})^3}\right) = \frac{1}{(U^{\underline{a}} U_{\underline{a}})^2}\,(u^0 - 3\hat{u}^{(3)}),$$

where

$$U^{\underline{a}} l_{\underline{a}} = u^0 - \hat{u}^{(3)}.$$

Proof: From Lemma A1.1, one has:

$$T\left(\frac{1}{(U^{\underline{a}} l_{\underline{a}})^3}\right) = \frac{1}{4\pi} \int \frac{1}{(U^{\underline{a}} l_{\underline{a}})^3}\,(l^0 l_0 - 3 l^i(\zeta', \overline{\zeta'}) l^i(\zeta, \overline{\zeta}))\, dS'^2 = \frac{1}{(U^{\underline{a}} U_{\underline{a}})^2}\,(u^0 - 3\hat{u}^{(3)}),$$

where $i = 1,2,3$. ■

Then it follows that

$$T(\Psi) = -T(K^3 M).$$

APPENDIX B

Let us note that for any function $\Gamma(\zeta, \overline{\zeta}) \in Y$ one has

$$|\Gamma(\zeta,\bar{\zeta})|^2=\left|\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}Y_{lm}(\zeta,\bar{\zeta})\right|^2$$
$$=\left|\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}\frac{(l-1)l(l+1)(l+2)}{4}\frac{4}{(l-1)l(l+1)(l+2)}Y_{lm}(\zeta,\bar{\zeta})\right|^2$$
$$\leqslant\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}|\Gamma^{lm}|^2\left[\frac{(l-1)l(l+1)(l+2)}{4}\right]^2\right]$$
$$\times\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\left[\frac{4}{(l-1)l(l+1)(l+2)}\right]^2|Y_{lm}(\zeta,\bar{\zeta})|^2\right].$$

Now, using the addition theorem for the spherical harmonics, one can prove that

$$\sum_{m=-l}^{l}|Y_{lm}(\zeta,\bar{\zeta})|^2\leqslant\frac{2l+1}{4\pi}.$$

Also one can see[3] that

$$\|\Gamma\|_D^2=\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}|\Gamma^{lm}|^2\left[\frac{(l-1)l(l+1)(l+2)}{4}\right]^2\right]$$

from which it is deduced that

$$|\Gamma|^2\leqslant\frac{1}{4\pi}\|\Gamma\|_D^2\left\{\sum_{l=2}^{\infty}\left[\frac{4}{(l-1)l(l+1)(l+2)}\right]^2(2l+1)\right\}=\frac{1}{4\pi}\frac{4}{27}\|\Gamma\|_D^2.$$

Since the modulus $\|\Gamma\|$ is bounded by its maximum value, namely

$$\|\Gamma\|^2\leqslant 4\pi\ \sup\|\Gamma(\zeta,\bar{\zeta})|^2;$$

it is deduced the inequality:

$$\|\Gamma\|^2\leqslant\tfrac{4}{27}\|\Gamma\|_D^2. \tag{B1}$$

APPENDIX C

Since the peculiar value of λ_* depends on the technique of the proof of Theorem 5.1, we present here three other proofs of this result.

Proof 1: Let us observe that

$$\langle\Gamma|\eth^2\bar{\eth}^2\Gamma\rangle=\langle\bar{\eth}^2\Gamma|\bar{\eth}^2\Gamma\rangle=\langle\eth^2\Gamma|\eth^2\Gamma\rangle;$$

the question arises then whether there is an extension of the Sobolev inequality that will relate $|\Gamma(\zeta,\bar{\zeta})|$ with the previous expression. We will now prove that there is such a relation.

Let the regular function Γ be given by

$$\Gamma=\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}Y_{lm}(\zeta,\bar{\zeta}),$$

then one has

$$
\begin{aligned}
|\Gamma(\zeta,\bar{\zeta})|^2 &= \left| \sum_{l=2}^{\infty} \sum_{m=-l}^{l} \Gamma^{lm} Y_{lm}(\zeta,\bar{\zeta}) \right|^2 \\
&= \left| \sum_{l=2}^{\infty} \sum_{m=-l}^{l} \Gamma^{lm} \sqrt{\frac{(l-1)l(l+1)(l+2)}{4}} \sqrt{\frac{4}{(l-1)l(l+1)(l+2)}}\, Y_{lm}(\zeta,\bar{\zeta}) \right|^2 \\
&\leqslant \left[\sum_{l=2}^{\infty} \sum_{m=-l}^{l} |\Gamma^{lm}|^2 \left[\frac{(l-1)l(l+1)(l+2)}{4} \right] \right] \\
&\quad \times \left[\sum_{l=2}^{\infty} \sum_{m=-l}^{l} \left[\frac{4}{(l-1)l(l+1)(l+2)} \right] |Y_{lm}(\zeta,\bar{\zeta})|^2 \right].
\end{aligned}
$$

Now, using the inequality implied by the addition theorem for the spherical harmonics, and that[3]

$$
\langle \Gamma | \eth^2 \bar{\eth}^2 \Gamma \rangle = \left[\sum_{l=2}^{\infty} \sum_{m=-l}^{l} |\Gamma^{lm}|^2 \left[\frac{(l-1)l(l+1)(l+2)}{4} \right] \right],
$$

one deduces

$$
|\Gamma|^2 \leqslant \frac{1}{4\pi} \langle \Gamma | \eth^2 \bar{\eth}^2 \Gamma \rangle \left[\sum_{l=2}^{\infty} \frac{4(2l+1)}{(l-1)l(l+1)(l+2)} \right] = \frac{1}{4\pi} \frac{4}{3} \langle \Gamma | \eth^2 \bar{\eth}^2 \Gamma \rangle.
$$

One also has that

$$
\langle \Gamma | \eth^2 \bar{\eth}^2 \Gamma \rangle = \langle \Gamma | \dot{\Psi}_0 \delta x \rangle + \langle \Gamma^2 | \dot{\Psi}_0 \rangle,
$$

where we have used that Γ is orthogonal to the last term appearing in the linear nice section equation.

Let us note that from the Schwarz inequality one obtains

$$
|\langle \Gamma | \dot{\Psi}_0 \delta x \rangle| \leqslant \|\Gamma\| \lambda \| \delta x \|;
$$

and similarly

$$
\langle \Gamma^2 | \dot{\Psi}_0 \rangle \leqslant \lambda \|\Gamma\|^2,
$$

from which it is deduced

$$
\sup |\Gamma|^2 \leqslant \frac{1}{4\pi} \frac{4}{3} (\lambda \|\Gamma\|^2 + \|\Gamma\| \lambda \| \delta x \|).
$$

It follows from the techniques used in Ref. 2 and in Sec. V that Γ is expressible as a series of the form:

$$
\Gamma = \Gamma_1 \lambda + \Gamma_2 \lambda^2 + \cdots,
$$

which satisfies

$$
\|\Gamma_n \lambda^n\| \leqslant \left(\lambda \sqrt{\frac{4}{27}} \right)^{n-1} \|\Gamma_1 \lambda\|
$$

with

$$
\|\Gamma_1 \lambda\| \leqslant \sqrt{\frac{4}{27}} \|\eth^2 \bar{\eth}^2 \Gamma_1 \lambda\| = \sqrt{\frac{4}{27}} \|\dot{\Psi}_0 \delta x + \Psi_0 + K_1^3 M_1\| \leqslant \sqrt{\frac{4}{27}} \|\dot{\Psi}_0 \delta x\| \leqslant \lambda \sqrt{\frac{4}{27}} \|\delta x\|.
$$

Then the modulus of Γ is bounded by

$$\|\Gamma\| \leqslant \sum_{n=1}^{\infty} \|\Gamma_n \lambda^n\| \leqslant \sum_{n=1}^{\infty} \left(\lambda \sqrt{\frac{4}{27}}\right)^{n-1} \|\Gamma_1 \lambda\| \leqslant \sum_{n=1}^{\infty} \left(\lambda \sqrt{\frac{4}{27}}\right)^{n} \|\delta x\| = \frac{\lambda \sqrt{\frac{4}{27}}}{1-\lambda \sqrt{\frac{4}{27}}} \|\delta x\|,$$

for $\lambda < \sqrt{\frac{27}{4}} \cong 2.598$.

Using this in our main inequality one obtains

$$\sup|\Gamma|^2 \leqslant \frac{1}{4\pi} \frac{4}{3} \sqrt{\frac{27}{4}} \left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2 \|\delta x\|^2 = \frac{1}{\pi} \sqrt{\frac{3}{4}} \left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2 \|\delta x\|^2.$$

Taking $\delta x = \delta x_0 > 0$ where δx_0 is constant, then one has that

$$\|\delta x\|^2 = 4\pi (\delta x_0)^2$$

and so:

$$\sup|\Gamma|^2 \leqslant 2\sqrt{3} \left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2 (\delta x_0)^2.$$

One can see that $\delta g(\delta x_0)$ will be positive if

$$\sup|\Gamma|^2 < (\delta x_0)^2,$$

which is equivalent to

$$2\sqrt{3} \left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2 < 1.$$

This inequality is satisfied for $\lambda \in (0, \lambda_*)$ with

$$\lambda_* = \sqrt{\frac{27}{4}} \frac{1}{1+\sqrt{2}\sqrt{3}} = 0.9080.$$

Proof 2: From the Sobolev inequality one obtains

$$\sup|\Gamma|^2 \leqslant \frac{1}{4\pi} \frac{1}{4} \|\Delta \Gamma\|^2,$$

as can be deduced from the following argument.

Let the regular function Γ be given by

$$\Gamma = \sum_{l=2}^{\infty} \sum_{m=-l}^{l} \Gamma^{lm} Y_{lm}(\zeta, \overline{\zeta}),$$

then one has

$$|\Gamma(\zeta,\bar{\zeta})|^2=\left|\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}Y_{lm}(\zeta,\bar{\zeta})\right|^2=\left|\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}l(l+1)\,\frac{1}{l(l+1)}\,Y_{lm}(\zeta,\bar{\zeta})\right|^2$$
$$\leqslant\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}|\Gamma^{lm}|^2[l(l+1)]^2\right]\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\left[\frac{1}{l(l+1)}\right]^2|Y_{lm}(\zeta,\bar{\zeta})|^2\right].$$

Using the inequality deduced from the addition theorem for spherical harmonics and that

$$\|\Delta\Gamma\|^2=\left[\sum_{l=2}^{\infty}\sum_{m=-l}^{l}|\Gamma^{lm}|^2[l(l+1)]^2\right],$$

one deduces

$$|\Gamma|^2\leqslant\frac{1}{4\pi}\|\Delta\Gamma\|^2\left[\sum_{l=2}^{\infty}\frac{(2l+1)}{(l(l+1))^2}\right]=\frac{1}{4\pi}\frac{1}{4}\|\Delta\Gamma\|^2.$$

Then from the linear nice section equation, one obtains

$$\langle\Gamma|\eth^2\bar{\eth}^2\Gamma\rangle=\frac{1}{4}\|\Delta\Gamma\|^2+\frac{1}{2}\int\nabla_a\Gamma\nabla^a\Gamma\;dS^2=\langle\Gamma|\dot{\Psi}_0\delta x\rangle+\langle\Gamma^2|\dot{\Psi}_0\rangle,$$

where we have used that Γ is orthogonal to the last term appearing in the linear nice section equation.

Using the inequalities mentioned in Proof 1 and

$$\|\Gamma\|^2\leqslant(2\pi)^2\int\nabla_a\Gamma\nabla^a\Gamma\;dS^2,$$

coming from the Poincaré lemma, one can deduce

$$4\pi\,\sup|\Gamma|^2+\frac{1}{2(2\pi)^2}\|\Gamma\|^2\leqslant\lambda\|\Gamma\|^2+\|\Gamma\|\lambda\|\delta x\|;$$

or equivalently

$$\sup|\Gamma|^2\leqslant\frac{1}{4\pi}\left\{\|\Gamma\|^2\left(\lambda-\frac{1}{8\pi^2}\right)+\|\Gamma\|\lambda\|\delta x\|\right\}.$$

Substituting the bound for $\|\Gamma\|$ found above, one obtains

$$\sup|\Gamma|^2\leqslant\frac{1}{4\pi}\left\{\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\|\delta x\|\right)^2\left(\lambda-\frac{1}{8\pi^2}\right)+\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\|\delta x\|\right)\lambda\|\delta x\|\right\},$$

from which it is deduced

$$\sup|\Gamma|^2\leqslant\frac{1}{4\pi}\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2\left(\sqrt{\frac{27}{4}}-\frac{1}{8\pi^2}\right)\|\delta x\|^2.$$

Taking $\delta x=\delta x_0>0$ where δx_0 is constant, one has

$$\sup|\Gamma|^2\leqslant\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2\left(\sqrt{\frac{27}{4}}-\frac{1}{8\pi^2}\right)(\delta x_0)^2.$$

Then in order for $\delta g(\delta x_0)$ to be positive, one needs

$$\sup|\Gamma|^2<(\delta x_0)^2,$$

or equivalently

$$\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2\left(\sqrt{\frac{27}{4}}-\frac{1}{8\pi^2}\right)<1.$$

This inequality is satisfied for $\lambda\in(0,\lambda_*)$ with

$$\lambda_*=\sqrt{\frac{27}{4}}\left(\frac{1}{1+\sqrt{\sqrt{\frac{27}{4}}-\frac{1}{8\pi^2}}}\right)\cong 0.9962.$$

Proof 3: Let the regular function Γ be given by

$$\Gamma=\sum_{l=2}^{\infty}\sum_{m=-l}^{l}\Gamma^{lm}Y_{lm}(\zeta,\overline{\zeta}).$$

It was proved in Appendix B that

$$|\Gamma(\zeta,\overline{\zeta})|^2\leqslant\frac{1}{4\pi}\frac{4}{27}\langle D\Gamma|D\Gamma\rangle=\frac{1}{4\pi}\frac{4}{27}\|\Gamma\|_D^2.$$

If Γ is a solution of the linear nice section equation then

$$\|\Gamma\|_D^2\leqslant\|\dot{\Psi}_0(\delta x+\Gamma)\|^2\leqslant\lambda^2\|\delta x+\Gamma\|^2\leqslant\lambda^2\|\Gamma\|^2+\lambda^2\|\delta x\|^2;$$

as was explained above. Then using the bound

$$\|\Gamma\|\leqslant\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\|\delta x\|,$$

for $(\lambda\sqrt{\frac{4}{27}})<1$; one obtains

$$\sup|\Gamma|^2\leqslant\frac{1}{4\pi}\frac{4}{27}\left\{\lambda^2\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2\|\delta x\|^2+\lambda^2\|\delta x\|^2\right\}.$$

Taking $\delta x=\delta x_0>0$ where δx_0 is constant, then one has that

$$\|\delta x\|^2=4\pi(\delta x_0)^2$$

and so:

$$\sup|\Gamma|^2\leqslant\frac{4}{27}\lambda^2\left\{\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2+1\right\}(\delta x_0)^2.$$

One can see that $\delta g(\delta x_0)$ will be positive if

$$\sup|\Gamma|^2<(\delta x_0)^2,$$

which is equivalent to

$$\frac{4}{27}\lambda^2\left\{\left(\frac{\lambda}{\sqrt{\frac{27}{4}}-\lambda}\right)^2+1\right\}<1.$$

This inequality is satisfied for $\lambda\in(0,\lambda_*)$ with $\lambda_*\cong 1.5134$.

[1] O. M. Moreschi, ''On the limitation of approximation methods around stationary backgrounds for isolated systems,'' Class. Quantum Grav. **5**, L53–L57 (1988).
[2] O. M. Moreschi, ''Supercenter of mass system at future null infinity,'' Class. Quantum Grav. **5**, 423–435 (1988).
[3] O. M. Moreschi, ''On angular momentum at future null infinity,'' Class. Quantum Grav. **3**, 503–525 (1986).
[4] R. Geroch, A. Held, and R. Penrose, ''A space-time calculus based on pairs of null directions,'' J. Math. Phys. **14**, 874–881 (1973).
[5] O. M. Moreschi, ''General future asymptotically flat spacetimes,'' Class. Quantum Grav. **4**, 1063–1084 (1987).
[6] M. S. Berger, *Nonlinearity and Functional Analysis* (Academic, New York, 1977).
[7] T. Aubin, *Nonlinear Analysis on Manifolds. Monge-Ampère Equations* (Springer-Verlag, Berlin, 1982).
[8] A. Friedman, *Partial Differential Equations* (Holt, Rinenhart and Winston, New York, 1969).
[9] A. D. Rendall, ''Existence and asymptotic properties of global solutions of the Robinson-Trautman equation,'' Class. Quantum Grav. **5**, 1339–1347 (1988).
[10] Y. Choquet-Bruhat and D. Chirstodoulou, ''Elliptic systems in $H_{s,\delta}$ spaces on manifolds which are Euclidean at infinity,'' Acta Math. **146**, 129–150 (1981).

On stationary black holes of the Einstein conformally invariant scalar system

T. Zannias[a)]
Instituto de Física y Matematicas, Univer. Michoacana SNH, Morelia Mich. Mexico, C.P. 58040 AP 2-82 Mexico

(Received 25 June 1997; accepted for publication 24 June 1998)

The circularity property of stationary-axisymmetric asymptotically flat solutions of the coupled Einstein conformally invariant scalar field equations is investigated. It is proven that as a result of the coupling of the scalar field to the background scalar curvature, in general such solutions fail to be circular. Consequently, it is argued that the space of stationary, black hole equilibrium states may contain states which are not circular. However, the circular stationary-axisymmetric sector is nonempty and in fact it is proven that: All circular, stationary-axisymmetric asymptotically flat black hole solutions are those and only those black holes states described by the two parameter family of Kerr black holes. © *1998 American Institute of Physics.* [S0022-2488(98)02910-7]

I. INTRODUCTION

The purpose of the present paper is to discuss stationary, asymptotically flat, black hole solutions admitted by the coupled Einstein conformally invariant massless scalar field equations. In a previous paper,[1] considerations were restricted to the static black hole sector of the theory. It was proven that the above system, admits no other static, asymptotically flat, black holes apart from the Schwarzschild class. Furthermore, the peculiar structure of the conformal stress tensor and its impact on the spectrum of possible static black hole states of the system was briefly discussed.[2]

Presently we shall extend the inquires of Ref. 1 to the nonstatic case. Even though the present work is a natural extension of the work presented in Ref. 1, the transition to the study of the nonstatic black hole sector involves rather slippery steps. At first recall that the coupled Einstein conformally invariant scalar field system is described by

$$(1-\alpha\Phi^2)R_{mn}=\alpha[4\nabla_m\Phi\nabla_n\Phi-2\Phi\nabla_m\nabla_n\Phi-g_{mn}\nabla^l\Phi\nabla_l\Phi], \tag{1a}$$

$$\nabla^m\nabla_m\Phi=0, \tag{1b}$$

where all indices are four-dimensional $\alpha=k/6$ with k signifying the gravitational coupling constant. Due to the fact that the right-hand side of (1a) contains second derivatives of the scalar field, as well as the presence of the factor $1-\alpha\Phi^2$, the conformal stress tensor fails to satisfy the familiar energy conditions. In particular the dominant energy condition as well as the null convergence conditions fail. Recall that validity of the null convergence conditions implies $R_{ab}l^al^b \geqslant 0$ for every null vector l. From (1a) one can easily deduce that $R_{ab}l^al^b$ has indefinite sign, signaling failure of the null convergence condition. The above mentioned failures are not without consequences. Let us first recall a few fundamental facts and some recent developments concerning black hole equilibrium states.

Hawking's original proof of the rigidity theorem[3] (see also Ref. 4) intended to guarantee three important attributes of any asymptotically flat black hole solution (possessing no ergoregion disconnected from the horizon) within Einstein (or Einstein Brans–Dicke[5]) gravity

(a) The Killing property of the event horizon,
(b) the axisymmetry property of the domain of outer communication, and

a)Electronic mail: zannias@ifm1.ifm.umich.mx

 © 1998 American Institute of Physics

(c) the topologically spherical like cross sections of the horizon.

In the original proof of the rigidity theorem, Hawking assumed validity of Einstein's field equations under the condition that the energy momentum of the matter-fields satisfies the dominant energy condition. Furthermore it was assumed by Hawking, that suitable hyperbolic equations govern the evolution of matter fields, and analyticity for geometry matter fields holds from infinity up to and including the horizon. One of the key steps in Hawking's proof of the rigidity theorem, involved the extension of the horizon isometries as a local space–time isometry group valid in the near vicinity of the horizon. Although such local extension does not lead into problems, its subsequent extension via analytical continuation as an isometry over all the domain of outer communication was problematic. This last step appeared as not well justified and concerns expressed in Ref. 6. Very recently, work by Chruściel[7,8] has offered a refinement in the proof of the rigidity theorem. The new approach aims at arriving at conclusions (a) and (b) but at the same time it avoids some of the weak points of the initial proof.[9] The key ingredient in this new formulation is a theorem established by Chruściel (abbreviated here after as CT) which, leaving aside precise mathematical definitions, states the following:[7,8]

Suppose a space–time satisfies the following conditions:

(1) It is analytic and asymptotically flat;
(2) possesses a nonempty globally hyperbolic and simply connected domain of outer communication $\langle\langle M\rangle\rangle$;
(3) admits a Killing vector field X with complete orbits;
(4) its ADM four momentum is timelike;
(5) in addition to the original Killing field X, there exist another distinct, (i.e., not constant multiple of X) Killing field Y, defined only in some neighborhood of $\langle\langle M\rangle\rangle$.

Then: The isometry group of $\langle\langle M\rangle\rangle$ must contains an $RxU(1)$.

Based on this theorem, the validity of which is independent of any field equations and energy conditions, it was shown in Ref. 7 that any stationary electrovacuum black hole space–time necessary satisfies properties (a) and (b). [Due to the fact that the Ricci tensor for electrovacuum satisfies the null convergence condition then property (c) above follows and the simply connectedness property of the domain of outer communication can be removed from condition (2) of CT (see comments in Ref. 9).] We may add parenthetically that as a bonus of the new reformulation, uniqueness of the Pappapetrou–Majumbar class of extreme black holes has eventually proven true.[11]

However, attempts to use the abovementioned work to prove a rigidity theorem for the black holes solution of (1a) and (1b) run into obstacles. We shall now indicate that the peculiarities of systems (1a) and (1b) coupled with the failure of the dominant and null convergence condition, makes us unable to use CT as a mean of proving a rigidity theorem for (1a) and (1b).

Right at the start, one is confronted with the issue of the Arnowitt, Deser, Misner (ADM) four momentum, namely, whether all asymptotically flat solutions of (1a) and (1b) possess timelike four momenta. As is well known, every nonsingular asymptotically flat solution of Einstein nonvacuum equations possesses positive ADM mass and timelike ADM four momentum, provided the stress tensor satisfies the dominant energy condition. However, even though for (1a) and (1b), the dominant energy condition fails one should not rush to conclude that all asymptotically flat solutions of (1a) and (1b) possess nonpositive ADM mass. In fact as we shall indicate in Sec. III below, (1a) and (1b) admit initial data, ''close enough'' to Minkowski data, which possess timelike ADM four momenta. Of course there is no contradiction with the powerful conclusion of the positivity of gravitation mass theorem. Validity of the dominant energy condition serves only as sufficient condition for the positivity of the ADM mass (-momentum). This point as has been recently discussed in Ref. 12 and in fact, in the same reference it was pointed out that the positivity of ADM mass and timelike nature of the ADM momenta may follow without even imposing the full strength of the dominant energy condition. In view of this discussion, it is fair to state that at this point we do not know the character of the ADM four momentum for every asymptotically flat solution of (1a) and (1b). Beyond the issue of ADM momenta, a stronger reason that unable us to prove a rigidity type theorem for (1a) and (1b), is centered around the analyticity property of arbitrary asymptotically flat solutions of (1a) and (1b). We have no evi-

dence of analyticity. Although every static or stationary solution of the Einstein vacuum,[13] or for instance, of the Einstein minimally coupled massless scalar field equations[14] are analytic on an appropriately constructed coordinate atlas, this is not the case for systems (1a) and (1b). We may recall that analyticity of every static solution for the two former systems, follows as a consequence of the elliptical nature of the dynamical equations clearly exhibited in the specially constructed coordinate atlas. Attempts to prove the elliptic character of the dynamical equations (1a) and (1b) do not yield the desired conclusion and work under progress indicates that analyticity for arbitrary static solutions of (1a) and (1b) may be lost.[14] We shall also, further below, provide additional evidence supporting loss of analyticity. On the other hand, analyticity is a crucial condition for the validity of CT. In particular the extension of the locally defined Killing field Y into a global one over the entire $\langle\langle M\rangle\rangle$, is a sole consequence of the analyticity. [In this regard, the reader is urged to examine the content of theorem (2.1) quoted in Ref. 8.] But the strongest reason for the nonapplicability of CT for our purpose, is centered in the failure of the null convergence condition. As we have already discussed in Ref. 9, satisfaction of the null convergence condition, asymptotic flatness at null infinity J, combined with some additional form of weak regularity near J (see Refs. 7 and 8 for the precise definitions of the term weak regularity), guarantees the simply connectedness property of the domain of outer communication and the S^2xR topological nature of every connected component of the event horizon. In turn, in order to establish the rigidity theorem at least for the electrovacuum cases, the S^2xR character of the horizon appears mandatory (see details in Ref. 8). It is not clear at all whether one can even contemplate a rigidity theorem, if the horizon is allowed to possess arbitrary topologies, for instance, to be toroidal or possess more exotic topologies of higher than genus two.

Due to the above (actually interrelated) reasons and irrespective of the status of other conditions in CT, the relationships between the event horizon and the Killing horizon, the axisymmetry of $\langle\langle M\rangle\rangle$ and the allowed horizon topologies for the possible black hole solutions of systems (1a) and (1b) are presently unknown. Because of the luck of that vital information, the present paper intends to analyze only a special class of black hole solutions of (1a) and (1b). More precisely we shall probe the sector of black hole solutions of (1a) and (1b) covered by Carter's work.[15,16] Carter's work on black hole equilibrium states, begins with the reasonable assumption that $\langle\langle M\rangle\rangle$ does not admit compact causal curves[15] an assumption referred to as the causality axiom.[15] [In this section we shall employ closely the terminology used in Ref. 15. In particular (pseudo) stationary action refers to the action where the infinitesimal Killing generator is restricted to be timelike only in the asymptotic region.] For a weakly asymptotically simple, (pseudo)-stationary space–time M, the domain of outer communication $\langle\langle M\rangle\rangle$ is defined by Carter, as the set of points from which there exist both future and past directed timelike curves extending to arbitrarily large asymptotic distances. It is then proven[15,16] that $\langle\langle M\rangle\rangle$ is a maximal asymptotically flat domain having the property that for every point x in $\langle\langle M\rangle\rangle$, the trajectory or orbit of the (pseudo) stationary action will if extended far enough, enter and remains in the chronological future of x. Equivalently, $\langle\langle M\rangle\rangle$ possess the structure of a fiber bundle with a three-dimensional base manifold and fibers the trajectories of the (pseudo) stationary group action.[17] Although the construction of $\langle\langle M\rangle\rangle$ involves only the use of the (pseudo) stationary group action, asymptotic flatness and validity of the causality axiom, the identification of the event horizon, for the class of nonstatic black holes needs additional inputs. Partially motivated by astrophysical considerations, Carter restricted attention only to axisymmetric black holes. More precisely, in Refs. 15 and 16 attention was restricted to black holes stationary states where the domain of outer communication $\langle\langle M\rangle\rangle$ admits two commuting actions, one (pseudo) stationary and the other axisymmetric,[18,19] the horizon admits spherical-like cross sections and additionally the Ricci tensor obeys the circularity condition.[20] Under those assumptions, the event horizon of any black hole which is either static, or stationary-axisymmetric must coincide with the ''staticity limit'' or ''circularity limit,'' respectively. The staticity limit is defined as the boundary of the region where the hypersurface orthogonal Killing field is timelike while the circularity limit is defined to be the boundary where a particular combination of the timelike at infinity and the axial Killing fields is timelike. Thus Carter's work, even though broad enough, does not employ any particular set of field equations and makes no specific reference to energy conditions, it does cover, as far as the stationary class is concerned, only the class of stationary-axisymmetric and circular ones.

The above brief description of the space–times associated with stationary black holes indi-

cates that even though the benefits of a rigidity theorem for (1a) and (1b) would be very welcomed, axisymmetry follows rather than being imposed in advance, at this stage and for the reason discussed above we are unable to prove such a theorem. For the purpose of the present paper we shall employ Carter's work, with all the limitations that entails.

II. THE CIRCULARITY ISSUE

According to the above discussion, in order to get the benefits of Carter's work, we must first examine whether the Ricci tensor of (1a) is circular. Although vacuum or electrovacuo stationary-axisymmetric asymptotically flat solutions of Einstein's equations are necessary circular[15,16,20] nothing is known about the circularity properties of stationary-axisymmetric solutions of (1a) and (1b). The results of the present section are of interest in their own right and are actually independent of the other sections[21] of the paper.

Accordingly, let us consider an arbitrary stationary-axisymmetric solution of (1a) and (1b). Denote by ξ, ψ the two commuting Killing fields generators of the (pseudo) stationary and axisymmetric actions, respectively, with the former being timelike at infinity and the later spacelike possessing closed orbits. Asymptotic flatness implies the existence of an axis where ψ is vanishing, a property that we shall often use later on. In the terminology of Carter[15,16,20] circularity or orthogonal transitivity of the action generated by the two commuting Killing fields, guarantees that the distribution of the two-dimensional subspaces of the tangent spaces of the manifold, orthogonal to the planes spanned by the commuting Killing fields is integrable. Such integrability property implies that one may choose coordinates such that the metric takes the Lewis–Pappapetrou form,[22] a form that simplifies enormously the field equations. Referring the reader to the original work of Carter[15,16,20] or to Sec. 7.1 and Appendix B.3 of Walds book,[22] if by $\omega_{(\xi)a}=\epsilon_{abcd}\xi^b\nabla^c\xi^d$, $\omega_{(\psi)a}=\epsilon_{abcd}\psi^b\nabla^c\psi^d$ we denote the twists of the two Killing fields, then circularity in a domain containing the axis is ensured provided: $\omega_{(\xi)a}\psi^a=\omega_{(\psi)a}\xi^a=0$. A computation, based on the properties of the Killing equation alone, shows[22]

$$\nabla_m(\omega_{(\xi)a}\psi^a)=-2\epsilon_{mbcd}\psi^b\xi^c R^d_e\xi^e, \tag{2}$$

with a similar expression for $\nabla_m(\omega_{(\psi)a}\xi^a)$. Using (1a) in (2) one gets

$$\begin{aligned}(1-\alpha\Phi^2)\nabla_m(\omega_{(\xi)a}\psi^a)&=-2\alpha\epsilon_{mbcd}\psi^b\xi^c[(4\nabla^d\Phi\nabla_e\Phi-2\Phi\nabla^d\nabla_e\Phi-\delta^d_e\nabla^l\Phi\nabla_l\Phi)]\xi^e\\&=-4a\Phi\epsilon_{mbcd}\psi^b\xi^c(\nabla^d\xi^e)\nabla_e\Phi.\end{aligned} \tag{3}$$

The last part in Eq. (3) is obtained by explicitly taking into account that the Lie derivative of Φ along the two Killing fields is vanishing, the term proportional to the magnitude of the gradient of Φ gives zero contribution after antisymmetrization, and a differentiation by parts has been performed in the term containing the second derivatives of Φ. The right-hand side of Eq. (3) can be simplified further by virtue of the symmetries of Φ and the Killing equation. It is convenient at this point, to introduce a set of local (t,ϕ,x,y) coordinates adapted to the Killing fields i.e., $\xi=\partial/\partial t$, $\psi=\partial/\partial\phi$ and x,y for the moment arbitrary spacelike ones. In such coordinates, the scalar field as well as $\omega_{(\xi)a}\psi^a$ have nonvanishing gradient only along the two nonignorable coordinates i.e.,

$$\nabla_e\Phi=(\nabla_x\Phi)\,\delta^x_e+(\nabla_y\Phi)\,\delta^y_e,$$

$$\nabla_m(\omega_{(\xi)a}\psi^a)=\nabla_x(\omega_{(\xi)a}\psi^a)\,\delta^x_m+\nabla_y(\omega_{(\xi)a}\psi^a)\,\delta^y_m.$$

Therefore, one may easily see that $\nabla_m(\omega_{(\xi)a}\psi^a)$ is proportional to $\nabla_m\Phi$ and in fact Eq. (3) can be rewritten as

$$\begin{aligned}(1-a\Phi^2)\nabla_m(\omega_{(\xi)a}\psi^a)&=-4a\Phi\epsilon_{abcd}\psi^b\xi^c(\nabla^d\xi^a)\nabla_m\Phi\\&=4a\Phi(\omega_{(\xi)a}\psi^a)\nabla_m\Phi=2a(\omega_{(\xi)a}\psi^a)\nabla_m\Phi^2=-2(\omega_{(\xi)a}\psi^a)\nabla_m(1-a\Phi^2).\end{aligned} \tag{4}$$

The additional minus sign appearing in passing from the second to the third equality results from transposing the indices a to the last slot of the the antisymmetric tensor so that the twist of ξ is compatible with the one employed in the derivation of Eq. (2).

From Eq. (4) one infers that as long as $(1-a\Phi^2)$ is nonzero then.

$$\omega_{(\xi)a}\psi^a = C(1-a\Phi^2)^{-2}, \tag{4a}$$

with C an arbitrary space–time integration constant. Similar considerations apply for the term $\omega_{(\psi)a}\xi^a$ as well, and an equation analogous to (5) is also easily obtained. Because of the presence of the axis, one could have concluded from (4a) that it is necessary for $C=0$ everywhere. This implies that in a region containing the axis an arbitrary stationary-axisymmetric solution of Eqs. (1a) and (1b) is locally circular. However, to conclude that circularity holds globally is a more delicate matter. Implicit in the derivation of Eq. (4a) lies the assumption that in the region under consideration the factor $(1-a\Phi^2)$ is no where vanishing. However, extending this property globally is a strong assumption. There is no *a priori* reason that an arbitrary solution of the field Eqs. (1a) and (1b) will necessarily obey such a property. Therefore, we shall examine the analog of

$$(1-a\Phi^2)\nabla_m(\omega_{(\xi)a}\psi^a) = -2(\omega_{(\xi)a}\psi^a)\nabla_m(1-a\Phi^2), \tag{4b}$$

in a region of space–time where $(1-a\Phi^2)=0$. If we could prove that indeed in such regions circularity is preserved, then we can safely conclude that all stationary-axisymmetric solutions of Eqs. (1a) and (1b) are indeed globally circular. In turn to settle this question one first needs to address the following one: Suppose that (g,Φ) is a nonsingular solution of the field equations (1a) and (1b). What is the value of the Ricci tensor at points where $1-a\Phi^2=0$? This question is addressed in details in Appendix A. There it is shown that at a point (or points) where $(1-a\Phi^2)=0$, but the gradient of Φ is not identically zero,[23] the Einstein field equations (1a) imply the following expression for the Ricci tensor:

$$\Phi R_{mn} = \nabla_m\nabla_n\Phi - \frac{\nabla^k\Phi}{\nabla^l\Phi\nabla_l\Phi}\left[2\nabla_k\nabla_m\Phi\nabla_n\Phi + 2\nabla_k\nabla_n\Phi\nabla_m\Phi - \Phi\nabla_k\nabla_n\nabla_m\Phi - g_{mn}\nabla_k\nabla^a\Phi\nabla_a\Phi\right]. \tag{5}$$

The above form of the Ricci tensor is more complex that the one shown in Eq. (1a) and consequently it is expected that the final form of Eq. (2) to be significantly different than the form shown in Eq. (4b). Forming now $R_{mn}\xi^n$ by utilizing (5), one notices that the first term in the square bracket of the right-hand side of (5) gives zero contribution. The second term after a trivial rearrangement, due to the Killing equation satisfied by ξ, also contributes zero. Finally, the last term results to a contribution proportional to ξ_m, and since $R_{mn}\xi^n$ is antisymmetrized with another copy of ξ it eventually contributes to zero. Therefore, one now gets

$$\nabla_m(\omega_{(\xi)a}\psi^a) = -2\epsilon_{mbcd}\psi^b\xi^c\left[\frac{1}{\Phi}\,\xi^n\nabla^d\nabla_n\Phi + \frac{1}{\nabla^l\Phi\nabla_l\Phi}\,\xi^n\nabla^k\Phi\nabla_k\nabla^d\nabla_n\Phi\right]. \tag{5a}$$

Even as it stands, by comparison to the steps that lead to the derivation of Eq. (4b) from Eq. (3) one realizes that the last term in (5a) contributes additional terms resulting into a considerable modification of the gradient of $\omega_{(\xi)a}\psi^a$. In Appendix B a rearrangement of this last term is presented that makes its contribution more transparent. Eventually from (5a), we are lead to the following equation satisfied by $\omega_{(\xi)a}\psi^a$:

$$\nabla_m(\omega_{(\xi)a}\psi^a) = -2\epsilon_{mbcd}\psi^b\xi^c\Bigg[-\frac{1}{\Phi}\,(\nabla^d\xi^n)\nabla_n\Phi + \frac{1}{\nabla^l\Phi\nabla_l\Phi}\,(-\nabla^k\Phi(\nabla^d\xi^n)\nabla_k\nabla_n\Phi$$
$$-\nabla^k\Phi(\nabla_k\xi_n)\nabla^d\nabla^n\Phi - R_k{}^d{}_{rn}\nabla^k\Phi\nabla^r\Phi\xi^n)\Bigg]. \tag{5b}$$

Repeating the arguments that lead to the derivation of Eq. (4b) from Eq. (3), one now gets

$$\nabla_m(\omega_{(\xi)a}\psi^a)=2(\omega_{(\xi)a}\psi^a)\,\frac{\nabla_m\Phi}{\Phi}+\frac{1}{\nabla^l\Phi\nabla_l\Phi}\,(\omega_{(\xi)a}\psi^a)\nabla^k\Phi\nabla_k\nabla_m\Phi$$

$$+2\epsilon_{mbcd}\psi^b\xi^c\,\frac{1}{\nabla^l\Phi\nabla_l\Phi}\,[\nabla^k\Phi(\nabla_k\xi_n)\nabla^d\nabla^n\Phi+R_k{}^d{}_{rn}\nabla^k\Phi\nabla^r\Phi\xi^n]. \qquad (5c)$$

This is the differential equation satisfied by $\omega_{(\xi)a}\psi^a$ whenever locally $1-a\Phi^2=0$ but its gradient is not vanishing.[24] Similarly the scalar $\omega_{(\psi)a}\xi^a$, satisfies an equation analogous to Eq. (5c) with the sole exception the axial Killing is switched with the timelike one and vice versa.

Comparing Eq. (5c) to Eq. (4b) one immediately notices an important difference: Whereas Eq. (4b) implies for instance that along the integral curves of any nonsingular vector field, $\omega_{(\psi)a}\xi^a$ satisfies a linear homogeneous equation, that is no longer the case for (5c). Of course some care must be exercized in interpreting the content of (5c). From its derivation it is clear that if indeed we were dealing with a theory where the Ricci tensor was given globally by Eq. (5), then one would easily have concluded from (5c) that for such theory circularity was broken. Within the present content, Eq. (5c) provides information regarding the gradient of $\omega_{(\psi)a}\xi^a$ at point (or points) of the space–time where $1-a\Phi^2=0$. Nevertheless we shall prove below that the information is sufficient, under certain conditions, in proving that circularity is broken for the theory (1a), (1b). To do so it is helpful first to advance an equivalent representation for each of the scalars $\omega_{(\xi)a}\psi^a=\epsilon_{abcd}\psi^a\xi^b\nabla^c\xi^d$, $\omega_{(\psi)a}\xi^a=\epsilon_{abcd}\xi^a\psi^b\nabla^c\psi^d$ that will be proven to be very helpful for the calculations that follow. We shall present the detailed calculations leading to the new representation only for the first one, because identical derivation holds for the second one. Obviously we have

$$\omega_{(\xi)a}\psi^a=\epsilon_{abcd}\psi^a\xi^b\nabla^c\xi^d=\epsilon_{abcd}\psi^a\xi^b\nabla^{(c}\xi^{d)}+\epsilon_{abcd}\psi^a\xi^b\nabla^{[c}\xi^{d]}=\epsilon_{abcd}\psi^a\xi^b\nabla^{[c}\xi^{d]}, \qquad (6)$$

where in arriving at the final equality, we have used the Killing equation satisfied by ξ. In above and in the following equations indices enclosed within square brackets ought to be antisymmetrized. Furthermore, since our analysis is purely local, it is convenient to work within the (t,ϕ,x,y) coordinate patch introduced earlier on and presently denote by (ξ,ψ,X,Y) the associated coordinate basis vector fields, respectively. In terms of this coordinate basis we may expand the antisymmetric part of $\nabla^{[a}\xi^{b]}$ in the following manner:

$$\nabla^{[a}\xi^{b]}=a_1\xi^{[a}\psi^{b]}+a_2\xi^{[a}X^{b]}+a_3\xi^{[a}Y^{b]}+a_4\psi^{[a}X^{b]}+a_5\psi^{[a}Y^{b]}+a_6X^{[a}Y^{b]}, \qquad (6a)$$

where the expansion coefficients a_i, $i=(1,...,6)$, are local bounded functions carrying all the informations regarding $\nabla^{[a}\xi^{b]}$. Although we can compute them explicitly, for the purpose of the following considerations their form is not required. Substituting Eq. (6a) into Eq. (6) we obtain

$$\psi_{(\xi)a}\psi^a=\tfrac{1}{2}a_6\epsilon_{abcd}\psi^a\xi^bX^cY^d. \qquad (6b)$$

This representation of $\omega_{(\xi)a}\psi^a$ is rather insightful, in the sense that all information regarding the circularity–noncircularity aspect of the action generated by the the Killing fields (ξ,ψ), is encoded in the coefficient a_6. For our purpose Eq. (6b) combined with (6a) will shorten considerably an otherwise messy and lengthy set of calculations.

Keeping in mind (6a) and (6b), we shall show in details that Eq. (5c) implies loss of circularity. At this point we have to make some assumptions concerning the distribution of zeros of the $1-a\Phi^2$ factor. As we have already argued earlier on, there is no fundamental reason to expect that for all stationary-axisymmetric solutions of (1a) and (1b) the factor $1-a\Phi^2$ is no where vanishing. If on the other hand we assume that $1-a\Phi^2$ admits zeros, it is expected that this set in the (x,y)-''plane,'' to be made either by isolated points, one-dimensional segments, or closed loops (and perhaps their unions). Our task now is to investigate the consequences of such configurations upon the circularity property of the underlying solution.

For that we shall first assume that on the (x,y)-plane the set of points where $1-a\Phi^2=0$ forms a nonintersecting smooth S^1-loop (smooth in the sense that the loop is minimally everywhere C^1, although the considerations that follow will remain intact if the loop is piecewise C^1). Let us here after denote the above loop by Γ. Due to stationary and axisymmetry of the background, such a set

of zeros for the $1-a\Phi^2$ factor, generate on each three $t=\text{const}$ spacelike section a two-dimensional torus, while in the space–time generate the product of a two torus with the real line. We may without loss of generality take the two coordinate basis vector fields X,Y defined earlier on, so that the first one is tangential to Γ while the other one is normal to the loop pointing ''inwards.'' [Such arrangement may require a change of coordinates on the (x,y)-plane, but basically it can be arranged.] Utilizing Γ we may locally foliate its vicinity by a family of loops. The construction of this local foliation is essentially equivalent to the construction of the two-dimensional Gaussian coordinates based on Γ. Each point of Γ is pushed to a fixed infinitesimal parameter distance along the geodesic integral curves originating at the point under consideration and having initial tangent the (two-dimensional) normal vector field of Γ. Relative to this foliation or coordinates, one then has the following behavior of Φ on Γ: $X^a\nabla_a\Phi=0$ and $Y^a\nabla_a\Phi\neq 0$. The idea now is to use Eq. (5c) to compute the gradients of $\omega_{(\xi)a}\psi^a$, $\omega_{(\psi)a}\xi^a$ along the tangential and normal directions of Γ. Again, we shall restrict considerations only to $\omega_{(\xi)a}\psi^a$, with the understanding that identical considerations hold for the other scalars as well.

Let us first compute the gradient of $\omega_{(\xi)a}\psi^a$ along the tangential direction X of Γ, i.e., form $X^m\nabla_m(\omega_{(\xi)a}\psi^a)$. This is given by contracting both sides of (5c) with X. One immediately notices that the contribution of first term in the right-hand side of (5c) is vanishing. Let us now shift attention to the terms within the square bracket of (5c) by first examining the term involving the Riemann tensor. Ignoring for the moment the innocuous and by assumption nonvanishing $\nabla^l\Phi\nabla_l\Phi$ term, we are interest to compute on Γ the following term:

$$I_1=\epsilon_{mbcd}X^m\psi^b\xi^c R_l{}^d{}_{rn}\nabla^r\Phi\nabla^l\Phi\xi^n. \tag{7}$$

Due to the symmetries of the Riemann tensor, the assumption that the gradient of Φ lies on the normal direction and the fact that the only nonvanishing component of $\epsilon_{mbcd}X^m\psi^b\xi^c$ must be taken along the Y direction, one concludes that $I_1=0$. Let us now consider the other term within the square bracket of Eq. (5c), i.e.,

$$I_2=\epsilon_{mbc}{}^d X^m\psi^b\xi^c\nabla_l\Phi(\nabla^l\xi^n)\nabla_d\nabla_n\Phi. \tag{7a}$$

We shall now rearrange I_2 so its content will be more clear. This, however, is not a straightforward task. A brute force computation of the the right-hand side of (7a) is a tedious and very lengthy. However, the identities (6a) and (6b) provide a shortcut. To make use of them, at first we substitute $\nabla^l\xi^n$ in (7a) above, by its antisymmetric part given by (6b). Because one of the indices of $\nabla^l\xi^n$ is contracted with the gradient of Φ, only the terms proportional to the terms a_6, a_5, a_3 in the expansion (6a) will survive i.e., we have

$$I_2=-\tfrac{1}{2}\epsilon_{mbc}{}^d X^m\psi^b\xi^c[a_6X^n+a_5\psi^n+a_3\xi^n]\nabla_d\nabla_n\Phi(Y^l\nabla_l\Phi). \tag{7b}$$

Now we need to compute the contraction of $\nabla_d\nabla_n\Phi$ with the three basis vectors in the square bracket of (7b). For that we should keep in mind that $\epsilon_{mbc}{}^d X^m\psi^b\xi^c$ will yield a nonzero contribution only when $\nabla_d\nabla_n\Phi=cY_dA_n$ with c in general nonvanishing and bounded and A_n a linear combination of the basis vectors. Furthermore notice we can write

$$X^n\nabla_d\nabla_n\Phi=\nabla_d(X^n\nabla_n\Phi)-(\nabla_dX^n)\nabla_n\Phi. \tag{7c}$$

Identical expressions to (7c) valid also for the case where X is replaced the Killing fields ξ, ψ, respectively. However, for the latter case the net contribution to I_2 is vanishing. This can be easily deduced since obviously for the case of ξ we have: $(\nabla^d\xi^n)\nabla_n\Phi=(\nabla^{[d}\xi^{n]})\nabla_n\Phi$. Substituting the antisymmetric part of the covariant derivative of ξ with its equivalent expression (6a), one eventually concludes that the net contributions of the a_5, a_3 terms to I_2 is vanishing. However, as can be easily seen the contribution arising from the term proportional to a_6 is not vanishing and in general it can be written in the following form:

$$I_2=-\tfrac{1}{2}a_6f\epsilon_{mbcd}X^m\psi^b\xi^cY^d(Y^l\nabla_l\Phi)=\omega_{(\xi)a}\psi^aF, \tag{7d}$$

where in passing to the last equality use of the representation (6b) has been made and the multiplicative factor F stands for all additional contributions arising from the gradient of Φ. Finally,

notice that the second term in the right-hand side of (5c) contributes a term proportional to $\omega_{(\xi)a}\psi^a$ in $X^m\nabla_m(\omega_{(\xi)a}\psi^a)$. It is not necessary to work out its exact form but it is suffice to comment that the proportionality function is necessarily a well behaved one. In view of the considerations so far and with reference to Eq. (5c), we conclude that $\omega_{(\xi)a}\psi^a$ satisfies a linear a equation along the tangential direction of the loop Γ. We may symbolically write it as $X^m\nabla_m(\omega_{(\xi)a}\psi^a)=G(\omega_{(\xi)a}\psi^a)$ with G a bounded function on Γ.

So far we have said nothing about the location of Γ relative the origin of the (x,y)-plane i.e., relative to the intersection of of the axis with the (x,y)-plane. Let us first assume, that this origin lies outside the the loop Γ. Then as we have shown earlier $\omega_{(\xi)a}\psi^a=0$, along the integral curves of any smooth vector field passing through the origin of (x,y)-plane and reaching the loop without intersecting regions or points where the factor $1-a\Phi^2$ is vanishing. By assumption the geometry of the space–time is nonsingular on and in the vicinity of the loop. This nonsingularity property of the geometry necessary dictates that $\omega_{(\xi)a}\psi^a$ must be minimally continuous on and also across the loop. If for instance, one assumes that at a point $\omega_{(\xi)a}\psi^a$ fails to be continuous, then necessarily the first derivatives of the metric must be discontinuous. In turn that implies that the curvature must exhibit singularities. As a consequence of the continuity requirement of $\omega_{(\xi)a}\psi^a$ along and across Γ coupled with the fact that outside of the loop $\omega_{(\xi)a}\psi^a=0$ one concludes from $X^m\nabla_m(\omega_{(\xi)a}\psi^a)=G(\omega_{(\xi)a}\psi^a)$ that $\omega_{(\xi)a}\psi^a$ is identically vanishing along the entire loop Γ. Similar conclusions apply also for the case where the loop Γ encloses the origin or even passes through the origin. In that case, the continuity requirement of $\omega_{(\xi)a}\psi^a$ implies that $\omega_{(\xi)a}\psi^a=0$ at the interior and on the loop.

Let us now consider the behavior of $\omega_{(\xi)a}\psi^a$ along the normal directions. We shall follow similar procedure as we did in computing the gradient of $\omega_{(\xi)a}\psi^a$ along the tangential direction. Accordingly, using (5c) we are first computing $Y^m\nabla_m(\omega_{(\xi)a}\psi^a)$ on Γ. For the present case, the analog I_1 of Eq. (7) reads

$$I_1=\epsilon_{mbcd}Y^m\psi^b\xi^c R_l{}^d{}_{rn}\nabla^r\Phi\nabla^l\Phi\xi^n, \tag{8}$$

thus the sole difference to the right-hand side of (7) is the exchange of the basis vector X for Y. However that difference carries enormous consequences. Since now the totally antisymmetric tensor is contracted with (Y,ξ,ψ), the only available contraction with the Riemann tensor is via the X direction. Consequently, and even though the other two indices of the Riemann tensor are contracted with the gradient of Φ the net effect is that in general

$$I_1=\epsilon_{mbcx}Y^m\psi^b\xi^c R_l{}^x{}_{rn}\nabla^r\Phi\nabla^l\Phi\xi^n\neq 0. \tag{8a}$$

Similarly the analog of (7b) presently takes the form

$$I_2=-\tfrac{1}{2}\epsilon_{mbc}{}^dY^m\psi^b\xi^c[a_6X^n+a_5\psi^n+a_3\xi^n]\nabla_d\nabla_n\Phi(Y^l\nabla_l\Phi). \tag{8b}$$

However, now again for the same reasons leading to the nonvanishing of (7d), one concludes that I_2 is nonvanishing and the terms proportional to a_6, a_5, a_3 contribute a term proportional to $\omega_{(\xi)a}\psi^a$ in $Y^m\nabla_m(\omega_{(\xi)a}\psi^a)$. We may eventually write the gradient of $\omega_{(\xi)a}\psi^a$ along the normal direction of the loop Γ in the following form:

$$Y^m\nabla_m(\omega_{(\xi)a}\psi^a)=2\epsilon_{mbcd}Y^m\psi^b\xi^c\,\frac{1}{\nabla^l\Phi\nabla_l\Phi}\,[R_k{}^d{}_{\rho n}\nabla^\rho\Phi\nabla^k\Phi\xi^n]=\frac{A}{\nabla^l\Phi\nabla_l\Phi}, \tag{9}$$

where A stands symbolically for the contribution of the Riemann tensor and in (9) we have ignored the terms proportional to $(\omega_{(\xi)a}\psi^a)$ since, as we have already discussed earlier on, $\omega_{(\xi)a}\psi^a=0$ along Γ. It is worth noticing that for a general stationary axisymmetric solution of (1a) and (1b), in general the contribution of $R_y{}^x{}_{yt}$ that appears in Eq. (9) will be different than zero. This term would be zero provided the geometry is invariant under $t\to -t$ i.e., the timelike Killing field ξ is hypersurface orthogonal. Thus under the mere assumption of stationarity and axisymmetry, the right-hand side of (9) would be different than zero. Before we continue it is worth reminding the reader that even though we have concentrated on the behavior of $\omega_{(\xi)a}\psi^a$, similar equations valid for the other scalar $\omega_{(\psi)a}\xi^a$. The detailed derivations are essentially identical and are omitted.

Let us now consider the implications of (9) upon circularity. We shall consider first the case where the loop Γ does not enclose the origin of the (x,y)-plane and for simplicity the factor $1-a\Phi^2$ is no where vanishing within Γ. As we have already discussed above continuity arguments imply that outside and on the loop $\omega_{(\xi)a}\psi^a=0$. But let us now shift attention to the value of $\omega_{(\xi)a}\psi^a$ at points interior to the loop. Accordingly in the interior of the loop $\omega_{(\xi)a}\psi^a$ satisfies Eq. (4b) and for every two points (1,2) belonging to the interior of Γ one has

$$\omega_{(\xi)a}\psi^a(1-a\Phi^2)^2|_1=\omega_{(\xi)a}\psi^a(1-a\Phi^2)^2|_2. \tag{9a}$$

Those are the only information, under the present conditions, that Eq. (4b) can provide. Let us, however, take an arbitrary point M on the loop Γ and consider for simplicity the geodesic emanating from the point M with initial tangent equal to the (inward pointing) normal at the M. Furthermore, we affinely parameterize the geodesic with the proper distance s, normalized so that $s=0$ at M. According to Eq. (9) the normal derivative of $\omega_{(\xi)a}\psi^a$ at M is different than zero, and that implies for points near by M and along this geodesic we may write

$$\omega_{(\xi)a}\psi^a|_s=\left[\frac{A}{\nabla^l\Phi\nabla_l\Phi}\,s+O(s^2)\right]. \tag{9b}$$

In other words the nonvanishing normal gradient of $\omega_{(\xi)a}\psi^a$ on Γ implies that in the interior of the loop necessary $\omega_{(\xi)a}\psi^a$ is not identically zero. That is sufficient to infer that circularity is indeed lost.

Similar conclusion regarding the value of $\omega_{(\xi)a}\psi^a$ valid also for the case where the loop Γ encloses or passes through the origin of the (x,y)-plane. The sole difference with the above case, lies in the fact that now $\omega_{(\xi)a}\psi^a$ is identically zero in the interior of the loop, but different than zero in the exterior region.

Notice furthermore, even though according to (9), $\omega_{(\xi)a}\psi^a$ is continuous on and across the loop, it fails to be differentiable in the normal direction. As a consequence the second normal derivatives of the metric must exhibit discontinuities. The situation here is similar to the constant density spherical star solution of Einstein equations (see, for example, Ref. 22). In that solution, whenever the surface of the star is matched to a vacuum solution, the curvature (for instance scalar curvature) exhibits discontinuities. The normal second derivatives of the metric are discontinuous across the surface. The same situation is encountered here as well. In conclusion, at points where $1-a\Phi^2$ passes through zeros, the geometry is characterized by a low differentiability class. It is that evidence that lets us conclude that analyticity of static solutions of (1a) and (1b) is likely absent, with the implications that we have already discussed in Sec. I. We shall say more about that elsewhere.

In summary we have shown that as a consequence of Eq. (5c) and in the case where the topology of zeros of the factor $1-a\Phi^2$ forms a smooth closed loop then circularity is necessarily broken. In principle one may apply similar considerations to the case where the factor $1-a\Phi^2$ is zero, for instance, along a linear segment. With the help of (5c) one can still prove that the tangential derivative $\omega_{(\xi)a}\psi^a$ is zero while the norm is nonvanishing. However, the conclusion may be different than the loop case. Let us suppose that the line segment along which $1-a\Phi^2$ is vanishing, is surrounded by a region where $\omega_{(\xi)a}\psi^a=0$. In that case we lead to the conclusion that the geometry is singular on the segment. However, if the region around the segment is characterized by the vanishing $\omega_{(\xi)a}\psi^a$ (such a situation can arise for instance whenever the segment is embedded within a loop) the geometry can be perfectly regular on the segment. Similar conclusions are valid for the case where $1-a\Phi^2$ is zero at an isolated point. In short, the issue of the zeros of $1-a\Phi^2$ needs additional attention. We may only add here that the fact that an arbitrary asymptotically flat solution of (1a) and (1b) may allow regions where $1-a\Phi^2=0$ it lead us to the conclusion that not all stationary asymptotically flat solutions of the theory are circular. There may exist torilike regions around the symmetry axis where circularity fails. It is of interest to note here that the underlying reason for the failure of circularity, is not really the mere presence of the second derivatives of the scalar field Φ in the right hand side of Eq. (1a) but rather is due to in the uncontrollable sign of the $1-\alpha\Phi^2$ factor combined with the existence of second derivatives.[25]

The above considerations remain valid whenever one considers asymptotically flat, stationary axisymmetric black hole solutions of the theory. One has no *a priori* reason to expect that 1

$-\alpha\Phi^2$ remains nonvanishing over the entire domain of the black hole exterior. Consequently at this stage we are lead to the conclusion that systems (1a) and (1b) may admit noncircular stationary axisymmetric black hole states. The nature of those black holes (if indeed they exist), in view of the discussions of the previous section, is unknown. We cannot any longer appeal to Carter's theorem, to conclude that their event horizon is a Killing horizon nor that they are uniformly rotating.

III. STATIONARY AXISYMMETRIC CIRCULAR STATES

Let us, in this section, concentrate on the sector of circular black hole states i.e., on the space of stationary-axisymmetric black holes where the Ricci tensor is circular at least in the domain of outer communication. We already know that the space of such black hole solutions of (1a) and (1b) is nonempty. The Kerr black hole is a ''trivial'' solution of Eqs. (1a) and (1b) with the field Φ zero everywhere. But do there exist other stationary-axisymmetric circular black hole states admitting a nontrivial conformal field? If so, what is the corresponding space–time geometries? The present section is intended to settle these questions. We shall show that actually all such black hole states must necessarily belong to the Kerr family, and thus the conformal field outside the horizon must be trivial.

One way of proving the statement is to make use of the fact that Eqs. (1a) and 1(b), via the action of a suitable conformal transformation, can be recast into an equivalent Einstein minimally coupled massless scalar field equation. In fact the conformal transformation accomplishing that is given by $\bar{g}_{mn}=(1-\alpha\Phi^2)g_{mn}$.[26] With further analysis, one can show that the above conformal transformations maps black hole solutions of the conformal system to that of the minimal one and vice versa. However, in the present paper we shall not follow this route. Instead, we shall work with the original systems (1a) and (1b). As it will become clear further below, the method of establishing the claim, utilizes only the wave equation obeyed by the scalar field, the Killing property of the horizon and asymptotic flatness and thus is actually independent upon the precise form of the field equations.

Let us begin by assuming that (1a) and (1b) admit a stationary, axisymmetric, circular asymptotically flat black hole away from the Kerr family. According to Carter's theorem, the location of the event horizon is marked by the vanishing of the norm of Killing field $X_1=\xi+\Omega\psi$, with the nonzero constant Ω defining the angular velocity of the event horizon. From Eq. (1b) after multiplication by Φ one easily obtains

$$\nabla_a(\Phi\nabla^a\Phi)-\nabla^a\Phi\nabla_a\Phi=0. \tag{10}$$

We shall integrate the above equation over a suitable space–time region Ω. This region Ω is chosen as follows: First we consider an arbitrary spacelike hypersurface labeled as t_1, intersecting only the future event horizon H^+ and being deformed so that it becomes asymptotically null.[27] We translate t_1 along the orbits of the Killing field ξ by a specified amount of parameter distance and label the resulting hypersurface as t_2. (In the conformal completion of the space–time, such hypersurfaces intersect the future null infinity J^+ across regular two spheres.) The region Ω, is the region enclosed between the two hypersurfaces, the future event horizon H^+, and future null infinity J^+. By integrating Eq. (10) over Ω and with the help of Stoke's theorem one easily obtains

$$\int \nabla_a\Phi\nabla^a\Phi d\Omega|_\Omega=\int \Phi\nabla^a\Phi d\Sigma_a|_{t_1}+\int \Phi\nabla^a\Phi d\Sigma_a|_{t_2}+\oint \Phi\nabla^a\Phi d\Sigma_a|_H+\oint \Phi\nabla^a\Phi d\Sigma_a|_\infty. \tag{10a}$$

However, because of translational invariance the volume integral over the two hypersurfaces (t_1,t_2) cancel out. The only potentially nonvanishing contributions to the right-hand side of above, come from the integral along H^+ and J^+, respectively.[28] Let us first deal with the surface integral across the horizon. We shall employ as the measure $d\Sigma_A$ the standard one (see, for instance, Ref. 15) i.e.,

$$d\Sigma_a=dS_{ab}l^b=(-l_an_b+l_bn_a)l^bdSdu, \tag{10b}$$

where l is the tangent field to the null generator field of H^+ while n is the other null future directed field normalized according to $l^a n_a = -1$.

One then finds

$$\oint \Phi \nabla^a \Phi d\Sigma_a|_H = \oint \Phi \nabla^a \Phi(-l_a n_b + l_b n_a) l^b dS du = \oint \Phi[(-n_a l^a) L_l \Phi + (l^a l_a) L_n \Phi] dS du, \tag{11}$$

with L_x signifying Lie differentiation along x. On the horizon, by assumption the Killing field X_1 coincides with l. Consequently one may conclude that the first term in the integrand of (11) always vanishes, provided Φ is bounded on the horizon. We shall impose this requirement, but we may stress that by allowing an unbounded Φ on the horizon requires a separate treatment. Under this assumption, it would be sufficient to show that the Lie derivative of Φ along the other null vector field n orthogonal to l, computed on the horizon is vanishing or more precisely $l^a l_a L_n \Phi|_{H^+} = 0$. On physical grounds one expects that to be the case, since otherwise there would be fluxes of scalar radiation across the horizon, a situation that is not expected to occur for a stationary state. Indeed we shall explicitly verify that expectation, by introducing a set of local regular across the horizon coordinates, and then studying the behavior of Φ near the vicinity of the horizon. In Ref. 1 we have explicitly constructed such a coordinate system, which, however, made use of Einstein field equations via the employment of Israel's formalism[29,30] as well as the hypersurface orthogonality of the timelike at infinity Killing vector field. For the present case, such construction it is not easily adaptable. However, in Ref. 1 we have also pointed out that one could have employed one of the regular on the horizon coordinate systems discussed in the work of Racz and Wald[31] (see also the account of Carter in Ref. 15). The Racz–Wald coordinate systems are rather advantageous, since their construction is solely based on the geometrical structure of the horizon and are being entirely independent upon the underlying field equations. For the present purpose we shall only make use of their generalized advanced Eddington–Filkenstein type coordinate system and for completeness purposes a brief outline of the basic steps involved in the construction of these coordinates will be presented.

Since by assumption the horizon admits topological spherical cross sections one can always consider a small open neighborhood A lying on an arbitrary but fixed here after section of the horizon. Introduce an arbitrary local coordinate patch x^i, $i=1, 2$ covering A and denote by A the portion of the horizon lying on the null geodesics intersecting A. Lie-drag x^i along the null generators piercing A and introduce further the Killing parameter u so that on the horizon $l = \partial/\partial u$ with $-\infty < u < \infty$ and $u=0$ on A. For any point p in A consider the other null vector n defined by $n^a l_a = -1$ and being orthogonal to other two coordinate basis vectors attached to x^i. By construction u, x^i is a local coordinate covering of A. Next, consider the null geodesics emanating from the horizon, with initial tangent n and introduce an affine parameter r along them so that $r=0$ on A. Lie-drag u, x^i along those geodesics thus locally extending u, x^i of A. By construction (u, r, x^i) is a local coordinate system covering a small space–time neighborhood of the horizon. This coordinate system for the obvious reasons is termed in Ref. 31 as a (local) Eddington–Filkenstein coordinate system. After glopalizing the construction over the entire cross section of the horizon the space–time metric in this neighborhood reads, $V^2 = -X_1 \cdot X_1$[31]

$$ds^2 = -V^2 du^2 + 2 du dr + 2 g_{ui} du dx^i + g_{ij} dx^i dx^j. \tag{12}$$

Relative to such system, near the horizon the timelike Killing field X_1 admits the following representation: $X_1 = \partial/\partial u$. It is worth pointing out that the so far the construction is valid irrespective of whether the orbits of l are incomplete or complete null geodesics and consequently this local Eddington–Filkenstein coordinate system is valid for any horizon. (Consequences of this observation will be discussed further below.) Note also that the field $n = -\partial/\partial r$ satisfies $n^a n_a = 0$ and on the horizon $l^a n_a = -1$.

Consider now the following null vector field:

$$\underline{l} = \frac{\partial}{\partial u} + \frac{1}{2} V^2 \frac{\partial}{\partial r}. \tag{13}$$

It is well defined near the horizon and smoothly reduces to the Killing field l on the horizon. Because Φ remains invariant along the orbits of the Killing field $X_1=\partial/\partial u$ one deduces that on and of the horizon

$$\overline{l}(\Phi)=\frac{1}{2}\,V^2\,\frac{\partial\Phi}{\partial r}. \tag{14}$$

As however the horizon is approached the left-hand side of the above equation is by itself vanishing, implying that on the horizon

$$V^2\,\frac{\partial\Phi}{\partial r}=-V^2n^a\,\frac{\partial\Phi}{\partial x^a}=-V^2L_n\Phi=0. \tag{15}$$

In view of those considerations we conclude that

$$\oint \Phi\nabla^a\Phi\, d\Sigma_a|_H=0. \tag{16}$$

Let us now shift considerations to the other boundary integral, i.e.,

$$I=\oint \Phi\nabla^a\Phi\, d\Sigma_a|_\infty\,. \tag{17}$$

We shall prove that its contribution is vanishing as well. To do so, we shall exploit the fact that any asymptotically flat black hole solution by definition is asymptotically flat at future null infinity J^+.[32] Using the existence of J^+, in principle we could have treated the above integral exactly in the same manner as its counterpart along H^+. In fact in the unphysical manifold i.e., the conformal completion of the physical space–time, (17) is an integral along J^+ taking between two specified smooth cross sections of J^+. The appropriate measure along J^+ is identical to that (10b) with suitable reinterpretation of the null fields l and n, respectively. One simply has to transform the integrand over the unphysical space–time and proceed with its computation. Equally well, however, we could compute I directly over the physical space–time. We shall pursue the second approach. To do so, we shall introduce first, Bondi coordinates,[33] (u,r,x^i) covering the asymptotic region between and around the two asymptotically null hypersurfaces (t_1,t_2). Bondi coordinates in the present discussion will play the analogous role that the generalized advanced Eddington–Filkenstein coordinate system played earlier on. By definition, $u=con$ is a family of null hypersurfaces ''intersecting'' J^+, r is an affine parameter along the null geodesics of the u spaces, and x^i parametrize the null geodesics, respectively. For an asymptotic at null infinity flat space–time and relative to those coordinates, the nonvanishing components of the physical metric are exhibiting the following asymptotic fall of rates[33]

$$g^{rr}=O(r),\quad g^{ri}=O(1),\quad g^{ij}=O(r^{-2}), \tag{18}$$

and by construction over the entire domain of validity of the Bondi chart, $g^{ur}=1$. On the other hand if by $l^m=g^{mn}(\partial u/\partial x^n)$ denote the future point tangent to outgoing null geodesics, then asymptotic flatness at J^+[34] demands

$$R_{mn}l^ml^n=O(r^{-4}). \tag{19}$$

From Eq. (1a) and after ignoring for the moment the irrelevant $1-\alpha\Phi^2$ factor, the above condition implies that asymptotically

$$L(\Phi)=4\left(l^m\,\frac{\partial\Phi}{\partial x^m}\right)^2-2\Phi l^ml^n\,\frac{\partial^2\Phi}{\partial x^m\partial x^n}=O(r^{-4}). \tag{20}$$

In deriving this expression, we are taking into account that the covariant components of the asymptotic metric in Bondi coordinates have the same form as in Eq. (12), and consequently

$l^m l^n \Gamma^a_{mn}(\partial\Phi/\partial x^a)=0$. Therefore, any solution of the wave Equation (1a) compatible with asymptotic flatness at null infinity must exhibit the following asymptotic behavior:[35]

$$\Phi = O(r^{-1}), \quad l^m \frac{\partial \Phi}{\partial x^m} = O(r^{-2}), \quad l^m l^n \frac{\partial^2 \Phi}{\partial x^m \partial x^n} = O(r^{-3}). \tag{21}$$

Relative to those asymptotic coordinates the integral (17) can be written as

$$I = \oint \Phi \nabla^a \Phi d\Sigma_a|_\infty = \lim \oint \Phi \nabla^a \Phi d\Sigma_a = \lim \oint \Phi g^{ab} \frac{\partial \Phi}{\partial x^a} n_b dV, \tag{22}$$

where the limit is understood as the affine parameter r tends to infinity. The integral is computed along an $r=$const timelike three surface with a spacelike outgoing pointing unit normal $n_a = \delta^r_a/(g^{rr})^{1/2}$ and $dV=\frac{1}{6}(-g)^{1/2}e_{abc}dx^a dx^b dx^c$ with $(-g)$ the determinant of the induced three metric on the $r=con$ hypersurface. Taking into account the quoted fall of rates in (18), and the fact $(-g)^{1/2}=O(r^2)$, when the the integral in (22) is computed and one passes to the limit of infinity affine distance, gets $I=0$.

From the discussion so far and with reference to integrals (10a), (16), and (17) one concludes that the field Φ is trivial outside the horizon. In turn this implies that all stationary-axisymmetric black hole solutions of theories (1a) and (1b) are Ricci flat. Appealing to well-known uniqueness theorems[36] for the Einstein vacuo equations we conclude that the space of stationary-axisymmetric black hole solutions of systems (1a) and (1b) is described by the familiar Kerr family.

Notice that as a byproduct of the fact that the generalized Eddington–Filkenstein coordinates are entirely independent of the value of the surface gravity of the horizon, the above discussion shows that the conformal system (modulo the boundenes property) does not admit static black holes possessing degenerate horizons,[37] a conclusion that was not covered in the approach of Ref. 1. The stationary sector contains extremal states, namely the maximally rotating Kerr states.

IV. DISCUSSION

It is clear from the above sections that the conformal system brings out challenges requiring additional work. Primarily the present work does not exclude the existence of stationary but not necessarily axisymmetric black hole states. As we have emphasized the employment and conclusions of Carter's work preassumes that one is working with the stationary-axisymmetric circular sector of the theory. It may be that the stationary, nonaxisymmetric sector is empty particularly in view of CT discussed in the Introduction. The new approach initiated by Chruściel[7,8] is inspiring and highly motivating. Are the actions predicted by CT necessarily circular? Can the CT theorem be adopted to the peculiarities of the conformal systems (1a) and (1b)? What are the consequences of the failure of the null convergence condition for the topological and metrical properties of black hole space–times? It may be that some of the questions raised are trivial but for the moment the answers are not known. Closely related to those issues is a quest for a better understanding of the nature of the domain of outer communication of black hole solutions of (1a) and (1b). In many of the so far established uniqueness theorems[36] considerations are restricted to black hole states where the stationary, timelike at infinity Killing field, remains timelike at the entire domain of outer communication and only becomes null at the boundary i.e., on the event horizon. This assumption also employed in the construction of the original uniqueness theorems for the Einstein–Maxwell case, although convenient, leaves open the possibility that there may exist black hole characterized by a nontrivial ergoregion entirely disconnected from the event horizon. For the case of the Einstein–Maxwell system, only very recently the work of Sudarsky and Wald[38] settled the issue. For the Einstein conformally invariant system or other systems that issue remains open.

Beyond the above considerations, one would also have had to consider the challenge posed by the Bekenstein solution.[2] Recall that this solution is an exact static, spherically symmetric, asymptotically flat solution of the Einstein conformally invariant system characterized by the property that the scalar field becomes unbounded on a regular horizon. Ignoring the fact that it is unstable, the solution was interpreted as a black hole. Recently, however, a critical examination of the unboundedness property of the scalar field has been carried out in Ref. 39. It was proven in Ref.

39 that the unboundedness property carries severe consequences. The Einstein field equations fail to be satisfied on the horizon and consequently the solution does not represent a genuine black hole. For the stationary case, proving in the last section that all stationary-axisymmetric circular states belong to the Kerr family, we have made use of the boundedness assumption of the scalar field on the horizon. Preliminary work indicates that unbounded fields on the rotating horizon leads to the same type of problems as for the nonrotating case. Those issues, however, will be discussed elsewhere.

In conclusion we may add that with the removal of the Bekenstein solution as a possible black hole, the present work on the conformal system combined with the work of other researchers on the Einstein scalar field system characterized by a variance of couplings (for minimal coupling see Refs. 40 and 41, for conformal,[1] arbitrary coupled[42,43] and for particular kind of self interaction[44–47]), strongly suggest that the following conjecture put forward in Ref. 39 appears to be correct: Spherical black holes cannot be endowed by (classical) scalar hair alone.

ACKNOWLEDGMENTS

Part of this work was carried out while the author was supported by a NAFTA Res. Fellowship administered by the Canadian NSERC. It has been completed, via the support of a Catedral Patrimonial administered by Conacyt-Mexico, and by a grant from Coordination-Cientifica of UMSNH. The author is thankful to the above institutions. Thanks are also due to Daniel Sudarsky for many discussions related to the content of Ref. 38.

APPENDIX A: ZEROS OF $1-a\Phi^2$ AND THE VALUES OF RICCI

In this Appendix we shall examine the following question. Suppose in an open neighborhood of a space–time point P, (g,Φ) is a singularity free solution of the Eqs. (1a) and (1b). Further suppose at P or more generally in a vicinity of P, the scalar field Φ satisfies: $1-a\Phi^2=0$. What then would Eqs. (1a) and (1b) tell us about the components of the Ricci tensor at P?

To address this question, let us first denote by S_{mn} the right-hand side of (1a) i.e., define S_{mn} by

$$S_{mn}=\alpha[4\nabla_m\Phi\nabla_n\Phi-2\Phi\nabla_m\nabla_n\Phi-g_{mn}\nabla^l\Phi\nabla_l\Phi]. \tag{A1}$$

The singularity free character of the geometry would then demand that the tensor S_{mn} must be vanishing at the point P. Accordingly the true value of $R_{mn}(P)$ would be determined via a limiting procedure. Such procedure would be sufficient for the determination of $R_{mn}(P)$, since by assumption the geometry is regular at P and consequently the components of the Ricci are minimally continuous at P provided they are expressed in a regular set of coordinates covering the vicinity of P. Relative to such setting, the components of the Ricci are obtained by first performing a covariant type Taylor series expansion of the tensor field S_{mn} and the scalar $1-a\Phi^2$. If for instance, the coordinates of P and a nearby point Q collectively are denoted by (x,y), respectively, and consider the geodesic joining the two (assuming it exists and is unique), then one may compute S_{mn} and $1-a\Phi^2$ at Q, take their ratio, and then the limit of the ratio, as Q is sliding along the geodesic towards P. In that manner

$$R_{mn}(P)=\lim\frac{S_{m'n'}(Q)}{1-a\Phi(Q)^2}, \tag{A2}$$

where the prime over the tensor indices signifies a tensor indices at the point Q. Formally one would then develops a power series for $S_{m'n'}(Q)$ and $1-a\Phi(Q)^2$, i.e., write

$$\delta^{m'}_m\delta^{n'}_n S_{m'n'}(Q)=A_{mn}+A_{mnl}\sigma^l+O(\sigma^l)^3, \tag{A3}$$

with a similar expression for the $1-a\Phi(Q)^2$. In above, $\sigma^m=\nabla^m\sigma(P,Q)$ where $\sigma(P,Q)$ is half of the square of the geodesic distance between Q,P and $\delta^{n'}_m$ is the bitensor of parallel transport.[48] The expansion coefficients can then computed by differentiating the series and take the coincidence limits.[48] The above procedure yields the value of Ricci, but it preassumes analyticity for both the space–time geometry and the scalar field Φ. Such assumption is unjustified. Even though

vacuum static or stationary solution of Einstein equation respect analyticity,[13] at least in the domain where the Killing field is timelike, it is not known whether the same property holds true for the nonvacuum case as for the present situation. Currently we have the question of analyticity of solutions of (1a) and (1b) under scrutiny and hope to come back.[14] We shall avoid the assumption of analyticity by resorting to a different method.

Writing the field Eq. (1a) in the form

$$(1-a\Phi^2)R_{mn}=S_{mn}, \tag{A4}$$

and covariantly differentiating both sides we obtain

$$-2a\Phi\nabla_k\Phi R_{mn}+(1-a\Phi^2)\nabla_k R_{mn}=\nabla_k S_{mn}. \tag{A5}$$

Evaluating the above equation at the points where $1-a\Phi^2=0$ one gets

$$-2a\Phi\nabla_k\Phi R_{mn}=\nabla_k S_{mn}. \tag{A6}$$

Assuming that the gradients of Φ are nonvanishing and contracting the above equation with $\nabla^k\Phi$ one gets after a trivial rearengement

$$\Phi R_{mn}=\nabla_m\nabla_n\Phi-\frac{\nabla^k\Phi}{\nabla^l\Phi\nabla_l\Phi}[2\nabla_k\nabla_m\Phi\nabla_n\Phi+2\nabla_k\nabla_n\Phi\nabla_m\Phi-\Phi\nabla_k\nabla_n\nabla_m\Phi-g_{mn}\nabla_k\nabla^a\Phi\nabla_a\Phi], \tag{A7}$$

which is the representation for the Ricci tensor given in the text.

APPENDIX B: DERIVATIVES OF Φ AND THE RIEMANN TENSOR

In this Appendix we shall manipulate further the third derivatives of the field Φ appearing in the right-hand side of Eq. (7) and shall give all the intermediate calculations leading to the derivation of Eq. (8). Let us begin with the term $\nabla_k\nabla_d\nabla_n\Phi$ and contract it with the ξ^n. Commuting the two covariant derivatives using the well-known identity: $\nabla_a\nabla_b A_c-\nabla_b\nabla_a A_c=R_{abc}{}^d A_d$ we obtain the following series of operations:

$$\begin{aligned}
\xi^n\nabla_k\nabla_d\nabla_n\Phi&=\xi^n\nabla_d\nabla_k\nabla_n\Phi-\xi^n R_{kd\rho n}\nabla^\rho\Phi\\
&=\nabla_d(\xi^n\nabla_k\nabla_n\Phi)-(\nabla_d\xi^n)\nabla_k\nabla_n\Phi-\xi^n R_{kd\rho n}\nabla^\rho\Phi\\
&=\nabla_d[\nabla_k(\xi^n\nabla_n\Phi)]-\nabla_d[(\nabla_k\xi_n)\nabla^n\Phi]-(\nabla_d\xi^n)\nabla_k\nabla_n\Phi-R_{kd\rho n}\nabla^\rho\Phi\xi^n\\
&=-\nabla_d[(\nabla_k\xi_n)\nabla^n\Phi]-(\nabla_d\xi_n)\nabla_k\nabla^n\Phi-R_{kd\rho n}\nabla^\rho\Phi\xi^n\\
&=-(\nabla_d\nabla_k\xi_n)\nabla^n\phi-(\nabla_k\xi_n)\nabla_d\nabla^n\Phi-(\nabla_d\xi_n)\nabla_k\nabla^n\Phi-\xi^n R_{kd\rho n}\nabla^\rho\Phi.
\end{aligned}$$

Using further the well-known identity valid for any Killing vector field

$$\nabla_d\nabla_k\xi_n=R_{ldkn}\xi^l,$$

we can eliminate the second derivatives of ξ at the expense of introducing the Riemann tensor. We then obtain

$$\xi^n\nabla_k\nabla_d\nabla_n\Phi=-R_{ldkn}\xi^l\nabla^n\Phi-(\nabla_k\xi_n)\nabla_d\nabla^n\Phi-(\nabla_d\xi_n)\nabla_k\nabla^n\Phi-\xi^n R_{kd\rho n}\nabla^\rho\Phi.$$

The contribution of the first Riemann tensor to Eq. (7) is vanishing since it is additionally contracted with $\nabla^k\Phi$. Therefore

$$\xi^n\nabla^k\Phi\nabla_k\nabla_d\nabla_n\Phi=-\nabla^k\Phi[(\nabla_k\xi_n)\nabla_d\nabla^n\Phi+(\nabla_d\xi_n)\nabla_k\nabla^n\Phi]-R_{kd\rho n}\nabla^\rho\Phi\nabla^k\Phi\xi^n,$$

which is the expression used in the text.

[1] T. Zannias, J. Math. Phys. **9**, 2643 (1995).
[2] There has been a significant amount of work aiming to explore consequences of the coupling of the scalar field to background scalar curvature at lower than four, four, and higher than four dimensions. See, for instance, C. Martinez *et al.*, Phys. Rev. D **54**, 3830 (1996); C. Klimcic, J. Math. Phys. **34**, 1914 (1993); B. C. Xanthopoulos and A. Dialynas *ibid.* **33**, 1463 (1992); J. D. Bekenstein, Ann. Phys. (N.Y.) **82**, 535 (1974); **91**, 72 (1975); B. C. Xanthopoulos and T. Zannias, J. Math. Phys. **32**, 1875 (1991); **33**, 1462 (1992); K. Bronnikov and K. Kireyev, Phys. Lett. A **67**, 95 (1978).
[3] S. W. Hawking, Commun. Math. Phys. **25**, 152 (1972).
[4] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space–time* (Cambridge University Press, Cambridge 1973).
[5] S. W. Hawking, Commun. Math. Phys. **25**, 152 (1972).
[6] P. T. Chruściel, Contemp. Math. **170**, 23 (1994); P. T. Chruściel and R. M. Wald, Class. Quantum Grav. **147**, 11 (1994).
[7] P. T. Chruściel In Journées Relativistas, Ascona (1996) and gr-qc 9610010.
[8] P. T. Chruściel, gr-qc 9610011.
[9] Condition (c), i.e., the spherical like cross sections of the horizon, actually follows from other considerations of the CT. In particularly the CT assumes that the domain of outer communication is simply connected. Such property is guaranteed as long as the Ricci tensor satisfies the null convergence condition, and additionally the null infinity satisfies some additional conditions referred as regularity conditions (or some equivalent version as those for example spell out in the Ref. 10 below.) As a byproduct of the satisfaction of the null convergence condition and regularity of the null infinity it then follows that the event horizon necessarily possess spherical like cross sections. For a proof see P. T. Chruściel and R. M. Wald in Ref. 6 above as well as the following Ref. 10.
[10] G. Galloway, Class. Quantum Grav. **13**, L99 (1995).
[11] M. Heusler, gr-qc 96077001.
[12] R. Beig and P. T. Chruściel, gr-qc 9510015, in particular notice the assumptions underlying Theorem (4.1). Prerequisite for the timelike character of the ADM four momenta is the existence of appropriate spacelike slices where matter and geometry exhibit special fall of rates.
[13] M. zum Hagen, Proc. Cambridge Philos. Soc. **67**, 415 (1970); **68**, 187 (1970).
[14] J. Estevez Delgado and T. Zannias (in preparation).
[15] B. Carter, in *Black Holes*, edited by C. DeWitt and B. S. DeWitt (Gordon and Breach, New York, 1973).
[16] B. Carter, in *Gravitation in Astrophysics*, edited by B. Carter and J. B. Hartle (Kluwer, Gargese, 1986).
[17] The construction of $\langle\langle M\rangle\rangle$ used in Carter's work is to be contrasted with the corresponding definitions in Refs. 7 and 8. Chruściel makes heavy use of the concept of completeness of trajectories of the asymptotically timelike Killing vector field.
[18] There is no loss of generality in imposing the commuting property of the two actions. The justification for this can be found in B. Carter, Commun. Math. Phys. **17**, 233 (1970), In fact a slightly more general conclusion can be drawn whenever axisymmetry action is replaced by the closely related but not identical concept of cyclic symmetry. See, for instance, discussion in the following Ref. 19.
[19] L. B. Szabados, J. Math. Phys. **28**, 2688 (1987).
[20] B. Carter, J. Math. Phys. **10**, 70 (1969).
[21] Even though considerations in the present section are restricted to the conformal coupling, i.e., for $\alpha=k/6$, the same calculations of the present section can be repeated for the case whenever an arbitrary coupling of the scalar field Φ to the background scalar curvature is considered. For such arbitrary coupling one finds $\alpha=\xi k$ with ξ a real parameter and the structure of the field Eqs. (1a) and (1b) remain the same although numerical coefficients will be altered. The inclusion of an arbitrary potential $V(\Phi)$ adds a term $g_{mn}V(\Phi)$ to the right-hand side of (1a).
[22] R. M. Wald, *General Relativity* (Chicago University Press, Chicago, 1984).
[23] In the case where the gradient of Φ is also identically vanishing one can still apply the method of the Appendix A to compute the Ricci tensor. In such event Eq. (A4) of Appendix A is differentiated twice, and one follows the same steps leading to (A7). Note in that event, the resulting Ricci contains fourth-order derivatives of the field Φ.
[24] In the case of vanishing gradients, even though the resulting form of Eq. (9) is more complex it nevertheless maintains similar structure. An inhomogeneous term is also appearing.
[25] In this regard it is of interest to repeat the calculations for arbitrary coupling of the scalar field Φ to scalar curvature. In this case one expects, if ξ is chosen to be negative circularity ought to hold although this has to be verified by detailed computations.
[26] A more extensive discussion about this transformation can be found in J. D. Bekenstein in Ref. 2; B. C. Xanthopoulos and A. Dialynas in Ref. 2; T. Zannias in Ref. 1. We may parenthetically add here that because of this property of the Einstein conformal system, asymptotically flat nonsingular initial should exist, possessing positive ADM mass and timelike four momenta.
[27] We could have assumed that those hypersurfaces merely intersect null infinity. Their asymptotic identification to some retarded time coordinates simply makes the evaluation of integral of Eq. (17) cited below an easier task.
[28] Although we are working in the physical space–time occasionally we shall use terminology associated with the conformal completion of the physical space–time.
[29] W. Israel, Phys. Rev. **164**, 1776 (1967).
[30] W. Israel, Commun. Math. Phys. **8**, 245 (1968).
[31] I. Racz and R. M. Wald, Class. Quantum Grav. **9**, 2643 (1992).
[32] Although there exist diverse definitions of asymptotic flatness at null infinity (see for instance discussion in Chap. 11 of Walds book, Ref. 12) nevertheless all proposed definitions agree on the existence of a smooth null hypersurface J^+ constituting part of the space–time ''boundary.'' In the present context asymptotic flatness implies the existence of J^+. In turn for nonvacuum spaces that requires, besides the peeling property of the Weyl tensor, certain asymptotic fall of rates for the Ricci and scalar curvature. For a relevant discussion consult Refs. 33 and 34 below.
[33] See, for instance, E. Newman and R. Penrose, J. Math. Phys. **5**, 566 (1962).
[34] For a discussion concerning the asymptotic decay of the curvature for a vacuo, asymptotically flat at null infinity space

see Ref. 33, while for electrovac, or general arbitrary nonvacuo space–times consult the following Ref., respectively; A. R. Exton, E. Newman, and R. Penrose, J. Math. Phys. **10**, 1566 (1969); G. Ludwig, Gen. Relativ. Gravit. **7**, 293 (1976).

[35] One source of concern is the possible existence of solutions f of the wave equation satisfying additionally $L(f)=0$ and exhibiting fall of rates of the form $f=O(r^a)$ with a in general nonvanishing positive or negative exponent. If such solutions exist, then due to the nonlinear structure of the operator L there may exists combinations of the form $\Phi=f+\phi$ with ϕ fulfilling the asymptotic conditions (21), so that that $L(f+\phi)=O(r^{-4})$. In such an event the integral (17) may not be vanishing. However, such possibility cannot occur. One can easily show that all f satisfying $L(f)=0$ have the form: $f=1/[K_0(x^i)-K_1(x^i)r]$ with $K_a(x^i)$ $a=0,1$ arbitrary integration ''functions.'' A superposition then of the form $\Phi=f+\phi$ satisfies (21). The trivial combination $\Phi=1/K_0(x^i)+\phi$ is not compatible with (19) unless $1/K_0(x^i)=0$.

[36] For an overview, as well as an update on the various black hole uniqueness theorems, see, for instance, the recent monograph: M. Heusler, *Black Hole Uniqueness Theorems* (Cambridge Univ. Press, Cambridge, 1996).

[37] See also T. Zannias, *Talk delivered in, The Second Mexican School on Gravitation and Mathematical Physics*, edited by D. Nunez *et al.* (1997).

[38] D. Sudarsky and R. M. Wald, Phys. Rev. D **46**, 1453 (1992); **47**, R5209 (1993).

[39] D. Sudarsky and T. Zannias, Phys. Rev. D (in press, 1998).

[40] J. D. Bekenstein, Phys. Rev. D **5**, 2941 (1972).

[41] J. E. Chase, Commun. Math. Phys. **19**, 276 (1970).

[42] A. Saa, J. Math. Phys. **37**, 2346 (1996).

[43] A. Saa, Phys. Rev. D **53**, 7373 (1996).

[44] J. D. Bekenstein, Phys. Rev. D **51**, R6608 (1995); A. E. Mayo and J. D. Bekenstein, *ibid.* **54**, 5059 (1996).

[45] D. Sudarsky, Class. Quantum Grav. **12**, 579 (1995); H. Penna and D. Sudarsky [(Preprint, UNAM (1997)].

[46] M. Heusler, J. Math. Phys. **33**, 3497 (1992).

[47] M. Heusler, Class. Quantum Grav. **12**, 779 (1992).

[48] This object has been examined at length in, J. L. Synge, *Relativity. The General Theory* (North Holland, Amsterdam, 1960); Further applications of it, as well as the idea of covariant type Taylor series expansion can be found in: B. S. Dewitt, in *Relativity Groups and Topology* (Les Houche 1963 Gordon-Breach).

Integrable discretizations of the Euler top

A. I. Bobenko,[a] B. Lorbeer,[b] and Yu. B. Suris[c]
Fachbereich Mathematik, Technische Universität Berlin,
Str. 17 Juni 135, 10623 Berlin, Germany

(Received 15 June 1998; accepted for publication 17 July 1998)

Discretizations of the Euler top sharing the integrals of motion with the continuous time system are studied. Those of them which are also Poisson with respect to the invariant Poisson bracket of the Euler top are characterized. For all these Poisson discretizations a solution in terms of elliptic functions is found, allowing a direct comparison with the continuous time case. We demonstrate that the Veselov–Moser discretization also belongs to our family, and apply our methods to this particular example. © *1998 American Institute of Physics.*
[S0022-2488(98)04211-X]

I. INTRODUCTION

The subject of integrable discretizations of integrable dynamical systems is slightly more than twenty years old. During the first decade of its existence several different approaches were proposed for discretizing soliton equations,[1–5] but neither of them dealt with integrable systems of the classical mechanics. The first examples in this important subarea seem to appear about ten years ago.[6,7] Veselov's paper contained discrete time versions of such classical integrable systems as the Euler case of the rigid body motion and the Neumann system. The algebraic construction behind these examples was later elaborated in more detail in Ref. 8. However, the nature of these exceptionally beautiful examples remains somewhat mysterious. They still resist being included in any of the existing general frameworks for integrable discretizations (cf. Ref. 9). The origin of the Veselov–Moser discretizations, their place in the corresponding continuous time hierarchies, their relations to other existing examples, and several further points remain to be clarified. The problem of comparing the explicit solutions found in Refs. 6–8 with the classical continuous time counterparts was also left open (despite the fact that on the level of equations of motion the continuous limit is easy to perform).

The present work may be considered as a sort of extensive comment on Refs. 6 and 8 in the part concerned with the Euler top. We do not close all the open problems mentioned above, but we do hope to bring some light into some of them. In particular, we find formulas for the solution of the discrete time Euler top in the form which makes the comparison with the continuous time case immediate. Moreover, the Veselov–Moser system turns out to be by no means the only reasonable discretization of the Euler top. We introduce a whole family of discretizations sharing integrals of motion with the continuous time system, and characterize those of them which share also the underlying invariant Poisson structure. In this context the Veselov–Moser system becomes just one particular case, and not the most simple one.

It has to be mentioned that, unlike Refs. 6 and 8, our procedure for obtaining explicit solutions is rather pedestrian; it does not use such advanced tools as spectral theory of difference operators[6] or Baker–Akhiezer functions.[8] Actually, our construction is based on the addition formulas for elliptic functions, and could have been invented by Jacobi. However, as a heuristic tool for finding it we used the Lax representation of the Euler top in su(2) with a spectral parameter on an elliptic curve, which was certainly unknown in the times of Euler and Jacobi.

[a]Electronic mail: bobenko@math.tu-berlin.de
[b]Electronic mail: lorbeer@math.tu-berlin.de
[c]Electronic mail: suris@sfb288.math.tu-berlin.de

 © 1998 American Institute of Physics

II. THE EULER TOP

The famous Euler's equations describing the motion of a rigid body with the fixed center of mass (Euler top) read[10]

$$\begin{aligned}\dot{M}_1&=\left(\frac{1}{B}-\frac{1}{C}\right)M_2M_3,\\ \dot{M}_2&=\left(\frac{1}{C}-\frac{1}{A}\right)M_3M_1,\\ \dot{M}_3&=\left(\frac{1}{A}-\frac{1}{B}\right)M_1M_2.\end{aligned}\tag{1}$$

Here $\mathbf{M}=(M_1,M_2,M_3)^T$ is the kinetic momentum vector in the coordinate system attached firmly to the body; the axes of this system coincide with the principal axes of inertia, and the numbers $A,B,C>0$ are the corresponding moments of inertia.

The vector form of the equations (1) is

$$\dot{\mathbf{M}}=\mathbf{M}\times\boldsymbol{\Omega}(\mathbf{M}),\tag{2}$$

where the vector of the angular velocity is introduced:

$$\boldsymbol{\Omega}(\mathbf{M})=\left(\frac{M_1}{A},\frac{M_2}{B},\frac{M_3}{C}\right)^T.\tag{3}$$

The equations (1) are Hamiltonian with respect to the following Poisson bracket on $\mathbb{R}^3(\mathbf{M})$:

$$\{M_1,M_2\}=M_3,\quad \{M_2,M_3\}=M_1,\quad \{M_3,M_1\}=M_2,\tag{4}$$

with the Hamilton function $H(\mathbf{M})=E(\mathbf{M})/2$, where

$$E(\mathbf{M})=\frac{M_1^2}{A}+\frac{M_2^2}{B}+\frac{M_3^2}{C}.\tag{5}$$

The bracket (4) is degenerate and has one Casimir function,

$$M^2(\mathbf{M})=M_1^2+M_2^2+M_3^2.\tag{6}$$

Generic symplectic leaves of the bracket (4) are two-dimensional spheres $M^2(\mathbf{M})=\text{const}$, hence each system Hamiltonian with respect to this bracket is integrable in the Liouville–Arnold sense; in particular, the Euler top is integrable.

An explicit solution to the equations (1) can be given in terms of the Jacobi elliptic functions.[10] Suppose for the sake of definiteness that

$$A>B>C>0,\tag{7}$$

then

$$\frac{1}{A}\leqslant\frac{E}{M^2}\leqslant\frac{1}{C}.$$

The formulas for the solution look different depending on whether this quantity is greater or smaller than $1/B$:

$$M_1(t)=a\begin{Bmatrix}\operatorname{cn}\nu(t-t_0)\\ \operatorname{dn}\nu(t-t_0)\end{Bmatrix},\quad M_2(t)=b\operatorname{sn}\nu(t-t_0),\quad M_3(t)=c\begin{Bmatrix}\operatorname{dn}\nu(t-t_0)\\ \operatorname{cn}\nu(t-t_0)\end{Bmatrix}.\tag{8}$$

Here and below in similar situations the upper expressions in curly brackets refer to the case $1/B<E/M^2\leqslant 1/C$, while the lower ones refer to the case $1/A\leqslant E/M^2<1/B$. The module k of the elliptic functions in (8) is given by

$$k^2=\left\{\begin{array}{l}\dfrac{(A-B)(M^2-CE)}{(B-C)(AE-M^2)}\\[2ex]\dfrac{(B-C)(AE-M^2)}{(A-B)(M^2-CE)}\end{array}\right\},\tag{9}$$

so that we always have $0<k^2<1$. The coefficients a,b,c are defined by

$$a^2=A\frac{M^2-CE}{A-C},\quad b^2=\left\{\begin{array}{l}B\dfrac{M^2-CE}{B-C}\\[2ex]B\dfrac{AE-M^2}{A-B}\end{array}\right\},\quad c^2=C\frac{AE-M^2}{A-C}.\tag{10}$$

Finally, the frequency ν is given by

$$\nu^2=\left\{\begin{array}{l}\dfrac{(B-C)(AE-M^2)}{ABC}\\[2ex]\dfrac{(A-B)(M^2-CE)}{ABC}\end{array}\right\}.\tag{11}$$

In the case $E/M^2=1/B$ the elliptic functions degenerate to the hyperbolic ones, and we have

$$M_1(t)=\frac{a}{\cosh\nu(t-t_0)},\quad M_2(t)=b\tanh\nu(t-t_0),\quad M_3(t)=\frac{c}{\cosh\nu(t-t_0)},$$

where

$$a^2=\frac{A(B-C)}{B(A-C)}M^2,\quad b^2=M^2,\quad c^2=\frac{C(A-B)}{B(A-C)}M^2,$$

and

$$\nu^2=\frac{(A-B)(B-C)}{AB^2C}M^2.$$

In all three cases the numbers a,b,c,ν satisfy the condition $abc\nu>0$. In what follows, we shall denote by $\nu(M^2,E)$ the positive square root of the function given in (11), and hence we must have $abc>0$.

III. DISCRETIZATIONS

Our aim here is to find *integrable discretizations* of the Euler top.

Definition 1: An integrable discretization of the Euler top (2) is a one-parameter family of diffeomorphisms $F\colon \mathbb{R}^3\times[0,\epsilon)\mapsto\mathbb{R}^3$,

$$\hat{\mathbf{M}}=F(\mathbf{M},h),\tag{12}$$

such that each one of them $F(\cdot,h)\colon \mathbb{R}^3\mapsto\mathbb{R}^3$ *is Poisson with respect to the bracket* (4), *has two integrals of motion* M^2 *and* E, *i.e.,*

$$M^2(F(\mathbf{M},h))=M^2(\mathbf{M}),\quad E(F(\mathbf{M},h))=E(\mathbf{M}),\tag{13}$$

and the following asymptotics hold:

$$F(\mathbf{M},h)=\mathbf{M}+h\mathbf{M}\times\boldsymbol{\Omega}(\mathbf{M})+o(h),\quad h\to 0. \tag{14}$$

Actually, our maps always will be defined by *implicit* equations of motion:

$$\hat{\mathbf{M}}-\mathbf{M}=hf(\mathbf{M},\hat{\mathbf{M}},h). \tag{15}$$

Proposition 1: Let $f\colon \mathbb{R}^3\times\mathbb{R}^3\times[0,\epsilon)\mapsto\mathbb{R}^3$ *be a* C^1*-function in each argument such that*

$$f(\mathbf{M},\mathbf{M},0)=\mathbf{M}\times\boldsymbol{\Omega}(\mathbf{M}).$$

Then for h small enough the equation (15) defines a local diffeomorphism (12) satisfying (14).

Proof: This follows immediately from the implicit function theorem for all h satisfying $\det(I-h\partial f/\partial\hat{\mathbf{M}})\neq 0$, i.e., for all h small enough. ■

When speaking about discretizations, it is natural to think about $\mathbf{M}$ in (12) as about sequences $\mathbf{M}_m\colon\mathbb{Z}\mapsto\mathbb{R}^3$ approximating solutions $\mathbf{M}(t)\colon\mathbb{R}\mapsto\mathbb{R}^3$ of the Euler top (2) in the sense that $\mathbf{M}_m\approx\mathbf{M}(mh)$. In this context the formula (12) takes the form of the difference equation,

$$\mathbf{M}_{m+1}=F(\mathbf{M}_m,h),$$

and, similarly, the formula (15) has to be thought of as

$$\mathbf{M}_{m+1}-\mathbf{M}_m=hf(\mathbf{M}_m,\mathbf{M}_{m+1},h).$$

The approximation property $\mathbf{M}_m=\mathbf{M}(mh)+O(h)$ holds on finite time intervals and is assured by (14) (this is a standard fact from the numerical analysis).

IV. ELLIPTIC COORDINATES AND POISSON DISCRETIZATIONS

Now we would like to find a manageable criterium for a map (12) to be Poisson with respect to the bracket (4). The corresponding statement becomes rather transparent in a new coordinate system in $\mathbb{R}^3(\mathbf{M})$. The corresponding change of variables,

$$\Phi\colon(M^2,E,\varphi)\mapsto(M_1,M_2,M_3),$$

is suggested by the formulas (8) for a solution of the Euler top. In the above formula $(M_1,M_2,M_3)\in\mathbb{R}^3$, and $(M^2,E,\varphi)\in\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{T}$, where $\mathbb{T}=\mathbb{R}/(4K\mathbb{Z})$, and $K=K(k^2)$ is the full elliptic integral of the first kind corresponding to the value of k^2 given in (9). The formulas for Φ are different in two regions of $\mathbb{R}^3$ separated by the two planes:

$$\frac{E(\mathbf{M})}{M^2(\mathbf{M})}=\frac{1}{B}\quad\Leftrightarrow\quad\left(\frac{1}{B}-\frac{1}{A}\right)M_1^2=\left(\frac{1}{C}-\frac{1}{B}\right)M_3^2. \tag{16}$$

Each one of the sets $\{E(\mathbf{M})/M^2(\mathbf{M})>1/B\}$ and $\{E(\mathbf{M})/M^2(\mathbf{M})<1/B\}$ consists of two sectors, and each one of these sectors is bounded by two half-planes.

The *elliptic coordinates* in $\mathbb{R}^3(\mathbf{M})$ are introduced by

$$M_1=a\begin{Bmatrix}\operatorname{cn}\varphi\\ \operatorname{dn}\varphi\end{Bmatrix},\quad M_2=b\operatorname{sn}\varphi,\quad M_3=c\begin{Bmatrix}\operatorname{dn}\varphi\\ \operatorname{cn}\varphi\end{Bmatrix}, \tag{17}$$

with a,b,c given by (10), and the modulus k^2 of the elliptic functions given by (9). On the subset $\{E/M^2>1/B\}$ the sign of c coincides with the sign of M_3 (which is constant in each one of two sectors), and the signs of a,b satisfy the condition $\operatorname{sign}(ab)=\operatorname{sign}(c)$. Similarly, on the subset $\{E(\mathbf{M})/M^2(\mathbf{M})<1/B\}$ the sign of a coincides with the sign of M_1, while the signs of b,c satisfy $\operatorname{sign}(bc)=\operatorname{sign}(a)$. Finally, on the four half-planes described by (16) the elliptic coordinates are defined by continuity according to

FIG. 1. Different energy levels for constant M^2.

$$M_1=\frac{a}{\cosh\varphi},\quad M_2=b\tanh\varphi,\quad M_3=\frac{c}{\cosh\varphi},$$

the signs of a,c being the same as the signs of M_1,M_3, respectively, and $\operatorname{sign}(b)=\operatorname{sign}(ac)$.

Proposition 2: The formulas (17) define a valid change of variables (a local diffeomorphism) near each point of $\mathbb{R}^3$ *(see Fig. 1).*

Proof: It can be verified by a direct calculation that in the region $\{E(\mathbf{M})/M^2(\mathbf{M})\neq 1/B\}$ we have

$$\frac{\partial(M_1,M_2,M_3)}{\partial(M^2,E,\varphi)}=-\frac{1}{4\nu(M^2,E)}\operatorname{dn}^2\varphi\neq 0,$$

and that the partial derivatives of which this Jacobian is composed allow a continuation to the boundary $\{E(\mathbf{M})/M^2(\mathbf{M})=1/B\}$. ■

Now it is obvious that an arbitrary map (12) having two integrals of motion $M^2(\mathbf{M})$ and $E(\mathbf{M})$, can be cast, in the elliptic coordinates (M^2,E,φ), in the form

$$\hat{M}^2=M^2,\quad \hat{E}=E,\quad \hat{\varphi}=\hat{\varphi}(M^2,E,\varphi). \tag{18}$$

Proposition 3: The map (12) having two integrals of motion $M^2(\mathbf{M})$ *and* $E(\mathbf{M})$ *is Poisson with respect to the bracket (4) if in the elliptic coordinates* (M^2,E,φ) *it takes the form*

$$\hat{M}^2=M^2,\quad \hat{E}=E,\quad \hat{\varphi}=\varphi+g(M^2,E), \tag{19}$$

with the function g not depending on φ.

Proof: To prove this statement, we have to calculate the Poisson bracket (4) in the coordinates (M^2,E,φ). The corresponding formulas read:

$$\{M^2,E\}=\{M^2,\varphi\}=0,\quad \{E,\varphi\}=-2\nu(M^2,E). \tag{20}$$

Indeed, the function M^2 is a Casimir function, hence it Poisson commutes with both E and φ. To calculate the bracket $\{E,\varphi\}$, we substitute (away from the boundary $E/M^2=1/B$) the expressions (17) into an arbitrary one of the formulas (4), and after straightforward calculations arrive at the expression given above.

Actually, the concrete expression for $\{E,\varphi\}$ is not essential for the proof of our proposition. The only important thing is that this Poisson bracket does not depend on φ. Indeed, we have

$$\{\hat{E},\hat{\varphi}\}=\frac{\partial\hat{\varphi}}{\partial\varphi}\{E,\varphi\}.$$

Since the Poisson bracket $\{E,\varphi\}$ depends only on M^2,E which are integrals of motion, we see that the necessary and sufficient condition for our map to be Poisson reads as $\partial\hat{\varphi}/\partial\varphi=1$, which is equivalent to the last equation in (19). ■

V. SPECIAL DISCRETIZATIONS

We derive now a family of discretizations of the Euler top. Our derivation is based on a Lax representation with a spectral parameter for the Euler top. We prefer to work with a su(2) Lax representation, since our experience in discretizing various geometric structures (see, for example, Ref. 11) convinced us that this procedure may be performed most straightforwardly when applied to su(2) Lax formulations.

To find a suitable Lax representation for the Euler top, we use a stationary version of the Lax representation of the chiral field model due to Cherednik.[12] Set

$$M(u)=\frac{1}{2i}\sum_{k=1}^{3} M_k w_k(u)\sigma_k, \tag{21}$$

where $\sigma_k\,(k=1,2,3)$ are the Pauli matrices, and $w_k(u)$ are the following elliptic functions:

$$w_1(u)=\rho\frac{1}{\mathrm{sn}(u,\kappa)},\quad w_2(u)=\rho\frac{\mathrm{dn}(u,\kappa)}{\mathrm{sn}(u,\kappa)},\quad w_3(u)=\rho\frac{\mathrm{cn}(u,\kappa)}{\mathrm{sn}(u,\kappa)}. \tag{22}$$

Here the parameter ρ and the module κ of the elliptic functions are defined by

$$\rho^2=\frac{A-C}{AC},\quad \kappa^2=\frac{C(A-B)}{B(A-C)}. \tag{23}$$

Further, for a vector $\mathbf{V}=(V_1,V_2,V_3)^{\mathrm{T}}\in\mathbb{R}^3$ set

$$V(u)=\frac{1}{2i}\sum_{k=1}^{3} V_k w_k(u-u_0)\sigma_k, \tag{24}$$

where the point u_0 is chosen so that

$$w_1(u_0)=A^{-1/2},\quad w_2(u_0)=B^{-1/2},\quad w_3(u_0)=C^{-1/2}. \tag{25}$$

Consider now the Lax equation,

$$\dot{M}(u)=[M(u),V(u)]. \tag{26}$$

With the help of identities

$$w_j(u)w_k(u-u_0)=w_j(u_0)w_l(u-u_0)-w_k(u_0)w_l(u), \tag{27}$$

where (j,k,l) is a permutation of (1,2,3), one sees that the matrix equation (26) is equivalent to the set of the following two equations:

$$\dot{\mathbf{M}}=\mathbf{M}\times\boldsymbol{\Omega}^{1/2}(\mathbf{V}), \tag{28}$$

$$0=\mathbf{V}\times\boldsymbol{\Omega}^{1/2}(\mathbf{M}). \tag{29}$$

Here the second equation is equivalent to $\mathbf{V}=\gamma\,\boldsymbol{\Omega}^{1/2}(\mathbf{M})$ with some $\gamma\in\mathbb{R}$, which, being substituted in the first equation, results in

$$\dot{\mathbf{M}}=\gamma\,\mathbf{M}\times\boldsymbol{\Omega}(\mathbf{M}). \tag{30}$$

Obviously, for a nonvanishing function $\gamma=\gamma(t)$ the latter equation is nothing but a time reparametrization of the Euler top (2). In this sense (26) is a Lax representation of the Euler top.

Discretizing the above construction in time, we introduce the matrix

$$\mathcal{V}(u)=I+\frac{h}{2i}\sum_{k=1}^{3} V_k w_k(u-u_0)\sigma_k, \tag{31}$$

and consider instead of (26) the discrete time Lax equation

$$\hat{M}(u)=\mathcal{V}^{-1}(u)M(u)\mathcal{V}(u). \tag{32}$$

(this equation may be interpreted as a stationary version of the lattice chiral field model by Ref. 13). Representing the latter equation as $\mathcal{V}(u)\hat{M}(u)=M(u)\mathcal{V}(u)$ and using the identities (27), we find that our matrix equation is equivalent to the set of the following three equations:

$$\hat{\mathbf{M}}-\mathbf{M}=\frac{h}{2}(\mathbf{M}+\hat{\mathbf{M}})\times\boldsymbol{\Omega}^{1/2}(\mathbf{V}), \tag{33}$$

$$0=\mathbf{V}\times\boldsymbol{\Omega}^{1/2}(\mathbf{M}+\hat{\mathbf{M}}), \tag{34}$$

and

$$\sum_{k=1}^{3}(\hat{M}_k-M_k)V_k w_k(u)w_k(u-u_0)=0. \tag{35}$$

Now (34) is equivalent to

$$\mathbf{V}=\tfrac{1}{2}\gamma\,\boldsymbol{\Omega}^{1/2}(\mathbf{M}+\hat{\mathbf{M}}), \tag{36}$$

with some $\gamma\in\mathbb{R}$, which, being substituted in (33), results in

$$\hat{\mathbf{M}}-\mathbf{M}=\tfrac{1}{4}h\gamma\,(\mathbf{M}+\hat{\mathbf{M}})\times\boldsymbol{\Omega}(\mathbf{M}+\hat{\mathbf{M}}). \tag{37}$$

It remains to notice that, plugging (36), (37) into (35), we bring the latter equation to the form

$$\sum_{k=1}^{3}w_k(u)w_k(u-u_0)w_k(u_0)(w_l^2(u_0)-w_j^2(u_0))=0,$$

which is automatically satisfied due to the identity

$$w_l(u)w_l(u-u_0)w_j(u_0)-w_j(u)w_j(u-u_0)w_l(u_0)=w_k(u_0)(w_l(u_0)-w_j(u_0)). \tag{38}$$

So, the matrix equation (32) is equivalent to (37) with some $\gamma\in\mathbb{R}$, the vector $\mathbf{V}$ being given by (36).

However, the equation (37) does not completely define the discretization due to arbitrariness of γ. In what follows we shall consider this equation with γ being a certain fixed function on $\mathbf{M}$, $\hat{\mathbf{M}}$. In other words, the subject of our further investigations will consist of discretizations governed by implicit equations of motion of the following special form:

$$\hat{\mathbf{M}}-\mathbf{M}=\tfrac{1}{4}h\gamma(\mathbf{M},\hat{\mathbf{M}},h)\,(\mathbf{M}+\hat{\mathbf{M}})\times\boldsymbol{\Omega}(\mathbf{M}+\hat{\mathbf{M}}). \tag{39}$$

We call them *special discretizations*. It is obvious that if

$$\gamma\colon \mathbb{R}^3\times\mathbb{R}^3\times[0,\epsilon)\mapsto\mathbb{R}_+$$

is a C^1–function of each argument satisfying

$$\gamma(\mathbf{M},\mathbf{M},0)=1, \tag{40}$$

then the equation (39) fulfills the conditions of Proposition 1, and therefore defines a map (12).

In components, special discretizations may be presented as

$$\hat{M}_1-M_1=\frac{1}{4}h\gamma(\mathbf{M},\hat{\mathbf{M}},h)\left(\frac{1}{B}-\frac{1}{C}\right)(M_2+\hat{M}_2)(M_3+\hat{M}_3), \tag{41}$$

$$\hat{M}_2 - M_2 = \frac{1}{4} h \gamma(\mathbf{M}, \hat{\mathbf{M}}, h) \left(\frac{1}{C} - \frac{1}{A} \right) (M_3 + \hat{M}_3)(M_1 + \hat{M}_1), \tag{42}$$

$$\hat{M}_3 - M_3 = \frac{1}{4} h \gamma(\mathbf{M}, \hat{\mathbf{M}}, h) \left(\frac{1}{A} - \frac{1}{B} \right) (M_1 + \hat{M}_1)(M_2 + \hat{M}_2). \tag{43}$$

VI. POISSON PROPERTY OF SPECIAL DISCRETIZATIONS [ ]

We investigate now the integrability properties of special discretizations (which are naturally expected due to the existence of a Lax representation with a spectral parameter).

Proposition 4: Maps defined by equations of motion (39) possess $M^2(\mathbf{M})$ *and* $E(\mathbf{M})$ *as integrals of motion.*

Proof: It follows from (39) that

$$\langle \hat{\mathbf{M}} - \mathbf{M}, \mathbf{M} + \hat{\mathbf{M}} \rangle = 0 \quad \text{and} \quad \langle \hat{\mathbf{M}} - \mathbf{M}, \mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}}) \rangle = 0, \tag{44}$$

or, equivalently,

$$\langle \hat{\mathbf{M}}, \hat{\mathbf{M}} \rangle = \langle \mathbf{M}, \mathbf{M} \rangle \quad \text{and} \quad \langle \hat{\mathbf{M}}, \mathbf{\Omega}(\hat{\mathbf{M}}) \rangle = \langle \mathbf{M}, \mathbf{\Omega}(\mathbf{M}) \rangle. \tag{45}$$

This proves our statement. ■

Actually, this statement can be almost inverted. Namely, consider an arbitrary discretization having $M^2(\mathbf{M})$ and $E(\mathbf{M})$ as integrals of motion. Then the pairs $(\mathbf{M}, \hat{\mathbf{M}})$ satisfy (45), which is equivalent to (44). If, in addition, $(\mathbf{M} + \hat{\mathbf{M}}) \times \mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}}) \neq 0$, then there exists a real number $\gamma = \gamma(\mathbf{M}, \hat{\mathbf{M}})$ such that (39) holds.

In what follows we shall also need the following technical and obvious statement.

Lemma 1: The set of fixed points of the maps (12) defined by (39) coincides with the set of the points $\mathbf{M}$ *for which* $(\mathbf{M} + \hat{\mathbf{M}}) \times \mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}}) = 0$*, and this coincides with the union of the coordinate axes,*

$$\{M_1 = M_2 = 0\} \cup \{M_2 = M_3 = 0\} \cup \{M_3 = M_1 = 0\}.$$

For all other points $\mathbf{M}$*, at least two of the three expressions* $M_1 + \hat{M}_1$*,* $M_2 + \hat{M}_2$*,* $M_3 + \hat{M}_3$ *do not vanish.*

Now we find conditions for a map defined by (39) to be Poisson. To this end notice that the pairs $(\mathbf{M}, \hat{\mathbf{M}}) = (\mathbf{M}, F(\mathbf{M}, h))$ belong to the subset of $\mathbb{R}^3(\mathbf{M}) \times \mathbb{R}^3(\hat{\mathbf{M}})$ singled out by the conditions $M^2(\mathbf{M}) = M^2(\hat{\mathbf{M}})$ and $E(\mathbf{M}) = E(\hat{\mathbf{M}})$. The elements of this subset may be parametrized by the quadruples $(M^2, E, \varphi, \hat{\varphi})$ according to

$$\mathbf{M} = \left(a \begin{Bmatrix} \operatorname{cn} \varphi \\ \operatorname{dn} \varphi \end{Bmatrix}, \; b \operatorname{sn} \varphi, \; c \begin{Bmatrix} \operatorname{dn} \varphi \\ \operatorname{cn} \varphi \end{Bmatrix} \right), \tag{46}$$

$$\hat{\mathbf{M}} = \left(a \begin{Bmatrix} \operatorname{cn} \hat{\varphi} \\ \operatorname{dn} \hat{\varphi} \end{Bmatrix}, \; b \operatorname{sn} \hat{\varphi}, \; c \begin{Bmatrix} \operatorname{dn} \hat{\varphi} \\ \operatorname{cn} \hat{\varphi} \end{Bmatrix} \right). \tag{47}$$

Equivalently, we can use the following coordinates: $(M^2, E, \overline{\varphi}, \Delta\varphi)$, where

$$\overline{\varphi} = \frac{\varphi + \hat{\varphi}}{2}, \quad \Delta\varphi = \frac{\hat{\varphi} - \varphi}{2}. \tag{48}$$

Let us denote from the above mentioned subset that

$$\gamma(\mathbf{M}, \hat{\mathbf{M}}, h) = \Gamma(M^2, E, \overline{\varphi}, \Delta\varphi, h). \tag{49}$$

Theorem 1: *The equations of motion of the special discretization (39) in the elliptic coordinates* (M^2,E,φ) *have the form (18), where the function* $\hat{\varphi}(M^2,E,\varphi,h)$ *is implicitly defined by the equation*

$$\frac{\Gamma(M^2,E,\overline{\varphi},\Delta\varphi,h)}{1-k^2\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi)}=\frac{2}{h\,\nu(M^2,E)}\frac{\mathrm{sn}(\Delta\varphi)}{\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)}. \tag{50}$$

The map defined by (39) is Poisson, iff the equation (50) may be solved for $\Delta\varphi$ *as*

$$\Delta\varphi=\delta(M^2,E,h)=\frac{h\,\nu(M^2,E)}{2}(1+O(h)), \tag{51}$$

with the function δ *not depending on* $\overline{\varphi}$.

Proof: Denoting

$$D=D(M^2,E,\overline{\varphi},\Delta\varphi)=1-k^2\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi), \tag{52}$$

we find from (46), (47) with the help of addition formulas for elliptic functions:

$$\hat{M}_1-M_1=-\frac{2a}{D}\begin{Bmatrix}\mathrm{sn}(\overline{\varphi})\mathrm{dn}(\overline{\varphi})\mathrm{sn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)\\ k^2\mathrm{sn}(\overline{\varphi})\mathrm{cn}(\overline{\varphi})\mathrm{sn}(\Delta\varphi)\mathrm{cn}(\Delta\varphi)\end{Bmatrix},\quad \hat{M}_1+M_1=\frac{2a}{D}\begin{Bmatrix}\mathrm{cn}(\overline{\varphi})\mathrm{cn}(\Delta\varphi)\\ \mathrm{dn}(\overline{\varphi})\mathrm{dn}(\Delta\varphi)\end{Bmatrix}, \tag{53}$$

$$\hat{M}_2-M_2=\frac{2b}{D}\mathrm{cn}(\overline{\varphi})\mathrm{dn}(\overline{\varphi})\mathrm{sn}(\Delta\varphi),\quad \hat{M}_2+M_2=\frac{2b}{D}\mathrm{sn}(\overline{\varphi})\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi), \tag{54}$$

$$\hat{M}_3-M_3=-\frac{2c}{D}\begin{Bmatrix}k^2\mathrm{sn}(\overline{\varphi})\mathrm{cn}(\overline{\varphi})\mathrm{sn}(\Delta\varphi)\mathrm{cn}(\Delta\varphi)\\ \mathrm{sn}(\overline{\varphi})\mathrm{dn}(\overline{\varphi})\mathrm{sn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)\end{Bmatrix},\quad \hat{M}_3+M_3=\frac{2c}{D}\begin{Bmatrix}\mathrm{dn}(\overline{\varphi})\mathrm{dn}(\Delta\varphi)\\ \mathrm{cn}(\overline{\varphi})\mathrm{cn}(\Delta\varphi)\end{Bmatrix}. \tag{55}$$

Plugging this into an arbitrary one of the equations of motion (41)–(43), we arrive after some cancellations at the equation (50). Notice that Lemma 1 assures that, away from the fixed points, these cancellations are legitimate in at least one of the equations of motion.

So, we have arrived at the equation (50) of the form

$$\Psi(M^2,E,\overline{\varphi},\Delta\varphi,h)=0, \tag{56}$$

which serves for determining the function $\hat{\varphi}(M^2,E,\varphi,h)$ for our map. The equation (56) may be rewritten as

$$\Psi\left(M^2,E,\frac{\varphi+\hat{\varphi}}{2},\frac{\hat{\varphi}-\varphi}{2},h\right)\overset{\text{def}}{=}\tilde{\Psi}(M^2,E,\varphi,\hat{\varphi},h)=0. \tag{57}$$

Proposition 3 gives a necessary and sufficient condition $\partial\hat{\varphi}/\partial\varphi=1$ for our map to be Poisson, and this is equivalent to the following condition:

$$\frac{\partial\tilde{\Psi}}{\partial\hat{\varphi}}+\frac{\partial\tilde{\Psi}}{\partial\varphi}=0\quad \text{on the solutions of (57).} \tag{58}$$

Since, obviously,

$$\frac{\partial\tilde{\Psi}}{\partial\hat{\varphi}}=\frac{1}{2}\left(\frac{\partial\Psi}{\partial\overline{\varphi}}+\frac{\partial\Psi}{\partial(\Delta\varphi)}\right),\qquad \frac{\partial\tilde{\Psi}}{\partial\varphi}=\frac{1}{2}\left(\frac{\partial\Psi}{\partial\overline{\varphi}}-\frac{\partial\Psi}{\partial(\Delta\varphi)}\right),$$

we find that the Poisson property is equivalent to the following condition:

$$\frac{\partial \Psi}{\partial \overline{\varphi}} = 0 \quad \text{on the solutions of (56).} \tag{59}$$

In turn, (59) assures that the solutions of (56) for $\Delta\varphi$ do not depend on $\overline{\varphi}$. ■

Corollary: In the conditions of the above theorem, solutions $\mathbf{M}_m = (M_{1,m}, M_{2,m}, M_{3,m})^{\mathrm{T}}$ to the difference equation,

$$\mathbf{M}_{m+1} - \mathbf{M}_m = \tfrac{1}{4} h\, \gamma(\mathbf{M}_m, \mathbf{M}_{m+1}, h)\, (\mathbf{M}_m + \mathbf{M}_{m+1}) \times \mathbf{\Omega}(\mathbf{M}_m + \mathbf{M}_{m+1}), \tag{60}$$

in the integrable (≡ Poisson) case are given by

$$M_{1,m} = a \begin{Bmatrix} \mathrm{cn}(2m\delta + \varphi_0) \\ \mathrm{dn}(2m\delta + \varphi_0) \end{Bmatrix}, \quad M_{2,m} = b\, \mathrm{sn}(2m\delta + \varphi_0), \quad M_{3,m} = c \begin{Bmatrix} \mathrm{dn}(2m\delta + \varphi_0) \\ \mathrm{cn}(2m\delta + \varphi_0) \end{Bmatrix}. \tag{61}$$

VII. FIRST EXAMPLES OF INTEGRABLE DISCRETIZATIONS

We use Theorem 1 to investigate the Poisson property of several discretizations. We start with the following negative result.

Proposition 5: The function $\gamma_0(\mathbf{M}, \hat{\mathbf{M}}, h) \equiv 1$ *defines a non-Poisson map.*

The simplest way to find Poisson (and therefore integrable) discretizations is to assure that the left-hand side of the equation (50) does not depend on $\overline{\varphi}$.

Proposition 6: The functions

$$\gamma_1(\mathbf{M}, \hat{\mathbf{M}}, h) = \frac{4M^2}{\langle \mathbf{M} + \hat{\mathbf{M}}, \mathbf{M} + \hat{\mathbf{M}} \rangle} \tag{62}$$

and

$$\gamma_2(\mathbf{M}, \hat{\mathbf{M}}, h) = \frac{4E}{\langle \mathbf{M} + \hat{\mathbf{M}}, \mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}}) \rangle} \tag{63}$$

define Poisson maps.

Proof: Using the second expressions in (53)–(55), it is not difficult to derive the following formulas:

$$\frac{1}{4}\langle \mathbf{M} + \hat{\mathbf{M}}, \mathbf{M} + \hat{\mathbf{M}} \rangle = \frac{a^2 + c^2 - b^2 \mathrm{sn}^2(\Delta\varphi)}{1 - k^2 \mathrm{sn}^2(\overline{\varphi}) \mathrm{sn}^2(\Delta\varphi)},$$

$$\frac{1}{4}\langle \mathbf{M} + \hat{\mathbf{M}}, \mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}}) \rangle = \frac{\dfrac{a^2}{A} + \dfrac{c^2}{C} - \dfrac{b^2}{B} \mathrm{sn}^2(\Delta\varphi)}{1 - k^2 \mathrm{sn}^2(\overline{\varphi}) \mathrm{sn}^2(\Delta\varphi)}.$$

Hence, the left-hand side of the equation (50) for $\gamma = \gamma_{1,2}$ does not depend on $\overline{\varphi}$. Taking into account that, according to (10),

$$a^2 + c^2 = M^2, \qquad \frac{a^2}{A} + \frac{c^2}{C} = E,$$

we find the following expressions for this left-hand side:

$$1 - \alpha_{1,2}\, \mathrm{sn}^2(\Delta\varphi),$$

where

$$\alpha_1(M^2,E)=\frac{b^2}{M^2},\quad \alpha_2(M^2,E)=\frac{b^2}{BE} \tag{64}$$

[see (10) for the expression of b^2 through M^2 and E]. Therefore, the equation (50) in these two cases is equivalent to

$$\frac{\mathrm{sn}(\Delta\varphi)}{\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)}=\frac{h\nu}{2}\frac{1}{1-\alpha_{1,2}\,\mathrm{sn}^2(\Delta\varphi)}. \tag{65}$$

Obviously, its solutions satisfy

$$\Delta\varphi=\delta_{1,2}(M^2,E,h)=\frac{h\nu(M^2,E)}{2}(1+O(h)).$$

VIII. VESELOV–MOSER DISCRETIZATION

In order to introduce the Veselov–Moser discretization, we need, first of all, the matrix notation for the Euler top equation (2). Using the well-known isomorphism between the Lie algebra $(\mathbb{R}^3,\times)$ with the Lie algebra $(\mathrm{so}(3),[\cdot,\cdot])$ of 3×3 skew-symmetric matrices with the usual commutator, we can rewrite the equations of motion (2) also in the matrix form

$$\dot{M}=[M,\Omega(M)], \tag{66}$$

where

$$M=\begin{pmatrix} 0 & M_3 & -M_2 \\ -M_3 & 0 & M_1 \\ M_2 & -M_1 & 0 \end{pmatrix} \tag{67}$$

and

$$\Omega(M)=\begin{pmatrix} 0 & M_3/C & -M_2/B \\ -M_3/C & 0 & M_1/A \\ M_2/B & -M_1/A & 0 \end{pmatrix}. \tag{68}$$

The relation between the matrices M and $\Omega=\Omega(M)$ may be expressed as follows:

$$M=J\Omega+\Omega J, \tag{69}$$

where the entries of the diagonal matrix,

$$J=\mathrm{diag}(J_1,J_2,J_3), \tag{70}$$

are defined by the relations

$$A=J_2+J_3,\quad B=J_3+J_1,\quad C=J_1+J_2. \tag{71}$$

A spectral parameter dependent Lax representation for (66) is

$$(M+\lambda J^2)^{\cdot}=[M+\lambda J^2,\Omega(M)+\lambda J]. \tag{72}$$

Now we can describe the Veselov–Moser construction. The differential equation (66) is replaced by the difference one,

$$\hat{M}=\omega^T M\omega, \tag{73}$$

where $\omega \in \mathrm{SO}(3)$ is an orthogonal matrix related to M by means of the following relation, coming to replace, or to approximate, (69):

$$hM = \omega J - J\omega^T. \tag{74}$$

It is easy to see that the previous two relations imply also

$$h\hat{M} = J\omega - \omega^T J. \tag{75}$$

Moser and Veselov demonstrated that, in general, the equation (74) does not determine the matrix $\omega \in \mathrm{SO}(3)$ uniquely. This nonuniqueness may be described as follows. Notice that (74), (75) are equivalent to the following factorizations of matrix polynomials:

$$I - h\lambda M - \lambda^2 J^2 = (\omega + \lambda J)(\omega^T - \lambda J)$$

and

$$I - h\lambda \hat{M} - \lambda^2 J^2 = (\omega^T - \lambda J)(\omega + \lambda J),$$

respectively. Denote by S the set of six roots of the equation $\det(I - h\lambda M - \lambda^2 J^2) = 0$. Then to each decomposition $S = S_+ \cap S_-$ into two disjoint sets of three roots satisfying $S_+ = \overline{S_+} = -S_-$ there corresponds a unique solution $\omega \in \mathrm{SO}(3)$ of (74) such that S_+ is the set of the roots of $\det(\omega + \lambda J) = 0$, while S_- is the set of the roots of $\det(\omega^T - \lambda J) = 0$.

However, in the continuous limit, when h is supposed to be small, there exists a *unique* decomposition, for which the roots from S_- are positive real numbers $O(h)$–close to J_1^{-1}, J_2^{-1}, J_3^{-1}. This is the only decomposition for which the corresponding ω has the following asymptotics:

$$\omega = I + h\Omega(M) + O(h^2). \tag{76}$$

Supposing h small enough, we consider from now on only this choice of ω, and demonstrate that the discretization of Veselov and Moser also belongs to the class of special discretizations described by the equation (39).

Theorem 2: *The Veselov–Moser discretization for h small enough may be represented in the form (39) with*

$$\gamma_{\mathrm{VM}}(\mathbf{M}, \hat{\mathbf{M}}, h) = \frac{2}{1 + \sqrt{1 - \frac{1}{4}h^2 ||\mathbf{\Omega}(\mathbf{M} + \hat{\mathbf{M}})||^2}}. \tag{77}$$

Proof: We shall need the following lemma, proof of which is put in the Appendix.

Lemma 2: For a matrix $\omega \in \mathrm{SO}(3)$, $O(h)$—close to I, there exists a unique matrix $W \in \mathrm{so}(3)$ such that $W = O(h)$, and

$$\omega = I + W + \frac{\gamma}{2} W^2,$$

where

$$\gamma = \frac{2}{1 + \sqrt{1 - |W|^2}}.$$

Here

$$|\mathrm{W}|^2 = W_1^2 + W_2^2 + W_3^2, \quad \text{for} \quad W = \begin{pmatrix} 0 & W_3 & -W_2 \\ -W_3 & 0 & W_1 \\ W_2 & -W_1 & 0 \end{pmatrix} \in \mathrm{so}(3).$$

With the help of this lemma we derive the following two equations:

$$\omega-\omega^T=2W, \quad \omega+\omega^T=2I+\gamma W^2. \tag{78}$$

Now from (74), (75), and the first equation in (78) we find

$$h(M+\hat{M})=(\omega-\omega^T)J+J(\omega-\omega^T)=2(WJ+JW); \tag{79}$$

hence

$$W=\tfrac{1}{2}h\Omega(M+\hat{M}). \tag{80}$$

Further, from (74), (75), and the second equation in (78) we find

$$h(\hat{M}-M)=J(\omega+\omega^T)-(\omega+\omega^T)J=\gamma[J,W^2]=\gamma[WJ+JW,W]=\tfrac{1}{4}\gamma h^2[M+\hat{M},\Omega(M+\hat{M})], \tag{81}$$

where

$$\gamma=\frac{2}{1+\sqrt{1-\dfrac{1}{4}h^2|\Omega(M+\hat{M})|^2}}.$$

Using the above-mentioned isomorphism between $(\mathbb{R}^3,\times)$ and (so(3),$[\cdot,\cdot]$), we see that the theorem is proved. ■

It remains to reproduce from our point of view the result of Moser and Veselov concerning the Poisson property of their discretization. To this end, we demonstrate that the equation (50) with Γ_{VM} corresponding to γ_{VM} allows a $\overline{\varphi}$–independent solution for $\Delta\varphi$.

Theorem 3: *For the Veselov–Moser discretization the equation (50) is equivalent to*

$$\mathrm{sn}(\Delta\varphi)\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)=\frac{h\nu}{2}(1+\alpha\,\mathrm{sn}^2(\Delta\varphi)-\beta\,\mathrm{sn}^4(\Delta\varphi)), \tag{82}$$

where

$$\alpha=\left\{\begin{array}{l}\dfrac{2ACE-(A+C-B)M^2}{(B-C)(AE-M^2)}\\[2ex] \dfrac{2ACE-(A+C-B)M^2}{(A-B)(M^2-CE)}\end{array}\right\}, \quad \beta=\left\{\begin{array}{l}\dfrac{AC \quad M^2-CE}{(B-C)^2AE-M^2}\\[2ex] \dfrac{AC \quad AE-M^2}{(A-B)^2M^2-CE}\end{array}\right\}. \tag{83}$$

The solution of (82) satisfies

$$\Delta\varphi=\delta_{\mathrm{VM}}(M^2,E,h)=\frac{h\nu(M^2,E)}{2}(1+O(h)). \tag{84}$$

Proof: For the Veselov–Moser discretization the left-hand side of the equation (50) can be calculated with the help of (77) and the second expressions in (53)–(55):

$$\frac{\Gamma(M^2,E,\overline{\varphi},\Delta\varphi,h)}{1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi)}=\frac{2}{1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi)+\sqrt{(1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi))^2-h^2G}}, \tag{85}$$

where

$$G=\frac{a^2}{A^2}\begin{Bmatrix}\mathrm{cn}^2(\overline{\varphi})\mathrm{cn}^2(\Delta\varphi)\\ \mathrm{dn}^2(\overline{\varphi})\mathrm{dn}^2(\Delta\varphi)\end{Bmatrix}+\frac{b^2}{B^2}\mathrm{sn}^2(\overline{\varphi})\mathrm{cn}^2(\Delta\varphi)\mathrm{dn}^2(\Delta\varphi)+\frac{c^2}{C^2}\begin{Bmatrix}\mathrm{dn}^2(\overline{\varphi})\mathrm{dn}^2(\Delta\varphi)\\ \mathrm{cn}^2(\overline{\varphi})\mathrm{cn}^2(\Delta\varphi)\end{Bmatrix}. \quad (86)$$

Hence the equation (50) in the present case is equivalent to

$$1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi)+\sqrt{(1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi))^2-h^2G}=h\nu\frac{\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)}{\mathrm{sn}(\Delta\varphi)}.$$

We leave the radical alone on the left-hand side and square the resulting equation to derive

$$-h^2G=h^2\nu^2\frac{\mathrm{cn}^2(\Delta\varphi)\mathrm{dn}^2(\Delta\varphi)}{\mathrm{sn}^2(\Delta\varphi)}-2h\nu\frac{\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)}{\mathrm{sn}(\Delta\varphi)}(1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi)). \quad (87)$$

Obviously, we have

$$G=G_0(M^2,E,\Delta\varphi)-G_1(M^2,E,\Delta\varphi)\,\mathrm{sn}^2(\overline{\varphi}), \quad (88)$$

where

$$G_0=\frac{a^2}{A^2}\begin{Bmatrix}\mathrm{cn}^2(\Delta\varphi)\\ \mathrm{dn}^2(\Delta\varphi)\end{Bmatrix}+\frac{c^2}{C^2}\begin{Bmatrix}\mathrm{dn}^2(\Delta\varphi)\\ \mathrm{cn}^2(\Delta\varphi)\end{Bmatrix}, \quad (89)$$

$$G_1=\frac{a^2}{A^2}\begin{Bmatrix}\mathrm{cn}^2(\Delta\varphi)\\ k^2\,\mathrm{dn}^2(\Delta\varphi)\end{Bmatrix}+\frac{c^2}{C^2}\begin{Bmatrix}k^2\,\mathrm{dn}^2(\Delta\varphi)\\ \mathrm{cn}^2(\Delta\varphi)\end{Bmatrix}-\frac{b^2}{B^2}\mathrm{cn}^2(\Delta\varphi)\mathrm{dn}^2(\Delta\varphi). \quad (90)$$

We shall prove that

$$G_1-k^2\,\mathrm{sn}^2(\Delta\varphi)G_0=k^2\nu^2\,\mathrm{cn}^2(\Delta\varphi)\mathrm{dn}^2(\Delta\varphi). \quad (91)$$

This formula allows one to derive from (88),

$$G=\frac{G_1}{k^2\,\mathrm{sn}^2(\Delta\varphi)}(1-k^2\,\mathrm{sn}^2(\overline{\varphi})\mathrm{sn}^2(\Delta\varphi))-\nu^2\frac{\mathrm{cn}^2(\Delta\varphi)\mathrm{dn}^2(\Delta\varphi)}{\mathrm{sn}^2(\Delta\varphi)}, \quad (92)$$

so that (87) is equivalent to

$$\frac{h^2G_1}{k^2\,\mathrm{sn}^2(\Delta\varphi)}=2h\nu\frac{\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)}{\mathrm{sn}(\Delta\varphi)},$$

or

$$\mathrm{sn}(\Delta\varphi)\mathrm{cn}(\Delta\varphi)\mathrm{dn}(\Delta\varphi)=\frac{h}{2\nu k^2}G_1.$$

This is equivalent to (82), as one easily calculates from the definition (90) that

$$G_1=k^2\nu^2(1+\alpha\,\mathrm{sn}^2(\Delta\varphi)-\beta\,\mathrm{sn}^4(\Delta\varphi))$$

with α and β as in (83). It remains to prove the relation (91). But it follows from the definitions (89), (90):

$$G_1 - k^2 \operatorname{sn}^2(\Delta\varphi) G_0 = \left\{ \begin{array}{l} \dfrac{a^2}{A^2} + \dfrac{k^2 c^2}{C^2} - \dfrac{b^2}{B^2} \\ \\ \dfrac{k^2 a^2}{A^2} + \dfrac{c^2}{C^2} - \dfrac{b^2}{B^2} \end{array} \right\} \operatorname{cn}^2(\Delta\varphi)\operatorname{dn}^2(\Delta\varphi),$$

and the expressions in the curly brackets here in both cases are equal to $k^2\nu^2$, due to (10), (9), (11). ■

IX. CONCLUSIONS

In the present work we have introduced a large family of discretizations of the Euler top sharing the integrals of motion with the continuous time system. We characterized those of them which are also Poisson with respect to the invariant Poisson bracket of the Euler top. For all these Poisson discretizations we found a solution in terms of elliptic functions which allows a direct comparison with the continuous time case. We demonstrated that the Veselov–Moser discretization also belongs to our family, and applied our methods to this particular example.

Let us mention some of the possible directions for further progress. First of all, our construction works on the level of reduced equations of motion of the rigid body [in the Lie algebra $\mathbb{R}^3 \approx \mathrm{so}(3) \approx \mathrm{su}(2)$]. It would be important to lift it to the level of the corresponding Lie group SO(3) [or SU(2)]. In a more mechanical language, we want to discretize the equations of motion in the rest frame, and not only in the frame attached firmly to the body. Notice that the Veselov–Moser construction has a variational (Lagrangian) origin, which makes it valid in the group. Another example of such a discretization admitting a variational formulation in the corresponding group is the recently found discrete time Lagrange top.[14]

As a second problem, our construction has to be generalized to higher dimensions (for the general Euler-Manakov case of the multidimensional rigid body). Notice that the Veselov–Moser construction goes through in higher dimensions.

Further, it would be important to include our construction into the general r-matrix framework, connected to factorizations in the loop groups; cf. Ref. 15.

APPENDIX: PROOF OF LEMMA 2

Let $\omega \in \mathrm{SO}(3)$, $\omega = \mathrm{I} + O(h)$. Set $V = (\omega - \mathrm{I})/(\omega + \mathrm{I})$, then $V \in \mathrm{so}(3)$, $V = O(h)$, and

$$\omega = \frac{I+V}{I-V} = \frac{2}{I-V} - I.$$

But for $V \in \mathrm{so}(3)$ we have: $V^3 = -|V|^2 V$, and by induction,

$$V^{2k+1} = (-1)^k |V|^{2k} V, \quad V^{2k+2} = (-1)^k |V|^{2k} V^2, \quad k \geqslant 1,$$

so that

$$\frac{I}{I-V} = I + \frac{1}{1+|V|^2} V + \frac{1}{1+|V|^2} V^2,$$

hence

$$\omega = I + \frac{2}{1+|V|^2} V + \frac{2}{1+|V|^2} V^2.$$

Now setting

$$W = \frac{2}{1+|V|^2} V,$$

so that $W \in \mathrm{so}(3)$, $W = O(h)$, we find

$$\omega = I + W + \frac{\gamma}{2} W^2, \quad \text{where} \quad \gamma = 1 + |V|^2.$$

We have, obviously,

$$|W|^2 = \frac{4}{(1+|V|^2)^2} |V|^2 \ \Rightarrow |V|^2 = \frac{2 - |W|^2 - 2\sqrt{1-|W|^2}}{|W|^2},$$

and

$$\gamma = 1 + |V|^2 = \frac{2 - 2\sqrt{1-|W|^2}}{|W|^2} = \frac{2}{1+\sqrt{1-|W|^2}}.$$

This proves the lemma. ■

[1] M. Ablowitz and J. Ladik, ''A nonlinear difference scheme and inverse scattering,'' Stud. Appl. Math. **55**, 213–229 (1976); ''On solution of a class of nonlinear partial difference equations,'' *ibid.* **57**, 1–12 (1977).
[2] R. Hirota, ''Nonlinear partial difference equations. I–V,'' J. Phys. Soc. Jpn. **43**, 1423–1433, 2074–2078, 2079–2086 (1977); **45**, 321–32 (1978); **46**, 312–319 (1978).
[3] F. Date, M. Jimbo, and T. Miwa, ''Method for generating discrete soliton equations. I–IV,'' J. Phys. Soc. Jpn. **51**, 4116–4124,4125–4131 (1982); **52**, 761–765, 766-771 (1983).
[4] G. Quispel, F. Nijhoff, H. Capel, and J. Van der Linden, ''Linear integral equations and nonlinear differential-difference equations,'' Physica A **125**, 344–380 (1984).
[5] F. W. Nijhoff, H. Capel, and G. Wiersma, ''Integrable lattice systems in two and three dimensions,'' Lect. Notes Phys. **239**, 263–302 (1984); ''Linearizing integral transform for the multicomponent lattice KP,'' Physica A **138**, 76–99 (1986).
[6] A. P. Veselov, ''Integrable systems with discrete time and difference operators,'' Funct. Anal. Appl. **22**, 1–13 (1988).
[7] Yu. B. Suris, ''Generalized Toda chains in discrete time,'' Alg. Anal. **2**, 141–157 (1990); ''Discrete-time generalized Toda lattices: complete integrability and relation with relativistic Toda lattices,'' Phys. Lett. A **145**, 113–119 (1990).
[8] J. Moser and A. P. Veselov, ''Discrete versions of some classical integrable systems and factorization of matrix polynomials,'' Commun. Math. Phys. **139**, 217–243 (1991).
[9] P. Deift, L.-Ch. Li, and C. Tomei, ''Loop groups, discrete versions of some classical integrable systems, and rank 2 extensions,'' Mem. Am. Math. Soc. **479** (1992).
[10] V. V. Golubev, *Lectures on the Integration of the Equations of Motion of a Rigid Body About a Fixed Point* (State Publishing, Moscow, 1953).
[11] A. I. Bobenko, U. Pinkall, ''Discretization of surfaces and integrable systems, in *Discrete Integrable Geometry and Physics*, edited by A. Bobenko and R. Seiler (Oxford University Press, Oxford, 1998).
[12] I. V. Cherednik, ''On the integrability of the equation of a two-dimensional asymmetric chiral $O(3)$-field and its quantum analogue,'' Yad. Fiz. **33**, 278–282 (1981); English trans. in Sov. J. Nucl. Phys. **33** (1981).
[13] F. W. Nijhoff and V. Papageorgiou, ''Lattice equations associated with the Landau–Lifshitz equations,'' Phys. Lett. A **141**, 2169–274 (1989).
[14] A. I. Bobenko, ''Discrete integrable systems and geometry,'' to appear in *Proceedings of the International Congress of Mathematical Physics, '97*.
[15] Yu. B. Suris, ''*R*-matrices and integrable discretizations,'' in *Discrete Integrable Geometry and Physics*, edited by A. Bobenko and R. Seiler (Oxford University Press, Oxford, 1998).

On harmonic maps into gauge groups

Qing Ding[a)]
Institute of Mathematics and Laboratory of Mathematics for Nonlinear Sciences, Fudan University, Shanghai 200433, People's Republic of China

(Received 2 December 1997; accepted for publication 26 June 1998)

We continue the work of the paper [J. Math. Phys. **37**, 4076 (1996)] on the existence of globally defined harmonic maps from the Minkowski plane $R^{1,1}$ into an infinite-dimensional Hilbert Lie group. We prove that the Cauchy problem for harmonic maps from $R^{1,1}$ into Hilbert loop groups $H_s(LG)$ can be solved globally for all remaining cases $\frac{1}{2}<s\leqslant\frac{3}{4}$ and we obtain similar results for harmonic maps from $R^{1,1}$ into certain Hilbert Lie gauge groups $H_s(S^n,G)$ $(n\geqslant 2)$. These answer the two questions remaining from the above paper. © *1998 American Institute of Physics.* [S0022-2488(98)02612-7]

I. INTRODUCTION

An infinite-dimensional Hilbert Lie group $\tilde{G}$ is a smooth group which is at the same time an infinite-dimensional smooth Riemann–Hilbert manifold. The tangent space to $\tilde{G}$ at the identity element is its Lie algebra. The theory of finite-dimensional Lie groups is a classical subject and has been well developed since the fundamental works by Killing (Ref. 1) and Cartan (Ref. 2) at the end of the last century. However, infinite-dimensional Lie groups are much more complicated and the theory is still under development. Until now only some particular infinite-dimensional Lie groups have been studied. A good and classical example is mapping space $\mathrm{Map}(M,G)$, i.e., the maps from a compact Riemannian manifold M into a compact semi-simple Lie group G. Such infinite dimensional groups are called sometimes—and we will follow this custom—gauge groups. The name comes from the fact that such gauge groups arise in the representation theory of gauge quantum fields, in which case M has the interpretation of physical space. G has the interpretation of the structure Lie group for the principal fiber bundle whose connections define the classical gauge fields. It should be mentioned that the gauge groups were also introduced by Gelfand and his collaborators (Ref. 3) into the mathematical literature from the point of view of noncommutative distributions. The study of such groups and the associated (infinite-dimensional) Lie algebras is a topic of great interest in itself and because of its numerous and fruitful connections with many domains of mathematics and its applications in theoretical physics.

The theory of gauge groups and loop groups has been actively developed recently (Refs. 4–6). Let $H_s(M,G)=\{\gamma:M\to G|\,\gamma \text{ is } H_s \text{ integrable}\}$ $(s>\frac{1}{2}\dim M)$ be a real Hilbert Lie gauge group. It is well known that $H_s(M,G)$ admits a Riemannian Hilbert metric structure (see, for example, Ref. 7). In particular, the groups $H_s(S^1,G)$ [we will denote them by $H_s(LG)$ following the notation used in Ref. 8] are loop groups, which becomes a very important tool in solving problems both in mathematics and physics. Following this point of view, among all the gauge groups the most interesting and important class is the class of gauge groups $H_s(S^n,G)$ from an n-dimensional Euclidean sphere. On the other hand, roughly speaking, the study of the gauge groups $\mathrm{Map}(M,G)$ mainly involves analysis on the source M. The best candidate of a higher-dimensional compact Riemannian manifold with well known harmonic theory is undoubtedly a Euclidean sphere S^n $(n\geqslant 2)$.

In this paper, we continue the work of Ref. 8 in studying harmonic maps globally defined on the Minkowski plane $R^{1,1}$ (i.e., the two-dimensional space–time) to the Hilbert Lie gauge (loop) groups $H_s(S^n,G)$. In this direction of study, Gu (Ref. 9) in 1980, Ginibre and Velo (Ref. 10) in 1982 proved that the solution of the Cauchy problem for harmonic maps from the two-dimensional space–time (i.e., the Minkowski plane) into a complete Riemannian manifold (or a symmetric space such as complex projective space or Grassmann manifold) exist globally, which implies that if a (two-dimensional) σ-model is regular at a certain time, then it is regular in the past and future.

[a)]Electronic mail: qdingk@online.sh.cn

 © 1998 American Institute of Physics

Their works stimute the deeper study of the Cauchy problem for the higher-dimensional σ-model or two-dimensional σ-model defined on the curved Lorentzian manifolds (Refs. 11–14). Recently, the author and Lu (Ref. 8) gave a geometric description of certain Hilbert loop groups and showed that there is a wealth of globally defined harmonic maps from $R^{1,1}$ into the Hilbert loop groups $H_s(LG)$ $(s>\frac{3}{4})$. However, two natural questions are left in Ref. 8. One is the following: do there exist globally defined harmonic maps from $R^{1,1}$ into the Hilbert loop groups $H_s(LG)$ when $\frac{1}{2}<s\leqslant\frac{3}{4}$ too? On the other hand, the geometric characterization of the loop groups explored in Ref. 8 cannot be generally guaranteed on a Riemannian–Hilbert manifold (see Remark 2 in Ref. 8). So another question is the following: does the analogous result still hold for the Hilbert Lie gauge groups $H_s(M,G)$ with $\dim M\geqslant 2$? Our purpose in this paper is to give an affirmative answer to the first question and a positive answer to the second question in the most important case of $M=S^n(n\geqslant 2)$. Our discussions will strongly depend on the important relations [see (6) and (7) below] of the structure constants of the gauge algebras $H_s(S^n,g)$ displayed in Sec. II and certain geometric characterizations (see Lemmas 1–2 below) of the Hilbert Lie gauge groups $H_s(S^n,G)$ revealed in Sec. III.

This paper is organized as follows. In Sec. II we present some basic materials about the Hilbert Lie gauge groups $H_s(M,G)$ and some well known facts concerning the harmonic theory on the Euclidean sphere S^n. Then we calculate the structure constants of the infinite-dimensional Lie algebra of the smooth gauge group $\mathrm{Map}(S^n,G)$. In Sec. III we prove that the geometric characterizations of the uniform positive and homogeneity for the loop groups (see Lemma 1 in Ref. 8) still hold for a certain Hilbert–Sobolev metric on $H_s(S^n,G)$. These can be regarded as natural generalizations of the geometric properties of loop groups. In Sec. IV we present the main results of this paper, i.e., the global existence of the solutions of the Cauchy problem for harmonic maps from $R^{1,1}$ into the Hilbert loop groups $H_s(LG)$ $(s>1/2)$ and the Hilbert Lie gauge groups $H_s(S^n,G)$ $(n\geqslant 2, s>(n+1)/2)$. In Sec. V we give some remarks and problems.

II. LIE GAUGE ALGEBRAS AND GROUPS

Although our main objects are gauge groups $\mathrm{Map}(S^n,G)$, where G is a semi-simple compact Lie group, we would like to give a brief summary of some relevant results of general gauge groups $\mathrm{Map}(M,G)$ (the details may be find, for example, in Ref. 7). Gauge groups $\mathrm{Map}(M,G)$ enjoy a richer structure—they can be endowed with a class of Sobolev–Hilbert metrics. The local model for $\mathrm{Map}(M,G)$ is a completion of the Lie algebra of smooth maps $\mathrm{Map}(M,g)$, where g is the (finite-dimensional) Lie algebra of G. Let $(\cdot,\cdot)_g$ be an inner product on g given by minus the Killing form. Then for any real number s, the Sobolev–Hilbert H_s metric on $\mathrm{Map}(M,g)$ is defined by

$$(X,Y)_s=\int_M((1+\Delta)^sX(\theta),Y(\theta))_g dx,\quad X,Y\in Lg, \tag{1}$$

where dx is the volume form on M and Δ is the Laplacian operator on M. The Hilbert completion of smooth maps in this inner product is denoted by $H_s(M,g)$. The elements in $H_s(M,g)$ are continuous for $s>\frac{1}{2}\dim M$. In this range there are corresponding Hilbert Lie groups $H_s(M,G)$ modeled on $H_s(M,g)$. Of course, throughout this paper, we regard $H_s(M,G)$ as the connected component part containing the identity element. The exponential map $\exp: g\to G$ induces a map $\mathrm{Exp}: H_s(M,g)\to H_s(M,G)$, which gives a local chart near the identity element in $H_s(M,G)$. Left translations by smooth $f\in H_s(M,G)$ provide a system of coordinate charts covering the entire manifold. The above coordinate charts, according to Ref. 8, are also called the exponential coordinate charts in this paper.

Next we turn to the theory of spherical harmonics (the details are referred to Refs. 15–16). Let P_k be the space of homogeneous polynomials of degree k on the $n+1$-dimensional Euclidean space R^{n+1}, and let H_k be the subspace of harmonic polynomials in P_k. The restrictions of the element of H_k to the unit sphere S^n in R^{n+1} are called spherical harmonics of degree k. The space of spherical harmonics of degree k will be denoted by $\tilde{H}_k$, which are the eigenfunctions of the Laplace Δ on S^n with eigenvalue $\lambda_k=k(n+k-1)$. It is well known that $P_k=H_k\oplus r^2H_{k-2}\oplus r^4H_{k-4}\oplus\cdots$, where $r=|x|=\sqrt{x_1^2+\cdots+x_{n+1}^2}$ and hence the restriction to the unit sphere S^n of

any element of P_k is a sum of spherical harmonics of degree at most k. Furthermore, $L^2(S^n)=\oplus_0^\infty \tilde{H}_k$ and $\dim \tilde{H}_k=\dim P_k-\dim P_{k-2}:=d_k=(2k+n-1)(n+k-2)!/(n+1-k)!k!$ when $k\geqslant 1$ and $\dim \tilde{H}_0=1$.

There is an inner product in $\tilde{H}_k$ given by

$$\langle f,g\rangle=\int_{S^n} fg\,d\sigma,\quad \forall f,g\in\tilde{H}_k,$$

where $d\sigma$ is proportional to the volume form of the standard volume form dx on S^n and normalized in such a way that $\int_{S^n} d\sigma=\dim \tilde{H}_k$. Let $\{f_1(x),\ldots,f_{d_k}(x)\}$ be an orthonormal basis of $\tilde{H}_k$ $(k\geqslant 0)$. Then we have (Ref. 16)

$$\sum_{i=1}^{d_k} f_i^2=1,\quad \text{for all}\quad x\in S^n. \tag{2}$$

From now on, we fixed a system of the eigenfunctions $\{\phi_{k,\tau}\}$ $(k\in Z_+=\{0,1,2,\ldots\},1\leqslant\tau\leqslant d_k)$ of the Laplace operator Δ on S^n as above, i.e., for any $k\geqslant 0$,

$$\phi_{k,\tau}=f_\tau,\quad \tau=1,\ldots,d_k, \tag{3}$$

where f_τ are the spherical harmonics that appeared in (2). Note that $\phi_{0,1}=1$ is the first eigenfunction of Δ with eigenvalue $\lambda=0$ and $\int_{S^n}\phi_{k,\tau}\phi_{l,\sigma}=\delta_{kl}\delta_{\tau\sigma}\,\mathrm{vol}(S^n)/d_k$.

Throughout the paper, we always let G be a semi-simple Lie matrix group, especially the unitary groups $U(N)$, with the Lie algebra g on which there is minus the Killing metric $(\cdot,\cdot)_g$. From the fundamental results of finite-dimensional Lie theory, one may see that the above restriction of G is not serious but it simplifies our calculations and statements. It is well known that the gauge algebra $\mathrm{Map}(S^n,g)$ of the gauge group $\mathrm{Map}(S^n,G)$ can be identified with $\{\xi: S^n\to g\}$ which takes the form

$$\xi(x)=\sum_{k\geqslant 0,1\leqslant\tau\leqslant d_k}\xi_{k,\tau}\phi_{k,\tau}, \tag{4}$$

where $\phi_{k\tau}$ are the eigenfunctions that appeared in (3) and $\xi_{k,\tau}$ are g-valued constant coefficients. When $G=U(N)$, the coefficients $\xi_{k,\tau}$ are skew-Hermitian complex $N\times N$ matrices.

Let $T^1,\ldots,T^N$ be a basis of g with structure constants C_{ab}^c, $1\leqslant a,b,c\leqslant N$. Then, obviously, $\{T^a\phi_{k,\tau}: 1\leqslant a\leqslant N, k\in Z_+,\ 1\leqslant\tau\leqslant d_k\}$ spans the whole space $\mathrm{Map}(S^n,g)$. Set $T^a_{k,\tau}=T^a\phi_{k,\tau}$, thus $\{T^a_{k,\tau}: 1\leqslant a\leqslant N, k\in Z_+, 1\leqslant\tau\leqslant d_k\}$ is a basis of $\mathrm{Map}(S^n,g)$. Now we come to calculate the structure constants $\{C^{(ak\tau)}_{(bl\sigma)(cm\mu)}\}$ $(1\leqslant a,b,c\leqslant N, k,l,m\in Z_+, 1\leqslant\tau\leqslant d_k, 1\leqslant\sigma\leqslant d_l, 1\leqslant\mu\leqslant d_m)$ of the Lie algebra $\mathrm{Map}(S^n,g)$ with respect to the basis $\{T^a_{k,\tau}\}_{1\leqslant a\leqslant N; k\in Z_+, 1\leqslant\tau\leqslant d_k}$.

In order to do this, we first expand

$$\phi_{k,\tau}\phi_{l,\sigma}=\sum_{m\geqslant 0,1\leqslant\mu\leqslant d_m} A^{m,\mu}_{k,\tau;l,\sigma}\phi_{m,\mu},$$

where

$$A^{m,\mu}_{k,\tau;l,\sigma}=\begin{cases}\displaystyle\int_{S^n}\phi_{k,\tau}\phi_{l,\sigma}\phi_{m,\mu}dx, & m\leqslant k+l;\\ 0, & \text{else,}\end{cases}$$

since $\phi_{k,\tau}\phi_{l,\sigma}\in P_{k+l}$. Next we have

$$\begin{aligned}[T^a_{k,\tau},T^b_{l,\sigma}]=[T^a\phi_{k,\tau},T^b\phi_{l,\sigma}]&=\sum_{1\leqslant c\leqslant N,0\leqslant m\leqslant k+l,1\leqslant\mu\leqslant d_m} C^c_{ab}T^cA^{m,\mu}_{k,\tau;l,\sigma}\phi_{m,\mu}\\ &=\sum_{1\leqslant c\leqslant N,0\leqslant m\leqslant k+l,1\leqslant\mu\leqslant d_m} C^c_{ab}A^{m,\mu}_{k,\tau;l,\sigma}T^c_{m,\mu}.\end{aligned} \tag{5}$$

Hence

$$C^{(cm\mu)}_{(ak\tau)(bl\sigma)}=\begin{cases} C^c_{ab}A^{m,\mu}_{k,\tau;l,\sigma}, & \text{if } m\leqslant k+l, k\leqslant l+m, m\leqslant m+k;\\ 0, & \text{else,}\end{cases} \tag{6}$$

since $A^{m,\mu}_{k,\tau;l,\sigma}$ satisfying $A^{m,\mu}_{k,\tau;l,\sigma}=A^{k,\tau}_{l,\sigma;m,\mu}=A^{l,\sigma}_{m,\mu;k,\tau}$. Note that, for any fixed k,l,m, we have

$$\sum_{\tau,\sigma,\mu}|A^{m,\mu}_{k,\tau;l,\sigma}|^2\leqslant\sum_{\tau,\sigma,\mu}\int_{S^n}\phi^2_{m,\mu}\int_{S^n}\phi^2_{k,\tau}\phi^2_{l,\sigma}=\mathrm{vol}^2(S^n),$$

from (2) and (3) and hence

$$\sum_{a,b,c,\tau,\sigma,\mu}|C^{(cm\mu)}_{(ak\tau)(bl\sigma)}|^2\leqslant\sum_{a,b,c}|C^c_{ab}|^2\sum_{\tau,\sigma,\mu}|A^{m,\mu}_{k,\tau;l,\sigma}|^2\leqslant\sum_{a,b,c}|C^c_{ab}|^2\,\mathrm{vol}^2(S^n):=C_0. \tag{7}$$

Finally, we would like to point out that Lie algebras $H_s(S^n,g)$ also have $\{T^a_{k,\tau}\}$ as their basis and have the same structure constants expressed in (6). The relations (6) and (7) are very important for us to explore some geometric properties of the Hilbert gauge groups $H_s(S^n,G)$ in the next section.

III. LOCAL GEOMETRY OF HILBERT LIE GAUGE GROUPS

For simplicity, we denote by $\tilde{G}$ a Hilbert Lie gauge group $H_s(S^n,G)$ $(s>n/2)$. We have already seen that Exp:$\tilde{g}\to\tilde{G}$ gives a local chart near the identity element of $\tilde{G}$. As doing in the finite-dimensional case, we can also use the left translations $DL_w:T_e\tilde{G}\to T_w\tilde{G}$ (here $L_wx=w\cdot x$ is a left translation of $\tilde{G}$) to obtain a left-invariant Riemannian Hilbert metric on $\tilde{G}$ from any given inner product $\langle\cdot,\cdot\rangle$ on $T_e\tilde{G}$.

The Sobolev–Hilbert metric defined by (1) on $H_s(S^n,g)$ may be expressed equivalently as follows:

$$\left\langle\sum_{k,\tau}\xi_{k,\tau}\phi_{k,\tau},\sum_{l,\sigma}\eta_{l,\sigma}\phi_{l,\sigma}\right\rangle=\sum_{k,\tau}[(1+\lambda_k)^s/d_k]\langle\xi_{k,\tau},\eta_{k,\tau}\rangle_g,$$

where $\xi=\Sigma_{k,\tau}\xi_{k,\tau}\phi_{k,\tau}$, $\eta=\Sigma_{k,\tau}\eta_{k,\tau}\phi_{k,\tau}\in H_s(S^n,g)$, $\langle\cdot,\cdot\rangle_g$ on the right hand side is a bi-invariant metric (e.g., minus the Killing metric) on g and $\lambda_k=k(k+n-1)$ the eigenvalues of Δ on S^n.

For a finite-dimensional Riemannian manifold, the Levi-Civita theorem states that there is a unique torsion-free connection compatible with a given Riemannian metric. This theorem persists for infinite-dimensional Riemannian Hilbert manifolds on which the tangent space is complete with respect to the given Riemannian Hilbert metric. Associated with a given complete Riemannian Hilbert manifold $\tilde{G}$, let $(x_1,\ldots,x_k,\ldots)$ be its countable coordinates in a chart U_p of a point p with p as the origin. Then

$$h_{ij}(x)=h_x\left(\frac{\partial}{\partial x_i},\frac{\partial}{\partial x_j}\right)$$

is the metric tensor over the chart which is a linear inversible continuous (=bounded) transformation from $T_x\tilde{G}\to T_x\tilde{G}$. The inverse of the metric tensor $(h_{ij}(x))$ is denoted by $(h^{ij}(x))$.

Let ∇ be the torsion-free connection compatible with the Riemannian Hilbert metric h and we define Γ^k_{ij}, the coefficients of the Riemannian Hilbert connection or Christoffel symbols, by the following formula

$$\nabla_{\partial/\partial x_i}\frac{\partial}{\partial x_j}=\sum_{k=1}^{\infty}\Gamma^k_{ij}\frac{\partial}{\partial x_k}.$$

Then

$$\Gamma_{ij}^{k}=\frac{1}{2}h^{kl}\left(\frac{\partial h_{il}}{\partial x_j}+\frac{\partial h_{lj}}{\partial x_i}-\frac{\partial h_{ij}}{\partial x_l}\right) \tag{8}$$

if the right hand side of (8) makes sense (i.e., the series are convergent).

Now we return to the Hilbert Lie gauge groups $H_s(S^n,G)$. Let $\{T^1,\ldots,T^N\}$ be an orthonormal basis of g. The vectors $T_{k,\tau}^a=T^a\phi_{k,\tau}$ $(1\leqslant a\leqslant N;k\in Z_+,1\leqslant\tau\leqslant d_k)$ form a basis of the gauge algebra $\tilde{g}$ with the orthogonality relation $\langle T_{k,\tau}^a,T_{l,\sigma}^b\rangle=(1+\lambda_k)^s/d_k\delta^{ab}\delta_{kl}\delta_{\tau\sigma}$. On a local exponential coordinate chart U_e of the identity of $H_s(S^n,G)$, without loss of generality assuming $B_e(1)\subset U_e$, we introduce the exponential real coordinates $\{x_{k,\tau}^a\}$ $(1\leqslant a\leqslant N;0\leqslant k\leqslant\infty,1\leqslant\tau\leqslant d_k)$ as follows:

$$\{x_{k,\tau}^a\}\in H_s(S^n,g)\to\mathrm{Exp}\left(\sum_{a,k,\tau}x_{k,\tau}^aT_{k,\tau}^a\right)\in U_e.$$

With the above exponential real coordinates, let us denote by $(h_{(ak\tau)(bl\sigma)}(x))$ the Riemannian–Hilbert H_s-metric tensor on the chart.

The following lemmas give a geometric characterization of the Hilbert Lie gauge groups $H_s(S^n,G)$ $(s>(n+1)/2)$ as well as that of the Hilbert loop groups $H_s(LG)(s>3/4)$ (Ref. 8), which reflect some essential geometric features of these infinite-dimensional gauge groups and will play a very important role in establishing the global existence results of harmonic maps from $R^{1,1}$ into them.

Lemma 1: Let $\tilde{G}=H_s(S^n,G)$ $(s>(n+1)/2)$. Then

(a) there are positive numbers A and α such that

$$\sum_{a,b;k,l,\tau,\sigma}h_{(ak\tau)(bl\sigma)}(x)\xi_{k,\tau}^a\xi_{l,\sigma}^b\geqslant A\sum_{a,k,\tau}[(1+\lambda_k)^s/d_k](\xi_{k,\tau}^a)^2, \tag{9}$$

for all $x\in B_e(\alpha)\subset U_e$, and $(\xi_{k,\tau}^a)\in\tilde{g}=H_s(S^n,g)$, *the Lie algebra of* $\tilde{G}$.

(b) For any given point $w\in\tilde{G}$, *in the neighborhood* $B_w(\alpha)=w\cdot B_e(\alpha)$ *of* w, (9) *is still valid.*

Proof: We only prove that (a) is true. (b) comes directly from (a) and the isometricity of left translations $L_w:B_e(\alpha)\to B_w(\alpha)$ given by $x\to w\cdot x$ on $\tilde{G}$.

In order to give an explicit expression of the metric tensor $(h_{(ak\tau)(bl\sigma)}(x))$ over U_e, we come to calculate the left-invariant vector fields $X_{k,\tau}^a=dL_x\,\mathrm{Re}\,T_{k,\tau}^a$ $(1\leqslant a\leqslant N;k\in Z_+,1\leqslant\tau\leqslant d_k)$ in the chart. Let $\{f_{k,\tau}^a(x,y)\}$ be the exponential coordinates of the point $\mathrm{Exp}(\Sigma x_{k,\tau}^aT_{k,\tau}^a)\cdot\mathrm{Exp}(\Sigma y_{k,\tau}^aT_{k\tau}^a)$ in U_e, where two points $\{x_{k,\tau}^a\},\{y_{k,\tau}^a\}\in B_e(1)\subset U_e$. Since the exponential map $\mathrm{Exp}:\tilde{g}\to\tilde{G}$ of the gauge group $\tilde{G}$ is induced from the exponential map $\exp:g\to G$ of the finite-dimensional compact group G, we set $Z=\Sigma(x_{k,\tau}^aT_{k,\tau}^a),W=\Sigma(y_{k,\tau}^aT_{k,\tau}^a)$ which are regarded as elements in g. Applying the Campbell–Baker–Hausdorff formula

$$\begin{aligned}&\mathrm{Exp}\left(\sum x_{k,\tau}^aT_{k,\tau}^a\right)\cdot\mathrm{Exp}\left(\sum y_{k,\tau}^aT_{k,\tau}^a\right)\\&\quad=\exp Z\cdot\exp W\\&\quad=\exp\left(Z+\int_0^1\psi((\exp\mathrm{ad}\,Z)(\exp t\,\mathrm{ad}\,W))(W)dt\right),\end{aligned}$$

where

$$\psi(z)=1+\sum_{k=1}^{\infty}\frac{(-1)^{k-1}(z-1)^k}{k(k+1)};$$

we have

$$f^a_{k,\tau}(x,y)=(Z+W)^a_{k,\tau}+\tfrac{1}{2}(\mathrm{ad}\,Z(W))^a_{k,\tau}+\tfrac{1}{12}((\mathrm{ad}\,Z)^2(W)+(\mathrm{ad}\,W)^2(Z))^a_{k,\tau}+\cdots$$
$$=x^a_{k,\tau}+y^a_{k,\tau}+\tfrac{1}{2}x^c_{m,\mu}w^b_{l,\sigma}C^{(ak\tau)}_{(cm\mu)(bl\sigma)}+\tfrac{1}{12}(x^e_{q,\iota}x^c_{m,\mu}y^b_{l,\sigma}+y^e_{q,\iota}y^c_{m,\mu}x^b_{l,\sigma})$$
$$\times C^{(dp\kappa)}_{(cm\mu)(bl\sigma)}C^{(ak\tau)}_{(dp\kappa)(eq\iota)}+\cdots, \tag{10}$$

where $C^{(cm\mu)}_{(ak\tau)(bl\sigma)}$ are the structure constants that appeared in (6).

Notice that $|\psi(\exp z)|\leqslant 1+\Sigma_{k=1}^{\infty}e^k|z|^k$, $|z|\leqslant 1$ and

$$\max_{a,b,c;k,l,m,\tau,\sigma,\mu}\{(C^{(ak\tau)}_{(bl\sigma)(cm\mu)})^2\}\leqslant C_0:=C_1^2,$$

where C_0 is the constant that appeared in (7); we have

$$|f^a_{k,\tau}|\leqslant\|x^a_{k,\tau}+y^a_{k,\tau}\|+eC_1(\|x\|+\|y\|)^2+\cdots+(eC_1)^k(\|x\|+\|y\|)^{k+1}+\cdots, \tag{11}$$

where $\|\cdot\|$ denote the norm of the H_s-metric ($s>n/2$), which implies that $f^a_{k,\tau}$ are actually uniformal convergence series in a neighborhood $B_e(\beta)$ for some fixed number $\beta<\frac{1}{2}\min\{1,1/eC_1\}$. We shall reformulate (10) with aid of the convenient Laudau symbol $O(x^k)$:

$$f^a_{k,\tau}(x,y)=x^a_{k,\tau}+y^a_{k,\tau}+\tfrac{1}{2}x^c_{m,\mu}y^b_{l,\sigma}C^{(ak\tau)}_{(cm\mu)(bl\sigma)}$$
$$+\tfrac{1}{12}x^e_{q,\iota}x^c_{m,\mu}y^b_{l,\sigma}C^{(dq\kappa)}_{(cm\mu)(bl\sigma)}C^{(ak\tau)}_{(dp\kappa)(eq\iota)}+O(x^3y), \tag{12}$$

where $O(x^k)$ means any continuous map in a neighborhood of the identity satisfying $|O(x^k)/||x||^k|\leqslant C<\infty$ as $x\to 0$.

Hence we can calculate left-invariant vectors from (12) as follows:

$$X^a_{k,\tau}(x)=\sum_{b;l,\sigma}\frac{\partial f^b_{l,\sigma}(x,y)}{\partial y^a_{k,\tau}}\bigg|_{y=0}\frac{\partial}{\partial x^b_{l,\sigma}}=\sum_{b;l,\sigma}A_{(ak\tau)(bl\sigma)}(x)\frac{\partial}{\partial x^b_{l,\sigma}},$$

where $A_{(ak\tau)(bl\sigma)}(x)=\delta^{ab}\delta_{kl}\delta_{\tau\sigma}+\frac{1}{2}C^{(bl\sigma)}_{(cm\mu)(ak\tau)}x^c_{m,\mu}+O(x^2)$. From the definition of the (left-invariant) H_s-metric on $H_s(S^n,G)$, we have

$$[(1+\lambda_k)^s/d_k]\delta^{ab}\delta_{kl}\delta_{\tau\sigma}=\langle X^a_{k,\tau}(0),X^b_{l,\sigma}(0)\rangle$$
$$=\langle X^a_{k,\tau}(x),X^b_{l,sigma}(x)\rangle$$
$$=A_{(ak\tau)(dp\kappa)}(x)A_{(bl\sigma)(eq\iota)}(x)h_{(dp\kappa)(eq\iota)}(x),$$

where $h_{(dp\kappa)(eq\iota)}(x)=h(\partial/\partial x^d_{p,\kappa},\partial/\partial x^e_{q,\iota})$. Since $\{X^a_{k,\tau}\}$ and $\{\partial/\partial x^a_{k,\tau}$ are two independent vector fields spanning $H_s(S^n,g)$, we find $(A_{(ak\tau)(bl\sigma)}(x))\in GL(H_s(S^n,g))$ for $x\in B_e(\beta)$. Let $(A^{(ak\tau)(bl\sigma)}(x))$ be the inverse of $(A_{(ak\tau)(bl\sigma)}(x))$. Obviously, $A^{(ak\tau)(bl\sigma)}(x)=\delta_{ab}\delta^{kl}\delta^{\tau\sigma}-\frac{1}{2}C^{(bl\sigma)}_{(cm\mu)(ak\tau)}x^c_{m\mu}+O(x^2)$ which is also analytic and uniform convergence in $B_e(\beta)$. From this we have

$$h_{(dp\kappa)(eq\iota)}(x)=A^{(dp\kappa)(ak\tau)}A^{(eq\iota)(bl\sigma)}[(1+\lambda_k)^s/d_k]\delta^{ab}\delta_{kl}\delta_{\tau\sigma}$$
$$=[(1+\lambda_p)^s/d_p]\delta^{de}\delta_{pq}\delta_{\kappa\iota}-\tfrac{1}{2}([(1+\lambda_p)^s/d_p]C^{(dp\kappa)}_{(cm\mu)(eq\iota)} \tag{13}$$
$$+[(1+\lambda_q)^s/d_q]C^{(eq\iota)}_{(cm\mu)(dp\kappa)})x^c_{m,\mu}(1+O(x)).$$

Thus $h_{(ak\tau)(bl\sigma)}(x)$ are also analytic and uniformly convergent in $B_e(\beta)$.

Now for any $\{\xi^a_{k,\tau}\}\in H_s(S^n,g)$,

$$\left|\sum_{a,b;k,l,\tau\sigma}h_{(ak\tau)(bl\sigma)}(x)\xi^a_{k,\tau}\xi^b_{l,\sigma}-\sum_{a;k,\tau}[(1+\lambda_k)^s/d_k](\xi^a_{k,\tau})^2\right|$$

$$= \left| \sum_{a,b;k,l,\tau,\sigma} (h_{(ak\tau)(bl\sigma)}(x) - h_{(ak\tau)(bl\sigma)}(0)) \xi^a_{k,\tau} \xi^b_{l\sigma} \right|$$

$$= \frac{1}{2} \Bigg| \sum \left(\frac{(1+\lambda_k)^s}{d_k} C^{(ak\tau)}_{(cm\mu)(bl\sigma)} + \frac{(1+\lambda_l)^s}{d_l} C^{(bl\sigma)}_{(cm\mu)(ak\tau)} \right) x^c_{m,\mu} (1 + O(x)) \xi^a_{k,\tau} \xi^b_{l,\sigma} \Bigg|. \tag{14}$$

The crucial point is to give an estimate of the term

$$Q := \left| \sum \frac{1}{2} ([(1+\lambda_k)^s/d_k] C^{(ak\tau)}_{(cm\mu)(bl\sigma)} + [(1+\lambda_l)^s/d_l] C^{(bl\sigma)}_{(cm\mu)(ak\tau)}) x^c_{m,\mu} \xi^a_{k,\tau} \xi^b_{l,\sigma} \right|$$

in case of $s > (n+1)/2$. In fact, we see that

$$Q \leqslant \sum_{a,b,c;k,l,m,\tau,\sigma,\mu} [(1+\lambda_k)^s/d_k] |C^{(ak\tau)}_{(cm\mu)(bl\sigma)} x^c_{m,\mu} \xi^a_{k,\tau} \xi^b_{l,\sigma}|$$

$$\leqslant C_2 \sum_{a,b,c;k,l,m,\tau,\sigma,\mu} (1+k)^{2s-n+1} |C^{(ak\tau)}_{(cm\mu)(bl\sigma)} x^c_{m,\mu} \xi^a_{k,\tau} \xi^b_{l,\sigma}|,$$

for some absolute constant C_2 since $\lambda_k = k(k+n-1) \sim k^2$ and $d_k \sim k^{n-1}$ as $k \to \infty$. Put $\alpha = (2s - n + 1)/2$; then

$$\sum_{a,b,c;k,l,m,\tau,\sigma,\mu} (1+k)^{2\alpha} |C^{(ak\tau)}_{(cm\mu)(bl\sigma)} x^c_{m,\mu} \xi^a_{k,\tau} \xi^b_{l,\sigma}|$$

$$\leqslant \sqrt{\sum_{a;k,\tau} \left[\sum_{b,c;l,m,\sigma,\mu} \frac{(1+k)^{2\alpha}}{(1+l)^{2\alpha}(1+m)^{2\alpha}} \left| C^{(ak\tau)}_{(cm\mu)(bl\sigma)} \right|^2 \right] (1+k)^{2\alpha} |\xi^a_{k,\tau}|^2}$$

$$\cdot \sqrt{\sum_{b;l,\sigma} (1+l)^{2\alpha} |\xi^b_{l,\sigma}|^2} \cdot \sqrt{\sum_{c;m,\mu} (1+m)^{2\alpha} |x^c_{m,\mu}|^2}.$$

Now we must have to indicate that $\Sigma_{b,c;l,m,\sigma,\mu} [(1+k)^{2\alpha}/(1+l)^{2\alpha}(1+m)^{2\alpha}] |C^{(ak\tau)}_{(cm\mu)(bl\sigma)}|^2$ are uniformly bounded for arbitrary k under the restriction $s > (n+1)/2$. Indeed, we have $\forall k$:

$$\sum_{b,c;l,m,\sigma,\mu} \frac{(1+k)^{2\alpha}}{(1+l)^{2\alpha}(1+m)^{2\alpha}} |C^{(ak\tau)}_{(cm\mu)(bl\sigma)}|^2$$

$$\leqslant \sum_{l+m \geqslant k, k+l \geqslant m, m+k \geqslant l} C_0 \frac{(1+k)^{2\alpha}}{(1+l)^{2\alpha}(1+m)^{2\alpha}} \quad [\text{using} \quad (6),\ (7)]$$

$$\leqslant C_0 \sum_{m \geqslant k/2; k \leqslant l+m, l \leqslant k+m, m \leqslant k+l} \frac{(1+k)^{2\alpha}}{(1+l)^{2\alpha}(1+m)^{2\alpha}}$$

$$+ C_0 \sum_{l \geqslant k/2; k \leqslant l+m, m \leqslant k+l, l \leqslant k+m} \frac{(1+k)^{2\alpha}}{(1+l)^{2\alpha}(1+m)^{2\alpha}}$$

$$\leqslant 2C_0 (1+k)^{2\alpha} \sum_{m=0}^{\infty} \frac{1}{(1+m)^{2\alpha}} \frac{2m}{(1+k/2)^{2\alpha}}$$

$$\leqslant C_3 \quad [\text{when} \quad s > (n+1)/2], \tag{15}$$

where C_3 is some absolute constant. Furthermore, note that there exists an absolute constant $C_4 > 1$ such that

$$C_4^{-1}\sqrt{\sum_{a;k,\tau}(1+k)^{2\alpha}|\xi^a_{k,\tau}|^2}\leqslant\|\xi\|:=\sqrt{\sum_{a;k,\tau}[(1+\lambda_k)^s/d_k]|\xi^a_{k,\tau}|^2}$$
$$\leqslant C_4\sqrt{\sum_{a;k,\tau}(1+k)^{2\alpha}|\xi^a_{k,\tau}|^2};$$

we arrive the estimate

$$Q\leqslant C_5\|x\|\|\xi\|^2,$$

for some absolute constant C_5 from the above estimates when $s>(n+1)/2$. This gives

$$\left|\sum_{a,b;k,l,\tau,\sigma}h_{(al\tau)(bl\sigma)}(x)\xi^a_{k,\tau}\xi^b_{l,\sigma}-\sum_{a;k,\tau}[(1+\lambda_k)^s/d_k](\xi^a_{k,\tau})^2\right|\leqslant(C_6+O(\|x\|))\|x\|\|\xi\|^2,$$

for some absolute constant C_6 from (14).

Hence there exist suitably small positive constants A,α $(0<\alpha\leqslant\beta)$ such that

$$\sum_{a,b;k,l,\tau,\sigma}h_{(ak\tau)(bl\sigma)}(x)\xi^a_{k,\tau}\xi^b_{l,\sigma}\geqslant A\|\xi\|^2,\quad\forall x\in B_e(\alpha).$$

The proof of Lemma 1 is completed.

Remark 1: The property of Lemma 1 (a) is called the uniform positive of the metric and the property of Lemma 1 (b) the homogeneity of the metric. As showing in Remark 2 in Ref. 8, the uniform positive and homogeneity of the metric can not be guaranteed on a Riemann–Hilbert manifold in general. This is one of the reasons why we restrict our attention to gauge groups.

Substituting (13) into (8) and by a direct calculation, we have

$$\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(x)=\frac{d_m}{2|1+\lambda_m|^s}\left(\frac{1}{2}B^{(ak\tau)}_{(bl\sigma)(cm\mu)}+\frac{1}{2}B^{(cm\mu)}_{(ak\tau)(bl\sigma)}-\frac{1}{2}B^{(bl\sigma)}_{(cm\mu)(ak\tau)}\right)(1+O(x)),\quad(16)$$

where $B^{(cm\mu)}_{(ak\tau)(bl\sigma)}=-[(1+\lambda_m)^s/d_m]C^{(cm\mu)}_{(ak\tau)(bl\sigma)}-[(1+\lambda_l)^s/d_l]C^{(bl\sigma)}_{(ak\tau)(cm\mu)}$. Using the same method used in proving Lemma 1 above, from the expression (16), we may have the following (the details are omitted)

Lemma 2: Under the same assumptions as in Lemma 1, we have

(i)

$$\|\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(x)\xi^a_{k\tau}\eta^b_{l,\sigma}\|^2:=\sum_{c;m,\mu}[(1+\lambda_m)^s/d_m]\left(\sum_{a,b;k,l,\tau,\sigma}\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(x)\xi^a_{k,\tau}\eta^b_{l,\sigma}\right)^2$$
$$\leqslant C_7\|\xi\|^2\|\eta\|^2,\qquad x\in B_e(\alpha),$$

for some constant $C_7>0$;

(ii)

$$\|(\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(x)-\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(y))\xi^a_{k,\tau}\eta^b_{l,\sigma}\|^2\leqslant C_8\|x-y\|^2\|\xi\|^2\|\eta\|^2,\qquad x,y\in B_e(\alpha),$$

for some positive constant C_8.

It seems that gauge groups $\mathrm{Map}(M,G)$ for higher-dimensional M can not be rewritten by a complexification with some reality condition in general. However, loop groups $\mathrm{Map}(S^1,G)$ does (see Refs. 5,8) and it is very convenient for calculations and statements. With the help of its complexification (with a reality condition), we can prove that Lemmas 1 and 2 are still valid for all the Hilbert loop groups $H_s(LG)$ $(s>1/2)$. Following the same notations in Ref. 8 (note that the symbol n denotes a index this time), we have Lemma $1'$.

Lemma $1'$: Let $\tilde{G}=H_s(LG)$ $(s>1/2)$. Then

(a) there are positive numbers A and α such that

$$\sum_{a,b,n,m} h_{(na)(mb)}(x)\lambda_n^a\lambda_m^b \geqslant A\sum_{a,n} (|n|^s\lambda_n^a)^2,$$

for all $x\in B_e(\alpha)\subset U_e$, and $(\lambda_n^a)\in\tilde{g}=H_s(Lg)$, the Lie algebra of $\tilde{G}$.

(b) For any given point $w\in\tilde{G}$, in the neighborhood $B_w(\alpha)=w\cdot B_e(\alpha)$ of w, the estimate in (a) above is still valid.

Proof: With the same methods used in the proof of Lemma 1 in Ref. 8, the key step is to prove that the estimate (11) in Ref. 8 is still valid for all the fixed $s>1/2$. We can arrive this estimate by the following two observations:

(α) An inequality $|l+n|\leqslant 2|l||n|$ for all nonzero integers l,m and hence

$$|l+n|^s\leqslant 2^s|l|^s|n|^s, \quad l,n\neq 0.$$

(β)

$$\begin{aligned}
I &\leqslant \sum_{a,b,c,n,m,l} |m|^{2s}|\alpha_{(lc)(na)}^{(mb)}x_l^c\lambda_n^a\lambda_m^b| \\
&\leqslant 3C_0\sum_{a,b,c,l,n} |l+n|^{2s}|x_l^c\lambda_n^a\lambda_{l+n}^b| \quad [\text{using} \quad (3),(4) \quad \text{in} \quad \text{Ref. 8}] \\
&\leqslant 3C_0 2^s\sum |l+n|^s|l|^s|n|^s|x_l^c\lambda_n^a\lambda_{l+n}^b| \\
&\leqslant 3\cdot 2^sC_0\sqrt{\sum_{a,n}|n|^{2s}|\lambda_n^a|^2}\sqrt{\sum_{c,l}|l|^{2s}|x_l^c|^2}\sqrt{\sum_{b,l,n}|l+n|^{2s}|\lambda_{l+n}^b|^2} \\
&\leqslant C_2\|x\|\|\lambda\|^2,
\end{aligned}$$

for some constant C_2 [this estimate holds similarly if we instead have $|\cdot|^{2s}$ by $(1+|\cdot|^2)^s$]. This establishes (11) for all $s>1/2$. The proof of Lemma $1'$ is completed.

Using the same argument (β) above with the aid of the inequality (α), we can also have Lemma $2'$.

Lemma $2'$: Under the same assumptions as in Lemma $1'$, we have

(i)

$$\|\Gamma_{(nk)(bm)}^{(cl)}(x)\lambda_n^a\mu_m^b\|^2\leqslant C_7\|\xi\|^2\|\eta\|^2, \qquad x\in B_e(\alpha),$$

for some constant $C_7>0$;

(ii)

$$\|(\Gamma_{(an)(bm)}^{(cl)}(x)-\Gamma_{(an)(bm)}^{(cl)}(y))\lambda_n^a\mu_m^b\|^2\leqslant C_8\|x-y\|^2\|\xi\|^2\|\eta\|^2, \qquad x,y\in B_e(\alpha),$$

for some positive constant C_8.

IV. HARMONIC MAPS FROM $R^{1,1}$ INTO GAUGE GROUPS

Let $R^{1,1}=\{(t,x):ds^2=dt^2-dx^2\}$ be the Minkowski plane (throughout this section x denotes the second coordinate of $R^{1,1}$). We represent a smooth map $u:R^{1,1}\to\tilde{G}=H_s(S^n,G)$ in an exponential coordinate chart as $u=\{u_{k,\tau}^a\}$ $(1\leqslant a\leqslant N;k\in Z_+,1\leqslant\tau\leqslant d_k)$. u is called a harmonic map if and only if u satisfies the following (hyperbolic) harmonic systems:

$$\frac{\partial^2u_{m,\mu}^c}{\partial t^2}-\frac{\partial^2u_{n,\mu}^c}{\partial x^2}+\sum \Gamma_{(ak\tau)(bl\sigma)}^{(cm\mu)}(u)\left(\frac{\partial u_{k,\tau}^a}{\partial t}\frac{\partial u_{l,\sigma}^b}{dt}-\frac{\partial u_{k,\tau}^a}{\partial x}\frac{\partial u_{l,\sigma}^b}{\partial x}\right)=0,\forall c;m,\mu. \tag{17}$$

In fact, (17) is the Lagrange–Euler equations of the (formal) energy $E(u)$,

$$E(u)=\int_{R^{1,1}} e(u)dt\,dx,$$

where $e(u)$ is the energy density:

$$e(u)=\left(\frac{\partial u^a_{k,\tau}}{\partial t}\frac{\partial u^b_{l,\sigma}}{dt}-\frac{\partial u^a_{k,\tau}}{\partial x}\frac{\partial u^b_{l,\sigma}}{\partial x}\right)h_{(ak\tau)(bl\sigma)}.$$

If $\tilde{G}$ is a finite-dimensional Lie group, then (18) is nothing but the equations of a two-dimensional σ-model or chiral. Since (18) is hyperbolic, we also propose an initial condition (Ref. 8) by

$$u^c_{m,\mu}(0,x)=(\phi_0)^c_{m,\mu}(x),\quad \frac{\partial u^c_{m,\mu}}{\partial t}(0,x)=(\phi_1)^c_{m,\mu}(x), \tag{18}$$

where $\phi_0=\{(\phi_0)^c_{m,\mu}(x)\}:R\to\tilde{G}$ is a smooth curve on $\tilde{G}$ and $\phi_1=\{(\phi_1)^c_{m,\mu}(x)\}$ is a smooth vector field along the curve ϕ_0 (i.e., $\|\phi_1(x)\|<\infty$).

Without a loss of generality, we assume that $\phi_0(0)=id\in\tilde{G}$, and introduce the characteristic coordinates (ξ,η) of $R^{1,1}$ (throughout this section ξ,η always denote the characteristic coordinates of $R^{1,1}$):

$$\xi=\frac{t+x}{2},\quad \eta=\frac{t-x}{2}.$$

Hence harmonic systems (18) become

$$\frac{\partial^2 u^c_{m,\mu}}{\partial\xi\partial\eta}+\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(u)\frac{\partial u^a_{k,\tau}}{\partial\xi}\frac{\partial u^b_{l,\sigma}}{\partial\eta}=0,\quad \forall c;m,\mu. \tag{19}$$

Now can completely follow the program displayed in Ref. 8 IV to establish the global existence of harmonic map systems (17) with the initial condition (18). Here we give a sketch as follows

First of all we come to state a local existence result for hyperbolic harmonic maps from $R^{1,1}$ into $\tilde{G}=H_s(S^n,G)[n\geqslant 2,s>(n+1)/2$ or $n=1,s>1/2]$. For a given large positive number $L>0$, let M be a number defined by

$$\sup_{|x|\leqslant L}\left\{\left[h_{(ak\tau)(bl\sigma)}(\phi_0)\frac{\partial u^a_{k,\tau}}{\partial\xi}\bigg|_{t=0}\frac{\partial u^b_{l,\sigma}}{\partial\xi}\bigg|_{t=0}\right]^{1/2},\left[h_{(ak\tau)(bl\sigma)}(\phi_0)\frac{\partial u^a_{k,\tau}}{\partial\eta}\bigg|_{t=0}\frac{\partial u^b_{l,\sigma}}{\partial\eta}\bigg|_{t=0}\right]^{1/2}\right\}$$

from (18). Let Λ_ν be a characteristic triangle in $R^{1,1}$ such that $\frac{1}{2}\nu\geqslant\xi\geqslant-\eta\geqslant-\frac{1}{2}\nu$, where

$$\nu=\min\left\{L,\alpha\sqrt{\frac{A}{M}}\right\},$$

and α,A are the positive constants that appeared in Lemma 1.

Lemma 3: For s with the restriction as in Lemma 1 (or the restriction in Lemma 1′), there exists a unique solution of (17) or (19) on Λ_ν satisfying the initial condition (18).

Outline of Proof: We consider, in a similar way as done in Ref. 8, the first order infinite systems as follows:

$$\begin{gathered}\frac{\partial y^c_{m,\mu}}{\partial\eta}+\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(v)y^a_{k,\tau}z^b_{l,\sigma}=0,\quad \frac{\partial z^c_{m,\mu}}{\partial\xi}+\Gamma^{(cm\mu)}_{(ak\tau)(bl\sigma)}(u)z^a_{k,\tau}y^b_{l,\sigma}=0,\\ \frac{\partial u^c_{m,\mu}}{\partial\xi}=y^c_{m,\mu},\qquad \frac{\partial v^c_{m,\mu}}{\partial\eta}=z^c_{m,\mu},\end{gathered} \tag{20}$$

with the corresponding initial conditions:

$$u^c_{m,\mu}(0,x)=v^c_{m,\mu}(0,x)=(\phi_0)^c_{m,\mu}(x),\quad y^c_{m,\mu}(0,x)=\frac{\partial(\phi_0)^c_{m,\mu}(x)}{\partial x}+(\phi_1)^c_{m,\mu}(x),\tag{21}$$

$$z^c_{m,\mu}(0,x)=-\frac{\partial(\phi_0)^c_{m,\mu}(x)}{\partial x}+(\phi_1)^c_{m,\mu}(x).$$

It is easy to see that if we can construct a unique solution of (20), (21) satisfying $u^c_{m,\mu}(t,x)=v^c_{m,\mu}(t,x)$, $\forall c,m,\mu$, then $u=\{u^c_{m,\mu}\}$ is the unique solution of (17), (18) over Λ_ν.

Thus, our aim is to construct a solution of (20), (21) on Λ_ν and to prove that such a solution must satisfy $u^c_{m,\mu}=v^c_{m,\mu},\forall c;m,\mu$. In fact, by using the same methods used in proving Lemma 2 in Ref. 8 with the aid of Corollary 1 (i.e., a local existence result for a linear partial differential system in Hilbert space) stated in Sec. IV (A) in Ref. 8, an elementary result in Riemann–Hilbert geometry which says that the length of a parallel vector field is still preserved invariant and the completeness of Hilbert Lie groups, we may be able to arrive at this conclusion. It should be pointed out that Lemma 1 and Lemma 2 (or Lemma 1′ and Lemma 2′) have played an essential role in this proof.

Next we come to state our main existence results of globally defined harmonic maps from the Minkowski plane into Hilbert Lie loop groups $H_s(LG)(s>1/2)$ or Hilbert Lie gauge groups $H_s(S^n,G)$ $(n\geqslant 2;s>(n+1)/2)$. These answer the two questions remaining from Ref. 8.

Theorem 1: There exists a unique globally defined harmonic map $u:R^{1,1}\to H_s(LG)$ $(s>1/2)$ satisfying a given initial conditions (18).

Theorem 2: There exists a unique globally defined harmonic map $u:R^{1,1}\to H_s(S^n,G)$ $(n\geqslant 2;s>(n+1)/2)$ satisfying a given initial conditions (18).

Since the proofs of Theorem 1 or 2 mainly depend on Lemma 3, we can give a unified proof of Theorem 1 and 2 as follows:

Proof: It is sufficient to prove the existence of a solution for the Cauchy problem (17), (18) in a given characteristic triangle $\Lambda_L=\{-\frac{1}{2}L\leqslant-\eta\leqslant\xi\leqslant\frac{1}{2}L\}$, where L is an arbitrary fixed positive number.

We write the initial conditions in the form

$$u\Big|_{t=0}=\phi_0(x),\quad \frac{\partial u}{\partial t}\Big|_{t=0}=\phi_1(x),$$

where $\phi_0=\phi_0(x)$ is a curve γ in $\tilde{G}=H_s(S^n,G)(n\geqslant 2,s>(n+1)/2$ or $n=1,s>1/2)$ and $\phi_1=\phi_1(x)$ is a tangent vector field along γ. Let

$$M=\max\left\{\sup_{-L\leqslant x\leqslant L}\left\|\frac{\partial u}{\partial\xi}\right\|_{t=0},\ \sup_{-L\leqslant x\leqslant L}\left\|\frac{\partial u}{\partial\eta}\right\|_{t=0}\right\},$$

where $\|\cdot\|$ denotes the length of a vector field on $\tilde{G}$ with respect to the H_s-metric. Put $\rho=\min\{L,\alpha\sqrt{A/M}\}$ where α,A are the constants that appeared in Lemma 1 (or Lemma 1′).

From Lemma 3 we see that the Cauchy problem (17), (18) on each characteristic triangle whose base side is not larger than 2ρ and lies in the interval $-L\leqslant x\leqslant L$ admits a solution. By uniqueness, on any overlapping region of two such characteristic triangles the two solutions coincide. Hence the Cauchy problem is solvable up to $t=\rho$ and the solution $u=p(t,x)$ is smooth.

Next we consider the Cauchy problem for (17) or (19) with the initial conditions

$$u|_{t=\varsigma\rho}=p(\varsigma\rho,x),\quad \frac{\partial u}{\partial t}\Big|_{t=\varsigma\rho}=\frac{\partial p}{\partial t}(\varsigma\rho,x),\qquad -L+\varsigma\rho\leqslant x\leqslant L-\varsigma\rho,\tag{22}$$

where $\varsigma>0$ is a fixed positive integer, provided the problem is solved up to $t=\varsigma\rho$. Evidently, on the line segment $-L+\varsigma\rho\leqslant x\leqslant L-\varsigma\rho$, we have

$$\left\|\frac{\partial p}{\partial \xi}\right\|_{t=s\rho}=\left\|\frac{\partial p}{\partial \xi}\right\|_{t=0}\leqslant M,\quad \left\|\frac{\partial p}{\partial \xi}\right\|_{t=s\rho}=\left\|\frac{\partial p}{\partial \xi}\right\|_{t=0}\leqslant M,$$

by the conservational law of harmonic map systems on $R^{1,1}$ [refer to (24) in Ref. 8 for details]. Therefore, by using Lemma 1 (b) [or Lemma 1′ (b)], we may also choose the same $\rho=\min\{L,\alpha\sqrt{A/M}\}$ for obtaining the local existence result of the Cauchy problem (17), (22). With the same arguments above, we see that the Cauchy problem for (17), (22) is solvable up to $t=(s+1)\rho$. Using this procedure successively we conclude the existence of a global solution on Λ_L. The uniqueness of solutions is obvious. This completes the proofs of Theorems 1–2.

V. CONCLUSION AND REMARKS

In the present paper, we have proved the global existence of the Cauchy problem of harmonic maps from $R^{1,1}$ into Hilbert loop groups $H_s(LG)$ $(s>1/2)$ or into Hilbert Lie gauge groups $H_s(S^n,G)$ $(n\geqslant 2; s>(n+1)/2)$. The proof strongly depends on certain geometric properties of these infinite-dimensional groups (Lemmas 1–2). It is natural to ask the following question: do there exist globally defined harmonic maps from $R^{1,1}$ into $H_s(S^n,G)$ when $n\geqslant 2$ and $n/2<s\leqslant(n+1)/2$ too?

The analogous results to the Hilbert Lie gauge groups $H_s(M,G)$ for M to be a real projective space RP^n or complex projective space CP^n may be established in the same way, since one has also known the detailed information about the spectrum of the Laplace operator on RP^n or CP^n. However, the same geometric properties (Lemmas 1–2) of the Hilbert Lie gauge groups $H_s(M,G)$ $(\dim M\geqslant 2)$ for a general Riemannian manifold M are still to be characterized, since one has not had a full understanding of the harmonic theory on an arbitrary Riemannian manifold M yet.

Finally we would like to mention a related interesting conjecture proposed by Atiyah (Ref. 17) states that any harmonic map from S^2 to the based loop group $\Omega(SU(2))$ (with the $H_{1/2}$-metric) is either a holomorphic or anti-holomorphic map. Combining it with an excellent work by Sibner, Sibner, and Uhlenbeck (Ref. 18), we wonder if it is possible to construct a counterexample to Atiyah's conjecture. Similar to a story, which says that the study of Yang–Mills systems over the 4-dimensional Minkowski space $R^{3,1}$ can be transformed to the study of Yang–Mills systems over the 4-dimensional Euclidean space R^4 by the path integral approach, the study of globally defined harmonic maps from $R^{1,1}$ into loop groups treated in this paper perhaps relates to constructions of nontrivial harmonic maps from R^2 or S^2 (for $S^2=R^2\cup\{\infty\}$) into the loop group $\Omega(G)$.

ACKNOWLEDGMENTS

The author thanks Dr. Xiangao Liu for his interest and helpful discussions in this work. He is also very grateful to the referee for his/her helpful comments. This work is supported partially by NNSF (19531050) of China and STDF of Shanghai, China.

[1] W. Killing, Math. Ann. **31**, 252 (1888); **33**, 1 (1889); **36**, 161 (1894).
[2] E. Cartan, Sur la structure des groupes finis et continus, Thése, 1894.
[3] I. M. Gelfand and M. I. Graev, Funkt. Anal. Pril. **2**, 20 (1968).
[4] S. Albeverio *et al.*, *Noncommutative Distributions*, Unitary Representation of Gauge Groups and Algebras (Marcel Dekker, New York, 1993).
[5] A. Pressley and G. Segal, *Loop Groups* (Clarendon Press, Oxford, 1986).
[6] V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd ed. (Cambridge University Press, Cambridge, 1985).
[7] D. S. Freed, J. Diff. Geom. **28**, 223 (1988).
[8] Q. Ding and B. Lu, J. Math. Phys. **37**, 4076 (1996).
[9] C. H. Gu, Commun. Pure Appl. Math. **33**, 727 (1980).
[10] J. Ginibre and G. Velo, Ann. Phys. (N.Y.) **142**, 393 (1982).
[11] J. Shatah, Commun. Pure Appl. Math. **41**, 459 (1988).
[12] T. Sideris, Commun. Pure Appl. Math. **42**, 1 (1989).
[13] Y. Choquet-Bruhat and C. H. Gu, C. R. Acad. Sci. Paris **308**, 167 (1989).
[14] Q. Ding, Chinese J. Contemp. Math. **17**, 59 (1996).
[15] G. B. Folland, *Introduction to Partial Differential Equations* (Princeton University Press, Princeton, 1976).
[16] W. P. do Carmo and N. R. Wallach, Ann. Math. **93**, 43 (1971).
[17] M. F. Atiyah, Commun. Math. Phys. **93**, 437 (1984).
[18] L. M. Sibner, R. J. Sibner, and K. Uhlenbeck, Proc. Natl. Acad. Sci. USA **86**, 8610 (1989).

The topological structure of the space–time disclination

Yishi Duan[a)]
Institute of Theoretical Physics, Lanzhou University, Lanzhou, 730000, People's Republic of China

Sheng Li[b)]
Institute of Theoretical Physics, Academia Sinica, Beijing 100080, People's Republic of China

(Received 8 July 1998; accepted for publication 4 August 1998)

The space–time disclination is studied by making use of the decomposition theory of gauge potential in terms of the antisymmetric tensor field and ϕ-mapping method. It is shown that the self-dual and anti-self-dual parts of the curvature compose the space–time disclinations which are classified in terms of topological invariants—winding number. The projection of space–time disclination density along an antisymmetric tensor field is quantized topologically and characterized by Brouwer degree and Hopf index. © *1998 American Institute of Physics.* [S0022-2488(98)03412-4]

I. INTRODUCTION

The topological space–time defects, if they exist, may help to explain some of the largest-scale structures seen in the universe today. They may have been found at phase transitions in the early history of the universe, analogous to those found in some condensed matter systems—the vortex line in liquid helium, flux tubes in type-II superconductors, or disclination lines in liquid crystals.

As a kind of space–time defect, torsion plays an important role in modern physics. Recently, a lot of works on the spin and torsion have been done by many physicists.[1–11] Many of them[8–11] are focused on the so-called space–time dislocation by considering the effect of topology of torsion in the Riemann–Cartan manifold. The consideration partially comes from the gauge field theory of dislocation and disclination continuum about deformable material media.[12–16] These works are trying to clarify the quantization of the gravitation (i.e., to quantize the Riemann–Cartan space-time itself). Moreover, they hope to find that the dynamical relation between the stress-energy-momentum tensor and curvature is expressed in general relativity by Einstein equations.

On the other hand, there exists another kind of space–time defect called space–time disclination. This kind of defect has the same important meaning as that of the torsion. Space-time disclination reflects the intrinsic property of the curved space–time. It is caused by inserting solid angles into the flat space–time. In Riemann–Cartan geometry, this effect is shown by the integral of the affine curvature along a closed surface. Duan, Duan and Zhang[12] had discussed the disclinations in deformable material media by applying the gauge field theory and decomposition theory of gauge potential. In their works, the projection of disclination density along the gauge parallel vector was found to correspond to a set of isolated disclinations in the three-dimensional sense and to be topologically quantized. The decomposition theory of gauge potential and ϕ-mapping method that they used are good tools to study the problems of gauge field.

In this paper, we discuss the space–time disclinations in a similar way to that used in Duan *et al.*'s works on the condensed matter disclination.[12] The decomposition formulas of SO(4) and SU(2) gauge potentials are given in terms of the antisymmetric tensor field and unit vector field, respectively. Using this decomposition theory the space–time disclinations are classified into two classes which are marked by the self-dual and anti-self-dual parts of the curvature. The projection of space–time disclination density along the gauge parallel tensor can be written in terms of winding numbers. Moreover we show that the projection of space–time disclination density,

a)Electronic mail: ysduan@lzu.edu.cn
b)Author to whom all correspondence should be addressed; electronic mail: lisheng@itp.ac.cn

 © 1998 American Institute of Physics

which is quantized topologically, corresponds to two groups of isolated disclinations. The positions of the disclination vertices are determined by the zeroes of the self-dual or anti-self-dual part of an antisymmetric tensor. And the Hopf index and Brouwer degree classify the disclinations and characterize the local nature of the space–time disclinations. For this quantization of the space-time, disclinations are directly related to the space–time curvature. As will be shown in this paper, it perhaps relates to the quantization of space–time.

This paper is arranged as follows: In Sec. II, we discuss the representation of the space–time disclination. In Sec. III, the decomposition of the SO(4) gauge potential is given by considering the relationship between SO(4) gauge theory and SU(2) gauge theory. Finally, we discuss the topological quantization and the local topological property of the space–time disclination in Sec. IV.

II. THE REPRESENTATION OF SPACE-TIME DISCLINATION

Let $\mathbf{M}$ be a compact, oriented four-dimensional Riemannian–Cartan manifold and $P(\mathbf{M},G,\pi)$ be a principal bundle with structure group $G=\mathrm{SO}(4)$. As was shown in Ref. 12, the dislocation and disclination continuum can be described by the reference, deformed and natural states. For the natural state there is only an anholonomic rectangular coordinate Z^a $(a=1,2,3,4)$ and

$$\delta Z^a = e^a_\mu \, dx^\mu, \tag{1}$$

where e^a_μ is vielbein. The metric tensor of the Riemann–Cartan manifold of natural state is defined by

$$g_{\mu\nu} = e^a_\mu e^a_\nu. \tag{2}$$

We have known that the metric tensor $g_{\mu\nu}$ is invariant under the local SO(4) transformation of vielbein. The corresponding gauge covariant derivative one-form of an antisymmetric tensor field $\phi^{ab}=-\phi^{ba}$ on $\mathbf{M}$ is given as

$$D\phi^{ab} = d\phi^{ab} - \omega^{ac}\phi^{cb} - \omega^{bc}\phi^{ac}, \tag{3}$$

where ω^{ab} is the SO(4) spin connection one-form

$$\omega^{ab} = -\omega^{ba}, \quad \omega^{ab} = \omega^{ab}_\mu \, dx^\mu. \tag{4}$$

The affine connection of the Riemann–Cartan space is determined by

$$\Gamma^\lambda_{\mu\nu} = e^{\lambda a} D_\mu e^a_\nu. \tag{5}$$

The torsion tensor is the antisymmetric part of $\Gamma_{\mu\nu}$ and is expressed as

$$T^\lambda_{\mu\nu} = e^{\lambda a} T^a_{\mu\nu}, \tag{6}$$

where

$$T^a_{\mu\nu} = \tfrac{1}{2}(D_\mu e^a_\nu - D_\nu e^a_\mu). \tag{7}$$

The Riemannian curvature tensor is equivalent to the SO(4) gauge field strength tensor two-form F^{ab}, which is given by

$$F^{ab} = d\omega^{ab} - \omega^{ac}\wedge\omega^{cb}, \quad F^{ab} = \tfrac{1}{2}F^{ab}_{\mu\nu} \, dx^\mu \wedge dx^\nu \tag{8}$$

and relates to the Riemann curvature tensor by

$$F^{ab}_{\mu\nu} = -R^\lambda_{\mu\nu\sigma} e^a_\lambda e^{\sigma b}. \tag{9}$$

The dislocation density is defined by

$$\alpha^a = T^a, \quad T^a = \tfrac{1}{2}T^a_{\mu\nu} \, dx^\mu \wedge dx^\nu. \tag{10}$$

Analogous to the definition of the three-dimensional disclination density in the gauge field theory of condensed matter, we define the space–time disclination density as

$$\theta^{ab}=\tfrac{1}{2}R^{\lambda}_{\mu\nu\sigma}e^{a}_{\lambda}e^{\sigma b}\ dx^{\mu}\wedge dx^{\nu}=-F^{ab}. \tag{11}$$

The size of the space–time disclination can be represented by means of the surface integral of the projection of the space–time disclination density along an antisymmetric tensor field n^{ab},

$$\Omega=\oint_{\Sigma}\theta^{ab}n^{ab}=-\oint_{\Sigma}F^{ab}n^{ab}, \tag{12}$$

where Σ is a closed surface including the disclinations. The new quantity Ω defined by (12) is dimensionless. Using the so-called ϕ-mapping method and the decomposition of gauge potential Duan *et al*. have proved that the dislocation flux is quantized in units of the Planck length.[11] In this paper we will show the size of the space–time disclination (12) is topologically quantized and is the sum of two groups of solid angles representing the size of disclinations.

III. DECOMPOSITION THEORY OF SO(4) SPIN CONNECTION

In this section we will give the decomposition theory of the SO(4) gauge potential in terms of an antisymmetric tensor field on a compact and oriented four-dimensional manifold, which is the foundation of topological gauge field theory. It will be seen in the following discussion that this decomposition of the SO(4) gauge potential is expressed by the combination of the SU(2) gauge potential. Hence we must first study the decomposition of the SU(2) gauge potential in terms of the sphere bundle on a compact and oriented four-dimensional manifold.

Let n be a unit SU(2) Lie algebraic vector

$$n=n^{A}I_{A},\quad A=1,2,3, \tag{13}$$

and

$$n^{A}n^{A}=1, \tag{14}$$

in which I_A is the generator of the group SU(2), which satisfies the commutation relation

$$[I_A,I_B]=-\epsilon^{ABC}I_C. \tag{15}$$

The covariant derivative one-form of n is given by

$$D_{\mathrm{SU}(2)}n=dn-[\omega_{\mathrm{SU}(2)},n], \tag{16}$$

where $\omega_{\mathrm{SU}(2)}$ is the SU(2) gauge potential one-form:

$$\omega_{\mathrm{SU}(2)}=\omega^{A}_{\mathrm{SU}(2)}I_A, \tag{17}$$

and

$$\omega^{A}_{\mathrm{SU}(2)}=\omega^{A}_{\mathrm{SU}(2)\mu}\ dx^{\mu},\quad \mu=0,1,2,3. \tag{18}$$

The curvature two-form is

$$F_{\mathrm{SU}(2)}=F^{A}_{\mathrm{SU}(2)}I_A=d\omega_{\mathrm{SU}(2)}-\omega_{\mathrm{SU}(2)}\wedge\omega_{\mathrm{SU}(2)}, \tag{19}$$

and

$$F^{A}_{\mathrm{SU}(2)}=\tfrac{1}{2}F^{A}_{\mathrm{SU}(2)\mu\nu}\ dx^{\mu}\wedge dx^{\nu}=d\omega^{A}_{\mathrm{SU}(2)}+\tfrac{1}{2}\epsilon^{ABC}\omega^{B}_{\mathrm{SU}(2)}\wedge\omega^{C}_{\mathrm{SU}(2)}. \tag{20}$$

Let $m=m^{A}I_A$, $l=l^{A}I_A$ be another two unit SU(2) algebraic vector orthornormal to n and to each other, satisfying

$$n^A = \epsilon^{ABC} m^B l^C. \tag{21}$$

It can be proved that the SU(2) gauge potential can be decomposed by n, m, l as

$$\omega^A_{\mathrm{SU}(2)} = \epsilon^{ABC}(dn^B\, n^C - D_{\mathrm{SU}(2)} n^B n^C) - n^A A, \tag{22}$$

where

$$A = dm^A\, l^A - D_{\mathrm{SU}(2)} m^A l^A \tag{23}$$

is a U(1)-like gauge potential.

If n is taken as a gauge parallel vector, i.e.,

$$D_{\mathrm{SU}(2)} n = 0, \tag{24}$$

the SU(2) gauge potential becomes

$$\omega^A_{\mathrm{SU}(2)} = \epsilon^{ABC}\, dn^B\, n^C - n^A A. \tag{25}$$

Then we can easily get the curvature two-form $F^A_{\mathrm{SU}(2)}$ as

$$F^A_{\mathrm{SU}(2)} = n^A\, dA - \tfrac{1}{2}\epsilon^{ABC}\, dn^B \wedge dn^C. \tag{26}$$

Now let us consider the SO(4) gauge theory. Let the four-dimensional Dirac matrix γ_a ($a = 1,2,3,4$) be the basis of the Clifford algebra which satisfies

$$\gamma_a \gamma_b + \gamma_b \gamma_a = 2\delta_{ab}. \tag{27}$$

The antisymmetric tensor field ϕ^{ab} on $\mathbf{M}$ can be expressed in the following matrix form:

$$\phi = \tfrac{1}{2}\phi^{ab} I_{ab}, \tag{28}$$

in which I_{ab} is the generator of the group SO(4),

$$I_{ab} = \tfrac{1}{4}[\gamma_a, \gamma_b]. \tag{29}$$

Similarly, the spin connection one-form and curvature two-form can be expressed as

$$\omega = \tfrac{1}{2}\omega^{ab} I_{ab}, \quad F = \tfrac{1}{2} F^{ab} I_{ab}. \tag{30}$$

It is well known that the spin representation of the SO(4) group is hormorphic to the direct product of the representations of two SU(2) groups,[17]

$$\mathrm{so}(4) \cong \mathrm{su}(2) \otimes \mathrm{su}(2). \tag{31}$$

The generators of the SO(4) group can be divided into two terms. Each term is a generator of some SU(2) group. Define

$$I^1_{\pm} = \tfrac{1}{2} I_{23}(1 \mp \gamma_5) = \tfrac{1}{4}(\gamma_2 \gamma_3 \pm \gamma_1 \gamma_4), \tag{32}$$

$$I^2_{\pm} = \tfrac{1}{2} I_{31}(1 \mp \gamma_5) = \tfrac{1}{4}(\gamma_3 \gamma_1 \pm \gamma_2 \gamma_4), \tag{33}$$

$$I^3_{\pm} = \tfrac{1}{2} I_{12}(1 \mp \gamma_5) = \tfrac{1}{4}(\gamma_1 \gamma_2 \pm \gamma_3 \gamma_4), \tag{34}$$

in which $\gamma_5 = \gamma_1 \gamma_2 \gamma_3 \gamma_4$. It can be proved $I^A_{\pm} (A = 1,2,3)$ satisfy the commutation relation of group $\mathrm{SU}(2)_{\pm}$,

$$I^A_{\pm} = -\epsilon^{ABC}[I^B_{\pm}, I^C_{\pm}], \tag{35}$$

and

$$[I_{\pm}^{A}, I_{\mp}^{B}] = 0. \tag{36}$$

Therefore $I_{\pm}^{A}$ are the generators of two different groups $\mathrm{SU}(2)_{\pm}$, and they are the basis of $\mathrm{SU}(2)_{\pm}$ Lie algebraic spaces.

An arbitrary antisymmetric tensor field ϕ^{ab} can be decomposed as

$$\phi^{ab} = \phi_{+}^{ab} + \phi_{-}^{ab}, \tag{37}$$

where

$$\phi_{+}^{ab} = \tfrac{1}{2}(\phi^{ab} + \tfrac{1}{2}\epsilon^{abcd}\phi^{cd}), \quad \phi_{-}^{ab} = \tfrac{1}{2}(\phi^{ab} - \tfrac{1}{2}\epsilon^{abcd}\phi^{cd}) \tag{38}$$

are the self-dual and anti-self-dual parts of ϕ^{ab}. Define

$$\phi_{\pm}^{1} = \phi^{23} \pm \phi^{14}, \quad \phi_{\pm}^{2} = \phi^{31} \pm \phi^{24}, \quad \phi_{\pm}^{3} = \phi^{12} \pm \phi^{34}. \tag{39}$$

We can rewrite ϕ as

$$\phi = \tfrac{1}{2}\phi^{ab} I_{ab} = \phi_{+} + \phi_{-}, \tag{40}$$

and

$$\|\phi_{\pm}\|^{2} = \phi_{\pm}^{A}\phi_{\pm}^{A} = \phi_{\pm}^{ab}\phi_{\pm}^{ab}, \tag{41}$$

in which

$$\phi_{\pm} = \tfrac{1}{2}\phi(1 \mp \gamma_5) = \phi_{\pm}^{A} I_{\pm}^{A}. \tag{42}$$

From the above discussion, we see that $\phi_{\pm}$ are the $\mathrm{SU}(2)_{\pm}$ algebraic vectors. With a similar decomposition as above it can be verified that

$$D\phi_{+} + D\phi_{-} = d\phi_{+} - [\omega_{+}, \phi_{+}] + d\phi_{-} - [\omega_{-}, \phi_{-}]. \tag{43}$$

For the independent of I_{+} and I_{-} we have

$$D\phi_{\pm} = d\phi_{\pm} - [\omega_{\pm}, \phi_{\pm}] \quad or \quad D\phi_{\pm}^{A} = d\phi_{\pm}^{A} - \epsilon^{ABC}\omega_{\pm}^{C}\phi_{\pm}^{B}, \tag{44}$$

in which

$$\omega_{\pm} = \omega_{\pm}^{A} I_{\pm}^{A} \tag{45}$$

is the gauge potential of the $\mathrm{SU}(2)_{\pm}$ gauge field. Similarly curvature two-form can be decomposed as

$$F = F_{+} + F_{-}, \tag{46}$$

and

$$F_{\pm} = d\omega_{\pm} - \omega_{\pm} \wedge \omega_{\pm}. \tag{47}$$

$F_{\pm} = F_{\pm}^{A} I_{\pm}^{A}$ are the curvature two-forms of the $\mathrm{SU}(2)_{\pm}$ gauge field.

Now define an antisymmetric tensor

$$n^{ab} = \frac{\phi_{+}^{ab}}{\|\phi_{+}\|} + \frac{\phi_{-}^{ab}}{\|\phi_{-}\|}, \tag{48}$$

i.e.,

$$n_{+}^{ab} = \frac{\phi_{+}^{ab}}{\|\phi_{+}\|}, \quad n_{-}^{ab} = \frac{\phi_{-}^{ab}}{\|\phi_{-}\|}. \tag{49}$$

Then

$$n=\tfrac{1}{2}n^{ab}I_{ab}=n_{+}+n_{-} \tag{50}$$

and

$$n_{\pm}=\frac{1}{2\|\phi_{\pm}\|}\phi(1\mp\gamma_5)=n_{\pm}^{A}I_{\pm}^{A}\,. \tag{51}$$

It naturally guarantees the constraint

$$n_{\pm}^{A}n_{\pm}^{A}=1, \tag{52}$$

i.e., $n_{\pm}^{A}$ are the $\mathrm{SU}(2)_{\pm}$ lie algebraic unit vectors. If n is taken as a gauge parallel tensor field

$$Dn=0, \tag{53}$$

we have

$$Dn_{\pm}=dn_{\pm}-[\omega_{\pm},n_{\pm}]=0. \tag{54}$$

From the decomposition formula of the SU(2) gauge potential (26) we get

$$F_{\pm}^{A}=n_{\pm}^{A}\ dA_{\pm}-\tfrac{1}{2}\epsilon^{ABC}\ dn_{\pm}^{B}\wedge dn_{\pm}^{C}\,. \tag{55}$$

This formula is very useful in the following discussion of space–time disclinations.

IV. THE TOPOLOGICAL QUANTIZATION AND THE LOCAL TOPOLOGICAL STRUCTURE OF THE SPACE–TIME DISCLINATION

It is easy to see that

$$F^{ab}n^{ab}=F_{+}^{ab}n_{+}^{ab}+F_{-}^{ab}n_{-}^{ab} \tag{56}$$

and

$$F_{\pm}^{A}n_{\pm}^{A}=F_{\pm}^{ab}n_{\pm}^{ab}\,. \tag{57}$$

By considering (55) we have

$$F_{\pm}^{A}n_{\pm}^{A}=dA_{\pm}-\tfrac{1}{2}\epsilon^{ABC}n_{\pm}^{A}\ dn_{\pm}^{B}\wedge dn_{\pm}^{C}\,. \tag{58}$$

Define

$$\Omega_{\pm}=\oint_{\Sigma}\left(\frac{1}{2}\epsilon^{ABC}n_{\pm}^{A}\ dn_{\pm}^{B}\wedge dn_{\pm}^{C}-dA_{\pm}\right). \tag{59}$$

$\Omega_{\pm}$ represent the components of space–time disclination corresponding to a different gauge group $\mathrm{SU}(2)_{\pm}$. The space–time disclination can be expressed as

$$\Omega=\Omega_{+}+\Omega_{-}\,. \tag{60}$$

If the surface Σ is closed, the second term on the right-hand side of (59) contributes nothing to $\Omega_{\pm}$, i.e.,

$$\Omega_{\pm}=\oint_{\Sigma}\frac{1}{2}\epsilon^{ABC}n_{\pm}^{A}\ dn_{\pm}^{B}\wedge dn_{\pm}^{C}\,. \tag{61}$$

To study the relationship between the solid angle $\Omega_{\pm}$ and the local properties of the disclinations inside the closed surface Σ, using the Stokes formula, $\Omega_{\pm}$ can be expressed as

$$\Omega_{\pm}=\int_V \frac{1}{2}\epsilon^{ABC}\, dn_{\pm}^A\wedge dn_{\pm}^B\wedge dn_{\pm}^C, \tag{62}$$

in which $\partial V=\Sigma$. Let us choose coordinates $y=(u^1,u^2,u^3,v)$ on M such that $u=(u^1,u^2,u^3)$ are intrinsic coordinate on V. For the coordinate component, v does not belong to V. Then

$$\Omega_{\pm}=\int_V \frac{1}{2}\epsilon^{ABC}\partial_i n_{\pm}^A\partial_j n_{\pm}^B\partial_k n_{\pm}^C\, du^i\wedge du^j\wedge du^k=\int_V \frac{1}{2}\epsilon^{ijk}\epsilon^{ABC}\partial_i n_{\pm}^A\partial_j n_{\pm}^B\partial_k n_{\pm}^C d^3u, \tag{63}$$

where $i,j,k=1,2,3$ and $\partial_i=\partial/\partial u^i$. Define solid angle densities as

$$\rho_{\pm}=\tfrac{1}{2}\epsilon^{ijk}\epsilon^{ABC}\partial_i n_{\pm}^A\partial_j n_{\pm}^B\partial_k n_{\pm}^C. \tag{64}$$

Then we get

$$\Omega_{\pm}=\int_V \rho_{\pm}d^3u. \tag{65}$$

For Eq. (51), the unit vectors $n_{\pm}^A(x)$ can expressed as follows:

$$n_{\pm}^A=\frac{\phi_{\pm}^A}{\|\phi_{\pm}\|}. \tag{66}$$

Hence

$$dn_{\pm}^A=\frac{1}{\|\phi_{\pm}\|}d\phi_{\pm}^A+\phi_{\pm}^A d\left(\frac{1}{\|\phi_{\pm}\|}\right), \tag{67}$$

and

$$\frac{\partial}{\partial\phi_{\pm}^A}\left(\frac{1}{\|\phi_{\pm}\|}\right)=-\frac{\phi_{\pm}^A}{\|\phi_{\pm}\|^3}. \tag{68}$$

Substituting the above equations into the solid angle density (64) we obtain

$$\begin{aligned}\rho&=\frac{1}{2}\epsilon^{ABC}\epsilon^{ijk}\partial_i(n_{\pm}^A\partial_j n_{\pm}^B\partial_k n_{\pm}^C)\\&=\frac{1}{2}\epsilon^{ABC}\epsilon^{ijk}\partial_i\frac{\phi_{\pm}^A}{\|\phi_{\pm}\|^3}\partial_j\phi_{\pm}^B\partial_k\phi_{\pm}^C\\&=-\frac{1}{2}\epsilon^{ABC}\epsilon^{ijk}\frac{\partial}{\partial\phi_{\pm}^D}\frac{\partial}{\partial\phi_{\pm}^A}\left(\frac{1}{\|\phi_{\pm}\|}\right)\partial_i\phi_{\pm}^D\partial_j\phi_{\pm}^B\partial_k\phi_{\pm}^C.\end{aligned} \tag{69}$$

Define the Jacobian $J(\phi_{\pm}/u)$ as

$$\epsilon^{ABC}J\left(\frac{\phi_{\pm}}{u}\right)=\epsilon^{ijk}\partial_i\phi_{\pm}^A\partial_j\phi_{\pm}^B\partial_k\phi_{\pm}^C. \tag{70}$$

By making use of the Laplacian relation in ϕ-space,

$$\partial_A\partial_A\frac{1}{\|\phi_{\pm}\|}=-4\pi\delta^3(\phi_{\pm}),\quad \partial_A=\frac{\partial}{\partial\phi_{\pm}^A}, \tag{71}$$

we can write the density of the solid angle as the δ-like expression

$$\rho_{\pm}=4\pi\delta^3(\phi_{\pm})J\left(\frac{\phi_{\pm}}{u}\right) \tag{72}$$

and

$$\Omega_{\pm}=\int_V 4\pi\delta^3(\phi_{\pm})J\left(\frac{\phi_{\pm}}{u}\right)d^3u. \tag{73}$$

It is obvious that $\rho_{\pm}$ are nonzero only when $\phi_{\pm}=0$.

Suppose that $\phi_{\pm}^A(x)$ $(A=1,2,3)$ possess $K_{\pm}$ isolated zeros, according to the deduction of Ref. 18 and the implicit function theorem,[19] the solutions of $\phi_{\pm}(u^1,u^2,u^3,v)=0$ can be expressed in terms of $u=(u^1,u^2,u^3)$ as

$$u^i=z_{\pm}^i(v), \quad i=1,2,3 \tag{74}$$

and

$$\phi_{\pm}^A(z_l^1(v),z_l^2(v),z_l^3(v),v)\equiv 0, \tag{75}$$

where the subscript $l=1,2,\ldots,K_{\pm}$ represents the lth zero of $\phi_{\pm}^A$, i.e.,

$$\phi_{\pm}^A(z_{\pm l}^i)=0, \quad l=1,2,\ldots,K_{\pm}, \quad A=1,2,3. \tag{76}$$

It is easy to get the following formula from the ordinary theory of δ-function that

$$\delta^3(\phi_{\pm})J\left(\frac{\phi_{\pm}}{u}\right)=\sum_{i=1}^{K_{\pm}}\beta_{\pm l}\eta_{\pm l}\delta^3(u-z_{\pm l}) \tag{77}$$

in which

$$\eta_{\pm l}=\text{sign } J\left(\frac{\phi_{\pm}}{u}\right)\Bigg|_{x=z_{\pm l}}=\pm 1 \tag{78}$$

is the Brouwer degree of ϕ-mapping and $\beta_{\pm l}$ are positive integers called the Hopf index of map $\phi_{\pm}$. This means that while the point x covers the region neighboring the zero $x=z_{\pm l}$ once, $\phi_{\pm}$ covers the corresponding region $\beta_{\pm l}$ times. Therefore the solid angle density becomes

$$\rho_{\pm}=4\pi\sum_{l=1}^{K_{\pm}}\beta_{\pm l}\eta_{\pm l}\delta^3(u-z_{\pm l}) \tag{79}$$

and

$$\Omega_{\pm}=4\pi\sum_{l=1}^{K_{\pm}}\beta_{\pm l}\eta_{\pm l}\int_V\delta^3(u-z_{\pm l})d^3u=4\pi\sum_{l=1}^{K_{\pm}}\beta_{\pm l}\eta_{\pm l}. \tag{80}$$

Therefore the space–time disclination is

$$\Omega=4\pi\sum_{l=1}^{K_+}\beta_{+l}\eta_{+l}+4\pi\sum_{l=1}^{K_-}\beta_{-l}\eta_{-l}. \tag{81}$$

We find that (79) is the exact density of a system of K_+ and K_- classical pointlike objects with ''charge'' $\beta_{+l}\eta_{+l}$ and $\beta_{-l}\eta_{-l}$ in space–time, i.e., the topological structure of disclinations formally corresponds to a pointlike system. These point objects may be called disclination points as in nematic crystals.[20] In Ref. 21, it was shown that the existence of disclination points is related to a kind of broken symmetry. The dislocations and disclinations appear as singularities of dis-

tortions of an order parameter.[22] In our paper, the disclination points are identified with the isolated zero points of vector field $\phi_{\pm}^{A}(x)$. From (72) we know that these singularities are those of the disclination density as well.

On another hand, the winding number $W_{\pm}$ of the surface Σ and of the mapping $\phi_{\pm}$ is defined as[23]

$$W_{\pm}=\oint_{\Sigma}\frac{1}{8\pi}\epsilon^{ABC}\frac{\phi_{\pm}^{A}}{\|\phi_{\pm}\|^{3}}d\phi_{\pm}^{B}\wedge d\phi_{\pm}^{C}, \tag{82}$$

which is equal to the number of times Σ encloses (or, wraps around) the point $\phi_{\pm}=0$. Hence, the space–time disclinations is quantized by the winding numbers

$$\Omega_{\pm}=4\pi W_{\pm}. \tag{83}$$

The winding number $W_{\pm}$ of the surface Σ can be interpreted or, indeed, defined as the degree of the mappings $\phi_{\pm}$ onto Σ. By (62) and (72) we have

$$\Omega_{\pm}=4\pi\int_{V}\delta(\phi_{\pm})J_{\pm}\left(\frac{\phi_{\pm}}{u}\right)d^{3}u=4\pi\ \deg\ \phi_{\pm}\int_{\phi_{\pm}(V)}\delta(\phi_{\pm})d^{3}\phi_{\pm}=4\pi\ \deg\ \phi_{\pm}, \tag{84}$$

where deg $\phi_{\pm}$ are the degrees of map $\phi_{\pm}:V\to\phi_{\pm}(V)$. Comparing Eq. (84) with (83), it shows the degrees of map $\phi_{\pm}:V\to\phi_{\pm}(V)$ is just the winding number $W_{\pm}$ of surface Σ and map $\phi_{\pm}$, i.e.,

$$\deg\ \phi_{\pm}=W_{\pm}(\Sigma,\phi_{\pm}). \tag{85}$$

Then the space–time disclination is

$$\Omega=4\pi(\deg\ \phi_{+}+\deg\ \phi_{-})=4\pi(W_{+}+W_{-}). \tag{86}$$

Divide V by

$$V=\sum_{l=1}^{K_{\pm}}V_{\pm l} \tag{87}$$

and $V_{\pm l}$ includes only one zero $z_{\pm l}$ of $\phi_{\pm}$, i.e., $z_{\pm l}\in V_{\pm l}$. The winding number $W_{\pm}$ of the surface $\Sigma_{\pm l}=\partial V_{\pm l}$ and the mapping $\phi_{\pm}$ is defined as

$$W_{\pm l}=\oint_{\Sigma_{\pm l}}\frac{1}{8\pi}\epsilon^{ABC}\frac{\phi_{\pm}^{A}}{\|\phi_{\pm}\|^{3}}d\phi_{\pm}^{B}\wedge d\phi_{\pm}^{C}, \tag{88}$$

which is equal to the number of times $\Sigma_{\pm l}$ encloses (or, wraps around) the point $\phi_{\pm}=0$. It is easy to see that

$$W_{\pm}=\sum_{l=1}^{K_{\pm}}W_{\pm l} \tag{89}$$

and

$$|W_{\pm l}|=\beta_{\pm l}. \tag{90}$$

Then

$$\Omega=4\pi\sum_{l=1}^{K_{+}}W_{+l}+4\pi\sum_{l=1}^{K_{-}}W_{-l}. \tag{91}$$

Assuming that is proved in Ref. 24, the increment

$$\frac{1}{2}\epsilon^{ABC}\frac{\phi_{\pm}^{A}}{\|\phi_{\pm}\|^{3}}\, d\phi_{\pm}^{B}\wedge d\phi_{\pm}^{C}=\epsilon^{ABC}\epsilon_{ijk}\frac{\phi_{\pm}^{A}}{\|\phi_{\pm}\|^{3}}\partial_i\phi_{\pm}^{B}\partial_j\phi_{\pm}^{C}dS^{k} \tag{92}$$

may be regarded as the solid angle on $\phi_{\pm}(\Sigma_{\pm l})$ subtended by the surface element dS on $\Sigma_{\pm l}$. Now it is easy to see why the space–time disclination is caused by inserting solid angles into the flat space–time.

Also we can write

$$\Omega=4\pi(N_{+}^{+}+N_{-}^{+})-4\pi(N_{+}^{-}+N_{-}^{-}) \tag{93}$$

in which $N_{+}^{\pm}$ are the sums of the Hopf indexes with respect to $\eta_{+}=\pm 1$ and $N_{-}^{\pm}$ are the sums of the Hopf indexes with respect to $\eta_{-}=\pm 1$. We see that while x covers V once, ϕ_{+}^{A} must cover $\phi_{+}(V)$ N_{+}^{+} times with $\eta=1$ or N_{+}^{-} times with $\eta=-1$ while ϕ_{-}^{A} must cover $\phi_{-}(V)$ N_{-}^{+} times with $\eta=1$ or N_{-}^{-} times with $\eta=-1$. Therefore the topological disclinations are distinguished by the sign of the Jacobian $J(\phi_{\pm}/u)$ at the zero point $x^{\mu}=z_{\pm l}^{\mu}$. Consequently, the global topological property of disclinations is characterized by the Brouwer degrees and Hopf indices of each disclination point, which label the local structure of disclinations.

In this paper, we have studied the topological structure, global and local properties of space–time disclinations. The disclinations we discussed are the general four-dimensional case, which is characterized by the so-called topological points. The decomposition of gauge potential, introduction of the gauge parallel tensor field $n^{ab}(x)$ $(a,b=1,2,3,4)$ and the ϕ-mapping method play important roles in establishing our theory. The disclination points are the zeros of the self-dual or anti-self-dual parts of the tensor field $\phi^{ab}(x)$ and the topological characteristics of the space–time disclinations are determined by the winding numbers which can be represented by the Brouwer degrees and Hopf indices at the disclination points. From the definition of the space–time disclination (11), we know that the space–time disclination is expressed exactly by the curvature of space–time. Therefore the quantization of the space–time disclination has a close relationship with the quantization of space–time.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of P. R. China.

[1] R. Jha, Int. J. Mod. Phys. A **9**, 3959 (1994).
[2] L. L. Smalley and J. P. Krisch, Class. Quantum Grav. **11**, 2375 (1994).
[3] R. Hammond, Gen. Relativ. Gravit. **26**, 1107 (1994).
[4] C. Sivaram and L. C. Garica De Andrade, Astrophys. Space Sci. **201**, 131 (1993).
[5] J. Kánnár, Gen. Relativ. Gravit. **27**, 23 (1995).
[6] C. M. Zhang, Int. J. Mod. Phys. A **8**, 5095 (1993).
[7] V. de Sabbata and Y. Xin, Int. J. Mod. Phys. A **10**, 3663 (1995).
[8] V. de Sabbata, Nuovo Cimento A **107**, 363 (1994).
[9] S. Luo, Int. J. Theor. Phys. **34**, 2009 (1995).
[10] D. K. Ross, Int. J. Theor. Phys. **28**, 1333 (1989).
[11] Y. S. Duan, S. L. Zhang, and S. S. Feng, J. Math. Phys. **35**, 1 (1994); Y. S. Duan, G. H. Yang, and Y. Jiang, Gen. Relativ. Gravit. **29**, 715 (1997); Helv. Phys. Acta **70**, 565 (1997).
[12] Y. S. Duan and Z. P. Duan, Int. J. Eng. Sci. **24**, 513 (1986); Y. S. Duan and S. L. Zhang, *ibid.* **28**, 689 (1990); **29**, 153 (1991); **30**, 153 (1992).
[13] D. G. Edelen, Int. J. Eng. Sci. **18**, 1095 (1980); A. Kadic and D. G. B. Edelen, *ibid.* **20**, 443 (1982); in *A Gauge Field Theory of Dislocations and Disclinations*, Lecture Notes in Physics, No. 174 (Springer, Berlin, 1983).
[14] D. G. B. Edelen and D. C. Lagoudas, in *Gauge Theory and Defects in Solids* (North-Holland, Amsterdam, 1988).
[15] N. D. Mermin, Rev. Mod. Phys. **51**, 591 (1979); L. Michel, *ibid.* **52**, 617 (1980); H. R. Trebin, Adv. Phys. **31**, 195 (1982).
[16] B. K. D. Gairola, in *Dislocations in Solids*, edited by F. R. N. Nabarro (North-Holland, Amsterdam, 1979), Vol. 1.
[17] H. Boerner, *Representation of Groups* (North-Holland, Amsterdam, 1963).
[18] Y. S. Duan and J. C. Liu, in Proceedings of the Johns Hopkins Workshop on Current Problems in Particle Theory, 1987 (unpublished), Vol. 11, p. 183.
[19] É. Goursat, *A Course in Mathematical Analysis*, translated by E. R. Hedrick (1904), Vol. I.
[20] M. Kleman, in Ref. 16, Vol. 5.
[21] P. G. de Gennes, C. R. Acad. Sci. (Paris) **275**, 319 (1972).
[22] J. Friedel, *Dislocations in Solids*, edited by F. R. N. Nabarro (North-Holland, Amsterdam, 1979), Vol. 5.
[23] V. Guillemin and A. Pollack, *Differential Topology* (Prentice-Hall, Englewood Cliffs, NJ, 1974).
[24] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).

Maximum entropy in the generalized moment problem

M. Frontini and A. Tagliani
Dipartimento di Matematica, Politecnico di Milano, Via Bonardi 9, 20133 Milano, Italy

(Received 5 May 1998; accepted for publication 16 June 1998)

The generalized moment problem in the framework of the maximum entropy approach is considered. A proof for the existence conditions of the solution is provided. © *1998 American Institute of Physics.* [S0022-2488(98)01012-3]

I. INTRODUCTION

The problem of estimating an unknown probability density function on the basis of some generalized known moments is an underdetermined inverse problem which occurs frequently in the physical and economic sciences. In the probabilistic framework, for instance, most often all that we are able to obtain from the governing stochastic equations are the moments of the solution, or the equations for moments.

One widely used technique of solving such an underdetermined problem is that of maximum entropy (M.E.) to determine the approximate probability density. The idea of M.E. is simply to choose an approximant function which maximizes some measure of entropy, subject to given moment constraints.

Much attention is paid to reconstruct a probability density on a finite measurable space (Refs. 1–2), while scarce attention has been devoted to probability densities on an infinite measurable space.

In this paper we wish to show how the idea of maximum entropy can be used in the approximate characterization of a probability density function, defined on the interval $[0,+\infty)$, whose first generalized moments are assigned. Necessary and sufficient conditions guaranteeing existence of the solution will be provided.

Let $M+1$ generalized moments $g_0=1$ and g_k, $k=1,2,\ldots,M$ of the probability density $f(t)$ be known,

$$g_k=\int_0^\infty g_k(t)f(t)dt, \quad k=0,\ldots,M \tag{1}$$

[as we are interested in probability measures then $g_0(t)=1$ is assumed].

Define the entropy $H[f]$ of a positive function $f(t)$,

$$H[f]=-\int_0^\infty f(t)\log(f(t))dt, \tag{2}$$

one looks for a solution of the problem

$$\max_f\left\{H[f]-(\lambda_0-1)\left(\int_0^\infty g_0(t)f(t)dt-g_0\right)-\sum_{k=1}^M \lambda_k\left(\int_0^\infty g_k(t)f(t)dt-g_k\right)\right\}, \tag{3}$$

with the constraints

$$\int_0^\infty g_k(t)f(t)dt=g_k, \quad k=0,1,\ldots,M. \tag{4}$$

It is easy to verify that (Ref. 3) $f(t)$, if it exists, will assume the following expression:

$$f(t)=f_M(t)=\exp\left(-\lambda_0-\sum_{j=1}^M \lambda_j g_j(t)\right), \tag{5}$$

 © 1998 American Institute of Physics

constrained by

$$\int_0^\infty g_k(t) f_M(t)\,dt = g_k, \quad k=0,\ldots,M. \tag{6}$$

The unicity of the M.E. solution, if it exists, is guaranteed by convexity of the functional $\Gamma(\lambda_1,\ldots,\lambda_M)$ defined by (Ref. 3)

$$\Gamma(\lambda_1,\ldots,\lambda_M) = \sum_{k=1}^{M} \lambda_k g_k + \ln\left(\int_0^\infty \exp\left(-\sum_{k=1}^{M} \lambda_k g_k(t) \right) dt \right). \tag{7}$$

The existence proof of the M.E. solution will be carried out in two parts: in Sec. III we will show that the lower bound of the generalized moment g_j is analogous to those furnished by the general moment problem; in Sec. IV we will provide, when they exist, the upper bound for the moment g_j.

The proof is an extension of a technique previously introduced (Refs. 4 and 5) that allows us to furnish the lower and upper bounds (when these exist) of the moment g_j, after having transformed the problems (5), (6) in an equivalent system of differential equations.

II. BACKGROUND

Our technique to find the admissible values of the moments g_j, $j=1,...,M$ consists in transforming (6) in a system of differential equations as follows. Let us maintain the moments g_j, $j=1,...,M-1$ constant and let us vary g_M in continuity. Then $\lambda_0,\ldots,\lambda_M$ are functions of g_M. By differentiating both sides of (6) with respect to g_M we have the system of differential equations,

$$G_M \cdot \begin{bmatrix} d\lambda_0/dg_M \\ \vdots \\ \vdots \\ d\lambda_M/dg_M \end{bmatrix} = -\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{8}$$

where

$$G_M = \begin{bmatrix} g_0 & g_1 & \cdots & g_M \\ g_1 & g_{11} & \cdots & g_{1M} \\ \vdots & \vdots & \vdots & \vdots \\ g_M & g_{M1} & \cdots & g_{MM} \end{bmatrix}, \tag{9}$$

with $g_{ij} = \int_0^\infty g_i(t) g_j(t) f_M(t)\,dt$, $i,j=1,\ldots,M$, is the bordered cross-product matrix, in general positive definite being strictly related to the variance–covariance matrix.

A. The functions $g_j(t)$

The functions $g_j(t)$, $j=1,\ldots,M$, defined on the interval $[0,+\infty)$, are linearly independent continuous functions of a proper Hilbert space. In various concrete questions it is necessary to deal with systems of functions $g_j(t)$, $j=0,\ldots,M$ satisfying the following properties:

(1) $g_j(t)>0$ $(0\leqslant t<\infty)$;

(2) $\lim_{t\to\infty}[g_j(t)/g_M(t)]=0$, $j=0,\ldots,M-1$.

From the adopted conditions on $g_j(t)$, $j=1,\ldots,M$ then (5) satisfies the following constraint:

$$\lambda_M \geqslant 0, \tag{10}$$

and $\lambda_{M-1}>0$ whenever $\lambda_M=0$.

Let us consider $f_M(t)$ where the multipliers λ_j, $j=0,\ldots,M$ assume finite values and the measure $d\psi(t)=f_M(t)dt$. By taking into account (9) the further hypothesis is assumed $g_j(t) \in L^2_{d\psi}([0,\infty))$, where, as usual,

$$L^2_{d\psi}([0,\infty))=\left\{ g_j(t): \int_0^\infty g_j^2(t)d\psi(t)dt=\int_0^\infty g_j^2(t)f_M(t)dt<+\infty \right\}, \quad j=1,\ldots,M, M\geqslant 1$$

(thus $|G_M|$ represents the Gramian).

B. The moment space

$\{g_1(t),\ldots,g_M(t)\}$ being linearly independent on $[0,\infty)$ then the following:
(1) The Gramian is not singular (Ref. 6).
(2) The space V_M of moments, given by the convex hull generated by the points $\{g_1(t),\ldots,g_M(t)\}$, $t\in[0,\infty)$, has a nonempty interior (Ref. 7).
(3) If the sequence of prescribed moments $\{g_j\}$, $j=1,\ldots,M$ is an inner point of V_M then there are uncountably many probability measures $d\sigma(t)$ having such prescribed moments (Ref. 7).

Elsewhere if $\{g_1(t),\ldots,g_M(t)\}$ are linearly dependent then the following:
(1) The Gramian is singular.[6]
(2) The points $\{g_1(t),\ldots,g_M(t)\}$, $t\in[0,\infty)$ belong to the boundary of V_M (Ref. 7).
(3) If the sequence of prescribed moments $\{g_j\}$, $j=1,\ldots,M$ is a boundary point of V_M then there is a unique probability measure $d\sigma(t)$ having such prescribed moments. Such a measure is a step-function with a finite number of jumps (Ref. 7).

Remark 1: In the adopted procedure to obtain (8), as g_M varies continuously, both the measure $d\psi(t)=f_M(t)dt$ and the Hilbert space $L^2_{d\psi}([0,\infty))$ vary.

C. Domain $D^{(M)}_{g_j}$

Let us indicate with $D^{(M)}_{g_j}$ the domain of the acceptable values of g_j moment, when the first M moments are assigned. Considering $f_{M+1}(t)$ and assuming the value $\lambda_{M+1}=0$ then we came back to $f_M(t)$ where g_j, $j=1,\ldots,M$ are assigned. The following relation is therefore valid:

$$D^{(M)}_{g_j}\subseteq D^{(M+1)}_{g_j}, \quad j=1,\ldots,M. \tag{11}$$

D. An auxiliary result

Let $f_M(t)=\exp(-\Sigma_{j=0}^M\lambda_j g_j(t))$ given, to which correspond the moments $\{g_0,g_1,\ldots,g_M\}$, we will prove inductively the existence of the function

$$f^{(M)}(t)=\exp\left(-\sum_{j=0}^{M}\tilde{\lambda}_j g_j(t)-\sum_{j=M+1}^{M^*}\lambda_j^* g_j \right),$$

where $M\geqslant 1$, $M^*\geqslant M+1$ and λ_j^*, $j=M+1,\ldots,M^*$ are arbitrarily fixed and $\lambda^*_{M^*}>0$.

First of all we show that for $M=1$ there exists $f_1(t)$ and then $f^{(1)}(t)$.

(a) Existence of $f_1(t)$. $g_1>g_1^-$ assigned. To determine the admissible values, g_1 is varied continuously. Thus it results that $\lambda_j=\lambda_j(g_1)$, $j=0,1$. By differentiating both sides of (6) with respect to g_1, the following system of differential equations is obtained:

$$G_1\cdot\begin{bmatrix} d\lambda_0/dg_1 \\ d\lambda_1/dg_1 \end{bmatrix}=-\begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{12}$$

Equation (12) is solved starting from the following initial condition. Let us assign to λ_1 an arbitrary positive value then from the first equation of (6) we obtain $\tilde{\lambda}_0$ and from the second equation $\tilde{g}_1$. From (12) λ_1 is a monotonic decreasing function. The solution $\lambda_1=0$ is not allowed. Integrating (12) starting from the initial condition $\tilde{\lambda}_0$, λ_1 and $\tilde{g}_1$, previously found, we can reach the given g_1 value.

(b) Existence of $f^{(1)}(t)$. In a similar way, by considering (6) with $f_1(t)$ replaced by $f^{(1)}(t)$, differentiating with respect to g_1 we obtain again system (12) which we can integrate starting from a different initial condition $\hat{\lambda}_0$, λ_1 and $\hat{g}_1$ obtained analogously as done for $f_1(t)$.

To complete the induction we prove that if exists the function $f^{(M-1)}(t)$ with the moments $\{g_0,g_1,\ldots,g_{M-1}\}$ of $f_{M-1}(t)$ then exists the function $f^{(M)}(t)$ with the moments $\{g_0,g_1,\ldots,g_M\}$ of $f_M(t)$. From the existence of the function $f^{(M-1)}(t)$ we can obtain

$$\tilde{g}_M=\int_0^\infty g_M(t)f^{(M-1)}(t)dt.$$

Then exists $\tilde{f}^{(M)}(t)$ with $\tilde{\lambda}_0,\ldots,\tilde{\lambda}_{M-1},\tilde{\lambda}_M\equiv 0$, and λ_j^* fixed, which moments are $\{g_0,g_1,\ldots,g_{M-1},\tilde{g}_M\}$. Such a particular solution represents an initial condition to solve the system of differential equations obtained as follows. Let $\{g_0,g_1,\ldots,g_{M-1}\}$ be fixed and g_M vary continuously. By differentiating both sides of (6) with respect to g_M, where $f_M(t)$ is replaced by $f^{(M)}(t)$, then (8) is obtained. Such a system of differential equations is solved, starting from the initial condition $\tilde{f}^{(M)}(t)$ previously found, until the assigned value of g_M is reached. Thus $f^{(M)}(t)$ exists with the same moments $\{g_0,g_1,\ldots,g_{M-1},g_M\}$ of $f_M(t)$.

III. LOWER BOUND

Once assigned $\{g_1,\ldots,g_{M-1}\}$ we call g_M^- the lower bound of g_M in the moment space. We calculate g_M^- inductively. Let us suppose assigned $\{g_1,\ldots,g_{M-1}\}$ to which the function $f_{M-1}(t)$ corresponds. From $f_{M-1}(t)$ we calculate

$$\tilde{g}_M=\int_0^\infty g_M(t)f_{M-1}(t)dt,$$

where $\tilde{g}_M>g_M^-$. Next we solve (8), starting from the initial condition given by $\{g_1,\ldots,g_{M-1},\tilde{g}_M\}$. Then

$$\frac{d\lambda_M}{dg_M}=-\frac{|G_{M-1}|}{|G_M|}, \tag{13}$$

so that λ_M is a monotonic decreasing function.

Equation (8) is solved until $\lambda_M\to+\infty$. In such a case $|G_M|\to 0$, $|G_{M-1}|$ being bounded for each $g_M\leqslant\tilde{g}_M$ value. Then g_M assumes values greater than $\tilde{g}_M^-$, where $\tilde{g}_M^-$ is an opportune value, satisfying, in general, the relationship $\tilde{g}_M^->g_M^-$. When $g_M=\tilde{g}_M^-$, $|G_M|=0$, as the set of functions $\{g_1(t),\ldots,g_M(t)\}$ is linearly independent, consequently the point $\{g_1,\ldots,g_{M-1},\tilde{g}_M^-\}$ belongs to a boundary of the moment space [$d\psi(t)=f_M(t)dt$, is degenerated in a measure with a finite number of increasing points]. Thus we have

$$\tilde{g}_M^-=g_M^-, \tag{14}$$

where g_M^- depends on the assigned sequence $\{g_1,\ldots,g_{M-1}\}$.

Thus $g_M>g_M^-$, $M=1,2,\ldots$, is a necessary condition for the existence of M.E. distribution. It remains to prove $g_M>g_M^-$ is an even sufficient condition.

IV. UPPER BOUNDS

The following notation will be used:
$g_j^+=$upper bound of the moment g_j, $j=1,\ldots,M$.

A. $M=1$

In Sec. II D we proved that $f_1(t)$ exists $\forall g_1>g_1^-$. Then g_1 does not admit an upper bound. The admissible values of g_1 is then represented by $g_1>g_1^-$.

Given g_1, from (5) and (6) the following value $\tilde{g}_2$ is obtained:

$$\tilde{g}_2=\int_0^\infty g_2(t)f_1(t)dt, \tag{15}$$

and used in the next.

B. $M=2$

Let $g_1(>g_1^-)$ be fixed. We vary g_2 continuously to determine its admissible values. Thus it results that $\lambda_j=\lambda_j(g_2)$, $j=0,1,2$. By differentiating both sides of (6) with respect to g_2, the following system of differential equations is obtained:

$$G_2\cdot\begin{bmatrix} d\lambda_0/dg_2 \\ d\lambda_1/dg_2 \\ d\lambda_2/dg_2 \end{bmatrix}=-\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{16}$$

Thus λ_2 is a decreasing monotonic function that must be non-negative for the existence of the integrals (6). So, as $\lambda_2=0$ is an admissible solution which leads to $f_1(t)$, one deduces that an M.E. solution exists if $g_2\leqslant\tilde{g}_2=g_2^+$. Therefore (5),(6) admit a solution $f_2(t)$ if and only if g_2 satisfies the inequality

$$g_2^-<g_2\leqslant g_2^+. \tag{17}$$

By obtaining $f_2(t)$, the following value $\tilde{g}_3$ is obtained:

$$\tilde{g}_3=\int_0^\infty g_3(t)f_2(t)dt. \tag{18}$$

Before discussing the case $M=3$ we show a preliminary result.

For every $\lambda_k\geqslant 0$ we can construct the following auxiliary function:

$$f^{(1)}(t)=\exp(-\lambda_0-\lambda_1 g_1(t)-\lambda_k g_k(t)),\quad k>2, \tag{19}$$

with the following constraints:

$$\int_0^\infty g_j(t)f^{(1)}(t)dt=g_j,\quad j=0,1,2, \tag{20}$$

in which the moments g_j, $j=0,1$ are assigned and g_2 depends on λ_k [being $k>2$, (19) is not an M.E. solution].

Equation (20) admits the particular solution $f_1(t)$, which is of the form (19) with $\lambda_k=0$ (and such that g_0,g_1 are given and $g_2=\tilde{g}_2$).

Varying g_2 continuously, in its admissible range $g_2^-<g_2\leqslant g_2^+$, we obtain $\lambda_j=\lambda_j(g_2)$, $j=0,1,k$. Differentiating both sides of (20) we have

$$\begin{bmatrix} g_0 & g_1 & g_k \\ g_1 & g_{11} & g_{1k} \\ g_2 & g_{21} & g_{2k} \end{bmatrix}\cdot\begin{bmatrix} d\lambda_0/dg_2 \\ d\lambda_1/dg_2 \\ d\lambda_k/dg_2 \end{bmatrix}=-\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{21}$$

Thus taking into account (9) λ_k is a monotonic decreasing function as (20) admits the particular solution $f_1(t)$ with $\lambda_k=0$, $\lambda_1>0$, $g_2=g_2^+$ and as $g_2\to g_2^-$ then $\lambda_k\to+\infty$, from which we have the following relationship:

$$\begin{vmatrix} g_0 & g_1 & g_k \\ g_1 & g_{11} & g_{1k} \\ g_2 & g_{21} & g_{2k} \end{vmatrix}>0,\quad \forall k>2. \tag{22}$$

C. $M=3$

We would calculate the admissible domain of the moments g_2,g_3.

(a) Domain of g_2. Let us consider (19), with $k=3$. Let g_2^* be an assigned value of g_2 that satisfies (17) to which correspond three values $\lambda_0^*,\lambda_1^*,\lambda_3^*$ of $\lambda_0,\lambda_1,\lambda_3$, respectively.

For every $\lambda_3^*\geqslant 0$ we can construct the following auxiliary function:

$$f^{(2)}(t)=\exp(-\lambda_0-\lambda_1 g_1(t)-\lambda_2 g_2(t)-\lambda_3^* g_3(t)) \tag{23}$$

(where λ_3^* represents the previously fixed value), of which the moments g_j, $j=0,1,2$ are assigned. Thus (23) is not an M.E. function. Varying g_2 continuously, then $\lambda_j=\lambda_j(g_2)$, $j=0,1,2$ and differentiating both sides of (6) with $f_M(t)$ replaced by $f^{(2)}(t)$, (16) is still obtained. λ_2 is a monotonic decreasing function of g_2. By choosing as an initial condition for the integration of (16) the solution $\lambda_0^*,\lambda_1^*,\lambda_2^*=0,\lambda_3^*$, $g_2=g_2^*$ there exists, because of monotonicity of λ_2, values of $g_2>g_2^*$ in correspondence of which $\lambda_2<0$ holds. Therefore we have shown that, fixed $g_1(>g_1^-)$ and g_2 [verifying (17)], there exist functions of the form (23) with $\lambda_2<0$ having such moments.

Let us consider the auxiliary function,

$$f^{(3)}(t)=\exp(-\lambda_0-\lambda_1 g_1(t)-\lambda_2^* g_2(t)-\lambda_3 g_3(t)) \tag{24}$$

(with $\lambda_2=\lambda_2^*<0$ fixed) of which the moments g_j, $j=0,1$ are assigned [for each arbitrary value $\lambda_3=\lambda_3^*>0$ then $f^{(3)}(t)$ exists from the results of Sec. II D) (24) is not an M.E. function. Let us vary g_2 continuously; then the multipliers $\lambda_0,\lambda_1,\lambda_3$ will be functions of g_2. Differentiating both sides of (6) with respect to g_2, with $f_M(t)$ replaced by $f^{(3)}(t)$, (21) is obtained, with $k=3$. λ_3 results therefore, as λ_k in (21), a strictly positive monotonic decreasing function of g_2, being $\lambda_2=\lambda_2^*<0$. Therefore g_2 does not admit an upper bound and the domain of the admissible values of g_2 is

$$g_2>g_2^- . \tag{25}$$

(b) Domain of g_3. Let us consider (5) and let $g_2>g_2^-$, while g_3 varies continuously; then $\lambda_j=\lambda_j(g_3)$, $j=0,1,2,3$. Differentiating both sides of (6) with respect to g_3 we have the system

$$G_3\cdot\begin{bmatrix} d\lambda_0/dg_3 \\ \vdots \\ \vdots \\ d\lambda_3/dg_3 \end{bmatrix}=-\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{26}$$

Thus from (13) λ_3 is a monotonic decreasing function. To determine the possible upper bound of g_3 it is necessary to distinguish two subcases.

(i) $g_2\leqslant g_2^+$. In such a case (6) admits a solution of the form (5) with $\lambda_3=0$, a solution of the problem with $M=2$. Therefore for each $g_2\leqslant g_2^+$, g_3 admits an upper bound $g_3^+=g_3^*$, where g_3^* is provided by (18).

(ii) $g_2>g_2^+$. In such a case (6) does not admit the solution $\lambda_3=0$. In fact, in correspondence of that solution would bring us back to the case $M=2$, but with a value $g_2>g_2^+$, contradicting the existence of an upper bound when $M=2$. Excluding the hypothesis $\lambda_3=0$, λ_3 being monotonic decreasing, it follows that g_3 does not admit an upper bound.

Figure 1 illustrates in a qualitative manner the admissible values of the moments g_2,g_3.

Before discussing the case $M=4$ we show a preliminary result.

The auxiliary function is considered,

$$f^{(4)}(t)=\exp(-\lambda_0-\lambda_1 g_1(t)-\lambda_2 g_2(t)-\lambda_k g_k(t)),\quad k>3, \tag{27}$$

of which the moments g_j, $j=0,1,2$ are assigned [for each arbitrary value $\lambda_k=\lambda_k^*>0$ then $f^{(4)}(t)$ exists from the results of Sec. II D] [(27) is not of M.E., being $k>3$]. Let g_0,g_1,g_2 be fixed, while g_3 varies continuously. The multipliers in (27) are functions of g_3. Differentiating both sides of (6), with $f_M(t)$ replaced by $f^{(4)}(t)$, the system is obtained:

$$\begin{bmatrix} g_0 & g_1 & g_2 & g_k \\ g_1 & g_{11} & g_{12} & g_{1k} \\ g_2 & g_{21} & g_{22} & g_{2k} \\ g_3 & g_{31} & g_{32} & g_{3k} \end{bmatrix}\cdot\begin{bmatrix} d\lambda_0/dg_3 \\ d\lambda_1/dg_3 \\ d\lambda_2/dg_3 \\ d\lambda_k/dg_3 \end{bmatrix}=-\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{28}$$

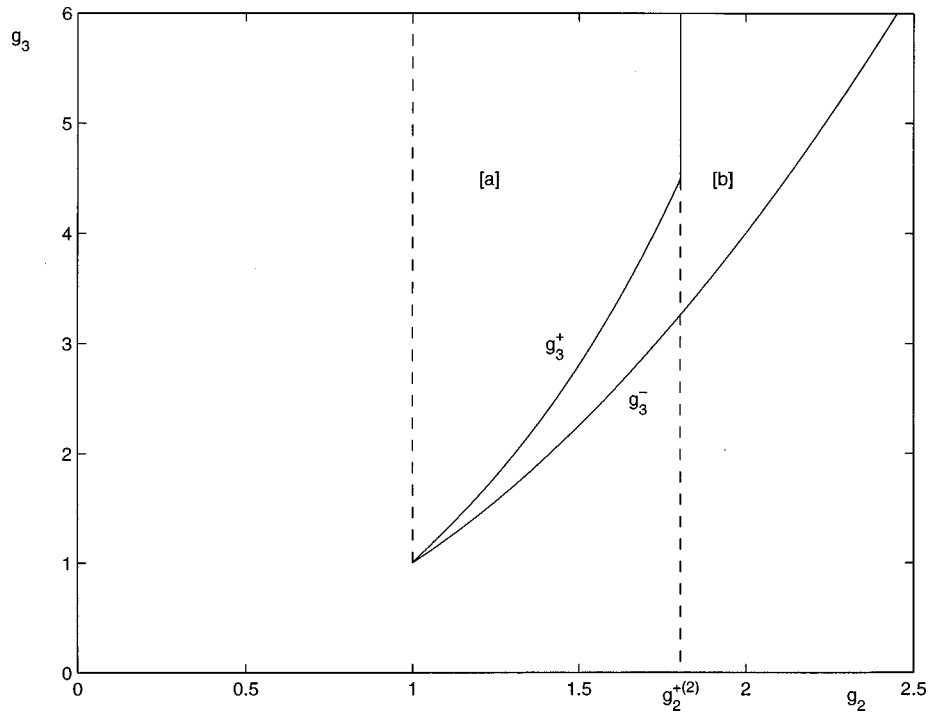

FIG. 1. $M=3$, admissible values of the moments g_2, g_3.

Thus λ_k is a monotonic function. The character of monotonicity of λ_k does not vary by varying g_2 as λ_k represents a family of disjointed curves. We suppose then $g_2 \leqslant g_2^+$, in such a case (6) admits the solution $\lambda_k=0$. Therefore λ_k is a monotonic decreasing for each value of g_2, from which the relationship,

$$\begin{vmatrix} g_0 & g_1 & g_2 & g_k \\ g_1 & g_{11} & g_{12} & g_{1k} \\ g_2 & g_{21} & g_{22} & g_{2k} \\ g_3 & g_{31} & g_{32} & g_{3k} \end{vmatrix} > 0, \quad k>3. \tag{29}$$

D. $M=4$

We wish to solve the domain of the admissible values of the moments g_3, g_4.

(a) *Domain of* g_3. Fixed $g_2>g_2^-$, $\lambda_3=\lambda_3^*<0$, varied g_3 continuously $\lambda_0,\lambda_1,\lambda_2,\lambda_4,g_4$ are functions of g_3. Differentiating both sides of (6) we have the system

$$\begin{bmatrix} g_0 & g_1 & g_2 & g_4 & 0 \\ g_1 & g_{11} & g_{12} & g_{14} & 0 \\ g_2 & g_{21} & g_{22} & g_{24} & 0 \\ g_3 & g_{31} & g_{32} & g_{34} & 0 \\ g_4 & g_{41} & g_{42} & g_{44} & 1 \end{bmatrix} \cdot \begin{bmatrix} d\lambda_0/dg_3 \\ d\lambda_1/dg_3 \\ d\lambda_2/dg_3 \\ d\lambda_4/dg_3 \\ dg_4/dg_3 \end{bmatrix} = - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \tag{30}$$

Thus λ_4 results monotonic decreasing with $\lambda_4>0$, the particular solution $\lambda_4=0$ being not admissible because of the condition $\lambda_3=\lambda_3^*<0$. Therefore, due to the arbitrarity of g_2 for each $g_2>g_2^-$, g_3 does not admit an upper bound.

(b) *Domain of* g_4. To underline possible upper bounds of g_4 it is necessary to show preliminarly that, having fixed an arbitrary couple (g_2,g_3), infinite solutions of the form (5) exist, characterized by large arbitrary values of λ_3.

(i1) Let $(g_2,g_3)\in[\mathrm{a}]$ (Fig. 1) a fixed couple and λ_3 vary continuously. $\lambda_0,\lambda_1,\lambda_2,\lambda_4,g_4$ are functions of λ_3. Differentiating (6) with respect to λ_3 we have the system

$$\begin{bmatrix} g_0 & g_1 & g_2 & g_4 & 0 \\ g_1 & g_{11} & g_{12} & g_{14} & 0 \\ g_2 & g_{21} & g_{22} & g_{24} & 0 \\ g_3 & g_{31} & g_{32} & g_{34} & 0 \\ g_4 & g_{41} & g_{42} & g_{44} & 1 \end{bmatrix} \cdot \begin{bmatrix} d\lambda_0/d\lambda_3 \\ d\lambda_1/d\lambda_3 \\ d\lambda_2/d\lambda_3 \\ d\lambda_4/d\lambda_3 \\ dg_4/d\lambda_3 \end{bmatrix} = - \begin{bmatrix} g_3 \\ g_{31} \\ g_{32} \\ g_{33} \\ g_{34} \end{bmatrix}. \tag{31}$$

Taking into account (29) λ_4 results in monotonic decreasing, while g_4 results in monotonic increasing. Nevertheless the solution $\lambda_4=0$ is not allowed as it would lead back to the case $M=3$, but in the presence of a couple (g_2,g_3) not belonging to the domain of the admissible values. Therefore between the infinite solutions obtained by varying λ_3 and corresponding to the same couple $(g_2,g_3)\in[\mathrm{a}]$, there exist some with large arbitrary values of λ_3.

(i2) Let $\lambda_k^{(M=3)}$, $k=0,1,2,3$ be the coefficients of the M.E. solution $f_3(t)$, in the case $M=3$, with $(g_2^{(b)},g_3^{(b)})\in[\mathrm{b}]$.

Let be $M=4$. $(g_2^{(a)},g_3^{(a)})\in[\mathrm{a}]$ a couple, with $g_3^{(a)}=g_3^{(b)}$. To the couple $(g_2^{(a)},g_3^{(a)})\in[\mathrm{a}]$, as previously shown, there are corresponding infinite solutions obtained by varying λ_3. We choose one of them characterized by

$$\lambda_3=\lambda_3^*>\lambda_3^{(M=3)}. \tag{32}$$

Let $g_3=g_3^{(a)}(=g_3^{(b)})$, $\lambda_3=\lambda_3^*>\lambda_3^{(M=3)}$ be fixed, while g_2 varies continuously in the interval $[g_2^{(a)},g_2^{(b)}]$. Then $\lambda_0,\lambda_1,\lambda_2,\lambda_4,g_4$ are functions of g_2. Differentiating both sides of (6) we have the following system:

$$\begin{bmatrix} g_0 & g_1 & g_2 & g_4 & 0 \\ g_1 & g_{11} & g_{12} & g_{14} & 0 \\ g_2 & g_{21} & g_{22} & g_{24} & 0 \\ g_3 & g_{31} & g_{32} & g_{34} & 0 \\ g_4 & g_{41} & g_{42} & g_{44} & 1 \end{bmatrix} \cdot \begin{bmatrix} d\lambda_0/dg_2 \\ d\lambda_1/dg_2 \\ d\lambda_2/dg_2 \\ d\lambda_4/dg_2 \\ dg_4/dg_2 \end{bmatrix} = - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{33}$$

By taking into account (22), (29) λ_4 is monotonic increasing. Besides, by increasing g_2, the couple $(g_2,g_3^{(a)})$, that initially belongs to the region [a], will belong subsequently to the region [b]. Therefore to each couple $(g_2,g_3^{(b)})\in[\mathrm{b}]$ corresponds to a solution with λ_3 given by (32). The achieved result is interpreted in the following way. When $M=4$ is also in the region [b], as in the [a], to each couple $(g_2,g_3^{(b)})\in[\mathrm{b}]$ correspond infinite solutions obtained by varying λ_3. A direct way of proof, analogous to that of the region $[a]$, is not immediate as in the region $[b]$, fixed a couple (g_2^*,g_3^*), varied λ_3 continuously, λ_4 is monotonic decreasing, but with solution $\lambda_4=0$ admissible, as to the solution it corresponds the value $\lambda_3^{(M=3)}$ of λ_3, not making evident the existence of solutions $\lambda_3>\lambda_3^{(M=3)}$.

Let $g_2=g_2^*$, $g_3=g_3^*$ be fixed within the the range of its admissible values. Let us vary g_4 continuously: then $\lambda_k=\lambda_k(g_4)$, $k=0,\ldots,4$. Differentiating both sides of (6) we have the system

$$G_4\cdot\begin{bmatrix} d\lambda_0/dg_4 \\ \vdots \\ \vdots \\ d\lambda_4/dg_4 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{34}$$

λ_4 is monotonic decreasing. To underline that there does not exist an upper bound of g_4 it is necessary to distinguish the two following cases concerning the domain of affiliation of the couple (g_2^*,g_3^*).

(i) $(g_2^*, g_3^*) \in [a]$. The solution $\lambda_4 = 0$ is not allowed as it would lead back to the case $M = 3$, but with a couple (g_2^*, g_3^*) not belonging to the domain of the admissible values of (g_2, g_3). Therefore, λ_4 being a monotonic decreasing and must be strictly positive, g_4 does not admit an upper bound.

(ii) $(g_2^*, g_3^*) \in [b]$. In such a case the solution $\lambda_4 = 0$ could be admissible, bringing us again to the case $M = 3$ and originating an upper bound of g_4. We wish, instead, to show the relationship $\lambda_4 > 0$. By choosing an initial condition for the solution of (34) verifying (32), the solution $\lambda_4 = 0$ is not allowed. In fact, if $\lambda_4 = 0$ we are led back to the case $M = 3$ but with a value $\lambda_3 > \lambda_3^{(M=3)}$, contradicting the hypothesis of unicity of the solution. From (31) g_4 being monotonic increasing function of λ_3, the moment g_4 does not admit an upper bound.

The relative analysis of the case $M = 4$ allows us to conclude that the moments g_3, g_4 does not admit an upper bound.

E. $M > 4$

The lack of upper bounds for the moments g_3, g_4 allows us to solve the general case $M > 4$, using for each M the results achieved in the case $M - 1$.

Let g_k, $k = 1, \ldots, M > 4$ be assigned.

(i) *Domain of* $g_1, \ldots, g_{M-1}$. By choosing $\lambda_M = 0$ one is led back to the case $M - 1$, where upper bounds of the moments $g_1, \ldots, g_{M-1}$ do not exist.

(ii) *Domain of* g_M. Let g_k, $k = 1, \ldots, M-1$ be assigned, $f_{M-1}(t)$ the corresponding density with multipliers that we will point out with $\lambda_k^{(M-1)}$, $k = 0, \ldots, M-1$.

By varying g_M, the family of functions $f_M(t)$ is considered. There exist between the infinite solutions that meet the moments g_k, $k = 1, \ldots, M-1$, also those characterized by large arbitrary values of λ_{M-1} [generalizing (31)]. We choose the one with $\lambda_{M-1} > \lambda_{M-1}^{(M-1)}$. Fixed g_k, $k = 0, \ldots, M-1$, g_M varied continuously then λ_k, $k = 0, \ldots, M$ are functions of g_M. Particularly, from the achieved system of differential equations (8), λ_M is monotonic decreasing, while λ_{M-1} is monotonic increasing. To solve such a system of differential equations one considers an initial condition characterized by $\lambda_{M-1} > \lambda_{M-1}^{(M-1)}$. An upper bound g_M does not exist as, in correspondence of such a bound, the solution $\lambda_M = 0$ holds. In such a case one is led back to the case $M - 1$, but with a value $\lambda_{M-1} > \lambda_{M-1}^{(M-1)}$, contradicting the unicity of the solution. Therefore g_M does not admit an upper bound.

V. CONCLUSIONS

Existence conditions of the generalized moment problem have been considered. From the achieved results it is deduced that, except the cases in which 2 or 3 moments are assigned, where unexpected upper bounds of the moments raise, in the remaining cases the necessary and sufficient conditions of existence of the M.E. solution are identical to ones of the moment problem.

[1] I. Csiszar, ''I-divergence geometry of probability distributions and minimization problem,'' Ann. Prob. **1**, 146–158 (1975).
[2] J. M. Borwein and A. S. Lewis, ''Convergence of best entropy estimates,'' SIAM J. Optimization **1**, 191–205 (1991).
[3] H. K. Kesavan and J. N. Kapur, *Entropy Optimization Principles with Applications* (Academic, London, 1992).
[4] A. Tagliani, ''On the application of maximum entropy to the moments problem,'' J. Math. Phys. **34**, 326–337 (1993).
[5] M. Frontini and A. Tagliani, ''Maximum entropy in the finite Stieltjes and Hamburger moment problem'' J. Math. Phys. **35**, 6748–6756 (1994).
[6] L. A. Lyusternik and A. R. Yanpol'skii, *Mathematical Analysis* (Pergamon, Oxford, 1965).
[7] M. G. Krein, ''The ideas of P. L. Cebysev and A. A. Markov in the theory of limiting values of integrals and their further development,'' Am. Math. Soc. Trans. **2**, 12 (1959).

On the use of a new integral theorem for the quantum mechanical treatment of electric circuits

W. Magnus and W. Schoenmaker
Advanced Semiconductor Processing Division, IMEC,
Kapeldreef 75, B-3001 Leuven, Belgium

(Received 16 March 1998; accepted for publication 12 August 1998)

We present a new integral theorem based on the well-known Gauss and Stokes theorems which provides a straightforward physical interpretation of the quantum mechanical energy balance equation governing the steady-state propagation of charge carriers through closed electric circuits. © *1998 American Institute of Physics.* [S0022-2488(98)03312-X]

I. INTRODUCTION

Recent developments in the field of mesoscopic physics and technology have lead to the fabrication of novel semiconductor structures and devices (quantum dots, quantum wires, quantum point contacts, single electron transistors) the working principles of which are directly based on generic quantum mechanical rules. Although thorough theoretical investigations of these new devices are being conducted on a world-wide scale, several fundamental features of quantum transport—such as the phenomenon of nonlocal resistance—are still not fully understood, especially when regarded from the side of nonequilibrium statistical mechanics. Contrary to the case of ''classical'' devices, the so-called active areas of quantum devices typically have sizes on the nanometer scale nanostructures, thereby allowing in principle for phase-coherent transport. This observation directly affects the strategy to be adopted when dynamical observables such as the electric current as a response to the external electric field are to be calculated: We believe that a reliable quantum mechanical calculation should encompass the entire, closed electric circuit without making any subdivision into spatial regions some of which could be taken into consideration for a semiclassical treatment. Our viewpoint is therefore completely opposed to the reservoir concept which considers the leads connecting the current source to the nanostructure[1,2] massive reservoirs where charge carriers have reached some local equilibrium and the presence of the current source is merely emulated by a chemical potential difference. On the other hand, within our approach a set of balance equations, based on the Heisenberg equations of motion obeyed by the charge field operators can be obtained to yield a dynamical characterization of a closed circuit containing one (or more) nanostructures or nanodevices.[3,4] One particular term arising in the energy balance equation and describing the electric power acquired by the charge carriers from the external field, can be given a straightforward interpretation for steady-state conditions (dc) with the help of a new integral theorem.

Section II illustrates the physical background which has led to the need of formulating a new integral theorem. The latter is proven in Sec. III and some concluding remarks are given in Sec. IV.

II. INTERPRETATION OF THE ENERGY BALANCE EQUATION

Representing a closed electric circuit by a toruslike region Ω confining the charge carriers but allowing exchange of energy with the environment, we have derived quantum mechanical momentum and energy balance equations describing the transport properties of the carrier gas residing in the interior of Ω. Taking the form of a rate equation, the energy balance equation reads:[3]

$$\frac{d\langle H_c\rangle_t}{dt} = \int_\Omega d\tau \langle \vec{J}(\vec{r})\rangle_t \cdot \vec{E}(\vec{r}) - \frac{i}{\hbar}\langle [H_c, H']\rangle_t, \tag{1}$$

 © 1998 American Institute of Physics

where H_c is the gauge invariant Hamiltonian of the carrier gas and H' represents the interaction of the carrier gas with all relevant scattering mechanisms (phonons, impurities, ...). $\vec{J}(\vec{r})$ denotes the electric current density operator whereas $\vec{E}(\vec{r})$ is the local, static electric field caused by an externally applied voltage in order to maintain a dc current through the circuit. The time dependence referred to by the subscript t in the statistical averages $\langle\ \rangle_t$ can in principle be obtained by combining the solution to the Heisenberg equations for all operators and the density matrix ρ_0 describing the circuit at the time the field was turned on, i.e.,

$$\langle X\rangle_t=\langle X(t)\rangle_0\equiv \mathrm{Tr}[\rho_0 X(t)] \quad \text{for } X=H_c,H',\vec{J}(\vec{r}). \tag{2}$$

Having attained a steady state for $t\to\infty$ the carrier gas may be characterized by a dc current and a stationary average energy. Hence the left-hand side of (1) vanishes and the energy balance equation reduces to

$$\int_\Omega d\tau\ \vec{j}(\vec{r})\cdot\vec{E}(\vec{r})=\frac{i}{\hbar}\lim_{t\to\infty}\langle[H_c,H']\rangle_t, \tag{3}$$

where $\vec{j}(\vec{r})\equiv\lim_{t\to\infty}\langle\vec{J}(\vec{r})\rangle_t$ denotes the current density vector.

Clearly, the left-hand side can be interpreted without further effort as the average power the carrier gas is extracting from the electric field in its steady state. However, the ultimate goal of circuit simulation programs is not to calculate the spatial distribution of the current *density* $\vec{j}(\vec{r})$ but rather to express the total current I as a function of measurable quantities such as the external voltage applied to the circuit. As is well-known from elementary electromagnetic theory,[5] the latter can be easily represented by the so-called electromotive force $\mathcal{E}$ which is nothing but the circulation of the electric field along a closed curve Γ lying in the circuit region:

$$\mathcal{E}=\oint_\Gamma \vec{E}\cdot d\vec{r}. \tag{4}$$

In general the circulation of a vector field depends on the curve Γ. In our case however, since no magnetic fields are involved, the third Maxwell equation leads to an irrotational electric field. At first sight, this conclusion seems to contradict the necessity of having a nonconservative field. However, an irrotational vector field needs not be conservative unless it is defined on a simply connected region, whereas the toruslike topology of an electric circuit implies only connectivity and the presence of at least one ''hole.'' Therefore, application of Stokes theorem to the circuit region Ω shows that the circulation of an irrotational vector field is merely independent of the choice of the circulation curve without having to be zero in the first place, thereby corroborating the intuitive result that an electromotive force may be considered a unique quantity for characterizing the strength of a nonconservative static electric field. The next step is to investigate how a stationary current I—unavoidably carried by a divergence-free current density vector—can go along with the concept of electromotive force and relate to power dissipation. More concretely, we would need a theorem from which it follows that the product $\mathcal{E}I$ equals the volume integral of $\vec{j}\cdot\vec{E}$ for any torus-shaped circuit. Such a theorem is provided by an integral theorem based on the familiar Gauss and Stokes theorems known from ordinary vector calculus.

III. INTEGRAL THEOREM

Proposition: Let Ω be a closed, multiply connected subset of $\mathbf{R}^3$ having one hole and being bounded by a surface $\partial\Omega$. If $\vec{J}$ and $\vec{E}$ are two differentiable vector fields defined on Ω, circulating around the hole and satisfying the conditions

$$\vec{\nabla}\cdot\vec{J}=0,$$

$$\vec{\nabla}\times\vec{E}=\vec{0}, \tag{5}$$

$$\vec{J}\parallel\partial\Omega \ \text{ or } \ \vec{J}=\vec{0} \ \textit{in each point of } \partial\Omega,$$

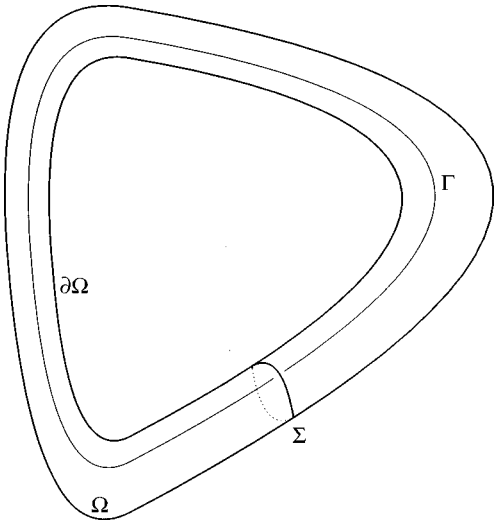

FIG. 1. Topology of the connected region Ω.

then

$$\int_{\Omega} \vec{J}\cdot\vec{E}\, d\tau = \left[\int_{\Sigma} \vec{J}\cdot d\vec{S}\right]\left[\oint_{\Gamma} \vec{E}\cdot d\vec{r}\right], \tag{6}$$

where Σ *is an arbitrary cross section, intersecting* Ω *only once and* Γ *is a simple closed curve, encircling the hole and lying within* Ω *but not intersecting* $\partial\Omega$ (see Fig. 1). The orientation of Σ is uniquely determined by the orientation of Γ.

Proof: Define curvilinear coordinates (u^1,u^2,u^3) and a corresponding basis $(\vec{a}_1,\vec{a}_2,\vec{a}_3)$ such that the coordinate curves $du^1=0$, $du^2=0$ coincide with the closed field lines of $\vec{J}$. Then, by definition, the field vector $\vec{J}$ has only a component along the third basis vector $\vec{a}_3$:

$$\vec{J}=J^k\vec{a}_k=J^3\vec{a}_3. \tag{7}$$

According to the last condition of (5), the field lines of $\vec{J}$ are nowhere intersecting the boundary surface $\partial\Omega$. Consequently, the particular choice of the curvilinear coordinates leads to the following observation:

(1) The outer surface $\partial\Omega$ is spanned by a subset of the curves $du^1=0$, $du^2=0$ and therefore, it must correspond to either $du^1=0$ or $du^2=0$.
(2) All internal points of Ω are addressed when u^1, u^2, and u^3 are bound to vary within three independent intervals A_1, A_2, and A_3.

Then, denoting the Jacobian determinant by $g(u^1,u^2,u^3)$ and using (7), the volume integral may be written

$$\int_{\Omega} \vec{J}\cdot\vec{E}\, d\tau = \int_{A_1} du^1 \int_{A_2} du^2 \int_{A_3} du^3\, g(u^1,u^2,u^3)\, J^3(u^1,u^2,u^3)\, E_3(u^1,u^2,u^3). \tag{8}$$

On the other hand, the divergence of $\vec{J}$ generally reads

$$\vec{\nabla}\cdot\vec{J}=\frac{1}{g}\frac{\partial}{\partial u^k}(gJ^k) \tag{9}$$

which, due to (7), reduces to

$$\vec{\nabla}\cdot\vec{J}=\frac{1}{g}\frac{\partial}{\partial u^3}(gJ^3). \tag{10}$$

In other words, the proposition $\vec{\nabla}\cdot\vec{J}=0$ implies that $K\equiv gJ^3$ is independent on u^3. Consequently,

$$\int_\Omega \vec{J}\cdot\vec{E}\; d\tau=\int_{A_1} du^1\int_{A_2} du^2 K(u^1,u^2)\int_{A_3} du^3 E_3(u^1,u^2,u^3). \tag{11}$$

For each fixed pair $(u^1,u^2)\in(A_1,A_2)$ we may define a coordinate curve $\Gamma(u^1,u^2)$, lying within Ω and being parametrized by u^3. The tangent vector along $\Gamma(u^1,u^2)$ is then given by

$$d\vec{r}=du^3\;\vec{a}_3, \tag{12}$$

whence

$$\int_{A_3} du^3\; E_3(u^1,u^2,u^3)=\oint_{\Gamma(u^1,u^2)}\vec{E}\cdot d\vec{r}. \tag{13}$$

Since $\vec{E}$ is irrotational, according to Stokes' theorem its circulation does not depend on the particular choice of the circulation curve as was already discussed in more detail in Sec. II. Consequently, $\Gamma(u^1,u^2)$ may be replaced by any other closed curve Γ encircling the hole region and sharing the traversal sense with $\Gamma(u^1,u^2)$:

$$\int_\Omega \vec{J}\cdot\vec{E}\; d\tau=\left[\oint_\Gamma \vec{E}\cdot d\vec{r}\right]\left[\int_{A_1} du^1\int_{A_2} du^2\; K(u^1,u^2)\right]. \tag{14}$$

Now, let u_0^3 be any number in the interval A_3 and let Σ be the piece of the surface $u^3=u_0^3$, cut off by the boundary surface $\partial\Omega$. Then we may define a vector surface element $d\vec{S}$ on Σ as follows:

$$d\vec{S}=\pm\vec{a}_1\times\vec{a}_2 du^1 du^2 \quad \text{for } (u^1,u^2)\in(A_1,A_2) \text{ and } u^3=u_0^3. \tag{15}$$

The plus sign corresponds with the correct orientation. Indeed, remembering that $d\vec{r}=du^3\;\vec{a}_3$ defines a traversal sense for the curve $\Gamma(u^1,u^2)$ intersecting Σ in the point (u^1,u^2,u_0^3) and that

$$g=(\vec{a}_1\times\vec{a}_2)\cdot\vec{a}_3 \quad \text{for } (u^1,u^2)\in(A_1,A_2) \text{ and } u^3=u_0^3. \tag{16}$$

We see that

$$d\vec{S}\cdot d\vec{r}=g du^1\; du^2\; du^3=d\tau \tag{17}$$

or that—for a positive volume element $d\tau$—$d\vec{S}$ and $d\vec{r}$ are pointing to the same side of Σ. Moreover, from (14) and (15) it becomes clear that

$$\int_{A_1} du^1\int_{A_2} du^2\; K(u^1,u^2)=\int_\Sigma \vec{J}\cdot d\vec{S}. \tag{18}$$

Finally, it should be noticed that the surface integral of $\vec{J}$ remains unchanged when Σ is replaced by another cross section Σ_1 which does not correspond to $du^3=0$. In such a case, one may always construct a Gaussian pill box from Σ, Σ_1 and the enclosed part of the boundary surface $\partial\Omega$. Applying Gauss' theorem to the pill box and bearing in mind that $\vec{\nabla}\cdot\vec{J}=0$, we conclude that the integral of $\vec{J}$ over the surface of the pill box vanishes, and since $\vec{J}$ itself is tangential along the surface $\partial\Omega$, it does not contibute to the surface integral so that

$$\int_\Sigma \vec{J}\cdot d\vec{S}=\int_{\Sigma_1}\vec{J}\cdot d\vec{S}_1 \tag{19}$$

for equal orientations of $d\vec{S}$ and $d\vec{S}_1$.
This completes the proof.

IV. DISCUSSION

Apart from the possibility of having other proofs and perhaps even more general formulations of the presented theorem, it should be noted that the present proof requires the field lines of $\vec{J}$ to be closed. Indeed, the theorem first envisages the circulation of $\vec{E}$ around a *coordinate curve* $\Gamma(u^1,u^2)$ before addressing the generalization to more general curves Γ, where $\Gamma(u^1,u^2)$ is by definition an arbitrary field line of $\vec{J}$. Strictly speaking from the point of view of mathematical rigor, the condition $\nabla\cdot\vec{J}=0$ does not generally lead to closed field lines for $\vec{J}$, and therefore the addition of the latter as a supplementary proposition may be required. On the other hand, we believe that in most of the physical applications for which the theorem has been formulated, the solenoidal character of the vector field $\vec{J}$ together with the condition that $\vec{J}$ ''follows'' the geometry of the finite and bounded region Ω inevitably leaves us with a pattern of closed field lines, at least if pathological field distributions are to be avoided.

V. CONCLUSION

The above discussed integral theorem provides a transparent physical interpretation of the quantum mechanical transport equation balancing power supply and power dissipation in a closed electric circuit. Although the evaluation of the scattering term may be quite elaborate, depending on the approximation scheme adopted, the theorem at least facilitates the expression of the transport equations in terms of macroscopic observables:

$$I\mathcal{E}=\frac{i}{\hbar}\lim_{t\leftrightarrow\infty}\langle[H_c,H']\rangle_t. \tag{20}$$

Furthermore, it indirectly supports the thesis that the electromotive force is the relevant parameter characterizing the strength of the applied electric field rather than the electrostatic or chemical potential differences existing between contact reservoirs.

ACKNOWLEDGMENT

We are indebted to Peter Bueken for critical reading of the manuscript.

[1] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge, 1995).
[2] H. U. Baranger and A. D. Stone, Phys. Rev. B **40**, 8169 (1989).
[3] W. Magnus and W. Schoenmaker (unpublished).
[4] W. Magnus and W. Schoenmaker, ''Fundamentals of dissipation in mesoscopic structures,'' presented at the first workshop of the ''Canadian European Research Initiative on Nanostructures'' (CERION98), Neuchâtel, 29 September 1998.
[5] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).

Solving Poisson's equation with interior conditions

J. E. M^{c}Carthy
Department of Mathematics, Washington University, St. Louis, Missouri 63130

E. Yu. Backhaus and J. Fajans
Department of Physics, University of California—Berkeley, Berkeley, California 94720-7300

(Received 8 August 1997; accepted for publication 11 August 1998)

We consider the problem of extending the solution of a particular two-dimensional Poisson equation to a larger domain. This problem is related to the problem of putting a non-neutral plasma into equilibrium by applying a suitable wall potential, and to similar problems in two-dimensional fluid dynamics. While one cannot always find an exact solution, one can always find an approximate solution if the plasma has no holes. © *1998 American Institute of Physics.* [S0022-2488(98)02312-3]

I. INTRODUCTION

We consider the Poisson equation

$$\nabla^2\Phi(z)=\begin{cases}-4\pi: z\in G,\\ 0: z\in D\backslash G,\end{cases}$$
$$\Phi|_{\partial G}=0, \tag{1}$$

where G is a planar domain, and D is a disk containing G. The unusual feature of Eq. (1) is that, while Φ is a function on D and its Laplacian there is given, the ''boundary condition'' is given on the interior curve ∂G, the boundary of G. It is routine to solve Eq. (1) in G, so one can view the problem as asking when a solution can be found that extends harmonically to the larger set D.

Indeed, the function

$$v(z)=-2\int_G \log|z-\zeta|\,dA(\zeta) \tag{2}$$

satisfies

$$\nabla^2 v(z)=\begin{cases}-4\pi: z\in G,\\ 0: z\in D\backslash G.\end{cases}$$

(Throughout the paper we shall use complex notation; $z=x+iy$ and $\zeta=\eta+i\xi$ will represent points in the plane, thought of as complex numbers. We write area integrals as $\int dA(\zeta)$ rather than $\int\int d\eta\, d\xi$.)

So the question becomes whether one can find a function u that is harmonic in all of D, and satisfies the ''interior condition'' $u=-v$ on ∂G, as then $\Phi=u+v$ will solve Eq. (1).

This problem originates in the study of equilibria of non-neutral plasmas.[1–3] Consider a flat-top (constant density) charged plasma that forms a (not necessarily symmetric) infinite cylinder. The plasma is enclosed by a circular wall, on which we can place an arbitrary potential, and is permeated by a uniform axial magnetic field **B**. Because the plasma and boundary are assumed to form infinite cylinders, this is really a two-dimensional system[4] in which the planar cross section of the plasma is G and the wall is ∂D. The charged plasma particles follow $\mathbf{E}\times\mathbf{B}$ orbits, where **E** is the sum of the field induced by the plasma itself and the field coming from the wall.

If we assume unity charge density, the plasma potential is given by Eq. (2). The wall potential will be some harmonic function u in D. Then the total electric field **E** will be the gradient of

 © 1998 American Institute of Physics

$u+v$. As the plasma particles travel perpendicularly to **E** and to **B**, they stay in the plane. The direction perpendicular to **E** defines the equipotentials of $u+v$; consequently the particles travel along these equipotentials. The plasma will be in equilibrium when the plasma boundary ∂G is an equipotential. Hence we are led to Eq. (1).

It was shown in Refs. 2 and 3 that most non-neutral plasmas without holes can be put into equilibrium, at least approximately, by the application of a suitable wall potential. This paper explores the mathematics behind solving this problem. Two-dimensional fluid systems share similar considerations; we hope that our results may have applications to other areas of physics.

In Sec. II, we discuss the exact solution for a class of shapes G, namely, polynomial images of the disk. We solve this problem completely, and show that the plasma can be put into an exact equilibrium if and only if the wall lies inside certain critical points. In Sec. III we show that for almost any plasma without holes, we can find an approximate equilibrium, in the sense that for an appropriate potential on the wall, the sum of this potential and the potential of the plasma itself has an equipotential that surrounds the plasma and is arbitrarily close to it.

We do not discuss the issue of stability of these plasma equilibria. Stability is considered in Refs. 2 and 5.

II. EXACT RESULTS

A. Mathematical setup

Let G be the planar cross section of the plasma, and let D be a disk containing G. We wish to place a potential on ∂D (the boundary of D) such that the sum of this potential and the potential from G itself is constant on ∂G. For simplicity, we shall assume that the charge density on G is 1.

The potential induced by the plasma itself is the function v from Eq. (2). The potential placed on ∂D will extend to a bounded harmonic function u on D. So we wish to find a function u, harmonic on D, that agrees with $-v$ on ∂G. [Note that the potential from the image charge, normally incorporated into v by replacing the integrand in Eq. (2) by the Green's function that vanishes on ∂D, has been incorporated into u instead.]

Thus we wish to solve the problem

$$\begin{aligned} \nabla^2 u &= 0 \quad \text{on } G, \\ u+v &= 0 \quad \text{on } \partial G. \end{aligned} \tag{3}$$

We know u on ∂G; it is uniquely determined inside G as the solution of the Dirichlet problem with those boundary values. Let $\phi: G \to \mathbf{D}$ be a Riemann map, that is, an analytic univalent function that maps G onto the unit disk **D** (such a map always exists for a simply connected domain G—see Ref. 6). Consider the function

$$\omega(\zeta,z) = \log \left| \frac{(z-\zeta)(1-\overline{\phi(\zeta)}\phi(z))}{\phi(z)-\phi(\zeta)} \right|. \tag{4}$$

The logarithm of the modulus of a nonvanishing analytic function is harmonic, so the function $\omega(\zeta,\cdot)$ is harmonic on G for any fixed ζ in G. Moreover, if ζ is in G and z is in ∂G, then $\omega(\zeta,z) = \log|z-\zeta|$ because $|\phi(z)| = 1$. So, inside G the function u is given by

$$u(z) = 2\int_G \log \left| \frac{(z-\zeta)(1-\overline{\phi(\zeta)}\phi(z))}{\phi(z)-\phi(\zeta)} \right| dA(\zeta). \tag{5}$$

[This function is harmonic, by differentiating under the integral sign, and has the right boundary values, by comparison with Eq. (2).]

There are two problems with extending the formula in Eq. (5) to values of z outside of G. The first is extending the function ϕ. The second is that, as soon as the point z is outside G, the function $\log|1-\overline{\phi(\zeta)}\phi(z)|$ will have a singularity for some ζ in G, and so Eq. (5) may no longer represent a harmonic function.

If the boundary of G is an analytic curve, then ϕ will extend to an analytic function on a neighborhood of ∂G. Moreover, the integral in Eq. (5) can be replaced by an integral (or distribution) over a compact subset of G, so that the new formula replacing Eq. (5) will represent a function that equals u on G, and is harmonic on a larger set containing G. Rather than explaining how to do this in general, using the theory of quadrature domains in the wide sense discussed in Ref. 7, it is sufficient for our purposes to discuss only the case where ϕ is the inverse of a polynomial mapping.

B. Polynomial images of disks

Let p be a polynomial of degree N that is univalent (one-to-one) on $\mathbf{D}$, and let G be the image of $\mathbf{D}$ under p. Assume $p(0)=0$. Then G is a *quadrature domain*, which means there are constants $a_0,a_1,\ldots,a_{N-1}$ and $b_1,\ldots,b_{N-1}$ such that, for any function h that is harmonic and integrable on G, the following quadrature formula holds:

$$\int_G h(z)\ dA(z)=a_0h(0)+\sum_{n=1}^{N-1} a_n\left(\frac{\partial^n h}{\partial x^n}\right)_{z=0}+b_n\left(\frac{\partial^n h}{\partial x^{n-1}\partial y}\right)_{z=0}. \tag{6}$$

If p is of degree 1, then G is a disk and Eq. (6) is the familiar mean-value property for harmonic functions. For how to calculate the coefficients a_n and b_n, see Appendix A.

The integrand in Eq. (5) is $\omega(\zeta,z)$, and we are integrating with respect to ζ. But for every fixed z in G, the function $\omega(\zeta,z)$ is harmonic in ζ for ζ in G, because Eq. (4) can be written as

$$\omega(\zeta,z)=\log\left|\frac{(z-\zeta)(1-\phi(\zeta)\overline{\phi(z)})}{\phi(z)-\phi(\zeta)}\right|,$$

i.e., as the logarithm of the modulus of a nonvanishing function that is analytic in ζ. Therefore, by combining Eqs. (5) and (6), we get that, for z in G,

$$u(z)=2\left[a_0\omega(0,z)+\sum_{n=1}^{N-1} a_n\left(\frac{\partial^n\omega(\zeta,z)}{\partial\eta^n}\right)_{\zeta=0}+b_n\left(\frac{\partial^n\omega(\zeta,z)}{\partial\eta^{n-1}\partial\xi}\right)_{\zeta=0}\right]. \tag{7}$$

The beauty of Eq. (7) is that, for any z *outside* G, as long as $\phi(z)$ is analytic and nonzero, the right-hand side of Eq. (7) is a harmonic function. As it agrees with u on G, this must give the harmonic extension of u to a domain containing G.

Now in our current case, ϕ is the inverse of the polynomial p, and maps G to D. Then ϕ will have branch points at the (at most) $N-1$ points outside G that it maps to a critical point of p (a point where p' vanishes). Choose nonintersecting branch cuts connecting the branch points of ϕ to ∞, and let Ω denote the plane minus these branch cuts. Then ϕ is a single-valued analytic function on Ω, and ϕ takes the value 0 only at 0. With this choice of ϕ, Eq. (7) gives the unique extension of u to Ω (unique because two harmonic functions on a connected set cannot agree on an open subset without agreeing everywhere).

A different choice of branch cuts could give rise to a different domain, Ω' say, and a different harmonic function u' on Ω'. But u and u' would have to agree on the connected component of $\Omega\cap\Omega'$ containing G.

It was shown by Ebenfelt[8,9] that one cannot continue u harmonically across the branch points. This can be seen by considering the behavior of ∇u near a branch point. By differentiating Eq. (7) one gets various terms in powers of $\phi(z)$, which remain bounded, and also terms in negative powers of z, which are also bounded (as we are only concerned with branch points outside G, and 0 is in G), and a term $-a_0\phi'(z)/\phi(z)$ which results from taking the $\partial/\partial z$ derivative of $2a_0\log|1/\phi(z)|$. As a_0 is never zero [indeed, it follows from Eq. (6) that a_0 is the area of G] and $\phi'(z)$ tends to ∞ at a branch point [because $p(\phi(z))=z$, so by the chain rule $\phi'(z)=1/p'(\phi(z))$, and $p'(\phi(z))$ tends to zero], it follows that ∇u tends to infinity as one tends to a branch point, so every branch point is a singularity of u.

We conclude the following: *It is impossible to extend u to be harmonic on any domain containing a branch point of ϕ, but this is the only obstruction. Therefore one can find a potential on a wall surrounding G that puts the plasma into equilibrium if and only if no branch points of ϕ lie within the wall.*

C. An example: Cardioids

The simplest example is when p is a second degree polynomial. Let β be a real number greater than or equal to 1, and let $p(z)=z^2+2\beta z$. The corresponding shapes are shown in Fig. 1 for different values of β.

One can calculate directly that $\phi(z)=-\beta+\sqrt{\beta^2+z}$. From Appendix A it follows that Eq. (6) becomes

$$\int_G h(z)\ dA(z)=2\pi(1+2\beta^2)h(0)+4\pi\beta^2\ \frac{\partial h}{\partial x}\ (0). \tag{8}$$

Therefore Eq. (7) becomes

$$u(z)=4\pi(1+2\beta^2)\ \log\left|\frac{z}{\phi(z)}\right|+4\pi\beta\ \mathrm{Re}\left[-\frac{2\beta}{z}+\frac{1}{\phi(z)}-\phi(z)\right]. \tag{9}$$

The function ϕ has a branch point at $-\beta^2$. Choose a branch of ϕ—the most obvious is to cut the plane along the negative real axis from $-\infty$ to $-\beta^2$. Then Eq. (9) defines a harmonic function on $\mathbf{C}\backslash(-\infty,-\beta^2]$ that agrees with $-v$ on the boundary of the plasma; but the function cannot be extended to be harmonic in any neighborhood of the point $-\beta^2$.

We conclude, therefore, that if the disk D does not contain $-\beta^2$ in its interior, then we can place a potential on ∂D that puts G into equilibrium. But if D does contain $-\beta^2$ in its interior, then it is impossible to put the plasma G into equilibrium by just putting a potential on ∂D.

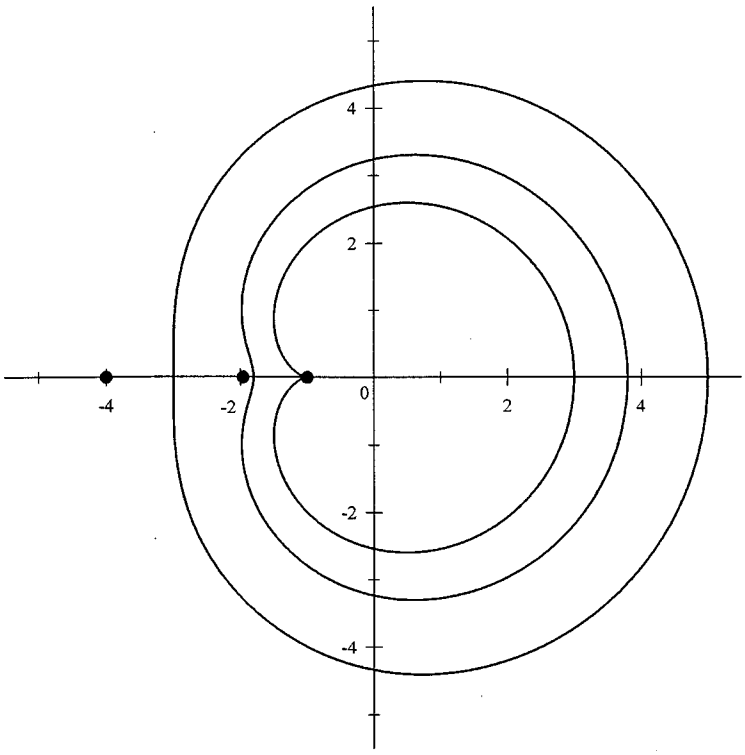

FIG. 1. Cardioids with β=2, 1.4, and 1. The corresponding critical points, indicated by the dots, are at -4, -1.96, and -1.

As β tends to 1, it becomes more and more difficult to put the plasma into equilibrium, because the branch point $-\beta^2$ gets closer and closer to the plasma (see Fig. 1). In particular, if $\beta<\sqrt{2}$, the branch point lies inside the convex hull of the plasma, so no disk D that surrounds the plasma can have the branch point outside it. (If one allowed a noncircular wall, it could surround the plasma but not the branch point, except in the limiting case $\beta=1$ when the branch point is on the boundary of the plasma.)

III. APPROXIMATE RESULTS

A. General theory

Despite the conclusion in Sec. II B above, it is nonetheless possible to put any simply connected plasma that does not have a pathological boundary into an *approximate* equilibrium, in the sense that one can distort the plasma by an arbitrarily small amount and then find a wall potential so that the sum of the wall potential and the potential from the plasma has an equipotential that is very close to the boundary of the plasma. (By very close we mean in the strong sense that the Hausdorff distance between the new plasma and the old plasma is very small, i.e., every point in one set is very close to some point in the other set.)

The technical condition we must impose on G is that its boundary be *locally connected*, which means the boundary is the continuous image of a circle. This implies that the inverse of the Riemann map, the analytic, univalent function p that maps $\mathbf{D}$ to G, extends to be continuous on the closure of $\mathbf{D}$.[10] It can therefore be uniformly approximated by polynomials, and the image of $\mathbf{D}$ under one of these approximating polynomials will be very close to G. Thus, by distorting the plasma by an arbitrarily small amount, we can assume that the map p is a polynomial, and that we are in the situation discussed in Sec. II B. Moreover, by another arbitrarily small perturbation if necessary, we can assume that the derivative of p does not vanish on ∂D, so ϕ has no branch points on ∂G.

We know that the exact solution is the function u from Eq. (7), but that this has singularities at the branch points of ϕ. Define G_ε to be the set of points that are within ε of G, and choose ε to be positive but small enough so that $G_{3\varepsilon}$ contains no branch points of ϕ, and so that the only point in $G_{3\varepsilon}$ where ϕ attains the value 0 is 0. The function u is harmonic on $G_{3\varepsilon}$ and gives the exact solution there.

Suppose one could find a function ψ, analytic on D and continuous on the closure of D, and very close to ϕ on $G_{2\varepsilon}$. Moreover, suppose $\psi(0)=\phi(0)$, $\psi'(0)=\phi'(0),\ldots,\psi^{(N-1)}(0)=\phi^{(N-1)}(0)$, and that ψ only takes the value 0 at 0. Then substitute ψ for ϕ in Eq. (4), and plug this into Eq. (7) to get a function $w(z)$ that is harmonic on D and very close to u on G_ε.

Now consider the set $H=\{z:v(z)+w(z)=0\}$. This will be a small distortion of the set $\{z:v(z)+u(z)=0\}$. The only way H could fail to contain a closed curve that is close to ∂G would be if the set $\{z:v(z)+u(z)=0\}$ contained, in addition to ∂G, a path connecting ∂G to ∂D [for then a small perturbation of the zero set of $u+v$ might no longer enclose (a small perturbation of) G]. But this cannot happen, as is shown in Appendix B. Therefore w restricted to the boundary of D will give a wall potential that produces an approximate solution.

How does one find ψ? Consider the function $\log(\phi(z)/z)$. This is analytic on $G_{3\varepsilon}$, so it can be uniformly approximated on $G_{2\varepsilon}$ by a polynomial q (by Runge's theorem[6]). Moreover, one can modify q slightly, if necessary, to ensure that the values of $q(0),\ldots,q^{(N-1)}(0)$ match those of $\log(\phi(z)/z)$ exactly. Then $\psi(z)=z\exp(q(z))$ will be analytic on all of $\mathbf{C}$, take the value 0 only at 0, be uniformly close to ϕ on $G_{2\varepsilon}$, and the first $N-1$ derivatives of ψ at 0 will match those of ϕ.

B. An example: Cardioids

Consider first the cardioid in Fig. 1, with $\beta=2$. As the branch point of ϕ lies outside the convex hull of G, we can approximate ϕ by a power series, centered at 1, say. Suppose we calculate the seventh-order Taylor polynomial for $\log(\phi(z)/z)$ about 1.

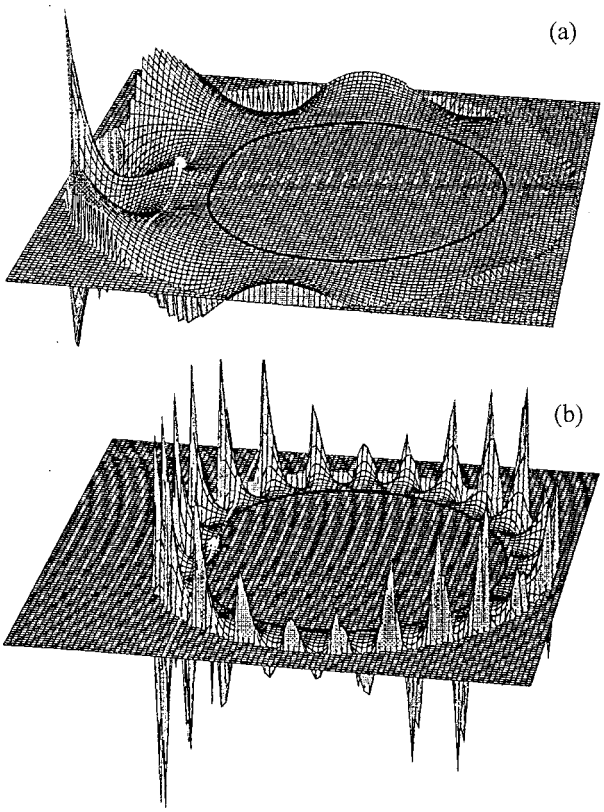

FIG. 2. (a) The difference between the exact solution $u(z;\beta=2)$ [Eq. (9)] and the approximate solution $w(z)$ [Eq. (10)], plotted out to a radius of 6.5. (b) The same difference for the $\beta=1.4$ cardioid, plotted out to a radius one unit larger than the radius of the cardioid. The outline of the appropriate cardioid is indicated by the black line. The corresponding critical points are indicated by the white dots; no exact solution valid in the cardioid interior can be analytic to the left of the white lines. Note that $u(z)$ and $w(z)$ agree closely inside the cardioids. There are more spikes in (b) that in (a) because of the increased number of terms used in (b).

We obtain

$$q_7(z) = -1.4436354 - 0.052786404(z-1) + 4.0325224\ 10^{-3}(z-1)^2 - 4.5228033\ 10^{-4}(z-1)^3 + 5.9701757\ 10^{-5}(z-1)^4 - 8.6302161\ 10^{-6}(z-1)^5 + 1.3221683\ 10^{-6}(z-1)^6 - 2.1090911\ 10^{-7}(z-1)^7.$$

Let

$$\psi(z) = z e^{q_7(z)}.$$

Then we obtain

$$\sup_{z \in G} |\psi(z) - \phi(z)| \leqslant 0.008.$$

So, if on any wall surrounding G we put the potential

$$w(z) = 36\pi \log \left| \frac{z}{\psi(z)} \right| + 8\pi \operatorname{Re} \left[-\frac{4}{z} + \frac{1}{\psi(z)} - \psi(z) \right], \tag{10}$$

then the equipotential $\{v + w = 0\}$ will be very close to the boundary of G. The difference between this approximate solution w and the exact, but divergent, solution u is plotted in Fig. 2(a). We can

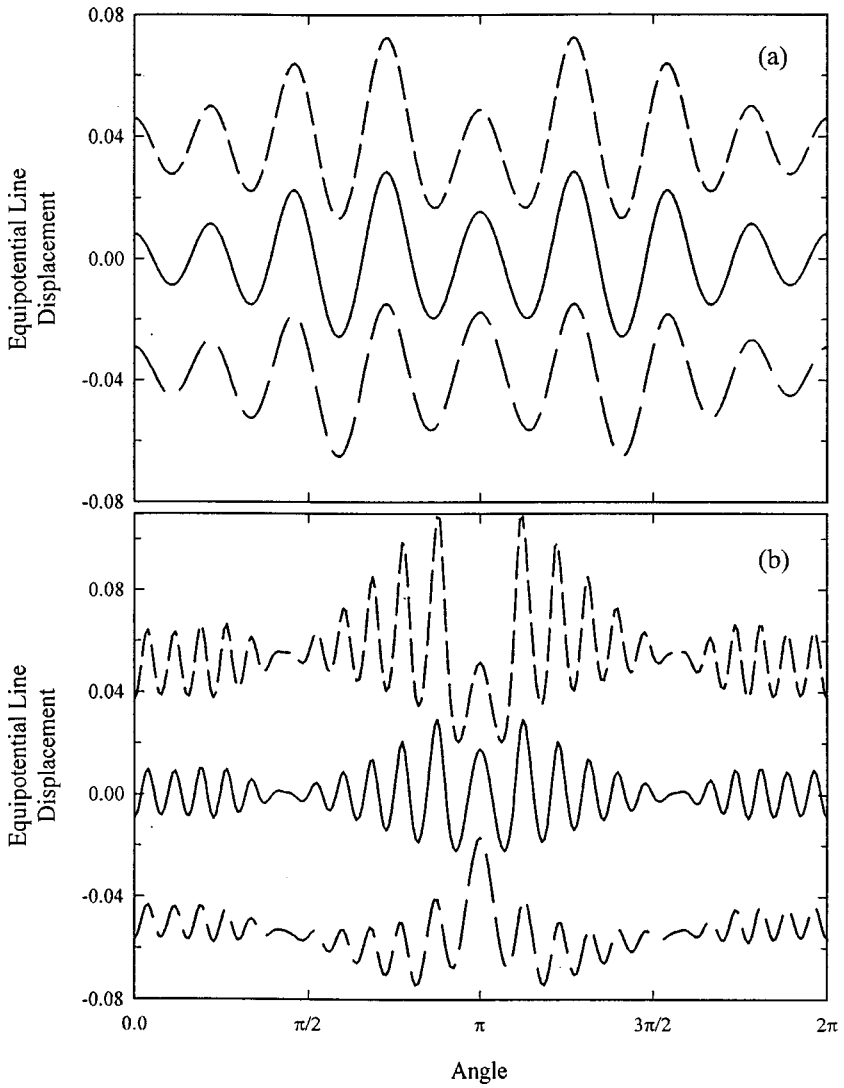

FIG. 3. (a) The distance of the $+1$ (the long dashed curve), the 0 (the solid curve), and the -1 (the short dashed curve) equipotentials of $v+w$ from the $\beta=2$ cardioid boundary, plotted as a function of angle around the cardioid. (b) The equivalent curves for the $\beta=1.4$ cardioid.

calculate that on ∂G, the function $v+w$ takes values between -0.66 and $+0.66$ (for reference, $v+w$ equals 53.4 at $r=0$). Figure 3(a) shows that the equipotential $v+w=0$ is very close to ∂G, and that the equipotentials $+1$ and -1 surround ∂G.

Note 1: The function w does not depend on the size of the disk D surrounding G. The size of the applied potential w on the wall of D does go up somewhat as D grows—in this case, if D is the disk centered at 1 of radius 6, then w ranges between -2224 and 679; if D is the disk of radius 12, then w ranges between -5893 and 3330.

Note 2: To increase accuracy, one need only include more terms in the Taylor expansion. This method will work whenever the branch points of ϕ lie outside some disk containing G, as the Taylor series for ϕ will converge on this disk.

Now consider the $\beta=1.4$ cardioid in Fig. 1. While we can still approximate $\log(\phi(z)/z)$ by polynomials on G, we can no longer do it by Taylor polynomials, and the construction of the polynomials is therefore much harder. One way to do it is to follow the usual construction in the proof of Runge's theorem, wherein path integrals are approximated by rational functions, and the poles of these functions are moved away to infinity.[6] This method is cumbersome, and will result in polynomials of very high degree.

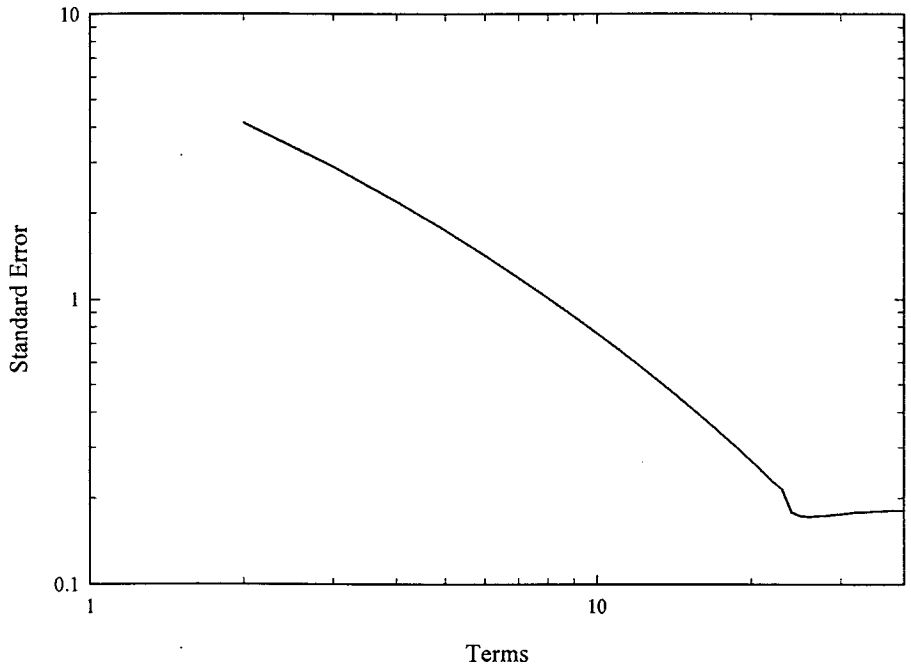

FIG. 4. The standard error (the standard deviation of $v+w$ evaluated on the cardiod boundary) of the $\beta=1.4$ expansion as a function of the number of terms in the expansion.

A better way is to pull the functions back to the disk. Consider not $f(z)=\log(\phi(z)/z)$ but the function $f(p(\zeta))=\log(\zeta/p(\zeta))$. Approximate this on the unit disk by polynomials in $p(\zeta)$. Do this by using Gramm–Schmidt to make an orthonormal basis from the functions $(p(\zeta))^n$, $n=0,1,2,...N$, and then finding the best L^2 approximant of $f(p(\zeta))$ with respect to this basis. If $\Sigma_{n=0}^{N} a_n(p(\zeta))^n$ approximates $f(p(\zeta))$, then $\Sigma_{n=0}^{N} a_n z^n$ will approximate $f(z)$ on G. Notice that following this procedure, if one increases N to improve the accuracy, all the coefficients a_n will change; one does not just add higher-order terms.

We used the above procedure to develop expansions for the $\beta=1.4$ cardioid shown in Fig. 1. The standard error of the expansion on the cardioid boundary ∂G is graphed in Fig. 4 as a function of the number of terms in the expansion. The error drops steadily with the number of terms until approximately 25 terms, at which point numeric errors limit the expansion accuracy. As for the $\beta=2$ cardioid, the difference between the approximate solution w and the exact, but divergent solution u is plotted in Fig. 2(b) for a 32-term expansion. On ∂G, the function $v+w$ takes values between -0.51 and $+0.52$ (for reference, $v+w$ equals 27.7 at $p=0$). Figure 3(b) shows that the equipotential $v+w=0$ is very close to ∂G, and that the equipotentials $+1$ and -1 surround ∂G.

IV. ACKNOWLEDGMENTS

We thank Jonathan Wurtele and Dmitry Khavinson for their helpful comments.

This work was supported by National Science Foundation Grant No. DMS 9531967 and the Office of Naval Research.

APPENDIX A: QUADRATURE DOMAINS

Let r be a rational function that is univalent on $\mathbf{D}$, and let $G=r(\mathbf{D})$. Such a G is the most general sort of simply connected quadrature domain. Let $\phi: G\to\mathbf{D}$ be the inverse (the Riemann map). Define the *Schwarz function* of G by

$$S(z)=\overline{r\left(\frac{1}{\phi(z)}\right)}.$$

Let $z_1,\ldots,z_K$ be the poles of S inside G. Then if f is any analytic function that is integrable on G, the following formula holds (Ref. 7, p. 19):

$$\int_G f(z)\ dA(z)=\pi\sum_{n=1}^{K}\ \mathrm{Res}\ (fS;z_n). \tag{A1}$$

Now, if r is actually a polynomial of degree N, and $r(0)=0$, then the Schwarz function of G will have a single pole of order N at 0. If the principal part of S at 0 is

$$\frac{\Sigma_{k=0}^{N-1}\ c_k z^k}{z^N},$$

then we obtain from Eq. (A1) that

$$\int_G \sum\ \alpha_k z^k\ dA(z)=\pi[\,\alpha_0 c_{N-1}+\alpha_1 c_{N-2}+\cdots+\alpha_{N-1}c_0].$$

Now suppose h is a real-valued harmonic function on G. Let $h=\mathrm{Re}\,(f)$ for some analytic function f (you can always do this when G is simply connected). With

$$\alpha_k=\frac{1}{k!}\frac{d^k f}{dz^k}\bigg|_{z=0}$$

we obtain that

$$\int_G h(z)\ dA(z)=\pi\ \mathrm{Re}[\,\alpha_0 c_{N-1}+\alpha_1 c_{N-2}+\cdots+\alpha_{N-1}c_0]. \tag{A2}$$

By the Cauchy–Riemann equations,

$$\frac{d^k f}{dz^k}=\frac{\partial^k h}{\partial x^k}-i\ \frac{\partial^k h}{\partial x^{k-1}\partial y}.$$

Substituting into Eq. (A2), we obtain

$$\int_G h(z)dA(z)=\pi\ \mathrm{Re}\left[\sum_{k=0}^{N-1}\ \frac{c^{N-k}}{k!}\left(\frac{\partial^k h}{\partial x^k}-i\ \frac{\partial^k h}{\partial x^{k-1}\partial y}\right)\right]. \tag{A3}$$

This is how one finds formula (6). One can calculate for the cardioids that the Schwarz function is

$$S(z)=\frac{2\beta^2+z+2\beta^2 z+2\beta(1+z)\sqrt{\beta^2+z}}{z^2},$$

and this has principal part

$$\frac{4\beta^2+(4\beta^2+2)z}{z^2}.$$

Combining this with Eq. (A3), one obtains Eq. (8).

APPENDIX B: EQUIPOTENTIALS AROUND *G* FORM SIMPLE CLOSED CURVES

We wish to show that ∂G is an isolated component of the set $\{z:u(z)+v(z)=0\}$, when G is a quadrature domain as in Sec. II B, and the boundary of G is smooth (as it is an algebraic curve,

we only need to avoid cusps like the point -1 in the $\beta=1$ cardioid in Fig. 1). This will mean that if one perturbs u slightly to get a function w, the zero set of $w+v$ will contain a closed curve close to the boundary of G.

As the equipotentials of $u+v$ are perpendicular to $\nabla(u+v)$, we will avoid degeneracy if we can show that $\nabla(u+v)$ is never 0 on ∂G. Let $\tilde{v}(\zeta)$ be the function we get by substituting $h(z) = -2\log|z-\zeta|$ into Eq. (6). Then $\tilde{v}$ has a singularity at 0, is harmonic everywhere else, and agrees with v outside of G. As both $u+v$ and $u+\overline{v}$ are superharmonic on G (because their Laplacian is nonpositive on G) and vanish on ∂G, they must be strictly positive on G. Therefore the zero set of $u+v$ is contained in $D\backslash G$, and equals the zero set of $u+\tilde{v}$. (For a discussion of superharmonic functions, see Ref. 11).

Let z_0 be a point in ∂G. Near z_0 the function $u+\overline{v}$ is harmonic, so it can be written as Re (f) for some analytic function f.[6] Without loss of generality, we can assume $f(z_0)=0$; if $f'(z_0)$ were zero, then $f(z)$ would behave like $(z-z_0)^n$ for some $n\geqslant 2$. But then f could not map a one-sided neighborhood of z_0 in G into the right half-plane (recall that ∂G has no cusp at z_0).

So we conclude that $f'(z_0)$ cannot be zero, and therefore

$$\nabla(u+\overline{v})\big|_{z_0}=\nabla(u+v)\big|_{z_0}=(\mathrm{Re}\,[f'(z_0)],\mathrm{Im}\,[f'(z_0)])\neq 0.$$

[1] J. Notte, A. J. Peurrung, J. Fajans, R. Chu, and J. Wurtele, Phys. Rev. Lett. **69**, 3056 (1992).
[2] R. Chu, J. S. Wurtele, J. Notte, A. J. Peurrung, and J. Fajans, Phys. Fluids B **5**, 2378 (1993).
[3] J. Fajans, E. Yu. Backhaus, and J. E. M^{c}Carthy, ''Equilibrium of highly asymmetric non-neutral plasmas,'' to appear in Phys. Plasmas.
[4] This assumption is true provided that the bounce motion along the magnetic field can be ignored, a condition which is frequently met.
[5] E. Yu. Backhaus, J. Fajans, and J. S. Wurtele, ''Stability of highly assymetric non-neutral plasmas,'' to appear in Phys. Plasmas.
[6] R. Greene and S. Krantz, *Function Theory of One Complex Variable* (Wiley, New York, 1997).
[7] H. Shapiro, *The Schwarz Function and its Generalization to Higher Dimensions*, University of Arkansas Lecture Notes (Wiley, New York, 1992).
[8] P. Ebenfelt, Compl. Var. **20**, 75 (1992).
[9] P. Ebenfelt, Compl. Var. **22**, 69 (1993).
[10] C. Pommerenke, *Boundary Behaviour of Conformal Maps* (Springer, Berlin, 1992).
[11] J. B. Garnett, *Bounded Analytic Functions* (Academic, New York, 1981).

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 39, NUMBER 12 DECEMBER 1998

On the triple sum formula for Wigner 9*j*-symbols

Hjalmar Rosengren[a)]
Department of Mathematics, Lund University, Box 118, S-221 00, Lund, Sweden

(Received 23 June 1998; accepted for publication 24 June 1998)

We give a new proof of the triple sum formula for Wigner 9*j*-symbols due to Ališauskas and Jucys. The proof uses explicit expressions for the coupling kernels recently introduced by the author. Parts of our results generalize to general recoupling coefficients. © *1998 American Institute of Physics.*
[S0022-2488(98)02112-4]

I. INTRODUCTION

The Wigner 9*j*-symbols arise in quantum mechanics as recoupling coefficients of four angular momenta. They were first introduced by Eugene Wigner in 1940. For the general theory of Wigner symbols and related objects we refer to the two-volume treatise of Biedenharn and Louck.[1,2]

There are many known expressions for 9*j*-symbols as multiple hypergeometric sums. Generally speaking, the larger the number of summation variables, the more symmetries these sums exhibit. The most compact known expression is the triple sum formula of Ališauskas and Jucys,[3] which does not exhibit any of the 72 symmetries of the 9*j*-symbol. Because of the low number of summation variables, this formula is useful in numerical computations,[4] and it has also been used to study the zeroes of 9*j*-symbols.[5,6] On the other hand, since the 72 symmetries are not exhibited in the triple sum formula, they give rise to nontrivial symmetries of this type of sums. This has been used to find several new hypergeometric summation and transformation formulas.[7–9]

The original proof in Ref. 3 uses a more complicated interpretation of 9*j*-symbols than the standard one, involving an embedding $\mathrm{SU}(2)\times\mathrm{SU}(2)\subseteq\mathrm{SO}(5)$. An elementary proof was given in Ref. 10. There the starting point is a formula expressing the 9*j*-symbol as a double sum, each term being the product of three 3*j*-symbols (Clebsch–Gordan coefficients). The triple sum formula is then derived using only elementary series manipulations. The proof is however quite tricky, and it seems very hard to find without knowledge of the end result.

In a previous paper,[11] we introduced some polynomials in $2n$ variables which were called coupling kernels. They are the reproducing kernels of certain spaces of coupling coefficients for n angular momenta, and may be expressed as n-variable hypergeometric sums. Using this explicit formula (for $n=4$) we will derive an expression for Wigner 9*j*-symbols as a double sum, each term being the product of only two 3*j*-symbols. After inserting appropriate single sum expressions for the 3*j*-symbols it is possible to perform the summation in one variable and arrive at a triple sum, which is equivalent to the one given in Ref. 3. Thus, in this case, there are no series manipulations involved and one can easily derive the formula without knowing it in advance.

We will follow the approach to coupling and recoupling coefficients of Refs. 11, 12. In particular, we will work with the Lie algebra $\mathfrak{su}(1,1)$ rather than with $\mathfrak{su}(2)$, which is more common in the literature. It is well-known that the Wigner symbols of $\mathfrak{su}(2)$ may be obtained from those of $\mathfrak{su}(1,1)$ by continuation. One may also generalize the relevant results of Refs. 11, 12 to $\mathfrak{su}(2)$, and then give a direct proof, analogous to the one below, of the triple sum formula in this case.

In Sec. II we survey parts of Refs. 11, 12. Most of this is reformulation of well-known facts in slightly different terms. The only new result that we need is the explicit expression for coupling kernels from Ref. 11 [equation (16) below]. In Sec. III we prove an expansion formula, given in Theorem 3.1, of which our double sum mentioned above is a special case. In Sec. IV we derive the Ališauskas–Jucys triple sum formula. Finally, in Sec. V we generalize Theorem 3.1 to arbitrary recoupling coefficients. We hope that this result, Theorem 5.5, may be useful also in the study of Wigner 3*nj*-symbols with $n>3$.

a)Electronic mail: hjalmar@maths.lth.se

 © 1998 American Institute of Physics

II. PRELIMINARIES

We will work with the holomorphic realization of highest weight representations of $\mathfrak{su}(1,1)$. Thus we let $\mathcal{A}^\nu$ ($\nu>0$) be the Hilbert space of analytic functions in the complex unit disc with reproducing kernel $(1-z\overline{w})^{-\nu}$. The norm of a function,

$$f(z)=\sum_{k=0}^{\infty} a_k z^k,$$

is then given by

$$\|f\|^2=\sum_{k=0}^{\infty} \frac{k!}{(\nu)_k}|a_k|^2,$$

where

$$(\nu)_k=\nu(\nu+1)\cdots(\nu+k-1), \quad (\nu)_0=1$$

is the Pochhammer symbol. This norm is invariant under the transformations,

$$f\mapsto f(\phi(z))\phi'(z)^{\nu/2},$$

where ϕ is an analytic automorphism of the disc. The derived representation is a unitary representation of $\mathfrak{su}(1,1)$ on $\mathcal{A}^\nu$ with the highest weight vector 1 of weight $-\nu$. Passing to the complexified Lie algebra $\mathfrak{sl}(2,\mathbb{C})$ yields the densely defined operators

$$E=-\frac{d}{dz}, \quad F=z^2\frac{d}{dz}+\nu z, \quad H=-\left(2z\frac{d}{dz}+\nu\right),$$

which satisfy the structure equations

$$[EF]=H, \quad [HE]=2E, \quad [HF]=-2F,$$

and the $\mathfrak{su}(1,1)$-reality conditions

$$E^*=-F, \quad F^*=-E, \quad H^*=H.$$

Consider a Hilbert tensor product $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of such spaces. We will use $\|\ \|$ and $\langle\ ,\ \rangle$ to denote its norm and scalar product. It decomposes under the Lie algebra action as

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}\simeq\bigoplus_{s=0}^{\infty}\binom{n+s-2}{n-2}\mathcal{A}^{|\nu|+2s}, \tag{1}$$

(where $|\nu|=\Sigma\nu_i$), that is, the space $\mathcal{A}^{|\nu|+2s}$ occurs with multiplicity $\binom{n+s-2}{n-2}$. A highest weight vector Q in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of weight $-(|\nu|+2s)$ is the image of the function 1 under an intertwining embedding,

$$\mathcal{A}^{|\nu|+2s}\to\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}.$$

In our realization, this means that Q is a homogeneous polynomial of degree s, which can be expressed in terms of the differences z_i-z_j of the coordinates. Equivalently, Q satisfies the identity

$$Q(az_1+b,\ldots,az_n+b)=a^sQ(z_1,\ldots,z_n), \quad a,b\in\mathbb{C}. \tag{2}$$

In agreement with (1), the space of such polynomials has dimension $\binom{n+s-2}{n-2}$.

Let $\mathcal{K}_Q$ be the embedding defined by $\mathcal{K}_Q(1)=Q$. In Ref. 12 it was proved that if

$$Q(z)=\sum_{|t|=s} c_t z_1^{t_1}\cdots z_n^{t_n}, \tag{3}$$

then the adjoint map $\mathcal{K}_Q^*$ is the differential operator given by

$$\mathcal{K}_Q^*(f_1\otimes\ldots\otimes f_n)(z)=\sum_{|t|=s}\overline{c_t}\frac{f_1^{(t_1)}(z)\cdots f_n^{(t_n)}(z)}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}. \tag{4}$$

In particular, writing $e_m(z)=z^m$,

$$\mathcal{K}_Q^*(e_{m_1}\otimes\ldots\otimes e_{m_n})=\sum_{|t|=s}\overline{c_t}\frac{(-m_1)_{t_1}\cdots(-m_n)_{t_n}}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}(-1)^s e_{|m|-s}=\overline{TQ(m_1,\ldots,m_n)}e_{|m|-s},$$

where TQ is the polynomial of degree s given by

$$TQ(m)=(-1)^s\sum_{|t|=s} c_t\frac{(-m_1)_{t_1}\cdots(-m_n)_{t_n}}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}.$$

This exhibits TQ as a coupling coefficient of our representation. If Q and Q' are two highest weight vectors, the corresponding coupling coefficients satisfy

$$\sum_{|m|=k+s}\frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}TQ(m)\,\overline{TQ'(m)}=\frac{(|\nu|+2s)_k}{k!}\langle Q,Q'\rangle, \tag{5}$$

for $k=0,1,2,\ldots$. Here s is the degree of Q or Q'; if their degrees are different both sides vanish.

For $n=2$, the space of highest weight vectors of degree s is one-dimensional, and spanned by $(z_1-z_2)^s$. The norm of this vector is given by

$$\|(z_1-z_2)^s\|^2=\frac{s!(\nu_1+\nu_2+s-1)_s}{(\nu_1)_s(\nu_2)_s}. \tag{6}$$

It will be convenient to write

$$\begin{aligned}P_s^{(\nu_1,\nu_2)}(m_1,m_2)&=(\nu_1)_s(\nu_2)_s(T(z_1-z_2)^s)(m_1,m_2)\\&=\sum_{k=\max(0,\,s-m_2)}^{\min(s,\,m_1)}\binom{s}{k}(-1)^k\frac{(\nu_1)_s(\nu_2)_s}{(\nu_1)_k(\nu_2)_{s-k}}(-m_1)_k(-m_2)_{s-k}.\end{aligned} \tag{7}$$

These polynomials are essentially Clebsch–Gordan coefficients, or Wigner $3j$-symbols, of our representation. As is well-known, they can be expressed in terms of Hahn polynomials; in standard notation,[13]

$$P_s^{(\nu_1,\nu_2)}(m_1,m_2)=(\nu_1)_s(-m_1-m_2)_s\,Q_s(m_1;\nu_1-1,\nu_2-1,m_1+m_2).$$

We will use some well-known properties of Clebsch–Gordan coefficients. First, it is clear from (7) that

$$P_s^{(\nu_1,\nu_2)}(m_1,m_2)=\begin{cases}(-1)^{m_2}s!\dfrac{(\nu_1)_s(\nu_2)_s}{(\nu_1)_{m_1}(\nu_2)_{m_2}}, & m_1+m_2=s,\\ 0, & m_1+m_2<s.\end{cases} \tag{8}$$

From (5) and (6) one obtains the orthogonality relations,

$$\sum_{m_1+m_2=s+k}\frac{(\nu_1)_{m_1}(\nu_2)_{m_2}}{m_1!\,m_2!}P_s^{(\nu_1,\nu_2)}(m_1,m_2)P_t^{(\nu_1,\nu_2)}(m_1,m_2)=\frac{(\nu_1+\nu_2+2s)_k}{k!}C_s^{\nu_1\nu_2}\delta_{st},\quad(9)$$

where

$$C_s^{\nu_1\nu_2}=s!(\nu_1)_s(\nu_2)_s(\nu_1+\nu_2+s-1)_s.\quad(10)$$

The expression (7) may be rewritten in terms of a terminating ${}_3F_2(1)$-series,

$${}_3F_2\left(\left.\begin{matrix}-n,a,b\\c,d\end{matrix}\right|1\right)=\sum_{k=0}^{n}\frac{(-n)_k(a)_k(b)_k}{(c)_k(d)_k k!}.$$

Because of the known symmetries of such series, cf. Ref. 14, p. 59, there is a large number of similar expressions, such as

$$P_s^{(\nu_1,\nu_2)}(m_1,m_2)=\sum_{k=0}^{s}\binom{s}{k}(-1)^{s-k}\frac{(\nu_1)_s(\nu_2)_s}{(\nu_1)_k(\nu_2)_{s-k}}(\nu_1+m_1)_k(\nu_2+m_2)_{s-k},\quad(11)$$

$$=(-1)^s\frac{m_2!}{(m_1+m_2-s)!}(\nu_1)_s(\nu_2+m_2)_{m_1}\,{}_3F_2\left(\left.\begin{matrix}-m_1,\nu_1+s,1-\nu_2-s\\\nu_1,1-\nu_2-m_1-m_2\end{matrix}\right|1\right),\quad(12)$$

$$=(\nu_1)_s(-m_1-m_2)_s\,{}_3F_2\left(\left.\begin{matrix}-m_1,-s,\nu_1+\nu_2+s-1\\\nu_1,-m_1-m_2\end{matrix}\right|1\right).\quad(13)$$

From the transformation formulas used to derive (12) it follows that

$${}_3F_2\left(\left.\begin{matrix}-m_1,\nu_1+s,1-\nu_2-s\\\nu_1,1-\nu_2-m_1-m_2\end{matrix}\right|1\right)=0,\quad(14)$$

if m_1, m_2, and s are integers with

$$m_1,s\geqslant 0,\quad m_1+m_2\geqslant s,\quad m_2<0.$$

This is a special case of Karlsson's summation formula.[15]

We will now discuss the coupling kernels introduced in Ref. 11. Let us consider the Hilbert subspace of highest weight vectors of degree s in $\mathcal{A}^{\nu_1}\otimes\ldots\otimes\mathcal{A}^{\nu_n}$. It has a reproducing kernel,

$$K_s(z,\overline{w})=\sum_{|t|=|u|=s}c_{t,u}z_1^{t_1}\cdots z_n^{t_n}\overline{w}_1^{u_1}\cdots\overline{w}_n^{u_n}.$$

The coupling kernels are obtained by applying the transform T in each of the two variables z and w; thus we define

$$\mathbf{P}_s(i,j)=\sum_{|t|=|u|=s}c_{t,u}\frac{(-i)_t(-j)_u}{(\nu)_t(\nu)_u},$$

where we use multi-index notation such as

$$(-i)_t=(-i_1)_{t_1}\cdots(-i_n)_{t_n}.$$

This is a reproducing kernel for the space of coupling coefficients of degree s; in fact,

$$\frac{(|\nu|+2s)_k}{k!}TQ(l)=\sum_{|m|=k+s}\frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}TQ(m)\mathbf{P}_s(l,m),$$

for each $k=0,1,2,\ldots$, and each highest weight vector Q of degree s. If (Q_k) and $(\tilde{Q}_l)$ are any two orthogonal bases of this space, then

$$\mathbf{P}_s(i,j)=\sum_{k,\,l}\frac{\langle Q_k,\tilde{Q}_l\rangle}{\|Q_k\|^2\|\tilde{Q}_l\|^2}\overline{TQ_k(i)}\,T\tilde{Q}_l(j). \tag{15}$$

In Ref. 11 an explicit expression for the coupling kernels was found, namely,

$$\mathbf{P}_s(i,j)=\frac{(-1)^s(-|i|)_s(-|j|)_s}{(|\nu|+s-1)_s s!}\sum_{m_1,\ldots,m_n=0}^{\infty}\frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}(-i)_m(-j)_m}{(-|i|)_{|m|}(-|j|)_{|m|}m!(\nu)_m} \tag{16}$$

(with the convention $0/0=0$, so that the sum is actually finite). Our proof of the triple sum formula will use this expression, though in fact we only need the case $|i|=|j|=s$, which is slightly easier to prove.

Let us now see how to introduce recoupling coefficients, and in particular Wigner $9j$-symbols, in the present approach. First we define the binary coupling of two highest weight vectors. Suppose that we decompose the tensor product as

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}\simeq(\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_k})\otimes(\mathcal{A}^{\nu_{k+1}}\otimes\cdots\otimes\mathcal{A}^{\nu_n}),$$

where $1\leqslant k\leqslant n-1$. Let Q_1 be a highest weight vector in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_k}$ of weight $-\mu_1$ and Q_2 one in $\mathcal{A}^{\nu_{k+1}}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of weight $-\mu_2$. Thus

$$\mu_1=\nu_1+\ldots+\nu_k+2t_1,\quad \mu_2=\nu_{k+1}+\ldots+\nu_n+2t_2,$$

where t_1 and t_2 are the degrees of Q_1 and Q_2. For s a non-negative integer we define

$$(Q_1,Q_2)_s=(\mu_1)_s(\mu_2)_s(\mathcal{K}_{Q_1}\otimes\mathcal{K}_{Q_2})(z_1-z_2)^s,$$

which is a highest weight vector in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of weight $-(\mu_1+\mu_2+2s)$ and degree t_1+t_2+s. Equivalently, $(Q_1,Q_2)_s$ is determined by

$$\mathcal{K}^*_{(Q_1,Q_2)_s}(f_1\otimes\ldots\otimes f_n)=(\mu_1)_s(\mu_2)_s\mathcal{K}^*_{(z_1-z_2)^s}(\mathcal{K}^*_{Q_1}(f_1\otimes\ldots\otimes f_k)\otimes\mathcal{K}^*_{Q_2}(f_{k+1}\otimes\ldots\otimes f_n)).$$

Choosing $f_i(z)=z^{m_i}$ gives the important formula

$$T(Q_1,Q_2)_s(m)=TQ_1(m')\,TQ_2(m'')\,P_s^{(\mu_1,\mu_2)}(|m'|-t_1,|m''|-t_2), \tag{17}$$

where

$$m'=(m_1,\ldots,m_k),\quad m''=(m_{k+1},\ldots,m_n).$$

Moreover,

$$\|(Q_1,Q_2)_s\|^2=C_s^{\mu_1\mu_2}\|Q_1\|^2\|Q_2\|^2,$$

with the constant as in (10).

We also need a more general type of binary coupling, corresponding to a decomposition,

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}\simeq(\mathcal{A}^{\nu_{\sigma_1}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_k}})\otimes(\mathcal{A}^{\nu_{\sigma_{k+1}}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_n}}),$$

where σ is a permutation of $(1,2,\ldots,n)$. If Q_1 and Q_2 are highest weight vectors in $\mathcal{A}^{\nu_{\sigma_1}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_k}}$ and $\mathcal{A}^{\nu_{\sigma_{k+1}}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_n}}$ respectively, we define

$$(Q_1,Q_2)_s^{\sigma}(z_1,\ldots,z_n)=(Q_1,Q_2)_s(z_{\sigma_1},\ldots,z_{\sigma_n}).$$

Then $(Q_1,Q_2)_s^{\sigma}$ is a highest weight vector in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$.

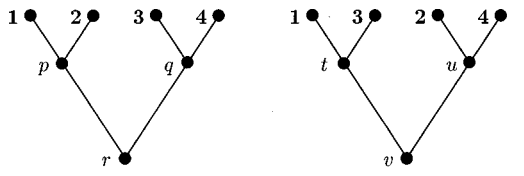

FIG. 1.

Now let $n=4$ and $\sigma=(1,3,2,4)$. We consider the two orthogonal systems of highest weight vectors in $\mathcal{A}^{\nu_1}\otimes\mathcal{A}^{\nu_2}\otimes\mathcal{A}^{\nu_3}\otimes\mathcal{A}^{\nu_4}$ defined by

$$\begin{aligned} Q_{pqr}&=((1,1)_p,(1,1)_q)_r,\\ \tilde{Q}_{tuv}&=((1,1)_t,(1,1)_u)_v^{\sigma}. \end{aligned} \tag{18}$$

Graphically, these iterated couplings may be represented as in Figure 1.

Let

$$W_{pqr}^{tuv}=\langle Q_{pqr},\tilde{Q}_{tuv}\rangle,$$

where the parameters ν_i are suppressed from the notation. These are essentially the Wigner $9j$-coefficients of our representation. Since W_{pqr}^{tuv} vanishes if $p+q+r\neq t+u+v$, we may assume that

$$p+q+r=t+u+v=s.$$

In more standard notation for $9j$-symbols, we have

$$\begin{Bmatrix} j_1 & j_2 & j_{12}\\ j_3 & j_4 & j_{34}\\ j_{13} & j_{24} & j \end{Bmatrix}=\frac{1}{\|Q_{pqr}\|\|\tilde{Q}_{tuv}\|}W_{pqr}^{tuv},$$

where the ''angular momenta'' are given by

$$j_1=-\tfrac{1}{2}\nu_1,\quad j_2=-\tfrac{1}{2}\nu_2,\quad j_3=-\tfrac{1}{2}\nu_3,\quad j_4=-\tfrac{1}{2}\nu_4,$$

$$j_{12}=-\tfrac{1}{2}(\nu_1+\nu_2+2p),\quad j_{34}=-\tfrac{1}{2}(\nu_3+\nu_4+2q),$$

$$j_{13}=-\tfrac{1}{2}(\nu_1+\nu_3+2t),\quad j_{24}=-\tfrac{1}{2}(\nu_2+\nu_4+2u),$$

$$j=-\tfrac{1}{2}(\nu_1+\nu_2+\nu_3+\nu_4+2s),$$

and the normalizing factor by

$$\|Q_{pqr}\|\|\tilde{Q}_{tuv}\|=(C_p^{\nu_1\nu_2}C_q^{\nu_3\nu_4}C_r^{\nu_1+\nu_2+2p,\ \nu_3+\nu_4+2q}C_t^{\nu_1\nu_3}C_u^{\nu_2\nu_4}C_v^{\nu_1+\nu_3+2t,\ \nu_2+\nu_4+2u})^{1/2}.$$

We conclude this section with a remark on transvectants and Wigner symbols. Let us consider the case $n=2$ of (4). It will be convenient to write

$$\begin{aligned} \mathcal{J}_s^{\nu_1\nu_2}(f_1,f_2)(z)&=(\nu_1)_s(\nu_2)_s\,\mathcal{K}^{*}_{(z_1-z_2)^s}(f_1\otimes f_2)(z)\\ &=\sum_{k=0}^{s}\binom{s}{k}(-1)^{s-k}\frac{(\nu_1)_s(\nu_2)_s}{(\nu_1)_k(\nu_2)_{s-k}}f_1^{(k)}(z)f_2^{(s-k)}(z), \end{aligned}$$

so that, again with $e_m(z)=z^m$,

$$\mathcal{J}_s^{\nu_1\nu_2}(e_{m_1},e_{m_2})=P_s^{(\nu_1,\nu_2)}(m_1,m_2)\,e_{m_1+m_2-s}.$$

This operator occurs in classical invariant theory and is known as the transvectant. It has also been called Cohen's operator and the Rankin–Cohen bracket.[16] It is clear from the above that the Wigner $9j$-symbols occur in the expansion

$$\mathcal{J}_r^{\nu_1+\nu_2+2p,\,\nu_3+\nu_4+2q}(\mathcal{J}_p^{\nu_1\nu_2}(f_1,f_2),\mathcal{J}_q^{\nu_3\nu_4}(f_3,f_4))$$
$$=\sum_{t+u+v=p+q+r}\frac{W_{pqr}^{tuv}}{\|\tilde{Q}_{tuv}\|^2}\,\mathcal{J}_v^{\nu_1+\nu_3+2t,\,\nu_2+\nu_4+2u}(\mathcal{J}_t^{\nu_1\nu_3}(f_1,f_3),\mathcal{J}_u^{\nu_2\nu_4}(f_2,f_4)).$$

It is interesting that in a recent paper[17] the coefficients occurring in

$$\mathcal{J}_r^{\nu_1+\nu_2+2q,\,\nu_3}(\mathcal{J}_q^{\nu_1\nu_2}(f_1,f_2),f_3)=\sum_{t+u=q+r}R_{qr}^{tu}\,\mathcal{J}_u^{\nu_1,\,\nu_2+\nu_3+2t}(f_1,\mathcal{J}_t^{\nu_2\nu_3}(f_2,f_3))$$

arise in a seemingly very different context. These are, up to a normalizing factor, the Racah coefficients, or Wigner $6j$-symbols, of our representation.

III. AN EXPANSION FORMULA

In this section we will derive a formula expressing the product of one $9j$- and two $3j$-symbols as a quadruple sum, each term being a product of four $3j$-symbols. As a special case we get an expression for the $9j$-symbols as a double sum of products of two $3j$-symbols, of which the triple sum formula is a rather easy consequence.

Let us choose Q and $\tilde{Q}$ in (15) as the bases given by (18). This gives

$$\mathbf{P}_s(i,j)=\sum_{\substack{p+q+r=s\\t+u+v=s}}\frac{W_{pqr}^{tuv}}{\|Q_{pqr}\|^2\|\tilde{Q}_{tuv}\|^2}\,\overline{TQ_{pqr}(i)}\,T\tilde{Q}_{tuv}(j). \tag{19}$$

Moreover, by (17),

$$TQ_{pqr}(i)=P_p^{(\nu_1,\nu_2)}(i_1,i_2)\,P_q^{(\nu_3,\nu_4)}(i_3,i_4)\,P_r^{(\nu_1+\nu_2+2p,\nu_3+\nu_4+2q)}(i_1+i_2-p,i_3+i_4-q),$$

$$T\tilde{Q}_{tuv}(j)=P_t^{(\nu_1,\nu_3)}(j_1,j_3)\,P_u^{(\nu_2,\nu_4)}(j_2,j_4)\,P_v^{(\nu_1+\nu_3+2t,\nu_2+\nu_4+2u)}(j_1+j_3-t,j_2+j_4-u).$$

Since these are real polynomials, we may ignore the conjugate bar in (19). Let us temporarily write $P_p=P_p^{(\nu_1,\nu_2)}$ and so on. We fix eight non-negative integers,

$$k_{12},k_{34},k_{13},k_{24},p',q',t',u',$$

and apply, to both sides of (19), the summation operator,

$$\sum_{\substack{i_1+i_2=p'+k_{12}\\i_3+i_4=q'+k_{34}}}\frac{(\nu)_i}{i!}P_{p'}(i_1,i_2)P_{q'}(i_3,i_4)\sum_{\substack{j_1+j_3=t'+k_{13}\\j_2+j_4=u'+k_{24}}}\frac{(\nu)_j}{j!}P_{t'}(j_1,j_3)P_{u'}(j_2,j_4).$$

By the orthogonality of Clebsch–Gordan coefficients (9), the right hand side vanishes unless

$$(p',q',t',u')=(p,q,t,u),$$

and we obtain

$$\sum_{\substack{i_1+i_2=p+k_{12}\\ i_3+i_4=q+k_{34}}}\sum_{\substack{j_1+j_3=t+k_{13}\\ j_2+j_4=u+k_{24}}}\frac{(\nu)_i(\nu)_j}{i!\,j!}\mathbf{P}_s(i,j)P_p(i_1,i_2)P_q(i_3,i_4)P_t(j_1,j_3)P_u(j_2,j_4)$$

$$=\frac{(\nu_1+\nu_2+2p)_{k_{12}}(\nu_3+\nu_4+2q)_{k_{34}}(\nu_1+\nu_3+2t)_{k_{13}}(\nu_2+\nu_4+2u)_{k_{24}}}{k_{12}!\,k_{34}!\,k_{13}!\,k_{24}!\,C_r^{\nu_1+\nu_2+2p,\,\nu_3+\nu_4+2q}C_v^{\nu_1+\nu_3+2t,\,\nu_2+\nu_4+2u}}$$

$$\times W_{pqr}^{tuv}P_r(k_{12},k_{34})P_v(k_{13},k_{24}). \tag{20}$$

Now we expand $\mathbf{P}_s$ according to (16). Then the left hand side of (20) takes the form

$$(-1)^s\frac{(-p-q-k_{12}-k_{34})_s(-t-u-k_{13}-k_{24})_s}{(|\nu|+s-1)_s s!}$$

$$\times\sum_{m_1,m_2,m_3,m_4=0}^{\infty}\frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}}{(-p-q-k_{12}-k_{34})_{|m|}(-t-u-k_{13}-k_{24})_{|m|}m!(\nu)_m}$$

$$\times\sum_{i_1+i_2=p+k_{12}}\frac{(\nu_1)_{i_1}(\nu_2)_{i_2}}{i_1!\,i_2!}(-i_1)_{m_1}(-i_2)_{m_2}P_p(i_1,i_2)$$

$$\times\sum_{i_3+i_4=q+k_{34}}\frac{(\nu_3)_{i_3}(\nu_4)_{i_4}}{i_3!\,i_4!}(-i_3)_{m_3}(-i_4)_{m_4}P_q(i_3,i_4)$$

$$\times\sum_{j_1+j_3=t+k_{13}}\frac{(\nu_1)_{j_1}(\nu_3)_{j_3}}{j_1!\,j_3!}(-j_1)_{m_1}(-j_3)_{m_3}P_t(j_1,j_3)$$

$$\times\sum_{j_2+j_4=u+k_{24}}\frac{(\nu_2)_{j_2}(\nu_4)_{j_4}}{j_2!\,j_4!}(-j_2)_{m_2}(-j_4)_{m_4}P_u(j_2,j_4). \tag{21}$$

The four inner sums may be computed using the probably well-known formula

$$\sum_{i_1+i_2=s+k}\frac{(\nu_1)_{i_1}(\nu_2)_{i_2}}{i_1!\,i_2!}(-i_1)_{m_1}(-i_2)_{m_2}P_s^{(\nu_1,\nu_2)}(i_1,i_2)$$

$$=(-1)^{m_1+m_2}\frac{(\nu_1+\nu_2+s+m_1+m_2)_{k+s-m_1-m_2}}{(k+s-m_1-m_2)!}(\nu_1)_{m_1}(\nu_2)_{m_2}P_s^{(\nu_1,\nu_2)}(m_1,m_2). \tag{22}$$

Note that, by (8), the left hand side vanishes unless $m_1+m_2\leqslant s+k$, and the right hand side unless $m_1+m_2\geqslant s$. One may prove (22) by inserting expression (11) for the Clebsch–Gordan coefficients (see the proof of Lemma 5.4 below).

It follows that the expression (21) equals

$$
\begin{aligned}
&(-1)^s \frac{(-p-q-k_{12}-k_{34})_s(-t-u-k_{13}-k_{24})_s}{(|\nu|+s-1)_s s!} \\
&\times \sum_{m_1,m_2,m_3,m_4=0}^{\infty} \frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}(\nu)_m}{(-p-q-k_{12}-k_{34})_{|m|}(-t-u-k_{13}-k_{24})_{|m|} m!} \\
&\times \frac{(\nu_1+\nu_2+p+m_1+m_2)_{k_{12}+p-m_1-m_2}(\nu_3+\nu_4+q+m_3+m_4)_{k_{34}+q-m_3-m_4}}{(k_{12}+p-m_1-m_2)!(k_{34}+q-m_3-m_4)!} \\
&\times \frac{(\nu_1+\nu_3+t+m_1+m_3)_{k_{13}+t-m_1-m_3}(\nu_2+\nu_4+u+m_2+m_4)_{k_{24}+u-m_2-m_4}}{(k_{13}+t-m_1-m_3)!(k_{24}+u-m_2-m_4)!} \\
&\times P_p(m_1,m_2)P_q(m_3,m_4)P_t(m_1,m_3)P_u(m_2,m_4).
\end{aligned}
$$

On the other hand, this is equal to the right hand side of (20). After some cancellations, this gives the following formula.

Theorem 3.1: *Let* k_{12}, k_{34}, k_{13}, k_{24}, p, q, r, t, u, and v be non-negative integers with

$$p+q+r=t+u+v=s.$$

Then, in the notation above, the equality

$$
\begin{aligned}
&\frac{1}{k_{12}!\,k_{34}!\,k_{13}!\,k_{24}!\,C_r^{\nu_1+\nu_2+2p,\,\nu_3+\nu_4+2q} C_v^{\nu_1+\nu_3+2t,\,\nu_2+\nu_4+2u}} \\
&\quad \times W_{pqr}^{tuv} P_r^{(\nu_1+\nu_2+2p,\nu_3+\nu_4+2q)}(k_{12},k_{34}) P_v^{(\nu_1+\nu_3+2t,\nu_2+\nu_4+2u)}(k_{13},k_{24}) \\
&= (-1)^s \frac{(-p-q-k_{12}-k_{34})_s(-t-u-k_{13}-k_{24})_s}{(|\nu|+s-1)_s s!} \\
&\quad \times \sum_{m_1,m_2,m_3,m_4} \frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}(\nu)_m}{(-p-q-k_{12}-k_{34})_{|m|}(-t-u-k_{13}-k_{24})_{|m|} m!} \\
&\quad \times \frac{1}{(k_{12}+p-m_1-m_2)!(k_{34}+q-m_3-m_4)!(k_{13}+t-m_1-m_3)!(k_{24}+u-m_2-m_4)!} \\
&\quad \times \frac{1}{(\nu_1+\nu_2+2p)_{m_1+m_2-p}(\nu_3+\nu_4+2q)_{m_3+m_4-q}(\nu_1+\nu_3+2t)_{m_1+m_3-t}(\nu_2+\nu_4+2u)_{m_2+m_4-u}} \\
&\quad \times P_p^{(\nu_1,\nu_2)}(m_1,m_2) P_q^{(\nu_3,\nu_4)}(m_3,m_4) P_t^{(\nu_1,\nu_3)}(m_1,m_3) P_u^{(\nu_2,\nu_4)}(m_2,m_4),
\end{aligned}
$$

holds, where the sum is over all non-negative integers such that

$$
\begin{aligned}
&p \leqslant m_1+m_2 \leqslant p+k_{12}, \\
&q \leqslant m_3+m_4 \leqslant q+k_{34}, \\
&t \leqslant m_1+m_3 \leqslant t+k_{13}, \\
&u \leqslant m_2+m_4 \leqslant u+k_{24}, \\
&m_1+m_2+m_3+m_4 \leqslant s.
\end{aligned}
$$

To prove the triple sum formula we only need the case

$$k_{12}=k_{13}=0, \quad k_{34}=r, \quad k_{24}=v.$$

Then the two conditions,

$$m_1+m_2=p, \quad m_1+m_3=t,$$

hold, so that the sum in the theorem reduces to a double sum. Moreover, the factors P_r, P_v, P_p, and P_t become trivial by (8), so we obtain a formula expressing the Wigner $9j$-symbol as a double sum of products of two Clebsch–Gordan coefficients. After further simplifications, such as writing

$$k_{34}+q-m_3-m_4=s-|m|,$$

we obtain the following formula.

Corollary 3.2: For p, q, r, t, u, and v non-negative integers with

$$p+q+r=t+u+v=s,$$

the equality

$$\begin{aligned}W_{pqr}^{tuv}=&(-1)^{r+v+s}\frac{p!\,r!\,t!\,v!(\nu_1)_p(\nu_2)_p(\nu_1)_t(\nu_3)_t}{(|\nu|+s-1)_s}\\&\times(|\nu|+2p+2q+r-1)_r(|\nu|+2t+2u+v-1)_v\\&\times\sum_{m_1,m_2,m_3,m_4}\frac{(|\nu|+s-1)_{|m|}(\nu_4)_{m_4}}{m!(s-|m|)!(\nu_1)_{m_1}}(-1)^{m_1+m_4}P_q^{(\nu_3,\nu_4)}(m_3,m_4)P_u^{(\nu_2,\nu_4)}(m_2,m_4)\\&\times(\nu_2+\nu_4+u+m_2+m_4)_{s-|m|}(\nu_3+\nu_4+q+m_3+m_4)_{s-|m|},\end{aligned}$$

holds, where the sum is over all non-negative integers such that

$$\begin{aligned}&m_1+m_2=p,\\&q\leqslant m_3+m_4\leqslant q+r,\\&m_1+m_3=t,\\&u\leqslant m_2+m_4\leqslant u+v.\end{aligned}$$

IV. THE TRIPLE SUM FORMULA

We will now prove the triple sum formula. Let us rewrite the sum in Corollary 3.2 using $j=s-|m|$ and m_1 as summation variables, thus replacing m_2, m_3, and m_4 by

$$p-m_1,\quad t-m_1,\quad s+m_1-p-t-j,$$

respectively. Using the identities

$$(a)_{s-k}=(-1)^k\frac{(a)_s}{(1-a-s)_k},$$

$$(a)_k=(-1)^k(1-a-k)_k,$$

we obtain

$$\begin{aligned}W_{pqr}^{tuv}=&(-1)^{q+u}(|\nu|+2p+2q+r-1)_r(|\nu|+2t+2u+v-1)_v\\&\times r!\,v!(\nu_1)_p(\nu_2)_p(\nu_1)_t(\nu_3)_t\sum_{j,m_1}\frac{(1-\nu_2-\nu_4-2u-v)_j(1-\nu_3-\nu_4-2q-r)_j}{j!(2-|\nu|-2s)_j}\\&\times\frac{(-p)_{m_1}(-t)_{m_1}}{m_1!(\nu_1)_{m_1}}\frac{(\nu_4)_{s+m_1-p-t-j}}{(s+m_1-p-t-j)!}\\&\times P_q^{(\nu_3,\nu_4)}(t-m_1,s+m_1-p-t-j)\,P_u^{(\nu_2,\nu_4)}(p-m_1,s+m_1-p-t-j),\end{aligned}$$

where the sum is over j and m_1 such that

$$0\leqslant m_1\leqslant \min(p,t),$$
$$0\leqslant j\leqslant \min(r,v),$$
$$p+t+j\leqslant s+m_1.$$

Now let us expand P_q and P_u using formulas (12) and (13), respectively. Thus we write

$$P_q=(-1)^q\frac{(s+m_1-p-t-j)!}{(r-j)!}(\nu_3)_q(\nu_4+s+m_1-p-t-j)_{t-m_1}$$
$$\times\sum_{i=0}^{t-m_1}\frac{(m_1-t)_i(\nu_3+q)_i(1-\nu_4-q)_i}{i!(\nu_3)_i(1-\nu_4-q-r+j)_i},$$

$$P_u=(\nu_2)_u(j-u-v)_u\sum_{k=0}^{\min(u,\,p-m_1)}\frac{(-u)_k(m_1-p)_k(\nu_2+\nu_4+u-1)_k}{k!(j-u-v)_k(\nu_2)_k}.$$

We then change the order of summation, pulling m_1 inside. We also manipulate the Pochhammer symbols to make i,j,k appear only as subscripts. This gives

$$W^{tuv}_{pqr}=(u+v)!(\nu_1)_p(\nu_1)_t(\nu_2)_p(\nu_2)_u(\nu_3)_q(\nu_3)_t(\nu_4)_{q+r}$$
$$\times(|\nu|+2p+2q+r-1)_r(|\nu|+2t+2u+v-1)_v$$
$$\times\sum_{i=0}^{t}\sum_{j=0}^{\min(r,\,v)}\sum_{k=0}^{\min(p,\,u)}\frac{1}{i!\,j!\,k!}\frac{1}{(1-\nu_4-q-r)_{i+j}(-u-v)_{j+k}}$$
$$\times\frac{(-r)_j(-v)_j(1-\nu_3-\nu_4-2q-r)_j(1-\nu_2-\nu_4-2u-v)_j}{(2-|\nu|-2s)_j}$$
$$\times\frac{(-t)_i(\nu_3+q)_i(1-\nu_4-q)_i}{(\nu_3)_i}\frac{(-p)_k(-u)_k(\nu_2+\nu_4+u-1)_k}{(\nu_2)_k}$$
$$\times\sum_{m_1=\max(0,\,p+t+j-s)}^{\min(t-i,\,p-k)}\frac{(i-t)_{m_1}(k-p)_{m_1}}{m_1!(\nu_1)_{m_1}}.$$

However, it follows from (14), that if we instead first sum over i, then the sums for which $m_1\leqslant p+t+j-s$ (if there are any such) vanish. Thus we can replace the lower limit in the inner sum by 0, and then compute it using the Chu–Vandermonde formula:

$$\sum_{m_1}\frac{(i-t)_{m_1}(k-p)_{m_1}}{m_1!(\nu_1)_{m_1}}=\frac{(\nu_1)_{p+t-i-k}}{(\nu_1)_{t-i}(\nu_1)_{p-k}}=\frac{(\nu_1)_{p+t}}{(\nu_1)_p(\nu_1)_t}\frac{(1-\nu_1-t)_i(1-\nu_1-p)_k}{(1-\nu_1-p-t)_{i+k}}.$$

This gives the following result, which is equivalent to the Ališauskas–Jucys triple sum formula. As remarked in Ref. 18, the sum may be expressed in terms of generalized hypergeometric functions introduced by Srivastava.[19]

Theorem 4.3: *The Wigner* $9j$*-symbols for* $\mathfrak{su}(1,1)$ *are given by*

$$
\begin{aligned}
W_{pqr}^{tuv}=&(u+v)!(\nu_1)_{p+t}(\nu_2)_p(\nu_2)_u(\nu_3)_q(\nu_3)_t(\nu_4)_{q+r}\\
&\times(|\nu|+2p+2q+r-1)_r(|\nu|+2t+2u+v-1)_v\\
&\times\sum_{i=0}^{t}\sum_{j=0}^{\min(r,\,v)}\sum_{k=0}^{\min(p,\,u)}\frac{1}{i!\,j!\,k!}\,\frac{1}{(1-\nu_4-q-r)_{i+j}(1-\nu_1-p-t)_{i+k}(-u-v)_{j+k}}\\
&\times\frac{(-t)_i(1-\nu_1-t)_i(\nu_3+q)_i(1-\nu_4-q)_i}{(\nu_3)_i}\\
&\times\frac{(-r)_j(-v)_j(1-\nu_3-\nu_4-2q-r)_j(1-\nu_2-\nu_4-2u-v)_j}{(2-|\nu|-2s)_j}\\
&\times\frac{(-p)_k(-u)_k(1-\nu_1-p)_k(\nu_2+\nu_4+u-1)_k}{(\nu_2)_k},
\end{aligned}
$$

where $p+q+r=t+u+v=s$.

This formula might look awkward, but we stress that it is the most compact known expression for Wigner $9j$-symbols. Note that the parameter occurring in the Pochhammer symbol $(\,)_{i+j}$ may be written in two ways:

$$1-\nu_4-q-r=(1-\nu_4-q)+(-r)=(\nu_3+q)+(1-\nu_3-\nu_4-2q-r),$$

as a sum of those occurring in $(\,)_i$ and $(\,)_j$ in the numerator. This is also true for $(\,)_{i+k}$ and $(\,)_{j+k}$, and each of the twelve numerator parameters occur once in the corresponding six sums.

V. GENERAL RECOUPLING COEFFICIENTS

In this section we will extend Theorem 3.1 to arbitrary recoupling coefficients. Thus we consider two decompositions:

$$
\begin{aligned}
\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}&\simeq(\mathcal{A}^{\nu_{\sigma_1}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_l}})\otimes(\mathcal{A}^{\nu_{\sigma_{l+1}}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\sigma_n}})=B_1\otimes B_2\\
&\simeq(\mathcal{A}^{\nu_{\tau_1}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\tau_{l'}}})\otimes(\mathcal{A}^{\nu_{\tau_{l'+1}}}\otimes\cdots\otimes\mathcal{A}^{\nu_{\tau_n}})=B_1'\otimes B_2',
\end{aligned}
$$

where σ and τ are permutations and $1\leqslant l, l'\leqslant n-1$. We choose arbitrary highest weight vectors,

$$Q_1\in B_1,\quad Q_2\in B_2,\quad Q_1'\in B_1',\quad Q_2'\in B_2',$$

of respective degree,

$$s_1,\quad s_2,\quad s_1',\quad s_2',$$

and weight,

$$-\mu_1,\quad -\mu_2,\quad -\mu_1',\quad -\mu_2'.$$

For a multi-index $(i_1,\ldots,i_n)$ we write

$$i_1^\sigma=(i_{\sigma_1},\ldots,i_{\sigma_l}),\quad i_2^\sigma=(i_{\sigma_{l+1}},\ldots,i_{\sigma_n}),$$

$$i_1^\tau=(i_{\tau_1},\ldots,i_{\tau_{l'}}),\quad i_2^\tau=(i_{\tau_{l'+1}},\ldots,i_{\tau_n}),$$

so that

$$\mu_1=|\nu_1^\sigma|+2s_1,\quad \mu_2=|\nu_2^\sigma|+2s_2,$$

$$\mu_1'=|\nu_1^\tau|+2s_1',\quad \mu_2'=|\nu_2^\tau|+2s_2'.$$

We will study the recoupling coefficient,

$$\langle (Q_1,Q_2)^{\sigma}_{s_3},(Q_1',Q_2')^{\tau}_{s_3'}\rangle,$$

where we may assume that

$$s_1+s_2+s_3=s_1'+s_2'+s_3'=s.$$

In (19) we replace Q_{pqr} and $\tilde{Q}_{tuv}$ by suitable orthogonal systems of highest weight vectors containing $(Q_1,Q_2)^{\sigma}_{s_3}$ and $(Q_1',Q_2')^{\tau}_{s_3'}$. Using (5) instead of (9) we obtain

$$\sum_{\substack{|i_1^{\sigma}|=s_1+k_1,\ |i_2^{\sigma}|=s_2+k_2,\\ |j_1^{\tau}|=s_1'+k_1',\ |j_2^{\tau}|=s_2'+k_2'}} \frac{(\nu)_i(\nu)_j}{i!\,j!}\mathbf{P}_s(i,j)\,TQ_1(i_1^{\sigma})\,TQ_2(i_2^{\sigma})\,\overline{TQ_1'(j_1^{\tau})\,TQ_2'(j_2^{\tau})}$$

$$=\frac{(\mu_1)_{k_1}(\mu_2)_{k_2}(\mu_1')_{k_1'}(\mu_2')_{k_2'}}{k_1!\,k_2!\,k_1'!\,k_2'!\,C_{s_3}^{\mu_1\mu_2}C_{s_3'}^{\mu_1'\mu_2'}}\langle (Q_1,Q_2)^{\sigma}_{s_3},(Q_1',Q_2')^{\tau}_{s_3'}\rangle$$

$$\times P_{s_3}^{(\mu_1,\mu_2)}(k_1,k_2)P_{s_3'}^{(\mu_1',\mu_2')}(k_1',k_2'),$$

where k_1,k_2,k_1',k_2' are arbitrary non-negative integers. By (16), the left hand side equals

$$(-1)^s\frac{(-s_1-s_2-k_1-k_2)_s(-s_1'-s_2'-k_1'-k_2')_s}{(|\nu|+s-1)_s s!}$$

$$\times\sum_m \frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}}{(-s_1-s_2-k_1-k_2)_{|m|}(-s_1'-s_2'-k_1'-k_2')_{|m|}m!(\nu)_m}$$

$$\times\sum_{|i_1^{\sigma}|=s_1+k_1}\frac{(\nu_1^{\sigma})_{i_1^{\sigma}}}{i_1^{\sigma}!}(-i_1^{\sigma})_{m_1^{\sigma}}TQ_1(i_1^{\sigma})\sum_{|i_2^{\sigma}|=s_2+k_2}\frac{(\nu_2^{\sigma})_{i_2^{\sigma}}}{i_2^{\sigma}!}(-i_2^{\sigma})_{m_2^{\sigma}}TQ_2(i_2^{\sigma})$$

$$\times\sum_{|j_1^{\tau}|=s_1'+k_1'}\frac{(\nu_1^{\tau})_{j_1^{\tau}}}{j_1^{\tau}!}(-j_1^{\tau})_{m_1^{\tau}}\overline{TQ_1'(j_1^{\tau})}\sum_{|j_2^{\tau}|=s_2'+k_2'}\frac{(\nu_2^{\tau})_{j_2^{\tau}}}{j_2^{\tau}!}(-j_2^{\tau})_{m_2^{\tau}}\overline{TQ_2'(j_2^{\tau})}.$$

To compute the inner sums we need instead of (22) the following lemma.

Lemma 5.4: If Q is a highest weight vector of degree s in $\mathcal{A}^{\nu_1}\otimes\ldots\otimes\mathcal{A}^{\nu_n}$, then

$$\sum_{|i|=s+k}\frac{(\nu)_i}{i!}(-i)_m\,TQ(i)=(-1)^{|m|}\frac{(|\nu|+s+|m|)_{k+s-|m|}}{(k+s-|m|)!}(\nu)_m\,TQ(m).$$

In Ref. 11 this is derived as a limit case of a more general formula. We will give an independent proof here.

Proof: Let Q be given by (3). Then, by (2),

$$\sum_{|t|=s}c_t z^t=\sum_{|t|=s}c_t(1+z_1)^{t_1}\cdots(1+z_n)^{t_n}$$

$$=\sum_{|t|=s}c_t\sum_{k_1,\ldots,k_n\geqslant 0}\frac{(-t)_k}{k!}(-1)^{|k|}z^k.$$

Applying T on both sides gives

$$TQ(m)=\sum_{|t|=s}c_t\sum_{k_1,\ldots,k_n\geqslant 0}\frac{(-t)_k(-m)_k}{k!(\nu)_k}=\sum_{|t|=s}c_t\frac{(\nu+m)_t}{(\nu)_t}, \tag{23}$$

by the Chu–Vandermonde formula; that is $TQ(m)=(-1)^sTQ(-\nu-m)$. This generalizes expression (11). Thus we have

$$\sum_{|i|=s+k}\frac{(\nu)_i}{i!}(-i)_m\,TQ(i)=\sum_{|i|=s+k}\frac{(\nu)_i}{i!}(-i)_m\sum_{|t|=s}c_t\frac{(\nu+i)_t}{(\nu)_t}$$

$$=\sum_{|t|=s}c_t\sum_{|i|=s+k}\frac{(\nu+t)_i}{i!}(-i)_m.$$

Replacing i by $i+m$ in the inner sum gives

$$\sum_{|i|=s+k}\frac{(\nu+t)_i}{i!}(-i)_m=\sum_{|i|=s+k-|m|}\frac{(\nu+t)_{i+m}}{(i+m)!}(-i-m)_m$$

$$=(-1)^{|m|}(\nu+t)_m\sum_{|i|=s+k-|m|}\frac{(\nu+t+m)_i}{i!}$$

$$=(-1)^{|m|}(\nu+t)_m\frac{(|\nu|+|t|+|m|)_{s+k-|m|}}{(s+k-|m|)!},$$

where the last step is the multivariable Chu–Vandermonde formula. This gives

$$\sum_{|i|=s+k}\frac{(\nu)_i}{i!}(-i)_m\,TQ(i)=(-1)^{|m|}(\nu)_m\frac{(|\nu|+s+|m|)_{s+k-|m|}}{(s+k-|m|)!}\sum_{|t|=s}c_t\frac{(\nu+m)_t}{(\nu)_t},$$

which, again by (23), proves the lemma. □

Applying this lemma gives the following generalization of Theorem 3.1.

Theorem 5.5: *In the notation above, the equality*

$$\frac{1}{k_1!\,k_2!\,k_1'!\,k_2'!\,C_{s_3}^{\mu_1\mu_2}C_{s_3'}^{\mu_1'\mu_2'}}\langle(Q_1,Q_2)_{s_3}^{\sigma},(Q_1',Q_2')_{s_3'}^{\tau}\rangle\,P_{s_3}^{(\mu_1,\mu_2)}(k_1,k_2)P_{s_3'}^{(\mu_1',\mu_2')}(k_1',k_2')$$

$$=(-1)^s\frac{(-s_1-s_2-k_1-k_2)_s(-s_1'-s_2'-k_1'-k_2')_s}{(|\nu|+s-1)_s s!}$$

$$\times\sum_{m_1,\dots,m_n}\frac{(|\nu|+s-1)_{|m|}(-s)_{|m|}(\nu)_m}{(-s_1-s_2-k_1-k_2)_{|m|}(-s_1'-s_2'-k_1'-k_2')_{|m|}m!}$$

$$\times\frac{1}{(k_1+s_1-|m_1^{\sigma}|)!(k_2+s_2-|m_2^{\sigma}|)!(k_1'+s_1'-|m_1^{\tau}|)!(k_2'+s_2'-|m_2^{\tau}|)!}$$

$$\times\frac{1}{(\mu_1)_{|m_1^{\sigma}|-s_1}(\mu_2)_{|m_2^{\sigma}|-s_2}(\mu_1')_{|m_1^{\tau}|-s_1'}(\mu_2')_{|m_2^{\tau}|-s_2'}}$$

$$\times TQ_1(m_1^{\sigma})\,TQ_2(m_2^{\sigma})\,\overline{TQ_1'(m_1^{\tau})\,TQ_2'(m_2^{\tau})},$$

holds, where the sum is over all non-negative integers such that

$$s_1\leqslant|m_1^{\sigma}|\leqslant s_1+k_1,$$
$$s_2\leqslant|m_2^{\sigma}|\leqslant s_2+k_2,$$
$$s_1'\leqslant|m_1^{\tau}|\leqslant s_1'+k_1',$$
$$s_2'\leqslant|m_2^{\tau}|\leqslant s_2'+k_2'.$$

We remark that for Wigner $6j$-symbols, that is for the recoupling coefficient,

$$\langle((1,1)_q,1)_r,(1,(1,1)_t)_u\rangle,$$

then in certain cases such as

$$k_1 = k_2' = 0, \quad k_2 = r, \quad k_1' = u,$$

the sum in the theorem reduces to a single hypergeometric sum, which is in fact a terminating balanced ${}_4F_3(1)$-series. This gives a new proof of Racah's single sum formula for $6j$-symbols.

[1] L. Biedenharn and J. Louck, *Angular Momentum in Quantum Physics* (Addison–Wesley, Reading, MA, 1981).
[2] L. Biedenharn and J. Louck, *The Racah–Wigner Algebra in Quantum Theory* (Addison–Wesley, Reading, MA, 1981).
[3] S. J. Ališauskas and A. P. Jucys, ''Weight lowering operators and the multiplicity-free isoscalar factors of the group R_5,'' J. Math. Phys. **12**, 594–605 (1971).
[4] K. Srinivasa Rao, V. Rajeswari, and C. B. Chiu, ''A new Fortran program for the 9-j angular momentum coefficient,'' Comput. Phys. Commun. **56**, 231–248 (1989).
[5] K. Srinivasa Rao and V. Rajeswari, ''Polynomial zeros of the 9-j coefficient,'' J. Phys. A **21**, 4255–4264 (1988).
[6] K. Srinivasa Rao, S. N. Pitre, and J. Van der Jeugt, ''The polynomial zeros of degree 2 of the 9-j coefficient,'' Rev. Mex. Fís. **42**, 179–192 (1996).
[7] J. Van der Jeugt, S. N. Pitre, and K. Srinivasa Rao, ''Multiple hypergeometric functions and 9-j coefficients,'' J. Phys. A **27**, 861–876 (1994).
[8] S. N. Pitre and J. Van der Jeugt, ''Transformation and summation formulas for Kampé de Fériet series $F_{1:1}^{0:3}(1,1)$,'' J. Math. Anal. Appl. **202**, 121–132 (1996).
[9] J. Van der Jeugt, S. N. Pitre, and K. Srinivasa Rao, ''Transformation and summation formulas for double hypergeometric series,'' J. Comput. Appl. Math. **83**, 185–193 (1997).
[10] A. P. Jucys and A. A. Bandzaitis, *Theory of Angular Momentum in Quantum Mechanics*, 2nd ed. (Vilnius, City, 1977) (in Russian).
[11] H. Rosengren, ''Multivariable orthogonal polynomials and coupling coefficients for discrete series representations,'' SIAM J. Math. Anal. (to appear).
[12] H. Rosengren, ''Multilinear Hankel forms of higher order and orthogonal polynomials,'' Math. Scand. **82**, 53–88 (1998).
[13] R. Koekoek and R. F. Swarttouw, ''The Askey scheme of hypergeometric orthogonal polynomials and its q-analogue,'' Delft University of Technology, 1998, available at `ftp://unvie6.un.or.at/siam/opsf/koekoek_swarttouw.`
[14] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge University Press, Cambridge, 1990).
[15] P. W. Karlsson, ''Hypergeometric functions with integral parameter differences,'' J. Math. Phys. **12**, 270–271 (1971).
[16] D. Zagier, ''Introduction to modular forms,'' in *From Number Theory to Physics* (Springer-Verlag, Berlin, 1992), pp. 238–291.
[17] P. B. Cohen, Yu. Manin, and D. Zagier, ''Automorphic pseudodifferential operators,'' in *Algebraic Aspects of Integrable Systems* (Birkhäuser, Boston, 1997), pp. 17–47.
[18] K. Srinivasa Rao and V. Rajeswari, ''A note on the triple sum series for the 9j coefficient,'' J. Math. Phys. **30**, 1016–1017 (1989).
[19] H. M. Srivastava, ''Generalized Neumann expansions involving hypergeometric functions,'' Proc. Cambridge Philos. Soc. **63**, 425–429 (1967).

Lie symmetry and integrability of ordinary differential equations

R. Z. Zhdanov[a)]
Institute of Mathematics, 3 Tereshchenkivska Street, 252004 Kyiv, Ukraine

(Received 21 January 1998; accepted for publication 8 July 1998)

Combining a Lie algebraic approach that is due to Wei and Norman [J. Math. Phys. **4**, 475 (1963)] and the ideas suggested by Drach [Compt. Rend. **168**, 337 (1919)], we have constructed several classes of systems of linear ordinary differential equations that are integrable by quadratures. Their integrability is ensured by integrability of the corresponding stationary cubic Schrödinger, KdV, and Harry–Dym equations. Next, we obtain a hierarchy of integrable reductions of the Dirac equation of an electron moving in the external field. Their integrability is shown to be in correspondence with integrability of the stationary mKdV hierarchy. © *1998 American Institute of Physics.* [S0022-2488(98)04311-4]

I. INTRODUCTION

The object of the study in the present paper is the system of first-order ordinary differential equations (ODEs) of the following structure:

$$\frac{d\psi}{dx}+L(x)\psi=0. \tag{1}$$

Here $\psi(x)$ is an n-component real-valued function column and $L(x)$ is an $n\times n$ matrix function taking values in some real r-dimensional matrix Lie algebra g, namely,

$$L(x)=\sum_{k=1}^{r} f_k(x)Q_k, \tag{2}$$

where Q_k are constant $n\times n$ matrices fulfilling the commutation relations,

$$[Q_i,Q_j]=\sum_{k=1}^{r} C_{ij}^k Q_k. \tag{3}$$

The general Lie algebraic approach to integrating systems of ordinary differential equations having the structure (1) has been developed by Wei and Norman.[1,2] The principal idea of their approach is a proper utilization of the Baker–Campbell–Hausdorff formula in order to represent the general solution of (1) in the form

$$\psi(x)=\exp\left\{\sum_{i=1}^{r} l_i(x)Q_i\right\}\chi, \tag{4}$$

where χ is an arbitrary constant n-component column.

However, with all its elegance and simplicity the approach is still too general and gives a kind of an existence theorem of a special representation of the general solution. The principal problem is that after representing the general solution of (1) in the form (4), we have to integrate a system of *nonlinear* ODEs for unknown functions $l_1(x),\ldots,l_r(x)$. On the other hand, there is the well-known inverse scattering method which reduces a problem of finding special solutions of nonlinear partial differential equations to an auxiliary linear problem (see, e.g., Refs. 3–5). So it is only

a)Electronic mail: rzhdanov@apmat.freenet.kiev.ua

 © 1998 American Institute of Physics

natural to make an attempt of combining both approaches in order to develop a routine for choosing special subclasses of ODEs (1) that can be reduced to solving some nonlinear (stationary) solitonic equations. Namely, we pick out specific subclasses of equations of the form (1) such that their integrability by quadratures is ensured by integrability of the corresponding stationary solitonic equations. To this end we consider systems of ODEs of the following special form:

$$\mathcal{L}\psi \equiv \left(\frac{d}{dx} + \lambda F(x) + G(x) \right) \psi = 0, \tag{5}$$

where $F(x)$, $G(x)$ are matrix functions taking values in a real r-dimensional matrix Lie algebra g. The parameter λ may be thought of as an auxiliary independent variable, i.e., $\psi = \psi(x,\lambda)$. Note that the functions $F(x)$, $G(x)$ are independent of λ.

Systems of ODEs (1) are of the utmost importance for the separation of variables in systems of partial differential equations (say, in the Dirac equation of an electron). This is due to the fact that after separating variables in a given system of partial differential equations, we have to integrate separated equations that are exactly of the form (1).

Since we deal with more specific models as compared with ones considered by Wei and Norman, it is possible to input more information into the ansatz (4). To this end we make use of its ''infinitesimal analog'' in the form of the Lie symmetry of the system of ODEs (5), whose coefficients are required to be polynomials in λ. The last restriction is crucial and provides a link of the system of ODEs (5) to stationary solitonic equations whose integrability is ensured by the inverse scattering method. As we learned recently, a similar idea was used about 80 years ago by Drach[6] in order to integrate by quadratures the one-dimensional stationary Schrödinger equation,

$$\psi'' - (\lambda + V(x))\psi = 0.$$

He had made an ansatz for a solution of the above equation by a proper fixing of its dependence on λ. As a result, he discovered a remarkable class of integrable stationary Schrödinger equations. Namely, it had been proved that, provided $V(x)$ is a solution of the nonlinear ODE called now the equation of the stationary KdV hierarchy, the stationary Schrödinger equation is integrable by quadratures. Moreover, Drach, in fact, suggested the method for integrating the stationary KdV hierarchy and discovered in this way the basics of the theory of the finite-gap solutions of the KdV equation. As shown in Ref. 7 the results of Ref. 6 can be obtained with the use of Lie symmetry of the stationary Schrödinger equation. One of the aims of the present paper is developing a technique that is a proper synthesis of the methods by Drach and Wei and Norman in order to reduce the problem of integrability by quadratures of systems of ODEs (5) to integrable solitonic hierarchies. A principal tool to be used in this respect is the Lie symmetry of (5).

The paper is organized as follows. The second section is devoted to a description of our approach to integrating systems of ODEs of the form (5). The approach is based on the utilization of symmetry properties of these systems within the class of Lie symmetries that are spanned by the basis elements of the Lie algebra g, with coefficients being polynomials in parameter λ. The problem of constructing such Lie symmetries is shown to be reduced to integrating some systems of nonlinear ODEs. Remarkably, for many interesting cases these ODEs are nothing more than equations of the stationary solitonic hierarchies (which is not unexpected in view of what was said above). Utilizing this approach we give in Sec. III, a systematic treatment of the case when the Lie algebra g is three dimensional. In this way, we find a number of linear systems of ODEs integrable in quadratures due to integrability of the stationary cubic Schrödinger, KdV, and Harry–Dym equations. Section IV is devoted to an analysis of a reduction of the Dirac equation for a particle moving in a specific electric field that is integrable with the help of the suggested procedure. It occurs that the reduced system of ODEs is integrable provided the nonzero component of the electromagnetic field satisfy one of the equations of the stationary mKdV hierarchy. The last section contains a brief discussion of the results of the paper.

II. THE GENERAL SCHEME

We are reminded that a Lie vector field,

$$X=\xi(x,\lambda)\,\frac{d}{dx}+\eta(x,\lambda),$$

where ξ is a smooth scalar function and η is an $n\times n$ matrix whose entries are smooth functions of x, is called the (Lie) symmetry of system (5) if X transforms the set of its solutions into itself, i.e.,

$$\mathcal{L}\psi=0\Rightarrow\mathcal{L}X\psi=0.$$

The above relation can be represented in the form[8]

$$[\mathcal{L},X]=R(x,\lambda)\mathcal{L}, \tag{6}$$

where R is some $n\times n$ matrix function. This operator equality is to be understood in the following way: the operators on the left- and right-hand sides should give the same result when acting on any continuously differentiable function.

Making use of the formula (6) it is easy to become convinced of the fact that if X is a symmetry of the system of ODEs (5), then $X+\Xi(x,\lambda)\mathcal{L}$ with an arbitrary smooth function Ξ is a symmetry as well. Consequently, without a loss of generality we can restrict our considerations to symmetries of the form

$$X=\eta(x,\lambda). \tag{7}$$

Importantly, operator (7) is the symmetry of the system (5) if and only if in (6) $R=0$, i.e.,

$$[\mathcal{L},X]=0 \tag{8}$$

(this is proved by direct computation).

The key idea of our approach is to fix *a priori* dependence of a symmetry on the parameter λ and to consider the case when η is a polynomial in λ of the order N with matrix coefficients.

So a Lie symmetry of system (5) is looked for as a polynomial in λ which coefficients are linear combinations of the basis elements $Q_1,\ldots,Q_r$ of the Lie algebra g,

$$X=\sum_{l=1}^{N}\sum_{k=1}^{r}s_{kl}(x)Q_k\lambda^l, \tag{9}$$

where $s_{kl}(x)$ are sufficiently smooth functions. From the invariance criterion (8) we get the following relation:

$$\sum_{l=1}^{N}\sum_{k=1}^{r}s'_{kl}Q_k\lambda^l+\sum_{l=1}^{N}\left\{\sum_{i,j,k=1}^{r}(f_is_{jl}C_{ij}^kQ_k+\lambda g_is_{jl}C_{ij}^kQ_k)\right\}\lambda^l=0.$$

Splitting it with respect to the powers of λ yields

$$\lambda^{N+1}:\quad\sum_{i,j=1}^{r}g_is_{jN}C_{ij}^k=0, \tag{10}$$

$$\lambda^{l}:\quad s'_{kl}+\sum_{i,j=1}^{r}(f_is_{jl}+g_is_{jl-1})C_{ij}^k=0, \tag{11}$$

$$\lambda^{0}:\quad s'_{k0}+\sum_{i,j=1}^{r}f_is_{j0}C_{ij}^k=0, \tag{12}$$

where $k=1,...,r,\ l=1,...,N$.

Equations (10) are purely algebraic. Solving these, we obtain recursively from Eqs. (11) the remaining coefficients of the symmetry operator X. Inserting the obtained results into (12), we get a system of relations for functions f_i, g_i that form a system of nonlinear ODEs. The structure of this system is determined both by the form of the initial system of ODEs and by the form of the commutation relations of the Lie algebra g.

The next step depends on the kind of the problem we are dealing with. If the problem is to check whether a given system of ODEs is integrable within the framework of our approach, then the only thing to be done is to verify whether the obtained system of ODEs is identically satisfied by the coefficients of system (5). If, on the contrary, we have to solve a classification problem, i.e., the one of describing functions f_i, g_i such that system (5) is integrable, then we have to find a (general or particular) solution of the mentioned system of nonlinear ODEs. What makes the whole procedure efficient is the fact that the nonlinear ODEs obtained are often the well-studied stationary solitonic equations.

Now with a Lie symmetry of system of ODEs (5) in hand, we can integrate it with the use of the following procedure. We diagonalize the operator X with the help of a properly chosen linear transformation of the dependent variables. The initial system of ODEs (5) being transformed in this way simplifies substantially and can be integrated by quadratures (at least for the low-dimensional Lie algebras).

The above procedure proves to be efficient, not only for integrating specific ODEs, but also for classification of systems of ODEs integrable by quadratures.

III. INTEGRABLE ODEs

We apply the above described method to classify integrable systems of ODEs of the form

$$\mathcal{L}\psi \equiv \left(\frac{d}{dx} + \sum_{a=1}^{3} (f_a(x) + \lambda g_a(x)) Q_a \right) \psi = 0, \tag{13}$$

where Q_1, Q_2, Q_3 are basis elements of a real three-dimensional Lie algebra g. It is not difficult to show that if a three-dimensional Lie algebra g is a direct sum of the lower-dimensional Lie algebras, then the corresponding system of ODEs (13) is integrated by quadratures. According to Ref. 9, the list of real inequivalent Lie algebras of the dimension three that are not direct sums of lower-dimensional Lie algebras is exhausted by the following algebras:

$$A_1: [Q_2, Q_3] = 0,$$

$$A_2: [Q_1, Q_3] = Q_1, \quad [Q_2, Q_3] = Q_1 + Q_2,$$

$$A_3: [Q_1, Q_3] = Q_1, \quad [Q_2, Q_3] = Q_2,$$

$$A_4: [Q_1, Q_3] = Q_1, \quad [Q_2, Q_3] = -Q_2,$$

$$A_5: [Q_1, Q_3] = Q_1, \quad [Q_2, Q_3] = aQ_2 \quad (0 < |a| < 1),$$

$$A_6: [Q_1, Q_3] = -Q_2, \quad [Q_2, Q_3] = Q_1,$$

$$A_7: [Q_1, Q_3] = aQ_1 - Q_2, \quad [Q_2, Q_3] = Q_1 + aQ_2 \quad (a > 0),$$

$$A_8: [Q_1, Q_2] = -Q_3, \quad [Q_1, Q_3] = -Q_2, \quad [Q_2, Q_3] = Q_1,$$

$$A_9: [Q_1, Q_2] = Q_3, \quad [Q_2, Q_3] = Q_1, \quad [Q_3, Q_1] = Q_2,$$

the remaining commutation relations being zero.

For the algebras $A_1 - A_7$ system of ODEs (13) is integrated by quadratures with arbitrary f_a, g_a. This is due to the fact that making a transformation,

$$\psi \to \tilde{\psi} = \exp\{F(x)Q_1 + G(x)Q_2\},$$

with properly chosen functions F,G, we can always reduce (13) to the system of ODEs of the form

$$\frac{d\tilde{\psi}}{dx}+f(x,\lambda)Q_3\tilde{\psi}=0,$$

whose general solution is given by the quadrature

$$\tilde{\psi}=\exp\left\{-Q_3\int f(x,\lambda)dx\right\}\chi,$$

where χ is an arbitrary constant column.

To complete the classification of systems of ODEs (13) integrable by quadratures within the framework of the above suggested approach, we have to consider the two remaining algebras $A_8=so(2,1)$ and $A_9=so(3)$. Both algebras lead to nontrivial results, even if we restrict our considerations to Lie symmetries that are second-order polynomials in λ.

We consider first the case when coefficients of system (13) take values in the $so(2,1)$ algebra. Using a sequence of transformations,

$$\psi\to\tilde{\psi}=\exp\{\alpha(x)Q\}, \tag{14}$$

where Q is one of the operators Q_1,Q_2,Q_3, together with a transformation of the independent variable,

$$x\to\tilde{x}=f(x),$$

we can reduce (13) to the form $\tilde{\mathcal{L}}\tilde{\psi}=0$ with

$$\tilde{\mathcal{L}}=\begin{cases}\dfrac{d}{dx}+V_2(x)Q_2+V_3(x)Q_3+\lambda Q_1, & \text{if } \Delta>0,\\ \dfrac{d}{dx}+V_1(x)Q_1+V_2(x)Q_2+\lambda Q_3, & \text{if } \Delta<0,\\ \dfrac{d}{dx}+V_1(x)Q_1+V_2(x)(Q_2+Q_3)+\lambda(Q_2-Q_3), & \text{if } \Delta=0.\end{cases}$$

Here Δ stands for $V_1^2+V_2^2-V_3^2$.

We restrict our analysis of symmetries of the above enumerated systems to the class of the second-order polynomials in λ. This means that symmetries of the systems of ODEs in question are looked for in the form (9) under $N=2$. Omitting the details of calculations we present below the explicit forms of symmetries and corresponding nonlinear ODEs for the ''potentials'' $V_1(x)$, $V_2(x)$.

Case 1. $\tilde{\mathcal{L}}=d/dx+V_2(x)Q_2+V_3(x)Q_3+\lambda Q_1$.

In this case the symmetry X reads as

$$X=\lambda^2C_1Q_3+\lambda C_2Q_1+\lambda C_1(V_2(x)Q_2+V_3(x)Q_3)+(\tfrac{1}{2}C_1(V_2^2-V_3^2)+C_3)Q_1$$
$$+(C_1V_3'+C_2V_2)Q_2+(C_1V_2'+C_2V_3)Q_3,$$

and what is more, the functions V_2, V_3 satisfy the hyperbolic stationary cubic Schrödinger equation,

$$C_1V_2''+C_2V_3'+(\tfrac{1}{2}C_1(V_2^2-V_3^2)+C_3)V_2=0,$$

$$C_1V_3''+C_2V_2'+(\tfrac{1}{2}C_1(V_2^2-V_3^2)+C_3)V_3=0.$$

As usual, when talking about stationary solitonic equations, we mean the ODEs obtained from standard $(1+1)$-dimensional solitonic equations via the ansatz $u(t,x)=u(x+Ct)$, $C=\text{const}$.

Case 2. $\tilde{\mathcal{L}}=d/dx+V_1(x)Q_1+V_2(x)Q_2+\lambda Q_3$.

In this case the operator X has the form

$$X=\lambda^2 C_1 Q_1+\lambda C_1 V_1 Q_1+\lambda C_1 V_2(x) Q_2+\lambda C_2 Q_3+(-C_1 V_2'+C_2 V_1) Q_1$$
$$+(C_1 V_1'+C_2 V_2) Q_2+(\tfrac{1}{2} C_1(V_1^2+V_2^2)+C_3) Q_3,$$

and what is more, the functions V_1, V_2 satisfy the stationary cubic Schrödinger equation,

$$C_1 V_1''+C_2 V_2'-(\tfrac{1}{2} C_1(V_1^2+V_2^2)+C_3) V_1=0,$$

$$C_1 V_2''-C_2 V_1'-(\tfrac{1}{2} C_1(V_1^2+V_2^2)+C_3) V_2=0.$$

Case 3. $\widetilde{\mathcal{L}}=d/dx+V_1(x) Q_1+V_2(x)(Q_2+Q_3)+\lambda(Q_2-Q_3)$.

Provided $V_2=0$, the corresponding system of ODEs is integrable by quadratures with an arbitrary V_1. If $V_2\neq 0$, then we can transform the operator $\widetilde{\mathcal{L}}$ to obtain

$$\frac{d}{dx}+V_1(x)(Q_2+Q_3)+V_2(x)(Q_2-Q_3)+\lambda(Q_2-Q_3).$$

We have not succeeded in constructing the general form of V_1, V_2 and therefore restrict our considerations to the particular cases (i) $V_1=\text{const}$, (ii) $V_2=0$.

Subcase 3.1. $V_1(x)=\alpha\equiv\text{const}$.

The Lie symmetry for this case reads as

$$X=\lambda^2(Q_2-Q_3) C_1+\lambda(\alpha C_1+C_2+\tfrac{1}{2} C_1 V_2) Q_2+\lambda(\alpha C_1-C_2-\tfrac{1}{2} C_1 V_2) Q_3$$
$$+\frac{1}{2} C_1 V_2' Q_1+\left(\alpha C_2+\left(C_2-\frac{1}{2}\alpha C_1\right) V_2-\frac{1}{2} C_1 V_2^2+\frac{1}{2\alpha} C_1 V_2''\right) Q_2$$
$$+\left(\alpha C_2-\left(C_2+\frac{1}{2}\alpha C_1\right) V_2+\frac{1}{2} C_1 V_2^2+\frac{1}{2\alpha} C_1 V_2''\right) Q_3,$$

the function $V_1(x)$ being a solution of the stationary KdV equation,

$$\frac{1}{2\alpha} C_1 V_2'''-3 C_1 V_2' V_2+2 C_2 V_2'=0.$$

Subcase 3.2. $V_2(x)=0$.

The Lie symmetry for this case is of the form

$$X=\lambda^2 C_1 V_1^{-1/2}(Q_2-Q_3)+\tfrac{1}{2}\lambda C_1 V_1' V_1^{-3/2} Q_1+\lambda(C_2 x+C_3+C_1 V_1^{1/2}) Q_2$$
$$-\lambda(C_2 x+C_3-C_1 V_1^{1/2}) Q_3+C_2 Q_1+\{(C_2 x+C_3) V_1+\tfrac{1}{8} C_1(2 V_1'' V_1^{-3/2}-3(V_1')^2 V_1^{-5/2})\} Q_2$$
$$+\{(C_2 x+C_3) V_1+\tfrac{1}{8} C_1(2 V_1'' V_1^{-3/2}-3(V_1')^2 V_1^{-5/2})\} Q_3,$$

the function $V_1(x)$ satisfying the nonlinear ODE,

$$\tfrac{1}{2} C_1 V_1''' V_1^{-3/2}-\tfrac{9}{4} C_1 V_1'' V_1' V_1^{-5/2}+\tfrac{15}{8} C_1(V_1')^3 V_1^{-7/2}+2(C_2 x+C_3) V_1'+4 C_2 V_1=0.$$

With $C_2=0$, the above equation is nothing other than the stationary Harry Dym equation,

$$-C_1(V_1^{-1/2})'''+2 C_3 V_1'=0,$$

which is known to be integrable by quadratures and, furthermore, under $C_3=0$ possesses solutions in terms of elementary functions,

$$V_1(x)=(\alpha_1 x^2+\alpha_2 x+\alpha_3)^{-2},$$

where α_1, α_2, α_3 are arbitrary real constants.

Now we turn to the case of the algebra $A_9=so(3)$ and consider a system of ODEs (13), where Q_1, Q_2, Q_3 are constant matrices satisfying the commutation relations

$$[Q_a,Q_b]=Q_c, \quad (a,b,c)=\text{cycle}(1,2,3).$$

Using a sequence of transformations (14), where Q is one of the operators Q_1, Q_2, Q_3, we can substantially simplify system (13) and reduce it to the following equivalent form:

$$\mathcal{L}\psi\equiv\left(\frac{d}{dx}+V_1(x)Q_1+V_2(x)Q_2+\lambda Q_3\right)\psi=0. \tag{15}$$

As above, a Lie symmetry X of system (15) is looked for as a second-order polynomial in λ with matrix coefficients. Inserting X into the invariance criterion $[\mathcal{L},X]=0$ yields a system of determining equations for its coefficients. Having solved these, we obtain the explicit form of the Lie symmetry admitted by a system of ODEs (15),

$$X=\lambda^2C_1Q_3+\lambda C_1(V_1Q_1+V_2Q_2)+\lambda C_2Q_3+(-C_1V_2'+C_2V_1)Q_1$$
$$+(C_1V_1'+C_2V_2)Q_2+\left(\frac{1}{2}(V_1^2+V_2^2)+C_3\right)Q_3,$$

where $V_1(x)$, $V_2(x)$ are solutions of the stationary cubic Schrödinger equation,

$$C_1V_1''-C_2V_2'-\left(\frac{1}{2}C_1(V_1^2+V_2^2)+C_3\right)V_1=0,$$

$$C_1V_2''+C_2V_1'-\left(\frac{1}{2}C_1(V_1^2+V_2^2)+C_3\right)V_2=0,$$

and C_1, C_2, C_3 are arbitrary constants.

In what follows, we will briefly discuss a method for integrating systems of ODEs (13), (15) based on their symmetry properties. Generally speaking, information about a Lie symmetry admitted by a system of ODEs is not sufficient to provide its integrability by quadratures within the Lie group approach (for more details see, e.g., Refs. 10–13). However, due to the remarkable algebraic structure of the ODEs under study, knowledge of a Lie symmetry makes it possible to construct their general solutions by quadratures. Namely, the following assertions hold true.

Lemma 1: Let the system of ODEs,

$$\mathcal{L}\psi\equiv\left(\frac{d}{dx}+f_a(x)Q_a\right)\psi=0, \tag{16}$$

where Q_1, Q_2, Q_3 are constant matrices forming a basis of the Lie algebra so(2,1), admit a Lie symmetry,

$$X=\sum_{a=1}^{3}g_a(x)Q_a. \tag{17}$$

Then it is integrable by quadratures.

Lemma 2: Let the system of ODEs (16), where Q_1, Q_2, Q_3 are constant matrices forming a basis of the Lie algebra so(3), admit a Lie symmetry (17) Then it is integrable by quadratures.

We adduce the proof of Lemma 2 (the first lemma is proved in a similar way). Making a change of dependent variables,

$$\psi\to\tilde{\psi}=\mathcal{V}(x)\psi, \quad \mathcal{V}(x)=\exp\left\{\sum_{a=1}^{3}h_a(x)Q_a\right\},$$

we can always transform the operator X to become

$$\tilde{X}=\mathcal{V}^{-1}X\mathcal{V}=g(x)Q_1, \quad g(x)\neq 0,$$

and what is more, this transformation preserves the structure of system (16). The invariance criterion $[\tilde{\mathcal{L}},\tilde{X}]=0$, where

$$\tilde{\mathcal{L}}=\mathcal{V}^{-1}\mathcal{L}\mathcal{V}=\frac{d}{dx}+\sum_{a=1}^{3}\tilde{f}_a(x)Q_a,$$

implies that

$$g(\tilde{f}_2Q_3-\tilde{f}_3Q_2)-g'Q_1=0.$$

As the matrices Q_1, Q_2, Q_3 are linearly independent, hence it follows that

$$\tilde{f}_2=0, \quad \tilde{f}_3=0, \quad g=\text{const}.$$

Consequently, the transformed system of ODEs necessarily takes the form

$$\left(\frac{d}{dx}+\tilde{f}_1Q_1\right)\psi=0,$$

and is evidently integrable by quadratures. Lemma 2 is proved.

Consequently, all the systems of linear ODEs considered in Sec. III possessing nontrivial symmetries of the form (9) can be integrated by quadratures with the use of the algebraic procedure described above. This is due to the well-known fact that the stationary solitonic equations arising as the invariance conditions are integrable by quadratures (see, e.g., Refs. 14, 15).

IV. THE HIERARCHY OF INTEGRABLE REDUCTIONS OF THE DIRAC EQUATION

The technique developed above applies straightforwardly to systems of ODEs having complex-valued coefficients. In this section we use this technique for obtaining integrable reductions of the Dirac equation of an electron,

$$i\sum_{\mu=0}^{3}\gamma_\mu\psi_{x_\mu}-\left(e\sum_{\mu=0}^{3}\gamma_\mu A^\mu+m\right)\psi=0, \tag{18}$$

moving in the electric field,

$$A_0=A_0(x_3), \quad A_1=A_2=A_3=0.$$

In the formulas (18) γ_μ are 4×4 Dirac matrices, $\psi=\psi(x_0,x_1,x_2,x_3)$ is a four-component complex-valued function, and e, m are constants.

The form of the vector-potential A_μ imply the following ansatz for the spinor field $\psi(x)$:

$$\psi(x)=\varphi(x_3).$$

Inserting this expression into the Dirac equation (18) yields a system of ordinary differential equations for the four-component function $\varphi(x_3)$,

$$\varphi'-(ie\gamma_3\gamma_0A_0-im\gamma_3)\varphi=0. \tag{19}$$

Denoting

$$x=2x_3, \quad V(x)=eA_0(x_3),$$

$$J_1=\frac{1}{2}\gamma_0 \quad J_2=\frac{i}{2}\gamma_3, \quad J_3=\frac{i}{2}\gamma_3\gamma_0,$$

we rewrite (19) in the following form:

$$\mathcal{L}\varphi\equiv(D_x-V(x)J_3-mJ_2)\varphi=0,\quad D_x=\frac{d}{dx}. \tag{20}$$

Note that the 4×4 matrices J_1, J_2, J_3 satisfy the commutation relations of the Lie algebra $so(3)$. Consequently, we can apply the routine developed in Sec. I in order to construct a Lie symmetry admitted by (20).

First we consider the case when coefficients of a Lie symmetry Q are third-order polynomials in m, and choose

$$X=\sum_{k=1}^{3}(a_k(x)+b_k(x)m+c_k(x)m^2+d_k(x)m^3)J_k, \tag{21}$$

where a_k, b_k, c_k, d_k are some smooth complex-valued functions.

Inserting the expression for X into the invariance criterion $[\mathcal{L},X]=0$ and splitting with respect to the powers of m and then with respect to linearly independent matrices J_1, J_2, J_3, we get the system of determining equations for the functions a_k, b_k, c_k, d_k,

$$d_1=0,\quad d_3=0,$$

$$d_2'-Vd_1=0,\quad d_1'+Vd_2-c_3=0,\quad d_3'-c_1=0,$$

$$c_2'-Vc_1=0,\quad c_1'+Vc_2-b_3=0,\quad d_3'-b_1=0,$$

$$b_2'-Vb_1=0,\quad b_1'+Vb_2-a_3=0,\quad c_3'-a_1=0,$$

$$a_2'-Va_1=0,\quad a_1'+Va_2=0,\quad a_3'=0.$$

Integrating the above system of ODEs yields

$$d_1=0,\quad d_2=C_1,\quad d_3=0,$$

$$c_1=0,\quad c_2=C_2,\quad c_3=C_1V,$$

$$b_1=C_1V',\quad b_2=\frac{1}{2}C_1V^2+C_3,\quad b_3=C_2V,$$

$$a_1=C_2V',\quad a_2=\frac{1}{2}C_2V^2+C_4,\quad a_3=C_1\left(V''+\frac{1}{2}V^3\right)+C_3V,$$

where C_1, C_2, C_3, C_4 are arbitrary constants and, furthermore, the potential $V(x)$ has to satisfy the following nonlinear ODEs:

$$C_1(V'''+\tfrac{3}{2}V^2V')+C_3V'=0,\quad C_2(V''+\tfrac{1}{2}V^3)+C_4V=0.$$

Thus, we have established that if the function $V(x)$ is a solution of the stationary mKdV equation,

$$C_1(V'''+\tfrac{3}{2}V^2V')+C_3V'=0, \tag{22}$$

then the initial system of ODEs (20) admits the Lie symmetry,

$$Q=C_1J_2m^3+C_1VJ_3m^2+C_1V'J_1m+(\tfrac{1}{2}C_1V^2+C_3)J_2m+(C_1(V''+\tfrac{1}{2}V^3)+C_3V)J_3.$$

This symmetry solves the problem of the integrability of the system of ODEs (20) by quadratures due to Lemma 2. Hence, we conclude that provided $V(x)$ is a solution of the stationary mKdV equation (22), then the system of ODEs (20) is integrable by quadratures.

Now we turn to the case when a Lie symmetry is looked for as a polynomial in m of an arbitrary order n,

$$X=\sum_{k=0}^{n}\sum_{a=1}^{3} f_a^k(x)J_a m^{n-k}.$$

The invariance criterion $[\mathcal{L},X]=0$ yields the following system of determining equations for the coefficients of the operator Q:

$$f_1^0=0, \quad f_3^0=0,$$

$$(f_3^k)'-f_1^{k+1}=0, \quad (f_1^k)'-Vf_2^k=0,$$

$$(f_2^k)'+Vf_1^k-f_3^{k+1}=0, \quad k=0,...,n-1,$$

$$(f_3^n)'=0, \quad (f_1^n)'-Vf_2^n=0, \quad (f_2^n)'+Vf_1^n=0.$$

We have obtained the two classes of solutions of the above system of ODEs that are given below

(1) $n=2N+1, \quad N\in\mathbf{N}$,

$$f_1^0=0, \quad f_2^0=1, \quad f_3^0=0,$$

$$f_1^{2k+1}=f_2^{2k+1}=0, \quad f_3^{2k+1}=R_k, \quad k=1,...,N,$$

$$f_1^{2k+2}=D_xR_k, \quad f_1^{2k+2}=(V-D_x^{-1}V')R_k,$$

$$f_3^{2k+2}=0, \quad k=0,...,N-1,$$

and the equation

$$D_xR_N=0, \tag{23}$$

holds.

(2) $n=2N+2, \quad N\in\mathbf{N}$

$$f_1^0=0, \quad f_2^0=1, \quad f_3^0=0,$$

$$f_1^{2k+1}=f_2^{2k+1}=0, \quad f_3^{2k+1}=R_k, \quad k=1,...,N,$$

$$f_1^{2k+2}=D_xR_k, \quad f_2^{2k+2}=(V-D_x^{-1}V')R_k,$$

$$f_3^{2k+2}=0, \quad k=0,...,N,$$

and the equation

$$R_{N+1}=0,$$

holds.

In the above formulas we make use of the following notations:

$$R_k=\sum_{j=0}^{k} C_j(D_x^2+V^2-VD_x^{-1}V')^jV, \quad k=0,...,N+1,$$

where $C_0,\ldots,C_{N+1}$ are arbitrary real constants and D_x^{-1} is the inverse of D_x.

A reader familiar with the soliton theory will immediately recognize the operator $\mathcal{X}=D_x^2+V^2-VD_x^{-1}V_x$ as the recursion operator for the mKdV equation,[5,11]

$$V_t+V_{xxx}+\frac{3}{2}V^2V_x=0.$$

Acting repeatedly with the operator $\mathcal{X}$ on the trivial conserved density $I_0=V$, we get the whole set of conserved densities of the mKdV equation. Next, the operator,

$$\mathcal{Y}=D_x\mathcal{X}D_x^{-1}\equiv D_x^2+V^2+V_xD_x^{-1}V$$

is the second recursion operator for the mKdV equation. Its repeated action on the trivial Lie symmetry $S_0=V_x$ yields the whole hierarchy of the higher symmetries of the mKdV equation. Hence, it follows, in particular, that condition (23) is rewritten in the form

$$\sum_{k=0}^{N}C_kS_k=0,\quad S_k=\mathcal{Y}^kV'. \tag{24}$$

The above equation is nothing other than the higher stationary mKdV equation. Provided $N=1$, it reduces to the standard stationary mKdV equation (22).

Hence, due to Lemma 2, the validity of the following assertion follows.

Theorem 1: *Let the function $V(x)$ satisfy the higher stationary mKdV equation (24) with some fixed N and $C_0,\ldots,C_N$. Then, the system of ODEs (20) is integrable by quadratures.*

It is common knowledge that the stationary mKdV hierarchy is reduced to the stationary KdV hierarchy with the help of the Miura transformation (see, e.g., Refs. 4, 5). Furthermore, the latter are integrated in terms of θ functions.[6,15] Consequently, the system of ODEs (19) is also integrable by quadratures, thus giving rise to exact solutions of the initial Dirac equation (18).

V. CONCLUSIONS

Here we have demonstrated how one can make use of powerful ideas and methods developed in the theory of solitons in order to integrate systems of ODEs of the form (5). Here we have to give its due to the paper by Drach,[6] where almost all the principal ideas necessary for consistent implementation of the technique described in Sec. II can be found (though Drach did not use Lie symmetries explicitly).

One more important remark is that the Lie symmetries of the reduced Dirac equation (19) constructed in Sec. IV have nothing to do with the maximal symmetry group $C(1,3)\otimes U(1)$ admitted by the initial Dirac equation. This means that they cannot be obtained by an appropriate reduction of the group $C(1,3)\otimes U(1)$. These symmetries correspond to conditional symmetry of the Dirac equation (for more details about conditional symmetries of linear and nonlinear Dirac equations, see Ref. 16). We believe that there exist numerous reductions of the Dirac equations that could be integrated in the way described in Sec. IV, thus giving principally new exact solutions to (18) with specific electromagnetic fields. A further possible application of the results of the paper is integrating systems of ODEs obtained after separating variables in the Dirac equation (18).

The next remark is concerned with the restriction of our considerations to the case of three-dimensional Lie algebras. This restriction is not at all essential, and it is quite clear how to include into consideration the higher-dimensional Lie algebras. The peculiarity of the three-dimensional case is that all the systems of ODEs of the form (13) possessing a Lie symmetry (7) are integrable by quadratures. For the case of a higher-dimensional Lie algebra, this might not be the case.

ACKNOWLEDGMENTS

It is a pleasure for the author to thank R. Conte, who has drawn the author's attention to the paper by Jules Drach.[6] Special thanks are addressed to A. Nikitin for stimulating discussions of the results presented in the paper. The author is supported by the Ukrainian DFFD Foundation (Project No. 1.4/356).

[1] J. Wei and E. Norman, J. Math. Phys. **4**, 575 (1963).
[2] J. Wei and E. Norman, Proc. Am. Math. Soc. **15**, 327 (1964).
[3] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevski, *Theory of Solitons: the Inverse Scattering Method* (Consultants Bureau, New York, 1980).
[4] G. L. Lamb, Jr., *Elements of Soliton Theory* (Wiley, New York, 1980).
[5] M. J. Ablovitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[6] J. Drach, Compt. Rend. **168**, 337 (1919).
[7] R. Z. Zhdanov, J. Math. Phys. **37**, 3198 (1996).
[8] W. I. Fushchych and A. G. Nikitin, *Symmetries of Equations of Quantum Mechanics* (Allerton Press, New York, 1994).
[9] G. M. Mubarakzyanov, Izv. Vyssh. Uchebn. Zaved. Mat. **32**, 114 (1963).
[10] G. Bluman and S. Kumei, *Symmetries and Differential Equations* (Springer-Verlag, New York, 1989).
[11] N. Kh. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Reidel, Dordrecht, 1985).
[12] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer Verlag, New York, 1986).
[13] L. V. Ovsjannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[14] B. A. Dubrovin, V. B. Matveev, and S. P. Novikov, Usp. Mat. Nauk **31**, 55 (1976).
[15] S. P. Novikov, Funct. Anal. Appl. **8**, 54 (1974).
[16] W. I. Fushchych and R. Z. Zhdanov, *Symmetries and Exact Solutions of Nonlinear Dirac Equations* (Naukova Ukraina, Kyiv, 1997).